Proposal

This is meant to be an analysis of Uber and yellow cabs in the Manhattan borough of New York City. We will gain insight into patterns in pickup locations, trip distances, and peak travel hours. At the end of the analysis, we will develop a machine learning that can predict the duration of a cab ride.

Data Aquisition

Two datasets were used for this project. They were acquired from FiveThirtyEight who in turn obtained the data from the NYC Taxi & Limousine Commission (TLC) by submitting a Freedom of Information Law request on July 20, 2015.

The Uber pickup data is from April 2014 to September 2014 and, Jan 2015 to June 2015.

The yellow cab data is just for April 2014.

Data Cleaning

Uber

The 2014 data is read in individually by month We use pd.concat to combine the datasets into one pandas data frame called ‘df’. The dataframe contains 4,534,327 entries and 4 columns. The columns contain the following information;

1. Date/Time, object type.
2. Latitude, float type
3. Longitude, float type
4. Base, object type. There are only 5 different dispatching bases for NYC in this dataset.

The 2014 data contains no NA values.

‘df\_15’ the dataframe with the 2015 data contains 14,270,479 entries and 4 columns. The columns contain the following information;

1. Base number, object type. There are only 8 different dispatching bases for NYC in this dataset.
2. Pickup date, object type
3. Affiliated base number, which is the same as the base number, object type.
4. Location ID, int type.

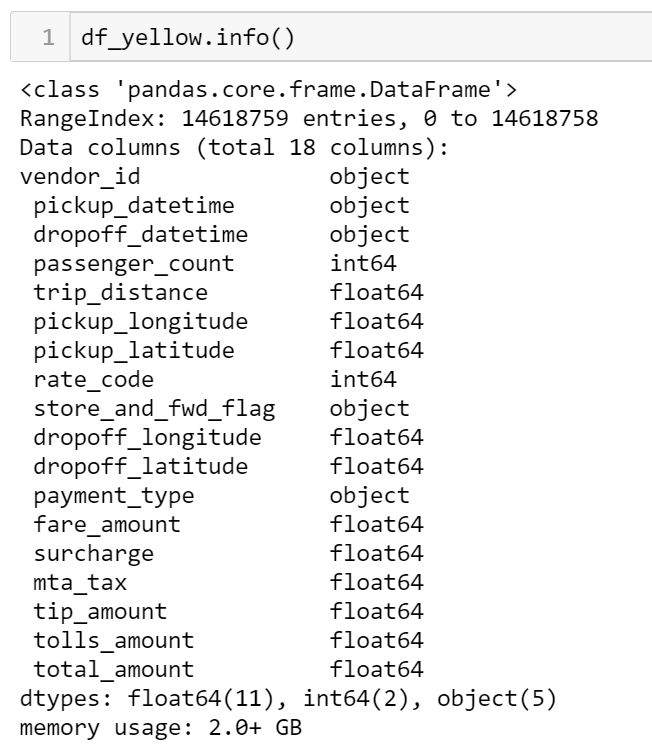
df\_15 doesn’t contain any NA values either.

We decided to use only the 2014 dataframe ‘df’ for further analysis because both data frames do not contain the same information and ‘df\_15’ does not contain latitude or longitude information which could be valuable information later in the project.

The ‘Date/Time’ column in df contains entries that are of type string. We will convert these entries to datetime objects as they will be more convenient to work with and allow us to perform time series analysis. To do this we first slice df into only its first 500 entries this is more convenient as it improves computation time and we can scale it up for the full data frame later. We use dt.strptime to convert one entry into datetime type, once we are sure that it works we create a function that iterates over df500 converting the Date/Time entries from a string type to a datetime object which is then stored in a column called ‘datetime’. We then drop the original column ‘Date/Time’ and are left with 4 feature columns as before. This clean dataset is then saved as ‘uber\_clean\_data.pkl’.

Yellow Cab

The yellow cab data we downloaded is for the month of April in the year 2014, we did this so we would have a similar time frame as the uber data we are working with.



There was a space in front of all the column names which we fixed using pandas’ rename function.

We dropped the ‘store\_and\_fwd\_flag’ column. And converted the drop off and pickup time columns to datetimes so they are easier to work with. We also used pandas value counts to find any outlier values that need to be omitted.

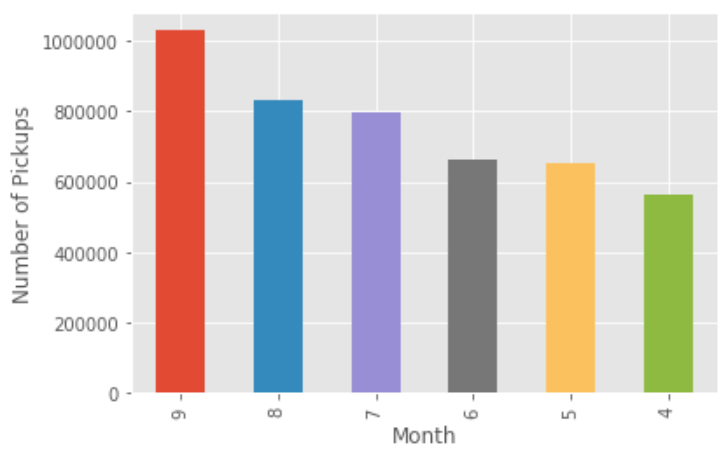
We also added a column called ‘total\_time’ that is the difference between the dropoff and pickup datetime values. ‘Total\_time’ contains time delta values

Exploratory Data Analysis

Uber

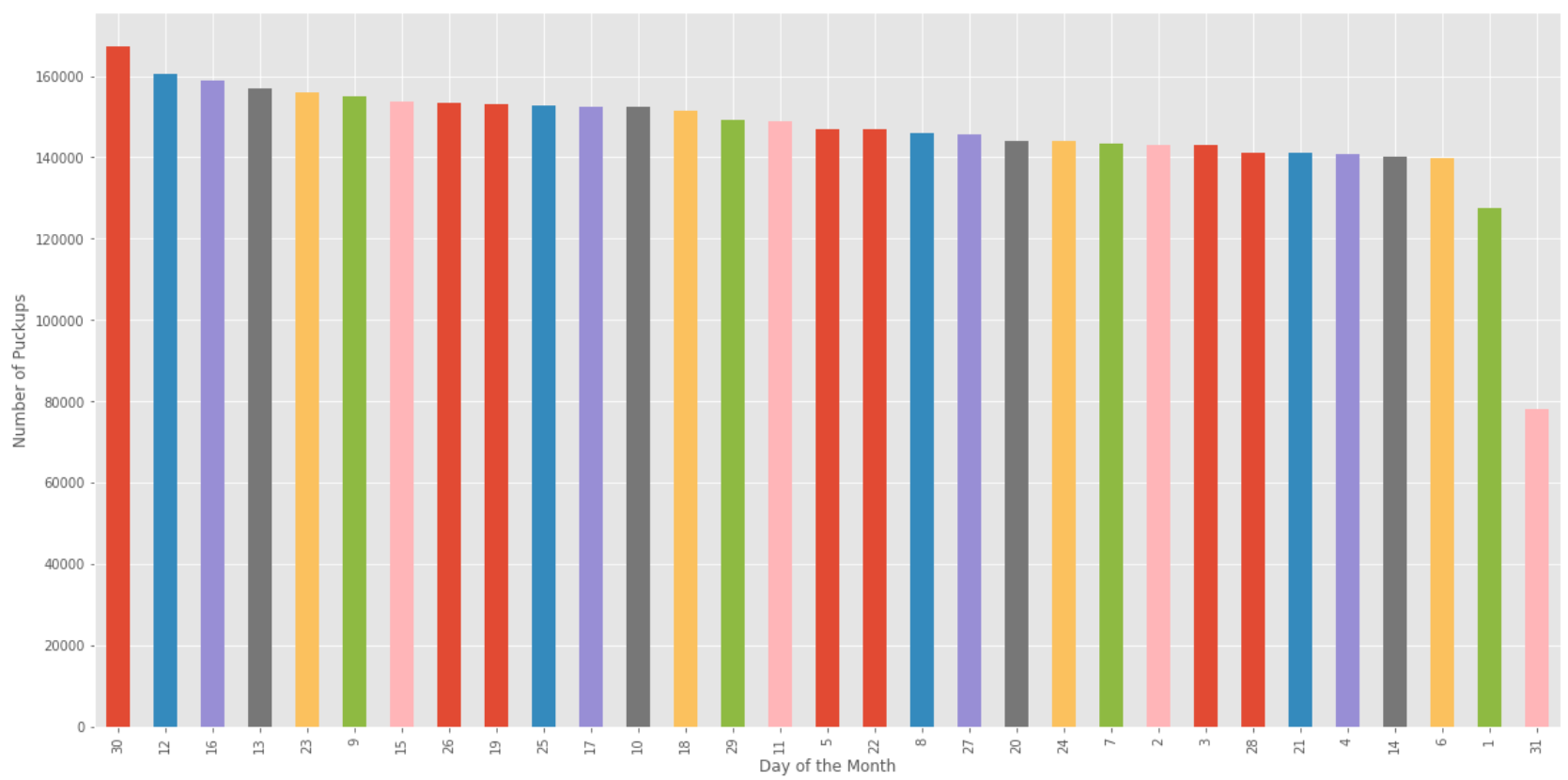
Plotted from smaller dataset

**Number of pickups by month**



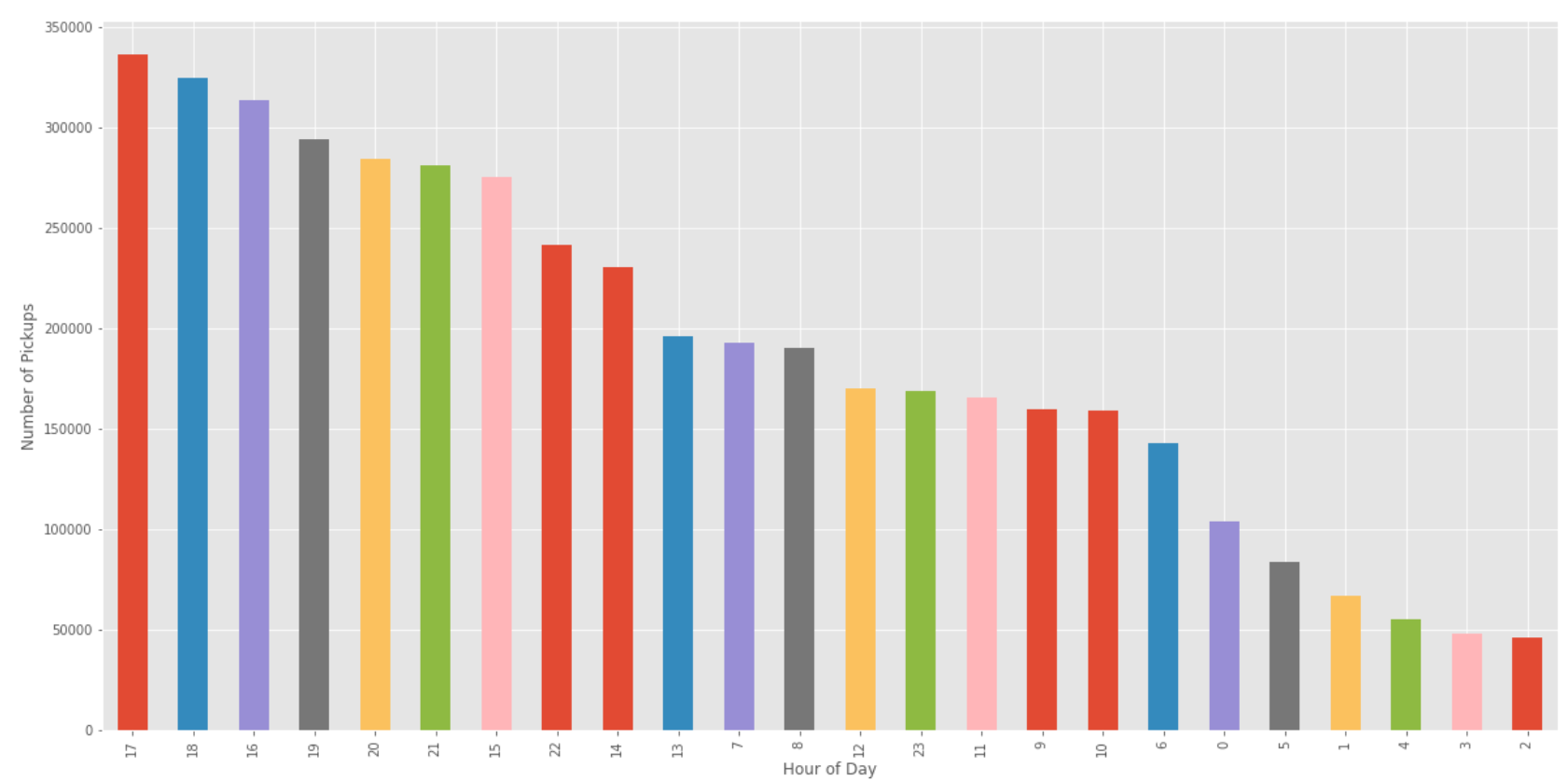
* This graph shows the number of total pickups for each month
* September has the most pickups while April has the least.
* The general trend is an increase in uber pickups every month. This maybe because of change in weather or an overall increase in awareness of the service over the months, or even a combination of both.

**Number of pickups by day**



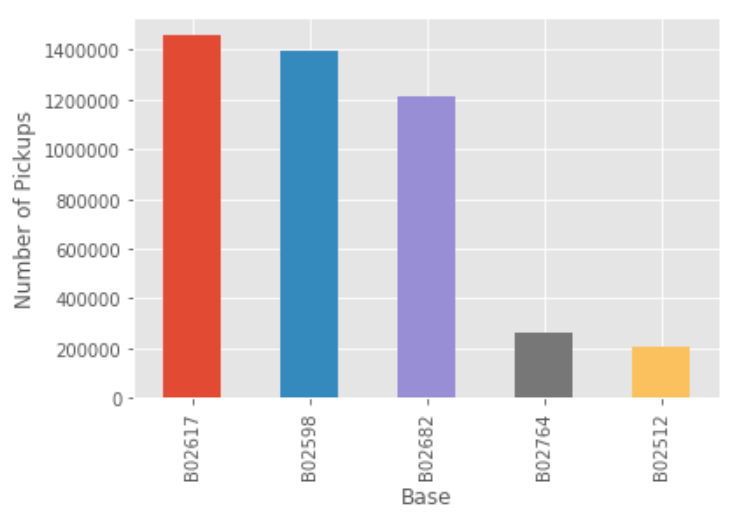
* This graph shows the number of pickups for each day in the months available in the dataset.
* The 30th seems to be the busiest day of every month, while the 31st is the least busy. However, it is important to note that only half the months have 31 days. So it would be sensible to exclude it and instead state that the 1st of each month is the lest busy.
* There is no general trend of increase or decrease in the number of pickups over the days of the months.

**Number of pickups by the hour**



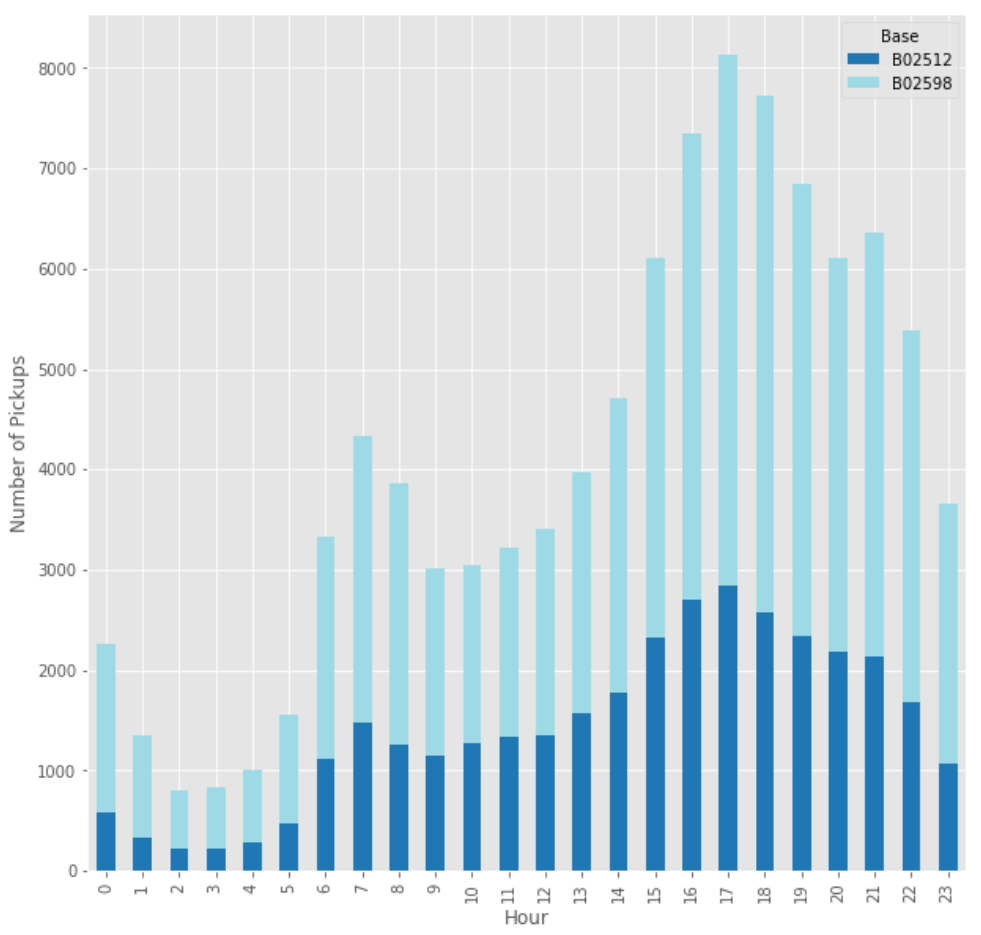
* 5 pm is the busiest time of day for uber pickups and that could be associated with that being the end of the workday. 2 am seems to be the least busy time.
* Overall it seems like the evening time around 5 pm is the busiest time of the day, and there is a decrease in pickups past midnight to 5 am.

**Number of pickups by the base**

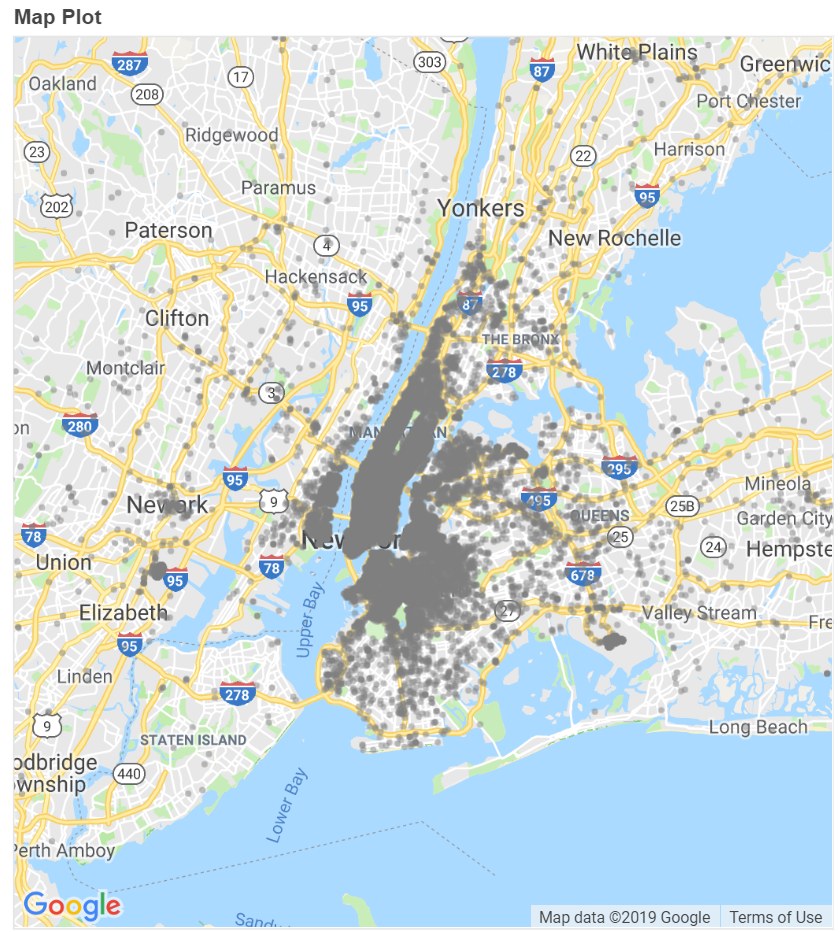


Base’s B02617, B02598, B02682 are where the majority of pickups are being dispatched from. With B02617 with the most pickups and B02512 with the least.

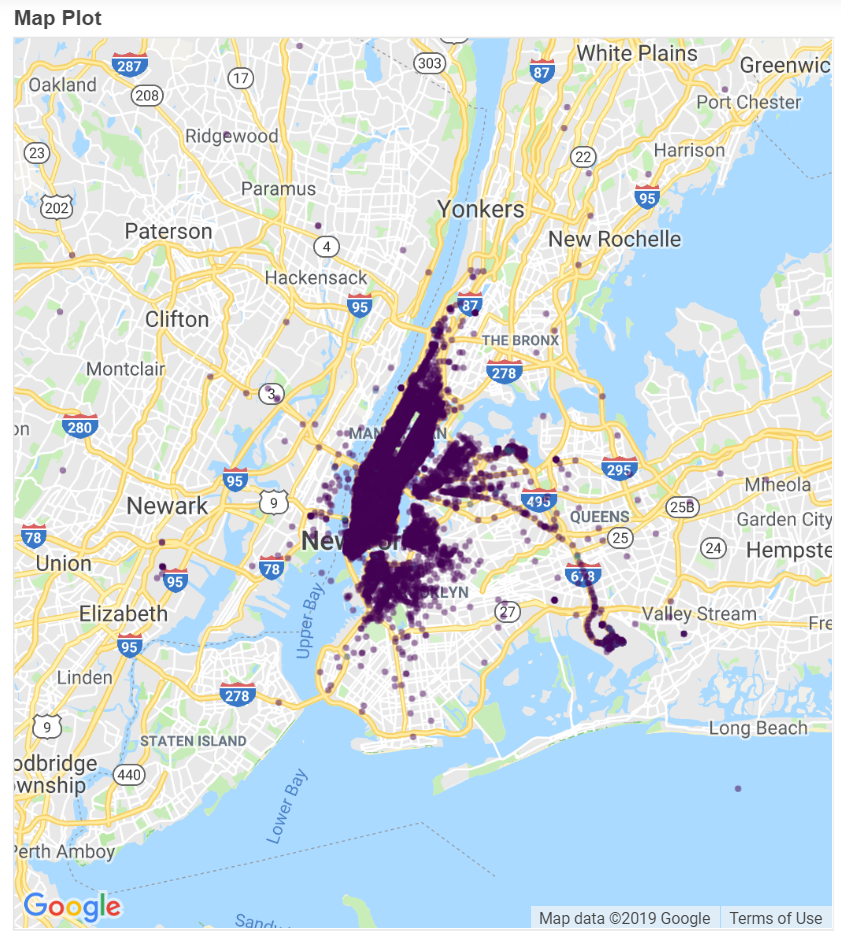
**Number of pickups by the hour of the day and colored by the base**



I created 2 density plots below that show the density of pickup loctions on the map of manhattan. I used google’s GMap API to do this.

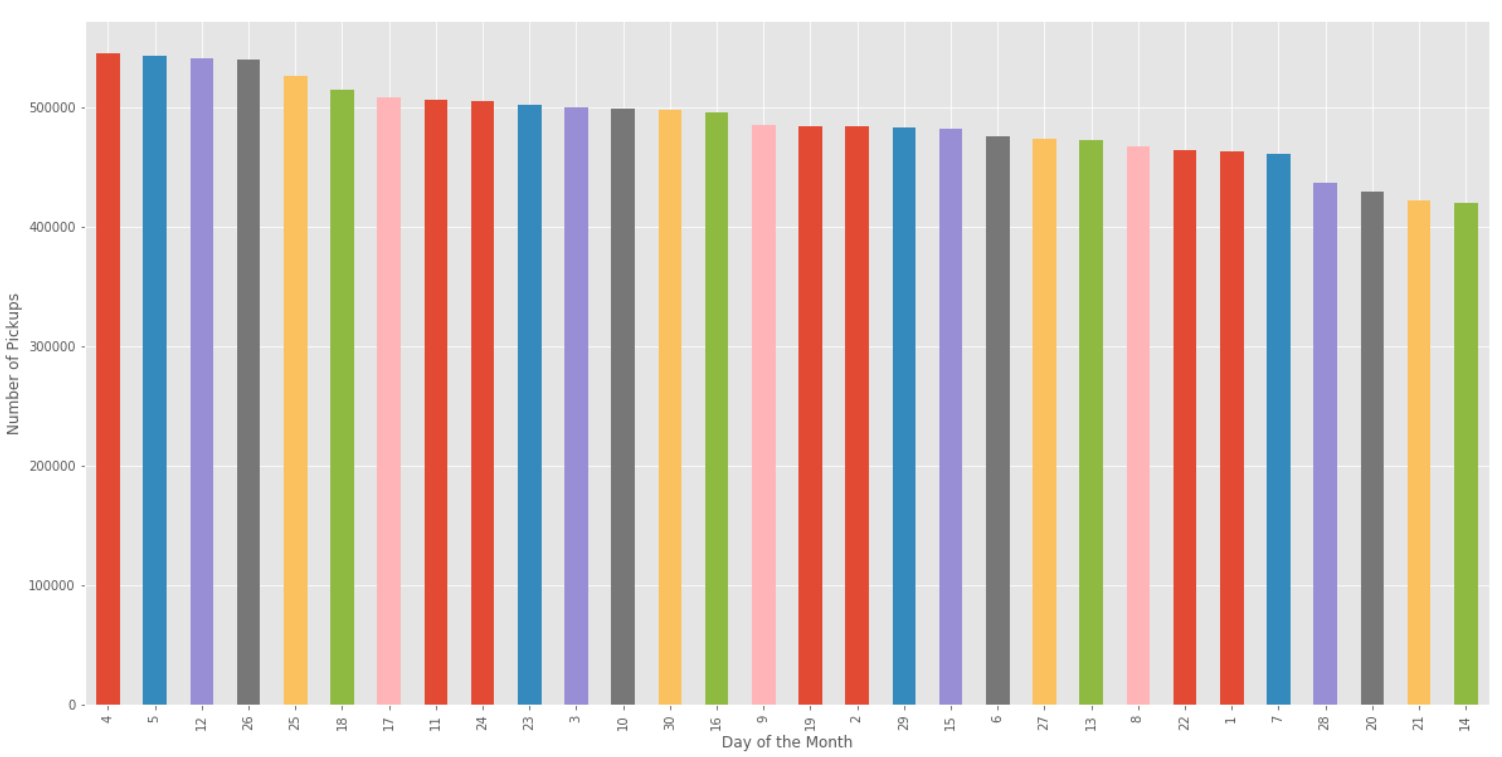


Yellow Cab

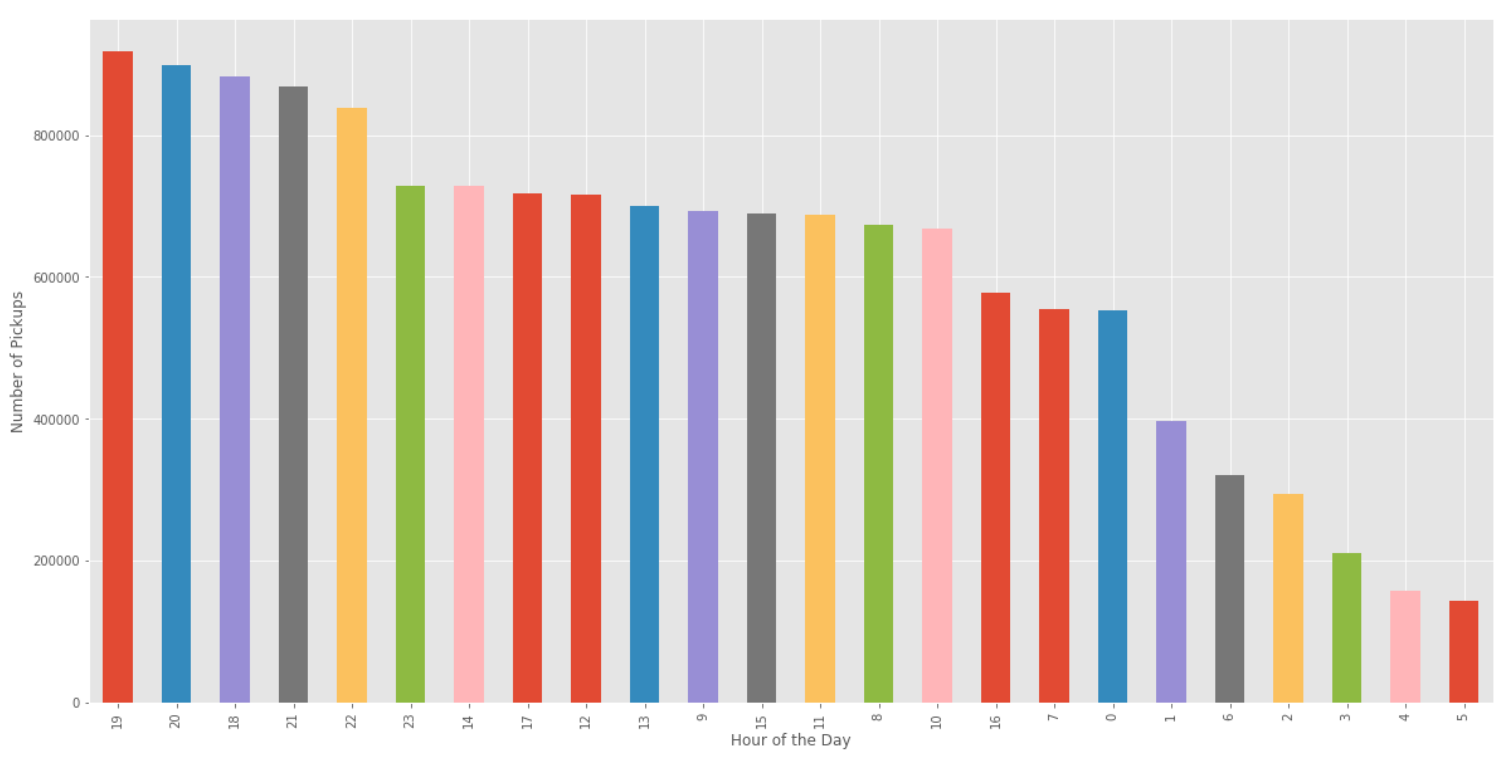


Comparing the two plots above we can see that there are more pickups outside the concentrated region of Manhattan for Uber

**Number of pickups by the day of the month**



**Number of pickups by the hour of the day**



Feature Engineering

For feature engineering, we focus our attention to the yellow cab data because that is what we will use to make the predictive model.

Since, during the machine learning stage will need to use the time values we will need to convert the times into different features that are not of DateTime type to be used by the model.

We add a column called ‘trip\_time’ this is a column which contains the time delta values from the ‘total\_time’ column but is an integer of the total number of minutes the trip took.

We then create 3 more features:

**‘Pickup\_day’** This is a column containing the number of the day of the week for each pickup.

**‘Pickup\_hour**’ This is a column containing the hour of the day of the week for each pickup.

**‘Pickup\_minute’** This is a column containing the minute of the day of the week for each pickup.

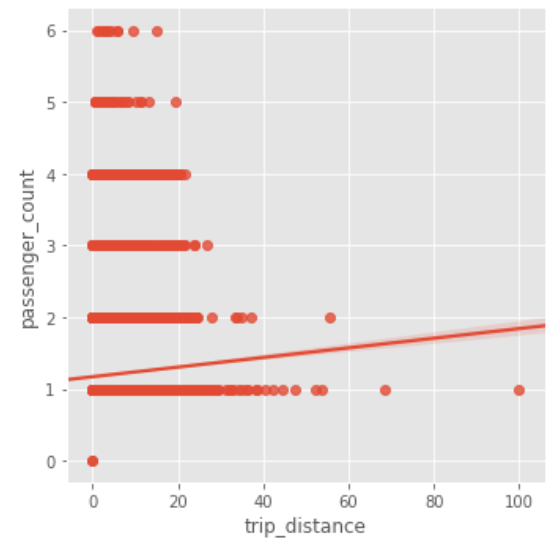
All the engineered columns contain values that are integers.

Inferential Statistics

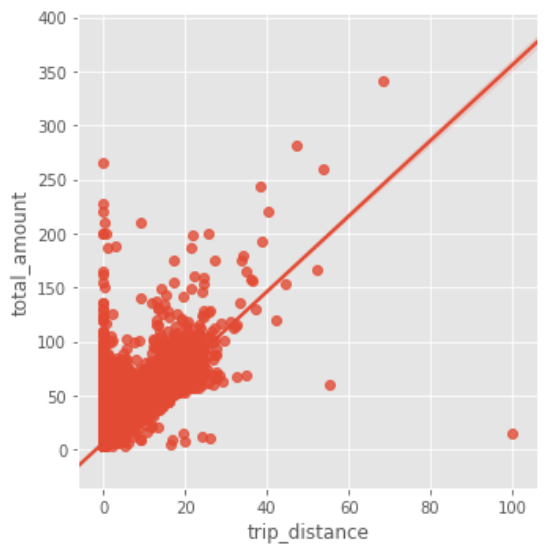
For the Uber dataset, we performed a t-test to see if the different bases had identical values for the hour of the day

We got a p-value of less than 0.05 so we reject the null hypithesis that the two bases have identical average values.

For the yellow cab data, we performed a correlation test between trip distance and passenger count. We got a correlation coefficient of 0.018 and a p-value of 0.0



We also performed a Pearson correlation test for trip distance and total amount here we got a correlation value of 0.9.



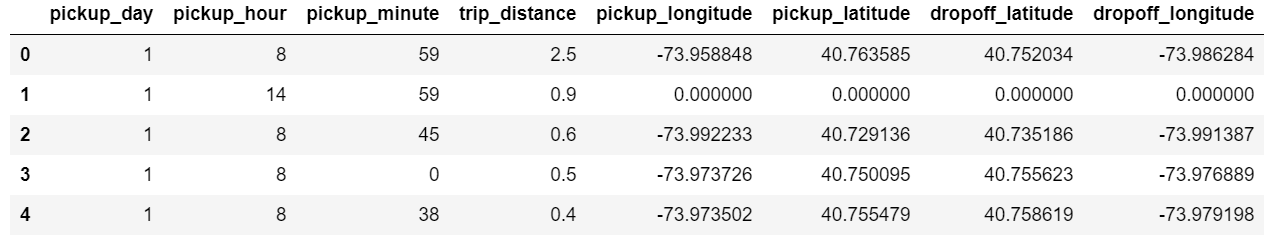
We performed a t-test for payment types of card and cash to see if they have average trip distances. We got a p-value of 0.0 so it is not statistically significant,

Machine Learning

For the machine learning, we only used the yellow cab data for April and not the uber data because the yellow cab has more useful features especially dropoff time. For the machine learning model, we are trying to predict ‘he duration of the trip which is a column labeled ‘trip\_time’

I performed a train test split on a smaller data frame called df\_small which contains only 100000 values instead of the full 14 million. I did this to improve computation time. I used 70% of this data for training and the remaining 30% for testing

I initially used KNN and Random forest classifier and ran into a lot of issues. I then realized I had to use regression because of the duration of the trip is a range of possible values. SO I then used a KNN regressor model.



The above data frame shows all the variables included in the prediction. These were selected by me by manually adding and removing certain columns from the full data frame to get the highest score.

With the best model, I got a score of 0.61 where the score is the r2 value. I got this score using 9 nearest neighbors as the parameter in the model.

I also performed a 3 fold cross validation but was unable t get a score higher than 0.61.

Future Work

EDA

I would like to make more plots on the new york map that compare differences between uber and yellow cab pickups. I would also like to include some comparisons using time as one of the variables when plotting on the map to gain some time-dependent insights.

I would also want to dive deeper into plotting more of the features for the yellow cab data.

Statistical Inferences

For this section, I would like to perform more statistical tests, especially for the yellow cab data. This will give me the insight to improve the predictive capabilities for the machine learning model.

Machine Learning

For this section, I want to spend some time on feature selection in the future instead of manually adding and dropping columns to improve the model.

I would also like to use a random forest regressor model to see if that gives me better results.