Student Behavior

Members: Meghana Srinivasa, Frederick Kusumo, and Elaine Wu

Statement of Purpose

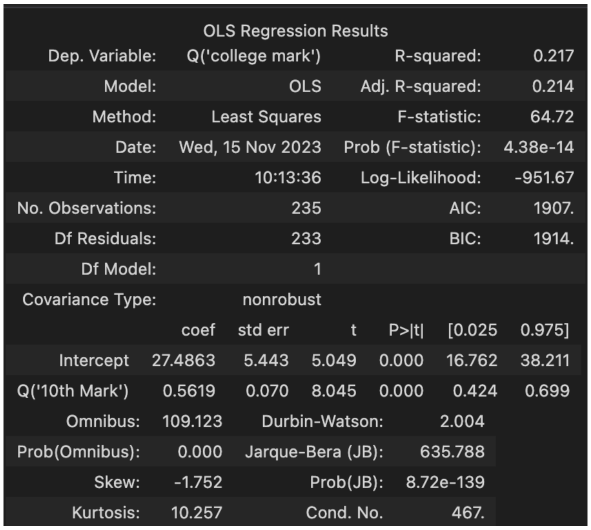
The dataset we are exploring is the Student Behavior dataset from Kaggle which was collected from university students through a Google form. It contains information such as grades in 10th and 12th grade, college grades, hobbies, daily studying time, salary expectations, satisfaction with their degree, stress levels, financial status, and more.

Our purpose is to learn how different facets of student behavior and life influence students' current and future outcomes. Knowing this is important to both universities and students because it provides insights for universities on how to create a happier student body and for students on how to manage stress. Specifically, our objectives and problem statements are to find out whether students who score good grades in 10th grade score good grades in college, whether certain hobbies are better for stress, whether more daily studying leads to better salary expectations, and whether stress levels affect college performance and a student's willingness to pursue their field of study.

Question 1

The first question we wanted to explore was: do students who score high grades in 10th grade score high grades in college? This is an interesting question because colleges often use a student's high school grades as indicators of what grades that student will receive at their college, and this becomes a large factor in admissions decisions. Determining whether there is a strong correlation between 10th grade and college scores will help determine whether this should be a large factor in college admissions.

The analytical technique we used was regression analysis. We performed an OLS regression with college scores as the dependent variable and 10th grade scores as the independent variable. We analyzed the R-squared, p-value, and coefficient of the independent variable to determine whether or not there was a relationship between 10th grade scores and college scores.



OLS Regression Results

Dep. Variable:	Q('college mark')	R-squared:	0.217
Model:	OLS	Adj. R-squared:	0.214
Method:	Least Squares	F-statistic:	64.72
Date:	Wed, 15 Nov 2023	Prob (F-statistic):	4.38e-14
Time:	10:13:36	Log-Likelihood:	-951.67
No. Observations:	235	AIC:	1907.
Df Residuals:	233	BIC:	1914.
Df Model:	1		
Covariance Type:	nonrobust		

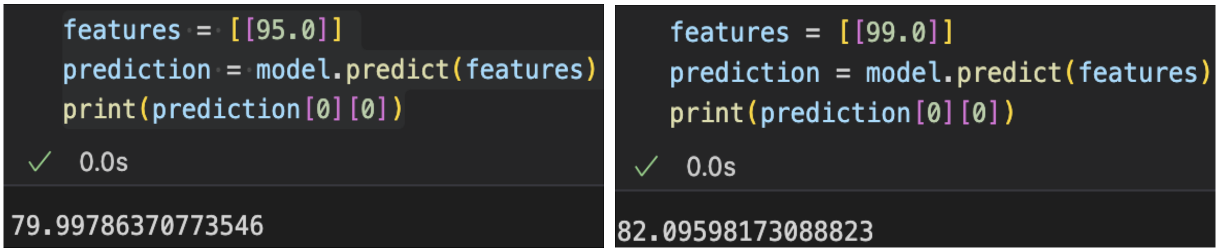
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	27.4863	5.443	5.049	0.000	16.762	38.211
Q('10th Mark')	0.5619	0.070	8.045	0.000	0.424	0.699

Omnibus:	109.123	Durbin-Watson:	2.004
Prob(Omnibus):	0.000	Jarque-Bera (JB):	635.788
Skew:	-1.752	Prob(JB):	8.72e-139
Kurtosis:	10.257	Cond. No.	467.

Figure 1: OLS Regression Results from 10th grade scores regressed on college scores

The coefficient of determination, R-squared, is 0.217. This means the model explains about 21.7% of the variance. The model is statistically significant because the p-value, Prob(F-statistic), is less than 0.05. We can reject the null hypothesis that 10th grade scores have no effect on college scores. From the coefficient of the 10th Mark, it seems that as a student's 10th grade score increases, their college scores increase as well. For example, if a student ended their 10th grade with a 95, their predicted college score would be 27.4863+(0.5619)(95) = 80.87. But if they ended their 10th grade with a 99, their predicted college score would be 27.4863+(0.5619)(99) = 83.11.

To check whether these predictions are accurate, we developed a machine-learning model. We split our data into training and testing sets, created a linear regression model, and fit our data onto the model. We analyzed both the predictions and the mean squared error value.



```
features = [[95.0]]
prediction = model.predict(features)
print(prediction[0][0])
✓ 0.0s

79.99786370773546
```

```
features = [[99.0]]
prediction = model.predict(features)
print(prediction[0][0])
✓ 0.0s

82.09598173088823
```

Figure 2: Predictions from the Linear Regression Model

The root mean squared error was about 15.66. This means that the model could use work because the lower the root squared mean value the better. The RMSE shows how far predictions fall from true values. We want predictions to be as close to the real values as possible, which is why a small RMSE means a good model. The OLS regression predicted that when a student scored a 95 at the end of 10th grade their college score would be about 80.9 and when they scored a 99 at the end of 10th grade their college score would be about 83.1. The linear regression model predicted that when a student scored a 95 their college score would be about 80.0 and when they scored a 99 their college score would be about 82.1. The linear regression model predicted one point lower for the college scores compared to the OLS regression. Since both model predictions were similar to each other, we can have more confidence in the accuracy of our predictions.

The visualization technique we used to portray the relationship between 10th grade and college scores was a heatmap.

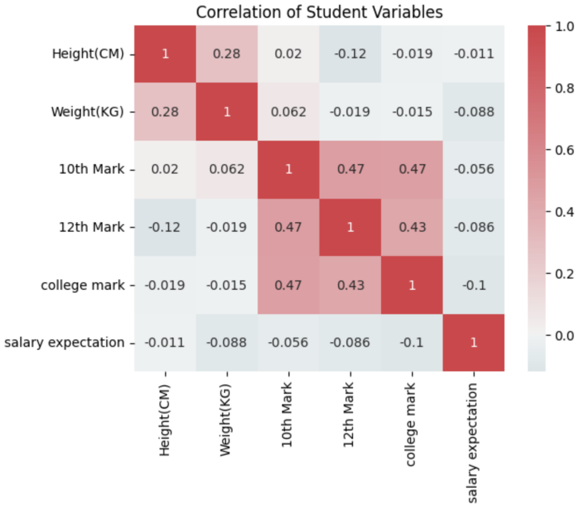


Figure 3: Heatmap Depicting Correlation of Student Variables

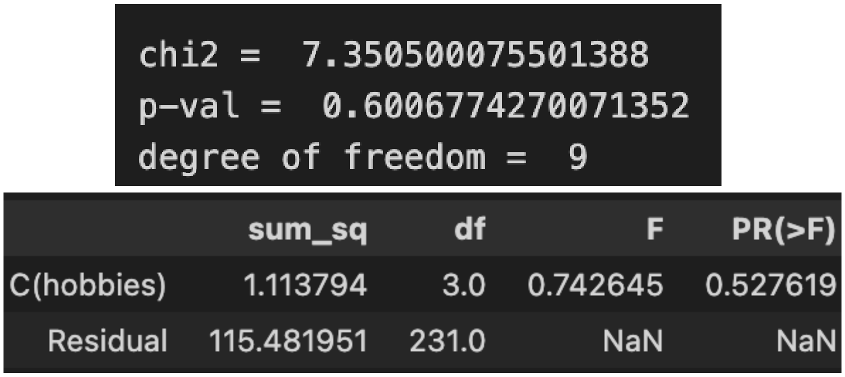
The highest correlations were between 10th grade scores and college scores and 10th grade scores and 12th grade scores. 10th grade scores and college scores having one of the highest correlations is evidence that 10th grade scores can be a good predictor of college scores. It is surprising that 12th grade scores had a lower correlation to college scores than 10th grade scores. Since 12th grade is right before college, we would have expected it to have a higher correlation. One possible explanation is that seniors in high school get burned out and work less once they have received college admissions, so their grades decline a little.

Overall, both analytical and visual techniques point to 10th grade scores being correlated to college scores. Thus, students who score high grades in 10th grade are more likely to also score high grades in college.

Question 2

The second question we wanted to explore was: are certain hobbies better for stress? This is a relevant question because most students are stressed and time-crunched. If there is one hobby that is better at dealing with stress than others, it would be efficient for college students to spend their limited time on that hobby to alleviate their stress.

The analytical technique we used was a chi-squared and ANOVA test. The chi-squared analysis helped us understand whether there was a significant association between certain hobbies and stress levels. We analyzed the chi-squared statistic and the p-value to learn if there was an association between certain hobbies and stress. The ANOVA test helped determine if there was a statistically significant difference in the mean stress levels among hobbies, and we analyzed the p-value to learn this.



```
chi2 =  7.350500075501388
p-val =  0.6006774270071352
degree of freedom =  9
```

	sum_sq	df	F	PR(>F)
C(hobbies)	1.113794	3.0	0.742645	0.527619
Residual	115.481951	231.0	NaN	NaN

Figure 4: Results from Chi-Squared and ANOVA tests

The chi-squared value was about 7.35. This value represents the sum of the differences between actual and expected data. The closer the number is to 0, the less difference between the actual and expected data. The p-value for the chi-squared test was 0.6 which is greater than 0.05. Even though there was a difference of 7.35 among stress levels for different hobbies, the results are not statistically significant. This means that we fail to reject the null hypothesis that there is a difference between stress levels for different hobbies. The p-value for the ANOVA test was about 0.527. Since the p-value is greater than 0.05, we fail to reject the null hypothesis that there is not a significant difference between stress levels for different hobbies. The model is not statistically significant.

The visualization technique used to portray stress levels among different hobbies was a count plot and a bar chart.

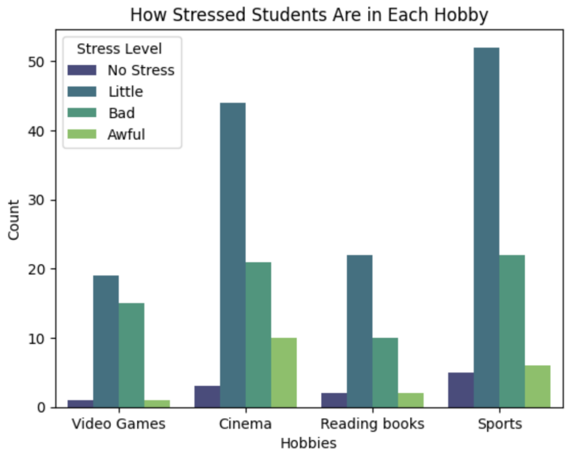


Figure 5: Countplot of How Stressed Students Are in Each Hobby

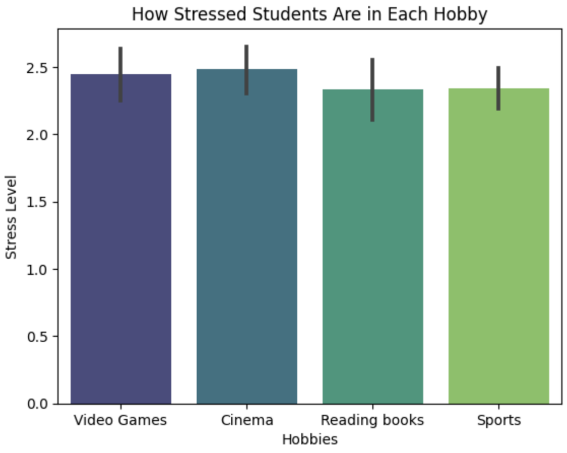


Figure 6: Barchart of Average Stress Level for Each Hobby

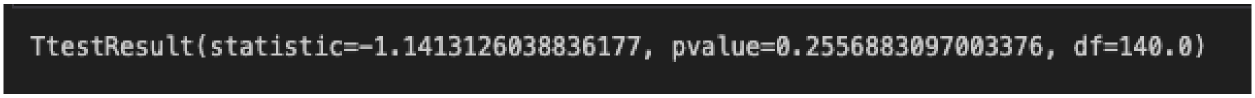
The count plot shows that for all hobbies, the most common level of stress was a little stress. The second most common level of stress for all hobbies was a bad level of stress. For the bar chart, we converted the stress levels into numeric values. 0 represents no stress while 4 represents an awful level of stress. Students who use cinema as their hobby have the highest average level of stress, and students who read books as their hobby have the lowest level of stress. However, it does not seem like these findings are statistically significant.

Overall, both analytical and visual techniques point to no hobby being better for stress than another. A possible explanation for this is that hobbies are personal, so what one student finds as a relaxing hobby might not be the same for another student.

Question 3

The third question we wanted to explore was: does higher daily studying time correlate with higher salary expectations? The question is relevant because students would be more motivated to study if they knew it would lead to higher salary expectations. We expect that a student who spends long hours studying will have a higher salary expectation due to their diligence in investing in long hours for better performance.

The technique we used for this specific question is the t-test statistic analysis. The t-test is used to compare the means of two groups. It is a hypothesis test to determine whether two groups are different from one another. We analyze the p-value in the t-test to see if there is evidence against the null hypothesis. The p-value from the statistical computation result will tell if there is a significant difference between the two variables we chose.



```
TtestResult(statistic=-1.1413126038836177, pvalue=0.2556883097003376, df=140.0)
```

Figure 7: t-test Result on Daily Studying Time to Salary Expectation

The t-test result from Figure 7 shows the statistical value is -1.141. This value represents the mean of the two sample sets we created for experimentation in finding the correlation. The value shows that the data lies to the left of the mean in a distribution. The p-value was 0.255 from the t-test, which is greater than 0.05. The result was not statistically significant, and we failed to reject the null hypothesis that there is a difference between the daily study time and salary expectations. Given that, the model is not statistically significant.

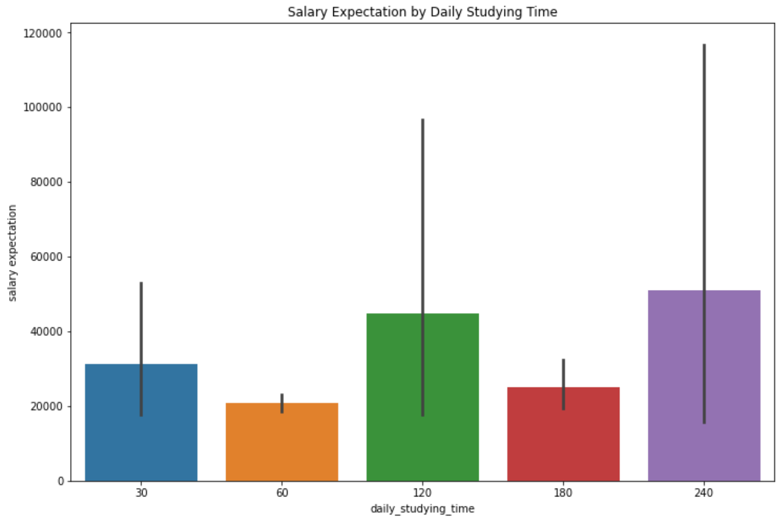
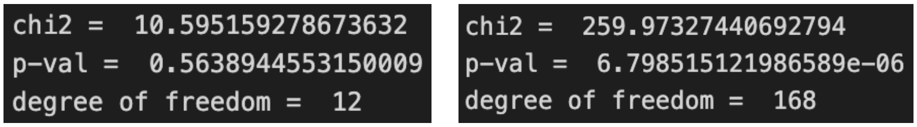


Figure 8: Salary Expectation by Daily Studying Time

Based on the visualization above, it shows the count of students who spend a certain amount of time studying according to their salary expectations. The graph does not present the linear correlation that we hypothesized initially. The visual shows that students who spent 60 minutes and 180 minutes studying have lower salary expectations than students who spent 30 minutes or 120 minutes studying. Yet, the students who spend more than two hours are overall the highest among others, but it does not show a high correlation between the two variables. Overall, both analytic and visual techniques point to no correlation between studying time and salary expectations, as they are not statistically significant. A possible explanation is that each student's academic ability and learning style are different. They all manage their study time differently, which is unique to every other student.

Question 4

The fourth question we wanted to explore was: do stress level and college performance affect the willingness of students to pursue their field of study? This is a relevant question because there are many variables that can affect a student's stress and their grades. Getting stressed or having bad grades does not necessarily mean that the student will not want to pursue what he or she is studying.



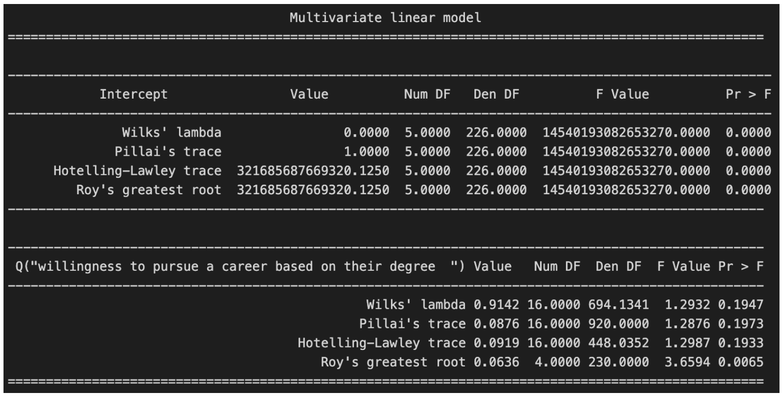
```
chi2 =  10.595159278673632
p-val =  0.5638944553150009
degree of freedom =  12
```

```
chi2 =  259.97327440692794
p-val =  6.798515121986589e-06
degree of freedom =  168
```

Figure 9: Chi-squared, left (Stress Level vs Willingness to pursue) and right (College Mark vs Willingness to pursue)

The technique we used is a chi-squared analysis and MANOVA. The chi-squared analysis is used to determine whether the variables have homogenous or heterogenous variance which is going to dictate which findings to look at based on the MANOVA, where if they are homogenous Pillai's trace better represents the data and if heterogenous Wilks' lambda better represents the data. We used MANOVA instead of ANOVA because we have two independent variables and one dependent variable.

According to the article "Heterogeneity and Homogenous Data in Statistics" by Stepanie Glen, the lower the p-value from the chi-test, the heterogeneity in the data is more significant. Based on the p-values of the chi-test, Stress Level and Willingness to pursue have a high p-value meaning that they have homogenous variance, while College Mark and Willingness to pursue have a low p-value meaning that they have heterogeneous variance. This would mean it would be better to observe both Pillai's trace and Wilk's lambda.



```
                                     Multivariate linear model
=================================================================================================

-------------------------------------------------------------------------------------------------
            Intercept                Value          Num DF   Den DF          F Value         Pr > F
-------------------------------------------------------------------------------------------------
               Wilks' lambda               0.0000  5.0000  226.0000  14540193082653270.0000  0.0000
              Pillai's trace               1.0000  5.0000  226.0000  14540193082653270.0000  0.0000
      Hotelling-Lawley trace  321685687669320.1250  5.0000  226.0000  14540193082653270.0000  0.0000
         Roy's greatest root  321685687669320.1250  5.0000  226.0000  14540193082653270.0000  0.0000
=================================================================================================

-------------------------------------------------------------------------------------------------
 Q("willingness to pursue a career based on their degree  ") Value   Num DF  Den DF  F Value Pr > F
-------------------------------------------------------------------------------------------------
                                         Wilks' lambda 0.9142 16.0000 694.1341  1.2932 0.1947
                                        Pillai's trace 0.0876 16.0000 920.0000  1.2876 0.1973
                                Hotelling-Lawley trace 0.0919 16.0000 448.0352  1.2987 0.1933
                                   Roy's greatest root 0.0636  4.0000 230.0000  3.6594 0.0065
=================================================================================================
```

Figure 10: Result from MANOVA test

Both p-values of Wilk’s lambda and Pillai’s trace are larger than 0.05 meaning that we fail to reject the null hypothesis, so the results are not statistically significant which means that there are more variables involved. The visualizations below provide new perspectives.

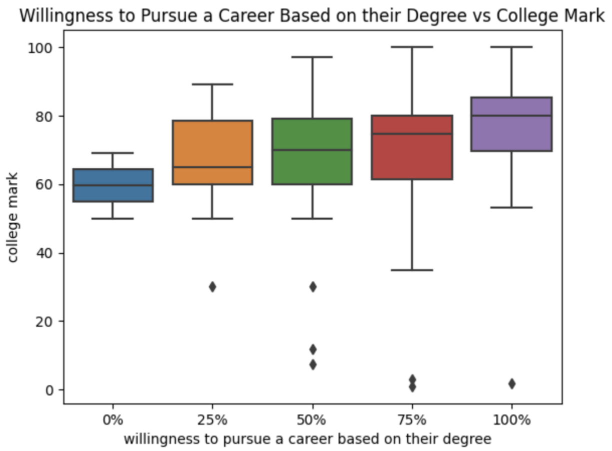


Figure 11: Boxplot of College Mark and Willingness to pursue a career based on their degree

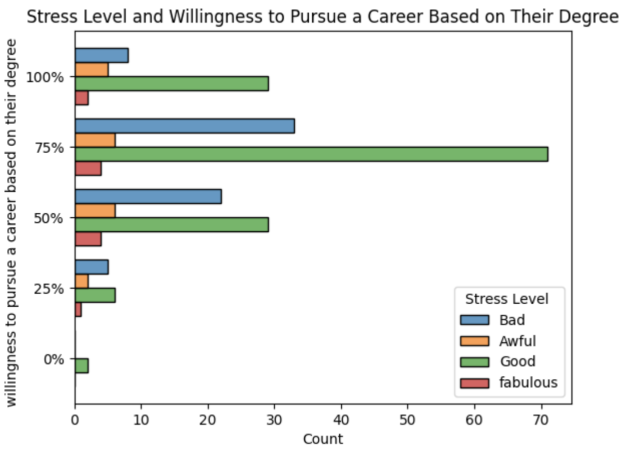


Figure 12: Countplot of Stress Level and Willingness to pursue career based on their degree

From the boxplot, it can be seen that as the mean of the college score increases, the desire to pursue a career related to the degree increases. The same can be seen from the count plot as 75% willingness to pursue has a higher bad stress level than 50%. Moreover, it can be seen from the overall count that 75% is higher than the remaining willingness to pursue which can also mean that students are generally not 100% sure that they are willing to pursue a career based on their degree yet.

Overall, both analytical and visual techniques are not pointing in a clear direction which tells us that this matter requires more variables to be taken into account to actually predict or observe statistical significance. However, it can be hypothesized that college marks have a higher effect on the willingness to pursue a career in the degree than stress level.

Limitations and Challenges

An overall limitation we had was the size of our data. Our dataset only has 235 rows, which means we only have data from 235 students. 235 students is not enough data to accurately extrapolate to all students. Additionally, there needed to be more information available on where this data was collected. If all the data was collected at one university, the data is then only applicable to students at that university rather than all students.

A specific limitation for Question 1 was that we wanted to analyze the correlation between scores from 9th to 11th grade because those grades are what colleges consider in admission decisions. However, the dataset only included scores for 10th and 12th grades. Another challenge was that the RMSE value for the linear regression model was not low enough for us to have significant confidence in the model. A challenge for Question 2 was doing data analysis by comparing two categorical groups. The first group was stress levels which were labels rather than numbers, and the second group was hobby type. To allow for multiple types of analyses, we changed the stress levels from labels to numeric values. A challenge for Question 3 was analyzing the correlation between daily studying time and salary expectations because the comparison of the two variables was numerical to the categorical group. To allow for statistical analyses, we needed to change the categorical to numerical values. A challenge for Question 4 is our unfamiliarity with MANOVA. It was difficult to interpret the findings because there are more types of intercepts produced compared to an ANOVA, which then required a chi-test to dictate which one to use.

Recommendations for Future Work

A recommendation for future work for Question 1 is exploring why 12th grade and college scores had a lower correlation than 10th grade and college scores, especially when 10th and 12th grade scores had a high correlation. Additionally, an area for future research would be lowering the RMSE value for the linear regression model to create a more accurate model.
An area for future work for Question 3 is to explore the reasoning behind why certain amounts of studying time have lower or higher expected salaries. A recommendation for Question 4 would be learning more background information regarding a MANOVA.

Citations

Ateş, Can, et al. “Comparison of Test Statistics of Nonnormal and Unbalanced Samples for Multivariate Analysis of Variance in Terms of Type-I Error Rates.” *Computational and Mathematical Methods in Medicine*, Hindawi, 18 July 2019, www.hindawi.com/journals/cmmm/2019/2173638/.

Stephanie. “Heterogeneity and Heterogeneous Data in Statistics.” *Statistics How To*, 10 Dec. 2020,www.statisticshowto.com/heterogeneity/#:~:text=A%20heterogeneous%20population%20or%20sample,would%20be%20heterogeneous%20for%20height.