Explonatory data analysis

none

12/5/2021

Dataset

This data set total consist of 1000 rows with 8 columns. A short description of the data is.

```
df %>%
  summary()
```

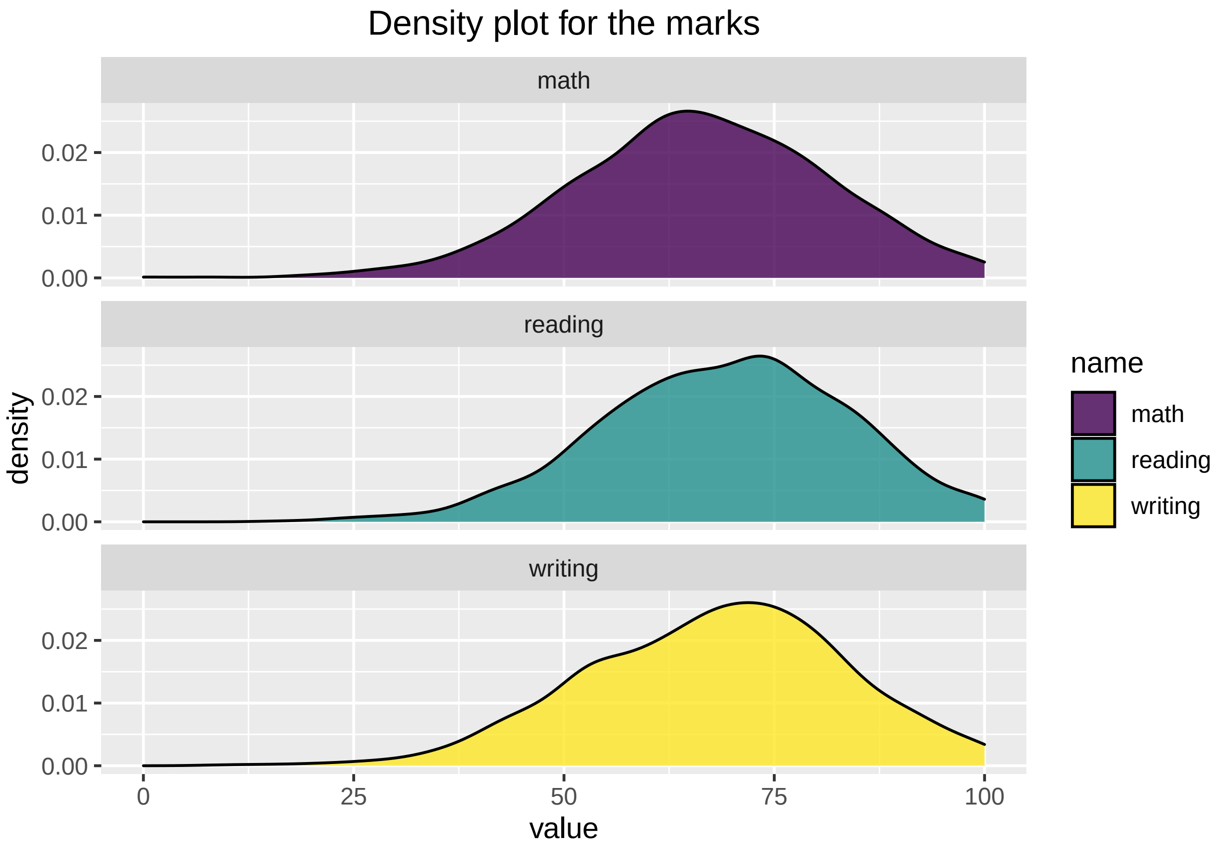
```
##     gender              race           parental level of education
##  Length:1000        Length:1000        Length:1000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##     lunch           test preparation course   math score     reading score
##  Length:1000        Length:1000             Min.   :  0.00   Min.   : 17.00
##  Class :character   Class :character        1st Qu.: 57.00   1st Qu.: 59.00
##  Mode  :character   Mode  :character        Median : 66.00   Median : 70.00
##                                             Mean   : 66.09   Mean   : 69.17
##                                             3rd Qu.: 77.00   3rd Qu.: 79.00
##                                             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
```

The source of the data is the “kaggle” website.

Distribution of the subjective scores

We will show the math reading and writing score. Usually if a distribution followes normal distribution then it think to be a perfect score.

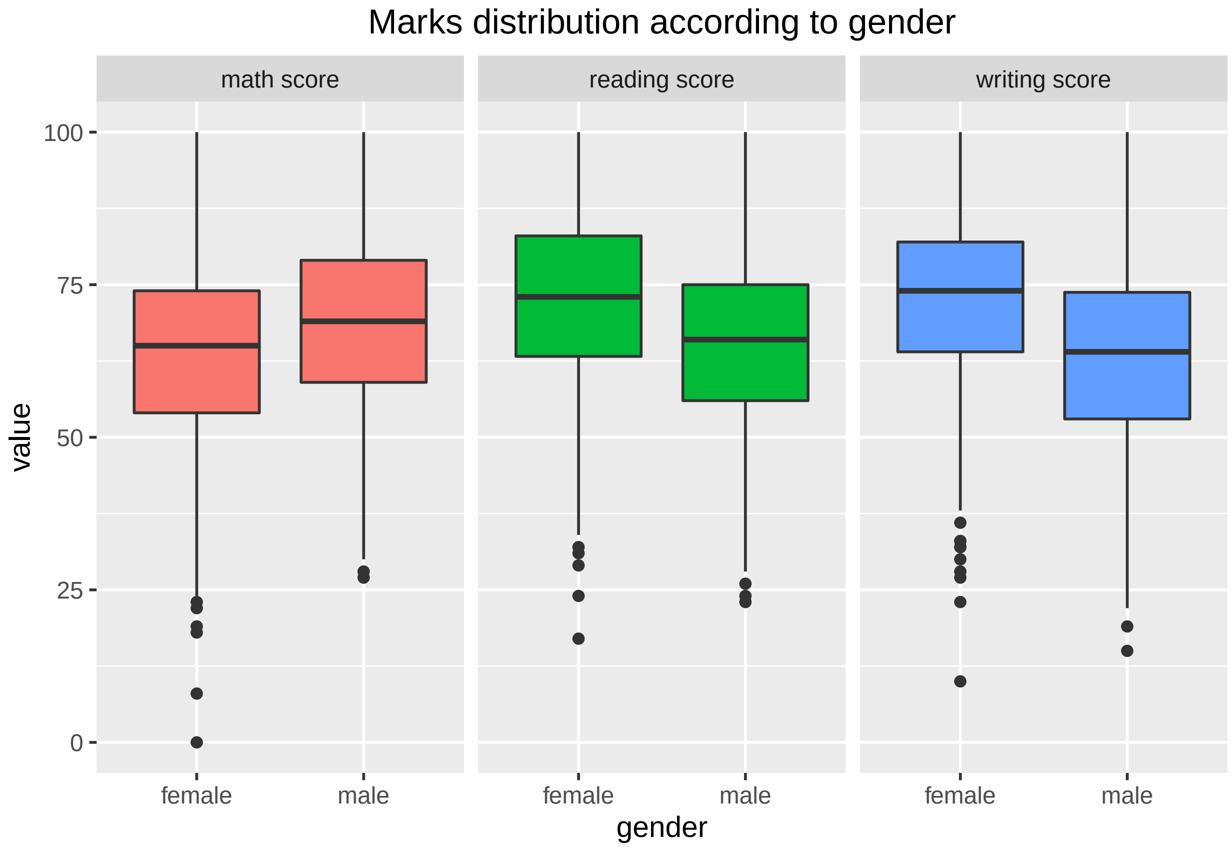
```
df %>%
    select(is.numeric) %>%
    pivot_longer(everything()) %>%
    mutate(name = str_extract(name, "^\\w+")) %>%
    ggplot(aes(value, fill = name)) +
    geom_density(alpha = .8) +
    facet_wrap(~name, nrow = 3) +
    scale_fill_viridis_d() +
    labs(title = "Density plot for the marks")
```



We can see that the graphs are roughly symmetric and also implies that the exams are unbiased (in favor of the good students, moderate and backbenchers) as the marks follow a symmetric shape.

Distribution of scores according to gender

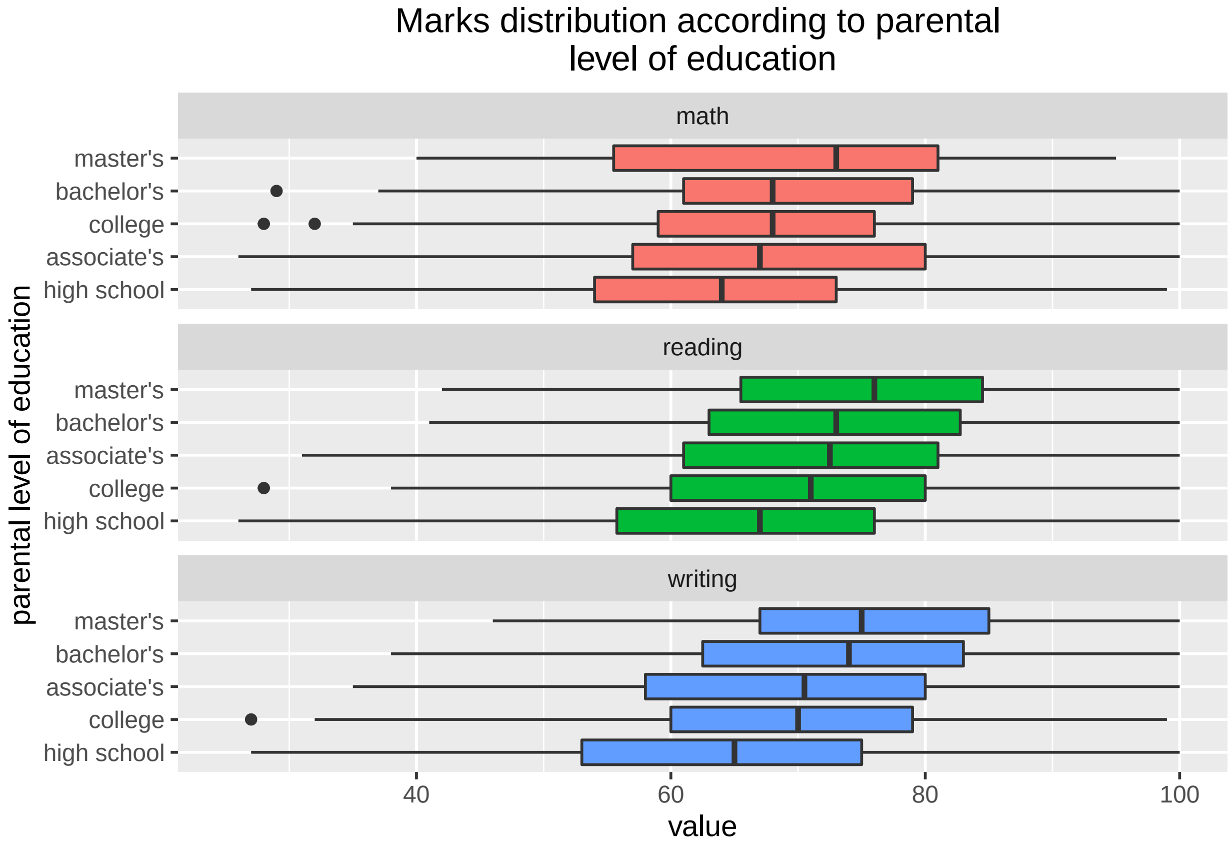
```
df %>%
    select(gender, is.numeric) %>%
    pivot_longer(-gender) %>%
    ggplot(aes(value, gender, fill = name)) +
    geom_boxplot(show.legend = F) +
    facet_wrap(~name) +
    coord_flip() +
    labs(title = "Marks distribution according to gender")
```



So, from this graph we can see that a male group has a tendency to do better in math and female group has a higher tendency to score more in reading and writing. We will perform statistical test for further evidence.

Distribution of scores according to parental level of education.

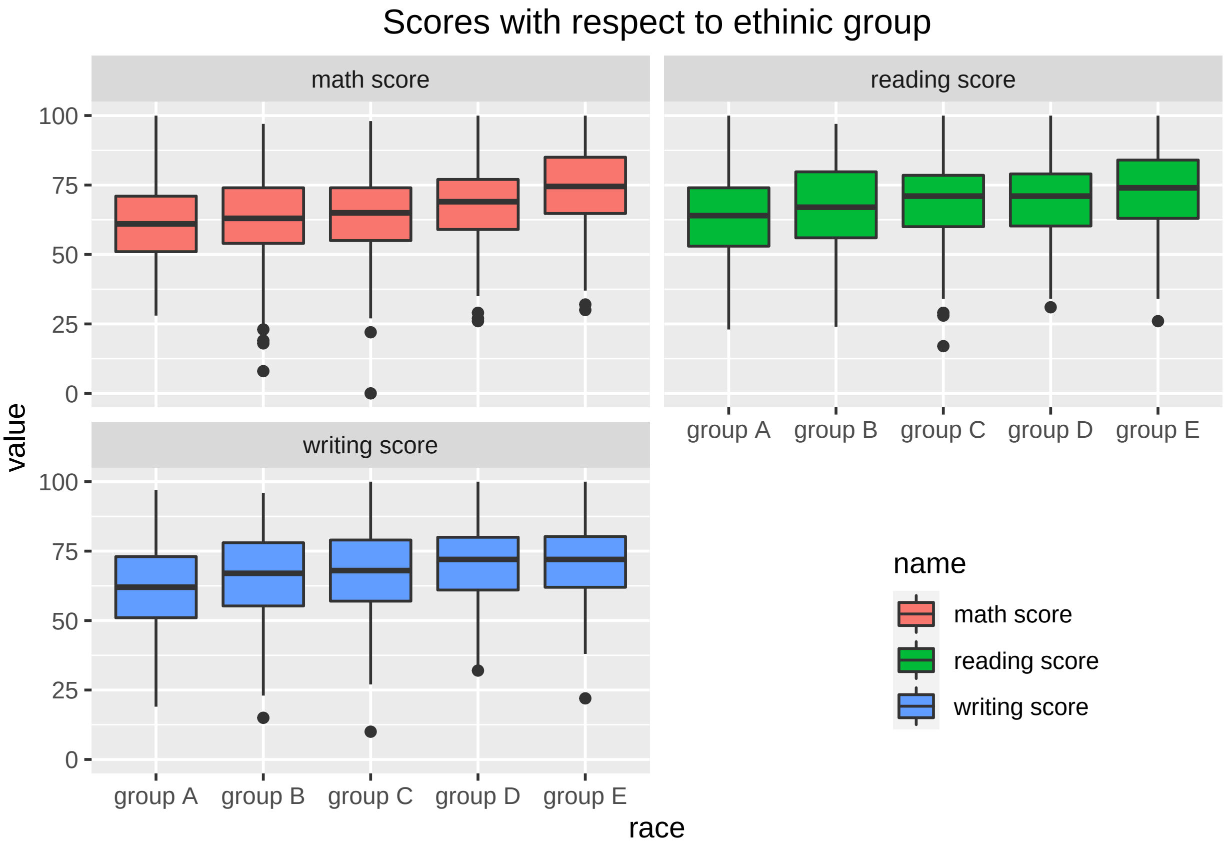
```
df %>%
    select(`parental level of education`, is.numeric) %>%
    pivot_longer(-`parental level of education`) %>%
    # group_by(`parental level of education`) %>%
    mutate(name = str_extract(name, "^\\w+"),
           `parental level of education` =
               str_remove(`parental level of education`, "degree|some"),
           `parental level of education` = str_trim(`parental level of education`),

           ) %>%
    nest(value) %>%
    mutate(by = map_dbl(data, ~median(pull(.x))),
           `parental level of education` =
               tidytext::reorder_within(`parental level of education`,
                                        by = by , within = name)
           ) %>%
    unnest() %>%

    ggplot(aes(`parental level of education`, value, fill = name)) +
    geom_boxplot(show.legend = F) +
    facet_wrap(~name, nrow = 3, scales = "free_y") +
    coord_flip() +
    tidytext::scale_x_reordered() +
    lims(y = c(25,100)) +
    labs(title = "Marks distribution according to parental \nlevel of education")
```



Those graphs don't show that parietal education level plays an important role for the children scores. If parents have a higher degree then the score would also higher indifference of the subjects.

Do ethinic groups play any role?

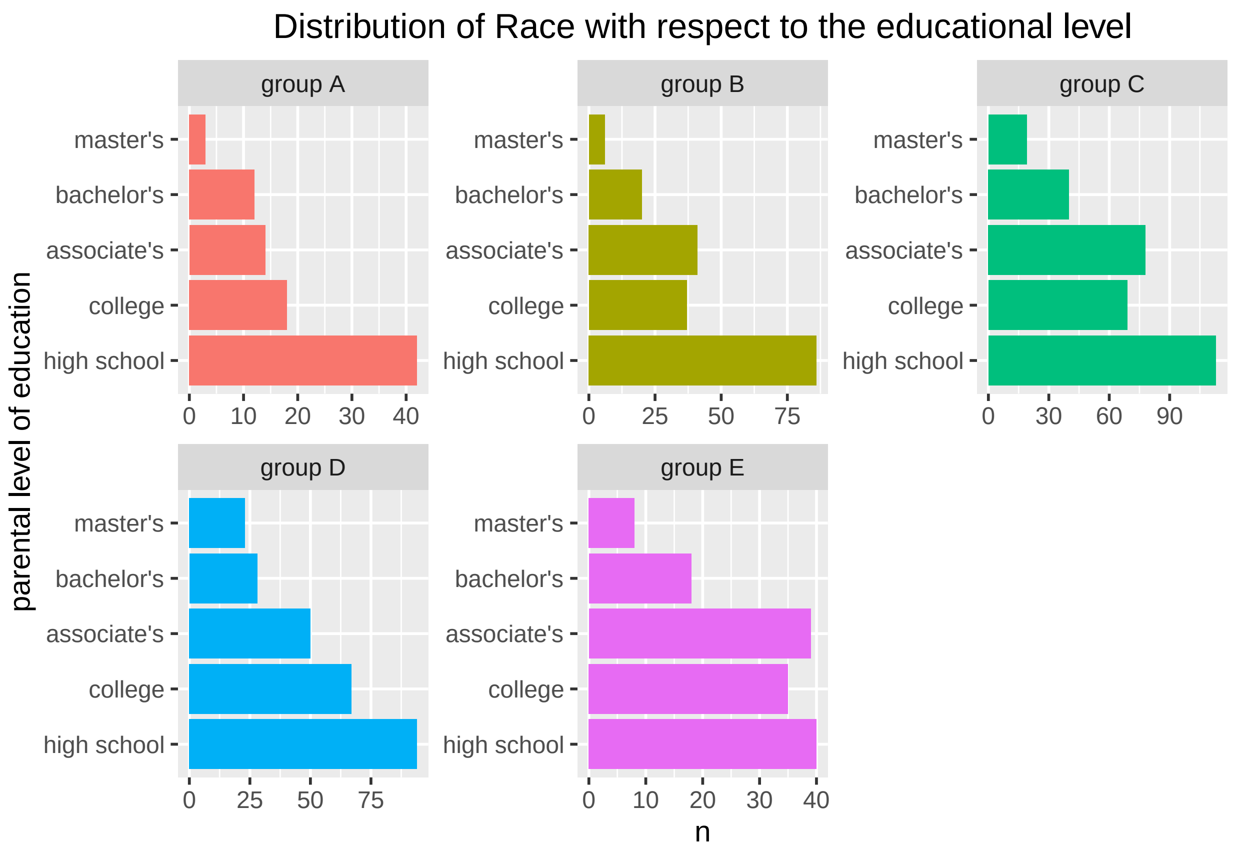
```
df %>%
    select(race, is.numeric) %>%
    pivot_longer(-race) %>%
    ggplot(aes(race, value, fill = name)) +
    geom_boxplot() +
    facet_wrap(~name, ncol = 2) +
    theme(legend.position = c(.8,.2)) +
    labs(title = "Scores with respect to ethinic group")
```



The students belongs to group E on average perform better than other race/ ethnic groups. But there may be some paradox. It my possible that the people from group 5 may be more educate than other ethnic group. We will try to find the graphical solution for that problem in the next graph.

Distribution of Race with respect to the educational level

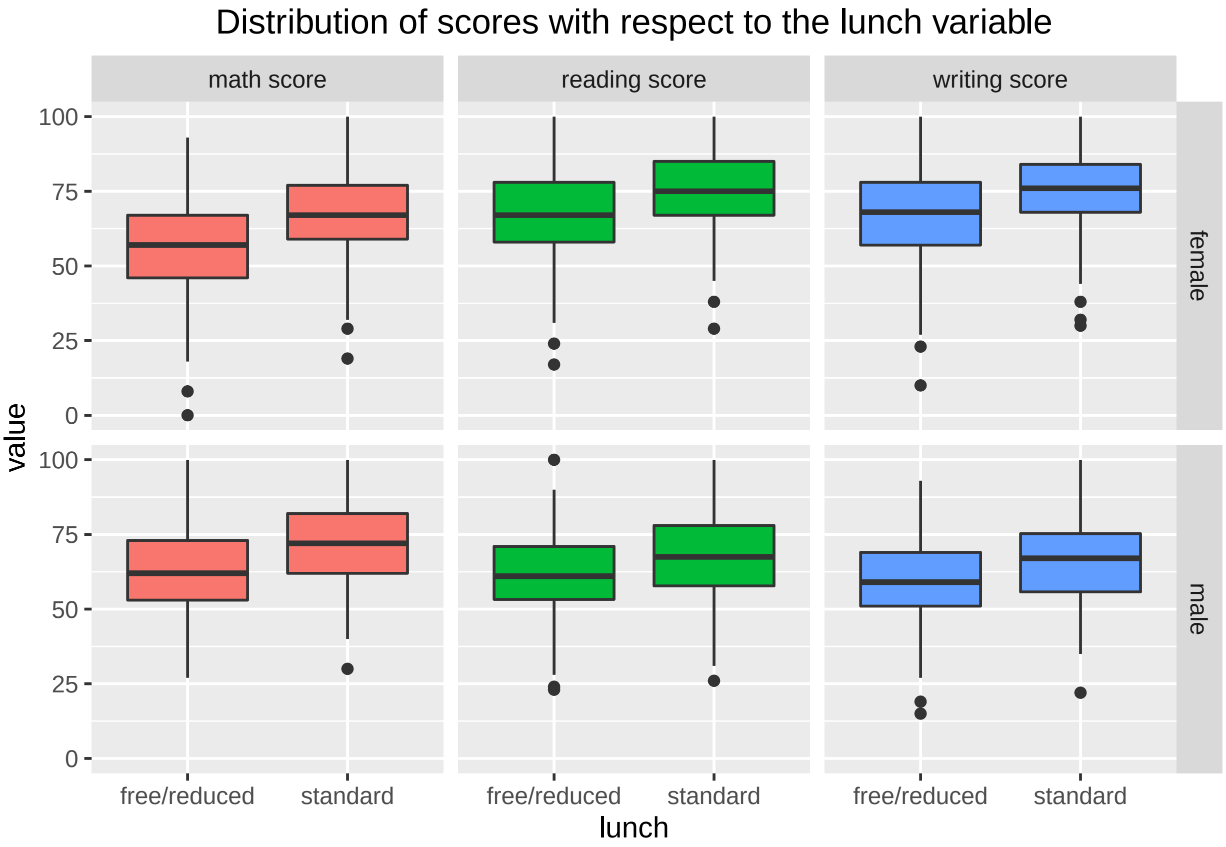
```
df %>%
    select(`parental level of education`, race) %>%
    count(`parental level of education`, race) %>%
    mutate(`parental level of education` =
               str_remove(`parental level of education`, "degree|some"),
           `parental level of education` = str_trim(`parental level of education`),
           `parental level of education` =
               factor(`parental level of education`,
                      c("high school","college","associate's","bachelor's","master's"))
           ) %>%
    ggplot(aes(`parental level of education`, n, fill = race)) +
    geom_col(show.legend = F) +
    facet_wrap(~race, scales = "free") +
    coord_flip() +
    labs(title = "Distribution of Race with respect to the educational level")
```



From this graph we can say that our initial guess about the more proportion of people from group E have higher education is true. The proportion of having master's degree is higher that any other group.

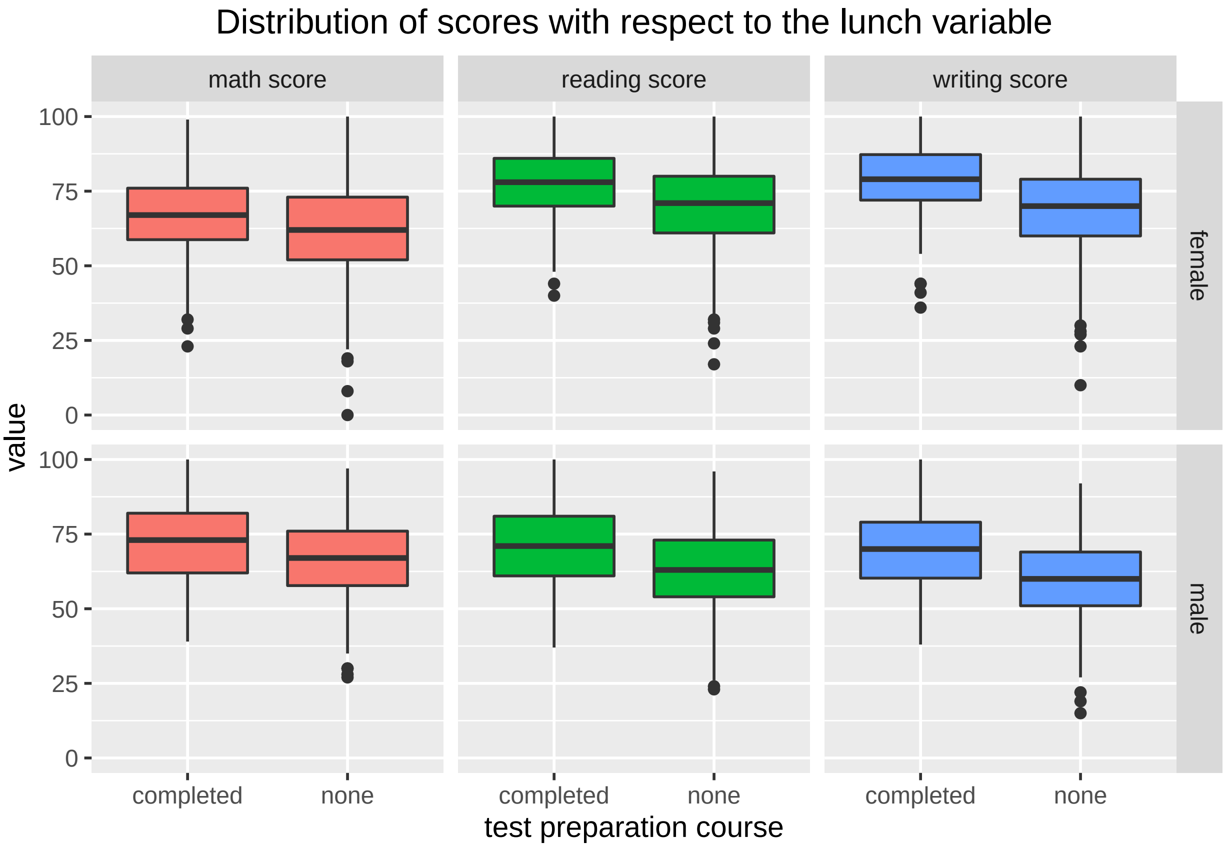
Distribution of scores with respect to the lunch variable

```
df %>%
    select(gender, lunch, is.numeric) %>%
    pivot_longer(-c(lunch,gender)) %>%
    ggplot(aes(lunch, value, fill = name)) +
    geom_boxplot(show.legend = F) +
    facet_grid(gender~name) +
    labs(title = "Distribution of scores with respect to the lunch variable")
```



From thsi graph we can see that, indifference of the male and the female group the standard lunch taking group has the advantage over the free/ reduced lunch student group.

```
df %>%
    select(`test preparation course`, gender, is.numeric) %>%
    pivot_longer(-c(`test preparation course`,gender)) %>%
    ggplot(aes(`test preparation course`, value, fill = name)) +
    geom_boxplot(show.legend = F) +
    facet_grid(gender~name) +
    labs(title = "Distribution of scores with respect to the lunch variable")
```



From this graph we can see that indifference of the male or female group both have an higher average number who have complected the test preparation course that the group who have not taken any kind of that course. But we thing this course boost uo the score of the female group much drastic way that the male group. we will check the statistical valitity.

```
options(scipen = -999, digits = 4)
df %>%
    select(gender, `test preparation course`, is.numeric) %>%
    pivot_longer(3:5) %>%
    nest(-gender) %>%
    mutate(model = map(
        data,
        ~ ( .x %>%
                mutate(
                    `test preparation course` = as.factor(`test preparation course`),
                    `test preparation course` = relevel(`test preparation course`,
                                                        ref = "none")
                ) %>%
                lm(formula = value ~ .) %>%
                summary()
            )$coefficients %>%
            as.data.frame() %>%
            rownames_to_column())
        ) %>%
    select(-data) %>%
    unnest() %>%
    filter(str_detect(rowname, "test")) %>%
    select(-c(4:5)) %>%
    pander::pander()
```

gender	rowname	Estimate	Pr(>|t|)
female	`test preparation course`completed	7.576e+00	2.302e-22
male	`test preparation course`completed	7.737e+00	2.65e-24

Here both p.value for the male and female group is less than .01 so they are statistically significant. However the magnitude is more or less same. So, taking some test preparation course will affect the male and female group in an equal way.