Student Grade Influencers

Data Visualization
Final Report
Fall 2017

Sreya Chakrabarti
Arpitha Kashyap

Abstract

Many practical studies are carried out to investigate factors affecting college students' performance. The focus of this project is to explore various factors that can influence student's quality of learning and grades. Our primary measure of academic achievement is the student's grade abstracted from the data set. We try to comprehend how these factors affect the students and their grades. We use several visualization techniques to explore these factors and determine the most influential factor among them all. In the process, we look into several techniques and plots that surprise us on the effects they have on the grades. Our results reveal that Alcohol is the highest influencer of student grades. Increase in alcohol consumption results in small yet statistically significant reductions in grades for both male and female students. But this is not the only important factor.

Contents

Introduction

Education is the foundation of a successful career and life. Especially the education we get in school, as it shapes a student's personality. Also school is a place where the student takes his/her first decisions with respect to their career. They make crucial decisions on what courses they should further study or realize what kind of courses or majors they would be interested in the long run. Therefore Grades in school and study habits developed in school are highly important. However, they face several distractions and do not give their best shot during school exams. Our project aims to find a few reasons that would affect a student's grade.

Our dataset is a survey of students from Math and Portuguese class from secondary school. There are a lot of attributes in the dataset which help in finding out what factors affect a student's grade positively or negatively. Using various visualization methods, we have tried to find the factors that are highly impactful in a student's academic performance.

We generally assume that most of the students are underperforming because he or she is hanging out too much or facing too many distractions. In this project we get the actual statistics of the causes. Each individual student is unique and hence it wouldn't be right to compare them or base our studies of individual students. We look at around 400 students and based on their performance in exams, we try to determine causes to the best of our ability.

Some of these factors have been explored already in several research papers. These papers also indicate that it is highly possible that more than one of these factors can be responsible for a student's failure or success. The papers also talk about the positive and negative effects of these factors. If there is a negative factor, there might also be a positive factor that neutralizes the negative one and vice versa. Is it not an easy task determining one factor that affects the grades alone. Since each student differs in several aspects from one another, there can be different factors affecting each one of them. We consider factors that affect the students as a whole or the general set of students.

Secondary and high school students give equal importance to friends and studies as pointed out in a research paper. Your peer group sometimes defines they type of person you are. We talk more about this is the jupyter notebook attached to this file. What we are trying to achieve is figuring out the right influencers for the right kind of students. We take this as a challenge as we are students and have been through the same phase before. The further sections talk about these influencers more in detail.

Background

There are several resources, blogs etc. that talk about how students are performing in their academics and what affects their academic performance the most. We have used a dataset from Kaggle for our analysis. Using Machine learning algorithms various estimations have been conducted on the same dataset. Some visualizations have been created based on certain categorical factors like romantic relationship, activities, internet etc. It's impact on academic grades has been explored through visualizations. Kaggle has a lot of relevant exploratory Data analysis and visualizations. There are several existing visualization techniques used like box plots, bar charts and histograms in R as well.

The data were obtained in a survey of student's Math and Portuguese language courses in secondary school. It contains a lot of interesting social, gender and study information about students. These courses were chosen because we want to consider a class that requires logical thinking and another one that requires memorization. These two classes are not very similar and would help us with in depth analysis of certain factors.

The data set consists of the following: There are 382 students in both courses.

Attributes for both Math course and Portuguese language course :

- 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 - 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)

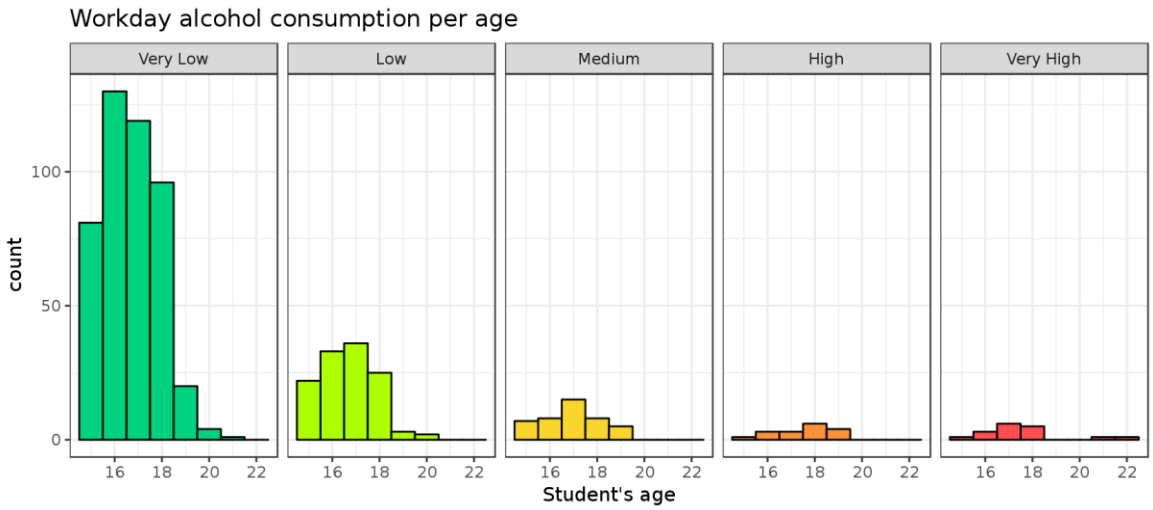
These grades are related with the course subject, Math or Portuguese:

- 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)

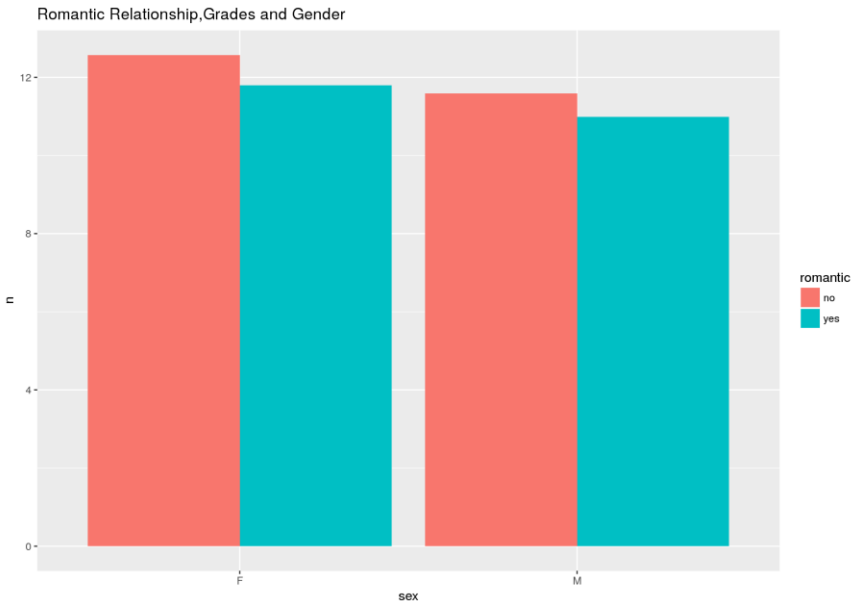
Since the data was obtained from a survey, the above points are the factors considered for this dataset. There are several factors both numerical and categorical. These factors were found to be closely related to how students perform in academics and some of them are the major causes. The dataset considers all the above factors because each student may be facing a different kind of problem that affects their mindset for studying. There have been several papers that talk about how internet has strongly affected the concentration of students. Especially these days, where the internet is everything. However, there are certain factors like alcohol, parent’s education, going out, hobbies and outdoor activities that also serve as factors that can be highly related to the grade.

Related Work

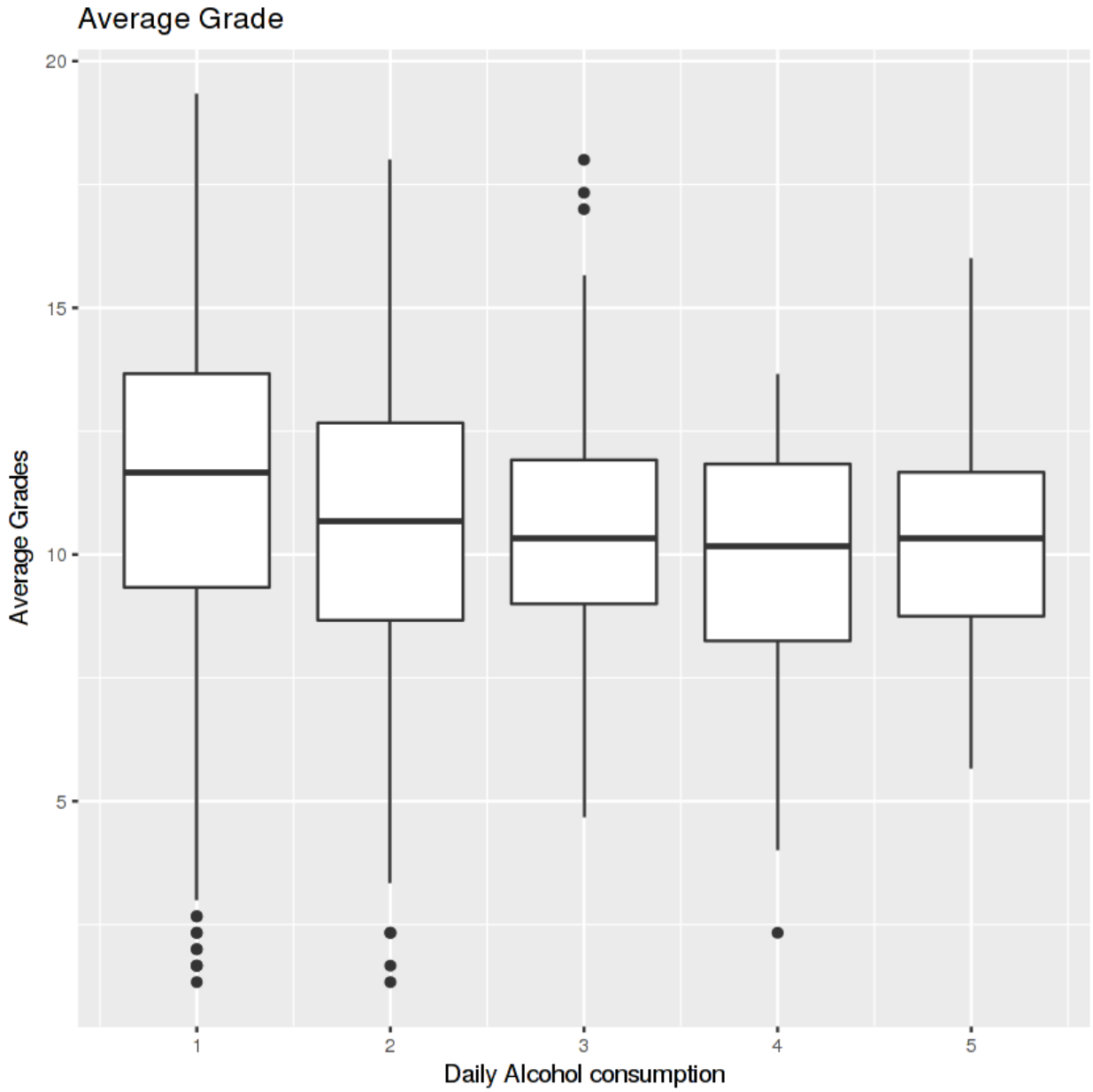
There are several visualizations that are already present for the given dataset. Of course, each of these visualizations are for only a few factors. There are some research papers showing visualization techniques and some discussion boards as well. Here, we will talk about several techniques we have come across.



The graph above is one of the simple visualizations we found in a discussion board. This plot shows the consumption of alcohol according to age of the student. The colors are not very visually pleasing, and this doesn't tell much about any factors and its effect on student's grades.

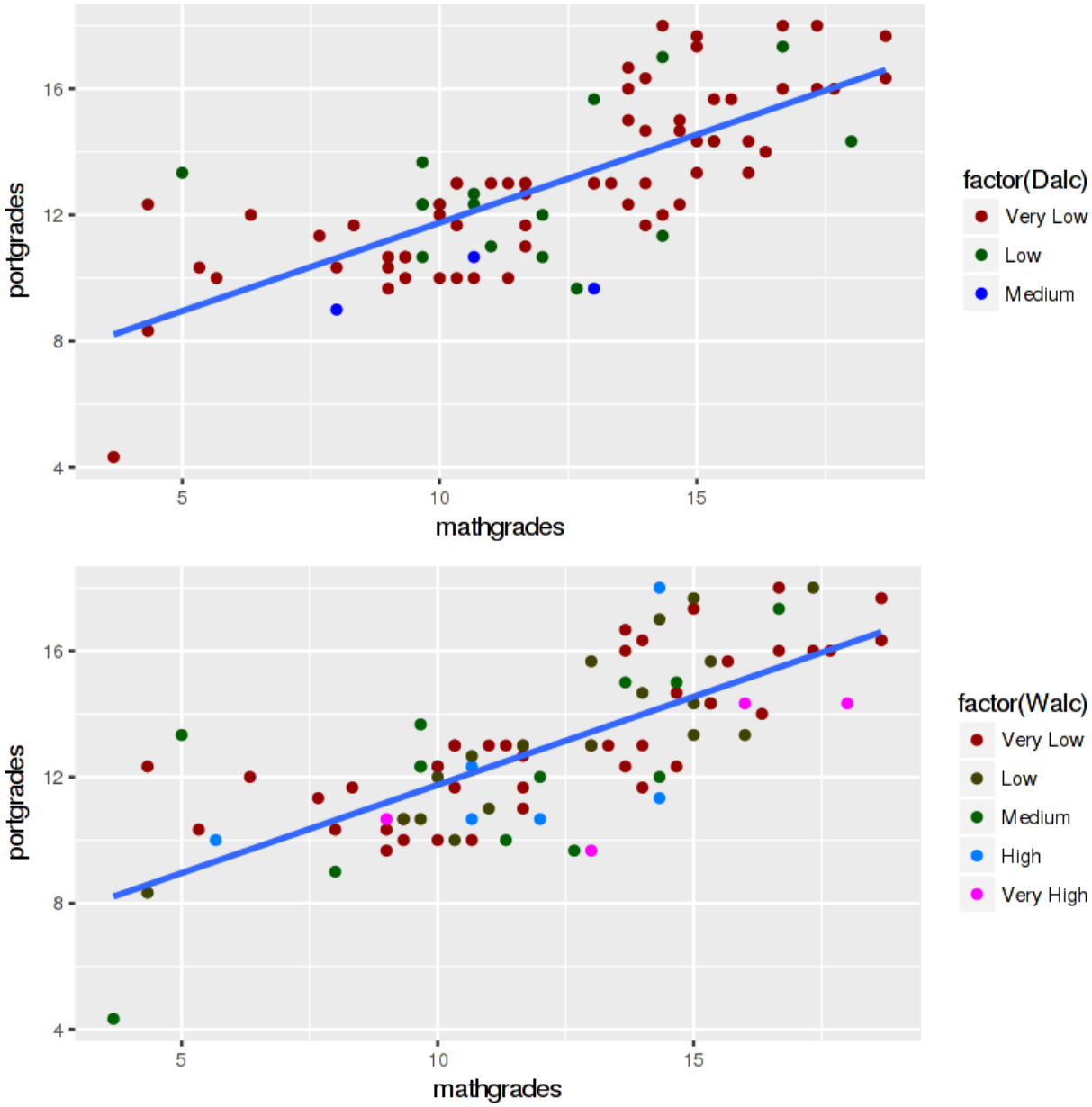


The plot above shows how romantic relationship affects the grades of a student based on gender. Romantic relationships are always a distraction especially in secondary or high school students. It has proven to be a major factor affecting concentration and hence grades several times.



The above boxplots represent the average grade with daily alcohol consumption. The grade seems to be highly affected by alcohol. It's hard to tell whether alcohol has a negative or positive affect. It also depends on the sex. While boxplots are a good way to visualize data, for this dataset, it doesn't give us a lot of information.

The plot below is a scatter plot of Daily and weekly alcohol consumption for Math and Portuguese class. This was another visualization used several times to describe the dataset.



Research Question

When it comes to understanding the causes affecting academic grade we mostly assume several reasons. For example, we might think that internet is the major cause for affecting students or having too much leisure time is or even getting poor quality education is a factor behind a student's academic performance. It is important to ask ourselves and figure out the factors which are impactful and to what extent.

In the project we expect to find some unusual factors that affect a student's grade. It is both interesting and curious to know the causes and effects of various factors that might be very simple for some of us. We are looking forward to find out the truth behind the factors and explore much more.

Process

Analysis of data:

We looked into several different techniques to visualize the given data. The analysis involved figuring out various techniques that represented how several factors affected the student's grades. We explored various plots by plotting them using python in jupyter notebook. We found that there were no null values in the dataset. It has both numerical and categorical values.

Visualization Methods:

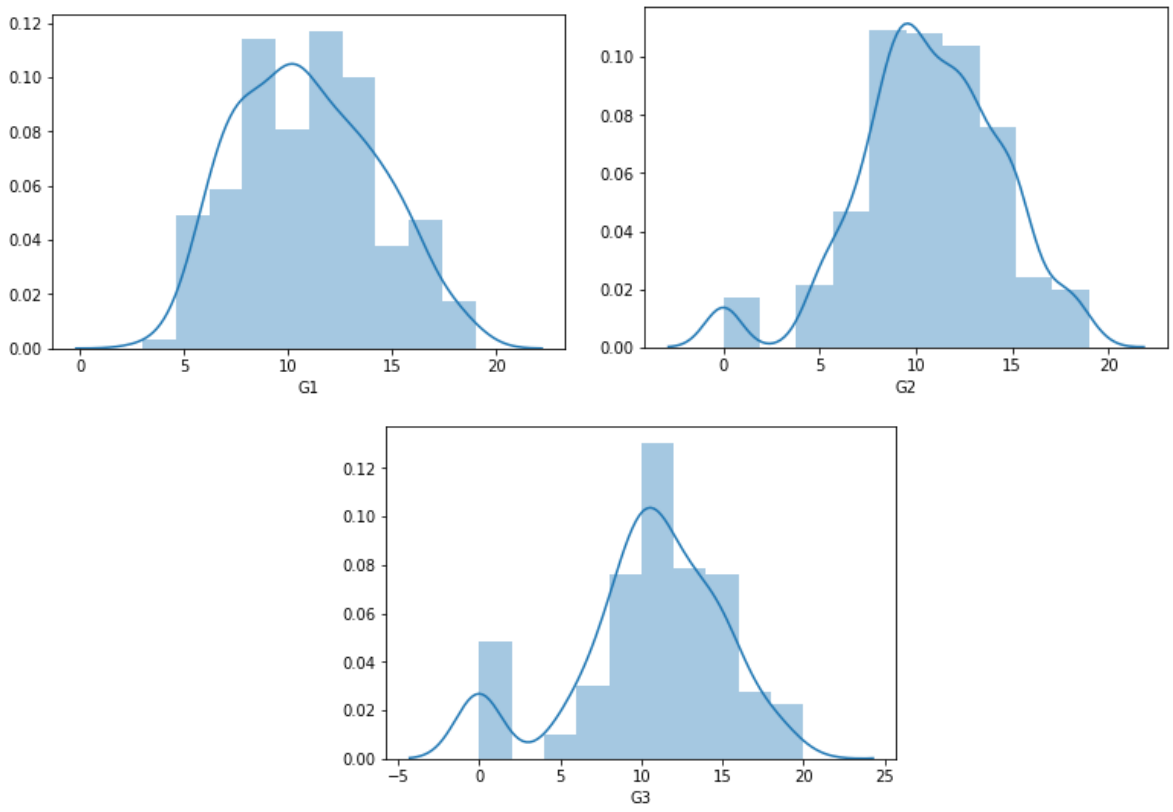
We used different types of plots like sector plots, correlation plot, bar charts with error bars, KDE plots and so on. We tried other techniques as well like histograms, boxplots and so on but those techniques already existed for the dataset.

We tried to use techniques and colors that looked visually pleasing. Most of our plots have good colors and represent the data quite well. We used a sector plot to represent how many hours students spend studying per week. We use a correlation plot to figure out what factors affect grades the most. We figure out the top five factors and draw more plots to find the one factor that affects the grade the most. We draw bar charts to explore how alcohol affects the grades. We also consider KDE plots and factor plots that show us how parent's education affects the grades.

The colors we used are also color-blind friendly. We have a story to tell through our visualizations. We also explored some existing visualizations that help us understand the data better.

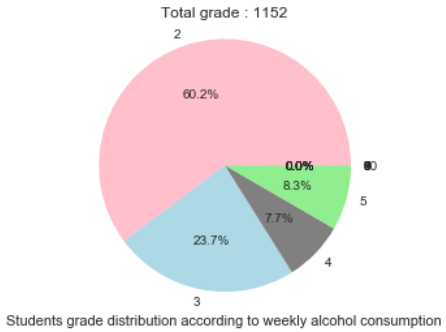
Failed Experiments:

Some of the failed experiments were drawing Histogram for the dataset. We created histograms of the grades to get some insight of grade distribution for different periods. However, it was not very useful.



The plots above show that all the three periods have a high score between 8 and 12. And a low of zero. Even on looking at the distributions, which varies a little amongst the three plots we can see that the change in distribution is not major.

We created a pie chart to see the distribution of alcohol consumption with grades. However, it was not very effective in giving much insights.

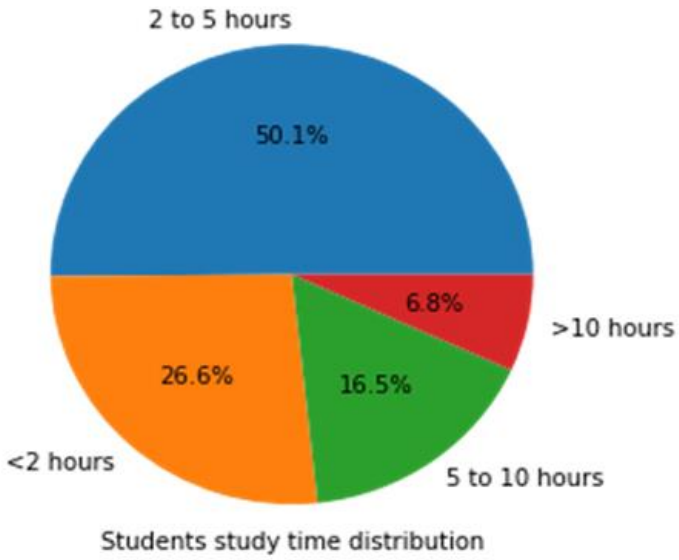


The information obtained from the pie chart was not very relevant to or the project goals. Yes, it talks about alcohol affecting grades, but nothing beyond what we already know. It also displays the total grade, which is not necessary. These were some of the experiments that failed during our trial and errors efforts to figure out important information and insights.

Results and Insights

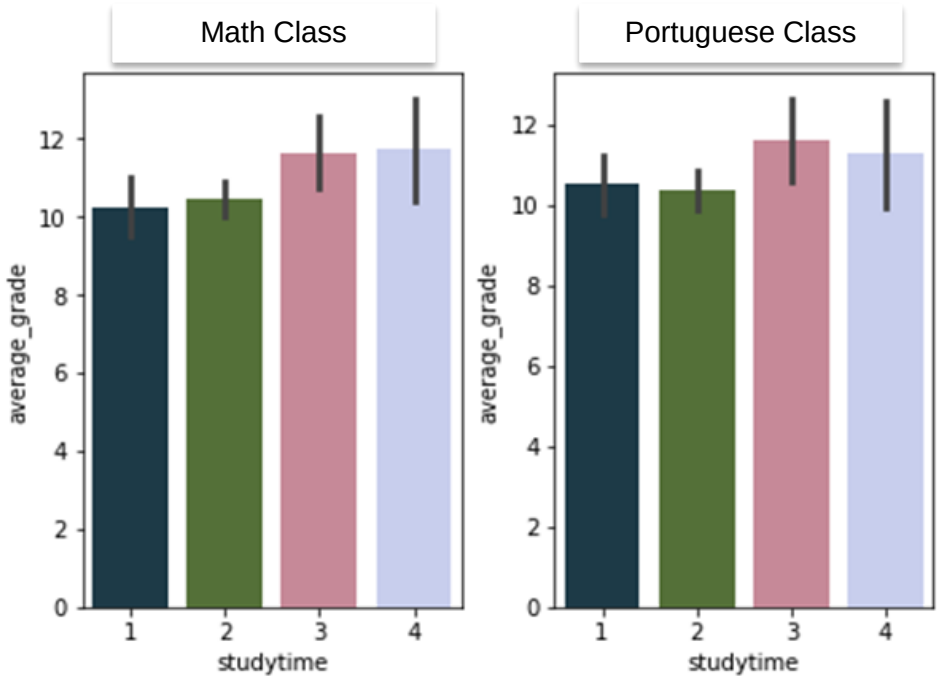
We first explore various factors and introduce the dataset. We start of with simple visualization techniques. The dataset has students in both Math and Portuguese class.

From this sector graph, we can see that majority of the students contribute a decent amount of time to study. We can see that 50% of the students spend about 2-5 hours in studying on a regular basis. More than half of the class spends very little time to study on a weekly basis. This alone looks like the most responsible factor that affects the grades. However, there are other factors too that influence their academic performance. We shall see more of the factors in this project.



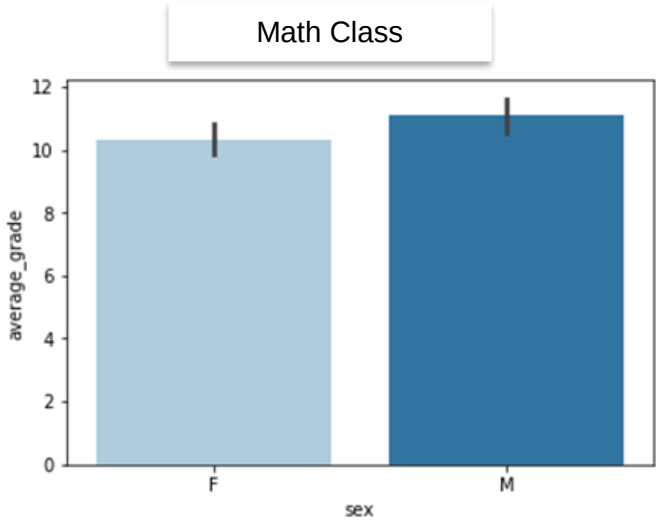
Students study time distribution

Now, considering both classes separately, we look at how study time affects the average grade. With experience, we must say that it does affect quite a lot. We also used error bars to represent how far from the reported value, the true value is. From the above graph, we can see that a moderate study time of 5-10 hours per week gives good grades relatively.



We consider another factor, 'sex'. Does it matter? Yes, it does! The graphs below show how the average grade is different for Males and Females. In both the graphs we can see that male has a higher average grade. This can be due to two factors one, the count of males in a class is higher than females. Second based on our research there are few factors like concentration, dedication, focusing on one thing at a time, confidence etc that are more dominant in males than females.

Also, some research shows that men concentrate on one thing at the time unlike women who can do many things at a time. This might be advantageous to men and hence helps them concentrate better for courses.

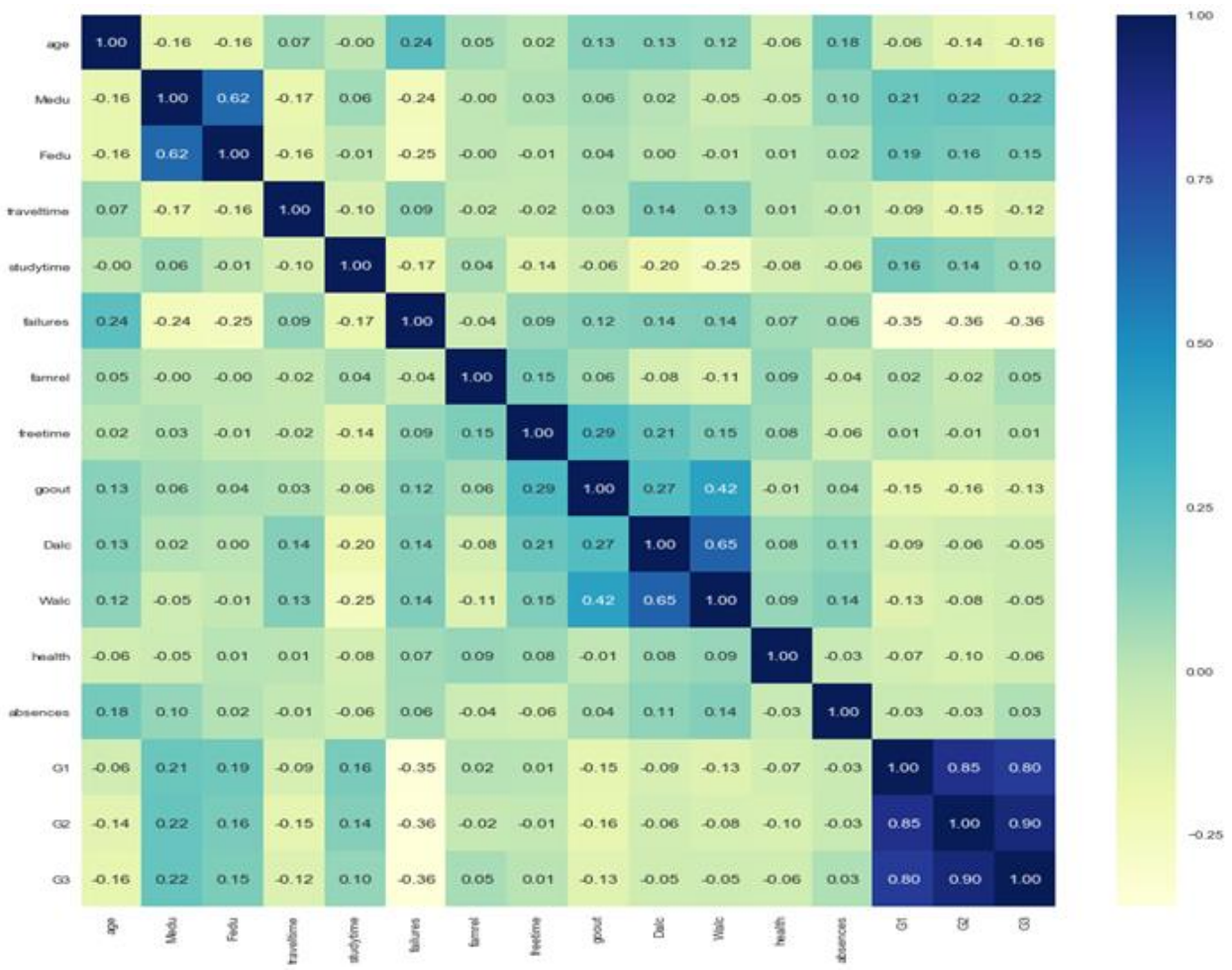


Based on the correlation plot below, we can see that grades G1, G2 and G3 are highly correlated to each other. Therefore, we are using G3 to represent all grades in some of the visualizations. There might be several other factors individually or combined which can influence the grades. However, we are looking at the ones which have a high correlation values because these seem to be the ones directly affecting grades in the dataset.

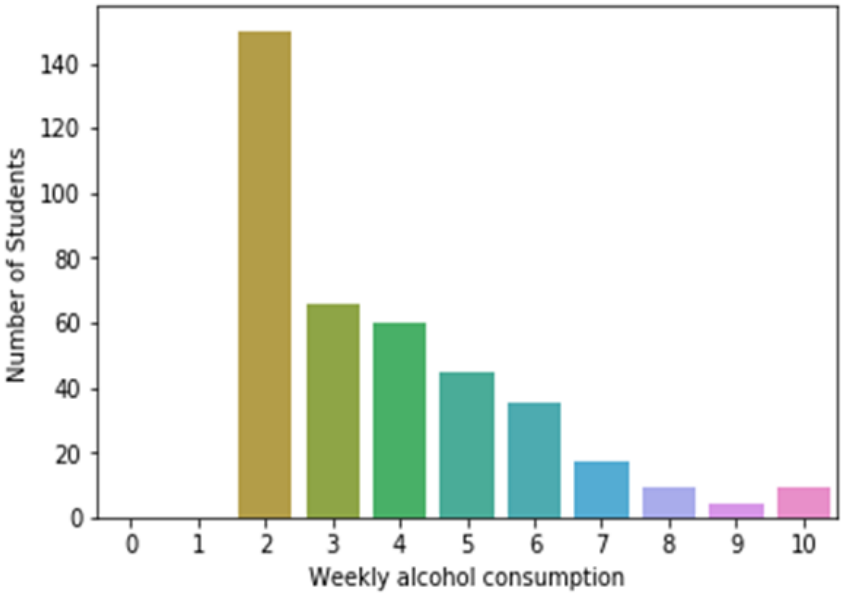
From the plot below, we can also see that the following are some of the factors that seem to affect the grades the most:

- Alcohol Consumption on weekdays
- Alcohol Consumption on weekends
- Mother's Education
- Father's Education
- Going out and
- Free Time

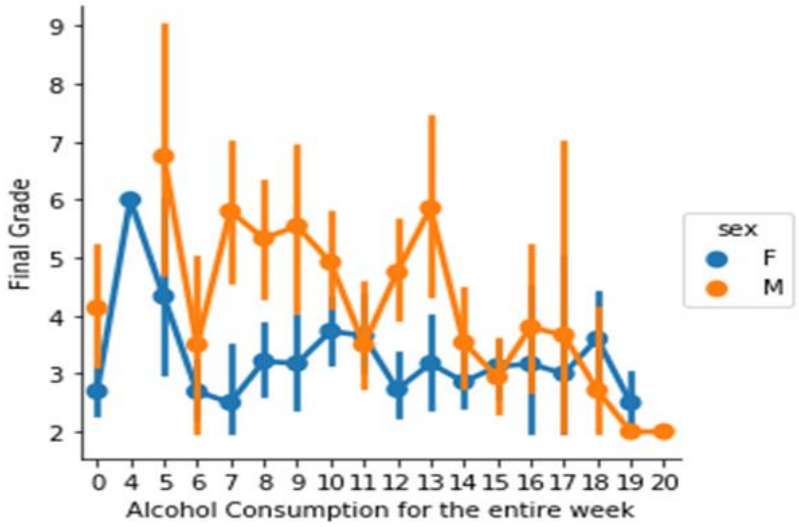
Since we observe that these factors are the ones that make a difference, we explore more of them. Alcohol seems to have the highest effect based on the correlation plot. Let's look into that first.



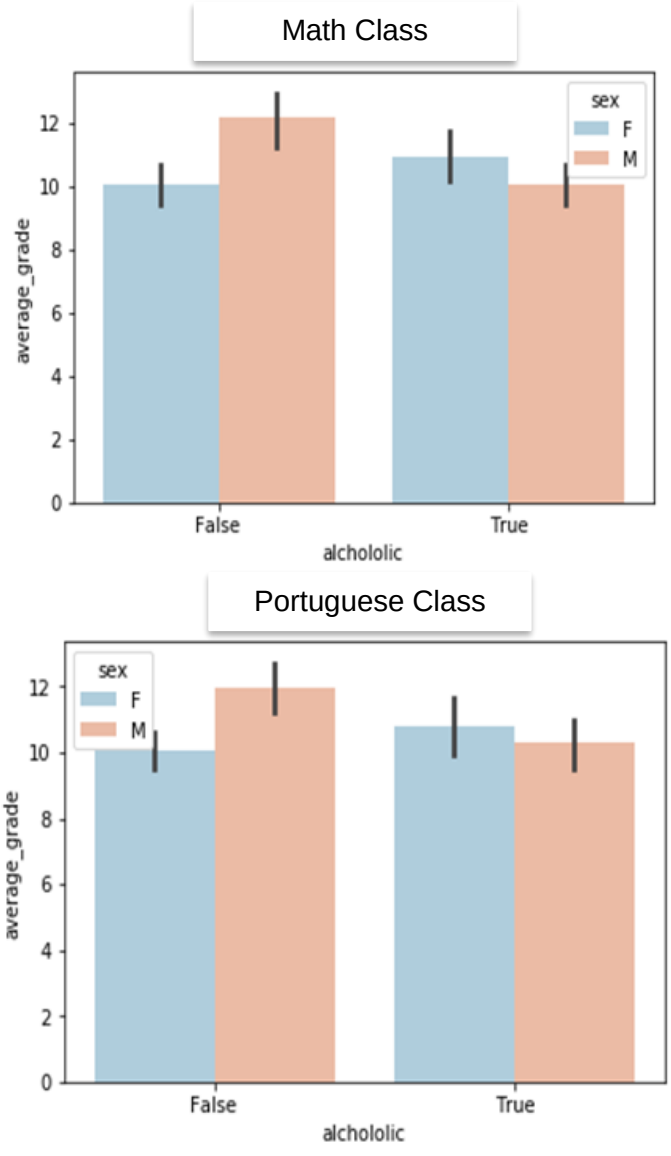
From the data set it is clear that there are no students that do not consume alcohol. In the graph below the x axis represents the number of times students consume alcohol per week. The minimum amount of alcohol consumed is twice a week. We can also see that max number of students consume alcohol twice a week. So around 150 students consume alcohol at least 2 times a week. Around 65 students consume alcohol at least 3 times a week and so on. In the plot the number of students consuming alcohol is decreasing with the number of times they have it in a week.



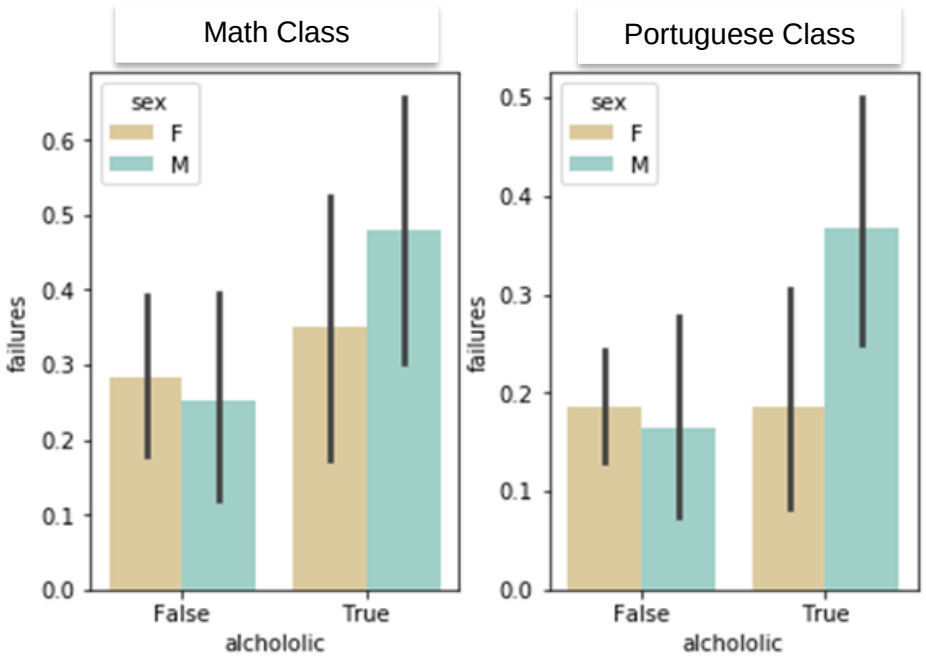
Now we can see how Alcohol consumption affects the Final grade. The plot also shows a range for which each of the values can vary. We can see that the influence of alcohol on males and females are different. In case of females, a small amount of alcohol consumption per week is helping them with their grades. However, an increase in consumption is causing a descent in their grades. However, for males, a high level of consumption leads to good grades but excess drinking drops their grades. Based on this we can say that, a decent amount of liquor consumption gives certain amount of relaxation in a stressful week, however if the quantity increases, it affects them negatively.



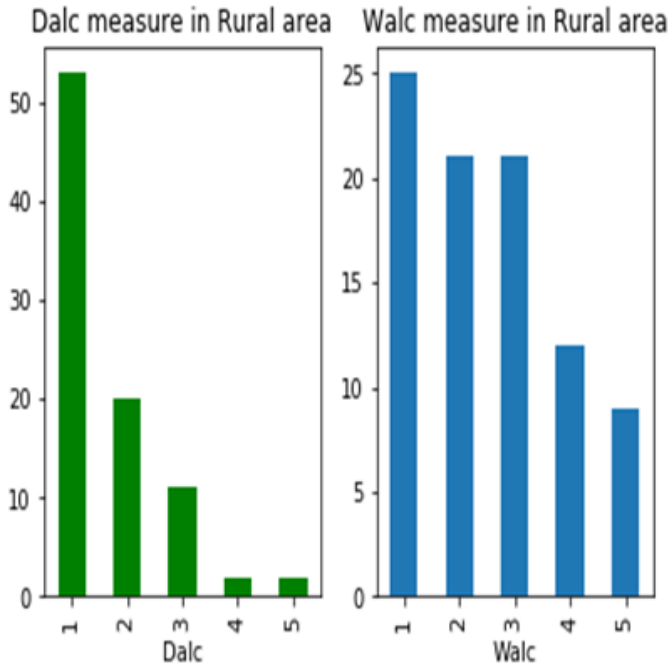
We have two graphs, one for Math class and the other for Portuguese class. The graphs below show that when females consume some amount of alcohol, their grade increases. However, for men, it's the opposite. This might be because it helps the females to relax and hence perform better. There might be several other reasons as to how the brain reacts to alcohol consumption for females and so on.

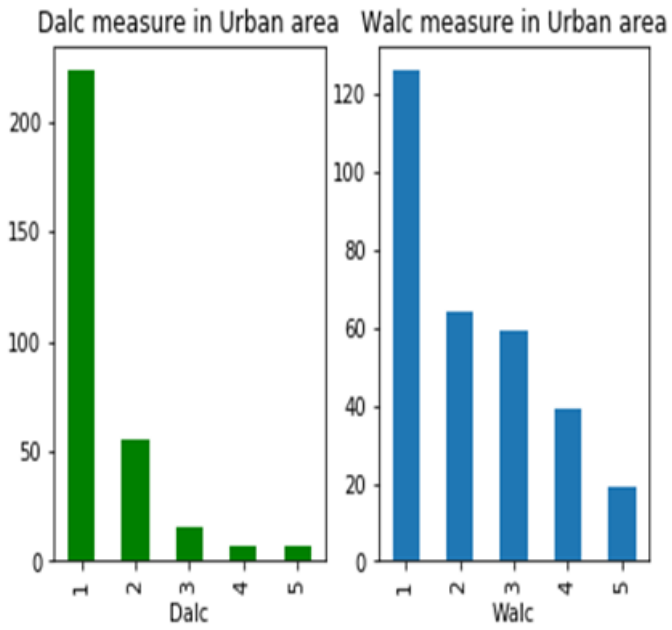


The plots below represent the failures due to alcohol consumption for both Math and Portuguese classes. The failures are especially high for males as consumption of alcohol seems to affect them more than females. From the above graph, considering classes that involve logic and memory, alcohol seem to affect males more than females. They have a higher negative impact for males when compared to females.



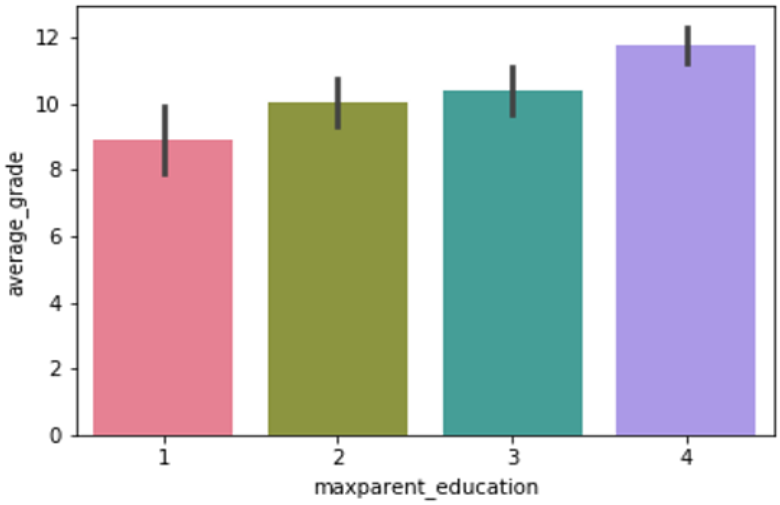
The consumption of alcohol is high in Urban areas whether it is a weekday or the weekend. Research also shows how student's grades are lesser in Urban areas when compared to rural areas. In urban areas, there are many distractions like the internet, malls, restaurants, and fancy places to visit. However, a simple life(financially and socially) in the rural area helps students stay more focused and consume less liquor.





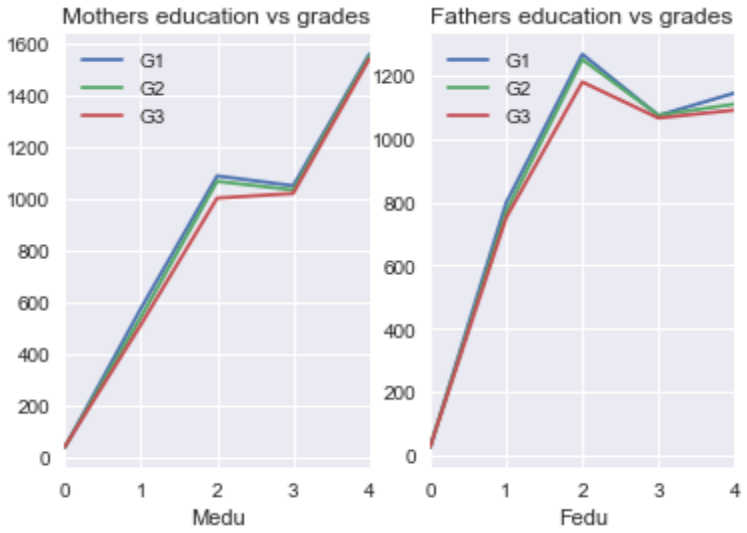
Parents qualification is also a very important factor influencing a child's academic performance. If the parents are educated they will encourage the child to study better, get good grades and make a good career.

In the bar chart below we can see the distribution of a student's grade with the parents education level. With 0 indicating no education,1 indicating studied till 4th grade, 2 indicating studied between 5th to 9th grade,3 indicating finishing secondary education and 4 indicating completing higher education.



From the plot we can clearly see the higher the parent's education the higher the grades. So, we can say that parents education is a positive factor influencing students grades.

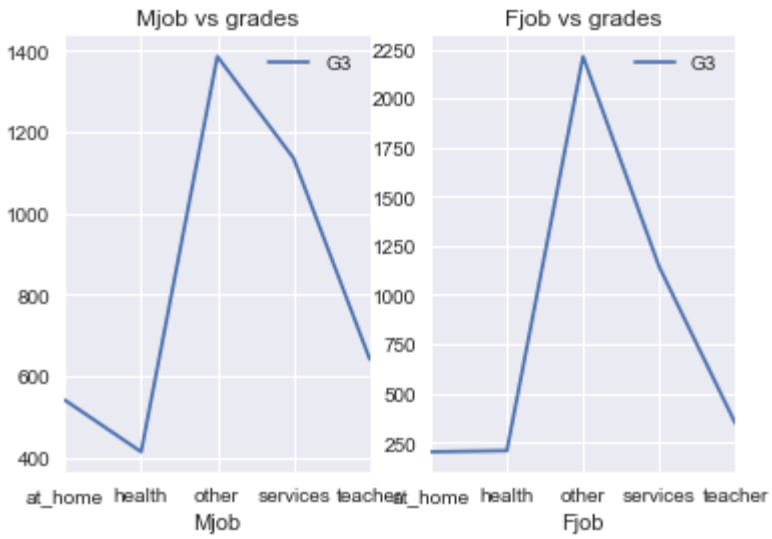
Now, lets see how they individually affect the child's grades.



In the plot Medu is for Mothers education and Fedu is for Fathers education.
We can see that Mother's education is linearly related to the child's grade. The higher the Mother is educated the better the child performs at school. On the other hand, we can see that if a Father is averagely educated the child does well in academics. Thus, a mother's education impacts the child more.

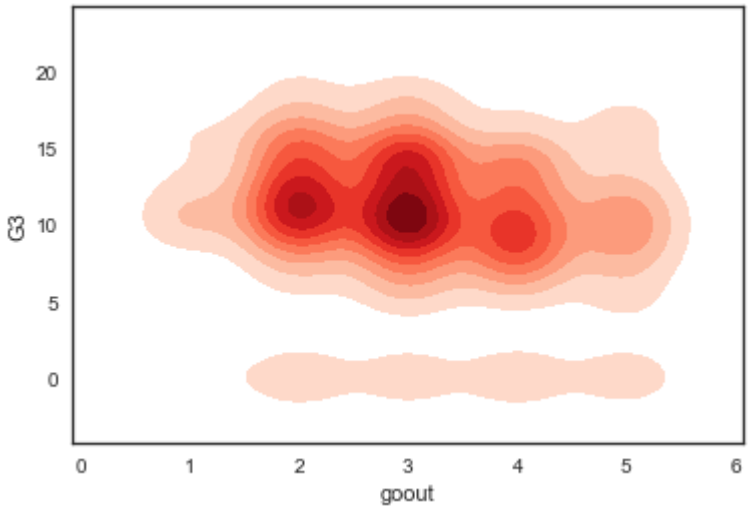
Parents job is also an essential factor in a child's life. If both the parents are very busy at their job the child does not get proper attention while growing up and that makes the kid indulge in bad activities or suffer from ignorance. On the other hand, if the parents are not working there would be poverty in the house and that would affect the child's upbringing negatively.

So, let's see from the plot below how much a parent's education affects the child's grades.



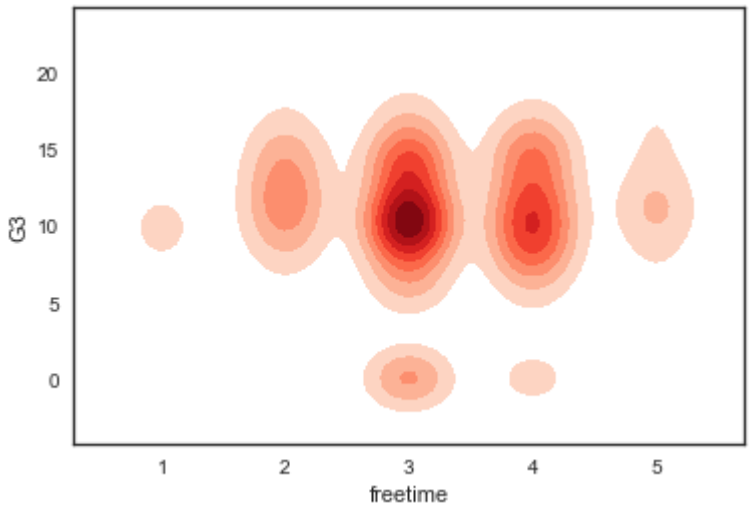
We can see that a Fathers job is more impactful in the child's academic performance. If a mother is at home the child still gets a better grade than if the father is at home. So, indicating that a Father's job affects a students' grade way more than a mother's' job.

It is assumed that hanging out might affect a student's grade negatively. From the plot below, we can get a statistical conclusion to this.



From the KDE plot above, we can see that Going out does affect the grades but not necessary it is negatively. A moderate amount of going out gives a decent grade in student's performance. Probably because a student spends some time playing and chilling out with their friends before he can come home refreshed to study. However, too much of hanging out leads to a decline in grades.

Let’s look at freetime’s impact on students grades from the plot below:



Here also we can see a moderate amount of freetime is good for getting a decent grade. A student needs to have time to pursue their hobby or relax. Probably being able to do that helps them concentrate better when they sit to study.

Conclusion

Among the various factors that we visualized, alcohol seems to be the dominant one. Alcohol consumption is a lifestyle which majority of the people adapt to. So, we conclude that Alcohol is the major factor influencing student's grades. There is a mixed review on the positive and negative effects of alcohol on student's grades.

Future Work

- We would like to research on how alcohol's effect varies on male and female. Some of the visualization we did so far indicated that there can be more to this.
- Analyzing parent’s influence on the kid’s grades would also be a good point to continue working on.
- Exploring different categorical variables especially the internet and whether it has a good or bad effect on student’s grades.
- Also exploring different models and try to predict final grade based on the factors.
- Certain important factors like parent’s income is not used in the dataset. Hence, we would like to explore outside the dataset as well and look into writing a paper about our findings.

References

[1]https://pdfs.semanticscholar.org/3c90/55c02a3fd87de21d1d20536335a8364882fa.pdf

[2]http://www.csus.edu/faculty/m/fred.molitor/docs/student%20performance.pdf

[3]https://www.kaggle.com/mukultaneja/analysis-student-alcohol-consumption

[4https://www.kaggle.com/calcifer/alcohol-consumption-and-average-grades

[5]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3843305/

[6]http://www.campusanswers.com/why-college-students-drink/