MWSUG 2016 – Paper AA22

Analyzing the effect of Weather on Uber Ridership

Snigdha Gutha, Oklahoma State University
Anusha Mamillapalli, Oklahoma State University

ABSTRACT

Uber has changed the face of taxi ridership, making it more convenient and comfortable for riders. But, there are times when customers are left unsatisfied because of shortage of vehicles which ultimately led to Uber adopting surge pricing. It's a very difficult task to forecast number of riders at different locations in a city at different points in time. This gets more complicated with changes in weather. In this paper we attempt to estimate the number of trips per borough on a daily basis in New York City. We add an exogenous factor, weather to this analysis to see how it impacts the changes in number of trips. We fetched six month worth data (approximately 1 million records) of Uber rides in New York City ranging from January 2015 to June 2015 from GitHub. We also gathered weather data (from Weather Underground) of New York City Borough wise for the same period of six months from Jan 2015 to June 2015. In this poster, we attempted to analyze Uber data and weather data together to estimate the change in the number of trips per borough due to changing weather conditions. We used SAS® Forecast Studio and built a model to predict the number of trips per day for the one week ahead forecast for each borough of the New York City.

INTRODUCTION

The purpose of this project is to analyze the impact of weather on Uber Ridership. We chose this topic in particular due to the raising concern among customers about Uber adopting surge pricing to deal with the growing demand.

We hypothesized that change in weather will affect the number of Uber rides and through our analysis, we found out that the number of rides increased on an average basis in case of any weather event than a normal day.

The dataset compiled for this project serves as a foundation for additional research. Analyzing at least a year worth of data will bring further insights. Demand can be more accurately predicted if the actual number of rides requested information is available along with the number of live rides.

PROJECT CONSIDERATIONS

IDENTIFICATION OF POTENTIAL BENEFACTORS

This study will particularly benefit Uber and its customers. In general any taxi service can benefit through this study. Analyzing the demand at borough level will also aid in the optimal utilization of existing resources. It can assist Uber in the design of optimal incentive programs which motivates Uber drivers to move to a different borough in the case of increased demand, which in turn can reduce surge pricing to an extent.

CONSTRAINTS AND LIMITATIONS

One of the major concerns is the change in number of rides is mainly attributed to the changing weather here and the results can be stated more confidently provided we had at least a year worth of data. The effect can be more precisely estimated if the information about actual number of rides requested is available along with the number of live rides. Also, the number of rides for certain days are missing and we imputed those days with the average of immediate previous non missing day's information and immediate next non missing day's rides.

DATA COLLECTION, CLEANING AND CONSOLIDATION

Uber rides data is collected from GitHub. The data ranged from January 2015 to June 2015 and covers ridership information for all the five boroughs of New York City. We gathered Weather data from Weather

Underground for the five boroughs of New York for the same tenor. Uber rides data is at date time level, and we couldn't find hourly weather data, so we aggregated Uber rides to date level. We then merged Uber dataset with Taxi_lookup_zone dataset to bring the Uber ridership data to borough level. We then merged the Merged1 dataset with Weather dataset to obtain the final dataset.

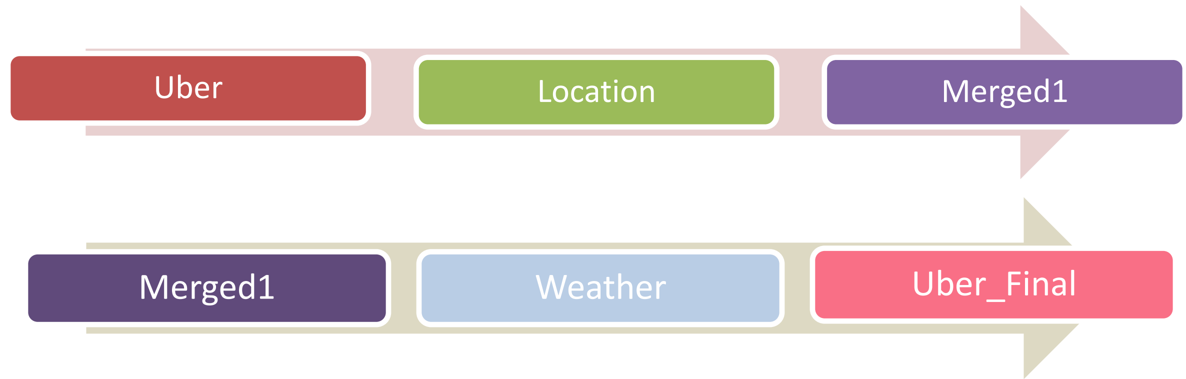

Figure 1. Data Preparation Flow

DATA DICTIONARY

Dataset	Variables
Uber_PickUp_data	Dispatching_Base_Num, Date_of_Pickup,Time, Affiliated_Base_Num, LocationID
Taxi_Lookup_Zone	Location_ID, Borough, Zone
Weather	Date ,Temperature, Humidity, Sea_level_pressure, Precipitation (Max, Mean and Min), Cloud_Cover, Wind, Event

Table 1. Data Dictionary

DATA CLEANING AND TRANSFORMATION

Uber_Pickup_Data has ridership information missing for certain days. We used average of immediate previous interval non missing value and immediate next interval non missing values to fill in the missing values. We used SAS Enterprise Guide's create time series data node to produce the final time series data on a daily level. To deal with Skewness and high kurtosis and bring the data to normality, we transformed weather metrics and the target variable number of rides using SAS Enterprise Miner's transform node. We applied Max Normal transformation which selects the most appropriate transformation for a given distribution to all the variables in the initial stage and later replaced the Max Normal transformation with the selected transformation respectively for all the variables.

Figure 2. Data Transformation

EXPLORATORY ANALYSIS

The initial exploratory analysis revealed following insights.

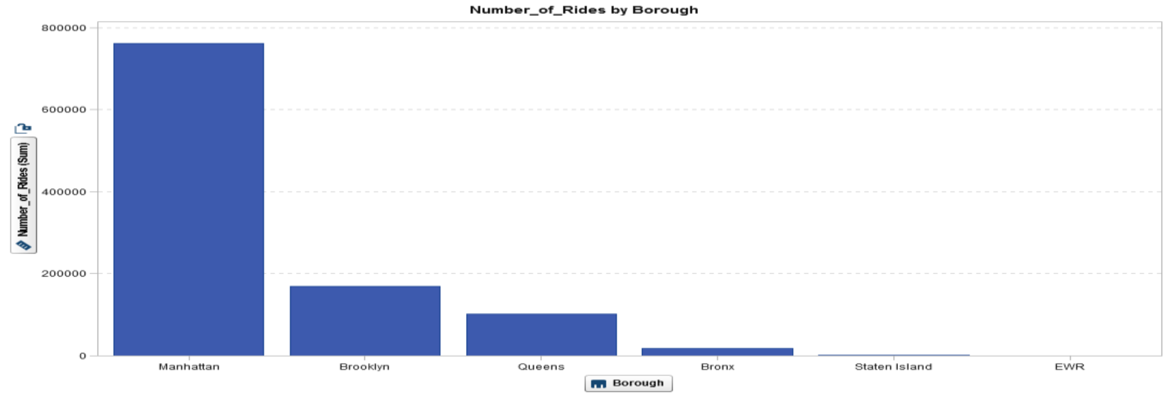

Figure 3. Number of Rides by Borough

Manhattan has the largest user base being the most populous county and financial capital of world followed by Brooklyn and Queens.

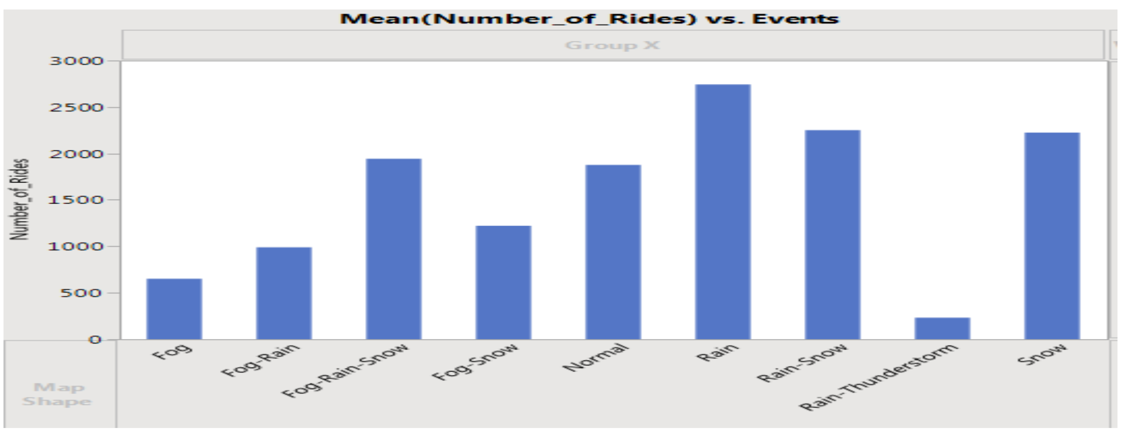

Figure 4. Average Number of Rides by Event

Overall, on an average a rainy day has reported more number of rides followed by snow than a normal day.

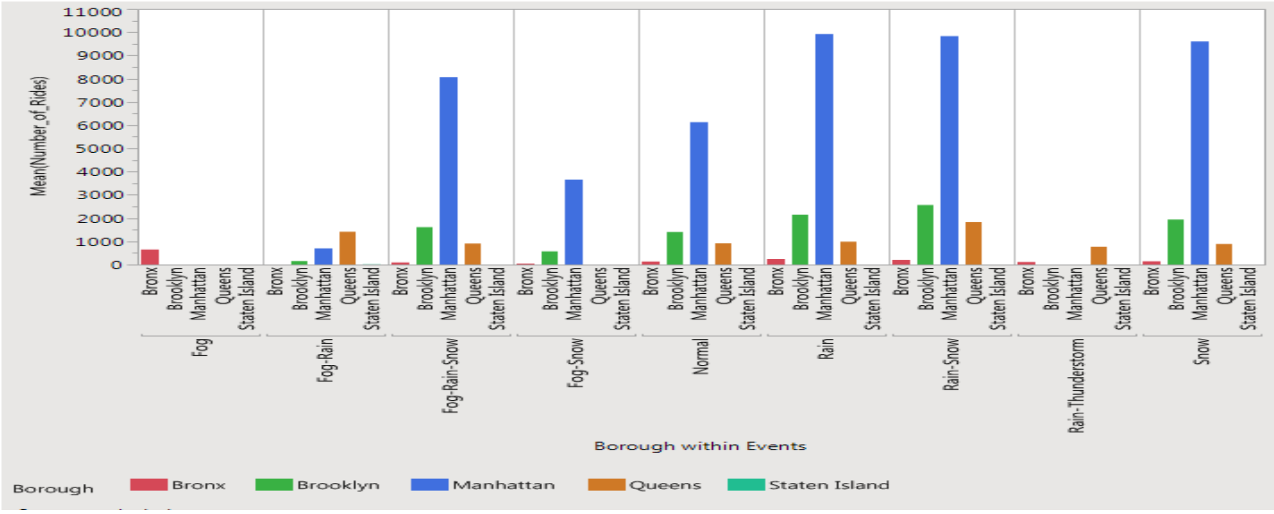

Figure 5. Average Number of Rides by Borough and Event

Fog seems to impact the number of rides in Bronx more than any other borough. Fog-Rain and Rain-Thunderstorms are influencing the number of rides in Queens the most. Rain and Snow is increasing the number of rides in Manhattan more than any other event.

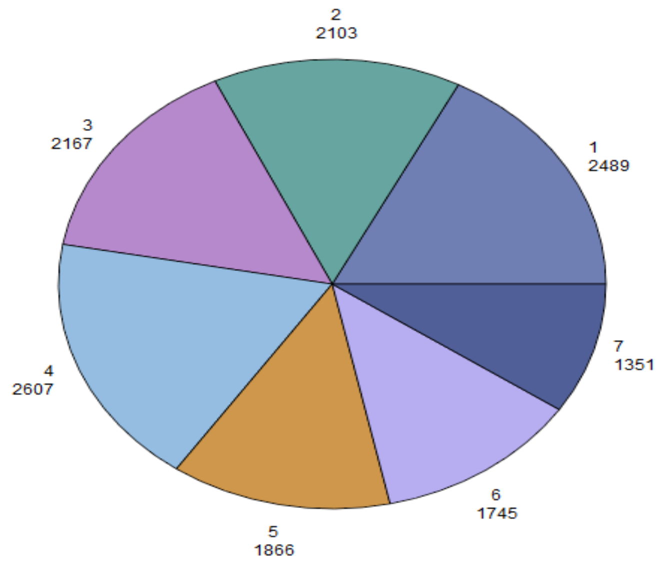

Figure 6. Mean Number of Rides by Day of Week

Considering all five boroughs, Wednesdays are the busiest day representing mid-week followed by Sunday (weekend). Saturdays have the least turnout.

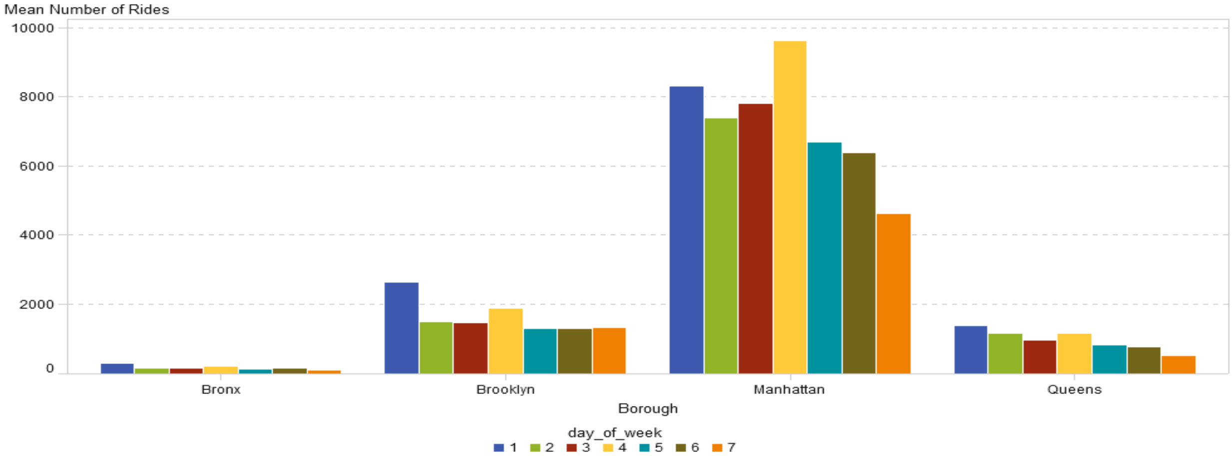

Figure 7. Mean Number of Rides by Borough by Day of Week

Wednesdays are the busiest day for Manhattan followed by Sunday. Whereas Sundays turn out to be the busiest day in other boroughs.

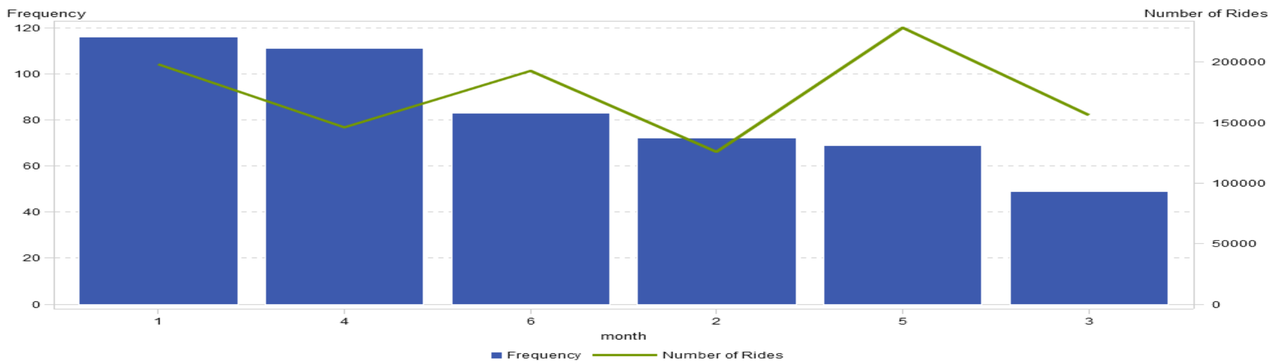

Figure 8. Number of Rides by Month

The month of May shows the peak number of rides when compared with the other months. April has the least number of rides.

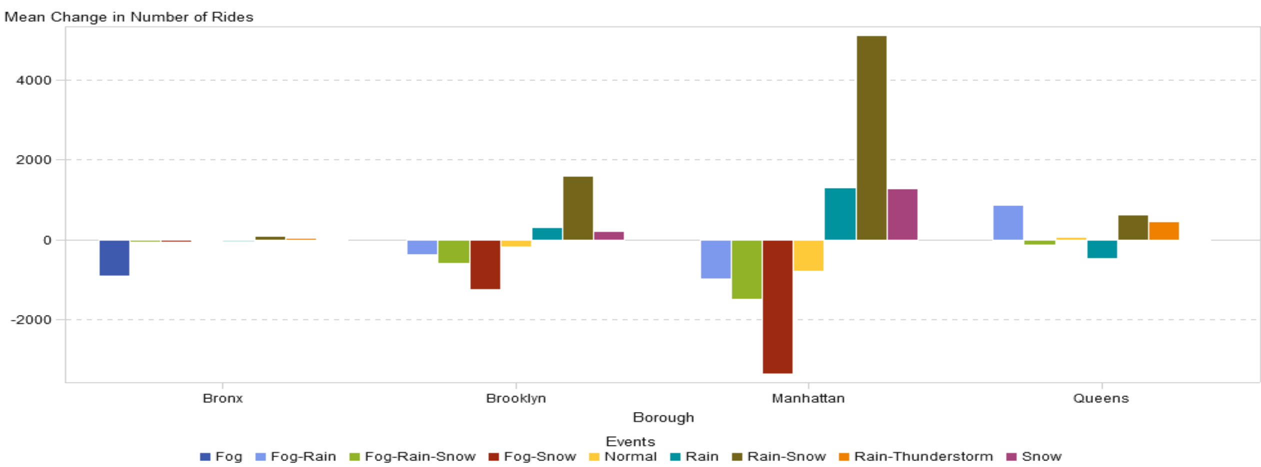

Figure 9. Mean Change in Number of Rides by Borough by Event

Mean change in number of rides is highest on a Rainy day with Snow in every borough except Bronx. In Bronx the change in number of rides is highest on a foggy day.

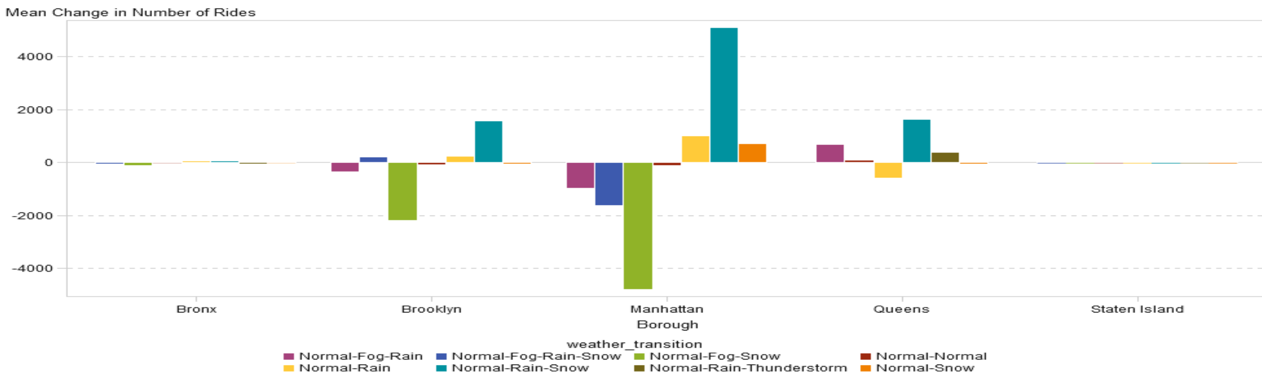

Figure 10. Mean Change in Number of Rides by Borough by Weather Transition

On an average change in number of rides is most prevalent when there is a transition of weather from normal to rain-snow or fog-snow.

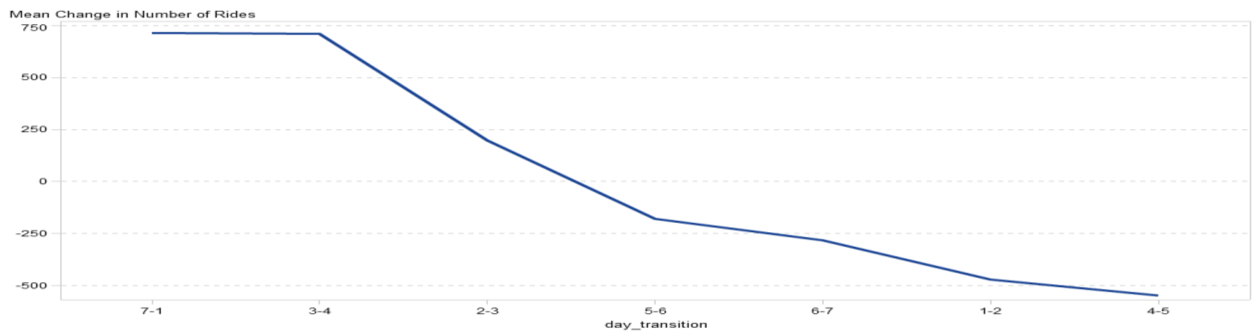

Figure 11. Mean Change in Number of Rides by Day Transition

Mean change in number of rides is highest when compared between Saturday vs. Sunday and Tuesday vs. Wednesday. There is not much change between Wednesday and Thursday.

TIME SERIES FORECASTING

We used SAS Forecast Studio for Desktop 14.1 version to perform time series forecasting. We considered one month of rides (June) as the holdout sample and borough as a by group variable to generate forecasts at borough level. We chose the cutoff to 2% for the detection of outliers. Log transformation is applied on the target variable (Number of Rides) to remove skewness. The interval chosen here is daily. As we found seasonality every week, we configured the seasonality variable to 7. The screenshots reporting the configurations are given below. The transformed target variable along with the transformed explanatory (Weather metrics) variables are fed into ARIMA models.

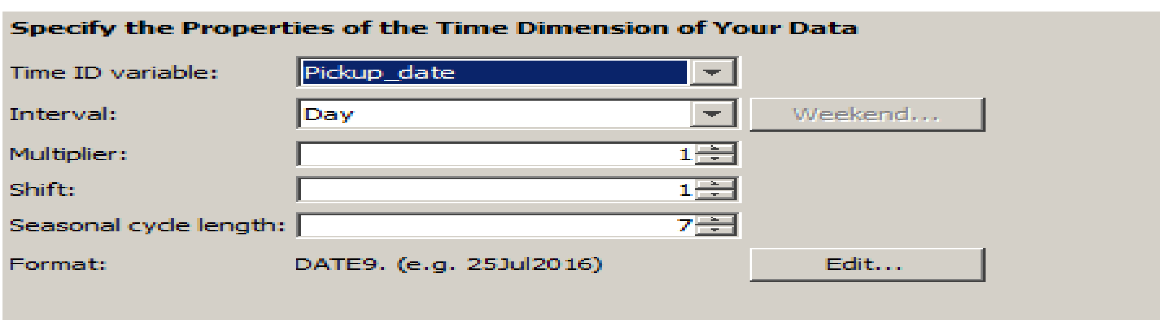

Figure 12. Seasonality Configuration

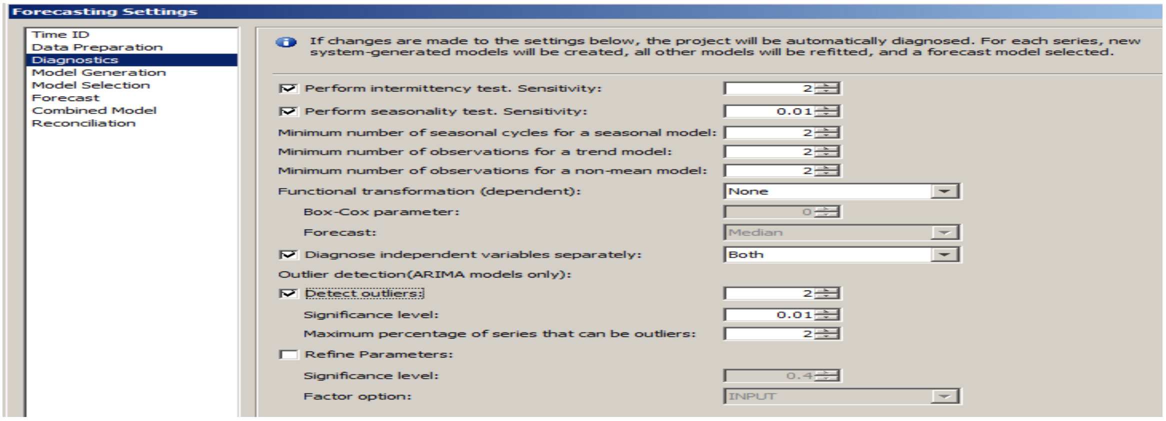

Figure 13. Diagnostic Configuration

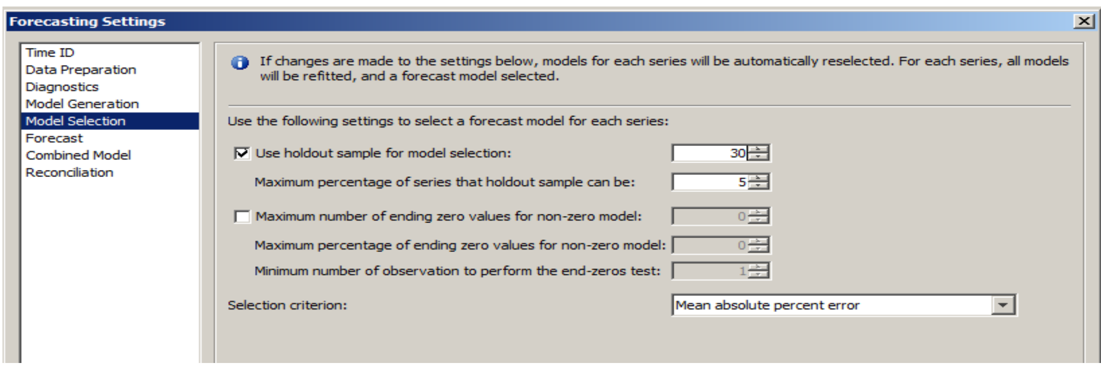

Figure 14. Hold out sample selection

ARIMA Model (Top_1) has been chosen as the best model based on holdout MAPE. Mean Temperature, Mean Visibility Miles, Wind direction, cloud cover and sea pressure turned out to be significant in predicting the number of rides.

BOROUGH	MAPE	Rec. MAPE
Manhattan	5.93	7.08
Queens	9.16	8.94
Brooklyn	10.89	9.14
Bronx	14.97	14.00
Staten Island	21.68	21.71

Figure 15. MAPE values for models at Borough levels

ARIMA Models selected at different borough levels

Bronx: Log_number_of_rides~ 2+ Lag(7)Mean Temperature+ Lag(1)Exp_Mean_VisibilityMiles

Brooklyn: Log_number_of_rides~2+ Lag(10)CloudCover+ Lag(1)Exp_Mean_VisibilityMiles+ Lag(1)Sqrt_Mean_Humidity+ Lag(7)Sqr_WindDirDegrees

Manhattan: Log_number_of_rides~2+ Lag(7)Sqr_windDirDegrees+ Lag(1)Exp_Mean_VisibilityMiles+ AO10MAR2015D+AO21APR2015D

Queens: Log_number_of_rides~Lag(11)Dif(7)Mean_Sea_Level_PressureIn+ AO21APR2015D+AO10MAR2015D

Staten Island : Log_number_of_rides~ Dif(1) Mean_TemperatureF+ Dif(1)Mean_Dew_PointF+ Dif(1)CloudCover+ Dif(1) SQRT_Mean_Wind_SpeedMPH+ Dif(1) SQR_WindDirDegrees

DIAGNOSTIC PLOTS

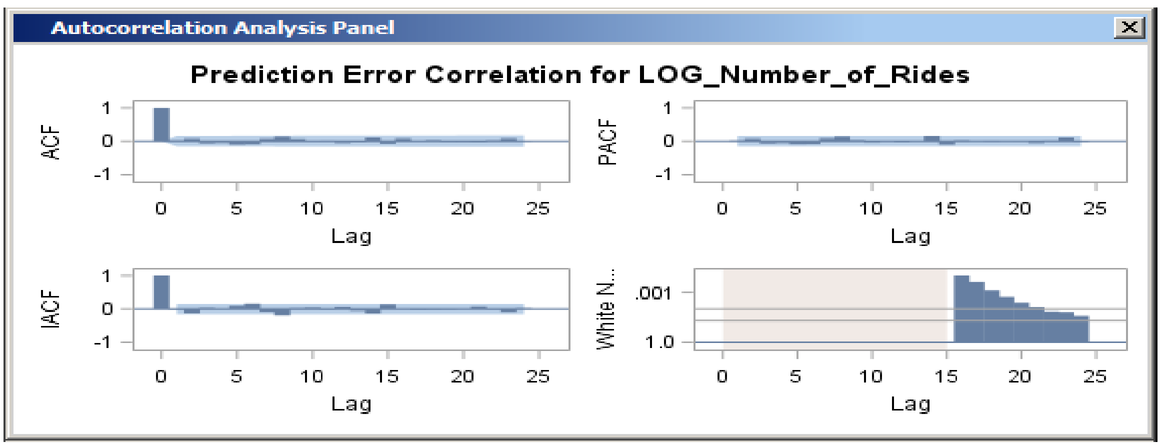

Figure 16. Diagnostic Plots

There appears to be no significant correlation in the residuals.

FORECAST RESULTS

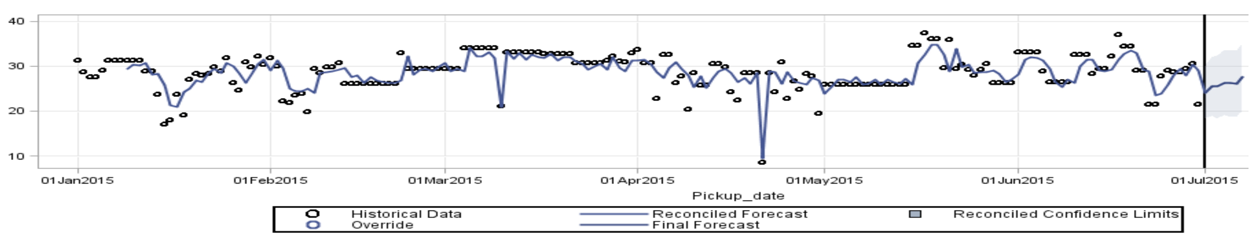

Figure 17. Forecast Output

From the plot we can see that the forecasted output is fitting the actual values very well.

Pickup_date	number_of_rides
01JUL2015	35
02JUL2015	58
03JUL2015	73
04JUL2015	111
05JUL2015	81
06JUL2015	65
07JUL2015	65

Figure 16. Bronx Forecasted Rides

Pickup_date	number_of_rides
01JUL2015	423
02JUL2015	378
03JUL2015	433
04JUL2015	476
05JUL2015	517
06JUL2015	565
07JUL2015	1331

Figure 19.17 Brooklyn Forecasted Rides

Pickup_date	number_of_rides
01JUL2015	360
02JUL2015	575
03JUL2015	384
04JUL2015	359
05JUL2015	372
06JUL2015	410
07JUL2015	280

Figure 180. Queens Forecasted Rides

Pickup_date	number_of_rides
01JUL2015	996
02JUL2015	1672
03JUL2015	1722
04JUL2015	2545
05JUL2015	2623
06JUL2015	2420
07JUL2015	5124

Figure 191. Manhattan Forecasted Rides

Pickup_date	number_of_rides
01JUL2015	4
02JUL2015	5
03JUL2015	4
04JUL2015	5
05JUL2015	5
06JUL2015	5
07JUL2015	6

Figure 22. Staten Island Forecasted Rides

The tables listed above gives the number of rides at borough level for the next one week. The count is obtained by taking the weather metrics for the same duration and performing scenario analysis with those values. The resulting output is log transformed, actual counts are obtained by the exponentiation of the output.

CONCLUSION

- The above results show that weather does influence Uber ridership, especially when it is little distracted from normal but not too extreme again. Manhattan being the most populous city is bringing a huge variation in the number of rides especially on a rainy day.
- The demand is highest during Wednesdays in Manhattan and Sundays in other boroughs.
- The insights derived from exploratory analysis and forecasting model can be very helpful in the optimal utilization of the existing resources as weather can change with time and borough.
- These insights can also be used in the design of incentive programs to uber drivers.
- Staten Island has the least Uber user base, Hence the Mean Absolute Percent Error turned out to be highest as there are not many rides to predict.

FUTURE WORK

The scope of the project can be extended to hourly analysis. Hourly analysis can bring more insights in terms of optimization. Analyzing at least a year worth of data will bring further insights about seasonality. Demand can be more accurately predicted if the actual number of rides requested information is available along with the live number of rides.

REFERENCES

http://toddwschneider.com/posts/a-tale-of-twenty-two-million-citi-bikes-analyzing-the-nyc-bike-share-system/%20-%20citibike-weather

https://newsroom.uber.com/uberdata-uber-for-style-and-comfort/

ACKNOWLEDGEMENT

We thank Dr. Goutam Chakraborty, Professor, Department of Marketing, Director of Business Analytics program- Oklahoma State University for his continuous support and guidance.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Snigdha Gutha

Masters in Business Analytics Program

Phone: 405-780-5622

Email: Snigdha.gutha@okstate.edu