DATA 100 Final Project (Traffic) Design Document

1. Group Members:

- Shiqi Zhang (3035916454)
- Tina Chen (3035602660)
- Jerry Pan (3035788742)

2. Overview from Part 3 Open-Ended EDA:

- Covid greatly affect people in different ways. In this project, we will examine how Covid influences people's travel time within the Bay area. Specifically, we want to explore how people's travel time varies between weekdays and weekends before and after the lockdown. Through using histograms and heatmaps, we will visually compare the differences on travel time between weekdays and weekends before and after the Covid lockdown. Moreover, we will also examine the areas that have had the most changes before and after the Covid lockdown.

3. Hypothesis & its Motivation:

a. Hypothesis

    - We believe the travel time differences between weekdays and weekends decrease after the lockdown. In other words, people take more time to travel in the same areas on weekdays compared to weekends before the lockdown and take less time to travel in the same area after the lockdown..
    - We also believe that Covid lockdown decreases people's travel time on both weekdays and weekends.

b. Motivation

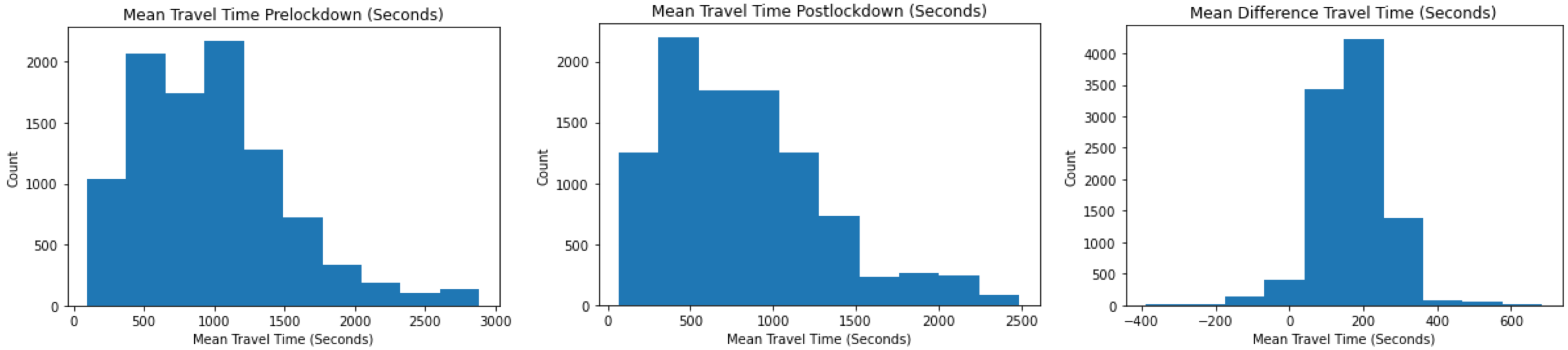
We want to see how Covid has impacted SF transportations and to what extent Covid influences people's traffic speed and time. Through generating heatmaps and visualizations, we will examine the top areas that are affected by the

Covid shutdown. Moreover, we can also predict people's traveling habits to see the travel routes changes before and after the lockdown.

4. Rationale Behind Our Hypothesis

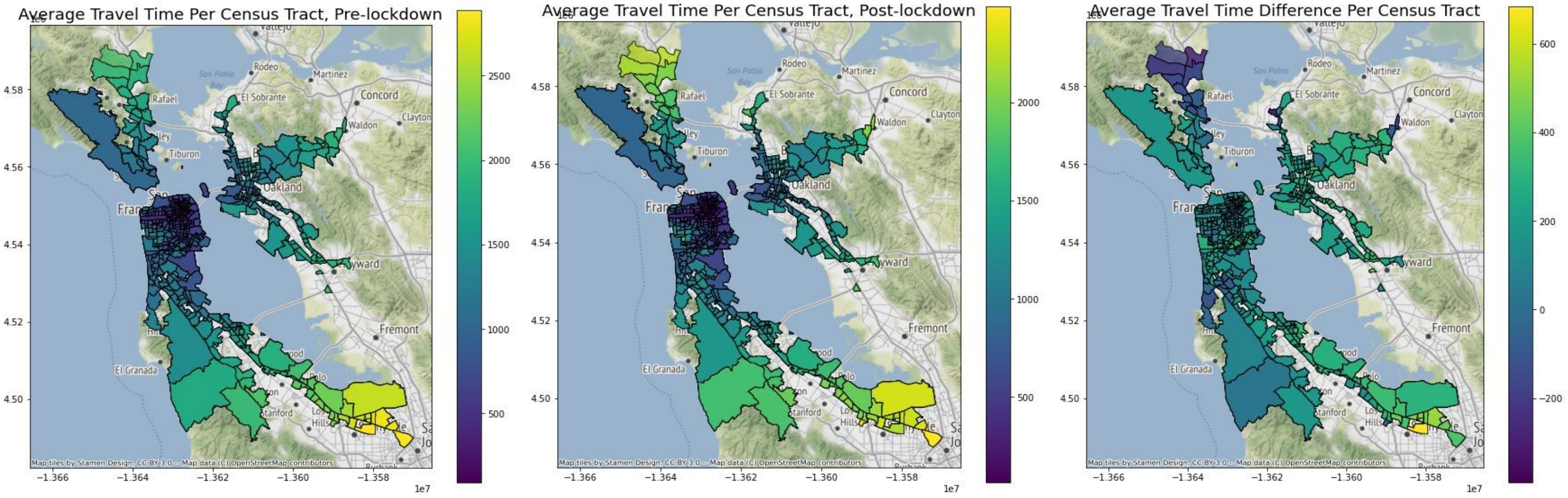
Mean Travel Time in General:

a. Pre/Post/Difference Lockdown Travel Time Histogram



The mean travel time in the pre-lockdown graph is centered around 250 to 1250 seconds with a distribution that is skewed to the right. The highest mean travel times are within the range of 1800 to 2800. On the other hand, the mean travel time in the post-lockdown graph centered around 250 to 1000 seconds. The distribution is also skewed to the right. However, more data is centered around 500 seconds than the pre-lockdown distribution. The highest mean travel times are within the range of 1500 to 2500. The two distributions indicate that the mean travel times decrease after the lockdown. This is shown by the fact that the mean of the mean travel time decreases, and the longest travel time also decreases. This result is further supported by the Mean Difference Travel Time histogram. Since the difference is calculated by the pre-lockdown mean travel time minus the post-lockdown mean travel time, the mostly positive difference indicates that the travel time has generally decreased. The difference in mean travel time centers around 200 seconds. Only a small portion of the census tracts have higher travel time during post-lockdown compared to pre-lockdown.

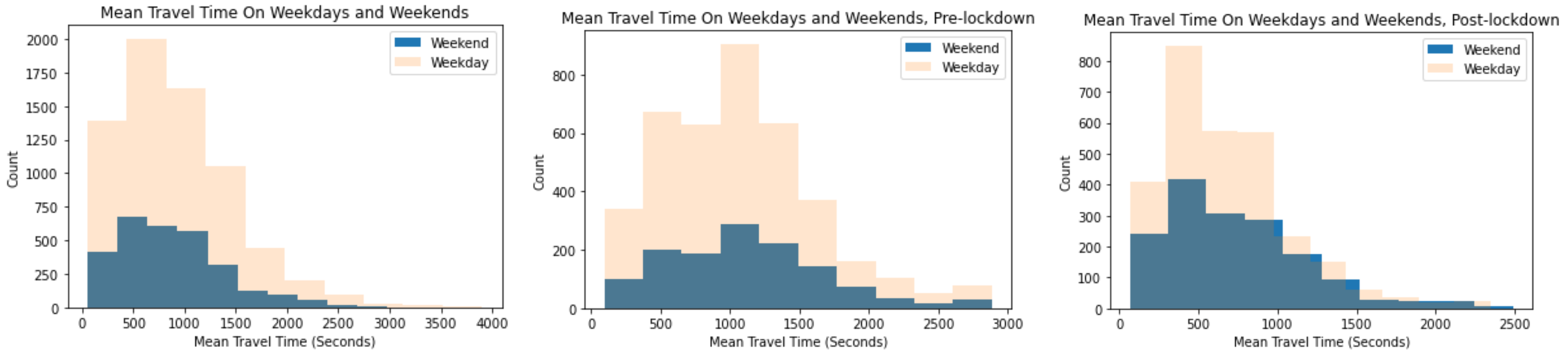
b. Pre/Post/Difference Lockdown Travel Time HeatMap


Average Travel Time Per Census Tract, Pre-lockdown
Average Travel Time Per Census Tract, Post-lockdown
Average Travel Time Difference Per Census Tract
1e6
4.58
4.56
4.54
4.52
4.50
−1.366
−1.364
−1.362
−1.360
−1.358
1e7
2500
2000
1500
1000
500
600
400
200
0
−200
Vallejo
Rodeo
Martinez
Concord
Clayton
El Sobrante
Waldon
Rafael
Tiburon
Oakland
Hayward
Fremont
El Granada
Stanford
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors

In the graph, darker colors represent a slower travel time whereas lighter colors represent a faster travel time. Generally, the colors in the Post-Lockdown HeatMap are lighter than the colors in Pre-Lockdown HeatMap, indicating the mean travel time is faster after the lockdown. This color change is especially noticeable around SF downtown areas and San Jose areas, because most areas contain big companies. After the Lockdown, most people choose to work from home, which decreases the population on the street. This becomes the reason why the travel time decreases. Interestingly, the travel time in San Rafael areas and San Pedro areas increases after the Lockdown. We believe that because people have more free time after the shutdown, people spend most of their time at home. They might choose to hike in the mountain areas to breathe the fresh air. This explains why those areas' travel time increases after the lockdown.

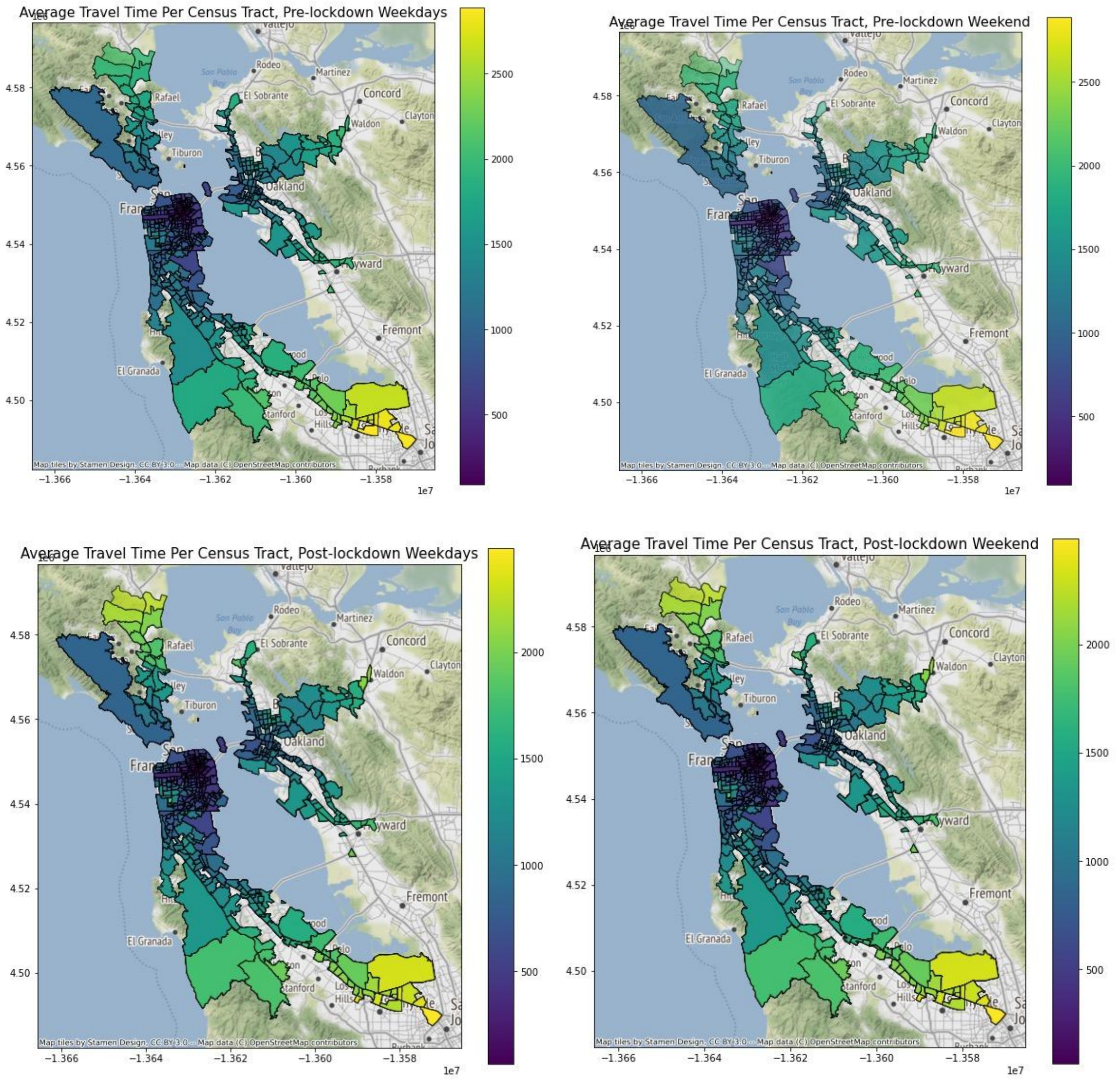
Weekday and Weekend Travel Time:

a. Pre/Post/Difference Lockdown Weekday Travel Time Histogram



The mean travel time in the pre-lockdown on weekdays and weekends graph is centered around 1000 to 1250 seconds with a distribution that is skewed to the right. On the other hand, the mean travel time in the post-lockdown graph centered around 250 to 500 seconds with the distribution skewed to the right. The two distributions indicate that the mean travel times decrease after the lockdown. The Mean Travel Time On Weekdays and Weekends histogram displays the disparity between mean travel times on weekdays and weekends. In general, weekdays have greater mean travel time because people go to work, and this increase of volume causes travel conjunctions. On the other hand, weekends have shorter mean travel time compared to weekdays. This might be caused by the fact that people often go to work or get off from work around the same time, which intensifies the traffic, while people go out at random times over the weekends, and the traffic is more distributed. Comparing the pre-lockdown and the post-lockdown histograms, the pre-lockdown one has greater difference between mean travel time on weekdays and weekends. The post-lockdown histogram shows a smaller difference between weekdays and weekends. This might be due to the fact that people started to work from home since the lockdown, so they have a more flexible schedule to go out.

b. Pre/Post/Difference Lockdown Weekday Travel Time HeatMap


Average Travel Time Per Census Tract, Pre-lockdown Weekdays
Average Travel Time Per Census Tract, Pre-lockdown Weekend
Average Travel Time Per Census Tract, Post-lockdown Weekdays
Average Travel Time Per Census Tract, Post-lockdown Weekend
1e6
4.58
4.56
4.54
4.52
4.50
−1.366
−1.364
−1.362
−1.360
−1.358
1e7
2500
2000
1500
1000
500
Vallejo
Rodeo
Martinez
Concord
Clayton
Walnut
El Sobrante
San Pablo Bay
Rafael
Tiburon
Oakland
Hayward
Fremont
El Granada
Stanford
Los Altos Hills
San Jose
Map tiles by Stamen Design, CC BY 3.0 — Map data (C) OpenStreetMap contributors

In the graph, darker colors represent a slower travel time whereas lighter colors represent a faster travel time. Generally, the colors in the Post-Lockdown HeatMap are lighter than the colors in Pre-Lockdown HeatMap, indicating the mean travel time is faster after the lockdown. The four heatmaps together indicates a general trend of mean travel time deduction in pre vs. post lockdown periods. The pre-lockdown mean travel time on weekdays is higher than the pre-lockdown travel time on weekends. Moreover, it is also greater than post-lockdown mean travel time on weekdays and on weekends. However, the dataset on weekends only contains limited points. This means that the differences between each group is difficult to directly visualize through heatmaps with human eyes. To address this limitation, we plan to further calculate the difference between pre-lockdown mean travel time on weekdays and pre-lockdown mean travel time on weekends, pre-lockdown mean travel time on weekdays and post-lockdown mean travel time on weekdays and then to visualize the difference through heatmaps.

5. Methodology
    a. Data

- As we are given two datasets “Daily Movement Speeds in San Francisco”, “Daily Travel Times in San Francisco” in March 2020, and one XML mapping from OSM nodes to GPS coordinates, we effectively join the relevant datasets and dataframes based on the mutual unique geographical identifier.
- Columns in “Daily Movement Speeds in San Francisco”: ['osm_start_node_id', 'osm_end_node_id', 'day', 'speed_mph_mean']
- Columns in “Daily Travel Times in San Francisco”: ['Origin Movement ID', 'Origin Display Name', 'Destination Movement ID', 'Destination Display Name', 'Date Range', 'Mean Travel Time (Seconds)', 'Range - Lower Bound Travel Time (Seconds)', 'Range - Upper Bound Travel Time (Seconds)', 'day']

b. Experiment

- The granularity of the dataset “Daily Movement Speeds in San Francisco” is the speed mean in Mile Per Hour (MPH) in each day along each route. Given the OSM start and end node ID, we made use of different geographical partition mechanisms, including Google Plus Codes and Census Tracts. With different groupby methods Google Plus Codes vs Census Tracts, we visualize and analyzed the histogram distribution, mean, and variance of mean travel speed pre-lockdown vs post-lockdown, and drew the conclusion that grouping by census tract is a better choice as with small in-subpopulation variance compared to greater across group variance.

c. Evaluation

- As the purpose and focus of this project is to compare the travel time/speed between pre-lockdown and post-lockdown, we visualized and analyzed a series of comparisons based on different locations and routes with GeoDataFrame map visualization.

6. Modeling

a. Purpose

- Since we want to figure out the speed/travel time difference between pre-lockdown and post-lockdown in different locations or routes, we want to predict the post-lockdown speed/travel time based on the pre-lockdown speed/travel time for each route during weekdays and weekends respectively.

b. Techniques

- Multiple linear regression would be used given all relevant features available, including but not limited to route destinations, streets, weekdays/weekends.
- No imputation is needed given no null values in our cleaned dataset.

c. Expectation & Limitations

- Given our prior EDA and visualization results that lead to our conclusions above, we expected our model prediction to have reasonable accuracy given significant differences between different routes and whether it's on weekdays or weekends.
- Limitations: given we only have 4 weekend days after grouping by MOVEMENT_ID, our model could potentially have high model bias compared to our model prediction on weekdays. Due to limited travel information from the original dataset, we will search for potential new datasets to compensate for other factors that might be impacting our hypothesis in part two.

7. Open-Ended Questions About Our Dataset

   a. Which specific areas are being affected more prominently by the lockdown?

   b. Do these areas suggest any implications for the socioeconomic status of neighborhoods?

   c. Are there any confounding variables that contribute to the change in average travel time between pre and post lockdown?