Stat 311 Homework 2

This assignment has some problems related to Lesson2 and emphasizes exploratory data analysis (EDA)—visualization and numeric summaries for qualitative and quantitative data. We recommend that you create a new folder for this assignment. Download the data files and the Rmd template to this folder before you begin. This template only provides the header and setup code, and the headers for the main problems. You need to add everything else. Use the same formatting structure as you did for Homework 1, using #### a), etc. to label subparts of problems. Check out the .Rmd files that appear on the Lesson 2 Presentations page—they contain code examples for different types of summaries that were presented in the lectures. Upload your knitted HTML file to Canvas.

Problems 1 – 3 do not require any code. Simply type your answers into the .Rmd file. Problems 4 and 5 require the use of R code.

To reinforce the concepts in the Lesson 2 lectures and for extra practice with R commands, I recommend that you try some of the OpenIntro tutorials that I linked on the Readings page for Lesson 2.

1. For each part, compare the distributions, A and B, based on means/SDs and medians/IQRs. Do not show any calculations. Simply state how the means/SDs or means/IQRs compare. Make sure to explain your reasoning.
   a) Compare the means/SDs for **A:** 3, 5, 5, 5, 8, 11, 11, 11, 13 and **B:** 3, 5, 5, 5, 8, 11, 11, 11, 20 M=8 SD=3.39 M=8.78 SD=4.92
   b) Compare the means/SDs for **A:** 0, 2, 4, 6 8, 10 and **B:** 20, 22, 24, 26, 28, 30 M=5 SD=3.42 M=25 SD=3.42
   c) Compare the medians/IQRs for **A:** 3, 5, 6, 7, 9 and **B:** 3, 5, 7, 8, 9 Med=6 IQR=4 Med=7 IQR=4.5
   d) Compare the medians/IQRs for **A:** 3, 5, 6, 7, 9 and **B:** 3, 5, 6, 7, 20 Med=6 IQR=4 Med=6 IQR=4.5

2. The average on a history exam (scored out of 100 points) was 85, with a standard deviation of 15. Is the distribution of the scores on this exam symmetric? If not, what shape would you expect this distribution to have? Explain your reasoning. SD of 15 is pretty high

3. Facebook data indicate that 50% of Facebook users have 100 or more friends, and that the average friend count of users is 190. What do these findings suggest about the shape of the distribution of number of friends of Facebook users (Backstrom 2011)?

Stat 311 Homework 2

4. This problem uses the same data regarding environmental policy versus economic policy that were presented in Lesson 2, Lecture 1, except the data are categorized by education level or party identity.
   a) Read in GallupByEd.csv and GallupByPI.csv, creating two separate objects that store the data. Convert variables in each object to factors as needed. Think about the order in which the factors will appear on output and adjust if you think reordering will make things clearer. How many observations are there in each data set?
   b) Produce two two-way contingency tables, one for each data set, with education or party ID in rows and the response in the columns.
   c) What is the joint percentage of people who favor environmental policy and who identify as independents?
   d) What is the marginal distribution (as percentages) for education? [Hint: you will be reporting three percentages]
   e) Produce two more tables that show row conditional percentages instead of counts. What is the conditional distribution for Response for those participants who identify as republican?
   f) Pick one of the two data sets and create two bar graphs (your choice of versions) to explore the association between either education or party ID and Response. Make sure the axes are appropriately labeled. Summarize the information you glean from the bar graphs.
   g) For the data set you picked, does the row variable (either education or party ID) appear to be associated with the response. Explain. [Note: this is a qualitative answer based on data visualization.]
5. Complete the following parts using a data set about popular diets (PopularDiets.csv). The data dictionary for the data set is found in the file DietDataDescription.pdf. The journal article that explains the study with results is in the file JournalArticleForDietStudy_joc40214. You will need to browse the journal article to answer parts a) and c).
   a) Was this study an observational study or an experiment? Briefly explain.
   b) What participants were sampled for this study?
   c) What do you believe to be the population of interest? Do you think the results can be generalized to the population of interest or some other population? Explain.
   d) Read in the data. Set variables to factors as needed. How many observations are in the data set? How many of the subjects completed the study?
   e) Explore the weight loss variable. Present summary statistics and two graphs (your choice) that provide insight into the distribution of this variable. Summarize your overall findings for the distribution of weight loss by describing the information you get from the summary statistics and graphs. What numeric statistics do you think are best to use to summarize the distribution? Explain.
   f) Explore how weight loss varies by type of diet. [Hint: use comparative box plots; maybe facet histograms by diet type] Make a qualitative assessment regarding differences in the distribution of weight loss by diet type.