Shiva Chakravarthy Gollapudi
Sb1111
11468697

Final Project

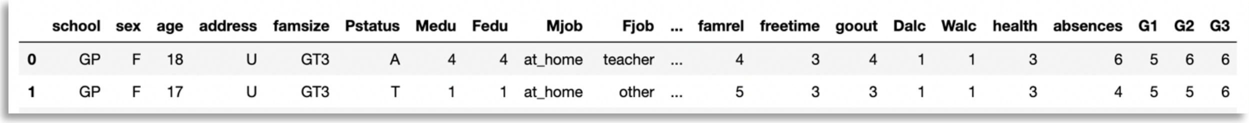
Exploratory Data Analysis

Introduction:

Student dataset is based on students who consume alcohol and their health status. 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. This dataset consists of 395 × 33 values.

❖ Exploring the Dataset:

The Dataset consists of many attributes and instances. Looking through the dataset one usually ends up in a web of questions about how this information is helpful in formulating an analysis. Let us take the very first instance from the dataset.



	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	6	5	6	6
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	4	5	5	6

The above table gives us the details about the student whose consume alcohol on daily basis and weekly basis. What were the health status of the student who consume alcohol on daily and weekly basis? Do they have any failures in exams based on the absent history? Which age group of people consume high percentage of alcohol? These questions can be answered by taking other instances into consideration and comparing them. A visual Analysis of this information helps us understand and gain insights about the data.

Methods:

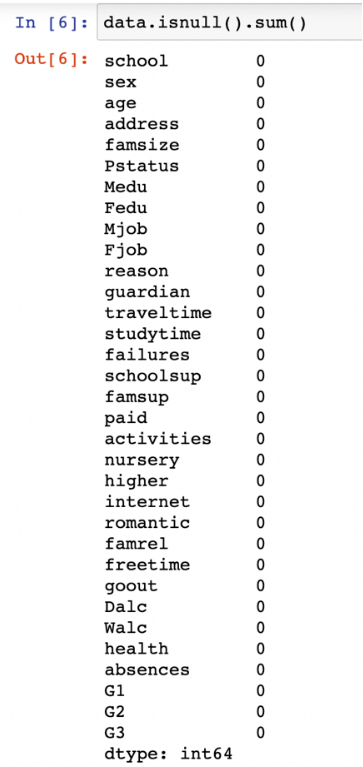
I have used python coding and Jupyter open-Source software to find the missing data and drop few columns which are not useful for analysis and visualization. Data preprocessing using python is popular, both in terms of pay it offers and popularity amongst recruiters looking for Python skills. With a rise in technologies like machine learning, artificial intelligence and predictive analytics, the need for professionals with a thorough knowledge of Python skills are much in demand. Apart from its general-purpose use for web development, it is widely used in scientific computing, data visualization, data mining and others. For this project I have used only the Jupyter notebook for the visualization.

Questions:

- Number of Students who are consume alcohol daily (Low to High) and weekly (Low to High) compared with students who are living with parents or not?
- Which age group students consume alcohol high and low on daily and weekly basis?
- What were the health status of the student who consume alcohol on daily and weekly basis?
- Compare first period grade, second period grade and final grade with failures?

Dataset Pre-Processing:

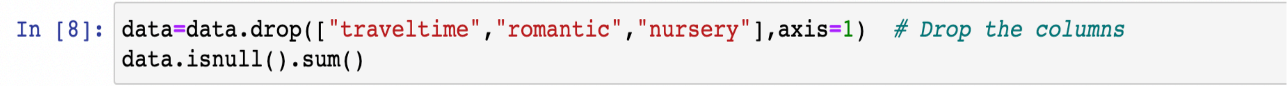
The Original Dataset found on the website. I have used python to look at the missing values. The count of Missing values can be displayed by using the below code.



```
In [6]: data.isnull().sum()

Out[6]: school        0
        sex           0
        age           0
        address       0
        famsize       0
        Pstatus       0
        Medu          0
        Fedu          0
        Mjob          0
        Fjob          0
        reason        0
        guardian      0
        traveltime    0
        studytime     0
        failures      0
        schoolsup     0
        famsup        0
        paid          0
        activities    0
        nursery       0
        higher        0
        internet      0
        romantic      0
        famrel        0
        freetime      0
        goout         0
        Dalc          0
        Walc          0
        health        0
        absences      0
        G1            0
        G2            0
        G3            0
        dtype: int64
```

Data Cleaning:

As the don't have any missing values, we are dropping few columns which are not useful for analysis and visualization.



```
In [8]: data=data.drop(["traveltime","romantic","nursery"],axis=1)  # Drop the columns
        data.isnull().sum()
```

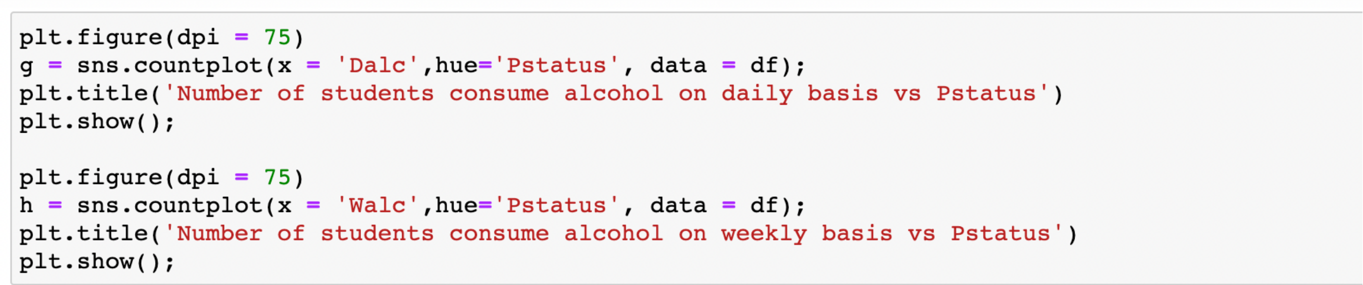
Data Visualization:

Question 1: Number of Students who consume alcohol on daily (Low to High) and weekly (Low to High) compared with students who are living with parents or not?

Results:

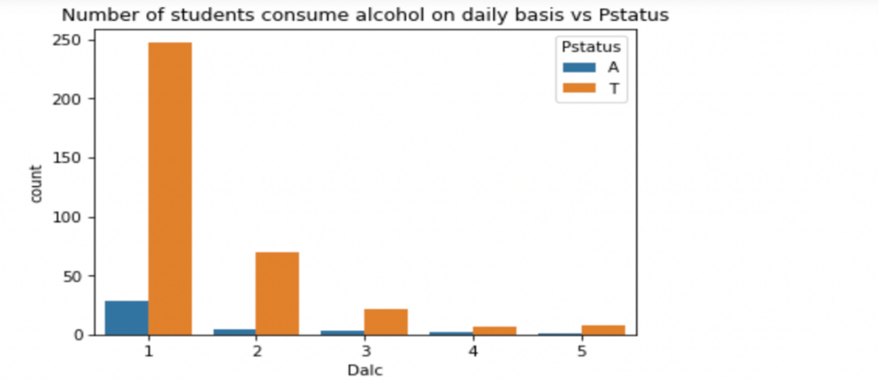
To visualize and evaluate the question we need to investigate the data. The data has an attribute called Dalc means workday alcohol consumption and Walc means weekend alcohol consumption compared with PStatus, which represents parent's cohabitation status, the attribute has T' - living together or 'A' – apart. Now, comparing this information to another attribute which is Dalc and Walc, which also has 5 values in it, 1 - very low to 5 - very high. We can plot a count plot graph for the count of number of students consume alcohol low to high on daily and weekly basis.

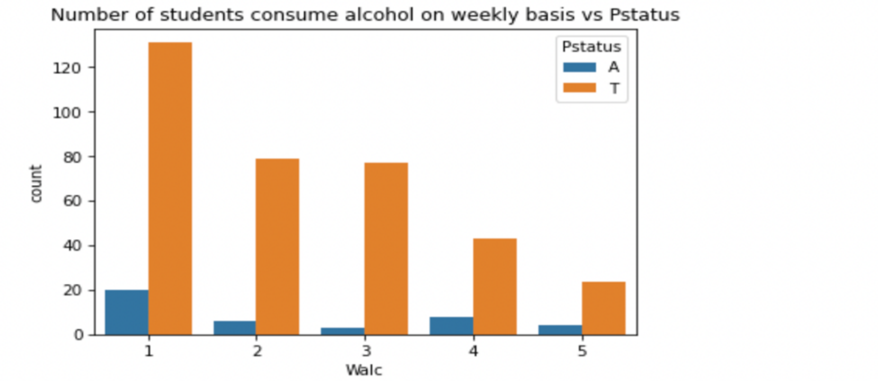
Code:



```
plt.figure(dpi = 75)
g = sns.countplot(x = 'Dalc',hue='Pstatus', data = df);
plt.title('Number of students consume alcohol on daily basis vs Pstatus')
plt.show();

plt.figure(dpi = 75)
h = sns.countplot(x = 'Walc',hue='Pstatus', data = df);
plt.title('Number of students consume alcohol on weekly basis vs Pstatus')
plt.show();
```

Visualization:





From the above visualization, we can conclude that students who are living with the parents consumes more alcohol compared to the students apart from their parents on both daily and weekly basis. students who consume less alcohol on daily and weekly are more are more compared to average and high consumed students.

Rationale:

The purpose of this graph is to convey relational information quickly as the bars display the quantity for a particular category. On Y-axis the count of the student will be displayed, while the X-axis is the Dalc and Walc attributes.

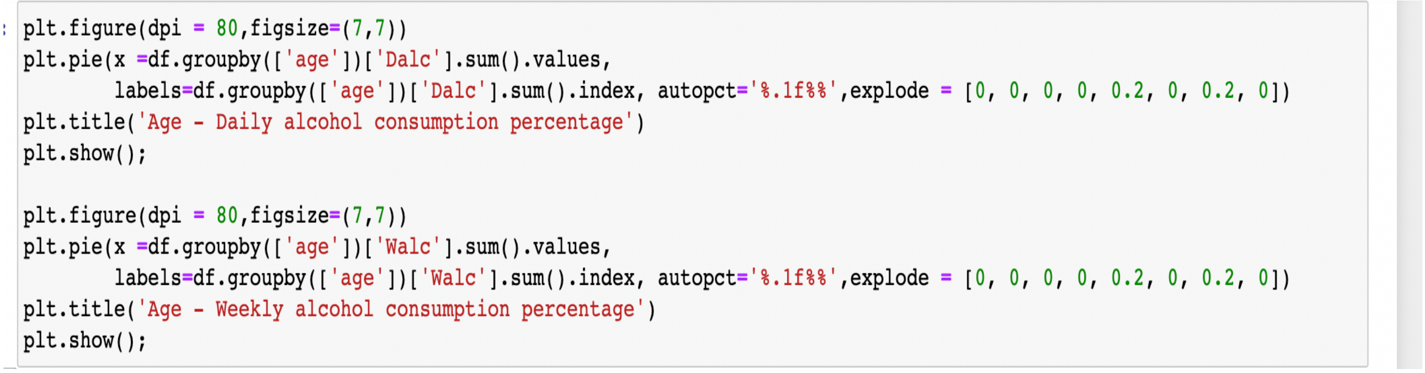
When interpreting a bar graph, the length of the bars/columns determines the count of the students as described on the Y-axis. The X-axis is the variables of the Dalc and Walc, on the scale of 1- very low to 5 – very high. Colors will be distinguishing between the students who lives with parents and apart from their parents.

Question 2: Which age group students consume alcohol high and low on daily and weekly basis?

Results:

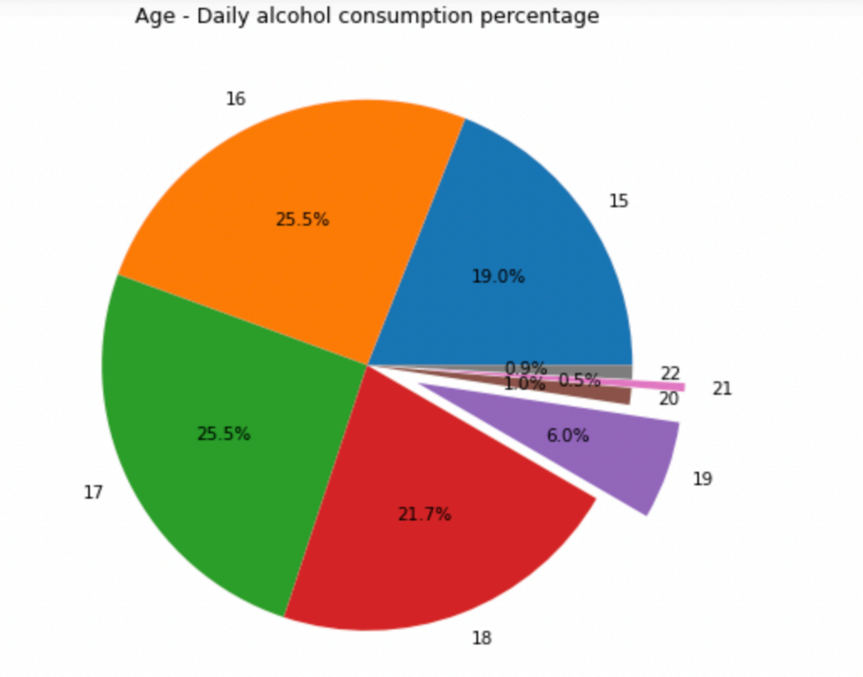
To visualize and evaluate this question we need to investigate the data. The data has an attribute called Dalc means workday alcohol consumption and Walc means weekend alcohol consumption compared with age, which represents age of the students. We can plot a pie chart for the percentage of the students which age group has high consumption of alcohol.

Code:

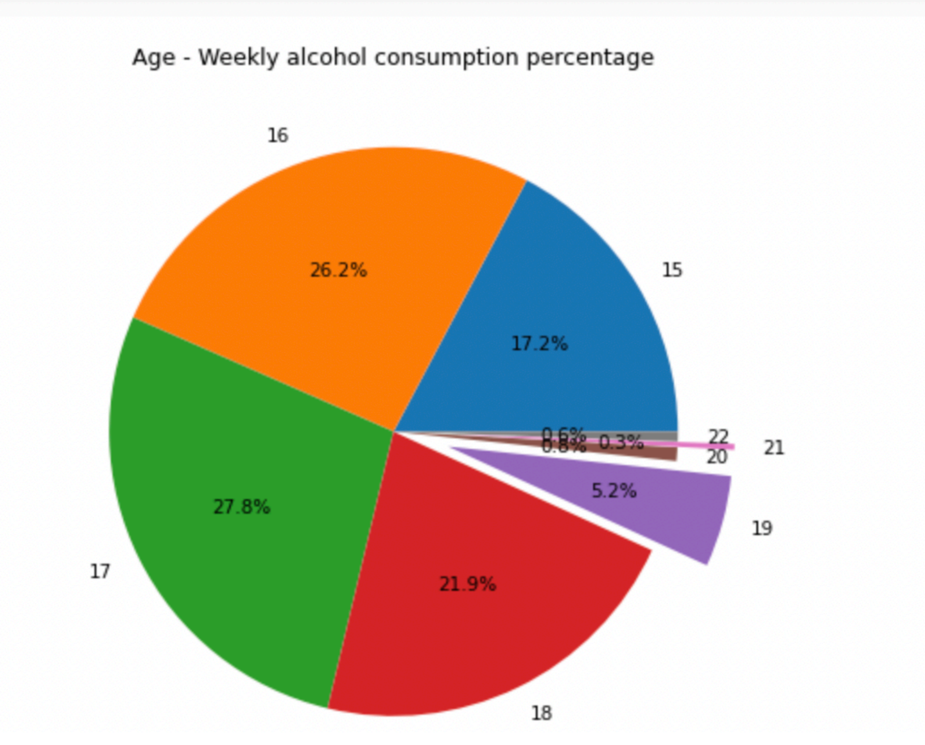


```
plt.figure(dpi = 80,figsize=(7,7))
plt.pie(x =df.groupby(['age'])['Dalc'].sum().values,
       labels=df.groupby(['age'])['Dalc'].sum().index, autopct='%.1f%%',explode = [0, 0, 0, 0, 0.2, 0, 0.2, 0])
plt.title('Age - Daily alcohol consumption percentage')
plt.show();

plt.figure(dpi = 80,figsize=(7,7))
plt.pie(x =df.groupby(['age'])['Walc'].sum().values,
       labels=df.groupby(['age'])['Walc'].sum().index, autopct='%.1f%%',explode = [0, 0, 0, 0, 0.2, 0, 0.2, 0])
plt.title('Age - Weekly alcohol consumption percentage')
plt.show();
```

Visualization:



Age - Daily alcohol consumption percentage



Age - Weekly alcohol consumption percentage

The above visualization clearly shows that the age group of 16 and 17 has equal percentage and high compared to other age groups of students who consume daily and the least is the age group of 21. In the second visualization, age group of 17 having percentage of the students who consumes alcohol weekly. If we look in deep age 17 is high percentage of students consumes alcohol in both daily and weekly and 21 is the least.

Rationale:

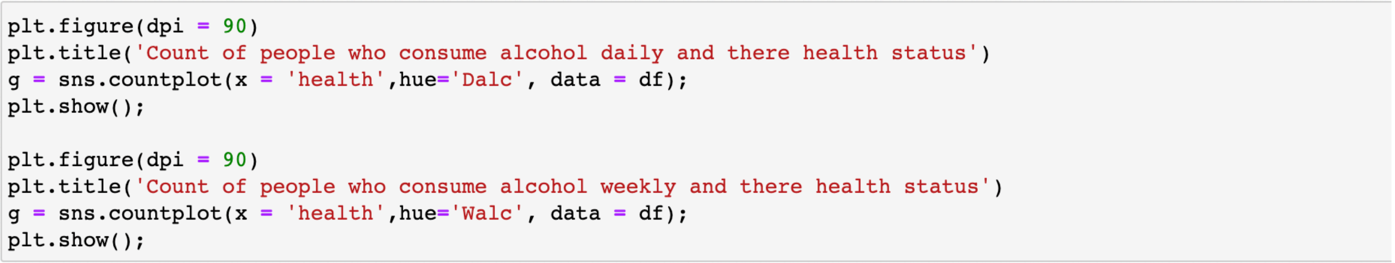
The main purpose of the pie chart is to show part-whole relationships. On X-axis we have grouped the age and added the number of students who consume alcohol on daily and weekly basis. The chart distinguished the percentage of the students who consumes alcohol clearly. Colors will be distinguishing between the among the different age group of students.

Question 3: What was the health status of the people who consume alcohol?

Results:

To visualize and evaluate the question we need to investigate the data. The data has an attribute called health means current health status from 1 - very bad to 5 - very good versus with attribute Dalc means students who consume alcohol consumption daily and Walc means weekend alcohol consumption We can plot a count plot graph for the count of number of students consume alcohol low to high on daily and weekly basis.

Code:



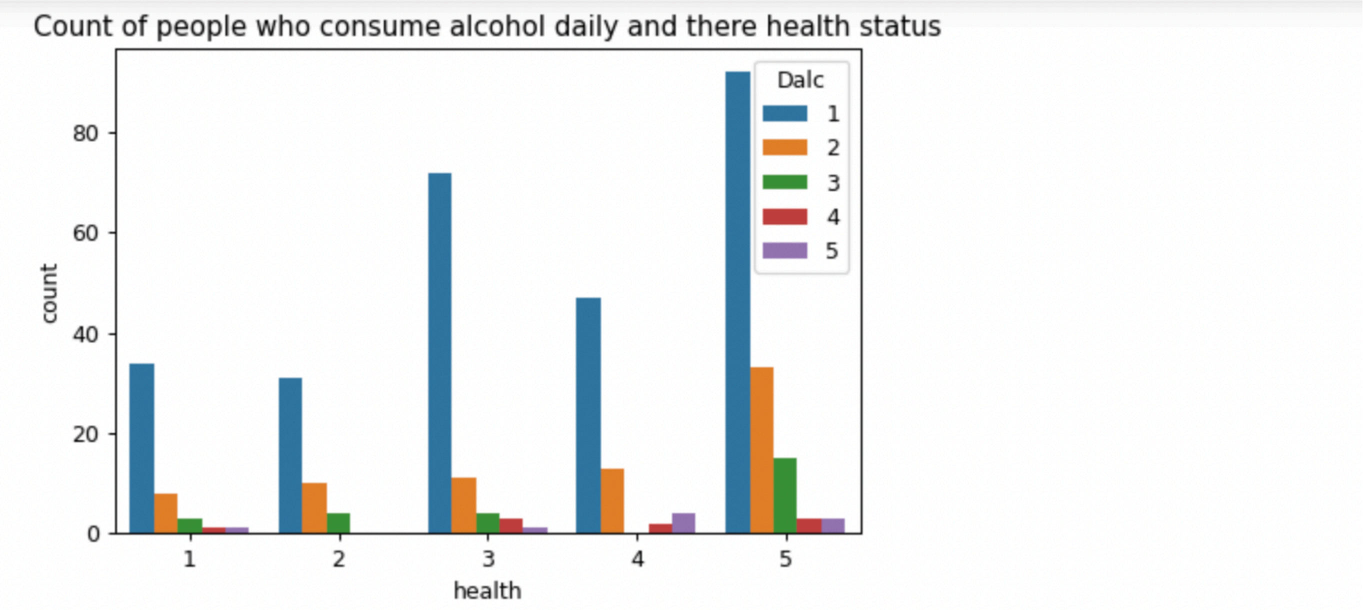
```
plt.figure(dpi = 90)
plt.title('Count of people who consume alcohol daily and there health status')
g = sns.countplot(x = 'health',hue='Dalc', data = df);
plt.show();

plt.figure(dpi = 90)
plt.title('Count of people who consume alcohol weekly and there health status')
g = sns.countplot(x = 'health',hue='Walc', data = df);
plt.show();
```

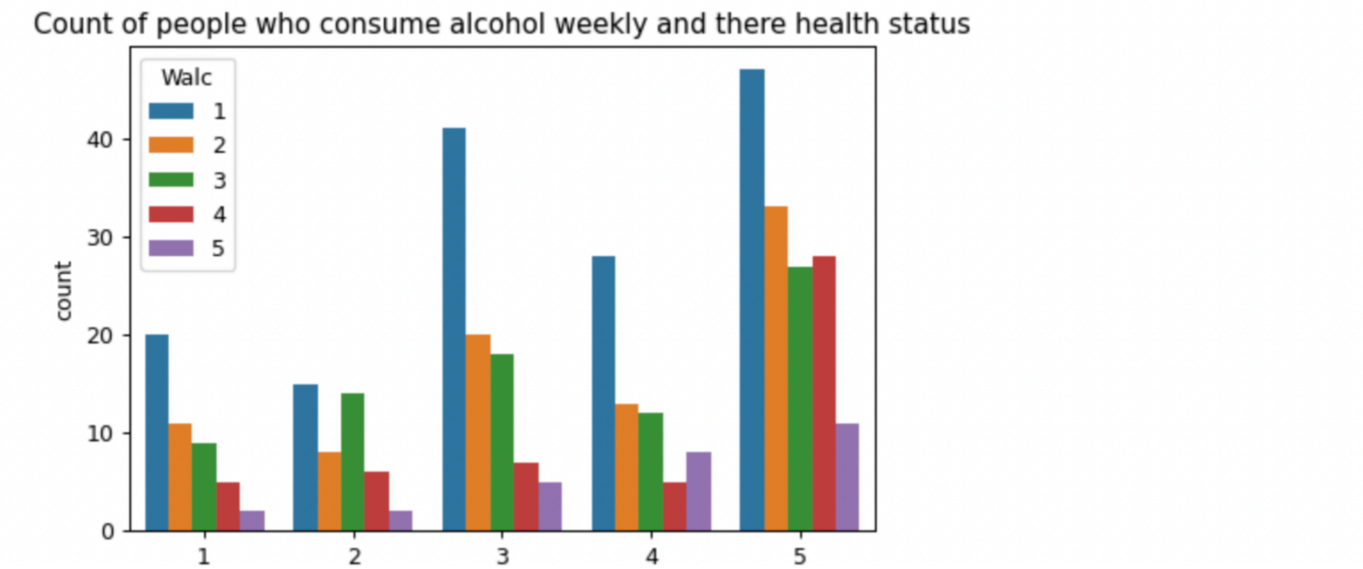
Rationale:

The purpose of this graph is to convey relational information quickly as the bars display the quantity for a particular category. On X-axis the of the student depends on the health (1-5) will be displayed, while the X-Count of the students consume alcohol bases Dalc and Walc attributes.

When interpreting a bar graph, the length of the bars/columns determines the count of the students as described on the X-axis. The Y-axis is the variables of the Dalc and Walc, on the scale of 1- very low to 5 – very high. Colors will be distinguishing between the students Dalc and Walc who consume alcohol. (1 to 5).

Visualization:


Count of people who consume alcohol daily and there health status


Count of people who consume alcohol weekly and there health status

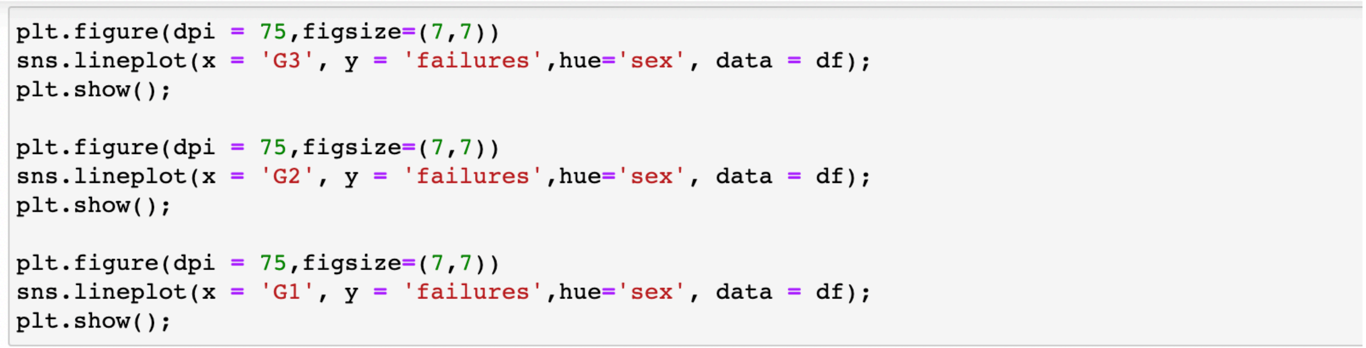
The above visualization clearly shows the students who consume low alcohol having very good health condition in both daily and weekly basis. It clearly shows that the health of the student is not depend on the consumption of the alcohol either daily or weekly.

Question 4: Compare first period grade, second period grade and final grade with failures?

Results:

To visualize this scenario, the data has an attribute called G1 means first period grade (numeric: from 0 to 20), G2 means second period grade (numeric: from 0 to 20), G3 means G3 - final grade (numeric: from 0 to 20, output target) versus with the failure count. We can plot a line graph to know the trend of the failures with respect to their grades.

Code:



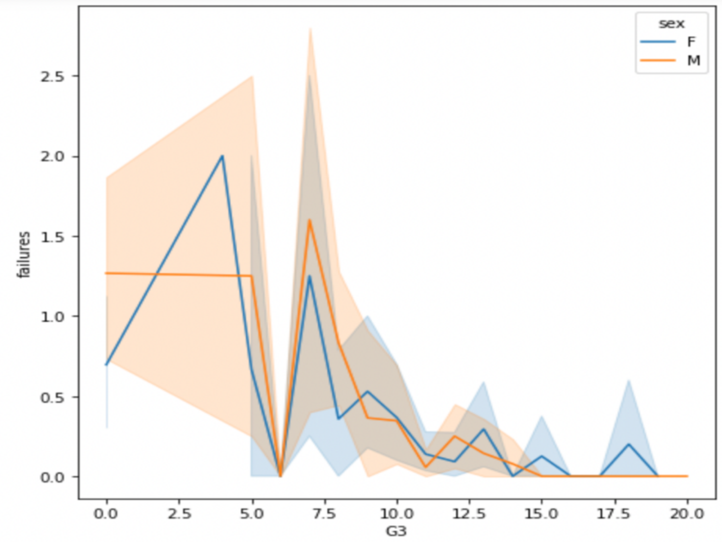
```
plt.figure(dpi = 75,figsize=(7,7))
sns.lineplot(x = 'G3', y = 'failures',hue='sex', data = df);
plt.show();

plt.figure(dpi = 75,figsize=(7,7))
sns.lineplot(x = 'G2', y = 'failures',hue='sex', data = df);
plt.show();

plt.figure(dpi = 75,figsize=(7,7))
sns.lineplot(x = 'G1', y = 'failures',hue='sex', data = df);
plt.show();
```

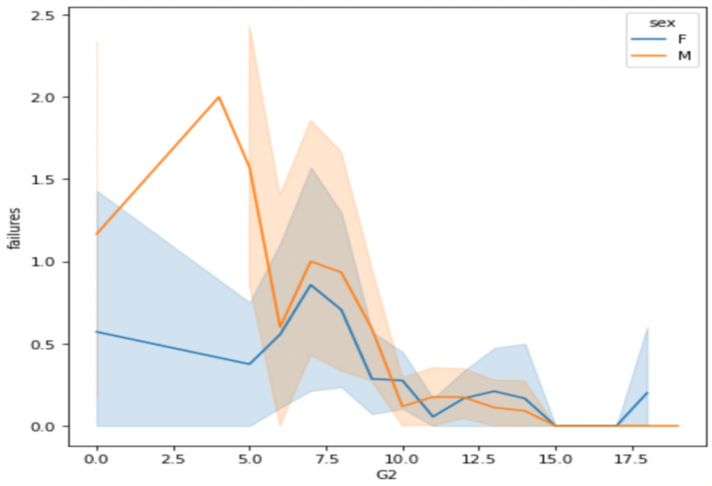
Rationale:

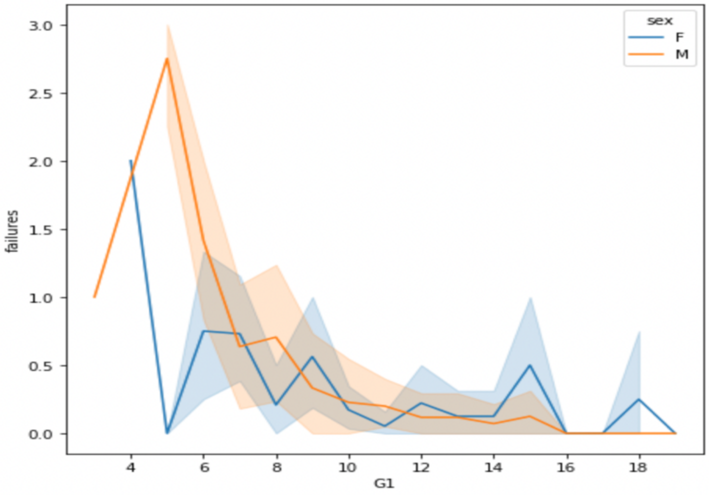
A line graph is commonly used to display change over time as a series of data points connected by straight line segments on two axes. The line graph therefore helps to determine the relationship between two sets of values, with one data set always being dependent on the other set. In this graph X-axis have taken the grades and the Y-axis is the failures and the color difference is the gender of the student.

The below visualization clearly explains there are high failures, below 5 grades for first, second and final exam. In first period grades having female student are high compared to male, but in the second and final exams male students are more failures compared to female students.

Visualization:







Software’s and Tools Used:

Python: Data Cleaning and Fixing the Missing Values and the Visualization in Jupyter notebook.

Jupyter Notebook: Jupyter Notebook is a web-based interactive development environment for Jupyter notebooks, code, and data. Jupyter is flexible: configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning. Jupyter is extensible and modular: write plugins that add new components and integrate with existing ones.

References: https://www.cs.ubc.ca/~tmm/courses/547-14/

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