KNOWLEDGE ENGINEERING LAB (CSE 4.1.7)

1. **Introduction to exploratory data analysis using R**
   - i. **Load the 'iris. CSV' file and display the names and type of each column.**
   - ii. **Find statistics such as min, max, range, mean, median, variance, standard deviation for each column of data.**
   - iii. **Generate histograms and density plots for each sepal length, sepal width, petal length, petal width.**
   - iv. **Generate box plots for each of the numerical attributes. Identify the attribute with the highest variance.**

Exploratory data:

Exploratory data Analysis is an approach to analyzing data to summarize their main characteristics, often with visual techniques.

About the Dataset:

The Iris flower data set consists of 50 samples from each of three species of Iris Flowers: Iris Setosa, Iris Virginica and Iris Versicolor . The Iris flower data set was introduced by the British statistician and biologist Ronald (1).

Dataset contains four features measured from each sample are sepal length, sepal width, petal length and petal width, in centimeters. Iris data is publicly available to use and is one of the most widely used data set, mostly by the beginners in the area of Data Science & Machine Learning. It consists of a set of 150 records under 5 attributes — Sepal length, Sepal width, Petal length, Petal width and Class-Labels (Species)

In Machine learning terminology, the observed features like sepal length, sepal width, petal length and petal width are called independent variables while the class-label which is to be determined is called dependent variable.

Download the dataset: Here

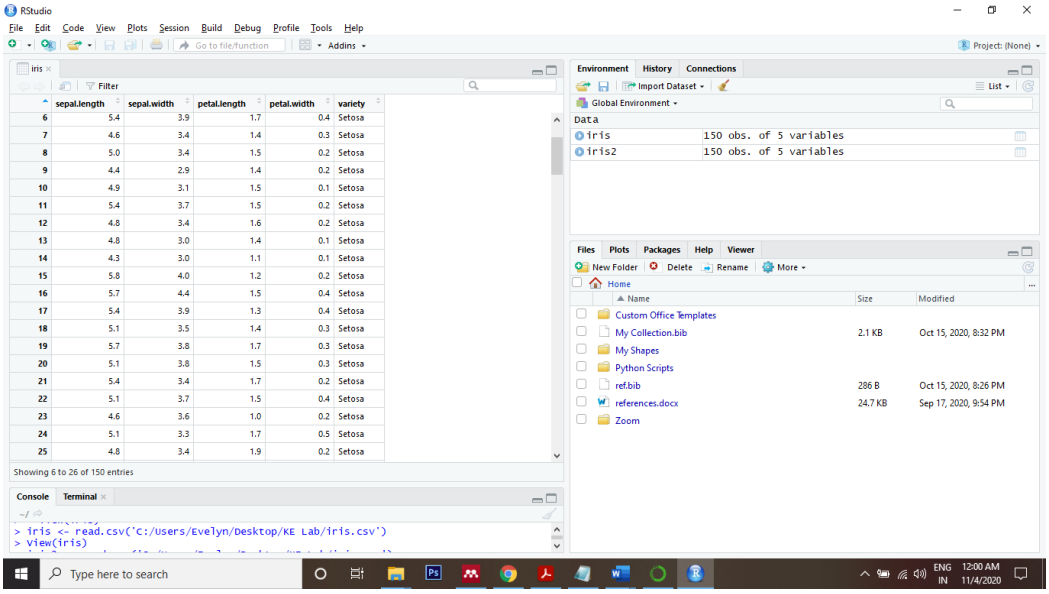
i. **Load the 'iris. CSV' file and display the names and type of each column.**

- Open R studio
- Load the iris dataset

```
iris <- read.csv("C:/Users/Evelyn/Desktop/KE Lab/iris.csv")
```

- View the dataset

```
View(iris)
```



- **Names and types of each column in the dataset**

```
str(iris)
'data.frame':    150 obs. of  5 variables:
 $ sepal.length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ sepal.width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ petal.length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ petal.width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ variety     : Factor w/ 3 levels "Setosa","Versicolor",..: 1 1 1 ...
```

ii. **Find statistics such as min, max, range, mean, median, variance, standard deviation for each column of data.**

- **Minimum value of each column in the dataset.**

  Minimum value of the feature or attribute in the dataset

```
min(iris$sepal.length)
[1] 4.3
min(iris$sepal.width)
[1] 2
min(iris$petal.length)
[1] 1
min(iris$petal.width)
[1] 0.1
```

- **Max value of each column in the dataset.**

  Maximum value of the feature or attribute in the dataset

```
max(iris$sepal.length)
[1] 7.9
max(iris$sepal.width)
[1] 4.4
max(iris$petal.length)
[1] 6.9
max(iris$petal.width)
[1] 2.5
```

- **Mean value of each column in the dataset.**

  The sample mean, also called the sample arithmetic mean or simply the average, is the arithmetic average of all the items in a dataset.

  Given, {2, 4, 6, 8, 2, 10, 12} is a set of data.

  Mean = 2+4+6+8+2+10+12/7 = 44/7

```
mean(iris$sepal.length)
[1] 5.843333
mean(iris$sepal.width)
[1] 3.057333
mean(iris$petal.length)
[1] 3.758
mean(iris$petal.width)
[1] 1.199333
```

- **Range of each column in the dataset.**

  Range value returns the minimum, and maximum value in the feature in the dataset.

```
range(iris$sepal.length)
[1] 4.3 7.9
range(iris$sepal.width)
[1] 2.0 4.4
range(iris$petal.length)
[1] 1.0 6.9
range(iris$petal.width)
[1] 0.1 2.5
```

- **Median of each column in the dataset.**

  The sample median is the middle element of a sorted dataset.

  Given, {2, 4, 6, 8, 2, 10, 12} is a set of data.

  To find median we have to first arrange the given data in ascending or descending order

  So, {2,2,4,6,8,10,12}. Thus,

  Median = 6

```
median(iris$sepal.length)
```

```
[1] 5.8
median(iris$sepal.width)
[1] 3
median(iris$petal.length)
[1] 4.35
median(iris$petal.width)
[1] 1.3
```

- **Variance of each column in the dataset.**

  The variance for a discrete variable made up of n observations is defined as:

  $$Variance = \frac{\sum(x - \bar{x})^2}{n}$$

  Where, x is the data point, $\bar{x}$ is the mean of all data points, n is number of data points

```
var(iris$sepal.length)
[1] 0.6856935
var(iris$sepal.width)
[1] 0.1899794
var(iris$petal.length)
[1] 3.116278
var(iris$petal.width)
[1] 0.5810063
```

- **Standard Deviation of each column in the dataset.**

  The standard deviation for a discrete variable made up of n observations is the positive square root of the variance and is defined as:

  $$Standard\ deviation = \sqrt{\frac{\sum(x - \bar{x})^2}{n}}$$

  Example

  A hen lays eight eggs. Each egg was weighed and recorded as follows:

  60 g, 56 g, 61 g, 68 g, 51 g, 53 g, 69 g, 54 g.

  First, calculate the mean:

  $\bar{x} = \frac{\sum x}{n}$ =59

  Calculation of the mean for example 1 a.

  Now, find the standard deviation.

  Table 1. Weight of eggs, in grams

Weight (x)	(x - mean)	(x - mean)2
60	1	1

56	-3	9
61	2	4
68	9	81
51	-8	64
53	-6	36
69	10	100
54	-5	25
472		320

Using the information from the above table, we can see that

$\sum(x - \bar{x})^2 = 320$

Variance is $\frac{320}{8} = 40$, and Standard deviation is $\sqrt{\frac{320}{8}} = 6.32$ (2)

```
sd(iris$sepal.length)
[1] 0.8280661
sd(iris$sepal.width)
[1] 0.4358663
sd(iris$petal.length)
[1] 1.765298
sd(iris$petal.width)
[1] 0.7622377
```

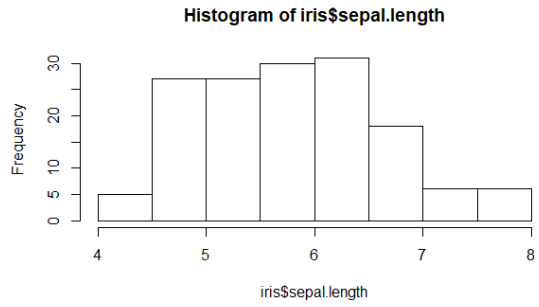
iii. Generate histograms and density plots for each sepal length, sepal width, petal length, petal width.

- **Histogram**

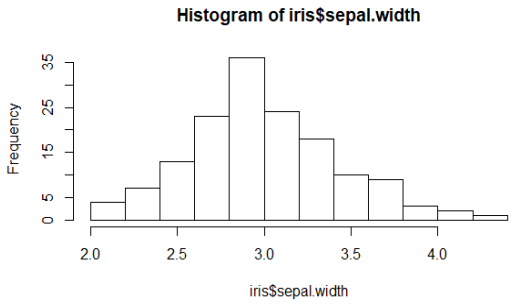
  Histogram are frequently used in data analys for visualizing the data. Through histogram, we can identify the distribution and frequency of the data. Histogram divide the continues variable into groups (x-axis) and gives the frequency (y-axis) in each group. The function that histogram use is hist(). With the breaks argument we can specify the number of cells we want in the histogram. (3).

  In simple words, A histogram represents counts within given intervals by the height of the bars.

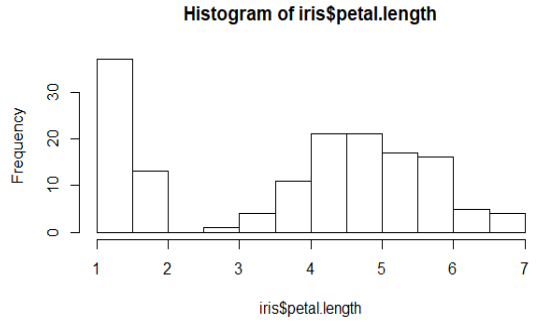
```
hist(iris$sepal.length)
```



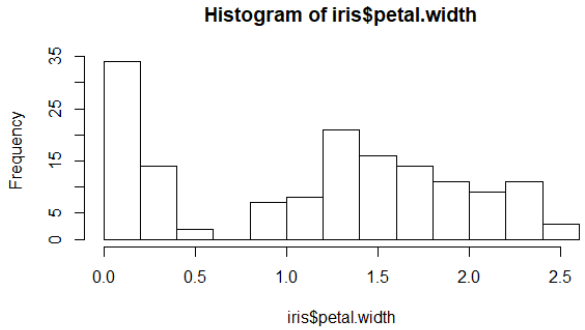

```
hist(iris$sepal.width)
```




```
hist(iris$petal.length)
```




```
hist(iris$petal.width)
```



- **Density Plots**

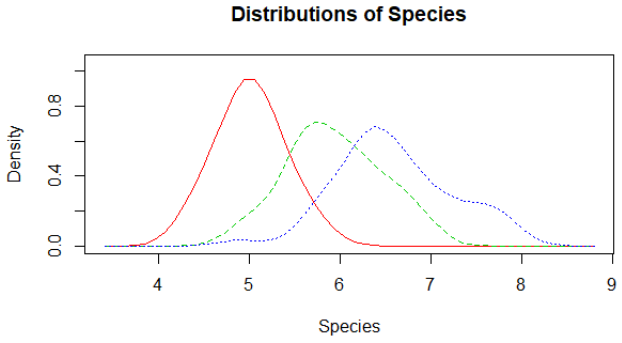
  Similar to the histogram, the density plots are used to show the distribution of data. Additionally, density plots are especially useful for comparison of distributions (4).

  Install a package called sm

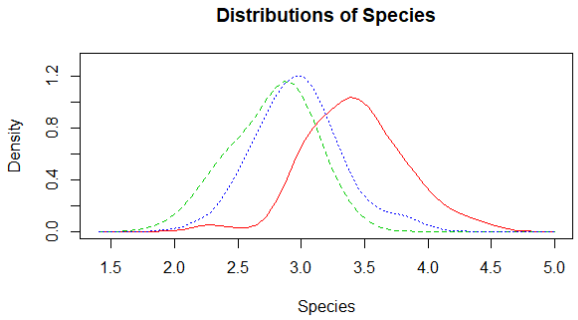
```
install.packages("sm")
library(sm)
```

  Density plots for iris

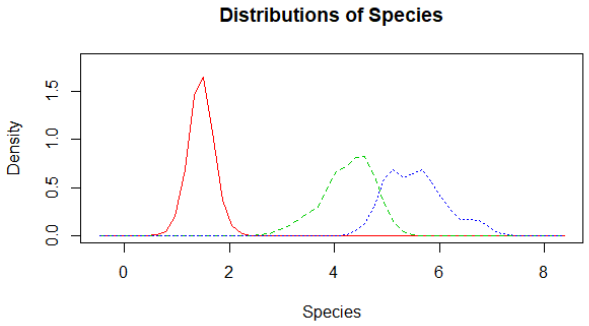
```
sm.density.compare(iris$sepal.length, iris$variety, xlab="Species")
title(main="Distributions of Species")
```



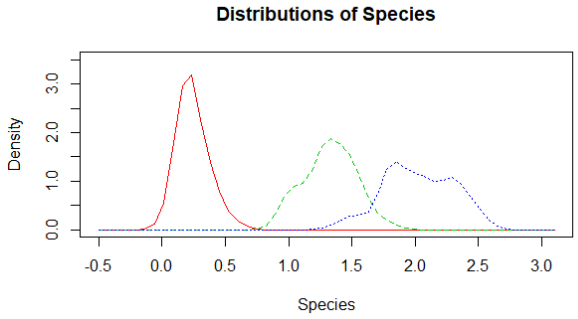

```
sm.density.compare(iris$sepal.width, iris$variety, xlab="Species")
title(main="Distributions of Species")
```




```
sm.density.compare(iris$petal.length, iris$variety, xlab="Species")
title(main="Distributions of Species")
```

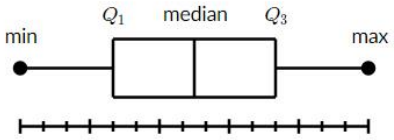



```
sm.density.compare(iris$petal.width, iris$variety, xlab="Species")
title(main="Distributions of Species")
```



iv. Generate box plots for each of the numerical attributes. Identify the attribute with the highest variance.

The five-number summary is the minimum, first quartile, median, third quartile, and maximum. In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median. The whiskers go from each quartile to the minimum or maximum (5).



Example:

A sample of 101010 boxes of raisins has these weights (in grams):

25, 28, 29, 29, 30, 34, 35, 35, 37, 38

1. Order the data from smallest to largest.

   25, 28, 29, 29, 30, 34, 35, 35, 37, 38

2. Find the median

   The median is the mean of the middle two numbers:

   25, 28, 29, 29, **30, 34**, 35, 35, 37, 38

   $$\frac{30 + 34}{2} = 32$$

3. Find the quartile

   First Quartile Q_1: The first quartile is the median of the data points to the left of the median.

   25, 28, **29**, 29, 30

   Q_1=29

   Third Quartile Q_3: The third quartile is the median of the data points to the right of the median.

   34, 35, **35**, 37, 38

   Q_3=29

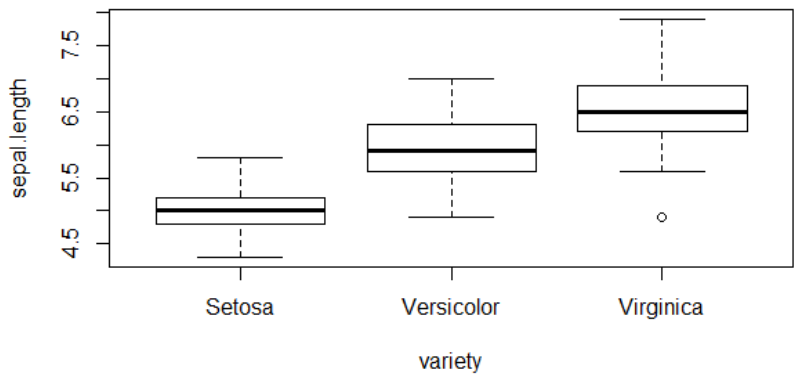
4. Complete the five-number summary by finding the min and the max.

   The min is the smallest data point, which is 25.

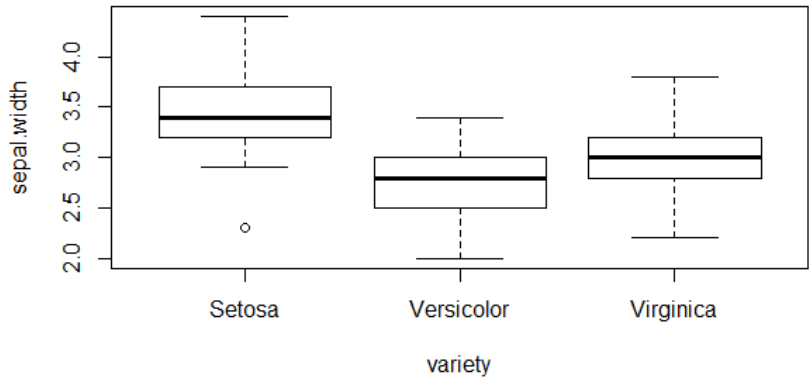
   The max is the largest data point, which is 38.

   The five-number summary is 25, 29, 32, 35, 38.

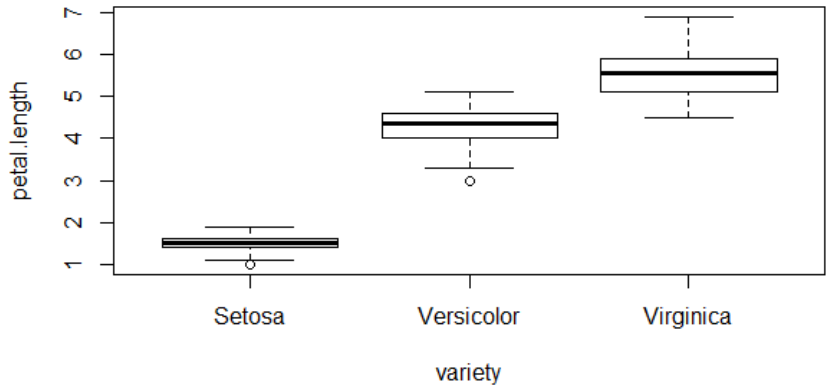
```
boxplot(sepal.length ~ variety, data=iris)
```



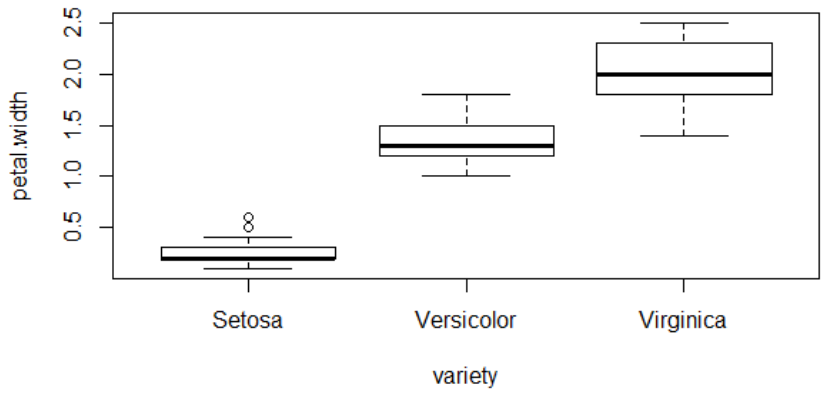

```
boxplot(sepal.width ~ variety, data=iris)
```




```
boxplot(petal.length ~ variety, data=iris)
```




```
boxplot(petal.width ~ variety, data=iris)
```



References

1. FISHER RA. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. Ann Eugen [Internet]. 1936 Sep;7(2):179–88. Available from: http://doi.wiley.com/10.1111/j.1469-1809.1936.tb02137.x
2. Variance and standard deviation [Internet]. Available from: https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch12/5214891-eng.htm#a3
3. Klodian Dh. How to make Histogram with R [Internet]. Available from: https://datascienceplus.com/histogram-with-r/
4. Klodian Dh. How to Compare Distribution by Using Density Plots in R [Internet]. Available from: https://datascienceplus.com/compare-distribution-by-density-plots/
5. Boxplot review [Internet]. Available from: https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/box-plot-review