Project 2

Juliesen Jaime, Rachel Asher, Caden Campbell

2023-11-10

```
##    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  
```

Exploratory Data Analysis

(a) Briefly discuss the question you are trying to answer.

Our team is interested in answering how GNP and past employment affects employment in Orange County.

Question: How do GNP and past employment levels influence present employment in Orange County?

(b) Cite the dataset and give a summary of what the dataset is about

We will be using the data set “OrangeCounty” which includes 2 variables: employment and gnp. Employment is measured quarterly in Orange county from 1965–1983. GNP is also measured quarterly and refers to real GNP.

(c) First check for completeness and consistency of the data (if there are NAs or missing observations, replace with the value of the previous observation; make a note of this)

After checking for completeness and missing observations, we have none in employment and gnp.

(d) Provide descriptive analyses of your variables. This should include the histogram with overlying density, boxplots, cross correlation. All figures/statistics must include comments.

For the histogram for employment, (orange.ts[,1]), there is a possibilty that it is slightly skewed right with larger values of data ranging from 300-600 and a significant portion in values 800-900. The histogram for GNP, (orange.ts[,2]), is normally distributed and shows some symmetry. However, the data fluctuates from low to high amounts as shown by the decreases and increases in columns. The range of the employment boxplot is 595.1 and the average value is 584.0. The range of the GNP boxplot is 665.3 and the average value is 1252.2. The GNP boxplot is fairly symmetrical which indicates a likelihood the distribution is normal while the employment boxplot is slightly skewed right.The cross-correlation plot measures the correlation between the current time period and a lag in either direction. It is symmetrical because a lag is equivalent to a forecast. As shown, the further away from the current time period, the less autocorrelation exists. However, it is worth nothing that this plot is measuring the ACF, so it includes the contributions of all intermediate periods between a lag and time 0. To give an example, the lag of -4 exhibits a correlation of approximately 0.4 with the current time period.

```r
#c
missing_values <- sum(is.na(OrangeCounty))

missing_values_per_column <- colSums(is.na(OrangeCounty))

cat("Total missing values in the dataset: ", missing_values, "\n")
```

```
## Total missing values in the dataset:  0
```

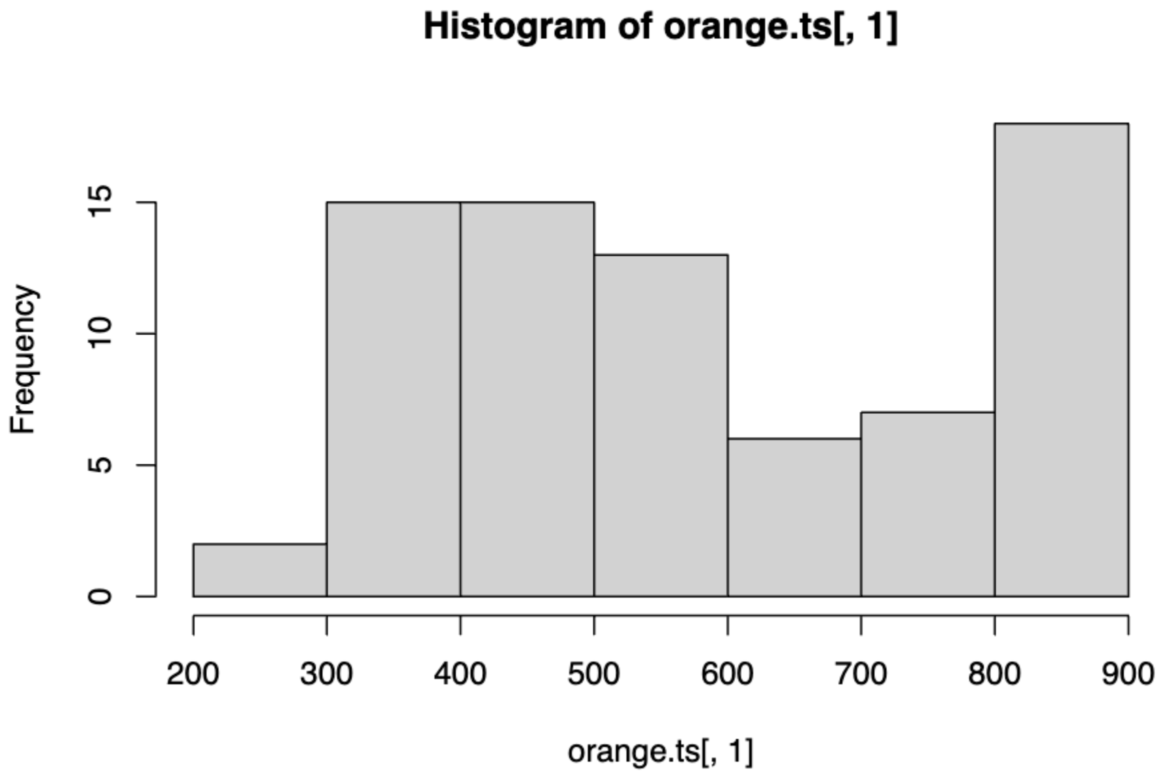
```r
cat("Missing values in each column:\n")
```

```
## Missing values in each column:
```

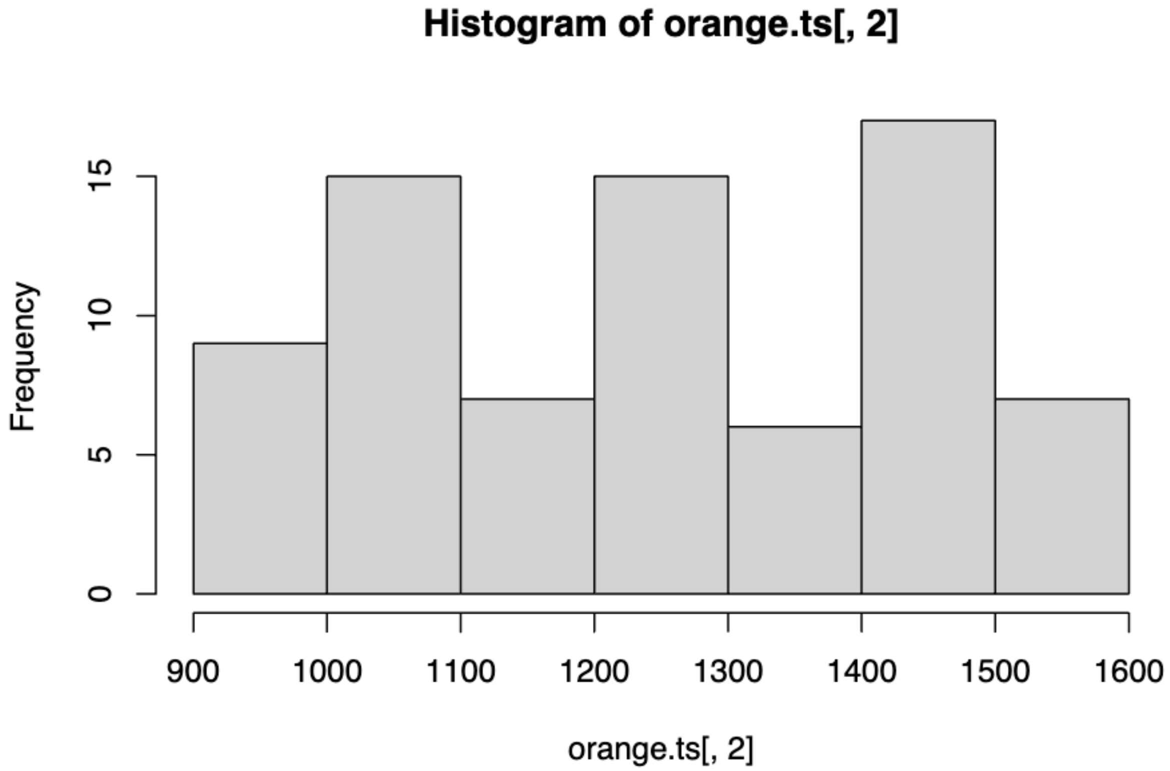
```r
print(missing_values_per_column)
```

```
## employment        gnp
##          0          0
```

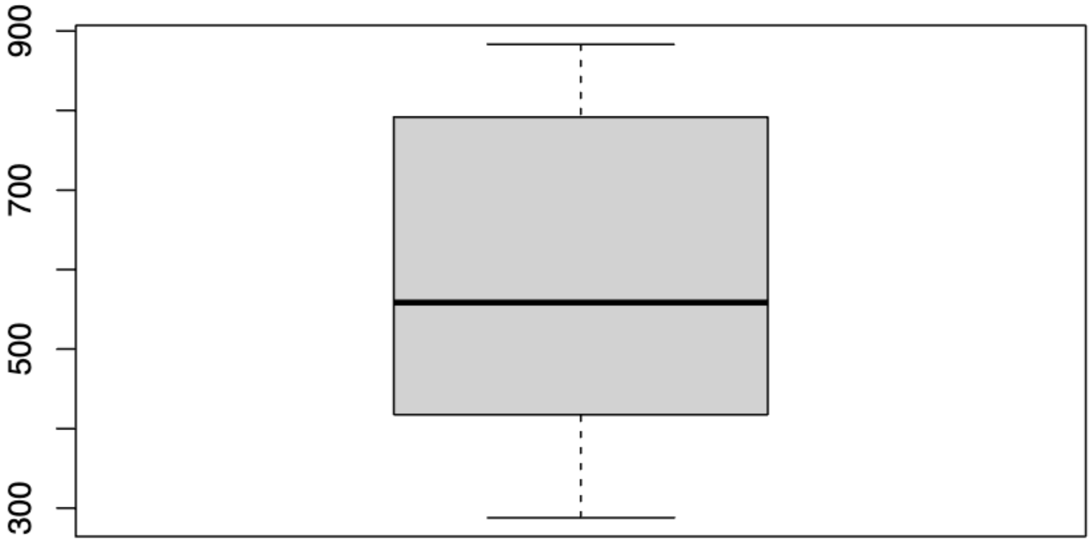
```r
#d

#histogram
hist(orange.ts[,1])
```



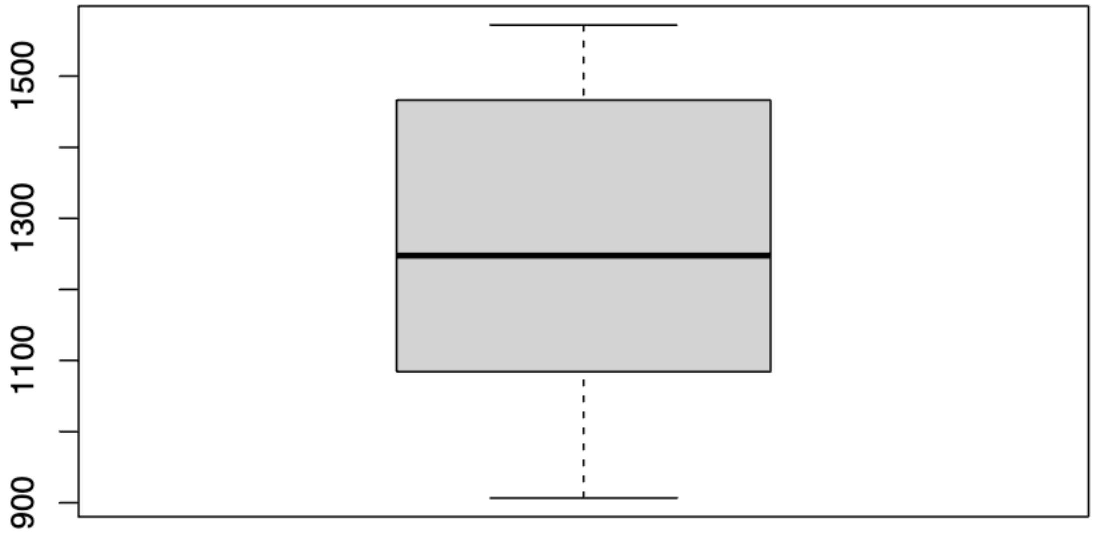
```r
hist(orange.ts[,2])
```



```r
#boxplot
boxplot(orange.ts[,1])
```



```
boxplot(orange.ts[,2])
```



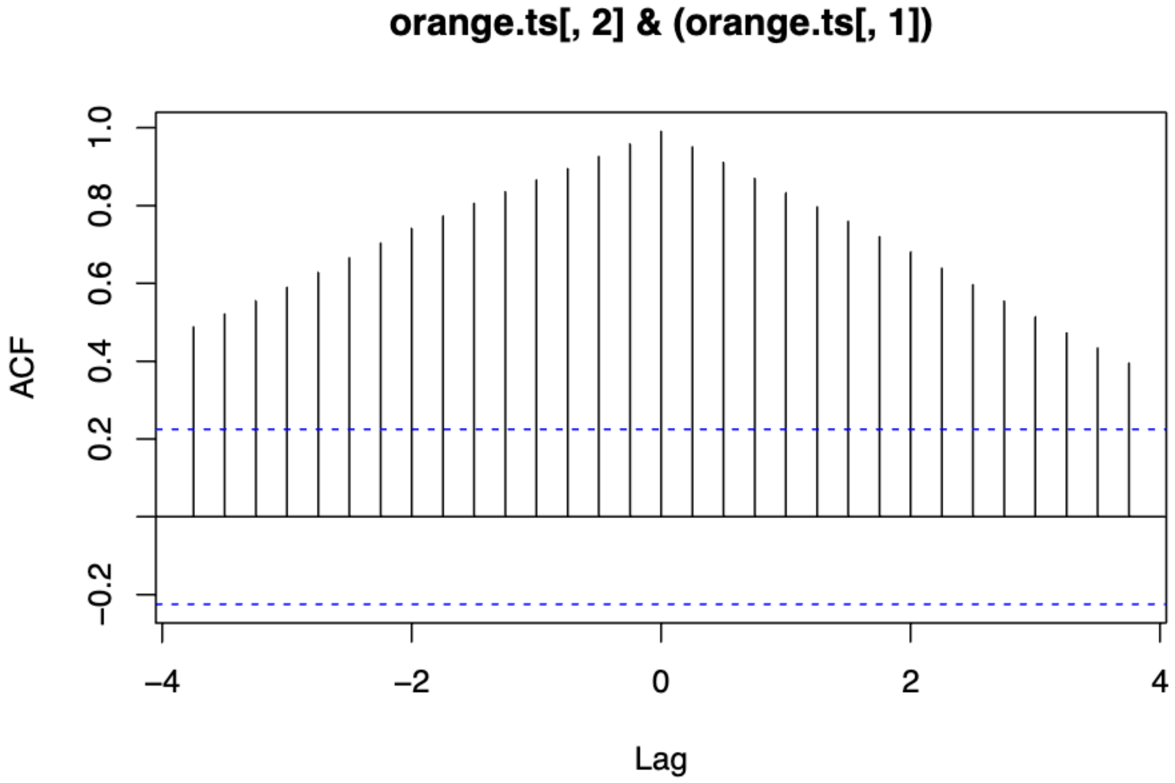
```
summary(orange.ts[,1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   288.0   417.9   558.5   584.0   788.1   883.1
```

```
summary(orange.ts[,2])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   906.6  1084.5  1247.6  1252.2  1465.3  1571.9
```

```
#cross correlation
ccf(orange.ts[,2],(orange.ts[,1]))
```



Data PreProcessing

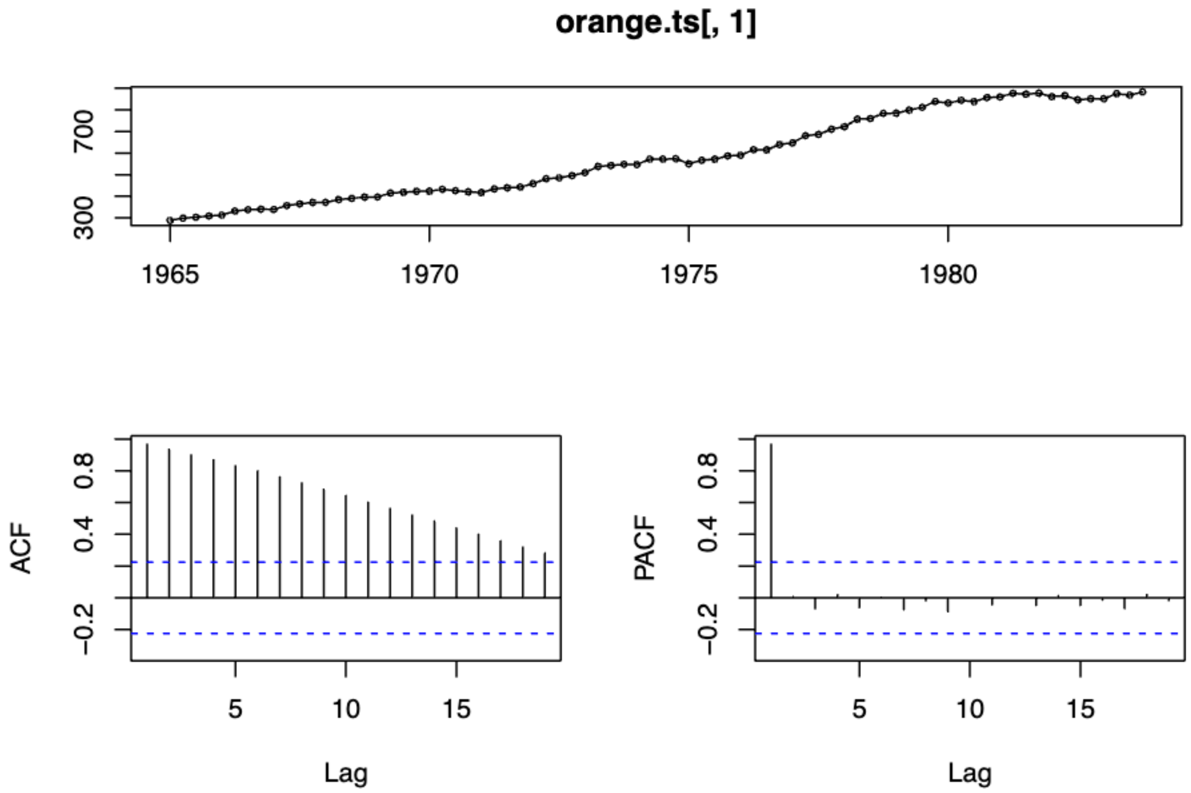
(a) With tsdisplay or ggtsdisplay, for each variable, use its time series plot, ACF and PACF to comment on its stationarity (you can also decompose the time series; note if there is seasonality). To supplement this, use the appropriate Dickey-Fuller (unit root) test, to determine whether or not it is stationary. Note using its PACF what the suspected order might be.

After using the tsdisplay to analyze the data, we can conclude that both variables, employment and GNP, are not stationary because there is an upwards trend and not white noise. Also, the ACF takes forever to die off which is an indicator of non-stationarity because the sample auto correlations remain large at long lags. Also, we used the ADF test in which both employment (p-value:0.09167) and GNP (p-value:0.08576) indicate that we fail to reject the null, meaning there is non-stationarity. Using the PACF, we can suspect that the order may be AR(1).

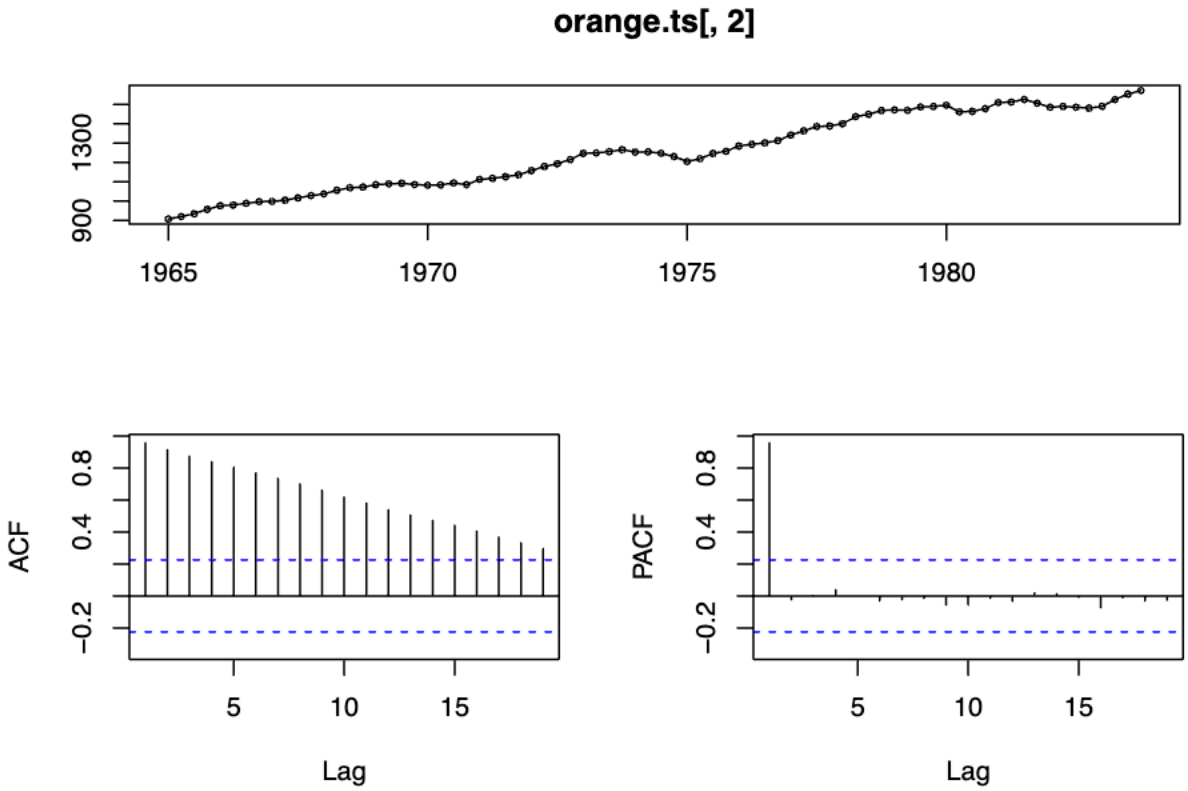
(b) If it is not stationary, determine the level of differencing to make our series stationary. We can use the ndiffs function which performs a unit-root test to determine this. After this, difference your data to ascertain a stationary time series. Re-do part a) for your differenced time series and comment on the time series plot, ACF and PACF. Recall that the time series models we've observed rely on stationarity.

Using the ndiffs function, we determined that both variables need to be differenced once. The plot now does not have an upward trend and is instead static. The ACF also does not take a long time to die off which indicates stationarity. The PACF for employment shows that this variable can be an AR(5) model. The PACF for GNP shows that it can be an AR(1) model. Although lag 12 is statistically significant, all the intermediate lags are not so it is not counted.

```
# A
data("OrangeCounty", package = "AER")

orange.ts <- ts(OrangeCounty, start=c(1965,1), end=c(1983,4),frequency=4)
orange.mod <- dynlm(d(gnp) ~ L(employment,0:4), orange.ts)
tsdisplay(orange.ts[,1])
```



```
tsdisplay(orange.ts[,2])
```



```
# test for stationary

adf.test(orange.ts[,1])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  orange.ts[, 1]
## Dickey-Fuller = -3.2166, Lag order = 4, p-value = 0.09167
## alternative hypothesis: stationary
```

```
adf.test(orange.ts[,2])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  orange.ts[, 2]
## Dickey-Fuller = -3.2533, Lag order = 4, p-value = 0.08576
## alternative hypothesis: stationary
```

```
# fail to reject null so it is non stationary

# B
```

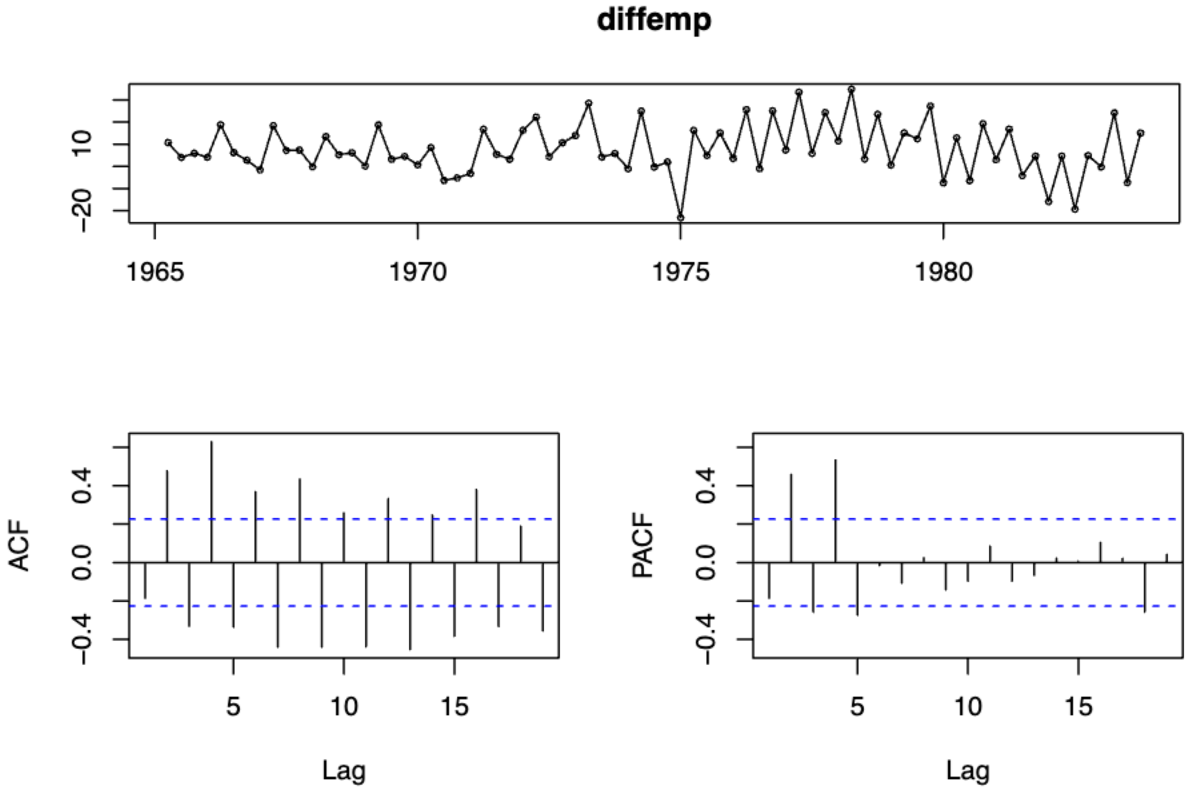
```
#number of times to be differences
ndiffs(orange.ts[,1])
```

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

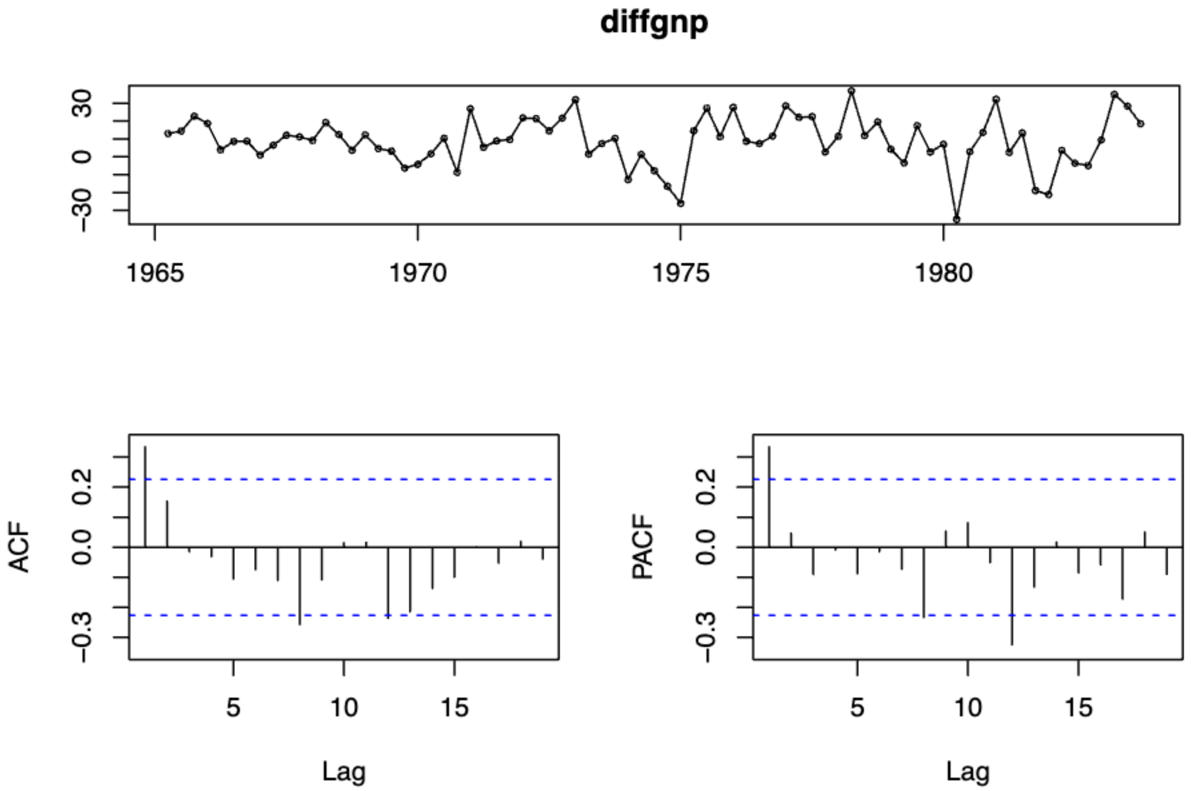
```
ndiffs(orange.ts[,2])
```

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

```
#difference the data
diffemp <- diff(orange.ts[,1])
diffgnp <- diff(orange.ts[,2])

tsdisplay(diffemp)
```



```
tsdisplay(diffgnp)
```



3. Feature Generation, Model Testing and Forecasting.

(a) Fit an AR(p) model to the data (using part 2(a), AIC or some built in R function)

After testing several AR models ranging from lags 2 to 5, we found that AR(5) had the lowest AIC criteria of 489.8945.This indicates that the AR(5) model would be the best fit for the data.

(b) Plot and comment on the ACF of the residuals of the model chosen in 3(a). If the model is properly fit, then we should see no autocorrelations in the residuals. Carry out a formal test for autocorrelation and comment on the results.

When looking at the ACF of the residuals, none of the lags are statistically significant which means there is no autocorrelation of the errors present. Also, we used the BG Test and got a p-value of 0.1443, meaning we failed to reject the null that there is no autocorrelation at the 5% significance level. Therefore, our errors are uncorrelated.

(c) Using the appropriate predictors, fit an ARDL(p,q) model to the data and repeat step (b) in part 3.

After fitting an ARDL model using the variable GNP from lags 1 to 5, we found that the ARDL model with the lowest AIC was ARDL (5,1) because it had the lowest value of 460.8806. This ARDL model fits our calculations of the PACF for each variable above where we indicated that employment should have 5 lags and GNP should have 1 lag.

```r
#A

#AR(2) model
employment.ar2 <- dynlm(d(employment) ~ L(d(employment),1:2), data = OrangeCounty)
AIC(employment.ar2)
```

```
## [1] 550.7529
```

```r
#AR(3) model
employment.ar3 <- dynlm(d(employment) ~ L(d(employment),1:3), data = OrangeCounty)
AIC(employment.ar3)
```

```
## [1] 539.9559
```

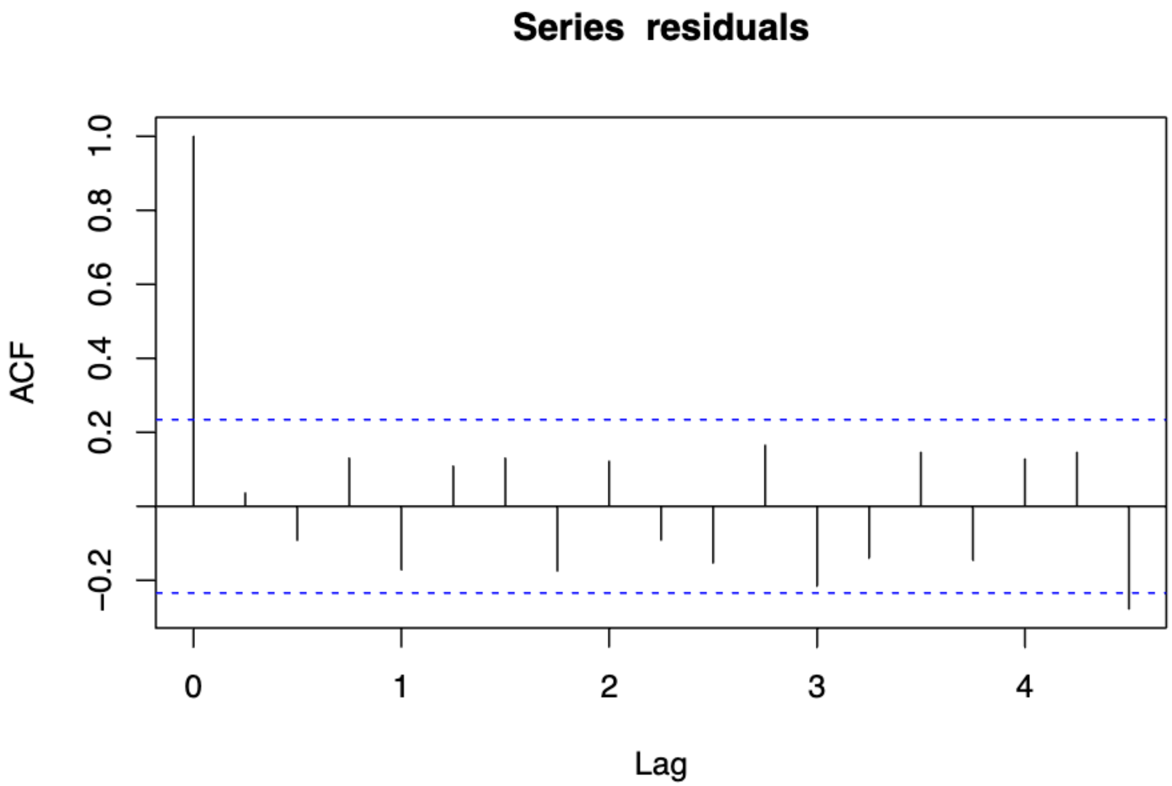
```r
#AR(4) model
employment.ar4 <- dynlm(d(employment) ~ L(d(employment),1:4), data = OrangeCounty)
AIC(employment.ar4)
```

```
## [1] 505.8736
```

```r
#AR(5) model
employment.ar5 <- dynlm(d(employment) ~ L(d(employment),1:5), data = OrangeCounty)
AIC(employment.ar5)
```

```
## [1] 489.8945
```

```r
#B

# are residuals correlated?
residuals <- resid(employment.ar5)
acf(residuals)
```



```r
bgtest(employment.ar5, order = 5)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 5
##
## data:  employment.ar5
## LM test = 8.2241, df = 5, p-value = 0.1443
```

```r
#C
gnp <- dynlm(d(gnp) ~ L(d(employment,1)), data = OrangeCounty)

employment.ard1 <- dynlm(d(employment) ~ L(d(employment),1:5) + L(d(gnp), 0:1), data = OrangeCounty)
employment.ard2 <- dynlm(d(employment) ~ L(d(employment),1:5) + L(d(gnp), 0:2), data = OrangeCounty)
employment.ard3 <- dynlm(d(employment) ~ L(d(employment),1:5) + L(d(gnp), 0:3), data = OrangeCounty)
employment.ard4 <- dynlm(d(employment) ~ L(d(employment),1:5) + L(d(gnp), 0:4), data = OrangeCounty)
employment.ard5 <- dynlm(d(employment) ~ L(d(employment),1:5) + L(d(gnp), 0:5), data = OrangeCounty)


AIC(employment.ard1)
```

```
## [1] 460.8806
```

```r
AIC(employment.ard2)
```

```
## [1] 462.0379
```

```r
AIC(employment.ard3)
```

```
## [1] 462.1871
```

```r
AIC(employment.ard4)
```

```
## [1] 463.7127
```

```r
AIC(employment.ard5)
```

```
## [1] 465.7119
```

4. Provide a brief summary of your findings and state which model performs better.

Our findings showed that the best model of fit was an ARDL model of (5,1) and the best model AR model for unemployment was an AR(5). This means that 5 lags of employment are significant and 1 lag of GNP is significant in affecting employment. Logically, this makes sense since past employment levels can affect current employment levels today. For example, if Orange County was in an economic downturn we can expect high unemployment rates for a few years. Furthermore, for GNP, the one lag may showcase that previous GNP from the previous year affects employment currently because the state of the economy as a whole reflects in the employment rates. Also, non-stationarity would be common in this data as shown in our findings since GNP and employment rates have increased over time due to technological innovations and increased economies of scales.

5. Suggest any limitations faced or improvements which could've been made to the model based on your findings, which should be supplemented with statistical tests(eg. degree of freedom restrictions, reverse causality).

One limitation of our model is that we didn't account for seasonality when evaluating our different model criteria. We found seasonality affecting the employment variable when looking at the Arima code which indicated that one degree of seasonality differncing is needed. We, however, we only differenced employment normally and not based on seasonailty which may affect our data and best model of fit.

```r
auto.arima(orange.ts[,1])
```

```
## Series: orange.ts[, 1]
## ARIMA(2,0,0)(0,1,1)[4] with drift
##
## Coefficients:
##          ar1      ar2     sma1   drift
##       1.4880  -0.5600  -0.4591  7.8204
## s.e.  0.0959   0.0972   0.1530  1.5542
##
## sigma^2 = 53.83:  log likelihood = -245.37
## AIC=500.74   AICc=501.65   BIC=512.13
```

```r
auto.arima(orange.ts[,2])
```

```
## Series: orange.ts[, 2]
## ARIMA(1,1,0) with drift
##
## Coefficients:
##          ar1   drift
##       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
```