CPSC 4300/6300: Assignment 1

Instructions

- Please submit your solution by 11:59PM September 15.
- Submissions should be made through **Canvas**.
- Please complete homework individually.
- Please submit a **single PDF/Word file** including your solutions and the code of your solutions. Feel free to use any third-party libraries.

Nutrition fact for McDonald's Menu data: Please download menu.csv from the following link: https://www.kaggle.com/mcdonalds/nutrition-facts

1. **Data Exploration (40 points, 8 points per sub question):**

   (a) Identity the type of the features.

   (b) Plot the histogram of Calories. What type of distribution it is?

   (c) Plot the correlation heatmap between features and Calories. You may notice that the diagonal elements are always 1. Explain the reason.

   (d) List the features which have the second and third largest positive correlation with Calories. Note: if you encounter multiple features actually mean the same thing, only list the feature with the largest correlation. For instance, if you see Sodium and Sodium (%Daily Value) and the former has larger correlation with Calories, list Sodium only.

   (e) Report all features which have negative correlation with Calories. Does your result meet your expectation?

2. **Plotting (30 points, 15 points per sub question):**

   (a) Plot the scatter plot for 'features vs. Calories' for all features found in 1(d).

   (b) Plot the box plot for all features found in 1(d) and 1(e) correspondingly.

3. **Data Pre-processing: missing values (30 points, 15 points per sub question):**

   (a) Report the median and standard deviation for all numerical features.

   (b) Write the code to replace the missing values with mean values. Report the median and standard deviation. Compare your result with (a) and write one sentence to explain your discovery.