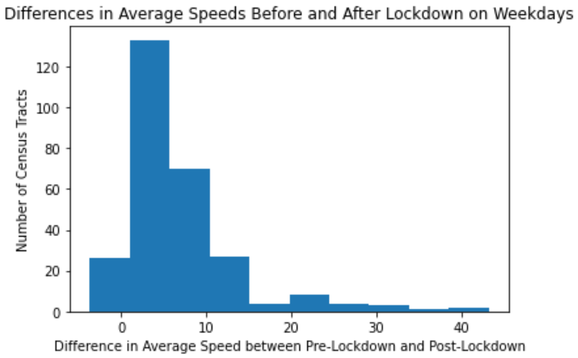
Jonah Grossman
Mitchell Bloemker
Sifath Mannan

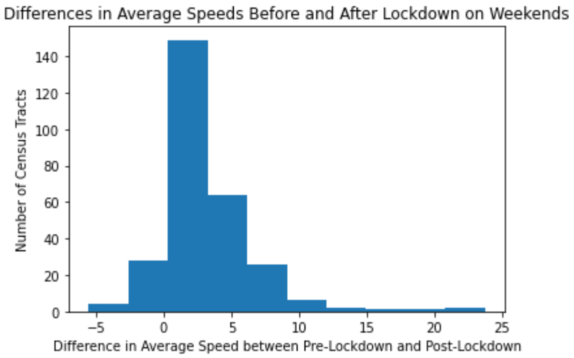
Data 100 Final Project Design Doc

In part 1 of the project, we mapped traffic speeds to Google Plus Codes as well as census tracts, finding that census tracts capture more meaningful subpopulations because they divide space according to neighborhood type and population instead of arbitrary rectangles across the city. We then began to tackle the question of how did the covid-19 lockdown affect traffic speed?

First, we split the dataset into pre-lockdown and post-lockdown, grouping by census tracts. This gave us average speeds, pre-lockdown and post-lockdown, and we created histograms to plot how average speeds were different. We also plotted the computed difference in post-lockdown vs pre-lockdown. We then looked at the impact of pre-lockdown average speed and post-lockdown average speeds on post-lockdown *change* in speed, computing the two correlations. Before running any code, our group expected both to have a positive correlation because we thought roads with high post-lockdown speeds would have seen the largest increase in speeds compared to pre-lockdown. As we expected, we found that the correlation between change in speed and the post-lockdown average speed was more strongly positively correlated compared to that of pre-lockdown average speed.

One intriguing portion of part 1 was visualizing the average speed by day, across all segments. After creating this line plot, our group became curious about all the fluctuations in the speed daily plot and began brainstorming some questions. The main question that our group began to focus on is whether or not there were differences in average speed on weekends versus weekdays. We then considered if there were large changes in traffic times pre-lockdown compared to post-lockdown on weekends and pre-lockdown compared to post-lockdown on weekdays. These initial ideas allowed us to use the given data from earlier in the project to split and manipulate our given dataframe into 4 new dataframes: pre-lockdown weekends, post-lockdown weekends, pre-lockdown weekdays, and post-lockdown weekdays. The next step we took was to take the difference between weekend average speed post-lockdown and weekend average speed pre-lockdown, as well as the difference between weekday average speed post-lockdown and weekday average speed pre-lockdown. We created the following histograms to visualize these differences.


Differences in Average Speeds Before and After Lockdown on Weekdays


Differences in Average Speeds Before and After Lockdown on Weekends

The histogram on the left shows the difference in average speed pre and post-lockdown on weekdays. Looking at the plot, we notice that average speed seemed to increase on weekdays post-lockdown because a majority of the census tracts had positive differences. Something in particular I noticed about this plot is that there were a lot of census tracts whose difference in average speed was larger than 25 mph. This is fascinating because if we compare this to the weekend histogram to the right we notice that there were larger differences in average speed on weekdays compared to weekends. A possible explanation for this could be that pre-lockdown people were driving to work which causes traffic congestion, which is also not a factor on weekends. However, when the lockdown began, people were not driving to work anymore, increasing average speed drastically on weekdays.

The histogram on the right shows the difference in average speed pre and post-lockdown on weekends. Similarly, this plot shows that the average speed generally increased on weekends post-lockdown. One interesting thing we noticed was that there are some negative differences in average speed on weekends, which would mean that the average speed pre-lockdown was greater than that of post-lockdown for some census tracts. Even though the weekends histogram did not have as large of differences as that of the weekdays histogram, there were more census tracts on weekends who had a difference in average speed around 0 to 10. One explanation for these negative differences could be that, post-lockdown, people used weekends to go out and get essential resources, creating more traffic congestion in certain commercial census tracts.

An open ended question we had was is there a way that we could classify each census tract by a category that would tell us more about what areas went through the biggest differences post-lockdown? For instance, could we cluster similar census tracts, such as highways or residential areas? Also, if we continued to expand further using weekends versus weekdays, could we find data on the average time for public transportation, such as the BART, to get from census tract to census tract? Would these times be different on weekends compared to weekdays? We also considered finding a way to map each destination's census tract to a category such as school zone, residential district, commercial, downtown, etc. Can we find correlations between "differences in bounds" (from Hayes Valley dataframe), i.e how much the route's travel time varies throughout the day, based on the destination's type of neighborhood?

In part 2 of the project, our group trained a model to predict traffic speed. Before building a model to predict daily traffic speed in San Francisco, we first created helper functions to assemble a dataset to predict daily traffic speed. After creating these helper functions, we converted our 'time series' dataframe to a numpy dataset and obtained our X_train, y_train, X_val, y_val, which was used to train and evaluate our linear model on pre-lockdown data. By doing so, we found that our first model is quantitatively and qualitatively accurate. Next, our group used the previously trained linear regression model and evaluated on post-lockdown data. Because the given dataset is distributed spatially and temporally, our model fails on post-lockdown data. To help visualize and understand where our model fails, we made a line plot showing performance of the original model throughout all days in March 2020. After doing some thorough analysis on this plot, our group concluded that this model was a bad predictor of model performance temporally. To give an example of why we concluded this, we chose to investigate the lowest point on the plot with the worst model performance, which was March 17th. Because the model we used to predict uses the traffic speeds of the last 5 days and for March 17th, 2 of the speeds included were prior to the shutdown, 2 of them were post shutdown and one of them was the exact day of the shutdown, the 17th had the worst model performance. We then began to combine our insights in model errors with EDA insights to produce a “fix” model on post-lockdown data. We did so by creating a new time series dataset using our daily average speed from pre-lockdown and then used this new model, time_series_delta, to train and evaluate our new model on pre-lockdown data. We found that the delta model was extremely effective for some days in the month but there are some days where the model is ineffective (e.g. March 14th). Doing this part of the project has taught us to try different models to see which model best predicts traffic speed and has the best accuracy.

Looking at the Hayes Valley dataset given at the end of part 1, we were curious how upper and lower bound travel time to the destination census tracts played into how these routes were affected by the lockdown. We saw that some destinations had a much larger difference in upper and lower bound travel times, and we were curious if the difference in these bounds said anything meaningful about the route from Hayes Valley to the various destinations. For each destination tract, we interpreted the difference in upper bound and lower bound as a measure of consistency in traffic speeds for that route throughout the given day. A low difference in bounds indicates the travel time was more consistent between the fastest and slowest times of the day. On the other hand, a large difference in bounds indicates that, over the course of the given day, the route was very slow at some times and much quicker at others, suggesting the route was affected by traffic congestion.

Drawing from the insight generated in part 1, we knew that overall traffic speeds increased in San Francisco post-lockdown, so we expected to find lower bound sizes after the lockdown. For instance, routes with highways would likely see a decrease in bound size post-lockdown because traffic congestion would decrease, and thus travel times would be more consistent throughout the day. With this in mind, our initial EDA led us to question whether, after the lockdown, routes with low day-to-day congestion (lower bound sizes) had experienced a larger decrease in bound sizes compared to pre-lockdown.

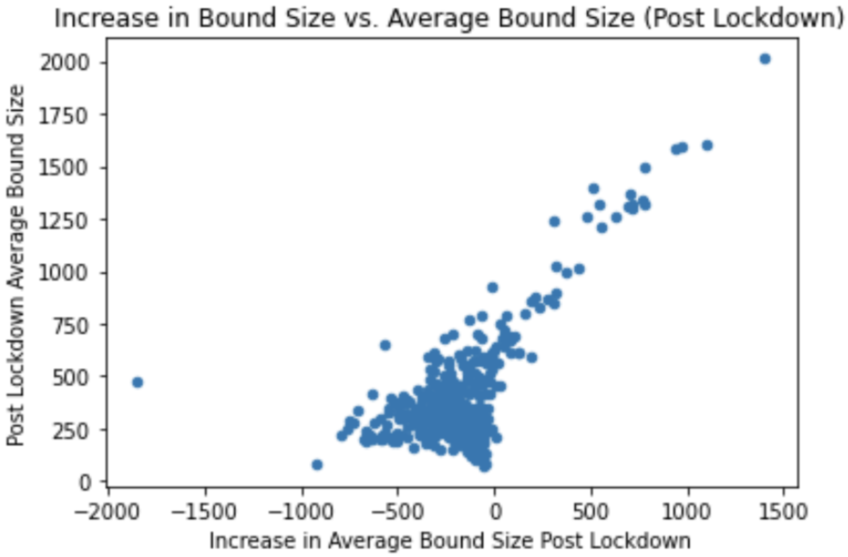
We hypothesized that these routes with the lowest differences in bounds (i.e. the routes with more consistent speeds), would have seen the largest decrease in bounds compared to pre-lockdown. Specifically, we hypothesized that there would be a high (> 0.8) positive correlation between each route's difference in upper & lower bounds (averaged post lockdown), and the route’s average decrease in bound differences before and after the lockdown.

In order to accept or reject this hypothesis, we would compare the correlation coefficient we obtain using the two desired features with a critical value of 0.8. In other words, if our obtained correlation coefficient is greater than 0.8 we would consider our test to be statistically significant and therefore would accept the null hypothesis. If our correlation coefficient is less than 0.8 then our test would not be considered significant and therefore we would fail to reject the null hypothesis.

To do this, we would need to manipulate the data and create several features using existing datasets. We started by computing the difference in upper bound and lower bound travel times for each destination tract on each day in the Hayes Valley dataset, recording the value in a column named “Bound Size”. We then split the dataframe into two sets, pre-lockdown and post-lockdown, and grouped them both by “Destination Movement ID”, aggregating by mean. This gave us each route’s average difference in upper and lower travel time bounds, averaged across the pre-lockdown and post-lockdown eras.

We then merged both sets back into a final dataframe and computed the difference between pre-lockdown and post-lockdown average bound sizes for each route. This value, stored in a column "Increase in Average Bound Size Post Lockdown", would essentially measure how much each route’s travel time went from being consistent throughout the day to more “stop & go” post-lockdown. As we expected, the general trend was that traffic became more consistent, i.e. less affected by “stop & go” congestion throughout the day, as indicated by an average *decrease* in bound size from 573 seconds to 386 seconds.

However, in order to accept or reject our hypothesis, we computed the correlation between “Increase in Average Bound Size Post Lockdown" and "Post Lockdown Average Bound Size". A high correlation would indicate that, post-lockdown, routes with low bound sizes (i.e more consistent speeds) would have seen larger decreases in bound size after the lockdown. We computed a correlation of 0.706, which is strong, but still lower than 0.8, so we rejected our hypothesis.



The biggest drawback of this model is that it only used simple correlations instead of a predictive model to make valuable predictions using training and test data. In other words, while we were able to establish a moderately high correlation of 0.7, it failed to establish statistical significance (> 0.8) , and didn’t allow us to make predictions on unseen data. Looking forward, we wanted to be able to test the accuracy of our model, which would mean training and testing it on subsets of the original dataset. This would allow us to generate more meaningful insight instead of simply finding the association between computed features.

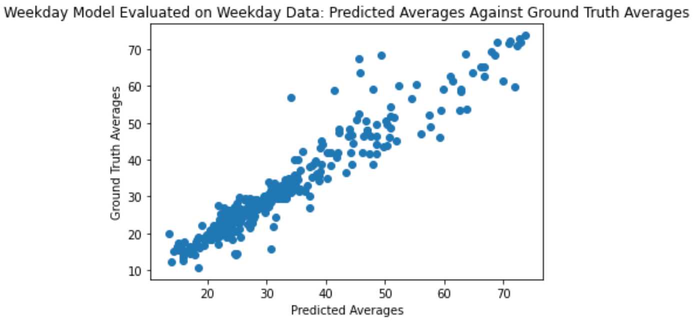
We realized that there are some improvements that could be made in our original hypothesis and tried to focus more on combining our original EDA insights, which investigated pre-lockdown and post-lockdown average speed for weekends versus weekdays, with a new hypothesis that we could tested by creating a model, evaluating it on weekend and weekday data and observe the accuracy of this new model.

As our group continued to make new models and to use these models to train and evaluate on certain data, it made us wonder if we could combine our insights in model errors with our original EDA insights about weekends versus weekdays. In particular, could we use a certain model to predict traffic speeds on any given weekday using only weekday data or predict traffic speeds on any given weekend using only weekend data.

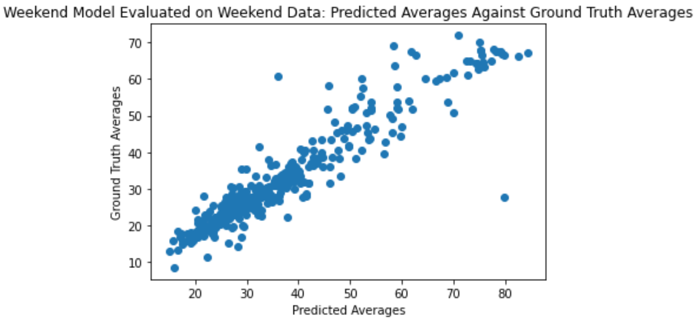
This invoked our group to come up with a new null hypothesis, which is that we believe that a weekday dataset can predict weekday or weekend average speeds greater than a weekend dataset can predict weekday or weekend average speeds. In other words, creating a model using a weekday only dataset can predict weekday and weekend traffic speeds close to a 90% accuracy. When testing this, we would fail to reject our null hypothesis if our weekday dataset predicts average speeds close to a 90% accuracy and reject our null hypothesis if our weekday dataset predicts average speeds much less than 90% and/or the weekend dataset predicts average speeds at a much higher accuracy than the weekday dataset.

In order to carry out this hypothesis, we first began by creating two different time series; one being a weekend only time series and the other one being a weekdays only time series. With these two time series, we then used our time_series_to_dataset helper function to create a dataset for our weekday model and for our weekend model. Once these were computed, we trained our weekday model on weekday data and on weekend data and evaluated our weekday model on both weekday and weekend data. We then did the same thing but for our weekend model, where we trained our weekend model on weekday data and on weekend data and evaluated our weekend model on both weekday and weekend data.

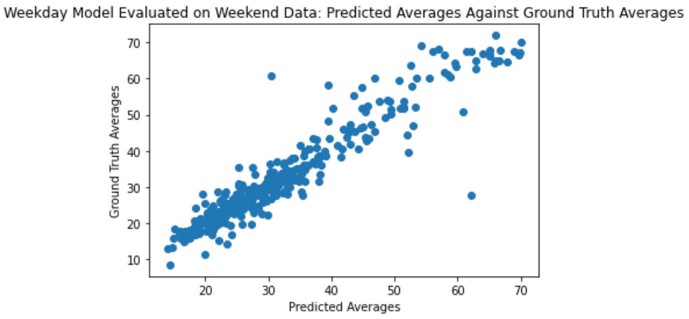
The first model we produced was a weekday model trained to a linear regression model. Once we evaluated our weekday model on weekday data and set our model to a trained linear model, we evaluated our accuracy by using the score function and to visualize the predicted averages against ground truth averages, we used the predict function in order for our model to predict average speeds. We then did the same exact steps to produce a weekend model trained to a linear regression model evaluated on weekend data and obtained our model accuracy and visualized our predictions. After seeing that the weekend model produced a low accuracy when evaluated on weekend data, we became curious about whether or not the weekend model was a good predictor of weekday data. Hence, we carried out the same steps to produce a weekend model trained to a linear regression model evaluated on weekday data and obtained our model accuracy and visualized our predictions. Once again, we obtained a low model accuracy when using our weekend only model. Therefore, to gather the last piece to test our hypothesis, we produced a weekday model trained to a linear regression model evaluated on weekend data and obtained our model accuracy and to visualize our predictions.


Weekday Model Evaluated on Weekday Data: Predicted Averages Against Ground Truth Averages

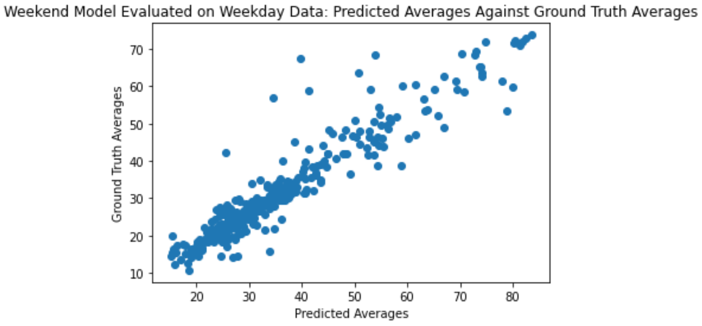
Score = 0.9066917388342355


Weekend Model Evaluated on Weekend Data: Predicted Averages Against Ground Truth Averages

Score = 0.7517028676741524


Weekday Model Evaluated on Weekend Data: Predicted Averages Against Ground Truth Averages

Score = 0.8984138177424913


Weekend Model Evaluated on Weekday Data: Predicted Averages Against Ground Truth Averages

Score = 0.7746658781192943

For each model we noticed a linear trend between our predicted average speeds and the ground truth average speeds. Given that these two variables were used as inputs, we used linear regression to model this relationship for all 4 models we created.

When analyzing the visualizations seen above, we noticed that the weekday model was much more accurate when evaluating on weekday data and weekend data compared to that of the weekend model. To prove this, we obtained our model accuracy, which is the score, and visualized our model predictions using scatter plots of predicted averages and ground truth averages. When evaluating our weekday model on weekday data we received an accuracy of 0.9066917388342355 which is greater than our null hypothesis accuracy. Next, we had to check how accurate the weekend model was when evaluating on weekend data. This accuracy was 0.7517028676741524, which was much lower than the weekday model evaluated on weekday data. This made us curious so we wanted to check how accurately the models would predict average speeds on the opposite dataset. Hence, when we trained our weekday model on weekend data and evaluated this model on weekend data, we obtained an accuracy of 0.8984138177424913. This proves that the weekday model was a much better predictor of average speed compared to the weekend model. Lastly, to make sure our theories were correct, we trained our weekend model on weekday data and evaluated this model on weekday data and obtained an accuracy of 0.7746658781192943. This once again proves that the weekday model was a good predictor for average speed, obtaining accuracies close to and greater than 90%. Also, because our weekend model has much lower accuracy than our weekday model, we know that our weekday model is a much better predictor. Therefore, our resulting accuracy for our weekday models were both close to 90%, which is what our null hypothesis stated, and we can see that our weekday model was far more accurate being evaluated on both weekday and weekend data compared to our weekend model. Hence, we fail to reject our null hypothesis.

A key takeaway we gained from these models is that our weekday model has is a better predictor of average speeds pre and post-lockdown compared to our weekend model. However, there are some further model improvements that could be made to possibly fix our less effective weekend model. First off, a major improvement that could be made to the weekend model would be to add average speeds on weekends for the month before the lockdown and for the month after the lockdown. We believe the weekend model was a bad model because it didn't have enough data to be a good predictor, so if we included more data of average speed on weekends before and after the lockdown this model would be a much better predictor.

Looking back at both our initial and improved models, we found ways to sort and manipulate data in order to better understand and predict real world phenomena. If we were to continue the work completed in this project, we could combine these features we used for our initial and improved models, bound size and weekend/weekday, to create a more detailed profile for each data point (each route). Then, we could run a multiple regression model, assigning varying weights to these features and using loss functions to optimize the accuracy of our model. Furthermore, we could even bring in data from other sources to increase the number of features and thereby increase accuracy even more. For instance, we could incorporate a time series detailing weather conditions in San Francisco. Some of these features could include wind speeds and inches of rain for each given day. We could then incorporate these quantitative variables in a multiple regression, alongside our "bound size", which would allow us to create an even more detailed profile for each route, further improving the accuracy of our model and helping us better predict how traffic might change with future lockdowns.