STA 141A Fall 2019 - Homework 1

Quilvio Hernandez - A01

due October 14

Honor Code: The codes and results derived by using these codes constitutemy own work. I have consulted the following resources regarding thisassignment: Listed at the bottom

1. For each of the following case, describe the best possible data structure (e.g., array, data frame, list, table etc.) for representing the data. Also, write appropriate R codes to answer the questions that follow, treating the data as if it is given.

   (a) **Data:** A study of health effects of air quality in 10 major cities of the world involves daily measurements on the four variables: average temperature (temp), total precipitation (precip), maximum PM10 concentration (PM10) and number of deaths among elderly population (death). Measurements are available for five years.

   The best possible data structure would be a data frame with six columns, one for each variable (city, temp, precip, PM10, death) and one more for the days, and each row representing a different day. The total dataframe would be a 18260x6 data frame.

      i. What is the average number of deaths for each of the cities on days where the PM10 concentration is greater than 20 ?

      ```
      > prob1ai_df <- prob1_df[which(pm10 > 20), ]
      > aggregate(prob1a_df$deaths, list(prob1ai_df$cities), mean)
      ```

      ii. What is the average PM10 concentration for each of the cities on days with no precipitation and average temperature above 80 degrees F ?

      ```
      > prob1aii_df <- prob1_df[which((precip == 0) & (temp>80)), ]
      > aggregate(prob1aii_df$pm10, list(prob1aii_df$cities), mean)
      ```

   (b) **Data:** The data consist of records of patients' visit to a clinic. The measurements for each patient are: date of visit (visit), age of patient in years (age), gender with values M or F (gender), weight in lb (weight), systolic blood pressure (BP.sys), diastolic blood pressure (BP.dia), blood glucose level in mg/dl (glucose). For blood pressure levels, the unit is standard and the value is numeric with range between 0 and 600.

   The best possible data structure would be a data frame with eight columns, one for each variable (visit, age, gender, weight, BP.sys, BP.dia, glucose) and one more for the patient id, and each row representing a different visit. The patient id column would be generated using the a concatenation of age, gender and weight. This would be used to track the patient between visits.

      i. How many times did each patient visit the clinic ?

```
> prob2_pid <- cbind(prob2_df, paste(prob2_df$age, prob2_df$gender, prob2_df$we
> colnames(prob2_pid) <- c('visit', 'age', 'gender', 'weight', 'BP.sys', 'BP.di
> table(prob2_pid$`patient id`)
```

ii. What is the average systolic blood pressure level for each of the patients with maximum weight (during the study period) greater than 180 lb ?

```
> probs2_ii <- prob2_pid[which(weight > 180), ]
> aggregate(probs2_ii$BP.sys, list(probs2_ii$`patient id`), mean)
```

iii. What is the average blood glucose level for each of the patients with age at least 40 years at the first visit ?

```
> prob2_iii <- prob2_pid[which(age >= 40), ]
> aggregate(prob2_iii$glucose, list(prob2_iii$`patient id`), mean)
```

2. Suppose you have:

- four types of animals: cat, dog, cow, squirrel;
- four possible colors: white, black, brown, red;
- five possible attribute: big, small, angry, cute, finicky.

(a) Generate random samples, with replacement, of size 100 from each of the types. Call the resulting vectors of character strings as: Animal, Color, Attribute.

```
> adj_list = c('big', 'small', 'angry', 'cute', 'finicky')
> colors_list = c('white', 'black', 'brown', 'red')
> animal_list = c('cat', 'dog', 'cow', 'squirrel')
> Attribute = sample(adj_list, size = 100, replace = TRUE)
> Color = sample(colors_list, size = 100, replace = TRUE)
> Animal = sample(animal_list, size = 100, replace = TRUE)
```

(b) Write an R code to combine the results to produce phrases (character strings) describing the animals, as in this example: big white dog.

```
> strings = paste(Attribute, Color, Animal, sep = ' ')
```

(c) Create a frequency distribution (or contingency table) of the different types of animals together with colors and attributes based on the sampled data.

```
> var_table <- table(Attribute, Color, Animal)
> var_table
```

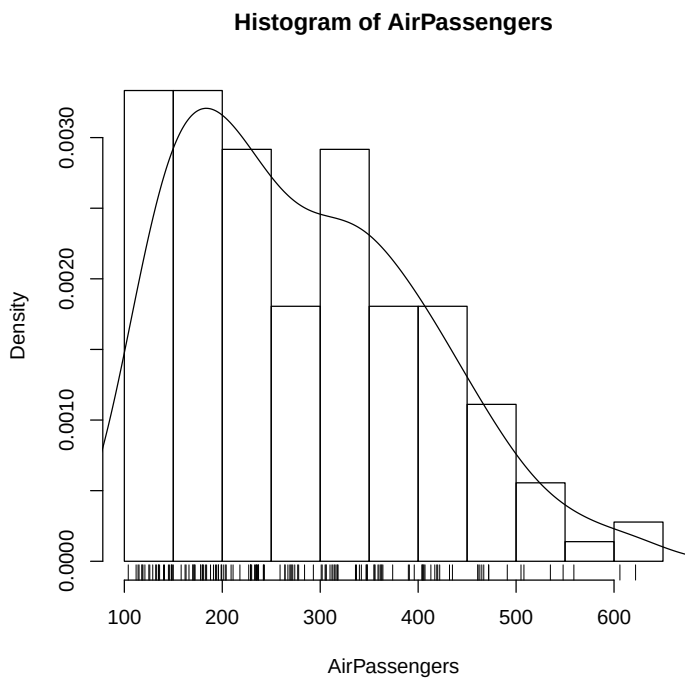
(d) Use the result in part (c) to obtain the frequency distribution of: (i) Animal vs. Color; (ii) Animal vs. Attribute; (iii) Animal.

```
> var_table <- table(data.frame(Attribute, Color, Animal))
> var_table
> table(data.frame(Color, Animal))
> table(data.frame(Attribute, Animal))
> library(plyr)
> count(data.frame(Animal))
```

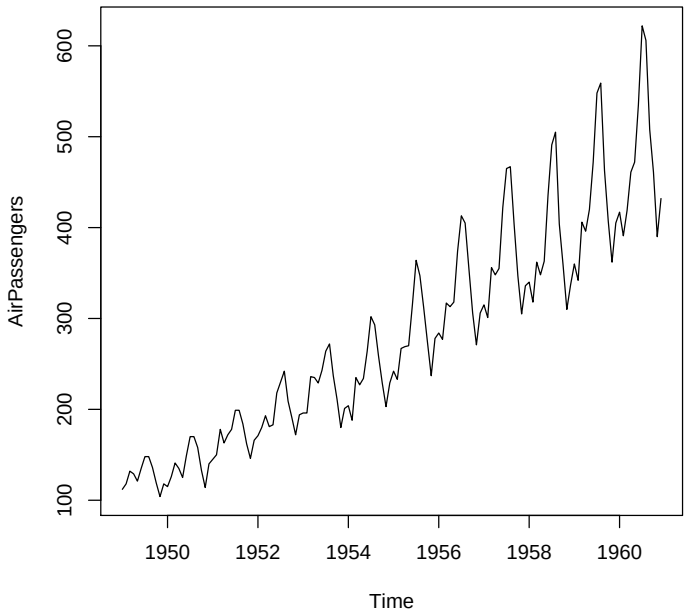
3. Give an informative graphical statistical summary of the following datasets (available with base R). In each case, write very brief (maximum of 100 words) description highlighting the findings. You may use up to 2 plots for illustrating the features of each data set.

   (a) `AirPassengers` : Monthly airline passenger numbers during 1949–1960.

```
> hist(AirPassengers, probability = TRUE)
> lines(density(AirPassengers))
> rug(AirPassengers)
```



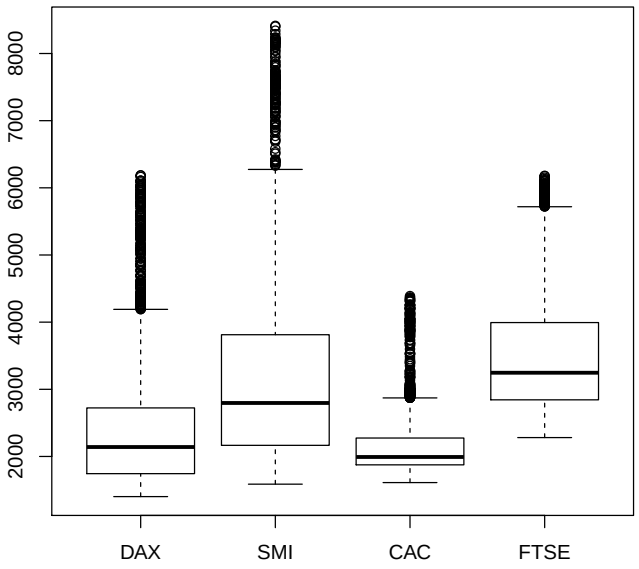

```
> plot(AirPassengers)
```



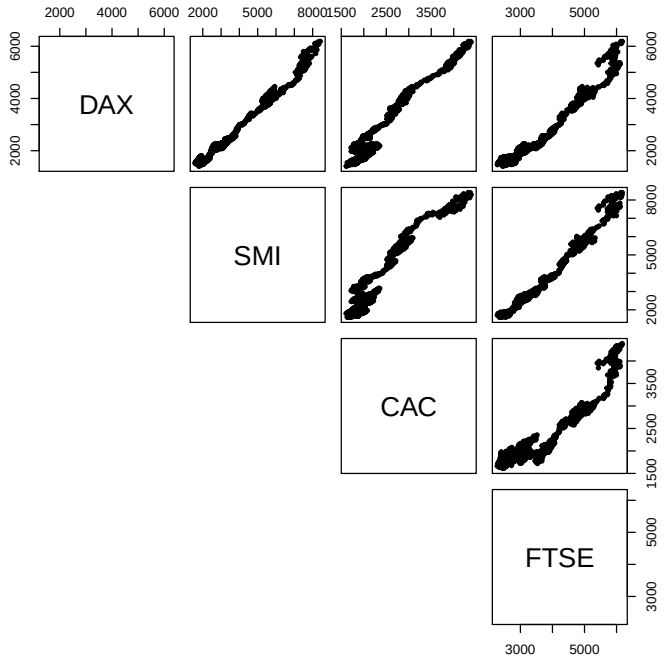
The `AirPassengers` dataset contains monthly totals of international airline passengers from 1949 to 1960, in thousands. The dimensions of the dataset are 144x1, with a minimum of 104, maximum of 622, mean of 280, and a median value of 265. The standard deviation is 199 and the IQR is 180.5. A boxplot (not shown) would demonstrate there are no outliers. The histogram reveals a right skewedness in the datset. The time series plot over time shows an overall increase between 1950 and 1960 with yearly spikes, possibly due to holiday travel.

(b) `EuStockMarkets` : Daily closing prices of major European stock indices during 1991–1998.

```
> boxplot(EuStockMarkets)
```



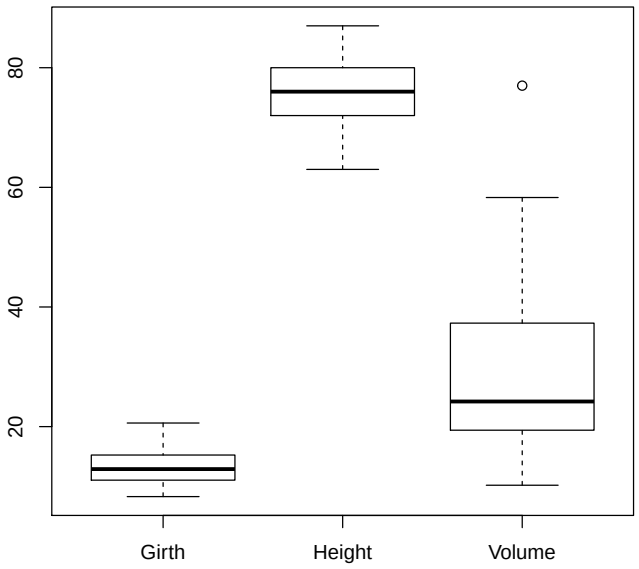

```
> pairs(EuStockMarkets, pch = 20, lower.panel = NULL)
```



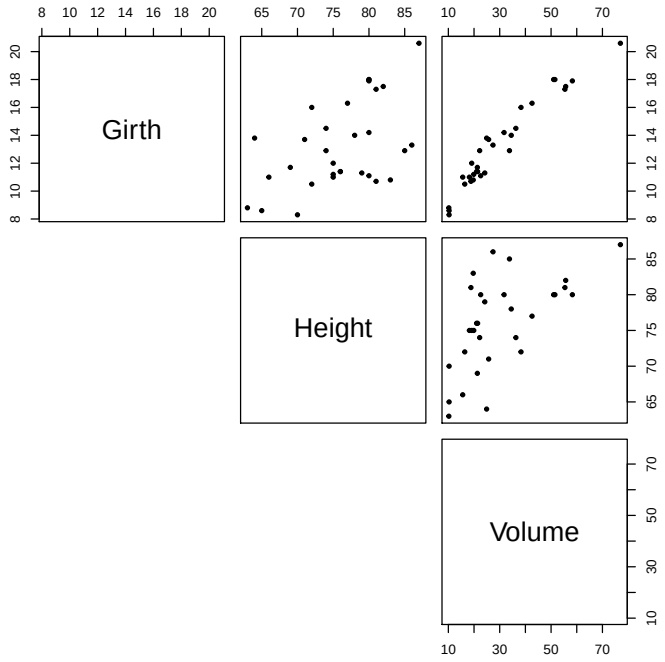
The `EuStockMarkets` dataset contains the daily closing prices of major European stock indices between 1991-1998. A examination of the correlation plot (not shown) would show that the 4 indices are all highly correlated. This is further emphasized by the strong positive correlations in all the pairs maps. The boxplots of the 4 variables would suggest that there the is a large spread in all 4 variables. SMI has the largest values and variance while CAC has the smallest values.

(c) `trees` : Girth, weight and volume for Black Cherry trees.

```
> boxplot(trees)
```




```
> pairs(trees, pch = 20, lower.panel = NULL)
```



The `trees` gives the girth, weight and volume for Black Cherry trees. The shape of the boxplots are symmetric. This would suggest that the data for all three variables appears to be relatively normally distributed around the mean, with a single outlier for volume. Examining the pairs plots would infer that there is a strong positive correlation between girth and volume, while there is a moderately positive correlation between girth and height, and volume and height. This is to be expected given any one of these variables could be dependent on the other two.

Code Appendix

```
> ##1a
> temp <- runif(1826, min = 60, max = 110)
> precip <- sample(rpois(10*(365*5+1), .5), 10*(365*5+1), replace = TRUE)
> deaths <- sample(0:100, 10*(365*5+1), replace = TRUE)
> pm10 <- sample(0:30, 10*(365*5+1), replace = TRUE)
> prob1_df <- data.frame(c(rep("San Diego",1826),rep("San Francisco",1826),rep("New York",18
+                          rep("Los Angeles",1826),rep("Austin",1826),rep("Seattle",1826),
+                          rep("Portland",1826),rep("Tampa Bay",1826),rep("Lansing",1826),
+                          rep("Philadelphia",1826)), temp, precip, deaths, pm10, rep(1:1826
> colnames(prob1_df) <- c('cities', 'temp', 'precip', 'deaths', 'pm10')
```

```
> prob1ai_df <- prob1_df[which(pm10 > 20), ]
> aggregate(prob1a_df$deaths, list(prob1a_df$cities), mean)
> prob1aii_df <- prob1_df[which((precip == 0) & ((aggregate(prob1_df$temp, list(prob1_df$cit
> aggregate(prob1aii_df$pm10, list(prob1aii_df$cities), mean)
> ##1b
> visit <- sample(seq(as.Date('1999/01/01'), as.Date('2000/01/01'), by="day"), 30, replace =
> age <- sample(0:95, 30, replace = TRUE)
> gender <- sample(c("M", "F"), 30, replace = TRUE)
> weight <- sample(0:300, 30, replace = TRUE)
> BP.sys <- sample(0:600, 30, replace = TRUE)
> BP.dia <- sample(0:600, 30, replace = TRUE)
> glucose <- sample(70:250, 30, replace = TRUE)
> prob2_df <- data.frame(visit, age, gender, weight, BP.sys, BP.dia, glucose)
> prob2_pid <- cbind(prob2_df, paste(prob2_df$age, prob2_df$gender, prob2_df$weight, sep = '
> colnames(prob2_pid) <- c('visit', 'age', 'gender', 'weight', 'BP.sys', 'BP.dia', 'glucose'
> table(prob2_pid$`patient id`)
> probs2_ii <- prob2_pid[which(weight > 180), ]
> aggregate(probs2_ii$BP.sys, list(probs2_ii$`patient id`), mean)
> prob2_iii <- prob2_pid[which(age >= 40), ]
> aggregate(prob2_iii$glucose, list(prob2_iii$`patient id`), mean)
> ##2
> adj_list = c('big', 'small', 'angry', 'cute', 'finicky')
> colors_list = c('white', 'black', 'brown', 'red')
> animal_list = c('cat', 'dog', 'cow', 'squirrel')
> Attribute = sample(adj_list, size = 100, replace = TRUE)
> Color = sample(colors_list, size = 100, replace = TRUE)
> Animal = sample(animal_list, size = 100, replace = TRUE)
> strings = paste(Attribute, Color, Animal, sep = ' ')
> var_table <- table(data.frame(Attribute, Color, Animal))
> var_table
> table(data.frame(Color, Animal))
> table(data.frame(Attribute, Animal))
> library(plyr)
> count(data.frame(Animal))
> ##3
> data("AirPassengers")
> data("EuStockMarkets")
> data("trees")
> summary(AirPassengers)
> boxplot(AirPassengers)
> hist(AirPassengers, probability = TRUE)
> lines(density(AirPassengers))
> rug(AirPassengers)
> plot(AirPassengers)
> # stripchart(AirPassengers, method = 'stack')
> # dotchart(as.numeric(AirPassengers))
```

```
>
>
> summary(EuStockMarkets)
> boxplot(EuStockMarkets)
> pairs(EuStockMarkets, pch = 20, lower.panel = NULL)
> corrplot(cor(EuStockMarkets), type = 'upper', method = 'color', col = brewer.pal(10, "PiYG
> summary(trees)
> boxplot(trees)
> pairs(trees, pch = 20, lower.panel = NULL)
> corrplot(cor(trees), type = 'upper', method = 'color', col = brewer.pal(10, "PiYG"))
>
```

References

- https://www.latex-tutorial.com/tutorials/lists/
- https://www.overleaf.com/learn/latex/Font_typefaces
- https://piazza.com/class/k0wv3t2y44e6vc?cid=15
- https://stackoverflow.com/questions/21502332/generating-random-dates
- https://stackoverflow.com/questions/21982987/mean-per-group-in-a-data-frame
- https://moderndata.plot.ly/create-colorful-graphs-in-r-with-rcolorbrewer-and-plotly/
- https://pjbartlein.github.io/GeogDataAnalysis/lec02.html
- https://yihui.name/knitr/demo/minimal/
- https://chemicalstatistician.wordpress.com/2015/02/03/how-to-get-the-frequency-table-of-a-categorical-variable-as-a-data-frame-in-r/