STAT 443: Assignment 1

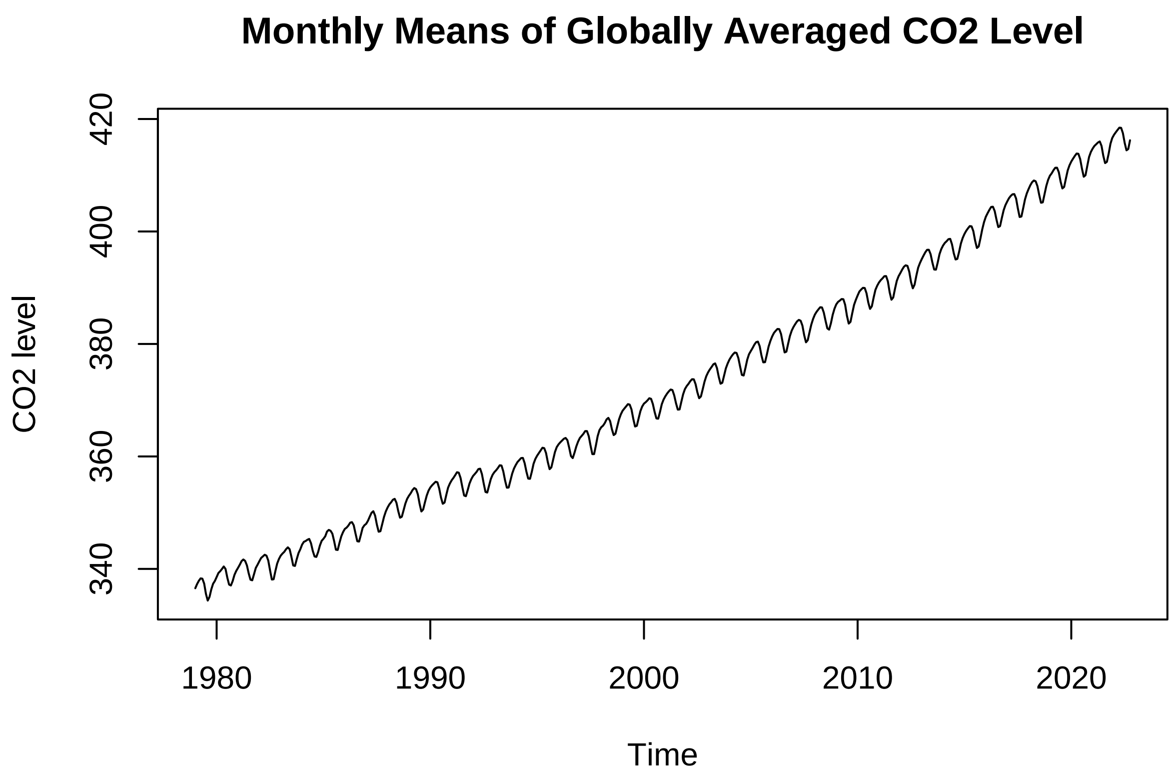
Wenxuan Zan (61336194)

2 Februray, 2022

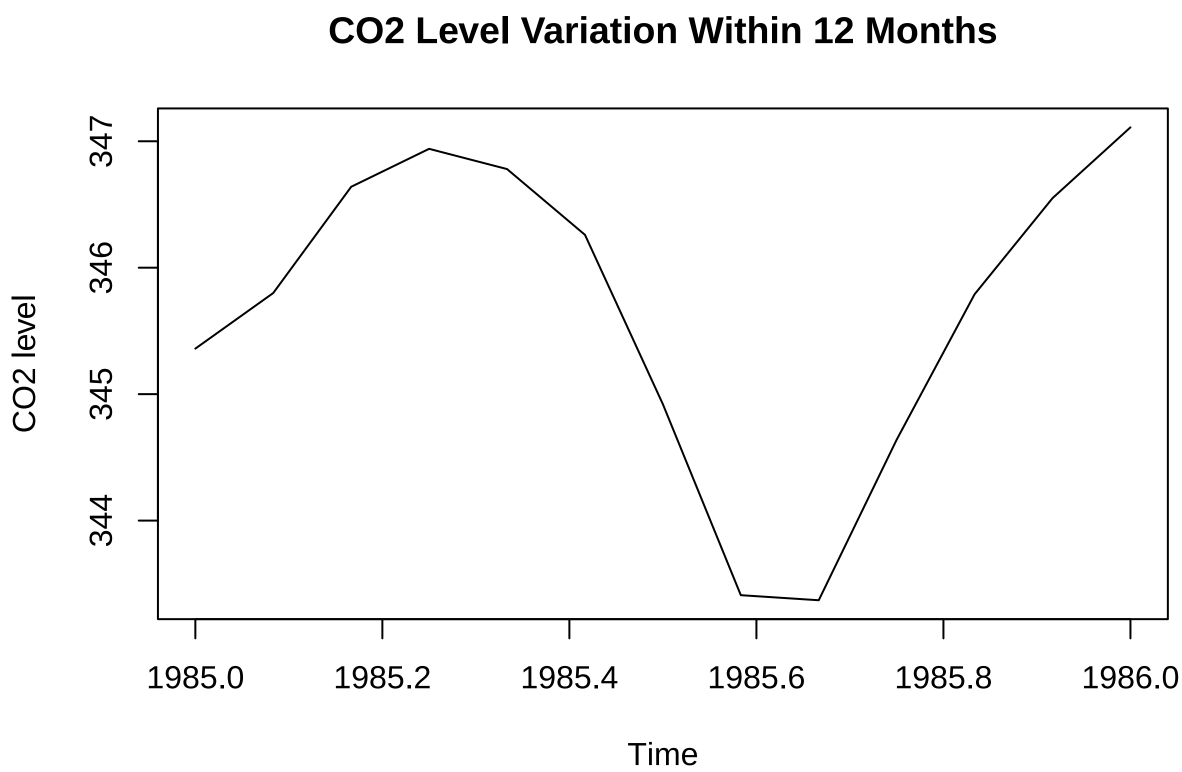
Question 1

a)

```
co2 <- read.csv("co2_mm_gl.csv", header = TRUE, skip = 55)
co2_ts <- ts(co2[,4], start = c(1979,1),frequency = 12)
plot(co2_ts,
     main = "Monthly Means of Globally Averaged CO2 Level",
     ylab = "CO2 level")
```




```
plot(window(co2_ts, start = c(1985,1),end = c(1986,1)),
     main = "CO2 Level Variation Within 12 Months",
     ylab = "CO2 level")
```



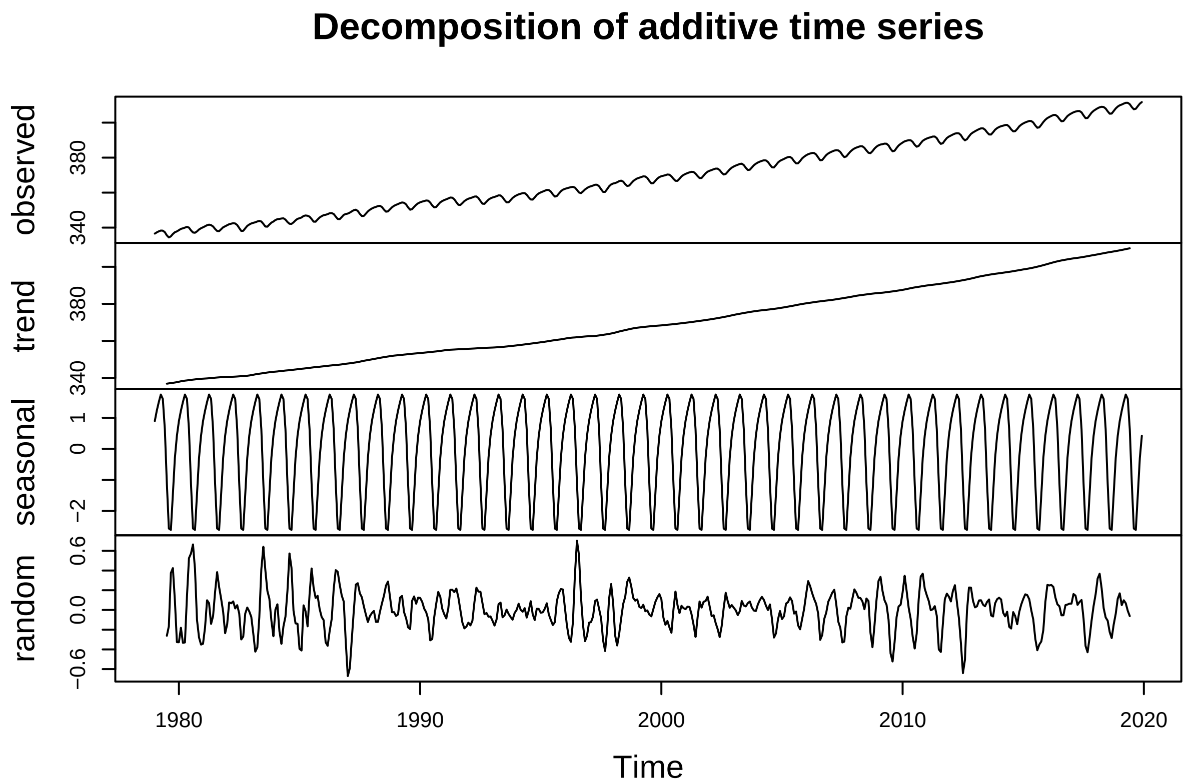
i) The above time series have a clear **upward trend** such that the monthly means of globally averaged CO2 level is increasing every year despite some variations.

ii) There appears to be **seasonal variations** in the monthly CO2 level as when we restrict the plot to display the average CO2 level over a 12-month period, we can clear see the CO2 level is high around March and December, low around June.**An additive model is more suitable** as we can see the seasonal effect remains constant over time and the error is also constant over time, therefore an additive model would be more appropriate than a multiplicative model.

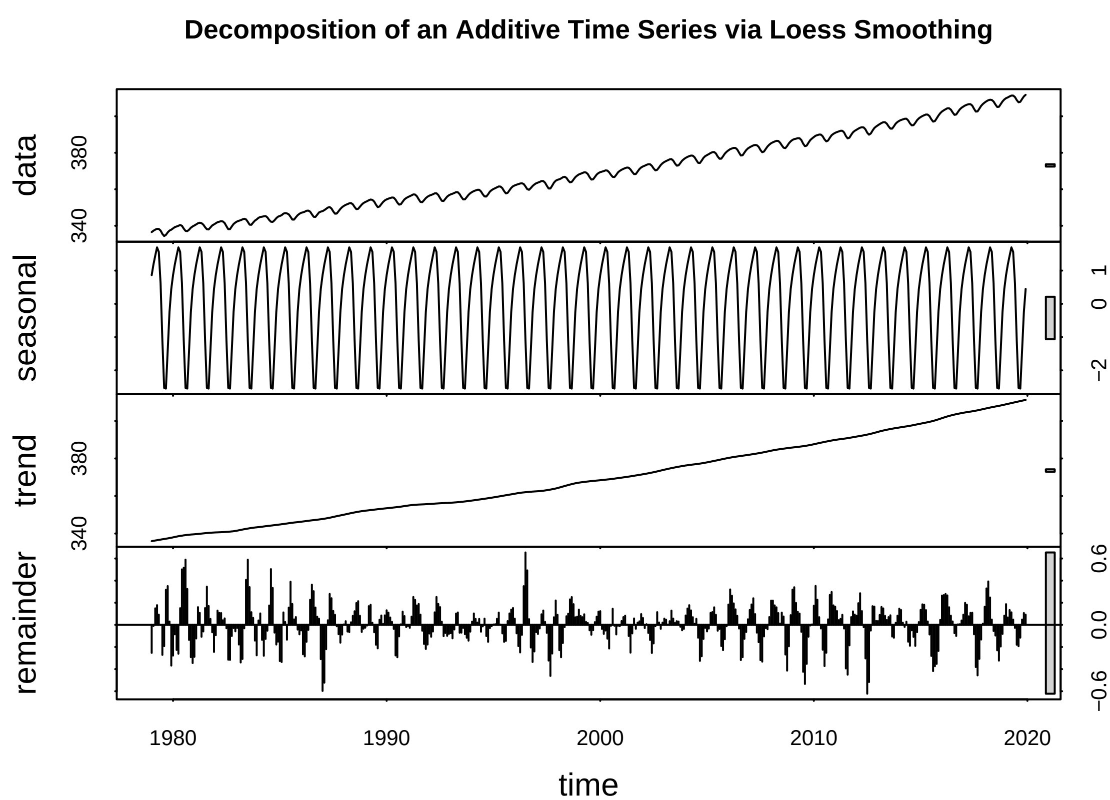
iii) **No**, the series have a clear upward increasing trend therefore it is not stationary.

b)

```
co2_train <- window(co2_ts,start = c(1979,1),end = c(2019,12),frequency = 12)
co2_test <- window(co2_ts, start = c(2020,1), end = c(2022,10), frequency = 12)

co2_train_decom <- decompose(co2_train, type = "additive")
plot(co2_train_decom)
```




```r
co2_train_loess <- stl(co2_train,s.window = "periodic")
plot(co2_train_loess,
     main = "Decomposition of an Additive Time Series via Loess Smoothing")
```



c)

```r
# MA method
ma_trend <- co2_train_decom$trend
ma_seas <- co2_train_decom$seasonal
ma_error <- co2_train_decom$random
# Loess smoothing
loess_trend <- co2_train_loess$time.series[,"trend"]
loess_seas <- co2_train_loess$time.series[,"seasonal"]
loess_error <- co2_train_loess$time.series[,"remainder"]
# creating data frame
lm_data <- data.frame(ma_trend = ma_trend,
                      loess_trend = loess_trend,
                      time = c(1:length(co2_train)))
# Fitted Models
lm_ma <- lm(ma_trend~time, data = lm_data)
lm_loess <- lm(loess_trend~time, data = lm_data)
summary(lm_ma)
```

```
##
## Call:
## lm(formula = ma_trend ~ time, data = lm_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9933 -1.6701 -0.3312  1.2281  4.7441
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.329e+02  1.819e-01  1830.0   <2e-16 ***
## time        1.488e-01  6.433e-04   231.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.953 on 478 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.9911, Adjusted R-squared:  0.9911
## F-statistic: 5.349e+04 on 1 and 478 DF,  p-value: < 2.2e-16
```

```r
summary(lm_loess)
```

```
##
## Call:
## lm(formula = loess_trend ~ time, data = lm_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0771 -1.7471 -0.3894  1.2344  4.9349
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.329e+02  1.832e-01  1817.6   <2e-16 ***
## time        1.491e-01  6.439e-04   231.5   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.028 on 490 degrees of freedom
## Multiple R-squared:  0.9909, Adjusted R-squared:  0.9909
## F-statistic: 5.361e+04 on 1 and 490 DF,  p-value: < 2.2e-16
```

i) Fitted model using MA method:

$$\hat{m_t} = 0.03329 + 0.1488 * t$$

Fitted model using loess smoothing method:

$$\hat{m_t} = 0.03329 + 0.1491 * t$$

ii) The trend component is significant at 95% confidence level under both method.

iii) I think the trend component is a good predictor because the trend is linear over the entire time range of the time series data looking at the trend component plot in part b); second, from the output of the linear models we observe that $R^2 > 0.99$f ro both models. This indicates that majority of the variation within the time series is explained by the trend component.

iv)

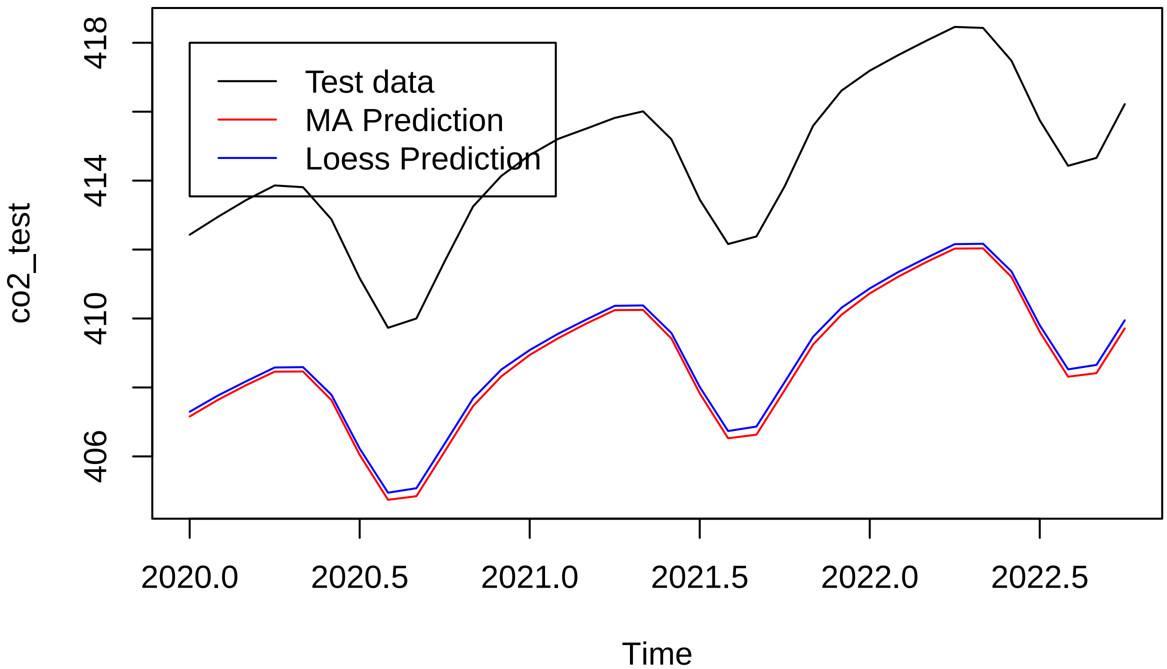
```r
t_test <- data.frame(time = max(lm_data$time) + c(1:length(co2_test)))
ma_trend_pred <- predict(lm_ma, newdata = t_test)
loess_trend_pred <- predict(lm_loess, newdata = t_test)

# estimate seasonal effect
seasonals <- tibble(ma_sea_hat = ma_seas,
                        loess_sea_hat = loess_seas,
                        season = rep(1:12,41)) %>%
  group_by(season) %>%
  summarise(shat_ma = mean(ma_sea_hat),
            shat_loess = mean(loess_sea_hat))

ma_pred <- ts(ma_trend_pred + seasonals$shat_ma,
              start = c(2020,1),
              frequency = 12)
```

```
## Warning in ma_trend_pred + seasonals$shat_ma: longer object length is not a
## multiple of shorter object length
```

```r
loess_pred <- ts(loess_trend_pred + seasonals$shat_loess,
                 start = c(2020,1),
                 frequency = 12)
```

```
## Warning in loess_trend_pred + seasonals$shat_loess: longer object length is not
## a multiple of shorter object length
```

```r
plot(co2_test, ylim = c(min(co2_test,ma_pred,loess_pred), max(co2_test,ma_pred,loess_pred)))
lines(ma_pred,col = "red")
lines(loess_pred, col = "blue")
legend(2020,
       418,
       legend = c("Test data", "MA Prediction", "Loess Prediction"),
       lty = c("solid","solid","solid"),
       col = c("black", "red","blue"))
```




```r
# MSPE
MSPE <- tibble(ma_pred = ma_pred,
               loess_pred = loess_pred,
               observed = co2_test) %>%
  mutate(sqr_diff_ma = (observed - ma_pred)^2,
         sqr_diff_loess = (observed - loess_pred)^2) %>%
  summarise(mspe_ma = mean(sqr_diff_ma),
            mspe_loess = mean(sqr_diff_loess))
MSPE
```

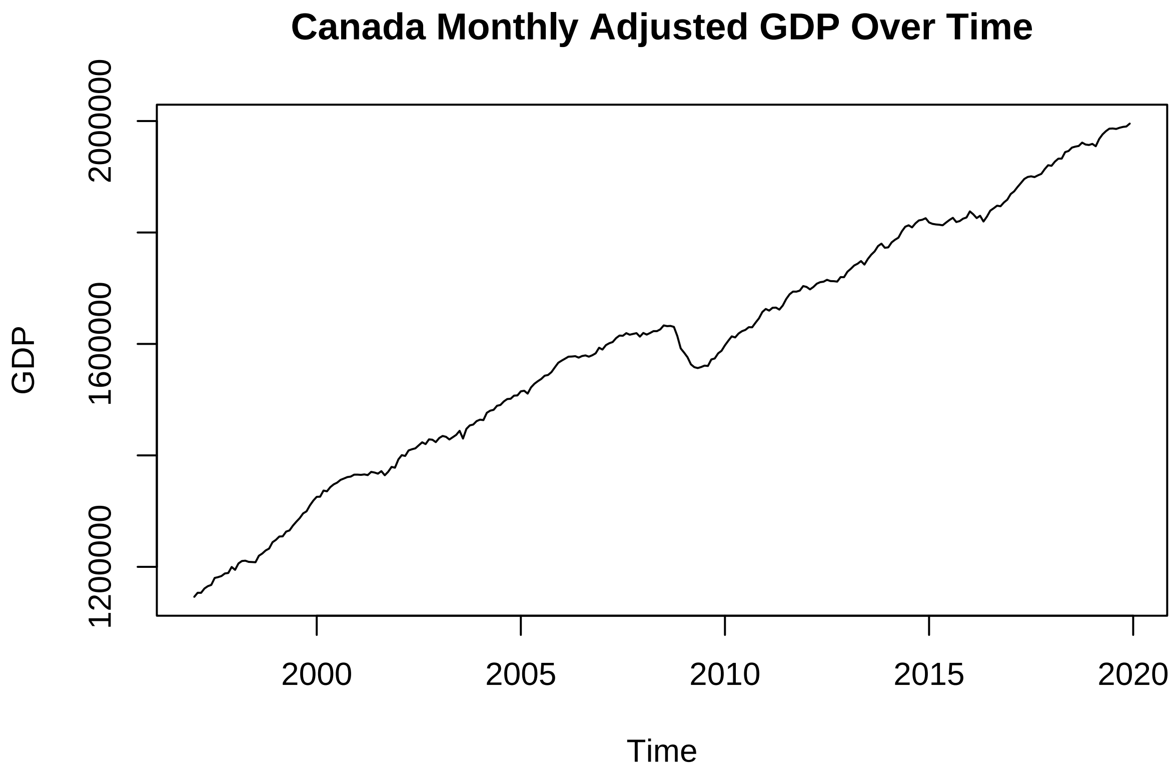
```
## # A tibble: 1 x 2
##   mspe_ma mspe_loess
##     <dbl>      <dbl>
## 1    34.1       32.1
```

Looking at the above table, the loess smoothing method result in lower MSPE on the test data. Therefore I would recommend to use the prediction model from the loess smoothing decomposition.

Question 2

a)

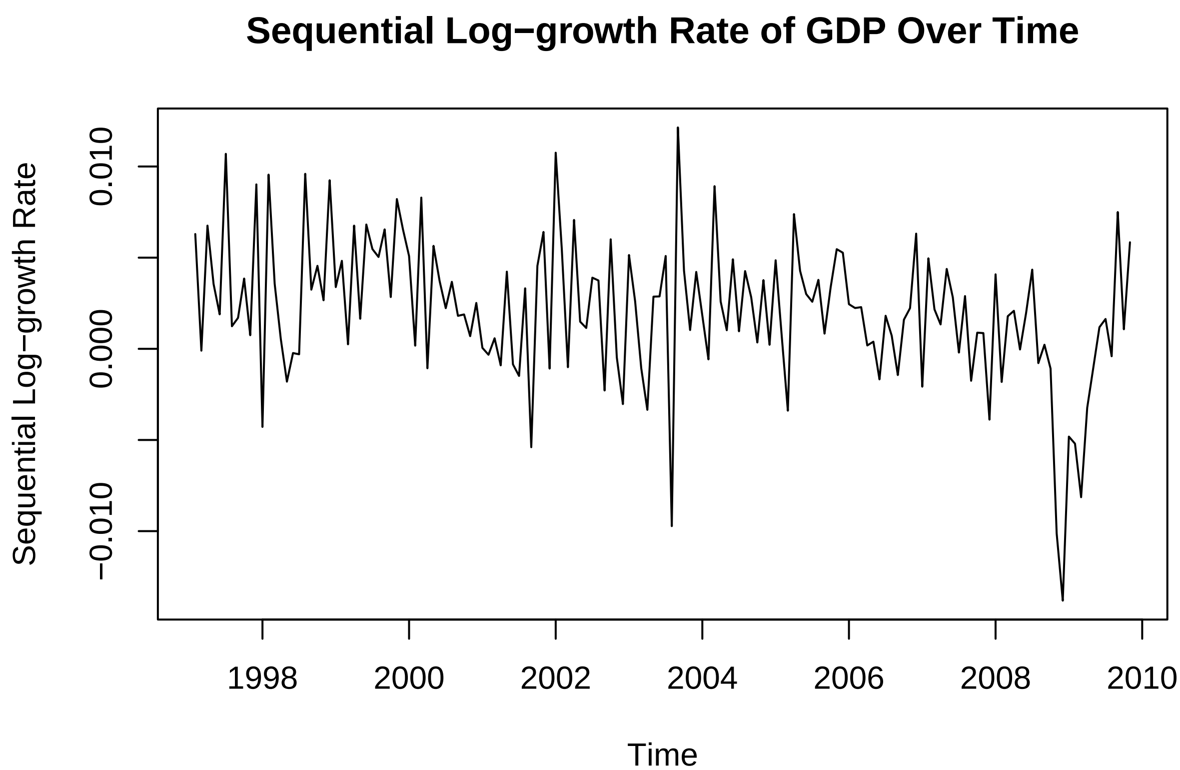
```r
gdp <- read_csv("CanandaGDP-2.csv",
                skip = 1,
                col_names = c("Time", "GDP"),
                show_col_types = FALSE)
gdp_ts <- ts(gdp[,2],start = c(1997,1), end = c(2019,12), frequency = 12)
plot(gdp_ts, main = "Canada Monthly Adjusted GDP Over Time")
```



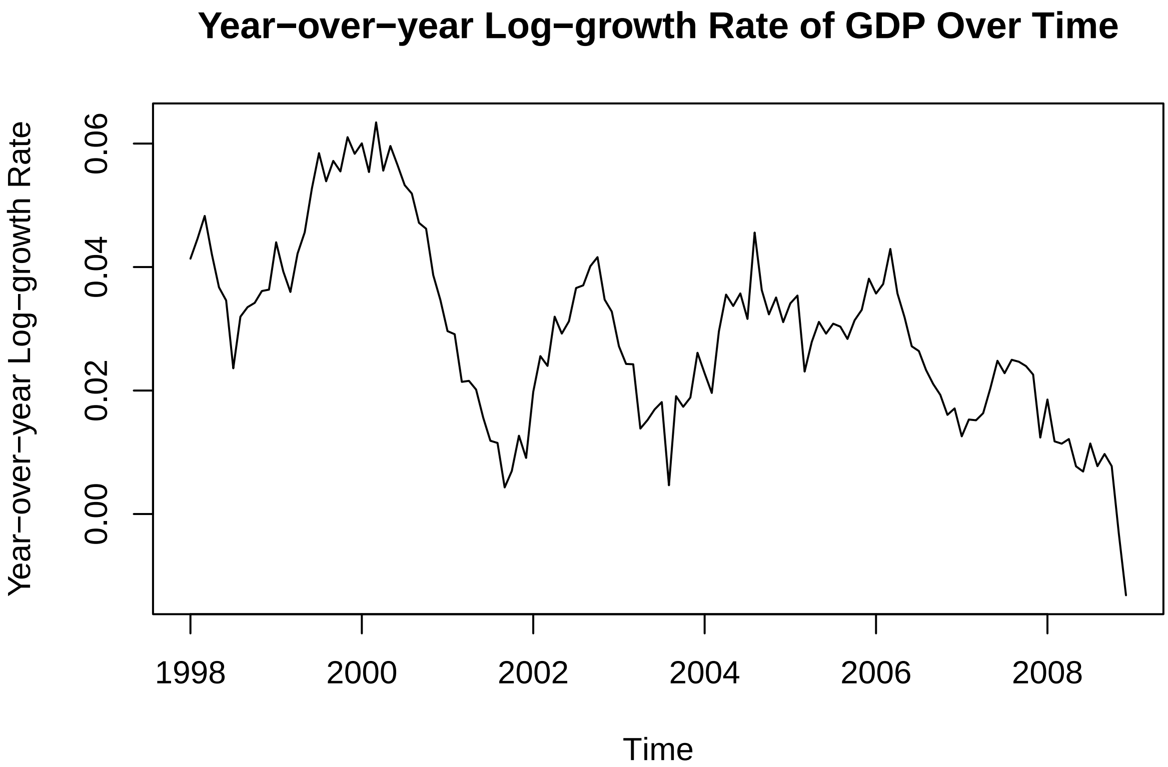
This time series is likely to be non-stationary because the present of a clear upward trend.

b)

```r
n = length(gdp$GDP)
seq_log_growth <- ts(log((gdp$GDP[-1])/(gdp$GDP[-n])),
                     start = c(1997,2),
                     end = c(2009,11),
                     frequency = 12)
plot(seq_log_growth,
     ylab = "Sequential Log-growth Rate",
     main = "Sequential Log-growth Rate of GDP Over Time")
```




```
year_over_year_log_growth <- ts(log((gdp$GDP[-c(1:12)])/(gdp$GDP[-c((n-11):n)])),
                                start = c(1998,1),
                                end = c(2008,12),
                                frequency = 12)
plot(year_over_year_log_growth,
     ylab = "Year-over-year Log-growth Rate",
     main = "Year-over-year Log-growth Rate of GDP Over Time")
```

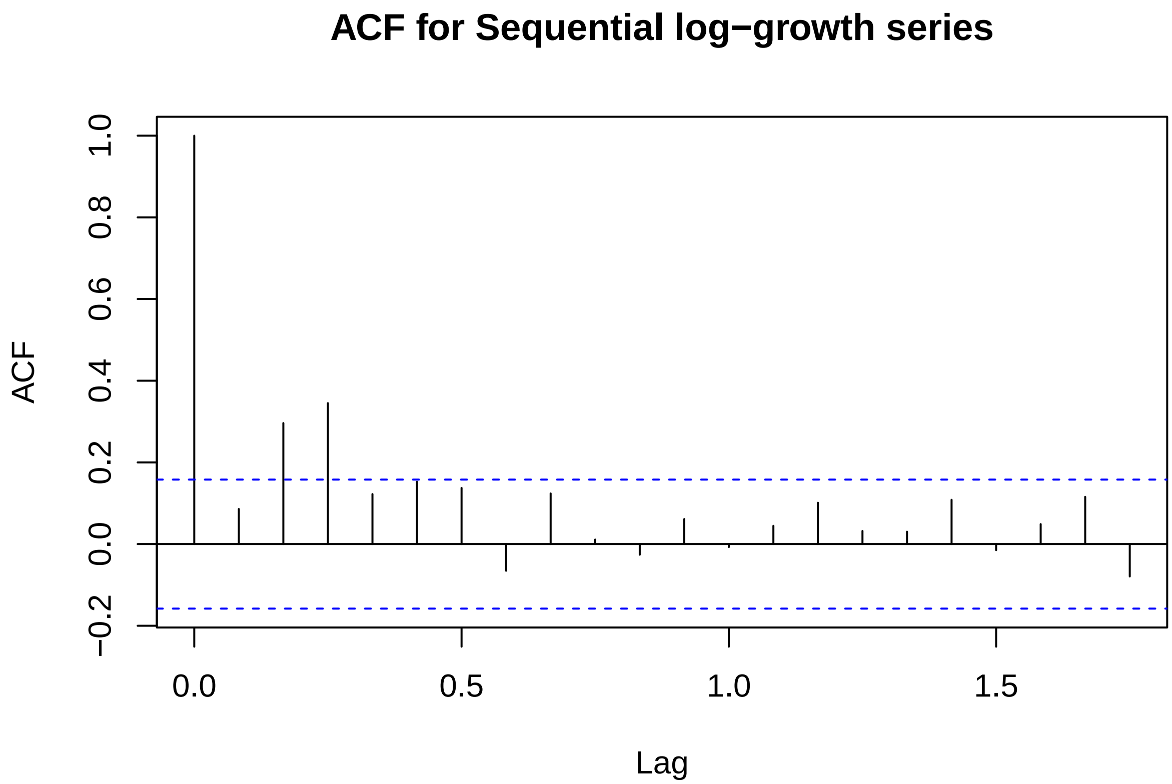


i) Looking at the above plots of log-growth rate series, the sequential log-growth rate series appear to be stationary because there is no apparent trend and the mean of log-growth rate appears to be constant. Whereas the year-over-year log-growth rate series appear to be non-stationary because the presence of a downward trend in log-growth rate over time.

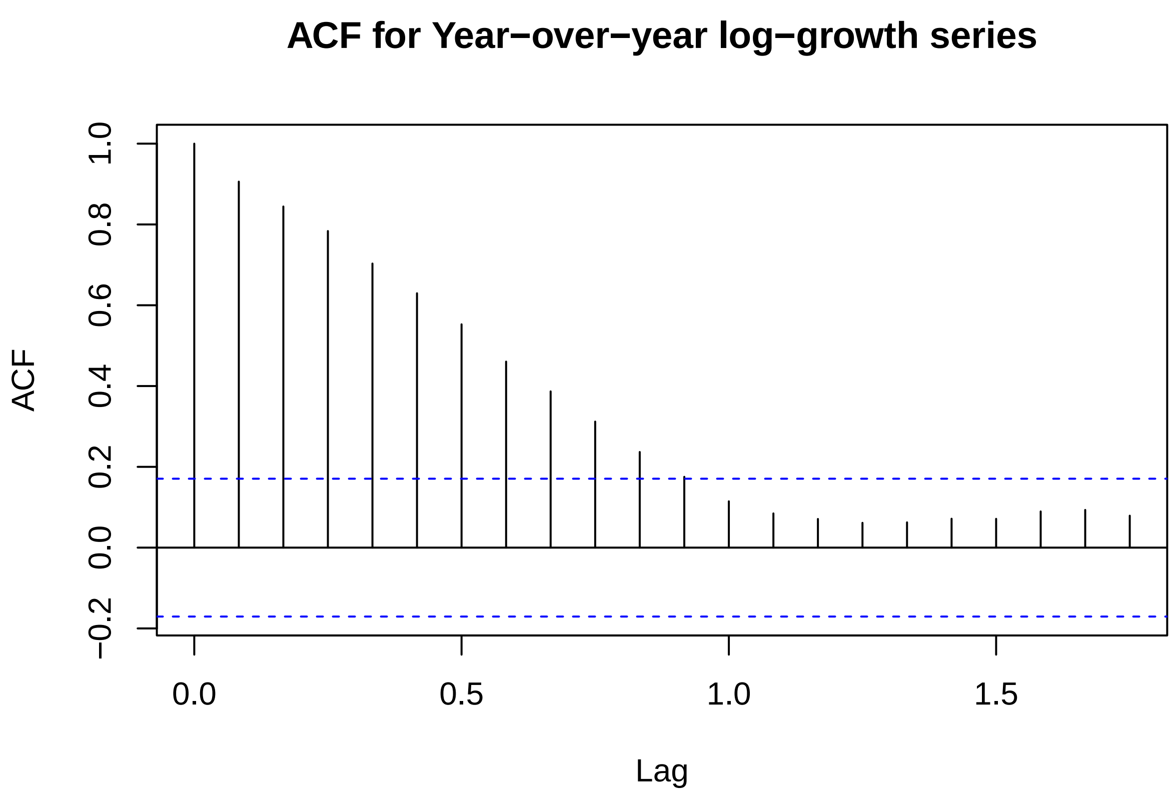
ii) The sequential log-growth rate definition is preferable, this is because the sequential log-growth transformation results in a new time series that is stationary.

iii)

```
acf(seq_log_growth,main = "ACF for Sequential log-growth series")
```




```
acf(year_over_year_log_growth, main = "ACF for Year-over-year log-growth series")
```



i) In the correlogram for sequential log-growth series, we observe the auto-correlation decays rapidly with lag and approaches toward 0, and most of the auto-correlations fall within the $\pm 2/\sqrt{n}$ bound. This suggests that there is likely no trend present in the series and the series is stationary.

ii) In the correlogram for year-over-year log-growth series, we observe that the auto-correlation decays slowly with lag. All of the auto-correlations are positive and many exceed the $\pm 2/\sqrt{n}$ bound. This suggests that there is strong positive temporal dependence or a upward trend present in the series and the series is likely non-stationary.