Problem 3 (6 credits)

HW1

Danny Moncada monca016

February 07, 2020

```
suppressWarnings(suppressPackageStartupMessages({
  library(TSA)
  library(forecast)
  library(ggplot2)
  library(dplyr)
}))
```

Boston Crime Data

Crime incident reports are provided by Boston Police Department (BPD) to document the initial details surrounding an incident to which BPD officers respond. This data set contains the date of all crimes from 6/15/2015 to 9/3/2018. We are interested in knowing the frequency of crimes changed over months.

Question 1 (1 credit)

Please change your working directory and load the data `crime.txt`. Report the dimension of the data.

Hints:

- Set `header=T`

```
#Please insert your code below
setwd("~/Masters - Business Analytics/Spring 2020/MSBA 6430 - Advanced Issues in Business Analytics/HW1
crime = read.table("crime.txt", header = TRUE)

paste("Number of rows:", dim(crime)[1])
```

```
## [1] "Number of rows: 319073"
```

```
paste("Number of columns:", dim(crime)[2])
```

```
## [1] "Number of columns: 3"
```

Question 2 (1 credit)

Please aggregate the data based on their date. That is, we should end up with a smaller dataset where each row contains year, month, day, and the frequency of crimes on that date. Report the dimension of the new dataset.

Hints:

- Create an all-one vector having the same length as the data, then consider the `aggregate` function where you could set the `list` option for grouping elements, and set the `FUN` option as `sum`.
- Aggregating data is an important skill for almost everyday data cleaning.

```
#Please insert your code below
crime_ts = crime %>%
              group_by(Year, Month, Day) %>%
```

```
                    summarise(crimes = n())

paste("Number of rows (aggregate df):", dim(crime_ts)[1])
```

```
## [1] "Number of rows (aggregate df): 1177"
```

```
paste("Number of columns (aggregate df):", dim(crime_ts)[2])
```

```
## [1] "Number of columns (aggregate df): 4"
```

Question 3 (1 credit)

Sort the data by Year, Month, and Day. Report the first ten rows of the sorted data

Hints:

- Consider the `order` function

```
#Please insert your code below
crime_ts = crime_ts %>% arrange(Year, Month, Day)

head(crime_ts)
```

```
## # A tibble: 6 x 4
## # Groups:   Year, Month [1]
##    Year Month   Day crimes
##   <int> <int> <int>  <int>
## 1  2015     6    15    249
## 2  2015     6    16    249
## 3  2015     6    17    234
## 4  2015     6    18    294
## 5  2015     6    19    289
## 6  2015     6    20    261
```

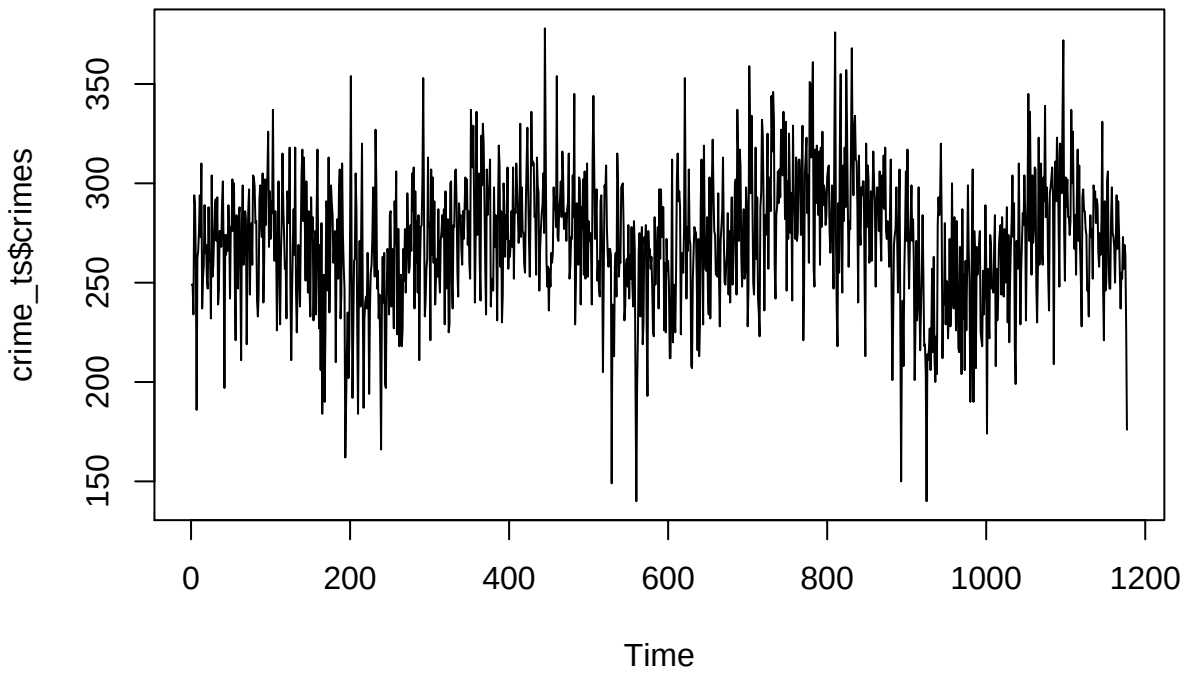
Question 4 (1 credit)

Plot the frequency of crimes by date. Do you see a pattern?

Hints:

- Consider the `ts.plot` function

```
#Please insert your code below
ts.plot(crime_ts$crimes)
```



Question 5 (1 credit)

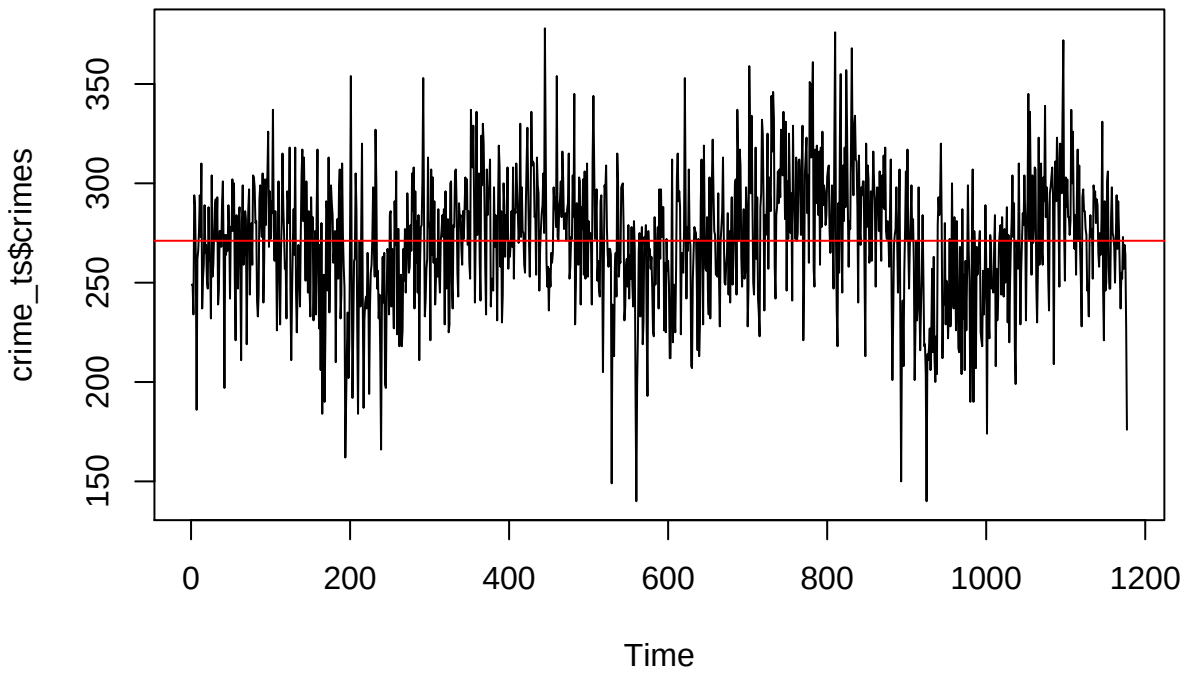
Is the time series stationary? Why?

Hints:

- Recall the definition of stationarity. What's the requirement on the mean function?

```
mu = mean(crime_ts$crimes)

ts.plot(crime_ts$crimes)
abline(h = mu, col = "red")
```



Question 6 (1 credit)

Which date has the highest crime frequency? How many crimes were reported on that day?

```
#Please insert your code and answer below

highest_crime = max(crime_ts$crimes)

highest_crime_date = crime_ts[crime_ts$crimes == highest_crime, ]
highest_crime_date
```

```
## # A tibble: 1 x 4
## # Groups:   Year, Month [1]
##    Year Month   Day crimes
##   <int> <int> <int>  <int>
## 1  2016     9     1    378
```

```
paste("This is how many crimes were reported on that day: ", highest_crime)
```

```
## [1] "This is how many crimes were reported on that day:  378"
```