Project Design Document - Traffic Dataset

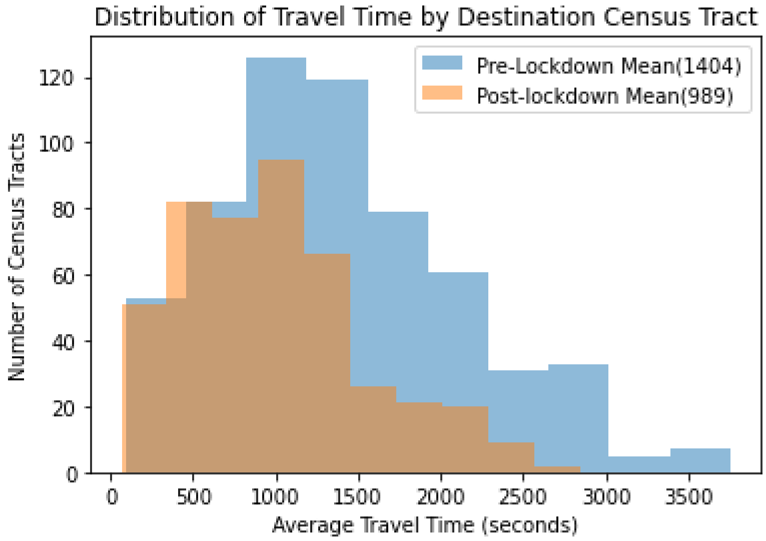
Cary Kuang, AJ Kumar, Raj Bhanushali, David Siu

Data

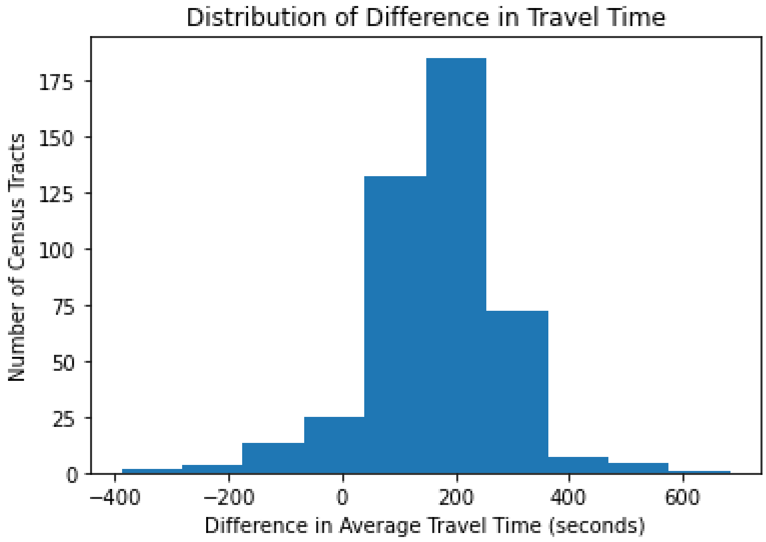
Our data is based on Uber's traffic speed data. We were initially given a dataset that provided the average speed for uber vehicles in mph for a specific road segment in San Francisco, with each row representing a unique road. With the latitude and longitude for each starting node, we were able to group rows into two main spatial categories (Google plus codes, Census tracts), thus clustering roads based on spatial categories allowing us to explore the data based on these new classifications.

EDA

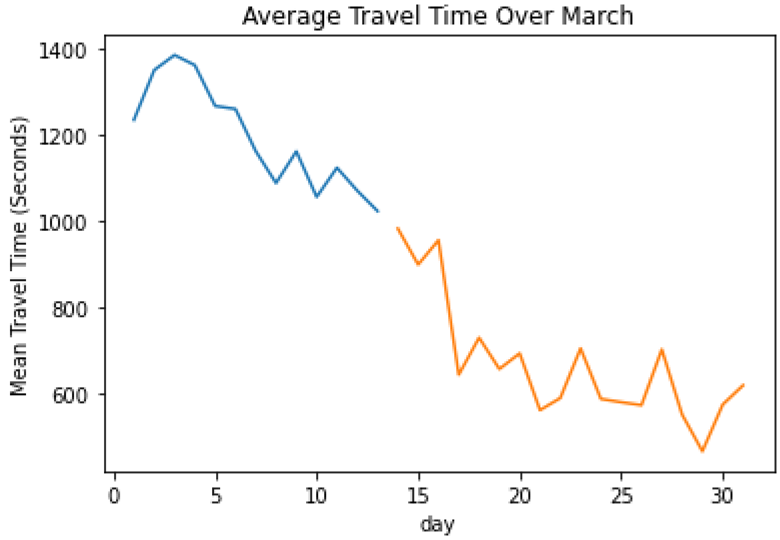
- Data
  - We performed EDA on two loaded data sets: daily travel times from Hayes Valley to all other census tracts around San Francisco and daily travel times from 300 Hayes St. to Golden Gate Park in San Francisco.
- EDA Workflow
  - In order to delineate the effect of lockdown on travel time, we split the data into pre and post lockdown by the lockdown date, March 13th, 2020
  - We compare the mean travel times pre and post lockdown by taking corresponding means of the mean travel time when the data is grouped by destination ID
    - We find the following distribution of mean travel times from Hayes Valley to other Destination Movement ID's pre- and post-lockdown.



    - Although fewer data points exist in the post-lockdown frame as compared to the pre-lockdown frame, there is a clear shift downwards in the distribution of mean travel times after lockdown.
- We compare the differences in mean travel times for each Destination Movement ID.
    - Differences that are positive indicate that the travel time decreases after lockdown, differences that are negative indicate that travel time increases after lockdown.



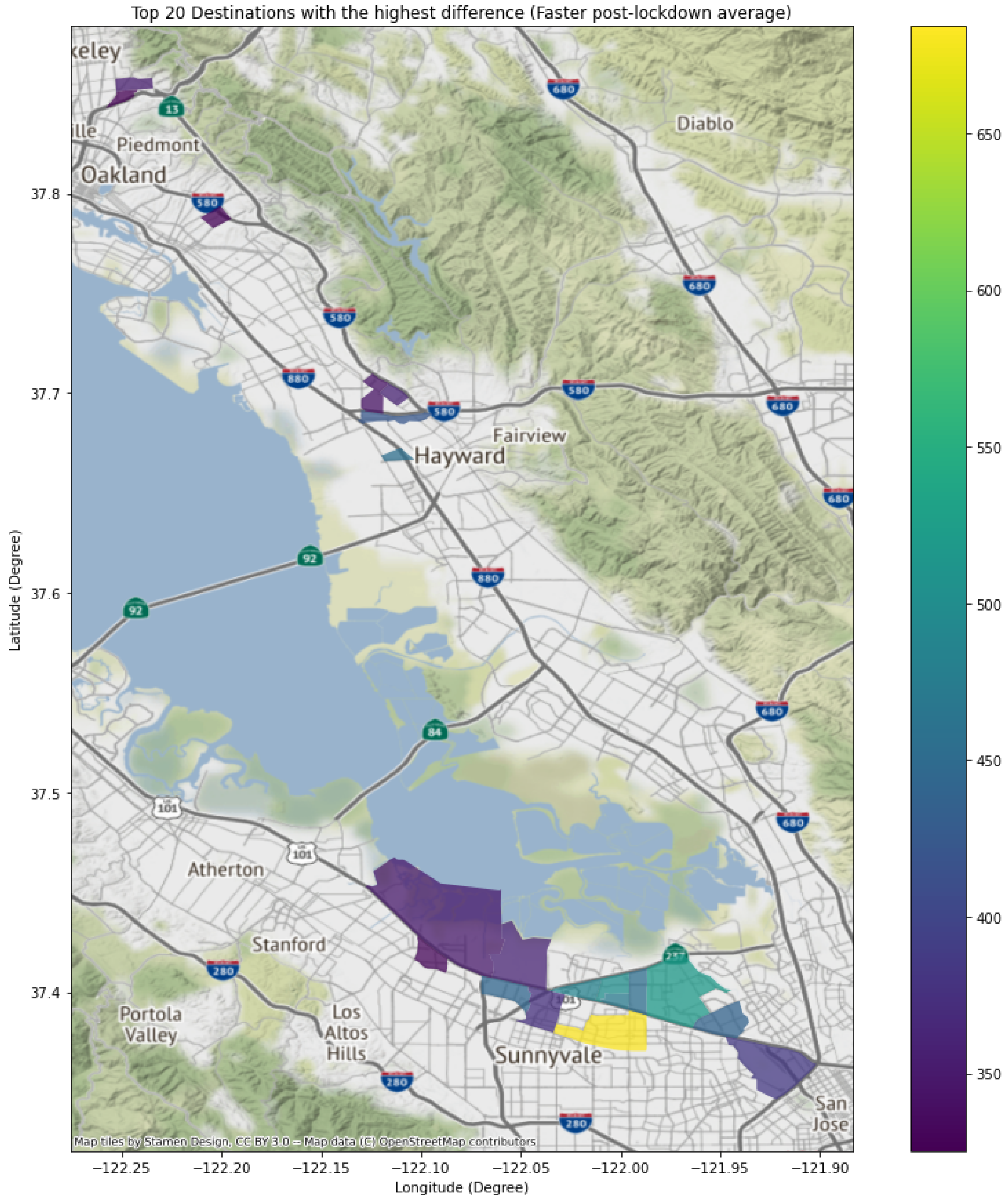
    -
    - The majority of data points lie above zero, indicating that pre-lockdown mean travel times were mostly higher than post-lockdown mean travel times.
    - We plot mean travel time on the same plot to observe the change in the trend over time.



    -
    - We observe that as time goes on, mean travel time decreases overall. Pre-lockdown, we see the mean travel times decrease, likely because

companies began to issue work from home orders before the full lockdown took effect. Post-lockdown, we observe that there is a sharp drop in mean travel times near the 15th of March, after which the mean travel time fluctuates a little bit but mostly stays low, indicating that there is a floor to the amount of decrease in mean travel time.

- We then compare the top 20 destinations with the highest difference (faster post-lockdown mean travel time)

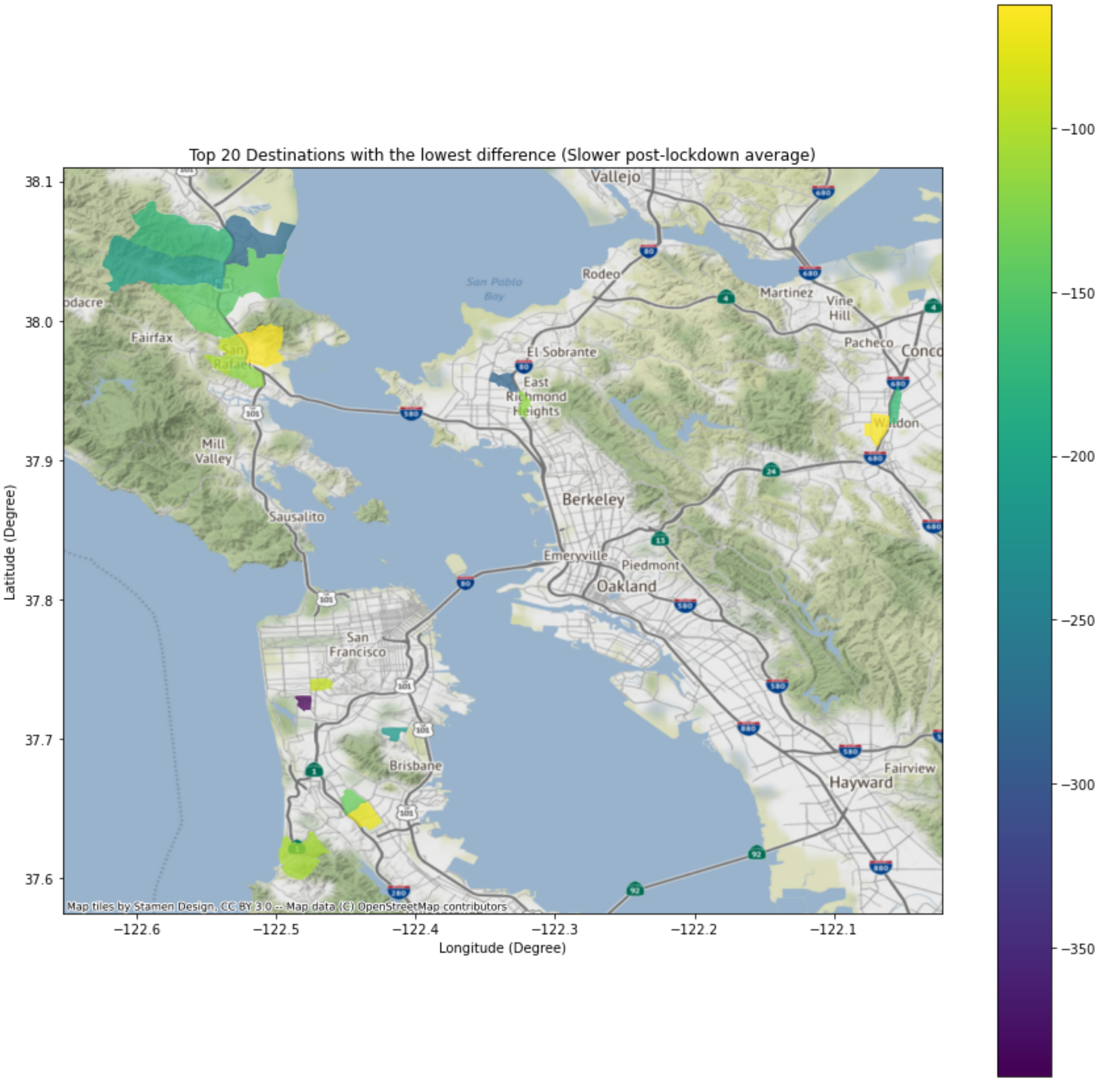


Top 20 Destinations with the highest difference (Faster post-lockdown average)

  -
  - It is observed that the biggest difference in mean travel time is to locations in the South Bay, such as Sunnyvale and San Jose. None of the top 20 destinations with the highest difference are in San Francisco. This may be due to the fact that these locations are further away from San

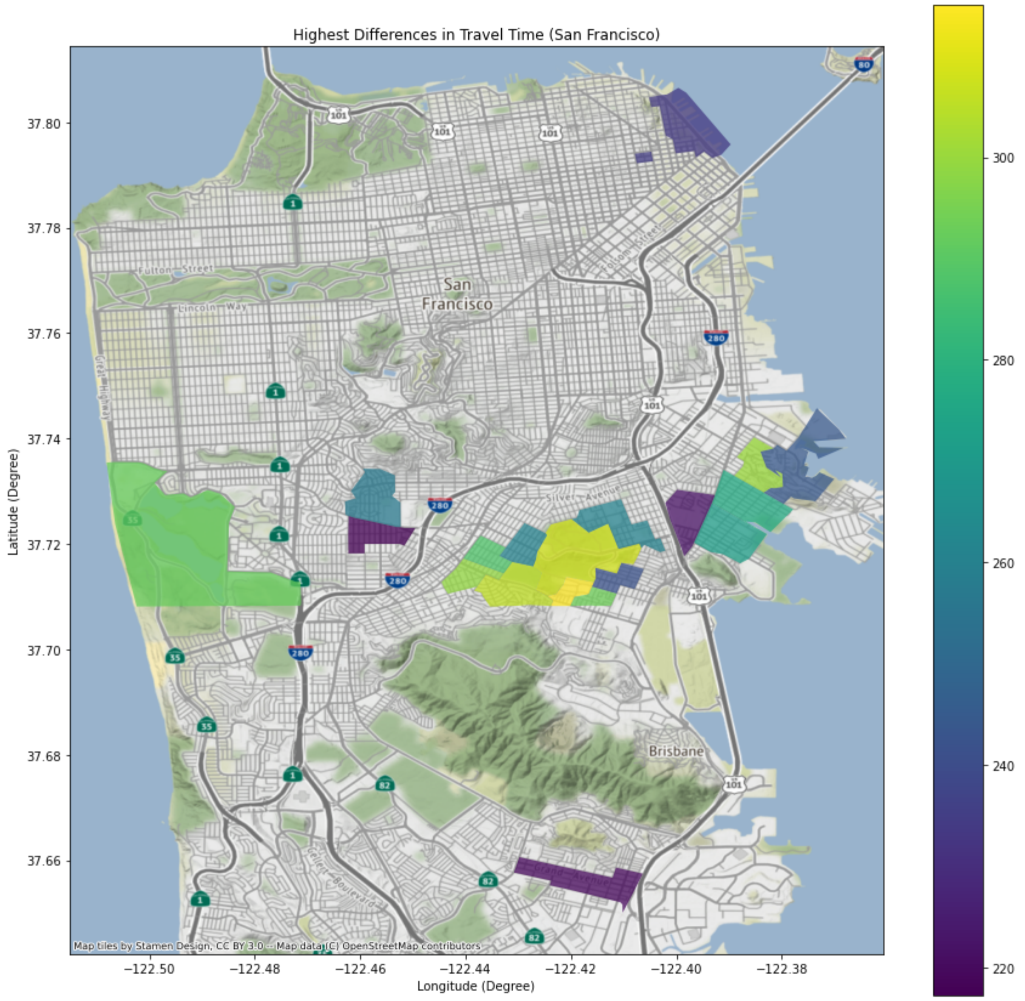
Francisco and are accessed by freeways which are more affected by lockdown status, but this does not account for variation in mean difference in regions that all appear in the South Bay, such as the difference between varying spaces in Sunnyvale.

- We then compare the top 20 destinations with greatest post-lockdown mean travel time increase.

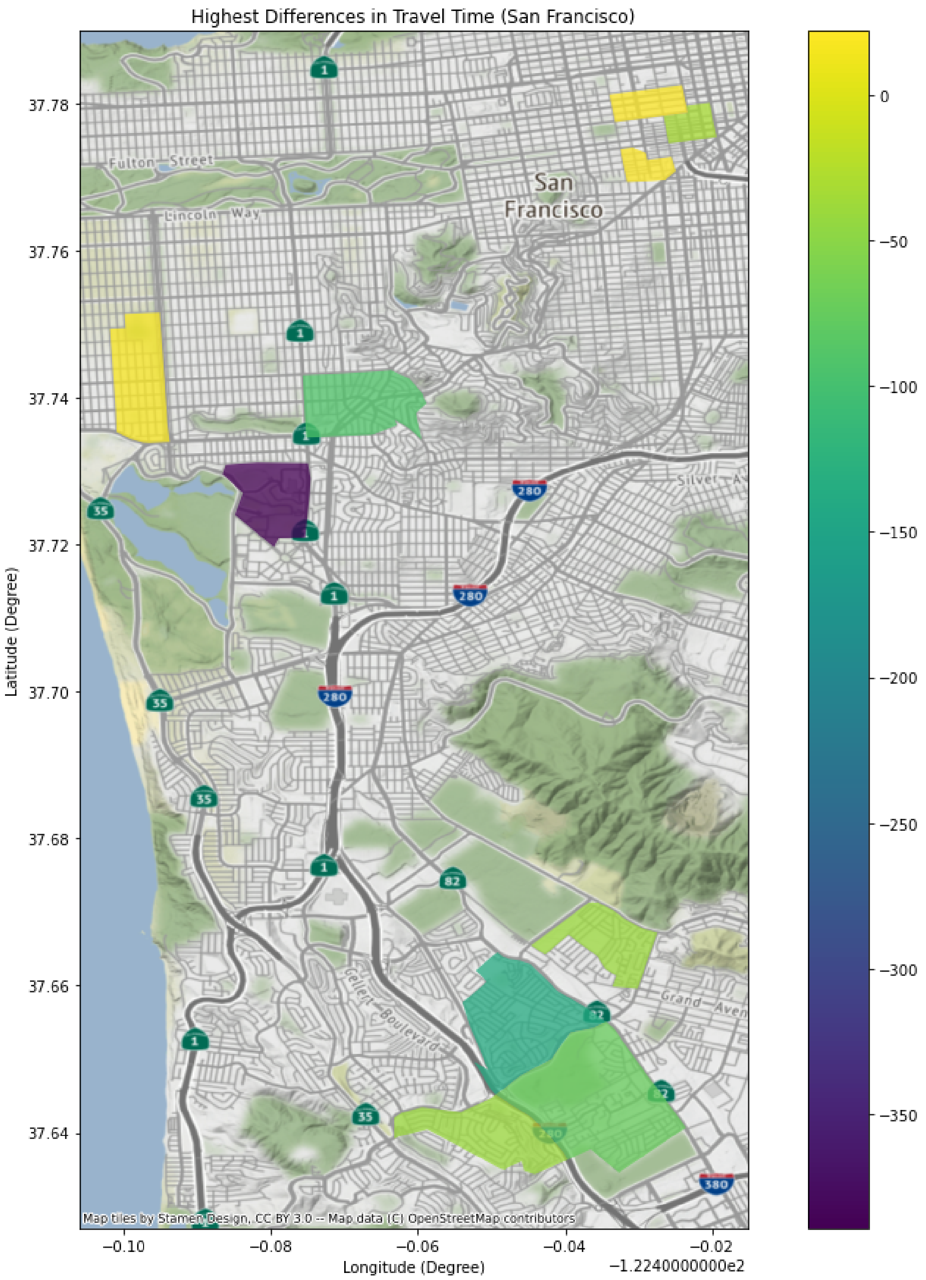


  -
  - It is observed that there is great variation in the locations that are accessed by this grouping, but it is noted that the majority of these areas are residential or off-branches of major highways. One possible explanation of this is that these are areas that are not as heavily traveled pre-lockdown, but were visited more after lockdown orders were implemented. Another possible explanation is that there is random noise in the data causing these variations.
- We then choose destination ID’s that are only in SF to decrease the scope of our data.

- We perform the same analysis of top 20 and bottom 20 mean travel times in only SF.



-
    - For locations that had the greatest difference in mean travel time, we observe that the majority of those locations are outside of the city in residential areas or near parks.

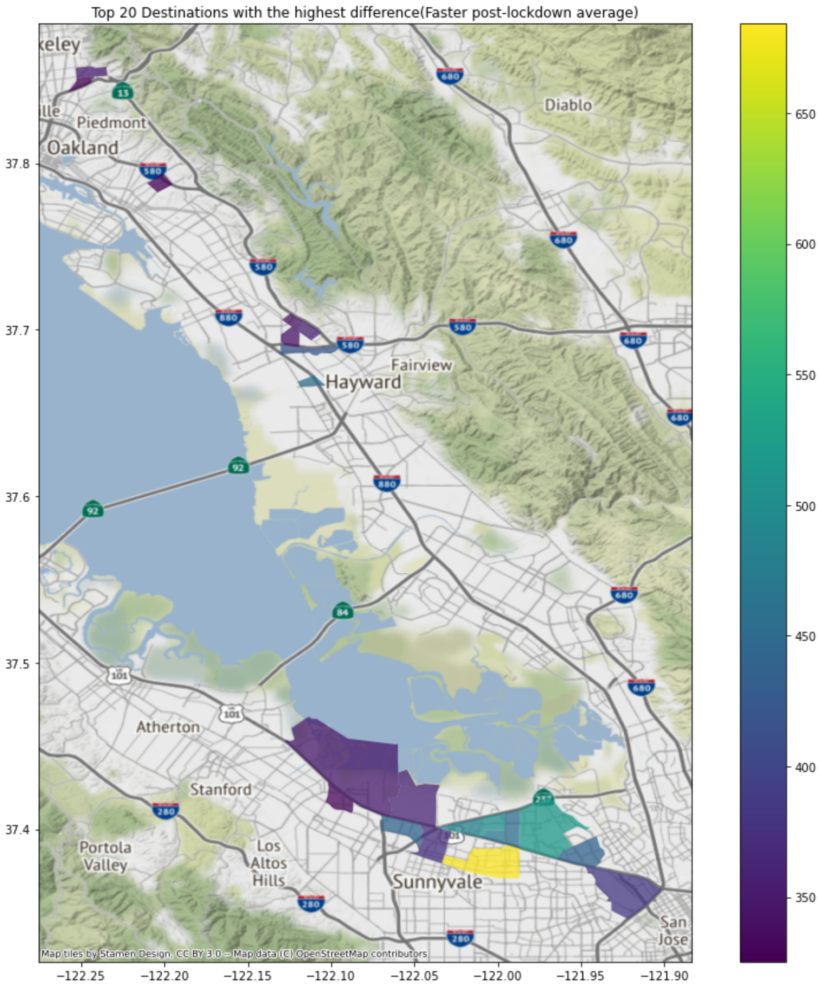


- For locations that had the lowest difference in mean travel time or increased travel time, we observe that the data is less consistent and seems to vary much more than the decreased travel times. This may again be due to random noise in our data, but we hope to be able to have a better understanding of this further on.

Model Proposal

The problem we would like to solve through modeling is predicting the travel time from 300 Hayes St to any location in the Bay Area, and ds1whether specific destinations were affected more significantly than others. Our hypothesis is that areas surrounding large corporate offices experienced the greatest drop in travel times after the lockdown, as the implementation of work-from-home policies greatly reduced the amount of workers commuting to their offices.

To test this hypothesis, we would like to create two models, one trained on pre-lockdown data and the other trained on post-lockdown data. We found this intriguing because we noticed the significant shift in travel times pre and post lockdown, and we also noticed the speedup varied greatly from area to area. For example, in our EDA we saw that travel times to the South Bay Area sped up the most after the COVID lockdown, and it would be fascinating to see whether our model can take this spatial information into account.


Top 20 Destinations with the highest difference(Faster post-lockdown average)

Our inputs for the model would be the location, and the distance from 300 Hayes St to the end destination. One idea we had to account for the spatial information in our model is to create dummy variables for each approximate tract. This way, we can look at the coefficients across the two models we create for each dummy variable to determine which areas experienced the greatest speedup pre and post lockdown. However, we are also looking into adding a feature to see whether our location is within a certain radius of a large office in the Bay Area. To do so, we require a dataset containing the top employing offices and their locations, which we have found but are still looking into acquiring access for. In the event that we are not able to get this data, we can manually look at the coefficients for specific regions and see whether those areas are near any major office locations.

We plan on using a linear regression model to predict travel times as we are solving a regression problem, not a classification problem. To do so, we will apply the concepts we have learned in class by minimizing the predicted travel times in our model with the observed travel times in each data point. As mentioned before, we plan to create dummy variables for each tract in the Bay Area to represent the location of our end destination. In addition, we will calculate the distance from 300 Hayes St to our final location using the latitudes and longitudes of both points. We feel these are the best features to add to our model because the location of where we are heading to makes a significant difference in travel time (as we noticed in our EDA), but distance is also an important feature to account for. Since we plan on using dummy variables, our model will have a lot of features so we feel that RIDGE regression would help as it can eliminate redundant features in our model and shrink the coefficients down to zero.