Correlation & Regression Analysis

Data for Relationship Analysis

Suppose, We have a data of students age.

Data for Relationship Analysis

Suppose, We have a data of students height.

Heigh (cm)	176	156	165	182	172

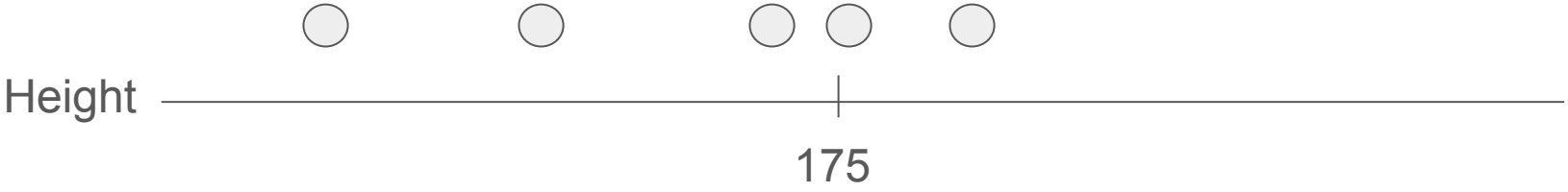
Univariate population: The population consisting of only one variable.

Data for Relationship Analysis

Suppose, We have a data of students height.

Height (cm)	176	156	165	182	172

Univariate population: The population consisting of only one variable.



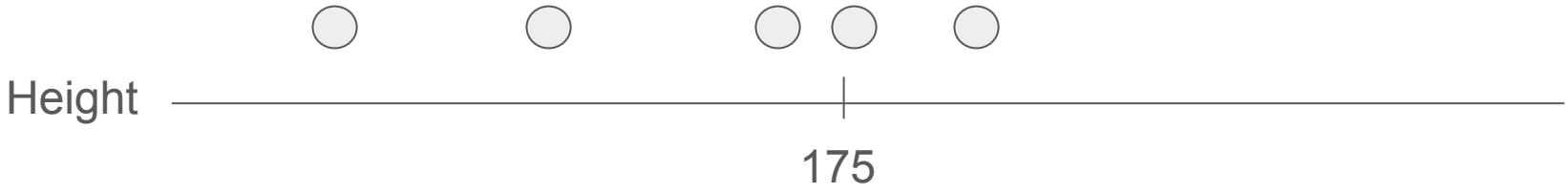
Can you see any correlation?

Data for Relationship Analysis

Suppose, We have a data of students height.

Height (cm)	176	156	165	182	172

Univariate population: The population consisting of only one variable.



Can you see any correlation?

No, Here, statistical measures suffer to find a relationship.

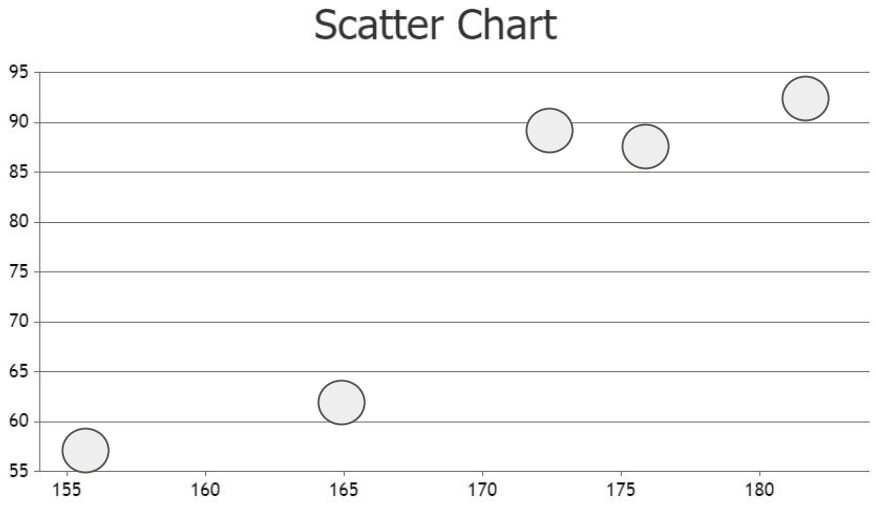
Data for Relationship Analysis

Now, Suppose, We have a data of students height and weight.

Height (cm)	176	156	165	182	172
Weight (kg)	86	56	62	92	89

Bivariate population: Here, the data happen to be with two variables.

Is there any relation?



Data for Relationship Analysis

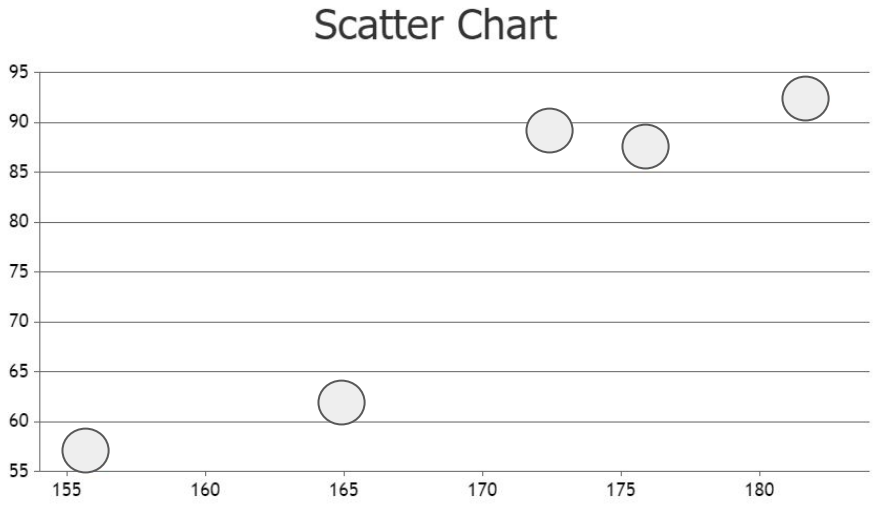
Now, Suppose, We have a data of students height and weight.

Height (cm)	176	156	165	182	172
Weight (kg)	86	56	62	92	89

Bivariate population: Here, the data happen to be with two variables.

Is there any relation?

Yes.



Data for Relationship Analysis

Now, Suppose, We have a data of students height, weight, and age.

Height (cm)	176	156	165	182	172
Weight (kg)	86	56	62	92	89
Age	25	14	218	23	26

Multivariate population: Here, the data happen to be when there are more than two variables..

Data for Relationship Analysis

Now, Suppose, We have a data of students height, weight, and age.

Height (cm)	176	156	165	182	172
Weight (kg)	86	56	62	92	89
Age	25	14	218	23	26

Multivariate population: Here, the data happen to be when there are more than two variables..

Is there any relation?

Maybe

Measures of Relationship

In case of bivariate and multivariate populations, usually, we have to answer two types of questions:

Q1: Does there exist relation between two variables (in case of bivariate population)?

If yes, of what degree?

Q2: Is there any relationship between one variable in one side and two or more variables on the other side (in case of multivariate population)?

If yes, of what degree and in which direction?

Measures of Relationship

In case of bivariate and multivariate populations, usually, we have to answer two types of questions:

Q1: Does there exist relation between two variables (in case of bivariate population) ?

Q2: Is there any relationship between one variable in one side and two or more variables on the other side (in case of multivariate population)?

Measures of Relationship

In case of bivariate and multivariate populations, usually, we have to answer two types of questions:

Q1: Does there exist relation between two variables (in case of bivariate population) ?

Q2: Is there any relationship between one variable in one side and two or more variables on the other side (in case of multivariate population)?

To find solutions to the above questions, two approaches are known.

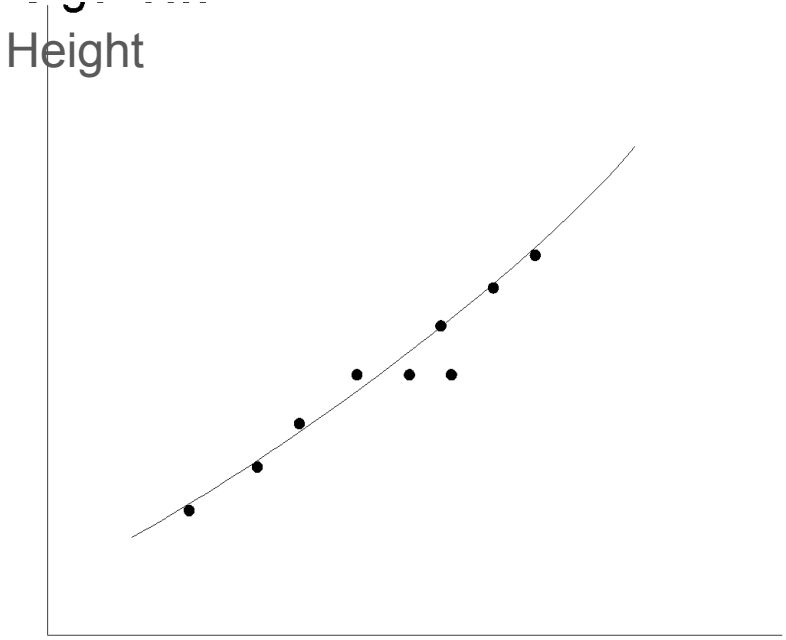
- **Correlation Analysis**
- **Regression Analysis**

Correlation Analyses

Correlation Analysis

In statistics, the word correlation is used to denote some form of association between two variables.

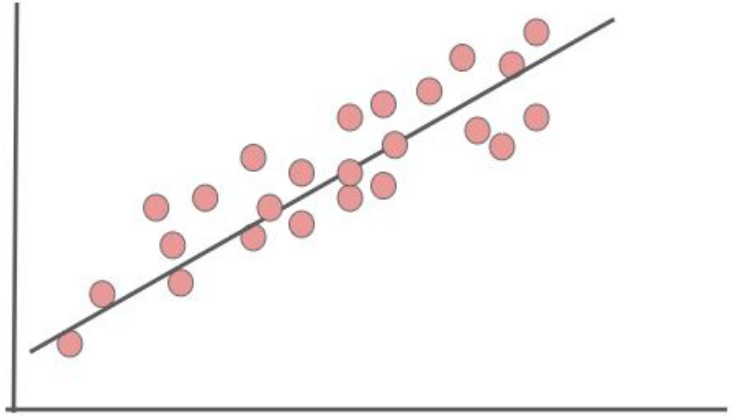
Example: Weight is correlated with height



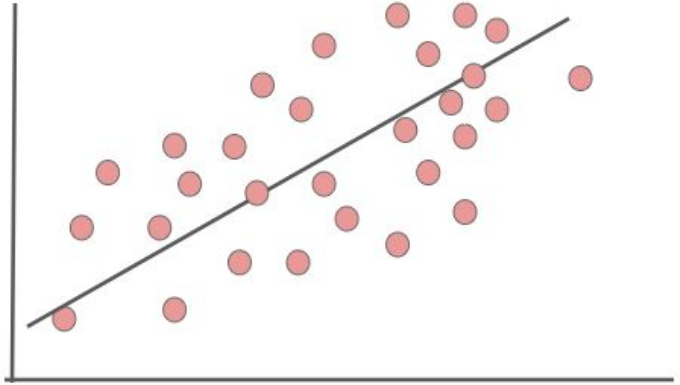
Correlation Type

There are couple of Correlation:

1) **Positive Correlation**
   a) **Strong Positive**
   b) **Weak Positive**



2) Negative Correlation
   a) Strong Negative
   b) Weak Negative

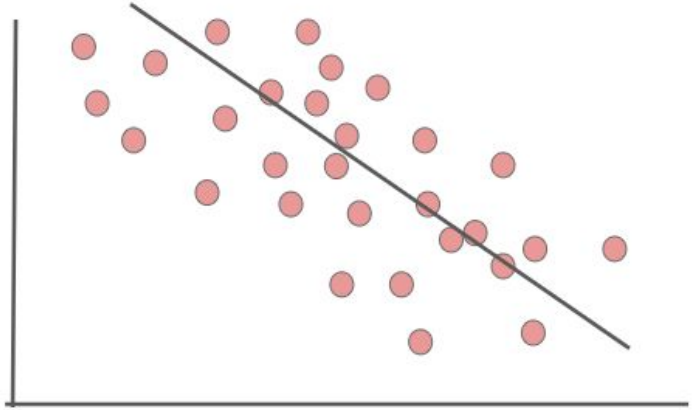


3) No Correlation

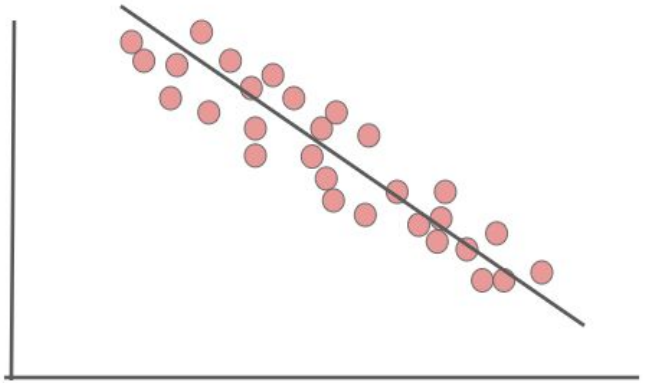
Correlation Type

There are couple of Correlation:

1) Positive Correlation
   a) Strong Positive
   b) Weak Positive



2) Negative Correlation
   a) Strong Negative
   b) Weak Negative

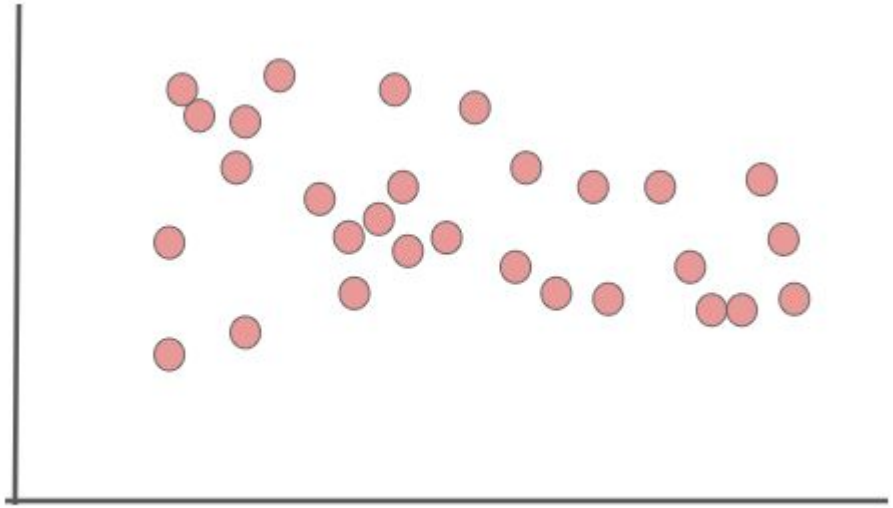


3) No Correlation

Correlation Type

There are couple of Correlation:

1) Positive Correlation
    a) Strong Positive
    b) Weak Positive



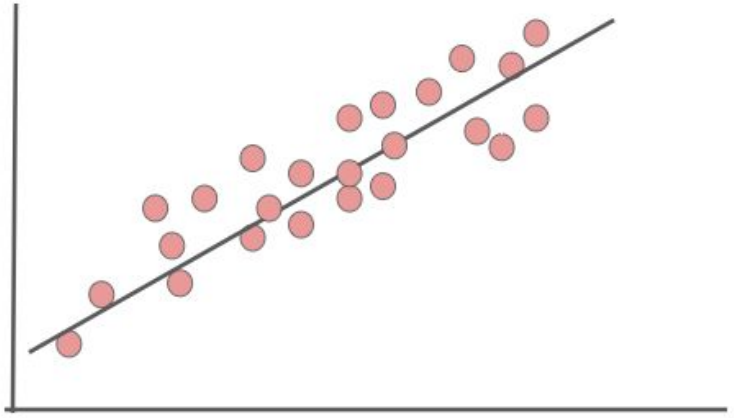
2) Negative Correlation
    a) Strong Negative
    b) Weak Negative

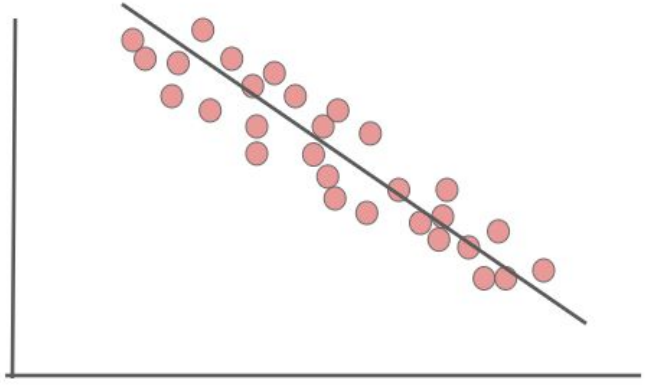
3) No Correlation

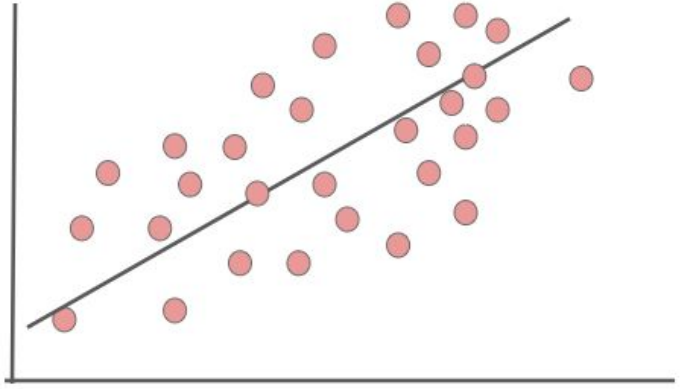
Correlation Form

Different form of Correlation:

1) **Linear Correlation**
2) No-Linear Correlation



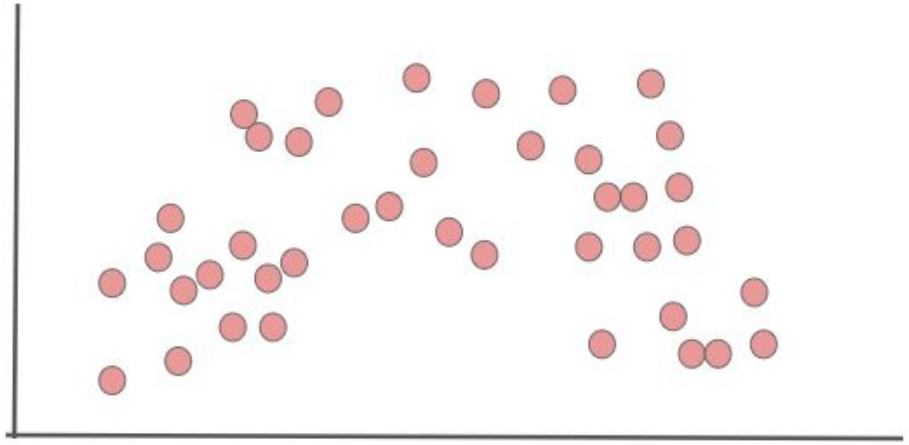




Correlation Form

Different form of Correlation:

1) Linear Correlation
2) **No-Linear Correlation**

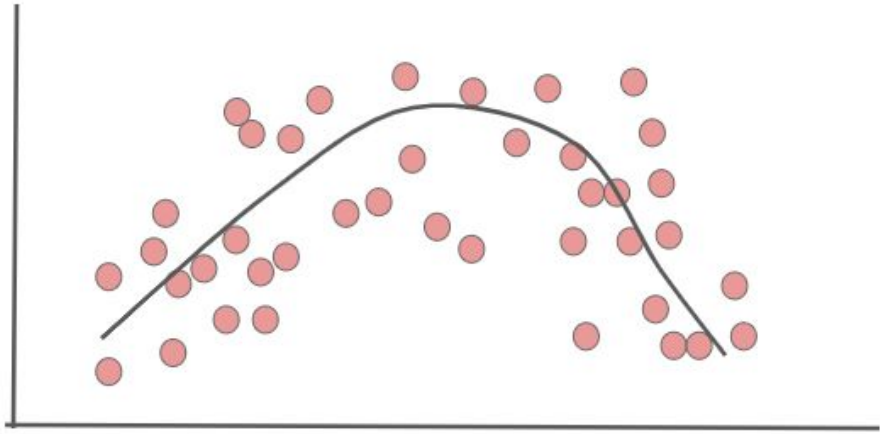


Correlation Form

Different form of Correlation:

1) Linear Correlation
2) **No-Linear Correlation**

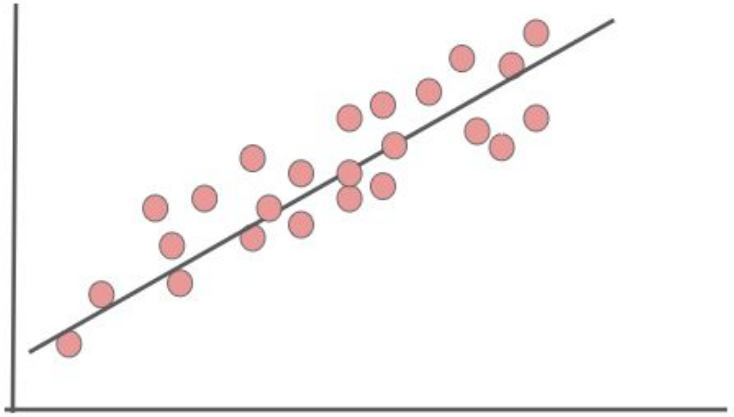
We may see a correlation, but not linear.

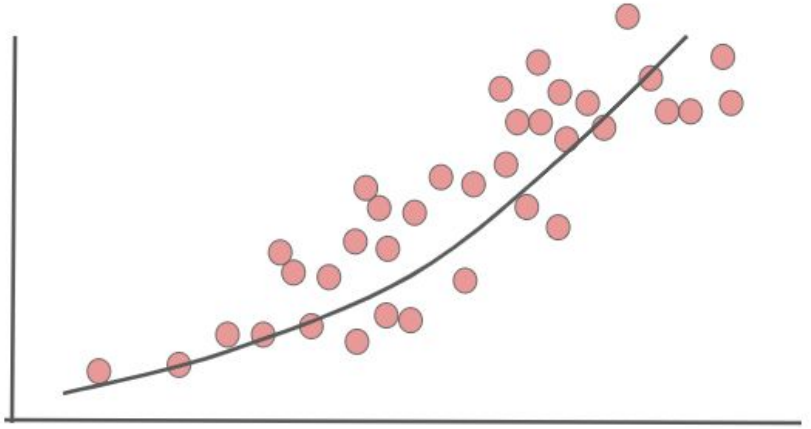


Correlation Form

Different form of Correlation:

1) **Monotonic**
2) Non-Monotonic

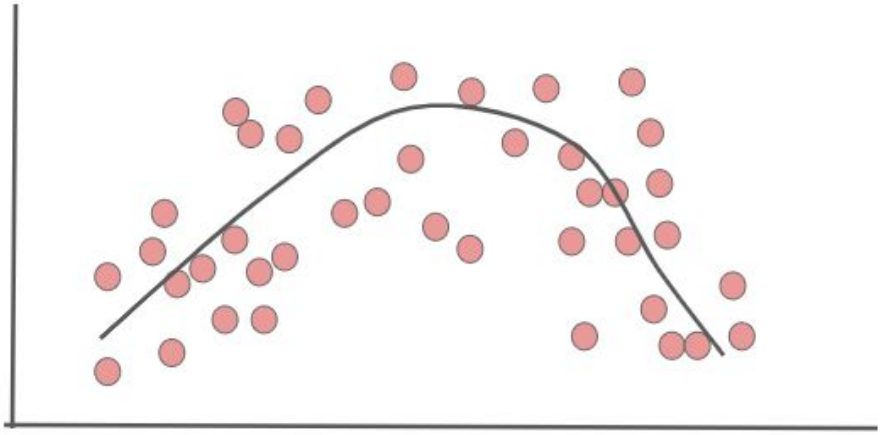




Correlation Form

Different form of Correlation:

1) Monotonic
2) Non-Monotonic



Correlation Coefficient

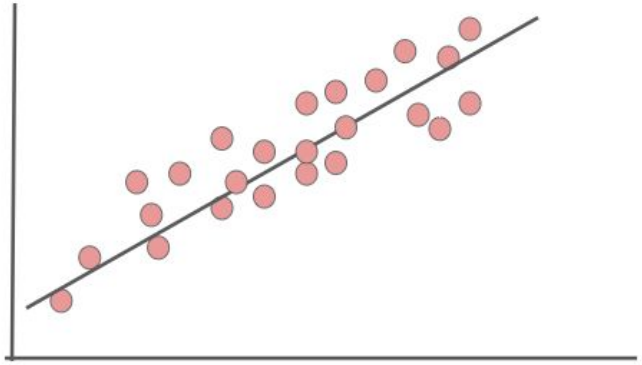
The correlation coefficient **(r)** is a numerical measure that describes the strength and direction of a linear relationship between two variables. It ranges from -1 to 1, where:

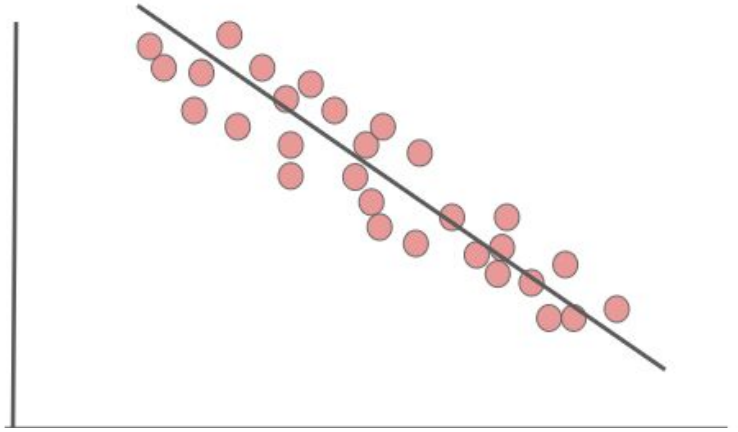
- 1 indicates a perfect positive linear relationship
- -1 indicates a perfect negative linear relationship
- 0 suggests no linear relationship between the variables.

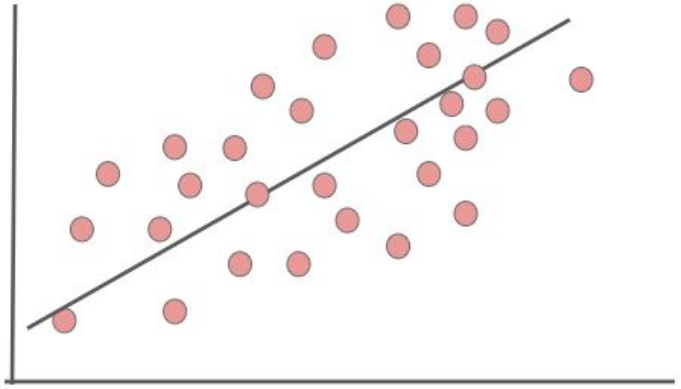
Values closer to -1 or 1 signify a stronger linear correlation, while values near 0 indicate a weak correlation.

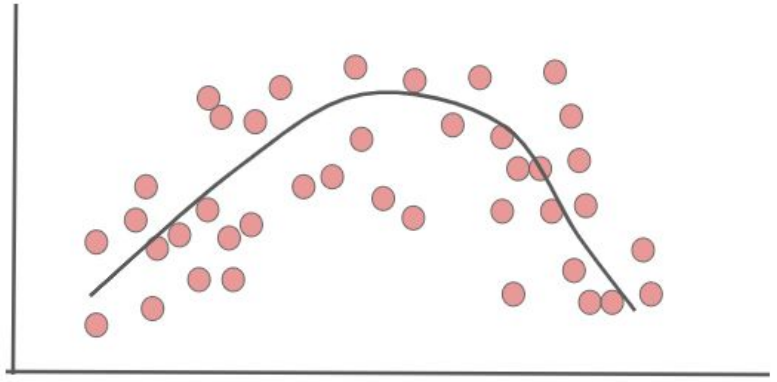
This coefficient is widely used in statistics to assess how closely data points fit a linear trend, and it forms the basis for many forms of predictive analysis and modeling.

Let’s guess the Correlation Coefficient









How to measure Correlation Coefficient

The methods of Correlation Coefficient

If two feature is **numerical**

If two feature is **ordinal**

If two feature is **nominal**

How to measure Correlation Coefficient

The methods of Correlation Coefficient

If two feature is **numerical**

Pearson Correlation

If two feature is **ordinal**

If two feature is **nominal**

How to measure Correlation Coefficient

The methods of Correlation Coefficient

If two feature is **numerical**

Pearson Correlation

If two feature is **ordinal**

Charles Spearman's Correlation

If two feature is **nominal**

How to measure Correlation Coefficient

The methods of Correlation Coefficient

If two feature is **numerical**

Pearson Correlation

If two feature is **ordinal**

Charles Spearman's Correlation

If two feature is **nominal**

Chi squared Correlation

Other Correlation Analysis

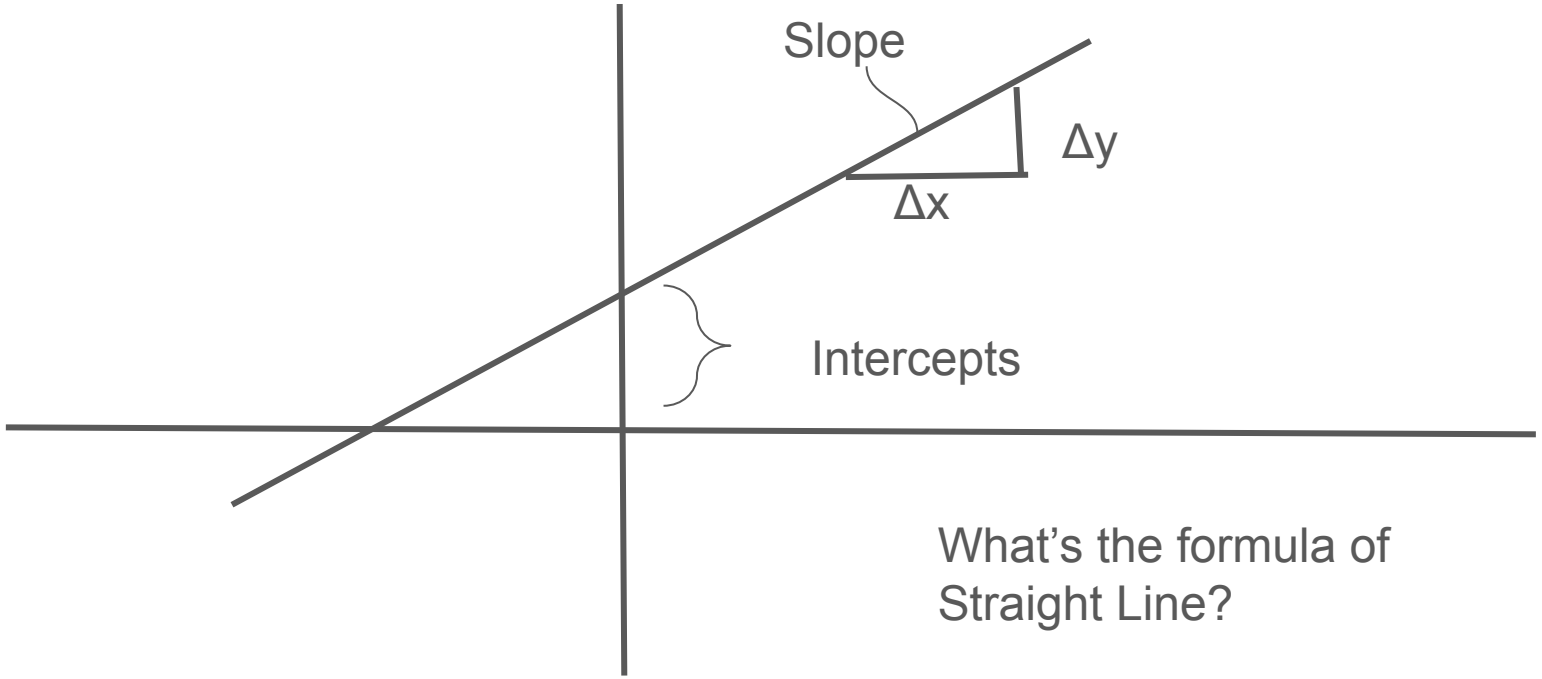
- Binary variable to binary variable correlation
  - Tetrachoric correlation

- Nominal/ categorical valued variable to binary variable correlation
  - Cramer’s V correlation

- Continuous variable to binary variable correlation
  - Point-biserial correlation

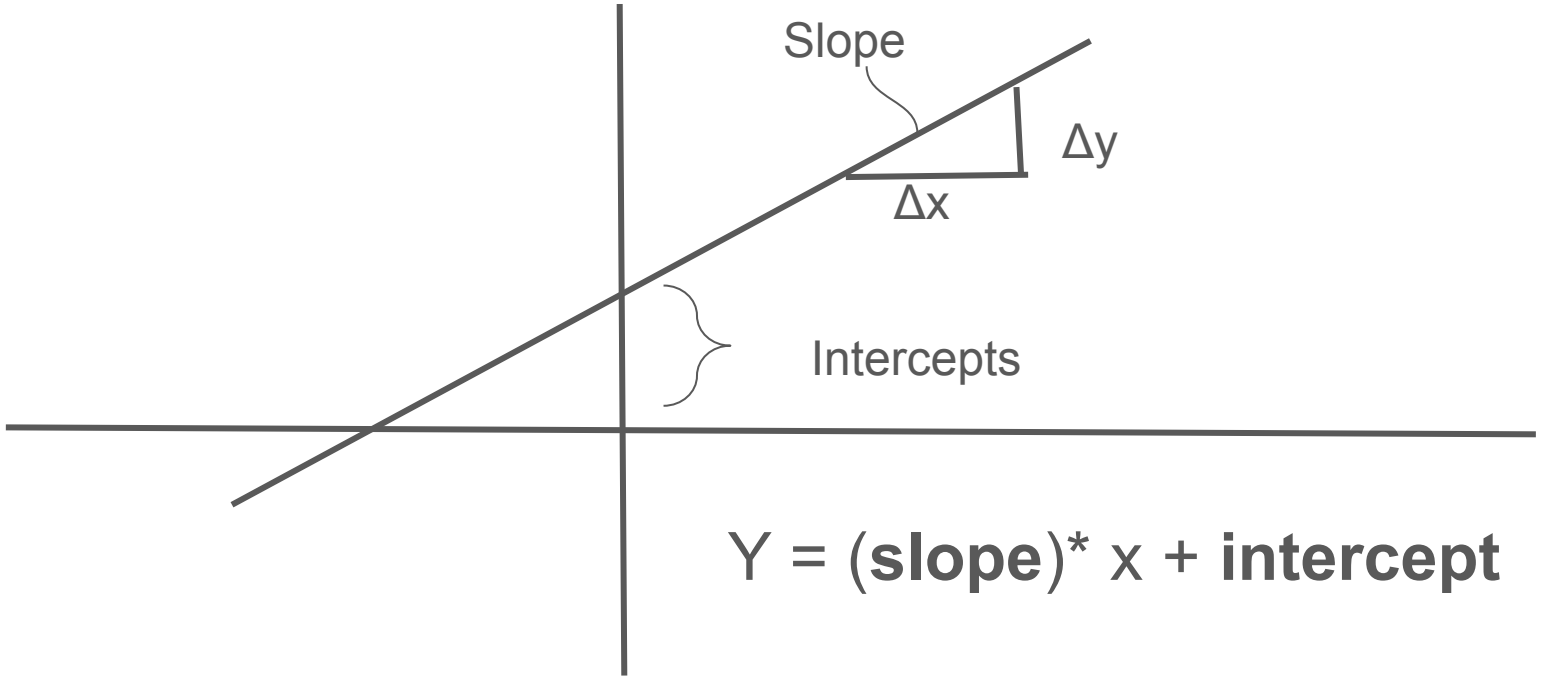
Regression Analysis

First talk about Straight Line



What's the formula of Straight Line?

First talk about Straight Line



$$Y = (\textbf{slope})^* x + \textbf{intercept}$$

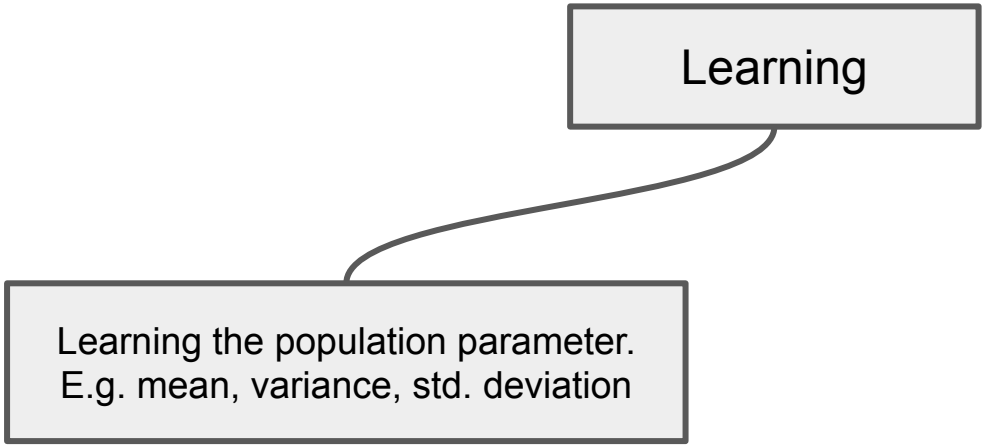
Now slope has an important use case

Because, this is the estimation of regression slope.

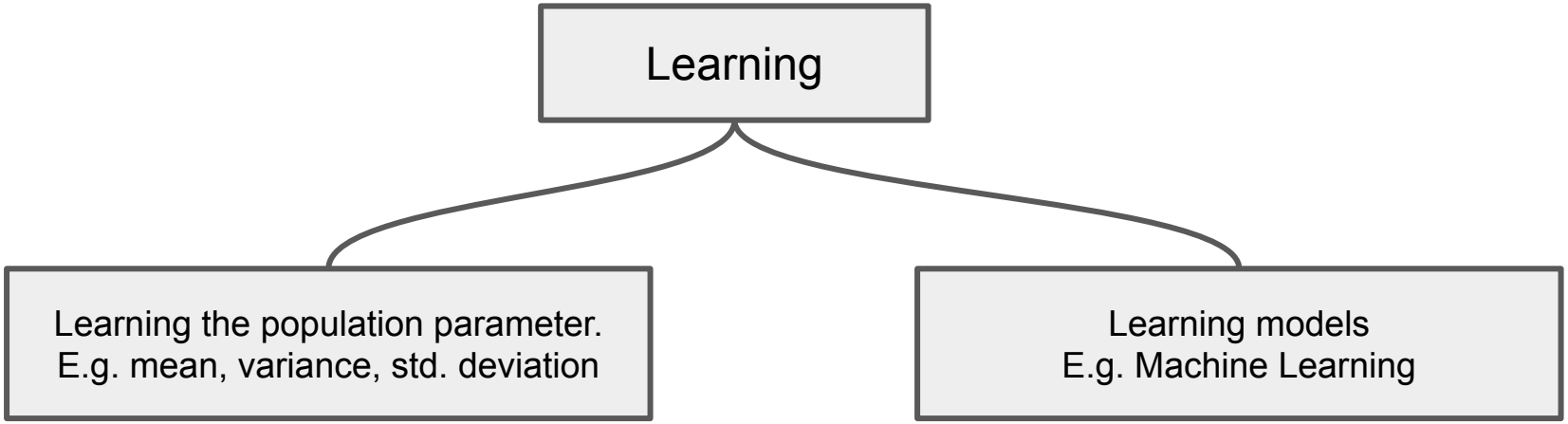
And, the x is the variable value (e.g. age, smoke)

And, y is the prediction.

Let’s talk about learning strategies



Let’s talk about learning strategies



For Statistical Learning

We assume:

- The observation come from a normal population.
- Sample size is small.
- Population parameters like mean, variance, etc. are hold good.
- Requires measurement equivalent to interval scaled data.

For Machine Learning

We assume:

Nothing!!!

For Machine Learning

We assume:

Nothing!!!

We only consider the dataset is large enough and have many attributes.

And, Machine will learn everything for us.

But there is a catch

We need to analyze the relationship between attributes.

For example:
Let's consider Wage of an employe example,

1) Do "Age" attributes related to "Wage"?
2) Do "Year of Experience" attributes related to "Wage"?
3) Do "Education Qualification" attributes related to "Wage"?

so on……

But, How can we analyze the relationship?

We need to find the correlation between them..

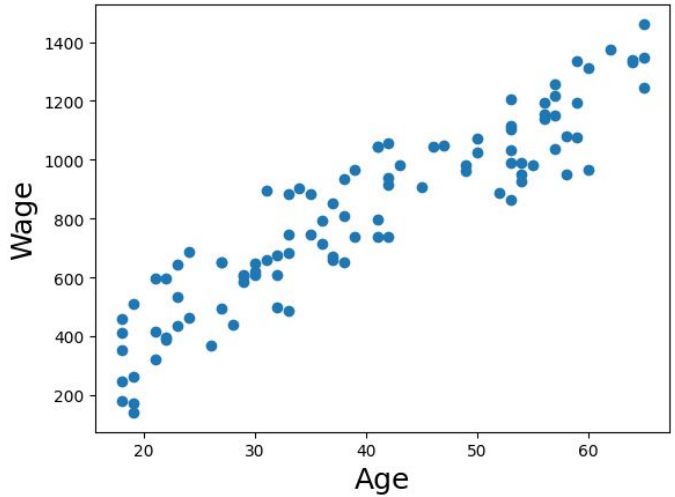
Like, Pearson Correlation, Spearman’s Correlation…

But, How can we analyze the relationship?

We need to find the correlation between them..

Like, Pearson Correlation, Spearman's Correlation…

If we plot the Wage vs Age plot..



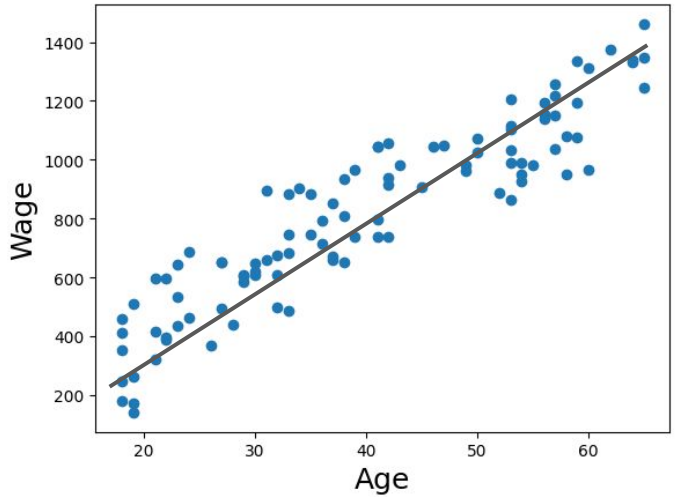
Is there any pattern?

But, How can we analyze the relationship?

We need to find the correlation between them..

Like, Pearson Correlation, Spearman's Correlation…

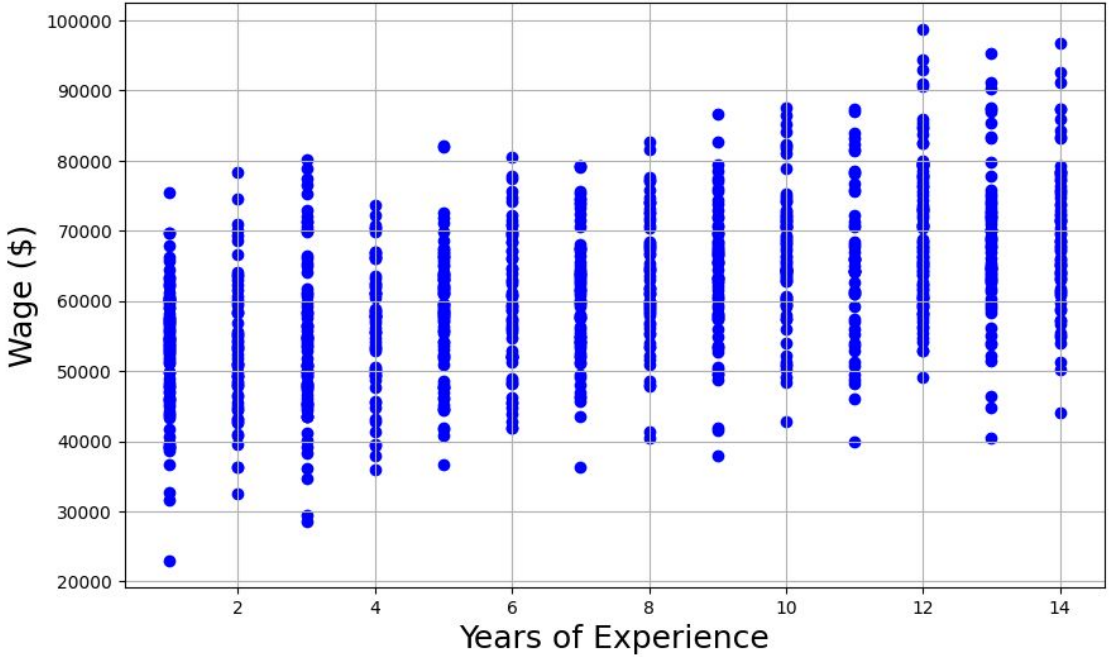
If we plot the Wage vs Age plot..



Yes, we can see a positive correlation

But, How can we analyze the relationship?

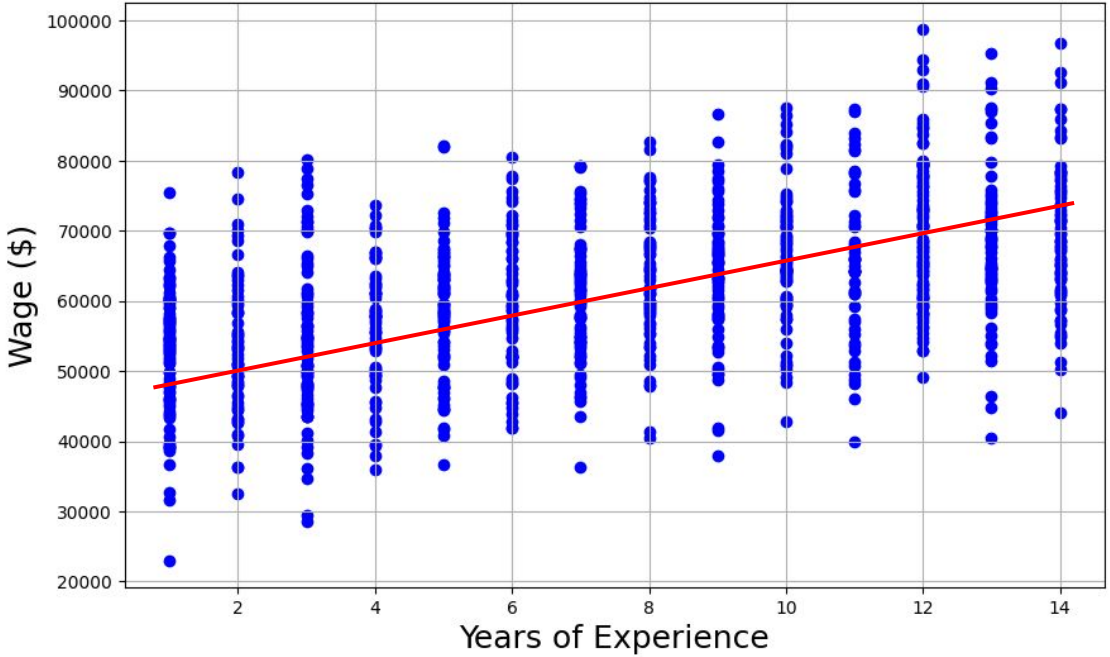
Again, If we plot the Wage vs Year of experiment plot..



Is there any pattern?

But, How can we analyze the relationship?

Again, If we plot the Wage vs Year of experiment plot..



Yes, Although the Years of Experience is discrete.

Now…

What if we consider Wage vs Age vs Year of Experience?

This is the concept of Machine Learning


THANK YOU