ECON 104 Project 2

Maritza Jimenez, Elaine Tran, Titania Le, Aanya Pramanik

2023-11-09

Exploratory Data Analysis

Question 1a

Despite employment and GNP performing relatively similar on the surface, are there any limitations in the data that may hinder your analysis of the variables?

Question 1b

Christian Kleiber and Achim Zeileis (2008). Applied Econometrics with R. New York: Springer-Verlag. ISBN 978-0-387-77316-2. URL https://CRAN.R-project.org/package=AER

The data set OrangeCounty from the AER package contains 2 variables, employment and gnp. Employment is the quarterly employment in the county and gnp is quarterly real GNP (Gross National Product). It contains a total of 76 observations and the data spans from 1965 to 1983. Employment values remain at 3 digits throughout with the min value being 288.000 and max value being 883.132. GNP ranges from 3-4 digits with the min value being 906.6016 and max value being 1571.9120.

Question 1c

The data set is complete and has no missing variables or inconsistencies as seen through the is.na function producing all FALSE results.

```
is.na(OrangeCounty.ts)
```

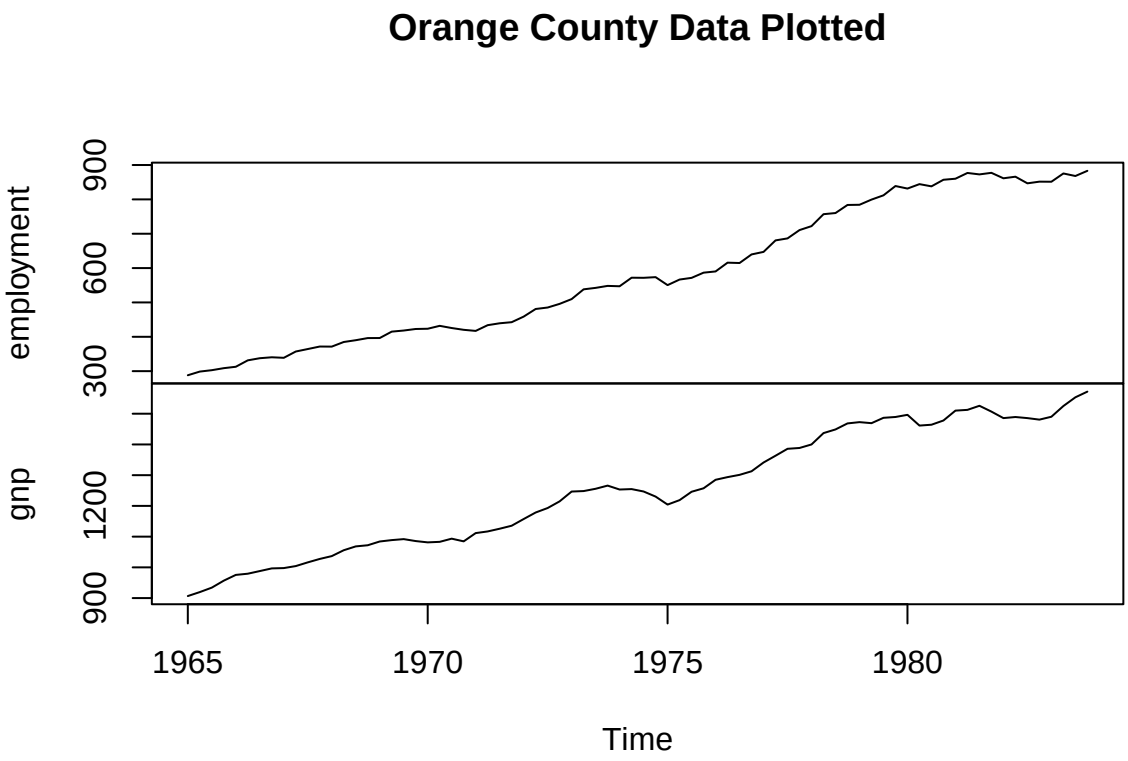
```
##       employment   gnp
##  [1,]      FALSE FALSE
##  [2,]      FALSE FALSE
##  [3,]      FALSE FALSE
##  [4,]      FALSE FALSE
##  [5,]      FALSE FALSE
##  [6,]      FALSE FALSE
##  [7,]      FALSE FALSE
##  [8,]      FALSE FALSE
##  [9,]      FALSE FALSE
## [10,]      FALSE FALSE
## [11,]      FALSE FALSE
## [12,]      FALSE FALSE
## [13,]      FALSE FALSE
```

```
## [14,]      FALSE FALSE
## [15,]      FALSE FALSE
## [16,]      FALSE FALSE
## [17,]      FALSE FALSE
## [18,]      FALSE FALSE
## [19,]      FALSE FALSE
## [20,]      FALSE FALSE
## [21,]      FALSE FALSE
## [22,]      FALSE FALSE
## [23,]      FALSE FALSE
## [24,]      FALSE FALSE
## [25,]      FALSE FALSE
## [26,]      FALSE FALSE
## [27,]      FALSE FALSE
## [28,]      FALSE FALSE
## [29,]      FALSE FALSE
## [30,]      FALSE FALSE
## [31,]      FALSE FALSE
## [32,]      FALSE FALSE
## [33,]      FALSE FALSE
## [34,]      FALSE FALSE
## [35,]      FALSE FALSE
## [36,]      FALSE FALSE
## [37,]      FALSE FALSE
## [38,]      FALSE FALSE
## [39,]      FALSE FALSE
## [40,]      FALSE FALSE
## [41,]      FALSE FALSE
## [42,]      FALSE FALSE
## [43,]      FALSE FALSE
## [44,]      FALSE FALSE
## [45,]      FALSE FALSE
## [46,]      FALSE FALSE
## [47,]      FALSE FALSE
## [48,]      FALSE FALSE
## [49,]      FALSE FALSE
## [50,]      FALSE FALSE
## [51,]      FALSE FALSE
## [52,]      FALSE FALSE
## [53,]      FALSE FALSE
## [54,]      FALSE FALSE
## [55,]      FALSE FALSE
## [56,]      FALSE FALSE
## [57,]      FALSE FALSE
## [58,]      FALSE FALSE
## [59,]      FALSE FALSE
## [60,]      FALSE FALSE
## [61,]      FALSE FALSE
## [62,]      FALSE FALSE
## [63,]      FALSE FALSE
## [64,]      FALSE FALSE
## [65,]      FALSE FALSE
## [66,]      FALSE FALSE
## [67,]      FALSE FALSE
```

```
## [68,]      FALSE FALSE
## [69,]      FALSE FALSE
## [70,]      FALSE FALSE
## [71,]      FALSE FALSE
## [72,]      FALSE FALSE
## [73,]      FALSE FALSE
## [74,]      FALSE FALSE
## [75,]      FALSE FALSE
## [76,]      FALSE FALSE
```

Question 1d

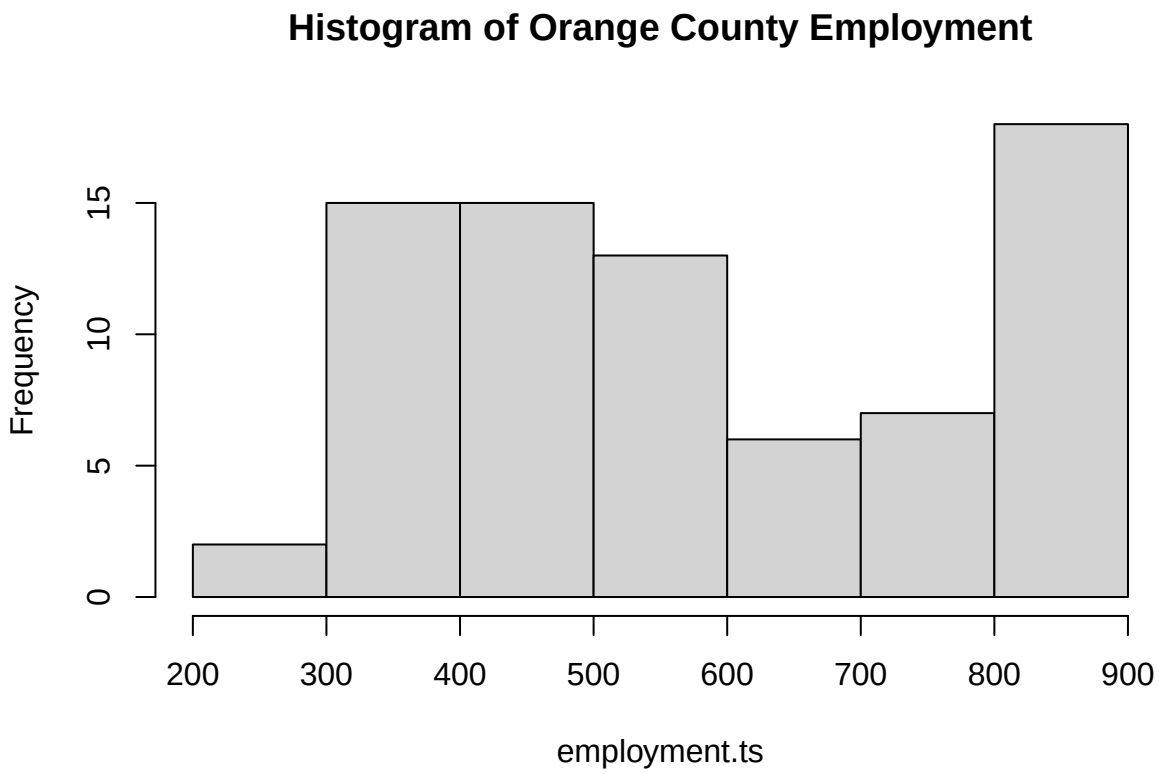
```
library(AER)
data("OrangeCounty")
OrangeCounty.ts <- ts(OrangeCounty, start = c(1965,1), end = c(1983,4), frequency = 4)
plot(OrangeCounty.ts, main = "Orange County Data Plotted")
```




```
employment.ts <- OrangeCounty.ts[,1]
gnp.ts <- OrangeCounty.ts[,2]
```

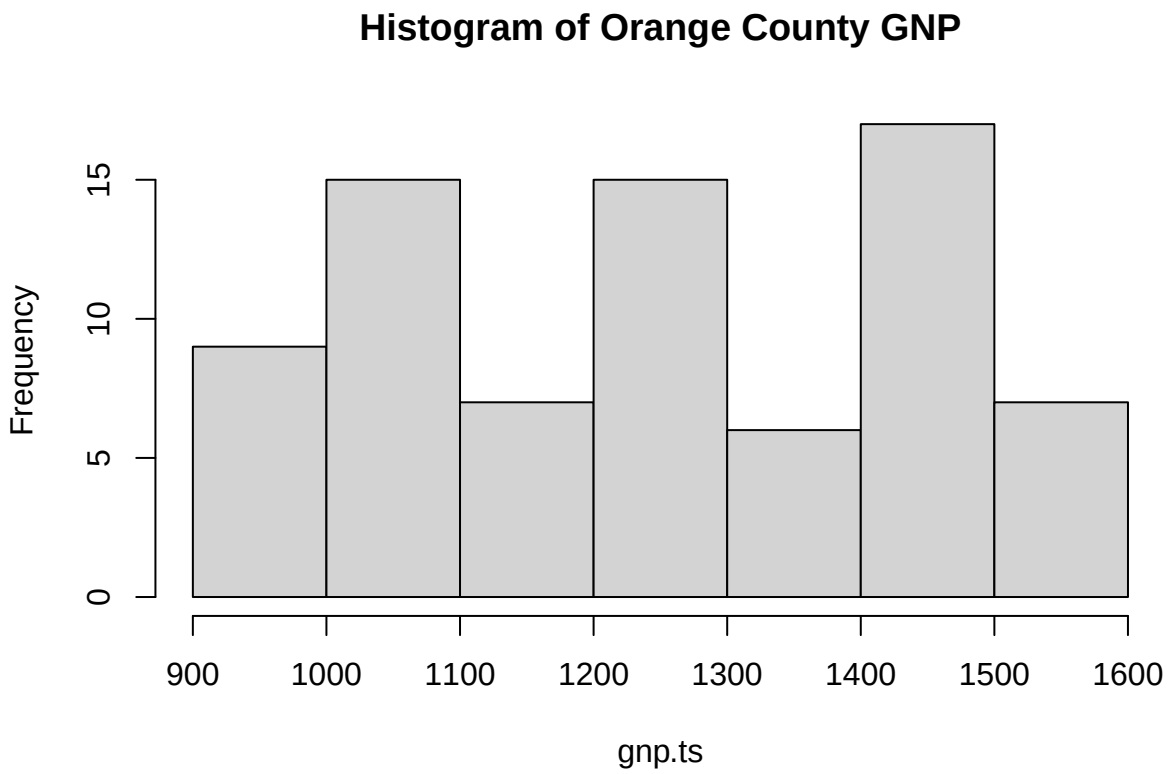
Employment and GNP appear to follow similar trends, with both variables showing dips and peaks concurrently. However, GNP exhibits peaks and dips with slightly greater magnitude. Employment, on the other hand, appears less stable, experiencing numerous changes—albeit very small ones—that GNP does not display.

```r
hist(employment.ts, main = "Histogram of Orange County Employment")
```



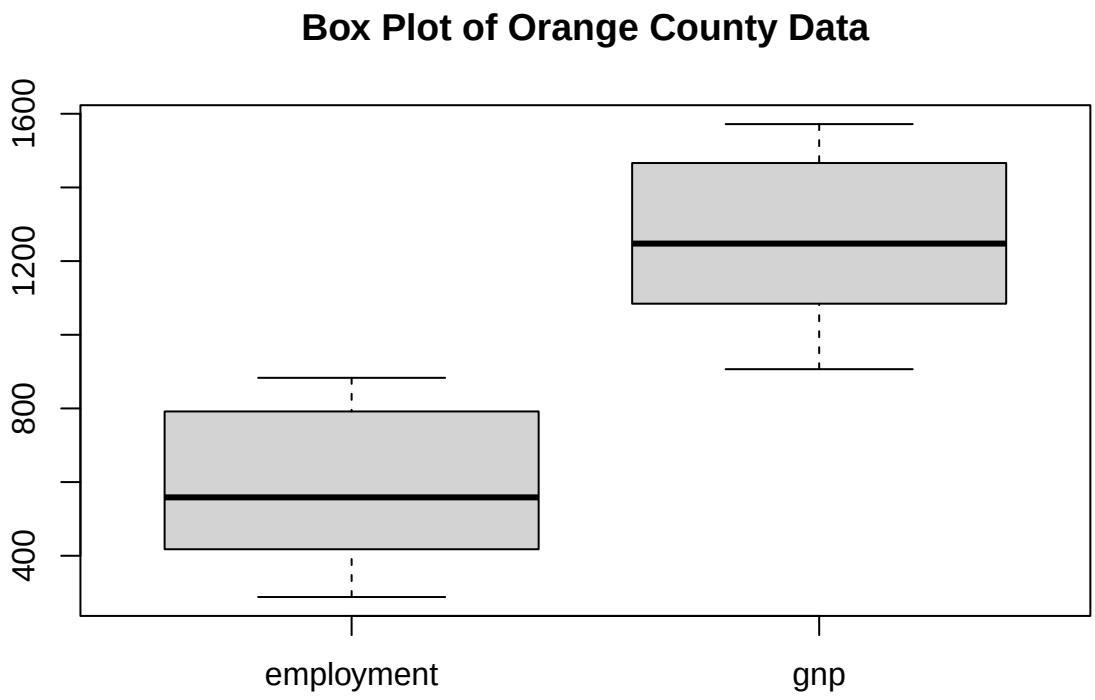
The histogram suggests a relatively balanced distribution, showing no significant skewness to the left or right. Quarterly employment in the lower range (200-300) and middle range (600-800) appears less frequent compared to other ranges, each with a frequency of approximately 10-15.

```r
hist(gnp.ts,  main = "Histogram of Orange County GNP")
```



The histogram depicting GNP data reveals a distinct pattern. In this representation, it appears that the data experiences jumps and falls in frequency. Specifically, GNP in the 900-1000 range has a frequency below 10, followed by the 1000-1100 range with a frequency of 15. This alternating pattern continues, with the subsequent range having a frequency below 10, and the following range registering a frequency of 15, and so on.

```r
boxplot(OrangeCounty.ts, main = "Box Plot of Orange County Data")
```

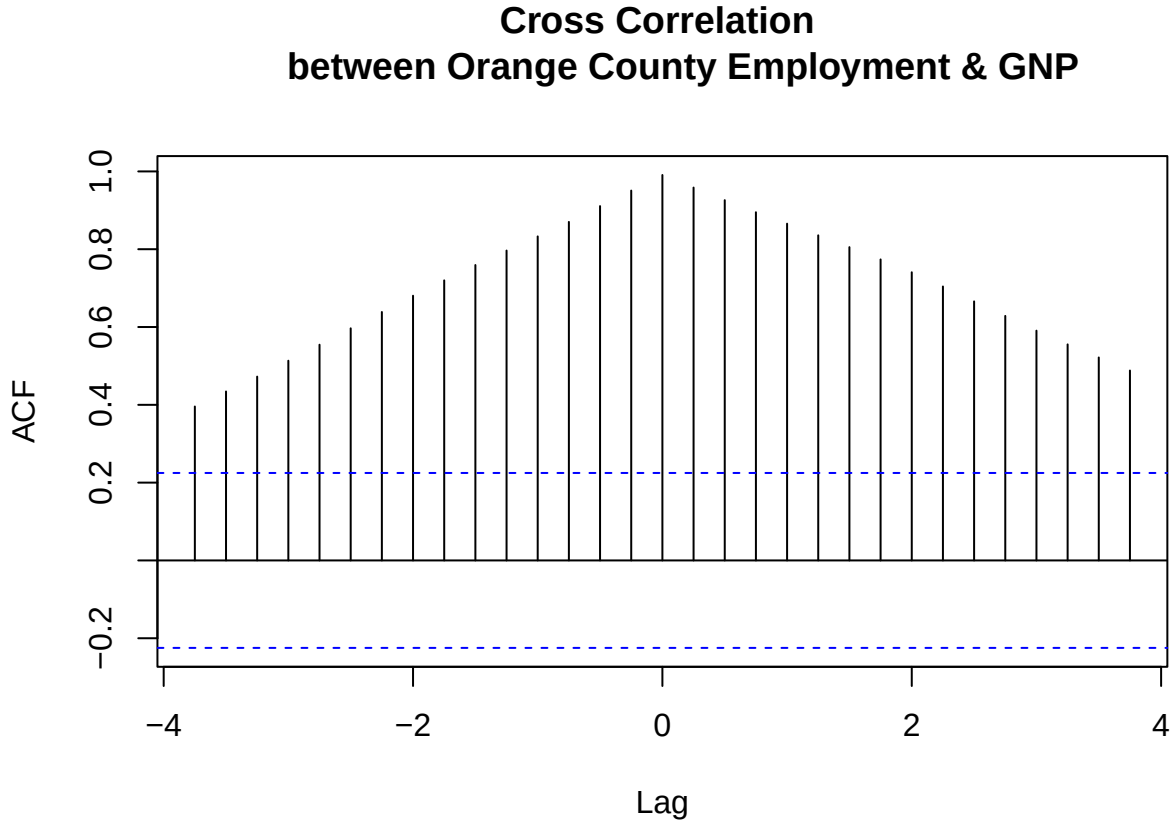



```r
summary(OrangeCounty.ts)
```

```
##    employment         gnp
##  Min.   :288.0   Min.   : 906.6
##  1st Qu.:417.9   1st Qu.:1084.5
##  Median :558.5   Median :1247.6
##  Mean   :584.0   Mean   :1252.2
##  3rd Qu.:788.1   3rd Qu.:1465.3
##  Max.   :883.1   Max.   :1571.9
```

The box plot reveals employment data ranging from a minimum of 288 to a maximum of 883.1, with a mean of approximately 580. Similarly, GNP data exhibits a minimum around 900, a maximum around 1570, and a mean of about 1250. The summary of the dataset aligns with the values depicted in the box plot.

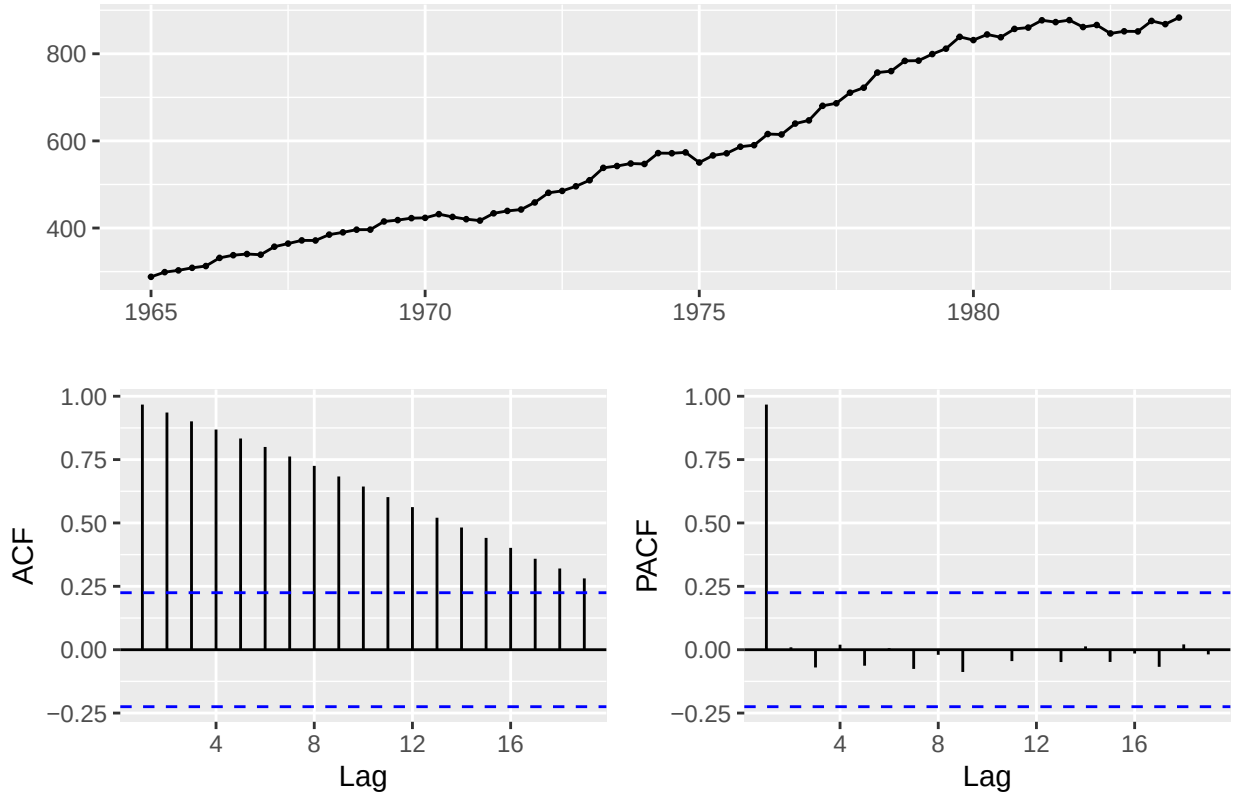
```r
ccf(OrangeCounty.ts[,1],(OrangeCounty.ts[,2]), main = "Cross Correlation
    between Orange County Employment & GNP")
```



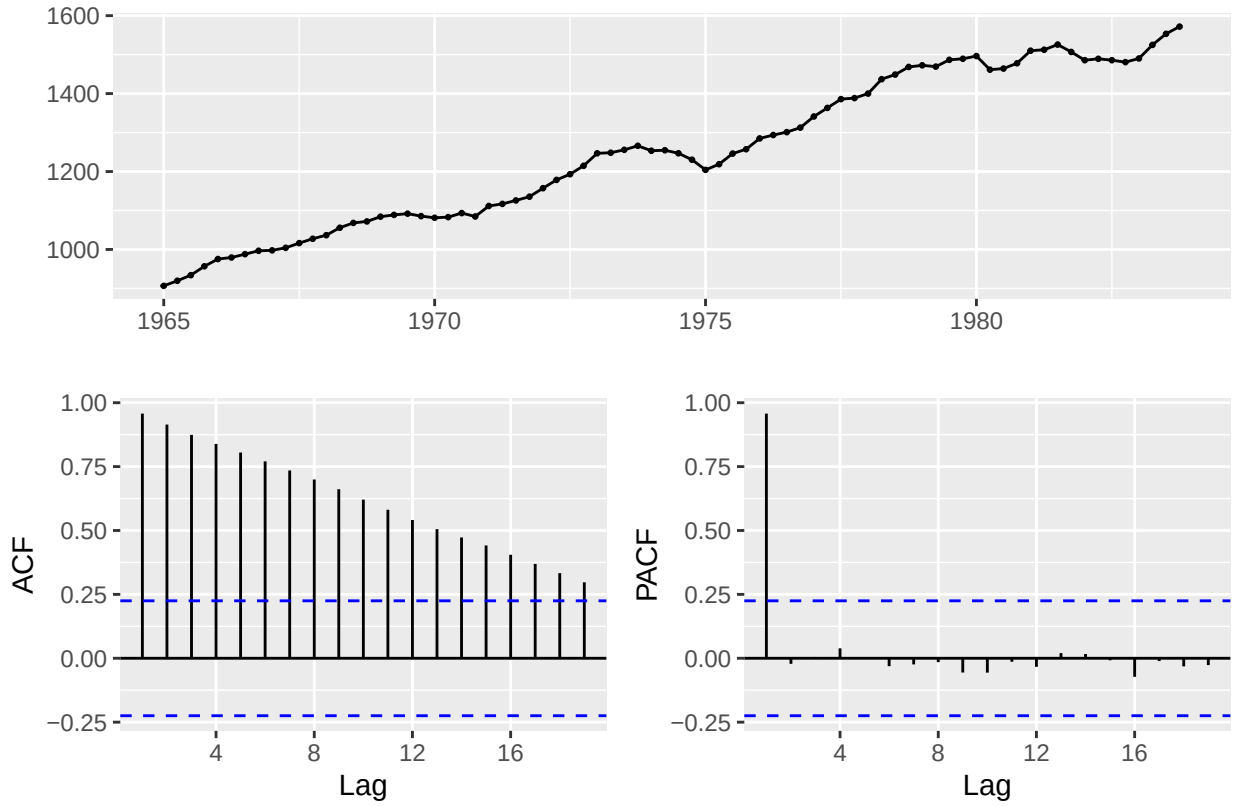
The cross-correlation plot highlights a robust correlation between our two variables, employment and GNP. This is intuitive, as employment and GNP frequently exhibit synchronized movements. For instance, during prosperous business periods, employers tend to hire more workers to meet increased demand.

Question 2a

```
library(forecast)
ggtsdisplay(employment.ts)
```



```
ggtsdisplay(gnp.ts)
```




```
library(tseries)
adf.test(employment.ts)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  employment.ts
## Dickey-Fuller = -3.2166, Lag order = 4, p-value = 0.09167
## alternative hypothesis: stationary
```

```
adf.test(gnp.ts)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  gnp.ts
## Dickey-Fuller = -3.2533, Lag order = 4, p-value = 0.08576
## alternative hypothesis: stationary
```

The ADF test results indicate p-values of 0.09167 for employment and 0.08576 for GNP, both exceeding the 0.05 threshold. Consequently, we fail to reject the null hypothesis of non-stationarity. To address this, differencing is required to achieve stationarity in our variables.

Question 2b

```
ndiffs(employment.ts)
```

```
## [1] 1
```

```
ndiffs(gnp.ts)
```

```
## [1] 1
```

The ndiffs command indicates that we need to difference both variables by 1 to account for the variables being non stationary.

```
library(base)
employment.ts2.1 <- diff(employment.ts)
adf.test(employment.ts2.1)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  employment.ts2.1
## Dickey-Fuller = -2.883, Lag order = 4, p-value = 0.2153
## alternative hypothesis: stationary
```

```
nsdiffs(employment.ts)
```

```
## [1] 1
```

```
gnp.ts2 <- diff(gnp.ts)
employment.ts2 <- diff(employment.ts, difference = 2)
adf.test(employment.ts2)
```

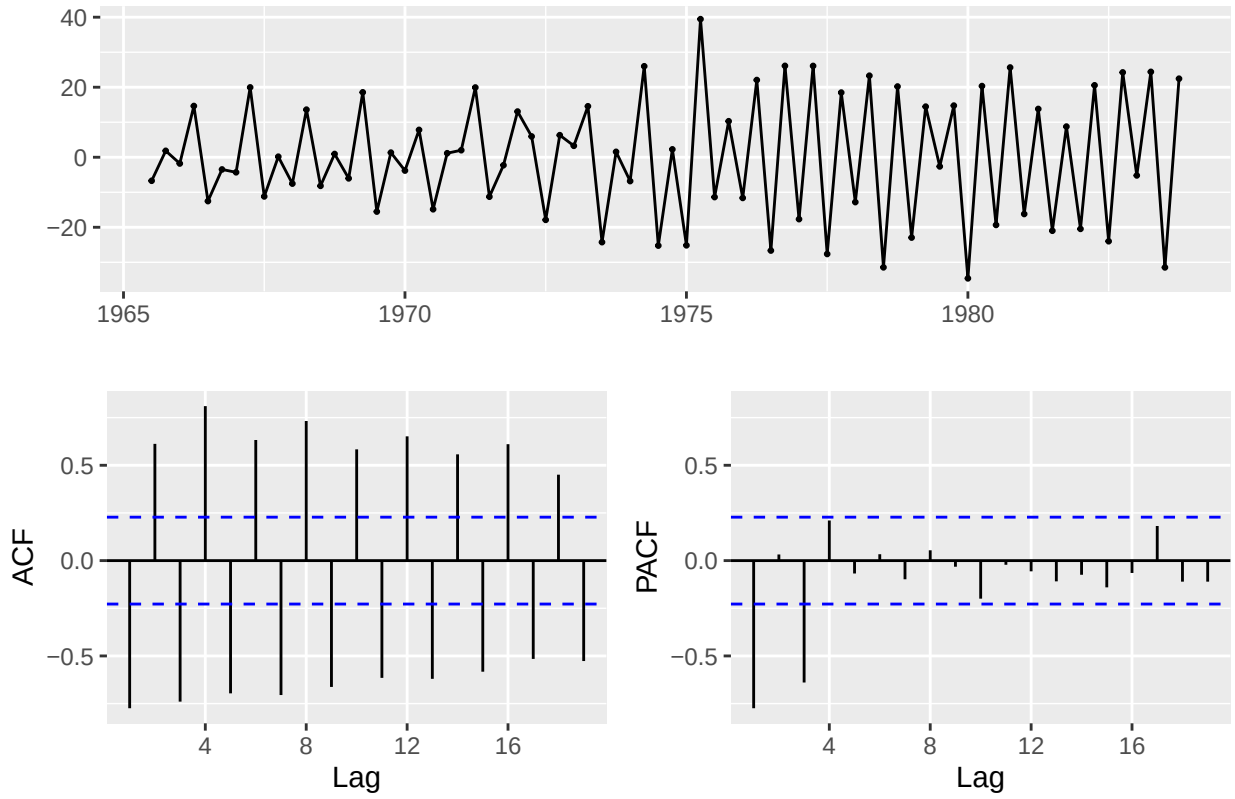
```
##
##  Augmented Dickey-Fuller Test
##
## data:  employment.ts2
## Dickey-Fuller = -4.9519, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

```
adf.test(gnp.ts2)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  gnp.ts2
## Dickey-Fuller = -3.5866, Lag order = 4, p-value = 0.04048
## alternative hypothesis: stationary
```

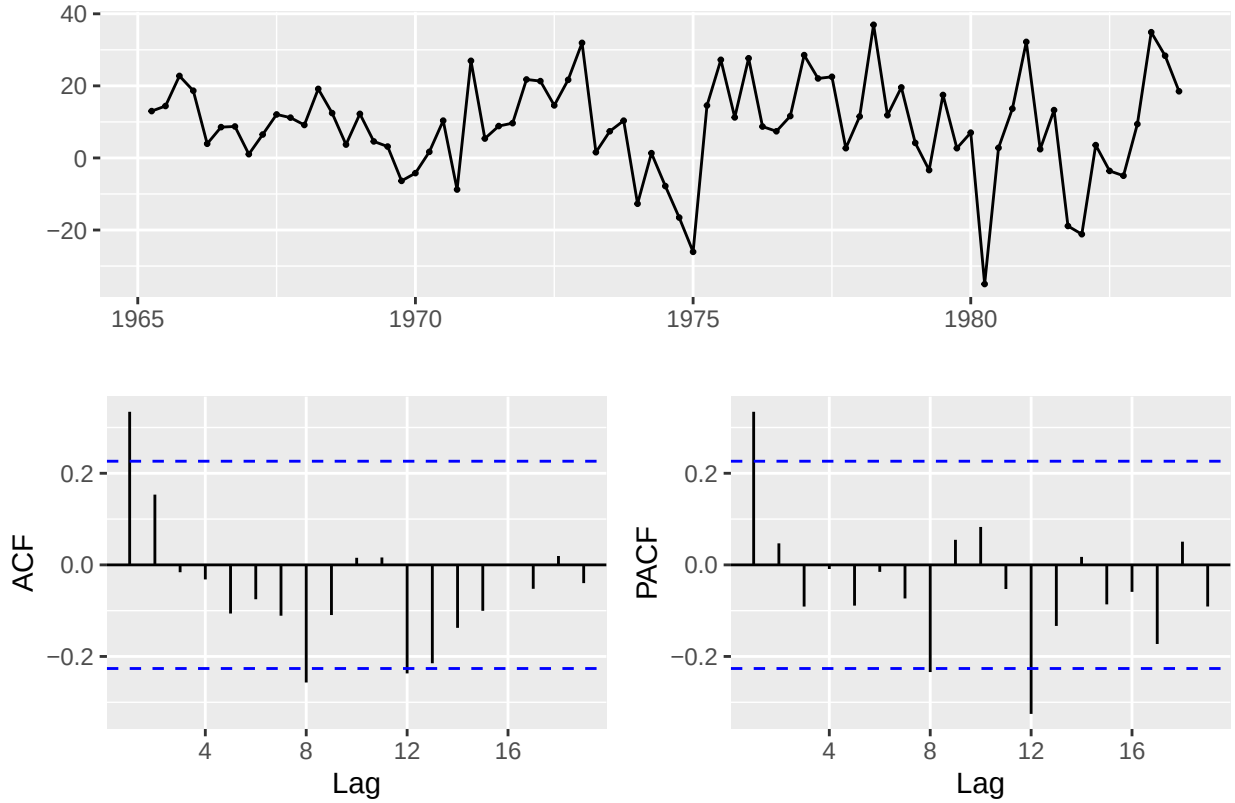
After differencing both variables by 1, the p-value for employment increases, indicating it remains non-stationary. In contrast, the p-value for GNP drops below 0.05, signifying stationarity. To assess potential seasonality in the employment variable, the `nsdiffs` command yields a result of 1, suggesting the presence of seasonality. To address this, we difference the employment variable twice, resulting in a p-value of 0.01—below 0.05—successfully achieving stationarity for both variables.

```
ggtsdisplay(employment.ts2)
```



Our plot now resembles white noise, displaying a more random pattern. With our variables now stationary, the ACF no longer exhibits a continuously declining pattern which is, at times, suggestive of non-stationarity. The PACF reveals significance at a lag of 3.

```
ggtsdisplay(gnp.ts2)
```



After differencing, the GNP graph now resembles white noise, exhibiting a random pattern. The ACF no longer continuously declines, a characteristic sometimes associated with non-stationarity. The PACF indicates potential significance at lags 8 and 12.

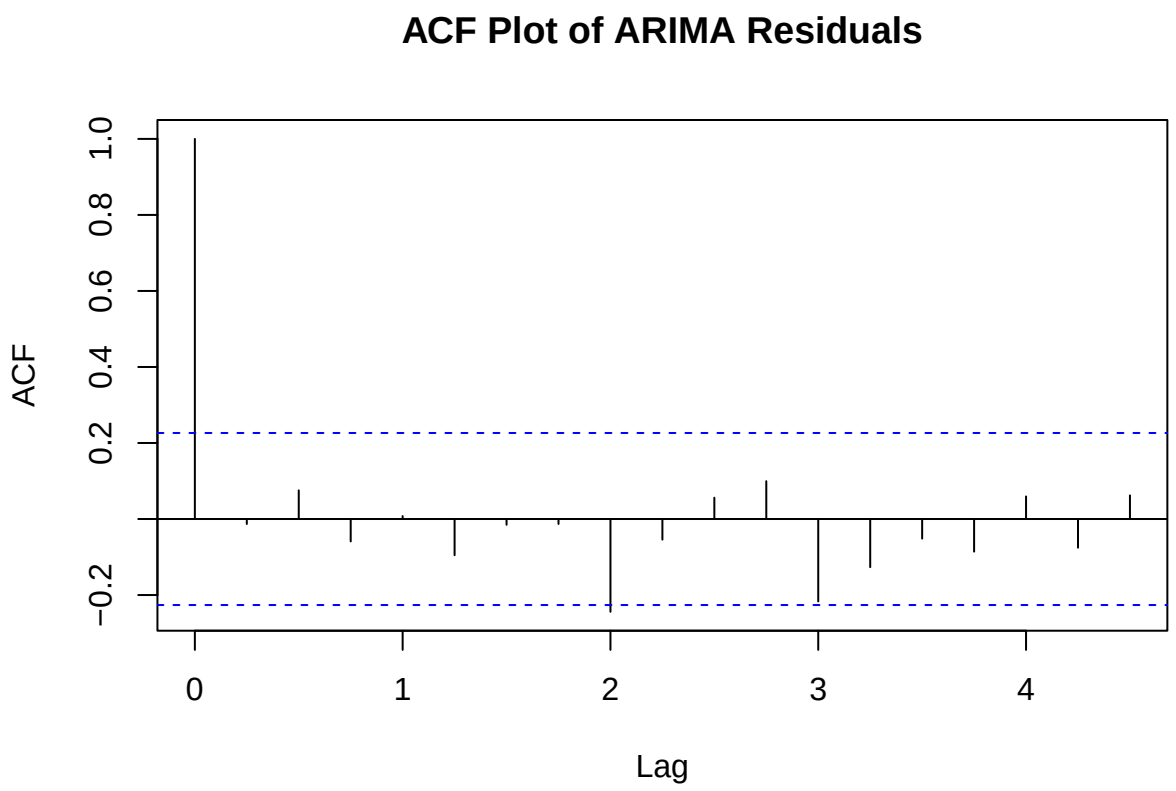
Question 3a

```
arima.mod<-auto.arima(gnp.ts2)
arima.mod
```

```
## Series: gnp.ts2
## ARIMA(1,0,0) with non-zero mean
##
## Coefficients:
##          ar1    mean
##       0.3325  8.9611
## s.e.  0.1083  2.2160
##
## sigma^2 = 170.8:  log likelihood = -298.23
## AIC=602.45   AICc=602.79   BIC=609.41
```

For this particular data set, GNP is our dependent variable. Using the auto.arima function in R, it can be concluded that the data follows an AR(1) model as the arima output is ARIMA (1,0,0).

```r
acf(arima.mod$residuals, main = "ACF Plot of ARIMA Residuals")
```



After plotting the ACF of the residuals for the AR(1) fitted model, there is still a slight indication of autocorrelation being present in the second lag. Therefore, a formal test must be carried out to confirm this.

Question 3b

```r
ljung_box_test <- Box.test(arima.mod$residuals, lag = 8, type = "Ljung-Box")
ljung_box_test
```

```
##
##  Box-Ljung test
##
## data:  arima.mod$residuals
## X-squared = 6.6662, df = 8, p-value = 0.573
```

After performing the Ljung_box test for the AR(1) model and testing it up to the second lag, our large p-value indicates that there is no evidence of autocorrelation as it is greater that our significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is no autocorrelation.

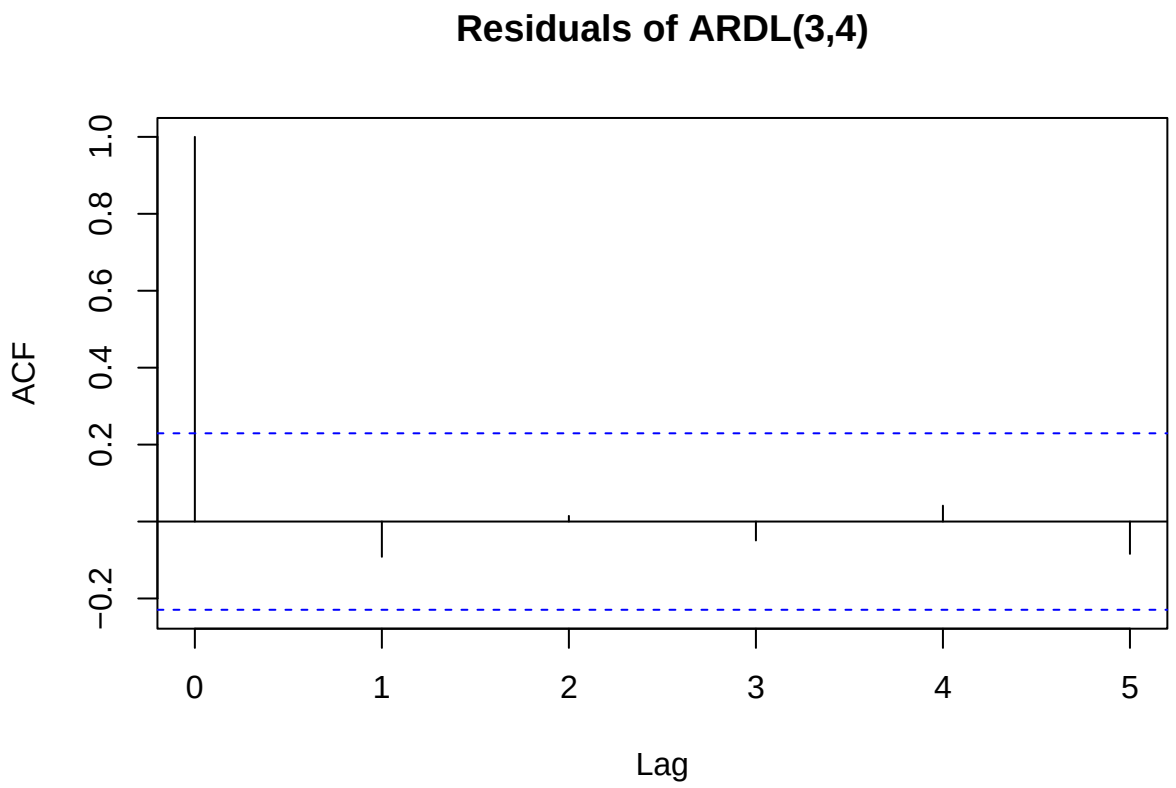
Question 3c

```
library(ARDL)
library(Hmisc)
data("OrangeCounty")
length(employment.ts2) <- length(gnp.ts2) # equalizes number of rows
employment.lag <- lag(employment.ts2)
gnp.lag <- lag(gnp.ts2)
OrangeCounty.diff <- ts(data.frame(employment = employment.ts2,
gnp = gnp.ts2, employment.lag, gnp.lag))
models <- auto_ardl(gnp ~ employment + gnp.lag + employment.lag,
data = OrangeCounty.diff, max_order = 12)
models$top_orders
```

```
##    gnp employment gnp.lag employment.lag       AIC
## 1    5          4       5              4 -4701.447
## 2    5          4       4              4 -4701.447
## 3    5          4       5              3 -4701.447
## 4    5          4       5              5 -4699.975
## 5    5          4       4              5 -4699.975
## 6    3          4       3              3 -4689.928
## 7    3          4       2              3 -4689.928
## 8    3          4       3              2 -4689.928
## 9    1          1       1              1 -4689.892
## 10   1          1       0              1 -4689.892
## 11   1          1       1              0 -4689.892
## 12   1          2       1              1 -4669.982
## 13   3          3       3              3 -4651.957
## 14   2          1       2              1 -4645.498
## 15   2          1       1              1 -4645.498
## 16   2          1       2              0 -4645.498
## 17   2          1       2              2 -4643.500
## 18   2          1       1              2 -4643.500
## 19   2          0       2              1 -4627.278
## 20   4          4       4              4 -4557.763
```

After looking at comparisons of different models, ARDL(5,4) had the lowest AIC and was therefore chosen as the ARDL model that best fits the data.

```r
ARDL <- ardl(gnp ~ employment, order(3, 4), data = OrangeCounty.diff)
ardlresid <- resid(ARDL)
acf(ardlresid, lag.max = 5, main = "Residuals of ARDL(3,4)")
```




```r
ljung_box_test3 <- Box.test(ardlresid, lag = 3, type = "Ljung-Box")
ljung_box_test3
```

```
##
##  Box-Ljung test
##
## data:  ardlresid
## X-squared = 0.84402, df = 3, p-value = 0.8389
```

The ACF of the residuals for the ARDL model has no significant lags, and therefore demonstrates no signs of autocorrelation. Since there is no autocorrelation in the residuals, the model is a good fit. The Box-Ljung test, a more formal test for autocorrelation, also confirms that there is no autocorrelation in the ARDL(5,4) model because the p-value (0.8389) is greater than 0.05.

Question 4

Based on the above data, we found an AR(1) model in addition to an ARDL(5,4) model. Testing the ARDL model using the Box-Ljung test, a more formal test for detecting autocorrelation in the residuals, resulted in a p-value of 0.9075 which is greater than the standard alpha value of 0.05. This indicates there is no autocorrelation in the ARDL model. The ACF of the residuals for the ARDL model does not depict any significant lags in the model as well as the AIC value for the ARDL model being the lowest at almost -5000, providing evidence that it is also a good fit for the data. As a result, we can make the conclusion that the ARDL(5,4) model fits our data better than the AR(1) model.

Question 5

Although initial ndiff() tests stated that employment only needed to be differenced once, employment still showed evidence of being non-stationary. As a result, it differenced once more to account for being potentially non-stationary as well as the possibility of seasonality. When creating our model, we used quarterly data which may include some component of seasonality that may not be accounted for which can skew our estimates. Seasonality can be improved, or normalized, by further differencing employment to sea a clearer trend analysis. This effect was seen with the generation of a p-value of 0.01, which is below the alpha value of 0.05, suggesting employment has become stationary. Furthermore, differencing employment twice instead of once (as suggested) could have created discrepancies that influenced the succeeding data and tests carried out.