Visualizing Quantitative and Qualitative Variables

John Karuitha

Tuesday, October 17, 2023

Background

As noted, frequency tables, relative frequency tables and two way contingency tables provides a useful way of summarising datasets that have one or two categorical variables. A pie chart or bar chart/ column chart are useful for visualizing these data.

When we have quantitative variables only, the best summaries are measures of center and measures of spread. The histogram and density plots (smoothed histograms) are useful in visualizing such data.

At times, you have a dataset that consists of both quantitative and Qualitative variables and want to summarise and visualize the datasets. Here you may employ measures of center and spread for the quantitative variable and frequency tables, relative frequency tables and contingency tables to summarise such datasets.

A boxplot is one of the useful tools for visualizing such data. Violin plots are a better alternative.

In this section, we examine how to summarise and visualize a dataset comprising of a quantitative variable and a categorical variable.

The data

Let us take a simple dataset of the heights (in Inches) of students in a class and the sex of students. The data comes from the `dslabs` package in `R`. The data represents the self-reported heights of students taking the Bachelors in Data science degree at Harvard University.

You can access the dataset in R as follows (ensure you are connected to the internet to install the package dslabs if it is not already installed. If dslabs is already available on your computer, you do not need the internet).

```
install.packages("dslabs") #For this you must connect to the internet
library(dslabs) ## Will work without internet
data(heights) ## Will work without internet
```

\begin{table}

\caption{Top 6 Observations of Heights vs Sex of a Sample of

```
    Students}
```

sex	height
Male	75
Male	70
Male	68
Male	74
Male	61
Female	65

\end{table}

Table 1: Bottom 6 Observations of Heights vs Sex of a Sample of Students

	sex	height
1045	Male	50.00
1046	Female	69.00
1047	Male	69.00
1048	Male	63.39
1049	Male	66.00
1050	Male	66.00

Describibining a Dataset

The dataset has 2 variables and 1050 observations. Sex is a categorical variable with 2 levels, male and female. Height is a continuous quantitative variable with a mean of 68.323 and a median of 68.5. The table below shows further summaries of the data.

```
##      sex
##  Female:238
##  Male  :812
```

There are 238 females and 812 males in the dataset.

```
##      height
##  Min.   :50.0
##  1st Qu.:66.0
##  Median :68.5
##  Mean   :68.3
##  3rd Qu.:71.0
##  Max.   :82.7
```

The Quartiles

We have already discussed the mean and the median. The minimum (min) is the smallest number in a range of quantitative variables. The Maximum (max) is the opposite of the minimum. The range is the difference between the minimum and the maximum and is a common measure of spread. The range is however affected a lot by extreme values. What we have not looked at are the quartiles. The median divides the data into two parts, and is usually a single figure. Quartiles, on the other hand divide the dataset into 4 equal parts. Unlike the median which is just one value, there are THREE values for quartiles.

- Quartile 1 (or 1st Quartile):
- The median (or the second quartile):
- Quartile 3 (or 3rd Quartile):

An easy way to compute the quartiles is as follows. First, get the median. The median will have divided the data into two halves. Note that the median is also the second quartile (quartile connotes a division by 4, so the second quartile is $\frac{2}{4} = \frac{1}{2}$).

Take the lower half of the data, including the median, and also get its median. That should give you the first quartile. Similarly, take the upper half of the dataset, including the median and compute its median. That is the 3rd quartile.

Let us use a simpler dataset to illustrate this. We shall revisit the heights dataset later.

Example

In this illustration, we create a fictitious dataset with 21 observations to illustrate quartiles. Compute the three quartiles for the given data.

```
##  [1] 47.49 50.66 49.61 54.43 50.58 51.59 47.09 53.57 45.87 48.20 50.45 50.48
## [13] 48.99 53.70 50.62 49.85 48.06 52.55 45.43 61.55 47.81
```

The most obvious quartile is the median, the second quartile. remember to get the median and all quartiles in general, the data must be in ascending or descending order. The median is the 11th element- 50.4494. Mark it!!

```
##  [1] 45.43 45.87 47.09 47.49 47.81 48.06 48.20 48.99 49.61 49.85 50.45 50.48
## [13] 50.58 50.62 50.66 51.59 52.55 53.57 53.70 54.43 61.55
```

Now, we focus on the the data part below the median (remember to also include the median there). The median of that lower part is 48.05573. This is the 1st quartile.

```
##   25%
## 48.06
```

Now we focus on the upper portion of the dataset, again including the median. The median of the upper part is 51.59315.

```
##   75%
## 51.59
```

Example Get the quartiles and draw a boxplot

Percentile

The first quartile (which we may write as $\frac{1}{4}$ is equivalent to the 25th percentile or $\frac{25}{100} = \frac{1}{4}$).

The median is the 50th percentile $\frac{50}{100} = \frac{1}{2}$.

The 3rd quartile is the 75th percentile $\frac{75}{100} = \frac{3}{4}$.

While quartiles divide the data into 4 equal parts, percentiles divide the dataset into 100 equal parts.

The Box plot.

The boxplot presents a five number summary in a plot. The summaries presented in a boxplot are the minimum, the maximum, the median, the first quartile and the third quartile. Note that the mean is not included here- why? It is considered not robust to extreme values or outliers.

The 1st quartile, median and third quartile are captured using a box. The maximum and minimum are captured using whiskers. The box plot is also called a box and whiskers plot as a result. The figure below serves to illustrate this point.

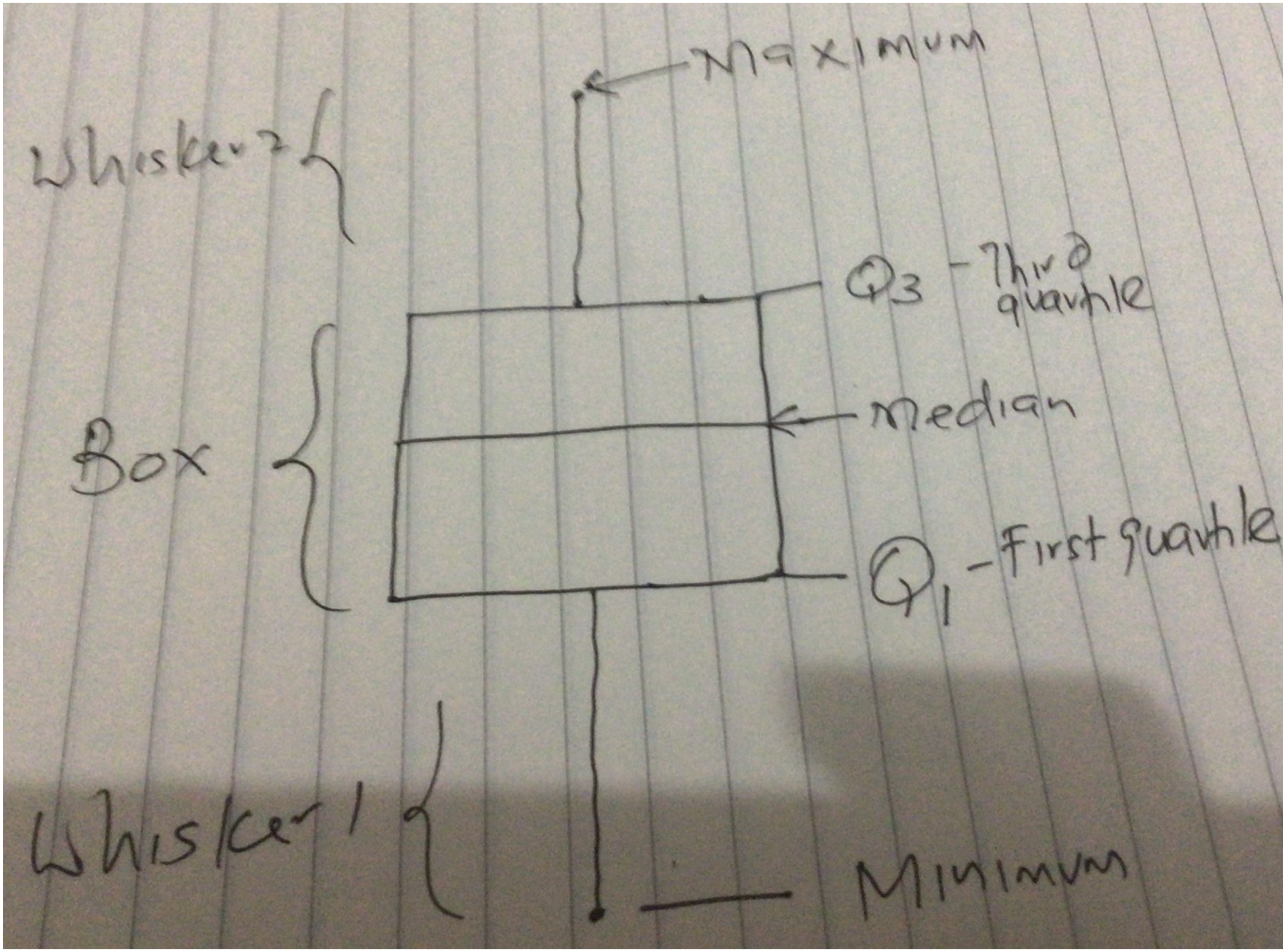
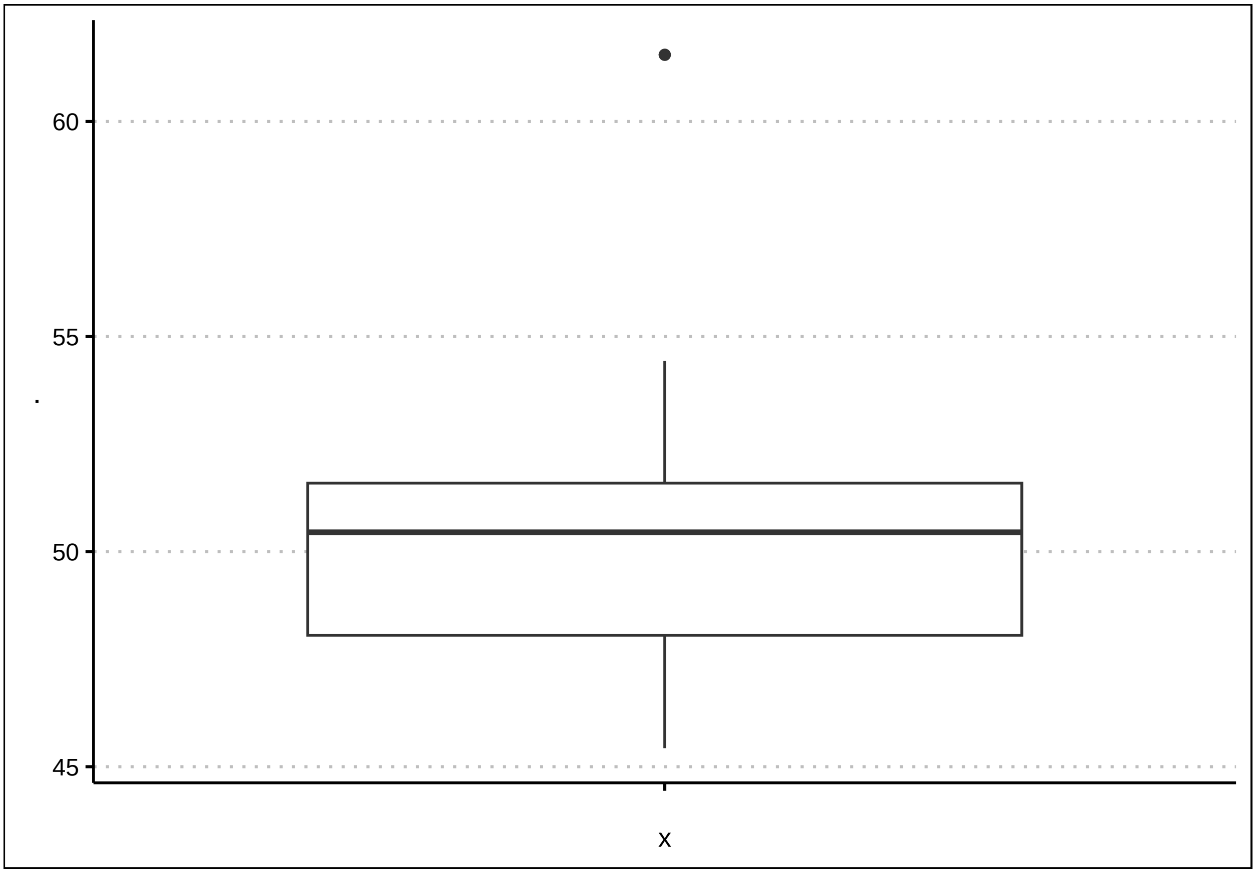


Figure 1: Box and Whiskers Plot

Now let us draw the boxplot for the data above



Note that in this case the maximum is represented outside the whiskers as a dot, meaning that it is an outlier. When an observation is too large or too small relative to other observations, it is presented separately as a dot. We shall not delve into outliers here, the box and whisker should do for us.

Now that we know what a box plot is, lets go back to our heights dataset.

Back to the Heights dataset

How well can we visualize this data? A box plot is a good choice given that it gives us five measures- minimum, Q1, median, Q3 and the maximum.

However, the boxplot tells us little about the distribution of the data or the shape. In this case, we could draw two boxplots side by side, one for males and one for females as follows.

1. Separate the data into males and females.
2. Arrange each of the two datasets in ascending or descending order.
3. Determine the minimum, Q1, median, Q3, and the maximum for each set of data; remember we now have two datasets male and female.
4. Draw the boxplots side by side with the x-axis representing sex (female versus male) and the y-axis representing the values above.

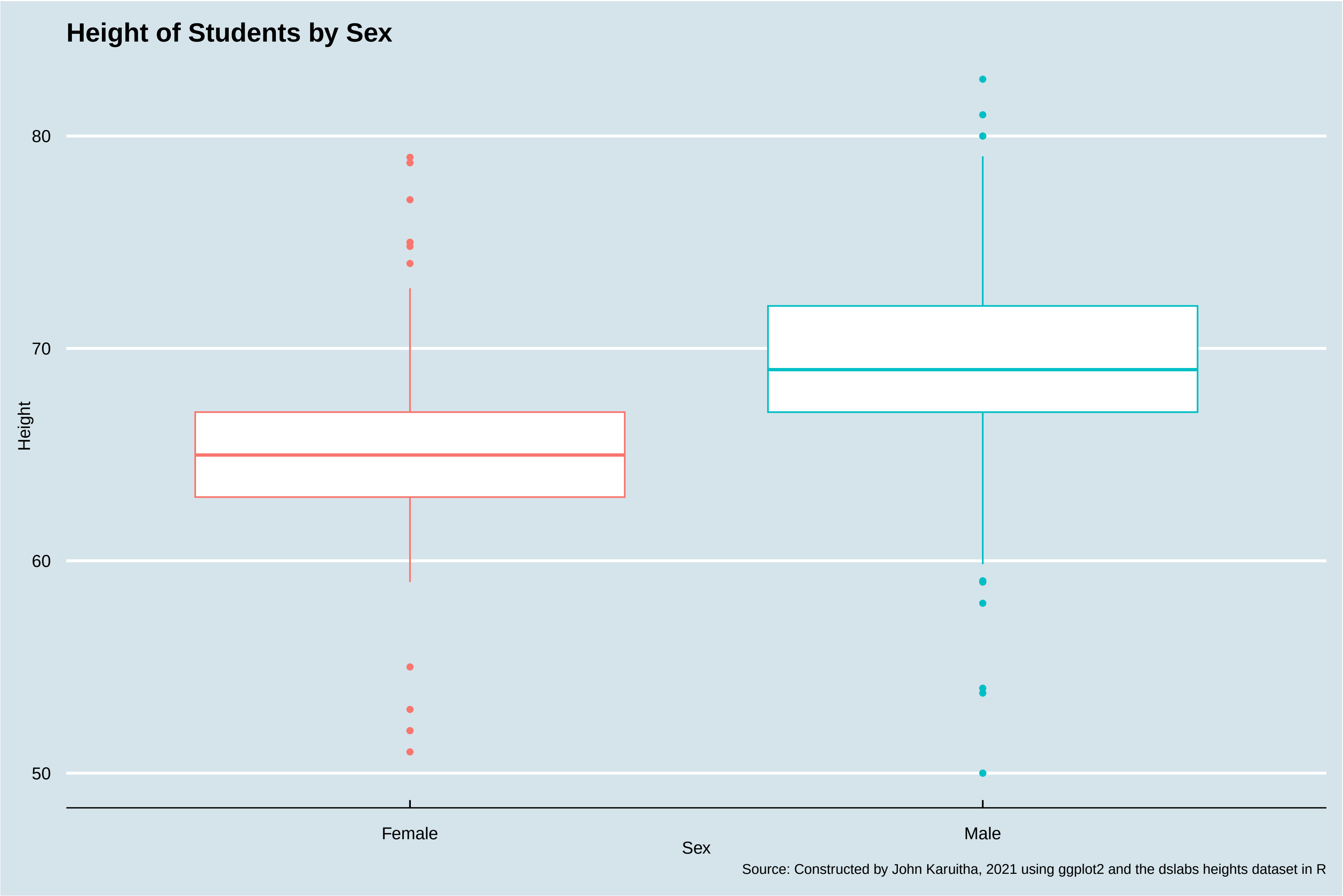
Below are summary statistics for the male and female datasets.

- Your sketch should be similar to this.

Table 2: Summary Statistics for Females and Males in the Heights Dataset

Variable	Sex	Mean	SD	Minimum	Q1	Median	Q3	Maximum
height	Female	64.94	3.761	51	63	64.98	67	79.00
height	Male	69.31	3.611	50	67	69.00	72	82.68

* Note that in all cases, males are on average taller than females. However, this does not mean that ALL males are taller than females as the chart on the next page shows.



Height of Students by Sex

Source: Constructed by John Karuitha, 2021 using ggplot2 and the dslabs heights dataset in R

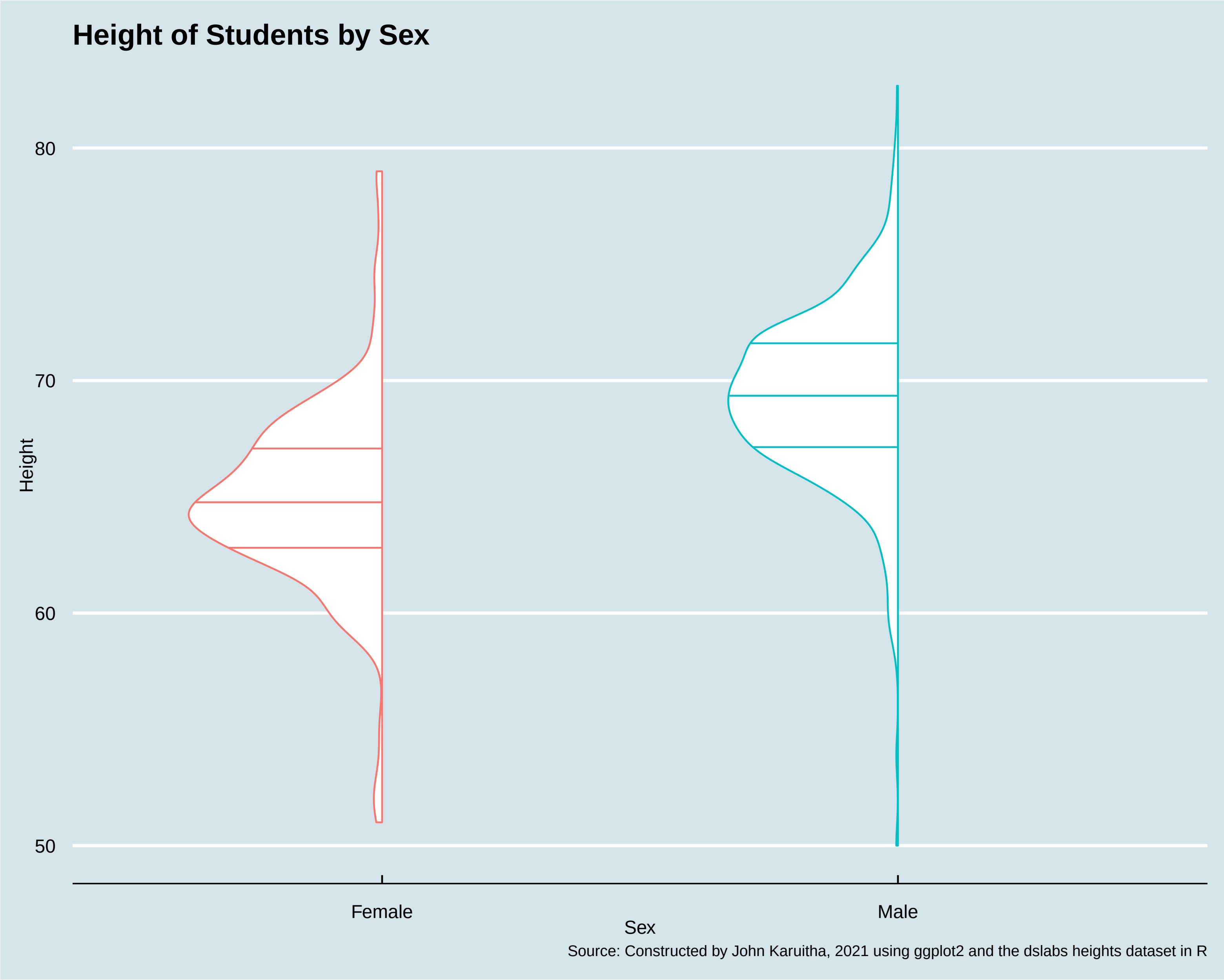
Interpreting the boxplots above

The interpretation could be something like this;

- Not all women are shorter than men.
- Not all men are taller than women.
- On average, men tend to be taller than women. Look at the median heights for men and women.
- The height data in this case has extreme values, that is, females and males who are extremely tall and extremely short than the average.

The Violin Plot

As noted earlier, the boxplot hides the distribution of the data. An improved version of a boxplot is the violin plot that combines histograms with the boxplot concept. While the violin plot is not examinable, it is useful to know it for your future research requirements. Turning the violin plot on its side gives you the density plots (smoothed histograms) for the dataset while retaining all the information contained in the boxplot.



Height of Students by Sex

Source: Constructed by John Karuitha, 2021 using ggplot2 and the dslabs heights dataset in R

Exercise

The dataset below shows the body mass index for two categories of people; those that exercise often versus those who do not exercise at all.

Table 3: Effects of Exercise on BMI

status	bmi
exercise	20.08
exercise	20.23
exercise	20.43
exercise	20.56
exercise	20.06
exercise	19.89
exercise	18.98
exercise	19.70
exercise	20.17
exercise	21.42
exercise	19.90
exercise	19.18
exercise	19.53
exercise	20.57
exercise	18.13
exercise	19.37
exercise	19.96
exercise	21.44
exercise	19.08
exercise	19.98
no_exercise	25.66
no_exercise	26.50
no_exercise	30.01
no_exercise	21.99
no_exercise	26.18
no_exercise	20.51
no_exercise	28.72
no_exercise	25.11
no_exercise	21.88
no_exercise	25.46
no_exercise	21.86
no_exercise	27.10
no_exercise	27.27
no_exercise	22.67
no_exercise	21.06
no_exercise	23.89
no_exercise	24.75
no_exercise	24.22
no_exercise	25.65
no_exercise	26.11
no_exercise	27.71

* source: Authors construction

REQUIRED

- Assume the data is not stratified into exercise versus no exercise, draw a box plot for the entire dataset.
- Divide the data into 2 groups of exercise versus no_exercise and draw the boxplots for both groups side by side. note that in this case, x will be the status, exercise vs no exercise, while y is the BMI.
- Based on the plot, describe the effects of exercise on BMI.