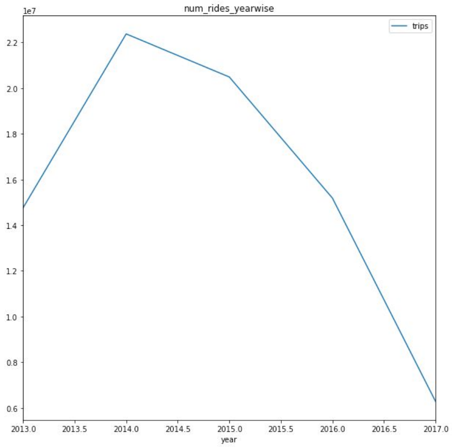
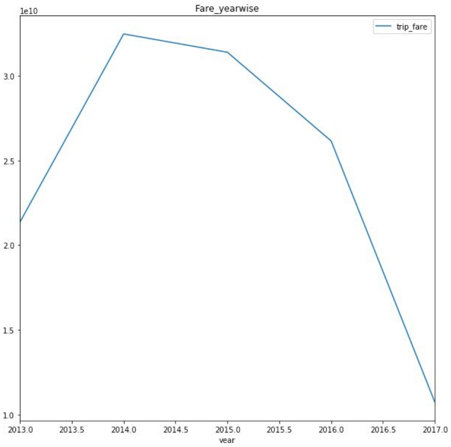
Colab Notebook
https://colab.research.google.com/drive/1hir7nv7ADCI2EBOKT_Zz4HgcSZ7ZIpB1

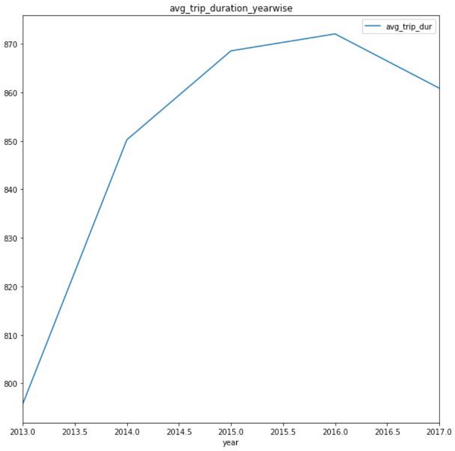
Exploring Chicago Cab Data

The first step in this study is Exploratory data analysis of the cab data. The aim is to analyze how different features in the data-set relates to the other variables. We want to identify the variables that have a predictive potency while predicting the daily and hourly trip duration of the taxis.

I begin by examining how Chicago cab business has changed over time, especially after the advent of the savvy competitors like Uber and Lyft. In Chicago, Uber and Lyft entered the cab business in 2011 and 2013 respectively. I plot the Total Yearly pickups Total Yearly Fare and Yearly Average trip duration graphs. From the first graph, it can be noticed that there is a sharp decline in the number of pickups with an annual rate of 35% after 2014. Total revenues per year follows the same trend with the total cab business at $18 million in 2016 as opposed to $35 million in 2013. This suggests Uber and Lyft has a grasp on the customer base with their competitive prices.

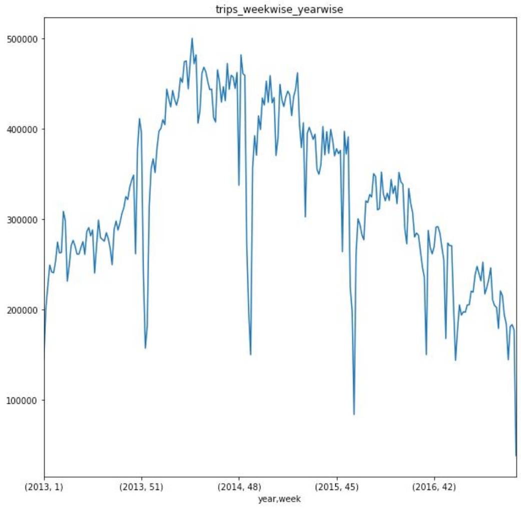




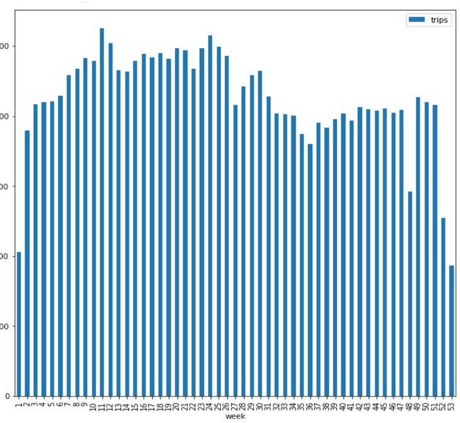


Surprisingly, we see a steady lift of 8.6% in the trip duration from 2013 till 2016 with a slight down adjustment towards end of 2016 or early 2017. This could be attributed to the fact that Chicago taxis are being booked only for long trips or outstation callings and the local commute needs are mostly met by Uber or Lyft.

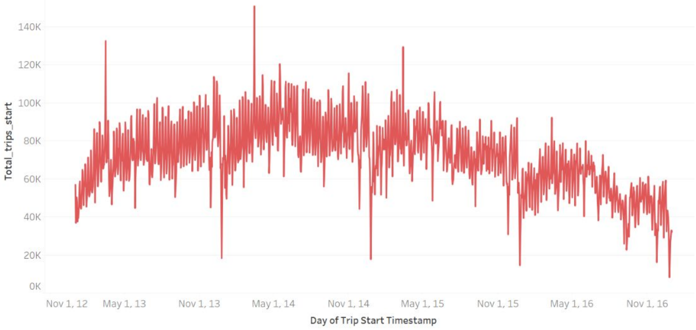
The decrease in taxi fares has caused huge economic burden on the cabbies as they aren't generating enough fares to keep up with their loan payments and meet their expenses. More than 350 foreclosure notices or foreclosure lawsuits have been initiated against medallion owners in the year 2017, compared to 266 in 2016 and 59 in 2015. Since October, lenders have filed lawsuits against at least 107 medallion owners who have fallen behind on loan payments, according to the union's count. The major reason behind this financial distress is that since the emergence of Uber & Lyft, Cabbies face an uneven playing field with the ride-share companies, who typically don't face the same permitting and fee rules.



Let's plot the total number of rides against the week number over the course of 4 years. After looking at the graph, its difficult to explain the variation in the number of rides every week of the year. But, we can notice a spike in week 10 and week 24. A similar downward spike can be noticed in the week 48 and 52. Week 10 contains the most important holiday celebrated widely in Chicago City i.e. St. Patrick’s Day. Week 48 and week 52 contains Thanksgiving and Christmas holiday.

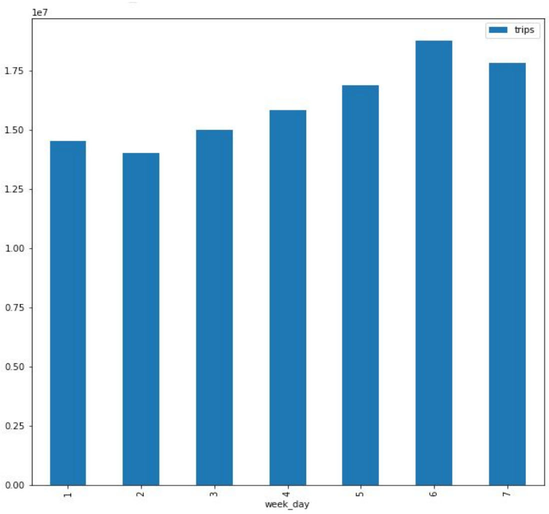


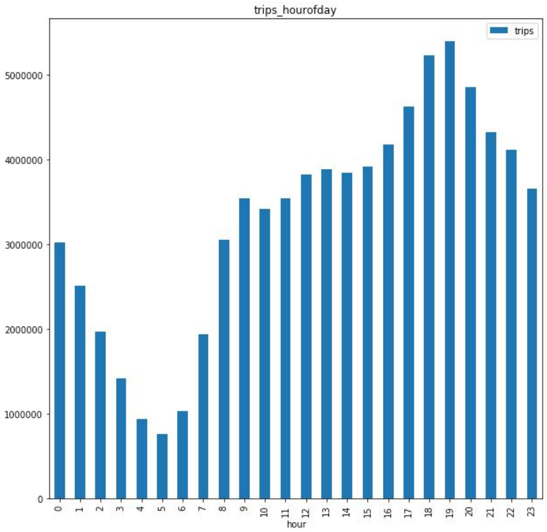
Holidays have always been special to the Chicago City. From raucous pub crawls to lively parades, there is nothing quite like St. Patrick's Day in Chicago. In March, Irish taverns are packed with revelers, jovial crowds jam the city streets and the Chicago River sparkles brilliant shades of emerald green. With so much going on, be it the downtown parade or dyeing of the Chicago river, Chicagoans travels a lot and they generally prefer public taxis.



It turns out that St. Patrick’s Day parade which is held in March of every year accounts for maximum number of taxi trips on a particular day.One of the important national holiday is Labor Day, created to celebrate the contributions of the American worker. It falls on the first Monday of September, resulting in a dearly coveted three-day weekend. Labor Day is essentially a day off for cabbies and this results in less number of ride on that day.

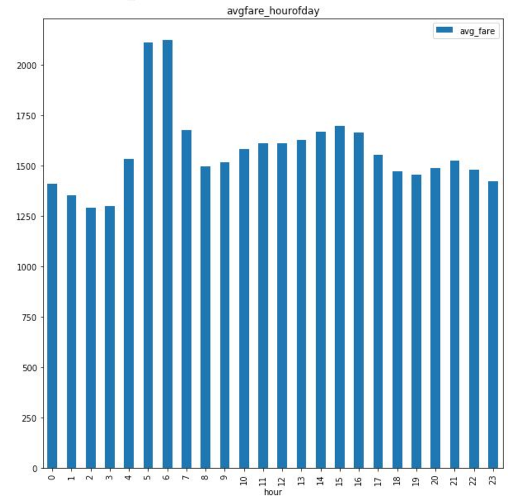
Memorial Day, Thanksgiving and Christmas of every year has the least number of taxi trips respectively. This is again due to the fact that there are few Taxi’s in service on these holidays. This motivates us to incorporate important holidays as parameters in our predictive model.

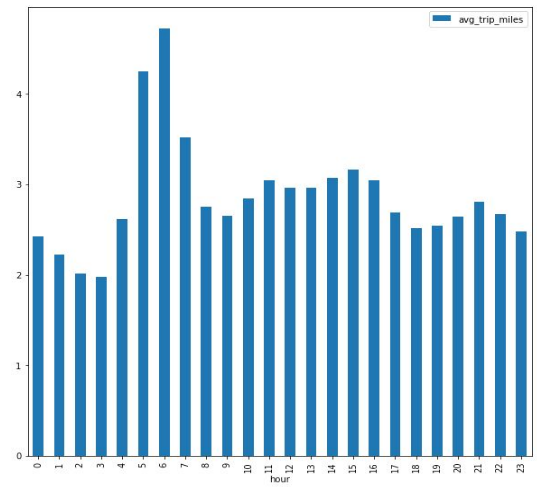




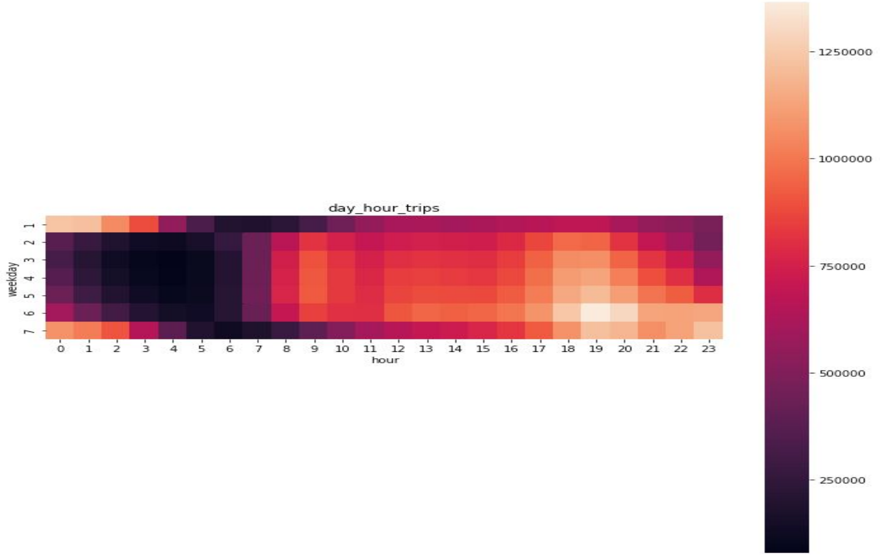
Now we analyze the number of pickups by the weekday and hour of the day. Friday receives the highest number of trips whereas Monday the lowest. Morning and evening rush hours are clearly visible, especially at 9 AM when people want to reach their offices and colleges, and at around 6-7 PM when they want to go back to their houses. A sharp decline in the pickups can be seen at 5th and 6th hours of the day.

To understand this better, I plotted the average fare vs hour of the day graph. Average fare per ride has peaked during the 5th-6th hour of the day. I analyzed this further to capture the reason for the increased prices during these hours by plotting the trip miles vs hour of the day.It was deduced why trip miles were skyrocketing during that interval. It turns out that number of drop offs at O'Hare International Airport (Second busiest airport worldwide) were maximum during the 6th hour of the day i.e. many Chicagoans uses taxi to board their early morning flights. Since O'Hare is located on the far Northwest Side of Chicago, Illinois, 14 miles northwest of Chicago's Loop business district (community responsible for highest number of pick-ups), trip miles and total fare are usually higher . Although the number of drops is greater in the evening as well, but the average remains low due to large number of rides.

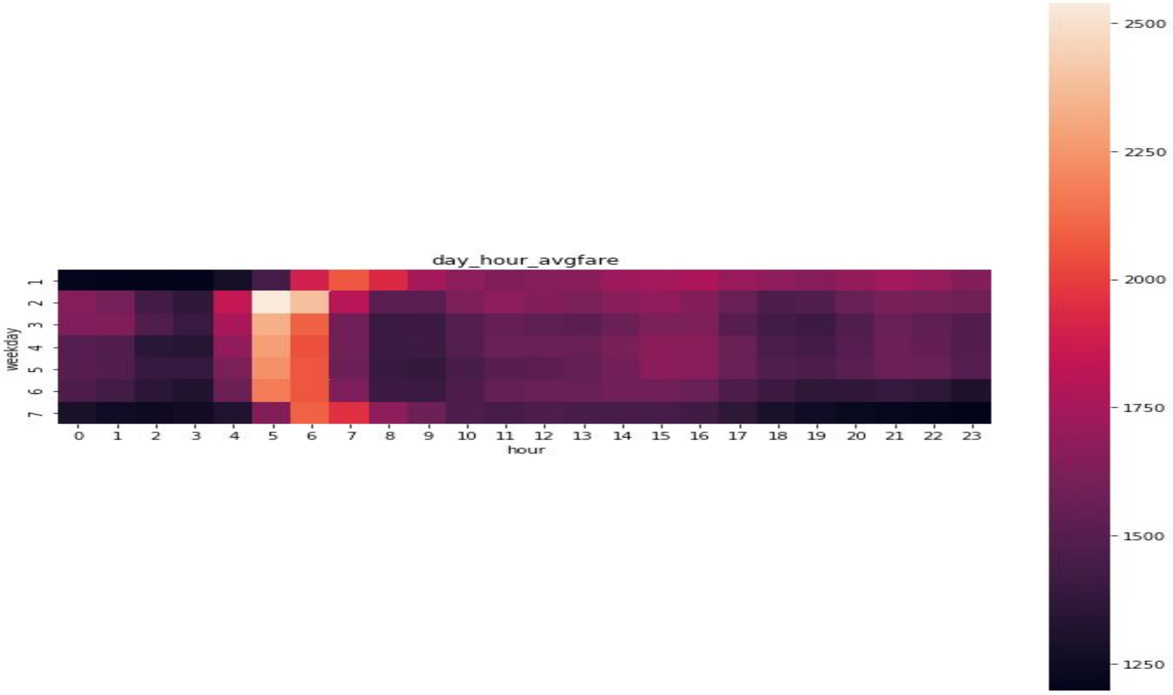




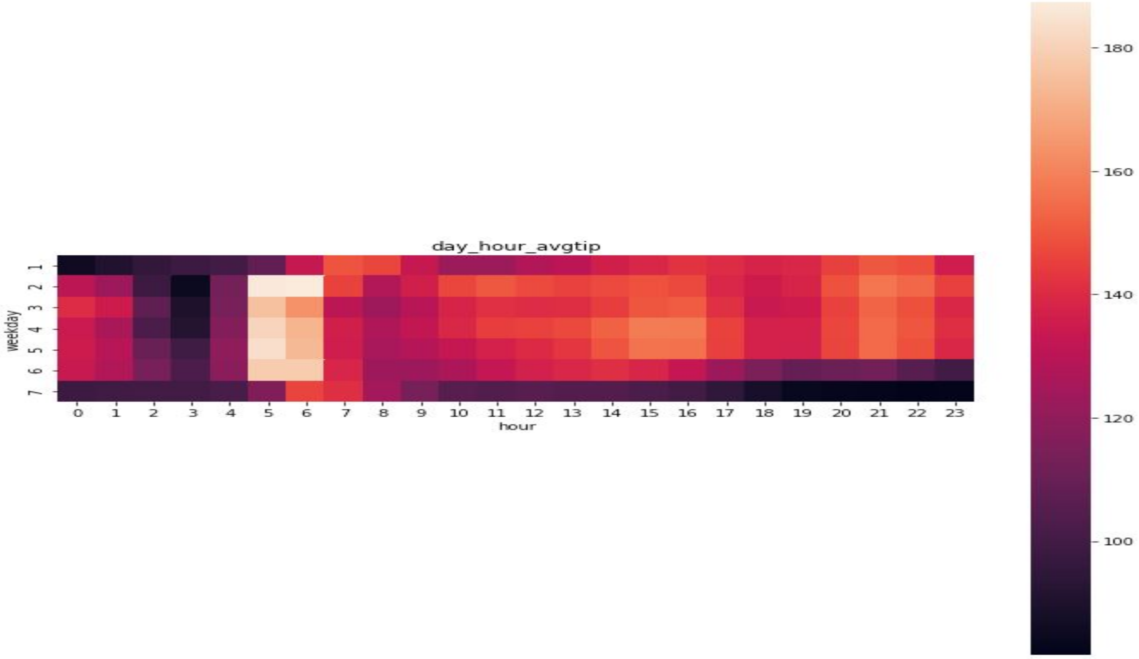
Weekly and hourly analysis made me curious and I wanted to see the combined effect of day of the week and hour of the day on number of rides. It would provide the exact day and hour at which Chicago Transport Authority (CTA) should have maximum number of active taxis. This heatmap gives a clear picture about dependence of both the parameters on number of rides. From 0-6 on Monday through Friday there are less number of rides, hence less number of taxis required to balance the demand. Chicagoans generally like to relax on Friday and Saturday, traveling a lot within the city, drinking which gives rise to more number of rides. Also, talking about allocation of resources CTA should have large number of taxis from 0th hour to 6th hour on Saturday and Sunday. Heatmap for Total trip is similar to the number of rides.



As mentioned previously, heatmap is entirely different for average fare. If some of the cabbies are looking to make more money per ride, following heatmap is very useful for them. It paints a profitable picture where there is high return on investment, if a cabby is willing to serve from 4AM-7AM. The traffic density at this time is less which aides higher fuel efficiency.

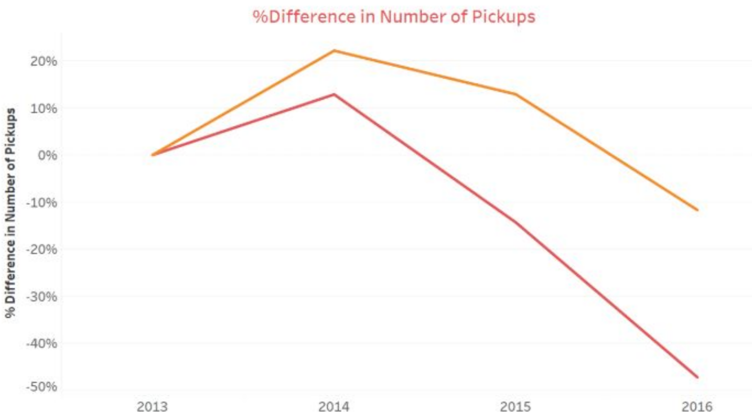


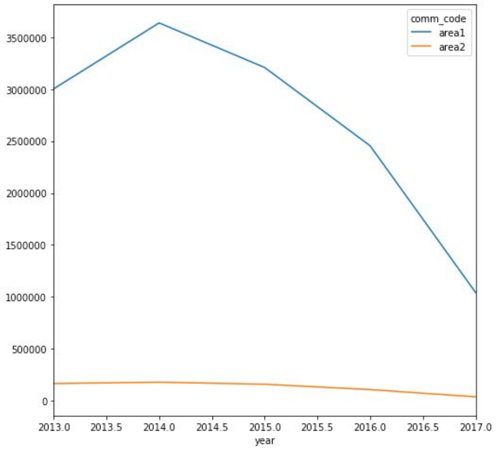
Also keeping in mind that since these rides are lengthier time and distance wise, from our own experiences, we have always tipped more in such situations. Hence, we plotted the similar heatmap for average tips, our intuition was right as many people appreciated the work done by cabbies and that too at 5AM or 6AM. Average tips were fairly high in the night time as well.



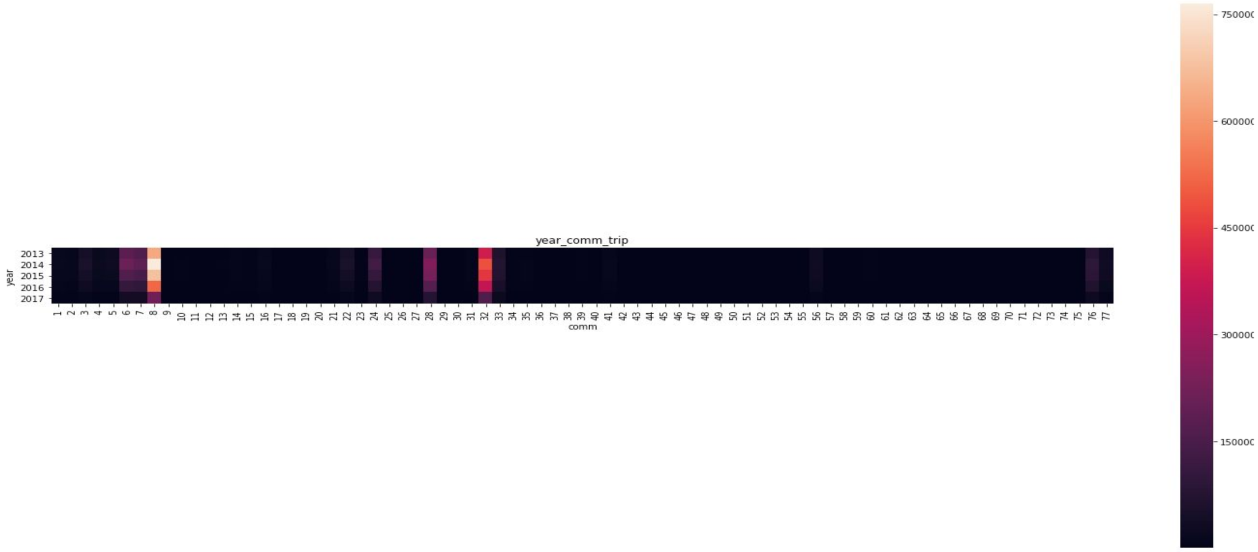
So far, we understood the importance of time features. Next, we wanted to understand the importance of pick up and drop off location.

Chicago’s taxi pickup declines are not evenly distributed among the city’s 77 community areas. For example, the Loop, Chicago’s central business district, shows a 23% annual decline, while Logan Square on the northwest side shows a 50% annual decline. In general, the areas located closest to the central business district show smaller declines in taxi activity.





We defined 5 particular community areas—the Loop, Near North Side, Near West Side, Near South Side, and O’Hare Airport—as “area1”, then compared pickups inside and outside of the area1(area2). As of November 2016, pickups inside area1 shows a 27% annual decline compared to a 42% annual decline outside area1. On a cumulative basis, area1 pickups have declined 39% since June 2014, while area2 pickups have declined a whopping 65%.



Model Development (Approaches, Results, Improvements)

Initial Preprocessing

Train data(2013-2016) and test data(2017) were filtered for non-zero values of trip miles and trip durations. Post filtration, Null rows appeared in pickup and drop-off coordinates (<7%), which were dropped. A standard scaler (on continuous data) and one-hot-encoding(on categorical data, such as day of week) is sought to be done given the data could be pulled in Colab. A PCA transformation on Pickup and Drop-off coordinates was attempted to label coordinates followed by clustering to group pick-up and drop-off areas.

1. Baseline Bigquery Linear Regression Model

For the first attempt at predicting trip duration using minimal features, I utilised Bigquery inbuild linear regression function. This model has the following input variables with their reason for selection

Variable	Source/Calculated From	Reason for selection
Trip Miles	Latitude, Longitude of pickup and dropoff	Intuitive, EDA revealed high importance of distance covered for trip durations[6]
Day of Week	Pickup Timestamp	Particular days of a week reacted differently to trip durations
Day	Pickup Timestamp	Particular days reacted differently to trip durations
Hour	Pickup Timestamp	Particular hours of a day reacted differently to trip durations

The model performance on 2017 data for prediction of trip duration were recorded as follows:

Baseline Model Performance on 2017 data

MAE	MSE	MSLE	Median Absolute Error	R2 Score	Explained Variance
508.3	934481.78	0.64	389.9	0.069	0.069

2.XGBoost Regressor Model trained on 2016 data

I trained one xgboost regressor on 2016 data with the same features as the baseline model. Starting with a random set of parameters (colsample_bytree = 0.3, learning_rate = 0.1,max_depth = 5, alpha = 10, n_estimators = 500) I achieved the following performance for a 70-30 train-test split on 2016 data

Training Results on 2016 data

RMSE	MAPE
780.54	66.41

The feature importance revealed the order **trip miles, hour, day of the week and day** w.r.t their **f-score** in the model.After that, I went to cross-validate this model with varying set of parameters and boosting rounds to achieve a stable performance across train and test splits.The final parameter observed were: *('colsample_bytree': 0.9,'learning_rate': 0.1, 'max_depth': 12, "tree_method": "gpu_hist")* & number of boosting round for optimal error values ~ 32.
The final cross-validation scores we as follows:

Cross-Validation Results on 2016 data

Test MAE mean	Test MAE std	Train MAE mean	Train MAE std
274.05	1.226	273.4	0.780

As we can see, the MAE with this model is much more lesser (around 300 less) than our baseline model. The performance of this model on 2017 data were as follows:

MAPE	**MAE**	**RMSE**
59.5681	267.379	774.32

3. XGBoost model trained on 2016 data with broader set of features (PCA on coordinates, clusters of PCA's, Speed, Direction, Holidays etc.)

A total of 54 features were curated to best fit the trip duration prediction. The preprocessing part was completed end-to-end and cross validation followed by grid-search to obtain optimal model parameters were initiated but the colab notebook collapsed multiple times owing to large RAM requirements in spite of turing the GPU mode on.

A detailed comment section for each step has been added for this part of the code and could be run error free given ample processing and storage space.

Final Results

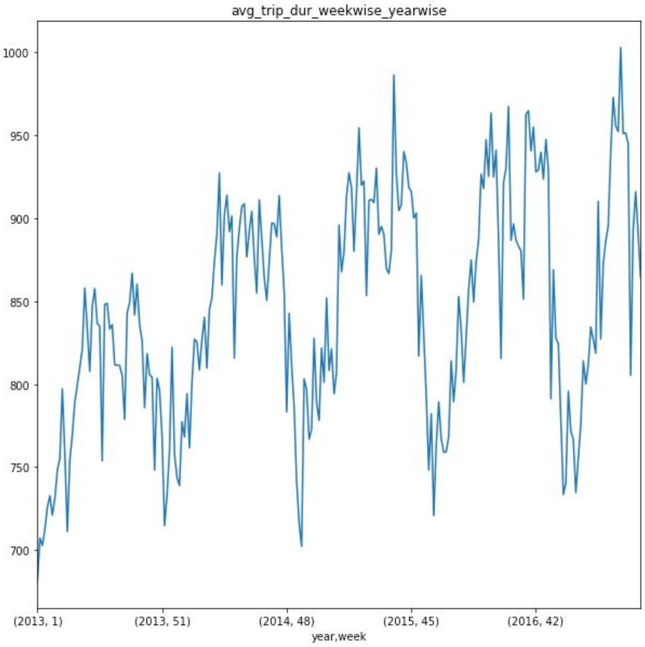
1. **The best MAPE observed for trained model was 59.5681 on 2017 dataset.**
2. **Only 1 year of data(2016) was required to build a model with consistent performances across unseen data sets.**

Areas of Improvements

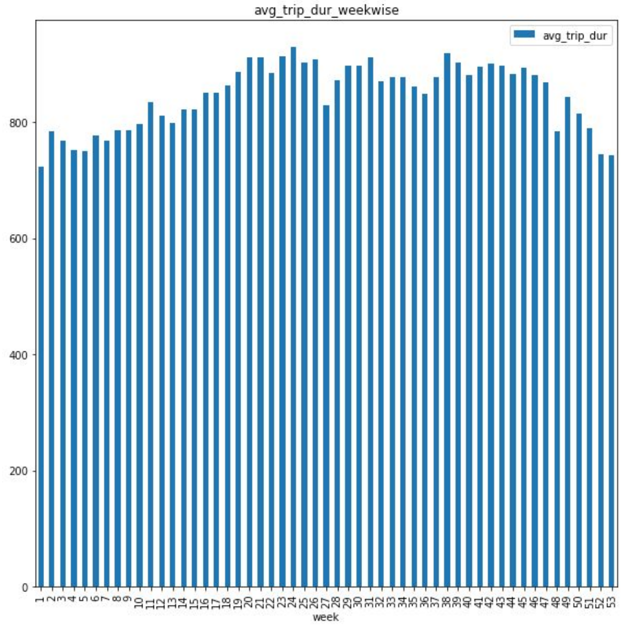
1. **Given better resources(RAM, Disk), the 3rd model would have been run end-to-end with an expected decrease in the MAPE value.**
2. **More concrete and diverse features, as attempted, would give more accurate interpretability towards trip duration.**

APPENDIX

Average trip duration Vs. Week of the year

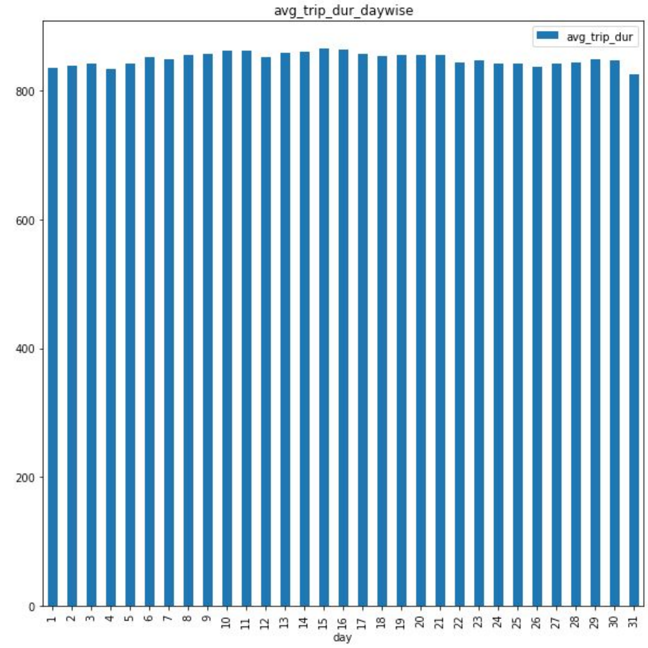


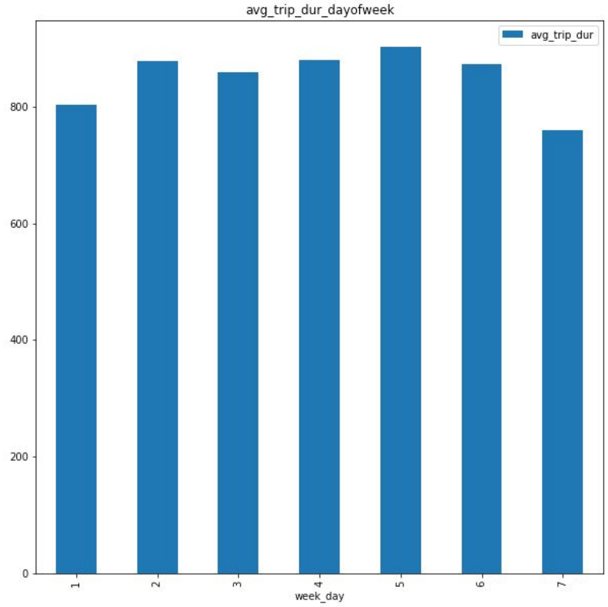
Average trip duration - Week wise



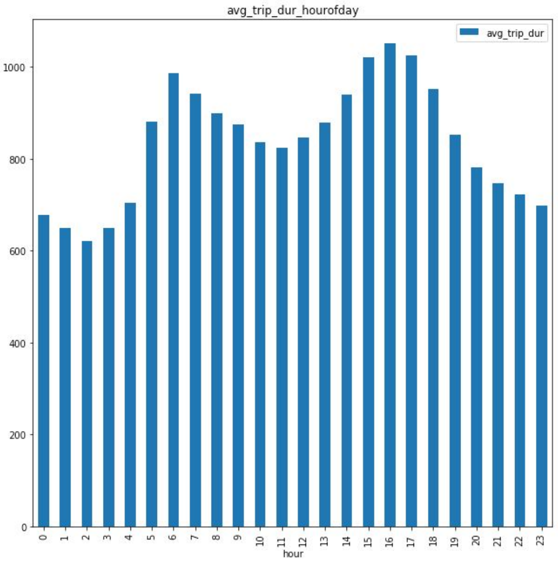
Average Trip Duration Vs. Day

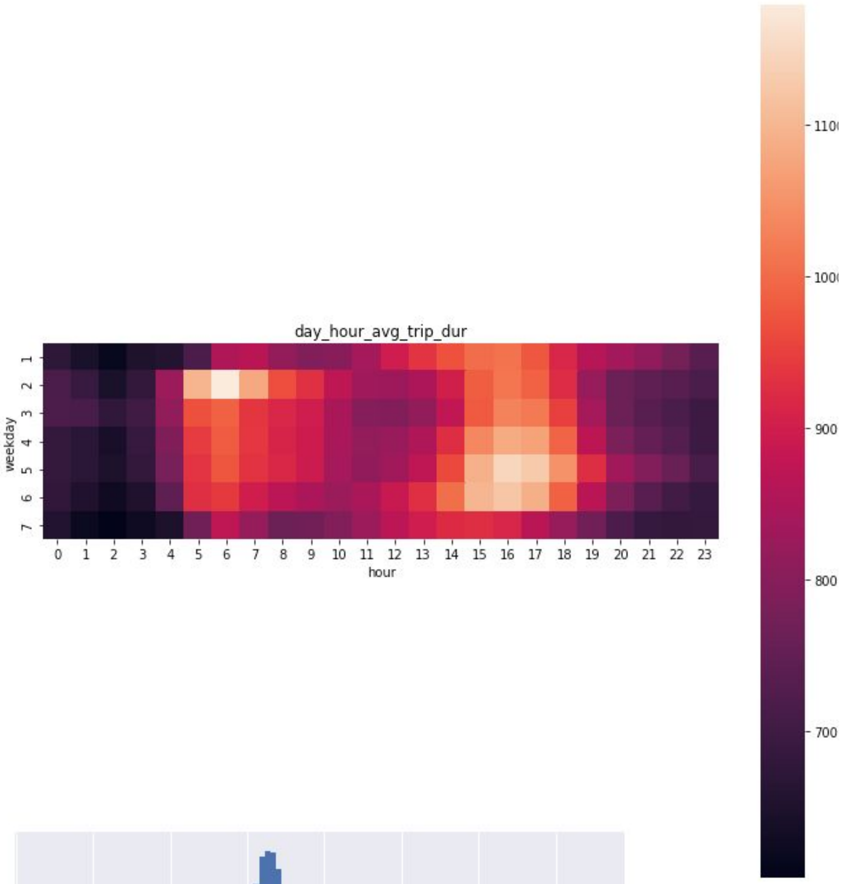
Average Trip duration Vs. Day of the Week



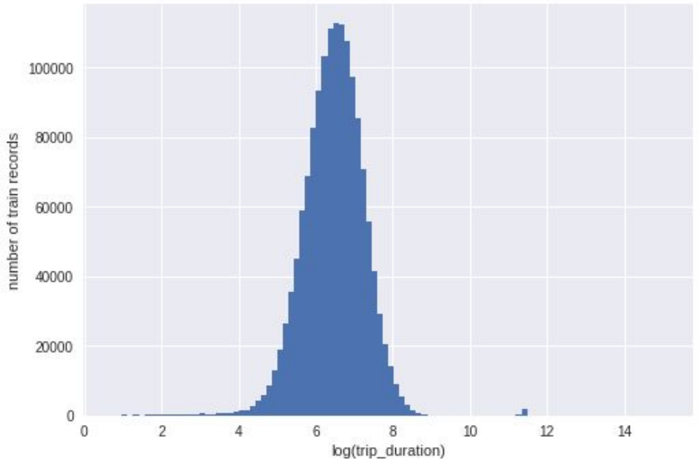


Average trip duration for hours



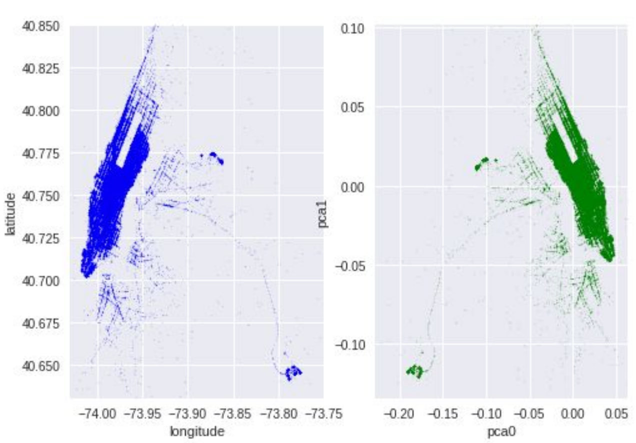


Log transform of trip duration

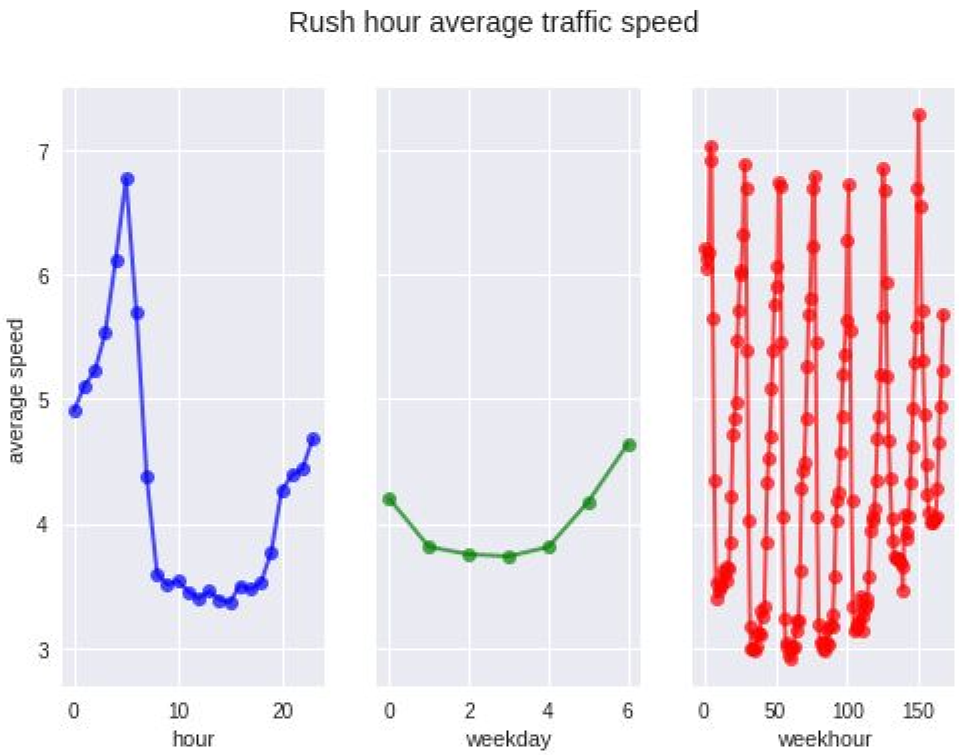


Original and PCA transformed Lat-Long

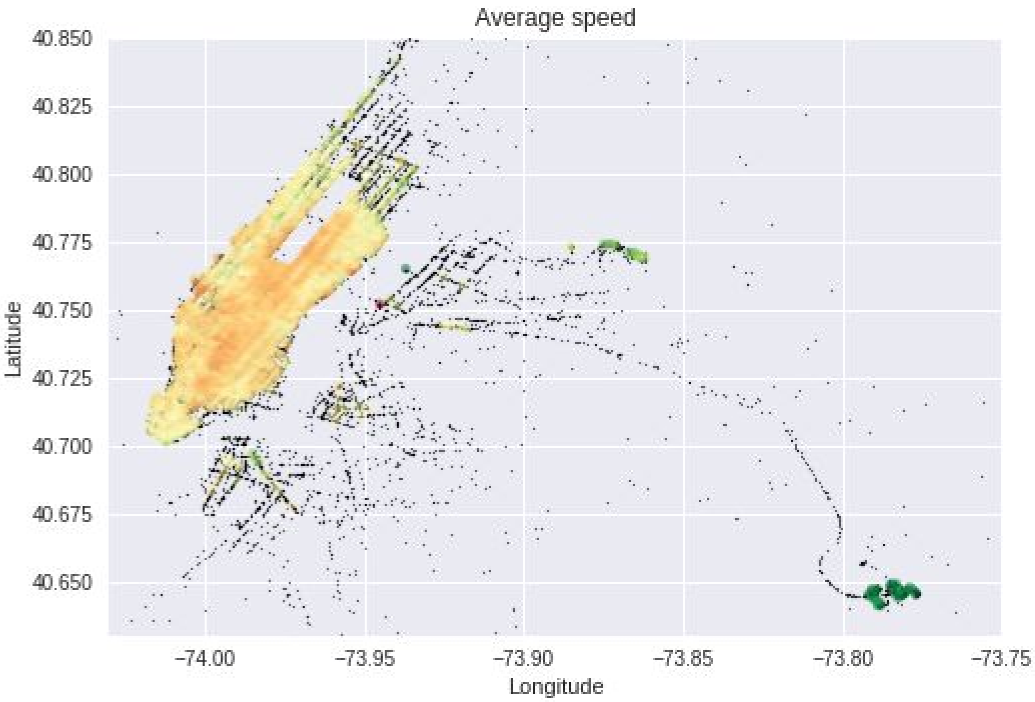
Pickup lat long coords and PCA transformed coords.



Speed



Rush hour average traffic speed
average speed
3
4
5
6
7
0
10
20
hour
0
2
4
6
weekday
0
50
100
150
weekhour



Average speed
Latitude
40.850
40.825
40.800
40.775
40.750
40.725
40.700
40.675
40.650
−74.00
−73.95
−73.90
−73.85
−73.80
−73.75
Longitude