Paige Hodgkinson
San Francisco State University
Math 448-01 Statistical Learning and Data Mining
Final Project
Prof Tao He

Abstract

This project investigates the relationship between bikeway infrastructure, bike commuting rates, and traffic safety in San Francisco, aiming to identify specific infrastructure features impacting crash occurrence. Utilizing integrated data from the US Census Bureau, SFMTA, and UC Berkeley TIMS, the analysis moves beyond simple correlations by considering bike commuter volume and socioeconomic factors like median income. Ordinary Least Squares and Ridge Regression models were employed to analyze the impact of various bikeway features on the number of crashes. Key findings from the Ridge Regression model indicate a strong positive association between intersection sharrows and increased crashes, particularly in areas with higher bike commuter volume, as further supported by OLS interaction analysis. Conversely, crossbikes showed an association with reduced crashes. While other features presented challenges due to multicollinearity, the results suggest that certain infrastructure designs have distinct safety implications. Recommendations include discouraging new intersection sharrows and exploring the safety benefits of crossbikes. Future work will focus on addressing data limitations and refining the analysis with more granular temporal and spatial data.

I. Introduction

As a frequent bicycle commuter and a former long-distance cyclist, I am curious about what data says about San Francisco's bicycle infrastructure decisions. This project investigates the complex relationship between bikeway infrastructure, bike commuting rates, and traffic safety in San Francisco, with a focus on identifying specific bikeway features that may increase or decrease crash occurrence. Recognizing the limitations of simple correlations, this analysis delves into the influence of bike commuter volume and socioeconomic factors, like median income, on these relationships. By analyzing cleaned and integrated data from the US Census Bureau, SFMTA, and UC Berkeley TIMS, the project seeks to move beyond general assessments of bikeway impact and provide nuanced insights into the safety implications of individual infrastructure designs. Ultimately, the goal is to inform evidence-based recommendations for bikeway planning and design that prioritize cyclist safety and mitigate potential risks associated with specific infrastructure types.

II. Data Sources and Cleaning

A. Data Sources

This project leverages four primary datasets.These datasets were sourced from the US Census Bureau data portal, data.sfgov data portal, and UC Berkeley TIMS system data filter exporter.

- The American Community Survey (ACS) 5-Year 2023 Means of Transport dataset provides estimates of commuting methods, including bike commuting, for all census tracts in San Francisco.
- The American Community Survey (ACS) 5-Year 2023 Geographic Mobility by Selected Characteristics in the United States gives the median income by census tract for San Francisco.
- The SFMTA Bikeway Network Point Features dataset contains detailed information about the locations and features of bikeways within San Francisco.
- The TIMS Berkeley dataset records information about traffic collisions in San Francisco, offering valuable insights into traffic safety dynamics.

-

B. Description of methodology for the loading and cleaning of the datasets.

1. ACS 5-Year 2025 Means of Transport Loading/Cleaning:

- **Loads ACS Data:** Reads a CSV file named `ACS_5Year_2023_Means_of_Transport.csv` into a pandas DataFrame.

- **Filters for Relevant Data:** Selects the 'Label (Grouping)' column and columns containing 'Estimate' (excluding 'Margin of Error'). Then, it specifically keeps only the rows related to 'Total:' commuters and 'Bicycle' commuters.
- **Calculates Percentage of Bike Commuters:** Converts relevant data to numeric, handles potential commas in numbers, and calculates the percentage of bike commuters for each census tract.
- **Extracts and Converts Census Tract GEOIDs:** Defines a function to extract the census tract number from the column names and convert it into a standardized 11-digit GEOID, which is then applied to rename the columns.

2. Crashes Loading/Cleaning:

- **Loads Crash Data:** Reads a CSV file named `Crashes.csv` into a pandas DataFrame.
- **Selects Relevant Columns:** Keeps a specific subset of columns related to crash details like time, location, severity, and involvement of alcohol.
- **Handles Missing Coordinate Data:** Addresses missing latitude and longitude values by attempting to fill them based on corresponding 'POINT_X' and 'POINT_Y' columns. For remaining missing coordinate pairs, it removes the original 'LATITUDE' and 'LONGITUDE' columns, renames 'POINT_X' and 'POINT_Y', and then manually imputes coordinates for a small number of specific rows using external information (like Google Maps).
- **Retrieves and Assigns Census Tracts:** Uses an external Census geocoder API to determine the census tract for each crash based on its latitude and longitude coordinates and adds this information to a new 'CENSUS_TRACT' column.

3. Bike Features Loading/Cleaning:

- **Loads Bikeway Data:** Reads a CSV file named `SFMTA_Bikeway_2025.csv` into a pandas DataFrame.
- **Extracts Latitude and Longitude:** Defines a function using regular expressions to extract latitude and longitude coordinates from a 'shape' column (likely containing Well-Known Text (WKT) geometry) and creates new 'LONGITUDE' and 'LATITUDE' columns.
- **Retrieves and Assigns Census Tracts:** Similar to the crashes data, it uses the Census geocoder API to determine the census tract for each bikeway segment based on its extracted latitude and longitude and adds this to a 'CENSUS_TRACT' column.
- **Saves Cleaned Data:** Saves the DataFrame with the added 'CENSUS_TRACT' information to a new CSV file named `Bikeways_cleaned.csv`.

4. ACS 5-year 2023 Median Income Loading/Cleaning:

- **Loads Income Data:** Reads a CSV file named `ACSST5Y2023.S0701.income.csv` into a pandas DataFrame.

- **Selects Median Income Column:** Identifies and selects the column containing the median individual income estimate.
- **Renames Columns and Removes Header Row:** Renames the first column to 'Census Tract' and the median income column to 'Median Income', and then removes the original header row.
- **Formats Census Tract IDs:** Defines a function to extract and format the census tract identifiers from the 'Census Tract' column to a consistent 11-digit format.

C. Description methodology for combining the different datasets. The process involves several key stages: data acquisition and initial loading, spatial unit standardization, feature engineering and integration, data cleaning and transformation, and final output.

Phase 1: Data Acquisition and Initial Loading

1. **Load Data :**
    - `ACS_cleaned.csv`: Contains demographic data from the American Community Survey, aggregated at the census tract level.
    - `Bikeways_cleaned.csv`: Contains information about bikeway infrastructure, including type and location.
    - `Crashes_cleaned.csv`: From TIMS data portal for traffic crash incidents and location in San Francisco.
    - `Income_cleaned.csv`: Provides median income data at census tract level.

Phase 2: Spatial Unit Standardization and Core DataFrame Creation

1. **Identification of Common Spatial Unit:** Use 'CENSUS_TRACT' as the primary common geographic identifier across the bikeway and crash datasets.
2. **Creation of a Unified Index:** Generate a comprehensive list of all unique census tract values present in the bikeway, crash, and ACS datasets.
3. **Establishment of a Base DataFrame:** Create a new pandas DataFrame (`combined_df`) with a single column containing the unified list of 'CENSUS_TRACT' values. This DataFrame serves as the foundation for integrating information from the other datasets.

Phase 3: Feature Engineering and Integration

1. **Incorporating Demographic Features:** Handle potential variations in column naming (e.g., combining 'CROSSBIKE' and 'CROSS BIKE').
2. **Mapping ACS Data to Census Tracts:** Transpose the `ACS_cleaned` DataFrame to align census tract identifiers as a column. Then, iterate through the `combined_df` and populate the demographic feature columns with the

corresponding values from the transposed ACS data based on matching 'CENSUS_TRACT' values.

3. **Integrating Crash Data:**
    - Initialize a 'Num_Crashes' column in `combined_df` to store the count of crashes within each census tract.
    - Iterate through the `Crashes_cleaned` DataFrame. For each crash record, identify the corresponding census tract in `combined_df` and increment the 'Num_Crashes' count for that tract.
4. **Integrating Bikeway Data:**
    - Select relevant columns ('DESCRIPT', 'COUNT', 'CENSUS_TRACT') from `Bikeways_cleaned`.
    - Handle missing descriptions by removing the corresponding row.
    - Categorize the different types of bikeway infrastructure ('DESCRIPT') by assigning a numerical identifier to each unique type.
    - Create new columns in `combined_df` for each bikeway category, initialized to 0.
    - Iterate through the categorized `Bikeways_cleaned` data and, for each record, identify the corresponding census tract and bikeway category in `combined_df`, then add the associated 'COUNT' to the respective bikeway category column.
    - Calculate the total count of bikeway infrastructure features within each census tract by summing the values across all the individual bikeway category columns and storing the result in the 'COUNT' column of `combined_df`.
5. **Creating a Relative Crash Metric:** Engineer a new feature, 'Relative_Crashes', by calculating the ratio of the number of crashes to the percentage of bike commuters within each census tract. Handle cases where the percentage of bike commuters is zero to avoid division by zero errors.
6. **Incorporating Income Data:**
    - Standardize the format of the census tract identifiers in both `combined_df` and `Income_cleaned` (ensuring they are strings, removing extra spaces, and padding with leading zeros for consistent length).
    - Merge `combined_df` with `Income_cleaned` based on the standardized census tract columns to add the 'Median Income' to the integrated dataset.
    - Remove the redundant census tract column from the merged DataFrame.

Phase 4: Data Cleaning and Transformation

1. **Handling Missing Demographic Data:** Remove any census tracts from `combined_df` that have missing values for the core demographic features (identified by NaN in the 'Total:' column after the ACS data integration).

2. **Handling Missing Values for Normalization:** Replace any non-numeric indicators (e.g., '-') with actual missing value representations (NaN) to enable numerical processing.
3. **Data Type Conversion:** Convert the relevant columns in `combined_df` to a numeric data type to prepare them for scaling.
4. **Data Scaling:** Apply Min-Max scaling to the numerical features (excluding the 'CENSUS_TRACT' identifier). This transformation scales the values of each feature to a fixed range (typically 0 to 1), which can be beneficial for various analytical techniques.
5. **Identifying Remaining Missing Values:** After the integration and transformation steps, identify and display any rows that still contain missing values to assess the completeness of the final dataset.

III. Data Visualization

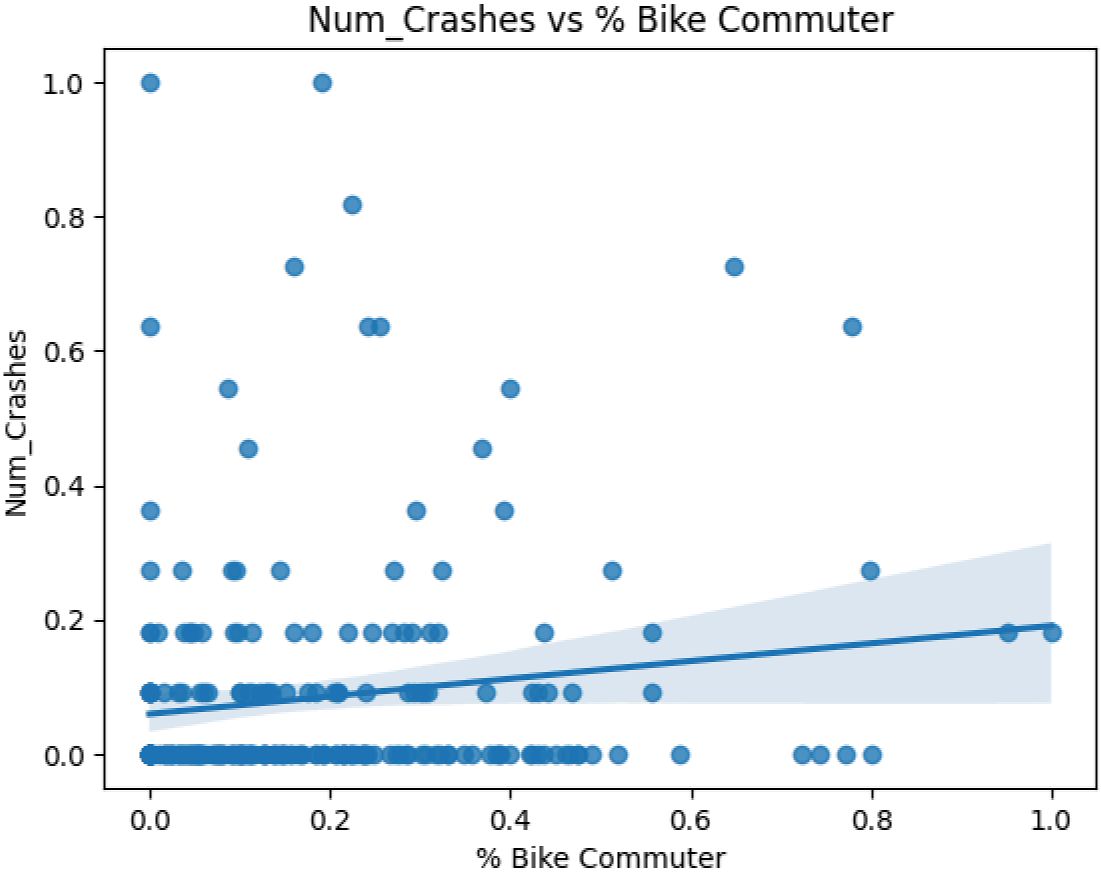
A. Linear Approaches

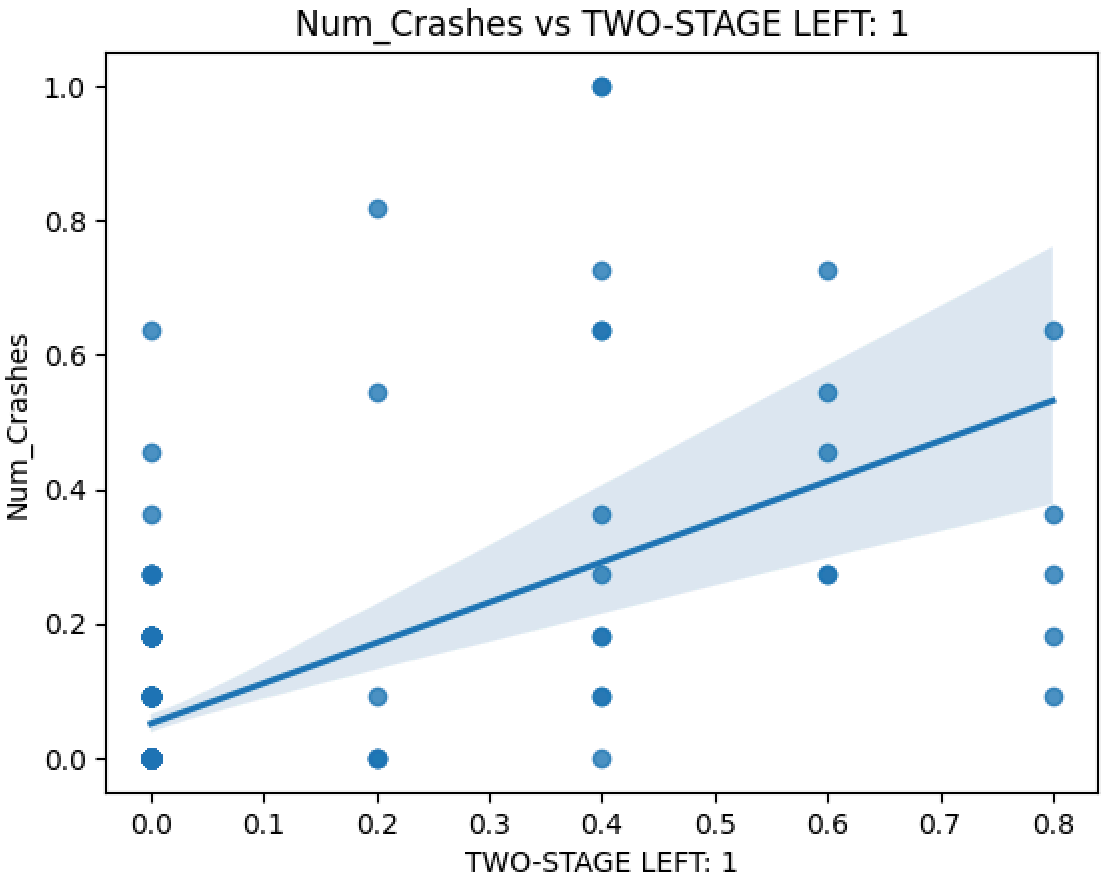
1. Exploratory Regression Plots- Set, % Bike Commuters with 1 Bike network feature.

The section used separate Ordinary Least Squares (OLS) regression models to predict the number of crashes based on the percentage of bike commuters, the presence of a specific bike network feature (as a 0/1 variable), and the interaction between these two factors. This allowed us to see how each bike feature and the level of bike commuting, both individually and together, relate to the number of crashes. The models incorporating 'Two-Stage Left Turn Lane' and 'Intersection Sharrow' exhibited the highest explanatory power for the number of crashes, with R-squared values of 0.352 and 0.402, respectively. The models for Painted Safety Zone, Protected Intersection, and Bike Channel demonstrated indicators of multicollinearity, leading to unreliable coefficient estimates and p-values for the intersection and interaction terms and so are not usable. Several models exhibited high skewness and kurtosis in their residuals, coupled with significant Jarque-Bera statistics, indicating not normally distributed residuals. Given these issues, and the poor fits of most of the models, I only included the 2 primary features here. Full results summary for each in set can be found in the output of provided .ipynb file.

Regression Model	R-squared	Prob (F-statistic)

Regression for Num_Crashes with % Bike Commuter, BIKE SIGNAL: 0, and interaction	0.171	8.42E-10
Regression for Num_Crashes with % Bike Commuter, TWO-STAGE LEFT: 1, and interaction	0.352	1.68E-22
Regression for Num_Crashes with % Bike Commuter, INTERSECTION SHARROW: 2, and interaction	0.402	1.26E-26
Regression for Num_Crashes with % Bike Commuter, BIKE BOX: 3, and interaction	0.226	2.51E-13
Regression for Num_Crashes with % Bike Commuter, MIXING ZONE: 4, and interaction	0.156	8.05E-09
Regression for Num_Crashes with % Bike Commuter, PAINTED SAFETY ZONE: 5, and interaction	0.021	0.0251
Regression for Num_Crashes with % Bike Commuter, CROSSBIKE: 6, and interaction	0.034	0.0386
Regression for Num_Crashes with % Bike Commuter, PROTECTED CORNER: 7, and interaction	0.035	0.0342
Regression for Num_Crashes with % Bike Commuter, PROTECTED INTERSECTION: 9, and interaction	0.021	0.0251
Regression for Num_Crashes with % Bike Commuter, BIKE CHANNEL: 10, and interaction	0.021	0.0251
Regression for Num_Crashes with % Bike Commuter, JUGHANDLE: 11, and interaction	0.063	0.000402
Regression for Num_Crashes with BIKE SIGNAL: 0, TWO-STAGE LEFT: 1, and interaction	0.427	8.52E-29


Num_Crashes vs % Bike Commuter
Num_Crashes
% Bike Commuter
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0



2. Exploratory Regression Plots- Coupled Bike Features

   The model explores the relationships between pairs of bike network features and the number of crashes using separate Ordinary Least Squares (OLS) regression models. For each unique combination of two different bike features (represented as binary variables), a regression model is created including both features and their interaction term to predict the number of crashes. This allows for the examination of how the presence of one bike feature might influence the effect of another on crash occurrence. The model summaries are printed to assess the statistical significance and strength of these relationships. For the most part, the combinations of intersection treatments examined in these models, including their interactions, do not statistically significantly predict the number of crashes. This is indicated by the consistently low R^2 values and the high p-values. The only statistically significant interaction term in this set is between TWO-STAGE LEFT and BIKE BOX, where the combined presence is associated with a slight increase in crashes compared to their individual non-significant effects. This happened for the previous linear set as well, and now is mentionable since it happened when examining interaction more closely. Several models suffer from perfect or severe multicollinearity and so are unusable, and these are also not normally distributed.Full results summary for each in set can be found in the output of provided .ipynb file.

3. Exploratory Regression Plots- Median Income, % Bike Commuters, Count

   This regression leveraged the overall count of bicycle network features of any type to see if infrastructure was always tightly correlated to the % Bike Commuters, or if median income can explain some of it. This model is already very limited in scope because of the highly collinear nature of median income tracts already having more cyclists. This is a commonly studied point, and I have my own model on this in another project.

   Median income variable is not statistically significant (p = 0.603). The percentage of bike commuters shows a statistically significant positive association with the number of crashes (p = 0.045). The coefficient of 0.1329 indicates that for every 1% increase in bike commuters, the number of crashes is estimated to increase by approximately 0.13, assuming all other variables remain constant. This suggests that higher rates of bike commuting are linked to a greater number of crashes, which is expected. The total count of bike features is a highly statistically significant predictor of the number of crashes (p < 0.001). The substantial positive coefficient of 1.2786 implies that each additional bike feature is associated with an increase of approximately 1.28 in the number of crashes, holding other variables constant. This suggests that a greater density of bike infrastructure is associated with a higher number of crashes, as expected. The interaction between the total count of bike features and median income is borderline statistically significant (p = 0.054). The negative coefficient of -1.3470 suggests a moderating effect of median income on the relationship between the count of bike features and crashes. Specifically, as median income increases, the positive association between the number of bike features and the number of crashes tends to weaken. In other words, the impact of adding more bike features on crash numbers is less pronounced in areas with higher median incomes.
4. Exploratory Regression Plots- Median income with bike features relative to total.

   This model used Ordinary Least Squares (OLS) regression to explore the relationship between median income and the proportion of each specific bike network feature relative to the total count of bike features at a location. For each bike feature, its count was divided by the total count of all bike features to represent its proportion. These proportions were then used as independent variables to predict median income.

The overall regression model was statistically significant (p = 0.00758) and explained approximately 28.4% of the variance in median income (R-squared = 0.284, Adjusted R-squared = 0.193). However, the model suffers from perfect multicollinearity, specifically involving PROTECTED INTERSECTION and BIKE CHANNEL: 10, as we saw with the other models. This makes sense to me because these features almost always occur in tandem, so they are not statistically usable here but that means nothing for the potential effect of these features. Despite this issue, the results suggest that the proportion of BIKE SIGNAL is statistically significantly and positively associated with median income, with p < 0.00. The coefficients for the other individual bike feature proportions were not statistically significant in predicting median income for this model.

B. Methodology

Unless specified, all non-linear approaches used the individual bike features and the % Bike Commuters as predictors, and did not use the Median Income or Count feature.

1. Lasso Regression

   The Lasso model, with an alpha of 0.1, was trained to identify the strongest predictors by potentially setting coefficients of less important features to zero. The resulting coefficients indicate the relationship between each predictor and crash numbers. Even with a low alpha, Lasso caused all our coefficients to go to zero due to colinearity. This is not statistically useful.

   ```
   Lasso Coefficients: [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
   ```

2. Ridge Regression

   Ridge regression (alpha=0.1) was used to predict crashes based on bike commuters and bike features, addressing potential collinearity. The model was trained, and coefficients were examined. Performance was evaluated on a train/test split using MSE, RMSE, and R-squared, with cross-validation for robustness.

   ```
   Intercept: 0.03287704328720133
   R-squared: 0.5242729185393799
   Mean Squared Error: 0.017443361530294
   Root Mean Squared Error: 0.1320732244450136
   R-squared: 0.5595640885786176
   ```

Average MSE (Cross-Validation): 0.021180910576971836
Average RMSE (Cross-Validation): 0.1455366296743601

Bike Network Feature	Coefficient
% Bike Commuter	0.04
BIKE SIGNAL	0.081
TWO-STAGE LEFT	0.367
INTERSECTION SHARROW	0.605
BIKE BOX	0.051
MIXING ZONE	-0.054
PAINTED SAFETY ZONE	0
CROSSBIKE	-0.17
PROTECTED CORNER	0.019
PROTECTED INTERSECTION	0
BIKE CHANNEL	0
JUGHANDLE	0.348

The analysis of the Ridge regression model revealed varying associations between bike network features and the number of crashes. The percentage of bike commuters showed a positive relationship, with an estimated increase of 0.04 crashes for every 1% rise in commuter numbers. Among the bike infrastructure, intersection sharrows exhibited the strongest positive association with crashes (0.605), followed by two-stage left turn lanes (0.367) and jughandles (0.348), suggesting these features may be linked to a higher incidence of crashes. Conversely, the presence of crossbikes (-0.17) and, to a lesser extent, mixing zones (-0.054) were associated with a potential reduction in crashes. Painted safety zones, protected intersections, and bike channels had coefficients of zero, likely due to collinearity within the model, indicating no unique linear relationship was detected for these features when others were considered. Overall, the model demonstrated reasonable predictive accuracy with a relatively low cross-validation mean squared error and a moderate fit to the data, as indicated by an R-squared of 0.56.

3. OLS with the top feature from Ridge, % Bike Commuter, interaction term.

   As a natural followup to the Ridge results, I tested the interaction term to account for the number of commuters. The presence of an intersection sharrow has a larger positive effect on crashes than the percentage of bike commuters. There is an interaction effect, meaning that the combined presence of an intersection sharrow and a higher percentage of bike commuters further increases the risk of crashes.

Parameter	Estimate
Intercept	0.06506171
Q("% Bike Commuter")	0.04097182
Q("INTERSECTION SHARROW: 2")	0.10739946
Q("INTERSECTION SHARROW: 2"):Q("% Bike Commuter")	0.02810106

4. Bootstrapped Ridge Regression

   The addition of bootstrapping aims to improve the reliability of the Ridge Regression results by assessing the stability of the predictions. By training multiple Ridge models on resampled data and averaging their outputs, it provides a more robust estimate of the model's performance and an indication of the prediction uncertainty.

   Mean Squared Error: 0.026220690723253944
   Root Mean Squared Error: 0.16192804180639603
   R-squared: 0.33794009302881456

   The non-bootstrapped Ridge Regression model has a substantially higher R-squared (0.524) compared to the bootstrapped model (0.338), and the bootstrapped model has lower MSE and RMSE values. The bootstrapped Ridge performed worse than the original.

5. GLM Exponential- Poisson with Log Link

Given the non-normal distribution implied by earlier models, I added a Poisson Regression model, starting with a log link. The relatively low values for MSE and RMSE indicate that, on average, the squared differences between the predicted and actual number of crashes are small. However, the R^2 value of -5.911 is extremely low and negative, implying that the model fits the data worse than a horizontal line for the mean of the dependent variable. This model is a poor fit and not usable.

```
Poisson Regression with Log Link Results:
Mean Squared Error: 0.273714428797652
Root Mean Squared Error: 0.5231772441512074
R-squared: -5.91115848850362
```

6. GLM Exponential- Poisson with Square Root Link

   I adjusted the previous model by replacing the log link with a sqrt() link. The square root link was specifically chosen to potentially stabilize the variance and ensure non-negative predictions for the count data. This should help stabilize the variance, and ensure that it predicts positive values. The square root link Poisson model also has a negative R^2 of -0.088, still indicating a poor fit, but substantially better than the log link model at -5.911 in terms of variance explained. The square root link model has considerably smaller MSE at 0.0431, than the log link model of 0.523. At this point, Ridge is still performing better than this Poisson.

7. GLM Exponential- Negative Binomial

   This model initializes to an alpha of 1, fits it, estimates the overdispersion parameter alpha, and refits the model if the estimated alpha significantly deviates from the initial value. Finally, it makes predictions on test data and evaluates the model using MSE, RMSE, and R-squared. The high MSE and RMSE values, along with a negative R-squared, suggest that the Negative Binomial Regression model is a poor model. Like with Poisson, negative R-squared indicates that the model explains less variance than simply using the mean as a predictor. This is a sign of a poor fit.

   ```
   Estimated alpha: 1.0
   Mean Squared Error: 0.38517883673653175
   Root Mean Squared Error: 0.62062777631728
   R-squared: -8.725581507694589
   ```

8. Polynomial Regression (with BoxCox Transform)

The MSE and RMSE values are extremely high compared to the other models. This suggests that the Polynomial Regression model is making large prediction errors. A negative R-squared value, especially one as large as -11939.305, is a strong indication of a poor model fit. This is the worst performing model after Lasso.

```
Mean Squared Error: 472.89233552976117

Root Mean Squared Error: 21.7460878212556

R-squared: -11939.305424164935
```

9. Support Vector Machine

This model uses support vector machine regression, simultaneously considering all specified bike infrastructure features and the percentage of bike commuters to make a prediction. A small grid with GridSearchCV was employed, testing three kernel types (linear, radial basis function (rbf), and polynomial), regularization strengths (0.1, 1, and 10), kernel coefficient gamma settings ('scale', 'auto', 0.1, and 1), and polynomial degrees (2 and 3).

```
Best Hyperparameters: {'C': 10, 'degree': 2, 'gamma': 'scale', 'kernel': 'rbf'}
Mean Squared Error: 0.03902913776365322
Root Mean Squared Error: 0.1975579352080124
R-squared: 0.014532927843360133
```

The R-squared value of 0.0145 indicates that the SVM model explains a very small proportion of the variance in Num_Crashes. This suggests that the model's fit is not very good. It's only slightly better than a model that simply predicts the mean. The MSE and RMSE values are relatively low, but not as low as the Poisson with square root link or Ridge Regression models. This suggests that the SVM model's predictions are not as accurate as those of the other models.

IV. Results and Conclusions

Model	R-squared	MSE
Ridge Regression	0.524	0.017443
Bootstrapped Ridge Regression	0.338	0.02622
GLM - Poisson (Log Link)	-5.911	0.273714
GLM - Poisson (Sqrt Link)	-0.088	0.0431
GLM - Negative Binomial	-8.725	0.385178
Polynomial Regression	-11939.305	472.892335
Support Vector Machine	0.0145	0.039029

Our best model was Ridge Regression standard. We'll explore the implications of the results from this model and also of the OLS with the interaction terms because that was so illuminating. We know the natural relationship between more commuters and more crashes, so we do not need to explore this relationship except when discussing the interaction between % Bike Commuters and the bike features.

1. BIKE SIGNAL (Coefficient: 0.08103647)
    - Analysis: A positive coefficient indicates that the presence of bike signals is associated with a moderate increase in crashes. Our MULTINOMIAL REGRESSION MODEL showed bike signals are more likely to be found in high income neighborhoods regardless of the % of Bike Commuters.
    - Recommendation:
        - Further analyze the relationship.
        - No recommendations can be made for this at this point.
2. TWO-STAGE LEFT (Coefficient: 0.36705863)
    - Analysis: A relatively **high** positive coefficient suggests a notable association between two-stage left turn lanes and increased crashes.
    - Recommendations:
        - Thoroughly investigate the safety performance of existing two-stage left turn lanes.
        - Consider design modifications to enhance visibility and reduce confusion.
        - Evaluate alternative intersection designs if safety concerns persist.
3. INTERSECTION SHARROW (Coefficient: 0.60540816)
    - Analysis: The highest positive coefficient in the Ridge model, indicating a strong positive relationship with crashes.
    - Additional Analysis (from OLS with Interaction): The OLS model further clarifies that intersection sharrows have a substantial positive effect on crashes (coefficient: 0.10739946), *and* that this effect is significantly increased when combined with higher percentages of bike commuters (interaction coefficient: 0.02810106). This means the risk associated with sharrows grows as bike commuter volume increases.
    - Recommendations:
        - Strongly discourage the installation of new intersection sharrows in favor of other bike features.
        - Systematically replace existing sharrows with protected bike lanes or other safer alternatives.
        - Where immediate replacement isn't feasible, implement urgent, enhanced safety measures, especially in areas with high bike commuter volume.
4. BIKE BOX (Coefficient: 0.05076579)
    - Analysis: A small positive coefficient suggests a slight association with increased crashes.
    - Recommendations:
        - Further analyze the relationship.
        - No recommendations can be made for this at this point.

5. MIXING ZONE (Coefficient: -0.05424209)
    - Analysis: A small negative coefficient suggests a slight association with a reduction in crashes.
    - Recommendations:
        - Further investigate the factors that contribute to the potential safety benefit- what kind of mixing zone?
        - I feel that mixing zones are very dangerous from considerable domain expertise.
6. PAINTED SAFETY ZONE (Coefficient: 0)
    - Analysis: A coefficient of 0 indicates that this feature did not show a unique linear relationship with crashes in this model, likely due to multicollinearity.
    - Recommendations:
        - We could not analyze this feature- further analysis is needed.
7. CROSSBIKE (Coefficient: -0.16992438)
    - Analysis: A moderate negative coefficient suggests an association with a reduction in crashes.
    - Recommendations:
        - Study the design and implementation of crossbikes to understand their safety benefits.
        - Promote their use and ensure clear visibility for both cyclists and drivers.
        - Consider wider adoption where appropriate.
8. PROTECTED CORNER (Coefficient: 0.01891967)
    - Analysis: A very small positive coefficient, indicating a minimal association with increased crashes.
    - Recommendations:
        - Monitor their safety performance.
        - Ensure designs adhere to best practice guidelines.
9. PROTECTED INTERSECTION (Coefficient: 0)
    - Analysis: A coefficient of 0, likely due to multicollinearity.
    - Recommendations:
        - Similar to painted safety zones, interpret with caution.
        - Further research is needed.
10. BIKE CHANNEL (Coefficient: 0)
    - Analysis: A coefficient of 0, likely due to multicollinearity.
    - Recommendations:
        - Similar to painted safety zones and protected intersections, interpret with caution.

- Further research is needed.

11. JUGHANDLE (Coefficient: 0.3484909)
    - Analysis: A relatively high positive coefficient, suggesting a notable association with increased crashes.
    - Recommendations:
        - In-depth safety analysis of jughandle intersections.
        - Consider design improvements or alternatives.
        - Implement driver and cyclist education programs focused on safe negotiation of jughandles.

V. Future Discussion and Changes

Going forward with this project, I have some initial plans.

Run this study again checking for entries that have combined bike channel and protected intersection, and if they're almost all together, just combine that into its own single value.

Checking **every** interaction term of features with % commuter would have given much more influential analysis for the ridge for the next few positive coefficients.

The date on features was spread across the 4 years- this is not a time sensitive model. Analyzing more time sensitive data on bike commuting, which we do not have access to but the city of San Francisco does, would allow us to use times of crashes and only weight those bike features which would have been installed after the crash.

The commuters living in an area isn't that predictive for us- the commuters passing through commuting corridors is more important. SF does collects that data in large amounts through auto counters and probably AI, but no longer publishes count data since 2019.

Code Files are attached along with this file.

References

San Francisco Municipal Transportation Agency (SFMTA). (n.d.). SFMTA Bikeway Network Point Features. [Data set]. Retrieved from data.sfgov data portal.

Transportation Injury Mapping System (TIMS), University of California, Berkeley. Transportation Injury Mapping System. 15 April 2025 from https://www.pa.gov/agencies/education/programs-and-services/educators/certification/teacher-information-management-system-tims.html. Data for years 2018-2023.

U.S. Census Bureau. (2023). American Community Survey 5-Year Estimates. [Data set]. Retrieved from https://www.census.gov/programs-surveys/acs/

U.S. Census Bureau. (2023). American Community Survey 5-Year Estimates. [Data set]. Retrieved from https://www.census.gov/programs-surveys/acs/