Using KDD and CRISP-DM data mining process to Analyse Student Alcohol Consumption

Affects of Alcohol Consumption in Academic's

Navjot Singh Virk: x13112406
Software Development, 4th Year BSc. Honors in Computing
National College of Ireland
Dublin, Ireland
Virksaabnavjot@gmail.com

In this project, for analysis purpose I have used 2 datasets of Portuguese student on 2 different courses Mathematics and Portuguese Language available at https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION

Abstract

The legal age for alcohol consumption is 18 years in most countries but teenagers find ways for consumption of this fluid and sometimes this leads in frustration and negative and uncontrolled behavior in our world future generation the teenagers. For this analysis the 2 datasets used are about Portuguese students in secondary school for two different courses (Mathematics and Portuguese language). The purpose of the analysis is to identify student alcohol consumption, the factors contributing to it and show the results how alcohol may affect the students in different ways and long term results of it and the potential risks. The chosen datasets contain student performance and other attributes that will help study and find out how alcohol consumption affects student grades, behavior like their relationship with the family and other other aspects of academic life. For, the project I had the choice to select any 2 datasets. The motivation for selecting the Alcohol consumption datasets was the results of this analysis can help understand student problems that may be leading to more consumption of this fluid and possibly suggest ways to fix them and help the society. The following analysis has been done using some aspects of KDD process and CRISP-DM process model for data mining.

I. Introduction

The value of education is high in everyone's life and a well educated person contributes to the society and live a happy life and school play a major role in our lives it's the place where everyone starts learning new things.

These days' alcohol has become a highly regarded party and enjoyment drink and people in social gatherings crave alcohol and people that don't drink it feel left alone and its kind of considered a taboo for people not drinking when they go out with friends. This kind of thinking have had bad impact on the today's teenagers as drinking these days is highly regarded as cool and no drinking not so cool. Which motivates student's teenagers to start consumption of alcohol at very early age as every student in secondary school wants to be fun and popular and no one wants to be left alone and also to not follow the whole group of people and stand for your decision takes a lot of self control on nerves which only few adults can successfully do hence the young people of the world these days are exposed to a mindset and thinking of "alcohol is life" that is regarded fun and aesthetic lifestyle but in reality its not so good.

Most countries around the world have age restrictions on alcohol consumption but young people find ways to work around. Early, exposure to alcohol in many cases leads to people becoming addicted to this fluid for life. Alcohol, not only costs money it costs life, in the past alcohol has been linked to suicide, self harm, loss of memory and many more devastating effects. It affects mood and its consumption makes a person less conscious and lead to many unwanted mistakes and people may face for life. Alcohol also contributes to bad decision making and young people taking bad steps, sexual frustration and unwanted pregnancy at very early age the time which should be consumed for education and building future.

Alcohol consumption has bad affect on students academically, like failing exams, bad behavior and relationship with their parents and health wise alcohol is deadly for young people as their bodies are delicate compared to adults and consumption of alcohol may affect their health and weaken their bodies and reduce immunity and may cause a lot of diseases and mental health problems.

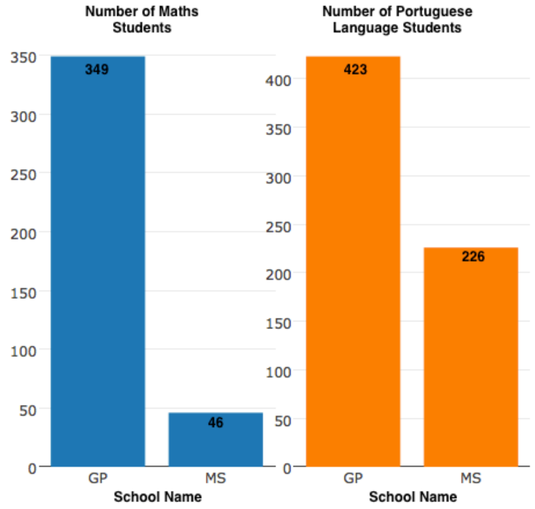
For the analysis, the chosen datasets are about Portuguese students in two schools Gabriel Pereira School and Mousinho da Silveira School and about two courses Mathematics and Portuguese language course which was composed by P. Cortez and A. Silva, at University of Minho in Portugal.

II. Data

There are two datasets used in the analysis are below are some quick pointers that will help understand this report and datasets –

Mathematics course dataset contains 393 rows (student data) and 33 columns (attributes) and out of 393 student data 208 females and 187 male students & GP school has 349 students in math’s course and MS has 46 students in its math’s course.

Portuguese course dataset contains 649 student records and same no. of attributes as the other course out of which are 383 females and 266 are male students & GP school has 423 students in math’s course and MS has 226 students in its math’s course.



Attribute	Description
school	Students 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 – 4)
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	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)
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)

And some analysis has been done by combining both the dataset and adding one new attribute called “subject” (course) in the combined version of the dataset and set it to numeric (1 – Math's and 2 – Portuguese language course).

III. Methodology

The two main methodologies that have been applied in this analysis are KDD (Process for useful knowledge extraction from data) and CRISP-DM (a process model for data mining) and their traces can be found in the whole analysis.

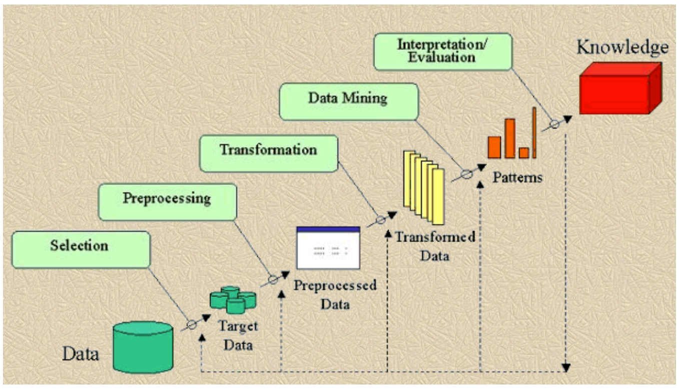


Figure: Steps of KDD Process

Image Source: https://goo.gl/3msNdV

The above figure shows the steps involved in KDD for extracting useful information and this analysis is a good example of KDD implementation. Following the process of KDD and little description and how they were implemented in this analysis (Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34):

1- Developing and Understanding of prior knowledge.

   Thinking what I want to achieve out of this analysis and if it could be beneficial to the society and read papers on KDD and CRISP-DM for Data mining.

2- Creating a target dataset: includes selecting dataset on which the discovery is to be made.

   Through research selected the alcohol consumption dataset it was not easy as there was a requirement for two datasets and most found datasets were single datasets.

3- Data Cleaning and Data Processing: removing any unnecessary data.

   The found datasets that I worked with were very well formatted already as some analysis had been performed on these datasets but my analysis finds more additional information on top of it and small data cleaning and processing were conducted.

4- Data Transformation: transforming data into forms appropriate for analysis.

   Used the csv file and imported the data in RStudio to start working with it.

5- Data Mining: apply intelligent method and extract data patterns.

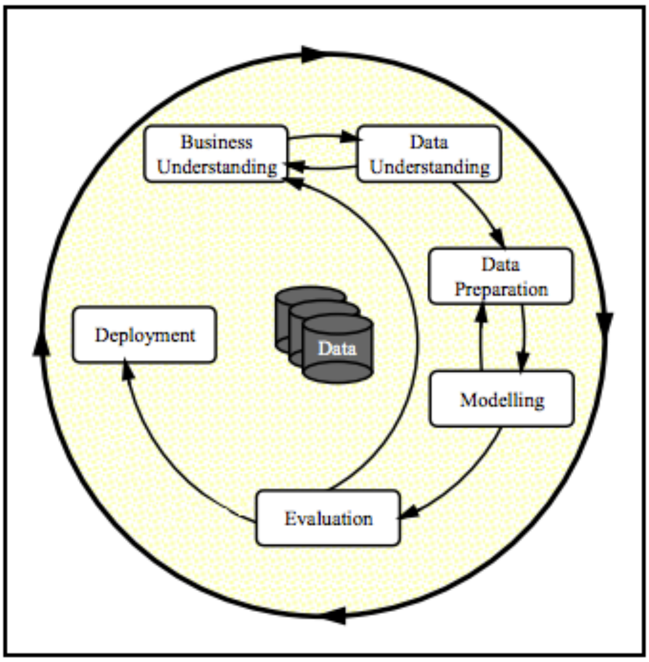
   Applied various methods to draw plot with data patterns.

6- Interpreting the mined data through mining process.

7- Knowledge Representation

   (as you will see in Results section)

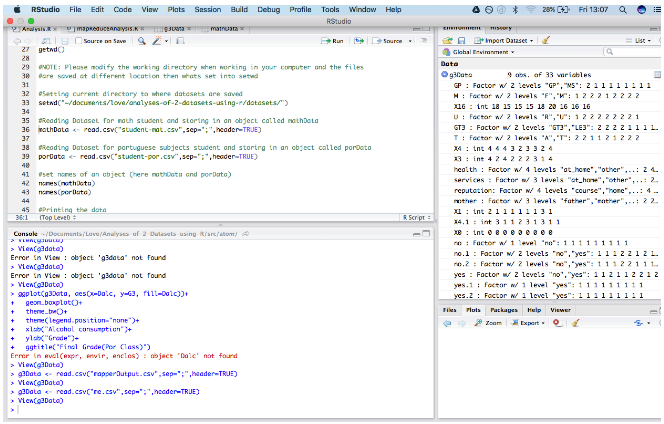
Know we will see how some CRISP-DM phases were used during analysis:



Like KDD, in CRISP-DM that the data selection is necessary and datasets were selected, data understanding was involved as can be seen in Data section of the paper, data preparation and modelling was involved to some stage and deployment (the final results of the analysis).

IV. Implementation and Artitecture

Once the datasets were selected, RStudio was used as the work environment for analysis and first step was to setup the current working directory and then both datasets were loaded and saved as R objects. The dataset were clean and didn't require much cleaning and then some testing was performed on the data to understand the data like the summary() function in R.



```r
getwd()

#NOTE: Please modify the working directory when working in your computer and the files
#are saved at different location then whats set into setwd

#Setting current directory to where datasets are saved
setwd("~/documents/love/analyses-of-2-datasets-using-r/datasets/")

#Reading Dataset for math student and storing in an object called mathData
mathData <- read.csv("student-mat.csv",sep=";",header=TRUE)

#Reading Dataset for portuguese subjects student and storing in an object called porData
porData <- read.csv("student-por.csv",sep=";",header=TRUE)

#set names of an object (here mathData and porData)
names(mathData)
names(porData)

#Printing the data
```

```
Console ~/Documents/Love/Analyses-of-2-Datasets-using-R/src/atom/
> View(g3Data)
> View(g3data)
Error in View : object 'g3data' not found
> View(g3data)
Error in View : object 'g3data' not found
> View(g3Data)
> ggplot(g3Data, aes(x=Dalc, y=G3, fill=Dalc))+
+   geom_boxplot()+
+   theme_bw()+
+   theme(legend.position="none")+
+   xlab("Alcohol consumption")+
+   ylab("Grade")+
+   ggtitle("Final Grade(Por Class)")
Error in eval(expr, envir, enclos) : object 'Dalc' not found
> View(g3Data)
> g3Data <- read.csv("mapperOutput.csv",sep=";",header=TRUE)
> View(g3Data)
> g3Data <- read.csv("me.csv",sep=";",header=TRUE)
> View(g3Data)
>
```

```
Data
g3Data          9 obs. of 33 variables
  GP : Factor w/ 2 levels "GP","MS": 2 1 1 1 1 1 1 1 1
  M : Factor w/ 2 levels "F","M": 1 2 2 2 1 2 2 2 2
  X16 : int 18 15 15 15 18 20 16 16 16
  U : Factor w/ 2 levels "R","U": 1 2 2 2 2 2 2 2 1
  GT3 : Factor w/ 2 levels "GT3","LE3": 2 2 2 2 1 1 1 1…
  T : Factor w/ 2 levels "A","T": 2 2 1 1 2 1 2 2 2
  X4 : int 4 4 4 3 2 3 3 2 4
  X3 : int 4 2 4 2 2 2 3 1 4
  health : Factor w/ 4 levels "at_home","other",..: 2 4…
  services : Factor w/ 3 levels "at_home","other",..: 2…
  reputation: Factor w/ 4 levels "course","home",..: 4 …
  mother : Factor w/ 3 levels "father","mother",..: 2 2…
  X1 : int 2 1 1 1 1 1 1 3 1
  X4.1 : int 3 1 1 2 3 1 3 1 1
  X0 : int 0 0 0 0 0 0 0 0 0
  no : Factor w/ 1 level "no": 1 1 1 1 1 1 1 1 1
  no.1 : Factor w/ 2 levels "no","yes": 1 1 1 2 2 1 2 1…
  no.2 : Factor w/ 2 levels "no","yes": 1 1 1 2 2 1 1 1…
  yes : Factor w/ 2 levels "no","yes": 1 1 2 1 1 2 2 1 2
  yes.1 : Factor w/ 1 level "yes": 1 1 1 1 1 1 1 1 1
  yes.2 : Factor w/ 1 level "yes": 1 1 1 1 1 1 1 1 1
```

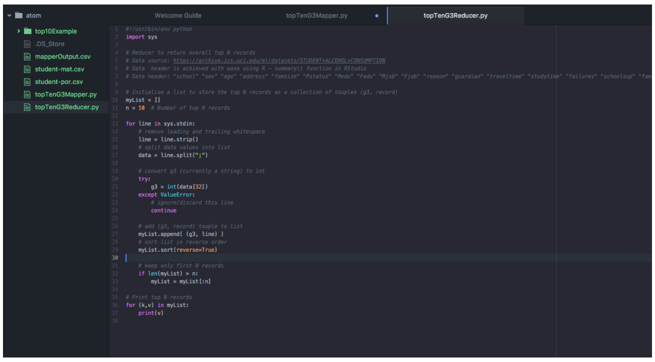
There are two types of analysis used during the analysis –

- Exploratory analysis - to find relations and connections in datasets and visualization.
- Descriptive analysis – used describe mostly the basic feature of the datasets and provide summary.

The analysis of the dataset is done in two ways –

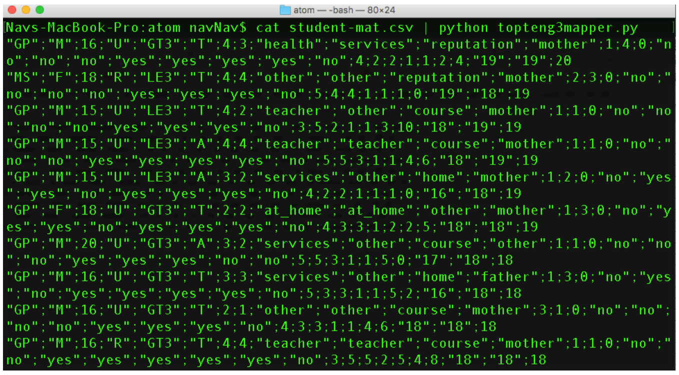
- Individually: Firstly, both the datasets are analyzed individually and results are produced (available in results section).
- Combined: Secondly, the data from both the datasets were combined and analysis was conducted and results were produced (available in results section).

Map Reduce: The mathematics dataset went through map reduce programming paradigm and top 10 student with best G3 (Final Grade were selected) and presented the whole map reduce process was done using Atom editor for .py files and mac terminal for passing dataset through mapper and reducer and outputting the results in a CSV file. And the found result was the student's getting highest grades were not consuming alcohol and hence had better results then other students.



```python
#!/usr/bin/env python
import sys

# Reducer to return overall top N records
# Data source: https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION
# Data  header is achieved with ease using R - summary() function in RStudio
# Data header: "school" "sex" "age" "address" "famsize" "Pstatus" "Medu" "Fedu" "Mjob" "Fjob" "reason" "guardian" "traveltime" "studytime" "failures" "schoolsup" "fam

# Initialise a list to store the top N records as a collection of touples (g3, record)
myList = []
n = 10  # Number of top N records

for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # split data values into list
    data = line.split(";")

    # convert g3 (currently a string) to int
    try:
        g3 = int(data[32])
    except ValueError:
        # ignore/discard this line
        continue

    # add (g3, record) touple to list
    myList.append( (g3, line) )
    # sort list in reverse order
    myList.sort(reverse=True)

    # keep only first N records
    if len(myList) > n:
        myList = myList[:n]

# Print top N records
for (k,v) in myList:
    print(v)
```

Screenshot of the environment for mapper and reducer.



```
Navs-MacBook-Pro:atom navNav$ cat student-mat.csv | python topteng3mapper.py
"GP";"M";16;"U";"GT3";"T";4;3;"health";"services";"reputation";"mother";1;4;0;"n
o";"no";"no";"yes";"yes";"yes";"yes";"no";4;2;2;1;1;2;4;"19";"19";20
"MS";"F";18;"R";"LE3";"T";4;4;"other";"other";"reputation";"mother";2;3;0;"no";"
no";"no";"no";"yes";"yes";"yes";"no";5;4;4;1;1;1;0;"19";"18";19
"GP";"M";15;"U";"LE3";"T";4;2;"teacher";"other";"course";"mother";1;1;0;"no";"no
";"no";"no";"yes";"yes";"yes";"no";3;5;2;1;1;3;10;"18";"19";19
"GP";"M";15;"U";"LE3";"A";4;4;"teacher";"teacher";"course";"mother";1;1;0;"no";"
no";"no";"yes";"yes";"yes";"yes";"no";5;5;3;1;1;4;6;"18";"19";19
"GP";"M";15;"U";"LE3";"A";3;2;"services";"other";"home";"mother";1;2;0;"no";"yes
";"yes";"no";"yes";"yes";"yes";"no";4;2;2;1;1;1;0;"16";"18";19
"GP";"F";18;"U";"GT3";"T";2;2;"at_home";"at_home";"other";"mother";1;3;0;"no";"y
es";"yes";"no";"yes";"yes";"yes";"no";4;3;3;1;2;2;5;"18";"18";19
"GP";"M";20;"U";"GT3";"A";3;2;"services";"other";"course";"other";1;1;0;"no";"no
";"no";"yes";"yes";"yes";"no";"no";5;5;3;1;1;5;0;"17";"18";18
"GP";"M";16;"U";"GT3";"T";3;3;"services";"other";"home";"father";1;3;0;"no";"yes
";"no";"yes";"yes";"yes";"yes";"no";5;3;3;1;1;5;2;"16";"18";18
"GP";"M";16;"U";"GT3";"T";2;1;"other";"other";"course";"mother";3;1;0;"no";"no";
"no";"no";"yes";"yes";"yes";"no";4;3;3;1;1;4;6;"18";"18";18
"GP";"M";16;"R";"GT3";"T";4;4;"teacher";"teacher";"course";"mother";1;1;0;"no";"
no";"yes";"yes";"yes";"yes";"yes";"no";3;5;5;2;5;4;8;"18";"18";18
```

Screenshot of passing dataset to mapper in mac terminal.

The list of analysis that is performed on the datasets –

For analysis of datasets individually -

- Finding the weekend alcohol consumption based on guardian of the student to see if living with a particular parent or other guardian has any impact on drinking habits.
- Consumption based on gender to see if boys or girls are drinking more and predict a reason why ? and a possible solution
- Consumption based on age to check and possibly control the level of drinking in students of age that's drinking the most.
- School Absences and Alcohol Consumption – to see if not going to school has any impact on consumption levels
- Alcohol Consumption and Grades – to see how alcohol consumption is impacting the studies and grades of the students.
- Weekly Study times and grades – compare grades of students with more study hours to the students with less study hours.

For analysis of data combined from both datasets –

- Find out weekend and workday alcohol consumption based on age of students
- Consumption based on age and gender
- Consumption based on family relations
- What % of students have access to internet and other activities.
- Alcohol consumption and desire for higher education

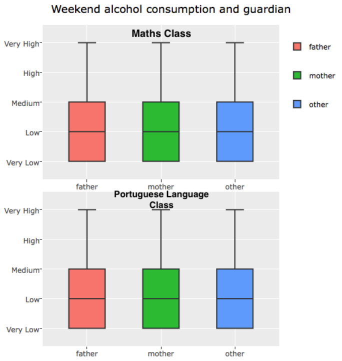
And finally, MapReduce programming paradigm to find records of students have Top 10, (Final Grade G3) and see if the students getting top results consumed what level of alcohol.

V. RESULTS

A. *Individual Analysis on Datasets (Results)*

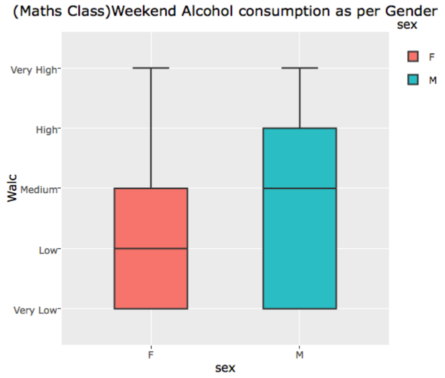
1) Weekend alcohol consumption based on if the student lives with their parents (mother or father) or other

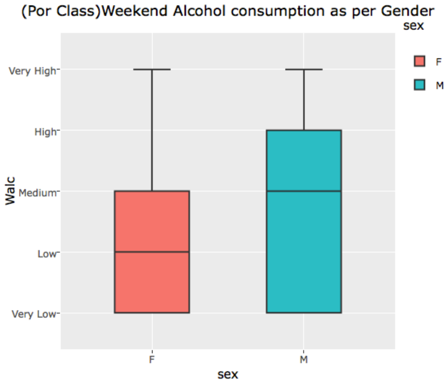
The results show that living with any of the guardian does not have huge impact on the alcohol consumption at least for these dataset (in the analysis its found very few students live with other person than there parents)



2) Weekend Alcohol Consumption based on Gender

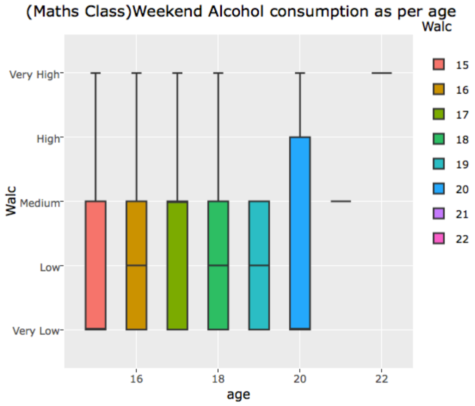
It can be easily seen from the two plots below that boys are drinking more than girls on weekends and girls highest is medium consumption level and on average low and for boys it high and on average it medium consumption level which shows boys are more prone hazards and are consuming a lot of alcohol.

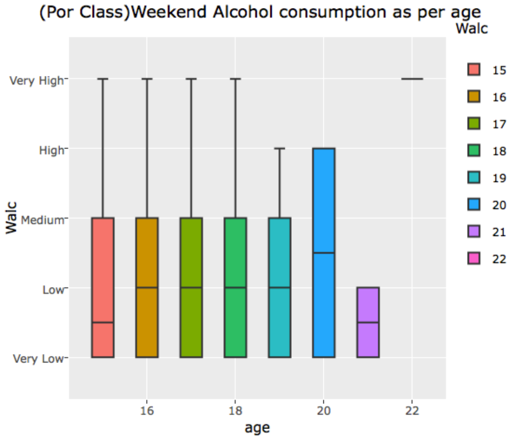




3) Weekend Alcohol based on age

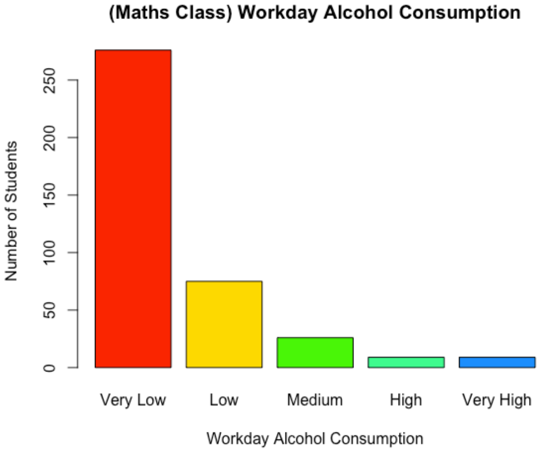
It can be seen in the results of both datasets that students of 20 years of age are consuming the highest level of alcohol compared to all students of other age.

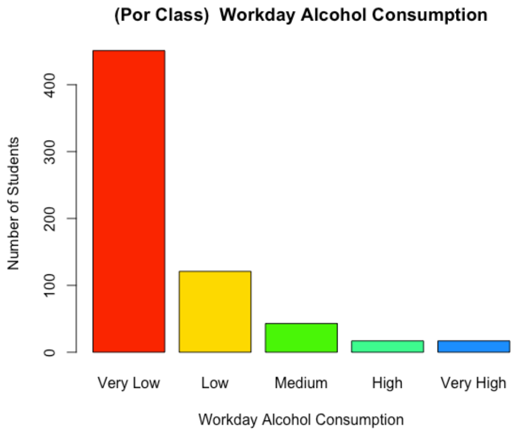




4) Workday Alcohol Consumption

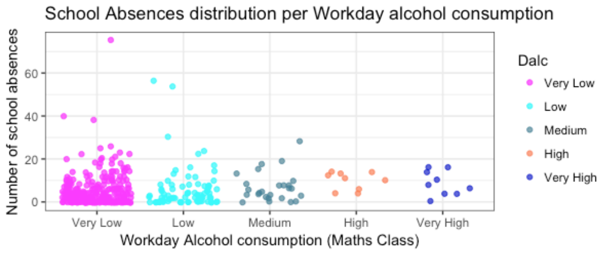
On workday the alcohol consumption of the majority is very low but still there are traces of medium and high consumption are found instead there should be no consumption at all as alcohol can ruin the life of these student.

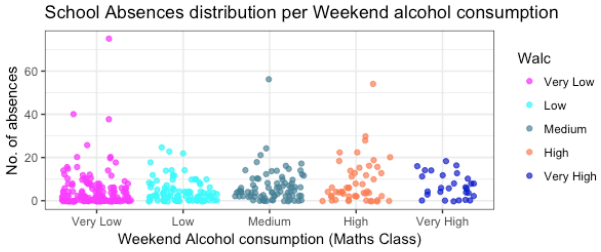


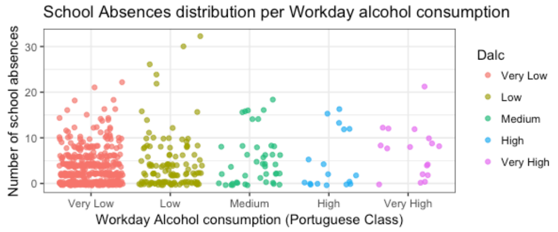


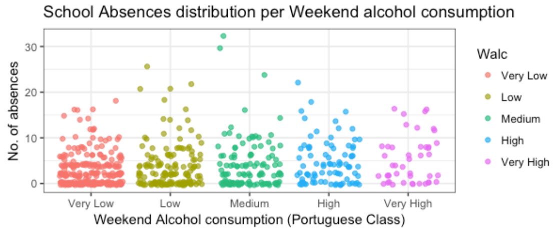
5) School Absences and Alcohol consumption

Getting absent in school can negatively impact student's studies.


School Absences distribution per Workday alcohol consumption

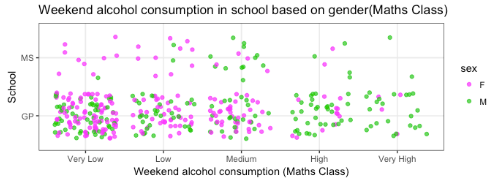

School Absences distribution per Weekend alcohol consumption

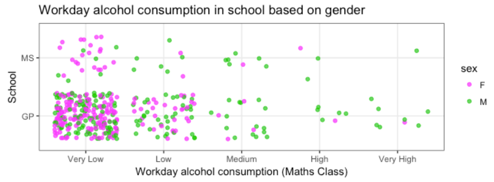

School Absences distribution per Workday alcohol consumption

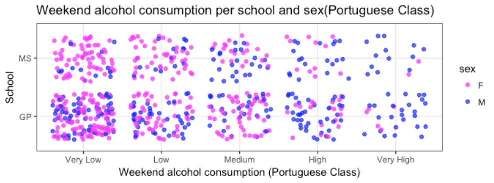

School Absences distribution per Weekend alcohol consumption

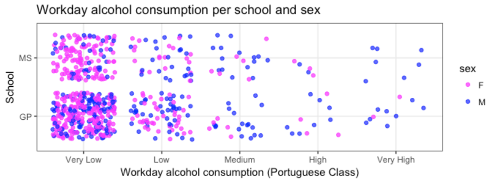
6) Weekend Alcohol Consumption based on gender (sex)

Majority of students are in very low consumption but in the plots it can be seen the very high level of alcohol consumption is done mainly by boys on weekends and there's also instance found many boys are consuming high level of alcohol on workdays which will have direct impact on their studies and the graphs below shows there is a good need for good direction for students of both schools.


Weekend alcohol consumption in school based on gender(Maths Class)

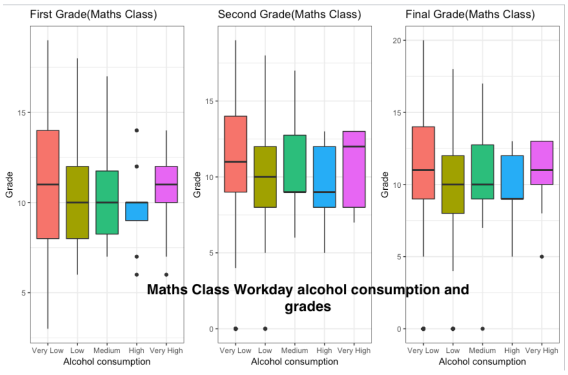

Workday alcohol consumption in school based on gender

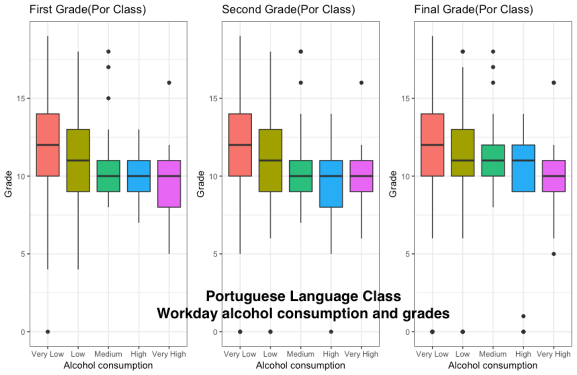

Weekend alcohol consumption per school and sex(Portuguese Class)


Workday alcohol consumption per school and sex

7) Workday Alcohol Consumption and Grades

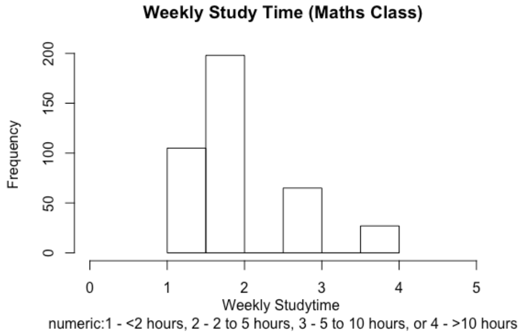
From the plot below it can be seen how alcohol has an impact on grades. The students consuming no alcohol / very low have better grades in all G1, G2, G3 and for both datasets

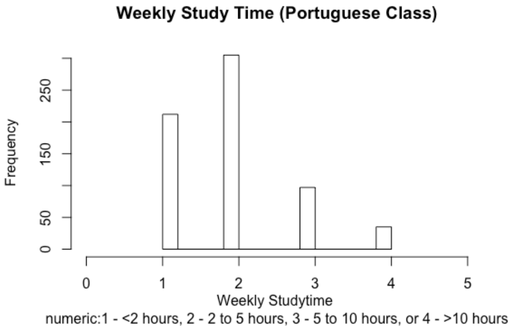

First Grade(Maths Class)
Second Grade(Maths Class)
Final Grade(Maths Class)
Maths Class Workday alcohol consumption and grades

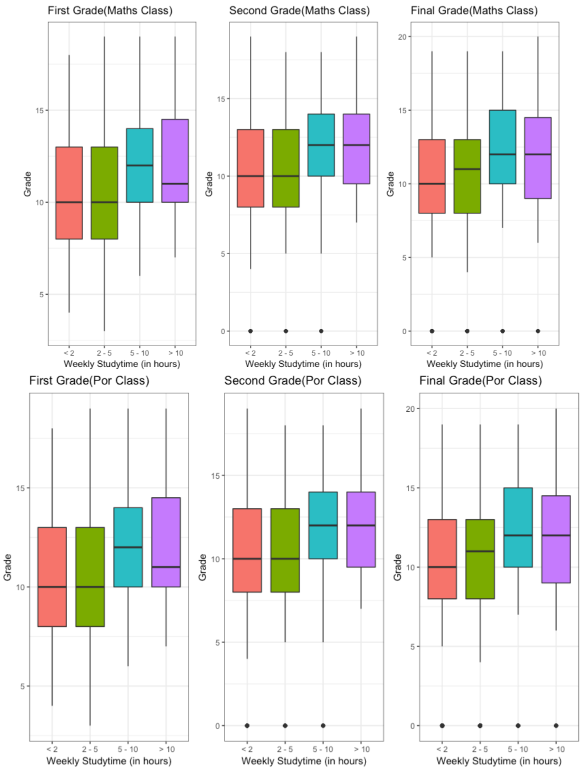

First Grade(Por Class)
Second Grade(Por Class)
Final Grade(Por Class)
Portuguese Language Class
Workday alcohol consumption and grades

8) *Weekly Study Time*

The histograms below show weekly study time of student in both dataset (on average its 2 -5 hours) more hours of study will result in better grades as we can see in plots below. So, the students should be suggested to study more and take part in activities and avoid alcohol



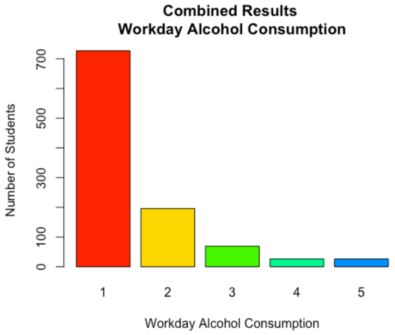




Better grades achieved by students when the weekly study hours are higher (for both datasets).

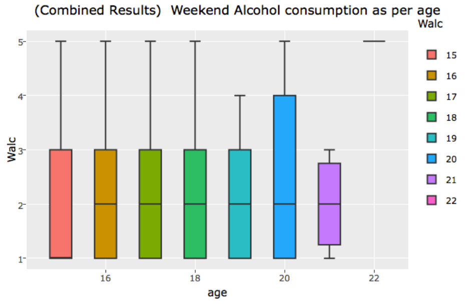
B. *Combined Data Analysis (Results)*

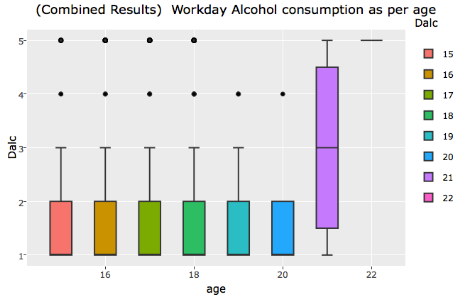
Workday Alcohol consumption histogram to show the alcohol consumption by the students of both courses.



1) *Weekend and Workday alcohol consumption as per age*

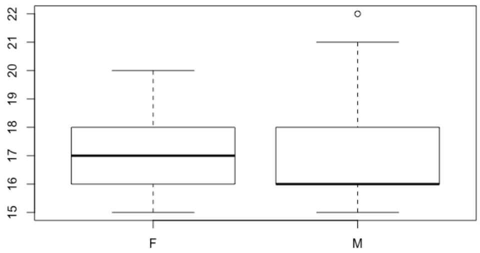
It can be seen on the weekends the 20 years old students are consuming high levels of alcohol and which is not a good sign as if they drink a lot on weekends their performance in school will definitely decrease for the coming week.





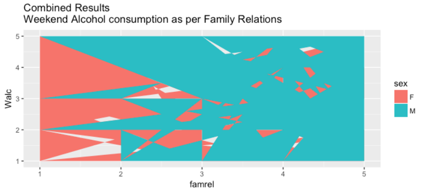
2) *Age and Gender (Sex)*

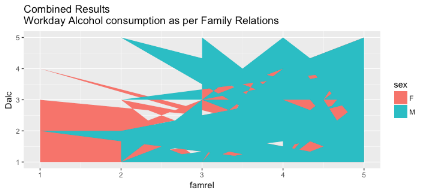
The plot below shows what is the age range of students in combined data and findings say on average the female students age is 17 years and for males its 16 years.



3) Alcohol Consumption distribution and family relations

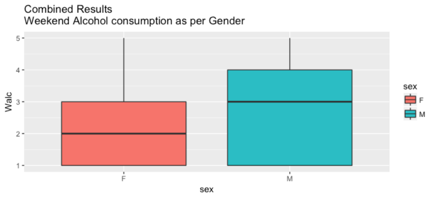
For the plot below we can see male students that have good family relations still do consume a lot of alcohol whereas for female students the results are much better than the other sex for both weekend and workday consumption distribution.

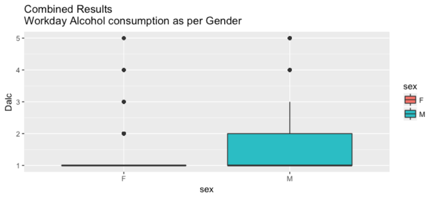

Combined Results
Weekend Alcohol consumption as per Family Relations


Combined Results
Workday Alcohol consumption as per Family Relations

4) Alcohol consumption based on gender

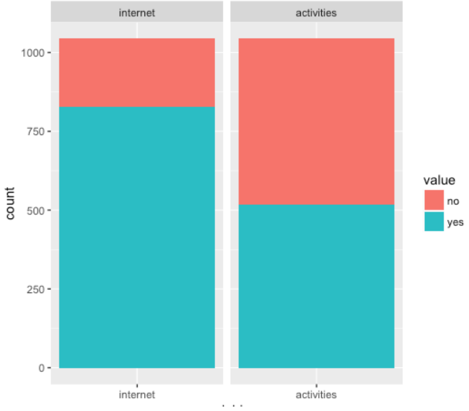
It can be seen for the plot below on weekends males are consuming the high level of alcohol and females on the other hand to medium range or less.


Combined Results
Weekend Alcohol consumption as per Gender


Combined Results
Workday Alcohol consumption as per Gender

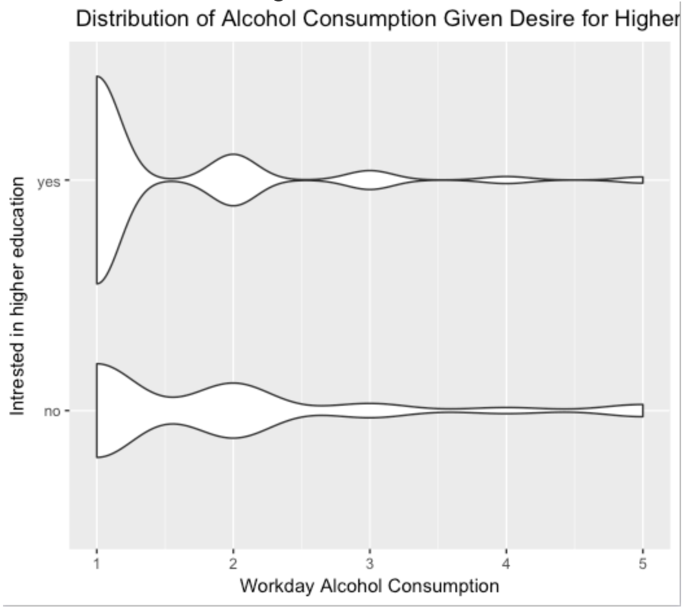
5) Plot to show to answer the question "How many students have access to internet and other activities

The results show that about 70-80% of the students have internet access which may be unsupervised and hence parental control on internet access should be strong so the students don't consume bad information from the internet and the level of activities is only about 50% which is somewhat less and students must be encouraged to take part in various activities.



6) Alcohol consumption based on desire for higher education

The plot shows the conditional distribution of workday alcohol consumption given the students desire for further higher education. There is a larger distribution of people in the very low alcohol consumption that want higher education than those who do not want higher education**.**


Distribution of Alcohol Consumption Given Desire for Higher

Reference: This plot has been done with the help of this site: https://www.kaggle.com/interkf/d/uciml/student-alcohol-consumption/alcohol-consumption-from-portuguese-school

Map Reduce results

The mapper and reducer was run on math's dataset and the findings were the students getting the top 10 final grades consumed very low (consumption = 1) and hence it showed alcohol do have bad impact on results and must be avoided.

VI. Conclusion and Future Work

The project was like a journey, during the project I used R programming language and RStudio for the first time and was a great learning experience and I speaking about the analysis results – it was found alcohol affected student grades and the more consumption of alcohol was majorly done by male students and it was shocking to know living with parents or others didn't had any impact on alcohol consumption (at least not in these dataset).Well, strong predictions cannot be made with this small sets of data and as I enjoyed the project and analysis phrase in future in free time I would love to do a similar analysis.

Acknowledgment

Similar work - Using data mining to predict secondary school student alcohol consumption by Fabio P. & M. Hossain at University of Camerino (Department of Computer Science)

VII. APPENDIX

1- Report of Analysis
2 - MapReduce Processing
These files are submitted along with the code deliverable on moodle.

REFERENCES

[1] F. Pagnotta and M. Amran Hossain. Using data mining to predict secondary school alcohol consumption

[2] 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- 9077381-39-7.

[3] "Alcohol And Mental Health". Drinkaware.co.uk. N.p., 2016. Web. 30 Dec. 2016.

[4] "Alcohol's Effects On The Body | National Institute On Alcohol Abuse And Alcoholism (NIAAA)". Niaaa.nih.gov. N.p., 2016. Web. 30 Dec. 2016.

[5] "CRISP-DM: Towards A Standard Process Model For Data Mining". citeseerx.ist.psu.edu. N.p., 2016. Web. 30 Dec. 2016.

[6] "Student Alcohol Consumption". Kaggle.com. N.p., 2016. Web. 18 Dec. 2016. "Student Alcohol Consumption | Kaggle". Kaggle.com. N.p., 2016. Web. 15 Dec. 2016.

[7] "The KDD Process For Extracting Useful Knowledge From Volumes Of Data". pbworks.com. N.p., 2016. Web. 30 Dec. 2016.

[8] "UCI Machine Learning Repository: STUDENT ALCOHOL CONSUMPTION Data Set". Archive.ics.uci.edu. N.p., 2016. Web. 5 Dec. 2016.

[9] “Data Mining: Knowledge Descovery” | Tutorialspoint.com N.p., 2016. Web. 28 Dec. 2016.