AAT

2024-12-30

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(tseries)
```

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

```
library(forecast)

data <- read.csv("C:\\Users\\sathw\\Downloads\\Thaitourism.csv")

cat("Dataset Overview:\n")
```

```
## Dataset Overview:
```

```
cat("Number of Rows:", nrow(data), "\n")
```

```
## Number of Rows: 4452
```

```
cat("Number of Columns:", ncol(data), "\n")
```

```
## Number of Columns: 5
```

```
cat("Column Names:\n")
```

```
## Column Names:
```

```r
print(names(data))
```

```
## [1] "region"      "nationality" "year"        "month"       "tourists"
```

```r
cat("\nSummary of the dataset:\n")
```

```
##
## Summary of the dataset:
```

```r
summary(data)
```

```
##     region          nationality             year          month
##  Length:4452        Length:4452        Min.   :2010   Min.   : 1.00
##  Class :character   Class :character   1st Qu.:2011   1st Qu.: 3.75
##  Mode  :character   Mode  :character   Median :2013   Median : 6.50
##                                        Mean   :2013   Mean   : 6.50
##                                        3rd Qu.:2015   3rd Qu.: 9.25
##                                        Max.   :2016   Max.   :12.00
##     tourists
##  Min.   :   104
##  1st Qu.:  5500
##  Median : 14216
##  Mean   : 38545
##  3rd Qu.: 49871
##  Max.   :958204
```

```r
# Assuming 'year' and 'month' are separate columns and the data is in "yyyy" and "mm" format
# Combine 'year' and 'month' into a date column, using the first day of the month
data$date <- as.Date(paste(data$year, data$month, "01", sep = "-"), format = "%Y-%m-%d")

cat("\nChecking for missing values:\n")
```

```
##
## Checking for missing values:
```

```r
print(sapply(data, function(x) sum(is.na(x))))
```

```
##      region nationality        year       month    tourists        date
##           0           0           0           0           0           0
```

```r
data$tourist[is.na(data$tourist)] <- 0  # Assuming the column 'tourist' needs to be filled

# Group by 'date' and summarize 'tourist' (total tourists per month)
time_series_data <- data %>%
  group_by(date) %>%
  summarise(tourist = sum(tourist))

# Create time series object for 'tourist' data, starting from the first available date
tourist_ts <- ts(time_series_data$tourist, start = c(year(min(time_series_data$date)), month(min(time_s
                frequency = 12)  # Monthly frequency

cat("\nDickey-Fuller Test:\n")
```

```
##
## Dickey-Fuller Test:
```

```
adf_test <- adf.test(tourist_ts)
```

```
## Warning in adf.test(tourist_ts): p-value smaller than printed p-value
```

```
print(adf_test)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  tourist_ts
## Dickey-Fuller = -4.1114, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

```
cat("\nKPSS Test:\n")
```

```
##
## KPSS Test:
```

```
kpss_test <- kpss.test(tourist_ts)
```

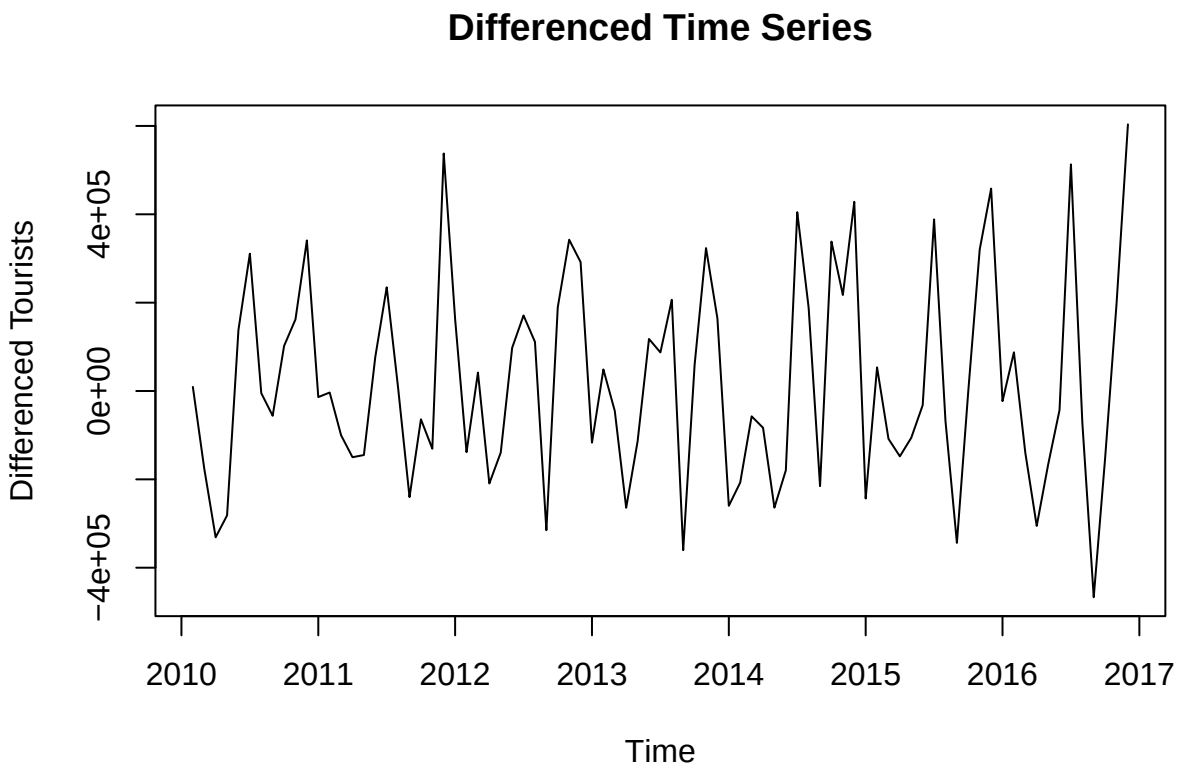
```
## Warning in kpss.test(tourist_ts): p-value smaller than printed p-value
```

```
print(kpss_test)
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  tourist_ts
## KPSS Level = 1.7931, Truncation lag parameter = 3, p-value = 0.01
```

```
# Differencing the series to make it stationary
differenced_tourist_ts <- diff(tourist_ts)

# Plot the differenced time series
plot(differenced_tourist_ts, main = "Differenced Time Series", ylab = "Differenced Tourists", xlab = "T
```



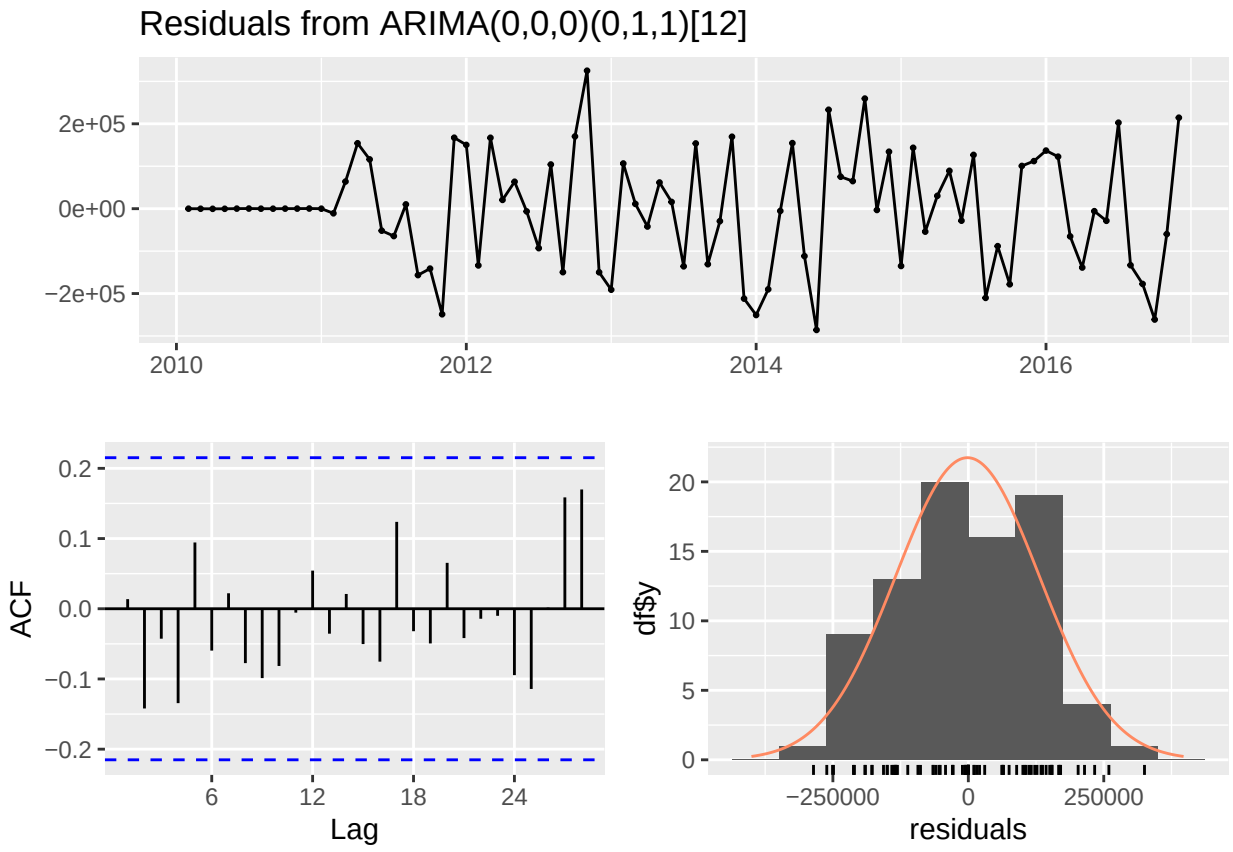

```
# Fit an ARIMA model to the differenced data
auto_model <- auto.arima(differenced_tourist_ts)
cat("\nSelected ARIMA Model:\n")
```

```
##
## Selected ARIMA Model:
```

```
print(auto_model)
```

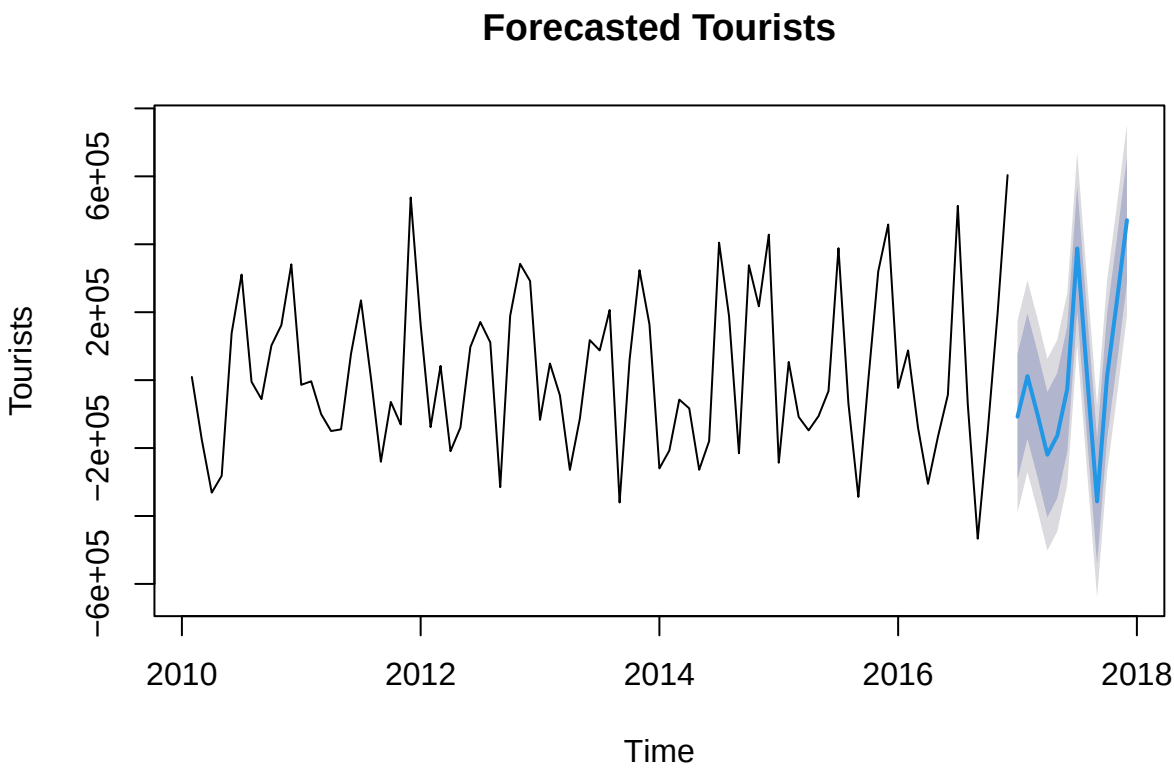
```
## Series: differenced_tourist_ts
## ARIMA(0,0,0)(0,1,1)[12]
##
## Coefficients:
##          sma1
##       -0.6193
## s.e.   0.1905
##
## sigma^2 = 2.07e+10:  log likelihood = -946.39
## AIC=1896.77   AICc=1896.95   BIC=1901.3
```

```
# Check residuals of the model
checkresiduals(auto_model)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,0,0)(0,1,1)[12]
## Q* = 9.8275, df = 16, p-value = 0.8755
##
## Model df: 1.   Total lags used: 17
```

```r
# Forecast the next 12 months
forecast_values <- forecast(auto_model, h = 12)
plot(forecast_values, main = "Forecasted Tourists", ylab = "Tourists", xlab = "Time")
```




```
cat("\nConclusions:\n")
```

```
##
## Conclusions:
```

```
cat("1. The dataset was preprocessed to handle missing values and converted into a time series object.\n
```

```
## 1. The dataset was preprocessed to handle missing values and converted into a time series object.
```

```
cat("2. Stationarity tests indicated (non-)stationarity, and differencing was applied to make the series
```

```
## 2. Stationarity tests indicated (non-)stationarity, and differencing was applied to make the series s
```

```
cat("3. An ARIMA model was fitted, and the residuals were validated to behave like white noise.\n")
```

```
## 3. An ARIMA model was fitted, and the residuals were validated to behave like white noise.
```

```
cat("4. Future tourist numbers were forecasted for the next 12 months.\n")
```

```
## 4. Future tourist numbers were forecasted for the next 12 months.
```