**Students Alcohol Consumption**

Performing an analysis on a dataset for some real-world outcomes.

Logo

Description automatically generated

**ST. CLAIR COLLEGE, MISSISSAUGA**

**Presenters:**

**Udit Arora 0769036**

**Manjit Singh 0765235**

**Komal Arora 0763537**

**Diksha Dalia 0766406**

**Sapna 0765232**

**Supervisor – Dr. Savita Seharawat**

**Abstract:**

We found this dataset very amazing. We downloaded it from Kaggle.

The reason behind using this data set is finding the reason regarding the consumption of alcohol among youth because in this present era students are more vulnerable to alcoholic substances, it is happening because of their surroundings, modernization and the company of their friends. Our focus is to focus over the effect on the performance of students due to alcohol in grades the course subject, Math, Which include the data regarding mother edu, Father edu , Family Support etc, Which provide the Overview of data of two School in which students get effected by alcohol in weekdays and weekends and how their marks get affected. In this dataset, we will also explain about their grades. Grades are divided into G1, G2, G3.

Using the attributes and performing the functions and other codes help us to find out the consumption of alcohol made by students. While performing on the datasets and finding the solutions to some questions

**1 Does alcohol effect success i.e., Grades and what factors contribute in that?**

This data will help us to find that what are the factors that affect the grades of students those could be the education level of parents, how often they absence from school, travel time and study time.

**2 Which gender, age group consume more alcohol?**

This data will help us to find that which age group and gender are addicted more towards the alcohol.

**3 Which school has better at grade performance among both schools?**

From the research of that questions, we will come to a point that which school have better grades from GP and MS.

**Objective**: Our main objective is to focus on what happen with the student’s success rate in grades who drink alcohol on weekdays and weekends as well as other factors affecting their grades like parents’ involvement and their education and the attendance, health along with what students do in their free time.

**Description:** Dataset contains 33 columns about students which explains data about their grades, performance, how often they are using internet, their attendance record and family background like their parents are living together or apart and other factors related to their personal life.

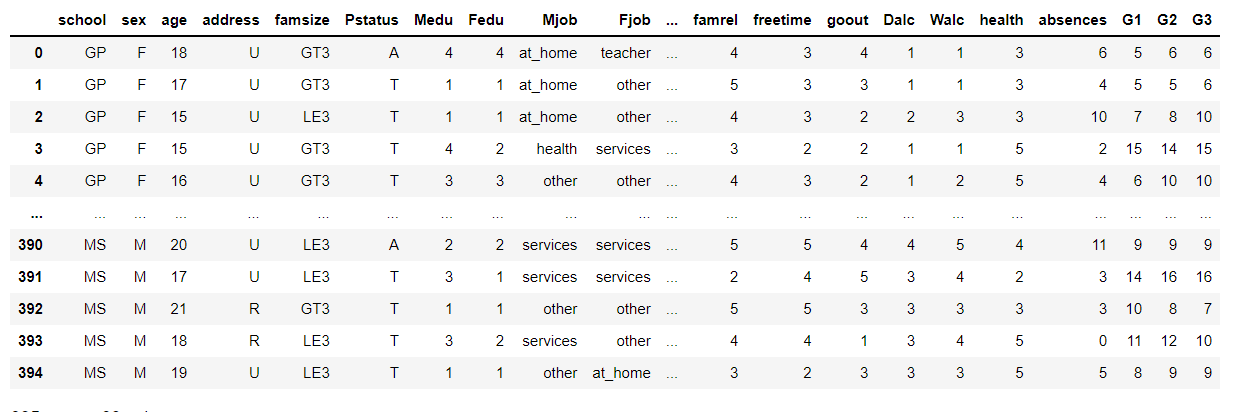
**Target attribute:** Grade (G3)

* The numerical variables are – age, absences, G1, G2, G3
* The categorical variables are – school, sex, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob, reason, guardian, traveltime, studytime, failures, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic, famrel, freetime, goout, health, alc.

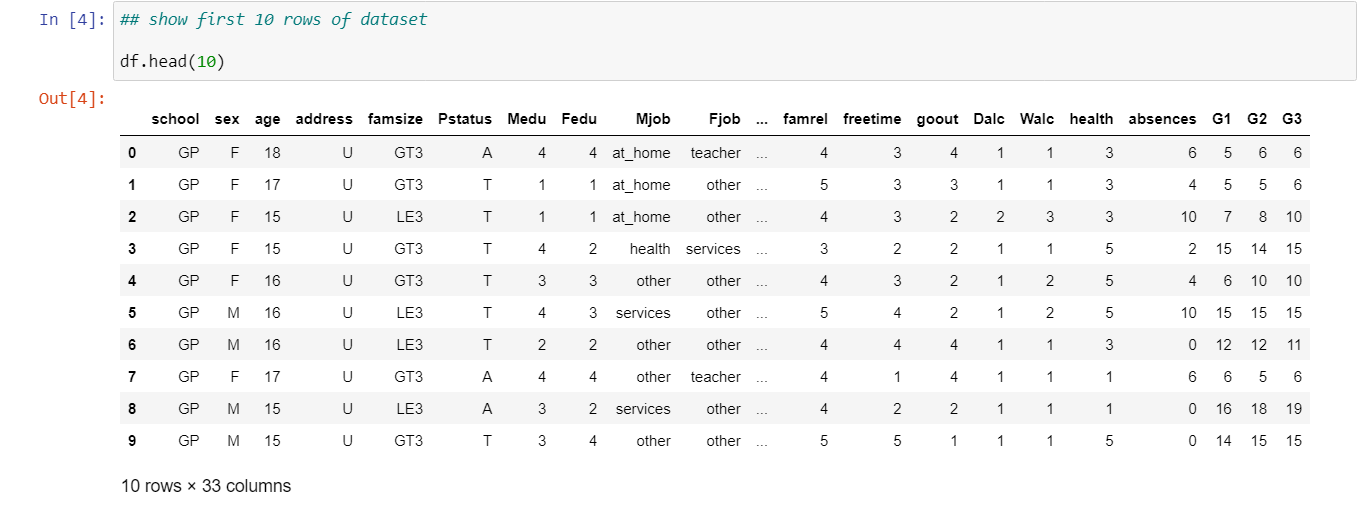
**Attributes in the dataset:**

|  |  |  |
| --- | --- | --- |
| S. No | VAR\_NAME | VAR\_DESCRIPTION |
| 1 | school | student's school ('GP' or 'MS’) |
| 2 | Sex | student's sex ('F' - female or 'M' – male) |
| 3 | Age | student's age (15 to 22) |
| 4 | address | student's home address type ('U' - urban or 'R' – rural) |
| 5 | famsize | family size ('LE3' - less or equal to 3 or 'GT3' - greater than 3) |
| 6 | Pstatus | parent's cohabitation status ('T' - living together or 'A' – apart) |
| 7 | Medu | mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education) |
| 8 | Fedu | father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education) |
| 9 | Mjob | mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g., administrative or police), 'at home' or 'other') |
| 10 | Fjob | father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g., administrative or police), 'at home' or 'other') |
| 11 | reason | reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other') |
| 12 | guardian | student's guardian (nominal: 'mother', 'father' or 'other') |
| 13 | 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) |
| 14 | studytime | weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours) |
| 15 | failures | number of past class failures |
| 16 | schoolsup | extra educational support (binary: yes or no) |
| 17 | famsup | family educational support (binary: yes or no) |
| 18 | paid | extra paid classes within the course subject (binary: yes or no) |
| 19 | activities | extra-curricular activities (binary: yes or no) |
| 20 | nursery | attended nursery school (binary: yes or no) |
| 21 | higher | wants to take higher education (binary: yes or no) |
| 22 | internet | Internet access at home (binary: yes or no) |
| 23 | romantic | with a romantic relationship (binary: yes or no) |
| 24 | famrel | quality of family relationships (numeric: from 1 - very bad to 5 - excellent) |
| 25 | freetime | free time after school (numeric: from 1 - very low to 5 - very high) |
| 26 | goout | going out with friends (numeric: from 1 - very low to 5 - very high) |
| 27 | Dalc | workday alcohol consumption (numeric: from 1 - very low to 5 - very high) |
| 28 | Walc | weekend alcohol consumption (numeric: from 1 - very low to 5 - very high) |
| 29 | health | current health status (numeric: from 1 - very bad to 5 - very good) |
| 30 | absences | number of school absences (numeric: from 0 to 93) |
| 31 | G1 | first period grade (numeric: from 0 to 20) |
| 32 | G2 | second period grade (numeric: from 0 to 20) |
| 33 | G3 | final grade (numeric: from 0 to 20, output target) |

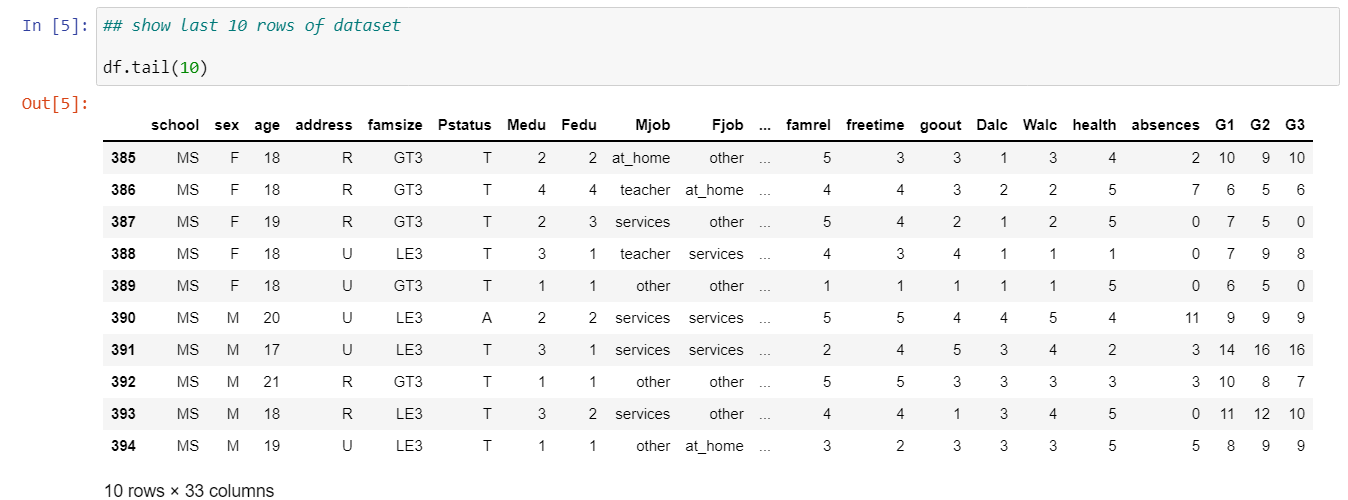
**View of our dataset**



**First 10 rows of dataset**



**Last 10 rows of dataset**

****

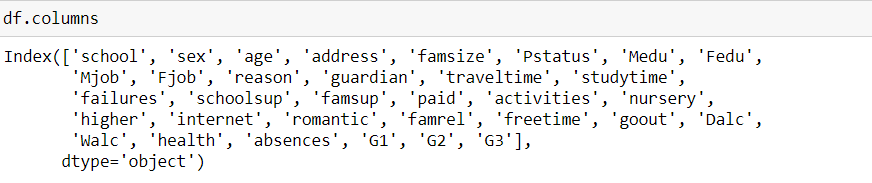
**Number of rows and columns**

|  |  |
| --- | --- |
| Number of rows | 395 |
| Number of columns | 33 |

****

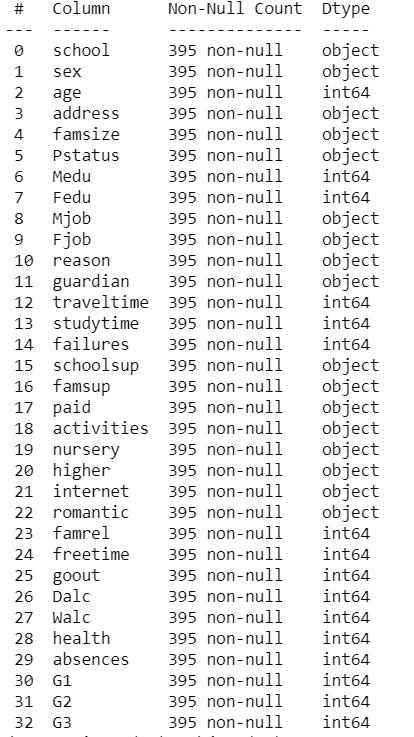
**Description of each attribute**

Column names

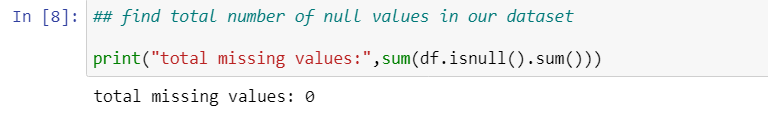


**Information about dataset**

From below output, it is clear that there is no null value in our dataset, and it also tells us about the datatype of each column



**Number of NA values**

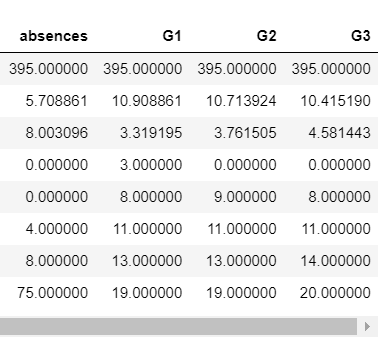


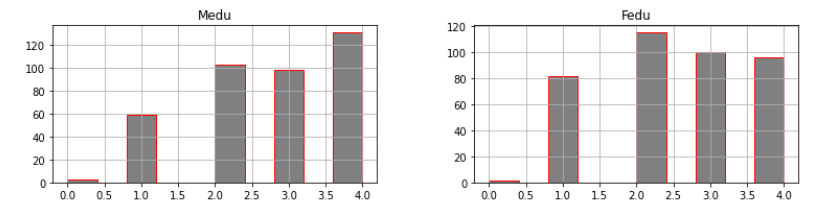
**Summary**

The structure of the data is to calculate basic statistics, such as the min, max, mean, and median, and missing value counts. Here we use summary command to check those statistics. The following shows basic statistics of each feature:

**Table

Description automatically generated**

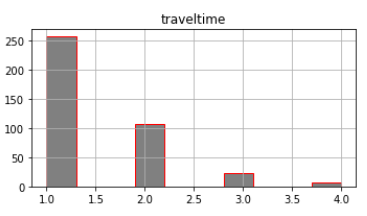
****

**Distribution of Categorical attributes**

**0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)**

The mother education histogram given above represent the axis ,Where X axis shows the level of education ,where Y axis shows the count, where majority of mothers lies within level 4 with the higher education.

The correspond graph against the mother education represent the father education where axis represent basically the level of father’s education but in this graph majority of father edu lies between level 2- 2.5



**Home to school travel time ( 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)**

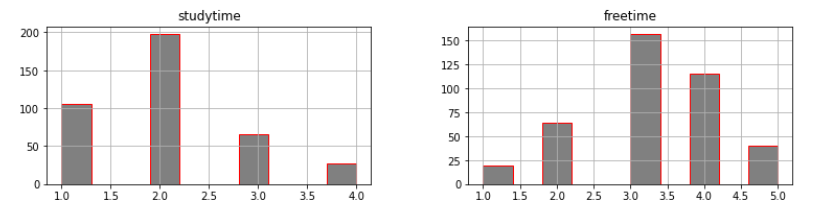
Travel time histogram represent the time taken to travel to school i.e 15mint.

to 1 hours which is represented in levels of 1,2,3,4 in the data set , where incorrospod to this the other axis shows the number of students i.e 50 to more than 240 etc.Chart, bar chart

Description automatically generated

**Going out with friends (from 1 - very low to 5 - very high)**

The go out time represent the information about students who go out with their family and friends. From this graph we can clearly seen that theless than 25 students goes out which are represented by 1.0( that define lowest no of going out) where as more than 40 students goes out very frequntly that lies on 5.0 on the graph (represents students go out at very high rate) and the time lies between 3.0-3.5 for majority of students who goes out not very frequent and not very low.`



**Weekly study time (1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)**

According to study time histogram representation most of the students study for 2 to 5 hours and very less no. of students study for more than 10 hours.

**Free time after school (from 1 - very low to 5 - very high)**

In the free time histogram it displays the data about the free time after studies. More than 150 students get average free time after studies.

Chart, line chart, box and whisker chart

Description automatically generated

**Workday alcohol consumption ( from 1 - very low to 5 - very high)**

**Weekend alcohol consumption ( from 1 - very low to 5 - very high)**

Comparision: On weekdays and weekends, majority of students drink very less alcohol. Very few number of students drink more alcohol on weekdays and weekends.

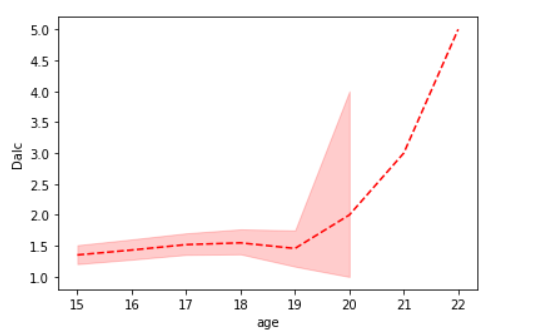
**Exploratory Analysis**

It usually refers to the process of performing initial investigation on data to discover patterns, check assumption with the help of summary statistics and graphical representation.

Now let’s look at graphs:

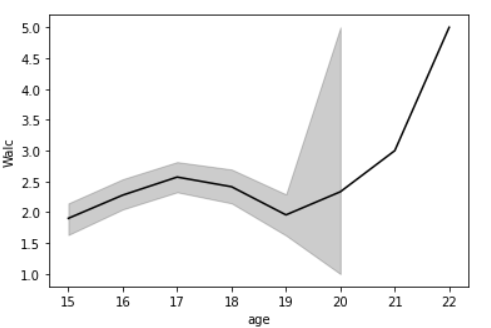
**Dalc Vs Age**

The first graph is for checking that which age group drink alcohol more on weekdays. As, we can clearly see that the age group of 22 drink alcohol more on workdays.

****

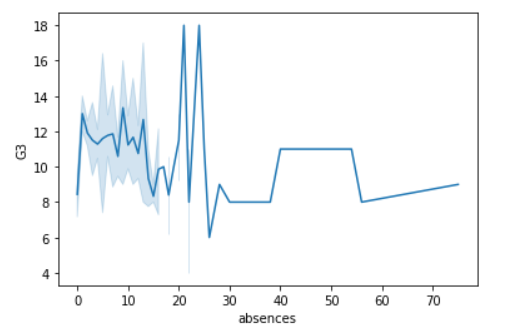
**Walc VS Age**

This graph represents the information about **alcohol consumption on weekends**. In this graph we can see that the alcohol consumption on weekends is less in age group of 15 and higher in age group of 22 on weekends.

****

**G3 Vs Absence**

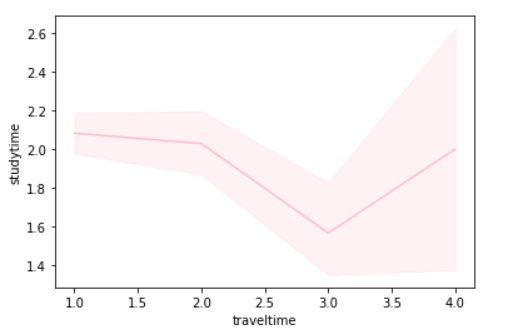
It shows fluctuation in this graph between G3 and absence of the students from the class.intially, when the number of absences is almost 0 graders are 8 but after that it start fluctuations till the count of absences is 25. As the absents from the class increased then the grades started decresed.

****

**Travel time VS Study Time**

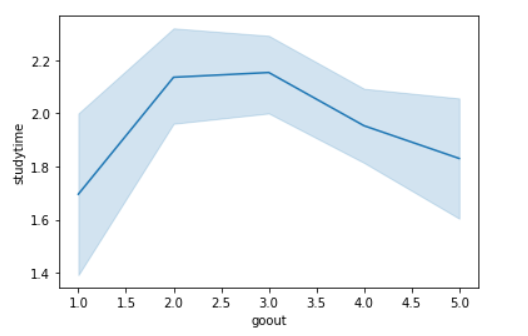
This graph is for checking the effect of travel time on study time

It is shown from the line plot that initially student spend less time in travelling and more in study on an average over the week. we can see that there is a drastic drop in the study time further to 3 hrs of travel time. so, the reason can be effective time management by students.

****

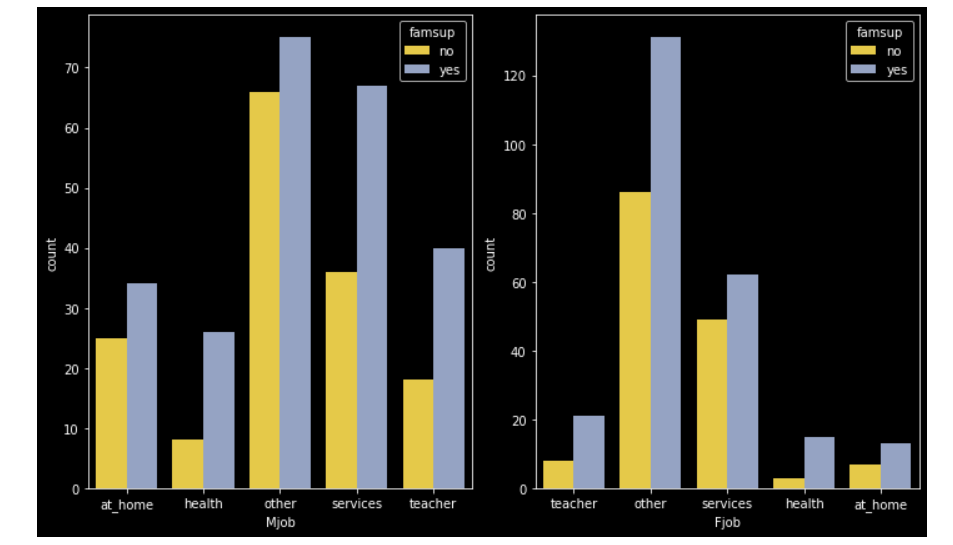
**Goout Vs Study time**

Initially when student go less outside, study time has seen a great increment but as the out time increases study time has decreased from 2.1 to 1.8(approximately)

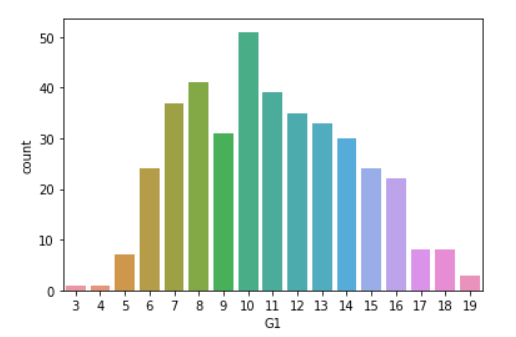
****

**Mother job with family support and Father job with family support**

Here we plot 2 bar graphs one for mother job with family support and another one is father job with family support to show that how much a student receives family support from parents’ side

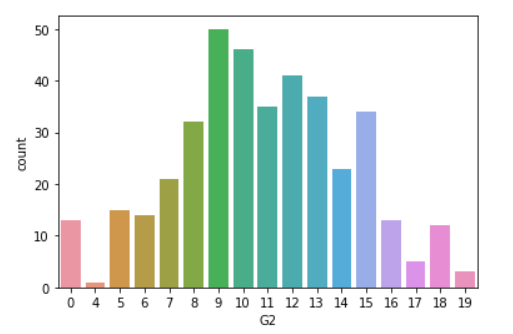
****

**In graph given below we can see that 50 students have 10 grades in term 1**

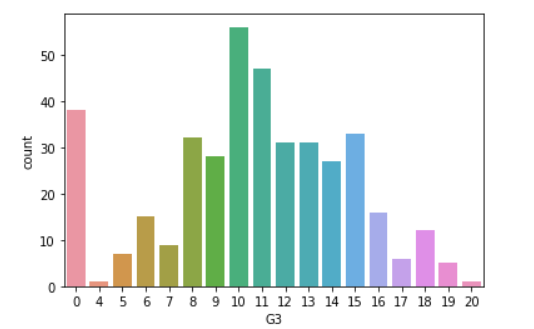
****

.

**In term 2, majority of students get 9 grades, and few students get 19 grades**

****

**Here in this graph we can clearly see that in grade 3,very less students get full grades.**

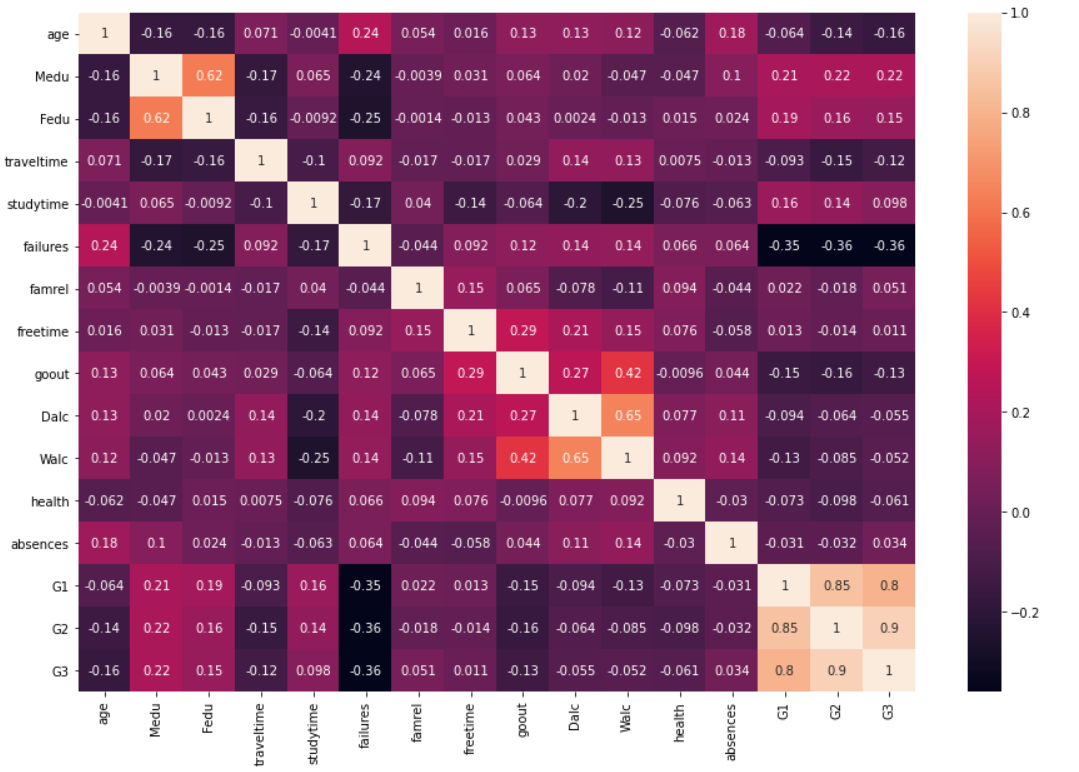
****

**Correlation of numerical values**

For numeric data, correlations are important to help determine that which attribute can be deleted from the dataset because if two variables are highly correlated, they are providing same information.

In this case, we see that the grades G1 and G2 are highly correlated, meaning the higher the grades in one session, the higher the grades in another session.

Now we must check the values of standard deviation of G1 and G2 attributes. The attribute with low standard deviation we can delete that one.



**Best fit line:**

The regression line is the “best fit” straight line. It is possible to draw when the data is obviously non- linear. As the below graph, we can see that this graph is about G2 and G3. In this graph, some of the points are not in line. These are outliers. Then we predict point on best fit line. The distance between the actual point and predicted points are our residuals. If the residuals are minimum it means our error chance is minimum.

Chart, scatter chart

Description automatically generated

**Splitting the dataset:**

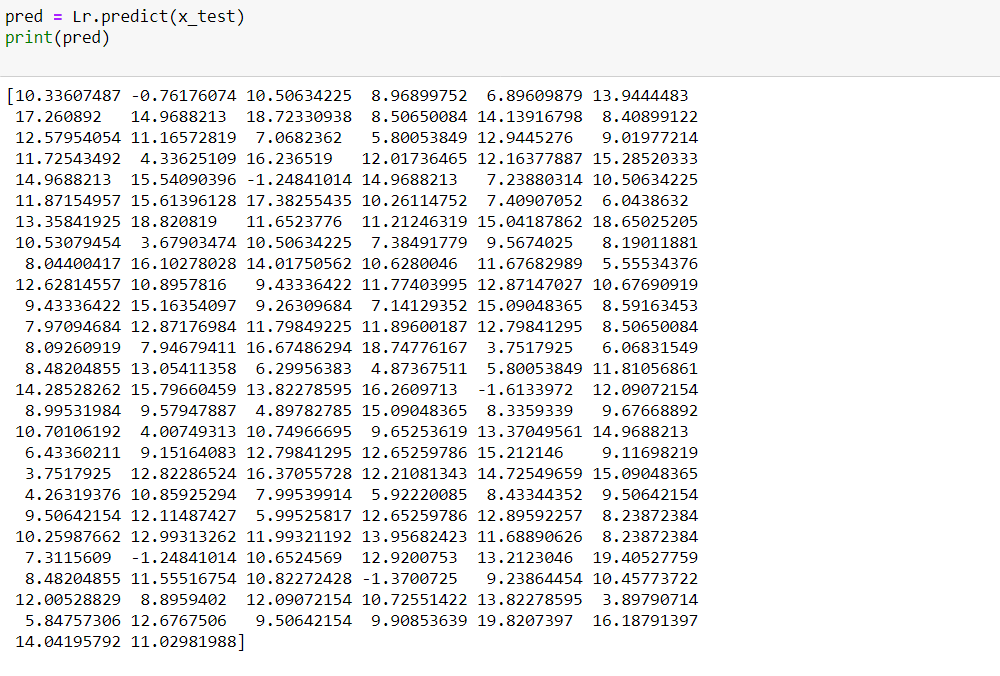
Here we Split the data in train and test set. 40% data in test set and 60% data in the train set. And set random state 40.

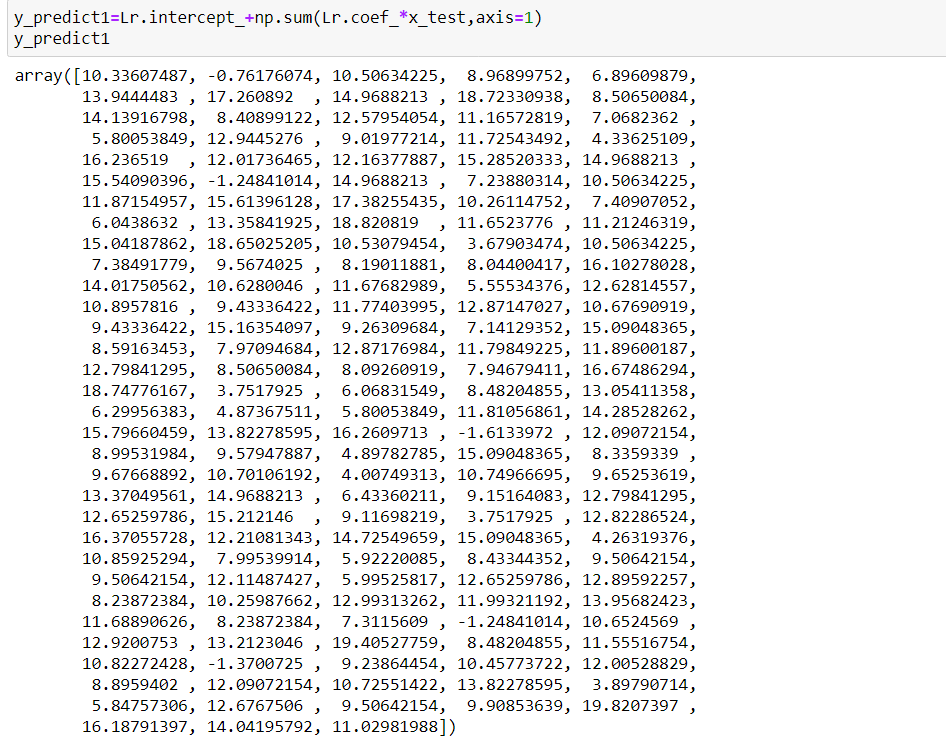
Text, application

Description automatically generated

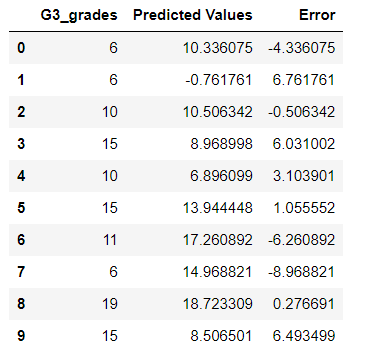
**Linear Model**

|  |  |
| --- | --- |
| **Intercept** | **-2.1000466086273537** |
| **Coefficient** | 0.1216623520737467 |
| **Accuracy** | 0.8244277193293813 |

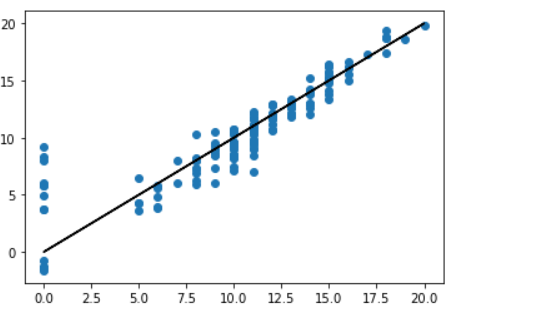
**s**

****

**error\_lr = preco\_lr-predict\_lr**

****

**Scatter plot of**

****

|  |  |
| --- | --- |
| **training accuracy of Linear regression** | **0.8244277193293813** |
| **testing accuracy of Linear regression** | **0.8272669754231312** |

**GitHub -** [Uditaror/Capstone-Project (github.com)](https://github.com/Uditaror/Capstone-Project)

**Reference: -** [**https://www.kaggle.com/uciml/student-alcohol-consumption**](https://www.kaggle.com/uciml/student-alcohol-consumption)