Assignment No.: _____

HELP University
university of achievers

Assignment Cover Sheet

Student Information (For group assignment, please state names of all members)		Grade/Marks
Name	**ID**	
Chin Zheng Yin	B2101086	

Module/Subject Information		Office Acknowledgement
Module/Subject Code	BDA205	
Module/Subject Name	Data Mining and Visualization	
Lecturer/Tutor/Facilitator	Ts. Dr. Yong Yoke Leng	
Due Date	21st October 2024	
Assignment Title/Topic	Assignment 1	
Intake (where applicable)	Semester 3, 2024	
Word Count	n/a	**Date/Time**

Declaration

- I/We have read and understood the Programme Handbook that explains on **plagiarism**, and I/we testify that, unless otherwise acknowledged, the work submitted herein is entirely my/our own.
- I/We declare that no part of this assignment has been written for me/us by any other person(s) except where such collaboration has been authorized by the lecturer concerned.
- I/We authorize the University to test any work submitted by me/us, using text comparison software, for instances of plagiarism. I/We understand this will involve the University or its contractors copying my/our work and storing it on a database to be used in future to test work submitted by others.

Note: 1) The attachment of this statement on any electronically submitted assignments will be deemed to have the same authority as a signed statement.
2) The Group Leader signs the declaration on behalf of all members.

Signature: Chin Zheng Yin	Date:
E-mail: B2101086@helplive.edu.my	

Assignment No.: _____

Feedback/Comments*
Main Strengths
Main Weaknesses
Suggestions for improvement

	Student acknowledge feedback/comments
Grader’s signature	Student’s signature:
Date:	Date:

Note:

1) A soft and hard copy of the assignment shall be submitted.
2) The signed copy of the assignment cover sheet shall be retained by the marker.
3) If the Turnitin report is required, students have to submit it with the assignment. However, departments may allow students up to **THREE** (3) working days after submission of the assignment to submit the Turnitin report. The assignment shall only be marked upon the submission of the Turnitin report.

*Use additional sheets if required.

Table of Contents

1.0 Introduction

1.1 Introduction

In the Student Performance Dataset, it basically emphasize on the alcohol consumption of students from secondary school. Specifically, this report mainly consists of students who take Mathematics course and Portuguese course. The datasets covers various areas that would possibly impact the alcohol consumption in students. This includes personal demographics, family background, personal lifestyle and academic result (grade). Selection of dataset is mainly motivated by the strong relationship between one's lifestyle choices and academic outcome. It is no doubt that adolescent is a crucial phase where their choices and decision will greatly impact their future. In this case, if a secondary school student gets addicted to alcohol at an early age, this will definitely affect his or her academic result and personal development negatively. Therefore, this dataset can clearly list out the statistics regarding on this matter.

1.1.1 Initial Observation

Initially, it is observed that this dataset covers many aspects of the students' life. This includes sex, age, living area, parent background, financial support, activities taken in school, family relationship, school life, study habit, academic grade and health status. In the end, everything will be linked to the final measure which is the workday alcohol consumption and weekend alcohol consumption of each student.

1.1.2 Significance of Initial Observation

These observations will be significant as they can be interpreted as factors causing alcohol consumption among secondary school students. They allow us to explore from different perspective and then find out the real cause of bad health among students. Upon understanding the true relationship between the students' alcohol consumption and academic result, schools can also take into consideration and try to educate the students on this matter. Overall, it would be clear that there are many factors that causes students to resort to alcohol which causes their health to be weaken. With these observations, effective actions can be taken at an early stage to ensure that these teenage students could have a better future ahead of them.

1.1.3 Curiosity of Selecting Dataset

Lastly, selecting this dataset is due to the curiosity of the relationship between an adolescent's personal choices and his or her academic achievement. With the Student Performance Dataset, unknown patterns can be uncovered which can greatly help in resolving this issue. For example, education of parent might be one of the factors for alcohol consumption among secondary student. This might not be obvious at first but this data, it can very much be shown. In short, this dataset can pretty much answer questions regarding on alcohol consumptions and academic result of a teenage student. With that, parties such as parents, educators and policymakers can come up with solutions which can improve the well-being and academic performance of these secondary school students.

1.2 Hypotheses Formulation/Assumptions

There are three hypotheses formulated in this report. Table 1 shows the all the hypotheses, the reason behind it and how it is related to the dataset.

Table 1

Hypothesis Formulated, Reason Behind Hypothesis and How it is Related to Dataset

Hypothesis	Implications
Students with a better family relationship would be less likely to engage in alcohol consumption	***Reason Behind Hypothesis*** For starters, a family with a good relationship provides sufficient emotional support and guidance for the children. According to Thomas et al. (2017), children who constantly receive support from their parents will have more self-esteem and self-worth. Naturally, this decreases the risk of the child having mental sickness such as depression or anxiety as they have someone they trust to lean to whenever they face any problem. On the other hand, children that suffers from mental sickness would offer find a coping mechanism outside of the family. They might resort to alcohol. Statistics has shown that the poorer the mental health, the more likely an adolescent would binge drinking (Holtes et al., 2015). Another reason would be the impact of parenting style. The way that parents treat the children is also very important for children. Authoritative parenting is much suggested in this modern era in which the parent focus on the children's feelings but also use positive discipline strategies at the same time (Pardee, 2024). This allows the children to have stronger relationship with parents as they know that their parents will always be there to listen to their children. However, implementing neglectful parenting would cause the opposite. This kind of parents would seldom check on their children. This usually causes these

	children to be rebellious and resort to use of substances at an early age. In this case, it will be very likely for them to resort to drinking during adolescent. ***How Hypothesis is Related to the Database*** There are three variables in the Student Performance Dataset that are able to test this hypothesis. The main dimension used will the "famrel". This stands for the quality of family relationship for each secondary school student. This variable literally connects with the strength of family bond in the students' family. This is an ordinal data which has a range between 1 to 5; 1 being to have a very bad relationship with family whereas 5 being to have an excellent relationship with family. This acts as the metric to analyze the family environment for each student. On the other hand, there two variables that can be the measure of the hypothesis: "dalc" and "walc". "Dalc" stands for the weekday alcohol consumption whereas "walc" stands for the weekdend alcohol consumption. This is also an ordinal data that has a range between 1 to 5. 1 being to have a low alcohol consumption whereas 5 being to have a very high alcohol consumption. These two variables can be added up and find the average of the rate of alcohol consumption on both weekday and weekends called "avg_alcohol". With the dimension and measures, we can easily analyze the correlation between the two variables. If the correlation between "famrel" and "avg_alcohol" is strong, this means the hypothesis is accepted. On the other, if the correlation between "famrel" and "avg_alcohol" are weak, this means that the hypothesis is not accepted.
Students who spend more time in studying will be less likely to engage in alcohol.	***Reason Behind Hypothesis*** Naturally, a student who spends most of his or her time studying is someone who values education a lot. Based on State (2022), students will only put effort and thrive for

something that they value. In that case, these students would not let anything distract them from learning. They know that in order to succeed, it is essential to be on their A game at all time. In that case, alcohol would be one of the things that they would try to avoid as it will cause them to lose focus.

Besides, Cargiulo (2007) also mentioned that alcohol dependance are usually associated with psychiatric conditions such as depression, panic disorder and anxiety. It would also cause other sickness such as brain damage and neurologic deficit If a student is eager to study and learn, they would be fully aware of these negative impacts that alcohol would bring to them. In fact, they would spend their time on their academic instead of thinking on doing anything that would disrupt their study rhythm. Therefore, this would be the reason why students who spend more time studying would be less likely to engage in alcohol consumption as they know that it would ruin them mentally and physically.

How Hypothesis is Related to Database

There are three variables in the Student Performance Dataset that are related to this hypothesis. The main variable would be "studytime". This is a ordinal data that ranges between 1 to 4, 1 being less than 2 hours of study time; 2 being 2 to 5 hours of study time; 3 being 5 to 10 hours of study time and 4 being more than 10 hours of study time. Alcohol consumption can be measured by using two variables: "dalc" (workday alcohol consumption) or "walc" (weekend alcohol consumption). To see the average alcohol consumption, "dalc" and "walc" can be added to calculate the mean of both variables, creating a new variable called "avg_alcohol". With these two variables, we can view the relationship between study time and alcohol consumption. If there is a strong relationship between "studytime" and "avg_alcohol", this means that the hypothesis is accepted; if

	there is a weak relationship between “studytime” and “avg_alcohol”, this means that the hypothesis is not accepted.
The presence of family support will greatly affect the students’ health	***Reason Behind Hypothesis*** Family supports includes having a good relationship with family, being able to afford a child’s needs and also having the support of parents towards a child’s decision. These are important as the lack of support of family on a child usually becomes a bearing towards a child’s achievement (Desforges & Abouchaar, 2003). This statement is also supported by Kapur (2023) who stated that parental support is needed for a child’s success. Children lacking with parental support do not have their parents to guide them during their development phase. Therefore, they would associate themselves with activities that will harm them in the long run. Also, parents that do not support their child would cause their child to be rebellious, resorting to unhealthy activities just to gain attention from his or her parents. On top of that, mental health is also a very significant aspect in health. The lack of family support would affect the students’ mental health negatively. Ong et al. (2021) stated that most people with mental illness tend to claim that it originates from their family members. This includes parental conflicts, lack of time spent with patents and also parents’ high expectation. Thus, students without the emotional support from their parents would tend to have a weaker mentality, causing them to give up easily in life. Consequently, they might resort to carrying out harmful activities such as doing drugs or drinking alcohol to make their pain go away. In the end, they would not know that they are mentally ill. ***How Hypothesis is Related to Database***

	For family support, the “famsup” variable in the dataset can be used as the dimension. It is a binary data where it only consists of values “yes” and “no”. “Yes” would mean that the student receives parental support while “No” would mean that the student do not receive parental support. To measure the students’ health, the “health” variable can be used. “Health” is an ordinal data ranging from 1 to 5. 1 would be a very bad health whereas 5 would be a very good health. To test the relationship between “famsup” and “health”, “health” can be modified into a categorical data which contains the categories “healthy” and “unhealthy”. Then the association between the combination of categories is examined. If there is a large number of the combination “yes” with “healthy” and a large number of the combination “no” with “unhealthy”, the hypothesis is accepted.

2.0 Data Mining Analysis

2.1 Data Import

The environment used to carry out explorative data analytics will be R studio. R studio is a good environment to merge datasets and also carry out data preprocessing. On top of that, complexed visualizations can be created using basic codes as well. All the codes to carry out data import process and the other initial data checks will be provided in Appendix A.

2.1.1 How Dataset is Imported into Environment

Initially, there are two separate csv format files which contain the data about student alcohol consumption. They are separated by the different courses each student take: Math and Portuguese. However, there are still students who belong to both datasets. Therefore, These two datasets will be merged with their common attributes which are “school”, “sex”, “age”, “address”, “famsize”, “Pstatus”, “Medu”, “Fedu”, “Mjob”, “Fjob”, “reason”, “nursery” and “internet”. To carry out this algorithm, “`merge()`” syntax. In the end, there should be 55 variables in total and 382 observations.

2.1.2 Challenges Encountered and How They are Resolved

There are several challenges when importing the student dataset. The first challenge would be merging two datasets together. The objective behind this step is to be able to analyze the data of students who are taking both Math and Portuguese courses. To resolve this challenge, R studio have a syntax that could easily join these datasets. For example, more complexed syntaxes such as `inner_join,left_join and outer_join` lets us to join the dataset with a unique key. However, in this case, the dataset do not have a specific unique to join but the variables and structure of variables in both datasets are the same. Therefore, we would be determining the common attributes of the dataset. Then, as shown in Figure 1 line 6, a simple `merge()` syntax can be used to join both datasets together, filtering out only the students who belong in both datasets.

The second challenge encountered would be a changing the csv file into a data frame in R studio. Incorrect file name, incorrect file format, delimiter mismatch and inconsistent headers often happens especially when we are uploading a csv file format. Therefore, when it is imported onto R studio, the file name with its format has to be double checked and keyed in correct, the separator that separates each data value must be specified and the header must be consistent. This is to make sure that the future merging process can also be carried out smoothly. If not, all the observations cannot be run onto the environment, causing missing values or inconsistent data. To resolve this, we must first examine the structure of the raw csv file and check for the delimiter and whether the header exists. In a case like this dataset, the delimiter for the values is a comma. There is also a consistent header in the dataset. Therefore, the delimiter has to be specified with syntax `sep = ","` and the header must be specified with syntax `header = TRUE.`

The third challenge encountered would be handling repeated same attributes that represents different dataset. For example, there is an attribute "G1" in the dataset of student from Math course. At the same time, there is also an attribute "G1" in the dataset of student from Portuguese course. Both attributes have the same name but represent different things. By default, R studio labels the columns of repeated attribute from Math course with ".x" while the columns of repeated attribute from Portuguese course will be labelled with ".y". In other words, "G1.x" would represent the grade for the students in their Math course whereas "G1.y" would represent the grade for the students in their Portuguese course. This is not practical as this might cause confusion when it comes to analysis in the future. To resolve this, these columns would be renamed and their course would be specified. For instance, "G1.x" would be renamed to "math_G1"; "G1.y" would be renamed to "port_G1".

The last challenge encountered would be modifying nominal variable types. It is important to make sure that the data types of each variable is correct and appropriate for upcoming EDA tasks. For example, sex in the dataset is a nominal data. However, since sex only consists of "F" and "M", the type of should be categorical data instead. On the other hand, there are also nominal data that should be converted as binary data instead. For example, the attribute “internet” only has the values “yes” and “no”. Thus, it would be more suitable to change it into binary data. This process is important so that all the data can easily be understood and be very straight forward. To resolve this issue, the syntax `as.factor()` can be used in R studio. The `as.factor()` syntax changes normal nominal data into categorical data. As for numerical data, R studio has already set the attributes with numerical values as a numerical variable type by default. Thus, no modification would be needed for numerical data.

2.1.3 Initial Data Checks Conducted

After data is successfully imported onto the environment, initial data checks are carried out to ensure the data is consistent. The first data check conducted would be data type check. This is to makes sure that the data types makes sense with the values. For example, the data type for “age” is integer; the data type for “guardian” is character, and so on. This data check is important as it ensures that data operation like calculations and filtering can be performed properly. Moreover, the data analysis such as regression models and association models would be easier since all the data is consistent. To conduct this data check, it would be conducted in R studio using the syntax “`str(<dataset>)`”.

Another initial data check is uniqueness check. This check is very important as well because it is to make sure that there are no duplicated data present. This process will also ease the analysis process as there will not be too much of the same data in the dataset. On top of that, the dataset would be clean before any other EDA tasks. To conduct this data check, it can easily be conducted in R studio using the syntax “`duplicated(<dataset>)`” while the syntax “`sum()`” can be used to view the total number of duplicated values.

Lastly, another initial data check that was conducted would be presence check. This would be to check whether are there any missing values in the dataset. If there are, actions must be taken to solve this issue. If not, analysis can then be proceeded. This check is needed as to improves data cleanliness. With missing data, inaccurate or biased results would be produced. With that, it distorts the analysis process, leading to an ineffective decision-making process. Missing data can easily be checked in R studio using the syntax “`is.na(<dataset>)`”. On top of that, the syntax “`sum()`” will be used to view the total number of missing data in the dataset.

Overall, initial data checks are important so that the data in the dataset is clean and ready to be analyzed. Checking for missing values, duplicates are the most basic steps but still the most important steps to be carried out. This is to ensure that the truthfulness of the result will be as accurate as possible, without any noisy data in the way. On top of that, checking data types is also important because this increases one’s understanding towards the dataset.

Determining the right data type for each dimension and measure is always important so that algorithms and syntaxes can be run smoothly without any errors.

2.2 Structural Investigation

In this section, it involves summarizing the structural features in the dataset. The section covers the types of all the variables and their roles. Then, it touches on the completeness of data and any metadata considered. In the end, the importance of understanding these features will be discussed. The process is still part of understanding data before any analysis happen. On another note, feature engineering will also be mentioned in this section. Codes written relevant to this section will be mentioned in Appendix B.

2.2.1 Types of Variables and Their Roles

As mentioned earlier, there were initially two datasets: student data from Math course and student data from Portuguese course, each dataset contains 33 variables. Upon merging these two datasets using the common variables, there are still 20 repeated variables available. The name of the variables may be the same but the both represent data from the student's Math course and Portuguese course separately. For example, a student's grade (G3) for Math course is 15 while his or her grade (also G3) for Portuguese course is 13. To solve this, as mentioned in the previous section. Repeated variables will be labelled "math_" or "port_". "math_" will be the variables for students in Math class while "port_" will be variables for students in Portuguese class. In the end, there will are 53 variables initially.

However, feature engineering is also carried out to improve the analysis of the target variable. There are 7 new variables created that are derived from the existing variables. For starters, a new column called "avg_alcohol" is created. This is to find the average rate of alcohol consumption of the students in the dataset. At first, "math_Dalc" and "port_Dalc" is added to find the average, creating a new variable called "avg_Dalc". Then, "math_Walc" and "port_Walc" is added to find the average to create the variable "avg_Walc". In the end, the mean of "avg_Dalc" and "avg_Walc" is calculated to find "avg_alcohol". Other than that, a new variable "avg_famrel" is also created. This is created using the average number between "math_famrel" and "port_famrel". With that, the average family relationship for student who is taking both Math and Portuguese course can be used as a new feature variable. Lastly, a new column called "avg_studytime" is also created. This is to find the average time students take to study for both Math and Portuguese course.

Other than creating new feature variables for numerical data, new categorical feature variable is created as well. Two new categorical variables: "math_health_cateogry" and "port_health_category" is created by deriving the "math_health" and "port_health" variables.

Initially, these variables are numerical ordinal data. Thus, they are derived to create categories of “healthy” and “unhealthy”. Rank of 1 and 2 would be categorized as unhealthy whereas rank of 3, 4 and 5 would be categorized as unhealthy. Thus, in the end, there will be 60 variables in the dataset in total.

Table 2 summarizes all the variables in the dataset, together with their respective types and roles.

Table 2

Types of all Variables in the Dataset and their Roles

Variables	Type	Roles
math_schoolsup / port_schoolsup	Categorical	Feature
math_paid / port_paid	Categorical	Feature
math_activities / port_activities	Categorical	Feature
Nursery	Categorical	Feature
math_higher / port_higher	Categorical	Feature
Internet	Categorical	Feature
math_romantic / port_romantic	Categorical	Feature
school	Categorical	Feature
sex	Categorical	Feature
address	Categorical	Feature
famsize	Categorical	Feature
Pstatus	Categorical	Feature
Mjob	Categorical	Feature
Fjob	Categorical	Feature
reason	Categorical	Feature
math_guardian / port_guardian	Categorical	Feature
math_famsup / port_famsup	Categorical	Feature
age	Numerical	Feature
Medu	Numerical	Feature
Fedu	Numerical	Feature
math_traveltime / port_traveltime	Numerical	Feature
math_studytime / port_studytime	Numerical	Feature
math_failures / port_failures	Numerical	Feature
math_famrel / port_famrel	Numerical	Feature
math_freetime / port_freetime	Numerical	Feature

math_goout / port_goout	Numerical	Feature
math_Dalc / port_Dalc	Numerical	Feature
math_Walc / port_Walc	Numerical	Feature
math_absences / port_absences	Numerical	Feature
math_G1 / port_G1	Numerical	Feature
math_G2 / port_G2	Numerical	Feature
avg_studytime	Numerical	Feature
avg_Dalc	Numerical	Feature
avg_Walc	Numerical	Feature
avg_famrel	Numerical	Feature
math_G3 / port_G3	Numerical	Feature
math_health / port_health	Numerical	Feature
math_health_category / port_health_category	Categorical	Target
avg_alcohol	Numerical	Target

Note. Table 2 shows all the variables available in the dataset. It is seen that there are 20 variables that are labelled with "math_" and "port_" separately. They grouped as the same variable as they have a same structural feature and same definition as well. The only thing that is differing the attributes would be the type of course.

2.2.2 Completeness of Data

In the dataset, many initial data checks are carried out to check the completeness of dataset. Fortunately, no missing values or duplicated data are present. On top of that, all data types are appropriate and are ready to be analyzed. Table 3 below shows a summary table of the completeness of each variable, showing the number of missing values, duplicates and inconsistencies

Table 3

Summary Table for Completeness of Data

Variables	Missing Data	Inconsistencies
math_schoolsup / port_schoolsup	None	None
math_paid / port_paid	None	None
math_activities / port_activities	None	None
Nursery	None	None
math_higher / port_higher	None	None
Internet	None	None
math_romantic / port_romantic	None	None

school	None	None
sex	None	None
address	None	None
famsize	None	None
Pstatus	None	None
Mjob	None	None
Fjob	None	None
reason	None	None
math_guardian / port_guardian	None	None
math_famsup / port_famsup	None	None
age	None	None
Medu	None	None
Fedu	None	None
math_traveltime / port_traveltime	None	None
math_studytime / port_studytime	None	None
math_failures / port_failures	None	None
math_famrel / port_famrel	None	None
math_freetime / port_freetime	None	None
math_goout / port_goout	None	None
math_Dalc / port_Dalc	None	None
math_Walc / port_Walc	None	None
math_absences / port_absences	None	None
math_G1 / port_G1	None	None
math_G2 / port_G2	None	None
avg_studytime	None	None
avg_Dalc	None	None
avg_Walc	None	None
avg_famrel	None	None
math_G3 / port_G3	None	None
math_health / port_health	None	None
math_health_category / port_health_category	None	None
avg_alcohol	None	None
Duplicated Entries	None	

Note. Table 3 shows the number of missing and inconsistent values in each variable of the dataset. In the end, the number of duplicated entries is also stated.

2.2.3 Metadata Considered

There metadata considered for the Student Performance Dataset is defining the variables. There are many variables that are phrased in short-form to decrease the naming redundancy. However, this might cause confusion for users as they do not know what some of the attributes mean. For example, the attribute "higher" would seem confusing in the dataset as it does not see relevant. Actually, "higher" in the dataset basically means whether the student would like to pursue on a higher education. Therefore, it is clear that without specifying the definition of each variable, some attributes might be confusing to the users. As a result, Table 4 is constructed to show all the definition of each variable.

Table 4

Definition of All Variables in the Dataset

Variable	Definition
math_schoolsup / port_schoolsup	Extra educational support
math_famsup / port_famsup	Family educational support
math_paid / port_paid	Extra paid classes within the course subject
math_activities / port_activities	Extra-curricular activities
Nursery	Attended nursery school
math_higher / port_higher	Wants to take a higher education
Internet	Internet access at home
math_romantic / port_romantic	In a romantic relationship
school	Students' school (GP: Gabriel Pereira, MD: Mousinho da Silveria
sex	Students' sex (F: female, M: male)
address	Students' address (U: urban, R: rural)
famsize	Family size (LE3: less than 3, GT3: greater than 3)
Pstatus	Parents' cohabitation status (T: together, A: apart)
Mjob	Mother's job
Fjob	Father's job
reason	Reason of choosing the school
math_guardian / port_guardian	Students' guardian
age	Students' age (between 15 to 22)

Medu	Mother’s education
Fedu	Father’s education
math_traveltime / port_traveltime	Home to school travel time (1: less than 15 minutes, 2: 15 to 30 minutes, 3: 30 minutes to 1 hour, 4: more than 1 hour)
math_studytime / port_studytime	Weekly study time (1: less than 2 hours, 2: 2 to 5 hours, 3: 5 to 10 hours, 4: more than 10 hours)
math_failures / port_failures	Number of past class failures
math_famrel / port_famrel	Quality of family relationship (1: very bad to 5: excellent)
math_freetime / port_freetime	Free time after school (1: very low to 5: very high)
math_goout / port_goout	Going out with friends (1: very low to 5: very high)
math_Dalc / port_Dalc	Workday alcohol consumption (1: very low to 5: very high)
math_Walc / port_Walc	Weekend alcohol consumption (1: very low to 5: very high)
math_health / port_health	Students’ current health status (1: very bad to 5: very good)
math_absences / port_absences	Number of school absences
math_G1 / port_G1	First period grade
math_G2 / port_G2	Second period grade
math_G3 / port_G3	Final grade
avg_Dalc	Average workday alcohol consumption of student from Math and Portuguese course.
avg_Walc	Average weekend alcohol consumption of student from Math and Portuguese course.
avg_famrel	Average family relationship of a student who is studying both Math and Portuguese
avg_alcohol	Average alcohol consumption in a week of student from Math and Portuguese course.
math_health_category / port_health_category	Health status if student from Math and Portuguese course (“Healthy”, “Unhealthy”)

Note. Table 4 shows the definition of all data in the dataset.

2.2.3 How Understanding Structural Feature Influence Approach to Analysis

Overall, having a deep understanding towards the structural feature in the dataset will influence the analytical technique used in the future. This is because different data types is suitable for certain techniques. With that, we can know which attribute can be used for certain specific analytical techniques. For example, binary data can be used to carry out classification or association technique. On the other hand, numerical data can be used for regression models or correlation analysis. Therefore, after understanding the variables types, the right analytical technique can be chosen to derive a more accurate result.

Understanding the structural features of the dataset will also help in ensuring model interpretability. Model interpretability is important so that we are able to know exactly which variable we need for analysis. This is why definition of the variables themselves have to be specified so that confusion would not happen. For example, "avg_alcohol" is the final alcohol consumption and the target variable. If it is not specified, other variables like "math_Dalc" or "port_Dalc" might be used instead, leading to inaccurate results. On top of that, with a good model interpretability, the variables can be easily explained to others.

Last but not least, understanding structural features of the dataset can help in identification of variable roles. Variables' roles are important so that we know exactly which variable will influence the final target. For instance, the target variable "Dalc" or "Walc" will be influenced by the feature variables such as "famsup", "higher" and so much more. It forms a guideline for us when we are performing the analysis process. We would know which variables will be the measure or depending variable and which variable will be the dimension or independent variable. Thus, this makes sure that we will not go out of scope during the analysis process.

2.3 Quality Investigation

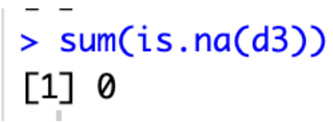
In this section, it would be the last step of the data preprocessing process. This step involves checking for any noisy data and trying to handle them. This step is very important as it ensures the dataset is clean and ready to be analyzed. With a clean dataset, analysis process can be carried out smoothly. Besides, results produced would be more accurate and effective. On that note, all the codes written relevant to this section will be stated in Appendix C.

2.3.1 Checking for Missing Data and Method of Handling

For missing data, the syntax `is.na()` and `sum(is.na())` can be used to view the missing data variables and also the total number of missing data in the whole dataset. Figure 1 below will later on show the final result of total number of missing data in the dataset.

Figure 1

Result of "sum(is.na())" Syntax



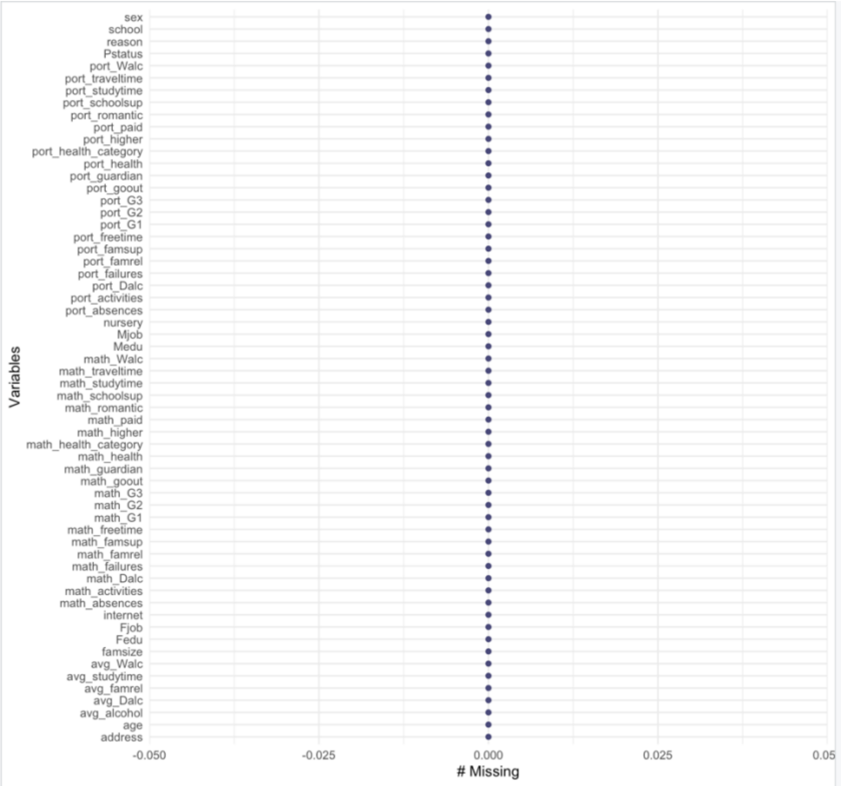
```
> sum(is.na(d3))
[1] 0
```

Note. Figure 1 shows the result showing the sum of missing data in the dataset. The result is zero, indicating that there are no missing values.

Other than the console output displayed by R Studio, visualizations are also created to check for missing values. A heatmap is produced as well to see the area where there are the greatest number of missing data. This will help us to clearly see the variable that needs to be handled regarding on missing data. The visualization will be shown in Figure 2 whereas Figure 3 will display the heatmap produced to show area of missing data.

Figure 2

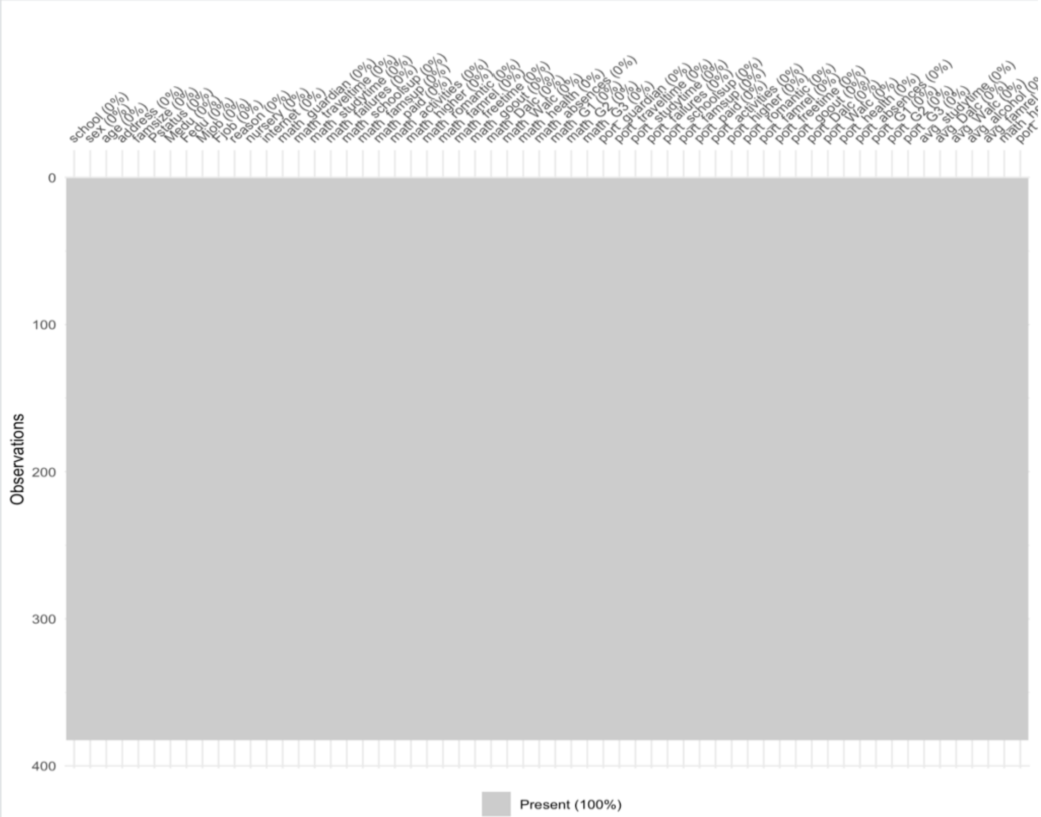
Result of Missing Value in the Dataset



Note. Figure 2 shows the number of missing data in the dataset. Clearly, there are no missing data in the dataset.

Figure 3

Heatmap Produced to Show the Area of Missing Data in the Dataset



Note. Figure 3 depicts the heatmap to show which part in the dataset has more missing values. In this case, it shows that there are no missing values in the dataset.

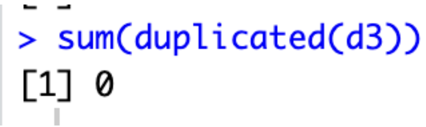
Based on the output and visualizations produced in Figure 1, Figure 2 and Figure 3, there are no missing values available in the dataset. Therefore, no measures need to be taken to handle it.

2.3.2 Checking for Duplicated Data and Method of Handling

To check for duplicated data, the syntax `duplicated()` and `sum(duplicated())` can be used to see the duplicated data in the dataset and the total number of it. Figure 4 shows the final result of the syntax.

Figure 4

Result of "sum(duplicated())" Syntax in R Studio



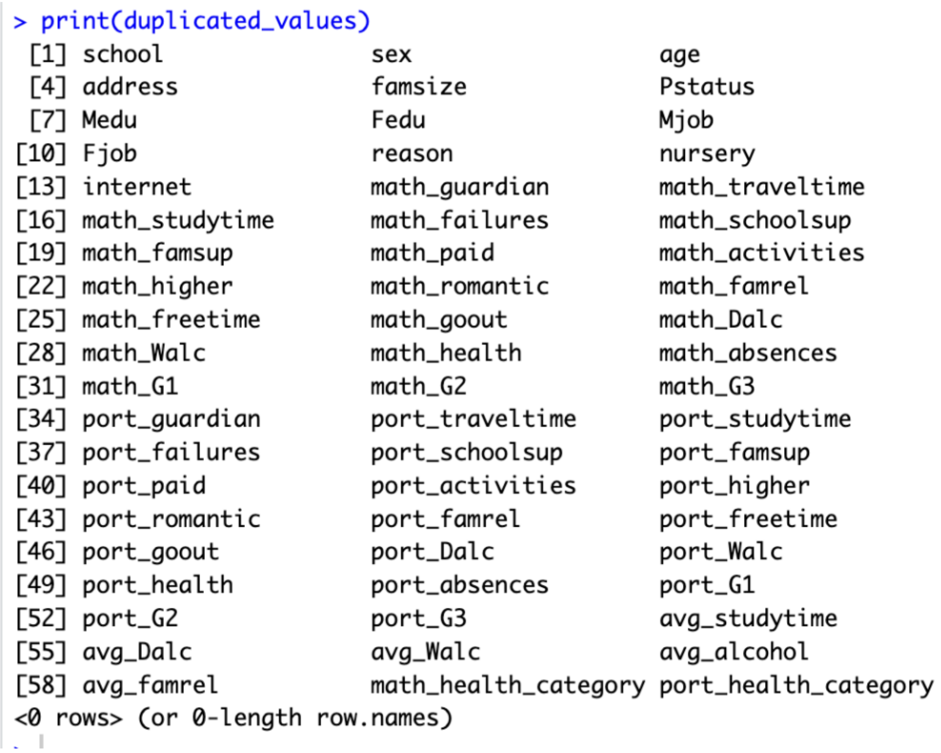
```
> sum(duplicated(d3))
[1] 0
```

Note. Figure 4 shows the total number of duplicated data in the dataset. From the result, it is shown that there are no duplicated data.

To double check whether are there any duplicated rows in the dataset, a new data frame called “duplicated_values” is created. This data frame is created and only consists of the duplicated rows. With this data frame, we can see which data entry has been duplicated and then find ways to handle it. After the data frame is created, it is then viewed. Figure 5 the result of the data frame “duplicated_value”.

Figure 5

Result of Values in Data Frame “duplicated_values”



```
> print(duplicated_values)
 [1] school               sex                  age                 
 [4] address              famsize              Pstatus             
 [7] Medu                 Fedu                 Mjob                
[10] Fjob                 reason               nursery             
[13] internet             math_guardian        math_traveltime     
[16] math_studytime       math_failures        math_schoolsup      
[19] math_famsup          math_paid            math_activities     
[22] math_higher          math_romantic        math_famrel         
[25] math_freetime        math_goout           math_Dalc           
[28] math_Walc            math_health          math_absences       
[31] math_G1              math_G2              math_G3             
[34] port_guardian        port_traveltime      port_studytime      
[37] port_failures        port_schoolsup       port_famsup         
[40] port_paid            port_activities      port_higher         
[43] port_romantic        port_famrel          port_freetime       
[46] port_goout           port_Dalc            port_Walc           
[49] port_health          port_absences        port_G1             
[52] port_G2              port_G3              avg_studytime       
[55] avg_Dalc             avg_Walc             avg_alcohol         
[58] avg_famrel           math_health_category port_health_category
<0 rows> (or 0-length row.names)
```

Note. Figure 5 depicts the result of the data frame created. It is shown that there are no observations in this data frame, indicating that there are no missing values in the whole dataset.

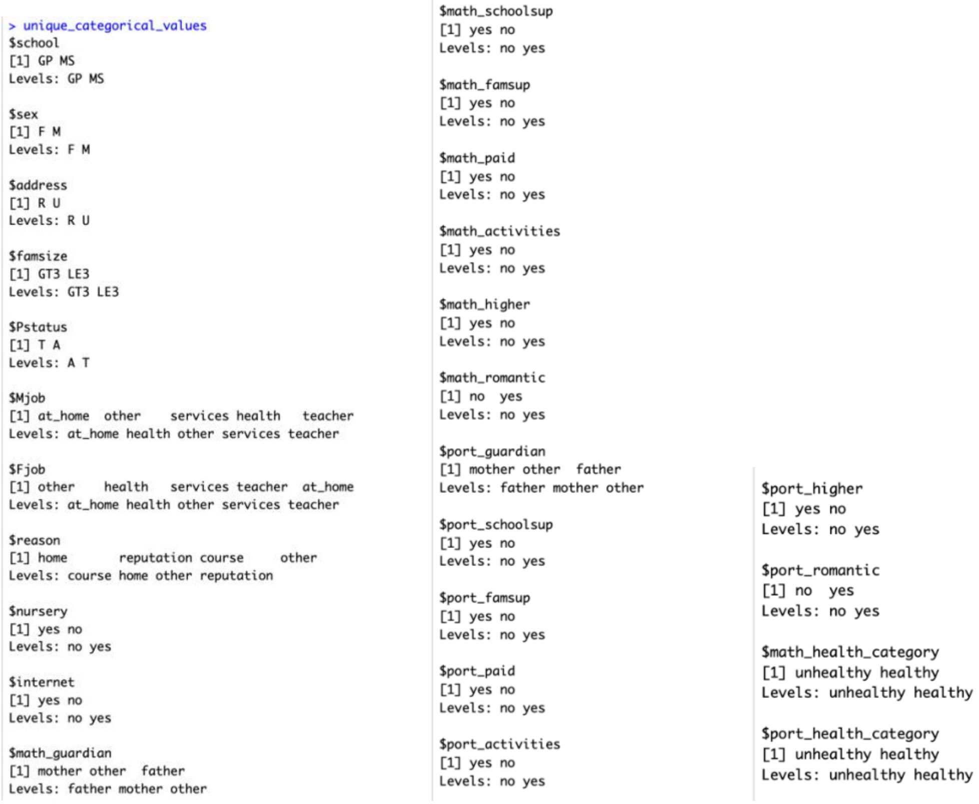
Since there are no duplicated data in the dataset, no actions needed to be taken to handle this issue.

2.3.3 Checking for Unique Values and Method of Handling

Checking on unique values usually applies on variables with specific number of levels. For example, sex in the dataset has only 2 levels: “F” and “M”. Therefore, it would not be unique anymore if there is an additional level present. The whole data would not make sense. In fact, it might cause skewness in the graph as well. Therefore, this matter needs to be checked and handled before the dataset is ready for analysis. Figure 6 shows the result of all levels of categorical data in the dataset. With that, we can check whether are there any unique and unnecessary values present in the variables

Figure 6

Results to Show all Levels of Categorical Data in the Dataset



```
> unique_categorical_values
$school
[1] GP MS
Levels: GP MS

$sex
[1] F M
Levels: F M

$address
[1] R U
Levels: R U

$famsize
[1] GT3 LE3
Levels: GT3 LE3

$Pstatus
[1] T A
Levels: A T

$Mjob
[1] at_home  other    services health   teacher
Levels: at_home health other services teacher

$Fjob
[1] other    health   services teacher  at_home
Levels: at_home health other services teacher

$reason
[1] home       reputation course     other
Levels: course home other reputation

$nursery
[1] yes no
Levels: no yes

$internet
[1] yes no
Levels: no yes

$math_guardian
[1] mother other  father
Levels: father mother other
```

```
$math_schoolsup
[1] yes no
Levels: no yes

$math_famsup
[1] yes no
Levels: no yes

$math_paid
[1] yes no
Levels: no yes

$math_activities
[1] yes no
Levels: no yes

$math_higher
[1] yes no
Levels: no yes

$math_romantic
[1] no  yes
Levels: no yes

$port_guardian
[1] mother other  father
Levels: father mother other

$port_schoolsup
[1] yes no
Levels: no yes

$port_famsup
[1] yes no
Levels: no yes

$port_paid
[1] yes no
Levels: no yes

$port_activities
[1] yes no
Levels: no yes
```

```
$port_higher
[1] yes no
Levels: no yes

$port_romantic
[1] no  yes
Levels: no yes

$math_health_category
[1] unhealthy healthy
Levels: unhealthy healthy

$port_health_category
[1] unhealthy healthy
Levels: unhealthy healthy
```

Note. Figure 6 depicts all the levels in the categorical data in the dataset. From what is shown, there are no unique values that are needed to be handled. All the variables are consistent and clean.

As a result, the dataset is clean and ready to be analyzed. There are no missing values, duplicated data or unique values that will skew the graph. In this process, no actions are needed to handle these issues. Thus, we can proceed to the analysis process.

3.0 Report Component

This section typically covers the analysis of the dataset. Key statistics and other interesting information are found and discussed in this section. Throughout the whole analysis process, it would be revolving around the hypothesis generated in the earlier section. In the end, it would be stated whether they are accepted or rejected. Before the report ends, all the key findings will be summarized and some recommendations will be given to conduct a better analysis in the future.

3.1 Content Investigation

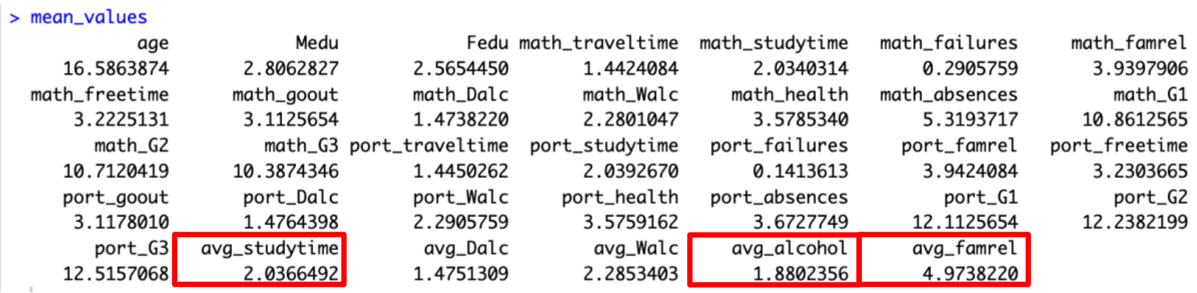
This subsection identifies the statistics of the overall data. However, the variables that are associated with the hypothesis will be focused more. In this report, the target variables would be “avg_alcohol”, “math_health_category” and “port_health_category”. On the other hand, the feature variables that will be used to see how they influence the target variable would be “famrel”, “avg_studytime”, “math_famsup” and “port_famsup”. Codes relevant to this section will be included in Appendix D.

3.1.1 Key Statistics in Dataset

The key statistics in the dataset can be easily calculated in R studio. The key statistics include mean, median, standard deviation and range of the data. Typically, this section mainly focuses on the numerical variables. The spread of data points for each variable is different and this section ought to uncover every one of them. Therefore, a new data frame “numerical_data” in R studio is created to store all the numerical data to better analyze the all the numerical data in the dataset all at once. After that, this data frame is then used to calculate the key statistics: mean, median, standard deviation and range of each variable. Figure 7 to Figure 10 shows each of the statistics calculated.

Figure 7

Results of Calculated Mean from Each Numerical Data

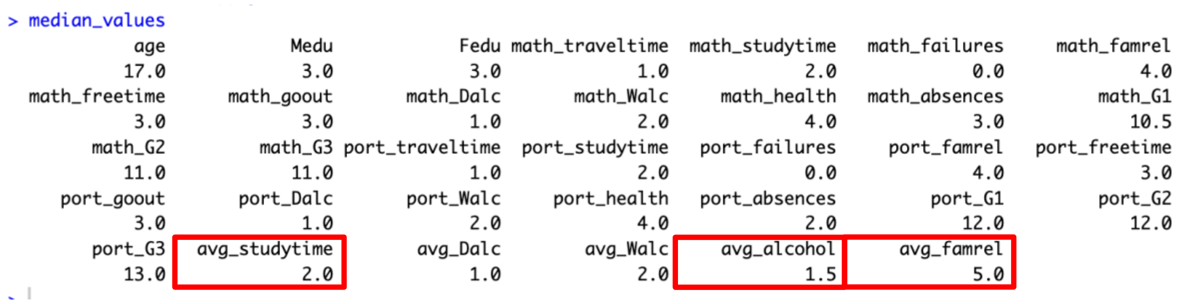


```
> mean_values
            age            Medu            Fedu math_traveltime  math_studytime   math_failures     math_famrel 
     16.5863874       2.8062827       2.5654450       1.4424084       2.0340314       0.2905759       3.9397906 
  math_freetime      math_goout       math_Dalc       math_Walc     math_health   math_absences         math_G1 
      3.2225131       3.1125654       1.4738220       2.2801047       3.5785340       5.3193717      10.8612565 
        math_G2         math_G3 port_traveltime  port_studytime   port_failures     port_famrel   port_freetime 
     10.7120419      10.3874346       1.4450262       2.0392670       0.1413613       3.9424084       3.2303665 
     port_goout       port_Dalc       port_Walc     port_health   port_absences         port_G1         port_G2 
      3.1178010       1.4764398       2.2905759       3.5759162       3.6727749      12.1125654      12.2382199 
        port_G3   avg_studytime        avg_Dalc        avg_Walc     avg_alcohol      avg_famrel 
     12.5157068       2.0366492       1.4751309       2.2853403       1.8802356       4.9738220 
```

Note. Figure 7 shows all the mean values from the numerical data in the dataset. The feature and target variables are highlighted using the red box.

Figure 8

Results of Calculated Median from Each Numerical Data



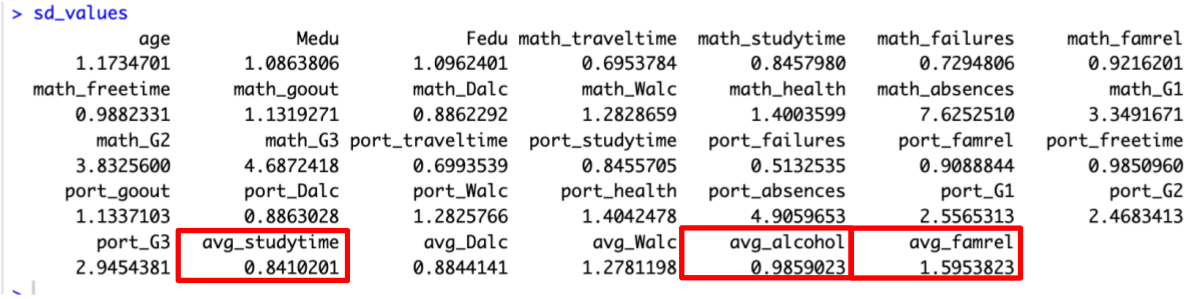
```
> median_values
            age            Medu            Fedu math_traveltime  math_studytime   math_failures     math_famrel 
           17.0             3.0             3.0             1.0             2.0             0.0             4.0 
  math_freetime      math_goout       math_Dalc       math_Walc     math_health   math_absences         math_G1 
            3.0             3.0             1.0             2.0             4.0             3.0            10.5 
        math_G2         math_G3 port_traveltime  port_studytime   port_failures     port_famrel   port_freetime 
           11.0            11.0             1.0             2.0             0.0             4.0             3.0 
     port_goout       port_Dalc       port_Walc     port_health   port_absences         port_G1         port_G2 
            3.0             1.0             2.0             4.0             2.0            12.0            12.0 
        port_G3   avg_studytime        avg_Dalc        avg_Walc     avg_alcohol      avg_famrel 
           13.0             2.0             1.0             2.0             1.5             5.0 
```

Note. Figure 8 shows all the median values from the numerical data in the dataset. The feature and target variables are highlighted using the red box.

From Figure 7 and Figure 8, it is shown that mean and median both the feature and target variables are do not differ much. Therefore, there might be no extreme outliers present in these relevant feature and target variables that will directly affect the value of the mean calculated.

Figure 9

Results of Calculated Standard Deviation Value from Each Numerical Data



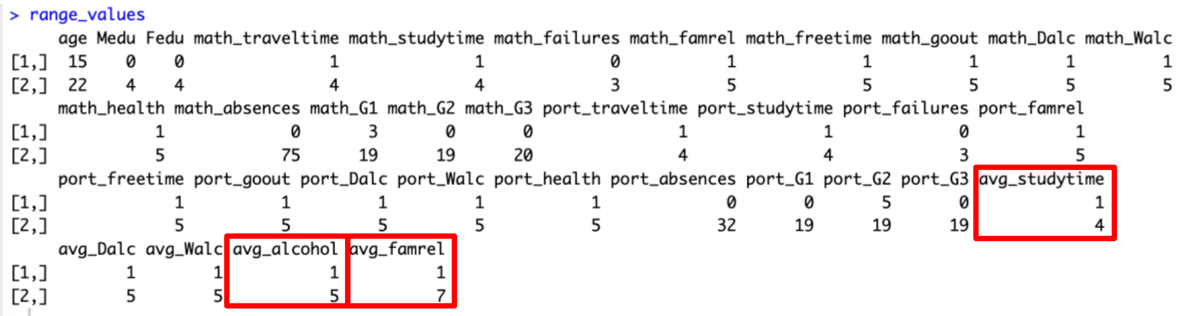
```
> sd_values
            age            Medu            Fedu math_traveltime  math_studytime   math_failures     math_famrel 
      1.1734701       1.0863806       1.0962401       0.6953784       0.8457980       0.7294806       0.9216201 
  math_freetime      math_goout       math_Dalc       math_Walc     math_health   math_absences         math_G1 
      0.9882331       1.1319271       0.8862292       1.2828659       1.4003599       7.6252510       3.3491671 
        math_G2         math_G3 port_traveltime  port_studytime   port_failures     port_famrel   port_freetime 
      3.8325600       4.6872418       0.6993539       0.8455705       0.5132535       0.9088844       0.9850960 
     port_goout       port_Dalc       port_Walc     port_health   port_absences         port_G1         port_G2 
      1.1337103       0.8863028       1.2825766       1.4042478       4.9059653       2.5565313       2.4683413 
        port_G3   avg_studytime        avg_Dalc        avg_Walc     avg_alcohol      avg_famrel 
      2.9454381       0.8410201       0.8844141       1.2781198       0.9859023       1.5953823 
```

Note. Figure 9 shows all the standard deviation values from the numerical data in the dataset. The feature and target variables are highlighted using the red box.

Through Figure 9, we can clearly see that the standard deviation of the relevant feature variables “avg_studytime” and “avg_famrel” together with one of the target variables “avg_alcohol” has a low standard deviation which is lesser than one. This means that the data point in these variables are not widely spread and relatively consistent. The distribution of these variables are more tightly packed and not far from the line of best fit.

Figure 10

Results of Calculated Range Value from Each Numerical Data



```
> range_values
     age Medu Fedu math_traveltime math_studytime math_failures math_famrel math_freetime math_goout math_Dalc math_Walc
[1,]  15    0    0               1              1             0           1             1          1         1         1
[2,]  22    4    4               4              4             3           5             5          5         5         5
     math_health math_absences math_G1 math_G2 math_G3 port_traveltime port_studytime port_failures port_famrel
[1,]           1             0       3       0       0               1              1             0           1
[2,]           5            75      19      19      20               4              4             3           5
     port_freetime port_goout port_Dalc port_Walc port_health port_absences port_G1 port_G2 port_G3 avg_studytime
[1,]             1          1         1         1           1             0       0       5       0             1
[2,]             5          5         5         5           5            32      19      19      19             4
     avg_Dalc avg_Walc avg_alcohol avg_famrel
[1,]        1        1           1          1
[2,]        5        5           5          7
```

Note. Figure 28 shows all the range values from the numerical data in the dataset. The feature and target variables are highlighted using the red box.

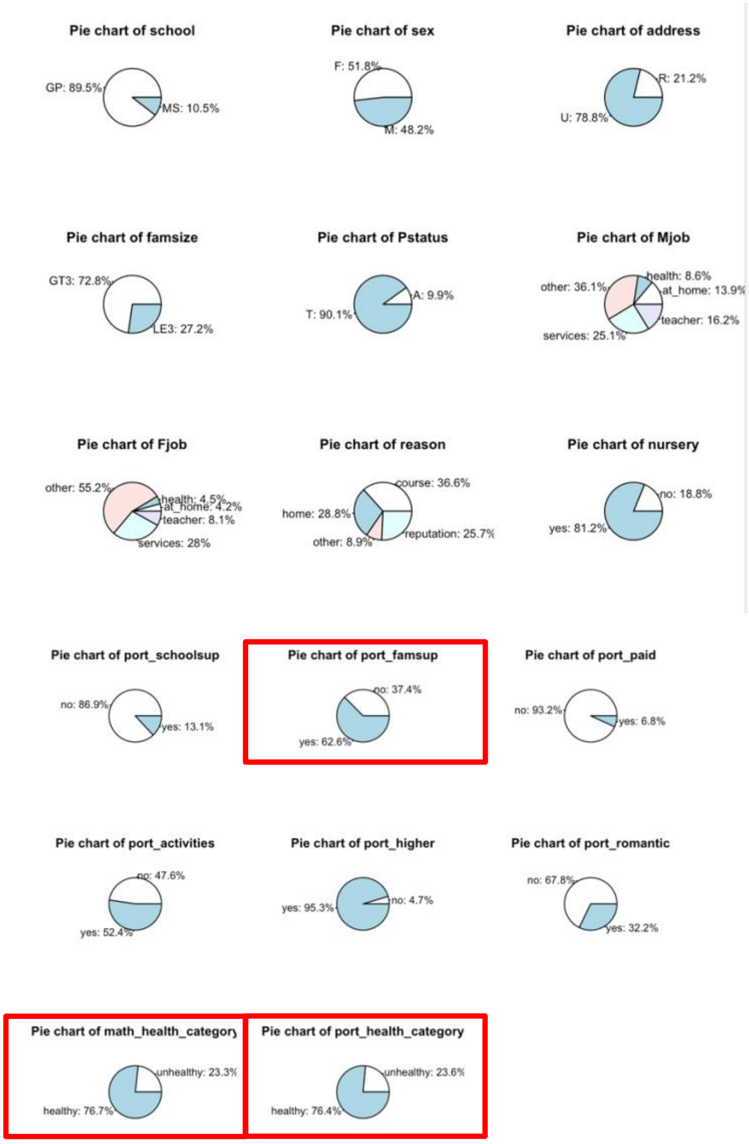
The range of “avg_studytime” is seen to be 3, followed by “avg_alcohol” and “avg_famrelwhich are both 4. This indicates that the range for these variables are not that high. Seeing them as ordinal data, they are not meant to be high in the first place.

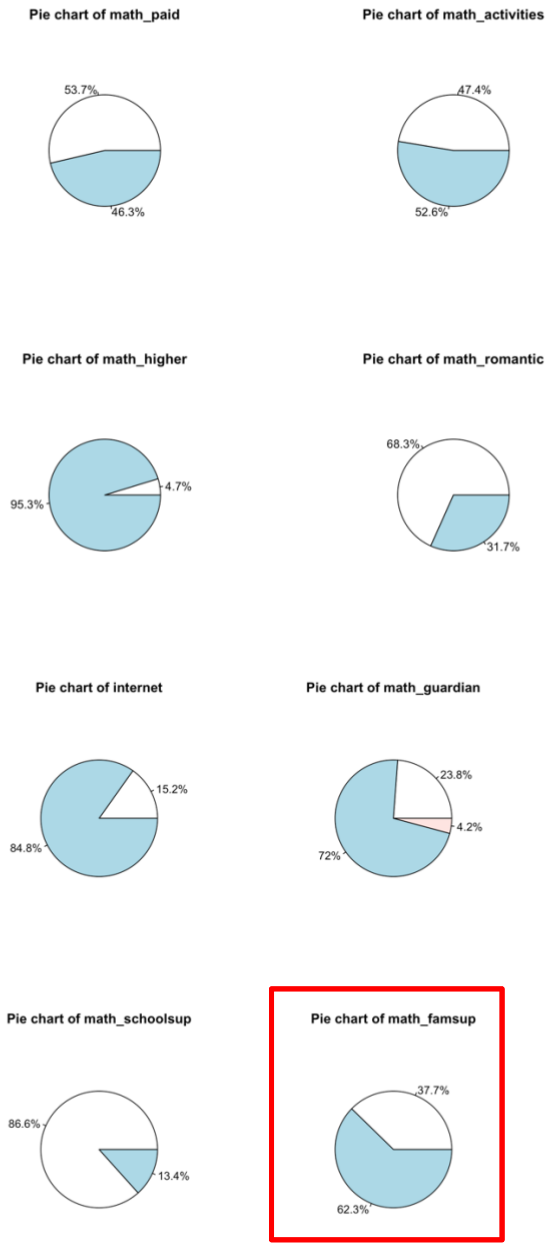
Other than numerical data, categorical data is also visualized using pie charts. The proportions of each pie charts are produced to see the percentage of each level of the variables. To execute this, a data frame called "categorical data" is created to store all the categorical data columns from the dataset. It would be easier to call the data frame as a whole.

After the data frame is created, multiple pie charts can be produced one at a time. The pie chart shows the percentage of each category in the variable clearly. Figure 11 shows the result and visualizations of the proportions of all 26 of the categorical data in the dataset.

Figure 11

Results Showing all Pie Charts Showing the Proportions of Each Categorical Variable in the Dataset





Note. Figure 11 shows all the proportions of the categorical variables in a pie chart visualization. The feature variables that are relevant to the hypothesis formulated are highlighted using a red box.

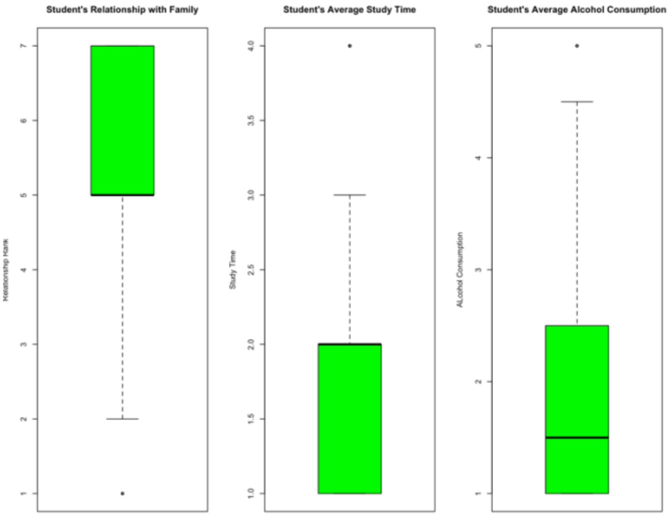
From Figure 11, it shows that there is a slight difference of proportion between "math_famsup" and "port_famsup". On the other hand, the proportion of "math_health_category" and "port_health_category" have a slight difference as well. This might mean the course that the students take does not really affect the presence of family support and also their health.

3.1.2 Outliers and Method to Handle them in the Dataset

There are many methods to identify outliers. However, in this report, box plots will be used. This is because box plots is easy to understand and also very clear in terms of visualization. Therefore, box plots will be used to identify which numerical variable contains outliers. This section will only focus on the numerical feature and target variables that are relevant to the hypothesis. In total, there will be 7 variables to be examined in this subsection. Figure 12 shows the box plots that are produced in R Studio for the relevant numerical variables "avg_studytime", "avg_alcohol" and "avg_famrel".

Figure 12

Results Showing Box Plot for "avg_studytime", "avg_alcohol" and "avg_famrel"



Note. Figure 12 shows the result of the box plot that shows each relevant feature and target variable.

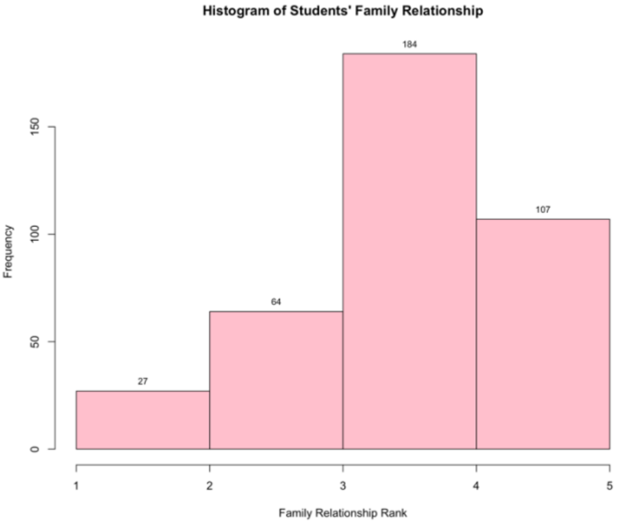
From Figure 12, it seems that there are certain outliers present in some of the relevant variables. However, it is decided that these outliers shall not be removed as they are valid data points. For example, a student with an exceptionally good or exceptionally bad relationship with the family might represent a very high or very low average alcohol consumption. Therefore, these outliers are seemed to be in interest with the hypothesis formulated. The same thing goes for the students' average study time. Thus, it would not be necessary to remove these outlier as they all are in interest to the analysis to be conducted.

3.1.3 Check for Imbalances

From Figure 12, all the variables seem to have potential imbalance in the dataset. Therefore, they will be further examined in this section. Histograms will be created to see the detailed skewness of each variable and then discussed. Figure 13 to Figure 15 shows the histogram created to view the skewness and spread of the variables “avg_famrel”, “avg_studytime” and “avg_alcohol”.

Figure 13

Histogram for “avg_famrel” Variable

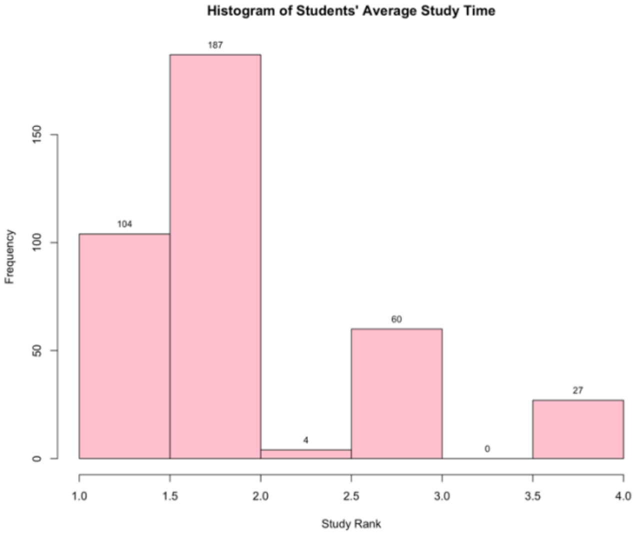


Note. Figure 13 depicts the histogram for the students’ average family relationship by its rank: 1 being very bad while 5 being very good.

From Figure 13, most of the student rate their family relationship higher than average, which is between 3 to 4. This might mean that the students in the dataset mostly have a good relationship with their family. Overall, the histogram is slightly left skewed. However, it is still balanced enough to carry out analysis smoothly.

Figure 14

Histogram for “avg_studytime” Variable



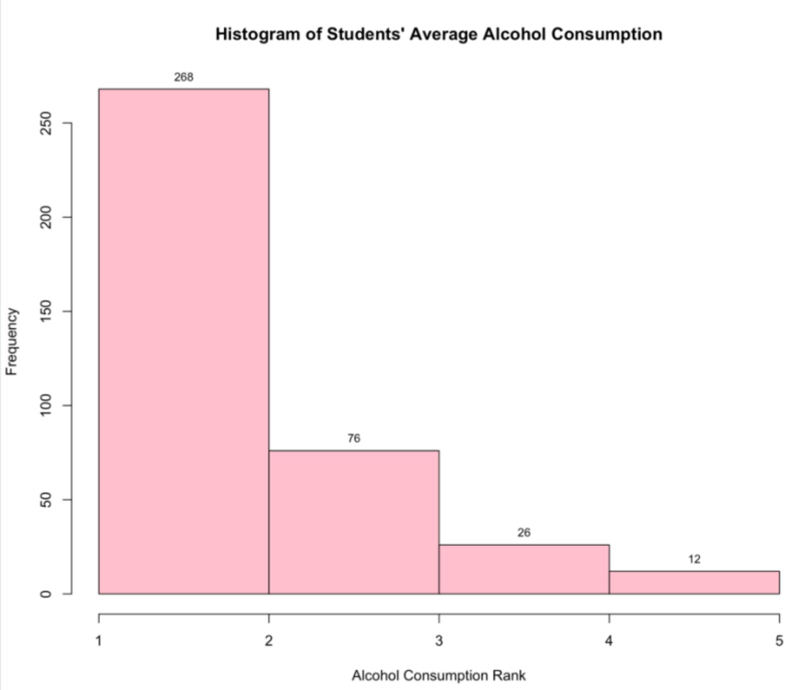
Note. Figure 14 depicts the histogram for the students' average study time for both Math and Portuguese course. 1 would be less than 2 hours, followed 2 being 2 to 5 hours, 3 being 5 to 10 hours and 4 being more than 10 hours.

The histogram in Figure 14 shows a heavily imbalanced graph. The students with a high study time rank 3 and 4 is clearly under represented. This imbalance might end up skewing the analysis towards a lower study time as there a very less data starting after rank 2, especially in classification tasks. Therefore, "avg_studytime" is a variable that is needed to be taken note on for classification or prediction tasks in the future.

Since this report will only be covering EDA tasks, classification and machine learning technique will not be used. Instead, this variable will be used for regression analysis later. Thus, handling this imbalance would not be much necessary as it would not affect the analysis much either. Even if there is an extreme outlier that cause the result to be slightly skewed, it still does not invalidate the result of the regression analysis.

Figure 15

Histogram for "avg_alcohol" Variable



Note. Figure 15 depicts the histogram for the rank of students' average alcohol consumption in a week. 1 would be the lowest rank while 5 will be the highest rank.

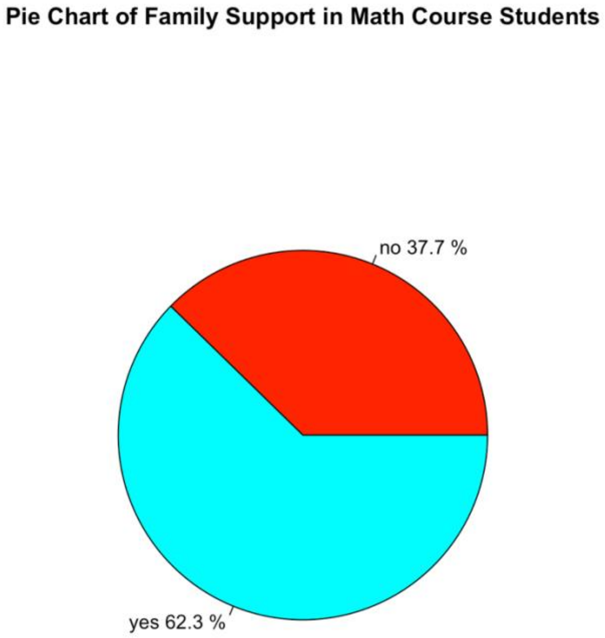
From Figure 15, it shows that the histogram is right skewed. The students with a high (4 to 5 rank) alcohol consumption rank is under represented. Therefore, imbalance does exist in this variable. It is shown that the alcohol consumption with rank 4 and 5 is under represented. With that, it might affect the results of prediction models or other machine learning tools. Therefore, this should be taken note on as well during these tasks later in the predictive analysis process.

In this EDA focused report, this imbalance will be overlooked in the meantime. This is because this variable will not be used for classification or any prediction rules where imbalances are greatly affected. This variable would be used for regression analysis instead. Therefore, the imbalance will not discredit the result. In fact, this histogram is showing that the majority of the students do not consume alcohol at a very high rate, which is also a good and normal sign.

As for categorical data, the relevant categorical data for this analysis would be “math_famsup” and “port_famsup”, “math_health_category” and “port_health_cateogry”. A pie chart is created to show the proportions these variables is examined. Figure 16 and Figure 19 shows the pie chart produced.

Figure 16

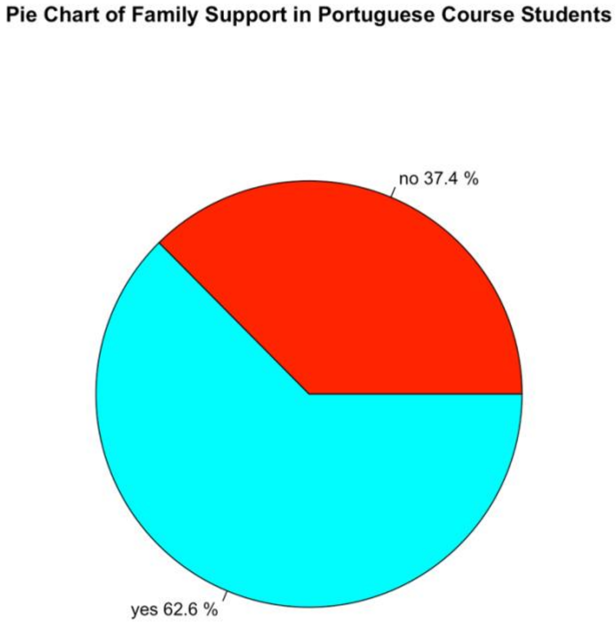
Pie Chart for “math_famsup”



Note. Figure 16 depicts the pie chart showing the proportion of family support among students in Math course. More than half of the students (62.3%) claim that they have received support from their family. On the other hand, there is also a percentage of 37.7% of the student who does not receive any support from their family. It is seen that the data points are relatively balanced and do not need much modification.

Figure 17

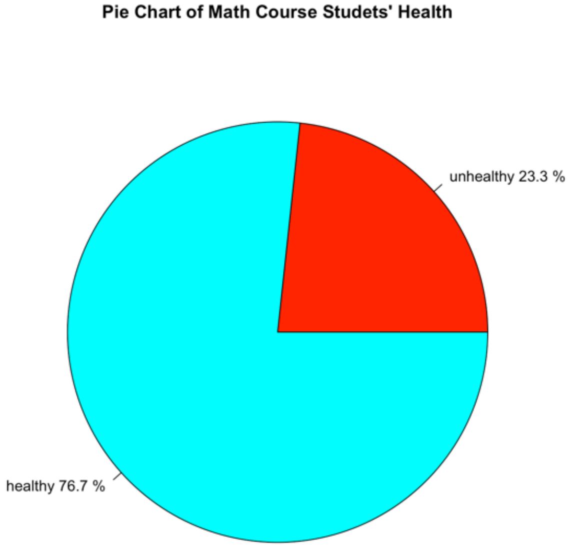
Pie Chart for "port_famsup"



Note. Figure 17 depicts the pie chart showing the proportion of family support among students in Portuguese course. Majority of the students (62.6%) form Portuguese course claim that they have received sufficient amount of family support. On the other hand, 37.4% of the students do not have support from their family. Therefore, it does not show imbalances in the pie chart, indicating that classification process can be proceeded.

Figure 18

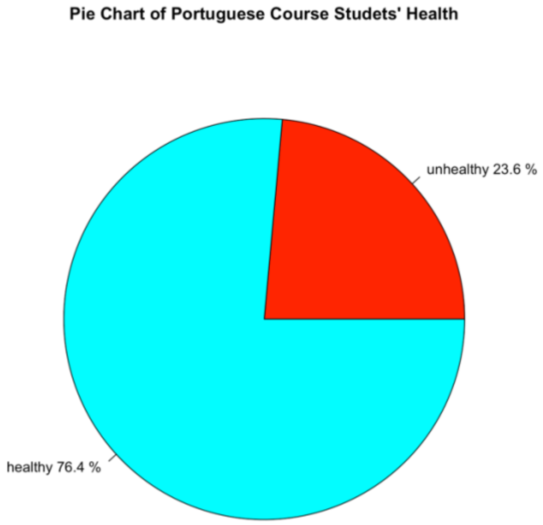
Pie Chart of "math_health_category"



Note. Figure 18 depicts the pie chart showing the proportion of the health of students taking Math course. Most of the students, 76.7% of them, claim that they are healthy where are 23.3% of the students claim that they are unhealthy. It might show imbalance, but each category is not overly representing one another. Therefore, no modifications are needed.

Figure 19

Pie Chart for "port_health_category"



Note. Figure 19 depicts the pie chart showing the proportion of the health of students taking Portuguese course. There 76.4% of students reported themselves as healthy whereas 23.6% of students reported themselves as unhealthy. Though it shows an imbalance, but the data is not overly skewed. Therefore, this imbalance does not need to be handled.

Overall, from Figure 16 to Figure 17 does not have much imbalances in the dataset. In fact, the relevant categorical variables are relatively balanced. On the other hand, from Figure 18 and Figure 19, imbalances do exist. This is something that needs to be taken note on ask well when carrying out prediction analysis such as classification. However, in this case, this imbalance is decided not to be handled as one of the categories is not overly representing the other. On top of that, since EDA is carried out, predictive analysis is not relevant in this report. Nevertheless, this still should be something to be handled before implementing any machine learning algorithms.

3.1.4 Relationship Between Different Variables

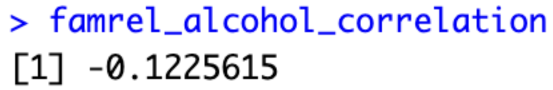
This section is divided into two parts. The first part would be discussing the relationship between variables relevant to the hypothesis. The second part would be further examing the relationships between variables that are not much related with the hypothesis.

3.1.4.1 Relationships Between Variables Relevant to Hypotheses. In this section, the correlation coefficients between relevant numerical variables: “avg_famrel”, “avg_studytime” and “avg_alcohol” will be used to view the relationship between all of them. In the end, the relationship between “avg_studytime” and “avg_alcohol” so as the relationship between “avg_famrel” and “avg_alcohol” will be highly focused on as these two relationships are highly relevant to the hypothesis formulated. Since all the variables used are ordinal data, they would not have a monotonic relationship in nature. Therefore, Spearman method will be the most suitable method to be used when calculating the correlation coefficient of these variables.

For starters, to further test the hypothesis, the correlation coefficient of “avg_famrel” and “avg_alcohol” is calculated. This tells us a more detailed type of relationship that these two variables have with each other. In the end, if the correlation coefficient is close to zero, this means that these two variables does not have any relationship with each other; if the value is close to one, this means that the two variables have a strong positive relationship; if the value is negative one, this would mean that the two variables have a strong negative relationship. Either way, a stronger negative relationship would solidify the hypothesis: the better the family relationship, the lower the alcohol consumption among students. Figure 20 shows the results of the calculated correlation coefficient between these two variables.

Figure 20

Result of Correlation Coefficient of Between “avg_famrel” and “avg_alcohol”



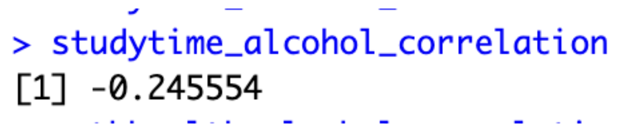
```
> famrel_alcohol_correlation
[1] -0.1225615
```

Note. Figure 20 depicts the correlation coefficient of between “avg_famrel” and “avg_alcohol”. The value is smaller than zero, indicating that it has a negative relationship. However, the value is relatively close to zero, indicating that this is a weak negative relationship. This means that the students’ relationship with their family does not strongly impact their average alcohol consumption.

Secondly, the correlation coefficient between “avg_studytime” and “avg_alcohol” is also calculated. The case would be the same as the correlation coefficient calculated for “avg_famrel” and “avg_alcohol”. It is assumed that the longer the average study time, the lower the alcohol consumption. Therefore, a strong negative relationship would mean that the hypotheses is accepted. Figure 21 shows the result of the calculated correlation coefficient between “avg_studytime” and “avg_alcohol”

Figure 21

Result of Correlation Coefficient of Between "avg_studytime" and "avg_alcohol"



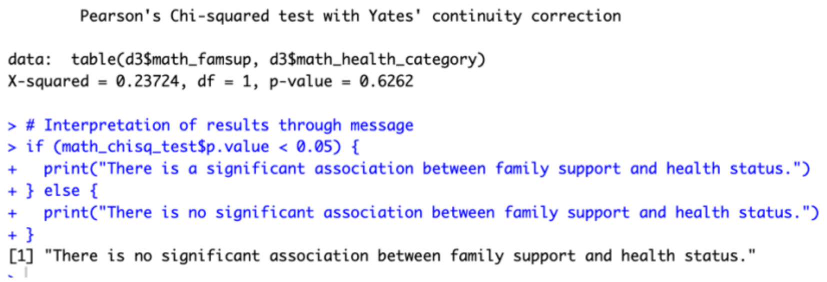
```
> studytime_alcohol_correlation
[1] -0.245554
```

Note. Figure 21 shows the correlation coefficient of Between "avg_studytime" and "avg_alcohol". The value is negative. This means that it certainly has a negative relationship between one another. Since the value is close to -0.25, this indicates that the negative relationship is moderate. The average study time of students might not directly affect their alcohol consumption, but it would be more noticeable. In other words, the study time of students moderately affects their alcohol consumption in a week.

Other than correlation coefficient between numerical data, a chi-squared test is carried out to test the association between the two sets of relevant categorical variables ("math_famsup" with "math_health_category" and "port_famsup" with "port_health_category"). For Math course, "math_famsup" will be the independent variable whereas "math_health" will be the dependant variable. As for Portuguese course, "port_famsup" will be the independent variable while "port_health_category" is the dependant variable. In the end, if the calculated chi-squared statistics is lower than the critical point (0.05), the hypothesis formulated will not be accepted. On the contrary, if the chi-squared statistics is higher than the critical point, the hypothesis formulated will be accepted. Figure 22 and Figure 23 shows the results of the chi-squared test carried out for "math_famsup" with "math_health_category" and "port_famsup" with "port_health_category" respectively.

Figure 22

Results of Chi-Squared Test on "math_famsup" and "math_health_category"

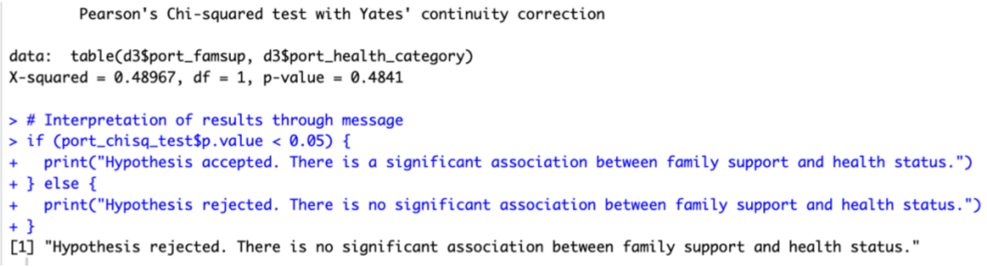


```
	Pearson's Chi-squared test with Yates' continuity correction

data:  table(d3$math_famsup, d3$math_health_category)
X-squared = 0.23724, df = 1, p-value = 0.6262

> # Interpretation of results through message
> if (math_chisq_test$p.value < 0.05) {
+   print("There is a significant association between family support and health status.")
+ } else {
+   print("There is no significant association between family support and health status.")
+ }
[1] "There is no significant association between family support and health status."
```

Note. Figure 22 depicts the results of chi-squared test on "math_famsup" and "math_health_category". It is shown that the p value is 0.6262, which larger than the significance level (0.05). Therefore, it is calculated that there is no significant association between family support and health status for students taking Math course. Therefore, the hypothesis is rejected.

Figure 23

Results of Chi-Squared Test on "port_famsup" and "port_health_category"



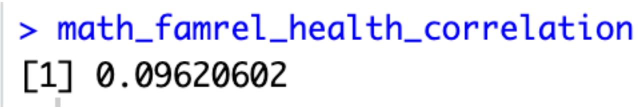
```
	Pearson's Chi-squared test with Yates' continuity correction

data:  table(d3$port_famsup, d3$port_health_category)
X-squared = 0.48967, df = 1, p-value = 0.4841

> # Interpretation of results through message
> if (port_chisq_test$p.value < 0.05) {
+   print("Hypothesis accepted. There is a significant association between family support and health status.")
+ } else {
+   print("Hypothesis rejected. There is no significant association between family support and health status.")
+ }
[1] "Hypothesis rejected. There is no significant association between family support and health status."
```

Note. Figure 23 depicts the results of chi-squared test on "port_famsup" and "port_health_category". After calculation, the p is 0.4841, which is also higher than the significance level (0.05). Therefore, it is safe to say the there is no significant association between family support and health status for students taking Portuguese course. In the end, the hypothesis will be rejected as well.

3.1.4.2 Relationships Between Variables Not Relevant to the Hypotheses. To test the relationships of these variables further, the correlation coefficient of the variables "math_famrel" with the variables "math_health". This to see the type and strength of relationships between these the family relationship and the health of students in Math course. Figure 24 shows the correlation coefficient value calculated in R Studio.

Figure 24

Correlation Coefficient of "math_famrel" and "math_health"



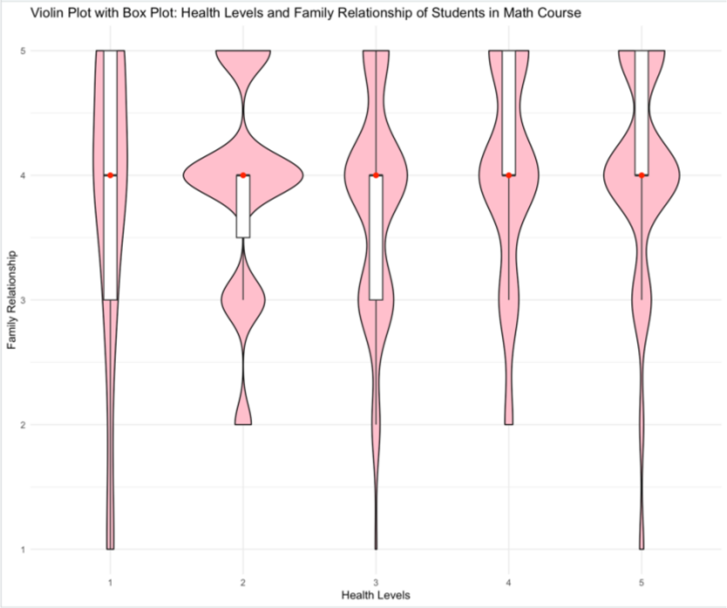
```
> math_famrel_health_correlation
[1] 0.09620602
```

Note. Figure 24 shows the correlation coefficient of the variables "math_famrel" and "math_health". The value is a positive value, indicating that there is a direct relationship between these two variables. In other words, the better the family relationship, the better the students' health. However, the value is also very close to zero. This would mean that there is almost no relationship between the variables.

This relationship is further examined using a violin plot in Figure 25 below. A box plot is also added on top of the violin plot to view the quartiles of the data.

Figure 25

Violin Plot of "math_famrel" Against "math_health"



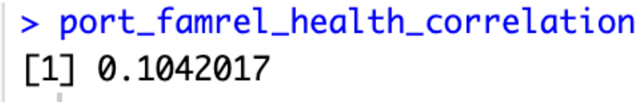
Note. Figure 25 depicts the violin plot together with a box plot that is created to see the relationship between family relationship and the health of students in Math course.

For the areas around health level 1 (bad health), it is shown that the distribution is skewed downwards. Majority of the students who rank their health as 1 have claimed that their family relationship is good (rank 4 to 5). The median (red dot) is also placed on the higher end, causing it to have a long "tail". For students who rank their health as 2, the violin plot is shorter, which means it has a smaller range. The concentration is the highest on the family relationship with rank 4, indicating that most of the students with health rank 2 rank their family relationship as 4. For the violin plots for health ranks 3, 4 and 5, their shape is somewhat similar. They show a higher concentration on family relationship of rank 4. This would indicate that students with health rank of 3 to 5 usually would rank their family relationship on a higher end (rank 4 and 5). Overall, it is found that the concentration on the family relationship with rank 4 is the highest among all health ranks. This means that even if the students do not have a good health, most of them still have a good family relationship.

Other than that, the relationship between "port_famrel" and "port_health" is also examined to see the relationship between family relationship and health of students in Portuguese course. Firstly, the correlation coefficient between these variables is calculated to view their exact relationship. Figure 26 shows the result of correlation coefficient between "port_famrel" and "port_health".

Figure 26

Correlation Coefficient between "port_famrel" and "port_health"



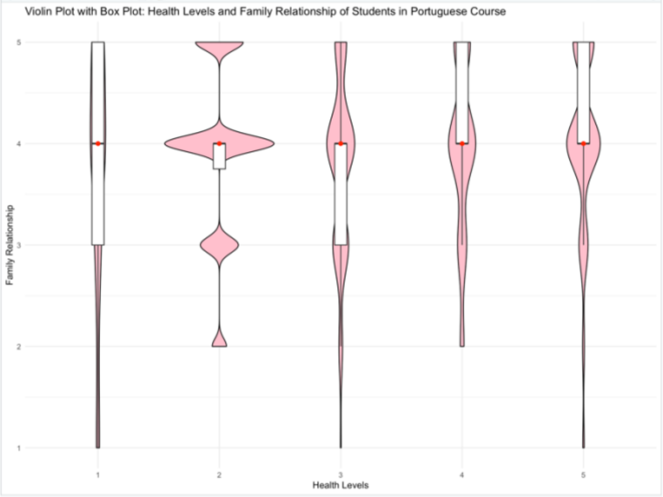
```
> port_famrel_health_correlation
[1] 0.1042017
```

Note. Figure 26 depicts the correlation coefficient calculated between the variables "port_famrel" and "port_health". The value calculated is a positive, which means that they have a direct relationship with each other. The better the family relationship, the better the health of students in Portuguese course. The value 0.1042017 calculated is also very close to 0. This would indicate that the relationship is weak. In other words, having a better relationship would not directly mean that the students would have a good health. It would be affected in some ways, but not much.

Other than correlation coefficient, a violin plot is also plotted to see the area where it has the highest rank of health. The violin plot produced is shown in Figure 27 below.

Figure 27

Violin Plot of "port_famrel" against "port_health"



Note. Figure 27 shows the violin plot produced to view the relationship between family relationship and health of students taking Portuguese course.

From Figure 27, it is seen that for health rank 1, most of the data points are located at the place with family relationship of rank 4 and 5. This would mean that even with a bad health, most students would still have a good relationship with their family. As for the health with rank 2, there is a high concentration on family with rank 4 and 3 as well. The distribution is more focused on the family relationship with rank 3 and 4. This means that most students with a health rank of 2 also claim that they have a relatively good relationship with their family. Other than that, the violin plots for health ranks 3, 4 and 5 are very similar. The shape of the violin plot is also narrow which would indicate that the distribution is quite even among the data. Nevertheless, most of the data concentrates on family relationship with rank of 4 and 5. This would state that a student who has a better health mostly will have a better relationship with his or her family. Overall, nevertheless the rank of the students' health, most of the student still claim that they have a high health rank. This would mean that their health do not significantly associate with their family relationship.

3.2 Visualization

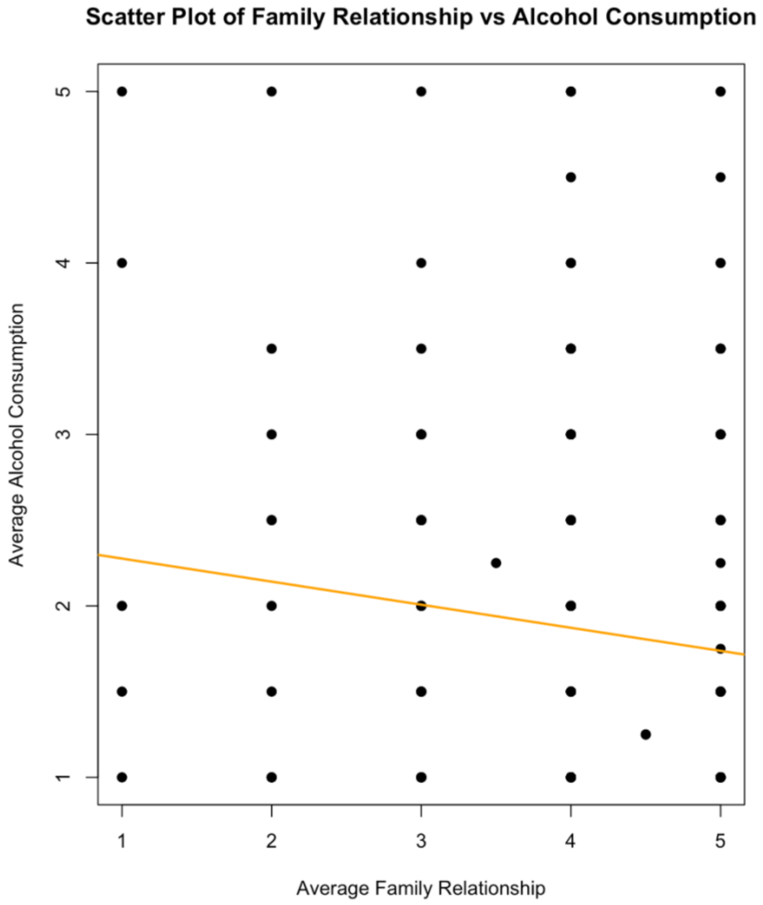
In this section, all the most relevant visualization from the analysis will be provided. This includes scatter plots that are highly related to the hypotheses and also a mosaic plot mainly for the target categorical data. This section will be divided into three parts where each part will be examining and giving a conclusion for each hypothesis. The relevant codes in R Studio to produce the visualizations in this section will be included in appendix E.

3.2.1 Students With a Better Family Relationship Would be Less Likely to Engage in Alcohol Consumption

To test this hypothesis, the relationship between two variables are examined. The variables are "avg_famrel" and "avg_alcohol". "Avg_famel" would be the independent variable whereas "avg_alcohol" would be the dependent variable. Thus, a scatter plot would be the most suited to see the relationship between these two variables. Figure 24 below will show the result of the scatter plot graph and a regression line on it, showing the type and strength of relationship between "avg_famrel" and "avg_alcohol".

Figure 24

Scatter Plot Between "avg_famrel" and "avg_alcohol"



Note. Figure 24 depicts the scatter plot between “avg_famrel” and “avg_alcohol”.

From Figure 24, the regression line formed in the scatter plot graph has a negative gradient. This means that these two variables have an inverse relationship. In other words, it can be said that the better the average family relationship, the lower the students’ alcohol consumption. However, the regression line in the figure is also relatively flat, which means the gradient value is not high. Therefore, the relationship between these two variables are not strong. This means that it is possible for one’s family relationship to affect a child’s alcohol consumption but it will not severely affect it to happen.

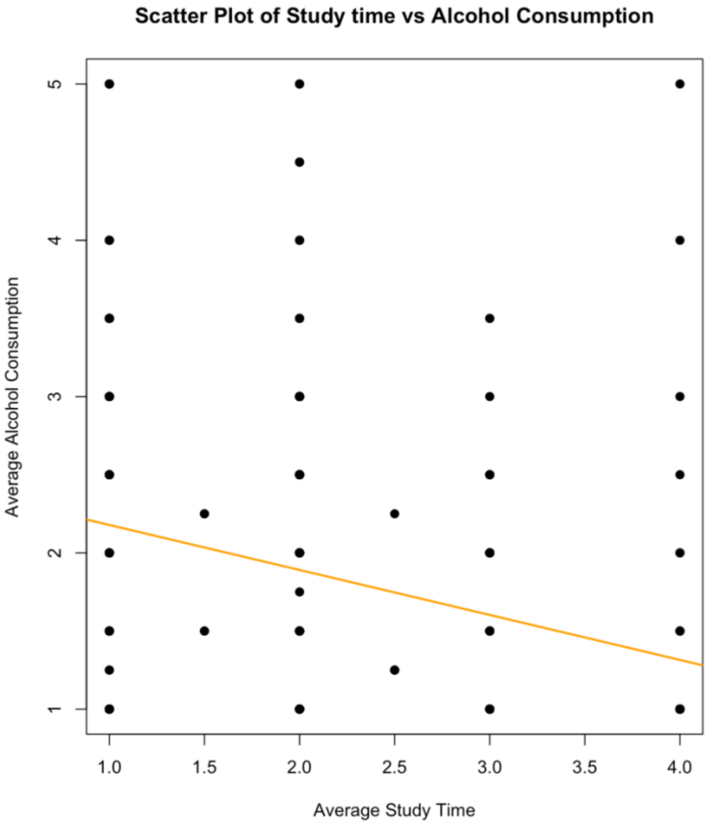
This could also be seen back in the previous section where the correlation coefficient value calculated is relatively low with the value of -0.1225615. The negative value indicates that there is definitely an inverse relationship between the two variables. However, the value is relatively far from the value -1. The further the correlation coefficient value is from -1, the weaker the relationship. Therefore, we can conclude that there is a weak negative relationship between students’ family relationship and their average alcohol consumption. Nevertheless, the hypothesis will be accepted.

3.2.2 Students Who Spend More Time in Studying Will be Less Likely to Engage in Alcohol

Firstly, the relevant variables that are used to test this hypothesis are “avg_studytime” and “avg_alcohol. In this case, the “avg_alcohol” would be depending on the “avg_studytime”. To test this hypothesis, the relationship between these two variables will be calculated and viewed in a scatter plot graph. At the same time, a regression line is created on the scatter plot to directly show the relationship and its strength. Figure 25 result of the scatter plot and regression line for the relationship between “avg_studytime” and “avg_alcohol” .

Figure 25

Scatter Plot Between “avg_studytime” and “avg_alcohol”



Note. Figure 25 depicts the scatter plot and regression line of the “avg_alcohol” against “avg_studytime” graph.

From Figure 25, the regression line has a negative gradient, indicating that these two variables have an inverse relationship. This would mean that the longer the study time, the lower the alcohol consumption. A for the strength of the relationship, the line is slightly slanted but not to a very high extent. Thus, it is safe to say that the relationship is moderate. In other words, the average time taken for students to study for school will be moderately affecting their average alcohol consumption. The possibility of this happening is slightly strong but it still would not directly affect the depending variable.

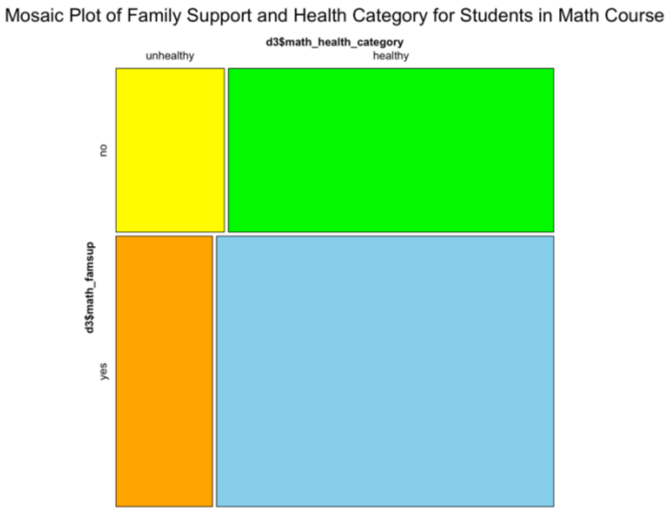
Moreover, from the correlation coefficient value (-0.245554) calculated earlier in the section, we can clearly see that the negative value validates the inverse relationship. As we all know, the closer the correlation coefficient value is to -1, the stronger the relationship. The value of approximately -0.25 is somewhat close to -1, letting the variables to have a moderate relationship with one another. Therefore, from the results given, the longer time taken for the students to study, the more unlikely they will engage in alcohol. In the end, the hypothesis is accepted.

3.2.2 Students' health is greatly affected by the presence of family support

For this hypothesis, the relationship between two sets of categorical variables will be examined. The first set of categorical variables that is to be examined is "math_famsup" and "math_health_category". These variables will target on the students taking Math course. In the end, we would get to see how family support affect a students' health who are taking Math course. To analyze the relationship between the both of them, a mosaic plot is created to visualize the relationship clearly. Since they are both categorical data, a regression analysis will not be suitable in this case. Figure 26 shows the mosaic plot created in R studio to view the relationship between "math_famsup" and "math_health_category".

Figure 26

Mosaic Plot to View Relationship Between "math_famsup" and "math_health_category"



Note. Figure 26 depicts the mosaic plot to view the relationship between "math_famsup" and "math_health_category"

From Figure 26, each rectangle of different color represents a particular combination of “math_famsup” and “math_health_category”. The yellow rectangle represent the students with no family support and is reported to be unhealthy; the green rectangle would be the students with no family support but is reported to be healthy; the orange rectangle represents the students with family support but is reported to be unhealthy; the blue rectangle would be the students with family support and is reported to be healthy.

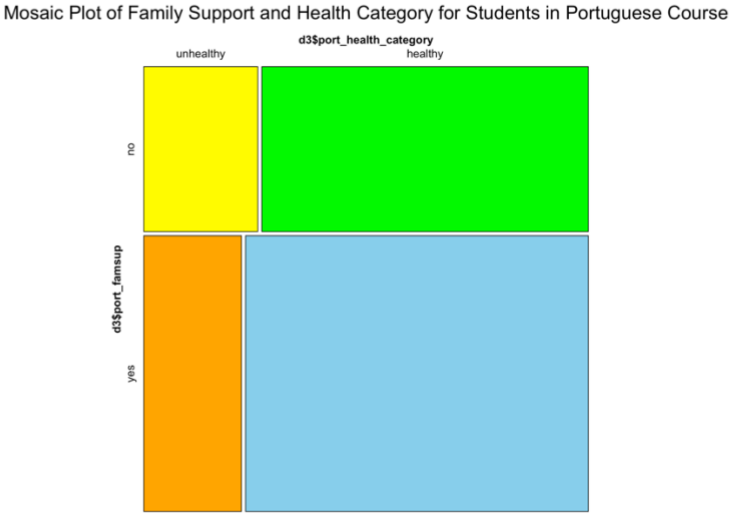
Based on the area of rectangles, the blue rectangle covers the largest area which indicates that majority of students who are healthy do receive support from their family. However, the green rectangle has the second largest area, indicating that there is also a large proportion of students who claims to be healthy even though they do not receive support from their family. On the other hand, the difference in number of unhealthy students who receive family support and the number of unhealthy students who do not receive family support is very little. Overall, it is clear that there is not much significance between the presence of family support and the health of students taking Math course.

On top of that, based on the chi-squared test, the calculated p-value between “math_famsup” and “math_health_category” is lower than the significance level (0.05). This would also mean that these two variables indeed do not have a significant association. Therefore, with the mosaic plot and chi-squared test conclusion provided, the hypothesis will be rejected.

As for the students who are taking Portuguese course, the relationship between “port_famsup” and “port_health_category” is also examined to see relationship between the two categorical variables for students in Portuguese course. A mosaic plot is also created test the association between two variables as regression analysis would not be suitable for categorical data. Figure 27 shows the mosaic plot of “port_famsup” and “port_health_category”.

Figure 27

Mosaic Plot of “port_famsup” and “port_health_category”



Note. Figure 27 depicts the mosaic plot created to examine the relationship between the presence of family support and the students' health category in the Portuguese course.

From Figure 27, each rectangle of different color represents a particular combination of "port_famsup" and "port_health_category". The yellow rectangle represent the students with no family support and is reported to be unhealthy; the green rectangle would be the students with no family support but is reported to be healthy; the orange rectangle represents the students with family support but is reported to be unhealthy; the blue rectangle would be the students with family support and is reported to be healthy.

Based on the area of rectangles, the majority of the students have received family support and is healthy (blue rectangle). However, there is also more than half the students with no family support but still reported to be healthy (green rectangle). In that case, the presence family support towards a student might not be a significant cause of his or her health. Even if the student do not receive family support, their health does not seem to be affected that much.

On top of that, the chi-squared test conducted before shows that the p-value is smaller than the significance level (0.05). Thus, it is concluded that there is no significance association between the presence of family support and the health of students who are taking Portuguese course. From both the conclusion made through chi-squared test and the mosaic plot created, the hypothesis will be rejected.

Overall, this hypothesis is rejected as both set of categories: Math course and Portuguese course have shown that the presence of family support will not greatly affect the students' health.

3.3 Conclusion and Recommendation

This section mainly aims to wrap up the entire report by summarizing the overall finding and their implications. Recommendations will also be provided in order to improve EDA processes. Before the report ends, challenges encountered in this analysis will be mentioned, together with they ways to overcome them. In a nutshell, the Student Performance Data is a dataset that lets us to analyze how different aspects of the students' life can affect their academic results and also alcohol consumption.

3.3.1 Summary of Overall Findings and Implications Related to Hypothesis

There are three findings throughout this EDA process. The first finding would be that the family relationship of a student has an inverse relationship with the alcohol consumption of the students themselves. From the regression analysis, the regression line produced is slightly slanted and has a negative gradient. Therefore, there is a weak negative relationship between the family relationship and alcohol consumption. In addition, the correlation coefficient calculated has -0.1225615: a negative number that is close to zero, indicating that the type of relationship is negative and the strength of it is weak as well. In that case, the result would show that though the quality of family relationship would not significantly affect the alcohol consumption of the student, they still have an impact towards it nevertheless. In the end, the initial hypothesis: students with a better family relationship would be less likely to engage in alcohol consumption, would be accepted.

The second finding would be that the average study time of a student also has an inverse relationship with their average alcohol consumption in a week. This finding is found be two ways: regression analysis and calculation of correlation coefficient. The regression analysis conducted shows that the regression line produced is slanted moderately and has a negative gradient. The moderate steepness represent the moderate strength of relationship between a students' average study time and their average alcohol consumption in a week. On the other hand, the negative gradient represent the negative relationship that these two variables have with each other. As for the calculation of correlation coefficient, the final value produced is -0.245554: a negative value that is relatively closer to zero. Though the value may be small, but it still indicates a moderate strength of relationship. In short, it is safe to say that the longer the study time, the more unlikely a students will engage in alcohol consumption. Thus, the initial hypothesis is accepted.

The last finding would be that there is no significant relationship between the presence of family support and the health of students who is studying both Math and Portuguese course. Upon conducting a chi-squared test towards the variables specifically from Math and Portuguese course, this finding is solidified. It is found that the p-value calculated through the test has exceeded 0.05 which is the significant level. This indicates that the null hypothesis (there is no significant association between family support and students' health) cannot be rejected. On top of that, mosaic plots are created for students specifically in Math course and Portuguese course as well. The plots produced showed that the number of healthy students who received family and the number of unhealthy students without family support is almost the same. Thus, there is indeed no significant association between the presence of family support and students' health. With that, the hypothesis is rejected.

3.3.2 Recommendation

The first recommendation would be to expand the scope of variables in the dataset. This dataset focuses more on the students' lifestyle, family status and also their demographics. Other external variables such as mental health, diet and socioeconomic status can also be used to test their relationship the target variables: average alcohol consumption and health status. By expanding the scope, more variables with a potentially high significance association can be used as the main predictor of the average alcohol consumption and health status. Other than that, it also gives a more comprehensive understanding towards a student's well-being.

The second recommendation would be to increase the use of visualizations during analysis. Other than box plots, histograms and scatter plots, there are many more types of visualization that can be used to view the relationship between the variables. In the future, more advanced visualizations can be considered to uncover more details of the datasets. For example, violin plots can also be used to visualize ordinal data across various categories. Moreover, clustering can also be carried out to see how family support, family relationship and average study time can affect each particular clusters. In other words, it is recommended to explore more on visualizations while maintaining the best practices of creating visualizations at the same time.

Lastly, it is recommended to utilize more on feature engineering. Having many variables in the dataset, more additional features variables can be derived using the existing ones. For example, a new variable called "work_life_balance" can be created by combining variables such as students' study time, free time and how often do they go out. These new variables can used to test the relationship between the target variables. There is a possibility where a stronger predictor for average alcohol consumption and health status can be found by carrying out feature engineering. Overall, this step is highly recommended as it dives in deeper into each feature variable to create a new variable that can be used to find calculate significance of the target variables.

3.3.3 Challenges Encountered and How They Were Overcome

One of the challenges encountered would be handling data type mismatch. The initial dataset had quite a number of mismatched data types. For example, the categorical data would be treated as strings whereas as the ordinal data would be treated as continuous data. Therefore, to handle this, the variables in the dataset have to be understood thoroughly. After that, their data types should be determined. Therefore, we would need to be very clear on what does each data mean and how it is interpreted. Then, basic codes in R Studio can be used to modify the data type. The syntax `as.factor()` and `as.numeric()` can be used to change the data types into categorical data and numerical data respectively.

Besides, another challenge would be understanding ordinal data. Ordinal data are data that has ranking and order. They are usually difficult to be used during analysis with categorical variables. For example, if we want to see the significance of the students' health (ordinal data) and the presence of family support (categorical data). This would be challenging as ordinal data does not have equal intervals like categorical data. Therefore, this could be handled by categorization. The ordinal data can be changed to broader categories so that it would be easier to run tests to see their relationship with other categorical variables. In this case, students' health can be categorized as "healthy" and "unhealthy" to carry out chi-squared test with family support.

Lastly, the last challenge would be finding the right visualization. The right visualization is very important during analysis as it aims to give the readers a clear picture of the data. If the visualization used is not suitable, then it might cause confusion or misinterpretation. For example, scatter plots would be suitable for test relationships between two numerical data while mosaic plots are more suitable to test relationships between to categorical data. If these visualizations are mixed up, the graph produced would be confusing and hard to understand, making it a bad analysis project. To handle this, we must first understand the structural feature of the dataset. Then, the use and properties of each visualization have to be understood too. With that, it would be easy to know which visualization highlights the key takeaway of the analysis.

References

Cargiulo, T (2007). Understanding the health impact of alcohol dependance. *American Journal of Health-System Pharmacy, 64*(5). https://academic.oup.com/ajhp/article-abstract/64/5_Supplement_3/S5/5135268

Desforges, C. & Abouchaar, A. (2003). *The impact of parental involvement, parental support and family education on pupil achievement and adjustment: A literature review* (Vol. 433). London: DfES. https://library.bsl.org.au/jspui/bitstream/1/3644/1/Impact%20of%20Parental%20Involvement_Desforges.pdf

Holtes, M., Bannink, R., Zwanenburg, E. J., As, E. V., Raat, H. & Broeren, S. (2015). Association of truancy, perceived school performance, and mental health with alcohol consumption among adolescents. *Journal of School Health, 85*(12), 852-860. https://onlinelibrary.wiley.com/doi/abs/10.1111/josh.12341

Kapur, R. (2023). *Lack of parental support: Unfavorable in Leading to Progression.* Research Gate. https://www.researchgate.net/publication/369479747_Lack_of_Parental_Support_Unfavourable_in_Leading_to_ProgressionThomas, P. A., Liu, H. & Umberson, D. (2017). Family relationship and well-being. *Innovation in Aging, 1*(3). https://academic.oup.com/innovateage/article/1/3/igx025/4617833

Ong, H. S., Fernandez, P. A. & Lim, H. K. (2021). Family engagement as part of managing patients with mental illness in primary care. *Singapore Medical Journal, 62*(5), 213-219. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8801858/

Pardee, L. (2024, February 22). *What is your parenting style, and why does it matter?* Parents. https://www.parents.com/parenting/better-parenting/style/parenting-styles-explained/#:~:text=Your%20parenting%20style%20can%20affect,the%20rest%20of%20their%20life.

State, L. (2022, August 12). *10 characteristics of a good student.* Linda State Education. https://lindastade.com/10-characteristics-of-a-good-student/

Appendix A. Data Import Codes

Code to Import Data and Merge Two Datasets Together

```
# import data
# add separate dataset ad d1 and d2
d1=read.table("student-mat.csv",sep=",",header=TRUE)
d2=read.table("student-por.csv",sep=",",header=TRUE)

# merge data that are available in both d1 and d2 as d3
d3=merge(d1,d2,by=c("school","sex","age","address","famsize","Pstatus","
Medu","Fedu","Mjob","Fjob","reason","nursery","internet"), all=FALSE)
```

Code Changing Repeated Column Names from Math Course

```
# rename same attributes from math course
colnames(d3)[colnames(d3) == "guardian.x"] <- "math_guardian"
colnames(d3)[colnames(d3) == "traveltime.x"] <- "math_traveltime"
colnames(d3)[colnames(d3) == "studytime.x"] <- "math_studytime"
colnames(d3)[colnames(d3) == "failures.x"] <- "math_failures"
colnames(d3)[colnames(d3) == "schoolsup.x"] <- "math_schoolsup"
colnames(d3)[colnames(d3) == "famsup.x"] <- "math_famsup"
colnames(d3)[colnames(d3) == "schoolsup.x"] <- "math_schoolsup"
colnames(d3)[colnames(d3) == "famsup.x"] <- "math_famsup"
colnames(d3)[colnames(d3) == "paid.x"] <- "math_paid"
colnames(d3)[colnames(d3) == "freetime.x"] <- "math_freetime"
colnames(d3)[colnames(d3) == "goout.x"] <- "math_goout"
colnames(d3)[colnames(d3) == "Dalc.x"] <- "math_Dalc"
colnames(d3)[colnames(d3) == "Walc.x"] <- "math_Walc"
colnames(d3)[colnames(d3) == "health.x"] <- "math_health"
colnames(d3)[colnames(d3) == "absences.x"] <- "math_absences"
colnames(d3)[colnames(d3) == "G1.x"] <- "math_G1"
colnames(d3)[colnames(d3) == "G2.x"] <- "math_G2"
colnames(d3)[colnames(d3) == "G3.x"] <- "math_G3"
colnames(d3)[colnames(d3) == "activities.x"] <- "math_activities"
colnames(d3)[colnames(d3) == "higher.x"] <- "math_higher"
colnames(d3)[colnames(d3) == "romantic.x"] <- "math_romantic"
colnames(d3)[colnames(d3) == "famrel.x"] <- "math_famrel"
```

Code Changing Repeated Column Names from Portuguese Course

```
# rename same attributes from Portuguese course
colnames(d3)[colnames(d3) == "guardian.y"] <- "port_guardian"
colnames(d3)[colnames(d3) == "traveltime.y"] <- "port_traveltime"
colnames(d3)[colnames(d3) == "studytime.y"] <- "port_studytime"
colnames(d3)[colnames(d3) == "failures.y"] <- "port_failures"
colnames(d3)[colnames(d3) == "schoolsup.y"] <- "port_schoolsup"
colnames(d3)[colnames(d3) == "famsup.y"] <- "port_famsup"
colnames(d3)[colnames(d3) == "schoolsup.y"] <- "port_schoolsup"
colnames(d3)[colnames(d3) == "famsup.y"] <- "port_famsup"
colnames(d3)[colnames(d3) == "paid.y"] <- "port_paid"
colnames(d3)[colnames(d3) == "freetime.y"] <- "port_freetime"
colnames(d3)[colnames(d3) == "goout.y"] <- "port_goout"
colnames(d3)[colnames(d3) == "Dalc.y"] <- "port_Dalc"
colnames(d3)[colnames(d3) == "Walc.y"] <- "port_Walc"
colnames(d3)[colnames(d3) == "health.y"] <- "port_health"
colnames(d3)[colnames(d3) == "absences.y"] <- "port_absences"
colnames(d3)[colnames(d3) == "G1.y"] <- "port_G1"
colnames(d3)[colnames(d3) == "G2.y"] <- "port_G2"
colnames(d3)[colnames(d3) == "G3.y"] <- "port_G3"
colnames(d3)[colnames(d3) == "activities.y"] <- "port_activities"
colnames(d3)[colnames(d3) == "higher.y"] <- "port_higher"
colnames(d3)[colnames(d3) == "romantic.y"] <- "port_romantic"
colnames(d3)[colnames(d3) == "famrel.y"] <- "port_famrel"
```

Code to Modify Nominal Data Types to Categorical Data Types in Dataset

```
# change nominal data to categorical data
d3$school <- as.factor(d3$school)
d3$sex <- as.factor(d3$sex)
d3$address <- as.factor(d3$address)
d3$famsize <- as.factor(d3$famsize)
d3$Pstatus <- as.factor(d3$Pstatus)
d3$Mjob <- as.factor(d3$Mjob)
d3$Fjob <- as.factor(d3$Fjob)
d3$reason <- as.factor(d3$reason)
d3$math_guardian <- as.factor(d3$math_guardian)
d3$port_guardian <- as.factor(d3$port_guardian)
d3$math_schoolsup <- as.factor(d3$math_schoolsup)
d3$port_schoolsup <- as.factor(d3$port_schoolsup)
d3$math_famsup <- as.factor(d3$math_famsup)
d3$port_famsup <- as.factor(d3$port_famsup)
d3$math_paid <- as.factor(d3$math_paid)
d3$port_paid <- as.factor(d3$port_paid)
d3$math_activities <- as.factor(d3$math_activities)
d3$port_activities <- as.factor(d3$port_activities)
d3$nursery <- as.factor(d3$nursery)
d3$math_higher <- as.factor(d3$math_higher)
d3$port_higher <- as.factor(d3$port_higher)
d3$internet <- as.factor(d3$internet)
d3$math_romantic <- as.factor(d3$math_romantic)
d3$port_romantic <- as.factor(d3$port_romantic)
d3$avg_famrel = as.numeric(d3$avg_famrel)
```

Code to View All Data Types in the Dataset in R Studio

```
# check for data types
str(d3)
```

Code to Check Duplicated Data and View Total Number of Duplicated Data in the Dataset

```
# check for duplicates
duplicated(d3)
# check for sum of duplicated data
sum(duplicated(d3))
```

Code to Check Missing Values and View the Total Number of Missing Values in the Dataset

```
# check for missing values
is.na(d3)
# check for sum of missing values
sum(is.na(d3))
```

Appendix B. Structural Investigation Codes

Code to Create New Variable “avg_alcohol”

```
# calculate average daily alcohol consumption and insert as new column
in d3
d3$avg_Dalc = rowMeans(d3[, c("math_Dalc", "port_Dalc")], na.rm = TRUE)
# calculate average weekend alcohol consumption and insert as new column
in d3
d3$avg_Walc = rowMeans(d3[, c("math_Walc", "port_Walc")], na.rm = TRUE)
# calculate average alcohol consumption in a week and insert as new
column in d3
d3$avg_alcohol = rowMeans(d3[, c("avg_Dalc", "avg_Walc")], na.rm=TRUE)
```

Code to Create New Variable “avg_famrel”

```
# calculate the average family relationship for a student and insert as
new column in d3
d3$avg_famrel = rowMeans(d3[, c("math_famrel", "port_famrel")], na.rm =
TRUE)
```

Code to Create New Variable “avg_studytime”

```
# calculate average study time and insert as new column in d3
d3$avg_studytime = rowMeans(d3[, c("math_studytime", "port_studytime")],
na.rm = TRUE)
```

Code to Create New Variable “math_health_cateogry”

```
# math_health
d3$math_health_category = cut(d3$math_health,
                              breaks = c(0, 2, 5),  # Define breaks for
                              the categories
                              labels = c("unhealthy", "healthy"),  #
                              Define labels
                              right = TRUE)  # Include the rightmost
                              interval
```

Appendix C. Quality Investigation Codes

Code to View Missing Data in the Dataset and Produce Visualization

```
# Install and load the package "naniar" to view missing data
install.packages("naniar")
library("naniar")
# visualize missing data
gg_miss_var(d3)
```

Code to Produce Heatmap for Missing Data

```
# produce heat map for missing values
vis_miss(d3)
```

Code to Created “duplicated_values” Data Frame with Only the Duplicated Rows

```
# Identify duplicated rows by forming a data frame
duplicated_values <- d3[duplicated(d3) | duplicated(d3, fromLast = TRUE),
]
# View the duplicated rows
print(duplicated_values)
```

Code to Show all Levels of Categorical Data in the Dataset to Check on Unique Values

```
# create value to store only categorical columns in the dataset
categorical_col = sapply(d3, function(x) is.factor(x))
# check for any unique values in all categorical columns of the dataset
unique_categorical_values = lapply(d3[, categorical_data], unique)
# Display the result
unique_categorical_values
```

Note. Line 139 first creates a value to only filter out the columns that stores categorical data. Then line 142 creates a data frame to store a list of the categorical columns and their respective levels. Line 145 is then run to show the results.

Appendix D. Content Investigation Codes

Code to Create a New Data Frame “numerical_data” to Store All Numerical Data in the Dataset

```
# create a data frame with only numeric variables inside to calculate
their key statistics
numerical_data = d3[, sapply(d3, is.numeric)]
```

Code to Calculate the Mean of Each Numerical Data

```
# calculate the mean of all numerical data and show result
mean_values = sapply(numerical_data, mean, na.rm = TRUE)
mean_values
```

Code to Calculate the Median of Each Numerical Data

```
# calculate the median of all numerical data and show result
median_values = sapply(numerical_data, median, na.rm = TRUE)
median_values
```

Code to Calculate the Standard Deviation of Each Numerical Data

```
# calculate the standard deviation of all numerical data and show result
sd_values = sapply(numerical_data, sd, na.rm = TRUE)
sd_values
```

Code to Calculate the Range of Each Numerical Data

```
# calculate the range of all numerical data and show result
range_values = sapply(numerical_data, range, na.rm = TRUE)
range_values
```

Code to Create a New Data Frame “categorical_data” to Store All Categorical Data in the Dataset

```
# create a data frame with only categorical variables inside to calculate
their proportions
categorical_data = sapply(d3, is.factor)
```

Code to Produce Pie Chart for All Categorical Data

```
# reason of doing this is because all 24 categorical data cannot be
shown at once
# create data frame to store first 12 variables to produce as a pie
chart
first_12_cat_data = names(categorical_data)[1:12]
# adjust display to show 12 pie charts proportions at once
par(mfrow = c(3, 3))
# Loop through each categorical variable and create pie charts with
percentage labels for first 12 data
lapply(first_12_cat_data, function(column) {
  if (is.factor(d3[[column]]) || is.character(d3[[column]])) {  # Ensure
it's a categorical variable
    counts <- table(d3[[column]])  # Get counts for each category
    percentages <- round(100 * counts / sum(counts), 1)  # Calculate
percentages
    labels <- paste0(names(counts), ": ", percentages, "%")  # Labels
with percentages

    pie(counts, labels = labels, main = paste("Pie chart of", column))
  }
})
# this execution cannot display the last eight variable, to solve this,
another code is run below
# run this code separately
# define the first remaining 4 variables that are unable to produce
specific_variables_1 = c("internet", "math_guardian", "math_schoolsup",
"math_famsup")
# adjust layout
par(mfrow=c(2, 2))
# Loop through the specific variables
for (var in specific_variables_1) {
  # Create a table of counts for the current variable
  counts <- table(d3[[var]])
  # Create the pie chart
  pie(counts, main = paste("Pie chart of", var), labels =
paste0(round(100 * prop.table(counts), 1), "%"))
}
# define the other remaining 4 variables that are unable to produce
specific_variables_2 = c("math_paid", "math_activities", "math_higher",
"math_romantic")
# adjust layout
par(mfrow=c(2, 2))
# Loop through the specific variables
for (var in specific_variables_2) {
  # Create a table of counts for the current variable
  counts <- table(d3[[var]])
  # Create the pie chart
  pie(counts, main = paste("Pie chart of", var), labels =
paste0(round(100 * prop.table(counts), 1), "%"))
}

# Adjust layout
par(mfrow = c(3, 3))
# Loop through each categorical variable and create pie charts with
percentage labels for last 12 data
lapply(names(d3)[categorical_data], function(column) {
  counts <- table(d3[[column]])  # Get the counts for each category
percentages <- round(100 * counts / sum(counts), 1)  # Calculate
percentages
  labels <- paste0(names(counts), ": ", percentages, "%")  # Create
labels with percentages
```

Note. This code divided into four part: extracting the first 9 variables and produce the pie charts of these variables, extracting the next 4 variables and producing the pie chart, extracting another 4 categorical variables and producing their pie chart and extracting the remaining 9 variables and producing the pie charts. The reason for this is because the display box is too small to produce all 26 categorical data. Thus, this action is taken.

Code to Produce Box Plot for “avg_studytime”, “avg_alcohol” and “avg_famrel”

```
boxplot(d3$avg_famrel, main = "Student's Relationship with Family", ylab
= "Relationship Rank", col = "green")
boxplot(d3$avg_studytime, main = "Student's Average Study Time", ylab =
"Study Time", col = "green")
boxplot(d3$avg_alcohol, main = "Student's Average Alcohol Consumption",
ylab = "ALcohol Consumption", col = "green")
```

Code to Create Histograms for Potential Imbalanced Relevant Variables

```
# produce histogram for potential imbalances
# avg_famrel
hist_avg_famrel =  hist(d3$avg_famrel,
                   main = "Histogram of Students' Family Relationship",
                   xlab = "Family Relationship Rank",
                   ylab = "Frequency",
                   col = "pink",
                   border = "black",
                   breaks = 5)
# Add text labels to the histogram
text(hist_avg_famrel$mids, hist_avg_famrel$counts, labels =
hist_avg_famrel$counts, pos = 3, cex = 0.8, col = "black")
# avg_studytime
hist_avg_studytime =  hist(d3$avg_studytime,
                      main = "Histogram of Students' Average Study
Time",
                      xlab = "Study Rank",
                      ylab = "Frequency",
                      col = "pink",
                      border = "black",
                      breaks = 5)
# Add text labels to the histogram
text(hist_avg_studytime$mids, hist_avg_studytime$counts, labels =
hist_avg_studytime$counts, pos = 3, cex = 0.8, col = "black")
# avg_alcohol
hist_avg_alcohol =  hist(d3$avg_alcohol,
                    main = "Histogram of Students' Average Alcohol
                    Consumption",
                    xlab = "Alcohol Consumption Rank",
                    ylab = "Frequency",
                    col = "pink",
                    border = "black",
                    breaks = 5)
# Add text labels to the histogram
text(hist_avg_alcohol$mids, hist_avg_alcohol$counts, labels =
hist_avg_alcohol$counts, pos = 3, cex = 0.8, col = "black")
```

Code in R Studio to Produce Pie Chart of Proportions for “math_famsup”, “port_famsup”, “math_health_category” and “port_health_category”

```
# check for categorical data imbalances using pie charts
# math_famsup
math_famsup_table = table(d3$math_famsup)
# Create the math_famsup pie chart with labels
pie(math_famsup_table,
    main = "Pie Chart of Family Support in Math Course Students",
    col = rainbow(length(math_famsup_table)),
    labels = paste(names(math_famsup_table),
                   round((math_famsup_table / sum(math_famsup_table)) *
100, 1), "%"))
# port_famsup
port_famsup_table = table(d3$port_famsup)
# Create the math_famsup pie chart with labels
pie(port_famsup_table,
    main = "Pie Chart of Family Support in Portuguese Course Students",
    col = rainbow(length(port_famsup_table)),
    labels = paste(names(port_famsup_table),
                   round((port_famsup_table / sum(port_famsup_table)) *
100, 1), "%"))
# math_health_category
math_health_category_table = table(d3$math_health_category)
# Create the math_famsup pie chart with labels
pie(math_health_category_table,
    main = "Pie Chart of Math Course Studets' Health",
    col = rainbow(length(math_health_category_table)),
    labels = paste(names(math_health_category_table),
                   round((math_health_category_table /
sum(math_health_category_table)) * 100, 1), "%"))
# port_health_category
port_health_category_table = table(d3$port_health_category)
# Create the math_famsup pie chart with labels
pie(port_health_category_table,
    main = "Pie Chart of Portuguese Course Studets' Health",
    col = rainbow(length(port_health_category_table)),
    labels = paste(names(port_health_category_table),
                   round((port_health_category_table /
sum(port_health_category_table)) * 100, 1), "%"))
```

Code to Find Correlation Coefficient of Between “avg_famrel” and “avg_alcohol”

```
# avg_famrel with avg_alcohol
famrel_alcohol_correlation <- cor(d3$avg_famrel, d3$avg_alcohol, method
= "pearson")
famrel_alcohol_correlation
```

Code to Find Correlation Coefficient of Between “avg_studytime” and “avg_alcohol”

```
# avg_studytime with avg_alcohol
studytime_alcohol_correlation <- cor(d3$avg_studytime, d3$avg_alcohol,
method = "pearson")
studytime_alcohol_correlation
```

Code in R Studio to Carry Out Chi-Squared Test on “math_famsup” with “math_health_category”

```
# chi-squared test on math_famsup with math_health_category
math_chisq_test = chisq.test(table(d3$math_famsup, 
d3$math_health_category))
print(math_chisq_test)
# Interpretation of results through message
if (math_chisq_test$p.value < 0.05) {
  print("Hypothesis accepted. There is a significant association between 
family support and health status.")
} else {
  print("Hypothesis rejected. There is no significant association 
between family support and health status.")
}
```

Code in R Studio to Carry Out Chi-Squared Test on “port_famsup” with “port_health_category”

```
# chi-squared test on port_famsup with port_math_category
port_chisq_test = chisq.test(table(d3$port_famsup, 
d3$port_health_category))
print(port_chisq_test)
# Interpretation of results through message
if (port_chisq_test$p.value < 0.05) {
  print("Hypothesis accepted. There is a significant association between 
family support and health status.")
} else {
  print("Hypothesis rejected. There is no significant association 
between family support and health status.")
}
```

Code to Find Correlation Coefficient of Between “math_famrel” and “math_health”

```
# math_famrel and math_health
math_famrel_health_correlation <- cor(d3$math_famrel, d3$math_health, 
method = "pearson")
math_famrel_health_correlation
```

Code Create Violin Plot for “math_famrel” against “fam_health”

```
ggplot(d3, aes(x = as.factor(math_health), y = math_famrel)) +
  geom_violin(fill = "pink", color = "black") +  # Violin plot for 
distribution
  geom_boxplot(width = 0.1, fill = "white", outlier.shape = NA) +  # Box 
plot for median and quartiles
  stat_summary(fun = median, geom = "point", size = 2, color = "red") + 
# Red point for the median
  xlab("Health Levels") +
  ylab("Family Relationship") +
  ggtitle("Violin Plot with Box Plot: Health Levels and Family 
Relationship of Students in Math Course") +
  theme_minimal()
```

Code to Find Correlation Coefficient of Between “port_famrel” and “port_health”

```
# port_famrel and port_health
port_famrel_health_correlation <- cor(d3$port_famrel, d3$port_health,
method = "pearson")
port_famrel_health_correlation
```

Code Create Violin Plot for “port_famrel” against “port_health”

```
# violin plot for port_famrel and port_health
# install package ggplot2
library(ggplot2)
ggplot(d3, aes(x = as.factor(port_health), y = port_famrel)) +
  geom_violin(fill = "pink", color = "black") +  # Violin plot for
distribution
  geom_boxplot(width = 0.1, fill = "white", outlier.shape = NA) +  # Box
plot for median and quartiles
  stat_summary(fun = median, geom = "point", size = 2, color = "red") +
# Red point for the median
  xlab("Health Levels") +
  ylab("Family Relationship") +
  ggtitle("Violin Plot with Box Plot: Health Levels and Family
Relationship of Students in Portuguese Course") +
  theme_minimal()
```

Appendix E. Visualization Codes

Code to Create a Scatter Plot Between “avg_famrel” and “avg_alcohol”

```
# Scatter plot to visualize the relationship between avg_famrel and
avg_alcohol
plot(d3$avg_famrel, d3$avg_alcohol,
     xlab = "Average Family Relationship",
     ylab = "Average Alcohol Consumption",
     main = "Scatter Plot of Family Relationship vs Alcohol
Consumption",
     pch = 19,
     col = "black")
# Fit a linear model
model_famrel_alcohol = lm(avg_alcohol ~ avg_famrel, data = d3)
# Add the regression line to the scatter plot
abline(model_famrel_alcohol, col = "orange", lwd = 2)
```

Code to Create Scatter Plot Between “avg_studytime” and “avg_alcohol”

```
# Scatter plot to visualize the relationship between avg_studytime and
avg_alcohol
plot(d3$avg_studytime, d3$avg_alcohol,
     xlab = "Average Study Time",
     ylab = "Average Alcohol Consumption",
     main = "Scatter Plot of Study time vs Alcohol Consumption",
     pch = 19,
     col = "black")
# Fit a linear model
model_studytime_alcohol = lm(avg_alcohol ~ avg_studytime, data = d3)
# Add the regression line to the scatter plot
abline(model_studytime_alcohol, col = "orange", lwd = 2)
```

Code to Create Mosaic Plot to View the Relationship Between “math_famsup” and “math_health_category”

```
# install vcd packages first
install.packages(vcd)
library(vcd)
mosaic(~ d3$math_famsup + d3$math_health_category,
       data = d3,
       gp = gpar(fill = c("yellow", "orange", "green","skyblue")),
       main = "Mosaic Plot of Family Support and Health Category for
Students in Math Course")
```

Code to Create Mosaic Plot to View the Relationship Between “port_famsup” and “port_health_category”

```
# create a mosaic plot for port_famsup and port_health_category
mosaic(~ d3$port_famsup + d3$port_health_category,
       data = d3,
       gp = gpar(fill = c("yellow", "orange", "green","skyblue")),
       main = "Mosaic Plot of Family Support and Health Category for
Students in Portuguese Course")
```

Appendix F. Data Set File

https://drive.google.com/drive/folders/1qcIrvuOn_rnzJZR3ZHEQvYbCNgC_l-W_?usp=share_link