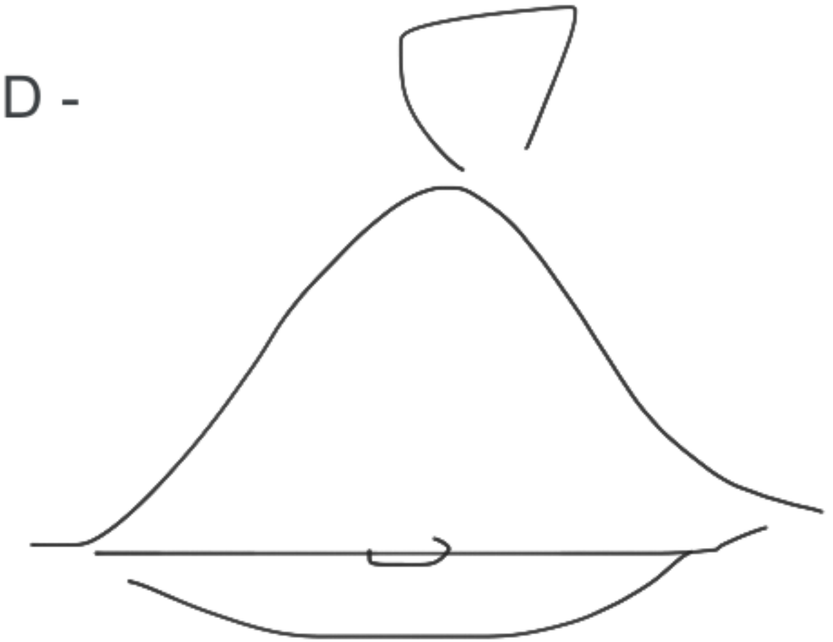
Stats

Type of Analysis
Types of Stats

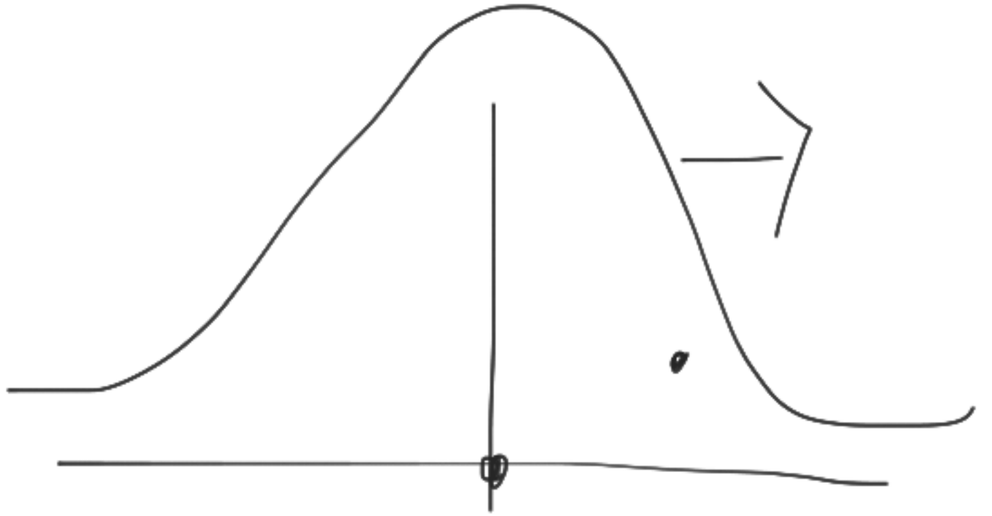
Measure of central Tendency - Mean, Median, Mode

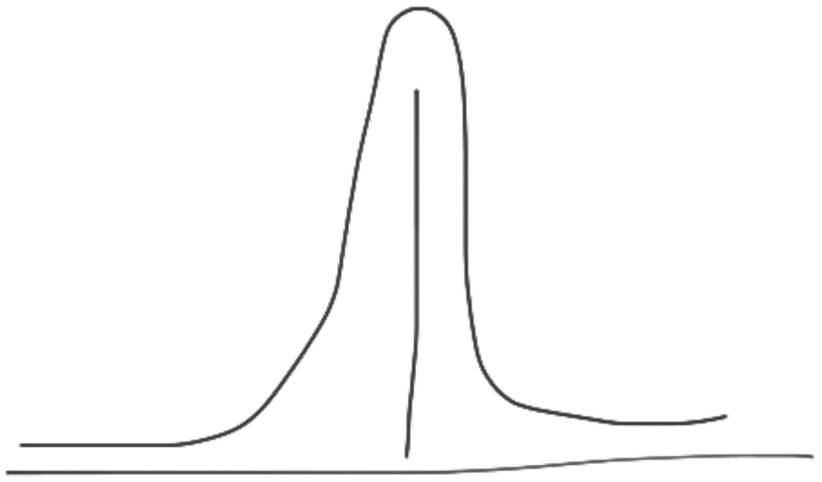
SD -

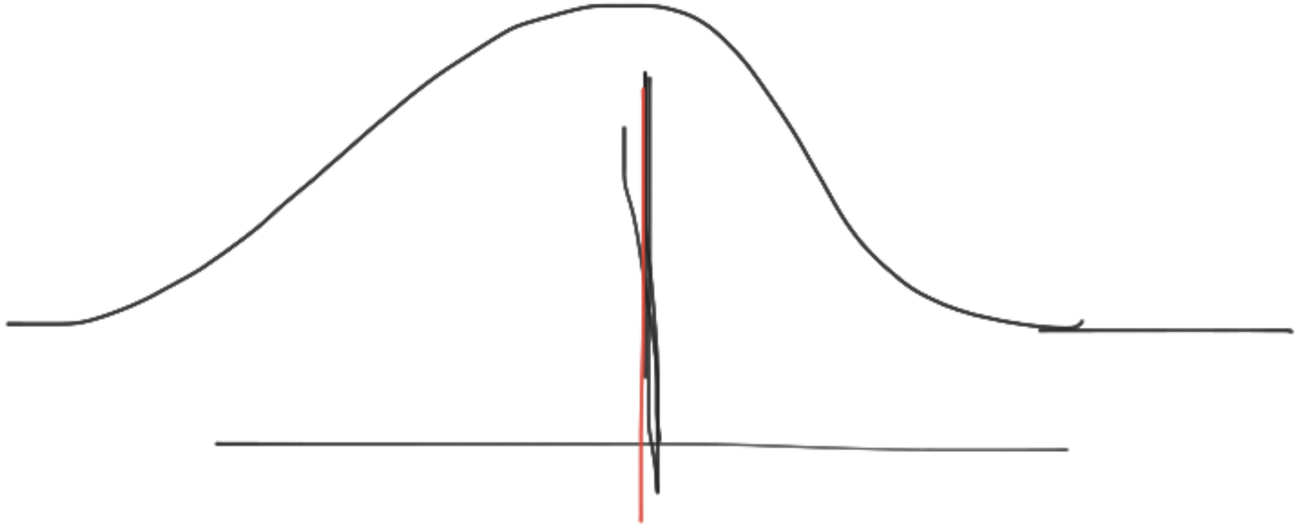


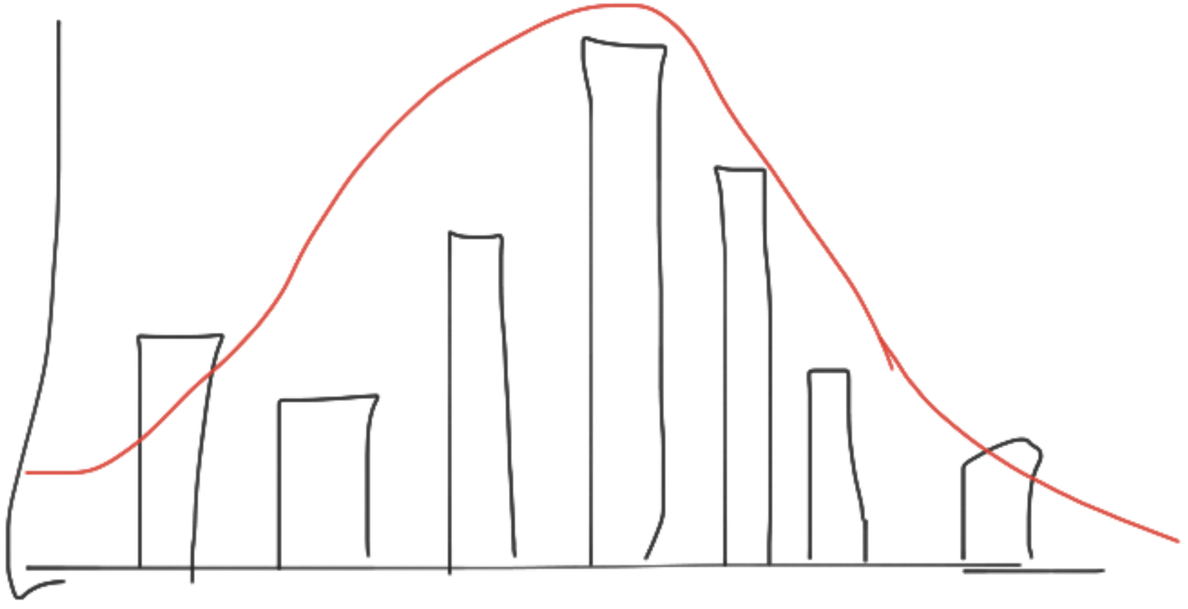
Measure of Variance

Vary

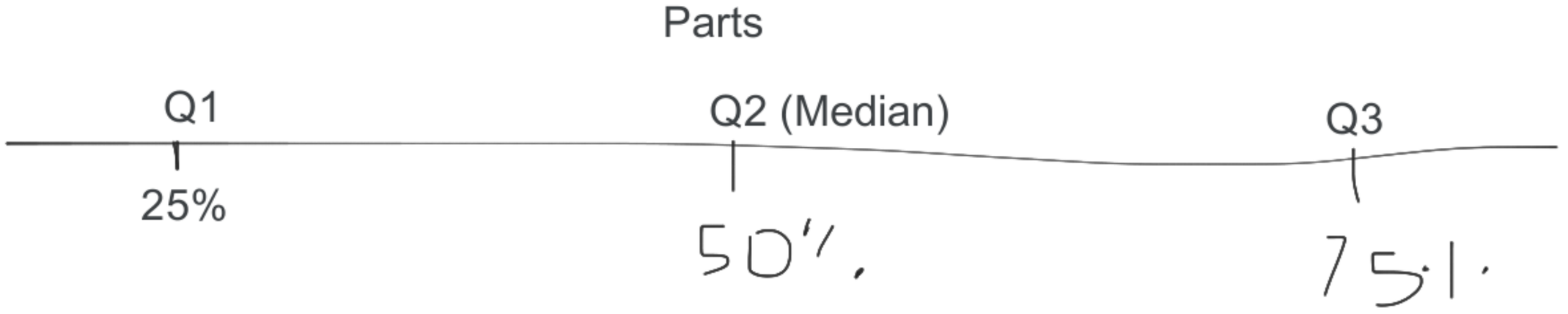








Inter Quartile Range IQR



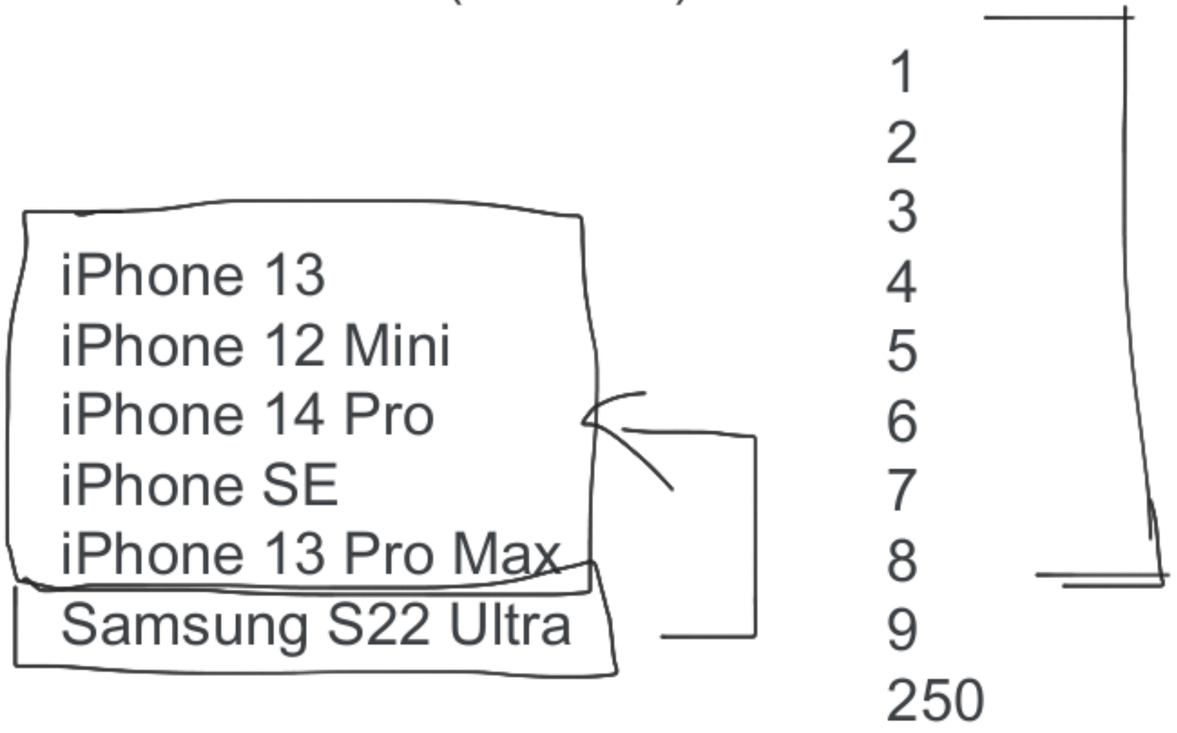
IQR - It is a measure based on the diving the dataset into quartiles (Parts). It also gives an idea where the bulk of the values lies. Plot -> Box Plot, Violin Plot
*It helps us to detect the outliers

IQR = Q3 - Q1

Q1 = 25%
Q2 = 50%
Q3 = 75%

Steps to find out the IQR

1. Arrange the data in the form of ascending order
2. Find Q2 (Median)
3. Find Q1 & Q3
4. Find IQR = (Q3 - Q1)
5. Upper Bound: Q3 - (1.5*IQR)
6. Lower Bound: Q1 - (1.5*IQR)

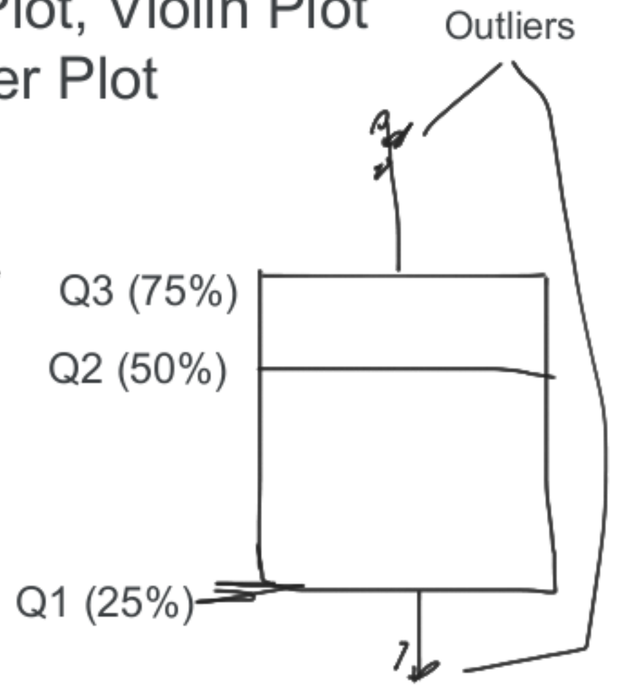


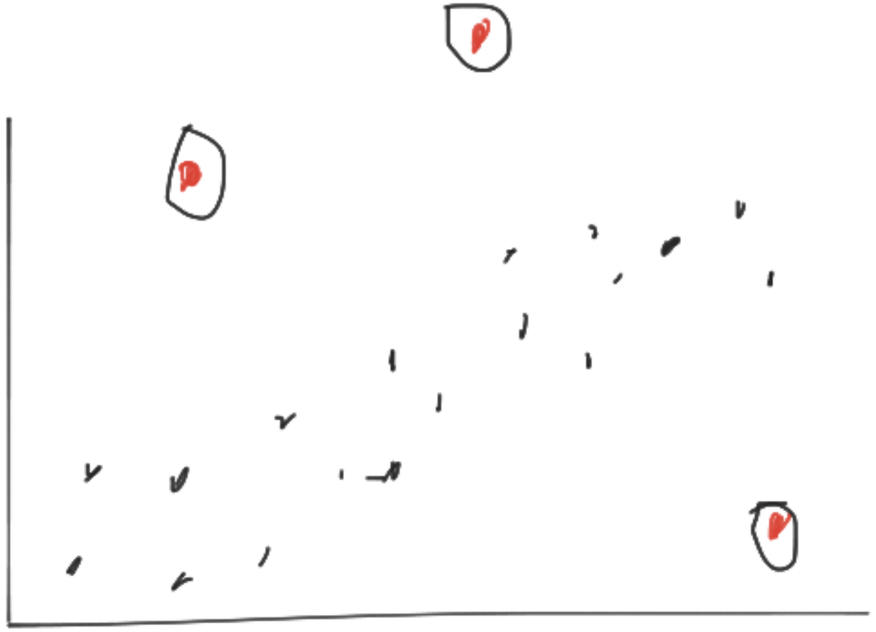
11, 20, 23, 24, 34, 25, 46, 90

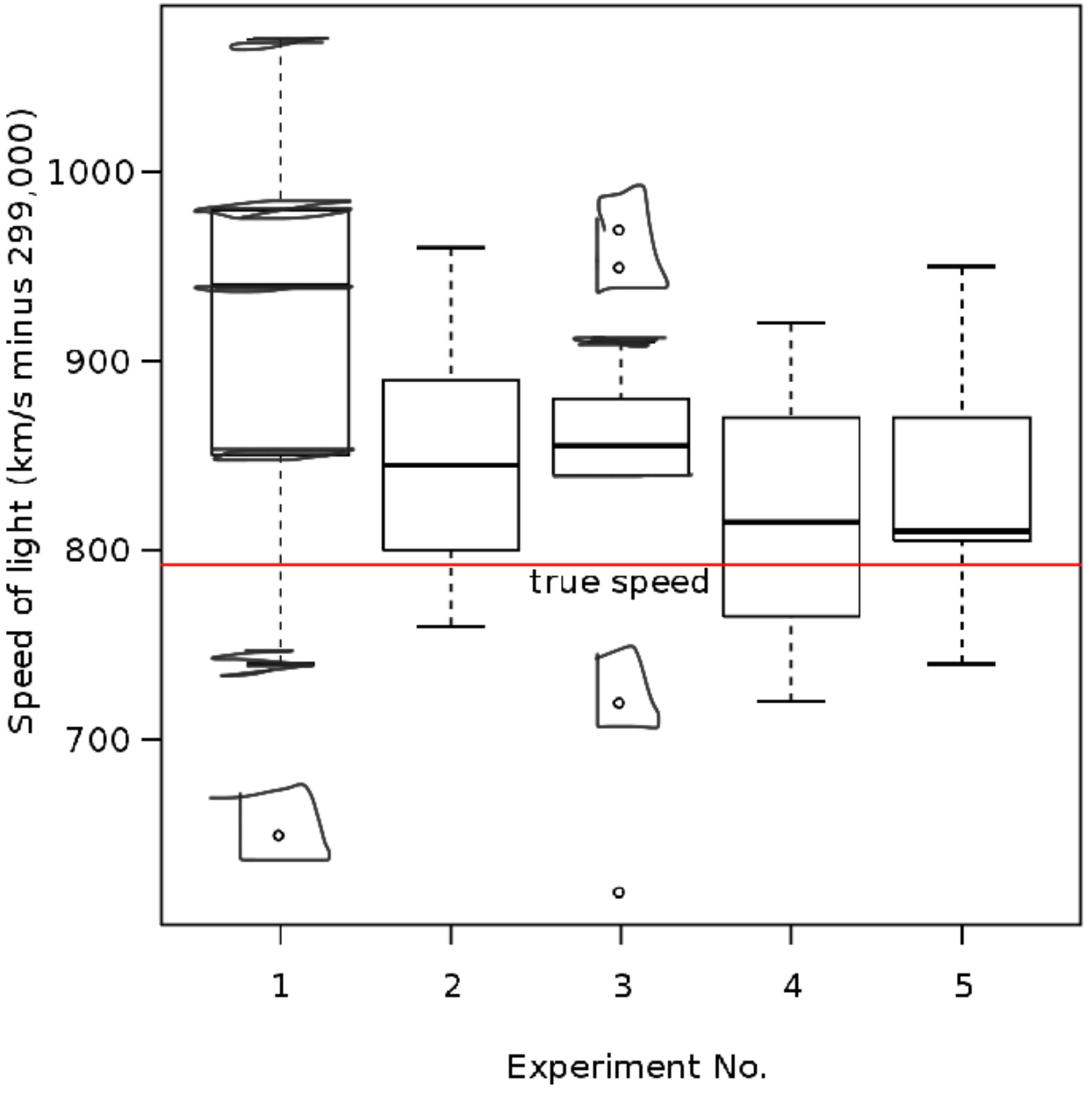
Note: Any datapoint thet lies outside the Lower & Upper Bound is an OUTLIER.

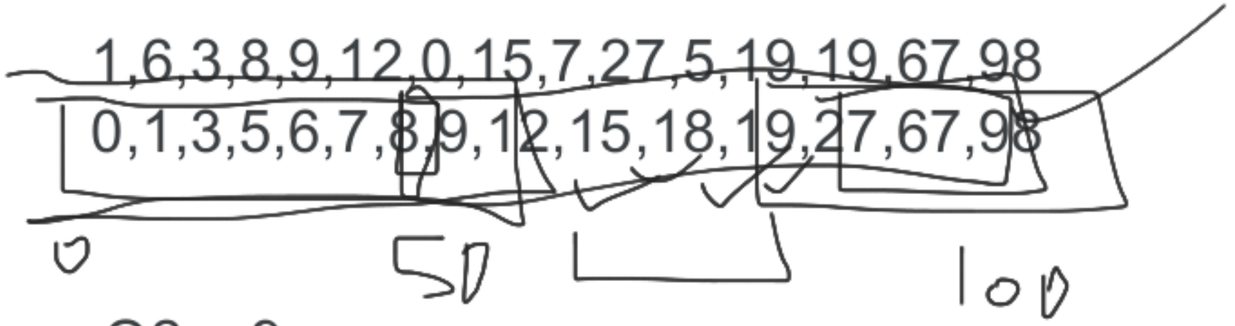
Ways to find out the outlier

1. Box Plot, Violin Plot
2. Scatter Plot
3. IQR
4. z-test



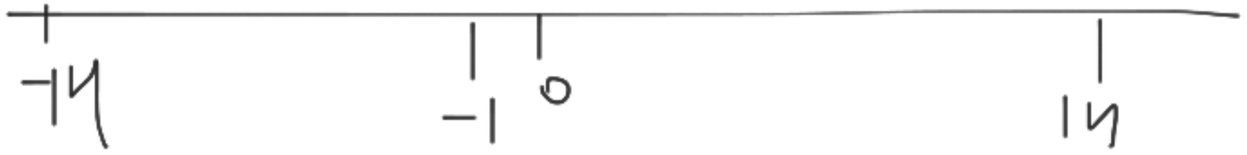



Speed of light (km/s minus 299,000)
1000
900
800
700
true speed
1
2
3
4
5
Experiment No.



Q2 = 9
Q1 = 5.5
Q3 = 18.5

IQR = Q3 - Q1 = 18.5 - 5.5 = 13



UB = 18.5 - (1.5*13) = -1
LB = 5.5 - (1.5*13) = -14 = +14

Random Variables

It is a variable that stores value that depends upon the outcomes of a random phenomeon.

Types of random Variable

Numerical RV

x = 25
x = 0.5
x = 92.9

Discrete RV

Should be a Whole number and cannot be -ve

Continuous RV

Can have any number that includes floating, decimal

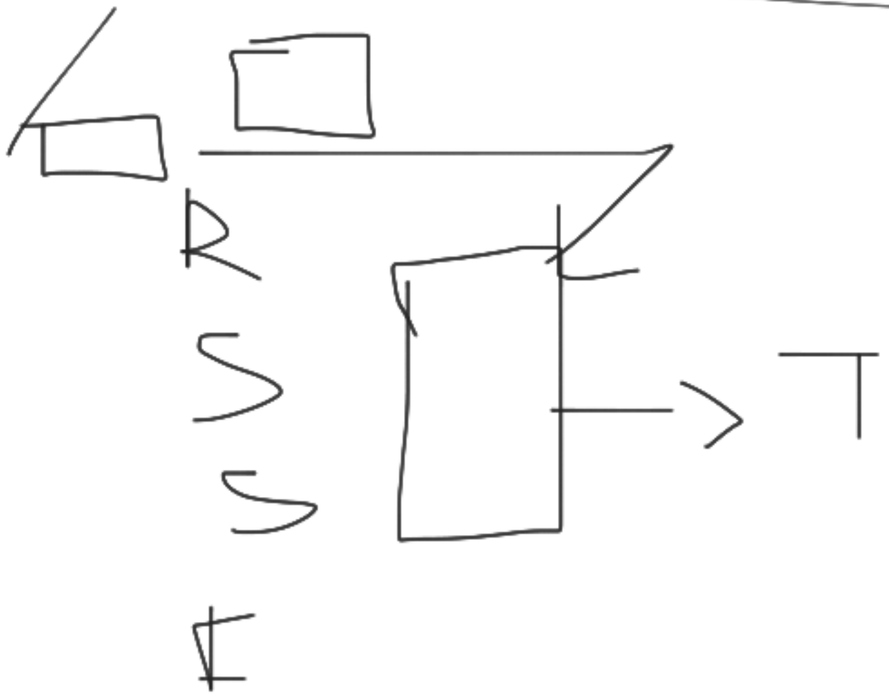
Categorical RV

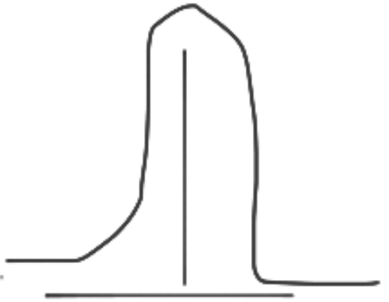
x = 'Male'
x = 'Electronics'
x = 'Domestic Animal'

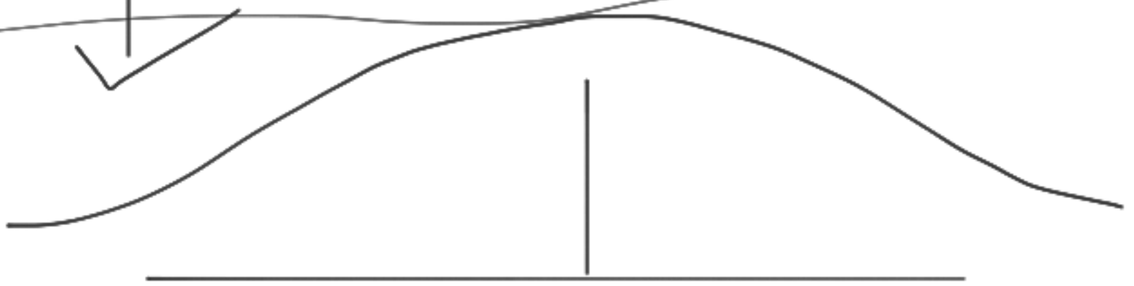
Feature Engineering (Feature Selection)

Feature
Target

A perfect model is a model which is having low bias & low variace







Covariance
Karl Pearson Correlation coefficient
Spearman Rank correleation

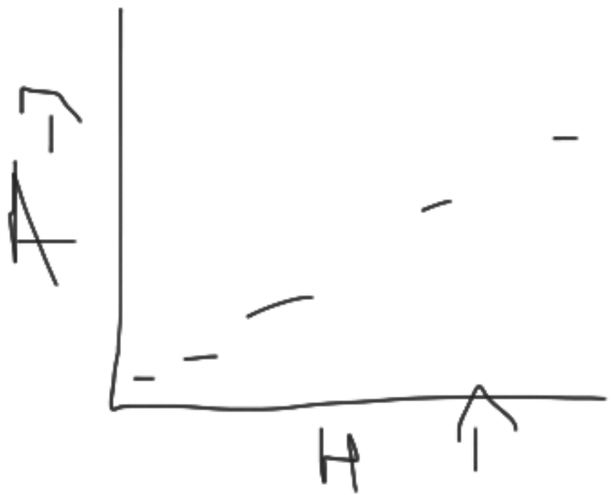
Covariance

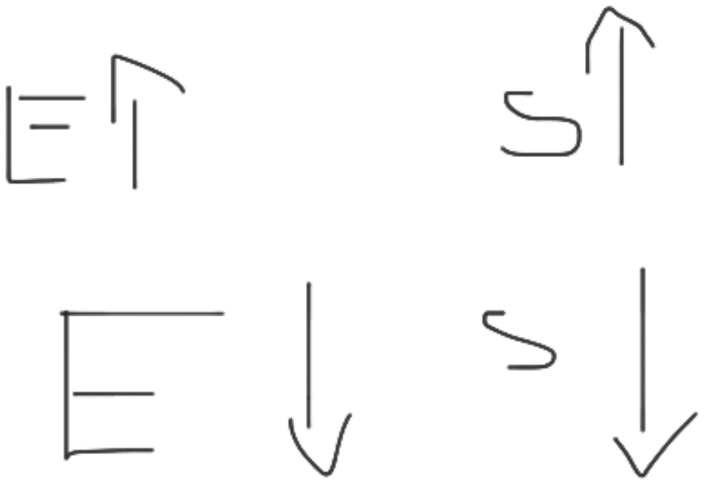
It is a measure to show the relationship between two random variables.
It will only help to find out the direction of the relationship (Disadvantage)

It can take any +ve & -ve values

Positive Covariance - When the outcome of the two RV tend to move in the same direction

Negative Covariance - When the outcome of the two RV tend to move in the different direction





$$Cov(x,y) = \sum \left(\frac{[x(i) - mean(x)][y(i) - mean(y)]}{n} \right)$$

```
def cov(x,y):
        mean_x = np.mean(x)
        mean_y = np.mean(y)
        count = 0
        for i,j in zip(x,y):
                result = (i - mean_x)*(j - mean_y)
                count += result
        return count/len(x)
```

```
np.cov(df.X, df.Y)
```

Pearson Correlation Coefficient

-1 to 1

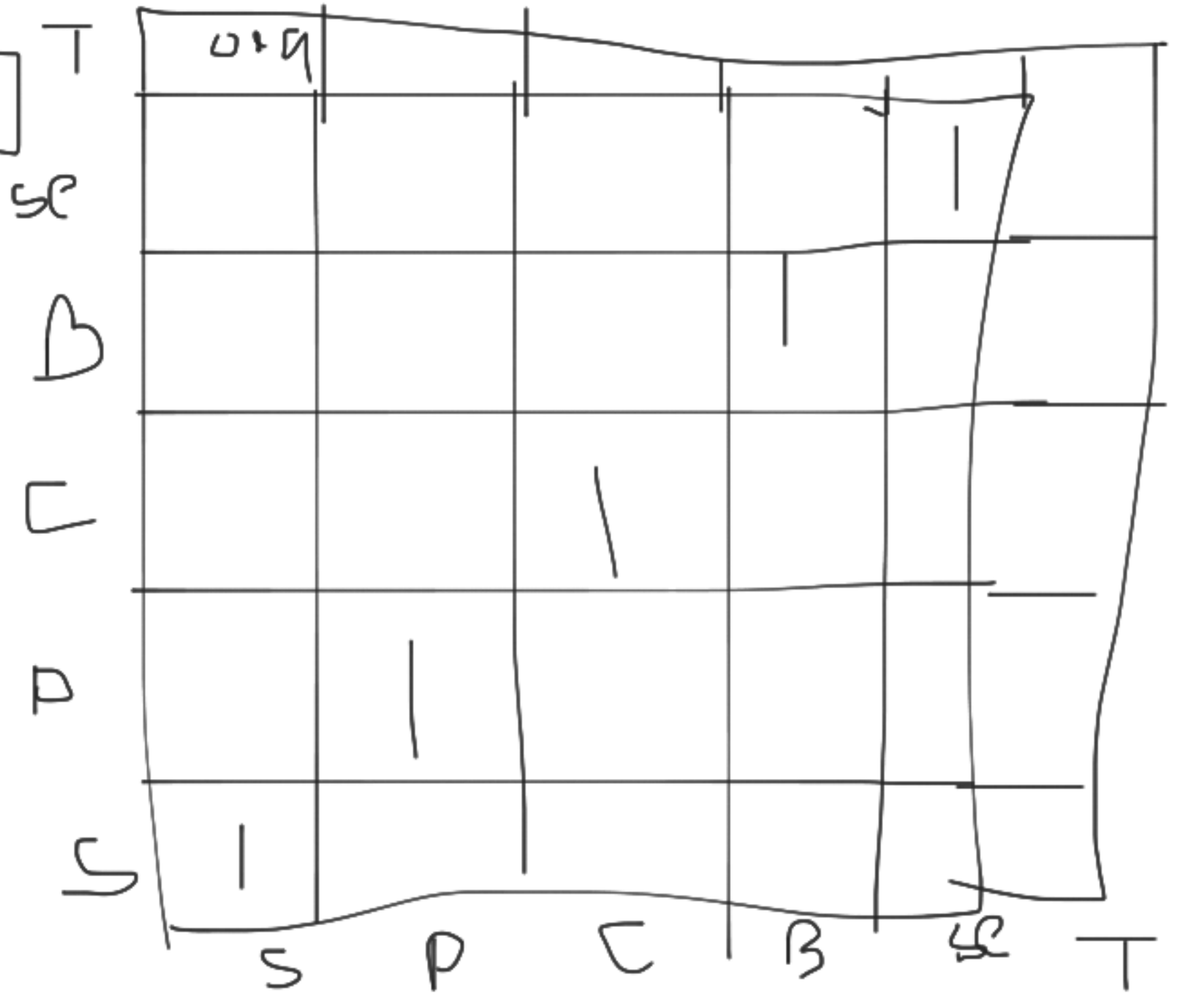
It is a measure to show the relationship between two random variables.
It not only helps to find out the direction of the relationship, but also the strength.

Top 5

Screen
Proccessors
Camera
Battery
Security

Apple -> New budget phone

1. Screen
2. Proccessors
3. Camera
4. Design
5. Battery
6. Storage
7. Heating reduction
8. Security
9. Color
10. Screen refresh rate



$$Corr(x,y) = \frac{Cov(x,y)}{SD(x) * SD(y)}$$

```
np.corrcoef(df.Age, df.Height)

df.corr()
```

1 -> Perfect Positive Correlation
0.99 - 0.79 -> High +ve Corr
0.78 - 0.5 -> Moderate +ve Corr
> 0.5 - 0 -> Low +ve corr

```
def corrcoef(x,y):
        mean_x = np.mean(x)
        mean_y = np.mean(y)
        count = 0
        for i,j in zip(x,y):
                result = (i - mean_x)*(j - mean_y)
                count += result
        cov = count/len(x)
        SD_x = np.std(x)
        SD_y = np.std(y)
        denominator = SD_x * SD_y
        return cov/denominator
```

Variable Measurement Scales

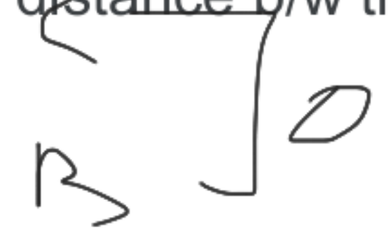
It is important beacause different types of data allow for different types of Data Analysis

1. Nominal - Qualitative Data. It is the data that split into categories-> Gender, Color, Car Type

2. Ordinal - Qualitative Data. It is the data where the order matters but the distance b/w the values doesn't matter.

3. Interval - Quantative Data. It is the data where the order matters and the distance b/w the values matters, and a natural zero (0) is not present.

4. Ratio - Quantative Data. It is the data where the order matters and the distance b/w the values matters, and a natural zero (0) is present



Spearman Rank Correlation

Data: Ordinal, Interval, Ratio

It is a measure to show the monotonic relationship between two random variables.
It not only helps to find out the direction of the relationship, but also the strength.

A monotonic relationship is where:

1. One variable increases and the other also increases, or
2. One variable decreases and the other also decreases

	A	B	C	D	E	F	G
1	Name	Physical activity (min)	Blood presure (mm Hg)	Physical activity (rank)	Blood presure (rank)	d	d_2
2	Alan	60	118	1	9	-8	64
3	Carl	55	117	2	10	-8	64
4	David	25	120	8	7	1	1
5	Don	50	121	3	6	-3	9
6	John	40	119	5	8	-3	9
7	Matt	45	122	4	5	-1	1
8	Mike	35	123	6	4	2	4
9	Neal	10	124	10	3	7	49
10	Rick	30	125	7	2	5	25
11	Rob	20	126	9	1	8	64
12							290
13	**Spearman correlation**		-0.757575758	=CORREL(D2:D11, E2:E11)			
14			-0.757575758	=1-(6*G12/(10*(10^2-1)))			

$$\rho = 1 - \frac{6 \Sigma d^2}{n(n^2-1)}$$

$$r = \frac{cov(r(x),r(y))}{SD(x) * SD(y)}$$

r(x) -> rank of x
r(y) -> rank of y

```
def spearman(x,y):
    rank_x = x.rank(ascending=False)
    rank_y = y.rank(ascending=False)
    covariance = cov(rank_x, rank_y)
    SD_x = np.std(x)
    SD_y = np.std(y)
    deno = SD_x * SD*y
    return covariance/deno
```

```
import scipy

scipy.stats.spearman(a,b)
```