CSCI N317 Spring 2018 Lab 5

**R Data Processing**

Find a large data set that is of your interest. They are plenty of public datasets available on the web. You can find a set of links in the file PublicDataSets.docx under Files -> SupplementaryMaterials.

Create an R program to perform a task or a set of tasks of your choice using this data set. Your program must meet the following requirements.

1. Have comments that
   1. describes the task(s) (5’)
   2. states the author of the program and the date that the program was written or completed (5’)
   3. describes the purpose for each code block (5’)
2. Import the dataset into R, from a local file on your computer or from a web link. Your data set must have more than 3000 rows. (15’)
3. Create user defined function(s) to process data. (10’)
4. New column(s) must be generated using original data and be added to the original dataset, e.g. if your data has a Price column and a Quantity column, you can create another column called Income whose values are generated through price \* quantity. Your new column(s) doesn’t have to have values in every cell. (20’)
5. Your task(s) must make sense. (10’)
6. Program runs and generates correct data in the new column(s). (10’)
7. Export the new data set into the original data format. E.g. if your original data set is in a file called myData.xlsx, your new data file that contains original data plus the new column(s) will still be saved as .xlsx file, e.g. newData.xlsx. (10’)
8. Zip up the folder with your script files, the original data file and the exported data file. Your code must be self-contained, i.e. the folder can be placed on any computer and run successfully. Name the zip file as lastname\_lab5.zip and submit this zip file to Canvas. (10’)