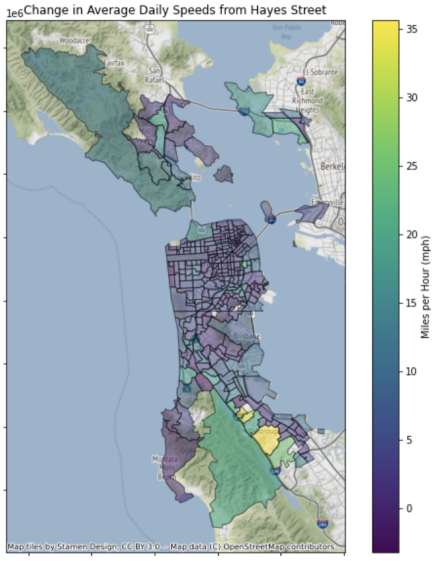
DATA100 Final Project Report

Dataset: Traffic / Lab Section: 104 / Group Members: Ariella Navarro, Chris Sung, Michael Yang

1. EDA

1.1. Visualizing the impact of lockdowns on traffic speeds

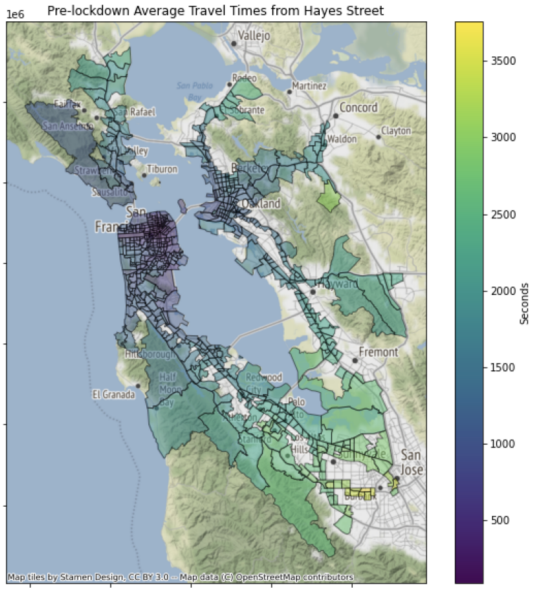
The primary data was from the Uber Movement Project, containing the daily average traffic speeds for each Open Street Maps (OSM) node in the Bay Area during March 2020 when the lockdown happened. During Part I, the OSM nodes were grouped into their respective census tracts, which are more useful for geospatial analysis given that they take into account human factors and have minimal self-variance. We then clustered the daily average speeds to census tracts, and mapped the change in speeds before and after the lockdown.



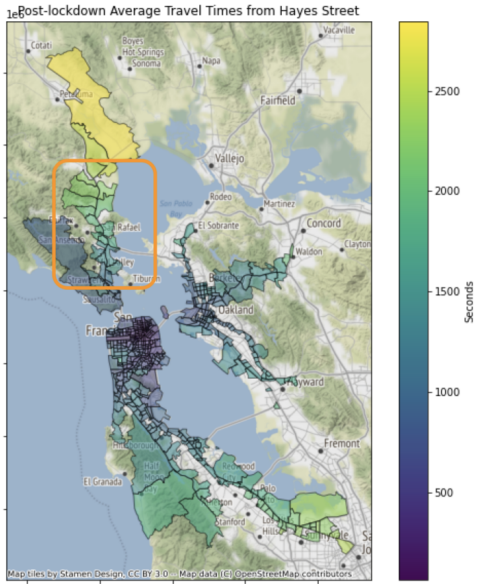
From the resulting map (on the right), we noticed to our surprise that the biggest change in average speed happened in census tracts between San Francisco and Silicon Valley, i.e., Hillsborough and San Mateo. Meanwhile, the average speeds within San Francisco had not changed significantly.

We conducted similar visualizations on average travel times of Uber trips originating from San Francisco, which also showed a similar result (as seen below): the changes were less significant for trips *within* San Francisco, and more noticeable for trips *to outside of San Francisco* (most noticeably in Marin County, marked in orange below).

Pre-lockdown



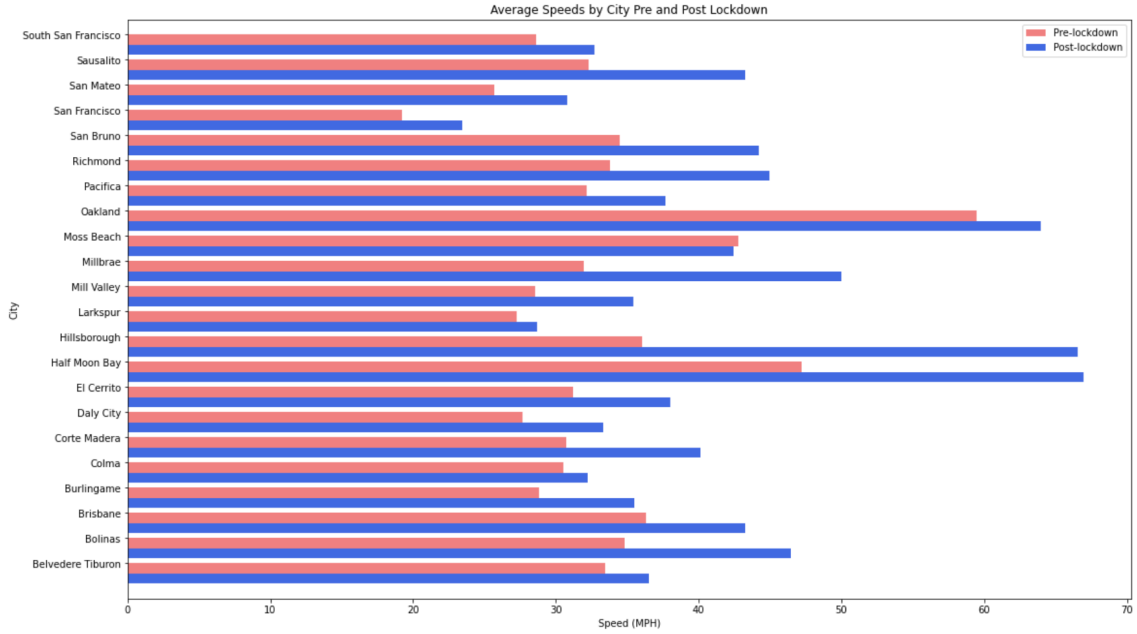
Post-lockdown

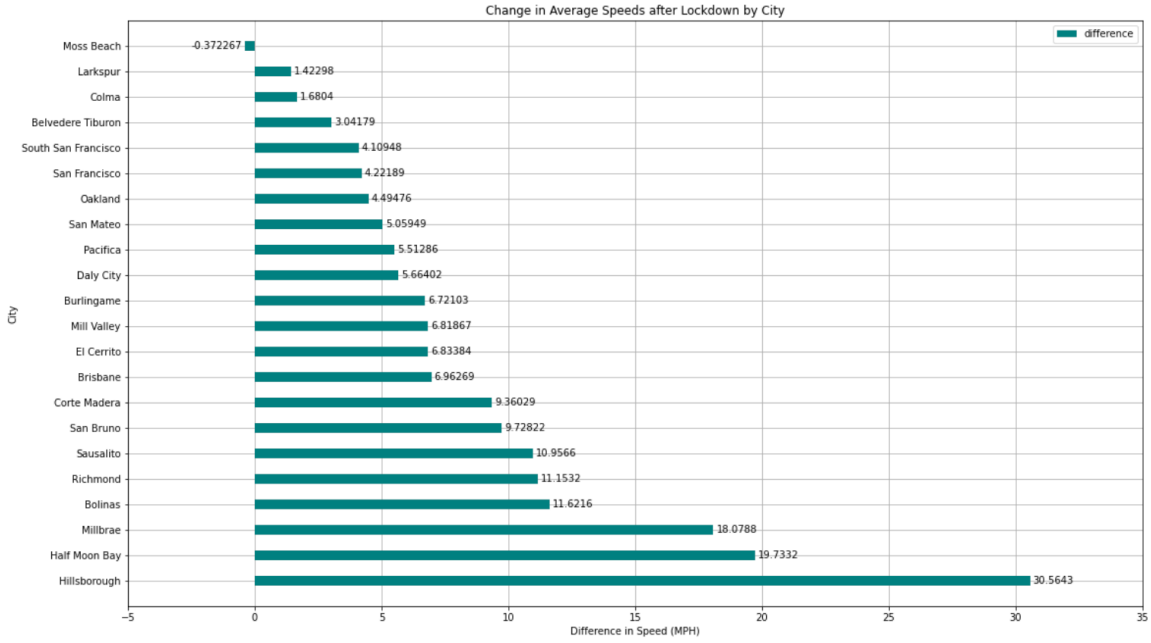


1.2. Analyzing the change in traffic speeds by city

This led us to think about the impact of the lockdown on traffic speeds in different regions of the Bay Area: ***Did the lockdown impact urban centers less than it impacted suburban areas?***

To quantify the impact in the City of San Francisco versus other areas, we grouped the census tracts and average speeds into cities, by extracting city names from the `display name` column in our original dataframe `speeds_to_tract` with average traffic speed for each census tract. **The resulting plot of average speeds pre- and post-lockdown by cities show that indeed San Francisco experienced less change in average traffic speeds than most other cities around the Bay Area.** Some other cities with less impact were nearby, like South San Francisco and Colma.



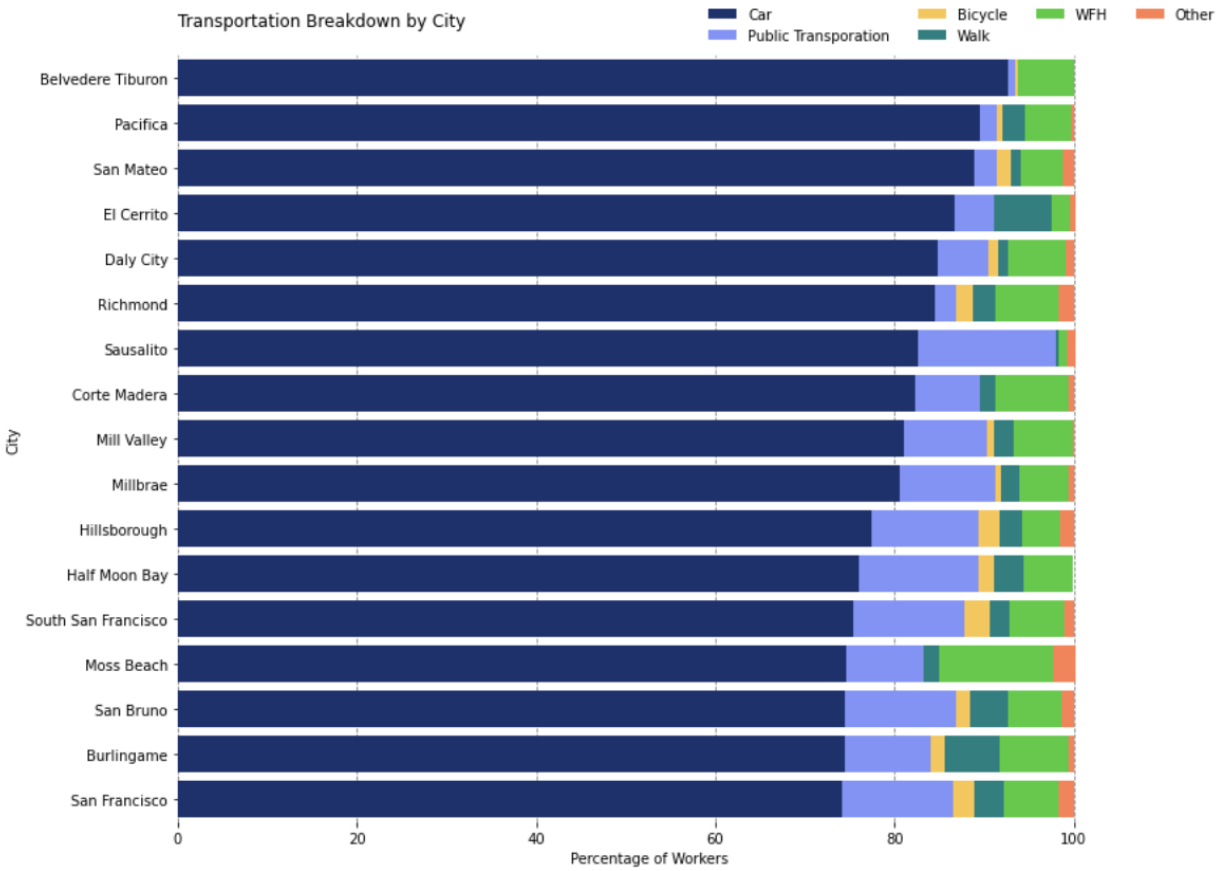


1.3. Considering the impact of means of transportations on the changes in traffic speed

The question arising from this result was: ***Why did the City of San Francisco experience less change in traffic speeds than other areas around the Bay?***

There can be numerous possible explanations to this question. One possibility is that the higher population density of San Francisco can lead to more vehicles on the streets, including buses that were still in operation after the lockdown, leading to less changes in the traffic speed.

From the 2019 American Community Survey (ACS), we obtained 1-year estimate data for the Bay area counties on "Means of Transportation to Work in California (S0802)".[1] Grouping the data by cities, the data showed that **San Francisco had the lowest proportion of cars as means of transportation and a bigger proportion of public transportation than other cities.** It is possible, then, that having a less proportion of cars and more public transportation and taxi use (coded as 'Other' below) can lead to less change in traffic speeds after the lockdown.



[1] U.S. Census Bureau (2020). Selected Means of Transportation to Work in California (S0802). American Community Survey 2019 1-year Estimates.

1.4. Asking other open-ended questions on the impact of lockdowns in traffic speeds

Yet this obviously is not the only factor that impacts the traffic speed after a lockdown. We formulated more open-ended questions as below, in relation to our primary questions of (1) *Did the lockdown impact traffic speeds in urban areas less than it impacted suburban areas?* and (2) *Why did the City of San Francisco experience less change in traffic speeds than other areas around the Bay?*:

1. *Socioeconomic Features:* What other socioeconomic features contribute to the impact of lockdowns on traffic speeds?
2. *Time Granularity & Frame:* How different are the speeds of different areas for (a) different times of the day, (b) weekday vs. weekend, and (c) after March?
3. *Regional Differences:* How does the lockdown's impact on average travel time and speed within San Francisco compare to different areas like East Bay, North Bay, and South Bay?
4. *Modeling & Optimization:* If our model has high RMSE, what statistics can we apply to understand the directionality of our error (since it is important to predict whether our predicted value should be a positive or negative change)?

In the following modeling, analysis, and further EDA, we investigate the first two questions.

For the first question, we introduce **external data from the US Census on sociodemographic factors** (i.e., population, race, school enrollment), **as well as commute-related factors** (i.e., number of vehicles and average commuting travel time) to train the model and determine whether or not these socioeconomic factors contribute to predicting the speeds.

For the second question, **we analyze whether the impact on travel speeds are different for weekends versus weekdays**, by separating the dataset into those categories.

2. Hypotheses

From the EDA, we formulated two hypotheses to test in our project.

A. The lockdown's impact on the average traffic speed was relatively small within San Francisco compared to the rest of the Bay Area

This alternate hypothesis will be accepted if the model better predicts smaller changes in speed in San Francisco than in other areas. The null hypothesis will be accepted if the changes in speed are the same for San Francisco than in other areas.

This will be confirmed by examining the errors, or the differences between predicted post-lockdown average speeds and actual post-lockdown average speeds for each census tract. If our hypothesis is incorrect or if our model underperforms, the error will have a large negative value, indicating that the predicted post-lockdown average speed for a tract was much smaller than the actual speed and that the tract was significantly impacted by the lockdown as it deviates from the model's expectations.

B. The change in average traffic speed is negatively correlated with the number of people using cars as the means of commuting

This alternate hypothesis will be accepted if the model feature for the number of people using cars is negatively correlated with the change in traffic speed. The null hypothesis will be accepted if there is no correlation between the change in traffic speed.

This will be confirmed by inspecting the model's coefficients corresponding to the number of vehicles in a given census tract. If this coefficient has a large value, it is reasonable to assume a correlation between the change in average traffic speed and the number of people using cars

Answers to these two hypotheses **can help city planners and ride-hailing companies like Uber to better plan for other lockdowns in the future by optimizing pricing and supply of drivers to urban areas** and other areas with less car usage and more public transportation and taxi usage.

3. Initial Model: Ridge Regression Model without Features

3.1. Rationale

Ridge regression was selected as the model for the experiment, since it is effective when the independent variables are highly correlated, which was discussed earlier when performing the guided EDA. Ridge regression model also introduces regularization, which reduces model complexity and helps avoid the risks of overplotting.

3.2. Design

The initial model was designed using dataframes from the guided EDA. For each census tract, the average speeds corresponding to days considered pre-lockdown (March 1st to March 13th) and post-lockdown (March 14th to March 31st) were grouped separately and averaged

respectively. We also incorporated the city name of each tract as a feature for this model to try to address our hypothesis A. Regex was used to extract the city name from a tract's `display_name` and was prepared for the model through one-hot encoding.

The objective of the model was to predict average post-lockdown speeds for each tract. To produce such a model, `RidgeCV` method from the scikit-learn library was used. We opted to incorporate cross-validation to ensure the optimal regularization parameter was being utilized. The training and test sets were produced through a 80%-20% split. The X matrix for each set contained average pre-lockdown speeds and city names, while the Y vector for each set contained average post-lockdown speeds.

3.3. Analysis

To analyze how well the model performed, we inspected the training accuracy of the model. Over multiple iterations of the model, the training accuracy ranged between 0.80 to 0.84. While this was a pretty good accuracy for a baseline model, there were several issues that needed to be addressed. Essentially only two features were being utilized (pre-lockdown average speeds and city name), thus leading to the risk that the model was underfitting. Additionally, the scope of the features used was very limited and did not provide any practical insight into sociodemographic factors that could explain the smaller change in speed in San Francisco as noticed in section 1.2. of our EDA. We decided to incorporate external datasets to expand the scope of the features and possibly gain more insight into the contributing factors.

4. Model Improvement: LASSO Regression Model with Features

4.1. Rationale

The first improvement to our model was introducing more features about the census tracts. However, we became concerned that all of the new features added to the model would result in overfitting. Additionally, not all of these features are necessarily relevant to predicting post-lockdown average speeds.

Accordingly, the second improvement to our model was opting for the LASSO regression model, which retains the regularization used in the ridge regression model, but works to eliminate features that have little to no impact on the model's predictions.

4.2. External Data Sources & Cleaning

We introduce the following several datasets from the 2020 US Decennial Census and 2019 American Community Survey to featurize for each census tracts:

- **Means of Transportation to Work:** Percent estimate of population by (a) means of transportation to work, (b) departure time, and (c) travel time.[2]
- **Housing and Occupancy**: Total estimated number of housing, occupied and vacant.[3]

[2] U.S. Census Bureau (2020). Selected Means of Transportation to Work in California (S0802). American Community Survey 2019 1-year Estimates.
[3] U.S. Census Bureau (2020). Selected Occupancy Status in California, Decennial Census, Redistricting Data (PL 94-171).

- **Population and Race:** Total estimated population and population by race.[4]
- **School Enrollment:** Total estimated population by school enrollment.[5]

We manually cleaned the dataset to make the column names more readable and exclude extraneous columns. For example, the dataset for population and race had columns for every combination of mixed races (e.g., White-Black, White-Asian, Black-Asian, etc.), and the means of transportation and school enrollment had subdivisions by gender which were irrelevant to our analysis. Additionally, columns in tables displayed as proportions were converted into raw totals to ensure all new features were displayed in a similar format.

4.3. Data Merging

Every dataset from the US Census uses `GEOID` as the unique ID for each tract, which is a concatenation of codes for the state, county, tract, and block.[6] However, our primary dataframe (`speeds_to_tract` from Part I with average daily traffic speeds for each tract) has arbitrarily numbered `MOVEMENT_ID` instead of `GEOID`, which both describe a singular tract in different formats.

In order to join our external data to the primary dataframe, we use the shapefile from the U.S. Census Bureau, which has `GEOID` and `Latitude` and `Longitude` for each census tract, as the key.[7] We spatially join the latitude/longitude coordinates in the shapefile to the multipolygons for each `MOVEMENT_ID` (Census Tracts) in the primary dataframe, and then join the shapefile and external datasets on the `GEOID`s, allowing for the new features to be described by `MOVEMENT_ID`.

During the process, 55 out of 285 census tracts in our original dataframe were dropped. This occurred most likely during the process of spatially joining, as the single latitude/longitude coordinates from the shapefile may not have aligned with the multipolygons in the original dataframe. This is especially likely for very small census tracts. This is a relatively small number of census tracts and will not have a significant impact in our model training.

[4] U.S. Census Bureau (2020). Selected Race in California, Decennial Census, Redistricting Data (PL 94-171).
[5] U.S. Census Bureau (2020). Selected School Enrollment in California (S1401). American Community Survey 2019 1-year Estimates.
[6] U.S. Census Bureau (2019). TIGER/Line® Shapefiles Technical Documentation, p. 23.
[7] U.S. Census Bureau (2020). 2020 TIGER/Line Shapefiles: Census Tracts.

4.4. Design

The model takes as inputs the expanded dataframe, with average traffic speeds and external features for each census tract. It outputs the predicted average post-lockdown traffic speed for each census tract.

We create 100 models by splitting the dataset randomly 100 times, taking 85% for training sets and 15% for test sets. All the results are then averaged within San Francisco and non-San Francisco groups.

The model takes in the following data as features:

Variable Name	Data (for each census tract)	Rationale
'speed_mph_mean_pre'	Average daily traffic speeds before lockdown	This is the primary feature to describe the prediction
'San Francisco'	Categorical variable (1 for SF and 0 for non-SF)	This is the categorical variable to group our predictions for SF and non-SF
'Commute|Car'	Est. # of people using cars to commute	More cars for commuting means more traffic in general, thus may relate to a smaller increase in speed.
'Commute|Car_Carpool'	Est. # of people carpooling to commute	More carpooling for commuting means more traffic in general, thus may relate to a smaller increase in speed
'Commute|Bicycle'	Est. # of people using bicycles to commute	More bikes for commuting suggests less car traffic, thus may relate to bigger change in speed
'TravelTime|AvgMins'	Est. Avg # of minutes for commute	Longer travel time means more traffic in general, thus may relate to a smaller increase in speed
'TravelTime|<10min'	Est. # of people with commutes less than 10 minutes	More people with short commuting time may relate to less use of cars, thus a bigger increase in speed
'Vehicles|3+'	Est. # of people with more than 3 cars	More people with more cars suggests more traffic in general, thus may relate to a smaller increase in speed
'H_Total'	# of housing	More housing may be related to the population density and demographics
'Est|K12',	Est. # of people enrolled in K-12	More people in K-12 school may relate to how 'suburban' the tract is, with more speed restrictions, thus a smaller change in speed
'Est|TotalInSchool'	Est. # of people enrolled in school (K-12, college, and graduate)	More people in school may relate to less people commuting for work, and also more speed restrictions around school zones, thus less change in speed
'Mixed'	Est. # of people who are mixed race	This is a unique feature about the racial demographics in the tract which increased the model accuracy
'Other'	Est. # of people who are of a race in the 'other' category	This is a unique feature about the racial demographics in the tract which increased the model accuracy

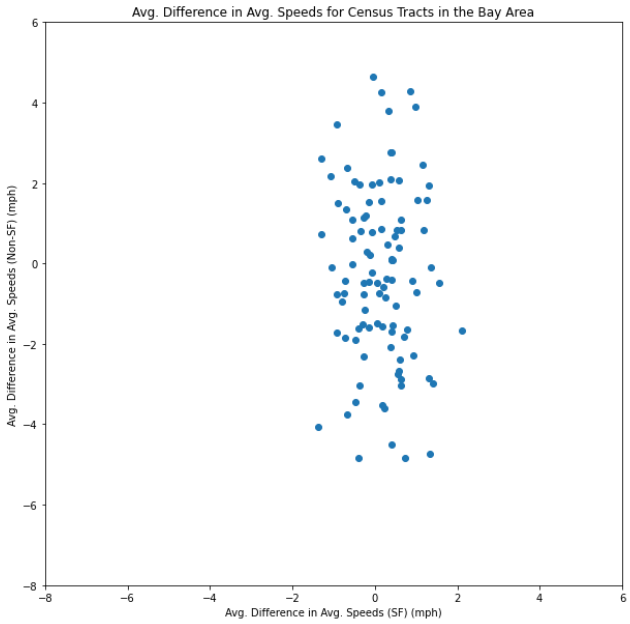
4.5. Evaluation & Analysis

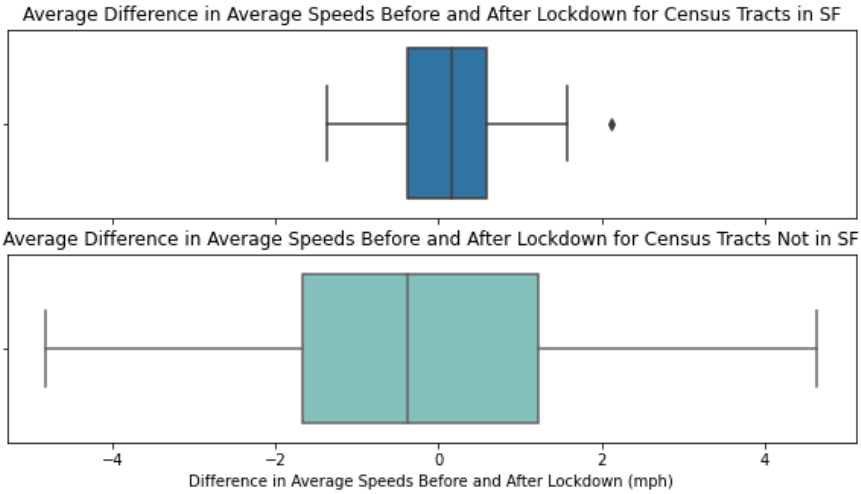
The range of training accuracy **for the model was 0.82~0.85**, which is an improvement from the initial ridge regression model's range. The range of **RMSE was approximately 3~6 mph.** While this high range of RMSE might be a cause for concern in regards to the accuracy of the model, our justification of hypothesis A relies on the residuals of each tract. Therefore, the range of RMSE is acceptable. Additionally, the city name features present in the ridge regression model were removed from the lasso regression model because they consistently illustrated no weight on the predictions computed by the model. `San Francisco` was kept solely to divide the data into San Francisco and non-San Francisco groups.

Hypothesis A: The lockdown's impact on the average traffic speed was relatively small within San Francisco compared to the rest of the Bay Area

The scatter plot of average difference in traffic speeds pre- and post-lockdown for San Francisco (SF) and elsewhere (non-SF) show that the errors had a narrower range for SF (narrow horizontal spread) and a wider range for non-SF (broad vertical spread). The box plots also show that the distribution of these differences are much narrower for SF and broader for non-SF.

Therefore, we accept our first alternate hypothesis as the average of predicted difference pre- and post-lockdown traffic speed in San Francisco has a much smaller range of approximately -2 to 2 mphs, compared to other areas in the Bay Area with a range of -4 to 4 mph.





Hypothesis B: The change in average traffic speed is negatively correlated with the proportion of the population using cars as the means of commuting

The correlation coefficients for each featurized parameters sorted in terms of absolute value are below. Note that these values are approximate and variable depending on the random training and test data splits when the model iterates 100 times.

Variable Name	Data (for each census tract)	Approx. Corr Coefs
'speed_mph_mean_pre'	Average daily traffic speeds before lockdown	1.40E+00
'TravelTime|AvgMins'	Est. Avg # of minutes for commute	-2.07E-01
'Commute|Bicycle',	Est. # of people using bicycles to commute	7.86E-03
'TravelTime|<10min'	Est. # of people with commutes less than 10 minutes	-5.66E-03
'Est|K12',	Est. # of people enrolled in K-12	4.55E-03
'Mixed'	Est. # of people who are mixed race	-4.06E-03
'Other'	Est. # of people who are of a race in the 'other' category	-2.42E-03
'Commute|Car'	# of people using cars to commute	-1.70E-03
'Vehicles|3+'	Est. # of people with more than 3 cars	1.51E-03
'H_Total'	# of housing	1.24E-03
'Commute|Car_Carpool'	Est. # of people carpooling to commute	1.13E-03
'Est|TotalInSchool'	Est. # of people enrolled in school (K-12, college, and graduate)	5.28E-04
'San Francisco'	Categorical variable (1 for San Francisco and 0 for elsewhere)	0.00E+00

We accept our second alternate hypothesis as our independent variable, the proportion of the population using cars to commute (`Commute|Car`), had a negative correlation coefficient of approximately -0.0017. While it is not the most significant factor among our features, the proportion of the population commuting in cars is negatively correlated with the change in speed after the lockdown. In other words, when more people in an area used cars for commuting, the traffic speed (as detected by the data from Uber rides) was slightly slower after the lockdown, possibly because there was more commuting traffic even post-lockdown.

Notably, a significant feature for predicting the traffic speeds after the lockdown was the average travel time for commuting, with a negative correlation coefficient of -0.207. This suggests that when the average travel time for commuting was longer, the traffic speed was slower after the lockdown.

5. Future Work

Extending Features: What other factors impact traffic speeds in lockdowns?

The internal accuracy of the model can be improved by investigating more features to be introduced to find better predictors of traffic speeds before and after the lockdown. These may be types of businesses in the area, figures on tourism, median age, household income, and population density, as these are all factors that may affect traffic patterns. One approach to identifying potential features could be comparing the characteristics of cities with low changes of average speed (e.g., San Francisco, Colma) to cities with high changes in average speed (e.g., Half Moon Bay, Hillsborough).

Expanding Scope: Are these factors shared among all cities?

The external validity of our model for predicting the impact of lockdowns in traffic speed can be tested by training the model to other metropolitan areas, like New York and Chicago. As the dataset is available from Uber Movement and the US Census, the model could easily be trained in a bigger scope of data. It is possible that the same features in our model cannot accurately predict the impact in other cities, and therefore leading to further investigation of different factors to improve our model.

Absolute vs. Relative: How do the absolute or relative values change the accuracy?

The features in our model have both absolute and relative or proportionate values. It is an open question whether having all absolute values or all relative values would improve or worsen the accuracy of our model.

Reducing Data Loss: How can we improve the data processing?

Improvements can be made in reducing data loss during the merging of our datasets by finding a better way to convert Movement ID to Census Tracts. This would prevent data from being dropped as a result of spatial joins and mismatched coordinates, as discussed in detail in *Data Merging* in section 4.

Increasing Granularity: Could the model predict speeds for smaller neighborhoods?

Alternatively, the model could take more granular data than census tracts, such as Google Pluscodes or OSM Nodes. This could allow more fine-grained analysis into different neighborhoods within the city, not just the city as a whole.

Taken together, these considerations could improve and expand the model to be used for different cities and smaller neighborhoods, and help us better understand the human factors that impact traffic before and after lockdowns. Such a model could help city planners, public health officials, and transportation companies to better plan and respond to other possible lockdowns in the future.