Project 1 Write Up

Executive Summary:

A dataset of 406 survey responses from students at Stellenbosch university was analyzed to examine the relationships between demographic information, academic performance, and alcohol habits. To perform this analysis, a dataset was cleaned, visualizations were created, and correlations and a regression model were developed. The key findings from the dataset indicate a difference in alcohol consumption habits between students on and off scholarship and that modules failed and classes missed are the best predictors of current GPA. All analysis was performed in Python using libraries including MatPlotLib, Seaborn, and SciKitLearn.

Introduction to Dataset:

Our dataset includes a survey of students enrolled at Stellenbosch University located in South Africa. The survey includes 406 responses received in the spring of 2024 the main aim of this survey was to predict students' academic performance based on the factors included such as drinking habits, scholarship status, residency status, economic factors, etc.

Research Questions and Areas of Interest:

The initial development of the research questions for the proposal focused on three areas:

1. What is the impact of drinking habits (both in volume consumed and number of days socializing per week) on academic performance and student outcomes?

2. What are the relationships between academic majors and alcohol habits?

3. Does scholarship status and/or living arrangement lessen the impact of alcohol on days of class missed, even for people consuming similar amounts?

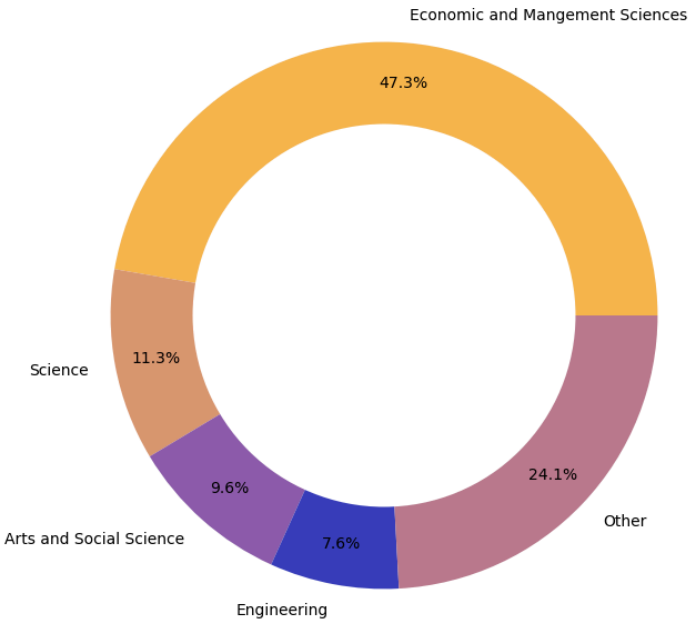
Throughout the process of developing the process, our focus slightly shifted and reframed the research to the following areas:

1. What are the patterns are present in the academic performance and alcohol consumption habits for participant when grouped by
    a. Major
    b. Scholarship Status
    c. Residential Status
2. Which factors gathered in this dataset are the best predictors of current GPA?

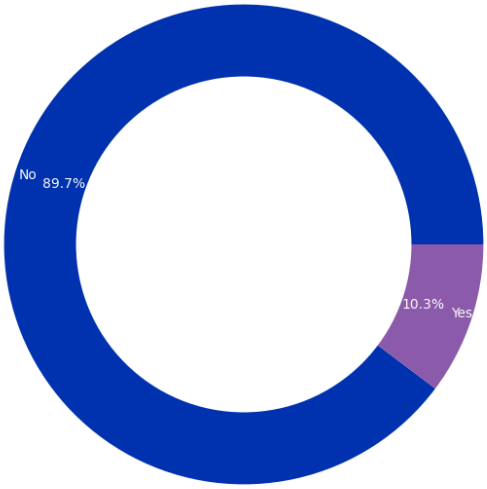
This shift in focus occurred due to both a downsizing in the number of team members of the project after an initial submission of the proposal, a better understanding of the information and data types available in the dataset, and a realization of realistic scope of work expectations amongst our team members.

To understand the makeup of our dataset amongst our categories of interest, donut charts were created showing the most popular majors, scholarship status, and residential status. The figures generated are displayed below:

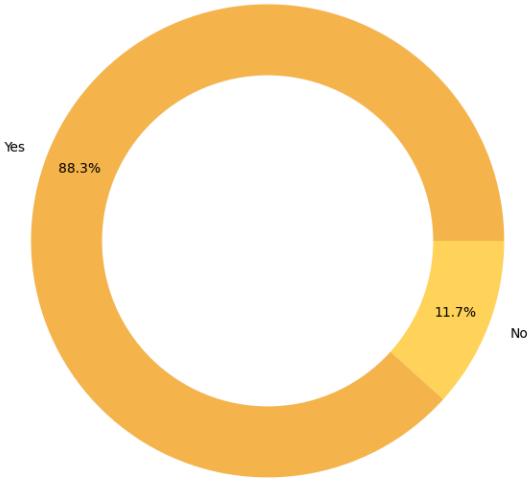
Majors of Study Participants



Status of Study Participants: On Scholarship?



Status of Study Participants: Off-Campus Residence?



It was recognized that the social and demographic factors of the dataset (including these areas of interest), the alcohol consumption habits of students, and the academic habits of students were all influential on each other, with none being truly independent. This was further explored in the regression analysis portion of the project.

Data Cleaning:

Our data set began with 406 entries, 17 columns (15 as objects, 2 as floats), with several null values scattered throughout as shown in Figure 1. During the data cleaning process all columns were renamed as the previous names/titles were the questions asked in the survey. Each column was assigned a name related to the questions asked.

```
orig_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 406 entries, 0 to 405
Data columns (total 17 columns):
 #   Column                                                                  Non-Null Count  Dtype  
---  ------                                                                  --------------  -----  
 0   Timestamp                                                               406 non-null    object 
 1   Your Sex?                                                               404 non-null    object 
 2   Your Matric (grade 12) Average/ GPA (in %)                              399 non-null    float64
 3   What year were you in last year (2023) ?                                333 non-null    object 
 4   What faculty does your degree fall under?                               399 non-null    object 
 5   Your 2023 academic year average/GPA in % (Ignore if you are 2024 1st year student)  320 non-null    float64
 6   Your Accommodation Status Last Year (2023)                              383 non-null    object 
 7   Monthly Allowance in 2023                                               375 non-null    object 
 8   Were you on scholarship/bursary in 2023?                                398 non-null    object 
 9   Additional amount of studying (in hrs) per week                         403 non-null    object 
 10  How often do you go out partying/socialising during the week?           404 non-null    object 
 11  On a night out, how many alcoholic drinks do you consume?               404 non-null    object 
 12  How many classes do you miss per week due to alcohol reasons, (i.e: being hungover or too tired?)  403 non-null    object 
 13  How many modules have you failed thus far into your studies?            403 non-null    object 
 14  Are you currently in a romantic relationship?                           403 non-null    object 
 15  Do your parents approve alcohol consumption?                            402 non-null    object 
 16  How strong is your relationship with your parent/s?                     403 non-null    object 
dtypes: float64(2), object(15)
memory usage: 54.0+ KB
```

Nulls were dropped for values in columns 'gender', 'hs_gpa', 'scholarship', 'study_hours','partying_frequency', 'drinks_consumed', 'classes_missed','modules_failed', 'relationship', ‘parents_approval’ and ‘parent_relationship’.

The survey included students who began attending in 2024 but had several questions regarding current GPA, field of study, accommodation status and financials that targeted students in 2023. This resulted in a lot of null values for the 2024 first year students. We did not want to lose the answers to the remaining questions regarding their study and drinking habits so we chose to replace their current gpa with a “0.0” placeholder.

The values in “year_of_study” were changed from objects to integers to try to add more numeric values to the data set. Students completing their first academic term in 2024 were replaced with “0”, "1st Year": 1, "2nd Year": 2, "3rd Year": 3, "4th Year": 4, and "Postgraduate": 5.

The null values in “field_of_study” were replaced with “Other”.

The null values in “off_campus” and “monthly_allowance” were replaced with “N/A”. Responses within ‘year_of_study’, ‘scholarship’, and ‘off_campus” were shortened from descriptive strings to one word, "Yes" or "No" ,answers since each column had two possible outcomes.

The rows with “N/A” values in columns “off_campus” and “monthly_allowance” were filtered out using a mask. We chose to use a filter instead of removing them all together so they could be added back later. At the time of cleaning, we had not fully decided which rows were the most necessary for our research. Using a mask to filter them out provided us more flexibility in case anything needed to be changed

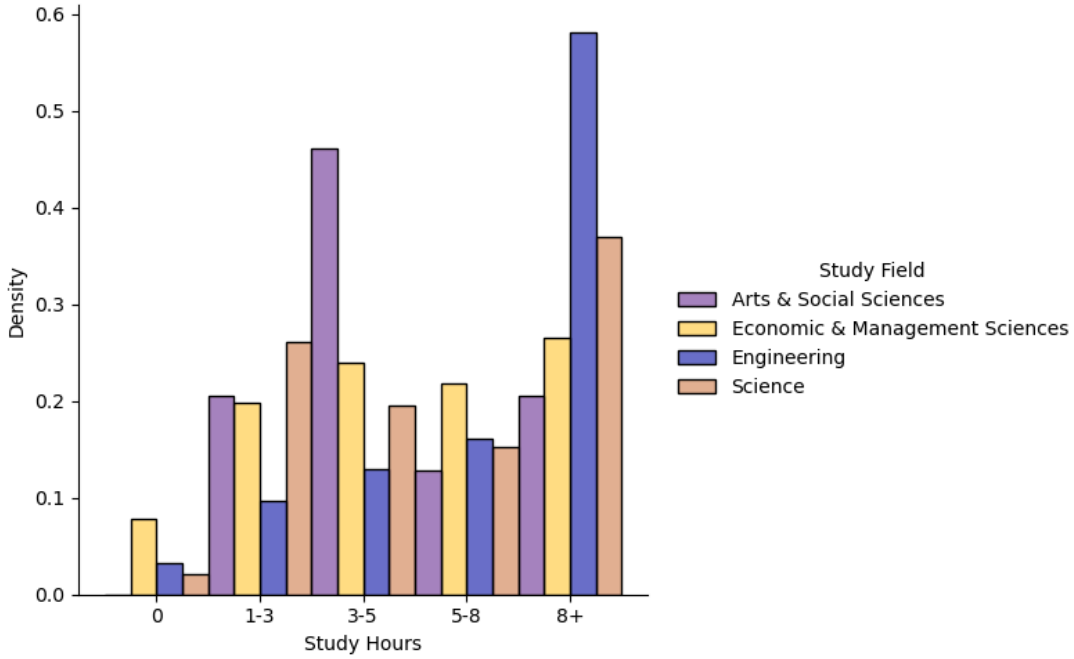
```
<class 'pandas.core.frame.DataFrame'>
Index: 351 entries, 0 to 405
Data columns (total 16 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   gender               351 non-null    object 
 1   hs_gpa               351 non-null    float64
 2   year_of_study        351 non-null    int64  
 3   field                351 non-null    object 
 4   current_gpa          351 non-null    float64
 5   off_campus           351 non-null    object 
 6   monthly_allowance    351 non-null    object 
 7   scholarship          351 non-null    object 
 8   study_hours          351 non-null    object 
 9   partying_frequency   351 non-null    object 
 10  drinks_consumed      351 non-null    object 
 11  classes_missed       351 non-null    object 
 12  modules_failed       351 non-null    object 
 13  relationship         351 non-null    object 
 14  parents_approval     351 non-null    object 
 15  parent_relationship  351 non-null    object 
dtypes: float64(2), int64(1), object(13)
memory usage: 46.6+ KB
```

At the end of the data cleaning process we had 351 entries, 16 columns with no null values.

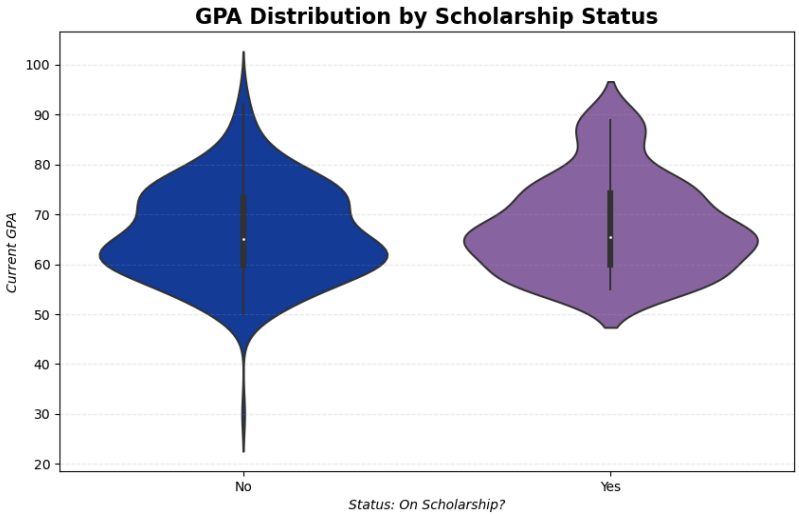
Academic Habits:

While analyzing the amount of time students spend studying per study field. We found that the majority of students majoring in Economic & Management Sciences, Engineering, and Science spend 8+ hours studying, while Arts & Social Sciences majors average higher in the 3-5 study hours per week category. We chose to analyze the students studying habits to see if certain majors might have less free time to spend on partying/drinking. However, due to the sample sizes of each study field category being completely uneven, it made this hard to determine. In order to get the distribution a column counting the amount of study hours for each category

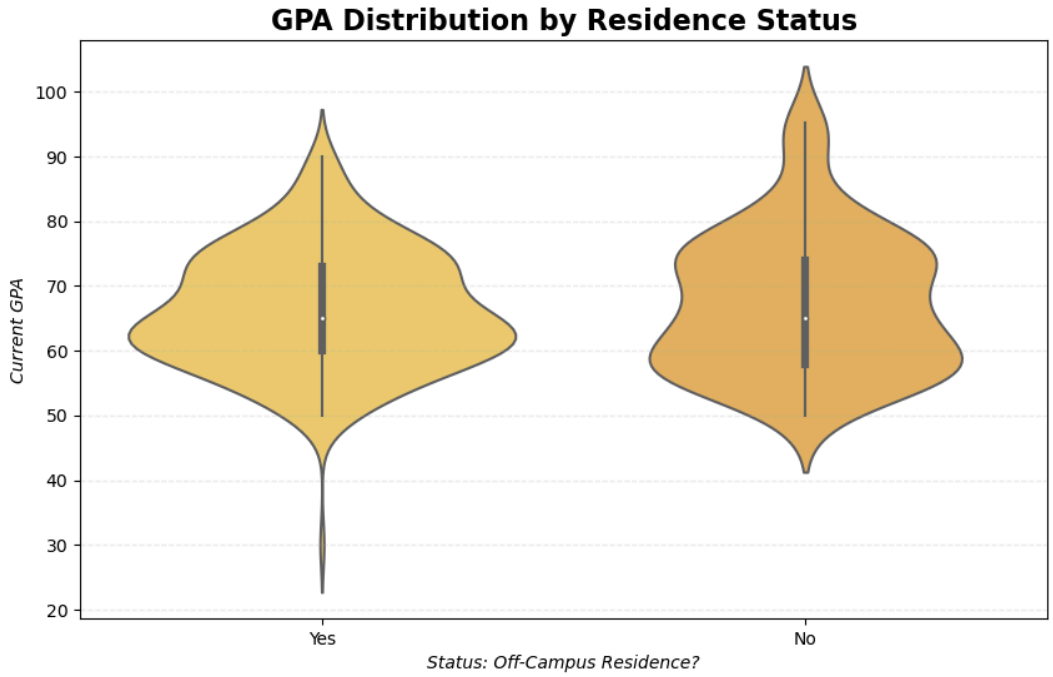
within their study fields was added so that a numerical value could be plotted with the two categorical values.



The current GPA distributions were generated for the dataset split by scholarship status and displayed in a violin plot. Both datasets had a median of GPA 65, with the group on scholarship having a mean GPA of 67.6 and the group not on scholarship having a median GPA of 66.4. To determine if this difference in mean was statistically significant, a t-test was conducted, but with a p-value of 0.52, the null hypothesis that a student's scholarship status has no effect on their GPA could not be rejected.

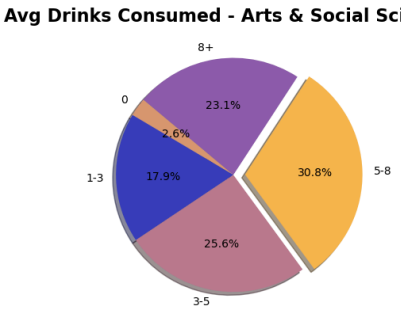


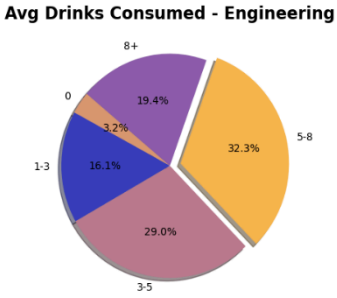
The same plot was created but using the residence type of the students. Both datasets again had a median of GPA 65, with the group living on campus having a mean GPA of 67.4 and the group living on campus having a median GPA of 66.4. The t-test performed yielded a p-value of 0.59, so the null hypothesis that a student's residence type has no effect on their GPA could not be rejected.

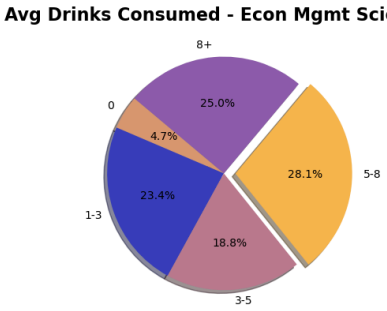


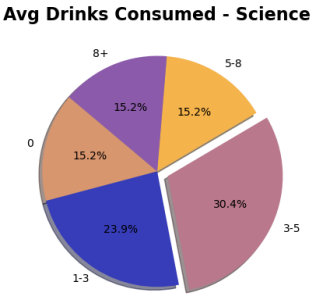
Alcohol Consumption Habits:

After analyzing the data, we found no relationship between the field of study and drinks consumed. With both categories being either categorical or pre-binned, we had to find a chart that would accurately depict the dataset without having to change the pre-binned data and to also allow the averages of the total instead of the count because the categories did not contain an even number of students. For the pie charts, a numerical column was added in order to count the amount of students in each category. This chart best depicted the distribution of students within each category separately.

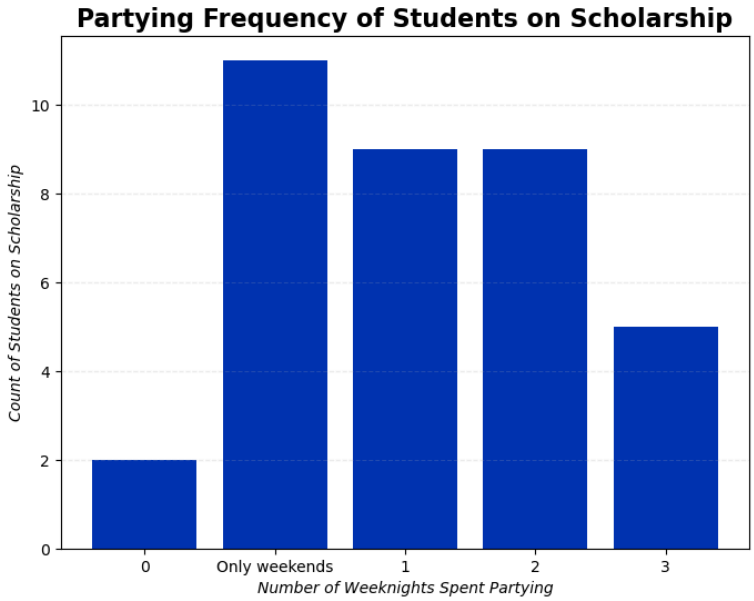


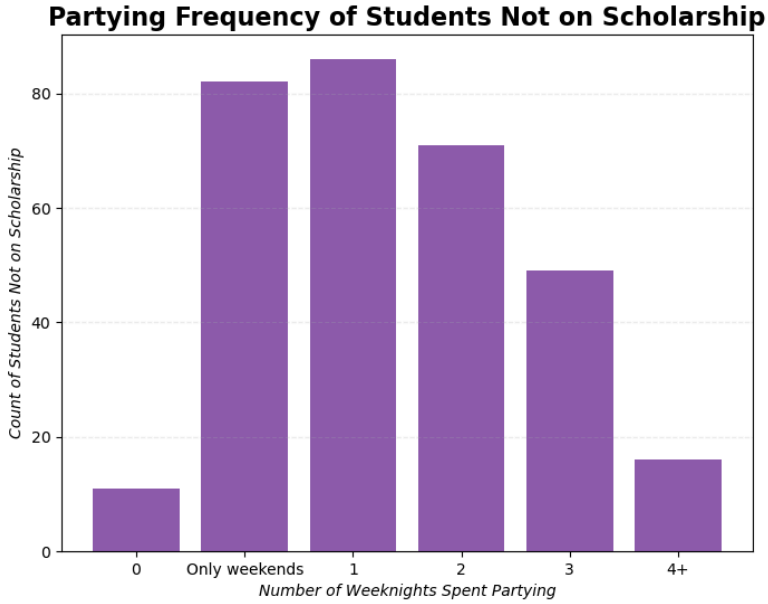






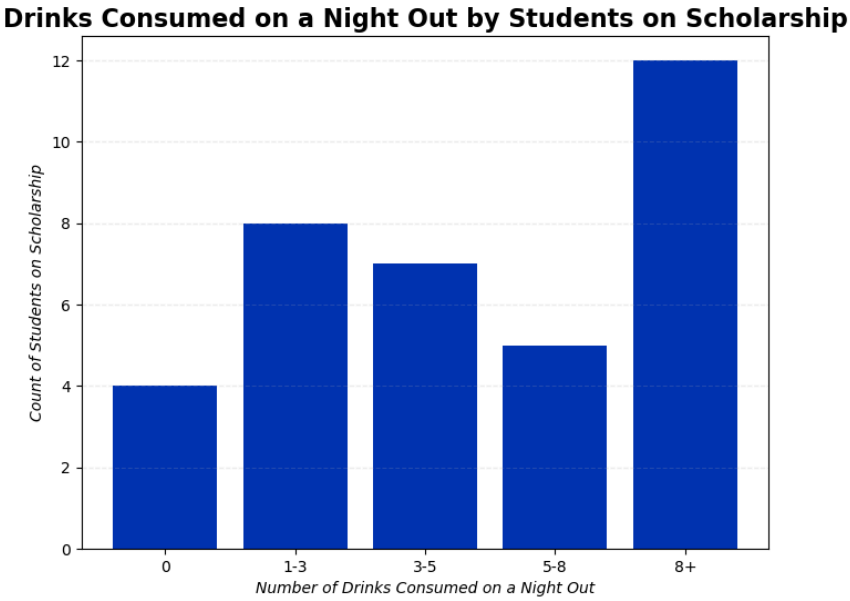
A data visualization for the average number of weeknights spent engaging in alcohol related social activity was created for both the students on scholarship and those not on scholarship.

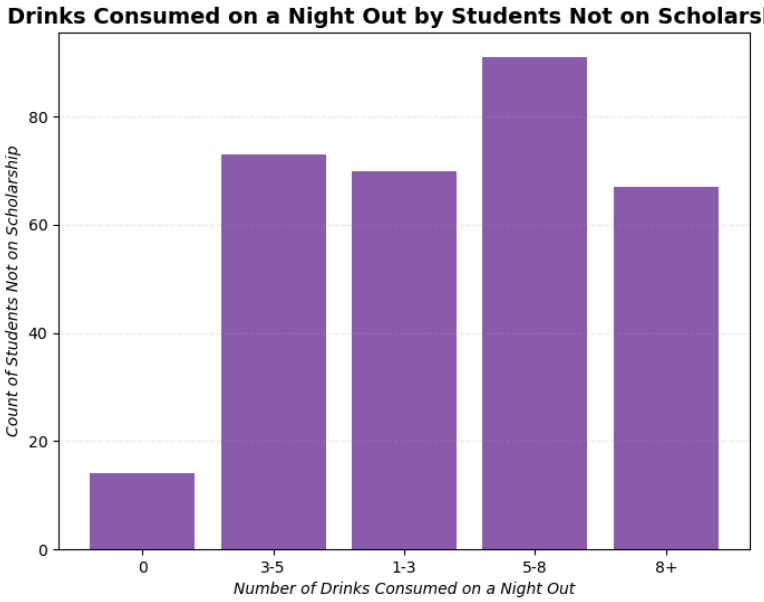




From the shape of the data, it can be seen that students not on scholarship tend to spend about a day more per week engaging in partying activity than there peers not on scholarship.

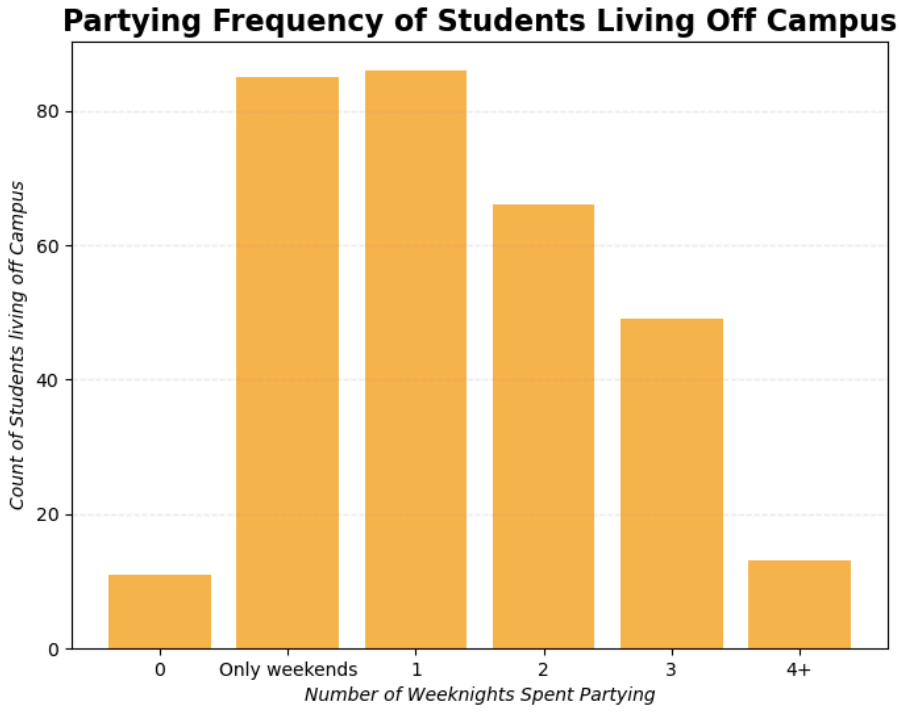
A similar visualization was generated for the typical number of drinks consumed on a night out for the two groups.

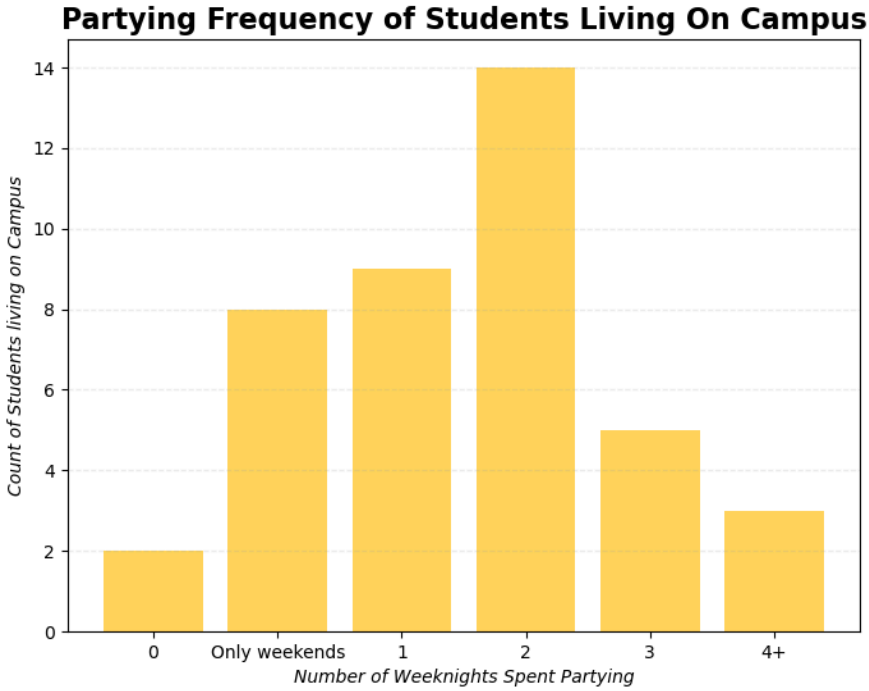


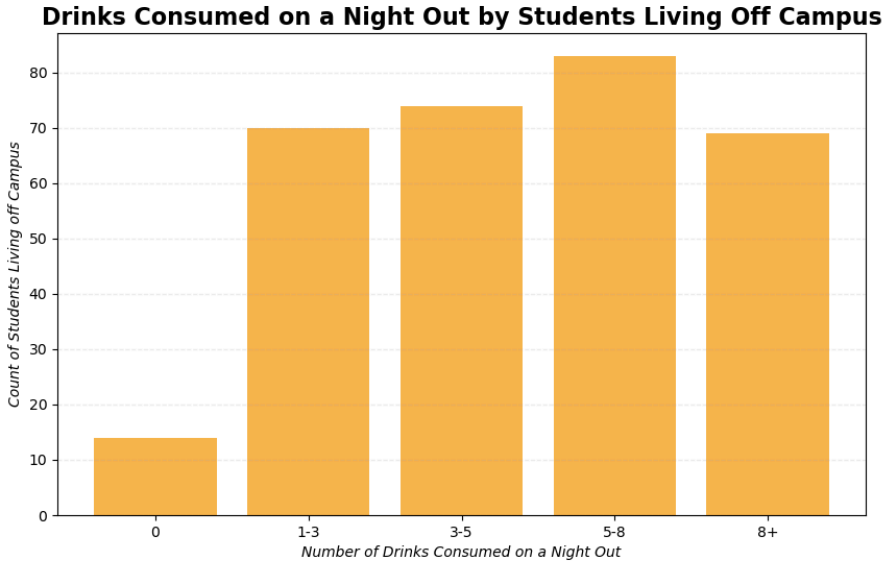


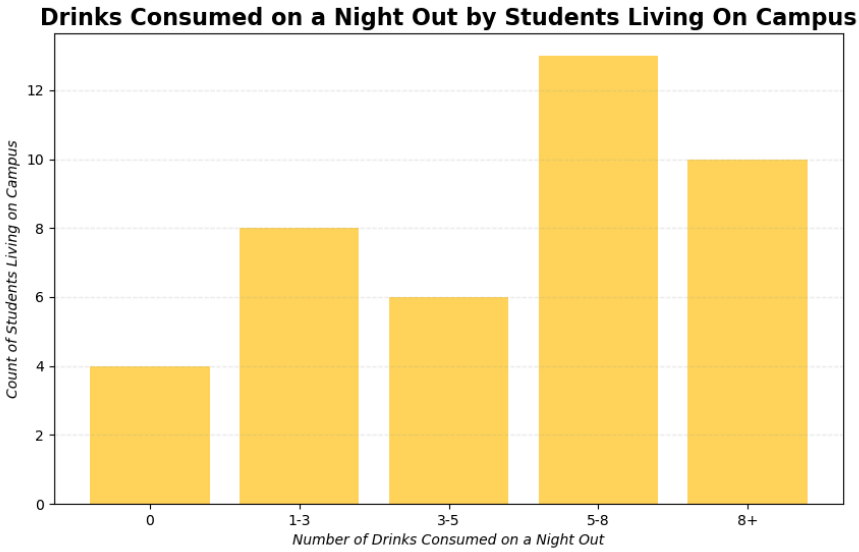
An interesting feature of note is the bimodal shape of the data for the students on scholarship. One peak in the data was for the 1-3 drinks category, with the second peak in the 8+ category. This, taken in consideration with the frequency of partying, is indicative that students on scholarship tend to binge drink more in less frequent intervals than their counterparts not on scholarship.

The same visualizations were generated for the data split based on students' residential type.


Partying Frequency of Students Living Off Campus
Count of Students living off Campus
0
20
40
60
80
0
Only weekends
1
2
3
4+
Number of Weeknights Spent Partying


Partying Frequency of Students Living On Campus
Count of Students living on Campus
0
2
4
6
8
10
12
14
0
Only weekends
1
2
3
4+
Number of Weeknights Spent Partying





The data indicates that that students living on campus tend to spend more nights partying those not living on campus, which is reasonable given their proximity. The differences in the number of drinks consumed on a night out are negligible between the two groups, especially considering the small sample size for the students living on campus.

One consideration was the overlap between the scholarship and residence statuses of students living on campus and on scholarship, and whether these two groups were nearly the same. Only 6 survey respondents belonged to both groups, leaving them functionally independent variables.

Regression Analysis:

In our regression analysis we investigated which variables within our data set had the greatest impact on student performance and utilized the findings to develop a model for predicting GPA. One of the challenges with this data set was only having two columns, High School GPA and Current GPA, with continuous data usable for regression. Additional data cleaning was performed in order to transform categorical data into numerical data.

Masks were created to filter out the rows with '0' in columns 'current_gpa' and 'year_of_study'. The data in these rows was not usable for predicting academic performance as the '0' represented students who had not yet completed a full term.

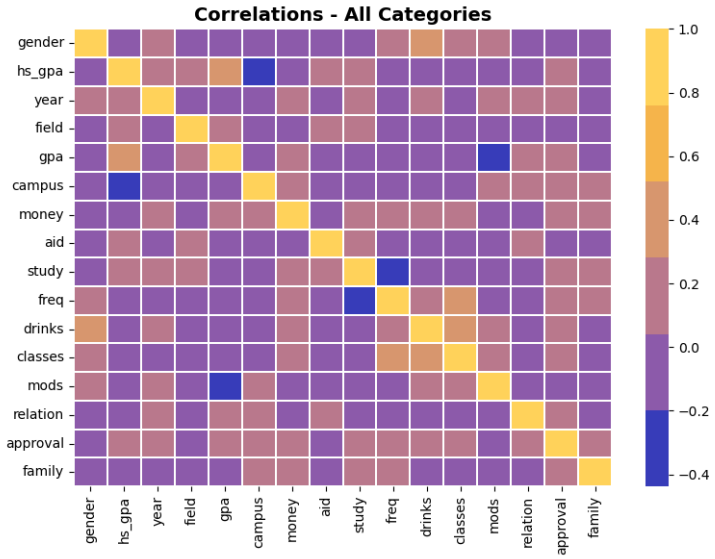
The relationships between columns 'drinks_consumed', 'study_hours', 'partying_frequency', 'modules_failed', 'classes_failed' and GPA were explored first as they were the fields believed to have the greatest impact on student outcome. Using the .replace method, each bin was encoded with a numerical value closest to the binned quantity ranging from -1 to 4.

```
cols={'0': 0,'1-3': 1,'3-5': 2,'5-8': 3,'8+': 4}

missed={'0': 0,'1': 1,'2': 2,'3': 3,'4+': 4}

new2={'Only weekends': -1,'0': 0, '1': 1, '2': 2,'3': 3,'4+': 4}

df3['drinks_consumed']=df2.drinks_consumed.replace(cols)
df3['study_hours']=df2.study_hours.replace(cols)
df3['partying_frequency']=df2.partying_frequency.replace(new2)
df3['modules_failed']=df2.modules_failed.replace(missed)
df3['classes_missed']=df2.classes_missed.replace(missed)
```

Sci-kit learn preprocessing LabelEncoder was used to transform the remaining categories into numerical data to identify and analyze relationships between all categories more effectively (GeekforGeeks).

	gender	hs_gpa	year	field	gpa	campus	money	aid	study	freq	drinks	classes	mods	relation	approval	family
0	0	76.0	2	1	72.0	1	0	0	4	-1	4	3	0	1	1	3
1	1	89.0	2	2	75.0	1	3	1	4	-1	2	4	0	0	1	3
2	1	76.0	1	0	55.0	1	0	0	2	2	4	3	0	0	1	3
3	1	89.0	2	4	84.0	1	2	0	2	3	4	2	0	1	1	3
4	0	74.0	2	1	52.0	1	0	0	2	-1	3	1	3	0	1	2

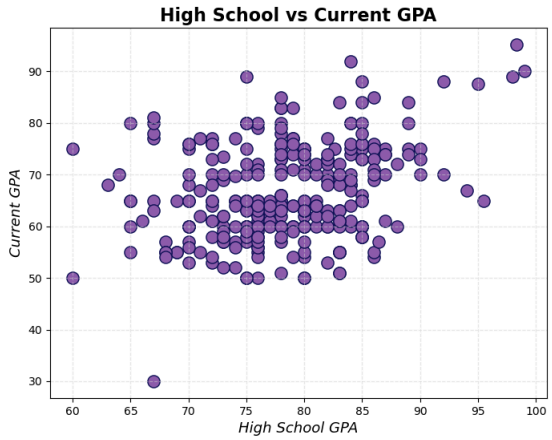
A heat map was created to illustrate the correlations between all variables. Most cells are pink or purple, indicating weak or no correlation among many variables.

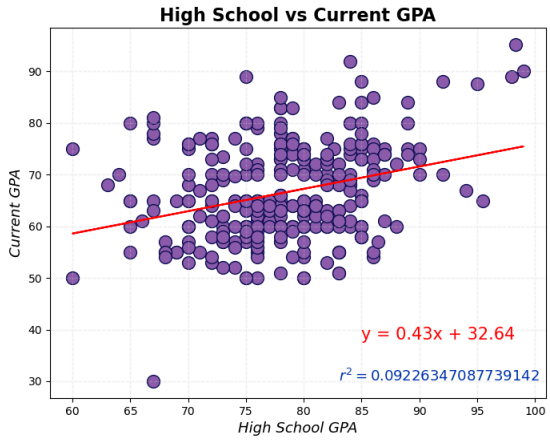


A violin plot was used to visualize the negative correlation between current GPA and the number of failed modules.

Using T-tests and ANOVA tests, we failed to reject our null hypothesis, which posits that there is no significant difference between the GPAs of students who have failed 0 modules and those who have failed 1 module.

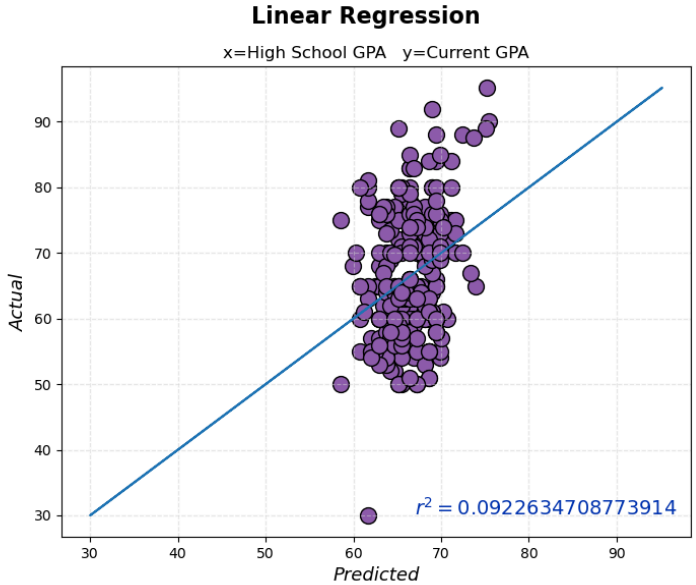
The initial linear regression indicated a very weak relationship between high school GPA and current GPA with a R^2 value of 0.09226347087739142.



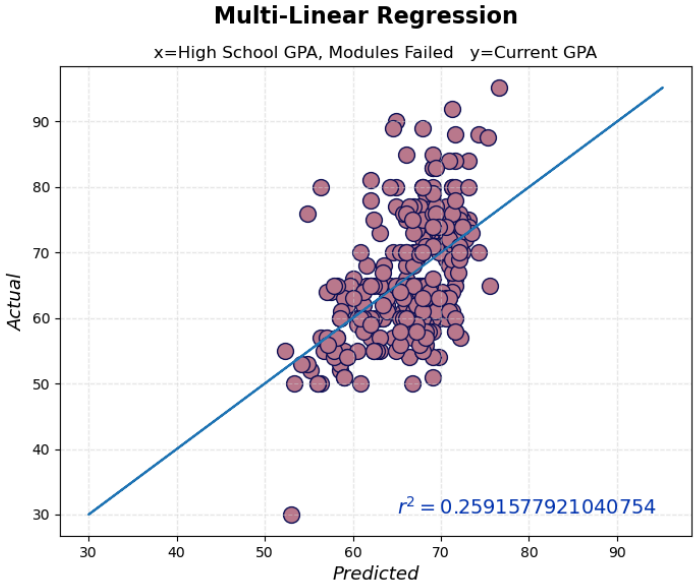


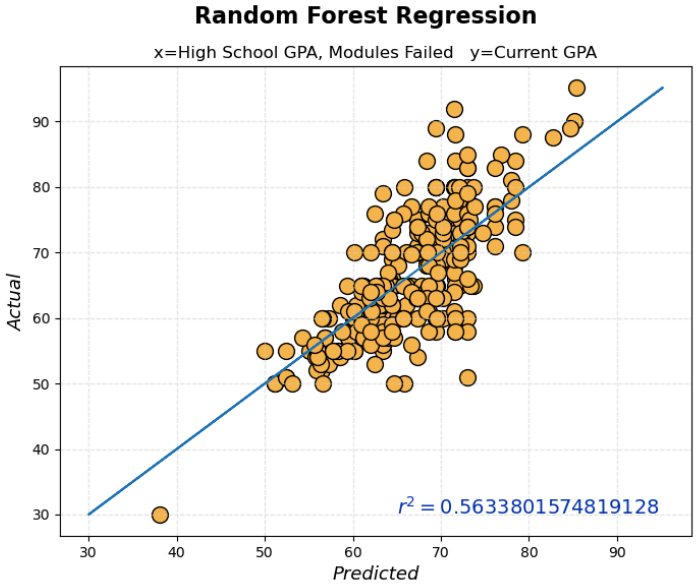
It was decided to use sci kit learn to build, train and evaluate Linear, Multiple Linear and Random Forest regression models to try to improve the accuracy of our predictions

The first model created using *scikit-learn* was a linear regression with high school GPA as the dependent variable and current GPA as the independent variable. The R^2 value for the linear regression created using *mathplotlib* and the model created using *scikit-learn* are the same, indicating that original analysis was done correctly.



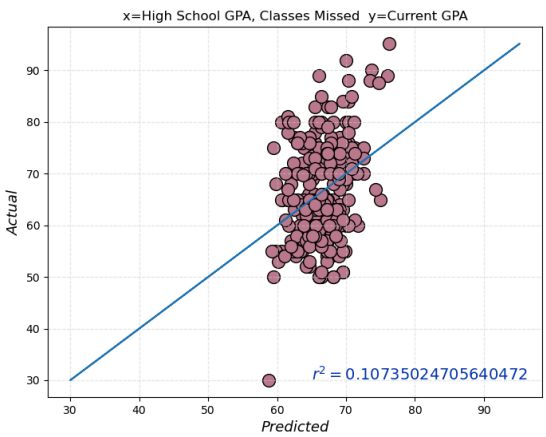
The multiple linear regression model was created next by adding a second dependent variable to the original model. Statistical testing and hypothesis testing done prior showed a significant relationship between the number of modules failed and a student's GPA. Therefore modules failed was used as the second dependent (using the numeric values representing the binned values of 0, 1, 2, 3 and 4+) . The improved R^2 value of 0.2591577921040754 shows that the multiple linear regression using high school GPA and the number of modules failed predicts a student's current GPA more accurately than a simple linear regression using only high school GPA and the number of modules failed.



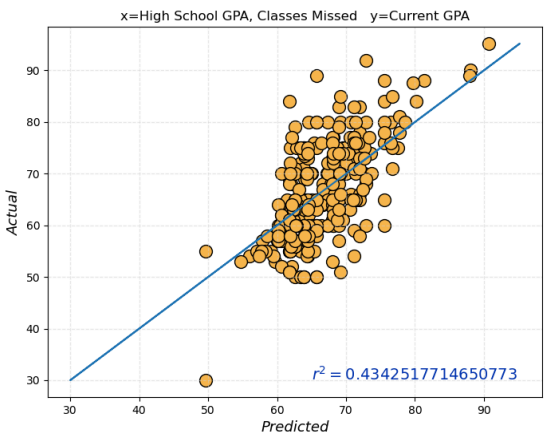


Finally a Random Forest regression model was created. This model uses an ensemble method to capture complex/non-linear relationships within the data to make more accurate predictions. Our model demonstrates this with a substantially improved R^2 value of 0.5633801574819128.

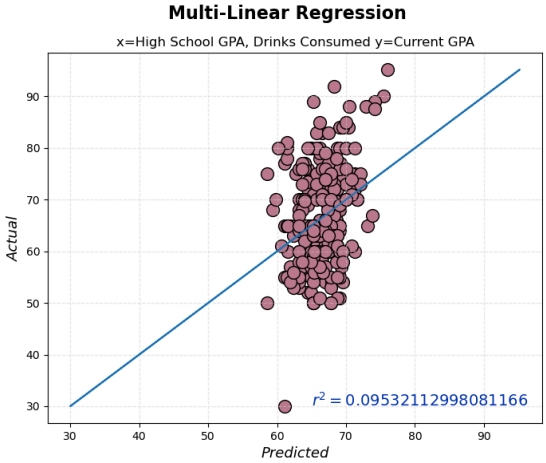
Multi-Linear Regression



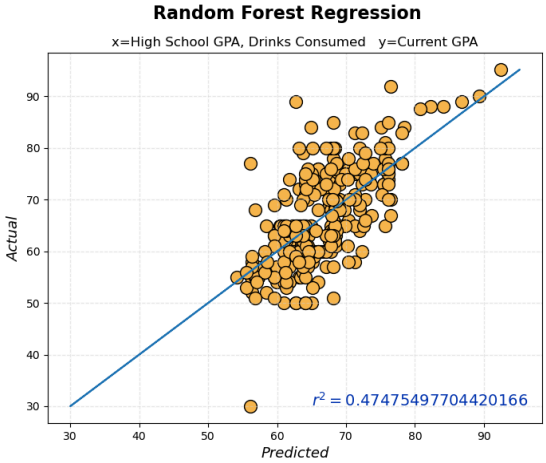
Random Forest Regression



Multi-Linear Regression

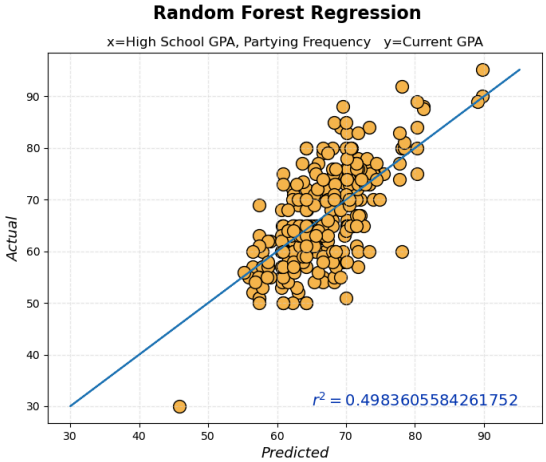


Random Forest Regression



Multi-Linear Regression

Random Forest Regression



We can hypothesize that the relationships between the variables in this dataset are indeed complex and nonlinear, given the significant improvement in R^2 values when using a Random Forest model to predict GPA. To further test this hypothesis, we developed both multiple linear and Random Forest models using the variables of classes missed, partying frequency, and drinks consumed to predict GPA.
Although linear and multiple linear regression can shed light on the relationships between variables, relying solely on these methods to predict student outcomes is insufficient. Human beings are complex, and predicting their outcomes necessitates a model like Random Forest, which can capture the intricate relationships among the variables that influence human experiences and in our case, student outcomes (Donges).

Limitations and Further Research:

The greatest limitations experienced with this project were the size of the dataset, the scope of the initial study and the method of data collection.

While the initial dataset provided 406 rows, after data cleaning and categorization, the number of datapoints under consideration for a particular question could fall significantly lower. For example, when looking at trends among majors, some of the top 4 majors only had between 30-40 data points to consider. Creating bar plots or histograms for these groups with such a limited amount of data raised some skepticism in the analysis as to whether the behaviors observed were truly representative. For example, was the shape of data in alcohol volume consumption of students on scholarship the result of a bimodal dataset or a poor sample size? The insights and conclusions derived from this team could have been bolstered by a larger dataset.

The scope of this dataset is also for only one university in South Africa. The results derived would be more applicable to the audience of this project if they were confirmed to be consistent amongst different campus backgrounds, especially those in the U.S. A recommendation would be to recreate this type of survey on more college campuses across different countries and normalize the GPAs based on the university's grade scale. The usability of the data would increase if the same patterns and correlations could be observed regardless of campus.

Finally, the largest challenge encountered during this project was the fact that data that could be numerical was reported in the original dataset as categorical, for example, on the question "how many drinks do you consume on a night out" data was given in ranges of 0, 1-3, 3-5, 5-8, and 8+. This is attributed to the method of data collection being a self-reported survey. While more difficult to administer, a more rigorous tracking of the specific number of drinks consumed on nights out over several weeks by study participants collected by an observed may not only allow for a true numeric datatype to be used in this analysis, but remove a source of bias in self-reporting.

Summary and Conclusions:

While analyzing the amount of time students spend studying per study field. We found that majority of students majoring in Economic & Management Sciences, Engineering, and Science spend 8+ hours studying, while Arts & Social Sciences majors average higher in the 3-5 study hours per week category. We chose to analyze the students studying habits to see if certain majors might have less free time to spend on partying/drinking. However, due to the sample sizes of each study field category being completely uneven, it made this hard to determine.

Works Cited:

Naude, Joshua and Bendeman, Jordan. *Effects of Alcohol on Student Performance*. Kaggle, 2024, https://www.kaggle.com/datasets/joshuanaude/effects-of-alcohol-on-student-performance.

"Set Order on sns.histplot." *Stack Overflow*, 16 Apr. 2021, https://stackoverflow.com/questions/67205522/set-order-on-sns-histplot.

Stellenbosch University. *Transcript Remarks: 1 March 2023*. 2023, https://www.sun.ac.za/english/SUInternational/Documents/Student%20fees/Transcript%20Remarks_1%20March%202023.pdf.

Donges, Niklas. "Random Forest Algorithm: How It Works and Why It Is So Effective." *Built In*, , https://builtin.com/data-science/random-forest-algorithm.

"ML | Label Encoding of Datasets in Python." *GeeksforGeeks*, https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/.