# STA 137 Final Project

Quynh Trinh,

2025-06-04

## Data Cleaning

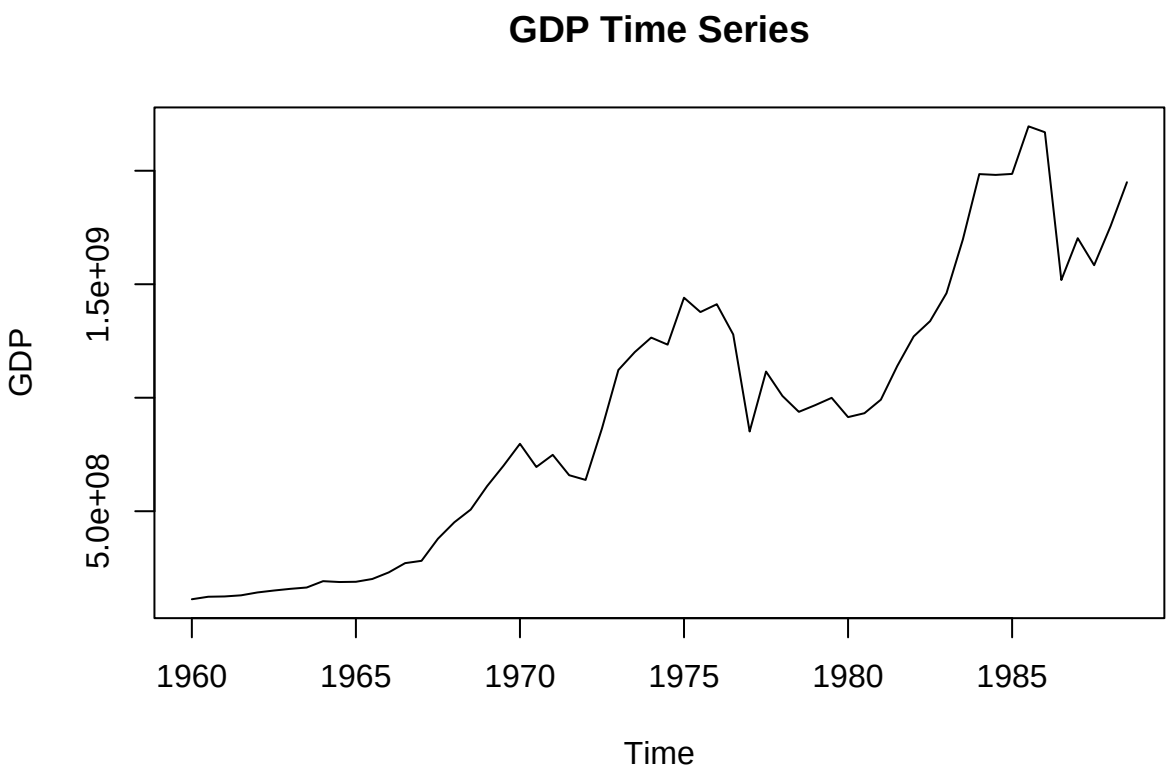
Keep Year, Imports, and GDP columns

```
finalPro_data <- finalPro_data[, c("Year", "Imports", "GDP")]
```

## I. GDP Time Series

### Plot GDP Time Series

```
# Plot GDP
gdp_ts <- ts(finalPro_data$GDP, start = 1960, frequency = 2)

ts.plot(gdp_ts, main="GDP Time Series", ylab="GDP")
```

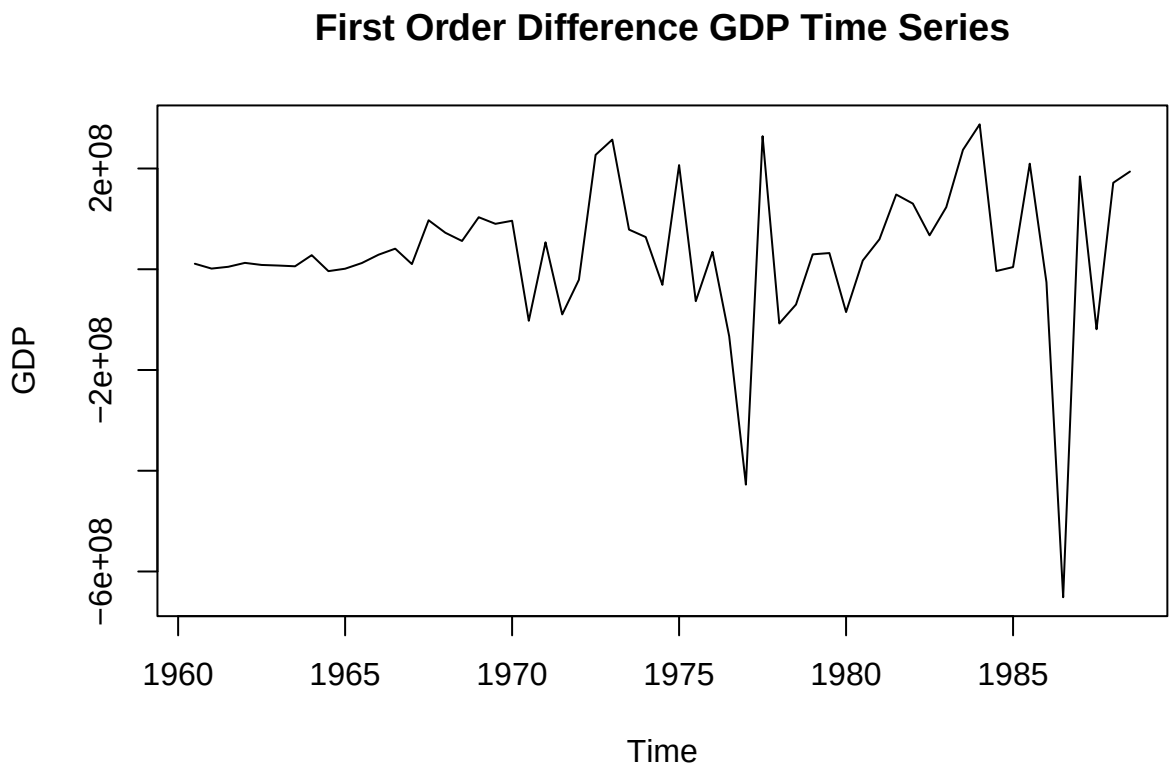



Summary:

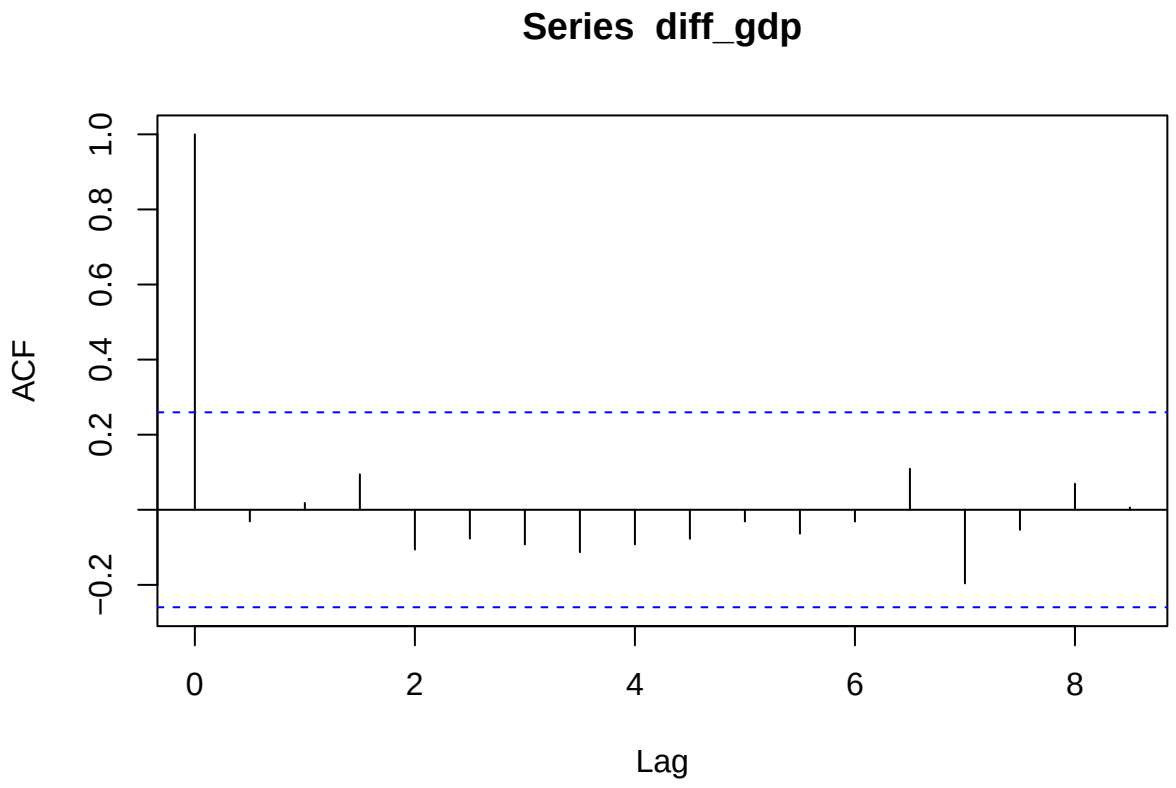
- GDP time series has upward trend, this shows this is non-stationary
- It has peaks around every 10 year: 1980, 1990, 2010

### Differencing GDP

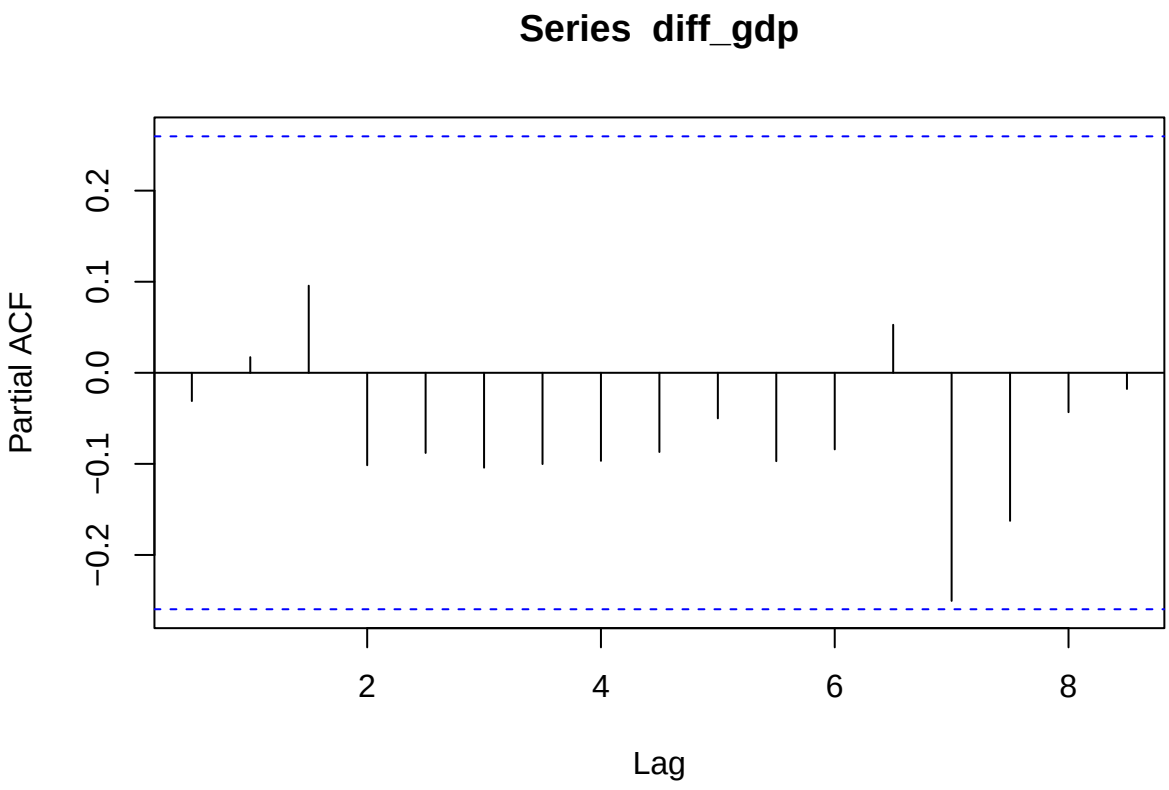
```
# 1st order difference
diff_gdp <- diff(gdp_ts)

ts.plot(diff_gdp, main="First Order Difference GDP Time Series", ylab="GDP")
```


First Order Difference GDP Time Series
GDP
2e+08
−2e+08
−6e+08
1960
1965
1970
1975
1980
1985
Time

```
acf(diff_gdp)
```




```
pacf(diff_gdp)
```




## Diagnostic GDP

### Coefficients

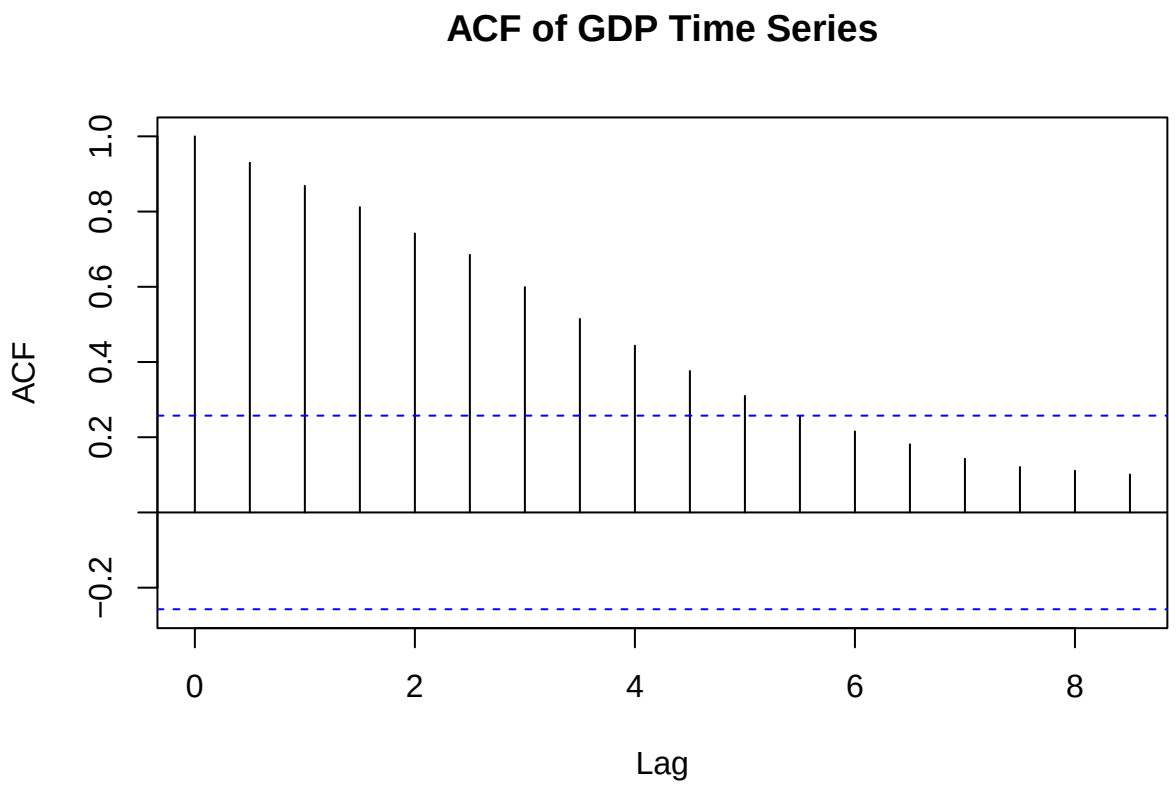
```
#
model_gdp <- lm(GDP ~ Year, data = finalPro_data)
```

(Note: default p-value = 0.1)

Significant: Year, CPI, Exports

### Residuals

```
# Residuals diagnostics for GDP
acf(gdp_ts, main = "ACF of GDP Time Series")
```



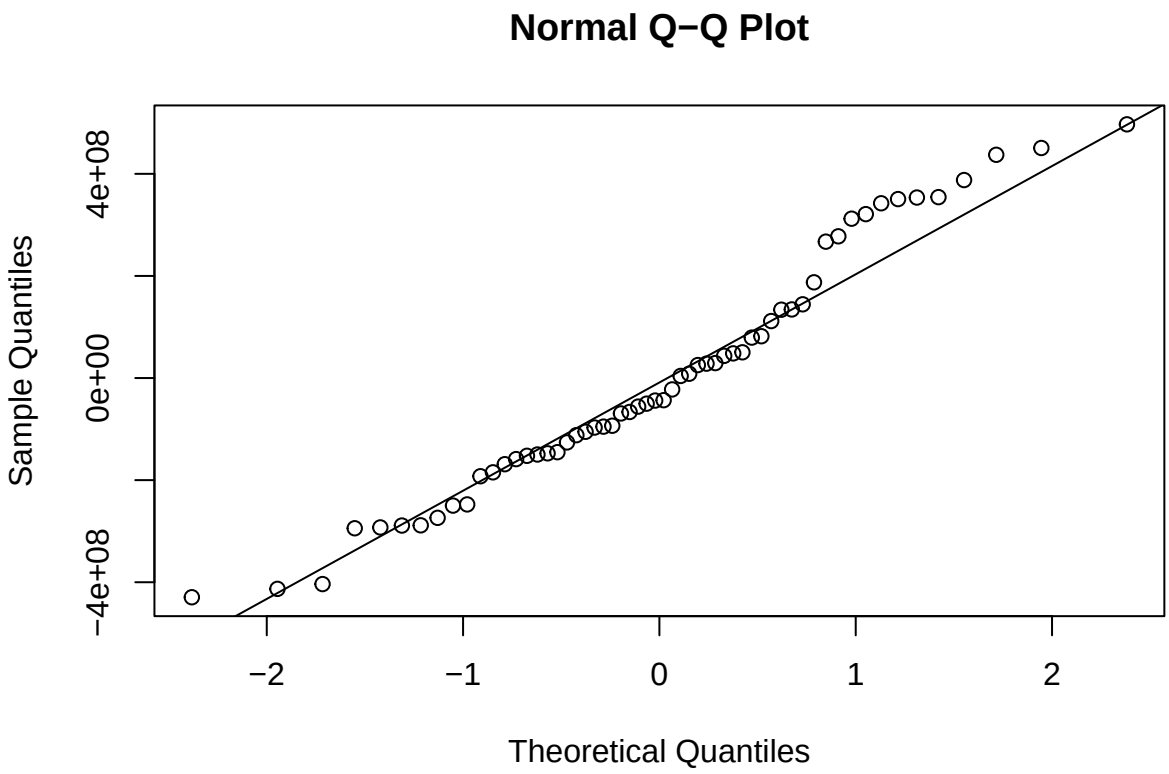

```
#Ljung test
#Box.test(gdp_ts)
```

ACF:

ACF values are decreasing gradually and stay above significance bounds, this means the time series is non-stationary, it likely has trend

### Check Normality and White Noise

```
resid_gdp <- residuals(model_gdp)

# Check normality
qqnorm(model_gdp$residuals)
qqline(model_gdp$residuals)
```




```
shapiro.test(resid_gdp) # Shapiro-Wilk test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid_gdp
## W = 0.96567, p-value = 0.09966
```

```
# Portmanteau Test (Ljung-Box)
Box.test(resid_gdp, lag = 10, type = "Ljung-Box")
```

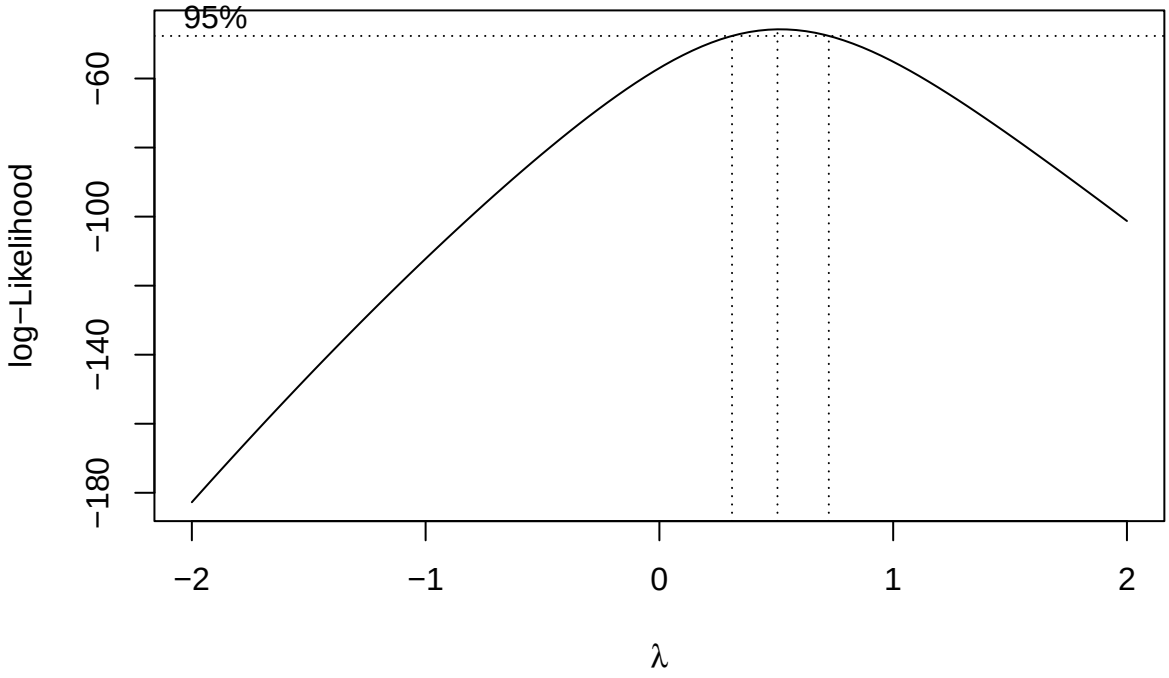
```
##
##  Box-Ljung test
##
## data:  resid_gdp
## X-squared = 120.16, df = 10, p-value < 2.2e-16
```

Normality:

Based on the QQ-plot for GDP time series, we see that there are most residuals are notclose to the line, which means the model might meet the assumption of normality. Then, we use Shapiro-Wilk test and got a p-value = 0.09966. Since the p-value is bigger than $\alpha = 0.05$, we conclude that our model for GDP is normally distributed.

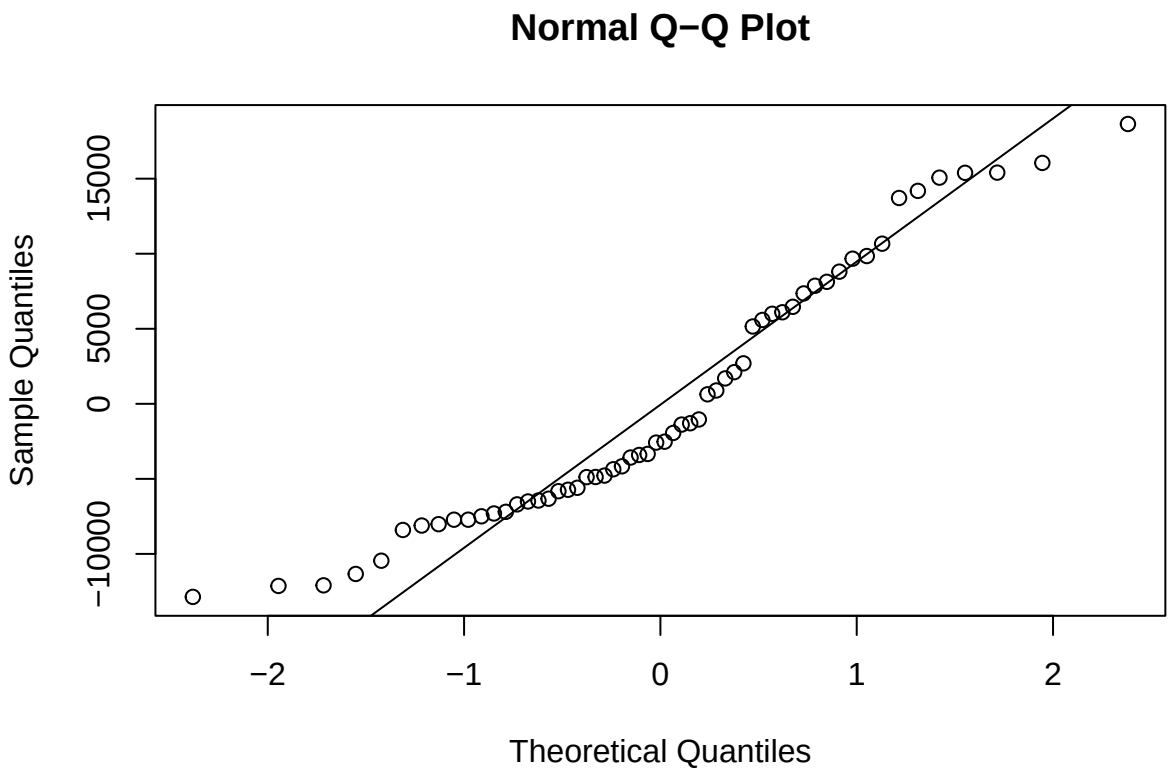
Next, we use Portmanteau Test (Ljung-Box) to test whether residuals are white noise. The null hypothesis for this test is residuals are white noise. After running the test, we got p-value $< 2.2e-16$. Since p-value $< 0.05$, we conclude that residuals are not white noise. Thus, we will transform our data to make it stationary.

**Diagnostic for Tranformation**




```
##
## Call:
## lm(formula = GDP_boxcox ~ Year, data = finalPro_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12863  -6490  -2553   6372  18644
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.647e+06  1.328e+05  -19.94   <2e-16 ***
## Year         1.363e+03  6.676e+01   20.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8511 on 56 degrees of freedom
## Multiple R-squared:  0.8816, Adjusted R-squared:  0.8794
## F-statistic: 416.8 on 1 and 56 DF,  p-value: < 2.2e-16
```

```r
# Normality box-cox
resid_gdp_boxcox <- residuals(model_gdp_boxcox)

# Check normality
qqnorm(model_gdp_boxcox$residuals)
qqline(model_gdp_boxcox$residuals)
```




```r
shapiro.test(resid_gdp_boxcox) # Shapiro-Wilk test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid_gdp_boxcox
## W = 0.93236, p-value = 0.003048
```

Normality and Constant Variance:

- Both assumptions are met after transforming

### Find the best ARIMA model for GDP Using auto.arima()

```r
arima_gdp <- auto.arima(gdp_ts)
arima_gdp
```

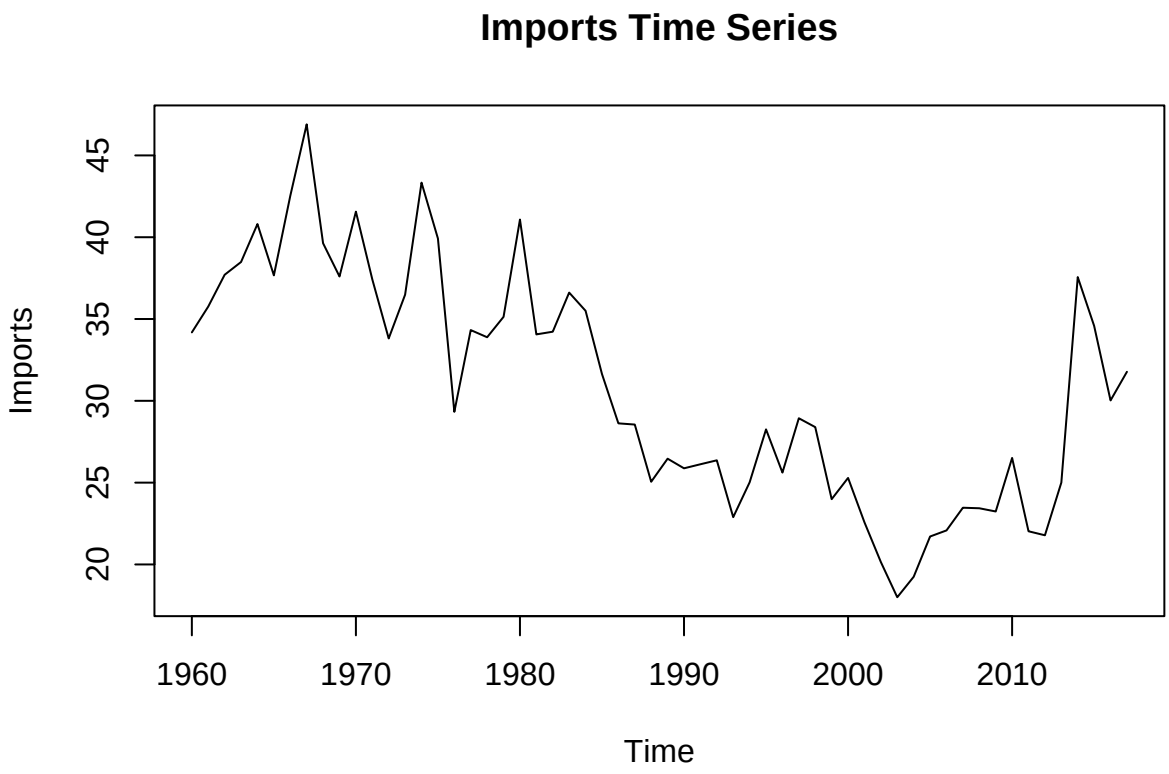
```
## Series: gdp_ts
## ARIMA(0,1,0) with drift
##
## Coefficients:
##          drift
##       32232562
## s.e.  19852873
##
## sigma^2 = 2.269e+16:  log likelihood = -1153.71
## AIC=2311.42   AICc=2311.64   BIC=2315.51
```

Suggested: ARIMA(0,1,0)

# II. Imports Time Series

## Plot Imports Time Series

```r
# Plot imports
imports_ts <- ts(finalPro_data$Imports, start = 1960, frequency = 1)

ts.plot(imports_ts, main="Imports Time Series", ylab="Imports")
```




Summary:

- The plot shows there is downward trend from 1960 to 2005, and increasing after that
- This means the Imports time series is non-stationary

## Diagnostics Imports

### Coefficients

```
model_imports <- lm(Imports ~ Year, data = finalPro_data)

summary(model_imports)
```

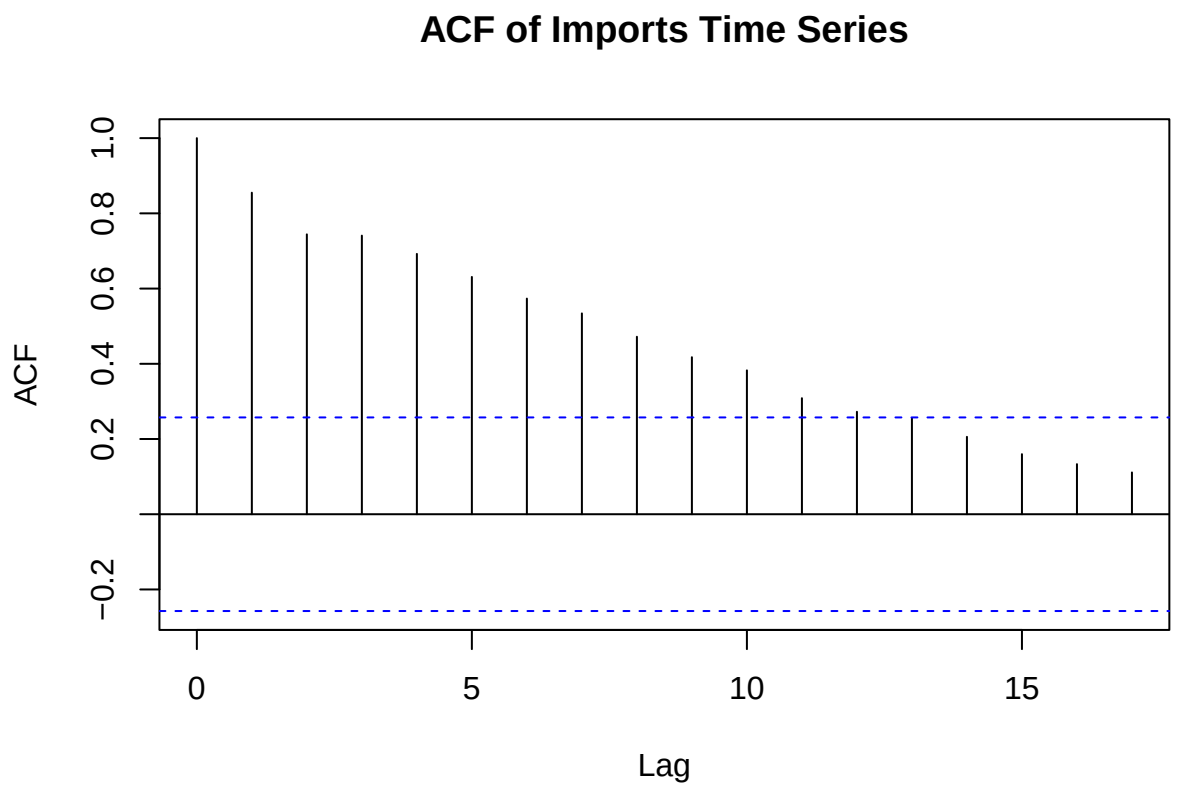
```
##
## Call:
## lm(formula = Imports ~ Year, data = finalPro_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3572 -3.5180 -0.4935  2.1746 14.6072
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 644.94984   76.49373   8.431 1.52e-11 ***
## Year         -0.30884    0.03847  -8.029 6.94e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.904 on 56 degrees of freedom
## Multiple R-squared:  0.5351, Adjusted R-squared:  0.5268
## F-statistic: 64.46 on 1 and 56 DF,  p-value: 6.937e-11
```

Significant: CPI, Exports

**Residuals**

```
# Residuals diagnostics for Imports
acf(imports_ts, main = "ACF of Imports Time Series")
```




```
#Ljung test
Box.test(imports_ts)
```

```
##
##  Box-Pierce test
##
## data:  imports_ts
## X-squared = 42.394, df = 1, p-value = 7.46e-11
```

ACF:

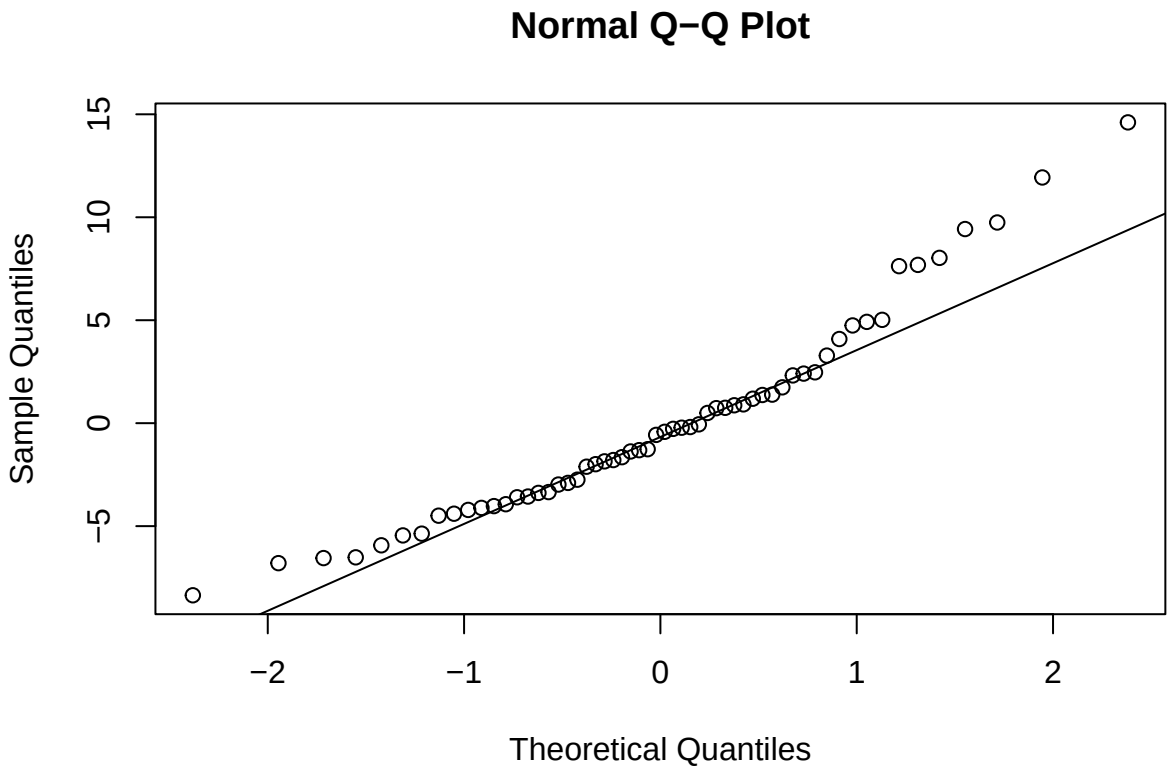
- ACF values decrease gradually and stay above significance bounds, this means the time series is non-stationary

Ljung test:

- p-value = 7.46e-11, this means the residuals are not independent

### Normality and Constant Variance

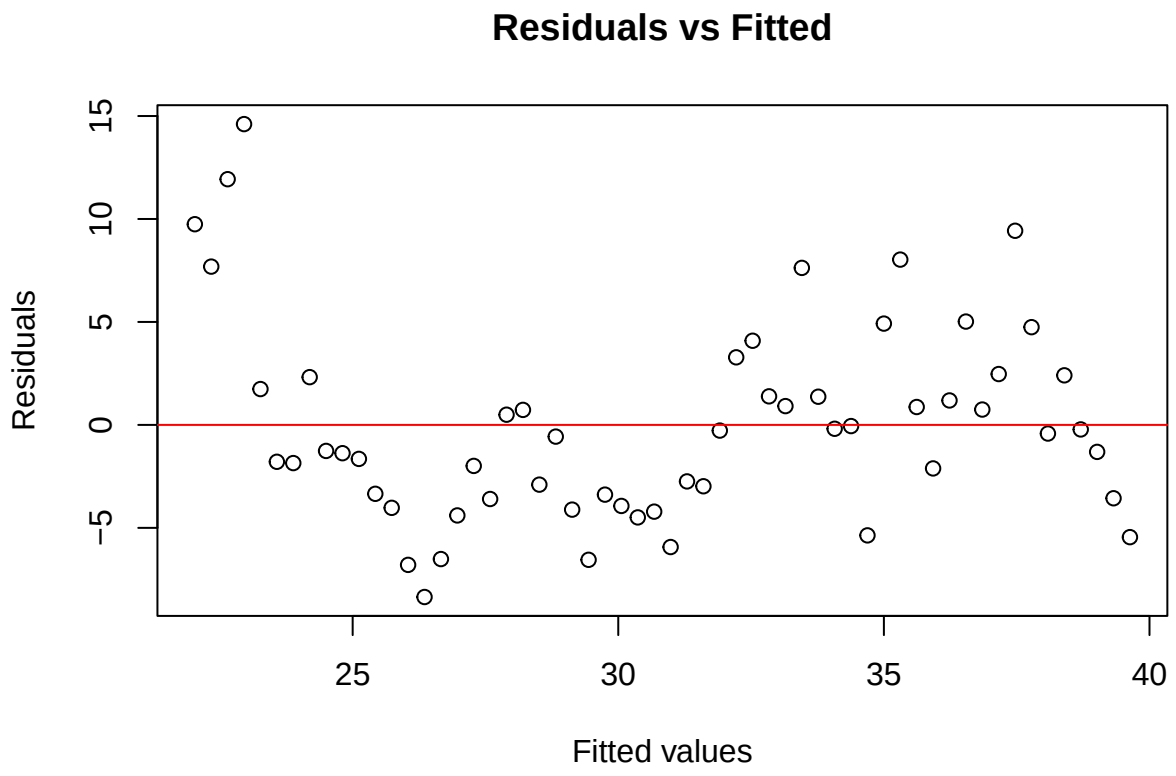
```
resid_imports <- residuals(model_imports)

# Check normality
qqnorm(model_imports$residuals)
qqline(model_imports$residuals)
```




```
shapiro.test(resid_imports) # Shapiro-Wilk test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid_imports
## W = 0.94415, p-value = 0.009884
```

```r
# Check constant variance
plot(fitted(model_imports), resid_imports,
     xlab = "Fitted values", ylab = "Residuals",
     main = "Residuals vs Fitted")
abline(h = 0, col = "red") # Residuals vs Fitted plot
```




```r
group_imports <- ifelse(fitted(model_imports) > median(fitted(model_imports)), "High", "Low")
leveneTest(resid_imports ~ group_imports, center=median)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.6171 0.4355
##       56
```

Normality:

- Most residuals are close to the line except some points at the right end of the plot. The data might meet normality
- Shapiro test: p-value = 0.1748. This means the data met normality

Constant Variance:

- The plot shows some patterns; most points are at the left side and waving pattern. This means the data might not meet constant variance
- Brown test: 0.07019 > 0.05. This means the data met constant variance

Thus, there is no needed for transforming Imports model.

### Find the Best ARIMA Model for Imports Using auto.arima()

```
arima_imports <- auto.arima(imports_ts)
arima_imports
```

```
## Series: imports_ts
## ARIMA(0,1,2)
##
## Coefficients:
##           ma1      ma2
##       -0.0463  -0.4473
## s.e.   0.1307   0.1361
##
## sigma^2 = 12.33:  log likelihood = -151.68
## AIC=309.37   AICc=309.82   BIC=315.5
```

Suggested: ARIMA(0,1,0)

## Save cleaned transformation

```
# Remove GDP, Country, Code
finalPro_data$GDP <- NULL
finalPro_data$Country <- NULL
finalPro_data$Code <- NULL

# Save the updated data frame to CSV
write.csv(finalPro_data, file = "finalPro_data_BoxCox.csv", row.names = FALSE)
```