**Citi Bike Big Data Analysis**

**698 Final**

Chunmei Zhu

**Introduction**

I am interesting in the life style of people in New York. Citi Bike Ride has been an essential behavior for some New Yorkers. Citi Bike is a largest bike share system in the nation, started to launch in May 2013 in New York City. It is available daily for use in 24 hours and riders have access to thousands of bikes at hundreds of stations across Manhattan, Brooklyn, Queens, Bronx, Staten Island and Jersey City. The bikes can be unlocked from one station and returned to any other station for one-way trips throughout the city.

Citi Bike is operated by Motivate, the global leader in bike share. Motivate works to re-envision how people experience and move around cities. In 2018 Citi Bike has over 143,000 members and grew to 12,000 bikes. Riders take their 50 million trips to 750 stations a year (1).

It was designed for quick trips with convenience in mind, and it’s a fun and affordable way to get around town. Single ride costs $3 up to 30 minutes and Day Pass costs $12 unlimited 30-minute rides in a 24-hour period. It also provides $14.95 monthly commitment and $169 annual membership for unlimited 45-minute rides (2).

**Challenges and Methodology**

In this project, the main challenges are data clearing for more than 17 million real time data in 2018, analysis rider behavior and performance by gender, predation on daily number of trips based by time series data from 2013 to 2019 using machine learning model and deep learning model.

The real time 2018 monthly data are zipped and posted in Citi Bike official website (3). Every file is near 1.5 million rows and 15 attributes, which includes Trip Duration (seconds), Start Time and Date, Stop Time and Date, Start Station Name, End Station Name, Station ID, Station Lat/Long, Bike ID, User Type (Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member), Gender (Zero=unknown; 1=male; 2=female), and Year of Birth.

It is impossible to open in Excel since data size is over the max number of rows in Excel and it would be truncated while it was opened in Excel. It only can be opened in text file format. Another, the links for these data file download are feed from amazon.com. It is not allowed me to do web scripting. So I need to download the zip files to my computer and wrote procedures to extract data from zip files and query data in python for my data cleaning, data mining in the feature engineering.

For user behavior analysis, I added daily NYC weather report from May 2013 to April 2019 and requested the data from xxx (4). It is a common sense that people are willing to go out in warn and nice weather and won’t like to bike during storm and freezing days. Average daily temperature, average daily snow depth and average wind speed are important factors impacted to biker’s behaviors.

**Citi Bike Main Finding in 2018**

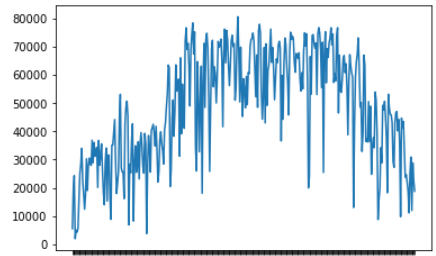
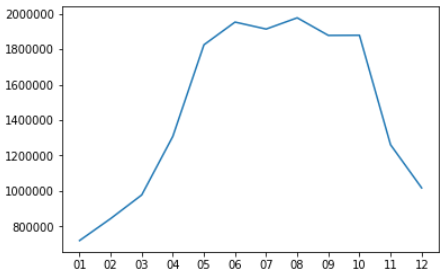
I curious some questions about the user behaviors, such as where the rides start to ride and where they go to, when they ride, which days they ride. My finding for these answer from Citi bike real time data in 2018 which published in its official website.

**Most Popular Start Stations and End Stations**

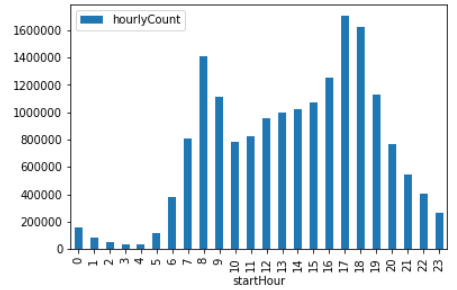
The most popular start stations in 2018 are From Central Park S & 6 Ave to Central Park S & 6 Ave, From E 7 St & Avenue A to Cooper Square & Astor Pl, From Grand Army Plaza & Central Park S to Grand Army Plaza & Central Park S, From Central Park S & 6 Ave to 5 Ave & E 88 St and From S 4 St & Wythe Ave to N 6 St & Bedford Ave.

**Daily Trip Counts**

People bike behavior has strong relation to weather condition. Per 2018 daily tip count on in the left chart and monthly trip count in right chart, the pick time duration is from May to October. The lowest volume is in first quarter and the total daily tips was less than 20,000. There are a large volatility in the daily trip counts. What make the impact in the daily base trips?

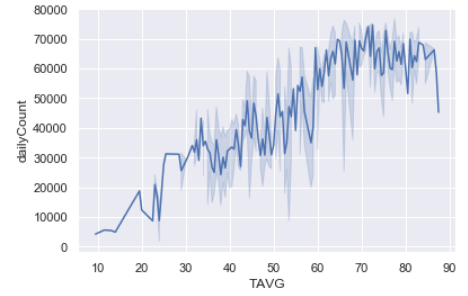
 

Per hourly trip count, most people like to bike during day light hours. The pick hours are from 8 am to 6 pm EST. Very few users biked at mid night before sun rise since it is not safe under critical visible condition on streets. There is longer day light hours the more Citi Bike tips.

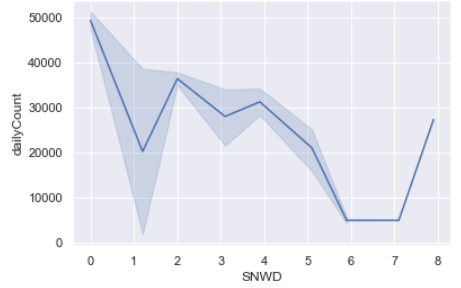


**Weather impact on Daily Trip Counts**

I want to know how weather effect the user behavior on biking. Per daily tip count in 2018 agains temperature, people love to bike during 65 to 75 degree in nice weather condition. The point of the turn over happenes at 85 degree that user won’t like to bike under more than 85 degree. Also people won’t like to bike under 25 degree. During 25 to 85 degree, number of daily trips has a linearly relationship with temperature, the more higher temparature the more Citi bike trips. Hower, I still see significan voritility due to high volumn changing from customers in daily base.



Looking into bad weather condition in snow days in 2018, number of daily trips has very sensitive when snowing. There is a deep drop in one inch snow depth. However, there were more people used Citi bike when snowed 2-4 inches. After review the snow day’s calendar, it happened in holiday season in December and January. New York travers bushed the number tips as customer user.



**Subscriber vs Customers Daily Tip Counts**

Who distribute the number of the trips? Since Citi Bike program provides different memberships for users, I split the data by user type and converted data-time value to month and weekday in the further analysis.

Both daily-trip counts by Subscribers and customers have same pick period in summer time. However, there are same critical daily-total counts of annual members against the counts from customers. What difference behaviors between annual membership users and customers?

Comparing monthly average weekday trip counts of subscribers and customers, subscriber users distributed less trips on Sunday and Monday. Subscriber users use Citi Bike similar times from Tuesday to Saterday.

Unlike subscriber users, there are more customers using Citi Bike on Sunday and Monday. The daily trip counts on these days are significant higher than other weekdays. Alos the daily total trips counts are more sensitive by the weather conditions. These users use 24-Hours pass or 3-Days pass. Most of them are either travlers in New York or active people living in New York.

Mean duration of subscriber is 13.2 mins and mean duration of customers is 43 mins. Over 70% trips of Subscribers are less then 15 minues and around 20% trips in 15 to 30 minues. However, nearly 50% trips of customers like to bike 15 to 30 minuse and less than 30% of them are in 15 minus since they use to bike for travels on Sunday and Monday.

|  |  |  |
| --- | --- | --- |
| Duration (minutes) | Subscriber Trips | Customer Trips |
| <15 | 11434025 | 570602 |
| 15-30 | 3257295 | 914763 |
| 30-45 | 788343 | 251066 |
| 45-60 | 89264 | 76668 |
| >60 | 45898 | 120415 |

**Citi Bike served in Long-Duration Trips**

Last year 2018, there are 14,894 Citi bikes served 332,245 trips which more than 45 minutes, which 59% of these trips from Customers and 41% from Subscribers. Were the bikes stolen issues? If the bike was stolen, the bike id would be showed in once in data. I need to remove the data from my analysis. If it wasn’t the issue, where they spend more than 45 minutes for a trip?

|  |  |
| --- | --- |
| **Trips >45 minutes durations** | **Count of trip** |
| From Grand Army Plaza & Central Park S to Grand Army Plaza & Central Park S | 3996 |
| From Central Park S & 6 Ave to Central Park S & 6 Ave | 3228 |
| From Grand Army Plaza & Central Park S to Central Park S & 6 Ave | 1968 |
| From Broadway & W 60 St to Broadway & W 60 St | 976 |
| From Soissons Landing to Soissons Landing | 772 |
| From Grand Army Plaza & Central Park S to Central Park West & W 72 St | 754 |
| From Grand Army Plaza & Central Park S to Broadway & W 60 St | 736 |
| From Centre St & Chambers St to Centre St & Chambers St | 733 |
| From Central Park West & W 72 St to Central Park West & W 72 St | 705 |
| From West Drive & Prospect Park West to West Drive & Prospect Park West | 627 |
| From 5 Ave & E 88 St to 5 Ave & E 88 St | 588 |
| From Yankee Ferry Terminal to Yankee Ferry Terminal | 551 |
| From 5 Ave & E 73 St to 5 Ave & E 73 St | 550 |
| From Central Park West & W 68 St to Central Park West & W 68 St | 529 |

I found these bikes were used multiple times in long-duration tips. There are 85% of those bicyles distribuded more than 10 times in more than 45 minutes tips. So it should not be in the issue of theft but in less than 10% total trips some users biked in long journeys.

Most long duration trips are circle tips. Frist two top favories tips are from Grand Army Plaze and Central Park S for travels. Another popular trip is at Uper Ease Manhattan nearly river side. Obviously, Citi Bike has very compatortive prices to the local small bicyle rental companies.

**Male vs Female Citi Bikers Performance**

This is a tricky question. One of well-known factors for bicycle performance is speed, which calculate by distance divide by trip duration in a same condition. There is not distance counted in these data sets. The challenges are how to get the distance and proper ride time.

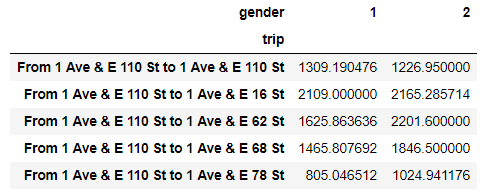
Someone suggests to get the distance traveled according to Google Maps. This probably understates the rider’s actual average bike speed since the riders may check next bike stations, wait for friends, check directions or deal with distractions. What’s more, the riders may not follow the route advised by Google Map.

Another, there is a common sense to most riders started from point A and end at point B as quickly as safely possible during weekdays. Riders will take longer time during pick hours of traffic and in bad weather conditions. Some brave or aggressive riders are not faired to ride on the busy traffic areas for short-cut but some do and take longer distance to the end point. Some riders ride for work and some ride for travel or exercise. Some riders ride individually and some ride with groups.

It is impossible to get the accurate number of speed by these limited raw data. However, to evaluation the performance by gender based on average trip distance, I would suggest to compare the data in a same trip – from same start station to same end station, similar day hours and weekdays or weekends.

The real task is to define the ‘same’ condition in the ride environment.

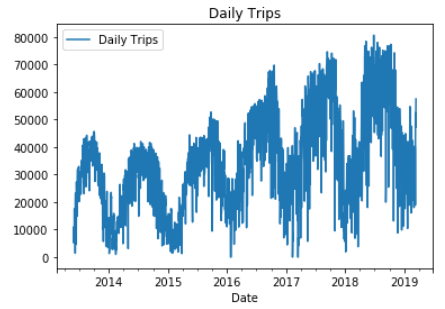
* Same trip – same start station to same destination
* Compare mean values of same trip by gender
* Number is the different of mean value of gender in different duration level of Male (Diff in seconds = mean tripDuration of Male – mean tripDuration of Female



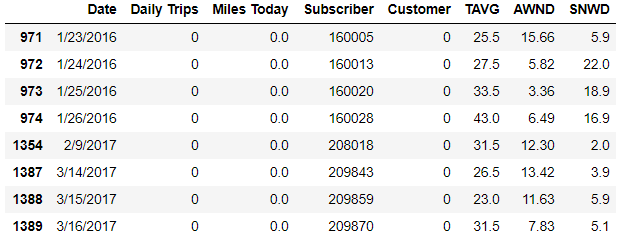
|  |  |  |
| --- | --- | --- |
| Sex\_Trip\_Dur\_Diff | Mean | Median |
| <30 mins | -409 | -104 |
| <45 mins | -527 | -83 |
| >45 mins | 5007 | 733 |

**Daily Trips of Citi Bike Prediction**

In May 2013, the number of annual membership increased from 0 to 30 thousands and the number of 24-hour pass and 7-Days pass customer were more than five thousands in the first week, which produced over nigh thousands in trips in a day. Nowadays, Citi Bike has more than 300 thousands subscribers and over a thousand daily base customers including 24-hour pass and 3-Days pass in March 2019.



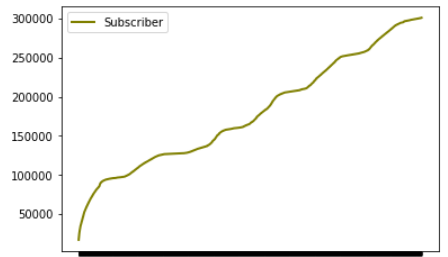
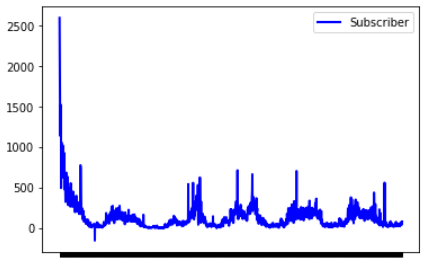
There are 8 days having zero tips in history data. Combining daily weather data (average of the day temperature, average wind speed and snow depth by inches), there were ether very windy, freezing temperature or heavy snow.



The number of daily trips has been observed seasonal pattern, pick duration from May to September and non-pick duration from October to March. It due to the trips from number of customer and weather condition effect, like temperature, rain, wind and snow.

**Subscriber**

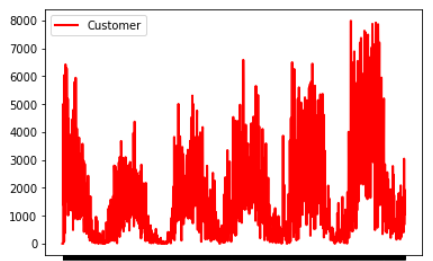
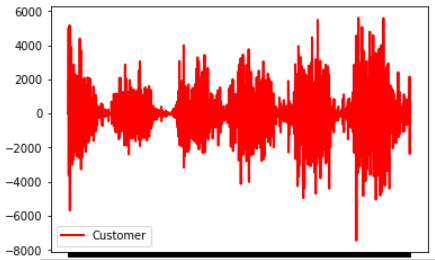
Per the chart of Daily Total Subscriber (left), the total number of the subscriber is continously growing during May 2013 to March 2019. There are 134 new members in daily average to subscribe the annual memberships of Citi Bike. In the chart of Daily Difference Total Subscriber, most new joiners singed in the program during summer time every year.

Daily Total Subscriber Daily Difference Total

**Customer**

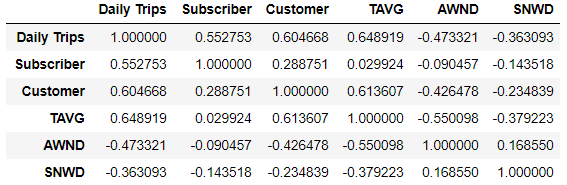
Citi Bike started its non-annual meberships as 24-Hours pass and 7-Days pass in 2013, then changed 7-Days pass to 3-Days pass June 2016, which provided more frexible to customer. Per the chart of Daily Customers (left), there is a clear seasonal patten in the curv with large daily volatirity. Per the chart of Daily Customers Difference (right), it also has obvisory patten.

Daily Customers Daily Customers Daily Difference

**Correlations**

Per the table of correlation bewteen number of daily trips and difference user types and three major weather features, tempurature and numbers of customer and Subscriber have very strong positive over 0.55 correlation, and averager wind speed and snow deepth have negative relationship to reduce the number tips.



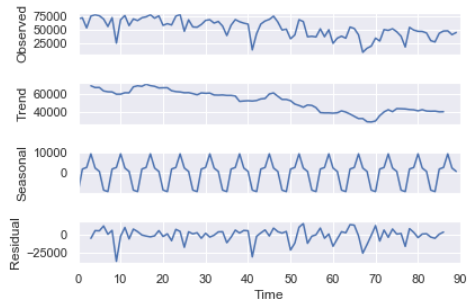
**Seasonal and Stationary Test**

This is a time series data which has time order. The character of time series data has non-stationary and seasonal effect in general. A model may fit the data during current period but may not in the future. Before to build a forecast model for prediction to aviod the misstake of modeling, seasonal and stationary test are necessary.

**Seasonal check for Daily Trip Counts**

There is statsmodels.tsa.seasonal package in python to check whether the data has seasonal patten. I also want to see residual patten of the data is random walk or white noise. If the residual is white noise with mean 0, it is stationary.

Zoomed in the graph from seasonal\_decompose for curve, trend, seasonal and residual on 2018 quarter 4th data set, the number of daily trips has down trend, seasonal patten, not white noise patten residual.



**Dickey-Fuller statistic for Stationary check**

I use Dickey-Fuller statistic to test stationry of the data. Formally, it tests the null hypothesis H0 that an autoregressive model has a unit root. Therefore, you have to be careful and understand that if you seek stationarity (which is usually the case), you want to reject H0. The test result on daily trip counts with 95% confident level is false to reject H0, which means the data is nonstationary.

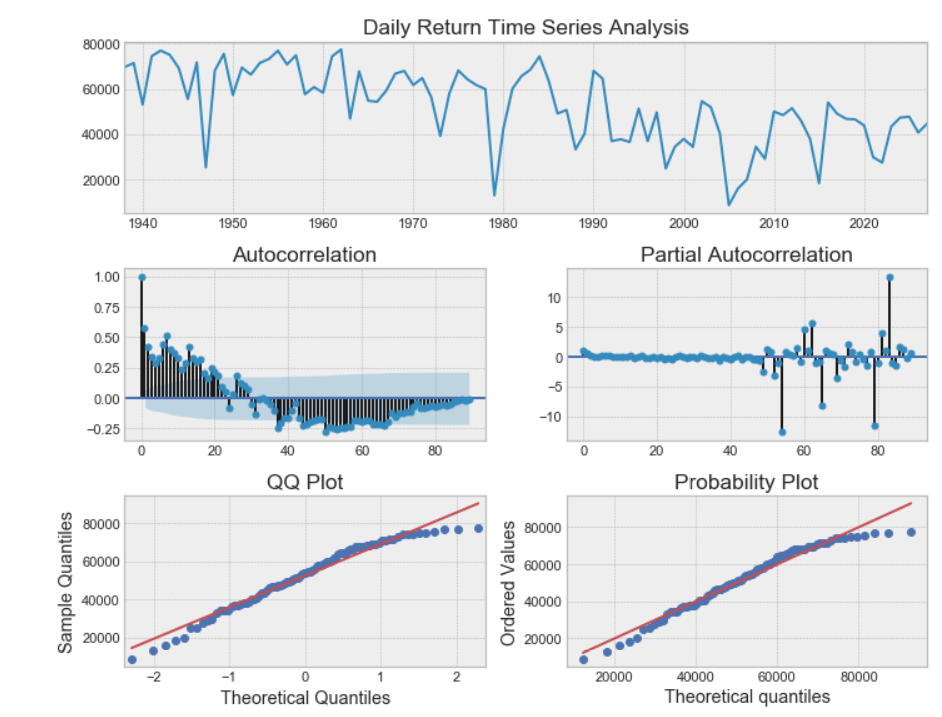
**Durbin-Watson Statistic for error correlation check**

Unfortunate, error of daily trips has correlate to the curve itself. That means the error of the time series data is dependent.

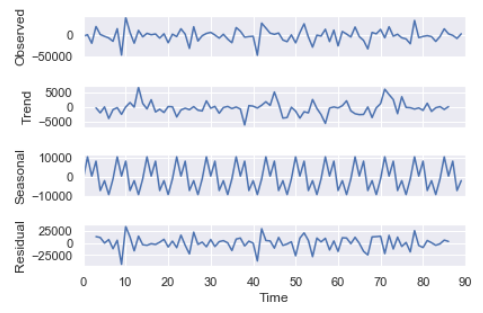
**P/ACF for lags independcy check**

To check whether the residuals are any significant corrections, I use auto-correlation function to compute the correlation between a variable and itself at each lag step up to some limit (in this case 40). The partial auto-correlation function is used to compute the correlation at each lag step that is NOT already explained by previous, lower-order lag steps.

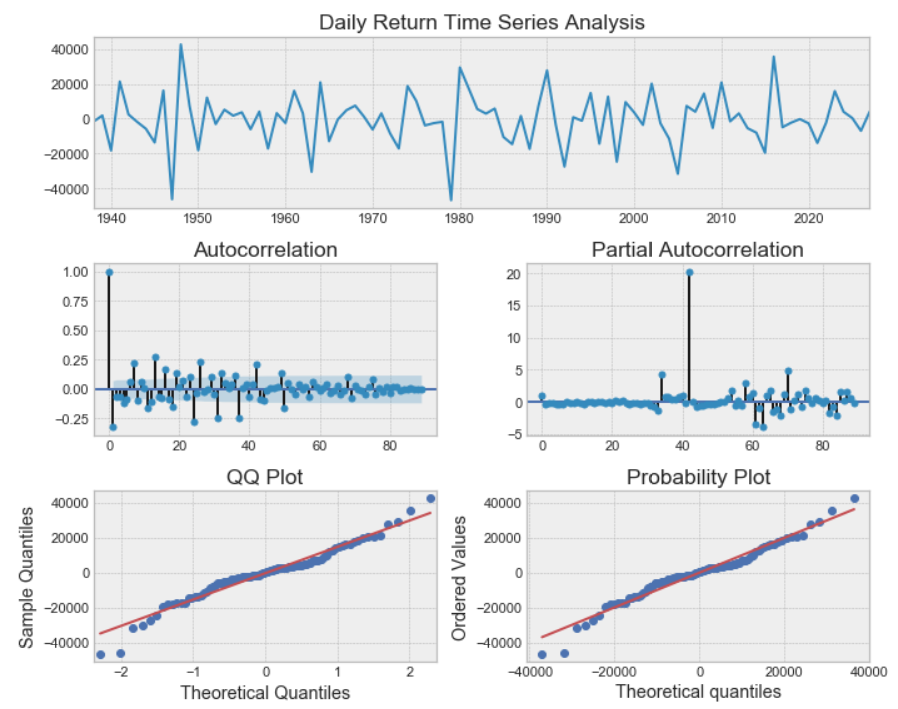
The auto-correlation and partial-autocorrelation results are very close to each other. What this shows is that there is no significant (> 0.2) correlation between the value at time t and at any time prior to t up to 40 steps behind. In order words, the series is a random walk.



I use the diff() in python to find the difference of number daily trips at time t-1 to the value at time t, which transforms the non-stationary data set to stationary data. The new data set passes all tests above.

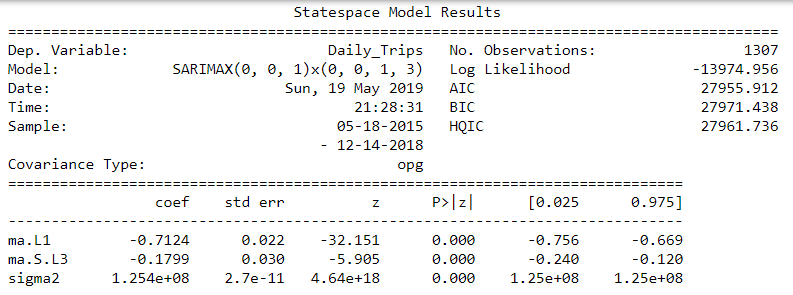
**Seasonal check for Daily Trip Counts Differenct** 

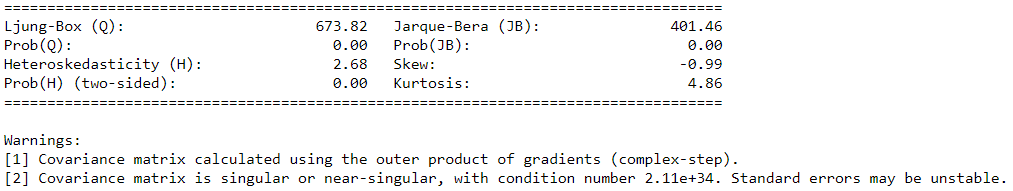
**P/ACF of Daily Trip Counts Differenct**

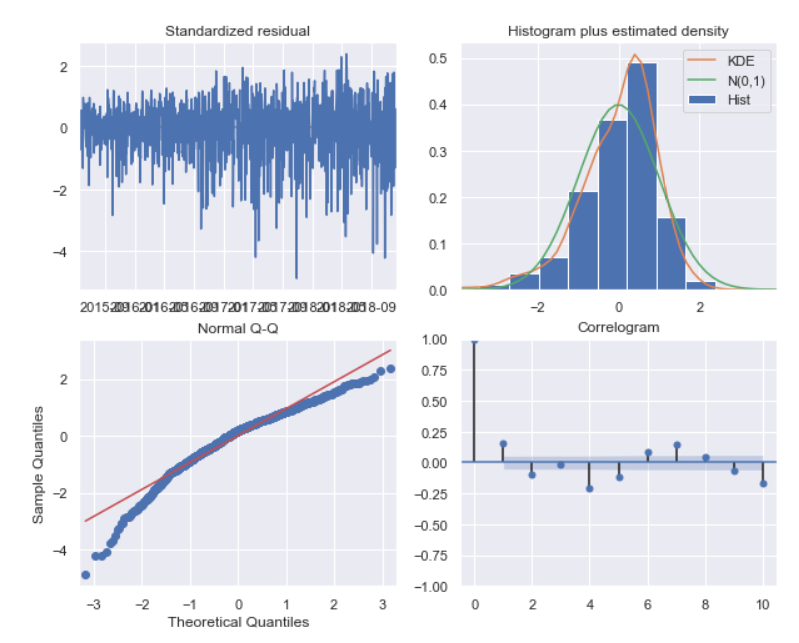


## Seasonal Auto-Regression Moving Average – SARIMA

I use seasonal auto-regression moving average (SARIMA) machine learning model for the time series data perdition. In this model, I wrote procedure to find the best model with the lowest AIC value in 95% confident level for seasonal model parameters. To predict daily trip counts during Oct 1, 2018 to Dec 12, 2018 by using daily counts from May 18,2015 , my best model is SARIMAX (0, 0, 1) x (0, 0, 1, 3), which mean MA (1). The model returns R2 0.2015211 and 156670405 RMSE score.









## Long short-term memory network - [LSTM Network](https://en.wikipedia.org/wiki/Long_short-term_memory) with Keras

I use long short-term memory network ([LSTM) Network](https://en.wikipedia.org/wiki/Long_short-term_memory) for daily difference of trip counts. It is input attributes Date and difference of one shift daily trip count as first layer. Then I designed four hidden layers to process learning with 70% train data set to predict 30% test data. It returns 12477.18 RMSE on test score which 8186.93 RMSE train score. This model improves the result of SARIMA model. Here is the result of the perdition of LSTM model.

## 

**Further considerations about business of Citi Bike in future**

If I were launched the Citi bike project, I would concern whether it could be at least making balance of earning and expending even it is supported by New York local government and Citi Bank advertisement fee. The demand, supply, risk and opportunity of Citi bike are the major factors in my consideration.

**Demand**

Per annual data of the trip counts, the demand of Citi bike is continuously growing after 2015 because of more stations opening, increasing number of bike and more compatible price then other transportation tools. It is hard to know the accurate demand of Citi bike base on the current data sets due to the limitation of number of stations, number of bikes in service by the real time.

It could be estimated by scurvies for the users to simulate to book bikes at desired start location and desired destinations at specified pick-up times and arrival times. What is more, it could be a better solution if there was a Citi bike app which allowed users to plan to book bikes at pacifist stations at pacifist real pick up times. To avoid the breaking plans from users, it could be allowed to do pre-paid or some amount of fee.

**Supply**

The major costs of launching the Citibike project are from the cost of bike including maintaining and fixing, storage fees of the bikes on the streets, digital system of bike parking, program management fees, project launching fee. The management fee includes to hire some agents to organize the amounts of bikes in the stations by watching number of available bikes at each stations by real time.

The marginal cost of each bike could be increased when the labor and material cost of Citibike plus an estimated portion of fixed costs (such as administration overheads and selling expenses) reach to the breakeven point. The increased basic salary by law and economic inflation in current years can reduces the amount of bike in supply.

**Risk**

Per risk from bikers, it is hard to be monitored the behaviors of every single biker. Some bikers bike at side walk and some streets without bike lanes. Another, bikers get hurt due to the poor conditions of old Citi bikes, that is possible to bring the cost of compensations from the law suits of Citi bike.

Per risk from local government, the government supervises what benefits and problems the Citi bike project brings to. The more Citi Bike trips the more space for bicycle lanes on the streets. It will be consider whether the increasing number of bicycle space would bring problems like reducing bus, taxi or other transportation speed and increasing car accidents causes by bikers. The government can limit the number of bike to be stored at certain stations and the number of bike station opening base on the balance of use from Citibike and other transportation tools.

**Opportunity**

Better environment for bikers – Current New York City government has passed a law which increase the fees to the private cars which enter Manhattan area to reduce the traffic problems. Furthermore, local government increases protected bicycle lance in New York. In 2019, there are 30.3 miles (4) for bicycle lance and majorly increate more space in Manhattan and Brooklyn Where young people live or have actives.

In 2019 MTA fare has been increased to $130 monthly pass, $3 a trip and will plan to continuous grow above $200. New York people have bearded the pain of slowness and high cost of Subway service. Even Umber has announced to reduce the cost per mile at beginning of this year, Citi Bike still has very competitive price to other competitors for short trips.

**Reference**

1. Citi Bike Trip Data: <https://www.citibikenyc.com/about>
2. Citi Bike Price: <https://www.citibikenyc.com/pricing>
3. NYC Weather: <https://www.ncdc.noaa.gov/cdo-web/confirmation>
4. *Protected Bicycle Lanes* [*https://www1.nyc.gov/html/dot/downloads/pdf/nycdot-protected-bike-lanes-2018-present.pdf*](https://www1.nyc.gov/html/dot/downloads/pdf/nycdot-protected-bike-lanes-2018-present.pdf)