**Group 6 -** Tyler Clyde, Sean King, Austin Liu, Nicholas Seah, and Kabir Wagle

# Project Phase 3: Muni Ridership Analysis.

# Introduction:

A strong transit system is essential for getting people around amidst the hustle and chaos of a major city. San Francisco is often overlooked when compared to the great public transport cities of the earth such as London and New York. SF has a large intercity transit network of its own referred to as the "Muni". The San Francisco Municipal Railway (Muni) has a special place amongst public transportation in American history as it is the first publicly-owned and operated transit agency in a major city. The Muni consists of a network of hybrid bus lines, cable cars, streetcars, electric trolleys, and seven light rail lines which connect the various neighborhoods of the city. While the city does have a large public transit network, many would argue that it is being underutilized by the population in favor of automobiles. A clear example of this can be seen when you look at the traffic issues that plague the city, with San Francisco having the 7th worst traffic congestion of any city in America. Given this underutilization, understanding data behind rider behavior and service utilization is essential in order to increase ridership and make public transportation a better option. By using information from large datasets like the San Francisco City Survey and detailed information on Muni stops, this report aims to understand ridership demographics and gain a better understanding of who is using the Muni. In doing this one can gain insights into what can be done in the future to increase ridership on the network.

By applying analytical models to our data we seek to answer the following questions:

1. Who are the frequent users of Muni, and what distinguishes them from less frequent riders?
2. How can underutilized segments be effectively targeted through strategic marketing?
3. Is there a link between the density of Muni stops and rider frequency?

By answering these questions we hope to provide meaningful insights for the Society for Metropolitan Analysis and Research of Transportation (SMART) regarding how it can make changes to make the Muni a better system.

<u>Supporting Articles</u>

Link: https://sf.gov/reports/april-2023/2023-city-survey-results
This article summarizes the results of the City survey. According to rider respondents, the average rating of the Muni was a B-, slight improvement since 2019. Trends show steady decline among infrequent riders since 2015.

# Data Collection and Summary:

## Dataset 1: Survey Responses

The first data source is survey data from San Francisco City resident respondents. The data was provided and curated by DataSF. The City Survey has been conducted annually, starting in 1996 and changing to biennial since 2005 (2021 missed due to COVID-19 pandemic). The survey asks city residents to rate their personal level of usage and satisfaction with essential city services and infrastructure. This includes things like the public library, public safety, energy, water, street cleanliness, and the most important for our purposes, the San Francisco Municipal Railway. In previous years, data was collected using mail-in forms and phone surveys. In the most recent data collection in spring of 2023, the city expanded to provide in-person and online options. Due to this inconsistent survey methodology, the data summary notes that comparing records from previous years should be interpreted with caution.

Our team decided on this data source for two main reasons:

- First, the dataset has a rich amount of variables covering different areas of interest.
- Secondly, the dataset was well organized and came from a reputable source.

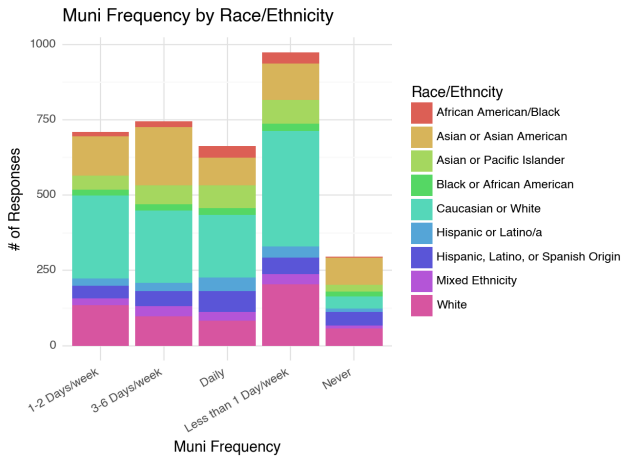


Figure 1: Muni Frequency by Race/Ethnicity

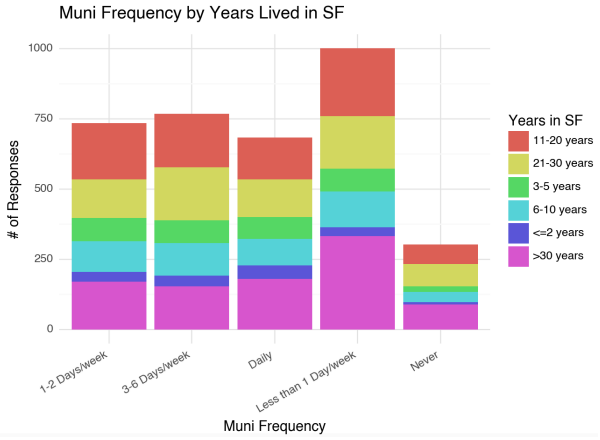


Figure 2: Muni Frequency by Years in SF

Figures 1 and 2 are both stacked column charts that demonstrate how survey responses for Muni Frequency can vary depending on their demographics (For illustrative purposes Races with < 100 responses were omitted). We often used contingency tables in order to gauge the relationship between Muni Frequency, and the numerous demographic variables we had at our disposal. Contingency tables are a great way to analyze the relationship between categorical variables, especially those with multiple levels. Contingency tables can easily be transformed into stacked column charts, just like the ones seen above.

## Dataset 2: Muni Stops

The second data source is information gathered by the city government detailing the location and feature data of all Muni stops. This dataset was also provided by DataSF. The purpose of the dataset is to gather data on all stops within the San Francisco Municipal

Transportation Agency system. This dataset was first created in November 2020 and most recently updated in December 2023, ensuring that the information is up to date. Stop data is continually updated from the "Trapeze" scheduling system and other critical info is updated every 3-6 months. When looking for a second dataset, we wanted to add an additional layer to our demographic analysis incorporating data regarding the presence of the Muni in different areas. The only geographic identifier present in dataset 1 (Survey Responses) was the BOS zoning code, an in-house supervisor district identifier that ranges from 1 - 11. Luckily, because the dataset 2 (Muni Stops) was also provided by the SF government, it included the BOS code identifier. This will allow us to merge the two datasets on the BOS code.

Our team decided on this data source for two main reasons:

- Firstly, the dataset contains information on the number of Muni stops in San Francisco.
- Secondly, the dataset contains the BOS zoning code identifier to merge datasets.

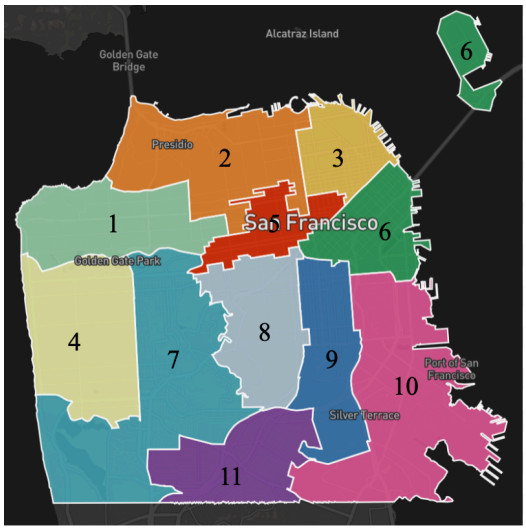


Figure 3: Map of BOS Districts in SF

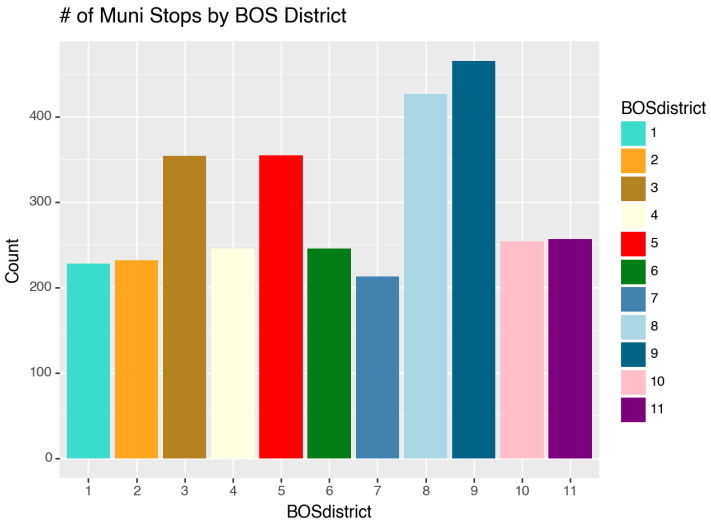


Figure 4: Frequency of Muni Stops by BOS District

Figures 3 and 4 illustrate the BOS codes used by the San Francisco government to segment the city. On the left is a map of the foggy city, with the BOS code overlaid and color coded for reference. On the right is the frequency of Muni stops by the BOS district, this allows us to identify the presence of Muni stops across the city. The benefit of this statistic is that it can be easily merged with dataset 1 (Survey Responses). This will allow us to measure the relationship between the Muni frequency and Muni stops.

# Model Procedure and Results:

## Research Question 1:

Our first research goal is to correctly classify Muni riders' frequency based on their demographic information in order to better market the less frequent riders. While LDA and QDA models were tested, a Decision Tree model was used because it performed the best in terms of

correctly identifying rider frequencies amongst all the other models, with an accuracy of 0.306. This means that when tested with the original data, the model classified 30.6% of the riders into the correct frequency category. Since there were five categories, if we were to randomly classify the respondents into frequencies, we would expect to correctly predict 20% of them, meaning our model did better than random selection. It was specified for the decision tree to have at least 20 observations per node, and that it does not grow beyond 4 levels to optimize the accuracy. The results from the model are shown below.

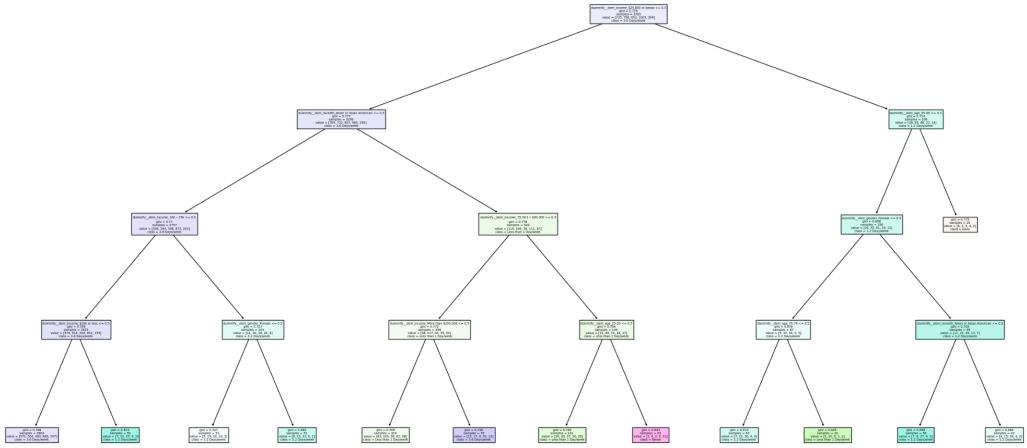

Figure 5: Multiclass Decision Tree

The resulting Decision Tree model provides us with a vast amount of information in terms of classifying riders into frequency categories. Traditionally, to navigate the decision tree, we consider the condition present in each node and move left if the statement is true for a respondent, and right otherwise. However, since the model included binary variables, decisions based on the nodes are now flipped. In other words, if the condition within each node is met, navigate right, otherwise go left.

The first split in the decision tree refers to whether or not a respondent's income fell below $25,000. Being the root node, this tells us that the $25,000 income threshold is a significant predictor of ridership frequency. If we descend down to the far left of the tree, we notice that the decision tree classifies riders into the “3-6 Days/week” frequency category. This part of the tree also corresponds to riders with income levels $10,000 or less, suggesting that those who fall in the lower income group tend to ride the Muni more frequently. This could be for a variety of factors, one being the increased cost of owning a car in the city of San Francisco. Furthermore, notice that the “Never” class appears for those who fall within the $75,000 to $100,000 income level between the ages 25 and 29. We can classify this group as higher-income Millennials who would probably prefer other forms of transportation including rideshare options like Lyft and Uber.

It is important to note the node classified as “Daily” which includes respondents between the ages of 65 and 69 years old on the right side of the tree. Initially, this seems counterintuitive; however upon research, this makes sense as the city of San Francisco incorporated the “Free

Muni for Seniors (Ages 65+)" program, which allows for riders within the 65 to 69 year old range to ride the MUNI for free.

## Research Question 2:

Our second research question asks us to identify the demographics of people that do and do not frequent the Muni so that appropriate marketing campaigns could be implemented to increase ridership. Frequent riders that consistently come back for the transportation service is a key performance indicator to the success of the public transportation system. Recurring riders allow us to gauge stability, reliability, efficiency, and quality of the transportation service.

To do this analysis, we created a new binary target variable that separates frequent (1-2 Days/week, 3-6 Days/week, Daily) vs infrequent riders (Never, Less than 1 Day/week). SVC and SVM models were considered, however Logistic Regression was selected as it had the highest Receiver Operating Curve, Area Under the Curve (ROC-AUC) score. The ROC-AUC metric is a single value that summarizes the overall performance of the model across various threshold settings ranging from 0 to 1, with higher scores indicating better model performance. The Logistic Regression model had an ROC-AUC score of 0.58, which indicates that the model is performing better than random chance. The results for select predictors from the Logistic Regression model are shown in the table below.

| **Predictor** | **Estimate** |
|---|---|
| Age between 25-29 years old | 0.019542 |
| Income less than $25,000 | 0.586981 |
| Household Size | -0.001744 |

Figure 6: Logistic Regression Output

From the Logistic Regression results, the following interpretations can be made:

- The odds of frequenting Muni for riders between 25-29 years old is 1.97% higher than for those 18-24 years old.
- The odds of frequenting Muni for riders with income less than $25,000 is 79.86% higher than for those with an income between $100,000-$200,000.
- The odds of frequenting Muni decrease by 0.17% for each one person increase in household size.

## Research Question 3:

Our third research question is to investigate the relationship between the number of Municipal stops and the frequency of riders. Our team decided that the best course of action was to repeat the analysis with the same 5 feature sets from Research Question 1. The only difference being that we would extend the models to include the effect of Stops. This decision was made to eliminate potential omitted variable bias and possible over or underestimation of the effect of

stops to classify riders by frequency. After tuning the minimum number of samples per leaf node and the max depth of the decision tree, we arrived at the same subset having the best performance of overall accuracy (26%).

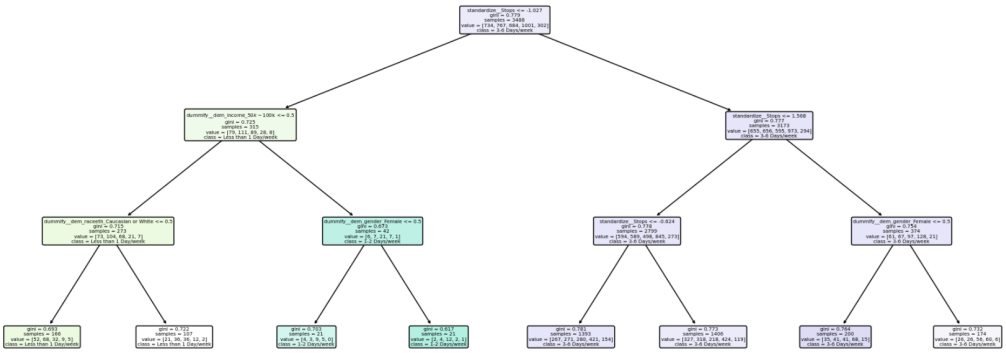

Figure 7: Decision Tree with Muni Stops

As we can see from the decision tree, the density of stops is a significant predictor as it is the variable in the root node. The number of stops is scaled to not overexert influence on the decisions and we lose some interpretability on the precise value. We can tell if an individual's board of supervisor district has more stops, it leads the classifier to classify cases into high frequency riders on the right half of the tree. On the other hand, with fewer stops in an individual's district, riders are more likely to be low frequency riders.

# Conclusion and Recommendations:

From the final models, we can categorize the demographics of individuals and their ridership frequency to determine possible underutilized segments. Based on the models, we can predict low frequency individuals based on demographics such as income, household size, age, and BOS district. The models indicate that low frequency individuals typically have an income of over $75,000, have a large household size, are between the ages of 18 and 30, and are located in a BOS district where the Muni stops are less dense versus the other districts. Since the odds of frequenting Muni decrease by 0.17% for each one person increase in household size, a larger household size indicates less frequency. With these estimates in mind, individuals that fit into these demographics should be the main targets for strategic marketing.

The goal of the marketing campaign is to increase the use frequency of the Muni. Understanding ages from 18 to 30 do not use the Muni very frequently, marketing over social media should be heavily considered since these ages typically spend an ample amount of time online. Also, it is predicted that many individuals with incomes over $75,000 do not frequent the Muni due to the convenience of more expensive options such as rideshares or commuting in a personal vehicle. For both of these demographics, a campaign that focuses on the efficiency of the Muni may help increase the ridership of these individuals. This can be done by showing a comparison of the time it takes to transport using the Muni versus the time it takes to commute by car during rush hour period. The goal would be to show how the Muni can save time and money for people who transport to work. Another display would be to show the ease of use of

the Muni system. Many people who are more accustomed to rideshare apps may choose to use them over the Muni based on their comfortability. By showing ways to use the Muni system and its ease of use may help these individuals see the Muni as a viable transportation option.

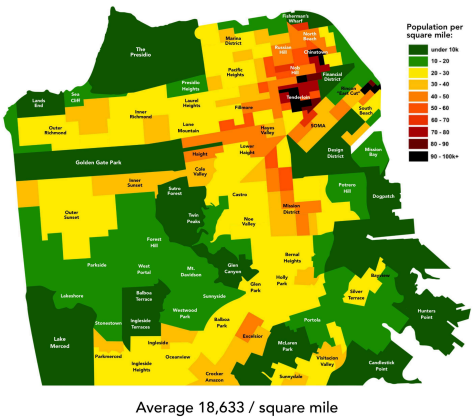


Figure 8: San Francisco Density Map

Another aspect SMART may want to consider is the cost benefit analysis of adding more transit stops in certain BOS districts. When comparing the density map of San Francisco to the BOS district locations, we can see some areas where an increase in stops may be beneficial. Based on the Muni's 2023 consensus, only 53% of individuals believe that the Muni does a good job serving the entire community. Districts to analyze include districts 1, 6, and 11. These districts show instances of highly populated areas, however, have a similar number of total Muni stops as districts with a lower population density. By possibly increasing the stops in these three districts, there may be an increase in both transit efficiency and total use frequency. The next step would be to conduct a cost analysis to see if the cost of increasing stops in these districts outweighs the possible benefits.

The overall goal of our analysis was to find trends between rider demographics and their Muni frequency and use them to determine ways to increase ridership and efficiency. Based on the conducted models, we observed that higher income and younger individuals do not frequently use the Muni. Also, individuals who live in a BOS district with a lower amount of Muni stops will have a harder time accessing the Muni and thus choose other transportation options. Between marketing campaigns and the analysis of adding more Muni stops, we believe these are viable options SMART should consider to possibly increase the ridership frequency of the less frequent riders.