IE5202 Applied Forecasting Techniques
Chong Woon Kiat
A0209349X
Project 2: Forecasting Highway Car Volumes

1. Data Exploration

Given 5 years information of holiday, temperature, rain amount, snow amount, cloud percentage, weather, we aim to build a model to predict the traffic volume for some missing data in 2018.

Observed in the train dataset that some hourly data are missing and other duplicate in hourly data when there is a change in weather within the hour. To create an evenly spaced dataset, a resampling is done by aggregating the duplicate to mean, and filling missing hourly data with preceding information with underlying assumption that the weather information has not changed.

Another large portion of data is missing from 2014 to 2015, as shown in figure 1. 2014-08-01 to 2015-08-01 is removed from the dataset. One year is removed to maintain continuity of season in the data. Some data beyond 2018 appears as 0 traffic (missing data to be predicted).

Holiday information is also converted to Boolean (True or False).

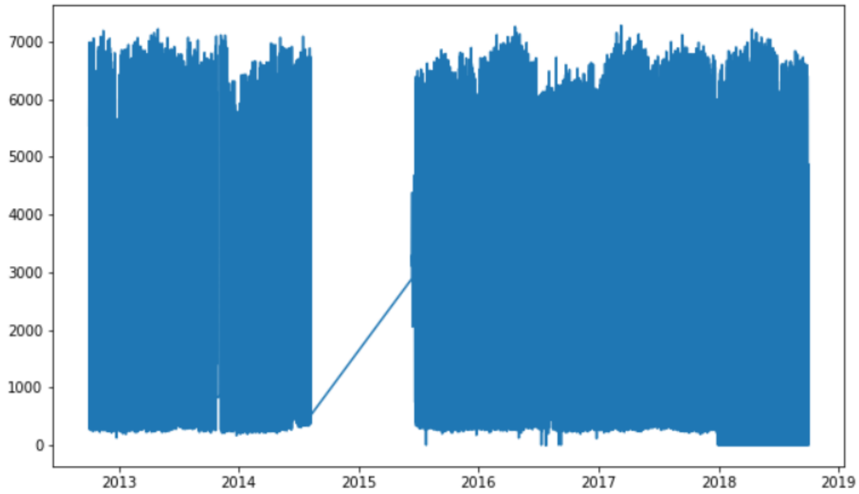
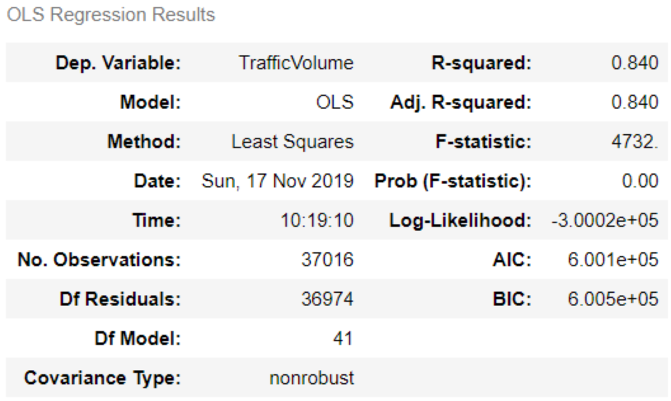


Figure 1: Traffic Volume Trend

2. Regression on Time

Linear Regression is built with predictors – Month, Hour, Isweekend, Weekday and trend, where trend is increasing hourly counter. Forward selection method is chosen to pick the predicted by minimizing AIC. Final model is TrafficVolume ~ C(Hour) + C(Weekday) + C(Weekday) + time.



OLS Regression Results

Dep. Variable:	TrafficVolume	**R-squared:**	0.840
Model:	OLS	**Adj. R-squared:**	0.840
Method:	Least Squares	**F-statistic:**	4732.
Date:	Sun, 17 Nov 2019	**Prob (F-statistic):**	0.00
Time:	10:19:10	**Log-Likelihood:**	-3.0002e+05
No. Observations:	37016	**AIC:**	6.001e+05
Df Residuals:	36974	**BIC:**	6.005e+05
Df Model:	41		
Covariance Type:	nonrobust		

Noticed from coefficients of hour, 7am and 5pm has the highest traffic volume with p-values of 0. Sunday has the least traffic volume with p-value of 0. Traffic also increased over time (coefficient of 0.0008/hr), due to increased ownership of car over time. The model has R-squared of 0.84 and adjusted R2 of 0.84, indicating that the model are able to explain 84% of the variability in traffic volume.

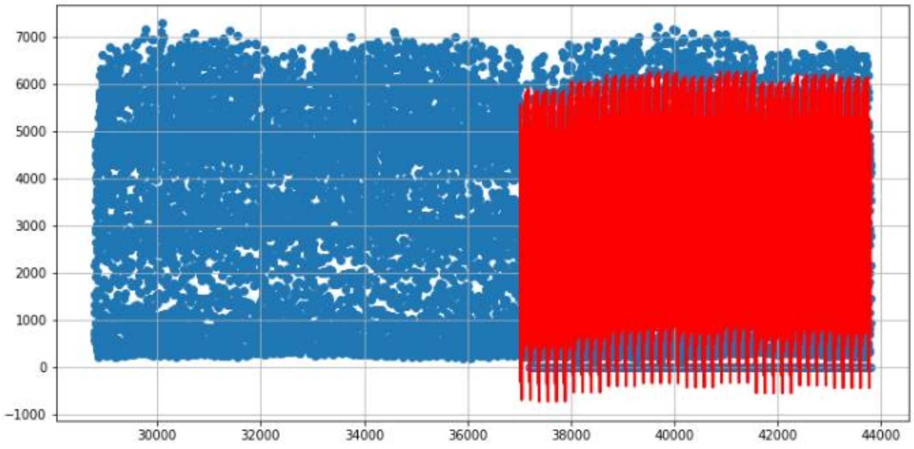


Figure2: Actual traffic volume (train and test dataset) versus predicted traffic volume (test dataset)

Obviously linear regression alone is not the best model for as time series data often has dependency on the value or noise from preceding data which is not taken into account in this model.

3. Exponential Smoothing

With Exponentially weighted moving average, the optimal alpha, is found to be 1 which minimal root mean square error, implying that the forecast value is exactly the same as last observed value.

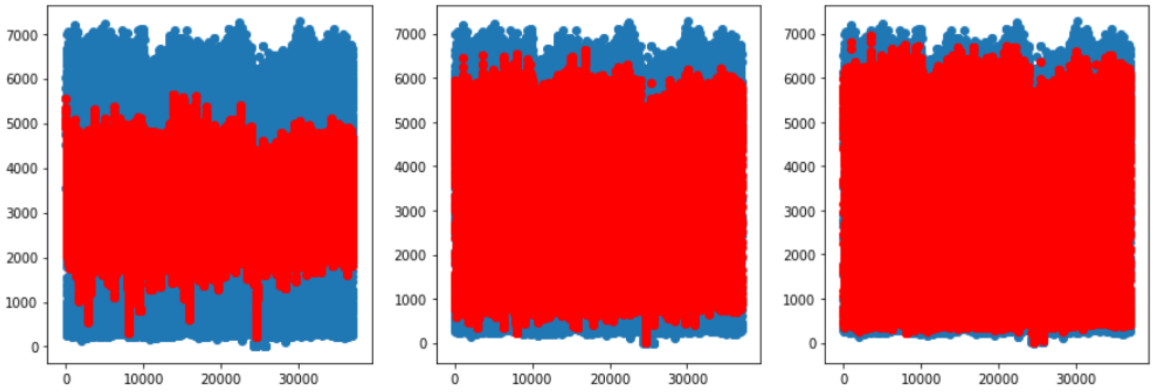


Figure 3: Predicted Traffic (Red) versus Actual (Blue) with alpha value of 0.1, 0.3, 0.5

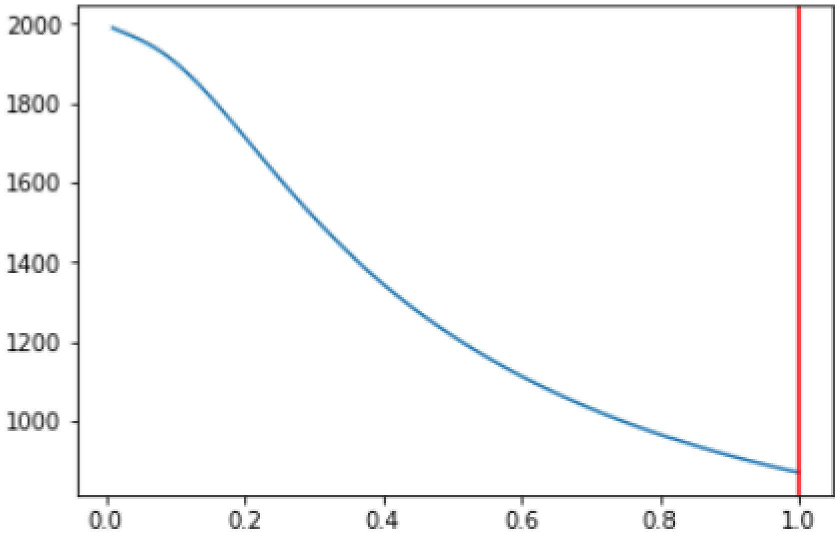


Figure 4: RMSE versus alpha - Optimal alpha (red)

With Holt's linear trend method, similar result for alpha is observed, 1, where beta is 0.01 which is, indicating that there is minimal linear trend in the dataset.

The drawback of these two method is that seasonal effect is not taken into account, which caused the alpha to be 1. Holt-Winter's seasonal method is a better choice of model for this dataset. However, there are multiple seasonality effects have to be taken care of (day, week, month).

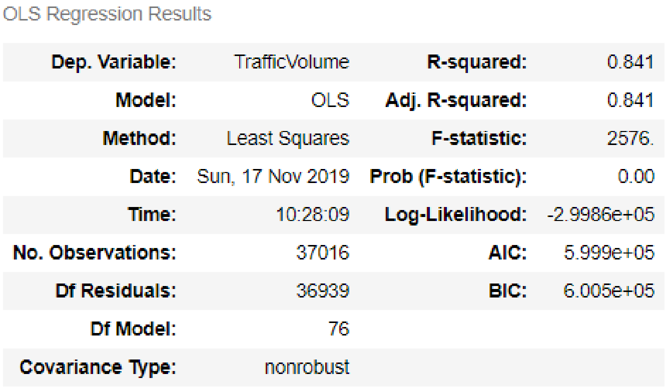
4. Free Form Forecasting

4.1. Linear Regression on predictors

Similar to part two, a linear regression model is built but with more non-time related predictors that can possibly influence traffic volume. Potential candidates are Isholiday, Month, Hour, Weekday, Isweekend, Temp, Rain, Snow, Cloud and Weather.

Forward selection method is used to pick the final model:

TrafficVolume ~ C(Hour) + Weekday + C(Month) + WeatherDescription + Temp + CloudsAll'



OLS Regression Results

Dep. Variable:	TrafficVolume	**R-squared:**	0.841
Model:	OLS	**Adj. R-squared:**	0.841
Method:	Least Squares	**F-statistic:**	2576.
Date:	Sun, 17 Nov 2019	**Prob (F-statistic):**	0.00
Time:	10:28:09	**Log-Likelihood:**	-2.9986e+05
No. Observations:	37016	**AIC:**	5.999e+05
Df Residuals:	36939	**BIC:**	6.005e+05
Df Model:	76		
Covariance Type:	nonrobust		

The model has R-squared and adj R-squared of 0.841, slightly higher than regression model on time, indicating that 84.1% of the variability is explainable by this model. The residual of the model is then fed into a time series model.

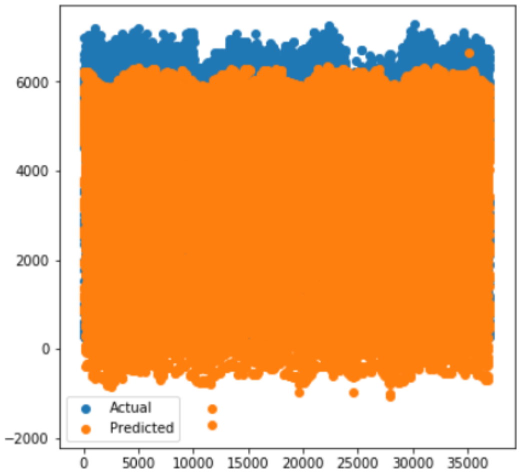


Figure 5: Actual versus predicted traffic volume of training dataset.

4.2. SARIMA model on residual from linear regression

Residual of the first 1000 and 100 data points from the linear regression model are plotted. It is obvious that the data is non-stationary, with 24-hour seasonal trend. A first order seasonal differencing of 24 hours is applied to the residual.

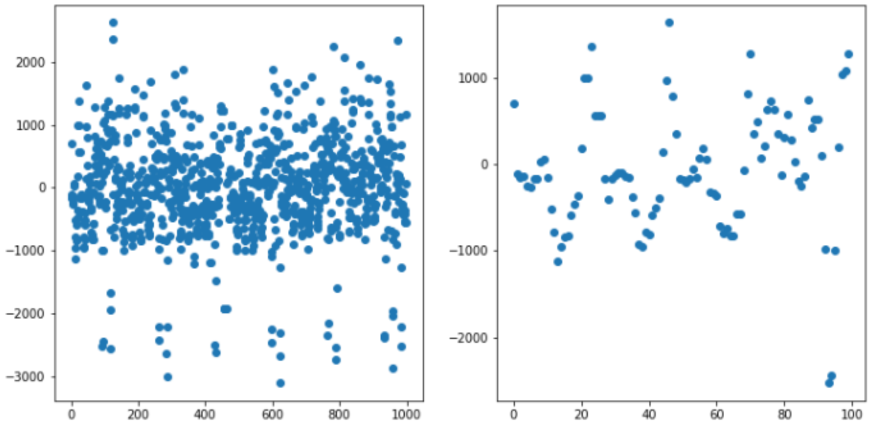


Figure 5: Residual from linear regression model

This data also appears to be non-stationary, additional first order non-seasonal differencing is applied.

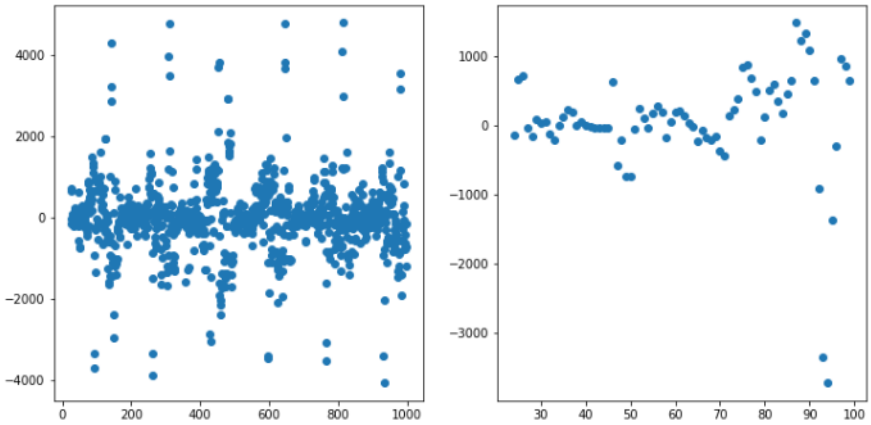


Figure 6: First order 24h seasonal differencing

The data appears more stationary after one seasonal and one non-seasonal differencing.

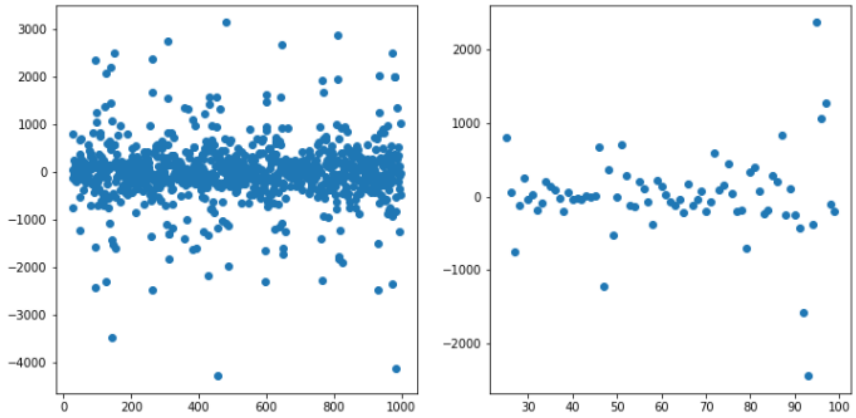


Figure 7: First order non-seasonal and first order 24h seasonal differencing

ACF and PACF of the differentiated residual is plotted.

Observed that for non-seasonal peaks, both ACF and PACF do not cut off at any lags, indicating that there is both AR and MA component in the dataset.

For seasonal peaks, PACF dies down and ACF cuts off after two peaks at 24 and 48, suggesting an seasonal MA(2) model.

A model is first built with ARIMA(1,1,1)(0,1,2)$_{24}$

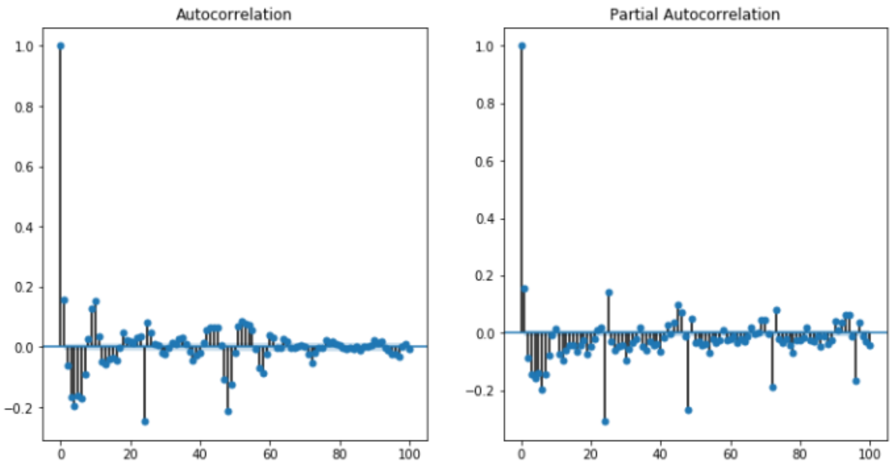
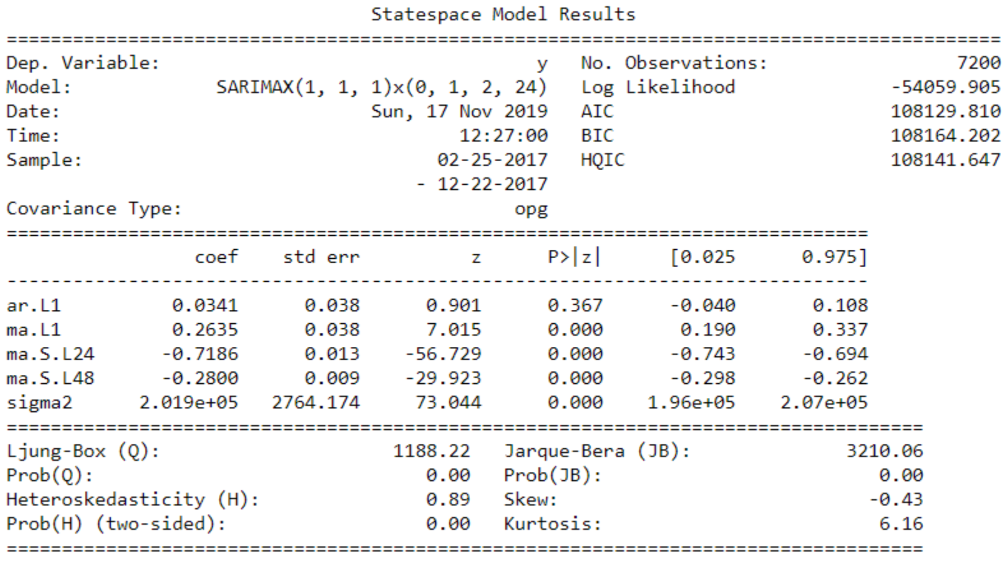


Figure 8: ACF and PACF



```
                                 Statespace Model Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                 7200
Model:             SARIMAX(1, 1, 1)x(0, 1, 2, 24)   Log Likelihood              -54059.905
Date:                            Sun, 17 Nov 2019   AIC                         108129.810
Time:                                    12:27:00   BIC                         108164.202
Sample:                                02-25-2017   HQIC                        108141.647
                                     - 12-22-2017
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0341      0.038      0.901      0.367      -0.040       0.108
ma.L1          0.2635      0.038      7.015      0.000       0.190       0.337
ma.S.L24      -0.7186      0.013    -56.729      0.000      -0.743      -0.694
ma.S.L48      -0.2800      0.009    -29.923      0.000      -0.298      -0.262
sigma2      2.019e+05   2764.174     73.044      0.000    1.96e+05    2.07e+05
===================================================================================
Ljung-Box (Q):                     1188.22   Jarque-Bera (JB):              3210.06
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.89   Skew:                            -0.43
Prob(H) (two-sided):                  0.00   Kurtosis:                         6.16
===================================================================================
```

The model is applied on test dataset. Predicted traffic volume is calculated by adding up predicted value from linear regression model and SARIMA model.

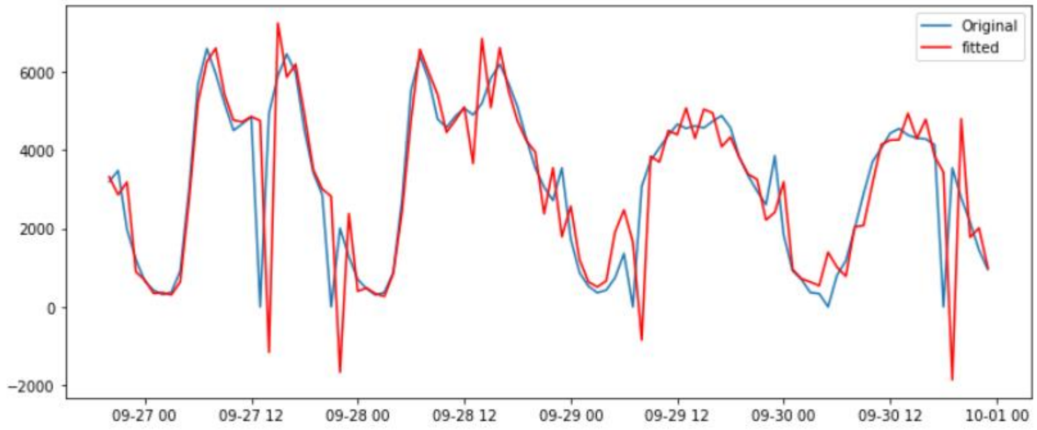


Figure 9: Last 100 predicted and actual traffic volume in test dataset

The RMSE of the model in test dataset (excluding missing data points to be predicted) is 1188.4

From QQ plot in figure 10 below, it is observed that traffic tends to be over or under predicted at extreme edge. This suggests that the linear regression model and SARIMA model can be further optimized.

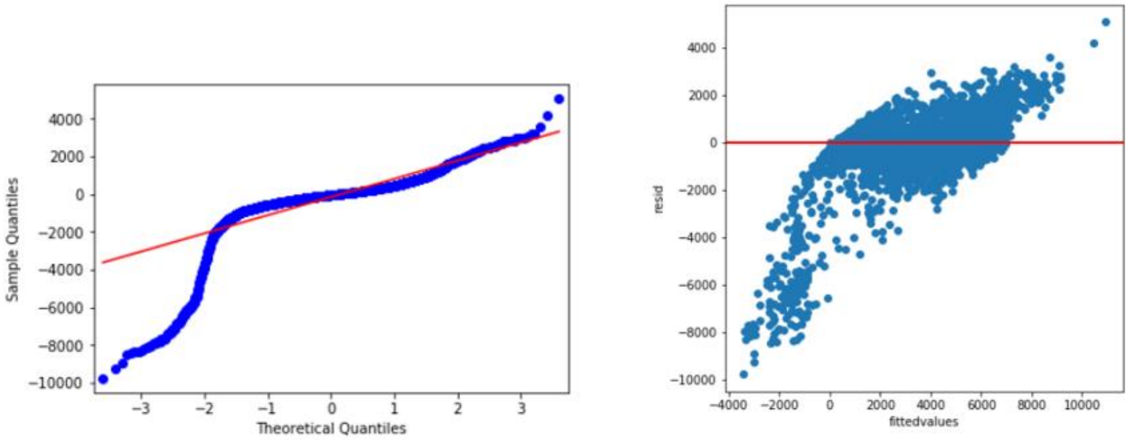


Figure 10: left - QQ plot of residual, right - Residual vs Predicted value

4.3. Model Tuning

Parameters of SARIMA model can then be optimized by changing the p and q value, the model is then fitted into test data and RMSE is evaluated, the optimal model should give the lowest RMSE.