CS 699
Assignment 1

Note: R is recommended since that is what is taught in class, but you may use any software you choose. As explained in submission guidelines on the next page, please provide your "final answers" in a document and provide a separate file containing your code or other detail that explains how you obtained your answers.

Problem 1 (70 points). Use the dataset *autism-adult.csv*. This data set has 704 observations and 19 features.

(1). Calculate the mean, median, and standard deviation (sample) of the *age* feature.
(2). Determine Q1, Q2, and Q3 of *age*.
(3). Plot the boxplot of the *age* feature.
(4). Implement min-max rescaling on the *age* feature. Replace the original *age* feature with the rescaled result. Provide the rescaled age for the seventh observation in the data.
(5). Determine the mode of the *country_of_res* feature.
(6). Review the *ethnicity* feature. You will notice several missing values in this feature. Determine a reasonable imputation for this feature. Explain what you are going to do and why. Then replace the original *ethnicity* feature with the imputed result.
(7). Create a bar graph of your imputed *ethnicity* feature.
(8). Implement dummy coding for the *gender* feature. Replace the original *gender* feature with the coded result. Provide the coded gender for the last ten observations in the data.
(9.) Identify which features in your data set are discrete and which are continuous.
(10.) Identify which features in your data set are numeric and which are non-numeric. Compare with the discrete/continuous classification you just made and discuss the similarities and/or differences you see.
(11.) After completing all requested tasks above, print the first 4 observations of the data.

Problem 2 (30 points). Use the dataset *correlation.csv*. This data set has 100 observations and 5 features.

(1). Create a scatterplot of feature *A1* vs. feature *A5*.
(2). Compute the correlation matrix for all five features in the data set.
(3). Identify the strongest correlation in the data set. Which factors are involved? Is it a positive correlation or a negative correlation?
(4). Implement z-score normalization on all features in the data set.
(5). Compute the correlation matrix for all five normalized features in the data set. Compare this correlation matrix with the matrix you obtained earlier and discuss the similarities and/or differences you see.

Submission Guidelines:

- **Submit the solutions in a single Word or PDF document and upload it to Blackboard. You should separately upload the code or another document that explains how you arrived at the answers. Make sure each filename includes your name.**
  - **If you need more than 3-4 separate files, you might ZIP/RAR these files. Please still upload the Word/PDF separately. The separate Word/PDF greatly simplifies the facilitator's ability to comment on your assignment.**
- **Your submission should be organized well and easy to follow. Clearly list which question you are answering and provide the answer requested (one or more numbers, a screenshot, a plot, an explanation, whatever the question asks you to do). If a question contains multiple parts, you should clearly provide your answer for each part.**
- **If your facilitator tells you to submit the files differently than the above guidelines, you are expected to respect your facilitator's wishes starting on the next assignment.**
- **Facilitators can deduct up to 20% if you fail to follow these requirements (more if the questions are not actually answered).**
- **Facilitators can deduct 5% for each day the assignment is late.**
- **Unless your facilitator or the professor agrees, your assignment will not be graded if it is more than 3 days late (e.g., no credit will be given after Friday at 6 AM Boston time). The professor will usually ask the facilitator to make the decision but in rare cases (<1% of the time) has overridden a facilitator. Do not expect the professor to override in most cases.**