MATH 448-01 AN INTRODUCTION TO STATISTICAL LEARNING DATA MINING
SAN FRANCISCO STATE UNIVERSITY
SPRING 2025
FINAL PROJECT

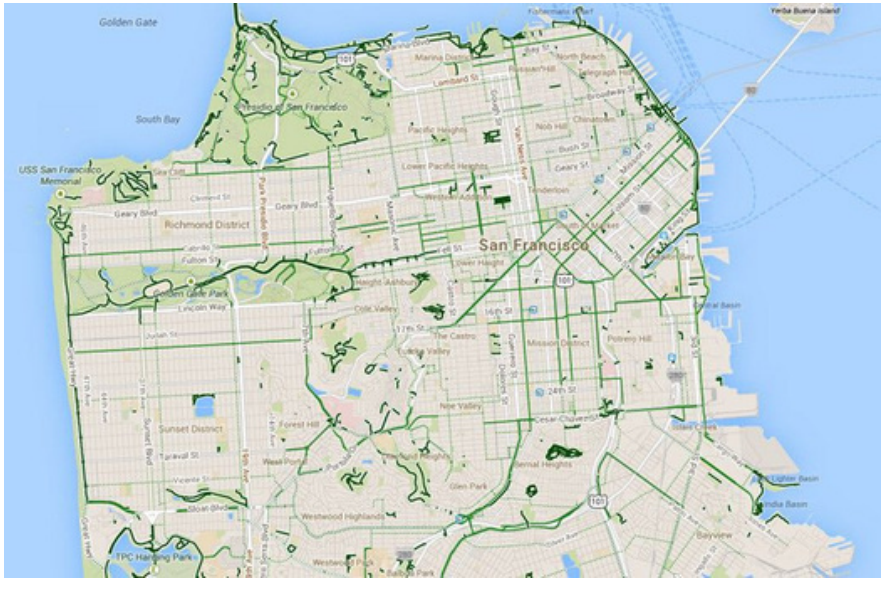
PAIGE HODGKINSON

DATA PROJECT

BICYCLE SAFETY ANALYSIS

A BY-CENSUS-TRACT ANALYSIS OF BICYCLE INFRASTRUCTURE, COMMUTERS, AND CRASHES IN SAN FRANCISCO 2018-2023

BACKGROUND



I AM A LOCAL CYCLIST

I am often made uncomfortable my work commute when it is too late in the day. I'd like to see what factors might explain why I feel unsafe.

I previous looked at a simple but statistically signifcant linear model between **crashes and median income** when accounting for the percent of bike commuters in another course. I wanted a more in-depth analysis of **why** that is.

San Francisco uses mass amounts of data to drive infrastructure decisions around reported bike passing through chokepoints.

Why and when to put different bike infrastructure?

DATASETS

AMERICAN COMMUNITY SURVEY (ACS) 5-YEAR 2023 MEANS OF TRANSPORT

Estimates of commuting methods, including bike commuting, for all census tracts in San Francisco.

AMERICAN COMMUNITY SURVEY (ACS) 5-YEAR 2023 GEOGRAPHIC MOBILITY BY SELECTED CHARACTERISTICS IN THE UNITED STATES

Gives us median income by census tract for San Francisco.

SFMTA BIKEWAY NETWORK POINT FEATURES

detailed information about the locations and features of bikeways within San Francisco.

TIMS BERKELEY

Records information about traffic collisions in San Francisco.

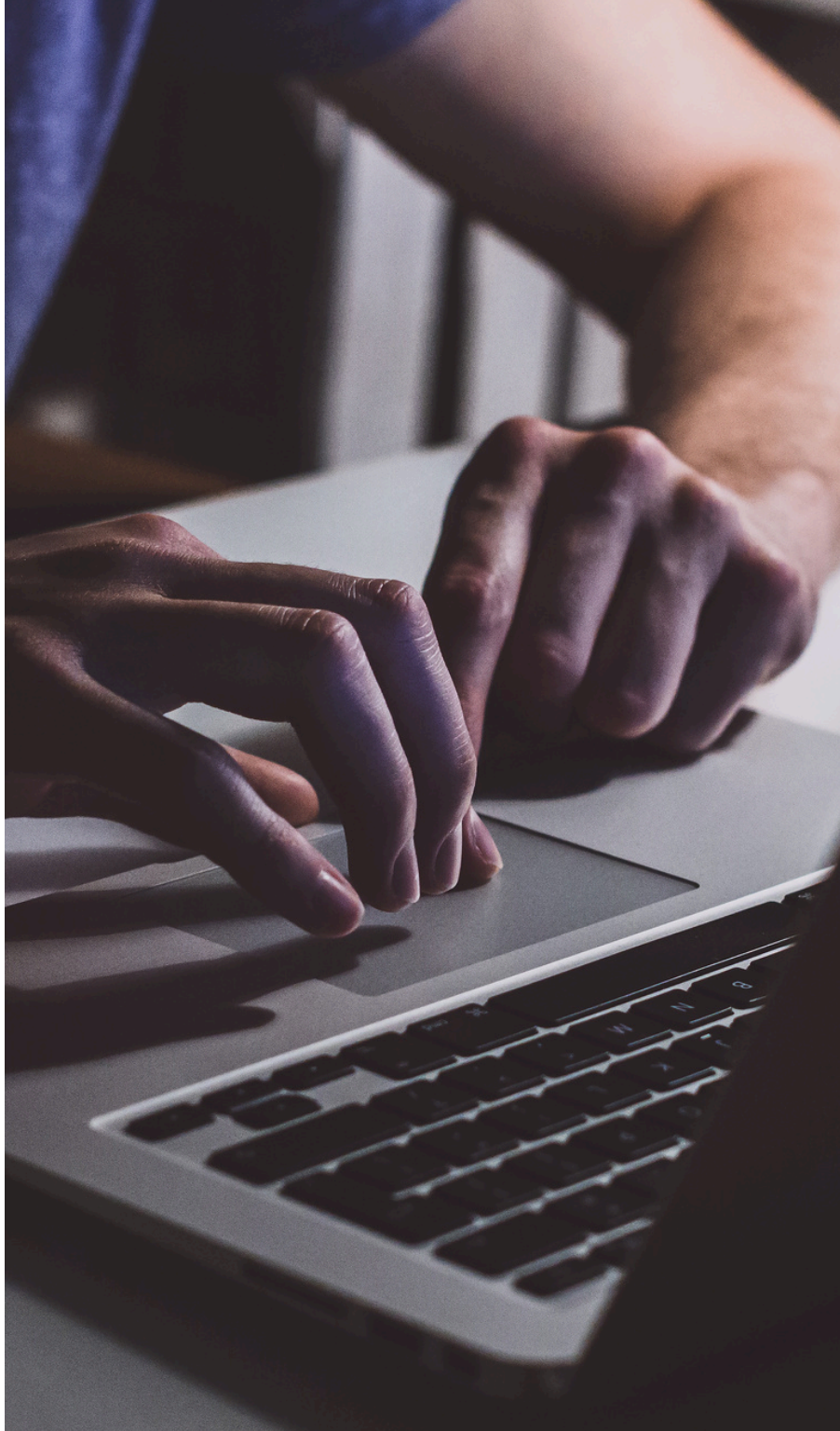
CLEANING

Loaded ACS data, filtered for total and bicycle commuters, calculated the percentage of bike commuters, and standardized census tract GEOIDs from column names.

Loaded crash data, selected relevant columns, handled missing coordinates (imputing some manually), and used a geocoding API to add the census tract for each crash.

Loaded income data, selected the median income estimate, renamed columns, removed the header, and formatted census tract IDs.

Loaded bikeway data, extracted latitude and longitude from a geometry column, and used a geocoding API to add the census tract for each bikeway segment.



QUESTIONS

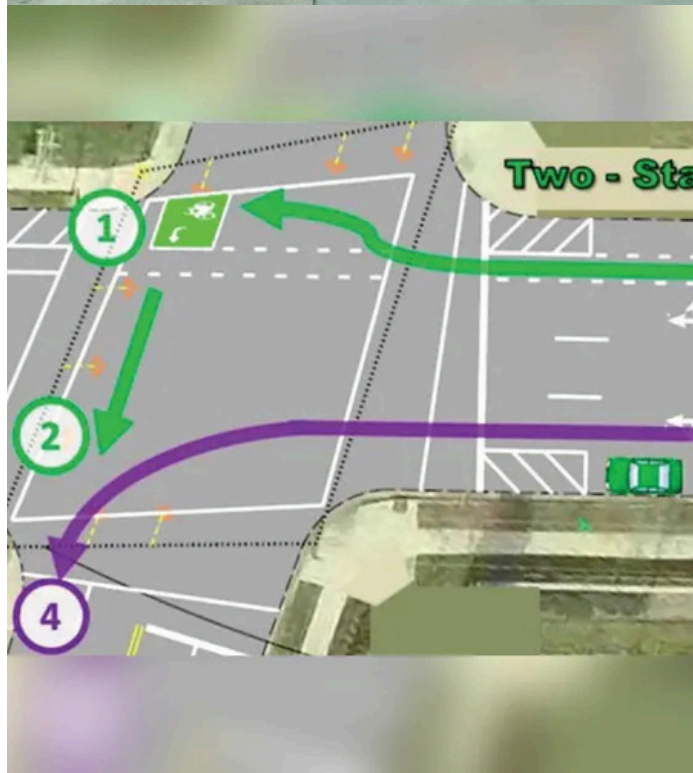
EXPLORATORY REGRESSION PLOTS

- Does the proportion of commuters affect the relationship of certain bike features to crashes?
- Do different bike features interact to affect their relationship to number of crashes?
- Is median income a predictor when using the overall density of bike features, when accounting for increase in the proportion of commuters??
- Do areas of certain income have association with certain bike features with accounting for increase in the proportion of commuters?

EXPLORATORY REGRESSION PLOTS

% BIKE COMMUTER + 1 BIKE FEATURE

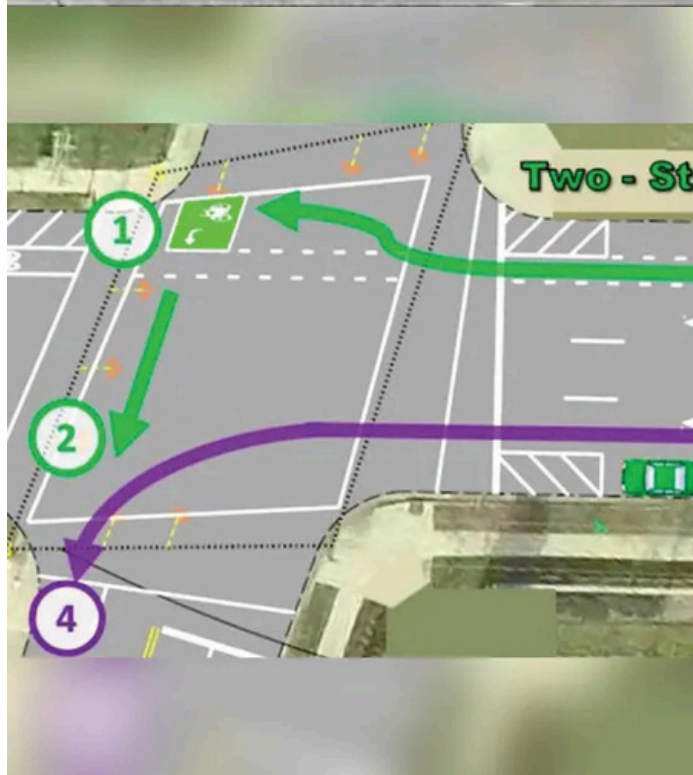
- Models for **TWO-STAGE LEFT** and **INTERSECTION SHARROW** explained the **most variance** in the number of crashes.
- Models for 'PAINTED SAFETY ZONE: 5', 'PROTECTED INTERSECTION: 9', and 'BIKE CHANNEL: 10' were unreliable due to multicollinearity.
- Most models exhibited high skewness and kurtosis 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.



EXPLORATORY REGRESSION PLOTS

COUPLED PAIRS OF BIKE FEATURES WITH INTERACTION TERM

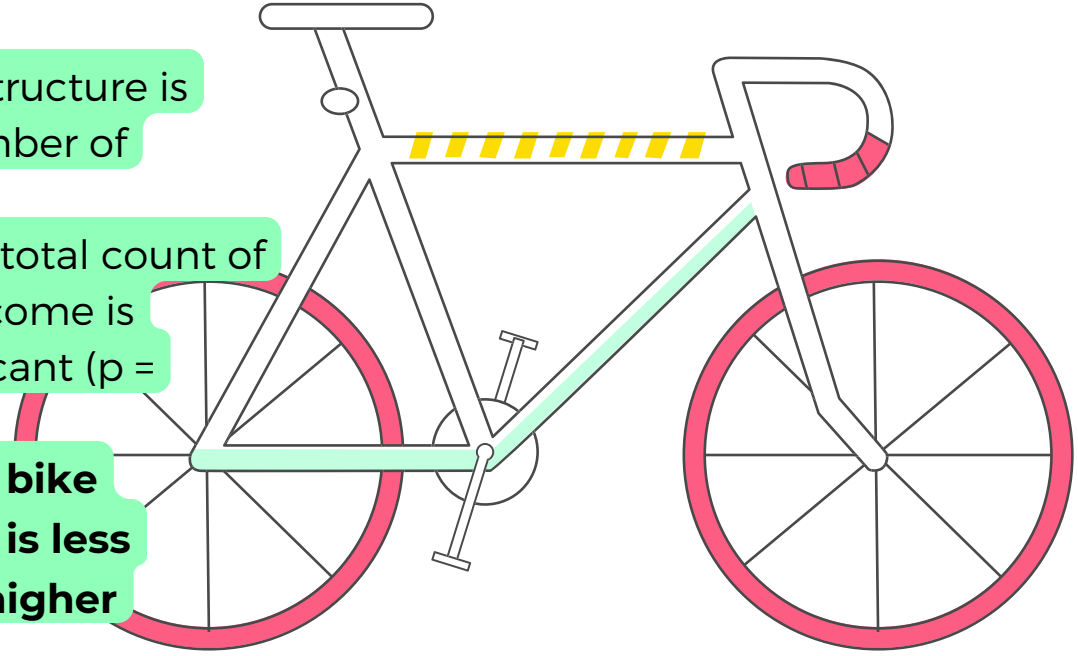
- Most were not usable.
- The only statistically significant interaction term in this set is between **TWO-STAGE LEFT and BIKE BOX** where the **combined presence is associated with a slight increase in crashes** compared to their individual non-significant effects.
- Probably just due to high crash intersections having more infrastructure.
- Again most exhibited high skewness and kurtosis in their residuals, coupled with significant Jarque-Bera statistics, indicating not normally distributed residuals.



EXPLORATORY REGRESSION PLOTS

EXPLORATORY REGRESSION PLOTS- MEDIAN INCOME, % BIKE COMMUTERS, COUNT

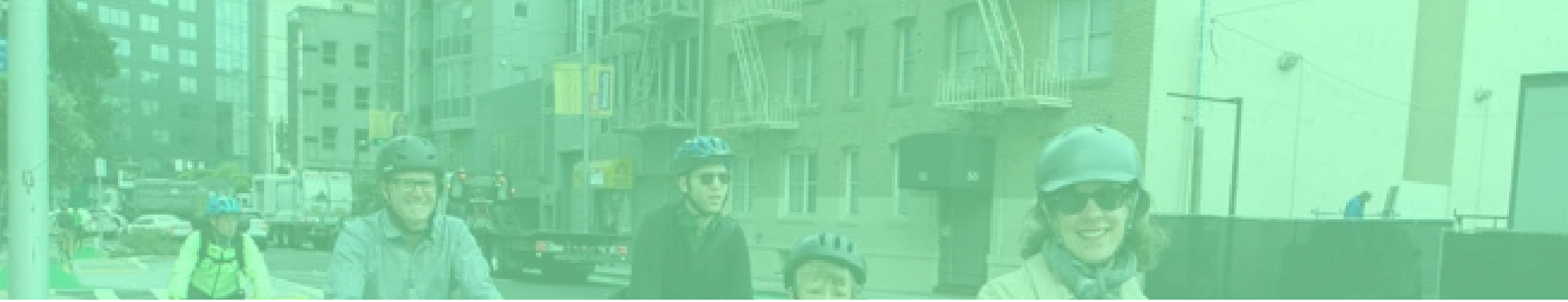
- Used overall count of bike features.
- Natural limitation- higher median income tracts have more cyclists.
- **Median income- not statistically significant.**
- Greater density of bike infrastructure is associated with a higher number of crashes, as expected.
- The interaction between the total count of bike features and median income is borderline statistically significant (p = 0.054).
- The **impact of adding more bike features on crash numbers is less pronounced in areas with higher median incomes.**



EXPLORATORY REGRESSION PLOTS

EXPLORATORY REGRESSION PLOTS- MEDIAN INCOME AS A RESPONSE FOR RELATIVE PROPORTION OF BIKE FEATURES

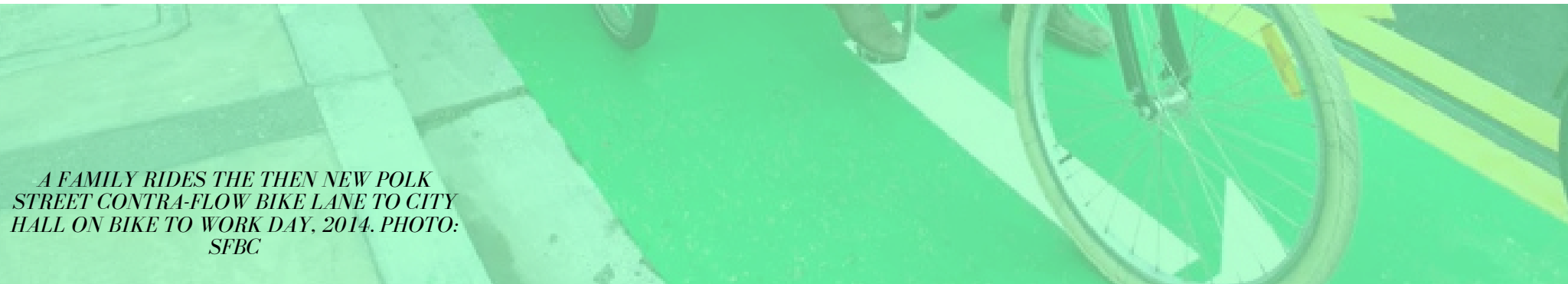
- For each bike feature, **its count was divided by the total count of all bike features** to represent its proportion for the predictors.
- The overall regression model was statistically significant ($p = 0.00758$) and explained approximately 28.4% of the variance in median income.
- However, the model suffers from perfect multicollinearity, specifically involving **PROTECTED INTERSECTION and BIKE CHANNEL**, which was seen in the other linear models as well.
- The only significant coefficient suggests that the proportion of **BIKE SIGNAL** is statistically significantly and positively associated with median income ($p < 0.001$).



LASSO REGRESSION

ALL COEFFICIENTS → 0

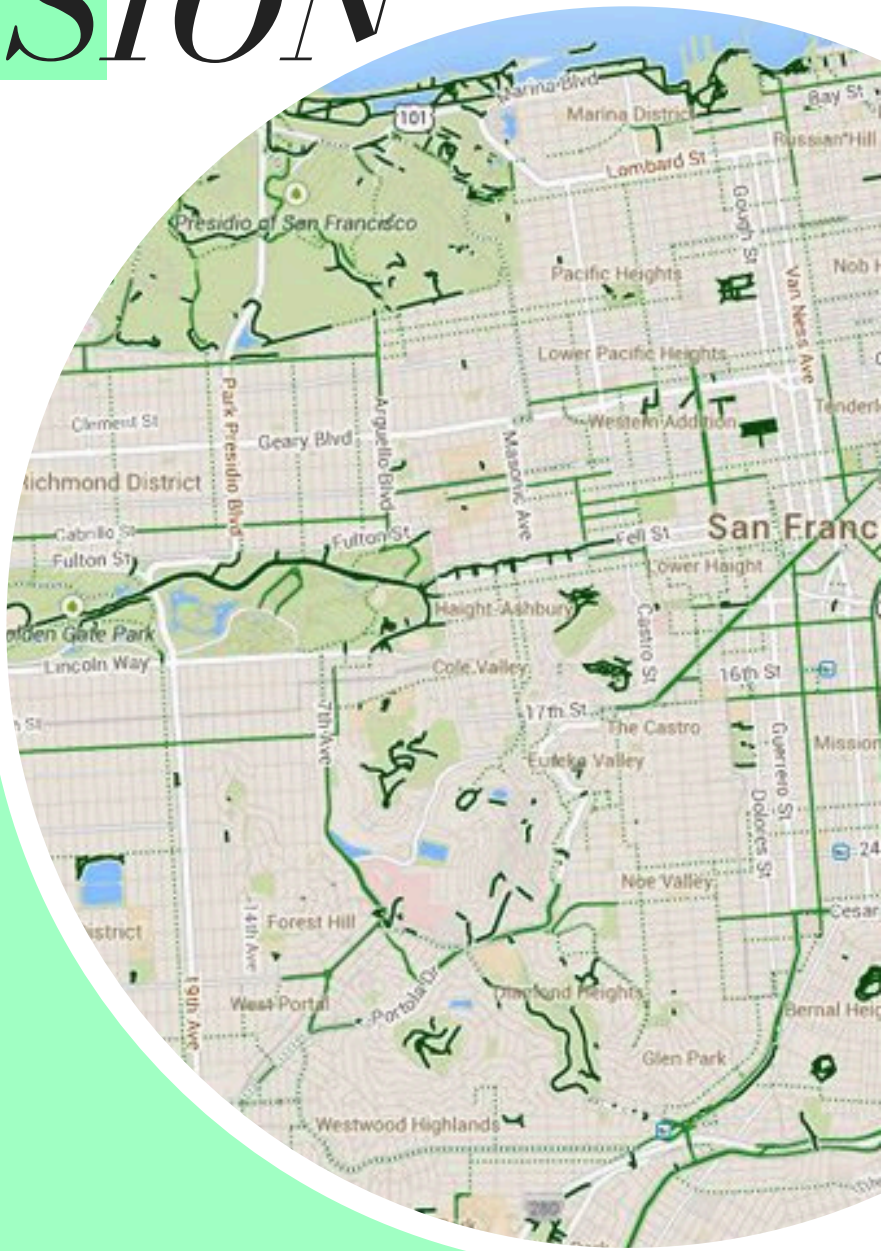
Did not handle the colinearity of our parameters.



A FAMILY RIDES THE THEN NEW POLK STREET CONTRA-FLOW BIKE LANE TO CITY HALL ON BIKE TO WORK DAY, 2014. PHOTO: SFBC

RIDGE REGRESSION

- **INTERSECTION SHARROW** has the largest positive coefficient (0.605), indicating the strongest positive association with crash numbers among the bike features.
- **TWO-STAGE LEFT** (0.367) and **JUGHANDLE** (0.348) also show substantial positive associations with crashes.
- **CROSSBIKE** (-0.17) and **MIXING ZONE** (-0.054) have negative coefficients, suggesting a potential association with a reduction in crashes.
- **PAINTED SAFETY ZONE, PROTECTED INTERSECTIO**N, and **BIKE CHANNEL** have coefficients of 0, likely due to collinearity within the model.
- R^2 : 0.5595640885786176
- MSE average: 0.021180910576971836



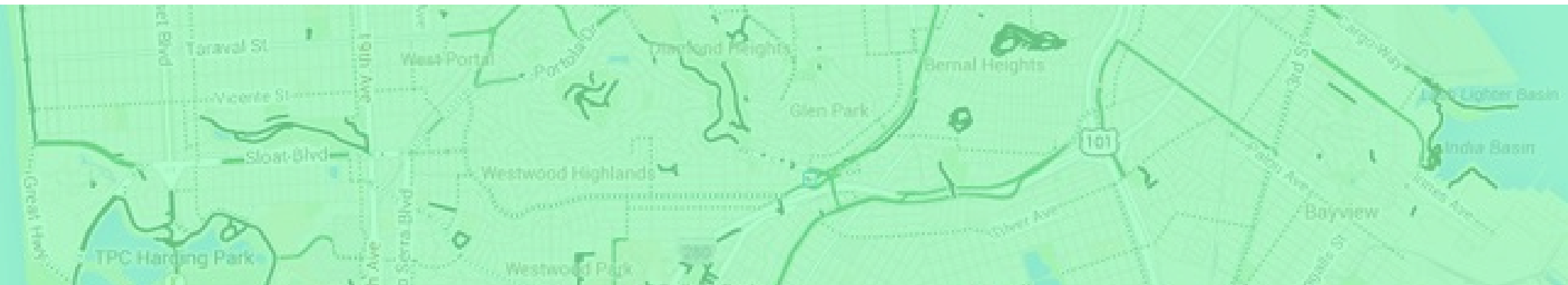
BOOTSTRAPPED RIDGE

R^2 : 0.338
MSE: 0.0262

VS

STANDARD RIDGE

R^2 : 0.524
MSE: 0.0212



ANALYSIS

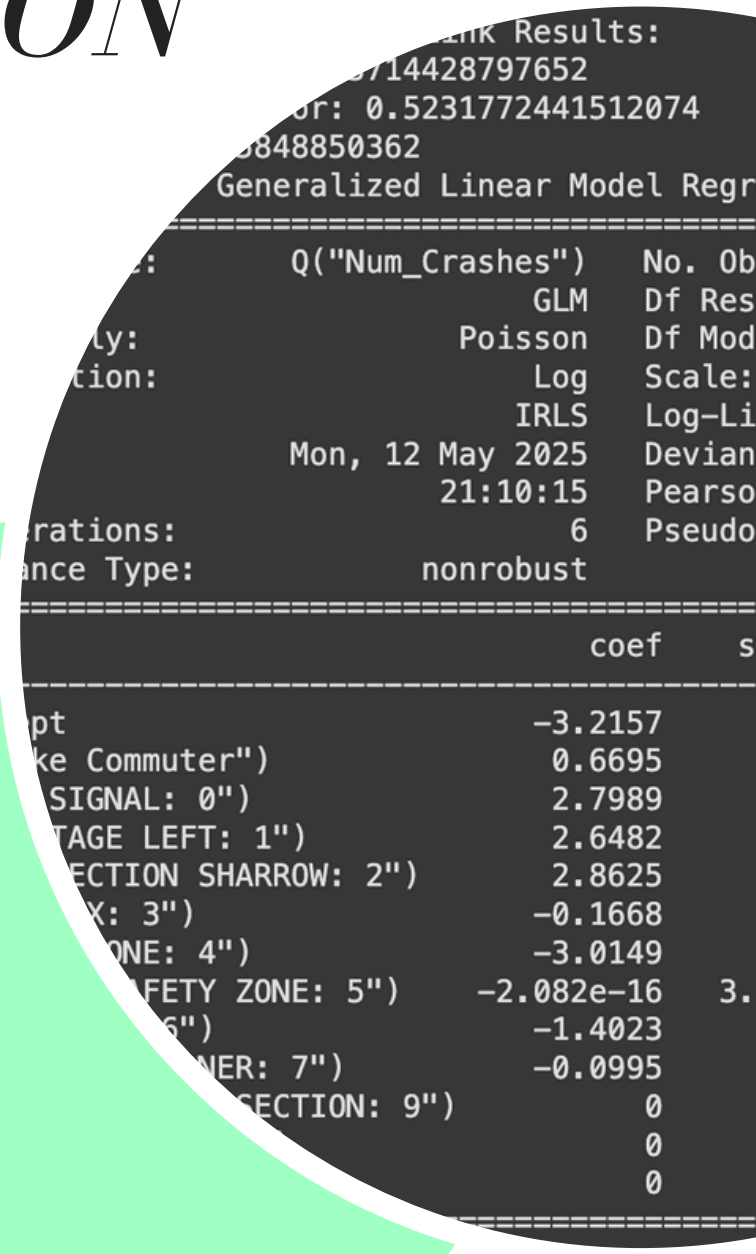
NON-LINEAR MODELS

Finding the best model fit as a priority..

POISSON REGRESSION

WITH LOG LINK

- The model was used due to the non-normal distribution observed in earlier models.
- It has a very low and negative R-squared of -5.911, indicating a poor fit.
- The model's predictions are **worse** than using the mean of the dependent variable.
- Mean Squared Error (MSE): 0.273714428797652
- Root Mean Squared Error (RMSE): 0.5231772441512074



POISSON REGRESSION

WITH SQUARE ROOT LINK

- The square root link was used to **stabilize variance** and ensure non-negative predictions.

- It has a **smaller MSE and larger R^2** compared to the log link model.

POISSON WITH LOG TRANSFORM LINK

VS

POISSON WITH SQUARE ROOT TRANSFORM LINK

R^2 : -5.9112
MSE: 0.274

R^2: -0.08796
MSE: 0.0431

RIDGE REGRESSION

VS

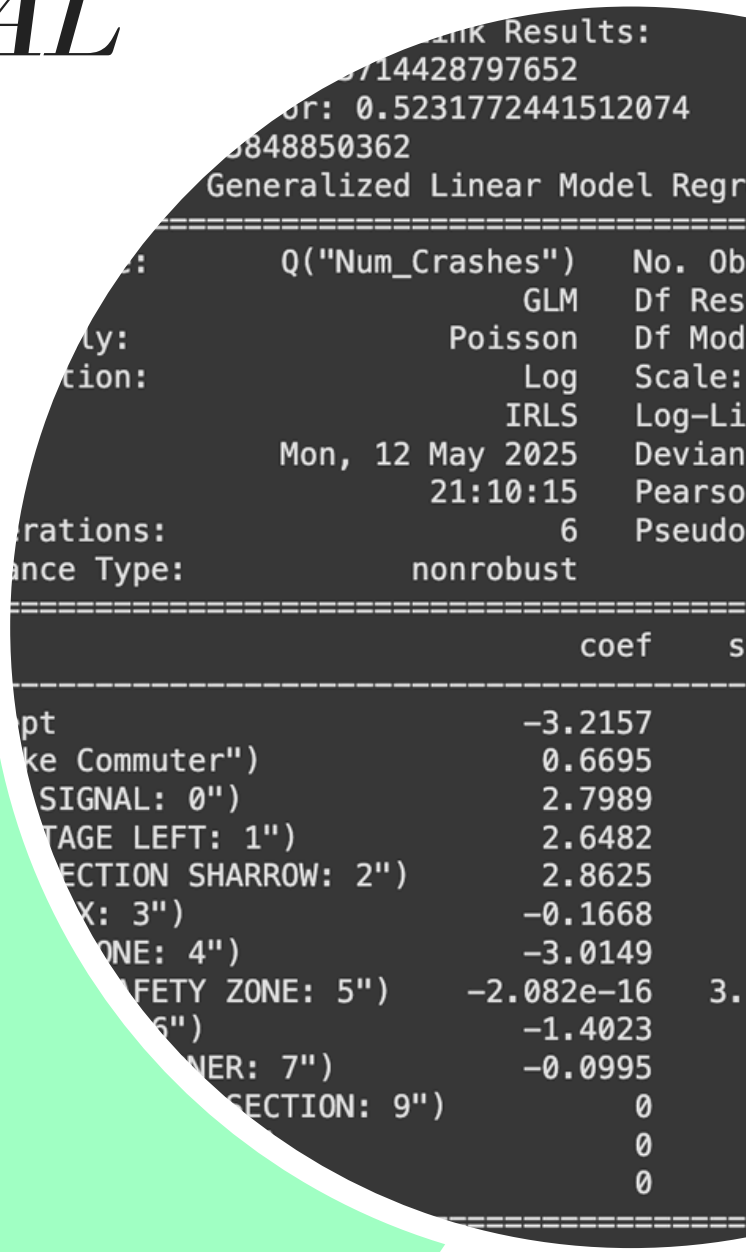
POISSON WITH SQUARE ROOT TRANSFORM LINK

R^2 : 0.5596
MSE: 0.0212

R^2: -0.08796
MSE: 0.0431

NEGATIVE BINOMIAL

- The model initializes to an alpha of 1, fits it, estimates the overdispersion parameter alpha, and refits the model.
- It has **high MSE** of 0.3851788673653175 and RMSE values of 0.62062777631728, and a negative R-squared of -8.725581507694589, indicating a **poor model**.
- The model explains less variance than using the mean as a predictor.



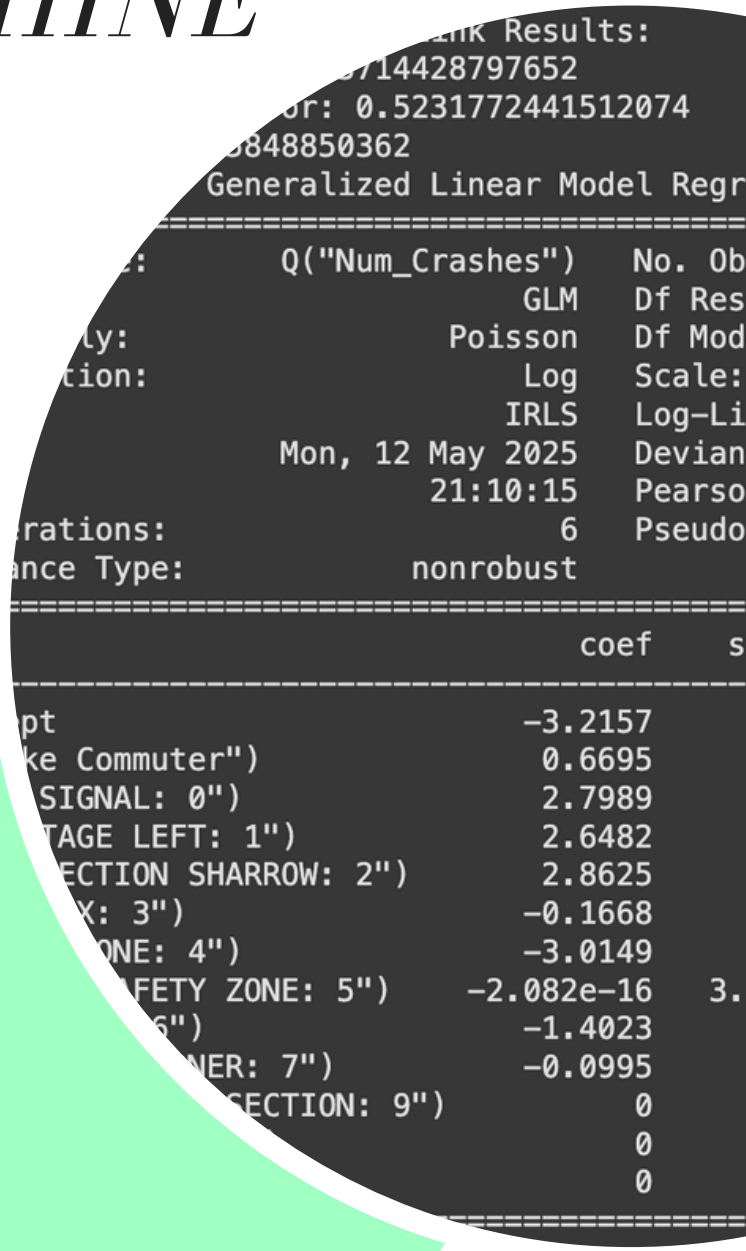
POLYNOMIAL REGRESSION

WITH BOXCOX TRANSFORM

- It has extremely **high MSE** of 472.89233552976117.
- It has a **large negative R^2** of -11939.305424164935, indicating a very **poor** model fit.
- It is the **worst-performing model after Lasso**.

SUPPORT VECTOR MACHINE

- Uses **gridsearchCV**, with the best hyperparameters being C': 10, 'degree': 2, 'gamma': 'scale', 'kernel': 'rbf, to predict crash numbers.
- It explains a very small proportion (0.0145) of the variance in crash numbers, indicating a **poor** fit.
- The model's predictions are **not as accurate as Poisson with square root link or Ridge Regression.**
- Mean Squared Error (MSE): 0.03902913776365322
- Root Mean Squared Error (RMSE): 0.1975579352080124
- R-squared: 0.014532927843360133



NON-LINEAR REVIEW

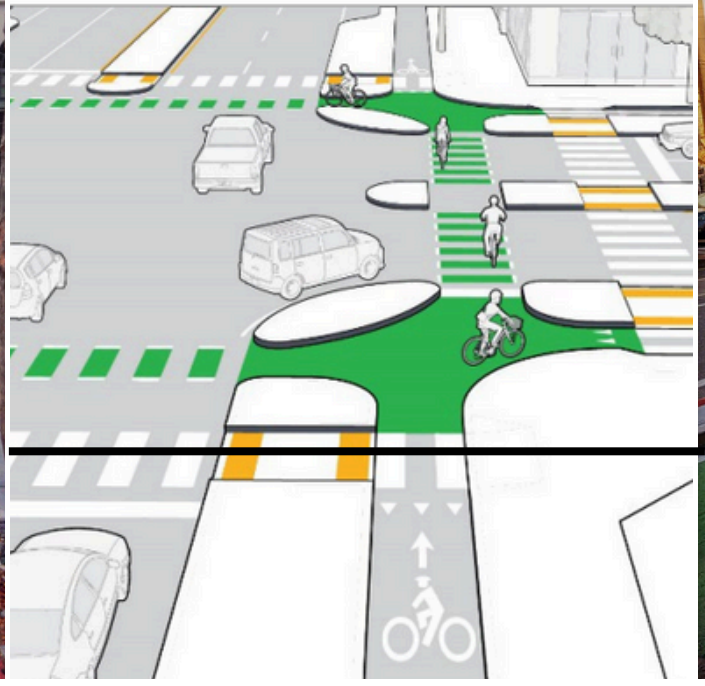
Model	R-squared	MSE
Ridge Regression	0.524	0.017443
Bootstrapped Ridge Regression	0.338	0.02622
GLM - Poisson (Log Link)	-5.911	0.273714
GLM - Poisson (Sqrt Link)	-0.088	0.0431
GLM - Negative Binomial	-8.725	0.385178
Polynomial Regression	-11939.305	472.892335
Support Vector Machine	0.0145	0.039029

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.

PROTECTED INTERSECTION

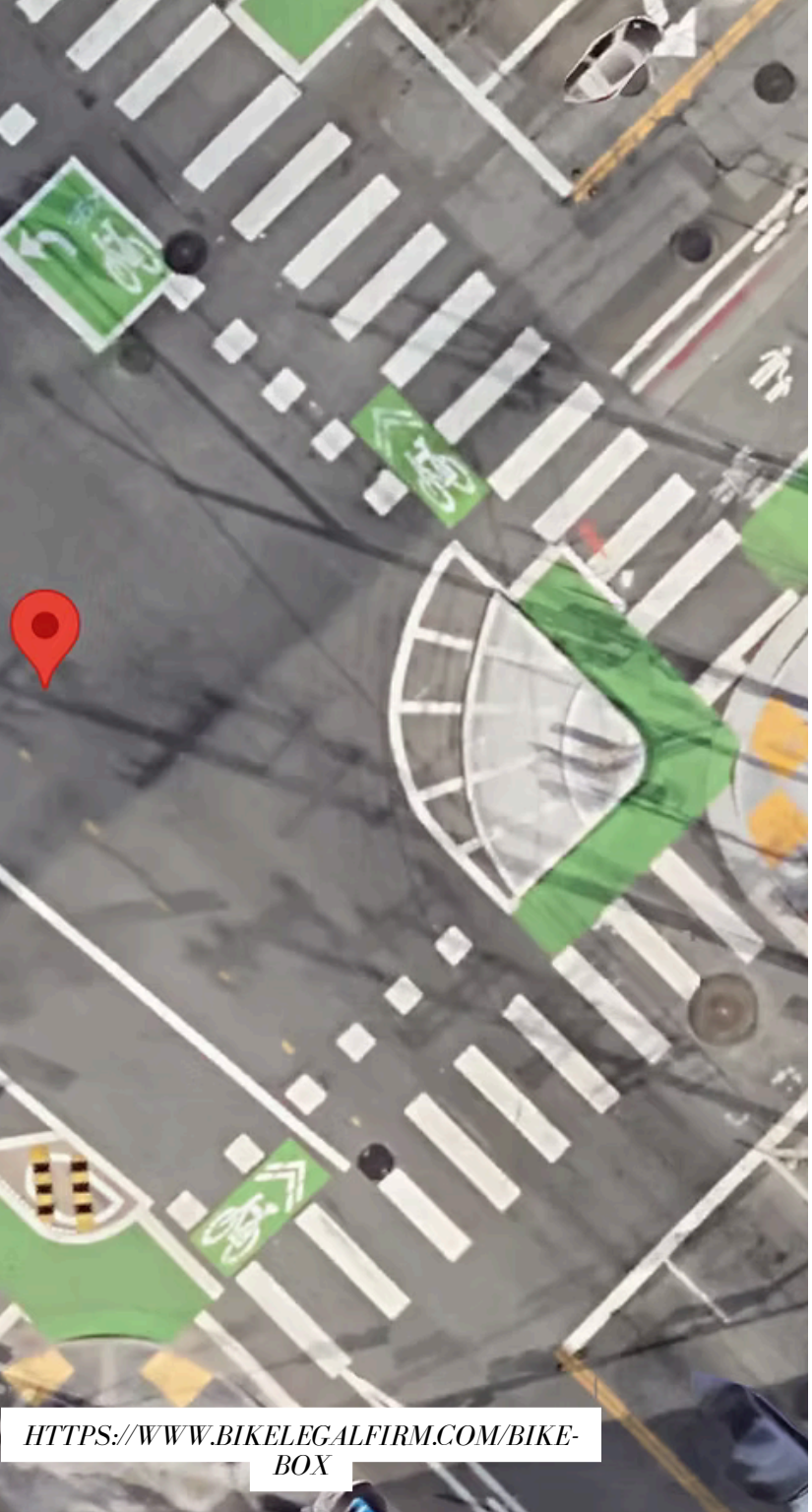


COEFFICIENT: 0

No recommendations.
We could not analyze this feature- further analysis i needed.

BIKE CHANNEL

COEFFICIENT: 0



HTTPS://WWW.BIKELEGALFIRM.COM/BIKE-BOX

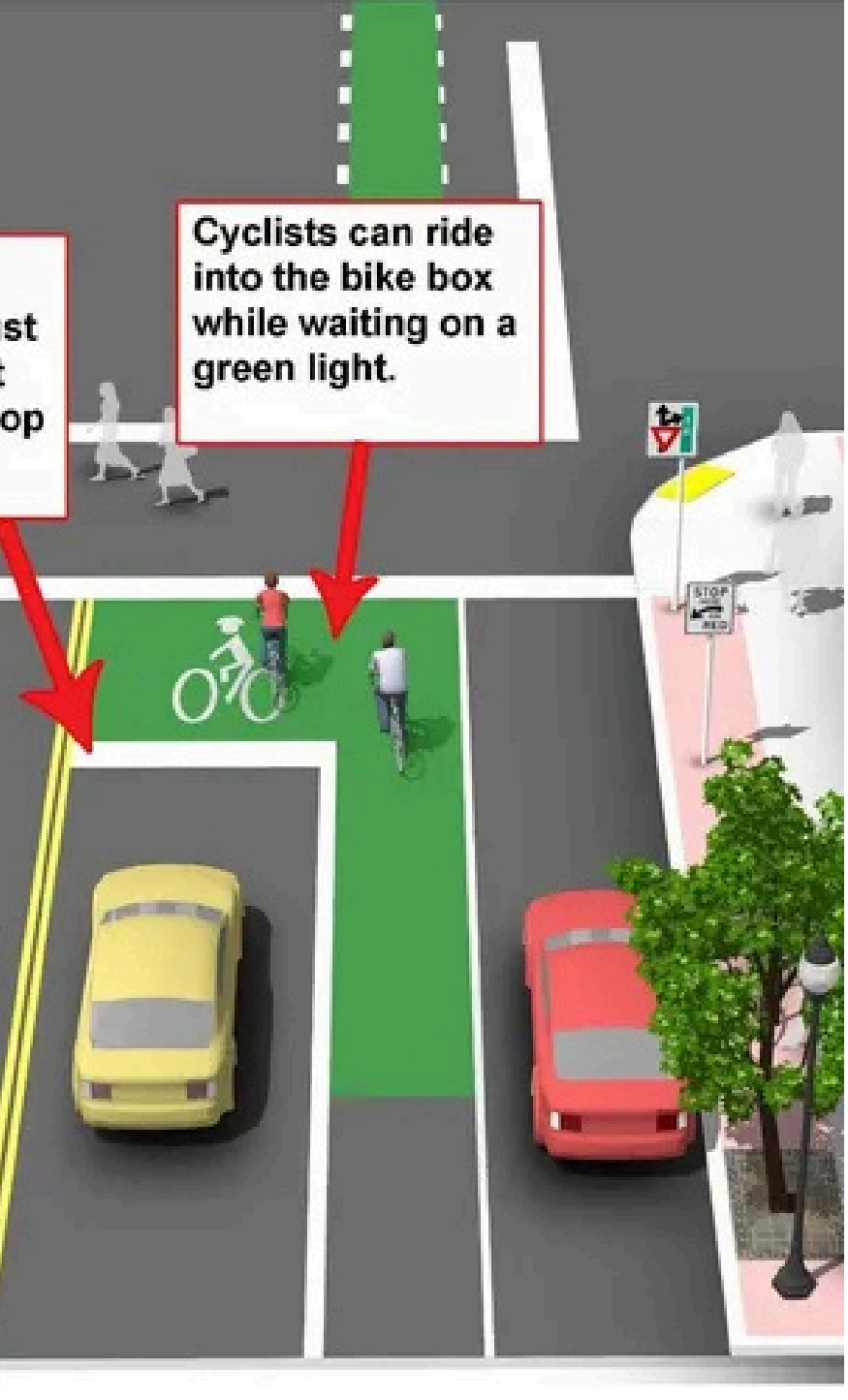
COEFFICIENT:
0.01891967

Analysis: A **small** positive coefficient suggests a slight association with increased crashes. Likely be due to increased volume of commuters.

Recommendations:
Further analyze the relationship.

No recommendations can be made for this at this point.

PROTECTED CORNER



HTTPS://WWW.BIKELEGALFIRM.COM/BIKE-BOX

COEFFICIENT:
0.05076579

Analysis: A **small** positive coefficient suggests a slight association with increased crashes. Could be due to increased volume of commuters.

Recommendations:
Further analyze the relationship.

No recommendations can be made for this at this point.

BIKE BOX

HTTPS://WWW.CAMBRIDGEMA.GOV/PARTICIPATORYBUDGETING/PBCYCLES/PBCYCLE7/BIKESIGNALSATBUSYINTERSECTIONS

COEFFICIENT:
0.08103647

The presence of bike signals is associated with a **moderate** increase in crashes. Could be due to increased volume of commuters.

Our MULTINOMIAL REGRESSION MODEL showed bike signals are **more likely to be found in high income neighborhoods regardless of the % of Bike Commuters.** Investigate this!

Recommendation:
Further analyze the relationship.
No recommendations can be made for this at this time.

BIKE SIGNAL

HTTPS://WWW.CAMBRIDGEMA.GOV/PARTICIPATORYBUDGETING/PBCYCLES/PBCYCLE7/BIKESIGNALSATBUSYINTERSECTIONS

COEFFICIENT:
0.3484909

Analysis: A relatively **high** positive coefficient, suggesting a notable association with increased crashes.
This could also be to increased commuters in these areas.

Recommendations:
In-depth safety analysis of jughandle intersections.
Consider design improvements or alternatives.

JUGHANDLE

COEFFICIENT: 0.60540816

Analysis: The **highest positive coefficient** indicating a strong positive relationship with crashes.

The OLS interaction term 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.**

INTERSECTION SHARROW

HTTPS://WWW.CAMBRIDGEMA.GOV/PARTICIPATORYBUDGETING/PBCYCLES/PBCYCLE7/BIKESIGNALSATBUSYINTERSECTIONS

COEFFICIENT: 0.60540816

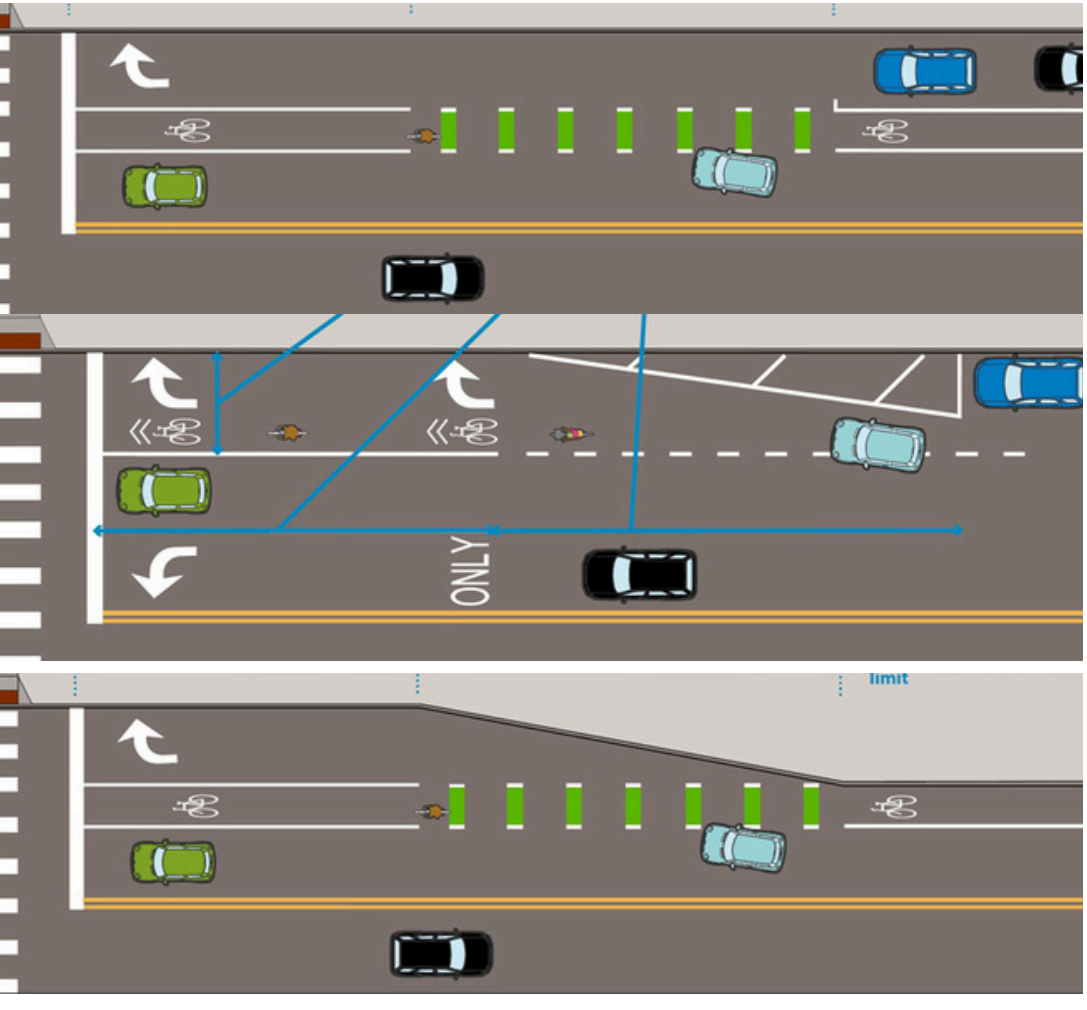
Recommendations:

Strongly discourage the installation of new intersection sharrows **in favor of other bike features- I say bike channels which we could not analyze.**

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.

INTERSECTION SHARROW



-COEFFICIENT:
-0.05424209

MIXING ZONE
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.

MIXING ZONE

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.

CROSS BIKE

CHECK ALL INTERACTION TERMS

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

DON'T USE % BIKE COMMUTERS OR TRACTS, ANALYZE CORRIDORS AND INTERSECTIONS

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.

REANALYZE THE COLLINEAR FEATURES

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.

A TIME SENSITIVE MODEL

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.

FUTURE

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] (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/

SF LOVES BIKES

BICYCLE SAFETY ANALYSIS

THANK YOU!