The Metro Interstate Traffic Volume Prediction

Instructor: Dr. Reza Jafari
By: Madhuri Yadav

Problem Statement

To Predict the The Metro Interstate Traffic Volume Prediction using Time Series Prediction models.

Applications:

1. Increase the efficiency and life of roads
2. Reduces traffic volume at a particular section
3. Provide better means for development of infrastructures
4. Provide better means to utilize other roads in case of special events in the city
5. Provide estimate of no vehicles against no of persons
6. During the given pandemic situations where social distancing has higher priority.

- Source: UCI Machine Learning Repository
- Shape: 433845 rows × 9 columns
- Important features:
  - holiday
  - temp
  - rain_1h
  - snow_1h
  - clouds_all
  - weather_main
  - weather_description
  - date_time
  - **traffic_volume**

Dataset

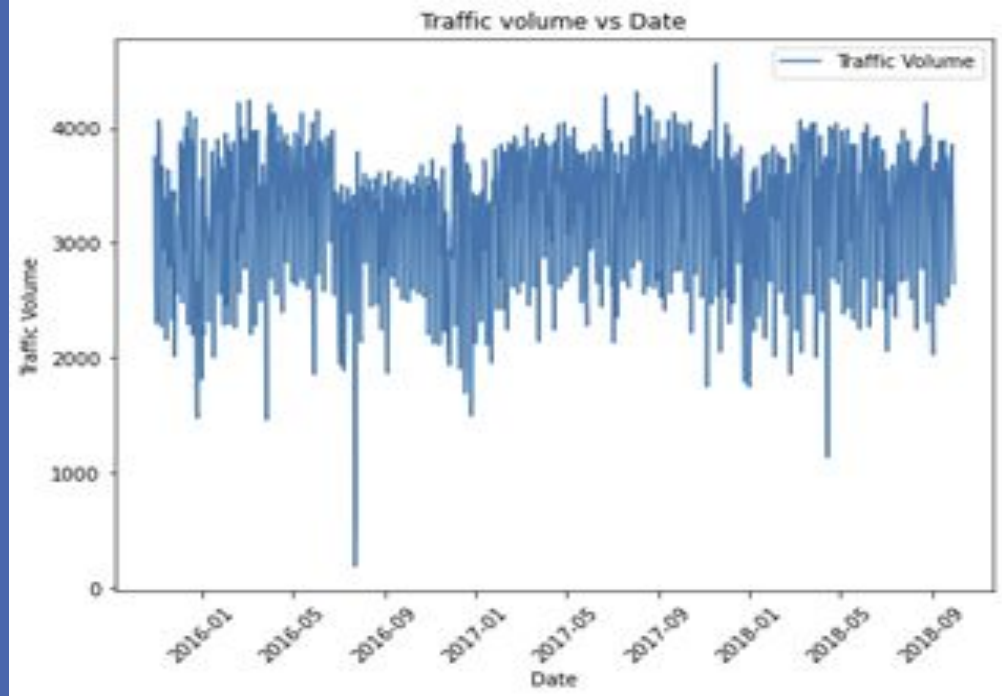
The Metro Interstate Traffic Volume

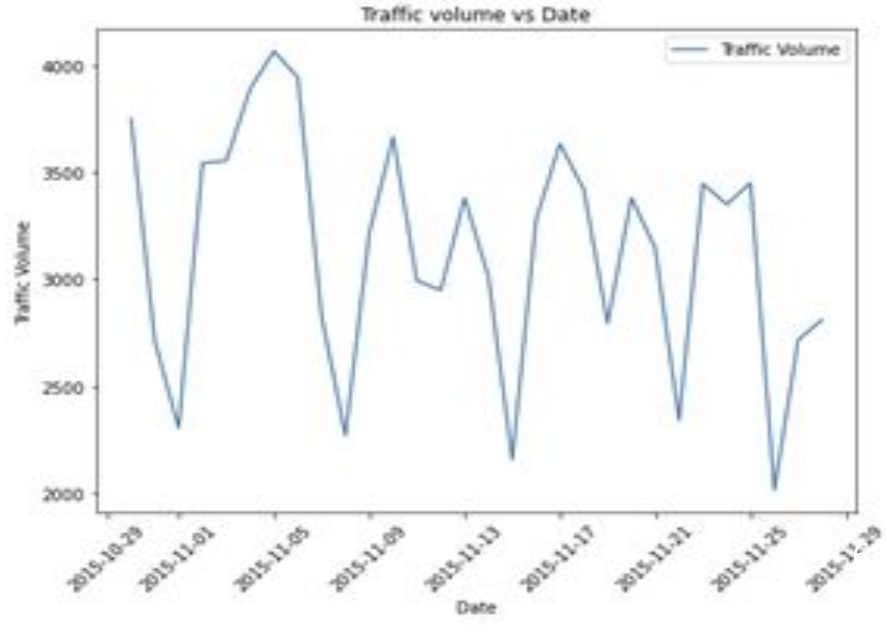
Data Preprocessing

1. Replacing None values with the nan.
2. "date_time" column is converted to datetime format
3. Aggregation of the data based on the date
4. Holiday data column is then converted to two values Yes and No.
5. Drizzle, Haze, Thunderstorm, Fog are changed to Rain and Mist accordingly foe weather column
6. Dataset Shape : 1067, 7

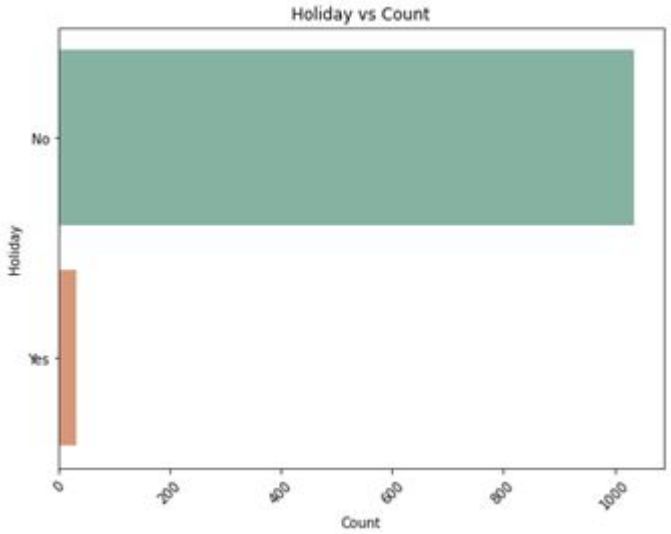
EDA 1

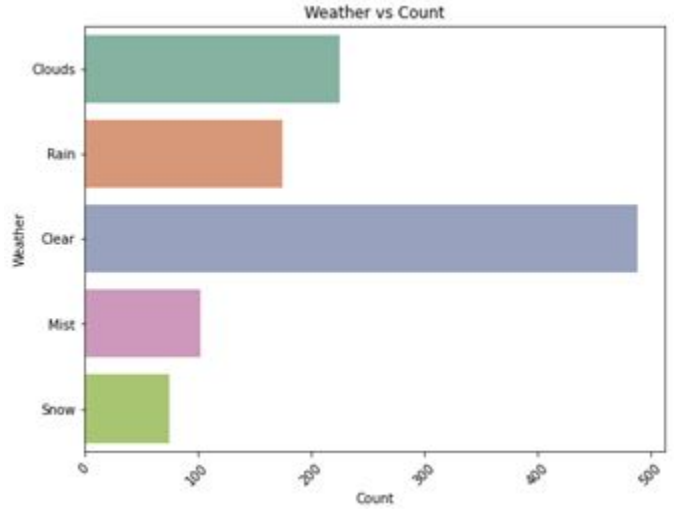
Distribution of target variable.

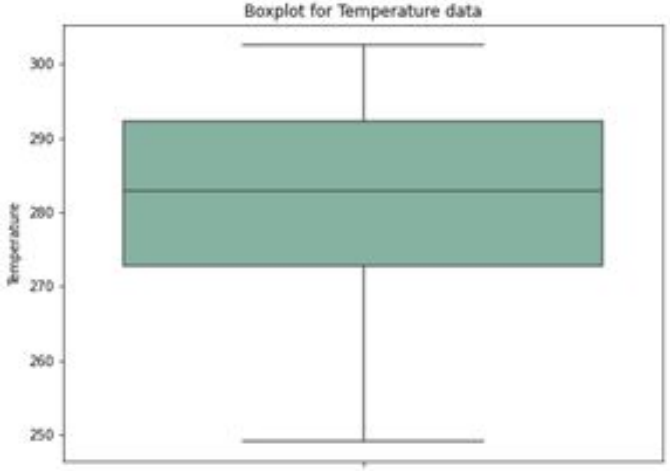




EDA 2

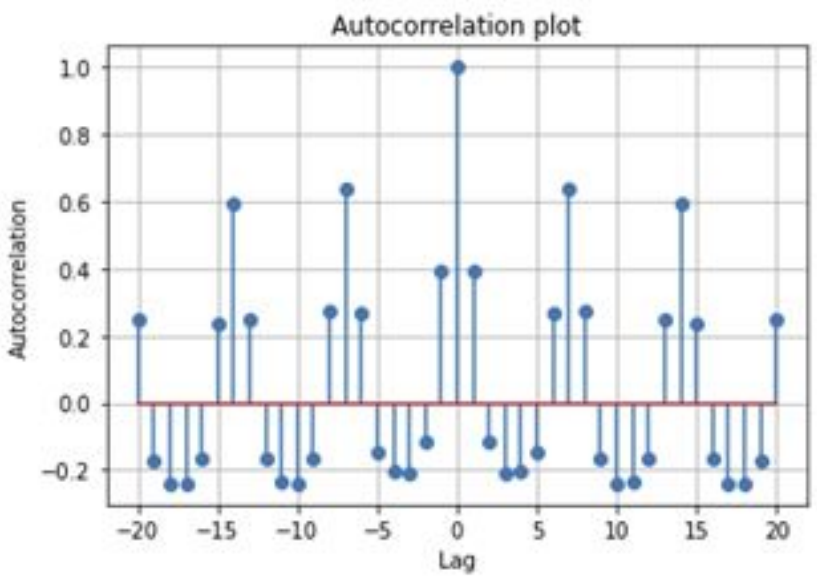

Holiday vs Count
Holiday
No
Yes
0
200
400
600
800
1000
Count

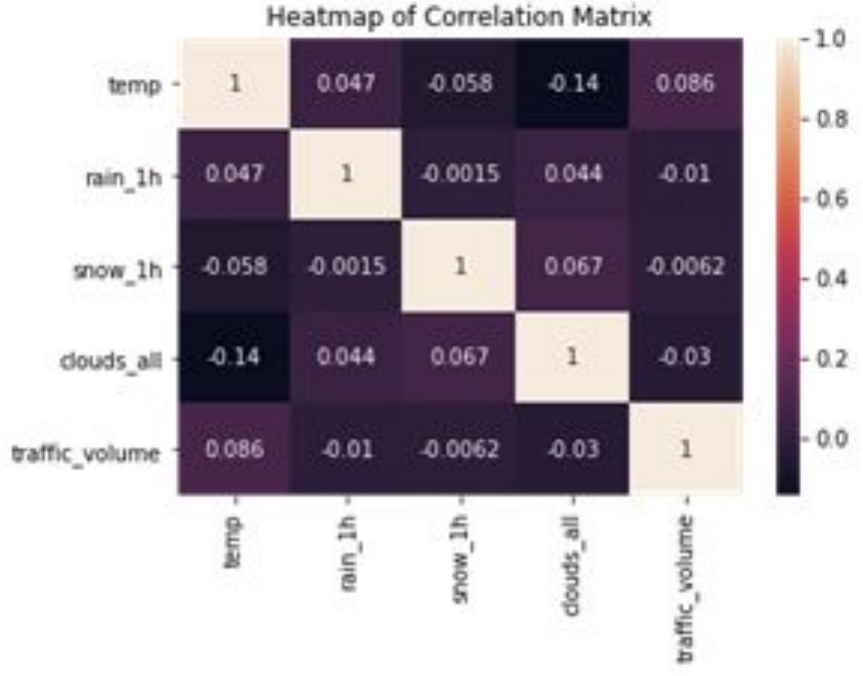

Weather vs Count
Weather
Clouds
Rain
Clear
Mist
Snow
0
100
200
300
400
500
Count


Boxplot for Temperature data
Temperature
300
290
280
270
260
250


Boxplot for Clouds data
Clouds
80
60
40
20
0

EDA 3


Autocorrelation plot
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−20
−15
−10
−5
0
5
10
15
20
Lag


Heatmap of Correlation Matrix
temp
rain_1h
snow_1h
clouds_all
traffic_volume
1
0.047
-0.058
-0.14
0.086
0.047
1
-0.0015
0.044
-0.01
-0.058
-0.0015
1
0.067
-0.0062
-0.14
0.044
0.067
1
-0.03
0.086
-0.01
-0.0062
-0.03
1
1.0
0.8
0.6
0.4
0.2
0.0

Train-Test Split

- Shape of the entire dataset: (1067, 7)
- Shape of the train set : (853, 7)
- Shape of the test set : (214, 7)

Stationarity

ADF test for traffic_volume

ADF Statistic: -4.711791

p-value: 0.000080

Critical Values:

1%: -3.437

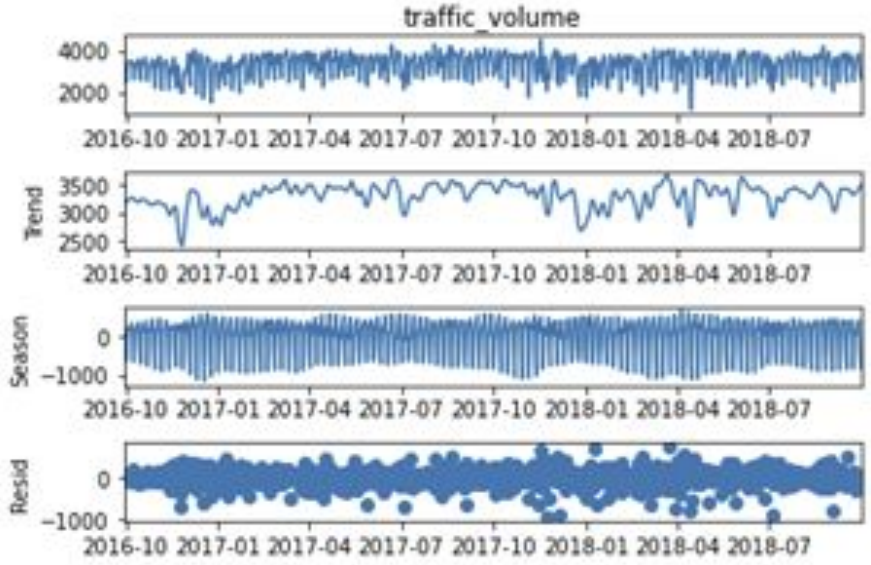
5%: -2.864

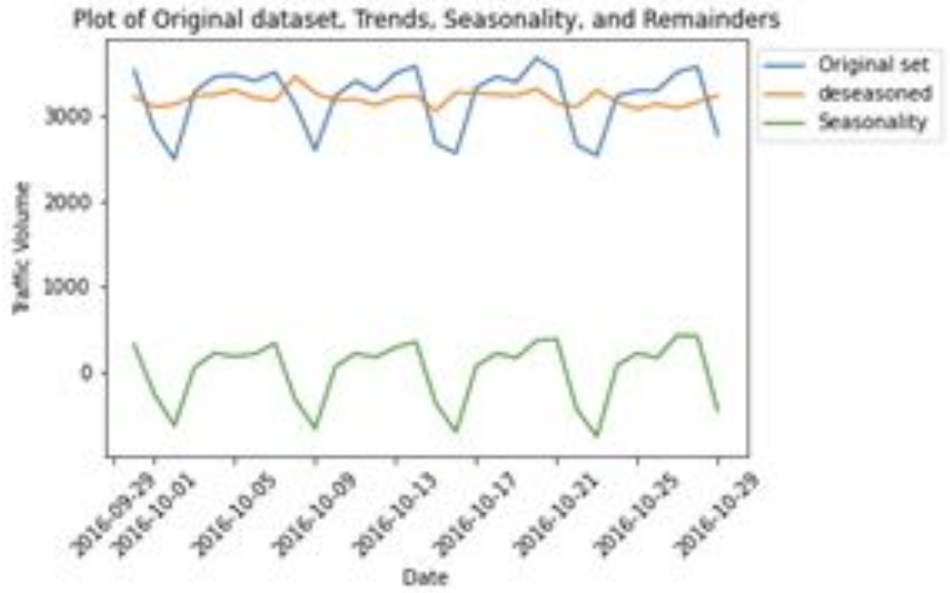
10%: -2.568

- **P-value is 0.000078 i.e. less than 0.05(95% or more confidence interval), hence we reject Null Hypothesis and conclude that the dependent variable is stationary.**

Time series Decomposition:

The strength of trend for this data set is: 0.54
The strength of seasonality for this data set is: 0.83





The summary of the model after features elimination.

OLS Regression Results

Dep. Variable:	traffic_volume	R-squared:	0.089
Model:	OLS	Adj. R-squared:	0.080
Method:	Least Squares	F-statistic:	9.443
Date:	Wed, 16 Dec 2020	Prob (F-statistic):	6.63e-10
Time:	19:47:31	Log-Likelihood:	-4506.3
No. Observations:	584	AIC:	9027.
Df Residuals:	577	BIC:	9057.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1717.2161	67.487	25.445	0.000	1584.665	1849.767
holiday_No	1161.6981	53.340	21.779	0.000	1056.934	1266.462
holiday_Yes	555.5181	89.513	6.206	0.000	379.708	731.329
temp	273.4530	110.391	2.477	0.014	56.636	490.270
weather_main_Clear	326.1443	83.552	3.903	0.000	162.041	490.248
weather_main_Clouds	362.0308	92.468	3.915	0.000	180.415	543.646
weather_main_Mist	241.5281	110.285	2.190	0.029	24.919	458.137
weather_main_Rain	174.5027	101.121	1.726	0.085	-24.108	373.113

Omnibus:	39.070	Durbin-Watson:	1.296
Prob(Omnibus):	0.000	Jarque-Bera (JB):	45.334
Skew:	-0.673	Prob(JB):	1.43e-10
Kurtosis:	2.772	Cond. No.	8.19e+15

Feature selection:

"snow_1h", "rain_1h", "clouds_all",

"weather_main_Snow"

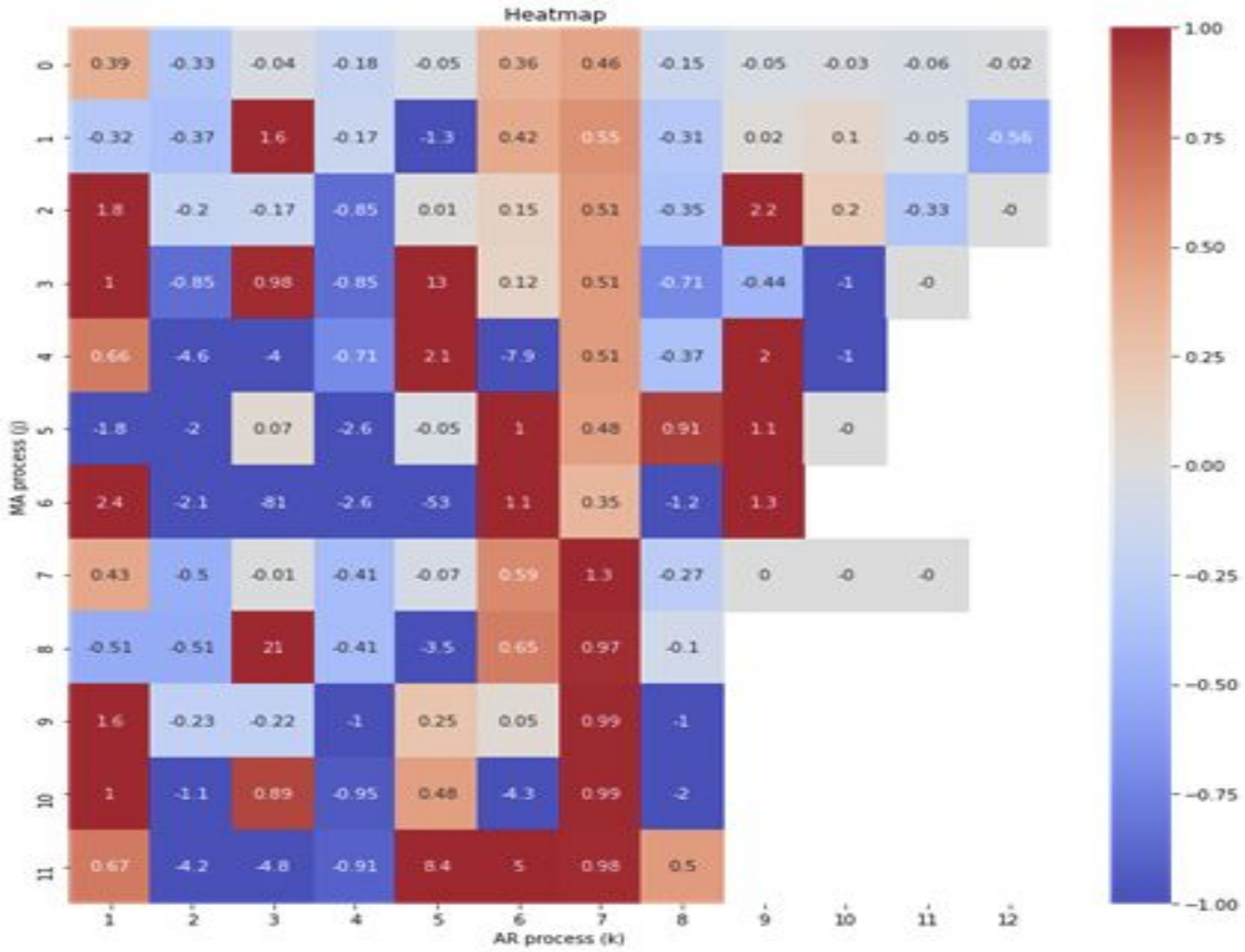
The Coefficients for the LSE model are:

const	33017.199475
holiday_No	-13654.444346
holiday_Yes	-8688.635940
temp	273.452969
weather_main_Clear	326.144272
weather_main_Clouds	362.030830
weather_main_Mist	241.528093
weather_main_Rain	174.502723

Name: traffic_volume, dtype: float64
-2451489.231002282 -6076.655145079008

Multiple Linear Regression

ARMA Model



[(2,0),(2,2),(4,0),(4,2),(4,7),(7,11),

(8,1),(8,8),(9,0),(9,5),(9,7),(10,1),

(10,3),(11,2)]

Levenberg Marquardt Algorithm:

The Model Summary is as Follows: ARMA (3,6)

Final parameters are: [-2.25 2.25 -1. -1.81 1.41 0.03 -0.69 0.38 -0.1]

[-2.2475869708739005, 2.2484706904368155, -1.0008095440306755, 0, 0, 0]

[-1.810799919627076, 1.4134001746938882, 0.026467957510958 04, -0.6864335286859218, 0.3841354061559829, -0.096063446819770 62]

Confidence Interval are:

-2.241 < a1 < -2.254

2.259 < a2 < 2.238

-0.995 < a3 < -1.007

-1.727 < b1 < -1.894

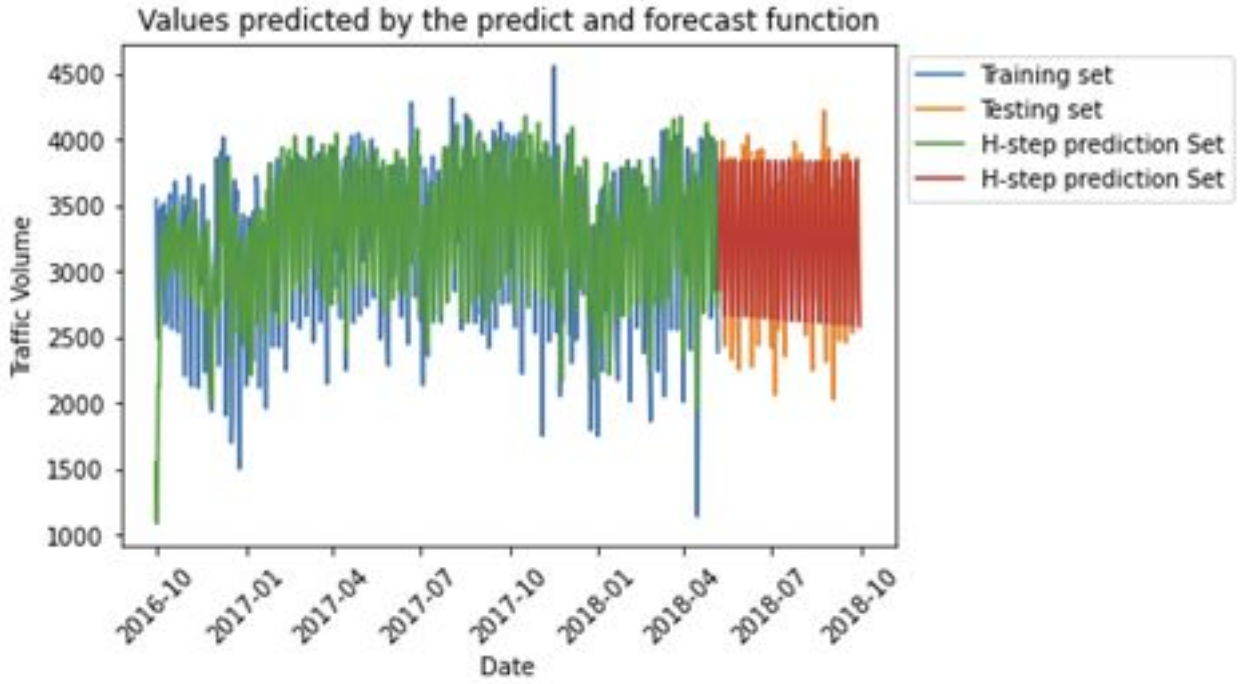
1.583 < b2 < 1.244

0.225 < b3 < -0.172

-0.488 < b4 < -0.885

0.554 < b5 < 0.214

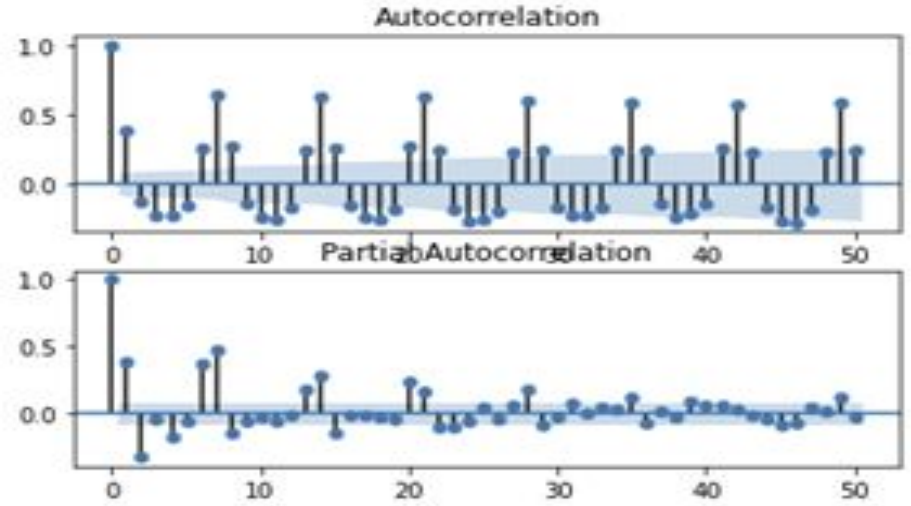
-0.013 < b6 < -0.180

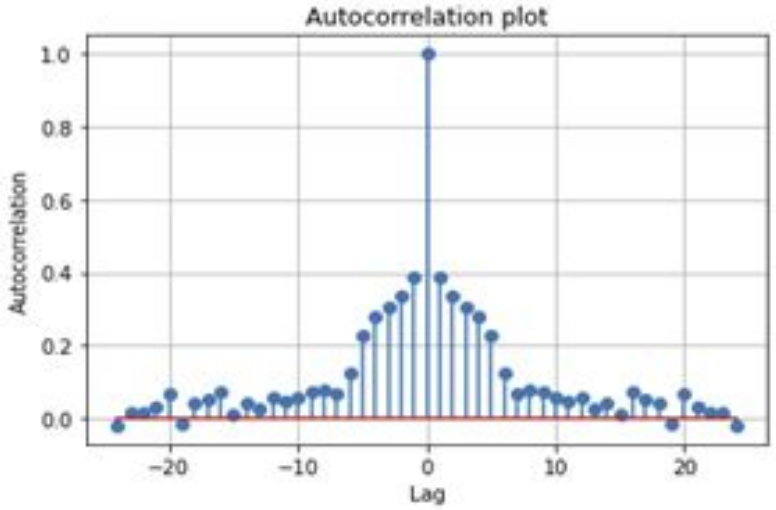


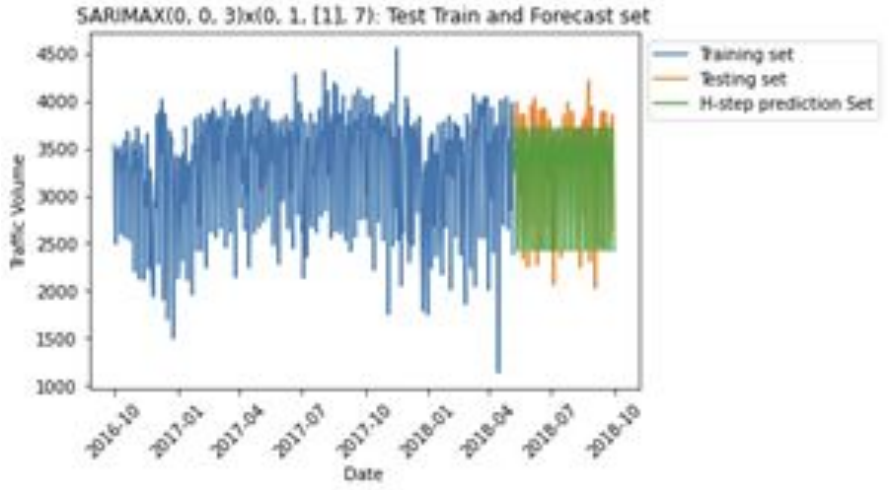
Continued...

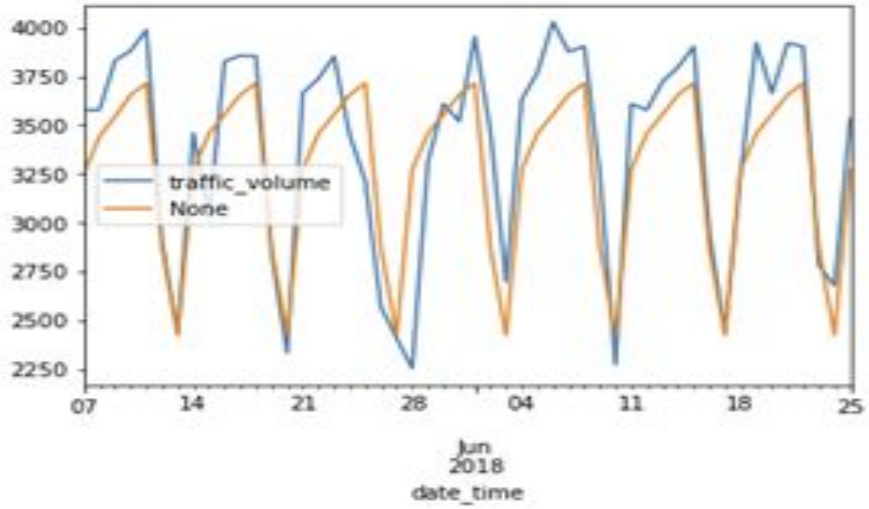
Best Model: SARIMAX

Best model: ARIMA(0,0,3)(0,1,1)[7] intercept
Total fit time: 36.500 seconds









Summary of all the Models

Methods/Values	Q-Value	R-value	MSE Prediction	MSE forecast	Mean Prediction Err	Mean Forecast Err	Var of Prediction Err	Var of Forecast Err
Average	893.88	1	327442.22	282937.54	58.03	55.48	324074.73	279859.84
Naive	644.57	1	393938.74	1191763.08	-1.98	954.94	393934.82	279859.84
Drift	641.83	0.99	398446.03	1497497.77	2.92	1101.43	398437.49	284349.23
SES	993.41	1	377483.02	492899.34	-2.27	461.56	377477.89	279859.84
Holt-Linear	937.24	1	316945.19	280808.81	8.99	30.81	316864.41	279859.84
Holt-Winter	84.35	0.41	124792.18	106484.61	-5.19	-117.35	124765.22	92714.55
SARIMA	296.25	0.56	222608.94	88359.25	93.23	62.67	213917.96	84431.72
ARMA(3,6)	62.5	0.68	184160.18	143525.41	8.68	8.95	184084.86	143445.32
ARMA(5,5)	14.32	0.43	148895.8	98287.35	5.85	-23.9	148861.59	97716.32
ARMA(5,7)	11.85	0.46	143495.5	95854.1	6.97	5.9	143446.93	95819.24

Conclusion & Challenges

- Since our data has strong seasonality of 0.83 as well as performed SARIMAX model gives the best results of all the other models we choose SARIMAX as our best model with MSE = 88359.25
- We could also perform further analysis of our model and implement SARIMA model to get the better results for prediction of our dataset. Also future work would include to work with prediction and forecast function for SARIMA.