Metro Traffic Volume Analysis and Prediction

Time Series Analysis and Modelling
Final Term Project

Jiwoo Suh

Contents

Data Description

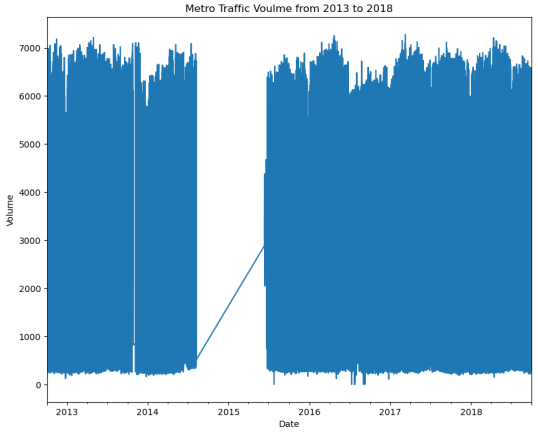
```
Int64Index: 40575 entries, 0 to 48203
Data columns (total 9 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   holiday              40575 non-null  object
 1   temp                 40575 non-null  float64
 2   rain_1h              40575 non-null  float64
 3   snow_1h              40575 non-null  float64
 4   clouds_all           40575 non-null  int64
 5   weather_main         40575 non-null  object
 6   weather_description  40575 non-null  object
 7   date_time            40575 non-null  object
 8   traffic_volume       40575 non-null  int64
dtypes: float64(3), int64(2), object(4)
memory usage: 3.1+ MB
None
```

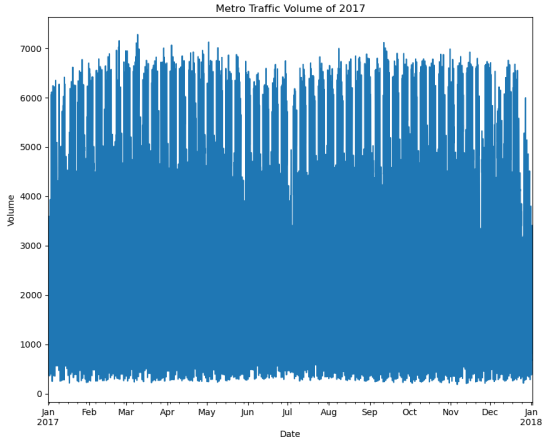
```
In [31]: print(data.head().to_string())
  holiday    temp  rain_1h  snow_1h  clouds_all weather_main weather_description            date_time  traffic_volume
0    None  288.28      0.0      0.0          40       Clouds    scattered clouds  2012-10-02 09:00:00            5545
1    None  289.36      0.0      0.0          75       Clouds      broken clouds  2012-10-02 10:00:00            4516
2    None  289.58      0.0      0.0          90       Clouds     overcast clouds  2012-10-02 11:00:00            4767
3    None  290.13      0.0      0.0          90       Clouds     overcast clouds  2012-10-02 12:00:00            5026
4    None  291.14      0.0      0.0          75       Clouds      broken clouds  2012-10-02 13:00:00            4918
```

- Number of Observation: 48203
- Target Variable: traffic_volume
- Numerical features: temp, rain_1h, snow_1h, clouds_all
- Categorical features: holiday, weather_main, weather_description

Data Description – Missing records

- There are some missing records: Interpolation is used
- Categorical features about Weather : df.fillna(method='ffill') method is used
- There are many duplicated or missing observations from 2014-2015 → Decided to take year 2017 only



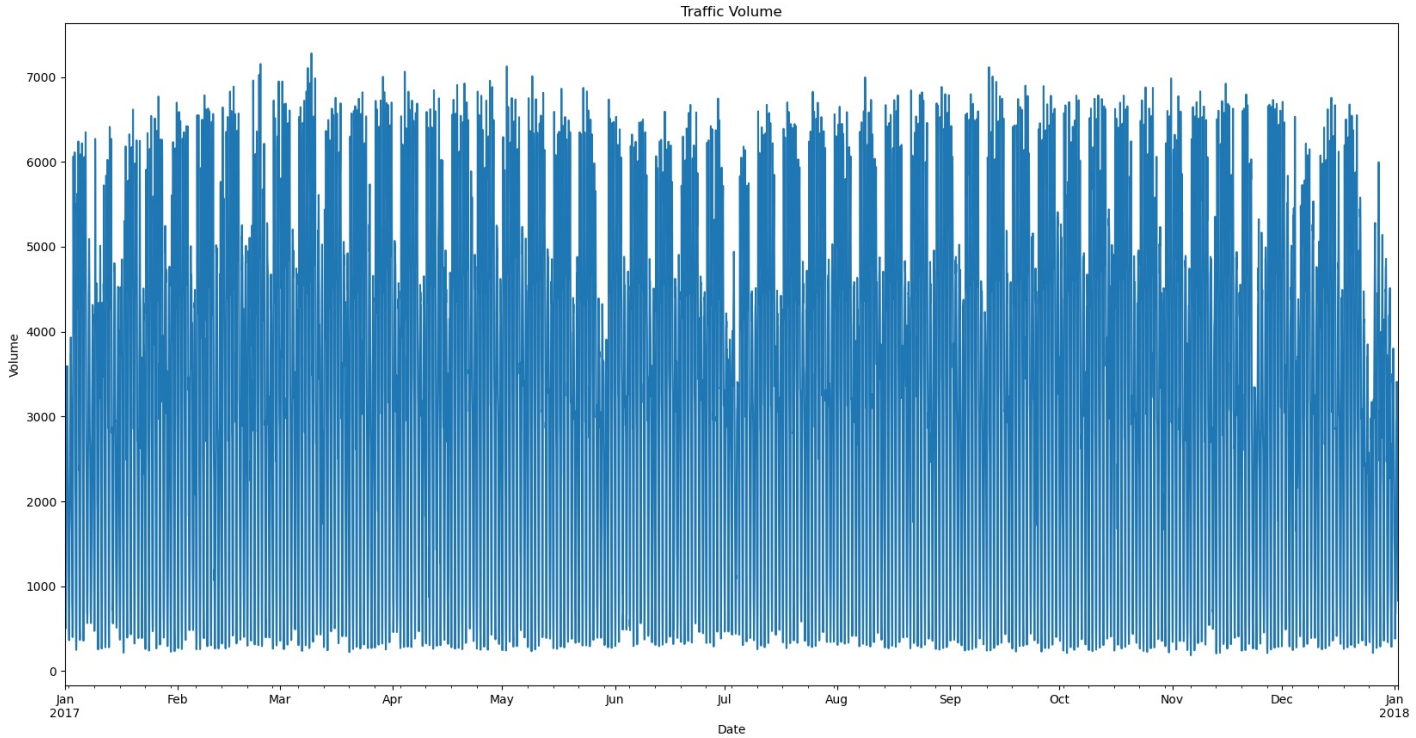


Data Description - Final Dataset(2017)

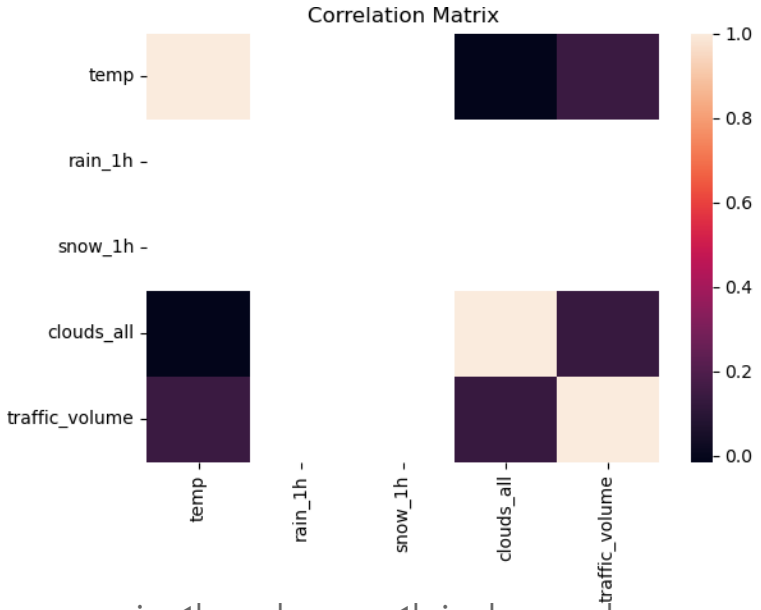
- Number of Observation: 8784
- Statistical description for Numerical features
- Every value in rain_1h, snow_1h categories at the year of 2017 is 0

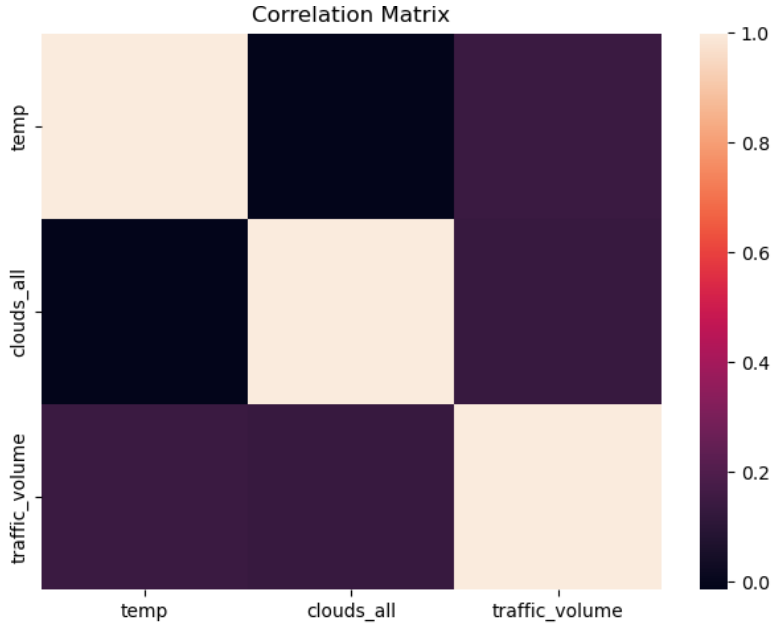
```
In [29]: df.describe()
Out[29]:
                temp  rain_1h  snow_1h   clouds_all  traffic_volume
count   8784.000000   8784.0   8784.0  8784.000000     8784.000000
mean     281.420064      0.0      0.0    43.781933     3369.585781
std       12.085219      0.0      0.0    39.161671     1981.338474
min      246.150000      0.0      0.0     0.000000      186.000000
25%      272.580000      0.0      0.0     1.000000     1351.500000
50%      282.600000      0.0      0.0    40.000000     3563.500000
75%      291.160000      0.0      0.0    90.000000     4990.000000
max      307.020000      0.0      0.0    92.000000     7280.000000
```

Data Description – Target Variable



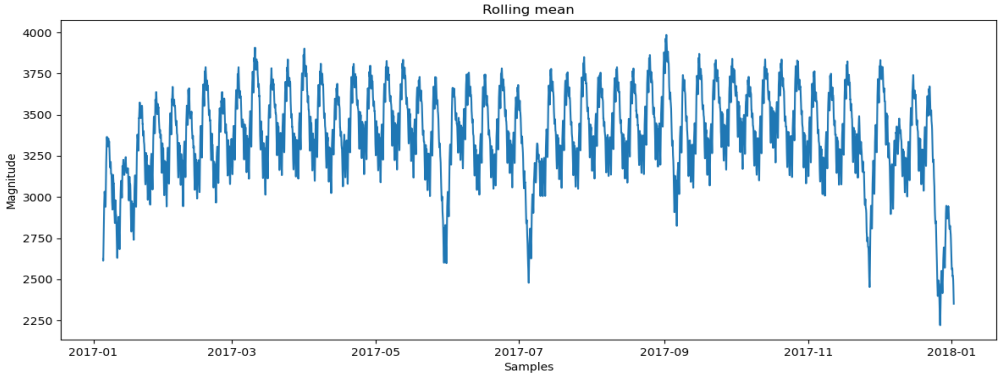
Data Description - Correlation Matrix

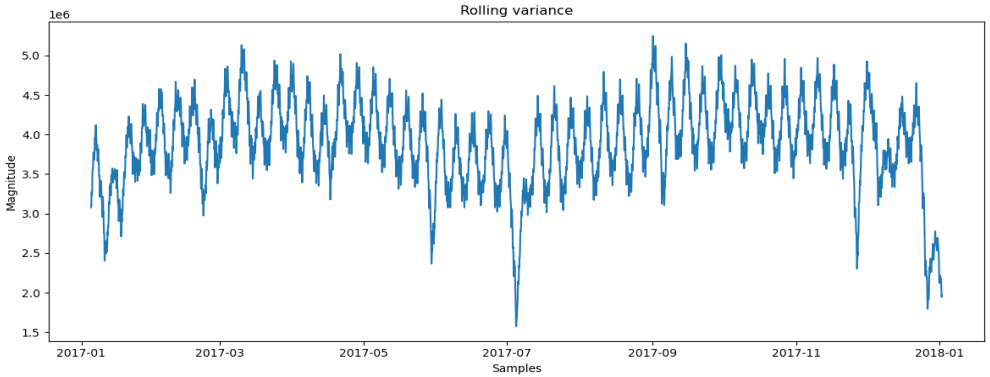




- rain_1h and snow_1h is dropped
- High correlation between numerical features and the target variable is not clearly identified
- High correlation within numerical features is not identified

Data Description – Rolling Mean & Variance





- From the plots of Rolling Mean and Variance, the dataset seems stationary

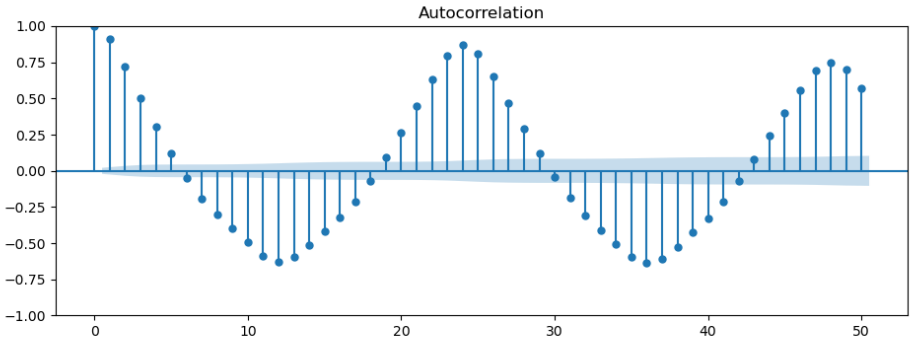
Data Description – ADF & KPSS test

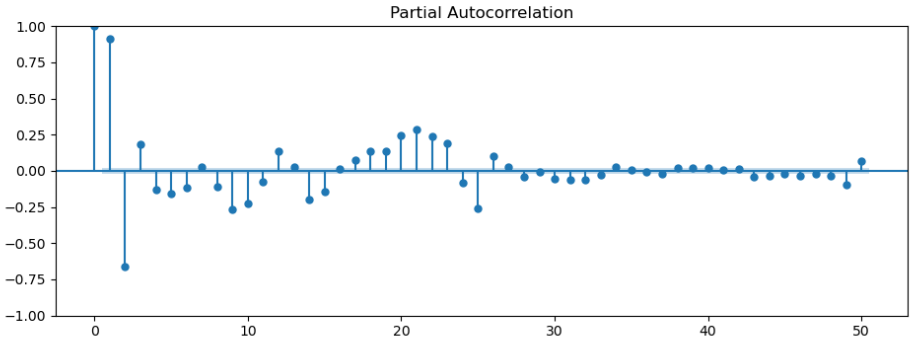
```
ADF Statistic: -11.254188
p-value: 0.000000
Critical Values:
    1%: -3.431
    5%: -2.862
    10%: -2.567
```

```
Results of KPSS Test:
Test Statistic            0.243493
p-value                   0.100000
Lags Used                29.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
dtype: float64
```

- P-value from ADF is lower than threshold(0.05), and P-value from KPSS is higher
- From the ADF and KPSS test, the dataset is stationary

Data Description – ACF / PACF





- ACF has oscillating shape
- The peak is 24, 48, ...
- The dataset has a severe seasonality

Data Description – ACF / PACF from Differencing

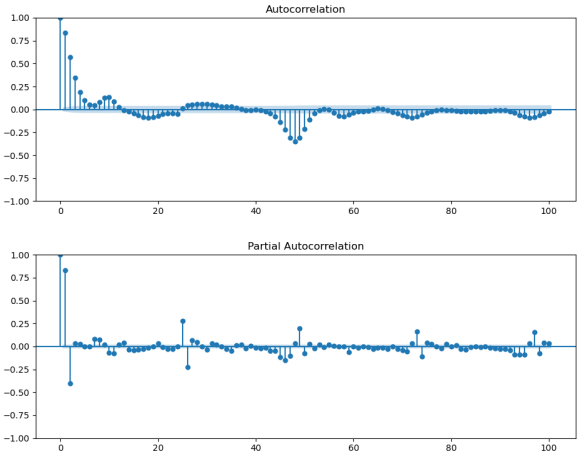


figure 1

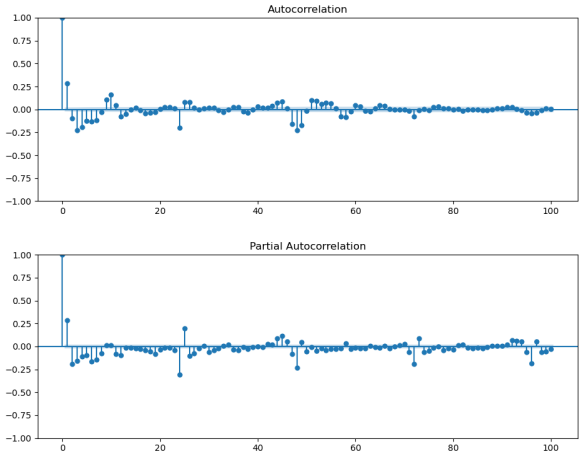


figure 2

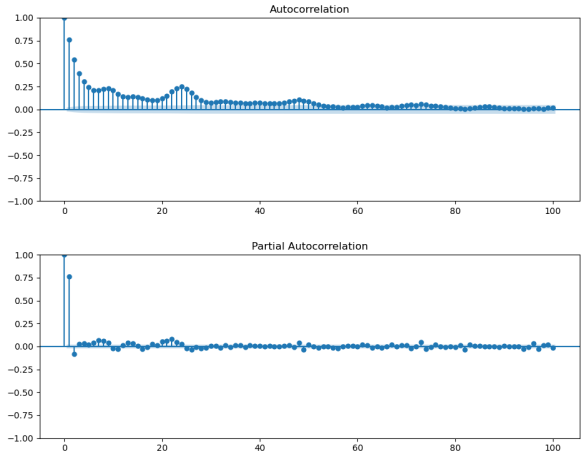
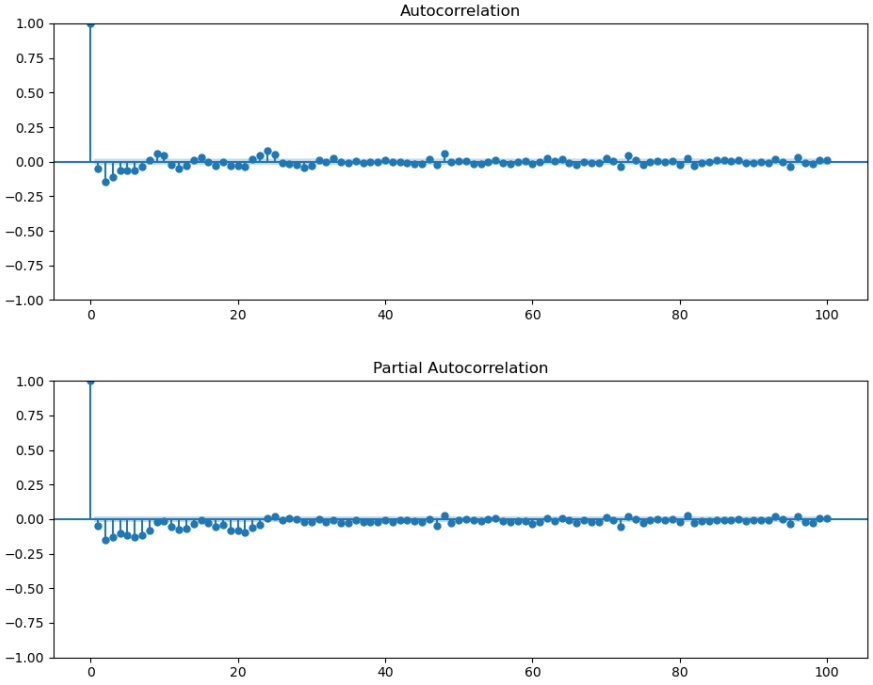


figure 3

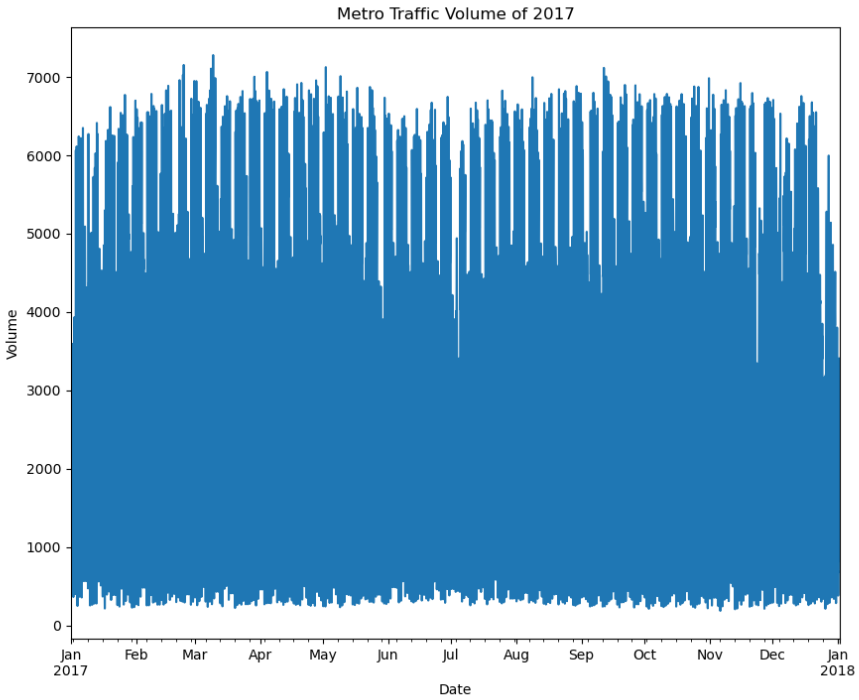
- Figure1 : Seasonal differencing with periods = 24
- Figure2 : Figure1 + First-order non-seasonal differencing
- Figure3: Seasonal differencing with periods = 168 (weekly seasonality)

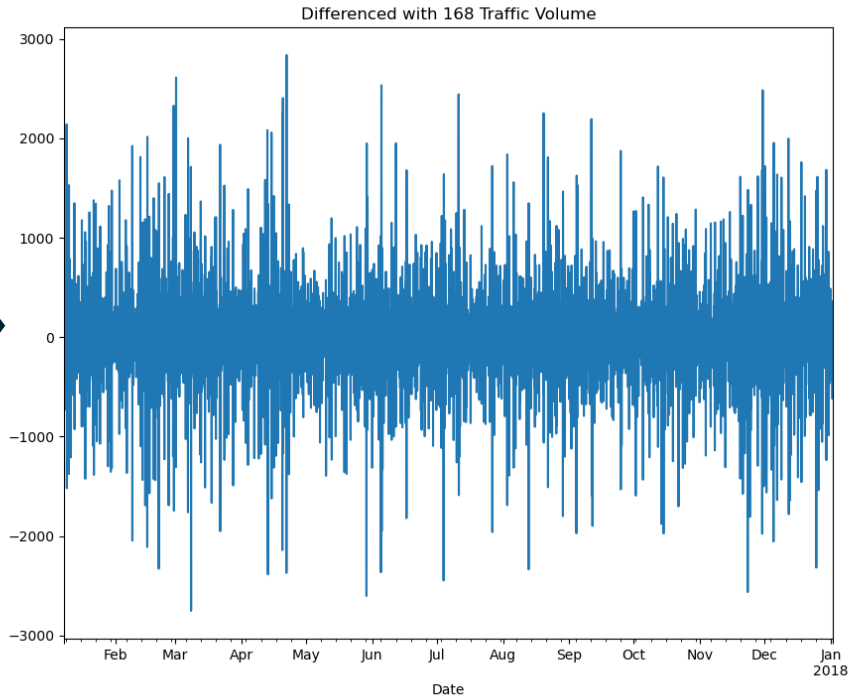
Data Description – ACF / PACF from Differencing



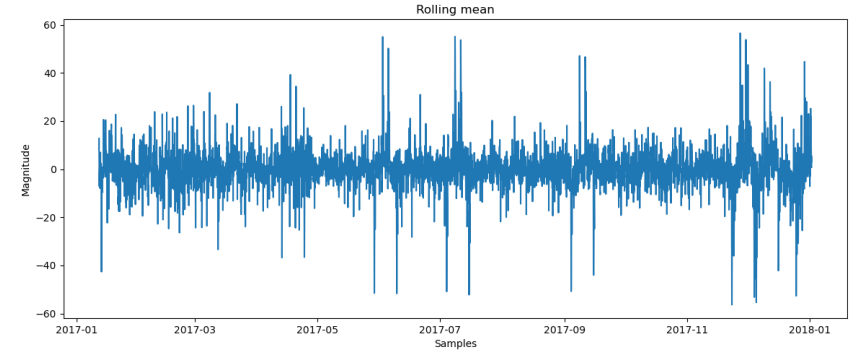
- Seasonal differencing with period = 168
- Non-seasonal first order differencing is followed
- Based on ACF, seasonal differencing with period of 168 followed by first order non-seasonal differencing is used to make dataset stationary

Data Description – Data after Differencing





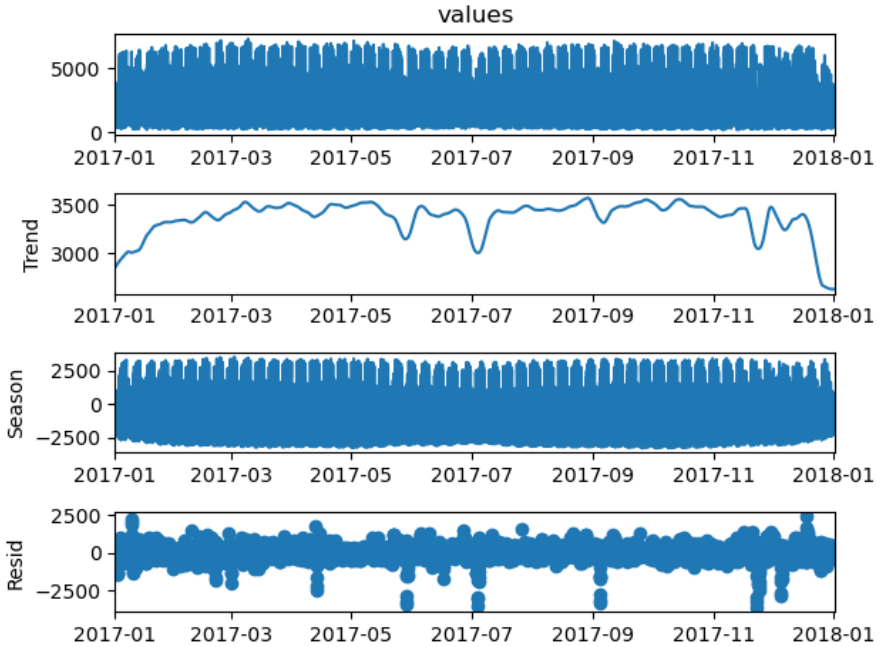
Data Description – Data after Differencing

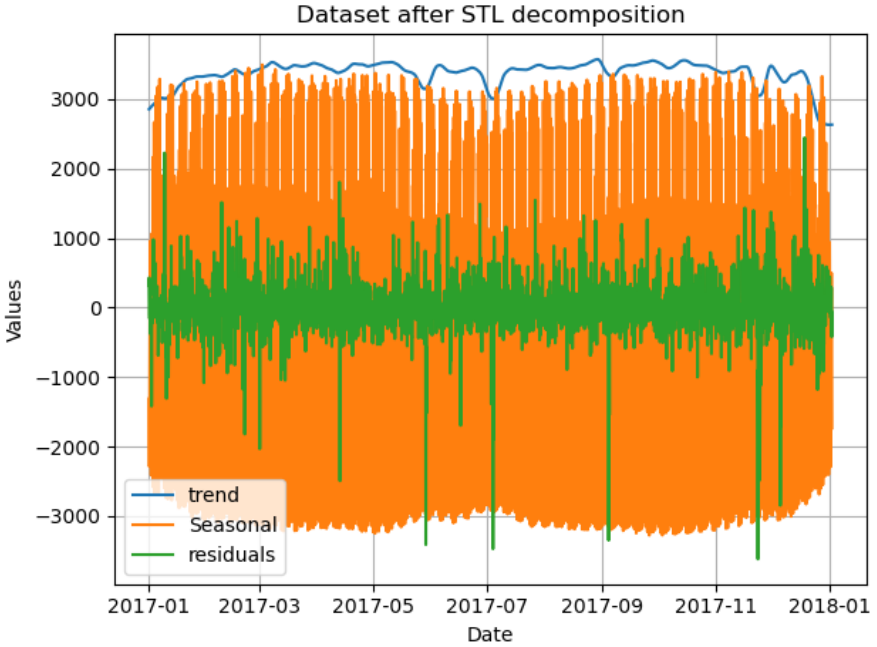



```
ADF Statistic: -22.576452
p-value: 0.000000
Critical Values:
    1%: -3.431
    5%: -2.862
    10%: -2.567
ADF for the transformed dataset: None
Results of KPSS Test:
Test Statistic            0.04072
p-value                   0.10000
Lags Used               380.00000
Critical Value (10%)      0.34700
Critical Value (5%)       0.46300
Critical Value (2.5%)     0.57400
Critical Value (1%)       0.73900
dtype: float64
KPSS for the transformed dataset: None
```

- Data after Differencing is also stationary

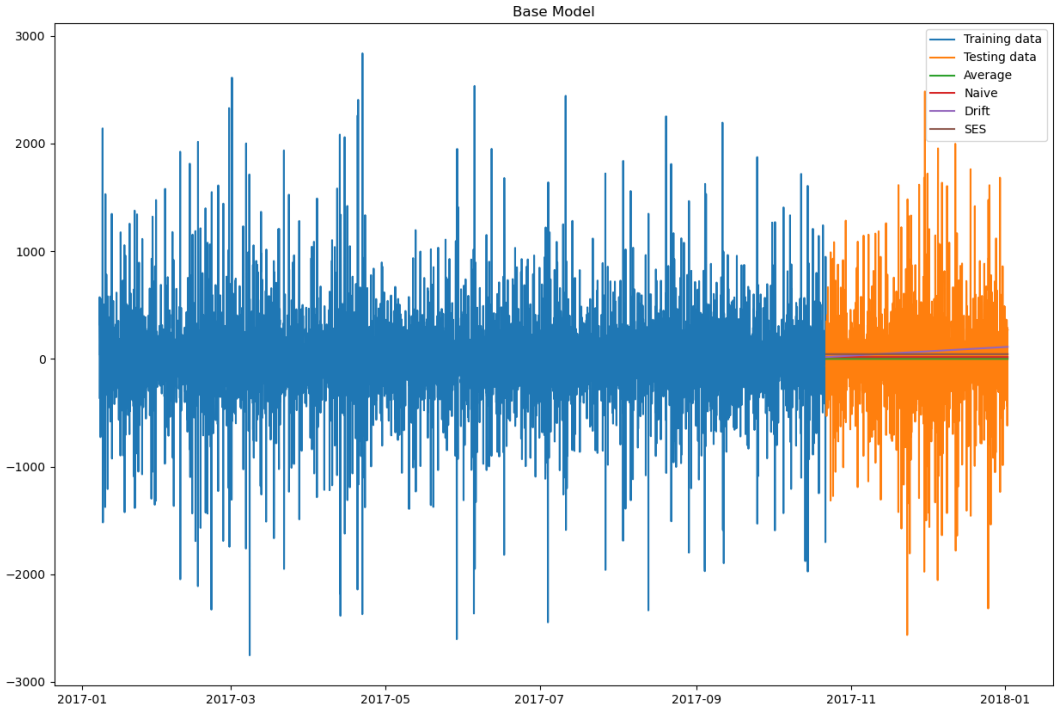
Data Description – STL Decomposition






```
The strength of trend for this data set is 0.29159030625136384
The strength of seasonality for this data set is 0.9762908411024094
```

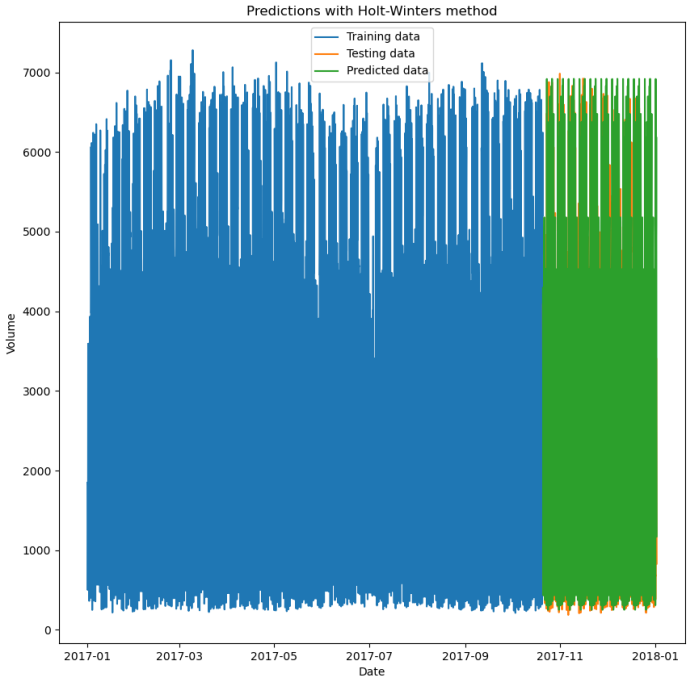
Model – Base Models



Best Base Model based on RMSE - **Average Model**

- RMSE for Average Model: 434.09354258940544
- RMSE for Naive Model: 434.43174536549657
- RMSE for Drift Model: 439.732105352747
- RMSE for SES Model: 436.33900696177295

Model – Holt-Winters Method



- seasonal='mul'
- seasonal_periods=168
- RMSE for Holt-Winters Model: 756.0131117149048

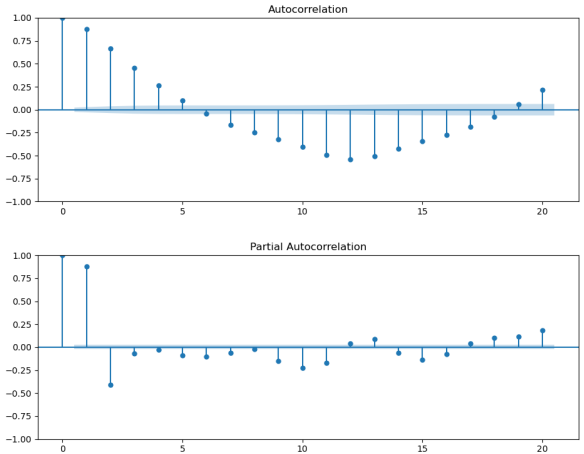
Model – OLS : Full Model

OLS Regression Results

Dep. Variable:	traffic_volume	R-squared:	0.072
Model:	OLS	Adj. R-squared:	0.067
Method:	Least Squares	F-statistic:	13.88
Date:	Tue, 09 May 2023	Prob (F-statistic):	6.32e-86
Time:	18:22:23	Log-Likelihood:	-63069.
No. Observations:	7027	AIC:	1.262e+05
Df Residuals:	6987	BIC:	1.265e+05
Df Model:	39		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3414.2349	29.159	117.092	0.000	3357.075	3471.395
x1	225.8202	25.094	8.999	0.000	176.628	275.013
x2	421.6712	48.431	8.707	0.000	326.733	516.610
x3	-1.231e+15	1.87e+15	-0.659	0.510	-4.89e+15	2.43e+15
x4	-1.231e+15	1.87e+15	-0.659	0.510	-4.89e+15	2.43e+15
x5	-1.231e+15	1.87e+15	-0.659	0.510	-4.89e+15	2.43e+15
x6	-1.231e+15	1.87e+15	-0.659	0.510	-4.89e+15	2.43e+15
x7	-1.231e+15	1.87e+15	-0.659	0.510	-4.89e+15	2.43e+15
x8	-1.231e+15	1.87e+15	-0.659	0.510	-4.89e+15	2.43e+15
x9	-3.481e+15	5.28e+15	-0.659	0.510	-1.38e+16	6.87e+15
x10	-1.231e+15	1.87e+15	-0.659	0.510	-4.89e+15	2.43e+15
x11	-1.294e+14	1.96e+14	-0.659	0.510	-5.14e+14	2.55e+14
x12	1.399e+13	2.12e+13	0.659	0.510	-2.76e+13	5.56e+13
x13	-1.231e+15	1.87e+15	-0.659	0.510	-4.89e+15	2.43e+15
x14	6.064e+14	9.2e+14	0.659	0.510	-1.2e+15	2.41e+15
x15	2.205e+14	3.35e+14	0.659	0.510	-4.35e+14	8.76e+14
x16	-1.502e+14	2.28e+14	-0.659	0.510	-5.97e+14	2.96e+14
x17	-2.913e+14	4.42e+14	-0.659	0.510	-1.16e+15	5.75e+14
x18	-3.396e+14	5.15e+14	-0.659	0.510	-1.35e+15	6.71e+14
x19	-3.76e+14	5.7e+14	-0.659	0.510	-1.49e+15	7.42e+14
x20	3.388e+14	5.14e+14	0.659	0.510	-6.69e+14	1.35e+15
x21	1.145e+15	1.74e+15	0.659	0.510	-2.26e+15	4.55e+15
x22	-3.887e+14	5.9e+14	-0.659	0.510	-1.54e+15	7.67e+14
x23	-1.26e+14	1.91e+14	-0.659	0.510	-5.01e+14	2.49e+14
x24	-2.137e+14	3.24e+14	-0.659	0.510	-8.49e+14	4.22e+14
x25	1.502e+14	2.28e+14	0.659	0.510	-2.96e+14	5.97e+14
x26	2.913e+14	4.42e+14	0.659	0.510	-5.75e+14	1.16e+15
x27	-4.47e+13	6.78e+13	-0.659	0.510	-1.78e+14	8.82e+13
x28	1.212e+14	1.84e+14	0.659	0.510	-2.39e+14	4.82e+14
x29	-5.198e+13	7.89e+13	-0.659	0.510	-2.07e+14	1.03e+14
x30	-1.767e+14	2.68e+14	-0.659	0.510	-7.02e+14	3.49e+14
x31	2.436e+13	3.7e+13	0.659	0.510	-4.81e+13	9.68e+13
x32	2.748e+14	4.17e+14	0.659	0.510	-5.43e+14	1.09e+15
x33	-3.288e+13	4.99e+13	-0.659	0.510	-1.31e+14	6.49e+13
x34	-0.0549	0.083	-0.659	0.510	-0.218	0.108
x35	-3.162e+14	4.8e+14	-0.659	0.510	-1.26e+15	6.24e+14
x36	3.396e+14	5.15e+14	0.659	0.510	-6.71e+14	1.35e+15
x37	2.267e+14	3.44e+14	0.659	0.510	-4.47e+14	9.01e+14
x38	-3.936e+14	5.97e+14	-0.659	0.510	-1.56e+15	7.77e+14
x39	1.03e+14	1.56e+14	0.659	0.510	-2.03e+14	4.09e+14
x40	-1.005e+15	1.53e+15	-0.659	0.510	-3.99e+15	1.98e+15
x41	-9.071e+13	1.38e+14	-0.659	0.510	-3.61e+14	1.79e+14
x42	-2.028e+14	3.08e+14	-0.659	0.510	-8.06e+14	4e+14
x43	-3.899e+14	5.92e+14	-0.659	0.510	-1.55e+15	7.7e+14
x44	-113.2027	99.089	-1.142	0.253	-307.447	81.041
x45	-2.325e+13	3.53e+13	-0.659	0.510	-9.24e+13	4.59e+13
x46	-1.064e+14	1.61e+14	-0.659	0.510	-4.23e+14	2.1e+14
x47	-4.151e+14	6.3e+14	-0.659	0.510	-1.65e+15	8.2e+14
x48	-1.814e+14	2.75e+14	-0.659	0.510	-7.21e+14	3.58e+14
x49	-1.283e+14	1.95e+14	-0.659	0.510	-5.1e+14	2.53e+14
x50	-1.814e+14	2.75e+14	-0.659	0.510	-7.21e+14	3.58e+14
x51	-9.071e+13	1.38e+14	-0.659	0.510	-3.61e+14	1.79e+14

- P-values for features except for first 2 are higher than 0.510



Model – OLS : Feature Selection

- Backward stepwise feature selection is used for feature selection

	AIC	BIC	Adj R^2	Prev. BIC	Prev. Adj R^2
24	126202.382899	126387.535808	0.067068	126394.602114	0.067173
25	126202.392573	126380.687967	0.066935	126387.535808	0.067068
23	126202.591689	126394.602114	0.067173	126401.832697	0.067255
26	126202.732032	126374.169911	0.066757	126380.687967	0.066935
22	126202.964758	126401.832697	0.067255	126409.187781	0.067321
21	126203.462326	126409.187781	0.067321	126416.597833	0.067380
27	126203.663992	126368.244356	0.066501	126374.169911	0.066757
20	126204.014863	126416.597833	0.067380	126424.211091	0.067412
19	126204.770606	126424.211091	0.067412	126431.907112	0.067433
28	126205.344846	126363.067695	0.066145	126368.244356	0.066501
18	126205.609112	126431.907112	0.067433	126431.907112	0.067433

- Features to keep: ['temp', 'clouds_all', 'weather_main_Clouds', 'weather_description_broken clouds', 'weather_description_drizzle', 'weather_description_few clouds', 'weather_description_heavy intensity drizzle', 'weather_description_heavy intensity rain', 'weather_description_light intensity drizzle', 'weather_description_light rain', 'weather_description_light rain and snow', 'weather_description_light snow', 'weather_description_mist', 'weather_description_moderate rain', 'weather_description_proximity thunderstorm', 'weather_description_proximity thunderstorm with rain', 'weather_description_scattered clouds', 'weather_description_sky is clear', 'weather_description_snow', 'weather_description_thunderstorm', 'weather_description_thunderstorm with light drizzle']
- Features to eliminate: ['weather_description_sleet', 'weather_main_Mist', 'weather_description_heavy snow', 'weather_main_Snow', 'holiday_Veterans Day', 'weather_main_Fog', 'weather_description_fog', 'weather_description_overcast clouds', 'weather_main_Drizzle', 'weather_description_haze', 'weather_main_Haze', 'weather_description_proximity thunderstorm with drizzle', 'weather_description_proximity shower rain', 'weather_main_Rain', 'holiday_None', 'weather_description_light shower snow', 'holiday_Martin Luther King Jr Day', 'weather_description_light intensity shower rain', 'holiday_Thanksgiving Day', 'holiday_Independence Day', 'weather_main_Thunderstorm', 'holiday_Labor Day', 'holiday_New Years Day', 'holiday_Memorial Day', 'holiday_Columbus Day', 'weather_description_thunderstorm with rain', 'holiday_State Fair', 'holiday_Washingtons Birthday', 'weather_description_thunderstorm with light rain', 'weather_description_thunderstorm with heavy rain']

Model – OLS: Multi-collinearity

	feature	VIF
0	temp	1.172962
1	clouds_all	4.008464
2	weather_main_Clouds	8.132498
3	weather_description_broken clouds	1.878463
4	weather_description_drizzle	1.172341
5	weather_description_few clouds	1.610834
6	weather_description_heavy intensity drizzle	1.020216
7	weather_description_heavy intensity rain	1.252323
8	weather_description_light intensity drizzle	1.297529
9	weather_description_light rain	2.185851
10	weather_description_light rain and snow	1.006923
11	weather_description_light snow	1.628223
12	weather_description_mist	2.539087
13	weather_description_moderate rain	1.883428
14	weather_description_proximity thunderstorm	1.394541
15	weather_description_proximity thunderstorm wit...	1.014823
16	weather_description_scattered clouds	2.448045
17	weather_description_sky is clear	6.054076
18	weather_description_snow	1.078326
19	weather_description_thunderstorm	1.065502
20	weather_description_thunderstorm with light dr...	1.007505

```
SingularValues =  [2.36529393e+04 2.06674058e+04 2.00305844e+04 1.89750342e+04
 1.80837385e+04 1.79407737e+04 1.78695095e+04 1.76470886e+04
 1.71115157e+04 1.09917415e+04 9.95726749e+03 9.30202604e+03
 9.18585138e+03 8.93128105e+03 8.92387914e+03 8.87402868e+03
 8.85318205e+03 8.82558169e+03 8.80142135e+03 8.79675880e+03
 8.79407464e+03 8.79381916e+03 8.78933437e+03 8.78870821e+03
 8.78806275e+03 8.78787631e+03 8.78643407e+03 8.78626963e+03
 8.78523799e+03 8.78516166e+03 8.78501698e+03 8.78500011e+03
 8.78500011e+03 8.78500011e+03 8.76670150e+03 8.74388789e+03
 8.71415017e+03 8.69908225e+03 8.60062439e+03 6.22119122e+03
 1.25070454e+03 3.74014989e+02 1.67038152e+02 3.30725782e-11
 2.92202838e-11 1.54935228e-11 5.45708318e-12 1.81533793e-12
 1.81533793e-12 1.81533793e-12 1.62549521e-12]
Condition Number= 3614978705674538.5
```

- There is no multi-collinearity in features

Model – OLS: Final Model

OLS Regression Results

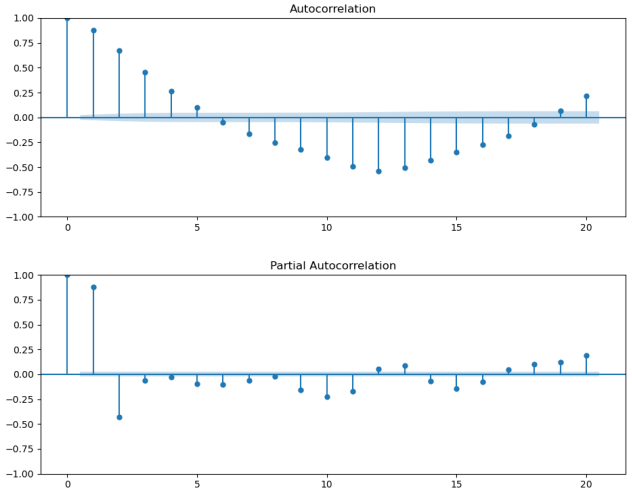
Dep. Variable:	traffic_volume	R-squared:	0.069
Model:	OLS	Adj. R-squared:	0.066
Method:	Least Squares	F-statistic:	24.57
Date:	Tue, 09 May 2023	Prob (F-statistic):	3.11e-92
Time:	18:43:18	Log-Likelihood:	-63081.
No. Observations:	7027	AIC:	1.262e+05
Df Residuals:	7005	BIC:	1.264e+05
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3402.3239	22.894	148.611	0.000	3357.444	3447.203
temp	222.2303	24.795	8.963	0.000	173.624	270.836
clouds_all	419.5406	45.837	9.153	0.000	329.686	509.395
weather_main_Clouds	-299.1868	65.289	-4.583	0.000	-427.172	-171.201
weather_description_broken clouds	137.5994	31.378	4.385	0.000	76.089	199.110
weather_description_drizzle	-103.0641	24.789	-4.158	0.000	-151.657	-54.471
weather_description_few clouds	202.3416	29.057	6.964	0.000	145.381	259.302
weather_description_heavy intensity drizzle	-45.3816	23.124	-1.962	0.050	-90.713	-0.051
weather_description_heavy intensity rain	-114.8374	25.620	-4.482	0.000	-165.061	-64.614
weather_description_light intensity drizzle	-109.2862	26.079	-4.191	0.000	-160.408	-58.164
weather_description_light rain	-195.1105	33.848	-5.764	0.000	-261.463	-128.758
weather_description_light rain and snow	-55.5572	22.973	-2.418	0.016	-100.592	-10.523
weather_description_light snow	-147.5406	29.213	-5.050	0.000	-204.808	-90.273
weather_description_mist	-255.0847	36.481	-6.992	0.000	-326.598	-183.571
weather_description_moderate rain	-164.2489	31.420	-5.228	0.000	-225.841	-102.657
weather_description_proximity thunderstorm	-160.4352	27.036	-5.934	0.000	-213.434	-107.437
weather_description_proximity thunderstorm with rain	-46.5996	23.063	-2.021	0.043	-91.811	-1.389
weather_description_scattered clouds	246.9102	35.821	6.893	0.000	176.691	317.130
weather_description_sky is clear	-115.5526	56.331	-2.051	0.040	-225.979	-5.126
weather_description_snow	-64.7941	23.774	-2.725	0.006	-111.398	-18.190
weather_description_thunderstorm	-82.7141	23.632	-3.500	0.000	-129.040	-36.388
weather_description_thunderstorm with light drizzle	-50.5480	22.980	-2.200	0.028	-95.596	-5.500

Omnibus:	2494.971	Durbin-Watson:	0.243
Prob(Omnibus):	0.000	Jarque-Bera (JB):	344.360
Skew:	-0.060	Prob(JB):	1.67e-75
Kurtosis:	1.922	Cond. No.	6.41

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

- Mean of residual_errors: -2.1288414479523444e-12
- Mean of forecast_errors: 70.6160590398032
- RMSE: 1883.637898288714
- The residual is NOT white
- R^2: 0.069



Model – ARIMA(2, 0, 6)

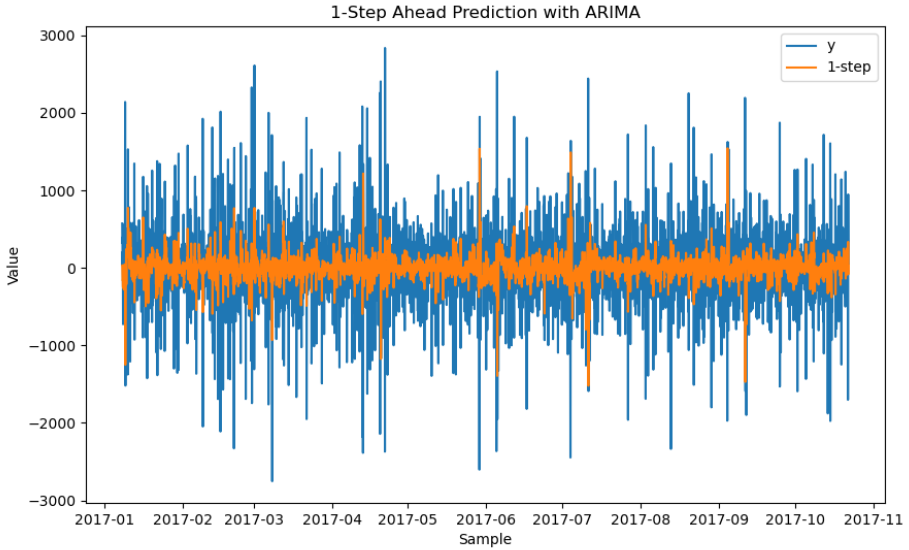
SARIMAX Results

Dep. Variable:	traffic_volume	No. Observations:	6892
Model:	ARIMA(2, 0, 6)	Log Likelihood	-50323.442
Date:	Tue, 09 May 2023	AIC	100666.884
Time:	22:54:50	BIC	100735.265
Sample:	01-08-2017	HQIC	100690.465
	- 10-22-2017		
Covariance Type:	opg		

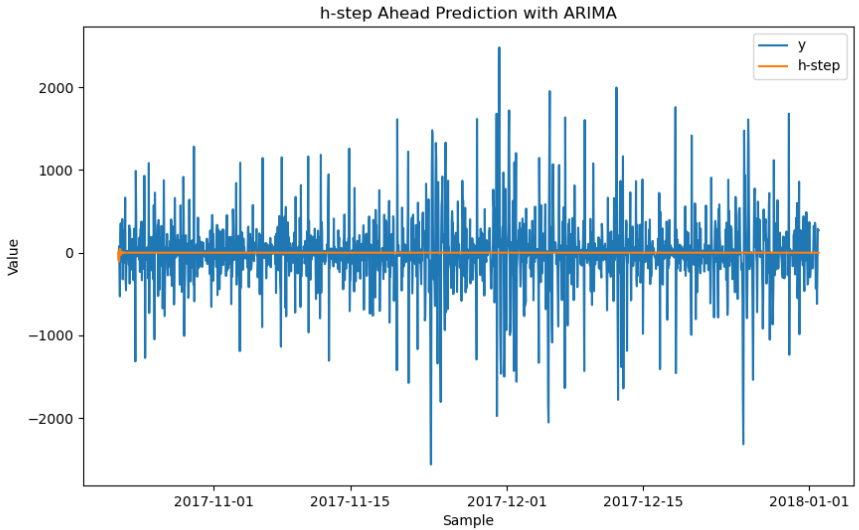
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.1640	0.365	-0.449	0.653	-0.880	0.552
ar.L1	0.8375	0.097	8.606	0.000	0.647	1.028
ar.L2	-0.5364	0.057	-9.437	0.000	-0.648	-0.425
ma.L1	-1.0759	0.098	-11.018	0.000	-1.267	-0.884
ma.L2	0.4731	0.078	6.072	0.000	0.320	0.626
ma.L3	-0.0726	0.028	-2.635	0.008	-0.127	-0.019
ma.L4	-0.0918	0.026	-3.566	0.000	-0.142	-0.041
ma.L5	-0.0940	0.020	-4.770	0.000	-0.133	-0.055
ma.L6	-0.0823	0.018	-4.564	0.000	-0.118	-0.047
sigma2	1.321e+05	999.809	132.170	0.000	1.3e+05	1.34e+05

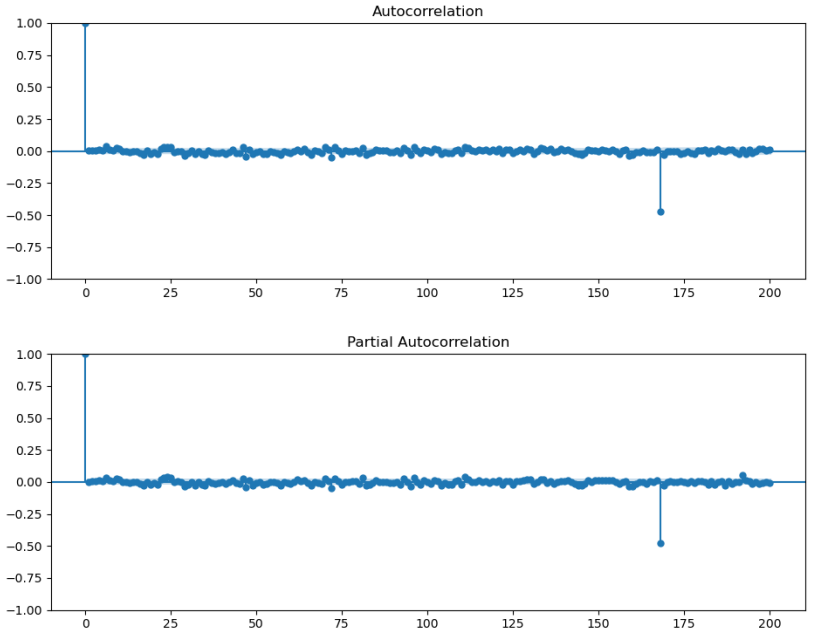
Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	35741.09
Prob(Q):	0.92	Prob(JB):	0.00
Heteroskedasticity (H):	0.78	Skew:	0.01
Prob(H) (two-sided):	0.00	Kurtosis:	14.16

- P-values from the estimated coefficients are less than 0.05
- RMSE for ARIMA: 434.1142080430148



Model – ACF of Residual from ARIMA(2, 0, 6)





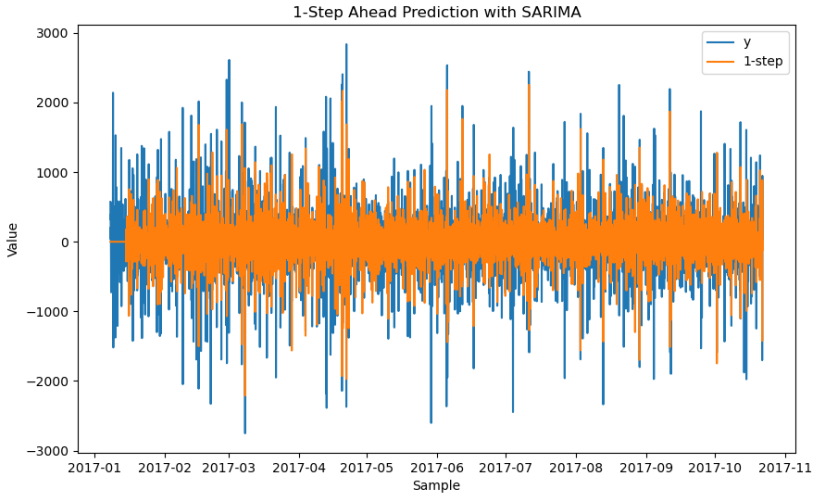
- ACF/PACF of residual has peak at lag=168

→ SARIMA is needed

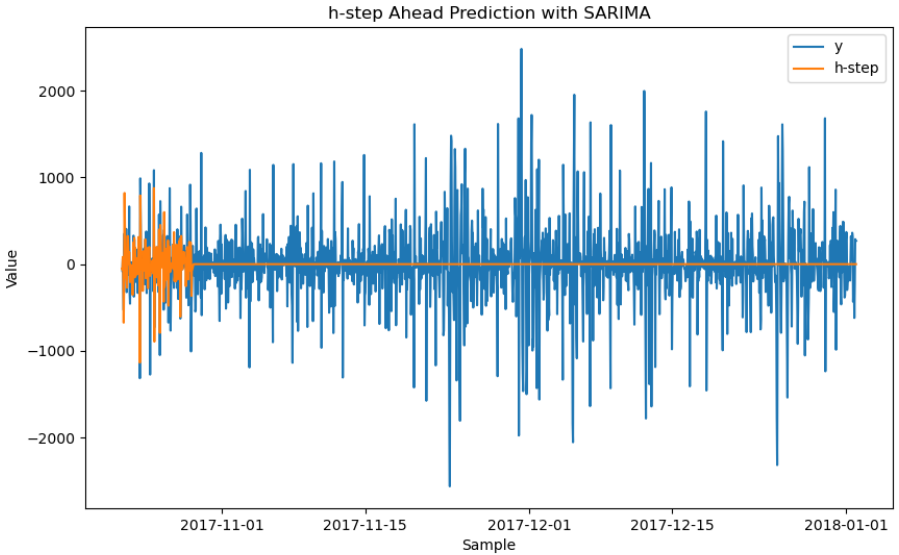
Model – SARIMA(0, 0, 1, 168)

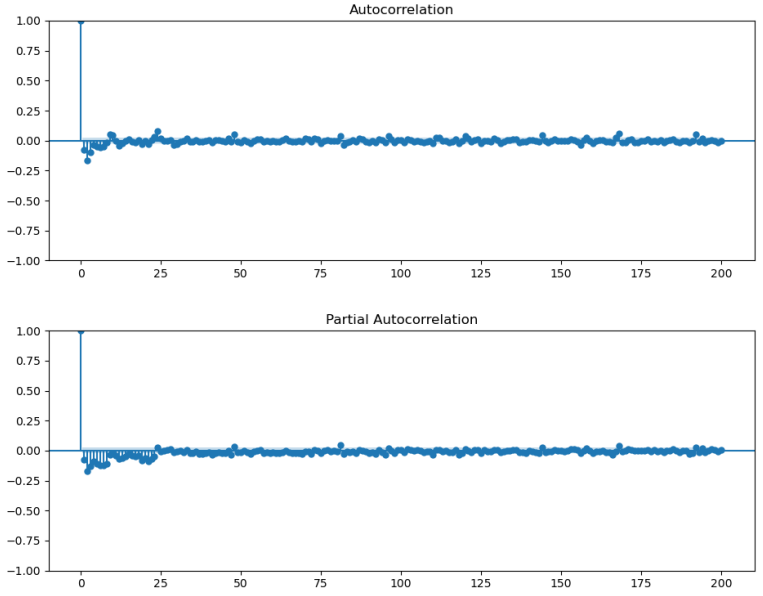
```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                  traffic_volume   No. Observations:                 6892
Model:             SARIMAX(0, 0, [1], 168)   Log Likelihood              -49162.934
Date:                     Tue, 09 May 2023   AIC                          98329.867
Time:                             15:12:25   BIC                          98343.543
Sample:                         01-08-2017   HQIC                         98334.583
                              - 10-22-2017
Covariance Type:                      opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.S.L168     -0.9007      0.006   -154.273      0.000      -0.912      -0.889
sigma2      8.827e+04    631.707    139.732      0.000     8.7e+04    8.95e+04
==========================================================================================
Ljung-Box (L1) (Q):                  47.08   Jarque-Bera (JB):             31726.57
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.71   Skew:                            -0.17
Prob(H) (two-sided):                  0.00   Kurtosis:                        13.51
==========================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
RUNNING THE L-BFGS-B CODE
```

- P-value from the estimated coefficient is less than 0.05
- RMSE for SARIMA: 427.19046452962493



Model – SARIMA(0, 0, 1, 168)





- There is no peak at 168 in ACF/PACF of residuals
- Mean of Residual: 0.36812733693893085
- Variance of Residual: 92899.74806144746

Final Model Selection – ACF/PACF of Residuals

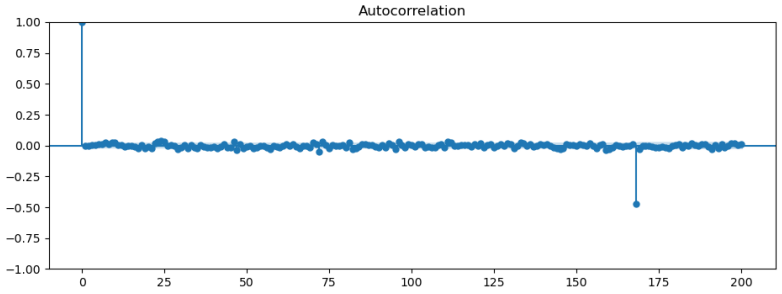


figure 1

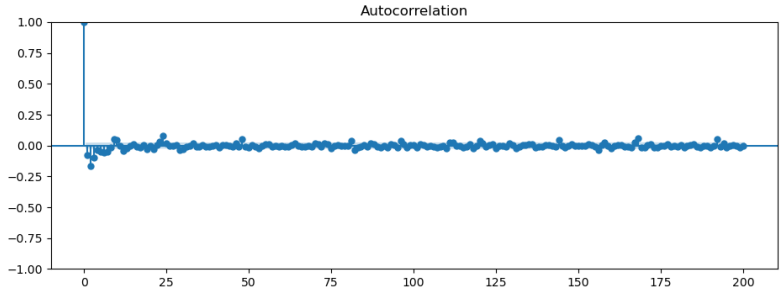


figure 2

- Figure1 : ACF/PACF of residual from ARIMA(2,0,7) model
- Figure2 : ACF/PACF of residual from SARIMA(0, 0, 1, 168) model

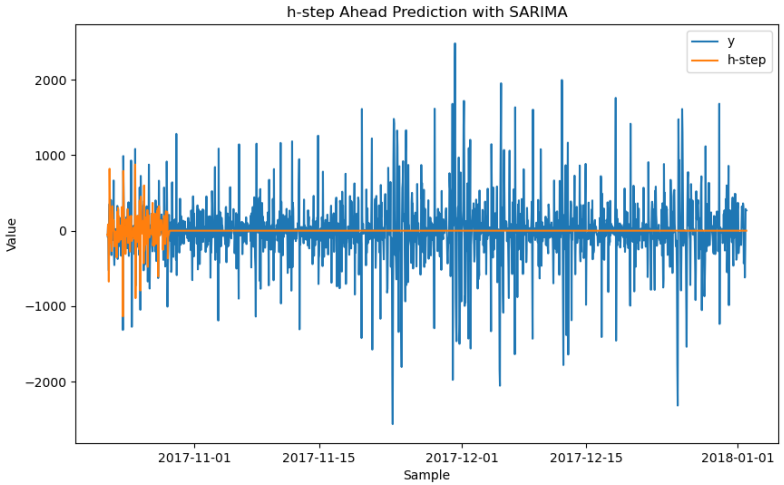
Final Model Selection

- RMSE for Average Model: 434.09354258940544
- RMSE for ARIMA: 434.13764539656626
- **RMSE for SARIMA: 427.19046452962493**

→ **SARIMA(0, 0, 1, 168)** is selected as a final model

Conclusion

- Based on RMSE and ACF/PACF of residual, **SARIMA(0, 0, 1, 168)** is selected as a final model
- RMSE: 427
- However, The residual is NOT white (by Ljung-Box test)
- Therefore, there is a possibility of better model



Thank You