M4 Forecasting Case Assignment

Michael Pelletier

```
library(fpp2)
```

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

```
## -- Attaching packages --------------------------------------------- fpp2
2.4 --
```

```
## v ggplot2   3.3.3     v fma       2.4
## v forecast  8.14      v expsmooth 2.3
```

```
##
```

```
library(forecast)
library(readxl)
library(urca)
library(ggplot2)
```

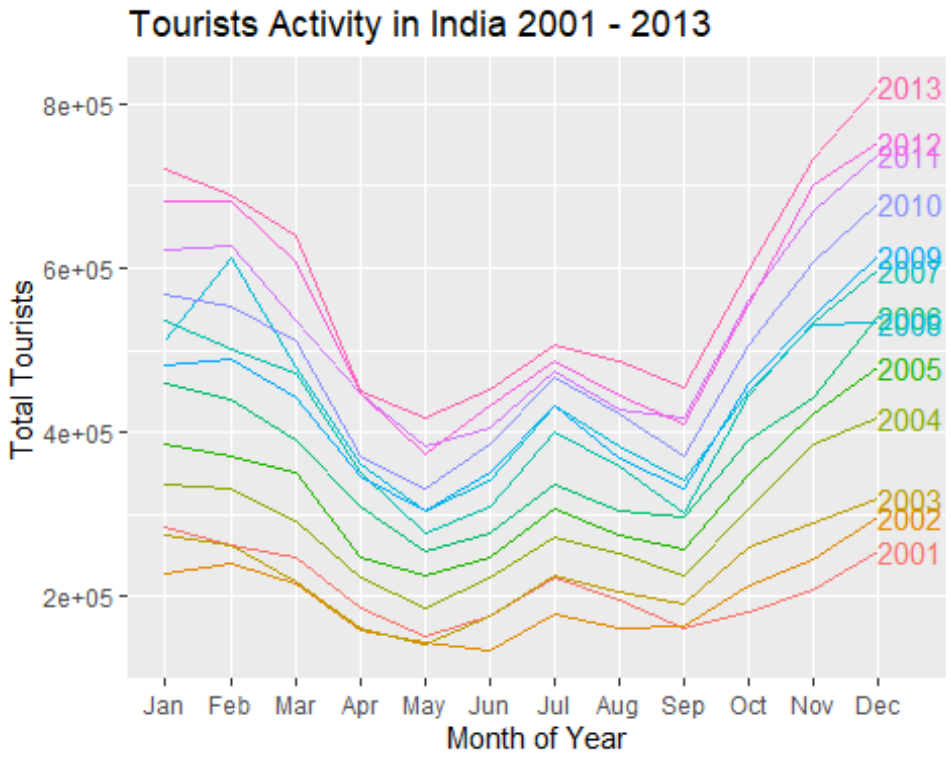
```
TouristVisit <- read.table(file = "clipboard", header = TRUE, sep = "\t")
ts.Touristvists <- ts(as.numeric(unlist(TouristVisit)), frequency = 12, start
= c(2001,1))
ts.Touristvists
```

```
##         Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct
## 2001 283750 262306 248965 185338 151098 176716 224432 196517 162326 181605
## 2002 228150 241133 216839 159789 144571 134566 178231 162594 163089 213267
## 2003 274215 262692 218473 160941 141508 176324 225359 204940 191339 260569
## 2004 337345 331697 293185 223884 185502 223122 272456 253301 226773 307447
## 2005 385977 369844 352094 248416 225394 246970 307870 273856 257184 347757
## 2006 459489 439090 391009 309208 255008 278370 337332 304387 297891 391399
## 2007 535631 501692 472494 350550 277017 310364 399866 358446 301892 444564
## 2008 511781 611493 479765 361101 304361 341539 431933 383337 341693 450013
## 2009 481308 489787 442062 347544 305183 352353 432900 369707 330707 458849
## 2010 568719 552152 512152 371956 332087 384642 466715 422173 369821 507093
## 2011 622713 627719 535613 446511 383439 405464 475544 428490 417478 559641
## 2012 681002 681193 606456 447581 374476 433390 485808 445632 411562 556488
## 2013 720321 688569 639530 450580 417453 451223 506427 486338 453561 598095
##         Nov    Dec
## 2001 209685 254544
## 2002 245661 296474
## 2003 290583 319271
## 2004 385238 417527
## 2005 423837 479411
## 2006 442413 541571
```

```
## 2007 532428 596560
## 2008 531683 533904
## 2009 541524 615775
## 2010 608178 680004
## 2011 669767 736843
## 2012 701185 752972
## 2013 733923 821581
```

Plot the series indicating the number of foreign tourists arriving in India each month.

```
ggseasonplot(ts.Touristvists, year.labels = TRUE)+xlab("Month of
Year")+ylab("Total Tourists")+ggtitle("Tourists Activity in India 2001 -
2013")
```



Looking at the plot, you can see that there is cyclicality and seasonality regarding month to month tourist activity. It seems that the tourist presence in India peaks in September then decreases after January most years. There is also a small jump in July.Lastly, you can see a pretty consistent increase in tourists each year.

Is the annual total of foreign tourists arriving in India correlated with the annual average exchange rate?

```
Exchange <-
read_excel("C:\\Users\\18046\\Documents\\DecModeling\\ExchangeRates.xlsx")
regression <- lm(Annual ~ Rate, data = Exchange)
summary(regression)

##
## Call:
```

```
## lm(formula = Annual ~ Rate, data = Exchange)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -205721  -76792   28563   97102  144539
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -180397     379657  -0.475    0.644
## Rate           12033       7995   1.505    0.160
##
## Residual standard error: 123800 on 11 degrees of freedom
## Multiple R-squared:  0.1708, Adjusted R-squared:  0.09538
## F-statistic: 2.265 on 1 and 11 DF,  p-value: 0.1605
```

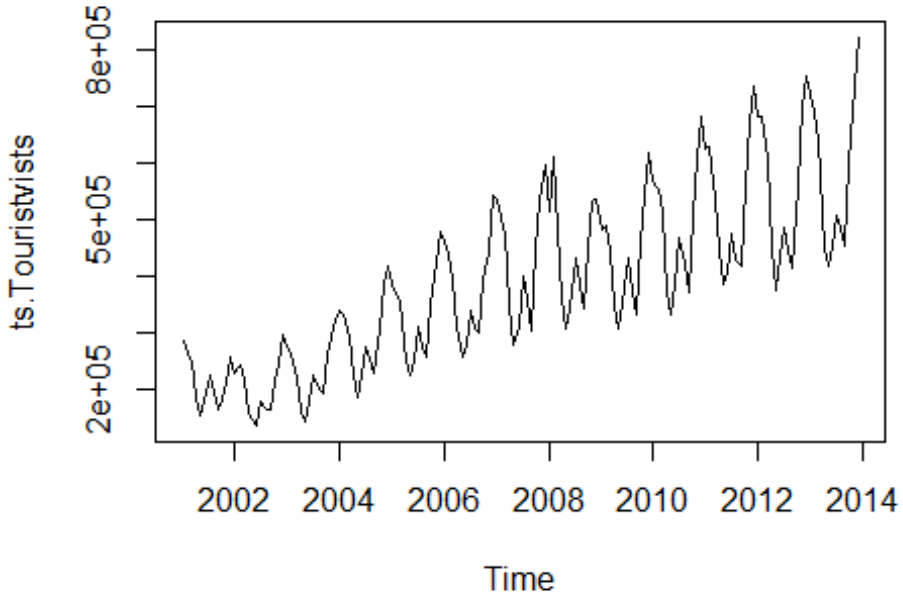
```
cor.test(Exchange$Annual,Exchange$Rate)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Exchange$Annual and Exchange$Rate
## t = 1.5051, df = 11, p-value = 0.1605
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1783577  0.7853974
## sample estimates:
##       cor
## 0.4132374
```

*I ran a regression model on the annual exchange rate and the annual tourist data from 2001-2013 and the model turned out insignificant. To further investigate, I conducted a Pearson's correlation test and it returned a correlation of 0.4132, which is not very indicative. Therefore, I can not confidently say there is very much correlation between tourists visiting India and the Exchange rate between India and US monetary systems.

Is there an inherent trend in the series?

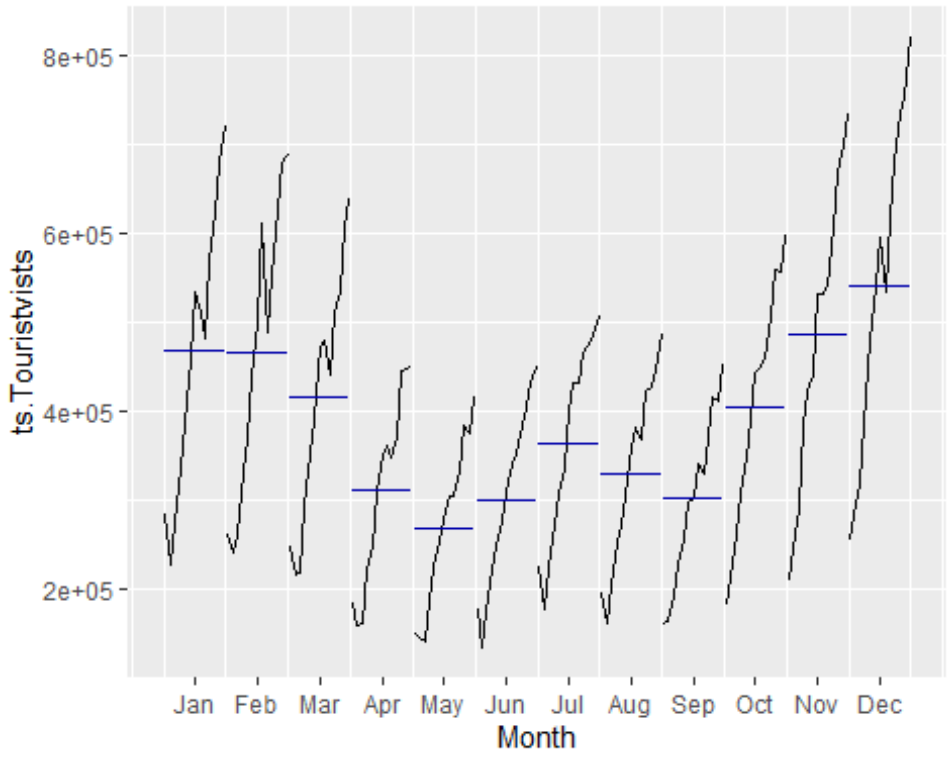
```
plot(ts.Touristvists)
```



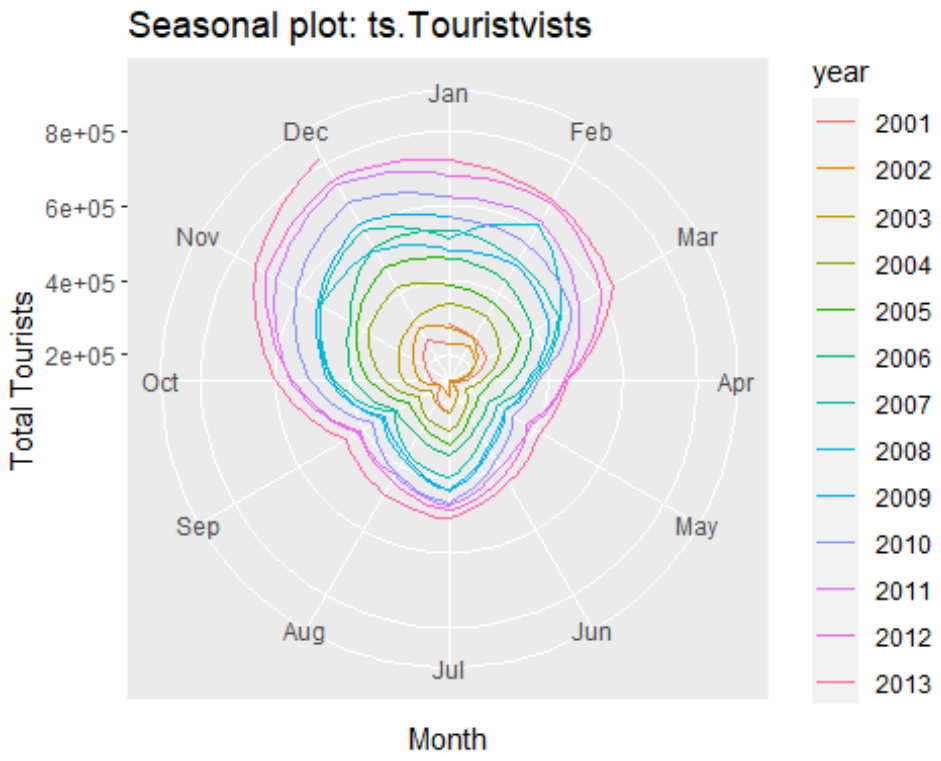
Looking at the Random Walk Model above, it is clear there is a positive drift within the data. Also, year by year the mean increases

Does the graph indicate any seasonal behavior?

```
ggsubseriesplot(ts.Touristvists)
```



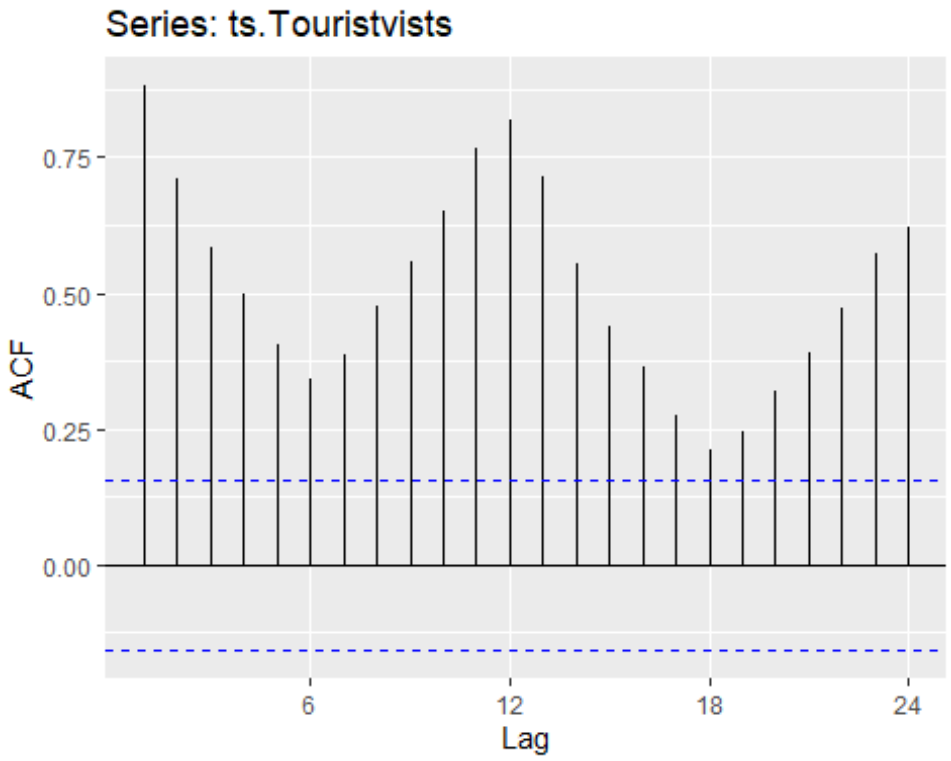

```
ggseasonplot(ts.Touristvists, polar = TRUE) + ylab("Total Tourists")
```



Above are two plots that show consistent seasonality within the data.Similar to our initial visualization of the data, both plots show peaks of tourism from September to February and a decrease in the summer, apart from July where is a small jump in tourism.

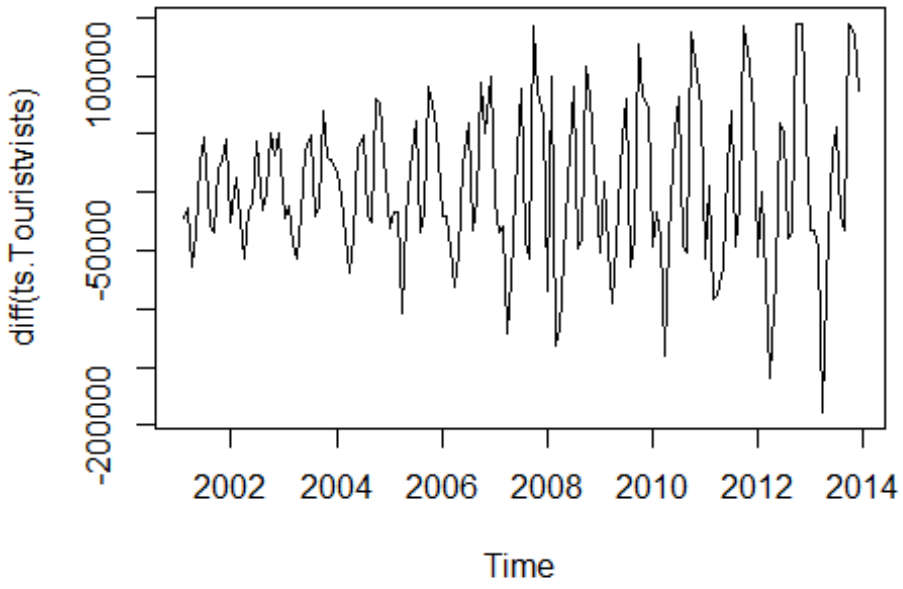
Is the series stationary? Does it need to be differenced? If yes, what are the implications?

```
ggAcf(ts.Touristvists)
```

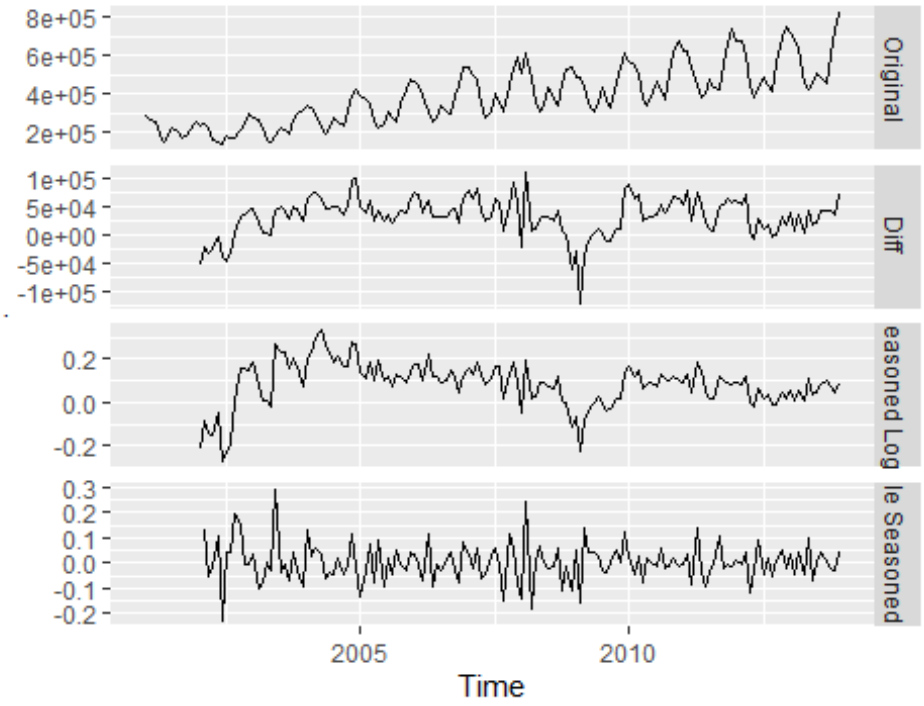



```
summary(ur.kpss(ts.Touristvists))

##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: mu with 4 lags.
##
## Value of test-statistic is: 2.5901
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.347 0.463  0.574 0.739

summary(ur.kpss(diff(ts.Touristvists)))

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

```
## ########################
##
## Test is of type: mu with 4 lags.
##
## Value of test-statistic is: 0.0433
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.347 0.463  0.574 0.739

plot(diff(ts.Touristvists))
```



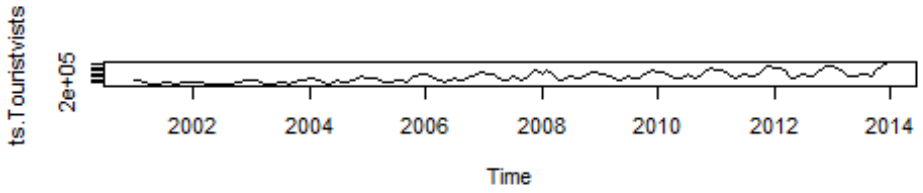

```
cbind("Original" = ts.Touristvists,
"Diff " = diff(ts.Touristvists,12),
"Seasoned Logs"= diff(log(ts.Touristvists),12),
"Double Seasoned Logs" = diff(diff(log(ts.Touristvists),12),1)) %>%
  autoplot(facets=TRUE)
```

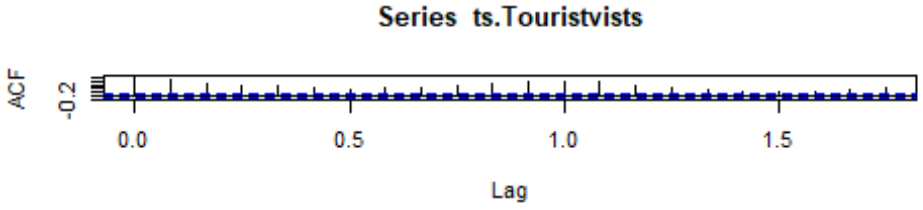


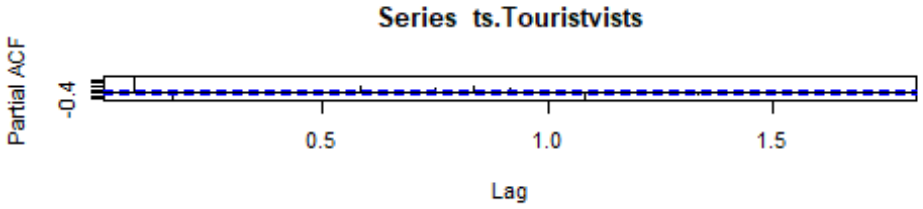
Other than the seasonality and trend observed above, there are a few more indicators that tell us this data is not stationary.First, our ACF gradually decreases and even increases at lag 12 which further proves seasonal behavior. Furthermore, I ran some KPSS tests and it returned a value of test-statistic at 2.5901, which is way above the critical value of significance, hence it did not pass the stationary test. I then differentiated the data and ran the same test which returned a test-statistic of 0.0433 which passes the stationary test.

Develop and ARIMA model to forecast the series.

```
par(mfrow = c(3,1))
plot(ts.Touristvists)
acf(ts.Touristvists)
pacf(ts.Touristvists)
```





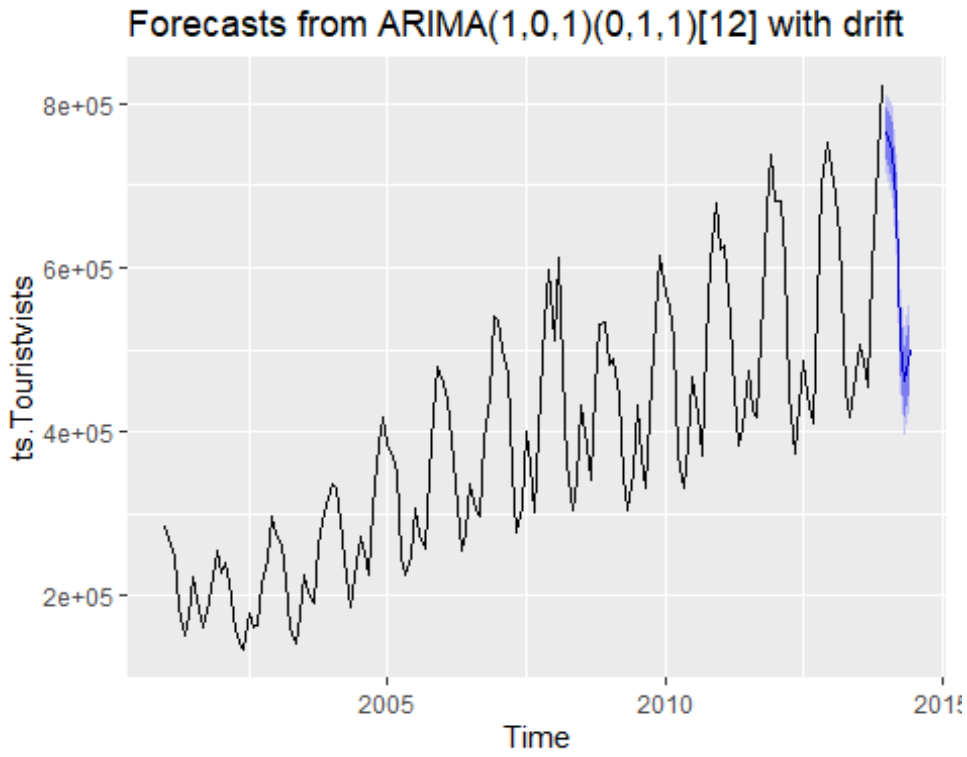



```
auto.arima(ts.Touristvists, seasonal = T)

## Series: ts.Touristvists
## ARIMA(1,0,1)(0,1,1)[12] with drift
##
## Coefficients:
##          ar1      ma1     sma1      drift
##       0.8055  -0.1971  -0.4528  2646.0503
## s.e.  0.0695   0.1106   0.0794   399.7329
##
## sigma^2 estimated as 605136671:  log likelihood=-1659.92
## AIC=3329.84   AICc=3330.28   BIC=3344.69
```

I found through trial and error with auto.arima I was using too many parameters. This is the model I found to have the best results.

Using the developed ARIMA model, forecast the expected tourist arrivals in India over the next six months.

```
fit <- auto.arima(ts.Touristvists, seasonal = T)
fit %>% forecast(h=6) %>% autoplot()
```

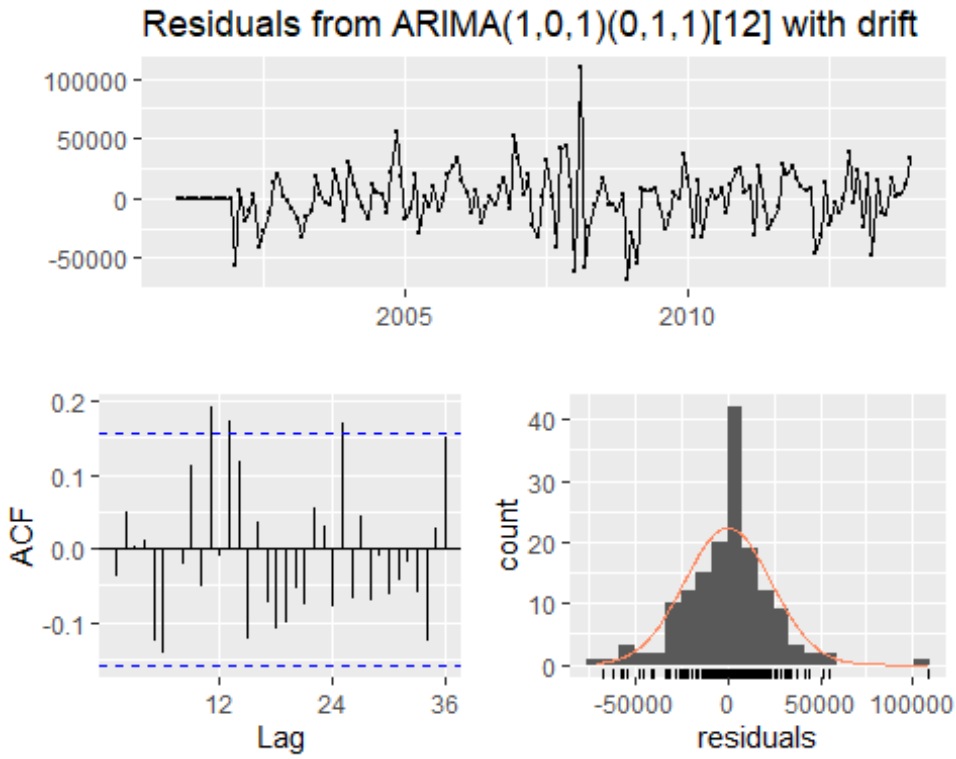



```
fit %>% forecast(h=6)

##          Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## Jan 2014       764760.8 733235.2 796286.3 716546.6 812974.9
## Feb 2014       743386.1 706484.2 780288.0 686949.5 799822.7
## Mar 2014       679063.1 639057.5 719068.7 617879.9 740246.4
## Apr 2014       511777.0 469880.5 553673.5 447701.8 575852.2
## May 2014       461862.8 418783.8 504941.8 395979.2 527746.5
## Jun 2014       499793.1 455963.9 543622.3 432762.1 566824.0
```

Forecasts for the next 6 months are listed above. Results look promising as at first glance, it has the amount of tourists decreasing heading into the summer months.

Evaluate the quality of your forecast and provide justification.

```
checkresiduals(fit)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,1)(0,1,1)[12] with drift
## Q* = 34.276, df = 20, p-value = 0.02432
##
## Model df: 4.   Total lags used: 24
```

Evaluating my results, the Ljung-Box test returns a p-value of 0.02432 which means are results are significant!