COSC 757 Data Mining Assignment 1

Instructions: This is an individual assignment. Use Blackboard to submit your document on the due date (no hard copies please). Late submissions will receive a zero grade.

Select a dataset from the UCI Machine Learning Repository that are classified for the task of Regression. Datasets can be found at the following link:
https://archive.ics.uci.edu/datasets

Exploratory data analysis:
Explore the dataset in R using visualization and descriptive statistics. You can use the functions provided in Chapter 3 of the text as an example. Write a brief report showing your exploratory data analysis. You should at least show descriptive statistics for the data including visualizations of the distribution of the attributes, relationships between attributes.

Data Preprocessing:

1) Some of the methods for data reduction require the data to be normalized (i.e. rescaling data measured in differing units. Use R to normalize the numerical attributes of this dataset using min-max normalization, z-scores, and decimal scaling
2) Select a continuous variable and experiment with at least two methods to bin the variable into discrete categories.
3) Find a variable that does not have a normal distribution based on your exploratory data analysis. Use the natural log, square root, and inverse square root transformations to make attempts to achieve normality. Report on your results

Regression Analysis:
Based on your exploratory data analysis of the dataset, come up with a prediction question and create a regression model to predict a dependent variable based on a set of dependent variables. For best results, make sure that the variables that you choose are numeric. If you insist on using a categorical variable, they will have to be converted to numeric variables.

The deliverable for this project will be a report that details your experiments. The report should be in either **ACM or IEEE conference paper format.** The document should have the following three sections: 1) Introduction; 2) Dataset Description; 3) Exploratory Data Analysis; 4) Data Preprocessing; 5) Regression Analysis; 6) Conclusion. Sections 3, 4, and 5 should provide details about the methods that you used to accomplish each task as well as the results that were produced by your experimentation. The conclusion should contain a discussion and interpretation of the result.

Links to format templates:
http://www.ieee.org/conferences_events/conferences/publishing/templates.html
https://www.acm.org/publications/proceedings-template