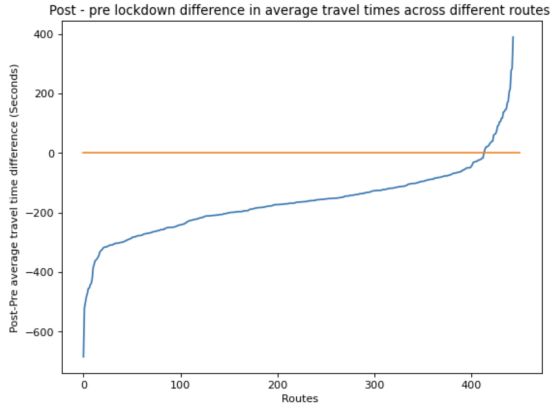
Data 100/200 Project Design Document
Traffic Dataset

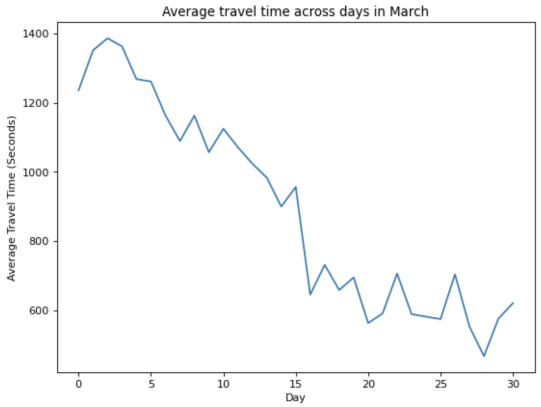
Eric Mahoney, Chris Liu, Varun Agarwal, Milena Rmus

Data: The traffic dataset consists of information about the speed and/or travel times in the different spatial regions of the San Francisco Bay Area, sampled in March 2020. Each row corresponds to a daily average of the appropriate traffic metric (speed (mph), travel times (number of seconds)) extracted from different spatial segments. Importantly, data columns contain information about 1) the spatial coordinates (latitude and longitude) of each area, 2) corresponding plus code of the given area, which can be derived from the spatial coordinates, and 3) the day on which the traffic information was collected. The dataset also offers start-destination locations, allowing the inspection of traffic properties along different routes. The dataset, therefore, offers a way to compare how different definitions of spatial segments/clusters affect our insights about the traffic properties of these areas, as well as how these properties changed from before to after the COVID-19 pandemic lockdown. Other information included in the dataset: the names of the spatial areas and the range of traffic metric values.

EDA: In the exploratory data analysis of part 1 (step 3) we focused on the following questions:

1. How does the travel time change across time (before and after lockdown, and just across different days)?

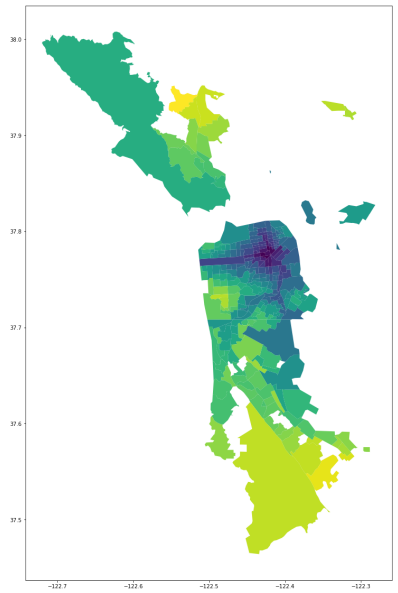


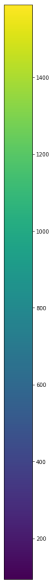


Insights: We plotted post-pre lockdown difference in travel time averages across different routes and found that the travel times reduced after lockdown (more negative difference values). The change across days of March 2020 also showed an overall reduction in average travel time from beginning to the end of the month.

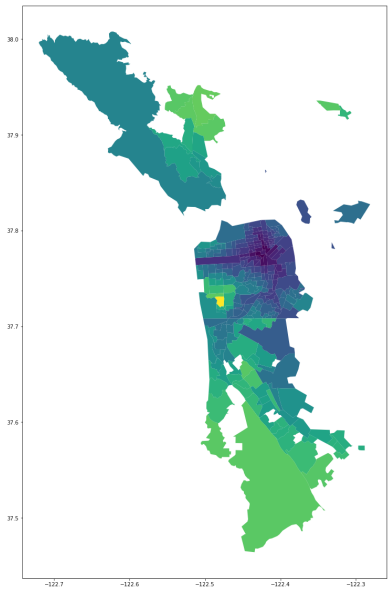
2. How did the travel times to different destinations from Hayes Valley change from before to after lockdown?
3. Is there any similarity/patterns in how the routes/destinations were affected?
4. What were the clusters that emerged?

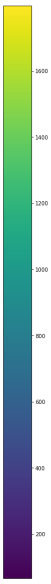
A) Pre lockdown travel times for different destinations



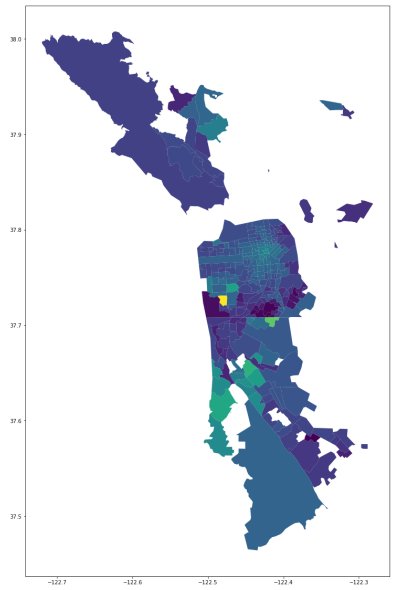


B) Post lockdown travel times for different destinations





C) Post - pre lockdown difference in travel times for different destinations



The first 5 rows show different destinations from Hayes Valley, sorted by pre-post lockdown absolute difference in travel times. The south bay regions (i.e. Sunnyvale) showed the greatest difference between pre-post lockdown travel times, in that the travel times decreased significantly after lockdown.

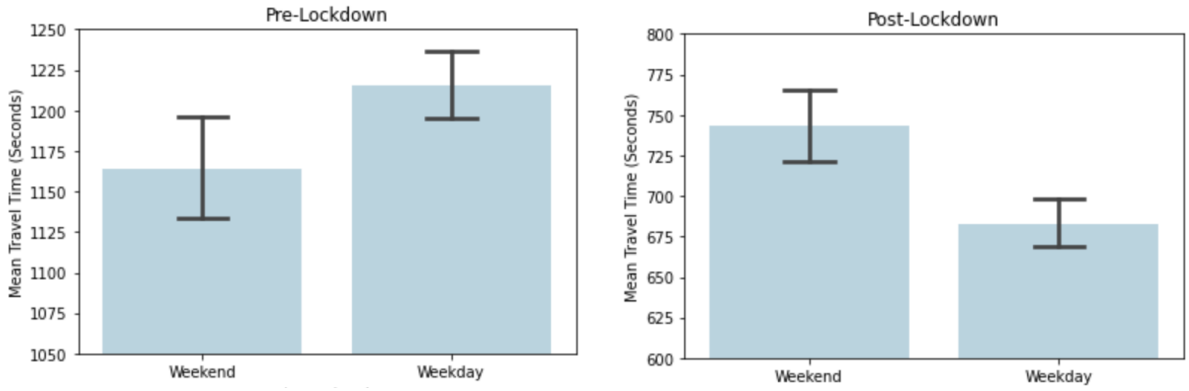
	index	Origin Movement ID	Destination Movement ID	Pre Lock Travel Time	Post Lock Travel Time	PrePostDifference	AbsDifference	Destination Display Name
0	7639	1277	1831	2887.111111	2203.000000	-684.111111	684.111111	1100 East Arques Avenue, Sunnyvale
1	5186	1277	1302	2816.076923	2296.000000	-520.076923	520.076923	Tasman Drive, Santa Clara
2	1929	1277	554	2701.538462	2200.333333	-501.205128	501.205128	200 2nd Street, Sunnyvale
3	7892	1277	2169	2663.153846	2180.000000	-483.153846	483.153846	1000 Liege Terrace, Morse Park, Sunnyvale
4	1000	1277	306	2703.416667	2231.666667	-471.750000	471.750000	200 Silverlake Drive, Lakewood, Sunnyvale

Insights: We grouped the daily average travel times by different destination IDs, and created a heatmap of average travel times 1) before lockdown 2) after lockdown, and 3) the difference in travel times post-before lockdown. Hayes Valley - South Bay (Sunnyvale/Mountain View), Hayes Valley - Sausalito, Hayes Valley - East Bay show similar patterns. The travel time is overall reduced post lockdown. These are likely freeways/bridges that were not busy post-lockdown, resulting in reduced travel times.

The areas closer to Hayes Valley in proximity show somewhat positive change post-lockdown (although this doesn't look super convincing), suggesting that residential areas might have had a slight increase in travel time (could be explained by people avoiding being around other people and opting to drive when necessary instead).

There are a few segments that actually increase in travel time post lockdown, which could possibly be a data artifact (i.e. not a lot of data points for this destination with some outliers, resulting in inflated averages).

5. Travel times dissociation between workdays and weekends - how does this differ pre/post lockdown?



Insights: We added an indicator variable to the Hayes Valley dataset to indicate whether the day was a weekend or a weekday, and then we plotted the mean travel times for each type of day. Before lockdown, weekends had lower travel times, on average. However, after lockdown, while travel times decreased dramatically, weekends had higher travel times on average. One possible explanation for this shift could be that weekday travel (i.e. travel for work) was impacted more than weekend travel (i.e. travel for leisure).

Propose a problem that you will address with modeling. Perhaps this is a problem you discovered while conducting EDA. Some potential questions to answer: What is the problem? Why is it relevant/intriguing? How will your model address this problem?

(Paraphrased from the Notebook)
Context: Because of the sudden change in traffic patterns, existing models to predict traffic (trained on pre-lockdown traffic) do not perform well anymore.

Goal: Create/train new models that can predict post-lockdown traffic given little training data. Consider problems with the existing models/original training data that we will need to address with our new model.

Problems with post-lockdown data:

- In the post-lockdown data, weekends no longer have shorter travel times, which could be a problem if the existing model takes the day of the week into account
-

The problem of *traffic prediction* is highly relevant for companies such as Uber/Lyft, as the properties of traffic largely contribute to determining the ride prices. One of the questions we

were interested in is: can we develop a model that can accurately predict the traffic speed, and what are the features that contribute the most to accurate speed predictions? Our model will include a number of features included in the speed datasets, and we will apply cross-validation and regularization to select the combination of features from the given dataset that constitute a model with the most accurate predictions. Our hypothesis is that …. [INSERT]

What sort of modeling do you plan on conducting? Carefully describe the methods you plan on using and why they would be appropriate for the question to be answered.

We will use a linear regression model to test our hypothesis. Given that we are not testing for classification or categorization, but are focusing on making predictions about the speed, the linear model seems like the most appropriate choice. In addition, the linear modeling approach lends itself nicely to our approach given the nature of the features we intend to include (speed as the feature we are trying to predict and a combination of quantitative continuous variables we plan to use for predicting).

We will start by using the domain knowledge to engineer a model with several different feature combinations that will likely yield accurate speed predictions. For instance, we can include the day of the month, latitude, longitude, the weekend/workday category, pre-post lockdown category, etc. For the categorical variables, we will use one-hot encoding. Prior to implementing the model, we will visualize the relationship between the data features, to determine whether we need to conduct some sort of transformation (i.e. taking the log / square root of the feature values). Next, we will use cross-validation and/or regularization to determine which one of the feature combinations yields the most accurate out-of-sample predictions of speed. Specifically, for each model we will use k-fold cross-validation to 1) compute the average model loss across k folds on validation sets, and 2) the model loss in making predictions about the held-out test data to account for the possibility of overfitting. We will make inferences about different models based on how well they reduce both bias and variance. In addition, we will also implement different types of regularization, to augment our feature selection process.

Some additional questions/modeling we'd consider exploring:

Logistic regression: can we train the logistic model to predict whether traffic properties (speed and/or travel times) were sampled before or after the lockdown? Here, we would classify traffic properties into 2 groups: before or after the lockdown. This classification problem might be useful, in that we can build an expectation of what different traffic properties will look like in the event of another lockdown.

Clustering: can we use an unsupervised learning algorithm (such as k-means clustering, or something similar) to construct traffic property-based clusters, that override the spatial segmentation (i.e. east bay and south bay might be the same cluster, even though they are not in close spatial proximity/in the same spatial cluster)? What clusters would we expect to

emerge? How do we think this would impact determining the ride pricing for companies like Uber/Lyft?