PROG8430 – Data Analysis, Modeling and Algorithms
Assignment #1: Exploratory Data Analysis with 'R' (10%)

## Submission and Grading Guidelines:

- All assignments must be submitted via the eConestoga course website before the due date into the assignment folder.
- You may make multiple submissions, but only the most current submission will be graded.
- Submissions:
    - **Do not use a Zip file to submit your assignment!**
    - Create a single text file (.txt) that includes all the R scripts you have written for the assignment. Copy and paste the contents of each script into this file.
    - Create a single PDF file (.pdf) that includes:
        - All your code and your answers
        - If you use R Markdown to create your report, you can directly convert the .rmd file to a .pdf file for submission.
- The marks on the assignment are generally awarded 50% for the actual R code and calculations and 50% for interpretation and demonstration that you understand what you have done.
- This is an individual assignment. Unauthorized collaboration is an academic offense. Your report and code files will be tested by Turnitin.
- All delay submission will be deducted 20% per day. **No excuse is accepted.**

## Assignment Tasks (30 points, 10% of final grade):

**PART 0** – Professionalism and Clarity (2 points) ;
Follow the submission instruction (2 points)

**PART 1 – Written Answers**

1. (4 points) You are working Streaming Service. The following statement is made by your manager. Based on the examples and discussion in Lecture 1, transform it into a question that can be answered with data analytics. Make sure you discuss the logic and reasoning you use to transform it and what problems you might define. (NOTE – This question is worth 4 marks, so answer appropriately). ***We have more customers than before, but our new customers are streaming less than before.***

2. (4 points) Consider the following three arrays of data. Each array is data for one file sharing site. The numbers in the array represent the number of downloads for each site in a day (for example, Site A had 28 downloads on the first day, 29 on the second and so on).
   Site A: (28 29 31 28 30 30 30 32 28 33)
   Site B: (23 19 23 33 32 27 20 24 42 32)
   Site C: (27 26 28 25 27 27 30 30 28 26)

   Based on the data provided, and using the skills learned in this class, answer the following questions. Make sure to provide *evidence* for your answers.

   a) (2 points) Which site has the least downloads on a typical day?
   b) (2 points) Which site has the most inconsistent usage?

**PART 2 (18 points) Please use the Dataset "PROG8430-23W-Assign01.txt" to answer following questions.** In this case, imagine you are working for a company that maintains three different data centres with multiple servers in each. The following tasks will seek to describe and explore some of the data which has been gathered by the company.
All variables as well as the descriptions are summarized in below table.

| Variable | Description |
|---|---|
| Manufacturer | Name of the Manufacturer of the Server |
| Server | Server Model Number |
| DC | Data Centre Name |
| SMBR | Server Message Blocks Received |
| SMBT | Server Message Blocks Transmitted |
| Conn | Connections Made |

| | | |
|---|---|---|
| 1 | Basic Manipulation<br>1. Read in the text file and change to a data frame<br>2. Append your initials to *all* variables in the data frame (Note – you will need to do this in *all* your subsequent assignments).<br>3. Change each character variable to a factor variable<br>4. What are the dimensions of the dataset (rows and columns)? | 4 points |
| 2 | Summarizing Data<br>1. Means and Standard Deviations<br><br>a. Calculate the mean and standard deviation for *Server Message Blocks Received*.<br>b. Use the results above to calculate the coefficient of variation (rounded to 3 decimal places).<br>c. Calculate the mean and standard deviation for *Server Message Blocks Transmitted*.<br>d. Also calculate the coefficient of variation (rounded to 3 decimal places).<br>e. Does the SMBT or SMBR have more variation?<br><br>2. Calculate the 45th percentile of the number of Server Message Blocks Transmitted. This calculation should be rounded to the nearest whole number (no decimal places). | 5 points<br><br>1 point |
| 3 | Organizing Data<br>1 Summary Table<br>a) Create a table showing the average *Server Message Blocks Transmitted* by *Manufacturer*. This should be rounded to two decimal places.<br>b) Which Manufacturer's Servers have, on average, transmitted the most server message blocks?<br><br>2. Cross Tabulation<br>a) Create a table counting all *Servers* by *Data Centre*.<br>b) Change the table to show the percentage of each *Server* in each *Data Centre* . This should be rounded to three decimal places.<br>c) What percentage of servers at Elmira are MG9696? | 2 points<br><br>3 points |

| | | |
|---|---|---|
| | 3. Histogram<br>a) Create a histogram of *Server Message Blocks Transmitted.*<br>b) The plot should be properly labelled and a unique colour and have 10 breaks.<br>c) Which range of SMBT is the most common? | 3 points |