Project: Capstone Project 1: Data Wrangling

1. What kind of cleaning steps did you perform?
    (1) Since I'm going to analyse the effect of different variables to the scores of different subjects, I used .dropna() to remove all the missing values.
    (2) Used .boxplot() function to check if there are significant outliers of the dataset
    (3) There were some outliers. So I removed all the outliers beyond 3 sigmas range.
    (4) Quantified categorical columns by using one hot encoding and quantified those can be ranked directly such as education level of parents
    (5) Dropped the score columns and made a propensity score matching of the control and test groups according to the variables. (Groups were classified based on the 2 types of lunch)
2. How did you deal with missing values, if any?

    I dealt with the missing values by using .dropna() and found out there wasn't missing value in the dataset.

3. Were there outliers, and how did you handle them?

    I used boxplot to check out and found that there were some outliers. I removed the outliers beyond the range of 3 sigmas for math_score, reading_score and writing_score respectively. The rest of the columns were all categorical variables.