Outlier Detection (Lab-7)

Code

Bibek Sapkota

Outlier Detection (Part-2)

BoxPlot

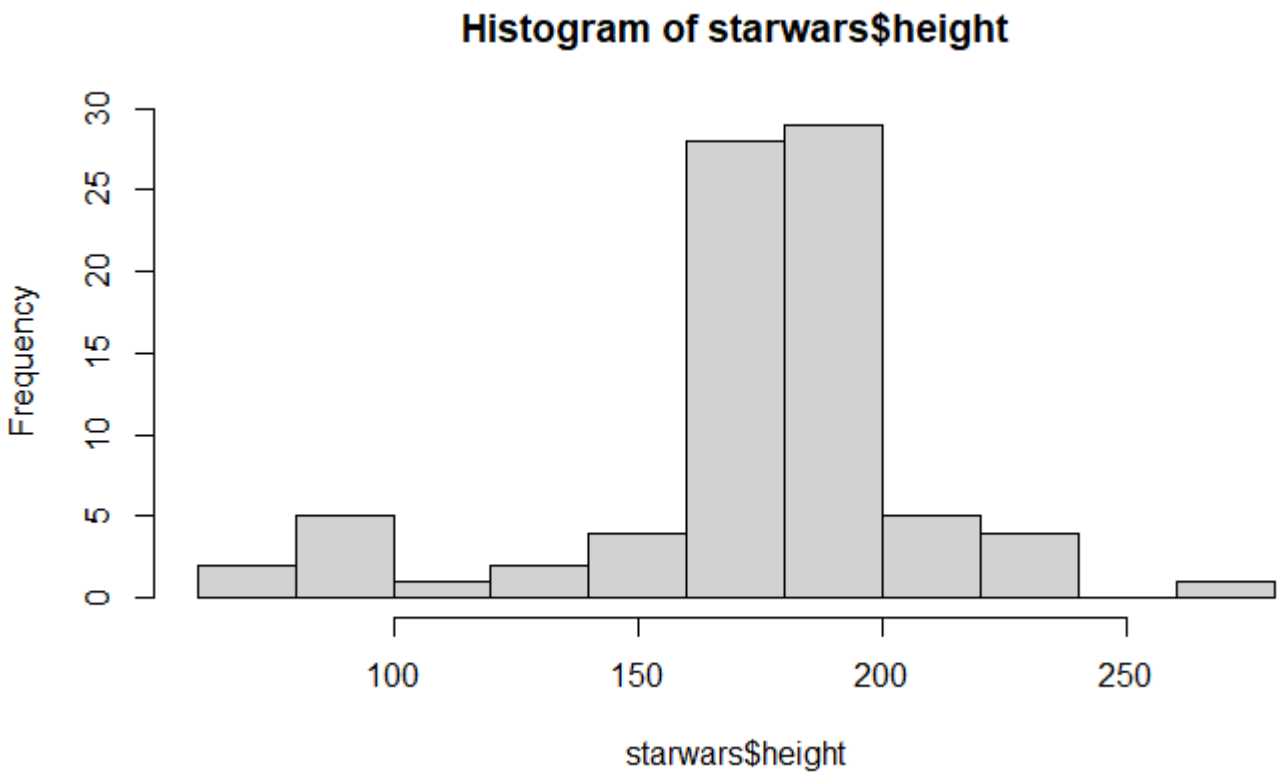
Hide

```
library("tidyverse")
```

task 1:Ploting the data.

Hide

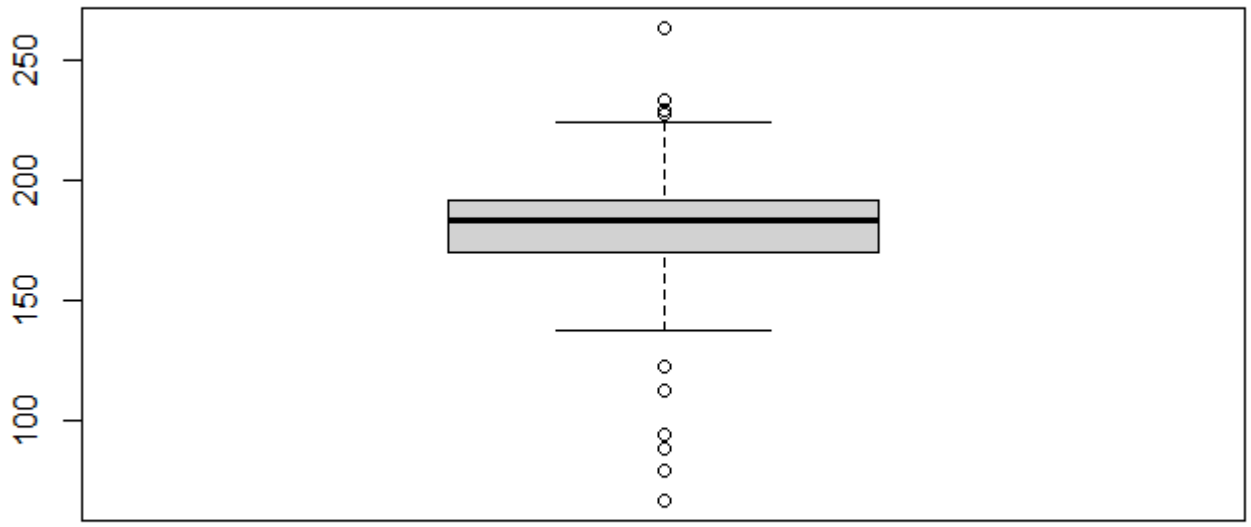
```
hist(starwars$height)
```



task 2:Creating a new dataset that only includes males and feamles and creating a boxplot using R

Hide

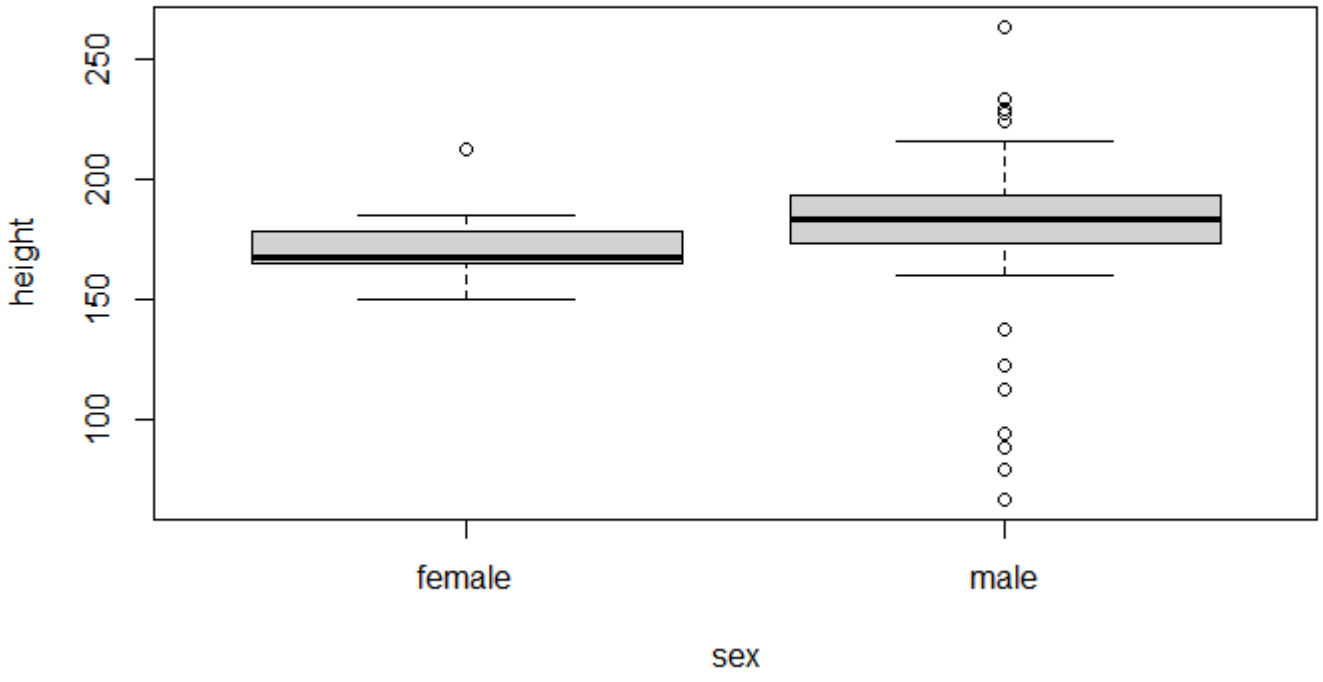
```
starwars_mf= starwars %>% filter(sex %in% c("male", "female"))

boxplot(starwars_mf$height)
```



task 3:creating seprate boxplot for both

Hide

```
boxplot(height~sex, data = starwars_mf)
```



Q4: How many outliers in males and females?

Ans- There is 1 outliers in females and 12 in males.

task 4:Filtering out outlier values

Hide

```
outliers <- boxplot(starwars_mf$height, plot=FALSE)$out
```

Q5: What does plot=FALSE do?

Ans- Boxplot function in R, the plot=FALSE parameter is used to suppress the creation of a plot.

Removing outliers from the dataset

task 1:First you need find in which rows the outliers are

Hide

```
starwars_mf[which(starwars_mf$height %in% outliers),]
```

name <chr>	**height** <int>	**m…** <dbl>	**hair_color** <chr>	**skin_color** <chr>	**eye_color** <chr>	**birth_year** <dbl>	**sex** <chr>	**ger** <ch
Chewbacca	228	112	brown	unknown	blue	200	male	mas
Yoda	66	17	white	green	brown	896	male	mas
Wicket Systri Warrick	88	20	brown	brown	brown	8	male	mas
Sebulba	112	40	none	grey, red	orange	*NA*	male	mas
Ratts Tyerel	79	15	none	grey, blue	unknown	*NA*	male	mas
Dud Bolt	94	45	none	blue, grey	yellow	*NA*	male	mas
Gasgano	122	*NA*	none	white, blue	black	*NA*	male	mas
Yarael Poof	264	*NA*	none	white	yellow	*NA*	male	mas
Lama Su	229	88	none	grey	black	*NA*	male	mas
Tarfful	234	136	brown	brown	blue	*NA*	male	mas

1-10 of 10 rows | 1-9 of 14 columns

task 2:you can remove the rows containing the outliers, one possible option is:

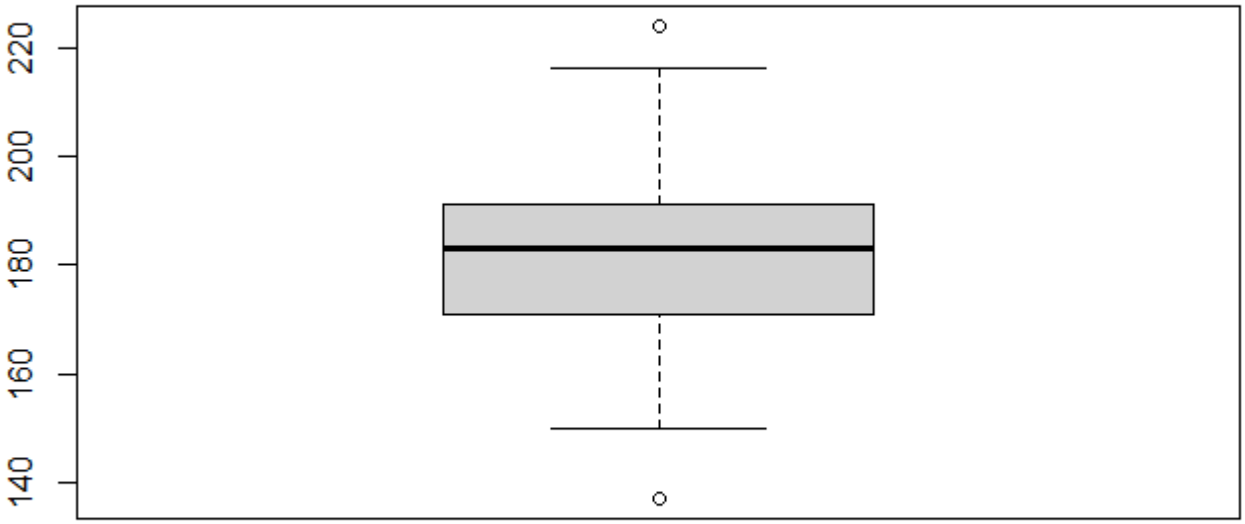
Hide

```
starwars_mf_new <- starwars_mf[-which(starwars_mf$ height %in% outliers),]
```

task 3:check outliers with boxplot

Hide

```
boxplot(starwars_mf_new$height)
```



3σ Rule

task 1: calculating standard deviation and mean

Hide

```
sd_value <- sd(starwars_mf$height, na.rm = TRUE)
sd_value
```

```
[1] 33.06843
```

Hide

```
mean_value <- mean(starwars_mf$height, na.rm = TRUE)
mean_value
```

```
[1] 177.6338
```

Q6: In above equations, what should be given as the value of na.rm and why?

Ans- In the given equations, the value of na.rm should be TRUE. This is because na.rm stands for “NA remove”, and setting it to TRUE will instruct the functions to remove any NA (missing) values before performing the calculation. If NA values are not removed, the functions sd and mean will return NA as the result because they cannot compute the standard deviation or mean with missing values present.

task 2:Then calculate upper and lower bounds

Hide

```
upper_bound <- mean_value + 3*sd_value
upper_bound
```

```
[1] 276.8391
```

Hide

```
lower_bound <- mean_value - 3*sd_value
lower_bound
```

```
[1] 78.42851
```

task 3: Extract outliers

Hide

```
outliers_sigma <-starwars_mf %>% filter((height > upper_bound)| (height < lower_bound))
outliers_sigma
```

n…	height	m…	hair_color	skin_color	eye_color	birth_year	sex	gender	homewo…
<chr>	<int>	<dbl>	<chr>	<chr>	<chr>	<dbl>	<chr>	<chr>	<chr>
Yoda	66	17	white	green	brown	896	male	masculine	*NA*

1 row | 1-10 of 14 columns

Hampel Identifier

task 1: Calculate median and MAD

Hide

```
median_value <- median(starwars_mf$height, na.rm = TRUE)
median_value
```

```
[1] 183
```

Hide

```
MAD_value <- mad (starwars_mf$height, na.rm = TRUE)
MAD_value
```

```
[1] 19.2738
```