**EBIS 3103 Introduction to Business Data Analytics -**

**Individual Assignment 1**

Instructor: Dr. Jingzhi Zhang

This assignment is designed to walk you through data manipulation and basic visualization techniques for given research questions. You will be also asked to answer several questions about data distribution using the real-world donorchoose.org data.

Now, suppose we are interested in projects sizes on the donorschoose.org platform. More in particular, we are interested in (a) how a teacher changes the project size as he/she repeatedly creates projects, and (b) how project sizes differ depending on the number of projects a teacher has created historically. Note the difference between the two questions. Understanding the difference is the key to this assignment.

Here are the steps you should follow:

1. Load projects.csv data. [5 marks]

2. Before analyzing the data, you need to create new variables that are needed for the analysis.

(1) Change the data type of the \*date\_posted\* column into date. (Hint use “lubridate” library). [5 marks]

(2) Create a new column called “project\_order” that shows how many projects a teacher has created including the given project. For example, the first project created by teacher A should have 1 as the value of this column, his/her second project should have 2, etc. Naturally, this column starts with 1 for every teacher. [5 marks]

(3) Recode all values of \*project\_order\* larger than 5 to 6. In other words, after the manipulation, the value of 6 in the \*project\_order\* column should mean 6 or larger numbers. [5 marks]

*Tips:*

* *for the step (3), you may use ifelse() in mutate () to recode the values.*

3. Graph the density of project sizes only for project\_order==1 and project\_order==6. As a variable for project sizes, use \*total\_price\_excluding\_optoinal\_support\*. If you use ggplot2 library, you can use the \*geom\_density\* function.

- Do you think they are close to normal density?

- Interpret any difference you notice from the two densities.

[A correct visualization: 10 marks / Interpretation: 10 marks]

*Tips:*

*- xlim(0,2000) can be used to set a limit of x-axis, up to 2000 for visualization. This action does not remove large values in the variable.*

4. Demonstrate that Chebychev’s inequality holds for the distribution of project sizes. [10 marks]

Tips:

* For k=2: at least 75% of the data lie within two standard deviations of the mean
* For k=3: at least 89% of the data lie within three standard deviations of the mean

5. Now using the \*projects\* table, it is time to create a new data set that you will name “teachers.” This table should have two columns for *each* teacher.

(a) \*avg\_project\*: the average of project sizes from each teacher.

(b) \*last\*: how many projects each teacher has created in the dataset.

As you did in step 2, recode all values of \*last\* column large than 5 to 6. In other words, after the manipulation, the value of 6 in the \*last\* column should mean 6 or larger numbers. [20 marks]

6. Using the new “teachers” table, graph the densities of average project size only for last==1 and last==6.

- Do you think they are close to normal density?

- Interpret any difference you notice from the two densities.

[A correct visualization: 10 marks / Interpretation: 10 marks]

*Tips for you:*

*- %in% can be used for matching values and returns a vector of the positions of matches of its first argument in its second.*

*- xlim(0,2000) can be used to set a limit of x-axis, up to 2000 for visualization. This action does not remove large values in the variable.*

7. Now, interpret the differences between outputs of step 3 and outputs of step 6. What would you conclude in relation to the given research interest? [10 marks]

Instructions:

1. Use R.

You do not have to use *tidyverse*, but it is recommended. If you are to use tidyverse, here’s a useful online resource that has most of what you need to finish this assignment:

https://r4ds.had.co.nz/transform.html

2. Use Rstudio and Rmarkdown, and compile your codes and explanation into an HTML or a PDF file. You should submit only one final compiled report file (other formats will **NOT** be graded).

3. Do not include more than twenty lines of output of your code. For example, if you want to show that your code successfully transforms the entire table, only shows, say, the first ten rows of the table.

4. Do not create a new data frame unless you are instructed to do so. When you create a new column, use the instructed column name. If you have to make your own, you need to justify.