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.

Project: Capstone Project 1: Exploratory Data Analysis

- Are there variables that are particularly significant in terms of explaining the answer to your project question?

  In this project, I used standard lunch and reduced lunch as 2 types of lunch to classify the dataset into 2 groups, control and test respectively. Besides, I used total score of the 3 subjects as the dependent variable to study the relation between lunch types and students' academic performance. In this case, the other variables are considered covariates, and propensity score matching is used to minimize the influence from the covariates such as gender, parental education level. The purpose of this project is to eliminate the selecting bias while doing control test.

  We can see the correlation r-value between lunch type and total score is -0.216, which means when lunch value is 1(test group: reduced lunch), students get lower score, and when lunch type is 0, students get higher score. we want to compare the difference between means of total score of these two groups. The means and standards deviation can both be obtained from our total_score columns. Thus, hypothesis test on difference of means can be a very effective way and t test is chosen in our case.

  (1) From scipy.stats import ttest_ind. (2) Calculate means and standard deviations of test and control groups. (3) Set lunch by assigning miu1, sigma1 and size 1000 to np.random.normal respectively. Set no_lunch by filling in np.random.normal with miu2, sigma2, 1000. (4) Do t test and obtain p-value. (5) Compare p-value with significant level 5%.

After doing t test based on the mean difference between the total score of control and test group. We rejected null hypothesis in favor of alternative hypothesis that the control group with standard lunch have higher mean of total score. This result shows that standard lunch can help students perform better in academic field.