Extra_Credit

Tulasi Janjanam

2024-11-12

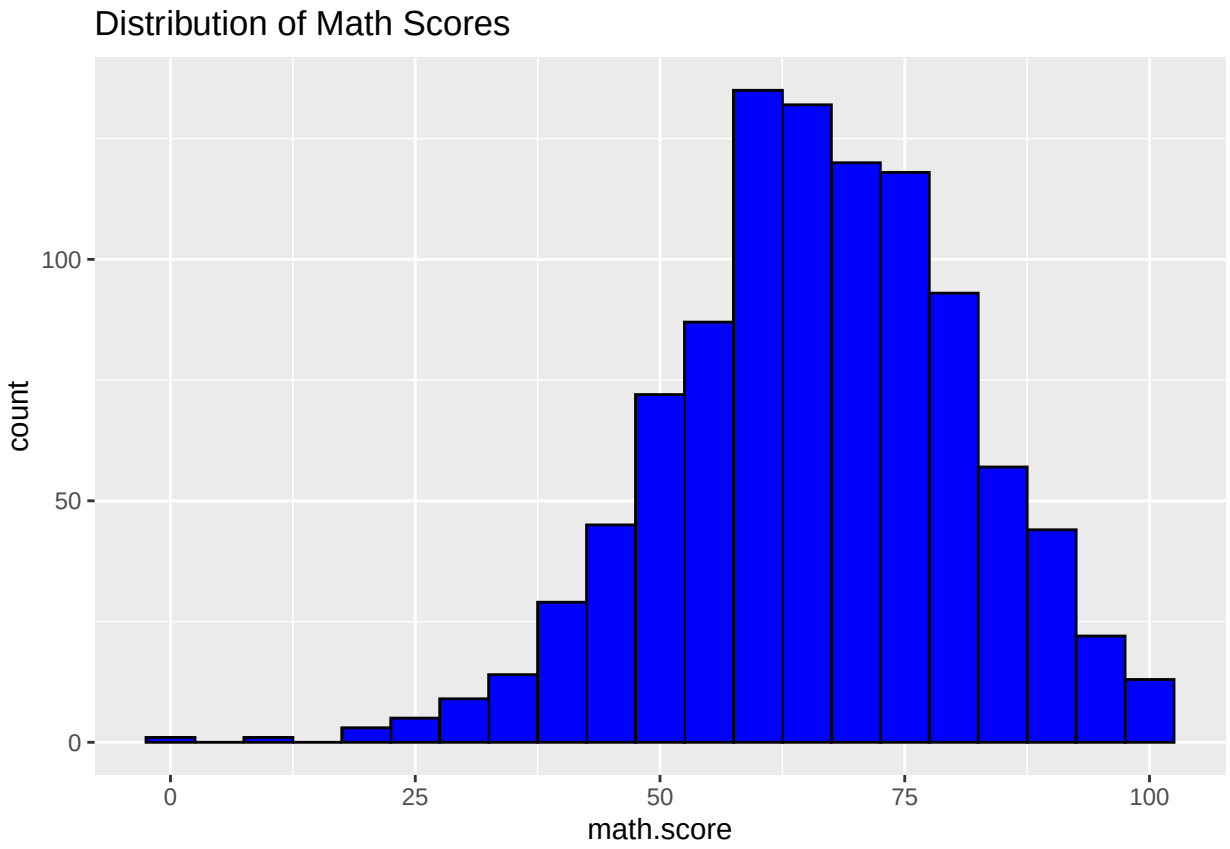
```
data <- read.csv("/Users/tulasijanjanam/Downloads/StudentsPerformance.csv", header = FALSE, skip = 1)
colnames(data) <- c("gender", "race.ethnicity", "parental.level.of.education",
                    "lunch", "test.preparation.course", "math.score",
                    "reading.score", "writing.score")
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
```

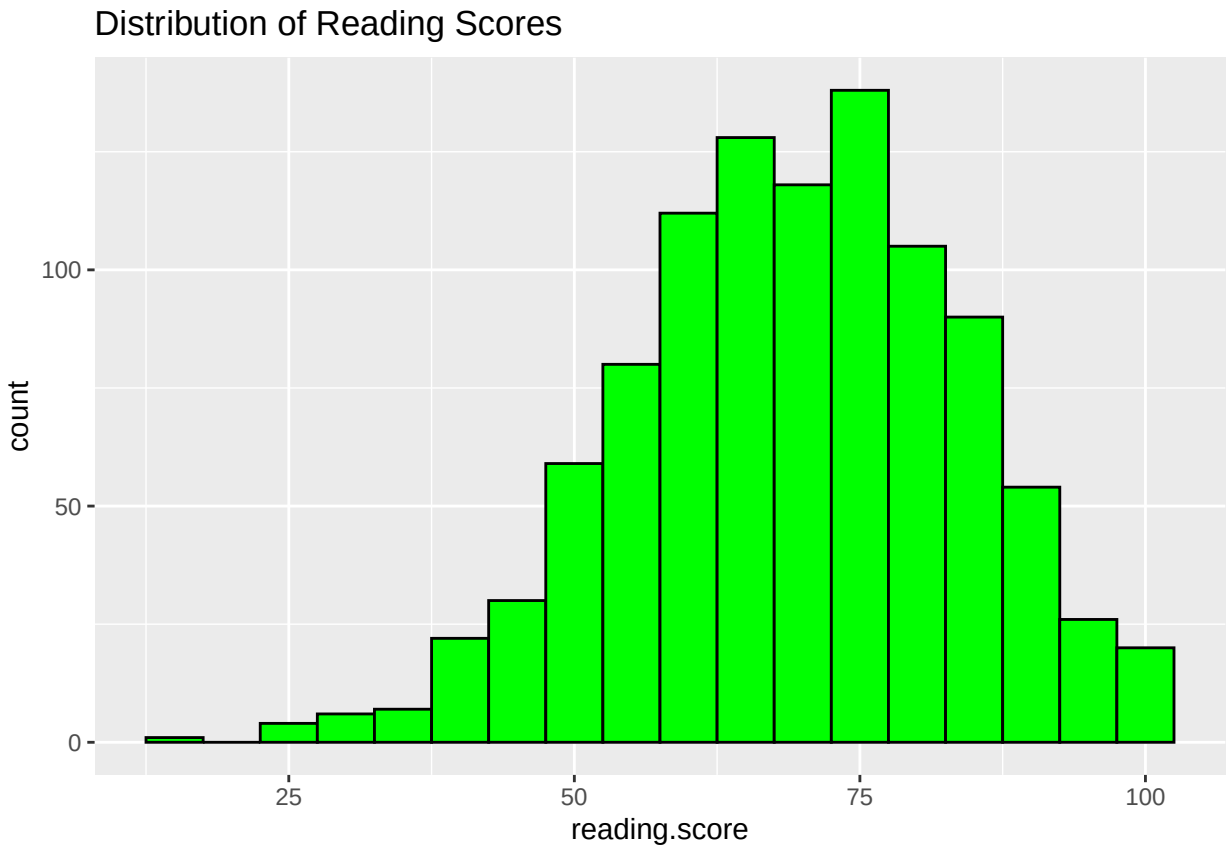
```
head(data)
```

```
##   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
```

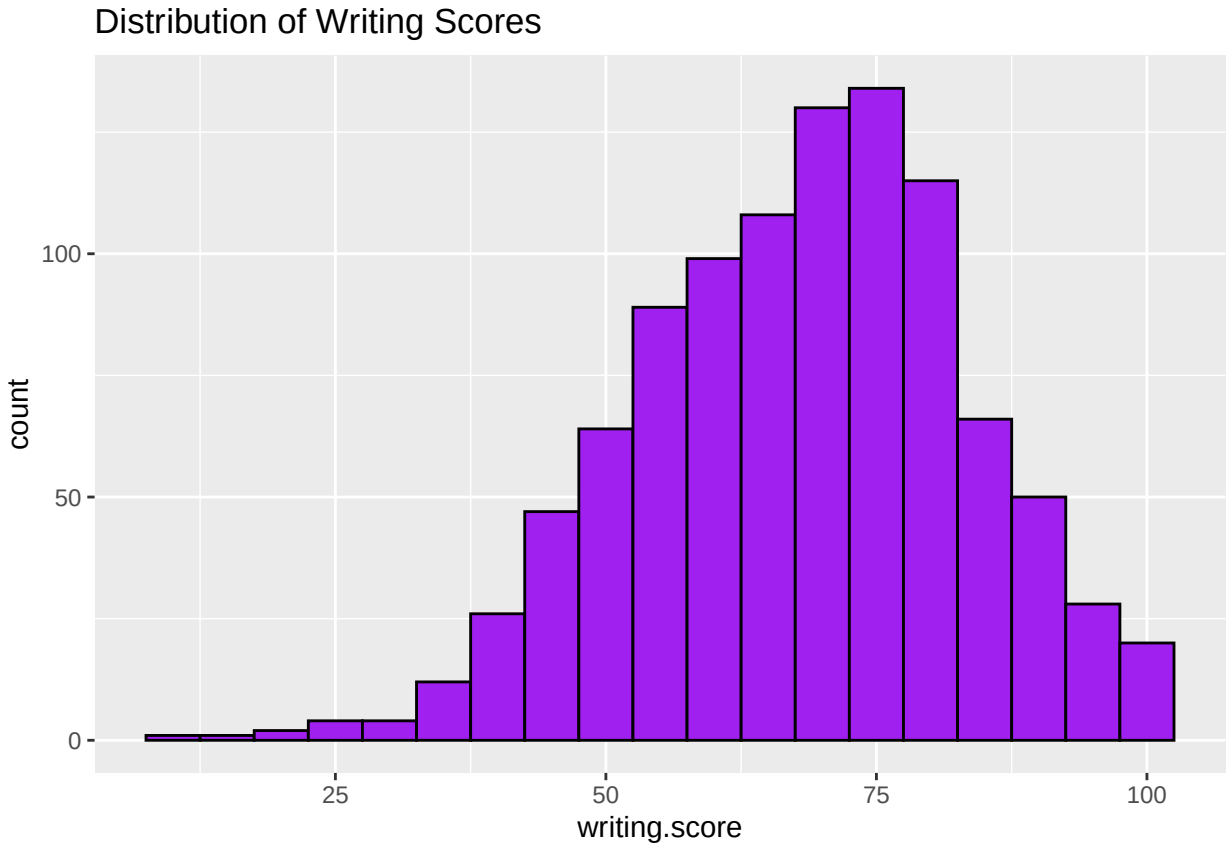
```
library(ggplot2)
ggplot(data, aes(x = math.score)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  ggtitle("Distribution of Math Scores")
```



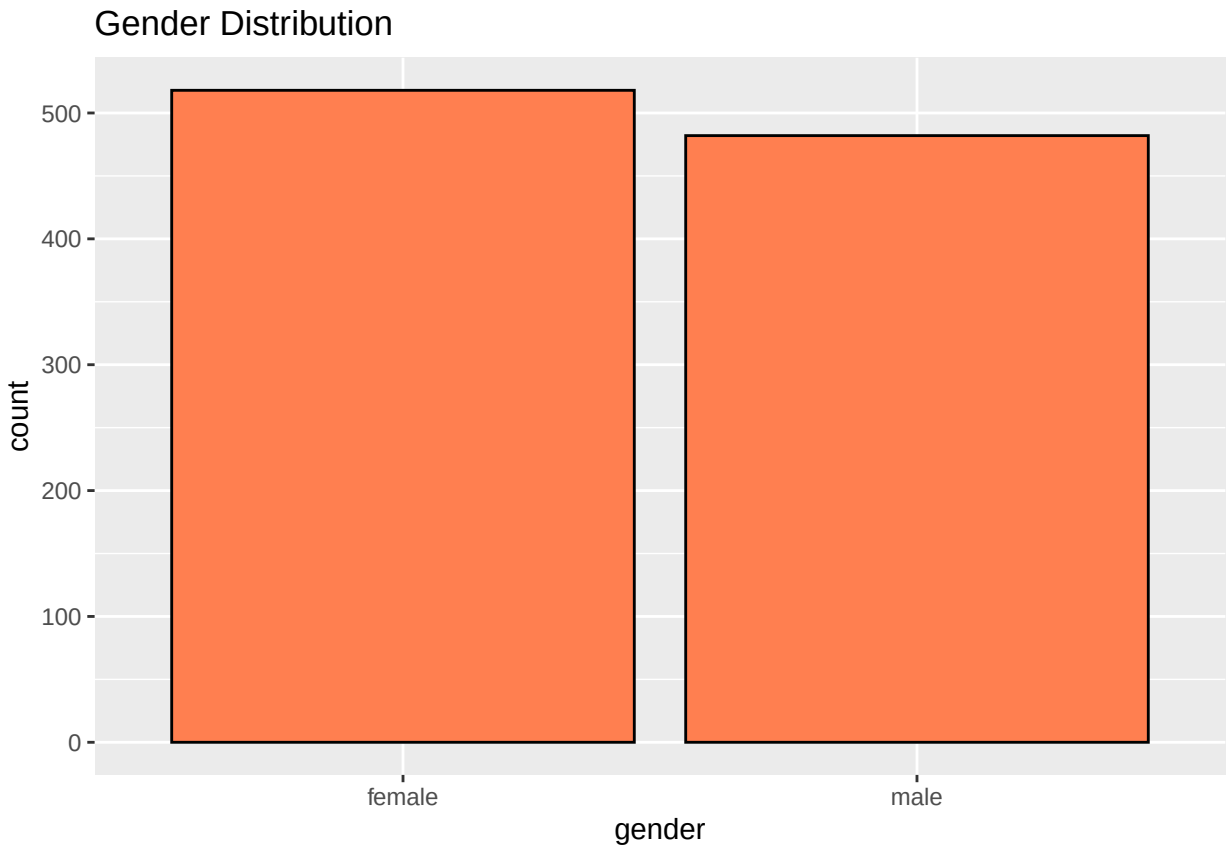

```
ggplot(data, aes(x = reading.score)) +
  geom_histogram(binwidth = 5, fill = "green", color = "black") +
  ggtitle("Distribution of Reading Scores")
```



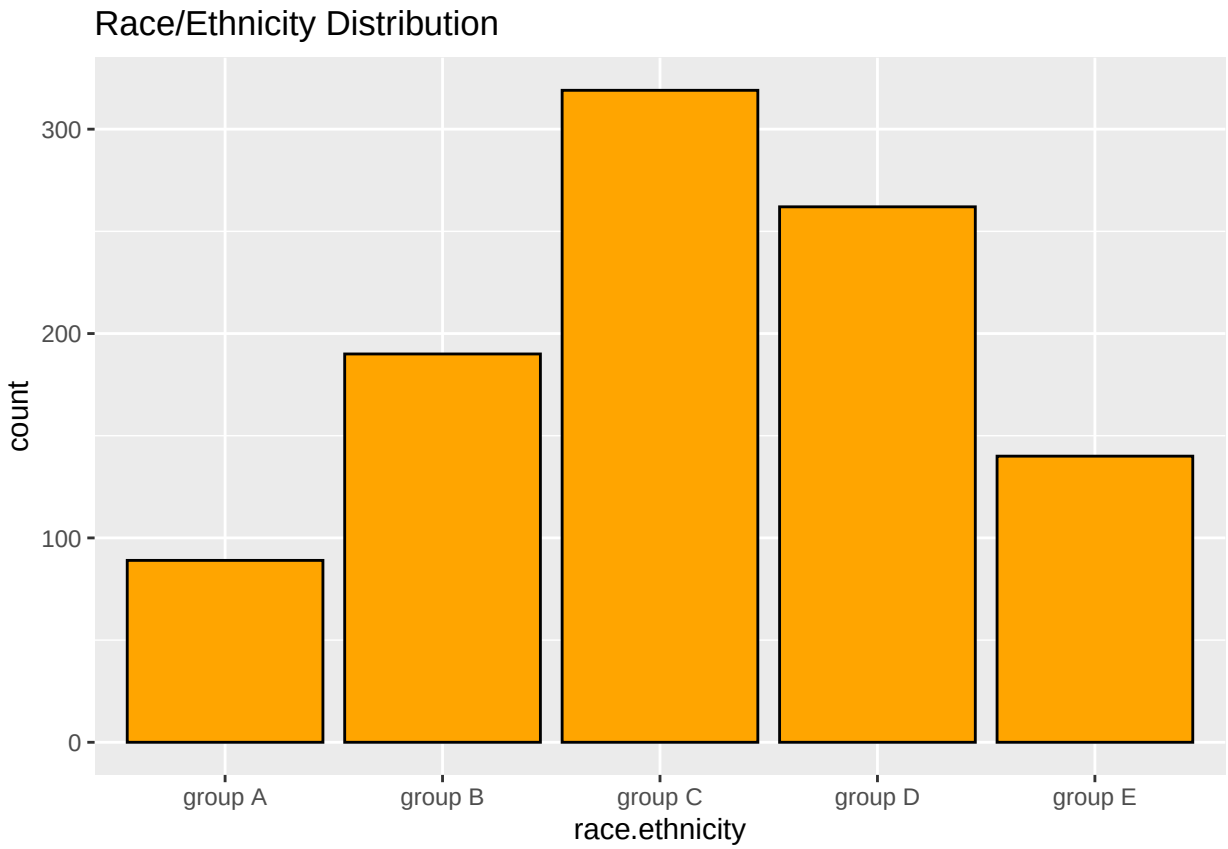

```
ggplot(data, aes(x = writing.score)) +
  geom_histogram(binwidth = 5, fill = "purple", color = "black") +
  ggtitle("Distribution of Writing Scores")
```



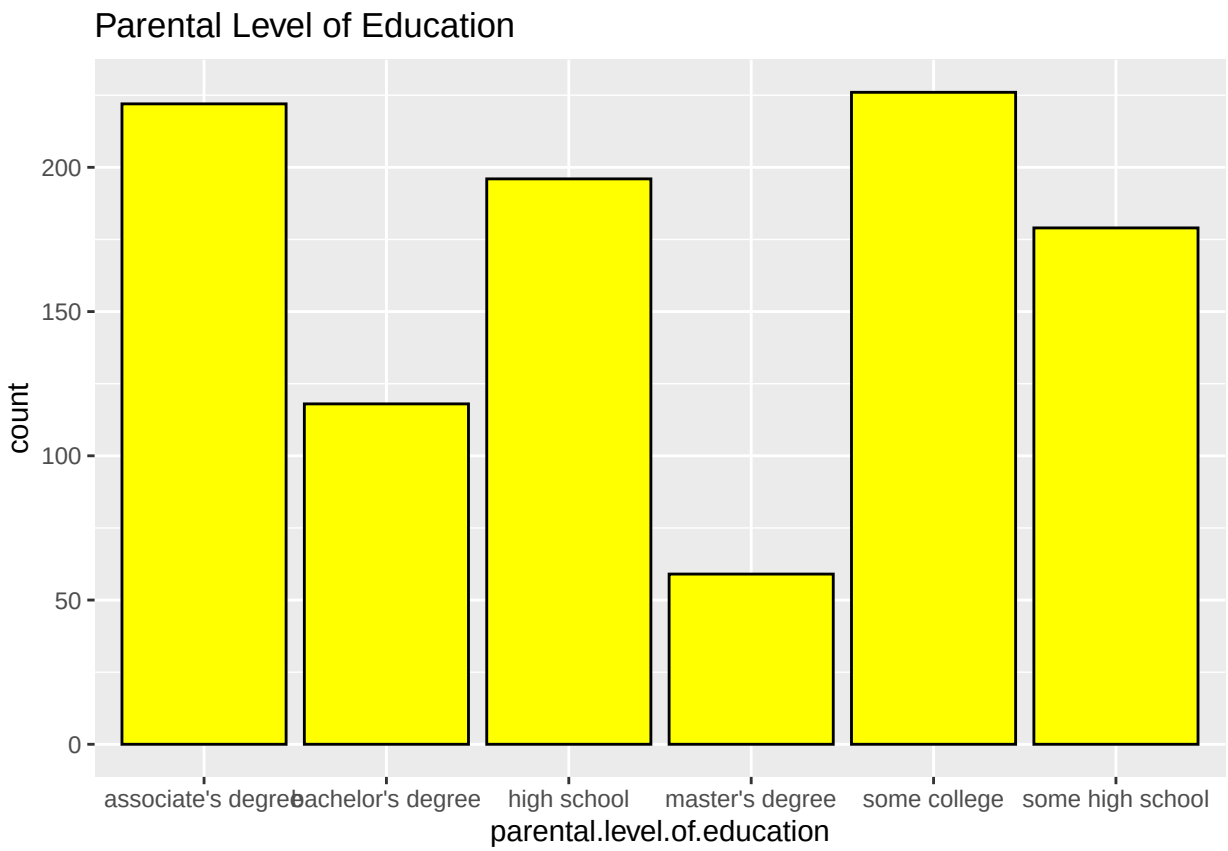

```
# Bar plots
ggplot(data, aes(x = gender)) +
  geom_bar(fill = "coral", color = "black") +
  ggtitle("Gender Distribution")
```



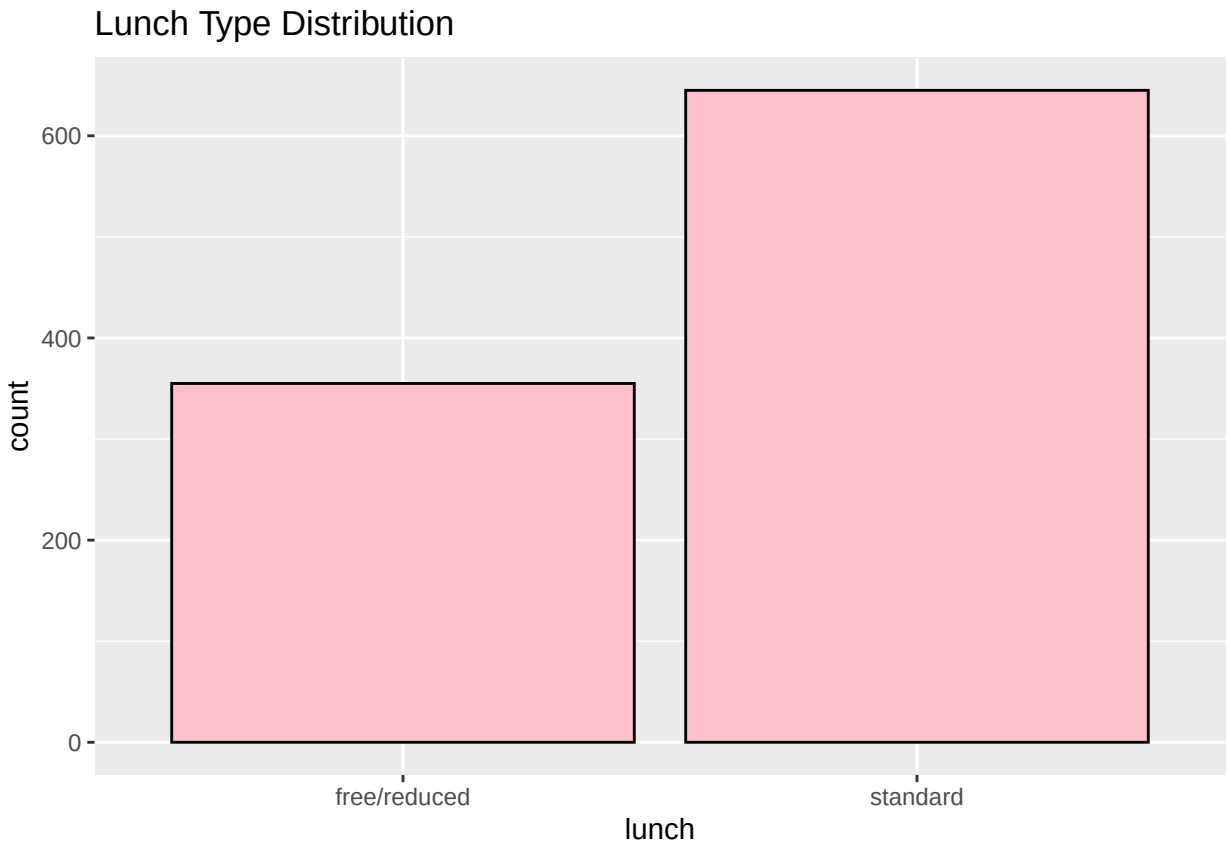

```
ggplot(data, aes(x = race.ethnicity)) +
  geom_bar(fill = "orange", color = "black") +
  ggtitle("Race/Ethnicity Distribution")
```



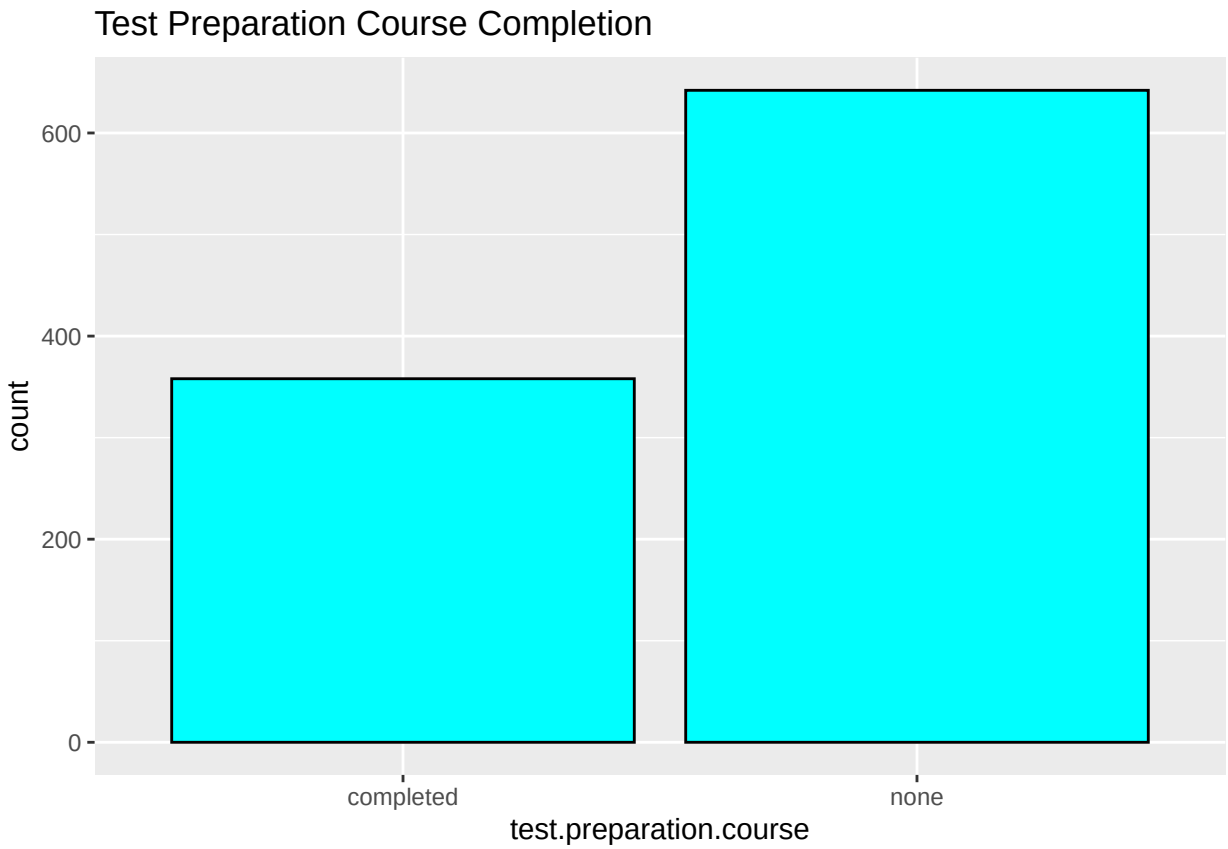

```
ggplot(data, aes(x = parental.level.of.education)) +
  geom_bar(fill = "yellow", color = "black") +
  ggtitle("Parental Level of Education")
```



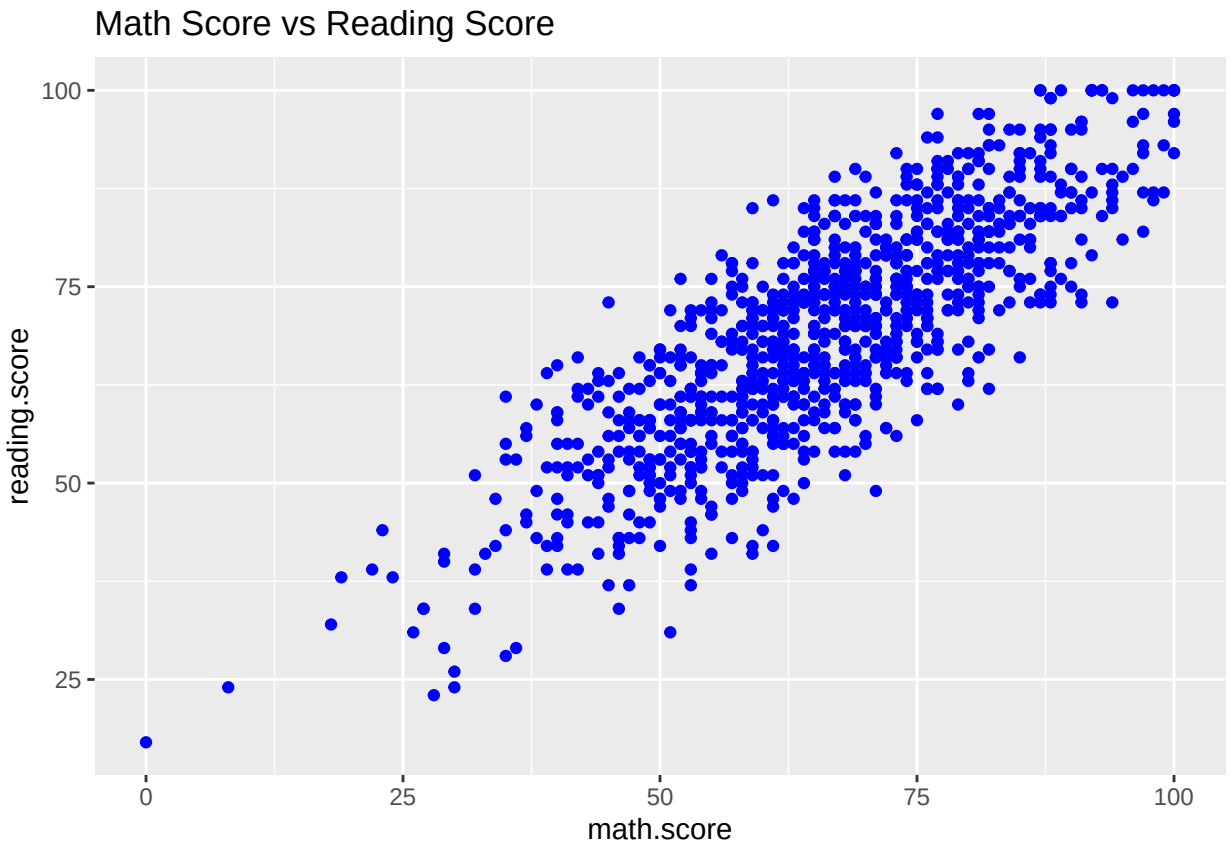

```
ggplot(data, aes(x = lunch)) +
  geom_bar(fill = "pink", color = "black") +
  ggtitle("Lunch Type Distribution")
```



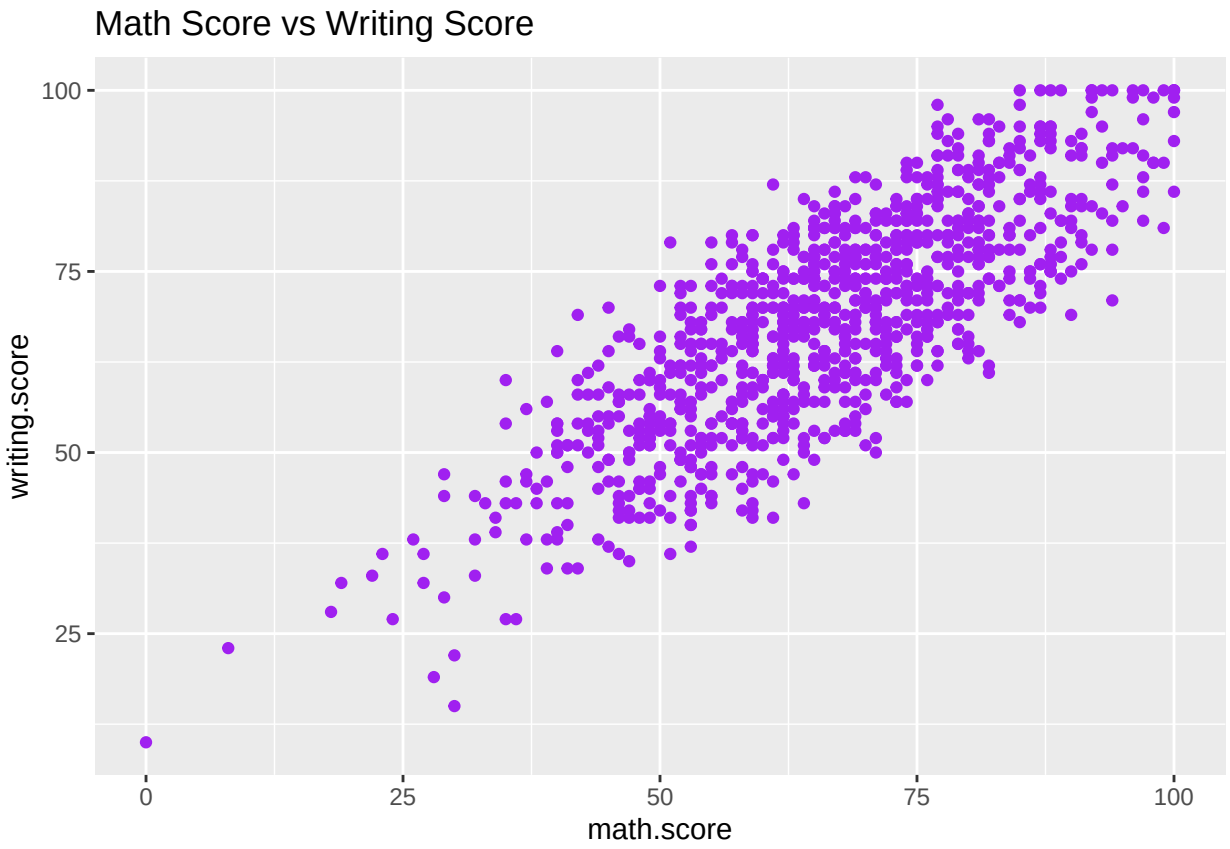

```
ggplot(data, aes(x = test.preparation.course)) +
  geom_bar(fill = "cyan", color = "black") +
  ggtitle("Test Preparation Course Completion")
```



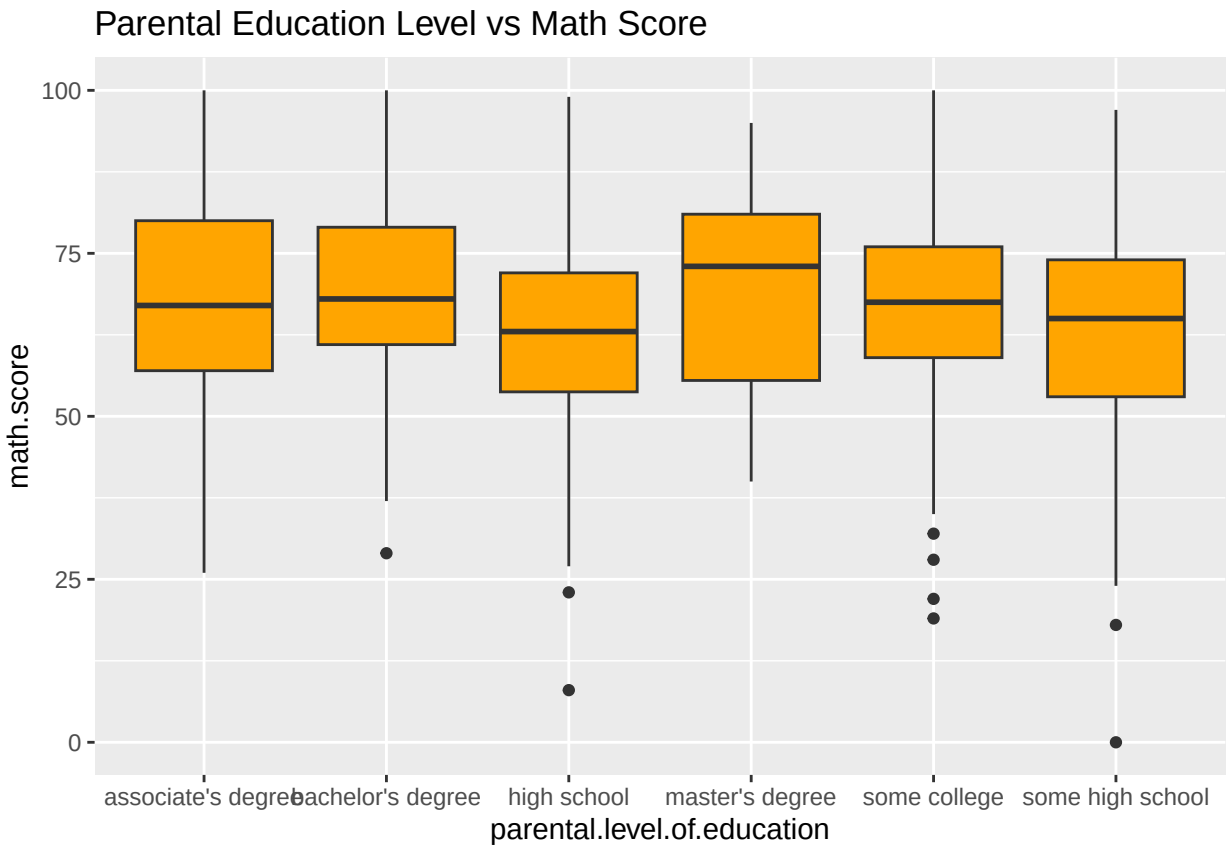

```
#Math Score and Reading Score
ggplot(data, aes(x = math.score, y = reading.score)) +
  geom_point(color = "blue") +
  ggtitle("Math Score vs Reading Score")
```




```
#Math Score and Writing Score
ggplot(data, aes(x = math.score, y = writing.score)) +
  geom_point(color = "purple") +
  ggtitle("Math Score vs Writing Score")
```




```
# Parental Education Level and Math Score
ggplot(data, aes(x = parental.level.of.education, y = math.score)) +
  geom_boxplot(fill = "orange") +
  ggtitle("Parental Education Level vs Math Score")
```




```
# Score variable
summary(data$math.score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   57.00   66.00   66.09   77.00  100.00
```

```
summary(data$reading.score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   59.00   70.00   69.17   79.00  100.00
```

```
summary(data$writing.score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   57.75   69.00   68.05   79.00  100.00
```

```
# Mean and standard deviation
mean_math <- mean(data$math.score)
mean_reading <- mean(data$reading.score)
mean_writing <- mean(data$writing.score)

sd_math <- sd(data$math.score)
sd_reading <- sd(data$reading.score)
```

```r
sd_writing <- sd(data$writing.score)

#Mean and standard deviation
mean_math
```

```
## [1] 66.089
```

```r
mean_reading
```

```
## [1] 69.169
```

```r
mean_writing
```

```
## [1] 68.054
```

```r
sd_math
```

```
## [1] 15.16308
```

```r
sd_reading
```

```
## [1] 14.60019
```

```r
sd_writing
```

```
## [1] 15.19566
```

```r
# Calculate the 50th, 90th, and 99th percentiles
math_percentiles <- quantile(data$math.score, probs = c(0.50, 0.90, 0.99))
reading_percentiles <- quantile(data$reading.score, probs = c(0.50, 0.90, 0.99))
writing_percentiles <- quantile(data$writing.score, probs = c(0.50, 0.90, 0.99))

#Percentiles
math_percentiles
```

```
##   50%   90%   99%
## 66.00 86.00 98.01
```

```r
reading_percentiles
```

```
##   50%   90%   99%
##  70.0  87.1 100.0
```

```r
writing_percentiles
```

```
## 50% 90% 99%
##  69  87 100
```

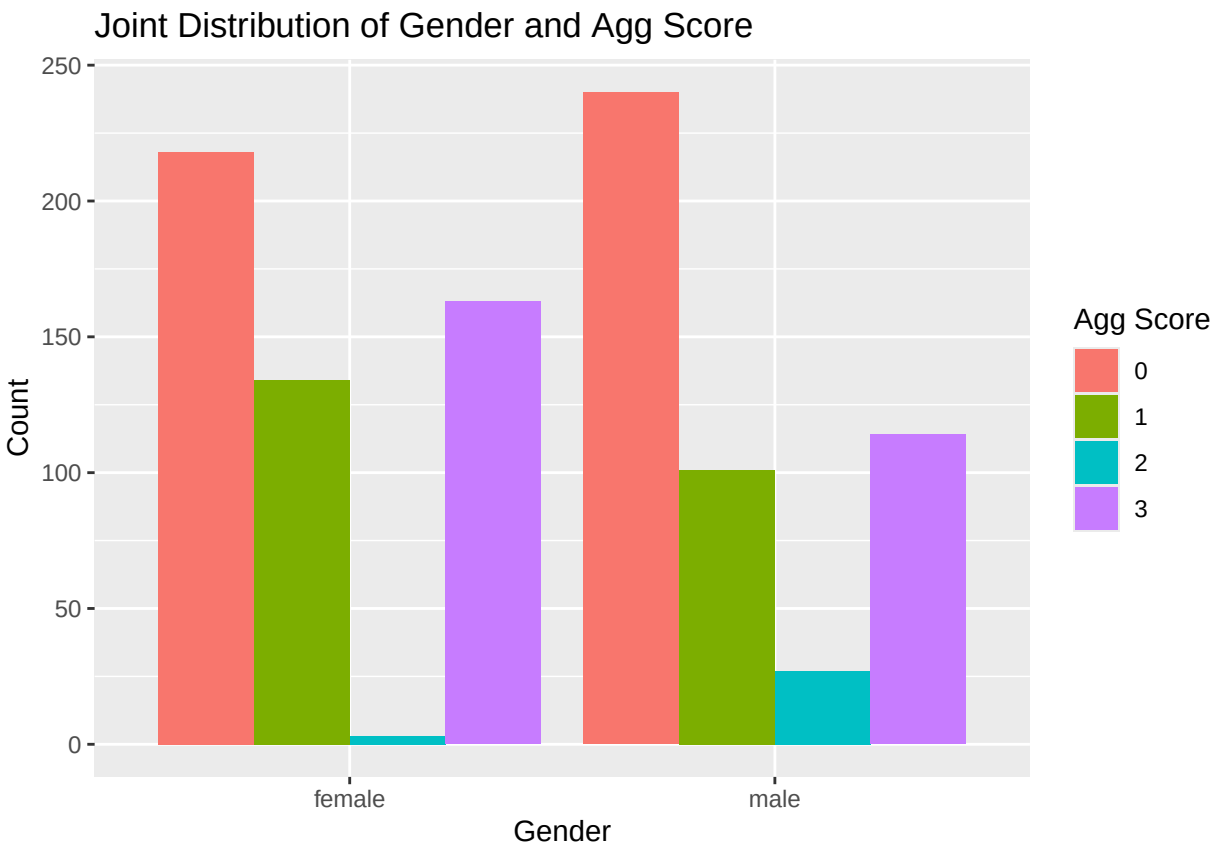
```r
# Convert parental.level.of.education
data$parental.level.of.education <- factor(data$parental.level.of.education,
                                           levels = c("some high school", "high school", "some college"
                                                      "associate's degree", "bachelor's degree", "maste
                                           ordered = TRUE)
head(data$parental.level.of.education)
```

```
## [1] bachelor's degree  some college       master's degree    associate's degree
## [5] some college       associate's degree
## 6 Levels: some high school < high school < ... < master's degree
```

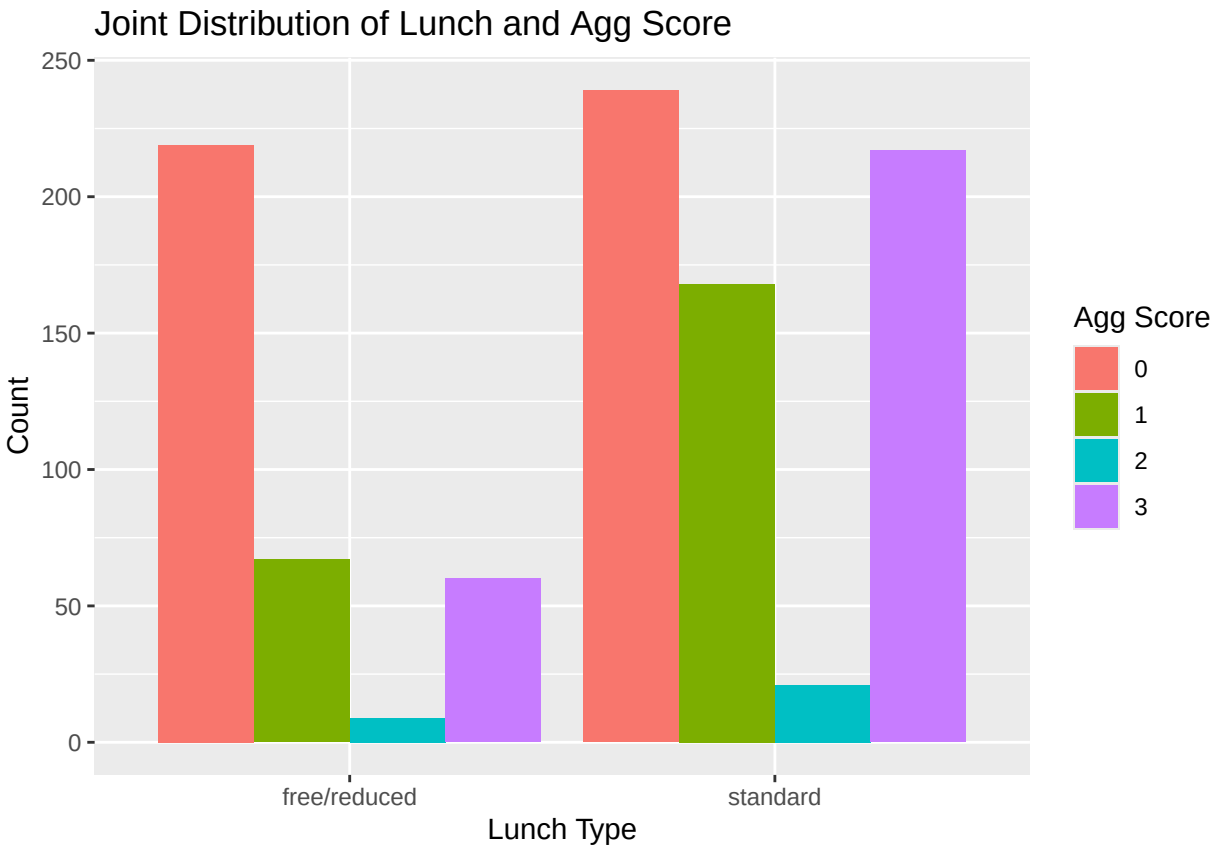
```r
# 60th percentile
math_60th <- quantile(data$math.score, 0.60)
reading_60th <- quantile(data$reading.score, 0.60)
writing_60th <- quantile(data$writing.score, 0.60)
#'agg_score' variable
data$agg_score <- ifelse(data$math.score >= math_60th & data$reading.score >= reading_60th & data$writi
                         ifelse(data$math.score >= math_60th & data$reading.score >= reading_60th, 2,
                                ifelse(data$math.score >= math_60th | data$reading.score >= reading_60
#Result
head(data$agg_score)
```

```
## [1] 1 1 3 0 3 3
```

```r
library(ggplot2)
# Plot 'gender' and 'agg_score'
ggplot(data, aes(x = gender, fill = factor(agg_score))) +
  geom_bar(position = "dodge") +
  labs(title = "Joint Distribution of Gender and Agg Score", x = "Gender", y = "Count", fill = "Agg Sco
```




```
# Plot 'lunch' and 'agg_score'
ggplot(data, aes(x = lunch, fill = factor(agg_score))) +
  geom_bar(position = "dodge") +
  labs(title = "Joint Distribution of Lunch and Agg Score", x = "Lunch Type", y = "Count", fill = "Agg S
```




```
# Did not score in the top 60th percentile
data_no_top_60 <- subset(data, agg_score == 0)

# Top 60th percentile for at least one subject
data_top_60 <- subset(data, agg_score >= 1)

head(data_no_top_60)
```

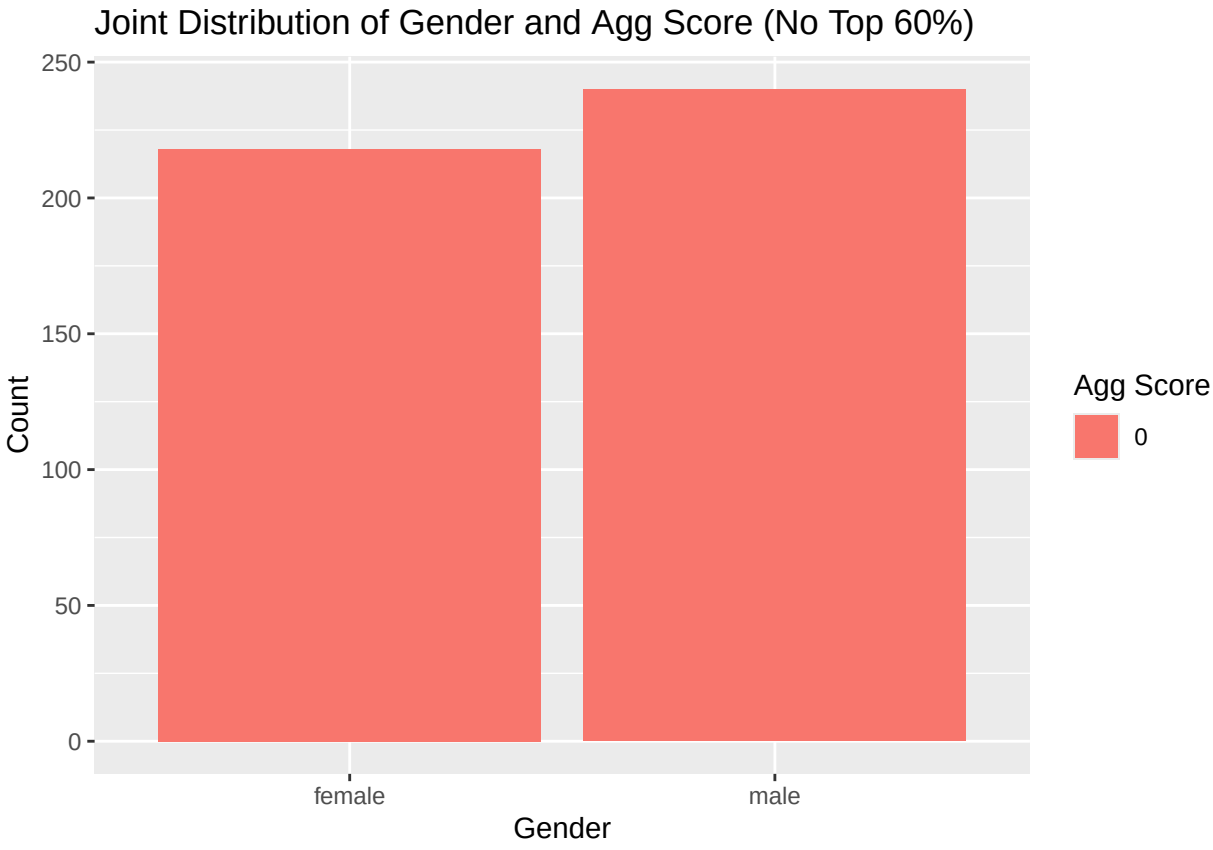
```
##    gender race.ethnicity parental.level.of.education        lunch
## 4    male        group A          associate's degree free/reduced
## 8    male        group B                some college free/reduced
## 9    male        group D                 high school free/reduced
## 10 female        group B                 high school free/reduced
## 11   male        group C          associate's degree     standard
## 12   male        group D          associate's degree     standard
##    test.preparation.course math.score reading.score writing.score agg_score
## 4                     none         47            57            44         0
## 8                     none         40            43            39         0
## 9                completed         64            64            67         0
## 10                    none         38            60            50         0
## 11                    none         58            54            52         0
## 12                    none         40            52            43         0
```

```
head(data_top_60)
```

```
##   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
## 5   male        group C                some college standard
## 6 female        group B            associate's degree standard
## 7 female        group B                some college standard
##   test.preparation.course math.score reading.score writing.score agg_score
## 1                    none         72            72            74         1
## 2               completed         69            90            88         1
## 3                    none         90            95            93         3
## 5                    none         76            78            75         3
## 6                    none         71            83            78         3
## 7               completed         88            95            92         3
```

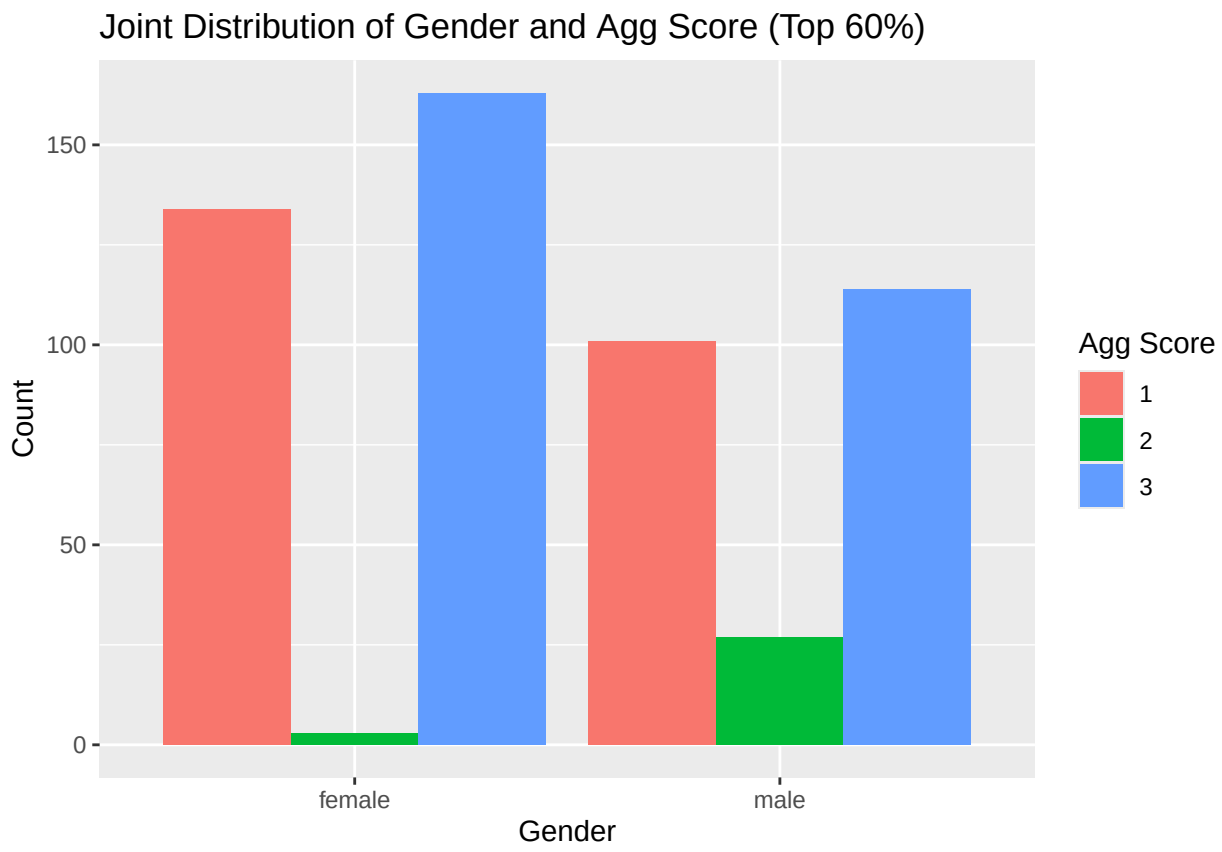
```r
library(ggplot2)

# Not in the top 60th percentile
ggplot(data_no_top_60, aes(x = gender, fill = factor(agg_score))) +
  geom_bar(position = "dodge") +
  labs(title = "Joint Distribution of Gender and Agg Score (No Top 60%)",
       x = "Gender", y = "Count", fill = "Agg Score")
```




```r
# Top 60th percentile for at least one subject
ggplot(data_top_60, aes(x = gender, fill = factor(agg_score))) +
```

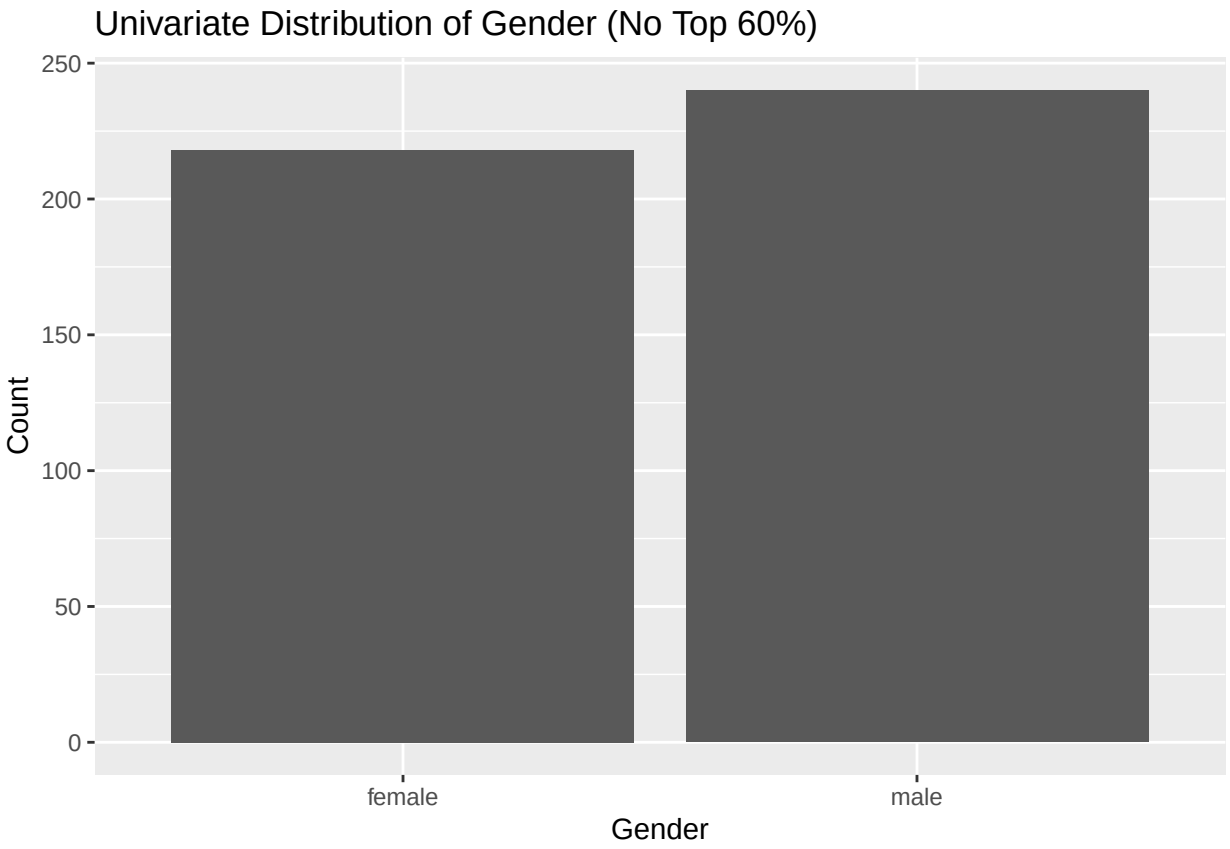
```
  geom_bar(position = "dodge") +
  labs(title = "Joint Distribution of Gender and Agg Score (Top 60%)",
       x = "Gender", y = "Count", fill = "Agg Score")
```



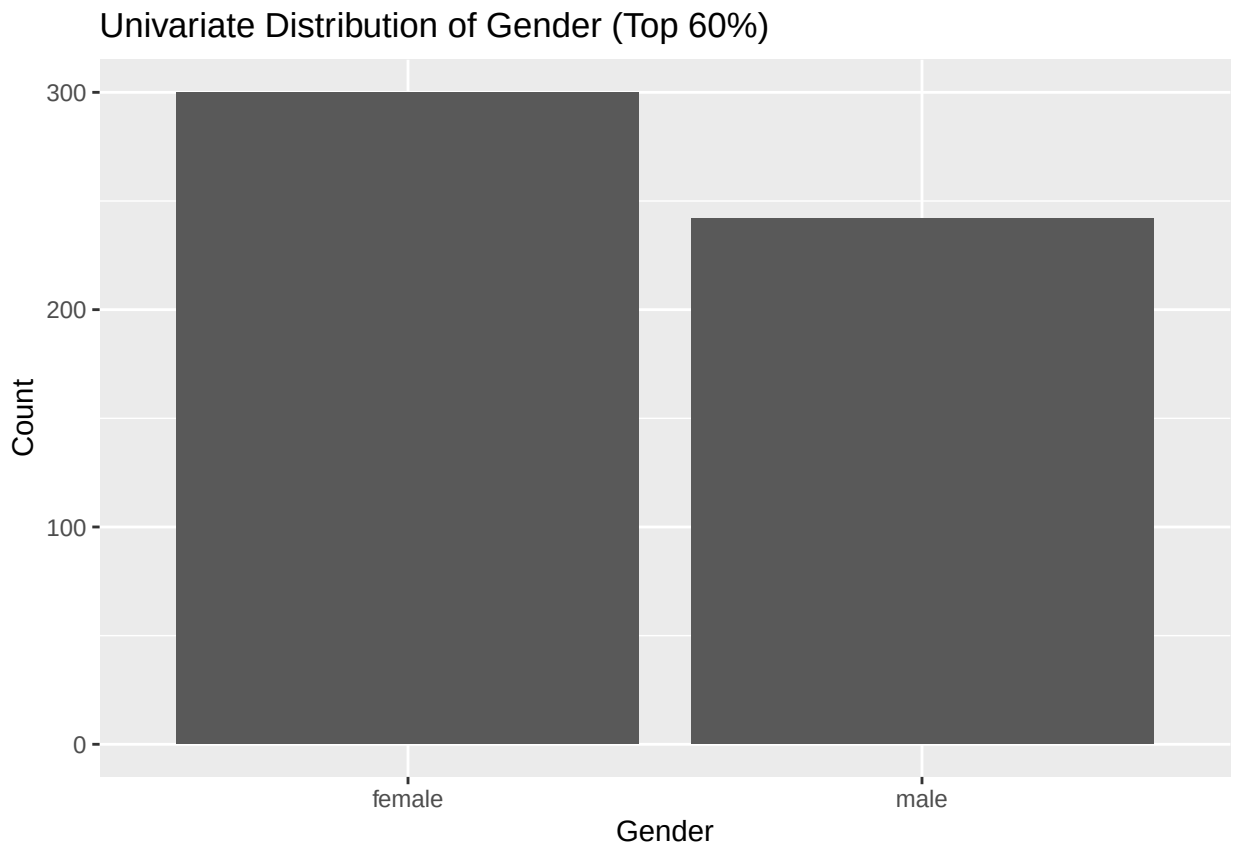

```
# Joint Distribution of 'lunch' and 'agg_score' for the two datasets
# For students who did not score in the top 60th percentile
ggplot(data_no_top_60, aes(x = lunch, fill = factor(agg_score))) +
  geom_bar(position = "dodge") +
  labs(title = "Joint Distribution of Lunch and Agg Score (No Top 60%)",
       x = "Lunch Type", y = "Count", fill = "Agg Score")
```


```r
# For students who scored in the top 60th percentile for at least one subject
ggplot(data_top_60, aes(x = lunch, fill = factor(agg_score))) +
  geom_bar(position = "dodge") +
  labs(title = "Joint Distribution of Lunch and Agg Score (Top 60%)",
       x = "Lunch Type", y = "Count", fill = "Agg Score")
```


```r
# Univariate plot
ggplot(data_no_top_60, aes(x = gender)) +
  geom_bar() +
  labs(title = "Univariate Distribution of Gender (No Top 60%)",
       x = "Gender", y = "Count")
```

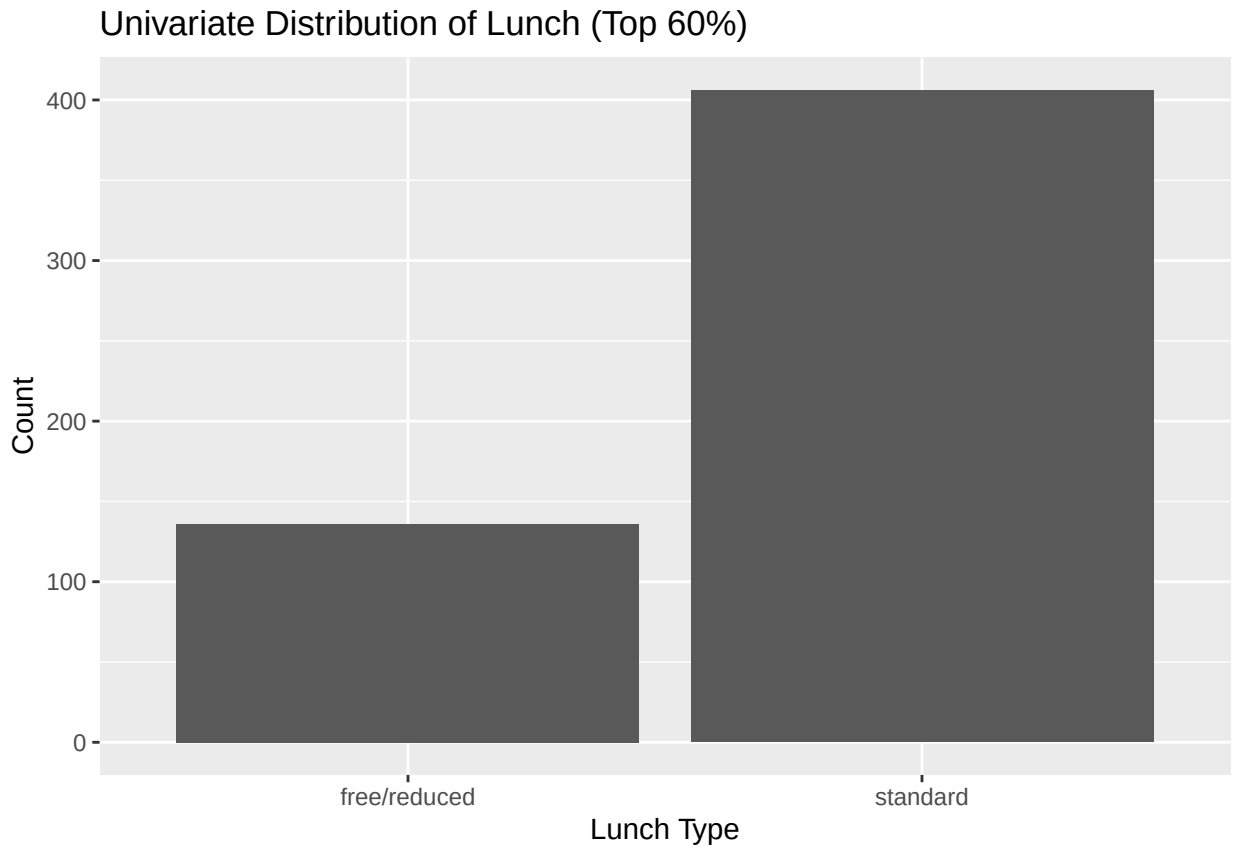



```
ggplot(data_top_60, aes(x = gender)) +
  geom_bar() +
  labs(title = "Univariate Distribution of Gender (Top 60%)",
       x = "Gender", y = "Count")
```




```
# Univariate plot of 'lunch'
ggplot(data_no_top_60, aes(x = lunch)) +
  geom_bar() +
  labs(title = "Univariate Distribution of Lunch (No Top 60%)",
       x = "Lunch Type", y = "Count")
```


```
ggplot(data_top_60, aes(x = lunch)) +
  geom_bar() +
  labs(title = "Univariate Distribution of Lunch (Top 60%)",
       x = "Lunch Type", y = "Count")
```




```
# Math
highest_math <- data[order(data$math.score, -data$reading.score)[1], ]
print(highest_math)
```

```
##    gender race.ethnicity parental.level.of.education        lunch
## 60 female        group C            some high school free/reduced
##    test.preparation.course math.score reading.score writing.score agg_score
## 60                    none          0            17            10         0
```

```
#Reading
highest_reading <- data[order(data$reading.score, -data$math.score)[1], ]
print(highest_reading)
```

```
##    gender race.ethnicity parental.level.of.education        lunch
## 60 female        group C            some high school free/reduced
##    test.preparation.course math.score reading.score writing.score agg_score
## 60                    none          0            17            10         0
```

```
# Writing
highest_writing <- data[order(data$writing.score, -data$math.score)[1], ]
print(highest_writing)
```

```
##    gender race.ethnicity parental.level.of.education        lunch
## 60 female        group C            some high school free/reduced
```

```
##    test.preparation.course math.score reading.score writing.score agg_score
## 60                    none          0            17            10         0
```

```
# Find the person who scored the lowest in Math
lowest_math <- data[order(data$math.score, data$reading.score)[1], ]

# Display the values of every variable for the person who scored lowest in Math
print(lowest_math)
```

```
##    gender race.ethnicity parental.level.of.education        lunch
## 60 female        group C            some high school free/reduced
##    test.preparation.course math.score reading.score writing.score agg_score
## 60                    none          0            17            10         0
```

```
# Find the person who scored the lowest in Reading
lowest_reading <- data[order(data$reading.score, data$math.score)[1], ]

# Display the values of every variable for the person who scored lowest in Reading
print(lowest_reading)
```

```
##    gender race.ethnicity parental.level.of.education        lunch
## 60 female        group C            some high school free/reduced
##    test.preparation.course math.score reading.score writing.score agg_score
## 60                    none          0            17            10         0
```

```
# Find the person who scored the lowest in Writing
lowest_writing <- data[order(data$writing.score, data$math.score)[1], ]

# Display the values of every variable for the person who scored lowest in Writing
print(lowest_writing)
```

```
##    gender race.ethnicity parental.level.of.education        lunch
## 60 female        group C            some high school free/reduced
##    test.preparation.course math.score reading.score writing.score agg_score
## 60                    none          0            17            10         0
```

```
#free/reduced lunch
free_reduced_lunch_data <- subset(data, lunch == "free/reduced")
head(free_reduced_lunch_data)
```

```
##    gender race.ethnicity parental.level.of.education        lunch
## 4    male        group A          associate's degree free/reduced
## 8    male        group B                some college free/reduced
## 9    male        group D                 high school free/reduced
## 10 female        group B                 high school free/reduced
## 18 female        group B            some high school free/reduced
## 19   male        group C             master's degree free/reduced
##    test.preparation.course math.score reading.score writing.score agg_score
## 4                     none         47            57            44         0
## 8                     none         40            43            39         0
## 9                completed         64            64            67         0
```

```
## 10                    none         38            60            50         0
## 18                    none         18            32            28         0
## 19               completed         46            42            46         0
```

```
# filtered dataset
write.csv(free_reduced_lunch_data, "free_reduced_lunch_students.csv", row.names = FALSE)
```

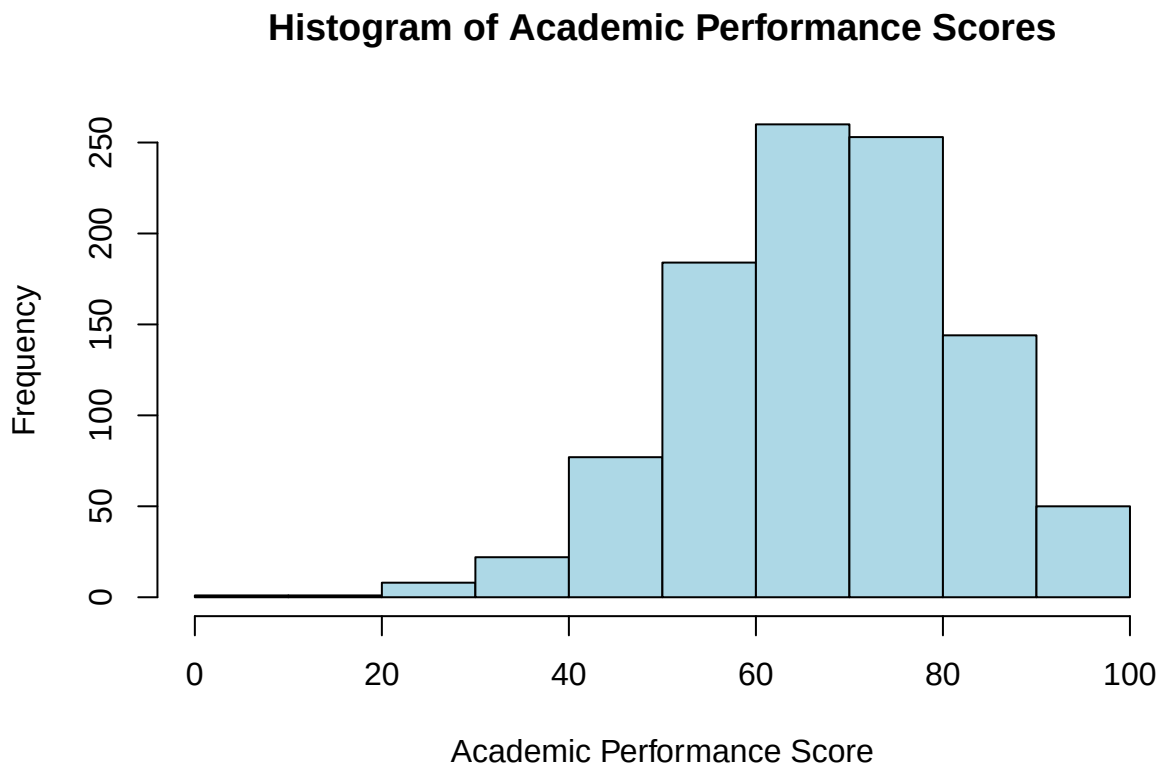
```
# Create a new variable 'aps_score' which is the average of the math, reading, and writing scores
data$aps_score <- rowMeans(data[, c("math.score", "reading.score", "writing.score")])
head(data)
```

```
##   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 agg_score
## 1                    none         72            72            74         1
## 2               completed         69            90            88         1
## 3                    none         90            95            93         3
## 4                    none         47            57            44         0
## 5                    none         76            78            75         3
## 6                    none         71            83            78         3
##   aps_score
## 1  72.66667
## 2  82.33333
## 3  92.66667
## 4  49.33333
## 5  76.33333
## 6  77.33333
```

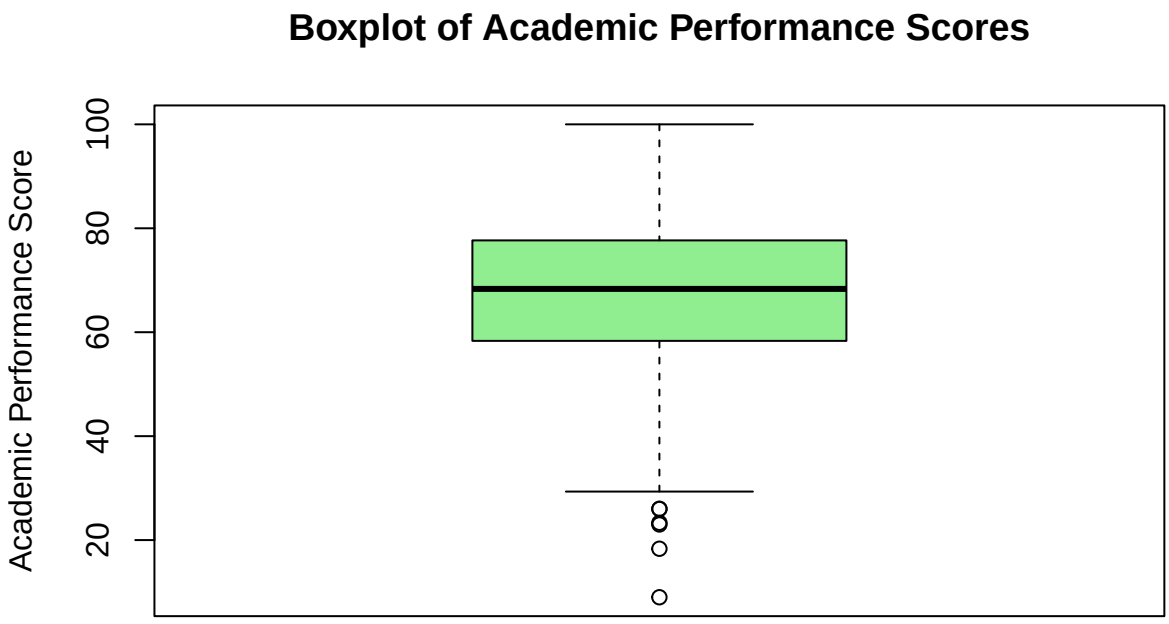
```
summary(data$aps_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   58.33   68.33   67.77   77.67  100.00
```

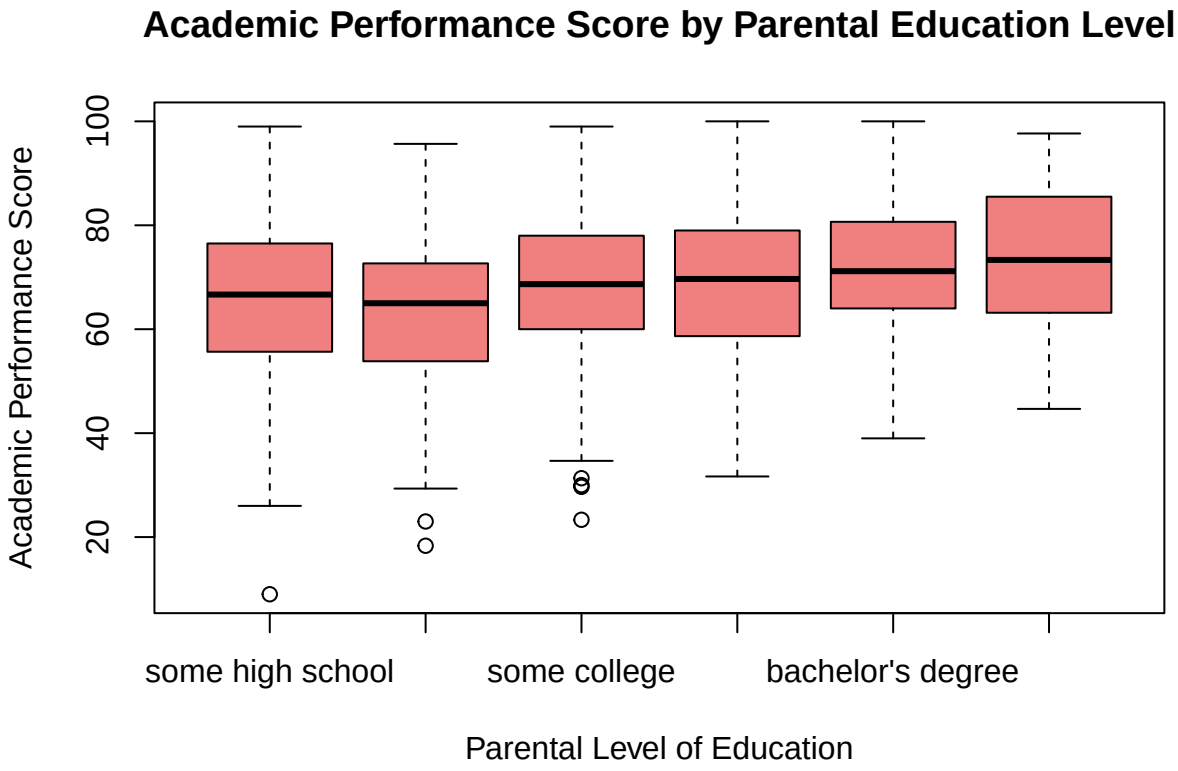
```
# Histogram
hist(data$aps_score, main = "Histogram of Academic Performance Scores",
     xlab = "Academic Performance Score", col = "lightblue", breaks = 10)
```



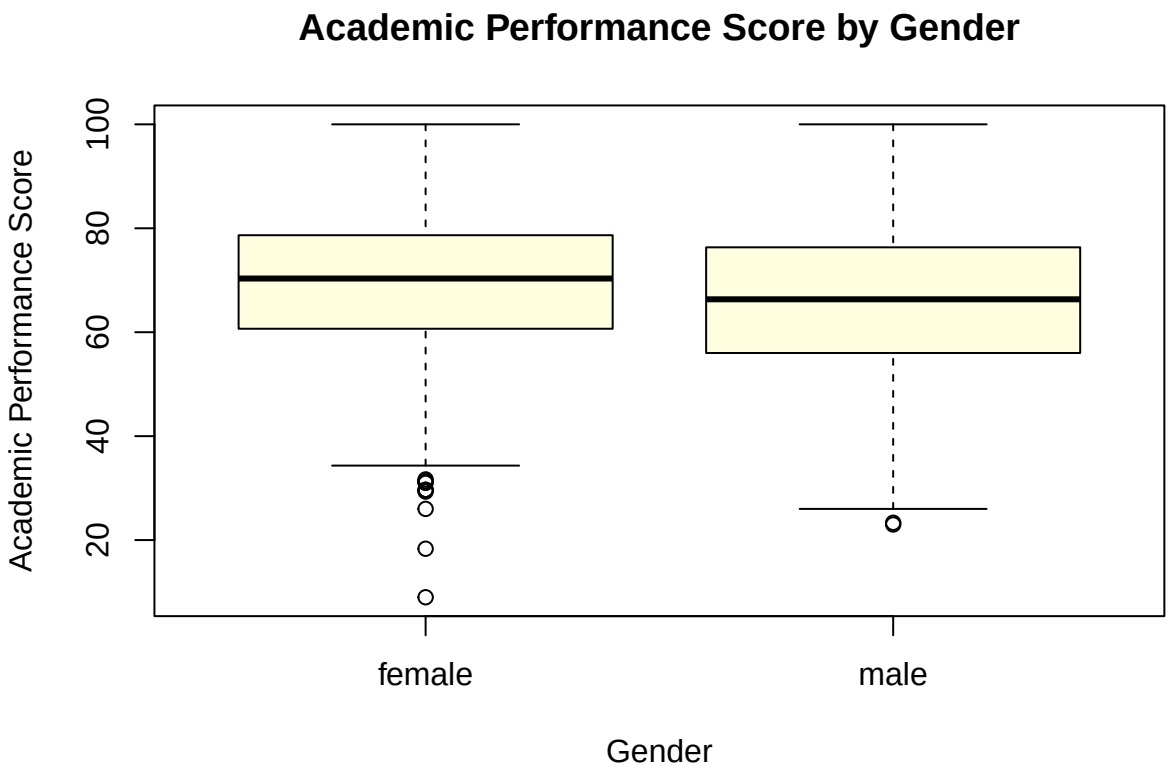

```r
# Boxplot
boxplot(data$aps_score, main = "Boxplot of Academic Performance Scores",
        ylab = "Academic Performance Score", col = "lightgreen")
```




```r
# Scatter plot
#parental.level.of.education = factor variable
data$parental.level.of.education <- factor(data$parental.level.of.education,
                                           levels = c("some high school", "high school", "some college"
                                                      "associate's degree", "bachelor's degree", "maste
# Plotting
boxplot(aps_score ~ parental.level.of.education, data = data,
        main = "Academic Performance Score by Parental Education Level",
        xlab = "Parental Level of Education", ylab = "Academic Performance Score",
        col = "lightcoral")
```




```r
# Boxplot
boxplot(aps_score ~ gender, data = data,
        main = "Academic Performance Score by Gender",
        xlab = "Gender", ylab = "Academic Performance Score",
        col = "lightyellow")
```




```
# Correlation matrix
cor(data[, c("aps_score", "math.score", "reading.score", "writing.score")])
```

```
##               aps_score math.score reading.score writing.score
## aps_score     1.0000000  0.9187458     0.9703307     0.9656672
## math.score    0.9187458  1.0000000     0.8175797     0.8026420
## reading.score 0.9703307  0.8175797     1.0000000     0.9545981
## writing.score 0.9656672  0.8026420     0.9545981     1.0000000
```