TIME SERIES ANALYSIS OF NEW YORK CITY YELLOW TAXI DATA FROM 2009-2015

Table of Contents

Introduction

In this age of digital technology where all transactions are stored in computers, there is more data available than ever before. With the advances made in computer technology, fast processing and large, cheap storage, it is getting increasingly more possible to process large amounts of data to gain insights.

Since companies use data to make decisions, it is logical for them to look at all these gigabytes of data as their next step in gauging customer behaviors to make more precise decisions on how to serve customers better.

In this project, I looked at NYC Taxi data to come up with a prediction for the year 2016. I started with doing some exploratory data analysis of the data to come up with some hidden insights in gigabytes of data. I plotted time series graph of daily taxi data from 2009-2015 and then also compared the pick-up and drop off location in New York City.

Then, I did time series forecasting by using appropriate model for the dataset. The only limitation of this was that the resources available to me was not enough to go into detailed analysis of the data. Given an opportunity, I would like to divulge deeper into this dataset to come up with even more finding such as which location/ area will have how many pick-ups on a given day

Data Collection

The New York City Taxi & Limousine Commission has taxi record available from 2009 to 2019. The taxi data is further divided into Yellow Taxi, Green Taxi and For Hire Vehicle. For this project, I decided to use the data of Yellow Taxi from 2009 to 2015. The entire dataset, comprising of just Yellow Taxi Data, from 2009 to 2015 have 1.1 billion rows. The dataset for further divided into months. I downloaded the monthly data from 2009 to 2015. The total size of the dataset from 2009-2015 was 173GB.

Each of the data file has the following columns

- Vendor-ID
- Trip Pickup Date-Time
- Trip Dropoff Date-Time
- Passenger Count
- Trip Distance
- Start Longitude
- Start Latitude
- End Longitude
- End Latitude
- Payment Type
- Total Amount (Fare Amount + Extra + MTA Tax + Toll Amount + Tip Amount)

Since our end goal was to do time series analysis of data and predict how many yellow taxi trips would occur in the year 2016, I decided to subset the data by Trip Pickup Date-Time. I got a count of trip completed based on Trip Pickup Date-Time for each day from 2009 to 2015 and saved it in a .csv file. This helped me in converting 173GB of data to just 41KB, as I just used the columns that I needed.

For plotting the graph of Trip Pick Up and Trip Drop Off location, I converted all the .csv file with its original columns (84 files/173gb) to. hdf5 file format and then merged all these. hdf5 files into one. This helped me plot the Yellow Taxi Drop Off and Pick up location on map.

Then, I wanted to cluster the areas with most yellow taxi trip pickup. To do this, I used H3 library, which is Uber's Hexagonal Hierarchical Spatial Index. By using it, I could easily cluster geospatial observations, and display them on a map. I did the clustering using just one month of data as doing it on the whole dataset was computationally impossible taking into consideration the fact that Pandas library is not useful when dealing with big data.

Data Analysis

At starting, I ran descriptive statistics to get a sense of average number of passenger present in each taxi trip and the average cost of a trip.

	Passenger Count	Trip Distance	Total Amount
Mean	1.67	14.68	16.06

Table 1

As we see from table 1, the average passenger count is 1.67. We can say that most of the yellow taxi trips have 2 passengers.

People taking a taxi tends to travel a relatively long distance. It could be that, from people who are using yellow taxi, most of them use yellow taxi to go to work from home or back home from work. Also, the average cost of yellow taxi trip is $16.06.

Now, Let's look at the trip year over year from 2009 to 2015.

Figure. 1

From figure 1, we see that there is a dip in yellow taxi trop from May to late August. The reason of this could be that a lot of college going students, from different part of the world, went back to home. Now, we look at the plot over time from 2009 to 2015.

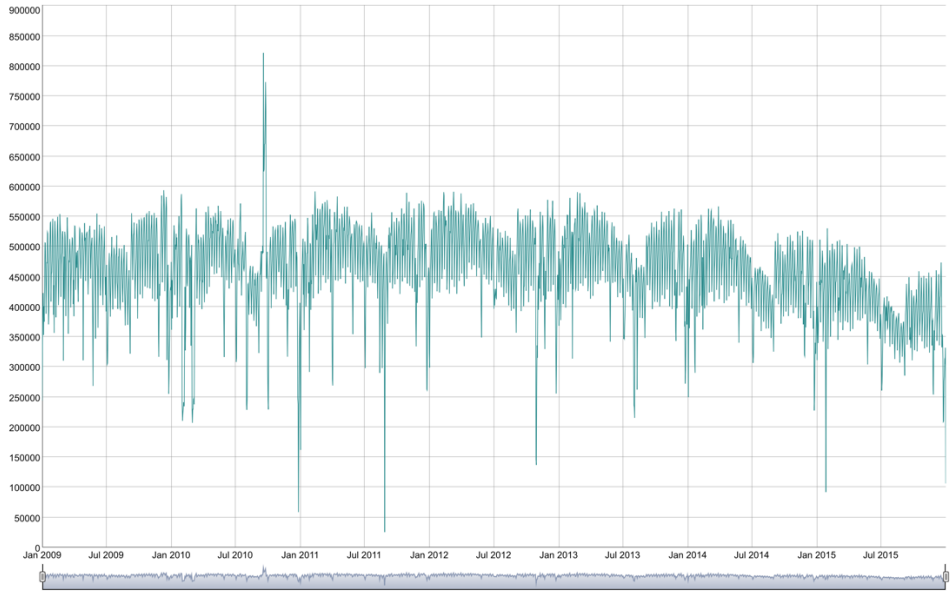


Figure 2

The plot (figure 2) shows the taxi trips from January 2009 to December 2015. We see that there is some outlier in the data. One of the interesting findings from this was the amount of yellow taxi trip on September 19, 2010. All the other outliers present in the graph are, according to my understanding, wrong entry of data, as these entries have one less zero at the end.

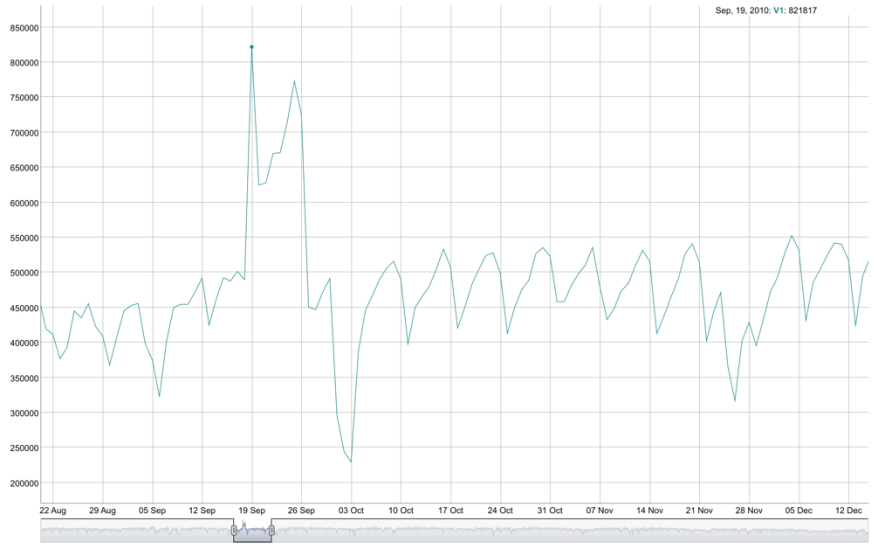


Figure 3

We see that there is a steep rise in taxi trip on 19th September (Sunday) 2010 (Figure 3). By doing further analysis, I found that there was a tornado in New York City on 16th September (Thursday) 2010. It was named “2010 Brooklyn/Queens Tornadoes”. I believe, this is the reason why there was a steep rise in yellow taxi trip on 19th of September.

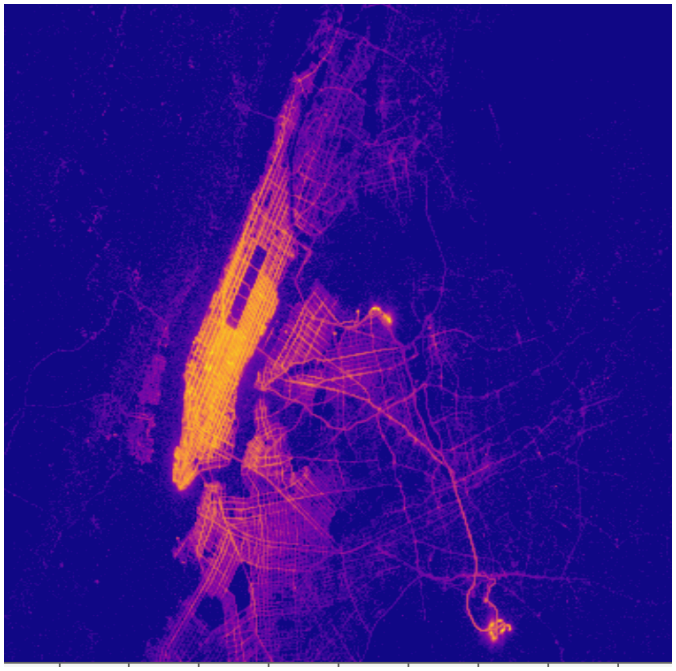


Figure 4 Trip Start

Figure 4 shows Trip Starting point ins New York City. The bright pink shows that there is more trip starting in that part of city. We see that most of the Yellow Taxi Pick Up are from Manhattan Area of New York City

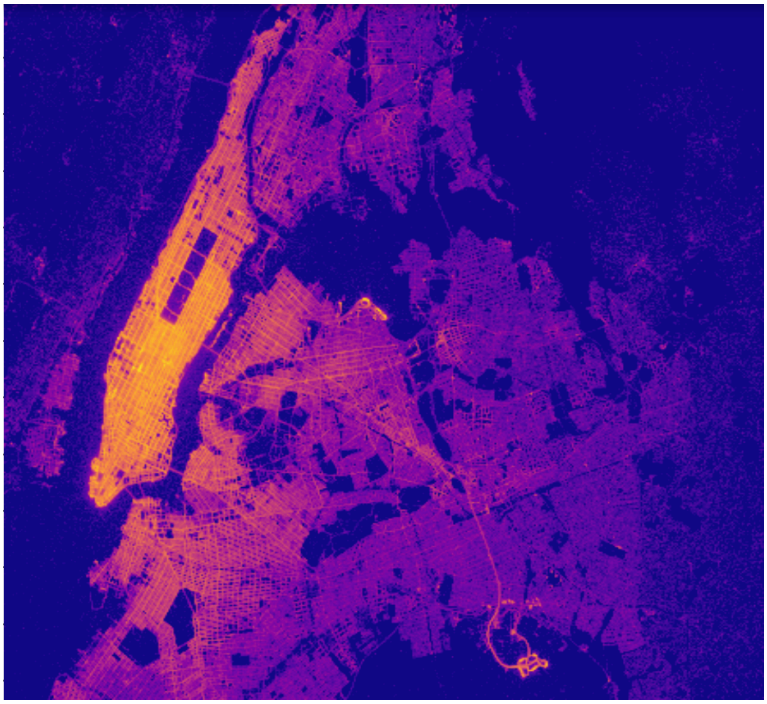
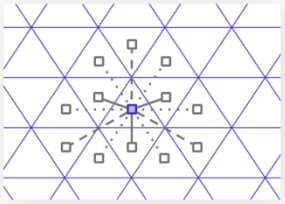
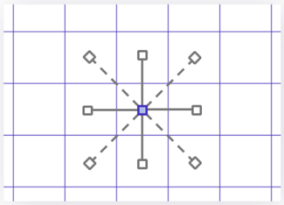
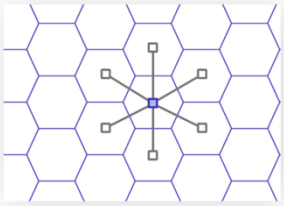


Figure 5 Trip End

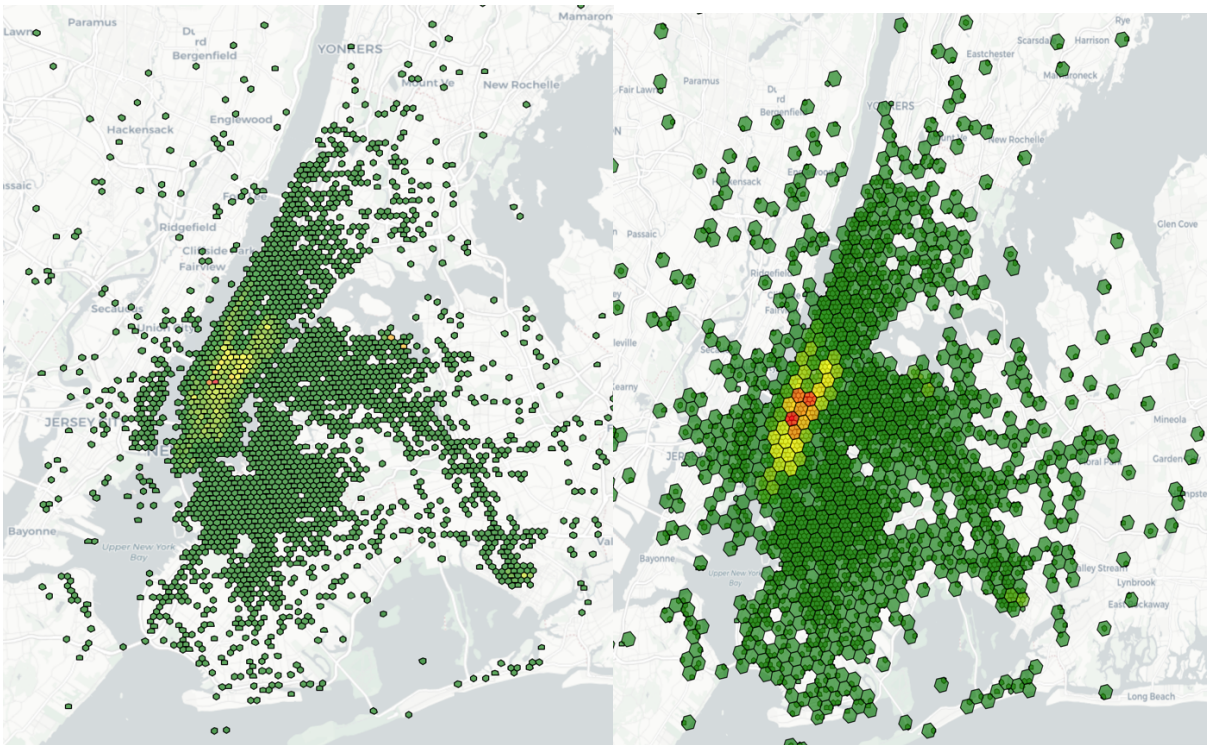
As we can see from figure 5, the drop off location is throughout the city. More Importantly, we see a lot more drop off in Queens area than compared to pick up in the very same area. It strengthens our assumption that a lot of people use Yellow Taxi to commute from work to home.

Next, lets' look at the clusters of area that had the most pickup and drop off

The H3 package from Uber is Unique in a way that it uses a hexagonal grid system. The advantage of hexagonal grid system is that, from a mid-point, the distance to all the edges are going to be the same, unlike square and triangle.

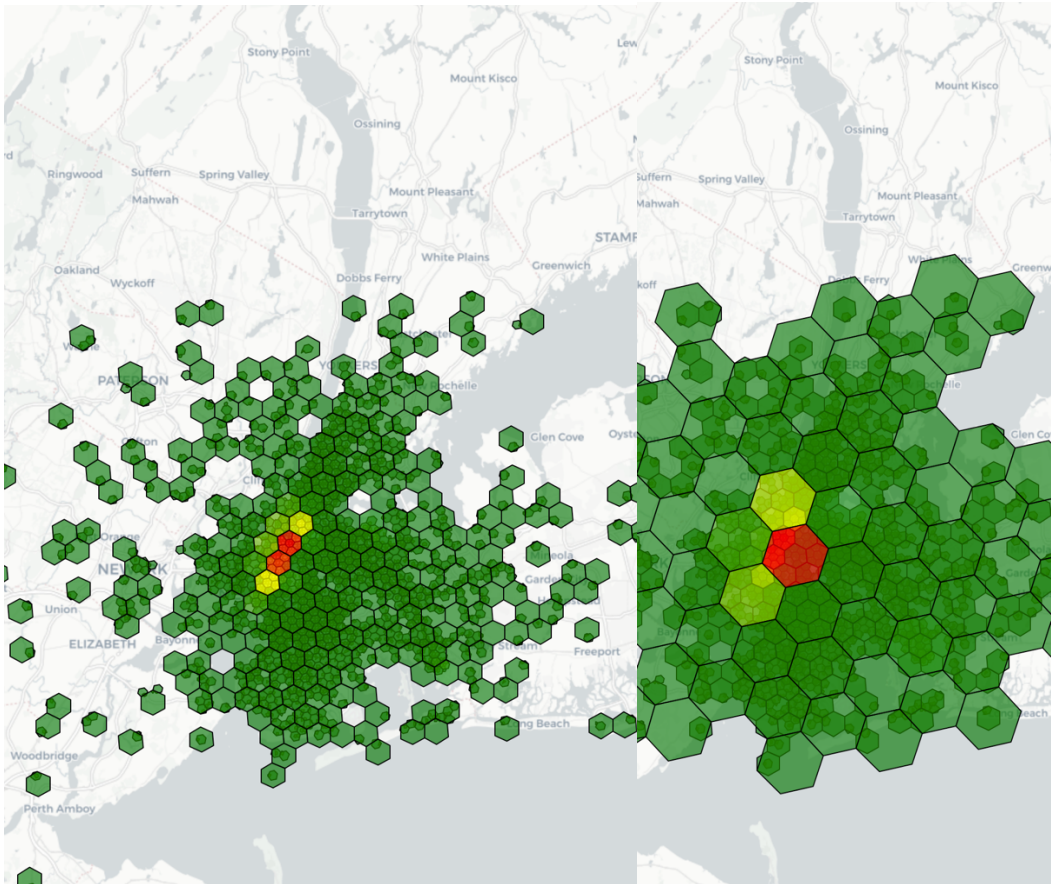




If the grid is all triangle, the distance between all the grid are going to be different. The surprising thing is that, the distance between all the square grids from mid-point is also different. But if you consider a hexagon, the distance between other hexagonal edges are the same. To analyze the pick-up point using hexagonal grid, I made a map, which is shown below.

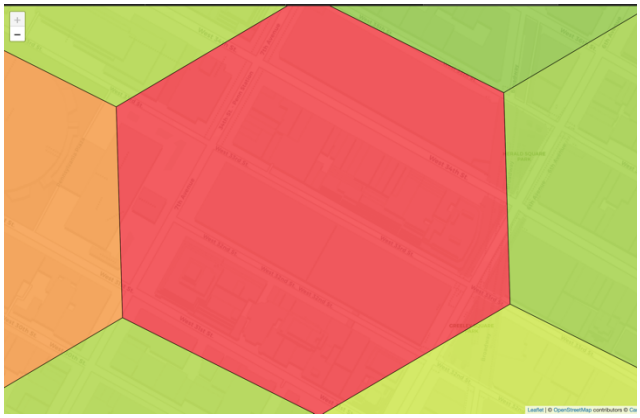


Figure

We can zoom out from street level to far off to see which area has the most pick-up based on the level that we want to see the pickup.



Figure



Figure

Let us now look at which area has the greatest number of Pick-up.
It looks like most of the pick-ups where from West 31st,32nd,33rd and 34th Streets in New York City, in the month of April 2015.

Discussion of Results

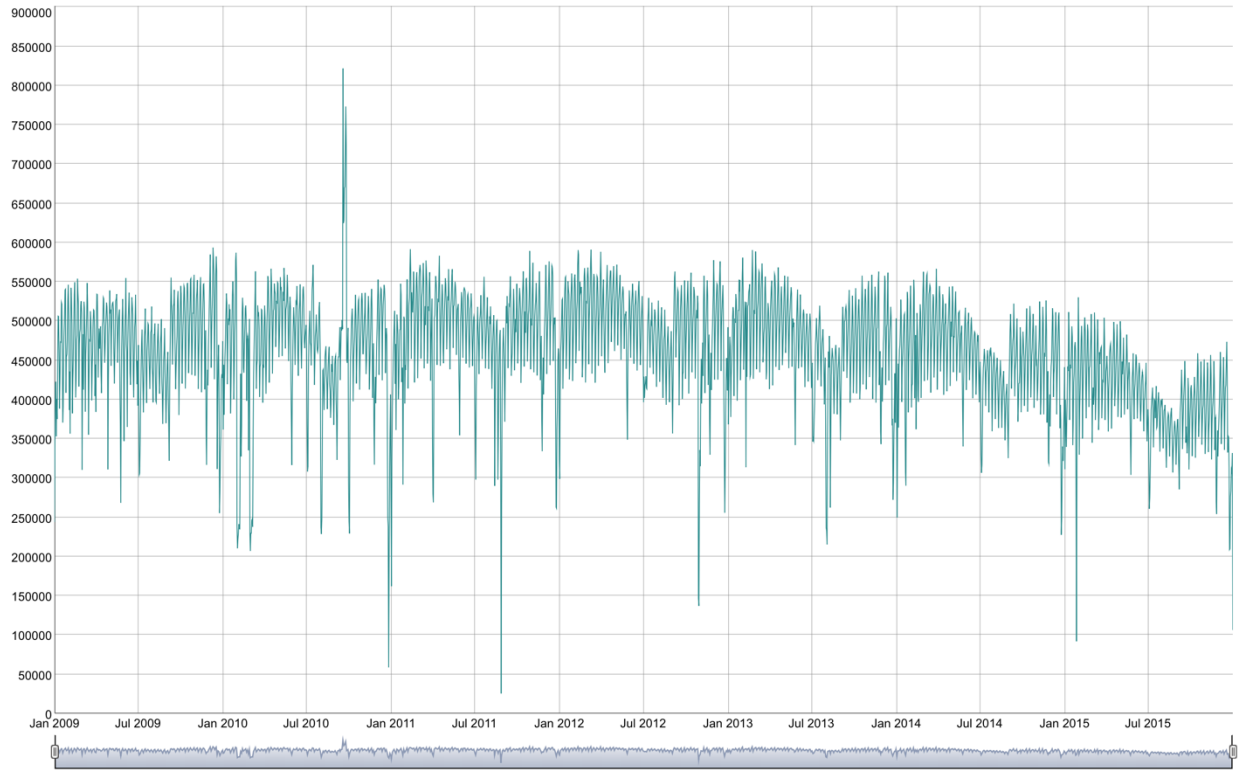


Figure 6.

By looking at the graph, we can say that the time series is weakly stationary. The reason to make this conclusion, considering the fact that there tends to be some seasonality in data that we can see by looking at it. But it is because of incorrect data.

To further check if it is stationary or not, I did Augmented Dickey-Fuller Test to check stationarity in data.

Dickey-Fuller = -7.262	Lag order = 13	p-value = 0.01	alternative hypothesis: stationary

Table 2. Stationary Check.

Based on table 2, we can say that the data is stationary as the p value is 0.01. The assumption/ null hypothesis for this test was that the series is nonstationary. But based on p-value, we can say that it is nonstationary.

To be double sure, I also checked for seasonality in the data. There was no seasonality present in the data based on test performed.

Now let's look at the ACF and PACF plot to decide which Arima model should we make.

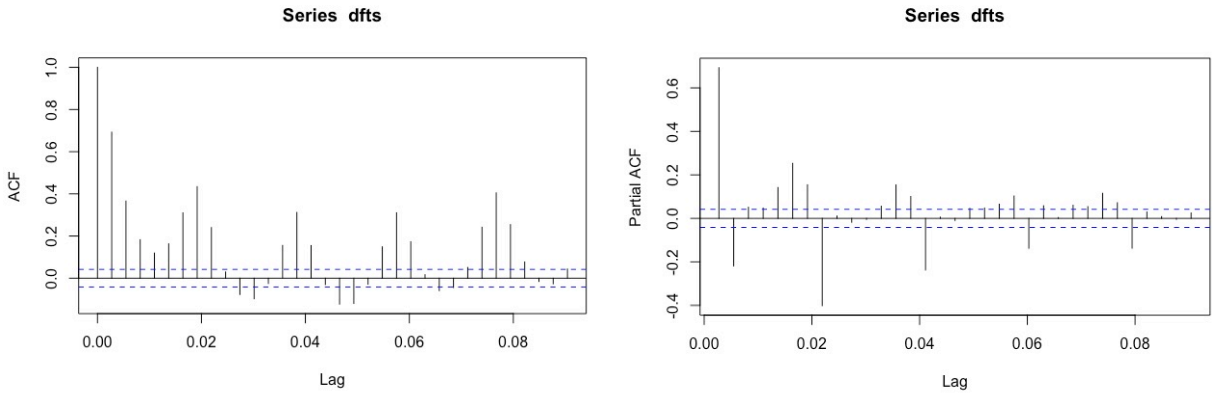


Figure 9

We see that both the ACF and PACF eventually dies down, so we should use ARMA model. Let's see which models performs better.

```
Series: dfts
ARIMA(5,1,2)

Coefficients:
         ar1      ar2      ar3      ar4      ar5      ma1     ma2
      0.2462  -0.6852  -0.1880  -0.2633  -0.3505  -0.4078  0.4481
s.e.  0.0860   0.0311   0.0369   0.0199   0.0382   0.0957  0.0351

sigma^2 estimated as 2e+09:  log likelihood=-26688.82
AIC=53393.63   AICc=53393.7   BIC=53439.2
```

Figure 8

The First Model that we made was ARIMA (5,1,2).

The AIC is 53393.63.

Let's look at diagnostic plot now.

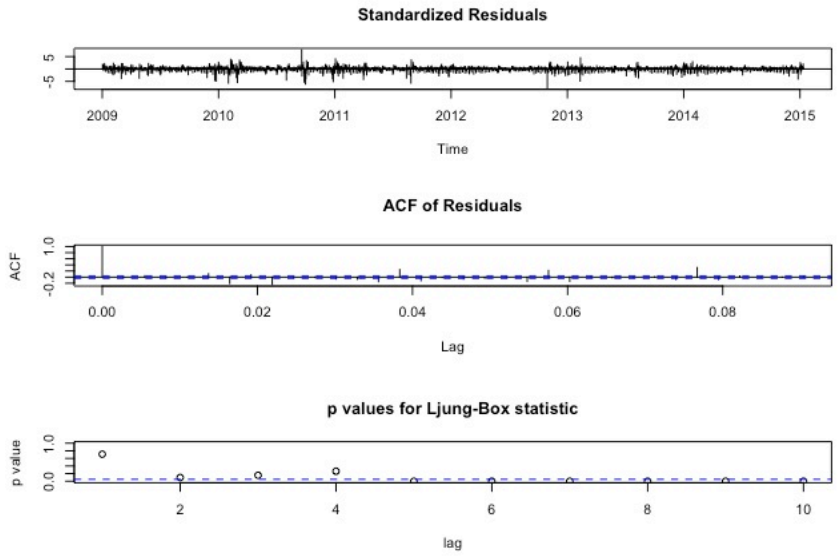


Figure 9

As we can see, the ACF plot looks good. The p-values for Ljung-Box test is above 0.05 for most part of it. Let's look at the other model now.

```
Call:
arima(x = dfts, order = c(3, 1, 2))

Coefficients:
          ar1     ar2      ar3      ma1      ma2
      -0.0876  0.4486  -0.1916  -0.0418  -0.8868
s.e.   0.0389  0.0364   0.0268   0.0319   0.0292

sigma^2 estimated as 2.207e+09:  log likelihood = -26801.14,  aic = 53614.27
```

Figure 10

The second model is ARIMA 3,1,2. The AIC value is 53614.27. Let's look at the diagnostic plots now.

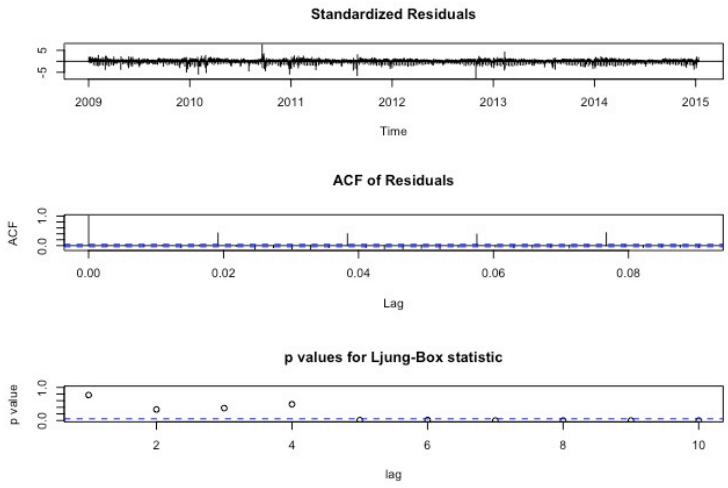


Figure 11

We can see from the diagnostic plot that ACF looks good and Ljung-Box also looks good. But based on AIC value, we will choose the ARIMA (5,1,2).

Also, let's look at the Errors for both of the models.

```
Training set error measures:
                    ME     RMSE      MAE       MPE    MAPE      MASE         ACF1
Training set 173.6402 44634.78 29554.18 -1.458322 7.67479 0.7530374 -0.007900224
```

Figure 12. ARIMA 5,1,2 Model

```
Training set error measures:
                   ME    RMSE      MAE       MPE  MAPE      MASE         ACF1
Training set 183.601 46968.5 33756.54 -1.855339 8.698 0.8601132 -0.006214779
```

Figure 13. ARIMA 3,1,2 Model

Based on the error measures too, we can say that ARIMA (5,1,2) is better model for this dataset.

Let's look at prediction for ARIMA (5,1,2) model now.

```
Forecasts:
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2015.0329       447163.6 389948.9 504378.4 359661.2 534666.1
2015.0356       433769.3 359106.3 508432.2 319582.2 547956.4
2015.0384       445282.4 364000.4 526564.4 320972.3 569592.5
2015.0411       444799.1 360101.7 529496.6 315265.5 574332.7
2015.0438       458256.7 371390.9 545122.4 325406.9 591106.4
2015.0466       477729.0 390374.1 565083.8 344131.3 611326.6
2015.0493       475056.0 385695.0 564417.0 338390.1 611721.9
2015.0521       454618.0 358023.4 551212.6 306889.3 602346.6
2015.0548       444383.7 339646.3 549121.1 284201.6 604565.8
2015.0575       446525.7 337116.6 555934.7 279198.9 613852.4
2015.0603       451785.1 339671.1 563899.0 280321.6 623248.5
2015.0630       459854.6 345969.4 573739.8 285682.2 634027.0
2015.0658       467694.0 352361.7 583026.2 291308.5 644079.4
2015.0685       466129.6 348492.2 583767.0 286218.7 646040.4
2015.0712       456720.6 334950.5 578490.6 270489.3 642951.8
```

Figure 14. Forecast

Based on the figure above, we can say that ARIMA 5,1,2 was a better model. But we did miss a key point. The data did not show any seasonality. But the data did have multiple seasonality in it.

To get an accurate result, we needed to use TBAT model, which is Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components.

Let's see the TBAT model now

```
Call: tbats(y = x.msts)

Parameters
  Alpha: 0.01348936
  Gamma-1 Values: 1.508705e-05
  Gamma-2 Values: 2.409704e-05
  AR coefficients: -0.039969 0.395124 -0.198851 -0.046961 0.050858
  MA coefficients: 0.857842
```

Figure 15. TBAT Model

The AIC for TBAT model was 74949. It is higher than the other two model, but it takes into consideration multiple seasonality in the data. Let's see the forecast of TBAT model.

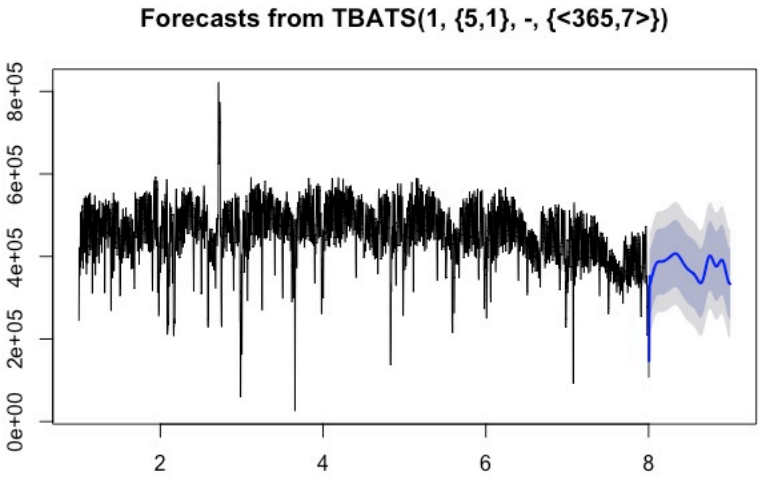


Figure 16. Forecast from TBAT Model

As, we can see, the forecast shows that the trip frequency each day, for yellow taxi are set to decrease.

Also, The auto.arima function also cannot detect multiple seasonality's as seen in this data, since these can only handle a single type of seasonality: *either* weekly *or* yearly. To get a correct predication, even though the AIC values of TBAT Model is high, I used TBAT model.

Lastly, I can say that TBAT model will give a better prediction and is more suitable for this kind of data that exhibit multiple seasonality in it.

Conclusions and Recommendations

Based on the analysis, we can say that the yellow taxi trip frequency is set to go down. It could be attributed to many factors like Uber and LYFT coming to New York City. We can clearly see from the data the forecast that it will reduce.

One of the other reasons for slow decline in Yellow taxi from 2014 could also be the green taxi that came around 2013 in New York City.

Yellow taxi needs to come up with competitive pricing and new marketing ideas to promote people in New York City to use Yellow Taxi Service instead of Uber and other Paid Taxi service.