UNDERSTANDING STUDENT PERFORMANCE

Authors: Sri Anirudh Reddy , Parimalnath Reddy.

Date: 04-15-2024

Executive Summary

This document presents a detailed analysis aimed at identifying key factors affecting student performance in Portuguese secondary schools. Utilizing a comprehensive dataset from the UCI Machine Learning Repository, this study employs statistical methods to explore the impact of various socio-economic, demographic, and educational factors on academic outcomes. The findings are intended to support educational policy makers and school administrators in making informed decisions to enhance student performance.

Table of Contents

1. Introduction

Purpose of the Study: The objective of this analysis is to uncover the underlying factors that significantly impact student performance, focusing on socio-economic indicators, parental involvement, and resource availability.
Background: Previous research indicates that student performance is influenced by a complex interplay of individual and institutional factors. This study builds on this foundation with a robust dataset to provide nuanced insights into these dynamics.
Scope of the Report: The report details the methodology used for data analysis, presents findings, discusses their implications, and offers evidence-based recommendations.

2. Data Overview

The dataset includes information from 650 students attending two different secondary schools in Portugal. It comprises 33 variables, categorized into binary, numeric, and nominal types, which provide a broad spectrum of data ranging from demographic details to academic performance metrics.

Variables Description:

- **Binary Variables:** These include sex (male or female), school (school GP or MS), and internet (yes or no), among others, indicating dichotomous attributes of the students.
- **Numeric Variables:** These encompass age, absences (number of school absences), and G1 to G3 (grades over three periods), offering quantitative measures that are essential for performance analysis.
- **Nominal Variables:** These include Mjob (mother's job) and Fjob (father's job), providing qualitative data without a quantitative order but with potential influence on student outcomes

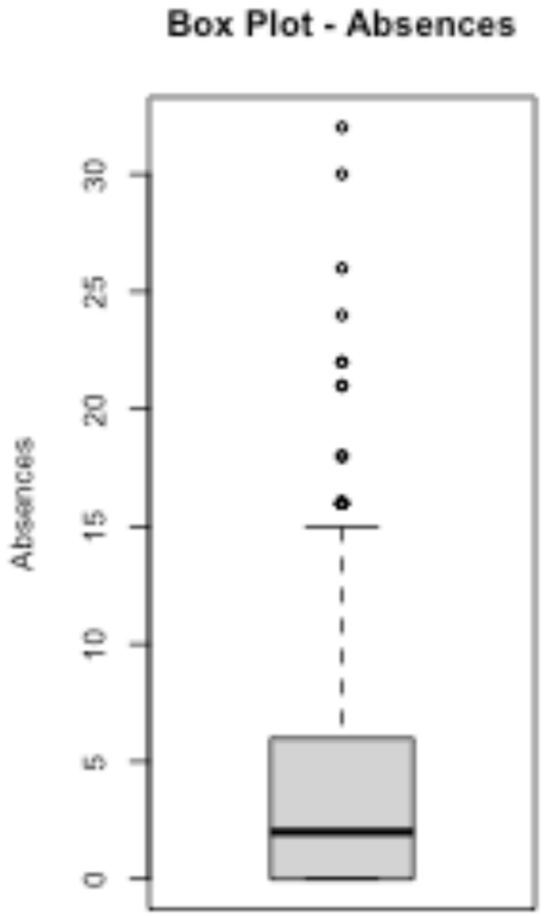
```
> binary_columns <- c("sex", "school", "address", "Pstatus", "famsize", "schoolsup",
"famsup","activities", "paid", "internet", "nursery", "higher", "romantic")
> numeric_columns <- c("age", "Medu", "Fedu", "traveltime", "studytime", "failures",
"famrel", "freetime", "goout", "Dalc", "Walc", "health", "absences")
> nominal_columns <- c("Mjob", "Fjob", "reason", "guardian")
```

Preprocessing Steps:

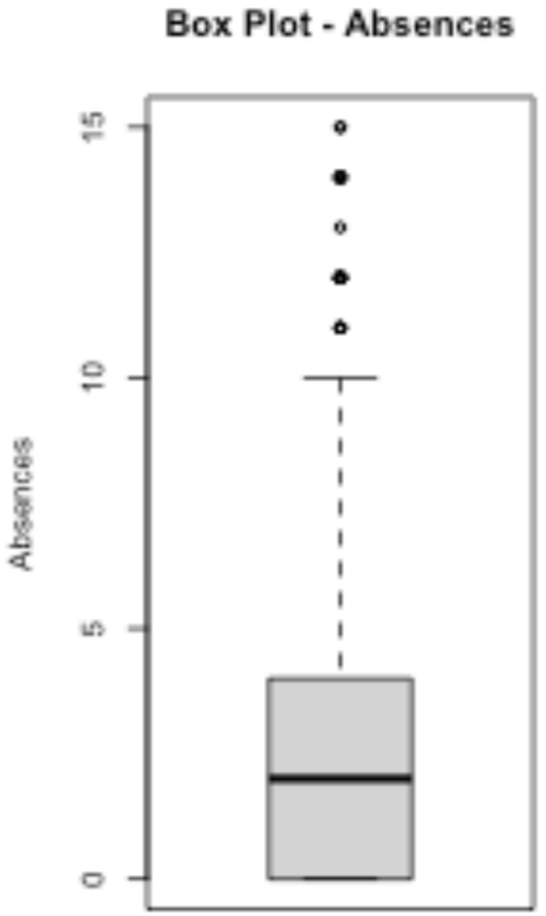
Outliers in the **absences** data were identified using the interquartile range method and replaced with the mean value to prevent skewed results. Grades were transformed into percentiles to normalize these figures across different exams, facilitating comparisons and trend identification.

3. Methodology:

- **Handling Outliers:** Outliers in absences were mitigated by capping values using predefined bounds based on the interquartile range, ensuring a more robust analysis.



Before



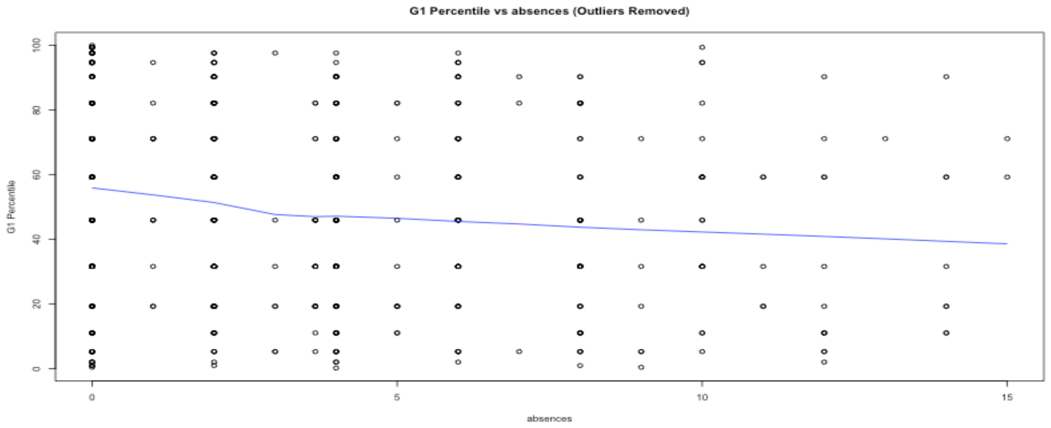
After

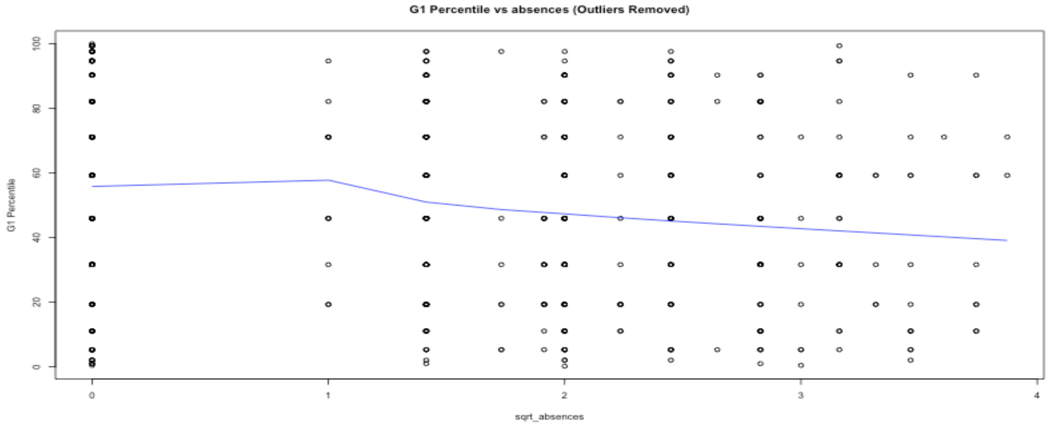
```
> #removing outliers from absences
> # Calculate the quartiles
> q1 <- quantile(d1[["absences"]], 0.25)
> q3 <- quantile(d1[["absences"]], 0.75)
>
> # Calculate the interquartile range (IQR)
> iqr <- q3 - q1
>
> # Define the lower and upper bounds for outliers
> lower_bound <- q1 - 1.5 * iqr
> upper_bound <- q3 + 1.5 * iqr
>
> # Replace outliers with mean value
> d1[["absences"]][d1[["absences"]] < lower_bound | d1[["absences"]] >
upper_bound] <- mean(d1[["absences"]], na.rm = TRUE)
```

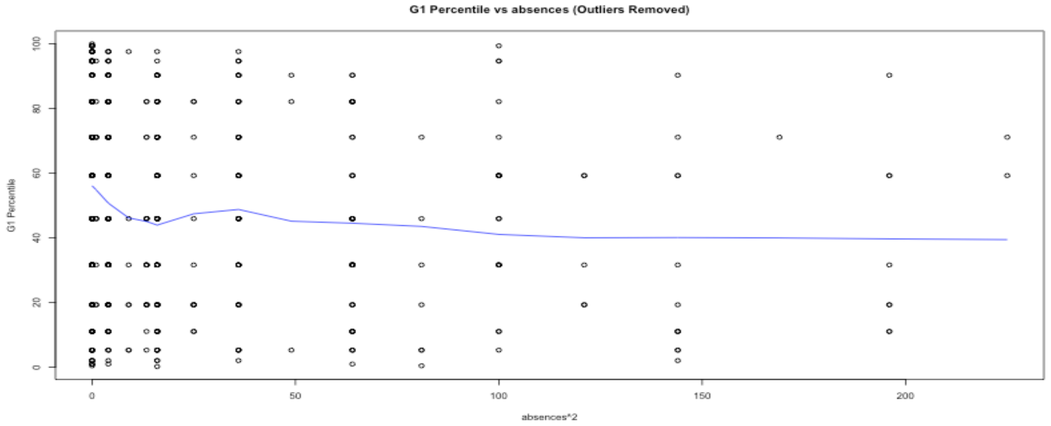
- **Data Transformation:**

Converting raw score marks into percentiles is a crucial step for normalizing the grades across different tests, making them comparable. This transformation is performed for three different grade variables (G1, G2, G3).

```
> #converts marks into percentile for improving data quality
> d1$G1_perc <- rank(d1$G1) / length(d1$G1) * 100
> d1$G2_perc <- rank(d1$G2) / length(d1$G2) * 100
> d1$G3_perc <- rank(d1$G3) / length(d1$G3) * 100
```



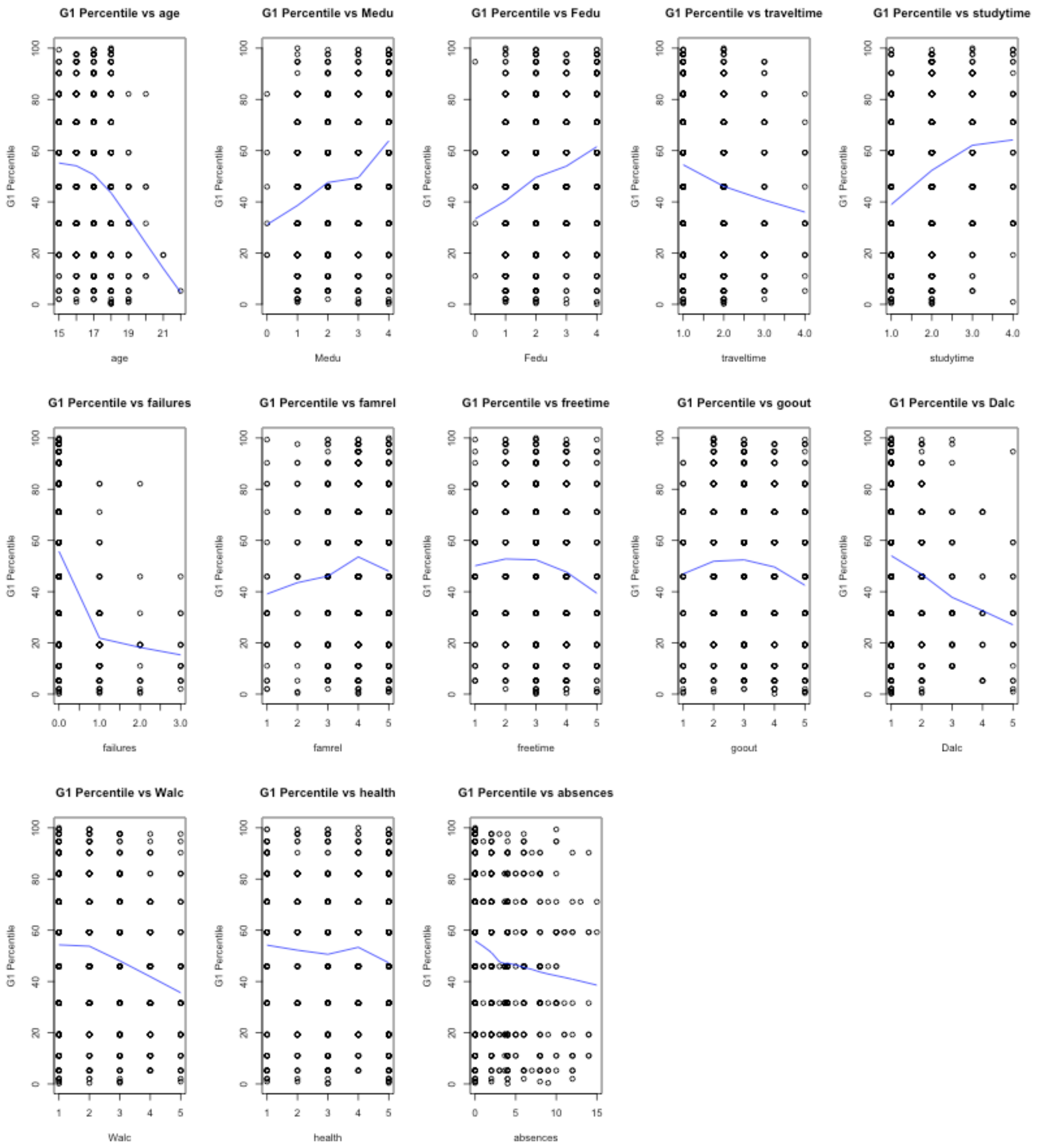




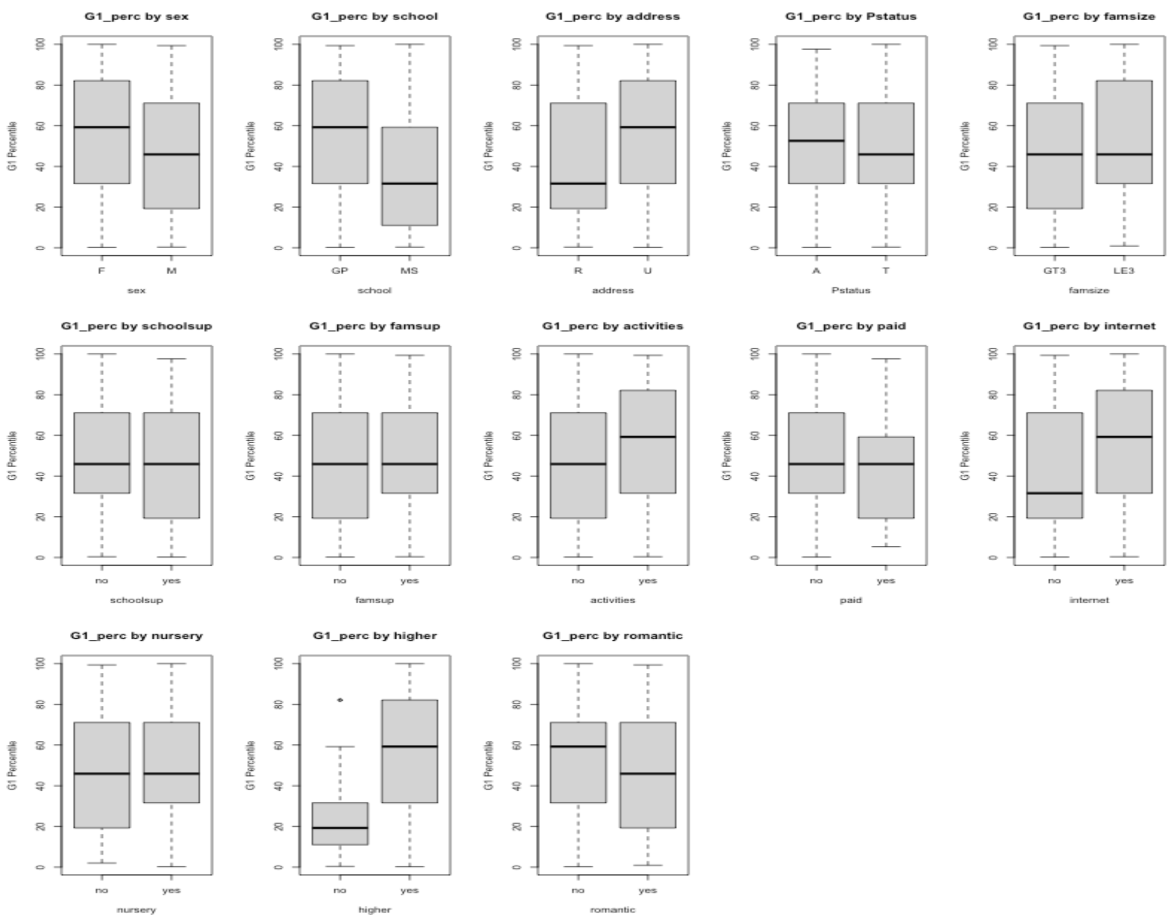
Exploratory Data Analysis: Visualizing Data

Multiple plots are generated to visually explore the relationships between numerical variables and the first-period grades percentile (G1_perc). Scatter plots with a lowess line (locally weighted scatterplot smoothing) are used to identify trends and patterns.

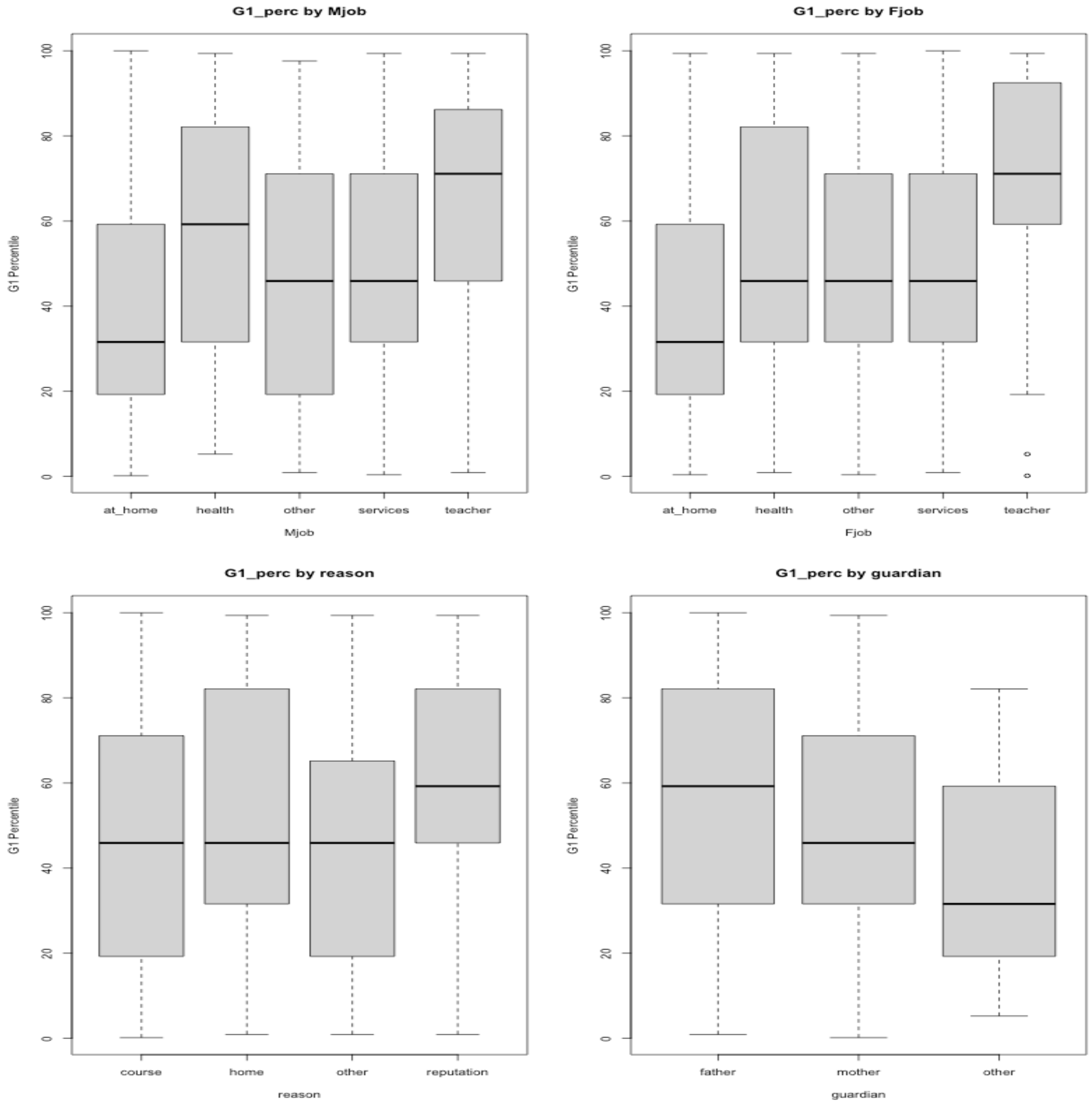
```
+ # Set up plotting area to accommodate all boxplots
+ par(mfrow=c(4, 5)) # Adjust as needed based on the number of binary variables
+
+ for (var in binary_columns) {
+   # Assuming the binary variables are already encoded as 0 and 1 in the data
+   # Create a temporary factor variable for plotting
+   factor_var <- factor(d1[[var]])
+
+   # Now draw the box plot
+   boxplot(G1_perc ~ d1[[var]], data=d1, main=paste("G1_perc by", var), xlab=var, ylab="G1 Percentile")
+ }
```



```
+ par(mfrow=c(4, 5)) # Adjust as needed based on the number of binary variables
+
+ for (var in binary_columns) {
+   # Assuming the binary variables are already encoded as 0 and 1 in the data
+   # Create a temporary factor variable for plotting
+   factor_var <- factor(d1[[var]])
+
+   # Now draw the box plot
+   boxplot(G1_perc ~ d1[[var]], data=d1, main=paste("G1_perc by", var), xlab=var,
ylab="G1 Percentile")
+ }
```



```
+ # Set up plotting area to accommodate all boxplots
+ par(mfrow=c(2, 2)) # Adjust as needed based on the number of binary variables
+
+ for (var in nominal_columns) {
+   # Assuming the binary variables are already encoded as 0 and 1 in the data
+   # Create a temporary factor variable for plotting
+   factor_var <- factor(d1[[var]])
+
+   # Now draw the box plot
+   boxplot(G1_perc ~ d1[[var]], data=d1, main=paste("G1_perc by", var), xlab=var,
ylab="G1 Percentile")
+ }
```



Linear Regression Modeling:

Objective: The goal here is to understand how various predictors (like failures, parental education, study time, etc.) influence the students' grade percentiles (G1 percentile used here).

Model Setup and Execution:

- **Model Construction**: The script constructs several linear regression models. Each model varies by the complexity and the predictors included. Initially, a model with selected predictors is built to see their individual impact on the G1 percentile.
- **Polynomial Terms**: To capture non-linear relationships, polynomial terms (e.g., square roots) are included for some predictors like absences. This approach helps in modeling more complex patterns that linear terms cannot capture effectively.

Linear Regression Models

```
> lm_model <- lm(G1_perc ~ failures + schoolsup + Medu + Fedu + studytime + goout +
Walc + traveltime + Dalc +famsup +reason+higher+Fjob+sex+absences, data =
new_df_for_lm)
> summary(lm_model)

Call:
lm(formula = G1_perc ~ failures + schoolsup + Medu + Fedu + studytime +
    goout + Walc + traveltime + Dalc + famsup + reason + higher +
    Fjob + sex + absences, data = new_df_for_lm)

Residuals:
    Min      1Q  Median      3Q     Max
-70.989 -17.625   0.439  18.512  50.902

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   24.79707    6.53405   3.795 0.000162 ***
failures     -11.47463    1.71098  -6.706 4.45e-11 ***
schoolsupyes -11.08163    3.13546  -3.534 0.000439 ***
Medu           2.53018    1.13184   2.235 0.025738 *
Fedu           1.26222    1.20239   1.050 0.294235
studytime      4.93236    1.22177   4.037 6.08e-05 ***
goout          0.07911    0.87184   0.091 0.927733
Walc          -0.45361    1.02305  -0.443 0.657636
traveltime    -2.44808    1.33077  -1.840 0.066300 .
```

```
Dalc             -2.82479    1.32626  -2.130 0.033569 *
famsupyes        -1.82938    1.99484  -0.917 0.359466
reasonhome        3.59908    2.48514   1.448 0.148048
reasonother      -2.01315    3.21037  -0.627 0.530837
reasonreputation   6.27082    2.54051   2.468 0.013840 *
higheryes        17.66263    3.32608   5.310 1.52e-07 ***
Fjobhealth        3.36276    6.47354   0.519 0.603621
Fjobother         6.32002    3.94003   1.604 0.109206
Fjobservices      4.25919    4.16683   1.022 0.307096
Fjobteacher      16.05423    5.82581   2.756 0.006026 **
sexM             -3.52524    2.10389  -1.676 0.094318 .
absences         -0.61009    0.27588  -2.211 0.027365 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.79 on 628 degrees of freedom
Multiple R-squared:  0.3347,    Adjusted R-squared:  0.3136
F-statistic:  15.8 on 20 and 628 DF,  p-value: < 2.2e-16

>
```

```
> lm_model_full <- lm(G1_perc ~ ., data = new_df_for_lm)
> summary(lm_model_full)

Call:
lm(formula = G1_perc ~ ., data = new_df_for_lm)

Residuals:
    Min      1Q  Median      3Q     Max
-69.642 -16.093  -0.461  16.406  63.621

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     47.7527    17.2855   2.763  0.00591 **
schoolMS       -11.8334     2.3088  -5.125 3.99e-07 ***
sexM            -5.8865     2.1761  -2.705  0.00702 **
age             -0.6107     0.8904  -0.686  0.49305
addressU         1.2993     2.2802   0.570  0.56902
```

```
famsizeLE3        2.7150     2.1375   1.270  0.20452
PstatusT          0.5464     3.0154   0.181  0.85626
Medu              1.4628     1.3184   1.110  0.26764
Fedu              1.0498     1.2008   0.874  0.38233
Mjobhealth        4.0558     4.6761   0.867  0.38609
Mjobother         2.6420     2.6388   1.001  0.31714
Mjobservices      3.9456     3.2506   1.214  0.22530
Mjobteacher       4.1444     4.3697   0.948  0.34328
Fjobhealth       -0.3264     6.5503  -0.050  0.96027
Fjobother         3.7064     3.9674   0.934  0.35056
Fjobservices      1.1112     4.1698   0.266  0.78995
Fjobteacher      12.5062     5.8648   2.132  0.03337 *
reasonhome        2.2312     2.4747   0.902  0.36762
reasonother       1.2396     3.2024   0.387  0.69884
reasonreputation  3.3609     2.5944   1.295  0.19565
guardianmother   -4.2413     2.3088  -1.837  0.06670 .
guardianother    -3.7596     4.6191  -0.814  0.41601
traveltime       -0.9553     1.3863  -0.689  0.49100
studytime         4.2893     1.2137   3.534  0.00044 ***
failures         -9.9866     1.7809  -5.608 3.11e-08 ***
schoolsupyes    -13.5090     3.1663  -4.267 2.30e-05 ***
famsupyes        -1.7416     1.9873  -0.876  0.38117
paidyes          -5.8400     4.0137  -1.455  0.14617
activitiesyes     3.2794     1.9459   1.685  0.09244 .
nurseryyes       -0.4629     2.3638  -0.196  0.84481
higheryes        15.9253     3.3322   4.779 2.21e-06 ***
internetyes       0.6205     2.4052   0.258  0.79650
romanticyes      -2.2735     1.9957  -1.139  0.25507
famrel            0.3824     1.0139   0.377  0.70619
freetime         -0.7942     0.9809  -0.810  0.41843
goout             0.3324     0.9375   0.355  0.72305
Dalc             -2.4395     1.3298  -1.834  0.06707 .
Walc             -0.5590     1.0317  -0.542  0.58810
health           -0.7403     0.6721  -1.102  0.27110
absences         -0.7666     0.2793  -2.745  0.00623 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 23.21 on 609 degrees of freedom
Multiple R-squared:  0.3857,    Adjusted R-squared:  0.3464
F-statistic: 9.806 on 39 and 609 DF,  p-value: < 2.2e-16

>
```

Anova table for comparision

```
> anova(lm_model, lm_model_full)
Analysis of Variance Table

Model 1: G1_perc ~ failures + schoolsup + Medu + Fedu + studytime + goout +
    Walc + traveltime + Dalc + famsup + reason + higher + Fjob +
    sex + absences
Model 2: G1_perc ~ school + sex + age + address + famsize + Pstatus +
    Medu + Fedu + Mjob + Fjob + reason + guardian + traveltime +
    studytime + failures + schoolsup + famsup + paid + activities +
    nursery + higher + internet + romantic + famrel + freetime +
    goout + Dalc + Walc + health + absences
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    628 355402
2    609 328163 19     27239 2.6605 0.0001699 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Using Squrt of abscense and doing lm on only filterd iportant columns

```
> lm_model_poly <- lm(G1_perc ~ failures + schoolsup + Medu + Fedu + studytime
+ goout + Walc + traveltime + Dalc +famsup
+reason+higher+Fjob+sex+sqrt(absences), data = new_df_for_lm)
summary(lm_model_poly)> summary(lm_model_poly)

Call:
lm(formula = G1_perc ~ failures + schoolsup + Medu + Fedu + studytime +
    goout + Walc + traveltime + Dalc + famsup + reason + higher +
    Fjob + sex + sqrt(absences), data = new_df_for_lm)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-72.112 -16.854   0.432  18.869  49.884

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      25.8509     6.5369   3.955 8.54e-05 ***
failures        -11.4085     1.7046  -6.693 4.85e-11 ***
schoolsupyes    -11.1870     3.1266  -3.578 0.000373 ***
Medu              2.5542     1.1281   2.264 0.023907 *
Fedu              1.3115     1.1986   1.094 0.274272
studytime         4.8561     1.2188   3.984 7.56e-05 ***
goout             0.1083     0.8694   0.125 0.900924
Walc             -0.3737     1.0209  -0.366 0.714462
traveltime       -2.3354     1.3274  -1.759 0.079011 .
Dalc             -2.8850     1.3208  -2.184 0.029308 *
famsupyes        -1.7609     1.9892  -0.885 0.376375
reasonhome        3.7568     2.4792   1.515 0.130190
reasonother      -2.1150     3.2010  -0.661 0.509024
reasonreputation  6.2583     2.5310   2.473 0.013676 *
higheryes        17.7387     3.3121   5.356 1.20e-07 ***
Fjobhealth        3.4303     6.4531   0.532 0.595204
Fjobother         6.1221     3.9298   1.558 0.119765
Fjobservices      3.8862     4.1591   0.934 0.350459
Fjobteacher      15.7538     5.8094   2.712 0.006876 **
sexM             -3.6210     2.0983  -1.726 0.084889 .
sqrt(absences)   -2.4169     0.8230  -2.937 0.003439 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.72 on 628 degrees of freedom
Multiple R-squared:  0.3386,    Adjusted R-squared:  0.3176
F-statistic: 16.08 on 20 and 628 DF,  p-value: < 2.2e-16
```

- 3

Interpretation of linear model:

- **R-squared: The Multiple R-squared of 0.3495 and Adjusted R-squared of 0.3288 indicate the model explains about 33-35% of the variance in the final grade percentile (G3_perc). This isn't extremely high, suggesting there's room for improvement.**

- **Individual predictors: Look at the p-values for each coefficient. Many variables (failures, schoolsup, Medu, etc.) are statistically significant, meaning they likely have a real relationship with G3_perc.**

What do we learn from this model?

- **Key factors: High importance variables include:**
- **Failures: Unsurprisingly, more failures lead to drastically lower G3_perc.**
- **Higher education goals: Students aiming for higher education tend to get better grades.**
- **Mother's Education (Medu): Higher maternal education is associated with better student performance.**
- **Father's Job (Fjobteacher): Students who have fathers working as teachers tend to have higher performance.**
- Study Time: As expected, studytime **has a positive impact on** G1_perc**, reinforcing the value of study habits.**
- **Potential surprises:**
- **School support (schoolsup): Students receiving extra educational support have lower G3_perc. This could mean support programs are targeted at struggling students.**
- **Reason for choosing school (reasonreputation): Students drawn to the school by its reputation seem to perform better.**

Failures,Higher education goals,School support ,Study Time are key indicators.

Logistic regression Models:

Logistic regression models to predict student consistency in performance (increasing or decreasing grades over time). This involves binary transformation of predictors, model fitting, and evaluation using confusion matrices and accuracy metrics.

```
>Logistic regression to predict consistency in student performance
glm_model <- glm(flag ~ ., data = new_df, family = binomial(link = "logit"))

> summary(glm_model)

Call:
glm(formula = flag ~ ., family = binomial(link = "logit"), data = new_df)
```

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.60081    2.81572  -1.279  0.20096
schoolMS         0.88285    0.38771   2.277  0.02278 *
sexM            -0.06442    0.36217  -0.178  0.85883
age              0.37685    0.14788   2.548  0.01082 *
addressU        -0.01146    0.38620  -0.030  0.97634
famsizeLE3      -0.23580    0.35400  -0.666  0.50535
PstatusT         0.42245    0.56027   0.754  0.45084
Medu             0.13046    0.23246   0.561  0.57464
Fedu            -0.08804    0.20365  -0.432  0.66549
Mjobhealth      -0.10104    0.78077  -0.129  0.89703
Mjobother       -0.01582    0.42855  -0.037  0.97055
Mjobservices     0.28021    0.55596   0.504  0.61425
Mjobteacher      0.51748    0.71446   0.724  0.46889
Fjobhealth      -1.48744    1.08383  -1.372  0.16994
Fjobother       -1.88739    0.74889  -2.520  0.01173 *
Fjobservices    -2.25411    0.77015  -2.927  0.00342 **
Fjobteacher     -2.66989    1.05724  -2.525  0.01156 *
reasonhome      -0.34115    0.40053  -0.852  0.39435
reasonother     -0.56110    0.56127  -1.000  0.31746
reasonreputation -0.84942    0.43321  -1.961  0.04991 *
guardianmother  -0.04452    0.35784  -0.124  0.90099
guardianother    0.66365    0.76049   0.873  0.38285
traveltime       0.16693    0.22241   0.751  0.45292
studytime        0.02604    0.20206   0.129  0.89747
failures        -0.36544    0.32384  -1.128  0.25912
schoolsupyes    -2.19839    1.10418  -1.991  0.04648 *
famsupyes       -0.26861    0.33965  -0.791  0.42903
paidyes         -1.52243    1.20398  -1.265  0.20605
activitiesyes    0.18036    0.33198   0.543  0.58693
nurseryyes      -0.15054    0.41153  -0.366  0.71451
higheryes       -1.13768    0.54574  -2.085  0.03710 *
internetyes      0.54865    0.42882   1.279  0.20074
romanticyes     -0.03153    0.31877  -0.099  0.92121
famrel           0.03544    0.16743   0.212  0.83237
freetime         0.03171    0.16514   0.192  0.84773
```

```
goout          -0.25749    0.16452  -1.565  0.11756
Dalc           -0.02939    0.22505  -0.131  0.89610
Walc            0.03009    0.17571   0.171  0.86402
health         -0.20099    0.11032  -1.822  0.06847 .
absences        0.05510    0.04870   1.131  0.25786
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 355.78  on 270  degrees of freedom
Residual deviance: 288.36  on 231  degrees of freedom
AIC: 368.36

Number of Fisher Scoring iterations: 5

>
```

VIF to check for multi collinearity

```
> vif(glm_model)
               GVIF Df GVIF^(1/(2*Df))
school     1.639071  1        1.280262
sex        1.511254  1        1.229331
age        1.308173  1        1.143754
address    1.576031  1        1.255401
famsize    1.277259  1        1.130159
Pstatus    1.273459  1        1.128476
Medu       3.316650  1        1.821167
Fedu       2.406049  1        1.551144
Mjob       4.078969  4        1.192117
Fjob       2.862592  4        1.140499
reason     1.898135  3        1.112725
guardian   1.709036  2        1.143373
traveltime 1.444574  1        1.201904
studytime  1.396223  1        1.181619
failures   1.363255  1        1.167585
schoolsup  1.077980  1        1.038258
famsup     1.326220  1        1.151616
```

```
paid      1.172113  1       1.082642
activities 1.339423  1       1.157334
nursery   1.230819  1       1.109423
higher    1.237895  1       1.112607
internet   1.350359  1       1.162049
romantic   1.206229  1       1.098285
famrel    1.239327  1       1.113251
freetime   1.376654  1       1.173309
goout     1.797416  1       1.340677
Dalc      1.966533  1       1.402331
Walc      2.134595  1       1.461025
health    1.213193  1       1.101451
absences   1.380643  1       1.175007
>


> #interpretation from vif values for correlation
> significant_vars <-
row.names(summary(glm_model)$coefficients[summary(glm_model)$coefficients
[, "Pr(>|z|)"] < 0.1, ])
print(significant_vars)
significant_vars <- row.names(significant_vars)
significant_vars> print(significant_vars)
[1] "schoolMS"        "age"            "Fjobother"       "Fjobservices"
[5] "Fjobteacher"      "reasonreputation" "schoolsupyes"     "higheryes"
[9] "health"

> glm_model_new <- glm(flag ~ school + age + Fjob + reason + schoolsup + higher
+ health + studytime + failures , data = new_df, family = binomial(link = "logit"))
> summary(glm_model_new)

Call:
glm(formula = flag ~ school + age + Fjob + reason + schoolsup +
    higher + health + studytime + failures, family = binomial(link = "logit"),
    data = new_df)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)      -3.9905     2.3396  -1.706  0.08808 .
schoolMS          0.5978     0.3153   1.896  0.05795 .
age               0.4097     0.1301   3.149  0.00164 **
Fjobhealth       -1.7009     0.9714  -1.751  0.07996 .
Fjobother        -1.8336     0.6988  -2.624  0.00869 **
Fjobservices     -2.1595     0.7273  -2.969  0.00299 **
Fjobteacher      -2.3158     0.9217  -2.513  0.01198 *
reasonhome       -0.2505     0.3689  -0.679  0.49713
reasonother      -0.5397     0.5141  -1.050  0.29387
reasonreputation  -0.6087     0.3824  -1.592  0.11148
schoolsupyes     -2.0573     1.0521  -1.955  0.05053 .
higheryes        -1.0150     0.5075  -2.000  0.04549 *
health           -0.1949     0.1004  -1.940  0.05233 .
studytime         0.1253     0.1804   0.695  0.48723
failures         -0.3223     0.2964  -1.087  0.27691
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 355.78  on 270  degrees of freedom
Residual deviance: 301.02  on 256  degrees of freedom
AIC: 331.02

Number of Fisher Scoring iterations: 5

>
```

Evaluation of Logistic regression model:

```
> #evaluate the model
> library(caret)
Loading required package: ggplot2

# Create indices for a 70-30 train-test split
set.seed(123)  # for reproducibility
splitIndex <- createDataPartition(new_df$flag, p = 0.7, list = FALSE)
```

```
# Create training and testing datasets
train_data <- new_df[splitIndex, ]
test_data <- new_df[-splitIndex, ]

# Fit model on training data
glm_train_model <- glm(flag ~ school + age + Fjob + reason + schoolsup + higher +
health+ studytime + failures ,
                       data = train_data, family = binomial(link = "logit"))

# Evaluate model on test data
prob_test <- predict(glm_train_model, test_data, type = "response")
predicted_test_classes <- ifelse(prob_test > 0.5, 1, 0)
confusionMatrix <- table(Predicted = predicted_test_classes, Actual =
test_data$flag)
print(confusionMatrix)

# Calculate accuracy
accuracy_test <- mean(predicted_test_classes == test_data$flag)
print(paste("Accuracy:", accuracy_test))
Loading required package: lattice
>
> # Create indices for a 70-30 train-test split
> set.seed(123)  # for reproducibility
> splitIndex <- createDataPartition(new_df$flag, p = 0.7, list = FALSE)
>
> # Create training and testing datasets
> train_data <- new_df[splitIndex, ]
> test_data <- new_df[-splitIndex, ]
>
> # Fit model on training data
> glm_train_model <- glm(flag ~ school + age + Fjob + reason + schoolsup + higher
+ health+ studytime + failures ,
+                        data = train_data, family = binomial(link = "logit"))
>
> # Evaluate model on test data
> prob_test <- predict(glm_train_model, test_data, type = "response")
> predicted_test_classes <- ifelse(prob_test > 0.5, 1, 0)
> confusionMatrix <- table(Predicted = predicted_test_classes, Actual =
test_data$flag)
> print(confusionMatrix)
         Actual
Predicted  0  1
        0 38 18
        1 15 10
>
> # Calculate accuracy
> accuracy_test <- mean(predicted_test_classes == test_data$flag)
> print(paste("Accuracy:", accuracy_test))
[1] "Accuracy: 0.592592592592593"
>
```

Intrepretation for Logistic Regression:

The model aims to predict if a student's performance will show continuous increase (flag=1) or continuous decrease (flag=0) based on various predictors.
Accuracy of the model is only 0.59. If I say all the values are true, then the accuracy would be 0.63. So, the model is not accurate. But the model is able to tell us factors that influence it.
School MS (`schoolMS`):

  - Odds Ratio (OR): 2.42. Students at Mousinho da Silveira school are about 142% more likely to show continuous improvement in performance compared to other schools.

- Age (`age`):
  - OR: 1.46. Each additional year in age is associated with a 46% increase in the likelihood of continuous performance improvement.
- Father's Job - Other (`Fjobother`):
  - OR: 0.15. If the father's job is categorized as 'other', there's an 85% decrease in the likelihood of the student's performance continuously improving.
- Father's Job - Services (`Fjobservices`):
  - OR: 0.10. Fathers working in services are associated with a 90% decrease in the likelihood of their children's performance improving.
- Father's Job - Teacher (`Fjobteacher`):
  - OR: 0.07. Fathers who are teachers are associated with a 93% decrease in the likelihood of their children's performance improving.
- Higher Education Aspiration (`higheryes`):
  - OR: 0.32. Students who aspire for higher education are 68% less likely to show continuous improvement in their performance.

- School Support (`schoolsupyes`):
    - OR: 0.11. Receiving school support is associated with an 89% decrease in the likelihood of continuous performance improvement.

School Reputation: Students who primarily chose the school for its reputation have about 41% lower odds of showing consistent improvement in performance (e^(-0.84942) = 0.427).