



Hourly Traffic Volume Forecasting

DATS 6450: Time Series Analysis and Modeling
Term Project
Rayna Liu

TABLE OF CONTENTS

01

Introduction

02

Description of the Dataset

03

Stationary

04

Time Series Decomposition

05

Holt-Winter Method

06

Multiple Linear Regression

07

ARMA, ARIMA, SARIMA Model

08

Based Models

09

Final Model Selection

10

Summary & Conclusion





01

Introduction

Traffic volume forecasts are used by many transportation analysis and management systems to better characterize and react to fluctuating traffic patterns. ¹

The purpose of this term project is to find the best model to forecast the hourly traffic volume by developing, analyzing and comparing couples of model.

Description of Dataset

- ❑ Data Preprocessing
- ❑ Dependent Variable v.s. Time
- ❑ ACF / PACF of Dependent Variable
- ❑ Correlation Matrix

02

An abstract graphic design featuring organic, flowing shapes in orange, olive green, and dark grey. A large orange shape on the left contains a white circle with the number '02'. To its right, a green shape contains a dark grey circle with a teal center. Several small teal and dark grey dots are scattered around the main shapes. A white teardrop shape points towards the bottom right.



About 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 [\[Here\]](#).

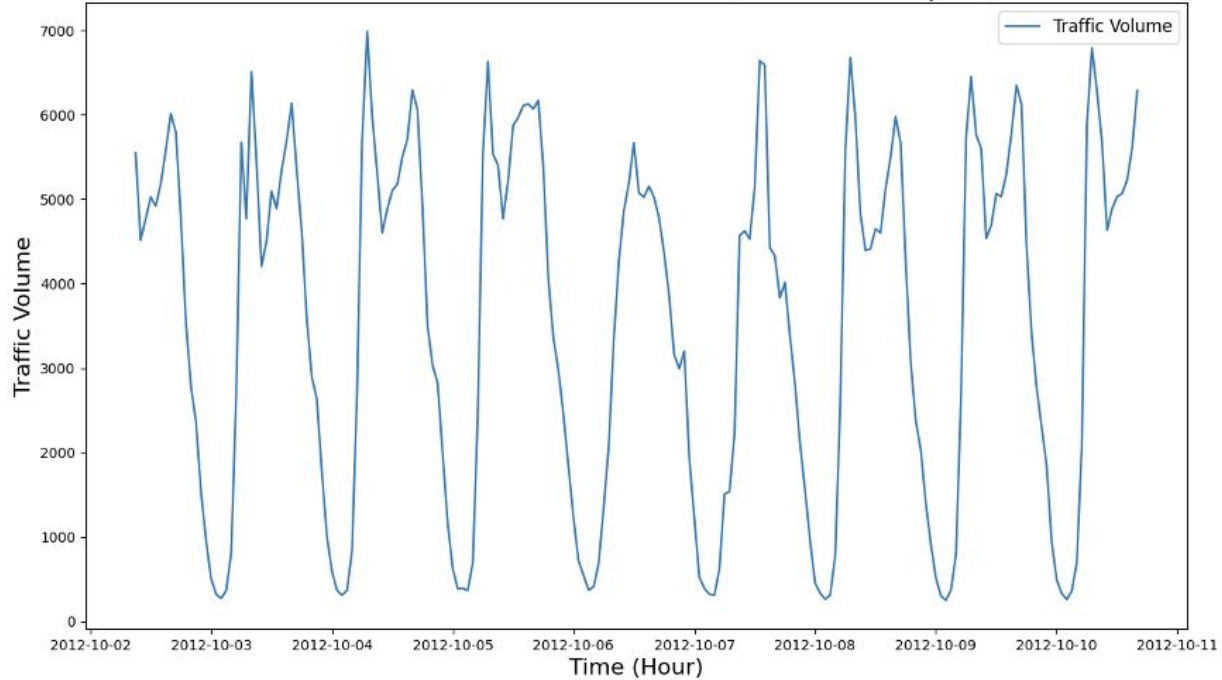
- ❑ Missing data -- Replace with the mean hourly traffic volume
- ❑ 40,575 instances, 9 attributes
- ❑ DateTime range from 2012-10-02 09:00:00 to 2018-09-30 23:00:00

Attribute Information:

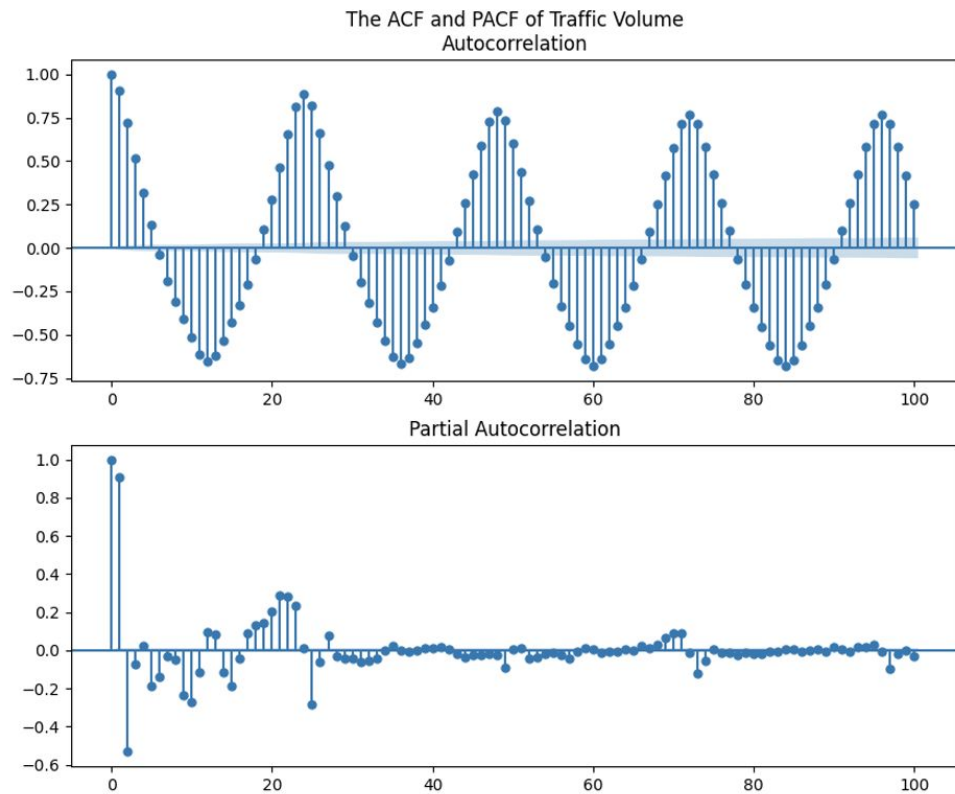
[holiday]	(Categorical) US National holidays plus regional holiday
[temp]	(Numeric) Average temperature in Kelvin
[rain_1h]	(Numeric) Amount in mm of rain that occurred in the hour
[snow_1h]	(Numeric) Amount in mm of snow that occurred in the hour
[clouds_all]	(Numeric) Percentage of cloud cover
[weather_main]	(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

Traffic Volume over Time

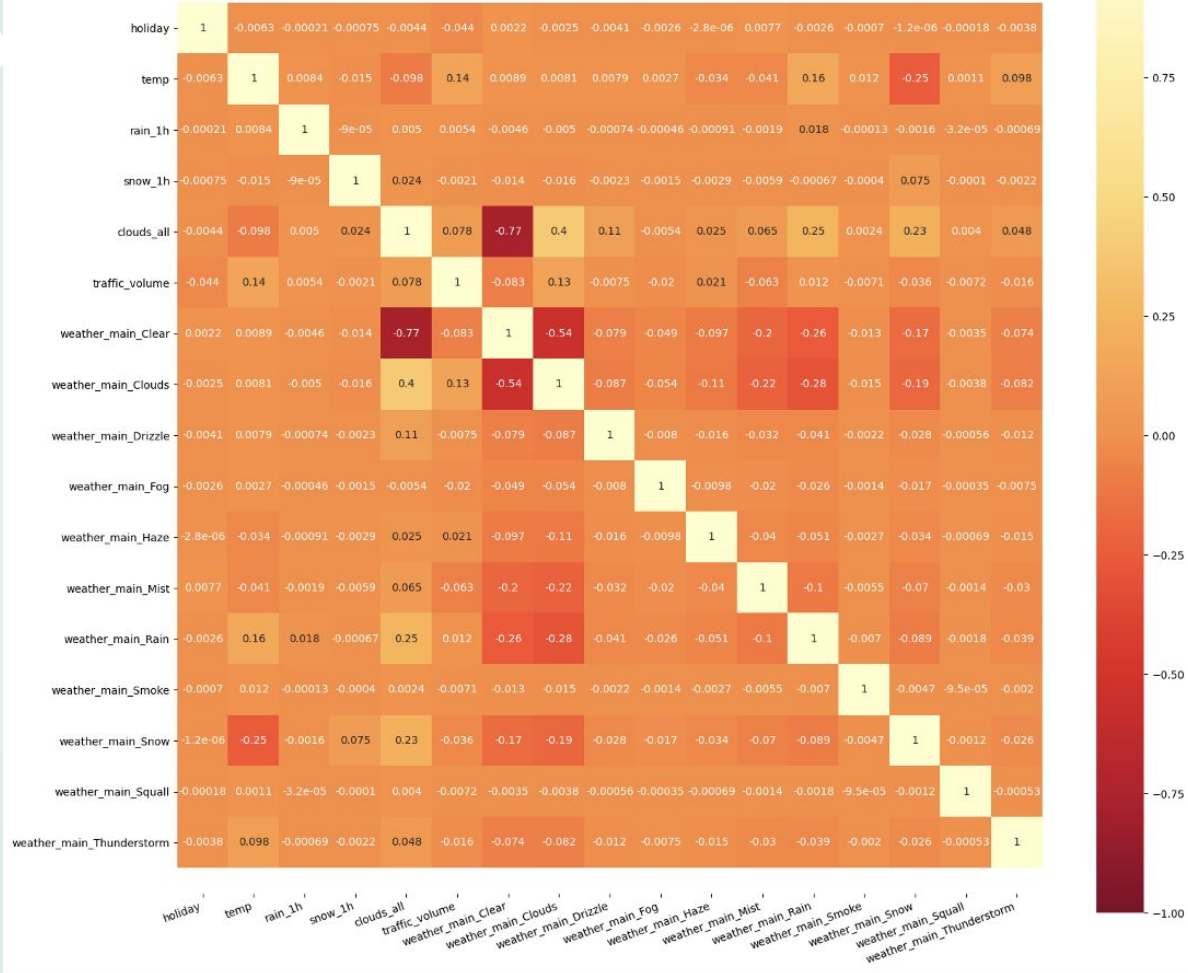
The Traffic Volume versus Time (First 200 samples)



ACF / PACF of Traffic Volume



Correlation Matrix of Traffic Volume Dataset

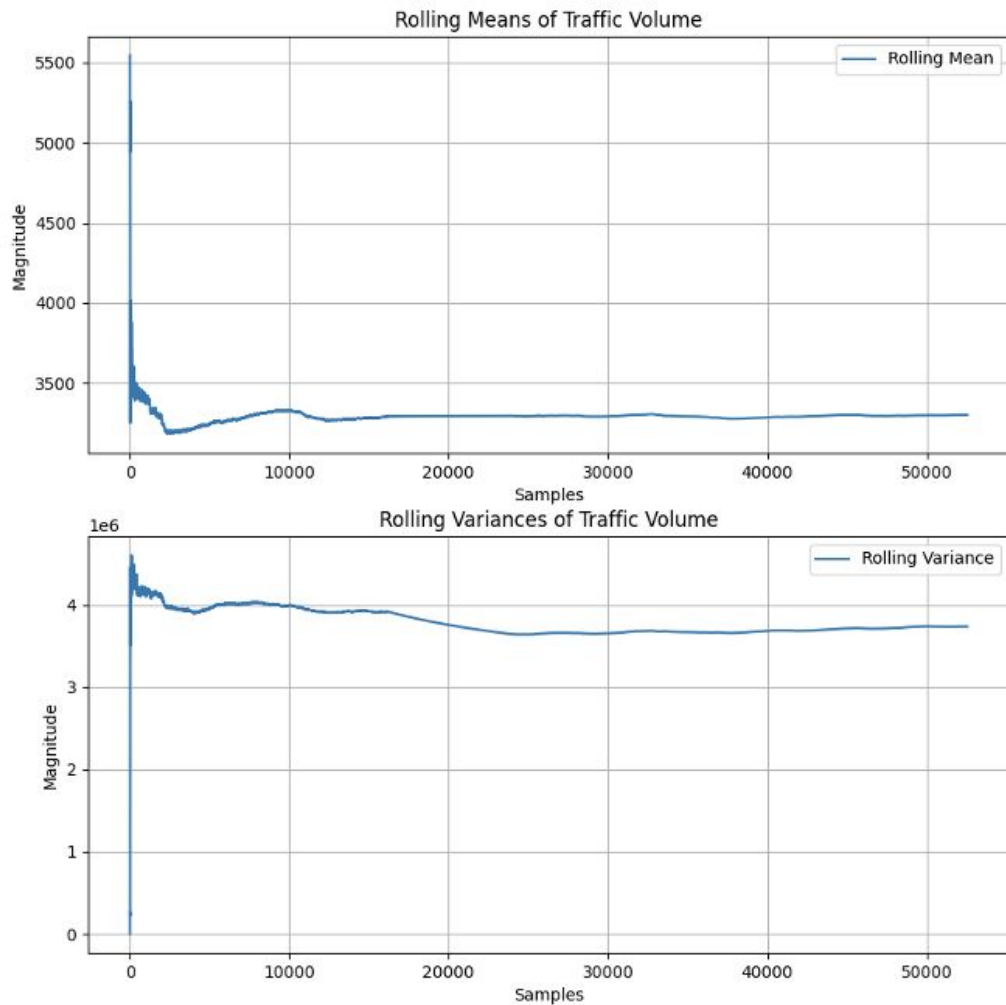


Stationary

- ❑ ADF Test
- ❑ Plot of Rolling Mean and Variance
- ❑ Seasonal / Non-Seasonal Differencing

03

An abstract graphic design featuring organic, flowing shapes in orange, teal, and white. A central teal circle with a black border contains the white number '03'. Other elements include a smaller teal circle with a black border, a white circle with a black border, and various smaller teal and white circles scattered around. The background is a light blue-grey color.



The ADF test of Traffic Volume:

ADF Statistic: -33.506261

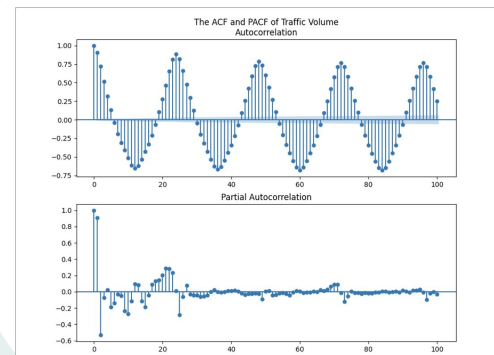
p-value: 0.000000

Critical Values:

1%: -3.430

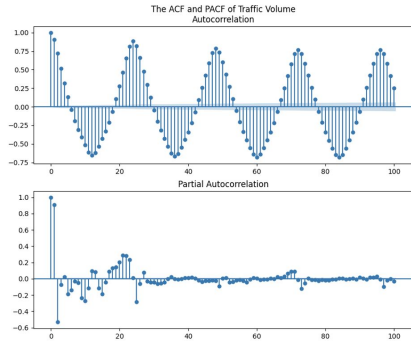
5%: -2.862

10%: -2.567



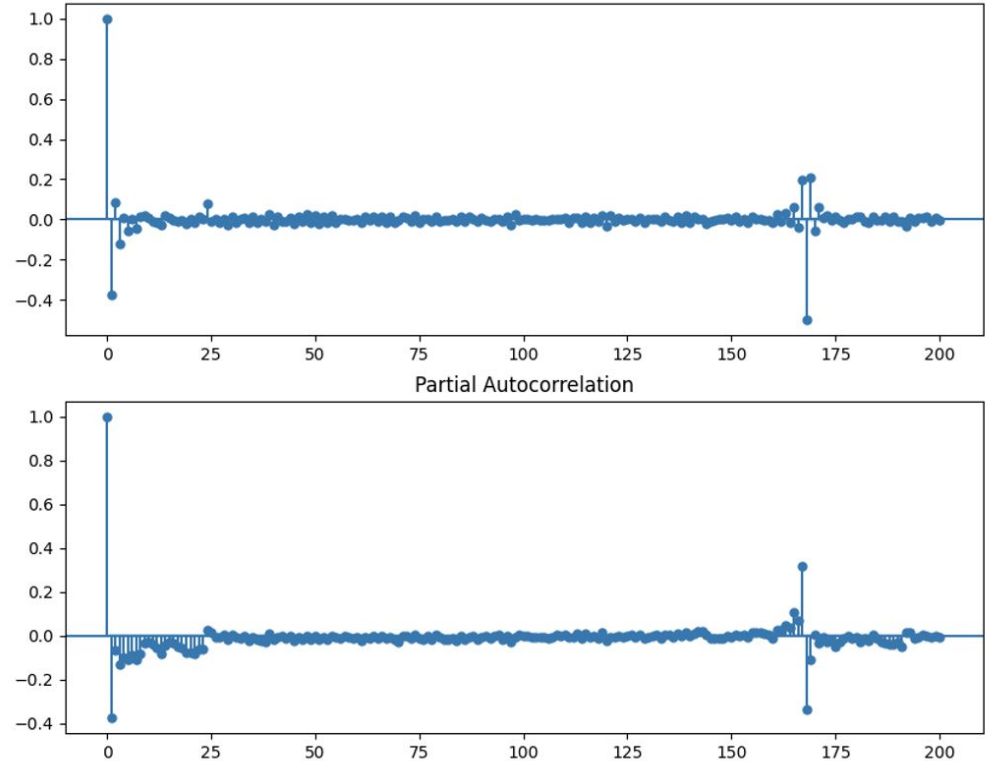
Apply Seasonal & Non-Seasonal Differencing

Before



After

The ACF and PACF of Seasonal and Non-Seasonal Differenced of Traffic Volume Autocorrelation



An abstract graphic on the left side of the slide. It features several organic, teardrop-like shapes in teal, orange, and dark grey. A central orange circle contains the white number '04'. Other shapes include a large orange circle with a dark grey outline, a white circle with a green outline, and a small green circle. The background is a light blue-grey.

04

Time Series Decomposition

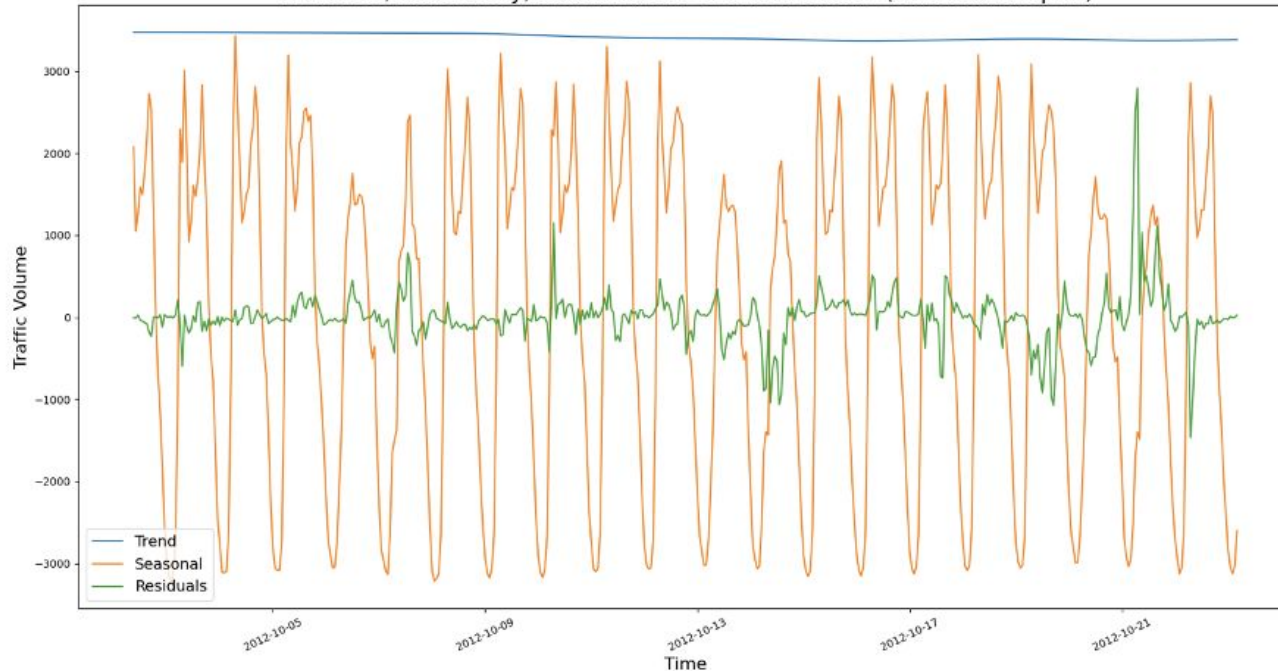
- ❑ STL Decomposition Method
- ❑ Strength of the Trend and Seasonality
- ❑ Plot of Raw Dataset v.s De-trended and Seasonally Adjusted Dataset

STL Decomposition Method

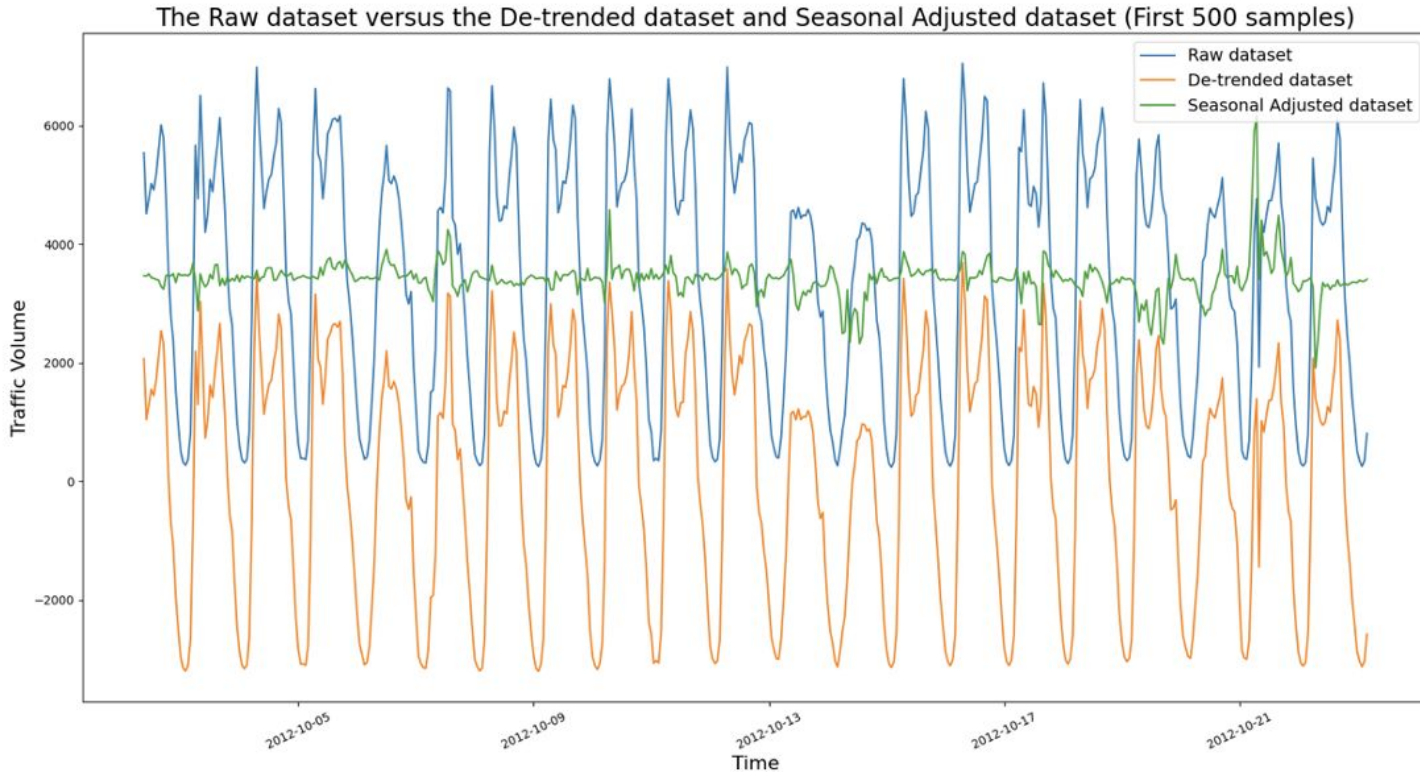
The strength of trend for the dataset is: 0.9999999022936712

The strength of seasonality for the dataset is: 0.9999995640033321

The Trend, Seasonality, and Reminder of Traffic Volume (First 500 samples)



Plot of Raw Dataset v.s. the De-trended and Seasonally Adjusted Dataset





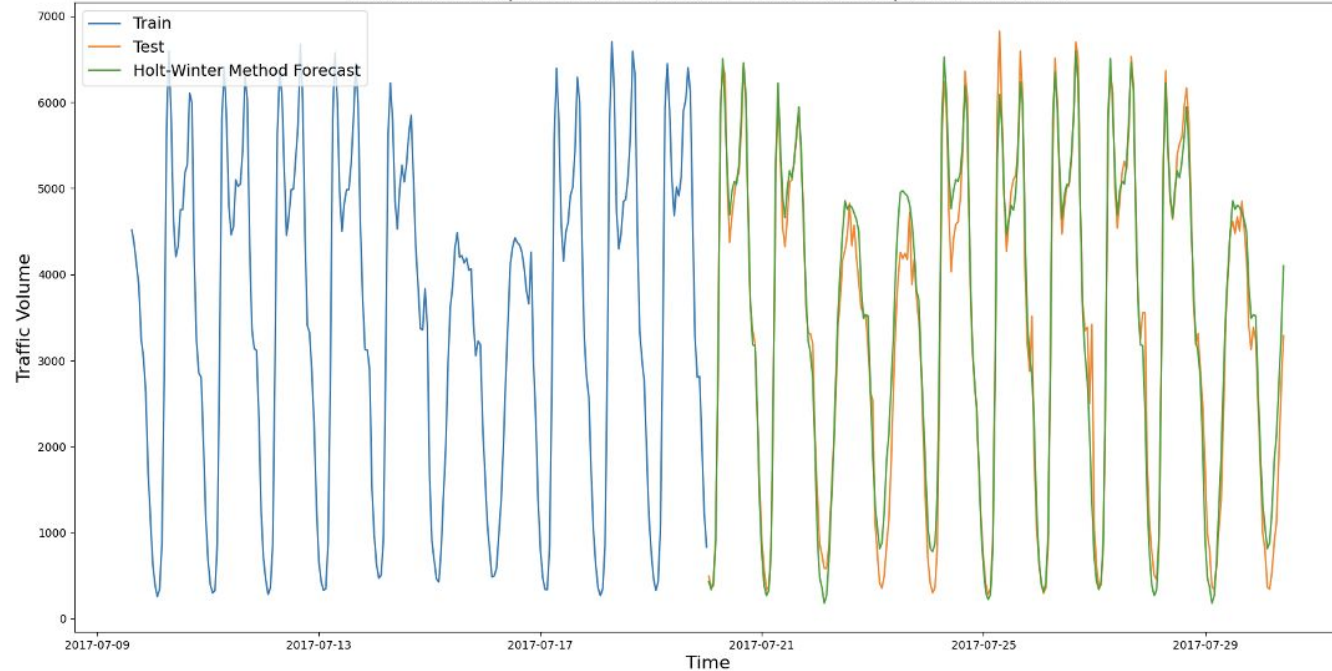
05

Holt-Winter Method

- ❑ Forecast v.s. Test set
- ❑ Forecast Analysis

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

The Holt-Winter Season Method Forecast with $MSE=323093.566$
(Last 250 samples of train set and first 250 samples of test set)



Multiple Linear Regression

- ❑ Collinearity Detection
- ❑ Feature Reduction
- ❑ Hypothesis Tests Analysis
- ❑ AIC,BIC,R-squared and Adjusted R-squared
- ❑ One-step ahead Forecast
- ❑ Residual Analysis

06

An abstract graphic design featuring organic, flowing shapes in orange, green, and teal. A central green circle contains the number '06' in white. Other shapes include a white circle with a green center, a large orange shape with a teal center, and various smaller circles and lines in the background.



Singular Values Analysis & Condition Number



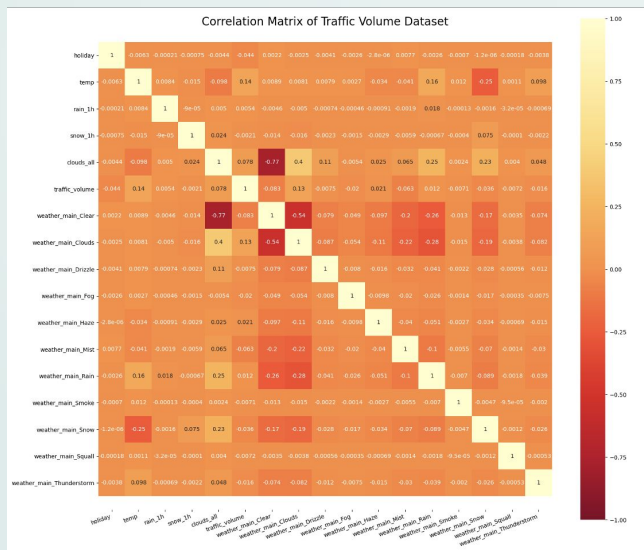
The Singular Values of the raw dataset is

```
[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]
```

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

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:             Wed, 05 May 2021   Prob (F-statistic): 6.61e-275
Time:             19:00:28          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.

Final Multiple Linear Regression Model

```
# ===== Final Model after Feature Selection =====
```

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:                   19:16:51          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] R^2 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.

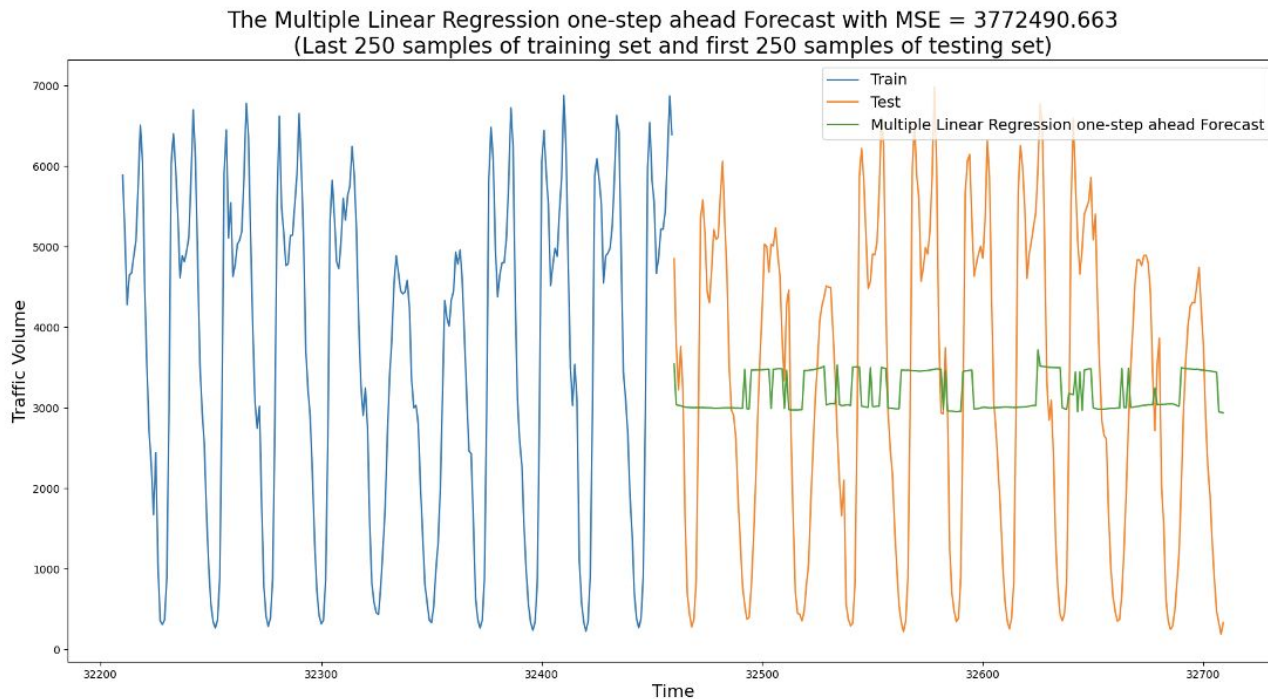
```
In[6]:
```

1-step ahead Forecast of MLR Model

The mean of the forecast error of MLR model is 72.238

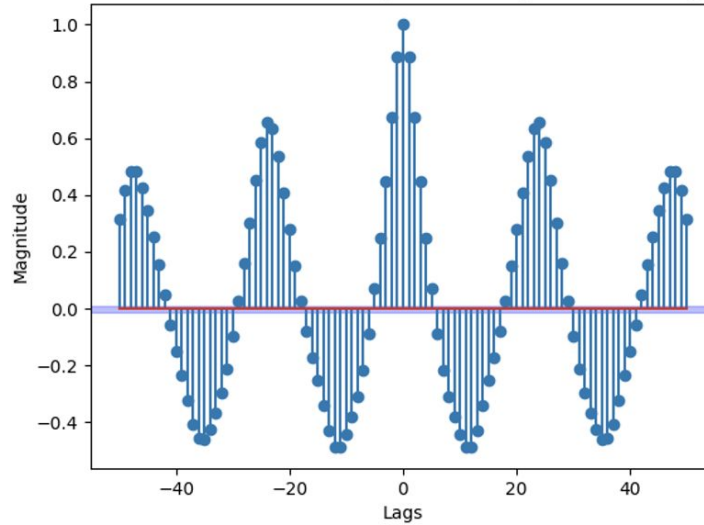
The variance of the forecast error of MLR model is 3767272.396

The Q value of the forecast error of MLR model is 162366.193



ACF of Residuals

The ACF of the Residuals of Multiple Linear Regression with lags = 50



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



07

ARMA, ARIMA, SARIMA

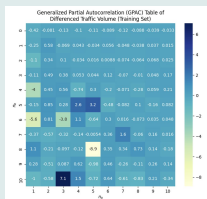
- ❑ Order Determination
- ❑ Estimated Parameters of ARMA Model
- ❑ Diagnostic Analysis

Order Determination

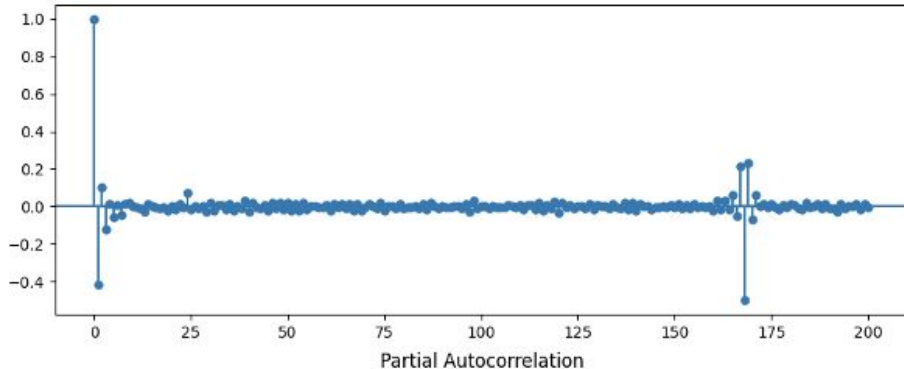
Estimate:

ARIMA(0,1,3) x ARIMA(0,1,1)₁₆₈

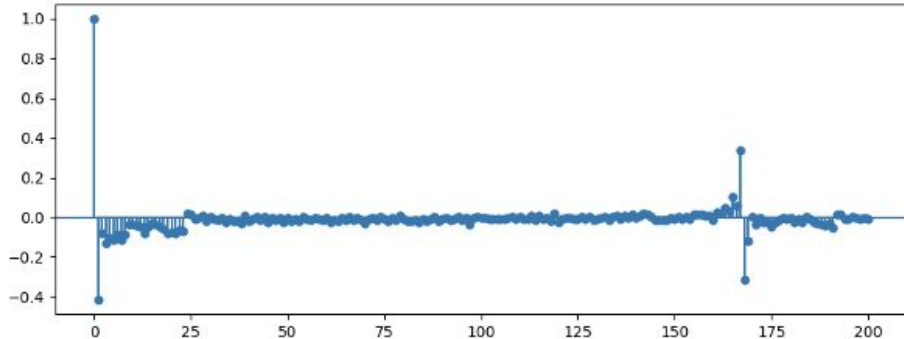
ARMA(0,171)



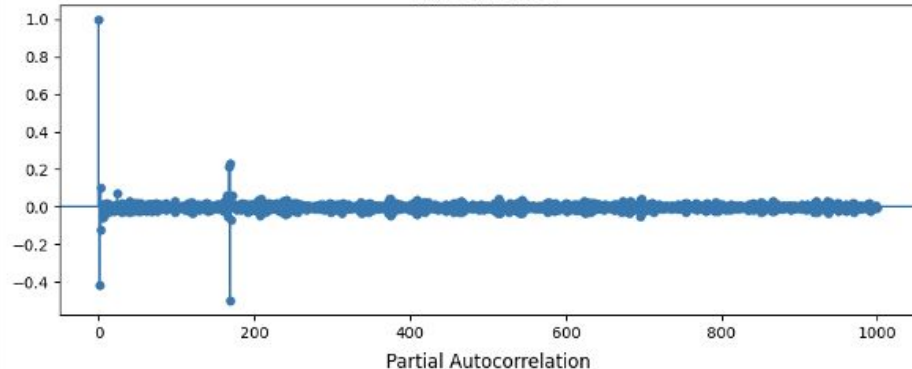
The ACF and PACF of Differenced Traffic Volume (Training Set)
Autocorrelation



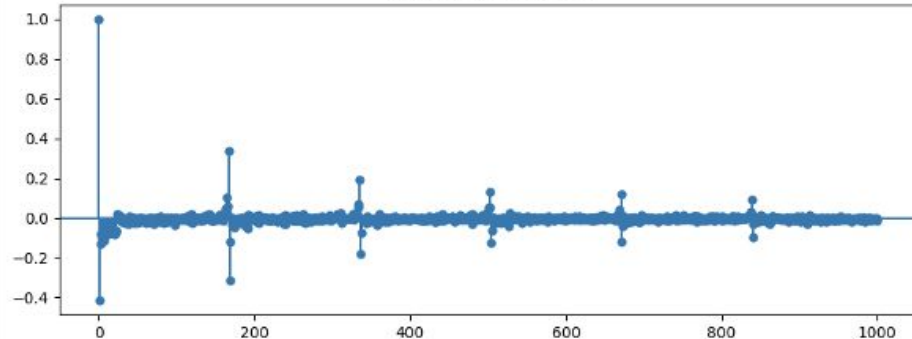
Partial Autocorrelation



The ACF and PACF of Differenced of Traffic Volume (Training Set)
Autocorrelation



Partial Autocorrelation



Estimated Parameters of ARMA Model

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
=====
```

The roots of numerator is

[0.97323702 -0.22183702]

The roots of denominator is

[-0.0967+0.57969743j -0.0967-0.57969743j]

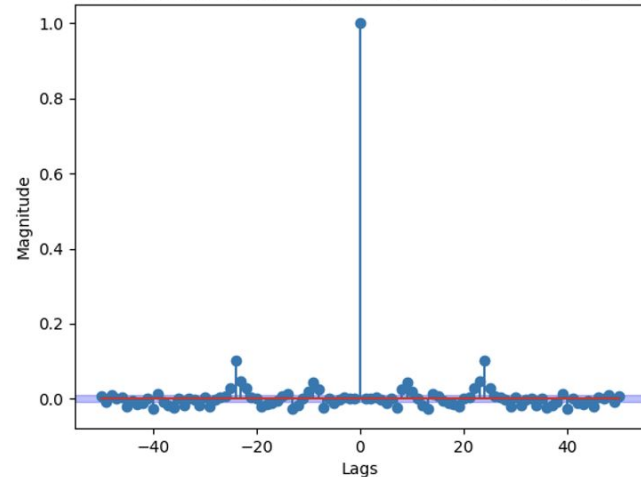
The residual is NOT white

The Q value is [1258.32575539]

The chi critical is 131.141216667052

The p-value of chi square test is [1.40757232e-199]

The ACF of the residuals of ARMA(2, 2) model with lags = 50



The estimated variance of error is 457221.63893733063

The estimated covariance of the a1 is 0.00052802731634219

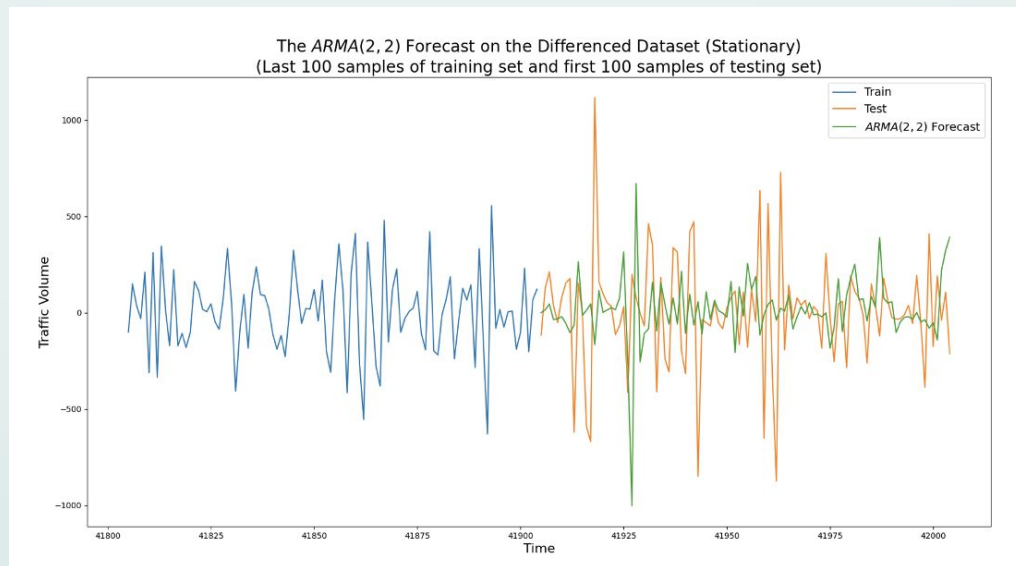
The estimated covariance of the a2 is 4.3615721515049795e-05

The estimated covariance of the a3 is 0.00039149247495931077

The estimated covariance of the b1 is 0.0003750385575653475

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

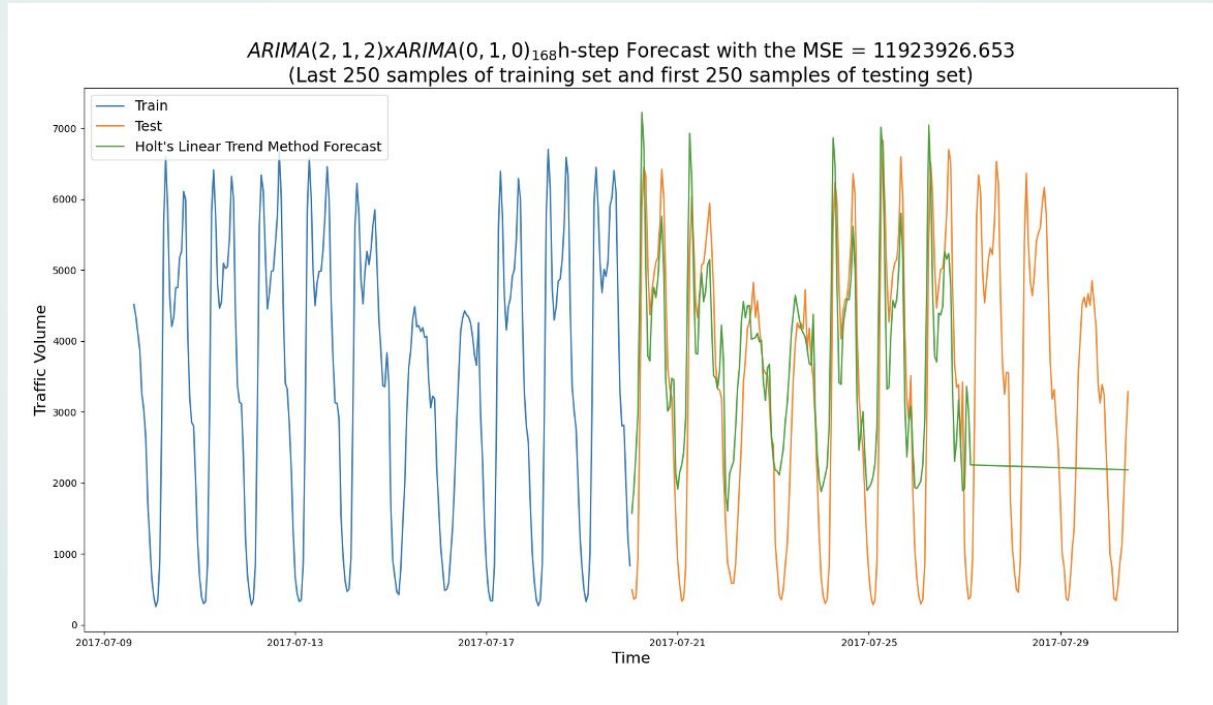
The variance of residual errors versus the variance of forecast errors is 0.935863574033126



ARMA(2,2) Forecast

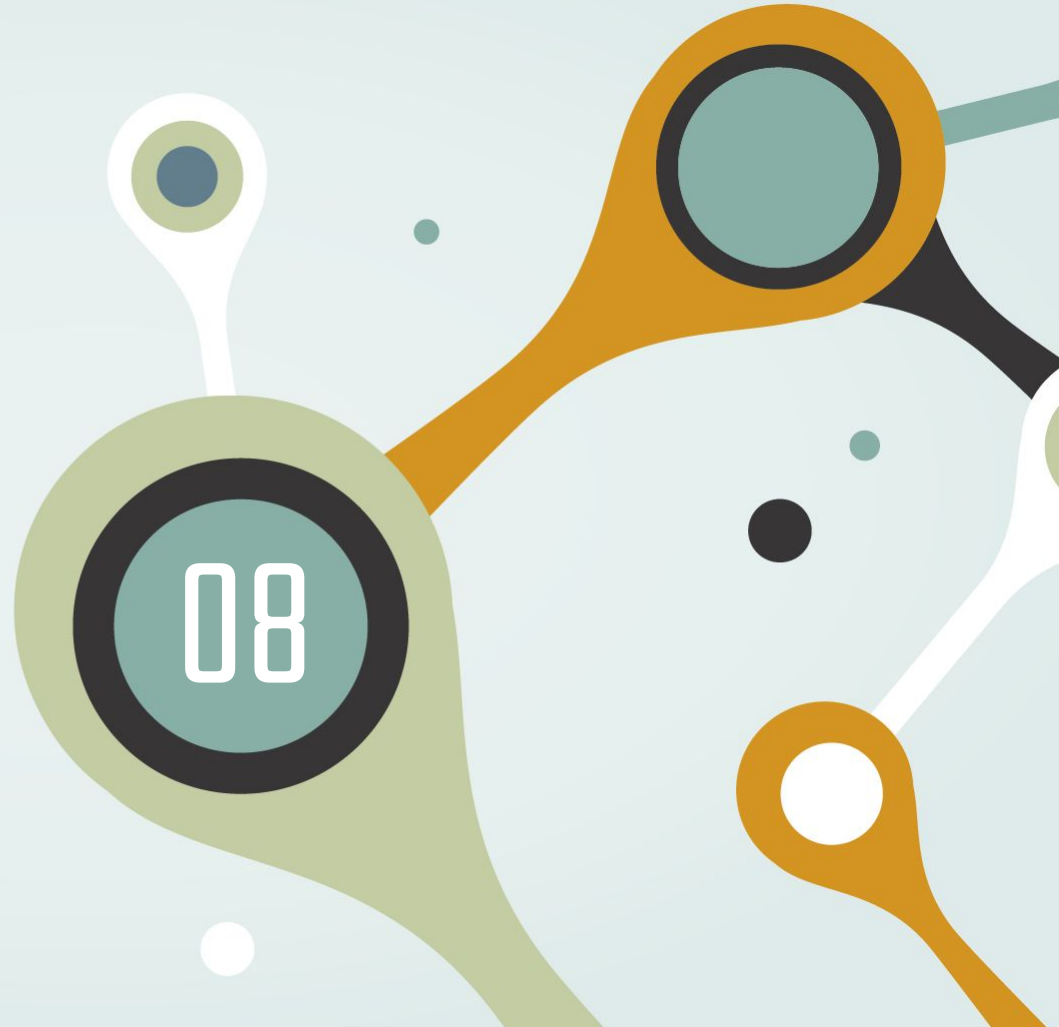
ARMA(2,2) model could only forecast the differenced dataset (stationary). Convert the ARMA(2,2) process to SARIMA model, which could use for forecast the traffic volume of the raw dataset (non-stationary).

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



Based Models

- ❑ Average Method
- ❑ Naive Method
- ❑ Seasonal Naive Method
- ❑ Drift Method
- ❑ Simple Exponential Smoothing
- ❑ Holt's Linear Trend Method



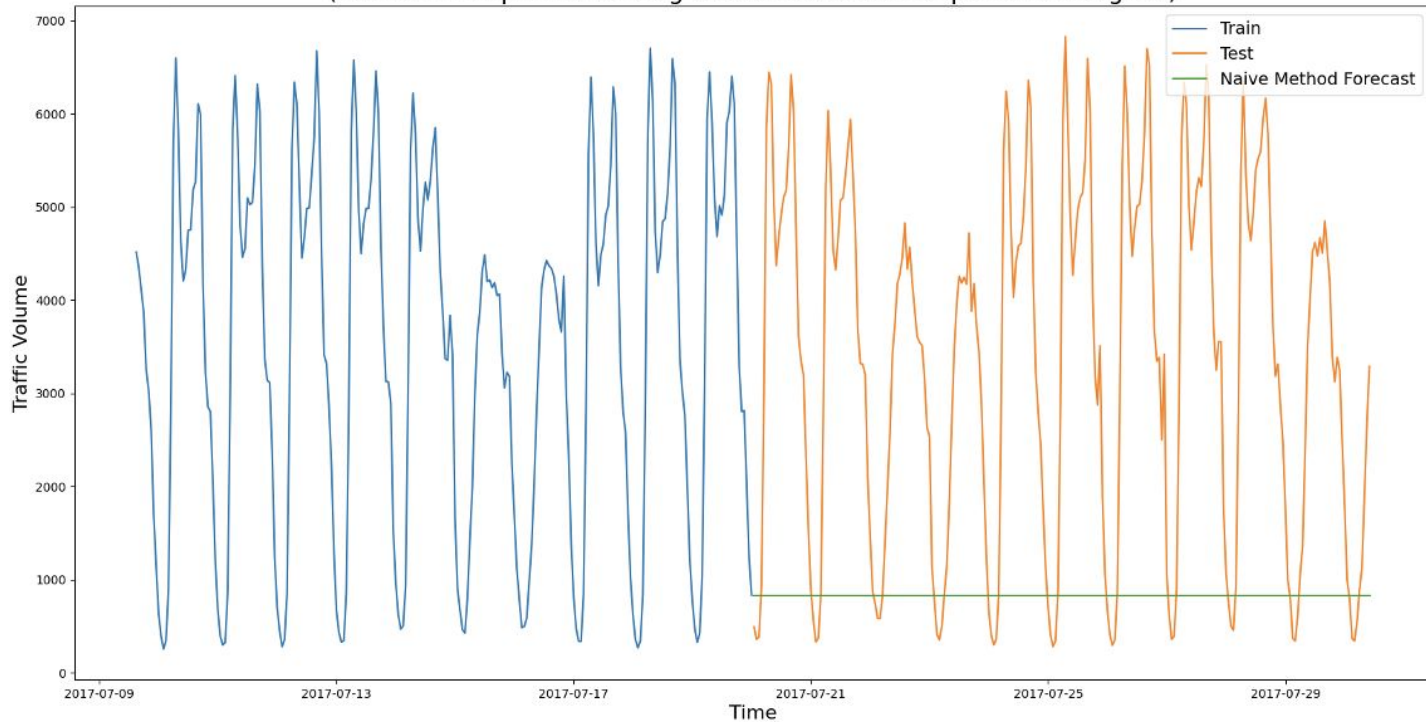
Average Method

Average Method h-step Forecast with MSE = 3945014.621
(Last 250 samples of training set and first 250 samples of testing set)



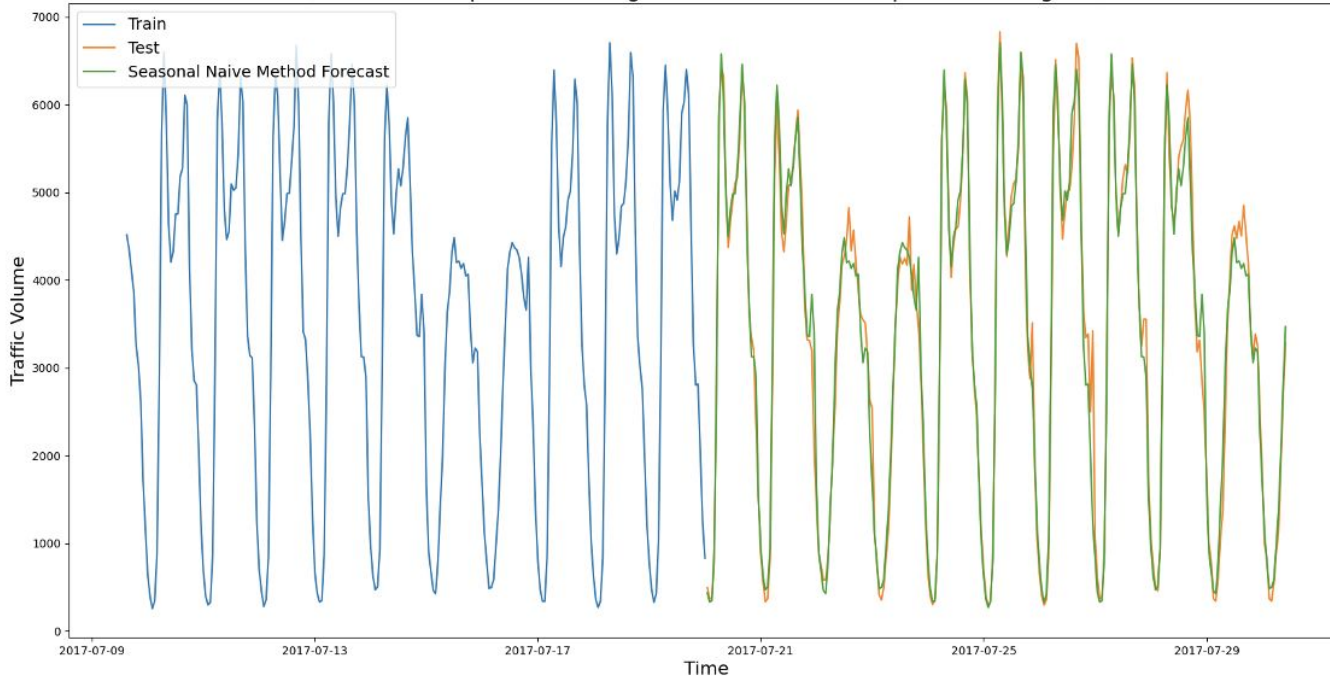
Naive Method

Naive Method h-step Forecast with MSE = 10256388.587
(Last 250 samples of training set and first 250 samples of testing set)



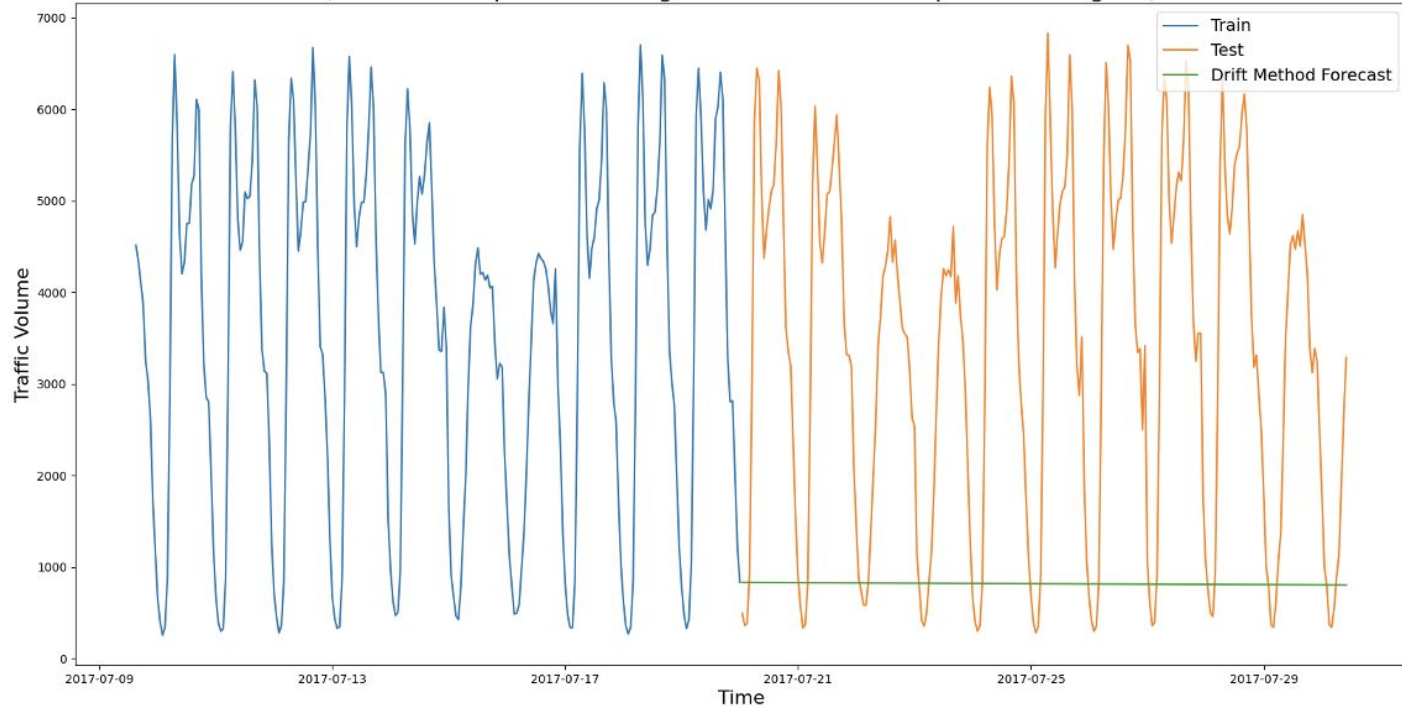
Seasonal Naive Method

Seasonal Naive Method h-step Forecast with MSE = 298706.555
(Last 250 samples of training set and first 250 samples of testing set)

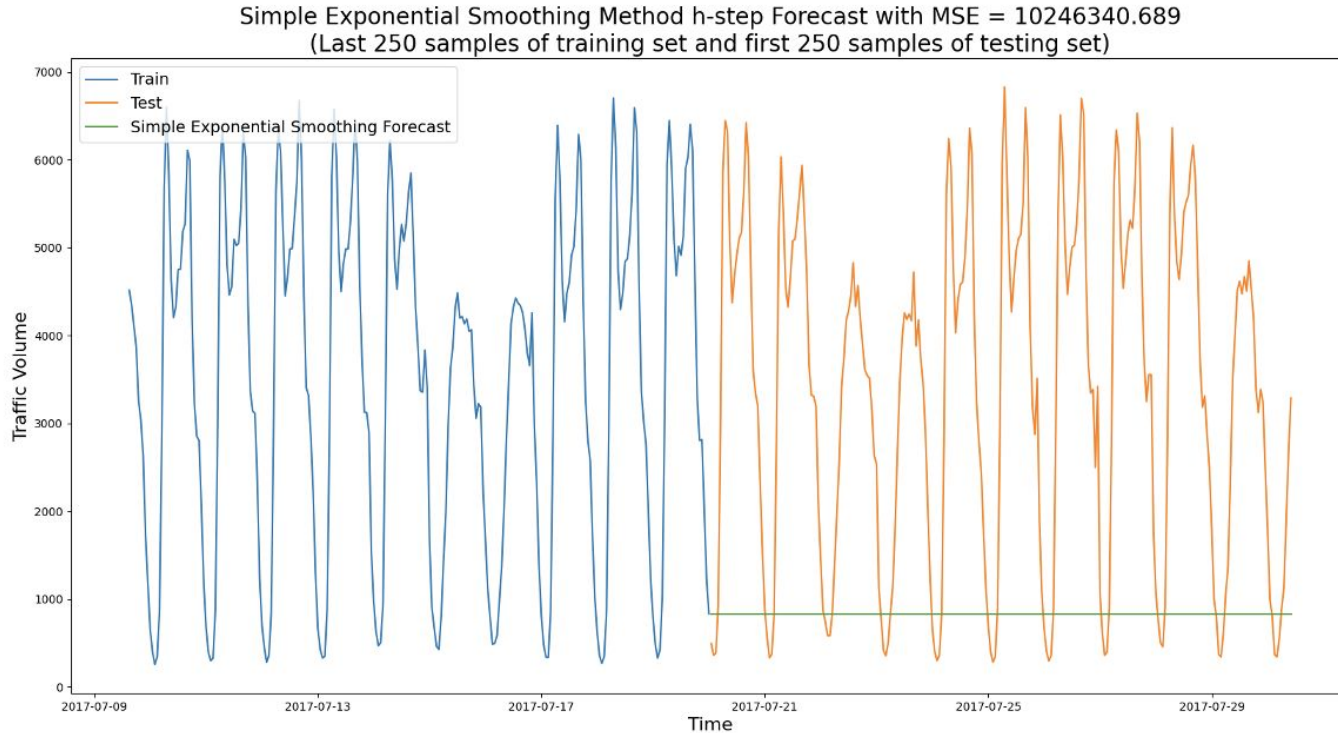


Drift Method

Drift Method h-step Forecast with $MSE = 13665170.833$
(Last 250 samples of training set and first 250 samples of testing set)

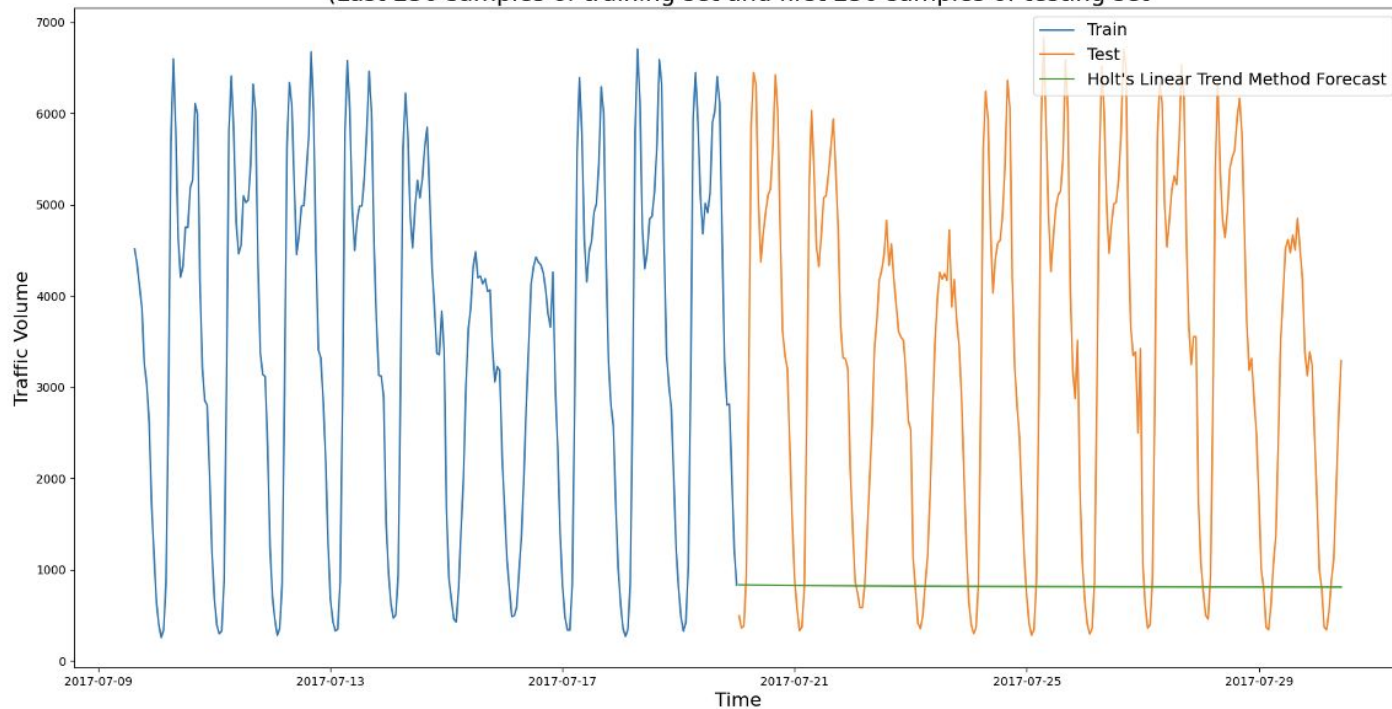


Simple Exponential Smoothing



Holt's Linear trend Method

Holt's Linear Trend h-step Forecast with $MSE = 10378338.047$
(Last 250 samples of training set and first 250 samples of testing set)





09

Final Model Selection

- ❑ Forecast Function and h-step Prediction

The Comparison of 9 Forecast Method

	↕ 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



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 $\frac{h-1}{m}$

An abstract graphic on the left side of the slide. It features a central orange circle with the number '10' in white. This circle is connected by a teal line to a larger teal shape above it. To the left, a dark grey shape contains an orange circle, with a teal line connecting it to the central '10' circle. Various other organic shapes in white, teal, and orange are scattered around the central elements.

10

Summary & Conclusion

The naive seasonal method is as helpful for highly seasonal data as my traffic volume data. 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.

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#>

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution

