MWSUG 2019 – Paper BL - 068

Taxi Ride Prediction: Does The Yellow Cab Supply Meet Customer Demands?

Sreejita Biswas, Oklahoma State University, Stillwater, Oklahoma

ABSTRACT

New York is the taxi capital of America and home to the classic yellow taxicabs. It would be beneficial to taxi companies and customers if rides are available whenever a customer is in need of one. To achieve this level of service, it is important to know how different factors influence the number of rides. This process gets complicated due to the effects of external forces such as weather and due to pricing strategies employed by other cab companies such as surge pricing during heavy demands times.

This paper attempts to predict the demand of a yellow taxi at a particular location, on a particular day and at a particular time. This will help to estimate the number of taxis that should be present at any given time or place. This project focuses on NY Yellow Taxis dispatched from a central facility in 2018. This paper will help to understand and predict the demand and supply of yellow taxicabs and help to improve customer satisfaction and the taxi cab industry's efficiency. Six months' worth of taxi rides related data (Jan'18 – June'18) from the New York City Taxi and Limousine Commission and weather data for the same period from NOAA (National Oceanic and Atmospheric Administration) were obtained for this research. Information included items such as pick-up/ drop-off locations, time/ date, distance, payment source, temperature, wind speed, and precipitation levels. Multiple models built using SAS® Forecast Studio to predict the demand and supply due to the variations in the weather over six months.

INTRODUCTION

With the introduction of non – traditional taxi services such as Uber and Lyft, it has been a tough ride for the yellow taxis to match up with their approach for pickups and drop-offs, leading to the decline of one of the oldest culture in New York. Recently there is an increase in dissatisfaction of the customers due to the surge prices applied by non - traditional cab companies and resulted in people expressing their desires to go back to yellow taxi services [1]. Even though customers want to use yellow taxi for their commute, the demand and supply of the taxis are not being met [2]. The purpose of this project is to analyze and find a solution for both the taxi companies and the customers where the taxis get benefited by revenue and the customers by service. Customers can pay less for their trips and need not wait for the taxis and the taxis can know where exactly they can get a customer. Also, the drivers of the yellow taxi are facing monetary problems as other cab companies are dominating the market [3]. To keep the yellow taxi tradition alive, this project will attempt to make their process more efficient.

The basic premises for this research are that changes in weather affects the number of rides, and customers care about the fare amount along with the availability of the ride [4]. The premise is based on the general believe that when it rains or when it is too hot to walk down, people tend to go for a cab rather than walking or going through a subway. Likewise, it is believed that during the peak office hours the demand would be more than the regular time and probably the demand would be more on weekend late night when people usually go out and would avoid driving back home.

DATA DICTIONARY

Data regarding taxi trips has been fetched from NYC Taxi and Limousine Commission(TLC). Weather data for the same period of January 2018 – June 2018 has been gathered from NOAA, a website storing all historical weather data.

The combined dataset has three major parts: Taxi trip data, weather data and taxi look up zone data

Dataset Source	Variables in the dataset
NYC Yellow Taxi Trip Dataset	Pickup and Drop off time, Pickup and Drop off location, trip distance, payment type, fare amount, tip, extra and total amount
Weather Dataset	Date, Temperature, precipitation, snow and wind
Taxi look up zone	The pickup and drop off location have IDs which are further divided into Location ID, Borough and Zone. This table helps us to identify the particular boroughs and zones.

Table 1: Datasets used for the paper

DATASET PREPARATION

Each month of taxi trip data had a count of 9 million rows approximately. Weather data obtained for this project was at the level of a day, while the taxi trips were at the level of each minute. Taxi trip data along with weather data was merged with taxi looks up zone data to get the boroughs for each trip. The trips were aggregated together at the day level for each borough. For example, on 2nd January, the total number of rides in the different boroughs with the weather condition for that particular day was prepared.

Because this paper focuses on effects on demand using weather and location, all the fair related data were deleted. There were few rows of taxi data which has locations outside the New York city. Again, because we focus only on NY rides, those rows were deleted. There were no missing values once the fair related data and the unknown locations were deleted.

A new column 'Total_Rides' was created aggregating rides and weather for each day, this variable was used as the target variable. All other variables will be used as input variable, except 'Time' which was used as TimeID for analysis.

Below is a snapshot of first and last few rows of the final dataset used for the paper.

Time	**Wind**	**Temp**	**Precipitation**	**Snow**	**Borough**	**Total_Rides**
1/1/2018 0:00	17.67	12	0	0	Bronx	39
1/1/2018 0:00	17.67	12	0	0	Brooklyn	285
1/1/2018 0:00	17.67	12	0	0	Manhattan	15308
1/1/2018 0:00	17.67	12	0	0	Queens	418
1/1/2018 1:00	17.67	12	0	0	Bronx	58
1/1/2018 1:00	17.67	12	0	0	Brooklyn	791
1/1/2018 1:00	17.67	12	0	0	Manhattan	17111
1/1/2018 1:00	17.67	12	0	0	Queens	539
1/1/2018 2:00	17.67	12	0	0	Bronx	58

....	....	....	....	....	....	....
6/30/2018 21:00	6.93	82	0	0	EWR	2
6/30/2018 21:00	6.93	82	0	0	Manhattan	11893
6/30/2018 21:00	6.93	82	0	0	Queens	783
6/30/2018 22:00	6.93	82	0	0	Bronx	33
6/30/2018 22:00	6.93	82	0	0	Brooklyn	243
6/30/2018 22:00	6.93	82	0	0	Manhattan	12852
6/30/2018 22:00	6.93	82	0	0	Queens	677
6/30/2018 23:00	6.93	82	0	0	Bronx	40
6/30/2018 23:00	6.93	82	0	0	Brooklyn	332
6/30/2018 23:00	6.93	82	0	0	Manhattan	12584
6/30/2018 23:00	6.93	82	0	0	Queens	708

Table 2: Final merged dataset used for analysis

EXPLORATORY ANALYSIS

Starting with the major boroughs in NY, below graph tells us that as expected, Manhattan has the highest number of trips over six months' duration, followed by Queens, Brooklyn and Bronx. The number of rides in EWR and State Island are almost negligible when compared to the other three boroughs.

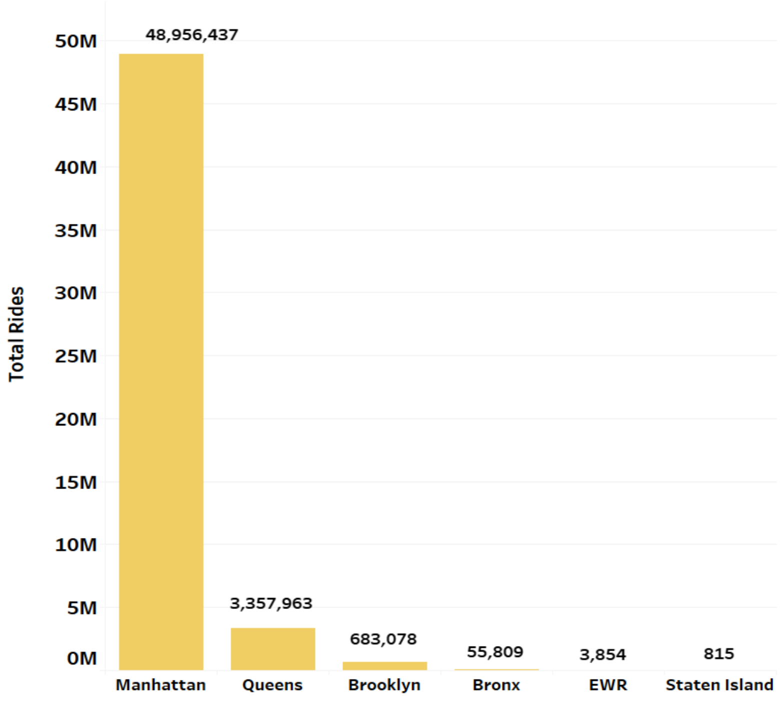


Figure 1: Distribution of total number of rides Borough wise

The month of March has the highest number of rides, followed by April, May, January, June and February. It's difficult to explain why March has the highest and June has the lowest number of rides. However, the colder months of January and February have fewer number of rides as expected.

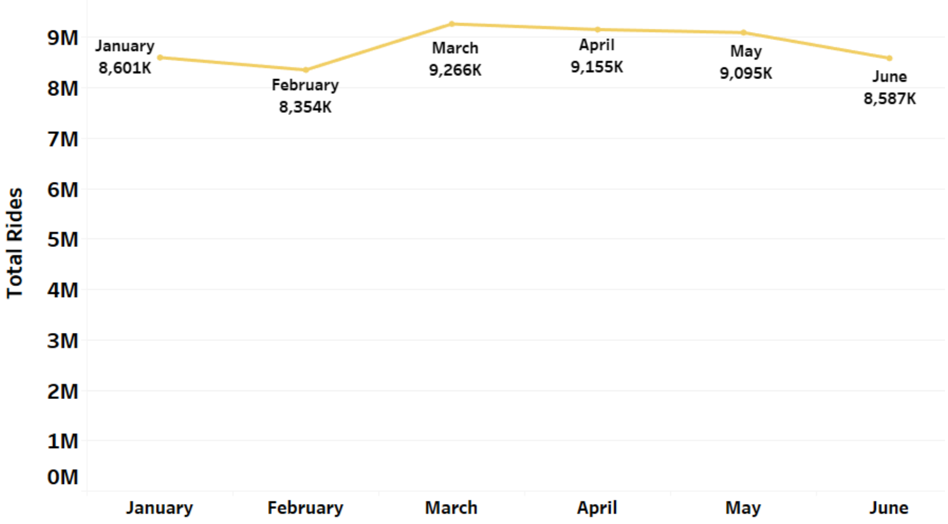


Figure 2: Distribution of total number of rides Month wise

The highest number of rides are usually taken on Friday, which also marks the beginning of the weekend so it makes sense if people try to get a taxi rather than driving themselves. It is followed by Thursday and Saturday. The least number of rides taken are on a Sunday, as most of the people try to get some rest before the next week starts.

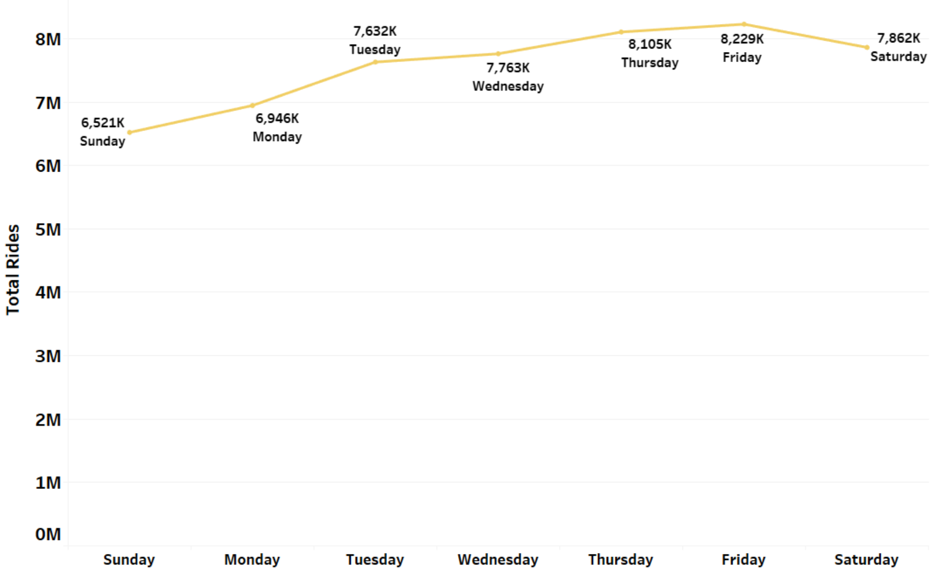


Figure 3: Distribution of total number of rides weekday wise

Evening time, specifically at 18:00 hours maximum rides take place. Perhaps this shows people getting off work and taking cabs to restaurants/bars etc. As expected, the lowest number of rides are the early morning at 4:00 hours.

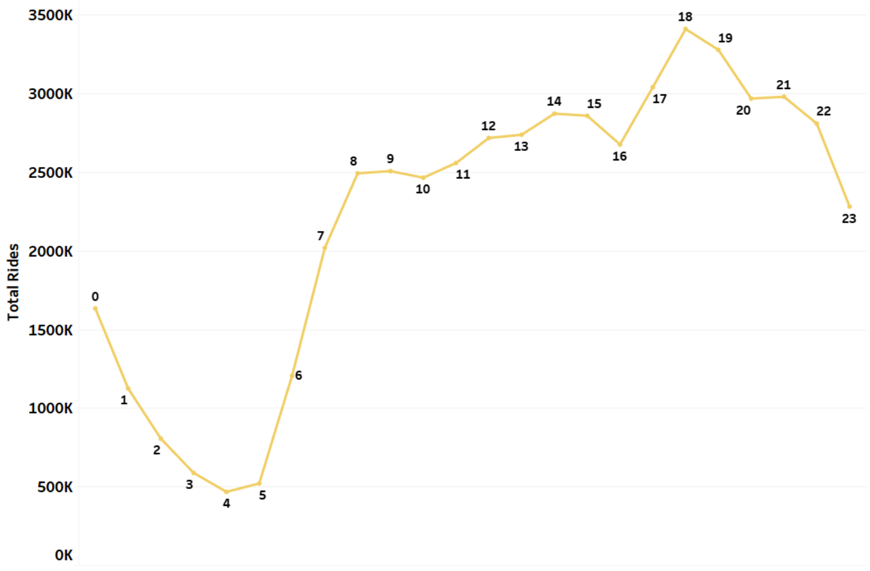


Figure 4: Distribution of total number of rides hour wise

For checking the effects of weather on the number of rides, only top three boroughs have been considered. From the below graph we can see the highest number of rides lie in the middle region of the temperature. This is in contrary to our assumption that when the temperatures are more extremes, people use more cabs. One possibility is that during extreme temperatures, people stay indoors and not use cabs at all. It is also possible that had we used event data such as rain, snow, ice along with temperatures, we would have found more nuanced relationships between weather and number of rides. In this data, the use of cabs seems more dependent on the what day of the week rather than the temperature of the day.

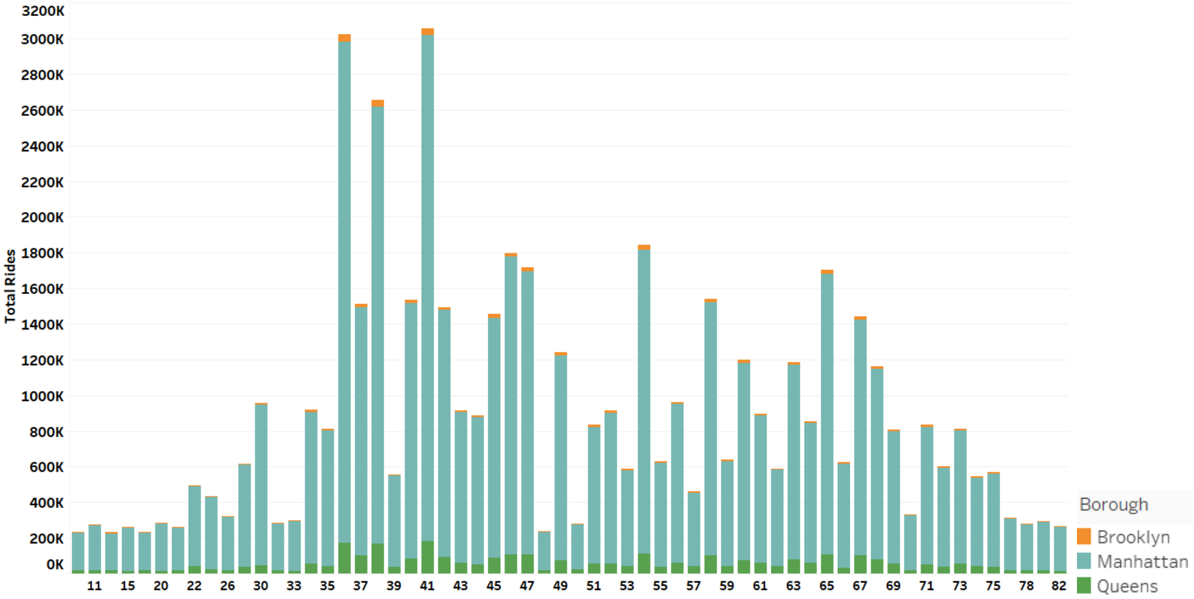


Figure 5: Distribution of total number of rides temperature wise

TIME SERIES FORECASTING

Before starting with the forecast, SAS® Enterprise Miner was used to prepare the dataset for the forecast. As explained in data preparation, all the independent variables had different units of aggregation (day, minute etc.), therefore the dataset was transformed to keep all the variables in the same unit for comparison and then the output was saved as a different dataset. This saved dataset was then used for Forecasting.

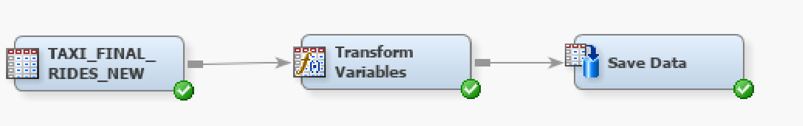
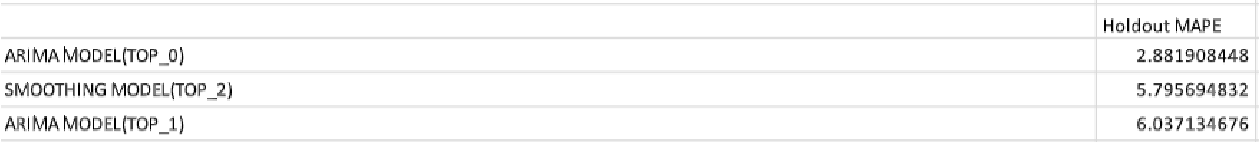


Figure 6: Transformation of variables in SAS Enterprise Miner

Once the data was transformed, it was modeled using SAS® Forecast Studio for time series forecasting. Since we have seen in our exploratory analysis that there is seasonality in week, therefore the season cycle was kept as 7. The month of June (30 days) was kept as the holdout sample for testing. A cutoff of 2% was used for outliers and the time interval was used as daily. This produces the prediction of total number of rides on daily basis. And, borough was used as the grouping variable so we can predict the daily number of rides borough wise. All the weather metrics were fed into multiple ARIMA models.

Based on the parameters set above, SAS® Forecast Studio gave three models. Out of the three tested models, ARIMA model TOP_0 was chosen for forecasting based on the holdout MAPE of 2.88%.



	Holdout MAPE
ARIMA MODEL(TOP_0)	2.881908448
SMOOTHING MODEL(TOP_2)	5.795694832
ARIMA MODEL(TOP_1)	6.037134676

Below are the screenshots of the best model's performance as a whole (not borough wise). The model seems to track the actual data closely for calibration/validation period (Jan – May) and also the holdout period (June). The parameters that were statistically significant were borough and snow, but in our dataset majority of the dataset has a snow value of zero. Therefore, snow coming as significant is not as useful. But two takeaway points from here are, as expected borough came as significant due to the number of rides and that temperature is not a significant factor determining the number of yellow taxi rides.

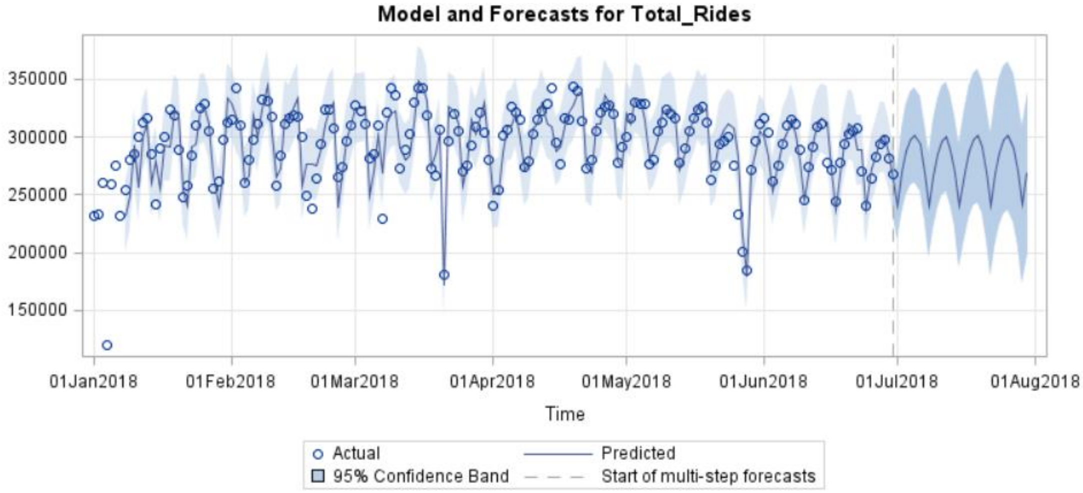


Figure 7: Time Series Forecast of the whole Model

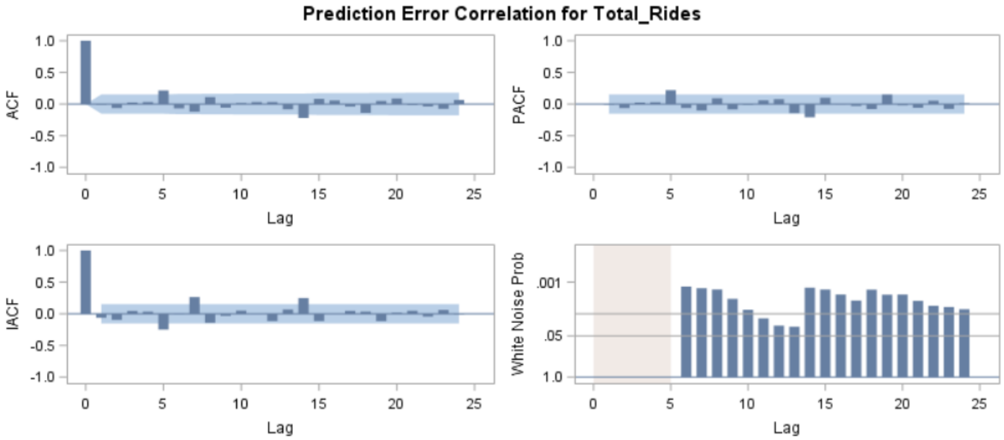


Figure 8: Prediction Error Correlation

From the above figure we can see that the autocorrelation values are mostly within the expected range and the fitted model exhibits expected white noise. While the model seems to work reasonably well as a whole, we also explored how it performed at a borough level. For this purpose, we first selected Manhattan because it is the most popular borough.

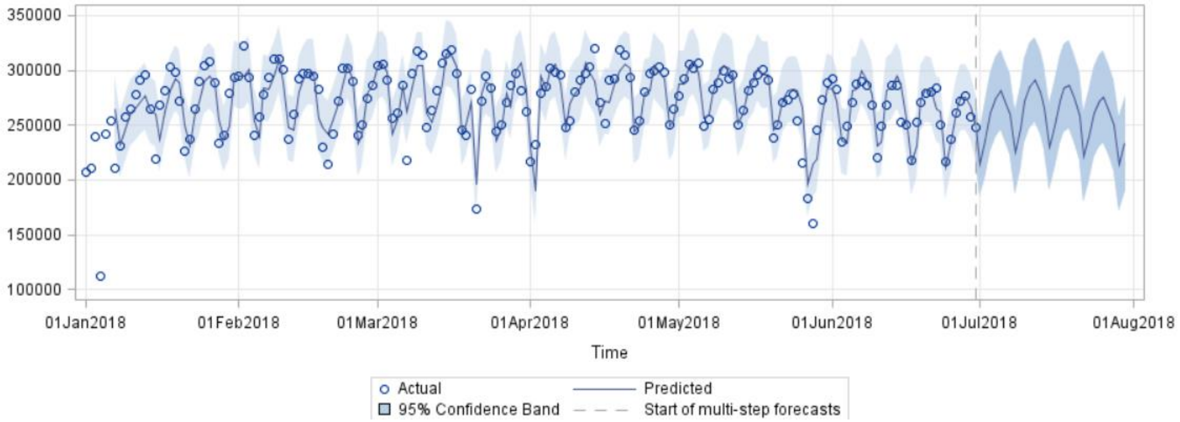
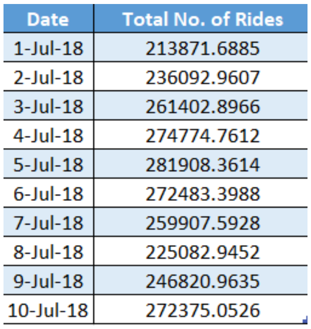


Figure 9: Time Series Forecast for Manhattan

Except few of the points the forecasted values are fitting well with the actual values. Below is a screenshot of the predicted number of rides for the month of July (test dataset).



Date	Total No. of Rides
1-Jul-18	213871.6885
2-Jul-18	236092.9607
3-Jul-18	261402.8966
4-Jul-18	274774.7612
5-Jul-18	281908.3614
6-Jul-18	272483.3988
7-Jul-18	259907.5928
8-Jul-18	225082.9452
9-Jul-18	246820.9635
10-Jul-18	272375.0526

Figure 10: Predicted Number of Rides in July for Manhattan

Since Manhattan has the highest number of rides, an hourly forecasting was done only for this borough so that better insights can be developed for the most important borough.

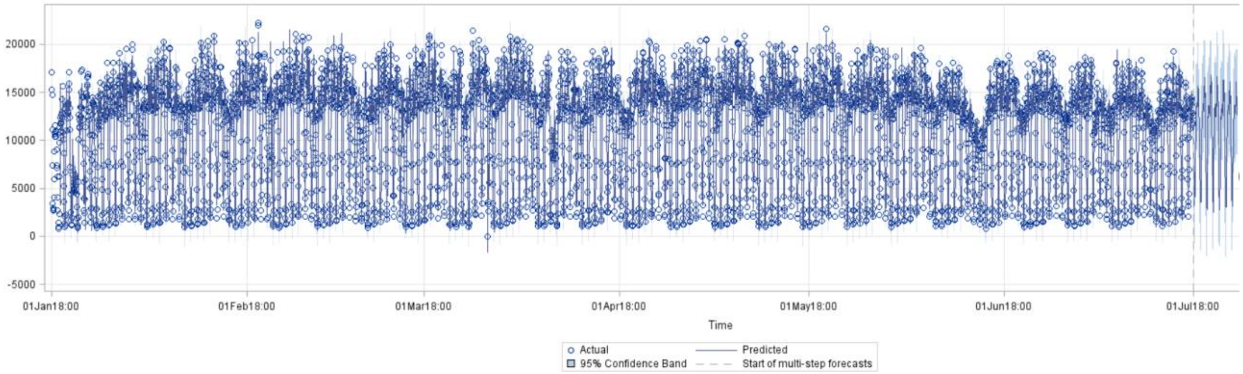
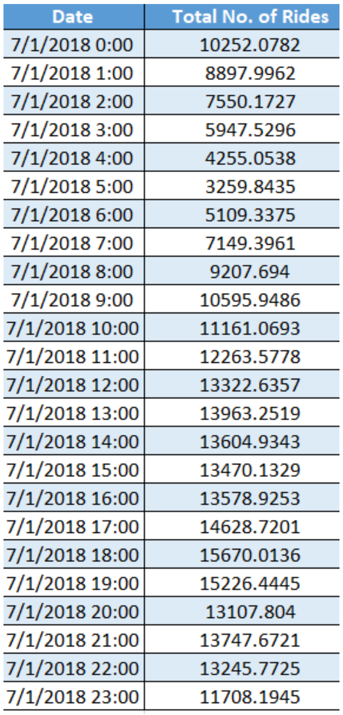


Figure 11: Time Series forecast for July in Manhattan

Looking at the hourly prediction, the forecast values are a good fit for the actual values as well.



Date	Total No. of Rides
7/1/2018 0:00	10252.0782
7/1/2018 1:00	8897.9962
7/1/2018 2:00	7550.1727
7/1/2018 3:00	5947.5296
7/1/2018 4:00	4255.0538
7/1/2018 5:00	3259.8435
7/1/2018 6:00	5109.3375
7/1/2018 7:00	7149.3961
7/1/2018 8:00	9207.694
7/1/2018 9:00	10595.9486
7/1/2018 10:00	11161.0693
7/1/2018 11:00	12263.5778
7/1/2018 12:00	13322.6357
7/1/2018 13:00	13963.2519
7/1/2018 14:00	13604.9343
7/1/2018 15:00	13470.1329
7/1/2018 16:00	13578.9253
7/1/2018 17:00	14628.7201
7/1/2018 18:00	15670.0136
7/1/2018 19:00	15226.4445
7/1/2018 20:00	13107.804
7/1/2018 21:00	13747.6721
7/1/2018 22:00	13245.7725
7/1/2018 23:00	11708.1945

Figure 12: Predicted Number of Rides on 1st July for Manhattan

We also explored how the model performs for the other boroughs as shown in screenshots below.

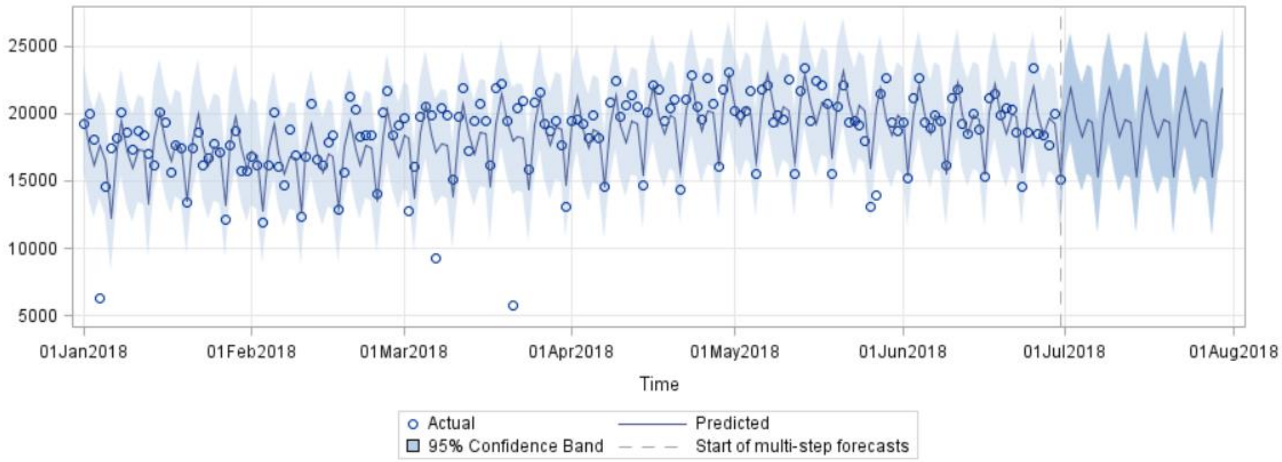


Figure 13: Time Series Forecast for Queens

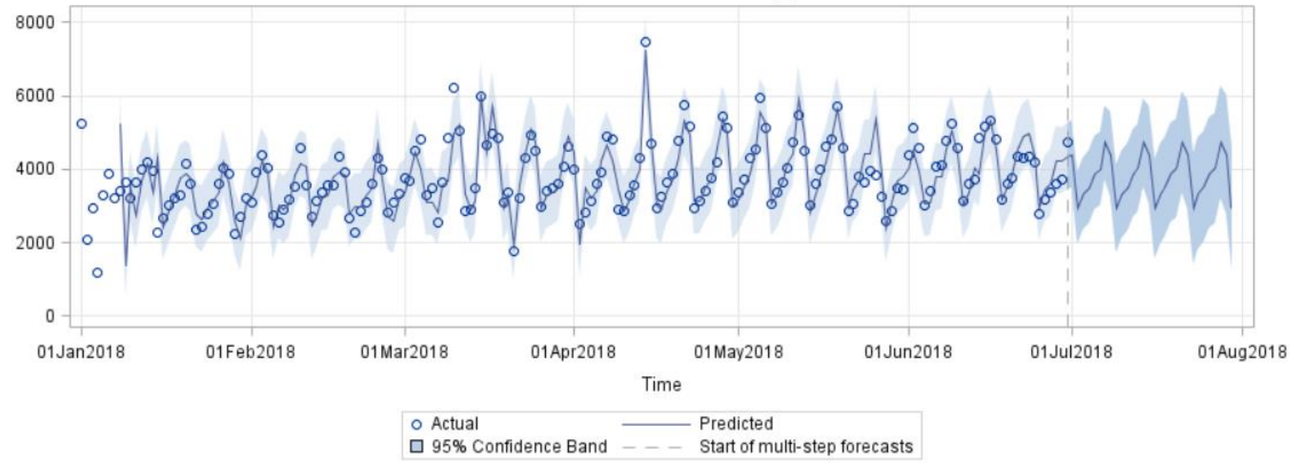


Figure 14: Time Series Forecast for Brooklyn

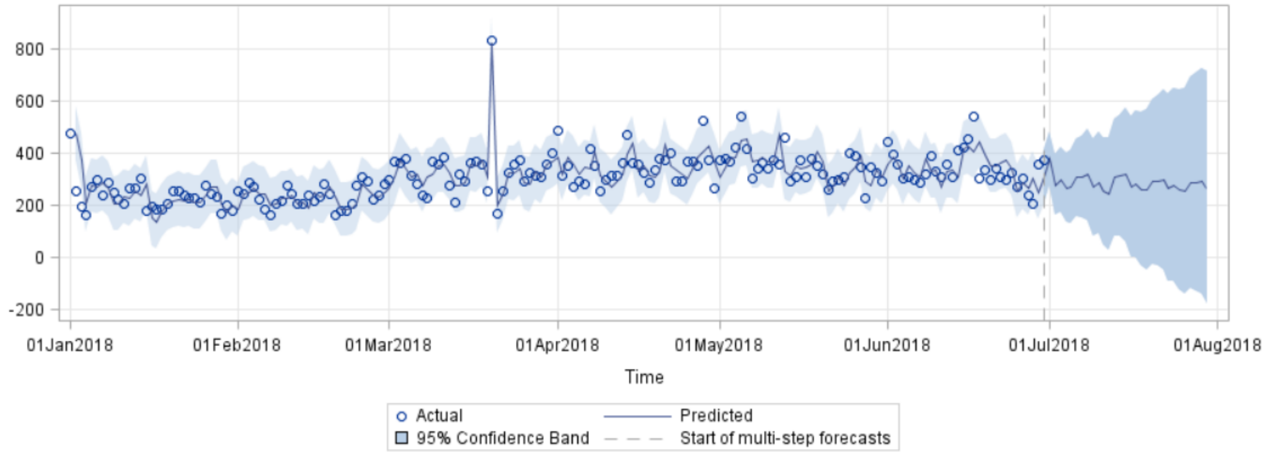
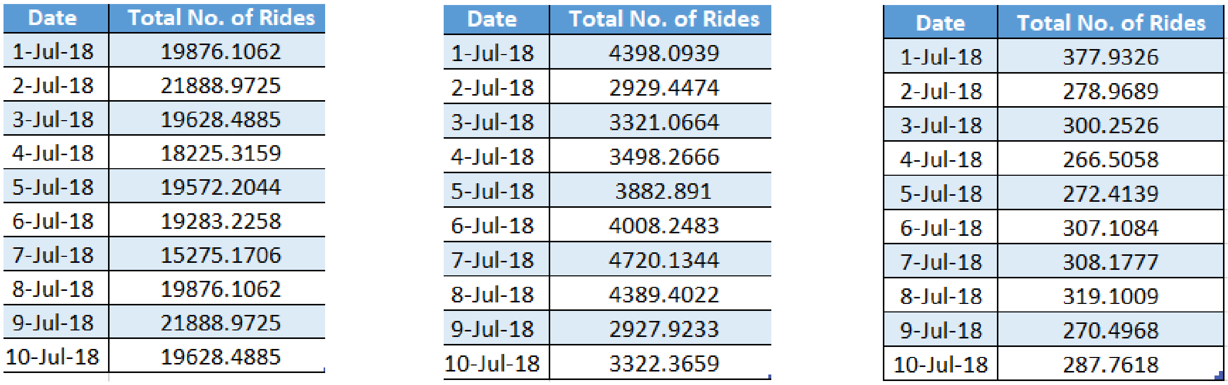


Figure 15: Time Series Forecast for Bronx



Date	Total No. of Rides
1-Jul-18	19876.1062
2-Jul-18	21888.9725
3-Jul-18	19628.4885
4-Jul-18	18225.3159
5-Jul-18	19572.2044
6-Jul-18	19283.2258
7-Jul-18	15275.1706
8-Jul-18	19876.1062
9-Jul-18	21888.9725
10-Jul-18	19628.4885

Date	Total No. of Rides
1-Jul-18	4398.0939
2-Jul-18	2929.4474
3-Jul-18	3321.0664
4-Jul-18	3498.2666
5-Jul-18	3882.891
6-Jul-18	4008.2483
7-Jul-18	4720.1344
8-Jul-18	4389.4022
9-Jul-18	2927.9233
10-Jul-18	3322.3659

Date	Total No. of Rides
1-Jul-18	377.9326
2-Jul-18	278.9689
3-Jul-18	300.2526
4-Jul-18	266.5058
5-Jul-18	272.4139
6-Jul-18	307.1084
7-Jul-18	308.1777
8-Jul-18	319.1009
9-Jul-18	270.4968
10-Jul-18	287.7618

Figure 16: Predicted Number of Rides in July for Queens, Brooklyn and Bronx respectively

State Island has the least number of rides and below is the forecast for that.

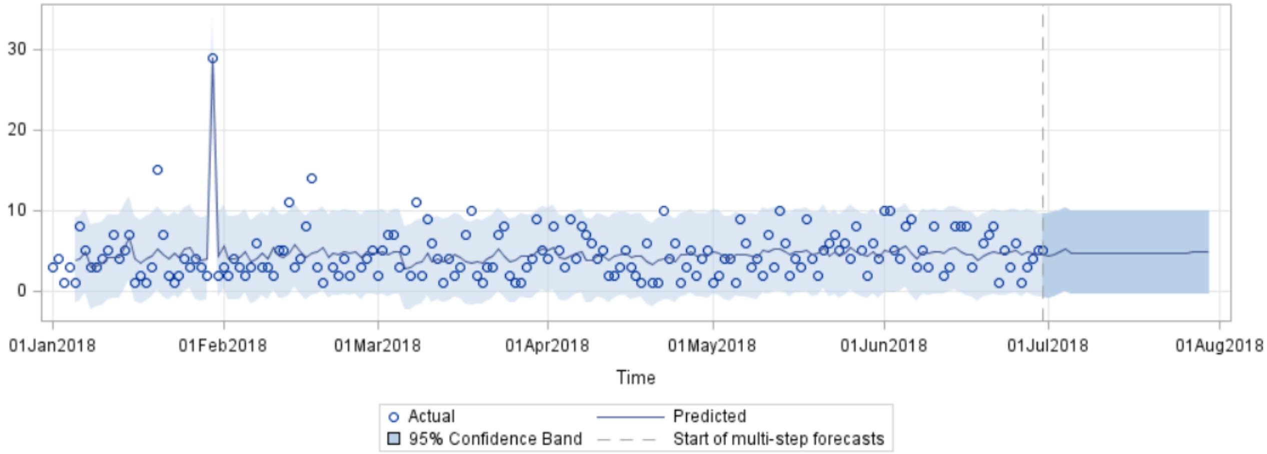
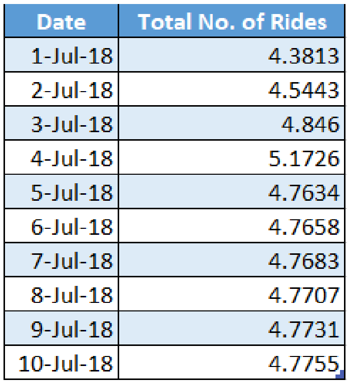


Figure 17: Time Series Forecast for State Island

From the above graph we can see that the forecasting values doesn't really fit with the actual values and this maybe because of the very less number of rides, SAS® Forecast Studio is not predicting it as good as other boroughs.



Date	Total No. of Rides
1-Jul-18	4.3813
2-Jul-18	4.5443
3-Jul-18	4.846
4-Jul-18	5.1726
5-Jul-18	4.7634
6-Jul-18	4.7658
7-Jul-18	4.7683
8-Jul-18	4.7707
9-Jul-18	4.7731
10-Jul-18	4.7755

Figure 18: Predicted No. of Rides for State Island

The above tables and forecasts give the number of rides for the month of July for the different boroughs. The number of rides have been calculated keeping the weather metrics for the same duration and performing the analysis with those values.

CONCLUSION

From the above analysis and forecasting, few insights were derived. Weather doesn't have a significant effect on the number of rides in any of the boroughs. Reason might be that Yellow taxis are hired on roads rather than other non-traditional taxis were home pickups are done. So whatever the weather condition is people take rides in yellow taxis. The demand of taxis is highest on Fridays followed by Thursdays and Saturdays, same as the pattern in Manhattan (majority of the rides are from Manhattan). The peak time for taxi rides are at 18:00 hours and the least is during 4:00 hours. Month of March has the highest number of rides followed by April and May. The forecasting is a good fit for all the boroughs except State Island since it has a very low number of rides. The hourly prediction of rides at the boroughs can really help the New York Yellow taxi cab company and the drivers to plan ahead and optimally utilize this forecasting to meet the demand of their customers.

This project had the limitations of lack of multivariate time series modeling. Also since yellow taxis doesn't have any surges depending on time of the day, fare has been excluded from the analysis. Hourly weather data with weather event of the day like rain, sunny etc. was not obtained for the project.

FUTURE WORK

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. Hourly weather data along with weather event might bring more insights if weather affects the number of rides. Also a forecast and prediction based on zones, along with boroughs will make it very easy for the drivers to be present at the location at any given point of time.

REFERENCES

[1]https://www.investopedia.com/articles/personal-finance/021015/uber-versus-yellow-cabs-new-york-city.asp
[2] https://www.nydailynews.com/opinion/ny-oped-how-the-yellow-cab-went-belly-up-20180803-story.html
[3] https://mic.com/articles/191390/new-york-city-cab-drivers-face-depression-and-debt-amid-increased-competition-from-uber-and-lyft#.F7d1BE7h9
[4] http://support.sas.com/resources/papers/proceedings17/1260-2017.pdf

https://www.investopedia.com/articles/personal-finance/021015/uber-versus-yellow-cabs-new-york-city.asp

https://www.nydailynews.com/opinion/ny-oped-how-the-yellow-cab-went-belly-up-20180803-story.html

https://www.mic.com/articles/191390/new-york-city-cab-drivers-face-depression-and-debt-amid-increased-competition-from-uber-and-lyft#.F7d1BE7h9

http://support.sas.com/resources/papers/proceedings17/1260-2017.pdf