STAT317 Assignment3

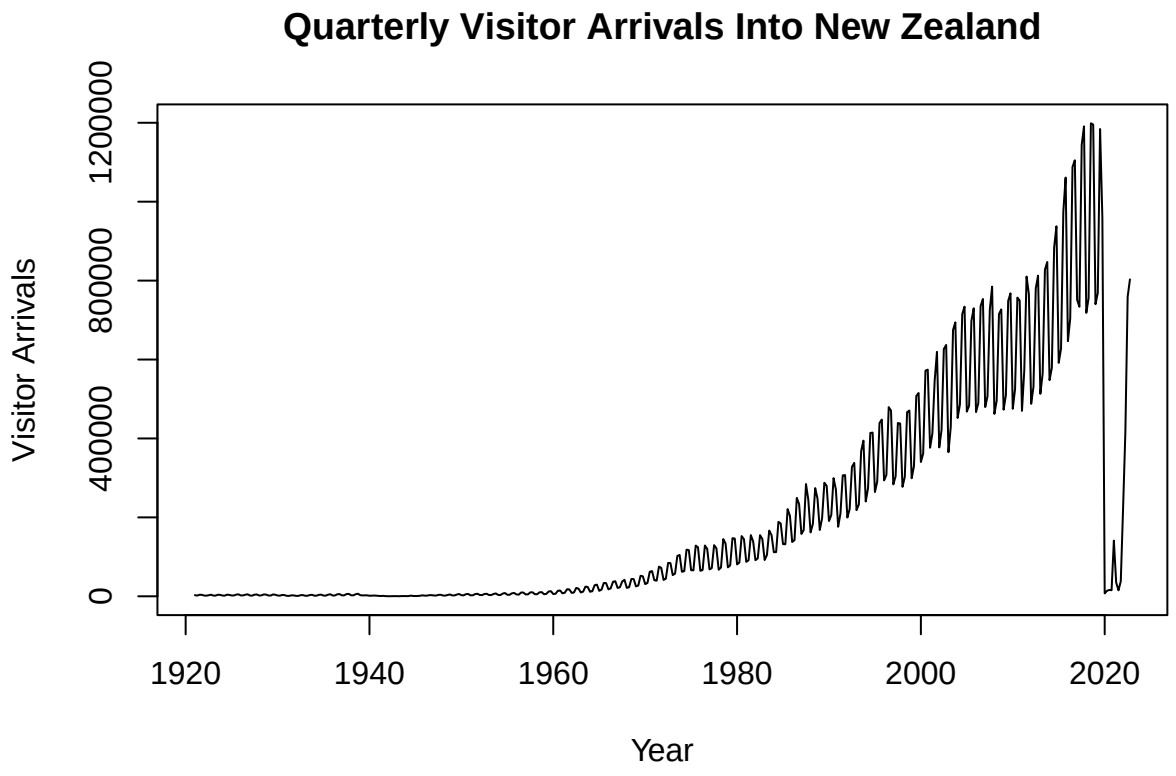
2023-09-29

```
library(forecast)
```

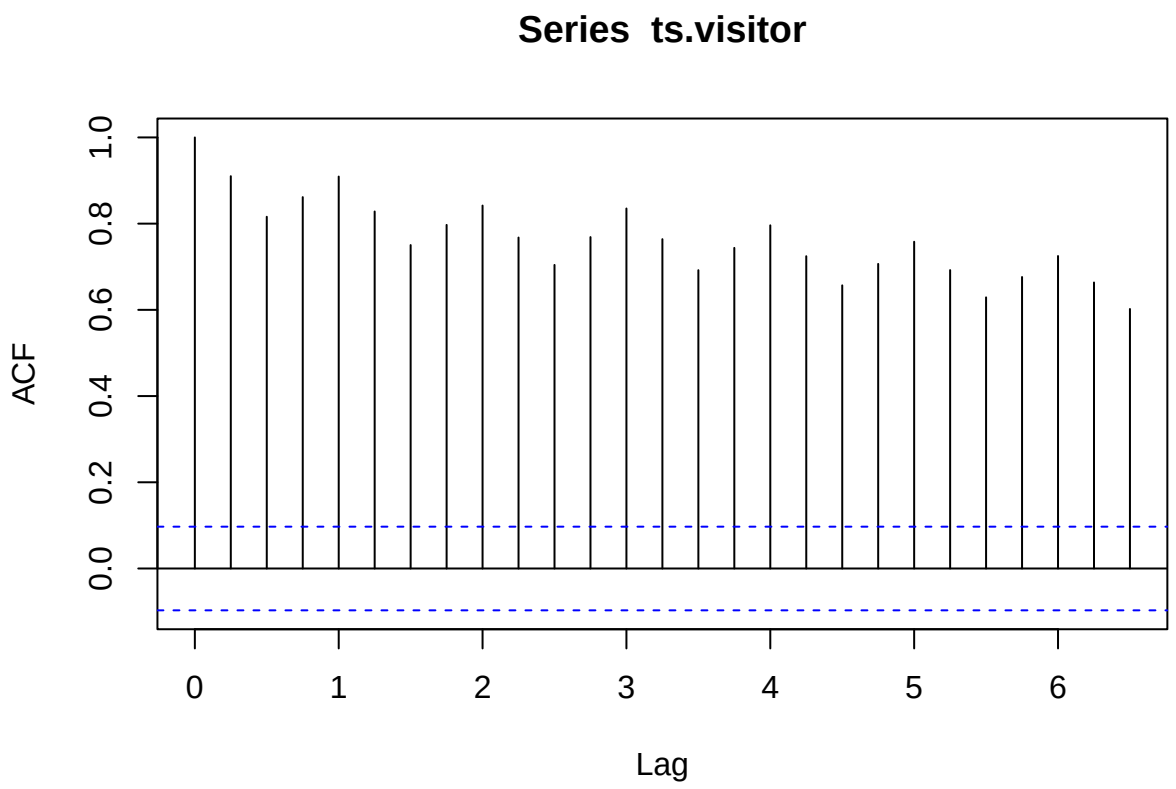
```
## Warning: package 'forecast' was built under R version 4.3.1
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

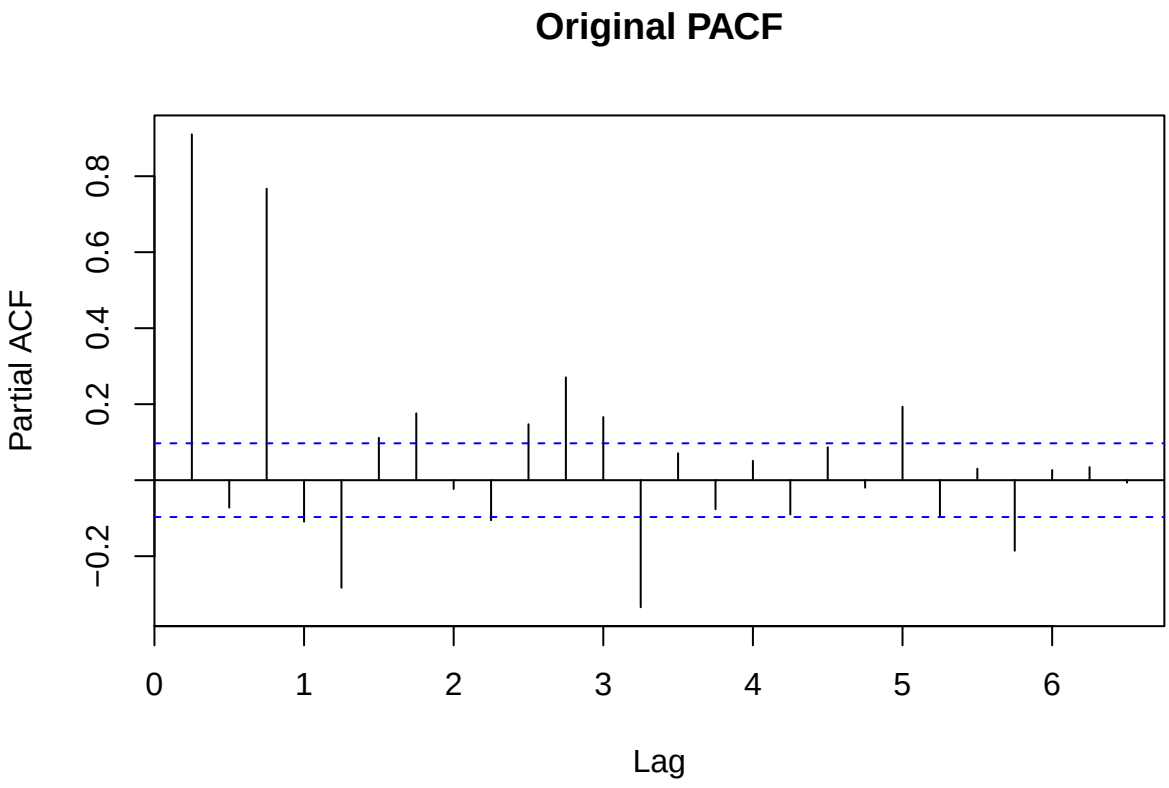
```
visitor <- read.csv("Visitors.csv")
ts.visitor <- ts(visitor$Actual.Counts, start= c(1921, 1), frequency = 4)
ts.plot(ts.visitor, xlab = "Year", ylab = "Visitor Arrivals", main =
"Quarterly Visitor Arrivals Into New Zealand")
```



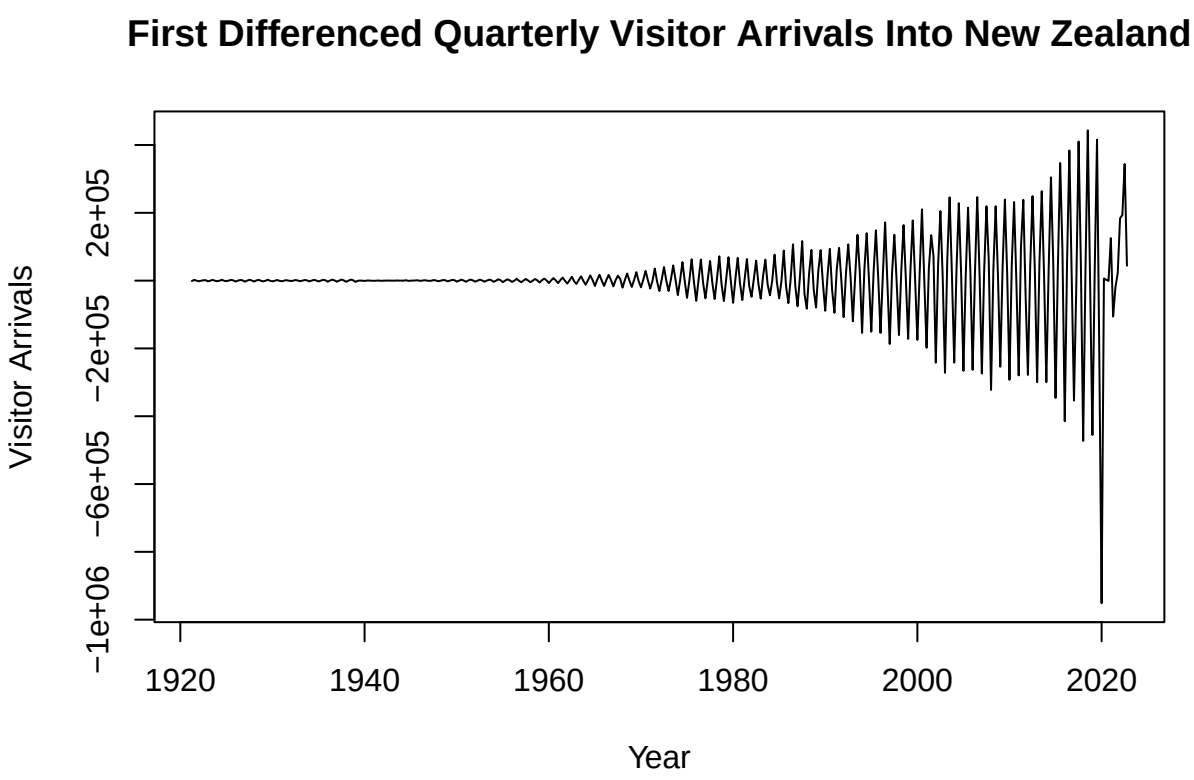
```
acf(ts.visitor)
```



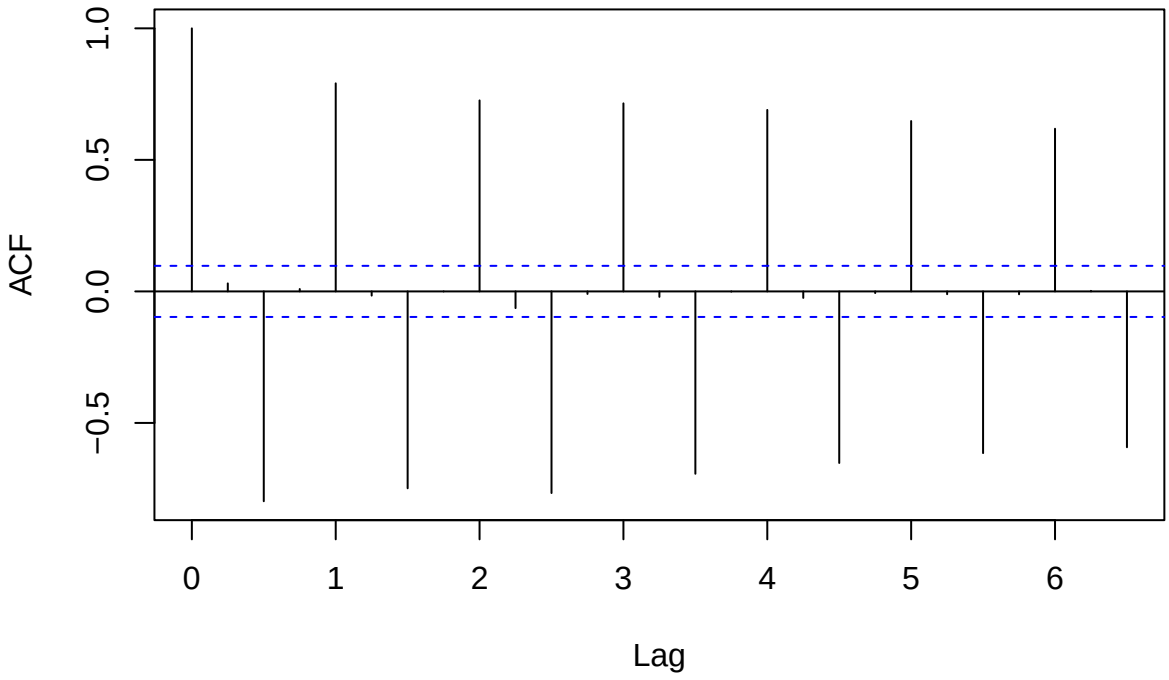

```
pacf(ts.visitor, main="Original PACF")
```



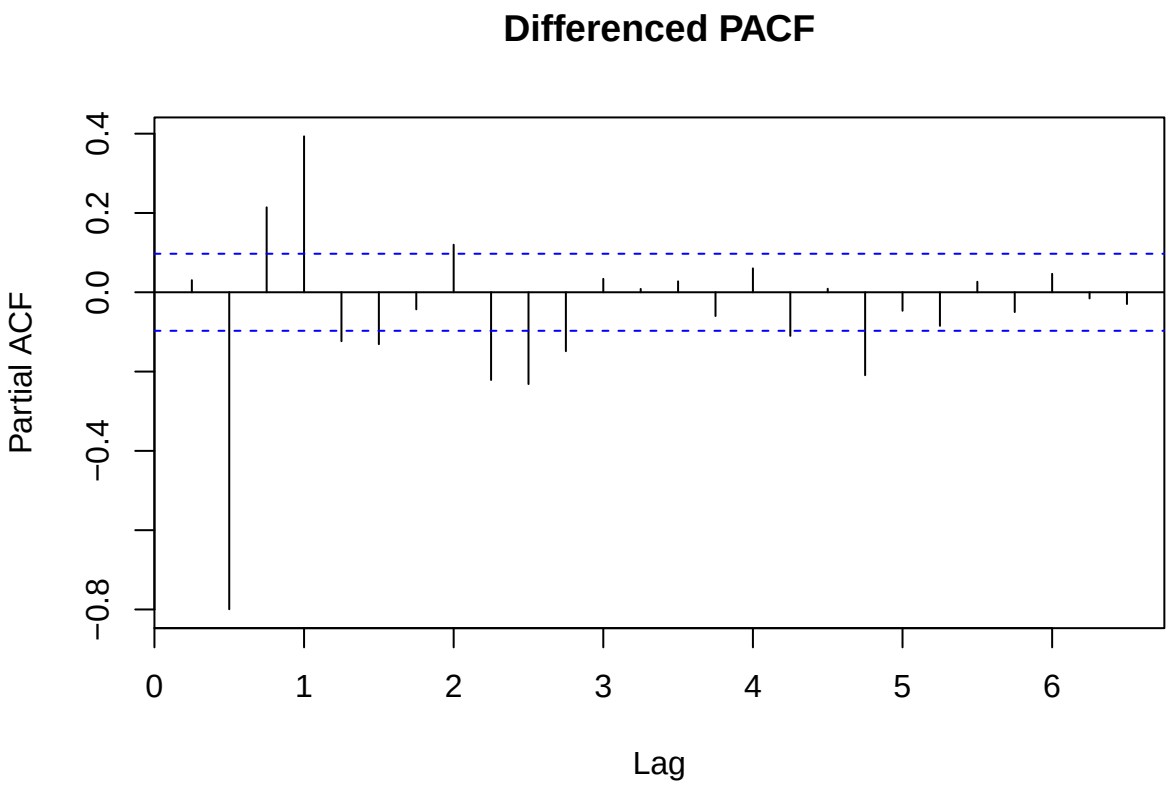

```
dvis <- diff(ts.visitor)
ts.plot(dvis, xlab = "Year", ylab = "Visitor Arrivals", main =
"First Differenced Quarterly Visitor Arrivals Into New Zealand")
```


First Differenced Quarterly Visitor Arrivals Into New Zealand
Visitor Arrivals
−1e+06
−6e+05
−2e+05
2e+05
1920
1940
1960
1980
2000
2020
Year

```
acf(dvis)
```




```
pacf(dvis, main = "Differenced PACF")
```



The diferenced plot does not have a positive trend like the original series. This is because the trend is removed when we take the first difference of a series.

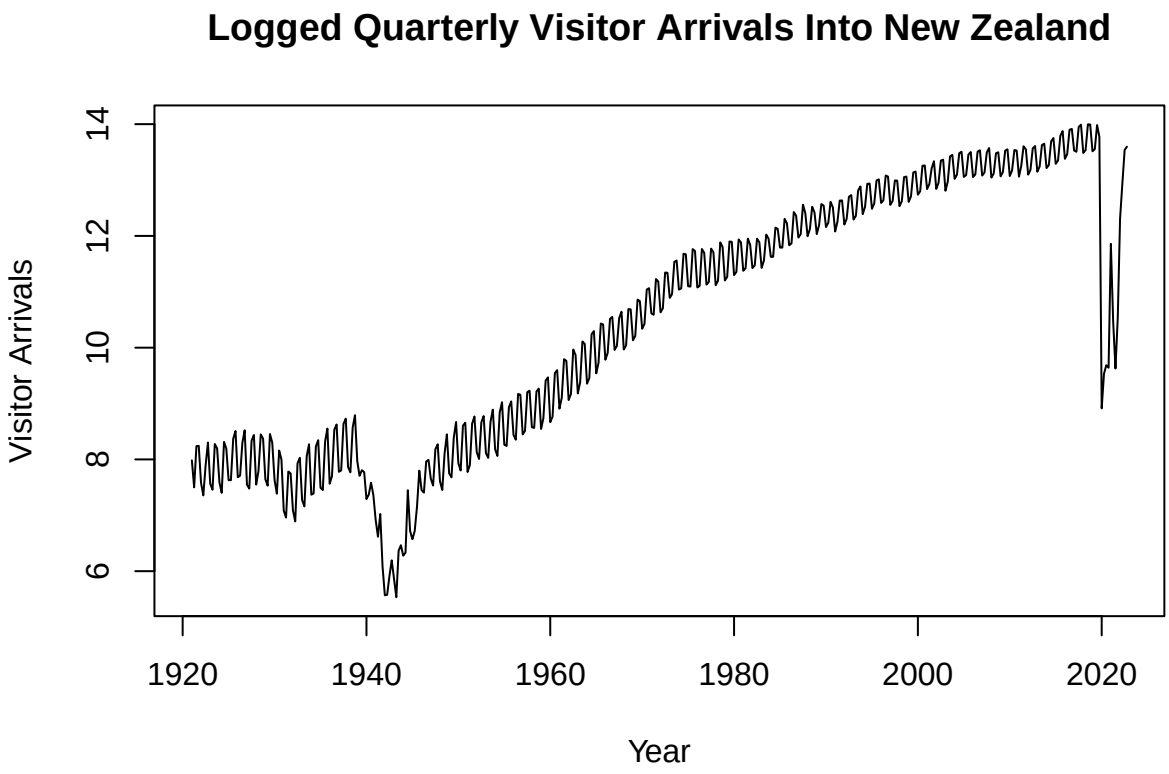
The time series plot in the differenced series the variance of visitor arrivals has consistently increased from 1960 until the present day. This would be due to improvements to technology in commercial air travel. These improvements have allowed for companies to have a larger, more efficient fleet of planes which would bring down prices of tickets for overseas travel.

All spikes in the original series are significant. This means that the current amount of visitor arrivals into New Zealand are highly correlated with all quaters before. This is due to the positive trend of the series. The spikes are gradually trending downwards. This means the number of visitor arrivals has less impact on the current number of visitor arrivals as we look further into the psat. There is a spike every fourth quater. This indicates a seasonal pattern to number of visitor arrivals. The spike occurs in the summer time. This is because christmas is in the summer, so any New Zealand citizens living overseas are coming to New Zealand for Christmas. It could also be because many of New Zealand's tourist attractions such as hiking and water based activities rely on warm weather.

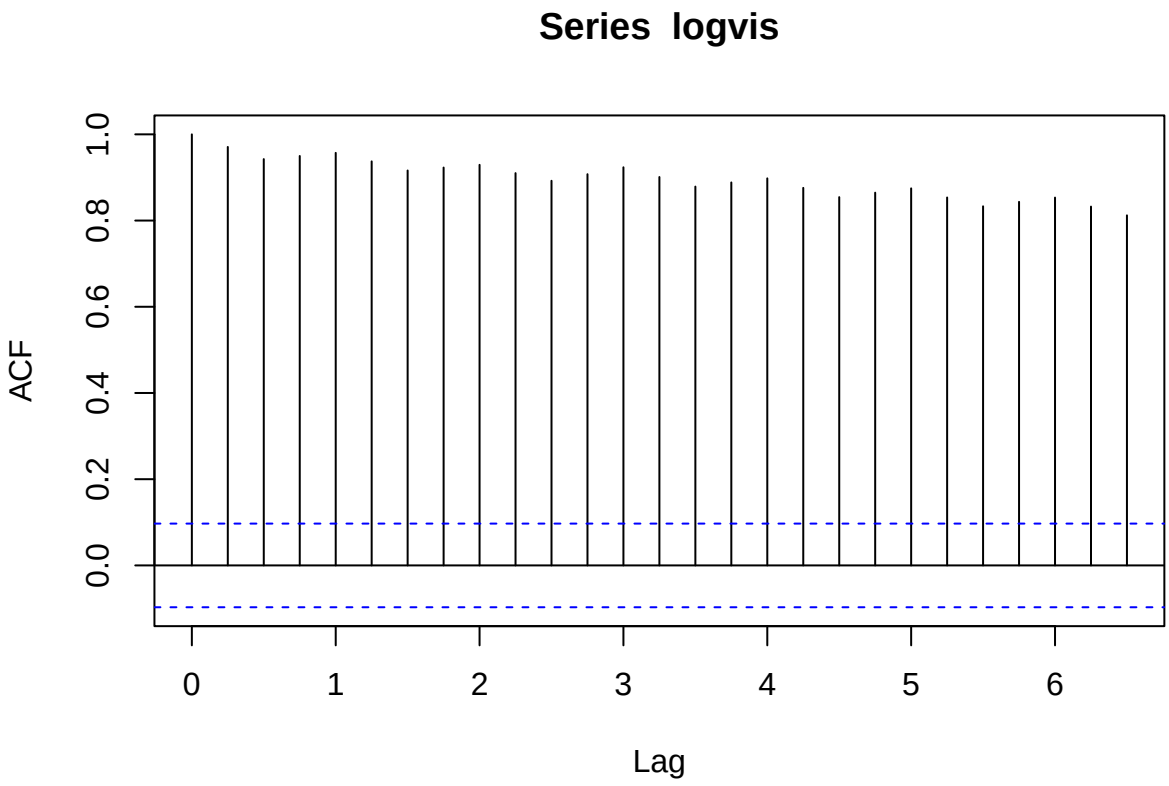
The ACF for the differenced series shows the current number of visitor arrivials is significantly correlated by the first quater of every year while every third spike is significantly negative. This suggests a seasonal pattern where there is highervisitor arrivials in the summer and lower visitor arrivials in the winter.

The first significant spike for the original pacf occurs at lag 1 and is positive while the first significant spike for the differenced plot occurs at lag 2 and is negative.

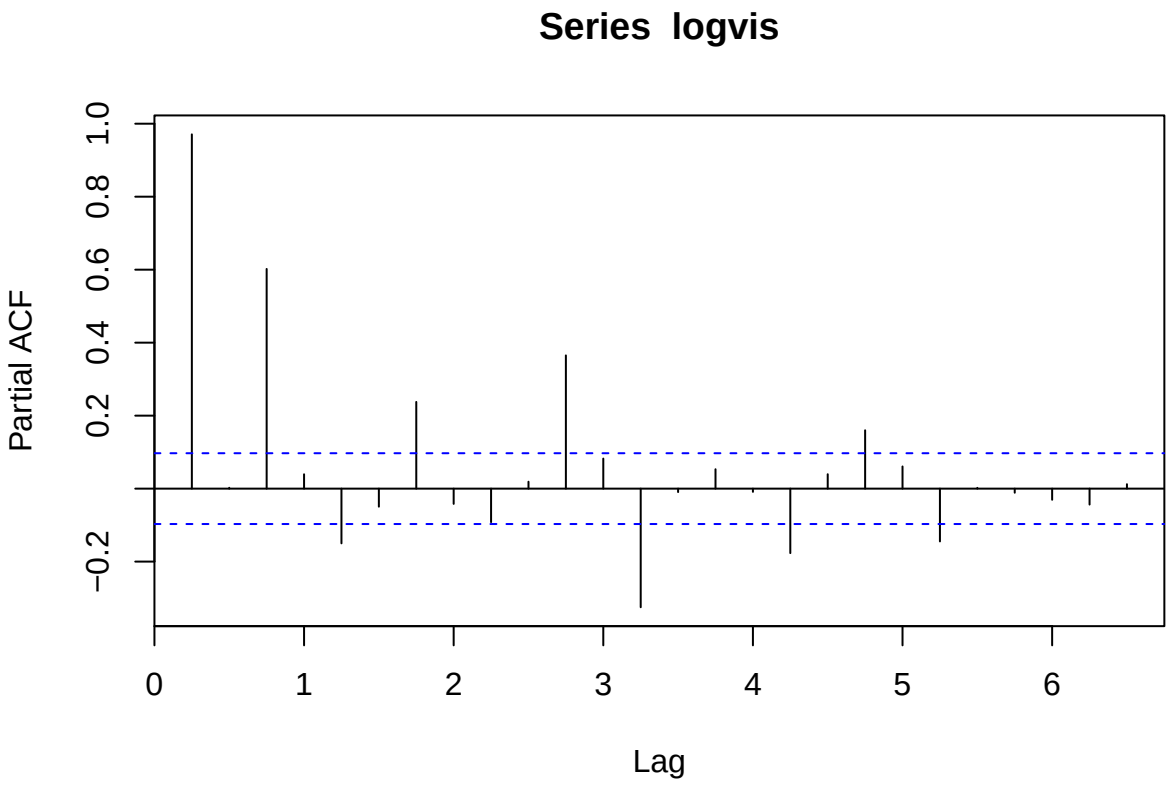
```
logvis <- log(ts.visitor)
ts.plot(logvis, xlab = "Year", ylab = "Visitor Arrivals", main =
"Logged Quarterly Visitor Arrivals Into New Zealand")
```


Logged Quarterly Visitor Arrivals Into New Zealand
Visitor Arrivals
6
8
10
12
14
1920
1940
1960
1980
2000
2020
Year

```
acf(logvis)
```




```
pacf(logvis)
```

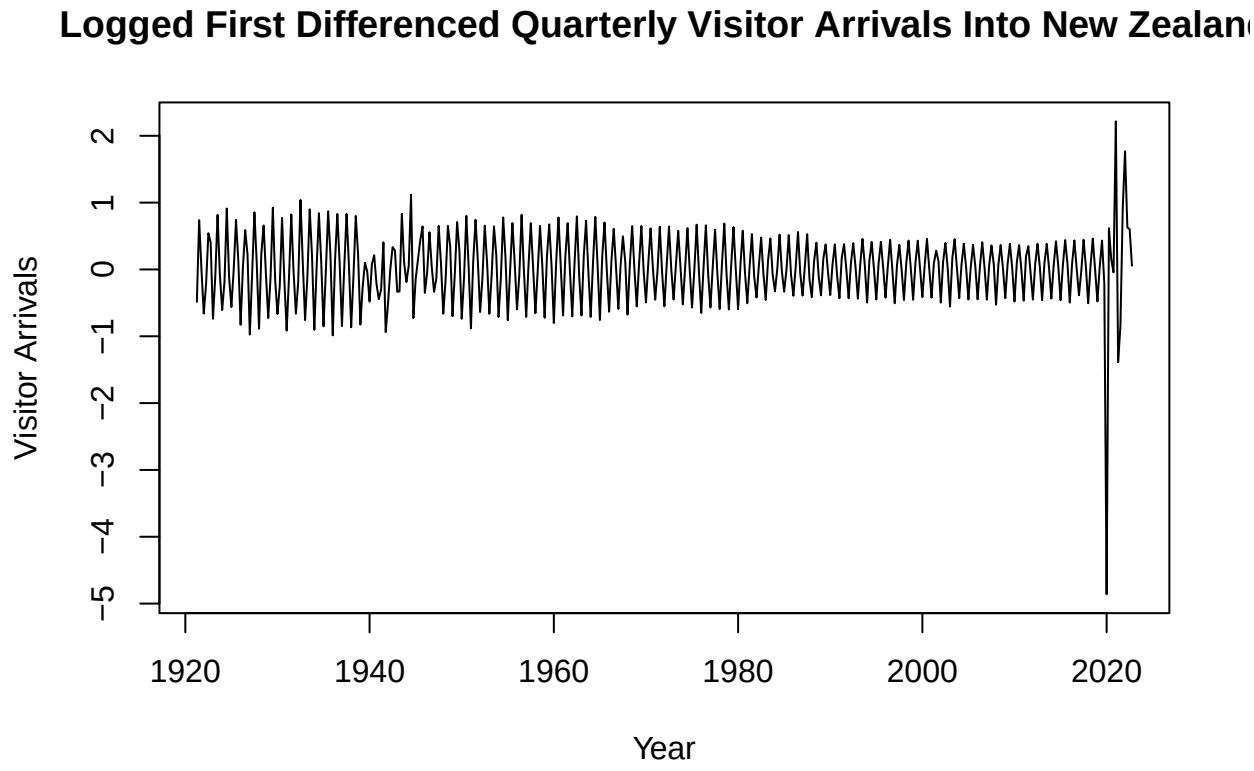


The series has an upwards trend and constant variance. Our first series had a changing mean and changing variance. We can see three distinct dips in this series that were not all visible in the first two plots. The first occurs around 1929 which can be explained by the great depression. The second is a large trough in 1939 which can be explained by the second world war. The last big trough occurs in 2020 which can be explained by COVID-19.

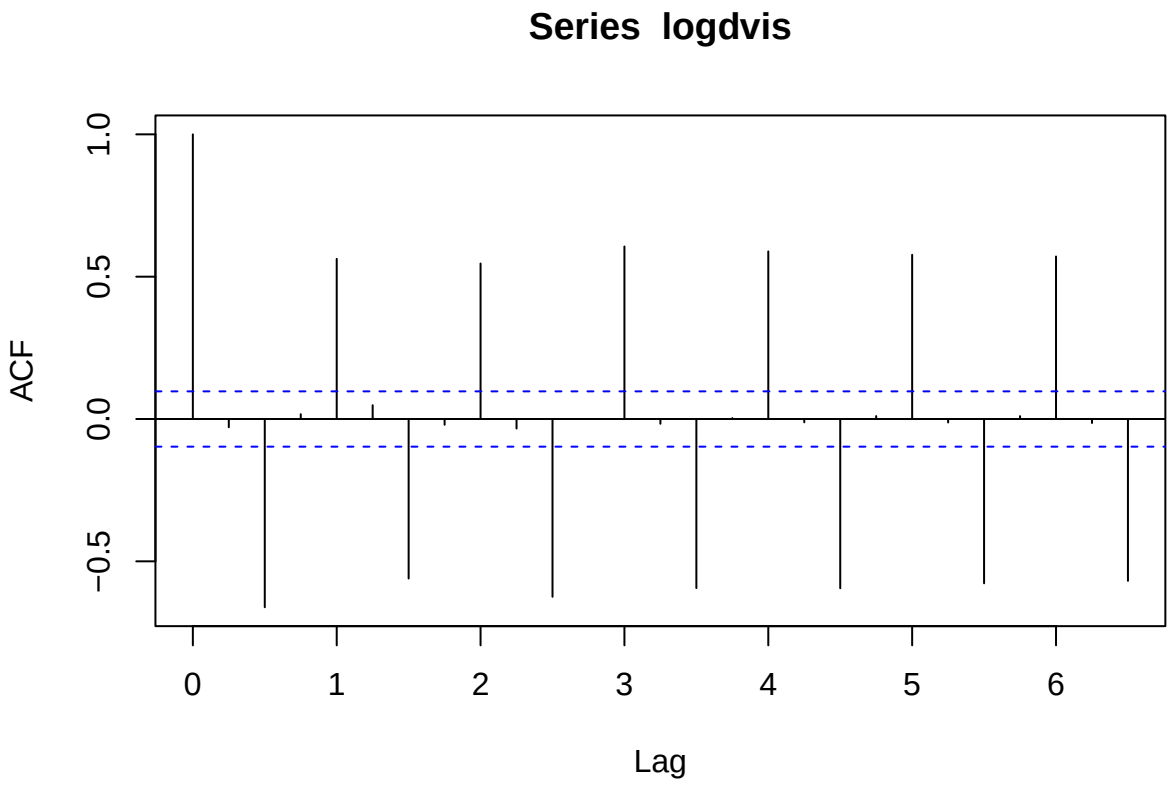
The ACF plot is very similar to our original ACF as all spikes are significantly positive, trending downwards whith a peak every four quaters. This is once again due to upwards trend in the series.

The pacf in the logged plot shows a clear seasonal pattern which was not shown on the previous plots. Every first and third spike is significant while every second and fourth spike is insignificant.

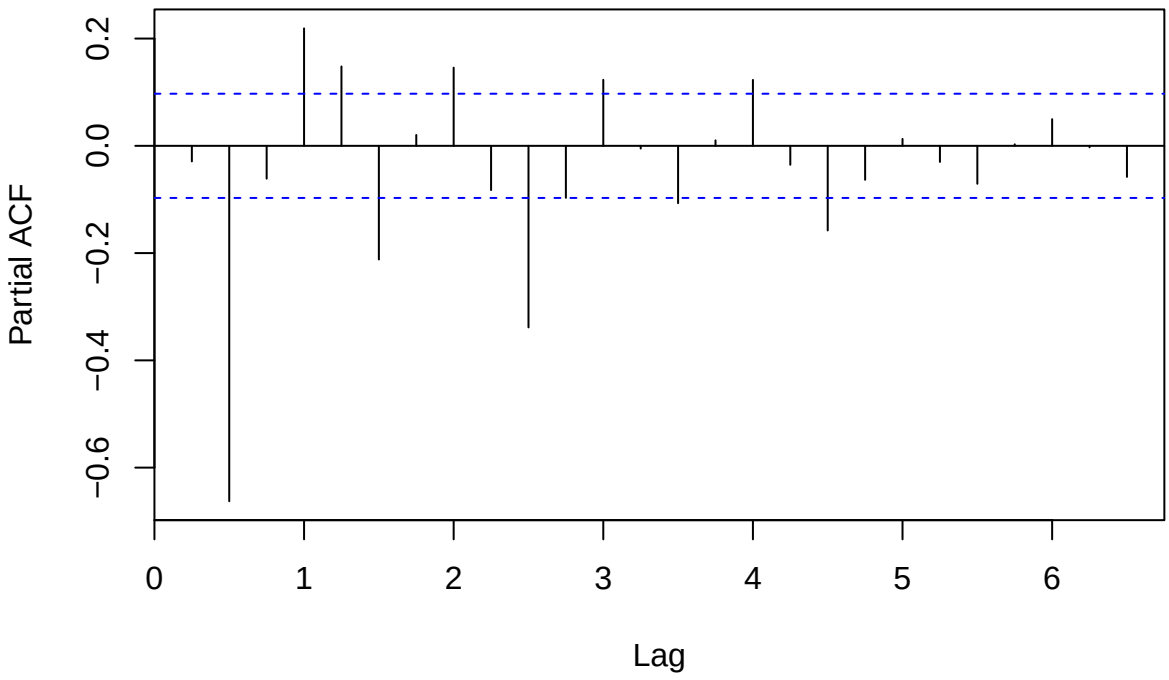
```
logdvis <- diff(logvis)
ts.plot(logdvis, xlab = "Year", ylab = "Visitor Arrivals", main =
"Logged First Differenced Quarterly Visitor Arrivals Into New Zealand")
```


Logged First Differenced Quarterly Visitor Arrivals Into New Zealan
Visitor Arrivals
2
1
0
−1
−2
−3
−4
−5
1920
1940
1960
1980
2000
2020
Year

```
acf(logdvis)
```



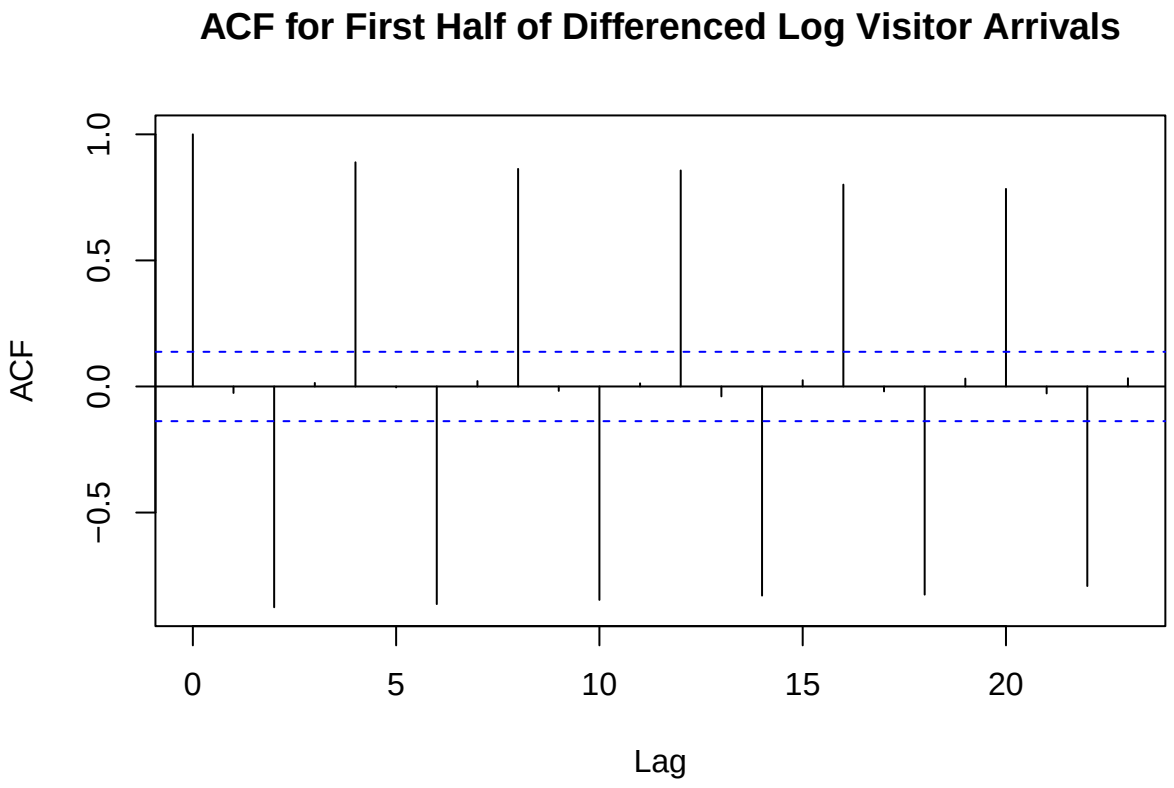

```
pacf(logdvis)
```



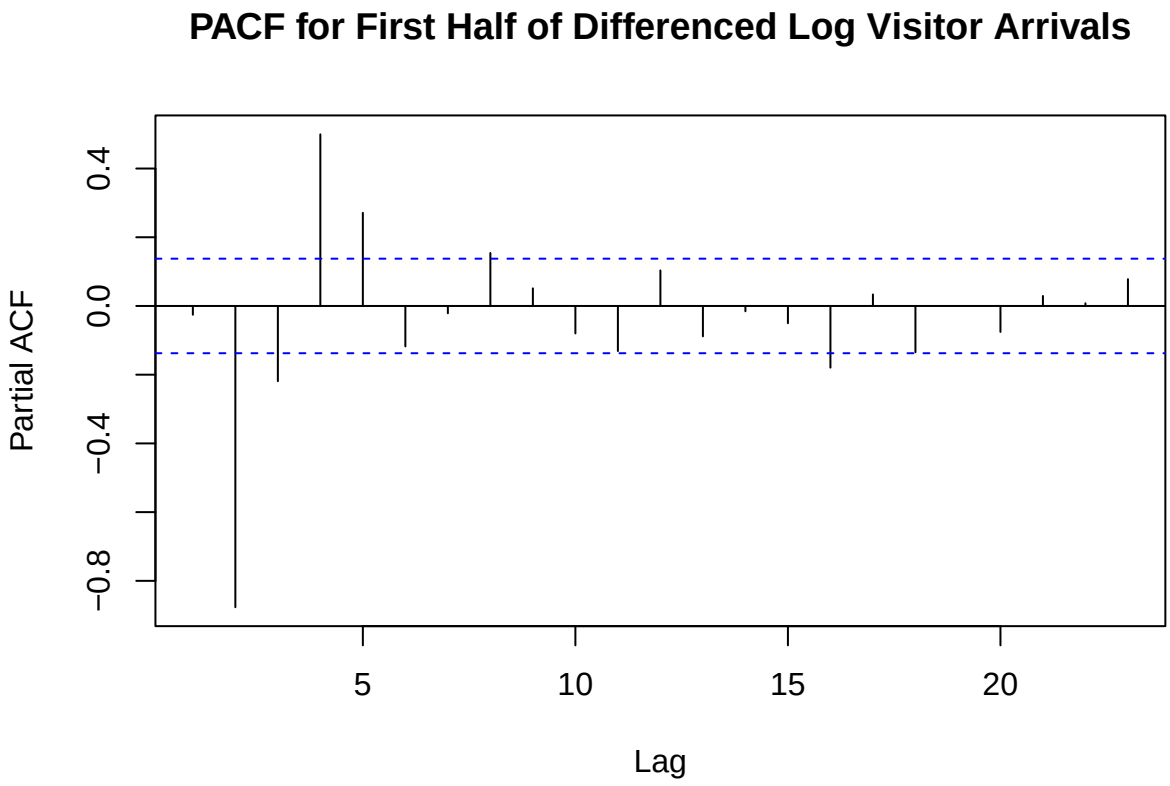
The logged differenced series has constant mean and constant variance. The variance from 1980 to the present day is less than the variance pre 1980. This suggests that the variation of the trend has decreased, meaning there was consistent growth of visitor arrivals during this time. There was also decreased variance around 1940 which was not shown on the other plots. The magnitude of the impact from covid is also shown to be march larger on this graph.

The acf is similar to the differenced acf as the significant spikes oscilate on the first and third lag values which indicates a seasonal pattern. Unlike the differenced acf the magnitude of the signicant spikes does not decrease after lag 5. This means the number of visitor arrivals from lag 5 has the same impact on visitor arrivals as lag 23.

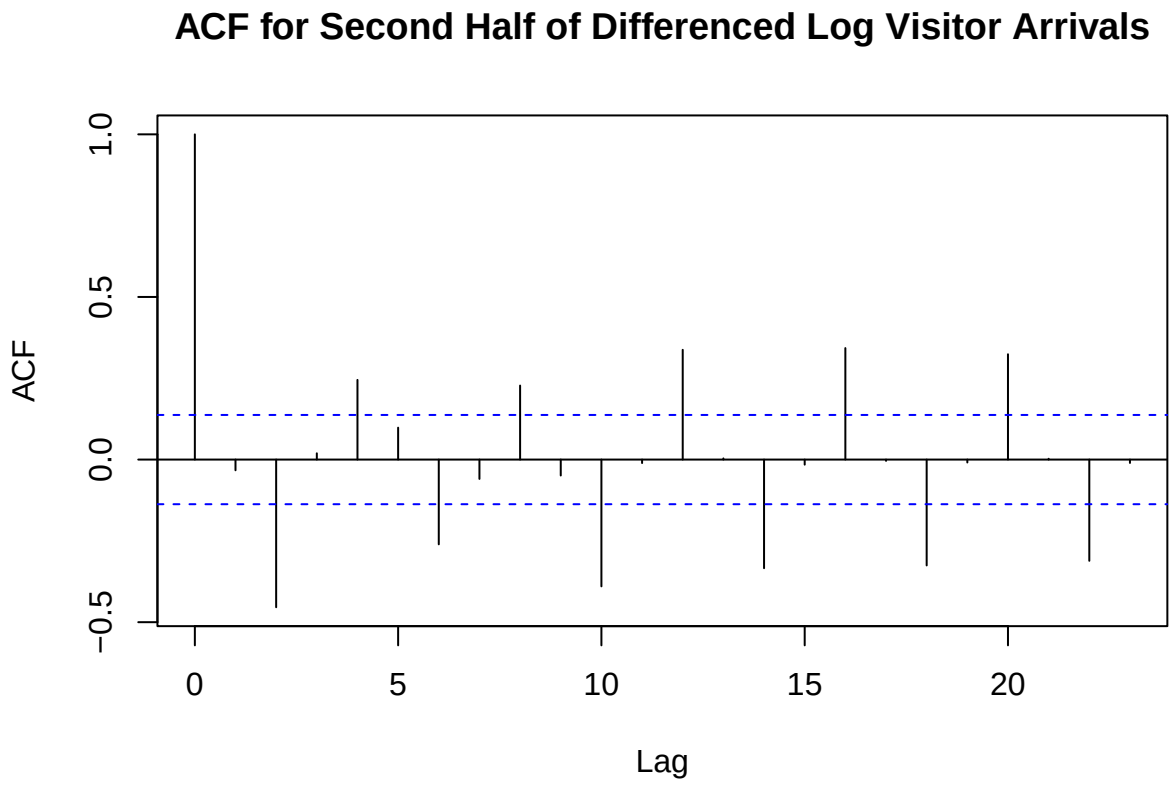
```
halfway <- floor(length(logdvis) / 2)
first_half_log_diff <- logdvis[1:halfway]
second_half_log_diff <- logdvis[(halfway+1):length(logdvis)]

acf(first_half_log_diff, main = "ACF for First Half of Differenced Log Visitor Arrivals")
```



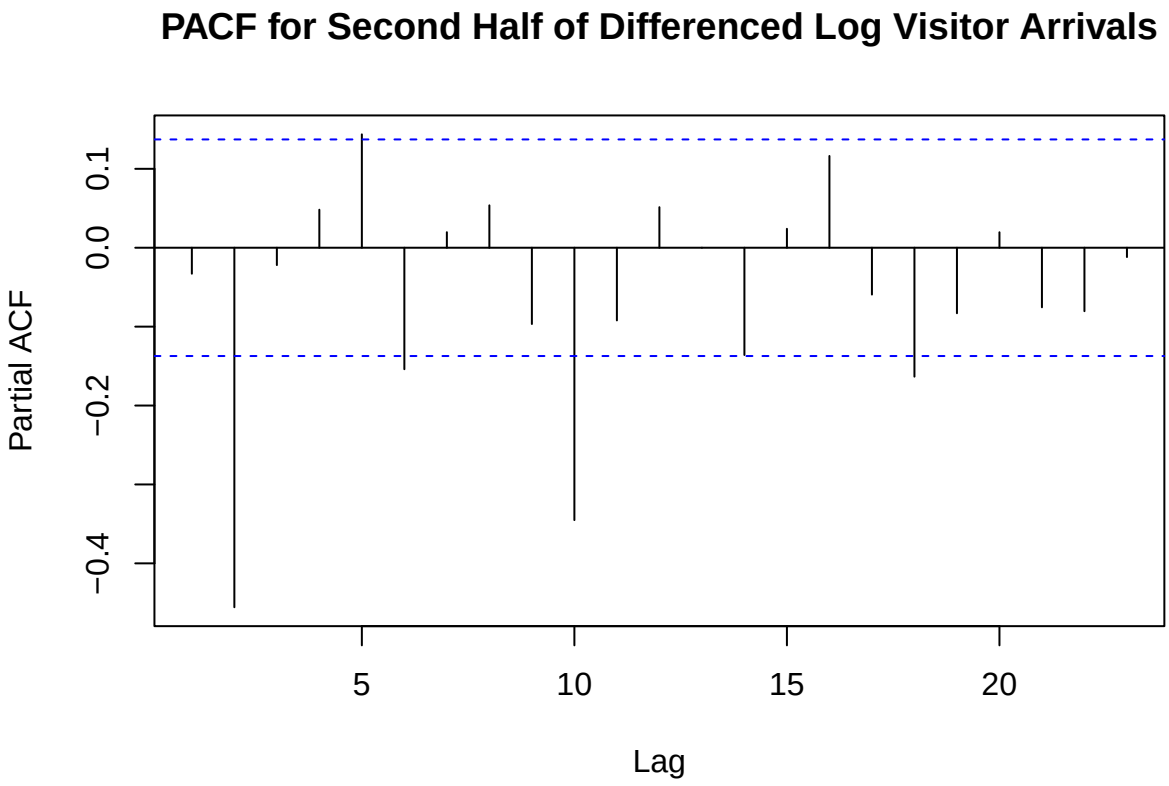

```
pacf(first_half_log_diff, main = "PACF for First Half of Differenced Log Visitor Arrivals")
```




```
acf(second_half_log_diff, main = "ACF for Second Half of Differenced Log Visitor Arrivals")
```




```
pacf(second_half_log_diff, main = "PACF for Second Half of Differenced Log Visitor Arrivals")
```

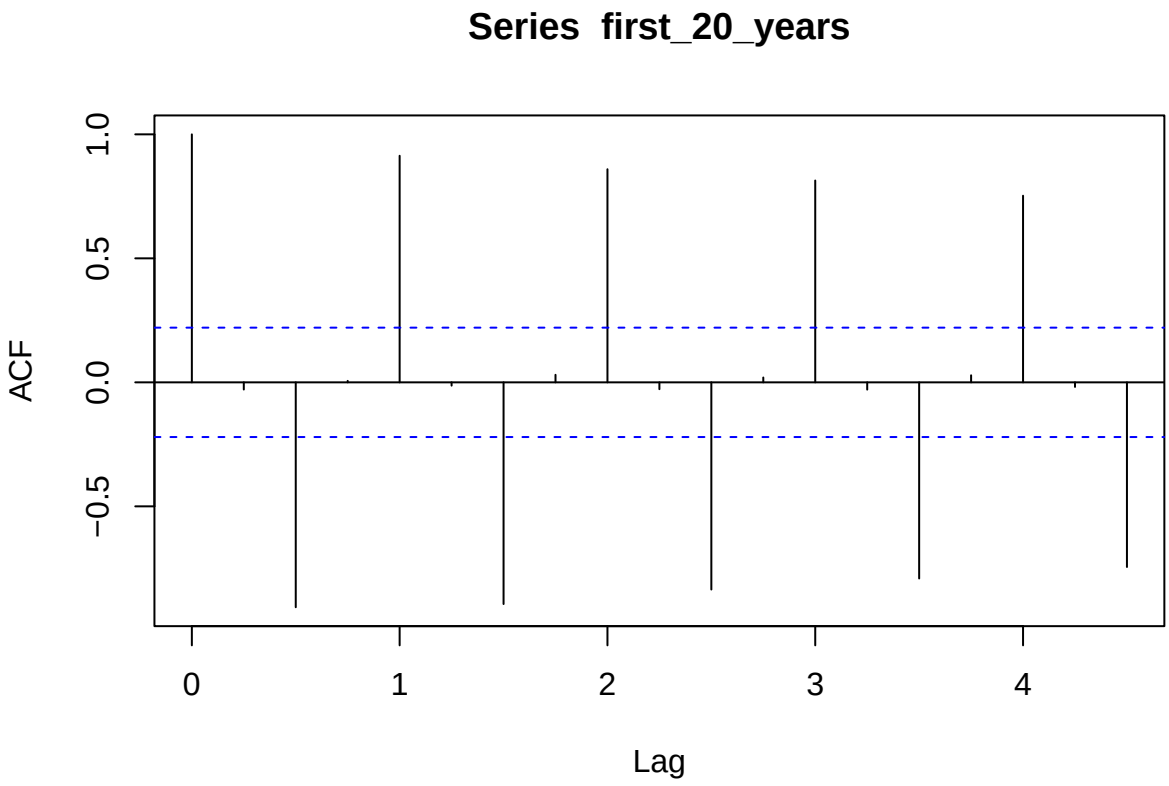


The ACFs from the first and second halves of the series are quite different. The ACF for the first half has significant spikes oscillating between positive and negative every second and fourth lag with large magnitudes. This means that the number of visitor arrivals is highly correlated with values from small and large lag values. This is because there was not much change in the number of visitor arrivals during the first half of the series (about 1920 to 1970) as travel was inaccessible to most people.

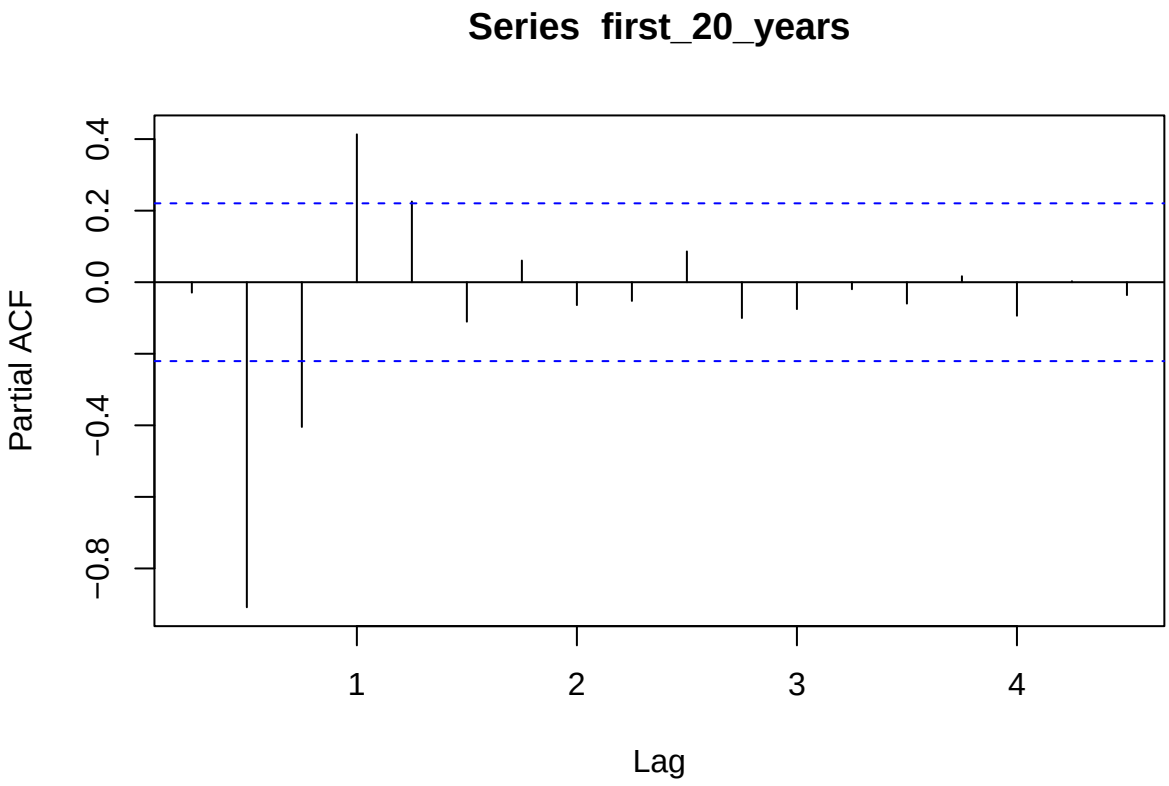
The ACF from the second half of the series also has significant spikes oscillating from positive to negative on the second an fourth lags but after lag 3 the magnitudes are much smaller than the magnitudes of the first half series. This means that the lag values have less of an impact of the number of arrivals when compared to the first half of the series. This is because there was a greater variation and trend of visitor arrivals in the second half of the series.

The PACFs are similar between the first and second halves of the series.

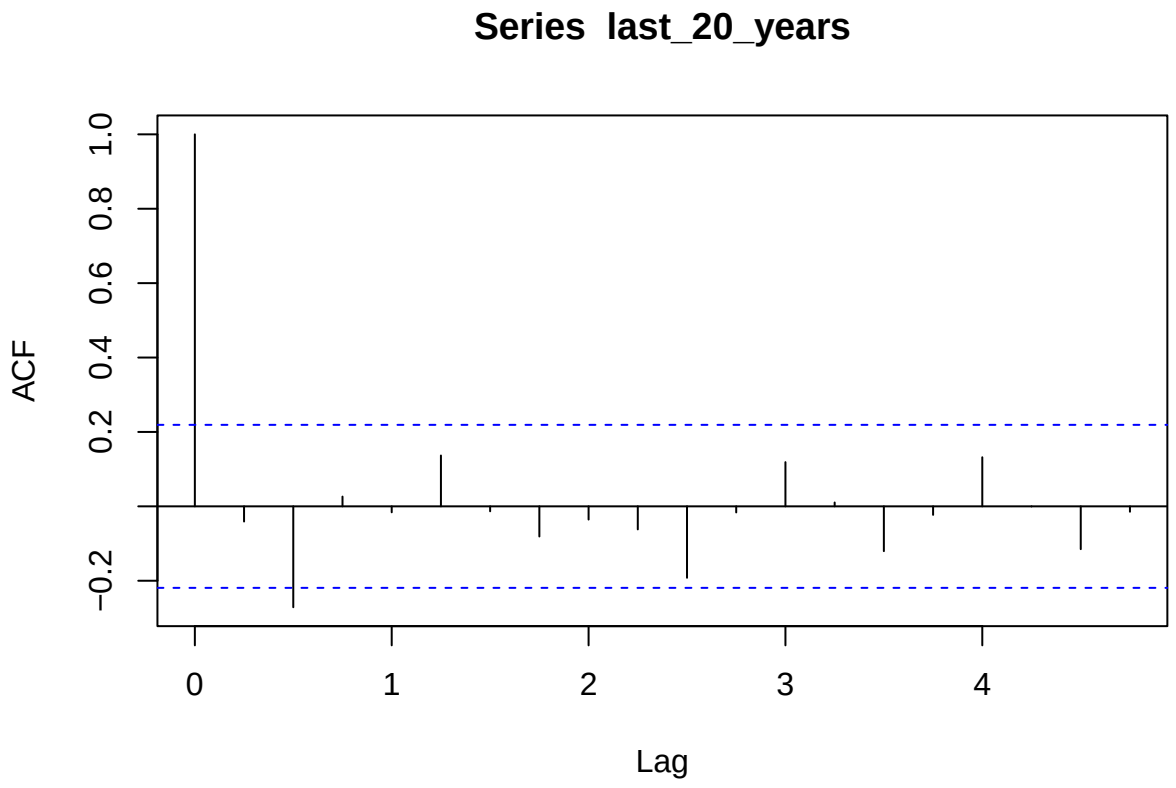
```
#Q4 b

first_20_years <- window(logdvis, start = start(logdvis), end = c(start(logdvis)[1] + 19, 4))
last_20_years <- window(logdvis, start = c(end(logdvis)[1] - 19, 1), end = end(logdvis))
acf(first_20_years)
```



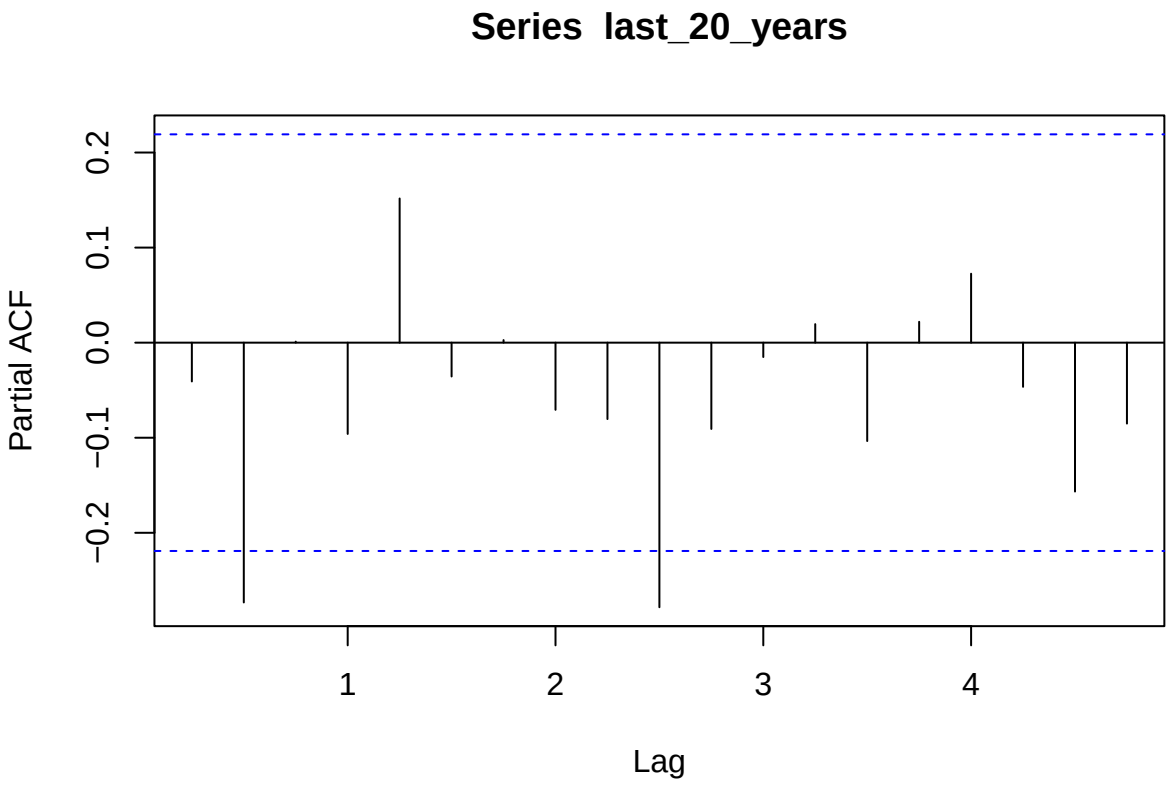

```
pacf(first_20_years)
```




```
acf(last_20_years)
```




```
pacf(last_20_years)
```



The ACFs from are quite different. The ACF from the first 20 years of the series has positive spikes oscillating from positive to negative on the second and fourth lags with a slight decrease in magnitude. This means values from the 18th spike have a high correlation with the current values. This is due to the small variance and horizontal trend in the first 20 years of the series.

The AFC from the last 20 years oscillates between positive and negative but none of the spikes are significant after the third spike. There is a still a slight peak every second and fourth lag indicating a seasonal trend. The non significant spikes mean values from lag 2 onwards do not have a significant impact the current number of arrivals. This is due to a large amount of variance in the number of arrivals in the last 20 years because of better techonology and covid19.

The pacf from the first 20 years shows a clear seasonal trend where every second and fourth spike has a peak. This is not shown in the pacf from the last 20 years.

The series second half of the series from 4a is the most informative as it shows a larger amount of time so we have more of an idea how the number of visitor arrivals has changed over the years and it is not as heavily impacted by covid 19 as the last 20 years series.

If we were asked to forecast the next two values we should use the logged differenced series. This is because our original series was not stationary as it has an increasing trend and the variance was not constant. taking the log and the difference of the series eliminated the trend and create constant variance. Once we have made our forecasts we can undo the transformations to get our forecasted number of visitor arrivals into New Zealand.