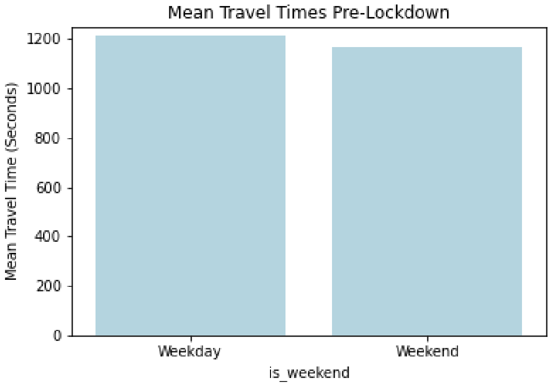
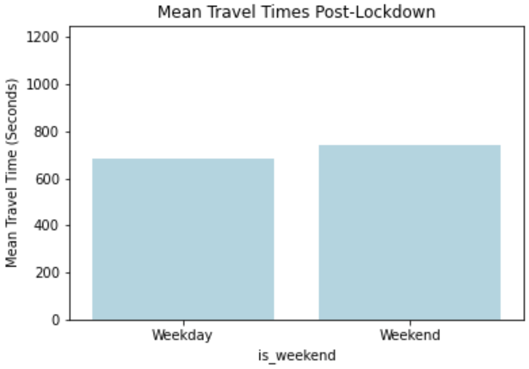
Predicting traffic speed on weekdays and weekends before and after the covid-19 lockdown

Data 100/200 Project Design Document
Traffic Dataset
Eric Mahoney, Chris Liu, Varun Agarwal, Milena Rmus

General question - How did the lockdown affect traffic properties (i.e. speed/travel time) on weekdays and weekends?





Previous EDA has shown that traffic properties differ on weekends and weekdays, before and after the lockdown.

Why is it relevant: For companies such as Lyft, making accurate traffic *predictions* (speed, expected travel time) is important, since traffic properties largely factor into determining ride prices. Knowing both where and *when* the traffic is going to be busier (i.e. based on the expected speed) is useful in predicting the duration of the ride, and consequently adjustment of the ride price.

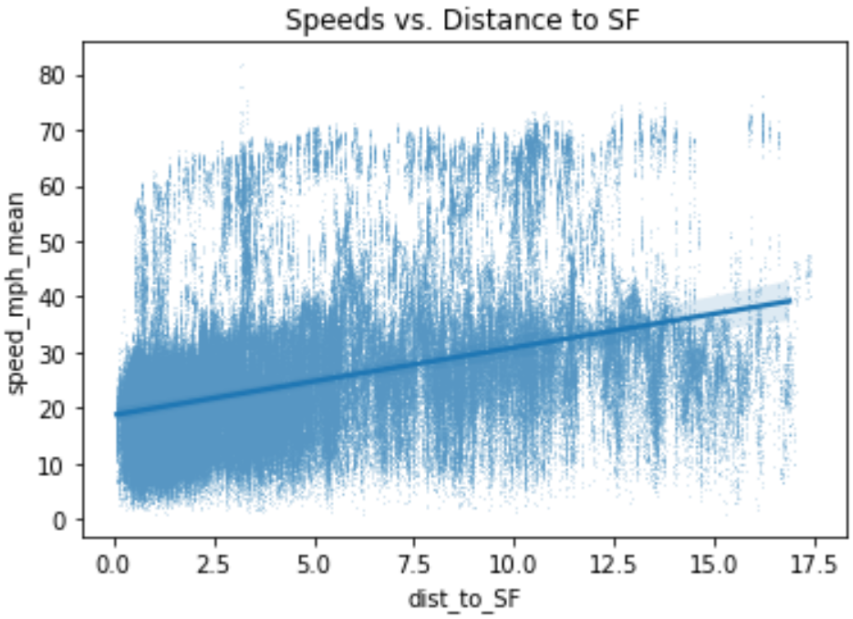
Hypothesis: Post-lockdown, weekday trips have higher average speeds (and hence lower travel times) than weekend trips. We believe that we can improve the accuracy of speed prediction from the baseline model (just distance and day) by adding the weekend/weekdays and before/after lockdown features to the model and/or by splitting the data into post/pre lockdown segments and training and evaluating the model on separate segments. We will test this hypothesis by attempting to construct a regression model that makes accurate predictions based on the temporal (day, weekend/weekday, pre/post lockdown) and spatial (the origin of trip, the distance of the trip) information. If the null hypothesis is true and weekday/weekend difference doesn't matter for traffic properties, then adding these features won't improve the model predictions.

Modeling Approach:
We used a linear regression model to test our hypothesis. We focused on making predictions about the speed from the combination of categorical and continuous features, and therefore the linear model seems like the most appropriate choice. To prevent overfitting, split the dataset into training and validation sets. We performed feature engineering to improve the model's performance on the training set, and once we reached an adequate error on the training set we checked the performance on the validation set to confirm that our model was not overfitted on the training data.

The modeling we did fell into two categories: low-resolution modeling, which consisted of making general predictions about average travel speeds for each day, and high-resolution modeling, which consisted of making predictions for each individual link in the dataset.

Adding features to the existing dataset: Because we were interested in how the trip distance factors into the speed predictions and the distance wasn't a part of the original data, we created an additional dimension of the dataset that contains information about the distance (in miles) between the origin and destination point. We use the geopy module to compute the distance between the start and end points of each link in the dataset.

We also used the coordinate feature to calculate the distance from the link in the dataset to downtown San Francisco (we chose the Civic Center BART station). Our EDA in Part 1 (the heatmaps) revealed that travel speeds were generally lower downtown, so this feature was useful in our model. The figure below also shows the correlation between this feature and travel speed.



Similarly, since the original speeds dataset does not contain information about whether the speeds were collected on the weekend or weekday, we sampled the information about which days were weekends in March 2020. We then combined this information with the original “day” variable to derive an additional feature which provides the label of whether a day is weekday or weekend.

Baseline Model

Because our question is largely focused on training the model to predict the speed as a function of temporal properties (weekend/weekday and pre/post lockdown), we aggregated the data across all days in March as that permits us to make a more direct comparison. Therefore, in the dataset we had average speeds and distance values for each day, along with an indication of whether the day is the weekday/weekend day, and whether it was before/after the lockdown.

Inputs to the model

We input the following features as part of the baseline model (see model improvement for more features we tested to improve the model): distance and day. We considered this to be the baseline model since it provides the basic level information required to make predictions about the speed.

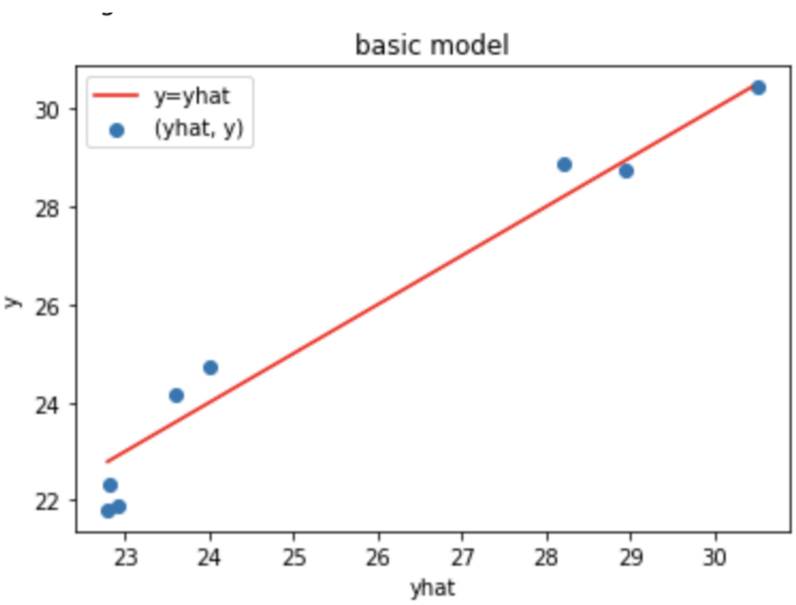
Model outputs

We used the *train_test_split* to divide data into the training and validation set. We then fit the model to the training set, and then evaluated the model using the testing set. We collected the following outputs from the model:

1. Model predictions (from using model.predict(validation data))
2. Loss metrics, including training/testing RMSE
3. Model score/R2 for both training and testing

These are the outputs we obtained from all of the following models we tested.

	RMSE (training)	RMSE (validation)	R2 (training)	R2(validation)
Baseline model	.64	.67	.96	.95

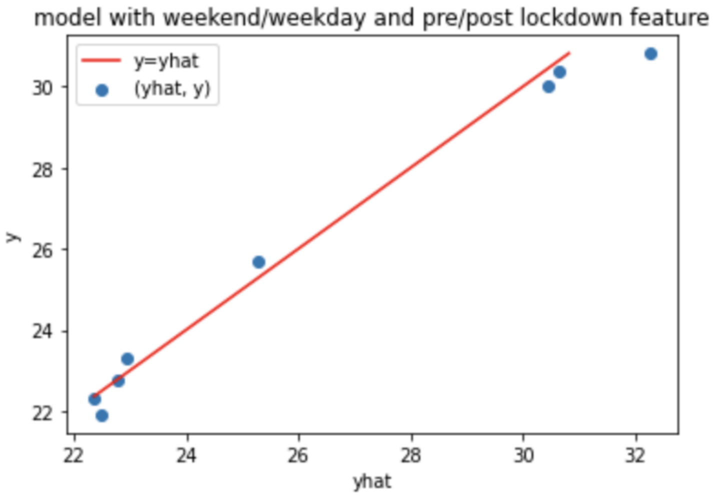


Prediction check for the model showing the relationship between the predicted and true speed values. The perfect performance would mean the dots lie on the red line.

Based on the figure showing the relationship between real speed values and the values predicted by the model (from validation), and the model metrics, the baseline model performs relatively well (**RMSE validation = .67, R^2 validation = .95**). However, we will test whether by adjusting the model by adding more features, and potentially changing the properties of the features (i.e some of which might not be linearly related and/or require transformations, like log transformations) we can get better performance.

Given our question, and the speed difference on weekends/weekdays and before/post lockdown, we wanted to test whether adding weekend/weekday and pre-post lockdown features improves our model predictions. Weekend/weekday and pre/post lockdown features are categorical; therefore, we used one-hot-encoding to convert categorical to integer data prior to fitting the new model.

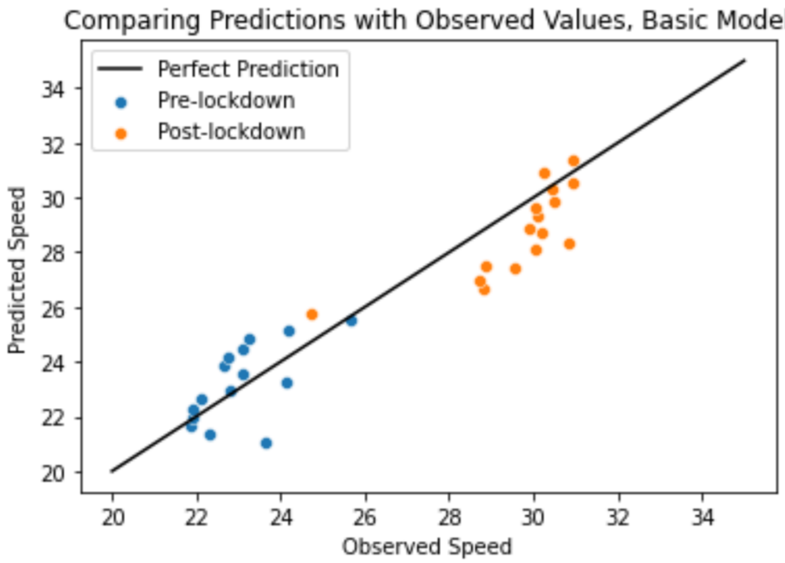
	RMSE (training)	RMSE (validation)	R2 (training)	R2(validation)
Improved model	.41	.60	.98	.97



Adding the weekend/weekday and pre/post lockdown further reduced the RMSE and increased R^2, relative to the baseline model (**RMSE validation = .60, R^2 validation = .97**)). This is in line with our prediction that considering the difference between average speeds on weekends and weekdays, before and after the lockdown can help fine-tune the model, such that we generate more accurate speed predictions.

Issues with Baseline Model

While the baseline model seems to offer decent predictions when we aggregate by day, when making higher resolution predictions (predicting travel speeds at the level each link and each day), our model's performance suffers. At this higher resolution, we have a training RMSE of **10.75 mph**, and an r^2 value of **0.198** on our training set. Additionally, the plot below reveals a more significant problem with this model.



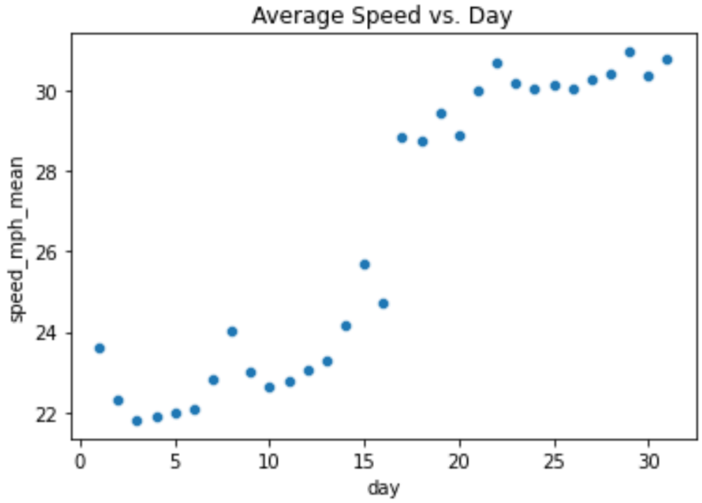
This plot compares the prediction to the actual observed value and aggregates by day. The black line of slope = 1 represents a perfect prediction. What is important to observe in this plot is that our model *does not* account very well for the difference conditions pre- and post-lockdown. The pre-lockdown predictions tend to be too high, and the post-lockdown predictions tend to be too low.

Therefore one of our goals when improving our model will be to account for this changepoint and reduce this systematic error in the model.

Improving the model

Day

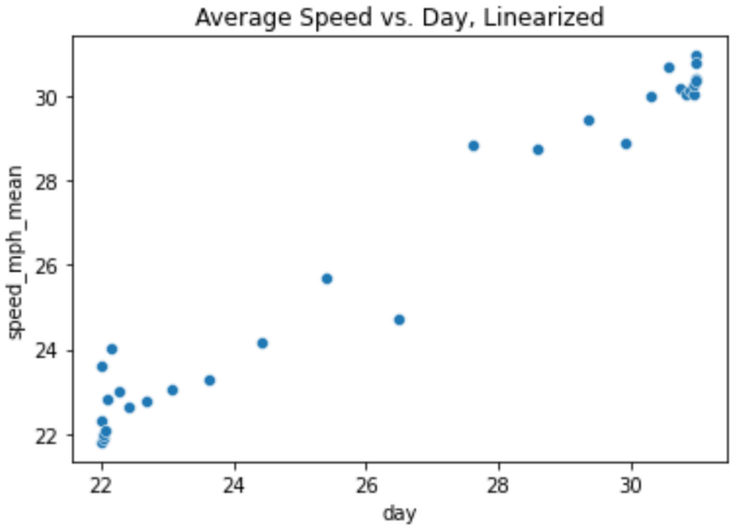
A plot of the average *speed_mph_mean* for each day is shown below. We noted that there appeared to be a sigmoidal relationship between the *day* feature and the travel speeds.



By applying this transformation to the *day* feature,

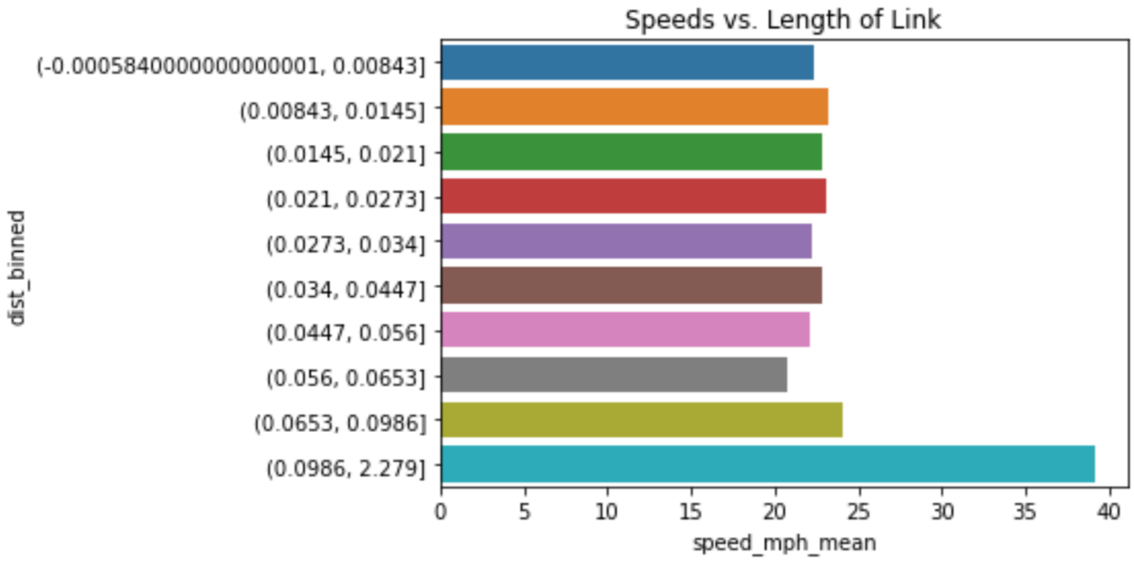
$$f(day) \; = \; \frac{9}{1+e^{(-\frac{day}{2}+8)}} + \; 22$$

we were able to linearize the relationship between *day* and *speed_mph_mean*, as shown in the plot below:



Distance

Next we looked at the *distance* feature. We binned our training data into 10 bins based on the distance of the link, and looked at the average travel speed for each bin. A bar chart summarizing these results is shown below.



Based on this EDA, we noted that links with a distance greater than 0.0986 miles had significantly higher travel speeds, and the rest of the links were about the same. Therefore we created a new feature, *is_long*, which is an indicator variable that is 1 if the link is longer than 0.0986 miles and 0 otherwise.

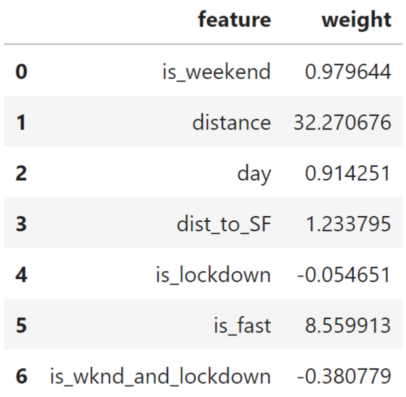
Temporal Features: Weekend and Lockdown

Our EDA in Part 1 revealed that there was a difference in travel speeds depending on whether it was a weekend or if the city was in lockdown. Therefore we created three new features that were indicator variables indicating whether or not it was a weekend, whether or not it was lockdown, and whether or not it was both a weekend AND in lockdown.

- For lockdown vs. non-lockdown, on average there was a **6 mph** difference between the two states.
- For weekend vs. non-weekend, on average there was a **1 mph** difference between the two states.
- For (is*_weekend && is_lockdown)*, on average there was a **3 mph** difference between the two states.

Evaluating Improved Model

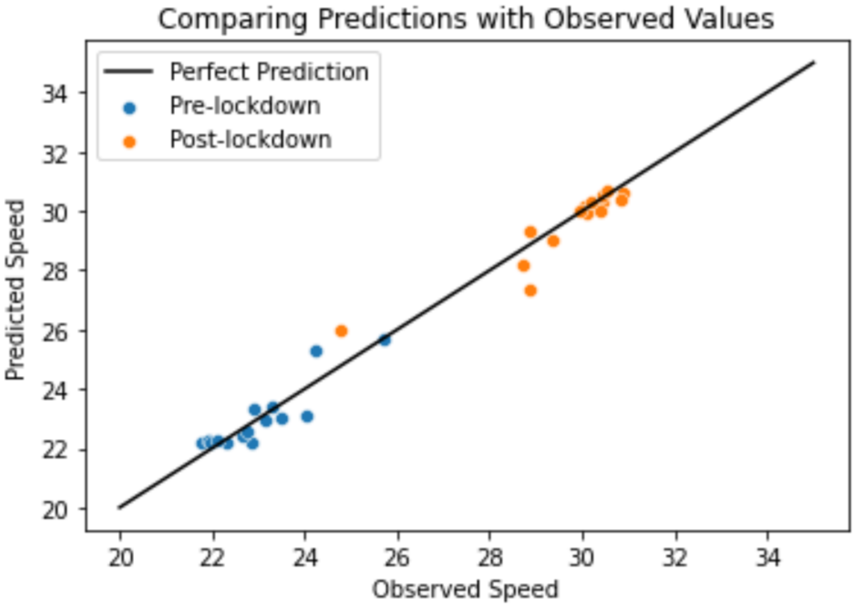
The features and weights of this improved model are summarized in the table below.



	feature	weight
0	is_weekend	0.979644
1	distance	32.270676
2	day	0.914251
3	dist_to_SF	1.233795
4	is_lockdown	-0.054651
5	is_fast	8.559913
6	is_wknd_and_lockdown	-0.380779

This model was used to predict travel speeds given a link the day, start point coordinates, and end point coordinates. The final model had a training RMSE of **9.33 mph**, and an r^2 of **0.394.** On the validation set we had an RMSE of **9.31 mph** and r^2 of **0.397.** We note that the new r^2 coefficient is a significant improvement over our baseline model.

We believe that our model has such low r^2 scores because there are many outliers and simply a very large amount of data in our training set. When we aggregate our predictions by day, our model performs reasonably well, as shown in the figure below. The black line with slope = 1 represents a perfect prediction. If points are close to this line it indicates that our model is performing well. Of importance, we note that there is not a difference in performance based on whether the day is pre- or post-lockdown, which is an improvement over our baseline model.



Next Steps and Future Modeling Work

In the current project, we tested whether we can make accurate predictions about speed, based on traveled distances, day of the month in March 2020, as well as features including whether the data was sampled on weekends/weekdays or pre-post lockdown. Our work has shown that on average, the regression model we used can provide reasonable predictions about speed. However, there's still room for improvement. Here are several suggestions for future work:

1. Separate data based on the types of areas (residential, industrial, highways, bridges, etc.). Current data is lacking such segregation. For instance, we only know speed information by the provided code, but we don't know what kind of area that code is located in. Getting a sense of how the type of region affects speed (and/or travel times) could help generate insights for determining length of trips in different areas, and subsequently the ride prices.
2. Related to our work, it would be interesting to further investigate how speed varies by weekday in different kinds of areas, and how it changed after the lockdown.
3. We focused on making the pre/post lockdown speed predictions by adding the feature that codes for whether the data was sampled before or after the lockdown. However, we can also estimate the average amount by how much the pre-lockdown speeds differ from post-lockdown speeds, and integrate that information into the data directly (such as normalizing by that amount). This can smooth out the differences between the speed values in the two time periods, and help the models predict the speed better (i.e. in the event of future lockdowns).
4. In this project, working with the full data set proved to be a significant challenge, as the data contained a lot of missing values, and likely many outliers. Nevertheless, in a dataset like this it is hard to tell whether a value is an outlier or simply reflects a non-outlier feature of the data (i.e. high speeds on freeways or bridges are much higher than the average speeds, but that is a feature of the data rather than an extreme deviation).
5. Distance-and-day model that's agnostic to whether pre-post lockdown distinction still segregates the predicted values into two clusters that likely correspond to speeds measured before and after the lockdown. That is an interesting pattern, given that we have not yet integrated the pre-post lockdown features, but adding these features actually does improve the model score. We have not addressed this in our project, but extensions of this work could shed more light on this point.