VISUALISATION PROTOCOL

STUDENT ALCOHOL CONSUMPTION

PRESENTED TO
PROF. GIOVANNI PROFETA

PREPARED BY

IVAN DUVNJAK
NAYANA MANJALI
DOMINIC STEFAN MEIER

Table of Contents

Introduction

The Student Alcohol consumption data that was generated on April 2008, was obtained from UC Irvine Machine Learning Repository . This data was collected through a survey of secondary school students from two different schools in Portugal: Gabriel Pereira (Évora) and Mousinho da Silveira (Portalegre), taking Mathematics and Portuguese language courses. The dataset is divided into two files according to these courses they take. Each dataset has essentially 33 fields that describe the student's background, interests, study information, grades, extracurricular activities, etc.

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

- **school** - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
- **sex** - student's sex (binary: 'F' - female or 'M' - male)
- **age** - student's age (numeric: from 15 to 22)
- **address** - student's home address type (binary: 'U' - urban or 'R' - rural)
- **famsize** - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
- **Pstatus** - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
- **Medu** - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- **Fedu** - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- **Mjob** - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- **Fjob** - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

- **reason** - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
- **guardian** - student's guardian (nominal: 'mother', 'father' or 'other')
- **traveltime** - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- **studytime** - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- **failures** - number of past class failures (numeric: n if 1<=n<3, else 4)
- **schoolsup** - extra educational support (binary: yes or no)
- **famsup** - family educational support (binary: yes or no)
- paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- **activities** - extra-curricular activities (binary: yes or no)
- **nursery** - attended nursery school (binary: yes or no)
- **higher** - wants to take higher education (binary: yes or no)
- **internet** - Internet access at home (binary: yes or no)
- **romantic** - with a romantic relationship (binary: yes or no)
- **famrel** - the quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- **freetime** - free time after school (numeric: from 1 - very low to 5 - very high)
- **goout** - going out with friends (numeric: from 1 - very low to 5 - very high)
- **Dalc** - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- **Walc** - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- **health** - current health status (numeric: from 1 - very bad to 5 - very good)
- **absences** - number of school absences (numeric: from 0 to 93)

These grades are related to the course subject, Math or Portuguese:

- **G1** - first-period grade (numeric: from 0 to 20)
- **G2** - second-period grade (numeric: from 0 to 20)
- **G3** - final grade (numeric: from 0 to 20, output target)

Abstract

The dataset[1] on students taking mathematics course has 395 rows whereas the one on the Portuguese language has 650 rows. Both of them contain 33 columns of integer, boolean, and string datatypes. There we 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.

Initially, we wanted to know how a student's surroundings can affect his/her alcohol consumption, for instance, if a student lives in a city with lots of pubs, clubs, and restaurants he/she would be more exposed to alcohol whereas one who has less or no pubs, clubs and such places in his/her city the student will have less access to alcohol. This could be visualised thanks to Mapbox. We looked for the city in which the schools are located and imported it into the website. The exact location of the schools was not found in the map therefore we had to add them manually. We then filtered out the places in which alcohol sale could be possible, for instance bars, restaurants, pubs etc. The map was then customized by adding icons, changing the colors, etc.

The second question that arose in our mind was if there were any particular relation between students' alcohol consumption and their performance in school. As it is a common belief that one who drinks a lot has a lower academic performance compared to one who avoids alcohol. For

[1] Data generated on April 2008

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April 2008, EUROSIS, ISBN 978-9077381-397.
Data Source: https://archive.ics.uci.edu/ml/datasets/student+performance

this visualiasation we mainly used the an open source tool for data visualization called Rawgraphs. We choose the columns, G3 (student's final grade,) Dalc (weekday alcohol consumption) and Walc (weekend alcohol consumption). In order to portray the visualization in a more understandable way we choose to merge the columns Dalc and Walc into one by taking the mean, creating a new column Alcohol consumption range. Furthermore, we also decided to group both the columns into different ranges. These operations were done using python. Then we proceeded by importing the modified dataset into Rawgraphs. We choose to present the data through alluvial diagram as it is a great tool for exploring categorical data.

List of all the actions you performed

Data Handling

- Importing the student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets in the python notebook
- Handling the datasets by checking for duplicates, null values
- Concatenate the two different datasets into one for an easier analysis.

Maps of the Cities

- Firstly, from the website from which we found the dataset, we searched for the exact location of the secondary schools, which where both located in Portugal.

- We looked for the two locations Évora and Portalegre in Mapbox in order to create new datasets out of them. Afterwards, we saved and exported the file, so it is later available in our map.
- Then we started with the basic template, we had to set the position of the schools manually, as it was not already present.
- Afterwards, we filtered out the position of the restaurants, pubs and bars and marked it in the map
- We then customized the map by adding separate icons, changing the colors etc.

The Alluvial Diagram of Grade and Alcohol Consumption

- Inorder to see the relationship between the alcohol consumption and student's scolastic performance, we thought to plot an alluvial plot, a form of sankey diagram that is a great tool to explore the categorical variables.
  - We choose the column G3 which provided the information on 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 thought to categorize them according to the Portughese grading system. For this we used python.

Under the Portuguese system, grades are given on a **scale from 0 to 20, the minimum passing grade being 10**.
The indications below can be followed as a reference:
18 to 20 - Excellent
16 and 17 - Very good
14 and 15 - Good
10 to 13 - Sufficient
0 to 9 - Fail

```
df['Grade_range']=df['g3']
for i in range(len(df['g3'])):
    if 18 <= int(df['g3'][i]) <= 20:
        df['Grade_range'][i]='Excellent'
    elif 16 <= int(df['g3'][i]) <= 17:
        df['Grade_range'][i]='Very Good'
    elif 14 <= int(df['g3'][i]) <= 15:
        df['Grade_range'][i]='Good'
    elif 10 <= int(df['g3'][i]) <= 13:
        df['Grade_range'][i]='Sufficient'
    else:
        df['Grade_range'][i]='Fail'
```

- 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 categorized the column into 5 different ranges of alcohol consumption as shown below.

```
df['alc_range']=df['average_alc']
for i in range(len(df['average_alc'])):
    if int(df['average_alc'][i])== 5 :
        df['alc_range'][i]='Very High'
    elif 4 <= int(df['average_alc'][i]) <= 4.5:
        df['alc_range'][i]='High'
    elif 3 <= int(df['average_alc'][i]) <= 3.5:
        df['alc_range'][i]='Medium'
    elif 2 <= int(df['average_alc'][i]) <= 2.5:
        df['alc_range'][i]='Low'
    elif 1 <= int(df['average_alc'][i]) <= 1.5:
        df['alc_range'][i]='Very Low'
```

- Next we imported the updated csv file of Rawgraphs, setting comma as the seperator
- We set the two selected columns as the steps and set an ordinal color scale (Light blue to Red) to highlight the level of alcohol consumption.