Dta analysis Final Project

2023-04-26

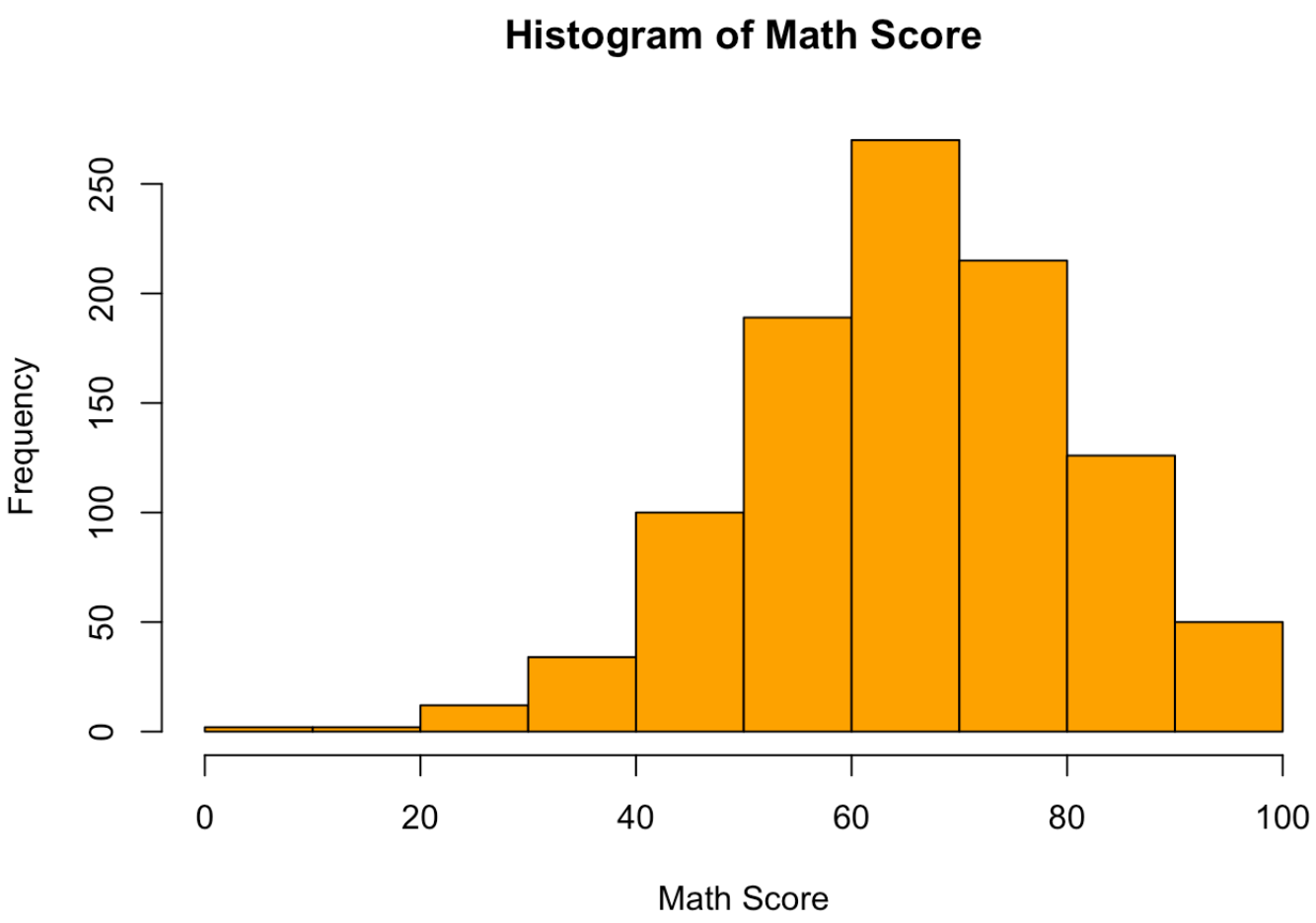
```
data <- read.csv("StudentsPerformance.csv",header=TRUE, sep=",")

summary(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
##  Min.   :0.000   Length:1000             Min.   :  0.00   Min.   : 17.00
##  1st Qu.:0.000   Class :character        1st Qu.: 57.00   1st Qu.: 59.00
##  Median :1.000   Mode  :character        Median : 66.00   Median : 70.00
##  Mean   :0.645                           Mean   : 66.09   Mean   : 69.17
##  3rd Qu.:1.000                           3rd Qu.: 77.00   3rd Qu.: 79.00
##  Max.   :1.000                           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
```

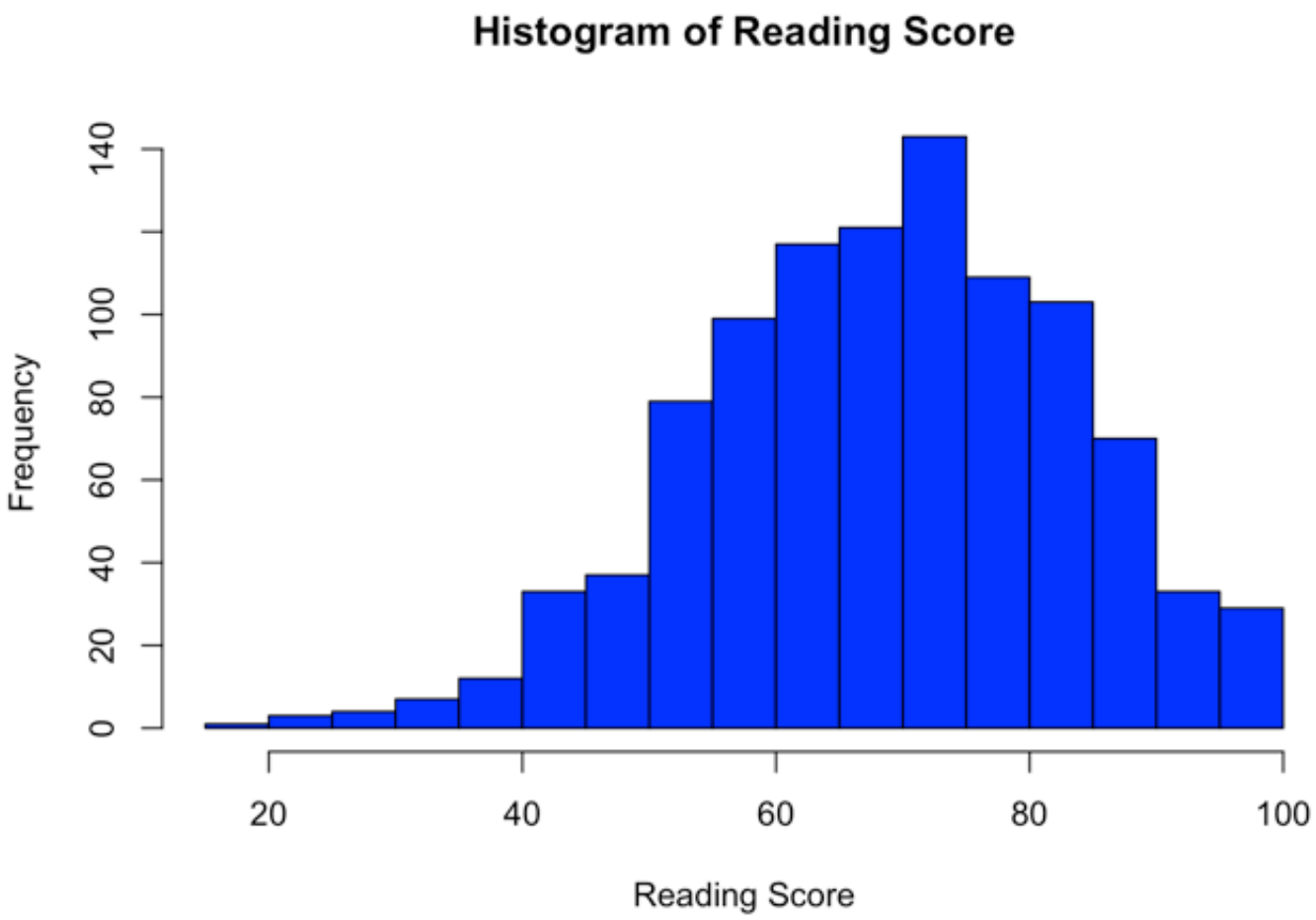
```
sum(is.na(data))
```

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

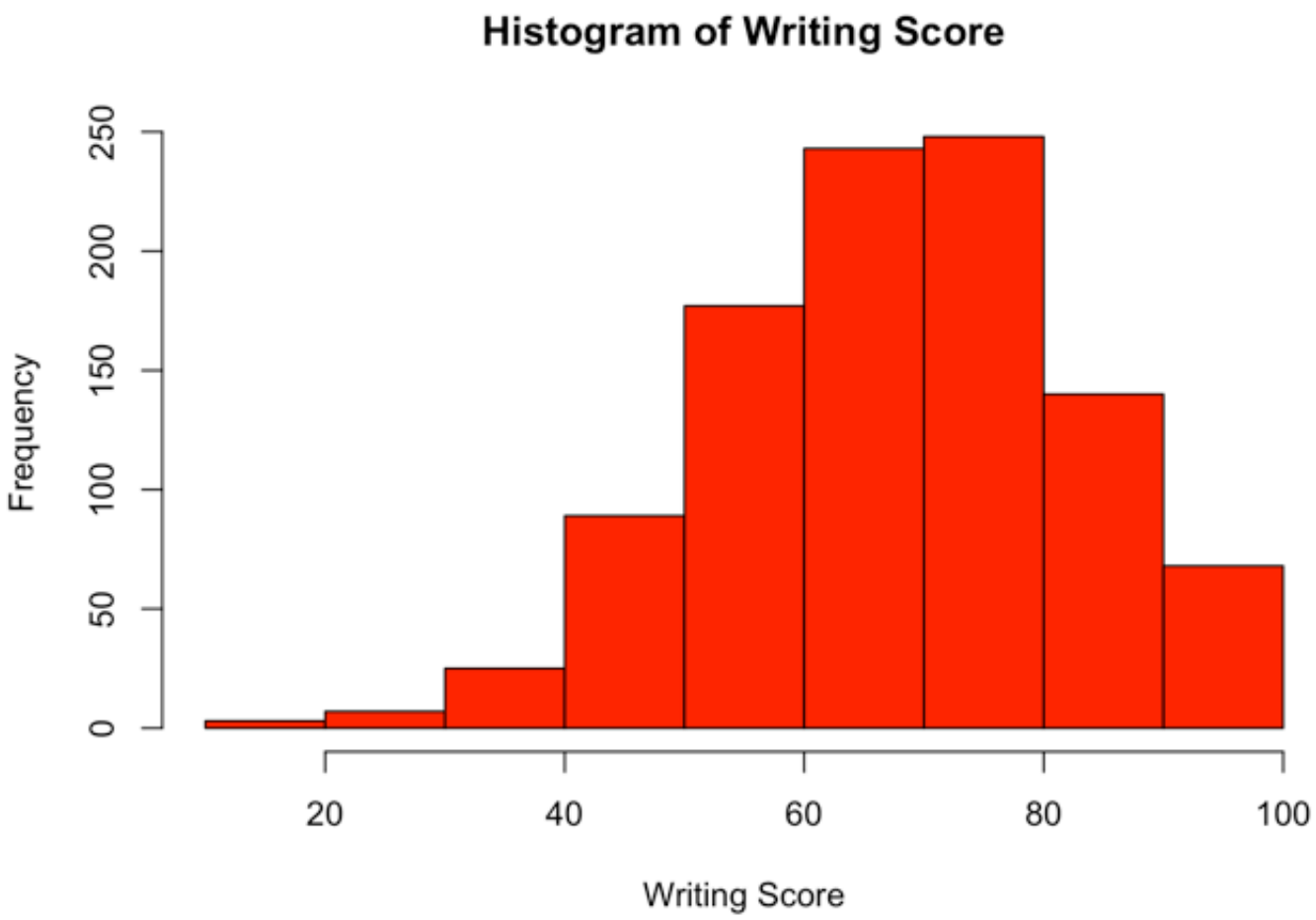
```
hist(
  data$math.score, col = "orange",
  main = "Histogram of Math Score",
  xlab= "Math Score",
  )
```



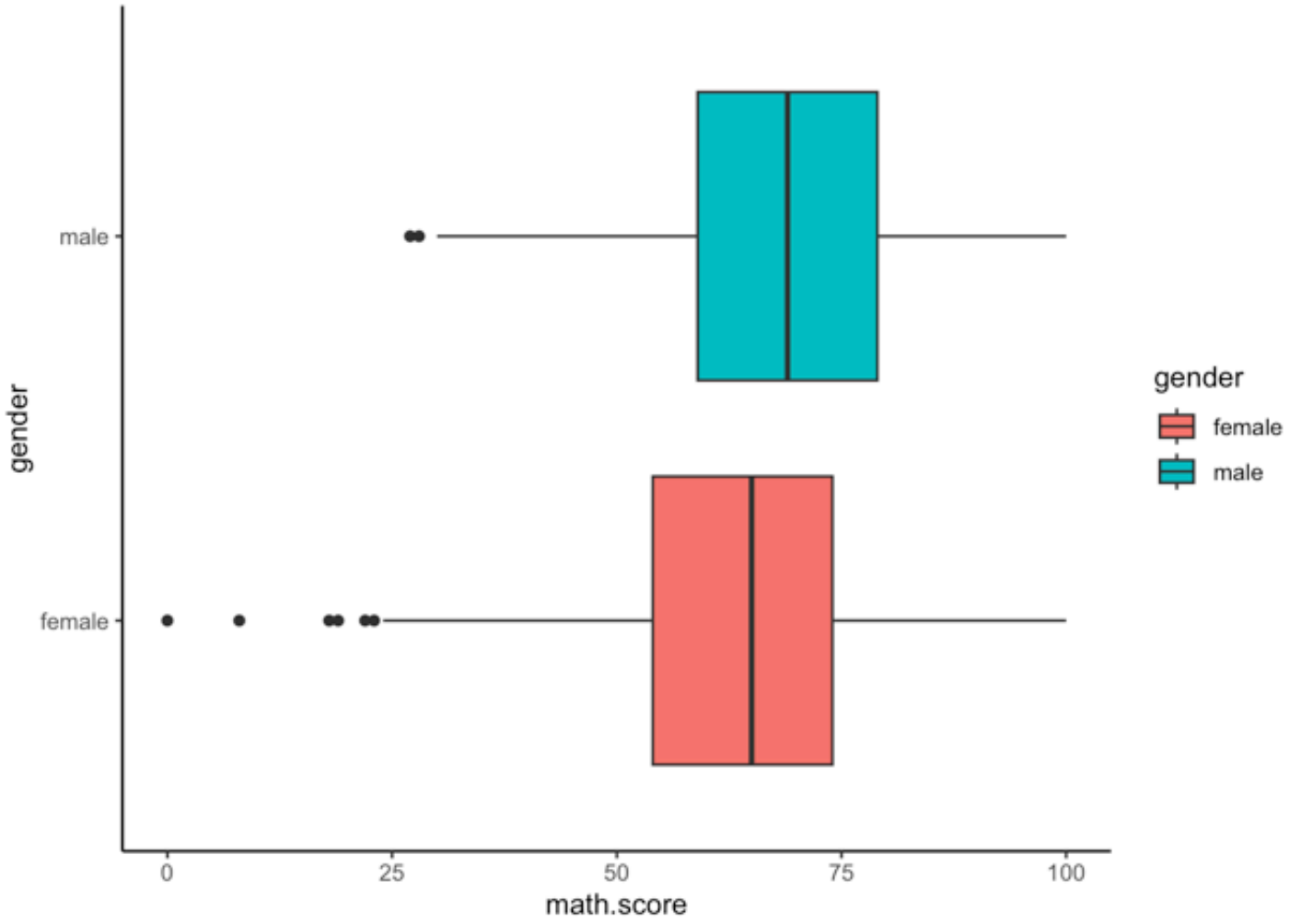

```
hist(
  data$reading.score, col = "blue",
  breaks = 20,
  main = "Histogram of Reading Score",
  xlab= "Reading Score",
  )
```



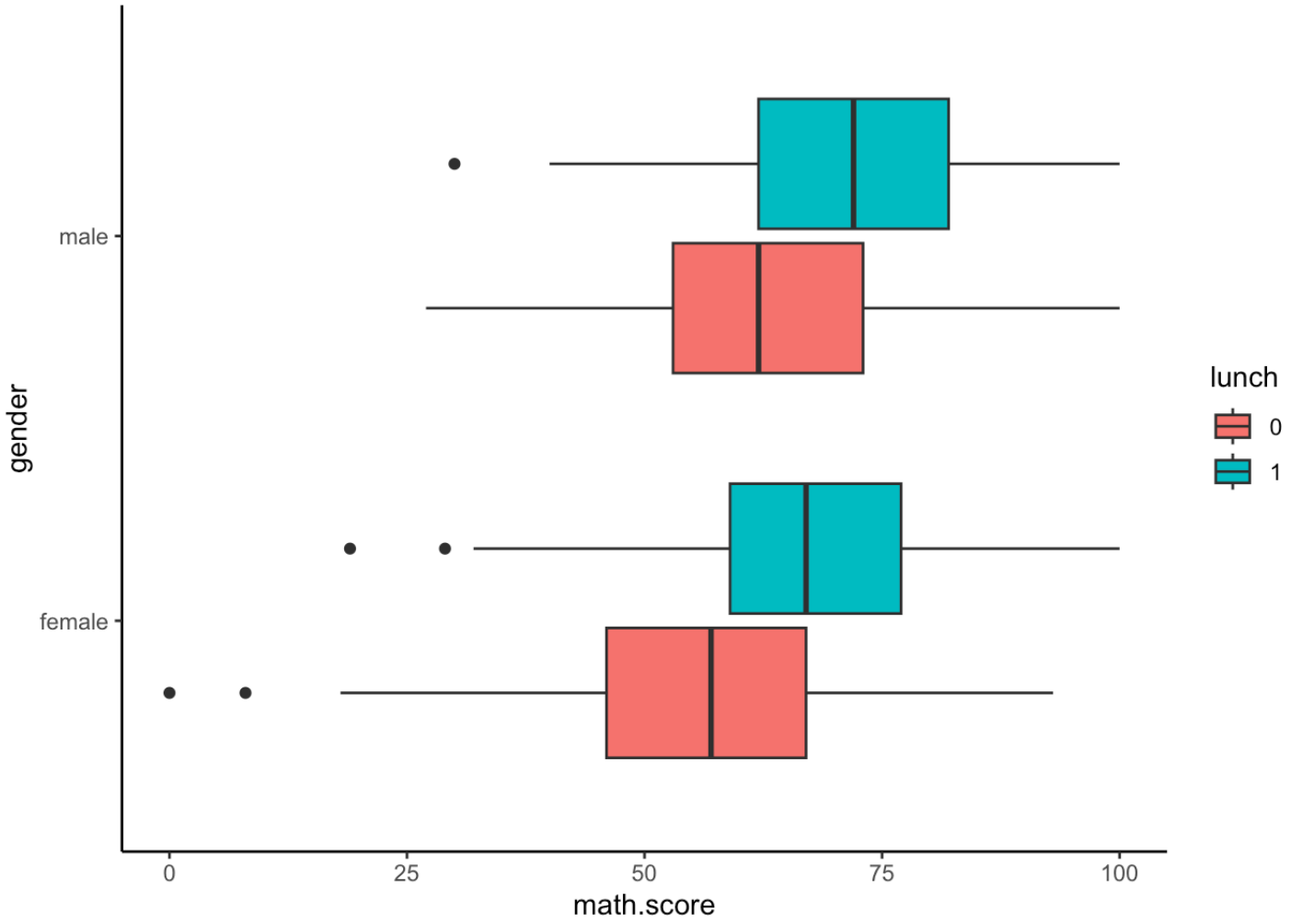

```
hist(
  data$writing.score, col = "red",
  breaks = 7,
  main = "Histogram of Writing Score",
  xlab= "Writing Score",
  )
```



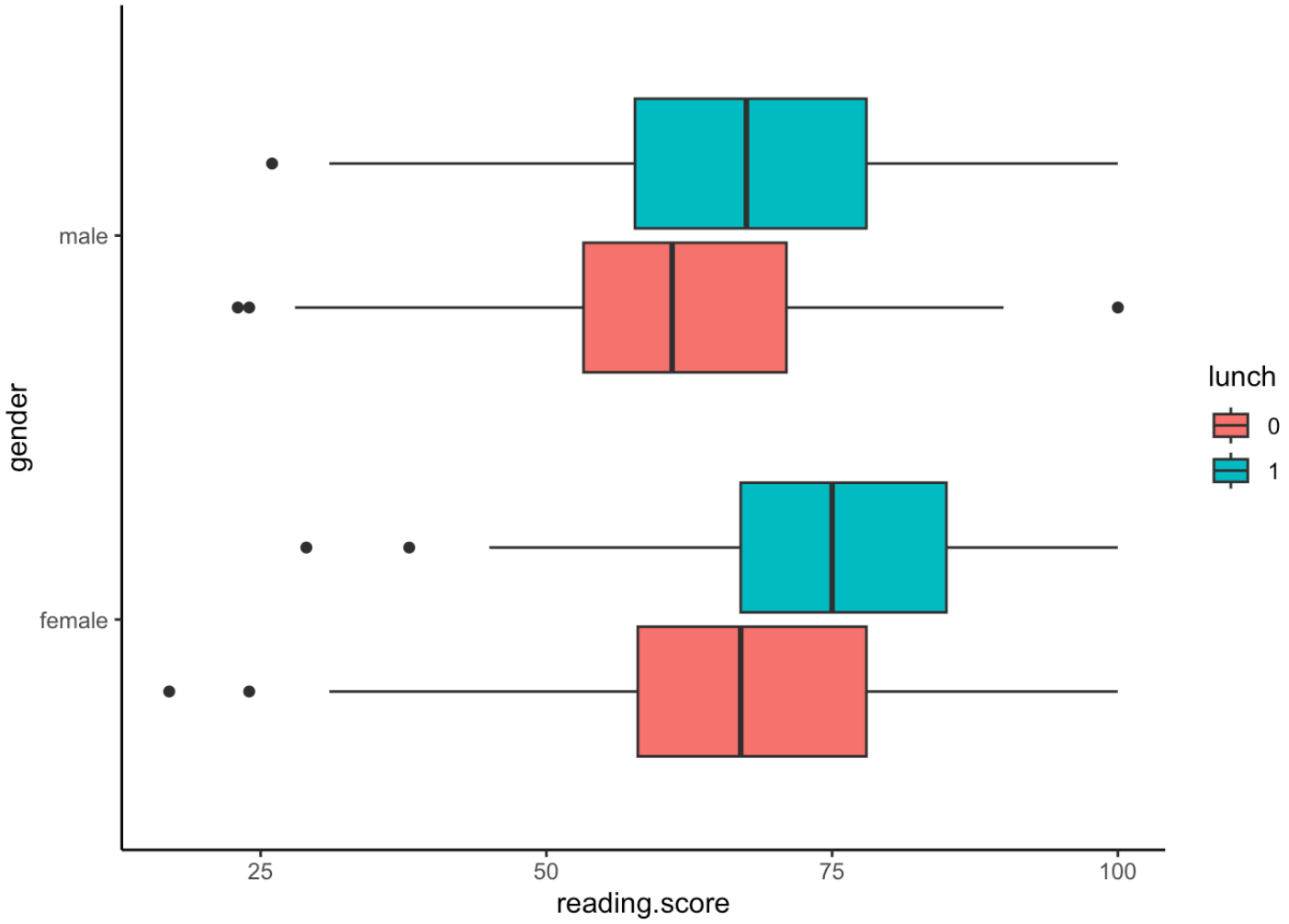

```
ggplot(data = data) +
  geom_boxplot(mapping = aes(x =gender , y =math.score, fill=gender)) +
  theme_classic()+
  scale_color_viridis_d()+
  coord_flip()
```



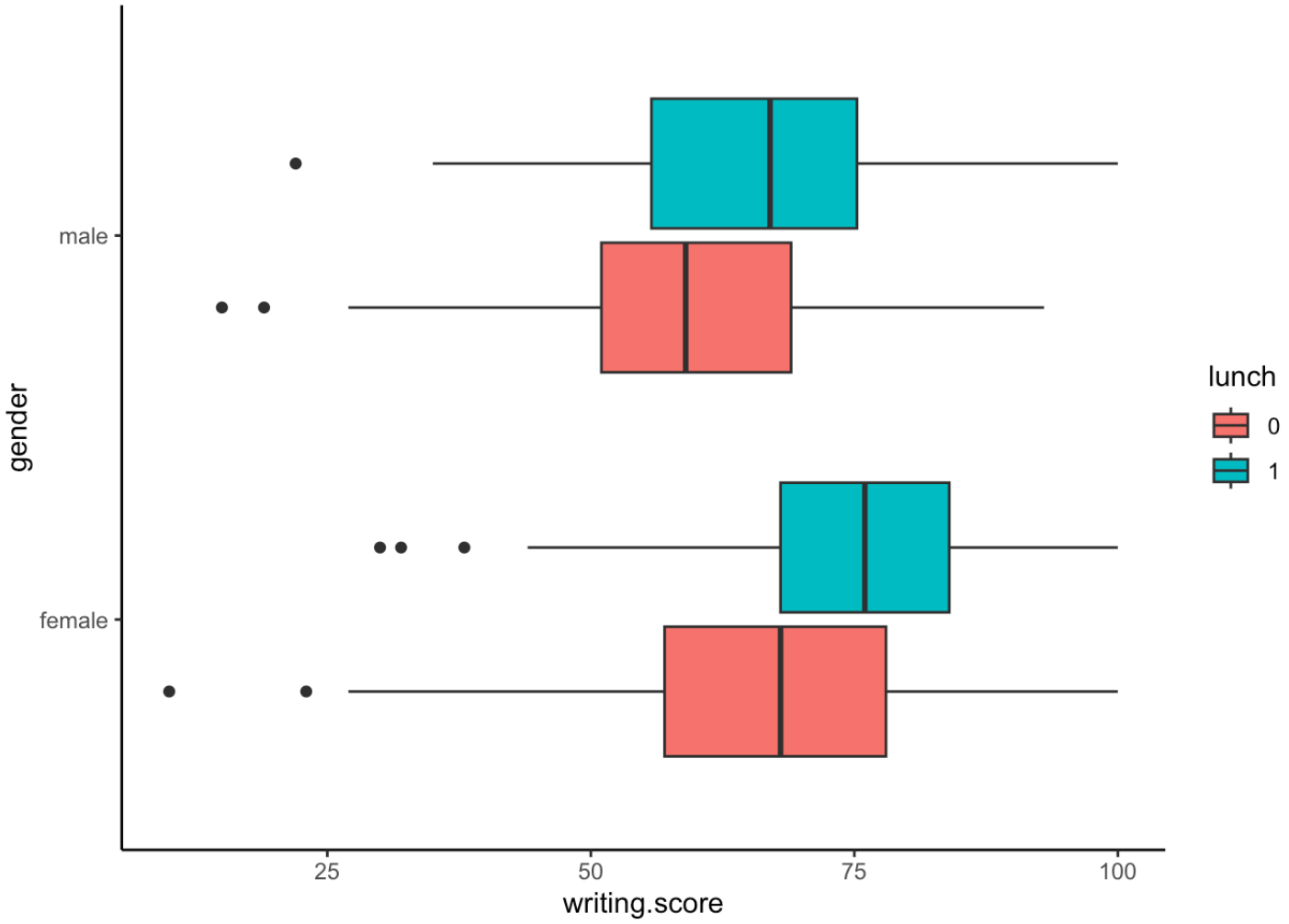

```
data$lunch<-factor(data$lunch)
ggplot(data = data) +
  geom_boxplot(mapping = aes(x =gender , y =math.score, fill=lunch)) +
  theme_classic()+
  scale_color_viridis_d()+
  coord_flip()
```




```
data$lunch<-factor(data$lunch)
ggplot(data = data) +
  geom_boxplot(mapping = aes(x =gender , y =reading.score, fill=lunch)) +
  theme_classic()+
  scale_color_viridis_d()+
  coord_flip()
```



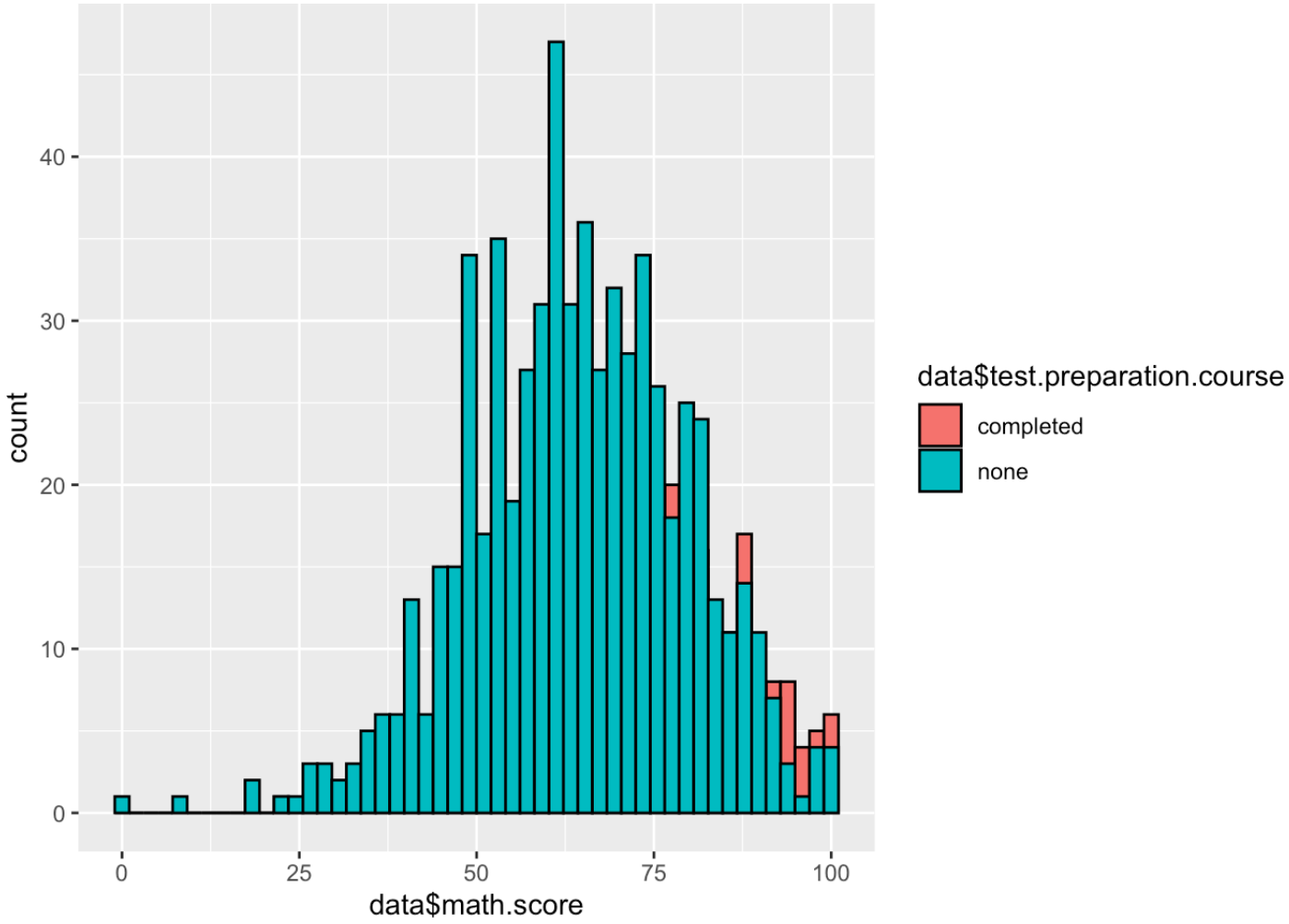

```
data$lunch<-factor(data$lunch)
ggplot(data = data) +
  geom_boxplot(mapping = aes(x =gender , y =writing.score, fill=lunch)) +
  theme_classic()+
  scale_color_viridis_d()+
  coord_flip()
```




```
ggplot(data, aes(x = data$math.score, fill = data$test.preparation.course)) +
  geom_histogram(position = "identity", alpha = 1.2, bins = 50, color="black")
```

```
## Warning: Use of `data$math.score` is discouraged.
## ℹ Use `math.score` instead.
```

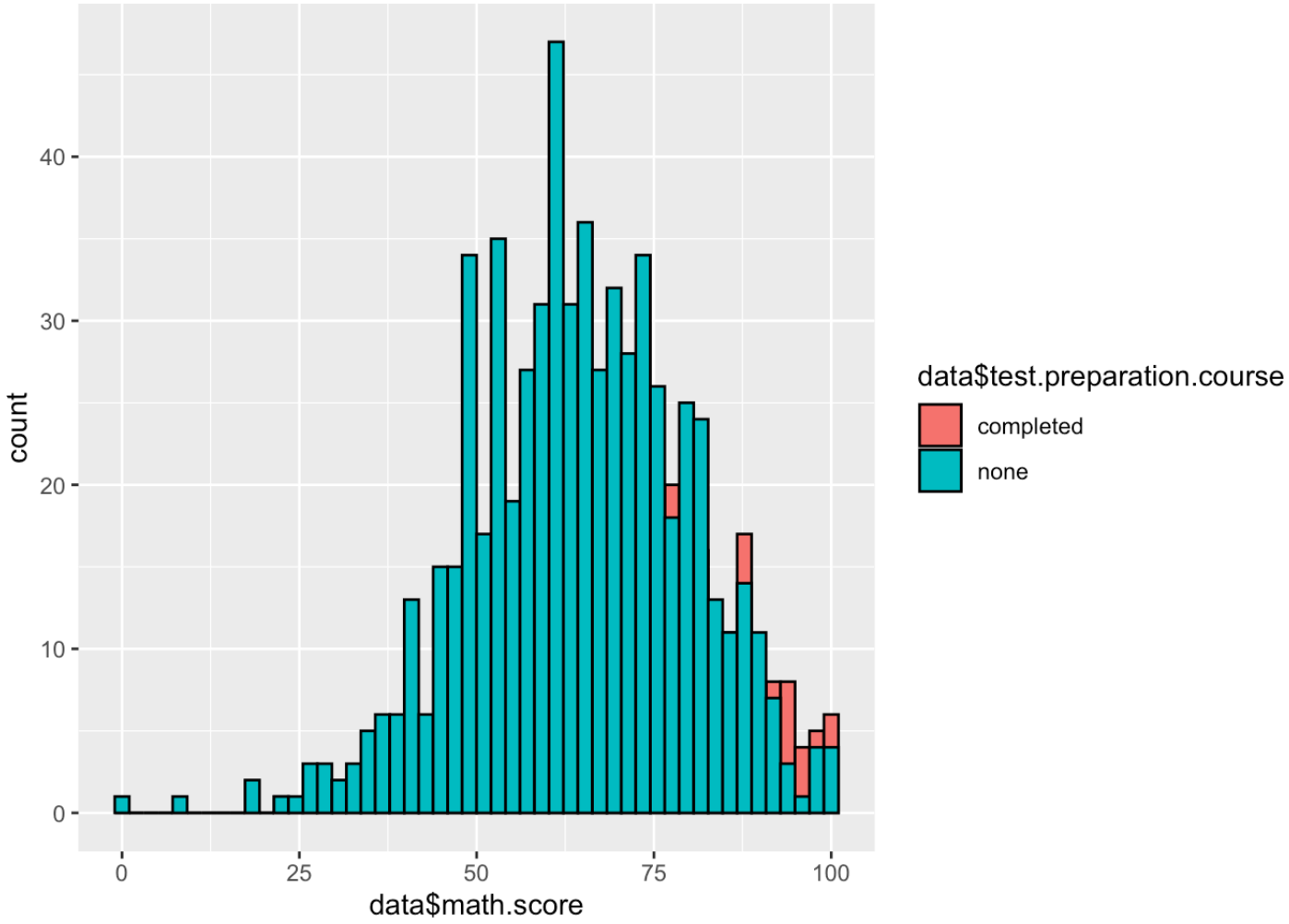
```
## Warning: Use of `data$test.preparation.course` is discouraged.
## ℹ Use `test.preparation.course` instead.
```




```
ggplot(data, aes(x = data$math.score, fill = data$test.preparation.course)) +
  geom_histogram(position = "identity", alpha = 1.2, bins = 50, color="black")
```

```
## Warning: Use of `data$math.score` is discouraged.
## ℹ Use `math.score` instead.
```

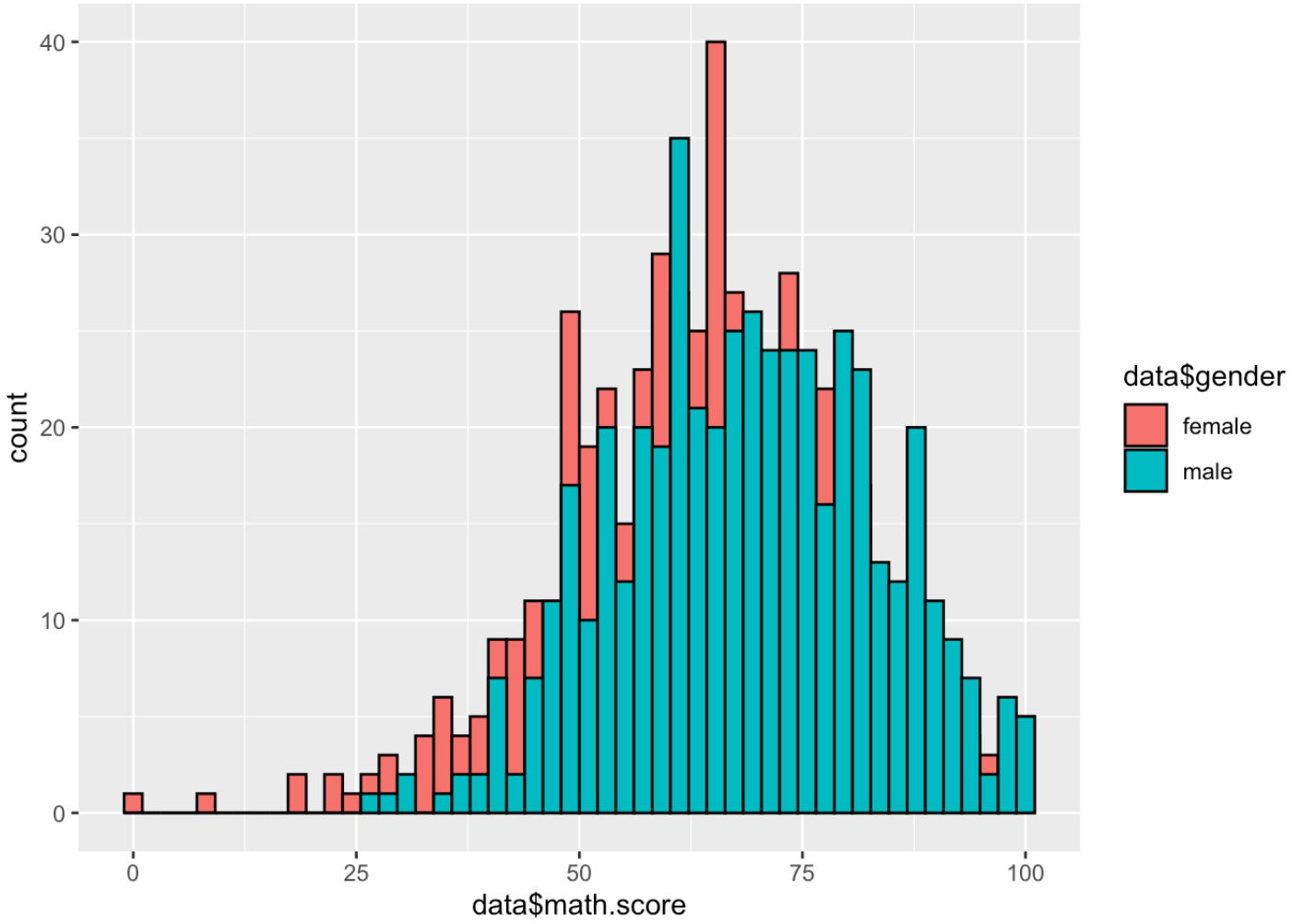
```
## Warning: Use of `data$test.preparation.course` is discouraged.
## ℹ Use `test.preparation.course` instead.
```




```
ggplot(data, aes(x = data$math.score, fill = data$gender)) +
  geom_histogram(position = "identity", alpha = 1.2, bins = 50, color="black")
```

```
## Warning: Use of `data$math.score` is discouraged.
## ℹ Use `math.score` instead.
```

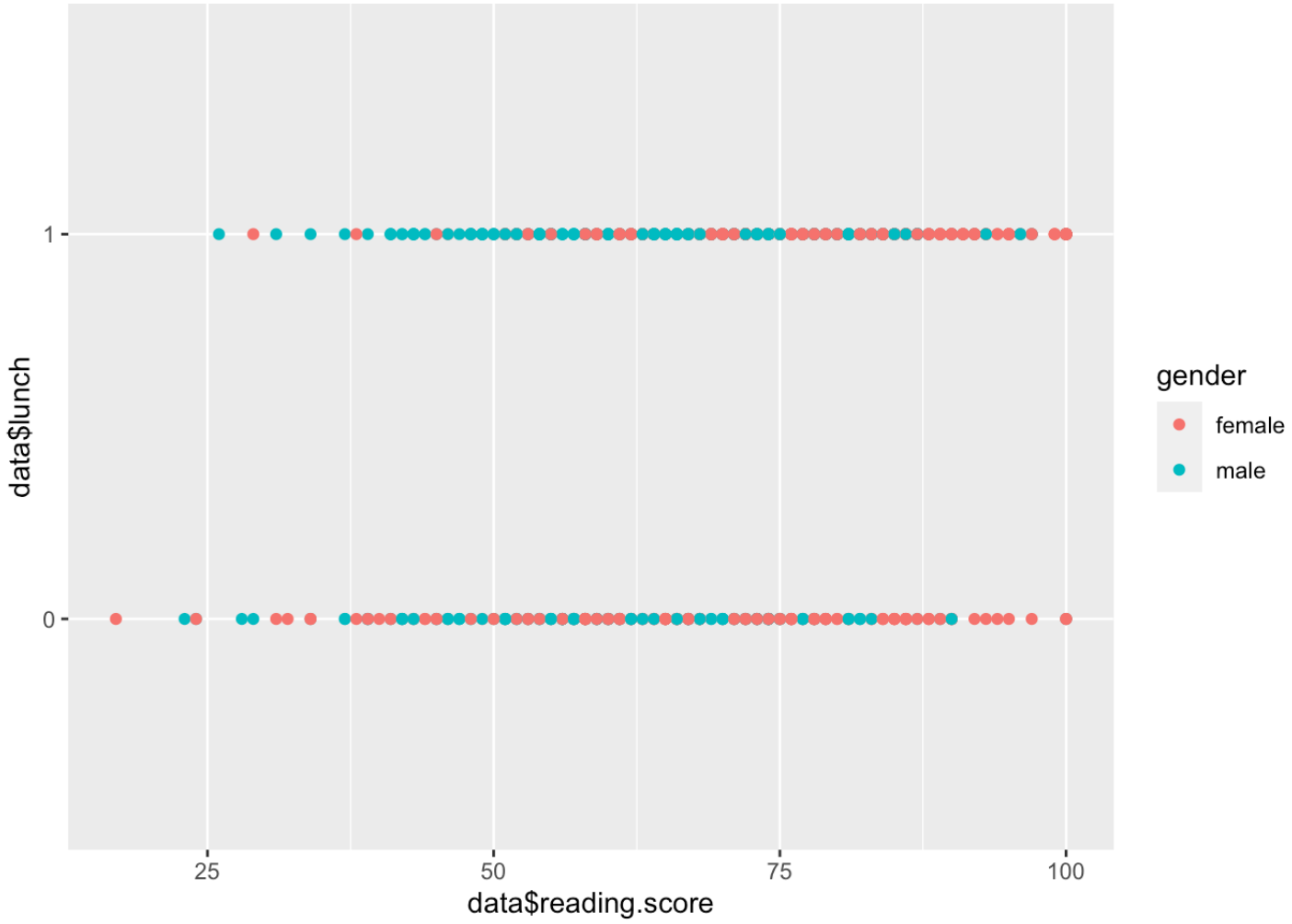
```
## Warning: Use of `data$gender` is discouraged.
## ℹ Use `gender` instead.
```




```
ggplot(data=data) +
 geom_point(mapping = aes(x=data$reading.score, y = data$lunch, colour = `gender`))
```

```
## Warning: Use of `data$reading.score` is discouraged.
## ℹ Use `reading.score` instead.
```

```
## Warning: Use of `data$lunch` is discouraged.
## ℹ Use `lunch` instead.
```




```
t.test(math.score ~ lunch, data= data, alternative = c("two.sided"), var.equal = TRUE
, conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  math.score by lunch
## t = -11.837, df = 998, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is no
t equal to 0
## 95 percent confidence interval:
##  -12.955269  -9.270694
## sample estimates:
## mean in group 0 mean in group 1
##        58.92113        70.03411
```

1 = standard 0= free/reduced

```
t.test(reading.score ~ lunch, data= data, alternative = c("two.sided"), var.equal = T
RUE, conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  reading.score by lunch
## t = -7.4511, df = 998, p-value = 2.003e-13
## alternative hypothesis: true difference in means between group 0 and group 1 is no
t equal to 0
## 95 percent confidence interval:
##  -8.844490 -5.156995
## sample estimates:
## mean in group 0 mean in group 1
##        64.65352        71.65426
```

```
t.test(writing.score ~ lunch, data= data, alternative = c("two.sided"), var.equal = T
RUE, conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  writing.score by lunch
## t = -8.0098, df = 998, p-value = 3.186e-15
## alternative hypothesis: true difference in means between group 0 and group 1 is no
t equal to 0
## 95 percent confidence interval:
##  -9.711845 -5.889596
## sample estimates:
## mean in group 0 mean in group 1
##        63.02254        70.82326
```

```
# Split the data into 80% training and 20% testing sets
set.seed(123)
train_index <- createDataPartition(data$math.score, p = 0.8, list = FALSE)
train_data <- data[train_index, ]
test_data <- data[-train_index, ]
```

```
model <- lm(math.score ~ gender+race.ethnicity +parental.level.of.education +lunch+te
st.preparation.course+reading.score + writing.score,data = data)
summary(model)
```

```
##
## Call:
## lm(formula = math.score ~ gender + race.ethnicity + parental.level.of.education +
##     lunch + test.preparation.course + reading.score + writing.score,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.4995  -3.6824   0.1218   3.3932  14.1178
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                 -11.60449    1.24479  -9.322
## gendermale                                   13.24045    0.37193  35.599
## race.ethnicitygroup B                         0.83537    0.69230   1.207
## race.ethnicitygroup C                         0.17823    0.64899   0.275
## race.ethnicitygroup D                         0.09840    0.67014   0.147
## race.ethnicitygroup E                         5.07770    0.73714   6.888
## parental.level.of.educationbachelor's degree -1.04690    0.61571  -1.700
## parental.level.of.educationhigh school        0.56773    0.53518   1.061
## parental.level.of.educationmaster's degree   -1.85607    0.79324  -2.340
## parental.level.of.educationsome college       0.40026    0.50814   0.788
## parental.level.of.educationsome high school   0.55216    0.54989   1.004
## lunch1                                        3.21271    0.37420   8.585
## test.preparation.coursenone                   3.50227    0.39658   8.831
## reading.score                                 0.26351    0.04205   6.266
## writing.score                                 0.70156    0.04352  16.120
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## gendermale                                   < 2e-16 ***
## race.ethnicitygroup B                         0.2279
## race.ethnicitygroup C                         0.7837
## race.ethnicitygroup D                         0.8833
## race.ethnicitygroup E                       1.00e-11 ***
## parental.level.of.educationbachelor's degree  0.0894 .
## parental.level.of.educationhigh school        0.2890
## parental.level.of.educationmaster's degree    0.0195 *
## parental.level.of.educationsome college       0.4311
## parental.level.of.educationsome high school   0.3156
## lunch1                                       < 2e-16 ***
## test.preparation.coursenone                  < 2e-16 ***
## reading.score                               5.52e-10 ***
## writing.score                                < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.362 on 985 degrees of freedom
```

```
## Multiple R-squared:  0.8767, Adjusted R-squared:  0.875
## F-statistic: 500.3 on 14 and 985 DF,  p-value: < 2.2e-16
```

```
predictions <- predict(model ,newdata = test_data)

RMSE <- sqrt(mean((test_data$math.score - predictions) ^ 2))
RMSE
```

```
## [1] 5.250177
```

```
# Create a null model
intercept_only <- lm(math.score ~ 1, data=data)
# Create a full model
all <- lm(math.score ~., data=data)
# perform forward step-wise regression
forward <- stepAIC (intercept_only, direction='forward',scope = formula(all))
```

```
## Start:  AIC=5438.73
## math.score ~ 1
##
##                              Df Sum of Sq    RSS    AIC
## + reading.score               1    153533  76157 4336.8
## + writing.score               1    147974  81716 4407.2
## + lunch                       1     28278 201411 5309.3
## + race.ethnicity              4     12729 216960 5389.7
## + test.preparation.course     1      7253 222436 5408.6
## + gender                      1      6481 223208 5412.1
## + parental.level.of.education  5      7296 222394 5416.4
## <none>                                    229689 5438.7
##
## Step:  AIC=4336.79
## math.score ~ reading.score
##
##                              Df Sum of Sq   RSS    AIC
## + gender                      1     33031 43126 3770.1
## + lunch                       1      6457 69699 4250.2
## + race.ethnicity              4      3955 72202 4291.5
## + writing.score               1      1274 74883 4321.9
## <none>                                    76157 4336.8
## + test.preparation.course     1        97 76059 4337.5
## + parental.level.of.education  5       407 75749 4341.4
##
## Step:  AIC=3770.12
## math.score ~ reading.score + gender
```

```
##
##                               Df Sum of Sq   RSS    AIC
## + writing.score                1    6519.2 36606 3608.2
## + lunch                        1    4311.7 38814 3666.8
## + race.ethnicity               4    2649.1 40477 3714.7
## + test.preparation.course      1     505.9 42620 3760.3
## <none>                                     43126 3770.1
## + parental.level.of.education  5     302.8 42823 3773.1
##
## Step:  AIC=3608.22
## math.score ~ reading.score + gender + writing.score
##
##                               Df Sum of Sq   RSS    AIC
## + lunch                        1    3223.1 33383 3518.1
## + race.ethnicity               4    2701.8 33905 3539.6
## + test.preparation.course      1    2497.4 34109 3539.6
## + parental.level.of.education  5     485.5 36121 3604.9
## <none>                                     36606 3608.2
##
## Step:  AIC=3518.06
## math.score ~ reading.score + gender + writing.score + lunch
##
##                               Df Sum of Sq   RSS    AIC
## + race.ethnicity               4   2531.69 30852 3447.2
## + test.preparation.course      1   1865.38 31518 3462.6
## <none>                                     33383 3518.1
## + parental.level.of.education  5    312.24 33071 3518.7
##
## Step:  AIC=3447.19
## math.score ~ reading.score + gender + writing.score + lunch +
##     race.ethnicity
##
##                               Df Sum of Sq   RSS    AIC
## + test.preparation.course      1   2091.06 28761 3379.0
## <none>                                     30852 3447.2
## + parental.level.of.education  5    290.27 30561 3447.7
##
## Step:  AIC=3379.01
## math.score ~ reading.score + gender + writing.score + lunch +
##     race.ethnicity + test.preparation.course
##
##                               Df Sum of Sq   RSS    AIC
## + parental.level.of.education  5    441.47 28319 3373.5
## <none>                                     28761 3379.0
##
## Step:  AIC=3373.54
## math.score ~ reading.score + gender + writing.score + lunch +
```

```
##     race.ethnicity + test.preparation.course + parental.level.of.education
```

```
# view results of forward stepwise regression
forward$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## math.score ~ 1
##
## Final Model:
## math.score ~ reading.score + gender + writing.score + lunch +
##     race.ethnicity + test.preparation.course + parental.level.of.education
##
##
##                             Step Df    Deviance Resid. Df Resid. Dev      AIC
## 1                                                     999  229689.08 5438.727
## 2               + reading.score  1 153532.5655       998   76156.51 4336.791
## 3                     + gender  1  33030.8930       997   43125.62 3770.117
## 4               + writing.score  1   6519.1863       996   36606.43 3608.224
## 5                       + lunch  1   3223.0611       995   33383.37 3518.058
## 6              + race.ethnicity  4   2531.6946       991   30851.68 3447.191
## 7     + test.preparation.course  1   2091.0583       990   28760.62 3379.007
## 8 + parental.level.of.education  5    441.4697       985   28319.15 3373.538
```

```
# view final model
summary(forward)
```

```
##
## Call:
## lm(formula = math.score ~ reading.score + gender + writing.score +
##     lunch + race.ethnicity + test.preparation.course + parental.level.of.education
,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.4995  -3.6824   0.1218   3.3932  14.1178
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             -11.60449    1.24479  -9.322
## reading.score                             0.26351    0.04205   6.266
```

```
## gendermale                                      13.24045    0.37193  35.599
## writing.score                                    0.70156    0.04352  16.120
## lunch1                                           3.21271    0.37420   8.585
## race.ethnicitygroup B                            0.83537    0.69230   1.207
## race.ethnicitygroup C                            0.17823    0.64899   0.275
## race.ethnicitygroup D                            0.09840    0.67014   0.147
## race.ethnicitygroup E                            5.07770    0.73714   6.888
## test.preparation.coursenone                      3.50227    0.39658   8.831
## parental.level.of.educationbachelor's degree    -1.04690    0.61571  -1.700
## parental.level.of.educationhigh school           0.56773    0.53518   1.061
## parental.level.of.educationmaster's degree      -1.85607    0.79324  -2.340
## parental.level.of.educationsome college          0.40026    0.50814   0.788
## parental.level.of.educationsome high school      0.55216    0.54989   1.004
##                                                Pr(>|t|)
## (Intercept)                                     < 2e-16 ***
## reading.score                                  5.52e-10 ***
## gendermale                                      < 2e-16 ***
## writing.score                                   < 2e-16 ***
## lunch1                                          < 2e-16 ***
## race.ethnicitygroup B                            0.2279
## race.ethnicitygroup C                            0.7837
## race.ethnicitygroup D                            0.8833
## race.ethnicitygroup E                          1.00e-11 ***
## test.preparation.coursenone                     < 2e-16 ***
## parental.level.of.educationbachelor's degree     0.0894 .
## parental.level.of.educationhigh school           0.2890
## parental.level.of.educationmaster's degree       0.0195 *
## parental.level.of.educationsome college          0.4311
## parental.level.of.educationsome high school      0.3156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.362 on 985 degrees of freedom
## Multiple R-squared:  0.8767, Adjusted R-squared:  0.875
## F-statistic: 500.3 on 14 and 985 DF,  p-value: < 2.2e-16
```

```
backward <- stepAIC (all, direction='backward')
```

```
## Start:  AIC=3373.54
## math.score ~ gender + race.ethnicity + parental.level.of.education +
##     lunch + test.preparation.course + reading.score + writing.score
##
##                              Df Sum of Sq   RSS    AIC
## <none>                                    28319 3373.5
## - parental.level.of.education  5       441 28761 3379.0
## - reading.score                1      1129 29448 3410.6
## - lunch                        1      2119 30438 3443.7
## - test.preparation.course      1      2242 30561 3447.7
## - race.ethnicity               4      2779 31098 3459.1
## - writing.score                1      7471 35790 3605.7
## - gender                       1     36436 64755 4198.6
```

```
backward$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## math.score ~ gender + race.ethnicity + parental.level.of.education +
##     lunch + test.preparation.course + reading.score + writing.score
##
## Final Model:
## math.score ~ gender + race.ethnicity + parental.level.of.education +
##     lunch + test.preparation.course + reading.score + writing.score
##
##
##   Step Df Deviance Resid. Df Resid. Dev      AIC
## 1                       985   28319.15 3373.538
```

```
summary(backward)
```

```
##
## Call:
## lm(formula = math.score ~ gender + race.ethnicity + parental.level.of.education +
##     lunch + test.preparation.course + reading.score + writing.score,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.4995  -3.6824   0.1218   3.3932  14.1178
##
## Coefficients:
```

```
##                                              Estimate Std. Error t value
## (Intercept)                                 -11.60449    1.24479  -9.322
## gendermale                                   13.24045    0.37193  35.599
## race.ethnicitygroup B                         0.83537    0.69230   1.207
## race.ethnicitygroup C                         0.17823    0.64899   0.275
## race.ethnicitygroup D                         0.09840    0.67014   0.147
## race.ethnicitygroup E                         5.07770    0.73714   6.888
## parental.level.of.educationbachelor's degree -1.04690    0.61571  -1.700
## parental.level.of.educationhigh school        0.56773    0.53518   1.061
## parental.level.of.educationmaster's degree   -1.85607    0.79324  -2.340
## parental.level.of.educationsome college       0.40026    0.50814   0.788
## parental.level.of.educationsome high school   0.55216    0.54989   1.004
## lunch1                                        3.21271    0.37420   8.585
## test.preparation.coursenone                   3.50227    0.39658   8.831
## reading.score                                 0.26351    0.04205   6.266
## writing.score                                 0.70156    0.04352  16.120
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## gendermale                                   < 2e-16 ***
## race.ethnicitygroup B                         0.2279
## race.ethnicitygroup C                         0.7837
## race.ethnicitygroup D                         0.8833
## race.ethnicitygroup E                       1.00e-11 ***
## parental.level.of.educationbachelor's degree  0.0894 .
## parental.level.of.educationhigh school        0.2890
## parental.level.of.educationmaster's degree    0.0195 *
## parental.level.of.educationsome college       0.4311
## parental.level.of.educationsome high school   0.3156
## lunch1                                       < 2e-16 ***
## test.preparation.coursenone                  < 2e-16 ***
## reading.score                               5.52e-10 ***
## writing.score                                < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.362 on 985 degrees of freedom
## Multiple R-squared:  0.8767, Adjusted R-squared:  0.875
## F-statistic: 500.3 on 14 and 985 DF,  p-value: < 2.2e-16
```