Introduction – Descriptive Statistics for Lung Capacity Dataset

This data is shows lungs capacity of smokers and non-smokers differ by age, gender and height

The dataset has six attributes:

1. LungCap(cc): Lung capacity in cubic centimeters
2. Age(years): Age of person in years
3. Height(inches): Height of person in years
4. Smoke: Does the person smoke
5. Gender: Gender of person.
6. Caesarean: the person's birth was normal or Caesarean.

Steps to be performed:

1. Import the dataset into R
2. Understand the structure of dataset
3. Graphical exploration of dataset
4. Descriptive statistics about the dataset
5. Insights from the dataset.

Import data into R:

```
library(readxl)
Dataset <- read_excel("LungCap Dataset.xls")
```

Understand the structure of dataset

- LungCap and height are numerical.
- Age is integer, but it should be considered as categorical.
- Smoke, Gender and Caesarean are categorical with two factors – Yes and No.

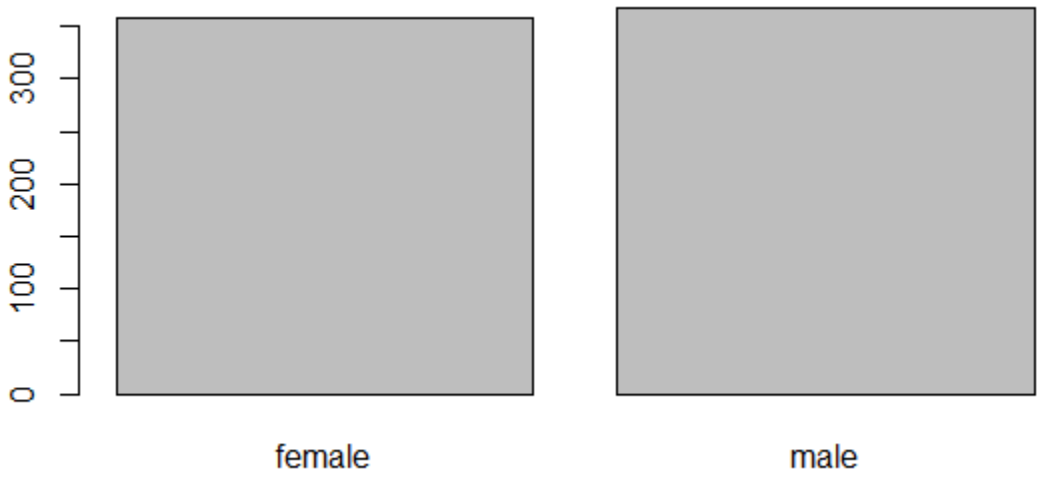
Graphical exploration of dataset

Out of six attributes lungCap and Height are numerical, and rest are categorical.

This data is showing lungs capacity of smokers and non-smokers by age, gender and height.

```
> names(Dataset)
[1] "LungCap.cc."    "Age..years."    "Height.inches." "Smoke"
[5] "Gender"         "Caesarean"

> table(Dataset$Gender) # Male:Female ratio

female   male
   358    367
```

```
> prop.table(table(Dataset$Gender)) #proporations

   female      male
0.4937931 0.5062069


> count = table(Dataset$Gender)
> barplot(count)
```

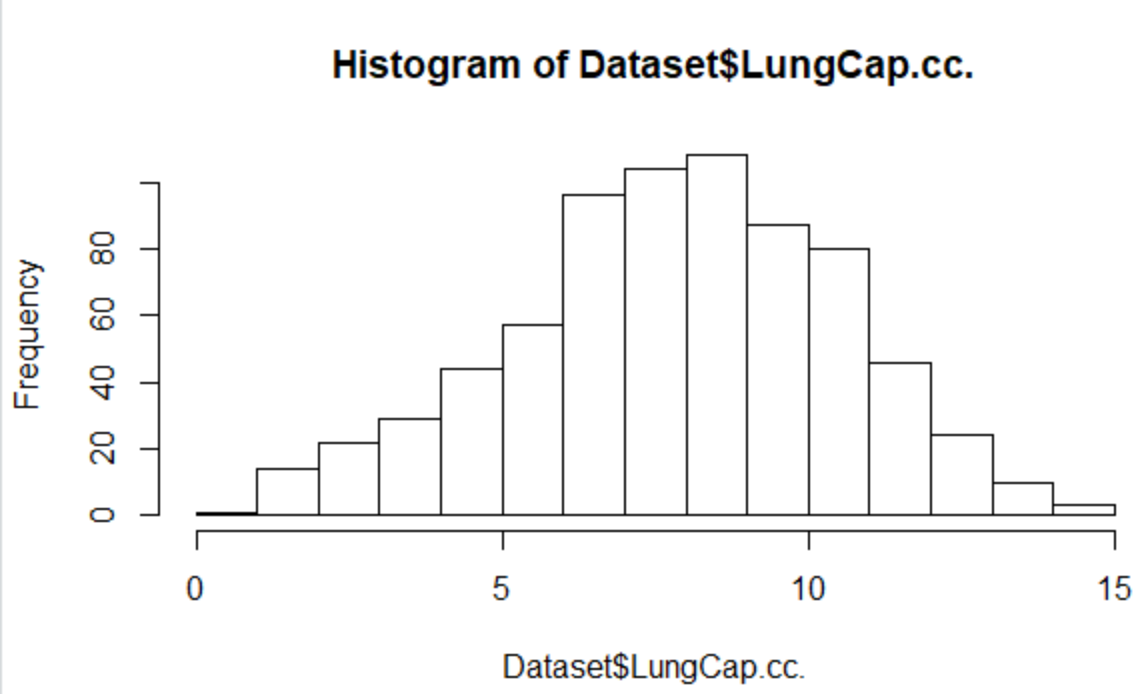


Inference: bar plot showing the percentage ratio between male and female in Gender variable (almost equal in proportions)

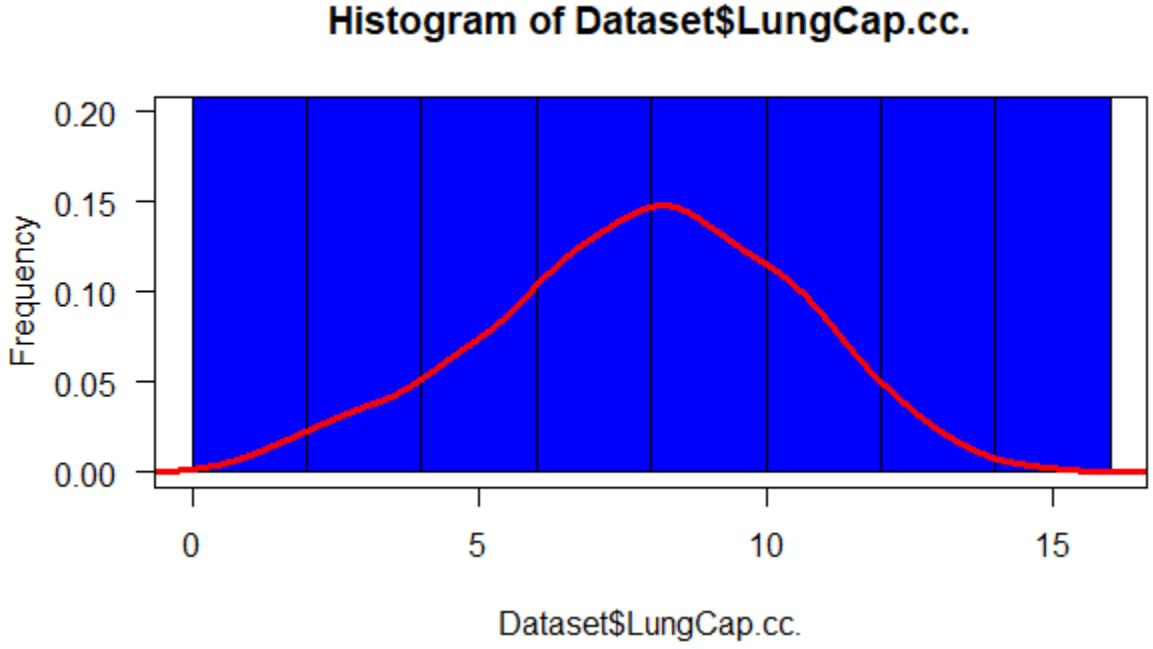
Histograms of Lungs Capacity:

Histograms is used for summarizing the distribution of a numeric variable.

```
> hist(Dataset$LungCap.cc.)
```



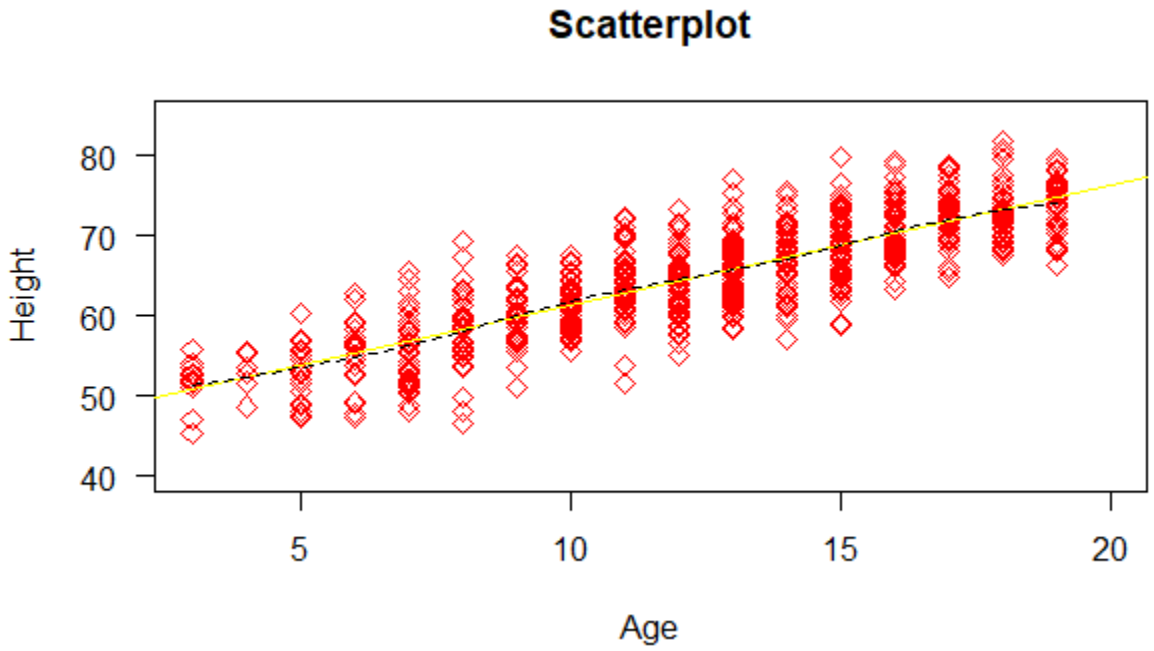

```
> hist(Dataset$LungCap.cc., ylim=c(0, 0.2), col=4, breaks=seq(from=0, to=16,
by=2), las=1, labels = TRUE )
> lines(density(Dataset$LungCap.cc.), col=2, lwd=3)
> box()
```



Inference: red line showing the Density for the variable Lung capacity

Scattered plot: which is shows lung capacity by age and abline showing the ratio of lung capacity by age

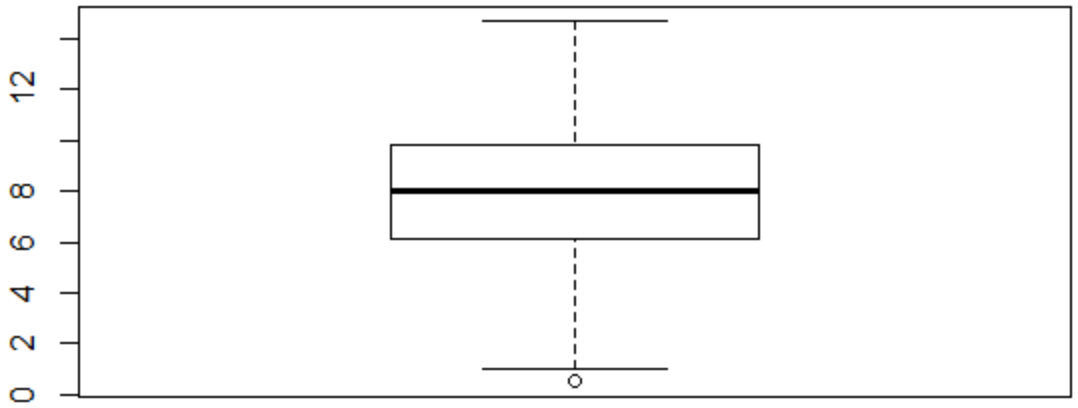
```
> plot(Age..years.,Height.inches.,main="Scatterplot",xlab="Age",ylab="Height"
,las=1,xlim=c(3,20),ylim=c(40,85),col=2,pch=5)
> abline(lm(Height.inches.~Age..years.),col=7)
> lines(smooth.spline(Age..years.,Height.inches.),lty=2,col=1)
```



Inference: There is a linear increase with respect to age and height of the person

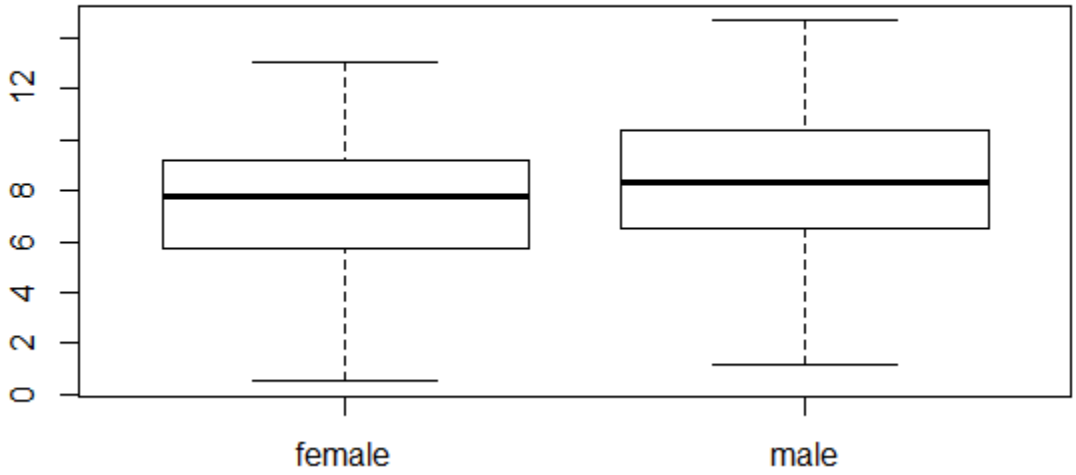
Boxplot for lungs capacity

```
> attach(Dataset)
> boxplot(LungCap.cc.)
```



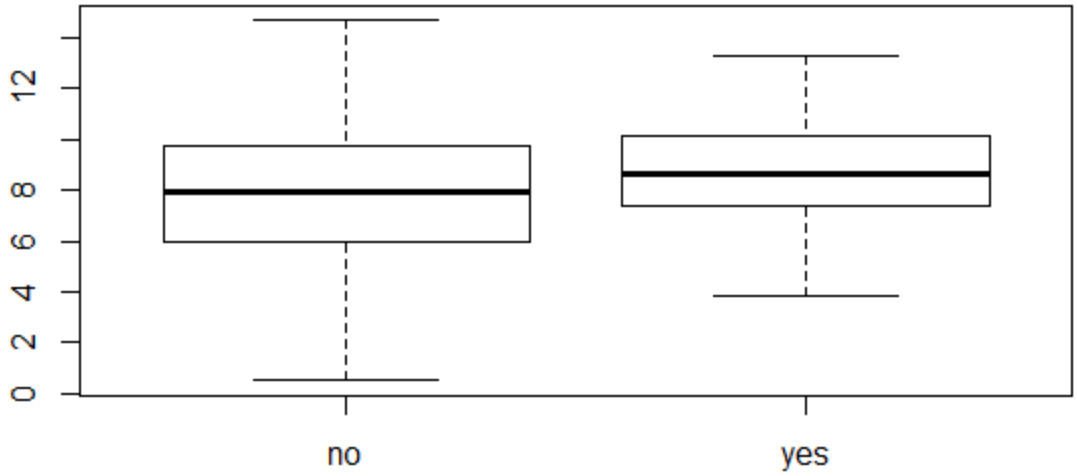
difference between lungs capacity between male and female.

```
> boxplot(LungCap.cc. ~ Gender)
```



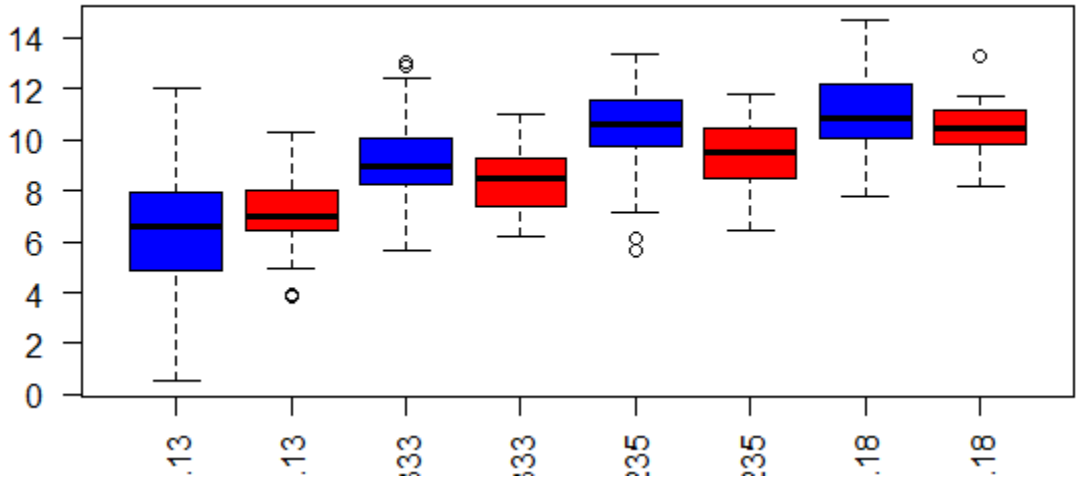
Boxplot to examine the relationship between smoker and non-smoker and lung capacity

```
> boxplot(LungCap.cc. ~ Smoke)
```



To examine relationship between Lung Capacity Vs. Smoker and non-smoker by Age Groups

```
> AgeGroups <- cut(Age..years., breaks=c(0,13,15,17,25), labels=c(13, 14/15,
16/17, 18))
> boxplot(LungCap.cc. ~ Smoke*AgeGroups,las=2,col=c(4,2))
```



Inference: the difference between the lung's capacity of the smokers and non-smokers by age groups. Red color showing smokers by age and blue color showing non-smoker by age. There is a clear cut trend between age and smoking

Descriptive statistics about the dataset

```
> summary(Dataset)
```

```
  LungCap.cc.      Age..years.    Height.inches.   Smoke        Gender    Caesarean
Min.   : 0.507   Min.   : 3.00   Min.   :45.30   no :648   female:358   no :561
1st Qu.: 6.150   1st Qu.: 9.00   1st Qu.:59.90   yes: 77   male  :367   yes:164
Median : 8.000   Median :13.00   Median :65.40
Mean   : 7.863   Mean   :12.33   Mean   :64.84
3rd Qu.: 9.800   3rd Qu.:15.00   3rd Qu.:70.30
Max.   :14.675   Max.   :19.00   Max.   :81.80
```

Inference:

- Lungs capacity distribution is identical to normal distribution with mean = 7.863
- Records with age 12 & 13 have highest number of observations
- Lungs capacity and Age are highly correlated.
- People with age 17 -19 have highest average lung capacity.