# DATS 6450: Time Series Analysis and Modeling

Instructor: Dr. Reza Jafari

Term Project

Rayna Liu

RL

05/04/2021

## Table of Contents


I. Introduction ---------- 6

II. Description of the Dataset ---------- 7

a. Data Preprocessing ---------- 7

b. Plot of the Dependent Variable versus Time ---------- 8

c. ACF/PACF of the Dependent Variable ---------- 9

d. Correlation Matrix ---------- 10

e. Split the Dependent Variable into Train and Test Set ---------- 11

III. Stationary ---------- 12

a. ADF Test ---------- 12

b. Plot of Rolling Mean and Rolling Variance ---------- 12

c. Seasonal / Non-Seasonal Differencing ---------- 13

IV. Time Series Decomposition ---------- 14

a. STL Decomposition Method ---------- 14

b. Plot of the Raw Dataset versus the De-trended Dataset and Seasonally Adjusted Dataset ---------- 14

c. Strength of the Trend and Seasonality ---------- 15

V. Holt-Winter Method ---------- 16

VI. Feature Selection ---------- 17

a. Split the Whole Dataset into Train and Test Set ---------- 17

b. Collinearity Detection ---------- 17

i. Singular Values Analysis ---------- 17

ii. Condition Number ---------- 17

c. Feature Reduction ---------- 18

VII. Multiple Linear Regression ---------- 21

a. Hypothesis Tests Analysis ---------- 21

i. F-test ---------- 21

ii. t-test ---------- 21

b. AIC, BIC, R-squared and Adjusted R-squared ---------- 22

c. Perform One-Step ahead Forecast ---------- 22

d. Residuals Analysis ---------- 23

i. RMSE ---------- 23

ii. ACF ---------- 23

iii. Mean, Variance, Q-value ---------- 24

VIII. ARMA, ARIMA, and SARIMA Model ---------- 25

a. Order Determination ---------- 25

i. GPAC Table ---------- 25

ii. ACF/PACF Plot ---------- 26

IX. Estimated Parameters of ARMA Model ---------- 27

X. Diagnostic Analysis ---------- 28

a. Diagnostic Tests ---------- 28

i. Confidence Intervals ---------- 28

ii. Zero/Pole Cancellation ---------- 28

iii. Chi-Square Test ---------- 28

iv. ACF of the Residuals ---------- 29

b. Estimated Variance of Error and Estimated Covariance of the Estimated Parameters ---------- 29

c. Biased or Unbiased Estimator ---------- 30

d. The Variance of Residual Errors versus the Variance of the Forecast Errors ---------- 30

e. Forecast Analysis ---------- 30

XI. Base-Models ---------- 32

a. Average Method ---------- 32

b. Naïve Method ---------- 33

c. Seasonal Naïve Method ---------- 34

d. Drift Method ---------- 35

e. Simple Exponential Smoothing ---------- 36

f. Holt's Linear Trend Method ---------- 37

XII. Final Model Selection ---------- 38

a. Forecast Function & h-Step ahead Prediction ---------- 39

XIII. Summary and Conclusion ---------- 39

XIV. References ---------- 40

# Abstract

Traffic volume forecasts are used by many transportation analysis and management systems to better characterize and react to fluctuating traffic patterns.[1] Develop a couple of models to forecast the hourly traffic volume and select the best model by observing, analyzing, and comparing these models' residual and forecast errors


# I. Introduction

The purpose of this term project is to find the best model to forecast the hourly traffic volume. First, do the data preprocessing to know the content of the dataset. Then, do the time series analysis to find the dataset's time-series properties, like stationarity, trend, and seasonality. Dealing with data non-stationary, strong seasonality, and strong trends if applicable. Developing multiple models such as Holt-Winter, Multiple Linear Regression, ARMA, ARIMA, SARIMA, and based models. By analyzing their residual and forecast error, the optimal model was selected to forecast traffic volume.

# II. Description of the Dataset

*Metro Interstate Traffic Volume Dataset* is about hourly Minneapolis-St Paul, MN traffic volume for westbound 1-94. It includes weather and holiday features from 2012-2018.[2] Link [Here].

There are no missing values, but there is missing time. Do the preprocessing procedures first to clean the dataset and make up these missing time series data.

## a. Data Preprocessing

After dropping the duplicated rows, I found out that there are over 10,000 hours of data missing. To handle these missing data, I separated them by hours and replaced the missing data with the mean hourly traffic volume. After the preprocessing procedures done, this dataset contains 40,575 instances and 9 attributes.

Attribute Information:

| | |
|---|---|
| [holiday] | (Categorical) US National holidays plus regional holiday |
| [temp] | (Numeric) Average temperature in Kelvin |
| [rain_1h] | (Numeric) Amount in mm of rain that occurred in the hour |
| [snow_1h] | (Numeric) Amount in mm of snow that occurred in the hour |
| [clouds_all] | (Numeric) Percentage of cloud cover |
| [weather_main] | (Categorical) Short textual description of the current weather |
| [weather description] | (Categorical) Longer textual description of the current weather |
| [date_time] | (DateTime) Hour of the data collected in local CST time |
| [traffic_volume] | (Numeric) Hourly I-94 ATR 301 reported westbound traffic volume |

The date range of this dataset starts from 2012-10-02 09:00:00 to 2018-09-30 23:00:00. The dependent variable is *traffic-volume*, and the independent variables are rest of attributes expect *date_time*.

| | holiday | temp | rain_1h | snow_1h | clouds_all | weather_main | weather_description | date_time | traffic_volume |
|---|---|---|---|---|---|---|---|---|---|
| 0 | None | 288.28000 | 0.00000 | 0.00000 | 40 | Clouds | scattered clouds | 2012-10-02 09... | 5545 |
| 1 | None | 289.36000 | 0.00000 | 0.00000 | 75 | Clouds | broken clouds | 2012-10-02 10:... | 4516 |
| 2 | None | 289.58000 | 0.00000 | 0.00000 | 90 | Clouds | overcast clouds | 2012-10-02 11:... | 4767 |
| 3 | None | 290.13000 | 0.00000 | 0.00000 | 90 | Clouds | overcast clouds | 2012-10-02 12:... | 5026 |
| 4 | None | 291.14000 | 0.00000 | 0.00000 | 75 | Clouds | broken clouds | 2012-10-02 13:... | 4918 |
| 5 | None | 291.72000 | 0.00000 | 0.00000 | 1 | Clear | sky is clear | 2012-10-02 14:... | 5181 |
| 6 | None | 293.17000 | 0.00000 | 0.00000 | 1 | Clear | sky is clear | 2012-10-02 15:... | 5584 |
| 7 | None | 293.86000 | 0.00000 | 0.00000 | 1 | Clear | sky is clear | 2012-10-02 16:... | 6015 |
| 8 | None | 294.14000 | 0.00000 | 0.00000 | 20 | Clouds | few clouds | 2012-10-02 17:... | 5791 |
| 9 | None | 293.10000 | 0.00000 | 0.00000 | 20 | Clouds | few clouds | 2012-10-02 18:... | 4770 |
| 10 | None | 290.97000 | 0.00000 | 0.00000 | 20 | Clouds | few clouds | 2012-10-02 19:... | 3539 |

***Figure 1***. First 10 rows of *Metro Interstate Traffic Volume Dataset*

## b. Dependent Variable versus Time

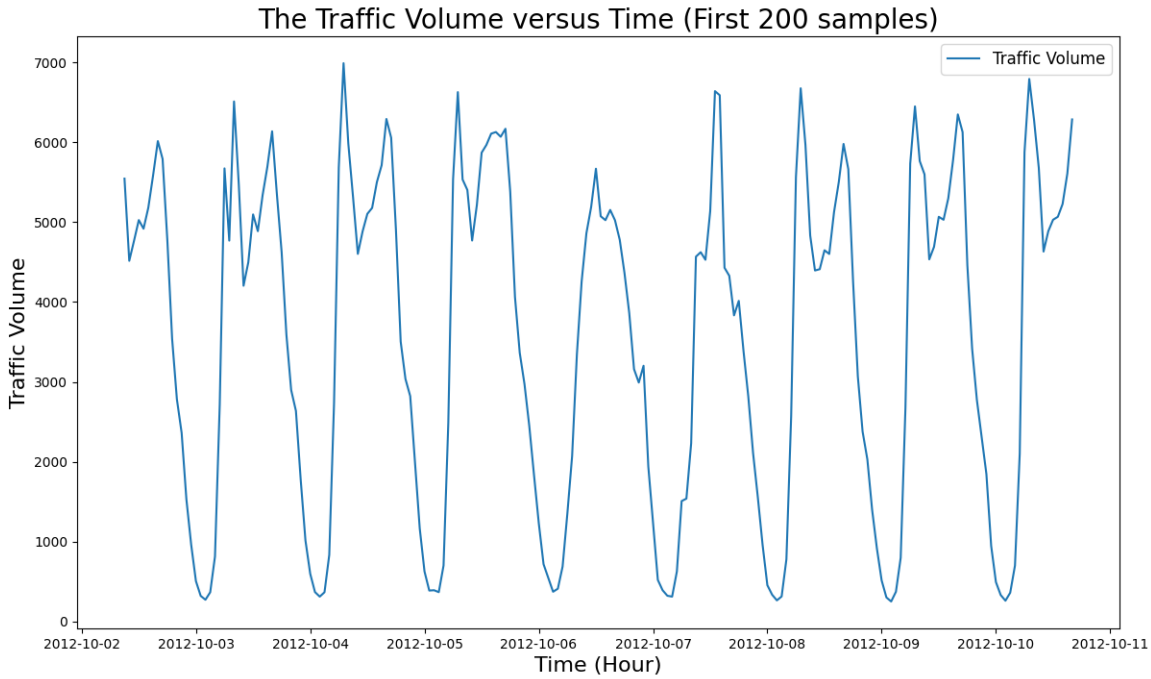



***Figure 2***. Plot of the Dependent Variable v.s. Time (First 200 instances)

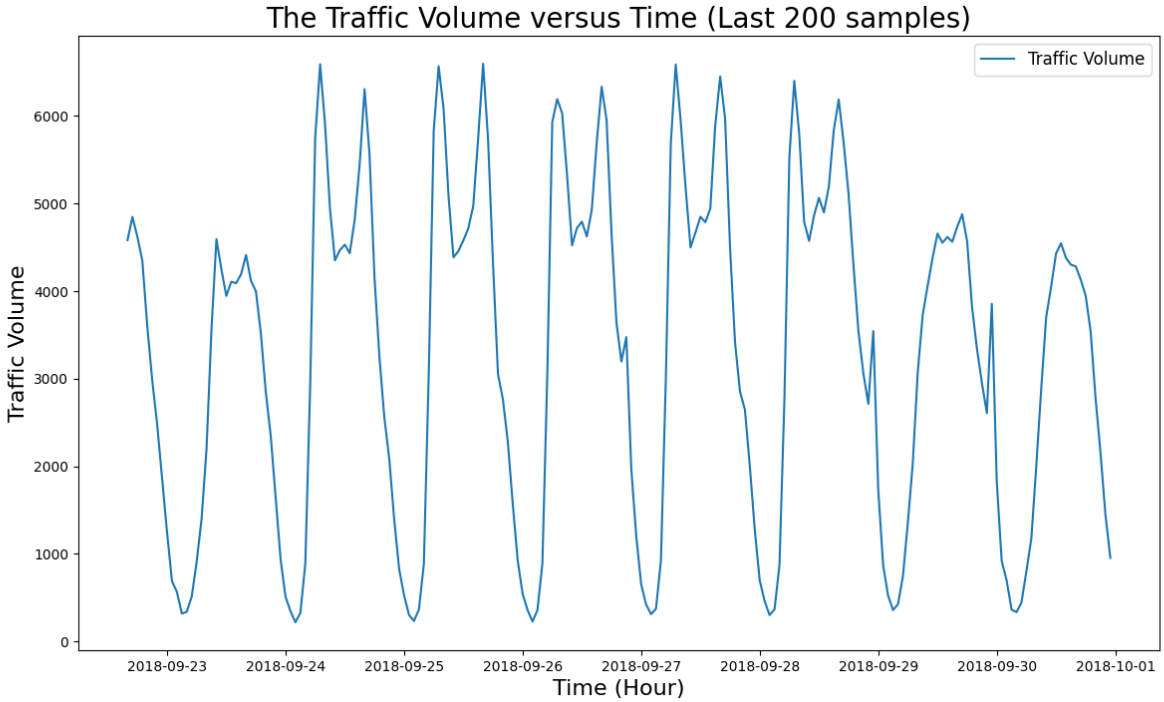



***Figure 3***. Plot of the Dependent Variable v.s. Time (Last 200 instances)

Since this is a large hourly change of the dataset, showing all the instances will obscure some information.

## c. ACF/PACF of the Dependent Variable

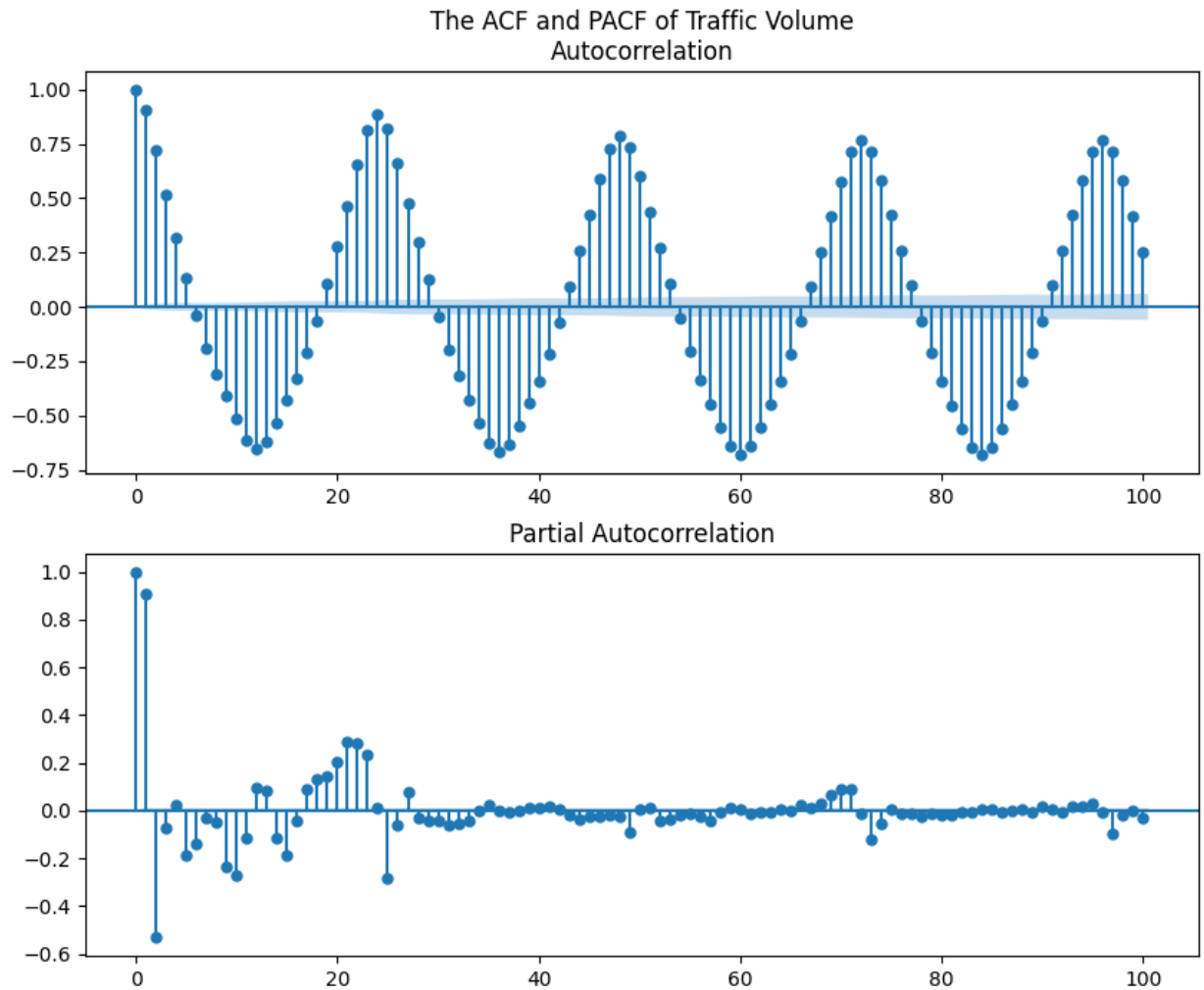



***Figure 4.*** ACF/PACF of the Dependent Variable with lags = 100

The ACF plot shows this is a seasonal dataset, which has a strong seasonal pattern with a period of 24 hours.

## d. Correlation Matrix

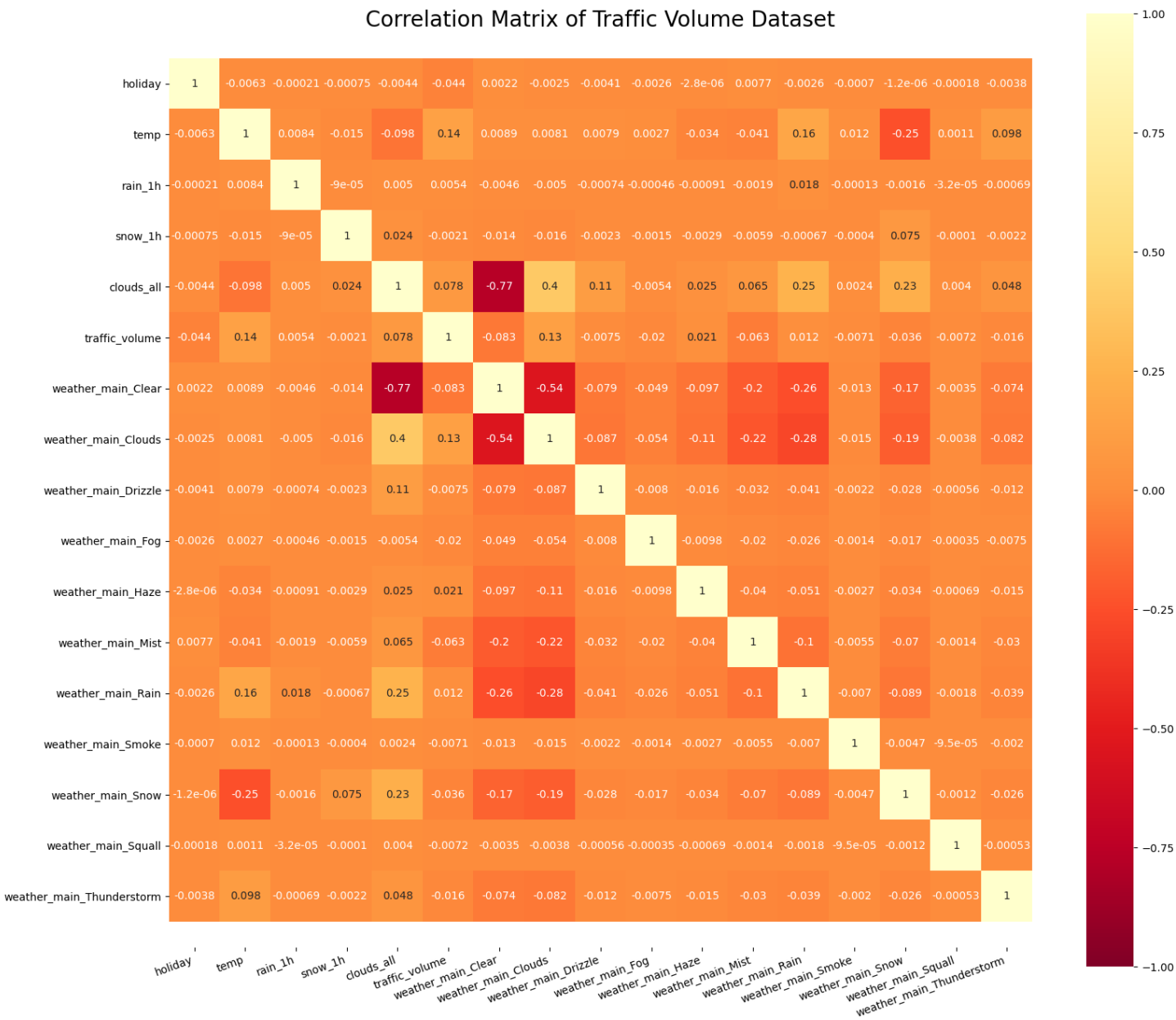



***Figure 5.*** Correlation Matrix with seaborn heatmap and Pearson's correlation coefficient

Before plotting the correlation matrix, I converted the categorical variable *holiday* and *weather_main* into indicator variables. The *weather_main* contains 11 weather categories: Clear, Clouds, Drizzle, Fog, Haze, Mist, Rain, Smoke, Snow, Squall, and Thunderstorm.

In *Figure 5*, the lightest and the deepest colors indicate the greatest correlation between the two variables. The *weather_main_Clear* has a strong negative correlation with *clouds_all* and a moderate negative correlation with *weather_main_Clouds*. There is slight correlation (correlation coefficient < 0.5) among other variables.

## e. Split the Dependent Variables into Train and Test Set

Split the dependent variable *traffic_volume* into train set (80%) and test set (20%). The *traffic_volume* with complete date time index in total 52,551 observations. The train set has 42,040 observations, and the test set has 10,511 observations.

```
In[8]: train
Out[8]:
2012-10-02 09:00:00    5545.0
2012-10-02 10:00:00    4516.0
2012-10-02 11:00:00    4767.0
2012-10-02 12:00:00    5026.0
2012-10-02 13:00:00    4918.0
                         ...
2017-07-19 20:00:00    2804.0
2017-07-19 21:00:00    2815.0
2017-07-19 22:00:00    2029.0
2017-07-19 23:00:00    1228.0
2017-07-20 00:00:00     832.0
Freq: H, Name: Traffic Volume,
```

```
In[9]: test
Out[9]:
2017-07-20 01:00:00     495.0
2017-07-20 02:00:00     361.0
2017-07-20 03:00:00     385.0
2017-07-20 04:00:00     930.0
2017-07-20 05:00:00    3008.0
                         ...
2018-09-30 19:00:00    3543.0
2018-09-30 20:00:00    2781.0
2018-09-30 21:00:00    2159.0
2018-09-30 22:00:00    1450.0
2018-09-30 23:00:00     954.0
Freq: H, Name: Traffic Volume,
```

***Figure 6.*** Train Set (80%) and Test Set (20%)

# III. Stationary

## a. ADF Test of the Raw Dataset

```
The ADF test of Traffic Volume:
ADF Statistic: -33.506261
p-value: 0.000000
Critical Values:
    1%: -3.430
    5%: -2.862
    10%: -2.567
```

***Figure 7.*** ADF Test of the Raw Dataset

## b. Plot of Rolling Mean and Rolling Variance of the Raw Dataset

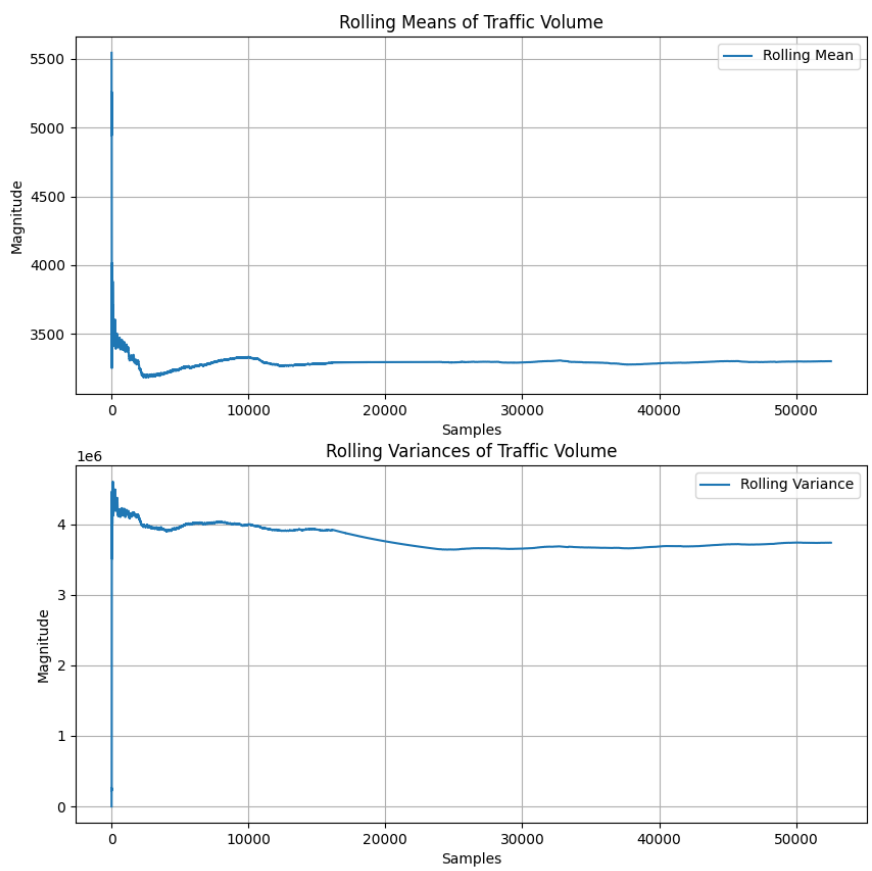



***Figure 8.*** Rolling Mean and Rolling Variance of the Raw Dataset

Although the p-value of ADF test is much less than 0.05, and the rolling mean and rolling variance are constant over time, the ACF in *Figure 4* (pg.9) shows the raw dataset has seasonality and is non-stationary.

## c. Seasonal / Non-Seasonal Differencing

Since this dataset has daily seasonal patterns (24 hours) and weekly seasonal patterns (7 days), applying the seasonal differencing with interval = 24 is unable to make the dataset completely stationary. Thus, using the seasonal differencing with interval = 7x24 = 168 plus the 1st order non-seasonal differencing would make the dataset completely stationary.

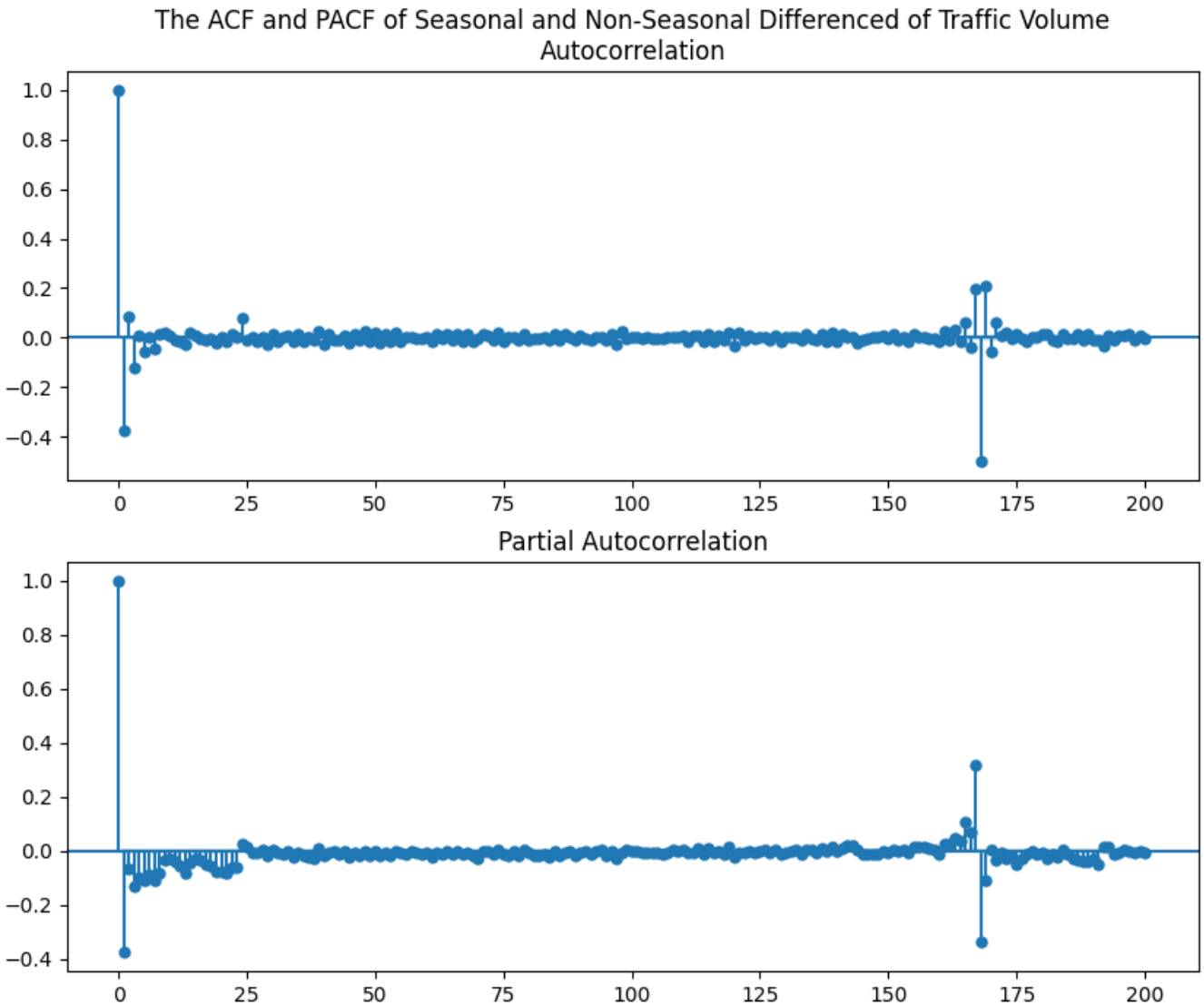



***Figure 9.*** ACF / PACF of Seasonal and Non-Seasonal Differenced Dataset with lags = 200

Compare to the ACF of the raw dataset in *Figure 4* (pg.9), this ACF no longer displays obvious patterns on every lag. It has obvious non-zero at lag = 1 and lag = 168, also some non-zeros around lag = 1 and lag = 168, and close to zeros at all other lags.

# IV. Time Series Decomposition

## a. STL Decomposition Method

Applying the STL (Seasonal and Trend Decomposition using Loss) decomposition method to approximate the trend, seasonality, and reminder.

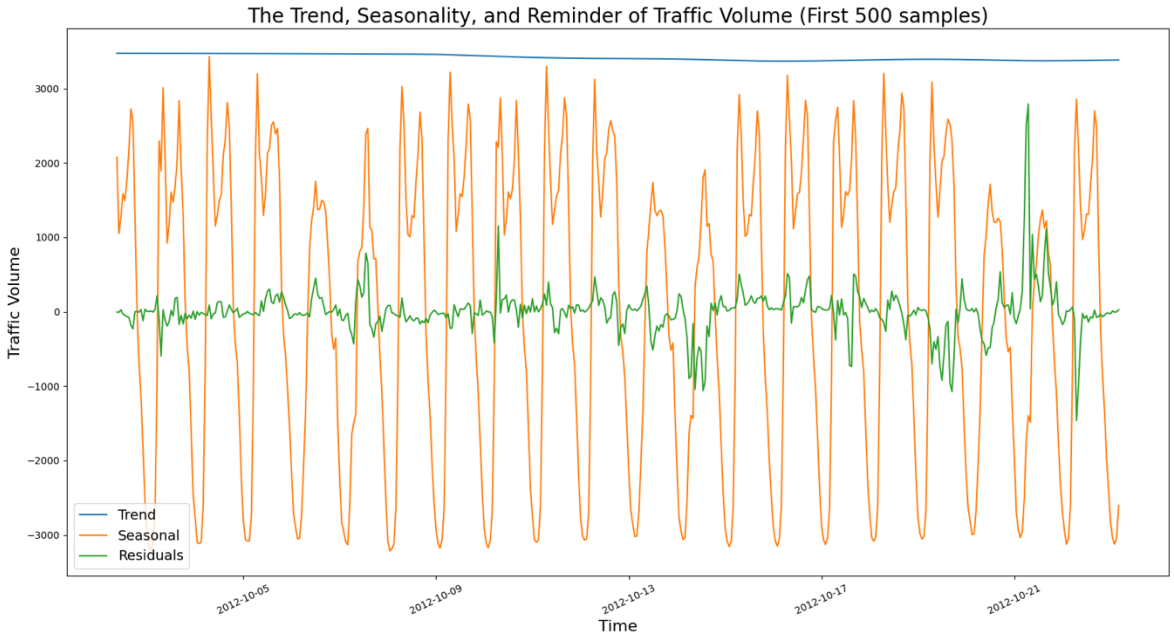



***Figure 10.*** Plot of the Trend, Seasonality, and Reminder of Traffic Volume (First 500 samples)

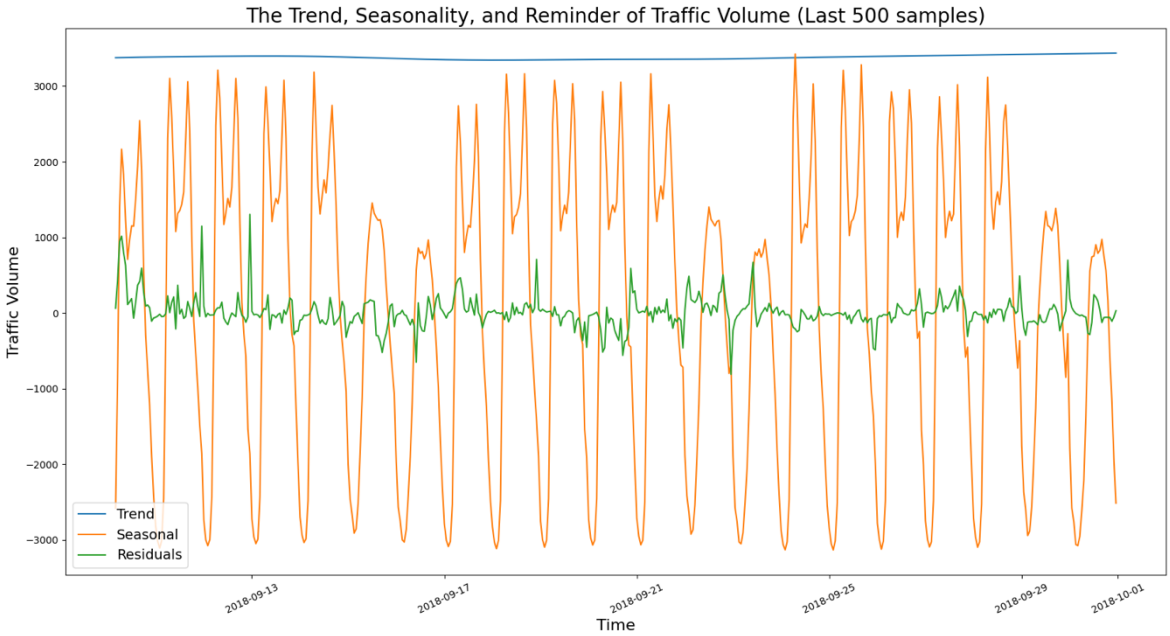



***Figure 11.*** Plot of the Trend, Seasonality, and Reminder of Traffic Volume (Last 500 samples)

From *Figure 10* and *Figure 11*, they show the trend is constant. Since the magnitude of seasonal fluctuation does not vary with level, it used additive decomposition. Thus, to generate the de-trended dataset and seasonally adjusted dataset by subtracting trend and seasonality from the raw dataset.

F

## b. Plot of the Raw Dataset versus the De-trended Dataset and Seasonally Adjusted Dataset

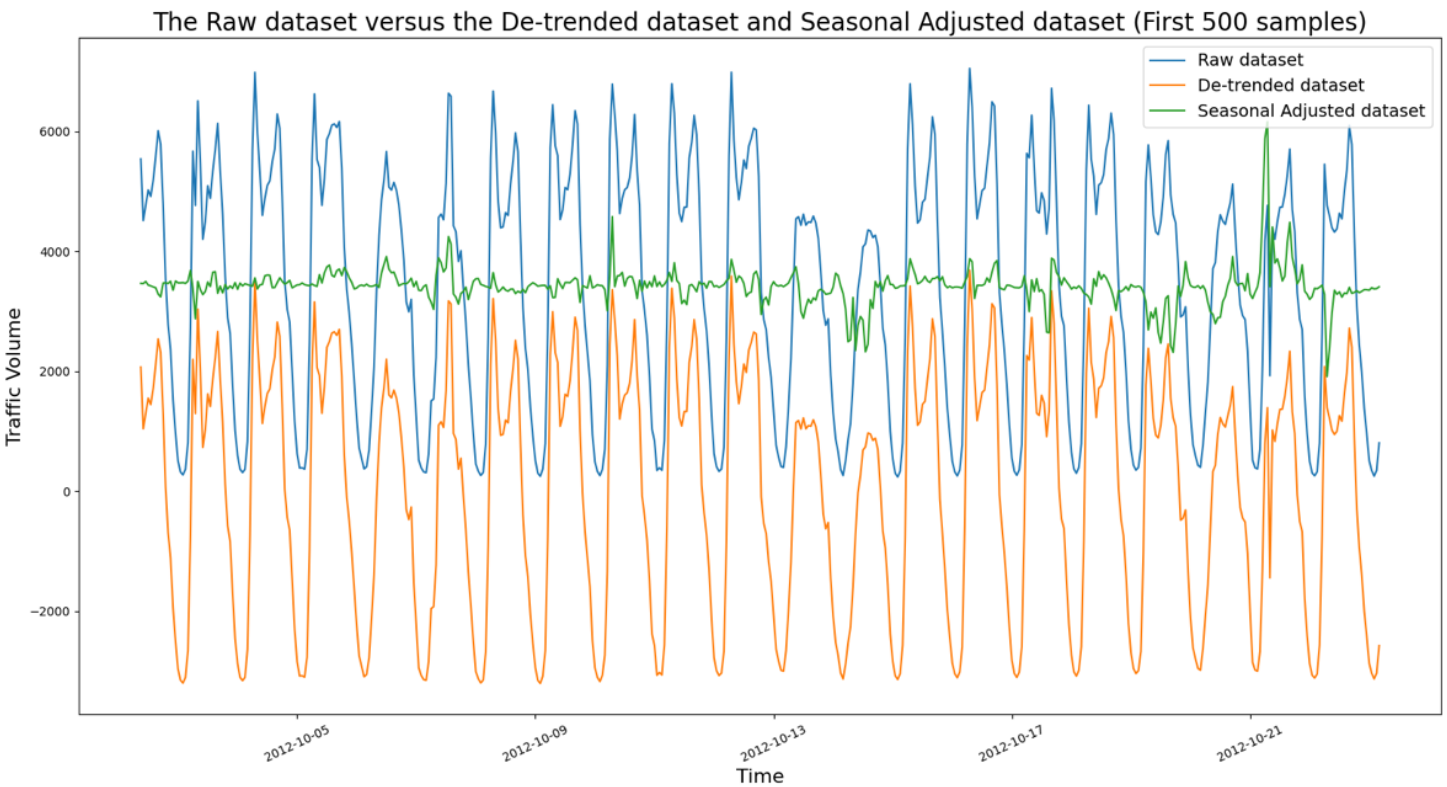



***Figure 12.*** Plot of the Raw Dataset versus the De-trended Dataset and Seasonally Adjusted Dataset

## c. Strength of the Trend and the Seasonality

```
The strength of trend for the dataset is: 0.9999999022936712
The strength of seasonality for the dataset is: 0.9999995640033321
```

***Figure 13.*** The Strength of the Trend and the Seasonality

The strength of the trend and seasonality both are close to 1. Thus, this dataset has a strong trend and seasonality.

# V. Holt-Winter Method

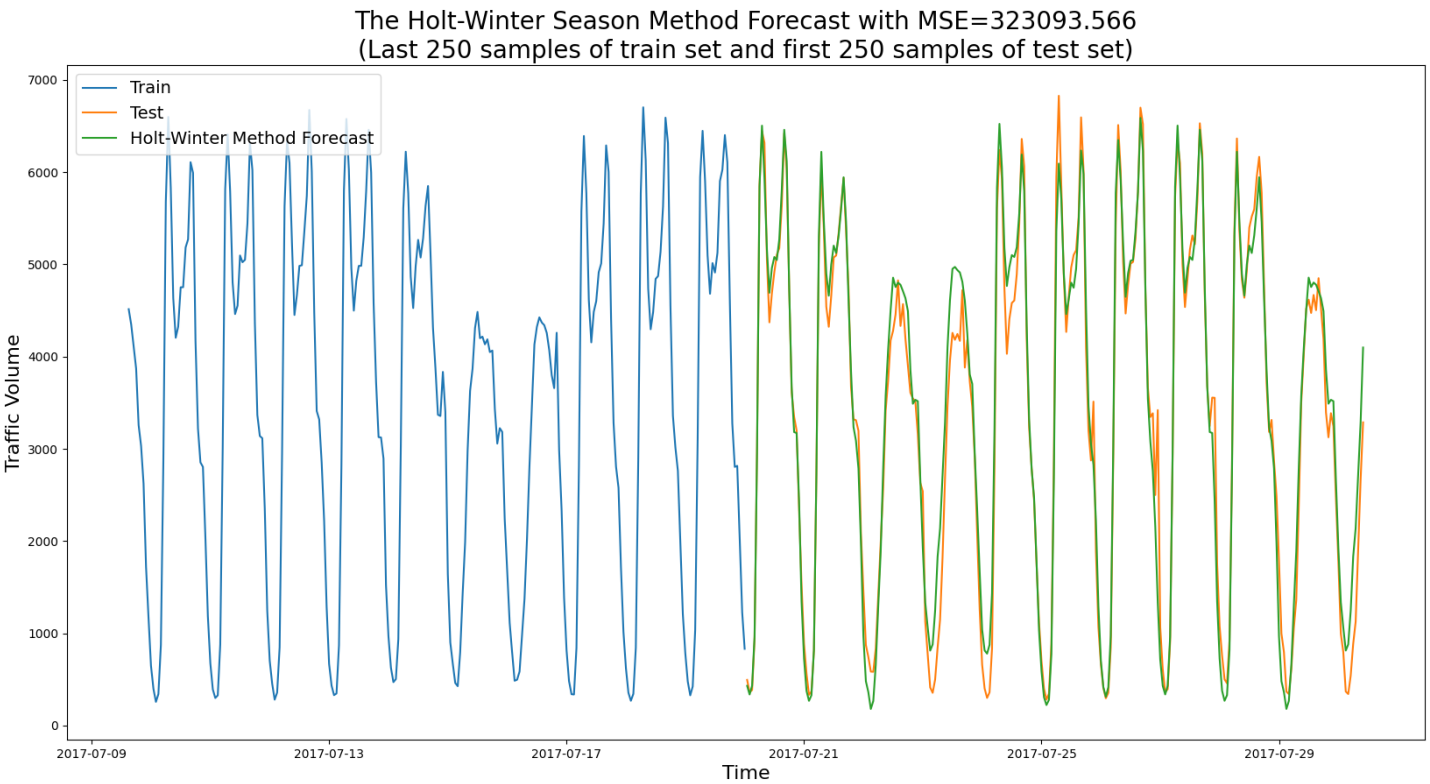



***Figure 14.*** Plot of the Holt-Winter Season Method Forecast

```
The mean of the forecast error of Holt-Winter Method is -173.642
The variance of the forecast error of Holt-Winter Method is 292941.916
The Q value of the forecast error of Holt-Winter Method is 42912.674
```

***Figure 15.*** Mean, Variance, and Q value of the Forecast Error of Holt-Winter Method

The Holt-Winter did a good job on forecasting traffic volume that exhibits both a trend and a seasonal variation. In *Figure 14*, the forecast is similar to the test set.

# VI. Feature Selection

## a. Split the Whole Dataset into Train and Test Set

Split the whole dataset into train set (80%) and test set (20%)

## b. Collinearity Detection

### i. Singular Values Analysis

```
The Singular Values of the raw dataset is
 [3.29877152e+09 9.66769254e+07 6.00854893e+07 9.89448666e+03
 4.43254690e+03 3.63818579e+03 2.24000481e+03 9.58183058e+02
 5.96537323e+02 4.76176197e+02 2.20376316e+02 9.08586789e+01
 5.29185584e+01 1.64223293e+01 1.29977258e+00 1.08999732e+00
 2.26365803e-10]
```

***Figure 16.*** The Singular Values of the Raw Dataset

There is one of singular value (2.26365803e-10) close to zero, which means that one or more feature is correlated. The collinearity exists in this dataset.

### ii. Condition Number

```
The condition number of the raw dataset is 4.050636093231158e+17
```

***Figure 17.*** The Condition Number of the Raw Dataset

The condition number is much higher than 1,000, which means there is severe degree of collinearity exists.

Based on the observation from the Heatmap *Figure 5* (pg.10), only two features have a relatively strong correlation with *traffic_volume*. There may be about two or three features left after feature reduction.

## c. Feature Reduction

Using a backward stepwise regression reduce the feature space dimension. First, generate the multiple linear regression model containing all potential predictors by using OLS function. Then, remove one predictor at a time.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         traffic_volume   R-squared:                       0.040
Model:                            OLS   Adj. R-squared:                  0.040
Method:                 Least Squares   F-statistic:                     90.64
Date:                Tue, 04 May 2021   Prob (F-statistic):          6.61e-275
Time:                        23:16:10   Log-Likelihood:            -2.9193e+05
No. Observations:               32460   AIC:                         5.839e+05
Df Residuals:                   32444   BIC:                         5.840e+05
Df Model:                          15
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                   -3274.6339    277.321    -11.808      0.000   -3818.194   -2731.074
holiday                 -2398.9414    297.356     -8.068      0.000   -2981.770   -1816.113
temp                       22.0121      0.829     26.561      0.000      20.388      23.636
rain_1h                     0.1759      0.198      0.888      0.375      -0.213       0.564
snow_1h                   -17.1519   1711.426     -0.010      0.992   -3371.610    3337.307
clouds_all                  2.7994      0.463      6.042      0.000       1.891       3.707
weather_main_Clear        145.1102    172.954      0.839      0.401    -193.887     484.107
weather_main_Clouds       469.9697    171.551      2.740      0.006     133.724     806.215
weather_main_Drizzle      -69.9105    193.583     -0.361      0.718    -449.341     309.520
weather_main_Fog         -336.9471    218.491     -1.542      0.123    -765.198      91.304
weather_main_Haze         665.3154    185.097      3.594      0.000     302.518    1028.113
weather_main_Mist        -171.3682    174.357     -0.983      0.326    -513.114     170.378
weather_main_Rain         102.6721    173.958      0.590      0.555    -238.293     443.637
weather_main_Smoke       -956.3906    522.266     -1.831      0.067   -1980.052      67.271
weather_main_Snow         158.6635    176.159      0.901      0.368    -186.615     503.942
weather_main_Squall      -2772.9132   1786.933     -1.552      0.121   -6275.368     729.542
weather_main_Thunderstorm -508.8354   199.626     -2.549      0.011    -900.109    -117.562
==============================================================================
Omnibus:                    23677.028   Durbin-Watson:                   0.242
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1897.051
Skew:                          -0.060   Prob(JB):                         0.00
Kurtosis:                       1.822   Cond. No.                     4.56e+17
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.27e-26. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

***Figure 18.*** The Multiple Linear Regression Model Containing all Potential Predictors

Step 1: Remove '*snow_1h'* feature because it has highest p-value (0.992) in the t-test, which means this coefficient is not significant.

Step 2: Remove '*weather_main_Drizzle'* feature because it has a highest p-value (0.718) in the t-test after previous step, which means this coefficient is not significant. After the first two step, the AIC, BIC, and Adjusted R-squared has no change. A huge drop of the condition number form 4.05e+17 to 5.15e+04.

Step 3: Remove '*rain_1h*' feature because it has a highest p-value (0.375) in the t-test after previous steps, which means this coefficient is not significant.

Step 4: Remove '*weather_main_Mist*' feature because it has a highest p-value (0.343) in the t-test after previous steps, which means this coefficient is not significant. The AIC, BIC, and Adjusted R-squared still has no change. The condition number decrease from 5.15e+04 to 5.14e+04.

Step 5: Remove '*weather_main_Fog*' feature because it has a highest p-value (0.245) in the t-test after previous steps, which means this coefficient is not significant.

Step 6: Remove '*weather_main_Squall*' feature because it has a highest p-value (0.181) in the t-test after previous steps, which means this coefficient is not significant. The AIC, BIC, and Adjusted R-squared still has no change, same as the begin. The condition number decrease from 5.15e+04 to 1.43e+04.

Step 7: Remove '*weather_main_Smoke*' feature because it has a highest p-value (0.146) in the t-test after previous steps, which means this coefficient is not significant. The AIC, BIC, and Adjusted R-squared still has no change, same as the begin. The condition number decrease from 1.43e+04 to 7.85e+03.

Step 8: Remove '*const'* feature because it's not a feature and has a high standard error (239.715). The AIC and BIC have slightly increase. A huge increase on the Adjusted R-squared which increase from 0.40 to 0.741.

Step 9: Remove '*weather_main_Thunderstorm'* feature because it's it has a highest p-value (0.036) and large standard error.

Step 10: Remove '*weather_main_Snow'* feature because it's it has a highest p-value (0.006) and large standard error.

Step 11: Remove '*clouds_all'* feature because it's it has a highest p-value (0.001) and it has strong correlation with two of the rest features. The AIC, BIC, and Adjusted R-squared has no change. The condition number decrease from 7.85e+03 to 7.75e+03.

Step 12: Remove '*weather_main_Clear*' feature because it has a highest p-value (0.030) in the t-test after previous steps.

Step 13: Remove '*weather_main_Clear*' feature because it has a highest p-value (0.030) in the t-test after previous steps.

Step 13: Remove '*holiday'* feature because it has a highest standard error (298.393) when all p-value are zero. The AIC has a slightly increase and the Adjusted R-squared decrease 0.001. The condition number decrease from 7.75e+03 to 2.11e+03.

Step 14: Remove '*weather_main_Haze'* feature because it has a highest standard error (81.030) when all p-value are zero. The BIC has a slightly increase. The condition number decrease from 2.11e+03 to 940.

Stop feature reduction because the condition number is lower than 1,000 and p-value of all coefficients are zero. Only three feature left in the end.

Final Multiple Linear Regression Model

```
                                 OLS Regression Results
==============================================================================
Dep. Variable:         traffic_volume   R-squared (uncentered):              0.740
Model:                            OLS   Adj. R-squared (uncentered):         0.740
Method:                 Least Squares   F-statistic:                     3.076e+04
Date:                Wed, 05 May 2021   Prob (F-statistic):                   0.00
Time:                        00:30:12   Log-Likelihood:                -2.9211e+05
No. Observations:               32460   AIC:                             5.842e+05
Df Residuals:                   32457   BIC:                             5.843e+05
Df Model:                           3
Covariance Type:            nonrobust
==========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
temp                    10.9367      0.055    198.485      0.000      10.829      11.045
weather_main_Clouds    486.3230     23.353     20.825      0.000     440.551     532.095
weather_main_Rain      203.2478     35.164      5.780      0.000     134.326     272.170
==============================================================================
Omnibus:                    45793.638   Durbin-Watson:                   0.231
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2128.915
Skew:                          -0.082   Prob(JB):                         0.00
Kurtosis:                       1.756   Cond. No.                         940.
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

***Figure 19.*** The Multiple Linear Regression Model after Features Reduction

```
The Singular Values of the feature reduction dataset is
 [2.57915630e+09 8.21201646e+03 2.92034943e+03]
```

```
The condition number of the feature reduction dataset is 939.7696617431238
```

***Figure 20.*** The Singular Values and Condition Number of the Feature Reduction Dataset

The final model does not have any singular value close to zero, and the condition number is less than 1000. It means the feature reduction dataset does not exist collinearity.

# VII. Multiple Linear Regression

## a. Hypothesis Tests Analysis

### i. F-test

```
F-statistic:                     3.076e+04
Prob (F-statistic):                   0.00
```

***Figure 21.*** The F-statistic and p-value of F-test

Since the F-statistic is large and the p-value of F-test is much less than 0.05, I have sufficient evidence to reject the null hypothesis at significant level 0.05. Therefore, I can conclude that my multiple linear regression model provides a better fit than the intercept-only.

### ii. T-test

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| temp | 10.9367 | 0.055 | 198.485 | 0.000 | 10.829 | 11.045 |
| weather_main_Clouds | 486.3230 | 23.353 | 20.825 | 0.000 | 440.551 | 532.095 |
| weather_main_Rain | 203.2478 | 35.164 | 5.780 | 0.000 | 134.326 | 272.170 |

***Figure 22.*** The t-statistic and p-value of t-test

Since the p-value of all coefficients are much less than 0.05, I have sufficient evident to reject the null hypothesis at significant level 0.05. Therefore, I can conclude that all the coefficients in my multiple linear regression model are significant.

## b. AIC, BIC, R-squared and Adjusted R-squared

```
The AIC of MLR model is 584233.285
The BIC of MLR model is 584258.448
```

***Figure 23.*** The AIC and BIC of the Multiple Linear Regression Model

```
R-squared (uncentered):                      0.740
Adj. R-squared (uncentered):                 0.740
```

***Figure 24.*** The R-squared and Adjusted R-squared of the Multiple Linear Regression Model

The AIC and BIC are large, which means this multiple linear regression model may not fits the dataset and not well for forecasting. The R-squared and Adjusted R-squared are 0.74, reveals that 74% of the data fit the multiple regression linear regression model.

## c. Perform One-Step ahead Forecast

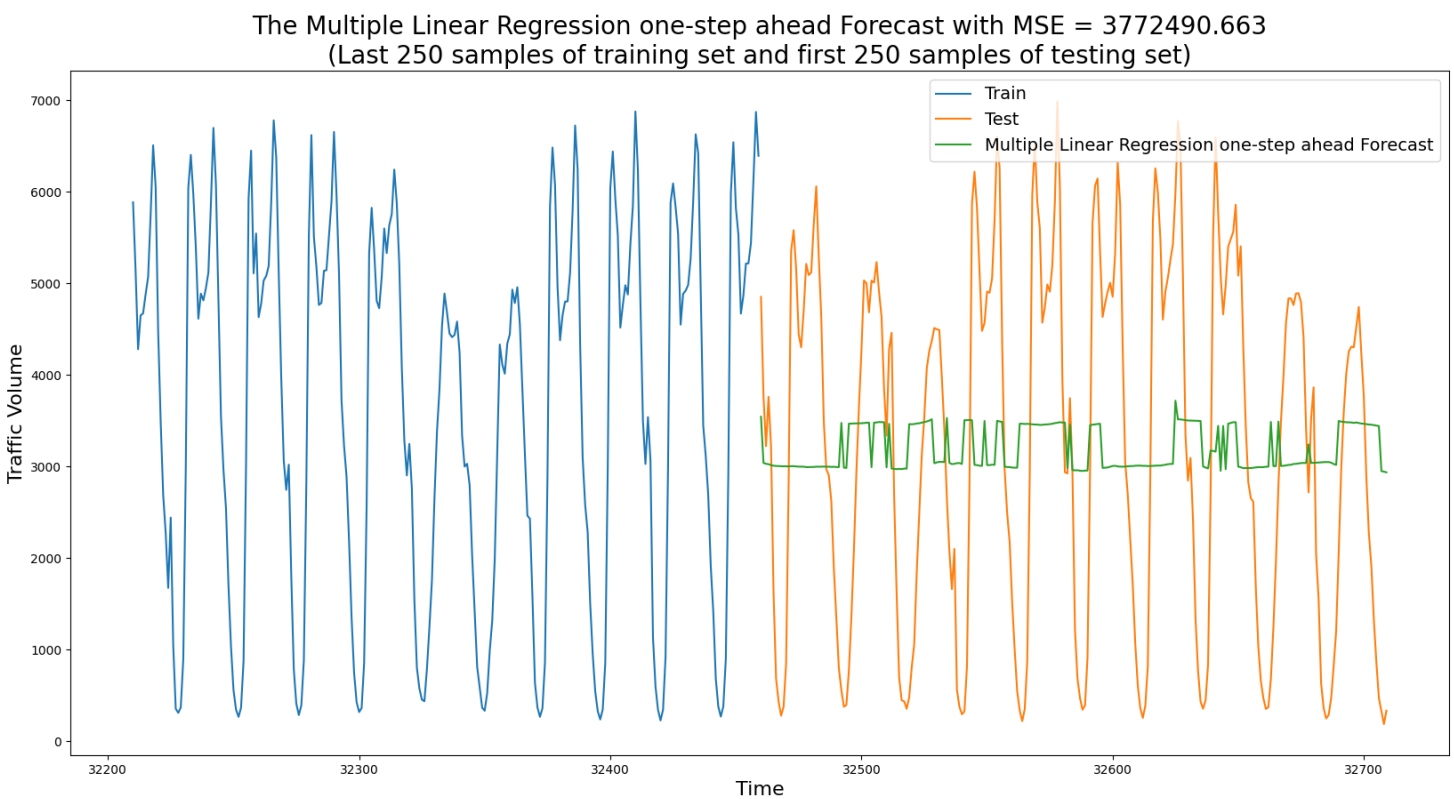



***Figure 25.*** Plot of the Multiple Linear Regression one-step ahead Forecast

```
The mean of the forecast error of MLR model is 72.238
The variance of the forecast error of MLR model is 3767272.396
The Q value of the forecast error of MLR model is 162366.193
```

***Figure 26.*** The Mean, Variance, and Q value of the Forecast error of the Multiple Linear Regression Model

## d. Residuals Analysis

### i. RMSE

```
The RMSE of MLR model is 1968.701
```

***Figure 27.*** The RMSE of Multiple Linear Regression Model

### ii. ACF of Residuals

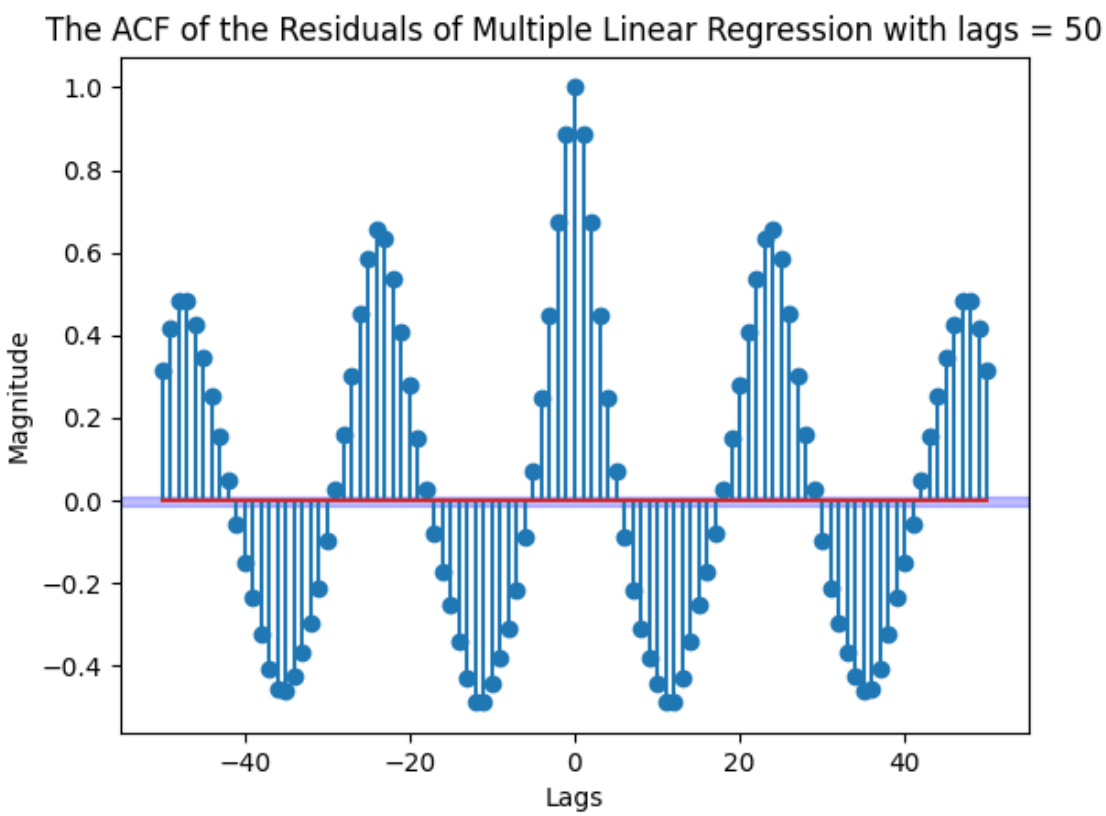



***Figure 28.*** The ACF of the Residuals of Multiple Linear Regression Model with lags = 50

iii. Mean, Variance and Q value of Residuals

```
The mean of the residuals of MLR model is -6.133
The variance of the residuals of MLR model is 3875784.079
The Q-value of the residuals of MLR model is 363383.16
```

***Figure 29.*** The Mean, Variance, and Q-value of the Residuals of Multiple Linear Regression Model

In conclusion, the multiple linear regression model does not do well on forecasting this dataset. First of all, the AIC and BIC of this MLR model are both very large. In addition, its R-squared and Adjusted R-squared show that only 74% of the data fit this MLR model. Secondly, from the analysis of forecast error. It can be clearly seen from *Figure 25* (pg.) that the forecast capability of this MLR model is weak, and the variance and Q-value of forecast error are large. From residuals analysis, its variance and Q-value are also large, and its ACF of the residuals still shows obvious patterns, which means that the MLR model has not captured the seasonality pattern in the dataset. This result is not unexpected. Multiple linear regression does not apply to this robust seasonal dataset with a weak correlation between features.

# VIII. ARMA, ARIMA, and SARIMA Model

## a. Order Determination

### i. GPAC Table

Since the seasonal period of this dataset is 168, a pattern with a constant column and a row of zeros must be displayed on columns and rows multiples of 168. It cannot show the GPAC table over 100 x 100. Therefore, for this dataset, I could not estimate the order of my SARIMA model by GPAC Table.

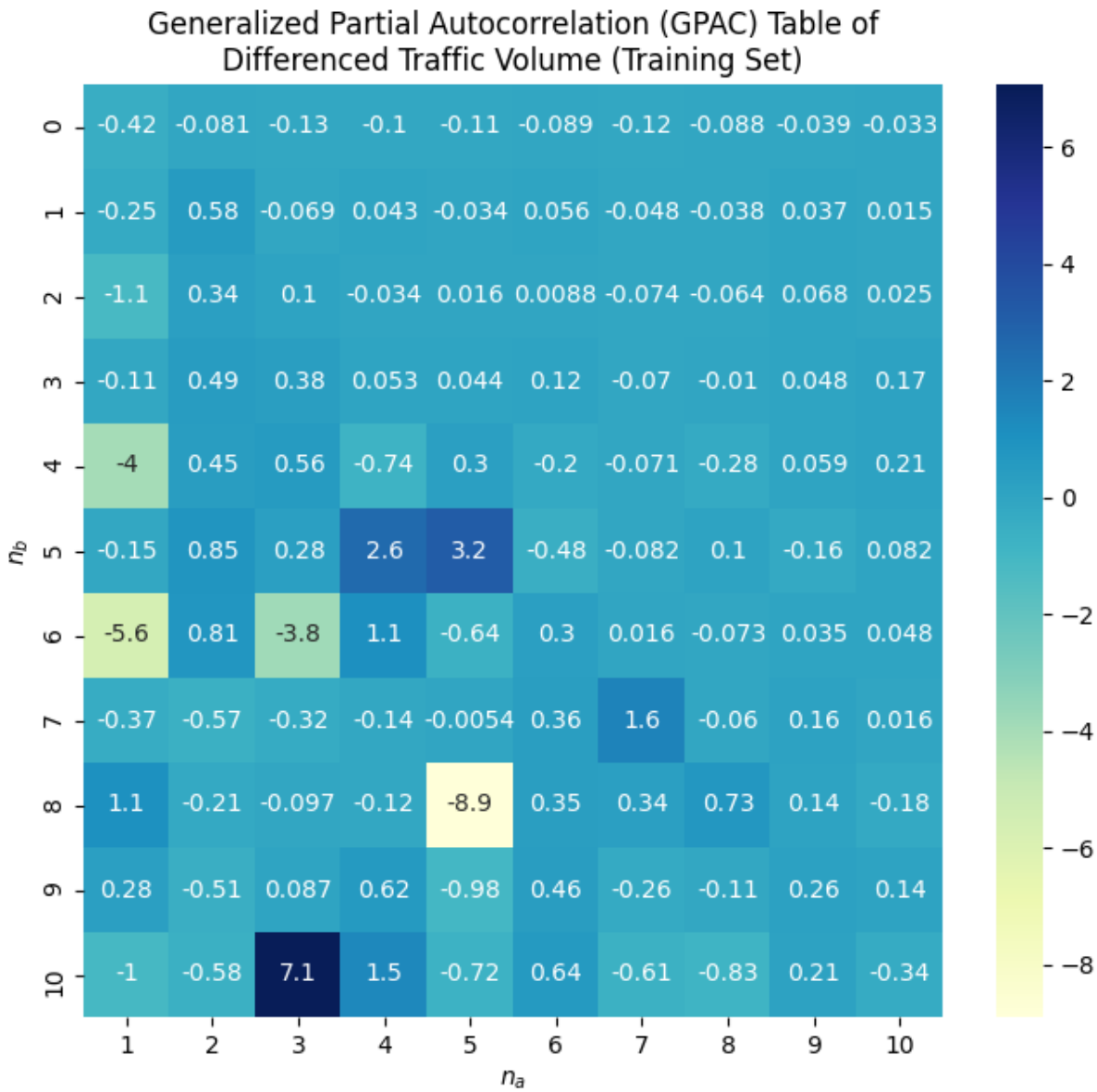



***Figure 30.*** The GPAC Table of the Differenced Dataset (Not useful for S = 168)

## ii. ACF/PACF Plot

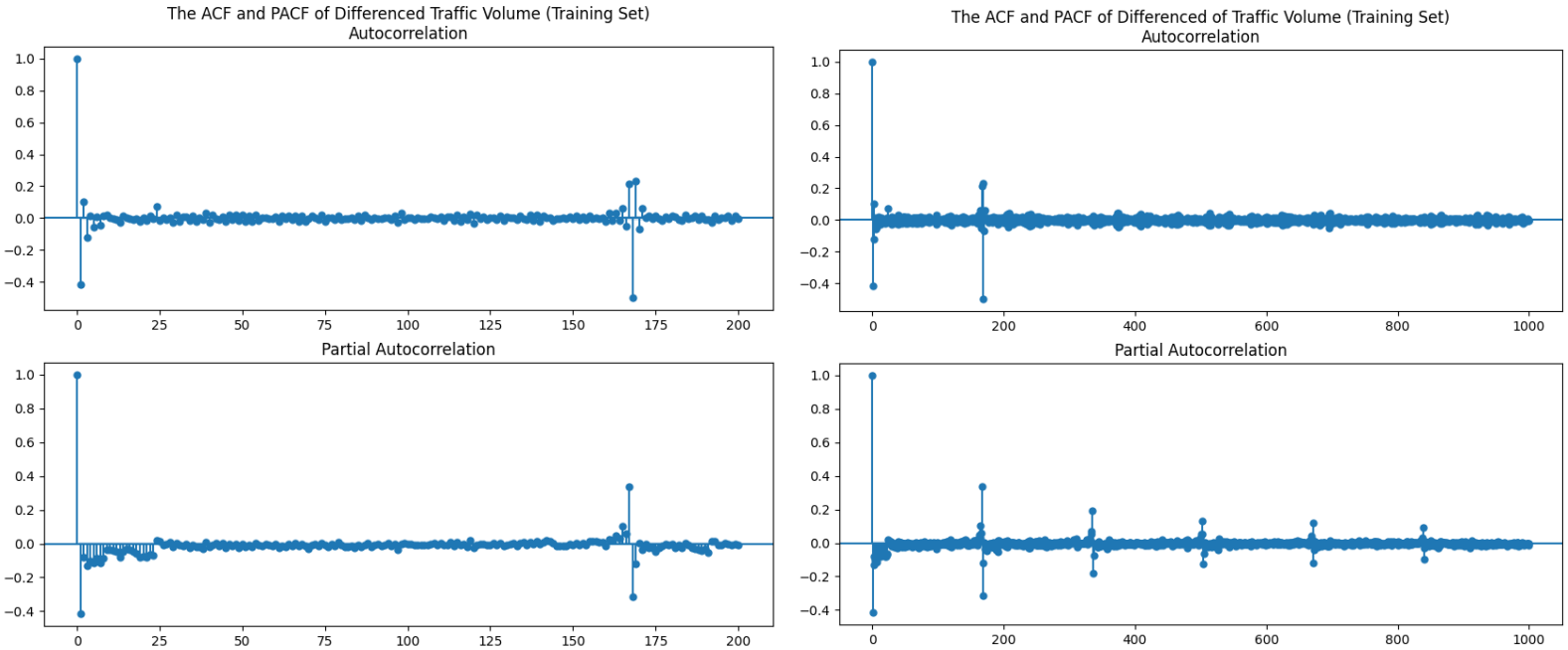



***Figure 31.*** Plot of the ACF /PACF of the Differenced Dataset with lags = 200 (left) and lag = 1000 (right)

The PACF (right) shows a decay across seasonal lags k = 168, 336, 504,…, and the ACF (right) shows nonzero autocorrelation at the first seasonal lag k = 168. This reflects the final model include at least a seasonal ARIMA model has $n_b$ order. Also, since the PACF and ACF has non-zero at lag k = 1, 2, 3 and additional nonzero autocorrelation at lags k = 165, 166, 167, 168, 169, 170, 171, I can estimated this is a multiplicative of non-seasonal ARIMA model and seasonal ARIMA model ARIMA(0,1,3) × ARIMA$(0,1,1)_{168}$. It is mathematically equivalent to a nonseasonal ARMA(0,171) process with parameters: $b_1 = b_2 = \ldots = b_3 = 0$, $b_{168} = B_1$, $b_{169} = b_1B_1$, $b_{170} = b_2B_1$, and $b_{171} = b_3B_1$.

# IX. Estimated Parameters of ARMA Model

Based on the estimated order from *Figure 31*, I tried to use the ARMA model to generate ARMA(0,171) process, but it took about three hours and could not return the result. I also tried to use the SARIMA model to generate ARIMA(0,1,3)×ARIMA(0,1,1)$_{168}$ process, in the same case, and could not return the result. The reason may be the number of orders is too large in the ARMA model and the seasonal period is too large in the SARIMA model. Therefore, I tried a smaller number in the ARMA model and found that ARMA(2,2) model has the best performance with the smallest AIC and Q-value of residuals.

```
                              ARMA Model Results
==============================================================================
Dep. Variable:                      y   No. Observations:                41905
Model:                     ARMA(2, 2)   Log Likelihood             -318255.367
Method:                       css-mle   S.D. of innovations            480.934
Date:                Wed, 05 May 2021   AIC                         636520.734
Time:                        09:31:33   BIC                         636563.950
Sample:                             0   HQIC                        636534.384

==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1.y        0.1934      0.020      9.708      0.000       0.154       0.232
ar.L2.y        0.3454      0.010     34.947      0.000       0.326       0.365
ma.L1.y       -0.7514      0.021    -35.902      0.000      -0.792      -0.710
ma.L2.y       -0.2159      0.020    -10.936      0.000      -0.255      -0.177
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.4445           +0.0000j            1.4445            0.0000
AR.2           -2.0045           +0.0000j            2.0045            0.5000
MA.1            1.0276           +0.0000j            1.0276            0.0000
MA.2           -4.5085           +0.0000j            4.5085            0.5000
-----------------------------------------------------------------------------
```

***Figure 32.*** The ARMA(2,2) Process

# X. Diagnostic Analysis

## a. Diagnostic Test

### i. Confidence Interval

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| ar.L1.y | 0.1934 | 0.020 | 9.708 | 0.000 | 0.154 | 0.232 |
| ar.L2.y | 0.3454 | 0.010 | 34.947 | 0.000 | 0.326 | 0.365 |
| ma.L1.y | -0.7514 | 0.021 | -35.902 | 0.000 | -0.792 | -0.710 |
| ma.L2.y | -0.2159 | 0.020 | -10.936 | 0.000 | -0.255 | -0.177 |

***Figure 33.*** The Estimated Parameters, the Standard Error, and the Confidence Intervals

The confidence intervals of all the estimated parameters do not contain zero, which means the estimated coefficients is important.

### ii. Zero / Poles Cancellation

```
The roots of numerator is
 [ 0.97323702 -0.22183702]
The roots of denominator is
 [-0.0967+0.57969743j -0.0967-0.57969743j]
```

***Figure 34.*** The Roots of Numerator and Denominator

There is no same roots of numerator and denominator, thus, no zero/pole cancellation.

### iii. Chi-Square Test

```
The residual is NOT white
The Q value is [1258.32575539]
The chi critical is 131.141216667052
The p-value of chi square test is [1.40757232e-199]
```

***Figure 35.*** Chi-Square Test of the Residuals of ARMA(2,2) Model

Since Q-value of the residual is larger than chi critical value and the p-value is much less than 0.05, we do not have sufficient evidence to show the residuals are white.

iv. <u>ACF of Residuals</u>

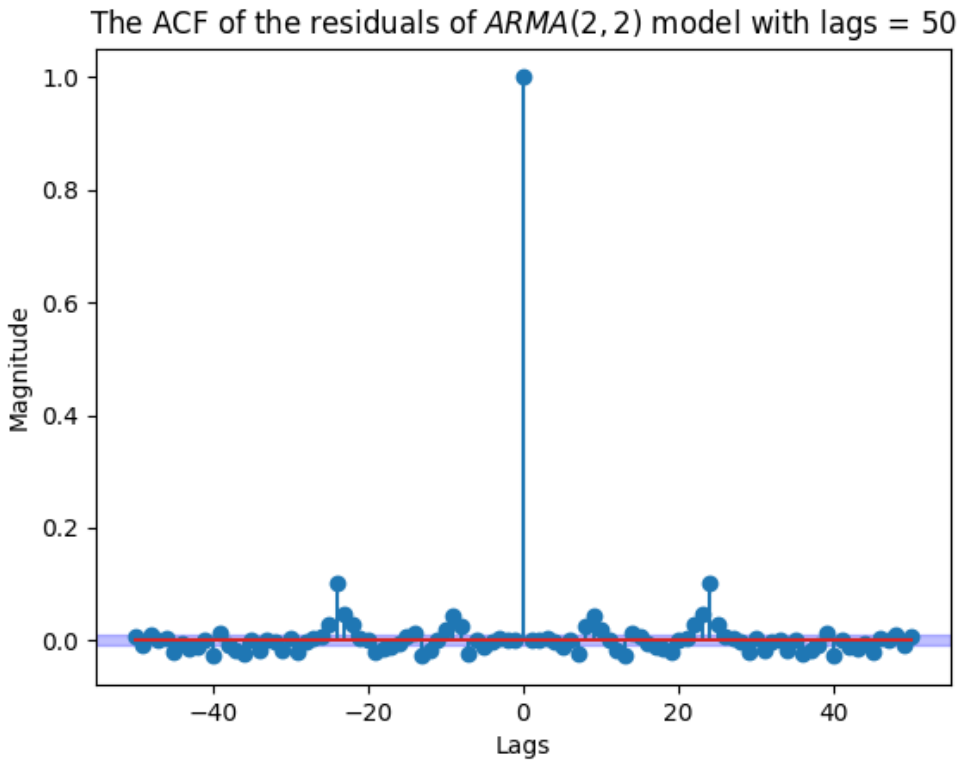



***Figure 36.*** The ACF of the Residuals of ARMA(2,2) Model with lags = 50

The ACF of residuals of ARMA(2,2) model is not white, which means that there exist a seasonal pattern is still not captured by the model.

## b. Estimated Variance of Error and Estimated Covariance of the Estimated Parameters

```
The estimated variance of error is 457221.63893733063
```

***Figure 37.*** The Estimated Variance of Error

```
The estimated covariance of the a1 is 0.00052802731634219
The estimated covariance of the a2 is 4.3615721515049795e-05
The estimated covariance of the a3 is 0.00039149247495931077
The estimated covariance of the b1 is 0.0003750385575653475
```

***Figure 38.*** The Estimated Covariance of the Estimated Parameters

## c. Biased or Unbiased Estimator

```
The mean of the residual of ARMA(2,2) is -0.039
```

***Figure 39.*** The Mean of Residual of ARMA(2,2) Process

Since the residuals of ARMA(2,2) process is close to zero, which means this is an unbiased estimator.

## d. The Variance of Residual Errors versus the Variance of the Forecast Errors

```
The variance of residual errors versus the variance of forecast errors is 0.935863574033126
```

***Figure 40.*** The Rate of the Variance of Residual Errors versus the Variance of Forecast Errors

The rate of the variance of residual errors versus the variance of forecast errors is close to 1, which means the model do a good job on forecasting.

## e. Forecast Analysis

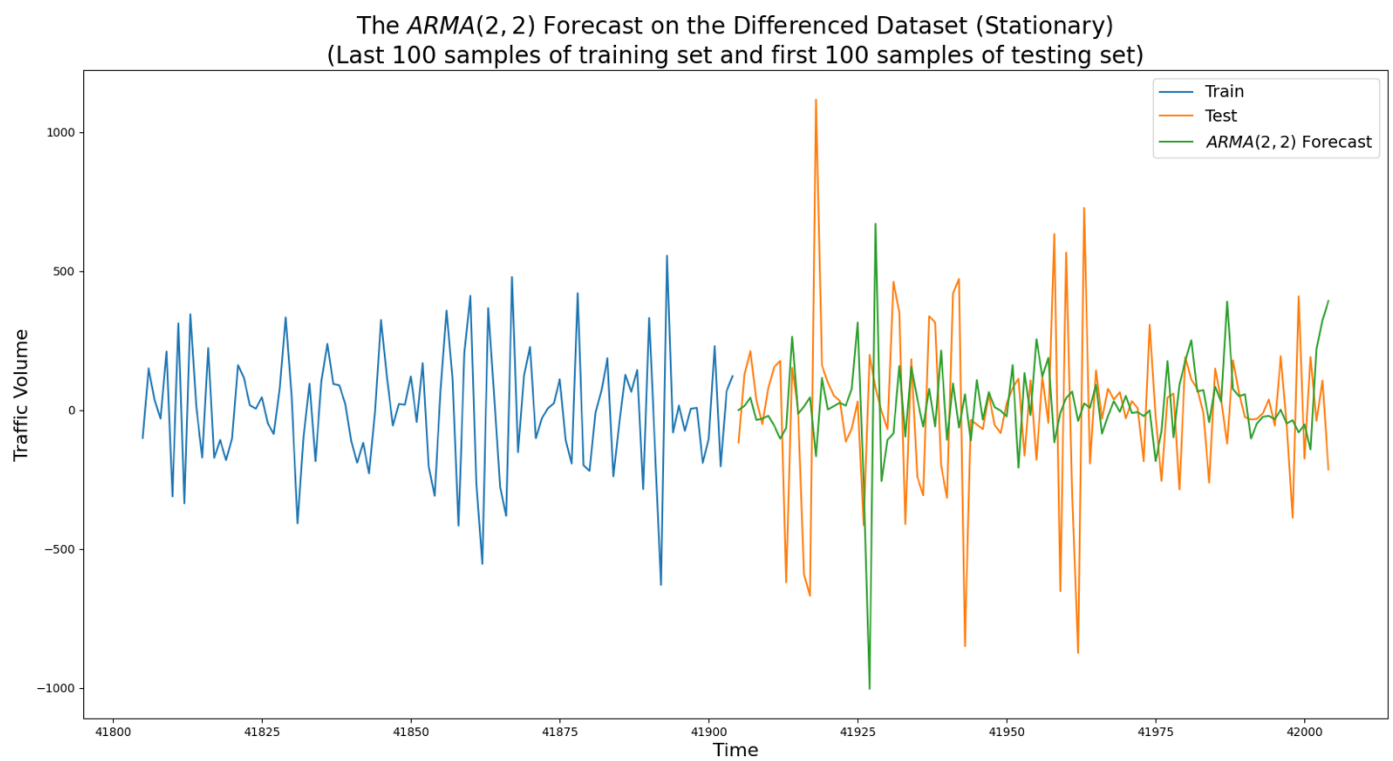



***Figure 41.*** The ARMA(2,2) Model Forecast on the Differenced Dataset (Stationary)

ARMA(2,2) model could only forecast the differenced dataset (stationary). Convert the ARMA(2,2) process to SARIMA model, which could use for forecast the traffic volume of the raw dataset (non-stationary).

The forecast function convert from ARMA(2,2) is ARIMA(2,1,2)×ARIMA(0,1,0)$_{168}$

$$(1-q^{-1})(1-q^{-168})(1-0.1934q^{-1}-0.3454q^{-2})y(t)=(1+0.7514q^{-1}+0.2159q^{-2})e(t)$$

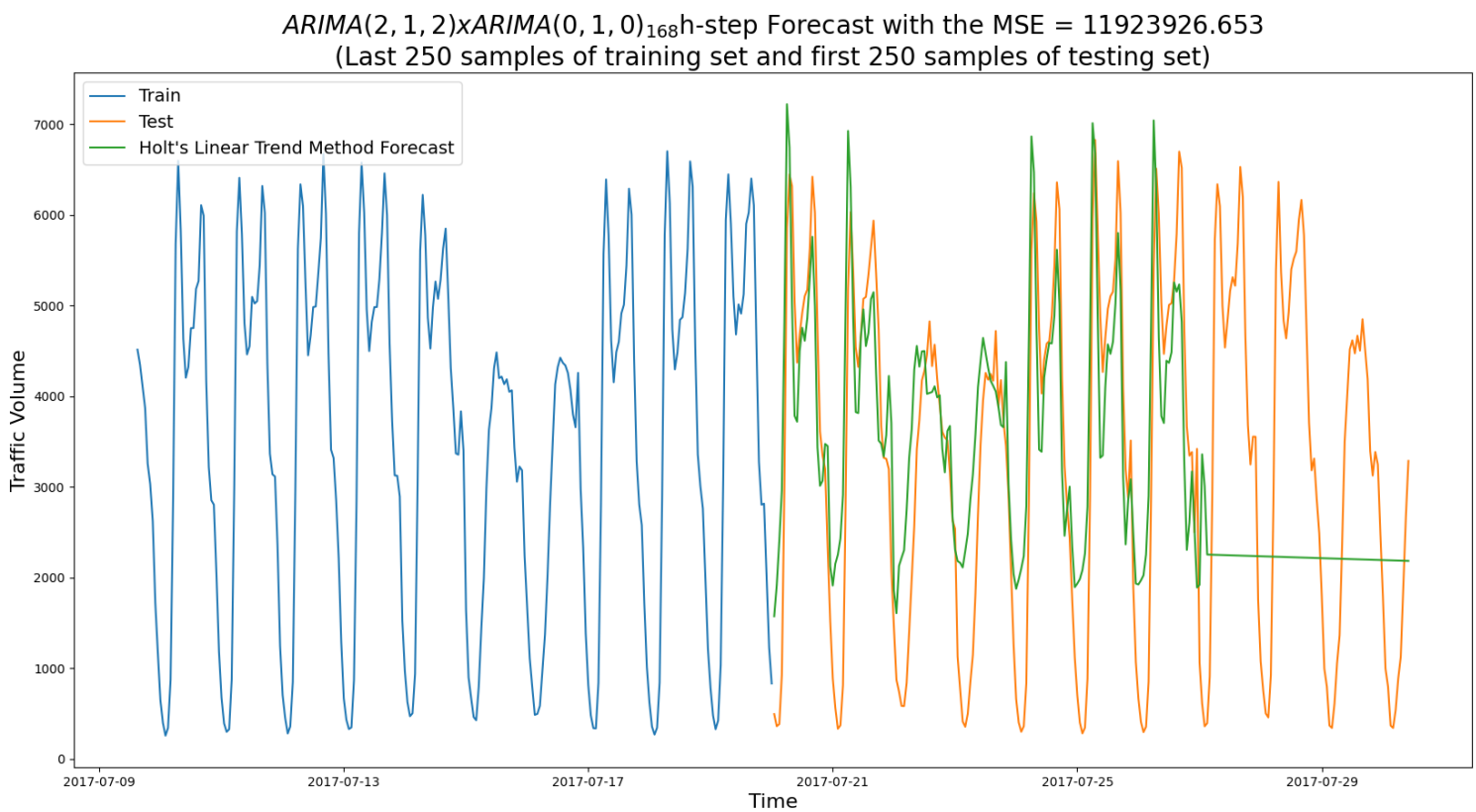



***Figure 42.*** The ARIMA(2,1,2)×ARIMA(0,1,0)$_{168}$ Model Forecast on the Raw Dataset (Non-Stationary)

```
The mean of the forecast error of ARIMA(2,1,2)xARIMA(0,1,0)_s168 is 2756.979
The variance of the forecast error of ARIMA(2,1,2)xARIMA(0,1,0)_s168 is 4322991.888
The Q value of the forecast error of ARIMA(2,1,2)xARIMA(0,1,0)_s168 is 200516.513
```

***Figure 43.*** The Mean, Variance, and Q-value of the Forecast Error of ARIMA(2,1,2)×ARIMA(0,1,0)$_{168}$ Model

# XI. Based Models

## a. Average Method

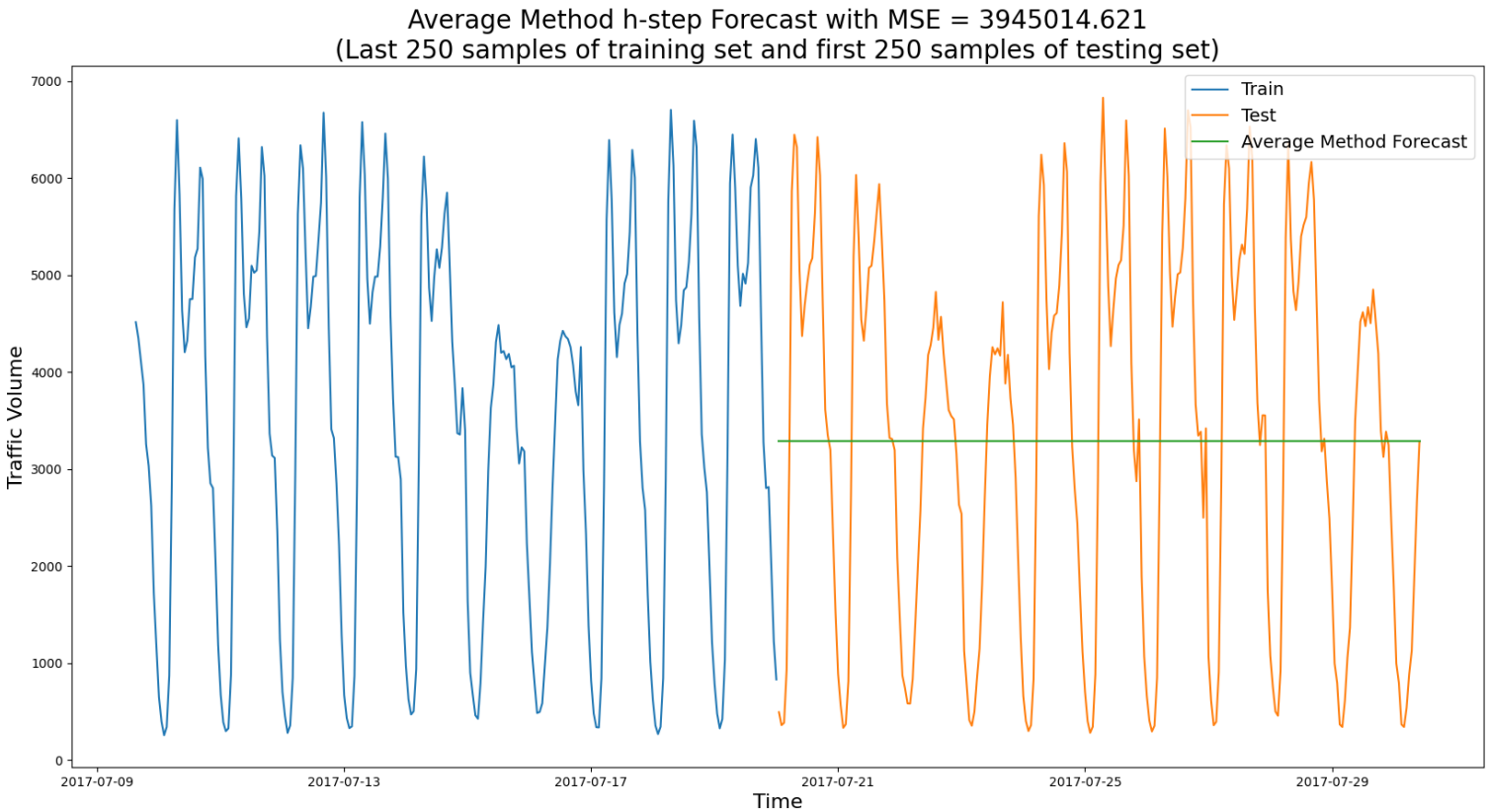



***Figure 44.*** The Average Method h-step Forecast

```
The mean of the forecast error of Average Method is 53.758
The variance of the forecast error of Average Method is 3942124.722
The Q value of the forecast error of Average Method is 230379.232
MSE for Average Method Forecast is 3945014.621
```

***Figure 45.*** The Mean, Variance, and Q-value of the Forecast Error of the Average Method

## b. Naïve Method

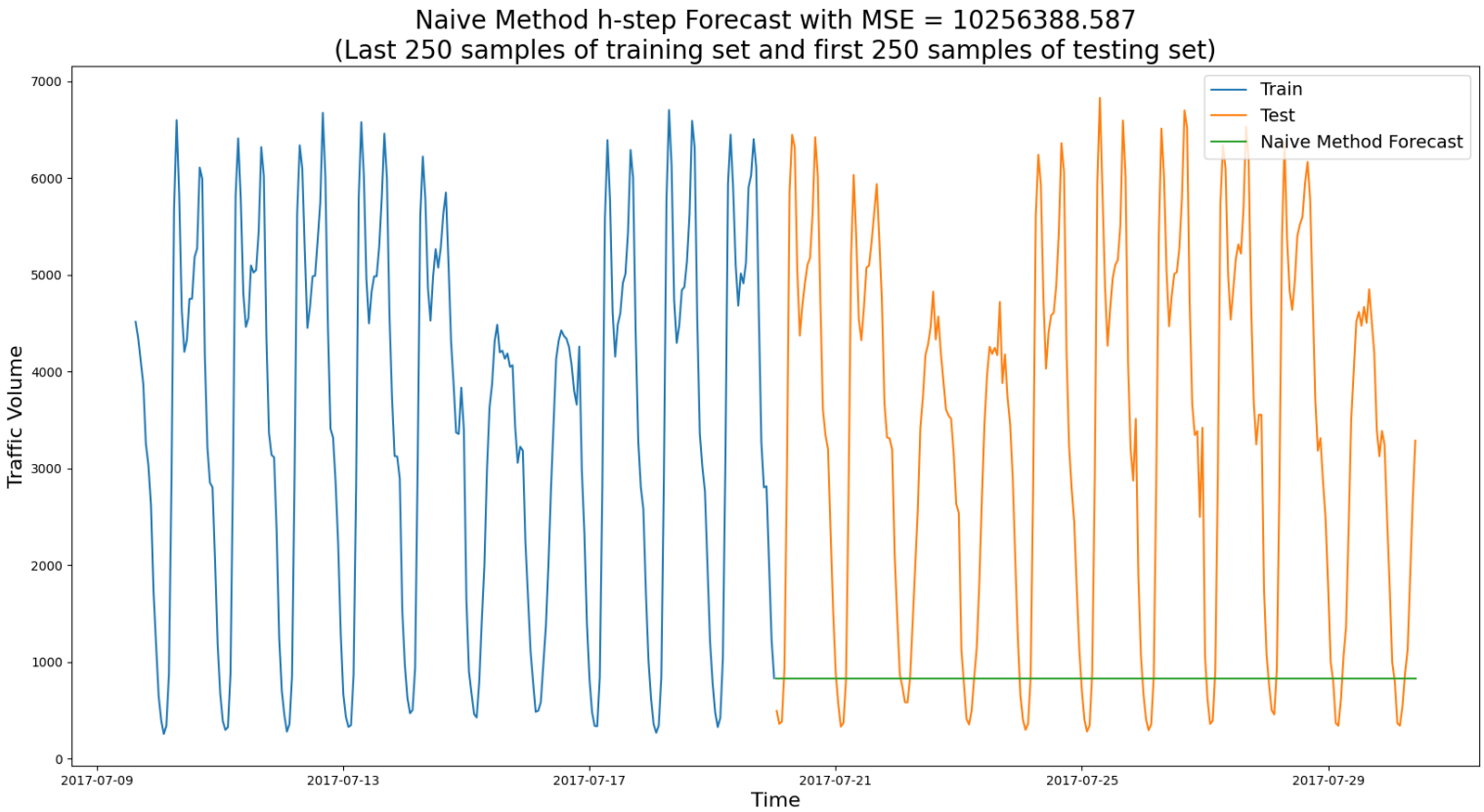
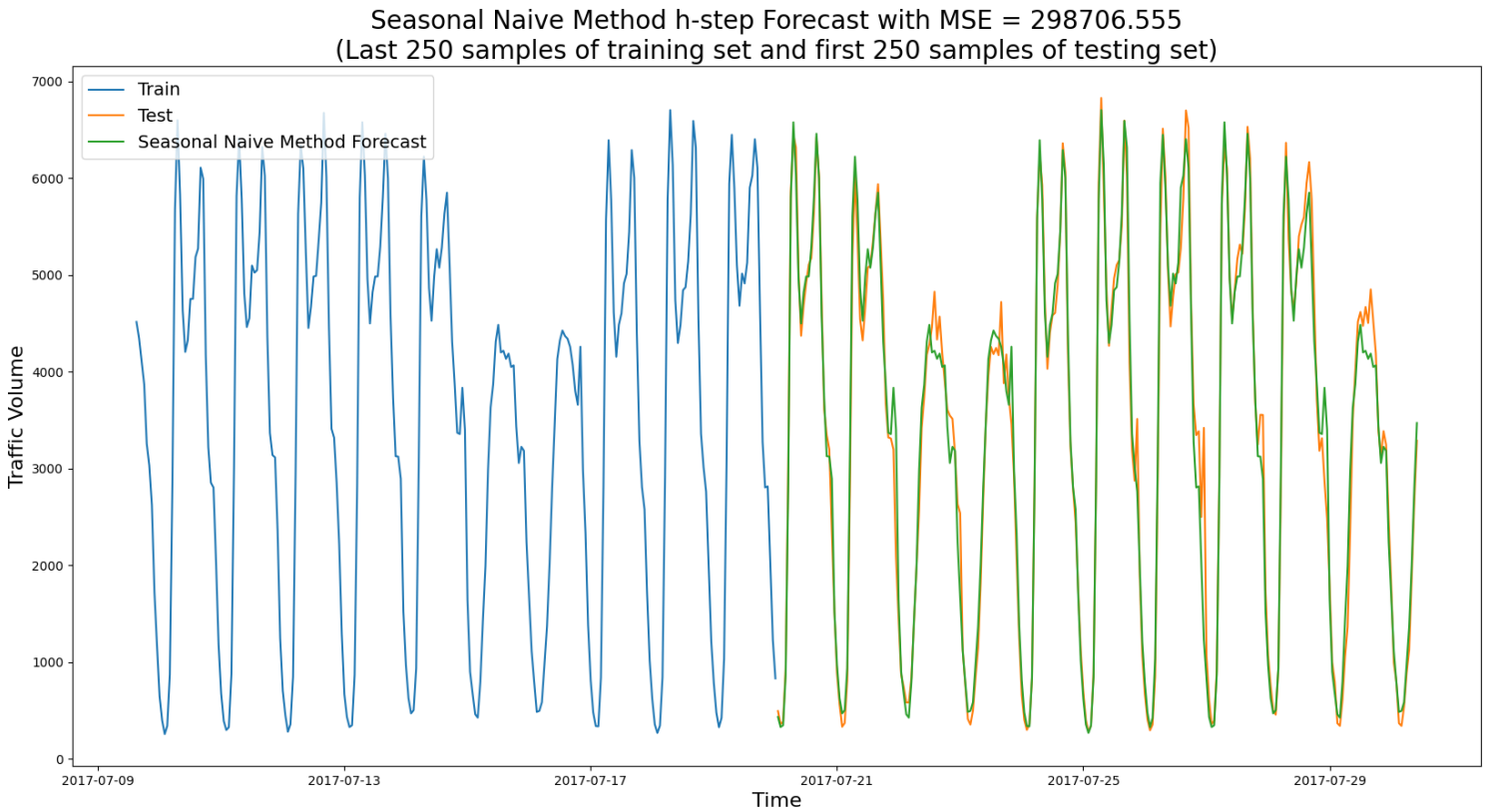



***Figure 46.*** The Naive Method h-step Forecast

```
The mean of the forecast error of Naive Method is 2512.82
The variance of the forecast error of Naive Method is 3942124.722
The Q value of the forecast error of Naive Method is 230379.232
```

***Figure 47.*** The Mean, Variance, and Q-value of the Forecast Error of the Naive Method

## c. Seasonal Naïve Method




## d.

***Figure 48.*** The Season Naive Method h-step Forecast

```
The mean of the forecast error of Seasonal Naive Method is -96.254
The variance of the forecast error of Seasonal Naive Method is 289441.664
The Q value of the forecast error of Seasonal Naive Method is 25900.393
```

***Figure 49.*** The Mean, Variance, and Q-value of the Forecast Error of the Seasonal Naive Method

## e. Drift Method

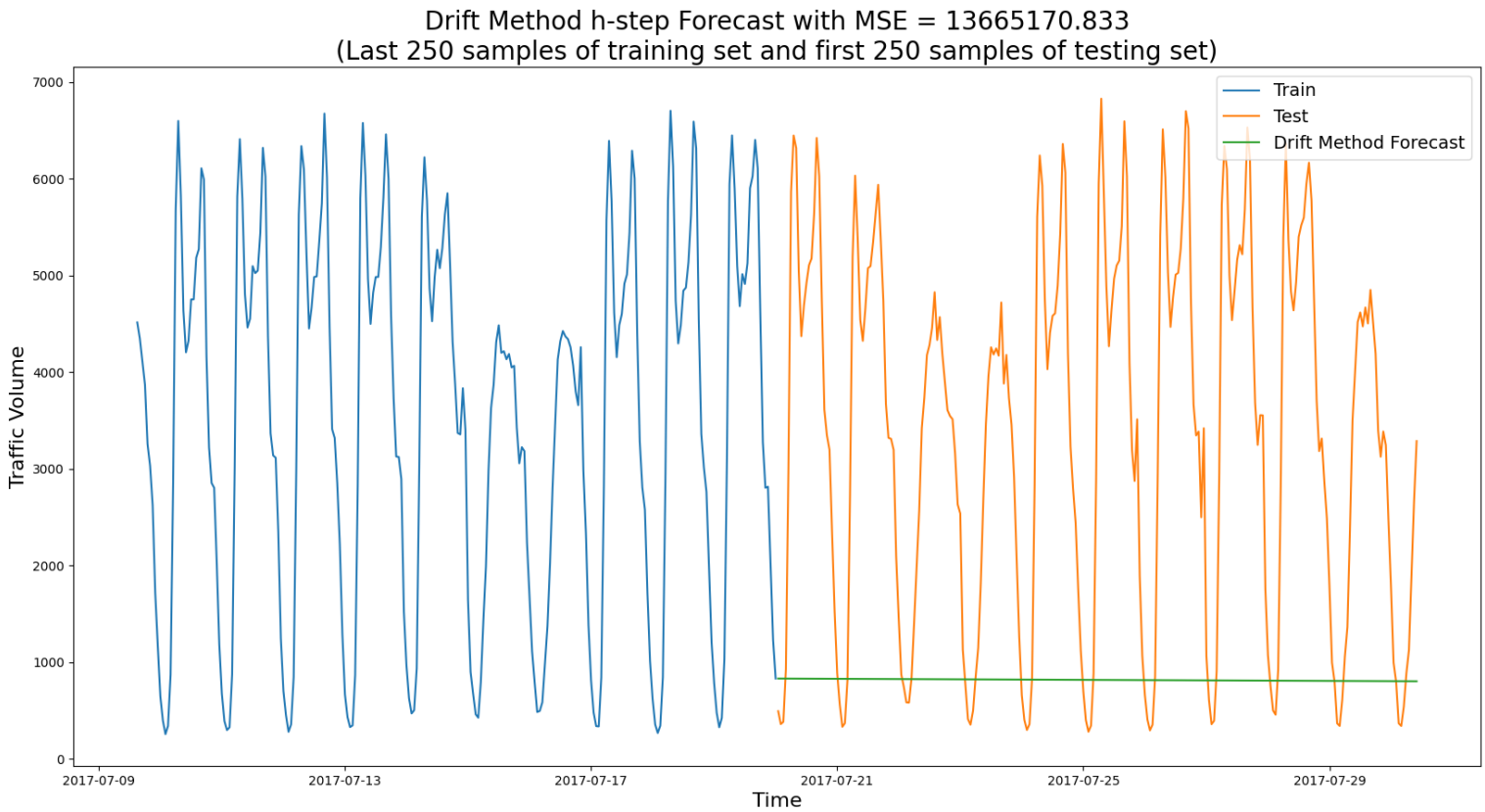



***Figure 50.*** The Drift Method h-step Forecast

```
The mean of the forecast error of Drift Method is 3102.071
The variance of the forecast error of Drift Method is 4042326.177
The Q value of the forecast error of Drift Method is 220901.753
```

***Figure 51.*** The Mean, Variance, and Q-value of the Forecast Error of the Drift Method

## f. Simple Exponential Smoothing

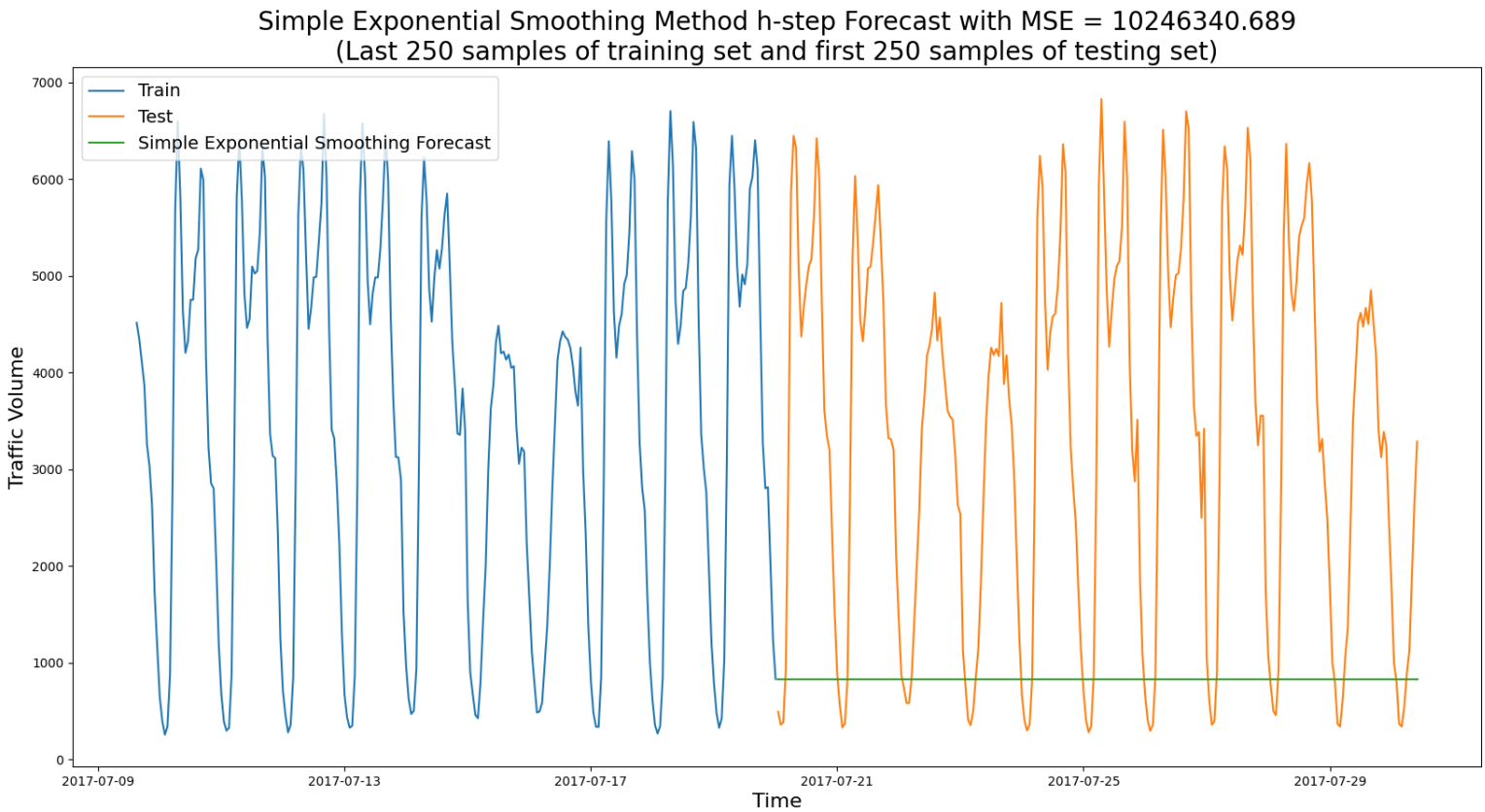



***Figure 52.*** The Simple Exponential Smoothing Method h-step Forecast

```
The mean of the forecast error of Simple Exponential Smoothing Method is 2510.82
The variance of the forecast error of Simple Exponential Smoothing Method is 3942124.722
The Q value of the forecast error of Simple Exponential Smoothing Method is 230379.232
```

***Figure 53.*** The Mean, Variance, and Q-value of the Forecast Error of the Simple Exponential Smoothing Method

## g. Holt's Linear Trend Method

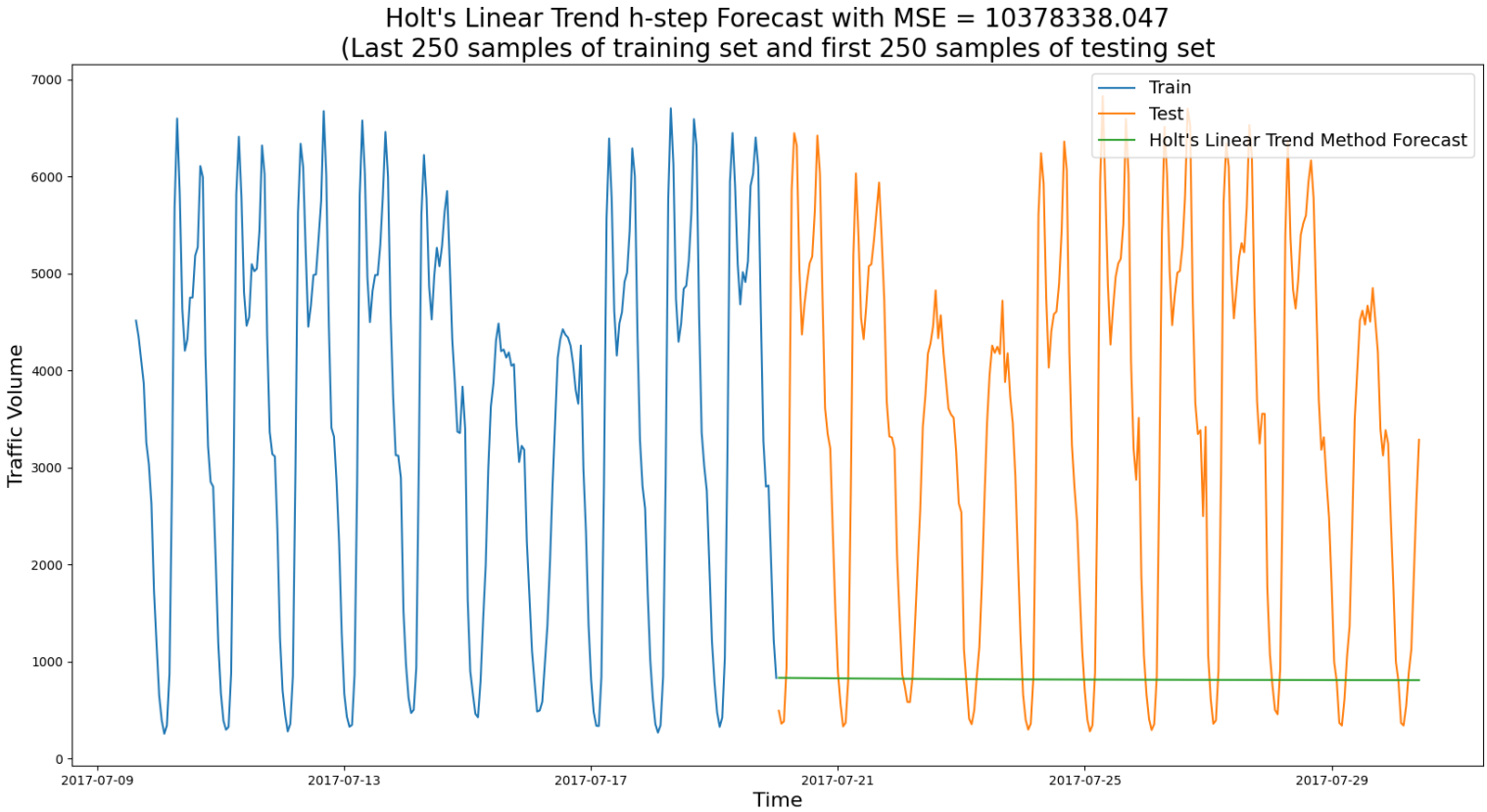



***Figure 54.*** The Holt's Linear Trend Smoothing Method h-step Forecast

```
The mean of the forecast error of Holt Linear Trend Method is 2536.976
The variance of the forecast error of Holt Linear Trend Method is 3942090.635
The Q value of the forecast error of Holt Linear Trend Method is 230383.63
```

***Figure 53.*** The Mean, Variance, and Q-value of the Forecast Error of the Holt's Linear Trend Method

# XII. Final Model Selection

| | Method | Q value | MSE | Mean of Prediction Error | Variance of Prediction Error |
|---|---|---|---|---|---|
| 0 | Average Method | 230379.23200 | 3945014.62100 | 53.75800 | 3942124.72200 |
| 1 | Naive Method | 230379.23200 | 10256388.58700 | 2512.82000 | 3942124.72200 |
| 2 | Seasonal Naive Method | 25900.39300 | 298706.55500 | -96.25400 | 289441.66400 |
| 3 | Drift Method | 220901.75300 | 13665170.83300 | 3102.07100 | 4042326.17700 |
| 4 | Simple Exponential Method | 230379.23200 | 10246340.68900 | 2510.82000 | 3942124.72200 |
| 5 | Holt's Linear Trend Method | 230383.63000 | 10378338.04700 | 2536.97600 | 3942090.63500 |
| 6 | Holt_Winter Method | 42912.67400 | 323093.56600 | -173.64200 | 292941.91600 |
| 7 | Multiple Linear Regression | 162366.19300 | 3772490.66300 | 72.23800 | 3767272.39600 |
| 8 | ARIMA(2,1,2)xARIMA(0,1,0)_s168 | 200516.51300 | 11923926.65300 | 2756.97900 | 4322991.88800 |

***Figure 53.*** The Comparison of 9 Forecast Method

From *Figure 53*, the Seasonal Naive Method has the minimum Q-value and MSE of the forecast error. It can be seen from *Figure 34* that the forecast can simulate the pattern of the test set very well. Therefore, the final Best Model is Seasonal Naive Method.

## a. Forecast Function and h-step ahead Prediction

In this case, I used the h-step ahead prediction function as the forecast function. For the Seasonal Naïve Method, the forecast for time $T + h$ is written as

$$\hat{y}_{T+h|T} = y_{T+h-m(k+1)}$$

where $m =$ seasonal period, and $k$ is the integer part of $\frac{h-1}{m}$

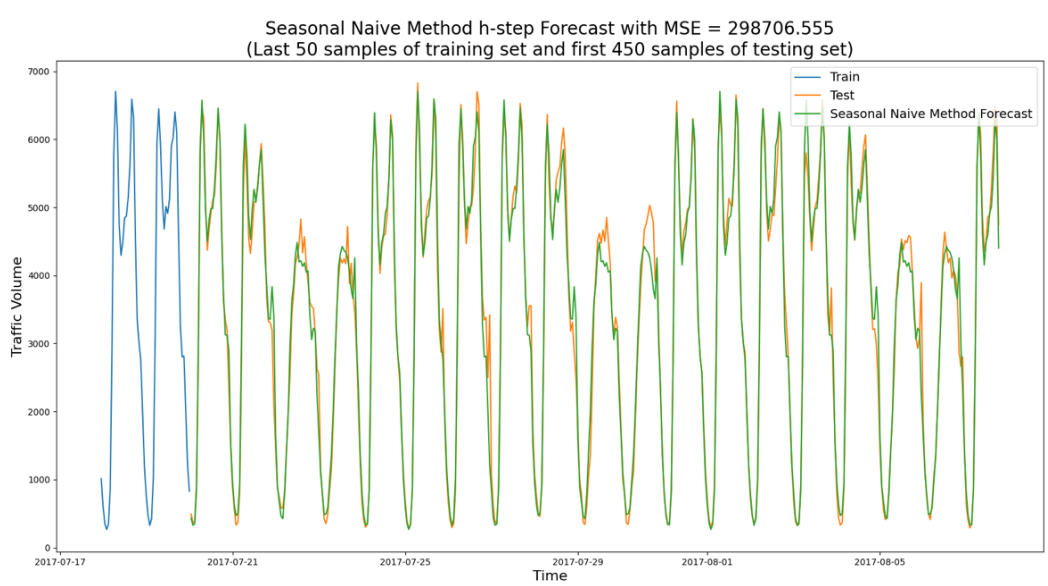



***Figure 54.*** Seasonal Naïve Method h-step Forecast

# X. Summary and Conclusion

The Seasonal Naive Method is as helpful for highly seasonal data as my traffic volume data. From *Figure 54*, my final model generated by the naive seasonal method can predict future values correctly. I think the limitation of this model is its accuracy based on accurate estimation of the seasonal period. For example, the seasonal period of my dataset can be daily (s =24) or weekly (s =168). Still, if the 24 is used as the seasonal period for prediction, the effect is not as good as if the seasonal period is 168. Another limitation is that this method cannot account for week-to-week changes in level.

# XIII. References


1. A Functional Data Analysis Approach to Traffic Volume Forecasting
   https://ieeexplore.ieee.org/document/7947181

2. Metro Interstate Traffic Volume Dataset
   https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume#