# IE 515 Transportation Analytics
# ANALYSIS OF NEW YORK CITY'S CITI BIKE SHARE DEMAND

Lavkesh Rajwani, Hitesh Gohil, Forum Bhutaiya, Ganesh Nagre

DECEMBER 13, 2018
UNIVERSITY AT BUFFALO

**ABSTRACT**

Bike Sharing Systems are booming in the last few years in most of the major cities of USA including, San Francisco, California, Washington D.C. and New York. Due to bustling city lives, people are switching to bike share for faster, economical & eco-friendly commutes. Increase in usage of bike sharing systems has led to challenging situations of bike unavailability and over-crowded dock stations. The objective of the study in this project is to predict the weekly bike demand for New York City's Citi BikeShare service for the coming year on the basis of the usage and demand of the previous stations. This helps not only in the future prediction of the bike demand but also provides data regarding the crowding at bike stations which helps in restoring of bikes depending on the demand forecasted. The collection and processing of the data are initially presented. The bicycle usage pattern is then analyzed and predicted using the various transportation models, of which the TBATS modeling provides the best fit. Results indicate the demand forecast between the busiest stations which helps Citi Bike to restore bikes when the demand gets low.

# Table of Contents



# List of Figures

# List of Tables

## RESPONSE TO REVIEWS

Here are the responses to the comments on our presentation:

1. What is the sampling rate? the research needs a centric topic and deep analysis.
    - We did not understand what you mean by Sampling Rate but we have worked towards a central topic for predicting bike demand at stations.

2. Need to compare different results from all the methods used
    - We have compared the goodness of fit for the three different models we have used.

3. Very interesting topic! The TBATS model seems to have a really good shape, it would be more persuading if it is compared with the observed data.
    - Thank you, the TBATS model predicted for the year 2017 is compared to the actual data of the year 2017 and it is found to be a very good fit. This was shown in the presentation as well and is depicted further in this report.

4. What is the final problem you are going to target? Are you also considering the demand based on hour of the day?
    - The target of the project is to forecast weekly demand of the bike share service. We considered the hourly demand but not demand based on hour of the day.

5. The group should have a solid base in the form of the literature reviews to verify the usage of models that they employ for prediction. Also, it can be useful to validate the results.
    - The current literature review provides precise details related to the topic and the usage of the models for prediction.

6. If only most important plots were shown.
    - The question of the group seems to be unclear. This report consists of all the important and useful plots based on the analysis, prediction, fit and accuracy of the data. We feel that these plots were important to showing why we rejected some models and how our models worked.

7. Most graphs were hard to read, didn't understand their conclusions.
    - We explained the plots in our verbal presentation. The group may not have paid attention towards it.

8. Did you try to improve the R-sq value of Linear Regression by using different factors?
    - The linear regression model is not a suitable model to be used so we moved away from that model and worked on others to get a better fit.

9. Time management could be improved for Presentation
    - Sorry for that, we will be careful next time.

10. Combine months of data and take a sample instead of only December data
    - The December data was for sampling. We have used the entire year's data for the prediction.

# 1. INTRODUCTION

Currently, vehicles are increasing day by day in the bustling cities like New York, San Francisco, Boston, Chicago etc. This leads to a rise in heavy traffic which creates the need for an alternate mode of transportation. Therefore, bike sharing is gaining popularity in a number of cities due to eco-friendly and a cost-effective alternative.

Citi Bike is a privately-owned bike-share program in New York which opened to the public in May 2013. The durable Citi Bikes and docking stations provide convenient and inexpensive mobility twenty-four hours a day, 365 days a year. Citi Bike is available for use 24 hours/day, 7 days/week, 365 days/year, and riders have access to thousands of bikes at hundreds of stations across Manhattan, Brooklyn, Queens, and Jersey City. They have currently 143,000 members which have traveled over 50 millionth rides since its inception. It owns nearly 15000 bikes distributed among 800 stations in and around the New York City.

Demand for bikes depends on many external factors like time of the day, the day of the week, type of user and also depends on several environmental factors such as temperature, humidity. Other factors that would affect are random factors like holidays and proximity to other bike stations, adding to the complexity of the problem.

Even though the bike sharing service provides convenient traveling facilities to the users, there are also many problems with which the customer complains. Through research, it was found that one of the major issues is the availability of parking at dock stations and also the availability of bikes at sparsely docked bike stations. This is due to the varying demand for bikes at the stations. Some station stays in peak demand at all the times whereas a few stations are seen to be docked nicely but the demand at those stations is not high. There are also times when the bikes are damaged and need service to hit the roads smoothly so the issue of rebalancing and restoring bikes at stations is aroused. Therefore, the aim is to find the weekly demand generated throughout the year at the highly demanding bike stations in NYC so that the bikes can be restored and rebalanced in the free time.

The approach to this problem is to forecast a demand using the mathematical model of the busiest stations which will help the CITI bike share service for rebalancing and restoring the bikes. This is done by combining historical usage patterns with the related information of users.

In this report, the prediction is done for the top 2 busiest route stations and this can be further implemented to find the rental demand for the remaining stations. The focus is concentrated on the weekly demand between the two stations ignoring the factors such as holiday, event and also the weeks into weekdays and weekends etc. Therefore, these factors can be taken into consideration for further study.

## 2. LITERATURE REVIEW

Significant efforts have been put in towards predicting demands of transport vehicles like bikes and cars to improve fleet management and minimize imbalances caused due to various physical and environmental factors. Over the last couple of years, the literature on bike sharing systems has emerged to address its various insufficient and needs for optimized systems. Few studies worth mentioning are:

Mitesh Gadgil et al. studied and evaluated various different approaches to forecast hourly demand of bikes at different stations in San Francisco Bay Area using different features like location, time of the day and also weather conditions to improve predictions. LASSO Regression was used to find top attributes of weather and the method of L1 regularization was used to retain only the features that are actually useful. Gradient boosting regression model was used to predict the hourly demand of the bikes. According to the authors, the data of the previous few hours along with the required particular hour' data was used for prediction to obtain proximity results.

Jayant Malani et al. studied and predicted the hourly demand of a large number of data points in the training set i.e. 443,591 and around 45 features of the Capital Bike share in Washington D.C. Various modeling regressors such as linear regressor, XGBoost regressor, Gradient Boosting regressor, Random Forest regressor, ExtraTree regressor, Bagging Regressor, Ensemble of Regressors and Time Series are used and compared. The XGBoost regressor outperformed all other regressors with Root mean square log error on the test set of 0.40505 and Average root mean square log error for 10-fold cross-validation of 0.40350.

Xiaomei Xu et al. proposed the users' demand prediction model using Back-Propagation Neural Networks to establish different demand prediction models according to different types of stations and days. A comparative analysis was employed to determine if the performance of prediction models improved by making a distinction among stations and working/nonworking days. The entire dataset was classified in 2 scenarios. Scenario 1 is the one which only makes a distinction

between working days and nonworking days and Scenario 2 which only makes a distinction among stations separately by using the real-life case of Wenzhou bike sharing system. It was found that in scenario 1, the R2 value of working days is 0.59 and for nonworking days is 0.52 whereas in scenario 2, the R2 value of cluster 1 is 0.71, and for cluster 2 is 0.64.

Leonardo Caggiani et al. presented a micro-simulation model for an optimal relocation of bikes in bike-sharing systems considering the dynamic variations of the demand for both bikes and free docking slot and micro-simulate the BSS (Bike Sharing Systems) in space and time. It determines the optimal repositioning flows, distribution patterns and time intervals between relocation operations by explicitly considering the route choice for trucks among the stations. The result shows that the relocation management increases users' satisfaction in terms of probability of finding bikes or free docking point. The author suggests that the proposed decision support system is more suitable for non-congested bike sharing systems and it is a modular method that can be used for wider systems.

Wen Wang studied and tested four modeling algorithms, that is, linear regression, neural network, decision tree, and random forest. On comparing the validation performance by 70/30 method, it was found that the random forest provided the best fit. The random forest model decreased RMSLE value from 0.449 to 0.265. Wen showed that Random forest is a good ensemble algorithm which combines both decision tree and bagging. He also states that there will always be a better model for prediction and no model is perfect.

## 3. DATA DESCRIPTION

Many bike sharing companies like Citi bike (New York), Divvy Bike (Chicago) have made their usage data available for open source use. The companies have been inclined to make their data open for public use to crowdsource solutions.

The NYC Bike Share operates Citi Bike program and generates data regarding the program, including trip records, a real time feed of station status and monthly reports. The Citi Bike program data is exclusively generated by the operator NYC Bike Share, a limited liability corporation solely owned by Motivate. The data has been processed to remove trips taken by staff for service reasons and trips that were below 60 seconds in length that is due to false starts or users trying to re-dock a bike to ensure if it's secure.

The historical yearly weather data was collected from Kaggle. The weather data of the year 2016 was available and it contained attributes such as temperature, humidity, wind speed, wind direction, visibility, pressure, heat index, precipitation, rain, fog and snow. Of all these data

attributes available, temperature, humidity and wind speed are used as their data was relatively dense.

The data of Citi Bike is available since its inception in May 2013 till date. It includes features such as Trip Duration, Start Time and Date, Stop Time and Date, Start Station Name, End Station Name, Station ID, Station Latitude/Longitude, Bike ID, User Type (Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member), Gender (Zero=unknown; 1=male; 2=female), Year of Birth.

The dataset and the variables in the dataset obtained from the website of Citi BikeShare as shown in the table below:

*Table 1: Dataset features and their format*

| # | FEATURE | DESCRIPTION | FORMAT |
|---|---|---|---|
| 1. | Trip duration | Duration of the bike trip. | Continuous (In Seconds) |
| 2. | Start time | The start time of the trip. | Timestamp |
| 3. | Stop time | The end time of the trip, | Timestamp |
| 4. | Start station ID | Station ID of the station where the trip started | Number |
| 5. | Start station name | Station name of the station where the trip ended | Category (String) |
| 6. | End station ID | Station ID of the station where the trip ended | Number |
| 7. | End station name | Station name of the station where the trip ended | String |
| 8. | Bike ID | Unique ID of the bike used for the trip | Number |
| 9. | Start/End Station Longitude | The exact longitude location on the map of the start and end stations | Number |
| 10. | Start/End Station Latitude | The exact latitude location on the map of the start and end stations | Number |

| | | | |
|---|---|---|---|
| 9. | User type | Subscriber = Annual Member<br><br>Customer = 24 Hour pass or 7 Day pass | Category (Subscriber/Customer) |
| 10. | Birth year | The birth Year of customer | Number |
| 11. | Gender | 0 = None, 1 = Male, 2 = Female | Category (String) |
| 12. | Temperature | Maximum and Minimum temperature in Fahrenheit | Number |
| 13. | Humidity | Mean Humidity in % | Number |
| 14. | Wind Speed | Maximum and Minimum Wind Speed in mph | Number |

The Citi Bike share data for the year 2015 and 2016 is used to predict the weekly bike count for the year 2017. Table 1 shows the data attributes actually used in this project.

### 3.1. DATA SUMMARY & STATISTICS

The entire data was vast and had many attributes that were not needed for this project. The major data relevant and important statistics are given below in the overview of the data related to the Citi Bike share:

*Table 2: Data Overview*

| Parameter | 2015 | 2016 | 2017 |
|---|---|---|---|
| Number of Trips | 937969 | 13845655 | 16364657 |
| Number of Bikes | 8477 | 10581 | 14204 |
| Number of Stations | 497 | 654 | 819 |

## 4. METHODOLOGY:

The aim of the project is to predict the bike usage data at popular stations of Citi Bike share service. There are various ways to predict and interpret the usage demand, each way being informative in some sense. Forecasting total demand would help the administrators of Citi Bike, see if their resources are being over or underutilized, and whether they would need to purchase additional bikes to cater to varying demands in the future. We will try different models and find the best suited one.

## 4.1. LINEAR REGRESSION

The Citi Bike data available to us had many variables. Potentially amongst them, the trip duration of a bike borrowed seemed like a good variable that can be estimated. The other variables included categorical variables logging the birth year, gender and user_type. However, with these variables, there was no literature which stated them to be good estimators for determining the trip duration. Besides that, the birth year registered in the data had no surety to be the real birth year of the user. There is no reliability over the demographic data that was present. The other variables which were defined within this dataset were the Starting location of the trip and the corresponding End location. This opened up a scope for analyzing the demand for the Citi Bikes across the city in a given month.

Another dataset available to us was through kaggle.com for the weather of NYC in the year 2016. This dataset included temperature, wind speed, humidity, and other categorical parameters like rain, visibility and snow, these entries were recorded every hour. We studied literature that used random forest classifier to identify the bike demand on an hourly basis and the estimators used were the temperatures, humidity, wind speed, rain, snow, and cloudy conditions. Therefore, we considered fitting the hourly demand of the bikes over all the stations in New York City over the month of December 2016. By analyzing the data, we realized the, it was largely sparse for categorical variables like rain and snow which could have been good estimators for the demand. However, we proceeded to try and model the hourly demand in the month of December against the temperature, humidity and wind speed which were continuous forms of predictors. The fit obtained from this was a poor one.

```
Call:
lm(formula = V1 ~ rain + tempi + hum, data = testdata)

Residuals:
    Min      1Q  Median      3Q     Max
-1570.0  -611.5  -145.9   368.4  3675.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1712.530    152.054  11.263  < 2e-16 ***
rain1       -540.784    181.915  -2.973  0.00308 **
tempi         13.542      2.339   5.790 1.18e-08 ***
hum          -15.669      2.132  -7.348 7.33e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 889.7 on 548 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.1606,   Adjusted R-squared:  0.156
F-statistic: 34.95 on 3 and 548 DF,  p-value: < 2.2e-16
```

The adjusted R-squared value obtained from this model indicates that the model is a terrible fit. However significant may be the temperature and humidity to the bike count per hour, it is still not a good predictive model, Hence, we did not proceed to test the accuracy of this model as the fit itself is below satisfactory. The plot justifies that the observations are not normally distributed, and the fitted vs. residual plot doesn't strictly follow a horizontal line through zero.

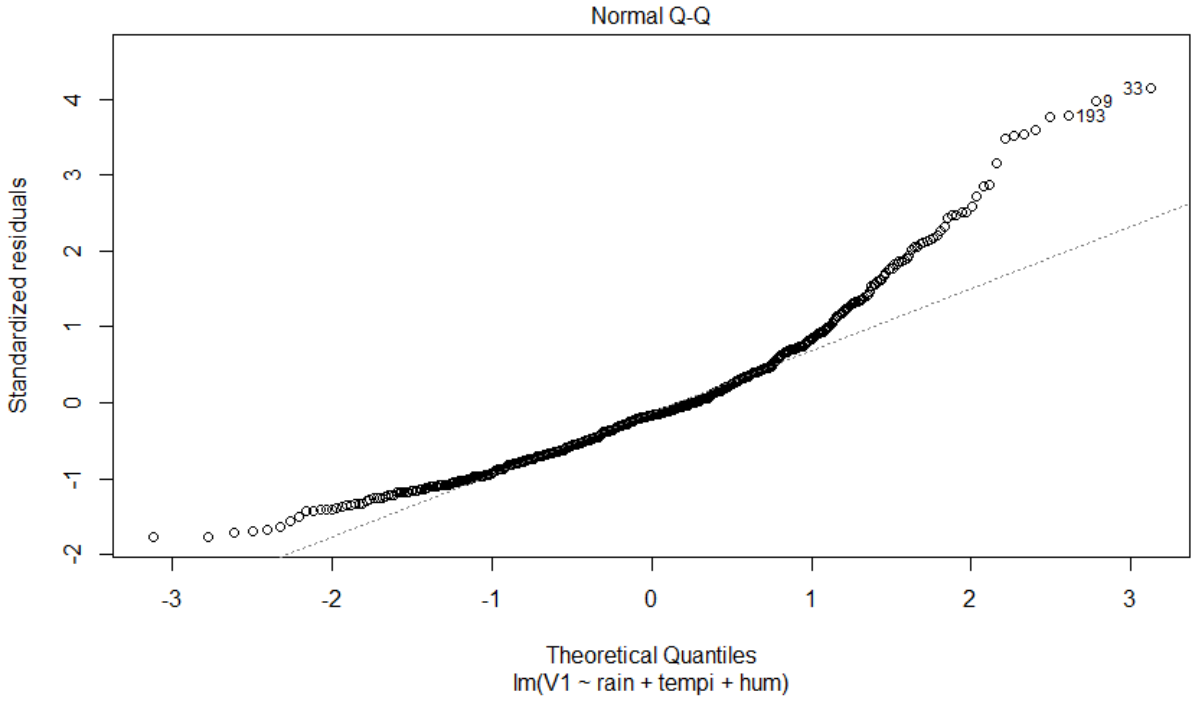



*Figure 1: Normal Quantile - Quantile Plot*

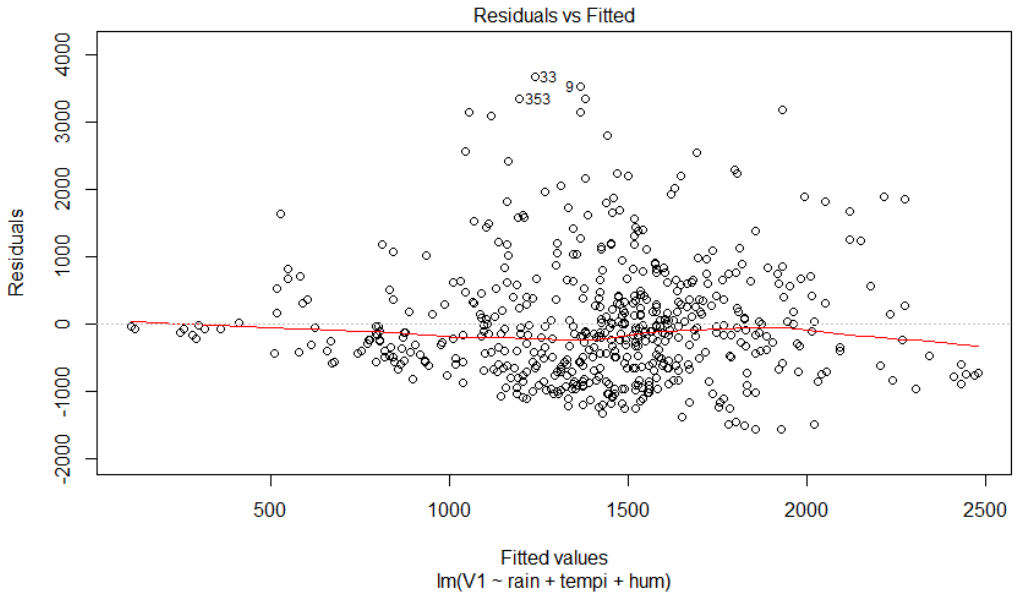



*Figure 2: Residuals vs Fitted Plot*

This is where we decide to terminate the regression approach and since demand is what we wanted to predict we wanted to see if we can model it as a time series, in order to have an autoregressive series for forecasting. In order to verify if this could be modeled as a time series object, we plotted the hourly demand of the month of December as a time series.

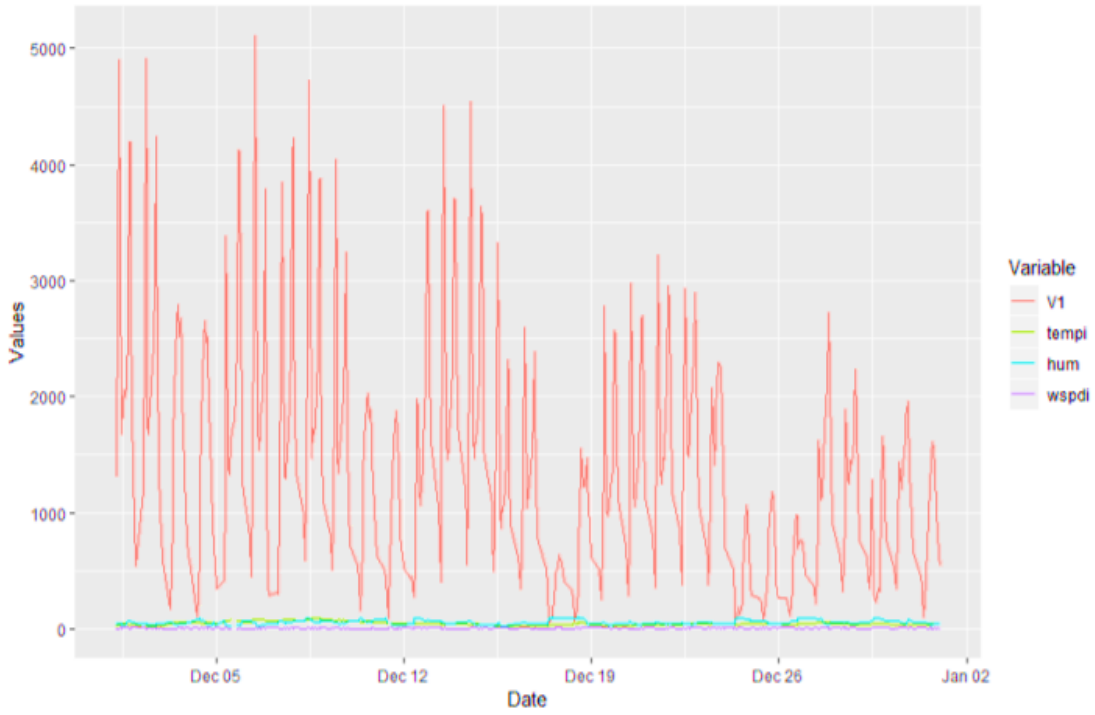



*Figure 3: Hourly Demand using Time Series for December 2016*

(In figure 3, V1 is the hourly demand)

We plotted the nature of varying hourly demand with the weather factors of temperature, humidity, and wind speed index. Here we observe that the bike demand does follow a trend. The month starts with high demand across the city however the demand pit around the holiday season. Since some pattern was visible for the hourly demand.

From our literature study, we have seen lasso regression, random forest, gradient boosting and XG Boost for predicting the hourly demand across the entire city. However, in this paper, we model this hourly demand as a time series model and we do so for individual stations across the city. This is crucial because while predicting the overall demand across the city, the prediction across all the stations is summed up and there is no way of telling which station will be busy at which time of the year and which will be not. Considering hourly demand across all the stations' shadows this individual station demand as it assumes that the trend that every station follows will be the same. Hence, further in the next session, we try to analyze the data in order to find out the busiest station in the year 2016. Then we move on to analyze which are the busiest routes in that year and are these busy stations a part of the busy routes.

## 4.2. DEMAND ANALYSIS

The demand for the bikes is not directly available within the dataset. However, we have assumed a few conditions in order to infer the hourly and the weekly demand for the bike at a given station, in this paper. This paper considers the outgoing bikes in a given unit of time as the demand or load on that station. Therefore, we infer the outflow of bikes as the demand of the station. However, same cannot be said for a bike station given the inflow, because the dataset only logs the entries of bikes being stored back in their docks, to enumerate the inflow at a given station. A person who doesn't find a place to dock his bike at a given station moves to a different station and docks his bike there, but this shadow inflow which was directed towards the 'full' station is not accounted for within the dataset. The dataset doesn't provide an estimate of available docks in an hour at the station, also the available and functional docks for individual stations might change throughout the year. Hence, it is difficult to quantify the inflow accurately. Thus, rebalancing a station as the difference between hourly outflow and inflow is not going to be pursued in this paper, instead, the fall in the predicted outflow from a station at a given hour will be considered as the appropriate time to rebalance the station if any is required.

We set "Start Station Name" in the dataset as a categorical variable and get 654 levels i.e. Different stations. The frequency of each level provides the total outflow of bikes from that particular station in that given year. This is how we decided to analyze candidates for most demanded stations.

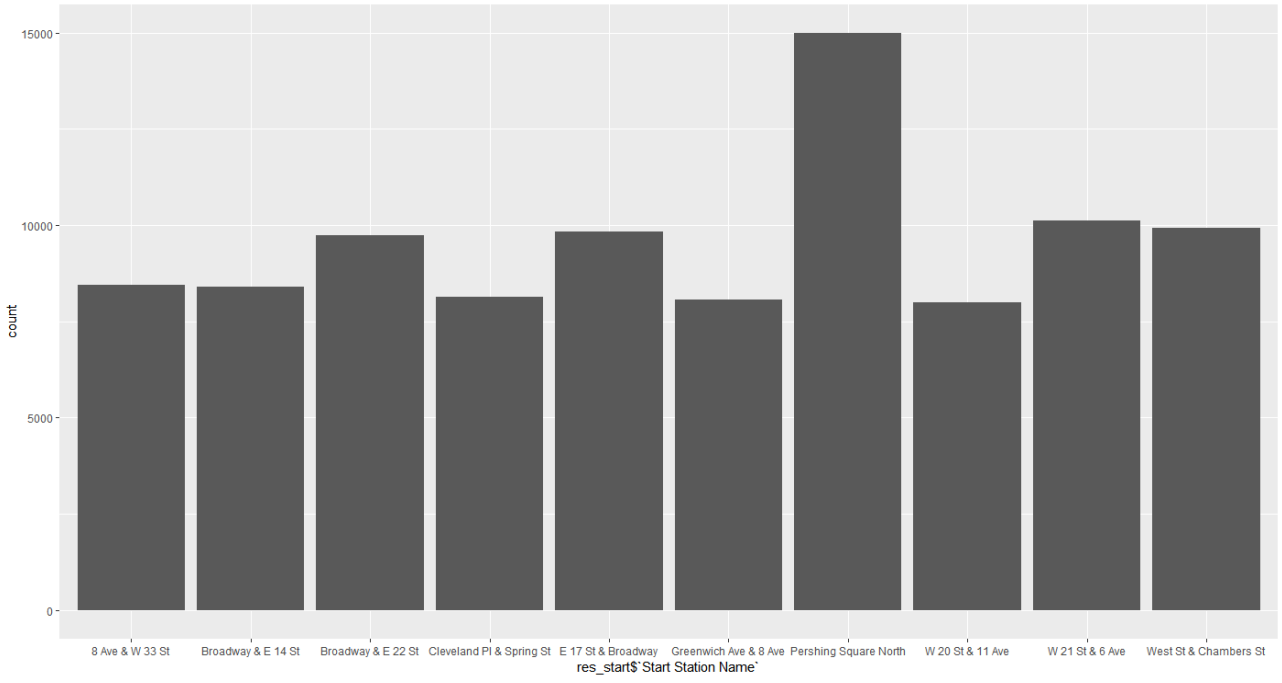



*Figure 4: Frequency of Bikes at 10 Most Popular Stations*

From demand bar graph for 2016 we can see the top 10 busiest stations and their yearly demand. Amongst these stations, Pershing Square North is significantly in demand.

Now moving on to check if these stations fall on the busiest routes as well or not. For that, demand heat map for the top 30 busiest routes is plotted. Here, 'To Station' is on the vertical axis and 'From Station' is on the horizontal axis.

In the figure 5, it can be observed that the busiest route starts from the Central Park station, however, it ends at the central park as well. Therefore, for this station it's assumed that the demand to be fairly much localized as the trips is from Central Park to Central Park, we assume these trips might be for leisure or for fitness purposes. It nullifies the overall demand of the cluster, hence we move on to the second busiest route which starts from Pershing Square North and the third busiest route from W 21 St & 6$^{th}$ Ave. Other stations which are part of the top ten busiest stations are not a part of where the top 30 busiest routes start from. Hence we continue further analysis with these two stations.

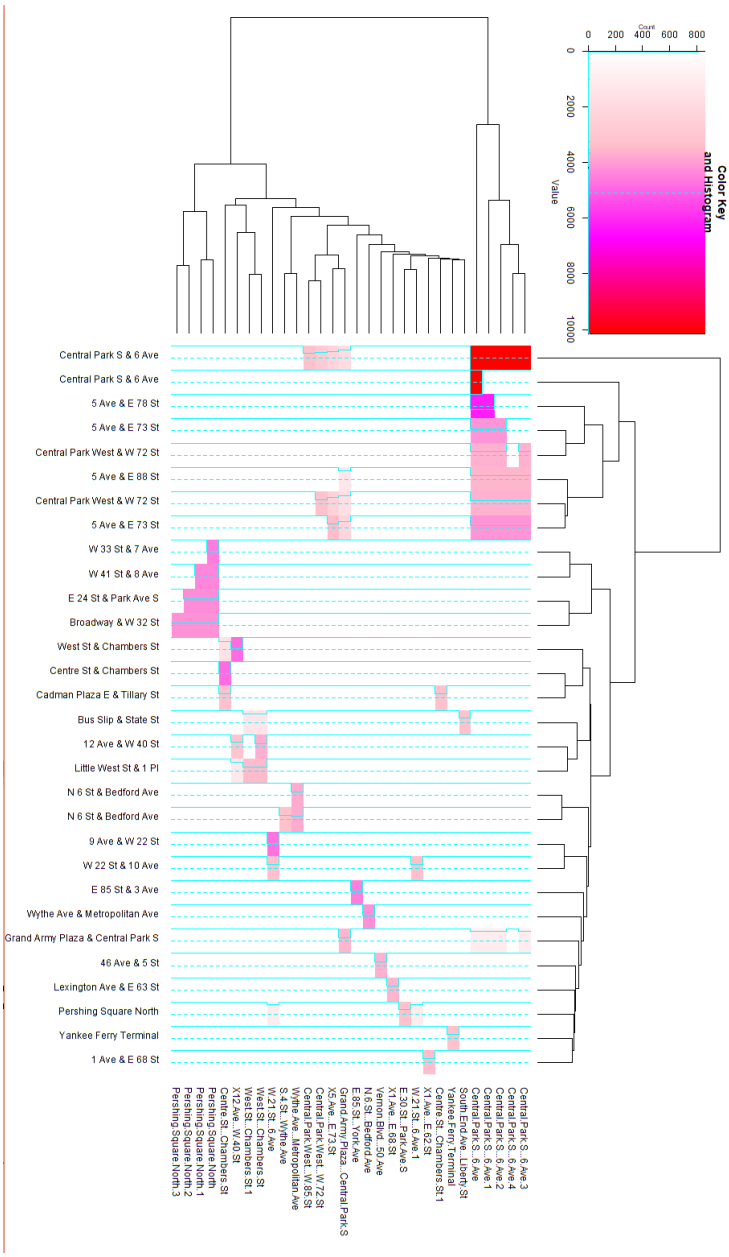
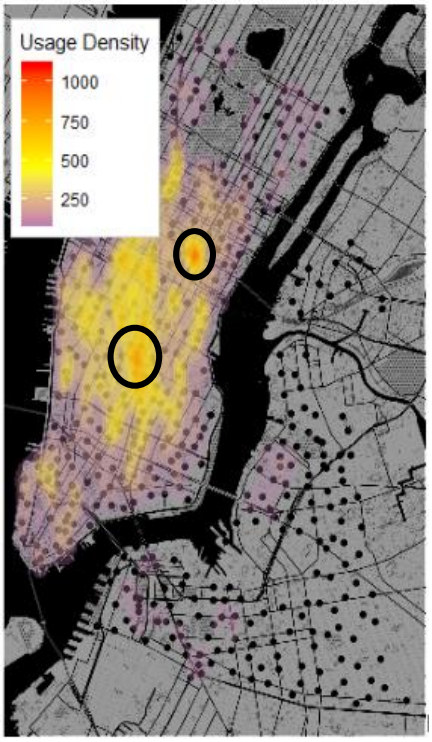
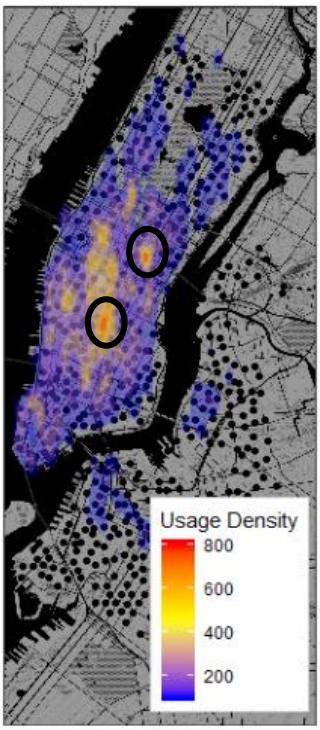



*Figure 5: Correlation Matrix of the Busiest 15 Routes*

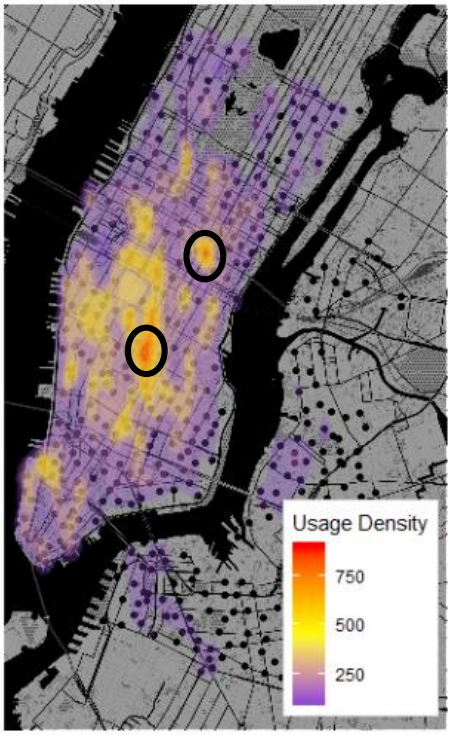
Now, combining the two analysis, we mark these stations on the map during different times of the year. A heat map is plotted considering every week's demand at the dock stations. Among all the 52 weeks' plots', we have shown here plots from each trimester of the year.



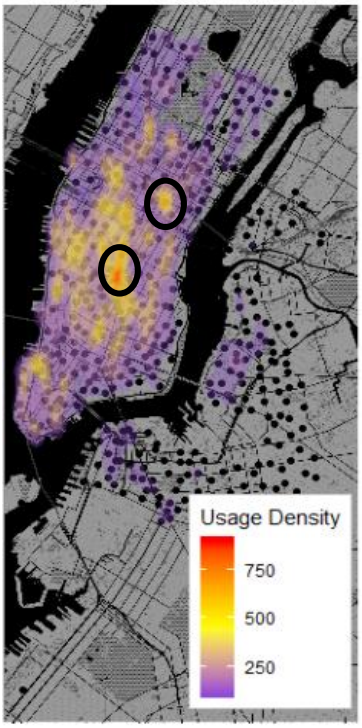

February




September




May




November

*Figure 6: Heat Map of the NYC Citi Bike Demand*

(Circles represent Pershing Square North Station (top circle) & W 21 St & 6 Ave Station (bottom circle))

## 4.3. TIME SERIES MODELING

### 4.3.1 Multi-Layer Perceptron model MLP

For the busiest individual station, Pershing Square North, we observed the hourly demand. We plot the ACF for Pershing Square North.

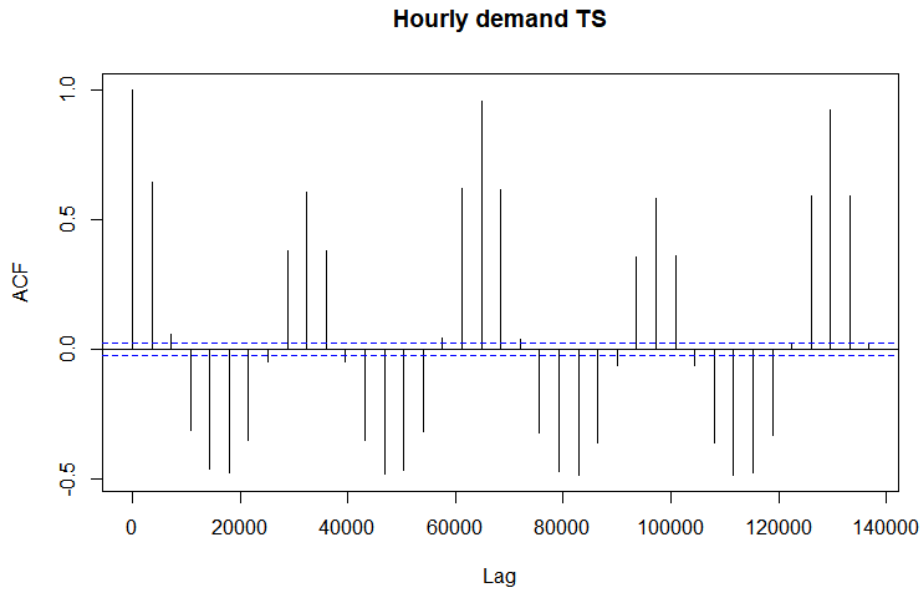



*Figure 7: ACF of Pershing Square North using Time Series*

In this ACF we can see a seasonality and therefore in order to model this we considered Multi-Layer Perceptron model. MLP has the capability to learn non-linear models and to learn models in real-time. We consider this model due to its self-learning neural network model. We used TSTools library and model the fit in the MLP function. We let the function automatically decide the number of hidden layers and detect the seasonality by learning. After fitting the model, we look at the fitting parameters

*Table 3: Details of Hidden Layer*

| | |
|---|---|
| Hidden Layers | 5 |
| MSE | 2261.77 |
| Lags | 4 |

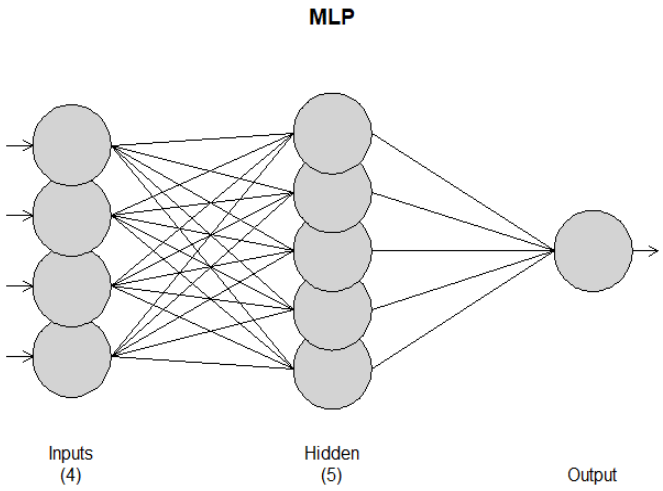



*Figure 8: Architecture of the network*

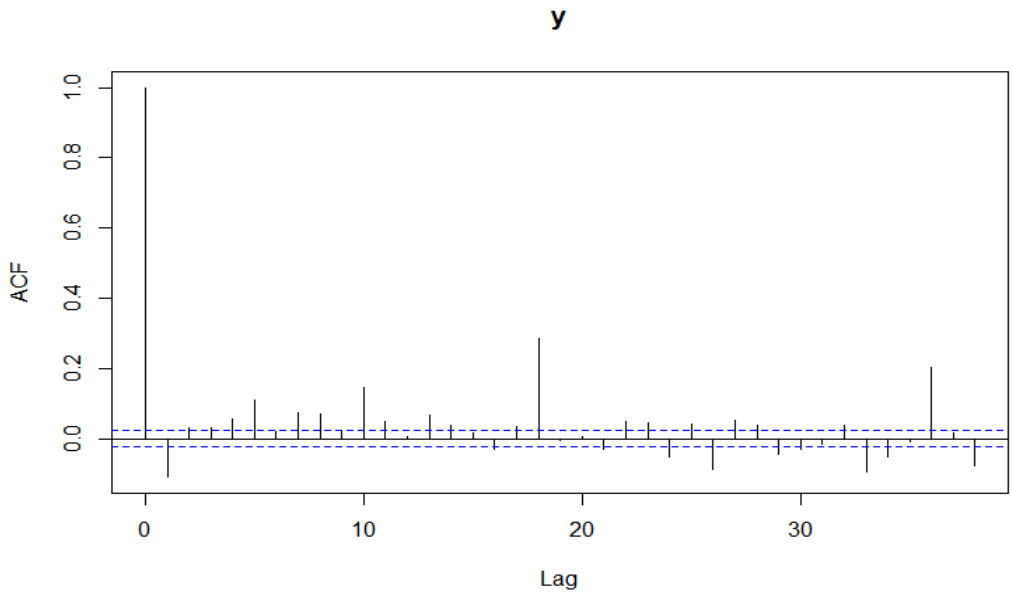



*Figure 9: ACF of MLP for Pershing Square North*

```
Box-Ljung test

data:  pred$residuals
X-squared = 81.067, df = 1, p-value < 2.2e-16
```

From the result obtained, we can say that it is very significant but fails Box-Ljung test.

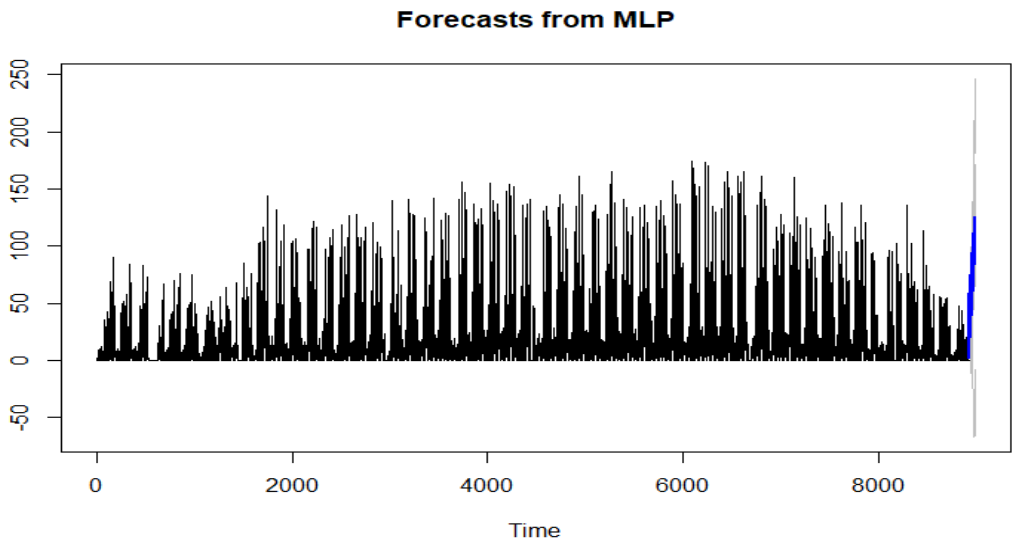



*Figure 10: Forecasts using MLP*

To compare our prediction, we plotted the actual values vs Forecasted values in the figure 10. The Black line represents the actual data for the 400 hours for which the value is forecasted. The blue line represents the forecasted values.

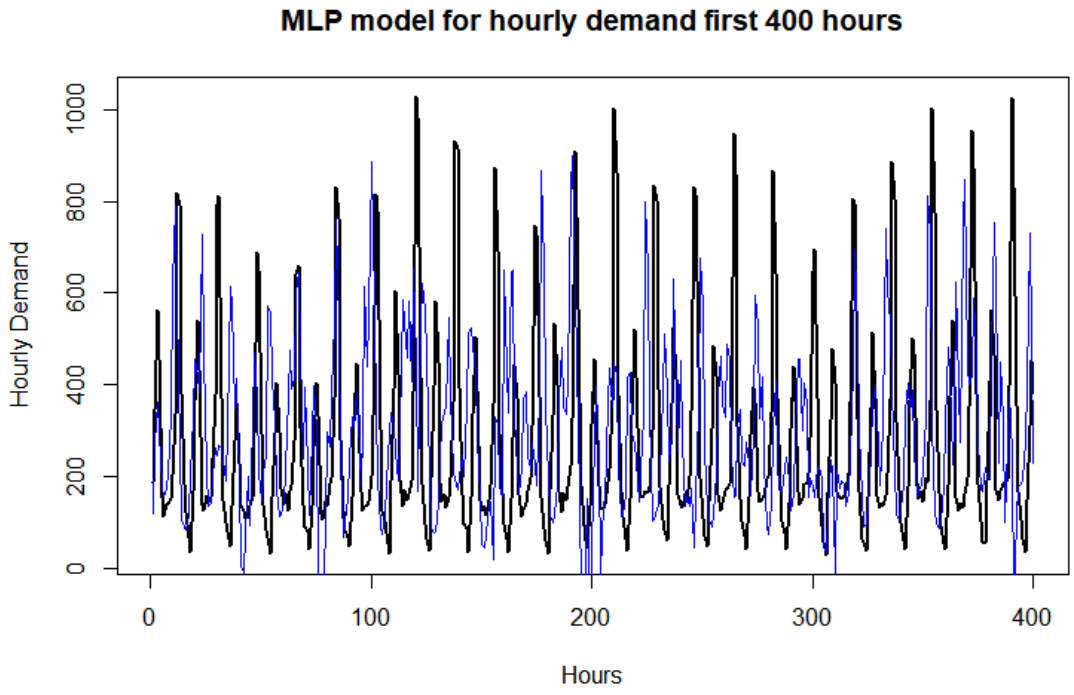



*Figure 11: Comparing Forecast vs Actual MLP model for hourly demand of first 400 hours*

Predicted Actual

It is unclear from the graph visualization how much the forecasted value differs from the actual value. Therefore, we calculated the error between them which gave the following result.

| | ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
|---|---|---|---|---|---|---|---|
| Training set | 1.563133e-04 | 47.55836 | 34.10382 | 0.969366 | 27.19956 | 0.2736108 | -0.1100391 |
| Test set | -3.998617e+01 | 240.87192 | 179.31933 | -97.036443 | 131.31761 | 1.4392186 | NA |

Looking at the MAPE value (131.31) of the test set, we can see that the accuracy of the model is very poor. So, we consider this model to have failed in the first year itself and thus, to increase our accuracy, we modeled other time series known as TBATS.

### 4.3.2. TBATS

Introduced by De Livera et al (2011), TBATS is an acronym that stands for:

T for Trigonometric regressors to model multiple-seasonalities
B for Box-Cox transformations
A for ARMA errors
T for Trend
S for Seasonality

TBATS uses State Space model that generalizes Exponential Smoothening and automates Box-Cox Transformation and ARMA errors. The TBATS model is helpful when there are seasonal changes over time. As we have seen in the ACF plot of the hourly demand TS, we observed a seasonal index over time. So, we applied the TBATS model for hourly demand. It gives two forecast prediction region for 95% and 80% accuracy. As seen below in the ACF for TBATS hourly demand, we observe that our model is not a white noise.

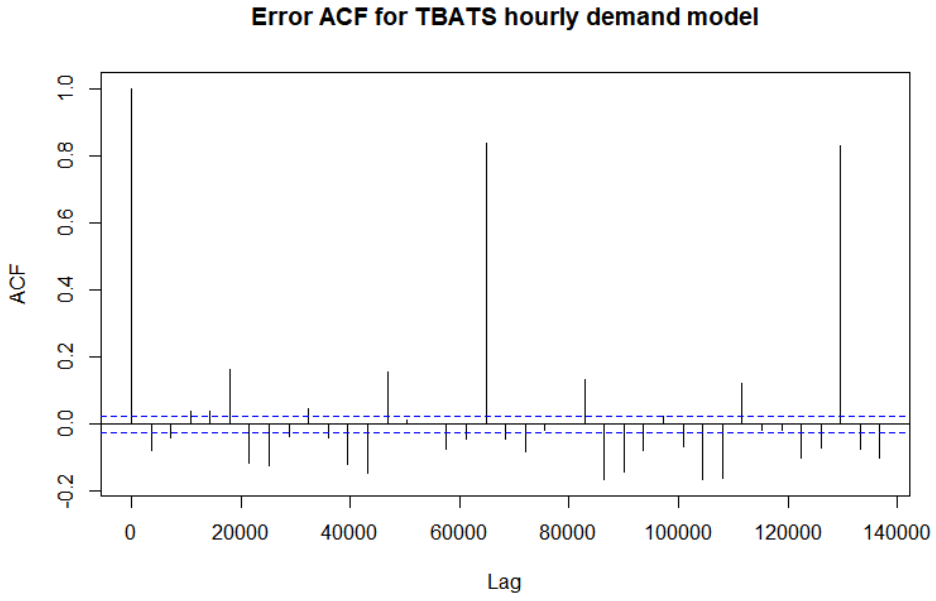



*Figure 12: Error ACF for TBATS Hourly Demand Model*

```
        Box-Ljung test

data:  fith1$errors
X-squared = 40.465, df = 1, p-value = 2.001e-10
```

As the p-value is very small, less than significant value 0.05, it fails the Box-Ljung test.

Therefore, it seems that it is difficult to predict the hourly demand for bikes and thus we predict weekly demand. It is also more feasible to restore bikes at a stand in a given week rather than every hour. We plot the weekly demand for our two busiest stations in the figure 13.

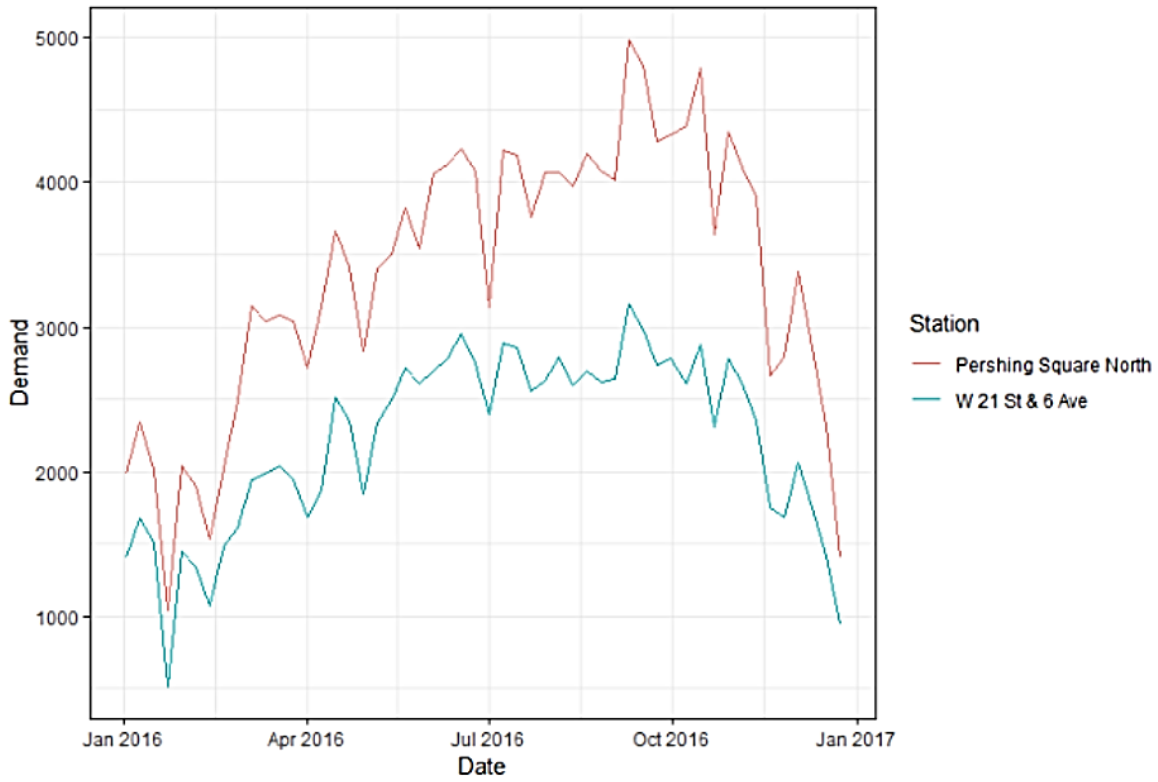



*Figure 13: Weekly Demand of 2 Busiest Stations*

Here we can see a trend which is similar between the two stations. We model this time series using **TBATS** and the seasonal periods are set for (1, 52) in "TBATS" function that means "observed every week for 52-week period".

We fit the model for our busiest station, Pershing Square North for 2015 data to predict the demand for the year 2016. Then, the fit is tested by calculating the ACF error.

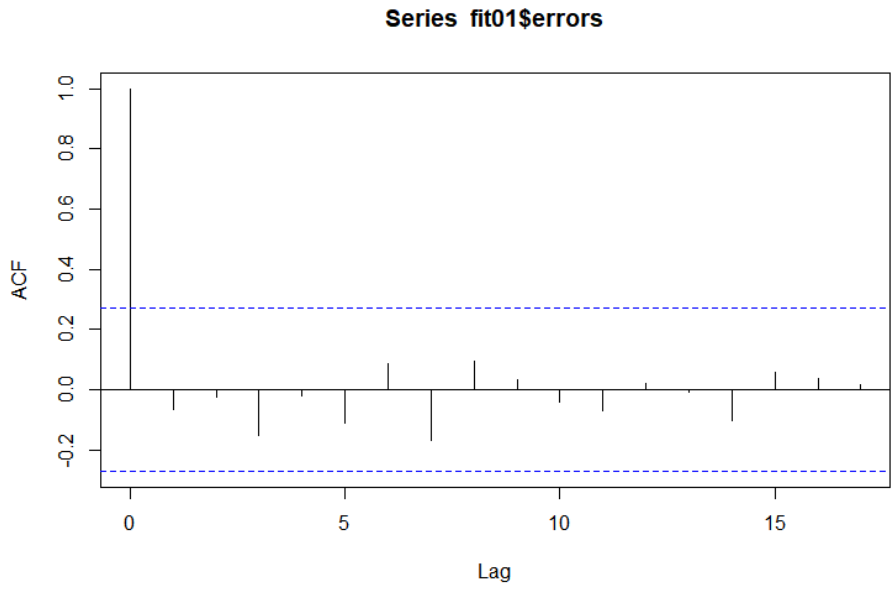



*Figure 14: ACF of Pershing Square North for 2015*

From figure 14 above, we observe that as ACF=1 at lag=0, it is a white noise.

```
        Box-Ljung test

data:  fit01$errors
X-squared = 0.23477, df = 1, p-value = 0.628
```

Therefore, the p-value is greater than the significant value (0.05) which means it clears the Box-Ljung test. Thus the model can be predicted for the next year (2016).
We plot the forecasted values in the figure below and then compare the forecasted values with the actual values for the year 2016.

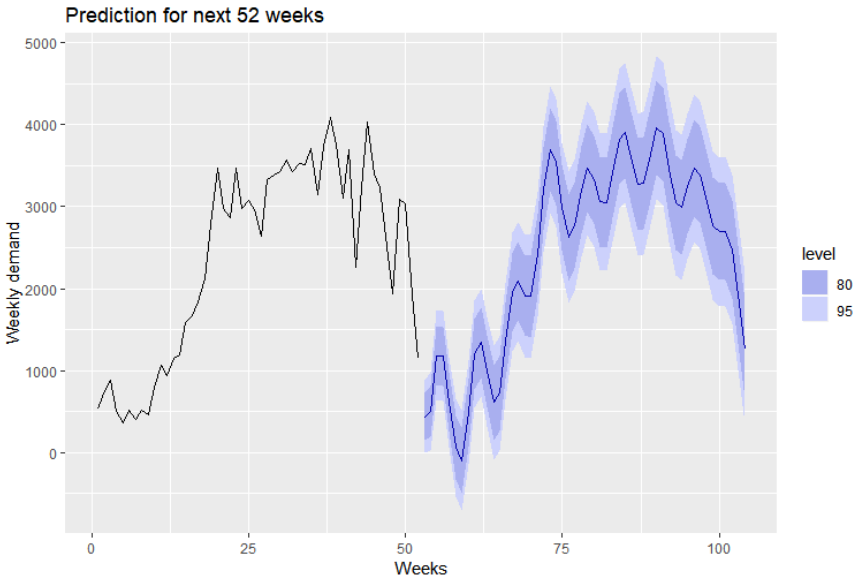



*Figure 15: Prediction for next 52 weeks of Pershing Square North for 2017*

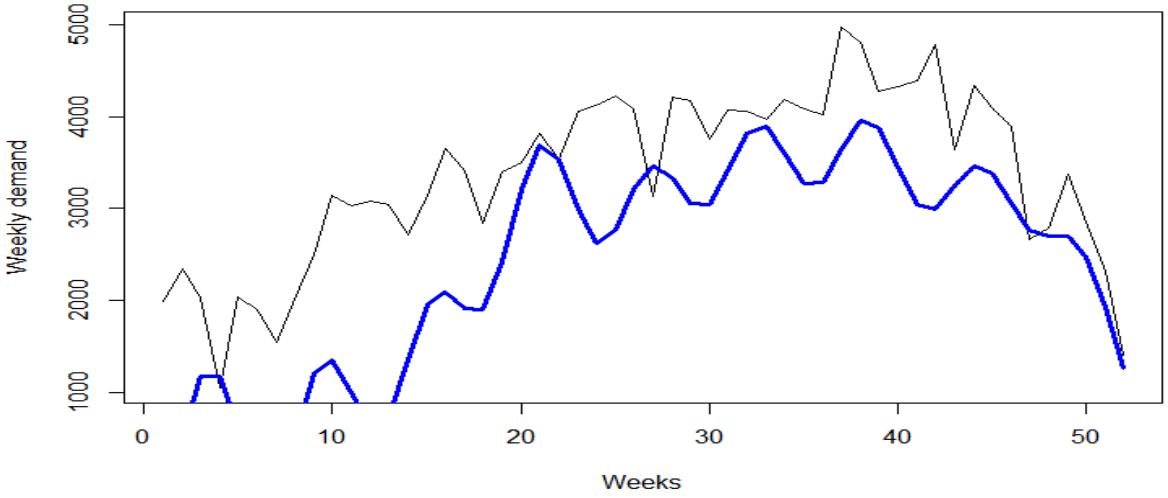



*Figure 16: Comparing Predicted vs Actual Demand of Pershing Square North for 2017*

Predicted Actual

| | ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 | Theil's U |
|---|---|---|---|---|---|---|---|---|
| Training set | 45.62033 | 224.1893 | 171.2715 | 2.15336 | 12.6873 | 0.445883 | -0.0652995 | NA |
| Test set | 958.31078 | 1158.4928 | 979.9270 | 31.51699 | 32.5483 | 2.551112 | 0.5765792 | 2.094935 |

We observe that the MAPE even though being lower than the one achieved in MLP, it is still significant. We intend to reduce this error and therefore make use of more data. So, we incorporate

the 2015 and 2016's actual data to predicted weekly demand, for the Pershing Square North Station for the year 2017.

The Model fit is as below:

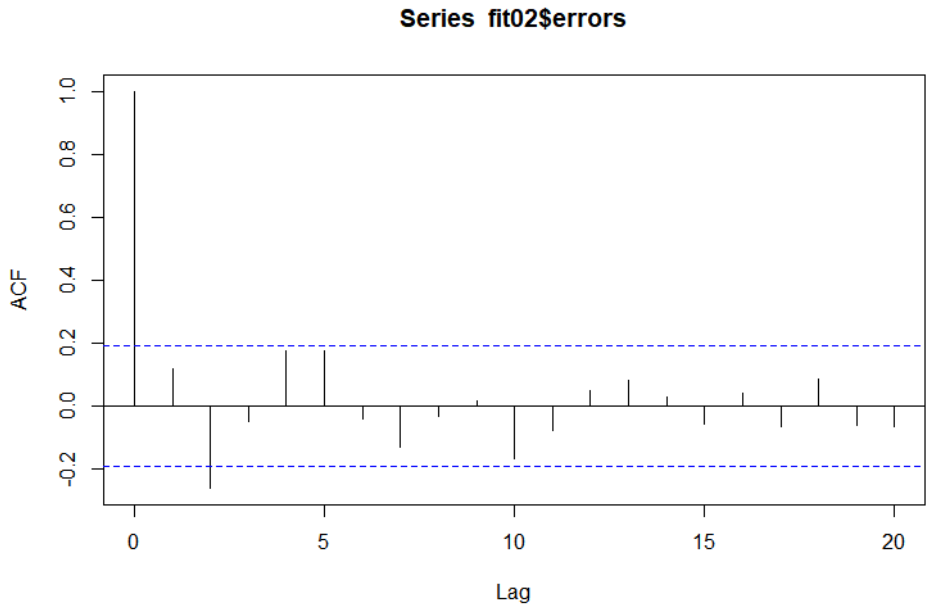



*Figure 17: ACF of Pershing Square North for 2017*

```
        Box-Ljung test

data:  fit02$errors
X-squared = 1.4427, df = 1, p-value = 0.2297
```

Here also, we can see the Box-Ljung test is cleared as the p-value is greater than significant value 0.05.

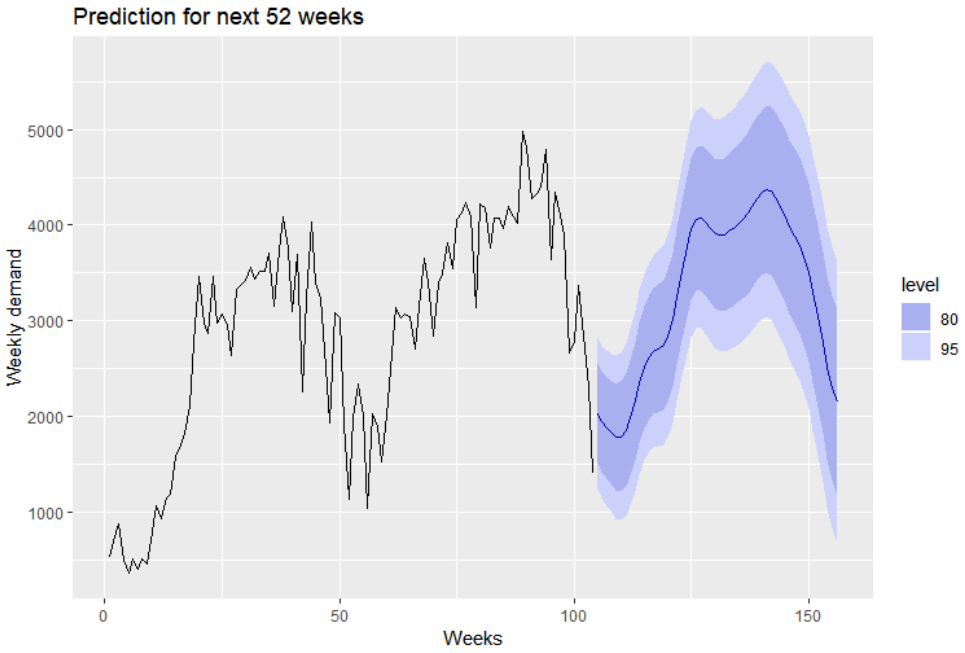



*Figure 18: TBATS Prediction of Pershing Square North for 2017*

Then the model for 2017 is predicted from the years 2015-16 and it fitted quite well.

We compared the predicted values with the actual values for the year 2017 and checked for errors.

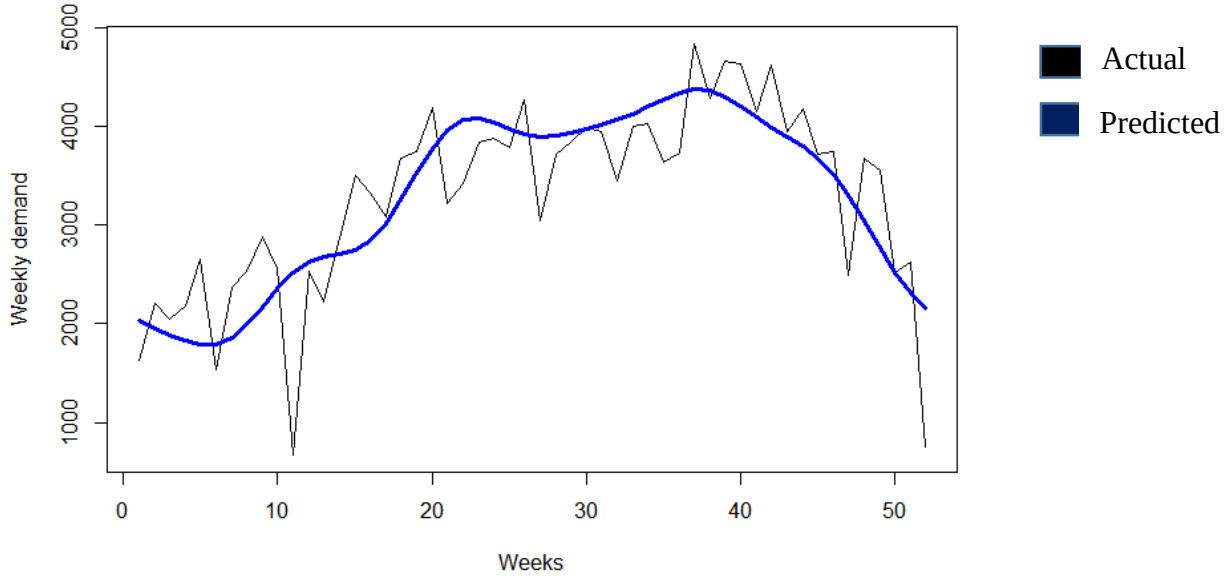



*Figure 19: Comparing Predicted vs actual plot of Pershing Square North for 2017*

| | ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
|---|---|---|---|---|---|---|---|
| Training set | 55.677780 | 404.8041 | 318.0360 | -1.766756 | 14.47038 | 0.7939724 | 0.1161032 |
| Test set | 3.924707 | 540.0779 | 414.2465 | -7.300841 | 20.15572 | 1.0341605 | NA |

The model gave a good MAPE with 80% accuracy and hence we proceed with this TBATS model and predict for the next busiest station, (W 21 St & 6 Av) in 2017 using 2015-16 data.

On receiving satisfactory result for station 1, we performed the same analysis for Station 2 i.e. W 21 St & 6 Ave.

```
        Box-Ljung test

data:  fit11$errors
X-squared = 0.00065762, df = 1, p-value = 0.9795
```

Since the p-value is quite significant, the Box-Ljung test is satisfied.

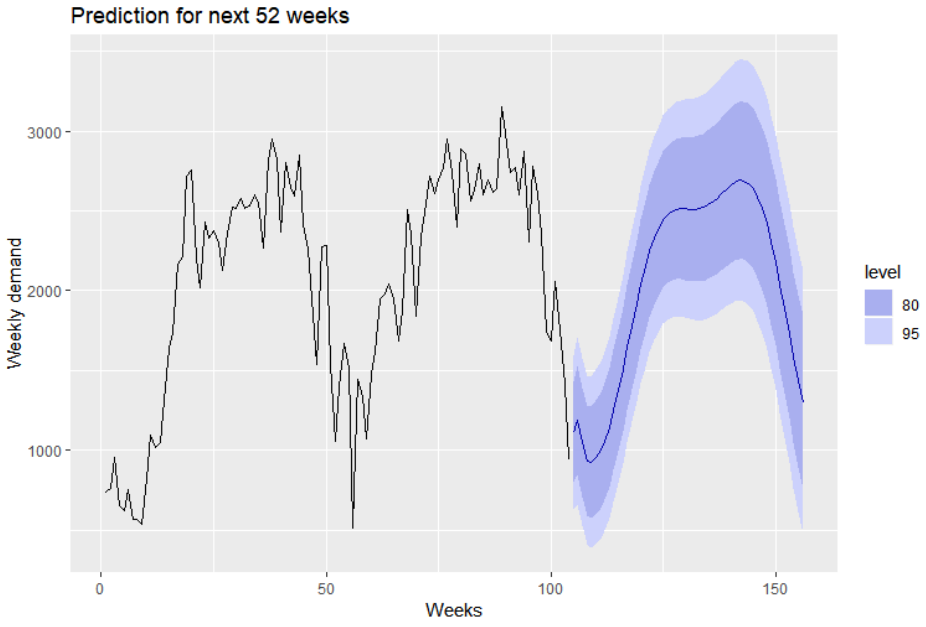



*Figure 20: TBATS prediction of W 21 St & 6 Ave for 2017*

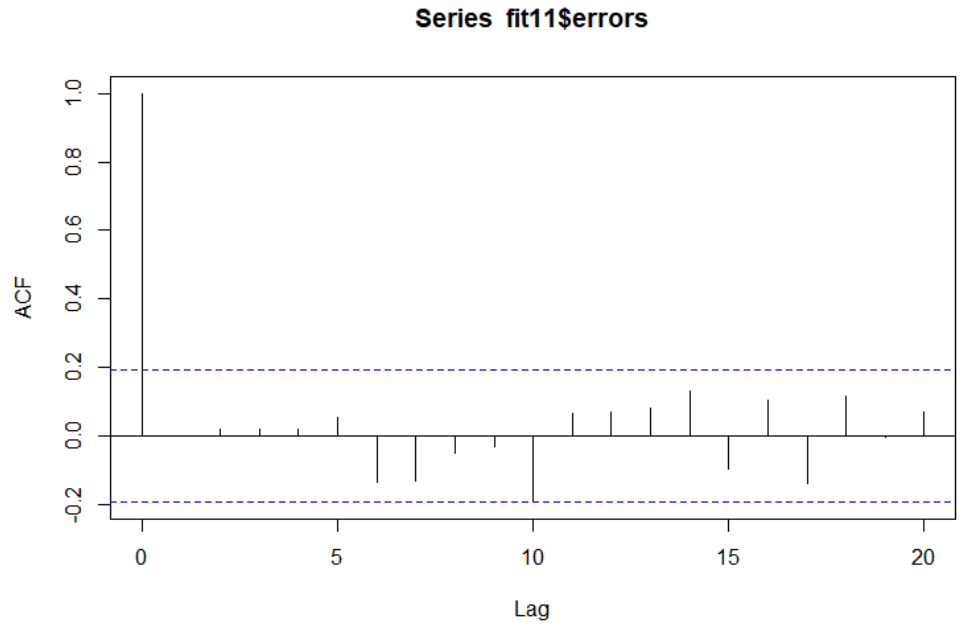



*Figure 21: ACF of W 21 St & 6 Ave for 2017*

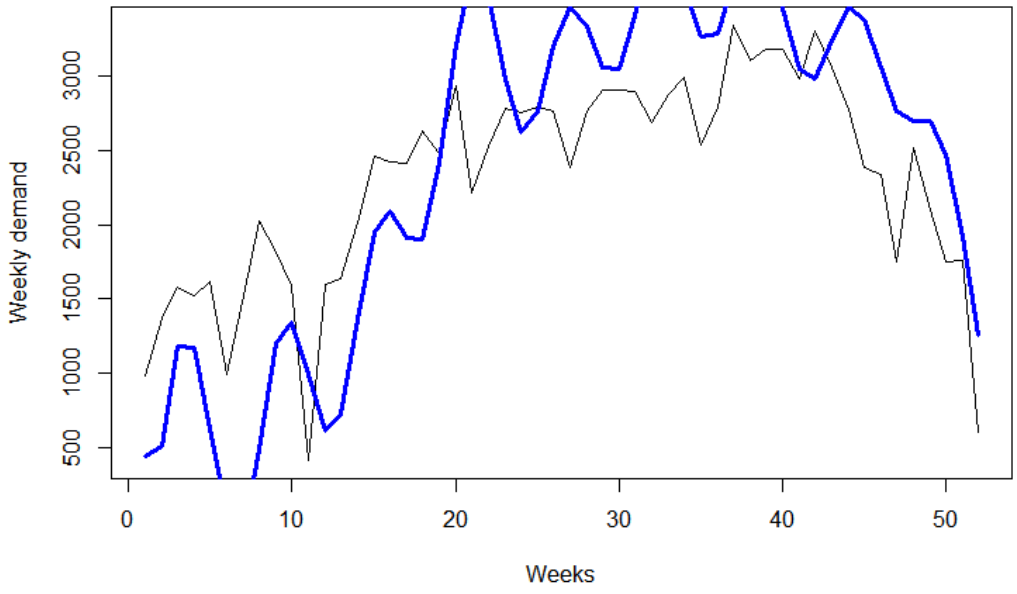



*Figure 22: Comparing Predicted vs Actual demand of W 21 St & 6 Ave for 2017*

Predicted Actual

| | ME | RMSE | MAE | MPE | MAPE | MASE | ACF1 |
|---|---|---|---|---|---|---|---|
| Training set | 14.19029 | 245.8324 | 191.051 | -1.879671 | 11.58603 | 0.7718477 | 0.002478764 |
| Test set | 266.42639 | 427.7336 | 364.036 | 5.929966 | 21.35327 | 1.4707084 | NA |

Thus for station 2 as well, the MAPE is significant with an accuracy of approximately 79%.

## 5. RESULT

From the analysis done, the results achieved are summarized in the table below:

*Table 4: Comparison of Modeling Methods*

| **Station** | **Data** | **Model** | **Model Fit** | **Prediction Accuracy (MAPE)** |
|---|---|---|---|---|
| Pershing Square North | Hourly Demand | Linear Regression | Bad with R sq. value 0.156 | - |
| Pershing Square North | Hourly Demand | Multi-Layer Perceptron (MLP) | Bad | - |
| Pershing Square North | Hourly Demand | TBATS | Bad | - |
| Pershing Square North | Weekly Demand | TBATS | Good | 80% |
| W 21 St & 6 Ave | Weekly Demand | TBATS | Good | 79% |

It was clearly observed from the analysis that TBATS provides a better fit and a good accuracy of 80% prediction.

## 6. CONCLUSION

While having a linear regression model, the R square value was very low which made our model not a good fit. Although the weather was significant for our sample data, due to lack of values in weather data for the whole year it was not possible to fit a time series model for the whole year with the weather as a factor. From the number of counts from demand and the relation between the stations, it was possible to find the busiest route between the two stations and we visualized it with a heat map. To predict the hourly demand for Pershing Square North station of the year 2016, multiple layer perceptron (MLP) time series was used which found to be a bad model as the MAPE error was very high. Therefore, to increase the accuracy TBATS time series was used wherein hourly demand for Pershing Square North station was calculated but resulted in the failure of the Box -Ljung test. The test cleared for weakly demand which meant it is feasible to restore bikes at a station weekly rather than hourly. Therefore, the demand for 2017 was forecasted by using 2015-16 data for the two busiest stations, Pershing Square North station and W21 Street & 6 Avenue,

with TBATS. We found that TBATS was 80% and 79% accurate for forecasting the weekly demand for those stations respectively. With the help of this model, CITI Bike share will be able to restore the docks by predicting the future demand in these two stations. Similarly, the demand can be forecasted for the rest of the stations.

## 7. FUTURE SCOPE

For better forecasting accuracy, the time can be categorized into different time periods such as weekdays and weekends, daytime and nighttime, and the inclusion of factors such as holidays or special events.

To increase the accuracy of the model, a larger data such as for 4-5 years can be taken. This would take into account multiple seasonalities and possible cyclicities. Such huge data can be handled by using machine learning techniques like XG boost or Random forest. The MLP model would also work better with a larger dataset.

## 8. REFERENCES


1. Mitesh Gadgil, Saurabh Kulkarni, Akshatha Gangadhariah, Anshul Jain, "Demand Forecasting on Bay Area BikeShare"
2. Jayant Malani, Neha Sinha, Nivedita Prasad, Vikas Lokesh "Forecasting Bike Sharing Demand"
3. Xiaomei Xu, Zhirui Ye, Jin Li, Mingtao Xu, "Understanding the Usage Patterns of Bicycle-Sharing Systems to Predict Users' Demand: A Case Study in Wenzhou, China", Computational Intelligence and Neuroscience; Volume 2018, Article ID 9892134
4. Leonardo Caggiani, Michele Ottomanelli, "A dynamic simulation based model for optima", SIDT Scientific Seminar 2012, Politecnico di Bari, via Orabona 4, Bari, 70125 Italy
5. Wen Wang, "Forecasting Bike Rental Demand Using New York Citi Bike Data", Dublin Institute of Technology
6. Thomas Nosal, Luis F. Miranda-Moreno, "The effect of weather on the use of North American bicycle facilities: A multi-city analysis using automatic counts", August 2014
7. Zhang Jiawei, Pan Xiao et al, "Bicycle-Sharing System Analysis and Trip Prediction IEEE MDM Conf. 2016", April 3
8. Hampshire Robert, Marla Lavanya, "An Analysis of Bike Sharing Usage: Explaining Trip Generation and Attraction from Observed Demand"
9. Dias Gabriel Martins, Bellalta Boris and Oechsner Simon, "Predicting Occupancy Trends in BarcelonaaAZs Bicycle Service Stations Using Open Data"
10. http://www.nyc.gov/html/dot/html/bicyclists/bicyclists.shtml
11. https://www.citibikenyc.com
12. https://www.kaggle.com/mathijs/weather-data-in-new-york-city-2016