CS544 Module 3 Assignment

Using R code, do the following:

Part 1) 25 points

Initialize the dataset about richest people in the Forbes billionaires list as shown below:

```
forbes <- read.csv("https://people.bu.edu/kalathur/datasets/forbes.csv")
```

a) Show the barplot of the frequencies of the number of rich people by country.

b) Using an appropriate plot, show the distribution of the females and males in the dataset.

c) Consider the top 5 categories in the dataset. Show how the females and males are distributed across these top 5 categories using the appropriate plot.

d) What inferences do you draw from the above plots?

Part 2) 25 points

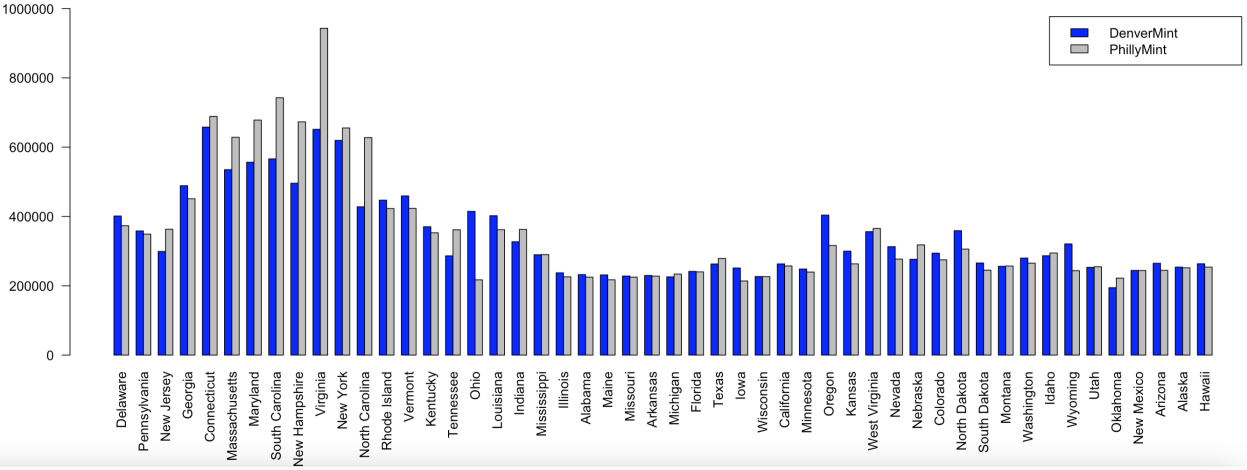
Initialize the dataset about the quarter coin productions of the 50 US states by the *DenverMint* and *PhillyMint*. The numbers in the dataset (in thousands) are the number of quarters minted. With the **R code** for the following:

us_quarters <- read.csv("https://people.bu.edu/kalathur/datasets/us_quarters.csv")

```
> head(us_quarters)
          State DenverMint PhillyMint
1      Delaware     401424     373400
2  Pennsylvania     358332     349000
3    New Jersey     299028     363200
4       Georgia     488744     451188
5   Connecticut     657880     688744
6 Massachusetts     535184     628600
```

a) For which state were the highest number of quarters produced by each mint? For which state were the lowest number of quarters produced by each mint?

b) Produce the following barplot from the data using the **R barplot** function with the data for the two mints as a matrix. Write any two striking inferences you can observe by looking at the plot.



c) Show the side-by-side box plots for the two mints. Write any two inferences for each of the box plots.

d) Using R code, what states would be considered as outliers for each of the two mints. Use the five number summary function to derive the outlier bounds.

Part 3) 25 points

Use the stocks dataset with the weekly closing values for the year 2021 initialized as shown below:

stocks <- read.csv("https://people.bu.edu/kalathur/datasets/stocks.csv")

```
> head(stocks)
        Date MSFT AAPL GOOG  FB AMZN TSLA
1 2021-01-01  216  130 1787 269 3162  816
2 2021-01-08  211  128 1740 246 3127  845
3 2021-01-15  223  136 1891 273 3307  845
4 2021-01-22  237  136 1863 265 3238  835
5 2021-01-29  240  137 2062 266 3331  850
6 2021-02-05  242  134 2096 270 3262  812
```

a) Show the pair wise plots for all the 6 stocks in the dataset in a single plot.
b) Show the correlation matrix for the 6 stocks in the dataset rounded to 2 decimals.
c) Provide at least 4 interpretations of the results.
d) Store the correlation matrix from **b)** in the variable ***cm*.** Using loops, for each stock in the dataset, show the top 3 correlated stocks for that respective stock. The code should work for any dataset with any number of stocks. Sample output for the given dataset is shown below:

```
Top 3 for Stock MSFT
GOOG AAPL TSLA
0.95 0.90 0.71

Top 3 for Stock AAPL
MSFT GOOG TSLA
0.90 0.79 0.73

Top 3 for Stock GOOG
MSFT   FB AAPL
0.95 0.85 0.79

Top 3 for Stock FB
GOOG MSFT AMZN
0.85 0.68 0.66

Top 3 for Stock AMZN
GOOG   FB MSFT
0.67 0.66 0.64

Top 3 for Stock TSLA
AAPL MSFT GOOG
0.73 0.71 0.47
```

Part 4) 25 points

Initialize the scores of 100 students as shown below:

scores <- read.csv("https://people.bu.edu/kalathur/datasets/scores.csv")

a) Show the default histogram of the student scores. Save the result of the histogram into a variable. Using only the **counts** and **breaks** property of this variable, write the R code to produce the following output. The code for the following output should not refer to the individual *scores* in the dataset.

```
 3 students in range (35,40]
 4 students in range (40,45]
10 students in range (45,50]
13 students in range (50,55]
17 students in range (55,60]
27 students in range (60,65]
13 students in range (65,70]
 8 students in range (70,75]
 3 students in range (75,80]
 2 students in range (80,85]
```

b) Using the breaks option of the histogram, show the histogram and the custom output as shown below so that students in the range (70,90] get an A grade, (50,70] get a B grade, and (30-50] get a C grade. The code for the following output should not refer to the individual scores and should be using only the **counts** and **breaks** of the histogram.

```
17 students in C grade range (30,50]
70 students in B grade range (50,70]
13 students in A grade range (70,90]
```

Submission:

When the term *lastName* is referenced, please replace it with your last name.

Provide all R code in a single file, **CS544_HW3_LastName.R**. Clearly mark each subpart of each question.

Provide the corresponding outputs from the R console and respective plots in a single PDF document, **CS544_HW3_LastName**.pdf.

Upload the two files to the Assignments section of Blackboard.

Note: Only ONE submission is allowed. Please be sure that what you are submitting is your final submission.