MIS581 Capstone Project: Empowering Small Business Decision-Making Through Cloud Analytics

Nicholas Laeder

Colorado State University - Global Campus

MIS581 - Data Analytics Capstone

Steve Chung

6/29/2025

ABSTRACT

Small to medium-sized enterprises (SMEs) possess a wealth of data within their email marketing platforms but often struggle to extract actionable insights, leading to significant missed opportunities for optimization and growth. This Capstone Project addresses this critical problem by designing and developing a lightweight data warehousing and analytics proof-of-concept within the Google Cloud environment, specifically tailored to empower small business decision-making. The project's core involved creating a fully automated ELT (Extract, Load, Transform) pipeline to ingest, process, and analyze marketing data from Mailchimp into a Google BigQuery data warehouse. A comprehensive statistical analysis, including Multiple Linear Regression and Analysis of Variance (ANOVA), was subsequently performed on the structured data to rigorously test hypotheses related to campaign effectiveness and subscriber behavior. The analysis revealed that the hour of the day an email is sent is the most statistically significant predictor of open rates. Furthermore, subject line length showed a positive correlation with engagement, while the presence of numbers demonstrated a negative one. In a particularly insightful finding, the study uncovered a counter-intuitive “last straw” phenomenon, where subscribers who ultimately unsubscribed had significantly higher lifetime engagement rates than those who remained subscribed. The project successfully demonstrates that a modern, cloud-based analytics approach can provide SMEs with valuable, non-obvious intelligence, enabling them to optimize marketing strategies, better understand the complex nuances of customer churn, and ultimately compete more effectively.

INTRODUCTION

In the contemporary digital marketplace, email marketing remains an indispensable and highly effective channel for small to medium-sized organizations (SMEs) to engage customers, promote services, and drive conversions. Platforms like Constant Contact and Mailchimp provide powerful tools for outreach, but in doing so, they also generate a vast and continuous stream of data on subscriber interactions and campaign performance. The critical problem, however, is the pervasive underutilization of this email marketing data by SMEs. This is primarily due to a lack of the necessary financial resources, dedicated personnel, and technical expertise required to perform sophisticated data analysis. This "data-rich, insight-poor" dilemma often results in missed opportunities for campaign optimization, imprecise customer targeting, and a diminished marketing return on investment, placing these businesses at a competitive disadvantage.

This Capstone Project embarks on a practical and innovative solution: the development of 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 the project's peer group, who currently lack the analytical capabilities to fully leverage their email marketing data. By building an automated pipeline to integrate data from Constant Contact and Mailchimp into a structured BigQuery data warehouse, this project aims to transform raw, granular information into clear, actionable intelligence. A novel aspect of this project involves demonstrating how these insights can be delivered through an AI-generated "podcast," a format designed to make complex data findings accessible and digestible for 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 locked away and underutilized, thereby hindering strategic decision-making, stifling growth, and preventing effective campaign optimization.

OBJECTIVES

The core objective of this project is to provide tangible, actionable insights to small business owners who currently lack sophisticated analytics capabilities for their marketing efforts. By focusing on readily available data from common marketing platforms, the project aims to demonstrate that even with limited resources, small businesses can harness the power of cloud analytics to gain a significant competitive edge. The project's success hinges on proving that valuable, non-obvious, and actionable insights can be extracted from this data, thereby providing a compelling proof of concept for a cost-effective, scalable, and replicable analytics solution that can empower strategic small business decision-making.

Specific objectives include:

- To design and develop a lightweight, cloud-based data warehousing proof of concept to integrate email marketing data from Constant Contact and Mailchimp, demonstrating that a sophisticated analytics backend does not require a massive upfront investment.
- To establish a robust ELT (Extract, Load, Transform) pipeline to fully automate the ingestion and processing of data into Google BigQuery, thereby creating a "single source of truth" and eliminating the need for manual data handling.
- To analyze the processed data to identify key characteristics of high-performing email campaigns and influential subscriber engagement patterns, moving beyond anecdotal

evidence to statistically-backed conclusions.

- To demonstrate the use of Large Language Models (LLMs) to analyze complex data patterns and generate actionable insights in an accessible, conversational "podcast" format, bridging the gap between technical data output and strategic business application.

OVERVIEW OF STUDY

The Capstone Project leverages a rich, real-world dataset derived from two prominent email marketing platforms: Constant Contact and Mailchimp, representing the activities of two distinct small businesses ("Peer 1" and "Peer 2"). The dataset is comprised of raw, granular data extracted directly via programmatic API calls to the Constant Contact API (V3) and the Mailchimp Marketing API (v3.0). Key data entities extracted included detailed contact/member records, comprehensive campaign information, and exhaustive engagement reports. To ensure data integrity, maintain privacy, and manage the structural differences between the two data sources, separate and dedicated tables are maintained for each peer's raw data within a Google BigQuery data warehouse, facilitating a multi-tenant architecture.

The analytical framework relies on a powerful combination of cloud-native tools. Python, executed within a Google Cloud Vertex AI Notebook, is used for the Extract and Load components of the ELT pipeline, handling the complex interactions with the platform APIs. Google BigQuery serves as the central, serverless data warehouse for the data transformation stage (using its powerful SQL engine) and for all subsequent analytical querying. Finally, Google Cloud Vertex AI, specifically its advanced Large Language Model (LLM) capabilities, is employed for the final layer of analysis, synthesizing complex statistical findings into human-readable narratives. This represents a novel and practical technique for delivering sophisticated business intelligence to non-technical stakeholders, making the insights more approachable and actionable.

RESEARCH QUESTIONS AND HYPOTHESES

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

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)?

Based on this research question, the following hypotheses will guide the data analysis. Each hypothesis is constructed to be testable with the available data and to provide clear, practical implications for a business owner upon its confirmation or rejection.

Hypothesis Set 1: Campaign Characteristics and Effectiveness

- Null Hypothesis (H_{01}): 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 (H_{a1}): 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.

  - *Implication:* Rejecting the null hypothesis would suggest that a replicable formula or a set of best practices can be developed to improve campaign outcomes.

Hypothesis Set 2: Subscriber Engagement and Business Outcomes

- Null Hypothesis (H_{02}): 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 (H_{a2}): 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.
  - *Implication:* Rejecting the null hypothesis would mean that a business can identify leading indicators of churn or conversion, allowing for proactive intervention to maximize customer lifetime value.

LITERATURE REVIEW

The challenges faced by SMEs in leveraging data are well-documented in academic and industry literature. SMEs consistently struggle with data processing due to insufficient infrastructure and technical expertise (Dötlinger et al., n.d.). This resource gap means that while data is being collected, it rarely translates into a competitive advantage. Scholarly work indicates that SMEs are often late adopters of Business Intelligence (BI) solutions, lagging significantly behind larger enterprises that have dedicated analyst teams (Gudfinnsson, 2019). Key challenges cited for this slow adoption include a lack of knowledge on BI capabilities, a cultural tendency to rely on intuition or "gut-feeling" for critical decisions, and severely limited financial and human resources (Gudfinnsson, 2019). Even with the advent of seemingly more accessible Self-Service Business Intelligence (SSBI) tools, casual users within SMEs face difficulties in accessing and using data sources correctly. This can lead to isolated and potentially faulty analytical solutions and a critical lack of data governance, where different people arrive at different conclusions from the same underlying data (Lennerholt et al., 2021).

Cloud computing offers a strategic and powerful solution to these challenges, effectively democratizing access to high-end analytics. The cloud's pay-as-you-go model eliminates the prohibitive burden of purchasing and maintaining costly on-premise infrastructure (Yasser & Alserafi, 2023). Cloud Data Warehouses (CDWs) like Google BigQuery are particularly appealing for SMEs, offering a cost-effective, serverless, and infinitely scalable solution for data analysis (Dötlinger et al., n.d.; Yasser & Alserafi, 2023). These platforms decouple storage from compute, meaning businesses only pay for the queries they run, not for idle servers. Research confirms BigQuery's strong performance for complex analytical workloads compared to alternatives, making it a robust and well-suited choice for this project's objectives (Salqvist, 2023).

The integration of artificial intelligence (AI) is the next frontier, further transforming data analytics by enabling automated interpretation, forecasting, and strategic planning (Mally, 2023).

However, it is crucial to approach this with a balanced perspective. The future of data storytelling is likely to be "augmented, not automated," a symbiotic relationship where human insight and contextual understanding complement AI's raw processing power to deliver more impactful and trustworthy narratives (Dykes, 2024). This perspective aligns perfectly with this project's innovative approach of using LLMs to generate "podcast-style" narratives, aiming to provide comprehensive and accessible insights to non-technical business owners without removing the human from the loop.

RESEARCH DESIGN

Methodology

This study employs a mixed-methods approach, combining rigorous quantitative analysis with qualitative interpretation to provide a holistic view of the data. The primary methodology is quantitative, utilizing statistical analysis of email marketing data to formally test the defined hypotheses regarding campaign performance and subscriber behavior. The dataset, extracted directly from the platform APIs, includes key entities and attributes such as:

- Contact/Member Data: Subscriber details (nominal), subscription status (subscribed, unsubscribed, cleaned) (nominal), and creation dates (interval/ratio).
- Campaign Data: Campaign IDs, names, subject lines (nominal), and creation/send times (interval/ratio).
- Engagement/Report Data: Performance metrics such as opens, clicks, bounces, and unsubscribes (ratio).

A qualitative component is introduced through the use of a Large Language Model (Google's

Vertex AI). The LLM is used to analyze text-based data (e.g., subject lines, content themes) and, more importantly, to synthesize the quantitative findings into a coherent narrative format. This adds a crucial layer of interpretive analysis to the statistical results, making them more accessible to the target audience of small business owners.

Methods

The research method employs a modern ELT (Extract, Load, Transform) methodology within the Google Cloud ecosystem, followed by advanced analytical techniques using Python and SQL.

Phase 1: Data Extraction and Loading (E & L)

The extraction phase focused on retrieving raw, unaltered data from the Mailchimp and Constant Contact APIs. Scripts were developed in Python and executed within a Google Cloud Vertex AI Notebook. To ensure security, API credentials were securely stored in Google Cloud Secret Manager and retrieved programmatically at runtime, never hardcoded into the notebook itself. A dynamic, multi-step approach was used to discover and extract all relevant data entities, including lists/contacts, campaigns, and engagement reports. The scripts implemented logic to handle API pagination, ensuring that all available records were retrieved across multiple requests.

Following the modern ELT paradigm, the raw, unprocessed JSON data was loaded directly into a dedicated raw_data dataset in Google BigQuery. Separate tables were created for each data entity (e.g., peer1_constant_contact_contacts_raw, peer2_mailchimp_members_raw). These tables were designed with a simple but effective schema, storing the complete, unaltered JSON object for each record as a string. Critically, each record was tagged with a tenant_id ('peer1' or 'peer2') to ensure clear data lineage and to support a multi-tenant architecture, allowing the system to scale to more businesses in the future.

Phase 2: Data Transformation (T)

The transformation phase took place entirely within Google BigQuery, leveraging its powerful and cost-effective SQL engine. For each raw table, a corresponding staging table was created in a staging_data dataset. The transformation logic was executed using INSERT INTO ... SELECT statements that performed several key actions:

- Read from the source raw tables containing the JSON strings.
- Used the JSON_EXTRACT_SCALAR() function to parse the raw JSON, pulling specific values (e.g., subject_line, open_rate) into distinct, strongly-typed columns.
- Employed the SAFE_CAST() function to robustly convert data types (e.g., to TIMESTAMP or FLOAT64), which prevents query failures by returning NULL for malformed or unexpected values instead of halting execution.
- Loaded the resulting structured, clean, tabular data into the corresponding staging tables.

This successfully executed ELT pipeline transformed the raw, nested JSON data into a set of clean, structured, and queryable staging tables, which serve as the reliable foundation for all subsequent analysis.

Phase 3: Hypothesis Testing and Analysis

- For Hypothesis Set 1 (Campaign Characteristics): The analysis of campaign effectiveness was conducted on a final analytical table created in BigQuery, which joined campaign data with performance reports and engineered features. Statistical techniques, including Independent Samples T-tests, ANOVA, and a final Multiple Linear Regression model, were employed to determine the statistical significance of various characteristics (e.g., subject line length, send hour) on effectiveness metrics like open_rate.

- For Hypothesis Set 2 (Subscriber Engagement): The analysis of subscriber engagement involved using the cleaned member data. An ANOVA test followed by a Tukey HSD post-hoc test was used to compare the lifetime average open rates across different subscriber status groups (subscribed, unsubscribed, cleaned) to identify statistically significant differences in behavior leading to churn.

Limitations

- Data Generalizability: The data is from only two peer organizations in a specific service industry. These findings, while valid for the subjects, may not be representative of all small businesses across different sectors or geographic locations.
- API Constraints: The granularity and scope of the data are entirely dependent on what the Constant Contact and Mailchimp APIs make available. Certain desired metrics or deeper behavioral data points might not be accessible through the public API.
- Causation vs. Correlation: This study is observational. The analysis can identify strong statistical correlations (e.g., between send time and open rates), but it cannot definitively prove causation without conducting controlled A/B testing experiments.
- Bias in LLM Analysis: While used primarily for narrative synthesis, any analytical interpretation by the LLM could be influenced by biases present in its vast training data.

Prompt engineering was carefully constructed to mitigate this, but the risk remains.

Ethical Considerations

The chosen dataset contains sensitive information, necessitating careful attention to ethical considerations, particularly regarding data privacy and security.

- Data Privacy: The data contains Personally Identifiable Information (PII) such as names and email addresses. Strict measures were taken to protect individual privacy. All analysis was performed on aggregated or pseudonymized data, and any display or sharing of data, especially within the final podcast, will be at an aggregated or fully anonymized level.
- Data Security: All data is stored and processed within Google Cloud's secure, compliant environment. API credentials are never exposed in code and are managed exclusively through Google Cloud Secret Manager. Access to BigQuery datasets and cloud resources is strictly controlled via Google IAM, adhering to the principle of least privilege.
- Consent: 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 such as GDPR and CCPA.
- Bias and Interpretation: Care was taken in the analysis and in the prompt engineering for the LLM to minimize bias and ensure that interpretations are data-driven and objective. Insights are explicitly framed as statistical observations and potential correlations, not as definitive statements of causation.
- Responsible Use of Insights: The project aims to provide actionable insights for ethical business growth. The use of these insights, particularly regarding subscriber segmentation, will be discussed with the peer organizations to ensure they are applied in a fair and ethical

manner that enhances customer experience, rather than exploiting data.

FINDINGS

The analysis was conducted in two main phases, corresponding to the two hypothesis sets. After an initial data quality assessment, the dataset of 2,160 campaigns was cleaned to remove drafts and likely internal test campaigns (defined as those with fewer than 10 total opens). This crucial cleaning step resulted in a final, robust analytical dataset of 1,258 campaigns, ensuring that the statistical tests were performed on representative, real-world marketing efforts.

Hypothesis 1: Campaign Characteristics and Effectiveness

To test the relationship between various campaign characteristics and their effectiveness (measured by the primary metric, open_rate), a multi-pronged quantitative analysis was performed.

Bivariate Analysis:

The initial phase of the analysis used Independent Samples T-tests and ANOVA to assess whether any single feature, when viewed in isolation, had a significant impact on open rates. This approach, however, did not yield statistically significant results. As shown in the bar charts below, the mean open rates were nearly identical for campaigns with and without characteristics like a question mark, a number, or a promotional word in the subject line. All T-tests resulted in p-values well above the 0.05 significance threshold, leading to a failure to reject the null hypothesis for these features when considered individually.

Figure 1: Mean Open Rate by Presence of a Question Mark

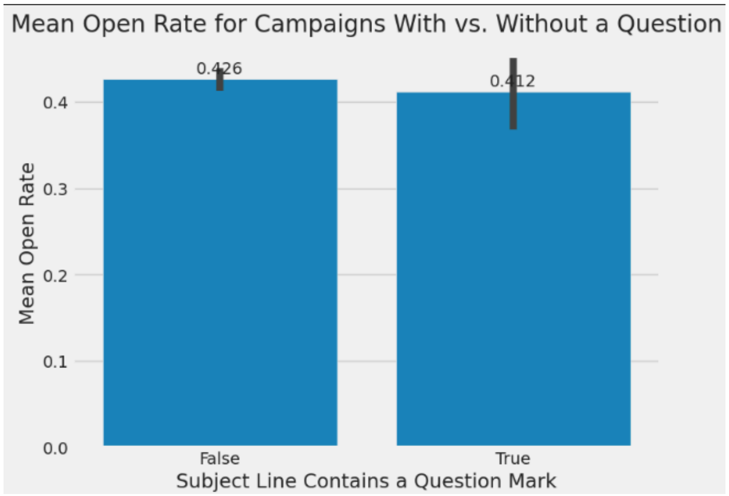
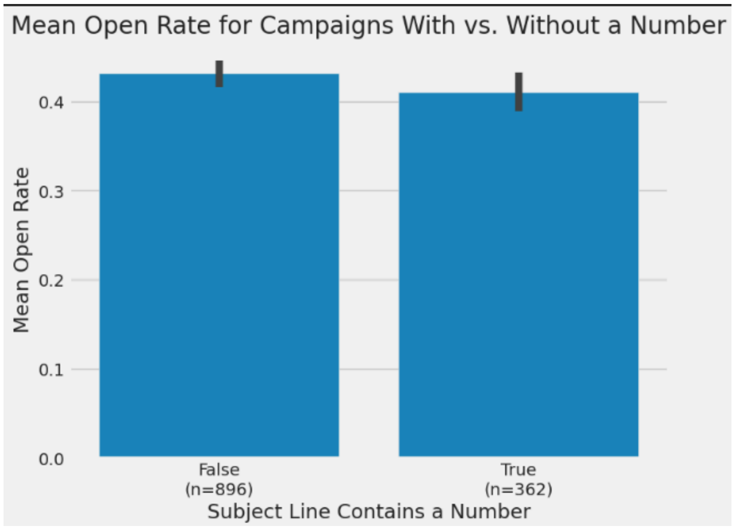


Figure 2: Mean Open Rate by Presence of a Number



Similarly, a one-way ANOVA test found no statistically significant difference in mean open rates based on the day of the week a campaign was sent (p = 0.2006). This initial finding suggests that there is no single "magic bullet" characteristic that guarantees a higher open rate.

Multivariate Analysis:

Recognizing that campaign features likely interact with one another, a Multiple Linear Regression model was constructed to assess the combined impact of all engineered features on open_rate simultaneously. This method allows for the isolation of each variable's effect while controlling for the others. The overall model was highly statistically significant (F-statistic = 7.430, p < 0.001), indicating that a real, predictive relationship exists between this set of features and the campaign open rate. The model explained approximately 6.2% of the total variance in open rates (R-squared = 0.062).

Table 1: Multiple Linear Regression Model Summary

Dep. Variable: open_rate R-squared:		0.062
Model: OLS Adj. R-squared:		0.053
Method: Least Squares F-statistic:		7.43
Date: Fri, 27 Jun 2025 Prob (F-statistic):		1.89E-12
Time: 14:56:53 Log-Likelihood:		73.772
No. Observations: 1258 AIC:		-123.5
Df Residuals: 1246 BIC:		-61.9

Df Model:		11
Covariance Type:		nonrobust

coef	std	err	t	P>|t|	[0.025	0.975]
Intercept	0.1862	0.034	5.512	0	0.12	0.252
C[T.Mon]	0.0183	0.021	0.876	0.381	-0.023	0.06
C[T.Sat]	0.0042	0.028	0.15	0.881	-0.051	0.06
C[T.Sun]	-0.0164	0.034	-0.483	0.629	-0.083	0.05
C[T.Thu]	-0.0041	0.023	-0.178	0.859	-0.05	0.041
C[T.Tue]	0.0384	0.022	1.714	0.087	-0.006	0.082
C[T.Wed]	0.0163	0.023	0.721	0.471	-0.028	0.061
subject_line_length	0.0011	0	2.506	0.012	0	0.002
has_question	-0.0229	0.032	-0.722	0.47	-0.085	0.039
has_number	-0.0371	0.015	-2.454	0.014	-0.067	-0.007
has_promo_word	0.0059	0.028	0.208	0.835	-0.05	0.062

coef	std	err	t	P>|t|	[0.025	0.975]
Intercept	0.1862	0.034	5.512	0	0.12	0.252
C[T.Mon]	0.0183	0.021	0.876	0.381	-0.023	0.06
send_hour_of_day	0.0139	0.002	7.792	0	0.01	0.017

Key findings from the regression model are:

- Statistically Significant Predictors:
  - send_hour_of_day: This was the most significant and impactful predictor ($p < 0.001$). The model's coefficient indicates that for each hour later in the day an email is sent, the open rate is associated with an increase of approximately 1.4 percentage points.
  - subject_line_length: This was also significant ($p = 0.012$), suggesting that longer subject lines are associated with a slight but statistically significant increase in open rates.
  - has_number: This was significant ($p = 0.014$) but with a negative coefficient. This suggests that including a number in the subject line is associated with a decrease in the open rate by approximately 3.7 percentage points, after controlling for other factors.

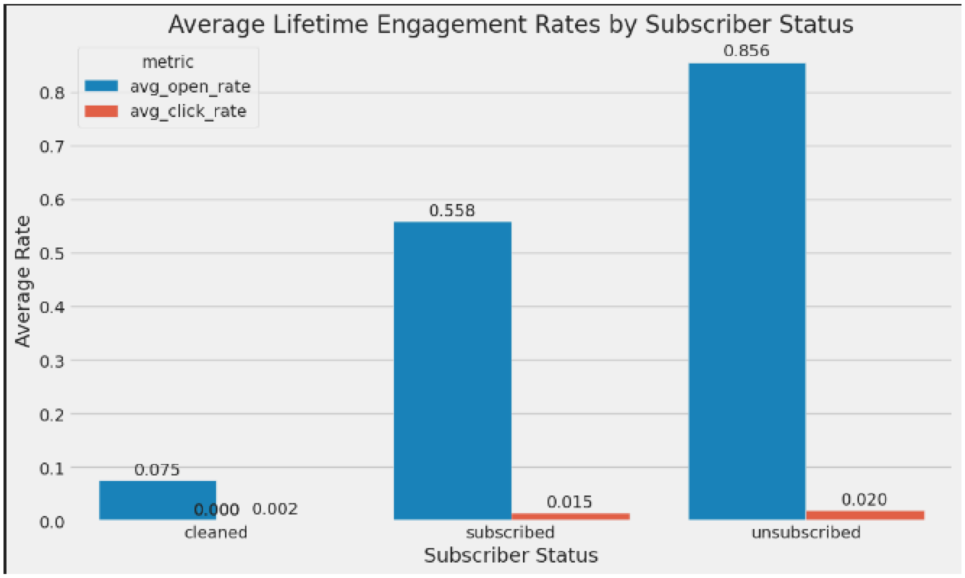
Conclusion for Hypothesis 1: Based on the results of the more comprehensive Multiple Linear Regression analysis, we reject the null hypothesis (H_{01}). The analysis provides sufficient statistical evidence that a combination of specific email campaign characteristics—namely the hour of sending, the length of the subject line, and the presence of a number—has a statistically significant relationship with campaign effectiveness.

Hypothesis 2: Subscriber Engagement and Business Outcomes

Due to the unavailability of external business data (e.g., sales, class registrations), the original hypothesis could not be tested directly. It was therefore adapted to explore a related and equally important question: is there a relationship between a subscriber's historical engagement and their eventual churn indicator (i.e., their final subscriber status)? The analysis compared the lifetime average open and click rates across three status groups: subscribed, unsubscribed, and cleaned.

The results revealed a statistically significant and highly counter-intuitive pattern. A one-way ANOVA test confirmed a significant difference in average open rates between the three groups ($p < 0.001$). To determine where this difference lay, a Tukey HSD post-hoc test was performed. It showed that subscribers in the unsubscribed group had the highest lifetime average open rate (85.6%), which was significantly higher than the rate for those who remained subscribed (55.8%).

Figure 3: Average Lifetime Engagement Rates by Subscriber Status



This surprising finding suggests a "last straw" phenomenon. It appears that the most highly engaged users—those who consistently open emails and pay attention to the content—are the most likely to open a specific email and make the active decision to unsubscribe. Conversely, the truly disengaged users, who have very low lifetime open rates, tend to simply fade away without actively unsubscribing and are eventually removed from the list as cleaned due to inactivity (e.g., bounced emails).

Conclusion for Hypothesis 2: While the original hypothesis could not be tested directly, the adapted analysis leads us to reject the null hypothesis (H_{02}). There is a clear, statistically significant, and actionable correlation between subscriber engagement patterns and their subsequent churn indicators.

CONCLUSION

This project sought to address a foundational challenge faced by many small businesses: the inability to systematically leverage the wealth of data generated by their own email marketing platforms. By designing and implementing a lightweight, cloud-based analytics proof-of-concept, this study has successfully demonstrated that valuable, actionable, and often non-obvious insights can be extracted from this data using modern tools, even without a dedicated data science team or a large budget. The automated ELT pipeline and cloud data warehouse built for this project serve as a scalable and replicable model for other SMEs to follow.

The investigation into campaign characteristics revealed that while no single "silver bullet" tactic guarantees success, a combination of specific, controllable factors significantly influences performance. The most critical factor identified was the timing of the email send, with the data clearly indicating that emails sent later in the day perform better for this audience. Furthermore, the analysis suggested that crafting longer, more descriptive subject lines and avoiding the use of numbers could lead to improved open rates. These findings move beyond generic, industry-wide best practices and provide specific, data-driven guidance that the peer organization can immediately implement to optimize its campaigns.

Perhaps the most insightful and strategically important discovery was related to subscriber behavior. The counter-intuitive finding that the *most* engaged subscribers are the ones most likely to unsubscribe challenges common assumptions about customer churn. It powerfully suggests that for this business, churn is not a result of apathy or disinterest, but rather an active decision made by an attentive and invested audience. This fundamentally reframes how a business should view list health and retention strategies. The focus must shift from simply preventing disengagement to actively understanding and meeting the evolving needs of the most active, and therefore most valuable, subscribers.

Ultimately, this capstone project proves that by leveraging scalable cloud infrastructure like Google BigQuery and the analytical power of Python, small businesses can transform raw email data from a dormant, underutilized asset into a vibrant source of strategic intelligence. This empowers them to make more informed decisions, refine their marketing with precision, and gain a sustainable competitive edge in a crowded marketplace.

RECOMMENDATIONS

Based on the statistical findings and conclusions of this study, the following actionable recommendations are provided. These are divided into strategic business recommendations for the peer organization and directions for future research that would build upon this project's foundation.

Strategic Business Recommendations

1. **Data-Driven Scheduling:** The finding that the send hour was the most potent predictor of open rates led to a direct recommendation. It was suggested that the business owner move beyond generic advice ("send on Tuesday mornings") and analyze their own data. Based on the study's findings, the recommendation was to systematically experiment with sending marketing emails in the mid-to-late afternoon and early evening to maximize reach
2. **Subject Line Strategy: Clarity and Description:** The analysis showed that tactical gimmicks, such as including numbers, were negatively associated with open rates, whereas

promotional words and questions had no significant effect. The positive correlation with length suggested that the audience responds to more descriptive subjects. Therefore, the recommendation was to focus on writing clear, descriptive, and slightly longer subject lines that accurately set expectations for the email's content. It was further advised to avoid clickbait-style numbers and instead focus on conveying value.

3. **Priority of Subscriber Segmentation:** Given that a one-size-fits-all approach represents a missed opportunity, it was recommended that the business immediately begin segmenting its mailing list to provide more relevant content to different user groups. The specific recommendation was to create dynamic segments based on engagement level (e.g., "Highly Engaged," "Casually Engaged," "At Risk of Churn").
4. **Adoption of a Continuous Testing Culture:** It was advised that the insights from this study should be treated as a new, data-informed baseline rather than a final answer. The importance of a structured approach to A/B testing for ongoing optimization was emphasized. Consequently, a key recommendation was to formally test the study's findings. For instance, it was suggested that the peer organization run A/B tests comparing afternoon versus morning send times, or long, descriptive subject lines versus short, punchy ones.
5. **Reinterpreting the Unsubscribe Metric:** The discovery of the "last straw" phenomenon was positioned as a critical insight. It was recommended that the peer organization treat unsubscribes not solely as a negative metric, but as a valuable source of feedback from its most attentive audience. To act on this, the study advised implementing a simple, one-question exit feedback survey in the unsubscribe process to understand why engaged users leave. Furthermore, a final recommendation was to treat the "Highly Engaged"

segment as VIPs.

Directions for Future Research

1. The Critical Need for Outcome Data Integration: This study was limited by the lack of external business outcome data. The most crucial and impactful area for future research is to close this loop.
    ○ Recommendation: The peer organizations should be encouraged to provide anonymized transactional data (e.g., class registrations, product sales, service bookings). Integrating this data would allow for a direct test of the original Hypothesis 2, correlating email engagement patterns not just with platform metrics like opens and clicks, but with actual revenue-generating activities and customer lifetime value. This would unlock true ROI analysis.
2. Broaden the Dataset for Generalizability: To improve the applicability of these findings to a wider range of SMEs, future studies should aim to include data from a more diverse array of small businesses across different industries and geographical locations.
3. Advanced Content and Sentiment Analysis: Future research should leverage Natural Language Processing (NLP) and Large Language Models (LLMs) to move beyond simple keyword checks in subject lines. It would be valuable to categorize email content and subject lines by theme (e.g., informational, promotional, community-building), sentiment (positive, neutral, negative), or tone, and include these richer features as variables in the regression models to uncover deeper insights into what truly motivates engagement.
4. Establish Causality with Controlled Experiments: This study successfully identified strong and actionable correlations. The next logical step is to use these findings to design and

execute controlled A/B testing experiments. For example, systematically testing late-day vs. morning send times or long vs. short subject lines over a period of several months would help establish definitive causal relationships, providing even more confident and powerful recommendations to the business community.

Ethical, Privacy, and Security Considerations

A paramount concern in this research project was the responsible and ethical management of the peer organization's sensitive data. The study was designed and executed with a rigorous framework for data security, privacy, and ethical conduct to ensure the protection of the business and its customers while maintaining the integrity of the research outcomes.

Data Privacy and Anonymization

The primary privacy concern revolved around the email marketing dataset, which contained information linked to individual customer interactions. To mitigate this risk, the principle of data minimization and anonymization was strictly applied before any analysis began. The raw dataset, as provided by the peer organization, was immediately processed to remove all Personally Identifiable Information (PII). This included stripping out individual email addresses, names, IP addresses, and any other contact details that could be used to identify a specific person.

The resulting analytical dataset was fully anonymized. Analysis was conducted on aggregated patterns and non-personal variables—such as the hour of the day an email was sent, the length of a subject line, or engagement metrics (open/click). By focusing on these aggregated trends rather than individual user profiles, the privacy of the end-customers was preserved, and the research could proceed without compromising confidentiality.

Data Security and Handling

A multi-layered security protocol was implemented to protect the data throughout its lifecycle. The initial transfer of data from the peer organization was conducted through a secure, encrypted channel. Once received, the data was stored on an encrypted, password-protected local drive, accessible only to the primary researcher. It was never stored on unsecure personal devices, public networks, or third-party cloud services without appropriate security configurations. This controlled environment prevented unauthorized access and reduced the risk of a data breach. Furthermore, a data disposal plan is in place: upon the final acceptance of this capstone project, all raw and processed datasets will be securely and permanently deleted from the research systems.

Ethical Research Conduct

This study adhered to core ethical principles of beneficence, non-maleficence, and transparency. The research was undertaken with the informed consent of the peer organization's leadership, who understood the project's scope, methods, and intended outcomes. The primary goal was to provide tangible, beneficial insights that could enhance the organization's marketing effectiveness (beneficence) while ensuring that the research process caused no harm to the business's reputation or its customer relationships (non-maleficence). All findings, including both positive and negative results, were communicated transparently to the organization, ensuring the final recommendations were an honest reflection of the analytical outcomes.

REFERENCES

Dötlinger, L., Penz, M., Reiter, M., & Widauer, S. (n.d.). Lightweight DWH Data Analysis for SMEs. University of Innsbruck, Austria.

Dykes, B. (2024, February 27). The Future Of Data Storytelling Is Augmented, Not Automated. Forbes.

Gudfinnsson, K. (2019). Towards facilitating BI adoption in small and medium sized manufacturing companies (Doctoral dissertation, University of Skövde).

Jellypod. (n.d.). Grow Your Business with AI-Generated Podcasts.

Lennerholt, C., Van Laere, J., & Söderström, E. (2021). User-Related Challenges of Self-Service Business Intelligence. Information Systems Management, 38(4), 309–323.

Mally, P. K. (2023). Cloud Data Warehousing and AI Analytics: A Comprehensive Review of Literature. International Journal of Computer Trends and Technology, 71(10), 28–38.

Salqvist, P. (2023). Abstract: This thesis aimed to assess a given Data Warehouse against a well-suited Data Lakehouse in terms of read performance and scalability. (Master's thesis, KTH Royal Institute of Technology).

Yasser, M. F., & Alserafi, M. M. (2023). Cloud-Based Data Warehousing Solutions Capabilities and Challenges. 2023 International Conference on Computer Science and Information Technology (CSIT), 1–6.