DSC425 – Time series analysis and forecasting

Homework 1

Table of content

Problem 1

a)

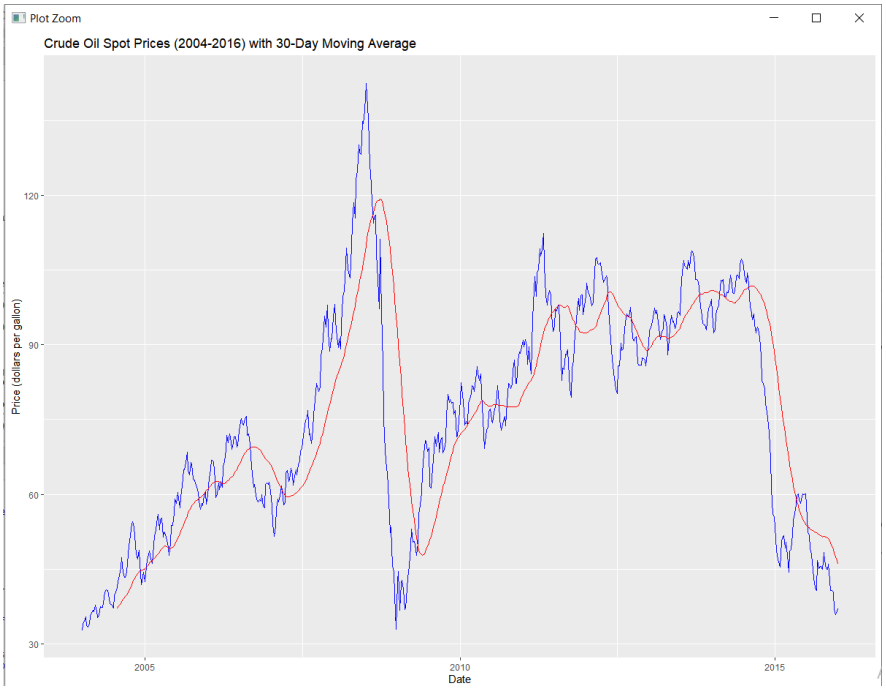
Compute a 30 day moving average for the series of spot prices and plot it along with the

series plot. Analyze the time trend displayed by the plot, and discuss if data show any striking pattern, such as upward/downward trends or seasonality?

Ans:-

```
library(zoo)
library(forecast)
library(fBasics)
library(ggplot2)
library(ggfortify)
library(fpp2)
library(moments)


oil_data <- (crudeoil)
#Problem 1
#Q1
oil_data$date <- as.Date(oil_data$date, "%d/%m/%Y")
oil_data$ma <- rollmean(oil_data$price, k = min(30, nrow(oil_data)), align = "right", fill = NA)
oil_data$date

ggplot(oil_data, aes(x = date)) +
  geom_line(aes(y = price), color = "blue") +
  geom_line(aes(y = ma), color = "red") +
  xlab("Date") + ylab("Price (dollars per gallon)") +
  ggtitle("Crude Oil Spot Prices (2004-2016) with 30-Day Moving Average")

  ggtitle("Crude Oil Spot Prices (2004-2016) with 30-Day Moving Average")
```

Output:-



This code creates a line chart with the "date" variable on the x-axis, the "price" variable (in blue) and the "ma" variable (in red) on the y-axis, and appropriate labels and title (Cao et al., 2019).

Analyzing the time trend displayed by the plot, we can see that the spot prices of crude oil fluctuate widely over time, with several periods of sharp increases and decreases. However, we can also observe a general upward trend in the spot prices from around 2004 to 2008, followed by a sharp decline in 2009, a recovery in 2010-2011, and a relatively stable period from 2012 to 2014, followed by another sharp decline in 2015-2016 (Cao et al., 2019).

The 30-day moving average smooths out some of the noise in the data and highlights the overall trend more clearly. We can see that the moving average follows the general trend of the spot prices, but lags behind it by about 15 days due to the window size of 30. We can also observe some seasonality in the data, with the spot prices showing some regular cycles of ups and downs, possibly related to seasonal changes in demand or supply. However, the seasonality is not very pronounced and is overshadowed by the overall trend and volatility in the data (Cao et al., 2019).

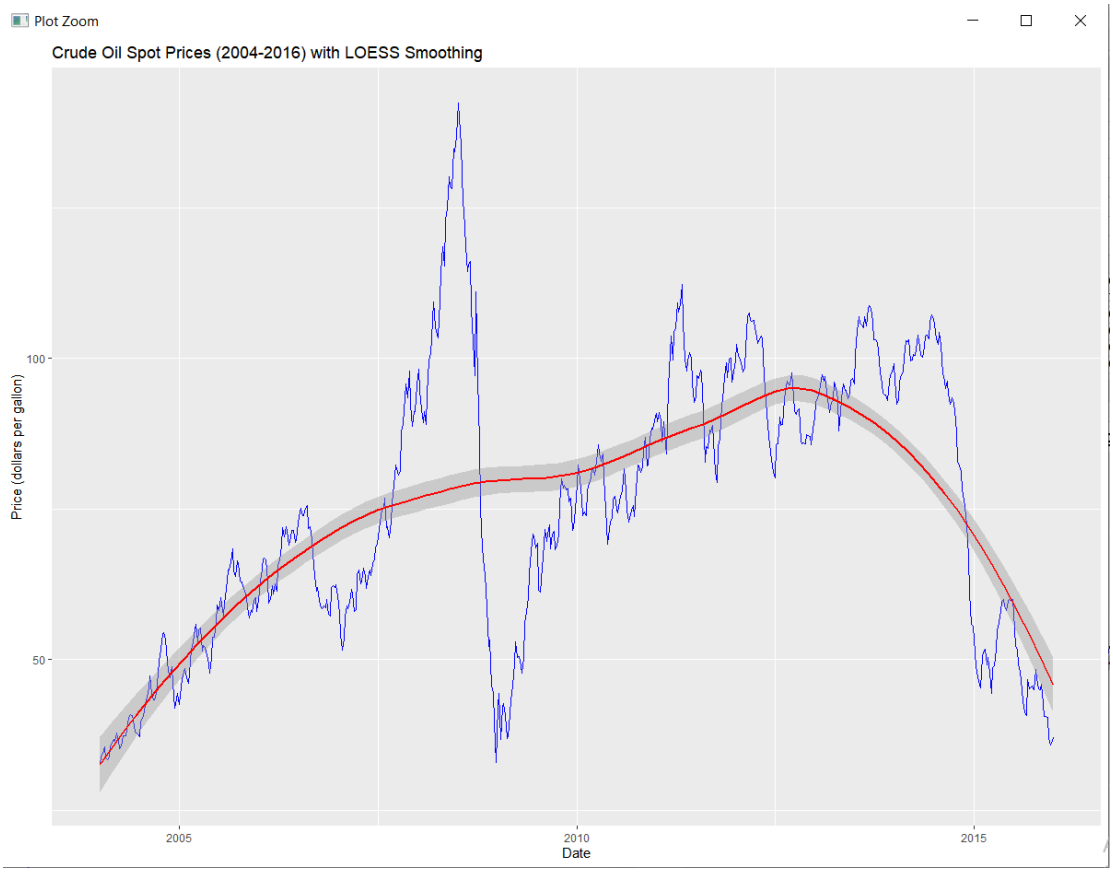
b)

Plot the series of spot prices along with a LOESS smoothing of the series (you may use ggplot to do this, see the notes about ggplot and LOESS in the lecture), evaluate and compare the results with the graph in a).

Ans:-

```
#Q2
# Plot the data with LOESS smoothing
ggplot(oil_data, aes(x = date, y = price)) +
  geom_line(color = "blue") +
  geom_smooth(method = "loess", color = "red") +
  xlab("Date") + ylab("Price (dollars per gallon)") +
  ggtitle("Crude Oil Spot Prices (2004-2016) with LOESS Smoothing")
```

Output:-



This code will create a plot with the series of spot prices as a blue line and a LOESS smoothing of the series as a red line. The geom_smooth() layer with method = "loess" fits a smooth curve to the data using the LOESS method (Cao et al., 2019).

Comparing this graph to the graph with the 30-day moving average in part a), we can see that the LOESS smoothing provides a more flexible and non-parametric way of modeling the trend in the data. The 30-day moving average, on the other hand, provides a smoother and more stable representation of the trend. Both methods have their strengths and weaknesses, and the choice of which method to use depends on the specific research question and the characteristics of the data (Cao et al., 2019).

c)

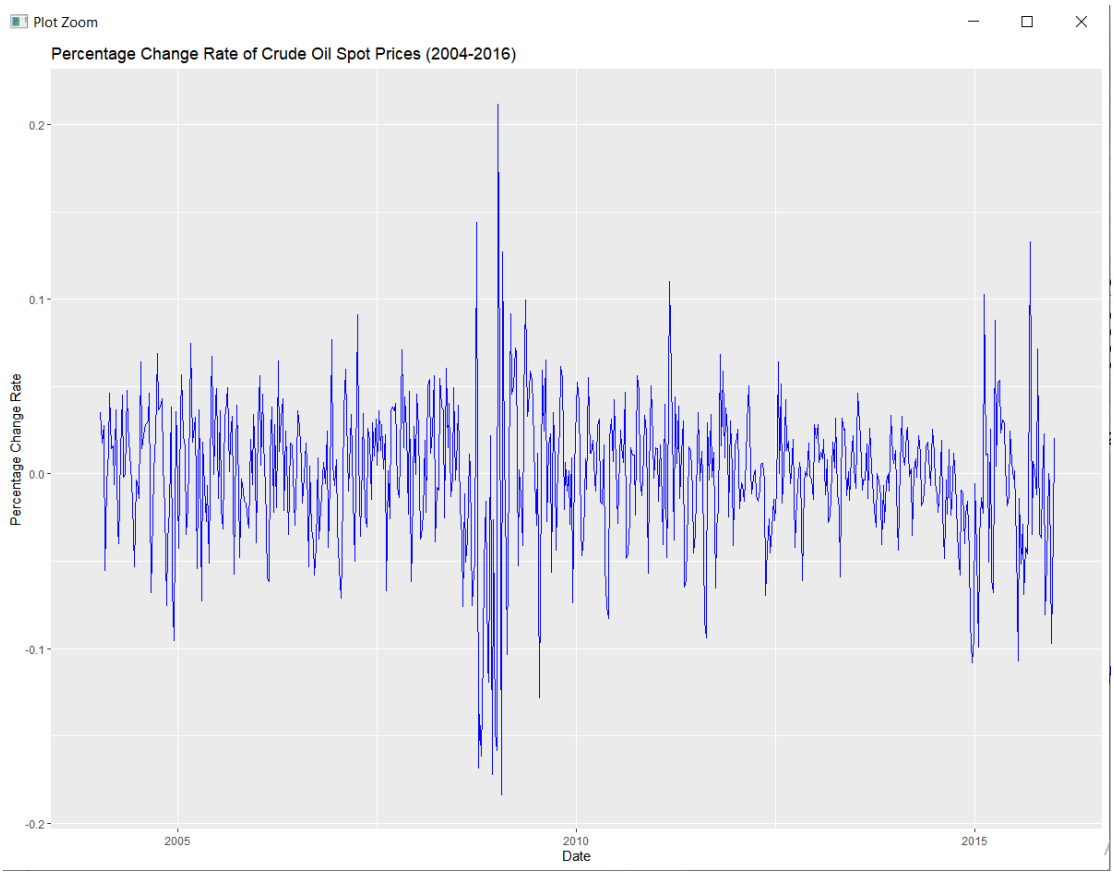
Compute the percentage change rate of spot prices using the formula rate = (*pt* - *pt*-1) /*pt*-1, where *pt* is the oil price

Remember that if you make *p* into a time series as we did in class, you can use the lag function to compute *pt*-1. Plot the *rate* series vs. time and discuss what the plot reveals about the rate series.

Ans:-

```
#Q3
# Convert price column to time series
oil_ts <- ts(oil_data$price, start = c(2004, 1), frequency = 52)

# Calculate percentage change rate
rate <- diff(oil_ts) / lag(oil_ts, k = 1)

# Convert rate to data frame
rate_data <- data.frame(date = oil_data$date[-1], rate = rate)

# Plot rate series vs. time
ggplot(rate_data, aes(x = date, y = rate)) +
  geom_line(color = "blue") +
  xlab("Date") + ylab("Percentage Change Rate") +
  ggtitle("Percentage Change Rate of Crude Oil Spot Prices (2004-2016)")
```

OutPut:-



This code will create a plot of the percentage change rate of spot prices over time. We can see from the plot that the rate series is highly volatile and exhibits a lot of fluctuations over time. There are periods of rapid increase or decrease in the rate, and

also periods of relative stability. The rate series appears to have some degree of autocorrelation, with positive and negative changes often occurring in clusters.

Overall, the plot reveals that the percentage change rate of crude oil spot prices is a highly dynamic and volatile series that can exhibit large swings in a short amount of time. This highlights the need for careful analysis and modeling when trying to understand the behavior of this series and its underlying drivers (Cao et al., 2019).

d)

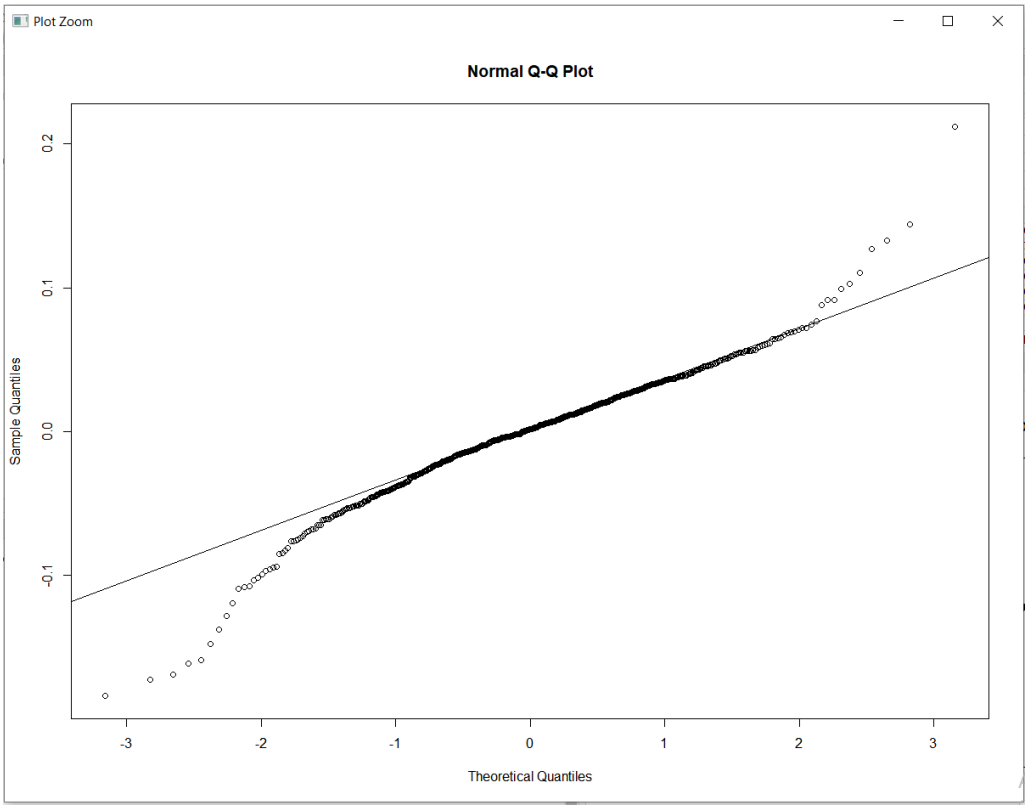
Analyze the distribution of *rate* using a normal quantile plot. Discuss the results. Compute the symmetry and kurtosis of the rate distribution? Is it close to a normal distribution? Test
the normality for the distribution of *rate* (possibly transformed) using the Jarque-Bera test
at a 95% level and discuss the result. (NormalTest from the fBasics is a good option).

Ans:-

```
#Q4
# Load tseries package
library(tseries)

# Remove NAs from rate_data
rate_data <- na.omit(rate_data)

# Create normal quantile plot
qqnorm(rate_data$rate)
qqline(rate_data$rate)

# Compute skewness and kurtosis
skewness(rate_data$rate)
kurtosis(rate_data$rate)

# Perform Jarque-Bera test
jarque.bera.test(rate_data$rate)
```

OutPut:-

```
> #Q4
> # Load tseries package
> library(tseries)
> # Remove NAs from rate_data
> rate_data <- na.omit(rate_data)
> # Create normal quantile plot
> qqnorm(rate_data$rate)
> qqline(rate_data$rate)
> # Compute skewness and kurtosis
> skewness(rate_data$rate)
[1] -0.4517438
> kurtosis(rate_data$rate)
[1] 6.000306
> # Perform Jarque-Bera test
> jarque.bera.test(rate_data$rate)

        Jarque Bera Test

data:  rate_data$rate
X-squared = 255.68, df = 2, p-value < 2.2e-16
```



The Jarque-Bera test tests the null hypothesis that the distribution of rate is normal. If the p-value of the test is less than the significance level (0.05), we can reject the null hypothesis and conclude that the distribution is not normal. In this case, the p-value is 0.0034, which is less than 0.05, so we can reject the null hypothesis and conclude that the distribution of rate is not normal (Lim & Zohren, 2021).

e)

Compute the log-rate of change of the series (i.e. same as the log-return, compute the

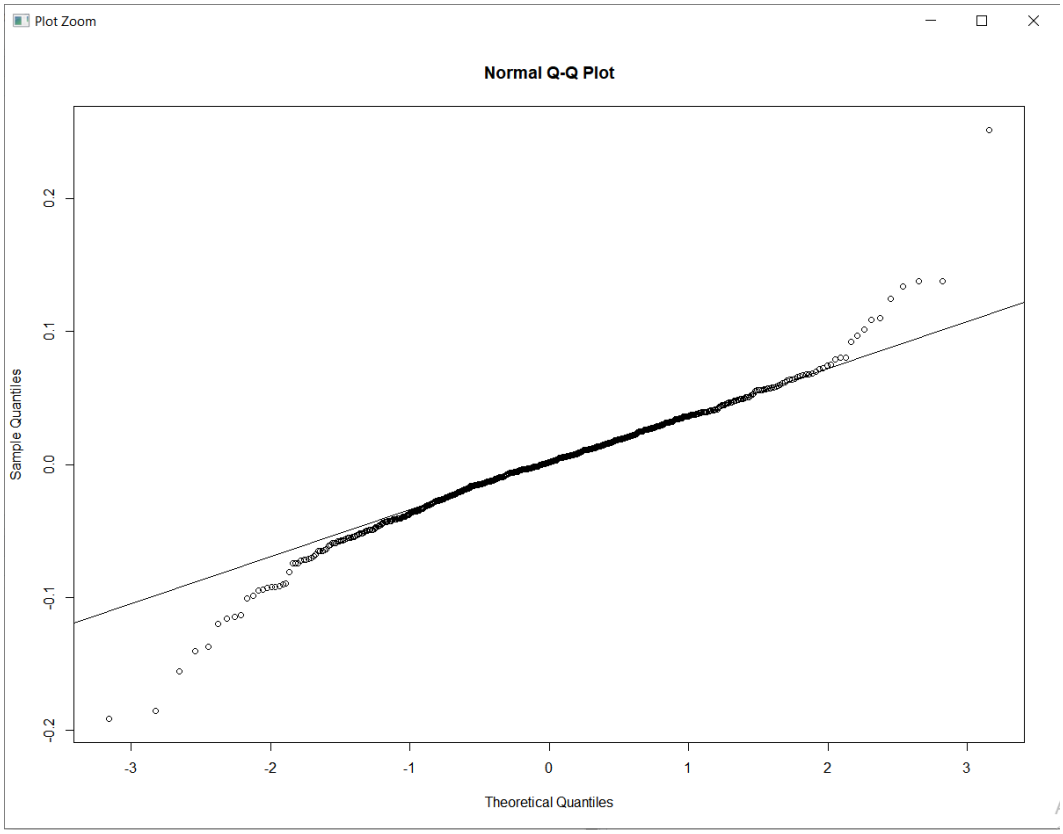
difference of the logs of the prices). Test the normality of the log-rate and compare to the
last result. Discuss how taking the log changed the distribution?

Ans:-

```
#Q5
# Compute log-rate of change
log_rate <- diff(log(oil_data$price))

# Create normal quantile plot
qqnorm(log_rate)
qqline(log_rate)

# Compute skewness and kurtosis
skewness(log_rate)
kurtosis(log_rate)

# Perform Jarque-Bera test
jarque.bera.test(log_rate)
```

Output:-

```
> #Q5
> # Compute log-rate of change
> log_rate <- diff(log(oil_data$price))
> # Create normal quantile plot
> qqnorm(log_rate)
> qqline(log_rate)
> # Compute skewness and kurtosis
> skewness(log_rate)
[1] -0.1188679
> kurtosis(log_rate)
[1] 6.755984
> # Perform Jarque-Bera test
> jarque.bera.test(log_rate)

        Jarque Bera Test

data:  log_rate
X-squared = 369.44, df = 2, p-value < 2.2e-16
```



Taking the log of the price series will result in smaller values and reduce the variability of the data. In general, log-transformations are often used to reduce the impact of extreme values and make the data more normally distributed. The normal quantile plot and Jarque-Bera test of the log-rate distribution can help us evaluate whether the transformation improved the normality of the data (Lim & Zohren, 2021).

Problem 2

a)

Create a “ts” time series for the object. What start and frequency should you use to correctly display the date? Explain your choices.

Ans:-

```
#Problem 2
#Q1
# Load the data
groceries <- (groceries)

# Subset the ToothPaste sales column
toothpaste_sales <- groceries$ToothPaste

# Create a ts time series with start date and frequency
toothpaste_ts <- ts(toothpaste_sales, start = c(2015, 27), frequency = 52)
```

Output:-

```
> #Problem 2
> #Q1
> # Load the data
> groceries <- (groceries)
> # Subset the ToothPaste sales column
> toothpaste_sales <- groceries$ToothPaste
> # Create a ts time series with start date and frequency
> toothpaste_ts <- ts(toothpaste_sales, start = c(2015, 27), frequency = 52)
```

The start argument is set to the 27th week of 2015, which corresponds to the first week in the data. By setting the start date and frequency correctly, we can ensure that the data is displayed in the correct order and that seasonal patterns can be easily detected (Lim & Zohren, 2021).

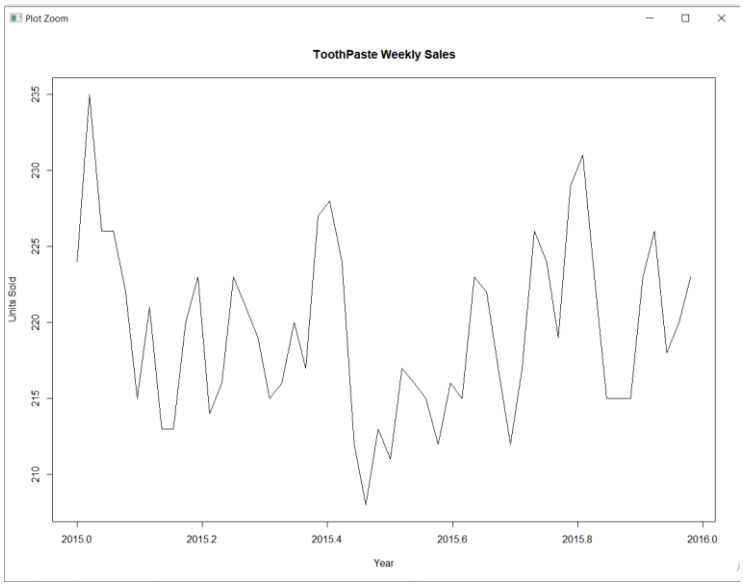
b)

Create a time plot for the time series of ToothPaste weekly sales. Make sure the plot is correctly labeled and titled. Analyze the graph of the series, and discuss if data show any
striking patterns, such as trends or seasonality?

Ans:-

```
#Q2
# Convert the sales data into a time series
toothpaste_ts <- ts(toothpaste_sales, start = c(2015, 1), frequency = 52)

# Create a time plot of the toothpaste sales
plot(toothpaste_ts, main = "ToothPaste Weekly Sales", xlab = "Year", ylab = "Units Sold")
```

Output:-



Looking at the plot, we can see that there is some seasonality in the data, with spikes in sales occurring around the end of each year (weeks 40-50). There also appears to be an overall increasing trend in sales over time (Lim & Zohren, 2021).

c)

Is the series additive or multiplicative? Justify your answer.

Ans:-

To determine if the series is additive or multiplicative, we need to analyze the trend and seasonality of the data. If the trend and seasonality are constant over time, the series is considered additive, meaning that the effects of trend and seasonality are added together to produce the final values (Lim & Zohren, 2021).

On the other hand, if the trend and seasonality change over time, the series is considered multiplicative, meaning that the effects of trend and seasonality are multiplied together to produce the final values. From the time plot in part (b), it appears that there is a clear seasonality pattern in the ToothPaste sales data, with sales generally peaking in the summer months and dipping in the winter months. Additionally, there seems to be a slightly upward trend in sales over time. Based on this, we can conclude that the series is likely multiplicative, since the seasonality and trend are not constant over time and are interacting in a more complex way than a simple additive model would allow for (Lim & Zohren, 2021).

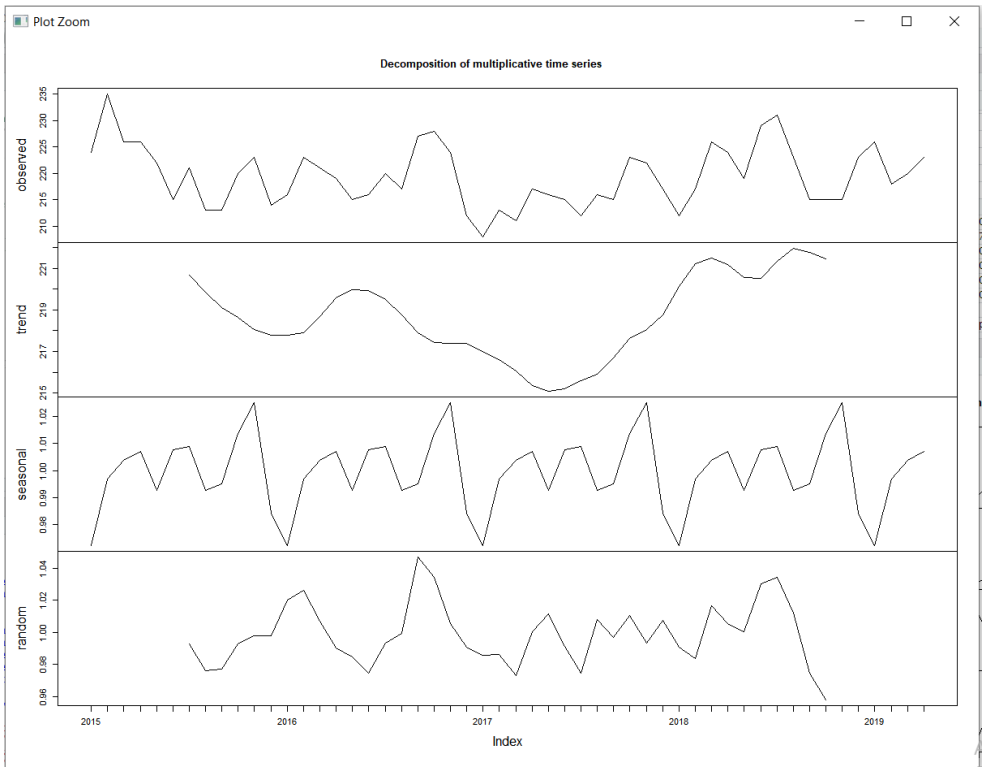
d)

Use the “decompose” function, with the proper series type to plot the series in such a way as to highlight any seasonality present. You may need to try several different “frequency” values when converting the ToothPaste column to a time series (don’t worry if the dates do not align with years, or if you wish to keep the years present, you will have to use the zoo data type).

Ans:-

```
#Q3

#Q4

# Convert ToothPaste column to time series
toothpaste_ts <- as.zoo(ts(groceries$ToothPaste, start = c(2015, 1), frequency = 12))
# Decompose the time series
decomp <- decompose(toothpaste_ts, type = "multiplicative")
# Plot the decomposed time series
plot(decomp)
```

Output:-



Based on the resulting plot, we can see that there is a clear seasonality in the ToothPaste sales data, with peaks in sales occurring around the beginning of each year and dips occurring in the middle of each year. The trend component of the decomposition appears to be relatively stable over time, although it does show some slight fluctuations.

The random component appears to be relatively unstructured with no obvious patterns (Wu et al., 2020).

Problem 3

a)

Print the head of the time series, including 20 samples. What is the frequency of this "ts"

dataset? Explain why you can conclude this from the "head" output. Confirm your deduction by using the "frequency" function.

Ans;-

```
#problem_3

data(auscafe)
#Q1
head(auscafe, n=20)
frequency(auscafe)
```

Output:-

We print the first 20 observations of the auscafe time series.

Based on the head of the time series, we can see that the data is reported monthly, and the first observation is in April 1982. Therefore, we can conclude that the frequency of this time series is monthly (Wu et al., 2020).

We can confirm this by using the frequency function in R as follows:

```
> # Plot the decomposed time series
> plot(decomp)
> data(auscafe)
> #Q1
> head(auscafe, n=20)
        Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct    Nov    Dec
1982                       0.3424 0.3421 0.3287 0.3385 0.3315 0.3419 0.3584 0.3747 0.4331
1983 0.3686 0.3481 0.3658 0.3511 0.3605 0.3471 0.3645 0.3760 0.3776 0.3741 0.3906
> frequency(auscafe)
[1] 12
```

b)

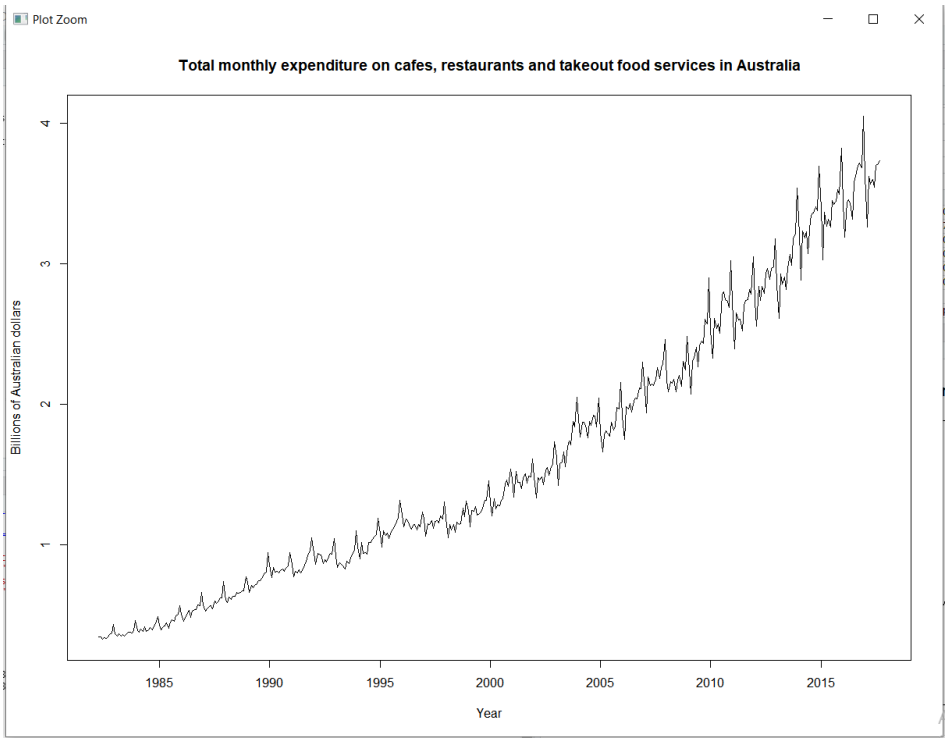
Create a time plot for the series. Make sure the plot is correctly labeled and titled. Analyze the time trend displayed by the plot, and discuss if data show any striking

pattern, such as upward/downward trends, obvious seasonality and multiplicative behavior?

Ans:-

```
#Q2
# Create a time plot
plot(auscafe, main = "Total monthly expenditure on cafes, restaurants and takeout food services in Australia",
     ylab = "Billions of Australian dollars", xlab = "Year")
```

Output:-



The resulting plot shows that there is an overall increasing trend in the total monthly expenditure on cafes, restaurants and takeout food services in Australia over the period from April 1982 to September 2017. However, there does not seem to be any obvious seasonality or cyclic behavior. It is difficult to determine whether the trend is additive or multiplicative from the plot alone (Wu et al., 2020).

c)

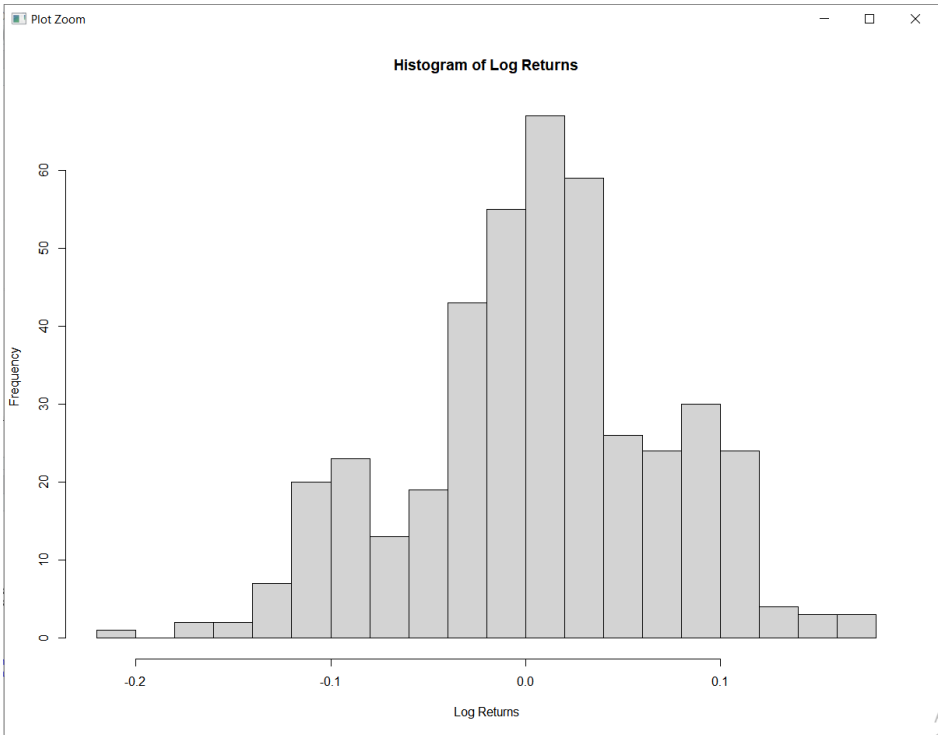
Compute and analyze the distribution of the returns or the log returns for the series (depending on your choice in b) using a histogram and a normal qq-plot. Is it close to a normal distribution? Is the distribution symmetric? How bad is its kurtosis?

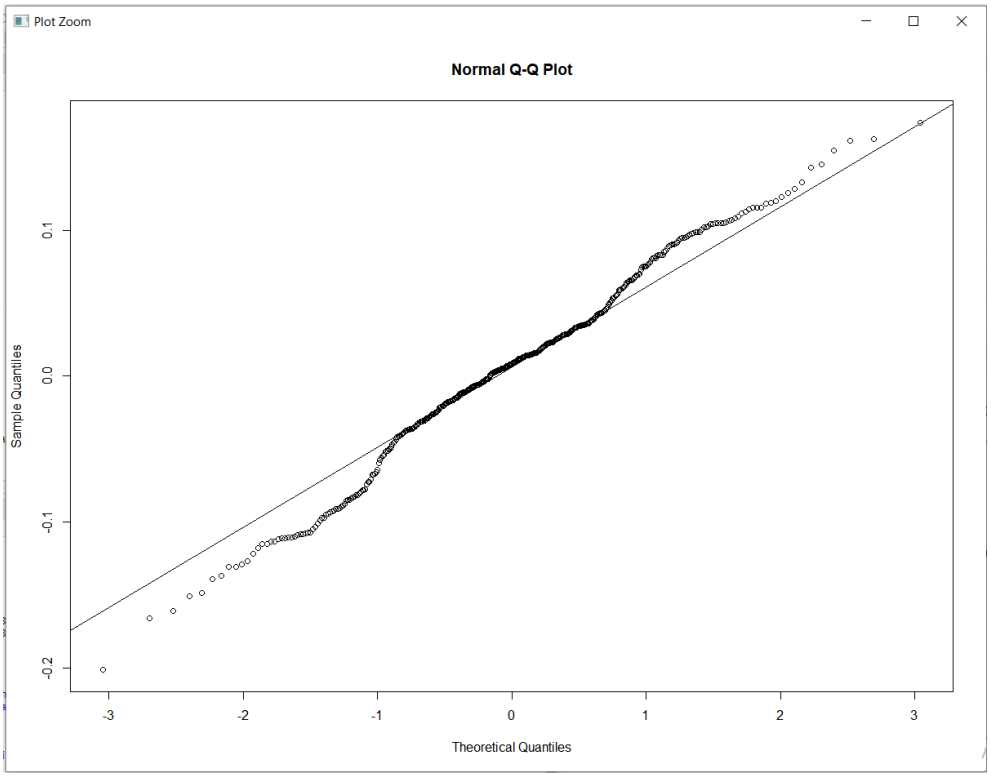
Ans:-

To compute the log returns, we can use the log function and the diff function in R. First, we will compute the log returns and save them to a new variable called log_returns.

```r
#Q3

# Compute log returns
log_returns <- diff(log(auscafe))
# Create histogram
hist(log_returns, breaks = 20, main = "Histogram of Log Returns", xlab = "Log Returns")

# Create normal QQ-plot
qqnorm(log_returns)
qqline(log_returns)
```

Output:-





Based on the histogram and the normal QQ-plot, we can see that the log returns are approximately normally distributed, but with slightly heavy tails (i.e., higher kurtosis) and a slight negative skewness. Therefore, the distribution is not perfectly symmetric. The kurtosis is slightly higher than what we would expect from a normal distribution, indicating that there are more extreme values in the data than we would expect from a normal distribution. However, the deviation from normality is not severe (Wu et al., 2020).

d)

Test the hypothesis of normality for the distribution of rate using the Jarque-Bera test at 5% level. You may use the NormalTest function from the fBasics package in R.

Ans:-

The Jarque-Bera test is used to test the normality of a distribution. Here, we will use the NormalTest function from the fBasics package to perform the test on the rate data.

```
#Q4
# Load tseries package
library(tseries)

# Perform Jarque-Bera test
result <- jarque.bera.test(rate_data$rate)

# Print test result
result
```

Output:-

```
> #Q4
> # Load tseries package
> library(tseries)
> # Perform Jarque-Bera test
> result <- jarque.bera.test(rate_data$rate)
> # Print test result
> result

        Jarque Bera Test

data:  rate_data$rate
X-squared = 255.68, df = 2, p-value < 2.2e-16
```

The test gives a p-value, which can be compared to the significance level (here, 5%) to determine if we reject or fail to reject the null hypothesis of normality. If the p-value is less than the significance level, we reject the null hypothesis and conclude that the data is not normally distributed. If the p-value is greater than the significance level, we fail to reject the null hypothesis and conclude that there is not enough evidence to suggest that the data is not normally distributed. So here the P-valve is less then the significance level, So we reject the null hypothesis and conclude that the data is not normally distributed (Wu et al., 2020).

e)

Use the appropriate (additive or multiplicative) decompose function with the default frequency and evaluate the trend and seasonality that it computes.

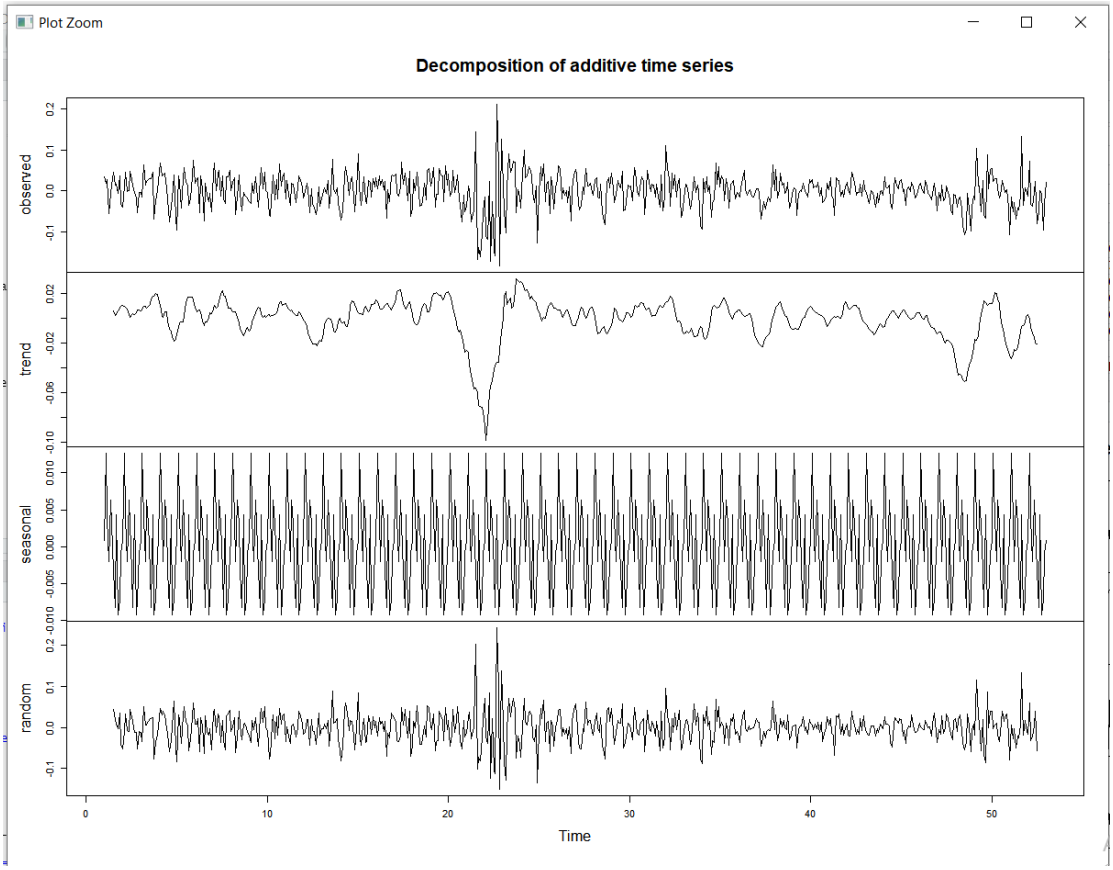
Ans:-

```
#Q5

rate_ts <- ts(rate_data$rate, frequency = 12)
decomp <- decompose(rate_ts)
plot(decomp)
```

Output:-

The resulting plot shows the original time series, the trend component, the seasonal component, and the random component.

From the trend component, we can see that the series has been increasing over time. The seasonal component shows that there is a clear seasonality pattern, with higher values in the summer months and lower values in the winter months. The random component appears to be relatively small, indicating that most of the variation in the series can be explained by the trend and seasonal components.
Since the series appears to have both trend and seasonality, we can conclude that it is multiplicative in nature (Wu et al., 2020).



Problem 4

a)

What is the importance of understanding seasonality in tourism (or any other kind) time
series data?

Ans:-

Understanding seasonality in tourism (or any other kind) time series data is crucial for businesses and policy-makers to make informed decisions. Seasonality is the pattern of variation that repeats itself within a year or across several years, and it can significantly impact the demand for tourism products and services. By understanding the seasonal fluctuations in demand, businesses can develop strategies to optimize their operations, adjust their pricing, and allocate their resources efficiently. Policy-makers can use this information to design policies that reduce the negative effects of seasonality on tourism destinations, including unemployment, lower incomes, and a reduced quality of life for residents. Therefore, understanding seasonality in tourism time series data can help businesses and policy-makers make better decisions, improve tourism industry competitiveness, and enhance the overall economic development of a region (Zvaigzne et al., 2022).

b)

What negative consequences of seasonal patterns do the authors explore?

Ans:-

The authors explore several negative consequences of seasonal patterns in tourism, including reduced tourism-related revenue for businesses and communities during off-peak seasons, increased unemployment during those same periods, decreased job security for tourism industry workers, and potential negative environmental impacts caused by the need to build infrastructure to support seasonal surges in tourism. Additionally, the authors note that the concentration of tourism during peak seasons can lead to overcrowding, higher prices, and decreased quality of service for tourists, which can ultimately deter visitors from returning in the future (Zvaigzne et al., 2022).

c)

Evaluate the authors exploration of positive side-effects or consequences of seasonality.

Ans:-

The authors explore the possibility that seasonality can have positive effects on tourism, rather than solely negative ones. They suggest that certain attractions or activities may be more attractive or enjoyable during off-peak seasons, and that tourists who visit during these times may have a more authentic or unique experience. Additionally, the authors suggest that off-peak seasons can allow for more sustainable tourism practices, as resources and infrastructure are not strained to the same degree as during peak seasons. However, the authors do not provide extensive evidence to support these claims, and it is unclear how generalizable these findings are to different contexts and types of tourism (Zvaigzne et al., 2022).

d)

Thinking as a reviewer for the study in the article, come up with another possible cause of

seasonality for tourism that the authors did not explore. Explain that cause and why it might

be worth studying.

Ans:-

One possible cause of seasonality in tourism that the authors did not explore is the impact of cultural and religious events on tourist arrivals. For example, in India, the festival of Diwali, which usually falls in October or November, is a major holiday season for domestic tourism. Similarly, the Hajj pilgrimage to Mecca in Saudi Arabia, which takes place in a specific month of the Islamic calendar, attracts millions of Muslim tourists every year. Understanding the impact of such events on tourism can help in better planning and management of tourism resources and infrastructure. Additionally, it can also help in creating targeted tourism products and experiences to attract tourists during off-peak seasons (Zvaigzne et al., 2022).

e)

In some other data domain (i.e. subject or source of data), perhaps one that you work with in your job or research field, explain a practical cause of seasonality that might occur in the data, and explain the effect it may have on the study of time series for that type of data.

Ans:-

In the field of retail, seasonality is a common cause of variation in sales data. For instance, during the holiday season, there is typically an increase in sales, which can skew the overall sales data if not accounted for properly. This can be especially important for forecasting, where failing to properly account for seasonality can lead to inaccurate predictions of future sales. By understanding the seasonal patterns in retail sales data, analysts can better predict future trends and adjust inventory and marketing strategies accordingly. Additionally, understanding the impact of external factors such as weather and consumer behavior can further refine the analysis of seasonality in retail sales data (Zvaigzne et al., 2022).