**Analyzing the Relationship Between Crime and Use of BART’s Public Transportation in San Francisco, CA**

**October 20th, 2023**

**Contributors & Team Roles:**

Kaden Buckley: Facilitator & Critic

Brogan Pietrocini: Gatekeeper & Notetaker

Ollie Sheehan: Coordinator & Compromiser

Ryan Perez: Coach & Arbitrator

Mehrpad Moattari: Evaluator & Initiator





**Table of Contents**

1. **Data**
2. **Research Questions**
3. **Ethics**

**I. Key Data Sources**

| **Primary Data** | **Links** |
| --- | --- |
| BART Ridership Reports (Entries and Exit per station) | [Ridership Reports | Bay Area Rapid Transit](https://www.bart.gov/about/reports/ridership) |
| Crime Rate per Police District in San Francisco | [San Francisco Crime Classification | Kaggle](https://www.kaggle.com/datasets/kaggle/san-francisco-crime-classification) |

| **Supplemental Data** | **Links** |
| --- | --- |
| BART Station-Specific Demographics | [BART Station Profiles](https://www.bart.gov/sites/default/files/docs/StationProfile2015_HomeOriginOnePagers_rev0629.pdf) |
| Public Transport Crime Statistics | 1. [BART Safety Compared To Other Transit Systems](https://bayareapunk.com/blog/transit-crime.html) 2. [BART Police Data](https://www.bart.gov/sites/default/files/docs/2023-01%20Monthly%20Chief%27s%20Report%20CRB%2003012023.pdf) |
| Bay Area Council BART Poll | [BART Poll](https://acrobat.adobe.com/link/review?uri=urn%3Aaaid%3Ascds%3AUS%3A0e107c3e-9903-3198-b075-ffdfc6749268) |

**Table 1: Sources of Data & Other Resources**

The data containing BART Entries and Exits per station is directly from the BART (Bay Area Rapid Transit) website. BART is a public transport system serving the San Francisco Bay Area. BART connects suburbs, urban centers, and airports. BART continually collects data on entries and exits for an array of reasons Some of which include the following: capacity planning and traffic management, infrastructure and maintenance decisions, safety preparedness, financial planning, fare pricing strategies, identifying rider patterns, and reporting and accountability reasons.

The crime rate per police district in San Francisco data is sourced from Kaggle. Kaggle is a platform for data science competitions and an online community for data scientists and machine learning practitioners. The data was collected for two reasons: to predict the category of crime that occurred, given the time and location, and to visualize the city and crimes.

The BART Entries and Exit per station data are collected with fare gates, automated passenger counters, ticket sales and reload data, surveys, special event data collections, integration with other systems, and data analysis and processing.

The crime rate per police district in San Francisco data was featured in Kaggle’s playground competition titled “San Francisco Crime Classification.” The original dataset is from SF OpenData, the central clearinghouse for data published by the City and County of San Francisco.

The cases in the BART Entries and Exit per station data are several statistics relating to ridership on BART over a 12-month period (October 2022 to September 2023). The statistics are grouped by monthly ridership, daily ridership, weekday ridership by time period, trip market as percentage of total exits, sub-region as percentage of total exits, and clipper metrics.

The cases in the crime rate per police district in San Francisco data are individual crimes reported in San Francisco with dates, days of the week, addresses, districts, and location coordinates.

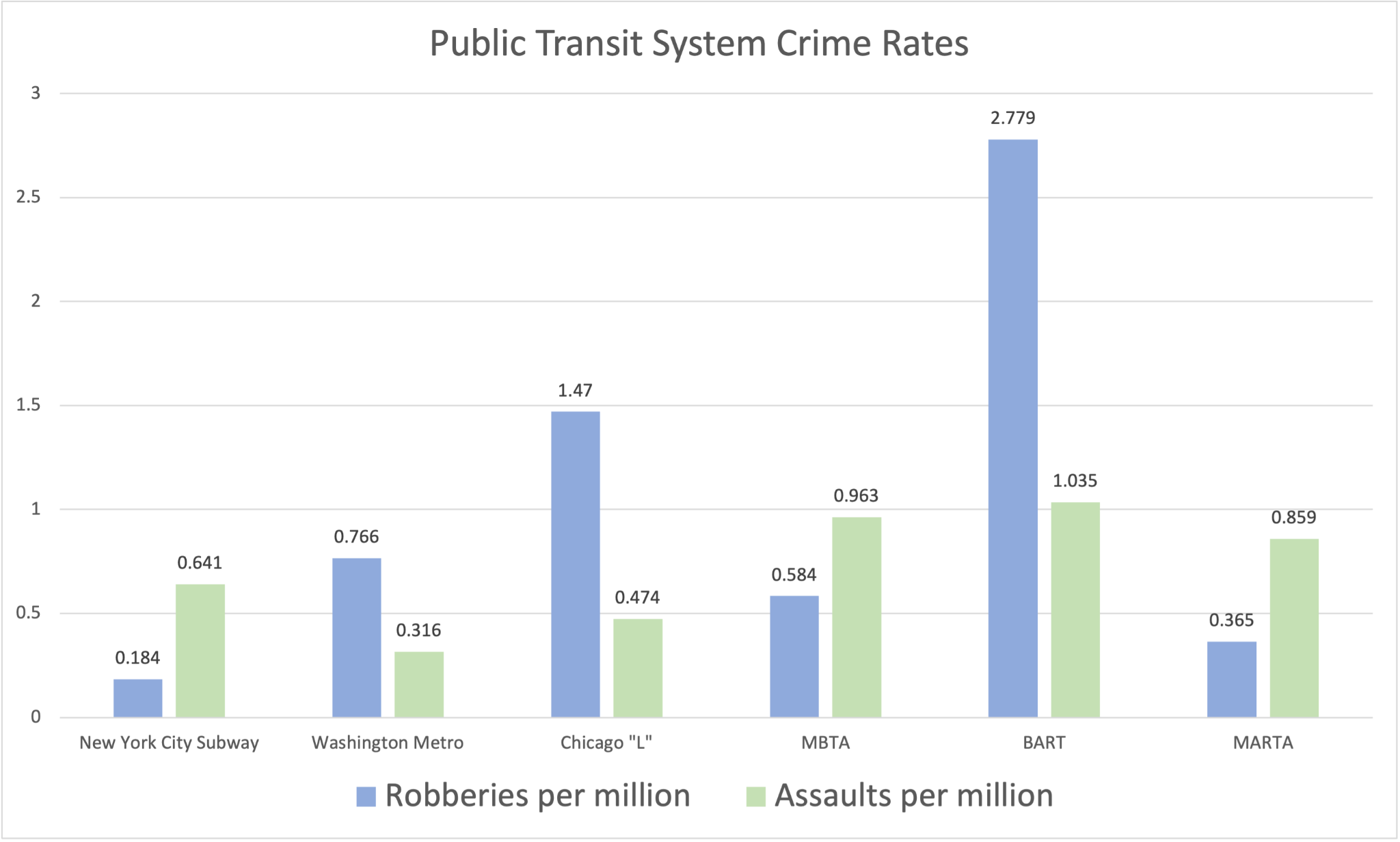
Our primary target variable will be entries per station and exits per station. Entries per station is the total number of individual riders that enter each BART station by month. Likewise, exits are the amount of times someone exits a BART station. These are a quantifiable measure of how many times someone uses each individual BART station.

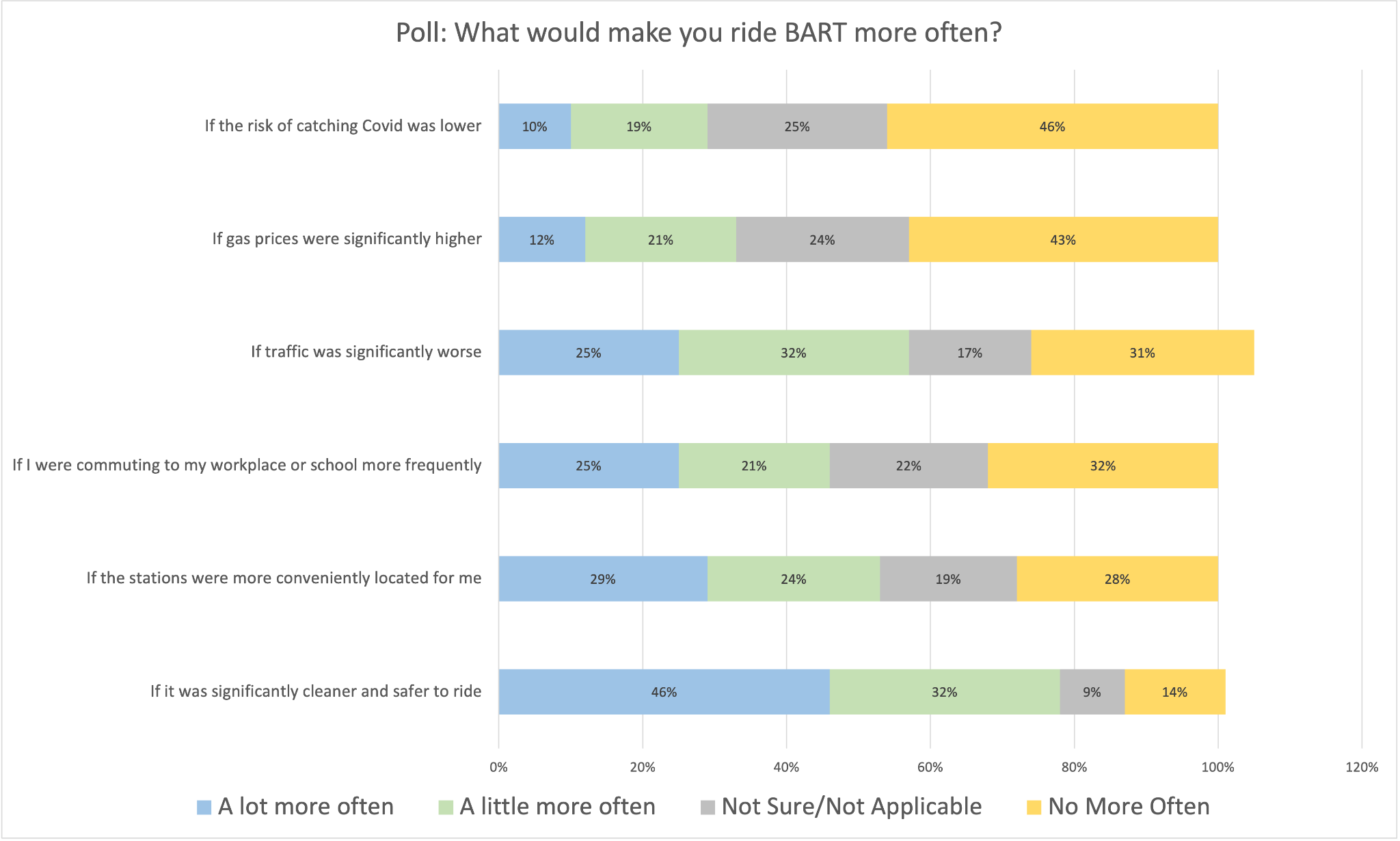
The first predictor variable of importance is the amount of crime in each BART station district. This would be a count of total crimes by district in San Francisco. We also want to analyze other potentially significant factors such as a district’s average household income and population to supplement our main dataset. We would scale the predictor variables based on population to correctly distinguish the relationship between crime and the use of various BART stations. In our analysis, we can determine the most important variables in modeling public transportation use to ultimately improve ridership in the environmental mission to reduce carbon emissions and positively influence society.

Entries per station are documented on a monthly basis in BART Ridership Reports. Each month is located on separate data frames, and therefore we will need to join the monthly data into one cohesive data frame. Entries per station are also documented in wide format and it will need to be converted to a tidy data format for proper analysis. Additionally, Public Transport Crime Statistics and Bart Police reports contain important data that is not available for download in a structured format. Therefore, we will need to either web scrape or manually convert this information into a structured format that can be analyzed. We have already done so for the purpose of this proposal. A summary of BART Ridership Reports can seen below in Table 2. Visualizations for Crime Statistics and BART Polls can be seen in Figure 1 and Figure 2.

| **Station Name** | **Sum of Entries** |
| --- | --- |
| **12** | **4907** |
| **16** | **5466** |
| **19** | **4601** |
| **24** | **5293** |
| **AN** | **1860** |
| **AS** | **1990** |
| **BE** | **1277** |
| **BF** | **2475** |
| **BK** | **4914** |
| **BP** | **4181** |
| **CC** | **9121** |
| **CL** | **2500** |
| **CM** | **1519** |

**Table 2: BART Ridership Report Summary (September, 2023)**

**Figure 1: Public Transit System Crime Rates**

**Figure 2: BART User Poll Results**

**II. Key Research Questions**

**1.) What major questions do you expect to be able to answer with this data?**

a. Is there a relationship between crime rates and public transportation usage?

b. Are there temporal patterns in crime rates relative to BART ridership?

c. How do other socioeconomic factors, such as average household income or demographic patterns, interact with crime rates and BART usage?

**2.) How will answers to those questions help the client make actionable decisions?**

**a. Resource Allocation and Security Planning:**

If certain types of crimes are identified as having a significant negative impact on ridership, BART authorities could allocate more resources towards preventing these crimes, perhaps through increased security measures, community outreach programs, or collaboration with local law enforcement.

**b. Strategic Development and Urban Planning:**

Understanding the relationship between socioeconomic factors and ridership could inform future urban development, ensuring that public transportation is accessible and appealing to a broad demographic, potentially encouraging a more equitable urban layout.

**c. Targeted Communication and Community Engagement:**

If specific areas are suffering from public transportation underutilization due to perceived or actual security concerns, targeted communication efforts could be initiated to address crime and safety issues, denounce any possible misconceptions, enhance rider safety education, and engage with the community to build trust.

**d. Policy Advocacy:**

Solid, data-backed evidence about the factors influencing public transportation usage can be a powerful tool in advocating for policy changes at higher governmental levels, potentially influencing broader initiatives related to urban safety and public transportation funding.

**3.) How will you validate that your answers to the questions are correct?**

**a. Statistical Rigor:**

Applying appropriate statistical analyses and tests to quantify the confidence in your results (e.g., correlation coefficients, significance testing, regression analysis). This involves testing for potential confounding variables that could affect the relationships you're studying.

**b. Scenario Analysis:**

Create hypothetical scenarios (e.g., a significant decrease or increase in crime rates) and use your model to predict their impact on BART usage. Comparing these predictions with stakeholders' expectations or historical data can reveal whether your model’s responses are plausible and consistent with real-world expectations.

**c. Error Analysis:**

Deliberately look for cases where your model fails or is less accurate and try to understand why. This process can uncover biases in the data, areas where your model might be overfitting, or other issues that you need to address to improve its accuracy.

**d. Sensitivity Analysis:**

Testing how your results respond to changes in the input data or assumptions (e.g., different time frames, excluding certain types of crimes) helps understand the robustness of your findings.

**III. Key Ethical Considerations**

Commuters and residents in the Bay Area who use BART for transportation will benefit from this analysis. In addition, local businesses in areas where BART stations are located may experience higher foot traffic. BART and other local agencies will benefit from improved public transportation and lowered crime rates as well.

One concern with this approach is the potential for communities associated with higher crime rates to be stigmatized or discriminated against. This could also lead to accelerated gentrification in different areas where BART usage might increase.

As always, the sources of our data need to be carefully considered, and we must acknowledge any potential biases in how it may have been collected. In particular, crime reporting can sometimes suffer from underreporting or misrepresentation, leading to skewed statistics on crime rates. Furthermore, if there is sampling bias that focuses on certain areas or demographics, it could lead to a misrepresentation of crime and transportation patterns. For example, we might assume that a particular station with high local crime rates is not operating at its fullest potential when in reality it is the main mode of transportation for residents in that area.

As we are trying to make predictions about BART usage based on crime rates, we want to be certain that we are not sharing personally identifiable information of anyone involved in reported crimes or those who are using BART. This would include names, specific addresses or locations, and any demographic information that could lead to the identification of individuals, especially in smaller communities.

As mentioned earlier, our main focus will be to ensure that our analysis does not reinforce stereotypes or prejudices against certain groups or communities. We need to prioritize transparency in the data collection process, as well as with how we analyze and report on this information. This, again, requires the protection of privacy for anyone in the dataset. In addition to our data collection process, we need to consider the impacts of our proposed interventions and avoid magnifying any inequalities that may already exist.