Week 1 Assignment: R

This assignment has two parts. Please complete both parts in the same .R file.

**** Always comment your code to make it clear and easy to read.**

PART 1

NOTES: If you find that your dataset is not sufficient to complete the requirements, get a more appropriate one. You must get a dataset that allows you to meet all requirements. If necessary, you can create a dataset or combine datasets to make a good one.

1) Install R Studio

2) Write an R program that does the following:

Find a dataset online (see note above). Your dataset must have at least 5 variables (columns/attributes) and should have at least 40 rows of data. Yes, it can be larger but not smaller. I recommend that you do not use a dataset that is too large as these can take time to process. Your dataset must have at least three quantitative variables and at least one qualitative/categorical variable. It can have more.

You will name the dataset: DATASET1.csv If the data is not in csv format, paste it into Excel and save it as csv. Read the dataset into R into a dataframe.

(a) Create a new feature via binning: Choose an appropriate variable in the dataset and group it into 5 groups (also called bins or categories). Create and add a new variable to your dataframe that represents this new categorical variable/feature that you created. This is called feature creation via binning. Once done, your dataframe will have an extra variable that is the result of the categorization of one of the quantitative variables.

Example: Here – the NumGrade variable is quantitative. A new variable or feature called LetterBin is created and each NumGrade is categorized into that new feature. There are many ways to do this and categories can also be numbers, such as 1 through 5 for example. Once the new feature is added, write the head of the dataframe to the outfile called MYOUTFILE.txt.

NumGrade	LetterBin
89.5	B
78.2	C
67	D
92	A
71.5	C
88.5	B

Next, create a second new feature that is binary (0 or 1). Use one of your quantitative variables to create this new binary feature.

Example: Notice that Over40 is a binary feature based on Age.

Age	Over40
32	0
70	1
12	0
67	1
32	0
45	1
50	1

At this point, you have created two new features. One is categorical and has 5 categories. The second is binary and has two categories.

(b) Create four functions. The first will calculate ANOVA, the second will perform an independent samples t test, the third will perform a basic z test, and the fourth will calculate summary statistics for a given variable. You determine the parameters needed to be sent to the functions. Choose appropriate variables from your dataset and apply each of the four functions. For example, you can apply the summary statistics function to any quantitative variable. To apply your z test function, you will need to choose a mean and variance and must choose a quantitative variable as your sample, etc…

Write (append) all results to a new file called MYOUTFILE.txt.

The MYOUTFILE.txt should contain the output that occurs when you call/apply your 4 functions appropriately.

(c) Choose appropriate variables from your dataset and create the following graphs:

- Boxplots for at least three variables on the same plot
- Histogram with normal curve
- Scatterplot between two variables
- Multiple bar graph

Copy and paste all graphs to a Word doc.

All Graphs MUST be properly labeled with title, and x and y axis labels. Graphs must use color.

PART 2

Part 2 of this assignment will help you to practice with cleaning data, pre-processing data, and performing initial EDA. Once you write this code, keep it in a location where you can find it later. You will certainly come back to this code during your MS at GU. In addition, **comment your code very well** and clearly. This will help you when you come back to it later.

You can find assistance for this in my shared R-code collection:

https://drive.google.com/drive/folders/1rXm4jTHMTTjFvHfJ3daCCWNmOdF9tNPl?usp=sharing

1) You will use the Kaggle Titanic Data for this assignment.
   Use this copy so that everyone uses the same data:
   https://drive.google.com/file/d/1iAGHV19PM8c92wWxAjDG9jDM_8YGt-iO/view?usp=sharing
2) Next, perform the following cleaning and prep on the dataset:
   a. Read in the Training dataset that I have given you above. Print the head of the data - specifically the first 10 rows.
   b. Next, print the structure of the dataset (using str in R)
   c. Next, change all char types to factors.
   d. Next, create a frequency table (using table) for all variables so that you can see what values are there and if any values look incorrect or odd.
   e. Next, for each variable, check is there are NA values (is.na) and also add up (sum) the NA values to present (print out) a total.
   f. Count the number of complete rows and print this. (Use complete.cases)
   g. Next, think of how to best clean the data. For example, if a given column/variable has many missing values AND does not seem useful in analysis, note this in the comments and remove that column. Otherwise, if a column is useful, such as Age, but also have incorrect and NA values (which is does) fix them and explain how you fixed them. For example, you might replace missing or incorrect ages with the median. Be sure to evaluate each column and either correct or remove it. Explain EVERYTHING that you do and why it is OK to do.
   h. Create a new column in the dataframe called BinnedAge. Use cut to discreize (bin or categorize) the Age values into 4 groups: 0 – 15, 16 – 29, 30 – 45, 46 –inf. Then, create a table that shows the counts for each group.
   i. As a last step, perform some exploratory data analysis (visual and statistical) on the dataset. I suggest boxplots, other plots, summary stats, etc.

Deliverables:

In ONE zip folder (NO TAR!) – include all of the following:

1) Your datasets (as .csv)

2) The out file you created: MYOUTFILE.txt for Part 1.

3) The Word Doc with the graphs for Part 1.

4) Your R code – which must fully run and be able to read in your dataset.

5) Suggested: Try also to create and save this as RMarkdown .html.