American University of Armenia

College of Science and Engineering

Time Series Forecasting: Final Project

Hakob Janesian, Emma Hovhannisyan

Fall 2022

Time Series Analysis of GG taxi data

Contents

List of figures

Abstract

In today's competitive business environment, taxi companies constantly seek ways to differentiate themselves and stand out in the market. One way they do this is by leveraging technology to gain a competitive edge over other firms. By investing in IT expertise, taxi firms are positioning themselves to stay ahead of the curve and remain competitive in an increasingly digital world. Taxi firms may benefit significantly from time series analysis since it enables them to track changes in customer demand over time, predict future demand, and adjust their operations accordingly. And our group believes that GG taxi would also benefit from the results of the time series analysis.

Introduction

The goal of the project is to do an analysis of time series data, which will help us to create models for prediction. Thus, the result of the project will be the prediction of the daily quantity of orders (completed/cancelled) of the GG taxi service. The data is derived from the taxi service GG, so the analysis will merely give insights into the company's business activity. Thus, the research results will be helpful for the stakeholder - GG taxi, to get the order estimation and the variance of the orders. This information can help GG taxi allocate its resources more efficiently, such as by deploying more drivers during peak demand periods or offering promotions to increase demand during slower times.

Literature review

Scholars D. Khryashchev and V. Huy described methods of predicting the number of taxi orders based on a given historical taxi demand data in a specific region. They have applied several methods as the Markov prediction algorithm (probabilistic model), ARIMA model (time-series forecasting), the Lempel-Ziv-Welch (sequence modeling), and deep learning Long Short-Term Memory (LSTM) models to solve the problem. They have done the testing on 14 million data samples, and as a result, they claim the Markov prediction algorithm and LSTM are the most efficient among all cases. Even they state that the simple Markov prediction algorithm might perform better than the deep learning (LSTM) method when the time series has a high predictability level. Moreover, in November 2020, S. Faghiha, A. Shahb, Z. Wangb, A. Safikhanic, and C. Kamgaa tried to predict the taxi demand of Manhattan taxis by combining time-series ARMA models with a linear regression model. They collected three months of data on yellow cubs, subway, Uber, temperature, and precipitation. Each variable was aggregated hourly. After removing the seasonality. These researchers believe that using only linear regression would be expensive because it would require more variables. They claim that by combining the linear regression and ARMA models, they were able to minimize the number of variables and get a high R-squared value since ARMA can simultaneously identify and model all variables that have been neglected.

Data description

The data is about the orders of GG taxi in the year of 2016. It includes 2 files (gg.rda, gg_all.rda). gg_all.rda contains 2661663 observations with 16 variables, and gg.rda contains 2661663 observations with 10 variables. As the two datasets describe the same events and are about the same period of time, we merged them in order to see the whole picture. As it includes only one year of data, we did data cleaning and visualized weekly and daily data.

The main variables that we focus on are in the following table.

Variable	Description
Number of completed orders	A variable that shows the number of completed orders
Number of cancelled orders	A variable that shows the number of cancelled orders
weekday	Weekdays from Monday to Sunday
created_at	Shows the date, when the order was created

Here is presented the mean, median, maximum and minimum values of completed orders for a day in 2016. **Mean = 4883, Median = 5012, Max = 9500, Min = 174.** Figure 1 represents the number of completed orders during a year, where we have a trend and seasonality.

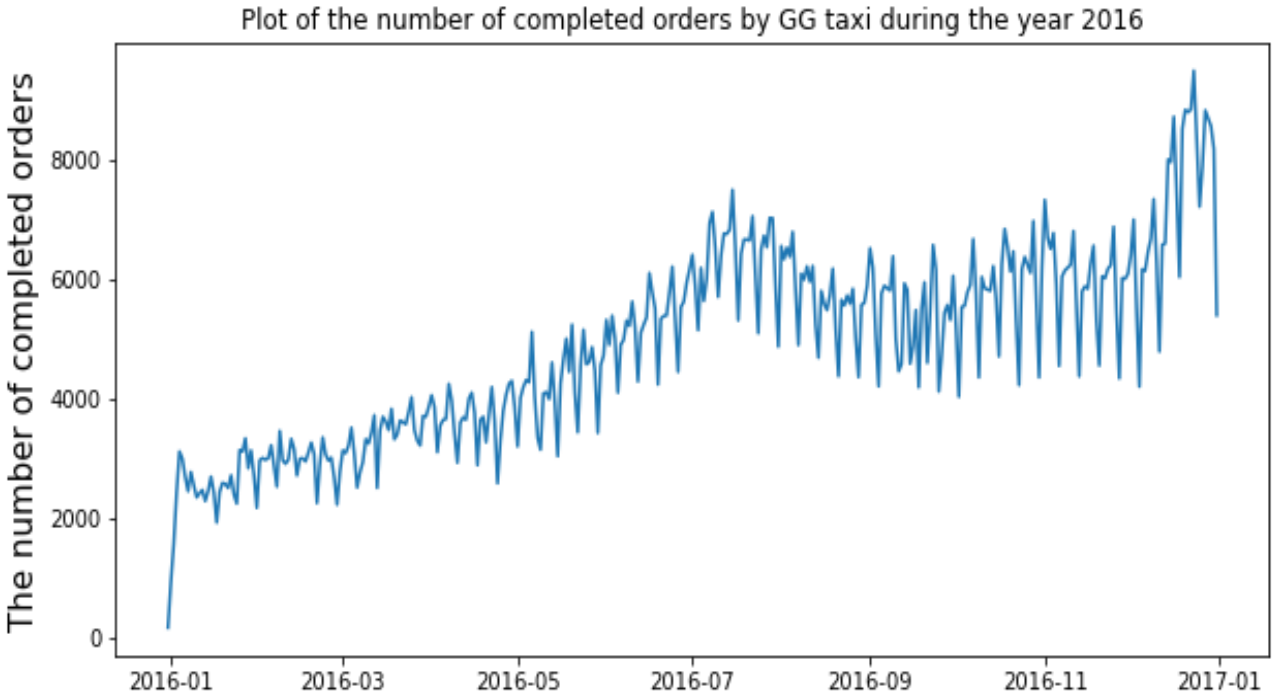

Figure 1

Below, we have presented the mean, median, maximum and minimum values of cancelled orders for a day in 2016. **Mean = 2370, Median = 1555, Max = 28064, Min = 277.** Figure 2 represents the number of cancelled orders during a year, where we can see some seasonality. In addition, in the last month of the year, there is a significant upward trend in cancellation rates.

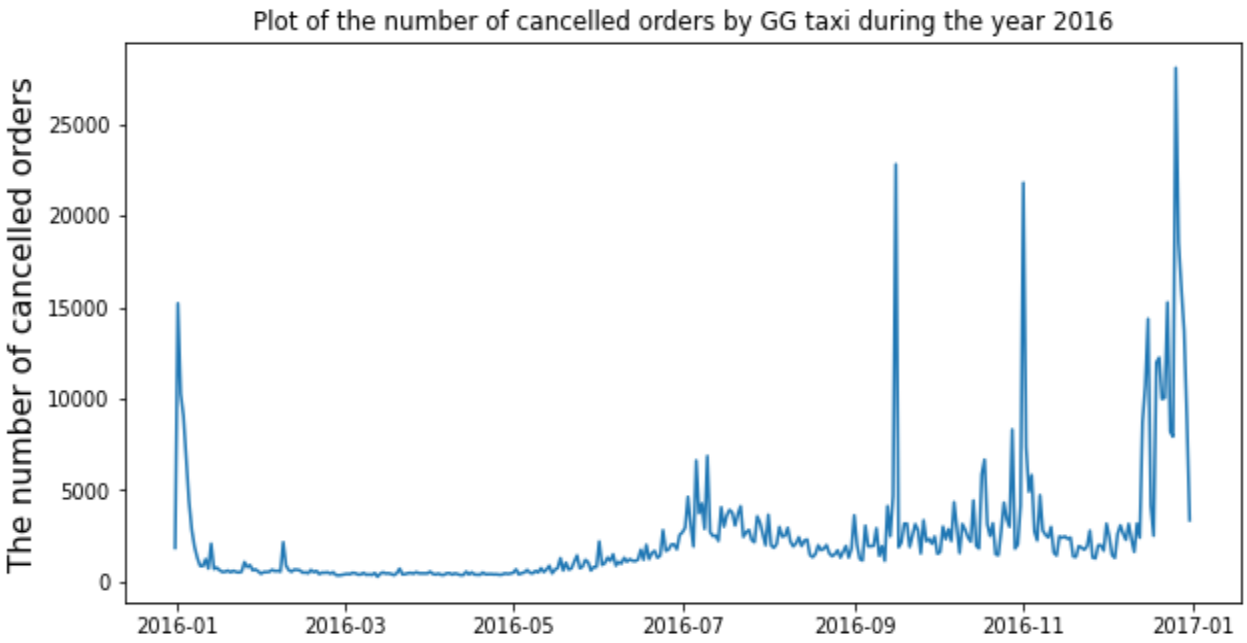

Figure 2

Moreover, we visualized a plot that represents the number of completed orders by weekdays in Figure 3.

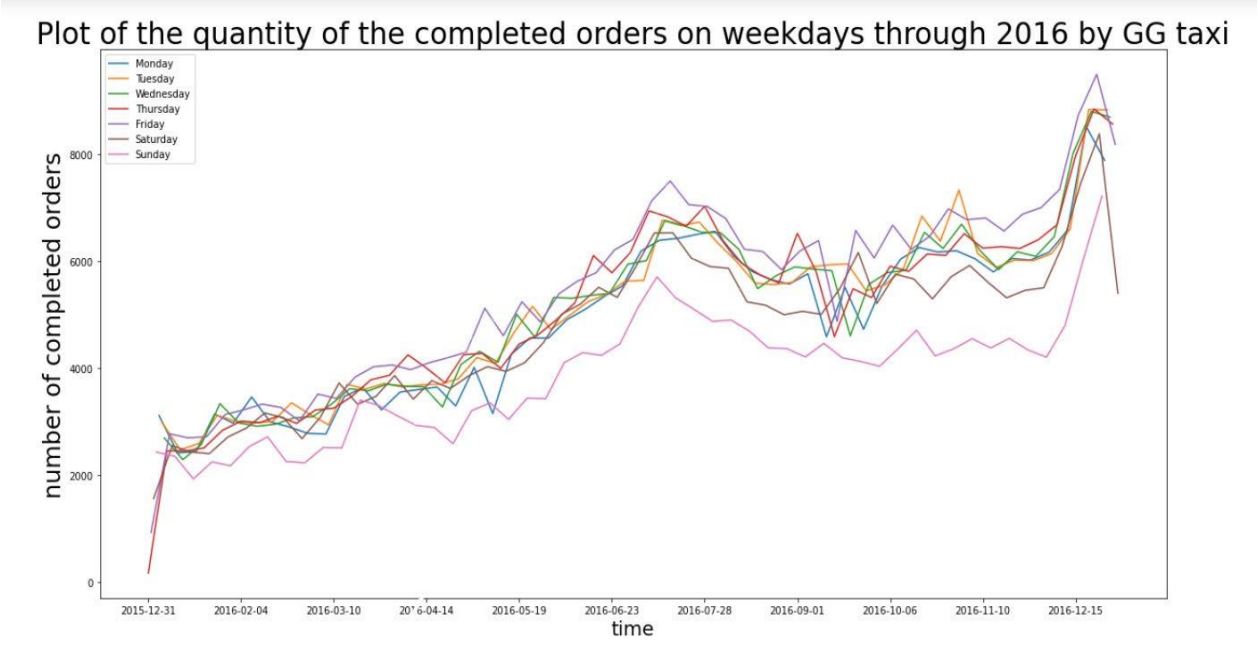

Figure 3

For all days of the week, we see an upward trend, starting from January 2016 till July 2016, then it is followed by a slight downward trend, and during the last two months of the year, there is a drastic upward trend. In addition, there is some seasonality. We can also confidently claim that in comparison with other weekdays, Sunday had the least amount of orders in 2016.

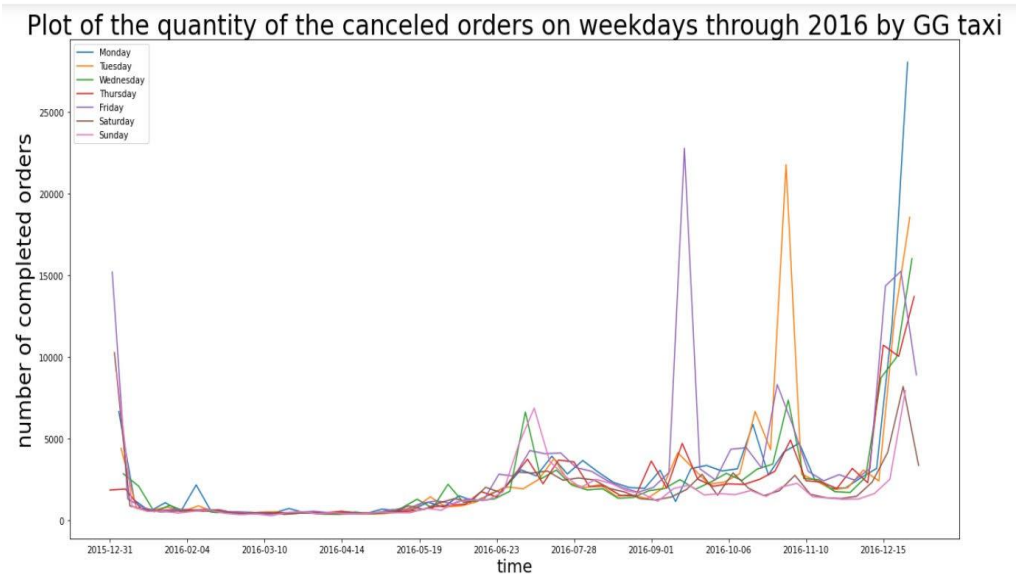

Figure 4

From the plot above, we conclude that in the first six months of 2016, more or less, the cancellation rates were similar on the weekdays. But in the second part of the year, we have some volatility and even spikes between September and October (Fridays), and also, there is a spike around November (Tuesdays). In addition, in the last month of the year, there is a significant upward trend in cancellation rates for all weekdays.

Day	Mean	Median	Max	Min
Monday	4888	5225.5	8526	2422
Tuesday	5084	5506	8843	2478
Wednesday	5104	5447	8806	2295
Thursday	5068	5489	8852	174
Friday	5434	5848	9500	933
Saturday	4773	5182	8387	1568
Sunday	3869	4204	7223	1934

Above, we have presented the mean, median, maximum, and minimum values of completed orders for weekdays, respectively, in 2016. Below, we presented the mean, median, maximum, and minimum values of cancelled orders for weekdays, respectively, in 2016.

Day	Mean	Median	Max	Min
Monday	2682	1970	28064	369
Tuesday	26189	1621	21788	365
Wednesday	2301	1765	16028	371
Thursday	2246	1852	13715	405
Friday	3285	2018	22798	372
Saturday	1760	1381	10270	361
Sunday	1690	1293	9061	277

Methods

Our group initially tested if the series was stationary or not. On top of that, we have done transformations which made the series stationary. Moreover, we have used SARIMA and exponential smoothing models to forecast the number of completed taxi orders for the following two weeks. Additionally, we have used the VAR (Vector Autoregression) model, which multivariate time series model, which means that it can analyze the relationships between multiple variables over time. In our case those variables are the number of completed orders and the number of cancelled orders.

Results

SARIMA model: Initially, we have done an ADF test on the original series and found out since the ADF statistics was greater than all 3 critical values (-1.42 > -3.45, -1.42 > -2.87, -1.42 > -2.57) and then we statistically proved that the time series was not stationary. then we did the KPSS test, and since the p-value was equal to 0.01, which was greater than 0.05, so we concluded that the series did not have a unit root, so we showed that the series is trend stationary. So, we concluded that the series was trend stationary but not stationary, and for that, we needed to do first-order differencing of the series in order to make it stationary. Still, the series was not stationary, and we did a seasonal differencing of the first-order difference data. After seasonal differencing, the ADF statistics was smaller than all 3 critical values (-5.36 < -3.45, - 5.36 < -2.87, -5.36 < -2.57), then we statistically proved that the first order differenced data was stationary. Then by ACF and PACF diagnostics, we identified the dependence orders of this model, which was SARIMA(1, 1, 1)(1, 1, 0)[7]. Moreover, we divided the data into the divided train (90%) and test (10%) and ran SARIMA(1, 1, 1)(1, 1, 0)[7] on the train data. We also performed residual

diagnostics, which you can see in the four plots below.

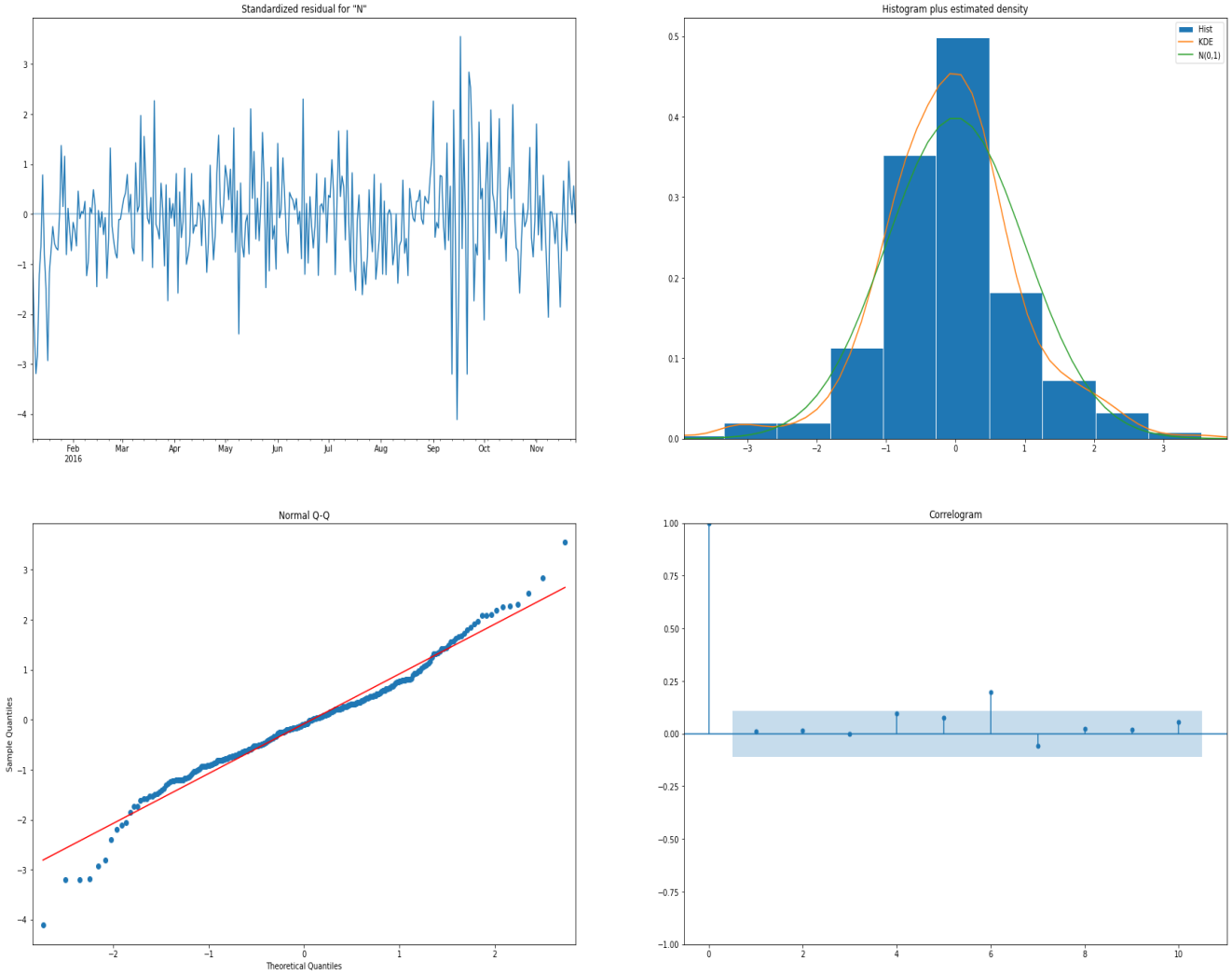

Figure 5

In the upper left plot, we see that standardized residuals are near zero and are not very volatile. In the histogram's plot, we see that the KDE line is close to the Normal (0, 1) bell shape curve. In the QQ plot, we see that points are near the Normal ab-line. And in the correlogram, we see that there is only one significant lag. To conclude, we can claim that the residuals are normally distributed. Then we did the Ljung-Box test, and the conclusion was that residuals were not identically and independently distributed. Additionally, we have estimated a new SARIMA model on a train set with auto.arima function. The output was SARIMA(5,1,0)(1,0,1)[7]. Besides, we have compared the AIC and BIC values of 2 SARIMA models. Considering the values of AIC, 4725.4 (SARIMA(1, 1, 1)(1, 1, 0)[7]) was less than 4789.5 (SARIMA(5,1,0)(1,0,1)[7]) and in case of BIC 4740.5 (SARIMA(1, 1, 1)(1, 1, 0)[7]) < 4819.9 (SARIMA(5,1,0)(1,0,1)[7]). Based on the aforementioned points, we have concluded that SARIMA(1, 1, 1)(1, 1, 0)[7] is a better model than SARIMA(5,1,0)(1,0,1)[7]. After both models had been processed by using the training data, we decided to test the models by making predictions on the test data.

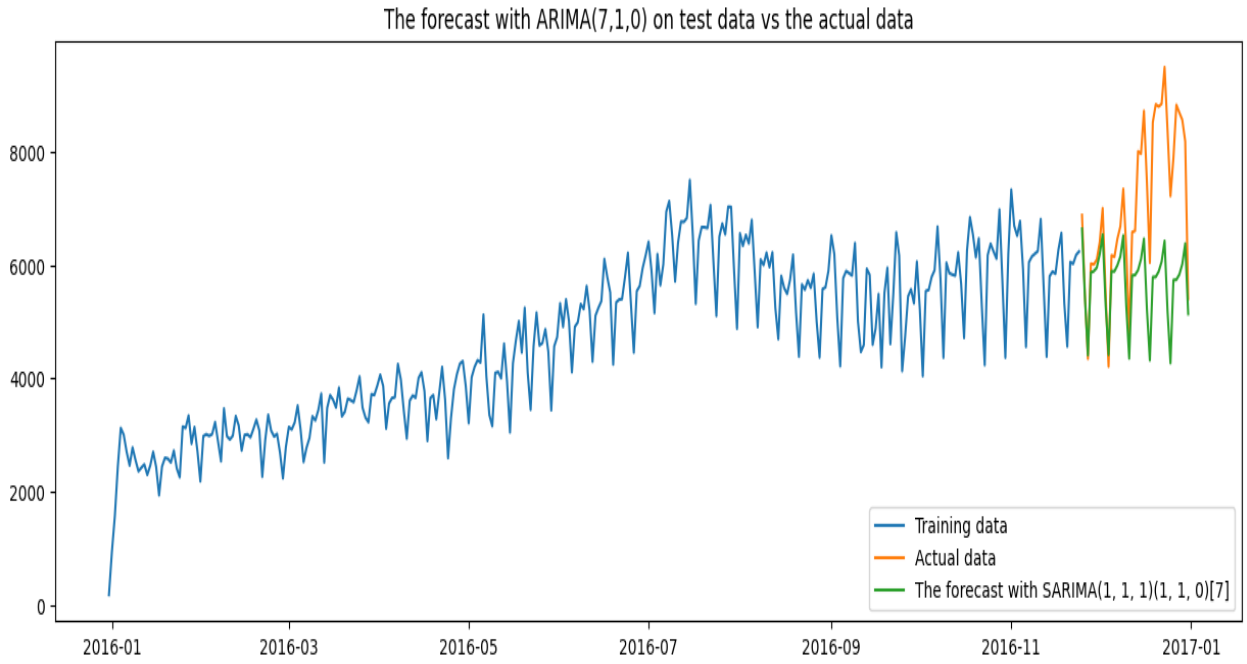

Figure 6

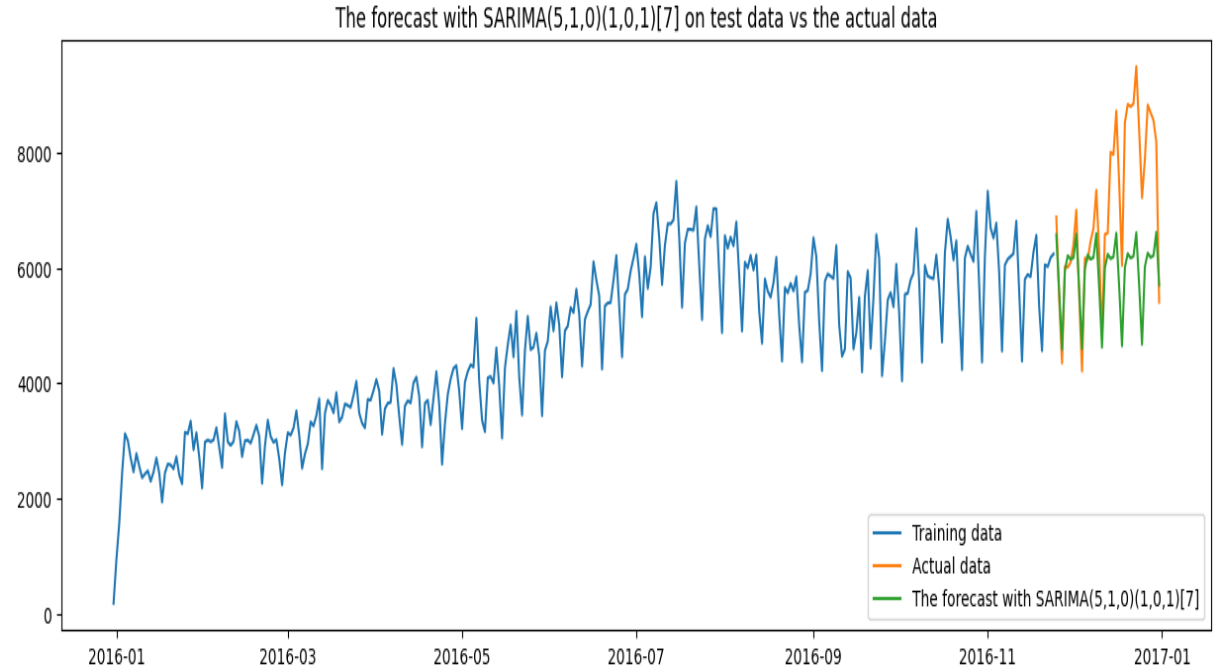

Figure 7

In Figure 6 and 7 , you can see the plots of forecasted values of the test data. Moreover, we looked at the MSE of both models, and as a result, the MSE of the SARIMA(5,1,0)(1,0,1)[7] model had a smaller MSE (2494755.172198142) which means it did the best forecast on the test data. Finally, we forecasted the

number of complete orders for the next two weeks with the SARIMA (5,1,0) (1,0,1) [7] model. You can see the forecast in Figure 8.

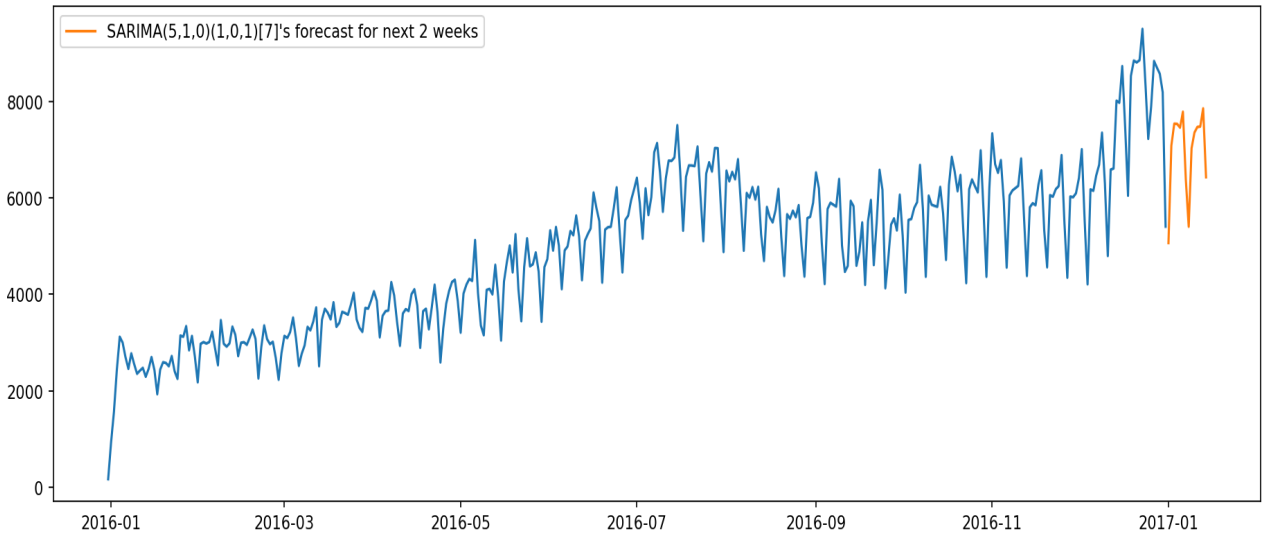

Figure 8

Holt Winter's : We have also performed Holt Winter's model. After checking several cases the group found out, that exponential smoothing with the trend set as none and seasonality set as additive was the best model. We got the forecast values of the exponential smoothing model on test data. Its MSE value was equal to 2628352.92.

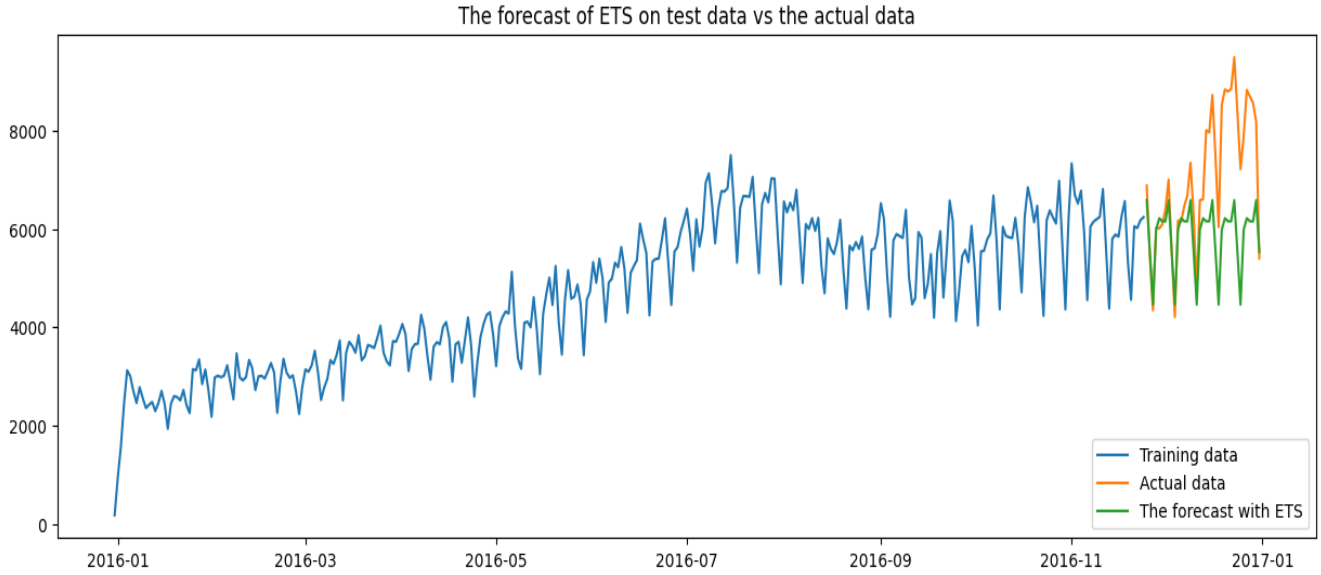

Figure 9

When we compare the EST's MSE value with SARIMA(5,1,0)(1,0,1)[7], we conclude that 2494755.17 is less than 2628352.92, so SARIMA(5,1,0)(1,0,1)[7] performs better and did a better job

on forecasting the test data than the ETS model.

In Figure 10, we see the plot of ETS forecasting completed taxi orders for the first two weeks of 2016.

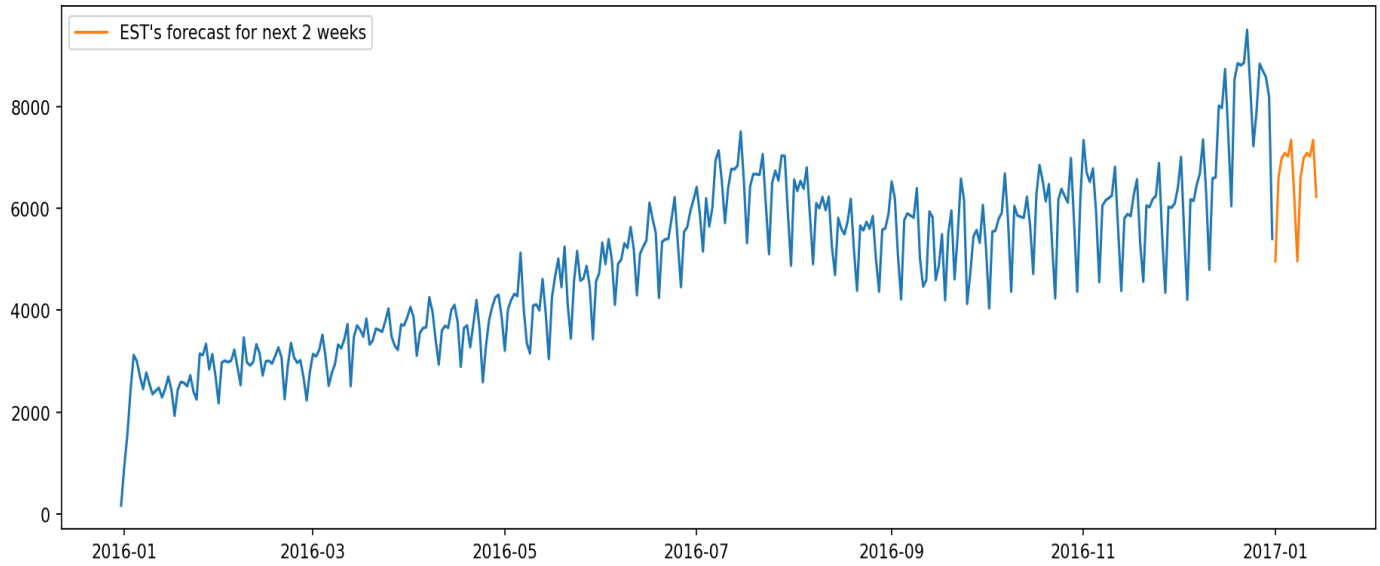

Figure 10

In Figure 11, in green color, we see the EST forecasts less number of completed taxi orders in the first two weeks of 2017 compared with the SARIMA model and based on the information the MSE of the SARIMA model while predicting the test data was less than the MSE of ETS, so we can also conclude that SARIMA(5,1,0)(1,0,1)[7] did the best job on forecasting the number of completed taxi orders of the first two weeks of 2017.

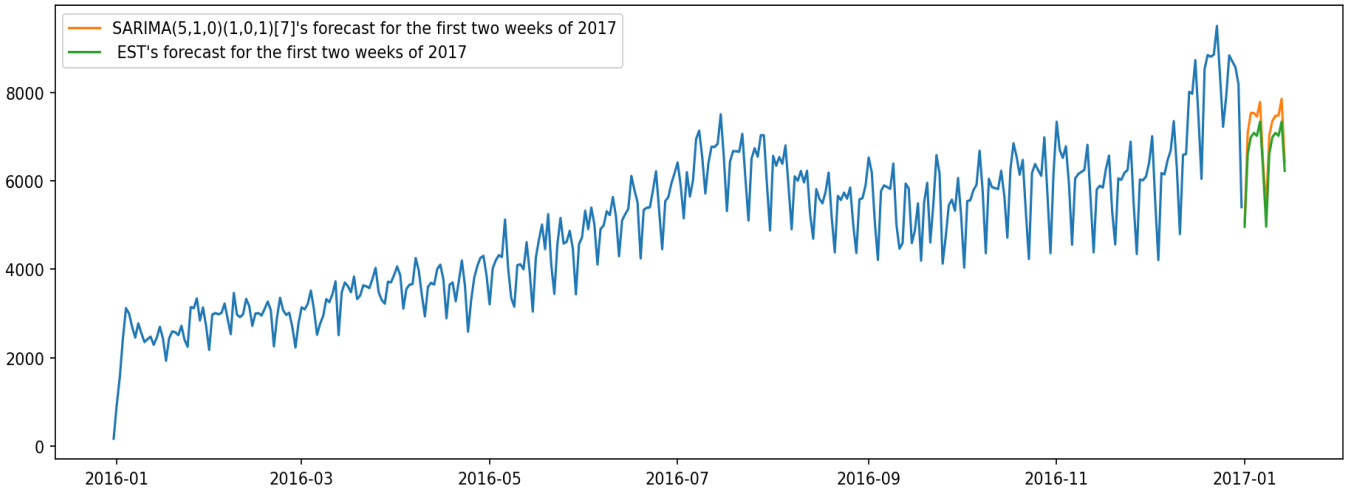

Figure 11

VAR Model

The main two variables are the number of completed orders and the number of cancelled orders. We calculated the first order difference of both variables just to make them stationary. In the figure below, one can see the plot of the percentage change of the "number of completed orders" of GG taxi over time.

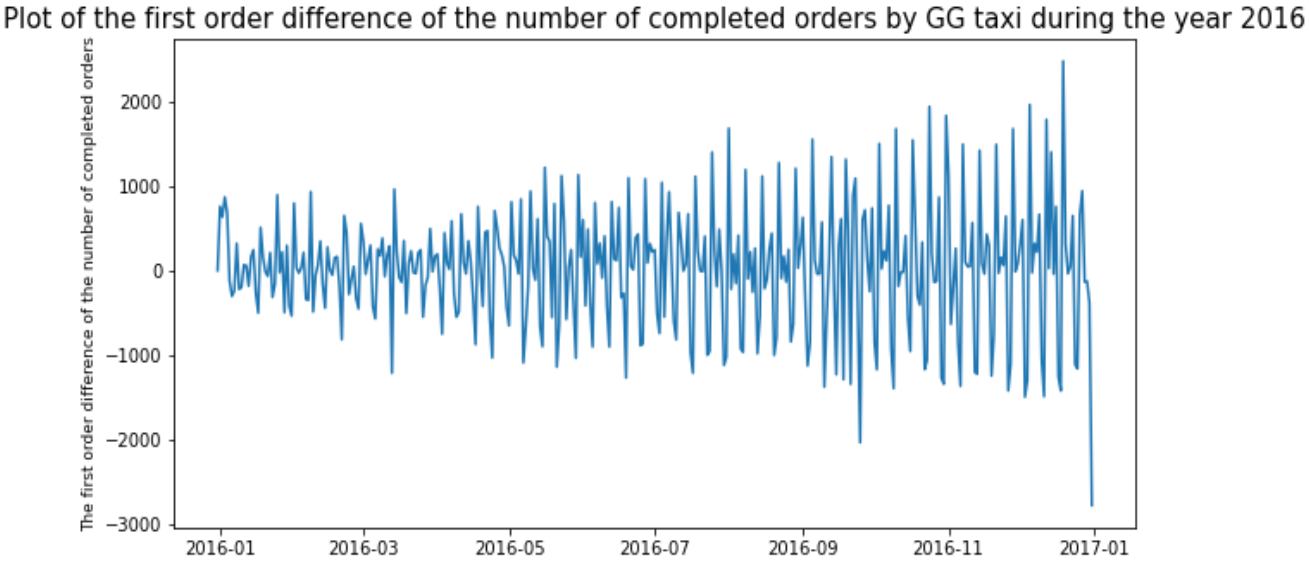

Figure 12

In Figure 12, we can understand that there is no trend and there is no seasonality. We have also conducted an ADF test to check the stationarity of the first order difference of the number of completed orders and as the result of the test showed that the new variable was stationary.

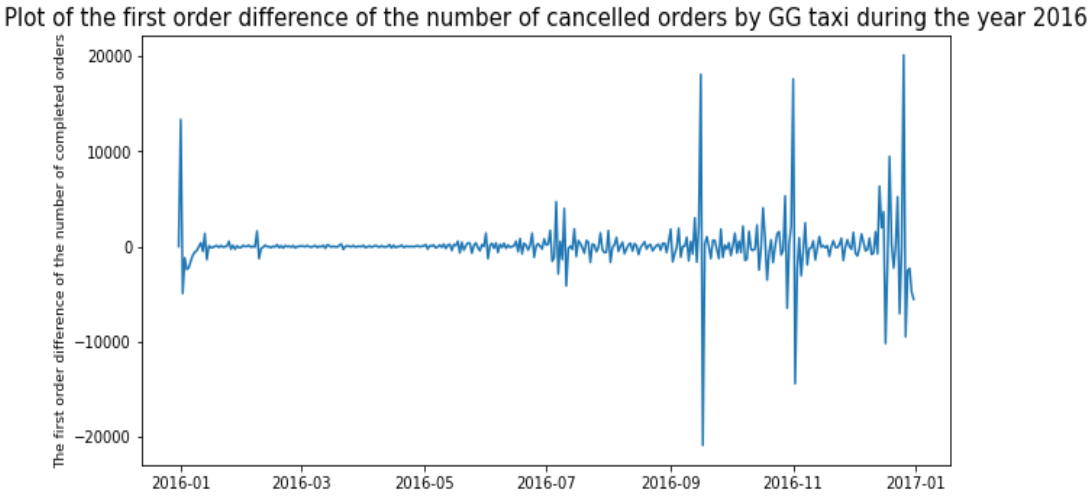

Figure 13

In Figure 13 , one can see the plot of the first order difference of the number of cancelled orders of GG taxi over time. We can claim that there is no trend and there is no seasonality. We also ran an ADF test to see if the first order difference in the number of cancelled orders was stationary, and the test's results revealed that the new variable was actually stationary.

It was crucial to find an appropriate order for the VAR model for these two variables. Thus, based on AIC (27.36) value the best order was the 7th lag. We have fitted VAR(7) on the two variables of interest. We also have done residual diagnostics for both variables. Based on the Ljung-Box test the VAR(7) model for the first order difference of the number of completed orders the residuals were not identically and independently distributed, while in case of the first order difference of the number of cancelled orders the residuals were identically and independently distributed.In the next step, our group has performed impulse response analysis of the model. Below one can see the output.

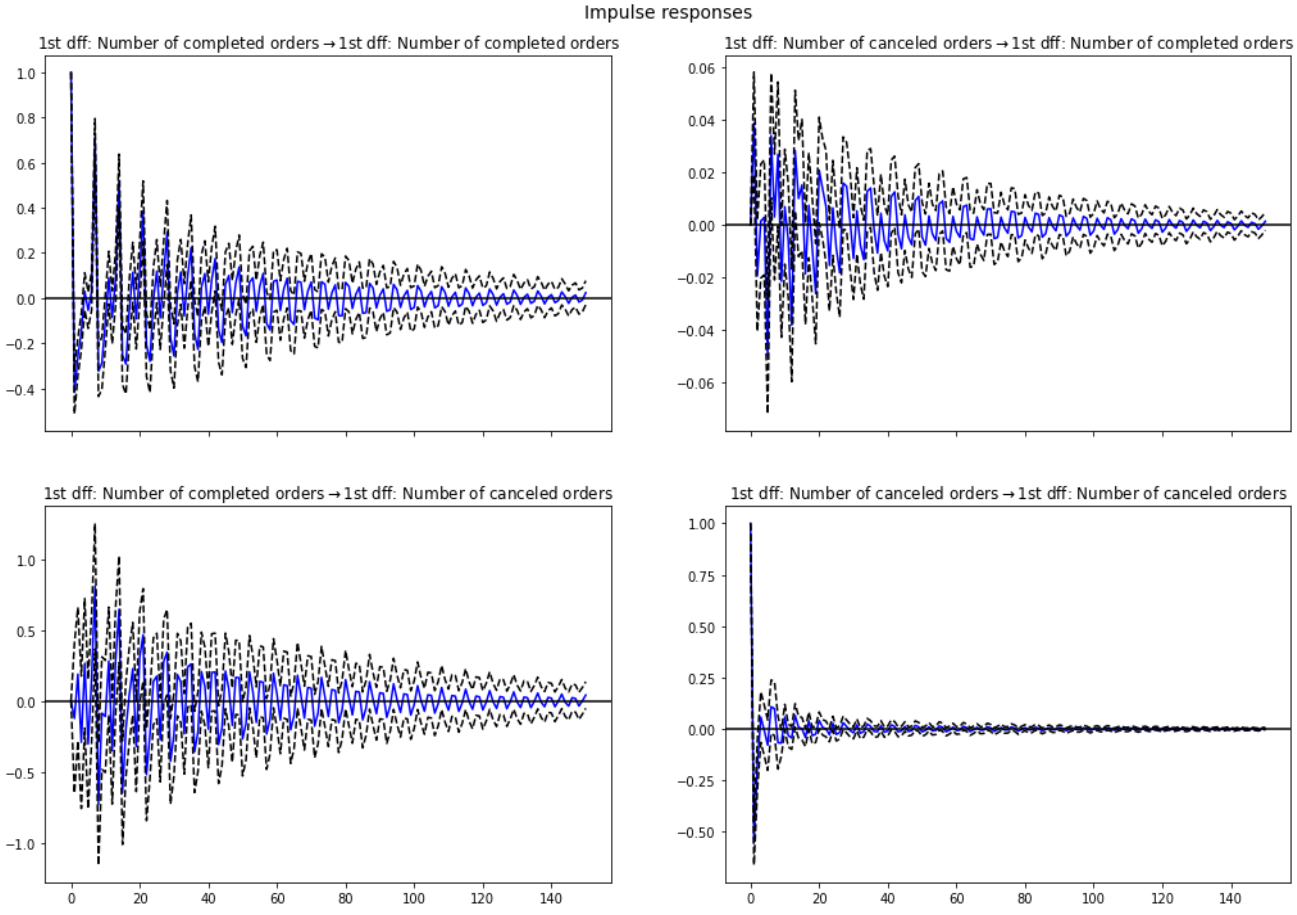

Figure 14

In the top left figure it is evident, that when one standard deviation shock is given to the “returns of the number of completed orders the returns” then the “ returns of the number of completed orders” itself

becomes very volatile, and the shock stays for a very long time and starts to fade away after the 100th lag. But in the contrary, in the bottom right plot it is obvious, that when one standard deviation shock is given to the “returns of the number of cancelled orders”, then the “returns of the number of cancelled orders” itself goes down below zero, then shows some small volatility, and disappears after the 20th lag.

From the top right plot , one can see that when one standard deviation shock was given to the “returns of the number of cancelled orders”, then the “returns of the number of completed orders” becomes so volatile, and the effects of the shock stays for a long time, and the shock begins to weaken after the 100th lag. From the bottom left plot, one can see that can see that when one standard deviation shock is given to the “returns of the number of completed orders” then the “returns of the number of cancelled orders” becomes so volatile, and the effects of the shock stays for a long time, and it starts to fade away after the 150th lag. After all, we have produced a 20 days ahead forecast with the help of the VAR(7) model. In the table below the results are presented.

	The first order difference of the number of cancelled orders forecasts	Number of completed orders percentage change
1	3927.810670	2519.258324
2	16600.745415	16285.474415
3	-6935.926401	-15455.536088
4	-3304.181769	1416.553323
5	-1644.967282	-1688.243420
6	-2318.975716	-190.940677
7	-5061.588234	-8914.450585
8	5028.316528	11242.919801
9	12459.183205	13217.200833
10	-5587.616236	-14429.319112
11	-4148.383552	-936.612253
12	-678.534358	833.047341
13	-1321.529792	-1015.061691
14	-4346.758601	-7822.354955
15	5262.990476	11801.932345
16	9466.556579	9408.640011
17	-4787.979693	-12073.138595
18	-4420.911034	-2300.202436
19	133.302786	2655.736941
20	-755.495909	-1463.214907

Conclusion

To summarize, taxi firms today are trying to use technology to gain an advantage over their competitors in both local and international markets. Taking into account this factor, our group performed a time series analysis on the data acquired from GG taxi. We designed our project by having the base of the approaches presented in the literature review. In order to predict the number of completed taxi orders for the upcoming two weeks, we applied the SARIMA and Holt Winter’s models. By the results we understood that SARIMA(5,1,0)(1,0,1)[7] did the best job on forecasting the number of completed taxi orders of the first two weeks of 2017. In the SARIMA’s forecast results, the taxi demand was actually decreased after the New Year period, which is really close to the reality. The VAR model was also used to examine the connection between the number of completed and cancelled taxi orders. To conclude the two variables were highly correlated, and the shocks stayed for a very long time.

Bibliography

1) Denis Khryashchev, Huy V., December 2019, “Predicting Taxi and Uber Demand in Cities: Approaching the Limit of Predictability.”

2) S. Faghiha, A. Shahb, Z. Wangb, A. Safikhanic, C. Kamgaa, November 2020, “Taxi and Mobility: Modeling Taxi Demand Using ARMA and Linear Regression ”