MIS581 Capstone Project: The Research Questions

Nicholas Laeder

Colorado State University - Global Campus

MIS581 - Data Analytics Capstone

Steve Chung

6/8//2025

Capstone Project: Research Questions, Hypotheses, and Research Methods

Introduction and Purpose of Research

My Capstone Project embarks on developing a lightweight data warehousing proof of concept within the Google Cloud and Google AI environment. The fundamental purpose of this research is to empower small business owners, specifically those in my peer group, who currently lack the analytical capabilities to fully leverage their email marketing data from platforms like Constant Contact and Mailchimp. By integrating this data into a structured data warehouse, we aim to transform raw information into actionable insights. A novel aspect of this project involves delivering these insights through an AI-generated podcast, making complex data accessible to a non-technical audience. This project seeks to address a critical gap in small business operations where valuable marketing data is collected but often remains underutilized, hindering strategic decision-making and campaign optimization.

Research Question

To address the overarching challenge of optimizing email marketing for small businesses, regardless of their specific platform, I've formulated a unified primary research question:

1. For small businesses utilizing email marketing platforms (such as Constant Contact and Mailchimp), what are the key characteristics of high-performing email campaigns, and how do identifiable subscriber engagement patterns influence desired business outcomes (e.g., conversions, retention, new sign-ups)?

This question aims to draw generalizable insights from diverse data sources, acknowledging that while the underlying data comes from different platforms, the fundamental analytical goals related to campaign effectiveness and subscriber behavior are consistent across small businesses.

Overview of Hypotheses

In quantitative research, a hypothesis is a testable statement that proposes a relationship between two or more variables. It is an educated guess or a provisional explanation that can be supported or refuted through empirical data analysis. Researchers typically formulate a null hypothesis (H0​), which represents a statement of no effect or no relationship, and an alternative hypothesis (Ha​ or H1​), which posits the existence of an effect or relationship. The goal of the research is often to gather sufficient evidence to reject the null hypothesis in favor of the alternative.

Hypotheses for Capstone Project

Based on the unified research question, the following hypotheses will guide the data analysis:

* Hypothesis Set 1: Campaign Characteristics and Effectiveness
  + Null Hypothesis (H01​​): There is no statistically significant relationship between specific email campaign characteristics (e.g., subject line elements, call-to-action design, content themes) and campaign effectiveness (e.g., open rates, click-through rates, conversion rates) for small businesses utilizing email marketing platforms.
  + Alternative Hypothesis (Ha1​​): Specific email campaign characteristics (e.g., subject line elements, call-to-action design, content themes) have a statistically significant relationship with campaign effectiveness (e.g., open rates, click-through rates, conversion rates) for small businesses utilizing email marketing platforms.
* Hypothesis Set 2: Subscriber Engagement and Business Outcomes
  + Null Hypothesis (H02​​): There is no statistically significant correlation between subscriber engagement patterns (e.g., open frequency, click behavior, churn indicators) and desired business outcomes (e.g., enrollment rates, new lesson sign-ups, customer retention) for small businesses utilizing email marketing platforms.
  + Alternative Hypothesis (Ha2​​): Specific subscriber engagement patterns (e.g., open frequency, click behavior, churn indicators) are statistically correlated with desired business outcomes (e.g., enrollment rates, new lesson sign-ups, customer retention) for small businesses utilizing email marketing platforms.

Research Method: Models and Tools for Hypothesis Testing

The research method employs an ELT (Extract, Load, Transform) methodology within the Google Cloud ecosystem, followed by advanced analytical techniques.

1. Data Acquisition and Preparation (Extract & Load): Python scripts, deployed as Google Cloud Functions, will be used to extract raw, granular data from the Constant Contact and Mailchimp APIs. This raw data will be loaded directly into separate, dedicated tables for each peer and platform within the raw\_data dataset in Google BigQuery. This ensures data purity, strict privacy isolation, and manages the inherent structural differences between the source systems. Google Cloud Secret Manager will securely store all API credentials.
2. Data Transformation and Querying (Transform): Transformations will occur within BigQuery using SQL. This involves cleaning, standardizing data types, handling missing values, and joining related entities (e.g., contacts with campaigns). A critical step here will be the creation of processed\_data tables where data from both sources, once transformed to a common conceptual schema, can potentially be combined or analyzed in parallel, still maintaining tenant\_id for distinct insights. This structured data will then be used for direct querying to prepare for hypothesis testing.
3. Hypothesis Testing and Analysis (Models and Tools):
   * For Hypothesis Set 1 (Campaign Characteristics): This analysis will involve identifying various campaign attributes (e.g., length of subject lines, presence of certain keywords, use of promotions). Statistical techniques such as regression analysis (if applicable to continuous outcomes like conversion rates) or comparative analysis (e.g., t-tests or ANOVA if comparing categorical characteristics) will be employed on processed data from both peers to determine the statistical significance of these characteristics on campaign effectiveness metrics (e.g., open rate, click-through rate). The LLM (Vertex AI) will also be used to qualitatively analyze content themes and suggest hidden commonalities in high-performing campaigns.
   * For Hypothesis Set 2 (Subscriber Engagement): This will involve analyzing metrics like open frequency, click density, and unsubscribe rates, and correlating them with the respective desired business outcomes (e.g., enrollment for Peer 1, new lesson sign-ups for Peer 2). Statistical correlation methods (e.g., Pearson's r, Spearman's ρ) will be calculated using BigQuery SQL or Python libraries on the processed data. Time-series analysis may be used to observe trends and identify temporal relationships.
   * Primary Tools: Google BigQuery (for SQL transformations and querying), Python (for scripting, API interaction, statistical analysis, and data preparation for LLM), and Google Cloud Vertex AI (for LLM-powered insight generation, summarization, and potentially qualitative analysis of text data).

Ethical Considerations

The chosen dataset, comprising email marketing data, necessitates careful attention to ethical considerations, particularly regarding data privacy and security.

* Data Privacy: The data contains Personally Identifiable Information (PII) such as email addresses. While the project focuses on aggregated insights and patterns, strict measures will be taken to protect individual privacy. Any display or sharing of data, especially within the podcast, will be at an aggregated, anonymized, or pseudonymized level, never revealing individual subscriber details. Direct access to raw PII will be limited to the project developer.
* Data Security: All data will be stored and processed within Google Cloud's secure environment. API credentials will be managed exclusively through Google Cloud Secret Manager. Access to BigQuery datasets will be strictly controlled via Google IAM, adhering to the principle of least privilege. The use of separate raw tables for each peer significantly enhances physical data segregation and minimizes the risk of accidental cross-tenant data exposure.
* Consent: It is assumed that the peer organizations have obtained appropriate consent from their subscribers for the collection and analysis of their email engagement data, in accordance with relevant privacy regulations (e.g., GDPR, CCPA). The project's scope is strictly analytical and non-marketing in nature.
* Bias and Interpretation: The LLM's analysis and podcast generation could potentially introduce biases present in its training data or in the prompts provided. Care will be taken in prompt engineering to minimize bias and ensure that interpretations are data-driven and objective. The insights provided will be clearly framed as statistical observations and potential correlations, not definitive causation, to avoid misinterpretation by the business owners.
* Responsible Use of Insights: The project aims to provide actionable insights for business growth. The use of these insights will be discussed with the peer organizations to ensure they are applied ethically and do not lead to manipulative or intrusive marketing practices.