Student Performance Analysis

Sri Darahas Koneru

Introduction

This project analyzes the Student Performance dataset from the UCI Machine Learning Repository. We explore how demographic, social, and academic factors impact final grades, focusing on G3 (final math grade).

Dataset Description

The dataset (`student-mat.csv`) contains 395 rows and 33 variables. Key variables include:

- `G1`, `G2`, `G3`: Grades in 3 periods (0–20)
- `studytime`: Weekly study time (1–4)
- `failures`: Number of past class failures
- `Medu`, `Fedu`: Mother's and father's education level
- `goout`: Going out with friends (1–5)
- `Dalc`, `Walc`: Alcohol consumption (1–5)

```
data <- read.csv("student-mat.csv", sep = ";")
str(data)
```

```
## 'data.frame':    395 obs. of  33 variables:
##  $ school    : chr  "GP" "GP" "GP" "GP" ...
##  $ sex       : chr  "F" "F" "F" "F" ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr  "U" "U" "U" "U" ...
##  $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr  "A" "T" "T" "T" ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr  "teacher" "other" "other" "services" ...
##  $ reason    : chr  "course" "course" "other" "home" ...
##  $ guardian  : chr  "mother" "father" "mother" "mother" ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup : chr  "yes" "no" "yes" "no" ...
##  $ famsup    : chr  "no" "yes" "no" "yes" ...
##  $ paid      : chr  "no" "no" "yes" "yes" ...
##  $ activities: chr  "no" "no" "no" "yes" ...
##  $ nursery   : chr  "yes" "no" "yes" "yes" ...
##  $ higher    : chr  "yes" "yes" "yes" "yes" ...
```

```
##  $ internet  : chr  "no" "yes" "yes" "yes" ...
##  $ romantic  : chr  "no" "no" "no" "yes" ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
##  $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
##  $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
##  $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
```

```
summary(data)
```

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

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

Data Cleaning

```r
sum(is.na(data))
```

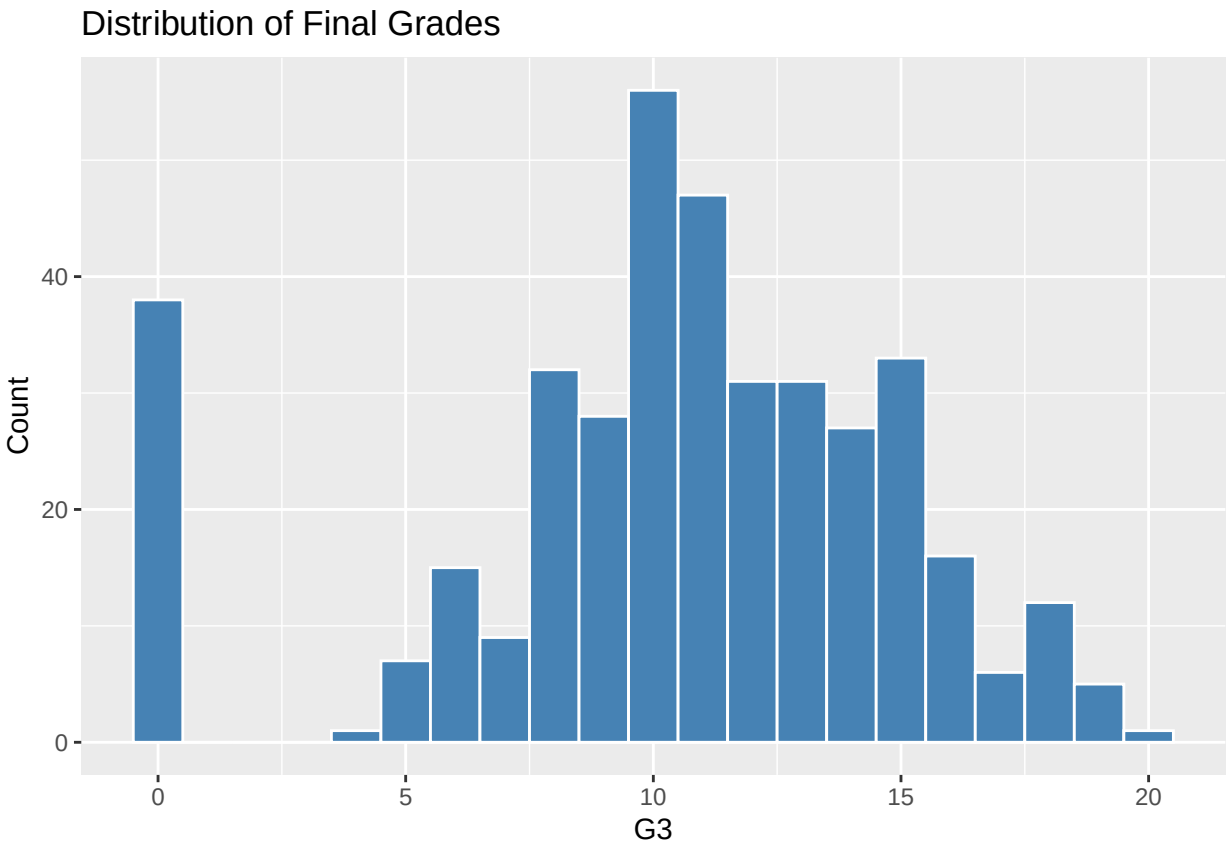
```
## [1] 0
```

```r
data <- data %>% mutate_if(is.character, as.factor)
```

Exploratory Data Analysis

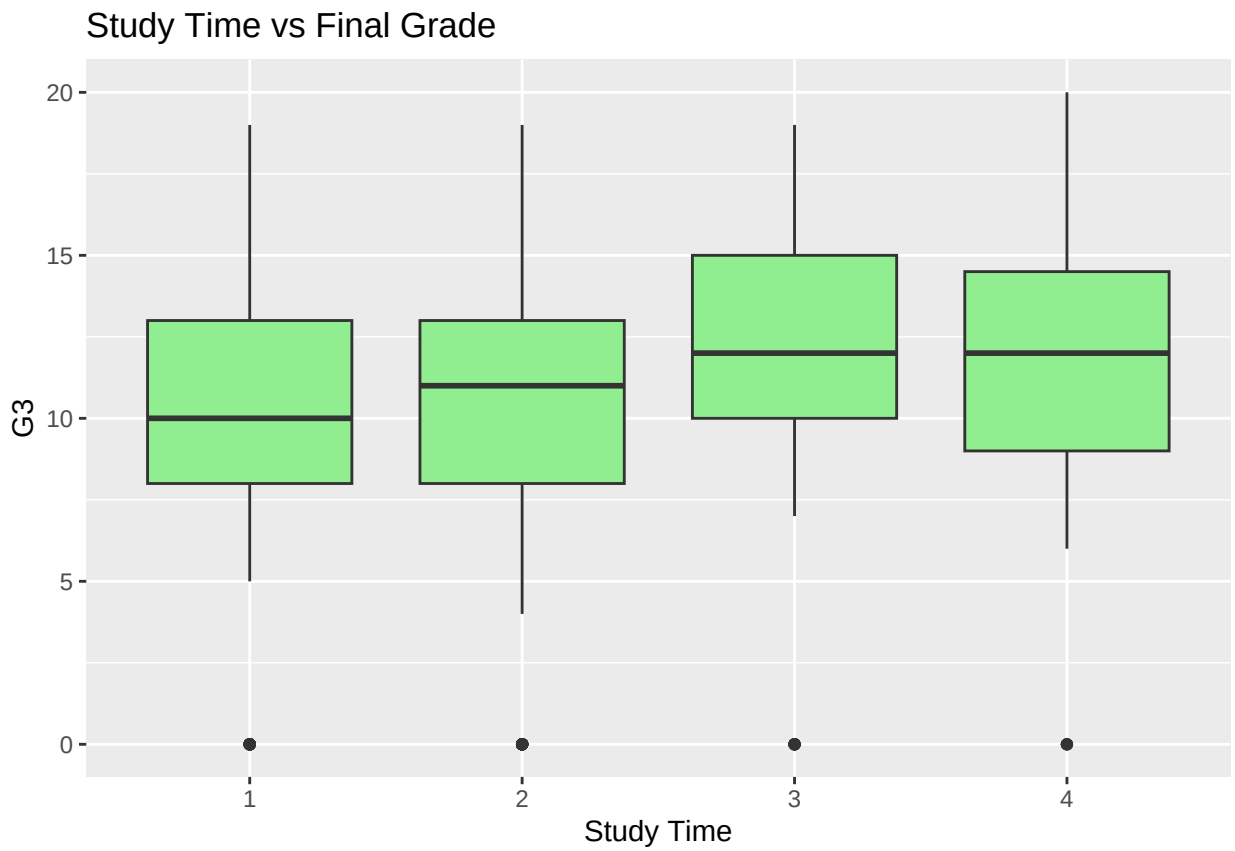
Final Grade Distribution

```r
ggplot(data, aes(x = G3)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "white") +
  labs(title = "Distribution of Final Grades", x = "G3", y = "Count")
```



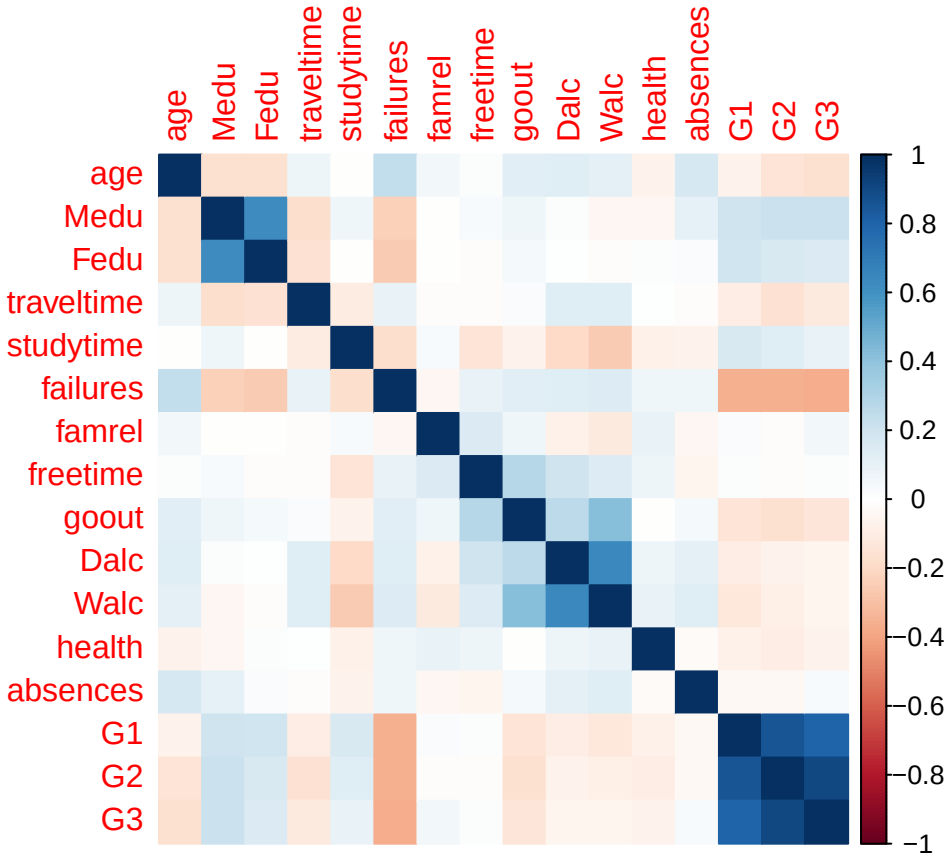
Study Time vs Final Grade

```
ggplot(data, aes(x = factor(studytime), y = G3)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Study Time vs Final Grade", x = "Study Time", y = "G3")
```



Correlation Matrix

```
numeric_data <- select_if(data, is.numeric)
cor_matrix <- cor(numeric_data)
corrplot(cor_matrix, method = "color")
```



Modeling (Machine Learning Enhancements)

Train/Test Split

```
set.seed(123)
n <- nrow(data)
train_index <- sample(1:n, 0.8 * n)
train <- data[train_index, ]
test <- data[-train_index, ]
```

Train Linear Regression Model

```
model <- lm(G3 ~ studytime + failures + Medu + Fedu, data = train)
summary(model)
```

```
##
## Call:
## lm(formula = G3 ~ studytime + failures + Medu + Fedu, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -12.3759  -1.8236   0.2213   2.8693   8.8737
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.8074     0.9282  10.567  < 2e-16 ***
## studytime     0.2276     0.2797   0.814   0.4165
## failures     -2.4373     0.3504  -6.955 2.09e-11 ***
## Medu          0.5117     0.2759   1.854   0.0646 .
## Fedu         -0.1295     0.2724  -0.475   0.6348
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.107 on 311 degrees of freedom
## Multiple R-squared:  0.1774, Adjusted R-squared:  0.1668
## F-statistic: 16.76 on 4 and 311 DF,  p-value: 1.869e-12
```

Predict and Evaluate Model

```
predictions <- predict(model, newdata = test)
rmse_val <- rmse(test$G3, predictions)
r2_val <- 1 - sum((predictions - test$G3)^2) / sum((mean(train$G3) - test$G3)^2)

rmse_val
```

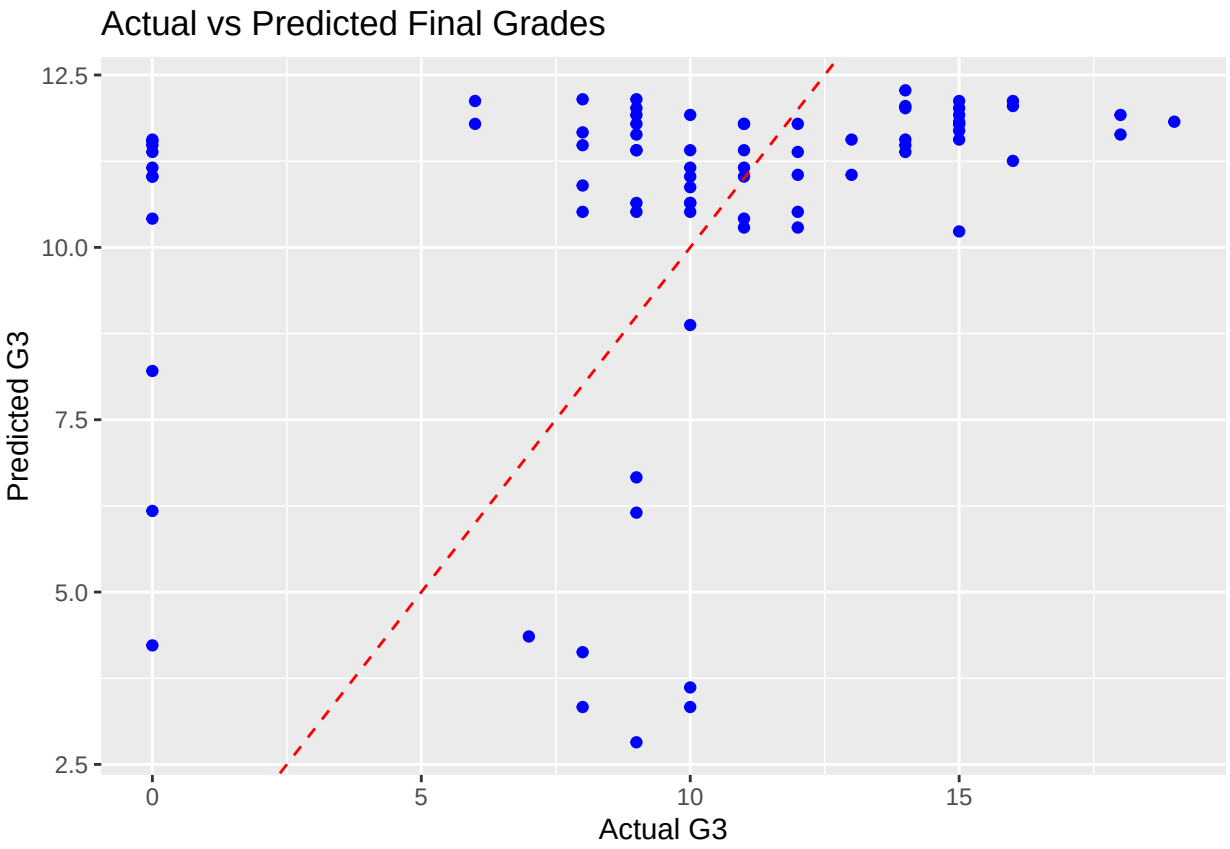
```
## [1] 4.821556
```

```
r2_val
```

```
## [1] 0.03434971
```

Actual vs Predicted Plot

```
ggplot(data.frame(actual = test$G3, predicted = predictions), aes(x = actual, y = predicted)) +
  geom_point(color = "blue") +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Actual vs Predicted Final Grades", x = "Actual G3", y = "Predicted G3")
```



Conclusion

Higher study time and parent education are linked to better grades.

More past failures hurt final scores.

A simple regression model gives a good early insight into academic predictors.

Adding model evaluation metrics (RMSE, R^2) and a prediction plot strengthens the analysis and better reflects real-world predictive modeling workflows.