VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Belagavi-590014, Karnataka

MINI PROJECT REPORT

OF

"STUDENT GRADE ANALYSIS & PREDICTION"

Submitted by

ABHISHEK S (3GN16CS002)

Under the guidance of

Prof. GURURAJ.S

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GURU NANAK DEV ENGINEERING COLLEGE

BIDAR-585403, KARNATAKA

2019-2020

VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELAGAVI

GURU NANAK DEV ENGINEERING COLLEGE

BIDAR-585401, KARNATAKA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the mini project report entitled "**STUDENT GRADE ANALYSIS & PREDICTION**" is a bonafide work carried out by ABHISHEK s (3GN16CS002) in practical fulfillment for the award of IA marks for **MACHINE LEARNIG (15CS73)** in COMPUTER SCIENCE AND ENGINEERING of the **GURU NANAK DEV ENGINEERING COLLEGE**, **BIDAR** during the year 2019-2020. It is certified that all the corrections/suggestions indicated for the internal assessment have been incorporated in the report. The Mini Project Report has been approved as it satisfies the academic requirements.

Signature of Guide .....

(Prof. GURURAJ.S)

MARKS AWARDED
ABHISHEK S (3GN16CS002) _____

STUDENT GRADE ANALYSIS & PREDICTION

1. Problem Statement

The problem statement can be defined as follows ”Given a dataset containing attribute of 396 Portuguese students where using the features available from dataset and define classification algorithms to identify whether the student performs good in final grade exam, also to evaluate different machine learning models on the dataset.”

2. Description of the Dataset

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school-related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two data sets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1.

This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).

2.1 Attribute Information:

- ✓ 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)

3. Methodology

Since universities are prestigious places of higher education, students' retention in these universities is a matter of high concern. It has been found that most of the students' drop-out from the universities during their first year is due to lack of proper support in undergraduate courses. Due to this reason, the first year of the undergraduate student is referred as a "make or break" year. Without getting any support on the course domain and its complexity, it may demotivate a student and can be the cause to withdraw the course.

There is a great need to develop an appropriate solution to assist students retention at higher education institutions. Early grade prediction is one of the solutions that have a tendency to monitor students' progress in the degree courses at the University and will lead to improving the students' learning process based on predicted grades.

Using machine learning with Educational Data Mining can improve the learning process of students. Different models can be developed to predict students' grades in the enrolled courses, which provide valuable information to facilitate students' retention in those courses. This information can be used to early identify students at-risk based on which a system can 1 suggest the instructors to provide special attention to those students. This information can also help in predicting the students' grades in different courses to monitor their performance in a better way that can enhance the students' retention rate of the universities.

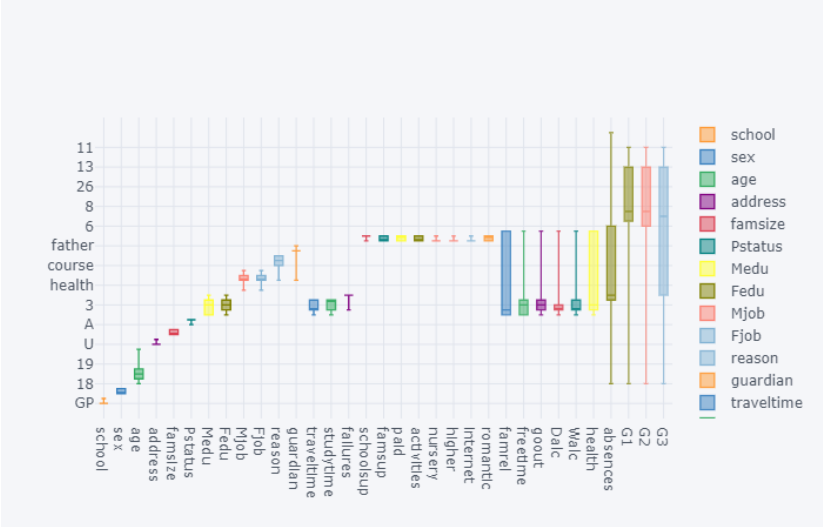
Using various packages such as cufflinks, seaborn & matplotlib to represent the data along with different attributes graphically or pictorially to analyse the dataset for predicting the Final Grade(G3).

4. Experimental Results

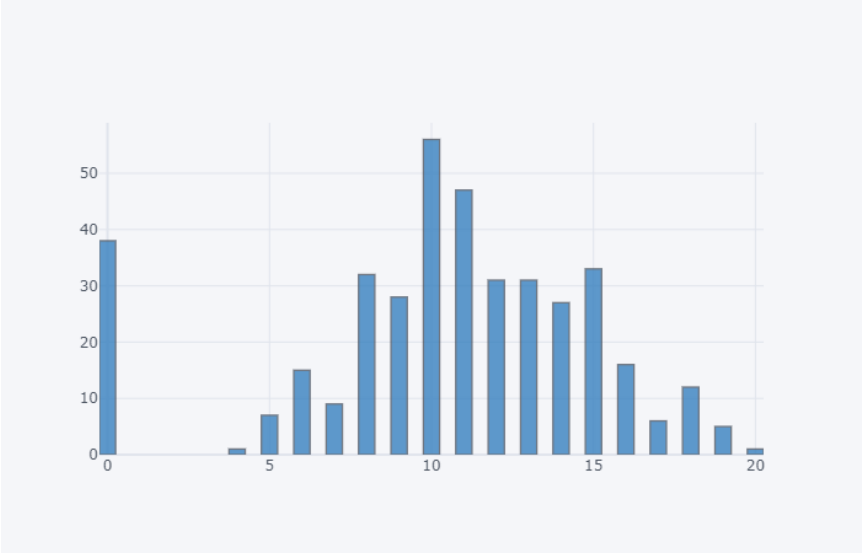
4.1 - KDE Plot to view all attributes using cufflinks

Observation: cufflink connects plotly with pandas to create graphs and charts of dataframes directly.

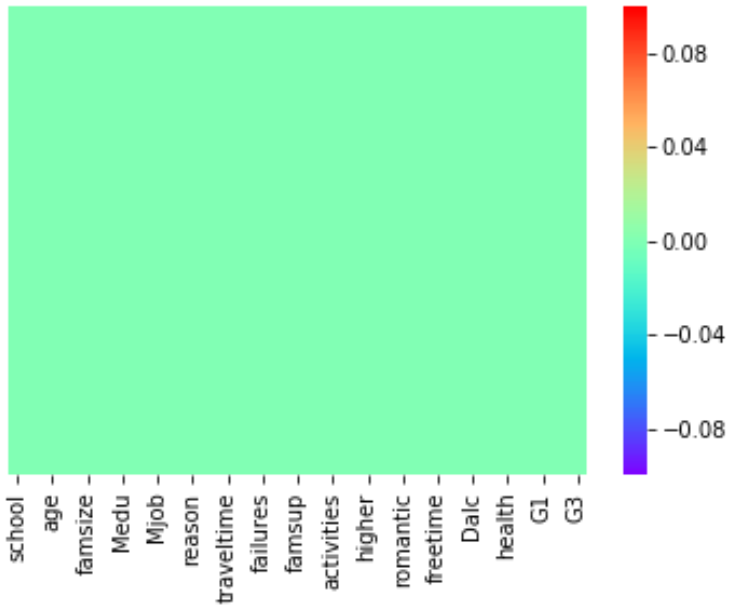
4.2 - Box Plot to view all attributes using cufflinks



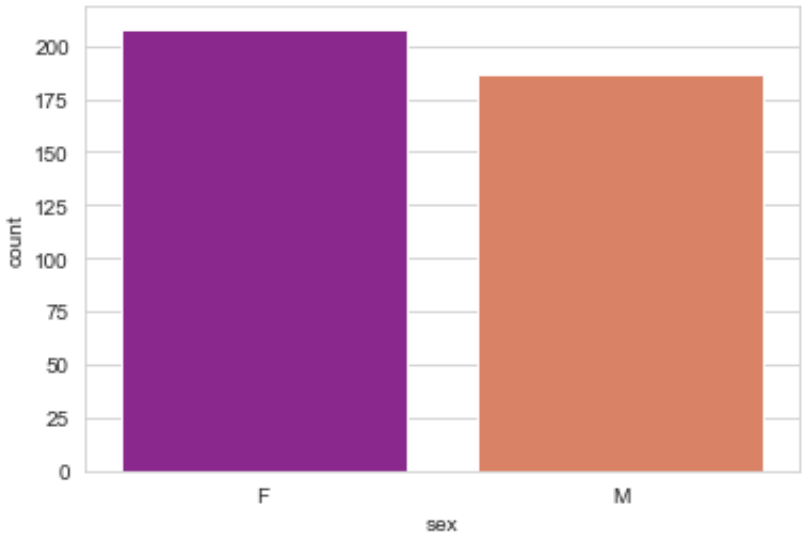
4.3 - Histogram Plot for G3 (Final Grade) using cufflinks



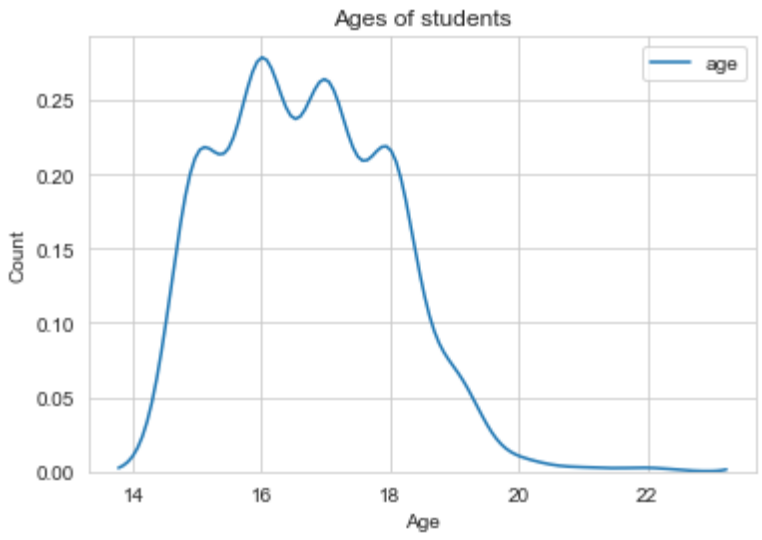
4.4 - Pictorial representation of any null data present in the dataset.



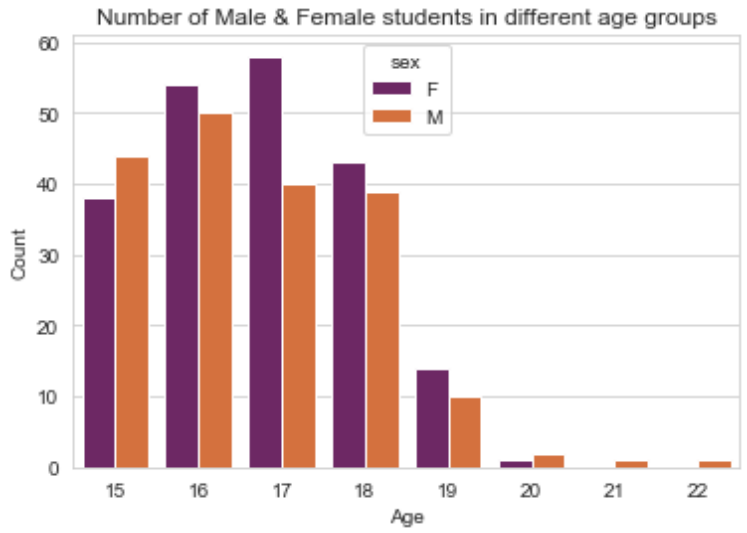
4.5 - Count Plot for Student Sex Attribute



4.6 - Kernel Density Estimation for Age of Students.

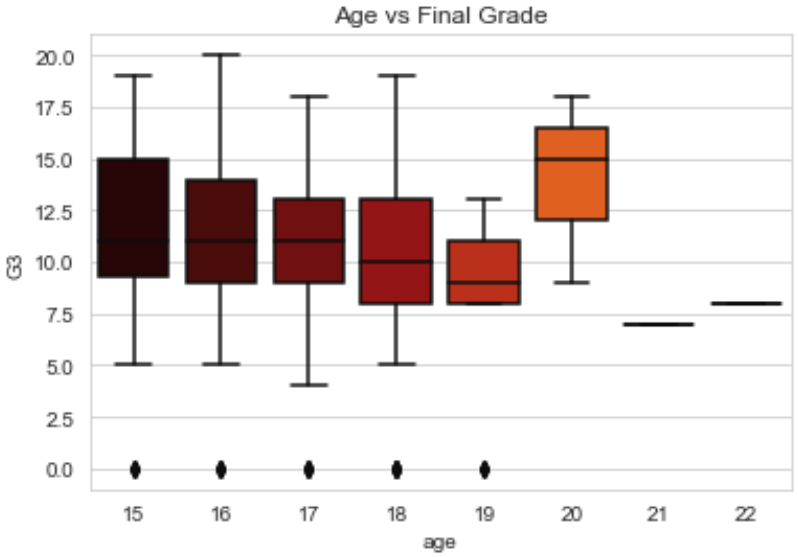


4.7 - Count PLot for Male & Female students in different age groups.



4.8 - Count Plot for students from Urban & Rural Region.

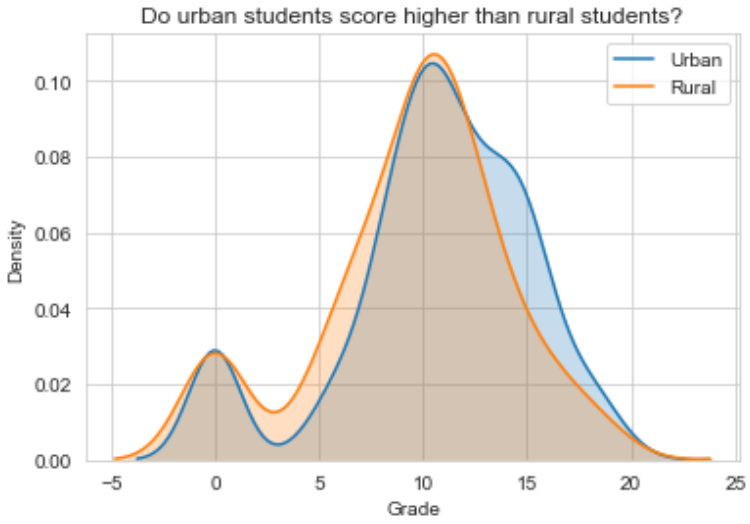
4.9 - Does age affect final grade?



Observation:

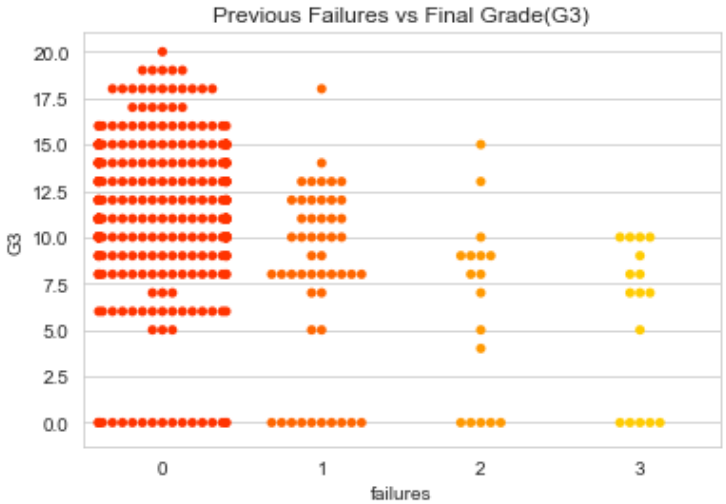
- Plotting the distribution rather than statistics would help us better understand the data.
- The above plot shows that the median grades of the three age groups(15,16,17) are similar. Note the skewness of age group 19. (may be due to sample size). Age group 20 seems to score highest grades among all.

4.10 - Do urban students perform better than rural students?



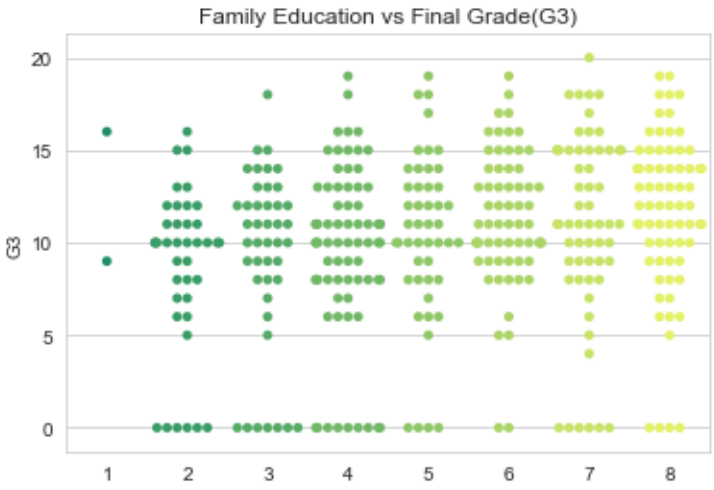
Observation: The above graph clearly shows there is not much difference between the grades based on location.

4.11 - Previous Failures vs Final Grade(G3)



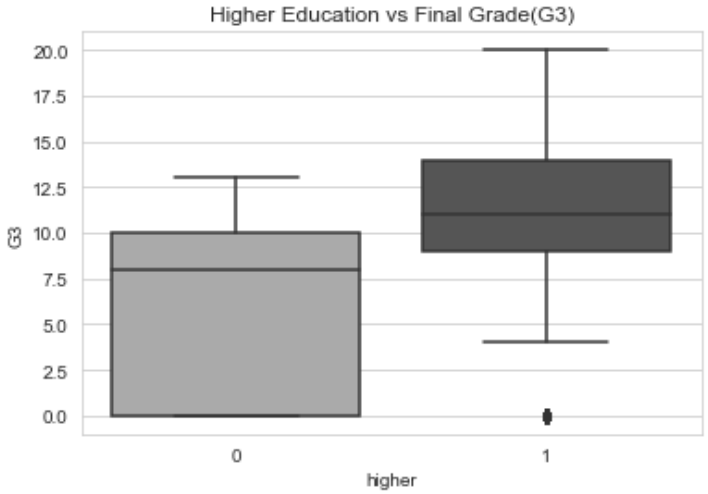
Observation: Student with less previous failures usually score higher.

4.12 - Family Education vs Final Grade(G3)



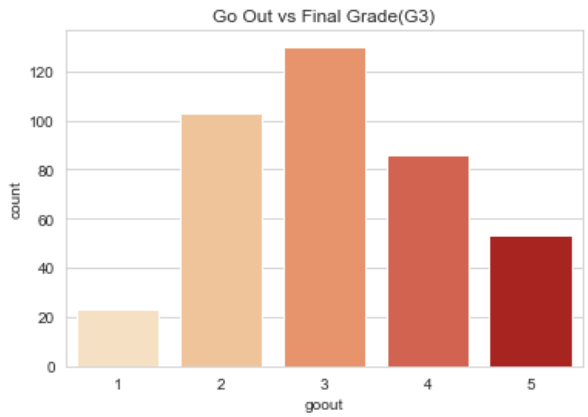
Observation: Educated families result in higher grades

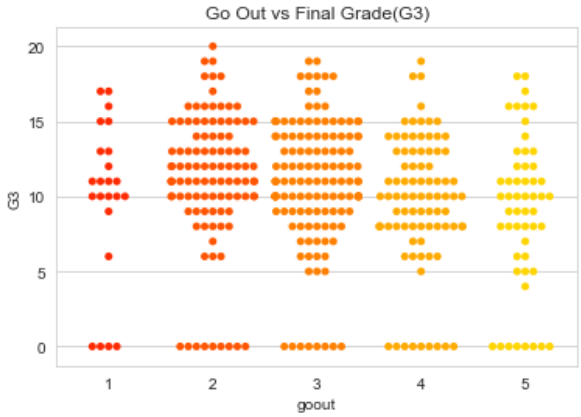
4.13 - Higher Education vs Final Grade(G3)



Observation: Students who wish to go for higher studies score more.

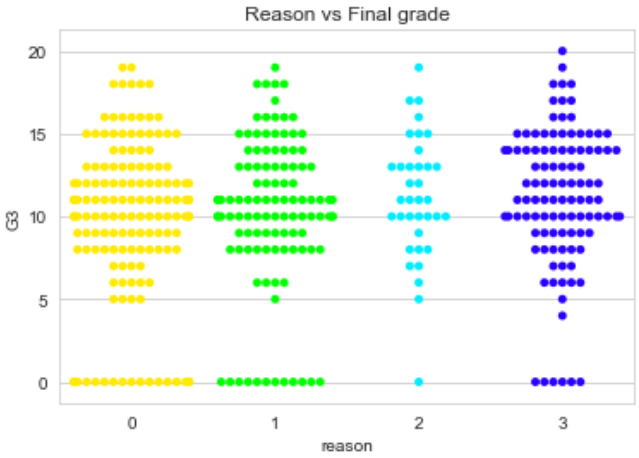
4.14 - Go Out vs Final Grade(G3)





Observation: The students have an average score when it comes to going out with friends & Students who go out a lot score less.

4.15 - Reason vs Students Count



Observation : The students have an equally distributed average score when it comes to reason attribute.

5. Conclusion

As we see both MAE & Model RMSE that the Linear Regression is performing the best in both cases.