Assignment_3

Olfa Jerbi 2021713094

Dataset visualization

```
Let's first visualize the dataset in hand
```

Data preview

holiday	Temp	rain_1h	snow_1h	clouds_all	weather_main	weather_description	date_time	traffic_volume
None	288.28	0	0	40	Clouds	scattered clouds	2012-10-02 09:00:00	5545
None	289.36	0	0	75	Clouds	broken clouds	2012-10-02 10:00:00	4516
None	289.58	0	0	90	Clouds	overcast clouds	2012-10-02 11:00:00	4767
None	290.13	0	0	90	Clouds	overcast clouds	2012-10-02 12:00:00	5026
None	291.14	0	0	75	Clouds	broken clouds	2012-10-02 13:00:00	4918

```
As seen above, some of variables of the time series are text, so let's encode them to numbers before the analysis
```

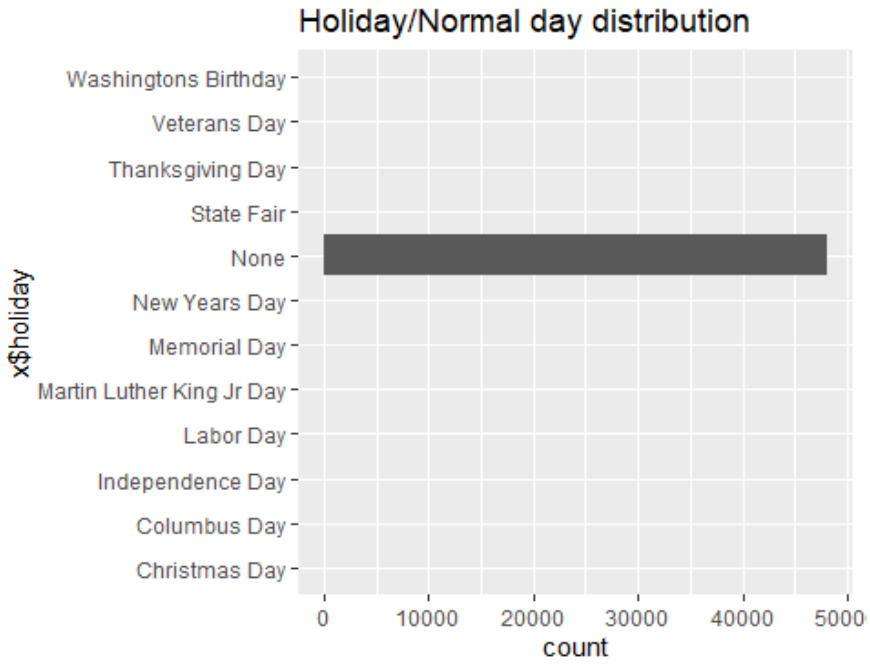
Encoded Data preview

holiday	temp	rain_1h	snow_1h	clouds_all	weather_main	weather_description	date_time	traffic_volume

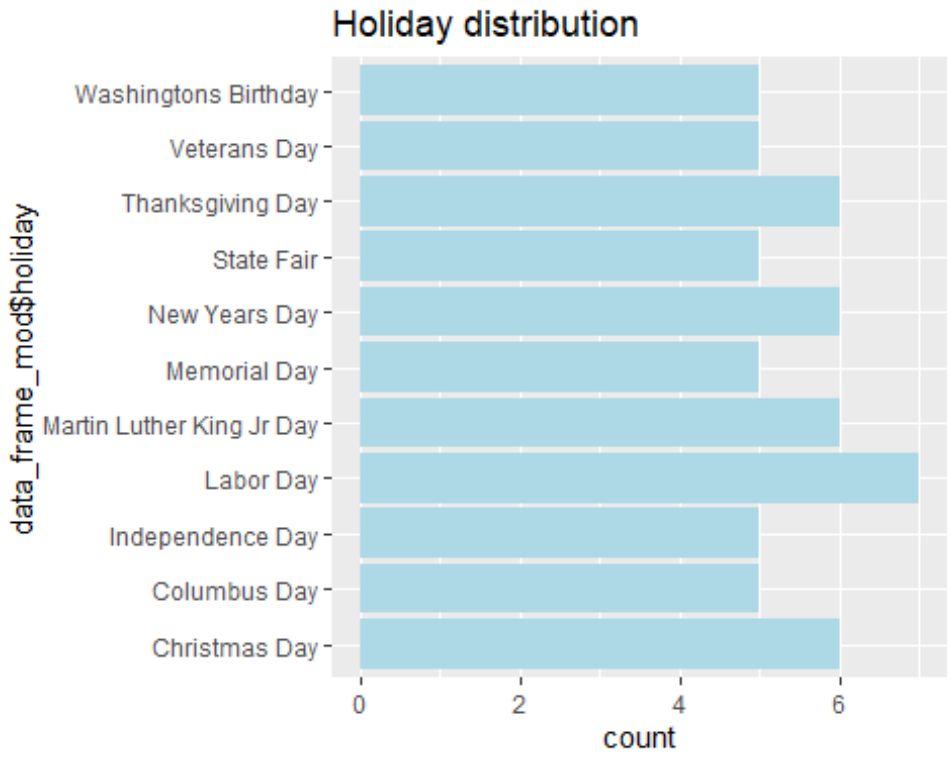
1	288.28	0	0	40	2	23	10/2/12 9:00	5545
1	289.36	0	0	75	2	1	10/2/12 10:00	4516
1	289.58	0	0	90	2	18	10/2/12 11:00	4767
1	290.13	0	0	90	2	18	10/2/12 12:00	5026
1	291.14	0	0	75	2	1	10/2/12 13:00	4918

Descriptive statistics

Let's start with observing the first variable "holiday" that describes whether the day is a holiday or not:

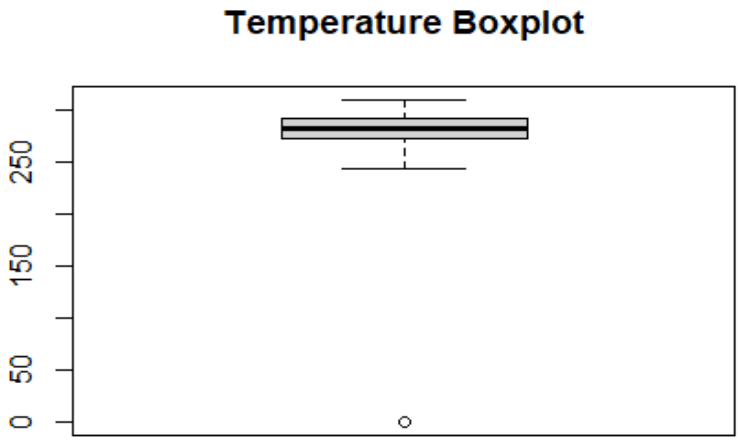


As normal days are much more frequent over the years of the data set, holidays are not as visible in the plot. To analyze them let's remove the normal days and plot the distribution again:



We can observe that Labor day holiday and the holidays related to Christmas, as well as the Martin Luther King Jr holiday are slightly longer than other holidays.

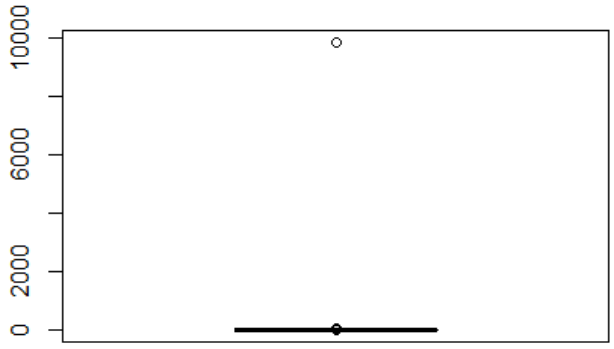
Let's now observe the variable temperature:



As seen in the box plot observed above, the temperature is between ~225 and 320 Kelvin, with the presence of an outlier around 0. Since the temperatures are in kelvin it is unlikely

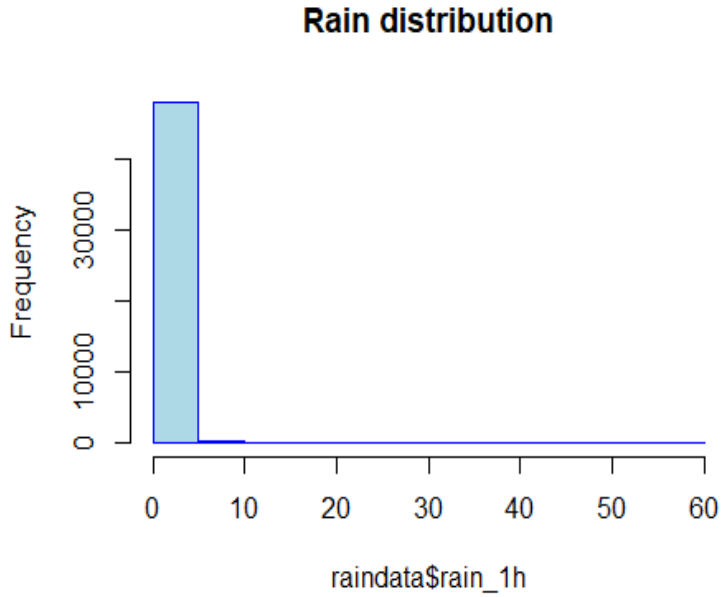
to find a value of 0 Kelvin (-273.15°C), this could be due to some error in capturing the temperature.

For rain variable:



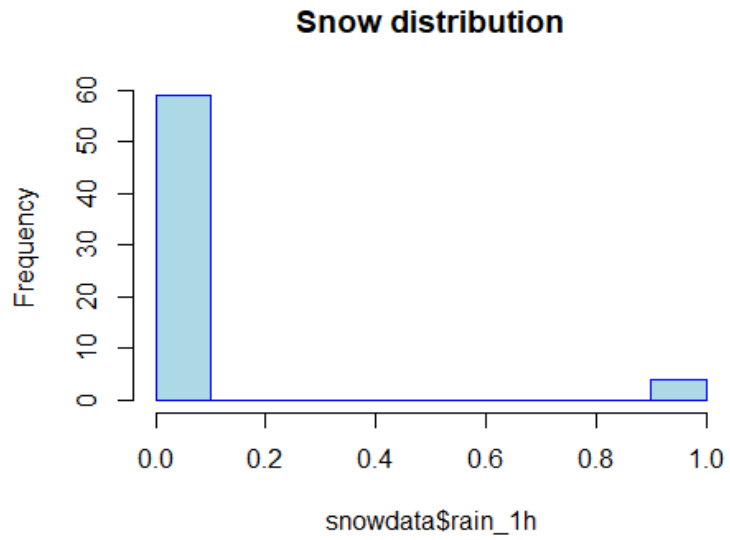
We can see that the values are condensed around the value of 50mm with an outlier of almost 10000mm. To be able to observe the regular values (the ones except the unusual case of very high precipitations), we remove that value and trace the histogram of the remaining values.

PS: the high precipitation rate was in July 11, 2016 that was characterized by a flash flooding and high winds in the US (explaining the high value).



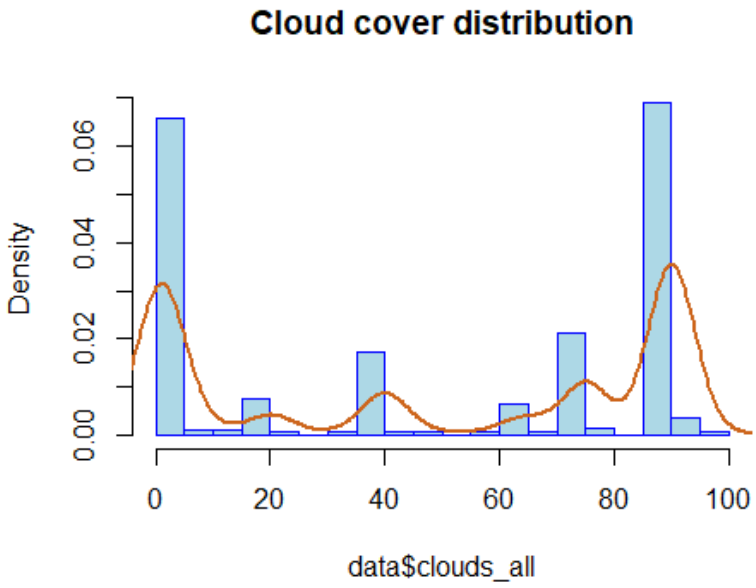
We can observe that a 0 to 5mm precipitation rate is the most frequent one. Other ranges are much less frequent especially when crossing the 10mm value.

For snow variable, since it evidently cannot snow all year round, we can remove all the 0 values and then observe the distribution of this variable:



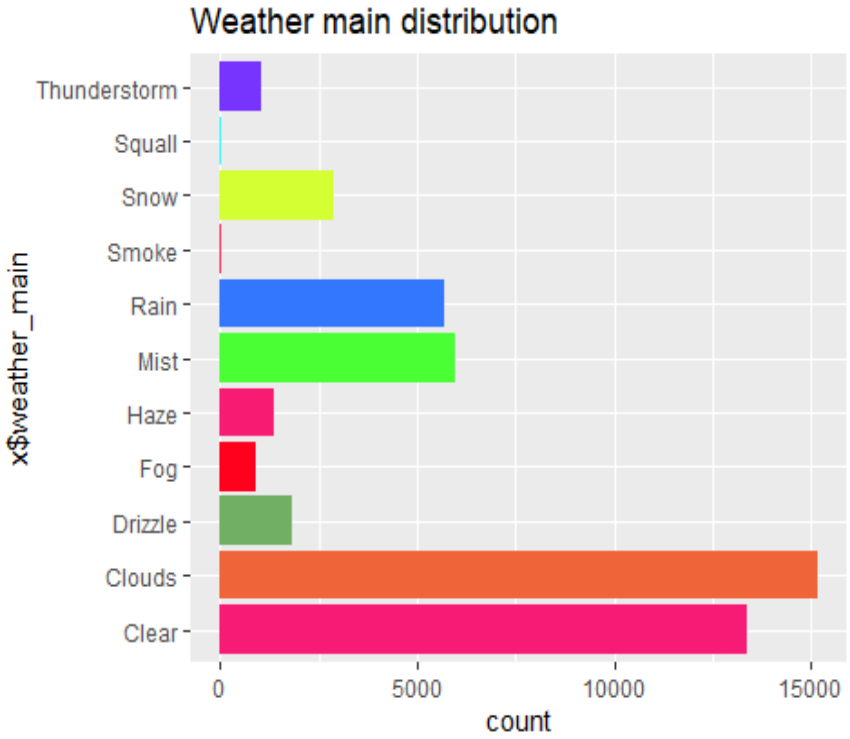
As observed above, the amount of snow is usually between 0.05 and 0.1mm (most frequent amount) and in cases of heavy snow it is between 0.9 and 1mm.

For the clound variable:

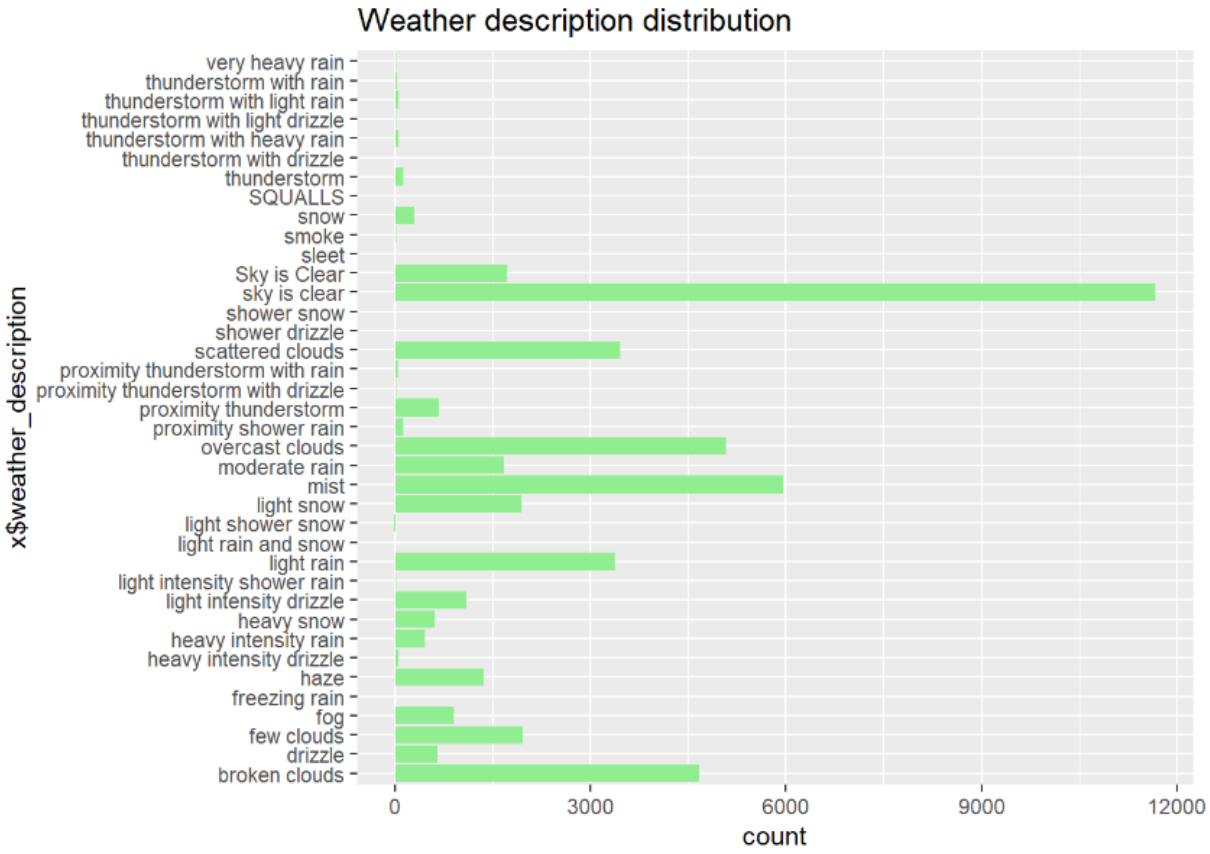


We can see that the variable “cloud” varies substantially where the value of around 0~5% and that of 85~90% are the most frequent ones (constituting the clear days and quite cloudy ones).

Let’s now observe the weather main and weather description variables. Starting with the main weather one:



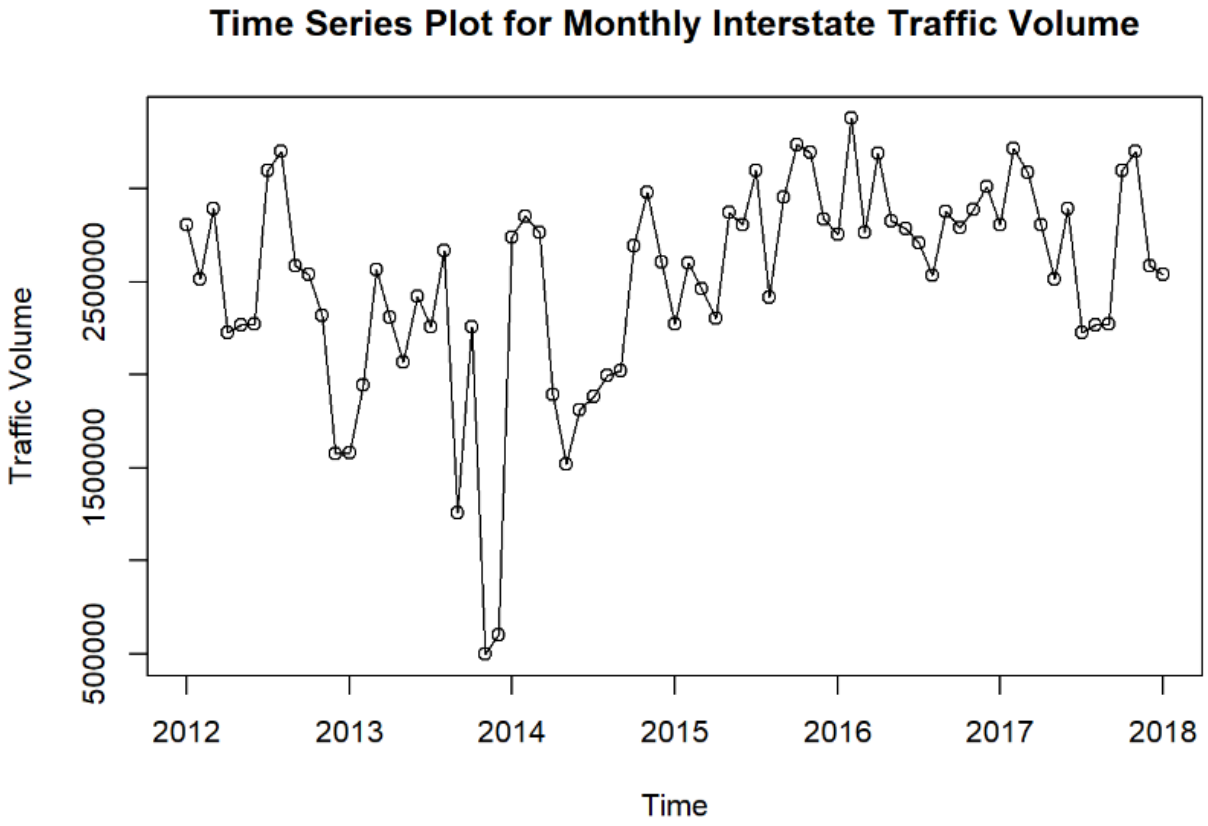
“Clouds” and “Clear” are the most frequent weather conditions, followed by “Mist”, “rain” and “snow”. Other weather conditions are way less frequent. Let’s now move on to observe the description of weather:



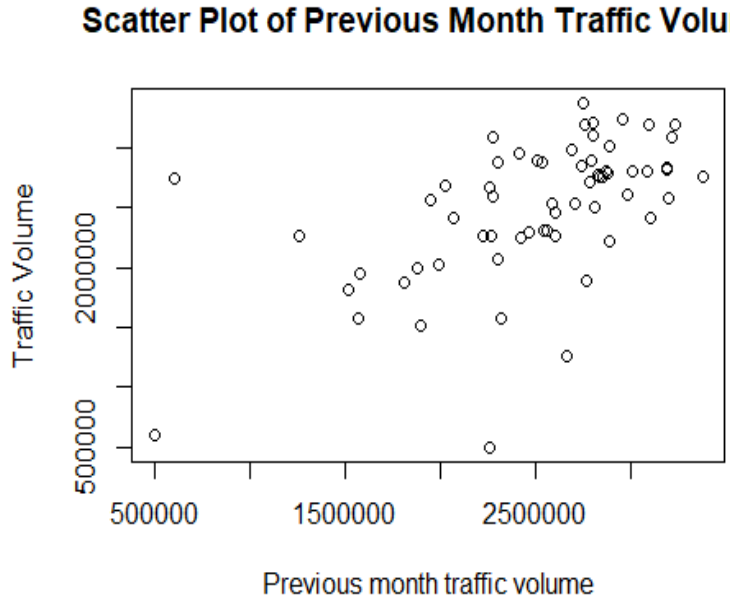
Time Series Exploration

Univariate analysis

Let's start by tracing our times series:

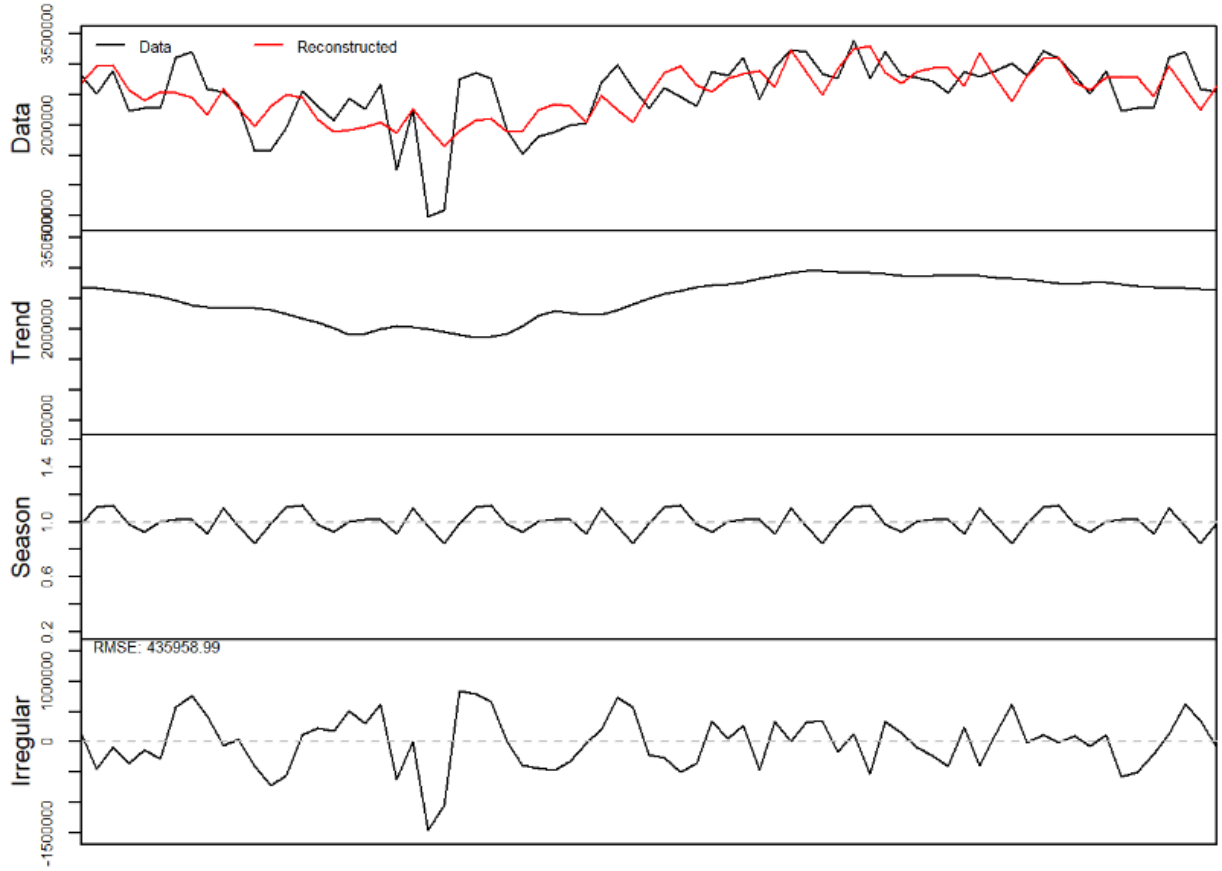


The data shows some seasonality trends and there is a huge decrease in the amount of traffic at the end of 2014 even though it is holiday season. We can also check whether the previous month's traffic affects the traffic volume of the next month.



We can observe that there is a weak (and positive) correlation between the previous month's traffic volume to the next month.

Let's now decompose the time series to observe its components (level,trend,seasonality and noise):



When observing this first decomposition, we can notice that the time series in hand presents a bit of a trend, where we can also notice the drop in the traffic volume (the one in 2014). It seems to also have weak to no seasonality. Lastly, it presents a high residual error. To be able to properly evaluate all the components, it is needed to perform the necessary tests.

Let's start with the trend component.

We consider the following pair of hypotheses:

H0: no trend

H1: linear trend,

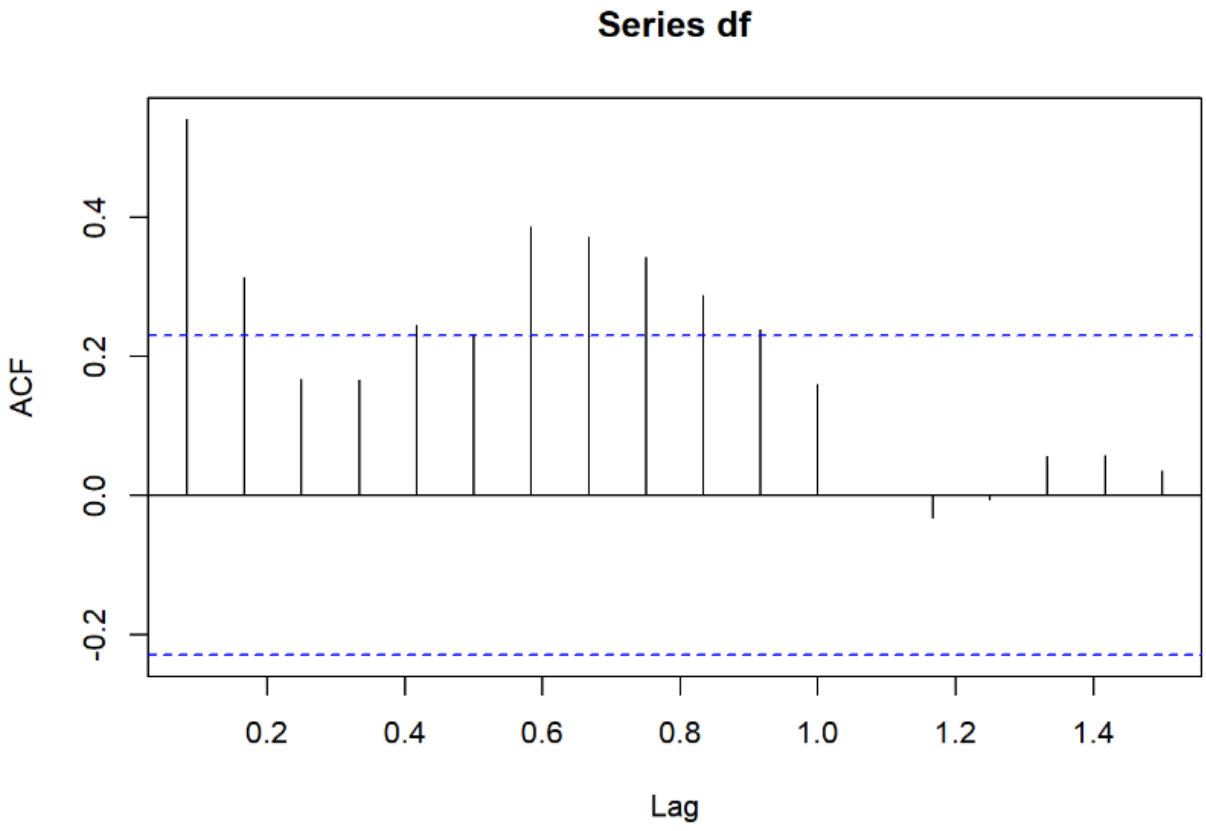
that can be tested specifically using t-test.

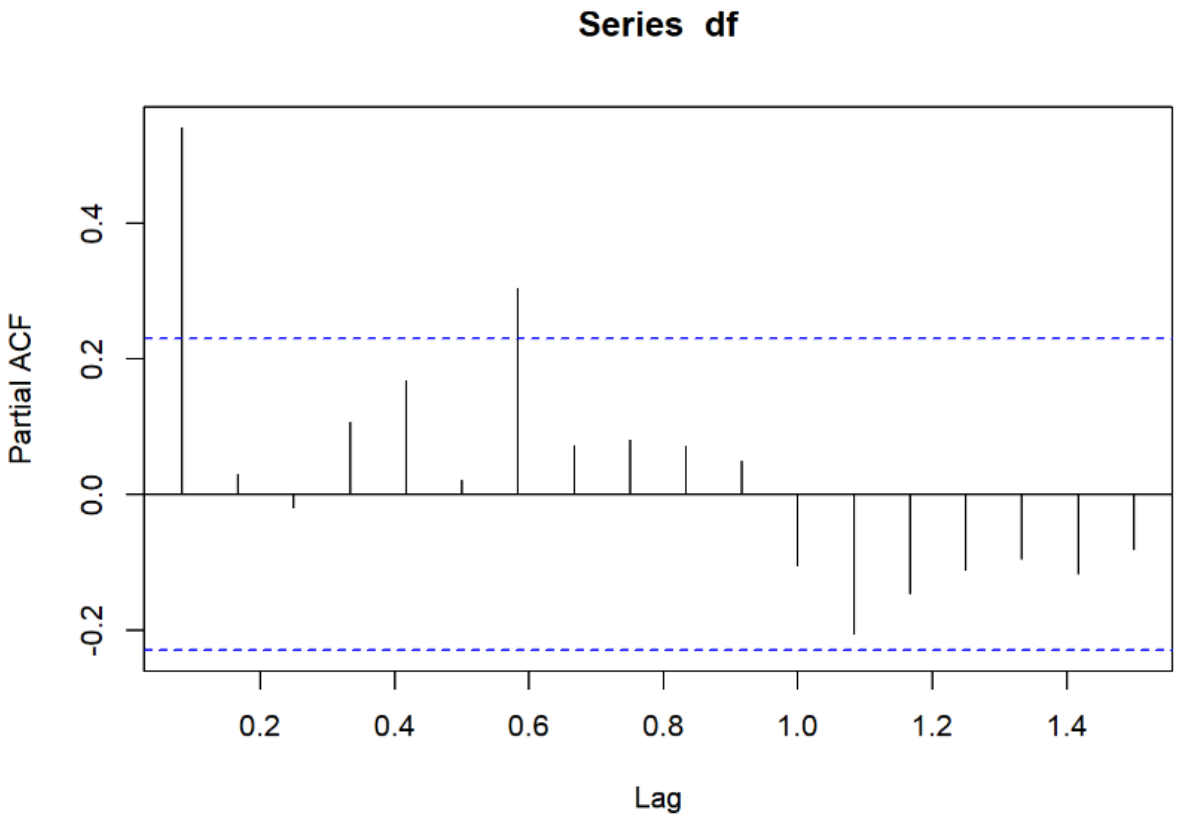
Assuming the time series may be auto-correlated (which is the usual case with observational data), we apply sieve-bootstrap version of the t-test, by adapting the approach of Noguchi, Gel, and Duguay (2011):

```
##
##  Sieve-bootstrap Student's t-test for a linear trend
##
## data:  df
## Student's t value = 3.3303, p-value = 0.047
## alternative hypothesis: linear trend.
## sample estimates:
## $AR_order
## [1] 1
##
## $AR_coefficients
##     phi_1
## 0.3781532
```

The small p-value (0.038 < 0.05) correctly indicates that there is enough evidence to reject the hypothesis of no trend in H0 in favor of the alternative hypothesis of a linear trend. In other words our time series does in fact have a trend.

Let's now check the seasonality component. Using the ACF and PACF graphs we can have an idea on whether a seasonality exists:





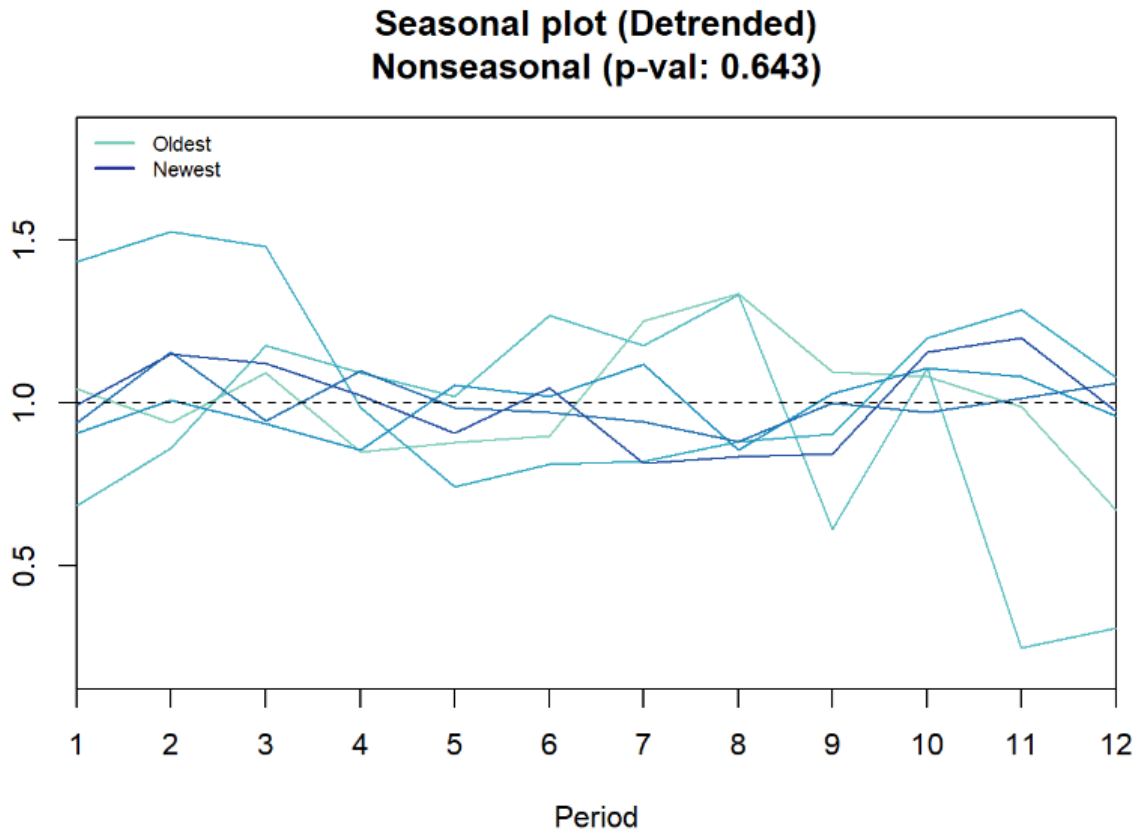
Observing the results of ACF and PACF, it is hard to observe any seasonality. To have definite results we can also try different seasonality tests such as Kruskall Wallis test, QS test, F-Test on seasonal dummies and Welch seasonality test. We can also use the function "isSeasonal" in R (with test=combined) that returns a bollean indicating the presence/absence of seasonality based on the combination of multiple tests. The results are shown bellow:

```
## Test used:  Kruskall Wallis
##
## Test statistic:  14.65
## P-value:  0.1993137

## Test used:  QS
##
## Test statistic:  0
## P-value:  1

## Test used:  SeasonalDummies
##
## Test statistic:  1.09
## P-value:  0.3868423

## Test used:  Kruskall Wallis
##
```

```
## Test statistic:  1.51
## P-value:  0.1921694

## [1] FALSE
```

All seasonality tests have high p-values (>.05) indicating the rejection of the hypothesis that the times series has a seasonality. Through the combined test too (result = FALSE) we can confirm the absence of a substantial seasonality in the traffic volume time series.

PS: we can also use "seasplot" function in R to check for both trend and seasonality in our time series (as shown below the previously found results are confirmed).




```
## Results of statistical testing
## Evidence of trend: TRUE  (pval: 0)
## Evidence of seasonality: FALSE  (pval: 0.643)
```

Let's now observe stationary: to do so we can use the Augmented Dickey-Fuller Test.

```
##
##  Augmented Dickey-Fuller Test
##
## data:  df
## Dickey-Fuller = -2.7739, Lag order = 4, p-value = 0.2601
## alternative hypothesis: stationary
```

The p-value from the test is high ($\alpha > 0.05$), then we cannot reject the null hypothesis and as a result we conclude that the time series is not stationary.

To conclude, our time series:

1- Has a trend (so needs de-trending)

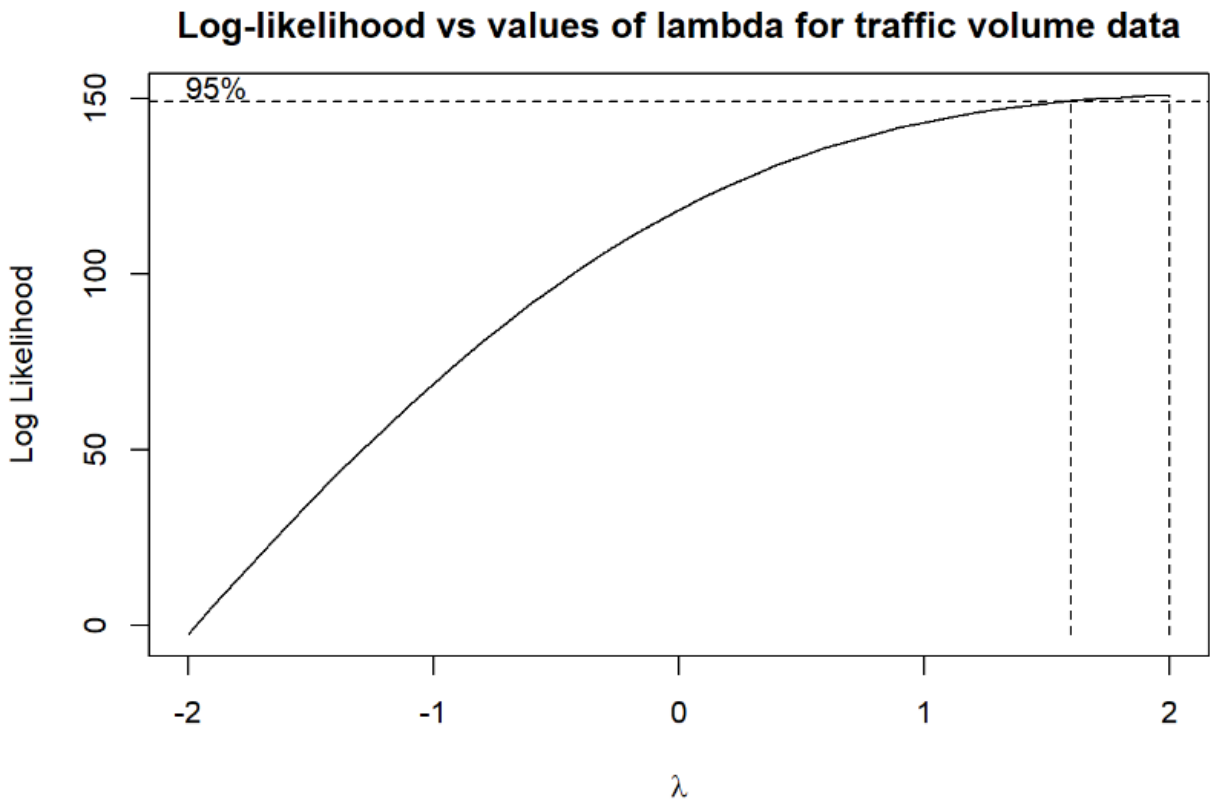
2- Has no seasonality

3- Non stationary (needs to be normalized)

4- Has high residual errors (needs some sort of filtering)

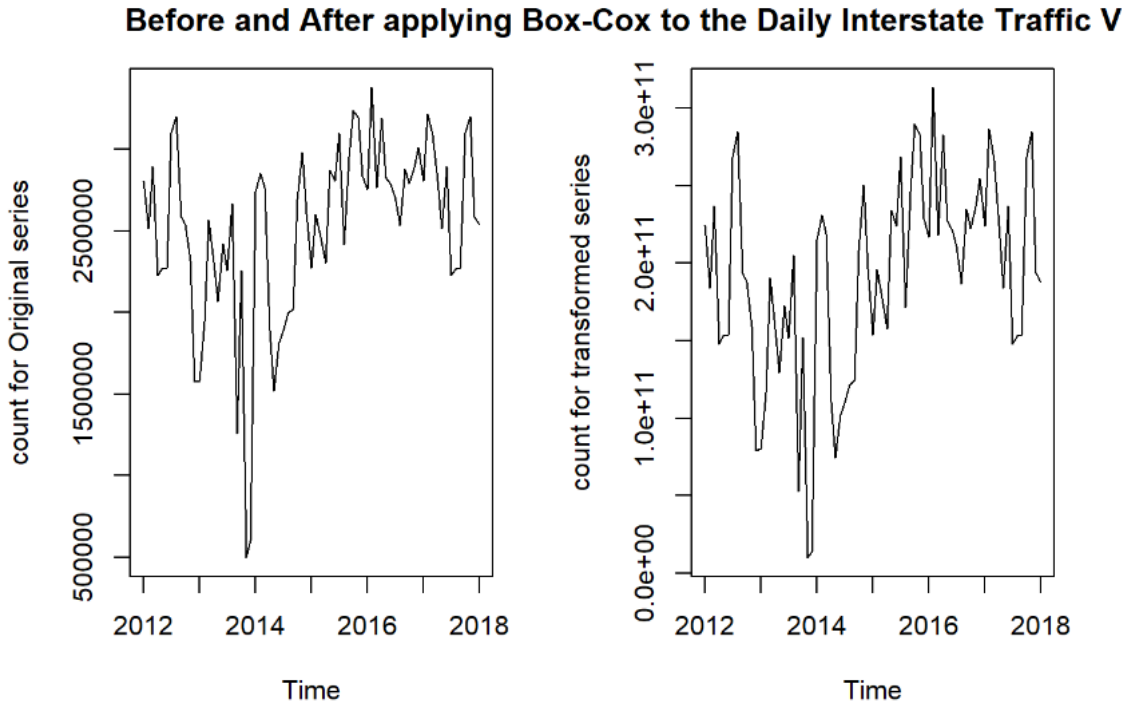
In other words our data needs pre-treatement before we can move on to modelling and prediction.

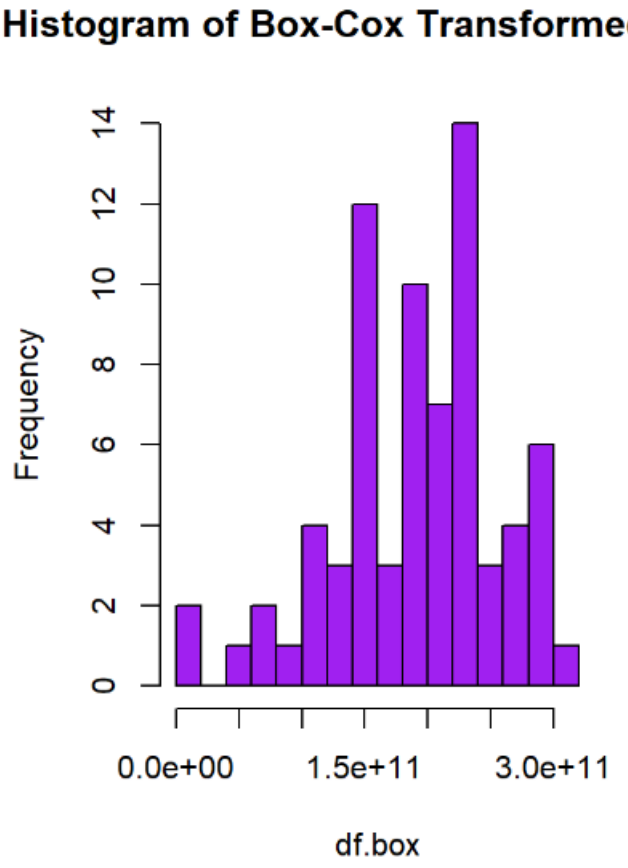
Starting with stationnarity, Box-cox transformation is applied to the series to help make the series stationary.




```
## [1] 1.6 2.0
```

The 95% confidence interval for lambda contains the values of lambda between 1.6 and 2.0. Here, we take the center of the two numbers 1.8 to apply it to the Box-Cox transformation.


Before and After applying Box-Cox to the Daily Interstate Traffic V
count for Original series
500000
1500000
2500000
2012
2014
2016
2018
Time
count for transformed series
0.0e+00
1.0e+11
2.0e+11
3.0e+11
2012
2014
2016
2018
Time


Histogram of Box-Cox Transformed
Frequency
0
2
4
6
8
10
12
14
0.0e+00
1.5e+11
3.0e+11
df.box

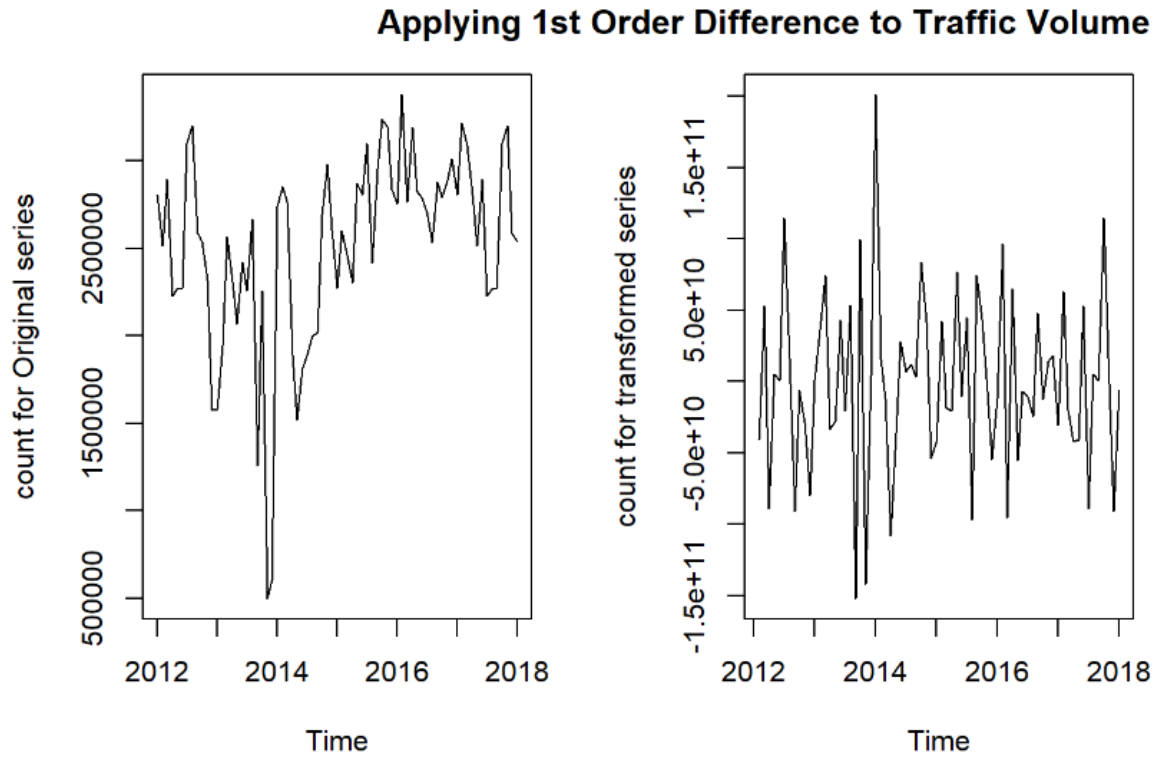
The data looks more normally distributed after the Box-Cox transformation as seen above. However, further Shapiro-Wilk normality test is applied to see if it passes the normality test.

```
##
##  Shapiro-Wilk normality test
##
## data:  df.box
## W = 0.96789, p-value = 0.05943
```

The Shapiro-Wilk normality test's p-value is greater than 0.05 and therefore, we can assume the new transformed data is normally distributed. After normalizing, the data is differentiated to make the series stationary.

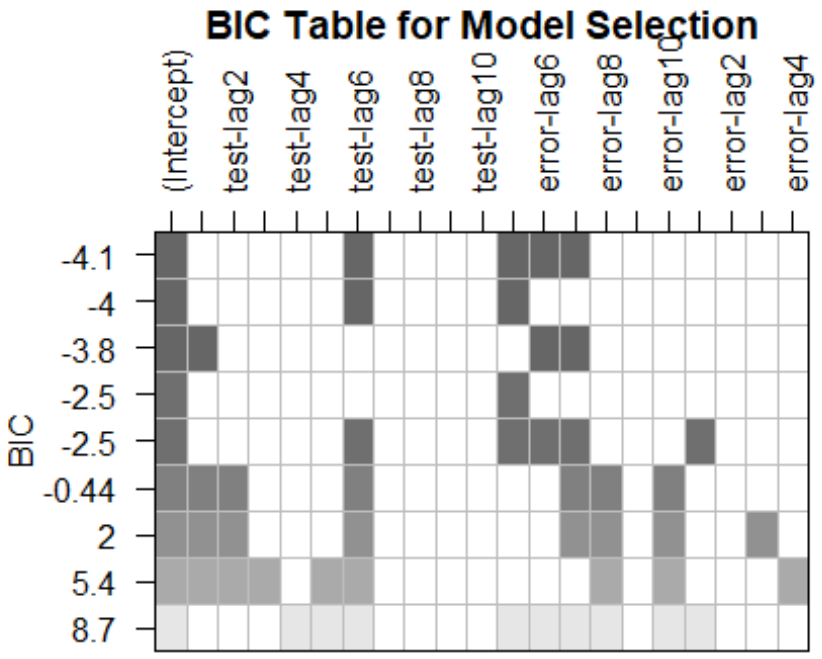
```
##
##  Augmented Dickey-Fuller Test
##
## data:  df.diff
## Dickey-Fuller = -5.2679, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

It is seen from the results that the p-value of the Dickey-Fuller test is < 0.01 and therefore the series is now stationary. Therefore, first order differentiating will be used to try to fit in the models. The following shows how the data looks like after applying Box-Cox and first order differentiating to the data. It can be seen that transformation also reduced the variance of the data as well.



As the series is stationary at this stage, EACF method will be tried on to help determine which models are to be tested out. According to the eacf table below, it is not clear which models are the best but we take ARMA(1,1), ARMA(1,2) and ARMA (1,3) as best possible candidates.

```
## AR/MA
##    0 1 2 3 4 5 6 7 8 9 10
## 0  x o o o o x o o o o o
## 1  x o o o o o o o o o o
## 2  x x o o o o o o o o o
## 3  x x x o o o o o o o o
## 4  o o x o o x o o o o o
## 5  o x x o o x o o o o o
## 6  o o o o x x o o o o o
## 7  x o o o o o o o o o o
## 8  x x o o o o o o o o o
## 9  x o o o o o o o o o o
## 10 x x x x o o o o o o o

## Reordering variables and trying again:
```



According to the BIC table, the smallest BIC are at lag 1,and 6 and at error lag 5,6 and 7. Therefore, we take ARIMA (1,1,1), ARIMA (1,1,2), ARIMA (1,1,3) , ARIMA (1,1,5) , ARIMA (1,1,6), ARIMA (6,1,5), and ARIMA(6,1,6) as possible candidates.

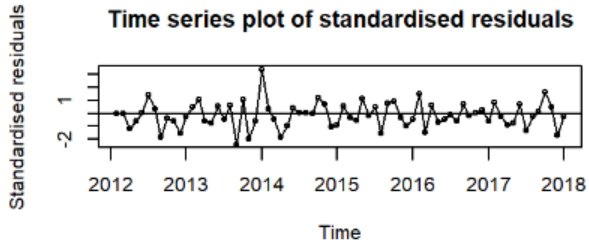
After building all the possible models we move towards selecting the best model. This can be done using the AIC and BIC scores:

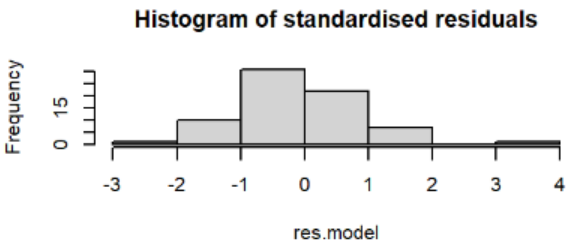
```
##              df      AIC
## model_615_ml 12 3725.523
## model_116_ml  8 3725.941
## model_112_ml  4 3728.780
## model_616_ml 13 3728.941
## model_115_ml  7 3730.317
## model_113_ml  5 3730.584
## model_111_ml  3 3737.229

##              df      BIC
## model_112_ml  4 3737.830
## model_113_ml  5 3741.897
## model_111_ml  3 3744.017
## model_116_ml  8 3744.043
## model_115_ml  7 3746.155
## model_615_ml 12 3752.675
## model_616_ml 13 3758.356
```

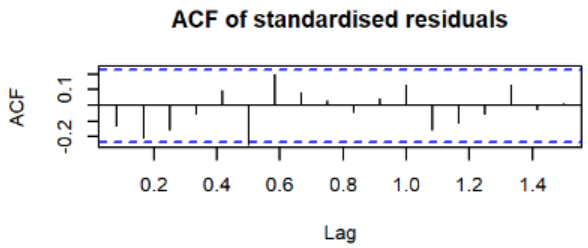
The two tests show contradictory results therefore, we will take ARIMA (6,1,5), ARIMA (1,1,2) ARIMA (1,1,3) and ARIMA (1,1,6) to test for residuals to see which model will be the best.

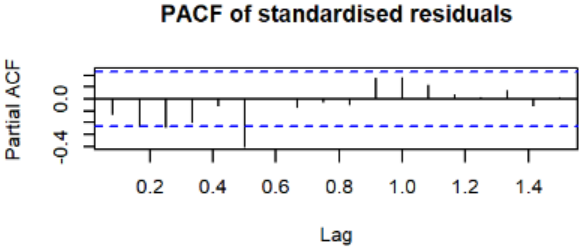
Residual analysis for ARIMA(1,1,2)

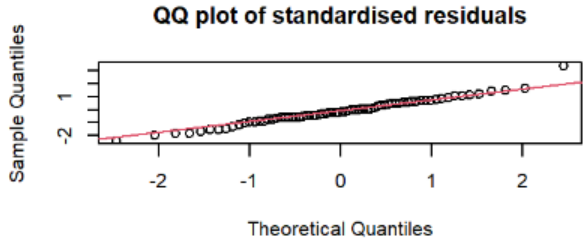
```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.97718, p-value = 0.2147
```





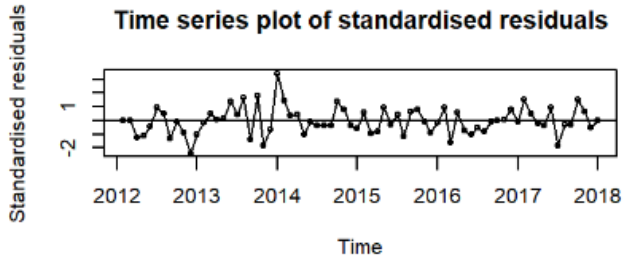


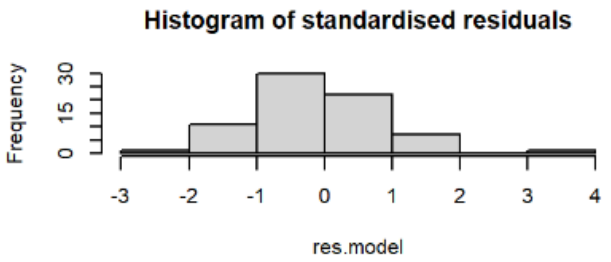


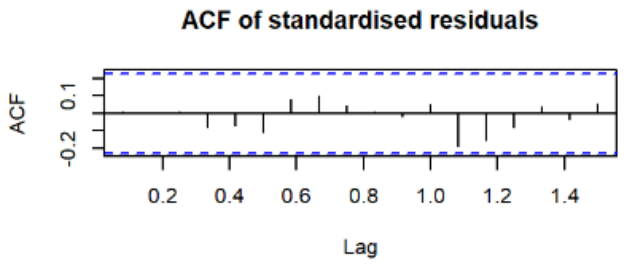


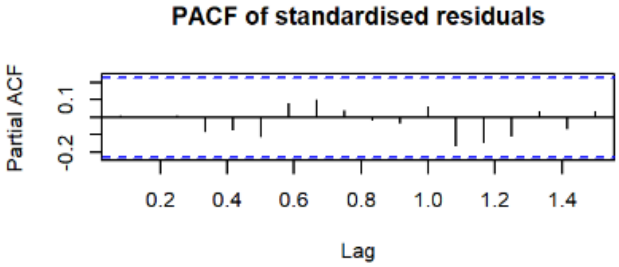
Residual analysis for ARIMA(1,1,6)

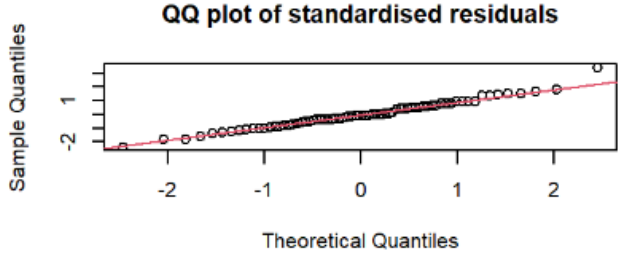
```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.98043, p-value = 0.3243
```





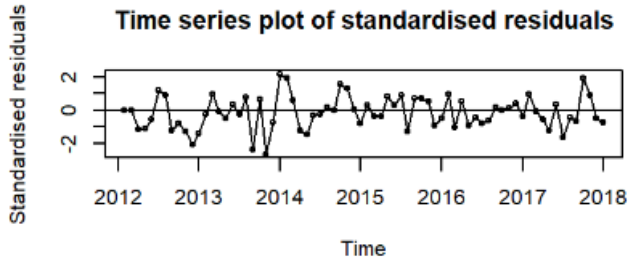


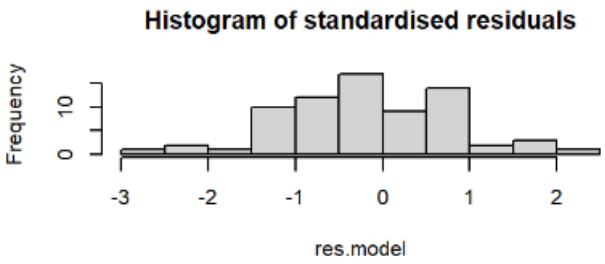


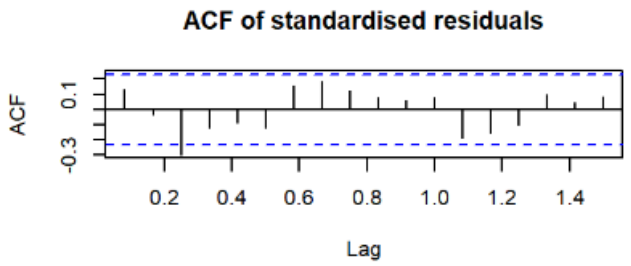


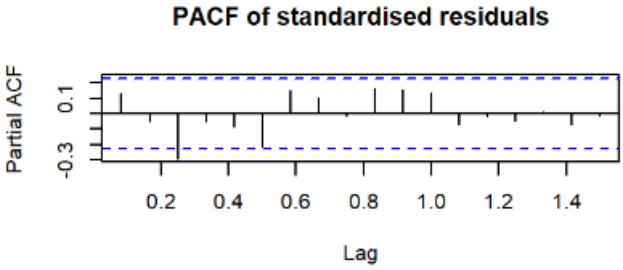
Residual analysis for ARIMA(1,1,3)

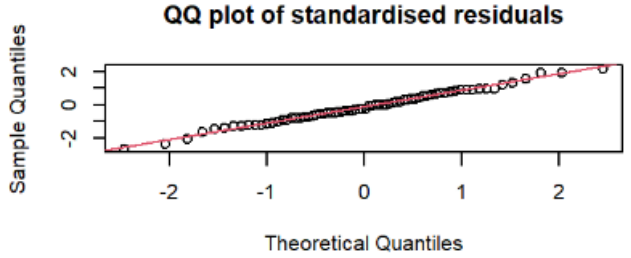
```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.99186, p-value = 0.9274
```





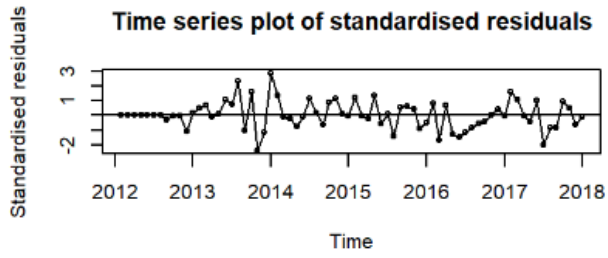


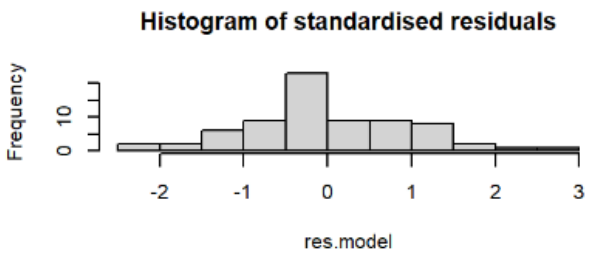


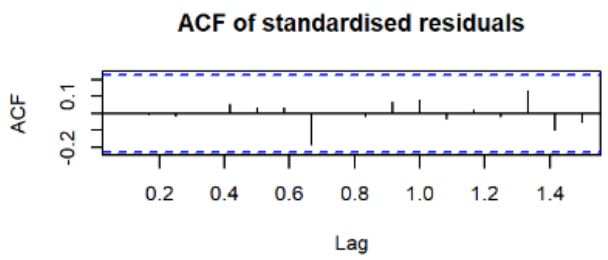


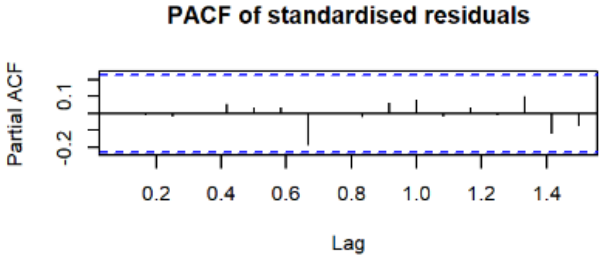
Residual analysis for ARIMA(6,1,5)

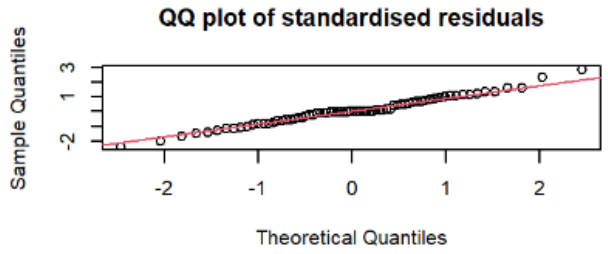
```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.98358, p-value = 0.471
```







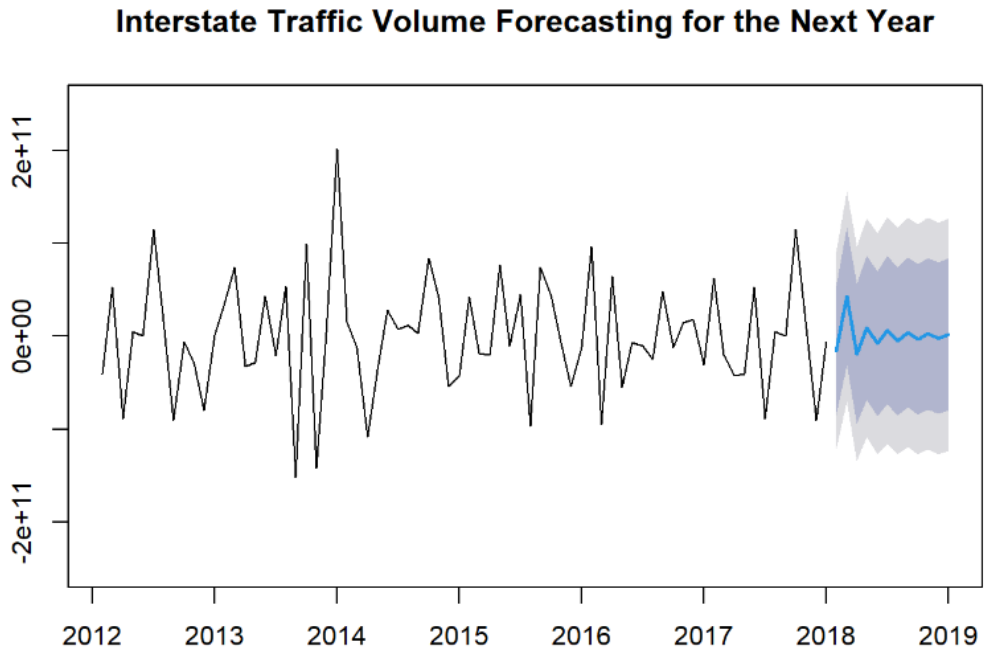




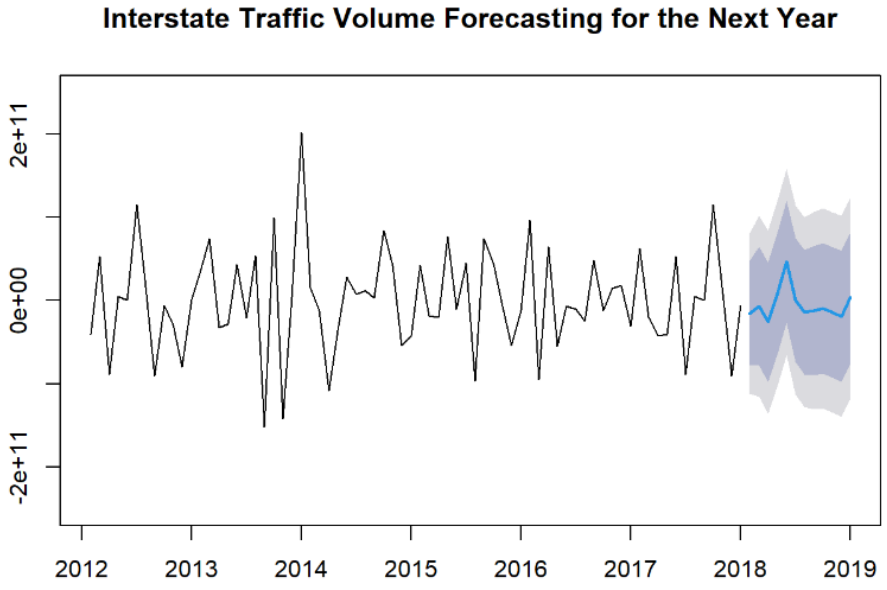
Only at ARIMA (1,1,6), and (6,1,5), the models' standardized residual plots show no trend nor changing variance meaning that the both of the models are supported. The histograms are also normally distributed and the QQ plot also shows that the data is normally distributed. Both the PACF and ACF plots do not show significant lags as well. The better performing model is ARIMA(6,1,5) but let's trace the next year's prediction of both models and observe.

Prediction

ARIMA (1,1,6) prediction



ARIMA (6,1,5) prediction

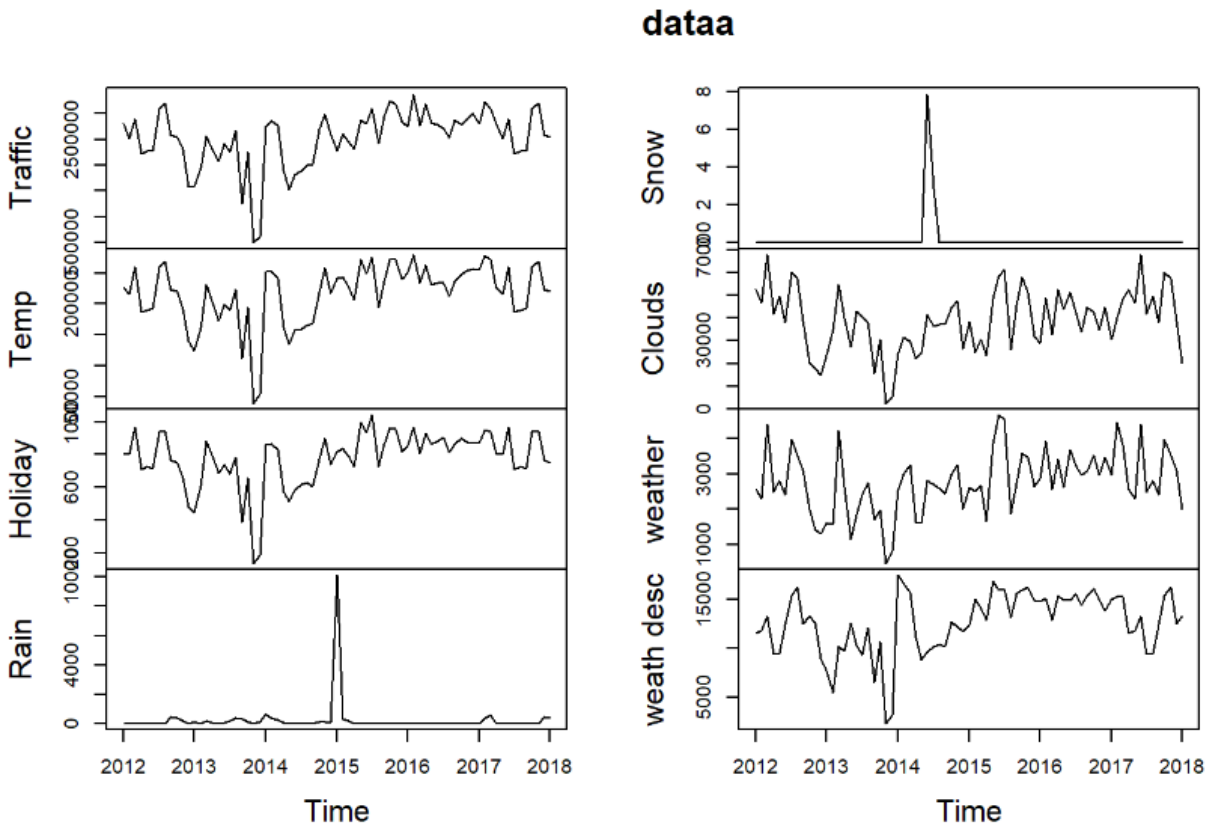


As observed above the ARIMA(6,1,5) is much better performing than ARIMA(1,1,6) so it is the model that we retain for the univariate time series modelling.

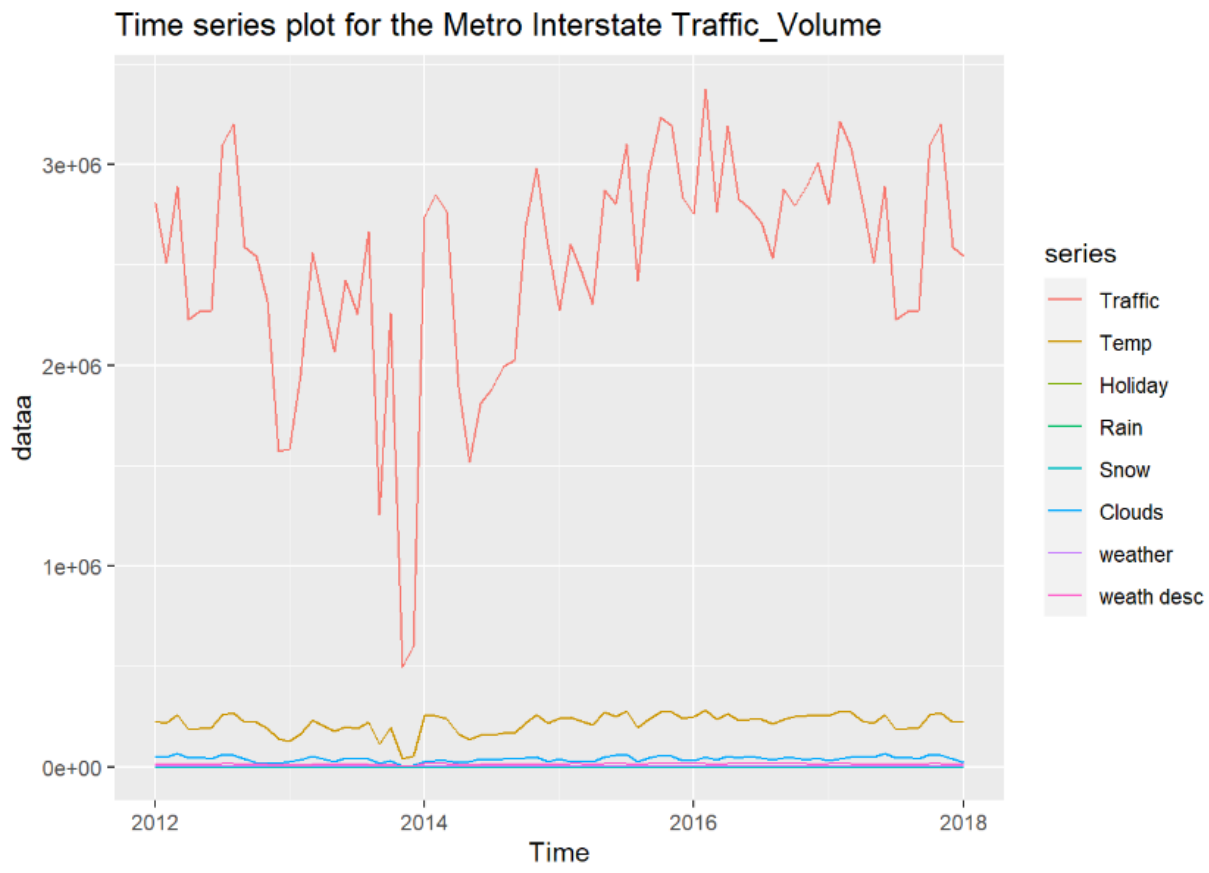
Multivariate analysis

Let's now move on to the multivariate time series modelling.

First of all we need to actually form the time series that pair the time stamp with each one of all the variables we have in our dataset. After forming these multiple time series we can plot each one of them as observed below:



Now we move on to forming one Multivariate time series that has all the possible univariate time series we have. We can observe it bellow:



After observing the time series we can observe if each one of the components is stationary or not. To do so we use adf test, the results are shown bellow:

```
## $Traffic
##
##  Augmented Dickey-Fuller Test
##
## data:  newX[, i]
## Dickey-Fuller = -2.7739, Lag order = 4, p-value = 0.2601
## alternative hypothesis: stationary
##
##
## $Temp
##
##  Augmented Dickey-Fuller Test
##
## data:  newX[, i]
## Dickey-Fuller = -2.8309, Lag order = 4, p-value = 0.2368
## alternative hypothesis: stationary
##
##
## $Holiday
##
```

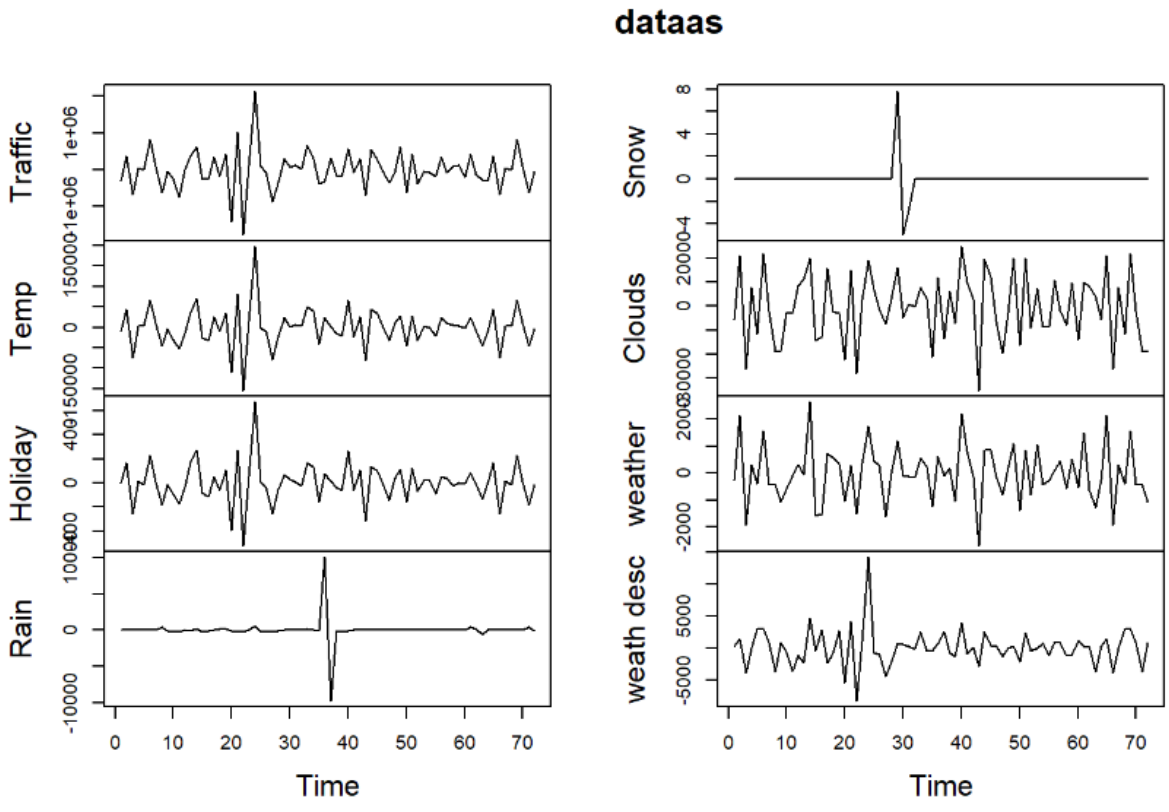
```
##  Augmented Dickey-Fuller Test
##
## data:  newX[, i]
## Dickey-Fuller = -2.7339, Lag order = 4, p-value = 0.2764
## alternative hypothesis: stationary
##
##
## $Rain
##
##  Augmented Dickey-Fuller Test
##
## data:  newX[, i]
## Dickey-Fuller = -3.7486, Lag order = 4, p-value = 0.02677
## alternative hypothesis: stationary
##
##
## $Snow
##
##  Augmented Dickey-Fuller Test
##
## data:  newX[, i]
## Dickey-Fuller = -3.5234, Lag order = 4, p-value = 0.0461
## alternative hypothesis: stationary
##
##
## $Clouds
##
##  Augmented Dickey-Fuller Test
##
## data:  newX[, i]
## Dickey-Fuller = -3.4606, Lag order = 4, p-value = 0.05279
## alternative hypothesis: stationary
##
##
## $weather
##
##  Augmented Dickey-Fuller Test
##
## data:  newX[, i]
## Dickey-Fuller = -3.2796, Lag order = 4, p-value = 0.08188
## alternative hypothesis: stationary
##
##
## $`weath desc`
##
##  Augmented Dickey-Fuller Test
##
## data:  newX[, i]
## Dickey-Fuller = -2.5931, Lag order = 4, p-value = 0.3337
## alternative hypothesis: stationary
```

As seen in the adf test results the only stationary time series are those of rain and snow (only ones with p-value < 0.05). In other words our time series is not stationary. Before moving on to forming a multivariate model (here we will be using VAR), it is important to prepare the time series, similar to what we did in the univariate analysis.

To make the time series stationary we of course differntiate, then we run the adf test again and observe the results:

```
## $Traffic
##
##  Augmented Dickey-Fuller Test
##
## data:  newX[, i]
## Dickey-Fuller = -5.5597, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
##
##
## $Temp
##
##  Augmented Dickey-Fuller Test
##
## data:  newX[, i]
## Dickey-Fuller = -5.4288, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
##
##
## $Holiday
##
##  Augmented Dickey-Fuller Test
##
## data:  newX[, i]
## Dickey-Fuller = -5.365, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
##
##
## $Rain
##
##  Augmented Dickey-Fuller Test
##
## data:  newX[, i]
## Dickey-Fuller = -5.8483, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
##
##
## $Snow
##
##  Augmented Dickey-Fuller Test
##
## data:  newX[, i]
## Dickey-Fuller = -5.7056, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
##
##
## $Clouds
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  newX[, i]
## Dickey-Fuller = -4.5554, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
##
##
## $weather
##
##  Augmented Dickey-Fuller Test
##
## data:  newX[, i]
## Dickey-Fuller = -6.0335, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
##
##
## $`weath desc`
##
##  Augmented Dickey-Fuller Test
##
## data:  newX[, i]
## Dickey-Fuller = -4.7033, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```



We can see now that our time series has become stationary as all the p-values observed above are significant (it is also observable in the traced graphs).

Now that out time series is ready we can start VAR modelling. To do so, we need to find the good parameter for our VAR modelling. One way to do so is through using the "VARselect" algorithm:

```
## $selection
## AIC(n)  HQ(n)  SC(n) FPE(n)
##      6      6      1      7
##
## $criteria
##                  1            2            3            4            5
## AIC(n) 1.106223e+02 1.110514e+02 1.115915e+02 1.102020e+02 1.082681e+02
## HQ(n)  1.115727e+02 1.128464e+02 1.142313e+02 1.136865e+02 1.125974e+02
## SC(n)  1.130309e+02 1.156009e+02 1.182819e+02 1.190334e+02 1.192404e+02
## FPE(n) 1.119144e+48 1.871640e+48 4.059258e+48 1.663130e+48 6.279966e+47
##                  6            7
## AIC(n) 1.049037e+02          NaN
## HQ(n)  1.100777e+02          NaN
## SC(n)  1.180169e+02          NaN
## FPE(n) 1.390281e+47 -5.287155e+16
```

Observing the selection algorithm results we can conclude that the lag.max=6 model is the optimal option as the selection criteria (AIC, HQ) are better for this one.

We can now create the VAR model with information obtained from the prior selection:

```
##
## VAR Estimation Results:
## ========================
## Endogenous variables: Traffic, Temp, Holiday, Rain, Snow, Clouds, weather,
weath.desc
## Deterministic variables: none
## Sample size: 66
## Log Likelihood: -3814.327
## Roots of the characteristic polynomial:
## 0.9815 0.9815 0.9803 0.9803 0.9682 0.9682 0.9639 0.9639 0.9564 0.9564 0.95
56 0.9556 0.9541 0.9541 0.9474 0.9474 0.9446 0.9446 0.9424 0.9424 0.9423 0.94
23 0.9329 0.9329 0.9325 0.9325 0.9256 0.9256 0.9222 0.9222 0.9163 0.9163 0.90
73 0.9073 0.8876 0.8876 0.882 0.882 0.8755 0.8755 0.8539 0.8539 0.8206 0.8206
0.5944 0.4811 0.4811 0.1734
## Call:
## vars::VAR(y = dataas, type = "none", lag.max = 6, ic = "AIC")
##
##
## Estimation results for equation Traffic:
## ========================================
## Traffic = Traffic.l1 + Temp.l1 + Holiday.l1 + Rain.l1 + Snow.l1 + Clouds.l
1 + weather.l1 + weath.desc.l1 + Traffic.l2 + Temp.l2 + Holiday.l2 + Rain.l2
+ Snow.l2 + Clouds.l2 + weather.l2 + weath.desc.l2 + Traffic.l3 + Temp.l3 + H
oliday.l3 + Rain.l3 + Snow.l3 + Clouds.l3 + weather.l3 + weath.desc.l3 + Traf
fic.l4 + Temp.l4 + Holiday.l4 + Rain.l4 + Snow.l4 + Clouds.l4 + weather.l4 +
weath.desc.l4 + Traffic.l5 + Temp.l5 + Holiday.l5 + Rain.l5 + Snow.l5 + Cloud
s.l5 + weather.l5 + weath.desc.l5 + Traffic.l6 + Temp.l6 + Holiday.l6 + Rain.
l6 + Snow.l6 + Clouds.l6 + weather.l6 + weath.desc.l6
##
##                  Estimate Std. Error t value Pr(>|t|)
## Traffic.l1     1.297e+00  1.350e+00   0.961   0.3494
## Temp.l1        1.674e+01  3.768e+01   0.444   0.6622
## Holiday.l1    -9.707e+03  9.711e+03  -1.000   0.3308
## Rain.l1        8.732e+01  1.093e+02   0.799   0.4348
## Snow.l1        9.369e+04  1.105e+05   0.848   0.4076
## Clouds.l1     -8.446e+00  2.980e+01  -0.283   0.7801
## weather.l1     2.292e+02  3.456e+02   0.663   0.5155
## weath.desc.l1  4.830e+00  8.155e+01   0.059   0.9534
## Traffic.l2     2.727e+00  1.556e+00   1.753   0.0967 .
## Temp.l2       -5.246e+01  4.492e+01  -1.168   0.2581
## Holiday.l2     3.499e+03  1.016e+04   0.344   0.7345
## Rain.l2        1.364e+02  1.276e+02   1.069   0.2994
## Snow.l2       -6.060e+04  9.934e+04  -0.610   0.5495
## Clouds.l2     -1.224e+01  2.314e+01  -0.529   0.6033
## weather.l2     4.621e+02  3.273e+02   1.412   0.1750
```

```
## weath.desc.l2  3.928e+01  7.718e+01   0.509   0.6170
## Traffic.l3    -8.979e-01  1.356e+00  -0.662   0.5164
## Temp.l3        1.992e+01  4.130e+01   0.482   0.6353
## Holiday.l3    -4.048e+03  1.079e+04  -0.375   0.7119
## Rain.l3        8.451e+00  1.072e+02   0.079   0.9381
## Snow.l3       -1.278e+05  1.248e+05  -1.024   0.3196
## Clouds.l3      2.384e+01  3.232e+01   0.738   0.4703
## weather.l3    -1.089e+02  3.140e+02  -0.347   0.7328
## weath.desc.l3 -5.094e+01  9.090e+01  -0.560   0.5821
## Traffic.l4    -1.499e+00  1.599e+00  -0.938   0.3608
## Temp.l4       -4.687e+00  3.458e+01  -0.136   0.8937
## Holiday.l4     9.422e+03  7.999e+03   1.178   0.2542
## Rain.l4        3.314e+01  1.077e+02   0.308   0.7619
## Snow.l4        1.716e+04  8.806e+04   0.195   0.8477
## Clouds.l4     -2.199e+01  3.353e+01  -0.656   0.5204
## weather.l4     5.896e+01  4.109e+02   0.143   0.8875
## weath.desc.l4 -2.232e+02  9.003e+01  -2.479   0.0233 *
## Traffic.l5     3.263e-01  1.757e+00   0.186   0.8547
## Temp.l5       -9.264e+00  3.943e+01  -0.235   0.8169
## Holiday.l5    -9.295e+02  8.525e+03  -0.109   0.9144
## Rain.l5        1.129e+02  1.235e+02   0.914   0.3729
## Snow.l5       -2.460e+04  8.451e+04  -0.291   0.7743
## Clouds.l5     -2.332e+01  2.349e+01  -0.993   0.3340
## weather.l5     3.605e+02  2.993e+02   1.205   0.2440
## weath.desc.l5  1.607e+02  1.005e+02   1.599   0.1272
## Traffic.l6     2.282e+00  2.090e+00   1.092   0.2892
## Temp.l6       -2.413e+01  4.344e+01  -0.556   0.5853
## Holiday.l6     3.611e+02  8.170e+03   0.044   0.9652
## Rain.l6        2.020e+02  1.166e+02   1.733   0.1002
## Snow.l6       -7.633e+04  8.936e+04  -0.854   0.4042
## Clouds.l6     -2.007e+01  2.208e+01  -0.909   0.3754
## weather.l6     1.867e+02  2.186e+02   0.854   0.4044
## weath.desc.l6 -1.333e+02  1.042e+02  -1.279   0.2171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Residual standard error: 434300 on 18 degrees of freedom
## Multiple R-Squared: 0.8242,  Adjusted R-squared: 0.3553
## F-statistic: 1.758 on 48 and 18 DF,  p-value: 0.09563
##
##
##
##
## Covariance matrix of residuals:
##               Traffic       Temp   Holiday       Rain       Snow    Clouds
## Traffic     1.886e+11  1.666e+10 5.606e+07 33200882.4 -1069.3243 1.964e+09
## Temp        1.666e+10  1.534e+09 5.150e+06  8258353.9  -325.9073 1.736e+08
## Holiday     5.606e+07  5.150e+06 1.736e+04    25015.1     1.0535 6.178e+05
## Rain        3.320e+07  8.258e+06 2.502e+04  1301569.5   -46.8031 2.419e+06
```

```
## Snow       -1.069e+03 -3.259e+02 1.054e+00      -46.8     0.7934 2.296e+02
## Clouds      1.964e+09  1.736e+08 6.178e+05  2418877.8   229.5735 7.736e+07
## weather     1.003e+08  9.582e+06 3.427e+04   137197.2    31.9098 3.895e+06
## weath.desc  1.014e+09  9.146e+07 3.041e+05  -166984.3   253.0464 2.486e+06
##               weather weath.desc
## Traffic    1.003e+08 1014065899
## Temp       9.582e+06   91458258
## Holiday    3.427e+04     304085
## Rain       1.372e+05    -166984
## Snow       3.191e+01        253
## Clouds     3.895e+06    2486028
## weather    2.802e+05     243943
## weath.desc 2.439e+05    7762275
##
## Correlation matrix of residuals:
##               Traffic      Temp  Holiday     Rain      Snow Clouds weather
## Traffic     1.000000  0.979615 0.979847  0.06701 -0.002764 0.5141 0.43634
## Temp        0.979615  1.000000 0.998072  0.18483 -0.009342 0.5039 0.46221
## Holiday     0.979847  0.998072 1.000000  0.16643  0.008977 0.5331 0.49146
## Rain        0.067009  0.184830 0.166430  1.00000 -0.046056 0.2411 0.22718
## Snow       -0.002764 -0.009342 0.008977 -0.04606  1.000000 0.0293 0.06768
## Clouds      0.514140  0.503906 0.533132  0.24106  0.029303 1.0000 0.83662
## weather     0.436345  0.462207 0.491457  0.22718  0.067676 0.8366 1.00000
## weath.desc  0.838084  0.838187 0.828447 -0.05253  0.101964 0.1015 0.16541
##             weath.desc
## Traffic        0.83808
## Temp           0.83819
## Holiday        0.82845
## Rain          -0.05253
## Snow           0.10196
## Clouds         0.10145
## weather        0.16541
## weath.desc     1.00000
```

The results above show the possible models for each one of variables. As we are only interested in the variable “Traffic volume”, we can focus only on that part. We can see that for the equation “Traffic” the variable weather description with a lag=4 is the only one with a significant p.value < 0.05. The variable Traffic with a lag=1 has a p.value of almost 0.1 and can then be accepted in a more comprehensive scenario.

In other words, the traffic volume can mostly be explained by the weather description of 6 hours ago and in a bit more stretched scenario also the traffic volume 2 hours ago.

Through the covariance matrix too, we can remark that Traffic has a high covariance with almost all the other variables except for snow and especially for Temperature, holiday and the two weather variables.

Also, through the correlation matrix we can observe that traffic is highly correlated with the variables Holiday, Temperature and Weather description.

We can also check the serial autocorrelation in the model residuals using the Portmantau test:

```
##
##  Portmanteau Test (asymptotic)
##
## data:  Residuals of VAR object var.m
## Chi-squared = 1238.8, df = 640, p-value < 2.2e-16
```

Our p-value is < 0.05 so there is strong evidence of autocorrelation among the VAR time series.

Let's now check the causality between the traffic volume and the other time series:

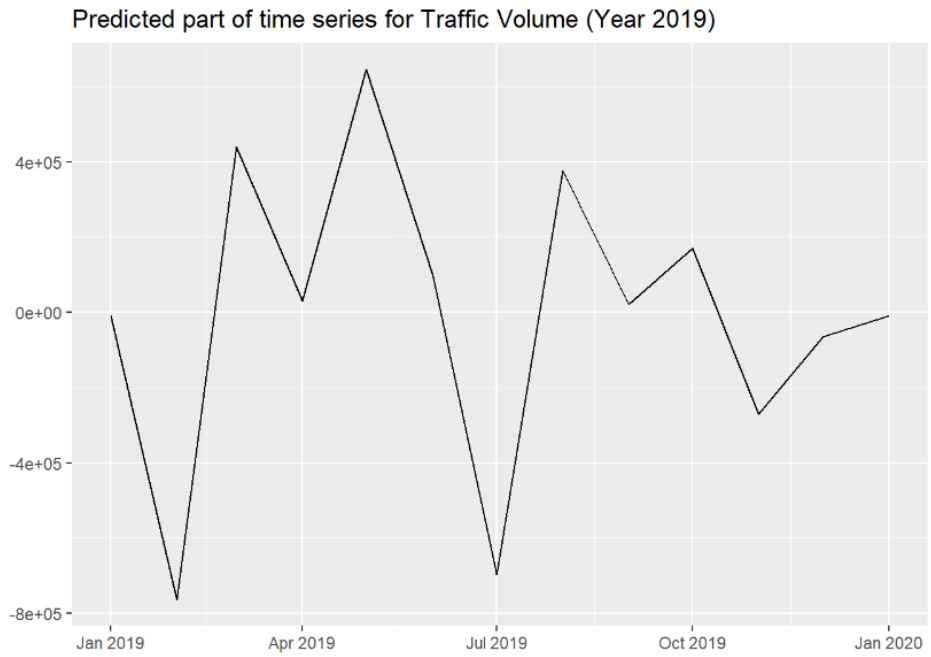
```
##
##  Granger causality H0: Temp Holiday Rain Snow Clouds weather weath.desc
##  do not Granger-cause Traffic
##
## data:  VAR object var.m
## F-Test = 1.2448, df1 = 42, df2 = 144, p-value = 0.1729
```

Our p-value here is high, meaning we should reject H1 and accept H0 that states that the variables (Temperature, Holiday, Rain, Snow, Clouds and weather) do not cause traffic.

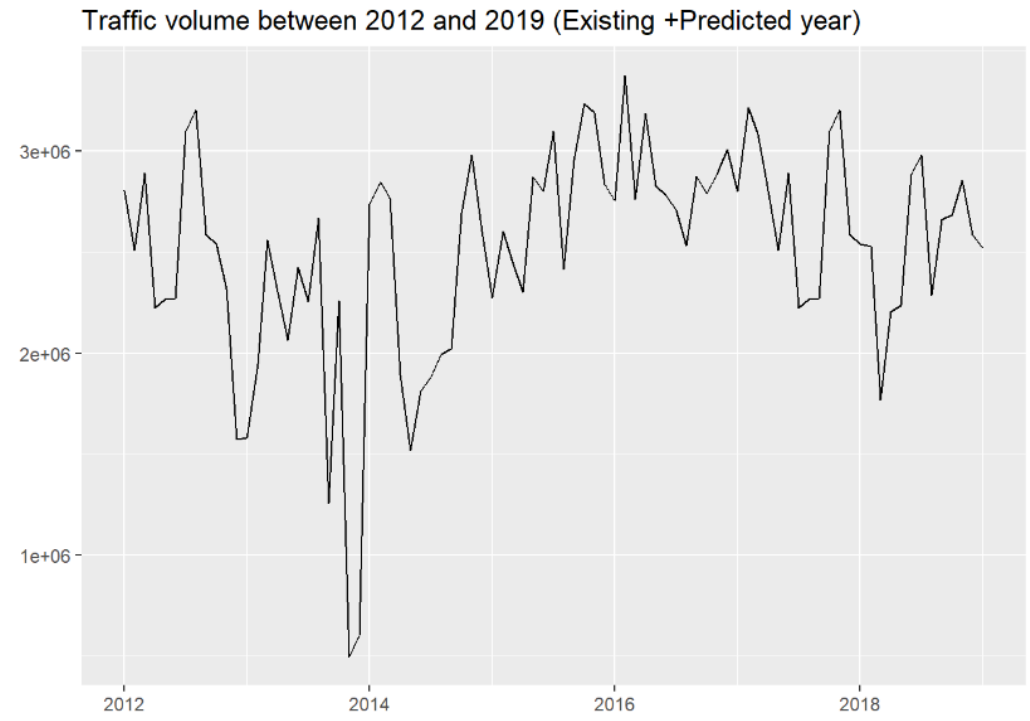
Prediction

Let's now move To Prediction for the next year of 2019:

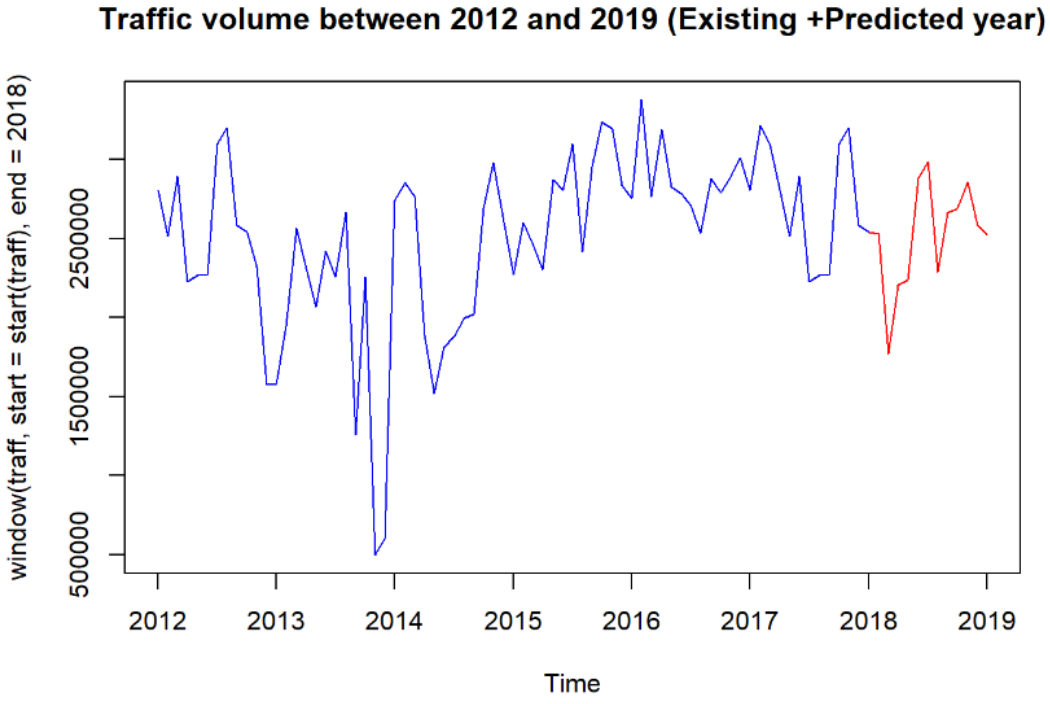
```
Here we see the predicted values of Traffic volume for the chosen one-year ra
nge
```



Here we can observe the traffic volume between the years 2012 and 2019 where 2019 values are the predicted one based on our VAR model

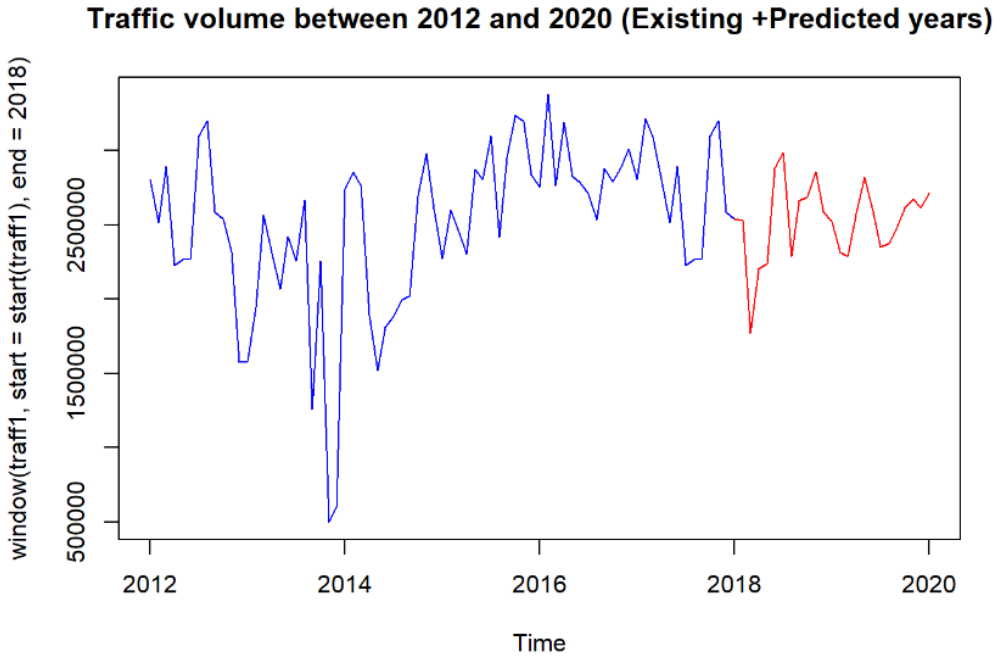


For better visualization we can trace in seperate colors



We can also predict the next 2 years and observe the results below:

```
Here we see the predicted values of Traffic volume for the chosen two years r
ange
```

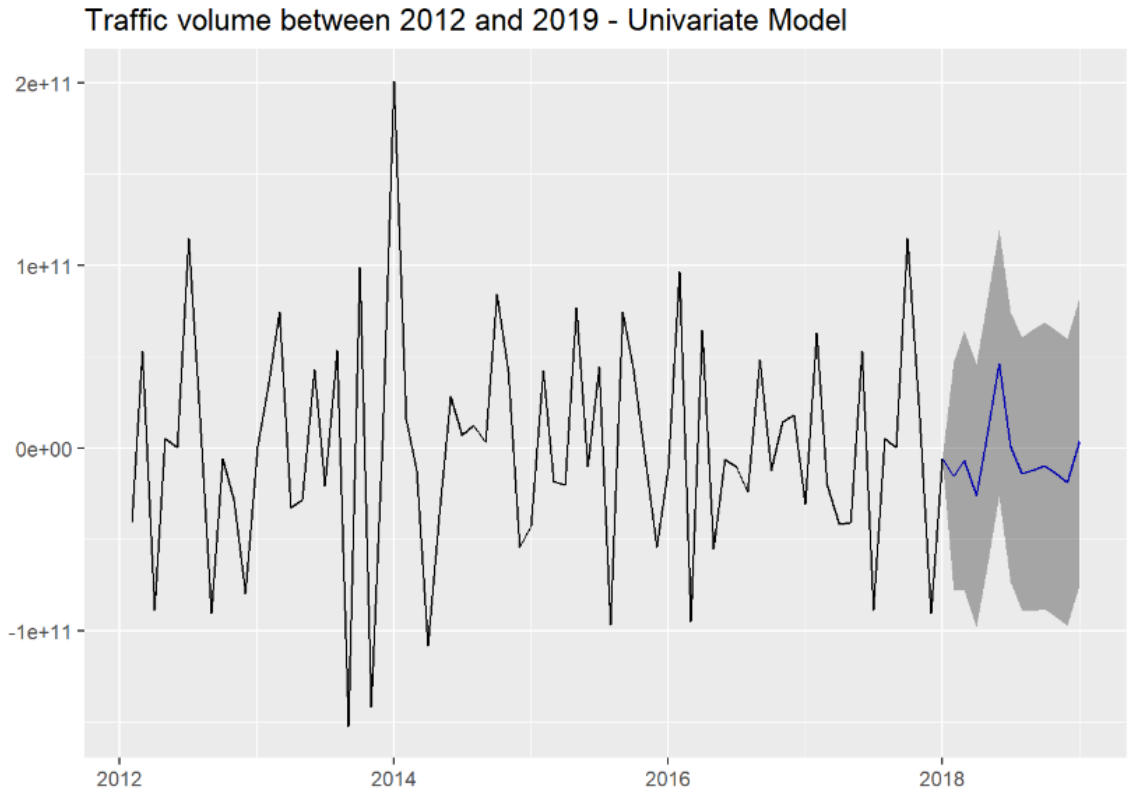


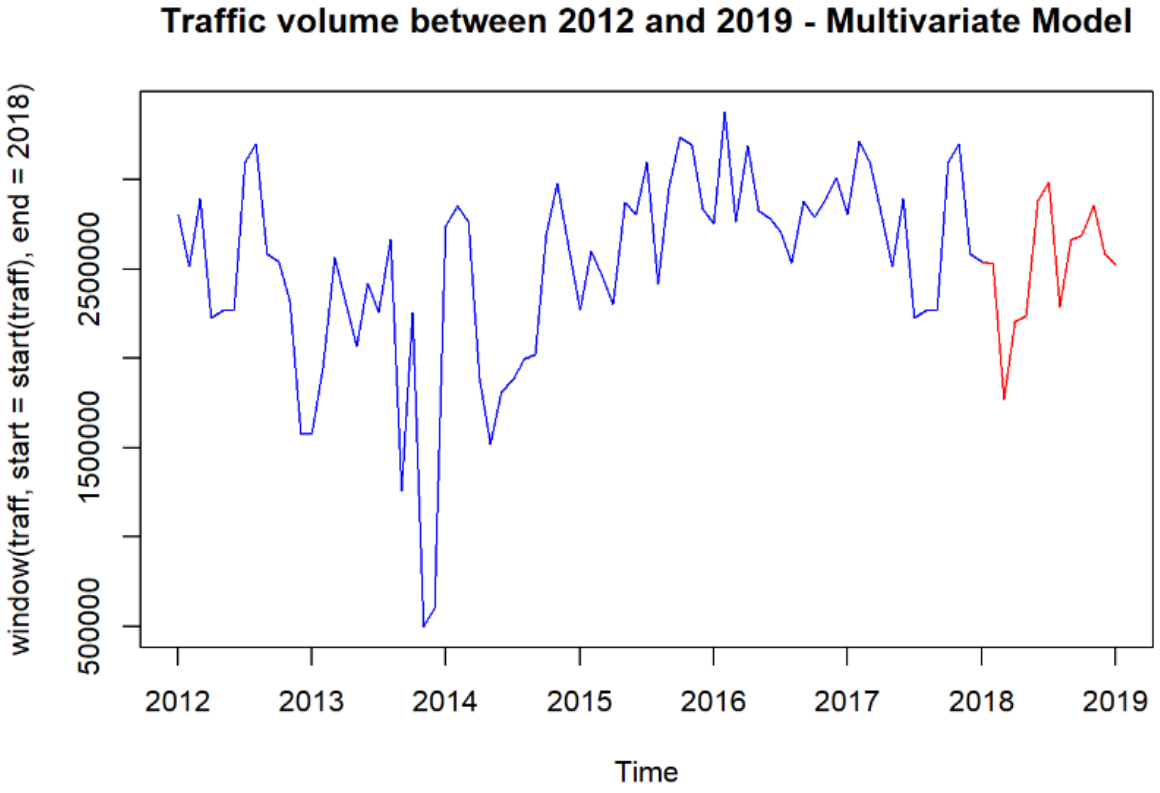
We can see that the quality of the prediction deteriorates when we increase the prediction period from 1 to 2 years.

Summary and conclusions

Observing the Traffic volume time series in hand, we found that it had no substantial seasonality which can contradict the assumption that traffic volume would present easily observable seasonality. However, throughout the analysis we can see how it is affected by (-in other words-correlate with) the holiday season, the temperature and overall weather condition. After analyzing the traffic volume time series' components, it is pre-processed to be smoothed and then the obtained treated time series is used to seek modelling (either univariate or multivariate) and of course prediction.

For the univariate modelling ARIMA(6,1,5) was chosen as the best model. We can trace one year prediction of this model and that of the multivariate model and compare:


Traffic volume between 2012 and 2019 - Univariate Model
2e+11
1e+11
0e+00
-1e+11
2012
2014
2016
2018


Traffic volume between 2012 and 2019 - Multivariate Model
window(traff, start = start(traff), end = 2018)
500000
1500000
2500000
2012
2013
2014
2015
2016
2017
2018
2019
Time

Observing the predictions from two best retained models, we can notice that the multivariate model shows rather better results than the simpler univariate one, as it seems to capture more of the variance of the traffic volume. As a result, the model obtained from the VAR modelling can be considered as the optimal one.

We must also note that with more prolonged data (larger data set) we can improve the performance of both models. Also, given the many irregularities in the present data set (especially the huge dip in the traffic volume around the end of 2014 that affected the trend), the obtained results can be considered as quite decent.

More models such as VARMA can also be explored in the future for testing better performing models. It is also important to note that SVAR and SARIMA models were not tested here, given the fact that the traffic volume data set was found to have no seasonality so it is rather meaningless to seek to build model that has one.