**IEOR4524 Phase 5 Report**

**Chilean Highway Toll Collection**

**Sponsor**: Ardian Private Equity

**Group 32:**

Michelle (Meixi) Peng ([mp4270@columbia.edu](mailto:mp4270@columbia.edu))

Kris (Tiantian) Chen (t[c3315@columbia.edu](mailto:c3315@columbia.edu))

Vidhi Agrawal ([va2504@columbia.edu](mailto:va2504@columbia.edu))

Howard (Haozhou) Yu ([hy2818@columbia.edu](mailto:hy2818@columbia.edu))

Raphael Barthes ([rpb2141@columbia.edu](mailto:rpb2141@columbia.edu))

1. **Introduction**

**1.1. About the Sponsor**

The project is initiated by a private equity firm with a substantial presence across diverse geographies. The organization focuses on creating infrastructure investments and stands out for its commitment to technological innovation. Their high level of involvement in this project is driven by a resolve to address pressing infrastructural issues with cutting-edge technological approaches.

**1.2. Problem Background**

This project confronts a critical challenge in urban toll road systems: the efficient and timely collection of tolls. Chile’s free-flow tolling systems display increasing instances of toll evasion (“leakage”). Prior efforts to mitigate this issue have been inadequate, and following the pandemic, default rates have increased from 4% to 20%.

There is a notable gap in utilizing advanced data analysis to predict toll payment behaviors and effectively manage toll collection. Our project will address the need for a data-driven approach that not only anticipates payment defaults but also informs the decision-making process regarding the involvement of collection agencies is thus critical for improving revenue streams and maintaining the sustainability of the system. Moreover, the project will explore the relationship between macroeconomic indicators and payment delays, aiming to provide evidence that could support renegotiations of the fixed toll price agreed upon with the Chilean government.

**1.3. Project Goal**

**1.3.1 Main Objectives**

The project's main goal is to improve urban toll management using data analytics. We aim to develop a predictive model using toll transaction data to forecast payment defaults and estimate payment timings. This model will help optimize collection agency involvement and enhance toll revenue efficiency. Additionally, we plan to analyze how macroeconomic trends influence payment delays, linking economic factors with payment behaviors to inform negotiations about toll pricing with the Chilean government.

**1.3.2 Measures of Success**

Success for the predictive model will be measured by its accuracy in forecasting defaults, improved prediction of payment timings, and enhanced revenue collection efficiency. For macroeconomic analysis, success will involve clear, strong correlations between economic factors and payment delays, aiding strategic toll rate negotiations. Key deliverables include an operational predictive model, a detailed correlation analysis report, a strategic implementation plan, and comprehensive presentations on both analyses. Outcomes include improved toll collection efficiency, evidence-based support for toll rate renegotiation, and an adaptable toll management framework that can be effectively used by the Chilean team in the future.

1. **Data and Resources**

**2.1. Data Collection Methodology**

Data collection utilized two channels. The primary source is the Vespucio dataset from the Chilean toll system, integrated into a cloud data warehouse on Snowflake. This dataset includes comprehensive information on vehicles, passengers, billing, and payment, facilitated by Ardian’s IT and Data Science team. Vespucio has also given additional information on credit scores by providing the Equifax dataset which is also on Snowflake.

The secondary channel involved open-source macroeconomic data, focusing on four areas: overall economic health, inflation, household conditions, and urban transportation. This data was collected from publicly accessible datasets provided by organizations such as The World Bank, International Monetary Fund (IMF), Central Bank of Chile, National Statistics Institute of Chile, OECD Stats, and CEIC data, primarily in CSV or XLSX formats.We considered but ultimately did not use APIs from the IMF and Trading Economics, as our use of macroeconomic data is limited to correlational analysis and not integrated into our predictive model.

**2.2. Quantitative Data**

**2.2.1 Open Source Macroeconomic Data**

As defined in the previous section, we identified 4 categories of macroeconomic data with various geographic scopes and timeframes:

**(1) Overall Economic Health:** For this section, GDP indicators and the unemployment rate were 2 essential factors underconsideration. Details are specified as follows:

| **GDP Indicators** | **Scope** | **Timeframe** | **Frequency** | **Source** |
| --- | --- | --- | --- | --- |
| GDP, current prices (billions of pesos) | Chile | 1997-2023 | Quarterly | Central Bank of Chile |
| GDP, current prices (billions of U.S. dollars; PPP; billions of international dollars) | Chile | 2018-2028E | Annually | International Monetary Fund |
| GDP per capita, current prices (PPP; international dollars per capita) | Chile | 2018-2028E | Annually | International Monetary Fund |
| Real GDP growth (annual percent change) | Chile | 2018-2028E | Annually | International Monetary Fund |
| GDP by economic activity (sectors), current prices, reference 2018 (billions of pesos) | Santiago | 2013-2022 | Annually | Central Bank of Chile |
| **Unemployment** | **Scope** | **Timeframe** | **Frequency** | **Source** |
| Unemployment rate (percent) | Chile | 1980-2028E | Annually | International Monetary Fund |
| Unemployment rate (percent) | Santiago | 1980-2023 | Monthly | CEIC (paid data source) |

**(2) Inflation:**

| **Inflation Indicator** | **Scope** | **Timeframe** | **Frequency** | **Source** |
| --- | --- | --- | --- | --- |
| CPI, CPI without volatiles and volatile CPI, base 2023=100, index, spliced ​​information | Chile | 1998-2024 | Monthly | Central Bank of Chile |
| Inflation rate, average consumer prices (annual % change) | Chile | 1980-2028E | Annually | International Monetary Fund |
| Inflation, GDP deflator (annual %) | Chile | 1961-2022 | Annually | Chile |
| Implied PPP conversion rate (national currency per international dollar) | Chile | 1980-2028E | Annually | International Monetary Fund |
| Transportation Cost Index (TCI), national coverage - base year 2018, analytical indices | Chile | 2019-2023 | Monthly | The National Statistics Institute of Chile |
| Supermarket sales index (ISUP) at constant prices, index, 2018=100 | Santiago | 2014-2023 | Annually | Central Bank of Chile |

**(3) Household conditions:**

| **Household** | **Scope** | **Timeframe** | **Frequency** | **Source** |
| --- | --- | --- | --- | --- |
| Annual national disposable income, net saving, lending and borrowing, in millions of Peso | Chile | 2019-2022 | Annually | OECD Stats |
| Household debt, loans and debt securities (% GDP) | Chile | 2002-2022 | Annually | International Monetary Fund |

**(4) Urban Transportation:**

| **Drivers, Vehicles, and Traffic Info** | **Scope** | **Timeframe** | **Frequency** | **Source** |
| --- | --- | --- | --- | --- |
| Number of driver's licenses processed (by class, type of processing, age, sex, and region) | Chile | 2011-2022 | Annually | The National Statistics Institute of Chile |
| Passage of vehicles through toll plazas and interurban highway gantries, units | Santiago | 2014-2023 | Annually | Central Bank of Chile |
| Passage of vehicles through urban highway portals, units | Santiago | 2014-2023 | Annually | Central Bank of Chile |
| Fleet / Park of taxi, buses, minibus, school bus, units | Santiago | 2014-2023 | Annually | Central Bank of Chile |

**2.1.2 Vespucio Dataset**

The Vespucio dataset is a structured dataset including tables on calendars, billing cycles, contracts, client information, credit information, costs, patents, billing documents, and more. Our analysis focuses on tables on billing documents, which include data points of clients and vehicles, as well as invoice and payment details such as invoice dates, payment dues, payment statuses, transaction amounts, and transaction generation types. The Vespucio dataset has 482M rows and 44 columns and ranges from 2006 to 2024.

**2.1.3. Equifax Dataset**

Equifax is a global data, analytics, and technology company known for providing information solutions to businesses and consumers. In Chile, Equifax operates as a credit reporting agency, offering credit information and risk management services to financial institutions, businesses, and consumers. The score is usually between 280 to 850, with basic interpretations as follows:

| 300 - 579 | 580-669 | 670-739 | 740-799 | 800+ |
| --- | --- | --- | --- | --- |
| POOR | FAIR | GOOD | VERY GOOD | EXCELLENT |

**2.3. Qualitative Data**

Our project exclusively utilizes structured data from the Vespucio dataset, Equifax credit information, and macroeconomic indicators. Given the quantitative nature of our objectives and the robustness of our data, qualitative data collection is not part of our consideration.

**2.4. Data Processing and Analysis**

**2.4.1 Open Source Macroeconomic Data**

We processed open-source macroeconomic data for correlation analysis by downloading it in CSV and XLSX formats and using AWS for processing. We formatted data on EasyMorph by converting it from horizontal to vertical format, matching dates with corresponding data descriptions and values. The reporting date was standardized across various data types, with an additional date column for consistency, crucial for integrating with the Vespucio dataset. We meticulously preprocessed data to remove empty fields and redundancies, appending an extraction date to each table. Lastly, we used EasyMorph connectors for bulk export to the Snowflake environment, preparing the datasets for further correlation analysis.

**2.4.2 Vespucio Dataset Exploratory Data Analysis**

The analysis is being carried out using python libraries like snowpark, pandas, seaborn, matplotlib and sweetviz

**(1) Equifax and Vespucio Data Joining:** Joined the Vespucio billing data with Equifax score information on RUT (the Chilean Unique identifier). This gives us a direct insight into the credibility of the person associated with the transaction and helps us analyze if Equifax has a correlation with payments made on time.

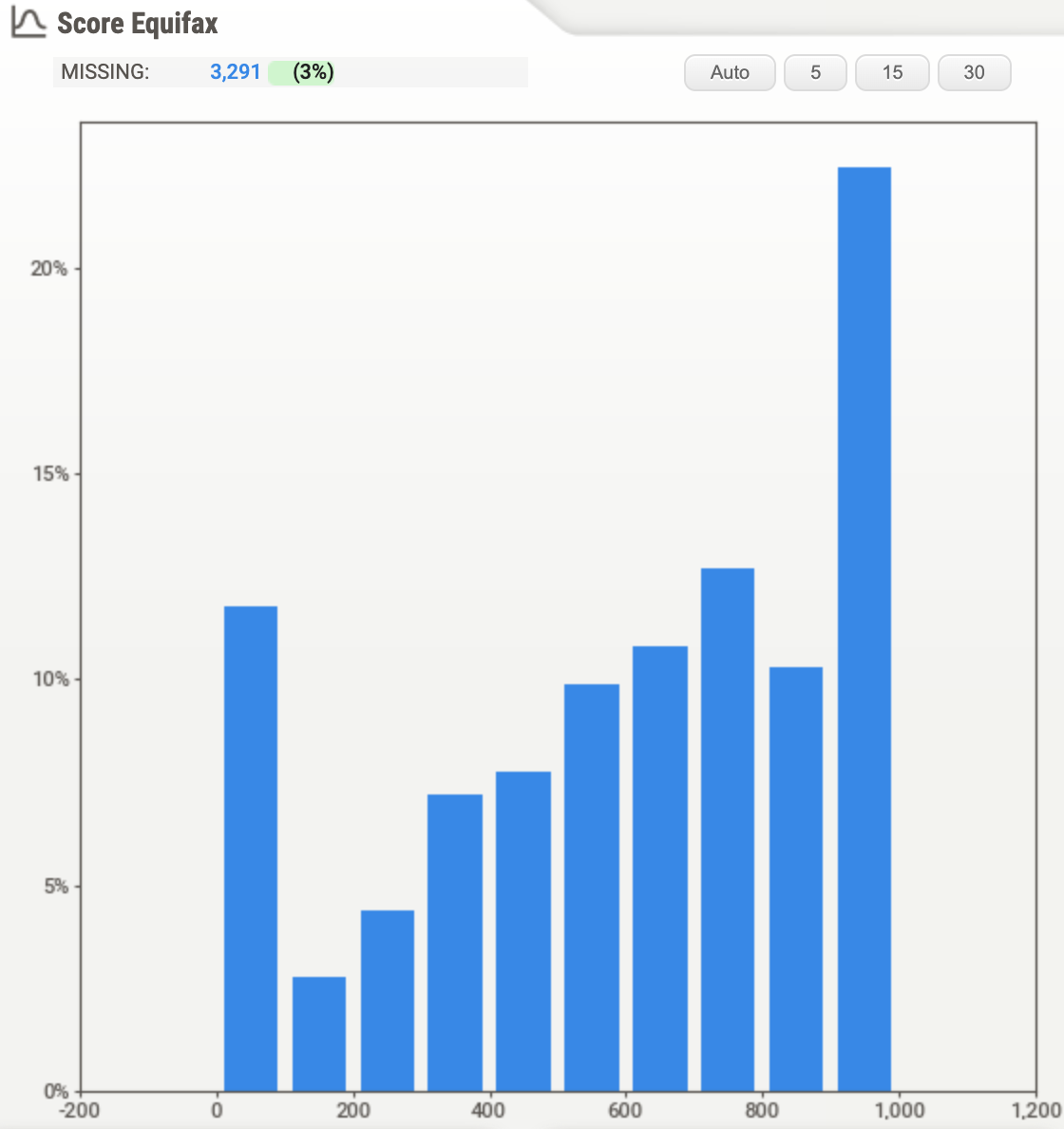
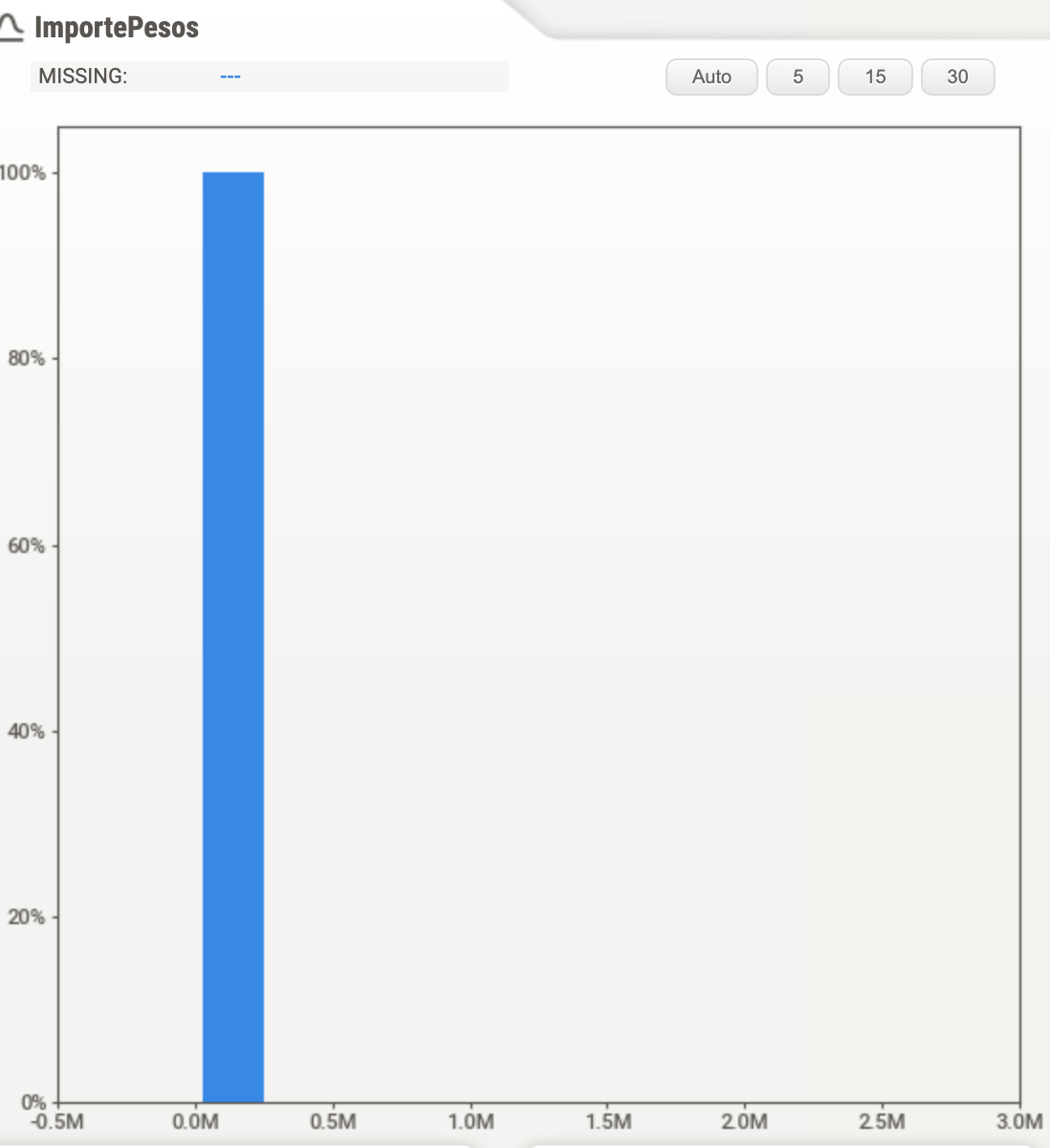
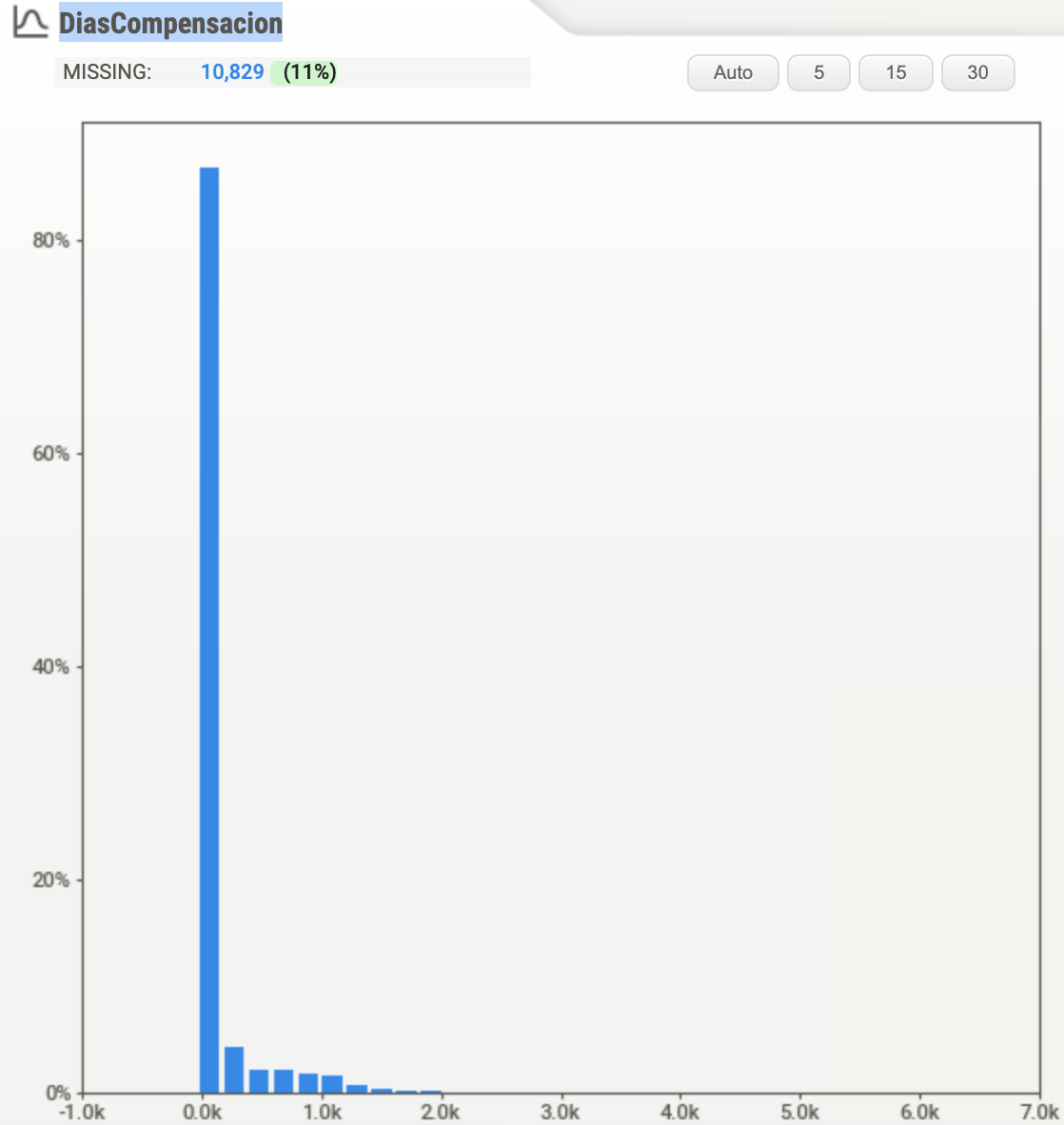
**(2) Missing values**: Checked for missing values in the dataset. Around eleven columns had missing values. Upon inspection we see three columns that have greater than 50% of the data missing. These columns are primarily date columns that are not relevant for the analysis and hence we decided to drop the three.

**(3) Duplicate Check:** No duplicates were observed in the entire dataset.

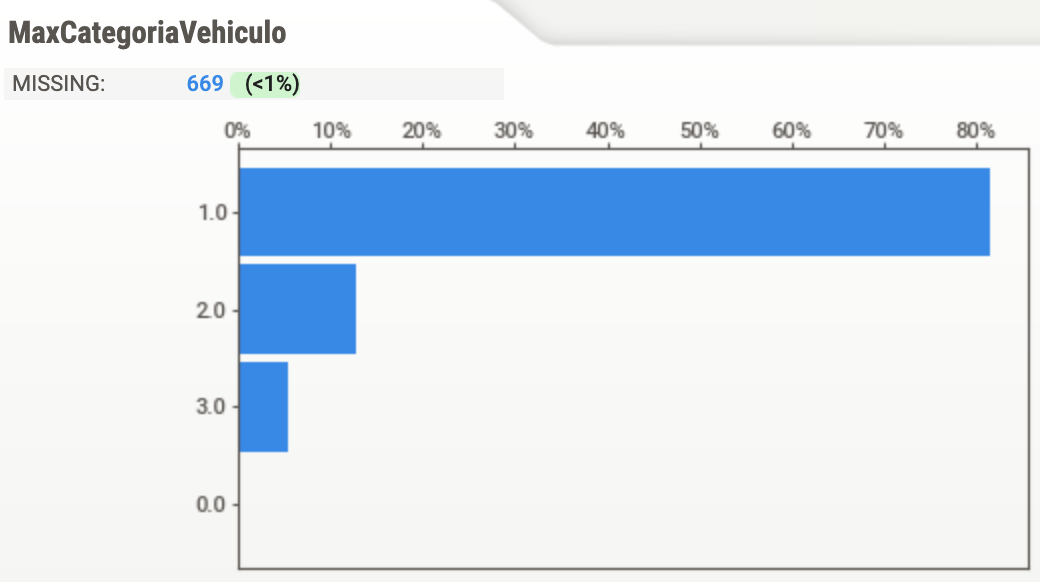
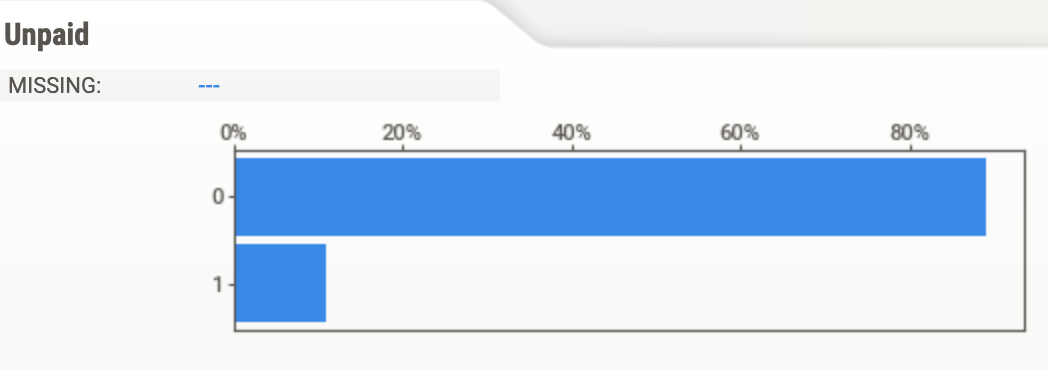
**(4) Sampling and Translation:** Extracted a 100,000 observation sample from the Snowpark data, which we then processed in Python, translating Spanish column names to English using the deep\_translator library. Augmented the dataset with custom variables tailored to our modeling needs.

**(5) Feature Creation:**Introduced a binary variable "Unpaid" to distinguish between paid and unpaid transactions to act as our target variable for the model. Incorporated a variable named "Days\_to\_Pay," indicating the number of days between the transaction emission date and the payment date (or the current date if the transaction remains unpaid). Extraction of year, month columns from Due-date and Payment- date columns.

**(6) Univariate Analysis**: We had around 41 columns to analyze and below we can see the distributions of a few important ones. Key numerical columns like ImportePesos(Amount due) and DiasCompensacion (payment date- due date) are heavily *skewed*, so we perform log transformations to reduce the skewness of numerical data. Equifax score looks decently distributed.

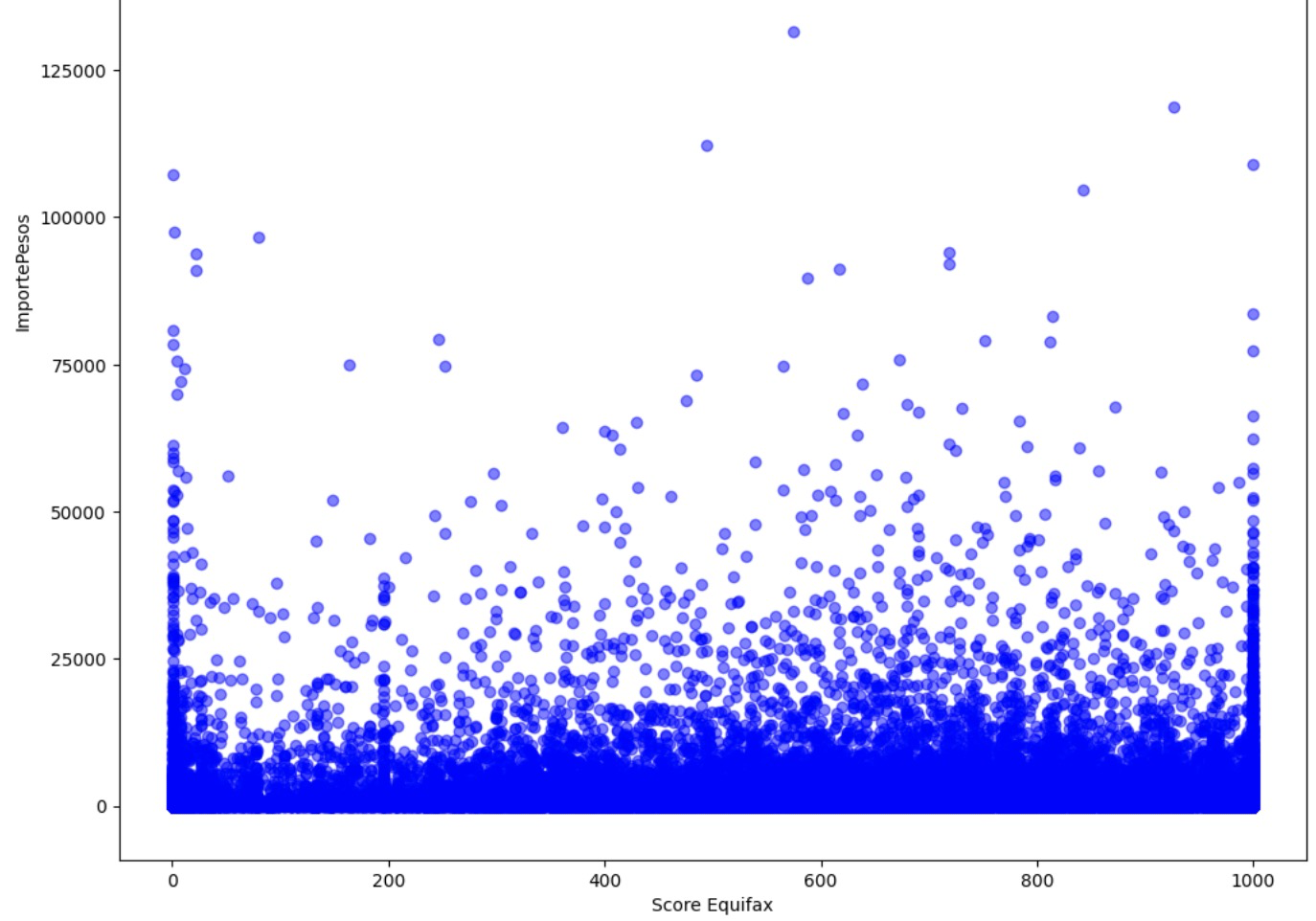
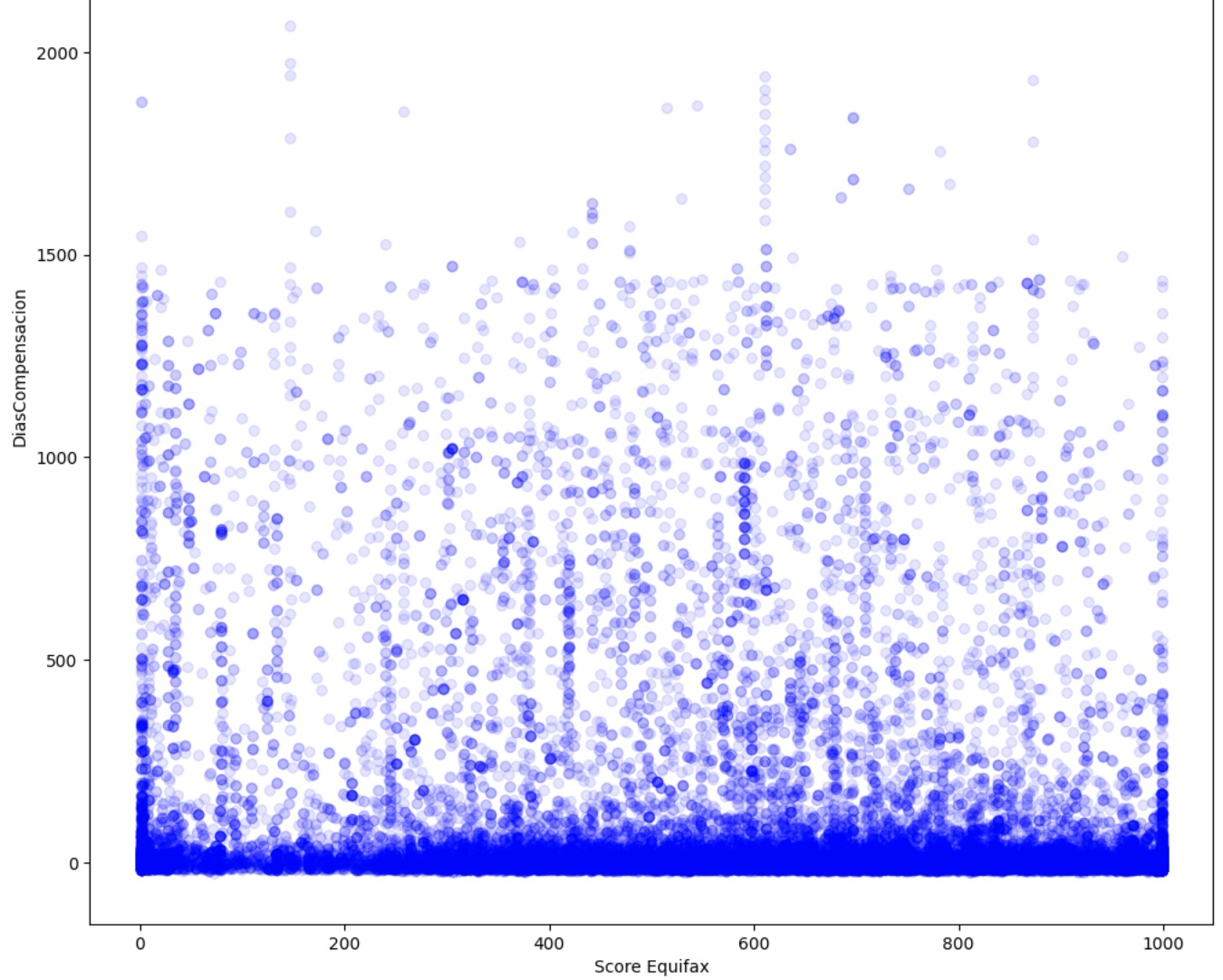
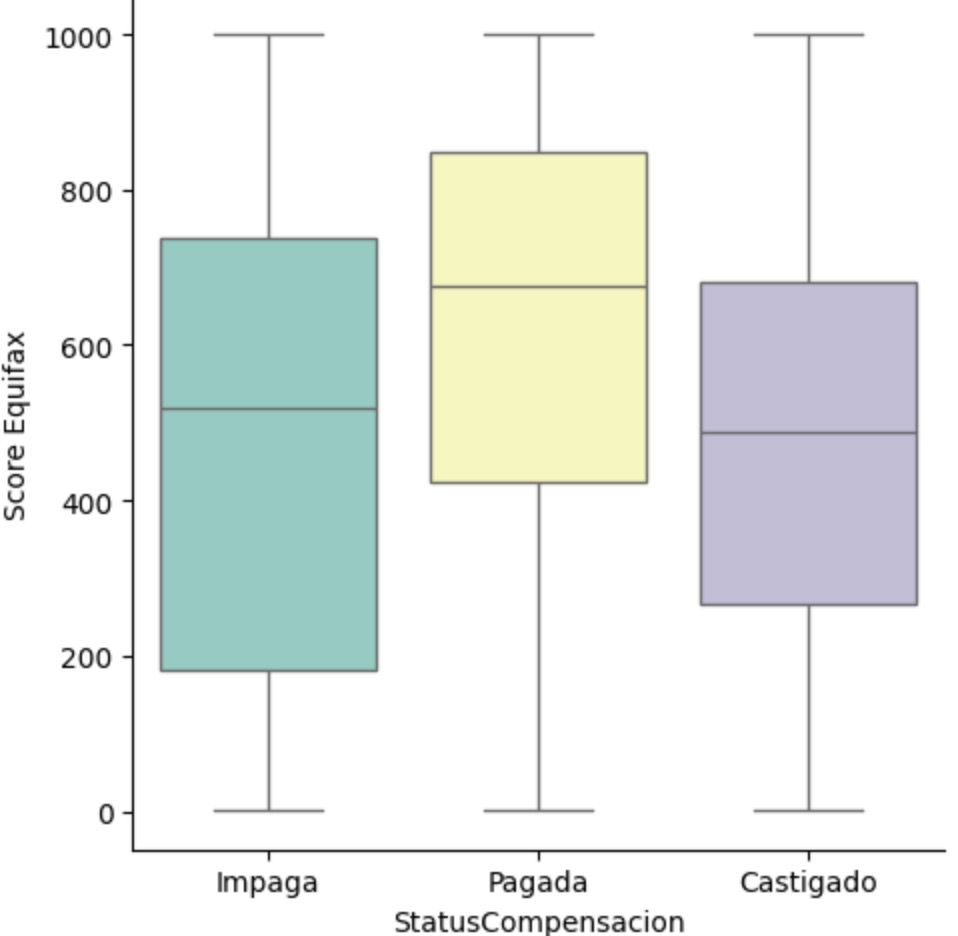


Similarly for a few important categorical variables, we get the following distribution. The target variable looks imbalanced with around 89% paid invoices and 11% unpaid invoices. The second graph shows the distribution of cities and the third shows the distribution of the category of motor vehicles.



(7)**Bivariate Analysis:** We ran a rigorous bivariate analysis and we can see a few associated with Equifax score below. According to the graph, the Pagada box apparently has a median higher than other two boxes, and the range of Pagada box is also located in a higher numerical position. Generally, we can conclude that the equifax score for Pagada is more likely to be higher than the other two categories.

In order to explore the relationship between numerical variables, we decided to use scatterplot. Unlike what we had expected, the days taken for payment and the payment amount don’t seem to have a relationship with the Equifax Score at first glance, and in fact, it seems that in both scatter plots, each Equifax Score has a similar pattern related to different numbers on y-axis, and most observations are concentrating in the relatively low range of the y-axis.



Furthermore, as the individuals in Equifax Score are mixed with natural individuals and firms, we divided the Equifax Score by types of individuals, and tried to explore if each type of individuals behave differently. The result of the grouped average within each type is as follows

|  | **Pagada** | **Impaga** | **Castigado** |
| --- | --- | --- | --- |
| **Natural** | 623.094449 | 496.404199 | 497.199828 |
| **Firms** | 581.999811 | 41.573864 | 283.550562 |

As shown above, both types have followed the trend that the Pagada group has a higher average score than other two groups. Meanwhile, for the firms, the distinction between Pagada and Impaga is sharper than natural individuals.

We also explore how frequency of unpaid transactions have changed over time and how the payment lag and amount due has changed over years using box plots. Our EDA is an extensive html report that can be discussed outside the report for in depth understanding.

**2.5. Feature Exploration and Statistical Analysis**

**2.5.1 Feature Exploration for Vespucio Data**

In addition to the univariate, bivariate analysis shown above,

**(1)** We have dropped identifier columns (DocumentID, ReceiptNumber, BusinessPartnerID, ClientID, RUT, ContractID) as of now as they are unique identifiers and will not be helpful in predicting the likelihood of defaulting on payments and the model will treat them as numerical values. If treated as categorical, the dataset will become excessively sparse. But we will still explore the possibility of inclusion during recursive feature elimination. Apart from this, we also ran a pearson and chi-squared correlation for numerical and categorical columns. Using the correlation coefficient, we eliminated columns which had high correlation with other predictor variables to avoid multicollinearity. For ex. in the format Dropped(Retained)- Region(City), StatusCompensacionTmp(StatusCompensacion), ​​OperacionParcial(OperacionPrincipal) and similar. We dropped 3 columns with over 50% missing values.

**(2)** Numerical columns with skewed distributions are normalized using log transformation methods

**3. Modeling**

Our current modeling efforts focus on macroeconomic data correlation analysis, and has provided a robust understanding of the influence of macroeconomic factors on key variables. Regarding the prediction model, it has progressed to a preliminary logistic regression model that serves the purpose of classifying the likelihood of due payments. It's important to note that this model is in its nascent stages. Comprehensive data processing and meticulous feature engineering will be continuously integrated enhancing the model’s predictive accuracy and reliability. The following sections will focus on our modeling efforts for the macroeconomic models. The challenges remain consistent for all parts of the project.

**3.1 Model Selection**

**3.1. 1 Models Considered**

Initially, the team selected **Ordinary Least Squares (OLS) Regression**.It’s interpretability and straightforward implementation made it an ideal choice for initial exploratory analysis, particularly useful in delineating the relationships between critical variables such as payment timings and macroeconomic factors, allowing the team to quickly derive preliminary insights. However, the OLS methodology encountered difficulties due to the extensive size of the dataset. The OLS approach, while beneficial for smaller datasets, proved inefficient when scaling up to the full scope of the project’s data. This necessitated a shift to a more robust solution that could handle scaling requirements without compromising the depth of analysis.

In response to these issues, the team pivoted to leveraging **Snowflake's built-in linear regression functions**. Snowflake's platform, designed for cloud-based, large-scale data operations, offered a suite of functions tailored for handling large datasets. Its linear regression capabilities provided the necessary computational power and scalability to analyze the voluminous data. The integration of Snowflake’s built-in functions enabled the team to conduct a comprehensive analysis with the robustness required by the project's expansive data demands.

**3.1.2 Rationale Behind Model Selection**

The transition from Python's in-memory processing to the more robust data warehousing capabilities of Snowflake was necessitated by the consideration of **data size and scalability**. Snowflake’s platform was adept at managing and processing this large dataset, ensuring that there was no compromise on performance.

Snowflake also offered **computational efficiency**, crucial for the timely analysis of large data volumes in a cloud-based environment. This was key to the project's success, where prompt data analysis was critical.

**Methodological balance** was achieved by leveraging Python's statistical libraries for their ease and depth in initial stages, and Snowflake's capabilities for processing the full dataset with complex computations. Despite Snowflake's limited detailed statistical outputs, a parallel analysis approach was adopted. The team compared Snowflake's coefficients and intercepts with those from Python's limited dataset analysis to infer statistical significance. This strategy allowed the effective use of both platforms, ensuring a comprehensive statistical analysis.

**3.2. Modeling Methodology**

**3.2.1 Development Process- Macroeconomic Indicator Correlations**

In the **initial data preparation stage**, our team standardized data formats and ensured consistency to lay the groundwork for accurate and reliable analyses.Next, we conducted an **initial analysis using Python's `statsmodels` library** for regression analysis on data subsets. This provided preliminary insights and helped in designing scalable models for Snowflake's environment, serving as a guide for developing more complex models. We then **integrated Snowflake's Snowpark library** for enhanced data frame operations within Snowflake. This strategic move allowed us to fully utilize Snowflake's data processing tools for an optimized analytical workflow. Adapting our methodology to Snowflake involved realigning data queries and analysis strategies to leverage its strengths, a shift that was both technical and strategic for handling complex data sets. The modeling process involved **iterative testing and refinement**, using initial Python findings as benchmarks to improve and validate models within Snowflake. This iterative approach was crucial for enhancing model accuracy and validity. Finally, a **hybrid approach** was used for model validation. Given Snowflake's limitations in detailed statistical outputs, we compared coefficients and intercepts from Snowflake's full dataset analysis with Python's limited data analysis. This ensured that significance inferred from smaller datasets could be extrapolated to the larger dataset analyzed in Snowflake.

**3.2.2 Challenges and Solutions**

Modeling within this project faced several challenges, notably **managing and analyzing large datasets efficiently**. The complexity and volume of the data necessitated a scalable, robust solution. The Snowflake platform, with its cloud-based data warehousing capabilities, was chosen for its ability to handle large datasets effectively, offering scalability in data management without sacrificing speed or accuracy. Another challenge was conducting **detailed statistical analysis within computational constraints**. Python's statistical libraries provide depth but struggle with scalability for large datasets. To address this, the team adopted a multifaced approach: detailed analysis on smaller samples using Python and broader analysis on the full dataset using Snowflake. This strategy maintained the depth of Python's analysis when scaling up to Snowflake's robust but less detailed environment. The team also faced the complexity of **adapting to Snowflake's unique coding and querying requirements**. Transitioning from traditional statistical programming languages to Snowflake's coding paradigm required a learning period. Despite this, the team effectively harnessed Snowflake's advanced data processing capabilities, integrating them into the project's workflow.

**3.3. Preliminary Results and Insights**

**3.3.1 Results and Insights on Macroeconomic Indicator Correlations**

In our macroeconomic data analysis, we focused on four sectors: General Economic Health, Inflation, Household Conditions, and Urban Transportation. We used Snowflake's linear regression and OLS regression on different data scales to ensure reliability and depth in our findings, tailored to these sectors.

**(1) General Economic Health:** Analysis revealed weak national correlations with toll payment timings, but stronger correlations in Santiago-specific data, suggesting localized economic impacts on payment behaviors. Notably low Durbin-Watson scores indicated time lags, and high Jarque-Bera test scores showed deviations from normal distribution, suggesting the need for potentially more complex models.

**(2) Inflation:** Inflation showed a significant correlation with toll payment delays across Chile, though complex with potential other influencing factors. Durbin-Watson scores indicated autocorrelation, and Jarque-Bera scores pointed to non-normal distribution of residuals, highlighting the complexity of this relationship.

**(3) Household Conditions:** Strong correlations were found between household conditions and toll payment timings in Chile, with potential autocorrelation and significant deviations from normal distribution in residuals, indicated by high Jarque-Bera scores. This suggests complex, non-linear, dynamics influencing toll payments.

**(4) Urban Transportation:** Transportation metrics showed a strong correlation with toll payment behaviors, with similar autocorrelation levels as seen in inflation data. Extremely high Jarque-Bera scores highlighted non-normal residual distributions, underscoring the complexity of urban transportation factors affecting toll payments.

# **4. Next Steps**

**4.1 Current Distribution of Responsibilities**

Our team's responsibilities are divided based on three key areas: macroeconomic correlation analysis, prediction model development, and Equifax data analysis.

**Kris**: Her responsibilities focused on building the data pipeline, including data collection, preprocessing, and storage. She gathered data from multiple trusted sources, verifying its credibility and ensuring its suitability for analysis. She then preprocessed the data to establish a clear structure and uniform format. In the AWS workspace, she built the data connector and imported datasets to Snowflake. She also worked collaboratively on running the models developed by Michelle.

**Michelle**: Took charge selection and model building and for both the full dataset and a limited subset. Her efforts led to the full utilization of snowflake data, the development of statistical models used to analyze the macroeconomic data, and the provision of generalizable code for the group. She also played a crucial role in liaising with sponsors and professors, contributed to writing general sections of the report, and was responsible for the review and editing process.

**Vidhi:** Was responsible for laying out the foundational code for Snowflake that helped the team connect to Snowflake instance and utilize the relevant data from the snowflake warehouse. She then worked on the entire exploratory data analysis, pre-processing(on-going) and feature engineering(on-going) for the Vespucio data which will then be utilized to build the prediction model.

**Raphael:** Was responsible for the preliminary logistic regression on the Vespucio Dataset.

**Howard:** Was responsible for analyzing the Equifax data. His role involved extracting insights from the credit information provided by Equifax, determining its relevance and impact on toll payment behavior.

**4.2 Next Steps for Macroeconomic Indicator Correlations**

Our analysis could be expanded by exploring additional datasets related to household conditions and urban transportation, which are complex due to detailed categorizations. For example, data on national disposable income, savings, lending, and borrowing could provide more insights into household financial health. Similarly, datasets categorizing driver's licenses by class, age, gender, and region could enhance our understanding of urban transportation. Based on our initial findings, we plan to explore the complex relationships between macroeconomic factors and toll payment behaviors more deeply. Our analysis has shown varying correlation levels across sectors, with patterns of autocorrelation and deviations from normal distribution in residuals. To address these, we aim to employ advanced modeling techniques, including non-linear and time-series analysis. The division of tasks is still to be finalized. Future steps and adjustments will be guided by feedback from our sponsor and our ongoing analysis.

**4.3 Nexts Steps for Vespucio Dataset**

As we got the complete data very recently, we’re still working on parts of data exploration and preprocessing. The dataset is huge(470M records) and hence demands more time for a rigorous analysis.

**Advanced Data Cleaning Techniques:** Beyond basic preprocessing, explore advanced cleaning techniques like encoding, anomaly detection and data imbalance handling using unsupervised/simulation methods. This can help in identifying and handling outliers more effectively. Additionally filter data based on business constraints for the prediction model.

**Advanced Feature Engineering:** Deepen focus on advanced feature engineering methods, including tree-based feature selection, addressing multicollinearity, and utilizing recursive feature elimination. These techniques can help in identifying the most impactful predictors and refining the model.

**Parallel Processing and Optimization:** Utilize Snowpark's capabilities for parallel processing to handle the large dataset. Consider optimizing queries and leveraging cloud-based compute resources to expedite data processing.

**Prediction model & Ensemble Methods:** Consider using ensemble methods which combine multiple models to improve predictive performance. Techniques like bagging, boosting, and stacking will be explored. Additionally, work on Hyperparameter tuning to fine-tune model metrics.

**Model Validation and Cross-Validation:** With such a large dataset, ensure rigorous model validation. Techniques like k-fold cross-validation can provide a better assessment of the model's performance and generalizability.

**Explainable AI (XAI) Techniques:** Integrate XAI methods to interpret complex model predictions. This is crucial for understanding model behavior and for making the model's decisions transparent to stakeholders.

**4.4 Next Steps for Presentations**

Moving forward,, we are focusing on two important presentations: the Mid-Term presentation scheduled for April 3rd with our sponsor, and the final presentation set for the week of May 6th.

During the **Mid-Term presentation**, our team will collaboratively develop a detailed report and presentation slides. Each member will highlight their specific analytical findings: Vidhi and Raphael will discuss feature engineering and predictive models, Kris and Michelle will address macroeconomic correlations, and Howard will explore the Equifax data insights. This session will primarily showcase our discoveries about how macroeconomic factors affect toll payment behavior. Each team member is responsible for drafting their report sections and slides, aiming for clear and engaging communication of our insights. The initial drafts are targeted for completion by March 28th, followed by a review and refinement period of one week. For the **final presentation**, we will provide a comprehensive summary of our project's results. This includes detailing our predictive model, the efficacy of different collection methods, and an in-depth analysis of macroeconomic factors' impact on payment behaviors. Team members will again be preparing their individual analysis sections for the report and presentation, ensuring thorough and expertly informed content for each part.