Forecasting Assignment

Phillip Frederick

September 18, 2020

```
library(fpp2)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
## v ggplot2   3.3.2     v fma       2.4
## v forecast  8.12      v expsmooth 2.3
## Warning: package 'forecast' was built under R version 3.5.3
## Warning: package 'fma' was built under R version 3.5.3
## Warning: package 'expsmooth' was built under R version 3.5.3
##
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.5.3
```

```
library(randtests)
```

```
## Warning: package 'randtests' was built under R version 3.5.2
```

```
library('readxl')
```

```
## Warning: package 'readxl' was built under R version 3.5.3
```

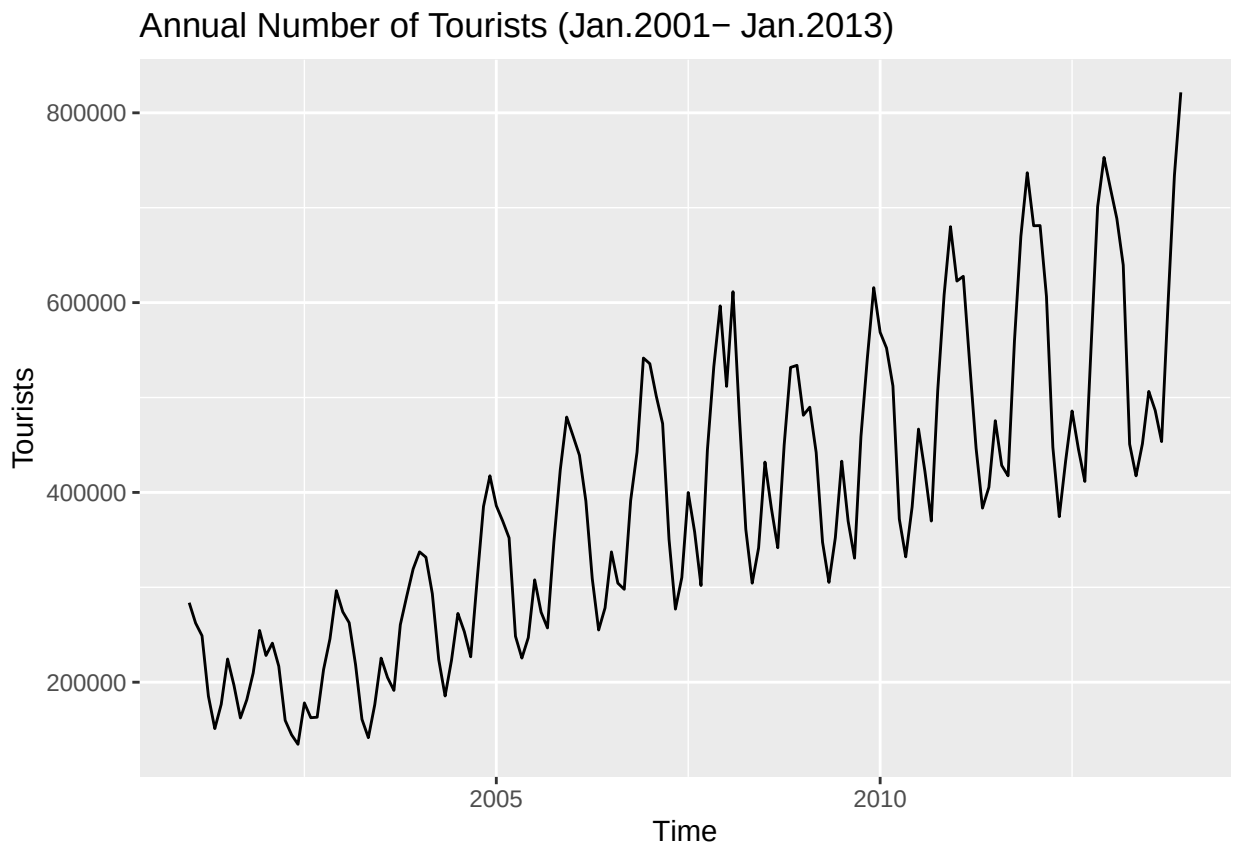
```
df<-read.table('clipboard',sep='\t',header=T,check.names=F)

tidy_df<-df %>% gather(Year,Tourists,"2001":"2013")
head(tidy_df)
```

```
##      Month Year Tourists
## 1  January 2001   283750
## 2 February 2001   262306
## 3    March 2001   248965
## 4    April 2001   185338
## 5      May 2001   151098
## 6     June 2001   176716
```

Reading in and tidying the data.

Question 1

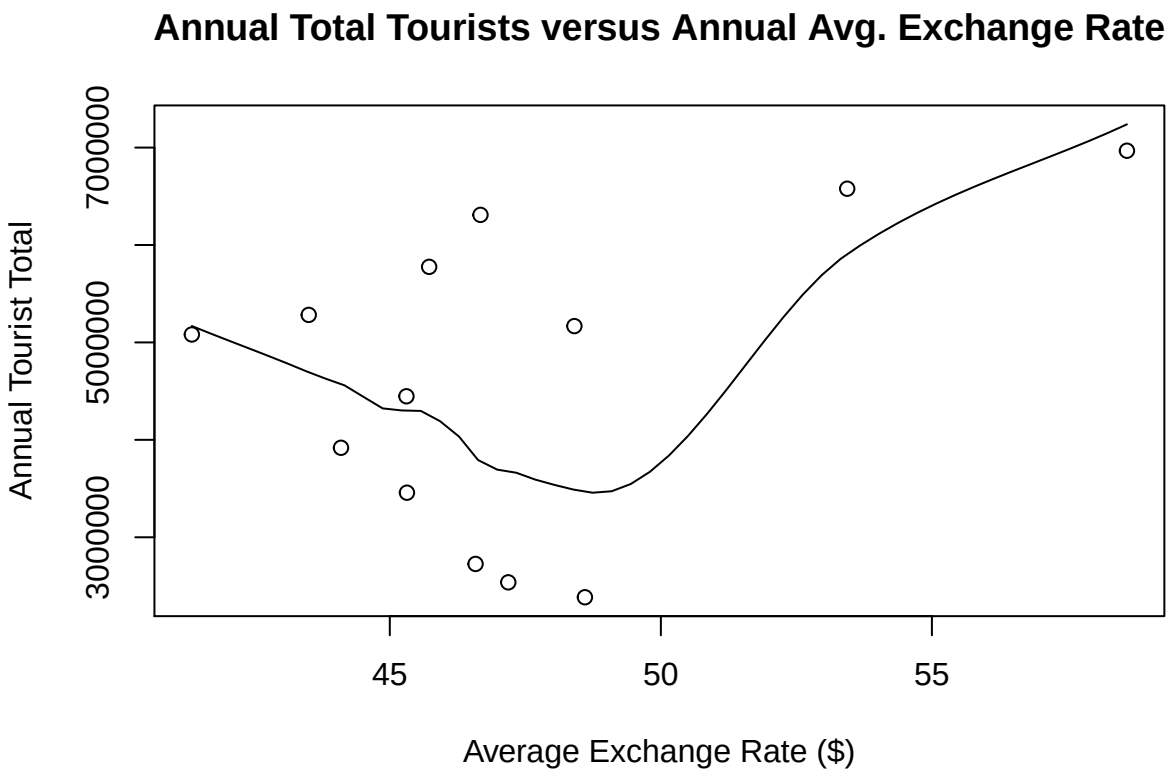
```
options(scipen=5)
autoplot(ts(tidy_df[,'Tourists'],frequency=12,start=c(2001,1))) +ylab('Tourists')+
  ggtitle("Annual Number of Tourists (Jan.2001- Jan.2013)")
```



The series exhibits a notable trend as well as seasonality. There is an obvious increase in variance with a non-constant mean.

Question 2

```
df2<-suppressMessages(read_excel('Indian_Holiday_File_Work.xlsx',sheet=2,skip=1))
colnames(df2)<-c('Year','Ann_Avg_Rate','Ann_Tot_Tourists')

scatter.smooth(x=df2$`Ann_Avg_Rate`,y=df2$Ann_Tot_Tourists,
               main="Annual Total Tourists versus Annual Avg. Exchange Rate",
               ylab="Annual Tourist Total",
               xlab="Average Exchange Rate ($)")
```




```
linear_model<-lm(Ann_Tot_Tourists~Ann_Avg_Rate ,data=df2)
summary(linear_model)
```

```
##
## Call:
## lm(formula = Ann_Tot_Tourists ~ Ann_Avg_Rate, data = df2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2468689  -921495   342720  1165250  1734454
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2165349    4555921  -0.475    0.644
## Ann_Avg_Rate   144414      95944   1.505    0.160
##
## Residual standard error: 1485000 on 11 degrees of freedom
## Multiple R-squared:  0.1708, Adjusted R-squared:  0.09541
## F-statistic: 2.266 on 1 and 11 DF,  p-value: 0.1604
```

The scatterplot highlights that there is no real correlation between annual average exchange rate and annual tourist totals.

This is further supported by conducting a simple linear regression and observing the R^2 term which shows that the average annual exchange rate explains only 17.1% of the variability in the annual tourists total. There is no real correlation.

Question 3

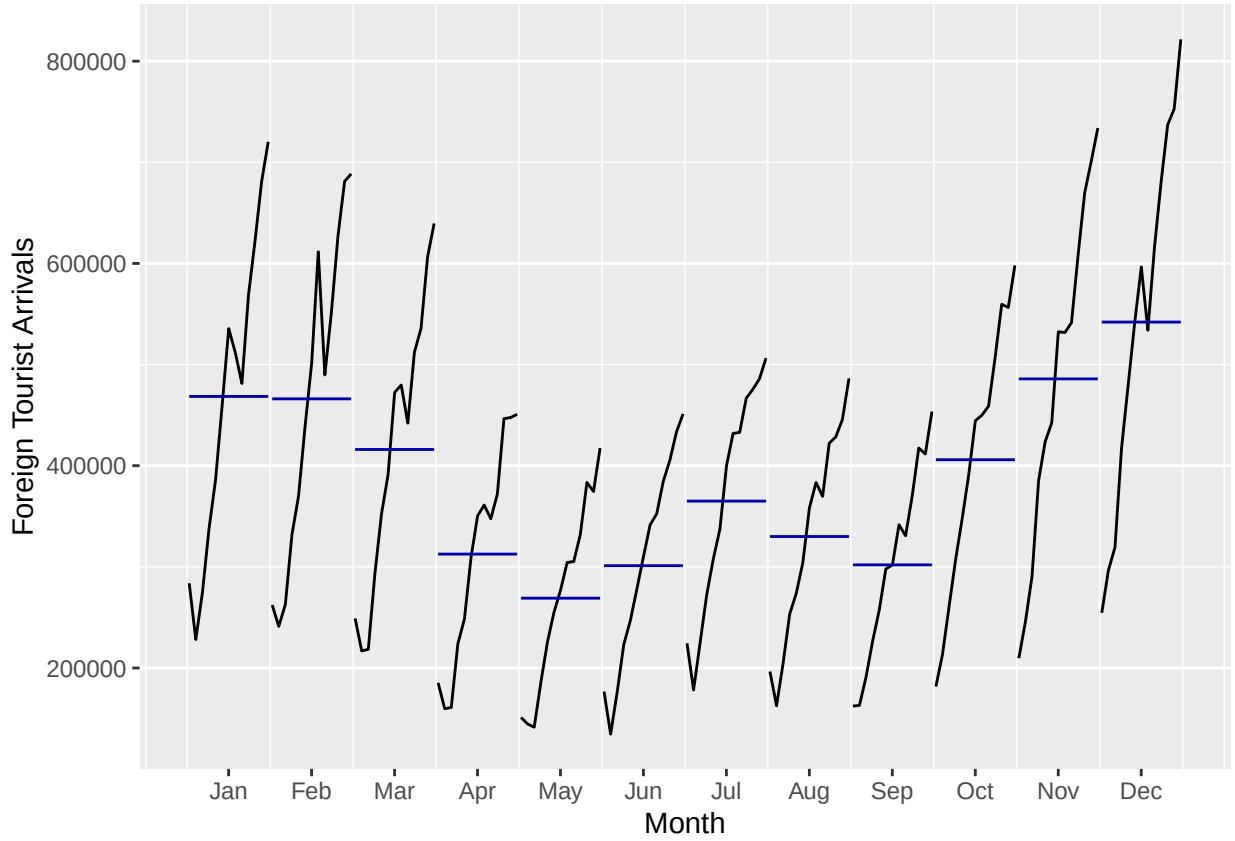
```
runs.test(tidy_df[,'Tourists'])
```

```
##
##  Runs Test
##
## data:  tidy_df[, "Tourists"]
## statistic = -9.1569, runs = 22, n1 = 78, n2 = 78, n = 156, p-value
## < 2.2e-16
## alternative hypothesis: nonrandomness
```

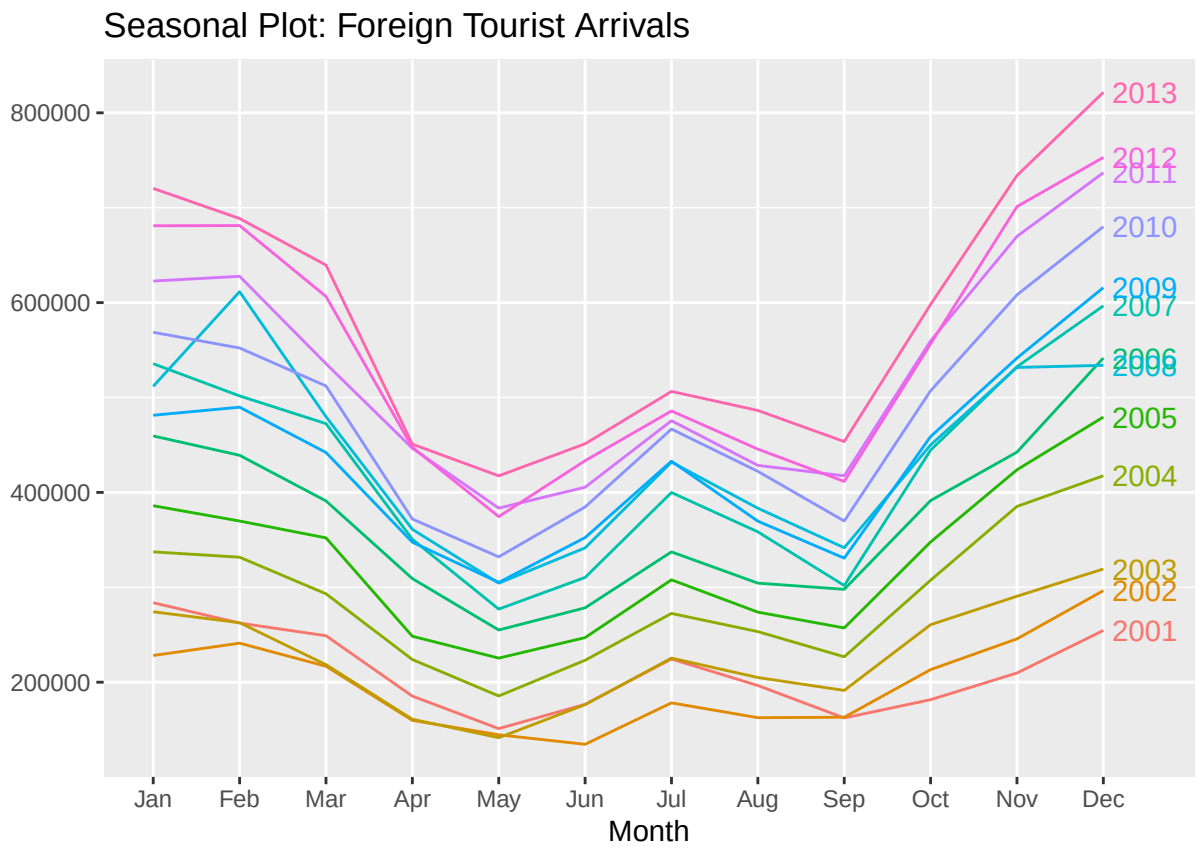
The p-value shows that we should reject the null hypotheisis that the data is random therefore supporting the notion that the data is serially correlated or trended.

Question 4

```
ts_obj<-ts(tidy_df[,'Tourists'],frequency=12,start=c(2001,1))

ggsubseriesplot(ts_obj)+
  ylab("Foreign Tourist Arrivals");
```




```
ggseasonplot(ts_obj,year.labels=TRUE)+
  ggtitle('Seasonal Plot: Foreign Tourist Arrivals')
```



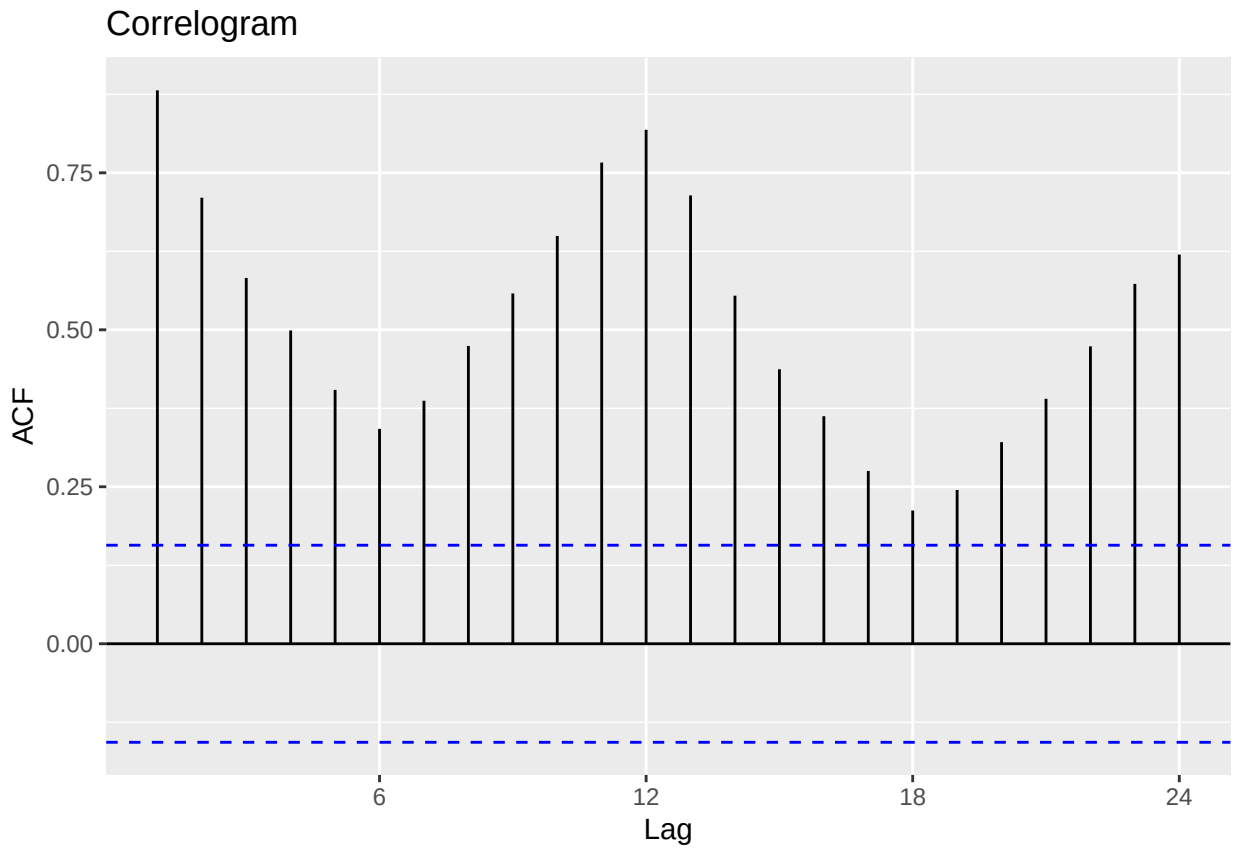
The arrival of foreign tourists is infact seasonal. The data confirms information from the case outline.

The warmer months of May thru August showcase lower foreign tourist arrival numbers on average than the cooler months of September thru February with December highlighting the most arrivals.

An interesting exception is July which shows a slight uptick in arrivals compared to other warm months.I believe this could be attributed to July being the coolest of the warmer months with an average high of 30.9 degrees celcius.

Question 5

```
ggAcf(ts(tidy_df[,'Tourists'],frequency=12,start=c(2001,1)))+ ggtitle('Correlogram')
```



The lagged components of the data show that as time progresses less recent values are less influential.

Spikes in the data reference seasonality and the downward, positive progression of correlated lags implies trend.

The data is non-stationary and does infact need to be differenced as well as log or square root transformed in order to account for the increasing variance observed.

The possible implications of differencing are namely the interpretation of the final results especially in addition to transforming the data.

Question 6

```
library(urca)
```

```
## Warning: package 'urca' was built under R version 3.5.3
```

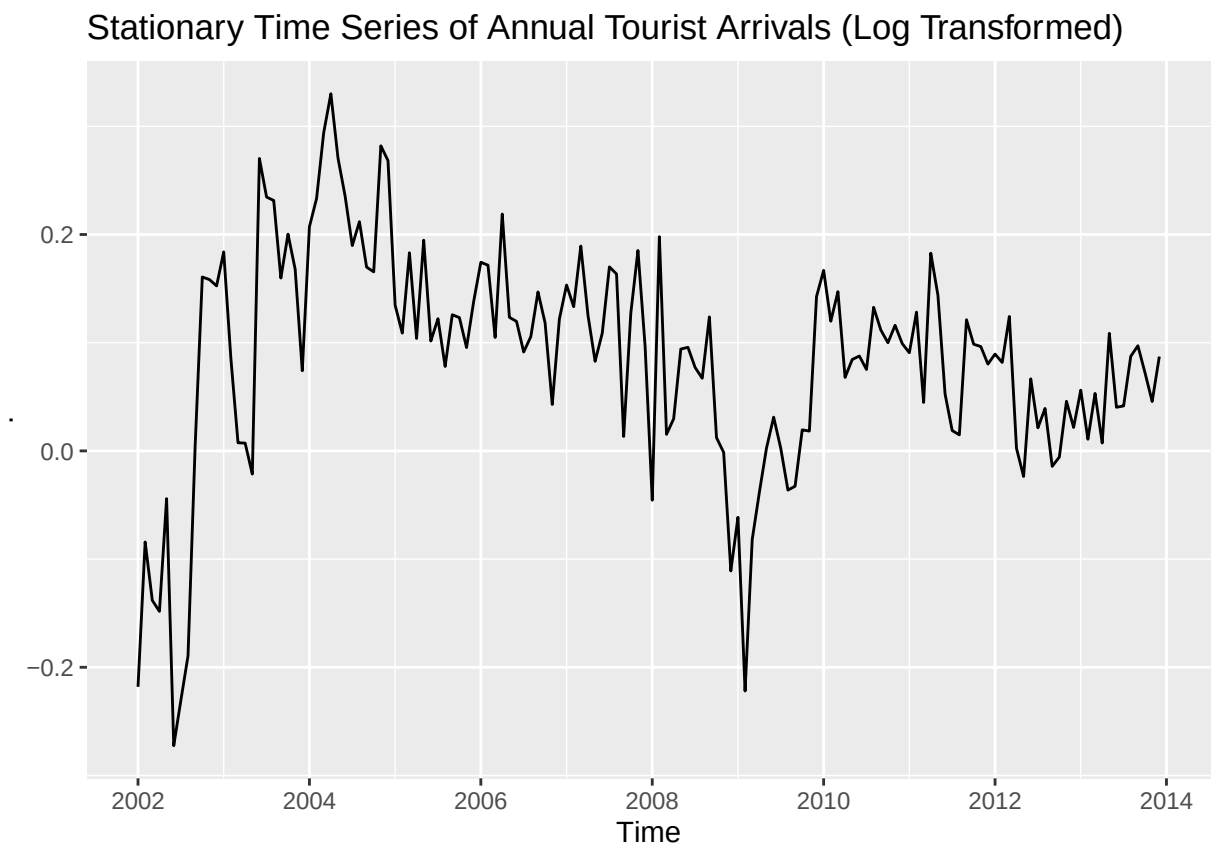
```
log_diff<-ts(tidy_df[,"Tourists"],frequency=12,start=c(2001,1)) %>% log() %>% diff(lag=12)
summary(ur.kpss(log_diff))
```

```
##
## ########################
## # KPSS Unit Root Test #
## ########################
```

```
##
## Test is of type: mu with 4 lags.
##
## Value of test-statistic is: 0.2966
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.347 0.463  0.574 0.739
```

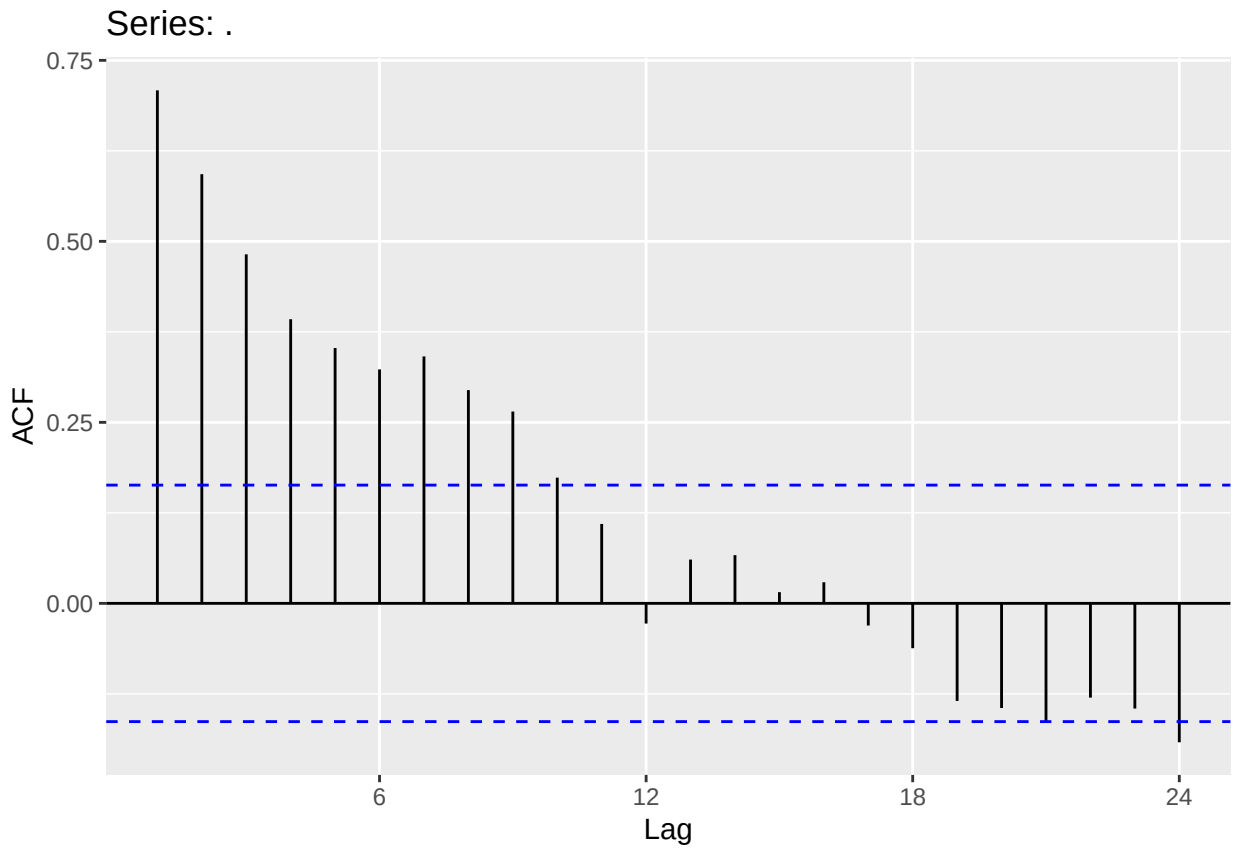
Log transformation of the data with a difference at lag 12 followed by confirmation of stationarity/non-seasonality using a KPSS test.

```
ts_obj %>% log() %>% diff(lag=12)%>%autoplot()+
  ggtitle('Stationary Time Series of Annual Tourist Arrivals (Log Transformed)')
```

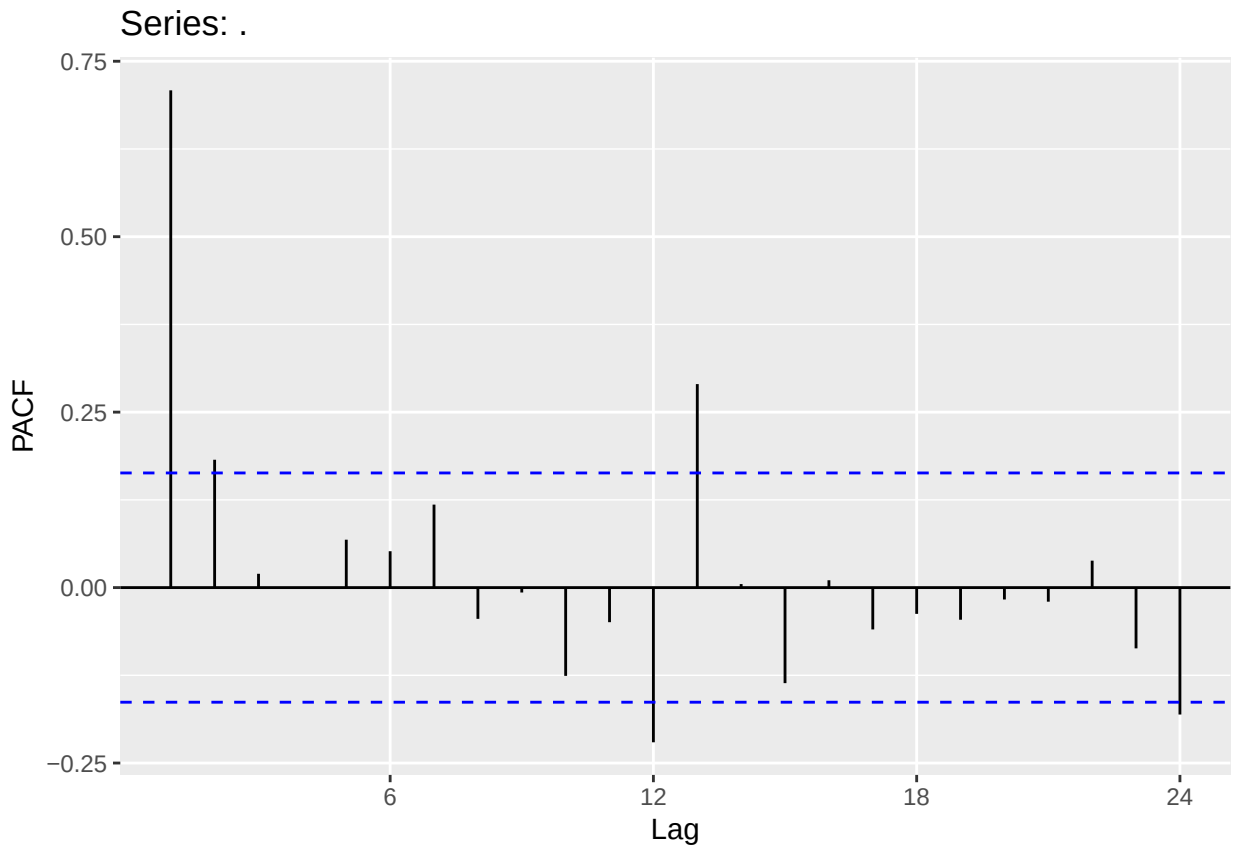


Visualization of stationary time series data.

```
log_diff%>%ggAcf()
```



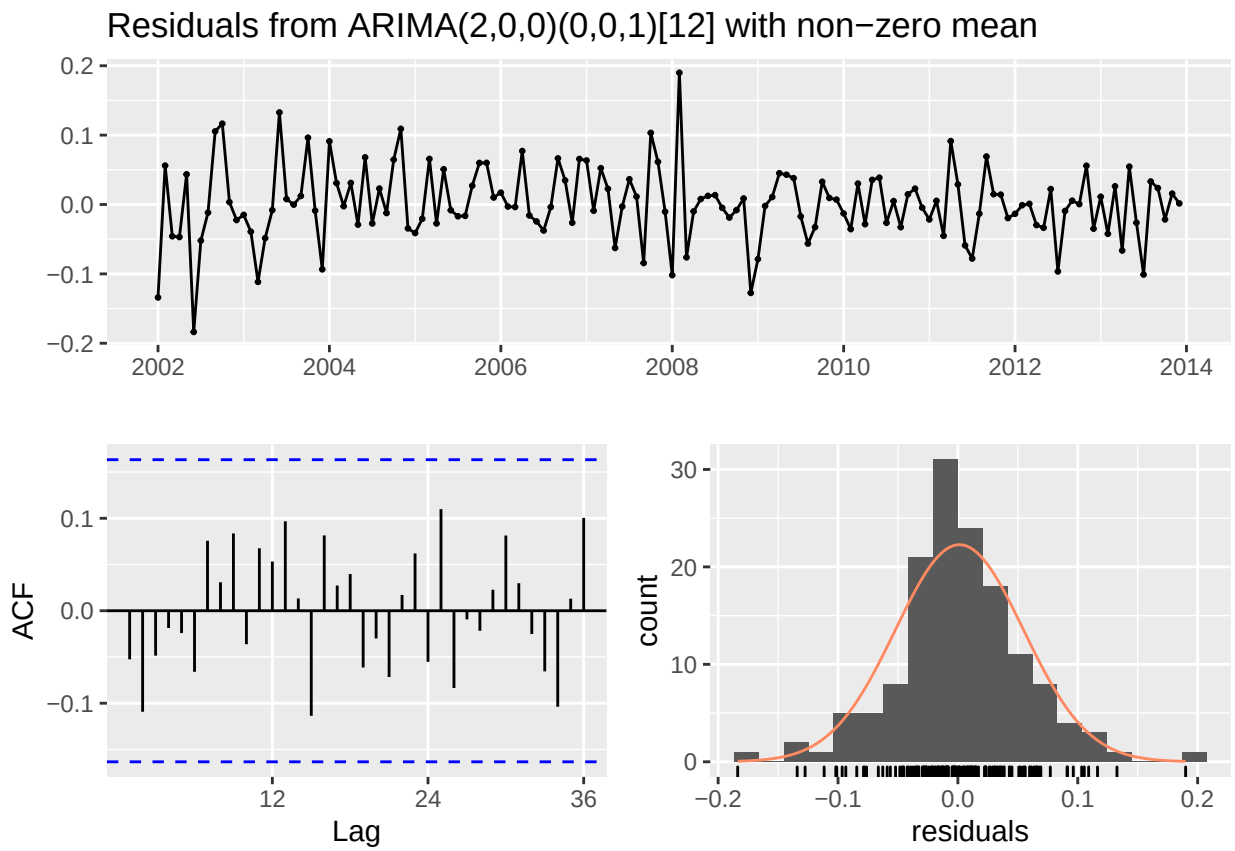

```
log_diff%>%ggPacf()
```



True significant lags are shown in lags 1,12 and 13 as shown in the partial autocorrelation function plot above.

This is suggestive of a possible AR(3) model.

```
fit<-auto.arima(log_diff)
checkresiduals(fit)
```

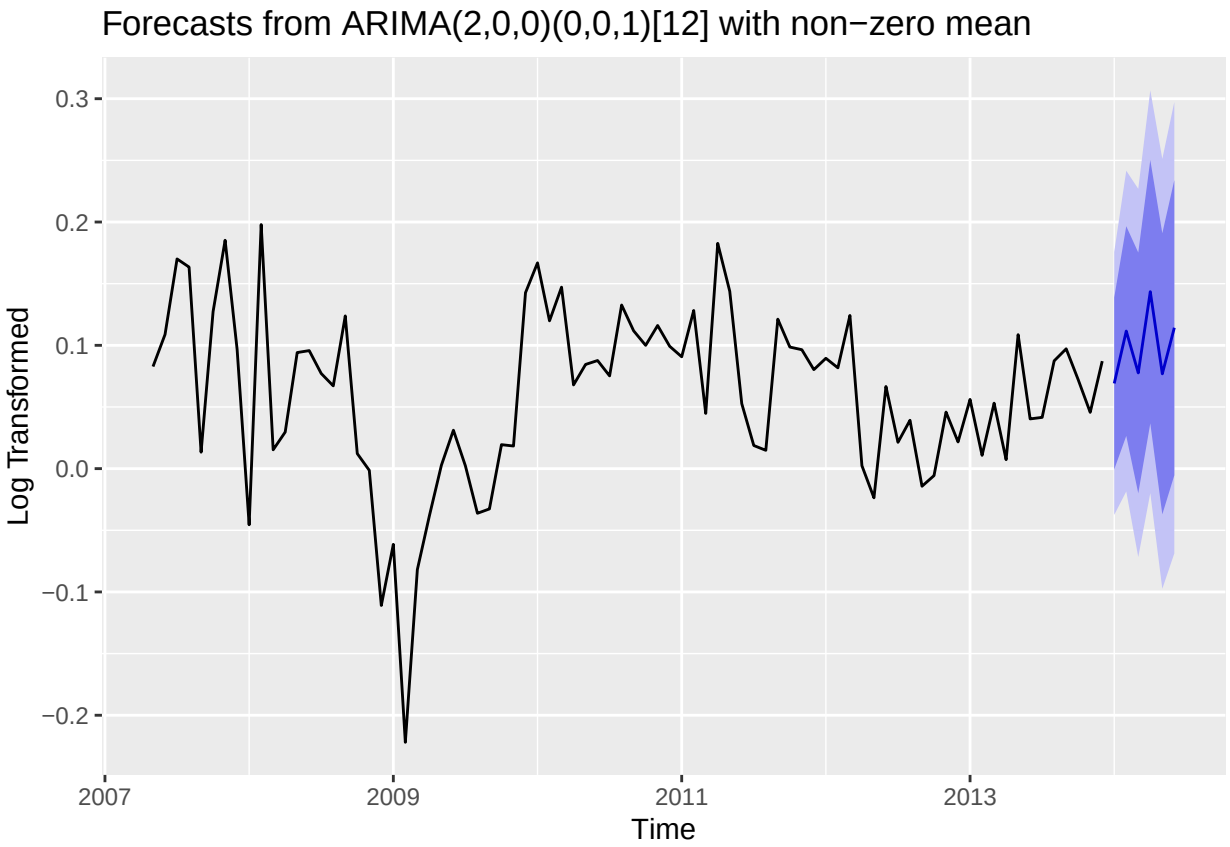



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,0,0)(0,0,1)[12] with non-zero mean
## Q* = 14.844, df = 20, p-value = 0.7852
##
## Model df: 4.   Total lags used: 24
```

The auto ARIMA function found that a model accounting for two lagged autoregressive terms and a MA(1) term associated with the seasonal component is the optimal model choice.

Question 7

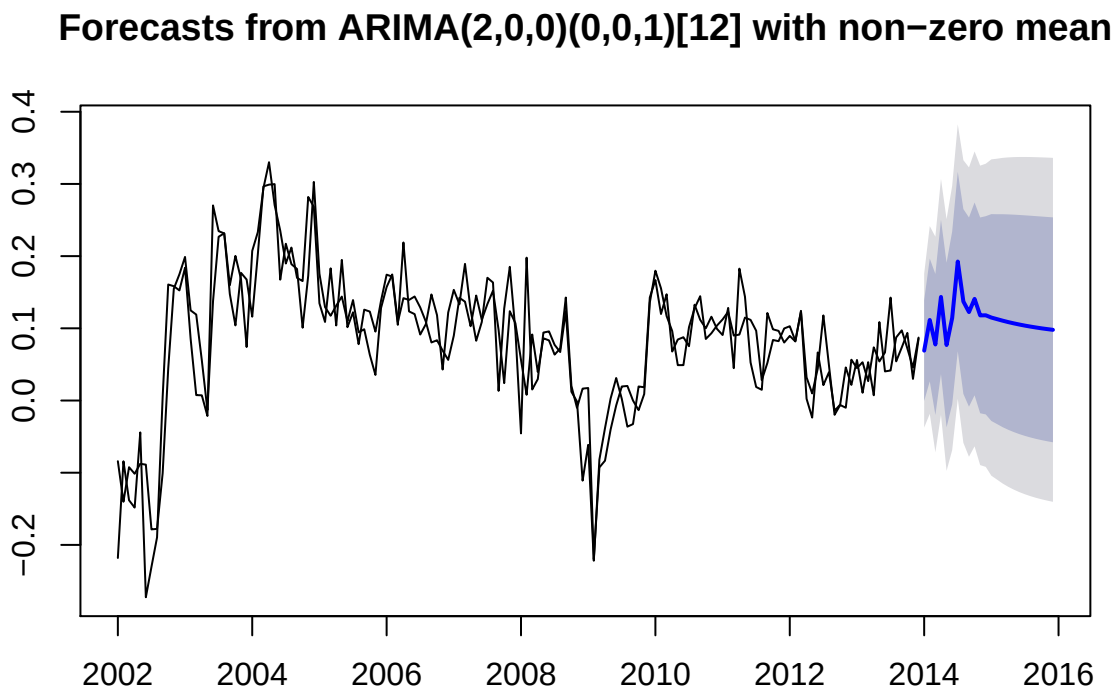
```
fit %>% forecast(h=6) %>% autoplot(include=80) + ylab('Log Transformed')
```



The forecast demonstrates an uptick in tourist arrivals midway thru the next six months.

Question 8

```
par(mfrow=c(1,1))
plot(forecast(fit))
lines(fitted(fit))
```



Judging from the superimposed visualization of the fitted and the actual values of the model it appears that our forecast does a decent job of extrapolating the underlying movement and patterns of our data.

Rao should expect to see an overall uptick of foreign tourist arrivals within the next six months. He is best suited to utilize an ARIMA(2,0,0)(0,0,1)[12]

model. It should be noted that while the model is accurate, based upon the fitted values chart, the ARIMA model does assume that historical patterns will not change during the forecast period in which case it should not be used as a guaranteed measure.

Overall the larger numbers of arrivals can be expected to transpire from December thru February and it should be noted that there is no evidence which suggest a strong correlation between

the annual average exchange rate and the total tourist arrival numbers.