Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA)

Univariate Analysis

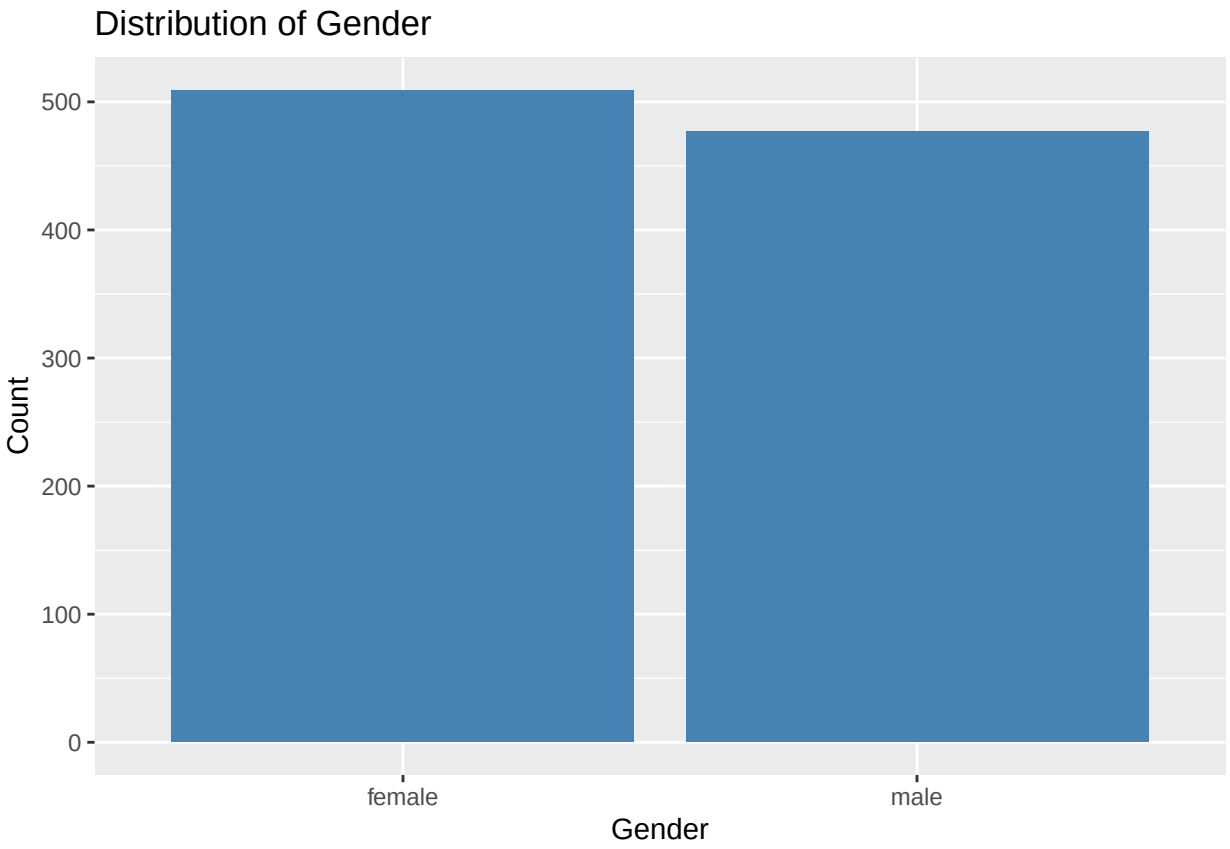
```
# Load the dataset
data <- read.csv("../data/study_performance_cleaned.csv")
head(data)
```

```
##   gender race_ethnicity parental_level_of_education        lunch
## 1 female        group B           bachelor's degree     standard
## 2 female        group C                some college     standard
## 3 female        group B             master's degree     standard
## 4   male        group A          associate's degree free/reduced
## 5   male        group C                some college     standard
## 6 female        group B          associate's degree     standard
##   test_preparation_course math_score reading_score writing_score
## 1                    none         72            72            74
## 2               completed         69            90            88
## 3                    none         90            95            93
## 4                    none         47            57            44
## 5                    none         76            78            75
## 6                    none         71            83            78
```

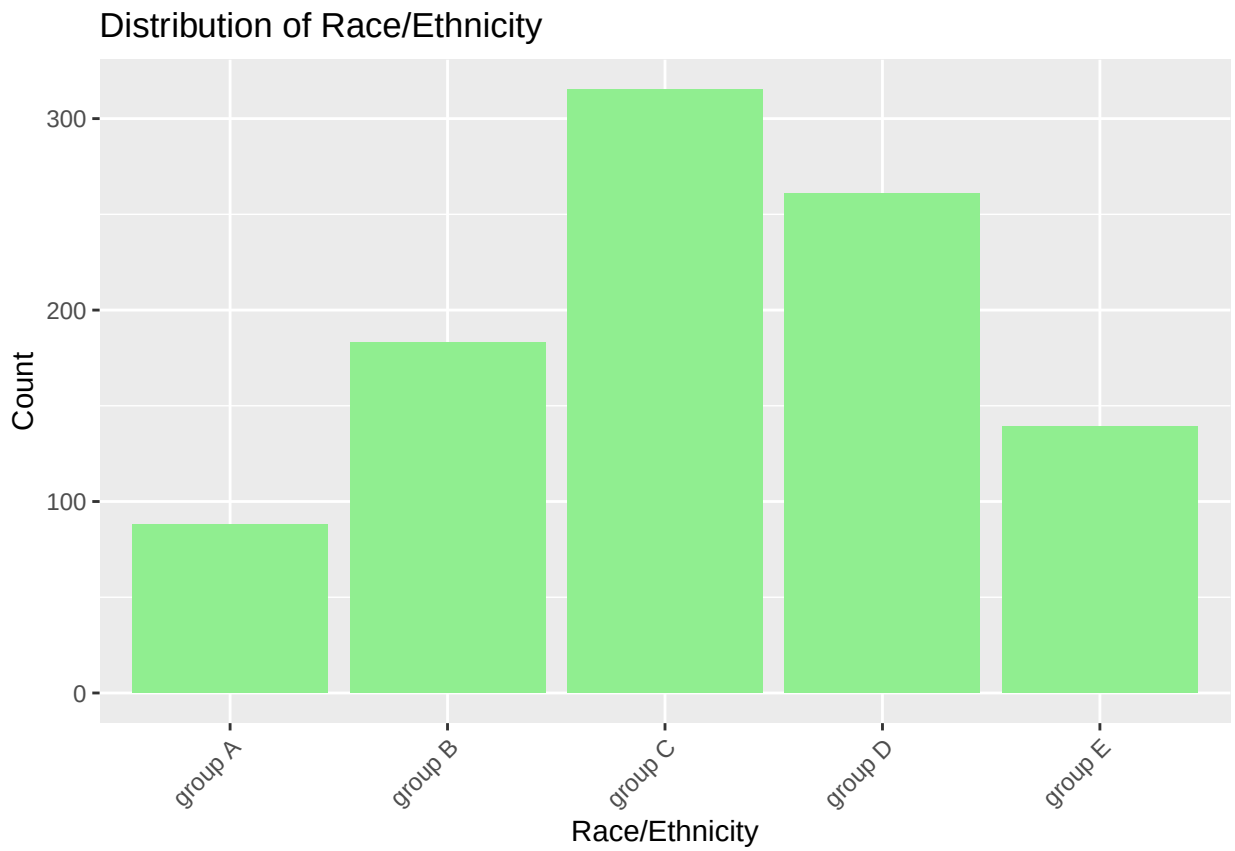
Explore distributions of numerical variables

Create bar plots for categorical variables

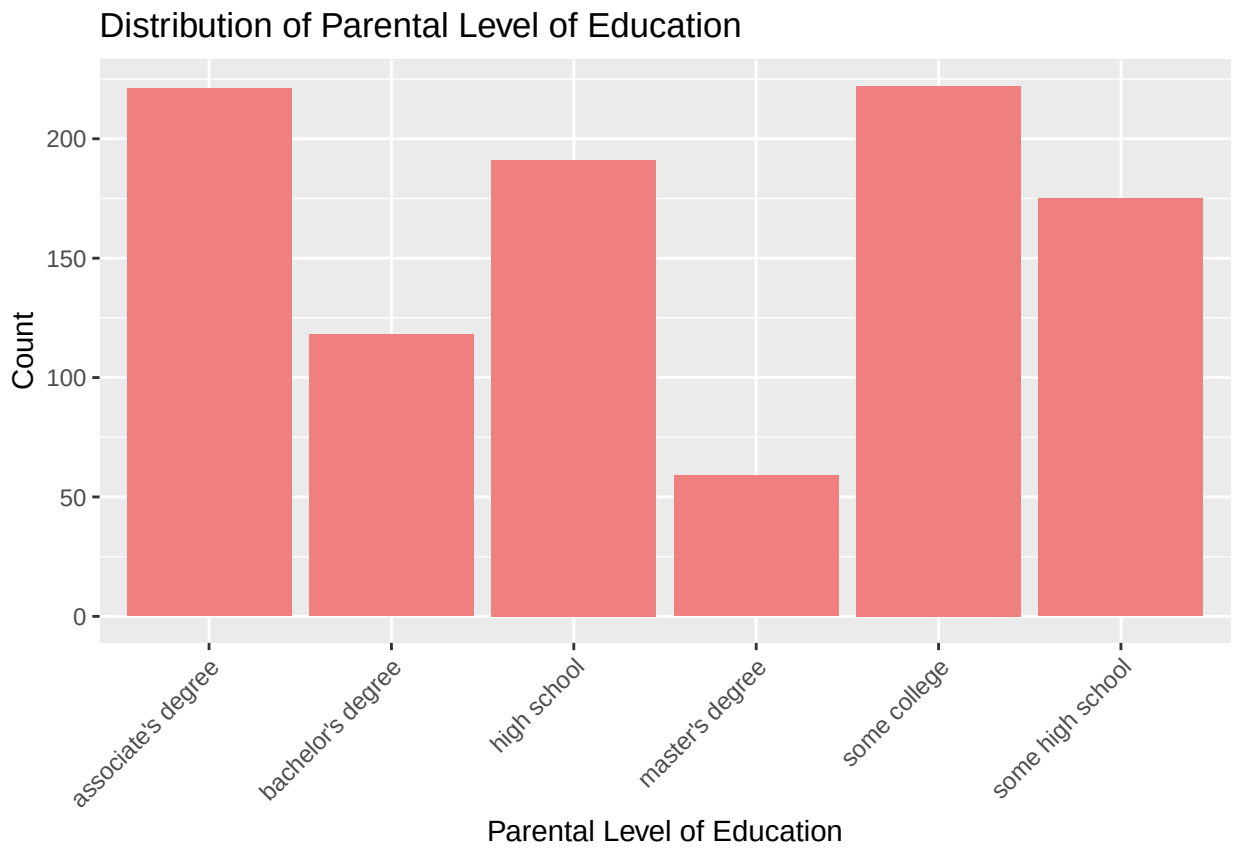
```
# Bar plot for gender
ggplot(data, aes(x = gender)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Distribution of Gender",
       x = "Gender",
       y = "Count")
```



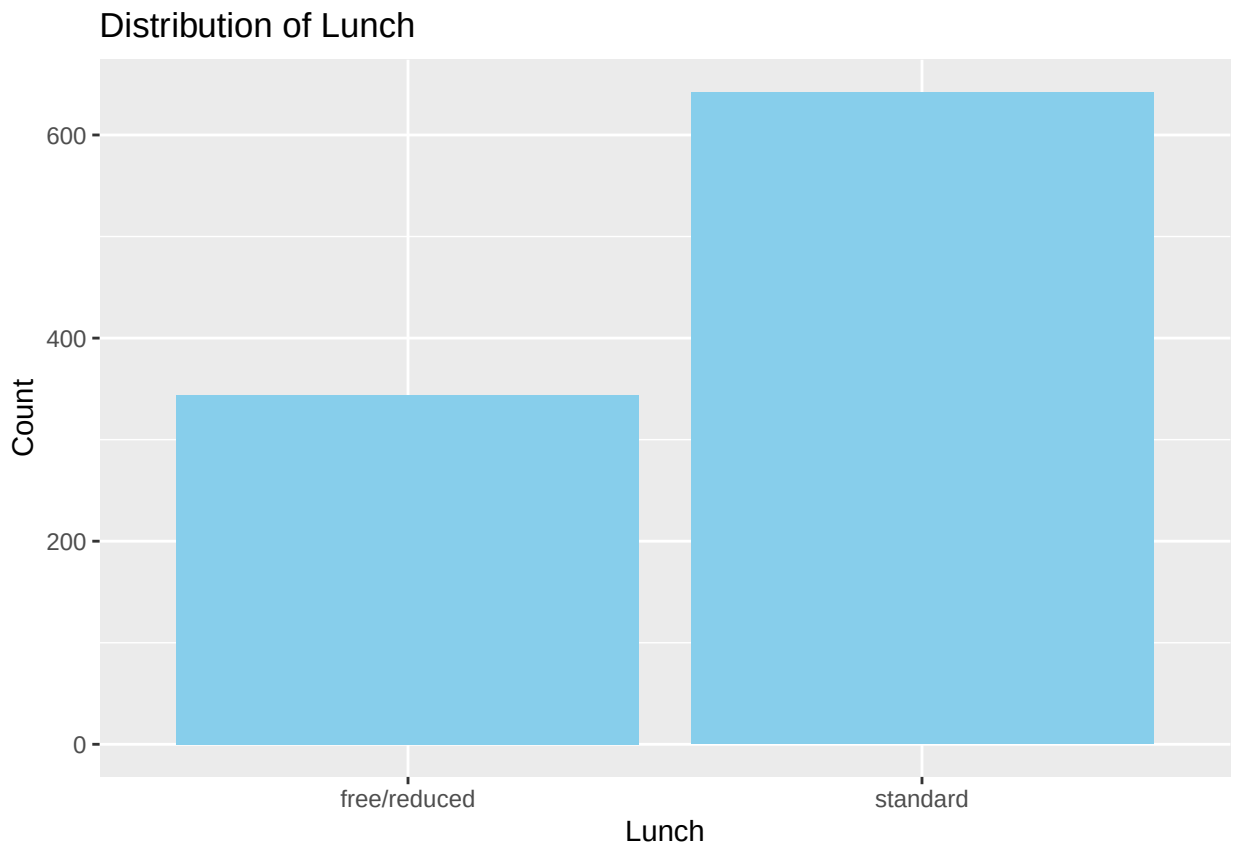

```
# Bar plot for race/ethnicity
ggplot(data, aes(x = race_ethnicity)) +
  geom_bar(fill = "lightgreen") +
  labs(title = "Distribution of Race/Ethnicity",
       x = "Race/Ethnicity",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



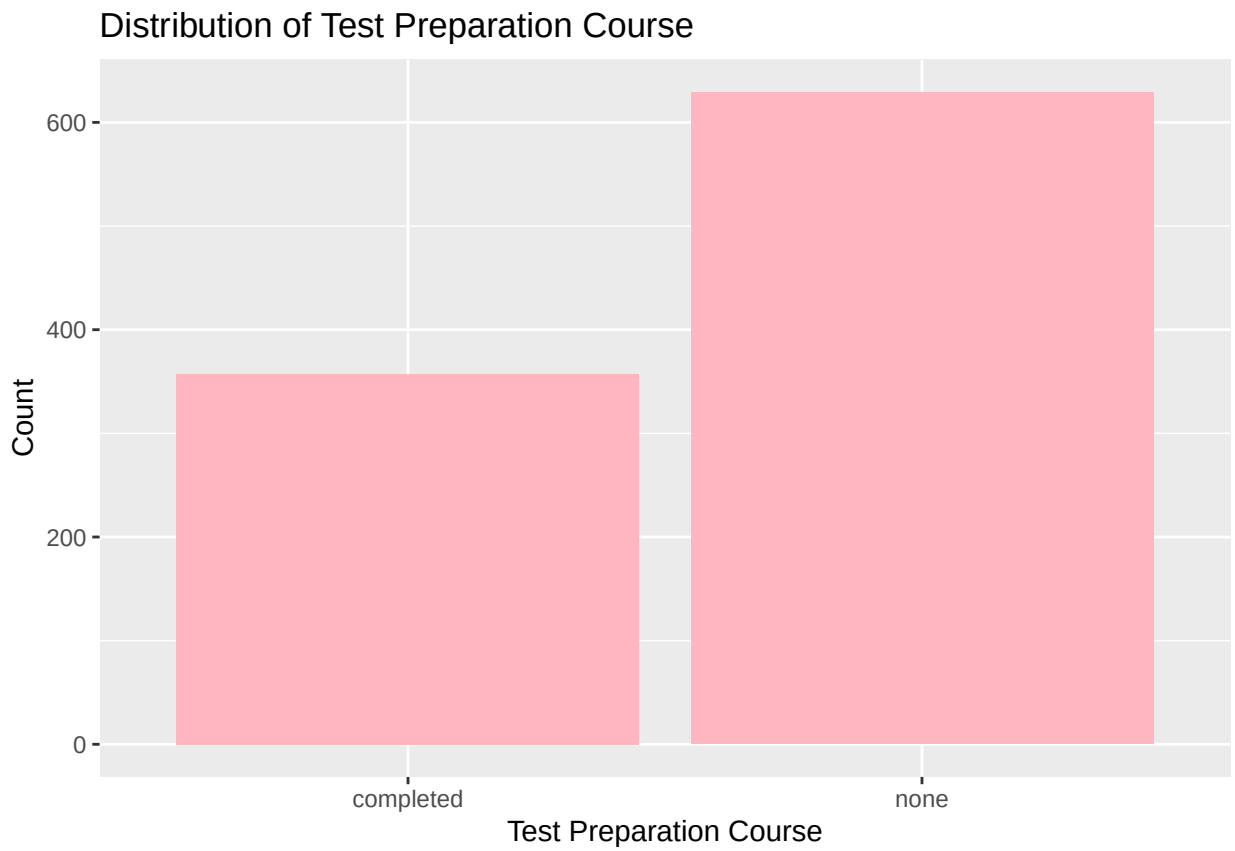

```
# Bar plot for parental level of education
ggplot(data, aes(x = parental_level_of_education)) +
  geom_bar(fill = "lightcoral") +
  labs(title = "Distribution of Parental Level of Education",
       x = "Parental Level of Education",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Bar plot for lunch
ggplot(data, aes(x = lunch)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Distribution of Lunch",
       x = "Lunch",
       y = "Count")
```

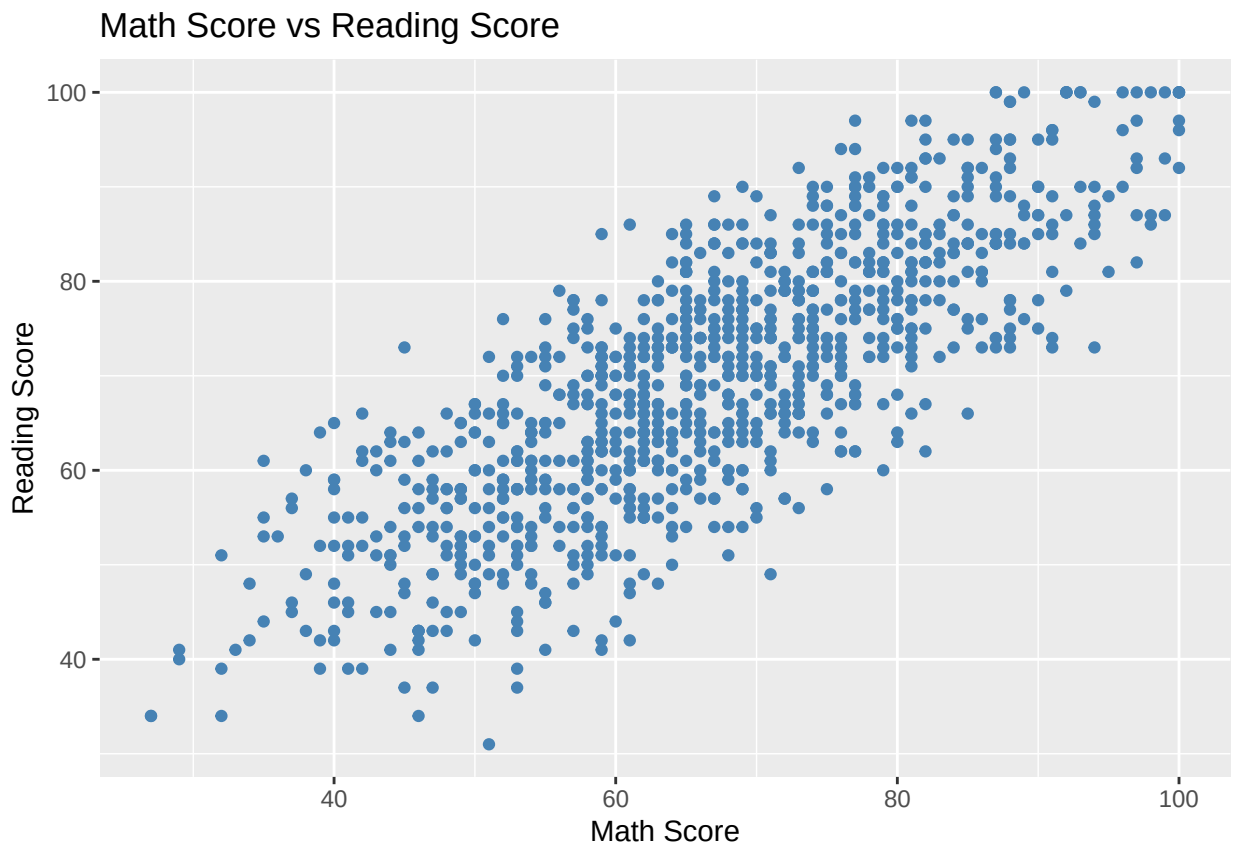



```
# Bar plot for test preparation course
ggplot(data, aes(x = test_preparation_course)) +
  geom_bar(fill = "lightpink") +
  labs(title = "Distribution of Test Preparation Course",
       x = "Test Preparation Course",
       y = "Count")
```

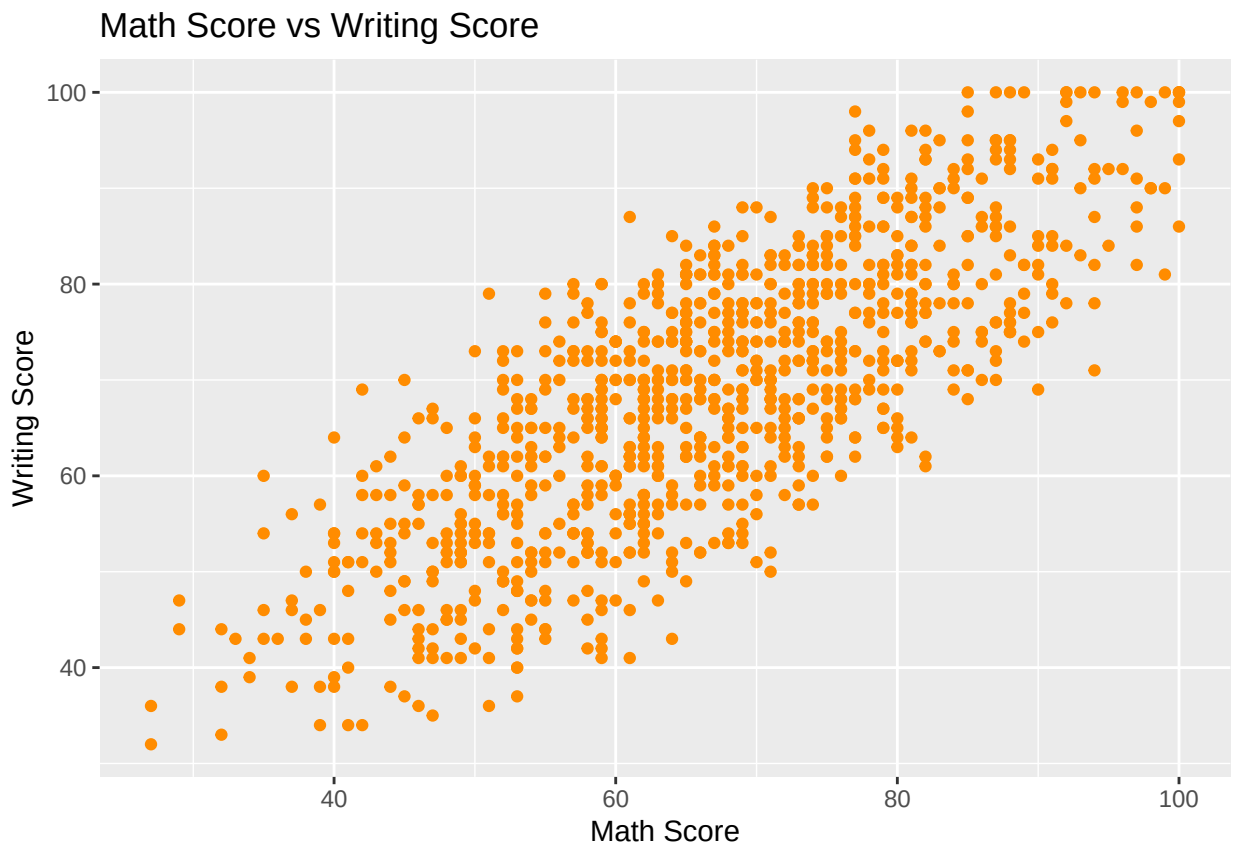


Bivariate Analysis ### Scatter plots for numerical variables

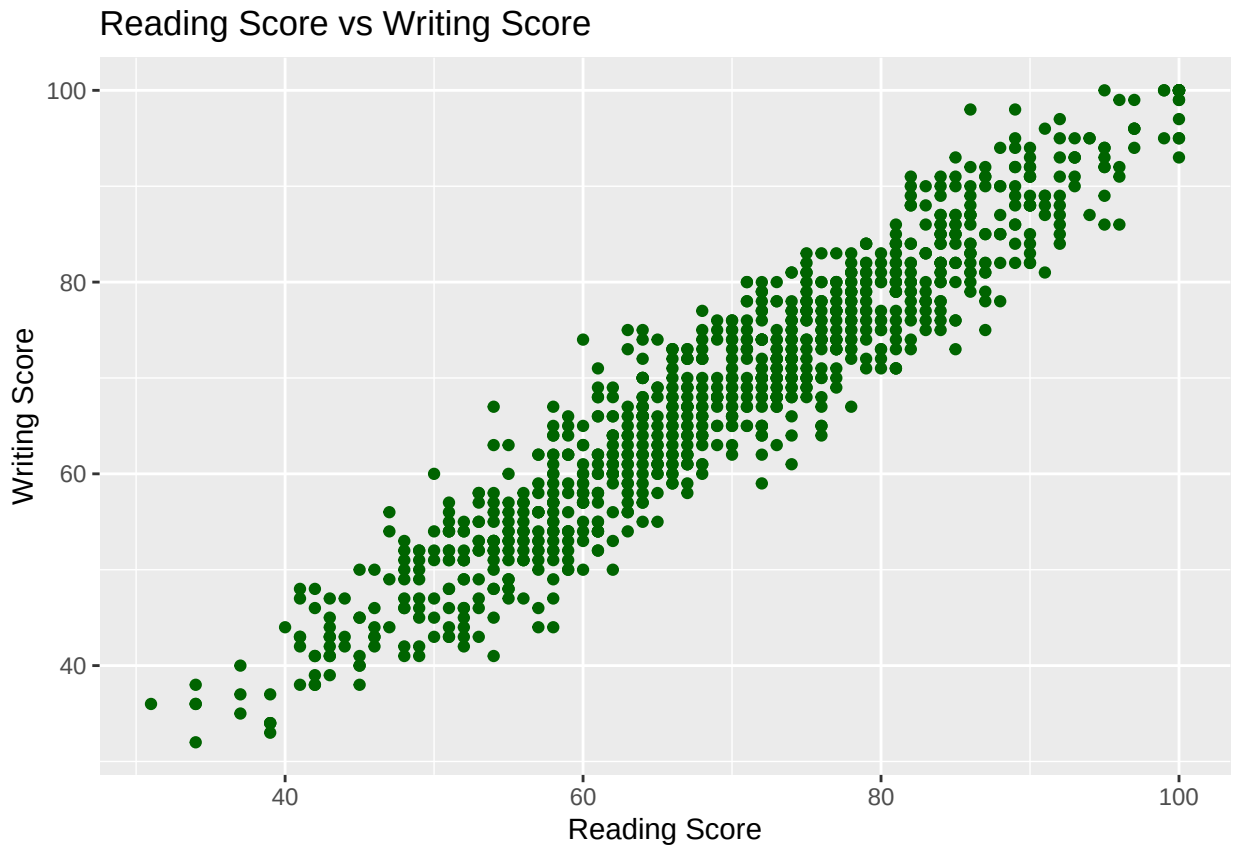
```
# Math score vs reading score
ggplot(data, aes(x = math_score, y = reading_score)) +
  geom_point(color = "steelblue") +
  labs(title = "Math Score vs Reading Score",
       x = "Math Score",
       y = "Reading Score")
```




```
# Math score vs writing score
ggplot(data, aes(x = math_score, y = writing_score)) +
  geom_point(color = "darkorange") +
  labs(title = "Math Score vs Writing Score",
       x = "Math Score",
       y = "Writing Score")
```

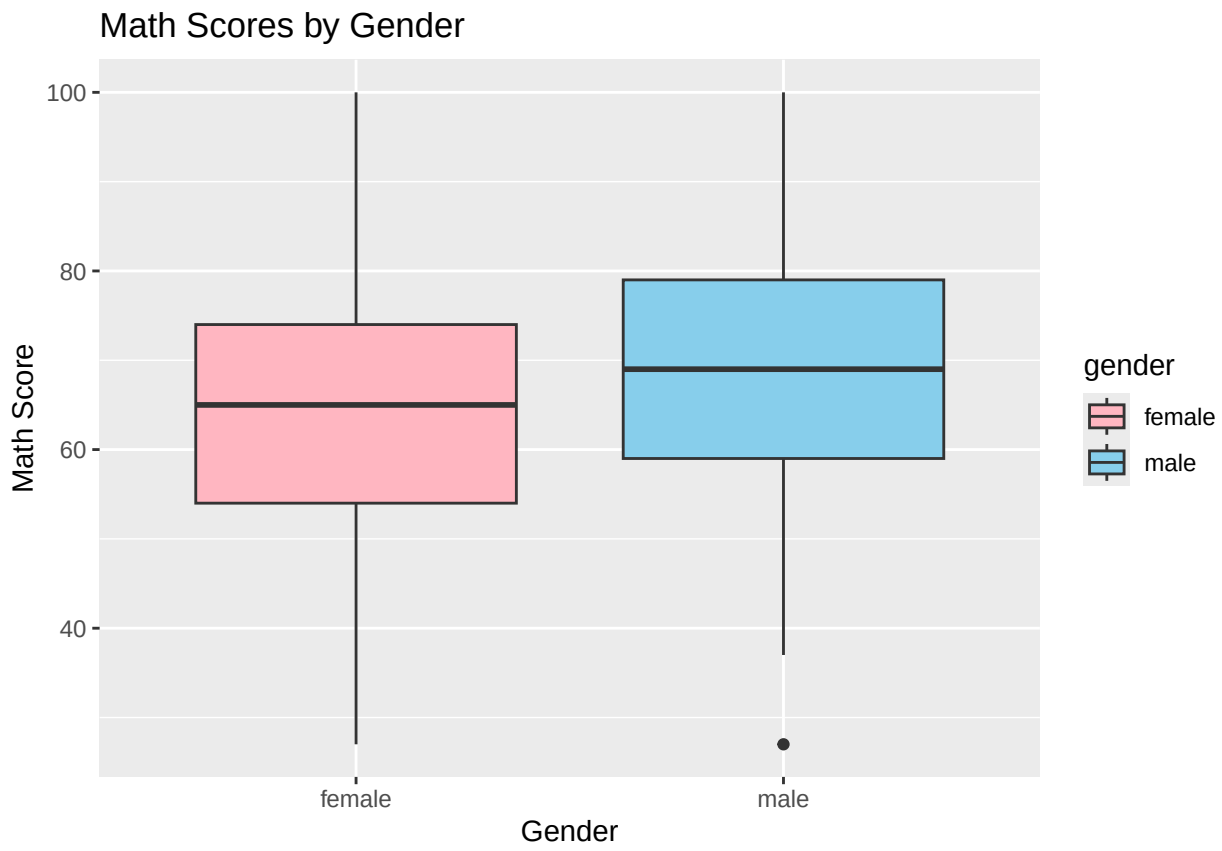



```
# Reading score vs writing score
ggplot(data, aes(x = reading_score, y = writing_score)) +
  geom_point(color = "darkgreen") +
  labs(title = "Reading Score vs Writing Score",
       x = "Reading Score",
       y = "Writing Score")
```

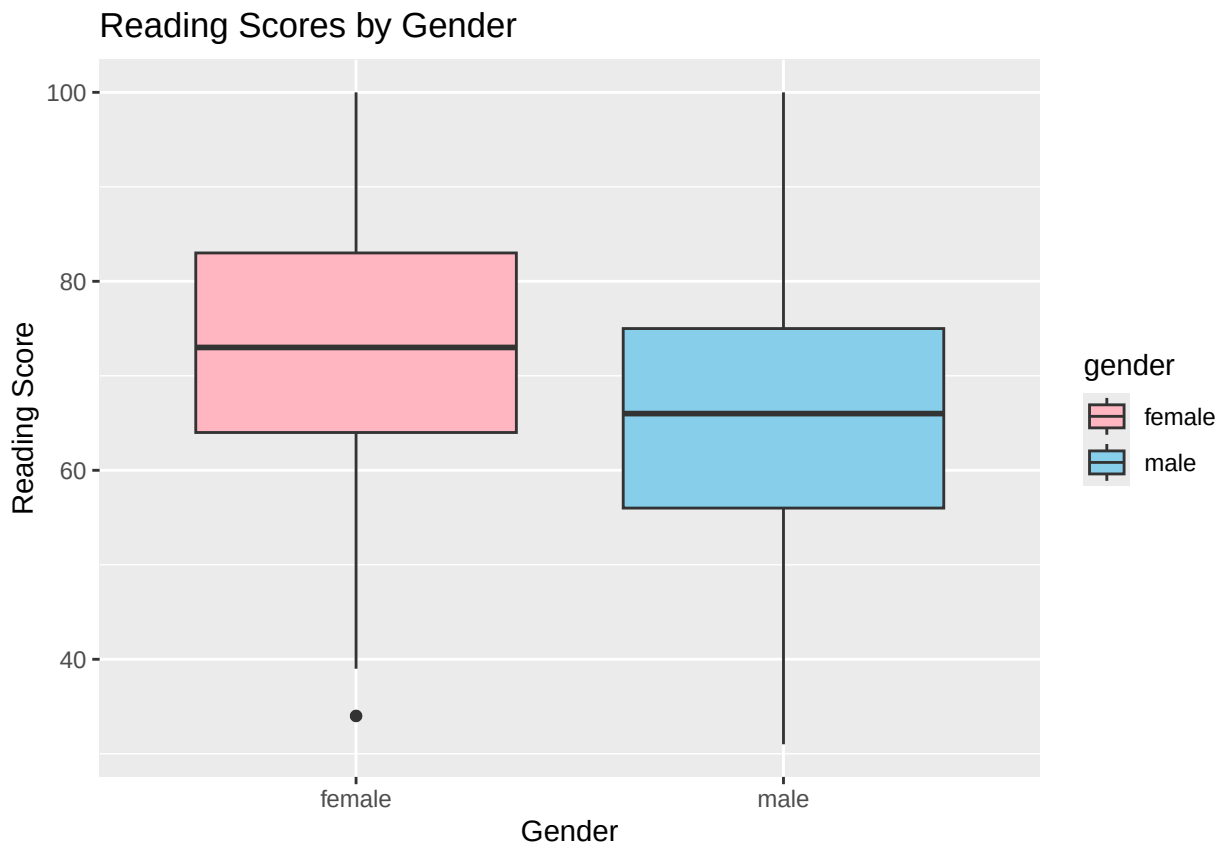


Box plots for comparing numerical variable across different categories

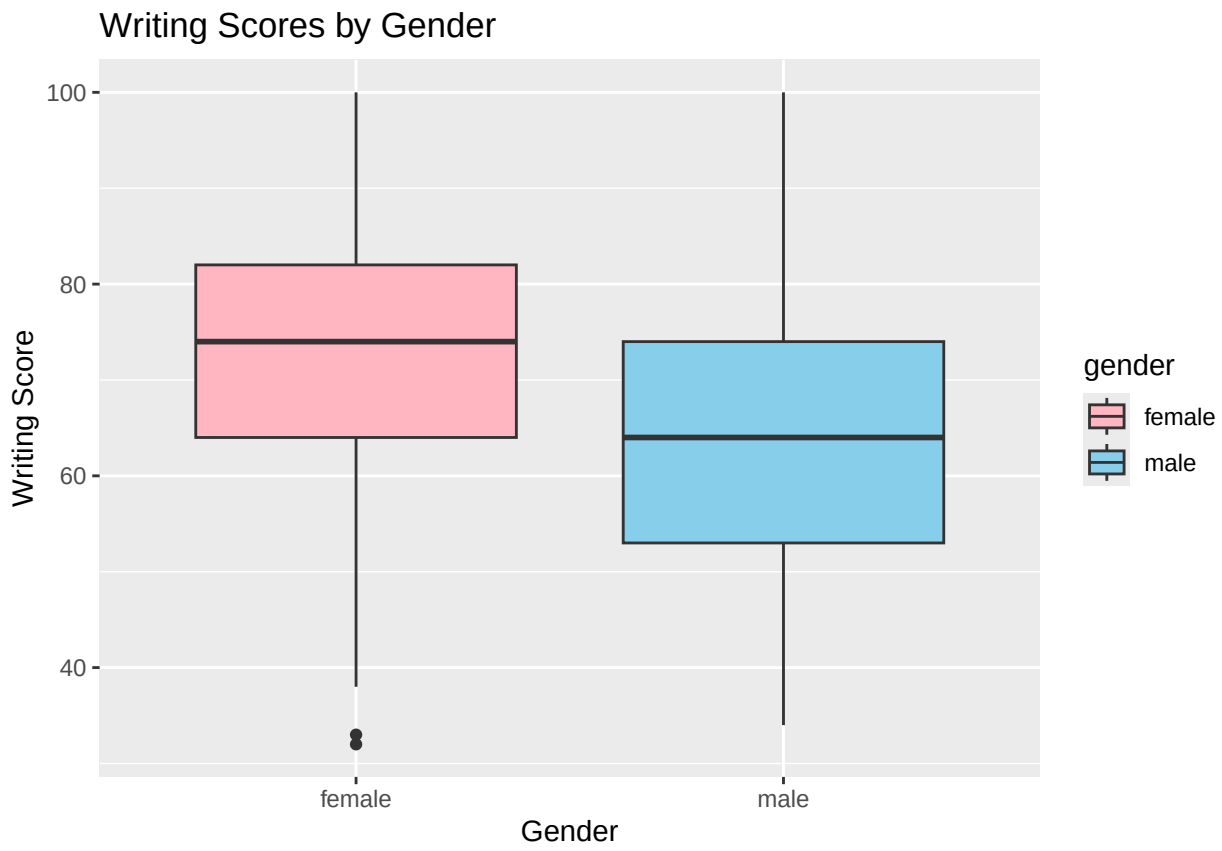
```
# Box plot of math scores by gender
ggplot(data, aes(x = gender, y = math_score, fill = gender)) +
  geom_boxplot() +
  labs(title = "Math Scores by Gender",
       x = "Gender",
       y = "Math Score") +
  scale_fill_manual(values = c("male" = "skyblue", "female" = "lightpink"))
```



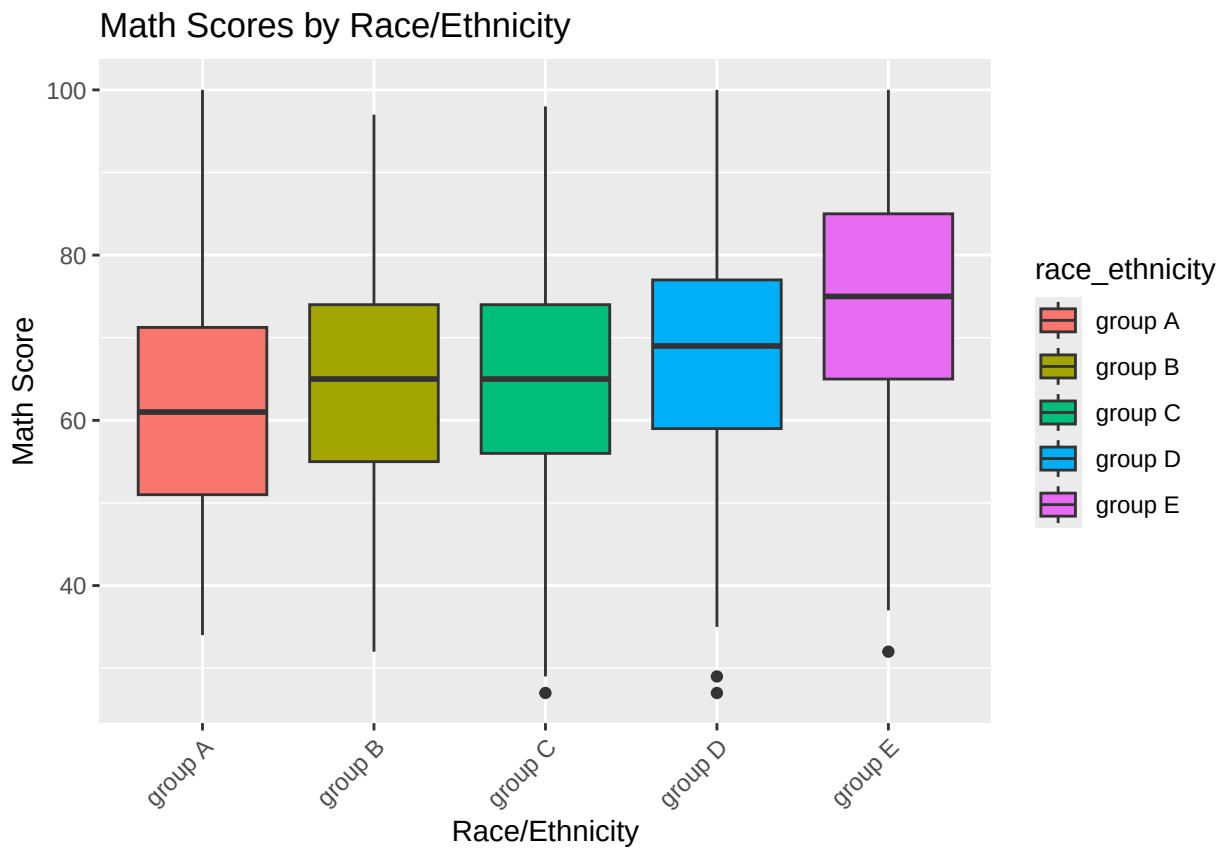

```
# Box plot of reading scores by gender
ggplot(data, aes(x = gender, y = reading_score, fill = gender)) +
  geom_boxplot() +
  labs(title = "Reading Scores by Gender",
       x = "Gender",
       y = "Reading Score") +
  scale_fill_manual(values = c("male" = "skyblue", "female" = "lightpink"))
```



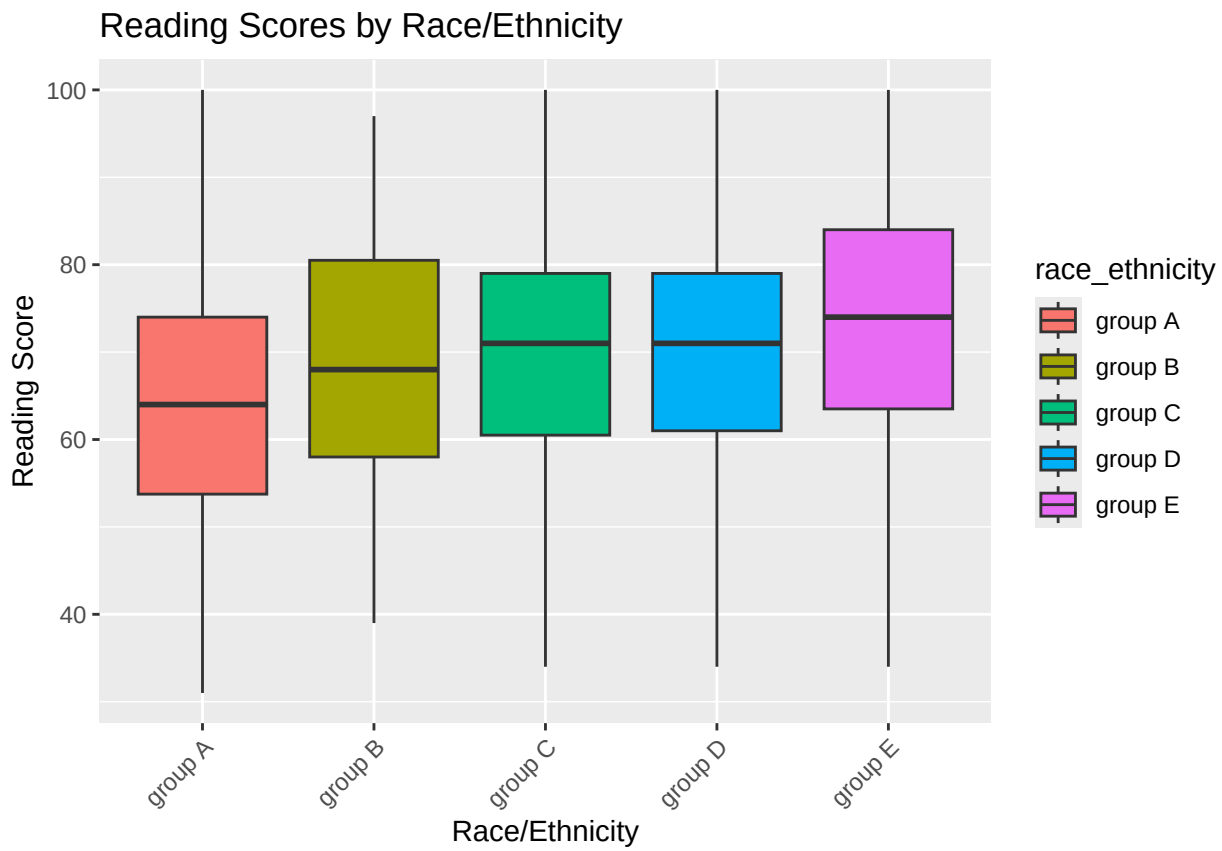

```
# Box plot of writing scores by gender
ggplot(data, aes(x = gender, y = writing_score, fill = gender)) +
  geom_boxplot() +
  labs(title = "Writing Scores by Gender",
       x = "Gender",
       y = "Writing Score") +
  scale_fill_manual(values = c("male" = "skyblue", "female" = "lightpink"))
```



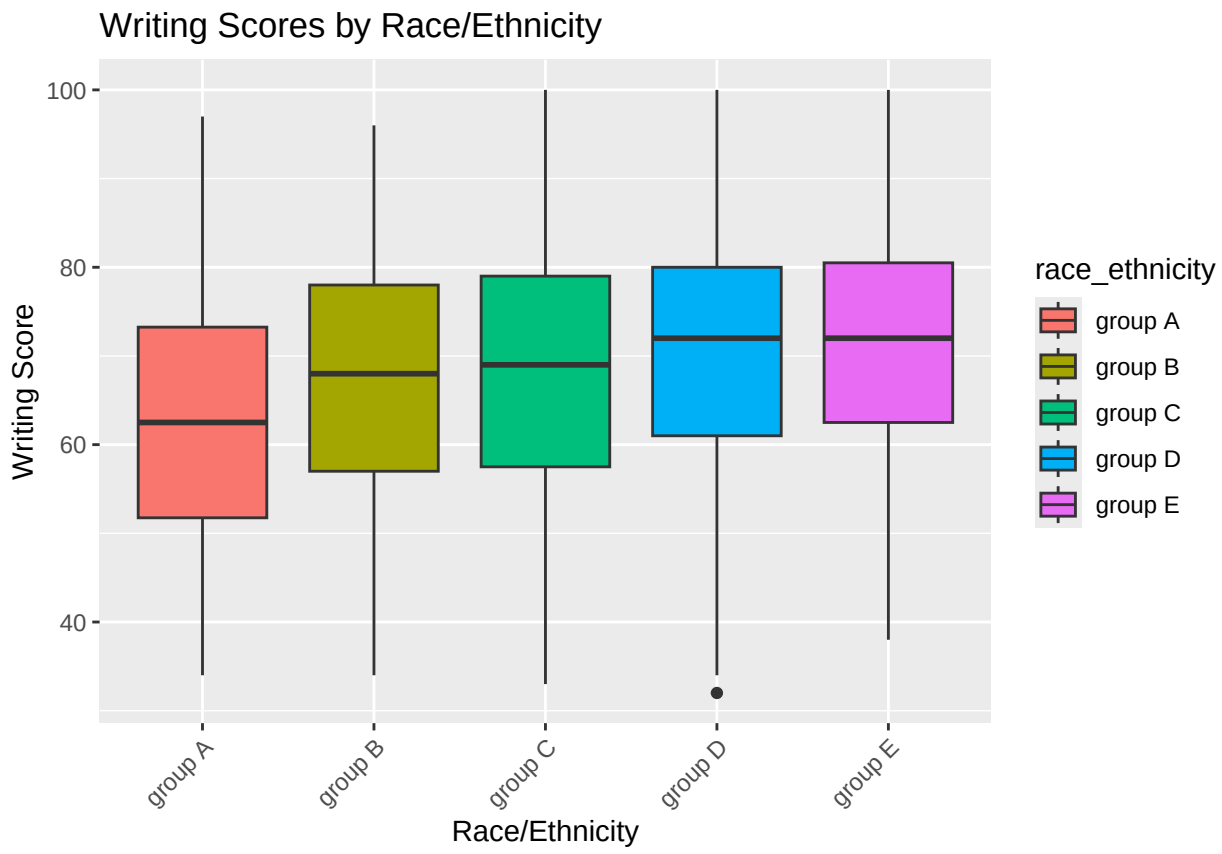

```r
# Box plot of math scores by race/ethnicity
ggplot(data, aes(x = race_ethnicity, y = math_score, fill = race_ethnicity)) +
  geom_boxplot() +
  labs(title = "Math Scores by Race/Ethnicity",
       x = "Race/Ethnicity",
       y = "Math Score") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# Box plot of reading scores by race/ethnicity
ggplot(data, aes(x = race_ethnicity, y = reading_score, fill = race_ethnicity)) +
  geom_boxplot() +
  labs(title = "Reading Scores by Race/Ethnicity",
       x = "Race/Ethnicity",
       y = "Reading Score") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Box plot of writing scores by race/ethnicity
ggplot(data, aes(x = race_ethnicity, y = writing_score, fill = race_ethnicity)) +
  geom_boxplot() +
  labs(title = "Writing Scores by Race/Ethnicity",
       x = "Race/Ethnicity",
       y = "Writing Score") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

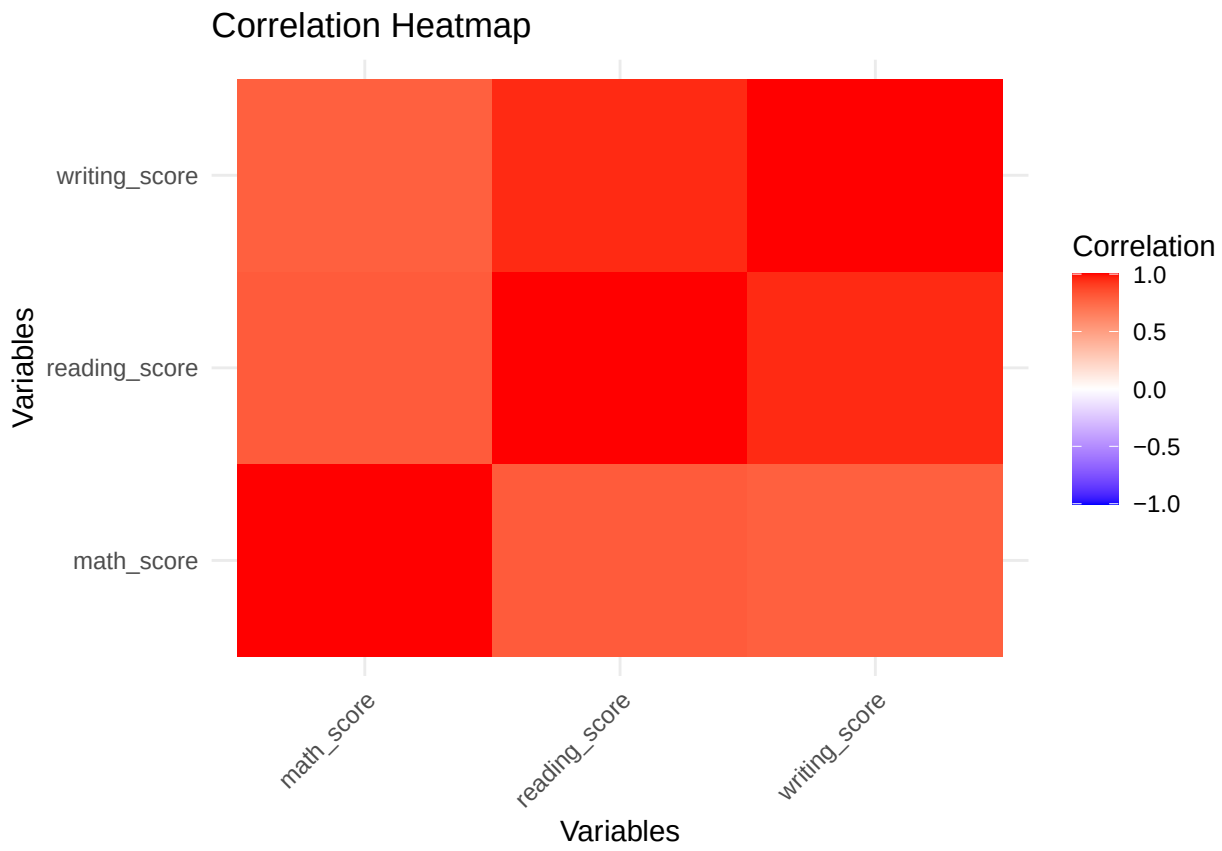



```
# Calculate correlation matrix
correlation_matrix <- cor(data[c("math_score", "reading_score", "writing_score")])

# Print correlation matrix
print(correlation_matrix)
```

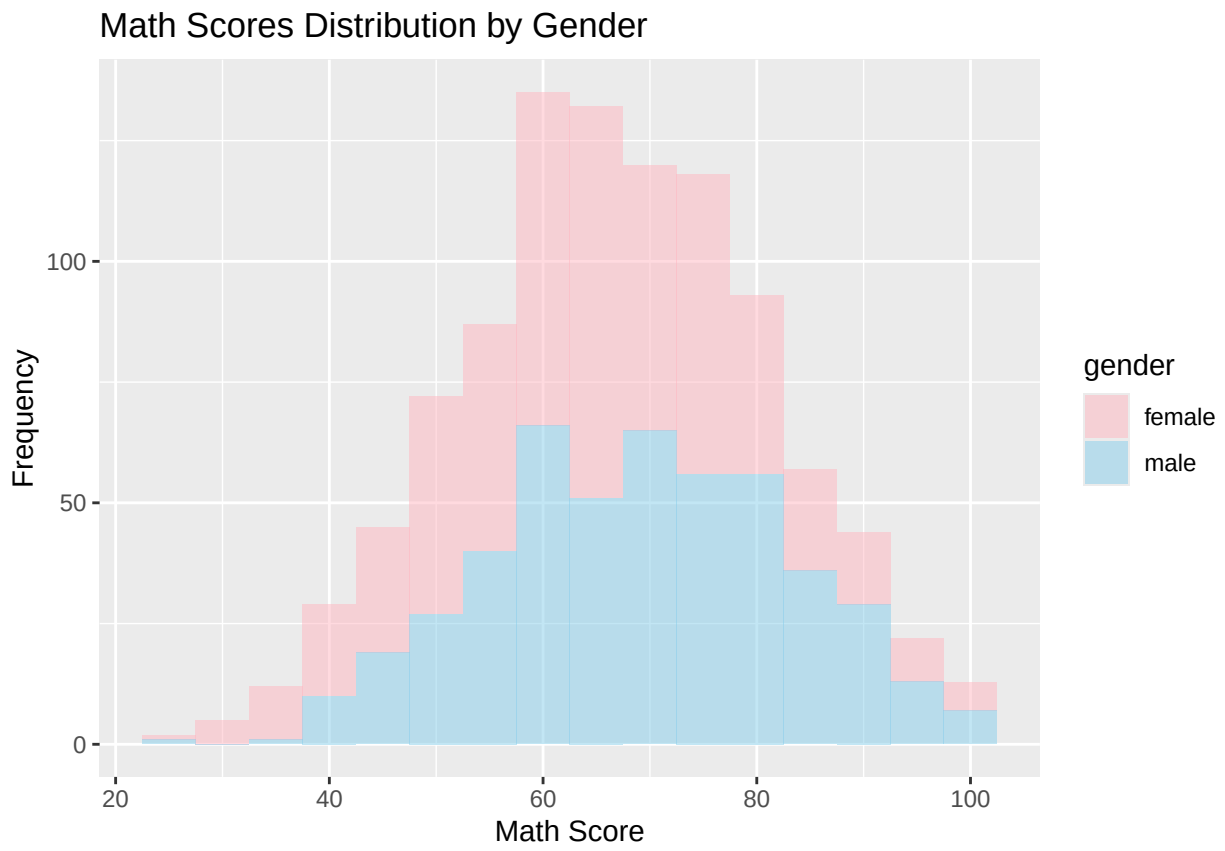
```
##               math_score reading_score writing_score
## math_score     1.0000000     0.7988810     0.7806676
## reading_score  0.7988810     1.0000000     0.9498439
## writing_score  0.7806676     0.9498439     1.0000000
```

```
# Visualize correlation matrix using a heatmap
ggplot(data = melt(correlation_matrix), aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0, limits = c(-1,1)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Correlation Heatmap",
       x = "Variables",
       y = "Variables",
       fill = "Correlation")
```

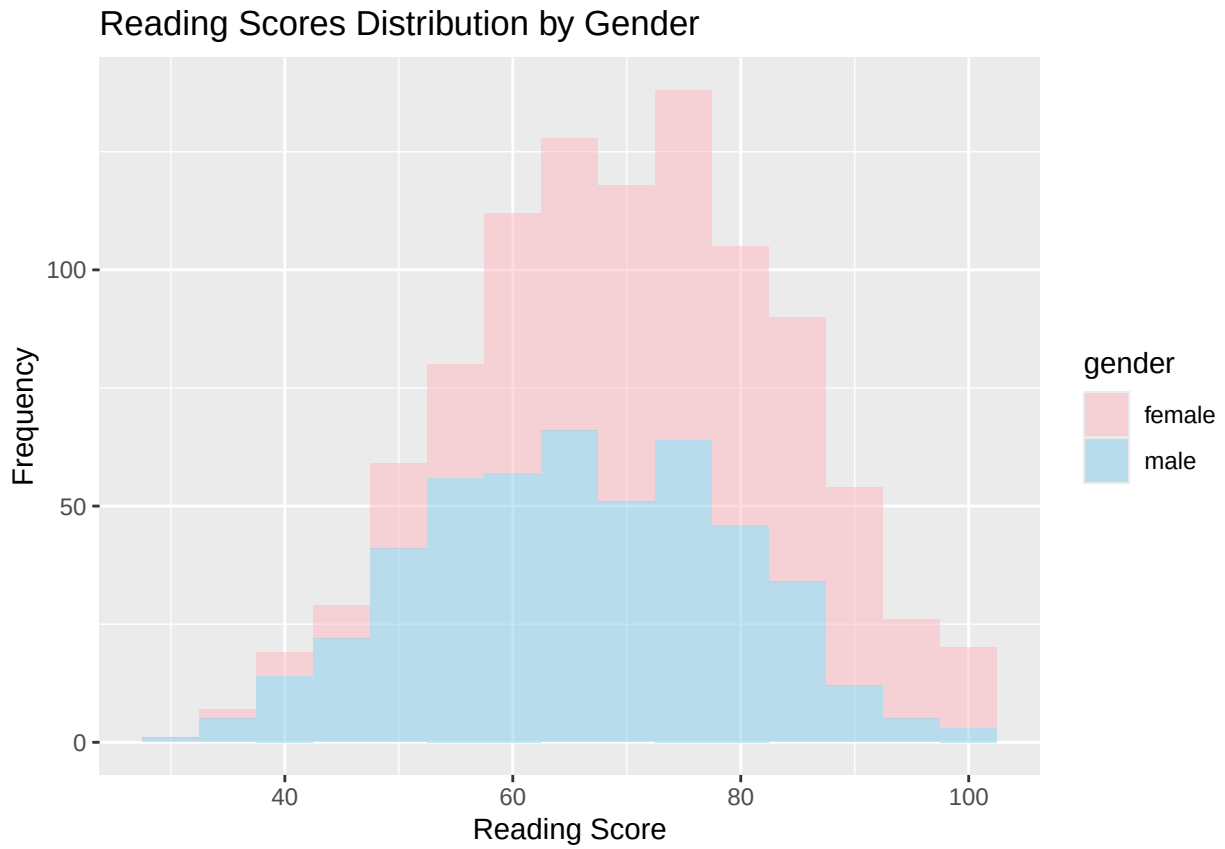


Compare distributions across different groups

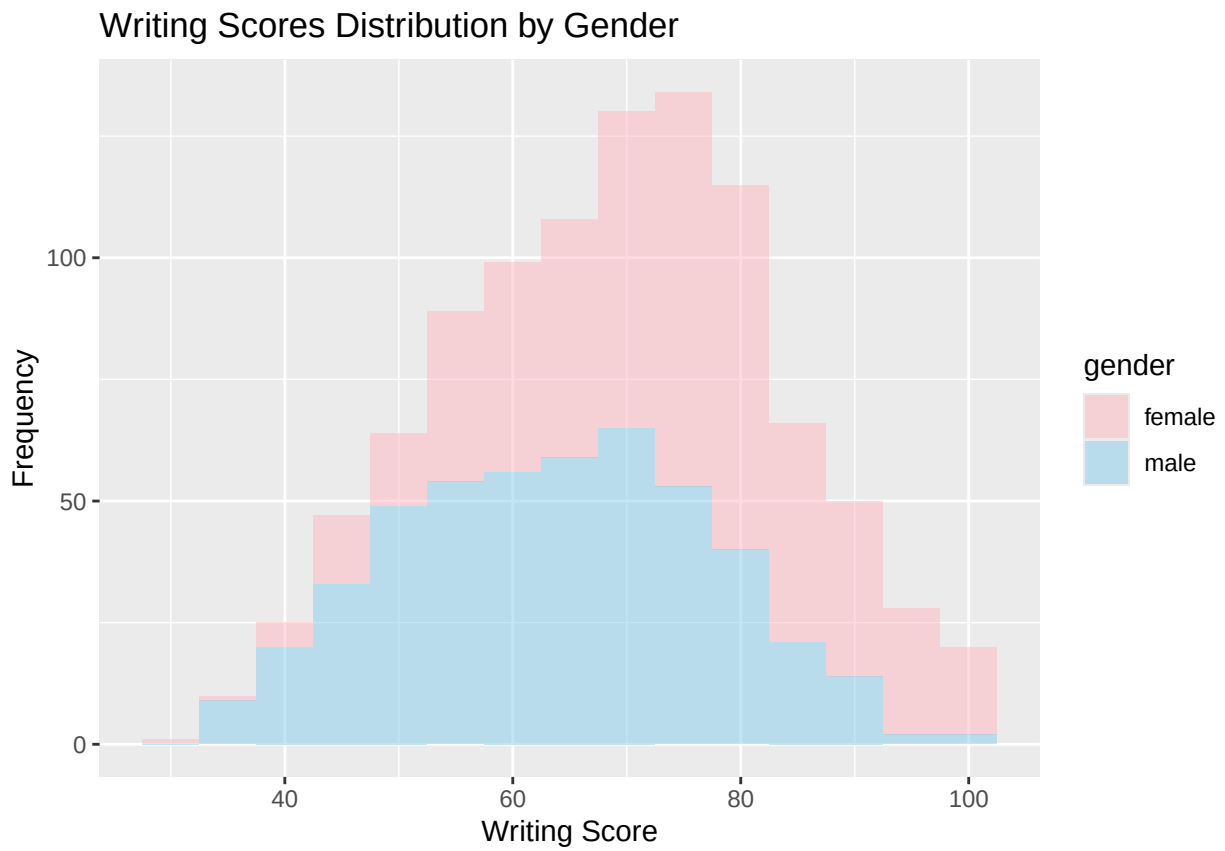
```
# Histogram of math scores by gender
ggplot(data, aes(x = math_score, fill = gender)) +
  geom_histogram(binwidth = 5, alpha = 0.5) +
  labs(title = "Math Scores Distribution by Gender",
       x = "Math Score",
       y = "Frequency") +
  scale_fill_manual(values = c("male" = "skyblue", "female" = "lightpink"))
```



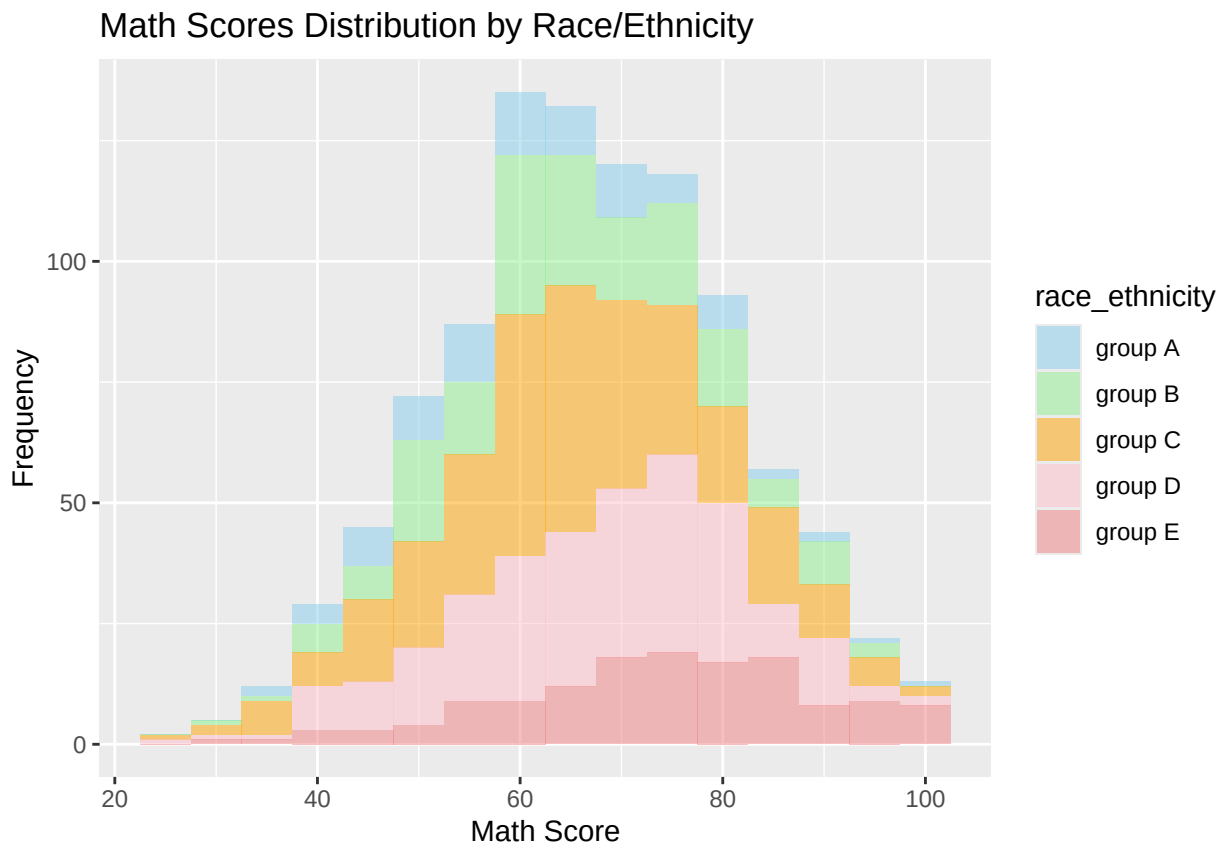

```r
# Histogram of reading scores by gender
ggplot(data, aes(x = reading_score, fill = gender)) +
  geom_histogram(binwidth = 5, alpha = 0.5) +
  labs(title = "Reading Scores Distribution by Gender",
       x = "Reading Score",
       y = "Frequency") +
  scale_fill_manual(values = c("male" = "skyblue", "female" = "lightpink"))
```



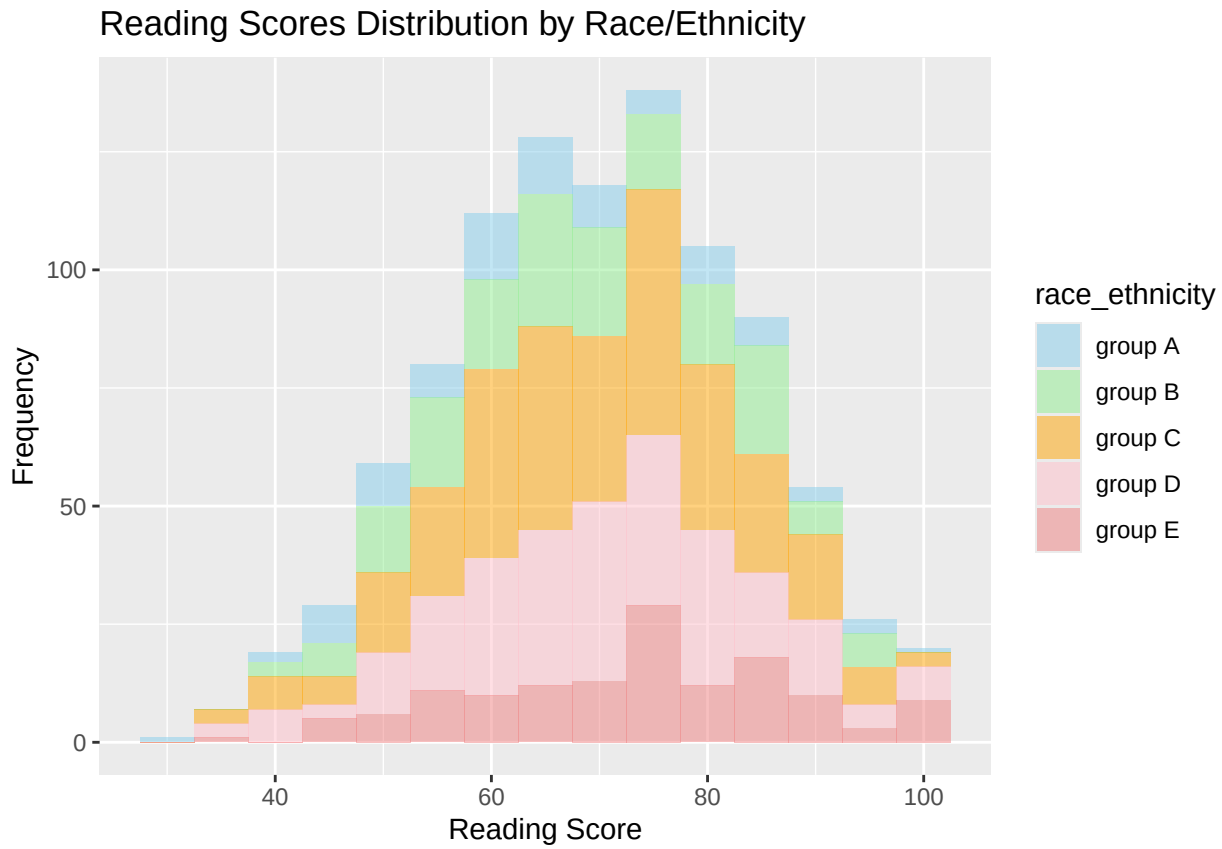

```r
# Histogram of writing scores by gender
ggplot(data, aes(x = writing_score, fill = gender)) +
  geom_histogram(binwidth = 5, alpha = 0.5) +
  labs(title = "Writing Scores Distribution by Gender",
       x = "Writing Score",
       y = "Frequency") +
  scale_fill_manual(values = c("male" = "skyblue", "female" = "lightpink"))
```



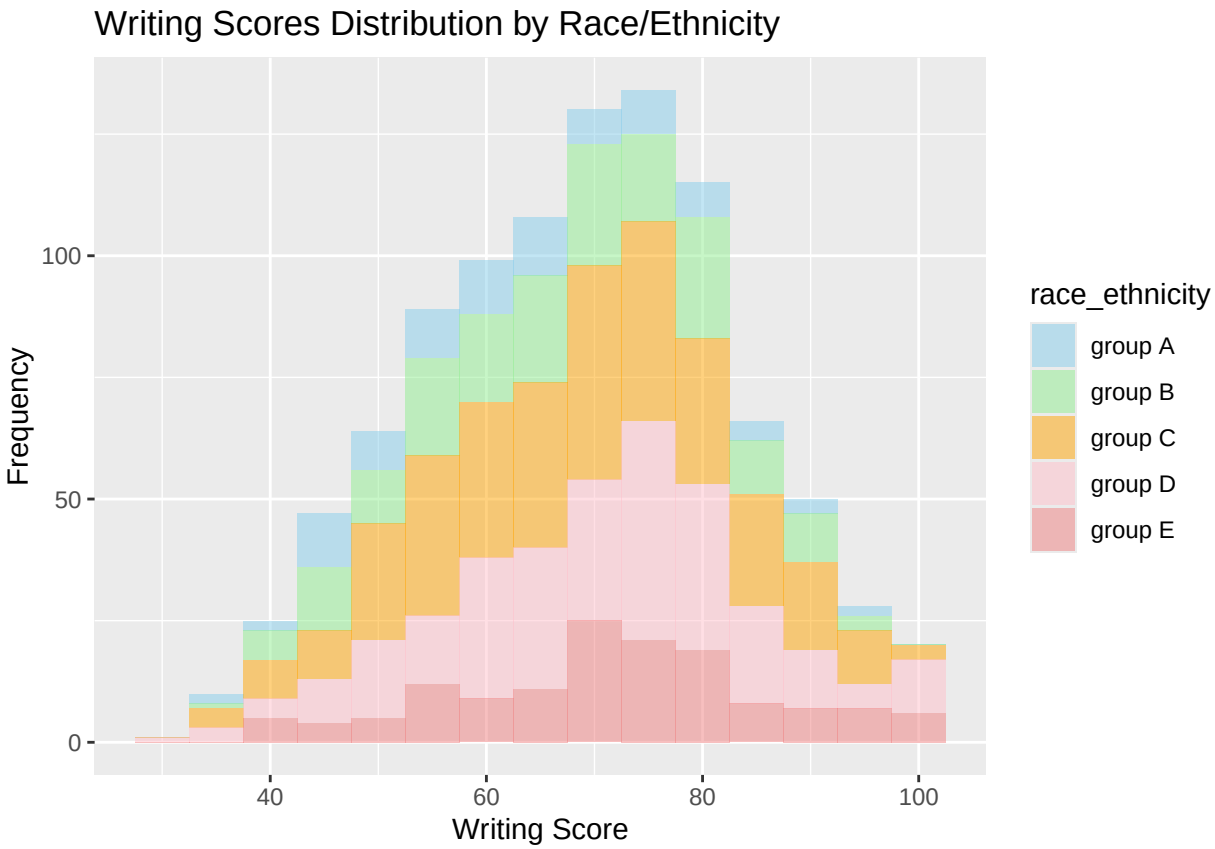

```r
# Histogram of math scores by race/ethnicity
ggplot(data, aes(x = math_score, fill = race_ethnicity)) +
  geom_histogram(binwidth = 5, alpha = 0.5) +
  labs(title = "Math Scores Distribution by Race/Ethnicity",
       x = "Math Score",
       y = "Frequency") +
  scale_fill_manual(values = c("group A" = "skyblue", "group B" = "lightgreen", "group C" = "orange", "g
```




```r
# Histogram of reading scores by race/ethnicity
ggplot(data, aes(x = reading_score, fill = race_ethnicity)) +
  geom_histogram(binwidth = 5, alpha = 0.5) +
  labs(title = "Reading Scores Distribution by Race/Ethnicity",
       x = "Reading Score",
       y = "Frequency") +
  scale_fill_manual(values = c("group A" = "skyblue", "group B" = "lightgreen", "group C" = "orange", "
```




```
# Histogram of writing scores by race/ethnicity
ggplot(data, aes(x = writing_score, fill = race_ethnicity)) +
  geom_histogram(binwidth = 5, alpha = 0.5) +
  labs(title = "Writing Scores Distribution by Race/Ethnicity",
       x = "Writing Score",
       y = "Frequency") +
  scale_fill_manual(values = c("group A" = "skyblue", "group B" = "lightgreen", "group C" = "orange", "
```


Writing Scores Distribution by Race/Ethnicity
Frequency
100
50
0
40
60
80
100
Writing Score
race_ethnicity
group A
group B
group C
group D
group E