Prediction Report

AUTHOR
Aaron Younger

PUBLISHED
December 10, 2025

Business Understanding

Business Problem

The data used in this analysis regard the performance in two distinct schools, mathematics and Portuguese. The primary goal of this analysis is to build prediction models that can predict the final score of a student using different variables. The ability for schools to be able to predict a students final grade can be highly valuable for several reasons. Schools can identify students early on that are most likely struggling. If educators understand which factors most influence final grades they can modify teaching strategies and course material.Predictive Models can also help schools in resource allocation giving them direction where to spend money.

Based on this information, the **business problem** is to determine whether a student's final grade can be accurately predicted using a set of explanatory variables.

Two Research Questions

Along with the business problem, this report explores two research questions:

1. Which of the two schools has a higher average G3 score, and is school membership a significant predictor in the regression model?

2. Does parental education level influence students' academic performance, as reflected in their G3 scores?

Now that the business problem and research questions have been clearly defined, the next step in this report is exploring the data to better understand its structure, key variables, and potential patterns relevant to students final grade. This **Data Understanding** phase provides the foundational insight needed for effective data preparation for modeling.

Data Understanding

```
# A tibble: 6 × 33
  school sex     age address famsize Pstatus  Medu  Fedu Mjob     Fjob    reason
  <chr>  <chr> <dbl> <chr>   <chr>   <chr>   <dbl> <dbl> <chr>    <chr>   <chr>
1 GP     F        18 U       GT3     A           4     4 at_home  teacher course
2 GP     F        17 U       GT3     T           1     1 at_home  other   course
3 GP     F        15 U       LE3     T           1     1 at_home  other   other
```

```
4 GP     F        15 U       GT3     T           4     2 health   servic… home
5 GP     F        16 U       GT3     T           3     3 other    other   home
6 GP     M        16 U       LE3     T           4     3 services other   reput…
# ℹ 22 more variables: guardian <chr>, traveltime <dbl>, studytime <dbl>,
#   failures <dbl>, schoolsup <chr>, famsup <chr>, paid <chr>,
#   activities <chr>, nursery <chr>, higher <chr>, internet <chr>,
#   romantic <chr>, famrel <dbl>, freetime <dbl>, goout <dbl>, Dalc <dbl>,
#   Walc <dbl>, health <dbl>, absences <dbl>, G1 <dbl>, G2 <dbl>, G3 <dbl>
```

Data Set Variable Key:

Categorical Variables:

1 School – student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira).
2 Sex – student's sex (binary: 'F' - female or 'M' - male).
3 Address – student's home address type (binary: 'U' - urban or 'R' - rural).
4 Famsize – family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3).
5 Pstatus – parent's cohabitation status (binary: 'T' - living together or 'A' - apart).
6 Mjob – mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home', or 'other').
7 Fjob – father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home', or 'other').
8 Reason – reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference, or 'other').
9 Guardian – student's guardian (nominal: 'mother', 'father', or 'other').
10 Schoolsup – extra educational support (binary: yes or no).
11 Famsup – family educational support (binary: yes or no).
12 Paid – extra paid classes within the course subject (Math or Portuguese) (binary: yes or no).
13 activities – extra-curricular activities (binary: yes or no).
14 nursery – attended nursery school (binary: yes or no).
15 higher – wants to take higher education (binary: yes or no).
16 internet – Internet access at home (binary: yes or no).
17 romantic – with a romantic relationship (binary: yes or no).

Numerical Variables:

18 Age – student's age (numeric: from 15 to 22).
19 Medu – mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education, 4 - higher education).
20 Fedu – father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education, 4 - higher education).
21 Traveltime – home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, 4 - >1 hour).
22 Studytime – weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, 4 - >10 hours).
23 Failures – number of past class failures (numeric: n if $1 \leq n < 3$, else 4).
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).

These grades are related with the course subject, Math or Portuguese:

31 G1 – first period grade (numeric: from 0 to 20).
32 G2 – second period grade (numeric: from 0 to 20).
33 G3 (Dependent Variable) – final grade (numeric: from 0 to 20, output target).

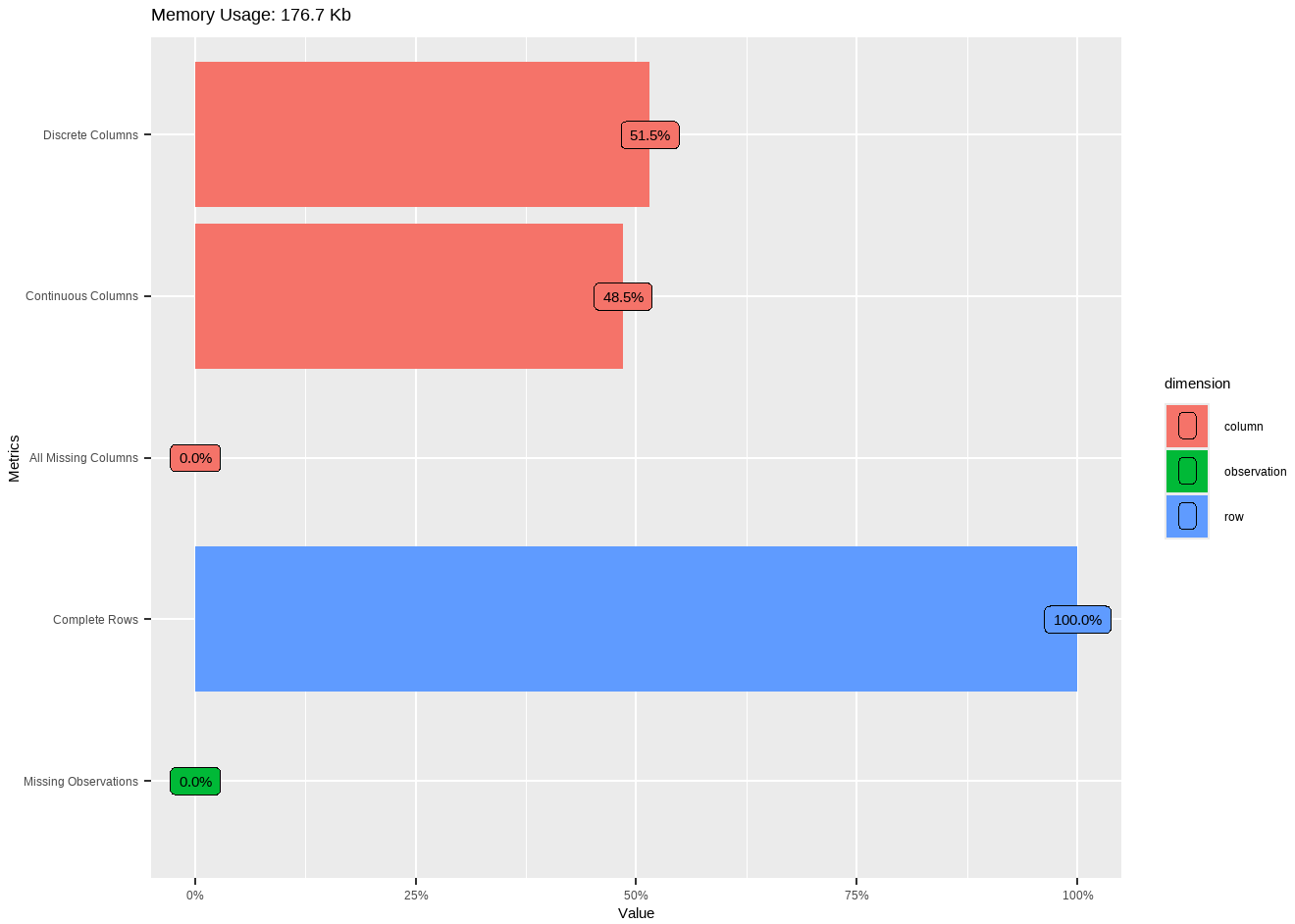
Dataset Exploration

Number of Rows and Columns

```
# A tibble: 2 × 2
  Statistic         Value
  <chr>             <int>
1 Number of Columns    33
2 Number of Rows      649
```

This dataset contains 33 variables, as indicated by the number of columns, and 649 observations, as indicated by the number of rows.

Data Structure and Completeness Overview



This graph shows that 17 variables are categorical variables and the remaining 16 variables are numeric. This dataset also contains no missing values.

Numeric EDA

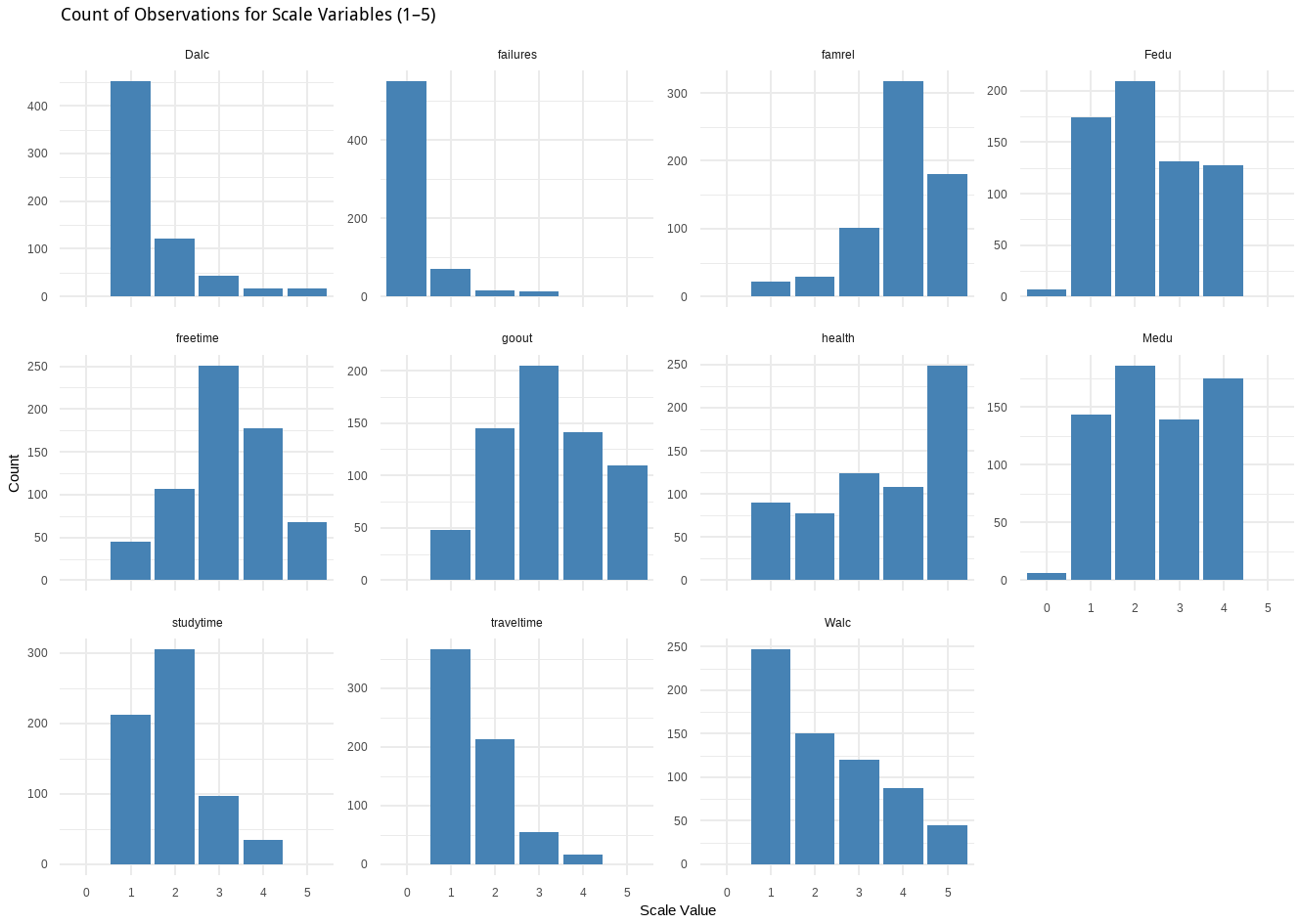
Since the dataset contains both categorical and numerical variables, the exploratory data analysis (EDA) will be conducted in two parts: one focusing on categorical variables and the other on numeric variables. First, this analysis will explore the relationships and patterns found in numeric variables.

Make Data set of all Numeric Variables

```
# A tibble: 6 × 16
    age  Medu  Fedu traveltime studytime failures famrel freetime goout  Dalc
  <dbl> <dbl> <dbl>      <dbl>     <dbl>    <dbl>  <dbl>    <dbl> <dbl> <dbl>
1    18     4     4          2         2        0      4        3     4     1
2    17     1     1          1         2        0      5        3     3     1
3    15     1     1          1         2        0      4        3     2     2
4    15     4     2          1         3        0      3        2     2     1
5    16     3     3          1         2        0      4        3     2     1
6    16     4     3          1         2        0      5        4     2     1
# ℹ 6 more variables: Walc <dbl>, health <dbl>, absences <dbl>, G1 <dbl>,
#   G2 <dbl>, G3 <dbl>
```

A subset containing only the numeric variables was created from the original dataset to perform numeric specific EDA.

Variables on a Scale



Some of the numerical variables were recorded using a scale. To investigate the distribution of observations per scale amount (0-5), column charts were created to display the frequency of observations within each level. This was to help identify if there was any level imbalance within variables.

The column charts indicated that all scaled variables exhibited considerable level imbalance and highlighted where scale levels could be merged. All eleven scale-based numeric variables were recoded into either binary or multi-level categorical variables to produce a more balanced distribution of observations.

Merge levels to Make Balanced predictors

```
# A tibble: 6 × 33
  school sex     age address famsize Pstatus Medu       Fedu  Mjob  Fjob  reason
  <chr>  <chr> <dbl> <chr>   <chr>   <chr>   <fct>      <fct> <chr> <chr> <chr>
1 GP     F        18 U       GT3     A       High Educ… High… at_h… teac… course
2 GP     F        17 U       GT3     T       Low Educa… Low … at_h… other course
3 GP     F        15 U       LE3     T       Low Educa… Low … at_h… other other
4 GP     F        15 U       GT3     T       High Educ… Low … heal… serv… home
5 GP     F        16 U       GT3     T       High Educ… High… other other home
6 GP     M        16 U       LE3     T       High Educ… High… serv… other reput…
```

```
# ℹ 22 more variables: guardian <chr>, traveltime <fct>, studytime <fct>,
#   failures <fct>, schoolsup <chr>, famsup <chr>, paid <chr>,
#   activities <chr>, nursery <chr>, higher <chr>, internet <chr>,
#   romantic <chr>, famrel <fct>, freetime <fct>, goout <fct>, Dalc <fct>,
#   Walc <fct>, health <fct>, absences <dbl>, G1 <dbl>, G2 <dbl>, G3 <dbl>
```

The reason for transforming these variables from numeric scales to categories is to produce better level balance. Creating level balance makes sure there is sufficient observation counts within all levels. Creating a more balanced observation count across levels within a variable will help improve model stability and reliability. Models that will be used in this project such as linear regression and decision trees rely on having adequate observation counts for each category. If one level has to little observations, the model cannot reliably estimate its effect.The transformations for these numeric variables included either binary or multi-leveled category transformations. Each variable that was recoded is detailed below, including the original scale values and the method used to collapse those values into the new categorical groupings.

- Failures (classes failed) was converted into a binary factor with two levels: "Failed" for students who had one or more past class failures, and "Not Failed" for those with no past failures. Before transformation, the variable ranged from 0 to 3, with the majority of observations concentrated at 0. Collapsing the scale into two categories increased the number of observations in the "Failed" group and reduced the degree of class imbalance. However, the "Not Failed" category still remains the dominant group, so some imbalance persists but less than before.

- Dalc (workday alcohol consumption) was converted into a binary factor with two levels: "1" for the lowest level of consumption and ">1" for any consumption greater than 1. Prior to transformation, the variable ranged from 1-5, with most of the observations in 1. Collapsing these categories helped increase the number of observations in the ">1" alcohol consumption category, reducing class imbalance.

- Medu (Mother's Education) was converted into a binary factor with two levels: "Low Education" which represented mothers with a primary or lower education and "High Education" which represents mothers with a secondary education or higher. Prior to transformation this variable ranged from 0-4. This variable's zero value had very small observations, so to mitigate that 0 was combined with 1-2. The new binary values of "Low Education" and "High Education" are very close in observation count meaning this variable now has good class balance.

- Fedu (Father's Education) was converted into a binary factor with two levels: "Low Education" which represented Fathers with a primary or lower education and "High Education" which represents fathers with a secondary education or higher. Prior to transformation this variable ranged from 0-4. This variable's zero value had very small observations, so to mitigate that 0 was combined with 1-2. The new binary values of "Low Education" and "High Education" are very close in observation count meaning this variable now has good class balance. "Low Education" has more observation counts than "High Education" however each class has an adequate amount of observations which prepares this variable for modeling.

- Freetime (free time after school) was converted into multiple categories: "Low" for <=2, "Medium" for = 3, and "High" for responses >= 4. Prior to transformation this variable ranged from 1-5, this transformation combined ranks 1-2 and 4-5 to have closer observation counts to rank 3 which had the

majority of observations. After transformation “Medium” and “High” have very similar observation counts with “Low” being the minority category. This variable now has an adequate amount of observations in each category for modeling.

- Gouut (going out with Friends) was converted into multiple categories: “Low” for <=2, “Medium” for = 3, and “High” for responses >= 4. Prior to transformation this variable ranged from 1-5, this transformation was to combine ranks 1-2 and 4-5 to have closer observation counts to rank 3 which had the majority of observations. After transformation, this variable has class balance with “High” having slightly higher observation counts than “Low” and “Medium”. This variable now has good class balance.

- Health (Current Health Status) was converted into multiple categories: “Low” for ranks <= 2, “Medium” for ranks between 3 and 4, and “High” for rank = to 5. Prior to transformation this variable ranged from 1-5 with 5 being the majority class. This transformation helped bring class balance by combining ranks 1-2 and 3-4 producing similar observation counts that rank 5 has.

- Studytime (weekly study time) was converted into multiple categories: “<2 Hours” for rank = 1, “2-5 Hours” for rank = 2, and “>5 Hours” for ranks 3 and 4. Prior to transformation this variable ranged from 1-4 with 2 being the majority class. This transformation was intended to combine ranks 3-4 to produce higher observation counts. Although there is still some class imbalance between categories there are now an adequate amount of observations in each category for modeling.

- Traveltime (home-to-school travel time) was converted into multiple categories: “<15 min” for rank = 1, “15–30 min” for rank = 2, and “>30 min” for ranks = 3-4. This transformation was intended to combine ranks 3 and 4 giving them higher observation counts. There is still class imbalance as “>30 min” has lower observations than the other two categories, however this transformation brings better class balance than before.

- Walc (Weekend Alcohol Consumption) was converted into multiple categories: “Low” for rank = 1, “Medium” for ranks 2-3, and “High” for ranks >= 4. Prior to this transformation this variable ranged from 1-5 with 1 being the majority class. This transformation was to give more observation counts to ranks 2-3 and ranks 4-5 to help even observation counts across categories. After transformation there is still some class imbalance as “High” has lower observation counts than “low” and “Medium” however after transformation there is better class balance than before.

- Famrel (Family Relationhsip Quality) was converted into a binary: “Good” for scores above rank 3, “Bad” for scores below rank 3. Prior to transformation this variable ranged from 1-5. After transformation there is still some class imbalance as “Good” has significantly more observations than “Bad”. However, there is an adequate amount of observations in both classes to be represented well in modeling.

Factor Conversion

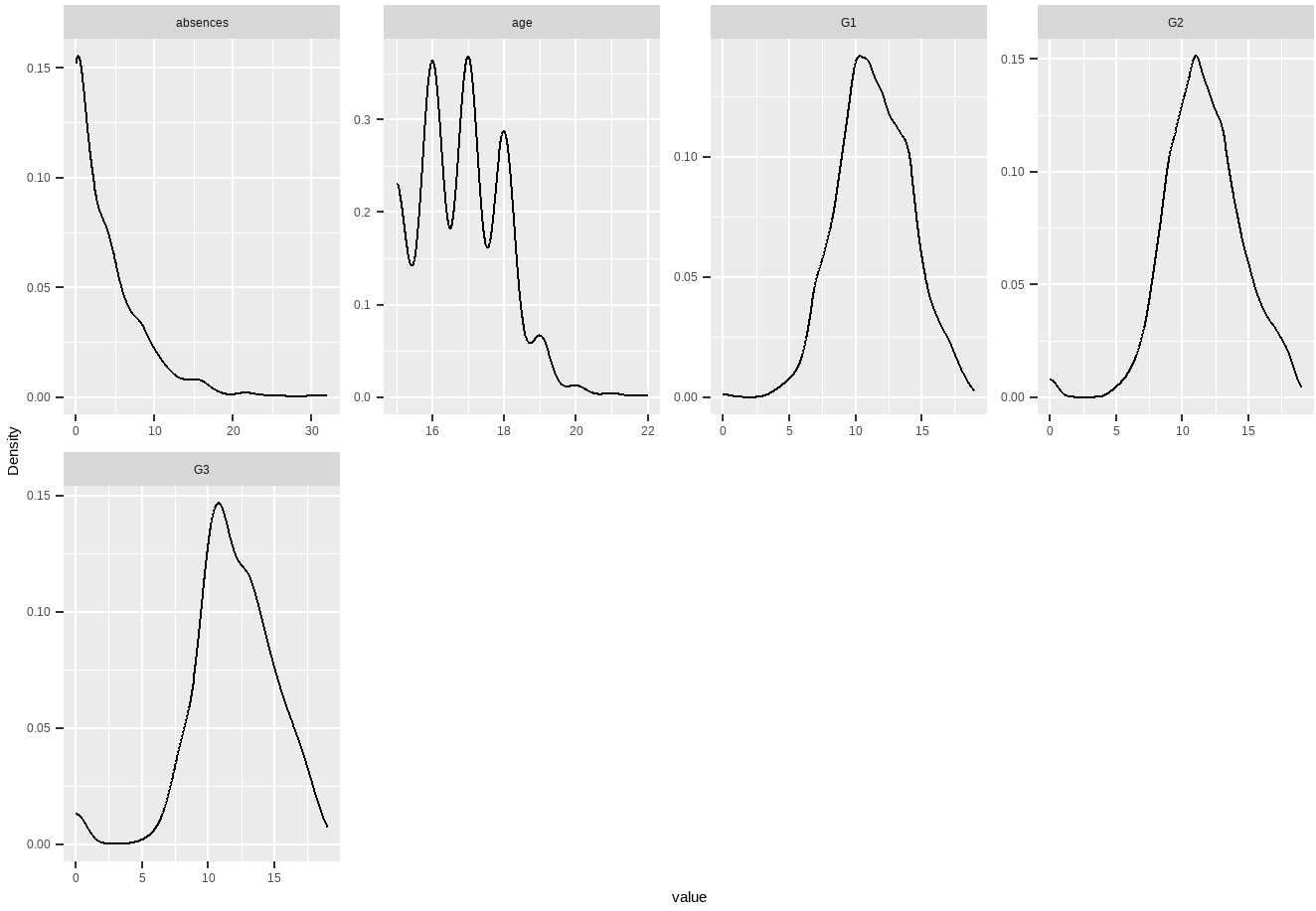
After each variable was transformed using categories, they were converted from numeric variables to factors as the variables did not imply mathematical distances and allows the model to estimate separate effects for each category.

Remove New Factor Variables from Numeric Only Dataset

```
# A tibble: 6 × 5
    age absences    G1    G2    G3
  <dbl>    <dbl> <dbl> <dbl> <dbl>
1    18        4     0    11    11
2    17        2     9    11    11
3    15        6    12    13    12
4    15        0    14    14    14
5    16        0    11    13    13
6    16        6    12    12    13
```

Since the numeric variables on a scale were grouped and changed to factors that left only five numeric variables left, age, absences, G1, G2, and G3.

Distribution of Numeric Variables



The distribution of the five numeric variables were checked using a density plot. G1, G2, and G3 had relatively normal distribution with slight left skewness, this is due to some of the scores in these variables being zero. Age shows slight right skewness. Absences shows more extreme right skewness that will need to be explored.

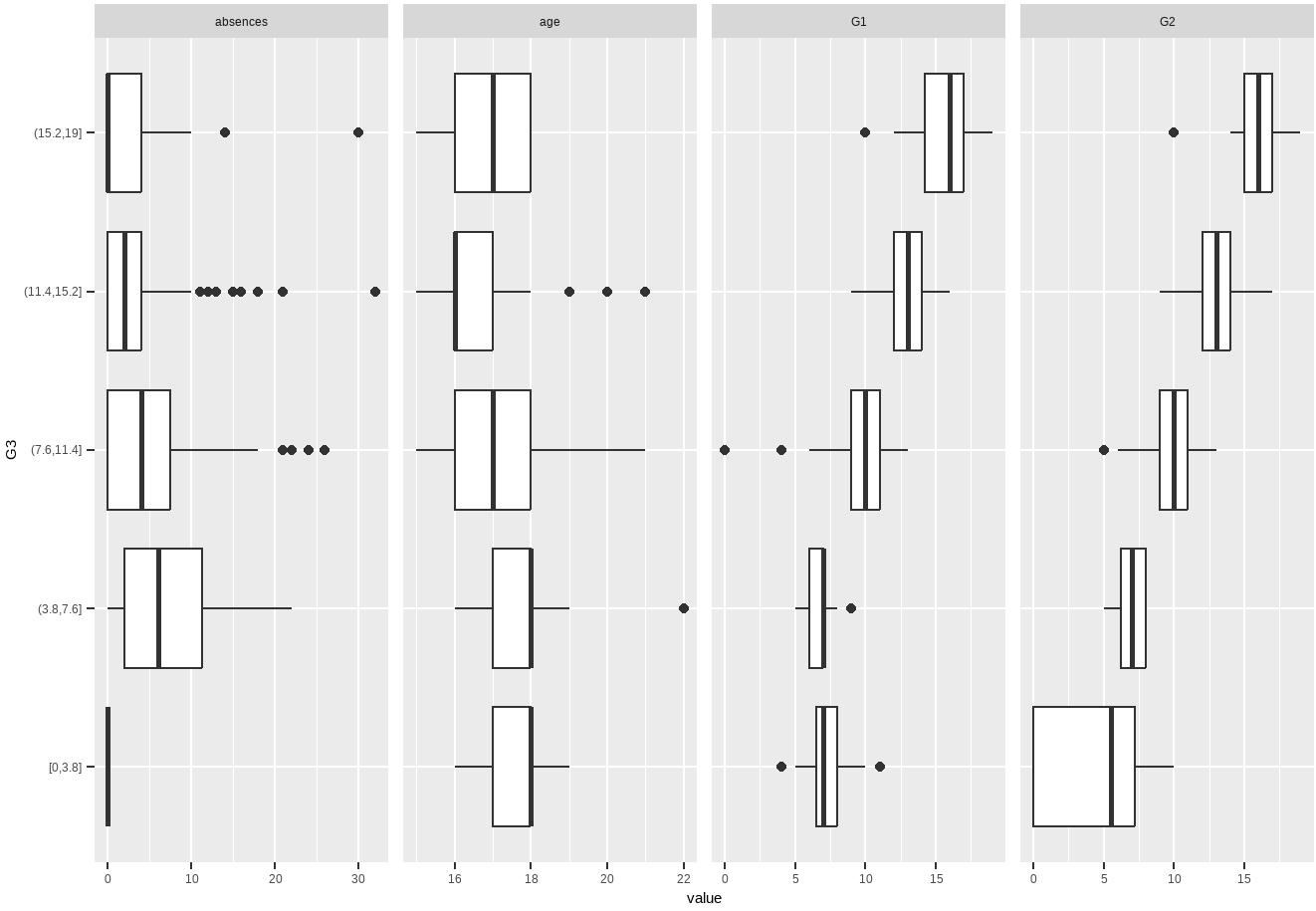
Outliers

```
# A tibble: 5 × 6
  variables outliers_cnt outliers_ratio outliers_mean with_mean without_mean
```

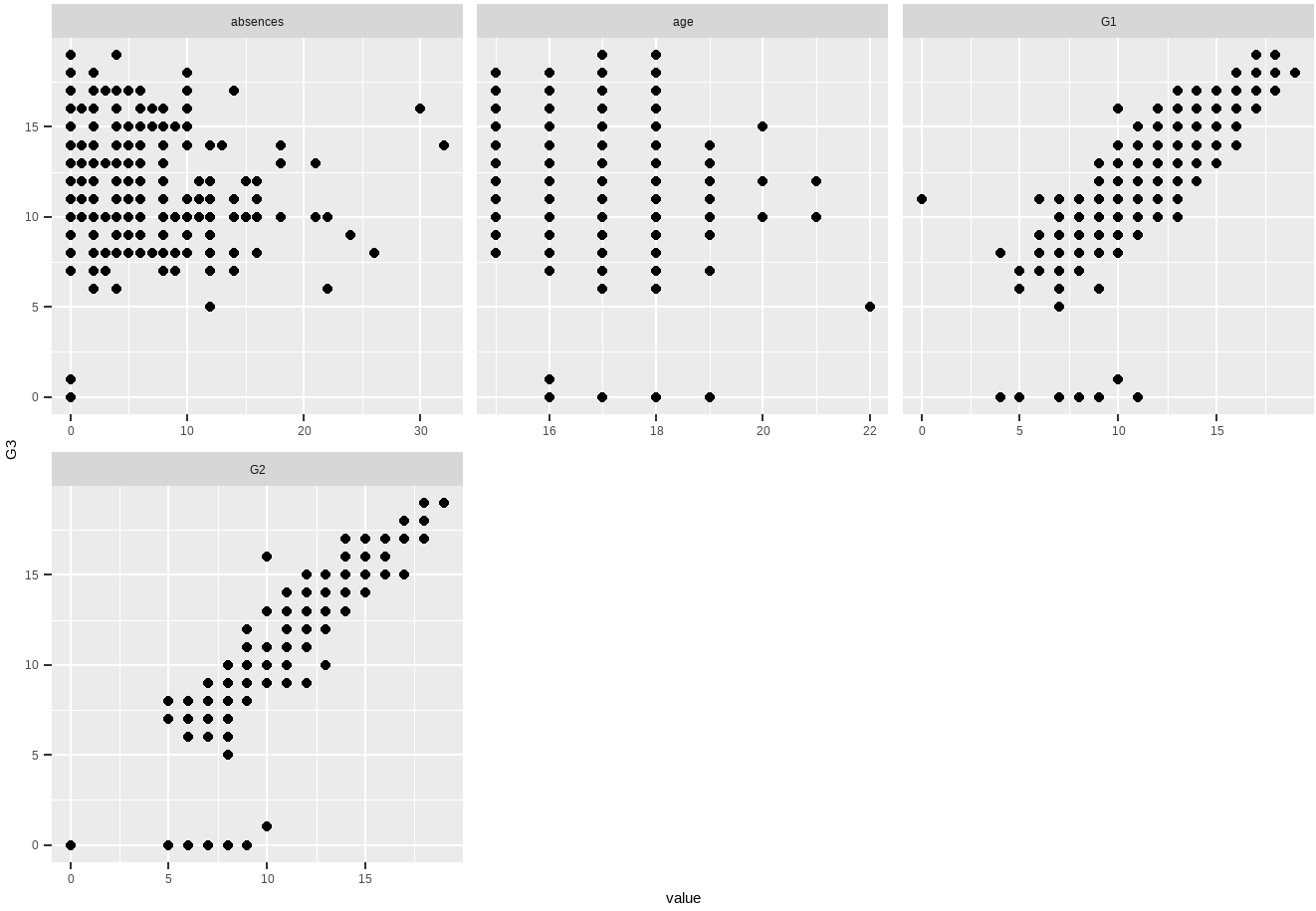
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
1	age	1	0.154	22	16.7	16.7
2	absences	21	3.24	19.6	3.66	3.13
3	G1	16	2.47	11.1	11.4	11.4
4	G2	25	3.85	11.4	11.6	11.6
5	G3	16	2.47	0.0625	11.9	12.2

Although all variables contain outliers, they do not pose a significant concern for this analysis. For the absences variable, the mean increases when outliers are included, indicating that extreme values are pulling the distribution to the right and contributing to its right skewness. In contrast, G1, G2, and G3 show lower means when outliers are included, consistent with their left-skewed distributions, where zero scores pull the data toward the lower end. Outliers will futher be explored and possibly removed in the data preperation phase.

Boxplots

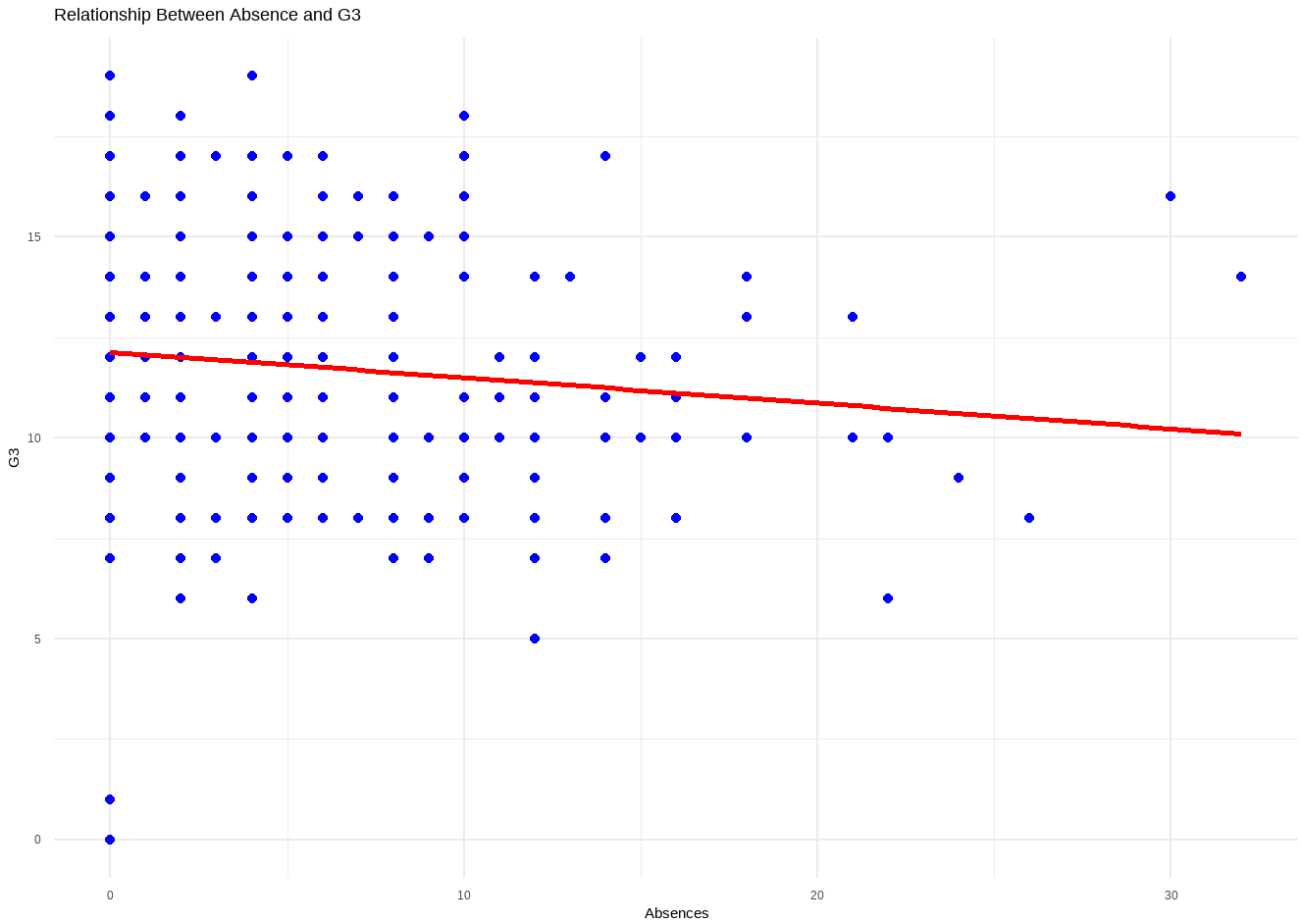


Scatterplots



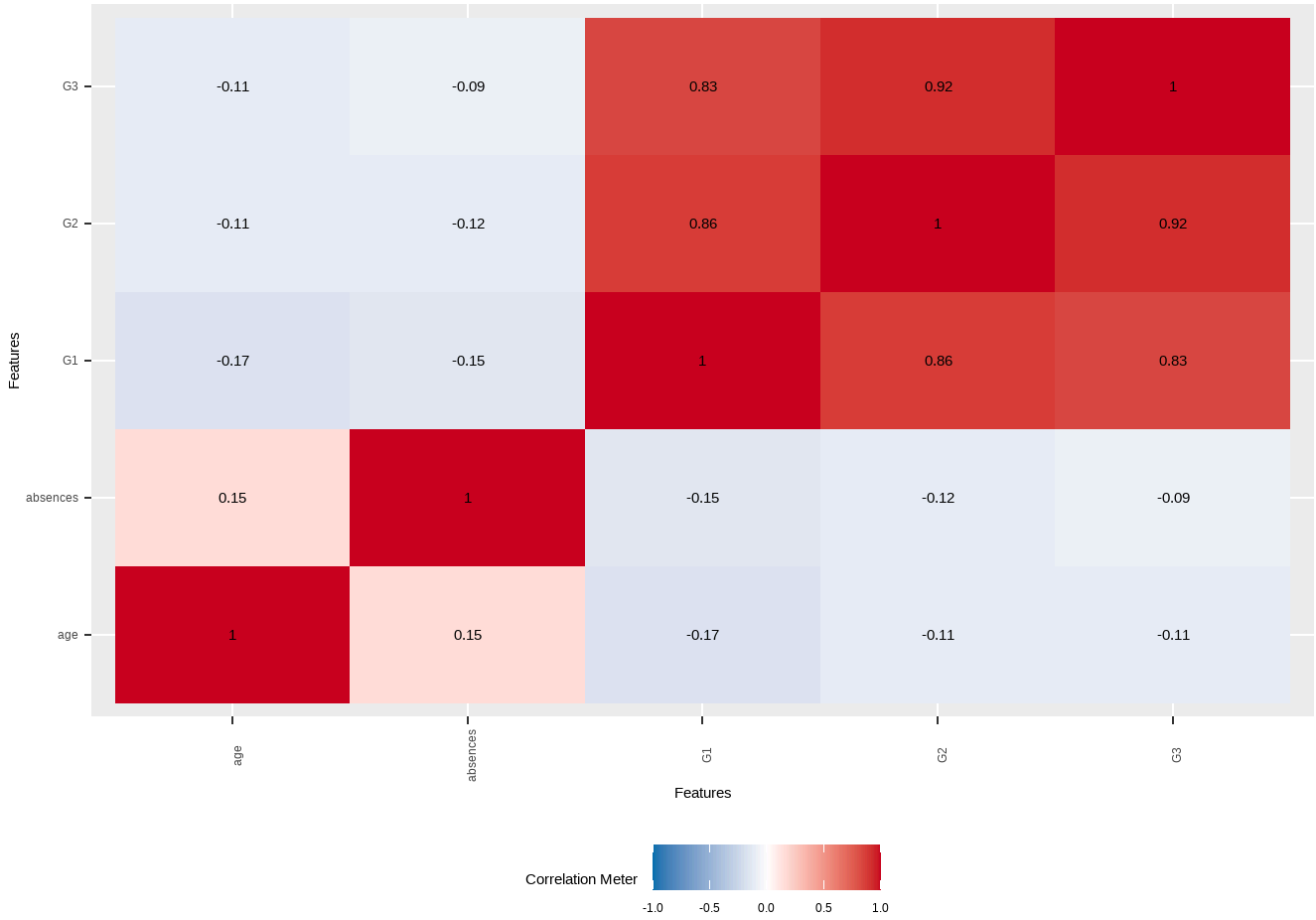
Comments on Scatterplots:
Based off the scatterplots, G1 and G2 have a positive linear relationship with G3. From the scatterplot it does not seem like absence has a relationship to G3, however due to uncertainty a more advanced scatterplot will be graphed comparing absence to G3 to see if there is a relationship.

Explore Variable Relationship with Depvar



Comments on Absence and G3:
Based off this scatterplot Absence and G3 have a slight negative relationship which makes sense, which aligns with the expectation that students with more absences tend to have lower final grades. However, the relationship is very weak and can be considered close to neutral.

Correlation Numeric Variables



Comments on Correlation Matrix:
All variables show weak correlation to the dependent variable except for G1 and G2 who have very strong positive correlation to the Dependent Variable, G3. Since both of these variables have high correlation with the dependent variable, there is a potential for multicollinearity among G1 and G2 variables. Note that Absences have a weak negative correlation with G3 which supports the findings in the scatterplot above.

Mean of Depvar G3

```
Mean of G3
```

```
[1] 11.90601
```

The mean of G3 will be calculated for both the training and test sets and compared to the mean of G3 in the full dataset. This comparison will help determine whether the train–test split is representative of the overall data and free from unintended bias. So the mean of G3 over the whole dataset will be used in the modeling phase.

Categorical EDA

A subset from the original dataset was made just containing categorical variables.

Convert all Categorical Variables to Factors

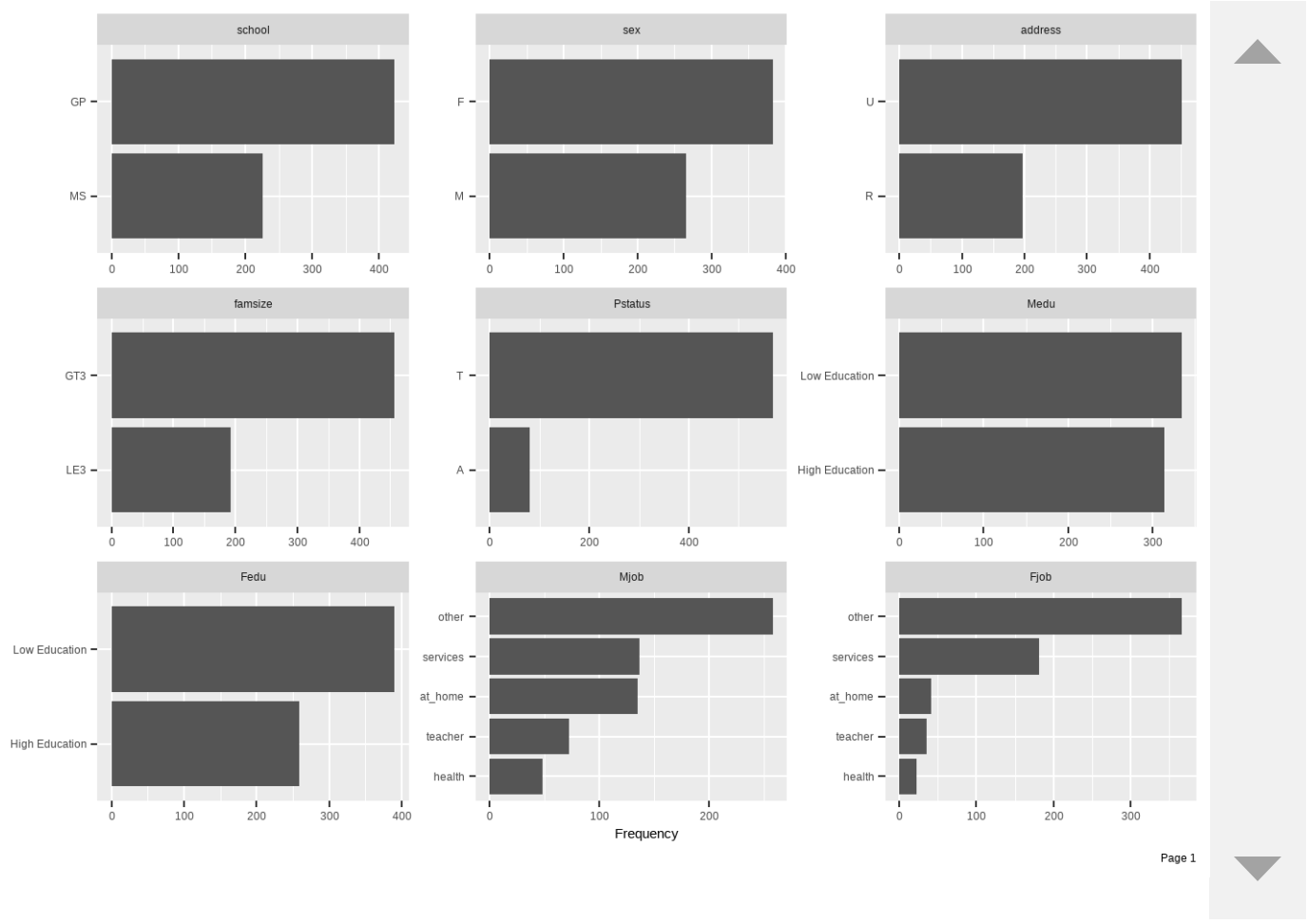
Categorical variables that were stored as characters were converted to factors so R could correctly interpret them as categorical data.

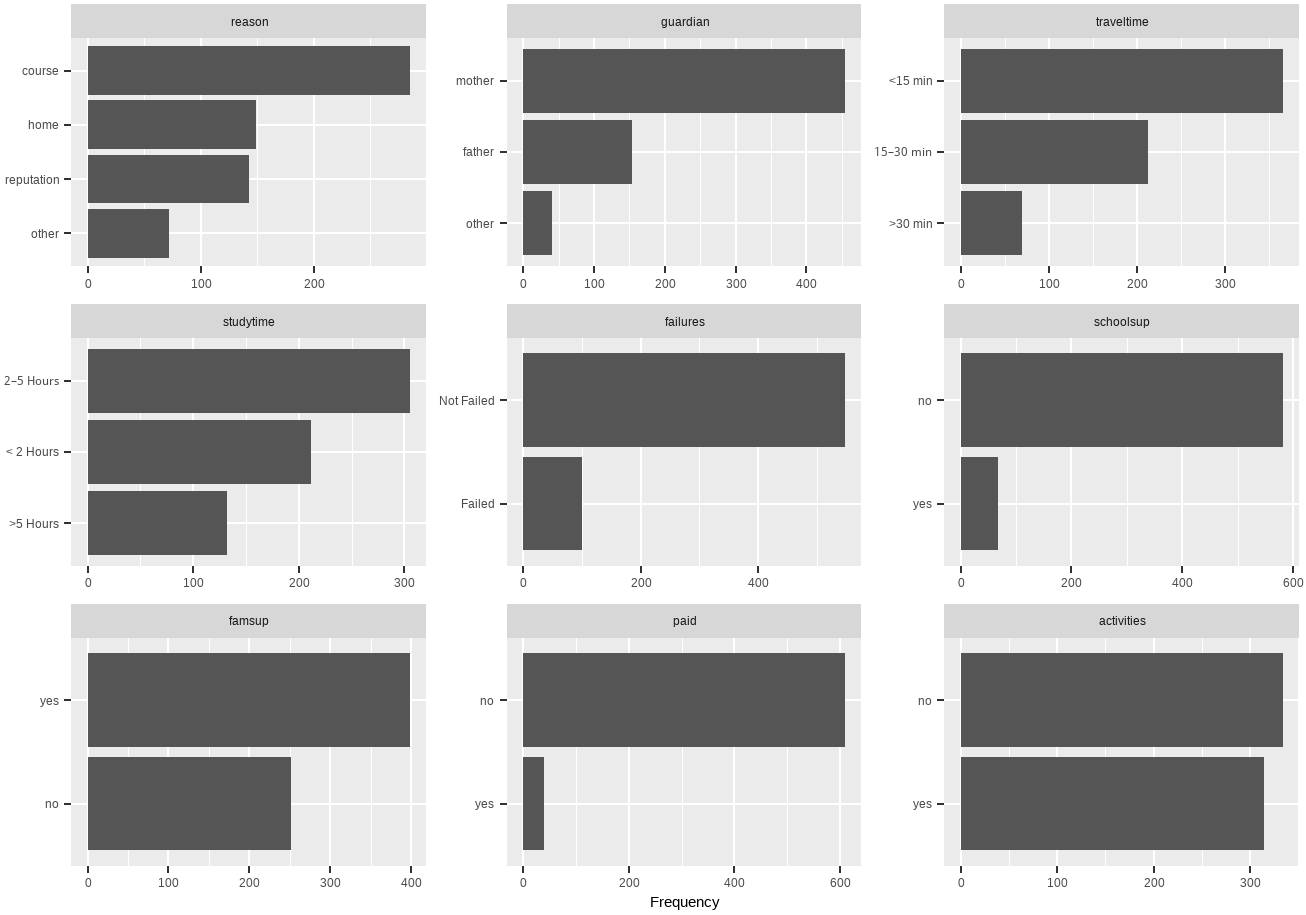
Make subset of all Categorical Variables

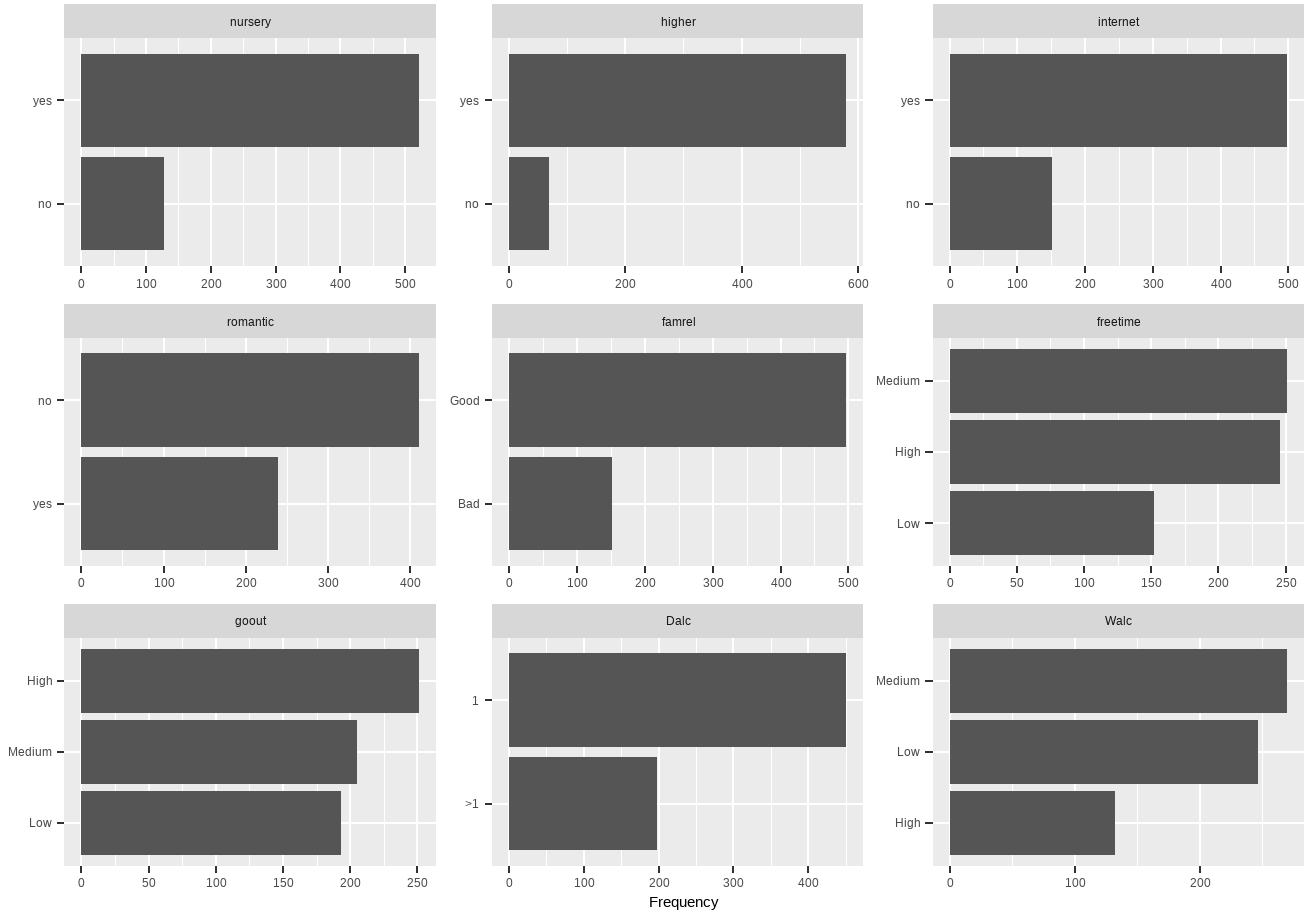
```
# A tibble: 6 × 29
  school sex   address famsize Pstatus Medu    Fedu  Mjob  Fjob  reason guardian
  <fct>  <fct> <fct>   <fct>   <fct>   <fct>   <fct> <fct> <fct> <fct>  <fct>
1 GP     F     U       GT3     A       High E… High… at_h… teac… course mother
2 GP     F     U       GT3     T       Low Ed… Low … at_h… other course father
3 GP     F     U       LE3     T       Low Ed… Low … at_h… other other  mother
4 GP     F     U       GT3     T       High E… Low … heal… serv… home   mother
5 GP     F     U       GT3     T       High E… High… other other home   father
6 GP     M     U       LE3     T       High E… High… serv… other reput… mother
# ℹ 18 more variables: traveltime <fct>, studytime <fct>, failures <fct>,
#   schoolsup <fct>, famsup <fct>, paid <fct>, activities <fct>, nursery <fct>,
#   higher <fct>, internet <fct>, romantic <fct>, famrel <fct>, freetime <fct>,
#   goout <fct>, Dalc <fct>, Walc <fct>, health <fct>, G3 <dbl>
```

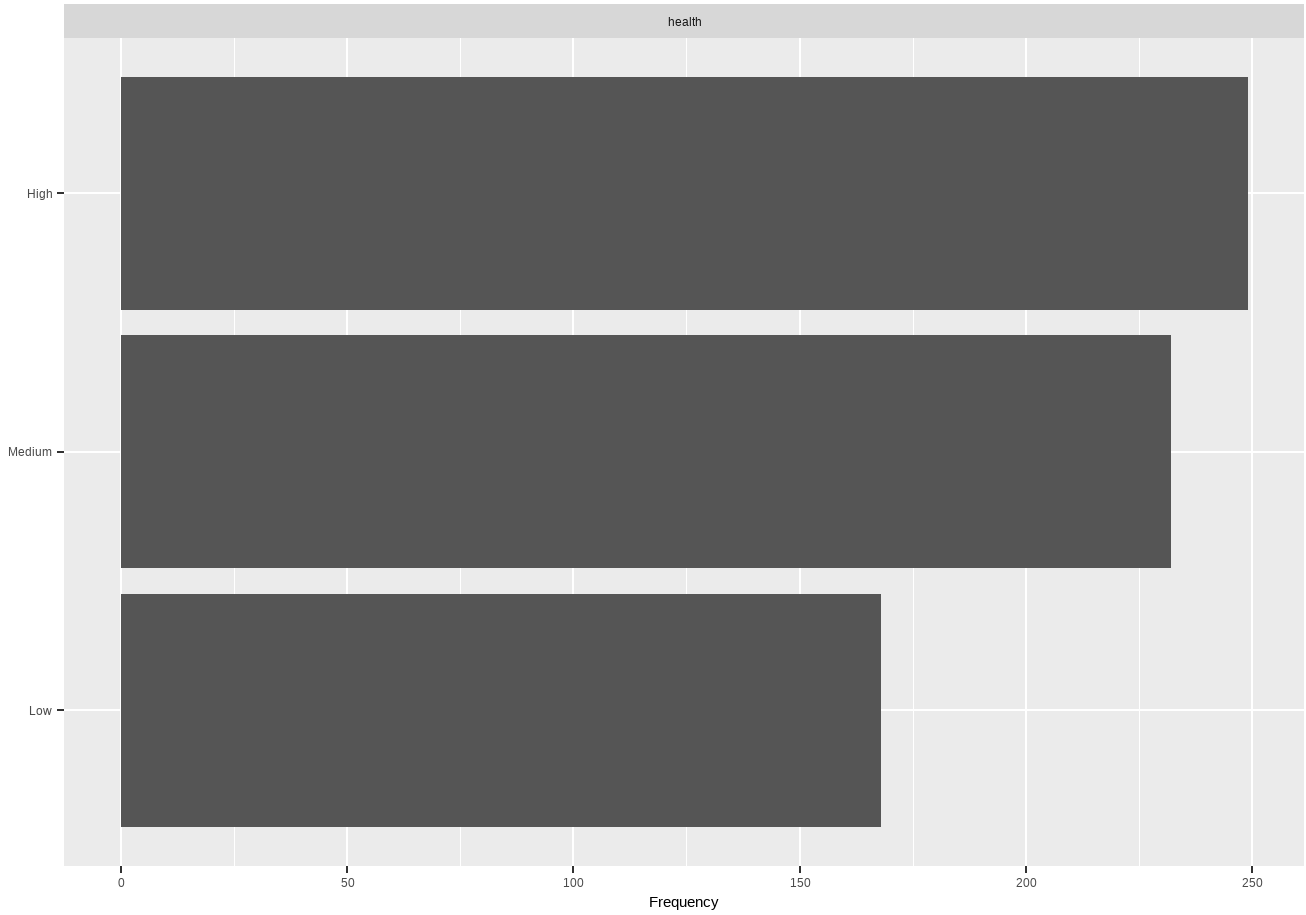
A subset containing only categorical variables was created from the original dataset to perform categorical specific EDA. The Dependent variable was also included in this subset so the relationship between categorical variables and the dependent variable could be explored.

Bar Plot of Categorical Variables




reason
course
home
reputation
other
0
100
200
guardian
mother
father
other
0
100
200
300
400
traveltime
<15 min
15–30 min
>30 min
0
100
200
300
studytime
2–5 Hours
< 2 Hours
>5 Hours
0
100
200
300
failures
Not Failed
Failed
0
200
400
schoolsup
no
yes
0
200
400
600
famsup
yes
no
0
100
200
300
400
paid
no
yes
0
200
400
600
activities
no
yes
0
100
200
300
Frequency


nursery
yes
no
0
100
200
300
400
500
higher
yes
no
0
200
400
600
internet
yes
no
0
100
200
300
400
500
romantic
no
yes
0
100
200
300
400
famrel
Good
Bad
0
100
200
300
400
500
freetime
Medium
High
Low
0
50
100
150
200
250
goout
High
Medium
Low
0
50
100
150
200
250
Dalc
1
>1
0
100
200
300
400
Walc
Medium
Low
High
0
100
200
Frequency



Majority of the categorical variables have good class balance and sufficient observations in each level for proper modeling, however there are some variables that struggle with class imbalance. These variables are explored using proportion tables.

Exploring Proportions and Imbalance in Categorical Variables

```
 Proportion of Parents Togther and Away
```

```
        A         T 
0.1232666 0.8767334 
```

```
 Proportion of the different jobs  Mothers Have
```

```
   at_home     health      other   services    teacher 
0.20801233 0.07395994 0.39753467 0.20955316 0.11093991 
```

```
 Proportion of the different jobs  Fathers Have
```

```
   at_home     health      other   services    teacher
0.06471495 0.03543914 0.56548536 0.27889060 0.05546995
```

```
 Proportion of Students who Have Failed
```

```
   Failed Not Failed
 0.1540832  0.8459168
```

```
 Proportion of students Travel Times
```

```
  <15 min 15–30 min   >30 min
0.5639445 0.3281972 0.1078582
```

```
 Proportion of Those Recieving Extra Educational Support
```

```
       no       yes
0.8952234 0.1047766
```

```
 Proportion of Students who Have Paid for Classes
```

```
       no       yes
0.8952234 0.1047766
```

```
 Proportion of Those Desiring Higher Education
```

```
       no       yes
0.1063174 0.8936826
```

The Variables that struggle from class imbalance are Parent Status, Mother Jobs, Father Jobs, Failures, Student Travel Time, Extra educational support, Shoolup, and higher education. All these variables have a level that represents less than 15% of the overall variable. This class imbalance is recognized and will be noted as a potential limitation to the models.

Simple Random Forest to see Categorical importance on the Depvar

A simple random forest model is made to explore categorical variable importance to the dependent variable G3. This can be used to help identify what categorical variables will most likely be important in modeling.

Display Important Variables

	Variable	%IncMSE	IncNodePurity
1	failures	33.4868337	962.84392
2	higher	18.4581504	360.93798
3	schoolsup	14.6059471	146.78509
4	school	9.7865979	304.08131
5	Fedu	8.6938104	131.52608
6	studytime	7.6951486	285.26023
7	sex	7.6842757	139.73208
8	Medu	6.0624249	138.26226
9	address	3.8536285	125.13976
10	goout	3.6874433	248.28190
11	Dalc	3.6002718	126.35240
12	Walc	3.4839511	234.94128
13	reason	2.9904544	380.94061
14	health	2.5639725	235.76500
15	famsup	2.3315875	153.50967
16	famrel	2.2021770	102.13679
17	nursery	2.2006402	110.03095
18	activities	1.9150415	120.63608
19	traveltime	1.3973149	206.23927
20	famsize	0.8355194	99.44191
21	Mjob	0.6379138	390.78312
22	Fjob	0.5722228	323.83669
23	internet	0.5516101	120.38446
24	freetime	0.5510254	234.77329
25	guardian	0.4466440	136.88895
26	Pstatus	-0.5346445	69.58376
27	paid	-0.7692341	37.01538
28	romantic	-1.8227569	116.37883

When looking at the output from the random forest we look at "%IncMSE" to identify important variables. The %IncMSE metric reflects the percentage increase in prediction error when a variable's values are randomly imputed. A higher %IncMSE indicates that the model relies more heavily on that variable, meaning it has a stronger influence over G3. In this case failures and higher education have the greatest impact on G3 whereas paid and romantic have the lowest affect on G3.

Data Understanding Summary

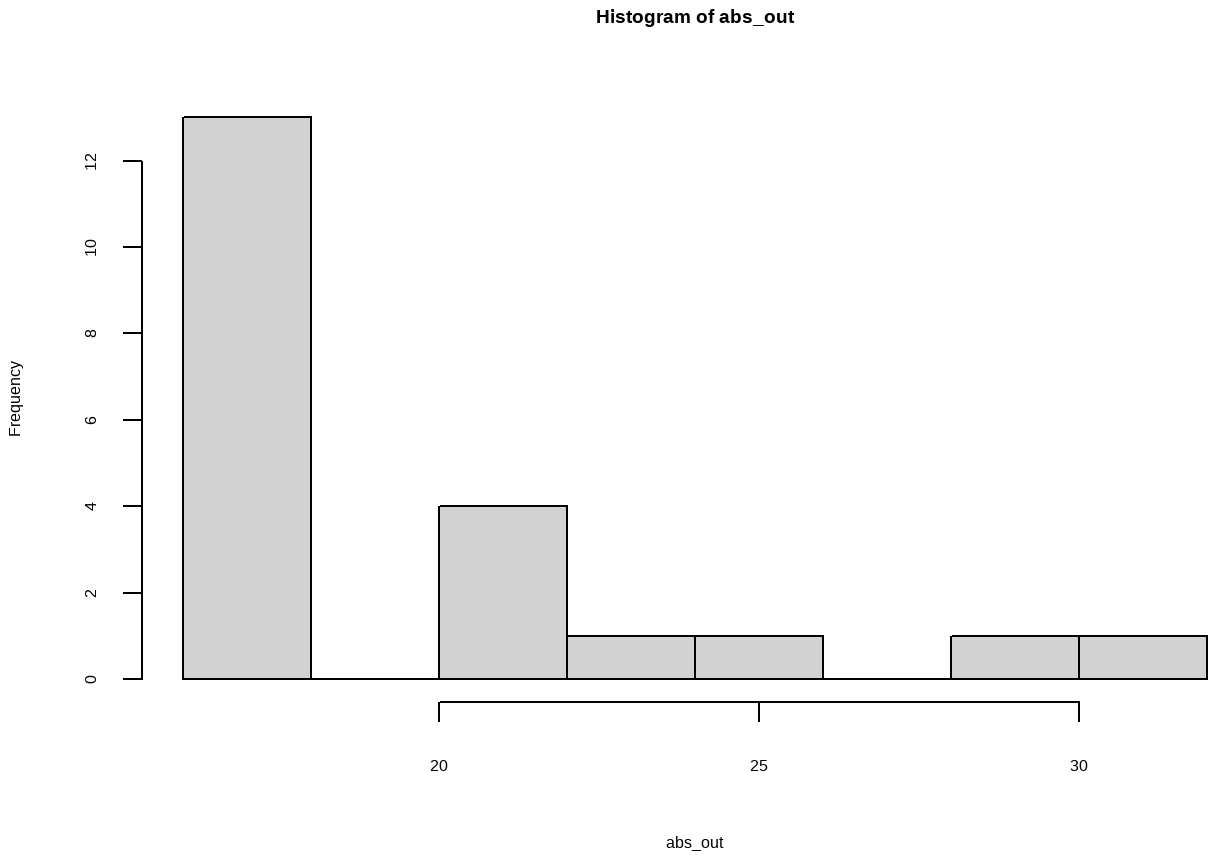
The Data Understanding phase resulted in several important transformations that will greatly support the modeling process. During the numeric EDA, all scaled numeric variables were binned into grouped categories to improve level balance, and these variables were converted into factors, leaving only five true numeric variables in the dataset. The numeric analysis also revealed that G1 and G2 are strongly correlated with G3, indicating a potential risk of multicollinearity that will need to be monitored during modeling. In the categorical EDA, most variables showed reasonable class balance, with only a few exhibiting imbalance. Additionally, a Random Forest model was used to assess categorical variable importance relative to the dependent variable, providing early insight into which predictors may be most influential. Overall, the Data

Understanding phase was highly valuable, as it improved variable balance and clarified which predictors are most closely related to the dependent variable.

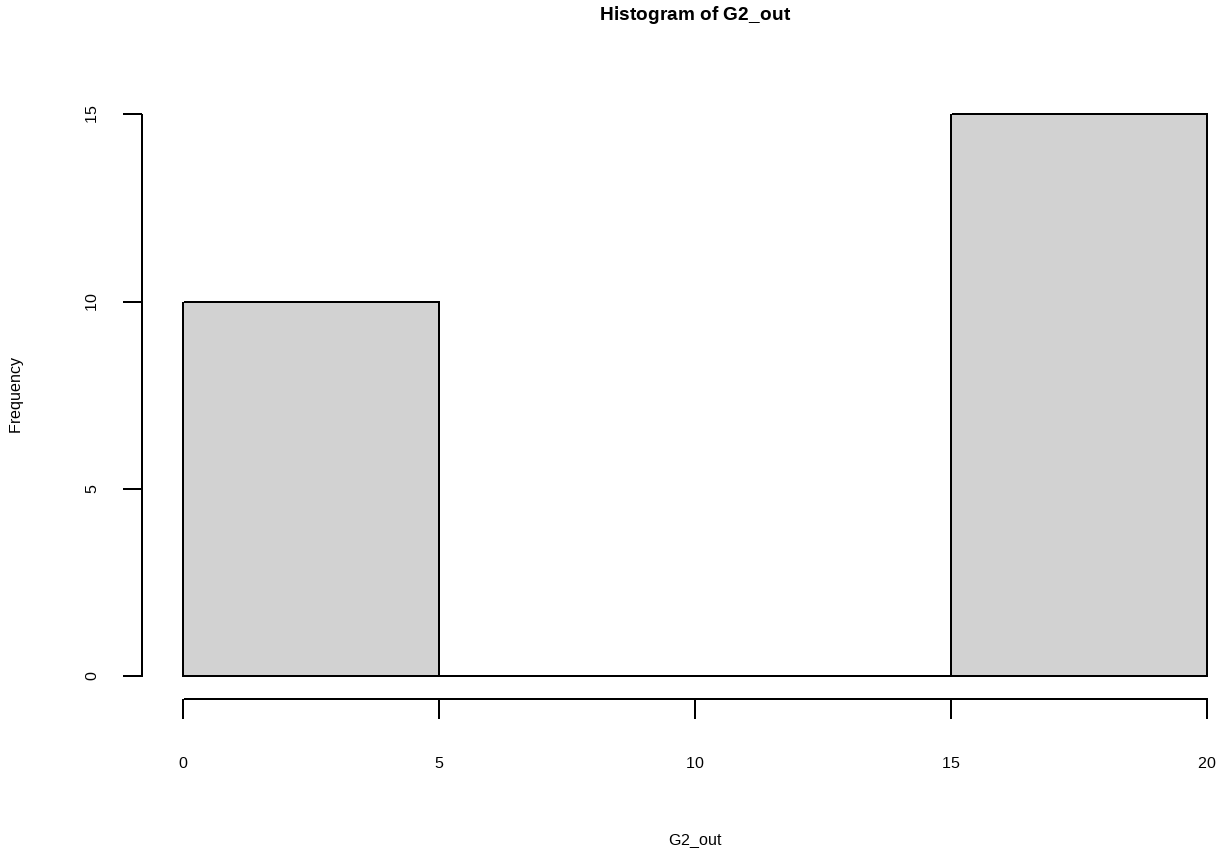
Data Preparation

Investigate Outliers

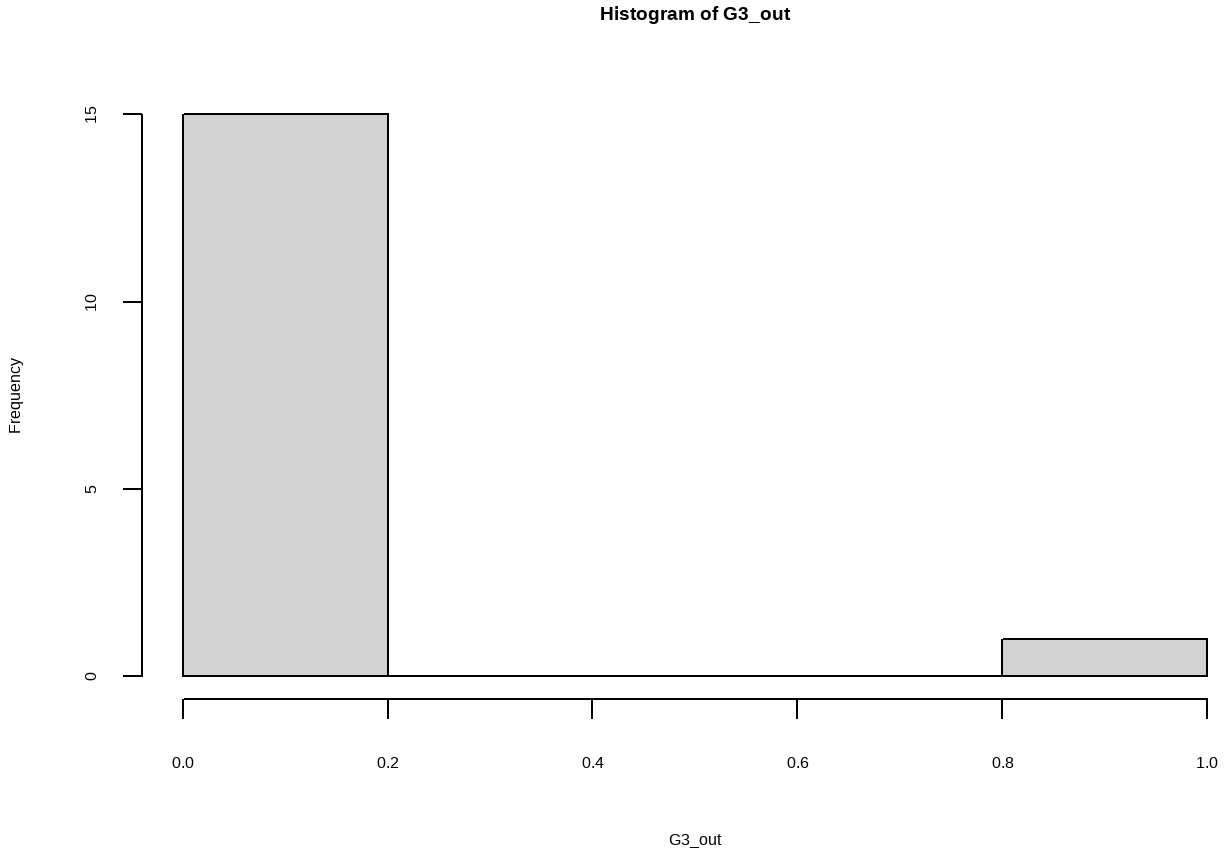
```
[1] 16 16 24 22 16 32 16 16 30 21 16 18 16 26 16 16 22 18 18 16 21
```




```
[1] 18 18 18 18 19 18 18 18 18 18  0  5 18  0  0  5 18 18  0  0  0 18  0  5 18
```




```
[1] 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```



Histogram and boxplot visualizations were used to examine the outlier observations across all numeric variables. For Age, G1, G2, and G3, no outliers were removed because the values fell within a reasonable range and did not appear to be extreme in the context of the data. However, the four highest values of absences—32, 30, 26, and 24—were removed because they were substantially higher than the rest of the distribution and each occurred only once.

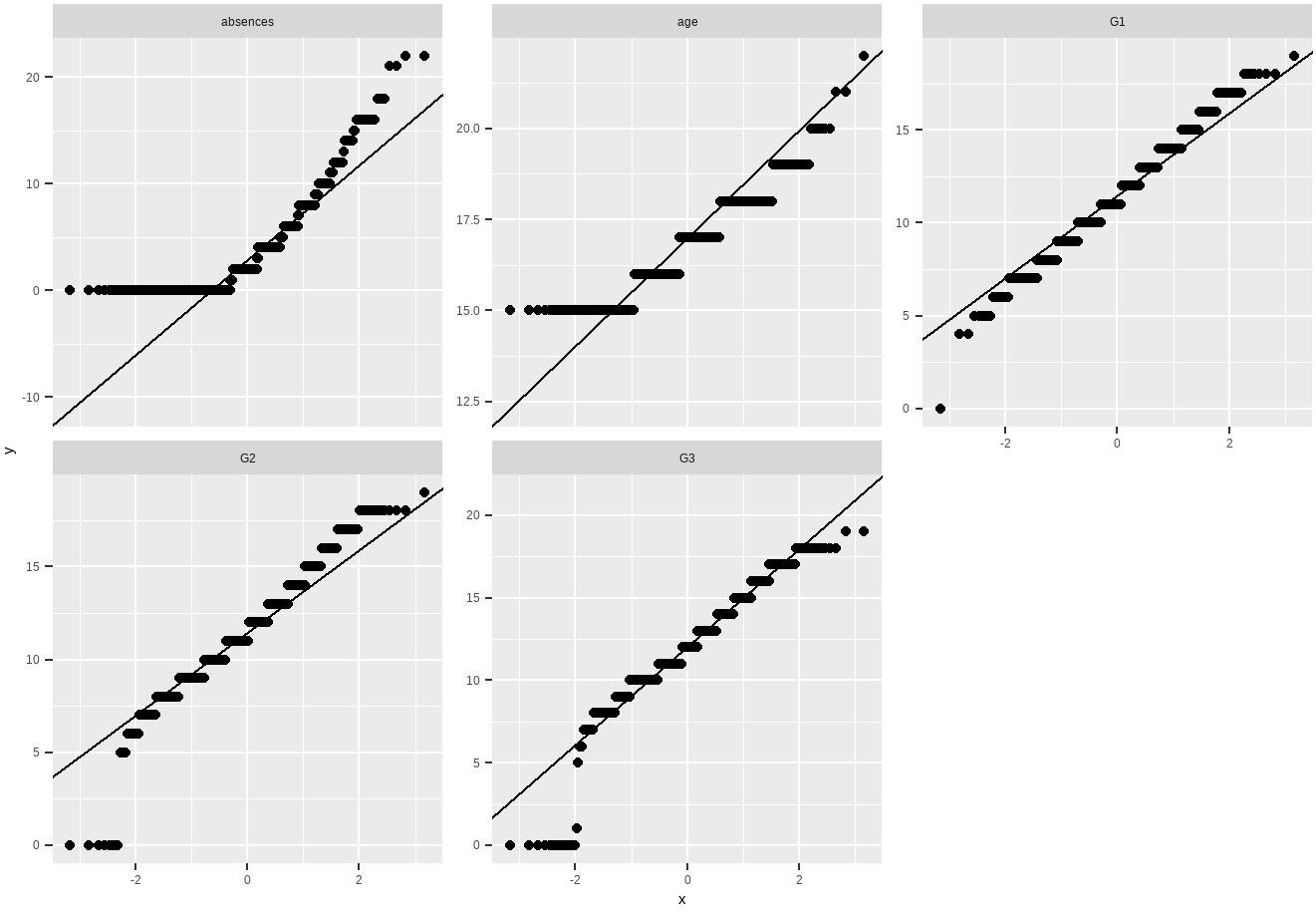
Remove Appropriate Observations

Skewness of Abscences

```
[1] 1.53647
```

After removing the four most extreme values, the absences variable still exhibits right skewness, but the level of skewness (1.54) is acceptable and no longer extreme.

Q-Q Plot



Although none of these variables have perfect normality, all variables have a majority of observations that fall close or on the Q-Q plot black line. Normality can be assumed as nothing drastic is happening to the distribution.

Look for Interaction Effects

Interaction effects occur when the impact of one predictor variable on the dependent variable depends on the level of another predictor variable. This is important to investigate because models such as linear regression assume that predictors operate independently. If strong interaction effects exist and are not accounted for models will suffer and produce power accuracy.

Lasso Model to look for Interaction Effects

To explore interaction effects, a Lasso regression model was used. By specifying the model formula as (.)^2, R expands the dataset to include all main effects and all possible two-way interaction terms. Lasso is well-suited for this task because it applies regularization that shrinks unimportant coefficients to zero, allowing meaningful interaction effects to stand out through non-zero coefficients.

```
# A tibble: 10 × 2
   term                              coef
   <chr>                            <dbl>
 1 Mjobteacher:schoolsupyes        -0.559
```

```
 2 Fjobteacher:traveltime>30 min   -0.382
 3 Mjobother:reasonother           -0.316
 4 schoolMS:healthHigh             -0.142
 5 reasonother:traveltime15–30 min -0.0483
 6 addressU:Dalc1                   0.0390
 7 failuresNot Failed:G1            0.0356
 8 schoolMS:sexM                   -0.0143
 9 higheryes:G1                     0.00470
10 age:G1                           0.00320
```

Based on the Lasso model results, several interaction terms were identified as potentially important because they retained non-zero coefficients. These include: Mjob:schoolsup, Mjob:Reason, G1:Failure, G2:Failure, and G1:G2. These interactions will be further examined by visualizing their relationships and evaluating model performance with and without these terms included.

Graph Interaction Effects

Mother's Job and School Up

```
# A tibble: 2 × 2
  Model                                 Adjusted_R2
  <chr>                                       <dbl>
1 Mjob * schoolsup (Interaction Model)      0.0427
2 Mjob + schoolsup (Additive Model)         0.0396
```

Including the interaction term in a regression model improves model fit, as evidenced by a higher adjusted R^2 compared to the model without the interaction. Therefore, this interaction effect will be included in the initial regression model.

Mother Job's and Reason

```
# A tibble: 2 × 2
  Model                               Adjusted_R2
  <chr>                                     <dbl>
1 Mjob * reason (Interaction Model)      0.0740
2 Mjob + reason (Additive Model)         0.0706
```

Including the interaction term in a regression model improves model fit, as evidenced by a higher adjusted R^2 compared to the model without the interaction. Therefore, this interaction effect will be included in the initial regression model.

G1 and Failure

```
# A tibble: 2 × 2
  Model                               Adjusted_R2
  <chr>                                     <dbl>
1 G1 * failures (Interaction Model)       0.691
2 G1 + failures (Additive Model)          0.692
```

Because the interaction reduces model performance, it will not be included in the final regression model.

G2 and Failure

```
# A tibble: 2 × 2
  Model                              Adjusted_R2
  <chr>                                    <dbl>
1 G2 * failures (Interaction Model)        0.845
2 G2 + failures (Additive Model)           0.845
```

Because the interaction reduces model performance, it will not be included in the final regression model.

G1 and G2

```
# A tibble: 2 × 2
  Model                          Adjusted_R2
  <chr>                                <dbl>
1 G1 * G2 (Interaction Model)          0.849
2 G1 + G2 (Additive Model)             0.846
```

Including the interaction term in a regression model improves model fit, as evidenced by a higher adjusted R^2 compared to the model without the interaction. Due to this the interaction effect between G1 and G2 will be included in the initial regression model.

Interaction Effects that will be used in the initial regression model is, Mother's Job: School up, Mother's Job: Reason, and G1: G2.

Correct Data Types

Data Partition: Regression Tree

This step ensures that all variables have the correct data types for regression tree modeling by converting any ordered factors into standard (unordered) factors and confirming that categorical variables are factors and numerical variables remain numeric.

```
Mean of G3
```

```
[1] 11.90698
```

```
 Mean of train set G3
```

```
[1] 11.95364
```

```
 Mean of test set G3
```

```
[1] 11.79688
```

The dataset was split into a 70% training set and a 30% testing set using random partitioning. To verify that the split did not introduce bias, the mean of G3 was compared across the full dataset, the training set, and the testing set. All three means are very similar, indicating that the partitioning preserved the overall distribution of G3 and that both subsets are representative of the original data.

Data Preparation: Linear Regression

```
Mean of G3

[1] 11.90698


 Mean of train set G3

[1] 11.88201


 Mean of test set G3

[1] 12.00781
```

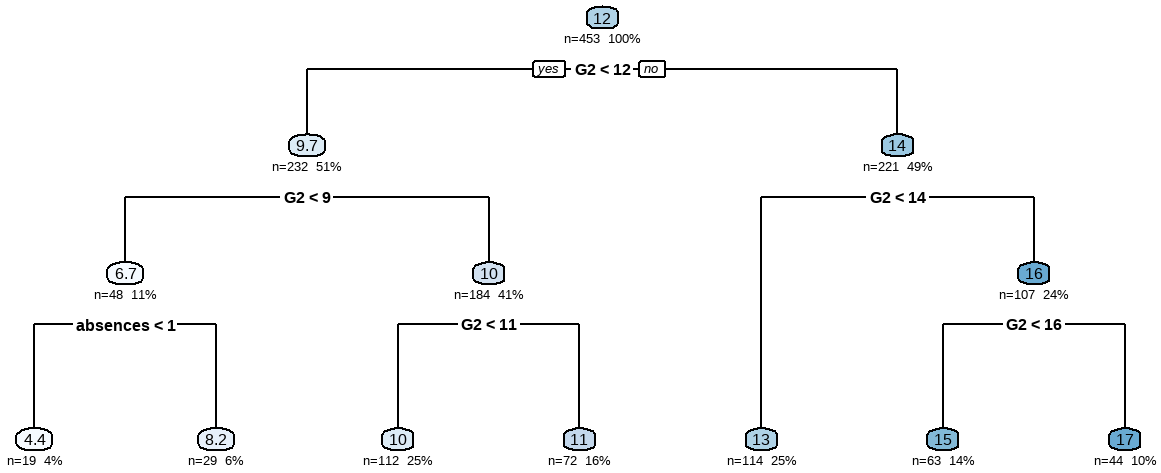
The dataset was split into an 80% training set and a 20% testing set using random partitioning based on the G3 variable. To confirm that this split did not introduce sampling bias, the mean of G3 was calculated for the full dataset, the training subset, and the testing subset. All three means are nearly identical, indicating that the train–test split preserved the overall distribution of G3. This suggests that both the training and testing sets are representative of the original data, making the partitioning appropriate for linear regression modeling.

Data Partition: Random Forest

Comments on Data Partitioning for the Random Forest:
The data partitioning step calculates the total number of predictor variables and uses that to determine a center value, the number of variables randomly sampled at each split. A small grid of values is then created around the center value(including one below, the center, one above, and boundary values), allowing the Random Forest model to test multiple candidates during tuning.

Model - Regression Tree

Default Tree



A default tree was made to see what variables are most important to predicting the values of G3, these variables being G2 and absences. G2 drives most of the splits in the default tree.

Full Tree

```
          CP nsplit rel error    xerror       xstd
1 0.52678032      0 1.0000000 1.0065851 0.10228374
2 0.11936029      1 0.4732197 0.4790166 0.06573209
3 0.10849877      2 0.3538594 0.4359840 0.06243761
4 0.04369563      3 0.2453606 0.2720445 0.04307994
5 0.02219416      4 0.2016650 0.2538961 0.04230471
6 0.01802516      5 0.1794708 0.2395513 0.04264682
```

```
          CP       nsplit    rel error       xerror         xstd
 0.004922738 10.000000000  0.100636869  0.233512682  0.049710415
```

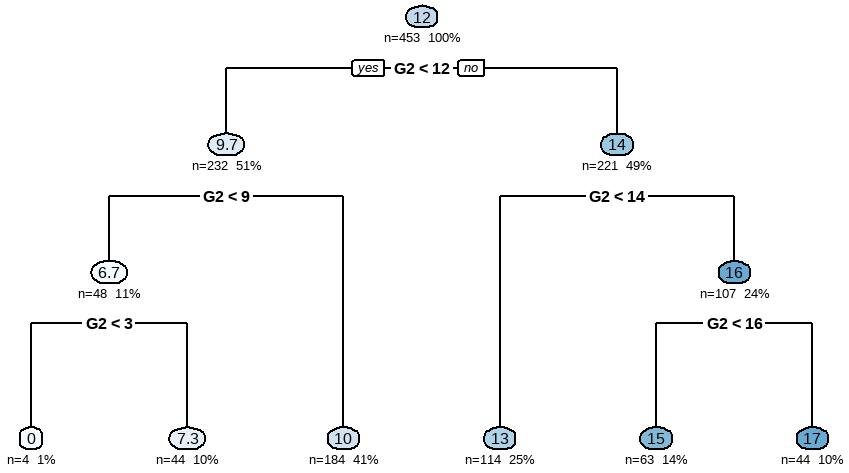
A full tree was made to produce a cptable to find the minimum error tree marked by the lowest “xerror”, the row with the minimum xerror value will be used to calculate the threshold value which will be used to prune the full tree.

Threshold

```
[1] 0.2832231

Best-Pruned Tree (1-SE Rule)

CP value: 0.04369563

Cross-validation error: 0.2720445

Number of splits: 3
```

The threshold value is determined by taking the minimum xerror and adding its corresponding xstd, creating an acceptable error range around the best-performing tree. Any tree whose xerror falls within this range is considered statistically indistinguishable from the optimal tree. This approach helps identify a set of candidate trees—ranging from simpler to more complex—that all perform comparably well. Based on this threshold, the most complex tree that still remains within the statistically acceptable range is the tree with ten splits.

Pruned Tree



Comments on Pruned Tree:
G2 accounts for all splits within the pruned tree. Due to this other predictors are not seen in how they may impact G3.

Best Tree

Create Best Tree

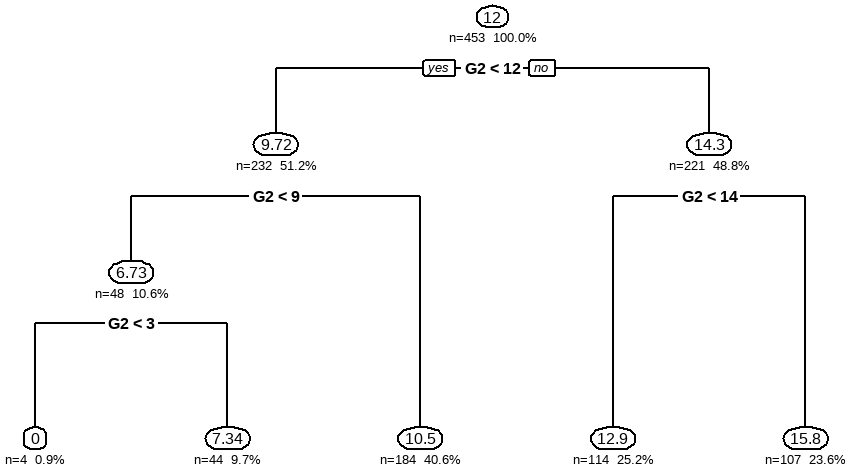
```
       RMSE  Rsquared       MAE Resample
1  1.248640 0.8011059 1.0188781   Fold05
2  1.368366 0.8257626 1.1230827   Fold02
3  1.442039 0.7940889 0.9053398   Fold06
4  1.469854 0.7327296 0.9958323   Fold03
5  1.506640 0.8634081 0.9818322   Fold09
6  1.537165 0.7239040 1.0206927   Fold04
7  1.583907 0.7783609 1.0949617   Fold08
8  1.677016 0.7109968 0.9766982   Fold01
9  1.877249 0.6980719 1.1101958   Fold10
10 1.969656 0.6839540 0.9435444   Fold07
```

Variable Importance

```
rpart variable importance

  only 20 most important variables shown (out of 48)

                     Overall
G2                   100.000
G1                    70.861
absences              18.558
failuresNot Failed    11.832
Mjobteacher            9.068
schoolMS               8.645
activitiesyes          7.740
higheryes              4.201
schoolsupyes           3.651
age                    2.404
paidyes                1.824
romanticyes            0.000
healthMedium           0.000
famsupyes              0.000
guardianother          0.000
PstatusT               0.000
`traveltime15-30 min`  0.000
gooutHigh              0.000
famsizeLE3             0.000
`FeduLow Education`    0.000
```

Comments on Variable Importance:
Variable Importance shows which variables have the most effect on G3. From this table it shows that G2, G1, Absences, and failures have the highest variable importance when relating to G3. G2 and G1 however are by far the strongest predictors within the Best Tree.

Print Tree

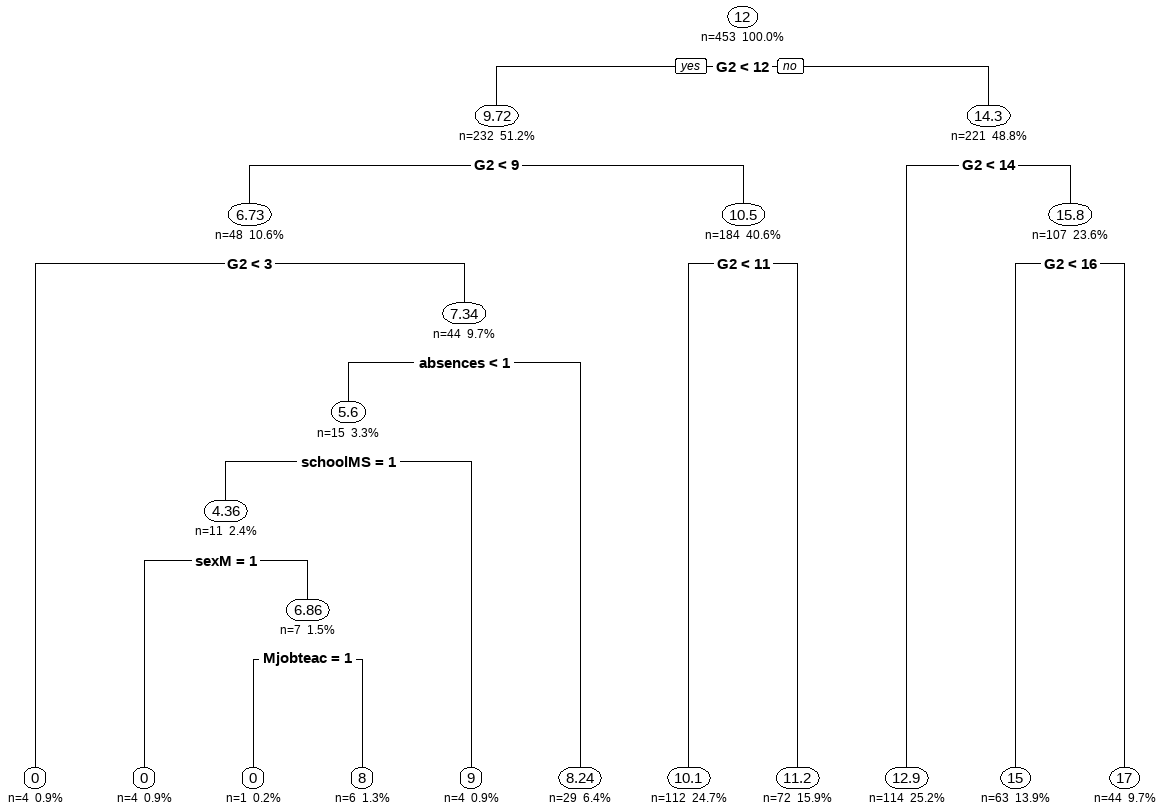


G2 takes up all the splits for the best tree, although this is helpful to know that G2 has the highest impact on G3 a more complex tree will be made to try and see how other predictors relate to G3.

Create More Complex Tree To See Further Splits

To make the most complex tree within statistical significance of the best tree, we will take the lowest cp value within the threshold range, the lowest cp value will render the most splits in the tree and give more opportunity for other predictors to be seen.

Print Complex Tree



Variable Importance Complex Tree

rpart variable importance

only 20 most important variables shown (out of 49)

	Overall
G2	100.0000
G1	75.2967
Mjobteacher	47.6850
sexM	32.3837
famrelGood	31.6760
activitiesyes	27.6272
absences	20.9934
schoolsupyes	20.3086
higheryes	20.3002
healthHigh	16.1091
schoolMS	15.9259
studytime2-5 Hours	14.6170
age	13.8151
famsizeLE3	11.2691
failuresNot Failed	10.9924
nurseryyes	1.7450
paidyes	1.3216

```
freetimeHigh         0.8332
reasonhome           0.7807
Dalc1                0.0000
```

Due to the increase in splits G2 is not the only predictor variable contributing to G3 in this tree. Although G2 still creates the most splits, Absences, School MS, sex, and Mother's that are teachers are predictors that now contribute to G3. It is surprising that the second most important variable,G1, does not contribute to a split within in the tree.

Evaluation: Regression Tree

```
             ME RMSE  MAE  MPE MAPE
Best Tree -0.06 1.53 1.01 -Inf  Inf

                ME RMSE  MAE  MPE MAPE
Complex Tree -0.14 1.44 0.81 -Inf  Inf
```

This evaluation compares the best tree to the complex tree. The complex tree has lower RMSE and MAE which means the complex tree predicts the values of G3 with slightly less average error than the best tree. MPE and MAPE appear as Inf/–Inf because G3 contains zero values, making percentage-based error measures undefined. Due to the complex tree having lower RMSE and MAE values this will be the chosen regression tree that will be compared to other regression models.

Model: Random Forest

```
Random Forest

453 samples
 32 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 407, 407, 408, 408, 408, 409, ...
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   MAE
   2    2.170987  0.6845458  1.4587455
  10    1.400154  0.8225581  0.8735182
  11    1.378205  0.8255381  0.8652854
  12    1.357500  0.8299388  0.8497624
  32    1.278574  0.8405733  0.8384674

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 32.
```

A Random Forest model was trained using 10-fold cross-validation to predict the final grade (G3). The model evaluated multiple mtry values—the number of predictors randomly sampled at each split—to identify the configuration that produced the lowest prediction error. Based on RMSE, the optimal model

used mtry = 32, indicating this setting achieved the best balance of predictive accuracy and model stability. The model was trained using 1,000 trees, and variable importance was recorded for interpretation in later steps.

Variable Importance

```
rf variable importance

  only 20 most important variables shown (out of 47)

                   Overall
G2                 100.000
G1                  33.269
absences            17.776
Mjobteacher         15.416
schoolMS            10.823
famrelGood          10.149
higheryes            9.850
sexM                 9.795
age                  9.702
MeduLow Education    9.253
failuresNot Failed   8.534
FeduLow Education    8.169
activitiesyes        7.607
schoolsupyes         7.411
healthHigh           6.649
reasonhome           5.978
Fjobservices         5.969
Fjobteacher          5.634
studytime2-5 Hours   5.442
WalcMedium           5.405
```

Identical to the Best tree, the most important variables for the random forest model are G2, G1, absences, and failures, with G2 and G1 being the most important variables relating to G3.

Evaluation Metrics: Random Forest

```
                ME RMSE  MAE  MPE MAPE
Random Forest 0.04 1.24 0.79 -Inf  Inf
```

The random forest has an RMSE value of 1.26 and MAE value of 0.78. The random forest will be compared to the best performing regression tree and linear regression model in the evaluation phase.

Model: Linear Regression- Assume a 10% Level of Significance

Cross Validation Control

Regression Model: All variables

```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7824 -0.4801  0.0259  0.6217  4.8106

Coefficients:
                             Estimate Std. Error t value
(Intercept)                 -2.855814   1.573539  -1.815
schoolMS                    -0.038057   0.162257  -0.235
sexM                        -0.074597   0.142685  -0.523
age                          0.050835   0.060101   0.846
addressU                     0.155506   0.146792   1.059
famsizeLE3                  -0.065075   0.136890  -0.475
PstatusT                    -0.226308   0.204916  -1.104
`MeduLow Education`          0.161562   0.164787   0.980
`FeduLow Education`          0.099189   0.156349   0.634
Mjobhealth                   0.433855   0.474281   0.915
Mjobother                   -0.153877   0.249477  -0.617
Mjobservices                 0.304384   0.297779   1.022
Mjobteacher                  0.426371   0.359964   1.184
Fjobhealth                  -0.524733   0.430109  -1.220
Fjobother                   -0.253083   0.248043  -1.020
Fjobservices                -0.394972   0.263898  -1.497
Fjobteacher                 -0.396251   0.374036  -1.059
reasonhome                  -0.051389   0.359796  -0.143
reasonother                 -0.183580   0.381768  -0.481
reasonreputation            -0.203416   0.426013  -0.477
guardianmother              -0.142665   0.151175  -0.944
guardianother                0.175951   0.310560   0.567
`traveltime15–30 min`       -0.049418   0.140508  -0.352
`traveltime>30 min`          0.418330   0.217559   1.923
`studytime2–5 Hours`         0.184374   0.152382   1.210
`studytime>5 Hours`          0.149310   0.190127   0.785
`failuresNot Failed`         0.566954   0.207986   2.726
schoolsupyes                -0.322216   0.465097  -0.693
famsupyes                    0.137694   0.129912   1.060
paidyes                     -0.175068   0.266546  -0.657
activitiesyes               -0.028836   0.127803  -0.226
nurseryyes                  -0.093607   0.149681  -0.625
higheryes                    0.001196   0.216313   0.006
internetyes                  0.074001   0.156364   0.473
```

```
romanticyes                      -0.109695   0.128742  -0.852
famrelGood                       -0.117085   0.145394  -0.805
freetimeMedium                   -0.234432   0.165354  -1.418
freetimeHigh                     -0.128431   0.175406  -0.732
gooutMedium                       0.108188   0.156196   0.693
gooutHigh                        -0.092874   0.175907  -0.528
Dalc1                             0.122557   0.166813   0.735
WalcMedium                       -0.033835   0.150315  -0.225
WalcHigh                          0.126017   0.227393   0.554
healthMedium                     -0.049859   0.159390  -0.313
healthHigh                       -0.195991   0.163099  -1.202
absences                          0.019498   0.015538   1.255
G1                                0.342456   0.105966   3.232
G2                                1.046334   0.077510  13.499
`Mjobhealth:schoolsupyes`         0.932344   1.529265   0.610
`Mjobother:schoolsupyes`          0.334447   0.540655   0.619
`Mjobservices:schoolsupyes`       0.512527   0.587270   0.873
`Mjobteacher:schoolsupyes`       -2.080219   0.854638  -2.434
`Mjobhealth:reasonhome`          -0.400603   0.752772  -0.532
`Mjobother:reasonhome`            0.189939   0.431587   0.440
`Mjobservices:reasonhome`        -0.150296   0.500766  -0.300
`Mjobteacher:reasonhome`         -0.567437   0.558131  -1.017
`Mjobhealth:reasonother`         -0.041687   0.792529  -0.053
`Mjobother:reasonother`          -0.764154   0.507544  -1.506
`Mjobservices:reasonother`       -0.199220   0.572639  -0.348
`Mjobteacher:reasonother`         0.690072   0.750776   0.919
`Mjobhealth:reasonreputation`    -0.027198   0.697956  -0.039
`Mjobother:reasonreputation`      0.120290   0.492570   0.244
`Mjobservices:reasonreputation` -0.106864   0.534029  -0.200
`Mjobteacher:reasonreputation`    0.514072   0.621179   0.828
`G1:G2`                          -0.015970   0.006542  -2.441
                                            Pr(>|t|)
(Intercept)                                  0.07020 .
schoolMS                                     0.81466
sexM                                         0.60136
age                                          0.39810
addressU                                     0.29000
famsizeLE3                                   0.63475
PstatusT                                     0.27001
`MeduLow Education`                          0.32740
`FeduLow Education`                          0.52614
Mjobhealth                                   0.36080
Mjobother                                    0.53768
Mjobservices                                 0.30724
Mjobteacher                                  0.23685
Fjobhealth                                   0.22310
Fjobother                                    0.30812
Fjobservices                                 0.13517
Fjobteacher                                  0.28999
reasonhome                                   0.88649
reasonother                                  0.63084
```

```
reasonreputation                          0.63325
guardianmother                            0.34582
guardianother                             0.57129
`traveltime15–30 min`                     0.72522
`traveltime>30 min`                       0.05513 .
`studytime2–5 Hours`                      0.22693
`studytime>5 Hours`                       0.43268
`failuresNot Failed`                      0.00666 **
schoolsupyes                              0.48880
famsupyes                                 0.28975
paidyes                                   0.51164
activitiesyes                             0.82159
nurseryyes                                0.53204
higheryes                                 0.99559
internetyes                               0.63626
romanticyes                               0.39464
famrelGood                                0.42107
freetimeMedium                            0.15695
freetimeHigh                              0.46443
gooutMedium                               0.48889
gooutHigh                                 0.59778
Dalc1                                     0.46291
WalcMedium                                0.82201
WalcHigh                                  0.57973
healthMedium                              0.75457
healthHigh                                0.23012
absences                                  0.21020
G1                                        0.00132 **
G2                                 < 0.0000000000000002 ***
`Mjobhealth:schoolsupyes`                 0.54239
`Mjobother:schoolsupyes`                  0.53649
`Mjobservices:schoolsupyes`               0.38327
`Mjobteacher:schoolsupyes`                0.01532 *
`Mjobhealth:reasonhome`                   0.59487
`Mjobother:reasonhome`                    0.66008
`Mjobservices:reasonhome`                 0.76421
`Mjobteacher:reasonhome`                  0.30985
`Mjobhealth:reasonother`                  0.95807
`Mjobother:reasonother`                   0.13287
`Mjobservices:reasonother`                0.72808
`Mjobteacher:reasonother`                 0.35851
`Mjobhealth:reasonreputation`             0.96893
`Mjobother:reasonreputation`              0.80718
`Mjobservices:reasonreputation`           0.84149
`Mjobteacher:reasonreputation`            0.40835
`G1:G2`                                   0.01502 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.314 on 452 degrees of freedom
```

```
Multiple R-squared:  0.8612,    Adjusted R-squared:  0.8415
F-statistic: 43.81 on 64 and 452 DF,  p-value: < 0.00000000000000022
```

Based off the initial regression model I will only keep the varialbes G1:G2, G1, G2, Mjobteacher:schoolsupyes, failures, and travel time. All other variables will be dropped due to their low significance in the model. Another regression model will be made with the variables that were seen significant in this model.

Regression Model: Only Significant Variables

```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0892 -0.5104  0.0028  0.6254  5.0779

Coefficients:
                            Estimate Std. Error t value             Pr(>|t|)
(Intercept)                -2.806667   0.799019  -3.513             0.000484
G1                          0.387526   0.093898   4.127             0.000043
G2                          1.083163   0.071425  15.165 < 0.0000000000000002
Mjobhealth                  0.063351   0.259620   0.244             0.807320
Mjobother                  -0.232076   0.169556  -1.369             0.171698
Mjobservices                0.056104   0.195482   0.287             0.774228
Mjobteacher                 0.146916   0.225886   0.650             0.515733
schoolsupyes               -0.300904   0.439359  -0.685             0.493743
`failuresNot Failed`        0.456346   0.191026   2.389             0.017267
`traveltime15-30 min`      -0.010829   0.129987  -0.083             0.933640
`traveltime>30 min`         0.320290   0.200511   1.597             0.110814
`G1:G2`                    -0.019023   0.005838  -3.258             0.001197
`Mjobhealth:schoolsupyes`   0.816763   1.419215   0.576             0.565209
`Mjobother:schoolsupyes`    0.374975   0.517446   0.725             0.468995
`Mjobservices:schoolsupyes` 0.392292   0.566636   0.692             0.489059
`Mjobteacher:schoolsupyes` -2.246900   0.810134  -2.773             0.005753

(Intercept)                ***
G1                         ***
G2                         ***
Mjobhealth
Mjobother
Mjobservices
Mjobteacher
schoolsupyes
`failuresNot Failed`       *
`traveltime15-30 min`
`traveltime>30 min`
`G1:G2`                    **
`Mjobhealth:schoolsupyes`
```

```
`Mjobother:schoolsupyes`
`Mjobservices:schoolsupyes`
`Mjobteacher:schoolsupyes`  **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.315 on 501 degrees of freedom
Multiple R-squared:  0.8458,    Adjusted R-squared:  0.8412
F-statistic: 183.2 on 15 and 501 DF,  p-value: < 0.00000000000000022
```

Compare Default Model and Refined Model

```
     adjr2_default adjr2_refined
[1,]     0.8415114     0.8411704
```

The Refined regression model while having significantly less variables only had a slight drop in adjusted R squared.

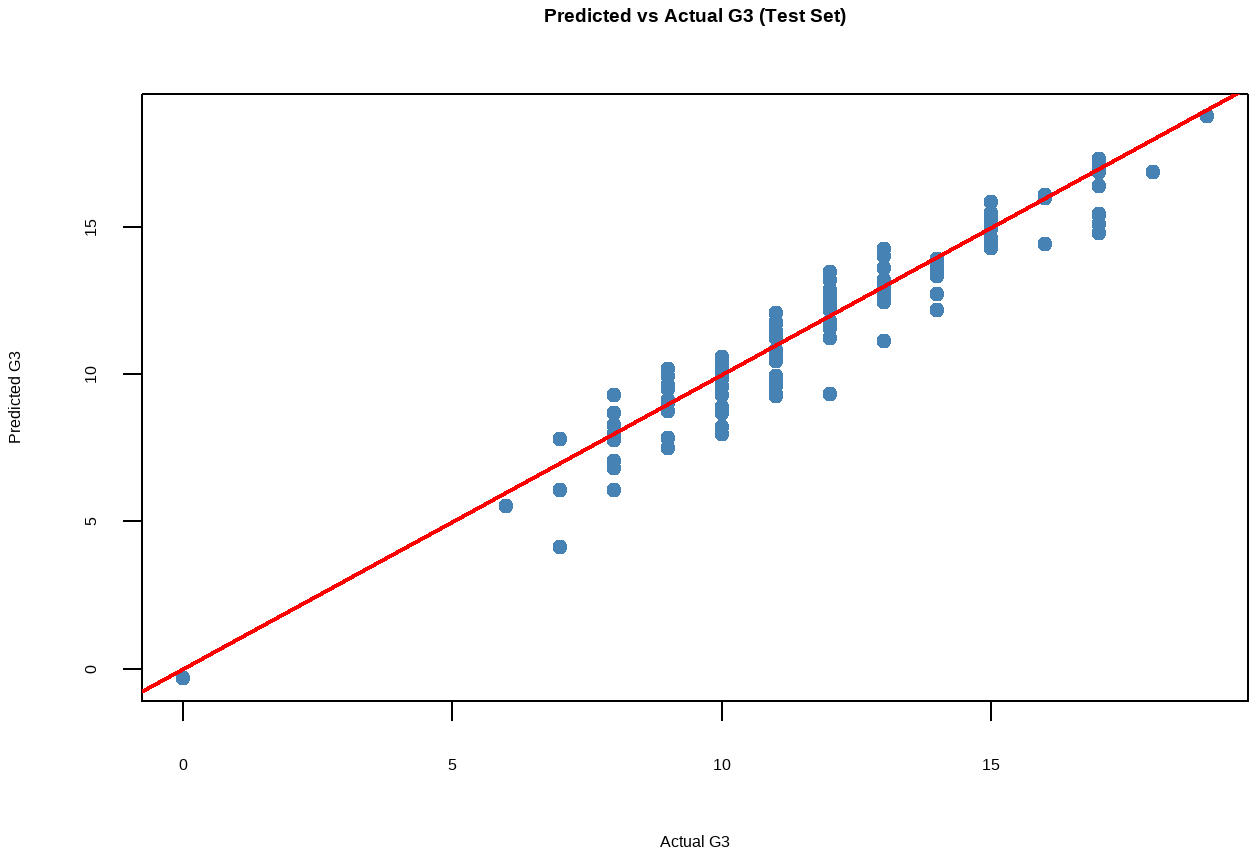
Evaluation Metrics: Linear Regression

```
          Model      RMSE      MAE       MAD MAPE
1 Refined Model 0.9128109 0.697347 0.7202631  Inf
```

The Linear Regression Model has an RMSE value of 0.91 and MAE value of 0.697. The refiend linear regression model will be compared to the best performing regression tree and random forest model in the evaluation phase.

Predict G3 Numbers

Predicted vs Actual G3 Values



The Predicted vs. Actual G3 plot shows that the regression model performs reasonably well on the test set. Most of the points fall close to the 45-degree reference line, indicating that the predicted G3 values closely match the actual scores. Although some deviation exists—particularly at lower and mid-range G3 values—the overall pattern suggests that the model captures the general trend of student performance. The alignment along the diagonal line reflects good predictive accuracy and indicates that the model generalizes well to new, unseen data.

Evaluation

Compare best Prediction Models

Model Performance Comparison (ME, RMSE, MAE)

Model	ME	RMSE	MAE
Refined Linear Regression	-0.28	0.91	0.70
Random Forest	0.04	1.24	0.79
Complex Tree	-0.14	1.44	0.81

The table compares the three best models- refined liner regression model, the random forest, and the complex tree -from the modeling phase. These models are compared using ME (Mean Error), RMSE (Root

Mean Squared Error), and MAE (Mean Absolute Error). Across all three of these evaluation metrics, the refined linear regression model performs the best in each one. The refined linear regression model having the smallest RMSE means it has the smallest average prediction error magnitude, making it the most accurate overall. The refined linear regression model having the Lowest MAE means the model's typical error in predicting a student's final grade is less than one point. And the ME value being close to zero shows that the model does not systematically overpredict or underpredict, meaning bias is minimal. Overall, the refined linear regression is the best out of the prediction models and will be the model used to answer the business problem and research questions.

Deployment

Answer to the Business Problem

The business problem guiding this analysis was to determine the final grades of students based off different variables. Based on the modeling and evaluation process, the refined linear regression model should be used as the tool to answer this business question. The refined linear regression model showed small average prediction error, with each prediction being within less than one point of the students grade. This model also showed minimal bias making it a perfect tool for school's to use.

Two Research Questions

1. Which factors influenced G3 the most?

The refined linear regression model identified several strong predictors of a student's final grade (G3). The most influential factors were G1 and G2, the student's earlier grading periods, which showed the strongest positive relationship with the final score. Additionally, an interaction effect: students whose mothers work as teachers and who also receive extra educational support tend to exhibit different performance patterns, indicating that the combination of home academic guidance and supplemental school support tend to lead to higher G3 values.

2. Does parental education level influence students' academic performance, as reflected in their G3 scores?

Based on the refined linear regression model, neither the mother's nor the father's level of education had a statistically significant effect on the student's final grade (G3). This suggests that, within this dataset, parental education does not meaningfully influence final academic performance.

Business Recommendations

If schools were to implement the refined linear regression model, three key recommendations emerge. First, prioritize early grade monitoring. Since G1 and G2 were the strongest predictors of G3, schools should focus on identifying students with low early-term performance and intervening before final grades are determined. Second, expand and strengthen educational assistance programs. The analysis showed that

students receiving extra academic support tend to achieve higher final grades, demonstrating the value of structured tutoring, supplemental instruction, and targeted academic resources. Finally, schools should develop support plans that help replicate aspects of a strong home academic environment. Students whose mothers worked as teachers tended to perform better, suggesting that guided academic support and structured study environments matter. Schools can help by offering supervised study halls and mentorship programs for students whose parents may be unable to provide academic assistance at home.

References

Citation of orignial data source with authors

Cortez, P., & Silva, A. (2014). Student Performance [Data set]. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/320/student+performance.

Citation of ChatGPT

OpenAI. (2025). ChatGPT (Version 5.1) [Large language model]. https://chat.openai.com/

R-Version citation

```
Version R version 4.5.2 (2025-10-31 ucrt)
```

Citation for all R packages

```
H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Brandon M. Greenwell (2017). pdp: An R Package for Constructing Partial Dependence Plots. The R Journal, 9(1), 421--436. URL https://journal.r-project.org/archive/2017/RJ-2017-016/index.html.
```