Code

Read and Clean Data

```
data <- read.csv("./data.csv") |>
  janitor::clean_names() |>
  mutate(
    gender = case_when(
      gender == "male" ~ 0,
      gender == "female" ~ 1,
      ),
    ethnic_group = case_when(
      ethnic_group == "group A" ~ 0,
      ethnic_group == "group B" ~ 1,
      ethnic_group == "group C" ~ 2,
      ethnic_group == "group D" ~ 3,
      ethnic_group == "group E" ~ 4,
      ),
    parent_educ = case_when(
      parent_educ == "some highschool" ~ 0,
      parent_educ == "some college" ~ 1,
      parent_educ == "associate's degree" ~ 2,
      parent_educ == "bachelor's degree" ~ 3,
      parent_educ == "master's degree" ~ 4,
      ),
    lunch_type = case_when(
      lunch_type == "standard" ~ 0,
      lunch_type == "free/reduced" ~ 1,
      ),
    test_prep = case_when(
      test_prep == "none" ~ 0,
      test_prep == "completed" ~ 1,
      ),
    parent_marital_status = case_when(
      parent_marital_status == "married" ~ 0,
      parent_marital_status == "single" ~ 1,
      parent_marital_status == "widowed" ~ 2,
      parent_marital_status == "divorced" ~ 3,
      ),
    practice_sport = case_when(
      practice_sport == "never" ~ 0,
      practice_sport == "sometimes" ~ 1,
      practice_sport == "regularly" ~ 2,
      ),
    is_first_child = case_when(
      is_first_child == "no" ~ 0,
```

```r
      is_first_child == "yes" ~ 1,
      ),
    transport_means = case_when(
      transport_means == "school_bus" ~ 0,
      transport_means == "private" ~ 1,
      ),
    wkly_study_hours = case_when(
      wkly_study_hours == "< 5" ~ 0,
      wkly_study_hours == "10-May" ~ 1,
      wkly_study_hours == "> 10" ~ 2,
      )
    )

# Deal with NA -- Calculate the column mean (round to integer) and plug it into NA cell
column_means <- round(colMeans(data, na.rm = TRUE), digits = 0)
for (col in names(data)) {
  data[[col]][is.na(data[[col]])] <- column_means[col]
  }

head(data)
```

```
##   gender ethnic_group parent_educ lunch_type test_prep parent_marital_status
## 1      1            2           3          0         0                     0
## 2      1            2           1          0         0                     0
## 3      1            1           4          0         0                     1
## 4      0            0           2          1         0                     0
## 5      0            2           1          0         0                     0
## 6      1            1           2          0         0                     0
##   practice_sport is_first_child nr_siblings transport_means wkly_study_hours
## 1              2              1           3               0                0
## 2              1              1           0               0                1
## 3              1              1           4               0                0
## 4              0              0           1               0                1
## 5              1              1           0               0                1
## 6              2              1           1               0                1
##   math_score reading_score writing_score
## 1         71            71            74
## 2         69            90            88
## 3         87            93            91
## 4         45            56            42
## 5         76            78            75
## 6         73            84            79
```

```r
# Another data set for EDA
data_long <- data |>
  pivot_longer(cols = c(math_score, reading_score, writing_score),
               names_to = "test", values_to = "score")
```

Summary

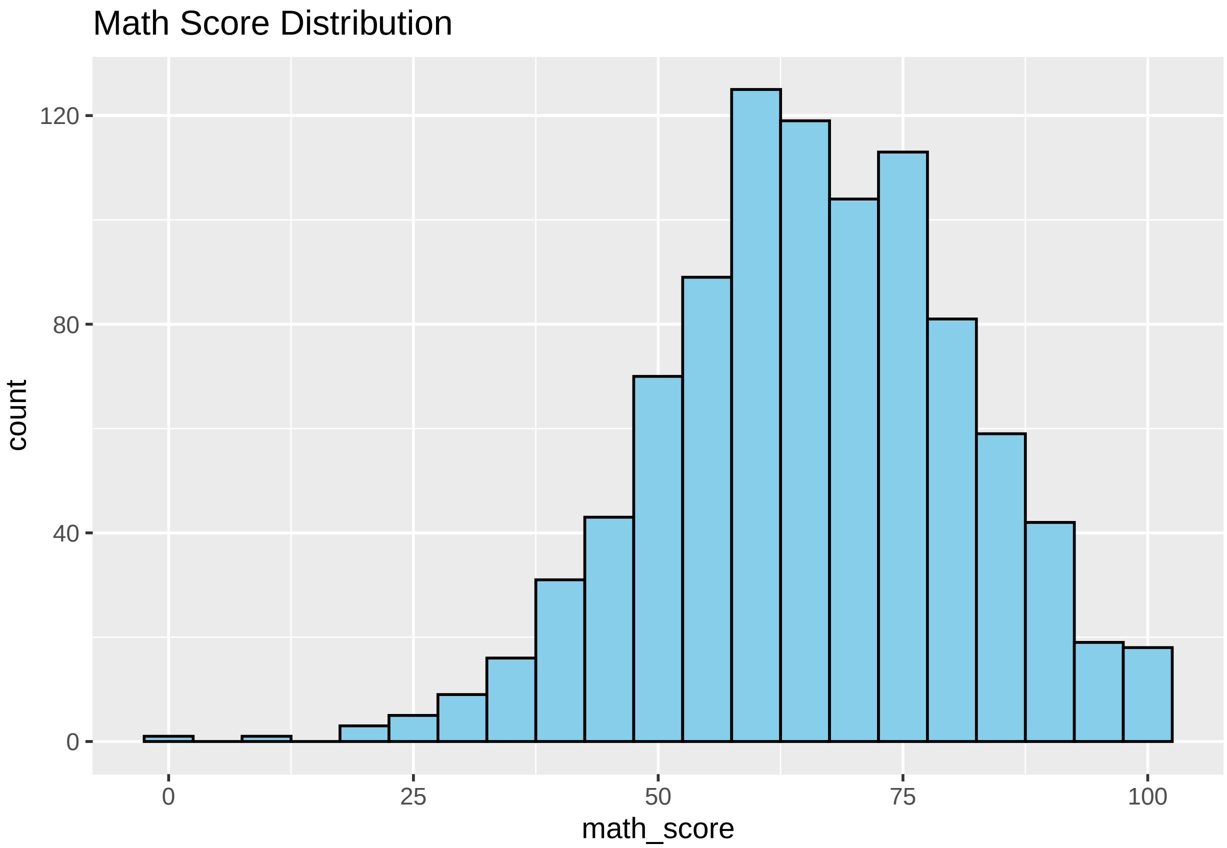
```r
transposed_summary <- t(summary(data))
knitr::kable(transposed_summary, caption = "Summary Statistics for Data", 2)
```

Table 1: Summary Statistics for Data

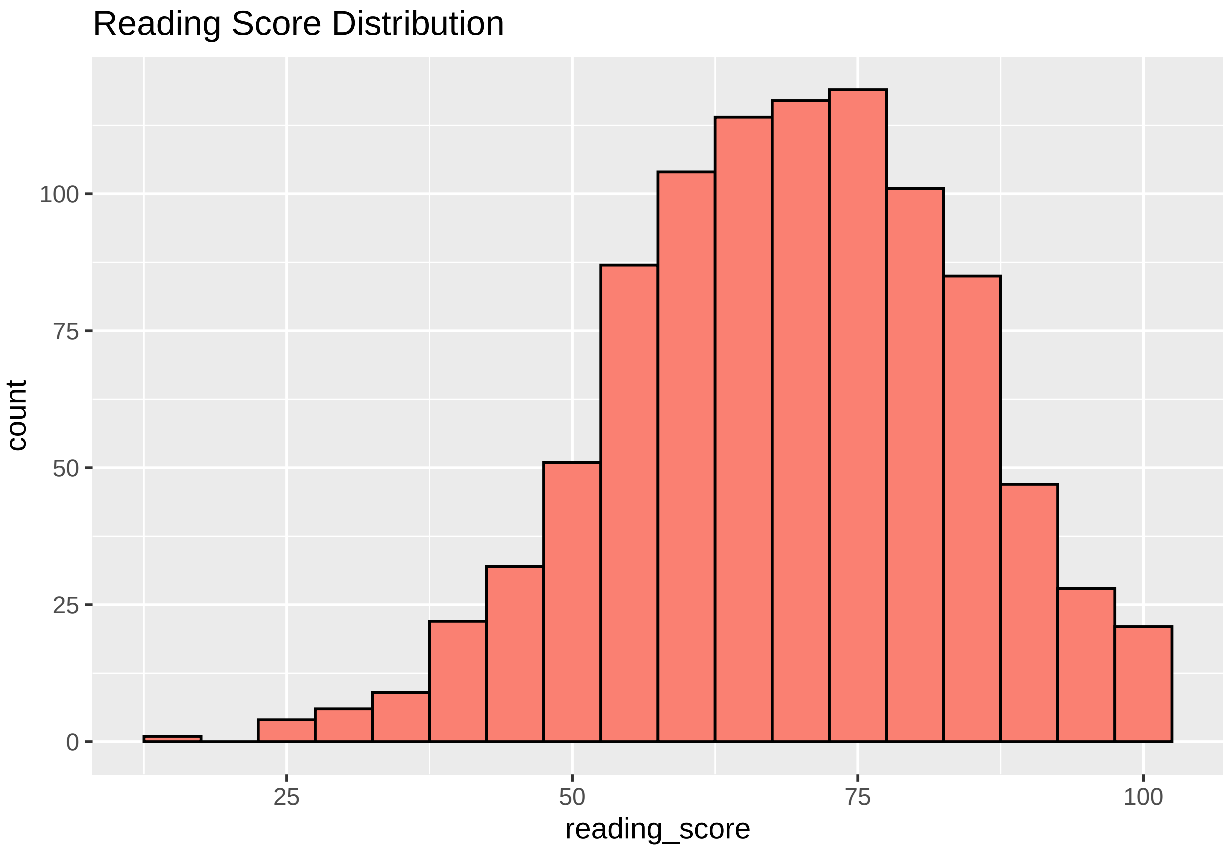
gender	Min. :0.0000	1st Qu.:0.0000	Median :1.0000	Mean :0.5148	3rd Qu.:1.0000	Max. :1.0000
ethnic_group	Min. :0.000	1st Qu.:1.000	Median :2.000	Mean :2.162	3rd Qu.:3.000	Max. :4.000
parent_educ	Min. :1.000	1st Qu.:2.000	Median :2.000	Mean :2.016	3rd Qu.:2.000	Max. :4.000
lunch_type	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.3492	3rd Qu.:1.0000	Max. :1.0000
test_prep	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.3397	3rd Qu.:1.0000	Max. :1.0000
parent_marital_status	Min. :0.000	1st Qu.:0.000	Median :0.000	Mean :0.789	3rd Qu.:1.000	Max. :3.000
practice_sport	Min. :0.000	1st Qu.:1.000	Median :1.000	Mean :1.244	3rd Qu.:2.000	Max. :2.000
is_first_child	Min. :0.0000	1st Qu.:0.0000	Median :1.0000	Mean :0.6688	3rd Qu.:1.0000	Max. :1.0000
nr_siblings	Min. :0.000	1st Qu.:1.000	Median :2.000	Mean :2.148	3rd Qu.:3.000	Max. :7.000
transport_means	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.3555	3rd Qu.:1.0000	Max. :1.0000
wkly_study_hours	Min. :0.0000	1st Qu.:0.0000	Median :1.0000	Mean :0.8914	3rd Qu.:1.0000	Max. :2.0000
math_score	Min. : 0.00	1st Qu.: 56.00	Median : 66.00	Mean : 65.98	3rd Qu.: 76.00	Max. :100.00
reading_score	Min. : 17.00	1st Qu.: 59.00	Median : 69.50	Mean : 68.84	3rd Qu.: 80.00	Max. :100.00
writing_score	Min. : 10.00	1st Qu.: 57.00	Median : 68.00	Mean : 67.93	3rd Qu.: 78.25	Max. :100.00

Histograms

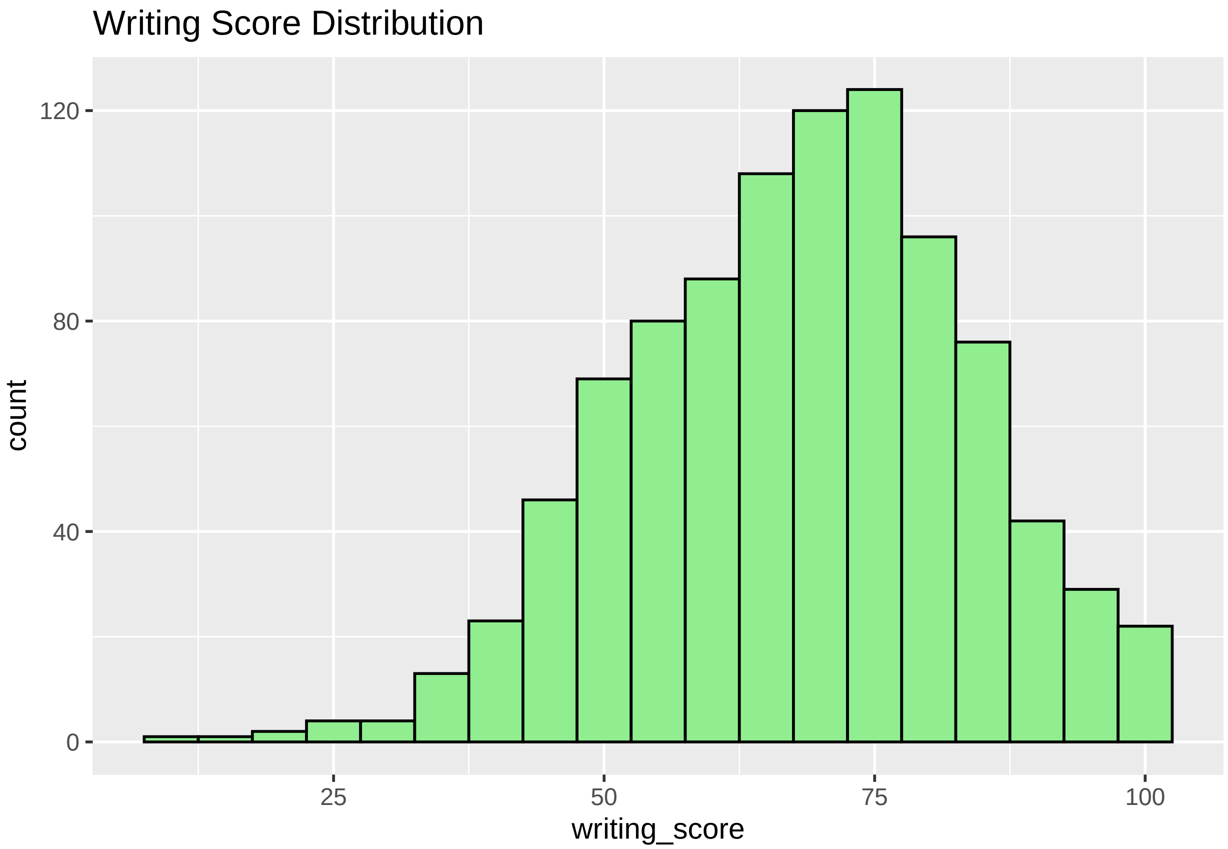
```r
ggplot(data, aes(x = math_score)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  labs(title = "Math Score Distribution")
```



```r
ggplot(data, aes(x = reading_score)) +
  geom_histogram(binwidth = 5, fill = "salmon", color = "black") +
  labs(title = "Reading Score Distribution")
```

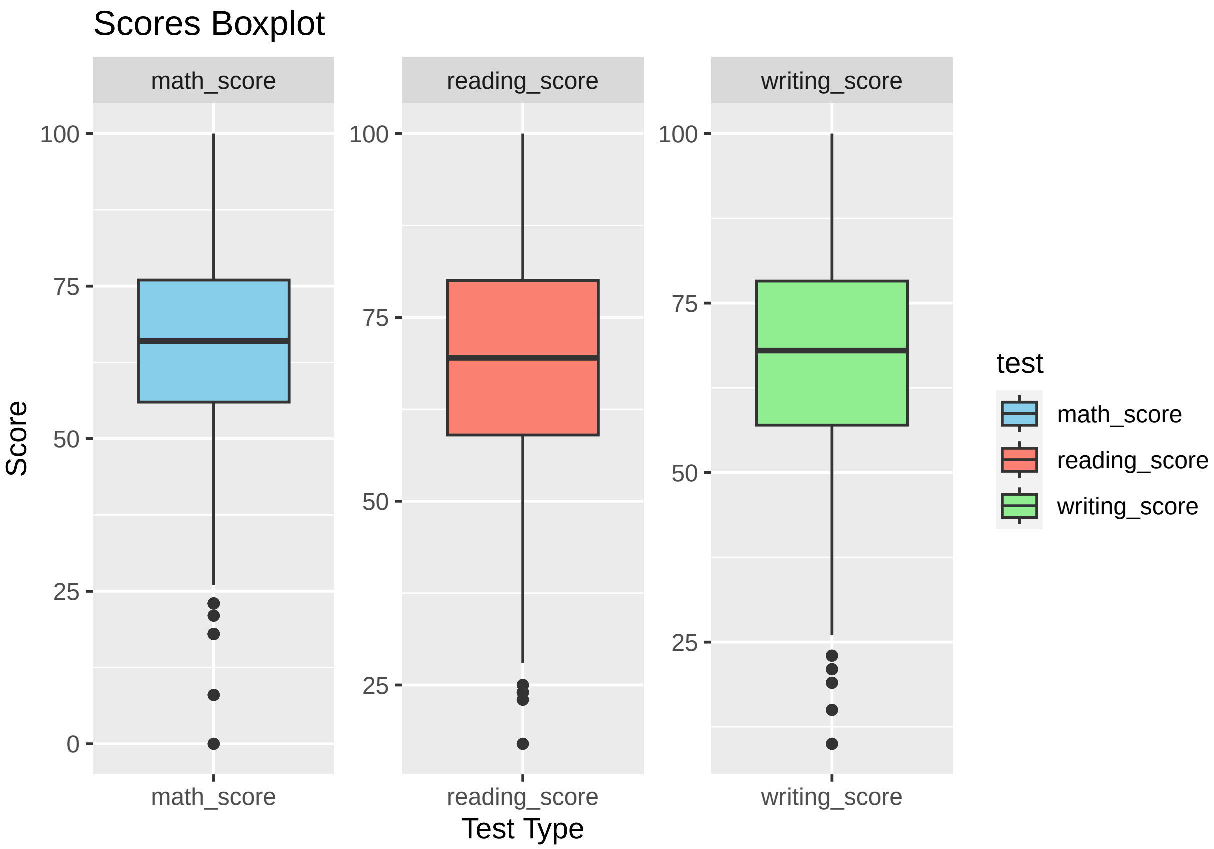


```r
ggplot(data, aes(x = writing_score)) +
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "black") +
  labs(title = "Writing Score Distribution")
```



Boxplots

```r
ggplot(data_long, aes(x = test, y = score, fill = test)) +
  geom_boxplot() +
  labs(title = "Scores Boxplot", x = "Test Type", y = "Score") +
  facet_wrap(~ test, scales = "free") +
  scale_fill_manual(values = c("skyblue", "salmon", "lightgreen"))
```


Scores Boxplot
math_score
reading_score
writing_score
100
75
50
25
0
Score
Test Type
test
math_score
reading_score
writing_score