Problem 3 (6 credits)

HW1

Michael Brucek (mbrucek), Eduardo Chavez (echavezh), Kevin Grady (grady133), Danny Moncada (monca016), Tian Zhang (zhan7003)

February 17, 2020

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

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.csv('crime.txt', header = T)
crime$Year.Month.Day = as.Date(crime$Year.Month.Day, format = "%Y %m %d")

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

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

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

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

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.

```r
#Please insert your code below
Frequency = rep(1, nrow(crime))
crime_ts = aggregate(Frequency, FUN = sum, by = list(Dates = crime$Year.Month.Day))
colnames(crime_ts)[2] <- "Frequency"

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

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

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

```
## [1] "Number of columns (aggregate df): 2"
```

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

```r
#Please insert your code below
crime_ts=crime_ts[order(crime_ts$Dates),]
head(crime_ts, 10)
```

```
##         Dates Frequency
## 1  2015-06-15       249
## 2  2015-06-16       249
## 3  2015-06-17       234
## 4  2015-06-18       294
## 5  2015-06-19       289
## 6  2015-06-20       261
## 7  2015-06-21       186
## 8  2015-06-22       262
## 9  2015-06-23       266
## 10 2015-06-24       276
```

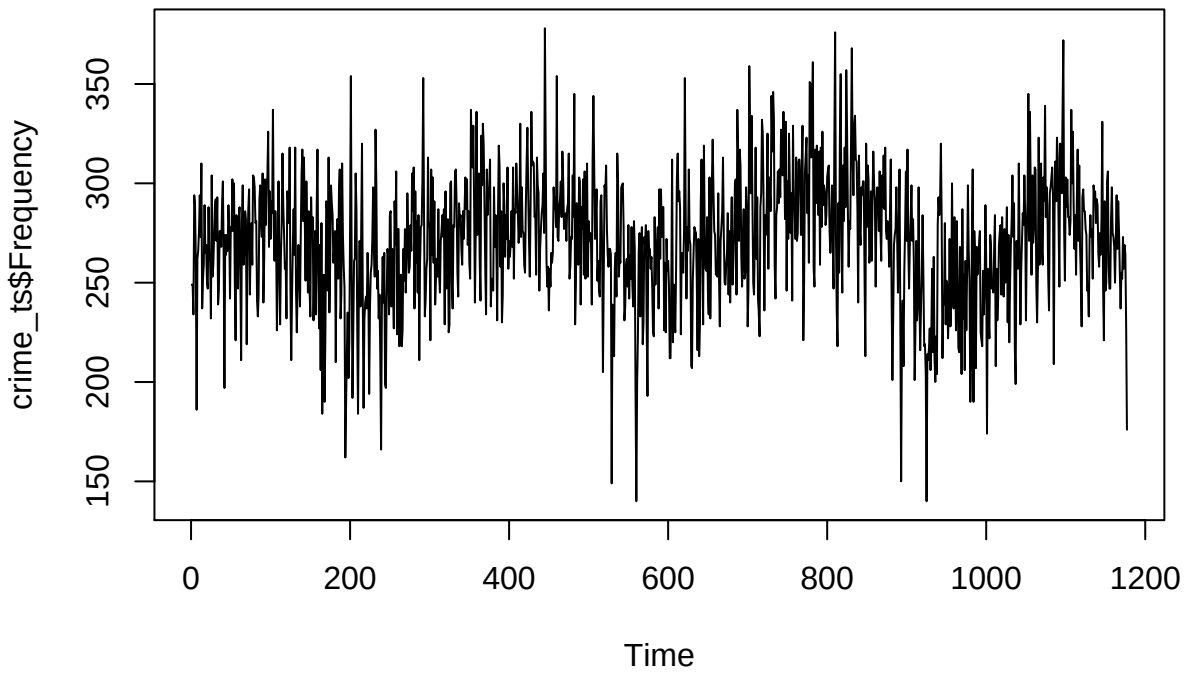
Question 4 (1 credit)

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

Hints:

- Consider the `ts.plot` function

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



Question 5 (1 credit)

Is the time series stationary? Why?

Hints:

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

```
#Please insert your answer below

"The requirement on the mean function for a stationary process is to remain constant."
```

```
## [1] "The requirement on the mean function for a stationary process is to remain constant."
```

```
paste("Overall mean:", mean(crime_ts$Frequency))
```

```
## [1] "Overall mean: 271.090059473237"
```

```
print("Mean in intervals of 200 days")
```

```
## [1] "Mean in intervals of 200 days"
```

```
mean(crime_ts$Frequency[1:200])
```

```
## [1] 266.94
```

```
mean(crime_ts$Frequency[201:400])
```

```
## [1] 267.68
```

```r
mean(crime_ts$Frequency[401:600])
```

```
## [1] 271.8
```

```r
mean(crime_ts$Frequency[601:800])
```

```
## [1] 281.38
```

```r
mean(crime_ts$Frequency[801:1000])
```

```
## [1] 265.22
```

```r
mean(crime_ts$Frequency[1001:1177])
```

```
## [1] 273.8362
```

```r
"The mean stays constant around the overall mean."
```

```
## [1] "The mean stays constant around the overall mean."
```

```r
######################

"It IS a stationary time series. From the plot we can tell its mean and variance +
stay constant or unconditional over time. And also, from the ADF test result below, +
we can safely say that we have no reason to believe it's non-stationary."
```

```
## [1] "It IS a stationary time series. From the plot we can tell its mean and variance +\nstay constan
```

```r
adf.test(crime_ts$Frequency)
```

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  crime_ts$Frequency
## Dickey-Fuller = -5.8543, Lag order = 10, p-value = 0.01
## alternative hypothesis: stationary
```

Question 6 (1 credit)

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

```r
#Please insert your code and answer below
head((crime_ts %>% arrange(desc(Frequency))),1)
```

```
##        Dates Frequency
## 1 2016-09-01       378
```

```r
"September 1st, 2016 had the highest crime frequency; 378 crimes were reported that day."
```

```
## [1] "September 1st, 2016 had the highest crime frequency; 378 crimes were reported that day."
```