Open-Ended Modeling Report

Erin Bhan, Arpna Ghanshani, Vidushee Mishra, Anika Tyagi

(Optional) Open-Ended EDA

Background on Data

The primary data given for this project was the Uber Movement data from March 2020 in San Francisco. The dataset was spatially grouped in two different ways. First, the data was partitioned by dividing the work into uniform slices with the Google Plus Codes Model. Then, the data was spatially partitioned by dividing the area according to population, using census tract information. In the data, we were provided with Open Street Maps (OSM) and Census Tract data. The OSM data provided us with nodes and ways. These IDs allowed for easy identification of streets and routes in the traffic speeds dataset. The census tract information geographically divided the area of interest - the county of San Francisco - according to 2010 US Census information. This allowed for easy identification of regions of differing travel times based on population. The first dataset contains information on 1586652 unique Uber trips during the month of March 2020, with features including OSM ID start and end nodes, the mean speed during the trip, and the day that the trip occurred. Additional datasets with information on San Francisco census tracts and GPS locations allowed us to also see where these Uber trips started and ended by breaking up geographies into plus code regions.

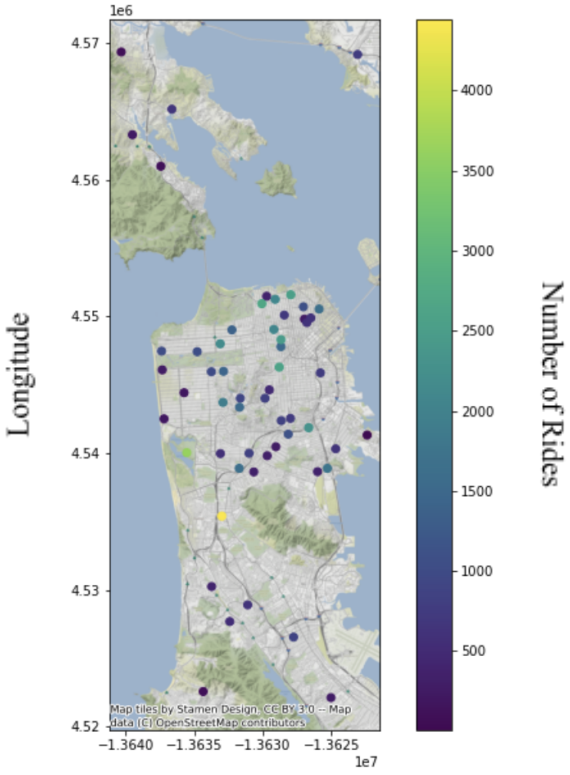
EDA Performed

The exploration we did in Section 3 Open-Ended EDA was looking at which destination plus code regions of trips from Hayes Valley experienced the biggest change in the number of rides taken pre vs post-lockdown. We hoped to visualize, on a heat map, which plus code regions experienced the highest change in the number of rides in order to see which regions were most affected by the stay-at-home orders and experienced less traffic. Thus, we hypothesized that the number of rides taken from Hayes Valley to most plus code regions clustered around downtown SF would decrease due to the decrease in people needing to get into the city for work and other activities.

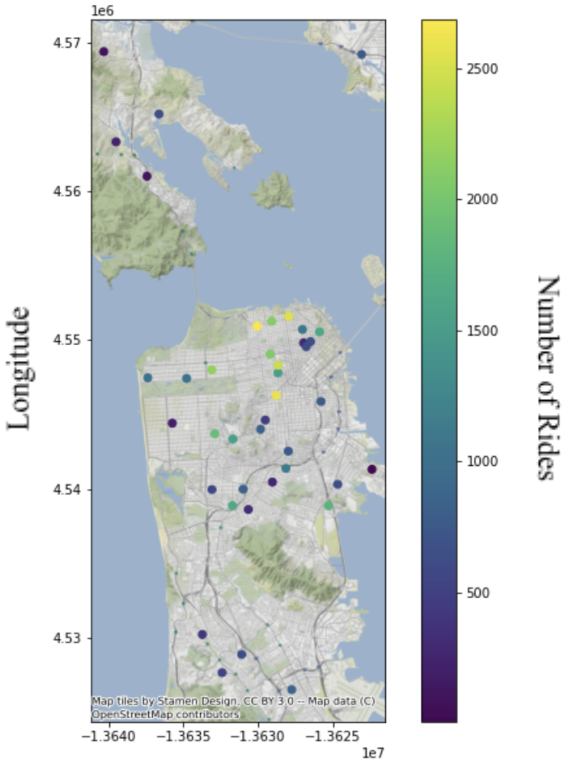
This hypothesis is supported by our heatmaps below, because the number of rides from the Hayes Valley to most plus code regions clustered around downtown SF decreased, as well as freeways going in and out of the city (dots got darker indicating fewer rides taken). However, there are also regions where the number of rides stayed the same, as shown by the data from the south of the city, with the plus codes around Golden Gate Park, hence not supporting our hypothesis. Finally, we observed one plus code that experienced a huge increase in the time it

takes to travel there from the Hayes Valley (this is the yellow plus code region in the right heatmap), and in our open-ended analysis, it would be interesting to see if there is any particular reason why this occurred or if it is just an outlier.

Pre-Lockdown Number of Rides



Post-Lockdown Number of Rides



Hypothesis for Part 2

Based on our previous exploration, we have a reasonable idea that the number of rides mainly decreased in dense areas when less traffic was on the road. The changes in speed that we were able to visualize in the first part of the project led us to question what other factors were also impacted by less traffic being on the road. In particular, did the frequency and clustering of car accidents change drastically pre vs. post lockdown? This question is especially interesting to explore given the panic that occurred when lockdowns were announced with many people frantically moving throughout the city to stock up on groceries and supplies. Breaking down our data to not just pre vs post lockdowns, but also looking at the frequency of car accidents the day the lockdown was announced could yield some interesting results. Additionally, splitting our data based on different time intervals, such as looking at weekdays vs weekends when considering the frequency of car crashes may help us better understand when these accidents occur and along which routes.

We found car accident data, with features like the severity of the accident, the longitude and latitude data of where the accident started and ended, the time when the accident occurred, and more for all car accidents in the United States from February 2016 to now. This dataset is US-Accidents: A Countrywide Traffic Accident Dataset from a data scientist at Lyft. Limiting this data to accidents that occurred in San Francisco during March 2020 will allow us to compare this data with our given Uber trip data.

DataSet: https://smoosavi.org/datasets/us_accidents

Using this data, we can hypothesize that the number of rides decreasing post lockdown positively correlates to a decrease in car accidents. The US-Accidents dataset along with the Uber trip data will also allow us to explore the regions where car crashes were more concentrated before and after lockdowns and compare the March 2020 car crash data to the March 2020 Uber trip data.

Questions for Further Open-Ended EDA in Part 2

1. Were car crashes more concentrated before or after lockdowns? If so, are there specific areas that are more concentrated than others?
2. Did the panic that occurred when lockdowns were announced affect traffic speeds?
3. Is there any correlation between traffic speeds and car crash concentrations?
4. Due to the panic at the beginning of the lockdown, did the number of car crashes increase, decrease, or stay the same?
5. How did the number and concentration of car crashes change pre and post-lockdown?
6. Where were car crashes more concentrated? Is there any correlation between the population density of those areas and the concentration of car crashes?

Problem

We revisited our hypothesis in the previous part of the project, and have decided to revise it to explore the following:

We hypothesize that we can more accurately predict the mean travel time of Uber rides taken in census tracts across San Francisco by using features such as frequency of car crashes, the severity of crashes, temperature, distance, time of day, and more.

By utilizing the Uber traffic dataset and the Lyft car crash dataset, we will be able to confirm this hypothesis if we see a positive correlation between the decrease in car crashes and the decrease in travel times, and we will reject the hypothesis if this is not the case. This question is particularly interesting as no other dataset with this direct correlation already exists, and many of

the features we are considering have varying degrees of impact on the likelihood of car crashes and hence average travel times. Factors such as whether it is day or night or the severity of crashes can be reasoned to increase likelihood of car accidents occurring, but also could affect the amount of traffic on the road and potentially decrease the likelihood of an accident. We hope to explore these questions and more through our EDA.

Modeling

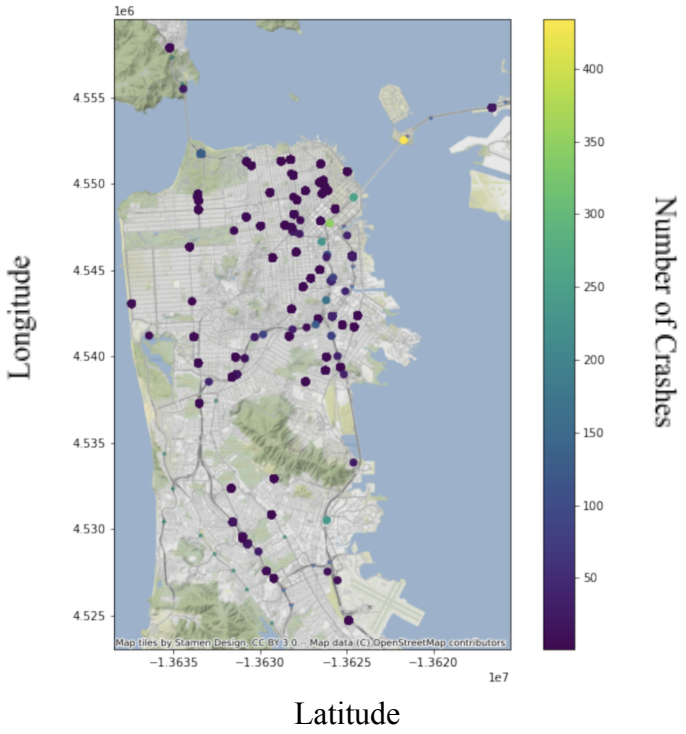
Baseline Model, Analysis, & Evaluation

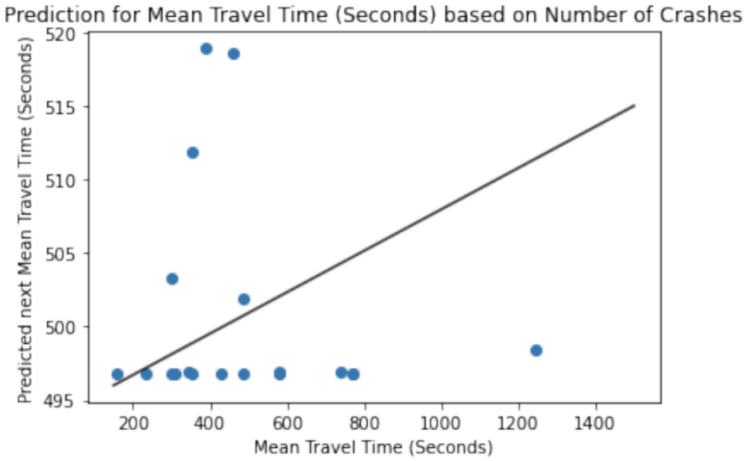
When building our first model to explore our hypothesis, we hoped to utilize a linear regression model to explore car crash frequency and travel times across SF. After training our data on this model, we hope to be able to predict mean travel times in order to best give Uber more accurate information on how long rides would take in different regions of SF. Inputs to this model are the frequency of car crashes and the output would be a prediction of mean travel time based on the aforementioned features.

Our reasoning for selecting this model over another one is that we would like to see if we can directly find a correlation between a decrease in car crashes and a decrease in mean travel times. Our first model utilized data from the whole month, in order to see if there is a correlation between car crashes and mean travel times both pre and post-lockdown. We chose to ignore the change point in this model because we hypothesized that we could linearly correlate crashes directly to travel times.

In order to achieve our model, we first had to process this additional dataset on car crashes across the US since 2016. First, we limited our data to only crashes that occurred in the city of ‘San Francisco’ or ‘South San Francisco’ and then limited to only crashes that occurred in March 2020. Additionally, this dataset had to be merged with our other data on Uber rides on census tracts, which was possible to do by utilizing Longitudes and Latitudes of census tracts and crashes, and grouping on this geographic level to get the frequency of crashes in each tract. We geographically plotted the car crash data on each census tract to give us an idea of how many accidents had occurred in each region.

Heat Map of Number of Crashes by Movement ID





After cleaning out data, training it on a linear regression model, and scoring it, we found an R^2 score of -0.0114. This is indicating that there is a negative correlation between mean travel time and car crashes. This tells us that when travel times are shorter between locations, speeds must be higher in order to decrease the time. Hence, we can say that shorter travel times are negatively correlated with more car crashes.

While this R^2 supports our hypothesis, the strength of this correlation is not particularly strong, and we cannot make a decisive conclusion that this is necessarily always the case. This can be

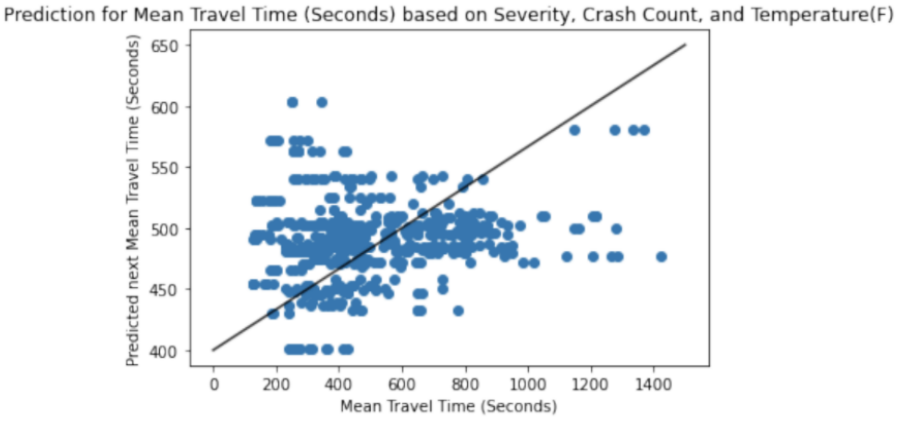
seen in the scatter plot of Predicted Mean Travel Time (seconds) based on Number of Crashes as the data is widely scattered and does not have any clear linearity. Moreover, the test RMSE of this model was quite high at 248.64, indicating that it is severely overfitting. This is likely because our car crash dataset was quite small, when there is a limited amount of data to train on it is likely to fit tightly to the training/test datasets on which our model was based. So, in our improvement of the model, we added other additional features so that the data would be more generalizable.

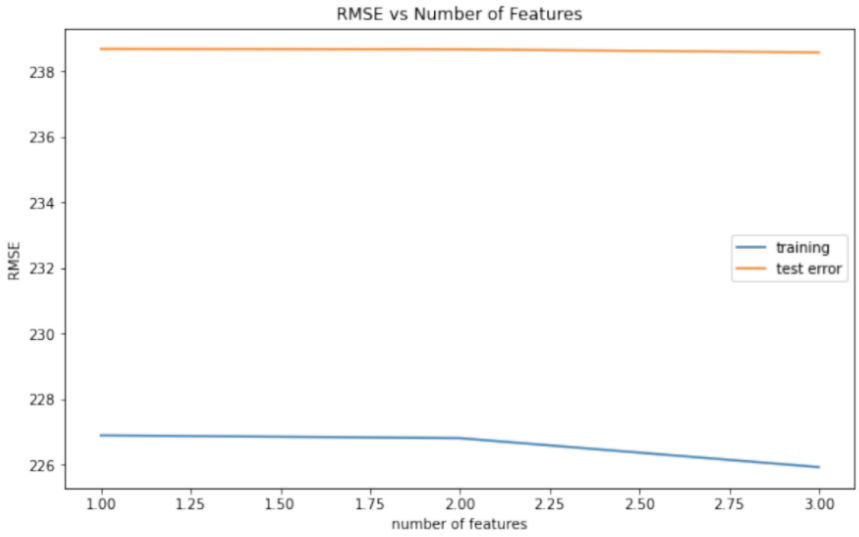
Improved Model 1 Analysis & Evaluation

In order to improve upon our first model, we decided to utilize feature engineering and conduct a multiple linear regression in order to more accurately predict mean travel times based on the frequency of crashes, the severity of crashes, and temperature. Our logic for including these features is as follows. We expect that more severe crashes would occur when there is more traffic on the road and more cars have the potential to be involved, leading to larger street closures and thus increasing travel times. We expect that when temperatures are higher, more people would be driving or taking an uber to avoid the bad weather, thus increasing traffic and likelihood of crashes and therefore mean travel times as well. Weighing all of these factors, we hope to see better predictions of mean travel time with our model.

We retrained the model on our new dataset and completed multiple linear regression to predict mean travel times, and plotted how our RMSE shifted with more features being added. Below we can see that unfortunately, our RMSE was still quite high regardless of the number of additional features, so this caused us to further consider what features would be more meaningful in our next improvement of our model.

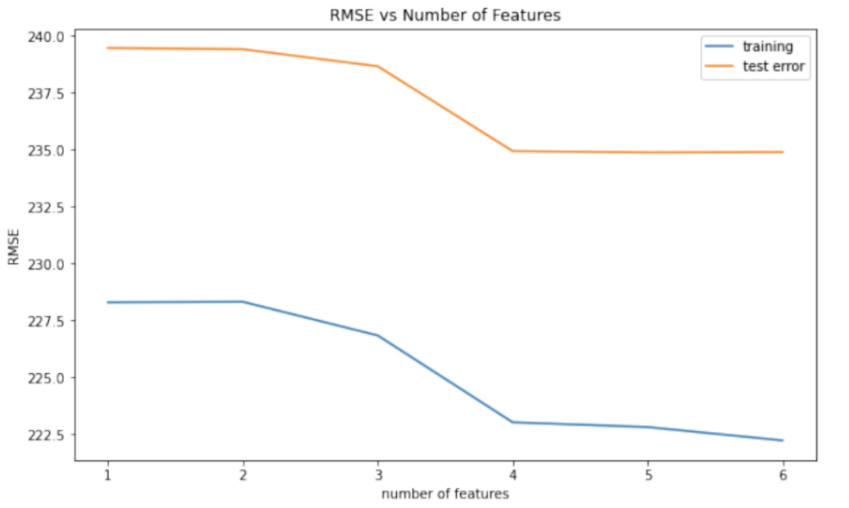
One benefit of this model, however, was the increase in the size of the data as can be seen by the scatter plot below. In order to preserve more data so that we didn't overfit our model, but hopefully decrease our RMSE with more features, we trained our multiple linear regression model on additional meaningful features in our final improvement.





Improved Model 2 Analysis & Evaluation

In the first improved model, we improved the baseline model by adding the severity of crashes and temperature features in our prediction of mean travel time. While the inclusion of more features in our multiple linear regression model slightly improved our baseline model, we felt as though there were not enough significant features and data to base our mean travel time predictions on. Thus, in order to improve this model, we identified certain features that we felt would provide a significant improvement to the model: whether it was day or night. Our logic for the inclusion of the time of day features is that we expect there to be more crashes at night, due to low visibility and an increase of exhaustion in drivers, thus correlating to higher travel times. With the inclusion of the time of day feature, we expected to visualize a more accurate prediction of mean travel time.



As shown in the model above, with the increase of the number of features through the inclusion of the time of day data, we were able to identify a decrease in our RMSE. With the inclusion of the day and night features to the model, we are predicting the model on six total features and we get a training RMSE of 222.208 and a test RMSE of 234.890. In comparison to the baseline model test RMSE of 248.64, the test RMSE of the improved model shows a significant decrease. In comparison to the first improved model, which had a test RMSE of 238.559, we can observe another slight decrease in RMSE. A lower RMSE correlates to a higher accuracy of data prediction for our multiple linear regression model. Thus, the decrease of RMSE in relation to the increase of the number of features is proof of an improved accuracy for the mean travel time predictions of our multiple linear regression model.

Answer

Based on the fact that our final model, which we created by adding in additional features to our baseline model, has a very weak positive correlation between mean travel time and predicted mean travel time, we cannot confirm our hypothesis. Although this model has a lower R^2 value than the previous ones that we created, the final value is still very high and our models are unable to show significant relationships between factors such as frequencies and severity of car crashes and decreased traffic times and slower average speeds, preventing us from proving the hypothesis correct.

Future Work

In terms of our current modeling research to examine the correlation of mean travel time and car crash data, we could improve our multiple regression models in the future through the inclusion of car crash data from many months before and after the lockdown. Currently, our models are

predicting the mean travel times of data from March 2020. We opted to only include the data of March 2020 in order to observe the difference of travel times in terms of car crash data before and after lockdown. However, since we are predicting travel times in terms of only one month of the year, there is not a significant amount of data to base our predictions off of. Through the inclusion of more months of car crash data, both before and after lockdown, we would be able to base our predictions off of more data of significance. This would allow for a more accurate and efficient multiple linear regression model.

To continue the work done in this project, we would also be interested in exploring the effect of the lockdown on the frequency of accidents in different regions. By studying Uber rideshare data, we found that traffic and the number of people taking trips decreased when the lockdown went in place, likely due to concerns about public health and the fact that their traveling needs were significantly reduced after offices and schools went remote. By using more detailed accident data, which contains information about what types of locations are the ones that have more severe and/or frequent accidents pre and post-lockdown, it will be interesting to see if these changes caused different levels of effects in varying types of regions such as downtown cities, suburban/residential areas, commercial regions, etc.

Not only that, but given the fact that many users utilize ridesharing or cars, in general, to travel to work in more urban areas, it will be intriguing to see the effects of remote work after the lockdown on the number and severity of accidents in these regions, especially. If there is more of a decrease in accidents in these areas than others, we can also further explore if this trend contributed to a decrease in traffic that is higher than what would have just happened due to decreased travel on its own.

Such data regarding the magnitude of changes in traffic that come from fewer cars on the road versus a reduction in severe accidents overall can be useful for both rideshare companies like Uber to change pricing models in certain areas to maximize profits while in traffic and also for urban development authorities in the government to improve infrastructure in accident-prone regions. Most importantly, analyzing changes in traffic and accidents based on the region can also boost our model's accuracy by helping us account for the time that would be spent in traffic and the amount of traffic in various regions, and how these factors would change with future lockdowns.