MSMS - 105

Ananda Biswas

Assignment 01

❖ **Task :** Collect a real data set belongs to your nearby. The sample size must be more than 20 with at least 4 different variables. Give the inference for this data using basic descriptive statistics and EDA approach.

➔ ***Data Description*** : A data-set has been created with help of the information obtained from students of Semester 1 of Statistics and Computing of DST-CIMS, BHU. A brief description of the data-set is as follows :

gender : gender of the student;

home_state : home state of the student;

CUET_score : score of the student in CUET PG Statistics 2024;

appeared_in_JAM : 1 if the student had appeared in JAM MS 2024, 0 otherwise;

JAM_score : score of an appearing student in JAM MS 2024;

coaching : 1 if the student had enrolled in any coaching institute for preparation of aforesaid examinations, 0 otherwise;

UG_CGPA : CGPA of the student in his/her undergraduate program;

UG_University_State : state of the university from where the student has completed his/her undergraduate program.

```
dim(raw_data)

## [1] 44  8
```

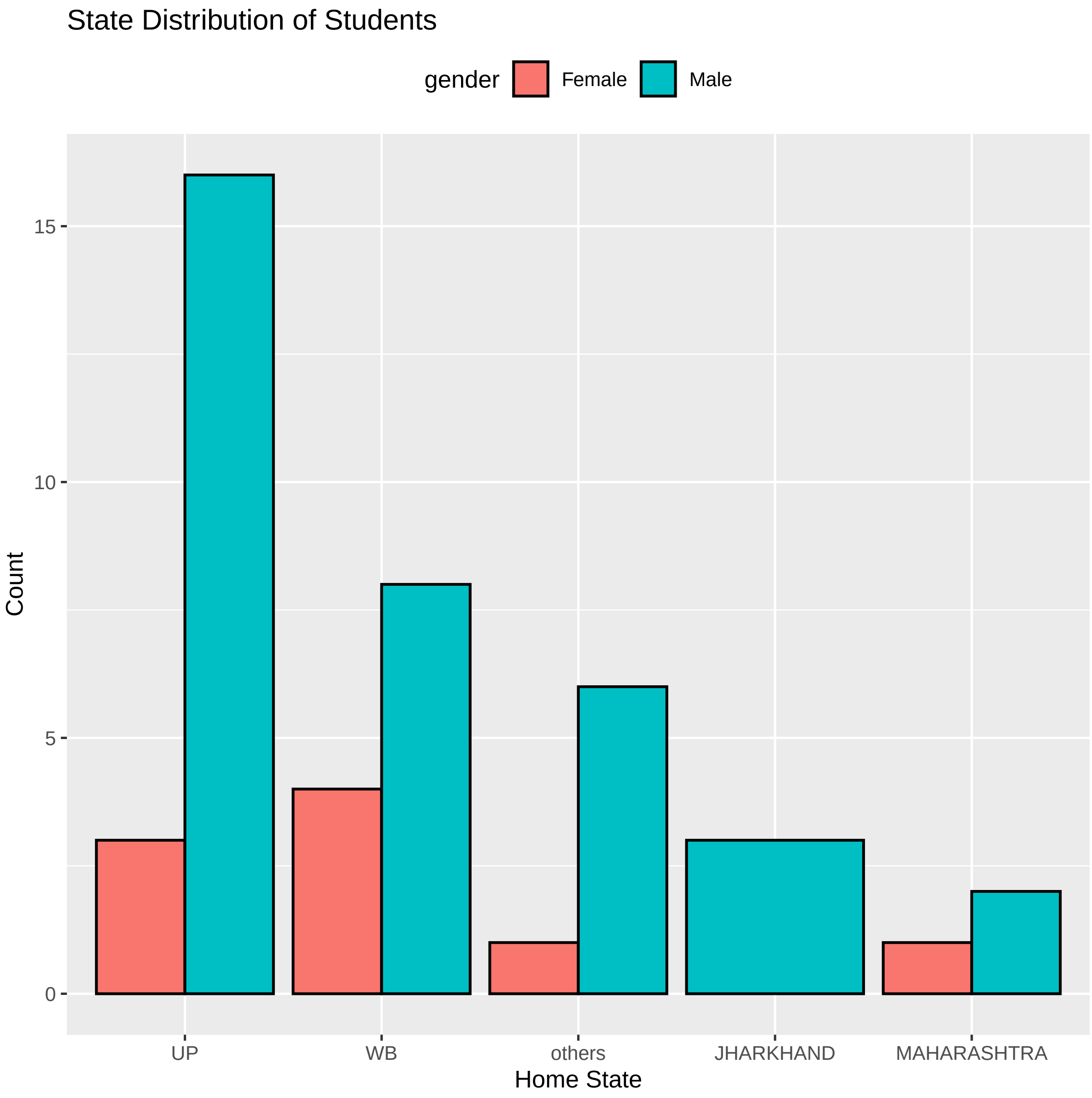
There are records of 44 students of the 8 variables as mentioned above.

```
names(raw_data)

## [1] "gender"              "home_state"          "CUET_score"
## [4] "appeared_in_JAM"     "JAM_score"           "coaching"
## [7] "UG_CGPA"             "UG_university_state"
```

Let us have a look how different states are represented by students grouped by gender.

```
home_state_and_gender %>%
  ggplot(aes(x = fct_infreq(home.state), fill = gender)) +
  geom_bar(position = "dodge", col = "black", linewidth = 0.6) +
  labs(x = "Home State", y = "Count",
       title = "State Distribution of Students") +
  theme(legend.position = "top")
```



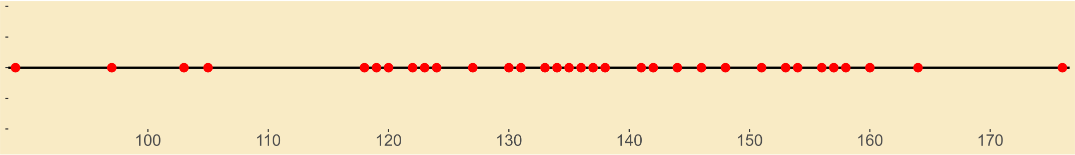
```
table(raw_data$gender)

##
## Female   Male
##      9     35
```

✎ Our data have 9 female students and 35 male students.

☞ *CUET_score*

Let us plot the values of ***CUET_score*** along the real line.



- Measure of Central Tendency : Mean CUET score of the students is 137.6136364.

- Measure of Dispersion : CUET score has a standard deviation of 18.9088584.

- Quartiles : The following are the quartiles of CUET score :

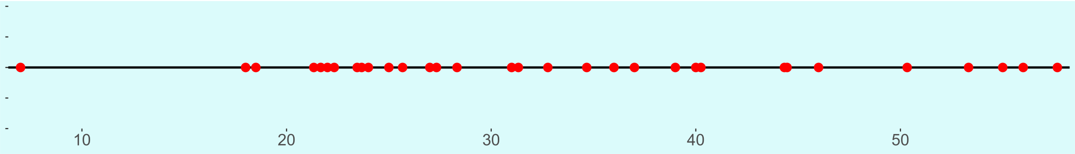
```
quantile(raw_data$CUET_score, probs = c(0.25, 0.5, 0.75))

##    25%    50%    75% 
## 124.00 138.00 153.25
```

☞ *JAM_score*

Let us plot the values of ***JAM_score*** along the real line.

```
df1 <- raw_data %>%
  drop_na()
```



- Measure of Central Tendency : Mean JAM score of the students is 33.2770588.

- Measure of Dispersion : JAM score has a standard deviation of 12.3388096.

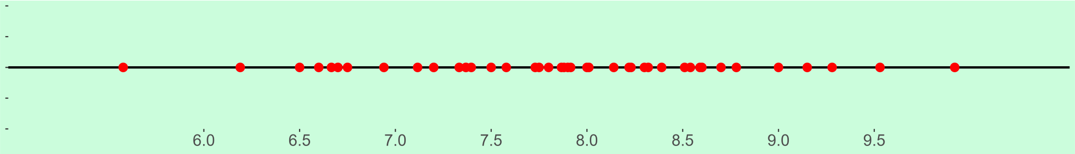
- Quartiles : The following are the quartiles of JAM score :

```
quantile(df1$JAM_score, probs = c(0.25, 0.5, 0.75))

##     25%     50%     75% 
## 23.7525 31.1650 40.1875
```

☞ ***UG_CGPA***

Let us plot the values of ***UG_CGPA*** along the real line.



- Measure of Central Tendency : Mean UG CGPA of the students is 7.8235682.

- Measure of Dispersion : UG CGPA has a standard deviation of 0.9138116.

- Quartiles : The following are the quartiles of UG CGPA :

```
quantile(raw_data$UG_CGPA, probs = c(0.25, 0.5, 0.75))

##    25%    50%    75%
## 7.2000 7.8735 8.4200
```

✎ 25% of the students have UG CGPA more than 8.42.

⚓ Coefficient of Variation

```
coefficient_of_variation <- function(x, na.rm = FALSE){
  return(sd(x, na.rm = na.rm) / mean(x, na.rm = na.rm))
}
```

```
coefficient_of_variation(raw_data$CUET_score)

## [1] 0.1374054
```

```
coefficient_of_variation(raw_data$JAM_score, na.rm = TRUE)

## [1] 0.3707903
```

```
coefficient_of_variation(raw_data$UG_CGPA)

## [1] 0.1168024
```

✎ So ***UG_CGPA*** has minimum variability.

⚓ Correlations

```
cor(df1$CUET_score, df1$JAM_score)

## [1] 0.4714785
```

```
cor(raw_data$CUET_score, raw_data$UG_CGPA)

## [1] 0.2304253
```

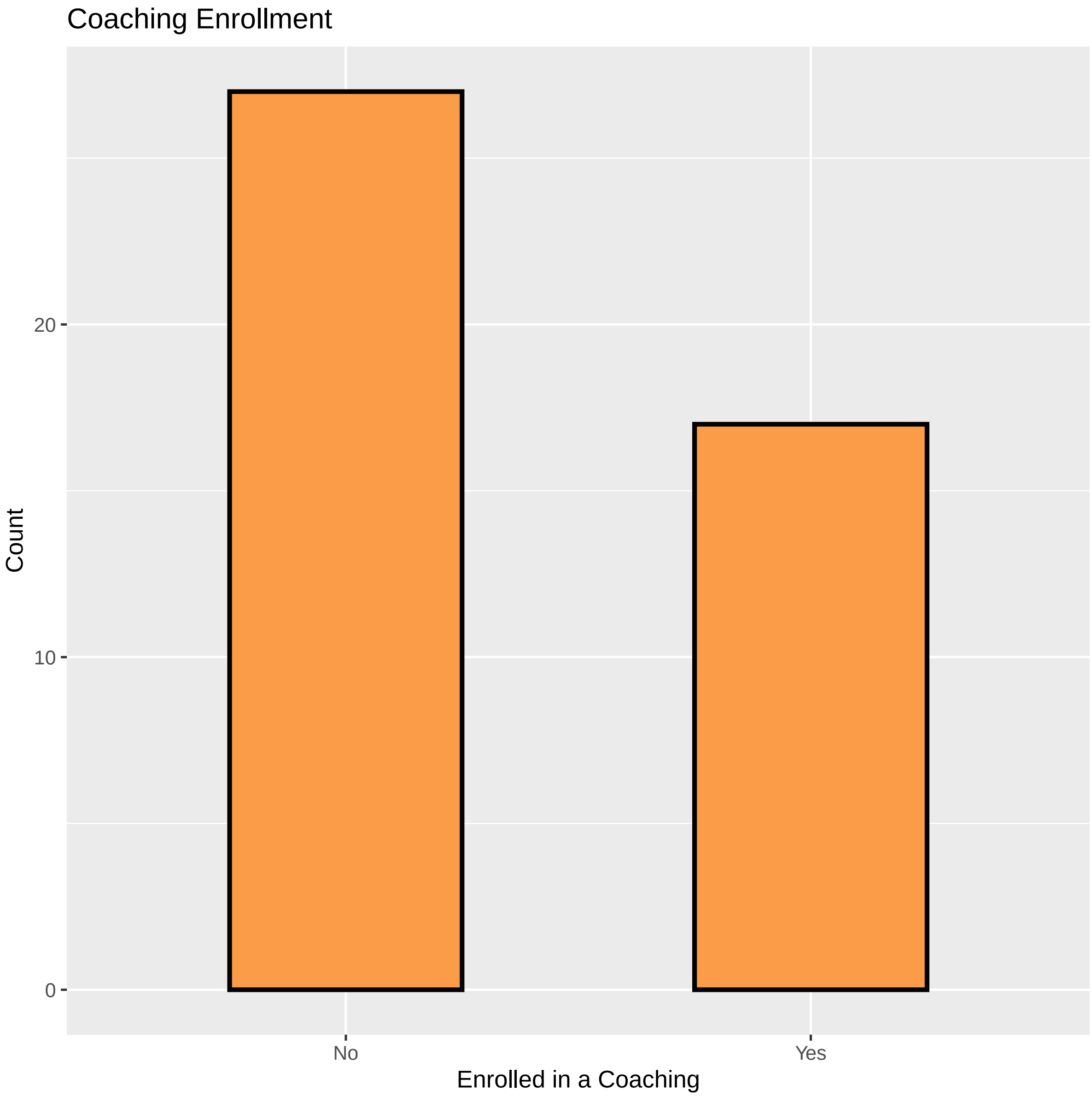
```
cor(df1$JAM_score, df1$UG_CGPA)

## [1] 0.07943115
```

✎ ***UG_CGPA*** and ***JAM_score*** have moderate correlation.

☞ Method of Preparation of the Students

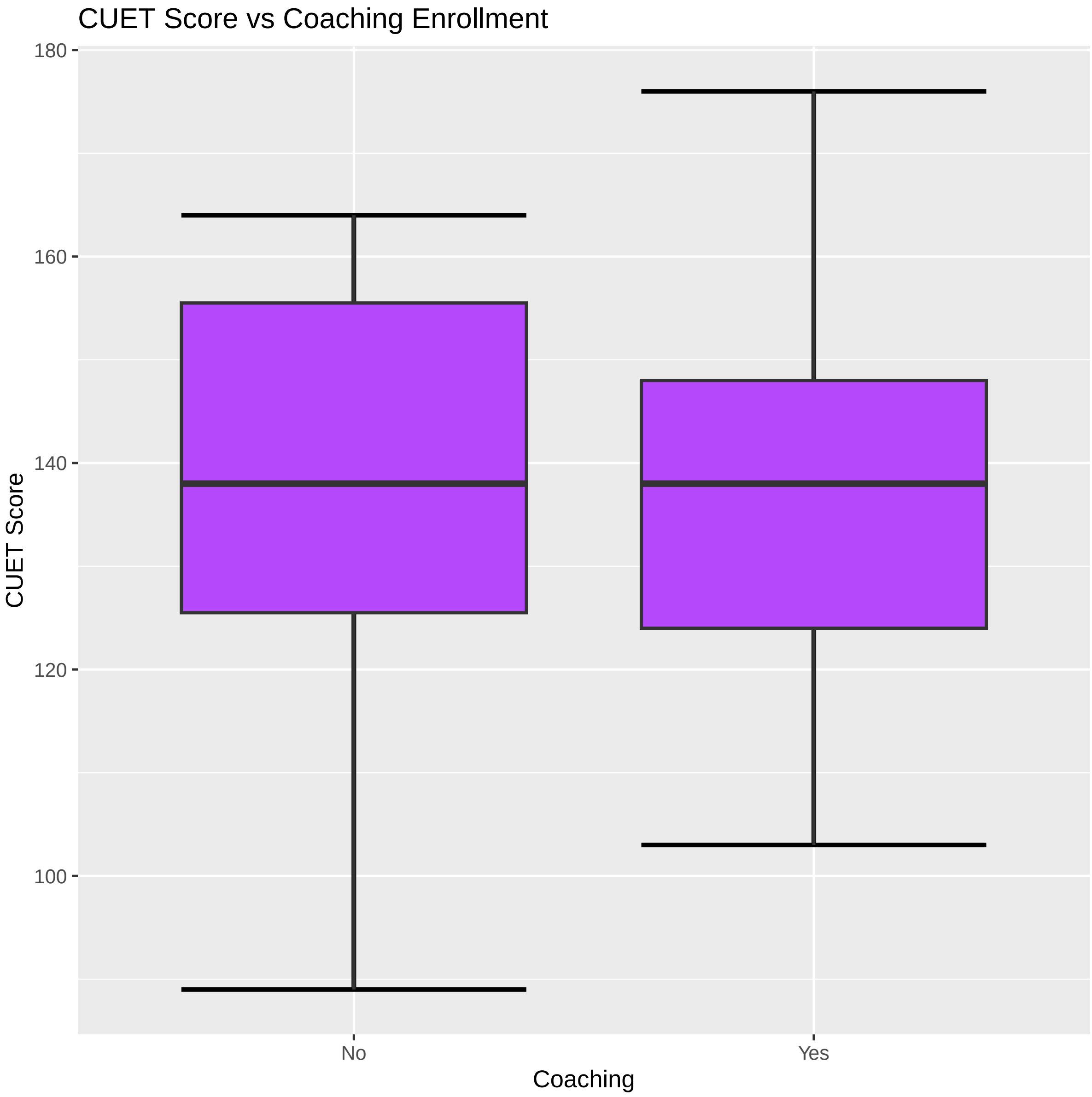
```
raw_data %>%
  ggplot(aes(x = fct_infreq(coaching))) +
  geom_bar(fill = "#fb9c48", width = 0.5, col = "black", linewidth = 1) +
  labs(x = "Enrolled in a Coaching", y = "Count", title = "Coaching Enrollment")
```



✎ So greater number of students prepared for competitive exam by self-study only.

☞ CUET Score for Two Methods of Preparation

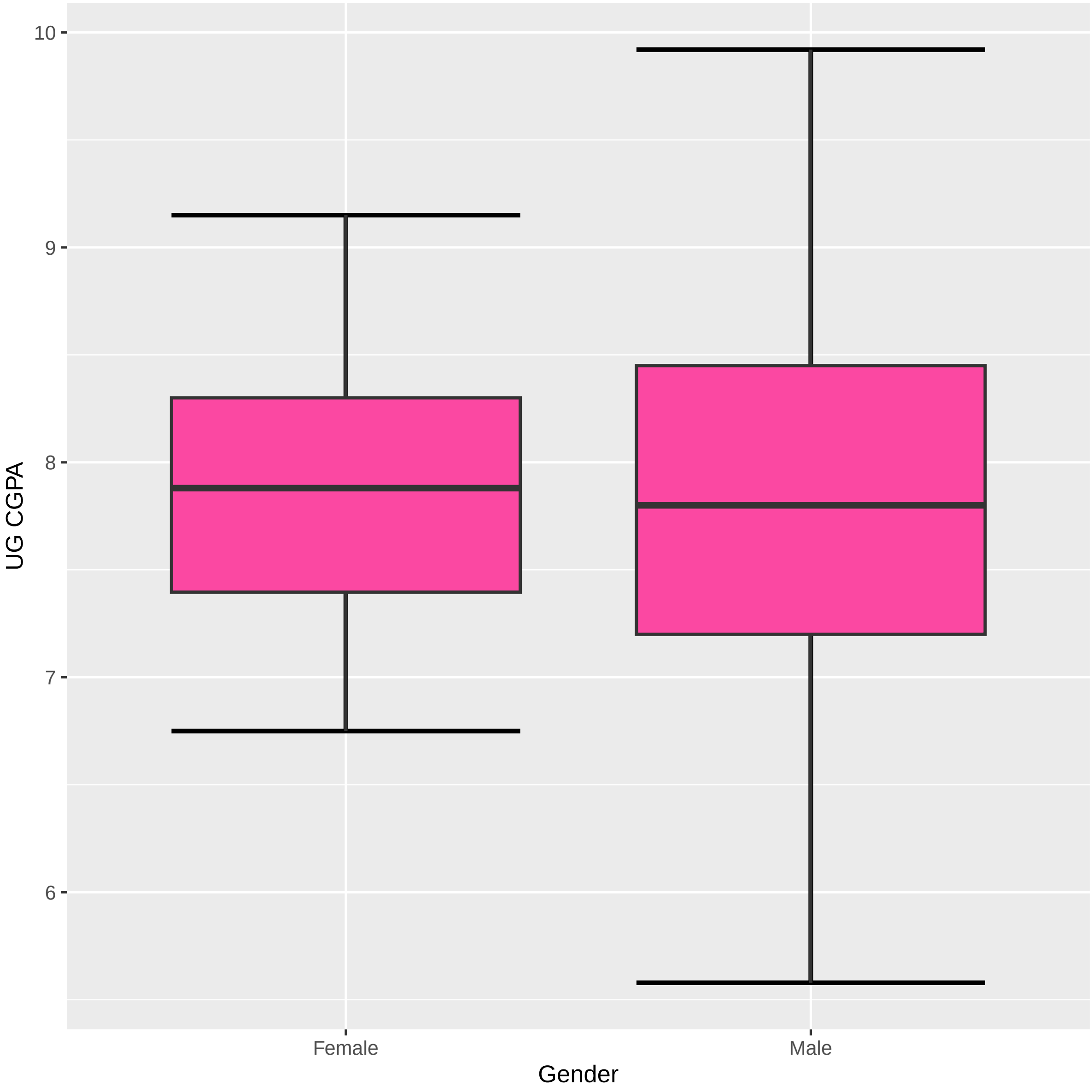
```r
raw_data %>%
  ggplot(aes(x = as.factor(coaching), y = CUET_score)) +
  stat_boxplot(geom = "errorbar", linewidth = 1) +
  geom_boxplot(fill = "#b548fb", linewidth = 0.7) +
  labs(x = "Coaching", y = "CUET Score", title = "CUET Score vs Coaching Enrollment")
```



✎ Median score of students who enrolled in a coaching institute and the ones who didn't are almost the same.

☞ UG CGPA for Male and Female Students

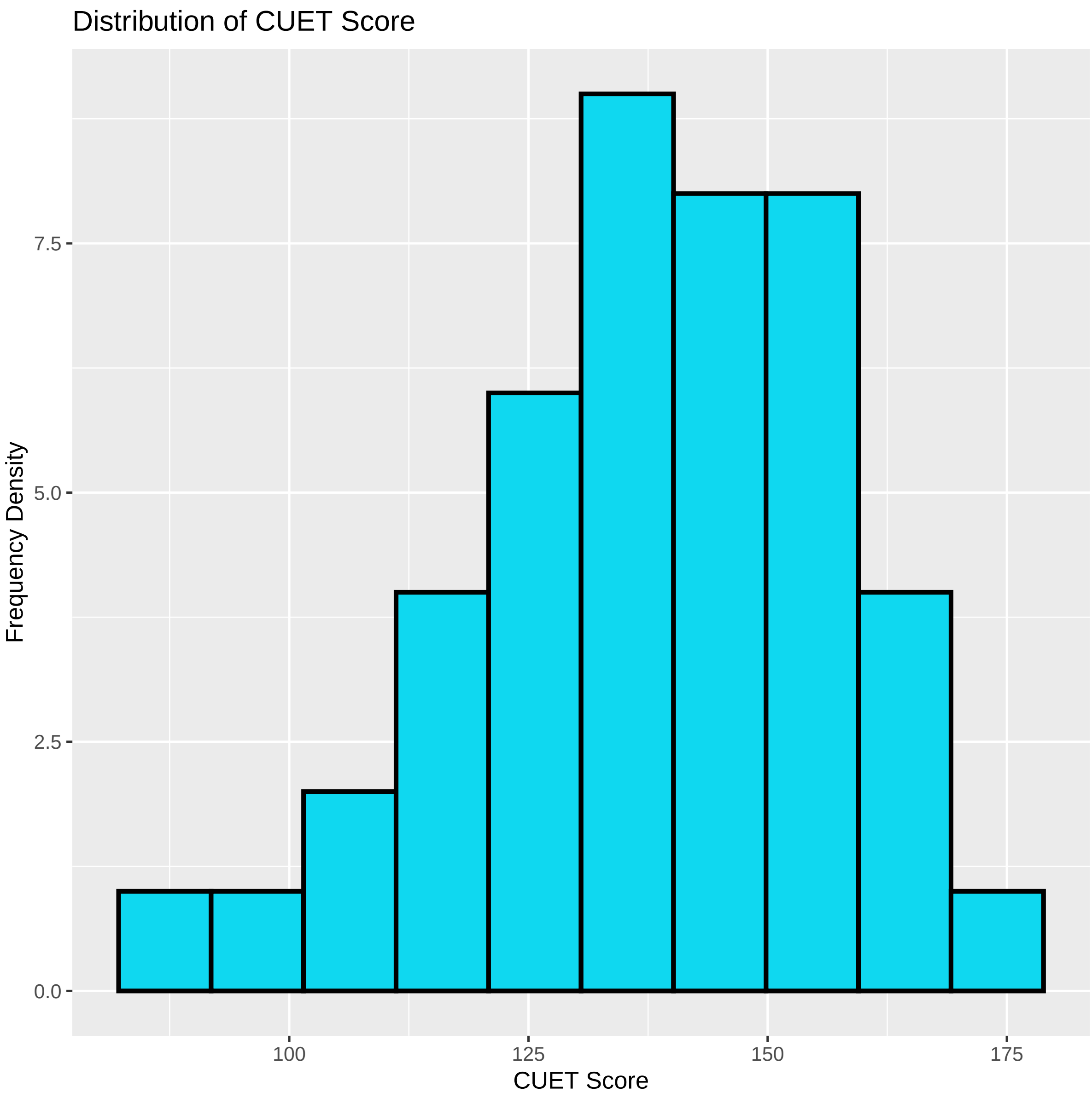
```
raw_data %>%
  ggplot(aes(x = gender, y = UG_CGPA)) +
  stat_boxplot(geom = "errorbar", linewidth = 1) +
  geom_boxplot(fill = "#fb48a2", linewidth = 0.7) +
  labs(x = "Gender", y = "UG CGPA")
```



✎ Average UG CGPA of female students is slightly higher than that of male students. Also the CGPAs of male students have greater dispersion.

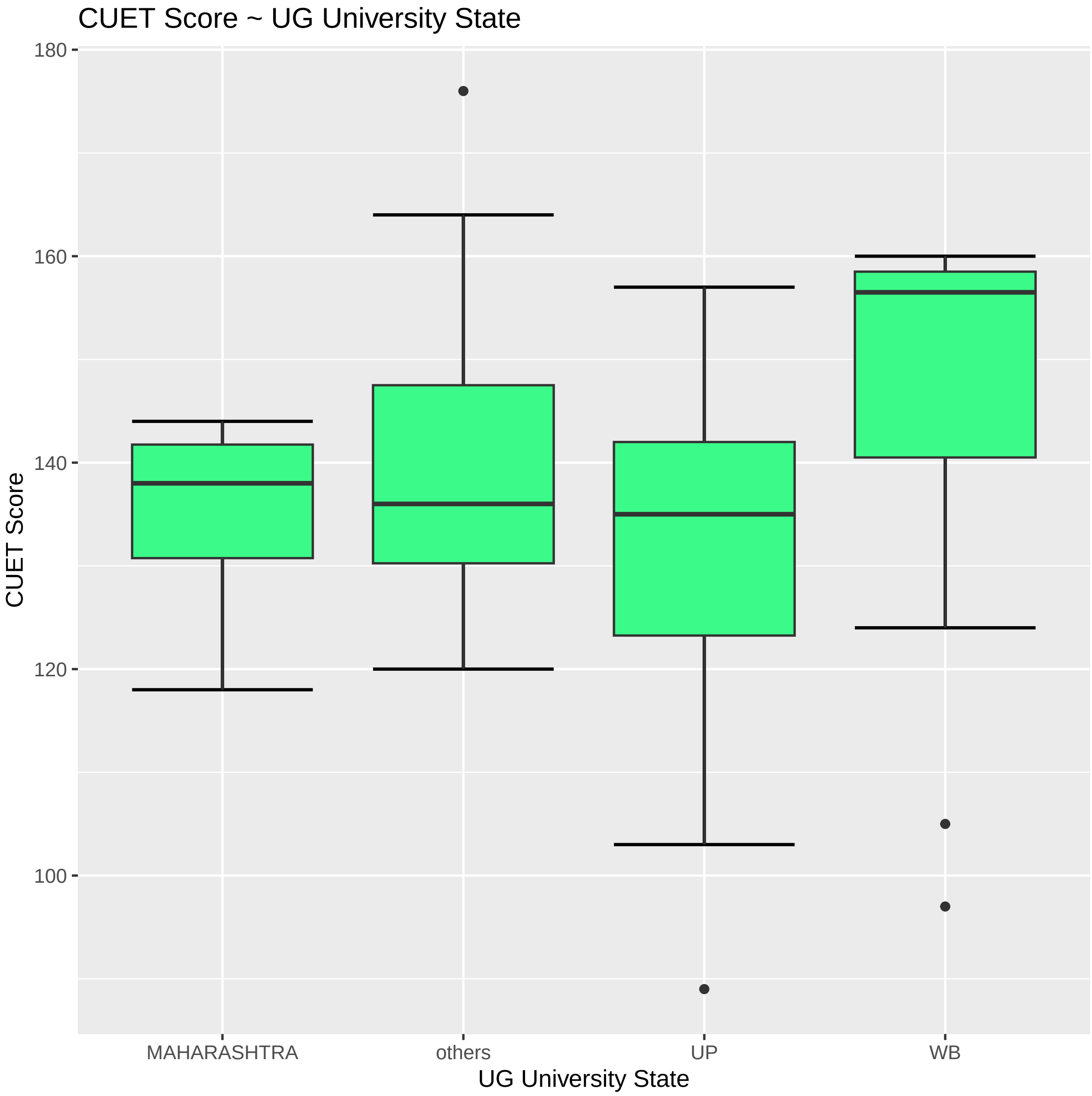
☞ Frequency Distribution of CUET Score

```r
raw_data %>%
  ggplot(aes(x = CUET_score)) +
  geom_histogram(fill = "#0FD8F0", bins = 10, col = "black", linewidth = 1) +
  labs(x = "CUET Score", y = "Frequency Density", title = "Distribution of CUET Score")
```



☞ CUET Score for Students of Different States

```r
score_univ_state %>%
  ggplot(aes(x = univ.state, y = score)) +
  stat_boxplot(geom = "errorbar", linewidth = 0.7) +
  geom_boxplot(fill = "#3bfa89", linewidth = 0.5) +
  labs(x = "UG University State", y = "CUET Score",
       title = "CUET Score ~ UG University State")
```



✎ Clearly, students from the Universities of West Bengal have better CUET Scores than rest of the students.