TMATH 390 R Lab 3

Continuing to use R for empirical exploratory data analysis

In this lab we will use R to continue to perform exploratory data analysis of the dataset you obtained in assignment 1. This dataset should contain 2 quantitative and 2 qualitative variables. For this computer lab, use the quantitative variable that you DIDN'T use in lab 2. You will be asked to compare your second quantitative variable to the one you chose in lab 2, so make sure to have those results handy.

C1. (4) Submit the R script you used to complete this computer lab assignment.

Read your dataset into the current R session. For demonstration this lab assumes you name your R data object data.df. Navigate to the directory in which your dataset is located using setwd, then read the dataset into R.

C2. (1) Describe your second quantitative variable, including its units of measurement.

C3. (1) Use the R summary command to produce a summary of your second quantitative variable. Copy and paste the results of the summary command in your R document.

C4. (1) Create a histogram for your second quantitative variable. Make sure that the histogram has a properly labeled axis and title. Copy and paste the histogram into your lab document.

C5. (2) Describe the histogram you produce. What shape does the distribution take? Are there any peculiarities in the distribution? How does it compare to the shape of the distribution for your first quantitative variable that you graphed in lab 2?

C6. (2) Provide the boxplot that compares the distribution of your second quantitative variable across the different levels (categories) of your qualitative variable (using the same qualitative variable as in lab 2).

C7. (2) Describe similarities and differences in the distributions of your second quantitative variable at different levels of your qualitative variable. Also compare this to what you observed in the boxplot of the first quantitative variable that you produced in lab 2.

C8. (2) Compute the sample mean, median, and sample standard deviation for your second quantitative variable for each level of your qualitative variable. **Copy and paste the R outputs that calculate the mean and standard deviation for your data at each level of your qualitative variable.**

C9. (1) We can also use R to compute sample quantiles. Compute the 0.1, 0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9 quantiles of you second quantitative variable. Use the code below to guide you. Report the values for the quantiles in your lab document.

```
# the quantile function returns values for the
# quantiles indicated by probs
# this example uses a data frame called data.df
# and calculates the quantiles for the column called V1
# Here the seq command is used. This produces a sequence
# that starts at the first number, ends at the second
# number, at the increment indicated by the third number
# you can try running the seq command by itself to see the result.
quantile(c=data.df$V1,probs=seq(0.1,0.9,0.1))
```

C10. (1) We can also produce summaries of our qualitative data. The table command returns a tally of the number of observations for each unique value of the variable. Use the table command to report the number of observations for each level of your qualitative variable. Use the code below.

```
# This assumes your data frame is called data.df, and
# the column for your qualitative variable is called V2
table(data.df$V2)
```

C11. (1) We can also create a bar graph of a qualitative variable using the R function barplot. Use the code below and provide a bar graph for your qualitative variable in your lab document.

```
# the barchart command takes the results of the table
# command and creates a barplot from them
barplot(table(data.df$V2),
        xlab="Write in name of your qualitative variable")
# if you want to make a Pareto graph, you can sort your
# table then create a bar graph using that.
# here we create an object for the table
table1.tab=table(data.df$V2)
# the sort function will rearrange this, and the decreasing
# argument tells the function to sort in descending order
table1a.tab=sort(table1.tab,decreasing=TRUE)
# Now create a Pareto graph using the sorted table object
barplot(table1a.tab,
        xlab="Write in name of your qualitative variable")
```

C12. (2) Use the numerical summaries and the plots to summarize your analysis. Use full sentences to write at least a paragraph and make sure to refer directly to your graphs and numerical summaries. In your summary comment on the following:

How are the distributions of your two quantitative variables similar? How do they differ? Consider the shape, variability, and measure(s) of location.

How do your two quantitative variables change at different levels of the qualitative variable? Do they change similarly?

Were your observations evenly dispersed across the levels of your qualitative variable? What do you think explains the distribution of your qualitative variable?