2020 East Coast Data Open

Team Number: 18
Team Members: Chenjian Wang, Yisu Zhou, Peter Wang, Ru Han

Executive Summary:

For bike-sharing companies like Citi Bike in NYC and Bay Area Bike Share (BABS) in San Francisco, rebalancing is an essential cost of operation. Because of the limited space available for docking stations, companies must use trucks to shuffle around bikes to ensure that customers looking for a bike will find one, and customers returning a bike will find an empty slot. When there is a large variation in weather, conventional ridership practices change, and different allocation methods must be employed. Our investigation looked at the changes in demand for shared bikes, and our topic question was: **what attributes can be used to predict demand, and how does it relate to the surrounding city infrastructure?** With a thorough understanding of the factors that affect bike demand, companies like Citi Bike and BABS will be able to more accurately use real time data to anticipate demand spikes or lulls.

Our main result is a model that predicts aggregate demand based on environmental and temporal data. This model was created based on our discovery that time of day was the most significant factor determining bike demand, unsurprisingly. There are two spikes in demand over the course of one day; one spike at around 8am, the other around 5pm -- regular working hours. Our exploratory data analysis also revealed that precipitation was actually less correlated with demand than temperature, suggesting that companies should pay more attention to weather events like squalls and polar vortices rather than sudden storms to anticipate demand.

Technical Investigation

Initial Exploration: Exploratory Data Analysis (EDA) and Data Visualization

Because the original NYC Bikeshare dataset had 27 million lines, we needed to take an unbiased sample as a representative dataset for our analyses. Since our exploration mainly focused on the hourly and daily basis of the general trend, we randomly sampled from the first two consecutive months of the data to form our dataset.

We first conducted EDA on the NYC Bikeshare dataset by aggregating the total record count and average trip duration by hour. We found several outliers for the hourly trip duration (a bike was “borrowed” for 4 million seconds), which we removed. We calculated that there were 110 trips per hour on average.

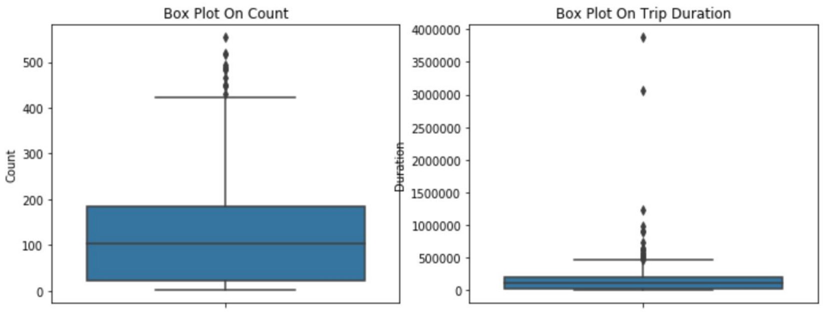


Figure 1: before removing outliers

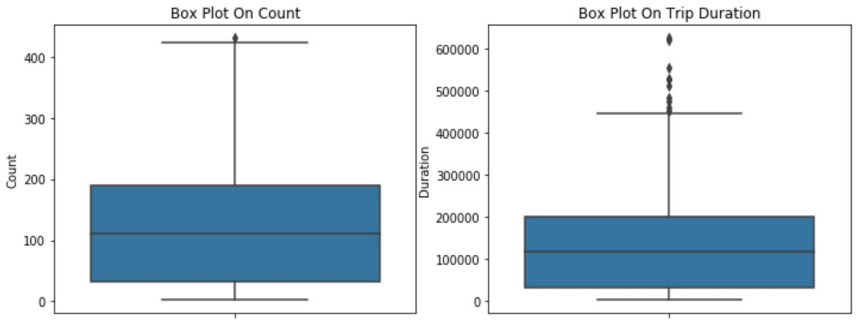


Figure 2: after removing outliers

To get a finer view into the change in demand over the course of a day, we aggregated the total record count and average trip duration on each day. The data revealed a bimodal distribution; the two peaks were at around 8am in the morning and 5pm in the afternoon, the traditional workday hours. We also found that workday demand was slightly higher than non-workday (weekends and holidays) demand. Furthermore, the non-workday trip duration distribution had a higher variance than the workday trip duration distribution. This implies that people take longer rides when they aren't working. When we plotted the demand over the course of a day for the seven days of the week, a significant difference in the hourly demand distribution between working and non-working days emerged. We see the two-peaked distribution on workdays that correspond to workers commuting to and from work, and we see a bell-shaped curve on non-work days that reflect the fact that people tend to use bikes throughout the day according to a roughly normal distribution.

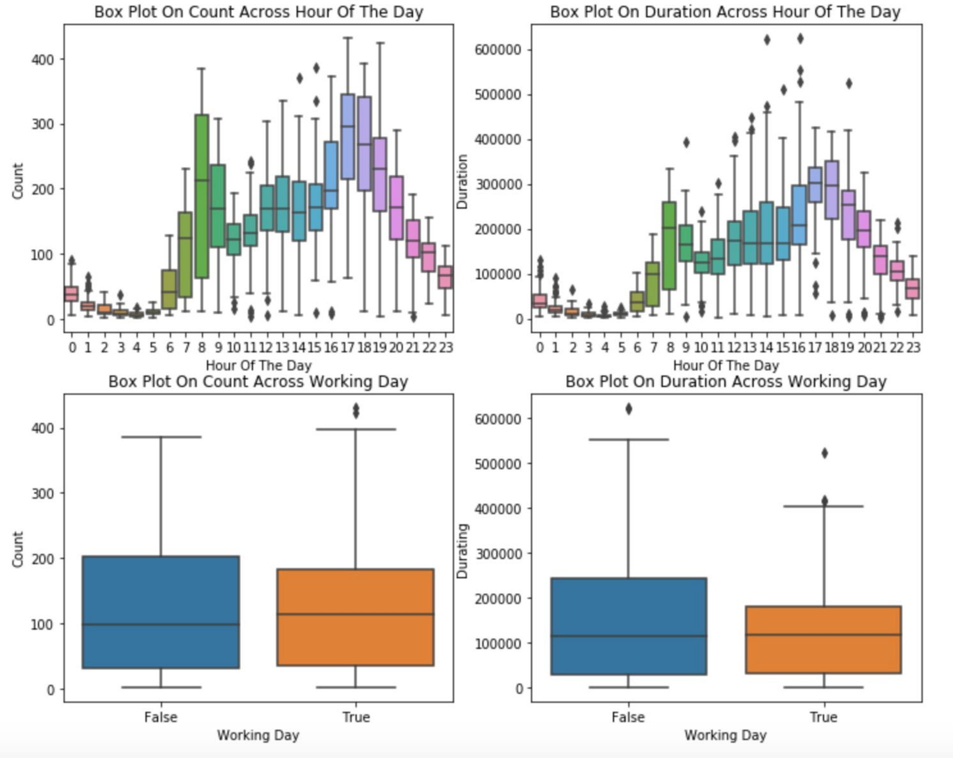


Figure 3: Count and Trip Duration by hour and working day(T/F)

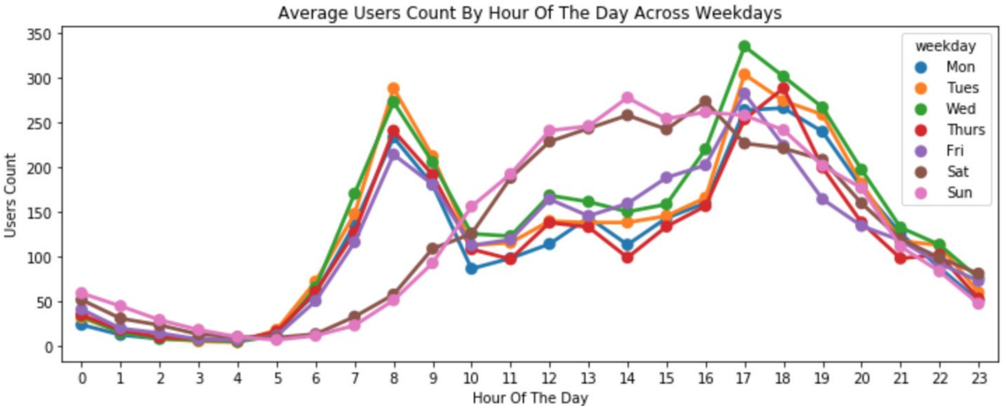


Figure 4: Avg user count by hour across weekdays

We then further tuned our analyses by splitting the data into the subscriber and non-subscriber groups (those with a Citi Bike subscription and those who bought 1-day/3-day passes) and looking at their hourly and daily trip durations. We noticed that the two-peaked trend in hourly demand only appeared in the subscriber's group. This strongly implies that most workers who use Citi Bike to commute to and from work are subscribers, and the non-subscribers are not beholden to the 8-5 work schedule. The non-subscriber group also tended to have longer trip durations across all hours compared to subscribers. The average trip time for a subscriber between the hours of 7-9am and 4-6pm was around 10 minutes. When a non-subscriber uses a

Citi Bike, which offers 30 minute rides before requiring the non-subscriber to pay extra, we see an average trip duration of 17 minutes no matter when the bike was checked out.

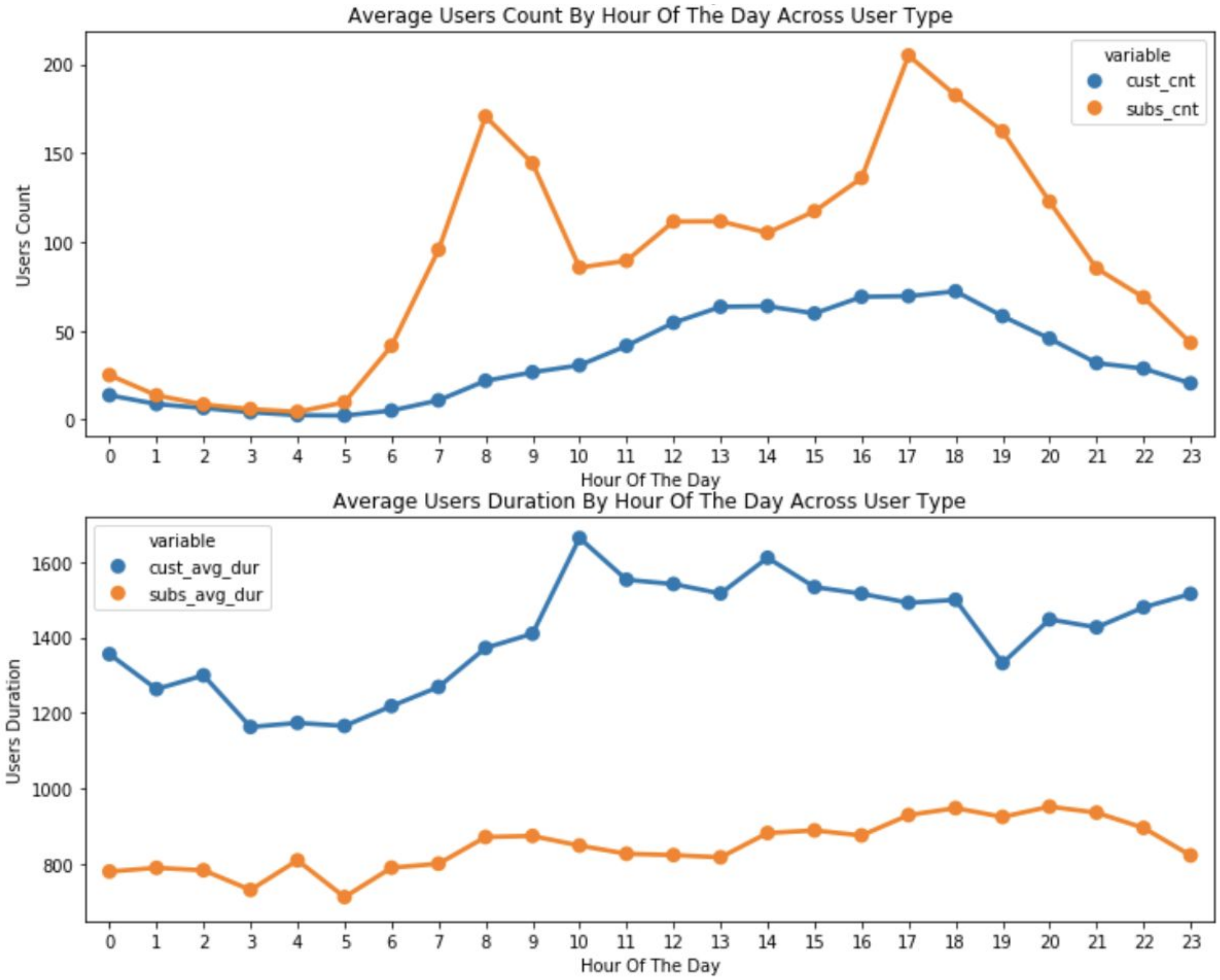


Figure 5: Avg user count by hour over the course of a day

Finally, we pulled in the weather data and merged it with the NYC Bike share dataset to explore the correlations between weather and demand. As shown by the correlation plot (figure 6), Dry Temperature and total trip counts has the highest correlation of 0.41. Unintuitively, precipitation has a very low correlation with demand.

We then plotted two regression plots to more deeply explore the relationship between dry temperature and trip count, and the relationship between dry temperature and trip duration. There seems to be a nonlinear relationship between the two variables.

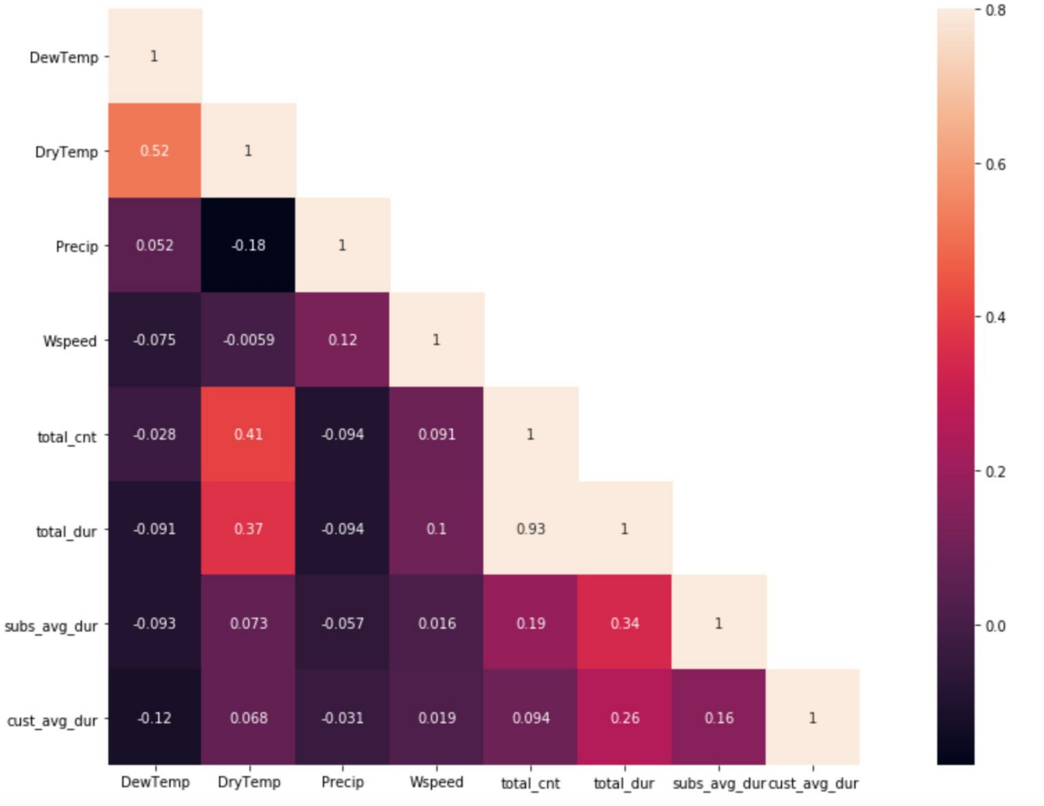


Figure 6: Correlation Plot of Weather Attributes vs. Demand

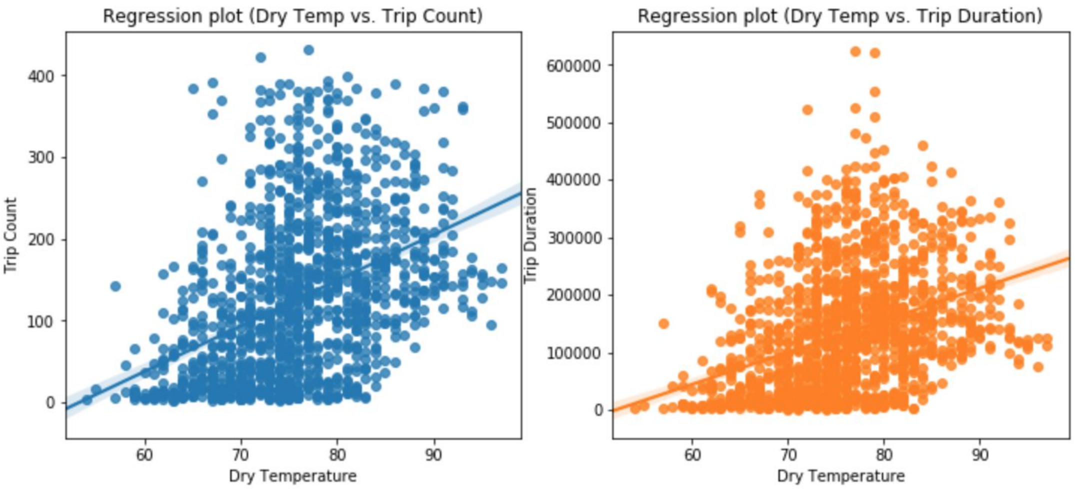
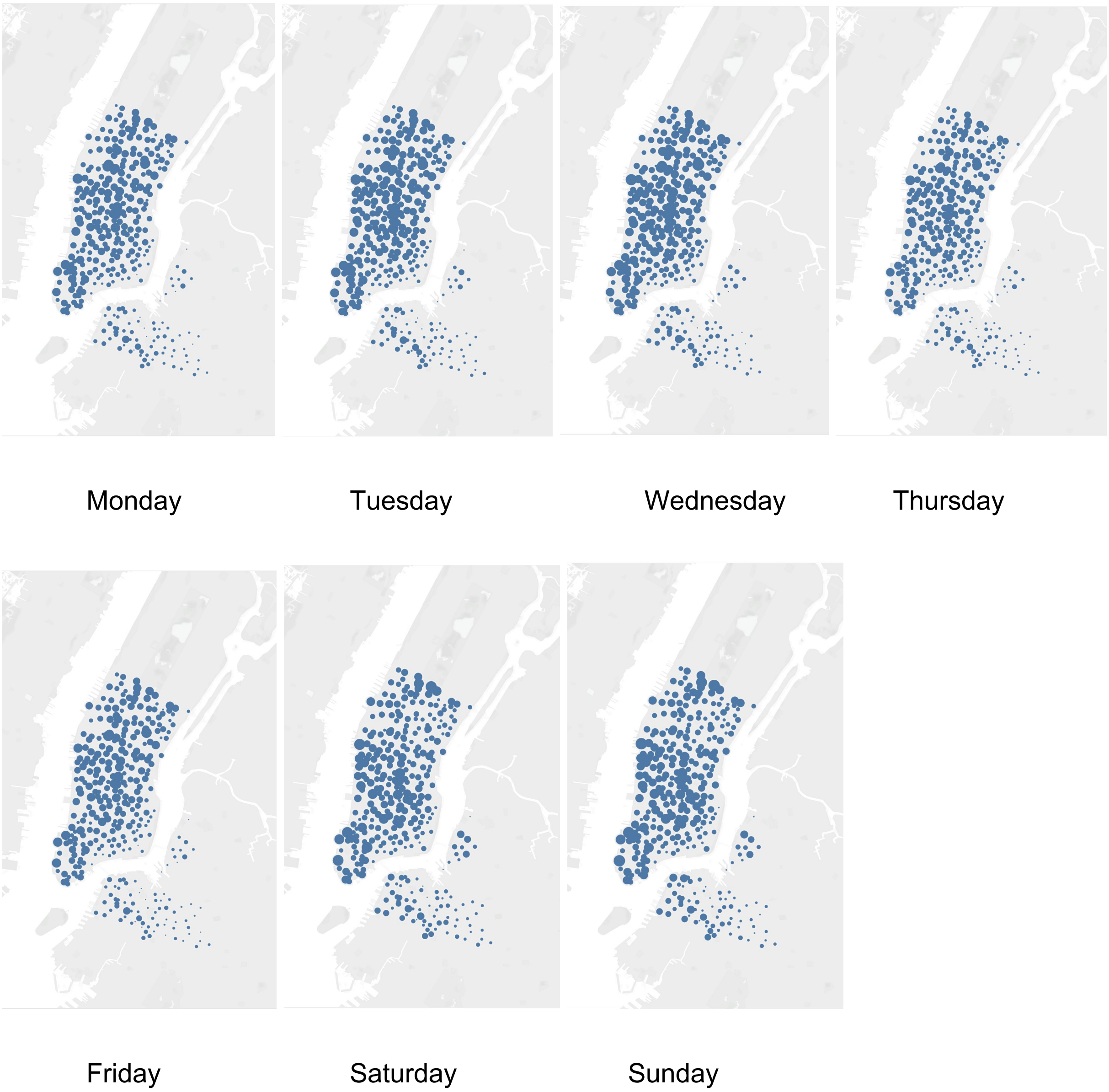


Figure 7: Regression Plot of Dry Temperature vs. Trip Count / Trip Duration

We also merged the data with the geographic dataset (latitude and longitude) to visualize the demand by docking station. We noticed that on that Manhattan island, there were several docking stations that had huge demand (measured by number of trips started) compared to other areas, but these spots changed as the week progressed. This provides indication for our modeling part that geographic information could play a huge role in predicting the demand. This data also greatly enhances our modeling results because with an understanding of aggregate demand, we can employ a simple procedure of allocating bikes to docking stations in proportion to their historical demand.

Figure 8: NYC Bike Share record counts Monday - Sunday by location



Monday

Tuesday

Wednesday

Thursday

Friday

Saturday

Sunday

Feature Engineering and Modeling

Model assumptions

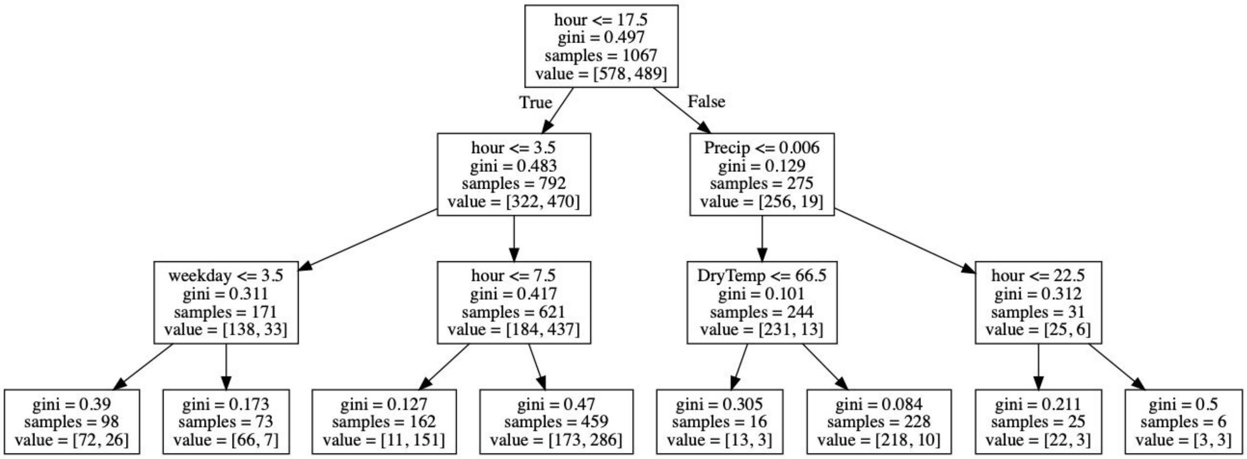
- There is nonlinear relationship between features and number of trips per hour (EDA demonstrate this)
- Markov property: we assume all the information is contained within the current state; in other words, past information doesn't matter in predicting future outcome.

Therefore, we chose lag(1) features and utilized random forest for modelling, since random forest possesses automatic feature engineering.

We start our modeling part by creating a response variable out of our merged and aggregated data. We take the difference of total counts between each two consecutive hours, and label it as either increase or decrease. If they are the same (only 1% of all the cases), we put it in the decrease category. We also created an additional feature called holiday to indicate whether that specific day is a holiday.

Then we start building our classification model to see whether we can predict the change of demand in the next hour using the features for the current hour in the original dataset (hour, day, month, weekday, holiday) and the features in the weather dataset (Wind speed, Precipitation and Temperature).

We split the data into train set (80%) and test set (20%) and fit a decision tree as a baseline model. The decision tree achieved 86.13% accuracy on train set and 86.11% accuracy on test set. By looking at the splitting criterion of the model, we can see the result matches with our previous hypothesis pretty well. The number one splitting criteria is the current hour. One thing we noticed that previously we found that precipitation has a very low correlation with demand, but it is actually a very important feature in the model. This is due to the fact that decision trees can detect interactions between features and use those interactions to predict the response variable.



And now we fit a Random Forest model which tends to be more robust and less overfitting to see if that could increase our test set accuracy. After fitting the model using the default parameters, we did see the test accuracy increased to 89.47%. By looking at the ROC curve, the model is very robust in terms of True Positive and and False Positive.

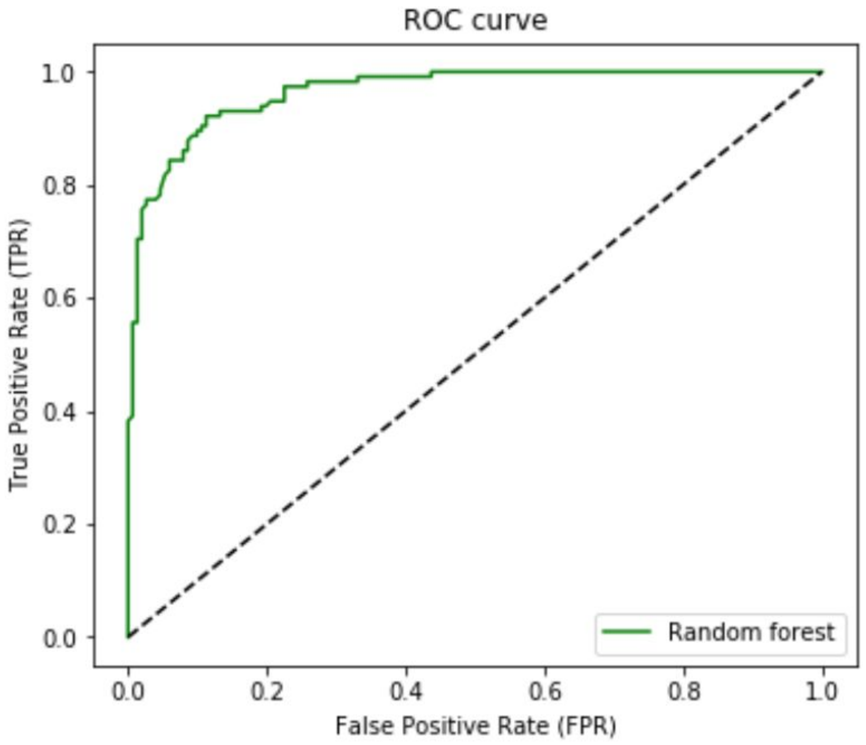
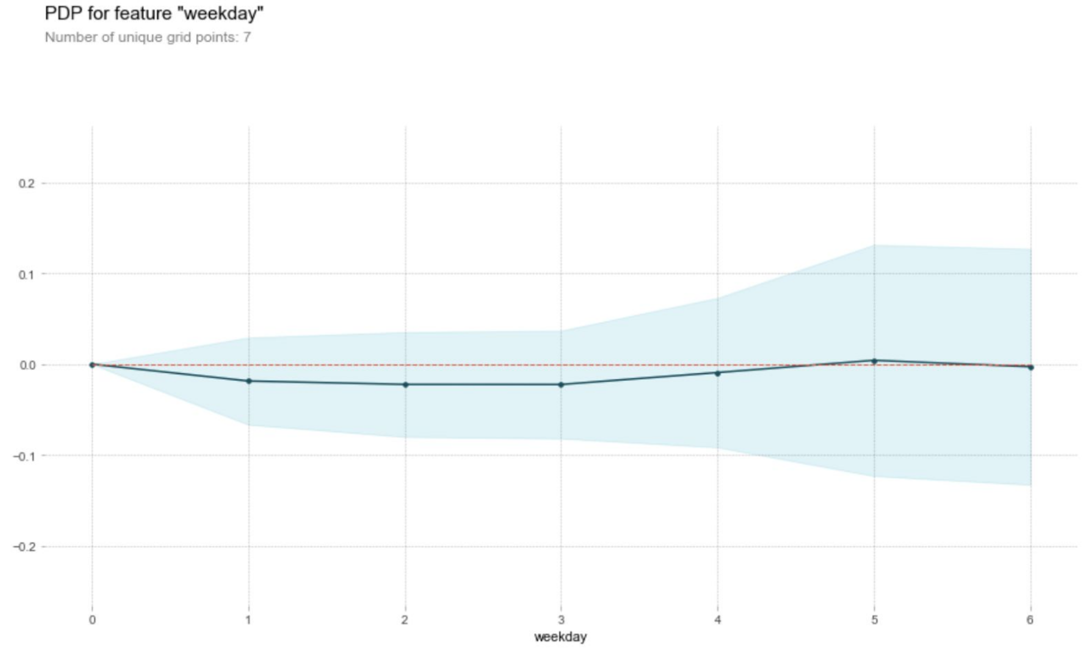
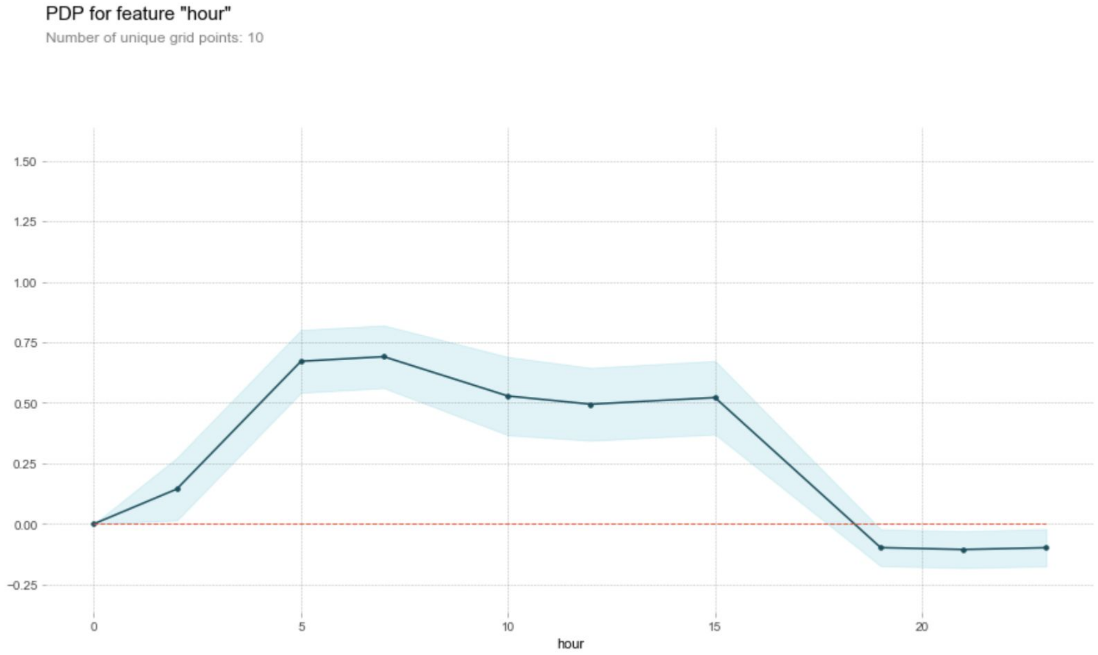


Figure 9: ROC curve

Model Interpretation:

By plotting out the partial dependence plot for the Random Forest model, we can see that the only two features that do not have a straight line partial dependence plot are weekday and hour features. The weekday feature has a minor effect on the next hour's demand compared to the hour feature. We can see that Tuesday, Wednesday and Thursday contribute to predicting a decrease compared to the other days. And all the hours before 7pm contribute to predicting an increase compared to the last several hours of the day.





Conclusion:

Our exploration started with some exploratory data analysis and data visualization, from which we learned the relationship between aggregate CitiBike demand and different environmental and temporal factors like work schedules, temperature, and precipitation.

This led to a more detailed analysis of CitiBike usage over the course of 24 hours on workdays and non-workdays, and analyzed the difference between the behavior of subscribers and non-subscribers. With our insights, we were able to create a random forest model to predict aggregate CitiBike demand in the next hour with the same environmental and temporal factors mentioned above as inputs. The classification model has an 89.5% accuracy.

Such a classification model could be used by CitiBike to create an optimal allocation policy that would minimize the times a potential customer is unable to take out a bike.

Future Work

We've conducted a thorough investigation of the environmental and temporal factors' effect on aggregate CitiBike demand, which are all used as inputs into the model. However, we must also investigate the effect of substitute transportation methods like ride-sharing apps, taxis, and the subway on aggregate demand. We conducted some preliminary research (see appendix A) on the relationship between subway use and CitiBike demand, but unfortunately were unable to incorporate the findings into the model.

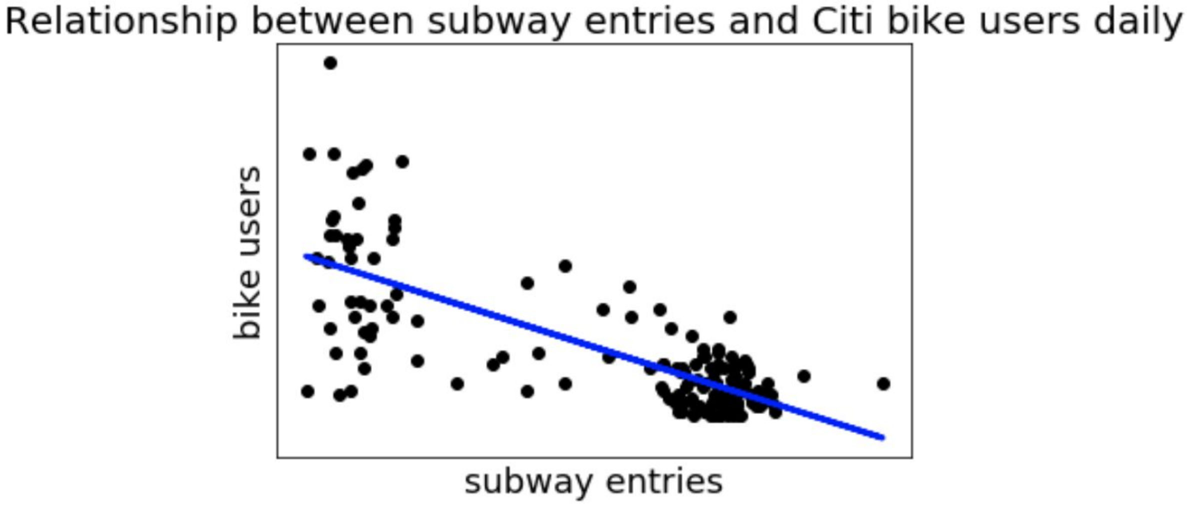
Our model also treats all CitiBike stations as an aggregate, which provides a good estimate of overall demand for bikes, but does not give a fine-grain allocation procedure that would tell CitiBike where and when bikes are needed in real time. With NTA data, we could perform the same analysis on a smaller scale, and draw more interesting conclusions about the different workday schedules of people living in different parts of Manhattan, which would again provide valuable insight into the overall ebb and flow of bike demand over a large time scale.

As with any model, we can also increase our sample size to create a more accurate depiction of reality - our current model includes just three months of sampled data because of computing and time constraints.

Appendix A

Relationship between subway entries and bike share usage

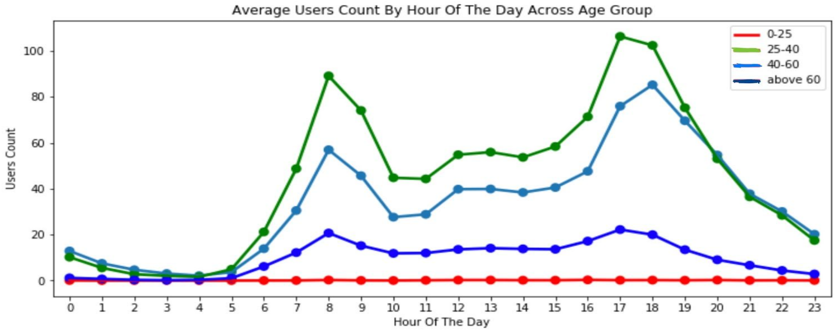
Since the bikeshare and subway datasets for the whole NYC are too big, we chose one specific subway station as an example: “34 ST-PENN STA” . We found the nearest Citi bike station is ‘station 3254’. After analyzing the total number of Citi bike users starting from ‘station 3254’ and total subway entries of ‘34 ST-PENN STA’ each day in year 2017, we found their correlation is -0.753. Following is the figure of their relationship.



It demonstrates that within a certain distance, when the number of subway entries increases/decreases, the number of Citi bike users will decrease/increase.

Appendix B

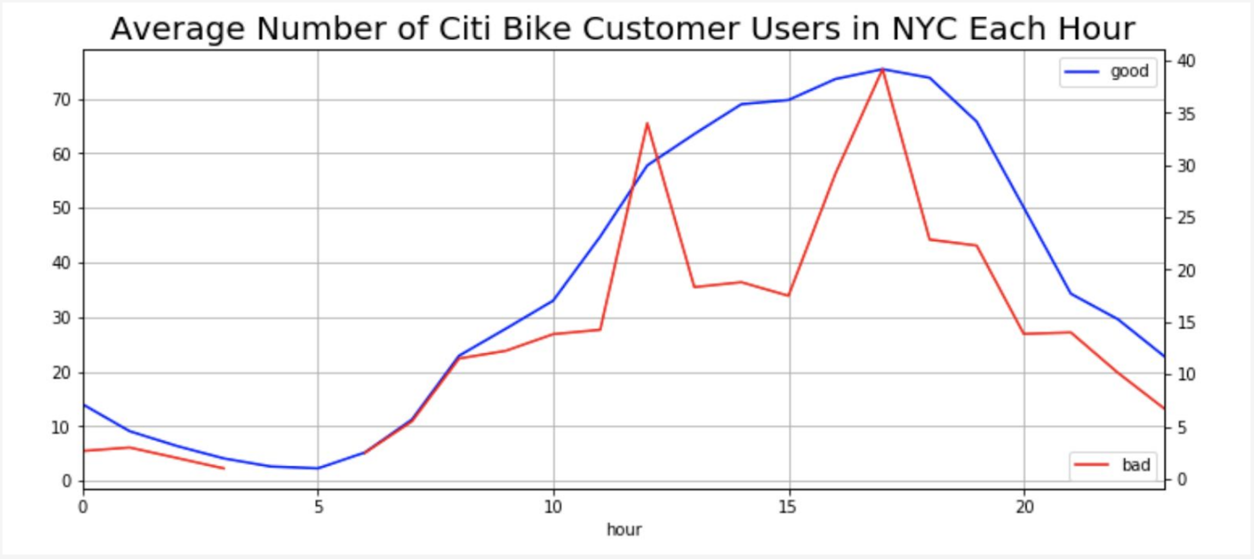
What about the demographic information? As we can see in the plot that the 0-25 and 25-40 age group has a very similar pattern compared to our previous findings. And the age group 40-60 has significantly lower usage. The above 60 age group does not ride bikes.

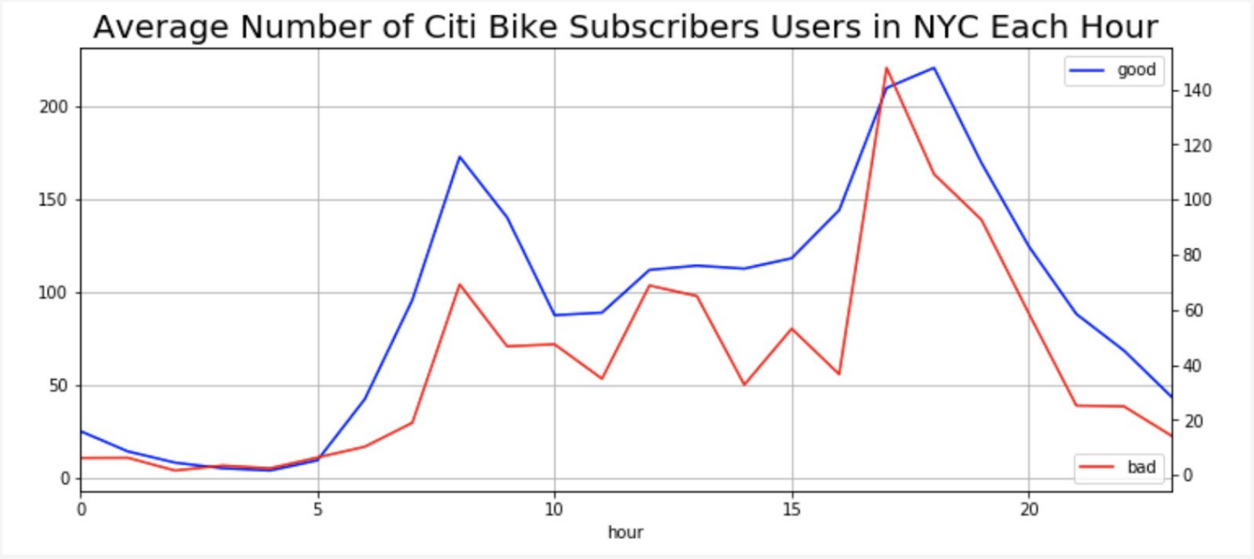


Appendix C

Influences of weather on Citi bike usage

We define the weather as ‘good’ if the dry temperature is larger than >50 and the hourly precipitation is lower than 0.01, and the weather as ‘bad’ if the dry temperature is lower than >50 and the hourly precipitation is larger than 0.01. Following are figures of average bike users(customers/subscribers) over hours under ‘good’/’bad’ weather.


Average Number of Citi Bike Customer Users in NYC Each Hour
good
bad
hour


Average Number of Citi Bike Subscribers Users in NYC Each Hour
good
bad
hour