p8130_final

Zihan Lin

R Markdown

```r
# Load the dataset
data <- read.csv("/Users/suwa/Desktop/p8130_final/data/Project_1_data.csv")

# Identify and encode binary categorical variables
# Check the structure of the data
str(data)
```

```
## 'data.frame':    948 obs. of  14 variables:
##  $ Gender             : chr  "female" "female" "female" "male" ...
##  $ EthnicGroup        : chr  "" "group C" "group B" "group A" ...
##  $ ParentEduc         : chr  "bachelor's degree" "some college" "master's degree" "associate's degre
##  $ LunchType          : chr  "standard" "standard" "standard" "free/reduced" ...
##  $ TestPrep           : chr  "none" "" "none" "none" ...
##  $ ParentMaritalStatus: chr  "married" "married" "single" "married" ...
##  $ PracticeSport      : chr  "regularly" "sometimes" "sometimes" "never" ...
##  $ IsFirstChild       : chr  "yes" "yes" "yes" "no" ...
##  $ NrSiblings         : int  3 0 4 1 0 1 1 1 3 NA ...
##  $ TransportMeans     : chr  "school_bus" "" "school_bus" "" ...
##  $ WklyStudyHours     : chr  "< 5" "10-May" "< 5" "10-May" ...
##  $ MathScore          : int  71 69 87 45 76 73 85 41 65 37 ...
##  $ ReadingScore       : int  71 90 93 56 78 84 93 43 64 59 ...
##  $ WritingScore       : int  74 88 91 42 75 79 89 39 68 50 ...
```

```r
# Convert binary categorical variables to 0/1
binary_vars <- c("Gender", "LunchType", "TestPrep", "IsFirstChild", "TransportMeans")
data <- data %>%
  mutate(across(all_of(binary_vars), ~ ifelse(. == levels(as.factor(.))[1], 0, 1)))

# Create dummy variables for multi-category variables
# Identify multi-category variables
multi_category_vars <- c("EthnicGroup", "ParentEduc", "ParentMaritalStatus", "PracticeSport", "WklyStud

# Generate dummy variables for multi-category variables
dummy_vars <- dummyVars("~ .", data = data, fullRank = TRUE)
data <- predict(dummy_vars, newdata = data) %>% as.data.frame()

# Handle missing values
# Remove rows with missing values
data <- data %>% drop_na()
```

```r
# Check the cleaned dataset
str(data)        # Check the structure of the cleaned dataset
```

```
## 'data.frame':    902 obs. of  30 variables:
##  $ Gender                      : num  0 0 0 1 1 0 0 1 1 1 ...
##  $ EthnicGroupgroup A          : num  0 0 0 1 0 0 0 0 0 0 ...
##  $ EthnicGroupgroup B          : num  0 0 1 0 0 1 1 1 0 0 ...
##  $ EthnicGroupgroup C          : num  0 1 0 0 1 0 0 0 0 1 ...
##  $ EthnicGroupgroup D          : num  0 0 0 0 0 0 0 0 1 0 ...
##  $ EthnicGroupgroup E          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ParentEducassociate's degree: num  0 0 0 1 0 1 0 0 0 1 ...
##  $ ParentEducbachelor's degree : num  1 0 0 0 0 0 0 0 0 0 ...
##  $ ParentEduchigh school       : num  0 0 0 0 0 0 0 0 1 0 ...
##  $ ParentEducmaster's degree   : num  0 0 1 0 0 0 0 0 0 0 ...
##  $ ParentEducsome college      : num  0 1 0 0 1 0 1 1 0 0 ...
##  $ ParentEducsome high school  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LunchType                   : num  1 1 1 0 1 1 1 0 0 1 ...
##  $ TestPrep                    : num  1 0 1 1 1 1 1 1 1 1 ...
##  $ ParentMaritalStatusdivorced : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ParentMaritalStatusmarried  : num  1 1 0 1 1 1 0 1 0 0 ...
##  $ ParentMaritalStatussingle   : num  0 0 1 0 0 0 0 0 1 0 ...
##  $ ParentMaritalStatuswidowed  : num  0 0 0 0 0 0 1 0 0 0 ...
##  $ PracticeSportnever          : num  0 0 0 1 0 0 1 0 0 0 ...
##  $ PracticeSportregularly      : num  1 0 0 0 0 1 0 0 0 0 ...
##  $ PracticeSportsometimes      : num  0 1 1 0 1 0 0 1 1 1 ...
##  $ IsFirstChild                : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ NrSiblings                  : num  3 0 4 1 0 1 1 1 3 1 ...
##  $ TransportMeans              : num  1 0 1 0 1 1 1 1 1 1 ...
##  $ WklyStudyHours< 5           : num  1 0 1 0 0 0 0 0 0 0 ...
##  $ WklyStudyHours> 10          : num  0 0 0 0 0 0 0 1 1 0 ...
##  $ WklyStudyHours10-May        : num  0 1 0 1 1 1 1 0 0 1 ...
##  $ MathScore                   : num  71 69 87 45 76 73 85 41 65 58 ...
##  $ ReadingScore                : num  71 90 93 56 78 84 93 43 64 54 ...
##  $ WritingScore                : num  74 88 91 42 75 79 89 39 68 52 ...
```

```r
summary(data)    # Summarize the cleaned data
```

```
##      Gender     EthnicGroupgroup A EthnicGroupgroup B EthnicGroupgroup C
##  Min.   :0.00   Min.   :0.00000    Min.   :0.000      Min.   :0.0000  
##  1st Qu.:0.00   1st Qu.:0.00000    1st Qu.:0.000      1st Qu.:0.0000  
##  Median :0.00   Median :0.00000    Median :0.000      Median :0.0000  
##  Mean   :0.49   Mean   :0.08204    Mean   :0.184      Mean   :0.2905  
##  3rd Qu.:1.00   3rd Qu.:0.00000    3rd Qu.:0.000      3rd Qu.:1.0000  
##  Max.   :1.00   Max.   :1.00000    Max.   :1.000      Max.   :1.0000  
##  EthnicGroupgroup D EthnicGroupgroup E ParentEducassociate's degree
##  Min.   :0.0000     Min.   :0.0000     Min.   :0.0000              
##  1st Qu.:0.0000     1st Qu.:0.0000     1st Qu.:0.0000              
##  Median :0.0000     Median :0.0000     Median :0.0000              
##  Mean   :0.2528     Mean   :0.1286     Mean   :0.2073              
##  3rd Qu.:1.0000     3rd Qu.:0.0000     3rd Qu.:0.0000              
##  Max.   :1.0000     Max.   :1.0000     Max.   :1.0000              
##  ParentEducbachelor's degree ParentEduchigh school ParentEducmaster's degree
##  Min.   :0.0000              Min.   :0.0000        Min.   :0.00000
##  1st Qu.:0.0000              1st Qu.:0.0000        1st Qu.:0.00000
##  Median :0.0000              Median :0.0000        Median :0.00000
##  Mean   :0.1075              Mean   :0.1863        Mean   :0.05765
##  3rd Qu.:0.0000              3rd Qu.:0.0000        3rd Qu.:0.00000
##  Max.   :1.0000              Max.   :1.0000        Max.   :1.00000
##  ParentEducsome college ParentEducsome high school   LunchType
##  Min.   :0.000          Min.   :0.0000             Min.   :0.0000
##  1st Qu.:0.000          1st Qu.:0.0000             1st Qu.:0.0000
##  Median :0.000          Median :0.0000             Median :1.0000
##  Mean   :0.214          Mean   :0.1718             Mean   :0.6475
##  3rd Qu.:0.000          3rd Qu.:0.0000             3rd Qu.:1.0000
##  Max.   :1.000          Max.   :1.0000             Max.   :1.0000
##     TestPrep      ParentMaritalStatusdivorced ParentMaritalStatusmarried
##  Min.   :0.0000   Min.   :0.000               Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:0.000               1st Qu.:0.0000
##  Median :1.0000   Median :0.000               Median :1.0000
##  Mean   :0.9412   Mean   :0.153               Mean   :0.5455
##  3rd Qu.:1.0000   3rd Qu.:0.000               3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.000               Max.   :1.0000
##  ParentMaritalStatussingle ParentMaritalStatuswidowed PracticeSportnever
##  Min.   :0.0000            Min.   :0.00000            Min.   :0.0000
##  1st Qu.:0.0000            1st Qu.:0.00000            1st Qu.:0.0000
##  Median :0.0000            Median :0.00000            Median :0.0000
##  Mean   :0.2251            Mean   :0.02661            Mean   :0.1175
##  3rd Qu.:0.0000            3rd Qu.:0.00000            3rd Qu.:0.0000
##  Max.   :1.0000            Max.   :1.00000            Max.   :1.0000
##  PracticeSportregularly PracticeSportsometimes  IsFirstChild     NrSiblings
##  Min.   :0.0000         Min.   :0.0000         Min.   :0.000   Min.   :0.000
##  1st Qu.:0.0000         1st Qu.:0.0000         1st Qu.:1.000   1st Qu.:1.000
##  Median :0.0000         Median :1.0000         Median :1.000   Median :2.000
##  Mean   :0.3636         Mean   :0.5022         Mean   :0.969   Mean   :2.155
##  3rd Qu.:1.0000         3rd Qu.:1.0000         3rd Qu.:1.000   3rd Qu.:3.000
##  Max.   :1.0000         Max.   :1.0000         Max.   :1.000   Max.   :7.000
##  TransportMeans   WklyStudyHours< 5 WklyStudyHours> 10 WklyStudyHours10-May
##  Min.   :0.0000   Min.   :0.0000    Min.   :0.0000     Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:0.0000    1st Qu.:0.0000     1st Qu.:0.0000
##  Median :1.0000   Median :0.0000    Median :0.0000     Median :1.0000
##  Mean   :0.8947   Mean   :0.2661    Mean   :0.1574     Mean   :0.5355
##  3rd Qu.:1.0000   3rd Qu.:1.0000    3rd Qu.:0.0000     3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000    Max.   :1.0000     Max.   :1.0000
##    MathScore       ReadingScore    WritingScore
##  Min.   :  0.00   Min.   : 17.0   Min.   : 10.00
##  1st Qu.: 56.00   1st Qu.: 59.0   1st Qu.: 57.00
##  Median : 66.00   Median : 69.0   Median : 68.00
##  Mean   : 66.03   Mean   : 68.8   Mean   : 67.85
##  3rd Qu.: 76.00   3rd Qu.: 79.0   3rd Qu.: 78.00
##  Max.   :100.00   Max.   :100.0   Max.   :100.00
```

```
head(data)       # View the first few rows of the cleaned data
```

```
##   Gender EthnicGroupgroup A EthnicGroupgroup B EthnicGroupgroup C
## 1      0                  0                  0                  0
## 2      0                  0                  0                  1
## 3      0                  0                  1                  0
## 4      1                  1                  0                  0
## 5      1                  0                  0                  1
## 6      0                  0                  1                  0
##   EthnicGroupgroup D EthnicGroupgroup E ParentEducassociate's degree
## 1                  0                  0                            0
## 2                  0                  0                            0
## 3                  0                  0                            0
## 4                  0                  0                            1
## 5                  0                  0                            0
## 6                  0                  0                            1
##   ParentEducbachelor's degree ParentEduchigh school ParentEducmaster's degree
## 1                           1                     0                         0
## 2                           0                     0                         0
## 3                           0                     0                         1
## 4                           0                     0                         0
## 5                           0                     0                         0
## 6                           0                     0                         0
##   ParentEducsome college ParentEducsome high school LunchType TestPrep
## 1                      0                          0         1        1
## 2                      1                          0         1        0
## 3                      0                          0         1        1
## 4                      0                          0         0        1
## 5                      1                          0         1        1
## 6                      0                          0         1        1
##   ParentMaritalStatusdivorced ParentMaritalStatusmarried
## 1                           0                          1
## 2                           0                          1
## 3                           0                          0
## 4                           0                          1
## 5                           0                          1
## 6                           0                          1
##   ParentMaritalStatussingle ParentMaritalStatuswidowed PracticeSportnever
## 1                         0                          0                  0
## 2                         0                          0                  0
## 3                         1                          0                  0
## 4                         0                          0                  1
## 5                         0                          0                  0
## 6                         0                          0                  0
##   PracticeSportregularly PracticeSportsometimes IsFirstChild NrSiblings
## 1                      1                      0            1          3
## 2                      0                      1            1          0
## 3                      0                      1            1          4
## 4                      0                      0            1          1
## 5                      0                      1            1          0
## 6                      1                      0            1          1
##   TransportMeans WklyStudyHours< 5 WklyStudyHours> 10 WklyStudyHours10-May
## 1              1                 1                  0                    0
## 2              0                 0                  0                    1
## 3              1                 1                  0                    0
## 4              0                 0                  0                    1
## 5              1                 0                  0                    1
## 6              1                 0                  0                    1
##   MathScore ReadingScore WritingScore
## 1        71           71           74
## 2        69           90           88
## 3        87           93           91
## 4        45           56           42
## 5        76           78           75
## 6        73           84           79
```

```
# Save Cleaned Data
write.csv(data, "/Users/suwa/Desktop/p8130_final/data/data_cleaned.csv", row.names = FALSE)
```

```
# Reload the dataset
data <- read.csv("/Users/suwa/Desktop/p8130_final/data/data_cleaned.csv")

# Generate a summary table
skim(data)
```

Table 1: Data summary

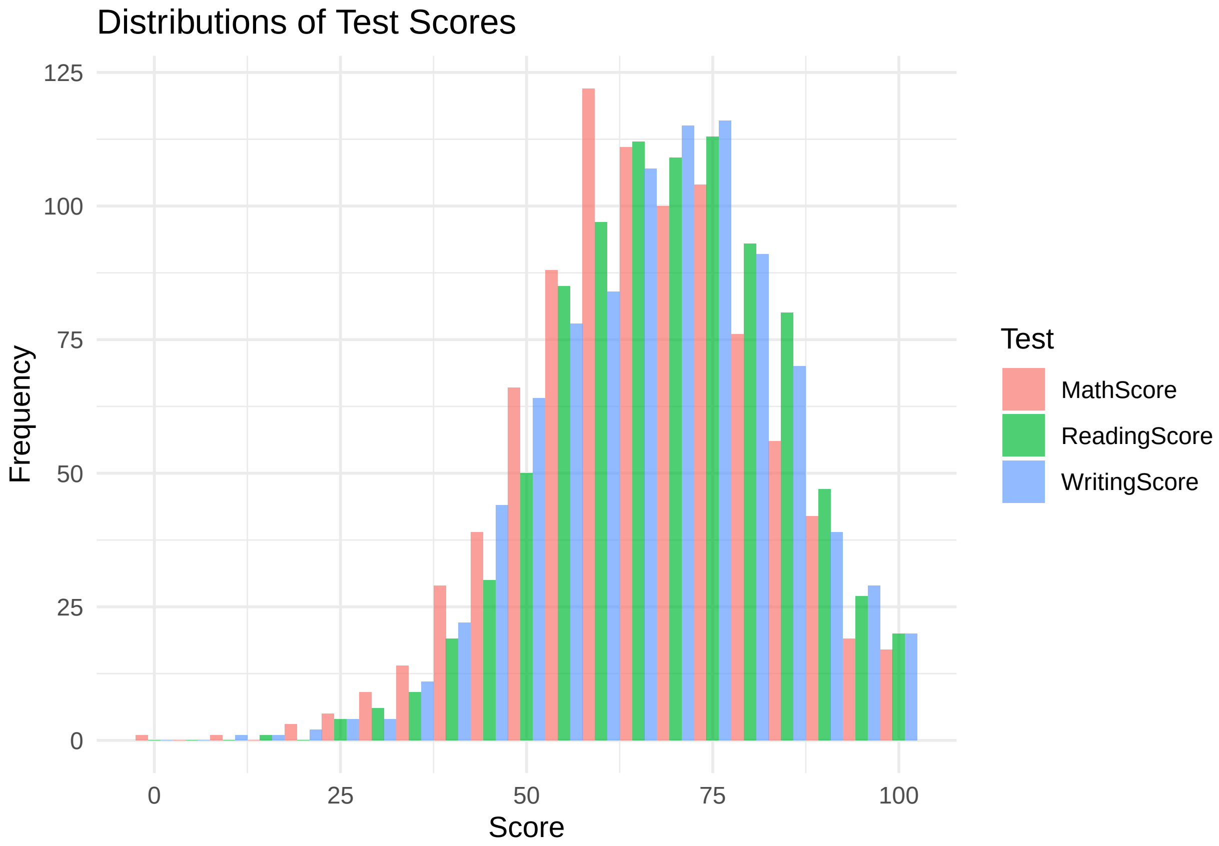
Name	data
Number of rows	902
Number of columns	30
Column type frequency:	
numeric	30
Group variables	None

Variable type: numeric

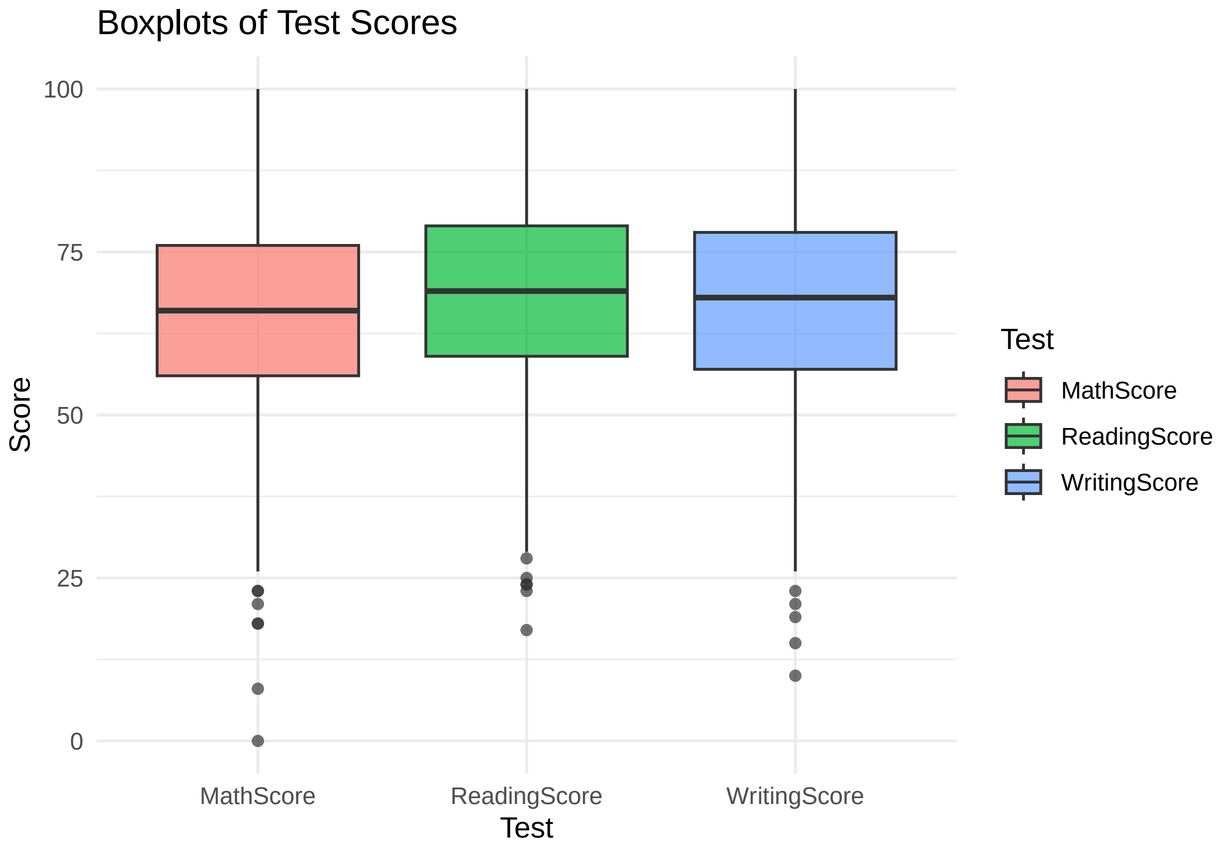
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Gender	0	1	0.49	0.50	0	0	0	1	1	
EthnicGroupgroup.A	0	1	0.08	0.27	0	0	0	0	1	
EthnicGroupgroup.B	0	1	0.18	0.39	0	0	0	0	1	
EthnicGroupgroup.C	0	1	0.29	0.45	0	0	0	1	1	
EthnicGroupgroup.D	0	1	0.25	0.43	0	0	0	1	1	
EthnicGroupgroup.E	0	1	0.13	0.33	0	0	0	0	1	
ParentEducassociate.s.degree	0	1	0.21	0.41	0	0	0	0	1	
ParentEducbachelor.s.degree	0	1	0.11	0.31	0	0	0	0	1	
ParentEduchigh.school	0	1	0.19	0.39	0	0	0	0	1	
ParentEducmaster.s.degree	0	1	0.06	0.23	0	0	0	0	1	
ParentEducsome.college	0	1	0.21	0.41	0	0	0	0	1	
ParentEducsome.high.school	0	1	0.17	0.38	0	0	0	0	1	
LunchType	0	1	0.65	0.48	0	0	1	1	1	
TestPrep	0	1	0.94	0.24	0	1	1	1	1	
ParentMaritalStatusdivorced	0	1	0.15	0.36	0	0	0	0	1	
ParentMaritalStatusmarried	0	1	0.55	0.50	0	0	1	1	1	
ParentMaritalStatussingle	0	1	0.23	0.42	0	0	0	0	1	
ParentMaritalStatuswidowed	0	1	0.03	0.16	0	0	0	0	1	
PracticeSportnever	0	1	0.12	0.32	0	0	0	0	1	
PracticeSportregularly	0	1	0.36	0.48	0	0	0	1	1	
PracticeSportsometimes	0	1	0.50	0.50	0	0	1	1	1	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
IsFirstChild	0	1	0.97	0.17	0	1	1	1	1	
NrSiblings	0	1	2.16	1.48	0	1	2	3	7	
TransportMeans	0	1	0.89	0.31	0	1	1	1	1	
WklyStudyHours..5	0	1	0.27	0.44	0	0	0	1	1	
WklyStudyHours..10	0	1	0.16	0.36	0	0	0	0	1	
WklyStudyHours10.May	0	1	0.54	0.50	0	0	1	1	1	
MathScore	0	1	66.03	15.55	0	56	66	76	100	
ReadingScore	0	1	68.80	14.82	17	59	69	79	100	
WritingScore	0	1	67.85	15.35	10	57	68	78	100	

```r
# Distributions of Test Scores
# Histograms for each test score
data %>%
  select(MathScore, ReadingScore, WritingScore) %>%
  pivot_longer(everything(), names_to = "Test", values_to = "Score") %>%
  ggplot(aes(x = Score, fill = Test)) +
  geom_histogram(binwidth = 5, alpha = 0.7, position = "dodge") +
  labs(title = "Distributions of Test Scores", x = "Score", y = "Frequency") +
  theme_minimal()
```



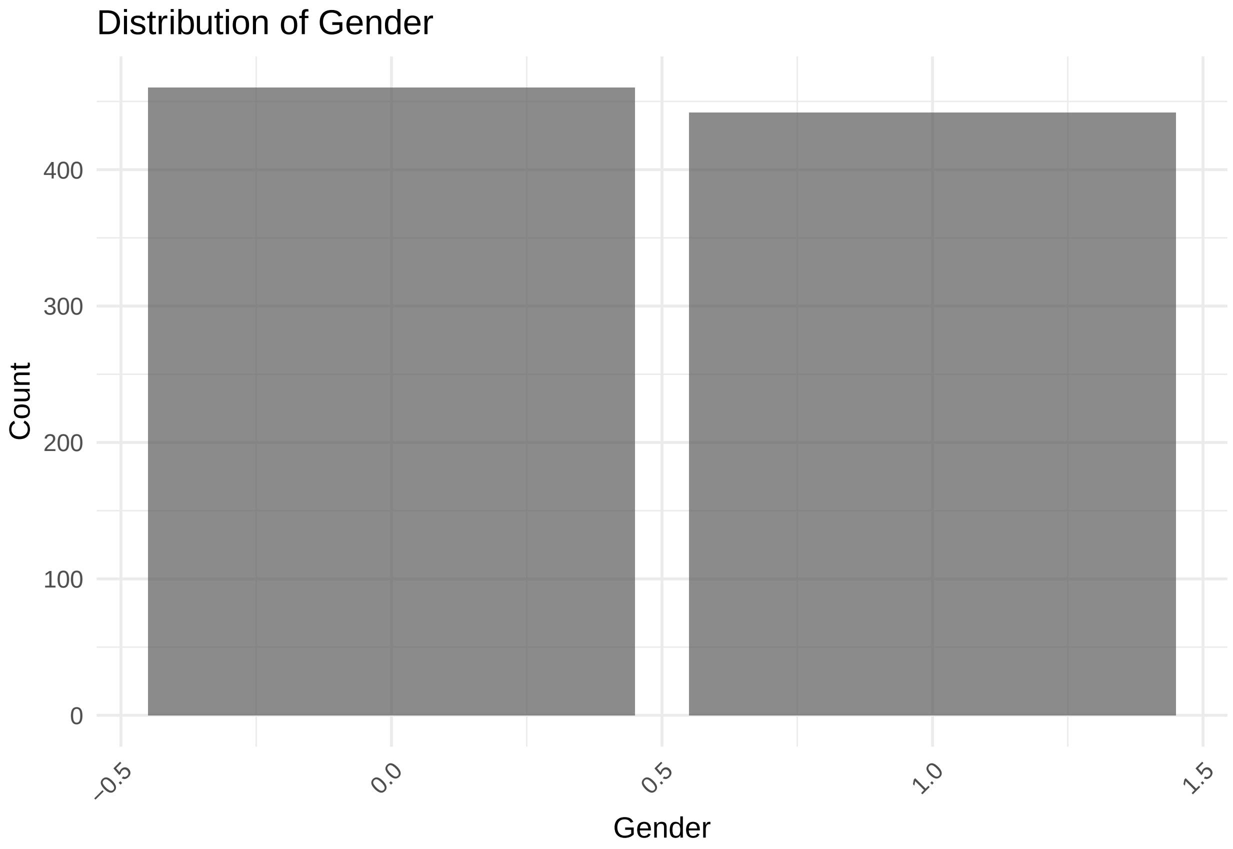
```r
# Boxplots for test scores
data %>%
  select(MathScore, ReadingScore, WritingScore) %>%
  pivot_longer(everything(), names_to = "Test", values_to = "Score") %>%
  ggplot(aes(x = Test, y = Score, fill = Test)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Boxplots of Test Scores", x = "Test", y = "Score") +
  theme_minimal()
```



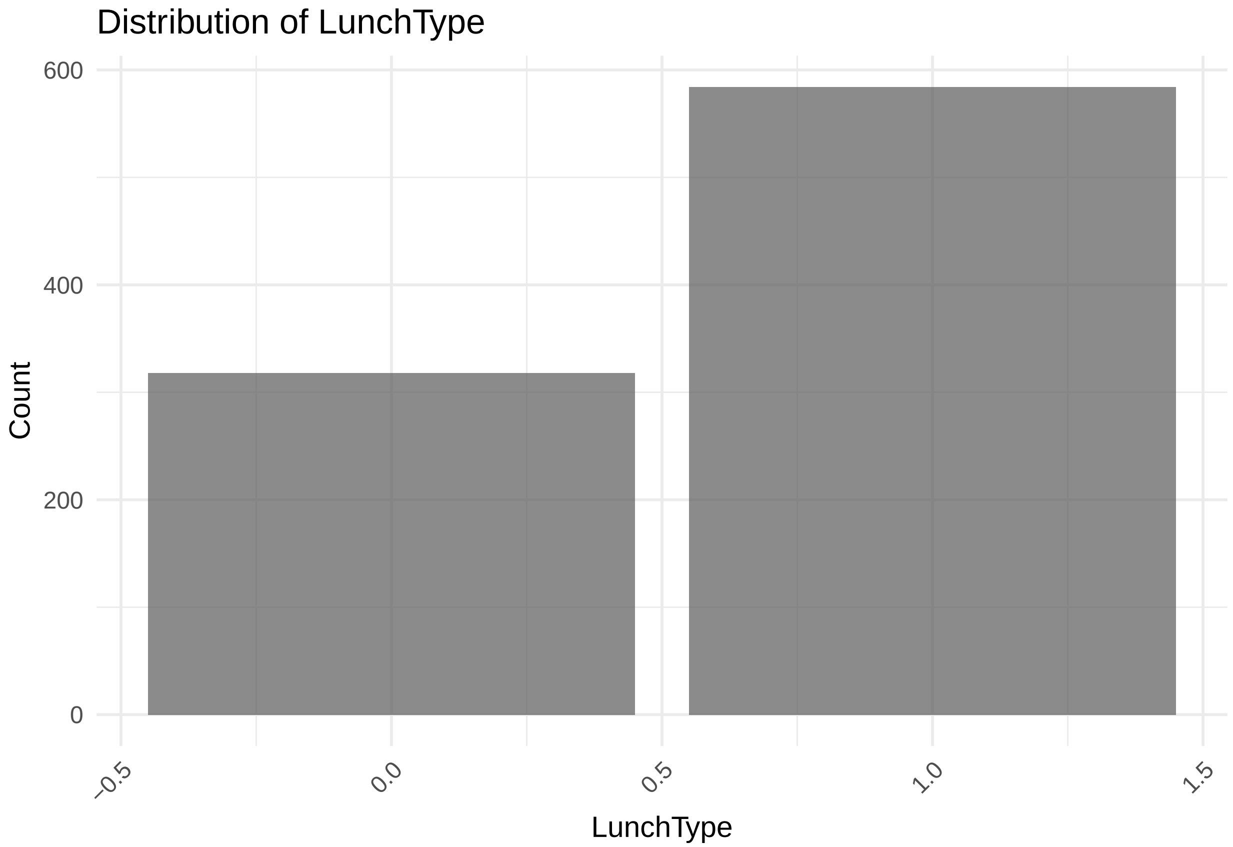
```r
# Distributions of Categorical Covariates
# Identify original categorical variables in the data
categorical_vars <- c("Gender", "LunchType", "TestPrep", "IsFirstChild")

# Bar plots for categorical variables
for (var in categorical_vars) {
  print(
    ggplot(data, aes_string(x = var, fill = var)) +
      geom_bar(alpha = 0.7) +
      labs(title = paste("Distribution of", var), x = var, y = "Count") +
      theme_minimal() +
      theme(axis.text.x = element_text(angle = 45, hjust = 1))
  )
}
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

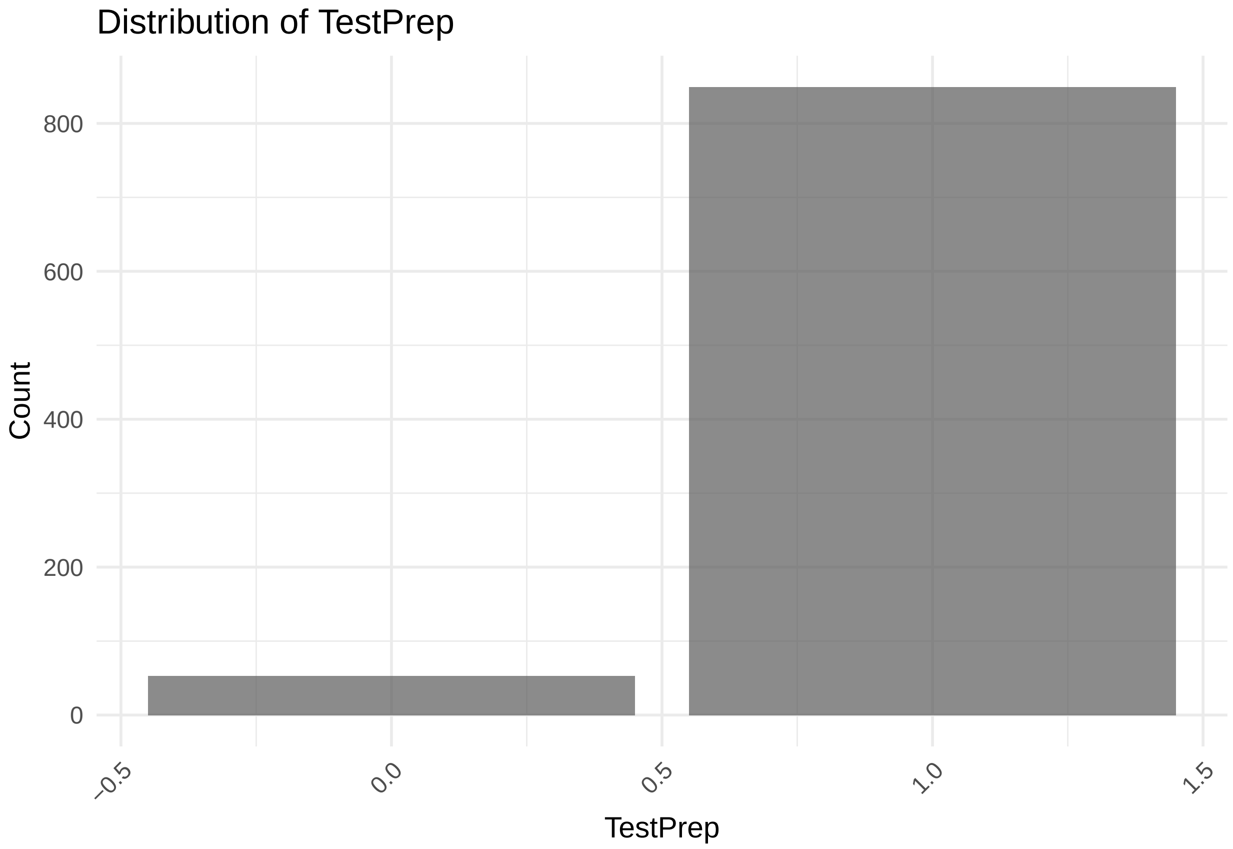
```
## Warning: The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



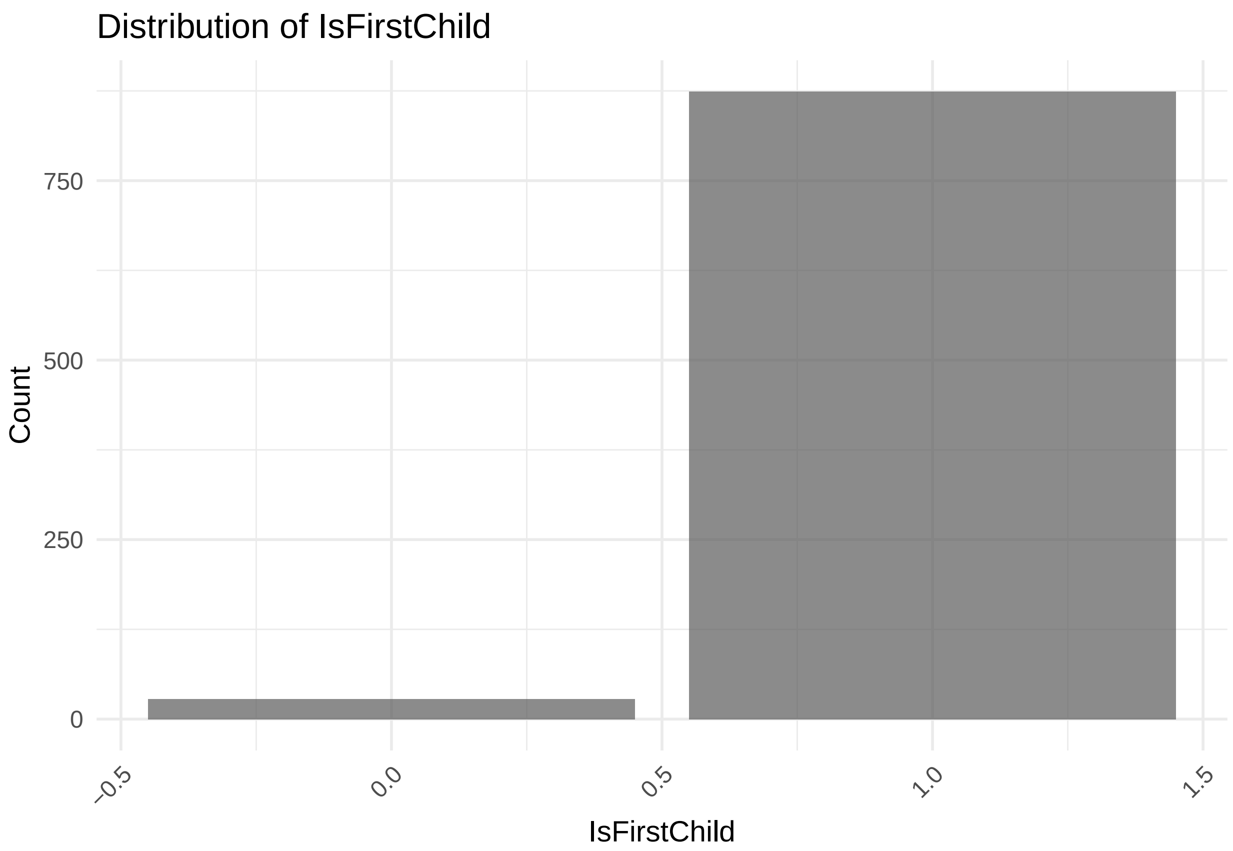
```
## Warning: The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



```
## Warning: The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



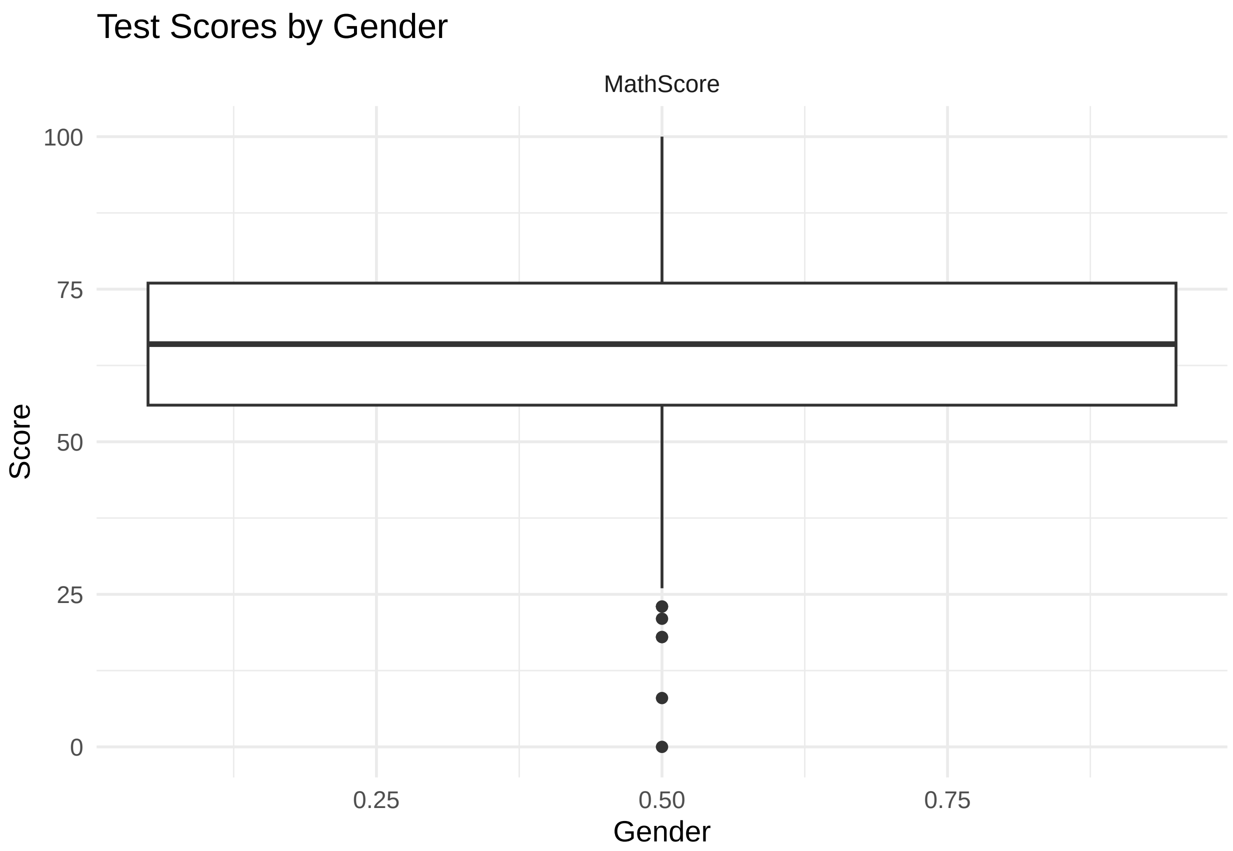
```
## Warning: The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



```r
# Relationships Between Test Scores and Key Covariates
# Boxplots for test scores by Gender
data %>%
  select(Gender, MathScore, ReadingScore, WritingScore) %>%
  pivot_longer(cols = starts_with("MathScore"), names_to = "Test", values_to = "Score") %>%
  ggplot(aes(x = Gender, y = Score, fill = Gender)) +
  geom_boxplot() +
  facet_wrap(~ Test) +
  labs(title = "Test Scores by Gender", x = "Gender", y = "Score") +
  theme_minimal()
```

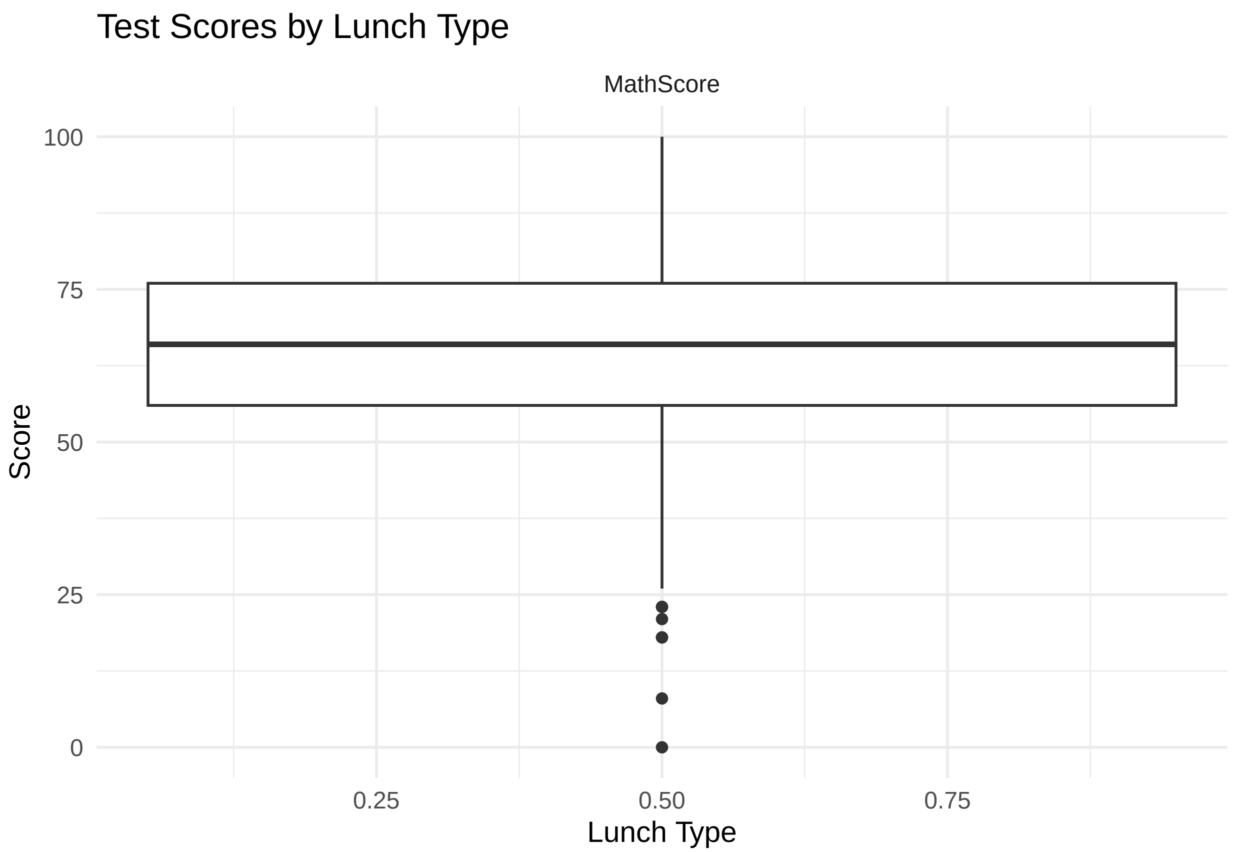
```
## Warning: Continuous x aesthetic
## i did you forget `aes(group = ...)`?
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

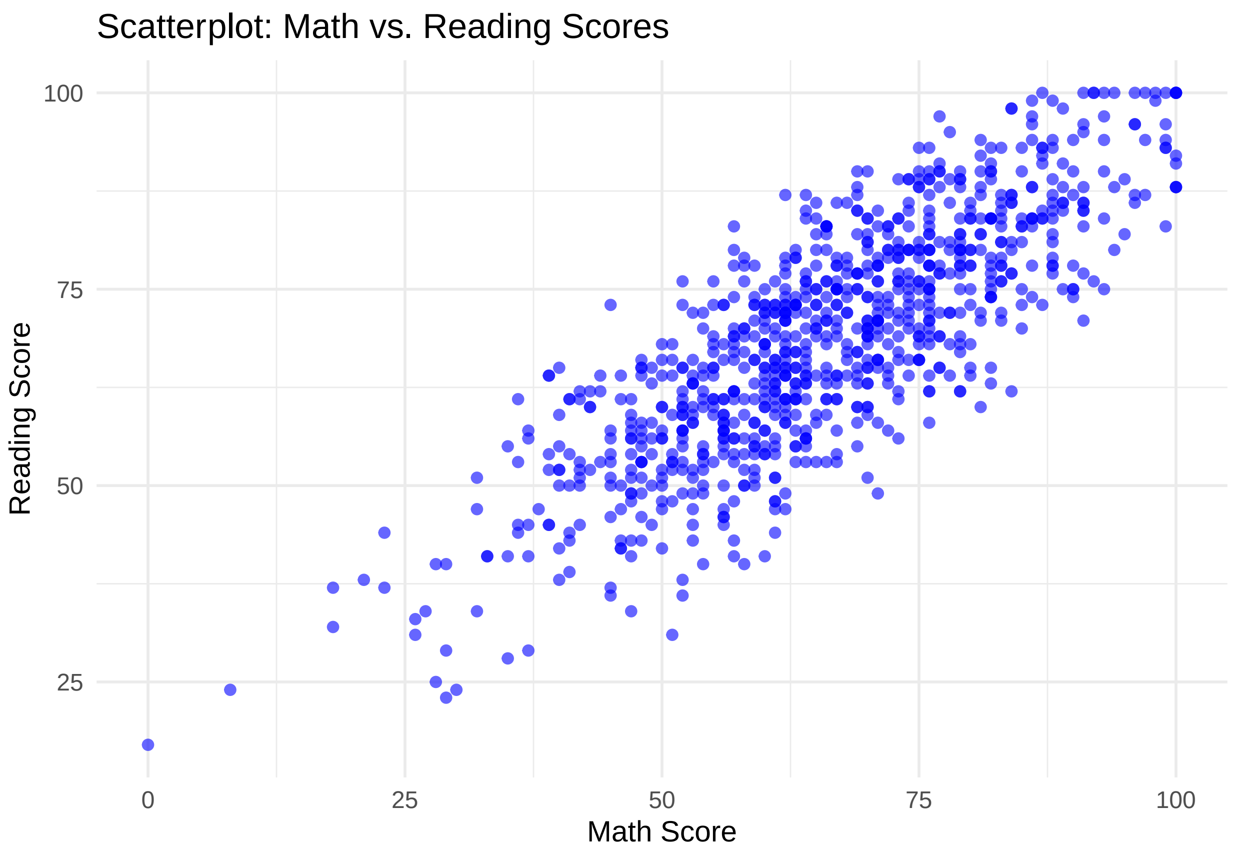


```
# Boxplots for test scores by LunchType
data %>%
  select(LunchType, MathScore, ReadingScore, WritingScore) %>%
  pivot_longer(cols = starts_with("MathScore"), names_to = "Test", values_to = "Score") %>%
  ggplot(aes(x = LunchType, y = Score, fill = LunchType)) +
  geom_boxplot() +
  facet_wrap(~ Test) +
  labs(title = "Test Scores by Lunch Type", x = "Lunch Type", y = "Score") +
  theme_minimal()
```

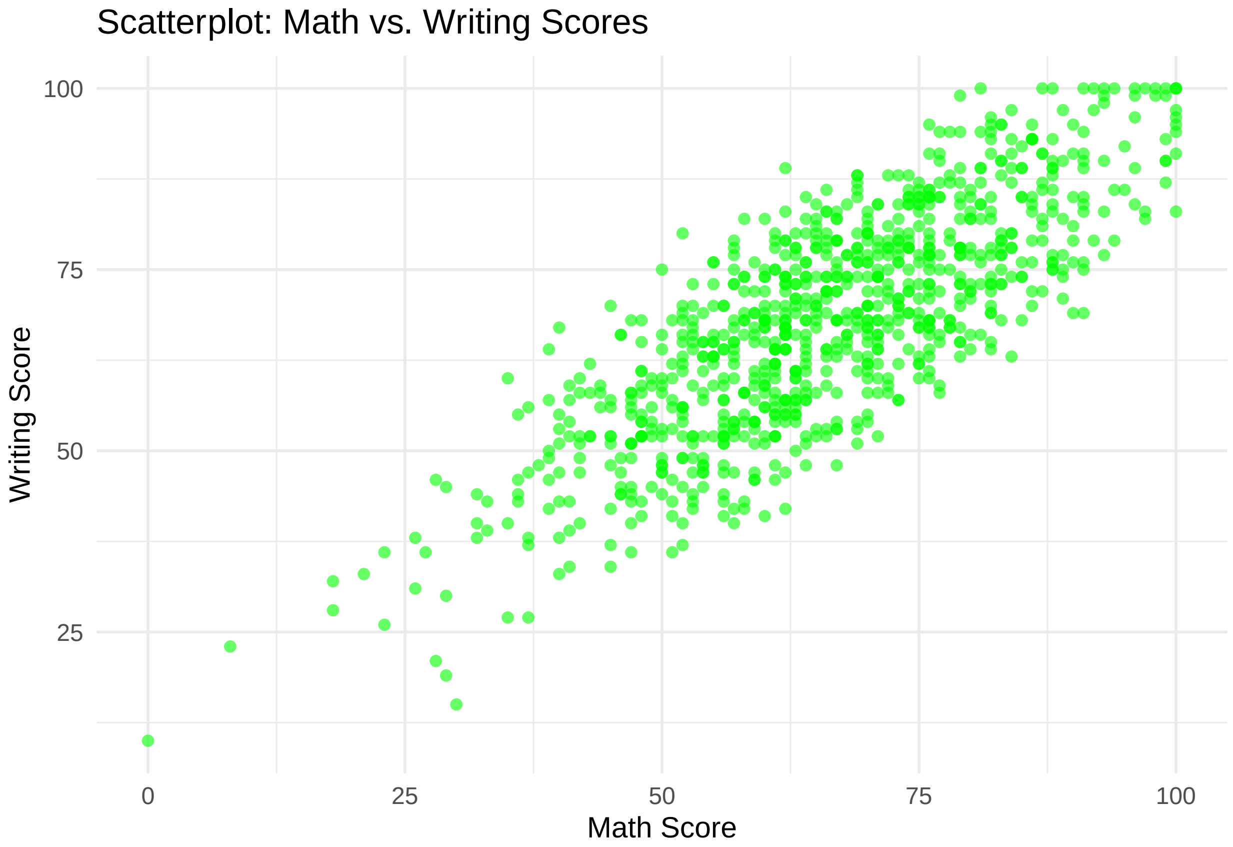
```
## Warning: Continuous x aesthetic
## i did you forget `aes(group = ...)`?
## The following aesthetics were dropped during statistical transformation: fill.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



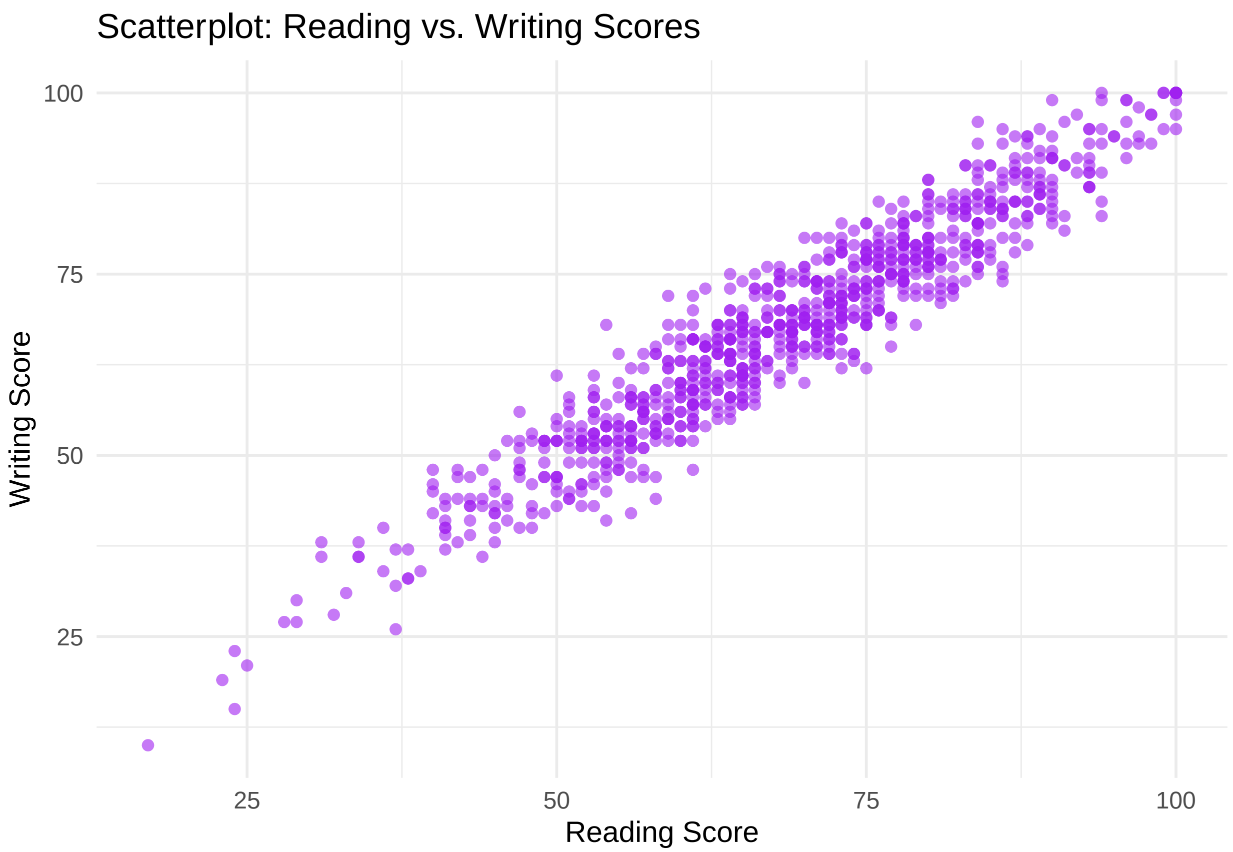
```r
# Pairwise Relationships
# Scatterplots of test scores
data %>%
  ggplot(aes(x = MathScore, y = ReadingScore)) +
  geom_point(alpha = 0.6, color = "blue") +
  labs(title = "Scatterplot: Math vs. Reading Scores", x = "Math Score", y = "Reading Score") +
  theme_minimal()
```



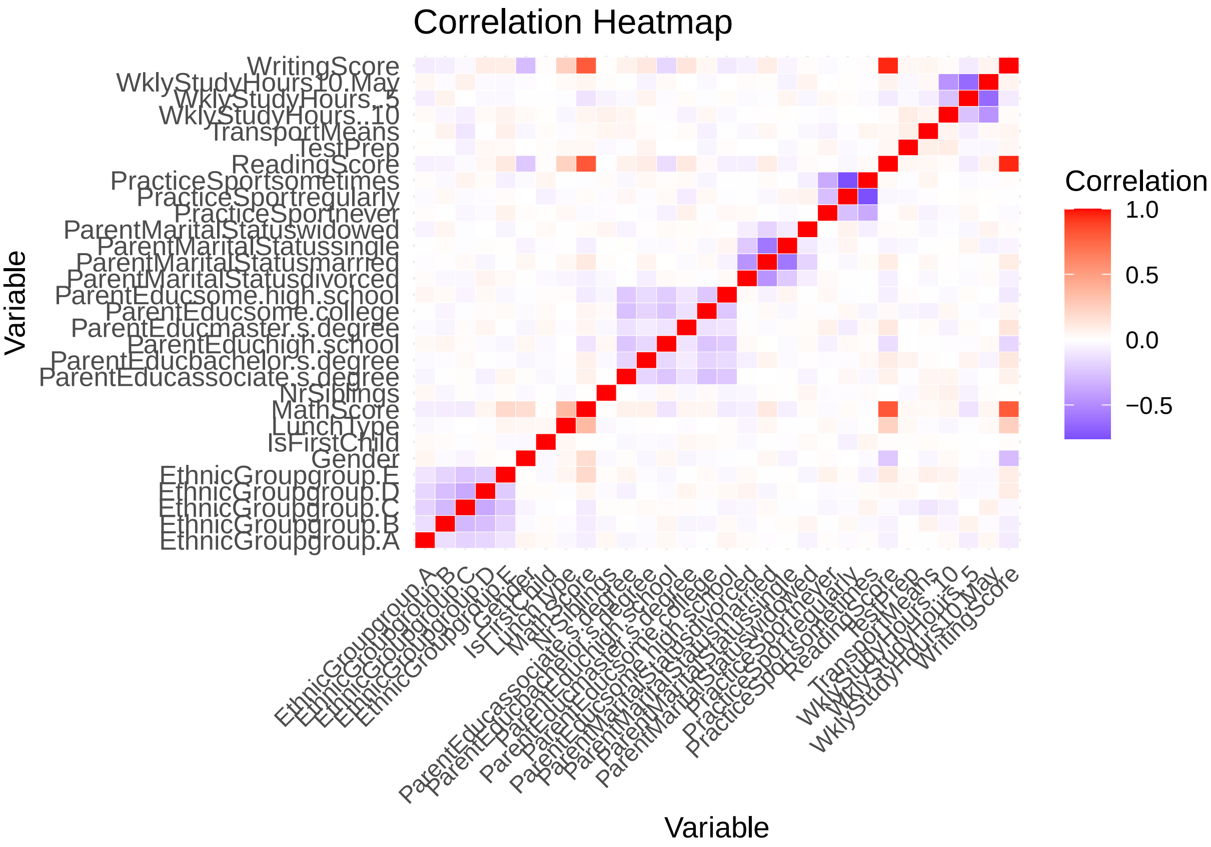
```
data %>%
  ggplot(aes(x = MathScore, y = WritingScore)) +
  geom_point(alpha = 0.6, color = "green") +
  labs(title = "Scatterplot: Math vs. Writing Scores", x = "Math Score", y = "Writing Score") +
  theme_minimal()
```



```r
data %>%
  ggplot(aes(x = ReadingScore, y = WritingScore)) +
  geom_point(alpha = 0.6, color = "purple") +
  labs(title = "Scatterplot: Reading vs. Writing Scores", x = "Reading Score", y = "Writing Score") +
  theme_minimal()
```



```r
# Correlation heatmap
numeric_vars <- data %>% select(where(is.numeric))
corr_matrix <- cor(numeric_vars, use = "complete.obs")

corr_matrix %>%
  as.data.frame() %>%
  rownames_to_column(var = "Variable1") %>%
  pivot_longer(cols = -Variable1, names_to = "Variable2", values_to = "Correlation") %>%
  ggplot(aes(x = Variable1, y = Variable2, fill = Correlation)) +
  geom_tile(color = "white") +  # Add grid lines
  scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0) +
  labs(title = "Correlation Heatmap", x = "Variable", y = "Variable", fill = "Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.text.y = element_text(size = 10))
```



Findings from Exploratory Data Analysis (EDA)

Pairwise Relationships:

Strong correlations between MathScore, ReadingScore, and WritingScore (r 0.95), suggesting redundancy in predictors for individual models. Weak correlation between NrSiblings and test scores. Visualizations indicate potential interaction effects, for example, between Gender and LunchType on MathScore.

Distributions:

Numeric variables like MathScore, ReadingScore, and WritingScore exhibit nearly normal distributions but with some skewness in scores below 50. NrSiblings is positively skewed with most values concentrated around 1 to 3.

Interactions and Covariate Effects:

Boxplots reveal that WklyStudyHours and EthnicGroup significantly impact test scores. Students with more than 10 hours of study time score higher across all test types.

Covariate Analysis:

Weekly study hours (WklyStudyHours) and test preparation (TestPrep) have clear separations in performance, suggesting strong predictive potential. Interaction plots highlight a differential impact of LunchType based on Gender.