TMATH 390 R Lab 2

Using R to graph data and find quantiles

In this lab we will use R to perform exploratory data analysis of the dataset you obtained in assignment 1. This dataset should contain 2 quantitative and 2 qualitative variables.

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

C2. (2) For this computer lab, choose 1 of your quantitative variables and 1 of your qualitative variables. Describe each variable you choose. For your qualitative variable, name the possible values (labels, categories) the variable can take, and the code that identifies each (for example, if your qualitative variable is Sex, then it might be coded as "M" for male and "F" for female).

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. For example, if my dataset is in a TMATH390 directory in my documents:

```
# change the working directory
setwd("c:/Users/mkenn/Documents/TMATH390")
# use read.csv, assuming your data is a csv file named TMATH390Data.csv:
data.df=read.csv("TMATH390Data.csv")
# you can also use file.choose() to open a window to find your data file
# in your computer system. If you successfully executed the command above,
# this command is not necessary.
data.df=read.csv(file.choose())
# use the "head" function to show the first 6 lines of your dataset:
head(data.df)
```

Quantitative variable

The R summary command produces a summary of a given dataset, depending on the kind of variable that is entered into the command.

C3. (1) Produce a summary of your chosen quantitative variable. Below is example code, assuming that the column of interest is called "V1". Copy and paste the results of the summary command in your R document.

```
# produce a summary of the V1 column in the data.df dataframe
summary(data.df$V1)
```

C4. (2) Describe what the summary command returns for a quantitative variable.

C5. (1) We use histograms to visualize the shape of the distribution for a quantitative variable. Create a histogram for your chosen quantitative variable. Copy and paste the histogram here. Below is an example code to create a histogram.

```
# creates a histogram of the column V1, in the dataframe data.df
# the xlab argument writes text to label the x-axis
# you should substitute the name of your quantitative variable
# and its units into the quotes
# and give your graph a title using the main argument
hist(data.df$V1,xlab="Name of quantitative variable (units)",main="Title")
```

C6. (2) Describe the histogram you produce. What shape does the variable take? Are there any peculiarities in the distribution?

Boxplots are useful to compare distributions. We will compare the distribution of your chosen quantitative variable across the different levels (categories) of your qualitative variable.

C7. (2) Provide the boxplot that compares the distribution of your chosen quantitative variable across the different levels (categories) of your qualitative variable. In the example below, we use the formula syntax to create the boxplot. Also given is an alternative syntax for the boxplot function. This code assumes that the column for your qualitative variable is called V1, and your qualitative variable takes values of either "M" or "F".

```
# first establish a graphing window with 1 rows and 2 columns,
# and las=1 sets axis labels to be horizontal
par(mfrow=c(1,2),las=1)
# Now create a boxplot using the formula syntax. The ~ represents a
# relationship between two variables, with the Y-variable on the
# left-hand side, and the X-variable on the right-hand side. In our case
# we want to compare the distribution of the quantitative variable (V1)
# across levels of the qualitative variable (V2)
boxplot(data.df$V1~data.df$V2)
# Here is the boxplot using an alternative syntax. You should obtain the
# same graph. The line "data.df$V1[data.df$V2=="F"]" tells R to look at the
# V1 column in the data.df dataframe, and the brackets tell R to only look
# at the rows in the data frame that have have a value of V2 = F
boxplot(data.df$V1[data.df$V2=="F"],data.df$V1[data.df$V2=="M"])
# For your boxplot, you would substitute the labels for your own categories.
# if you have > 2 levels, then you would add those to the boxplot command
```

C8. (2) Describe similarities and differences in the distributions of your quantitative variable at different levels of your qualitative variable.

C9. (2) We can also use R to compute individual summary statistics for your quantitative variable (beyond the summary command). Below is example code to compute the sample mean, median, and sample standard deviation for your 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.**

```
# first the mean
mean(data.df$V1[data.df$V2=="F"])
# and the median
median(data.df$V1[data.df$V2=="F"])
# and the standard deviation
sd(data.df$V1[data.df$V2=="F"])
# for your data replace the "F" with the different values your
# qualitative variable can take.
```

C10. (2) Use the numerical summaries and the plots to discuss how your quantitative variable compares among the different levels of your qualitative variable.