DATA 100:

Design Document

(Traffic Dataset)

Students: Sreeja Apparaju, James Jeong, Tanmay Vijaywargiya

Discussion Session 135 | Date: 17 Nov 2021

Introduction

We worked with Uber's Traffic Dataset for the DATA 100 Final Project. This data includes speed data for Uber rides during the month of when California-wide COVID lockdown measures began, which was in March 2020. Through this design document, we first reflect on our insights from Guided Data Cleaning and EDA given by the course staff. Then, we provide an insight into our Open EDA which we used to construct our hypothesis test. We conclude our design document by giving a jest on our modeling idea for Final Project Part 2.

Guided Data Cleaning & EDA

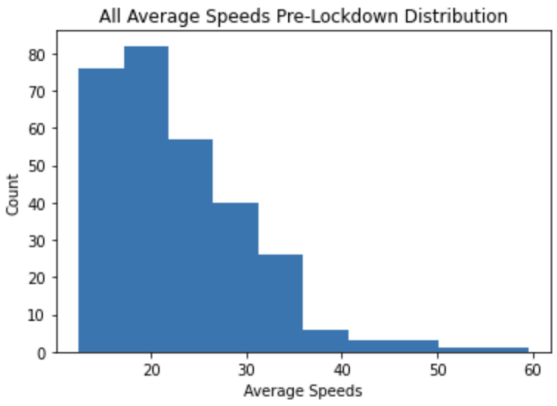
Following the instruction in the Traffic Jupyter Notebook, we first loaded the speeds_to_node dataset (Uber's traffic speeds dataset)

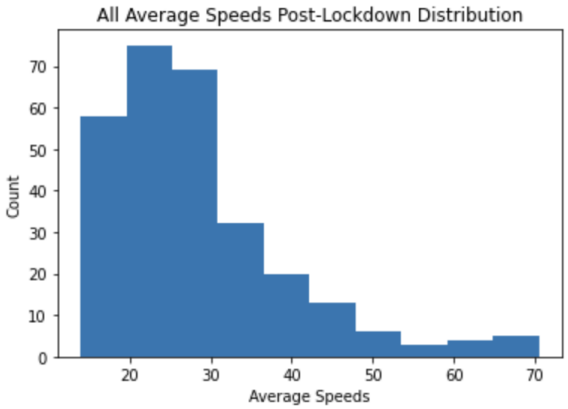
The data frame consists of columns in an OpenStreetMap (OSM) node format and we used Regex to extract the latitude, longitude, and node id and build a new dataframe node_to_gps. Next, we mapped traffic speeds to GPS coordinates to segment traffic speeds spatially. We were tasked to build Google Plus Codes for speeds_to_gps, by dividing the whole world into uniformly-sized squares, which are 0.012 degrees latitudinally and longitudinally. *Figure 1* shows speeds_to_gps with additional column plus_code representing our calculated 'Google Plus Code.'

	osm_start_node_id	osm_end_node_id	day	speed_mph_mean	osm_node_id	Latitude	Longitude	plus_latitude_idx	plus_longitude_idx	plus_code
0	26118026	259458979	1	64.478000	26118026	37.675280	-122.389194	18140	4801	18140_4801
1	26118026	259458979	2	62.868208	26118026	37.675280	-122.389194	18140	4801	18140_4801
2	26118026	259458979	3	62.211750	26118026	37.675280	-122.389194	18140	4801	18140_4801
3	26118026	259458979	4	62.192458	26118026	37.675280	-122.389194	18140	4801	18140_4801
4	26118026	259458979	5	61.913292	26118026	37.675280	-122.389194	18140	4801	18140_4801
...	...	...	...	...	...	...	...	...	...	...
417634	4069109544	615120176	30	38.956000	4069109544	37.732039	-122.507126	18145	4792	18145_4792
417635	5448539901	65446993	16	25.627000	5448539901	37.622476	-122.413763	18136	4799	18136_4799
417636	302964668	4069109544	19	40.802000	302964668	37.732418	-122.507206	18145	4792	18145_4792
417637	302964668	4069109544	20	36.076000	302964668	37.732418	-122.507206	18145	4792	18145_4792
417638	5022068066	302964668	19	39.592000	5022068066	37.733635	-122.507100	18145	4792	18145_4792

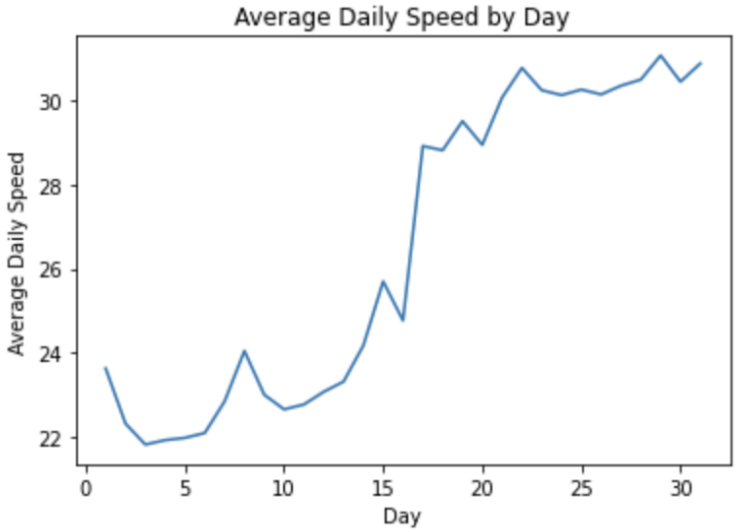
In the next steps, we perform a similar analysis of traffic speed by dividing our data into census tracts using speeds_to_tract dataframe.

In order to further understand the impact of lockdown on traffic speed, we sorted our census tract dataframe speeds_to_tract according to pre and post-lockdown period and computed the average speed per census tract. Both histograms indicated a heavy rightward skew, but the right histogram has a large portion of data points under 20 mph (second highest bar), and maxima at around 60 mph while the left histogram has the second-highest bar around 25 mph, and the maxima increased to around 70 mph.





Due to this rightward shift from pre-covid to post-covid and supposed increase in speeds, we further created a line graph to visualize the average speed from before covid to after covid.



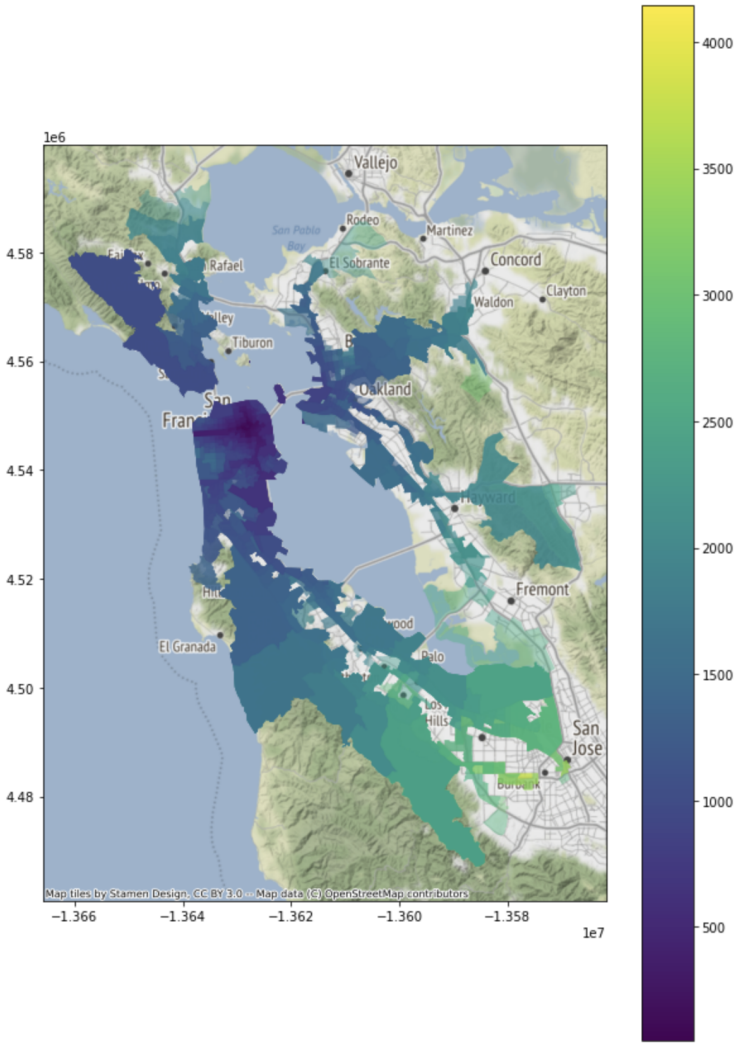
The graph above shows the immediate effect of a massive increase from the days preceding lockdown to the days immediately after. The average speed across the entire data spiked almost 5 mph over the course of 2 days. In the context of the dataset which has hundreds of thousands of data points, this is a massive step up in speed.

Open EDA

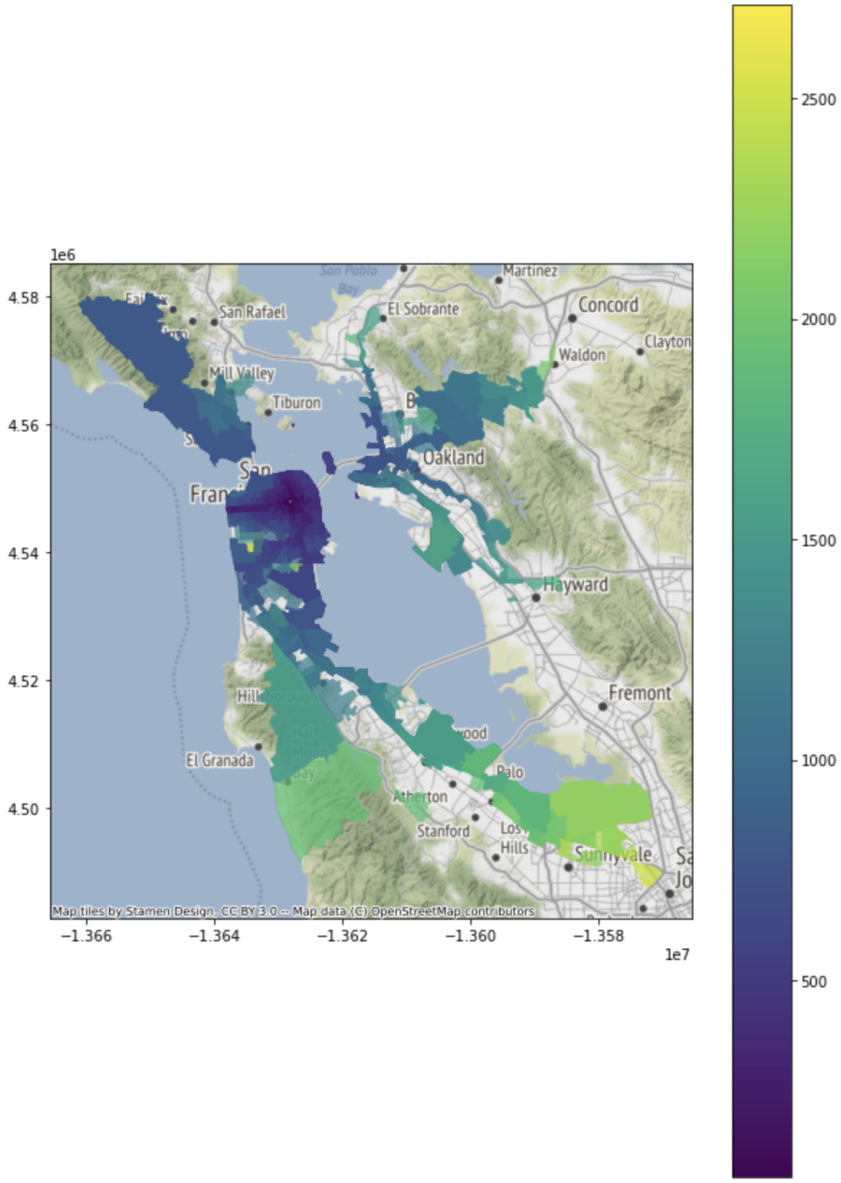
Since our increase in speed was dramatic, we wanted to explore whether a similar effect occurred with times traveled in the SF Bay Area. Using times_to_tract dataframe that contains the travel time from Hayes Valley in central San Francisco to every other census tract in the SF Bay Area, we used columns:

1. Destination Movement ID
2. Destination Display Name
3. Mean Travel Time (Seconds)
4. day

We converted the Pandas DataFrame into GeoPandas DataFrame to include geometric points and merged it with our previous tract_to_gps. Now, since we could map travel times visually our exploratory analysis could reach new levels of inference and intuitive conclusions.



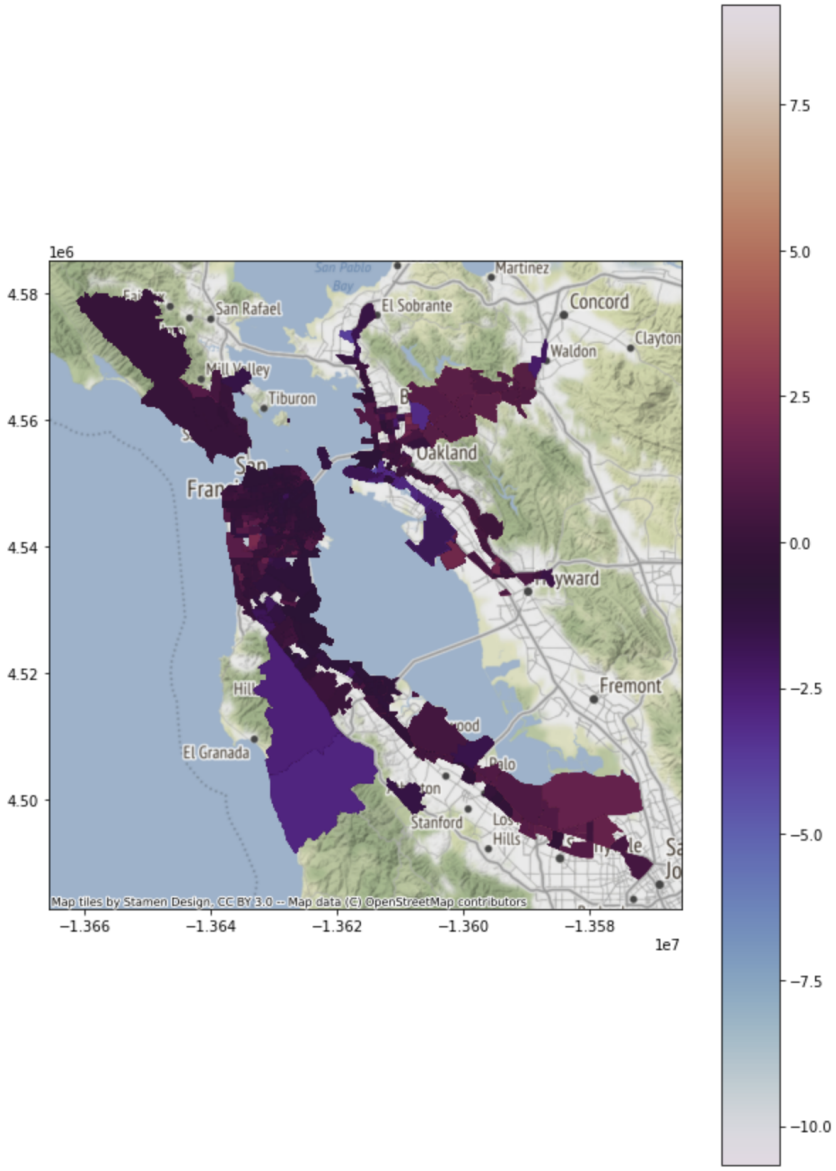
Heatmap of daily travel times from Hayes Valley to other census tracts
Pre-COVID lockdown



Heatmap of the daily travel times from Hayes Valley to other census tracts
Post-COVID lockdown

Splitting the data into pre-covid and post-covid, we produced choropleth maps. On the map, we noticed a dramatic effect on the travel times, most noticeably on the places farther away from Hayes Valley. Since the speeds increased as seen in the line graph, and since the travel times decreased, it is plausible to infer that the overall traffic in the San Francisco Bay Area reduced from pre-covid to post-covid.

However, we also wanted to see how much they decreased when compared to the average difference (pre and post-covid) overall. Therefore, we standardized the differences between pre and post-covid and plotted each average standard deviation onto a choropleth map. To our surprise, there was a unique pattern we witnessed that led to the formulation of our hypothesis.



Heatmap of the standardized difference in daily travel times from Hayes Valley to other census tracts Post-COVID lockdown

In our standardized differences choropleth graph, we noticed higher standard deviations (indicating a greater decrease in times traveled) with areas that were considered suburban or traditionally higher affluence in the SF Bay Area. Areas in South Bay (an affluent area of the Bay Area) had standard deviations much higher than the areas that were considered less affluent near San Francisco. In addition, areas near San Francisco in suburbs had higher standard deviations than those considered in the inner city of San Francisco.

This was eye opening information and we were quite intrigued in understanding whether trips to higher-income locations or lower-income locations have considerably different travel times. We would expect that travel time would generally be correlated with variables such as distance. However, we were now interested in discovering whether the location destination, whether lower or higher income could be relatively correlated with the travel time. In order to compute this over the time period of Covid, we would first find the rate of change in travel times for each location from pre-covid to post-covid. Since we witnessed lower income neighborhoods having lower standard deviations than higher income neighborhoods, we would expect that the lower income neighborhood destinations would have a lower rate of change in travel times than those in the higher income neighborhoods.

Rate of Change of the Differences:
(Post-Covid Travel Time – Pre-Covid Travel Time) / Pre-Covid Travel Time

Hypothesis, Testing & Modeling

Hypothesis (Ha): From pre-covid to post-covid, the rate of change in time traveled on average to reach a lower-income area should be more than the rate of change in time traveled on average to reach a higher-income area.

Alternative Hypothesis (Ho): From pre-covid to post-covid, the rate of change in time traveled on average to reach a lower-income area should NOT be more than the rate of change in time traveled on average to reach a higher-income area.

To bring income data into our study, we researched the US Census Tract by median income. We'll be able to relate this dataset to the current travel time dataset which is already combined geometrically with tract_to_gps. By using this data, we will have to partition the census tracts into higher and lower-income groups which will be separated by context.

Once we retain these groups, the dataset will be ready for more data manipulation. We will have to create a new column that will find the rate of change between pre-covid and post-covid per data point for each higher-income or lower-income census tract. After determining the rate of change, we will create a new column in our dataset and input our rate of changes per data point.

To begin the testing process we will first try to visually analyze and compute the distribution of our low-income rate of change in differences. To compare we will, in tandem, compute the distribution of our high-income rate of change in differences. Then, we will superimpose these distribution histograms to find the visual difference between each of the locations.

To understand the statistical significance between the high-income or low-income destinations, we will take the mean of the low-income rate of change in differences and standardize it to the distribution of the high-income rate of change in differences. If the t-statistic computed through this calculation is lower than the 2.5% level of significance one-sided (above +1.95 standard deviations from mean), then we will reject the null hypothesis. If the t-statistic computed through this calculation is greater than the 2.5% level of significance one-sided (below +1.95 standard deviations from mean), then we will fail to reject the null hypothesis in favor of the alternative hypothesis.

For our modeling portion, we will use the Multiple Linear Regression model (an extension of the OLS linear regression). The model will be used to regress travel time on the rate of change in differences from pre-covid/post-covid and whether the destination is higher income or lower income.

From our model, we can use the rate of change in differences from pre-covid to post-covid and the higher or lower-income area to predict a newer and more accurate travel time to the destination at hand currently.

We can further look into future questions such as the demographics of the region (race, gender, age), how impacted a certain community was from the lockdown, what the main occupation of individuals in a region was. These can all be used to understand travel times due to accessibility and traffic.