STATS 10 Assignment 1

Please submit both parts of the assignment in one single PDF file. You can use any PDF editor software to merge the two parts into one file. Please make sure that the questions are in the correct order and clearly labeled, and that the answers are legible and easy to read.

To submit your assignment, upload the PDF file under the designated assignment page on the course website before the deadline specified. Email or hard copy submissions are not accepted.

Part I

Include both the R commands and their corresponding outputs, results, or answers for all exercise questions in Part I.

1. Vectors:
   a. Create a vector named *heights* that contains the heights, in inches, of yourself and two students near you. Print the contents of this vector.
   b. Create a vector named *names* that contains the names of these people. Print the contents of this vector.
   c. Try typing cbind(heights, names). What did this command do? What class is this new object?
   Hint: Try the class() function.

2. Downloading data:
   a. Download the data set births.csv from the course site and upload it into RStudio. Name the data frame *NCbirths*.
   b. Demonstrate that you have been successful by typing head(NCbirths) and copying and pasting the output into your word processing document.

3. Package loading
   a. Install the maps package. Verify its installation by typing find.package("maps") and include the output in your answer.
   b. Type library(maps) to load up the package. Type map("state") and include the plot output in your answer.

Use the births data set for questions 4-11

4. Perform vector operations
   a. Extract the weight variable as a vector from the data frame
   b. What units do you think the weights are in?
   c. Create a new vector named weights_in_pounds which are the weights of the babies in pounds. You can look up conversion factors on the internet.
   d. Demonstrate your success by typing weights_in_pounds[1:20] and including the output in your word processing document.

5. What is the mean weight of the babies in pounds?
   a. What percentage of the mothers in the sample smoke? *Hint: use the tally function with the format argument. Use the help screen for guidance.*
   b. According to the Centers for Disease Control, approximately 21% of adult Americans are smokers. How far off is the percentage you found in 2 from the CDC's report?

6. Produce three different histograms of the weights in pounds. Use 3 bins, 20 bins, and 100 bins. Which histogram seems to give the best visualization, and why?

7. We can use the syntax boxplot(vector1, vector2) to make a side by side box plot. Create a side by side boxplot of the mother's ages and the father's ages. Which gender tends to be older?

8. Try typing histogram(~ weight | Habit, data = NCbirths, layout = c(1, 2)). Describe what this code does. Based on the graph, do you see any major differences between baby weights from smoking moms vs. non-smoking moms?

9. Produce a dot plot of the weights in pounds.

10. Consider the other categorical variables in this data. Of those that record the health of the baby, which do you think will be associated with the mother's smoking and why? Make a two-way Summary Table to check your hypothesis. Do you have evidence that this variable associated with smoking? Why?

11. Produce a nicely formatted scatter plot of the weight of the baby vs. the mother's age.

Part II

You may choose to type or write your answers electronically or scan your handwritten solutions. Please ensure that you show all steps and explanations to receive full credit, unless otherwise instructed.

1. A data set on Shark Attacks Worldwide posted on StatCrunch records data on all shark attacks in recorded history including attacks before 1800. The data set can be viewed here: https://www.statcrunch.com/app/index.html?dataid=2188687

   a. How many variables are contained in the data?

   b. Which of the following questions could not be answered using this data set? Briefly explain.

      i. In what month do most shark attacks occur?

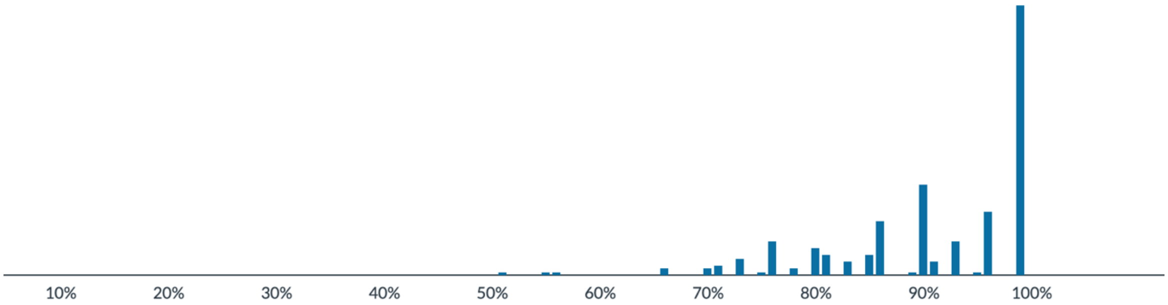
      ii. Are shark attacks more likely to occur in warm temperature or cooler temperatures?

      iii. Attacks by which species of shark are more likely to result in a fatality?

      iv. What country has the most shark attacks per year?

   c. A researcher wants to understand the age of the people in the data set and proposed some questions of interest: Are the reported cases are mostly younger people or older people? How is the age distributed? How would you help the research answer these questions? What statistical tools (e.g., graphs, measures) will you use? (You only need to describe your approach)

2. The scores of a quiz are displayed in the graph below.



   a. Describe the shape of distribution

   b. Would the mean score be greater than, less than, or about the same as the median score? Explain.

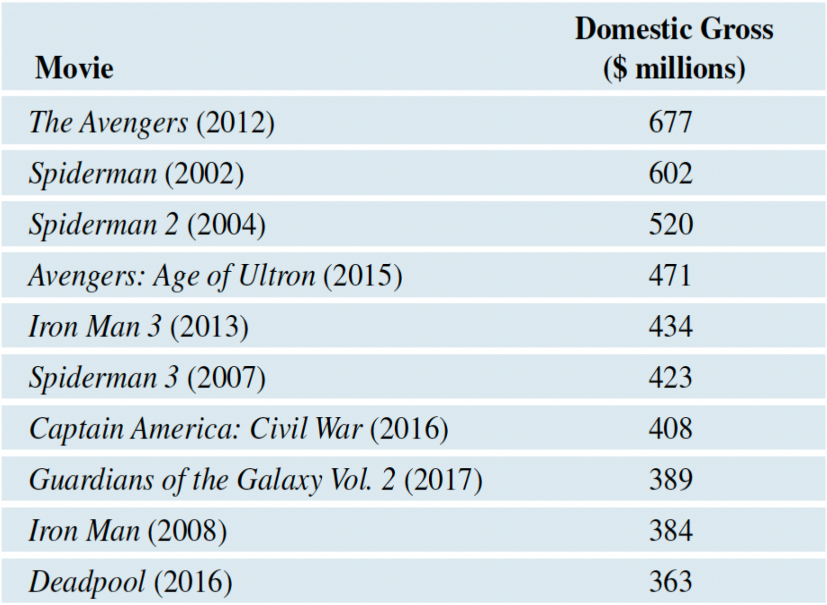
   c. What measures would you use to report the center and spread. Explain.

3. The distribution of test scores in a class is unimodal and symmetric with a mean of 80 pts and a standard deviation of 7pts. Based on the information, Adam estimated that his score is higher than approximately 97.5% of the students in class. What score did Adam receive? Explain.

4. Assume that both men and women's heights have symmetric and unimodal distributions. Women's distribution has a mean of 64 inches and a standard deviation of 2.5 inches. Men's distribution has a mean of 69 inches and a standard deviation of 3 inches.
   a. What women's height corresponds with a z-score of -1.50?

   b. Professional basketball player Evelyn Akhator is 75 inches tall and plays in the WNBA (women's league). Professional basketball player Draymond Green is 79 inches tall and plays in the NBA (men's league). Compared to their own peers, who is taller?

5. The top ten movies based on Marvel comic book characters for the U.S. box office as of fall 2017 are shown in the following table, with domestic gross rounded to the nearest hundred million. (Source: ultimatemovieranking.com)

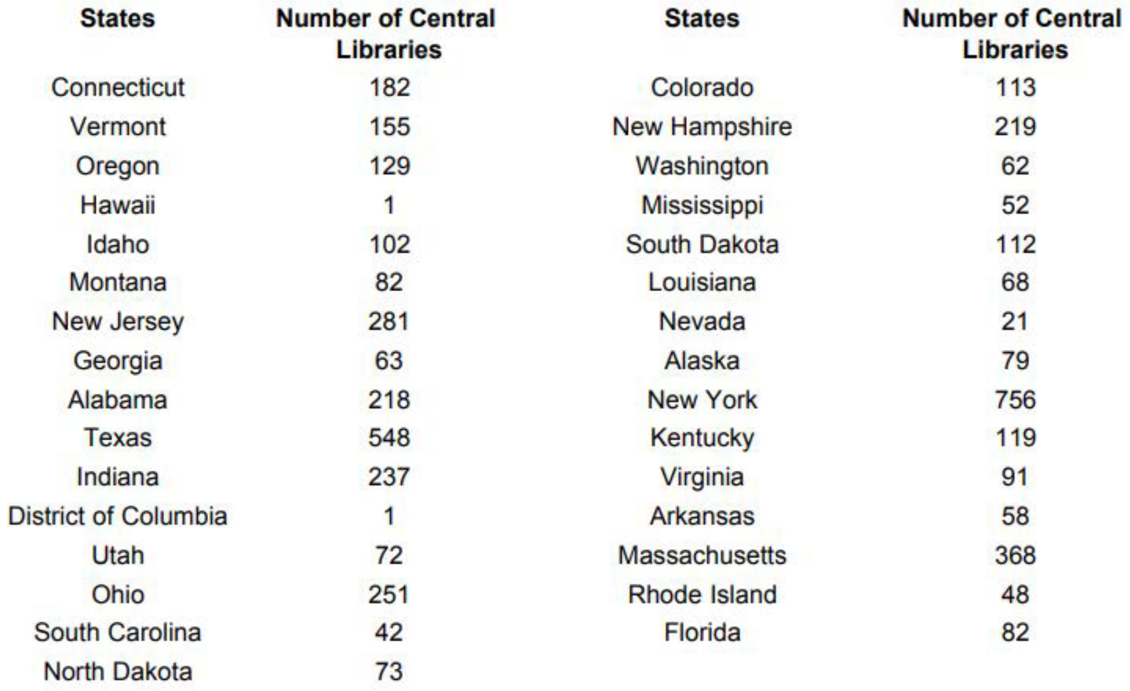


Movie	**Domestic Gross ($ millions)**
The Avengers (2012)	677
Spiderman (2002)	602
Spiderman 2 (2004)	520
Avengers: Age of Ultron (2015)	471
Iron Man 3 (2013)	434
Spiderman 3 (2007)	423
Captain America: Civil War (2016)	408
Guardians of the Galaxy Vol. 2 (2017)	389
Iron Man (2008)	384
Deadpool (2016)	363

   a. Report the five-number summary of the domestic gross income.

   b. Interpret the five-number summary in context, i.e., what information can you obtain about the distribution of the domestic gross income?

6. The data set below show the number of central public libraries in 32 states.



States	Number of Central Libraries	States	Number of Central Libraries
Connecticut	182	Colorado	113
Vermont	155	New Hampshire	219
Oregon	129	Washington	62
Hawaii	1	Mississippi	52
Idaho	102	South Dakota	112
Montana	82	Louisiana	68
New Jersey	281	Nevada	21
Georgia	63	Alaska	79
Alabama	218	New York	756
Texas	548	Kentucky	119
Indiana	237	Virginia	91
District of Columbia	1	Arkansas	58
Utah	72	Massachusetts	368
Ohio	251	Rhode Island	48
South Carolina	42	Florida	82
North Dakota	73		

The five number summary is given as:

Minimum	Q1	Median	Q3	Maximum
1	62	91	218	756

Sketch a boxplot using the five-number summary above and the data below.
Mark the values of the quartiles, the lower whisker, the upper whisker, and any potential outliers in the boxplot. Explain how you determined the length of the whiskers.
(The scale of the plot does not need to be accurate)