April 12, 2021

Yitong Sun

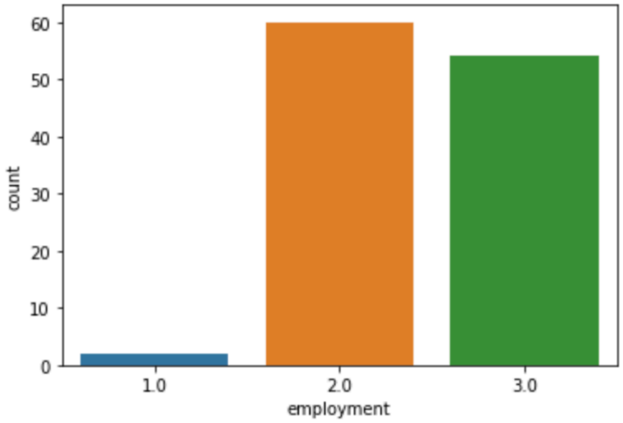
Tufts University

Report for RBS BI Co-op Assignment

- **Question 1: How does GPA impact on employment?**

During the early stage of the data exploration of the topic, I have identified that the "employment" field consists of three main categories of entries: Full-time Jobs, Part-time Jobs, and Unemployed, with Others being the outliers. By plotting the distribution of entries in the "employment" section on the bar chart, the visual data from the bar chart allow us to learn that:

- 1 entry for student who has a full-time jobs
- 0 entry for the Others category



After gaining the understanding of the data structure for 'employment', I used three different charts to demonstrate the impact of GPA on employment and the reverse cause-and-effect relationship as in below:

Fig. 1 Pair Plot

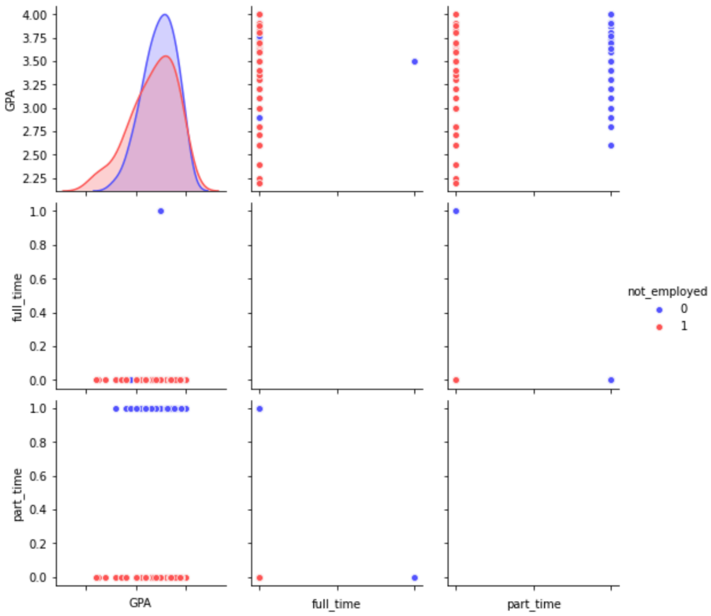


Fig. 2 Violin Plot

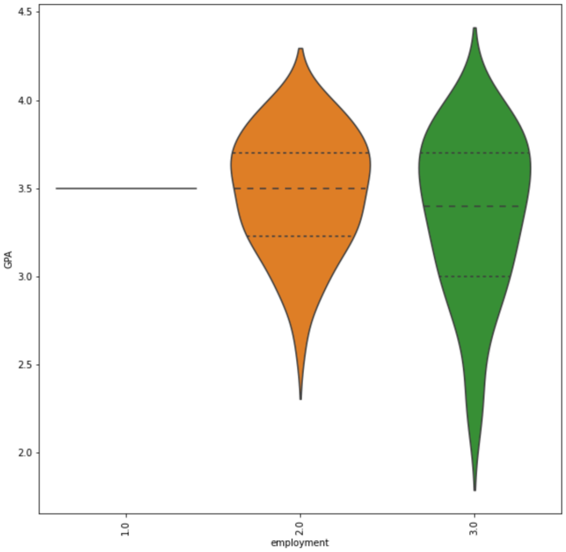
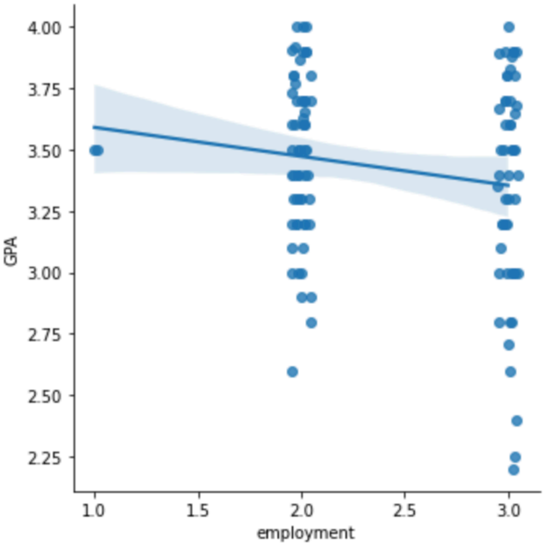


Fig. 3 Scatter Plot with Regression



To further explore whether the influence of GPA would impact employment, I used violin plot and scatter plots (Fig.2 and Fig.3) to analyze the relationship between employment and GPA. When the GPA is approximately less than 2.4, no student participates in full-time or part-time work. When the GPA is approximately greater than 3.0, there is no significant difference in employment. Students with lower GPA are not likely to find jobs, while those with higher GPA are more likely to find part-time jobs.

From Fig. 3, we could also see that employment is negatively correlated with GPA, which means having jobs does not impact GPA as what we would intuitively expect.

- **Question 2: Evaluate ‘weight’ Field**

The first step after cleaning the ‘weight’ column is to see the distribution of this field.

Fig. 4 Histogram on ‘weight’ Field

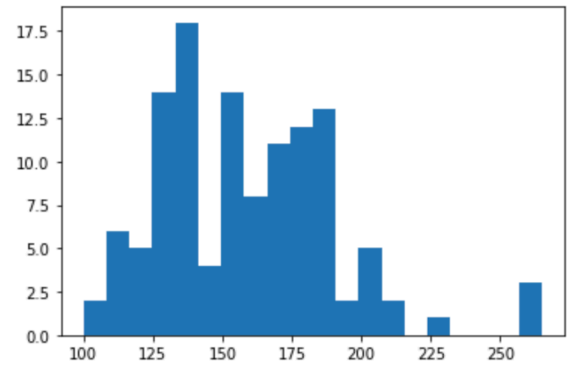
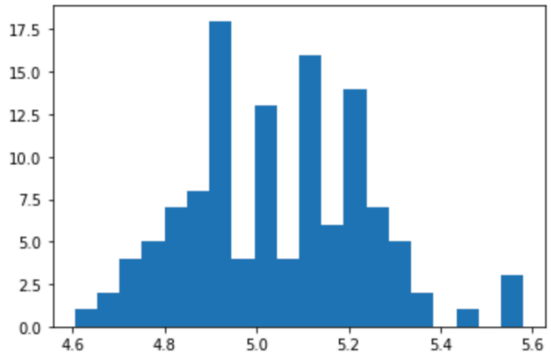


Fig. 5 Histogram on ‘weight’ Field after Log Normalization



It can be seen from the Fig. 4 that the weight data appears to the left, that is, the median weight is less than the average weight. Left-skewed data can easily lead to excessive consideration of left-skewed data in statistics, resulting in statistical errors and possible heteroscedasticity problems. Taking the logarithm can enlarge the value less than the median by a certain proportion, thus forming a normal distributed data as shown. However, problem still persists with the uneven data distribution and outliners.

- **Question 3: Potential Enhancement to Dataset**

Attributes could be added:

  - Scholarships
  - Students’ Free time, which can be used to explore the impact of free time on cooking habit, current diet, daily exercise and so on.
  - Academic Pressure

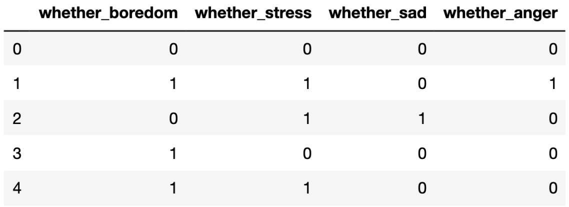
Attributes should be revised:

  - Income: students that are “unemployed” having an annual Income above $100,000, which is intuitively conflicting.
  - Survey Questions:
    - “Which of these pictures you associate with word ‘xyz’?” – The question itself appears to be ambiguous for thorough analysis.
    - An example of the questions should be asked instead: “Which food/drink would you prefer from the following pictures?”

- **Question 4 – Cleaning Column**

For processing of data surrounding comfort_food_reasons, dummy variable can be used, such as student’s boredom, stress, angry, etc. The result of the data analysis is then outputted as a table:

Table 1 Dummy Variables for Field ‘comfort_food_reasons’



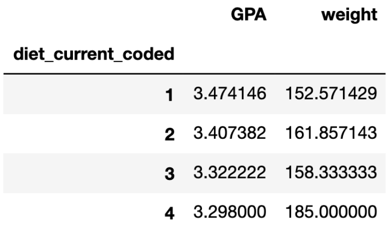
	whether_boredom	whether_stress	whether_sad	whether_anger
0	0	0	0	0
1	1	1	0	1
2	0	1	1	0
3	1	0	0	0
4	1	1	0	0

The specific data cleaning ideas are as follows: read the comfort_food_reasons of different students, use split to decompose the text, and calculate whether comfort_food_reasons appear in the comfort_food_reasons according to different emotions such as "sad", "stress" and other keywords. If they exist, record the dummy variable of the emotion as 1. In addition, I cleaned the data according to what the first word of the text is and compared the cleansing results with the comfort_food_reasons_coded in the data and found that most of the methods I used were the same, and a few of the comfort_food_reasons_coded classified by the text appeared to be different. Please see the code output (RBS_assignment.ipynb) for further information.

- **Question 5 – Further Exploration**

Part 1 - How do diets impact GPA and weight?

Table 2 Diets Impact on GPA and Weight



	GPA	weight
diet_current_coded		
1	3.474146	152.571429
2	3.407382	161.857143
3	3.322222	158.333333
4	3.298000	185.000000

As we group by the students' current diets, we are able to identify that students under more scientific diets have higher GPA and lower weight; students with more scientific diets have higher GPAs than students with unscientific diets, and the average weight is lower than those with unscientific diets.

Part 2 - Fligner's Test

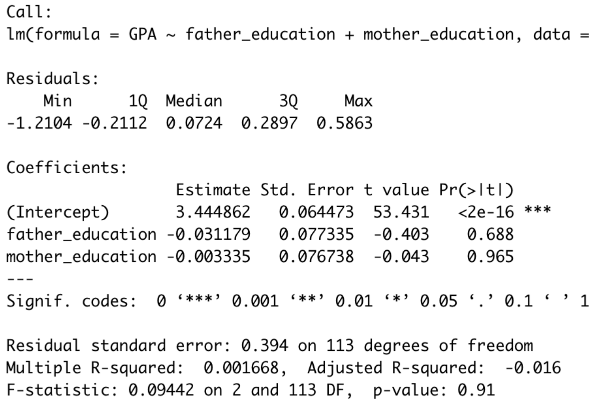
I run normal test on 'GPA' column. Given the P-value = 0.00313 < 0.05, it is confirmed that 'GPA' data conforms normal distribution with statistics value of 11.53. Thus, I use Fligner-Killeen to tests the null hypothesis that all input data are from populations with equal variances with alpha = 0.05 for 'Gender' and 'employment'. And the test results are as in below:

For 'Gender' data, P-value = 0.60862 > 0.05, do not reject null hypothesis – there is no significant difference on GPA between female and male students.

For 'employment' data, P-value = 0.038929 < 0.05, reject null hypothesis – there is a significant difference on GPA among students who have full-time jobs, part-time jobs and no jobs.

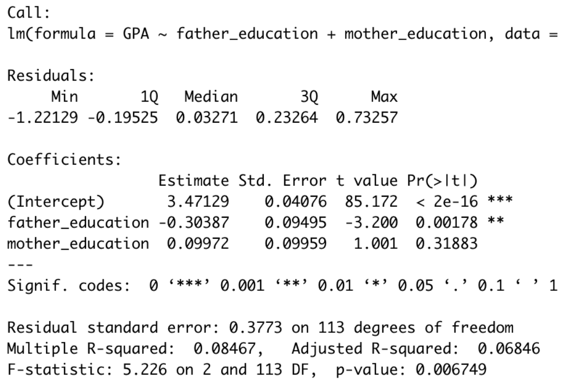
Part 3 - Ordinary Least Squares Modeling

I would like to determine if parents' education backgrounds would affect students' GPA. First, I converted the 'father_education' data and 'mother_education' data into binary data by dividing the education background into either below college degree or college degree and above. Thus, I build the first OLS model (Ordinary Least Squares model). The summary of the modeling is as in below:



```
Call:
lm(formula = GPA ~ father_education + mother_education, data = gpa_college)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.2104 -0.2112  0.0724  0.2897  0.5863 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       3.444862   0.064473  53.431   <2e-16 ***
father_education -0.031179   0.077335  -0.403    0.688    
mother_education -0.003335   0.076738  -0.043    0.965    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.394 on 113 degrees of freedom
Multiple R-squared:  0.001668,  Adjusted R-squared:  -0.016 
F-statistic: 0.09442 on 2 and 113 DF,  p-value: 0.91
```

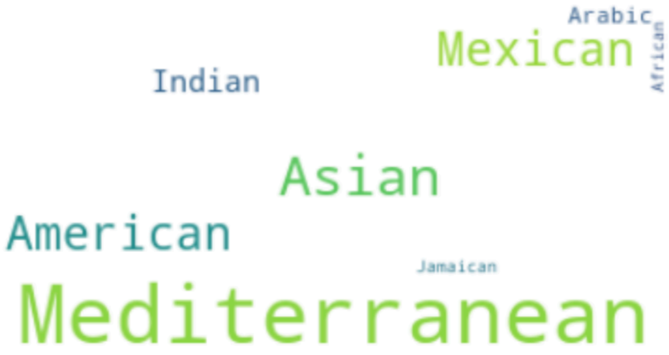
Given the p-value = 0.91, the first model is confirmed to be invalid, and it proves that parents’ college degree or above does not have a significant impact on their children’s GPA compared with parents with GED degree or some college degree. What if the education background goes up a little further? To further explore the impact of parents’ education background, I built a second model with binary education data but dividing the columns into either graduate degree or below this time. And the results as in below:



```
Call:
lm(formula = GPA ~ father_education + mother_education, data = gpa_advanced)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.22129 -0.19525  0.03271  0.23264  0.73257 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       3.47129    0.04076  85.172  < 2e-16 ***
father_education -0.30387    0.09495  -3.200  0.00178 ** 
mother_education  0.09972    0.09959   1.001  0.31883    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3773 on 113 degrees of freedom
Multiple R-squared:  0.08467,   Adjusted R-squared:  0.06846 
F-statistic: 5.226 on 2 and 113 DF,  p-value: 0.006749
```

Given the p-value = 0.0067, it is confirmed that this model is valid. Student whose father has graduate degree would have GPA 0.30 lower than peers whose fathers have college degrees or below, while students whose mother has graduate degree would have GPA 0.1 higher than peers whose mother doesn’t with all other variables being constant. For more information, please check the R script file (Question5_extended.rscript) in the assignment folder.

Part 4 – Word cloud of favorite cuisine



Challenge:

In Question 5, Part 2, I intuitively used t-test to test for statistical significance on variables, however, data would have sufficed both homogeneity of variance and normal distribution where the normal test result indicates it is not. Thus, I used Fligner-Killeen test instead since it does not have requirement for data distribution and variance.

Remarks

This dataset is great for exploring health status, health awareness and GPA of college students with respect to diet, weight, exercise, their perception of food. In previous data exploring, it is identified that students with higher GPA are more likely having full time or part time jobs compared with students with lower GPA. Having a job also will not affect students' GPA generally. Additionally, students under more scientific diets have higher GPA and lower weights compared with students who are not. Among all students, there is not much different in GPA between female and male, but there is a significant different between students who are employed and who are not. Last but not least, it appears students whose father has a graduate degree usually have 0.3 less GPA than students whose father has a college degree or below, while students whose mother has a graduate degree usually have 0.1 more GPA than students whose mother does not. This discrepancy showed from OLS model result is quite interesting and worth further research and exploration.