Notes on Basics of Statistics:

It is the science of collecting, organizing, summarizing, and analyzing data to get meaningful insight from it to draw an optimal conclusion.

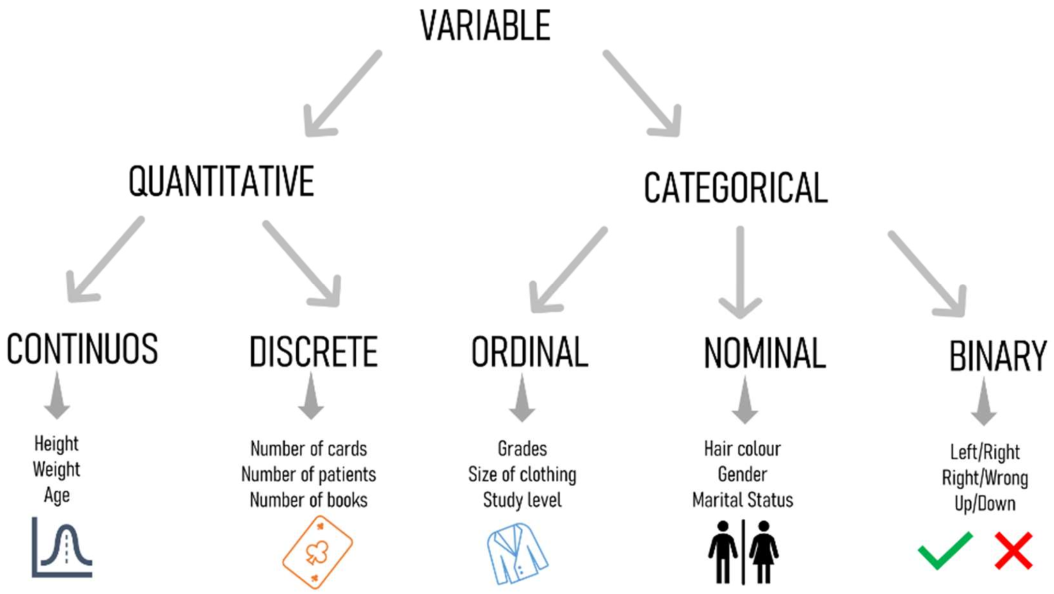
Type of Statistics:

1. **Descriptive Statistics:**
   This includes an analysis of the data at hand to understand the expected behavior using measures of central tendency (mean, median, mode, variance, standard deviation, skewness, and kurtosis).
2. **Inferential Statistics:**

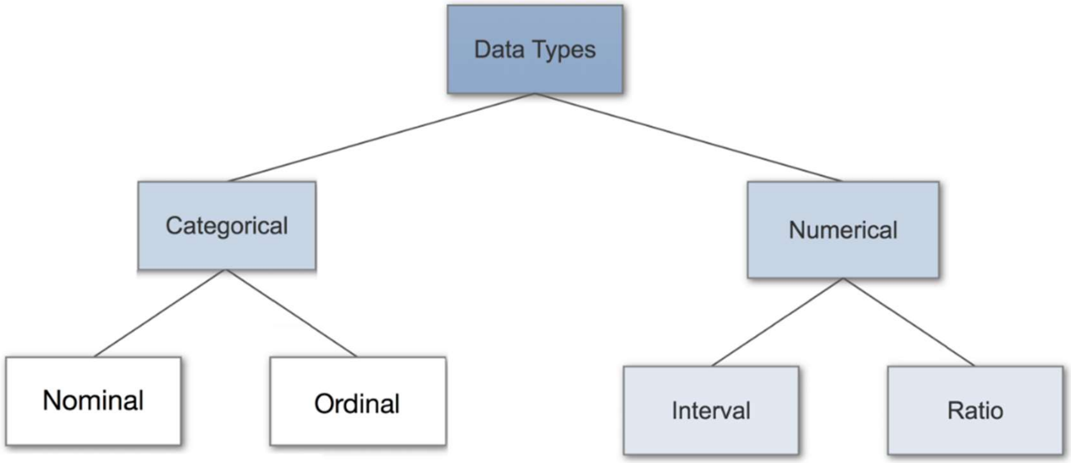
   This includes the selection of samples from the population to the analysis of said samples to draw conclusions about the population.

Variables & Its type:

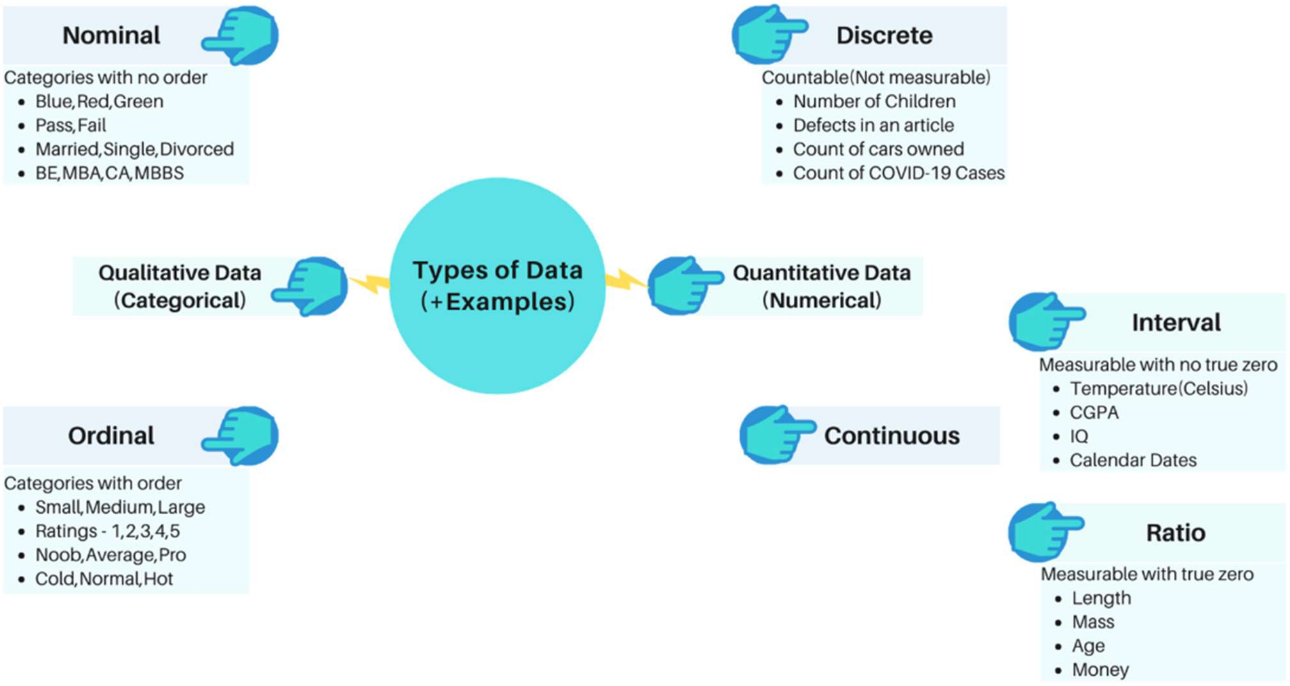
A variable is defined as the alphabetic character that expresses a numerical value or a number. In algebraic equations, a variable is used to represent an unknown quantity or quality of data.



Data Types:



Understanding Data and Datatype:


Types of Data (+Examples)
Qualitative Data (Categorical)
Quantitative Data (Numerical)
Continuous

Nominal

Categories with no order

- Blue,Red,Green
- Pass,Fail
- Married,Single,Divorced
- BE,MBA,CA,MBBS

Ordinal

Categories with order

- Small,Medium,Large
- Ratings - 1,2,3,4,5
- Noob,Average,Pro
- Cold,Normal,Hot

Discrete

Countable(Not measurable)

- Number of Children
- Defects in an article
- Count of cars owned
- Count of COVID-19 Cases

Interval

Measurable with no true zero

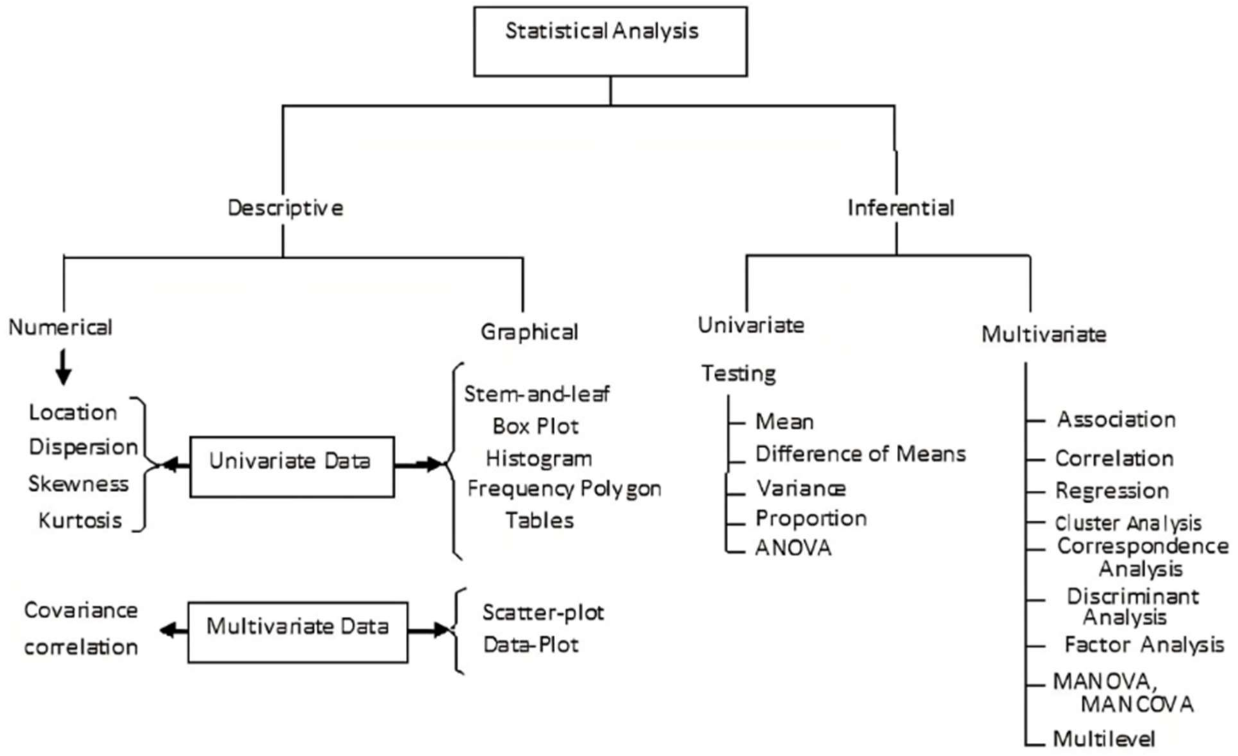
- Temperature(Celsius)
- CGPA
- IQ
- Calendar Dates

Ratio

Measurable with true zero

- Length
- Mass
- Age
- Money

Statistical Analysis



The measure of central Tendency

- **Mean** (Mean affected by magnitude (outlier)
  - Average of data point (sum of data point/ number of data points)
  - Affected by outlier

  Mean of a sample:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

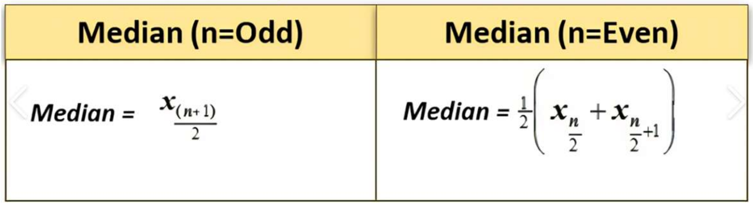
- **Median**
  - This can be defined as Middle positional Value of given data set and can be derived in two steps,

    (1) first by Arranging dataset (n) in ascending order then

    (2) **for odd numbers of observations,** by selecting the observation at (n+1)/2 positions for an odd number of observation

For an even number of observation, calculating the average of two data points at nth/2 term and (n+1)/2 term,
[Median Formula | How To Calculate Median (Calculator, Excel Template) (educba.com)]

- Not affected by outlier



Median (n=Odd)	Median (n=Even)
Median = $x_{\frac{(n+1)}{2}}$	*Median* = $\frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right)$

- **Mode:** Mode is one of the values that indicate a central tendency of a set of data. Mode or modal value gives us an idea about which of the items in a data set is more likely to occur frequently. It is the measure of Central Tendency other than Mean and Median.
- Depending on the type of dataset given, you can find one, two three or even multiple modal values. Some data sets may have no mode value at all.
  - Most occurring value in the data set
  - Can be unimodal, bimodal, multimodal

$$\text{Mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

Where,

l = lower limit of the modal class
h = size of the class interval
f_1 = frequency of the modal class
f_0 = frequency of the class preceding the modal class
f_2 = frequency of the class succeeding the modal class

([Mode Formula: Learn to Calculate Mode For Set of Data - Embibe])

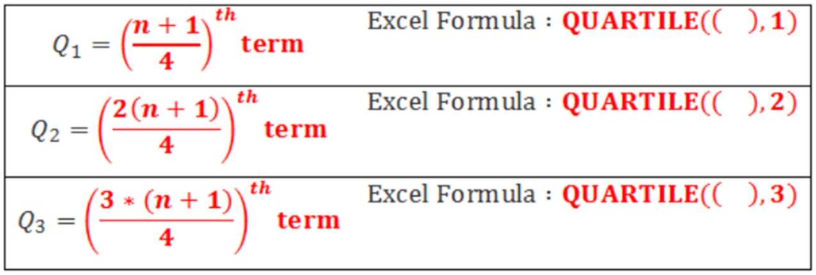
- Shows Expected Behavior of Data Set at the center

Spread (Overall spread of Data)

- **Range (Max-Min)**
  - The range for any distribution is given by = highest value – lowest value.
    Range = Highest Value – Lowest Value
  - Doesn't show variation in between

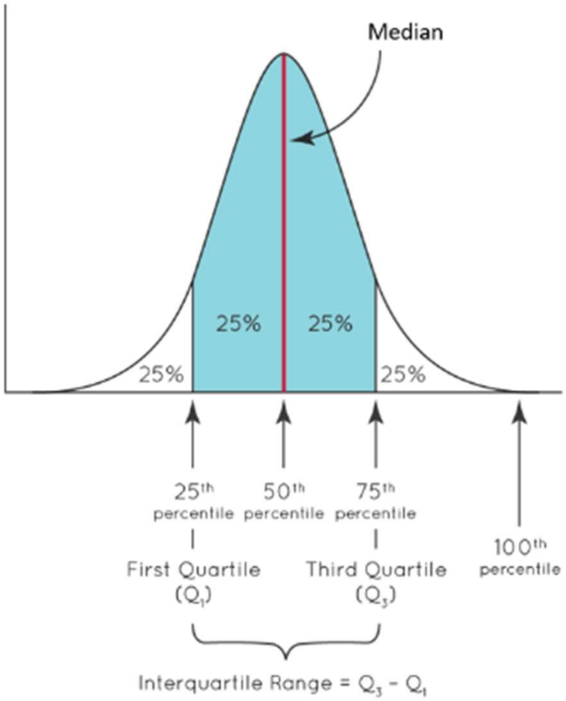
IQR (Inter Quartile Range)

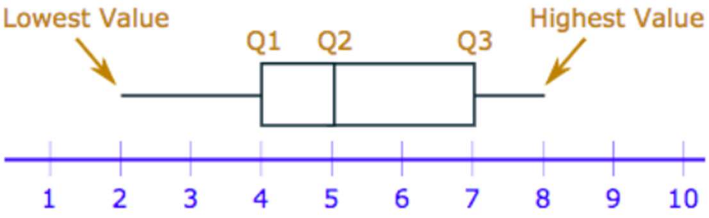
When the set of observations are arranged in ascending order the quartiles are represented as,



$Q_1 = \left(\frac{n+1}{4}\right)^{th}$ term	Excel Formula : QUARTILE(( ), 1)
$Q_2 = \left(\frac{2(n+1)}{4}\right)^{th}$ term	Excel Formula : QUARTILE(( ), 2)
$Q_3 = \left(\frac{3*(n+1)}{4}\right)^{th}$ term	Excel Formula : QUARTILE(( ), 3)

The interquartile range (IQR) = Upper Quartile (Q3) – Lower Quartile (Q1)
Lower Bound Limit = Q1 - 1.5 x IQR
Upper Bound Limit = Q3 + 1.5 x IQR





Measures of Dispersion:

- **Variance :**

  Average squared deviation of the data points from their mean.

  - Affected by extreme

Population variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

where μ is the population mean.

Sample variance:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

- **Standard Deviation (SD) : How data variating inside the range**

  Sample Standard Deviation:

$$\sigma = \sqrt{\frac{\Sigma\,(x - u)^2}{N}}$$

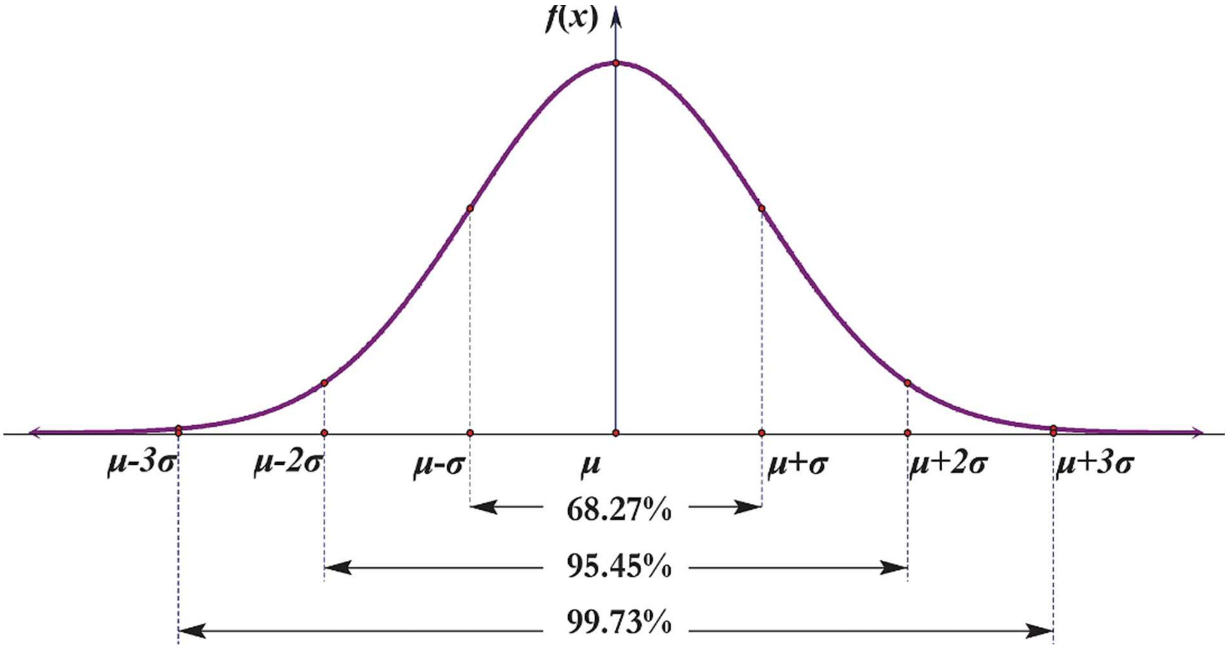
$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

(Note: for sample data n-1 is considered for **Basel correction**)

$1\sigma = 68.26\ \%$, Data spread both lef and right from center is 68.26 %

$2\sigma = 95.44\ \%$, Data spread both lef and right from center is 95.44 %

$3\sigma = 99.74\%$, Data spread both lef and right from center is 97.74 %

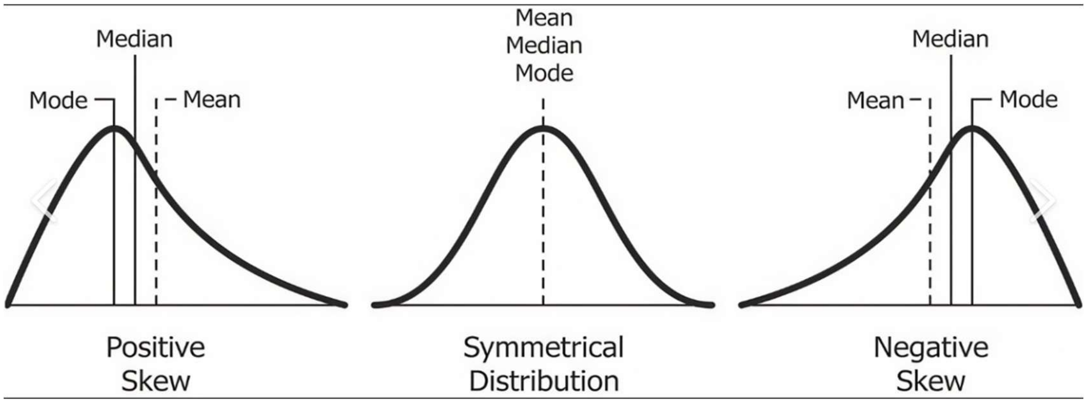


The larger the value of *n* is, the closer that the population and sample standard deviations will be.

Shape of Distribution

- **Symmetricity / Normal distribution**

  Mean = Median = Mode



$$\text{Skewness} = \frac{\sum_i^N (X_i - \bar{X})^3}{(N-1) * \sigma^3}$$

Skewness (skew: where it has been flattened)

- Positive skewed **(mean>median)/mean-median= + ve** (i.e. Salary in an Organization) - flattened at right
- **Negative Skewed (Median > Mean)/mean-median= - ve** (i.e. easy Exam)

Kurtosis: (Strength of Relationship with Standard Deviation)

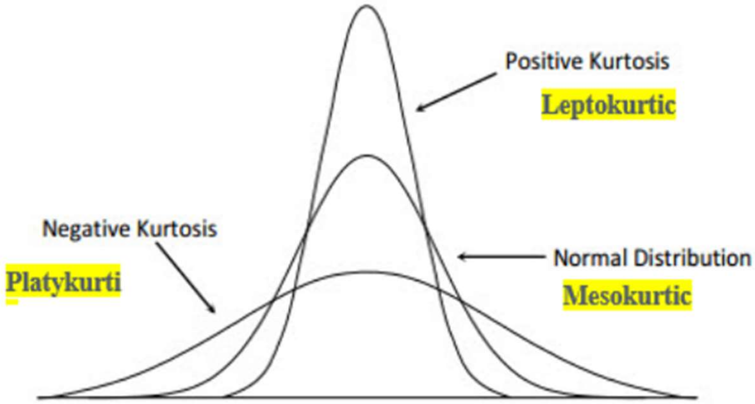
It is statistical measures which define shape of each tails of distribution where it is heavy tailed (presence of outliers) or light tailed (paucity of outliers) compared to normal distribution.

If **Kurtosis > ±3 SD then it is to be called Leptokurtic** (short tailed) with Low Standard Deviation

if **Kurtosis < ±3 SD then it is called Platykurtic** (long Tailed) with High Standard Deviation

If Kurtosis = ± 3 SD then it is called Mesokurtic

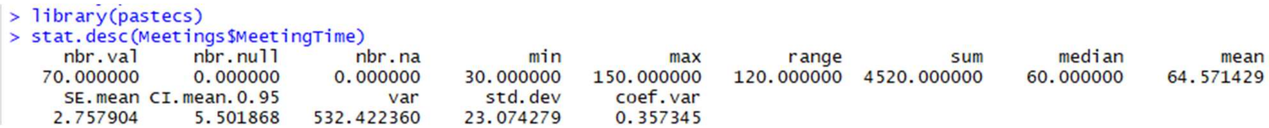
$$\text{Kurtosis} = n * \frac{\sum_i^n (Y_i - \bar{Y})^4}{\sum_i^n (Y_i - \bar{Y}^2)^2}$$



Getting Major Statistical Details in Excel / Python / R:

R-Program / Library (Pastecs)

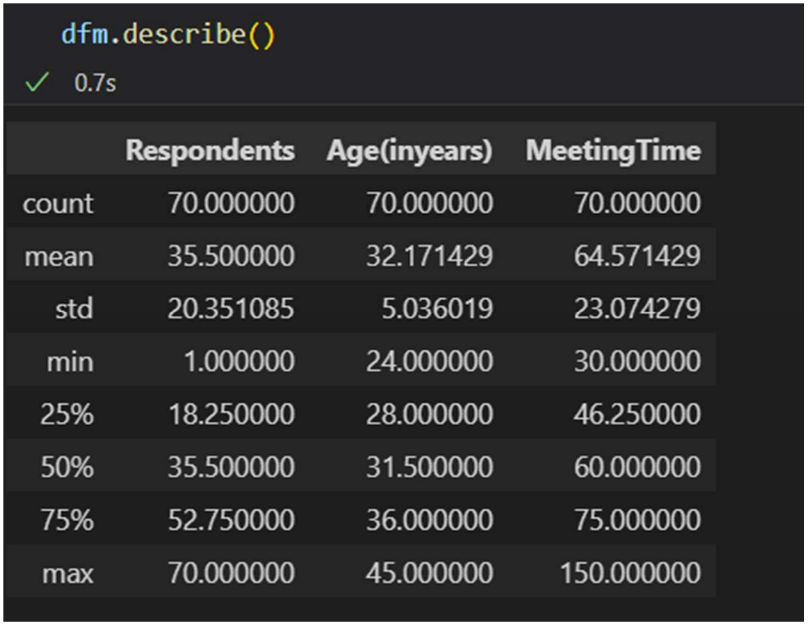
stat.desc(filename$column_name) to get all the statistical measures:



```
> library(pastecs)
> stat.desc(Meetings$MeetingTime)
     nbr.val     nbr.null       nbr.na          min          max        range          sum       median         mean 
   70.000000     0.000000     0.000000    30.000000   150.000000   120.000000  4520.000000    60.000000    64.571429 
     SE.mean CI.mean.0.95          var      std.dev     coef.var 
    2.757904     5.501868   532.422360    23.074279     0.357345 
```

Python /library (pandas)

df. describe() to get all the statistical measures:

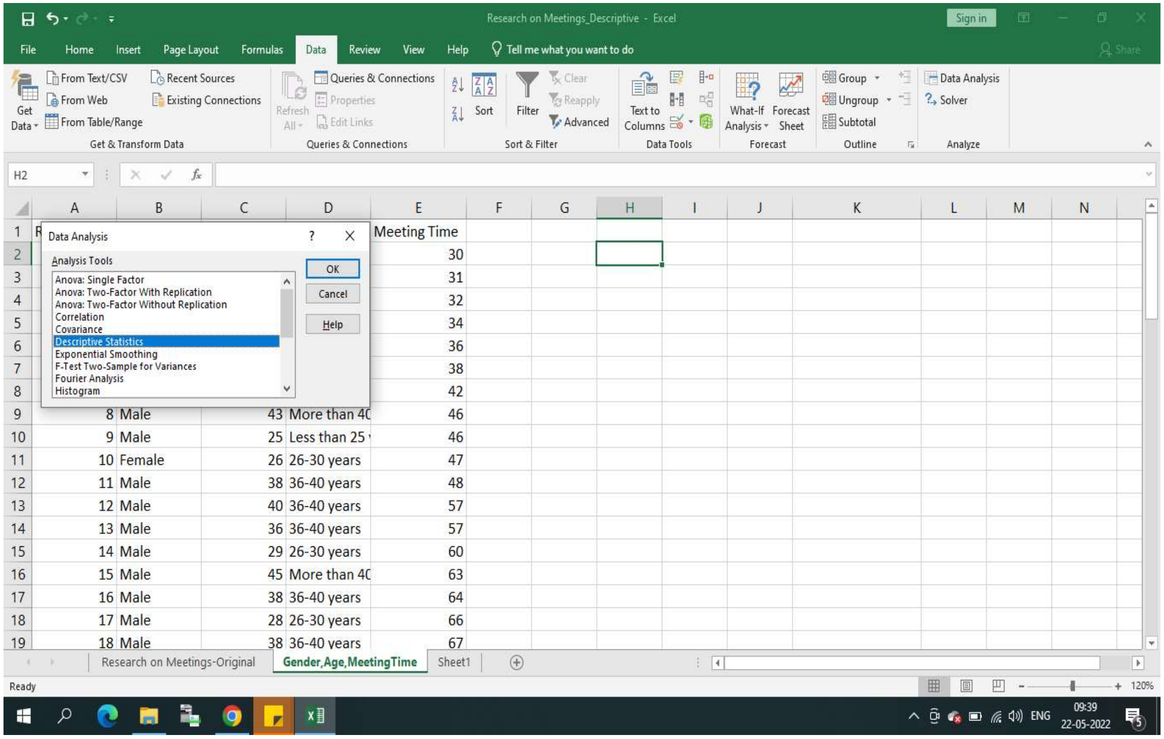


```
dfm.describe()
```

✓ 0.7s

	Respondents	Age(inyears)	MeetingTime
count	70.000000	70.000000	70.000000
mean	35.500000	32.171429	64.571429
std	20.351085	5.036019	23.074279
min	1.000000	24.000000	30.000000
25%	18.250000	28.000000	46.250000
50%	35.500000	31.500000	60.000000
75%	52.750000	36.000000	75.000000
max	70.000000	45.000000	150.000000

Step-1 (Importing Data Analysis and selecting Descriptive Statistics)



Data Analysis

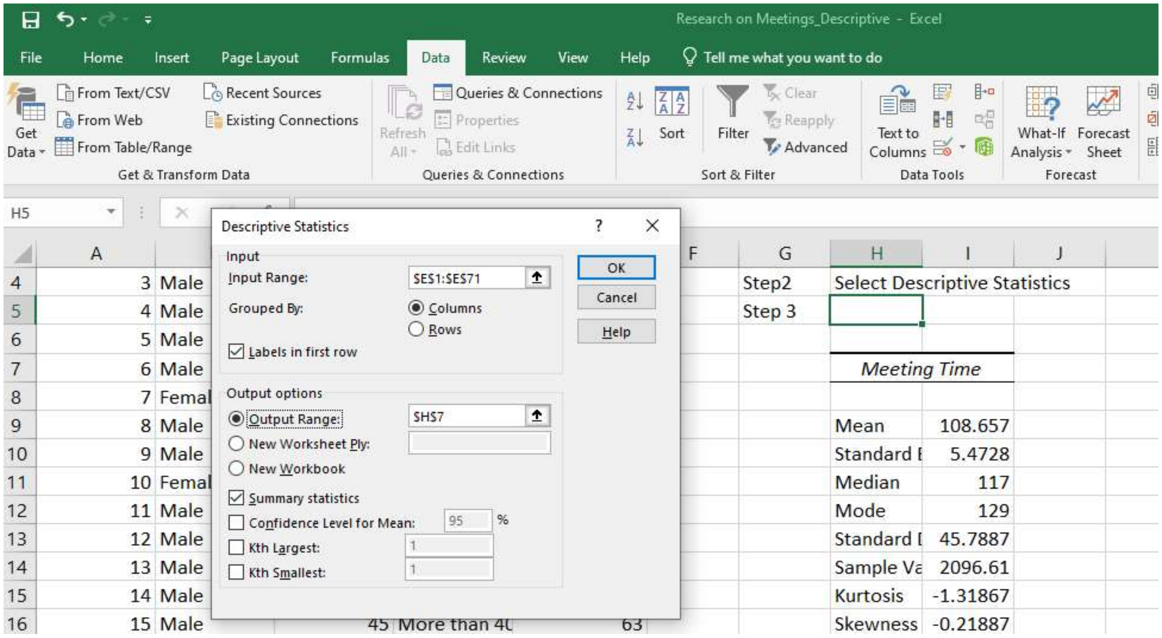
Analysis Tools

- Anova: Single Factor
- Anova: Two-Factor With Replication
- Anova: Two-Factor Without Replication
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram

OK | Cancel | Help

	A	B	C	D	E
1	R				Meeting Time
2					30
3					31
4					32
5					34
6					36
7					38
8					42
9	8	Male	43	More than 4C	46
10	9	Male	25	Less than 25 '	46
11	10	Female	26	26-30 years	47
12	11	Male	38	36-40 years	48
13	12	Male	40	36-40 years	57
14	13	Male	36	36-40 years	57
15	14	Male	29	26-30 years	60
16	15	Male	45	More than 4C	63
17	16	Male	38	36-40 years	64
18	17	Male	28	26-30 years	66
19	18	Male	38	36-40 years	67

Step-2 (selecting the input and output columns)



Descriptive Statistics

Input

Input Range: $E$1:$E$71

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Output options

● Output Range: $H$7

○ New Worksheet Ply:

○ New Workbook

☑ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

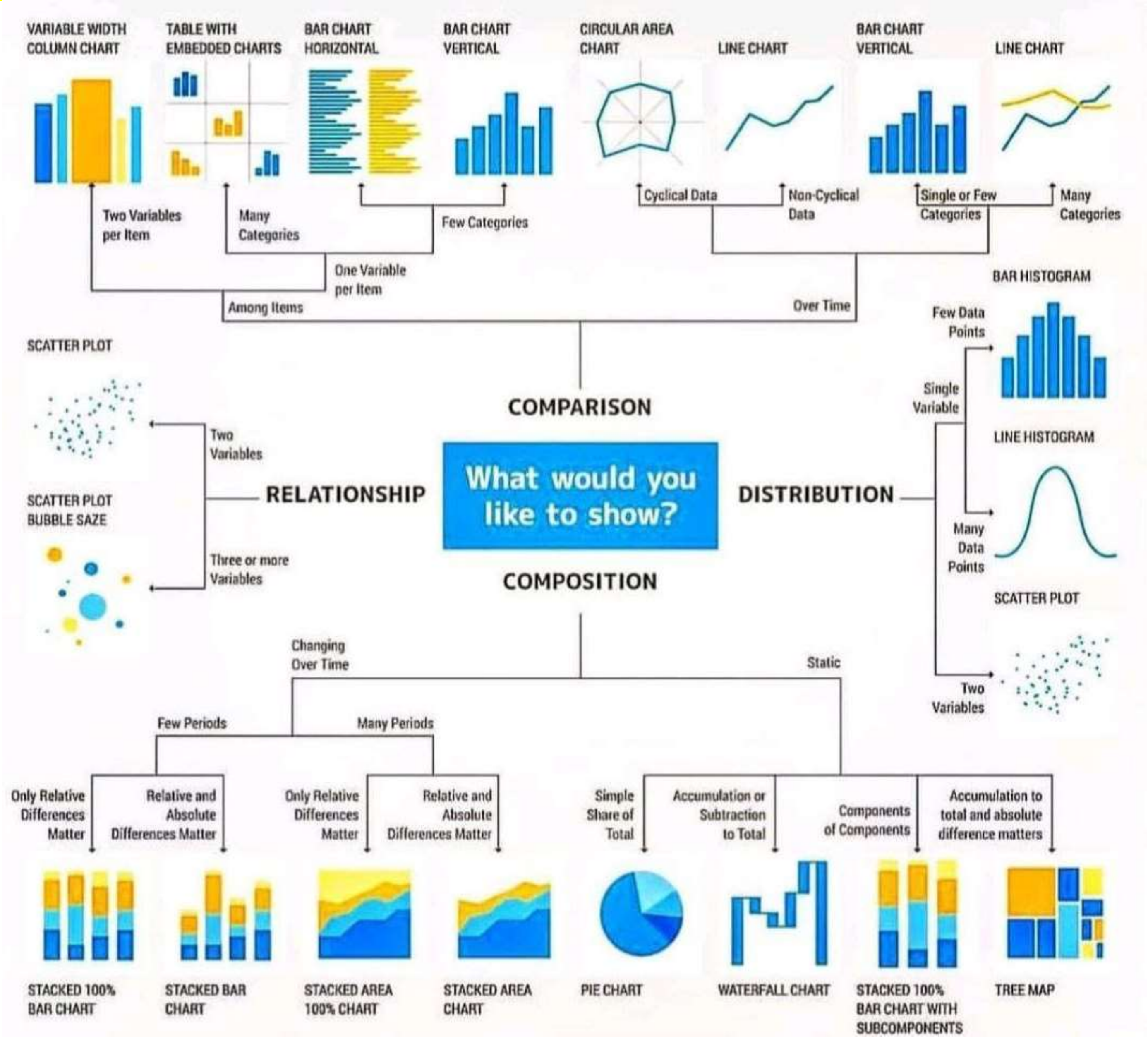
□ Kth Smallest: 1

OK | Cancel | Help

	A	B	G	H	I
4	3	Male	Step2	Select Descriptive Statistics	
5	4	Male	Step 3		
6	5	Male			
7	6	Male		*Meeting Time*	
8	7	Femal			
9	8	Male		Mean	108.657
10	9	Male		Standard I	5.4728
11	10	Femal		Median	117
12	11	Male		Mode	129
13	12	Male		Standard I	45.7887
14	13	Male		Sample Va	2096.61
15	14	Male		Kurtosis	-1.31867
16	15	Male		Skewness	-0.21887

45 More than 4C 63

Type of Plots


What would you like to show?
COMPARISON
Among Items
Two Variables per Item
VARIABLE WIDTH COLUMN CHART
One Variable per Item
Many Categories
TABLE WITH EMBEDDED CHARTS
Few Categories
BAR CHART HORIZONTAL
BAR CHART VERTICAL
Over Time
Cyclical Data
CIRCULAR AREA CHART
Non-Cyclical Data
LINE CHART
Single or Few Categories
BAR CHART VERTICAL
Many Categories
LINE CHART
RELATIONSHIP
Two Variables
SCATTER PLOT
Three or more Variables
SCATTER PLOT BUBBLE SAZE
DISTRIBUTION
Single Variable
Few Data Points
BAR HISTOGRAM
Many Data Points
LINE HISTOGRAM
Two Variables
SCATTER PLOT
COMPOSITION
Changing Over Time
Few Periods
Only Relative Differences Matter
STACKED 100% BAR CHART
Relative and Absolute Differences Matter
STACKED BAR CHART
Many Periods
Only Relative Differences Matter
STACKED AREA 100% CHART
Relative and Absolute Differences Matter
STACKED AREA CHART
Static
Simple Share of Total
PIE CHART
Accumulation or Subtraction to Total
WATERFALL CHART
Components of Components
STACKED 100% BAR CHART WITH SUBCOMPONENTS
Accumulation to total and absolute difference matters
TREE MAP