

## Project: Capstone Project 1: Exploratory Data Analysis

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.

Before doing inferential statistics to figure out what the outcome tells us, data wrangling and propensity score matching of the test and control groups are needed. In data wrangling, (1) Use `df.dropna()` to remove the missing values in the DataFrame `df`. (2) Remove the rows in `df` which contain value beyond the range of 3 standard deviation of either one of the 3 score columns. (3) For the columns whose values can be ranked by number directly, assign consecutive integer to each type of value based on the level from low to high. (4) For those columns that cannot be quantify directly, use one hot encoding and delete one column to avoid dummy variable trap afterwards. (5) Import `Matcher` and classify test and control group by lunch type free/reduced and standard respectively. (6) Set the 3 types of score as dependent variables, lunch types as independent variables and others as covariates. (7) Conduct PSM and we can obtain the pairs of matched test and control groups for inferential statistics.

Calculate correlation by importing `linregress` from `scipy.stats` and 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: free/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 an 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  $\mu_2$ ,  $\sigma_2$ , 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.