DS501 STATISTICAL AND MATHEMATICAL METHODS FOR DATA SCIENCE

Dr. Muhammad Wasim

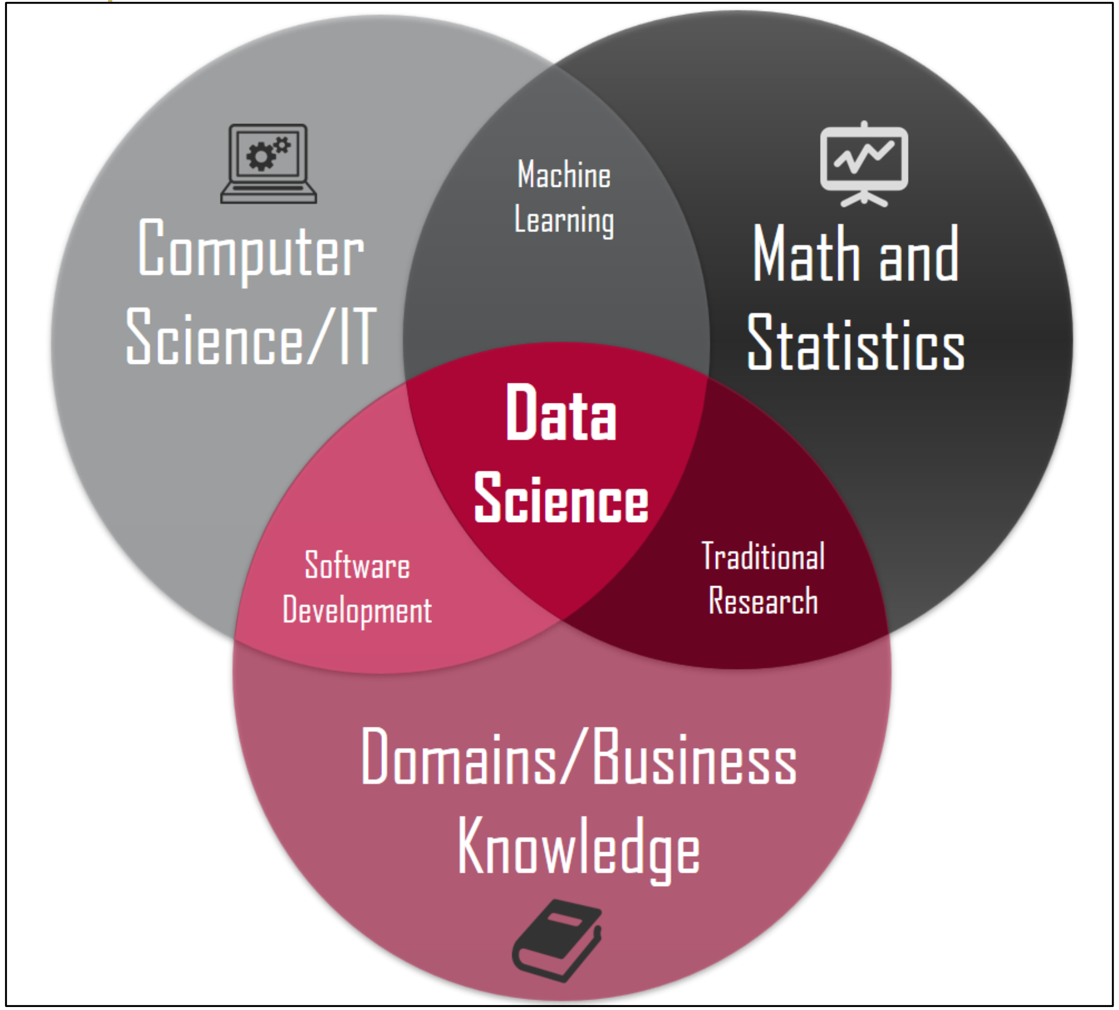
PhD, MS, M.Phil, M.Sc, MCS

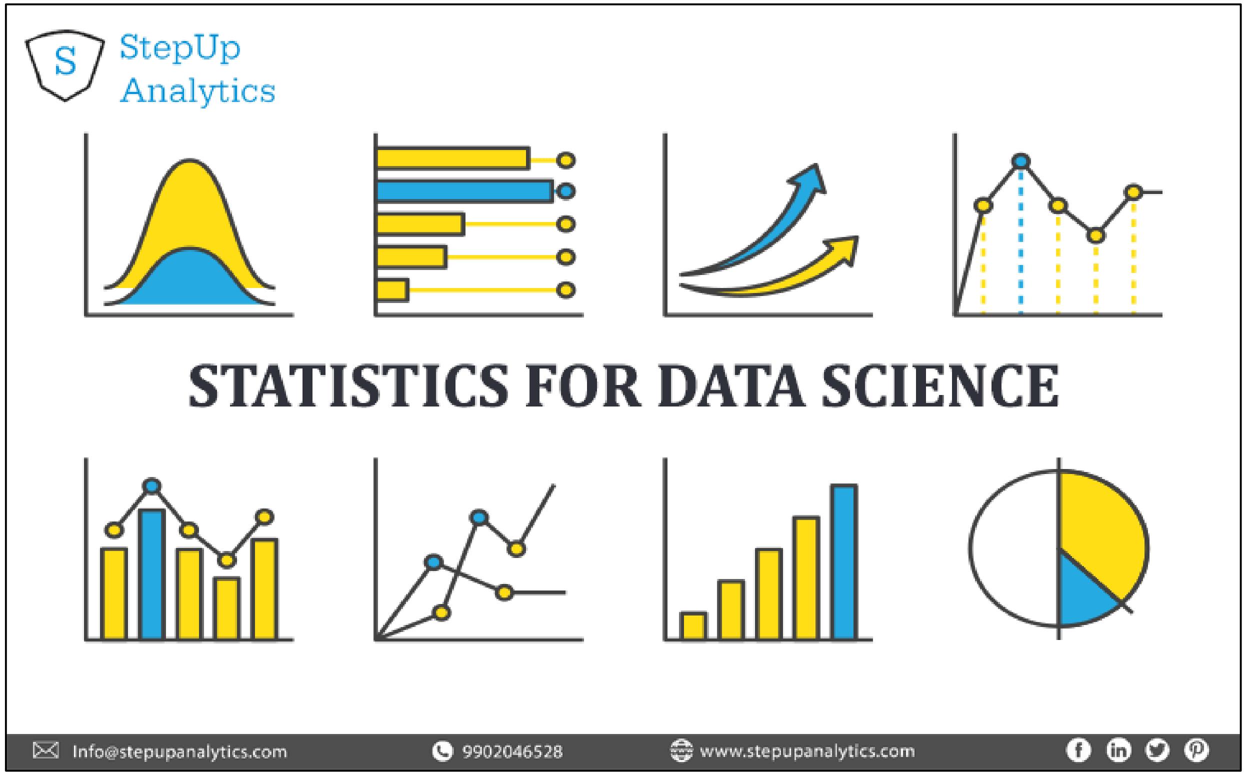
Certified Data Analyst [KARACHI.AI]

Lecture Week 05

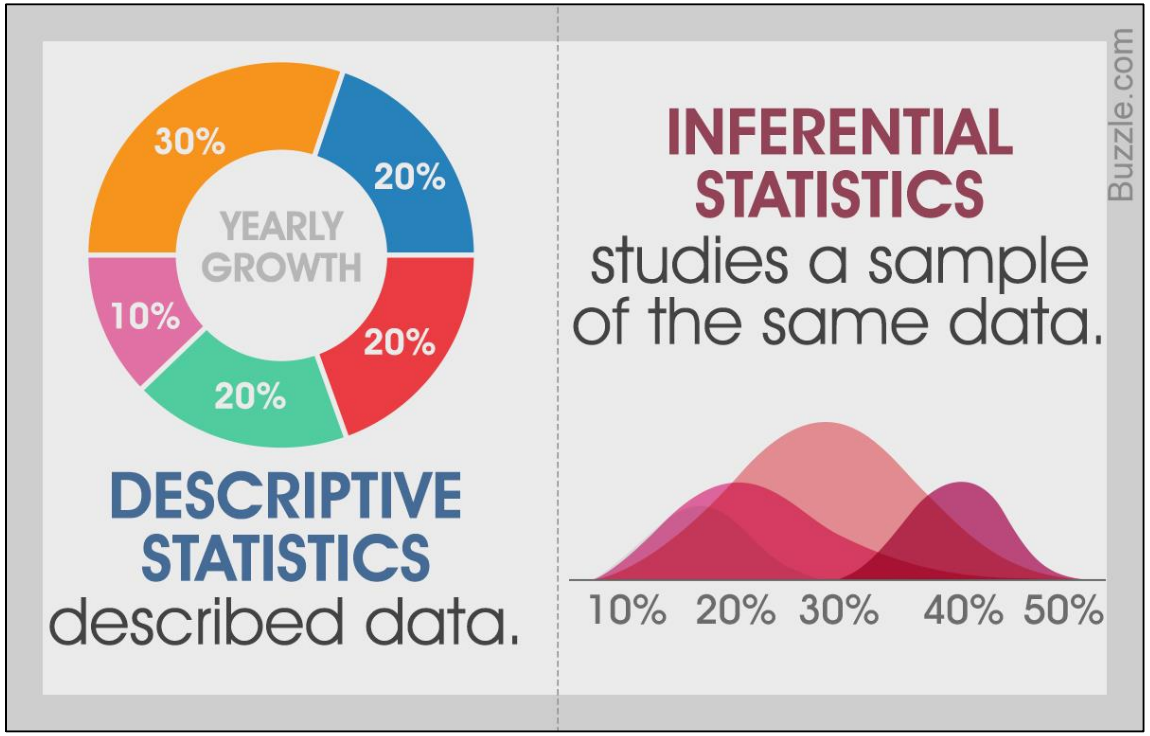
- Statistics for Data Science

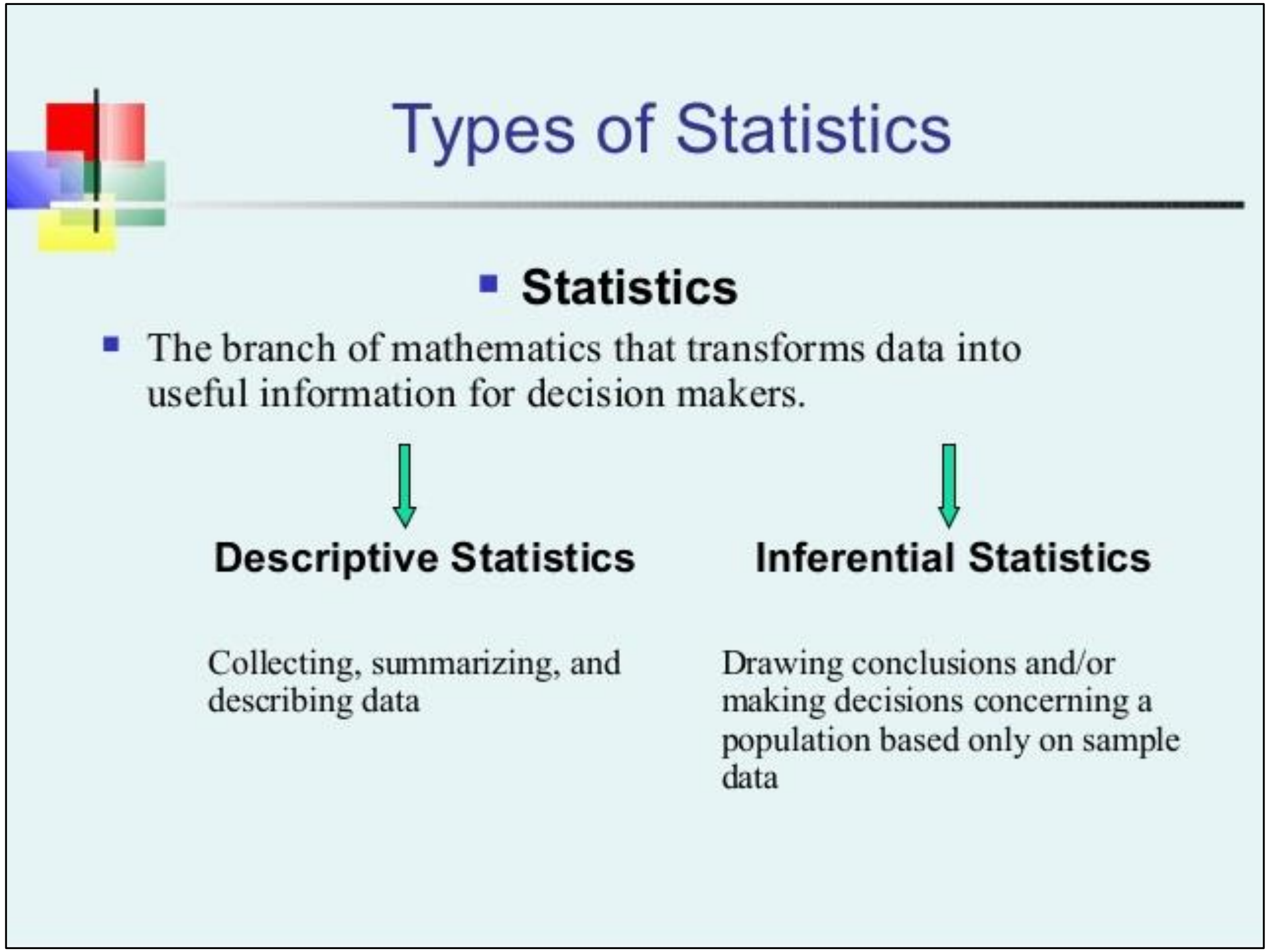
STATISTICS FOR DATA SCIENCE





STATISTICS FOR DATA SCIENCE





Types of Statistics

- **Statistics**

- The branch of mathematics that transforms data into useful information for decision makers.

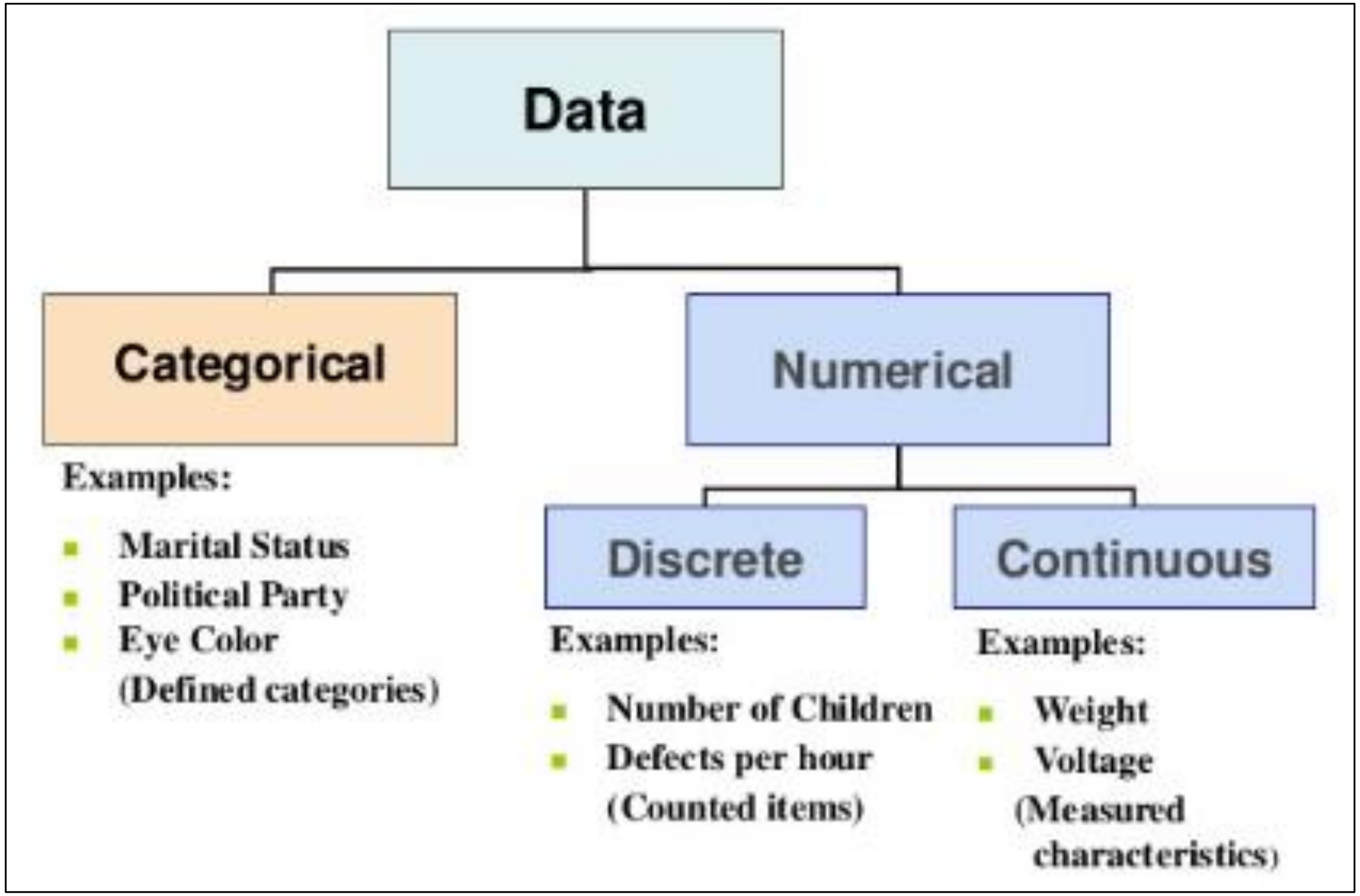
Descriptive Statistics

Collecting, summarizing, and describing data

Inferential Statistics

Drawing conclusions and/or making decisions concerning a population based only on sample data

STATISTICS FOR DATA SCIENCE



Data

Categorical

Examples:

- Marital Status
- Political Party
- Eye Color

(Defined categories)

Numerical

Discrete

Examples:

- Number of Children
- Defects per hour

(Counted items)

Continuous

Examples:

- Weight
- Voltage

(Measured characteristics)

STATISTICS FOR DATA SCIENCE

Measures Of The Spread:

Just like the measure of center, we also have measures of the spread, which comprises of the following measures:

Range: It is the given measure of how spread apart the values in a data set are.

Inter Quartile Range (IQR): It is the measure of variability, based on dividing a data set into quartiles.

Variance: It describes how much a random variable differs from its expected value. It entails computing squares of deviations.

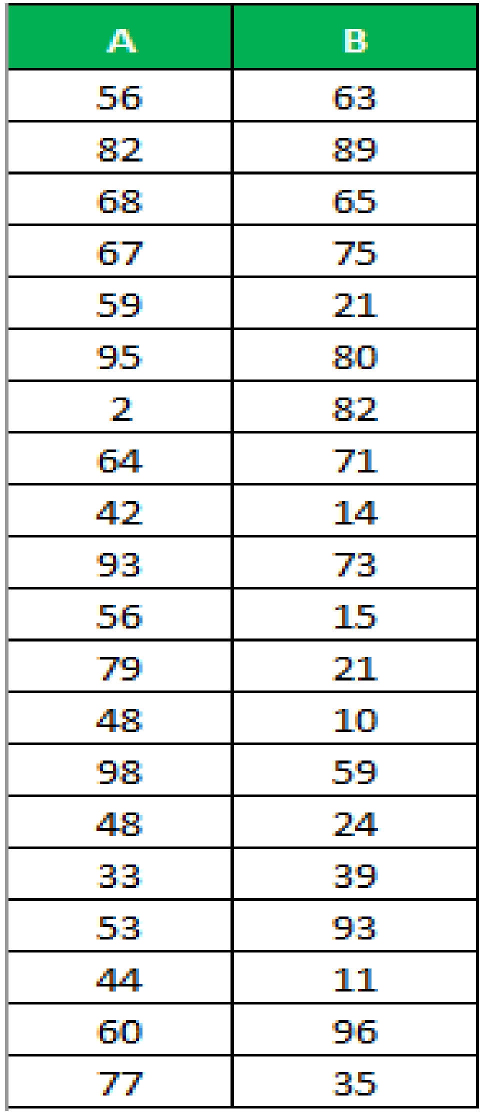
- ***Deviation*** *is the difference between each element from the mean.*
- ***Population Variance*** *is the average of squared deviations*
- ***Sample Variance*** *is the average of squared differences from the mean*

Standard Deviation: It is the measure of the dispersion of a set of data from its mean.

STATISTICS FOR DATA SCIENCE

1. Which class (A or B) is more consistent. How to study the variables by plotting a histogram

2. Write a python code for the solution with data visualizations.



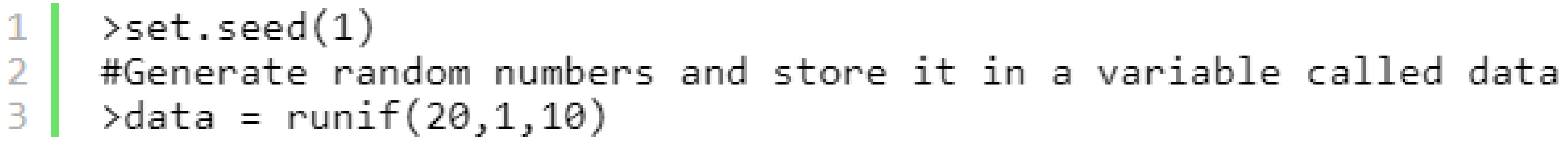
A	B
56	63
82	89
68	65
67	75
59	21
95	80
2	82
64	71
42	14
93	73
56	15
79	21
48	10
98	59
48	24
33	39
53	93
44	11
60	96
77	35

Resource Link:

https://www.datacamp.com/tracks/statistics-fundamentals-with-python

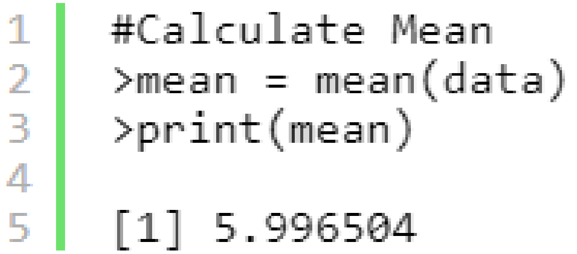
STATISTICS FOR DATA SCIENCE

Step 1: Import data for computation



```
>set.seed(1)
#Generate random numbers and store it in a variable called data
>data = runif(20,1,10)
```

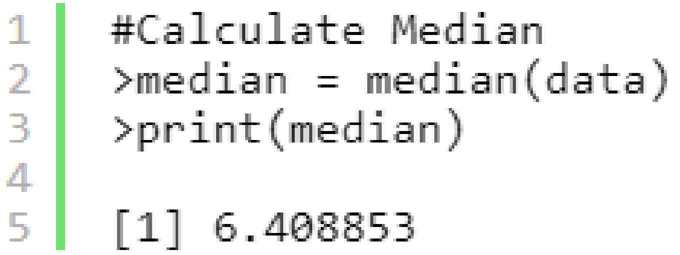
Step 2: Calculate Mean for the data



```
#Calculate Mean
>mean = mean(data)
>print(mean)

[1] 5.996504
```

Step 3: Calculate the Median for the data

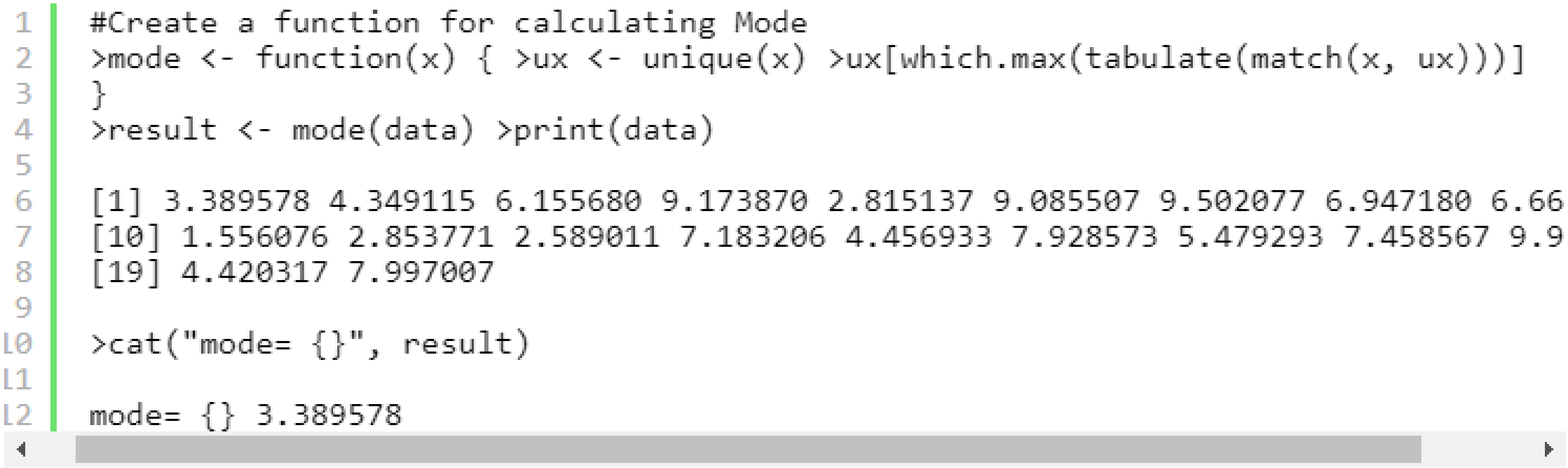


```
#Calculate Median
>median = median(data)
>print(median)

[1] 6.408853
```

STATISTICS FOR DATA SCIENCE

Step 4: Calculate Mode for the data



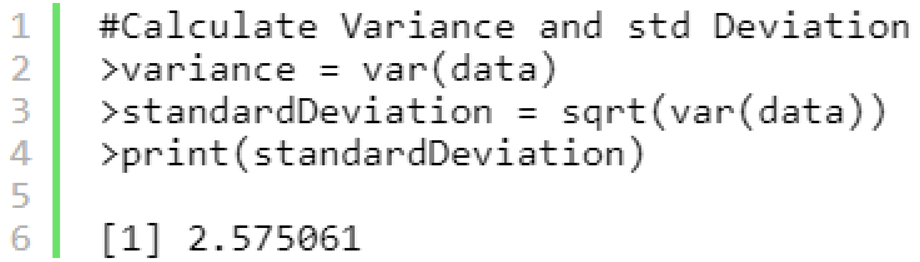
```
#Create a function for calculating Mode
>mode <- function(x) { >ux <- unique(x) >ux[which.max(tabulate(match(x, ux)))]
}
>result <- mode(data) >print(data)

[1] 3.389578 4.349115 6.155680 9.173870 2.815137 9.085507 9.502077 6.947180 6.66
[10] 1.556076 2.853771 2.589011 7.183206 4.456933 7.928573 5.479293 7.458567 9.9
[19] 4.420317 7.997007

>cat("mode= {}", result)

mode= {} 3.389578
```

Step 5: Calculate Variance & Std Deviation for the data

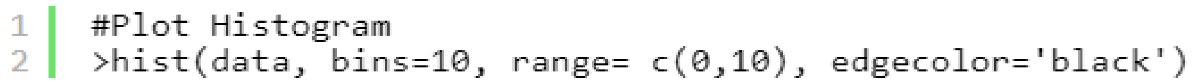


```
#Calculate Variance and std Deviation
>variance = var(data)
>standardDeviation = sqrt(var(data))
>print(standardDeviation)

[1] 2.575061
```

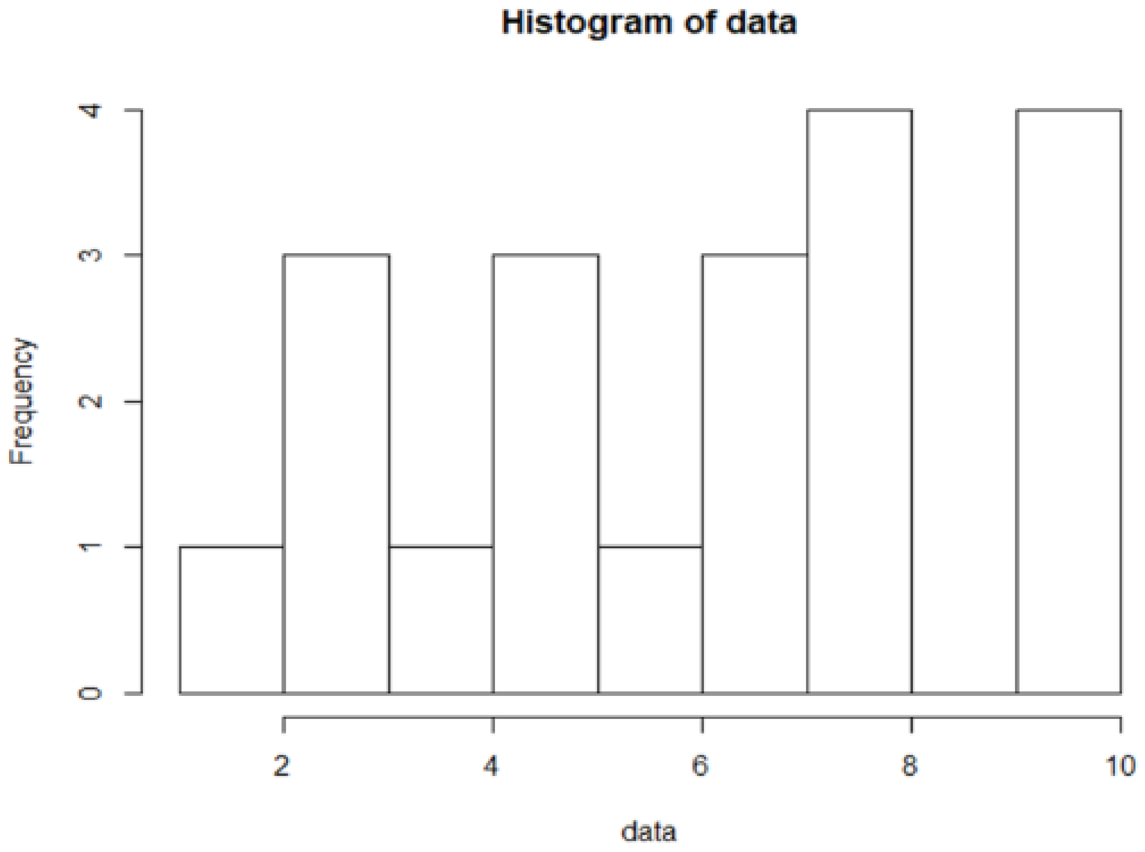
STATISTICS FOR DATA SCIENCE

Step 6: Plot a Histogram



```
#Plot Histogram
>hist(data, bins=10, range= c(0,10), edgecolor='black')
```

The Histogram is used to display the frequency of data points:


Histogram of data
Frequency
data