p8130_final

Zihan Lin

R Markdown

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

# View 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 degree
##  $ 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 ...
```

```
# Check for missing values
sum(is.na(data))
```

```
## [1] 46
```

```
# Summarize missing data by variable
colSums(is.na(data))
```

```
##             Gender         EthnicGroup          ParentEduc           LunchType
##                  0                   0                   0                   0
##            TestPrep ParentMaritalStatus       PracticeSport       IsFirstChild
##                  0                   0                   0                   0
##          NrSiblings      TransportMeans      WklyStudyHours           MathScore
##                 46                   0                   0                   0
##        ReadingScore        WritingScore
##                  0                   0
```

```r
# Handle missing values (example: mean imputation for numeric variables)
data <- data %>%
  mutate(across(where(is.numeric), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))

# Check for duplicate rows
duplicates <- data[duplicated(data), ]
print(duplicates)
```

```
##  [1] Gender              EthnicGroup         ParentEduc
##  [4] LunchType           TestPrep            ParentMaritalStatus
##  [7] PracticeSport       IsFirstChild        NrSiblings
## [10] TransportMeans      WklyStudyHours      MathScore
## [13] ReadingScore        WritingScore
## <0 rows> (or 0-length row.names)
```

```r
# Remove duplicate rows if any
data <- data[!duplicated(data), ]

# Ensure categorical variables are factors
data <- data %>%
  mutate(across(where(is.character), as.factor))

# Summary after cleaning
summary(data)
```

```
##     Gender     EthnicGroup                ParentEduc         LunchType
##  female:488          : 59                      : 53   free/reduced:331
##  male  :460   group A: 80   associate's degree:198   standard    :617
##               group B:171   bachelor's degree :104
##               group C:277   high school       :176
##               group D:237   master's degree   : 55
##               group E:124   some college      :199
##                             some high school  :163
##       TestPrep   ParentMaritalStatus   PracticeSport IsFirstChild
##           : 55           : 49                 : 16      : 30
##  completed:322   divorced:146        never    :112   no :314
##  none     :571   married :516        regularly:343   yes:604
##                  single  :213        sometimes:477
##                  widowed : 24
##
##
##    NrSiblings       TransportMeans WklyStudyHours   MathScore
##  Min.   :0.000             :102          : 37     Min.   :  0.00
##  1st Qu.:1.000   private   :337    < 5   :253     1st Qu.: 56.00
##  Median :2.000   school_bus:509    > 10  :150     Median : 66.00
##  Mean   :2.155                     10-May:508     Mean   : 65.98
##  3rd Qu.:3.000                                    3rd Qu.: 76.00
##  Max.   :7.000                                    Max.   :100.00
##
##   ReadingScore     WritingScore
##  Min.   : 17.00   Min.   : 10.00
##  1st Qu.: 59.00   1st Qu.: 57.00
```

```
##  Median : 69.50   Median : 68.00
##  Mean   : 68.84   Mean   : 67.93
##  3rd Qu.: 80.00   3rd Qu.: 78.25
##  Max.   :100.00   Max.   :100.00
##
```

```r
# Separate numeric and categorical variables
numeric_vars <- data %>%
  select(where(is.numeric))

categorical_vars <- data %>%
  select(where(is.factor))

# Summary for numeric variables
numeric_summary <- numeric_vars %>%
  summarise(across(everything(),
                   list(
                     Mean = ~mean(., na.rm = TRUE),
                     SD = ~sd(., na.rm = TRUE),
                     Min = ~min(., na.rm = TRUE),
                     Q1 = ~quantile(., 0.25, na.rm = TRUE),
                     Median = ~median(., na.rm = TRUE),
                     Q3 = ~quantile(., 0.75, na.rm = TRUE),
                     Max = ~max(., na.rm = TRUE)
                   ), .names = "{.col}_{.fn}")) %>%
  pivot_longer(cols = everything(),
               names_to = c("Variable", ".value"),
               names_sep = "_")

# Summary for categorical variables
categorical_summary <- categorical_vars %>%
  summarise(across(everything(),
                   ~paste(names(table(.)), ":", as.vector(table(.)), collapse = ", "),
                   .names = "{.col}")) %>%
  pivot_longer(cols = everything(),
               names_to = "Variable",
               values_to = "Levels_Values")

# Format numeric columns to two decimal places
format_numeric <- function(x) {
  if (is.numeric(x)) {
    round(x, 2)
  } else {
    x
  }
}

# Generate the summary table
summary_table <- bind_rows(
  numeric_vars %>%
    summarise(across(everything(),
                     list(Mean = ~mean(., na.rm = TRUE),
                          SD = ~sd(., na.rm = TRUE),
                          Min = ~min(., na.rm = TRUE),
```

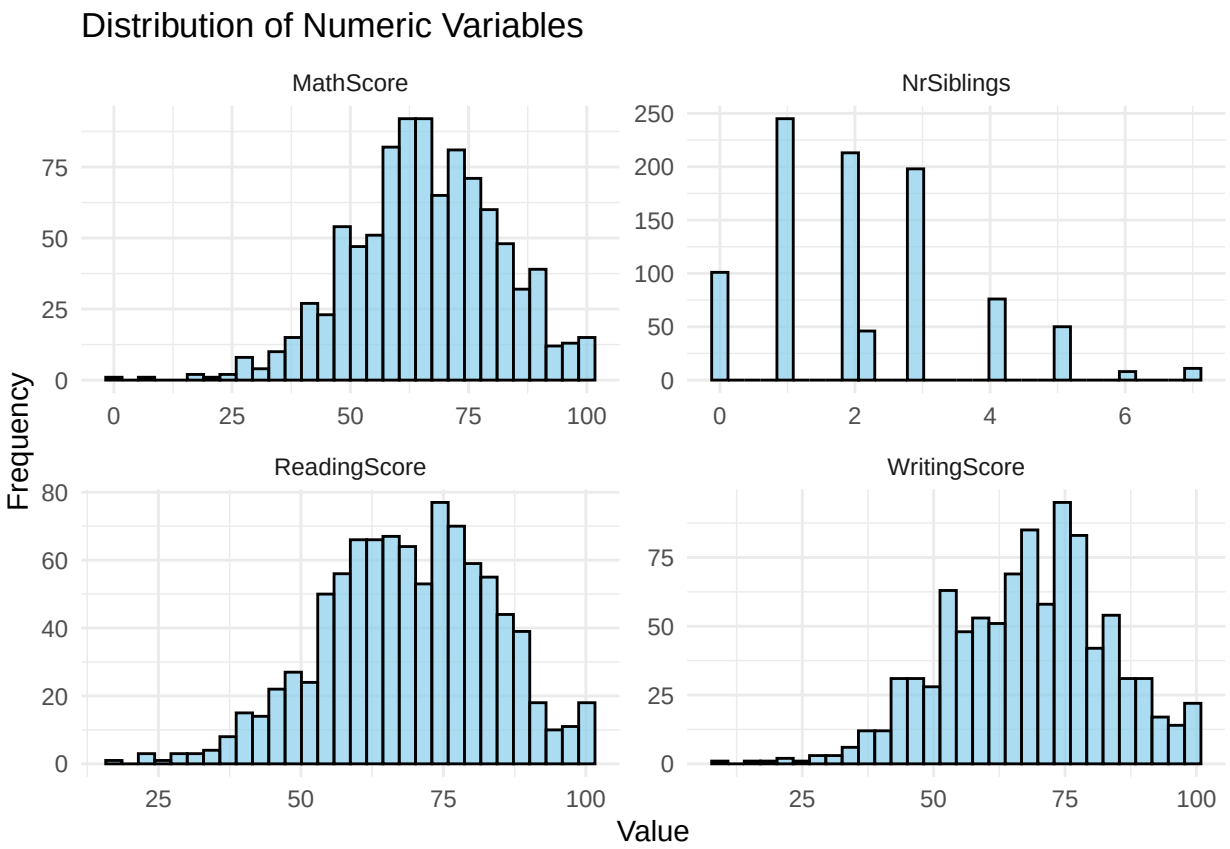
```
                          Q1 = ~quantile(., 0.25, na.rm = TRUE),
                          Median = ~median(., na.rm = TRUE),
                          Q3 = ~quantile(., 0.75, na.rm = TRUE),
                          Max = ~max(., na.rm = TRUE)),
                      .names = "{.col}_{.fn}")) %>%
    pivot_longer(cols = everything(),
                 names_to = c("Variable", ".value"),
                 names_sep = "_") %>%
    mutate(Variable_Type = "Numeric"),
  categorical_vars %>%
    summarise(across(everything(),
                     ~paste(names(table(.)), ":", as.vector(table(.)), collapse = "; "),
                     .names = "{.col}")) %>%
    pivot_longer(cols = everything(),
                 names_to = "Variable",
                 values_to = "Levels_Values") %>%
    mutate(Variable_Type = "Categorical", Mean = NA, SD = NA, Min = NA, Q1 = NA, Median = NA, Q3 = NA, M
)

# Wrap text in the Levels_Values column
summary_table$Levels_Values <- str_wrap(summary_table$Levels_Values, width = 40)

# Apply numeric formatting to two decimal places
summary_table <- summary_table %>%
  mutate(across(c(Mean, SD, Min, Q1, Median, Q3, Max), format_numeric))

# Create the final table
summary_table %>%
  select(Variable, Variable_Type, Levels_Values, Mean, SD, Min, Q1, Median, Q3, Max) %>%
  kable(format = "latex", booktabs = TRUE, caption = "Descriptive Summary Statistics") %>%
  kable_styling(latex_options = c("striped", "hold_position")) %>%
  column_spec(1, width = "3cm") %>%  # Adjust width for Variable
  column_spec(2, width = "2.5cm") %>%  # Adjust width for Variable_Type
  column_spec(3, width = "2.5cm") %>%  # Adjust width for Levels_Values
  column_spec(4:10, width = "0.8cm") %>%  # Adjust widths for numeric summaries
  row_spec(0, bold = TRUE) %>%  # Bold header row
  footnote(general = "Numeric variables display statistical summaries; categorical variables list levels
```

```
# Distribution of Numeric Variables
# Plot histograms and density plots for numeric variables
numeric_vars %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Value") %>%
  ggplot(aes(x = Value)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black", alpha = 0.7) +
  facet_wrap(~Variable, scales = "free", ncol = 2) +
  labs(title = "Distribution of Numeric Variables",
       x = "Value",
       y = "Frequency") +
  theme_minimal()
```

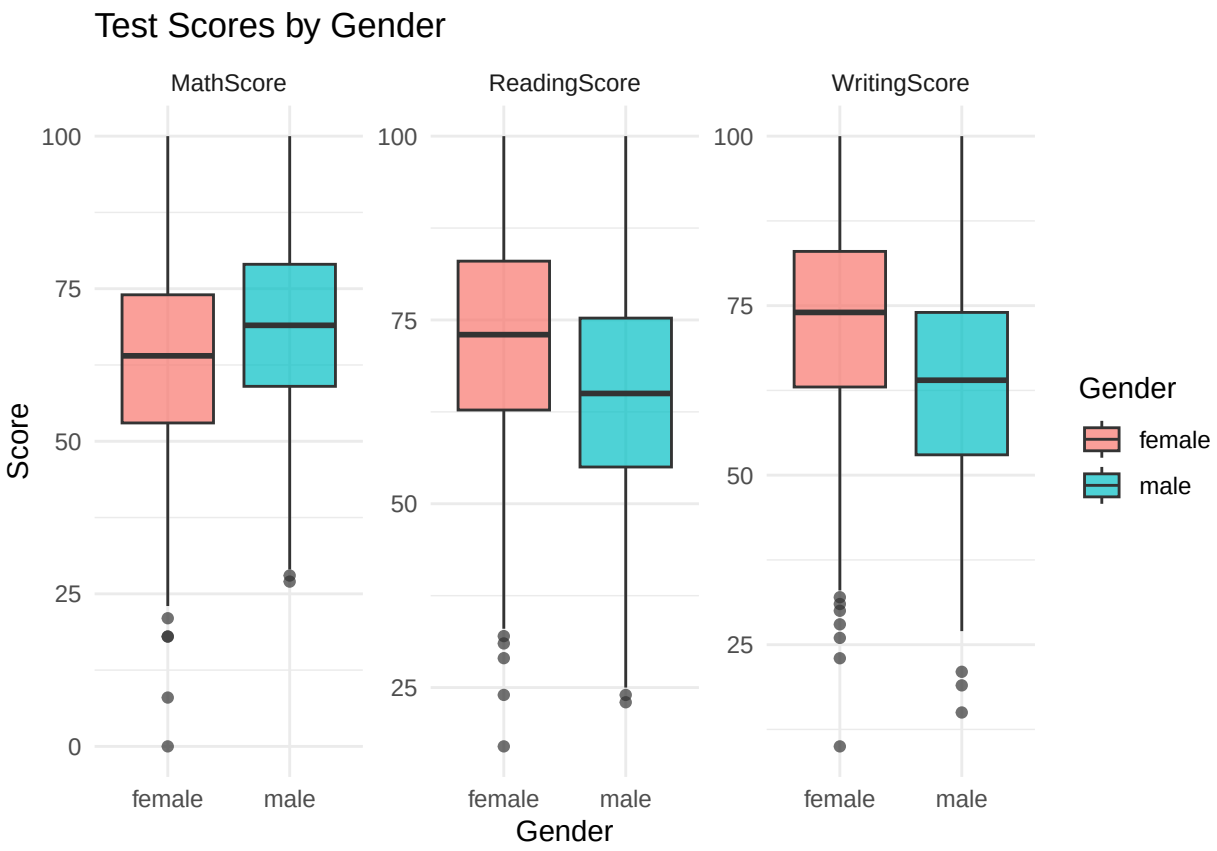
Table 1: Descriptive Summary Statistics

Variable	Variable_Type	Levels_Values	Mean	SD	Min	Q1	Median	Q3	Max
NrSiblings	Numeric	NA	2.16	1.45	0	1	2.0	3.00	7
MathScore	Numeric	NA	65.98	15.53	0	56	66.0	76.00	100
ReadingScore	Numeric	NA	68.84	14.80	17	59	69.5	80.00	100
WritingScore	Numeric	NA	67.93	15.41	10	57	68.0	78.25	100
Gender	Categorical	female : 488; male : 460	NA	NA	NA	NA	NA	NA	NA
EthnicGroup	Categorical	: 59; group A : 80; group B : 171; group C : 277; group D : 237; group E : 124	NA	NA	NA	NA	NA	NA	NA
ParentEduc	Categorical	: 53; associate's degree : 198; bachelor's degree : 104; high school : 176; master's degree : 55; some college : 199; some high school : 163	NA	NA	NA	NA	NA	NA	NA
LunchType	Categorical	free/reduced : 331; standard : 617	NA	NA	NA	NA	NA	NA	NA
TestPrep	Categorical	: 55; completed : 322; none : 571	NA	NA	NA	NA	NA	NA	NA
ParentMaritalStatus	Categorical	: 49; divorced : 146; married : 516; single : 213; widowed : 24	NA	NA	NA	NA	NA	NA	NA
PracticeSport	Categorical	: 16; never : 112; regularly : 343; sometimes : 477	NA	NA	NA	NA	NA	NA	NA
IsFirstChild	Categorical	: 30; no : 314; yes : 604	NA	NA	NA	NA	NA	NA	NA
TransportMeans	Categorical	: 102; private : 337; school_bus : 509	NA	NA	NA	NA	NA	NA	NA
WklyStudyHours	Categorical	: 37; < 5 : 253; > 10 : 150; 10-May : 508	NA	NA	NA	NA	NA	NA	NA

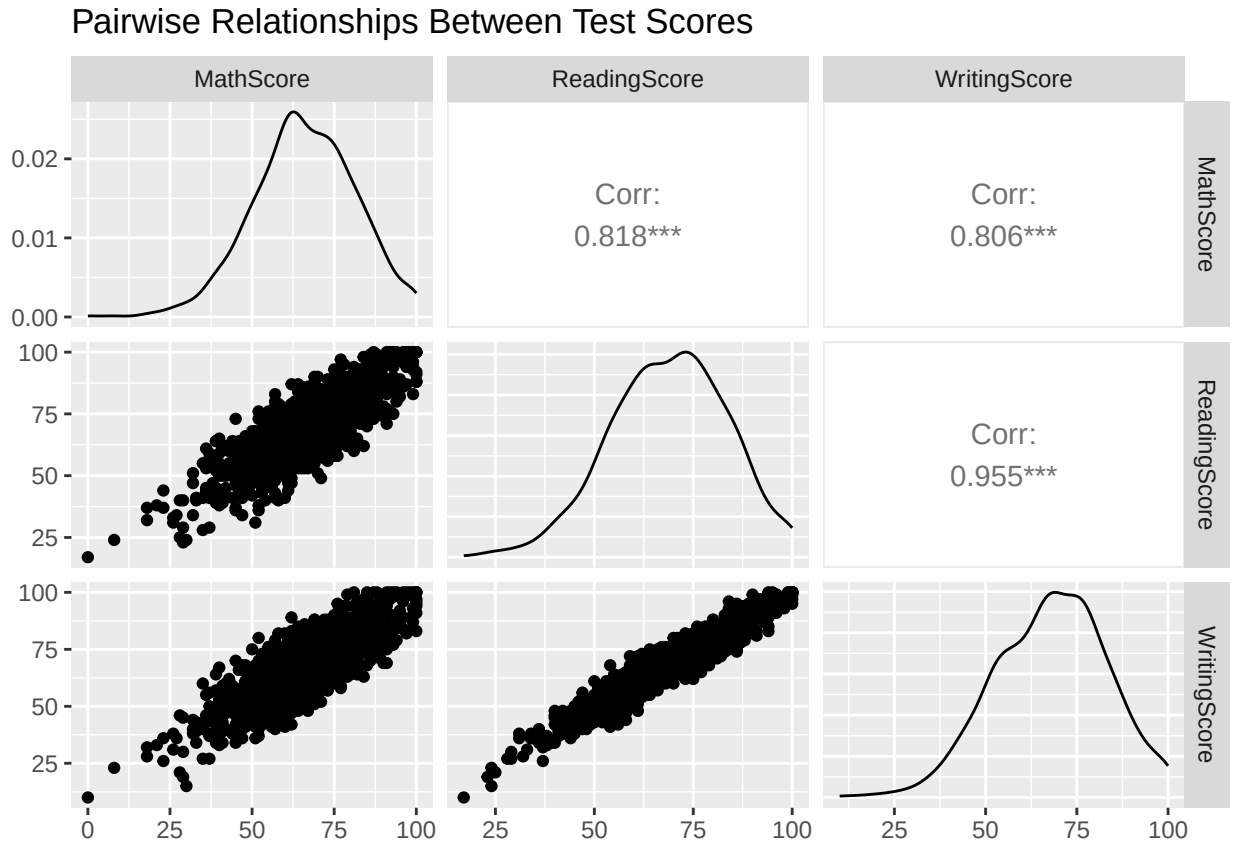
Note:
Numeric variables display statistical summaries; categorical variables list levels with counts.



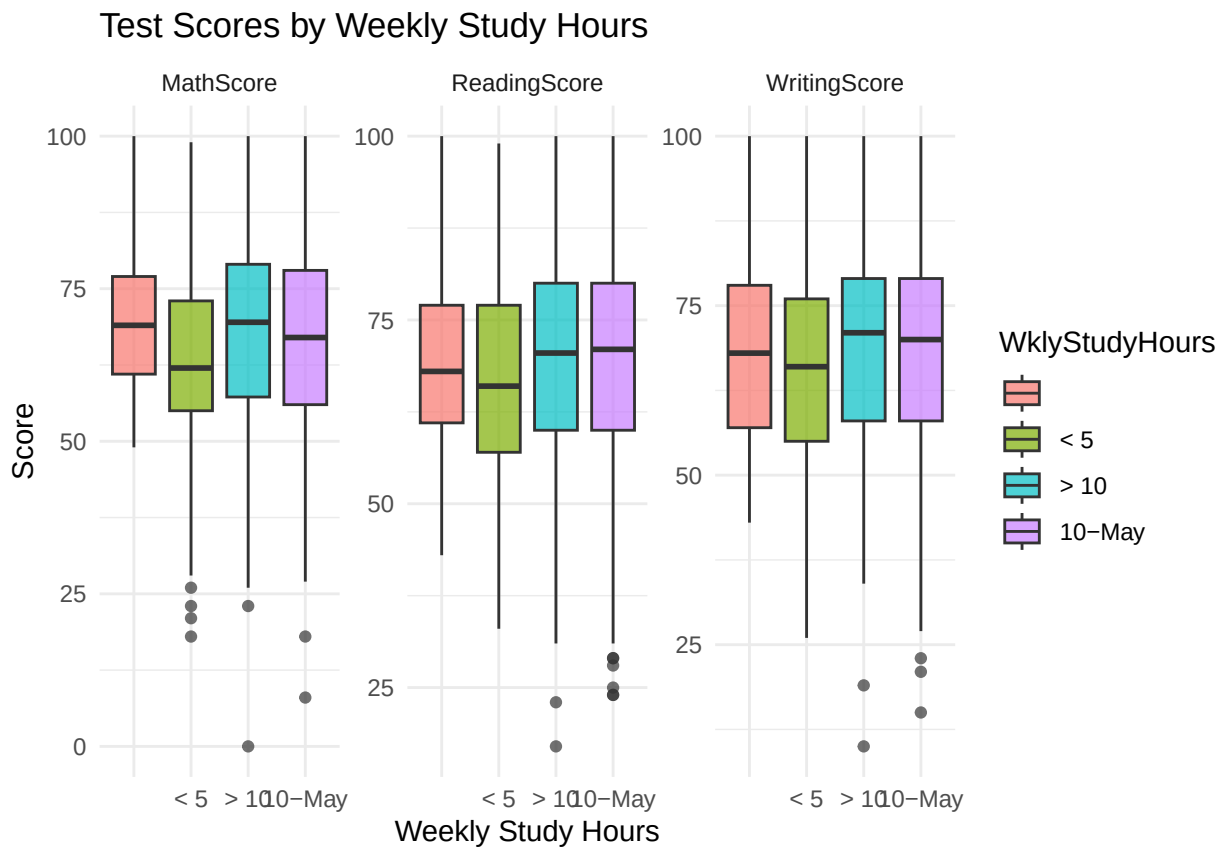

```
# Boxplots for Categorical Covariates vs Test Scores
# Explore relationships between categorical covariates and test scores
data %>%
  pivot_longer(cols = c(MathScore, ReadingScore, WritingScore),
               names_to = "TestType", values_to = "Score") %>%
  ggplot(aes(x = Gender, y = Score, fill = Gender)) +
  geom_boxplot(alpha = 0.7) +
  facet_wrap(~TestType, scales = "free") +
  labs(title = "Test Scores by Gender",
       x = "Gender",
       y = "Score") +
  theme_minimal()
```



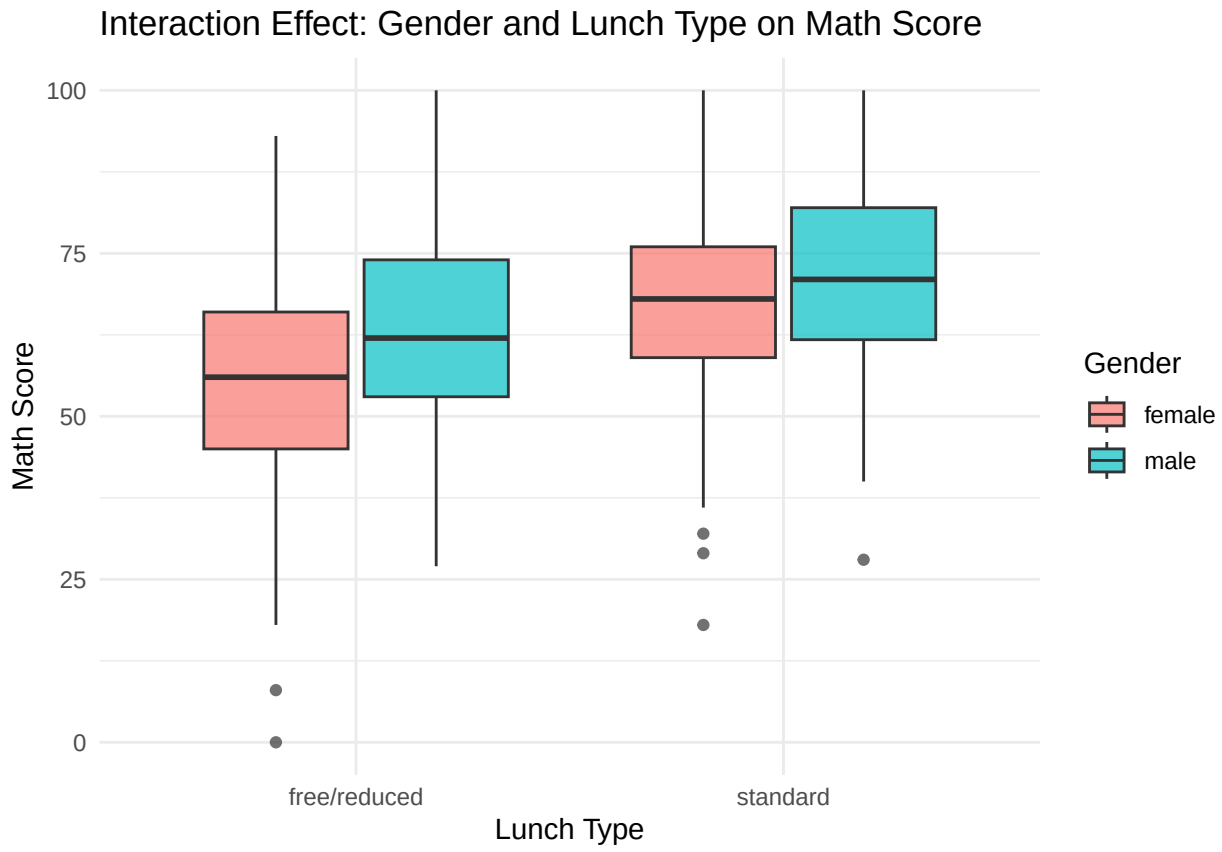

```
# Scatterplots to Explore Pairwise Relationships Between Test Scores
# Explore correlations between Math, Reading, and Writing Scores
ggpairs(data, columns = c("MathScore", "ReadingScore", "WritingScore"),
        title = "Pairwise Relationships Between Test Scores")
```



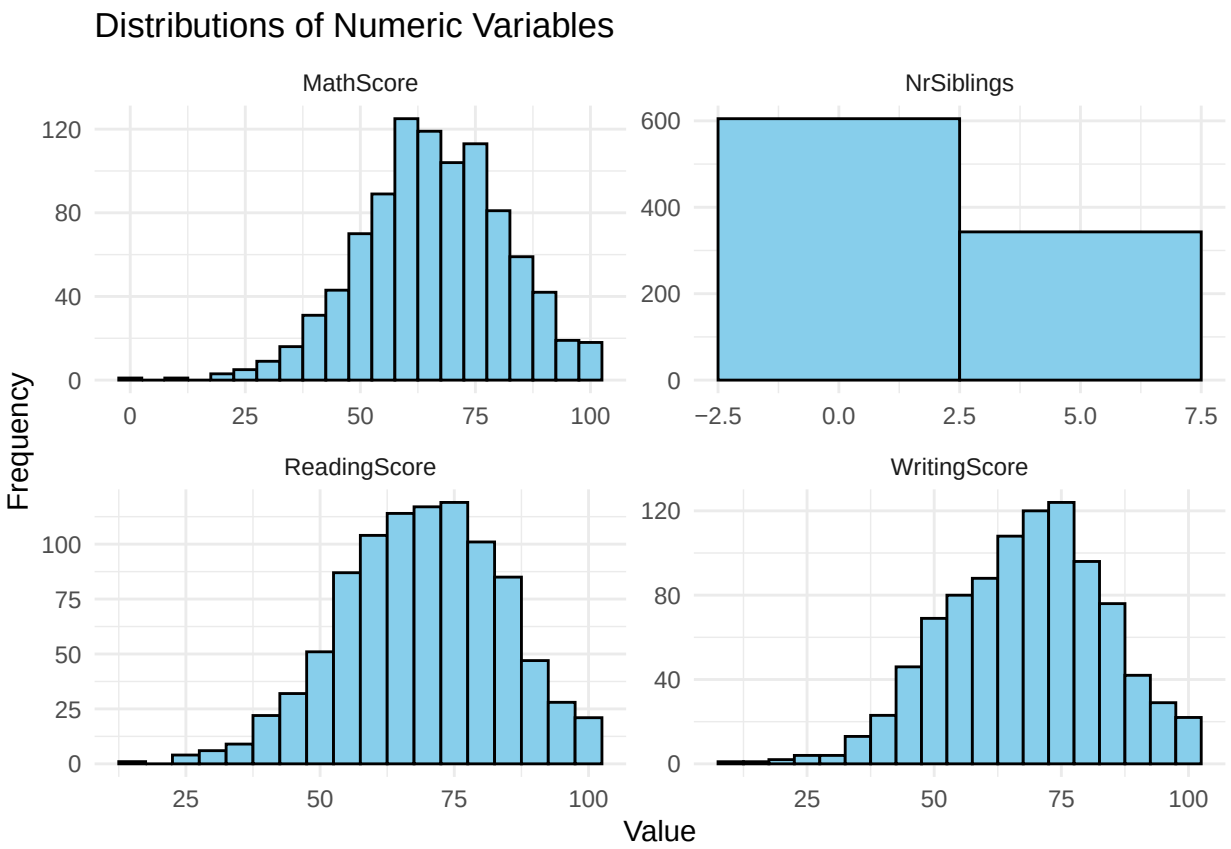

```
# Covariate vs. Weekly Study Hours
# Explore the effect of weekly study hours on test scores
data %>%
  pivot_longer(cols = c(MathScore, ReadingScore, WritingScore),
               names_to = "TestType", values_to = "Score") %>%
  ggplot(aes(x = WklyStudyHours, y = Score, fill = WklyStudyHours)) +
  geom_boxplot(alpha = 0.7) +
  facet_wrap(~TestType, scales = "free") +
  labs(title = "Test Scores by Weekly Study Hours",
       x = "Weekly Study Hours",
       y = "Score") +
  theme_minimal()
```



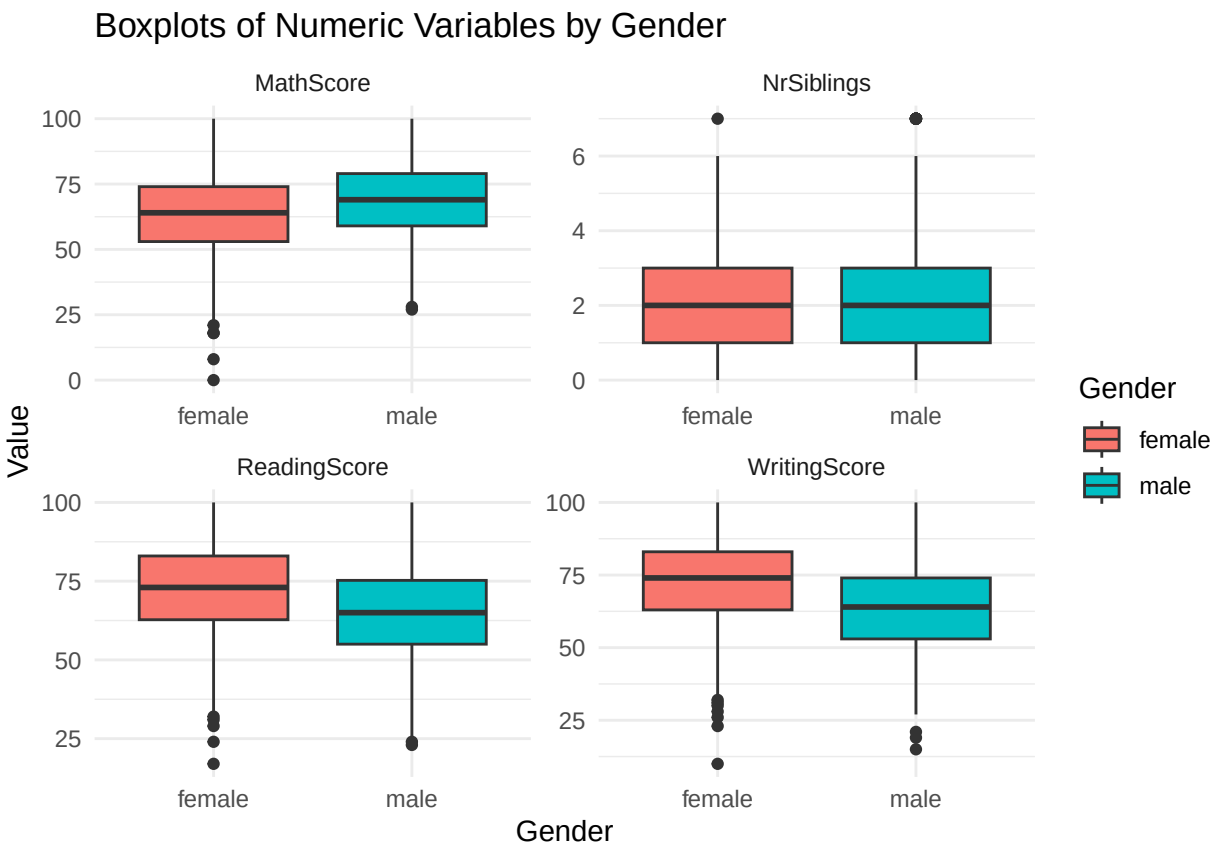

```r
# Check for Interaction Effects
# Example: Gender and LunchType interaction on MathScore
ggplot(data, aes(x = LunchType, y = MathScore, fill = Gender)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Interaction Effect: Gender and Lunch Type on Math Score",
       x = "Lunch Type",
       y = "Math Score") +
  theme_minimal()
```



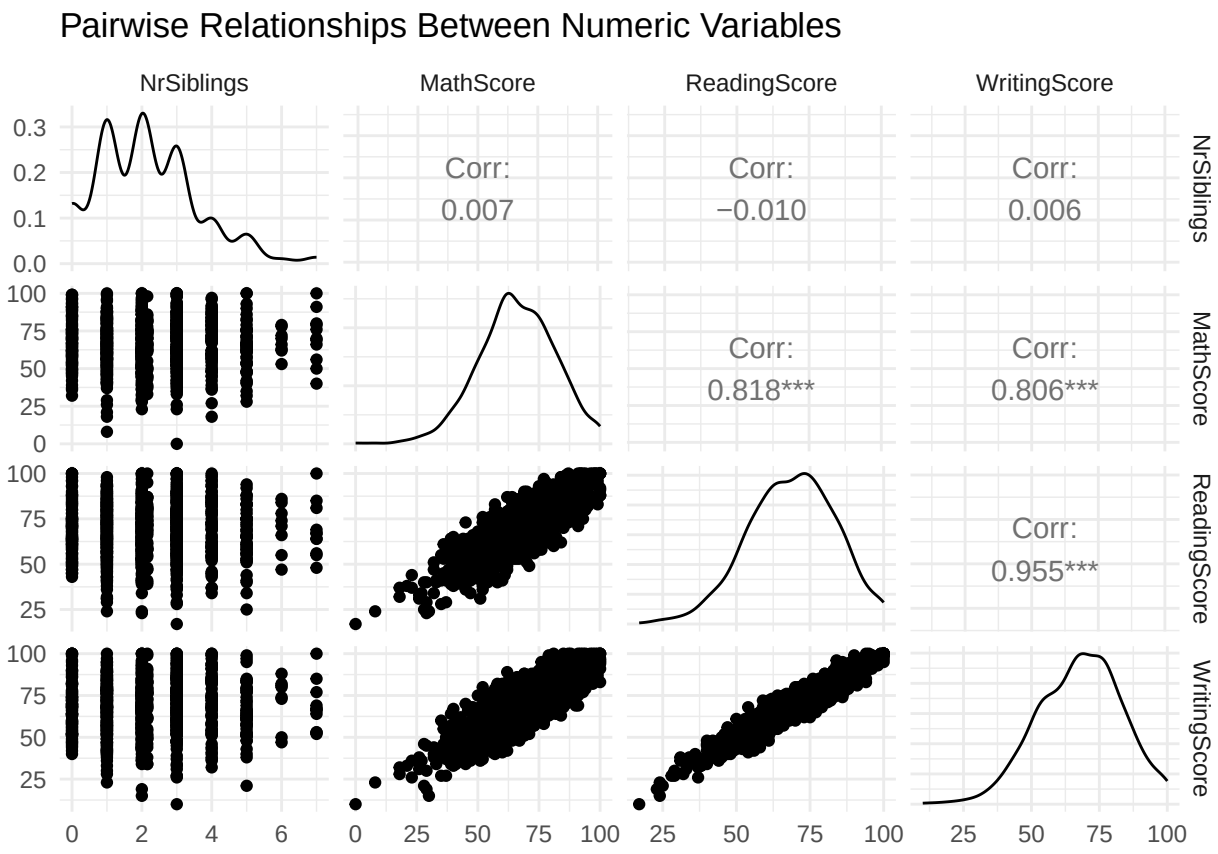

```
# Histograms for numeric variables
numeric_vars %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Value") %>%
  ggplot(aes(x = Value)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  facet_wrap(~ Variable, scales = "free", ncol = 2) +
  theme_minimal() +
  labs(title = "Distributions of Numeric Variables", x = "Value", y = "Frequency")
```



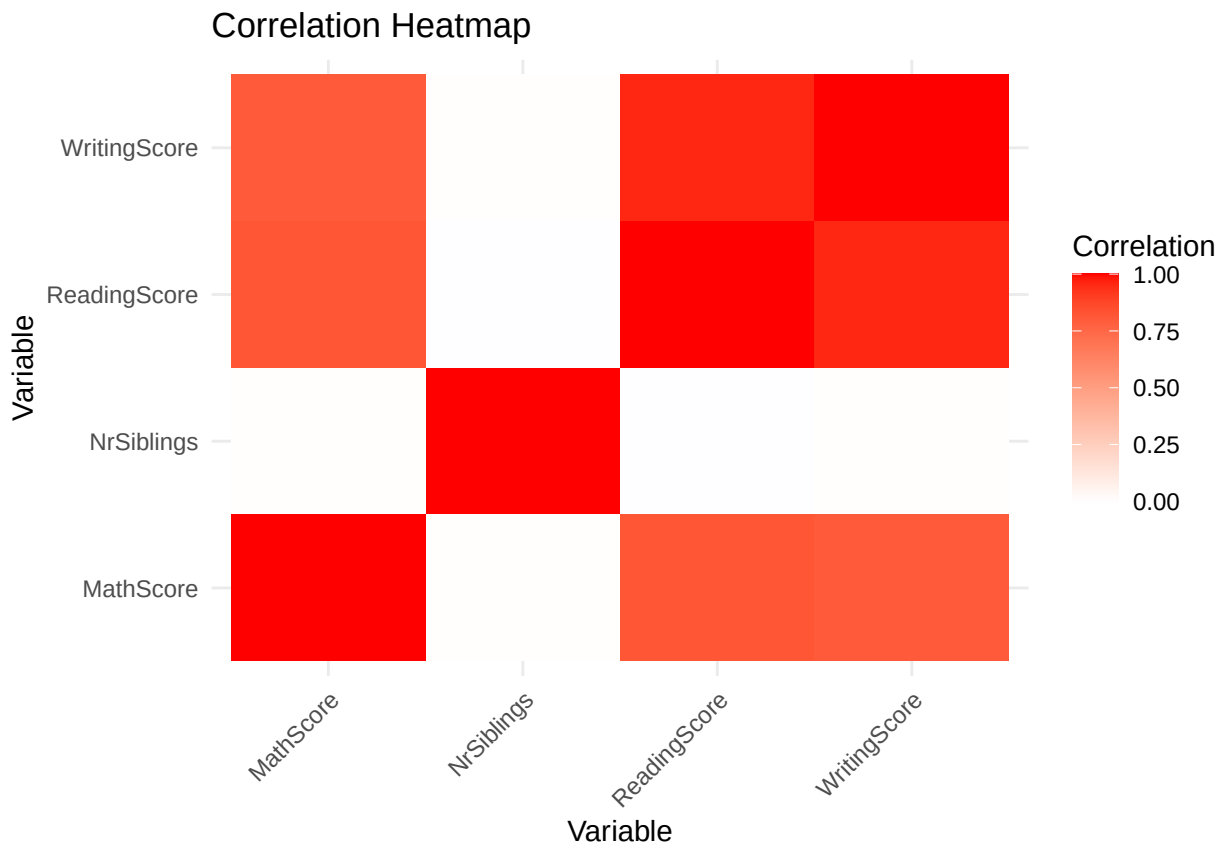

```r
# Boxplots for numeric variables by categorical covariates (e.g., Gender)
numeric_vars %>%
  bind_cols(data %>% select(Gender)) %>%
  pivot_longer(cols = -Gender, names_to = "Variable", values_to = "Value") %>%
  ggplot(aes(x = Gender, y = Value, fill = Gender)) +
  geom_boxplot() +
  facet_wrap(~ Variable, scales = "free", ncol = 2) +
  theme_minimal() +
  labs(title = "Boxplots of Numeric Variables by Gender", x = "Gender", y = "Value")
```



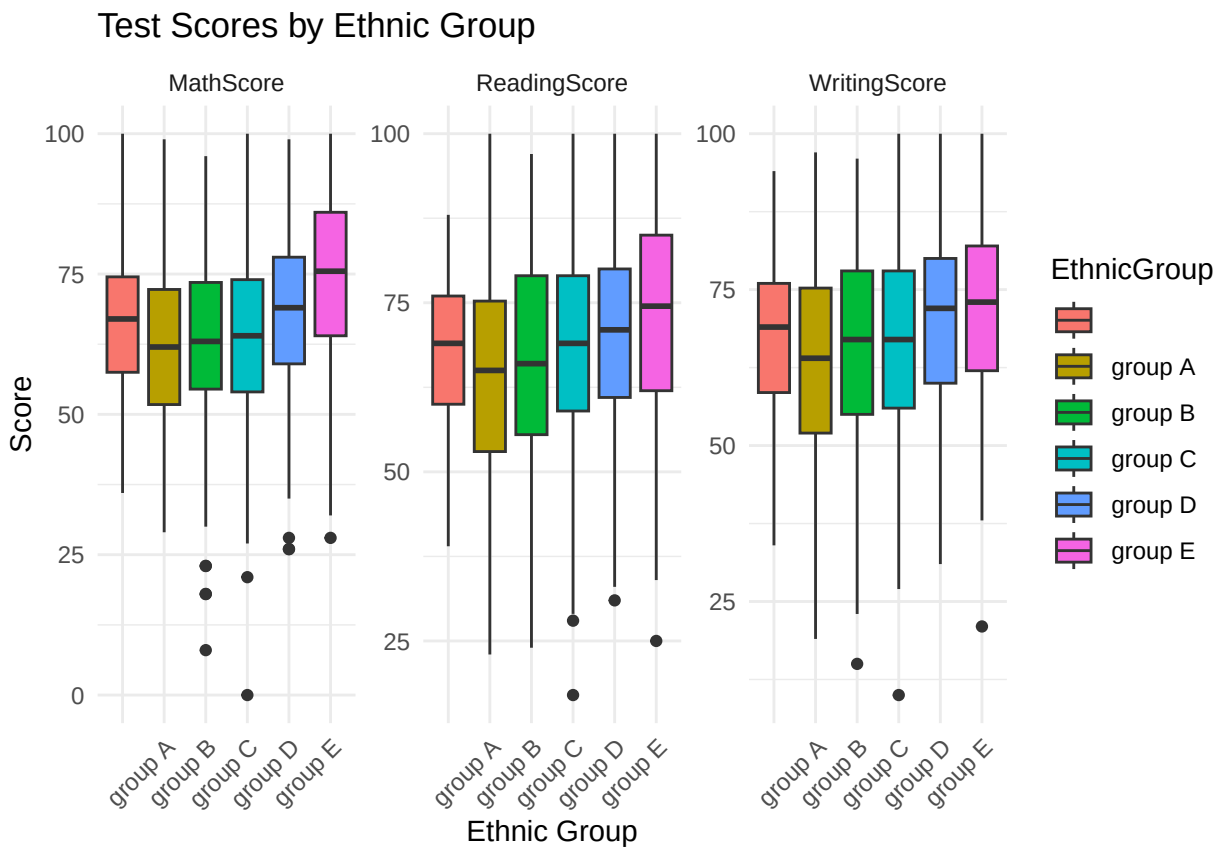

```
# Scatterplots for pairwise relationships
numeric_vars %>%
  GGally::ggpairs() +
  theme_minimal() +
  labs(title = "Pairwise Relationships Between Numeric Variables")
```




```r
# Correlation heatmap
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() +
  scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0) +
  theme_minimal() +
  labs(title = "Correlation Heatmap", x = "Variable", y = "Variable", fill = "Correlation") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Boxplots for test scores by EthnicGroup
data %>%
  pivot_longer(cols = c(MathScore, ReadingScore, WritingScore), names_to = "TestType", values_to = "Sco
  ggplot(aes(x = EthnicGroup, y = Score, fill = EthnicGroup)) +
  geom_boxplot() +
  facet_wrap(~ TestType, scales = "free") +
  theme_minimal() +
  labs(title = "Test Scores by Ethnic Group", x = "Ethnic Group", y = "Score") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


Test Scores by Ethnic Group
MathScore
ReadingScore
WritingScore
Score
100
75
50
25
0
group A
group B
group C
group D
group E
Ethnic Group
EthnicGroup