TOPIC: EXPLORING THE IMPACT OF ALCOHOL CONSUMPTION ON STUDENT ACADEMIC PERFORMANCE

Alcohol consumption among college students is a prevalent and concerning issue worldwide.Despite extensive research on the subject, there remains a need for comprehensive studies examining the direction correlation between alcohol use and academic performance. Project objectives: to investigate the relationship between alcohol consumption and academic performance among college students. I have downloaded my dataset from **dataworld**. The dataset name is **student-lpo** and it has 649 rows and 31 columns the. The dataset was not in english it was in portuguese so it was difficult to know the meaning of column name I had to search on internet the meaning of the column name in English then I changed the column name into English so that I can understand well the data and I can also get more insights from the data. Here are new names of dataset's column after changing them to English names: 'School' 'Sex', 'Age', 'Address', 'Family_Size', 'Parental_Status','Mother_Education', 'Father_Education', 'Mother_Job', 'Father_Job','Reason', 'Guardian', 'Travel_Time', 'Study_Time', 'School_Support','Family_Support', 'Paid_Classes', 'Activities', 'Nursery','Higher_Education', 'Internet_Access', 'Romantic_Relationship','Family_Relationship', 'Free_Time', 'Go_Out','Workday_Alcohol_Consumption', 'Weekend_Alcohol_Consumption','Health_Status', 'Absences', 'First_Semester_Grade','Second_Semester_Grade'. I have also put in consideration other variables that can affect student academic performance but my focus is on effects of alcohol consumption so in the dataset there is these two column: Workday_Alcohol_Consumption', 'Weekend_Alcohol_Consumption' explaining about alcohol been drunk during weekdays and the other one is alcohol been drunk during weekend by students

Loading the dataset

I have loaded the dataset using the pandas package that has the method pd.read_csv() used to read data that are in comma separated value. After loading the dataset I have checked the format of it by viewing five rows using head() function and the dataset is in wide format since every column represents a unique observation. Then I checked the number of rows and number of columns using the shape attribute of pandas. Then after I checked the data types of each column using the info() function of pandas, the dataset has integer data types, string data types and float data types.

Changing column names

The data was recorded in Portuguese as you can see the name of the dataset so other column names are not in english so I had to change them for better analysis. I used the rename() function in pandas that take parameter of new column names to change

Checking data types of columns

I have checked datatypes of each column using info() function in pandas it returns column name and it's data types

Dropping other columns that were not needed

I have dropped a column named "Address" and "school " from the dataset, these columns will not affect anything for analyzing the dataset

Checking and filling missing values in the dataset

I used this isnull() function and sum() function that return the number of missing values in each column and I used fillna() and the mean of each column to fill in the missing values.

Removing outliers

Firstly I have to visualize the data if it has outliers using boxplot. Boxplot is the best at showing outliers and spread of dataset

I have removing outliers using scatter plot method here I have calculated interquartile range using code and also I calculated lower bound and upper bound and then I have removed outliers, using code that checked every number in column if it is greater than upper bound i have removed that number and when a number is less than lower bound I have also removed it. Outliers prevent deep insights of the dataset. After removing outliers I plotted graph again to show if outliers are still exit but the plotted boxplot are not showing any outliers

Dealing with duplicates in the dataset

Duplicates are rows that are similar in the dataset they need to be removed so that they can not affect the process of getting insights from the data

I checked for duplicated rows using duplicated() function in pandas and after finding rows that are similar in the dataset then I removed them using drop_duplicates() function in pandas

Transforming dataset

I have changed the format of numerical columns to have a mean of 0 and standard deviation of 1 using z-score normalization. To transform numerical columns I used the standardscaler class from the sklearn package.

Creating new variables

I have introduced three new variables in the dataset using other existing columns in the dataset here are new introduced variables:

1. Total_alcohol_consumption

   Is the sum of workday and weekend alcohol consumption
2. **Average_village**

   Is the average of first and second semester grades
3. **Study_eeficient**

   Is the ratio of study time to travel time

Removing outliers in new introduced columns

I have also removed outliers for the new introduced columns

Encoding categorical variables

This is changing the categorical column's value into numbers that increase the robustness of the model. I have changed these columns **'Mothher_Job', 'Father_Job', 'Reason', 'Guardian', 'School_Support' into numbers** into numbers. To incode I used LabelEncoder class from the sklearn package.

Exploratory data analysis

1. **Finding summary statistics**

   This is finding mean, average, minimum value, maximum value, standard deviation, 25 percent, and 75 percent of the dataset. To do this I use describe() function in pandas

2. **Data visualization**

   - **Histograms**

     Histograms are useful for understanding the distribution of individual column

     Histograms are showing that semester grades have much contributed to the dataset second by average grades then absence from the school

   - **Boxplot**

     It is used to show spread or skewness of the column and it is also used to show if a column has outliers**.** After viewing boxplot it shows that columns of workday_alcohol_consumption and weekday_alcohol_consption are showing long whiskers in the left side this suggest low consumption of alcohol compared to the whole population

     First, second semester and average grades shows normal distribution it suggests balanced and effective education approach

     Study time,absence, and age also shows long left whisker suggesting small group of student study small hours, small group of people have some days not going to school and small group of people have low ages

   - **Correlation matrix**

     It helps to understand relationship between numerical columns by showing numbers between -1 ,0 to +1

     When it shows -1 it means there's a perfect negative correlation between those two columns, when it shows 0 it means there is no correlation and when it shows +1 it means there is perfect positive correlation of the column. From the graph number it is showing that alcohol consumption even workday or weekend have inverse proportion because correlation is negative meaning when alcohol consumption increases the lower grade you will get in first semester or second semester

     Study time and grade are showing post correlation number meaning increase in study time it causes increase in grade in all semester

     Absence and grade are showing negative correlation number meaning the more absent the lower grades you will get from all semester

- **Pair plots**

Pair plots provide scatter plots for each pair of numerical columns, useful for identifying relationships and patterns of numerical columns. From the plotted graph it is showing points forming upward sloping line for any column pair meaning positive correlation for any columns pair and if point would form downward sloping line I would say it's negative correlation and if also points would be in randomly scattered I would say there's no any relationship between columns

For example alcohol consumption in workday or weekend with semester grades in first or second semester is showing downward points meaning negative correlation increase in one cause decrease in other

Age and study time is showing no relationship mean increase in another variable has no effect in another variable

Study time time and semester grade has points pointing upward meaning increase in one variable has increase in another variable.

Saving transformed dataset

I have saved the transformed data in comma separated values format using to_csv() function found in pandas