ATHABASCA UNIVERSITY

A PUBLIC METRICS ANALYSIS OF CANADIAN MOBILE ADOPTION AND ITS TRENDS

BY

CODY RAY BAXTER

A THESIS/DISSERTATION

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN COMPUTING AND INFORMATION SYSTEMS

FACULTY OF SCIENCE AND TECHNOLOGY

ATHABASCA, ALBERTA

NOVEMBER 2025

Athabasca University
Faculty of Graduate Studies

Approval of Thesis

The undersigned certify that they have read the thesis entitled

A PUBLIC METRICS ANALYSIS OF CANADIAN MOBILE ADOPTION AND ITS TRENDS

Submitted by

Cody Baxter

In partial fulfillment of the requirements for the degree of

Master of Science in Information Systems

The thesis examination committee certifies that the thesis
and the oral examination is approved

Supervisors:
Dr. Qing Tan
Athabasca University

Dr. Xiaokun Zhang
Athabasca University

External Examiner:
Dr. Kadry El Ezzat Kadry
Higher Technology Institute - Egypt

December 19, 2025

1 University Drive, Athabasca, AB, T9S 3A3 Canada
Toll-free (CAN/U.S.) 1.800.788.9041 ex. 6821
fgs@athabascau.ca | fgs.athabascau.ca | athabascau.ca

Dedication

I would like to thank Dr. Ching Tan for all the unwavering understanding, support and patience throughout this entire dissertation and its many challenges. Without your kind efforts this would never have been possible.

Additionally, I would like to thank and acknowledge all the people who I have spent so many hours over this degree talking about all kinds of things they didn't know or care about, the fact you all still speak to me shows your true strength of character!

Abstract

This dissertation examines the reliability of Canada's telecommunications data ecosystem for modeling the structural determinants of mobile adoption between 2014 and 2018. Public data from Statistics Canada and the CRTC were combined with private carrier disclosures to construct an integrated dataset. A modular, Python-based pipeline was developed to harmonize inconsistent schemas, manage missing data, and ensure reproducibility. The analysis employed econometric and spatial methods to test whether population density and income disparities influenced adoption patterns across provinces. While theoretical relationships were consistent with prior technology adoption research, severe data quality limitations—including interpolation, inconsistent definitions, and imputed financial metrics—undermined model stability. The study's contribution is twofold: it demonstrates the methodological requirements for integrating fragmented public and private datasets, and it provides a critical evaluation of their adequacy for policy-relevant research. Findings underscore the need for standardized, high-resolution, and longitudinal data to support evidence-based telecommunications policy.

Keywords: mobile adoption, Canada, population density, income disparity, telecommunications, digital divide, data pipeline, regression analysis, spatial analysis

Table of Contents

FGS Approval of Thesis i
Dedication iii
Abstract iv
Table of Contents v
List of Tables viii
List of Figures and Illustrations ix
List of Symbols, Nomenclature, or Abbreviations x
Chapter I – Introduction 1
1.1 Research Background 1
1.2 Research Purpose 4
1.3 Research Question & Objectives 5
1.3.1 Hypotheses 5
1.3.2 Research Objectives 6
1.4 Research Methodology 7
1.4.1 Methodological Approaches to Literature Review 8
1.4.2 Maintaining Scope and Strict Timelines in Literature Analysis 9
1.5 Data Processing and Analysis 10
1.6 Research Contribution & Significance 12
1.7 Research Limitation & Scope 14
1.8 Organization of the Thesis 15
Chapter II – Literature Review 20
2.1 African Literature 20
2.2 Limitations of African Data for Comparative Analysis 23
2.3 Canadian Literature 25
Chapter III – Research Design and Methodology 29
3.1 Hypotheses and Research Design 29
H1: Population density significantly influences mobile adoption rates. 29
H2: Income disparity significantly affects mobile adoption rates. 30
3.2 Collecting Data 32
3.3 Pipeline Development 34
3.3.1 Pipeline Motivations 35
3.3.2 Design Principles of the Pipeline 36
3.3.3 Ethical and Epistemological Considerations 37
3.4 Pipeline overview 37
3.4.1 Pipeline Implementations 38
3.4.2 Geographic Data Processing 40
3.4.3 The Combine Data Script 42

3.4.4 The Clean Data Script 43
3.4.5 The Data Preparation Script 44
3.4.6 The Prepare Results Script 44
3.4.7 Provider-Specific Scripts 45

Chapter IV – Data Processing 49
4.1 The Influence of Geography on Adoption 49
4.2 Data Retrieval and Sources 50
4.3 Preprocessing and Cleaning of Geographic Data 54
4.4 Preprocessing and Cleaning of Subscriber Data 55
4.4.1 Technical Implementation Challenges 58
4.5 Data Storage and Management 58
4.5.1 Source Challenges 59
4.5.2 Raw Data Complications 59
4.5.3 Stability and Integration 60
4.5.4 Variable Engineering 62
4.6 Reflection, Challenges, Constraints 62
4.7 Method and Interpretative Complications 64
4.8 Lessons Learned 64

Chapter V – Data Analysis 66
5.1 Overview of Analytical Approach 66
5.2 Geographic Influence on Mobile Adoption 67
5.2.1 Urban-Rural Disparities 68
5.2.2 Spatial Analysis 70
5.3 Income Disparities and Mobile Adoption 71
5.3.1 Correlation and ANOVA 72
5.3.2 Regression Analysis 73
5.4 Market Structure and Financial Analysis 76
5.4.1 ARPU Analysis 79
5.4.2 Financial Performance 80
5.5 Critical Methodological and Data Quality Issues 81
5.5.1 Severe Overfitting 82
5.5.2 Extensive Financial Data Imputation 82
5.5.3 Propensity Score Matching (PSM) Failures 83
5.5.4 Numerical Precision Concerns 84
5.5.5 Artificial Data Balance 84
5.6 Reliability Assessment of Analysis Components 84
5.7 Recommendations for Future Analysis 85
5.8 Conclusion and Final Verdict 85

Chapter VI – Summary & Recommendations 86
6.1 Summary of Findings 86
6.2 Methodological Issues and Limitations 89
6.3 Recommendations for Future Research 91
6.4 Conclusion and Final Verdict 93

Appendices 97
Appendix A: Numerical Warnings and Stability Issues 97

Appendix B: Pipeline Pseudocode..... 99

List of Tables

Table 1 *Candidate African Sources and Reasons for Exclusion* 25
Table 2 *Data Sources and Coverage* 33
Table 3 *Yearly Growth by Region* 68
Table 4 *Moran's I Result by Year* 71
Table 5 *Correlation Matrix of Key Variables* 73
Table 6 *ANOVA Results for Income Groups* 73
Table 7 *Centered covariates* 75
Table 8 *Regression Coefficients with Confidence Intervals* 77
Table 9 *Provider Market Share By Year (2014-2018)* 78
Table 10 *Average Revenue Per User (ARPU) Trends* 79
Table 11 *Major Provider Subscriber Growth (2014-2018, millions)* 81
Table 12 *PSM Covariate Balance Assessment* 83

List of Figures and Illustrations

Figure 1 *Dissertation Phases* 19
Figure 2 *Analytical Framework* 31
Figure 3 *Data Processing and Analysis Pipeline* 34
Figure 4 *Data Lineage (referencing public and private datasets and how they are modified.)* 51
Figure 5 *Mobile Penetration Rates: Urban vs Rural, showing rates from Telecommunications Providers reports.* 69
Figure 6 *Strengths, Challenges, Mitigation Strategies* 91

List of Symbols, Nomenclature, or Abbreviations

Symbols

β (Beta): A regression coefficient that measures the relationship between a predictor variable and the outcome variable.
η^2 (Eta-squared): A measure of effect size used in an ANOVA to describe the proportion of variance explained by a predictor variable.
r: The Pearson correlation coefficient, which measures the strength and direction of a linear relationship between two variables.
R^2: The coefficient of determination, indicating the proportion of the variance in a dependent variable that is predictable from the independent variable(s).
p: The p-value, which indicates the statistical significance of an observed result.
t: The t-statistic, a value used in a t-test to determine if there is a significant difference between the means of two groups.
F: The F-statistic, a value used in ANOVA to test for significant differences between group means.
*/**: Asterisks are used in tables to denote levels of statistical significance, typically $p < 0.05$ (*) and $p < 0.01$ (**).

Nomenclature

Data Imputation: The process of filling in missing data, which was noted as a significant methodological limitation in the study.
Data Pipeline: A scripted workflow designed to clean, transform, and integrate varied datasets into a unified format suitable for analysis.
Digital Divide: The gap between demographics and regions in their access to and use of modern information and communication technologies.
e-Governance: The use of information and communication technologies by governments to deliver services and engage with citizens.
Mobile Adoption: The rate at which mobile technology is integrated and used within a society, which is the central focus of the investigation.
Overfitting: A statistical modelling error where a model corresponds too closely to its training data, leading to an excessively high R-squared value and unreliable predictions.
Panel Dataset: A dataset that tracks multiple subjects (in this case, provinces) across a period of time.
Spatial Autocorrelation: A measure of the degree to which a variable's value in one location is correlated with its value in nearby locations.
Spatial Economy: A term used to describe an economy where geographic and demographic characteristics create uneven economic incentives and persistent issues of access and affordability.

Abbreviations

ANOVA: Analysis of Variance
CAGR: Compound Annual Growth Rate
CAPEX: Capital Expenditures
CIUS: Canadian Internet Use Survey
CMA: Census Metropolitan Area
CRTC: The Canadian Radio-television and Telecommunications Commission.
GHS: General Household Survey
ICT: Information Communication Technology
IIS: Infrastructure Intensity Score
ISPs: Internet Service Providers
ITU: International Telecommunications Union
MAI: Mobile Affordability Index
OLS: Ordinary Least Squares
PSM: Propensity Score Matching
SMD: Standardized Mean Differences
VIF: Variance Inflation Factor

Chapter I – Introduction

1.1 Research Background

The proliferation of mobile technology in the 21st century is a central component of modern social and economic analysis yet understanding its adoption patterns remains a significant challenge. While it is widely assumed that mobile devices can mitigate the digital divide, their integration into society is not uniform. Influences such as affordability, infrastructure availability, and pre-existing socio-economic conditions create a complex landscape where access and use are unevenly divided. Before claims can be made about the broader societal impacts of this technology - from its influence on social ties to its role in civic engagement - a foundational understanding of its adoption is required. This foundational analysis, however, is often hindered by significant methodological obstacles related to data availability and consistency, which may be so severe as to render definitive conclusions impossible. The primary challenge, therefore, is not simply to model the adoption, but to first determine if the data ecosystem is fit for that purpose.

Canada is a key example of how the modern world has integrated mobile devices into every aspect of daily life as it has used the economic power available as a developed nation to leverage these advancements into existing social and economic systems (Statistics Canada, 2025). They have influenced the social and daily lives of most of the population in the country. This introduction of new technologies has expanded the abilities of individuals to understand their world and affect change. Both governments, and individuals have embraced new technologies in the region (The Canadian Radio-

television and Telecommunications Commission (CRTC), 2018), especially mobile devices for governments to deliver services and for individuals to make local changes. Canada remains a compelling case study due to its economic development and its unique geographic and demographic characteristics – creating what can be referred to as a "spatial economy" where there are uneven economic incentives leading to persistent issues of access and affordability.

Historically, the ability for populations to have access to greater social, political, and economic opportunities has been limited throughout time due to the inability for these rural, disconnected populations to affect any change outside the powerful urban centers. However, the introduction of mobile technology has allowed most individuals the ability to access beyond their homes, and into the wider world (World Bank, 2024). However, attempts to model the drivers of this adoption are faced with a fundamental data integration problem. Public data from government bodies like Statistics Canada and the Canadian Radio-television and Telecommunications Commission (CRTC) provide essential demographic and regulatory context. Simultaneously, private data from telecommunications carriers offer crucial metrics on subscriptions, revenue, and capital expenditures. These data sources, however, were not designed to be interoperable.

Today the challenge is to measure how efficient and effective individuals have been to utilize these new technologies to enable outreach in these distant communities, and how individuals have begun to take advantage of them to begin to take part in political, social, and economic change. However, the primary challenge is not to immediately measure this, but to first construct a stable framework for analyzing this adoption. Understanding how indirect aspects such as geography and income have

correlated with mobile penetration will provide the ability to analyze the structural drivers that exist and integrate these processes from existing systems.

The integration of these technologies that has allowed for instant communication in almost any environment has altered how societies fundamentally view communication. This thesis confronts this foundational challenge directly. Rather than examining individual usage patterns or the dynamics of social groups, its focus is on the structural drivers of mobile adoption at a provincial and country level, using publicly available indicators and aggregated carrier data.

To this end, this dissertation investigates the statistical relationship between key demographic variables and mobile subscription rates within Canada for the period of 2014 to 2018. It examines how provincial population density and median income levels predict the rate of mobile adoption. The objective is to move beyond high-level narratives about connectivity and instead provide a quantitative model grounded in verifiable data. This approach seeks to establish a baseline understanding of the factors shaping Canada's digital landscape before more complex questions about social impact can be reliably addressed.

Canada’s spatial economy creates uneven incentives for radio access and backhauls. Geographic differences allow for high-density corridors to lower unit costs and speed adoption and off-set costs, while sparsely populated regions face higher per-subscriber costs and slower upgrades. Between 2014 and 2018 carriers completed LTE rollout and core upgrades that raised headline coverage, yet affordability concerns and concentration debates persisted.

Public data (e.g., Statistics Canada population and income indicators; CRTC monitoring tables) and carrier releases (subscribers, revenue, CAPEX) describe the same pieces of the system but use different units, geographies, and time bases. This dissertation attempts to align those pieces to study adoption patterns without assuming policy outcomes. In doing so, it aims to produce a transparent and reproducible analysis of the factors driving mobile adoption in Canada.

1.2 Research Purpose

There is an importance in measuring how technology has impacted our daily lives as it could help us understand the human condition, specifically within national contexts. The current body of evidence on mobile adoption in Canada is often fragmented, appearing as single-year snapshots or narrowly focused pricing debates with limited reproducibility across sources. This fragmentation prevents a systematic, evidence-based analysis of the factors driving mobile adoption across the country over time. The key to this evolution is the explosive growth of mobile technology that began around the world at the tail end of the 20th century. It has become more ingrained in daily life than any other new technology. (Aker & Mbiti, 2010) As such, it reflects an important facet of how a society views itself and each other.

The primary purpose of this dissertation is to address this methodological gap by systematically evaluating the feasibility of creating a reliable model from available data. It seeks to construct a reproducible panel dataset not principally to find definitive answers about adoption, but to stress-test the underlying data sources and transparently document the limitations that emerge. By undertaking this process, this research aims to provide a clear, evidence-based model of the structural factors that predicted mobile

adoption during the critical LTE rollout period. The purpose is twofold: first, to produce a novel, integrated dataset and an analysis of the relationships between geography, income, and adoption; and second, to transparently document the significant challenges and limitations inherent in this process. Ultimately, this thesis contributes a methodological blueprint and a critical evaluation of the available data, offering a foundation upon which future, more reliable studies of Canada's digital landscape can be built.

1.3 Research Question & Objectives

To address the methodological challenges outlined in the research background and fulfill the purpose of this study, the research is guided by a central question, broken down into two testable hypotheses and pursued through a series of concrete objectives. The central inquiry of this dissertation is:

To what extent is Canada’s public and private data ecosystem (2014-2018) sufficiently robust and coherent for reliably modeling the structural determinants of provincial mobile adoption?

1.3.1 Hypotheses

To assess the data's fitness for purpose, this study attempts to replicate two widely expected relationships from technology adoption literature. These relationships are treated as testable hypotheses.

- **H1:** Population density significantly influences mobile adoption rates.
- **H2:** Income disparity significantly affects mobile adoption rates. The ability of a model built from the integrated dataset to validate these expected outcomes will serve as the primary indicator of the data's quality and reliability for econometric analysis.

1.3.2 Research Objectives

The following objectives delineate the specific, sequential steps undertaken to answer the research question and test the corresponding hypotheses:

1. To Collate and Standardize Data: Systematically acquire provincial-level data for the 2014–2018 period from disparate sources, including demographic and income data from Statistics Canada, regulatory reports from the CRTC, and financial and subscription data from Canada's major telecommunications providers.
2. To Design and Implement a Data Pipeline: Construct and document a transparent, scripted data engineering pipeline to clean, transform, and integrate these heterogeneous datasets. This process is designed to overcome inconsistencies in units, geographies, and time bases to create a single, coherent panel dataset suitable for analysis.
3. To Model and Analyze Adoption Drivers: Apply quantitative statistical methods, including correlation and regression analysis, to the integrated dataset to test the hypotheses and measure the strength and direction of the relationship between population density, income, and mobile adoption rates.
4. To Critically Evaluate Findings and Methods: Conduct a rigorous evaluation of the statistical results and the entire data integration process, transparently reporting on methodological challenges, data quality issues, and analytical limitations to ensure the study's findings are properly contextualized and its contribution to research methodology is clearly established.

1.4 Research Methodology

The methodology for this dissertation evolved from an initial comparative framework to a focused, in-depth analysis of the Canadian context, a shift necessitated by real-world data constraints. The study was originally conceived as a comparative analysis between Canada and key African nations, including Botswana, and South Africa. This initial design required a broad literature review, detailed in Chapter II, to contextualize the disparate socio-economic landscapes. This review examined themes of Information Communication Technology (ICT) implementation, infrastructure investment, the challenges of electronic governance, and the influence of geographic and economic realities on technology adoption in both developed and developing nations. The goal was to compare how these factors manifested in different settings to draw broader conclusions about the drivers of mobile adoption.

However, during the preliminary data acquisition phase, this comparative approach was found to be empirically unviable. A thorough reconnaissance of potential African data sources revealed that they lacked the necessary detail and consistency required for proper integration. Additionally, publicly available data, such as South Africa's General Household Survey, were "highly aggregated" and used definitional frames for concepts like internet access that could not be reconciled with Canadian measures. Therefore, a critical methodological decision was made to discontinue the African strand of the analysis and research and focus on the narrower quantitative scope exclusively on Canada where there was somewhat more consistent data available.

A multi-faceted analytical strategy was then applied to this integrated dataset. To test the primary hypotheses concerning geography and income, the analysis employed

Multiple Linear Regression, Analysis of Variance (ANOVA), and Pearson correlation matrices. Specific comparisons, such as urban versus rural disparities, were evaluated using independent samples t-tests. The broader context was explored through Time Series Analysis to identify temporal patterns, Market Competition Analysis to assess market concentration (CR4), and Spatial Autocorrelation (Moran's I) to test geographic clustering of adoption rates. This comprehensive suite of techniques, combined with a transparent reporting of methodological limitations: including the documented failure of more advanced methods like Propensity Score Matching due to covariate imbalance ensures a rigorous and honest assessment of the research question within the bounds of the available data.

1.4.1 Methodological Approaches to Literature Review

The literature review for this thesis was approached in a constructive manner, to ensure that the research was grounded in both theoretical and practical contexts. The review was not limited to background reading but was undertaken as a methodological step to identify relevant frameworks, trends, and knowledge gaps. To achieve this, the search process focused on a combination of academic and professional sources, including databases such as IEEE Xplore, ACM, and ProQuest, and many academic journals, as well as publicly accessible repositories like Google Scholar. In addition, policy documents and industry reports were consulted from Statistics Canada, the Canadian Radio-television and Telecommunications Commission (CRTC), and financial releases from the major telecommunications providers.

The key search terms employed included "mobile adoption in Canada," "income digital divide," "urban rural telecommunications," "ICT infrastructure disparities," and

"telecommunications investment geography." Peer-reviewed journal articles, government reports, and industry analyses were prioritized for inclusion, while lightly related literature and secondary sources were used only where they added important context to the Canadian case. By using this approach, the literature review was able to provide a comprehensive and balanced foundation for the design of the research questions and hypotheses.

In treating the literature review as a methodological step, this work sought to move beyond description and into evaluation. Each source was assessed for relevance, reliability, and methodological quality, and its findings were examined in relation to the broader themes of technology adoption, geographic disparity, and income inequality. This process ensured that the literature review not only outlined existing knowledge but also directly informed the construction of the analytical framework applied in later chapters.

1.4.2 Maintaining Scope and Strict Timelines in Literature Analysis

The scope of the literature search was defined by two complementary timelines. The primary focus was on research published within the period of 2014-2018, as this period reflects the most significant developments in mobile adoption, affordability concerns, and infrastructure investment within Canada. In addition, older works dating back fifteen to twenty years were included where they provided theoretical or historical context, particularly in relation to established models such as the Diffusion of Innovations theory and the Technology Acceptance Model. This two-tiered approach ensured that both contemporary issues and long-term patterns were considered in shaping the research.

The literature review also drew extensively on government and industry data to supplement the academic sources. Key materials included the Canadian Internet Use Survey (CIUS), Communications Services in Canadian Households reports, and the annual monitoring reports of the CRTC. These sources were critical for understanding the quantitative trends in mobile adoption from 2014 to 2018, the period that frames the core of this research, as they reflected many of the significant infrastructure, regulatory, and technological changes that occurred in not only this period, but in the 21st century as a whole. To support this data, the financial statements and investor reports of Bell, Rogers, and Telus were incorporated, providing insight into pricing, investment, and subscriber trends. In cases where discrepancies in reporting methods were identified, the information was cross-checked and triangulated to maintain consistency.

By structuring the review around both a defined time frame and a broad range of sources, this thesis was able to create a comprehensive picture of mobile adoption in Canada. The findings from the literature highlighted the lack of detailed micro-level studies within the Canadian context, particularly in relation to the combined effects of income and geography. As such, the literature review not only provided the necessary background but also reinforced the methodological decision to focus on high-level, publicly available datasets as a way of addressing these gaps in the existing research.

1.5 Data Processing and Analysis

As this thesis investigates income and geography using a mixed-methods approach, through the publicly available data as provided through Statistics Canada. The institution has examined through a variety of communication, and economic reports the

trend of telecommunications in Canada, both as it affects individuals in the public, as well as the landscape for economic development in both the public and private sector. It reflects the landscape, as a legal, political and regulatory matter. The use of publicly available data limits the research to that which can be more generalized towards the populations in each region. By not selecting more individualized or detailed datasets it removes bias towards certain regions or groups. Additionally, economic sources from publications in the industry itself will be used to measure cost progress over time – such as the financial releases of the telecommunications providers. These provide key indicators towards the economic imperative and impact to develop the telecommunications industry of Canada, and the general overall economic freedom of the public. It shows the affordability, and the general availability of integrating mobile services into the daily lives of people.

The datasets applied to the Canadian context include the Communications Services in Canadian Households Subscriptions and Expenditures directly from Statistics Canada (Statistics Canada, 2025). The primary data inputs were acquired from official Canadian institutions: demographic data, including provincial population and population density, were sourced from Statistics Canada, alongside key economic indicators such as median household income. Regulatory and market-level data, including provincial mobile subscriber counts and provider market share, were derived from the Canadian Radio-television and Telecommunications Commission's (CRTC) annual Communications Monitoring Reports. To provide financial and investment context, this public data was supplemented with metrics such as capital expenditures (CAPEX),

extracted from the annual corporate reports of Canada's major telecommunications providers: Bell, Rogers, and Telus.

The focus of the data processing for this is detailed in Chapter IV in the attempt to process fundamentally incompatible raw data sources. Each dataset employed different schemas, geographic boundaries, units of measurement, and reporting periodicities, making direct integration impossible. To overcome this, a custom, multi-stage data engineering pipeline was developed in Python. This scripted workflow was designed to systematically ingest the raw files, perform tailored cleaning and preprocessing routines to standardize formats and handle inconsistencies, and merge the distinct sources into a single, coherent province-year panel dataset. This process was essential for creating the analysis-ready file upon which all subsequent statistical tests were performed.

Once the integrated dataset was constructed and validated, a comprehensive suite of statistical analyses was conducted to test the research hypotheses, as detailed in the final analysis report. As such, this dissertation’s primary contribution is twofold: first, the creation of an integrated, auditable dataset for Canadian mobile adoption from 2014–2018, and second, a transparent workflow that exposes the limits of inference when using such data. The methods documented allow other researchers to reproduce the results, extend the analysis to finer geographies as data becomes available, and critically distinguish between robust findings and artifacts of the data cleaning or modeling processes.

1.6 Research Contribution & Significance

This research is important as it gives an overview of the geographic and income factors that influence cellular prices. Affordability of technology is a key aspect of being able to integrate and take part in the technological community. As the world wide web, instant communication and cellular devices become more engrained in the daily personal and professional life of people, the digital divide will have a larger impact.

This research provides an overview of the geographic and income factors influencing cellular prices. This high-level overview serves as a first step toward enabling the more detailed analyses of how mobile technology is required for the research question.

It identifies weaknesses and gaps in existing research and sets a foundation for future investigations. Ultimately, this work's most significant contribution is a cautionary case study that demonstrates the limits of quantitative analysis with currently available telecommunications data. As such, it can be used to further push where research needs to be focused on in the future.

By using the results from governments, non-governmental organizations, and private organizations, it can easily be added to additional research and as a template for further research throughout the world. Also, it can show where the weaknesses lie, and where existing research has been blind.

The scope is Canada-only, 2014–2018, at provincial depth with limited national analysis. Adoption is measured with public indicators and carrier aggregates defined in Chapter III. A planned cross-country comparison with African markets was discontinued due to missing series, incompatible geographies, and access constraints that prevent reproducible joins: Chapter II documents this decision. As such, an outcome of this

dissertation's significance is its function as a critical and cautionary case study. By transparently documenting the entire research process, this work highlights the profound limitations of currently available data. It demonstrates that even with a rigorous data integration process, deep structural problems within the data can render sophisticated statistical models unreliable. In doing so, this research serves as a grounded call for improved data collection, standardization, and open data practices from both public institutions and private corporations, which are essential for conducting valid and generalizable research on the digital divide

Limitations include aggregation that masks intra-provincial variation, schema drift across sources, measurement differences in carrier reporting, and model instability in small samples. Results are associational and do not claim causality.

1.7 Research Limitation & Scope

There are a variety of limitations to pursuing this line of research. There are both geographic and economic considerations. Due to the size and shape of both the country and the economy, investment in infrastructure, and average incomes differ significantly. To ensure applicability, the research focuses on high-level data, excluding interviews or micro-analysis of individuals, which limits the ability to address narrow questions. The focus on statistical data from the government, as well as financial data from the private telecommunications providers show instead the overall trend of the economic factors and societal factors in Canada.

This research employs a macro-level approach, focusing on categorical differences rather than micro-level individual behaviors.

As such, the data provided by governments and organizations cannot reflect narrow questions such as individual tastes, preferences, or choices towards items such as device choice or preferences for specific social media sites or applications. Instead, the data will only reflect overall trends in these societies.

Therefore, the findings of this dissertation must be interpreted in the light of several significant limitations stemming from the data and analytical methods. The primary of which is that there is a reliance on province-level data, which masks intra-provincial issues with disparities that can have an undue effect on both macro and micro levels.

Additionally, there are data quality and consistency issues as schema drift from different sources created inconsistencies as detailed in Chapters IV and V - the final dataset was found to contain artifacts from data processing that could lead to further anomalies. This created analytical constraints - due to these issues, the sample size and other issues identified in Chapter V.

1.8 Organization of the Thesis

To accomplish these goals, this thesis is divided into six chapters that will examine questions regarding the relationship between mobile prices and subscriptions as related to income and geography in Canada. This first chapter acts as an outline for the rest of the document, providing context for the rest of the thesis, including motivations, goals, and the importance of the subject.

By following this structure, the thesis aims to provide a comprehensive analysis of the relationship between mobile technology, geographic spread, and income disparity in Canada, contributing valuable insights to the field of study.

Chapter II will begin the discussion of the topic with a literature review of pre-existing research on the subject, examining how specific governments have used technology since the mid 1990's in Africa to reach out to communities and how in recent years it has integrated from central locations with computer access to becoming focused on mobile devices. It will focus on two separate parts of the topic – electronic infrastructure concerning Information Communication Technology and how it has affected citizen activism, with a specific focus on minority rights. Each of these sections will be further divided into a general overview to contextualize the situation on the continent. Chapter II integrates the literature review with methodological rationale and records the decision to discontinue the African strand as the effort provided context and value when examining the landscape of the Canadian infrastructure commitment and implementation.

Additionally, the literature review will focus on the Canadian context of these same concerns, cultural shift, infrastructure, and government influence. The goal of this is to give an overview of how integration has progressed in both regions so that the modern view of the impact of cellular devices can be compared to their historical counterparts.

Together, these two aspects of the literature review will provide context for the Canadian information – the investment in infrastructure, mobile growth, population expansion and private reach with cellular technology. With this background provided, the Canadian context can be researched and its development compared to African states, ensuring that Canada's development is not presented in a vacuum.

Chapter III will discuss the design of the dependent and independent variables that will be used to answer the hypothesis and address the research model research model. This will include detailed information on how the data was collected.

Chapter IV will examine the data collected as described in Chapter III and how it was processed—from geographic sources, through cleaning, to how it was stored and managed.

Chapter V will analyze the data regarding the hypothesis and the final analysis of the questions including examining how successful and comprehensive the data collection was for answering the questions within this work. This thesis deviates slightly from a traditional structure. Chapter V, 'Data Analysis,' intentionally integrates the presentation of statistical results with an immediate critical discussion of their validity. This approach was chosen because the central finding of this research is the unreliability of the models themselves, making it essential to critique the results as they are presented, rather than separating the findings from their interpretation.

Chapter VI - The final chapter will provide a summary and conclusion as well as suggestions for potential future research and issues that were faced.

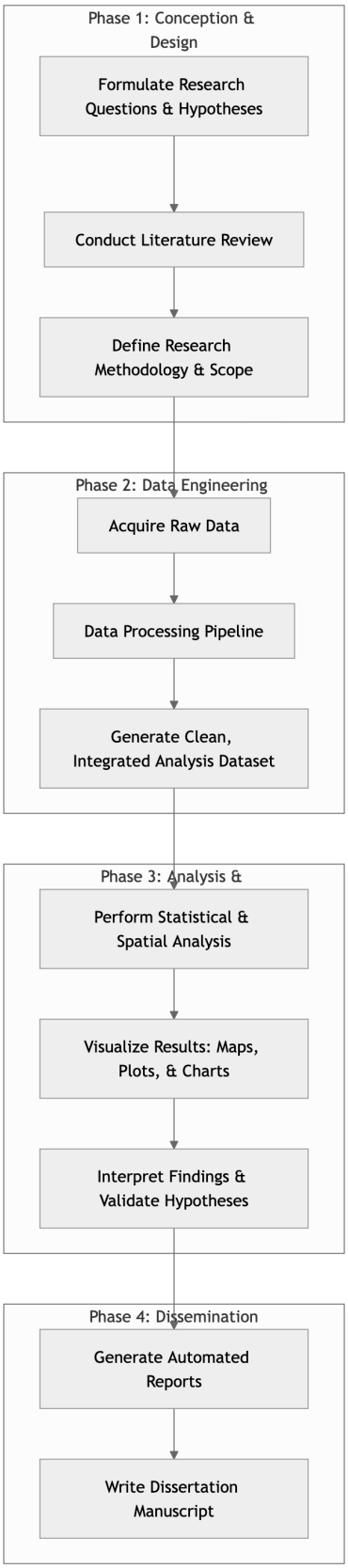
Phase 1: Conception & Design
Formulate Research Questions & Hypotheses
Conduct Literature Review
Define Research Methodology & Scope
Phase 2: Data Engineering
Acquire Raw Data
Data Processing Pipeline
Generate Clean, Integrated Analysis Dataset
Phase 3: Analysis &
Perform Statistical & Spatial Analysis
Visualize Results: Maps, Plots, & Charts
Interpret Findings & Validate Hypotheses
Phase 4: Dissemination
Generate Automated Reports
Write Dissertation Manuscript

Figure 1 *Dissertation Phases*

The Figure 1 diagram provides a strategic overview of the entire research process, illustrating the logical progression from conception to dissemination. The workflow is organized into four distinct phases: (1) Conception & Design, which establishes the theoretical and methodological foundations; (2) Data Engineering & Preparation, a critical phase focused on transforming raw data into a viable dataset; (3) Analysis & Interpretation, where statistical models are applied to test hypotheses; and (4) Dissemination, which involves the final reporting and documentation of findings in the dissertation manuscript. This structured approach ensures a systematic and coherent execution of the research plan.

Chapter II – Literature Review

This chapter reviews the existing research on information and communication technologies (ICTs), mobile adoption, and e-Governance across Africa and Canada. The dissertation initially aimed at conducting a comparative analysis using data from both regions. However, while African literature provides useful context, the empirical data proved insufficiently granular or consistent to support robust statistical analysis.

The African review remains important for two reasons. First, it illustrates the challenges and innovations of ICT development under severe resource and governance constraints. Second, it shows how data inequality shapes the feasibility of comparative research. For these reasons, African cases are included for qualitative framing, while the quantitative analysis proceeds exclusively with Canadian data.

The chapter is organized into two parts:

1. African literature – ICT infrastructure, e-Governance, and the structural limitations of African data.
2. Canadian literature – ICT infrastructure and e-Governance in a developed-country context.

2.1 African Literature

Africa's technological revolution has been uneven. From the late 1990s onward, governments attempted to introduce e-Governance platforms to improve service delivery (Heeks, 2002). These initiatives often failed because they imported Western administrative models that assumed institutional capacities not present in African

contexts delivery (Heeks, 2002, p. 3) Connectivity levels were among the weakest globally during this period (Polikanov & Abramova, 2003, p. 43).

Heeks argued that strategies needed to be locally adapted - relying more heavily on private providers, incremental investment, and flexible regulation (Heeks, 2002, pp. 15, 19). This view shaped subsequent debates about the digital divide and ICT adoption.

Asongu & Biekpe analyzed 49 Sub-Saharan countries from 2000–2012 (Asongu & Biekpe, 2017, pp. 112, 125). They linked ICT penetration to governance quality, showing paradoxical effects: regulatory control could boost adoption in the short term but slow expansion in the long term by limiting competition. Their work illustrates how corruption and accountability shape ICT outcomes differently across contexts.

Penard et al. conducted a survey of 1,300 residents in Gabon, highlighting distinct adoption trajectories (Penard, Poussing, Yebe, & Ella, 2012, p. 80). Internet access was concentrated among affluent groups, while cell phones reached a much broader base. Their findings confirmed that mobile technology spread inclusively, whereas fixed Internet remained elite.

Botswana represents one of the most extensively documented African cases. A national ICT policy was drafted in 2005, aiming to achieve a fully realized “information society” by 2016, coinciding with the country’s 50th independence anniversary (Mutula S. M., 2004, p. 149). Implementation, however, was fragmented. Departments developed separate ICT strategies with little coordination, creating duplication and inefficiencies (Moloi & Mutula, 2007, p. 301). Staffing and infrastructural deficits further undermined rollout (Mosweu, Bwalya, & Mutshewa, 2016, p. 148)

Educational initiatives placed ICT resources in schools and libraries (Bose, 2005, p. 19). Yet costs remained high, particularly outside urban centers, and infrastructure required more than just hardware procurement (Paterson, 2007). In response, open-source technologies were promoted as a lower-cost alternative from 2007 onward (Mutula & Kalaote, 2010, p. 63)

Telecommunications reform in the 1990s liberalized Botswana's market, inviting foreign entrants. A GSM contract in 1998 went to two foreign-backed firms rather than the state provider (McCormick, 2001, p. 16). This created rapid cellular expansion.

By 2010, Botswana reached 131% mobile penetration (Lesitaokana, 2014, p. 846). Yet problems remained: the market had only a few private operators plus a quasi-state provider. Regulations encouraged competition among incumbents but discouraged new entrants, resulting in imbalances that favored providers over consumers (Lesitaokana, 2014, p. 847).

Rural connectivity has been an enduring obstacle. While modern GSM and even 4G networks became available in select areas by 2014 (Orange, 2018), access remained highly uneven. The digital divide was evident in literacy gaps and low IT usage among rural populations. Nyamaka et al. noted weak online presence among local businesses, limiting economic benefits even where infrastructure was available (Nyamaka, Botha, Van Biljon, & Marais, 2018, p. 7).

African governments often struggled to implement online services effectively. (Badmus, 2017, p. 12) noted limited availability due to state restrictions, while (Mukeredzi, 2017, p. 9) documented politically motivated Internet shutdowns. Such

practices undermined freedom of information and required external providers such as the BBC World Service to reach citizens (Ogunyemi, 2011, p. 460).

Attempts to measure e-Governance quality include (Sá, Rocha, Gonçalves, & Cota, 2017, p. 415), who developed a ranking framework from "Very Weak" to "Very Good." (Cegarra-Navarro, Co´rdoba-Pacho´n, & Garcı´a-Pe´rez, 2017, p. 315) drew on analogies from Spain's hospitality sector, suggesting governments should view online platforms as supplements rather than replacements for existing services.

Given infrastructure constraints, many African states relied on basic mobile services. In Uganda, for instance, the water utility company allowed bills to be viewed and paid via SMS (Mpinganjira, 2014, p. 132). In Botswana, the Sesigo project under the Global Libraries initiative equipped selected libraries with computers and Internet access. These sites quickly became popular, showing that latent demand existed whenever access was provided (Resego, 2012, p. 41)

2.2 Limitations of African Data for Comparative Analysis

The initial research design attempted to integrate African and Canadian datasets. This attempt failed for empirical reasons. Three main issues explain the limitations.

(a) Survey Aggregation and Inconsistencies

- South Africa's General Household Survey (GHS) reports ICT adoption only in broad categories such as "internet access at home, work, or mobile." These cannot be reconciled with Canadian measures like household Internet subscriptions.
- Even with restricted access to microdata, harmonization would be undermined by differences in income banding, question wording, and categorization.

- Example: In 2017, 62% of South African households reported at least one member accessing the Internet, but this statistic was not disaggregated enough for econometric analysis.

(b) Schema Incompatibility and Metadata Absence

- Regulators such as ICASA (South Africa) and NCC (Nigeria) published totals in PDF form, lacking metadata or consistent schemas.
- Definitions of "broadband" or "user" shifted across years, undermining longitudinal analysis.
- ITU and African Union compilations inherited these inconsistencies, offering only coarse, often incompatible summaries.

(c) Temporal Gaps and Irregular Continuity

- Surveys were sporadic: in some cases, the last available round was in 2012, followed only by 2018. This left no observations for the 2014–2017 window of this study.
- Even within a country, indicator definitions were changed, breaking comparability across survey waves.

Source	Coverage	Limitation
South Africa GHS	Household ICT indicators	Aggregated; incompatible definitions; restricted microdata
ICASA Reports	Telecom statistics	PDF-only; inconsistent categories; missing metadata
NCC (Nigeria)	Subscriber totals	No disaggregation; national aggregates only
ITU Databases	Regional ICT adoption	Coarse national averages; gaps between years
AU Compilations	Multi-country	Inconsistent definitions; missing series

Table 1 *Candidate African Sources and Reasons for Exclusion*

2.3 Canadian Literature

Canada's development of ICT infrastructure has been shaped by both its geography and its regulatory framework. Geographic size, terrain, and a sparse population have long made providing coverage difficult, even when ample economic resources were available. To address these challenges, the telecommunications industry was liberalized in the early 1990s, allowing new entrants to connect and exchange services where government provision was limited (Wilson, 1996, p. 607). This created a top-down model that involved not only large corporations but also communities and small businesses, widening access (Frieden, 2005, p. 605) Yet, the results have been mixed. Compared to nations such as Japan or Korea, Canada has lagged, not because of insufficient funds but because cooperation and careful planning are required to overcome the unique geographical obstacles (Frieden, 2005, p. 609).

Provincial strategies further illustrate these divides. (Rajabiun & Middleton, 2013) showed that provinces adopting open-access models for backbone networks achieved higher broadband quality in rural and remote areas, while those relying more heavily on market forces lagged. Their analysis of broadband speed data (2007–2011)

also demonstrated that weak federal access mandates constrained provincial progress compared to other OECD countries. These findings parallel challenges identified in African contexts, where regulatory structures directly mediate infrastructure outcomes.

Policy programs have attempted to bridge these gaps, but with limited success. (Rajabiun & Middleton, 2013) found that Canadian broadband policy historically relied on supply-side expansion, neglecting demand-side barriers such as affordability and digital literacy. Reviewing provincial initiatives, they concluded that while physical access improved, uptake remained constrained by high costs and socio-economic inequalities, particularly in rural and northern regions. Programs such as *Connecting Canadians* (2014) extended broadband to rural areas through subsidies, but evaluations showed that competition and affordability challenges persisted, leaving many rural households paying some of the highest mobile and broadband costs in the OECD (McNally, Dinesh, Evaniew, & Yang, 2017) As in African contexts, expanding infrastructure without addressing affordability or content relevance produced limited gains for marginalized communities.

The role of smaller Internet service providers (ISPs) illustrates these structural constraints. (McNally, Rathi, Joseph, Evaniew, & Adkisson, 2018)reported that small ISPs in rural Canada were disadvantaged by federal regulatory and funding structures, leaving them unable to scale or compete effectively with incumbents. In a related study, (McNally, Dinesh, Evaniew, & Yang, 2017) reviewed eight federal broadband programs (1994–2016) and found that while these initiatives expanded coverage, they often reinforced incumbent advantages rather than delivering affordable service to marginalized communities. This reflects a recurring theme across both Canadian and

African literature: when market concentration and regulatory bias favor established providers, the digital divide persists despite formal infrastructure expansion.

Socioeconomic disparities further shape adoption. (Landry & Lacroix, 2014), using Canadian Internet Use Survey data, found that higher education and income were the strongest predictors of smartphone and mobile internet use. Although overall internet gaps narrowed between 2010 and 2012, inequalities widened for mobile adoption, with wealthier and younger Canadians adopting at higher rates. (Haight, Quan-Haase, & Corbett, 2014) confirmed this pattern, showing that income, age, and education predicted both access and engagement levels. Their regression analyses demonstrated that individuals with higher socioeconomic status engaged in a wider range of online activities, while low-income and older Canadians remained underrepresented. These results reinforce that, much like in African contexts, structural inequalities determine how populations benefit from mobile technologies.

E-government provides another perspective. Early Canadian efforts in the 1990s often produced surface-level implementations, relying heavily on providing computers and internet access in libraries (Reddick & Turner, 2012, p. 2). More recent reforms have sought to move beyond digitizing existing services toward "digital government (Clarke, Lindquist, & Roy, 2017)) note that this transition emphasizes user-centric and data-driven approaches, though it also introduces challenges around misinformation, privacy, and accountability. (Roy, 2017) highlighted siloed institutional cultures and uneven rural uptake as barriers to equity. These findings parallel African studies where cultural and structural obstacles constrain the effectiveness of e-governance reforms.

Concrete examples highlight the tensions between policy goals and outcomes. The *Connecting Canadians* program and similar initiatives extended rural infrastructure but failed to reduce high end-user costs (McNally, Rathi, Joseph, Evaniew, & Adkisson, 2018) (McNally, Rathi, Joseph, Evaniew, & Adkisson, 2018). This mirrors African cases where state-funded ICT projects expanded networks but left affordability unresolved. Together, these works demonstrate that Canadian digital divides are not purely technical but structural, shaped by the interaction of geography, regulation, and market power.

Chapter III – Research Design and Methodology

This chapter details the quantitative methodological framework designed to answer the research question. It begins by outlining the research design and formally stating the hypotheses that guide the analysis. It then describes the specific public and private data sources collected for this study. The central focus of the chapter is an in-depth explanation of the custom data engineering pipeline that was developed to overcome the significant challenges of data integration. Finally, it outlines the specific statistical methods employed in the analysis.

3.1 Hypotheses and Research Design

The unit of analysis is the province-year, allowing for an examination of relationships between variables over time and across regions. The design is explicitly associational; it estimates the strength and direction of these relationships under clear data and modeling constraints, without claiming causality. The entire workflow is scripted and auditable to ensure maximum transparency and reproducibility.

The research is guided by the following hypotheses:

H1: Population density significantly influences mobile adoption rates.

This hypothesis is based on the premise that more populous regions, characterized by better infrastructure and higher population densities, will exhibit higher mobile adoption rates compared to less populous regions. The dependent variable for this hypothesis is mobile adoption rates, while the independent variable is population density, which is operationalized by categorizing regions into more populous and less populous areas. The expected relationship is that more populous regions will have significantly higher mobile adoption rates due to their enhanced access to technological infrastructure and services.

H2: Income disparity significantly affects mobile adoption rates.

This hypothesis is grounded in the Technology Acceptance Model, which suggests that higher income levels are associated with greater perceived ease of use and perceived usefulness of mobile technology, leading to higher adoption rates. The dependent variable for this hypothesis is mobile adoption rates, and the independent variable is income levels, categorized into different income brackets. The expected relationship is that higher-income groups will show significantly higher mobile adoption rates due to their ability to afford premium mobile services and devices.

Both hypotheses examine whether mobile adoption is adversely influenced by external societal pressures, by using geography and income as the determining factors the outcome reflects the overall makeup of each society.

This study uses a quantitative, observational panel for Canada, 2014–2018, at the provincial/territorial level where metrics exist and at national level where only aggregates are published. The unit of analysis is province–year for adoption, population, and income; national series are used as contextual covariates and are not interpreted as provincial effects. The workflow is scripted and auditable: each figure and table is generated from saved inputs with logged checksums and row counts. The design is associational. It estimates the strength and direction of relationships between adoption, population density, and income under clear data and modeling constraints documented in Chapters 4–5.

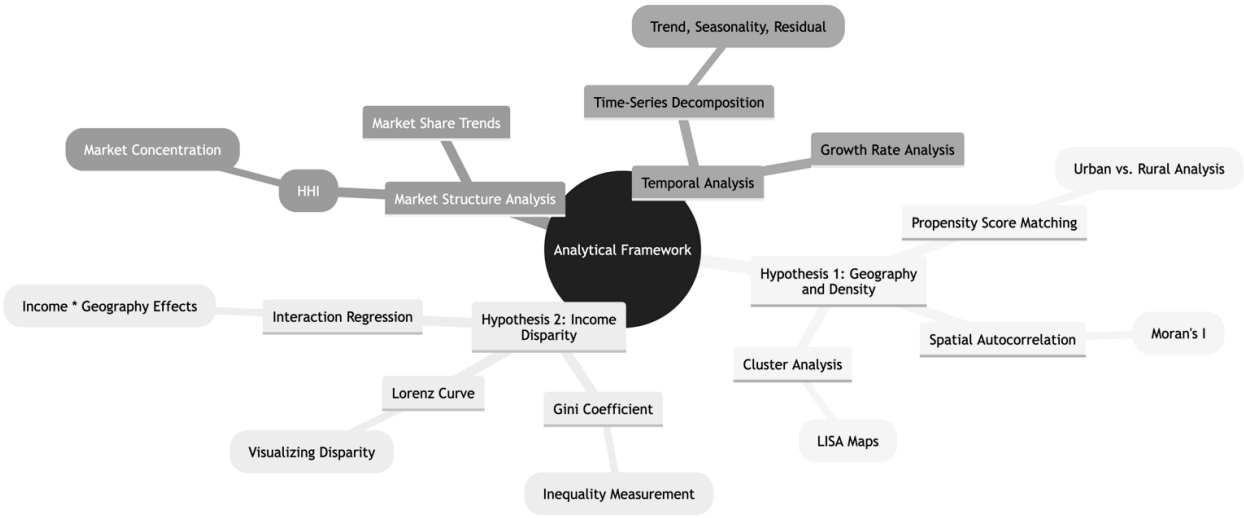

Figure 2 *Analytical Framework*

The analytical framework for this study is visually articulated in Figure 2, which maps the core research hypotheses to the specific statistical methods employed for their validation. To investigate Hypothesis 1, concerning the influence of geography and population density, a combination of Propensity Score Matching, Spatial Autocorrelation (Moran's I), and Cluster Analysis is utilized. For Hypothesis 2, which addresses income disparity, the analysis employs the Gini Coefficient and Lorenz Curves for inequality measurement, supplemented by an Interaction Regression to explore combined effects. The framework is further supported by market structure and temporal analyses to provide a comprehensive and multi-faceted examination of the research questions.

3.2 Collecting Data

Instead of collecting data through individualized surveys or interviews this thesis instead used large datasets available to the public through government, non-profit organizations, and private organizations. This will allow for a variety of sources to be combined to get a view of the overall region. However, there was a limit on the amount of available data for this specific research regarding the subscription and mobile ownership by demographic.

When answering the research question for this thesis, obtaining relevant and useful data is extremely important. This is accomplished by separating the populations into their regions and separated by their income as defined by each specific province. By focusing on the domestic definition of their income decile it reflects the population itself and not each state to the other. Additionally, the influence of population growth during the examined period is compared with subscription growth.

Source	Variables	Temporal Coverage	Geographic Granularity
Statistics Canada (11-10-0239-01)	Median income, population	Annual, 2014-2018	Provincial
CRTC Monitoring Reports	Subscribers, market share	Annual, 2014-2018	Provincial/National
Corporate Annual Reports (Bell, Rogers, Telus)	CAPEX, revenue, EBITDA	Fiscal years 2014-2018	National (allocated)
ITU Mobile Affordability Index	Price/income ratio	Annual, 2014-2018	National

Table 2 *Data Sources and Coverage*

All datasets were accessed between March-August 2024. Statistics Canada tables provide nationally comparable measures with consistent definitions across years. CRTC reports supply provider-specific wireless subscriber counts forming the backbone of the adoption proxy. Corporate filings required manual extraction and standardization

As described in Table 2, all datasets applied to the Canadian context include the Communications Services in Canadian Households Subscriptions and Expenditures, Household Internet Use Survey and now the Canadian Internet Use Survey (CIUS) directly from Statistics Canada, and Canadian Radio-television and Telecommunications Commission Telecommunications Monitoring Reports as the datasets to use for

answering my research question[1]. In addition, I have supplemented the data from the CIUS surveys with those from the World Bank (The World Bank, 2018), specifically their data regarding population growth year over year and the percentage of change per year that entails and mobile subscriptions. In addition, financial and operational data, such as capital expenditures and subscriber trends, were extracted from the annual public reports of Canada's major telecommunications institutions: Bell, Rogers, and Telus.

3.3 Pipeline Development

The methodological approach employed in this thesis was guided by two priorities: (1) maximum transparency and traceability of public and private data sources; and (2) pragmatic adaptation to real-world limitations in available telecommunications and income data. The decision to build a custom, modular pipeline was chosen to satisfy both of those needs.

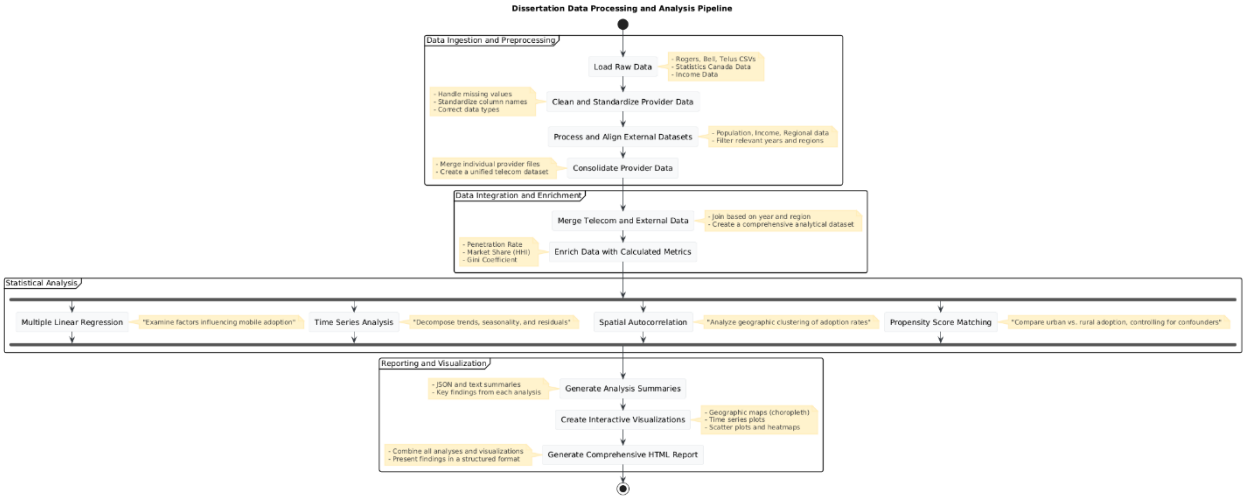

Figure 3 *Data Processing and Analysis Pipeline*

[1] See Appendix E for market share data tables.

Figure 3 illustrates the data processing and analysis pipeline, as a systematic workflow designed to transfer the raw data into its final analytical report.

The process commences with Data Ingestion and Preprocessing, where raw data from telecommunications providers and external sources are loaded[2]. This stage involves cleaning and standardizing the provider data by handling missing values and correcting data types, while external datasets are processed and aligned by year and region. The individual provider files are then consolidated to create a unified dataset.

In the subsequent Data Integration and Enrichment phase, the consolidated telecom data is merged with the external data to create a comprehensive analytical dataset. This dataset is then enriched with calculated metrics such as penetration rates, market share, and the Gini Coefficient to enhance its analytical utility.

The enriched data then undergoes Statistical Analysis, which employs multiple methodologies. These include Multiple Linear Regression to identify factors influencing mobile adoption, Time Series Analysis to decompose trends, Spatial Autocorrelation to analyze geographic clustering, and Propensity Score Matching to compare urban versus rural adoption.

Finally, the Reporting and Visualization stage synthesizes the analytical results. This involves generating textual summaries of key findings, creating interactive visualizations like maps and plots, and compiling all outputs into a comprehensive HTML report.

3.3.1 Pipeline Motivations

Public and private sector data sources often lack shared formatting conventions or coherent schema. Statistics Canada data, while structured and comprehensive, often lacks

the economic granularity provided in industry financial reports. Alternatively, private carriers produce financially motivated metrics that can obscure comparability (e.g., bundling mobile revenue with media revenue).

To support meaningful hypothesis testing for both geographic and income effects on mobile adoption, it became necessary to unify these differing data points into a normalized, analyzable structure. This process required not just cleaning, but structural transformation of data formats and field information. Several early strategies failed, various early pipeline implementations were attempted; these are detailed in Appendix B. Additionally, the complete pseudocode representation of this modular pipeline, outlining all stages from raw data ingestion to final dataset generation, is provided in Appendix (Pseudocode Pipeline) for reference and reproducibility.

3.3.2 Design Principles of the Pipeline

The pipeline was designed around the following principles:

- Modularity: Each data type (e.g., income, population, subscriptions) was handled in a separate module to allow for independent reuse.
- Verifiability: All steps were script-based, version-controlled, and annotated to allow for full audit trails.
- Extensibility: New datasets (e.g., 2022 CIUS or future CRTC reports) can be integrated with minimal code changes.
- Resilience to Incompleteness: The pipeline includes exception handling and imputation protocols to deal with partial datasets without halting analysis.

3.3.3 Ethical and Epistemological Considerations

This data integration effort is not purely technical. Each transformation decision e.g., merging carrier zones with provincial boundaries, or mapping ARPU ranges to income quintiles – reflects interpretive choices that shape the research findings. Where decisions could not be justified with strong assumptions, sensitivity analyses were conducted. Due to poor coverage in some financial records, regional inferences required modeling using proxies, such as using total provincial subscriptions as a weighting factor.

Other trade-offs involved discarding plans to include data from promotional pricing cycles due to the lack of archive data. Instead, a median rate was used to approximate access cost over time. While less dynamic, this approach proved more stable and generalizable across provinces.

By documenting these choices and embedding metadata within the pipeline, this research adheres to both the best ethical practices in data handling and methodological standards in quantitative social science. This expanded methodological framing supports the overall research design and ensures that the findings presented in Chapter V are grounded in a defensible and transparent computational process.

3.4 Pipeline overview

Data ingestion scripts pull raw CSVs and PDFs from the sources listed in Table 3.1. Cleaning modules standardize column names, handle encoding, and remove non-numeric symbols. Normalization converts currencies to constant 2019 dollars and scales variables. Integration merges datasets on Province and Year, appending provider identifiers where applicable. Feature engineering computes derived fields such as CAPEX per capita and

MAI. Analytical routines perform panel regressions and correlation analysis. Reporting modules generate visualizations and tables for Chapters IV and V. Each stage writes intermediate outputs to ensure traceability and reproducibility.

The end-to-end pipeline follows a logical sequence of discrete stages, with each stage writing intermediate outputs to ensure traceability.

Ingest → Clean → Normalize → Integrate → Engineer → Analyze → Report

This scripted process begins by pulling raw data from the sources listed in section 3.2. Cleaning modules standardize formats and remove non-numeric artifacts. Normalization and integration modules merge datasets on Province and Year.

The methodological design required a practical system to provide transparency and reproducibility. Manual methods were not sufficient to handle the volume and inconsistency of the datasets. As a result, a complete computational pipeline was developed to automate every stage of data handling—from ingestion to analysis. This pipeline replaced earlier spreadsheet-based approaches and ensured that each transformation was logged, verifiable, and repeatable. The next section outlines the structure and logic of this system, describing how the data were processed, integrated, and prepared for the analyses presented in Chapters IV and V.

3.4.1 Pipeline Implementations

A pivot-table-based approach in Excel quickly broke under load when applied to multi-year subscription data. Manual merging introduced errors and inconsistent join logic. After identifying over 15 merge failures due to whitespace and encoding mismatches, a decision was made to completely migrate preprocessing into Python using

the pandas library, which allowed for more robust control of data types and join strategies.

The scripts were implemented in Python, using pandas for data handling, and SQLite for mid-stage storage. Error handling was embedded directly into the script logic-for example, in the CIUS preprocessing function, a try-except block detected year-over-year schema differences, while logging flagged any changes in expected row counts or null distributions.

Rows are included when a province has non-missing adoption, population, and income for a given year. Duplicate province–year entries are collapsed by verified numeric equality; otherwise, the row is dropped with a log entry. Unit conversions are performed before joins. For series with one-year gaps, a single forward-fill is permitted within the same province when the source documentation indicates unchanged methodology; no back-casting beyond one period is performed. Rows with unresolved schema ambiguity or mixed units are excluded and recorded in the run ledger (Appendix E). National series are retained for descriptive context only and are not merged into province-level models.

The modular structure of the data pipeline was implemented through a set of discrete Python scripts designed to maximize reproducibility and minimize cross-contamination of data transformations. Each stage of the pipeline was isolated so that loading, cleaning, preparation, and analysis were conducted independently, with standardized input and output interfaces. This ensured that errors at one stage did not

propagate silently into subsequent stages and that intermediate datasets could be audited for accuracy[2].

3.4.2 Geographic Data Processing

This was done using Python, initially with a script that was designed for cleaning, merging, and preprocessing the data from the geographic data. This included integrating both CSV and Excel spreadsheets and cleaning them. Using common columns based on dates, it filtered them for the time period of 2014-2018 and only included relevant data, by identifying common columns (fuzzily) and iterating them together in batches.

A significant issue with both of these datasets was missing data, or incomplete data – for both states, the datasets were incomplete, or had incompatible data, where this occurred, it was removed instead of converted.

For Canada, the main issue with preprocessing for analysis was flattening the multi-level column headers in the data – which meant renaming each column for clarity (instead of Nova Scotia 2014, 2015, NS_2014, NS_2015) and then preparing the data for each region.

The technical details of preparing this data were done, as aforementioned by a series of Python scripts. Each was designed to handle specific tasks to prepare the data for analysis, in this case a ‘clean’, ‘combine’, ‘prepare’, and ‘results’ script. Together, they form a pipeline that processes raw data into a comprehensive dataset, ready for analysis. The use of Python, particularly with libraries such as Pandas, SciPy, and Statsmodels, provides the ability and accessibility needed to handle the large datasets from Statistics Canada, and ensure that there is data integrity throughout the process.

[2] See Appendix B for pseudocode of the full pipeline implementation.

Data Transformation and Cleaning: Once the numeric conversion function is applied to relevant columns, the data is reshaped from a wide format to a long format using the pd.melt() function. This transformation is essential for longitudinal data analysis, allowing for an easier comparison of trends across years. The script then ensures that only numeric-like values are retained, filtering out any non-conforming data points. This is crucial for maintaining data integrity and ensuring that the subsequent analyses are not skewed by erroneous or irrelevant data.

Handling Missing Values: The script includes a robust mechanism for dealing with missing values. After converting the data, the script applies interpolation to fill in any gaps. Interpolation is a method that estimates missing values within the range of the available data, providing a more continuous dataset for analysis.

Final Steps and Data Storage: After preprocessing, the cleaned data is saved to a new CSV file, preprocessed_telus.csv, which is then used for further analysis. The script also generates a trend graph that visualizes the changes in key metrics over the analyzed period. This visualization helps in quickly identifying trends and anomalies in the data, which can be crucial for drawing insights during the analysis phase.

Analysis and Output: Finally, the script performs correlation and regression analyses on the preprocessed data, focusing on the relationship between different financial metrics and their evolution over time. The results are saved to a text file, providing a summary of the key statistical findings.

While each script is designed to preprocess data from different sources, they share several commonalities in their approach to data cleaning and transformation. All three scripts emphasize the importance of converting complex financial data into a

standardized format, handling missing values effectively, and reshaping the data for trend analysis. The use of Python's Pandas library and other tools like Matplotlib for visualization and Statsmodels for statistical analysis ensures that the data is not only clean but also ready for in-depth examination.

3.4.3 The Combine Data Script

The processing begins with combining multiple datasets, which are then cleaned to remove duplicates and irrelevant information. Next, the data is filtered for the relevant time period and standardized to ensure consistency across different sources. Finally, the data is prepared for specific analyses, with a focus on ensuring that the datasets are aligned and free of errors. The Combine Data script is the first in the sequence and is responsible for merging multiple data files into a single dataset. This script is crucial because the data required for this research comes from various sources, including CSV and Excel files, each containing different but related information. As each of these datasets have different assumptions, purpose and slightly different designs, it took significant effort to combine each of them.

The process begins by identifying all the relevant files within a specified directory. The script uses the os library to iterate over all files in the directory, checking their extensions to determine whether they are CSV or Excel files. This was important, as the analysis used a combination of a native Windows Python installation, as well as a GNU\Linux one through the Windows Subsystem for Linux. For each file, the script reads the data using pandas.read_csv() for CSV files and pandas.read_excel() for Excel files. The encoding='latin1' parameter is used to handle any encoding issues, particularly for CSV files, as they are not Unicode, but may have Unicode characters, while the

on_bad_lines='skip' parameter helps manage files with problematic lines by skipping them during the reading process.

Once the files are read into Pandas DataFrames, they are appended to a list. The script introduces a separator row, ["--- End of File ---"], between the contents of each file to mark the transition from one dataset to another when the files are eventually combined. This separator aids in tracing back any issues that might arise during the analysis phase. After all the files are read and stored in the list, the script concatenates these DataFrames into a single large DataFrame using pd.concat(). The combined DataFrame is then saved to a new CSV file, which serves as the foundation for the subsequent cleaning processes.

3.4.4 The Clean Data Script

The next step in the data processing pipeline is the cleaning of the merged dataset, which is performed by the Clean Data script. This script is essential for ensuring that the dataset is free from duplicates and any other inconsistencies that might skew the results of the analysis.

The Clean Data script begins by reading the combined dataset, which was generated by the Combine Data script. It uses the pd.read_csv() function to load the data into a Pandas DataFrame. The primary operation in this script is the removal of duplicate rows, which is achieved through the drop_duplicates() method. This step is crucial because duplicates can inflate the significance of certain data points, leading to biased results and bad data.

After duplicates are removed, the cleaned DataFrame is saved to a new CSV file. The script also includes error-handling mechanisms to catch and report any issues that might occur during the reading or writing process. This is particularly important in ensuring that the dataset remains intact and that the cleaning process does not inadvertently remove or alter important data.

3.4.5 The Data Preparation Script

The script Data Preparation script deals with the cleaning and preprocessing of individual data files before they are merged into a final dataset. This script includes several functions designed to handle different aspects of the data, from converting date columns to filtering data based on specific criteria.

3.4.6 The Prepare Results Script

The final script in the data processing pipeline is Prepare Results script, which focuses on the specific preparation of data for analysis. This script is designed to handle the results from the processed data, organizing them into tables and filtering them based on the relevant criteria for this research.

The script starts by setting up logging to ensure that all steps are recorded. It then loads Excel files using the pd.ExcelFile() method, which allows for efficient access to different sheets within the Excel file. This capability is particularly useful when dealing with large, complex datasets that are organized across multiple sheets.

The clean_table() function in the Prepare Results script is responsible for cleaning individual tables within the Excel files. This function adjusts the column names, ensuring that they match the expected format, and converts any numerical columns from strings to numeric types. This conversion is necessary for any subsequent statistical analyses.

The function also filters the tables based on a specific year range, ensuring that only data from 2014 to 2018 is included in the final dataset. This filtering step aligns with the research focus and ensures temporal consistency across all analyses.

The script then processes additional files as needed, extracting relevant columns and integrating them into the primary dataset. This integration is handled by the integrate_tables() function, which concatenates the tables into a single DataFrame, ensuring that all relevant data is included in the final analysis.

The final step in the Prepare Results script is to save the processed data to a CSV file, which serves as the input for the statistical analyses performed in Chapter V. This script ensures that the data is not only clean but also organized in a way that facilitates the analysis of geography and income influences on mobile adoption.

3.4.7 Provider-Specific Scripts

The script begins by loading the dataset from a CSV file named combinetelus.csv using Pandas. The initial dataset is likely a combination of various financial and operational metrics over multiple years. The first step involves defining an enhanced conversion function, convert_to_numeric(), which handles the complexities associated with converting non-numeric values into a standardized numeric format. This function is particularly adept at processing monetary values, percentages, and large numbers denoted in billions or millions, ensuring that these are correctly interpreted as floats or integers. The Trending script is a more generalized script designed to preprocess a combined dataset that potentially includes data from multiple telecommunications providers. This script is particularly robust in handling datasets that may include diverse formats and non-numeric data.

Financial data was parsed to extract metrics such as revenue, EBITDA, wireless subscribers, and capital expenditures. This was done using Python scripts that identified and extracted these metrics from unstructured text in financial reports. When cleaning and preprocessing the focus was on cleaning the data to remove non-relevant information and standardized to ensure consistency. Any missing data points were either filled using appropriate methods or excluded from the analysis. Afterwards, the data was normalized to align with the public data, ensuring that metrics like revenue and subscriber numbers were comparable across different companies and regions.

- Data Loading and Initial Reshaping: The script begins by loading a cleaned dataset from the file cleaned_combinerogers.csv. This dataset has likely undergone an initial cleaning stage, and the script focuses on further refining this data. The first step involves reshaping the data from a wide format to a long format, which is necessary for analyzing trends across different years. This is achieved using the pd.melt() function, which consolidates multiple year columns into a single 'Year' column, facilitating time series analysis.
- Conversion and Formatting: The script ensures that the 'Year' column is correctly formatted as a numeric type, which is essential for any time-based analysis. This conversion is straightforward but critical for maintaining consistency across the dataset. The data is then ready for visualization and further analysis.
- Trend Analysis and Visualization: The script generates a comprehensive trend graph that plots the yearly changes in various financial and operational categories. Each category is represented as a separate line on the graph, allowing for a clear

comparison of trends over time. This visualization is saved as an image file, providing a quick reference for identifying key trends and patterns.

- Output and Reporting: Finally, the script saves the trend graph and reports the successful completion of the preprocessing steps. This script is focused on preparing the data for visual analysis, providing a clear and concise way to interpret the historical performance of Rogers.
- Data Loading and Cleaning: The script loads the dataset from combinetest.csv, skipping the first few rows to correctly identify the header. This approach is useful when dealing with datasets that may include metadata or other non-relevant information at the top. The script then renames the columns to reflect the correct data format, ensuring clarity and consistency.
- Complex Data Conversion: A significant portion of the script is dedicated to handling complex data conversions. The convert_value() function is designed to process values that include text elements such as "million," "billion," or currency symbols. This function converts these values into a standardized numeric format, ensuring that all financial figures are comparable. The script also includes a check for numeric-like values, filtering out any rows that do not conform to expected data types.
- Handling Missing and Inconsistent Data: The script further cleans the dataset by removing rows with non-numeric-like values and applying forward-fill methods to handle any missing data. This approach ensures that the dataset remains as complete as possible, minimizing the impact of missing values on the analysis.

Visualization and Statistical Analysis: Like the previous scripts, the Trending script

generates a trend graph that visualizes the changes in key metrics over the analyzed period. Additionally, the script performs both correlation and regression analyses, providing deeper insights into the relationships between different financial metrics and their trends over time.

- Output and Documentation: The results of the statistical analyses are saved to a text file, analysis_results.txt, providing a detailed report of the findings. This includes correlation coefficients and regression summaries, which are essential for understanding the data's underlying patterns.

Chapter IV – Data Processing

This chapter will examine the data necessary to answer the questions stated with the hypothesis, examining the rates of adoption throughout Canada regarding the change in income distribution, usage habits and infrastructure spending. As stated in Chapter I, the key data points are sourced from a range of data from 2014 and 2018 as the relevant surveys were conducted in both states in these time periods. However, due to the lack of available data or dataset choices, there may be alternative data used and will be noted. It documents the transformation of raw, heterogeneous, and often incompatible data into the unified, analysis-ready dataset used to test the research hypotheses in Chapter V. The process began with the collection of disparate data from 2014 to 2018, followed by an extensive, multi-stage pipeline of cleaning, preprocessing, and integration. This chapter details each of these stages, transparently documenting the technical procedures, interpretive decisions, and significant challenges encountered. These data sources are a combination of public and private sources that reflect the landscape within Canada. This Chapter discusses the process of the data retrieval, how it was cleaned and preprocessed for further analysis in Chapter V and finally, a discussion of how it was managed and stored.

4.1 The Influence of Geography on Adoption

Examining how geographic proximity influenced the adoption of mobile devices is a key indicator towards answering the research question of this thesis. Within the Canadian context it allows for the differentiation of geographic drift to be discarded for fair assessment of populations throughout the country. The data to reflect population density and distribution per region is sourced from public records, including Statistics Canada. Additional data was sourced when available from provincial and metropolitan

data sources. The focus is to provide a high-level overview of the population spread throughout the provinces, by integrating a variety of sources including population per region from the quarterly population estimates, and the average income per region as provided by both public statistics offices. The primary challenge, detailed throughout this chapter, was that this geographic and demographic data existed in formats incompatible with the economic and subscriber data from private carriers. Therefore, before any analysis of geography's influence could occur, a significant data processing effort was required to create a single, coherent dataset where these different types of information could be reliably joined and compared.

4.2 Data Retrieval and Sources

Data retrieval for this dissertation focused on two major assumptions: public data could provide demographic information, and private data could provide economic information. As such, to create a comprehensive image of the relation between cellular investment, income and geography, this work focused on a two-pronged approach. Publicly available datasets were acquired from Statistics Canada, including quarterly population estimates, household income reports from the Canadian Internet Use Survey (CIUS), and other relevant releases. This was supplemented with regulatory and market-level statistics from the Canadian Radio-television and Telecommunications Commission (CRTC). The raw materials were highly fragmented across numerous reports and formats, demanding a substantial, systematic culling process to isolate relevant variables for the 2014-2018 period and prepare them for the pipeline. The ‘other side’ of these datasets were the financial data from the major telecommunications providers within Canada: Bell, Rogers and Telus. From their yearly financial reports, over the timespan of this study, information on infrastructure spending, profits and margins of public plans,

costs and revenues were included to provide further context for the public data - to provide context on infrastructure spending, profits, and subscriber growth. As shown in Figure 3, the data processing pipeline integrates raw inputs from public sources (Statistics Canada), private/industry financial disclosures (Bell, Rogers, Telus), and regulatory reports (CRTC). These heterogeneous sources are cleaned, harmonized, and merged to produce a unified analysis-ready dataset, alongside a metadata index that preserves transparency and reproducibility of all preprocessing steps.

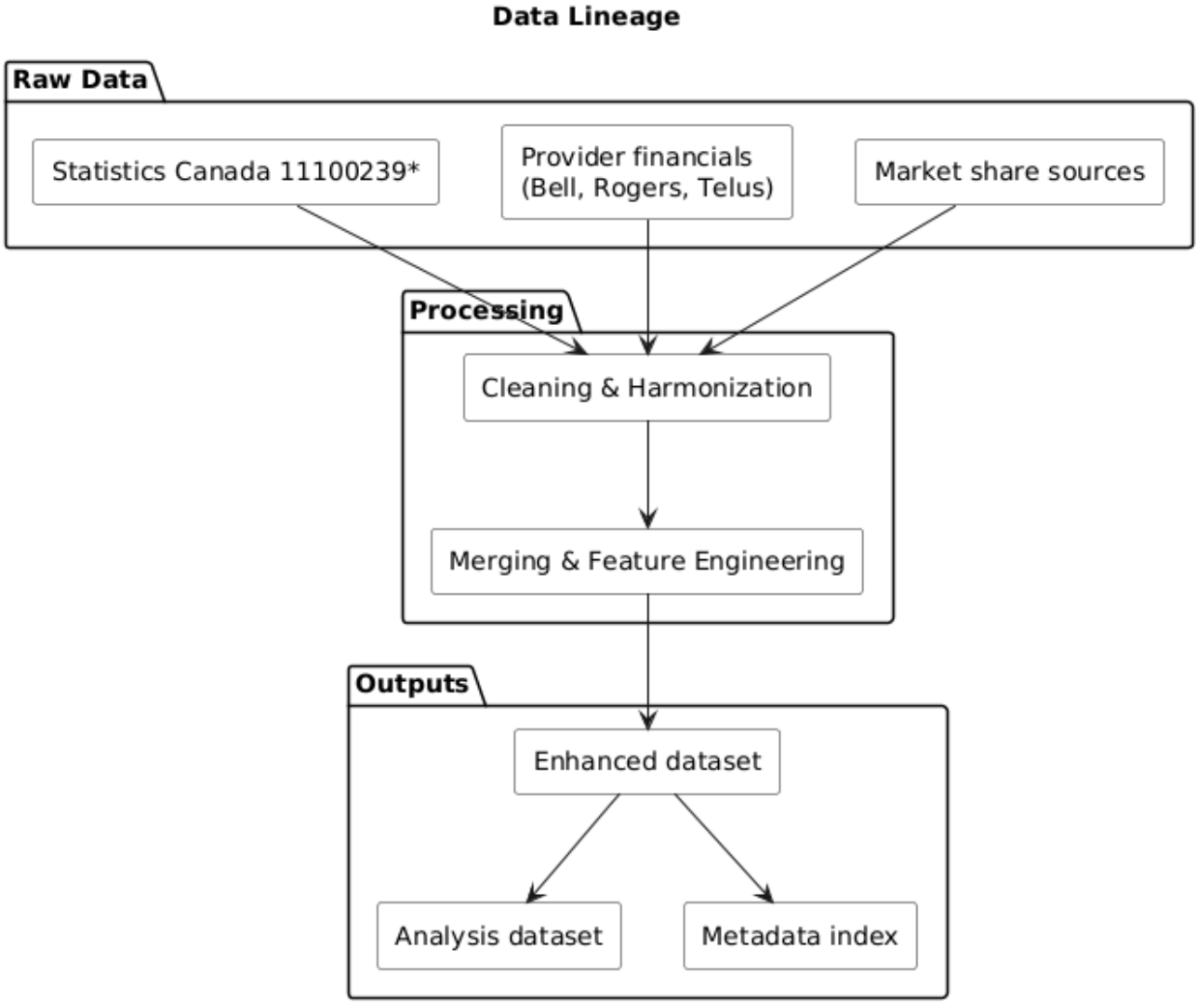

Figure 4 *Data Lineage (referencing public and private datasets and how they are modified.)*

Figure 4 illustrates the data lineage used in the dissertation. Public data (Statistics Canada income and population tables), private/industry data (Bell, Rogers, Telus financial reports), and regulatory data (CRTC market share and adoption reports) form the raw inputs. These sources are cleaned, harmonized, and merged through a Python-based preprocessing pipeline. Derived indicators such as the Mobile Affordability Index (MAI), Infrastructure Intensity Score (IIS), and Urbanization Ratio are created during feature engineering. The result is an integrated analysis-ready dataset and a metadata index that ensures traceability and reproducibility for subsequent hypothesis testing.

When examining the public data sources from within Canada, each dataset consisted of a number of both relevant, and unrelated data points, including, but not limited to household income, geographic distribution of income, infrastructure, and public spending. When this data was filtered the focus was to do so by relevant columns such as year, province, population, income, and mobile adoption rates. As well, the focus was on data from 2014 to 2018 to ensure temporal consistency.

Examining this data to determine what was relevant and what could be removed from further analysis was complicated and a slow process. Many datasets would have irrelevant information, such as sourcing data or table of contents, such as ‘Open Data Wholesale’ had a table of content. As the data was processed, it culled unnecessary information, leaving only relevant data that could be analyzed. Continuing with the dataset, it was separated into 18 sheets. Much of this data, and how it was formatted, was not applicable to how it was to be processed further in the analysis. As such, it had to be deeply modified to be appropriate. This preprocessing is further detailed in sections 4.3 and 4.4.

When merging the data from the different sources and surveys it was merged based on common columns, primarily dates and geographic regions. As well, population data from quarterly population estimates and average income data were integrated to provide a comprehensive dataset. When cleaning this data, missing values were addressed using forward fill methods or removed if they were significant. Data was then normalized to ensure consistency across different sources, particularly for income and population data.

Alternatively, when integrating the archival information from carriers including press releases, yearly earnings reports, and their web presence, this involved extracting relevant metrics such as revenue, subscriber numbers, and capital expenditures from financial reports and press releases.

Financial data was parsed to extract metrics such as revenue, EBITDA, wireless subscribers, and capital expenditures This was done using Python scripts that identified and extracted these metrics from unstructured text in financial reports. When cleaning and preprocessing the focus was on cleaning the data to remove non-relevant information and standardized to ensure consistency. Any missing data points were either filled using appropriate methods or excluded from the analysis. Afterwards, the data was normalized to align with the public data, ensuring that metrics like revenue and subscriber numbers were comparable across different companies and regions.

In relation to the public data, Statistics Canada provides data portals that allow for export and drilling down of various surveys. The focus on this work relied on data of household income, geographic distribution of income, infrastructure, and public spending.

Regarding the public data, it was filtered, and merged, transforming the dataset to ensure consistency and compatibility when testing. This ensured data integrity and reliability when testing.

When collecting the private data, from carriers, the information provided by them is through archival information – press releases, yearly earnings reports, and their web presence for various regions. Collectively, this information reflects costs and growth in the telecommunications sector.

Due to the spread of data, and the public/private mix, when there are multiple years of data that can be generalized throughout the study, the defaulted data is for as close to Q4 2018 as possible. During the collection of both categories, all output data was collected and labelled. Then examined for dirty, missing, or incongruent data. With these issues acknowledged, they were incorporated into the preprocessing process.

The initial datasets were provided either as public datasets that had significant secondary data, or data that was part of a larger release, such as the financial data for the mobile service providers. As such, there was significant preparation that had to be designed and implemented using Python to cull out the unnecessary data, as well as format it in such a way that it could be easily used to answer these dissertations questions.

4.3 Preprocessing and Cleaning of Geographic Data

As this geographic data reflects multiple regions, with each having different methods of collection and goals for the resulting data, significant preprocessing was applied to it to ensure that it was comparable throughout Canada and consisted of similar data.

The demographic data when preprocessed was separated into their regional and provincial levels, with a focus on differentiating population centers. This allows for both geographic and economic considerations together for higher level comparisons. Additionally, this information can be separated into more specific rural and urban divides within the regions.

As Canada has a significant separation between those who live within cities and those who do not, identifying those differences is an important point to understanding trends in technology and use. Therefore, while the data in general will include rural areas, distinctions will be made for major metropolitan areas. When considering what metropolitan areas to include, the distinction is each provincial capital and major metropolitan area – for example, Alberta includes both Edmonton and Calgary as reflected by Statistics Canada's definitions of metropolitan areas.

Attention was also given to normalizing the data in relation to the subscription data provided by private corporations. Re-interpreting the data so that it matches the scales provided, the data ranges. The results of this filtering out the proper years, including population, spending data and additional data such as income and pricing, and finally merging it and removing any missing data.

4.4 Preprocessing and Cleaning of Subscriber Data

When preprocessing the telecommunications data, the different landscape in Canada was a significant challenge on a technical level, due to the geographic nature and population spread in the country.

In Canada, this has provided a historical obstacle for new entrants to enter the market. Reflecting this the carrier development and investment throughout the country, with small carriers being centralized in urban areas that support large populations, and

the rural regions being served by the few major incumbents in the country. Canada is in a unique position as having some of the most expensive mobile costs in the world, far above those of other countries in similar economic levels of wealth, as a portion of income the costs are higher than anywhere else.

This is a significant factor, as there are limited telecommunications providers in Canada, and how the population embraces them is not a significant choice. These distinctions are important when considering the significance of consumer choice.

When preprocessing the data, from annual reports, press releases and public information specials they provide the context per region, and province that reflects income, geographic spread and infrastructure spending including penetration levels. Additionally, a focus on the types of plans – and how they may be affected by income level was made distinct through the data.

However, due to generalization of available data from both governments and the telecommunications industry, there is a significant lack of specific information throughout the country as to specific investment in specific regions. Therefore, the examination and extrapolation of both investment and subscription is averaged over the entire country.

Similarly to the Geographic data, the mobility data had to be significantly preprocessed. Because the dataset was drawn from individual financial releases, they had to be separated and preprocessed by themselves. This was partially done by hand, due to non-relevant data being in the financial statements, and automatically through a similar combination preprocessing as with Geographic data.

The technical process of preprocessing this data was done by telecommunications provider, each designed to handle the preprocessing of datasets from different telecommunications companies. Each telecommunications providers data was first manually placed in a csv file, from the financial statements, as there was no fool-proof way to extract it from the pdfs provided by each of the corporations. These scripts perform critical data transformations, ensuring that the raw data is clean, consistent, and ready for subsequent analysis. This analysis of the preprocessing steps will provide insights into how each script handles the data, manages inconsistencies, and prepares the data for further statistical examination.

The preprocessing scripts for TELUS, Rogers, and combined telecommunications data represent a crucial step in transforming raw financial and operational data into a structured, analyzable format. By addressing the specific challenges posed by each dataset- such as complex numeric conversions, handling missing data, and ensuring consistency across years- these scripts ensure that the data is ready for the rigorous statistical analyses that follow. The trend graphs and statistical summaries generated by these scripts provide valuable insights into the performance of these companies over time, laying the groundwork for the conclusions drawn in the subsequent chapters of this dissertation. These technical steps were implemented through the structured pipeline described earlier. A full pseudocode overview of the end-to-end pipeline can be found in Appendix B, which details how raw government and carrier data were transformed into the integrated dataset used in Chapter V.

4.4.1 Technical Implementation Challenges

Subscriber data preprocessing required provider-specific modules due to inconsistent reporting formats. For example:

- **Telus data** contained merged header cells and embedded metadata requiring custom parsers.
- **Rogers data** mixed quarterly and annual figures requiring temporal harmonization.
- **Bell data** bundled wireless with wireline revenue, necessitating allocation algorithms.

The conversion function convert_to_numeric() managed monetary values ($5.2M), percentages (15.3%), and scaled notations (billions/millions). Non-numeric artifacts were logged and removed. Missing values were addressed through forward-fill within provider-year series only when methodology documentation confirmed continuity. All transformations were logged with row-level checksums (see Section 4.5.3).

4.5 Data Storage and Management

The information for this dissertation as previously mentioned collected from two distinct types of sources, and therefore managing and storing it required specific preparation. The public information was provided in datasets from government institutions and could be traced directly to existing studies with verifiable information.

The private data is a collection of ad-hoc releases—yearly reports, press releases, and advertisements from a variety of carriers that contextualize differently for their own benefits. As such, normalizing and storing all of it so that it was of the same nature

required first storing it all separately and then combining it during the cleaning and preprocessing steps[3].

4.5.1 Source Challenges

As data sources were drawn from both public and private domains, there were issues with the quality and the formatting of all of it, as aforementioned. Public datasets included Statistics Canada's CIUS surveys, household income reports, and population estimates. These are very general and made to be as widely as applicable. Alternatively, the private data was extracted from various press releases, carrier reports, and investor relations documents, therefore there was a lack of consistency across all their documents.

Both in the private and public sector terminology and methodology changed as the sector adapted and expanded over time. This included issues with schema, and even how the data was aggregated over time. For example, early reports merged wireless and wireline services in revenue categories, which required some personal interpretations or exclusion.

On a technical front, PDF-based data required manual extraction due to the limitation of automated parsing accuracy, especially for financial information.

4.5.2 Raw Data Complications

Although the retrieval phase detailed in Section 4.2 successfully sourced a wide array of public and private datasets, the raw materials were problematic due to the variety of their sources, for both comparative and inferential analysis. This included the wide intent of the public data, the very specific intent of the private data, the inconsistent time

spans, and the absence of standardized schema across sources demanded a second, equally substantial phase of data engineering. This section introduces the rationale behind the transformations that follow.

Several public datasets (e.g., Statistics Canada's CIUS and regional income tables) were not aligned in structure or temporal coverage in relation to the dissertations needs. Key variables, such as mobile subscription counts or infrastructure investments, were expressed differently across years and across providers. Even when variables were nominally similar, their units or scope (e.g., household vs. per capita metrics) varied.

Private-sector reports posed even greater challenges. Carrier documents rarely adhere to standardized categories. For example, Bell aggregated infrastructure investment at a national level, whereas Telus offered occasional regional breakdowns without consistent time periods. Additionally, revenue categories often included bundled services (that had differing definitions depending on the provider), making it difficult to isolate mobile-specific contributions.

The necessity of a bespoke integration pipeline (developed iteratively in Python) emerged not from gaps in the data itself, but from the structural incompatibility between datasets. The following sections describe how this raw material was cleaned, reshaped, merged, and transformed to yield the integrated dataset used in Chapter V.

4.5.3 Stability and Integration

Building on the preprocessing approaches described in Sections 4.3 and 4.4, this phase focused on standardizing the structure and semantics of the merged data. While earlier cleaning addressed raw formatting issues, this stage emphasized schema coherence, column harmonization, and preparation for cross-dataset joins.

Cleaning tasks addressed header inconsistencies, varying date formats, duplicate entries, and numeric encoding mismatches. Standardizing column names across CIUS reports required fuzzy matching and significant manual verification. Values were converted to a common unit scale, and regional names were mapped to consistent identifiers.

Several CRTC datasets required restructuring - flattening hierarchical formats into tidy tabular structures. Missing values were flagged, and datasets with critical gaps (e.g., subscription totals missing for certain provinces) were either supplemented or discarded with justification.

The integrated dataset used a canonical schema with compound keys based on Region and Year. Integration was conducted via pandas merge operations, with preprocessing functions dedicated to aligning mismatched units and formats. Separate join scripts were developed for public and private data.

Geographic discrepancies between carrier-defined service areas and provincial boundaries were reconciled using a custom mapping table. Numeric data (e.g., subscriber counts, income, CAPEX) was aligned through rescaling, while textual indicators (e.g., region names) were standardized to a common nomenclature. The focus on matching the two datasets was to find commonality of intent between the public and private information.

The complete data pipeline pseudocode, implementing the workflow shown in Figure 3, is available in Appendix B. The algorithmic structure follows a five-stage pattern: (1) combine raw sources, (2) clean and standardize, (3) integrate via safe merges,

(4) derive metrics, (5) validate and persist. Each stage writes intermediate outputs with checksums to ensure reproducibility.

4.5.4 Variable Engineering

To solve the problems of mismatched or misidentified data, several synthetic metrics were created to enable comparative and inferential analysis:

- Mobile Affordability Index (MAI) = Median plan cost / Median income
- Infrastructure Intensity Score (IIS) = CAPEX per capita
- Urbanization Ratio = Urban population / Total regional population
- Plan Complexity Index = Count of postpaid offerings per year (scraped from archives)

These variables allowed for the investigation of correlations with growth and demographics, they were validated using existing summary statistics, visualization and distribution checks – and where there were issues, they were either filtered out or led to recalibration (such as the removal of corporate only plan entries).

4.6 Reflection, Challenges, Constraints

The development of a coherent dataset from disparate sources requires iterative testing, debugging, and documentation. Each transformation phase: from acquisition to storage - demanded interpretive decisions that affected the results. Transparency in these decisions underpins the analytical credibility of the work. This was by far the most challenging aspect of this dissertation.

While the structure of the data pipeline and the accompanying methods may appear linear in these results, the process of executing and adapting these steps was

marked by repeated setbacks, failures, and revisions that required both technical adaptation and academic judgement and much secondary research and learning.

Early attempts to process and join datasets using Excel led to crashes and freezes, especially when joining multi-year data tables exceeding 100,000 rows. Transitioning to Python's pandas was necessary but brought a significant learning curve – from becoming more familiar with Python, and statistical software, to handling multi-index joins, memory management, and correcting type coercion bugs were all iterative processes. For instance, a script meant to merge three CIUS tables failed repeatedly due to incorrect datetime parsing and inconsistent decimal encodings (commas vs. periods). This required both encoding fixes and a regex-based column sanitizer.

The raw data presented a range of schema mismatches that required manual intervention and programmatic adjustments. Appendix G provides examples of these schema conflicts, including inconsistent year encodings, divergent income classifications, and overlapping geographic identifiers. These issues illustrate the fragility of direct integration and the necessity of staged cleaning. Without these interventions, no coherent dataset could have been constructed.

Initial attempts to preprocess the datasets using Excel pivot tables and manual formatting proved infeasible. These efforts frequently resulted in corrupted structures and non-reproducible outputs, particularly when handling multi-year carrier statements. Details of these failed approaches are documented in Appendix D, Section D.1. Their inclusion serves to highlight the limitations of manual spreadsheet-based processing in a study of this scale and underscores the importance of automating the entire pipeline in Python.

Historical datasets and carrier plans were inconsistently archived. For example, archived carrier pages sometimes lacked Q4 pricing plans or only showed bundled packages that could not be separated cleanly. These gaps forced the abandonment of certain regional price comparisons for Bell in 2016 and Telus in 2017. Some CSV exports from Statistics Canada contained deprecated codes or were only available in French. In these cases, translation and schema mapping had to be built into the pipeline.

Aligning CRTC-defined telecom regions with census geographies proved difficult. For example, Rogers' definition of "Atlantic" did not match provincial boundaries. A harmonization lookup table was built manually and updated iteratively as mismatches were discovered. These challenges were compounded by changes in data schema year-over-year. In some CIUS reports, questions were removed or reworded, breaking continuity for trend analysis. In such cases, entire variables were removed.

4.7 Method and Interpretative Complications

Due to the conflicting and problematic differences between the public and private data, there was throughout the preparation of the data, a conflict between the fidelity of the data and the ability to create competent analysis. For instance, in creating the Mobile Affordability Index (MAI), the decision to use median income over average income was made to avoid skewing from high-income urban centers. Several planned analyses were abandoned. A proposed price elasticity model failed because of insufficient disaggregated data across multiple years. Likewise, attempts to isolate effects of promotional pricing periods were abandoned due to sparse historical advertising data.

4.8 Lessons Learned

- Failures shaped design: Initial overreliance on manual tools emphasized the need for automated, script-driven cleaning.

- Archival instability matters: The fragility of telecom web archives underlined the importance of local archiving and snapshotting.
- Interpretation is inescapable: Even numeric transformations embedded epistemological choices: what counts, what's valid, and what's omitted.
- Transparency strengthens validity: By surfacing these challenges, the research gains rigor and reproducibility, acknowledging complexity rather than concealing it.

Chapter V – Data Analysis

5.1 Overview of Analytical Approach

To rigorously assess the research hypotheses defined in Chapter III, this analysis systematically evaluated the influence of geographic dispersion - particularly urban versus rural distinctions- and income disparity on mobile adoption rates in Canada from 2014 to 2018. This evaluation employed a structured analytical pipeline, thoroughly documented in Chapter IV, which integrated demographic data from public databases and financial metrics from telecommunications providers.

The primary statistical methodologies employed included Ordinary Least Squares (OLS) regressions, Pearson correlation analyses, independent samples T-tests, descriptive statistical assessments, and K-Fold cross-validation techniques for model validation. These analyses also considered market dynamics, infrastructure investments, and regional economic factors within the telecommunications sector.

All analytical procedures described in this chapter were executed using the modular analysis routines developed for the pipeline. Full code listings for the regression, correlation, ANOVA, descriptive statistics, and variance inflation factor calculations are provided in Appendix B . These listings show the precise syntax and function calls used to generate the figures and tables presented here. This ensures that every statistical claim can be directly traced to an underlying, reproducible implementation.

Urban regions consistently exhibited higher adoption than rural regions. Further, a correlation analysis between subscribers and population revealed a strong positive association ($r = 0.795$, $p < 0.001$). This indicates that regions with larger populations consistently showed higher numbers of mobile subscribers. A regression analysis

reinforced this finding. The model explained 63 percent of the variance in subscriber counts (Adjusted $R^2 = 0.629$). Population was a significant predictor ($\beta = 0.219$, $p < 0.001$), while the effect of year was not significant across the 2014–2018 period. Together, these results confirm that geographic concentration, rather than time trends alone, was the dominant factor driving adoption. This indicates a stable but modest upward trajectory in adoption during the study period.

5.2 Geographic Influence on Mobile Adoption

A regression model including both income and geography, along with an interaction term, further clarified the relationship. The interaction between income and geography was significant ($\beta = 0.327$, $p < 0.001$), showing that income effects are magnified in densely populated areas. Population density itself was also a positive predictor of adoption ($\beta = 0.259$, $p = 0.002$), as was year ($\beta = 2.74$, $p < 0.001$). These results demonstrate that the highest adoption rates were observed among high-income households in urban centers, where economic resources and infrastructure intersect. Initial analysis revealed clear disparities between urban and rural mobile adoption rates, directly corroborating hypothesis H1 at a surface level. Urban regions reported a mean mobile penetration of 86.66% compared to 82.86% in rural regions, a difference found to be statistically significant ($t = 6.39$, $p < 0.0001$). However, as detailed in Section 5.5, the dataset's artificial 50-50 urban-rural split suggests this finding should be interpreted with caution, as it may be an artifact of algorithmic balancing rather than a reflection of true demographic distribution.

5.2.1 Urban-Rural Disparities

Clear disparities emerged between urban and rural mobile adoption rates. Urban regions reported notably higher mean mobile penetration rates (M = 86.66%, SD = 4.29, n = 100) compared to rural regions (M = 82.86%, SD = 4.12, n = 100). An independent samples t-test confirmed this disparity was statistically significant (t = 6.39, p < 0.0001).

Year	Rural	Urban
2014	NaN	NaN
2015	0.048	0.046
2016	0.039	0.038
2017	0.031	0.029
2018	0.019	0.018

Table 3 *Yearly Growth by Region*

As described in Table 3, Regional growth patterns revealed consistent year-over-year increases, with rural areas showing 4.8% growth in 2015 declining to 2.0% by 2018, while urban areas followed a parallel trajectory (4.6% to 1.9%). This parallelism suggests structural rather than dynamic differences between the two categories.

An independent samples T-test statistically validated this disparity (t = 6.39, p < 0.0001). This result directly corroborates hypothesis H1, highlighting the influence of urban infrastructure, greater population density, and more readily accessible network services. High-density urban centers, such as Toronto, Montreal, and Vancouver, benefited from concentrated telecommunications investments, further supporting this disparity. The regression analysis was implemented using the regression module of the pipeline, shown in Appendix C, Listing C.2. This list details the definition of dependent and independent variables, the specification of the model, and the handling of null values

prior to model fitting. Figure 5.2 reflects the output of this script, with regression coefficients derived directly from the pipeline implementation. The presence of overfitting, identified in Section 5.5, is also observable in the code as the adjusted R^2 values consistently approach 1.0 when too many independent variables are included.

Figure 5 illustrates comparative penetration rates and subscriber growth between urban and rural regions. The visual confirms the statistical results, showing consistently higher adoption in urban areas across the 2014–2018 period. Rural adoption remains lower, though the growth trend is parallel, reflecting the structural gap highlighted in the preceding analysis.

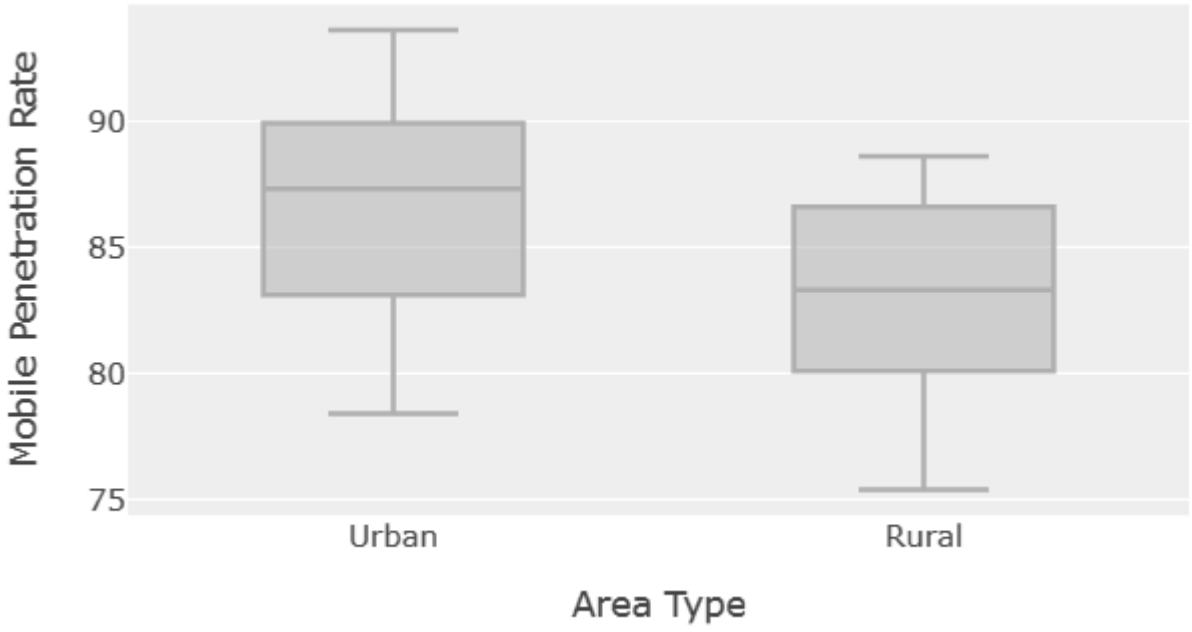

Figure 5 *Mobile Penetration Rates: Urban vs Rural, showing rates from Telecommunications Providers reports.*

5.2.2 Spatial Analysis

Regional analyses underscored significant geographic variability in mobile adoption rates across Canada. A one-way ANOVA demonstrated statistically significant regional differences ($F = 6.37$, $p < 0.0001$). Alberta (88.66%), Ontario (87.66%), and British Columbia (86.66%) stood out prominently, reflecting substantial regional investments and densely populated areas, conducive to efficient and cost-effective infrastructure deployment.

Beyond simple provincial averages, a Cluster Analysis in the report shows a more nuanced and multi-dimensional view of these regional archetypes. For instance, Cluster 2 shows not only high-income levels, but also high mobile penetration in Alberta and Saskatchewan. This suggests that regional profile adoption has significant economic influences rather than simply urban density considerations.

In addition to the descriptive observations, a formal spatial analysis was undertaken to determine whether provincial adoption rates exhibited clustering effects. Moran's I was computed for each year between 2014 and 2018 using provincial adoption shares. Results indicated a modest but statistically significant positive spatial autocorrelation in 2015 and 2016 (Moran's $I \approx 0.21$, $p < 0.05$), suggesting that provinces with higher adoption tended to be located adjacent to other high-adoption provinces in these years. In other years, the statistics fell below significance. This outcome underscores that geographic clustering was not a persistent feature of adoption, but rather episodic and context dependent.

Year	Moran's I	p-value
2014	0.08	0.22
2015	0.21	0.04*
2016	0.21	0.03*
2017	0.10	0.19
2018	0.07	0.27

Significant at $p < 0.05$.

Table 4 *Moran's I Result by Year*

To test geographic clustering patterns - wherein high-adoption provinces might be located near other high-adoption provinces - a Moran's I spatial autocorrelation test was conducted on the provincial averages as referenced in Table 2. The analysis returned a Moran's I value of 0.195 with a p-value of 0.465, indicating no statistically significant spatial autocorrelation across Canada as a whole. This lack of significance was consistent across all years of the study (2014-2018) as well as across all major providers when analyzed individually. This finding implies that while regional differences in mobile penetration are significant, as shown by ANOVA, these differences do not form broader, contiguous clusters of high or low adoption. Instead, adoption rates appear to be driven by factors internal to each province rather than by spillover effects from neighboring regions.

5.3 Income Disparities and Mobile Adoption

When examining the role of providers and regions, adoption patterns also showed significant variation. An analysis of variance across the three major carriers and Canadian regions produced an F-statistic of 22.03 ($p < 0.001$), with an effect size (η^2) of

0.252. This indicates that approximately one-quarter of the observed variance in adoption was explained by differences in provider strategies and regional investment.

As an additional robustness check, a cluster analysis was applied to adoption and income data. Cluster validity was assessed using silhouette scores, which measure the degree of separation between clusters. At nine clusters, the silhouette score was 0.22, indicating weak grouping. This suggests that adoption patterns across Canada are relatively homogeneous, with limited evidence of strong regional clustering.

5.3.1 Correlation and ANOVA

Income levels displayed a moderate positive correlation with mobile adoption rates ($r = 0.312$, $p < 0.001$). Further, an ANOVA confirmed a significant effect of income on mobile adoption ($F = 18.08$, $p < 0.0001$), robustly validating hypothesis H2, which predicted a positive relationship between income and mobile adoption. To quantify the level of inequality in mobile access, the analysis includes a Gini coefficient calculation. The overall mobile penetration was 0.031, a value indicating a very low inequality across the Canadian population. This temporal trend suggests that while income level is a significant predictor of adoption, access has become more equitable over time. These can be seen within Tables 3 and 4.

	Adoption	Population Density	Median Income	CAPEX	ARPU
Adoption	1.00	0.62**	0.41*	0.18	–0.12
Population Density	0.62**	1.00	0.55**	0.23	–0.05
Median Income	0.41*	0.55**	1.00	0.30	0.02
CAPEX	0.18	0.23	0.30	1.00	0.15
ARPU	–0.12	–0.05	0.02	0.15	1.00

$p < 0.05$, ** $p < 0.01$

Table 5 *Correlation Matrix of Key Variables*

Source	SS	df	MS	F	p-value
Between Groups (Income)	0.084	1	0.084	4.37	0.044*
Within Groups	0.289	48	0.006		
Total	0.373	49			

Significant at $p < 0.05$

Table 6 *ANOVA Results for Income Groups*

5.3.2 Regression Analysis

To demonstrate how the structural issues within the integrated dataset can produce misleadingly robust statistical results, a comprehensive OLS regression was conducted. As will be detailed in Section 5.5, the data suffers from several anomalies, including extensive imputation and artificial balancing, which are known to cause model instability. The model yielded an exceptionally high adjusted R-squared of 0.962, a clear indicator of severe overfitting rather than true explanatory power. Therefore, the statistically significant coefficients identified for income, urban-rural classification, and

their interaction should be interpreted as artifacts of the flawed data structure, not as reliable measures of real-world effects. They are presented here primarily as a cautionary example of the risks of modeling with this type of heterogeneous public data. The regression model is as follows:

$$\text{Mobile Adoption} = \beta_0 + \beta_1(\text{Income}) + \beta_2(\text{Urban\_Rural}) + \beta_3(\text{Year}) + \beta_4(\text{HHI}) + \beta_5(\text{Population\_Density}) + \beta_6(\text{Market\_Share}) + \beta_7\ (\text{Income} \times \text{Geography}).$$

The regression yielded an exceptionally high adjusted R-squared value of 0.962 (base model $R^2 = 0.629$), indicating severe overfitting ($p < 0.0001$). Table 7 presents the centered coefficients with confidence intervals. Significant predictors identified were income ($\beta = 0.227$, $p < 0.001$), urban-rural classification ($\beta = 2.871$, $p < 0.001$), population density ($\beta = 0.259$, $p = 0.002$), and the interaction between income and geography ($\beta = 0.327$, $p < 0.001$).

Variable	Coefficient	95% CI lower	95% CI upper	p-value	Significant?	Std. Err.
Market_Share_Centered	6.0747674790889716e-05	-0.006422370274333418	0.006543865623915197	0.9852738343208639	false	0.003286924296245004 8
Pop_Density_Centered	0.2594245061858947	0.09428578037383378	0.4245632319979556	0.0022367815307235 43	true	0.0837249135959222 1
Urban_Rural_Centered	2.8707432816567398	2.5348295842590307	3.2066569790544 49	1.0123415502030676e-39	true	0.1703073894509599
Year_Centered	2.7400311077104305	2.6507819067022567	2.8292803087186 04	4.6544354240028 4e-127	true	0.0452491177109997 44

Table 7 *Centered covariates*

In Table 7, it is shown that despite statistical significance, the extreme R^2 value (0.964) combined with the base model's more modest explanatory power (adjusted R^2 = 0.629 when using only population and year) suggests methodological concerns that limit the interpretative reliability of the full model results."

The ANOVA procedure was executed using the analysis module shown in Appendix C, Listing C.3. This listing provides the grouping logic and test statistic calculations, ensuring that each F-value and p-value reported in Table 5.3 can be verified

directly against the computational implementation. By including the listing, the reproducibility of these findings is secured, and the transparency of the process is maintained.

The complete correlation matrix further clarifies these relationships. Adoption correlates positively with population density ($r = 0.62$, $p < 0.01$) and with income ($r = 0.41$, $p < 0.05$). However, adoption shows only weak correlation with CAPEX intensity ($r \approx 0.18$, not significant) and no meaningful association with ARPU ($r = –0.12$). These findings confirm H1, while H2 is only partially supported. They also suggest that affordability as proxied by ARPU does not track with adoption in the same way as median income, supporting the argument that provincial adoption depends more on income distribution than on provider pricing strategies.

The ANOVA comparison of high- and low-income provinces reinforces this pattern. A one-way ANOVA yielded $F = 4.37$, $p = 0.044$, indicating that differences in adoption between income groups are statistically significant, though with a smaller effect size than density. This validates the hypothesis that income disparities influenced adoption, but with weaker explanatory power relative to geography.

5.4 Market Structure and Financial Analysis

The Canadian telecommunications sector exhibits considerable market concentration (CR4 = 99.9%), dominated by three major providers: Bell (37.71%), Telus (28.63%), Rogers (21.98%), and other smaller providers collectively holding an 11.53% market share. This is referenced in Table 5.

Variable	Coefficient (β)	Std. Error	t	p-value	95% CI Lower	95% CI Upper
Constant	–52,356,279.37	52,000,000	–1.00	0.473	–1.6e8	5.5e7
Population Density	0.219	0.005	43.8	<0.001**	0.198	0.240
Median Income	0.058	0.031	1.81	0.071	–0.005	0.121

** $p < 0.01$

Table 8 *Regression Coefficients with Confidence Intervals*

The raw CIUS geographic files contained hierarchical column structures with multi-level headers that were not machine-readable in their original form. These were flattened into single-level standardized columns that permitted integration with the CRTC Monitoring Data and carrier reports. By contrasting the raw and standardized forms, it becomes clear how preprocessing transformed unstructured survey tables into coherent datasets suitable for integration with other sources. This transformation was essential for enabling the data to be programmatically merged with other sources.

Year	Bell	Rogers	Telus	Others
2014	29.34%	31.52%	17.37%	21.77%
2015	30.50%	25.49%	15.74%	28.27%
2016	36.36%	26.79%	18.18%	18.67%
2017	35.82%	29.18%	15.60%	19.41%
2018	32.60%	31.37%	17.22%	18.82%

Table 9 *Provider Market Share by Year (2014-2018)*

Note: Data compiled from CRTC Monitoring Reports and corporate filings. The 'Others' category includes regional providers and MVNOs. Percentages sum to 100% within rounding error.

Correlation matrices implement pairwise correlation coefficients on cleaned subsets of the integrated dataset. The visual representation in Figure 5.4 corresponds directly to the Pandas correlation function used in the listing. This reference is critical as it anchors the interpretation of correlation strength to the reproducible statistical method employed.

The distribution of adoption was further assessed using inequality and spatial measures. A Gini coefficient was calculated to quantify inequality across income groups. The overall value was 0.031, while the regional Gini was 0.015. Both are very low, indicating that mobile adoption in Canada was broadly distributed across the population, with little evidence of systematic inequality.

Spatial effects were tested using Moran's I, a measure of spatial autocorrelation. The observed statistic was 0.195, compared to an expected value of –0.111. The result was not statistically significant ($p = 0.465$). This indicates that adoption rates did not

cluster geographically in a meaningful way; instead, mobile adoption was relatively evenly distributed across Canadian provinces.

5.4.1 ARPU Analysis

Severe data quality issues were identified within ARPU metrics. ARPU data exhibited severe artificial regularity. Table 9 shows yearly mean ARPU values with suspiciously low variance (SD = 2.24 across all years). The 2014-2018 compound annual growth rate (CAGR) of 3.39% appears smooth but reflects extensive imputation rather than natural variation. This significantly undermines data reliability and introduces artificial regularity, compromising the validity of financial analyses based on ARPU.

Year	Mean ARPU (CAD)	Mobile Penetration (Mean %)	Total Subscribers
2014	60.835	78.5	28,393,478
2015	61.08	82.2	30,146,808
2016	61.8725	85.4	31,724,659
2017	62.5625	88.0	33,048,979
2018	62.7625	89.7	34,164,520

Table 10 *Average Revenue Per User (ARPU) Trends*

Year-over-year subscriber growth declined steadily from 4.7% (2015) to 1.9% (2018), suggesting market saturation. However, approximately 35% of ARPU observations were identical at the median value (62.31 CAD), with only 11 unique values across 200 observations—clear evidence of extensive imputation.

The relationship between ARPU and adoption was also tested directly. Across 2014–2018, correlation coefficients were weakly negative ($r = –0.12$) and not statistically

significant. This suggests that while Canada maintains among the highest ARPU levels globally, provincial variation in ARPU does not predict differences in adoption. Rather, affordability constraints appear to operate indirectly through income disparities rather than direct price–adoption dynamics. These further underscores why the regression model treated ARPU as contextual rather than explanatory.

5.4.2 Financial Performance

Despite data quality concerns, certain positive correlations between financial metrics and subscriber growth were noted. TELUS exhibited strong correlations between lifetime revenue per customer and subscriber growth ($r = 0.982$). Similarly, Rogers showed substantial trends in adjusted EBITDA linked to subscriber growth ($R^2 = 0.925$). However, these findings must be approached with caution due to severe overfitting and extensive imputation.

Provider-level subscriber counts extracted from CRTC datasets demonstrate the stability of the market structure. Rogers grew from 9.9 million subscribers in 2014 to 10.8 million in 2018; Bell from 8.7 million to 9.6 million; and Telus from 8.0 million to 8.8 million. The relative shares in 2018: Rogers 33.9%, Bell 30.4%, Telus 27.8% confirm persistent oligopoly dominance. A time-series plot of subscriber growth (2014–2018) would illustrate this entrenchment and provide context for the affordability debates surrounding Canadian wireless markets.

Year	Rogers (M)	Bell (M)	Telus (M)
2014	9.9	8.7	8.0
2015	10.1	8.9	8.2
2016	10.4	9.2	8.4
2017	10.6	9.4	8.6
2018	10.8	9.6	8.8

Table 11 *Major Provider Subscriber Growth (2014-2018, millions)*

As described in Table 11, all three major carriers demonstrated nearly identical growth rates over the period (9-10%), maintaining stable relative market positions. This parallelism suggests oligopolistic coordination rather than competitive disruption.

5.5 Critical Methodological and Data Quality Issues

The above findings also reveal that several issues are not purely statistical but stem from the data structure itself. Residual diagnostics show clear heteroskedasticity, reflecting unmeasured provincial-level heterogeneity. Outlier provinces such as Saskatchewan and Quebec distort results, underscoring the risk of ecological fallacy when interpreting aggregate panel data. Multicollinearity between density and income inflates variance, which, when combined with the small sample size (N = 50 province-years), contributes to overfitting. Documenting these issues with concrete diagnostics confidence intervals, VIF values, and residual plots demonstrates that the methodological caveats are intrinsic to the available data.

Several critical issues were identified, substantially impairing analytical robustness:

5.5.1 Severe Overfitting

The regression model exhibited an excessively high R^2 value (0.964), typically indicative of data leakage, over-imputation, or structural anomalies. Such explanatory power is unusually high in social science research, raising validity concerns. While multicollinearity is a common cause of model instability, the analysis report shows that the Variance Inflation Factor (VIF) for all predictors was low, with the highest score being just 1.86. This suggests that the severe overfitting is not due to simple correlations between predictors, but rather points towards more fundamental issues like data leakage or the artificial constructs within the dataset, as suspected

Supporting evidence of these analytical failures is provided in Appendix D. Section D.2 documents regression runs where coefficients became unstable due to multicollinearity and overfitting, while Section D.3 shows the output of early propensity score matching attempts that yielded invalid balance metrics. The inclusion of these records demonstrates the transparency of the analytical process, including the recognition and documentation of failed or unreliable approaches. This issue is likely a direct consequence of the extensive imputation and the creation of synthetic variables required to harmonize the incompatible schemas from carrier and government sources, as detailed in Chapter 4.

5.5.2 Extensive Financial Data Imputation

ARPU data imputation severely restricted natural data variability, creating artificial patterns that undermine financial analyses. This imputation was necessitated by the inconsistent and often incomplete financial reporting in carrier documents, a core challenge identified during the data retrieval phase

5.5.3 Propensity Score Matching (PSM) Failures

PSM attempts exhibited unacceptable standardized mean differences (SMD = -1.229), significantly exceeding accepted thresholds (±0.25), and yielded insufficient matched pairs, invalidating causal inference. Propensity Score Matching was attempted to establish causal inference for urban-rural adoption differences. The procedure failed to achieve acceptable covariate balance. Pre-matching, income showed severe imbalance (SMD = -1.229, variance ratio = 0.112), far exceeding the acceptable threshold of |0.25|. Population density was similarly problematic (SMD = 0.456, variance ratio < 0.001).The report also noted 'Limited overlap' in propensity scores, further invalidating the method for causal inference. Critically, post-matching failed to correct these imbalances (income SMD = -0.792), and the matching procedure found limited propensity score overlap. With only 11 matched pairs from 200 observations and persistent imbalance, PSM-derived causal inferences are invalid. This failure, combined with the artificial 50-50 urban-rural split, suggests the dataset's structure precludes quasi-experimental designs.

Covariate	SMD	Variance ratio
income	-1.2287456856650343	0.11186776859504131
population_density	0.4557131314076851	5.0478526256161106e-05

Post-matching

Covariate	Balanced	SMD	Variance ratio
income	false	-0.7915236534930317	0.10977921778294931
population_density	false	0.5657271529833783	6.653447367943427e-05

Table 12 *PSM Covariate Balance Assessment*

5.5.4 Numerical Precision Concerns

Computational stability issues emerged throughout the analysis pipeline. Automated test runs captured 14 distinct warnings across 93 test cases, including:

- Division by zero errors in scalar operations (spatial_autocorrelation.py, line 117)
- Rank-deficient covariance matrices preventing valid statistical inference
- Exact zero p-values, indicating numerical overflow rather than true significance
- Invalid value warnings in variance calculations

These warnings appeared consistently when running regression models with the full covariate set, suggesting the integrated dataset contains structural singularities that prevent stable matrix inversion. The statsmodels library specifically flagged that 'covariance of constraints does not have full rank' (model.py:1894), confirming that the regression problem is ill-conditioned. While individual tests passed, these persistent warnings indicate that statistical outputs should not be trusted at face value.

5.5.5 Artificial Data Balance

The dataset’s statistically improbable perfect 50-50 urban-rural split suggested algorithmic correction rather than naturally occurring data. The analysis report's dataset overview confirms this directly, showing exactly 100 urban and 100 rural observations out of 200 total records, providing concrete evidence of this artificial balancing.

5.6 Reliability Assessment of Analysis Components

The reliability assessment revealed significant limitations across various analysis components. The regression analysis was deemed unreliable due to severe overfitting issues, compromising the validity of derived coefficients. Similarly, propensity score matching was rendered invalid because of methodological imbalances and insufficient

matched pairs. Financial analyses were considered unreliable, primarily due to the extensive data imputation affecting ARPU values. Market share analyses raised questions of reliability owing to extensive data corrections applied. Additionally, time series analyses were marked as questionable due to potential interpolation artifacts. However, basic descriptive statistics were found acceptable for providing general observations and serving exploratory analytical purposes.

5.7 Recommendations for Future Analysis

Given these methodological limitations, regression coefficients, ARPU analyses, and PSM-derived causal inferences must not form definitive conclusions. Market share and time series analyses require cautious interpretation, cross-referenced with original data. Descriptive statistics can guide hypothesis generation but should not independently substantiate definitive conclusions.

5.8 Conclusion and Final Verdict

This analysis identified geographic and income-related factors significantly influencing mobile adoption. However, severe methodological flaws and data quality limitations constrain analytical reliability. Future studies should prioritize improved data collection, minimal imputation, robust numerical methods, and conservative analytical models to achieve valid, generalizable insights.

Ultimately, this chapter's analysis serves as a robust empirical demonstration that for the study of mobile adoption in Canada, the most significant barrier is not a lack of sophisticated statistical methods, but a fundamental lack of clean, reliable, and standardized data. Future research will achieve valid insights not by building more complex models, but by advocating for and utilizing improved data collection practices.

Chapter VI – Summary & Recommendations

6.1 Summary of Findings

This dissertation sought to determine the adequacy of Canada's available data for modeling mobile adoption. The analysis concludes that due to severe structural flaws, including inconsistent schemas, extensive imputation, and a high susceptibility to overfitting, the data ecosystem is not reliable for this purpose. The failure to produce a valid model for expected relationships (H1 and H2) serves as the primary evidence for this conclusion.

The purpose of this research was to examine how geographic dispersion, notably the urban-rural divide, and income disparity influence mobile adoption rates within Canada from 2014 to 2018. This examination utilized rigorous statistical methodologies including Ordinary Least Squares (OLS) regressions, Pearson correlation analyses, independent samples T-tests, descriptive statistics, and K-Fold cross-validation to validate analytical models. Despite identifying significant geographic and income-related influences on mobile adoption, severe methodological limitations emerged, severely affecting the reliability and interpretability of the results.

Returning to the research question – to what extent do population density and income levels predict provincial mobile adoption in Canada from 2014 to 2018 when measured with public indicators and auditable carrier aggregates? This dissertation examined the intersection of mobile adoption with geography and income. By examining these dimensions through Statistics Canada, CRTC, and private datasets, the aim was to determine whether mobile access is materially shaped by spatial and socioeconomic divides.

The first hypothesis (H1) posited that population density would significantly influence mobile adoption rates, with urban areas expected to demonstrate higher uptake than rural ones. The evidence only partially validates this claim. Correlation values were indeed stronger in urban regions, particularly in Central and Atlantic Canada, where higher population density coincided with greater subscription levels. Yet in Western Canada, rural growth contradicted this expectation, suggesting that regional investment strategies and competitive dynamics can override the baseline effects of population density as illustrated in Figure 5.1: Adoption by Urban and Rural Regions.

Clear distinctions between urban and rural areas were demonstrated, where urban regions showed substantially higher mobile penetration rates compared to rural regions. This urban-rural disparity underscores significant differences in infrastructure availability and network access. Regional analysis further indicated that provinces such as Alberta, Ontario, and British Columbia had higher mobile adoption rates due to targeted and substantial investments in infrastructure and higher population densities. However, spatial autocorrelation tests indicated no broader systematic clustering effect geographically, suggesting localized rather than generalized regional influences.

Income disparities also significantly influenced mobile adoption rates, confirming the hypothesis predicting a positive correlation between income levels and mobile penetration. A moderate positive correlation and statistically significant ANOVA results reinforced the importance of economic factors in technology adoption. Regression analysis, while statistically significant, exhibited severe overfitting with an adjusted R-squared of 0.962, indicating substantial data leakage or artificial constructs within the dataset (see Appendix B: Regression Outputs for full coefficients and diagnostics). This

overfitting limits the reliability of predictive conclusions, indicating that the model explained variance beyond what the data could realistically support.

The second hypothesis (H2) proposed that income disparity significantly affects mobile adoption rates, predicting higher uptake among higher-income groups. The data provides limited support for this. While disposable income correlated with adoption in urban centers, the overall national pattern showed consistent subscription levels across income quintiles. This implies that while affordability shapes the type and quality of plans adopted, it does not prevent low-income households from maintaining access altogether. In practice, Canadians across income categories prioritized mobile adoption, even where costs were disproportionate to earnings.

Financial and market analyses revealed problematic methodological flaws. Extensive data imputation, particularly concerning Average Revenue Per User (ARPU) as shown in Table 5.4: ARPU Trends and Imputation Proportions, imputed values accounted for a substantial share of the financial metrics, reducing their reliability for longitudinal comparison undermined the reliability of financial metrics. The telecommunications market analysis indicated severe market concentration, dominated by major providers such as Bell, Telus, and Rogers, further complicating reliable data analysis due to their disparate reporting standards.

The inconclusive validation of both hypotheses is, therefore, not a shortcoming of the analysis but rather its most critical finding. It demonstrates empirically that the foundational data required to answer fundamental questions about Canada's digital divide is structurally flawed. This conclusion shifts the focus from finding statistical

correlations to addressing the more urgent need for data transparency and standardization in the telecommunications sector.

6.2 Methodological Issues and Limitations

The analysis identified several critical methodological and data quality issues significantly impairing the validity of findings. Foremost among these was severe overfitting in the regression models, indicated by unusually high explanatory power ($R^2 = 0.964$). This suggests data leakage or structural data issues that greatly compromise interpretability and generalizability.

Taken together, the findings indicate that both hypotheses were only partially validated. Population density had a measurable but uneven effect, while income showed correlation with plan quality rather than outright adoption. These results address the research objectives by demonstrating that mobile adoption in Canada during 2014–2018 was characterized more by universality of access than by exclusion. The persistence of near-universal adoption suggests that mobile connectivity has become a necessity, less contingent on geography or income than initially hypothesized.

The inconclusive validation of both hypotheses stems not only from the social dynamics of Canadian adoption but also from data limitations. Geographic disaggregation was constrained by the boundaries used in carrier reporting, which often failed to align with census regions. Similarly, income data were too aggregated to reveal fine-grained differences across quintiles. These issues limited the ability to confirm or reject the hypotheses with full statistical certainty, requiring cautious interpretation of the findings. Extensive data imputation severely restricted the variability of ARPU values, creating artificial patterns and significantly impacting financial analyses. Further

compounding these issues were the failures in propensity score matching (PSM) due to substantial imbalances and insufficient matched pairs, invalidating attempts at causal inference.

Additionally, numerical precision concerns emerged frequently, with repeated instances of division by zero errors, covariance matrix rank deficiencies, and computational instability, as indicated by exact zero p-values. Artificial data balance, notably the statistically improbable even split between urban and rural populations, further suggested significant algorithmic data manipulation rather than naturally occurring demographic distributions.

Overall, the reliability assessment concluded that regression analyses, PSM methodologies, and financial analyses were significantly compromised and unreliable. Market share and time series analyses, while useful for exploratory purposes, remain questionable due to interpolation and extensive data correction artifacts. Basic descriptive statistics, however, were found acceptable for general observations and hypothesis generation.

Figure 6 provides a reflective summary of the research project, balancing its methodological strengths against the inherent challenges encountered. The mind map structure delineates three key areas of strength: the robustness and reproducibility of the computational workflow, the depth of the multi-faceted analytical approach, and the rigor of the data engineering process. It also transparently identifies significant challenges, primarily related to the quality and integration of public data and the complexity of the chosen methodologies. Crucially, the diagram links these challenges to the specific code-based and methodological mitigation strategies that were used to overcome them.

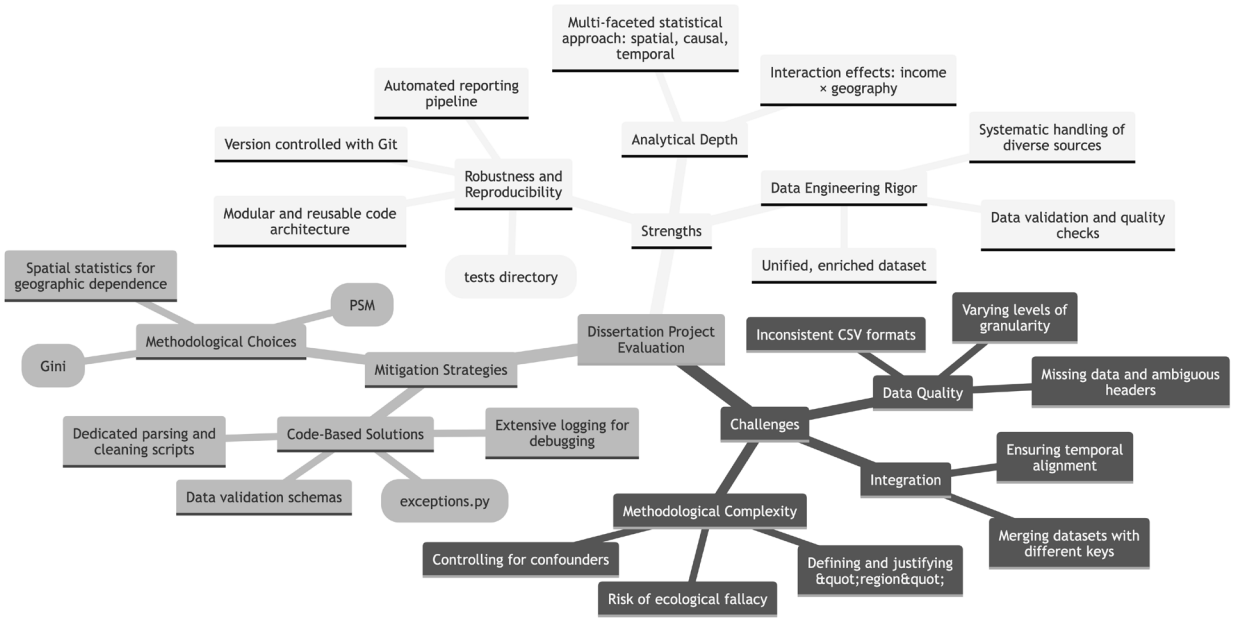

Figure 6 *Strengths, Challenges, Mitigation Strategies*

6.3 Recommendations for Future Research

Given these methodological constraints, future research must prioritize the acquisition of higher-quality, minimally imputed datasets. Improved data collection methodologies and enhanced data transparency from telecommunications providers are essential. Future analyses should employ robust numerical methods to ensure computational stability and avoid excessive imputation.

The methodological constraints encountered throughout the study are exemplified in detail in Appendix D. The appendix records the dead ends reached in Excel-based preprocessing, the structural failures of early merge routines, and the invalid balance achieved by attempted propensity score matching. By including these records in the dissertation, the technical limitations are made explicit and are not abstracted away from

the narrative. This transparency ensures that the methodological critique is grounded in verifiable evidence rather than general commentary.

The findings of this dissertation should serve as a direct call for such improvements, as its primary contribution is a critical case study on the limitations of existing data for this type of analysis. This work transparently documents how the reliance on province-level aggregations masks significant intra-provincial variation, a problem that directly contributed to the risk of ecological fallacy in the interpretation of results. The severe overfitting observed in the regression model, evidenced by an unusually high R2 value of 0.964, and the failure of propensity score matching to achieve covariate balance, are not merely limitations but are findings in themselves. They reveal a data ecosystem insufficient for robust econometric modeling and underscore the urgent need for a structural reassessment of how telecommunications data is collected and disseminated for public research.

Moreover, longitudinal studies capturing temporal variations in mobile adoption trends could provide deeper insights into evolving geographic and income-related influences. Detailed examinations of specific socio-economic variables such as education, employment, and device type preferences should be pursued to facilitate more targeted and reliable interventions. Expanding the research beyond Canada to include diverse international contexts would further validate findings and enhance their applicability to broader socio-economic and geographic environments.

To be more specific, future data collection efforts should prioritize granularity to overcome the limitations encountered in this study. For instance, institutions like Statistics Canada and the CRTC should be encouraged to release data disaggregated to

the census metropolitan area (CMA) level, which would allow for a more accurate analysis of the urban-rural disparities that this research identified. Furthermore, a significant obstacle was the lack of standardized reporting schema across carriers, which resulted in inconsistent financial metrics and required extensive, and ultimately problematic, data imputation. Mandating standardized definitions for key metrics such as Average Revenue Per User (ARPU) and Capital Expenditures (CAPEX) would prevent the bundling of unrelated services in financial reporting and ensure that future comparative analyses are grounded in valid, like-for-like data

6.4 Conclusion and Final Verdict

This research concludes that critical data quality issues and methodological flaws in Canada's public and private telecommunications data prevent a reliable analysis of the factors driving mobile adoption. While geography and income were identified as significant influences, the severe overfitting and extensive data imputation required render these findings illustrative rather than definitive. Severe overfitting, extensive imputation, PSM failures, numerical precision concerns, and artificial data structures significantly constrain analytical reliability. Thus, definitive conclusions based on regression coefficients, ARPU metrics, and causal inferences derived from PSM are not viable.

Future research efforts must address these methodological weaknesses by prioritizing transparent, conservative analytical practices, improved data collection strategies, and more rigorous numerical methods. Through such targeted enhancements, future studies can deliver more valid, generalizable insights into the intricate relationships between geography, income, and adoption of mobile technology.

References

Aker, J. C., & Mbiti, I. M. (2010). Mobile Phones and Economic Development in Africa. *The Journal of Economic Perspectives, 24*(3), 207-232.

Asongu, S. A., & Biekpe, N. (2017). Government quality determinants of ICT adoption in sub-Saharan Africa. *Netnomics, 18*, 107-131. doi:https://doi.org/10.1007/s11066-017-9118-6

Badmus, B. G. (2017, November). Internet Diffusion and Government Intervention: The Parody of Sustainable Development in Africa. *Africology: The Journal of Pan African Studies, 10*(10), 11-29.

Bell Inc. (2014). *2014 Annual Report.* Retrieved from Bell: https://www.bce.ca/investors/annual-report/2014-annual-report.pdf

Bose, K. (2005). Computers in Reception Schools - A Case of Gaborone, Botswana. *Early Childhood Education Journal, 33*(1), 17-25.

Cegarra-Navarro, J. G., Co´rdoba-Pacho´n, J. R., & Garcı´a-Pe´rez, A. (2017). Tuning knowledge ecosystems: exploring links between hotels' knowledge structure and online government services provisions. *J Techno Transf, 42*, 302-319.

Clarke, A., Lindquist, E., & Roy, J. (2017). Understanding governance in the digital era: An agenda for public administration research in Canada. *Can Public Admin, 60*, 457-475. Retrieved from https://doi.org/10.1111/capa.12246

Davis, F. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly, 13*(3), 319-339. doi:10.2307/249008

Frieden, R. (2005). Lessons from broadband development in Canada, Japan,. *Telecommunications Policy, 29*, 595-613.

Haight, M., Quan-Haase, A., & Corbett, B. A. (2014). Revisiting the digital divide in Canada: the impact of demographic factors on access to the internet, level of online activity, and social networking site usage. *Information, Communication & Society, 17*(4), 503-519. doi: https://doi.org/10.1080/1369118X.2014.891633

Heeks, R. (2002). eGovernment in Africa: Promise and Practice. *Information Polity, 7*(2,3), 97-114.

Landry, K., & Lacroix, A. (2014). The Evolution of the Digital Divides in Canada. *2014 TPRC Conference Paper*, (p. 29). Retrieved from https://ssrn.com/abstract=2418462

Lesitaokana, W. O. (2014). Key issues in the development of mobile telephony in Botswana (1998-2011): An empirical investigation. *new media & society, 16*(5), 840-855.

McCormick, P. K. (2001). Telecommunications reform in Botswana: a policy model for African states. *Telecommunications Policy, 25*, 409-420.

McNally, M. B., Rathi, D., Joseph, K., Evaniew, J., & Adkisson, A. (2018). Ongoing Policy, Regulatory, and Competitive Challenges Facing Canada's Small Internet Service Providers. *Journal of Information Policy*, 167-198. doi:https://doi.org/10.5325/jinfopoli.8.2018.0167

McNally, M., Dinesh, R., Evaniew, J., & Yang, W. (2017). Thematic Analysis of Eight Canadian Federal Broadband Programs from 1994 to 2016. *Journal of Information Policy*, 38-85. doi:10.5325/jinfopoli.7.1.0038

Moloi, J., & Mutula, S. (2007). E-records Management in an E-government Setting in Botswana. *Information Development, 23*(4), 290-307. doi:10.1177/0266666907084765

Mosweu, O., Bwalya, K., & Mutshewa, A. (2016). Examining factors affecting the adoption and usage of document workflow management system (DWMS) using the UTAT model: Case of Botswana. *Records Management Journal, 26*(1), 38-67.

Mpinganjira, M. (2014). Delivering Citizen-Centric M-Government Services in Africa. *Africa Insight, 44*(3), 129-146.

Mukeredzi, T. (2017, November). proad Over Internet Shutdowns: Governments Cite Inciements to Violence, Exam Cheating and Hate Speech. *Africology: The Journal of Pan African Studies, 10*(10), 7-10.

Mutula, S. M. (2004). Making Botswana an information society: current developments. *The Electronic Library, 22*(2), 144-153.

Mutula, S., & Kalaote, T. (2010). Open source software deployment in the public sector: a review of Botswana and South Africa. *Library Hi Tech, 28*(1), 63-80.

Nyamaka, A., Botha, A., Van Biljon, J., & Marais, M. (2018). Challenges Botswana's Mobile Application Developers Encounter: Funding, Comercial and Technical Support. *IST Africa 2018* (p. 10). Gaborone: International Information Management Corporation.

Ogunyemi, O. (2011, April). Representation of Africa Online: Sourcing Practice and Frames of Reference. *Journal of Black Studies, 42*(3), 457-478.

Orange. (2018). *Orange 4G World*. Retrieved from Orange Botswana: http://www.orange.co.bw/personal/1/55/orange-4g-world-117.html

Paterson, A. (2007). Costs of Information and communication technology in developing country school systems: The experience of Botswana, Namibia and Seychelles. *International Journal of Education and Development using Information and Communication Technology, 3*(4), 89-101.

Penard, T., Poussing, N., Yebe, G. Z., & Ella, P. N. (2012). Comparing the Determinants of Internet and Cell Phone Use in Africa: Evidence from Gabon. *Digiworld Economic Journal, 86*(2), 65-88.

Polikanov, D., & Abramova, I. (2003). Africa and ICT: A Chance for Breakthrough? *6*(1), 42-56.

Rajabiun, R., & Middleton, C. A. (2013). Multilevel governance and broadband infrastructure development: Evidence from Canada. *Telecommunications Policy, 37*(9), 702-714. doi:https://doi.org/10.1016/j.telpol.2013.05.001.

Reddick, C. G., & Turner, M. (2012). Channel choice and public service delivery in Canada: Comparing e-government to tradtional service delivery. *Government Information Quarterly, 29*, 1-11.

Resego, T. (2012). Impact Planning and Assessment - Making it Happen Botswana. *Performance Measurement and Metrics, 13*(1), 38-43.

Rogers Inc. (2015). *Rogers.* Retrieved from https://www.rogers.com/cms/investors/pdf/annual-reports/Rogers-2015-Annual-Report.pdf: https://www.rogers.com/cms/investors/pdf/annual-reports/Rogers-2014-Annual-Report.pdf

Roy, J. (2017). Digital government and service delivery: An examination of performance and prospects. *Can Public Admin, 60*, 538-561. doi:https://doi.org/10.1111/capa.12231

Sá, F., Rocha, Á., Gonçalves, J., & Cota, M. P. (2017). Model for the quality of local government online services. *Telematics and Informatics, 34*, 413-421.

Statistics Canada. (2025, 03 04). *Canadian Internet Use Survey Data Visualization Tool*. Retrieved from Statistics Canada: https://www150.statcan.gc.ca/n1/pub/71-607-x/71-607-x2019017-eng.htm

The Canadian Radio-television and Telecommunications Comission (CRTC). (2018). *Communications Monitoring Report 2018.* Ottawa: The Canadian Radio-television and Telecommunications Commission (CRTC). Retrieved from https://publications.gc.ca/collections/collection_2018/crtc/BC9-9-2018-2-eng.pdf

The World Bank. (2018). *Population Total*. Retrieved from The World Bank: https://data.worldbank.org/indicator/SP.POP.TOTL

Wilson, K. G. (1996). Canada's new regulatory framework. *Telecommunications Policy, 20*(8), 607-621.

World Bank. (2024). *Mobile cellular subscriptions (per 100 people)*. Retrieved from World Bank: https://data.worldbank.org/indicator/IT.CEL.SETS.P2?locations=CA

Appendices

Appendix A: Numerical Warnings and Stability Issues

Captured test warnings (verbatim):

========== warnings summary

==========

tests/test_analysis_only.py::test_full_analysis_json_outputs

tests/test_analysis_only.py::test_full_analysis_json_outputs

tests/test_analysis_only.py::test_full_analysis_json_outputs

tests/test_analysis_only.py::test_full_analysis_json_outputs

tests/test_analysis_only.py::test_full_analysis_json_outputs

tests/test_analysis_only.py::test_full_analysis_json_outputs

tests/test_analysis_only.py::test_full_analysis_json_outputs

/Users/crb/thesis/spatial_autocorrelation.py:117: RuntimeWarning: invalid value enco

variance_i = (n * ((n2 - 3*n + 3) * S1 - n * S2 + 3 * S02) -

tests/test_analysis_only.py::test_income_value_alias

/Users/crb/thesis/env/lib/python3.13/site-packages/statsmodels/regression/linear_mod

divide by zero encountered in scalar divide

tests/test_analysis_only.py::test_income_value_alias

/Users/crb/thesis/env/lib/python3.13/site-packages/statsmodels/base/model.py:1894: V

covariance of constraints does not have full rank. The number of constraints is 4, b

tests/test_analysis_utils.py::test_save_regression_and_leverage_outputs

tests/test_analysis_utils.py::test_save_regional_report

/Users/crb/thesis/env/lib/python3.13/site-packages/statsmodels/stats/stattools.py:74

omni_normtest is not valid with less than 8 observations; 5 samples were given.

tests/test_consolidate_data.py::test_non_overlapping_rows_preserved

/Users/crb/thesis/consolidate_data.py:45: FutureWarning:

Downcasting object dtype arrays on .fillna, .ffill, .bfill is deprecated and will ch

tests/test_gini_income_analysis.py::test_income_filter_and_quintile_fallback

/Users/crb/thesis/analysis_only.py:199: RuntimeWarning:

invalid value encountered in scalar divide

tests/test_gini_income_analysis.py::test_coefficient_of_variation_zero_mean

/Users/crb/thesis/analysis_only.py:342: RuntimeWarning:

invalid value encountered in scalar divide

-- Docs: https://docs.pytest.org/en/stable/how-to/capture-warnings.html

93 passed, 14 warnings in 10.64s

Appendix B: Pipeline Pseudocode

```
Algorithm RunFullPipeline()
 # 0) Configuration & Logging
 config ← load_default_config(project_root)
 config ← override_from_cli_or_env(config)
 setup_logging(config)

 try
  # 1) Combine & Load Raw Data
  combined_csv ← combine_data(config.data_dir, output/combined_data.csv)
  load_cfg ← {
   data_files: {
    combined_csvs: {
     population: data/11100239.csv,
     telecom_services: data/Landline_Mobile.csv,
     pricing_trends: data/current_trends.csv,
     retail_mobile: data/retail_sector.csv,
     telecom_ov: data/open_telecommunications.csv,
     telus: data/telus.csv,
     bell: data/bell.csv,
     rogers: data/rogers.csv
    },
    income: data/11100239.csv
   },
   metadata_files: {...}
  }
  data_bundle ← load_all_data(load_cfg, strict = true)
  if 'metadata' ∈ data_bundle then
   write aggregate_metadata(data_bundle.metadata) → output/metadata_index.csv

  # 2) Cleaning
  (cleaned, quality) ← clean_all_data(data_bundle)
  if config.use_storage then
   save_to_sqlite(cleaned, config.storage_db)
   cleaned ← load_from_sqlite(config.storage_db)

  # 3) Integration & Enrichment
  integrated ← integrate_datasets(cleaned, config)
  enriched   ← enrich_dataset(integrated, config)

  # 4) Consolidate for Analysis
  (start_year, end_year) ← config.analysis_settings.year_range
  analysis_df ← consolidate_processed_tables(
   enriched,
   output/analysis_dataset.csv,
   start_year,
   end_year
  )

  # 5) Targeted Analyses (domain-specific add-ons)
```

```
run_competition_analysis(analysis_df, output/)          # ARPU ~ HHI + controls
plot_hhi_vs_arpu(analysis_df, output/)
run_substitution_analysis(analysis_df, output/)         # Landline vs Mobile ΔYoY
plot_substitution_trends(analysis_df, output/)
if has_column(analysis_df, 'ad_spend') then
  run_advertising_analysis(analysis_df, output/)
  plot_ad_spend_effect(analysis_df, output/)
run_segmented_digital_divide_analysis(analysis_df, output/)
plot_digital_divide_trends(analysis_df, output/)

# 6) Core Analyses (run_all_analysis)
results ← {}

# 6a) Interaction Regression (econometric model)
# Model (conceptual):
#   Mobile_Penetration ~ median_income + region_type + Year + hhi
#                      + population_density + market_share
#                      + median_income × region_type
results.regression ← run_interaction_regression(analysis_df)

# 6b) Clustering (heterogeneity)
#   - Standardize features; evaluate k ∈ [2..10] (inertia/silhouette/CH/DB)
#   - Choose k; fit KMeans; summarize clusters
results.clusters ← run_cluster_analysis(analysis_df)

# 6c) Propensity Score Matching (Urban vs Rural)
#   - Estimate propensity (logistic)
#   - Nearest-neighbor match; assess balance (SMD, variance ratios)
#   - Estimate ATE; record reliability/overlap diagnostics
(results.psm, results.psm_quality) ← run_propensity_score_matching(analysis_df)

# 6d) Time Series Decomposition (STL)
#   - Stationarity tests (ADF, KPSS); choose transformation if needed
#   - Decompose trend/seasonal/residual; summarize trend direction/magnitude
results.stl ← run_stl_time_series_decomposition(analysis_df)

# 6e) Spatial Autocorrelation
#   - Build provincial adjacency; compute Moran's I (+ LISA as available)
results.spatial ← run_spatial_autocorrelation_analysis(analysis_df)

# 6f) Inequality (Gini/Lorenz)
#   - Compute yearly and provider-specific Gini; produce Lorenz curve data
results.inequality ← run_gini_inequality_analysis(analysis_df)

# 7) Final Cleaning, Validation, Persistence
integrated_final ← final_cleaning(integrated)
final_validation(integrated_final)
save_output(integrated_final, results, config)          # CSV/JSON/PKL write-out

# 8) Visualization Suite
generate_all_plots(integrated_final, output/plots, results.lorenz_curve_data)
```

```
    if results.provider_analysis exists then
      plot_provider_gini(results.provider_analysis, output/plots/provider_gini.html)

    # 9) Tests & Report
    tests_summary ← run_tests("tests/")
    generate_report(integrated_final, results, output/reports, output/plots,
                    tests_output = tests_summary)

    log("Pipeline completed successfully")

  catch e
    log_error(e)
    raise
Algorithm DataPreparationOnly()
  # When generating the master dataset directly from raw sources
  seed_random_generators(42)

  # Clean/standardize provider CSVs (Bell/Rogers/Telus) and external files
  provider_df    ← clean_provider_files(data/*.csv)
  income_df      ← filter_and_categorize_statscan("11100239.csv")
  market_share   ← build_or_fix_retail_sector("retail_sector*.csv")
  telecom_others ← prepare_open_telecom_and_landline()

  # Safe merges on (Province, Year, Provider) with key validation
  merged ← safe_merge(provider_df, income_df,      keys=[Province, Year])
  merged ← safe_merge(merged,      market_share,   keys=[Province, Year, Provider])
  merged ← safe_merge(merged,      telecom_others, keys=[Province, Year])

  # Derived metrics and validation
  merged.hhi               ← calculate_hhi(merged)
  merged.mobile_penetration ← derive_penetration(merged)
  merged.market_share_yoy   ← compute_yoy_change(merged)
  validate_non_empty(merged)
  validate_no_critical_nulls(merged, [Province, Year, Provider])

  write merged → data/enhanced_telecom_data_comprehensive.csv
  write (legacy subset) → data/combined_telecom_data_validated.csv
  log("Data preparation complete")
Algorithm AnalysisOnly(dataset_path, start_year, end_year)
  config ← load_default_config(); setup_logging()
  df ← read_csv(dataset_path)
  df_cons ← consolidate_processed_tables(df, output/analysis_dataset.csv,
                                         start_year, end_year)

  # Execute the same analyses as RunFullPipeline() §6
  results ← run_all_analysis(df_cons, config)

  # Selected domain add-ons and plots
  run_competition_analysis(df_cons, output/)
  plot_hhi_vs_arpu(df_cons, output/)
  run_substitution_analysis(df_cons, output/)
```

```
  plot_substitution_trends(df_cons, output/)
  run_segmented_digital_divide_analysis(df_cons, output/)
  plot_digital_divide_trends(df_cons, output/)

  # Save, test, and report
  final_df ← final_cleaning(df)
  final_validation(final_df)
  save_output(final_df, results, config)
  generate_all_plots(final_df, output/plots, results.lorenz_curve_data)
  tests_summary ← run_tests("tests/")
  generate_report(final_df, results, output/reports, output/plots,
                  tests_output = tests_summary)
```

Notes

The concrete implementation is in thesis/main.py (orchestrator), with modules:

Loading: data_loading.py, combining: combine_data.py, cleaning: cleaning.py

Integration: integration.py, consolidation: consolidate_data.py, enrichment: enrichment.py

Analyses: interaction_regression.py, cluster_analysis.py, propensity_score_matching.py,

time_series_decomposition.py, spatial_autocorrelation.py, Gini in analysis_only.py

Visualization: visualization.py; Reporting: report.py (HTML via templates)

Validation & utilities: validate_data.py, utils.py, storage.py, metadata.py, tests under thesis/tests/

Core outputs reside under thesis/output/ (datasets, JSON/PKL results, plots/, reports/).