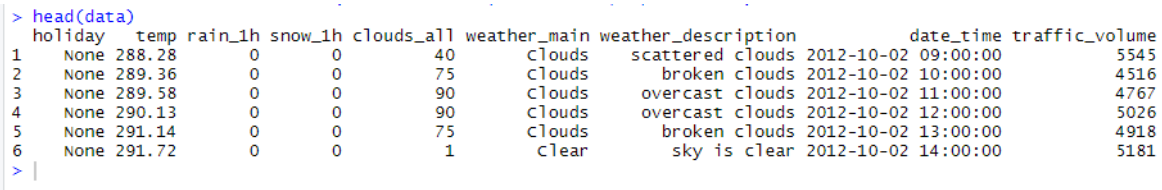
Nibedita Bal

Milestone 3: Data preparation and Exploratory Analysis

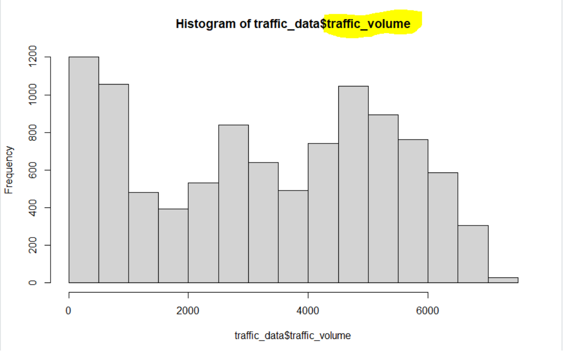
The Traffic Volume Data Set is collected from Kaggle and it contains 48,204 rows of hourly traffic data with a total of 9 attributes such as holiday type, temperature, rainfall, snow, clouds, weather, weather, date time hour of the data collected and traffic volume collected between October 2012 to January 2018. It is a headache for Minneapolis that during holiday seasons the traffic accidents on highways have been increasing. In this report, we will look up the traffic volume pattern and analyse the features carefully to see if the pattern is actually going up. In my individual part, I've used the first 10000 rows of observation.

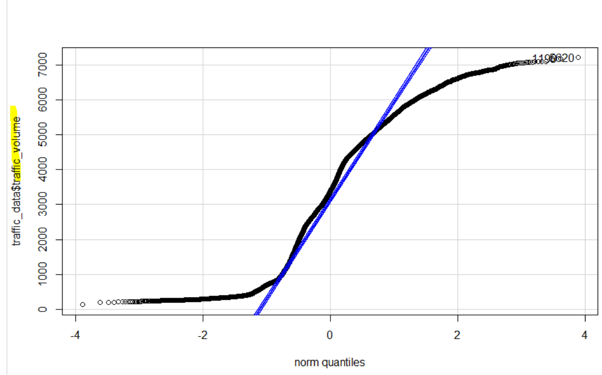
Just to get started below is the snapshot of features used:

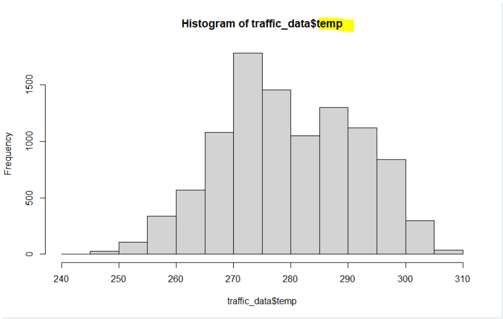


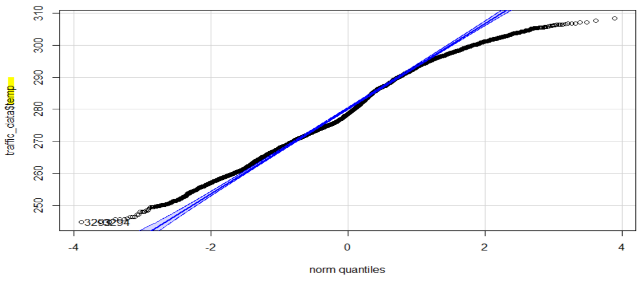
```
> head(data)
  holiday   temp rain_1h snow_1h clouds_all weather_main weather_description            date_time traffic_volume
1    None 288.28       0       0         40       Clouds    scattered clouds 2012-10-02 09:00:00           5545
2    None 289.36       0       0         75       Clouds       broken clouds 2012-10-02 10:00:00           4516
3    None 289.58       0       0         90       Clouds     overcast clouds 2012-10-02 11:00:00           4767
4    None 290.13       0       0         90       Clouds     overcast clouds 2012-10-02 12:00:00           5026
5    None 291.14       0       0         75       Clouds       broken clouds 2012-10-02 13:00:00           4918
6    None 291.72       0       0          1        Clear        sky is clear 2012-10-02 14:00:00           5181
> 
```

Histogram and QQ plot for traffic volume & temperature: The quantile plot for temperature does not look normal. Looks like there is a fatter tail. The bar of traffic volume is higher in the first 1000 data points and lower in the next 2000 data points.

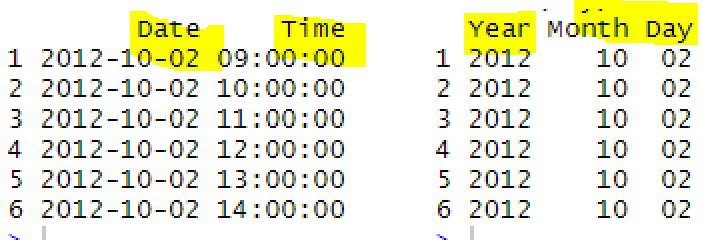






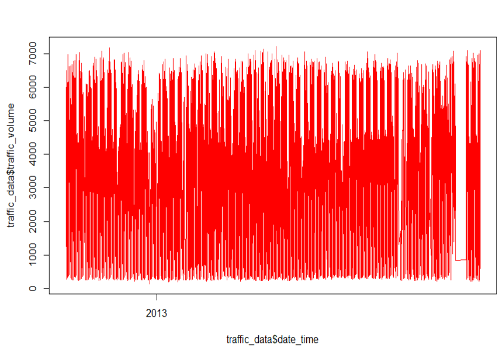


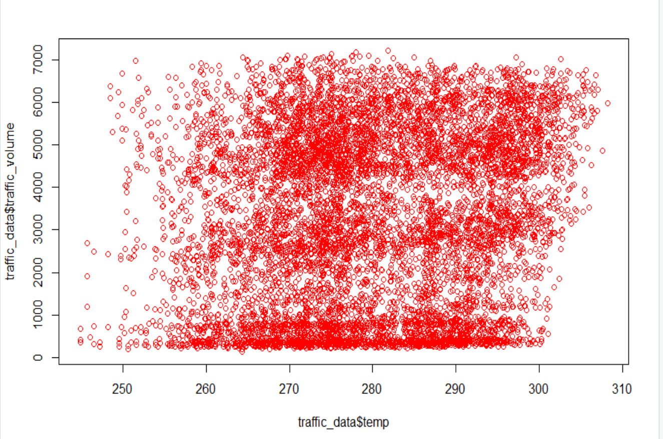
As a part of preprocessing, the date_time column has been splitted to date and time. Also, splitted date into year, month and day as below:



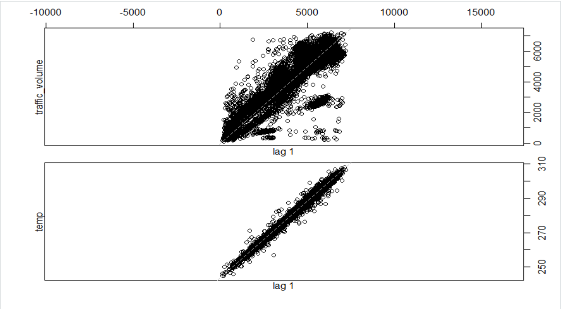
```
        Date     Time         Year Month Day
1 2012-10-02 09:00:00       1 2012    10  02
2 2012-10-02 10:00:00       2 2012    10  02
3 2012-10-02 11:00:00       3 2012    10  02
4 2012-10-02 12:00:00       4 2012    10  02
5 2012-10-02 13:00:00       5 2012    10  02
6 2012-10-02 14:00:00       6 2012    10  02
```

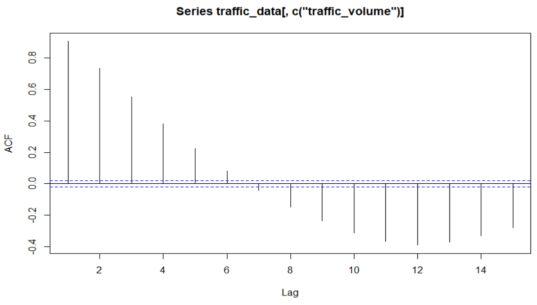
Generated traffic volume flow for a particular period of years:It seems to mean revert during the entire periods and the variance is same for traffic volume.





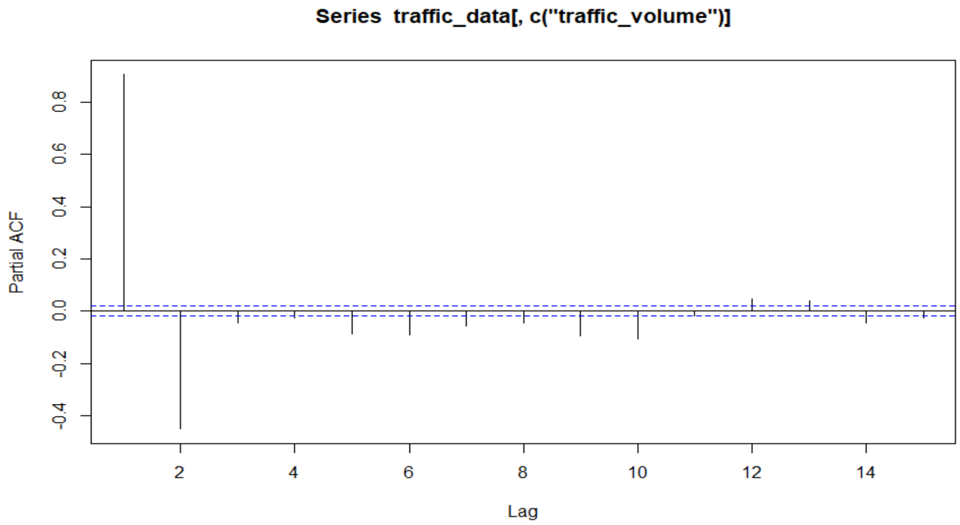
Acf plot of traffic volume and temperature: we don't see serial autocorrelation at lag 5.We see a very strong positive correlation between the temperature though.



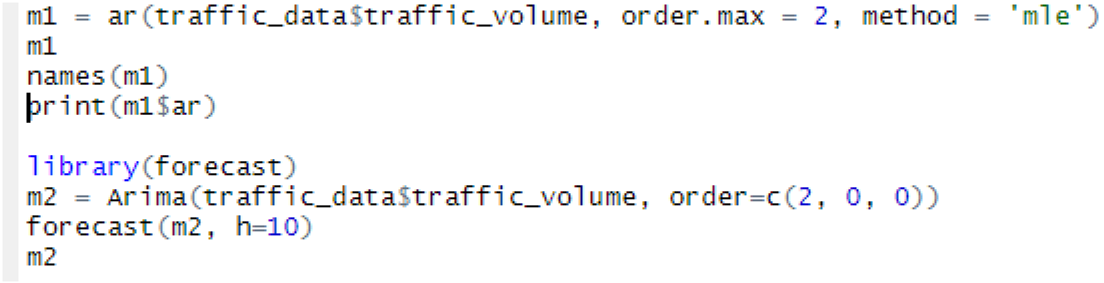


Box.test(traffic_data[,c('traffic_volume')], lag=5, type = "Ljung-Box")
The P-value <0.05 . So we could reject the null hypothesis that there is no autocorrelation.

Pacf: I see significant serial autocorrelation at point 0.01.



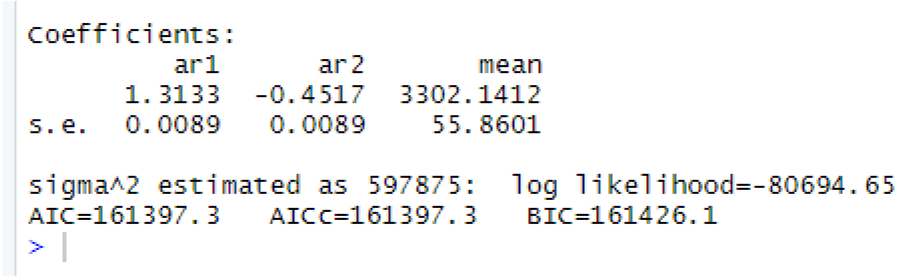
ARIMA:



```
m1 = ar(traffic_data$traffic_volume, order.max = 2, method = 'mle')
m1
names(m1)
print(m1$ar)

library(forecast)
m2 = Arima(traffic_data$traffic_volume, order=c(2, 0, 0))
forecast(m2, h=10)
m2
```

Output:



```
Coefficients:
         ar1      ar2       mean
      1.3133  -0.4517  3302.1412
s.e.  0.0089   0.0089    55.8601

sigma^2 estimated as 597875:  log likelihood=-80694.65
AIC=161397.3   AICc=161397.3   BIC=161426.1
> 
```