**CS 4331 – Data Mining**

Spring 2020

**Assignment 1 - 8 Points**

Feb 3, 2020

**Full Name: Dipendra Yadav Class Section: CS 4331-001**

Acknowledge your collaborators or source of solutions, if any. **Submission by 2/10/2020 is required.** This exercise is for getting familiar with *Data Preparation and Exploration*. For coding, use Python 3 or R. A subset of your answers will be graded. Include your .py or .r file with helpful comments included and no hard-coded file paths (only the file name should be present when reading or writing the file). All files should be submitted in a zip file and named CS4331\_00?\_eraiderlogin\_Vn.zip where ? is replaced with your class section number, eraiderlogin is your eraiderlogin, and n is replaced with the version. You may submit up to 3 times and only the last version will be graded.

1. Identify each column as nominal, ordinal, interval, or ratio in the Auto\_mpg\_raw.csv data set.
   1. Miles per gallon: Ratio
   2. Cylinders:
   3. Displacement:
   4. Horsepower:
   5. Weight:
   6. Acceleration:
   7. Model Year:
   8. Origin:
   9. Car Name:
2. Using the Auto\_mpg\_raw.csv file, create a histogram for each column except for the “Car name” column. Describe any misleading data points that you see from the histograms. Remove the misleading data points from the dataset, generate new histograms, and export as Auto\_mpg\_adjust.csv.
3. Use z-values to identify any outliers (if any) in the numeric data in Auto\_mpt\_adjust.csv and print them out.
4. Using the cereals.CSV data set,
   1. create a bar graph and normalized bar graph of the “Manuf” variable with “Type” overlay
   2. create a contingency table of “Manuf” and “Type”
   3. create a histogram and normalized histogram of “Calories” with “Manuf” overlay
   4. bin the “Calories” variable using bins for 0 to 90, 91-110, and over 110 calories and create a bar chart of the binned calories variable with “Manuf” overlay

**Source:** Chantal D. Larose, Daniel T. Larose, Data Science Using Python and R

*All assignments will be checked for academic misconduct (cheating, plagiarism, collusion, falsifying academic records, misrepresenting facts, violations of published professional ethics/standards, and any act or attempted act designed to give unfair academic advantage to oneself or another student) defined by “OP 34.12: Grading Procedures, Including Academic Integrity” of TTU.*