CS544 Module 3 Assignment

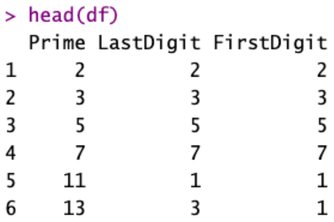
Using R code, do the following:

Part 1) 20 points

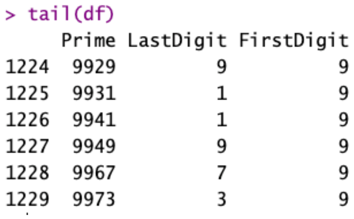
Initialize the dataset about prime numbers as shown below:

df <- read.csv("http://people.bu.edu/kalathur/datasets/myPrimes.csv")

The resulting data frame of the primes below 10000 along with their last and first digits is as shown below:



```
> head(df)
  Prime LastDigit FirstDigit
1     2         2          2
2     3         3          3
3     5         5          5
4     7         7          7
5    11         1          1
6    13         3          1
```



```
> tail(df)
     Prime LastDigit FirstDigit
1224  9929         9          9
1225  9931         1          9
1226  9941         1          9
1227  9949         9          9
1228  9967         7          9
1229  9973         3          9
```

a) Show the barplot of the frequencies for the last digit.

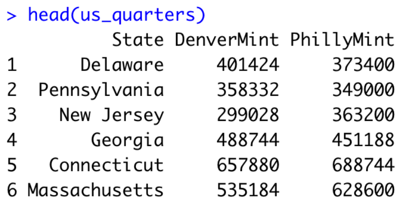
b) Show the barplot of the frequencies for the first digit.

c) What inferences do you draw from these two plots? (two inferences from each plot)

Part 2) 30 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("http://people.bu.edu/kalathur/datasets/us_quarters.csv")

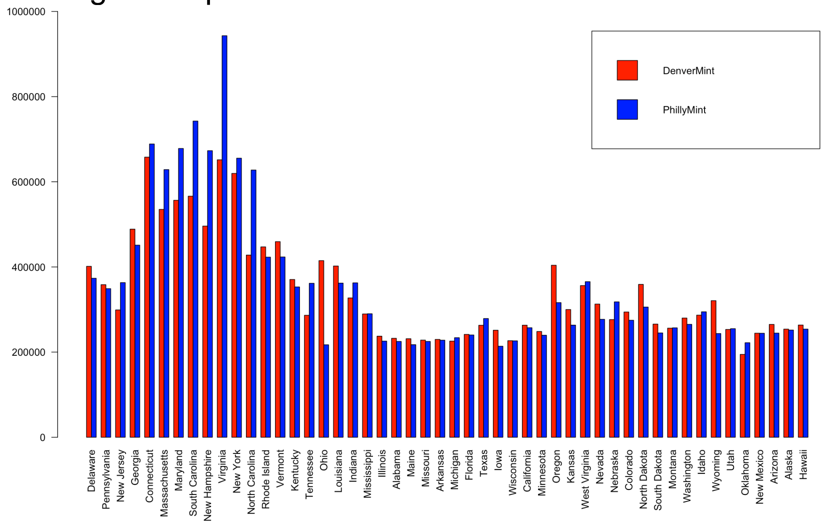


```
> head(us_quarters)
          State DenverMint PhillyMint
1      Delaware     401424     373400
2  Pennsylvania     358332     349000
3    New Jersey     299028     363200
4       Georgia     488744     451188
5   Connecticut     657880     688744
6 Massachusetts     535184     628600
```

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) What is the value of the total coins in dollars?

c) 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.



d) Show the scatter plot of the number of coins between the two mints. Write any two inferences you can observe looking at the plot.

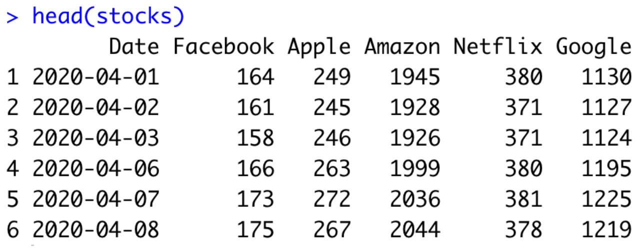
e) Show the side-by-side box plots for the two mints. Write any two inferences for each of the box plots.

f) 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) 20 points

Use the **FAANG** stocks dataset with the April daily High values initialized as shown below:

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



```
> head(stocks)
        Date Facebook Apple Amazon Netflix Google
1 2020-04-01      164   249   1945     380   1130
2 2020-04-02      161   245   1928     371   1127
3 2020-04-03      158   246   1926     371   1124
4 2020-04-06      166   263   1999     380   1195
5 2020-04-07      173   272   2036     381   1225
6 2020-04-08      175   267   2044     378   1219
```

a) Show the pair wise plots for all the 5 stocks in the dataset in a single plot.

b) Show the correlation matrix for the 5 stocks in the dataset.

c) Provide at least 4 interpretations of the results.

Part 4) 30 points

Initialize the scores of 100 students as shown below:

scores <- read.csv("http://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 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.

```
 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.

```
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:

- You must work on your assignments individually. You are not allowed to copy the answers from the others.
- Each assignment has a strict deadline. However, you are still allowed to submit your assignment within 2 days after the deadline with a penalty. 15% of the credit will be deducted unless you made previous arrangements with your facilitator and professor. Assignments submitted 2 days after the deadline will not be

Create a folder, CS544_HW3_lastName and place the following files in this folder.

Provide the R code, **HW3_lastName.R**, with each portion of the code clearly identified by the corresponding question. Prepare a corresponding Word/PDF document by pasting the output for each question including the plots and explanations.

Archive the folder (CS544_HW3_lastName.zip). Upload the zip file to the Assignments section of Blackboard.