**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 primary objective of this project is to enhance the strategic management of urban toll collections through data-driven analytics. The key endeavor will be to develop a predictive model informed by toll transaction data, engineered to forecast payment defaults and accurately estimate the timing of payments based on data obtained through the toll system. Through this predictive mechanism, we aim to empower decision-makers with insights to optimize the involvement of collection agencies, thereby bolstering the efficiency of toll revenue collection.

Another critical component of the project is to undertake an analysis of the interplay between macroeconomic trends and the occurrence of payment delays.By establishing a quantifiable link between economic factors and payment behaviors, the project seeks to provide a solid analytical foundation to support renegotiation of the fixed toll price currently set with the Chilean government.

**1.3.2 Measures of Success**

For the predictive model, success will be marked by its accuracy in forecasting payment defaults, an increase in the precision of payment timing predictions, and a tangible improvement in the efficiency of revenue collection. The impact of the predictive model will be quantifiable through a reduction in the rates of payment default and an increase in the consistency and timeliness of revenue streams.

For the macroeconomic analysis, success will be evaluated based on the clarity and strength of the correlations identified between economic factors and payment delays, expected to yield insights that could guide strategic negotiations for toll rate adjustments. The ultimate indicator of success for the correlation analysis will be its utilization in renegotiating toll rates, leading to a pricing structure that maintains the system's financial health.

Key deliverables for the project include a fully operational predictive model, an exhaustive correlation analysis report, a strategic plan for the model’s implementation, and a pair of in-depth presentations that explain both the predictive and correlation analyses. The expected improvements from these deliverables include a boost in toll collection efficiency, evidence-based support for toll rate renegotiation, and an adaptable prediction framework fostering the long-term sustainability of toll operations amidst fluctuating economic environments.

1. **Data and Resources**

**2.1. Data Collection Methodology**

Data collection was streamlined through two channels. The primary source is the Vespucio dataset, a product of the Chilean toll system that integrates highway transportation data into a single cloud data warehouse in the Snowflake Platform. Advised by Ardian’s IT and Data Science team, the system automates the collection and processing of toll-road passage records, encompassing detailed information on vehicles, passengers, billing, and payment collection activities. This makes the dataset readily accessible on Snowflake, ensuring the timely and efficient handling of data for analysis.

The second channel involved open-source macroeconomic databases. To initiate the collection process, we categorized relevant macroeconomic data into 4 principal areas: overall economic health, inflation, household conditions, and urban transportation information. Additionally, we sought to encompass two levels of geographic data: national-level data from Chile and city-level data from the Santiago Metropolitan Area, where the highway is located. We also collected datasets across various temporal dimensions: annually, quarterly, and monthly.

We successfully collected the macroeconomic data from publicly accessible datasets provided by esteemed organizations such as The World Bank, International Monetary Fund (IMF), Central Bank of Chile, National Statistics Institute of Chile, OECD Stats, and CEIC data. These datasets were typically available in CSV or XLSX format, allowing for straightforward integration and analysis.

We evaluated the use of APIs offered by the IMF and Trading Economics to streamline our data analysis process. However, we ultimately chose not to utilize these APIs because our current use of macroeconomic data is confined to correlational analysis: we do not plan to incorporate macroeconomic data into our predictive model, thus deeming the automation of data retrieval through APIs unnecessary.

**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 | The World Bank |
| 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**

To prepare the open-source macroeconomic datasets for correlation analysis in Snowflake, we built our ETL pipelines. We began by downloading the data in CSV and XLSX formats and processed them in the AWS environment. Our initial step was to employ EasyMorph to reformat the datasets, transforming them from a horizontal structure to a vertical one by unpivoting dates into a single column, and matching them with each row of accordance data descriptions and values. We then standardized the reporting date across various data types to ensure consistency, adding this as a new column in date format to be prepared to join with billing information from the Vespucio dataset. Additionally, we meticulously preprocessed the data to eliminate empty fields and redundancies and appended an extraction date to each table, achieving a structured dataset ready for analysis. Finally, we set up connectors in EasyMorph to bulk export the cleaned data into our Snowflake environment, ensuring that our datasets were primed for the subsequent stages of correlation exploration.

**2.4.2 Vespucio and Equifax Data**

**(1) Join Vespucio data and Equifax**: We 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.

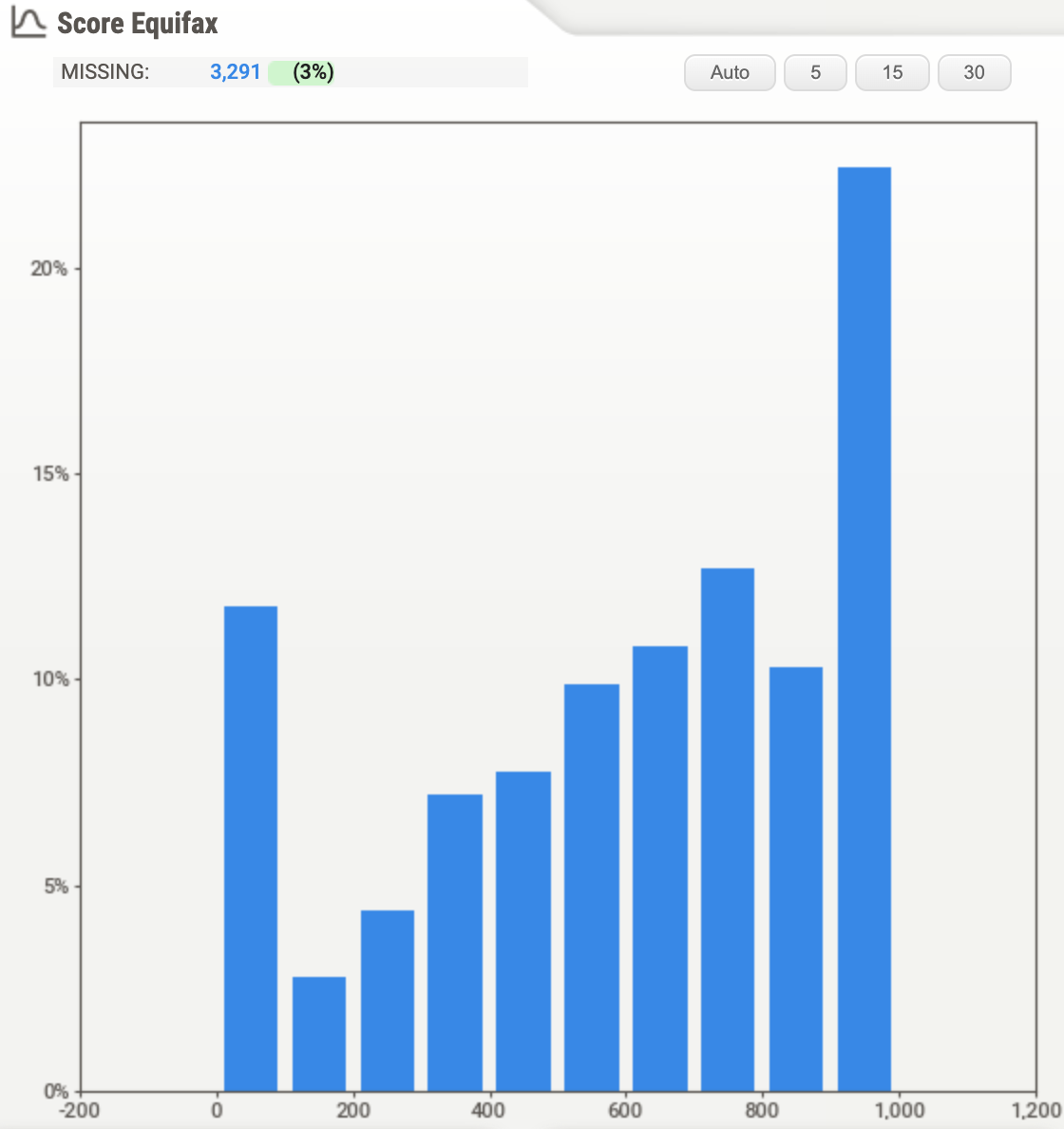
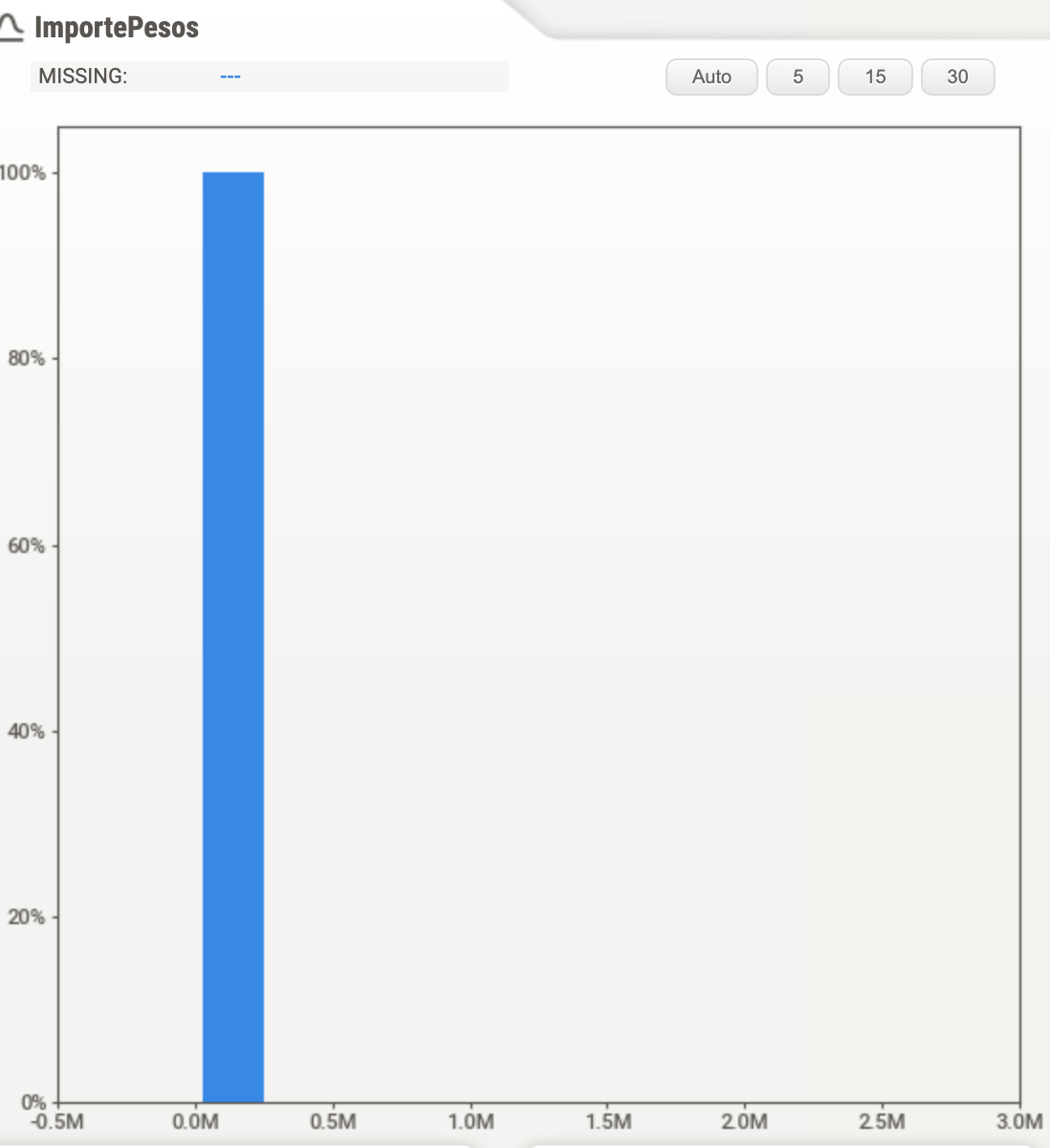
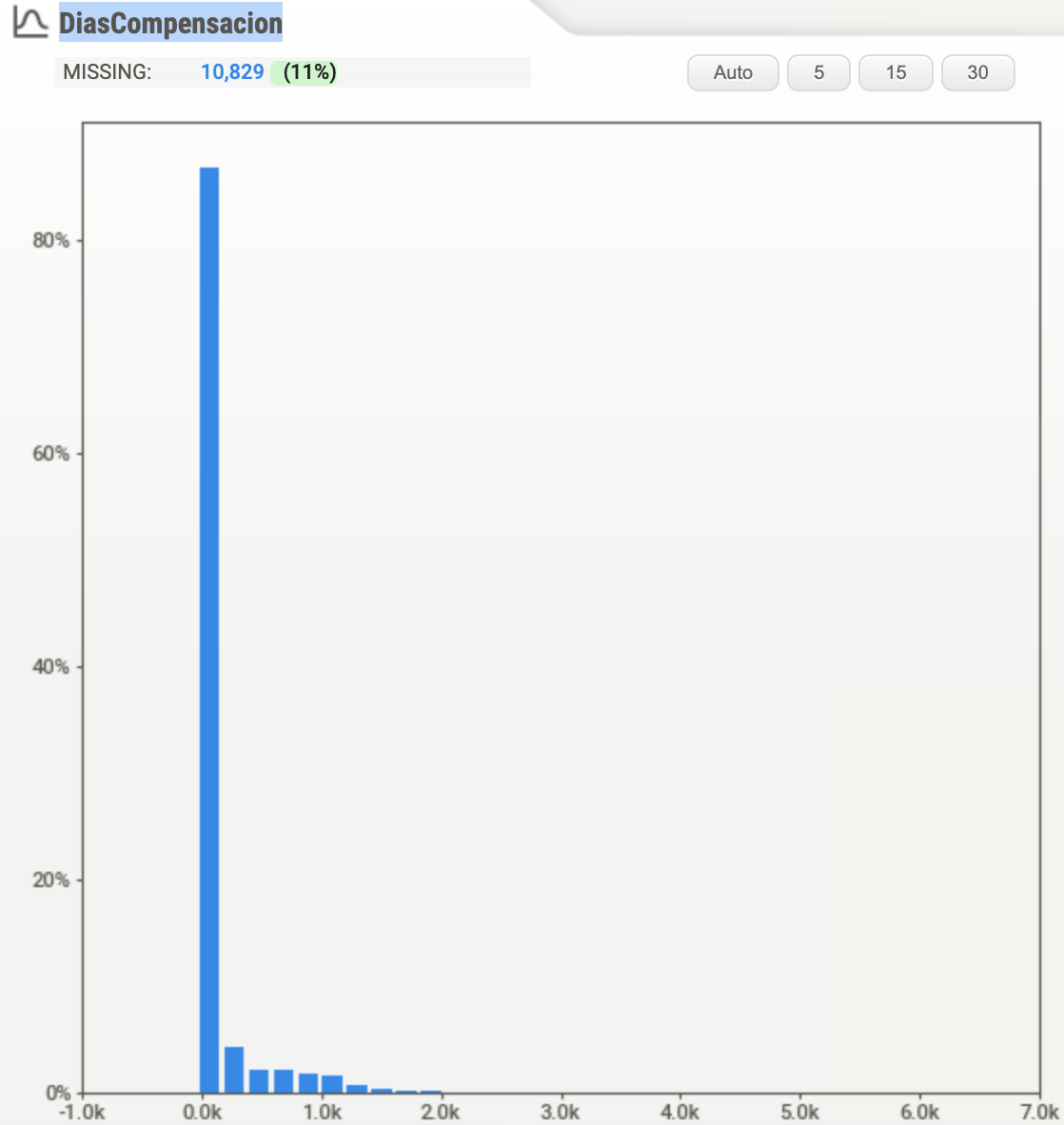
**(2) Check for missing values**: As the first step, we first checked for missing values in the dataset. Around eleven columns had missing values. Upon inspection we see three columns that have greater than 50% of the data missing. These columns are primarily date columns that are not relevant for the analysis and hence we decided to drop the three. (how should we impute this?)

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

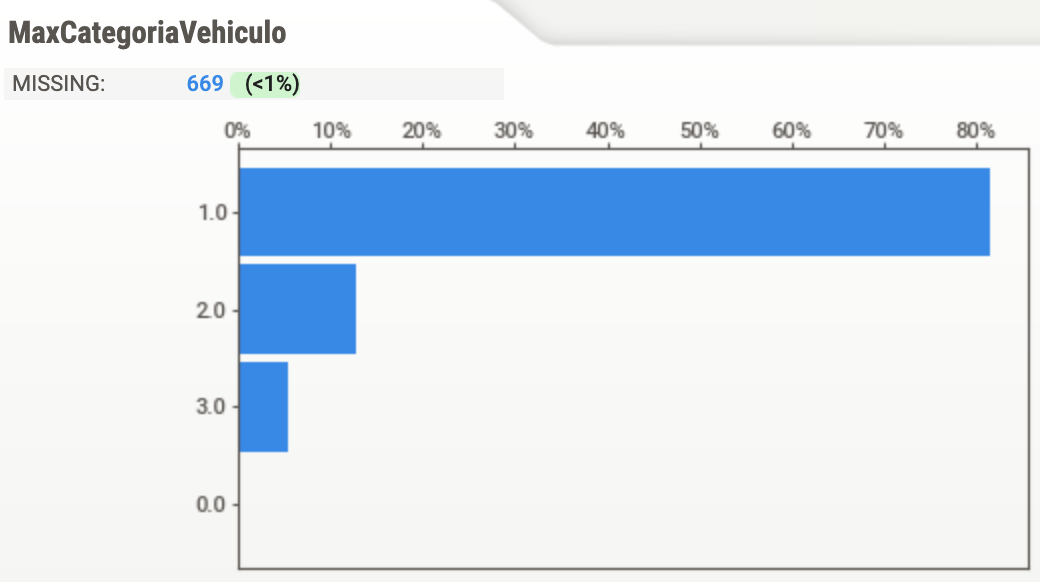
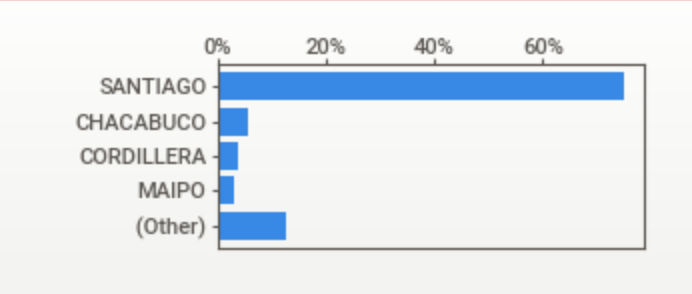
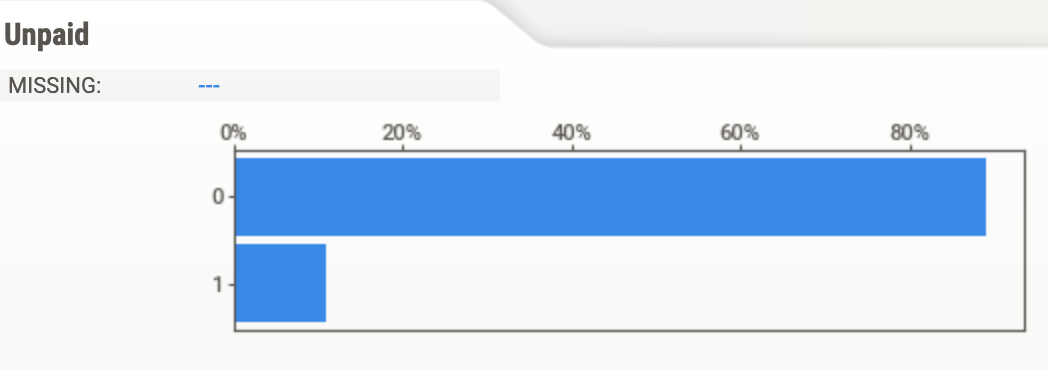
**(4) Sampling and Translation:** In the Vespucio dataset, we started off by extracting a subset of the Snowpark data containing 100,000 observations, which we then processed in Python. Since the data was in Spanish, we employed the deep\_translator library to accurately translate the column names into English, simplifying subsequent data manipulation tasks. Additionally, we augmented the dataset with custom variables tailored to our modeling needs.

**(5) New Feature Creation:**We introduced a binary variable "Unpaid" to distinguish between paid and unpaid transactions. This will act as our target variable for the model. Furthermore, we 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).

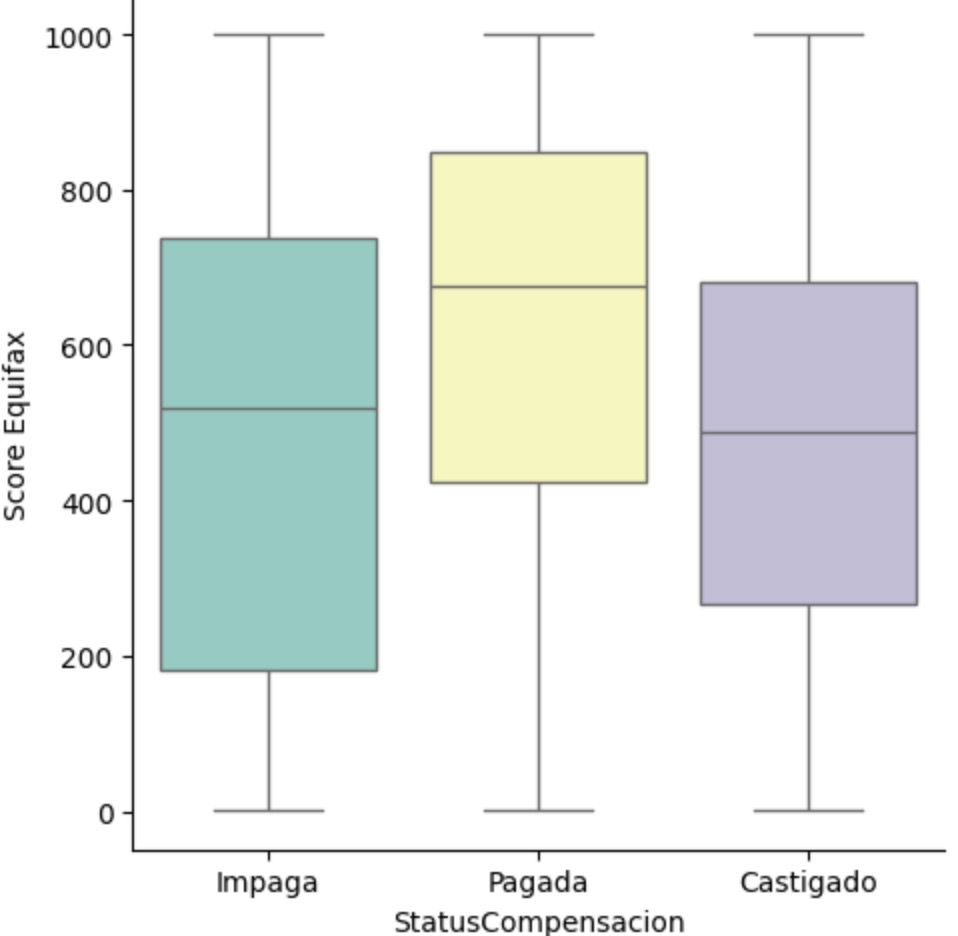
**(6) Univariate Analysis**: We had around 41 columns to analyze and below we can see the distributions of a few important ones. The important numerical columns like ImportePesos(Amount due) and DiasCompensacion (payment date- due date) are heavily skewed. We perform log transformations to reduce the 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.



Bivariate Analysis



**2.5. Feature Exploration and Statistical Analysis**

**2.5.1 Feature Exploration for Vespucio Data**

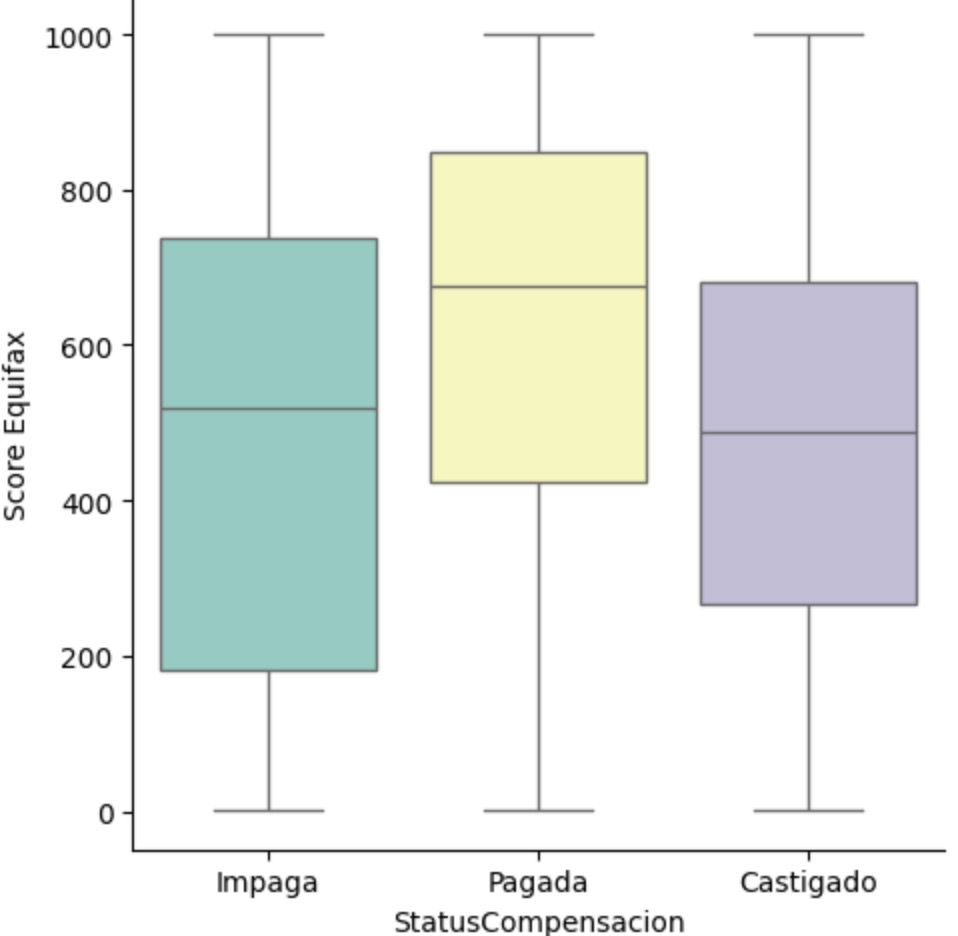
**2.5.2 for Equifax**

We assumed that the Equifax score, as a credibility score, will be a good predictor of an individual’s payment behavior, and we further assumed that for the group who paid the fee, individuals in that group should have a higher average Equifax score. We also assumed that the higher equifax score group should behave better in paying, meaning that they may have a shorter payment period, or have a higher payment amount. With these assumptions, we carried out basic explorations on the relationships between “Score Equifax” and “StatusCompansacion” (paid or not), “DiasCompansacion”(days to make payment), and “ImportePesos” (payment amount) respectively.

**2.6 Data Analysis**

**2.6.1 Score Equifax vs StatusCompansacion**

In order to show numerical values distributed across different categories, we decided to use box plot to show score equifax distribution across the three scenarios of impaga (unpaid), pagada (paid), castigado (cancelled). The graph generally coincides with what we expected. 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. Apart from the median in the graph, we further calculated the mean score of the three groups respectively, which were Castigado–467.394343, Impaga–467.523932, Pagada–614.726527, another evidence supporting our assumption on higher equifax score for group Pagada.

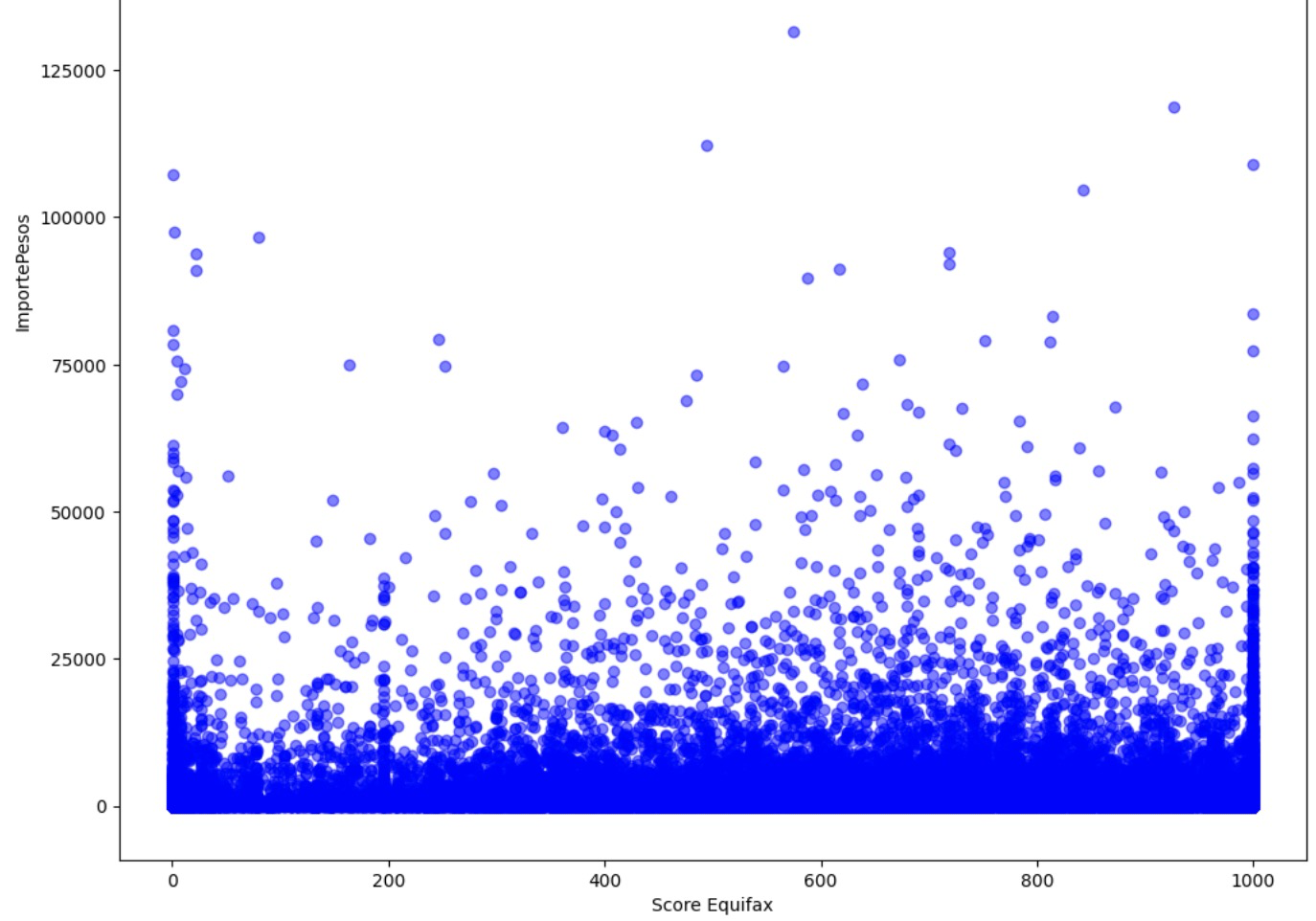
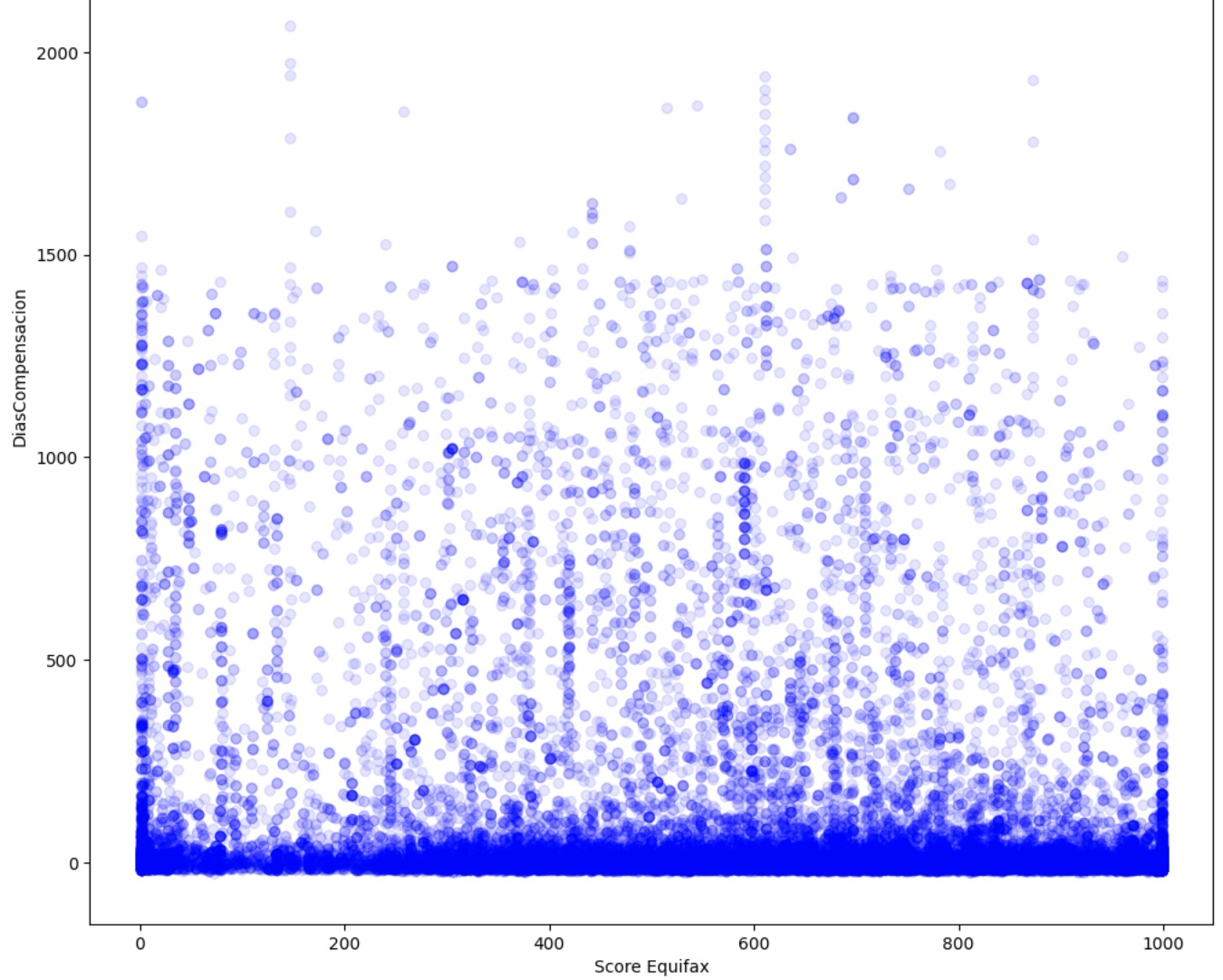


Furthermore, as the individuals in Equifax Score are mixed with natural individuals and firms, we divided the Equifax Score by types of individuals, and tried to explore if each type of individuals behave differently. The result of the grouped average within each type is as follows.

|  | **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.

**2.6.2 Score Equifax vs DiasCompensacion and Score Equifax vs ImportePesos**



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.

1. **Modeling**

**3.1 Model Selection**

**3.1. 1 Models Considered**

Initially, the team selected **Ordinary Least Squares (OLS) Regression** for its strong reputation and proven effectiveness in statistical analysis. Its capabilities for 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. The OLS method facilitated rapid hypothesis testing, allowing the team to quickly derive preliminary insights. However, as the analysis progressed, the OLS methodology encountered difficulties due to the extensive size of the dataset. The simplicity of the OLS approach, while beneficial for smaller datasets, proved to be a hindrance when scaling up to accommodate the full scope of the project’s data. This challenge necessitated a shift in strategy to a more robust solution that could handle the project's data requirements without compromising the depth of analysis.

In response to these scalability 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 advanced analytical functions tailored for handling large datasets. Its linear regression capabilities provided the necessary computational power and scalability to analyze the voluminous data efficiently. The integration of Snowflake’s built-in functions marked a significant strategic shift, enabling the team to conduct a comprehensive analysis with the robustness required by the project's expansive data demands.

We also applied **Logistic Regression** to predict whether a transaction would remain unpaid based on driver characteristics, with a dummy variable "Unpaid" indicating default payment. Although achieving an accuracy of 0.89, we observed a significant imbalance in predictions, with nearly all observations classified as "Paid" ( 9648 out of 9663 in our test set). To address this, we introduced class weights, giving greater importance to the minority class ("Unpaid" == 1). This balanced the predictions, increasing the identification of "Unpaid" observations to 1689 out of 9663, reducing overall accuracy slightly to 0.80. Our next step involves constructing a cost matrix to quantify the real costs associated with false negatives and false positives, enabling us to determine precise weights for each label.

In order to explore the relationship between the payment behavior and Equifax, we again applied Logistic Regression, choosing payment status as the dependent variable and Equifax score as the independent score. While the p-value was promising, the R-squared value was relatively small, and thus we may need to consider other models to refine the exploration.

Additionally, we attempted to predict the debt associated with default transactions using linear regression. However, the results did not yield statistically significant findings.

For the Equifax Score,

PLEASE INCLUDE ADDITIONAL MODELS USED FOR PREDICTION AND EQUIFAX

**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. This capability to handle data at scale was a critical factor in selecting Snowflake for the project, as it provided the necessary infrastructure to support the extensive data analysis required.

Another consideration was of **computational efficiency**.Snowflake's environment brought forth a significant advantage in terms of computational efficiency and performance. The platform's ability to analyze large volumes of data in a cloud-based environment was instrumental in meeting the project's demands for timely data processing. This was particularly valuable for a project where the timely analysis of data streams was integral to the overall success of the endeavor.

Achieving **methodological balance** was another pivotal aspect of the project's approach. Python's statistical libraries were highly valuable for their ease of use and depth of statistical analysis, especially in the early stages of the project. However, when it came to processing the full dataset with its complex computations, Snowflake's capabilities were more aligned with the project's needs. Despite the lack of detailed statistical outputs such as p-values in Snowflake’s regression functions, the team devised a solution to conduct parallel analyses. By comparing the coefficients and intercepts from Snowflake's analysis with those derived from Python's analysis on a more limited dataset, the team could infer the statistical significance of the results obtained from the full dataset. If the results from both platforms showed a close alignment, the p-values obtained from Python’s limited data analysis were considered reflective of what could be expected from the larger dataset analyzed in Snowflake. This dual-analysis strategy allowed the team to utilize the strengths of both platforms to achieve a comprehensive and insightful statistical analysis.

The reason for choosing logistic regression is that the payment condition is divided into categories, and as we are trying to explore the relationship between numerical variables and categorical variables, the logistic regression will be a reasonable choice. However, as noticed above, the model is not performing well in r-squared value, thus we may wish to seek other models such as tree-based models, GAM, and other methods.

PLEASE INCLUDE ADDITIONAL RATIONALE FOR PREDICTION AND EQUIFAX MODELS

**3.2. Modeling Methodology**

**3.2.1 Development Process**

In the first stage of **data preparation**, our team worked to standardize data formats and ensure consistency across the dataset, establishing a solid foundation for the analytical work to follow. This was to ensure the reliability and accuracy of the analyses that would be conducted subsequently.

Upon completing the data preparation, the team proceeded with an **initial analysis leveraging Python's `statsmodels` library**. This step was instrumental in performing regression analysis on selected data subsets, which provided preliminary insights and helped shape the design of more complex models that would later be executed within Snowflake's environment. The early-stage analysis acted as a litmus test, guiding the development of scalable models that would capitalize on Snowflake's robust capabilities.

Following the preliminary analysis, the team **integrated Snowflake’s Snowpark library** to facilitate direct data frame operations within the Snowflake ecosystem. This integration was a strategic move to enhance the team's ability to manage and analyze the full dataset. Snowflake's powerful data processing tools were fully harnessed, allowing for a highly efficient and optimized analytical workflow. The transition to utilizing Snowflake's built-in functions for data analysis represented a significant adaptation in our methodology. It required realignment of our data queries and analytical strategies to effectively exploit Snowflake's data processing strengths. This shift was not merely technical but also strategic, ensuring that the team's computational approach was aligned with Snowflake's advanced capabilities, which was crucial for managing and extracting insights from complex data sets.

The **testing and refinement stage** of the modeling process was characterized by iterative testing. The initial findings from Python's analysis were utilized as benchmarks to inform and refine the models within Snowflake. This iterative cycle was essential not only for improving the accuracy of the models but also for verifying their validity.

Finally, a **hybrid approach for model validation** was adopted. Due to Snowflake's limitations in providing detailed statistical outputs such as p-values, the team devised a strategy where the coefficients and intercepts derived from Snowflake's analysis of the full dataset were compared against those obtained from Python's analysis on limited data. This comparison was pivotal in ensuring that the significance inferred from the smaller dataset could be reasonably extrapolated to the larger dataset.

**3.2.2 Challenges and Solutions**

The process of modeling within the ambit of this project presented several challenges. One of the most significant challenges was the **efficient management and analysis of the expansive datasets** involved. The volume and complexity of the data required a solution that could offer scalability and robust computational power without compromising on the speed or accuracy of the analysis. The Snowflake platform, with its cloud-based data warehousing capabilities, rose to the occasion, providing the project with the necessary infrastructure to handle large datasets. It enabled the team to store, retrieve, and manipulate vast amounts of data effectively, addressing the critical need for scalability in data management.

Efficient data management brought forth another challenge: the need for **detailed statistical analysis within computational limits**. Python, with its extensive libraries and tools for statistical analysis, offers depth and nuance but often at the expense of scalability when dealing with large data volumes. To navigate this, the team employed a bifurcated strategy: conducting detailed statistical analyses on smaller data samples using Python, and expanding the analysis to the full dataset with Snowflake. This approach ensured that the insights gleaned from Python's intricate statistical outputs were not lost when scaling up the analysis within the more robust but less detailed Snowflake environment.

Adding to these challenges was the **complexity of adapting to Snowflake's unique coding and querying requirements**. Our team members, accustomed to the syntax and structures of traditional statistical programming languages, had to acclimate to Snowflake's particular coding paradigm. This necessitated a period of learning and adjustment, as the team became proficient in utilizing Snowflake's specific functions and commands to perform the required data analysis tasks. Despite the learning curve, the team was able to harness the full potential of Snowflake's platform, integrating its advanced data processing capabilities into the project's workflow.

**3.3. Preliminary Results and Insights**

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

As outlined in previous sections, we divided our macroeconomic data into four distinct sectors: General Economic Health, Inflation, Household Conditions, and Urban Transportation, to generate targeted insights into their varied influences on toll payment behaviors. In conducting this analysis, we employed Snowflake's built-in linear regression functions on the full dataset alongside Ordinary Least Squares (OLS) regression on a limited dataset. This dual approach enabled us to identify consistent patterns in the coefficients across different scales of data, providing a degree of reliability and depth to our findings, which have been specifically tailored to these key macroeconomic sectors.

**(1) General Economic Health**

Our analysis of general economic health indicators presented varied results. Nationally, the high p-values suggest a weak correlation with toll payment timings, indicating these general economic health metrics are not major influencers of toll payment behaviors at a broad scale. However, when we narrowed our focus to Santiago-specific economic indicators, the results showed significantly lower p-values, hinting at a more substantial relationship in this urban context. Notably, very low Durbin-Watson scores near zero were observed, suggesting the presence of time lag in the data. This could imply that the economic health of Santiago has a delayed effect on toll payment behaviors, warranting a potential exploration through time-series analysis later. Additionally, high Jarque-Bera test scores indicate that the residuals from these regressions deviate notably from a normal distribution, which calls for caution in interpreting these results and may signal the need for more complex model structures.

**(2) Inflation**

The correlation analysis between inflation indicators for overall Chile and toll payment delays resulted in very low p-values, indicating a statistically significant relationship. The Durbin-Watson scores, although less extreme than those observed in the general economic health sector, still indicate potential autocorrelation in the residuals. The Jarque-Bera test scores, while less pronounced than in the general economic health analysis, also point to non-normal distribution of residuals. This suggests that while inflation has a more pronounced impact on payment behaviors than broader economic indicators, the relationship is complex and may be influenced by other unaccounted factors.

**(3) Household Conditions**

In examining household conditions across Chile, the analysis revealed very low p-values, suggesting a strong correlation with toll payment timings. Durbin-Watson scores in this sector were similar to those observed in the inflation data, indicating potential autocorrelation. However, the household conditions data yielded particularly high Jarque-Bera test scores, which are indicative of a significant deviation from normal distribution in the model residuals. This suggests that while household conditions are a significant factor, the underlying dynamics influencing toll payments are highly complex and possibly non-linear.

**(4) Urban Transportation**

Our analysis within the urban transportation sector showed very low p-values, denoting a strong statistical significance in the relationship between transportation metrics and toll payment behaviors. The Durbin-Watson scores here mirrored those seen in the inflation data, suggesting a similar level of autocorrelation. The extremely high Jarque-Bera test scores once again pointed to the non-normal distribution of residuals, underlining the complexity of the factors influencing toll payments in the context of urban transportation.

**3.3.2 Equifax score**

Based on the current finding, the increase in equifax score will result in a higher probability of completing fee payment. However, we haven’t explored the quantitative relationship between Equifax and other variables, such as payment time, payment amount, etc.

PLEASE INCLUDE ADDITIONAL RESULTS AND INSIGHTS FOR PREDICTION AND EQUIFAX MODELS

# **4. Next Steps**

**4.1 Current Distribution of Responsibilities**

The division of responsibilities between our members is based upon our three main analyses: macroeconomic correlation analysis, prediction model development, and Equifax data analysis.

(1)Macroeconomic Correlation Analysis:

**Kris**: Her responsibilities focused on building the data ETL pipelines, 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.

(2)Prediction Model Development:

**Vidhi and Raphael** worked collaboratively on the data preprocessing and prediction model. Their responsibilities encompassed the entire lifecycle of the model development, from initial concept to testing and refinement. They applied various statistical techniques and models to predict toll payment behaviors, focusing on optimizing the model's accuracy and reliability.

(3)Equifax Data Analysis:

**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. Howard was tasked with understanding the intricacies of the Equifax data and integrating these insights into the broader analysis framework.

**4.2 Next Steps for Macroeconomic Indicator Correlations**

Beyond the indicators already under examination, there exist additional datasets pertinent to household conditions and urban transportation that we have yet to explore in depth. These datasets are notably more complex due to their intricate structures and detailed categorical distinctions. For instance, datasets detailing national disposable income, complemented by net savings, lending, and borrowing figures, could yield further insights into the financial health of households. Likewise, datasets that categorize drivers' licenses processed by variables such as class, age, gender, and region could enrich our understanding of urban transportation patterns. Incorporating these nuanced datasets could potentially unveil more comprehensive insights into the economic landscape we are studying.

Given our preliminary findings, the immediate next steps involve delving deeper into the complexities of the relationships we have identified between macroeconomic factors and toll payment behaviors. Our analysis has uncovered varying degrees of correlation across different sectors, with consistent patterns of autocorrelation (as indicated by low Durbin-Watson scores) and deviations from normal distribution in residuals (highlighted by high Jarque-Bera test scores). To address these findings, we plan to adopt more sophisticated modeling approaches and analytical methods including non-linear and time-series analysis. Kris and Michelle will spearhead this focused effort on macro-data correlation. However, the specific division of tasks remains to be determined. Further feedback and suggestions from our sponsor regarding our current analysis will guide the direction of our ongoing efforts, which may be subject to adjustments.

**4.4 Nexts Steps for Vespucio Dataset**

We will further explore the correlation between Equifax score and other demographic variables, so as to decide if Equifax score can be solely depended for effectively predicting the payment behavior. Meanwhile, we will further seek alternative models to refine the prediction result of models related to Equifax score.

**4.4 Next Steps for Presentations**

As we advance into the next phase of our project, a significant focus will be on preparing for two critical presentations: the Mid-Term presentation with our sponsor scheduled for April 3rd, and the final presentation planned for the week of May 6th.

For the **Mid-Term presentation**, our team will engage in a collaborative effort to develop a comprehensive report and accompanying slides, with each team member focusing on the results from their specific area of analysis. Vidhi and Raphael will concentrate on the prediction models, Kris and Michelle will handle the macroeconomic correlation aspects, and Howard will delve into the findings from the Equifax data. This presentation will primarily spotlight our correlation findings, offering our sponsor a detailed view of how macroeconomic factors interplay with toll payment behaviors.

In preparation for this, team members will be responsible for crafting the sections of the report and slides related to their analysis. This task will involve not only collating and summarizing the data but also ensuring that the insights are presented in a clear, concise, and impactful manner. Given the importance of this presentation in communicating our progress and findings to our sponsor, we aim to have the initial drafts of the report and slides ready for internal review by March 28th, allowing for a week of revisions and refinements.

Looking ahead to the **final presentation**, the objective will be to present a more rounded and complete view of our project's outcomes. This presentation will encompass a detailed report and slides showcasing our findings, including the model built to predict the likelihood of payment defaults, timing of payments, and the most effective collection approaches. Additionally, it will feature an expanded analysis of the correlation between macroeconomic data and payment behaviors.

Each team member will again be tasked with preparing the sections of the report and slides pertinent to their respective analyses. This approach ensures that each segment of the presentation is informed by deep, specialized knowledge and insights.