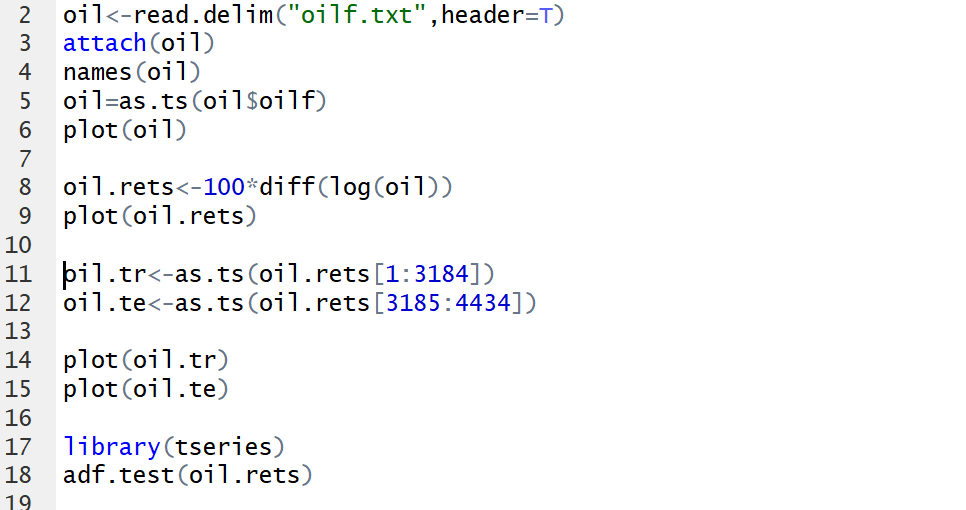
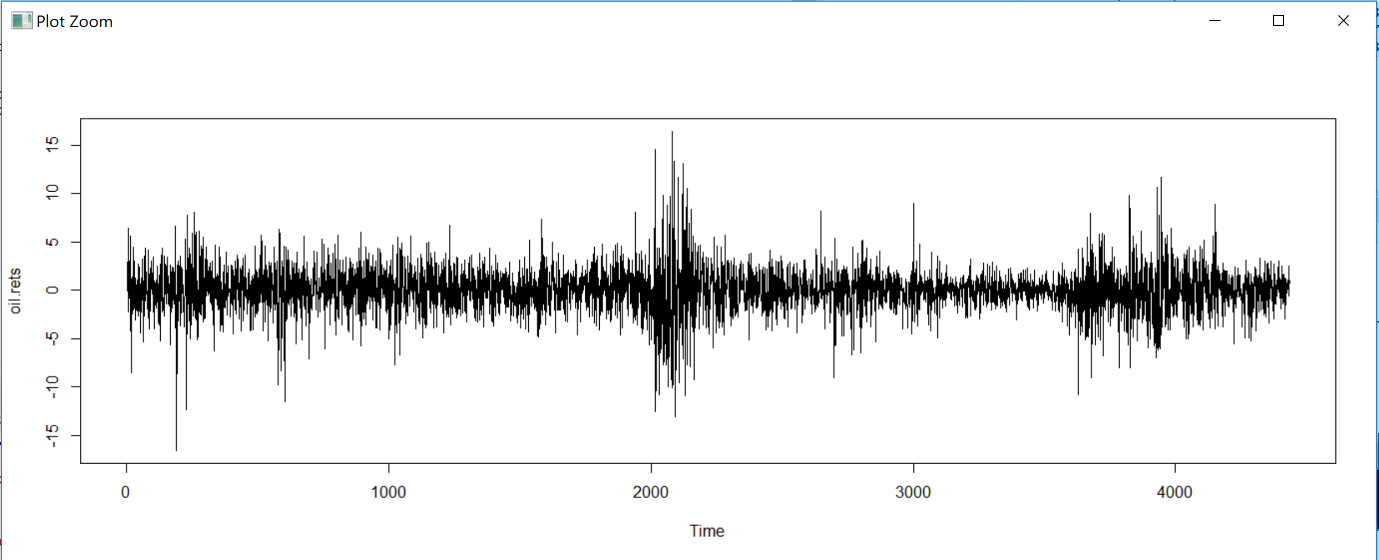
**Qa(i)**

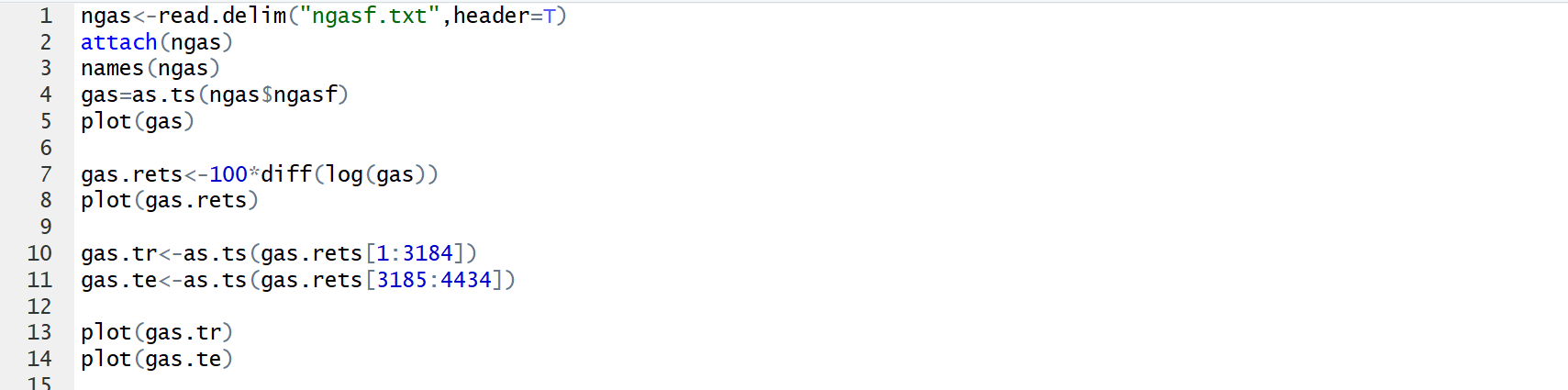
**Data Preparation**

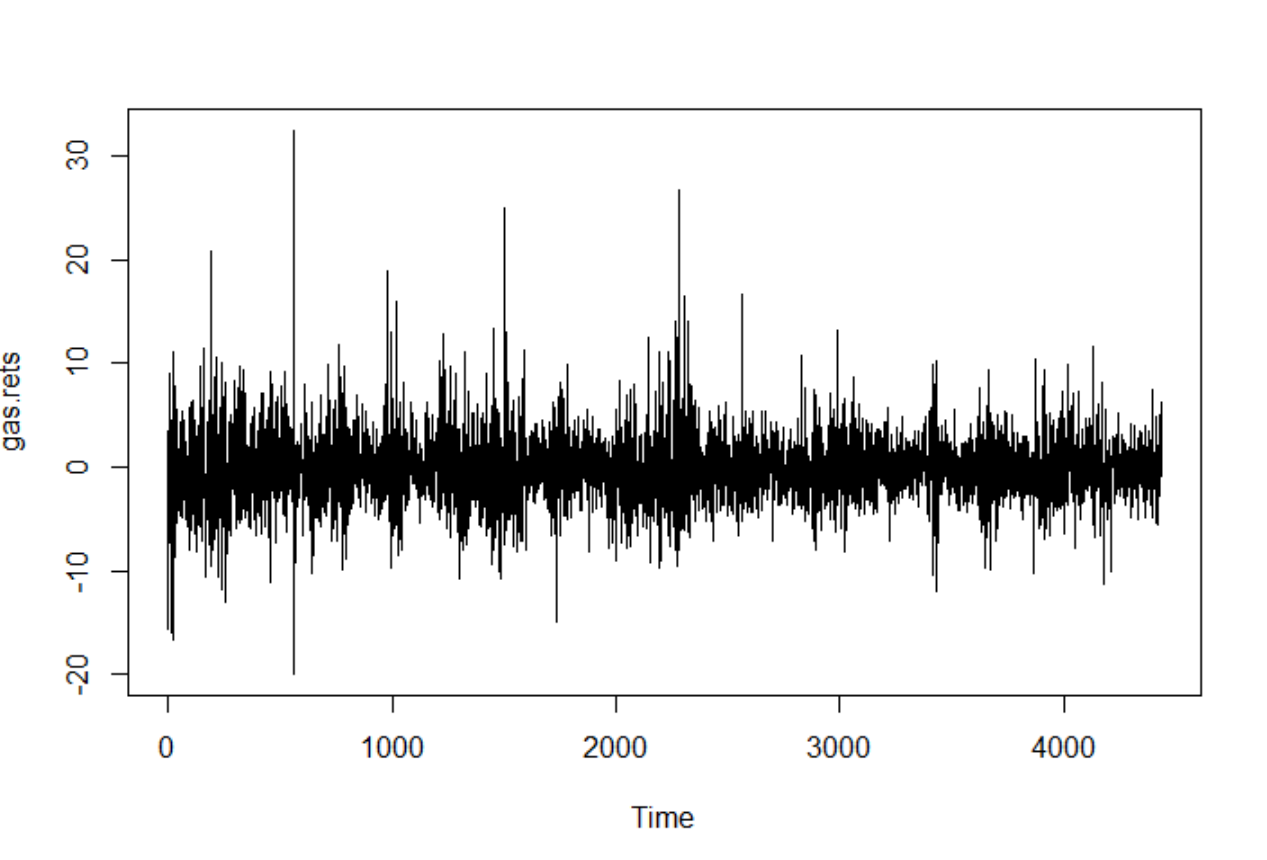
Oilf(Data Preparation)



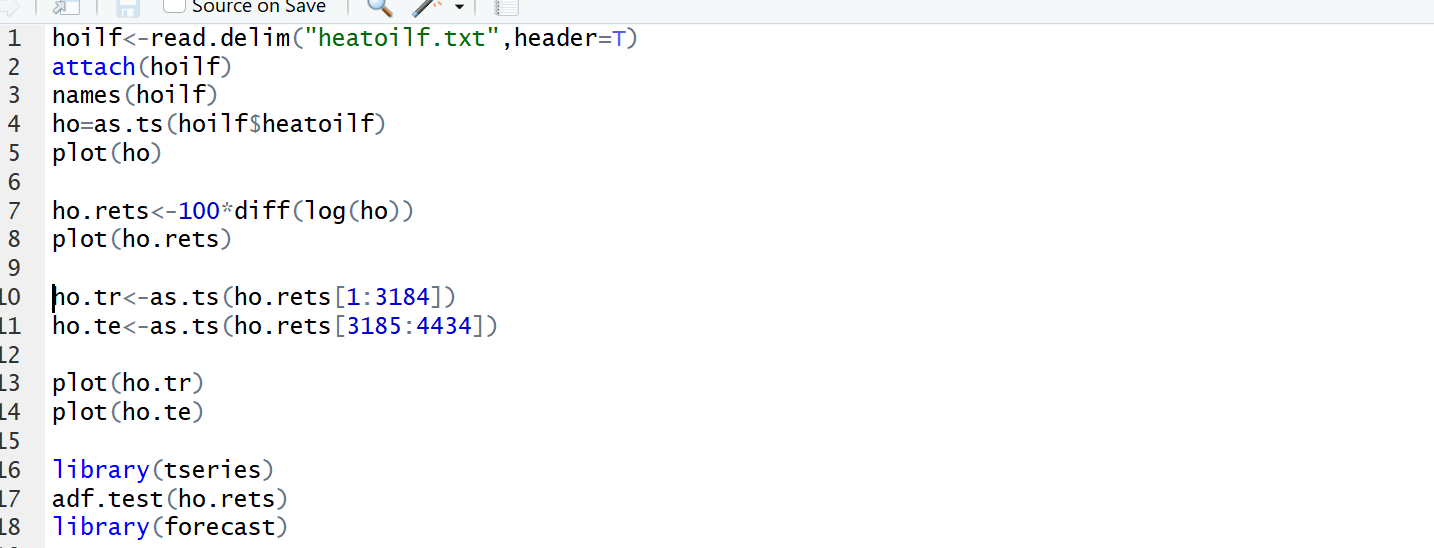


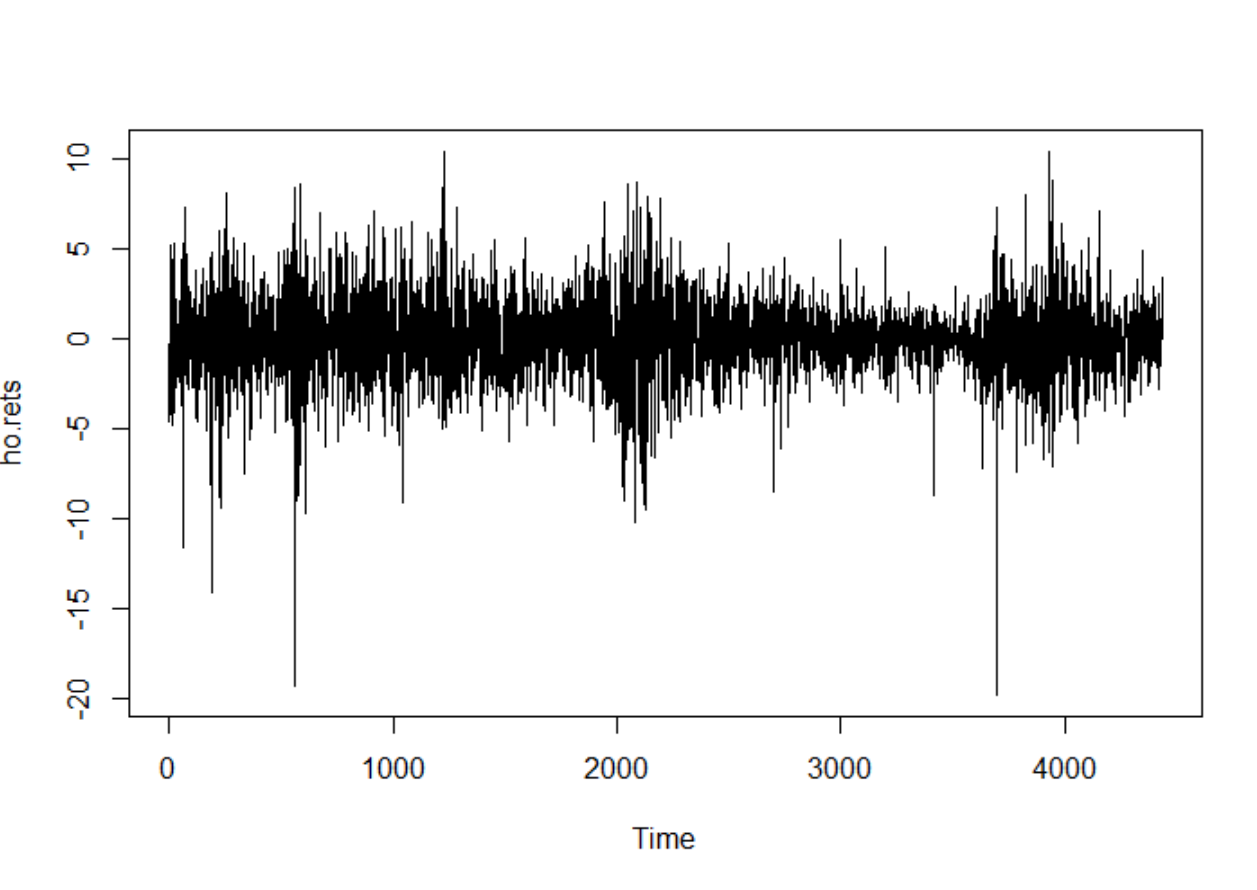
Ngasf data preparation





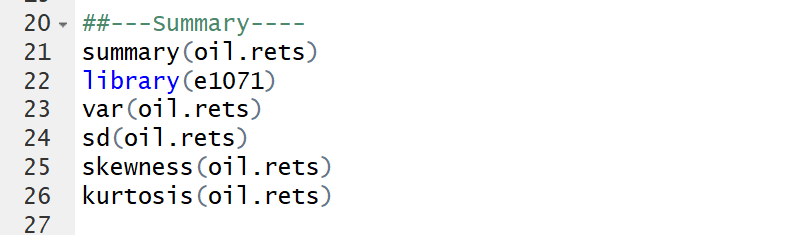
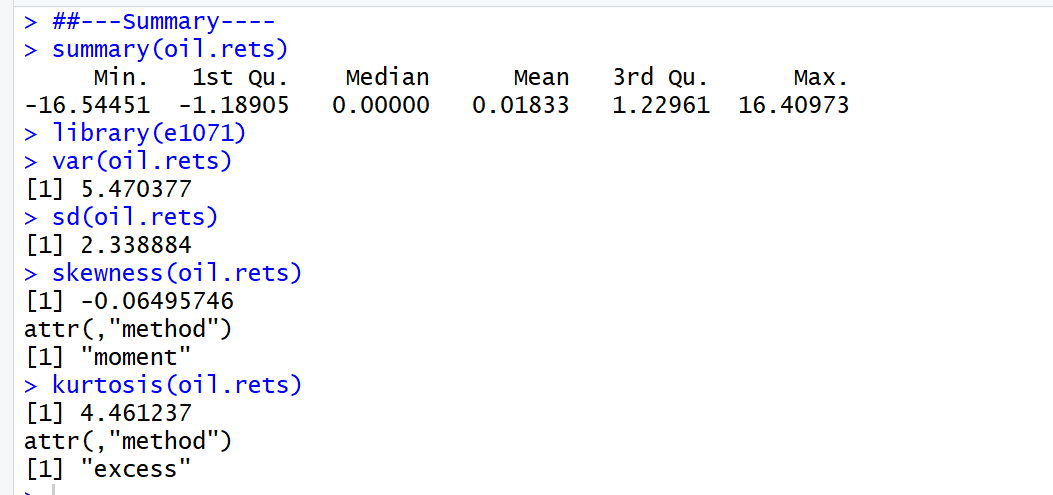
HeatOilf data preparation



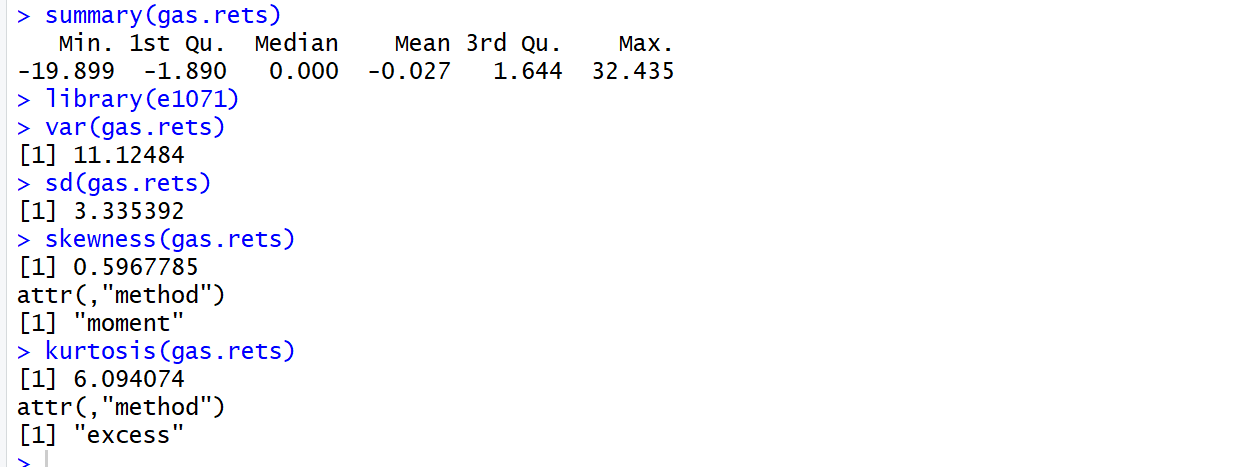


**Summary Stats**

Oil

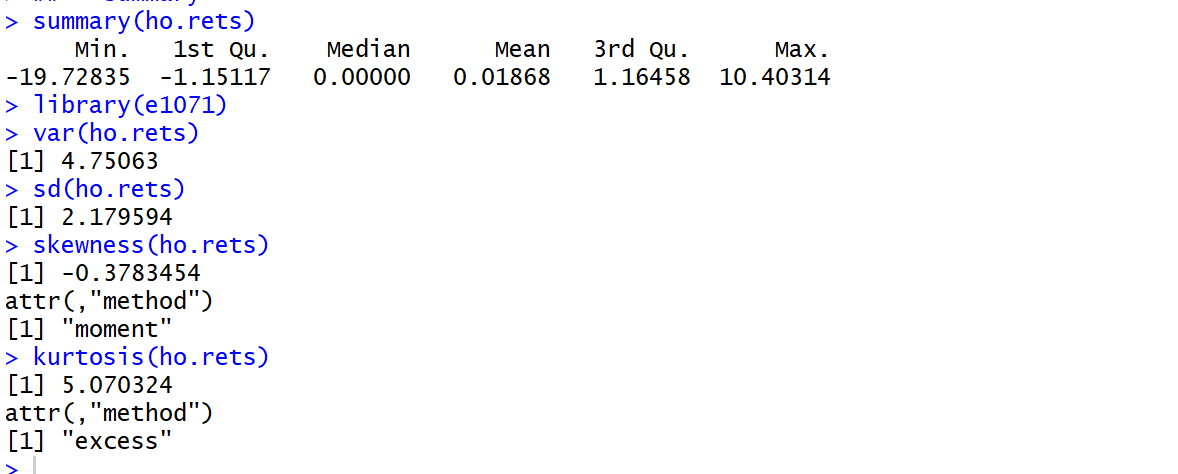


Gas

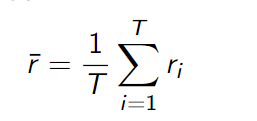


HeatOil





**Mean**

Sample mean captures central tendency of returns, if its value is positive, it shows that the investors have positive returns over the period on average. Whereas, if the sample mean value is negative, it shows that the investors have negative returns over the period on average. which is defined by

Oil: Sample mean of 0.01833 indicates that on average, daily return on oil price increased over the time by 0.01833%.

The average annual rate of return; 0.01833×261=4.78413%

Gas: Value is -0.027% that indicates on average, daily return on natural gas prices tend to be decreasing by 0.027% over the period.

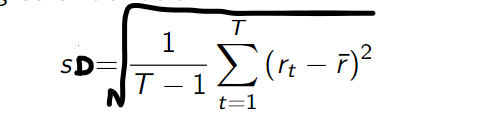
The average annual rate of return; -0.027×261=-7.047%

HeatOil: Value is 0.01868 that implies on average, daily return on heatoil prices are likely to increase by 0.01868 over that period.

The average annual rate of return; 0.01868×261=4.87548%

As conclusion, from 1st of January 2001 to 29th of December 2017, when we only look at mean, heating oil shows highest return, then Investors might prefer this.

**Standard Deviation**

The sample standard deviation captures the average distance away from the mean. This is used for risk measure. If the value of standard deviation, it implies that the return on that investment is very unstable (fluctuate a lot). Investors does prefer smaller value of it.

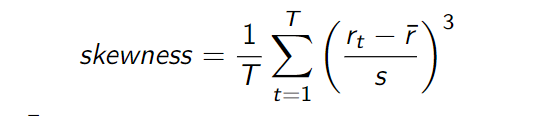
Oil: The daily return on oil price is deviated from the mean by 2.338884% on average over the time.

Gas: The daily return on natural gas price is deviated from the mean by 3.335392% on average during that time.

HeatOil: The daily return on heating oil prices is deviated from the mean by 2.179594% on average during that time.

As conclusion, from 1st of January 2001 to 29th of December 2017, investing on heating oil might be less risky than the other two.

**Skewness**

Investors are interested in whether the mean value are more likely to be negatively distributed or positively distributed too. The skewness measures the amount of asymmetry in a distribution. The larger the absolute size of the skewness, the more asymmetrical is the distribution. Skewness of a normal distribution is equal to 0.

Oil: The value is -0.064957 that is positive skewness. Investors expect frequent small gains and a few large losses.

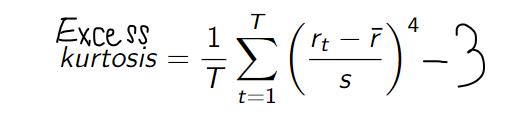
Gas: The value is 0.596 that is positive skewness. Investors would expect frequent small losses and few large gains.

HeatOil: The value is -0.3783 that is negative skewness. Investors would expect frequent small gains and a few large losses.

As conclusion, from 1st of January 2001 to 29th of December 2017, heating oil log returns are the most negatively skewed in this sample which indicates the investors can expect frequent small gains and a few large losses from their investment. Whereas for natural log returns, investors can expect frequent small losses and a few gains from their investment.

**Excess Kurtosis**

It tells us if there is a high possibility of extreme positive or negative excess returns, regardless of sign. It measures the thickness of the tail of a distribution. Investors can expect occasional extreme returns (positive or negative) if it has high kurtosis. If the calculated excess kurtosis is bigger than 0, it is considered as high kurtosis. If it is lesser than 0, it is considered as low kurtosis.

****

Oil: The value of 4.461237 implies that the high possibility of big positive and negative return realizations.

Gas: The value of 6.094 indicates that there might be a high likelihood of getting big positive or negative return realization.

HeatOil: The value of 5.0703 indicates that there might be a high likelihood of getting big positive or negative return realization.

If investors are looking for safety investment, they can engage in oil from 1st of January 2001 to 29th of December 2017

**Minimum**

It just tells us the minimum rate of return from 1st of January 2001 to 29th of December 2017

Oil: Minimum rate of return is -16.54451%

Gas: Minimum rate of return is -19.899%

HeatOil: Minimum rate of return is -19.7283%

**Maximum**

It tells us the maximum rate of return from 1st of January 2001 to 29th of December 2017

Oil: The maximum rate is 16.40973%

Gas: The maximum rate is 32.435%

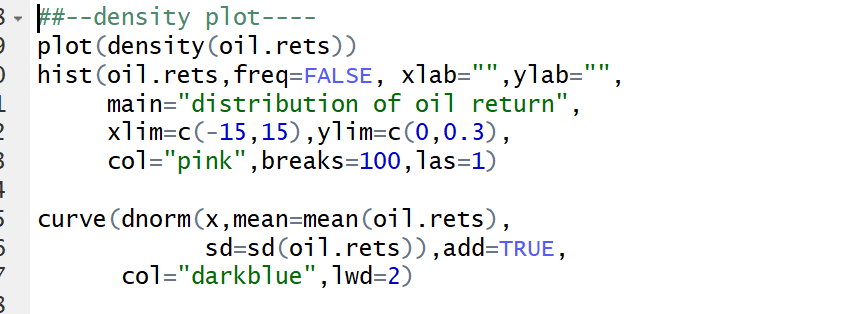
HeatOil: The maximum rate is 10.40314.

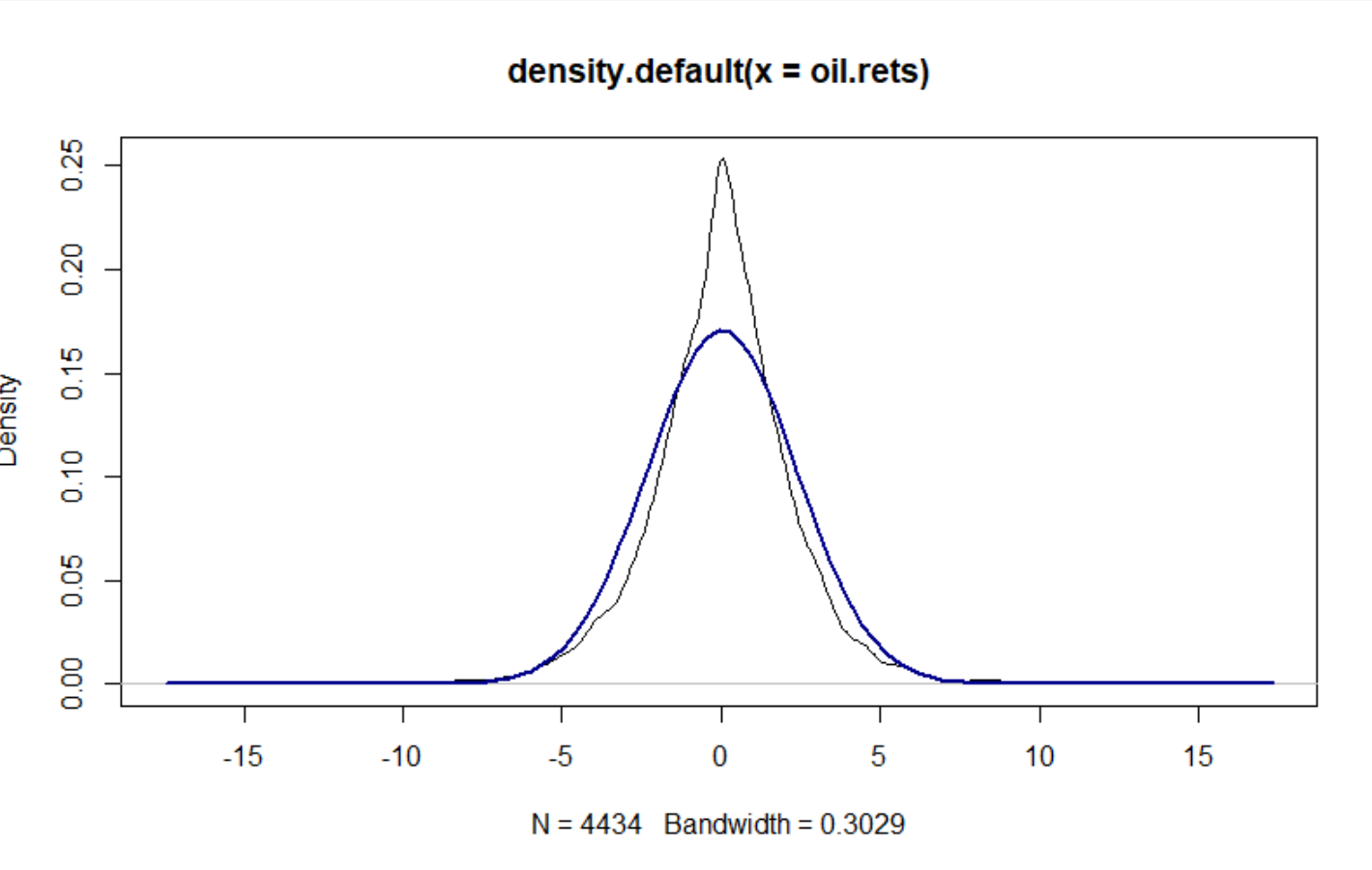
As, overall conclusion based on summary statisitcs, crude oil might be preferable, since it has negative skewness, smaller standard deviation, mean of return is positive and excess kurtosis is smaller.

Qa(ii)

**Oil**

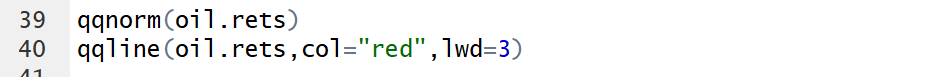
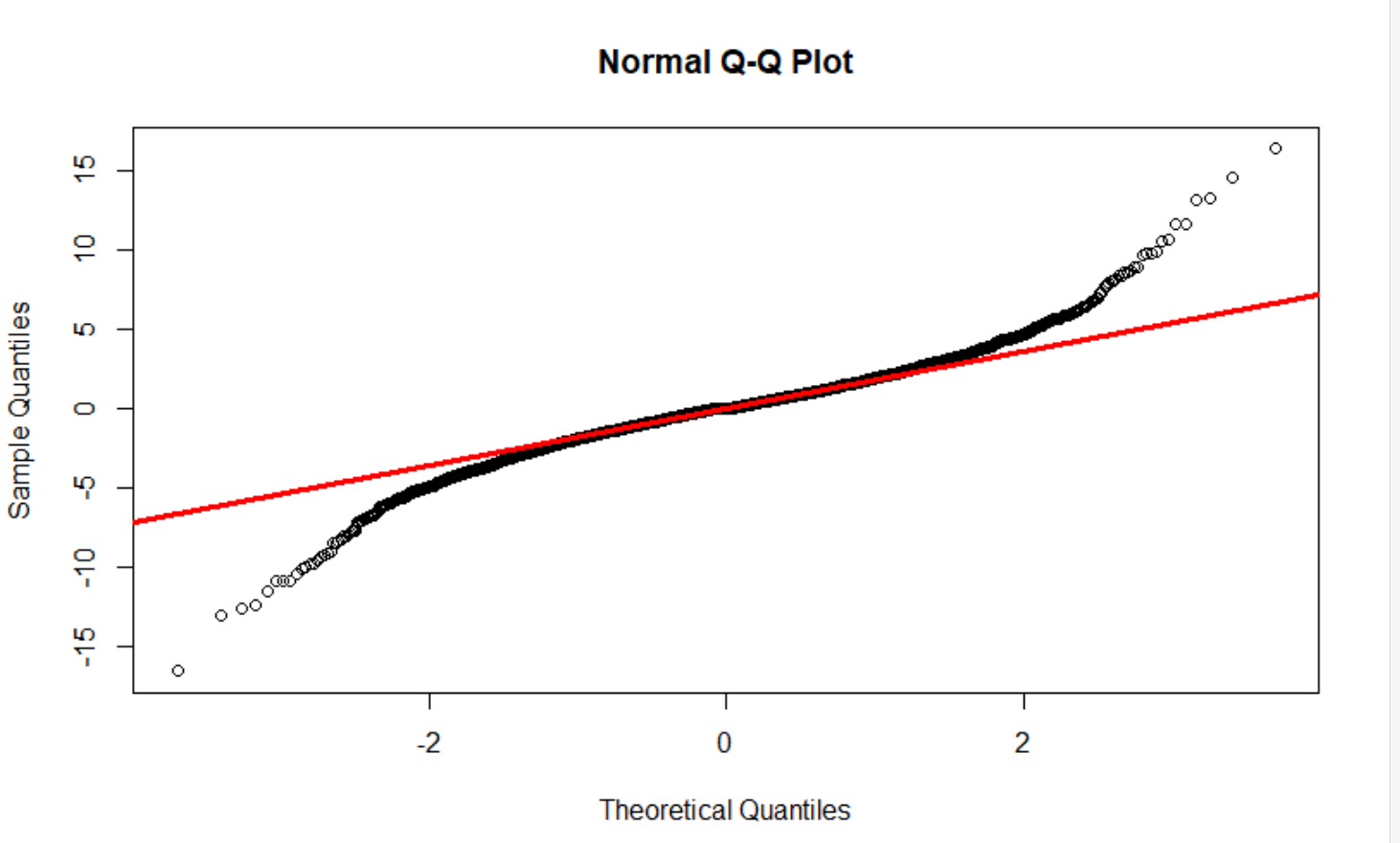
Density Plot





Based on the density plot compared with theoretical distribution, the oil log return does not seem to follow theoretical distribution.

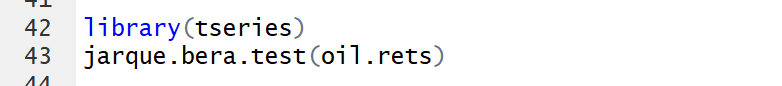
QQ Plot

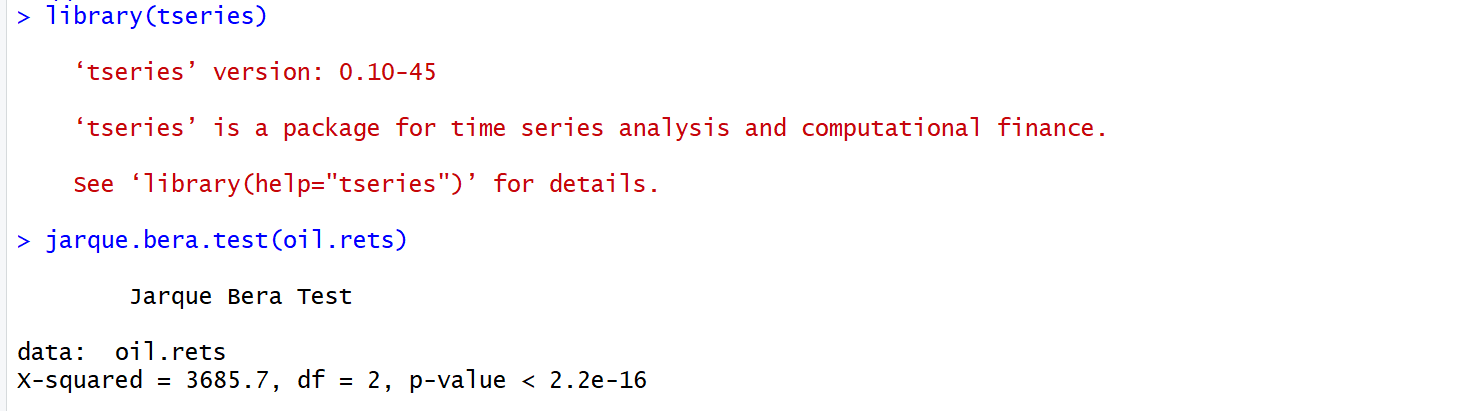


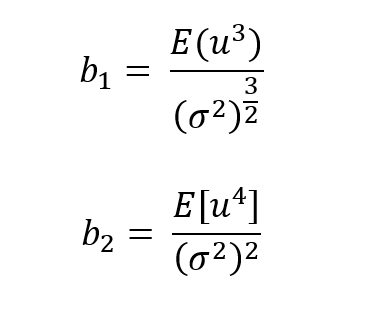
Based on QQ plot, again it might not follow normal distribution, since each data points are far from redline.

To ensure, we could perform jarque bera test.

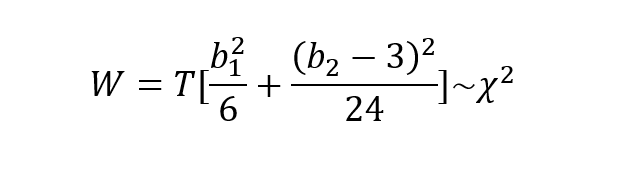
Jarque Bera Test





The coefficient can be expressed by:

The value of Jarque Bera statistics is defined by:



Null Hypothesis: Jointly skewness and kurtosis are Zero

Alternative Hypothesis: at least either skewness or Kurtosis is not Zero

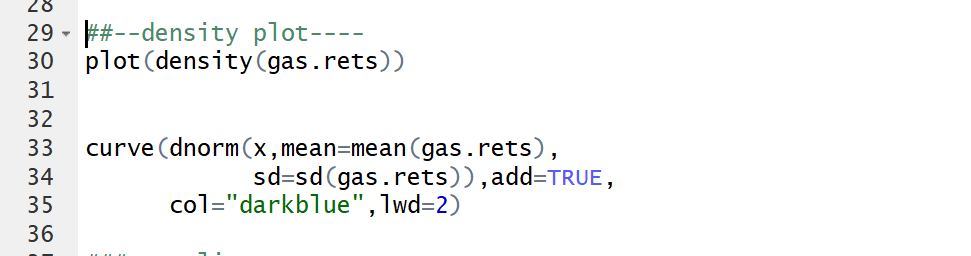
Tstats=3685.7

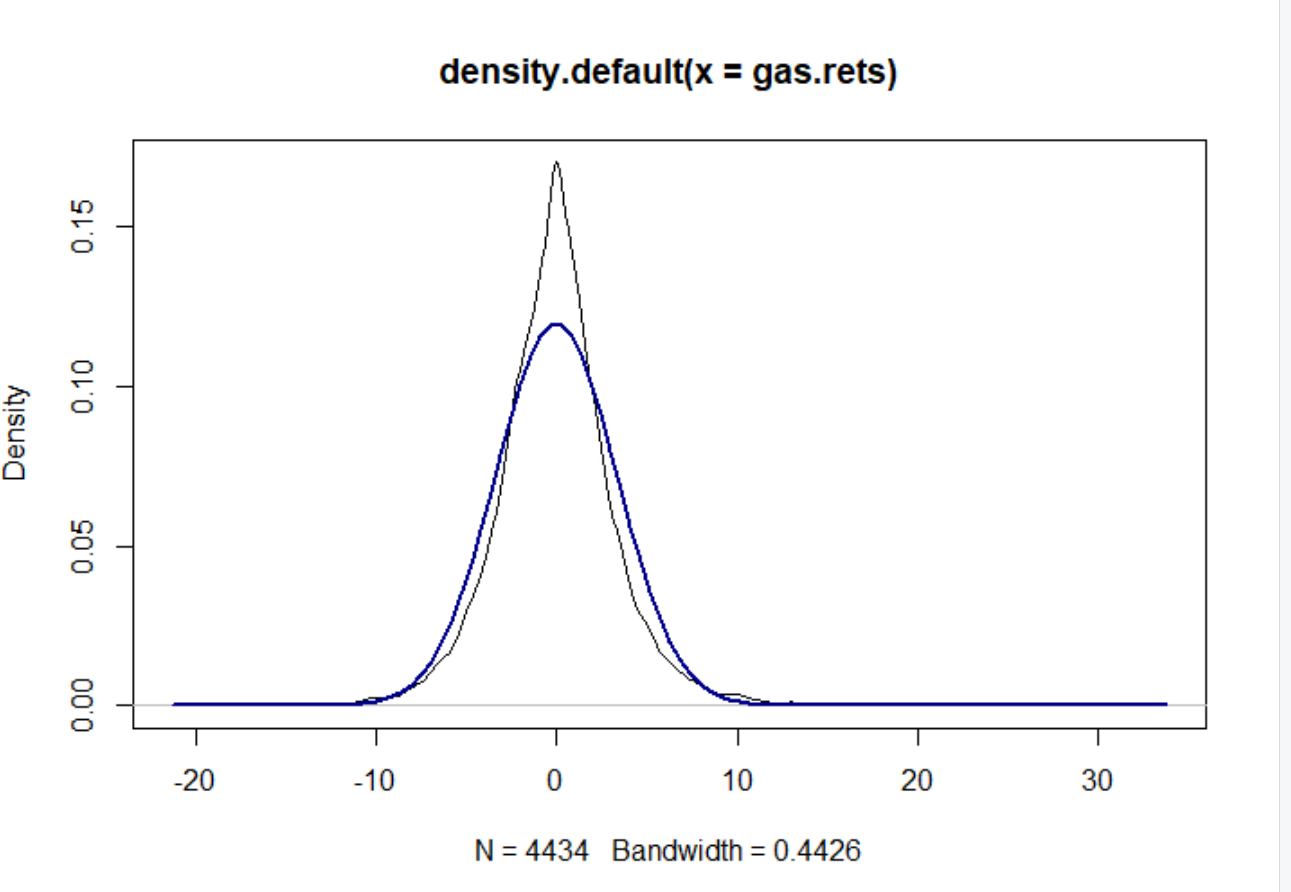
P-value=2.2e-16

Therefore, we conclude that there is enough evidence to reject the null hypothesis, and conclude that oil log return data is not normally distributed.

**Gas**

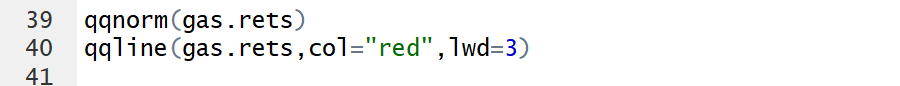
Density Plot

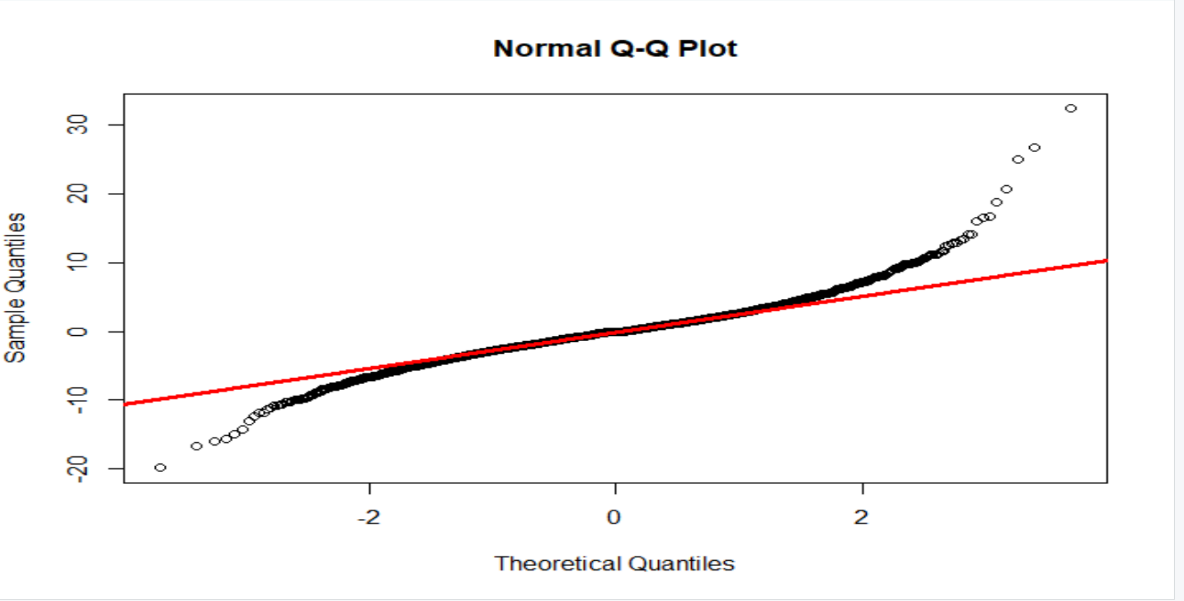




Based on the density plot it would not follow theoretical distribution.

QQplot

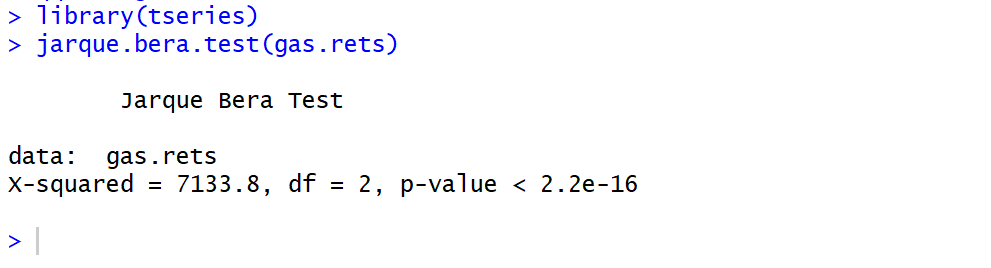


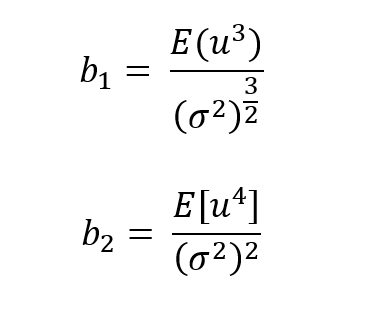


Again, each data points are far from redline. The distribution of log return on natural gas might not follow normal distribution.

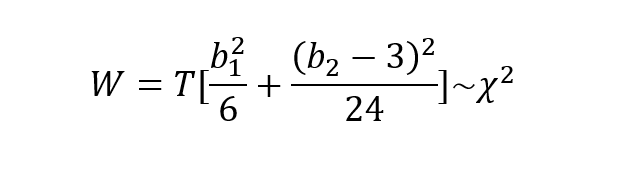
Jarque-Bera





The coefficient can be expressed by:

The value of Jarque Bera statistics is defined by:



Null Hypothesis: Jointly skewness and kurtosis are Zero

Alternative Hypothesis: at least either skewness or Kurtosis is not Zero

Tstats=7133.8

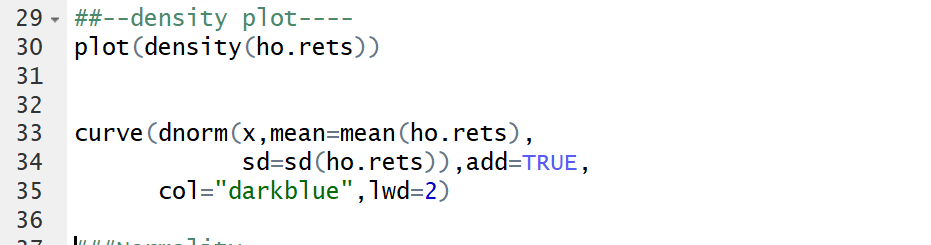
P-value=2.2e-16

Therefore, we conclude that there is enough evidence to reject the null hypothesis and conclude that natural gas log return data is not normally distributed.

.

**HeatOil**

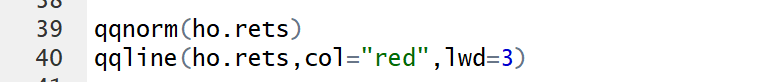
Denstity Pot

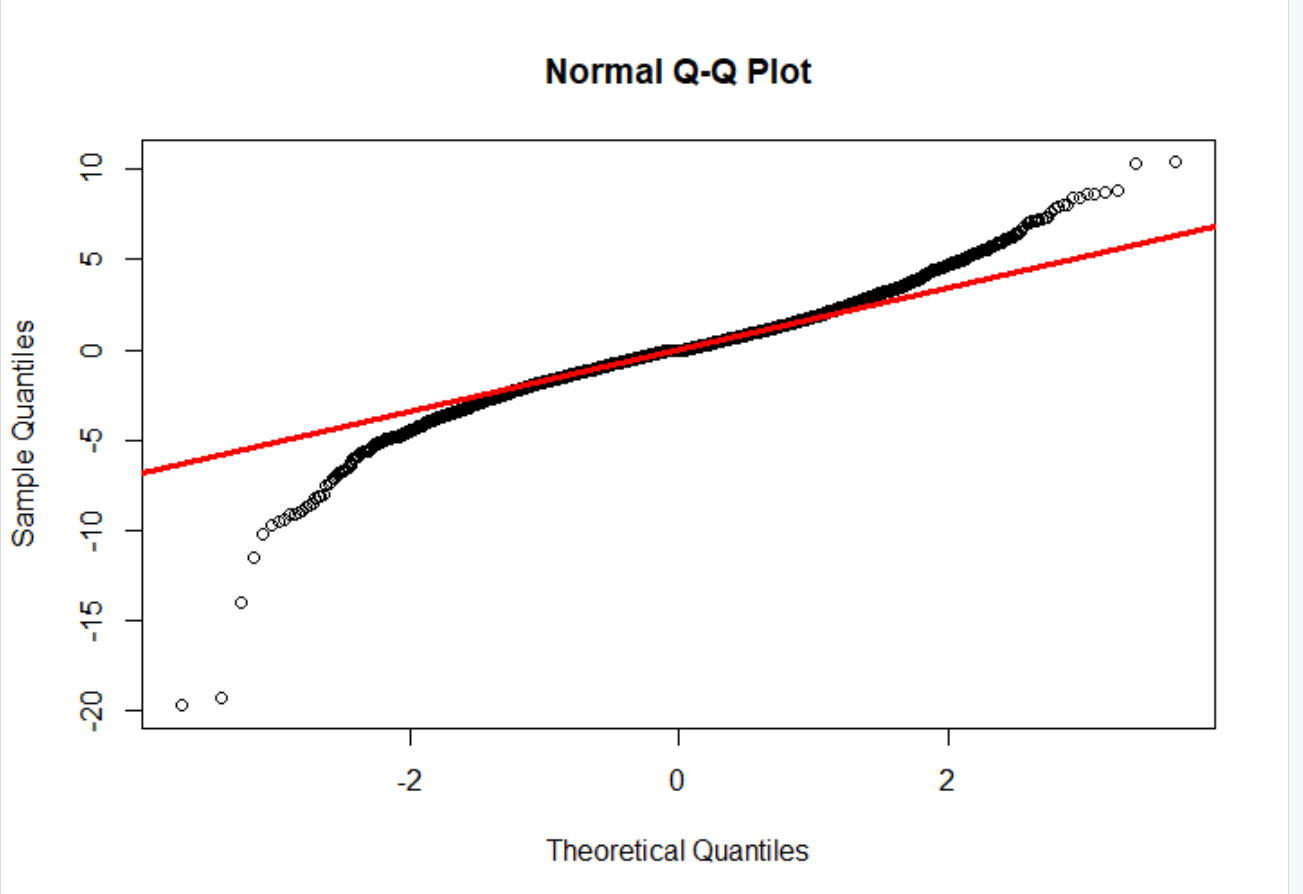
****

****

Based on the density plot, it would not follow theoretical distribution.

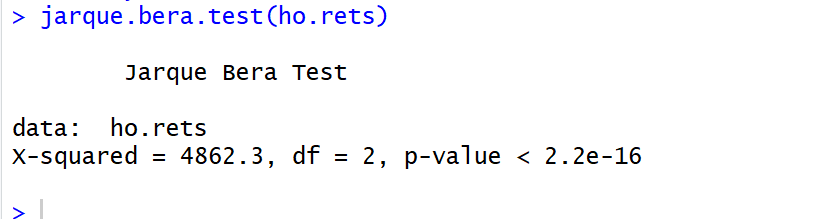
QQ-Plot

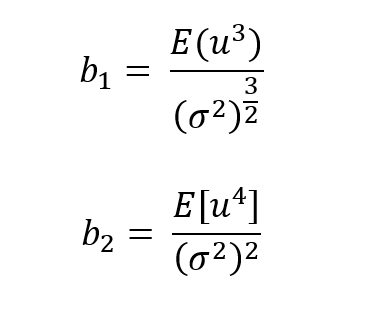




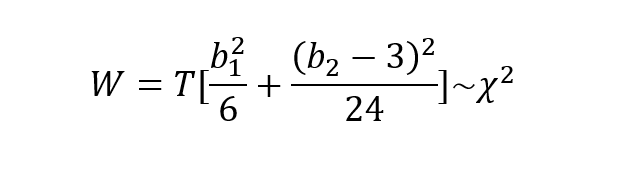
Again, each data points are far from redline. The distribution of log return on heating oil might not follow normal distribution.





The coefficient can be expressed by:

The value of Jarque Bera statistics is defined by:



Null Hypothesis: Jointly skewness and kurtosis are Zero

Alternative Hypothesis: at least either skewness or Kurtosis is not Zero

Tstats=4862.3

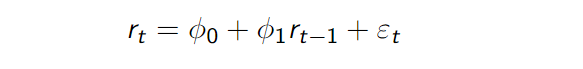
P-value=2.2e-16

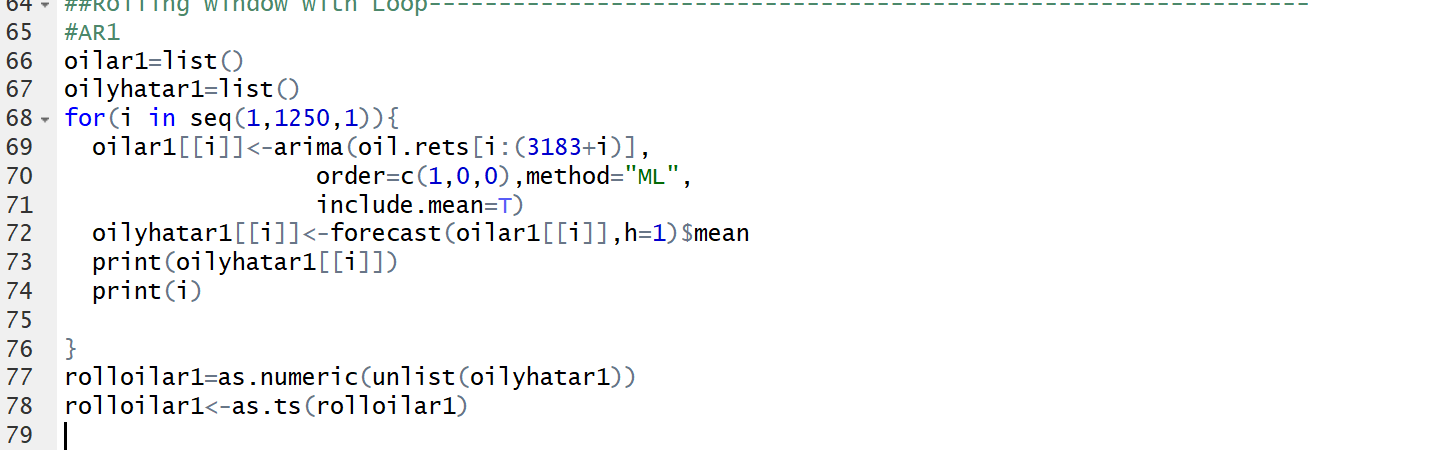
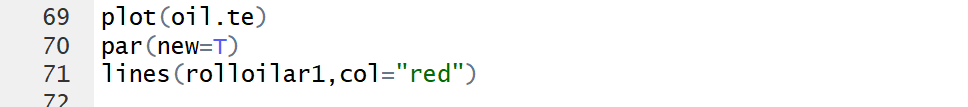
Therefore, we conclude that there is enough evidence to reject the null hypothesis and conclude that heating oil log return data is not normally distributed.

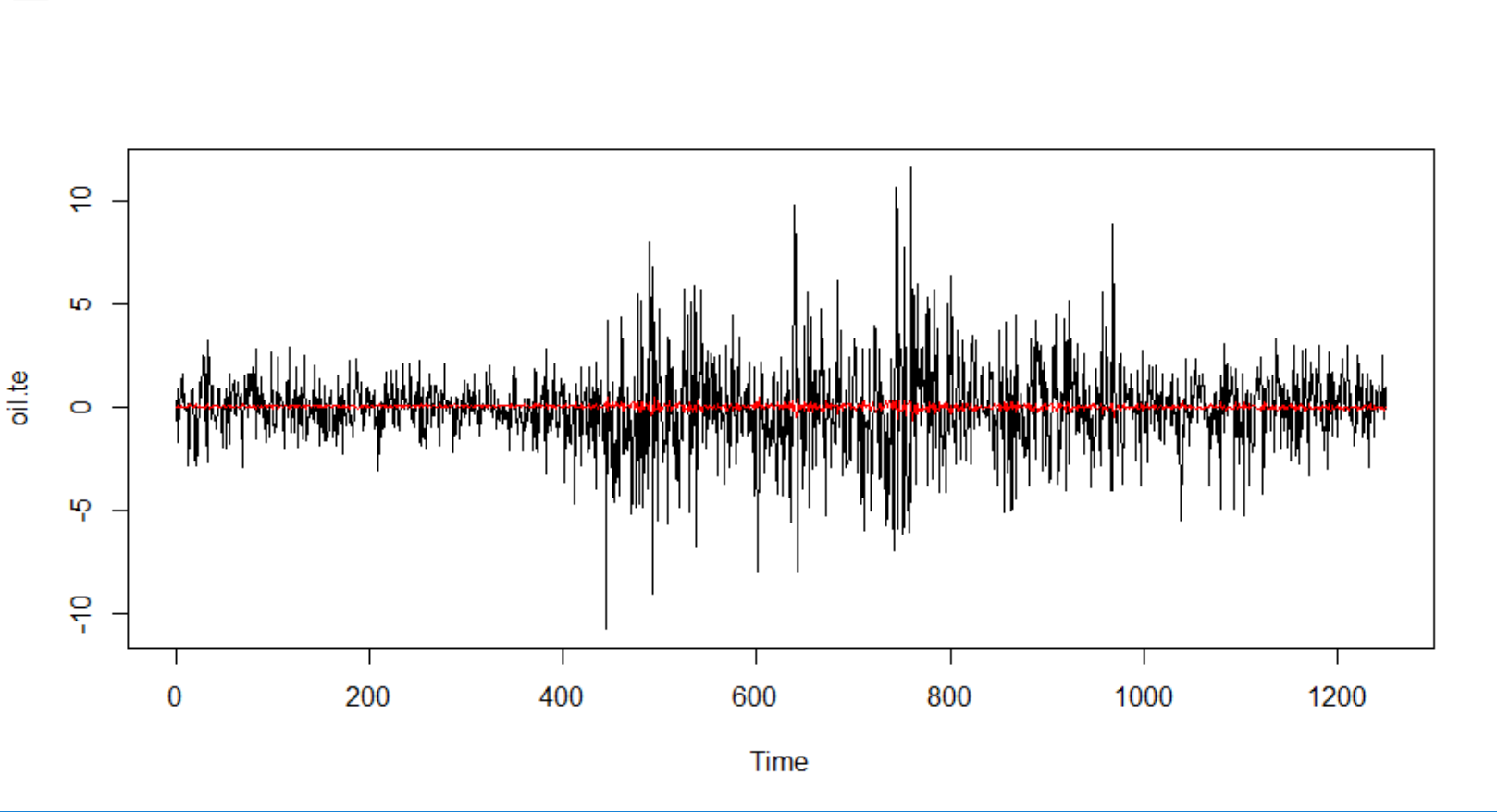
As a conclusion, all three data set that we are handling are not normally distributed.

**Qb(i)**

**Oil**

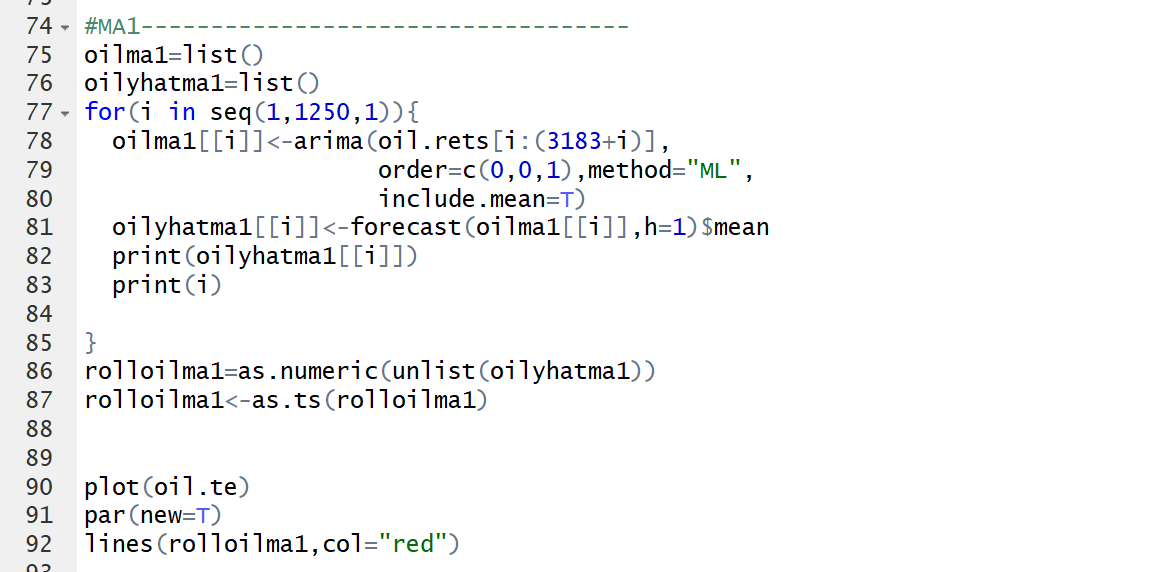
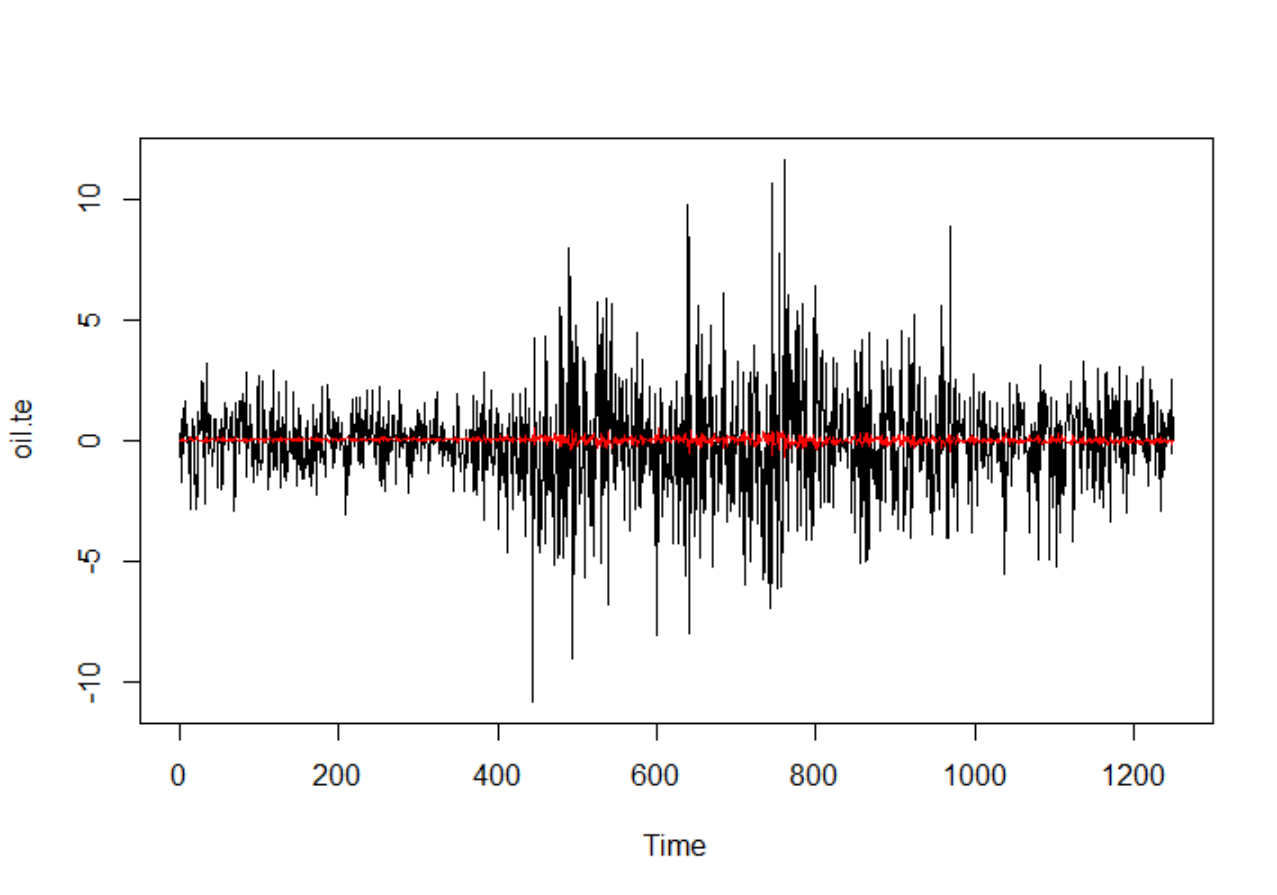
AR1: An AR(1) autoregressive process is the first order process, meaning that the current value is based on the immediately preceding value.

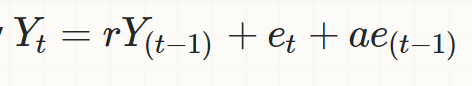




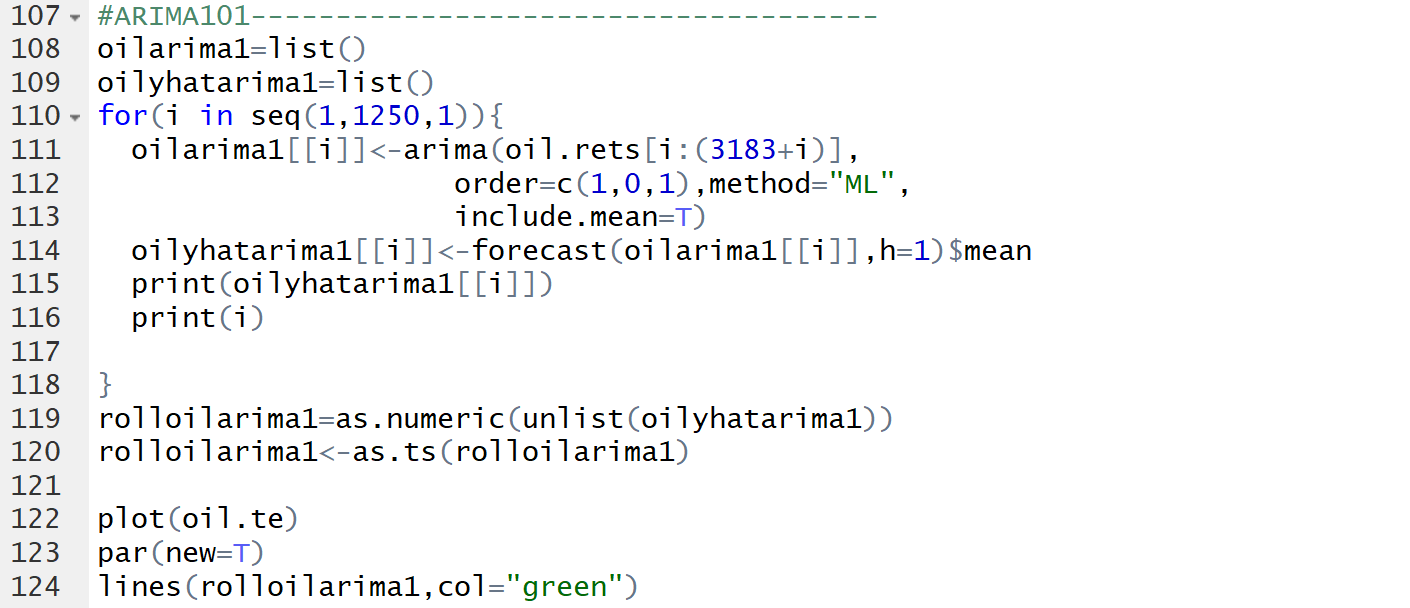
MA1

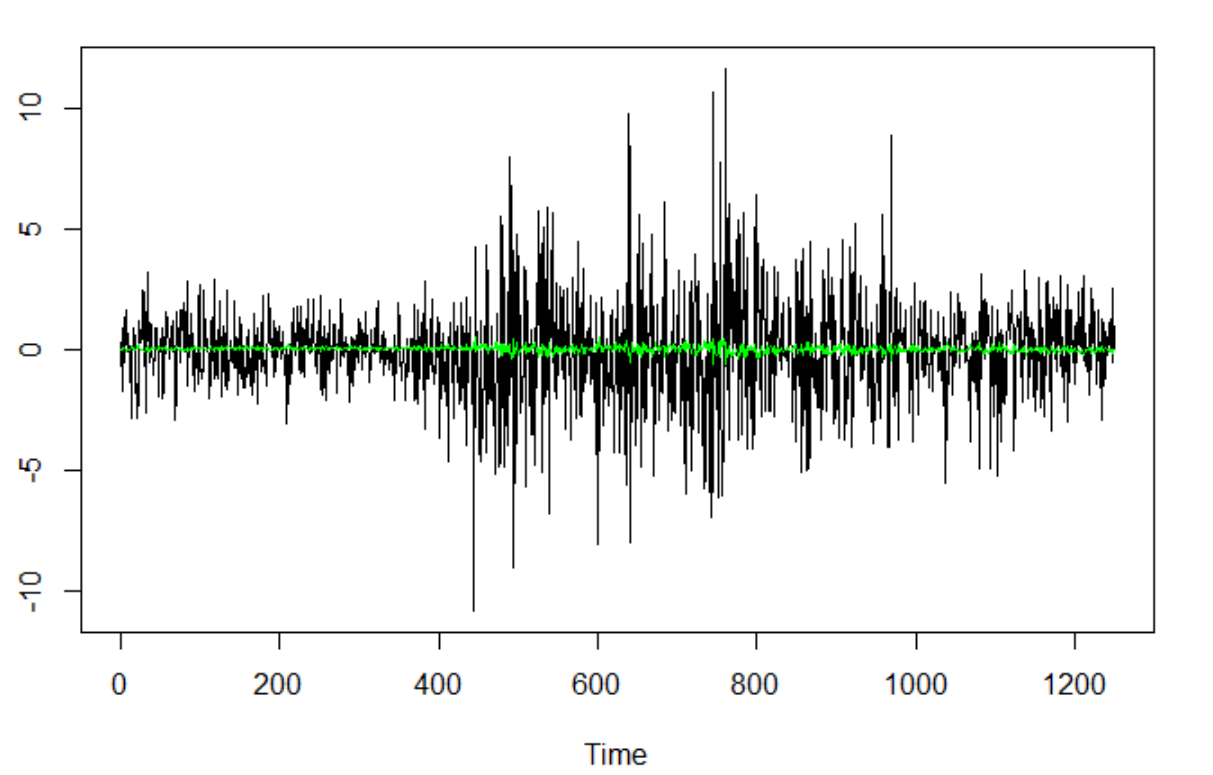
MA(1) has the previous term of error as predictor.

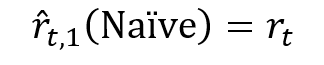
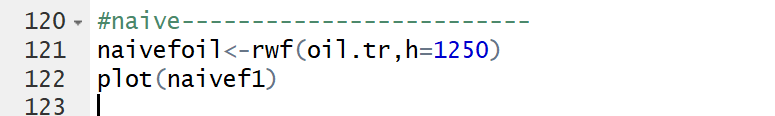


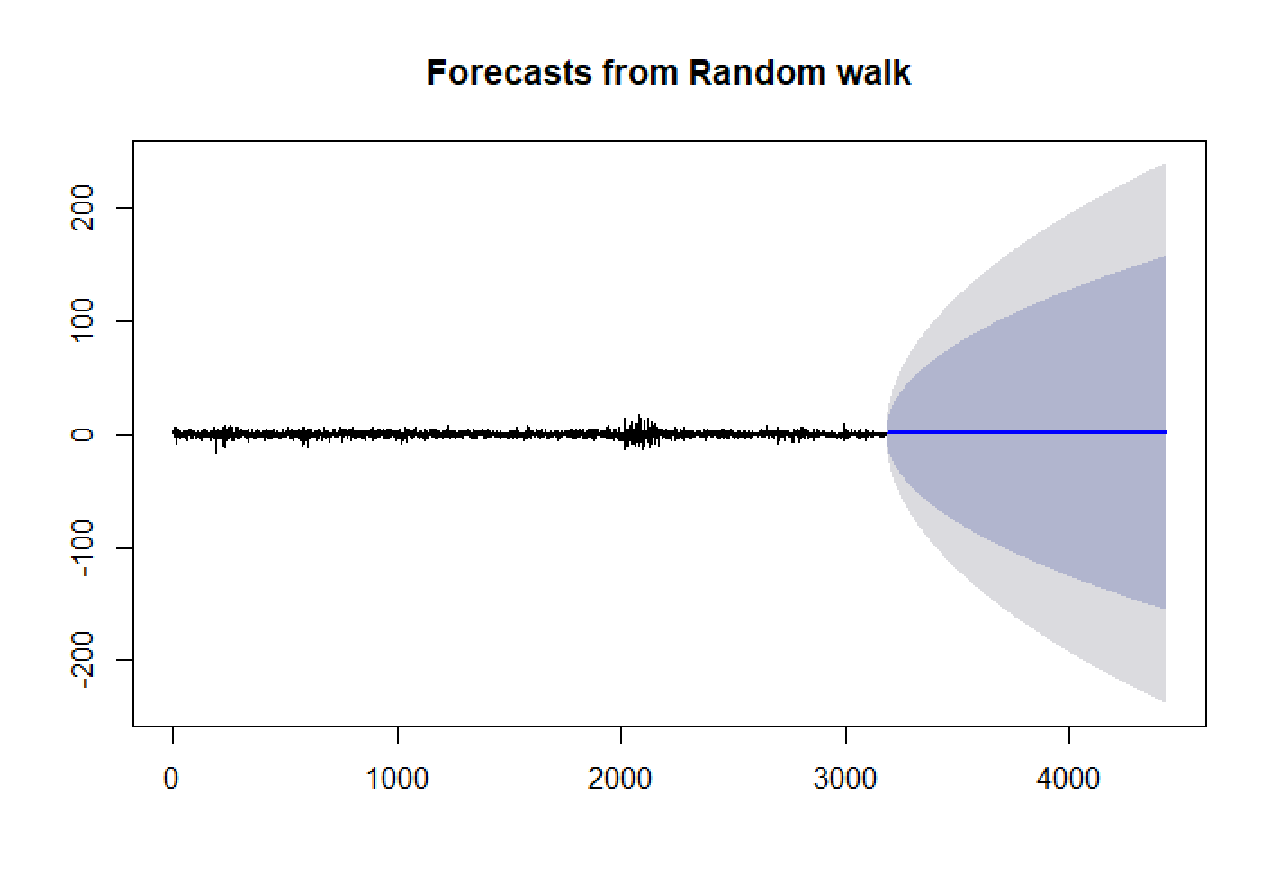
ARIMA101

This model is basically combination of AR(1) process and MA(1) process

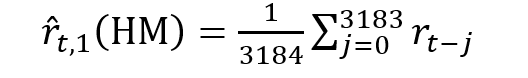




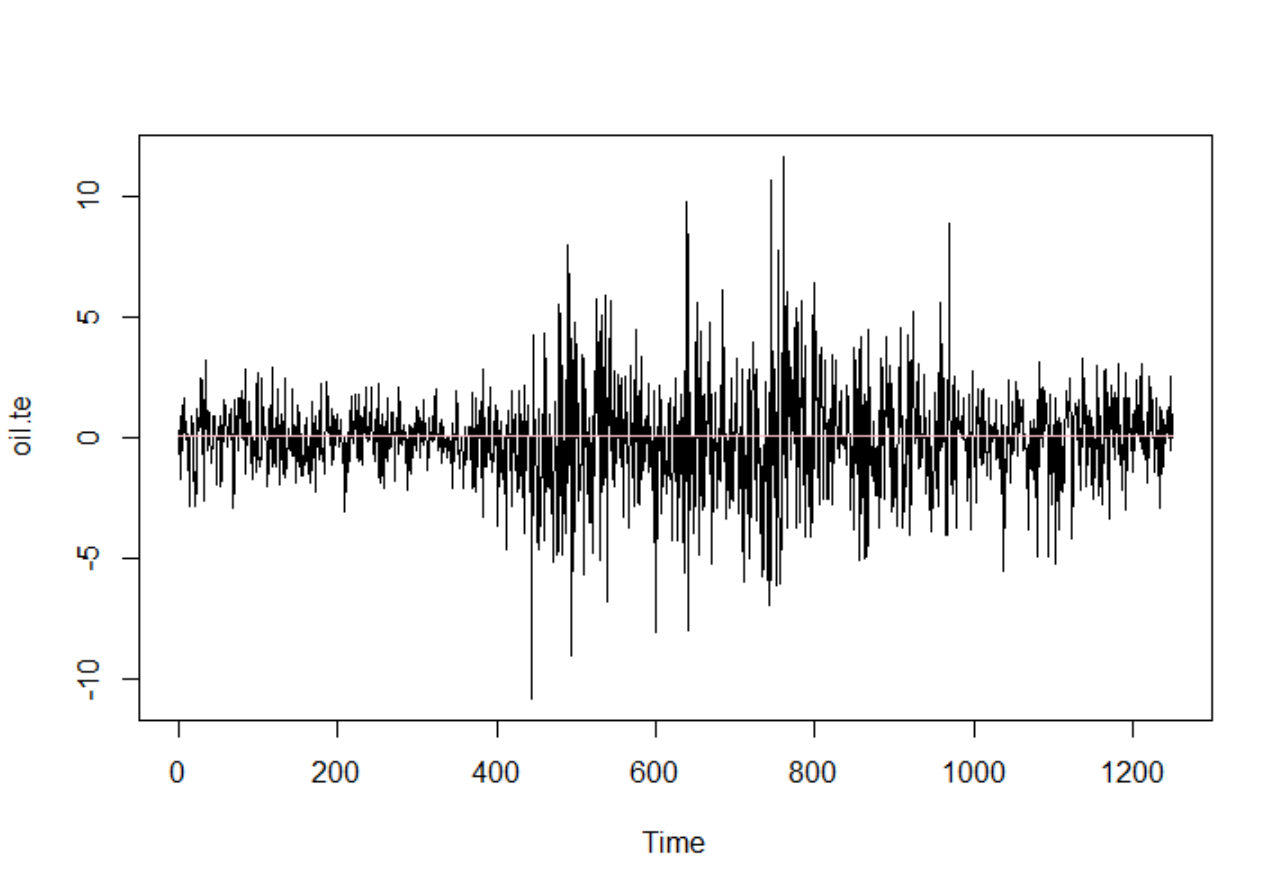
Naïve



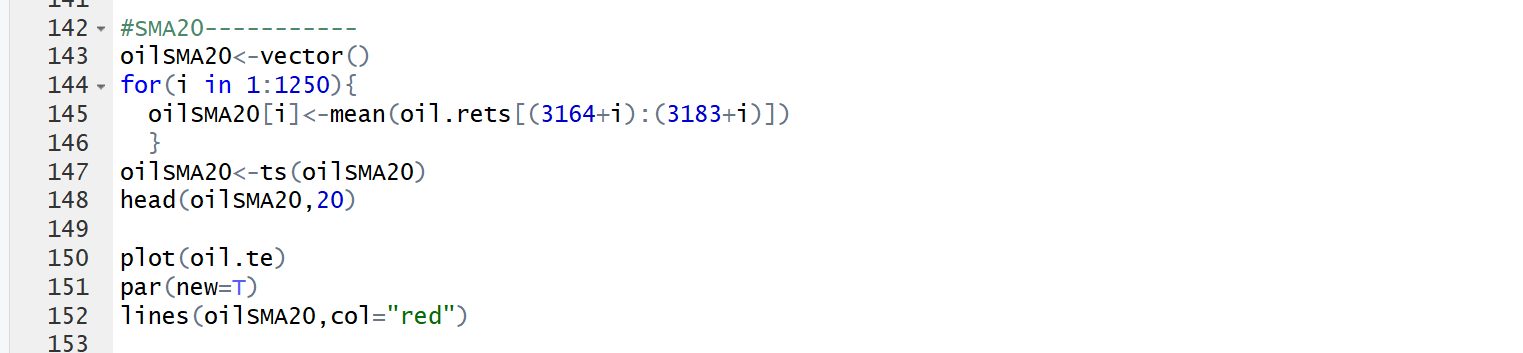
Historical Mean

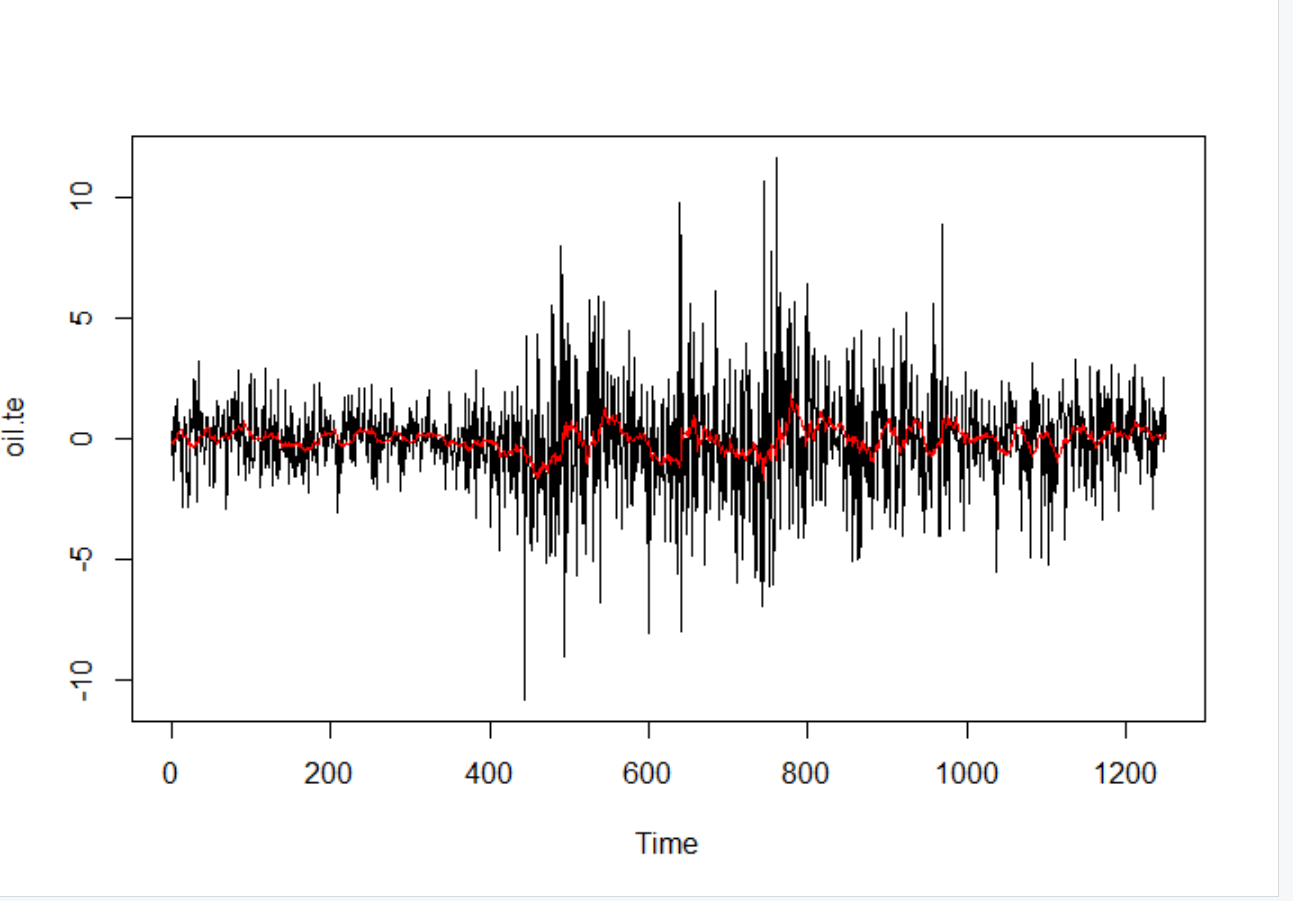




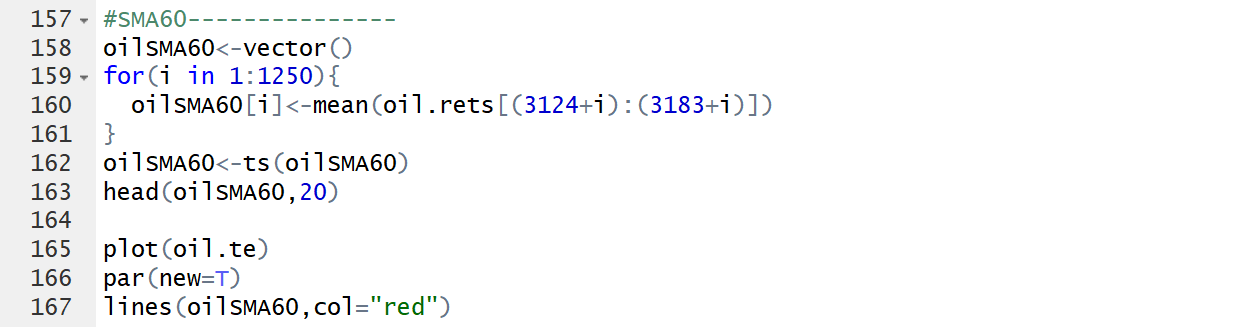


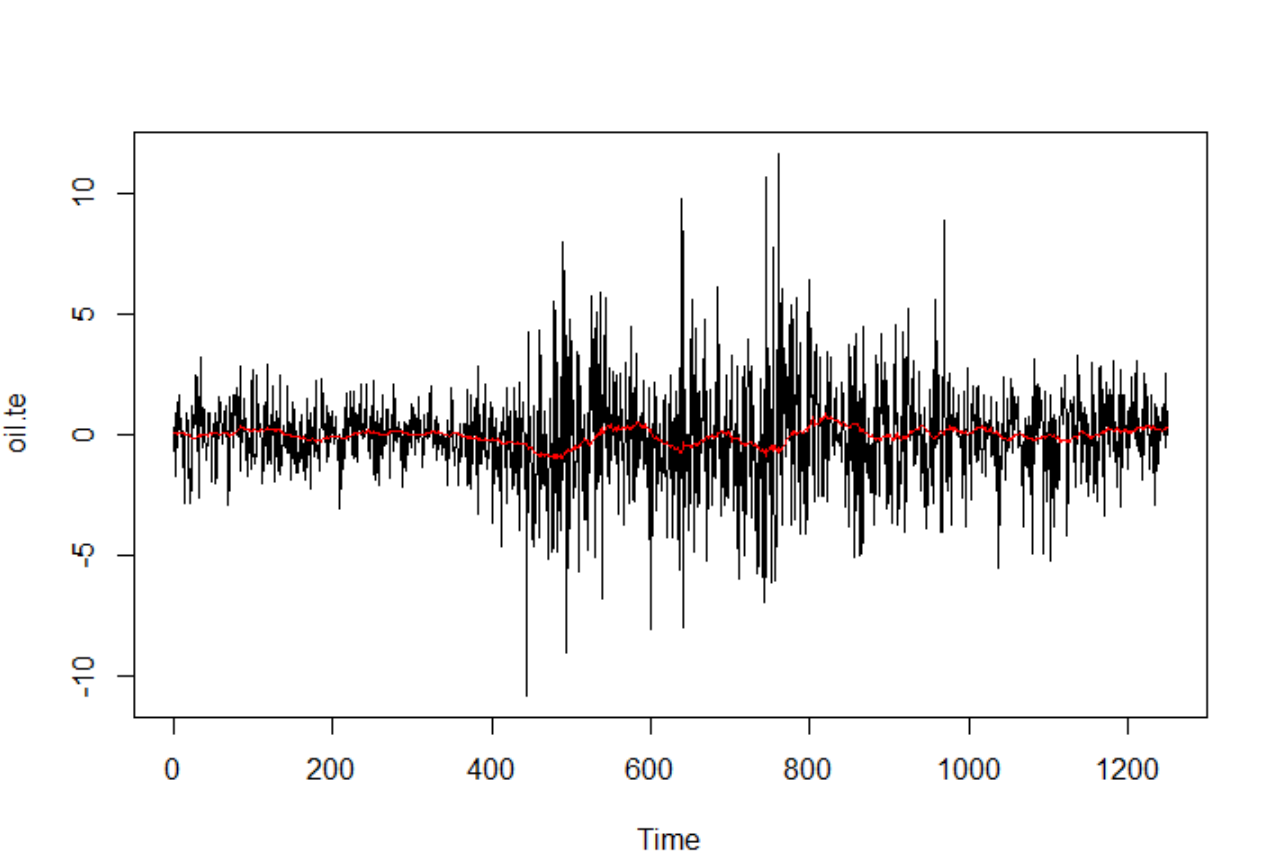
Simple Moving Average Lag=m

Simple Moving Average(lag=20)

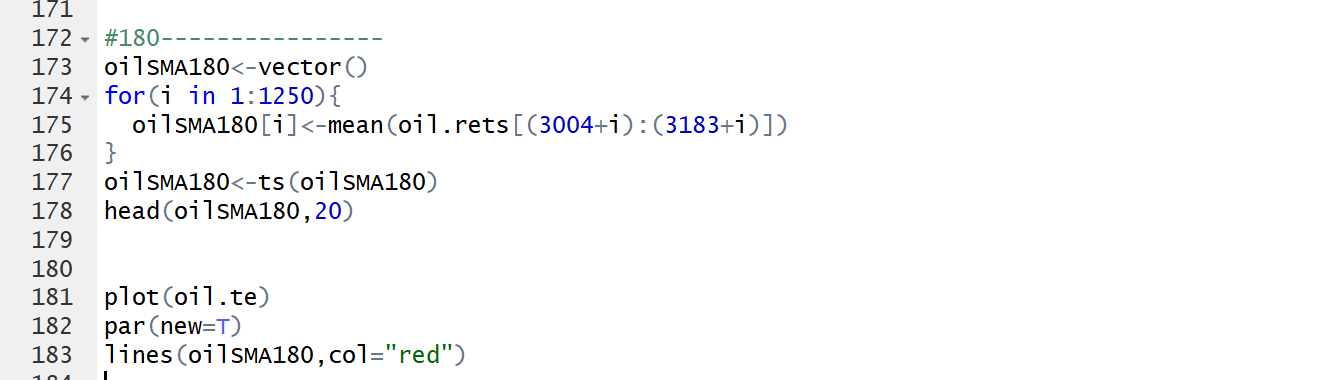
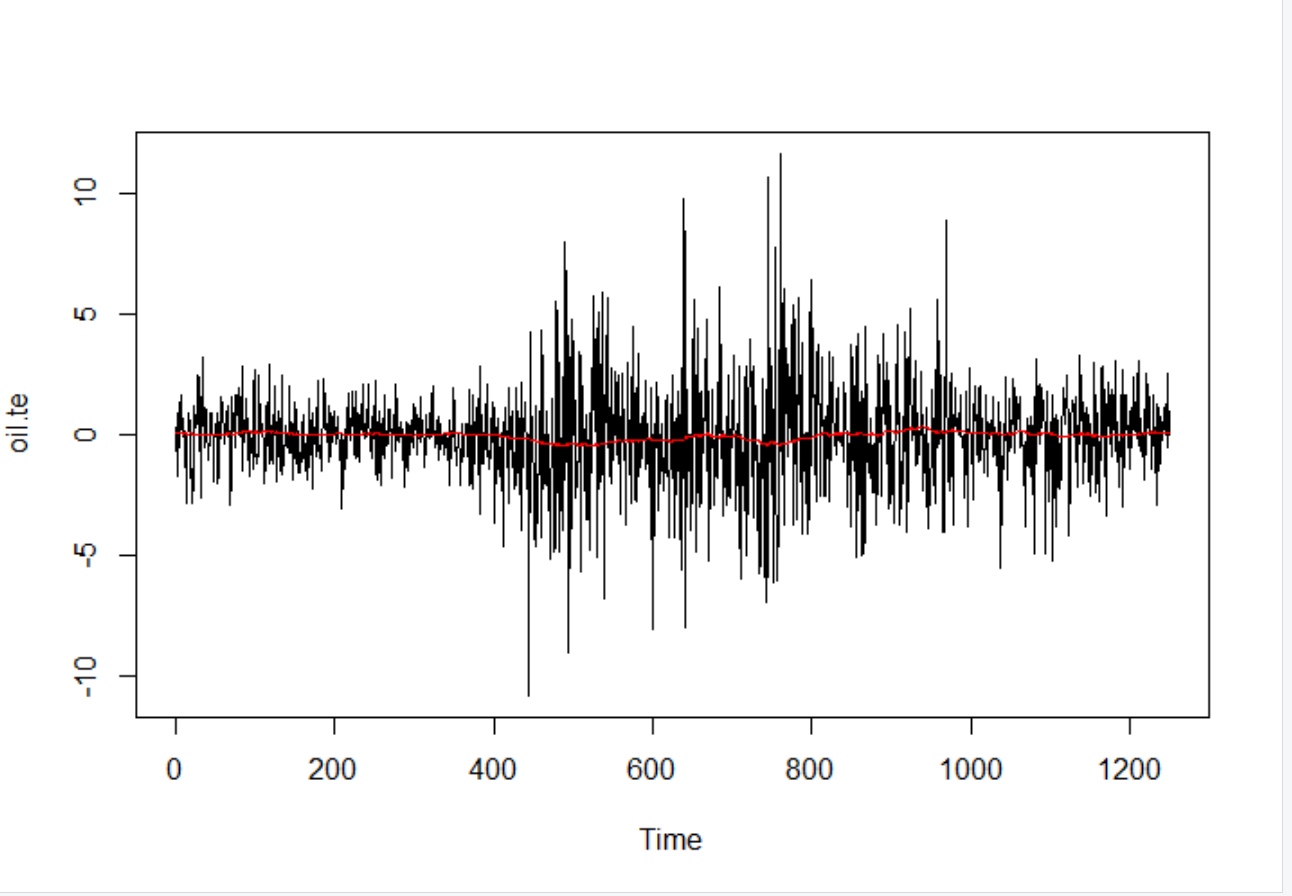


Simple Moving Average(lag=60)

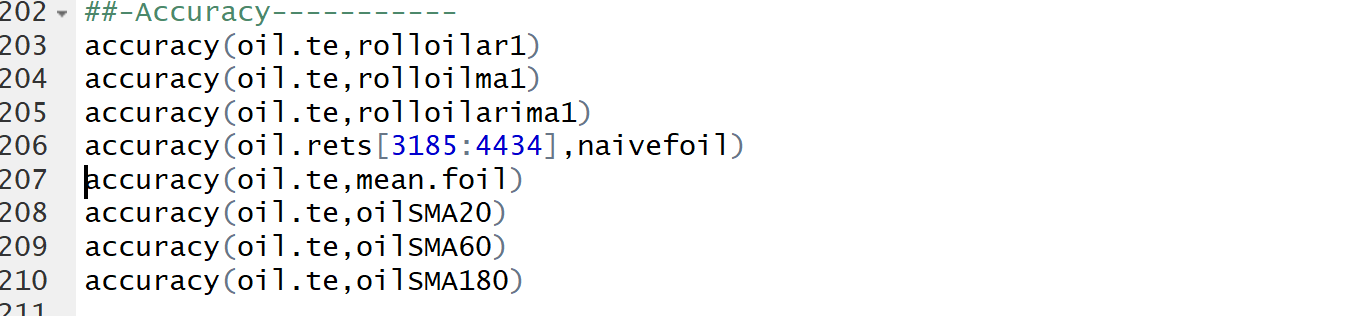


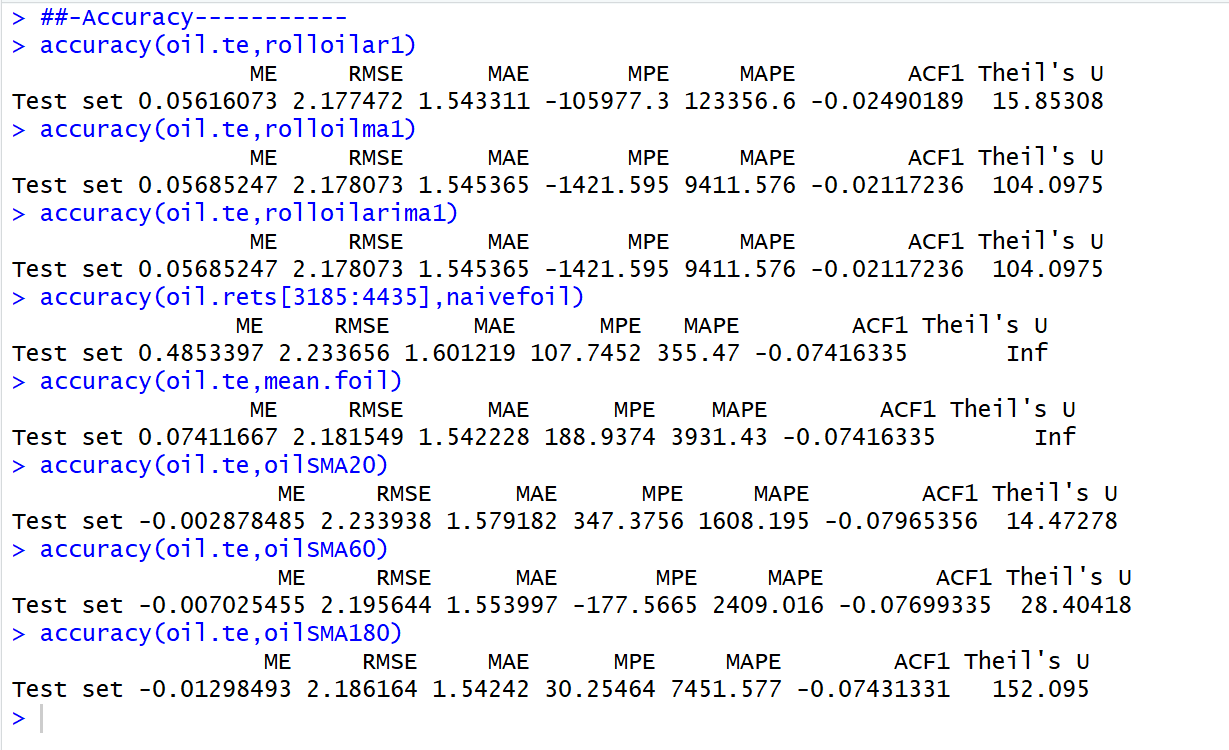


Simple Moving Average(lag=180)

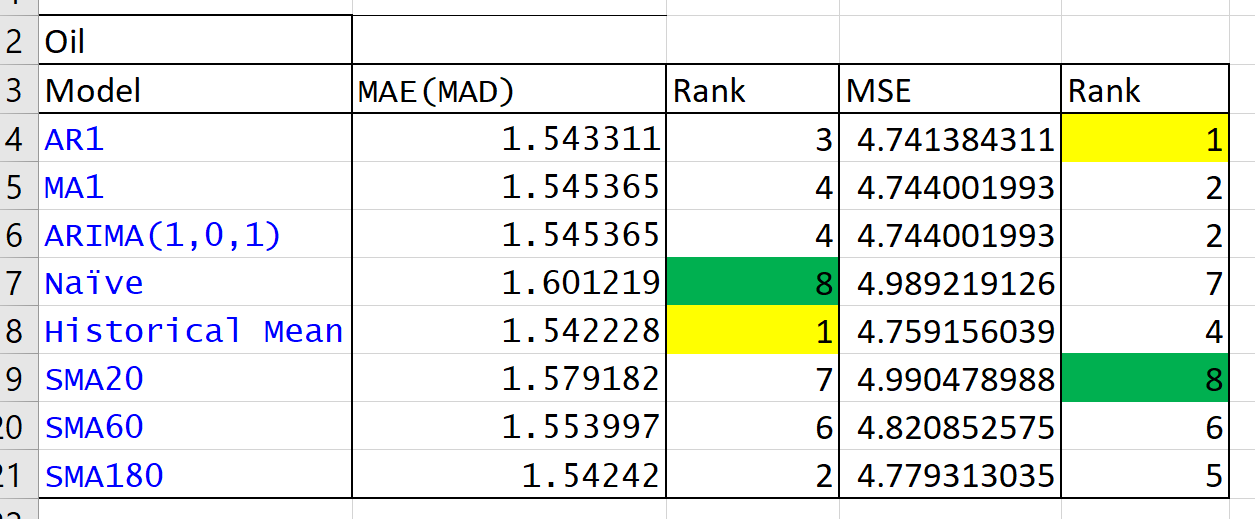


Accuracy





Based on the above output



This table can be driven. We usually prefer smaller value of MAE(MAD) and MSE. Since, smaller value of it indicates, a model has more capability of capturing data. Meanwhile, higher error signify that the variation of the data could not be captured by model’s predictors.

Based on MAE(MAD)

Best Model: Historical Mean Model

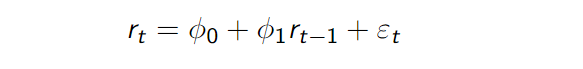
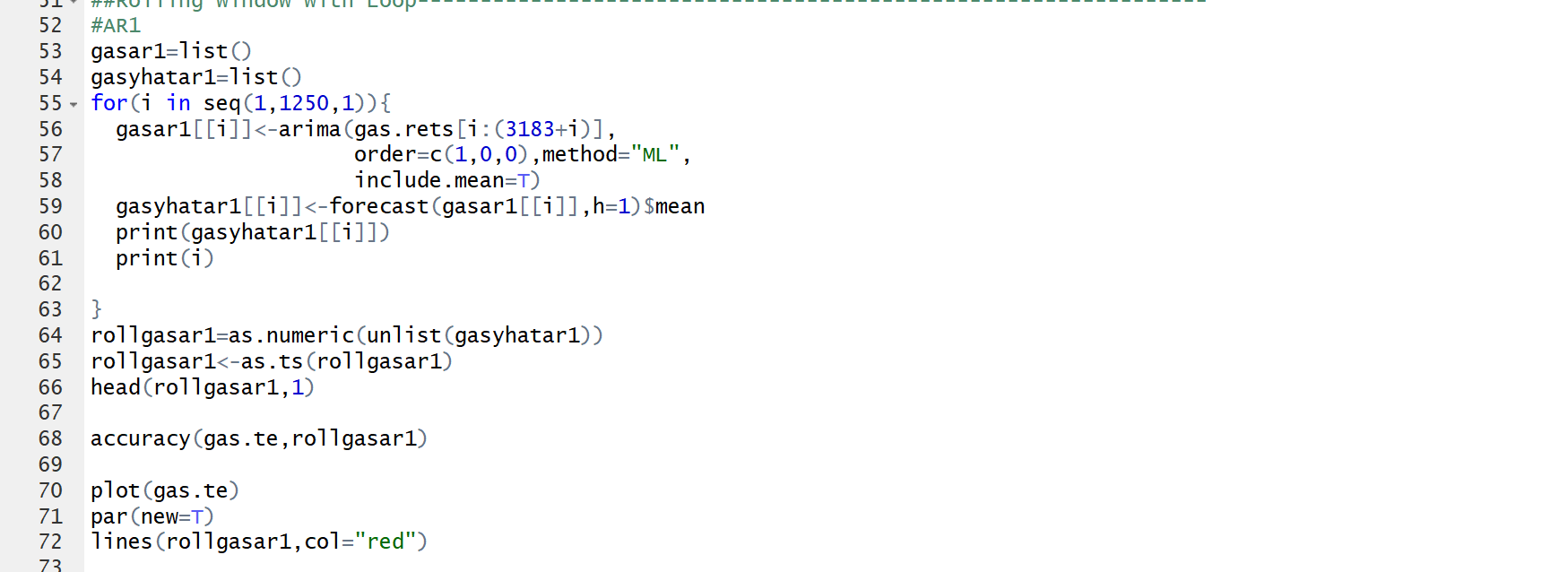
Worst Model: Naïve Model

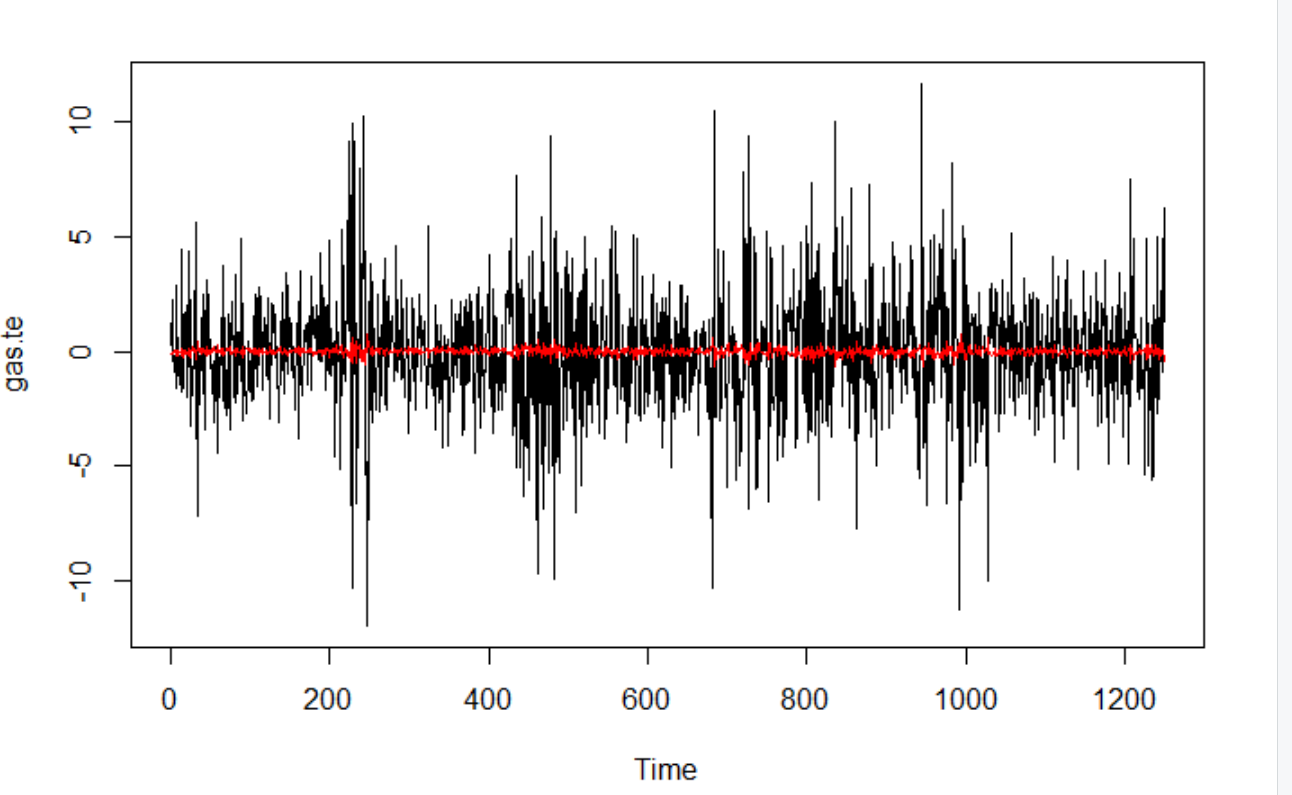
Based on MSE

Best Model: AR1 model

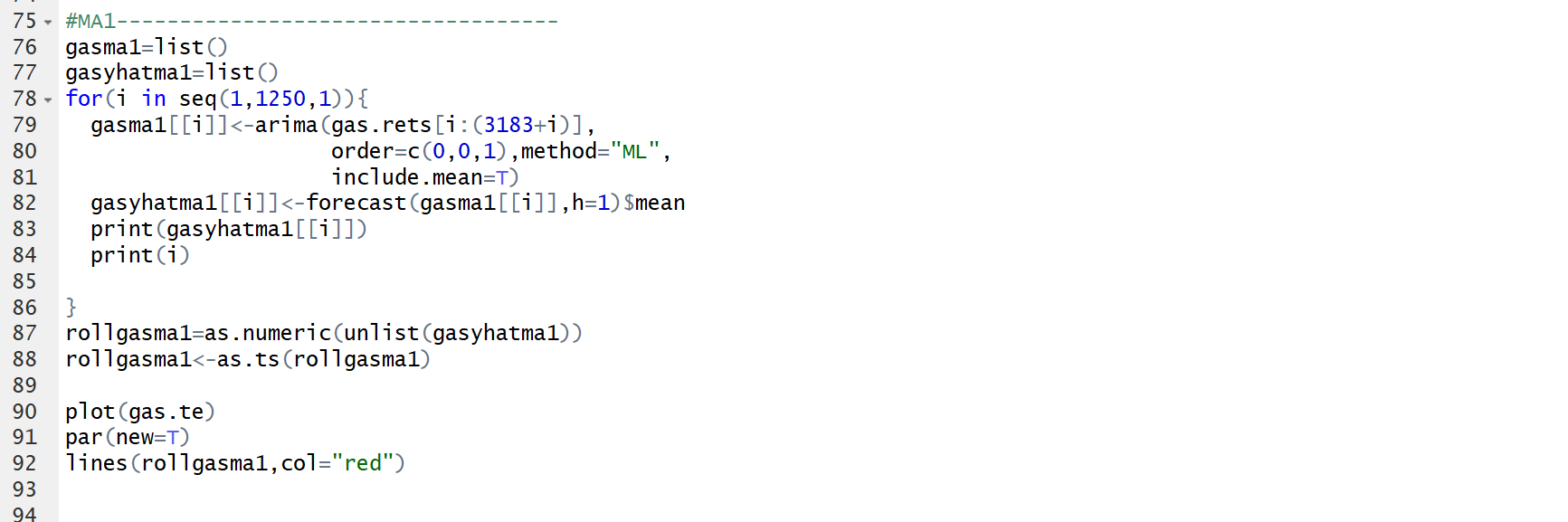
WORST Model: Simple moving average model(lag=20)

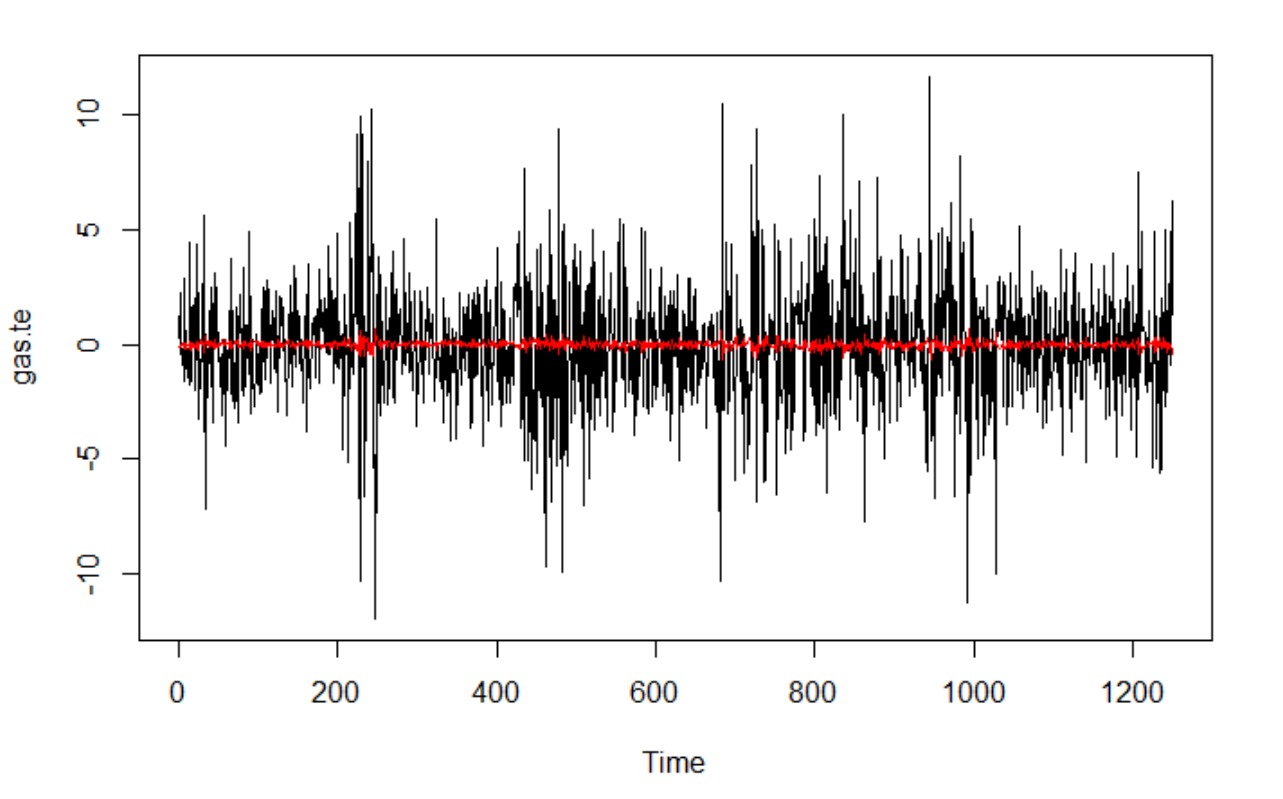
**Gas**

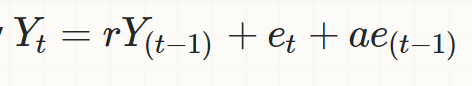
AR(1)

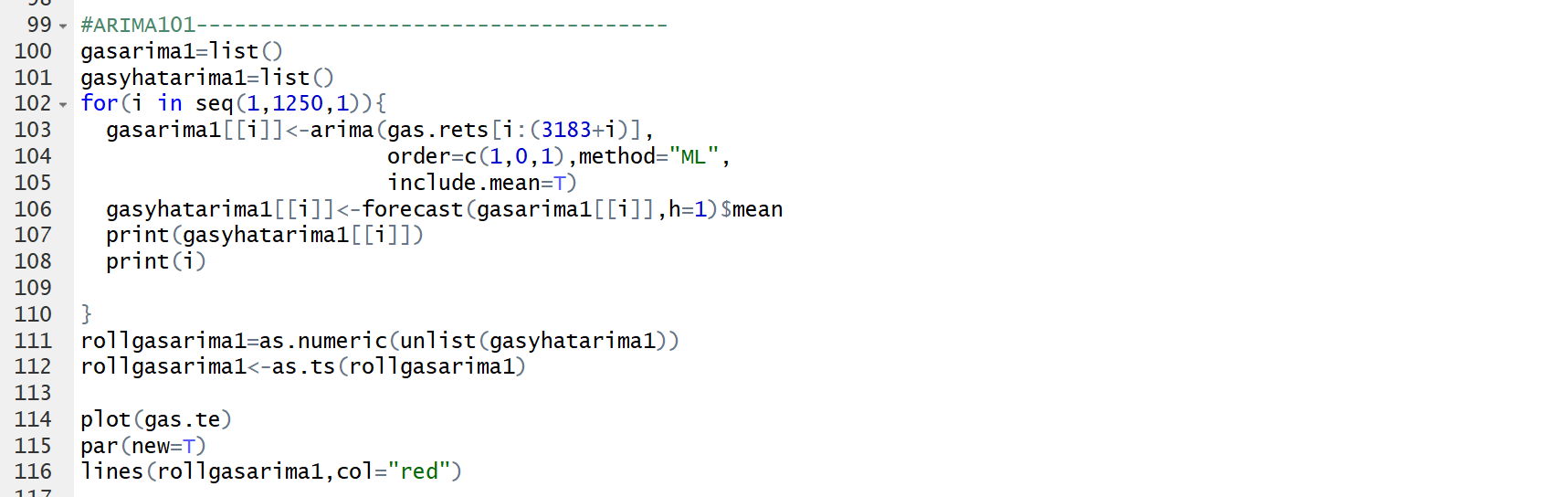


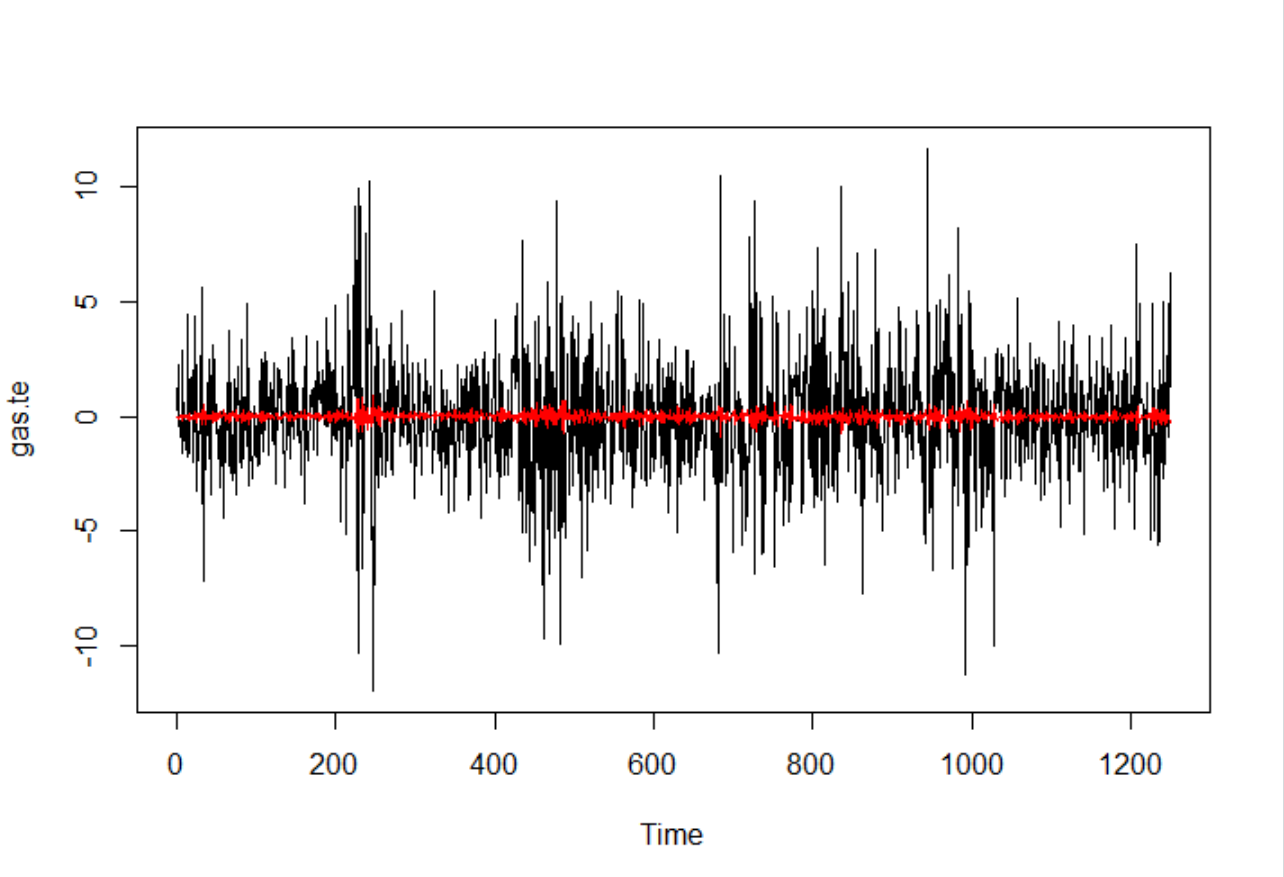
MA1



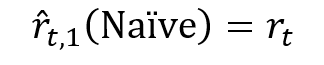


ARIMA101

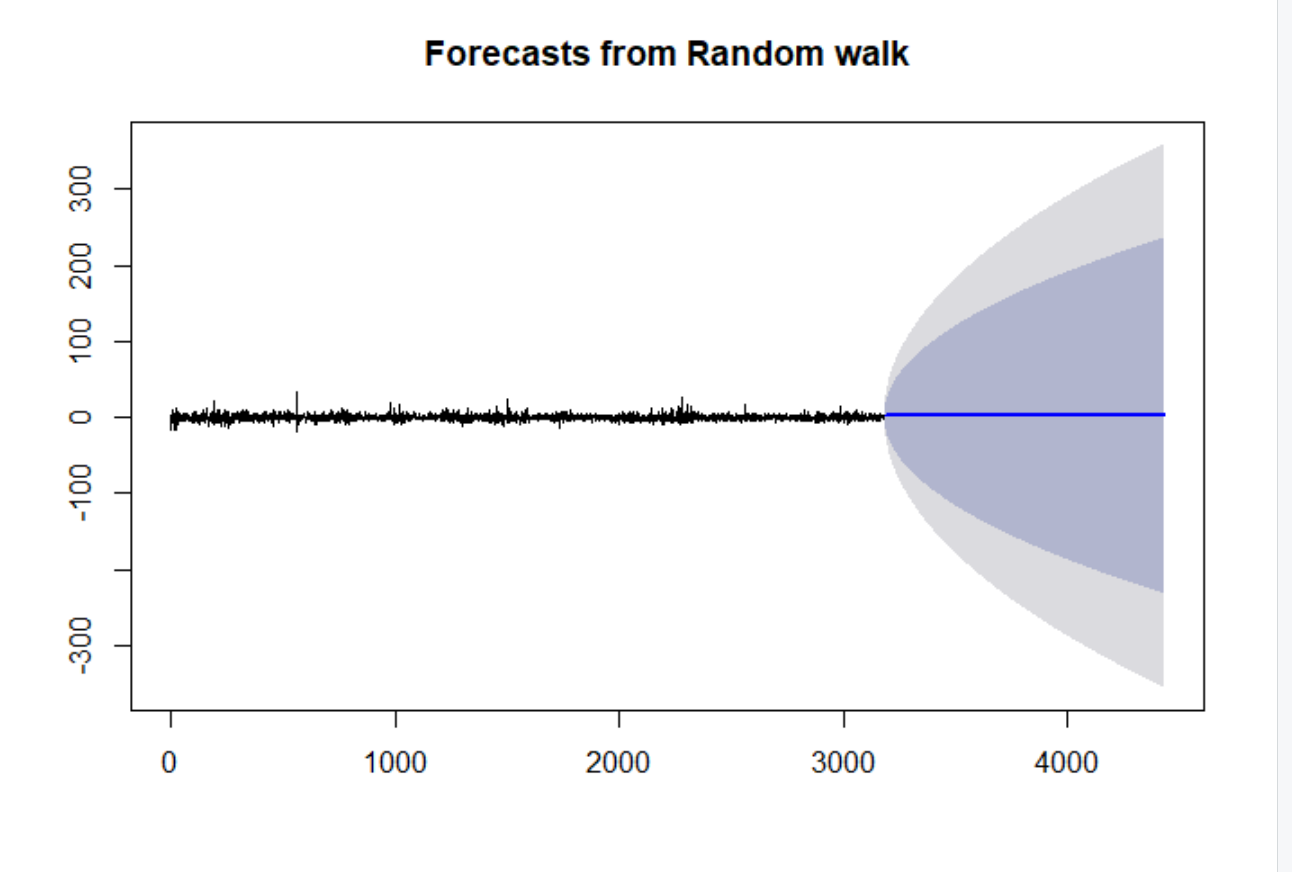




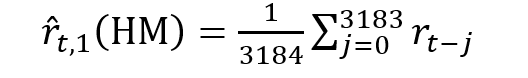
Naïve



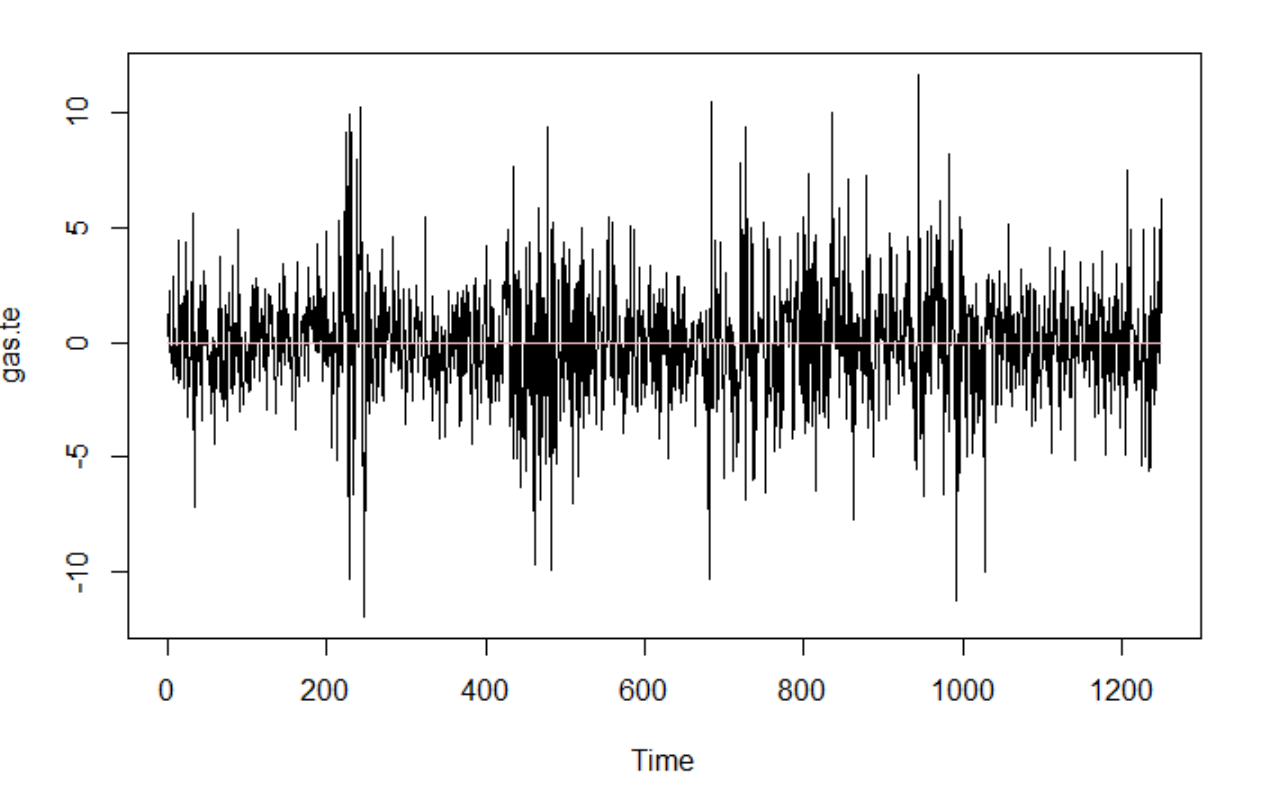




Historical Mean

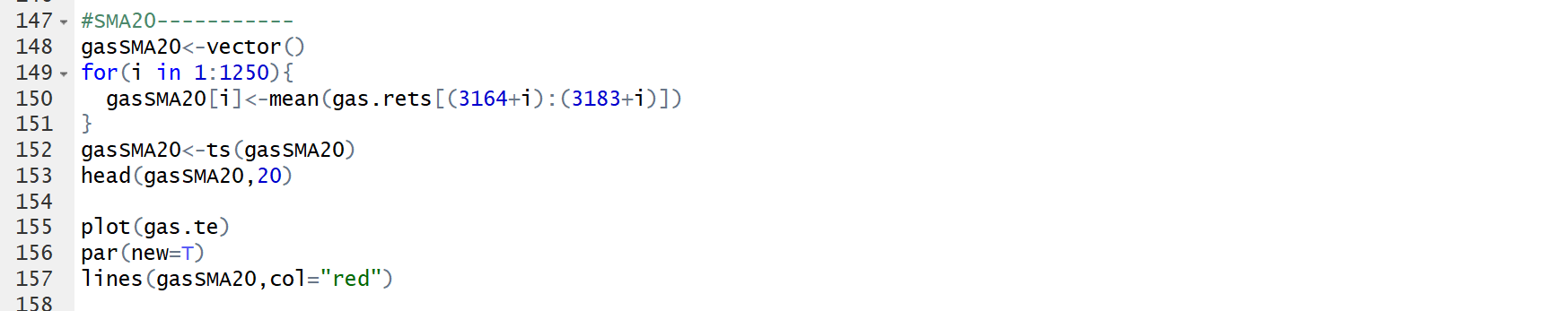


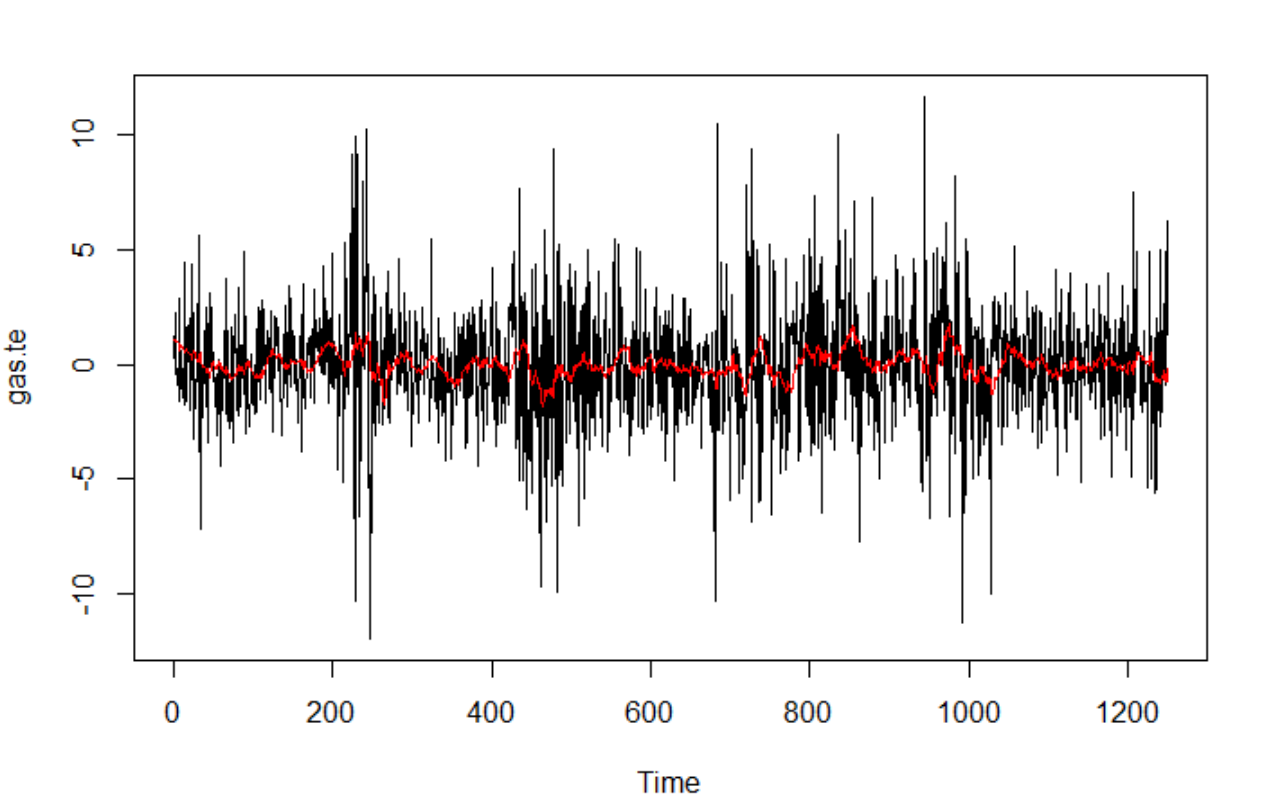




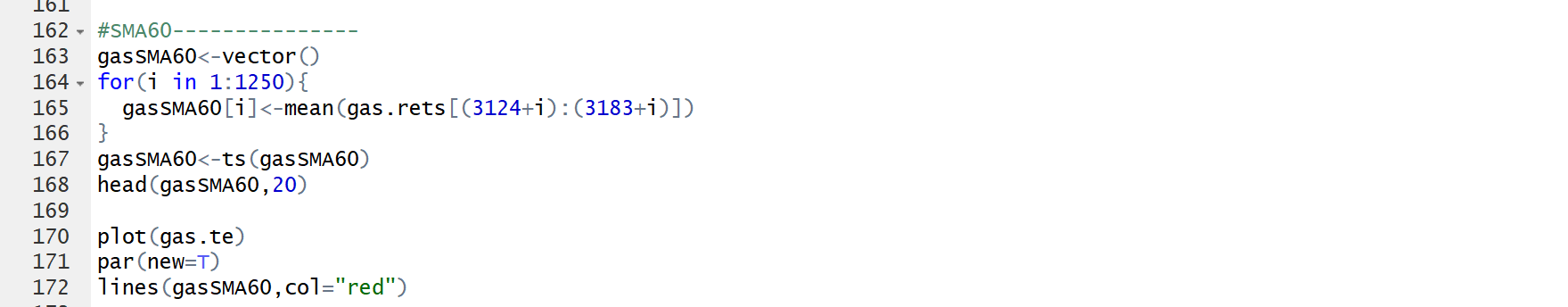
Simple Moving Average Lag=m

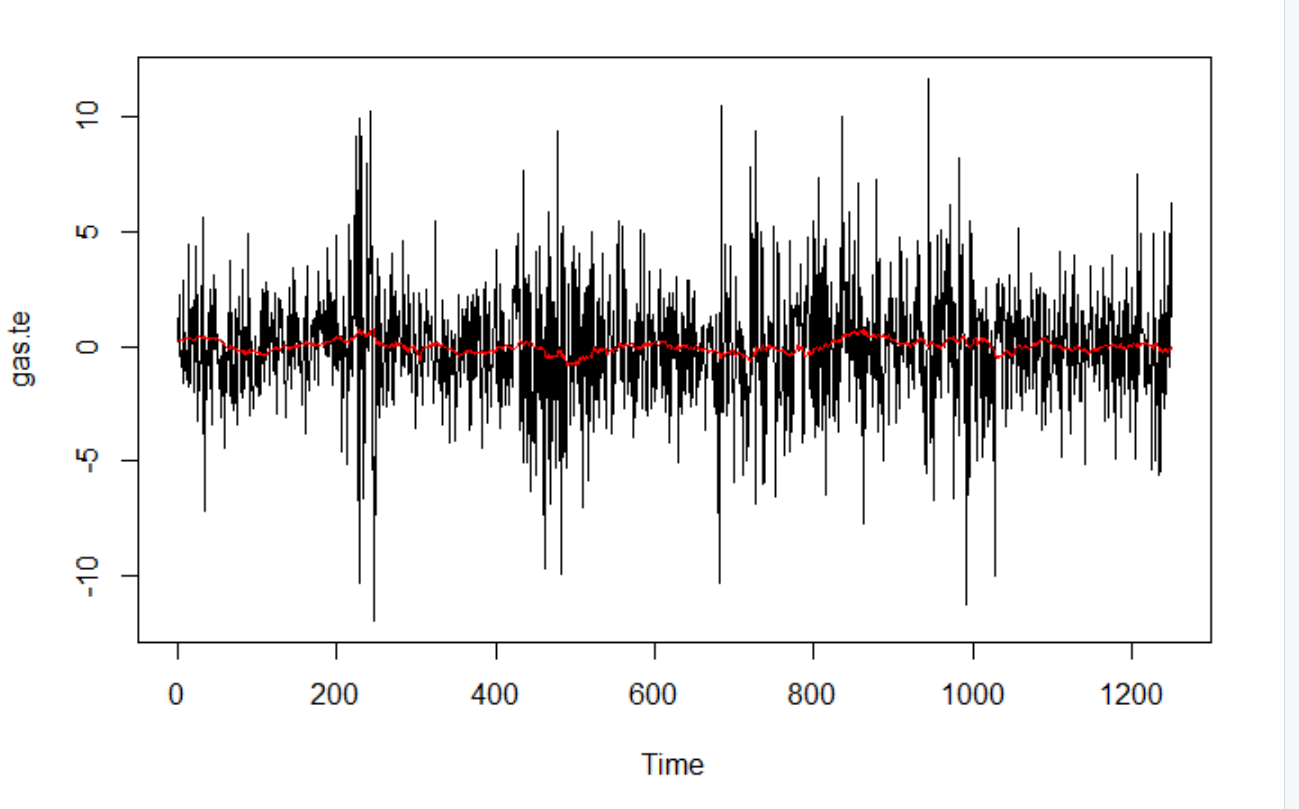
Simple Moving Average(lag=20)



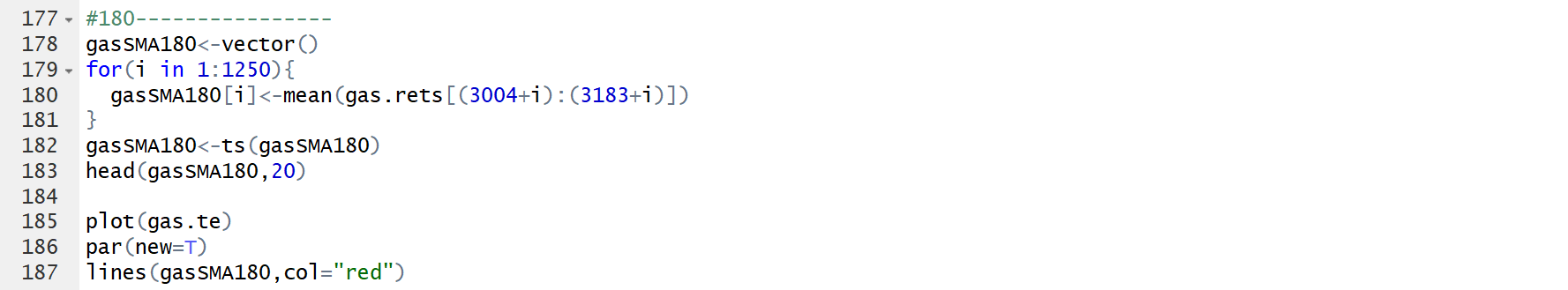
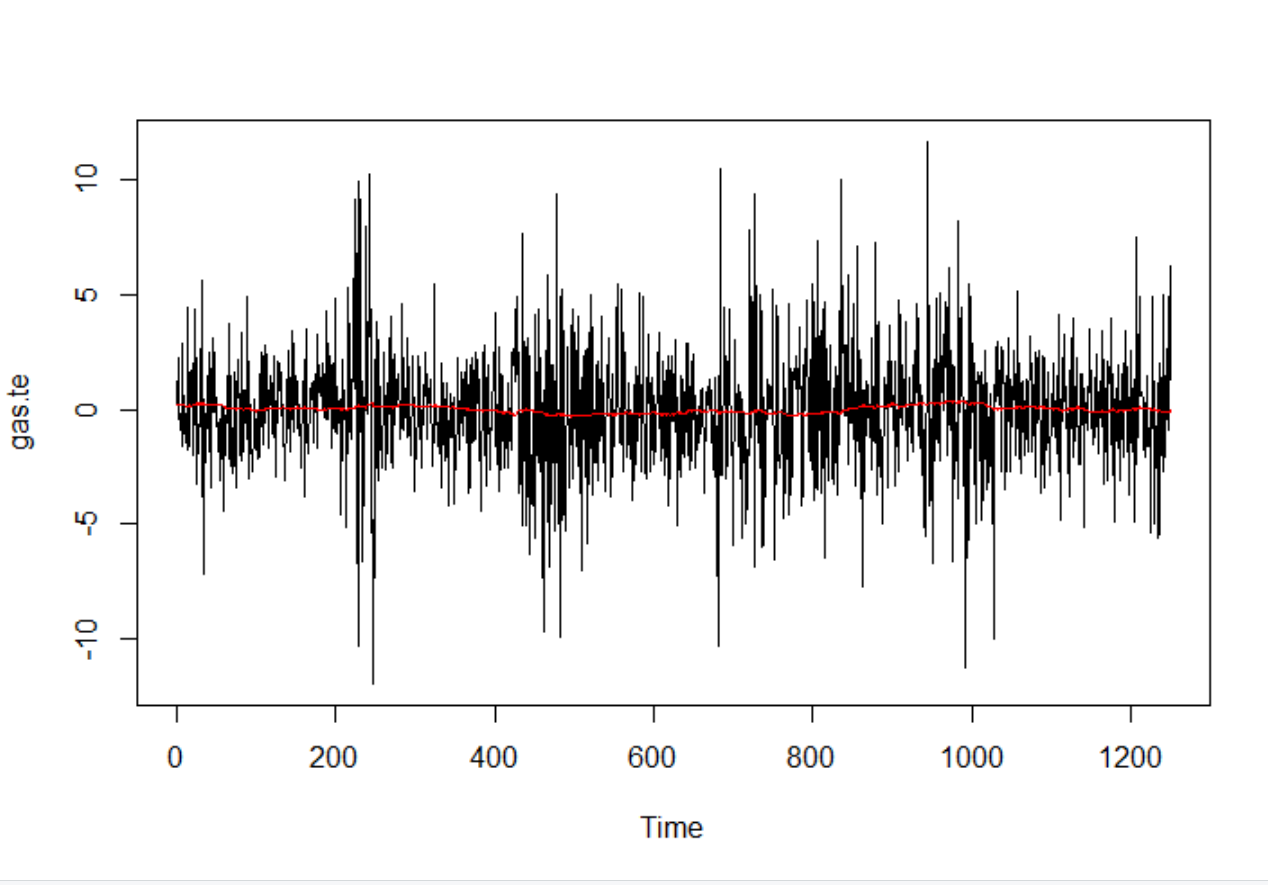


Simple Moving Average(lag=60)

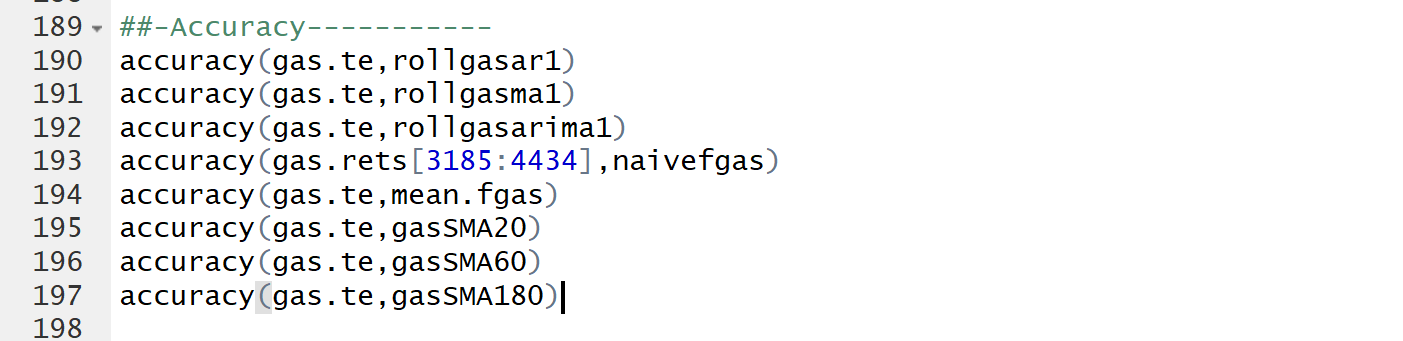


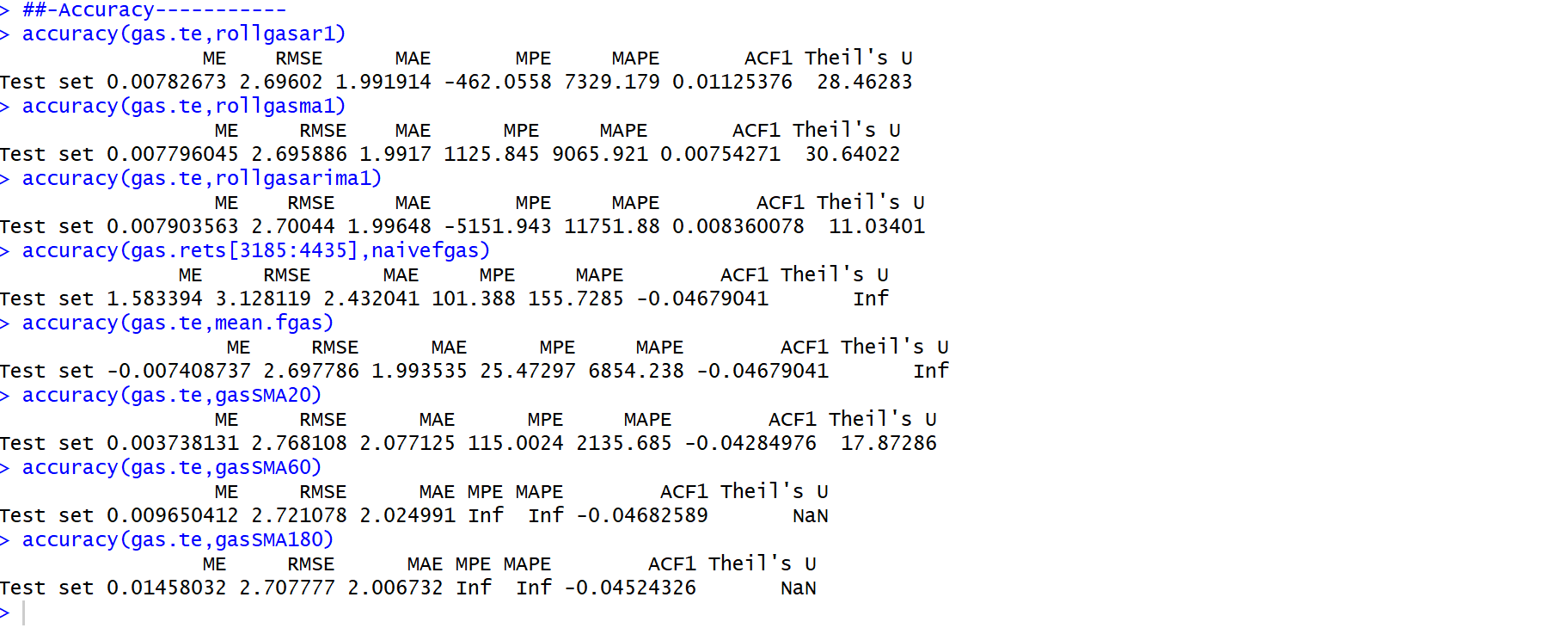


Simple Moving Average(lag=180)

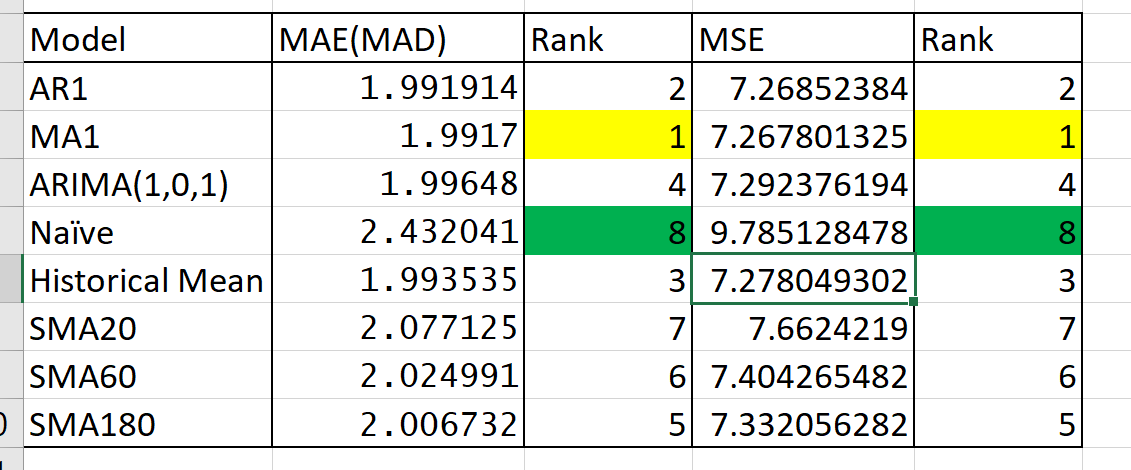


Accuracy





Based on this output, the able below can be driven



This table can be driven. We usually prefer smaller value of MAE(MAD) and MSE. Since, smaller value of it indicates, a model has more capability of capturing data. Meanwhile, higher error signify that the variation of the data could not be captured by model’s predictors.

Based on MAE(MAD)

Best Model: MA1 model

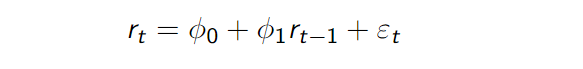
Worst Model: Naïve model

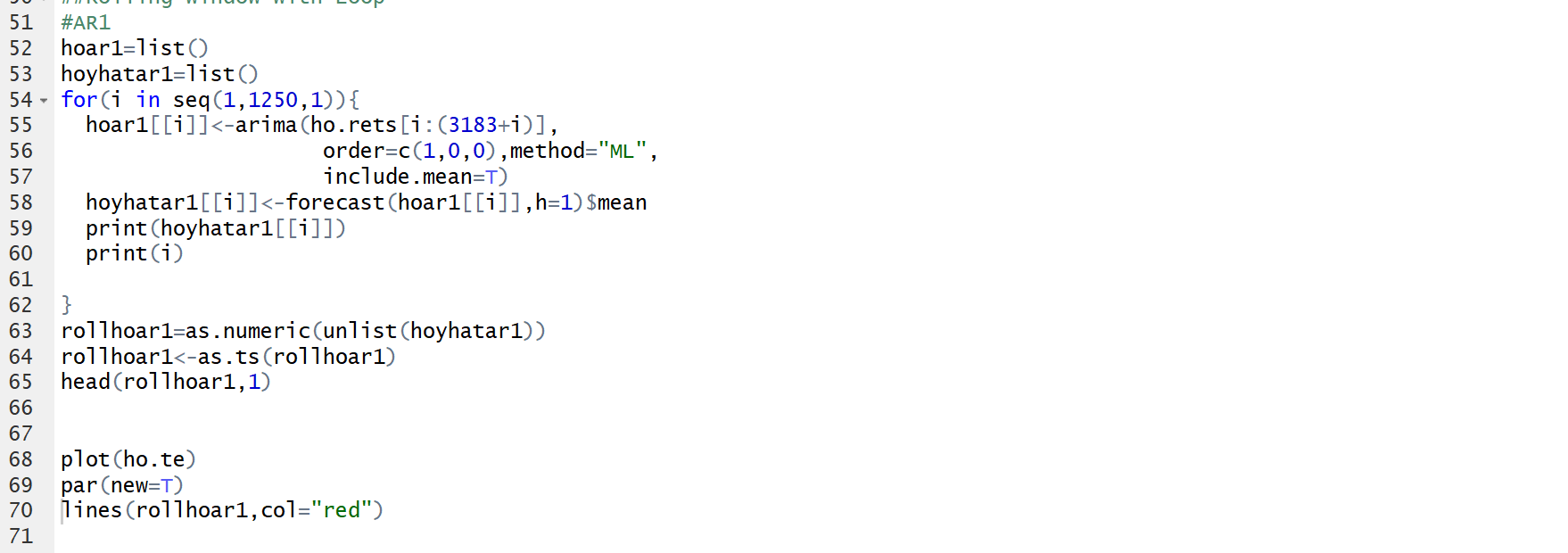
Based on MSE

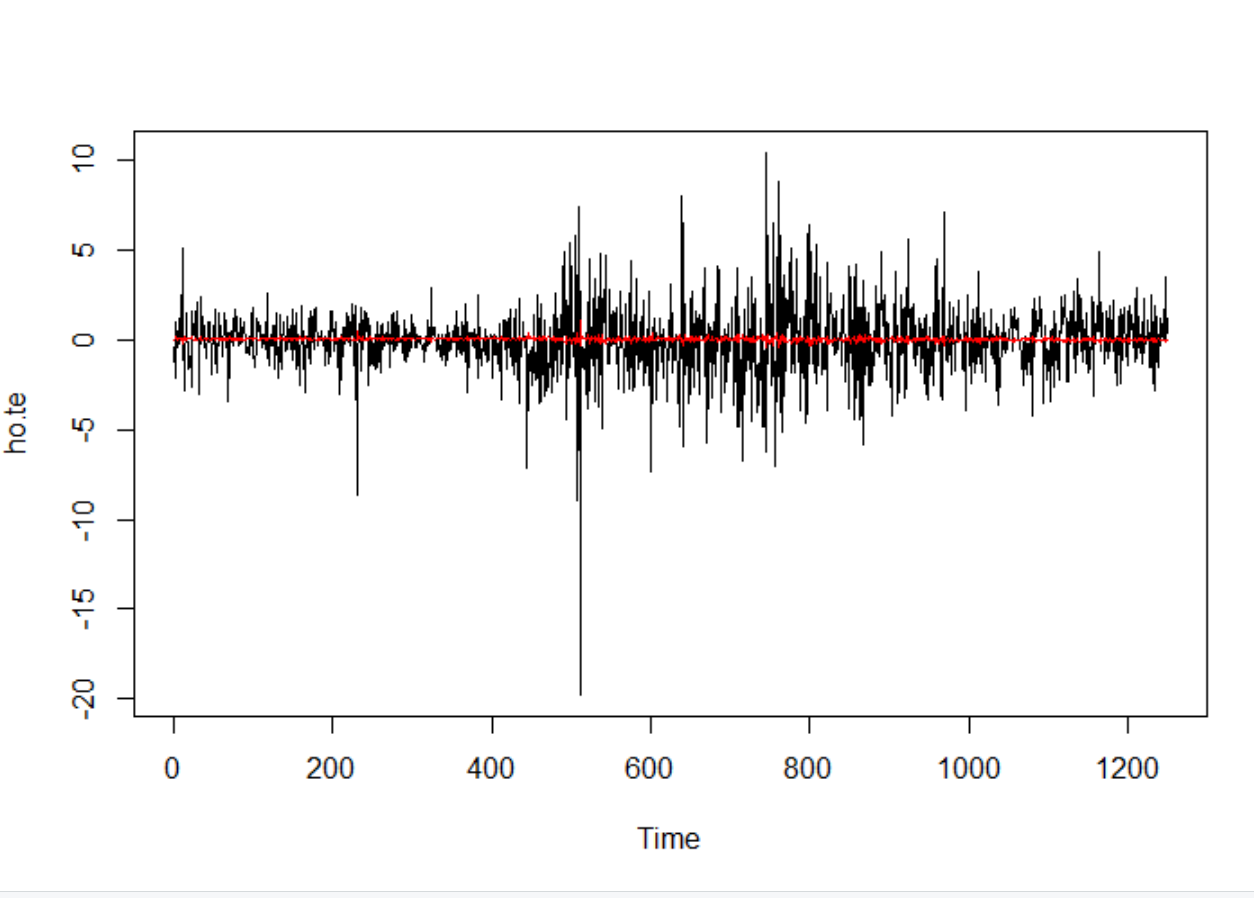
Best Model: MA1 model

WORST Model: Naïve model

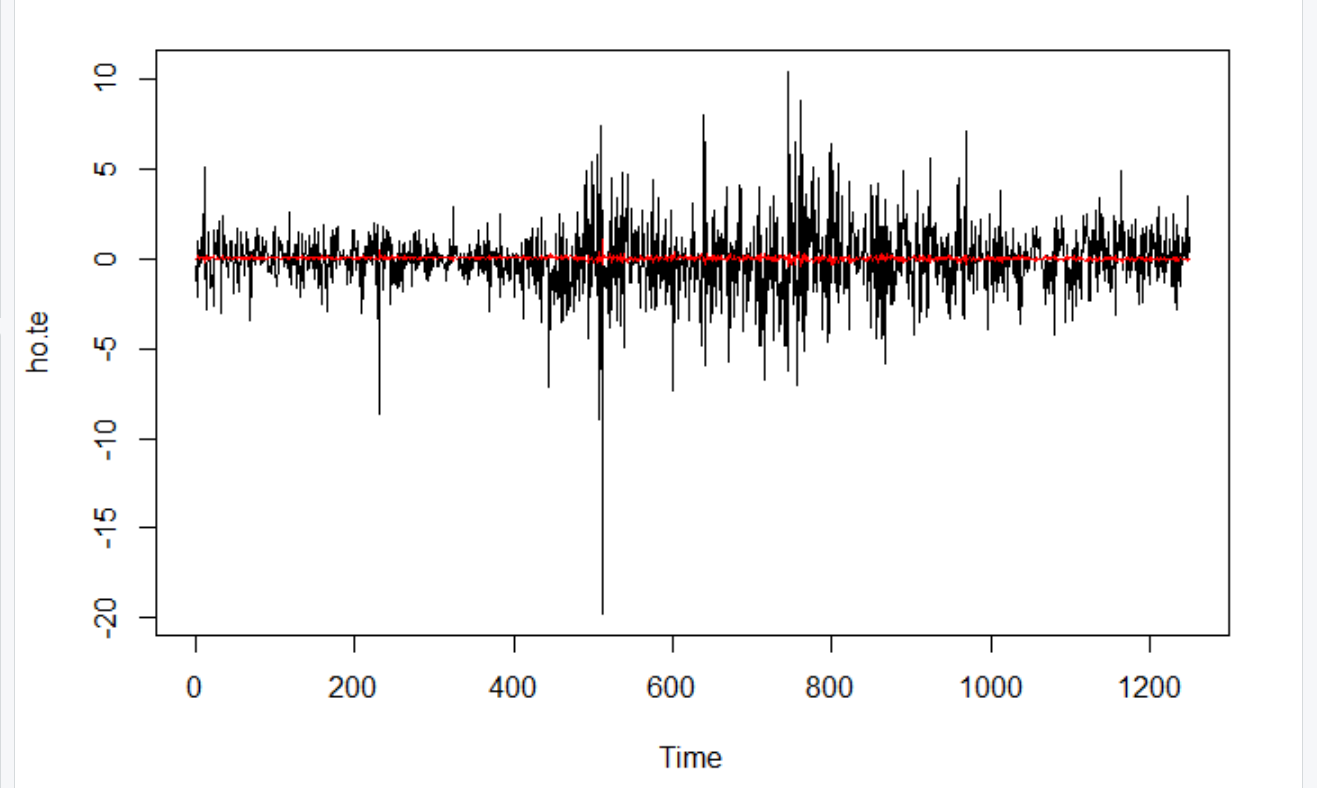
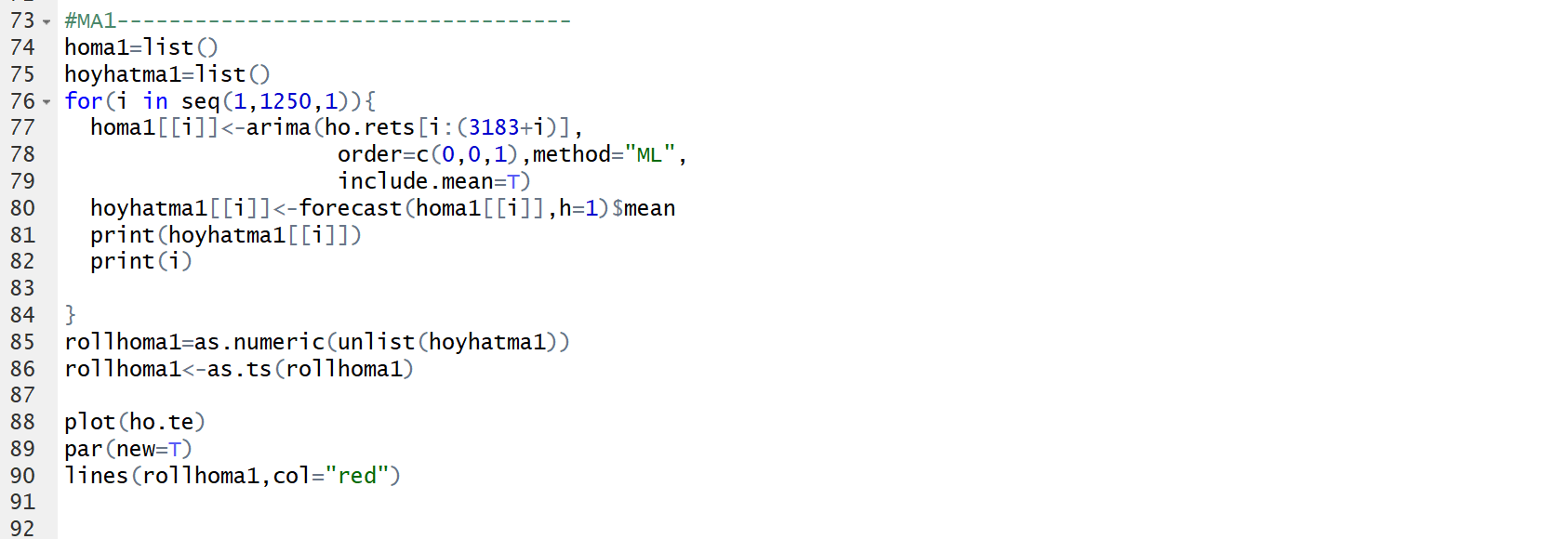
**Heating Oil**

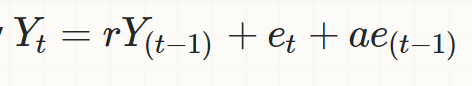
AR(1)

****

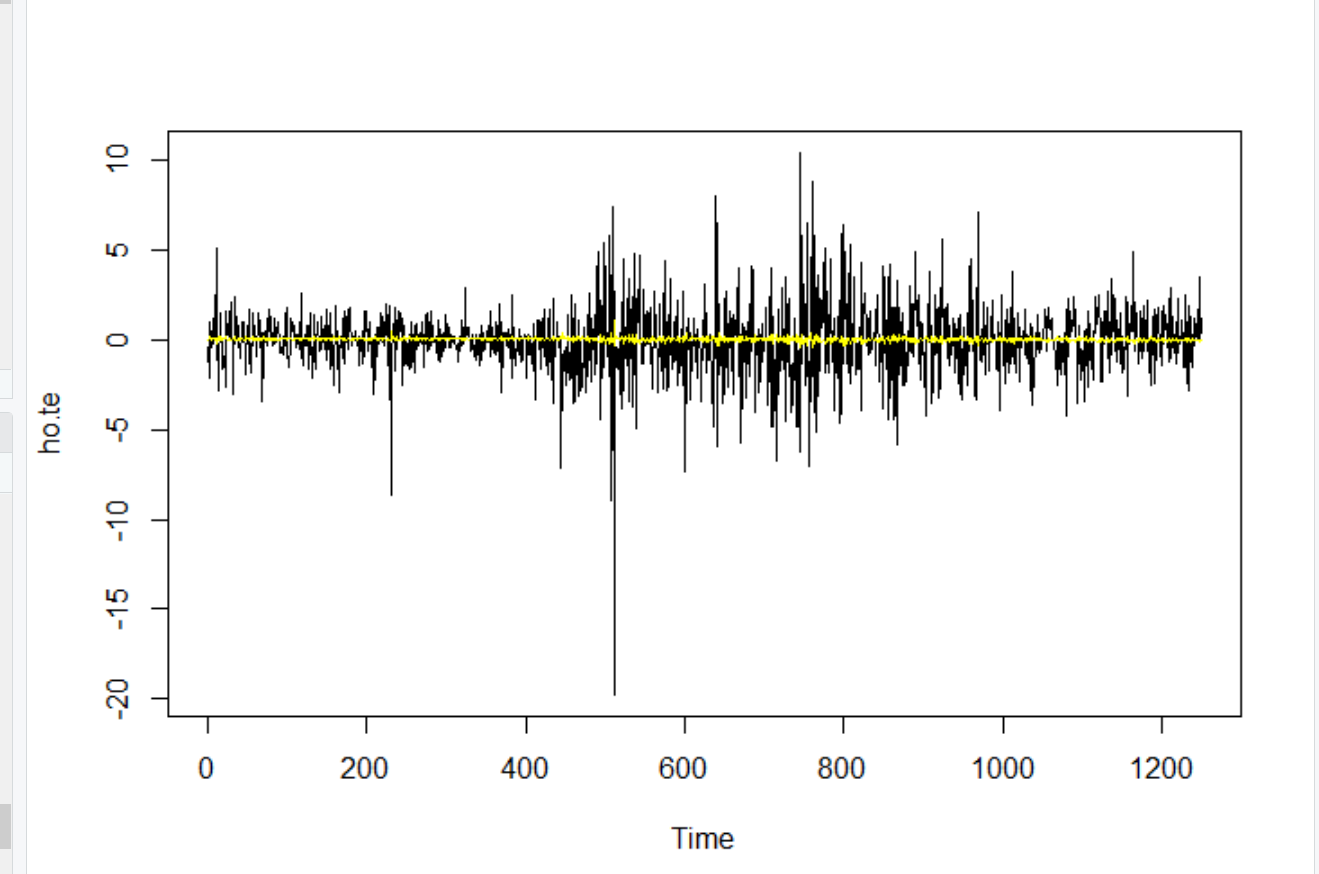
****

MA1

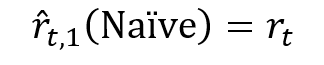
****

ARIMA101

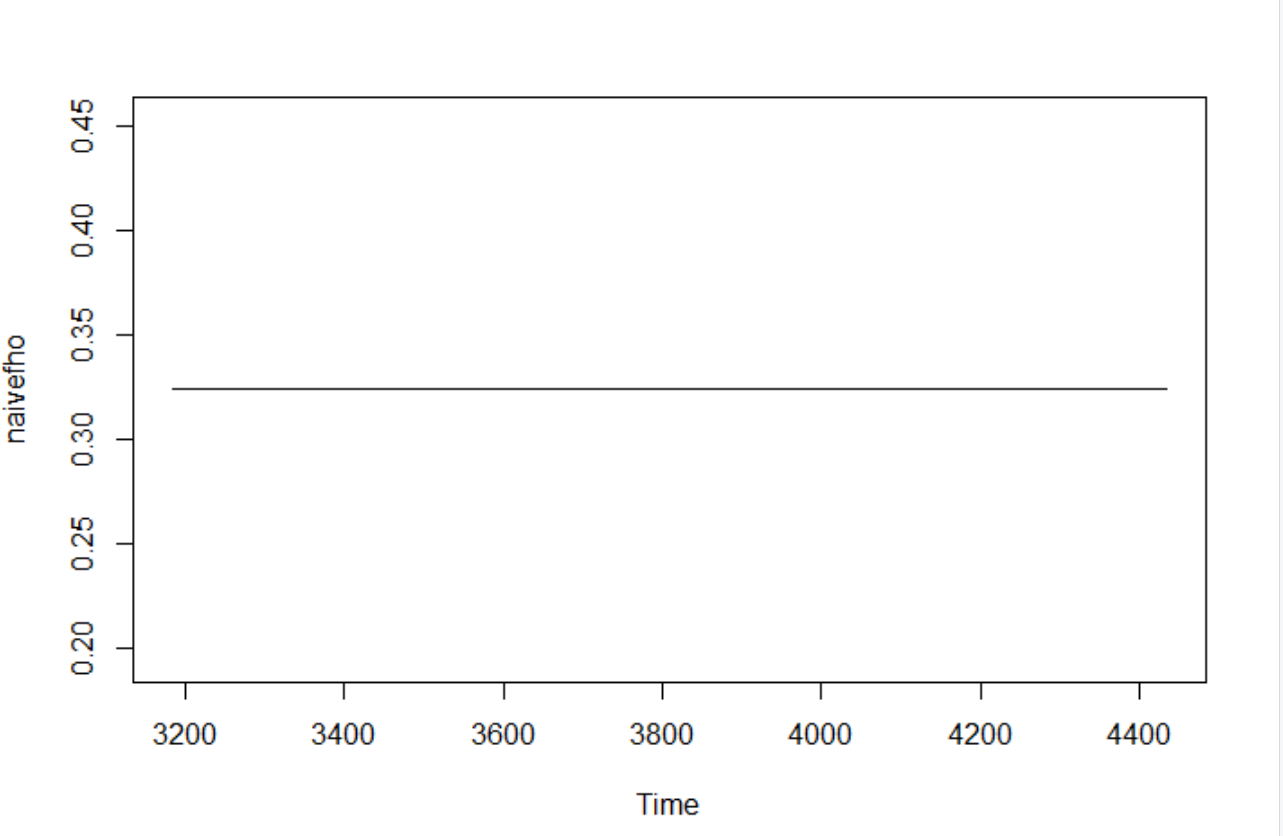
****

****

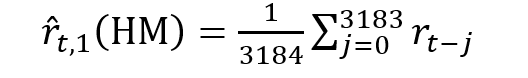
Naïve



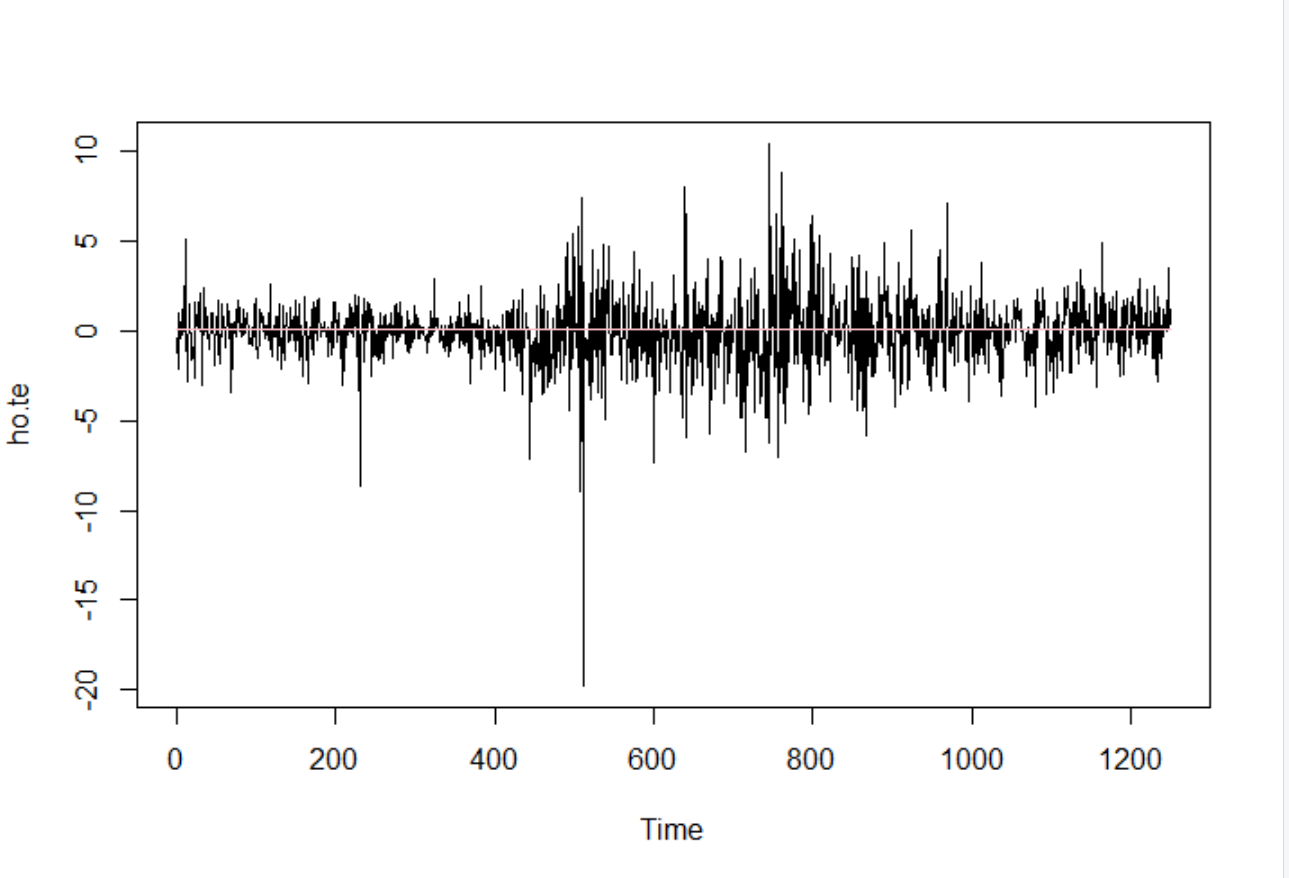
****

****

Historical Mean

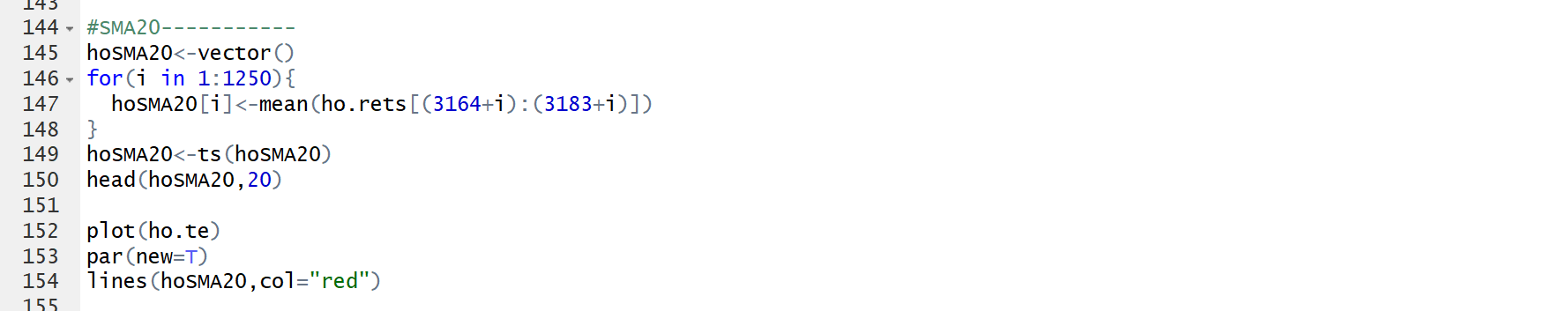


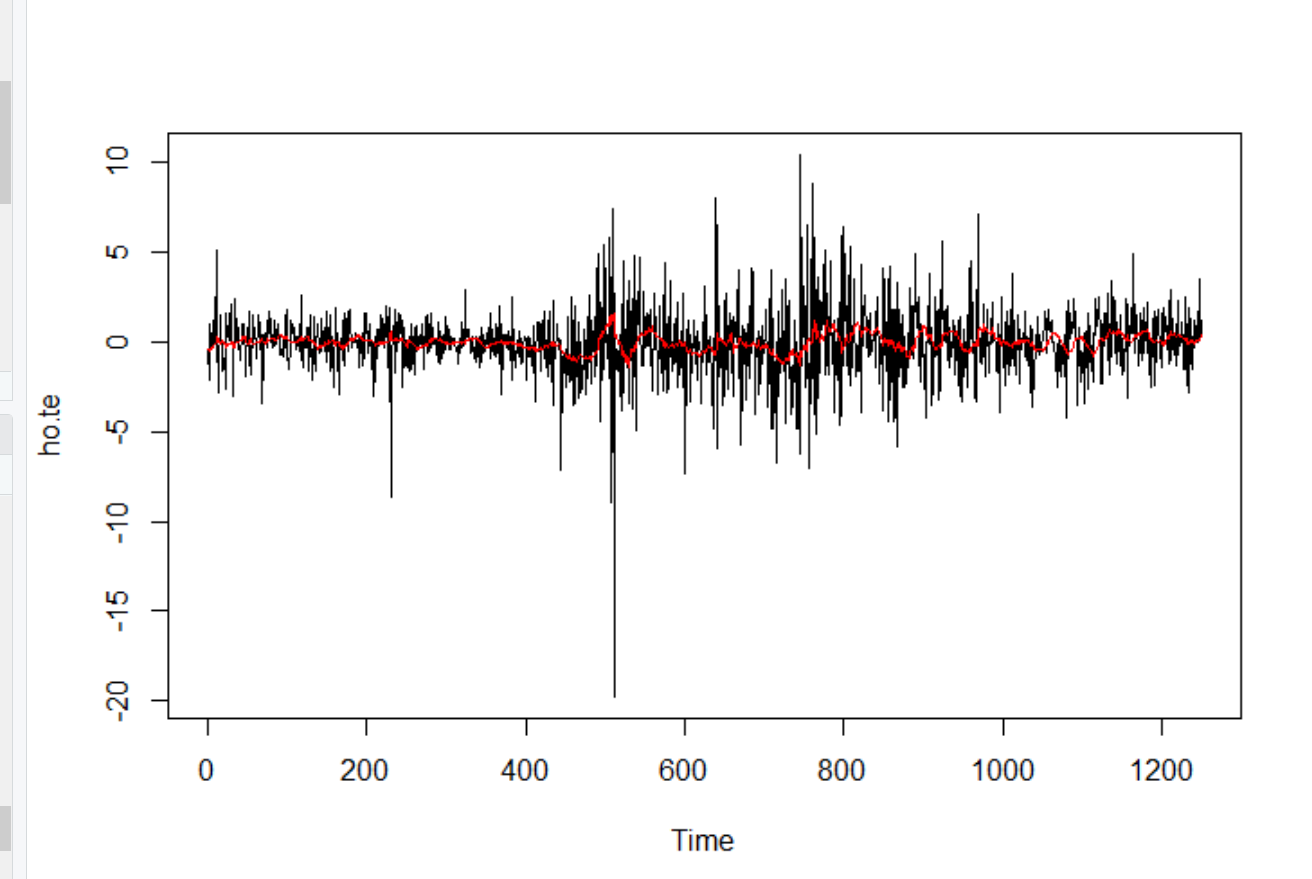
****

****

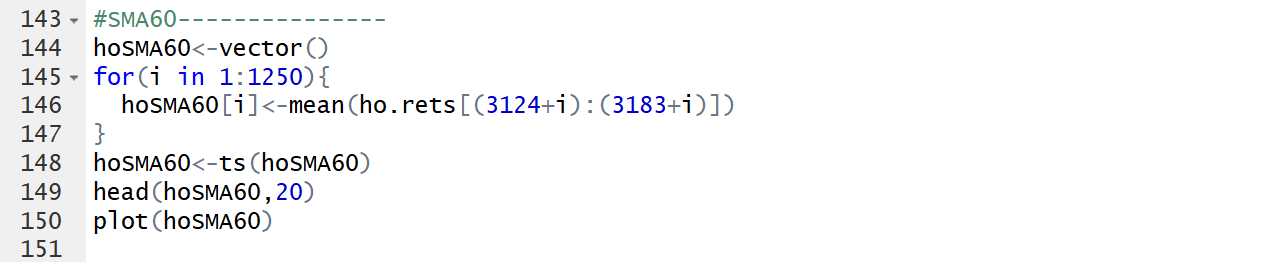
Simple Moving Average Lag=m

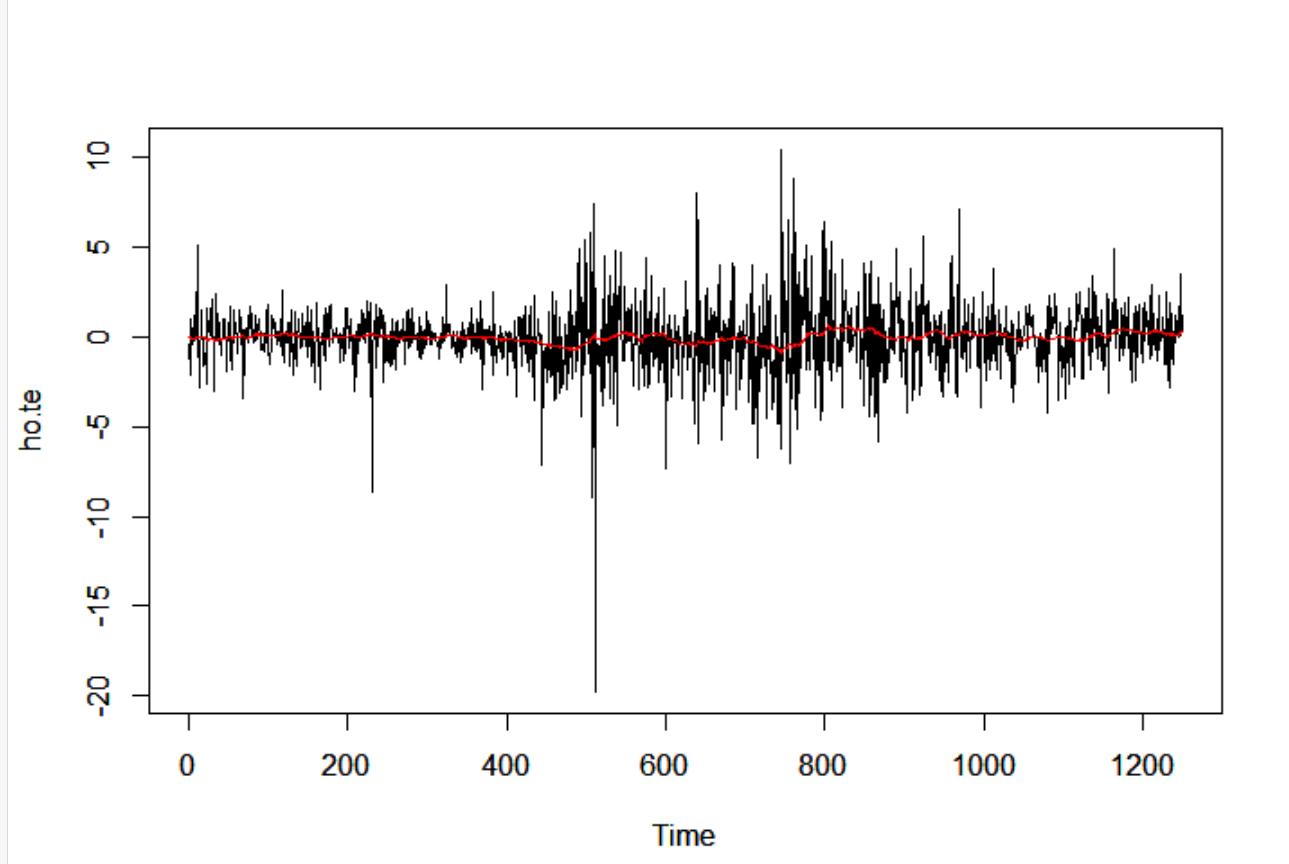
Simple Moving Average(lag=20)

****

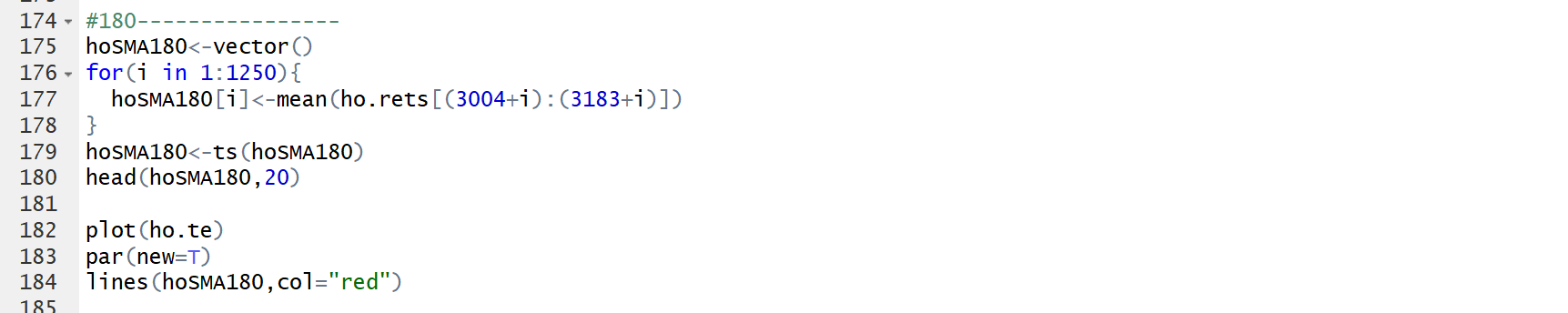
****

Simple Moving Average(lag=60)

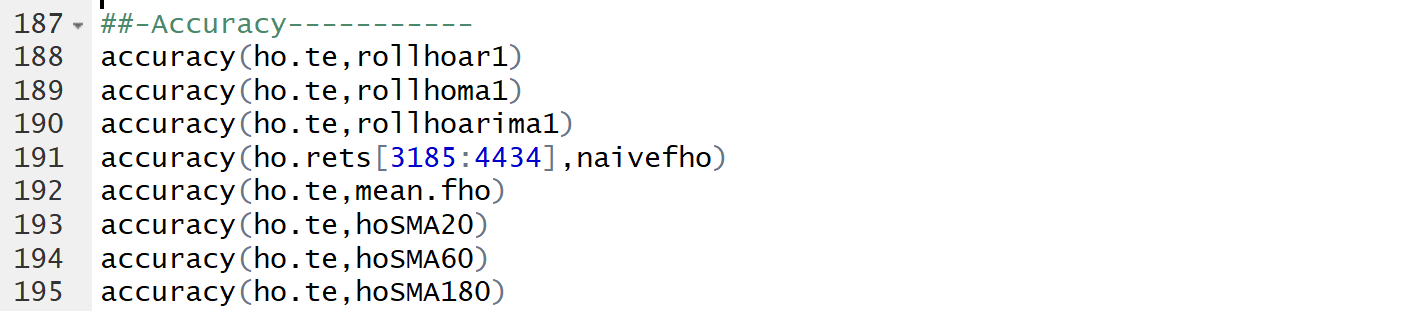


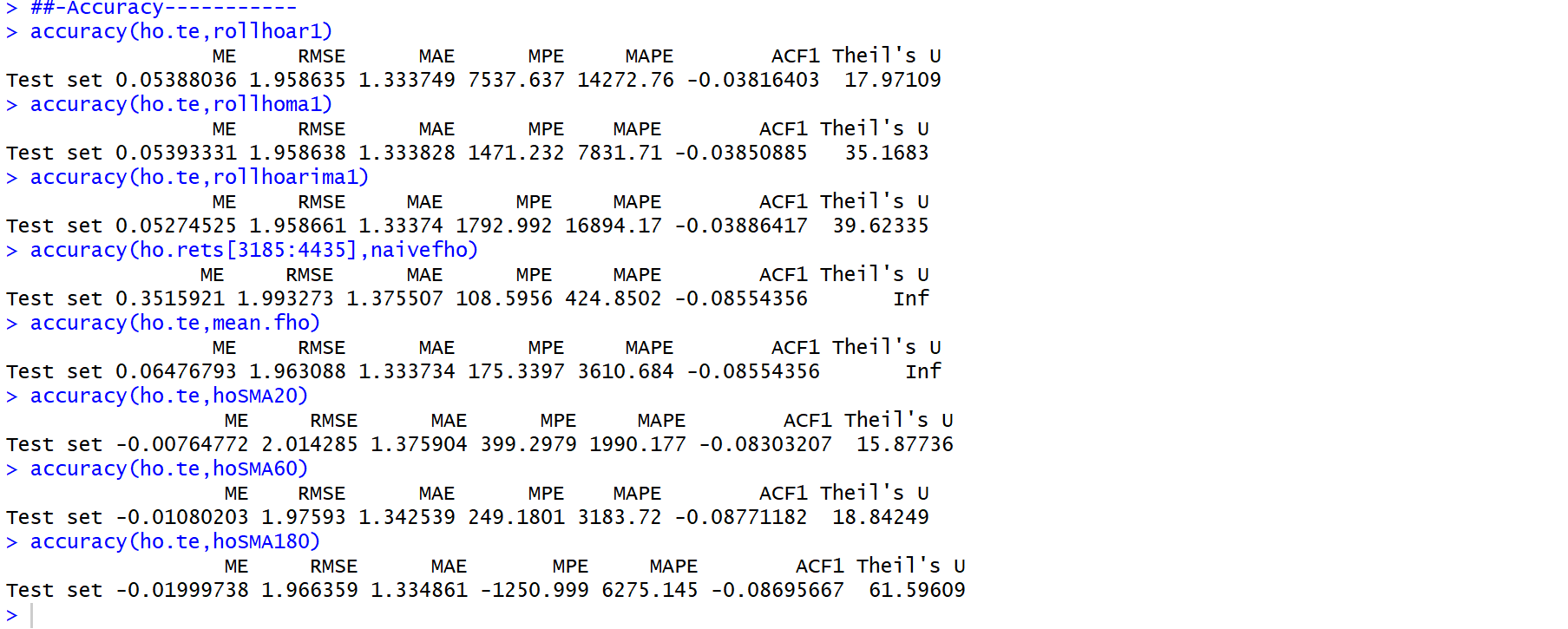
****

Simple Moving Average(lag=180)

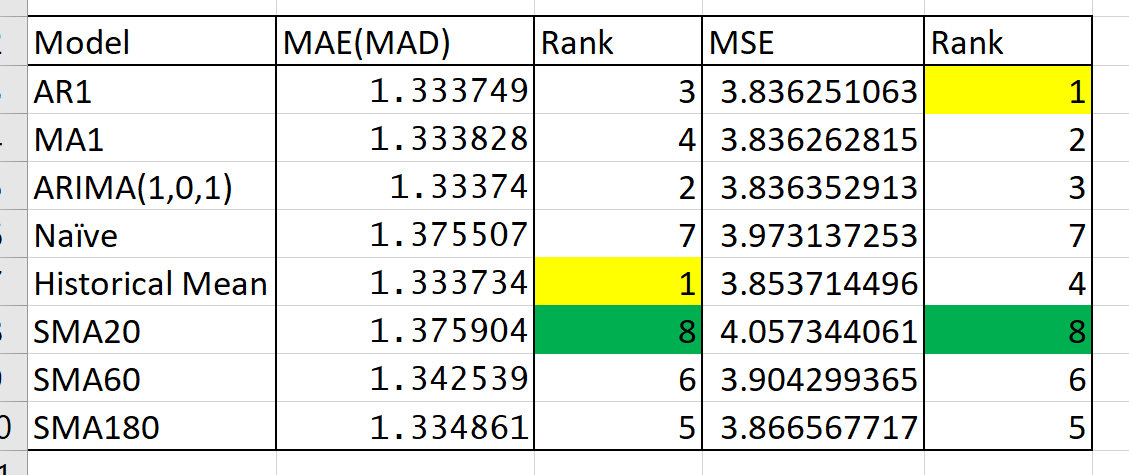


Accuracy





Based on the R output, the table below can be made



This table can be driven. We usually prefer smaller value of MAE(MAD) and MSE. Since, smaller value of it indicates, a model has more capability of capturing data. Meanwhile, higher error signify that the variation of the data could not be captured by model’s predictors.

Based on MAE(MAD)

Best Model: Historical mean model

Worst Model: Simple Moving Average Model(lag=20)

Based on MSE

Best Model: AR1 model

WORST Model: Simple Moving Average model(lag=20)

Qb(ii)

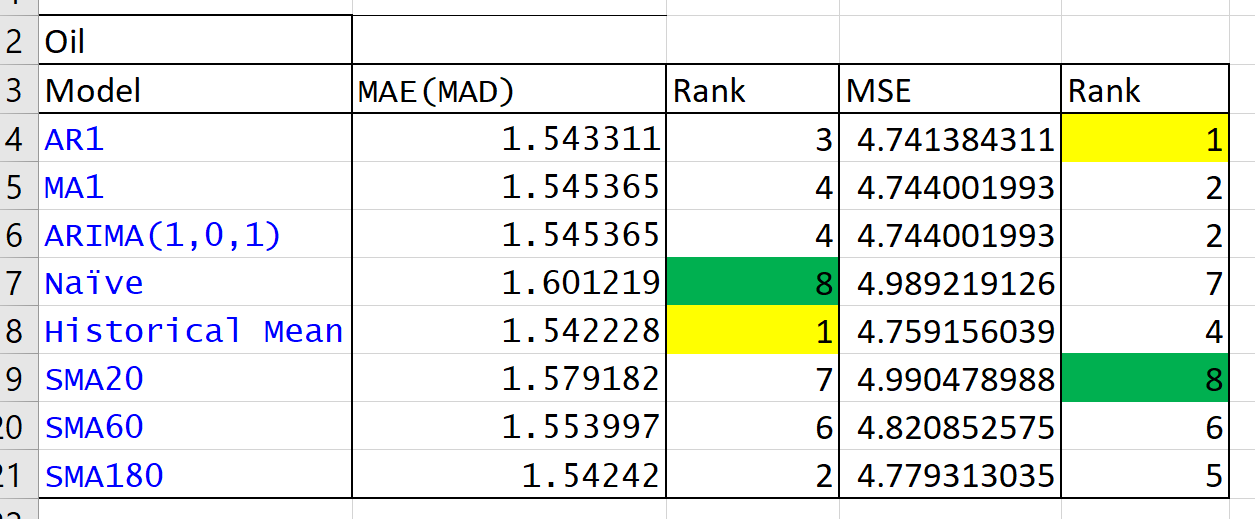
**MSE** and **MAD** are not the same error metrics, however as its smaller value, the model’s performance is considered as good. One of the reasons of it, based on the above 3 accuracy tables driven from R output, all the values are different each other resulted from mathematical differences. Even though, MSE and MAE(MAD) are closely related in terms of maximum likelihood estimation, the assumption of error term is a slightly different. Both two have own unique characteristic.

**MSE**: When MSE is minimized, squared error is minimized. Minimization of squared error equals to least square method where we assume maximum likelihood estimation with assumption of error term following the normal distribution.

**MAD**: When MAD is minimized, absolute error is also minimized. Minimization of absolute error is equal to the assumption of maximum likelihood estimation where error term is following the Laplace distribution. Generally speaking, the Laplace distribution has wider tails of distribution, so it observes more outliers. When we want to evaluate data that have a lot of outliers, or when we want to get robust accuracy evaluation that would not be affected by outliers, MAD might be better than MSE.

Qb(iii)

No. As a model becomes complex always there are high likelihood of facing with overfitting problem. In other words, model would have enough capability of predicting in-sample data whereas, that model does not have enough ability to forecast out-sample data.



For example, from the accuracy table of crude oil forecasting.

The value of MAE(MAD) from ARIMA (1,0,1) is larger than that value of Historical mean model. This indicates that complexed model is not always better that simple model. Meanwhile, based MSE, AR1 model shows higher performance compared to simple model. These two differences indicate, if complex model is better than simple model is depending on the which error metrics that you use, the way of building the model (including some parameters). However, for construction of complex model, as I stated in the above, we always must pay attention to over training of model.