Student Performance Analysis

Sanjay Chhetri

2025-04-27

Introduction

The goal of this project is to analyze how students’ reading and writing scores relate to their math scores and to build a simple predictive model based on these relationships.

Data Loading

```
# Load the dataset
student_data <- read_csv("StudentsPerformance.csv")
```

```
## Rows: 1000 Columns: 8
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (5): gender, race/ethnicity, parental level of education, lunch, test pr...
## dbl (3): math score, reading score, writing score
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# View the first few rows
head(student_data)
```

```
## # A tibble: 6 × 8
##   gender `race/ethnicity` parental level of educa…¹ lunch test preparation cou…²
##   <chr>  <chr>            <chr>                     <chr> <chr>
## 1 female group B          bachelor's degree         stan… none
## 2 female group C          some college              stan… completed
## 3 female group B          master's degree           stan… none
## 4 male   group A          associate's degree        free… none
## 5 male   group C          some college              stan… none
## 6 female group B          associate's degree        stan… none
## # ℹ abbreviated names: ¹`parental level of education`,
## #   ²`test preparation course`
## # ℹ 3 more variables: `math score` <dbl>, `reading score` <dbl>,
## #   `writing score` <dbl>
```

Data Exploration

```
# Structure of the dataset
glimpse(student_data)
```

```
## Rows: 1,000
## Columns: 8
## $ gender                        <chr> "female", "female", "female", "male", "m…
## $ `race/ethnicity`              <chr> "group B", "group C", "group B", "group …
## $ `parental level of education` <chr> "bachelor's degree", "some college", "ma…
## $ lunch                         <chr> "standard", "standard", "standard", "fre…
## $ `test preparation course`     <chr> "none", "completed", "none", "none", "no…
## $ `math score`                  <dbl> 72, 69, 90, 47, 76, 71, 88, 40, 64, 38, …
## $ `reading score`               <dbl> 72, 90, 95, 57, 78, 83, 95, 43, 64, 60, …
## $ `writing score`               <dbl> 74, 88, 93, 44, 75, 78, 92, 39, 67, 50, …
```

```
# Summary statistics
summary(student_data)
```

```
##     gender          race/ethnicity     parental level of education
##  Length:1000        Length:1000        Length:1000                
##  Class :character   Class :character   Class :character           
##  Mode  :character   Mode  :character   Mode  :character           
##                                                                   
##                                                                   
##                                                                   
##     lunch           test preparation course   math score     reading score   
##  Length:1000        Length:1000             Min.   :  0.00   Min.   : 17.00  
##  Class :character   Class :character        1st Qu.: 57.00   1st Qu.: 59.00  
##  Mode  :character   Mode  :character        Median : 66.00   Median : 70.00  
##                                             Mean   : 66.09   Mean   : 69.17  
##                                             3rd Qu.: 77.00   3rd Qu.: 79.00  
##                                             Max.   :100.00   Max.   :100.00  
##  writing score   
##  Min.   : 10.00  
##  1st Qu.: 57.75  
##  Median : 69.00  
##  Mean   : 68.05  
##  3rd Qu.: 79.00  
##  Max.   :100.00
```

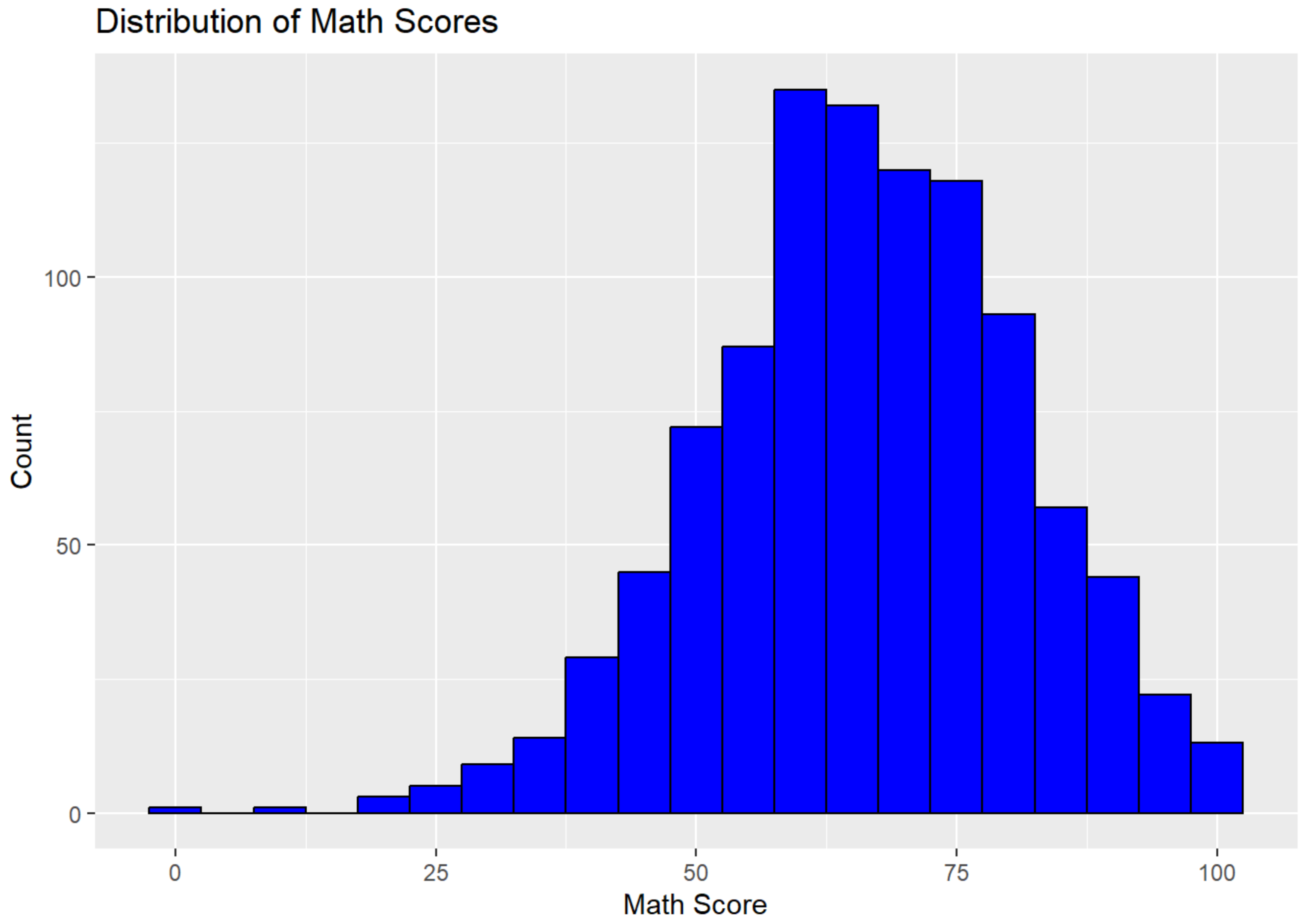
```
# Dataset dimensions
dim(student_data)
```

```
## [1] 1000    8
```

Visualizations

Distribution of Math Scores

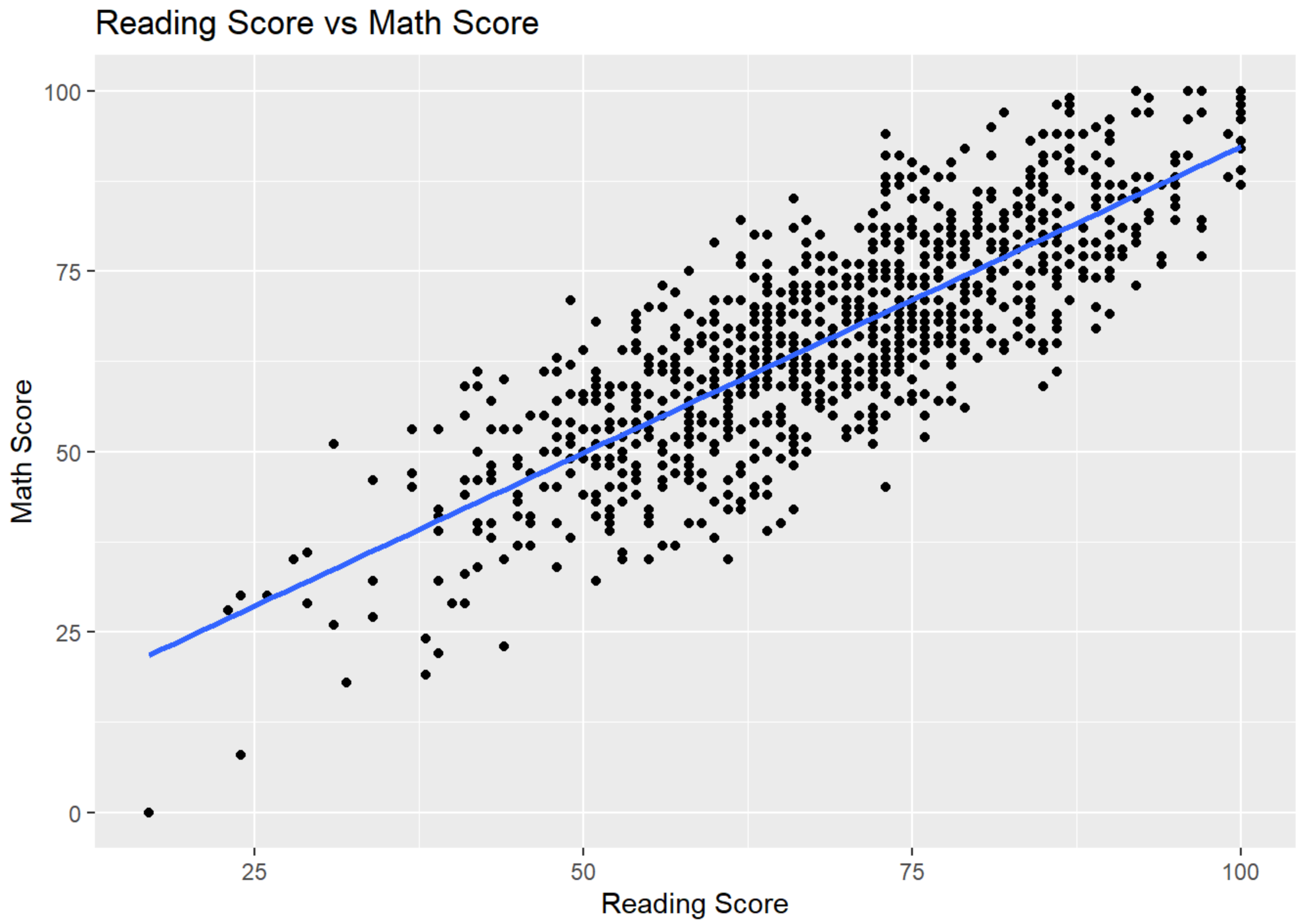
```
ggplot(student_data, aes(x = `math score`)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  labs(title = "Distribution of Math Scores", x = "Math Score", y = "Count")
```



Reading Score vs Math Score

```
ggplot(student_data, aes(x = `reading score`, y = `math score`)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Reading Score vs Math Score", x = "Reading Score", y = "Math Score")
```

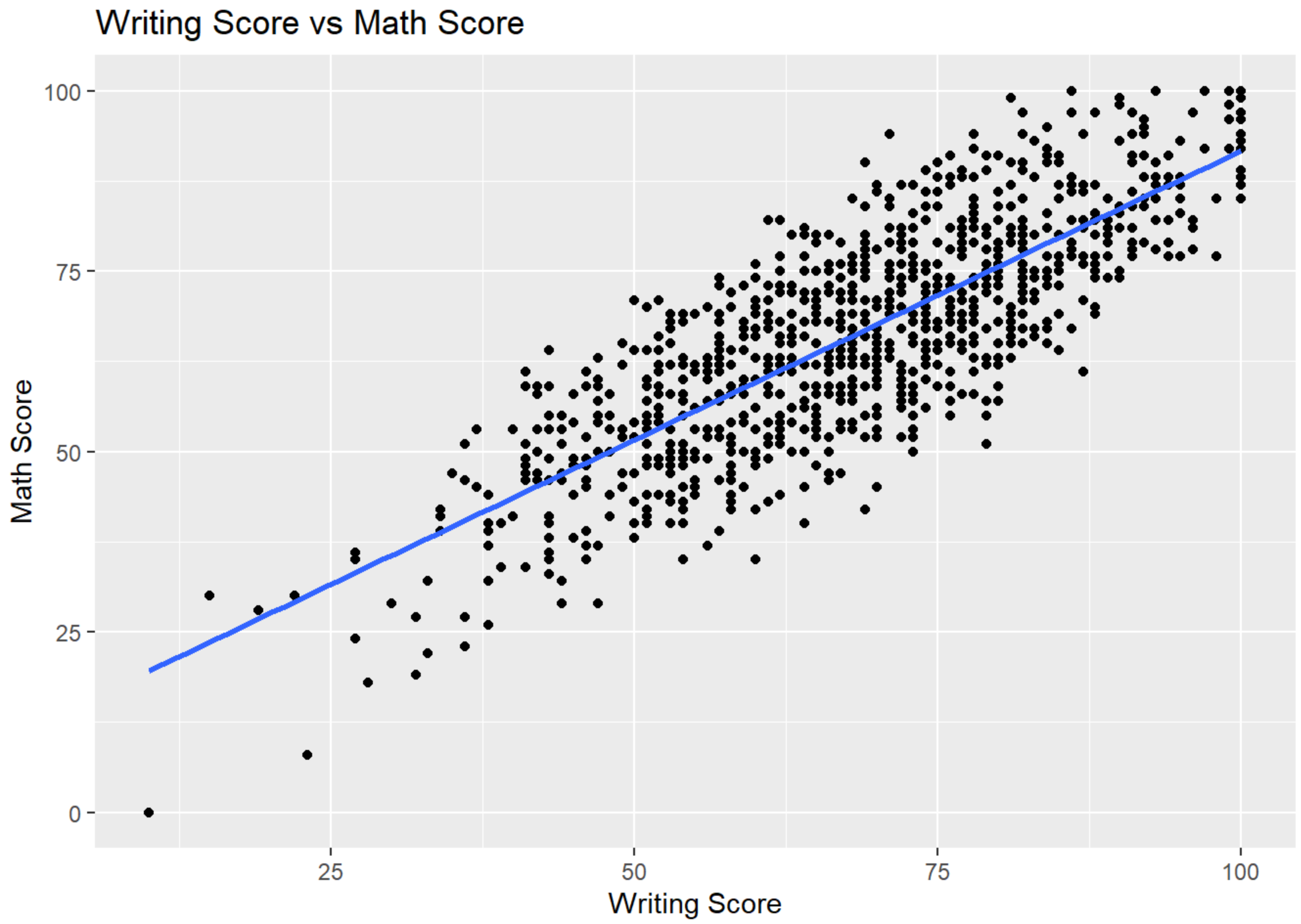
```
## `geom_smooth()` using formula = 'y ~ x'
```



Writing Score vs Math Score

```
ggplot(student_data, aes(x = `writing score`, y = `math score`)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Writing Score vs Math Score", x = "Writing Score", y = "Math Score")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Modeling

Linear Regression Model

```
# Build the linear model
model <- lm(`math score` ~ `reading score` + `writing score`, data = student_data)

# Model summary
summary(model)
```

```
##
## Call:
## lm(formula = `math score` ~ `reading score` + `writing score`, 
##     data = student_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -23.8779  -6.1750   0.2693   6.0184  24.8727 
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      7.52409    1.32823   5.665 1.93e-08 ***
## `reading score`  0.60129    0.06304   9.538  < 2e-16 ***
## `writing score`  0.24942    0.06057   4.118 4.14e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.667 on 997 degrees of freedom
## Multiple R-squared:  0.674,  Adjusted R-squared:  0.6733 
## F-statistic:  1031 on 2 and 997 DF,  p-value: < 2.2e-16
```

Results and Interpretation

- Both reading and writing scores are statistically significant predictors of math score.
- R-squared value is approximately 0.674, meaning about 67.4% of the variability in math scores can be explained by reading and writing scores.

Conclusions

- Students who perform better in reading and writing are likely to perform better in math.
- Reading and writing scores are strong predictors for math performance.

Future Work

- Incorporate more features like parental education, lunch status, and test preparation.
- Test model improvements using different regression techniques.
- Explore classification approaches (e.g., pass/fail prediction based on scores).