PROJECT
DOCUMENTATION

STUDENT ALCOHOL CONSUMPTION

PRESENTED TO PROF. GIOVANNI PROFETA
PRESENTED BY IVAN DUVNJAK
NAYANA MANJALI

Table of Contents

Abstract

Alcohol consumption by students is widely regarded as a socially relevant issue, and it is a source of concern for many parents and school administrators. A number of studies have indicated that it has negative consequences for students, such as leading to dropouts and poor academic performance. However, recent research indicates that alcohol does not always have the same effect on student performance. In fact, some studies suggest that drinking may even help improve their grades. This discrepancy in results has led to much debate about the pros and cons of alcohol use among college-aged individuals. Some believe that consuming alcoholic beverages poses significant risks to both the individual and society at large while others argue that moderate drinking doesn't necessarily lead to harmful outcomes.

The purpose of this visualisation is to determine whether or not drinking alcohol impacts students' academic grades, as well as how their surroundings can influence their alcohol use. We would like to, especially thank professor Giovanni for his guidance, reviews, and recommendations during the development of the project. Also for his availability outside of class hours.

Introduction and Interpretation

For this analysis, we took the dataset, Student Alcohol Consumption, from UC Irvine Machine Learning Repository. This data was gathered by a poll of secondary school students from two distinct schools in Portugal: Gabriel Pereira (Évora) and Mousinho da Silveira (Portalegre), who were studying Mathematics and Portuguese. From the result, a dataset was formed that describes the student's background, interests, study information, grades, extracurricular activities, etc.

First of all, we were curious to know how a student's surroundings can affect his/her alcohol consumption because we all have heard that a child's environment can influence their behavior. For instance, if a student lives in a city with several pubs, clubs, and restaurants, he/she will likely be more vulnerable and susceptible to alcohol, whilst a student who lives in a city with very few or no pubs, clubs, and such places will have less access to alcoholic beverages.

The city of Évora, where Gabriel Pereira (GP) school is located, is a medium-sized city with many historical sites. Monuments and structures are the major magnets for visitors due to their long and rich history. Furthermore, the municipality and other organisations support a significant variety of "Festas Populares" (popular celebrations) honouring saints, holidays, "Feiras" (fairs), and cultural events (like televised musical performances). On the other hand, Portalegre, the location of the other school, Mousinho da Silveira (MS), is a little city with religious monuments, museums, and so on. We used Mapbox to create maps of the two

cities and highlighted locations where alcohol may be sold, such as bars, restaurants, and pubs. Evora has more pubs, restaurants, and bars than Portalegre, as indicated on the map. This might be because Evora is a tourism destination and must suit the demands of visitors. We obtained a different outcome from the map's output. Students in the Portalegre area consume somewhat more alcohol than students in Evora.

The second question that sprang to mind was if there was any link between students' alcohol use and their academic performance. It is widely held that those who consume a lot of alcohol perform worse academically than those who do not. To test if the latter theory is correct, we chose to use Rawgraphs to create an alluvial plot.

The two factors we considered were the alcohol consumption range and the grade range. The alcohol consumption column was created by averaging the two columns Dalc (workday alcohol consumption) and Walc (weekend alcohol consumption), which were then sorted from high to low. The output of the graph sustains the aforementioned hypothesis. The students who drink too much often fail or score low marks. But this does not mean that the students who consume less or no alcohol always get a good score.

Data sources and Data pre-processing

The Student Alcohol consumption data that was generated in April 2008, was obtained from UC Irvine Machine Learning Repository.

The dataset is divided into two files, students taking mathematics csv and students taking the Portuguese language csv, or in other words according to the courses the students take. Each dataset has essentially 33 fields that describe the student's background, interests, study information, grades, extracurricular activities, etc. The dataset on students taking mathematics courses has 395 rows whereas the one on the Portuguese language has 650 rows. Both of them contain 33 columns of integer, boolean, and string data types.

In order to pre-process the data, we imported the CSV file in python. We then proceeded to clean the dataset. When checked there were no missing values or duplicated data present in the dataset. For a better analysis, we decided to concatenate both datasets which then resulted in a dataset with 1044 rows and 33 columns. We then analysed some columns of data with the aim to search for interesting questions.

For the second analysis question, to show the relationship between the grade and alcohol consumption, presenting the grade column clearly was necessary. For this, we choose to categorise the grade instead of keeping it as integers. To do this we looked for the Portuguese grading system and found that the grades were divided into 5 categories. So we decided to categorise the provided grades accordingly. Next, we had two columns of interest, Dalc (working day alcohol consumption) and Walc(weekend alcohol consumption).

To get them in a single column we took the average between the two. In order to have a better overview, we categorised the column into 5 different ranges of alcohol consumption.

Interface design

We put one of our data visualisations as the first thing to see when opening the site. The data visualisation shows two maps for each city that we studied. We put them next to each other so the user can confront them and have an immediate idea of what kind of city they are, this way the user should notice that one has more points of interest than the other.

After that we put two tables under each city map referencing each city's school that we analysed. They show information about the alcohol consumption regarding the students of those schools. Similarly to the maps the tables are put next to each other, this is done to help the comparison of the data and they are under each city to be able to tell more easily which city the table is referencing.

Next thing we start introducing the project, the first part references the first data visualisation, while the second part is about the second data visualisation which we put after the text but before the conclusion. We decided to put it after to enforce and prove what we said in the text earlier and to help the conclusion of the project.

This data visualisation is an alluvial diagram, we choose it because it shows that most students that have a high alcohol consumption usually do worse than the others. Since the thicker splits of “High” and “Very high” end up in failure or sufficient.

As said earlier we put the conclusion after and the links to the dataset and the visualisation protocol.

Data visualisations

The best method we could illustrate the correlation between the student's surroundings and alcohol use for the first question was through a map. As a preliminary analysis, before proceeding to plot the map we have gone through several websites such as the two cities' official websites, TripAdvisor, etc to get a deeper understanding of the localities. Later we were able to plot the map thanks to Mapbox. We looked for the two locations Évora and Portalegre in Mapbox in order to create new datasets out of them. Afterward, we saved and exported the file, which is later available in our map. We set the icons for the alcohol-accessible spots, and we had to set the position of the schools manually, as it was not already present. We then customised the map by adding separate icons, changing the colours, etc. We put the two maps on the website using a container to display them one next. Underneath two tables with the info about the sites

For our next question, we chose to plot an alluvial diagram using a website our teacher provided during the lecture, Rawgraphs. The biggest benefit of this diagram is its ability to depict complex, multivariate processes such as the allocation of funds in a concise, easy-to-grasp visual manner. We chose column G3 which provided the information on the student's final grade, therefore the summed average of G1 and G2. As the datatype of the interested column was integers (values from 0-20) we categorised them according to the Portuguese

grading system with the help of python. As said earlier in the document, in the processing part, we used the alcohol consumption as the other column. After plotting the graph in raw graphs, we fixed some details with svg viewer (labels size, position, ecc). As final step we added a script (provided by our teacher) to highlight the lines of the alluvial diagram.

Next Step

As mentioned before this dataset has many columns that describe each aspect of a student's life. It contains all details such as his/her age, family background, relationship background, academics, interests, and more. There is surely a lot more content for us to explore in the future with this dataset. For the time being, we were only able to select and analyse some of the essential columns that have a stronger relationship with students' alcohol consumption. In the future, it would be interesting to see how the emotional side of a student can influence the drinking habit. Also, it would be more explorative if we can compare this dataset with others and do some analysis.