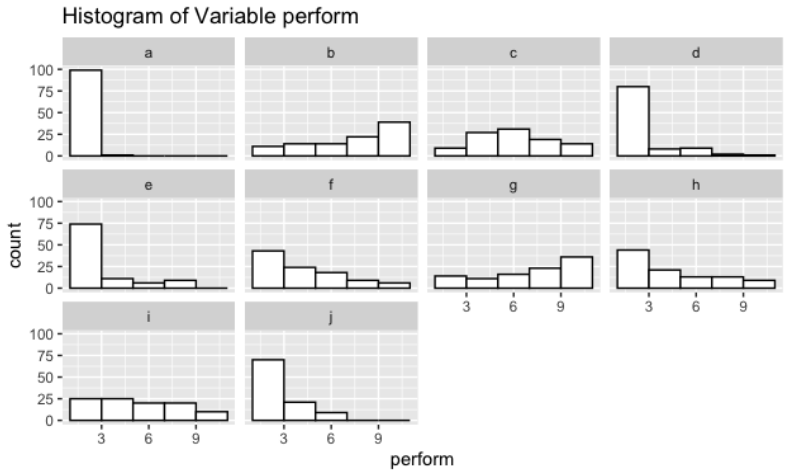
BUSI2045 – Homework 1

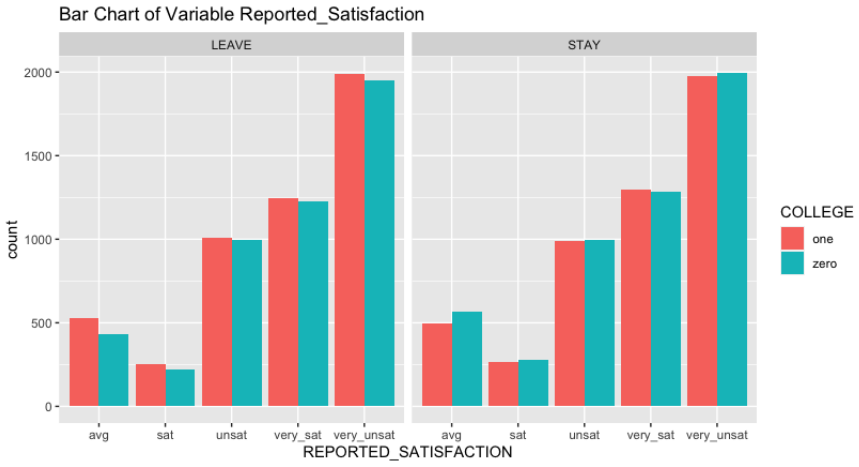
Note: Please include both R codes and results in your solutions. (You may use the "Compile Report" function under Menu "File" in RStudio to generate a Word/PDF report of both R codes and results)

Question 1: Data Exploration and Visualization (20 points)

(a) Read the dataset *brand_ratings.csv* into R. Construct a histogram plot (as below) using variable ***perform***.
Hints: you may need to (i.) adjust the ***binwidth***; (ii.) adjust the aesthetical attributes ***fill*** and ***color***; (iii.) facet the histogram in different panels according to the variable ***brand***.



(b) Load the dataset *churn.arff* into R. Create a bar plot using the variable ***REPORTED_SATISFACTION***. Your output should look similar as the below graph.
Hints: (i.) Specify the portions in color according to the ***COLLEGE***; (ii.) Separate the plot in different panels according to the variable ***LEAVE***.



Question 2 Describe Data (40 Points)

Read the file *Assignment1_Q2.csv* into R. Construct a subset called ***Q2_data*** such that ***X3*** has no missing value. Use ***Q2_data*** to answer following questions:

(a) How many observations in this data set? What are the types (numeric, integer, etc.) of these variables?

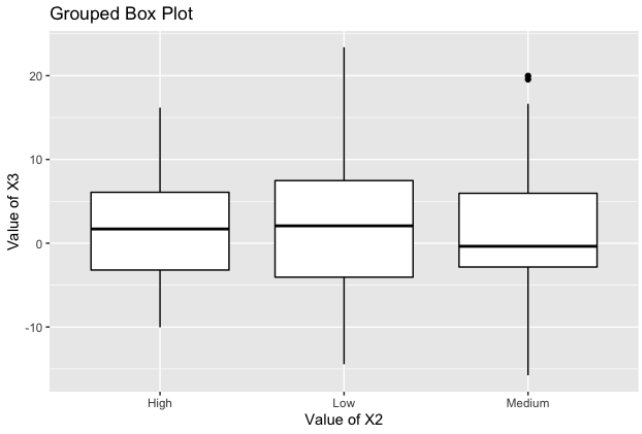
(b) Which variable(s) belong to the discrete variable? Check the unique values for these discrete variables. Which variable(s) belong to the continuous variable? Check the values of mean, standard deviation, and range for these continuous variables.

(c) Construct a frequency table as below.
Hints: You may inspect the dataset to find suitable variables first. Ignore the order of the columns and rows.

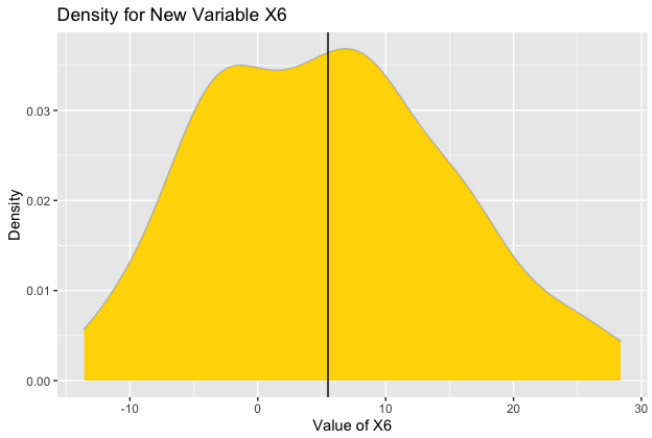
	North	South	West	East
High	12	7	9	11
Median	?	?	?	?
Low	?	?	?	?

(d) Is variable ***X4*** normally distributed? Use ggplot2 to create a QQ plot to help answer this question.

(e) Recreate the following boxplot for variable ***X3*** across the different levels of ***X2***. The result should look like the below.



(f) Create a new variable ***X6*** which is the sum of ***X3*** and ***X4***. Visualize the distribution of ***X6*** as below.
Hints: (i.) There are two layers in total, including a density plot, and a vertical line to mark the mean of ***X6***.



Question 3: Describe Data (40 Points)

(a) Read the file *marketing_campaign.csv* in R and construct a subset named ***df2_sub*** where the variable ***Income*** contains no missing value, and variables ***NumStorePurchases*** and ***NumWebPurchases*** are not equal to 0. How many observations and variables are in this subset?

Answer the following questions based on this new dataset.

(b) What are the values of 10%, 50%, 80% percentile for variable ***Income***?

(c) Write a named function to compute the ratio of the interquantile value (the difference between 75% and 25% quantile value) against the range (i.e., the difference between max and min value) of a variable. Apply that function to three variables (i.e., ***Income, NumStorePurchases*** and ***NumWebPurchases***) in the dataset.
Hint: you may need to use ***apply()*** *function.*

(d) Write an anonymous function to solve the above question.