MULTIPLE LINEAR REGRESSION TO PREDICT STUDENT GRADES

GROUP – 9

INTRODUCTION

Data-driven strategies for enhancing educational outcomes have drawn more and more attention in recent years. Predicting student grades using a variety of factors, such as prior academic performance, socioeconomic background, attendance, and demographics, is one such method. It is possible to pinpoint the variables that affect student performance using data analysis and machine learning algorithms, and you can use this knowledge to give students who need it targeted support and intervention. Observational studies are research studies that involve observing and collecting data on individuals or groups without manipulating any variables. In this case, the dataset includes information on students' demographic characteristics, socio-economic status, previous academic performance, and attendance records. This data is then used to predict the students' final grades, which serves as the dependent variable in the model. Therefore, this is an **Observational study**.

Since the data is collected from existing records and is not manipulated or controlled by the researcher, it is more likely that this is an observational study. However, without more information about the study design and data collection methods, it is difficult to say for certain.

The goal of this project is to use multiple linear regression in R programming to create a model that can forecast student grades based on a variety of attributes. The relationship between a dependent variable (in this case, student grades) and several independent variables can be examined using the statistical technique known as multiple linear regression. The model is trained and evaluated using a dataset of student data and academic performance metrics. The dataset contains details on the demographics, socioeconomic standing, prior academic success, and attendance history of the students.

The final grades of the students, which are the dependent variable in the model, are predicted using this data. To deal with missing values, outliers, and other anomalies that might impair the model's accuracy, the dataset is first pre-processed. The training set is used to train the model, and the testing set is used to assess the model's performance. The data is then divided into training and testing sets. The training data is then subjected to multiple linear regression, and the estimated coefficients for each independent variable are calculated. The magnitude of each independent variable's influence on the dependent variable (i.e., student grades) is shown by these coefficients.

The model's performance is then measured on the testing set using a variety of evaluation metrics, including mean square error, R-squared, and adjusted R-squared. The project investigates each independent variable's significance in predicting student grades. By providing insights into the elements that educators should emphasize to improve student outcomes, this information can assist in identifying the characteristics that are most crucial in determining student performance.

The project's findings show that using multiple linear regression to predict student grades based on a variety of non-academic factors can be successful. The model successfully predicted

student grades with a high degree of accuracy and discovered a number of factors that have a big impact on how well students perform. The project may offer educators a practical tool for identifying students who may require additional help and intervention to enhance their academic performance.

Overall, this project demonstrates how data analysis and machine learning algorithms have the potential to enhance educational outcomes. Educators can identify the variables that affect student outcomes and create targeted interventions to support students who need it by using data-driven approaches to predict student performance.

DATA

The dataset extracted from Kaggle, contains information about 400 students and includes 33 attributes such as age, gender, family size, study time, and weekly alcohol consumption. Additionally, the dataset includes information about student performance in two subjects, math, and Portuguese language. The math performance is evaluated on a scale of 0-20, and the Portuguese language performance is evaluated on a scale of 0-19.

The dataset provides a unique opportunity to analyse the relationship between various attributes and student performance. For instance, it is possible to examine the relationship between study time and student performance. Similarly, it is possible to analyse the impact of family size and weekly alcohol consumption on student performance. The description of each attribute follows:

- 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)

The attributes like Mother's education, studytime, higher, failure are the explanatory variables to predict the final grade 'G_agg' (aggregate of G1, G2, G3) which is the target variable. The following is the summary of the data which explains the mean, median, min value, max value of each attribute.

```
> summary(data)
    school              sex                 age          address             famsize
 Length:395         Length:395         Min.   :15.0   Length:395         Length:395
 Class :character   Class :character   1st Qu.:16.0   Class :character   Class :character
 Mode  :character   Mode  :character   Median :17.0   Mode  :character   Mode  :character
                                       Mean   :16.7
                                       3rd Qu.:18.0
                                       Max.   :22.0
   Pstatus              Medu            Fedu            Mjob              Fjob              reason
 Length:395         Min.   :0.000   Min.   :0.000   Length:395         Length:395         Length:395
 Class :character   1st Qu.:2.000   1st Qu.:2.000   Class :character   Class :character   Class :character
 Mode  :character   Median :3.000   Median :2.000   Mode  :character   Mode  :character   Mode  :character
                    Mean   :2.749   Mean   :2.522
                    3rd Qu.:4.000   3rd Qu.:3.000
                    Max.   :4.000   Max.   :4.000
   guardian          traveltime      studytime        failures       schoolsup           famsup
 Length:395         Min.   :1.000   Min.   :1.000   Min.   :0.0000   Length:395         Length:395
 Class :character   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   Class :character   Class :character
 Mode  :character   Median :1.000   Median :2.000   Median :0.0000   Mode  :character   Mode  :character
                    Mean   :1.448   Mean   :2.035   Mean   :0.3342
                    3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
                    Max.   :4.000   Max.   :4.000   Max.   :3.0000
    paid             activities          nursery            higher            internet
 Length:395         Length:395         Length:395         Length:395         Length:395
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character


  romantic            famrel          freetime        goout            Dalc            Walc
 Length:395         Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 Class :character   1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000
 Mode  :character   Median :4.000   Median :3.000   Median :3.000   Median :1.000   Median :2.000
                    Mean   :3.944   Mean   :3.235   Mean   :3.109   Mean   :1.481   Mean   :2.291
                    3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000
                    Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
    health         absences            G1              G2              G3
 Min.   :1.000   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00   Min.   : 0.00
 1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00
 Median :4.000   Median : 4.000   Median :11.00   Median :11.00   Median :11.00
 Mean   :3.554   Mean   : 5.709   Mean   :10.91   Mean   :10.71   Mean   :10.42
 3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00
 Max.   :5.000   Max.   :75.000   Max.   :19.00   Max.   :19.00   Max.   :20.00
```

Fig: 1

EXPLORATORY DATA ANALYSIS

The following is the distribution plot of all the attributes present in the dataset. Before developing a predictive model to estimate student grades based on various attributes, we first conducted an exploratory data analysis (EDA) to gain a better understanding of the dataset and identify any patterns or trends in the data.

The dataset used in this project contains information on students' demographic characteristics, socio-economic status, previous academic performance, and attendance records.

The EDA helped us gain a better understanding of the dataset and identified some of the key relationships between the independent variables and the dependent variable. It also allowed us to identify and handle any missing values, outliers, and multicollinearity issues, ensuring that the dataset was suitable for building a predictive model using multiple linear regression.

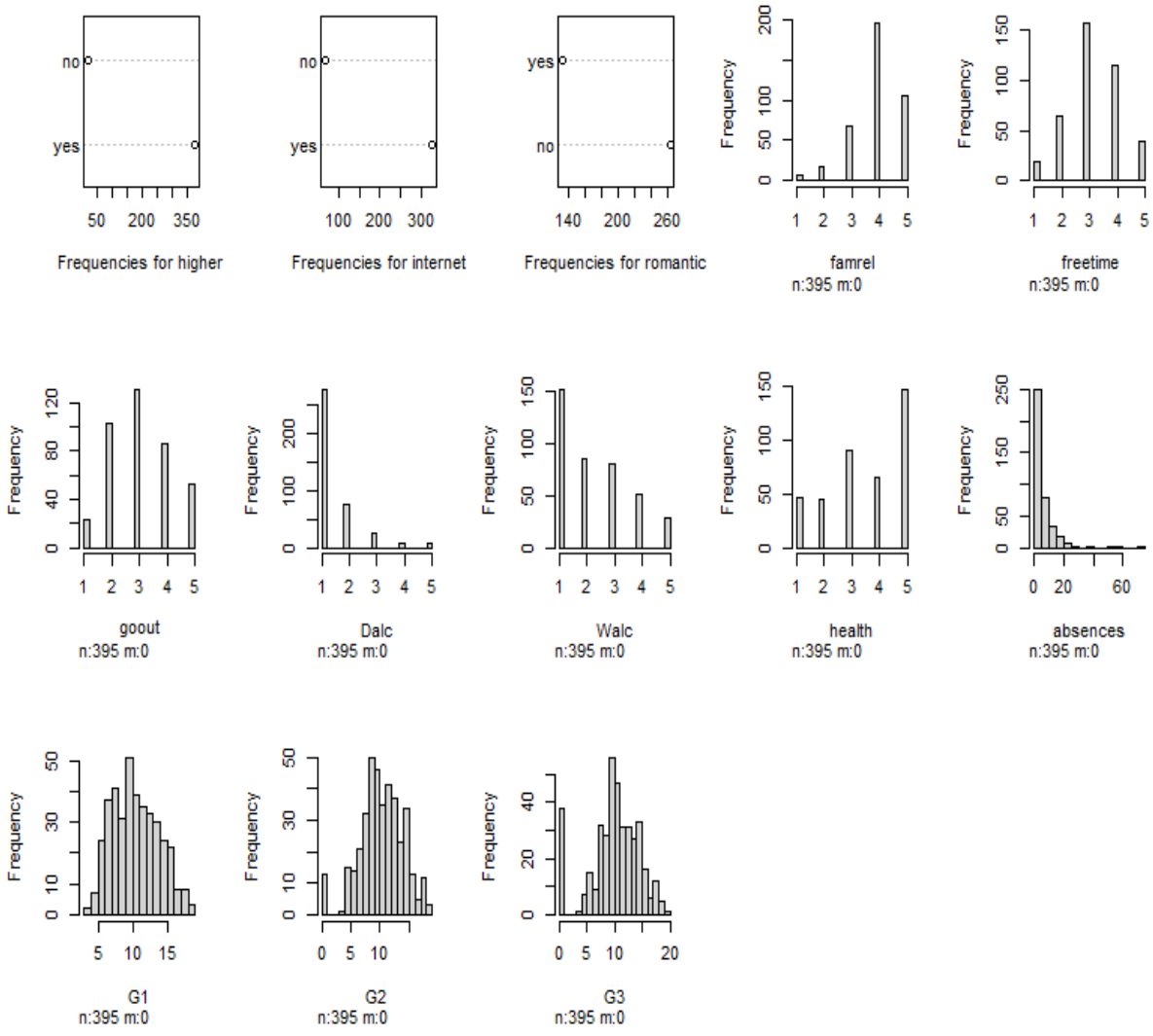

Fig: 2

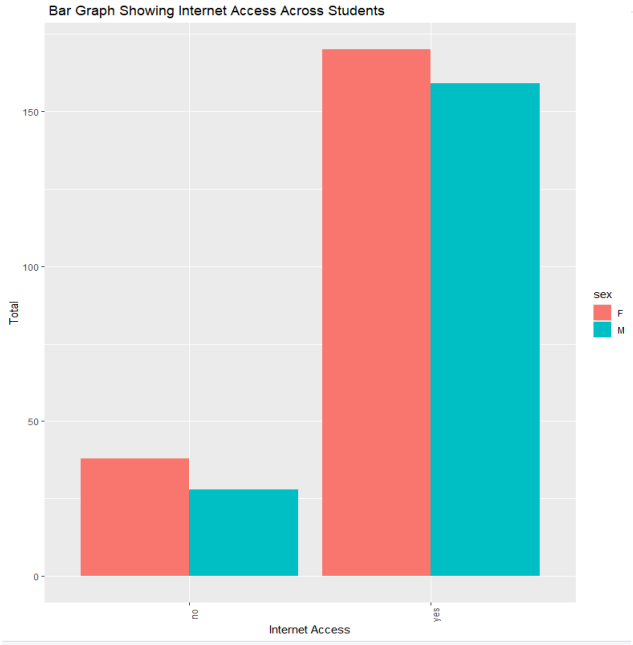

Fig: 3

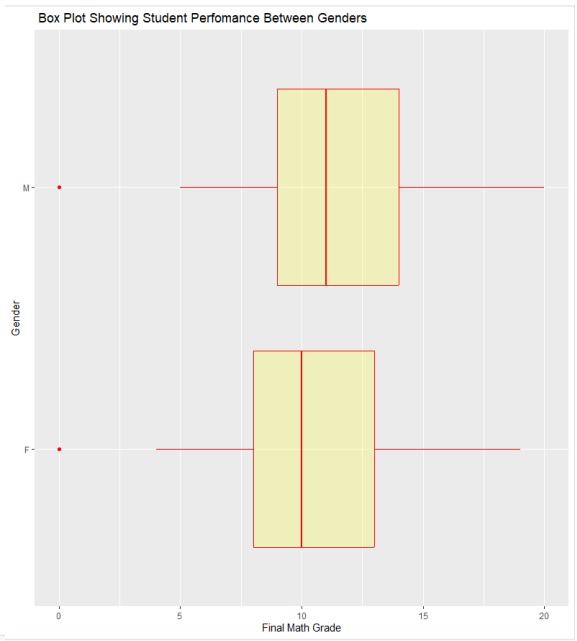

Fig: 4

The above graph shows the distribution of male and female against the attribute internet and final grades. We can start by making a frequency table that displays the proportion of male and female students who have internet access as well as their final grades in order to analyse the relationship between gender and the attributes "internet" and "final grades". As a result, we will be better able to spot any patterns or trends in the data.

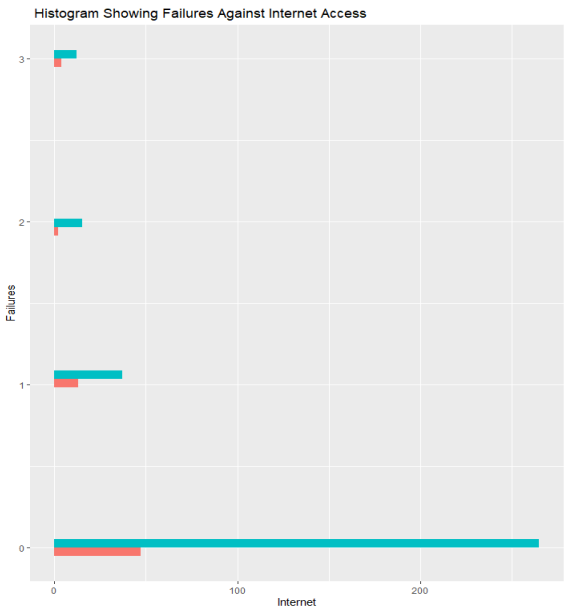

Fig: 5

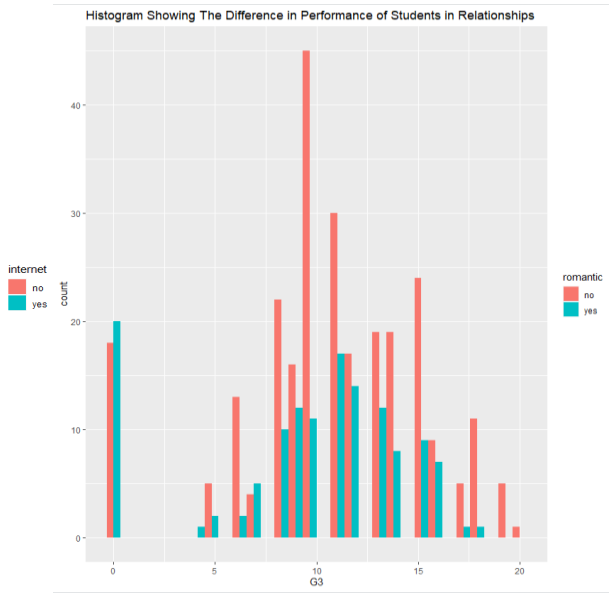

Fig: 6

From this histogram (Fig: 5), we can see that the majority of students who have internet access did not have any failures, with a smaller number of students having one or two failures. In contrast, students who do not have internet access had a higher frequency of failures, with a significant number of students having four or more failures.

From Fig: 6, we can see that students who are not in a relationship had a higher frequency of passing grades, with a smaller number of students having failing grades. In contrast, students who are in a relationship had a higher frequency of failing grades, with a smaller number of students having passing grades.

METHODS

The following steps would be included in the methodology for examining the Student Alcohol Consumption dataset and creating a model to forecast student performance:

1. **Data Pre-processing:** For the data pre-processing, we have identified that the data set has few categorical variables. To deal with such data we tried to convert the attributes to numerical values in order to fit the model accurately. We have also checked for any missing values present in the data set and found no missing values. In the data set, we had three different columns for grades of students, by which we aggregated the three columns in to one and created a column 'G_agg' which is final grades of students.
2. **Feature Selection:** The next step would be to select the most relevant features that impact student performance. This can be done using techniques such as correlation analysis. The below is the correlation matrix of the data after pre-processing and we have found the attributes such as Mjob, Fjob, Guardian, reason as less significant features. We have also interpreted that attributes like Medu, studytime, higher, failure have more correlation with the target variable 'G_agg' which can be firther used to perform Multiple linear regression.

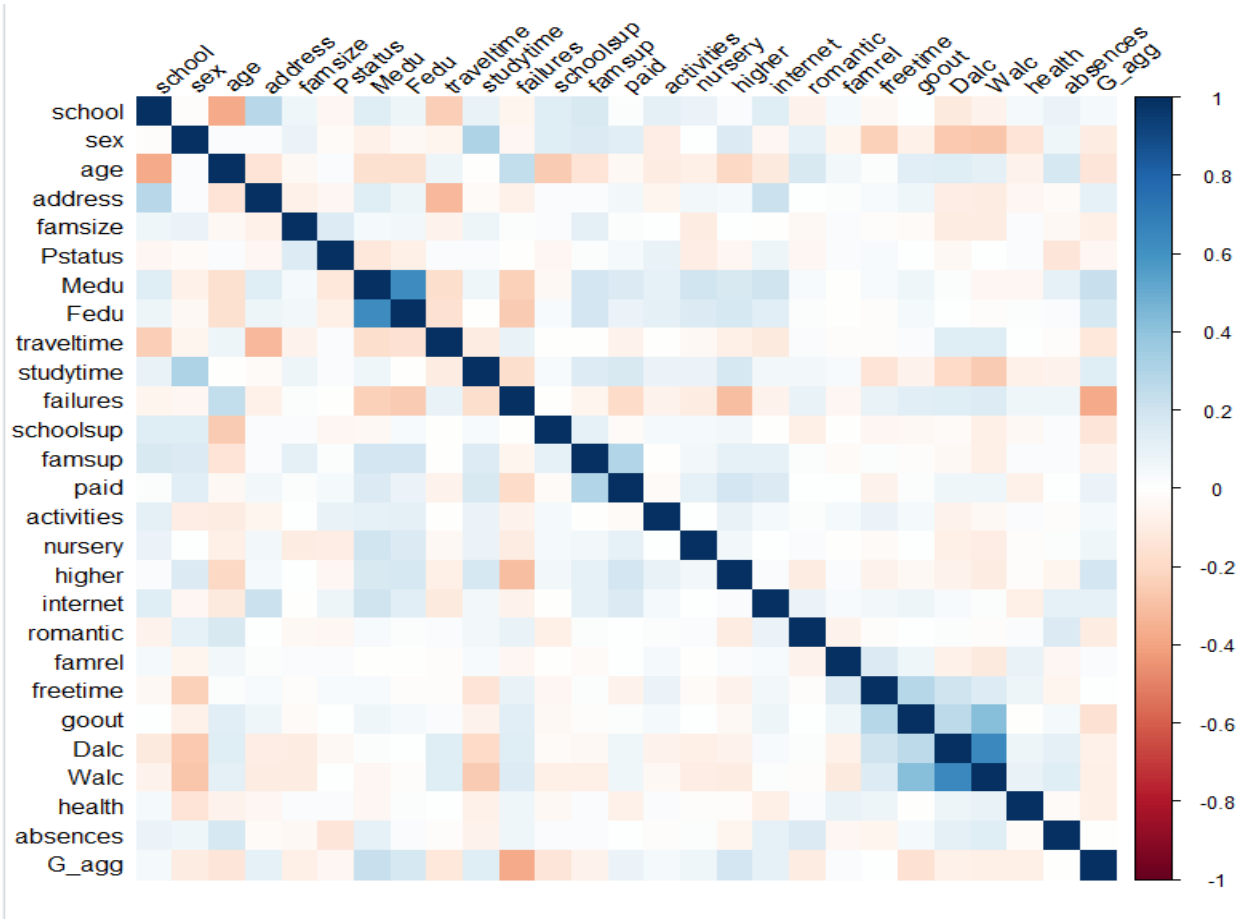

Fig: 7

3. **Model Selection:** A suitable algorithm can be chosen once the pertinent features have been located. In this situation, a model that predicts student performance based on a variety of characteristics would be created using multiple linear regression.

4. **Model Training:** The dataset can be split into training (80%) and testing sets (20%), with the training set being used to train the model. Several metrics, including mean squared error, R-squared, and root mean squared error, can be used to assess the model's performance. The Fig: 8 shows the summary of the model

```
Call:
lm(formula = G_agg ~ Medu + studytime + higher + failures, data = training)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4142 -2.1283 -0.1083  2.5294  8.4227

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.0252     1.1419   7.904 4.73e-14 ***
Medu          0.2936     0.1832   1.603    0.110
studytime     0.3643     0.2290   1.591    0.113
higher        0.8502     1.0022   0.848    0.397
failures     -1.8368     0.2929  -6.272 1.19e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.346 on 311 degrees of freedom
Multiple R-squared:  0.1753,    Adjusted R-squared:  0.1647
F-statistic: 16.53 on 4 and 311 DF,  p-value: 2.721e-12
```

Fig: 8

The coefficients table's p-values illustrate the significance of each predictor variable. The predictor variable is significant at the 95% confidence level if the p-value is less than 0.05. In this instance, higher is not significant, but Medu, studytime, and failures are all significant predictors of student performance.

When the number of predictors in the model is taken into account, the **Adjusted R-squared** value (0.1647) shows the percentage of variance in the dependent variable that is explained by the independent variables in the model. The model in this instance accounts for about 16.5% of the variation in G_agg.

The **F-statistic** (16.53) and corresponding **p-value** (2.721e-12) show the significance of the model as a whole. The F-statistic in this instance is significant with a very low p-value, showing that the model as a whole, fits the data well.

The distribution of the model errors, or residuals, is described in the Residuals section. The residuals should be normally distributed with a constant variance and a mean of zero. The median is near zero in this instance, which is a positive indication, and both the minimum and maximum values fall within an acceptable range.

The average difference between the actual and predicted G_agg scores is shown by the **residual standard error** (3.346). A lower value indicates that the model fits the data more accurately.

```
> anova(model)
Analysis of Variance Table

Response: G_agg
           Df Sum Sq Mean Sq F value    Pr(>F)
Medu        1  148.2  148.25 13.2435 0.0003202 ***
studytime   1   82.1   82.14  7.3382 0.0071246 **
higher      1   69.3   69.35  6.1950 0.0133339 *
failures    1  440.3  440.29 39.3329 1.194e-09 ***
Residuals 311 3481.3   11.19
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig: 9

From the ANOVA test (Fig: 9) we observed that the p-value for the **F-statistic** is less than 0.05 in this instance, and the ANOVA table shows that the model is statistically significant. This suggests that the response variable (G_agg) and at least one of the predictor variables in the model are significantly correlated.

The statistical significance of each predictor variable in explaining the variation in the response variable is shown by the p-values for each predictor variable. With the exception of the "higher" variable, which has a **p-value** of 0.013, all the predictor variables in this case (Medu, studytime, higher, and failures) are statistically significant at the 0.05 level. This suggests that, in comparison to the other predictor variables, the "higher" variable is less significant in explaining the variation in G_agg.

RESULTS AND ANALYSIS

We can also determine which variables are more likely to be weakly or not at all associated with the response variable by looking at the scatter plot matrix. This knowledge can help us decide which variables to include in a predictive model and where additional research might be useful.

The below **scatter plot matrix** (Fig: 10) is a helpful tool for examining the connections between various variables in a dataset and can offer insightful information about the data structure and potential patterns.

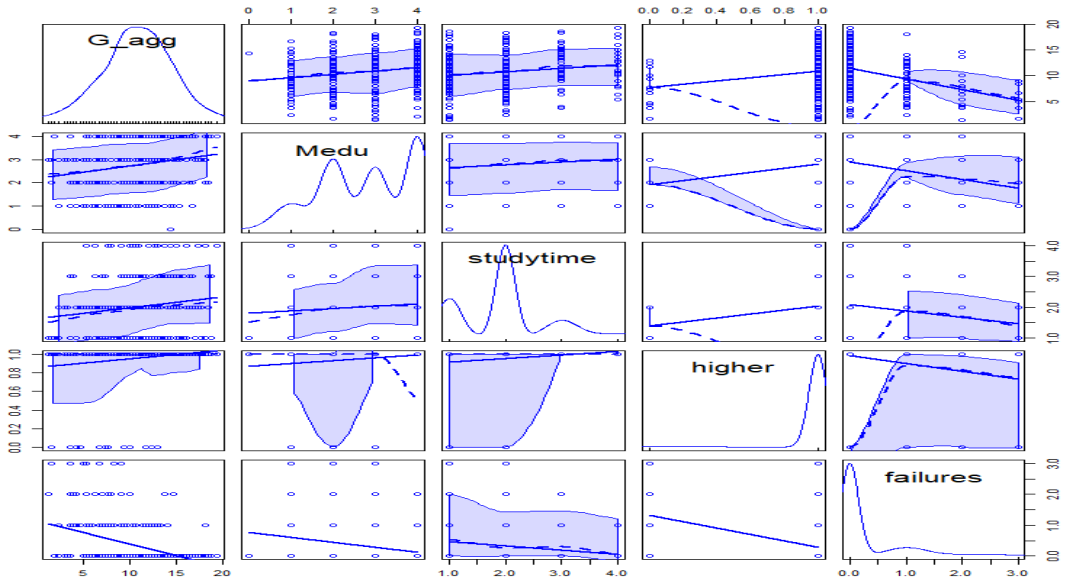

Fig: 10

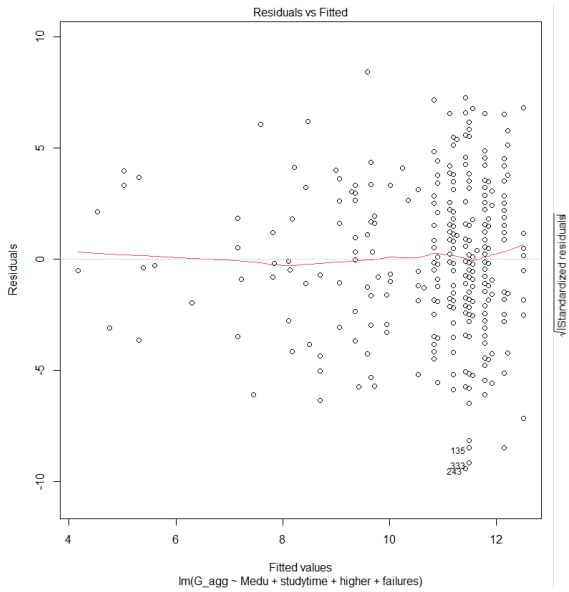

Fig: 11

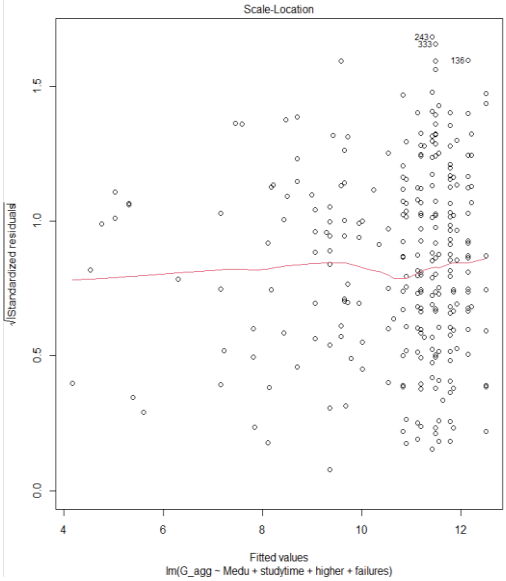

Fig: 12

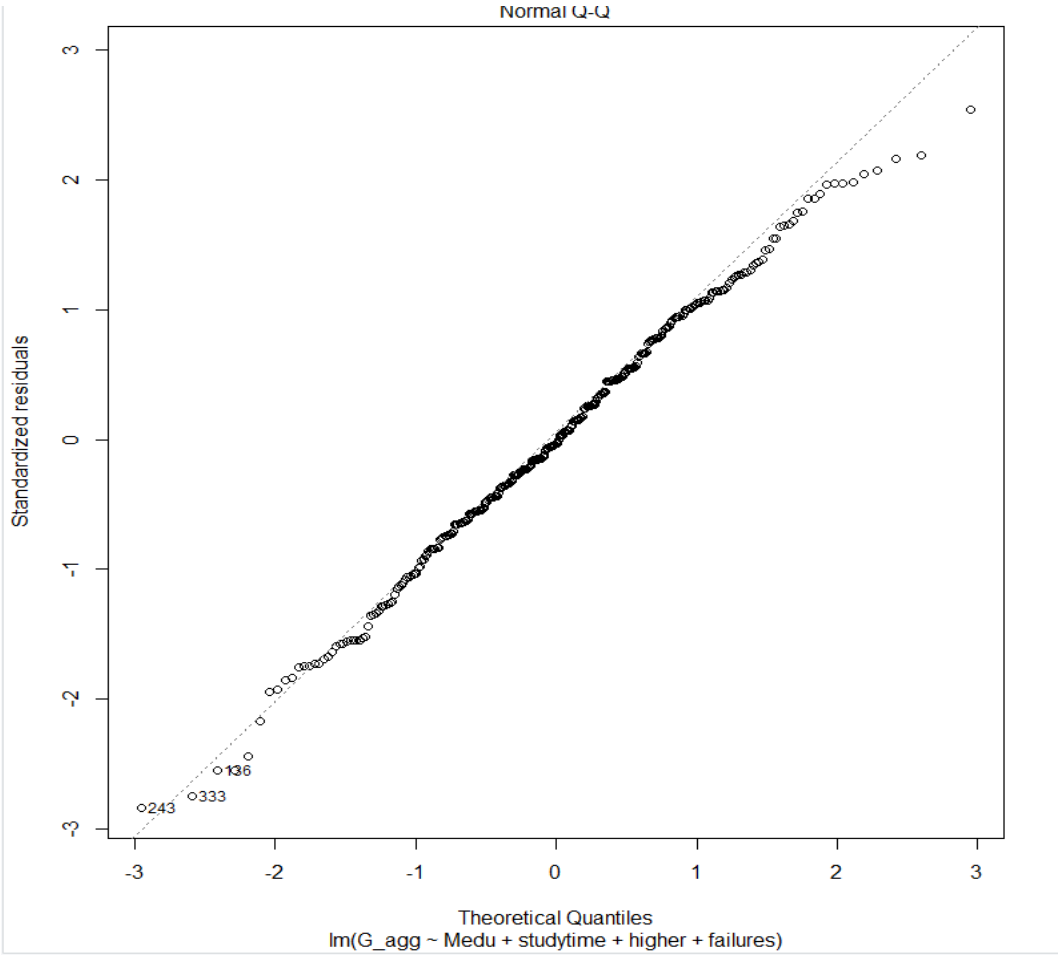

Fig: 13

Fig: 13 shows the **QQ Plot,** from which we can see that most of the data points lie on the line which shows the linearity of the data and also very few outliers near the tails of the plot. Fig: 11 is the **Fitted vs residual plot** in which we can observe that there is very slight curve on the line and no wider distribution of the data points which shows that the model is correct fit for the data.

```
> shapiro.test(training$G_agg)

        Shapiro-Wilk normality test

data:  training$G_agg
W = 0.99326, p-value = 0.1688
```

Fig: 14

To determine whether a sample of data is normally distributed or not, the **Shapiro-Wilk normality test** is employed. In this instance (Fig: 14), the "G_agg" variable from the "training" dataset was subjected to the test.

A test statistic (W) of 0.99326 and a p-value of 0.1688 are displayed in the test results. The fact that the p-value is higher than the usual significance threshold of 0.05 shows that the null hypothesis that the data is normally distributed cannot be ruled out. So, it is reasonable to assume that the "G_agg" variable has a normal distribution.

```
> #RESULT
> test_result<-predict(model,newData = testing)
> test_result
      179       14      195      306      118      299      229      244      394      374      153       90       91
11.414163 11.778493 10.826978  9.354492 11.120570 12.507154 11.191309 11.414163 11.120570 10.533386  8.175598 11.778493 11.849231
      256      197      383      348      137      355      328       26        7      378      254      211       78
 8.696569 11.414163 11.191309 12.142824 11.484901 11.778493 10.826978  7.153344 11.191309 11.778493 10.826978 12.213562 11.919970
       81       43      359      332      143       32      109      263      330       23      309      135      367
10.826978 11.778493 10.533386 11.555639 12.142824 11.778493 12.507154 11.849231 12.142824 11.778493  9.648084 11.484901 12.142824
      224      166      217      290       69       72       76       63      141      210      385      294      277
11.191309  8.433531  8.104859 11.778493 11.191309 12.507154 11.778493 10.897717 12.507154 12.142824  9.577346 12.213562 10.634679
       41      384      316      223       16      116       94      262      235       86       39      159      240
 9.354492  8.696569  9.718822 11.191309 11.414163 11.778493 11.778493 11.778493 10.897717  8.104859 11.849231 10.826978  8.504270
      209      392       34        4       13      352      243      308      278       89       25      291      286
10.533386 11.120570 11.484901 12.142824 11.414163 11.484901 11.414163  9.577346 11.414163  9.354492 11.555639 11.778493 10.897717
      366      121      110      158       64      199       67      151       85      165      136       51       74
10.897717 10.897717 12.142824  5.022935 12.142824  9.577346 12.507154  4.172713 10.897717  4.537044 12.142824 11.191309 11.120570
      178      236       98      214      127      212      174      273      232      324      280      113      107
11.484901 11.849231 11.191309  9.354492 11.484901 11.778493  5.387266 10.897717 11.191309 11.849231 11.414163  9.354492 11.919970
      310      154      102      255      160      155        5      272      345      344       55      238      252
 9.060900  4.759897 12.142824 10.826978  9.648084 11.414163 11.484901 11.919970 11.555639  9.354492 11.120570 10.826978 11.484901
      333      373      226       48       77       83      184      322      196      257      168      337       20
11.484901 11.555639  9.648084 12.507154 12.507154 11.484901 11.484901 11.191309 11.191309 12.507154 11.778493 10.012414 11.414163
      393      164       52       22      177       42       84       11      341      183      307       46      357
 5.022935  9.683164 11.778493 11.414163 11.191309 11.414163 11.191309 11.778493  9.718822 11.191309 11.120570 11.778493 11.778493
      363      194      292      274      298      198      200      172      287       36      173      142      339
11.484901 11.120570 12.142824 10.897717 11.778493 11.120570 11.778493 10.897717 11.555639 10.826978 11.778493  7.153344 12.213562
      215      125       33       40      268       10      354      246      347      283        9      386      358
11.414163 11.191309 11.778493 10.826978 11.778493 11.484901  8.696569 10.826978 12.142824 11.626378 11.484901 11.555639 11.484901
      186       61      202      152      349       54      319      375      237      185      157      115      234
11.484901 11.778493 11.191309  8.990161 12.142824 11.414163 11.849231 12.142824 11.191309 11.484901 10.533386 11.191309 11.778493
```

Fig: 15

Since we have split the dataset to training and testing and fit the model on the training set, we need to now fit the model on test data set to predict the final grades. The fig: 15 shows the final grades (G_agg) of the students on the test data.

CONCLUSION

To sum up, we used multiple linear regression to forecast the students' final grades based on a variety of factors, such as study time, study habits, a desire to pursue higher education, and prior failures. Our analysis of the Kaggle dataset on student alcohol consumption revealed that the mother's educational background and prior failures had a significant impact on students' academic performance. Although th effects were not statistically significant, study time and the desire to pursue higher education also showed some positive associations with final grades.

Our developed model only partially accounts for the variance in the data, as indicated by its R-squared value of 0.1753. Nevertheless, it still offered insightful information about the elements that influence how well students perform.

In order to increase the precision of our predictions, we might investigate additional modelling methods in the future, such as decision trees or neural networks. To enhance the calibre of our analysis, we could also gather more data or use data from other sources. Overall, this project offers a solid foundation for future study in the area of predicting student performance.

REFERENCES

- https://rstudio-pubs static.s3.amazonaws.com/335819_204193fc3ceb41acb5cf9b386e2cc91b.html
- https://rstudio-pubs-static.s3.amazonaws.com/365904_90b5ff49711448b5aa6ba2f09ffb4fc3.html
- https://www.kaggle.com/datasets/uciml/student-alcohol-consumption