Data Analysis Project

Student Alcohol Consumption

Piotr Wojciechowski – X00152561

Dataset: https://www.kaggle.com/uciml/student-alcohol-consumption

Dataset Categories used:

- Sex – binary – Student's sex
- Age – numerical – Student's age
- Goout - ordinal – Going out with friends
- Dalc - ordinal – Workday alcohol consumption
- Walc - ordinal – Weekend alcohol consumption
- Health - ordinal – Current health status
- Absences - numerical – Number of school absences
- G3 - numerical – Final grade

Research Questions:

1. Does alcohol consumption affect your grades in secondary school?
2. Which gender scores a higher grade?
3. How much alcohol is consumed by students of different ages?
4. Is health affecting their grades and causing absences from school?
5. Is consuming alcohol correlated with going out with friends?
6. Are students who drink less inclined to score more in their final grade?

Table of Contents

Introduction 3
Importing Libraries and Dataset 3
Brief look at the Dataset 4
Exploratory Data Analysis 5
Which gender scores a higher grade? 7
Is health affecting student grades and causing absences from school? 8
How much alcohol is consumed by students of different ages? 8
Is consuming alcohol correlated with going out with friends? 11
Does alcohol consumption affect your grades in secondary school? 14
Are students who drink less/more inclined to score more in their final grade? 16
Quick Correlation Check 18
Conclusion 19
References 20

Introduction

During our time in secondary school we often get stressed out due to work and exams that the teachers give us which everyone deals with in a different way, some might spend time playing games, going for walks, reading books and then there will be those who will result to drinking with friends or on their own just to escape the continuous grind of secondary school. The dataset I will be using contains many attributes of students who are enrolled in Mathematics and Portuguese classes in the school of Gabriel Pereira and Mousinho da Silveira. These attributes vary from family background, academic performance, personal preferences, health etc. For my analysis I will be analysing the area of alcohol consumption among students and see its effects on their grades in their mathematics class.

My hypothesis for this project is that higher alcohol consumption does not has no effect on student's final grades. I believe that students who do consume much more alcohol on weekdays and weekends are still capable of handling their studies and achieving an above average grade.

Importing Libraries and Dataset

Let us start by loading the dataset and the libraries we are going to use for our exploratory analysis into Jupyter Notebook. The Python libraries which we will use are:

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Scipy
- Statistics

Now we are going to specify which columns we will use since dropping the columns will result in a long line of code and in this project we are not interested in the different fields of study that the dataset gives us like parental relations, parental education, student ambition, personal activities etc.

The columns we will use from our csv file are going to be:

'sex', 'age', 'goout', 'Dalc', 'Walc', 'health', 'absences', 'G3'.

Once we do that let us print out the head of the data frame that we will be looking at.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
import statistics as st

columns = ['sex', 'age', 'goout', 'Dalc', 'Walc', 'health', 'absences', 'G3']
try:
    student = pd.read_csv('student-mat.csv', usecols = columns)
except pd.io.common.CParserError:
    print("Your data contained rows that could not be parsed.")
student.head()
```

	sex	age	goout	Dalc	Walc	health	absences	G3
0	F	18	4	1	1	3	6	6
1	F	17	3	1	1	3	4	6
2	F	15	2	2	3	3	10	10
3	F	15	2	1	1	5	2	15
4	F	16	2	1	2	5	4	10

Brief look at the Dataset

Before we continue let us look at the dataset and make sure it is clean and ready before we do any kind of analysis. This will also help us understand the nature of each column and make sure if we need to change anything.

```
student.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 395 entries, 0 to 394
Data columns (total 8 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   sex       395 non-null    object
 1   age       395 non-null    int64 
 2   goout     395 non-null    int64 
 3   Dalc      395 non-null    int64 
 4   Walc      395 non-null    int64 
 5   health    395 non-null    int64 
 6   absences  395 non-null    int64 
 7   G3        395 non-null    int64 
dtypes: int64(7), object(1)
memory usage: 24.8+ KB
```

Now let us call the describe method that pandas give us.

```
student.describe()
```

	age	goout	Dalc	Walc	health	absences	G3
count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000
mean	16.696203	3.108861	1.481013	2.291139	3.554430	5.708861	10.415190
std	1.276043	1.113278	0.890741	1.287897	1.390303	8.003096	4.581443
min	15.000000	1.000000	1.000000	1.000000	1.000000	0.000000	0.000000
25%	16.000000	2.000000	1.000000	1.000000	3.000000	0.000000	8.000000
50%	17.000000	3.000000	1.000000	2.000000	4.000000	4.000000	11.000000
75%	18.000000	4.000000	2.000000	3.000000	5.000000	8.000000	14.000000
max	22.000000	5.000000	5.000000	5.000000	5.000000	75.000000	20.000000

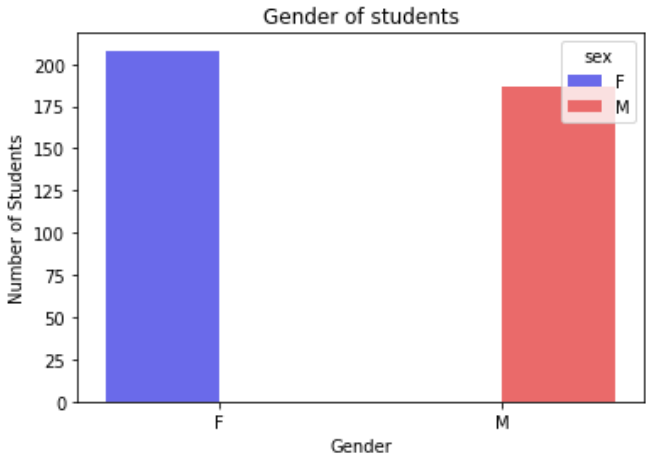
Looking at the descriptive statistics of our student's data frame, we can tell that

- Average age of students is 16
- Average alcohol consumption on the weekday is 1
- Average alcohol consumption on the weekend is 2
- Average health is 4
- Average final grade is 10

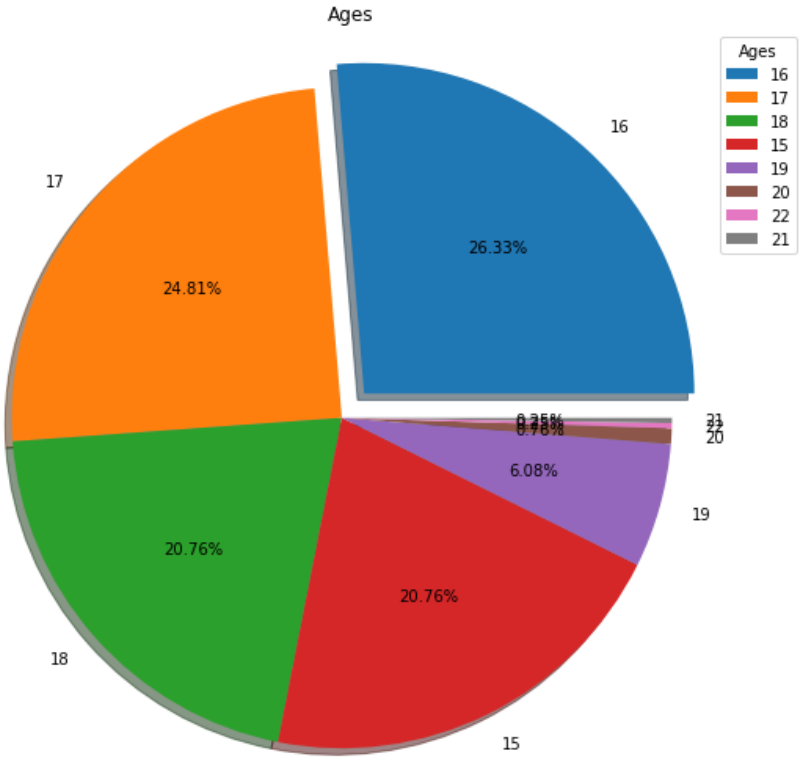
From our preliminary analysis on the dataset shows that our dataset has no missing values, is tidy and the features are of 2 datatypes int64 and one object. We are now ready to move onto our exploratory part of the project.

Exploratory Data Analysis

Firstly, we will check out the number of students who are part of the dataset and what binary sex they are.

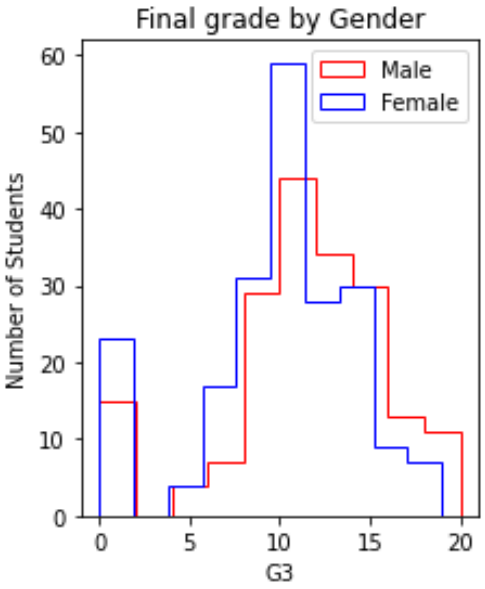


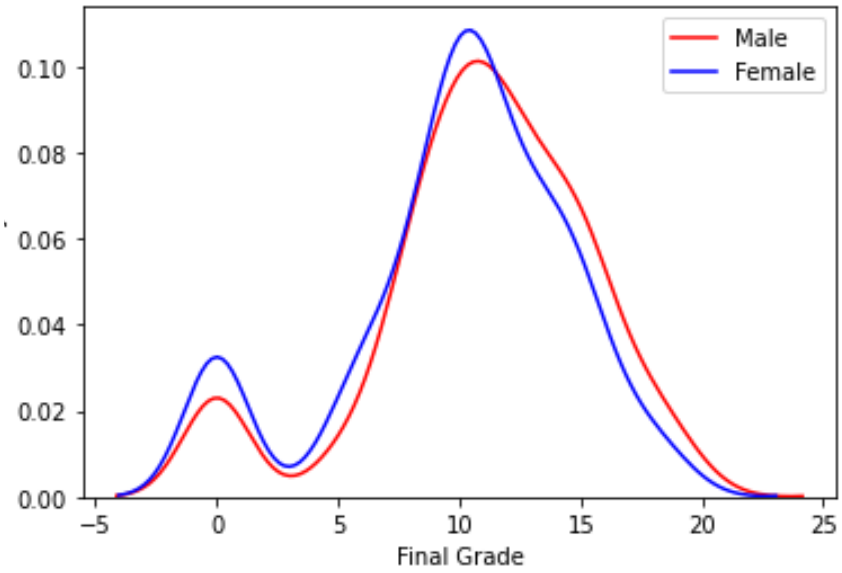
From our plot of our student dataframe we see the there are 208 females and 187 males. We will analyse the gender of students more in depth when we consider the final grades of students and their consumption of alcohol based on age. Let us look at the age of students using a pie chart. I assume that the younger you are the less likely you are to drink simply because of the drinking age in Portugal.



The majority of our are ages from 15-18 which is to be expected since we are dealing with secondary schools. After 18 though we see that there is a decrease in students aged 19 and most surprisingly we find out our dataset contains a small percent of our students being 20-22. When we performed a describe on our dataframe we found that the mean age was 16, which we can see clearly that this is the case.

Which gender scores a higher grade?





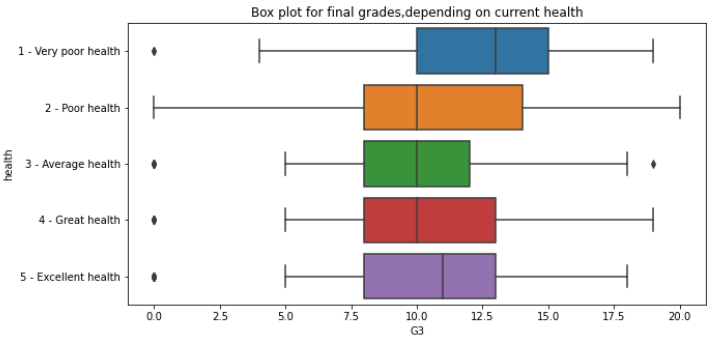

```
n_data = len(student)
average_grade = student['G3'].sum()/n_data
mgrade_df = student[student['sex']=='M']['G3']
fgrade_df = student[student['sex']=='F']['G3']
mgrade_mean = np.mean(mgrade_df)
fgrade_mean = np.mean(fgrade_df)

print("The average grade of male students is:", round(mgrade_mean,2))
print("The average grade of female students is:", round(fgrade_mean,2))
print("The average grade is:", round(average_grade,2))
print()
mgrade_med = np.median(mgrade_df)
fgrade_med = np.median(fgrade_df)
print("The median grade of male students is:", mgrade_med)
print("The median grade of female students is:", fgrade_med)

mgrade_stdev = st.stdev(mgrade_df)
fgrade_stdev = st.stdev(fgrade_df)
print("Standard Deviation of grades for male students:", mgrade_stdev)
print("Standard Deviation of grades for male students:", fgrade_stdev)
```

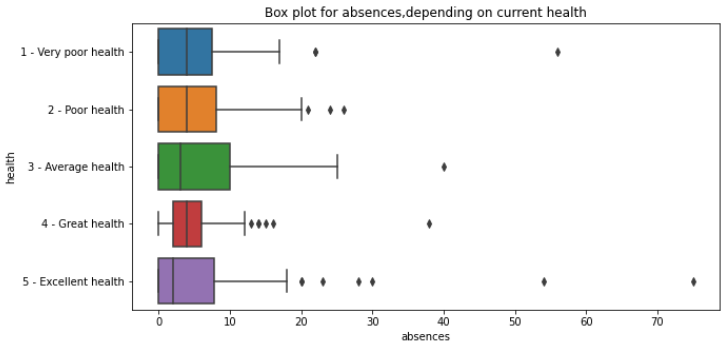
```
The average grade of male students is: 10.91
The average grade of female students is: 9.97
The average grade is: 10.42

The median grade of male students is: 11.0
The median grade of female students is: 10.0
Standard Deviation of grades for male students: 4.495296834986385
Standard Deviation of grades for male students: 4.622338337431134
```

When we perform plots for the final grade, we can see that the overall average of our students is 10.42. Male students have a higher average than females by almost one score and they achieve above a grade of 10 much more often than their female counterpart. The median grades for males and females are 10 and 11 respectively with the standard deviation being 4.49 and 4.62.

Is health affecting student grades and causing absences from school?

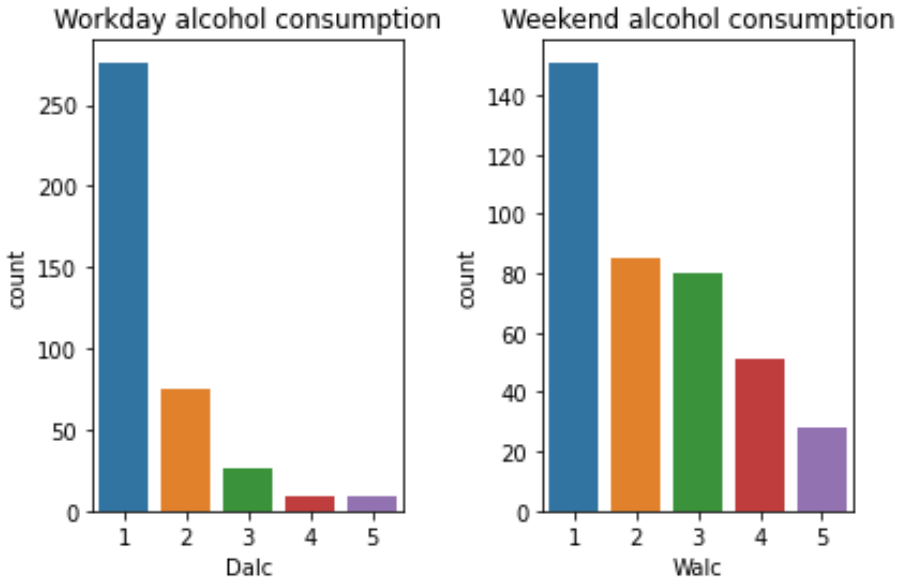
Now that we have done some basic exploration into our dataset finding out more about student's age and their final grades, we should check their health and how it affects final grades and their absences before we move onto the main area of our data. I believe that students who get a higher grade are much healthier and if they are not healthy, they might be absent from school and possibly will not know the math topic in detail for their exam.





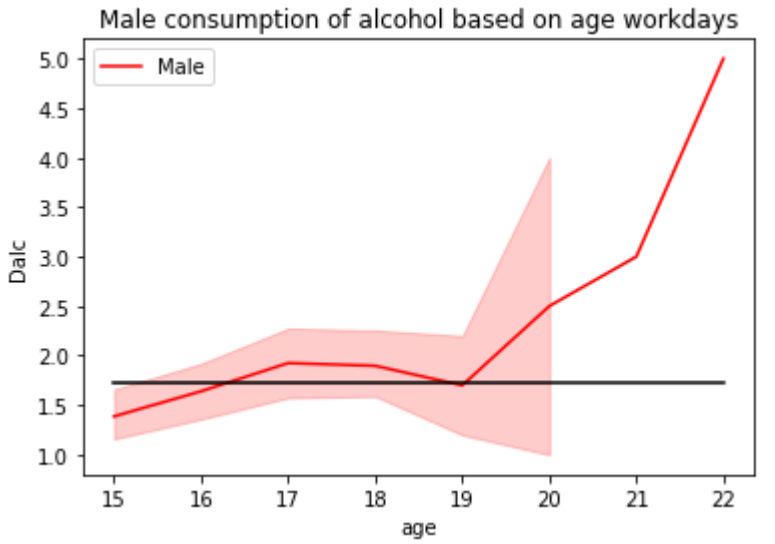
After plotting the data into boxplots, we can see that my hypothesis was wrong. From the students who enrolled into maths from both schools, 47 students have very poor health, but they appear to achieve a final grade of above 10. Although students with poor health (2) appear to be scoring just as well students with very poor health, not to say that students above average health are not doing well but there is a difference. When we check student's health against absences, we find that there is no solid connection between students being absent and their health. With our health analysis done we can now move onto the next are of study which is the alcohol consumption of students.

How much alcohol is consumed by students of different ages?



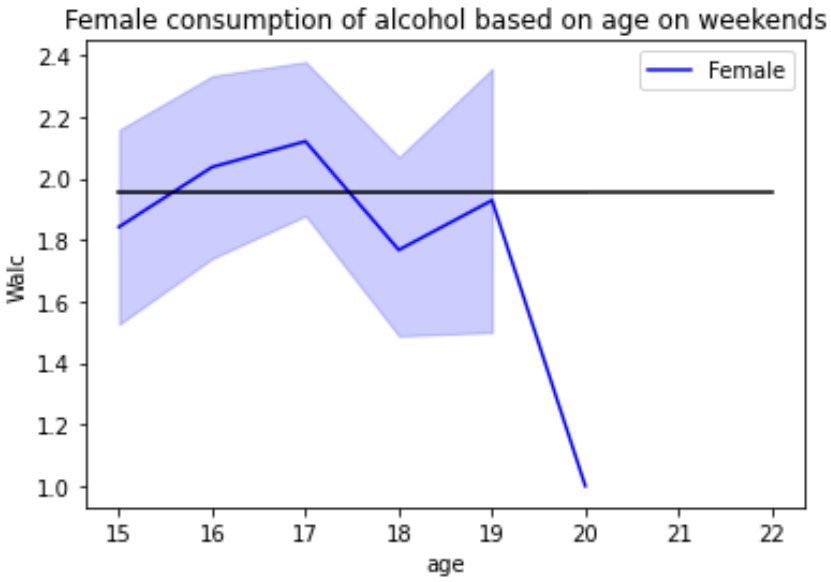
From our plot we see that no student said that they don't consume any alcohol during the week but we can also see that students tend be responsible when consuming alcohol when it's the a working

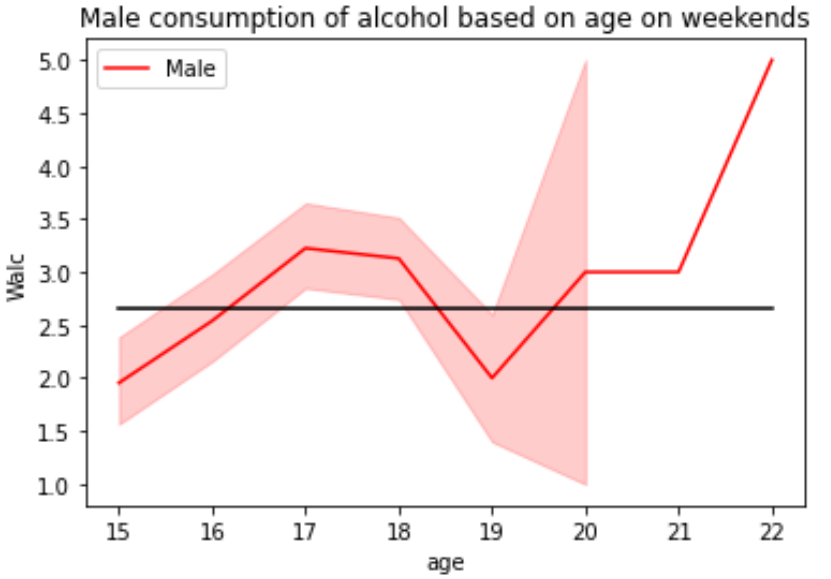
day because they know that they have school the next day so the majority of the students consume a small of amount of alcohol on weekdays. We do see a rise in the amount during the weekend which is to be expected since students want to simply relax after a hard week of school and knowing they do not have any important responsibilities they will drink a bit more. Let us check the consumption of alcohol of both sexes' based on age.



After performing a line plot on our data, we can see that females on average represented by the black line drink slightly less than males. Both sexes' tend to consume a slight amount of alcohol from 15-17 which is surprising since the drinking age in Portugal is 18 unless it is wine which then it's 16 so they are probably acquiring alcohol in some form or another. Once they become of legal drinking age which we can see a spike in both groups but a larger spike in the male group starting to drink more and more even if it is a workday! An interesting note here is the shade or count of people ending abruptly around 19-20, and that is because usually at that age students graduate and move on to a new stage in a life. This is clear for female students where you see a drop after 19 and only a few of them staying on. It is unfortunate to see an increase of alcohol usage for male students after most graduated. I believe that the reason for students having to continue their secondary studies instead of graduating at the average age of 18-19 is because of Portugal's overall education system being worse than other European countries. Only recently has it begun to improve its education

seeing as only in 2012 has seen a major improvement in mathematics, reading and science. In 2011 it was recorded that 58% of 25-34-year olds complete secondary education compared to the OCED average of 82% [1].

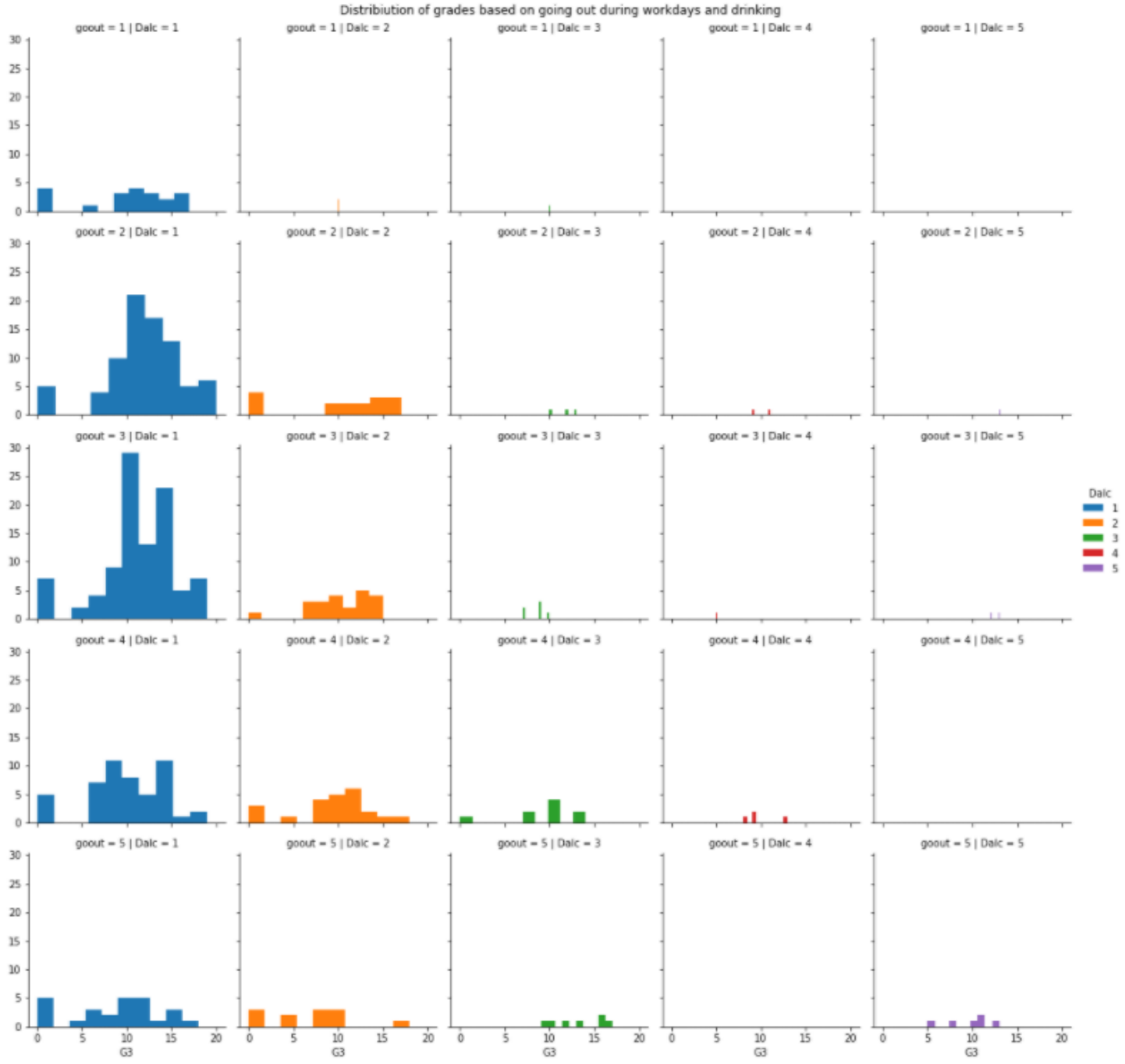


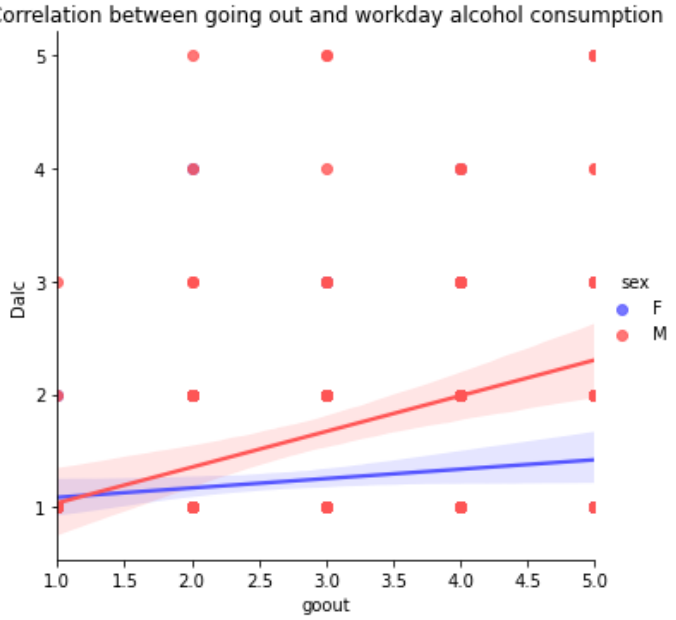


On weekends we can see the average consumption of alcohol, again represented by the black line go up which is to be expected since we saw in our bar plot that students drink more once they don't have to worry about school the next day. Male students continue to have a higher consumption average then women. It is surprising though to see so much more alcohol being drank by 15-17 years but after the age of 18 there is a slight drop in consumption before dropping down completely or increasing.

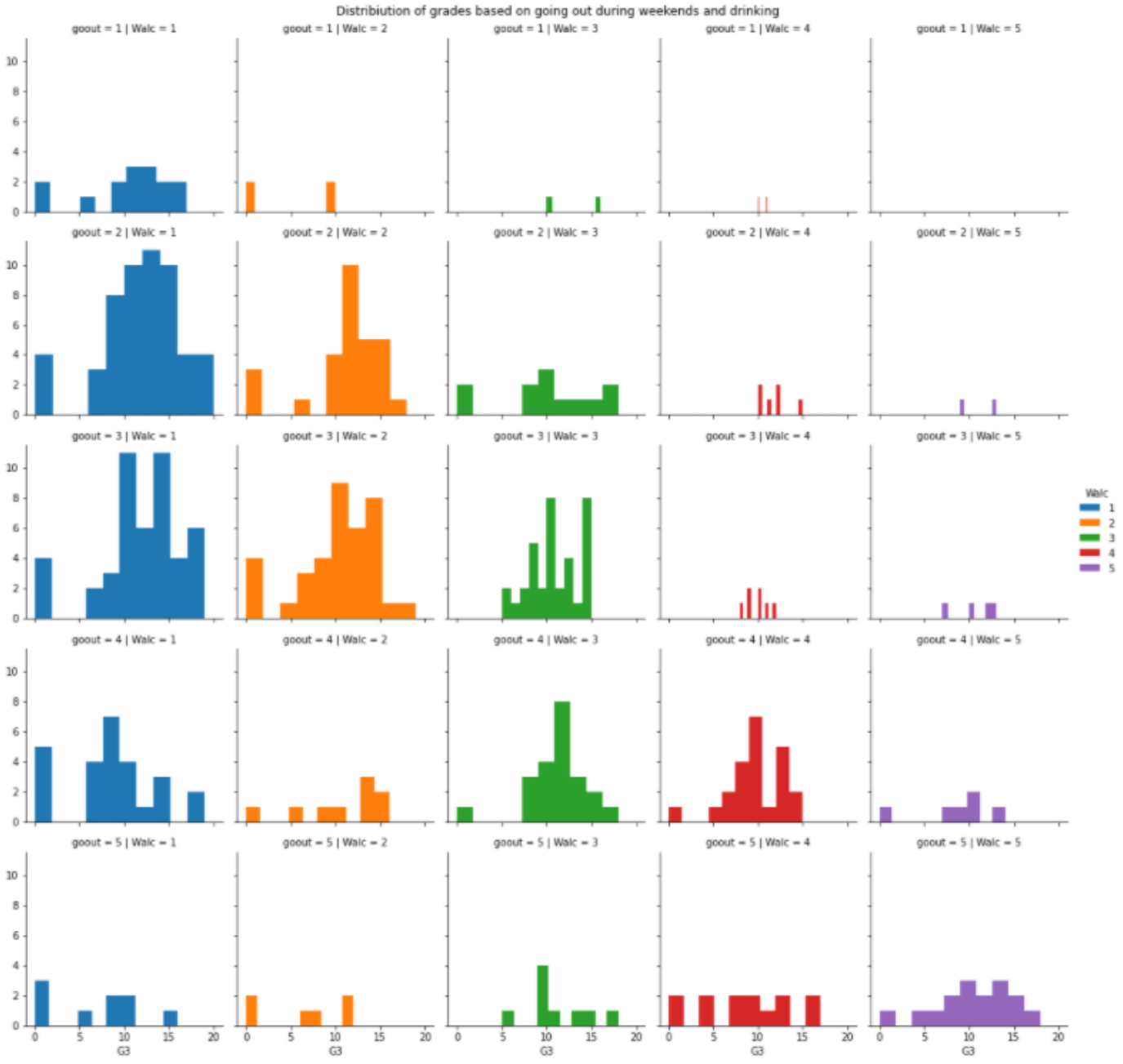
Is consuming alcohol correlated with going out with friends?

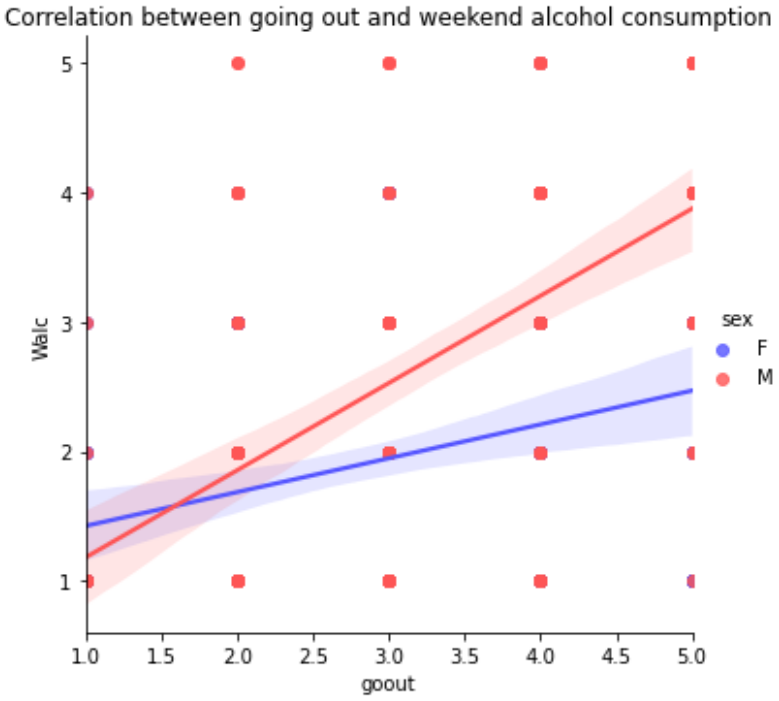
For this research question I would like to hypothesise if you are drinking alcohol on a weekend or weekday won't be correlated with hanging out with friends. To explore this area of our dataset we can use seaborn's Facet Grid which can help us map our dataset onto multiple grids that correspond to levels of different variables allowing us to see their relationship. In this case we are going to plot the column of going out with friends with both the weekday and weekend alcohol consumption columns plotting the grids against the final grade. On top of that we will check the correlation of the two columns we are testing with a lmplot.





From our plots we can gather that hanging out and consuming alcohol is different for both genders. It is clear that male students who hang out more with friends consume more alcohol when comparing it to the female group. Even though there is some correlation between hanging out and drinking it does not affect their grades.





As expected, the alcohol consumption among peers is quite different on weekends and much more prominent than on workdays. We saw in previous plots that both male and female students tend to consume much more alcohol, but we can also see now that during that time they spend more time with friends while doing so. However, it is not seriously affecting their grades at a larger scale but there is definitely more correlation between going out and drinking alcohol on the weekends rather than consuming it on weekdays.

If we compute the Pearson correlation coefficient on both areas and perform a chi-square test, we will find that:

```
# Compute Pearson correlation coefficient
corr, pval = stats.pearsonr(student['Dalc'], student['goout'])
corr = round(corr, 2)

print("Correlation between going out and consuming alcohol on workdays is:", corr)
print("The p-value between going out and consuming alcohol on workdays is:", pval)
```

```
Correlation between going out and consuming alcohol on workdays is: 0.27
The p-value between going out and consuming alcohol on workdays is: 7.136727169466113e-08
```

```
crosstab = pd.crosstab(student['Dalc'], student['goout'])
stats.chi2_contingency(crosstab)
```

```
(48.78628219209503,
 3.5714751508089954e-05,
 16,
 array([[16.07088608, 71.96962025, 90.83544304, 60.09113924, 37.03291139],
        [ 4.36708861, 19.55696203, 24.6835443 , 16.32911392, 10.06329114],
        [ 1.51392405,  6.77974684,  8.55696203,  5.66075949,  3.48860759],
        [ 0.52405063,  2.34683544,  2.96202532,  1.95949367,  1.20759494],
        [ 0.52405063,  2.34683544,  2.96202532,  1.95949367,  1.20759494]]))
```

```
# Compute Pearson correlation coefficient
corr, pval = stats.pearsonr(student['Walc'], student['goout'])
corr = round(corr, 2)

print("Correlation between going out and consuming alcohol on weekends is:", corr)
print("The p-value between going out and consuming alcohol on weekends is:", pval)
```

```
Correlation between going out and consuming alcohol on weekends is: 0.42
The p-value between going out and consuming alcohol on weekends is: 2.4016155113007542e-18
```

```
crosstab = pd.crosstab(student['Walc'], student['goout'])
stats.chi2_contingency(crosstab)
```

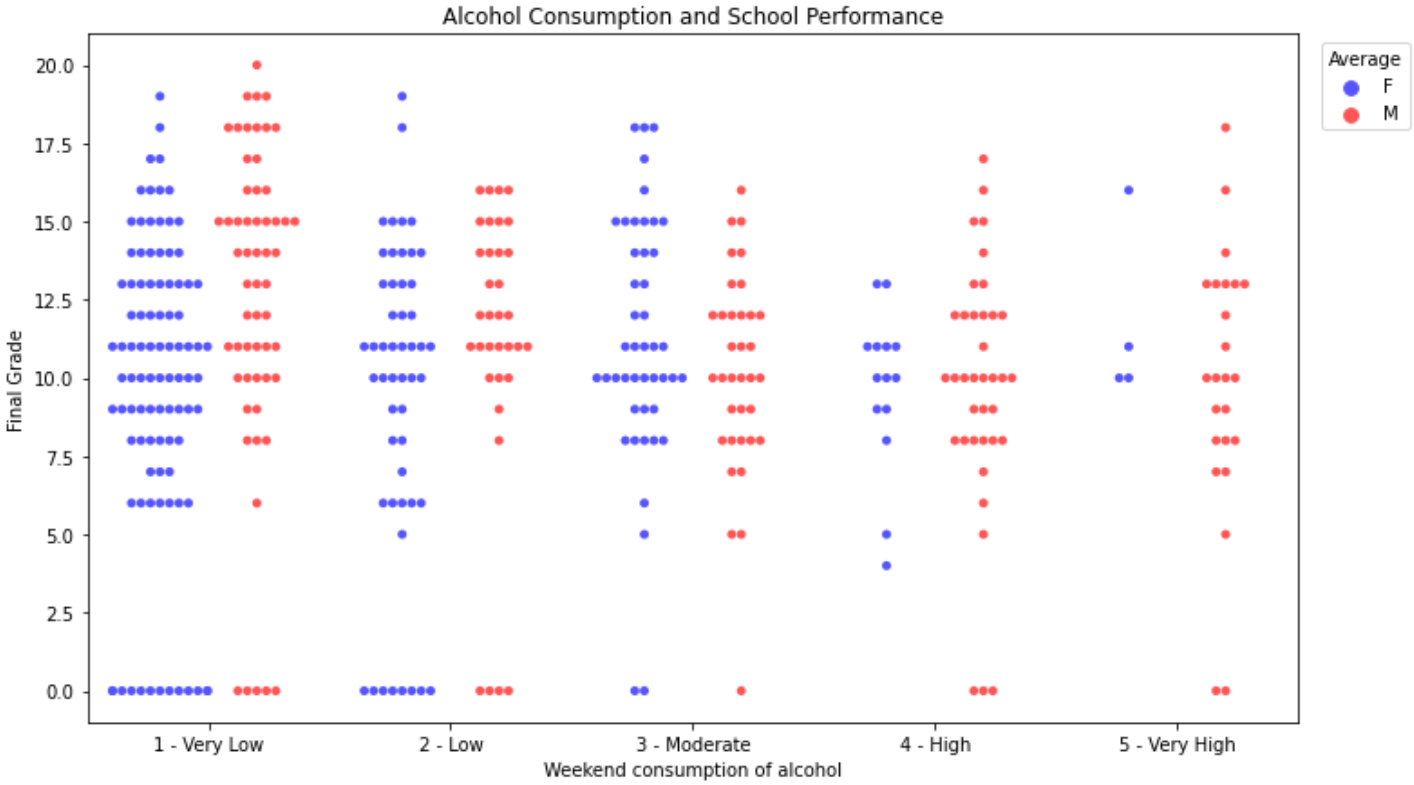
```
(116.56749687179874,
 2.50352859326213e-17,
 16,
 array([[ 8.79240506, 39.37468354, 49.69620253, 32.87594937, 20.26075949],
        [ 4.94936709, 22.16455696, 27.97468354, 18.50632911, 11.40506329],
        [ 4.65822785, 20.86075949, 26.32911392, 17.41772152, 10.73417722],
        [ 2.96962025, 13.29873418, 16.78481013, 11.10379747,  6.84303797],
        [ 1.63037975,  7.30126582,  9.21518987,  6.09620253,  3.75696203]]))
```

Our weekend corr value shows that there is a higher correlation of the two groups than our weekday corr value, meaning that our group of students prefer to spend their time alone on weekdays if they drink. In the first chi-square test we find our p-value to be (3.57e-05). Unfortunately, the majority of our frequencies are less than 5 so the results cannot be trusted fully. In the second chi-square test our p-value is (2.50e -17) and although there are a lot more frequencies that are greater than 5, some of our values are still less than the desired value. Nevertheless, from our two p-values it is clear that they are less than 0.05 meaning that our previously stated null hypothesis can be rejected. Thus, the results indicate that there is some relationship between going out and drinking but students who drink on weekdays seem to prefer to spend more time alone.

Does alcohol consumption affect your grades in secondary school?

Now that we have looked at the student's health, alcohol consumption between ages and if going out with friends has anything to do with the amount of alcohol students drink, we can now take a deeper look into their school performance while drinking. Here I will be returning to the main hypothesis of my projected that I stated in the introduction - To do this we can perform a swarm plot of both sexes' and see how they do depend on when they drink alcohol.

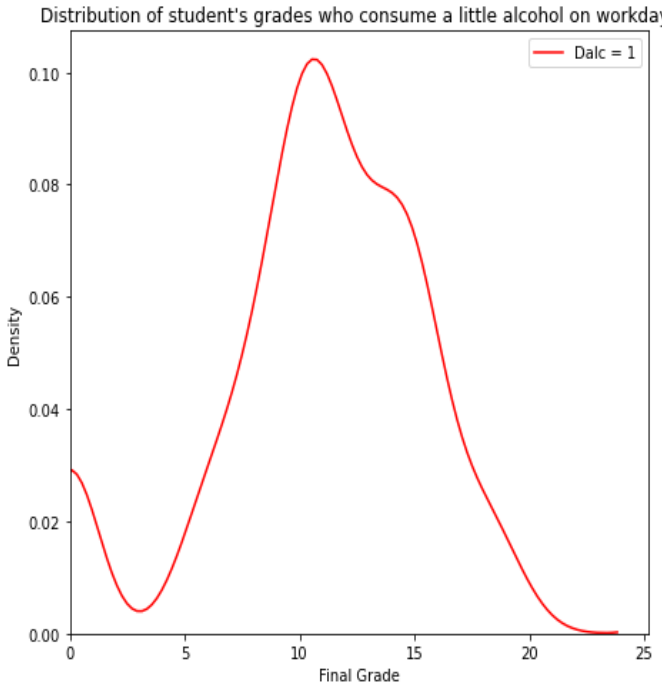
Here we can see our data is incredibly congregated on the left hand side which makes sense since we have previously noted that the students in our data don't consume that much alcohol when they know that they have school the next day. Majority of students tend to score above average when it comes to their final grade, but it can be seen that students who do consume more alcohol on workdays are scoring slightly lower scores than their peers.

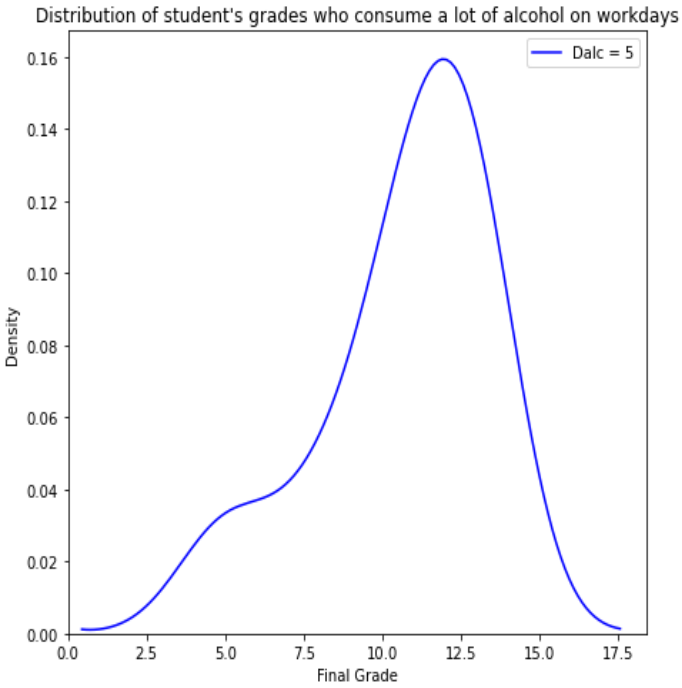


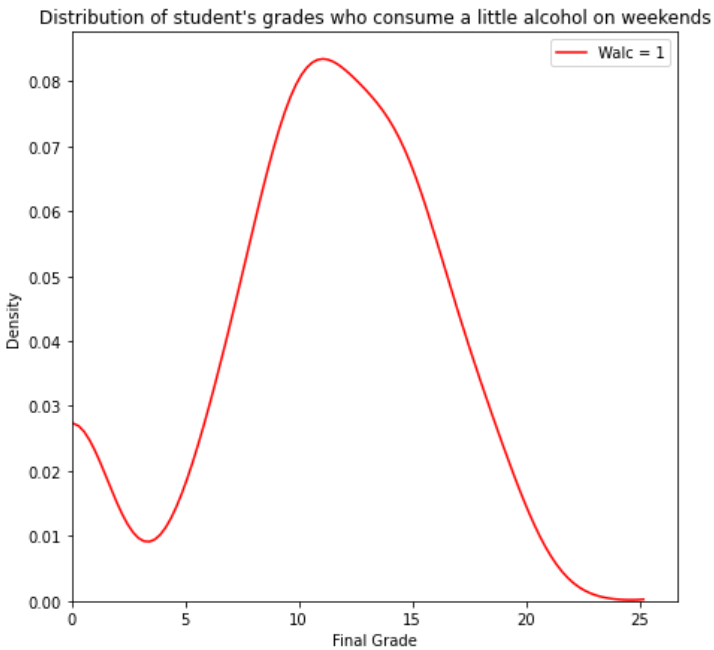
When we move onto the display of grades depending on weekend alcohol drinking our plot is a lot more evenly distributed. Both genders are doing well in their studies with the male group performing slightly better than their female peers. From the graph we can gather that the highest grade of 20 was achieved by a male student with a low alcohol consumption while the highest grade for the opposite sex was 19 who either consumed a very low or low amount of alcohol. This does not help prove our hypothesis, but it does tell us that students who drink less achieve better results.

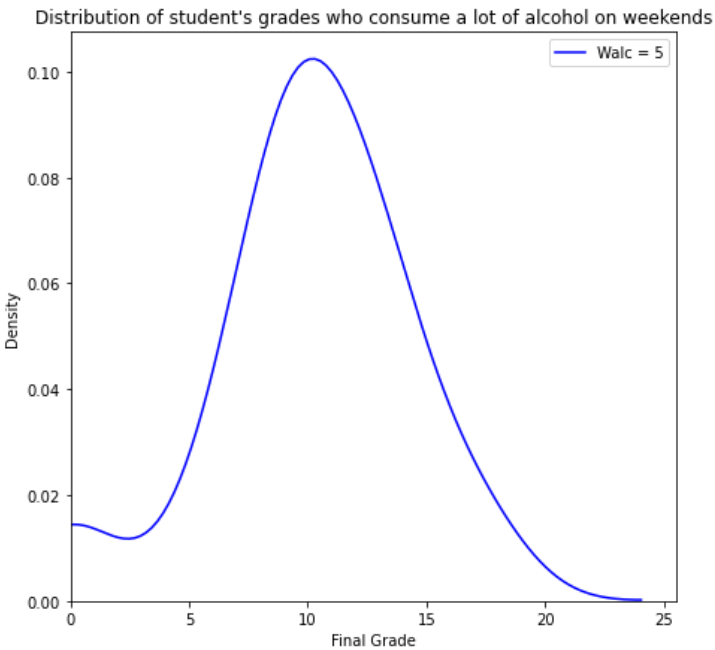
Are students who drink less/more inclined to score more in their final grade?

To make sure our hypothesis is proven properly we will create a normal distribution of student's grade and after that perform statistical tests on the data.









From the first set of distributions we can again see that those who love to drink on workdays is much less compared to the number of those who spend their weekdays sober. Those that do drink more their final grades do end up being significantly lower, ending at a 17.5.

In the second set the distributions are much more similar, and it seems it will not affect the grade at school even if some students decide to drink more during their time off from school. Now to compute the Pearson correlation coefficient.

```
# Compute Pearson correlation coefficient
corr, pval = stats.pearsonr(student['Dalc'], student['G3'])
corr = round(corr, 2)

print("Correlation between grade and consuming alcohol on workdays is:", corr)
print("The p-value between grade and consuming alcohol on workdays is:", pval)
```

```
Correlation between grade and consuming alcohol on workdays is: -0.05
The p-value between grade and consuming alcohol on workdays is: 0.2784914783598845
```

From our calculation it seems that the correlation of alcohol consumption on weekdays and the grade is -0.05, meaning that the two columns are moving in opposite directions with the same magnitude. The p-value we get is ~0.29 which is higher than 0.05 meaning it is not statistically significant indicating strong evidence for our null hypothesis. Before we decided if we reject the hypothesis lets take a look at the weekend column against the grade.

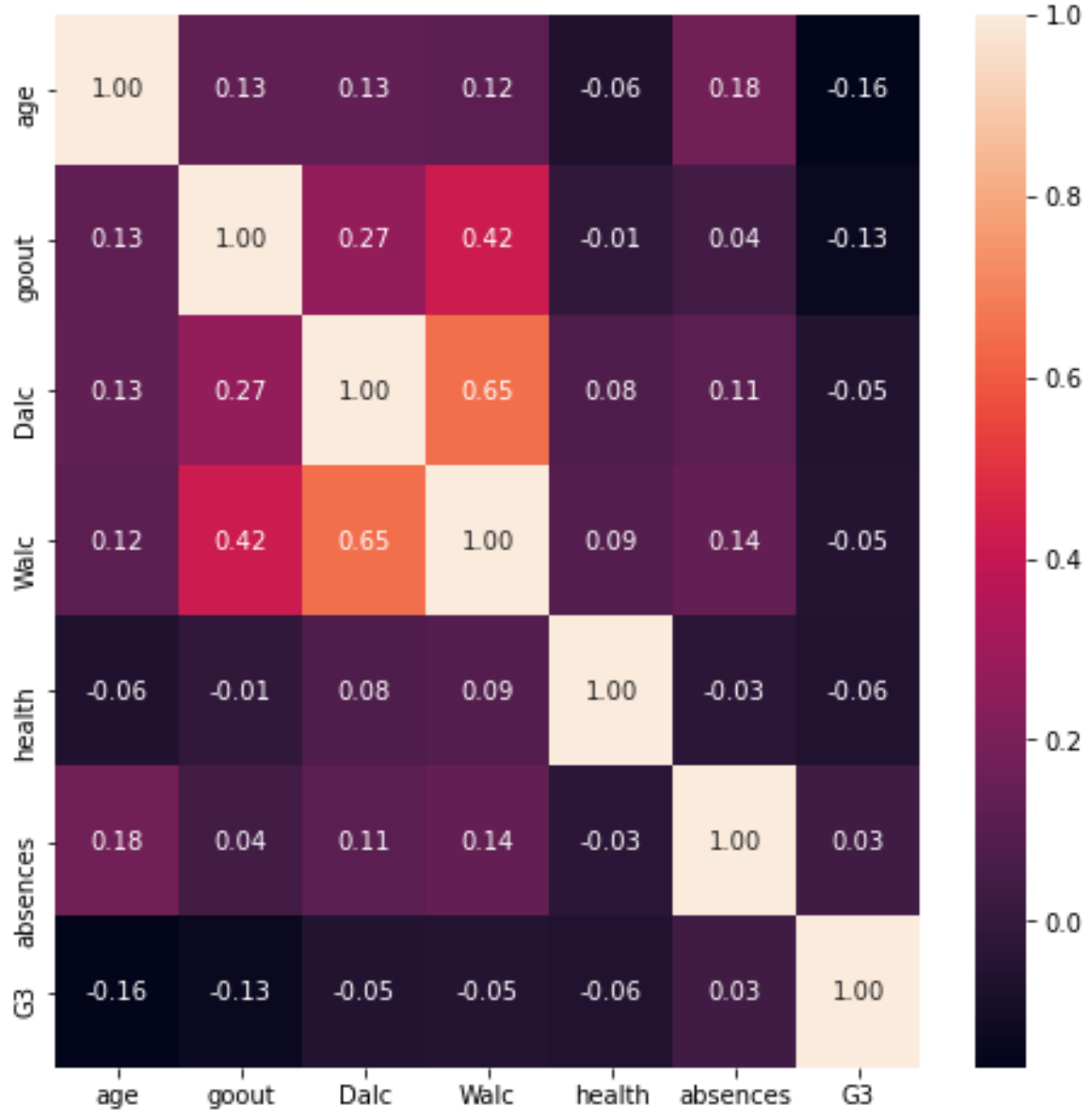
```
# Compute Pearson correlation coefficient
corr, pval = stats.pearsonr(student['Walc'], student['G3'])
corr = round(corr, 2)

print("Correlation between grade and consuming alcohol on weekends is:", corr)
print("The p-value between grade and consuming alcohol on weekends is:", pval)
```

```
Correlation between grade and consuming alcohol on weekends is: -0.05
The p-value between grade and consuming alcohol on weekends is: 0.30315210798427544
```

It appears that we have the exact same correlation between grade and weekend alcohol usage as we did in our workday calculations. Therefore, our columns will move in opposite directions with the same magnitude. The p-value we achieve is ~0.30, slightly higher than our previous p-value. This means that our area of study is not statistically significant, and we fail to reject the null hypothesis.

Quick Correlation Check

Now that we have performed our plots and tests, we are ready to draw up a conclusion for this data analysis. Before we do that though let us create a quick Pearson correlation so we can use it in the future.



Conclusion

This dataset has given us very interesting insights into the student's of Portugal. As the students age increases the failure raises. This might be due to some students hanging out more so they don't have time for studies but the Portuguese education system might be at fault here as well as we have seen from the article that can be found in the references page. Students health does not have an affect on their grades, it seems that it is precisely the opposite where students with below average health are getting better results than students who are in perfect condition. The amount of absences they have appears to have no correlation between their health as well. Students tend to be more responsible and do not consume too much alcohol when they have school the next day, but they know they can relax on the weekends. When relaxing on the weekends students hang out more frequently and consume more alcohol because of it but this not influencing their grades negatively. When looking at alcohol consumption and grades we found out that there is no effect on grades if the student drinks during the weekend but for those who consume a large amount during workdays we see a massive drop in score but is that what causes it? No, from our statistical analysis we found that the p-value in both sets was higher than 0.05 meaning we have to fail to reject the main hypothesis, meaning that higher alcohol consumption does not affect student grades.

References

Jeong Yee et.al (2014), “EDUCATION POLICY OUTLOOK PORTUGAL”, OCED [online]. Available from: http://www.oecd.org/education/EDUCATION%20POLICY%20OUTLOOK_PORTUGAL_EN.pdf [accessed 03 December 2020]

Dataset used to help with project [online]. Available from: https://www.kaggle.com/hely333/what-is-the-secret-of-academic-success [accessed 17 November 2020]

Dataset used to help with project [online]. Available from: https://www.kaggle.com/javidimail/effect-of-alcohol-use-on-gpa [accessed 20 November 2020]

Datacamp course used [online]. Available from: https://learn.datacamp.com/courses/statistical-thinking-in-python-part-1 [accessed 25 November 2020]