Capstone I – Milestone Report

1. Problem Statement: Why it is a useful question to answer and for whom.

In the paper "Obesity and unhealthy lifestyle associated with poor executive function among Malaysian adolescent"[1], it was found that high BMI scores are associated with low executive function. My focus in this project is to determine which other lifestyle factors are predictive of a low executive function score.

Executive function is a cluster of cognitive processes that underlie planning, organizing, and regulation in humans, and is responsible for the achievement of purposeful, goal-directed behaviors. In the literature, executive function is broken into three scores: inhibition (the ability to resist distraction and maintain focus), working memory (the ability to store, maintain, and manipulate information over a brief period of time), and cognitive flexibility (the ability to shift attention, select information, and alter response strategies in response to changing task demands).

Proper executive function is a very important factor in leading a healthy and balanced life through adolescence and adulthood. Problems with functioning can lead to detrimental effects – from not being able to meet the demands of a workforce successfully to struggling with mental health issues.

The problem is as follows: finding major lifestyle trends in the data set related to executive function, identifying potential factors related to low executive function, and using these factors to build a predictive model. The first step in the analysis of the dataset is to confirm whether the assumption regarding BMI scores is true, followed by an evaluation of lifestyle trends.

The prefrontal cortex develops during adolescence, which is the prime time to adopt healthy habits which will be maintained in adulthood. The state or local governments in Malaysia can use this data to decide whether actions should be taken to regulate controllable factors such as physical activity levels and proper lunch options at school.

This study is a small one, but its effects hold implications on a global scale. While this study took place in Malaysia, if a similar study could be replicated in the United States for example, it could help determine what lifestyle modifications could be implemented nationally to improve executive function in students and tackle childhood obesity.

2. Description of the dataset, how you obtained, cleaned, and wrangled it

The dataset was found in the Public Library of Science repository, a non-profit science and technology publisher with a library of open-access journals and literature. The dataset was available as an Excel file with 513 subjects from two separate high schools in Selangor, Malaysia. Any adolescents who suffered from neurological or psychiatric disorders, had medical conditions, learning disabilities, history of traumatic brain injuries, or had difficulties in performing physical activities were excluded from the study.

The students – aged 12 through 16 – provided answers to prompts regarding age, date of birth, sex, ethnicity, meal consumption patterns, physical activity, and sleep quality through self-administered questionnaires. The subjects' parents were given surveys to answer demographic information regarding household size, income category, father's years of education, and mother's years of education. Height and weight measurements were taken twice for each subject on-site at the time of experiment, with the mean value of each measurement used as the final measurement.

Three separate tests were conducted to measure executive function:

1. **Interference - (Stroop Color-Word Test)**
    a. Higher score = better capacity for inhibition responses, i.e. the ability to resist distraction and maintain focus

2. **Working Memory - (Digit Span Test)**
    a. Higher score = better working memory performance, i.e. the ability to store, maintain, and manipulate information over a brief period of time

3. **Cognitive Flexibility - (Trail-Making Test)**
    a. Lower score = better performance, i.e. the ability to shift attention, select information, and alter response strategies in response to changing task demands

Any duplicate data were dropped. Twenty-seven values from four separate columns were missing. The missing values were dropped and 486 observations – out of 513 - were retained for further exploration. No outliers were apparent.

3. Initial findings from exploratory analysis

An initial exploration of the dataset indicated that only 17 students were 14 years old (3.5%) and 33 were 16 years old (6.8%). The spread of 12, 13, and 15-year-olds was relatively even. There were also 102 more responses from female students than male students.

Visual EDA followed and box plots were created for the categorical variables in the dataset which included sex, physical fitness category, physical activity category, global sleep category, income category, and household size. The results were as follows:

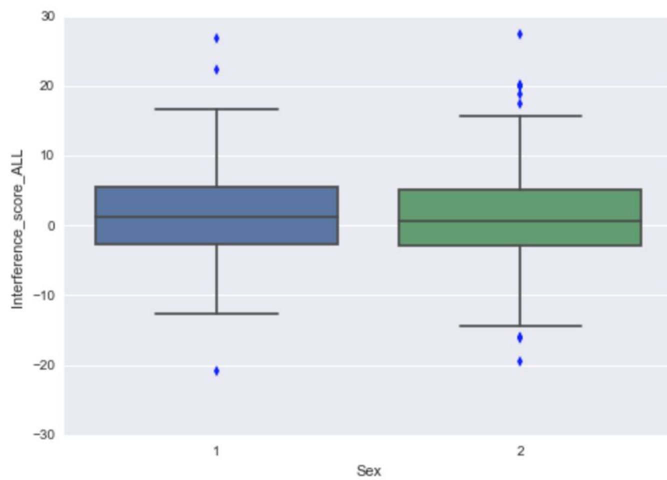

Figure 1.1. Sex vs. Interference. 1 = male; 2 = female.

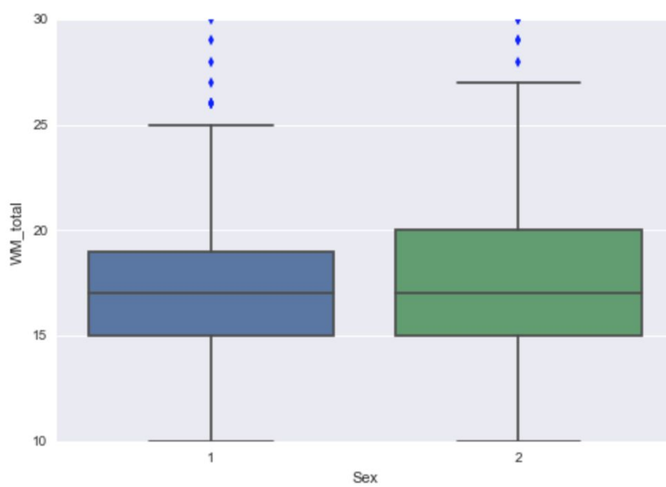

Figure 1.2. Sex vs. Working Memory. 1 = male; 2 = female.

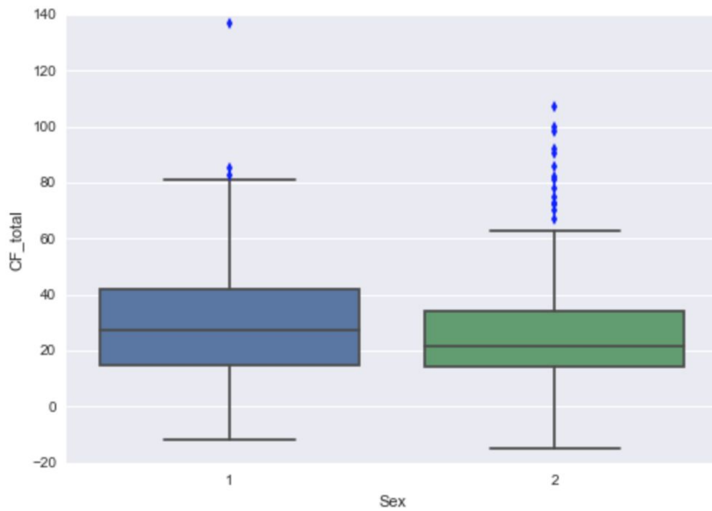

Figure 1.3. Sex vs. Cognitive Flexibility. 1 = male; 2 = female.

The interference distribution is very similar for males and females (Fig. 1.1). There is slightly more variation in the distribution for working memory, while it is clear that there is a significant difference between males and females when it comes to cognitive flexibility (Fig. 1.2, 1.3).

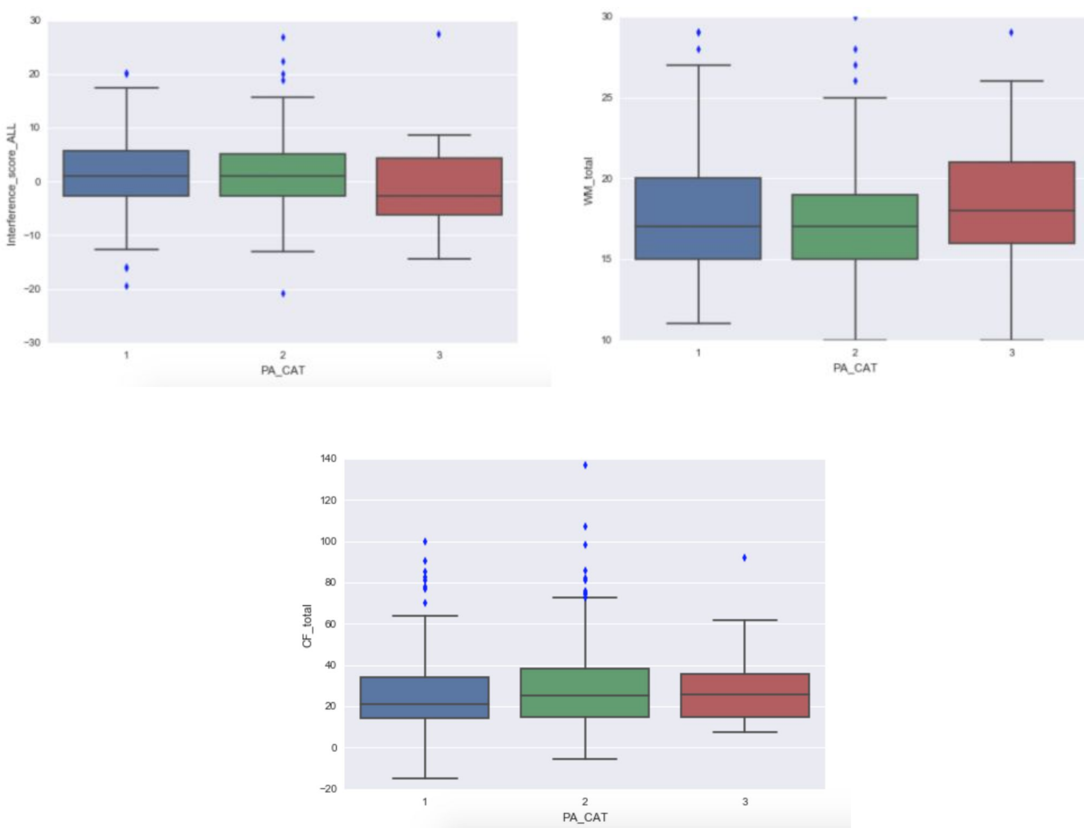

Figure 2. Physical Activity Category vs. Executive Function. 1 = low; 2 = moderate; 3 = high.

Similarly, the executive function distributions are similar for all physical fitness score categories, though the differences in physical activity distributions suggests there are differences between groups (Figure 2).

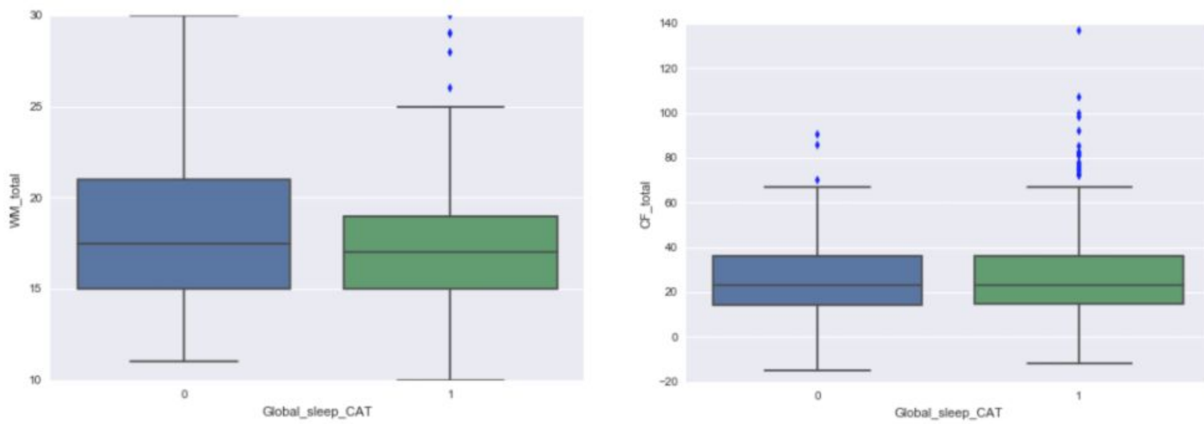

Figure 3. Global Sleep Category vs. Executive Functions (Working Memory and Cognitive Flexibility). 0 = poor sleep quality; 1 = good sleep quality.

Global sleep category is divided into two groups: poor sleep quality and good sleep quality. The working memory median is the same for both sleep category groups, but there is a large difference in their distributions (Fig. 3). There is no variation for cognitive flexibility.

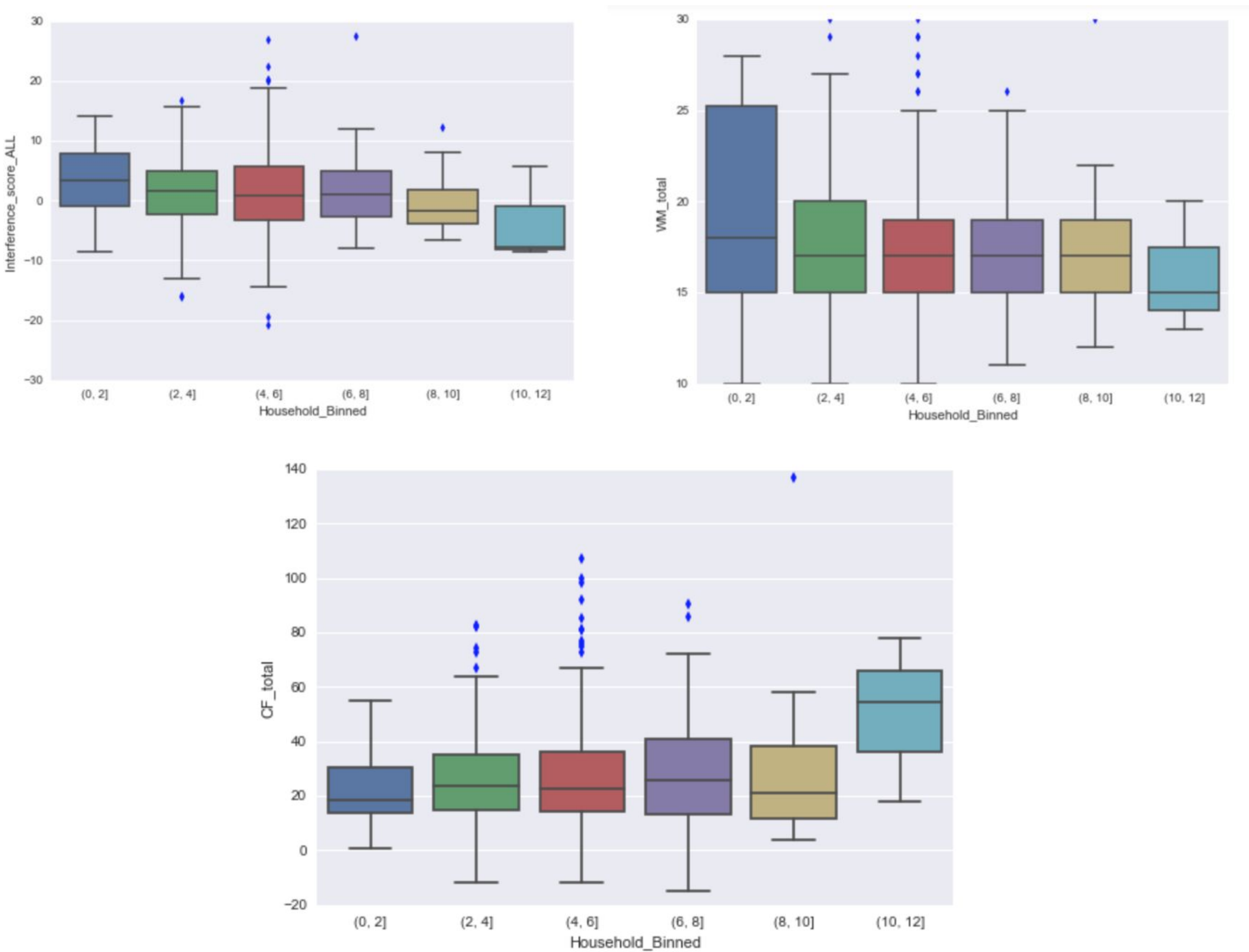

Figure 4. Household Size vs. Executive Functions.

Household size shows large discrepancies for all executive functions (Fig. 4).

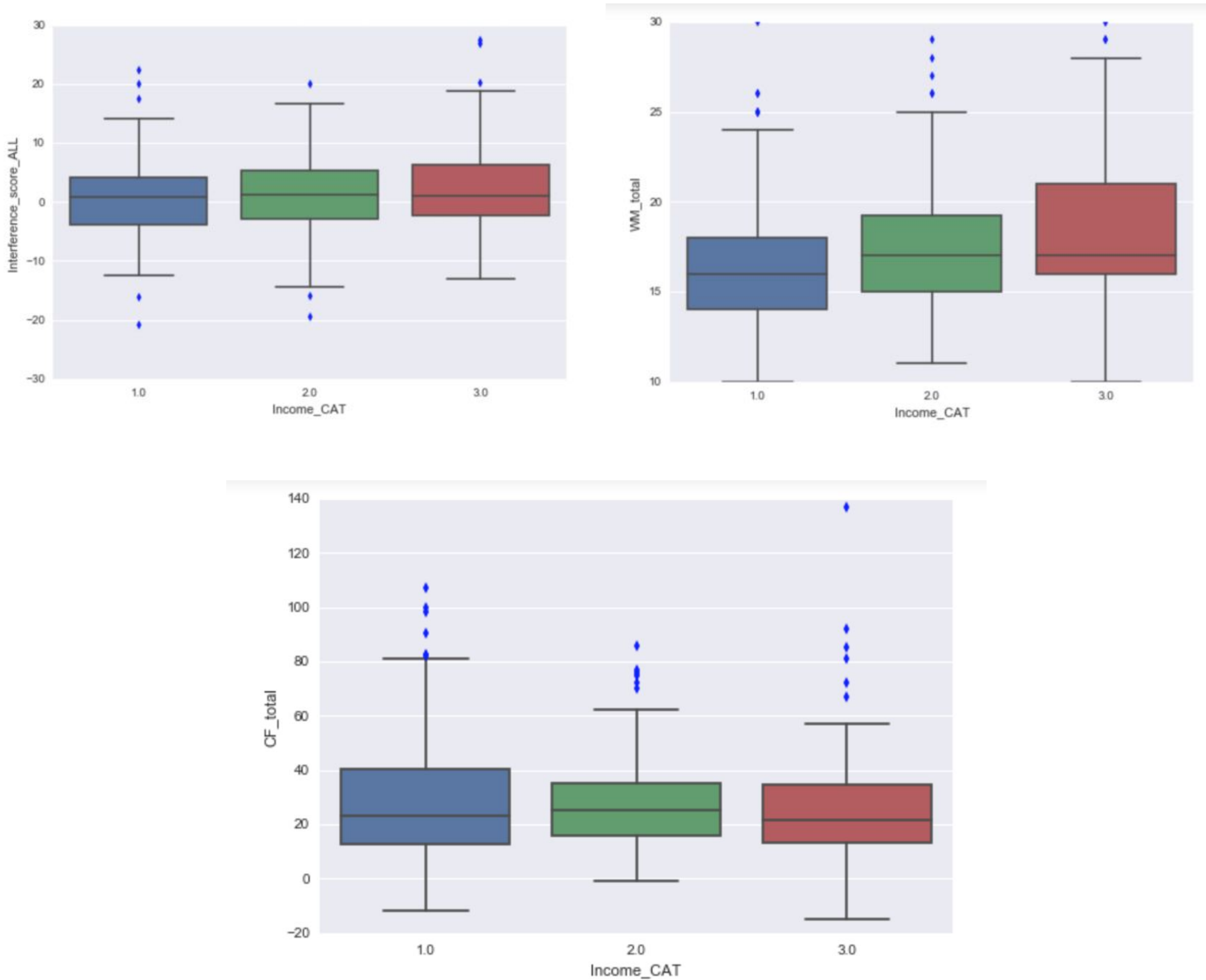

Figure 5. Monthly Household Income Category vs. Executive Function.
1 (RM < 2299); 2 (RM = 2300 - 5599); 3 (RM > 5600)

Monthly household income category is split into three groups: RM < 2299, RM = 2300 - 5599, and 3) RM > 5600, where RM stands for the Malaysian Ringgit, the country's official currency.

For reference, the annual salary for each group (in dollars) is as follows:

- Group 1 (RM < 2299) = USD < $6,760.00
- Group 2 (RM = 2300 - 5599) = USD = $6,763.00 - $16,462.00
- Group 3 (RM > 5600) = USD > $16,465.00

Income category shows a stable distribution for interference and cognitive flexibility, but displays a strong discrepancy when it comes to working memory (Fig. 5).

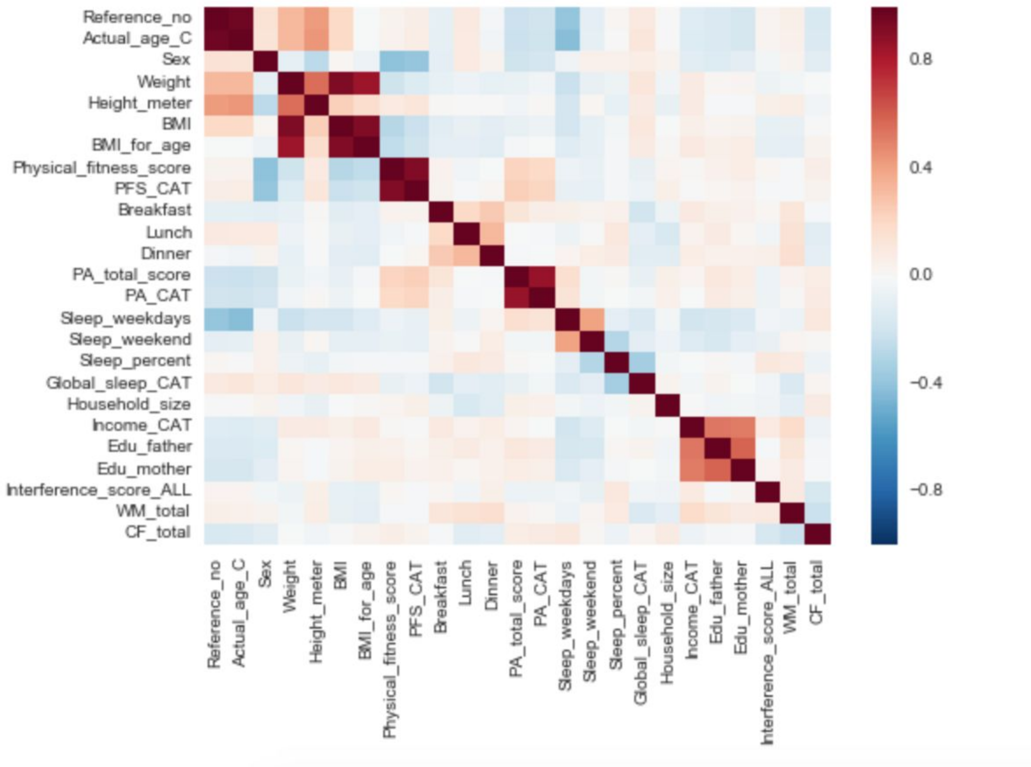

Figure 6. Standard correlation matrix for all variables.

A standard correlation matrix was graphed to get a better sense of the data, followed by statistical analysis to determine whether any perceived correlations were statistically significant (Fig. 6).

Two-sample t-tests were conducted for two categorical variables - sex and global sleep category. The p-value of 0.010 was calculated for sex with regard to cognitive flexibility (CF), leading to the rejection of the null hypothesis at the 0.05 significance level, signifying that mean CF scores are statistically different for males and females. A Bonferroni correction was performed on the sample, which gave a corrected significance value of 0.03, and reiterated the need to reject the null hypothesis in favor of the alternate.

The null hypothesis was also rejected for global sleep category with working memory (p-value = 0.0007; Bonferroni $\alpha = 0.0021$), suggesting that those who fall into the "good sleep quality" category exhibit statistically significant working memory scores than those who fall into the "poor sleep quality" category.

A different test was required for the remaining categorical variables, all of which had three or more categories within their respective variables. The One-Factor ANOVA was the most appropriate parametric test for determining statistical significance – however, the physical fitness and physical activity categories failed to meet the proper ANOVA conditions. As such, multiple non-parametric (Welch's) t-tests were conducted for both variables. The results were as follows:

- **PFS_CAT:** none of the categories have significantly different functioning scores than the means.
- **PA_CAT:** none of the categories have significantly different functioning scores than the means.

Multiple t-tests were conducted for income category with interference as well, and it was found that there is a significant difference between the means of Category 1 and Category 3 regarding interference (p-value = 0.047 at $\alpha = 0.05$). The Bonferroni correction produced a corrected α = 0.141 but the output stated that null hypothesis could not be rejected.

```
In [104]: # Bonferroni Correction

          p_income = [0.388, 0.047, 0.252]

          p_adjusted_income = multipletests(p_income, alpha=0.05, method='bonferroni')
          p_adjusted_income

Out[104]: (array([False, False, False], dtype=bool),
           array([ 1.   ,  0.141,  0.756]),
           0.016952427508441503,
           0.016666666666666666)
```

The income category/working memory and income category/cognitive flexibility parameters met the conditions for ANOVA and the following results were obtained:

1) There is a significant difference in mean working memory scores between students from low-income and high-income backgrounds

2) There is no significant difference between the three group means with respect to cognitive flexibility.

The Pearson correlation coefficient was a more apt test to determine the statistical significance of the continuous variables.

The following results were found:

- **Age:** There is a significant negative correlation between age and cognitive flexibility (r = -0.146; p = 0.012)

- **Weight:** There is no significant correlation between weight and executive function.

- **BMI:** There is no significant correlation between BMI and executive function.

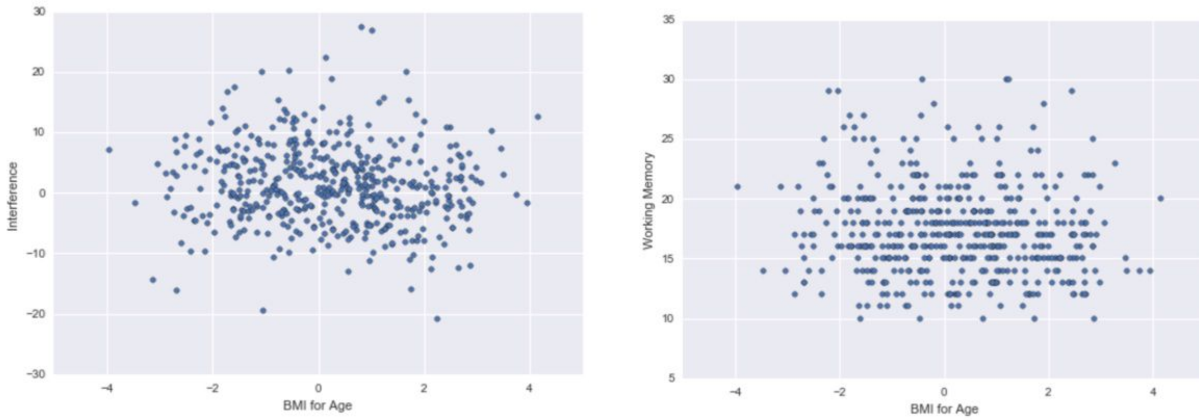

Figure 7. BMI-for-Age vs. Interference (left) and Working Memory (right).

- **BMI-for-Age (Figure 7):** There is a significant negative correlation between BMI-for-age and interference (r = -0.089 ; p = 0.048), as well as BMI-for-age and working memory (r = -0.094; p = 0.038). The correlation between BMI-for-age and cognitive flexibility is insignificant (r = 0.038; p = 0.396).

- **Physical_fitness_score:** There is no significant correlation between physical fitness scores and executive function.

- **PA_total_score:** There is no significant correlation between physical activity scores and executive function.

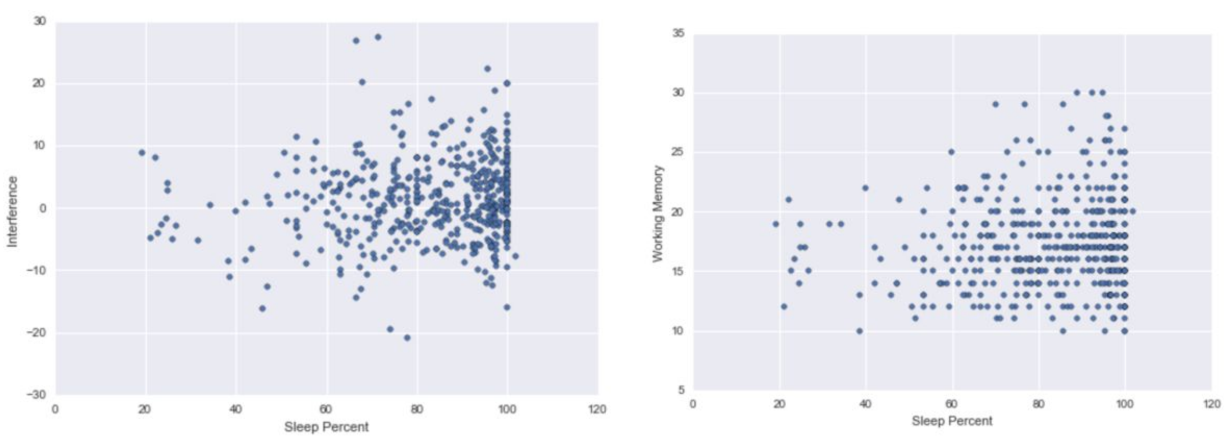

Figure 8. Sleep Percent vs. Interference (left) and Working Memory (right).

- **Sleep_percent (Figure 8):** There is a significant positive correlation between sleep percent and interference ($r = 0.110$; $p = 0.015$), as well as sleep percent and working memory ($r = 0.092$; $p = 0.042$).

- **Household_size (Figure 9):** There is a significant negative correlation between household size and working memory ($r = -0.098$; $p = 0.0306$).

_