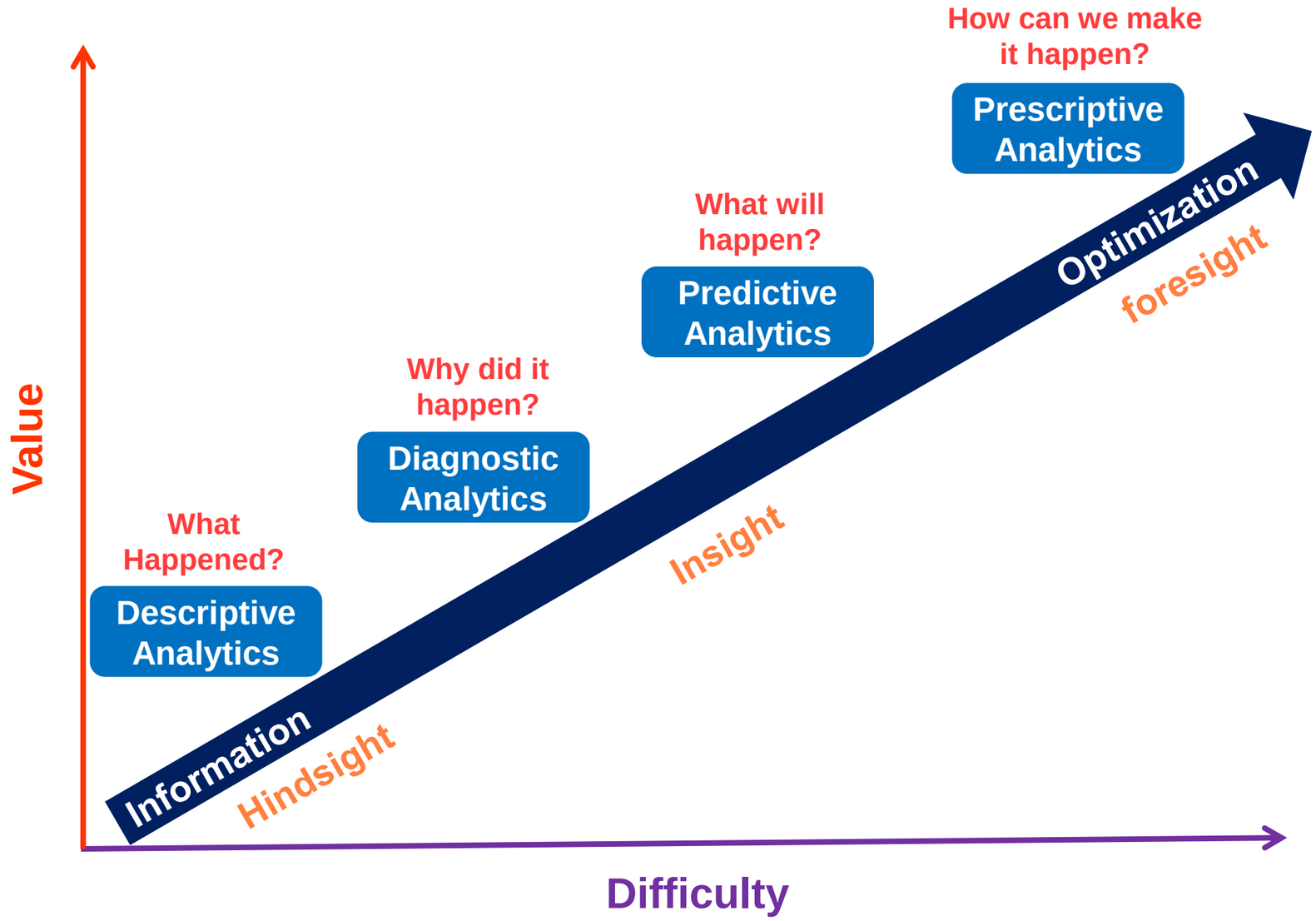
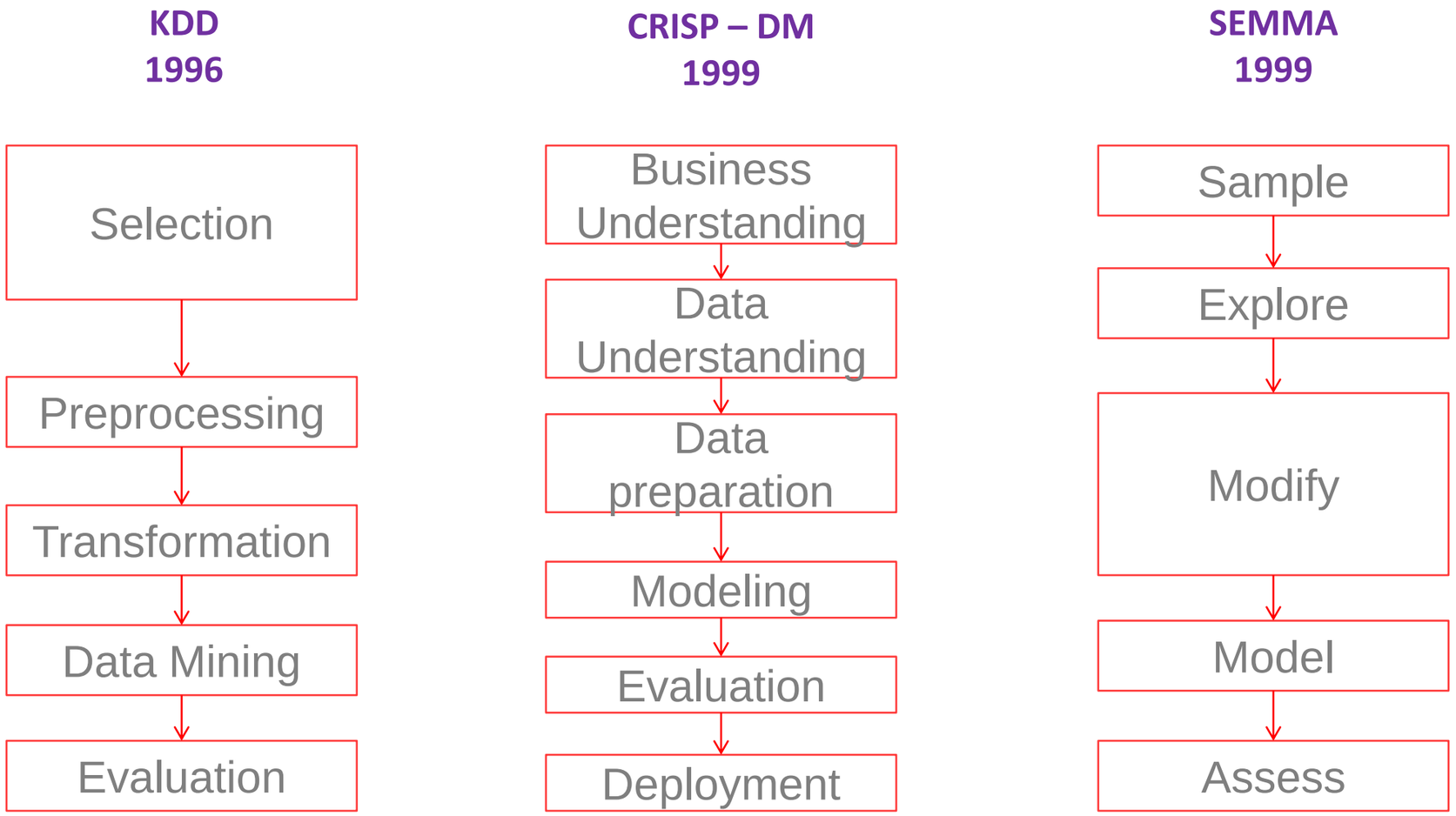

# Data Analysis /Data Science

# Analytics Continuum



# Overview of data mining Process

Summary of data Mining Frameworks



# Steps in Analytical Process

## 1. Business Understanding

- Identify the Business objective
- Assess the situation
- Determine the Analytical goals
- Produce a project plan

## 2. Data Understanding

- ✓ Collect the data
- ✓ Describe the data
- ✓ Explore the data
- ✓ Verify the data Quality

## 3. Data Preparation

- Select the data
- Clean the data
- Construct the data
- Integrate the data
- Format the data

## 4. Modeling

- Select a modeling technique
- Generate a test Design
- Build a model
- Assess a model

## 5. Evaluation

- ✓ Evaluate the results
- ✓ Review the process
- ✓ Determine the next steps

## 6. Deployment

- Deploying the plan
- Monitoring and maintenance of the plan
- Producing the final report
- Reviewing the project

# Typical Effort for each Process

- Business Understanding >> 5 to 15 %
- Data Understanding >> 5 to 10 %
- Data Preparation >> 50 to 60 %
- Modeling >> 5 to 15 %
- Evaluation >> 5 to 10 %
- Deployment >> 10 to 15 %

# Data understanding…

- Data Pre-processing
  - Use different data transformations in order to expose the structure of prediction problems in a better way
    - Standardizing
    - Normalizing etc.

# Modelling

- Based on the type of data, we need to select the modelling algorithm.

- Whether it is a
  - Regression Problem
  - Cluster Analysis
  - Classification Analysis
- Statistical Models are mainly divided into different types like,
  - Grouping / Predicting / Association
    - (based on the type of dependent/Independent variables, different methods exist and, we need to choose accordingly)

# Top 10 questions that we need to assess before starting a project

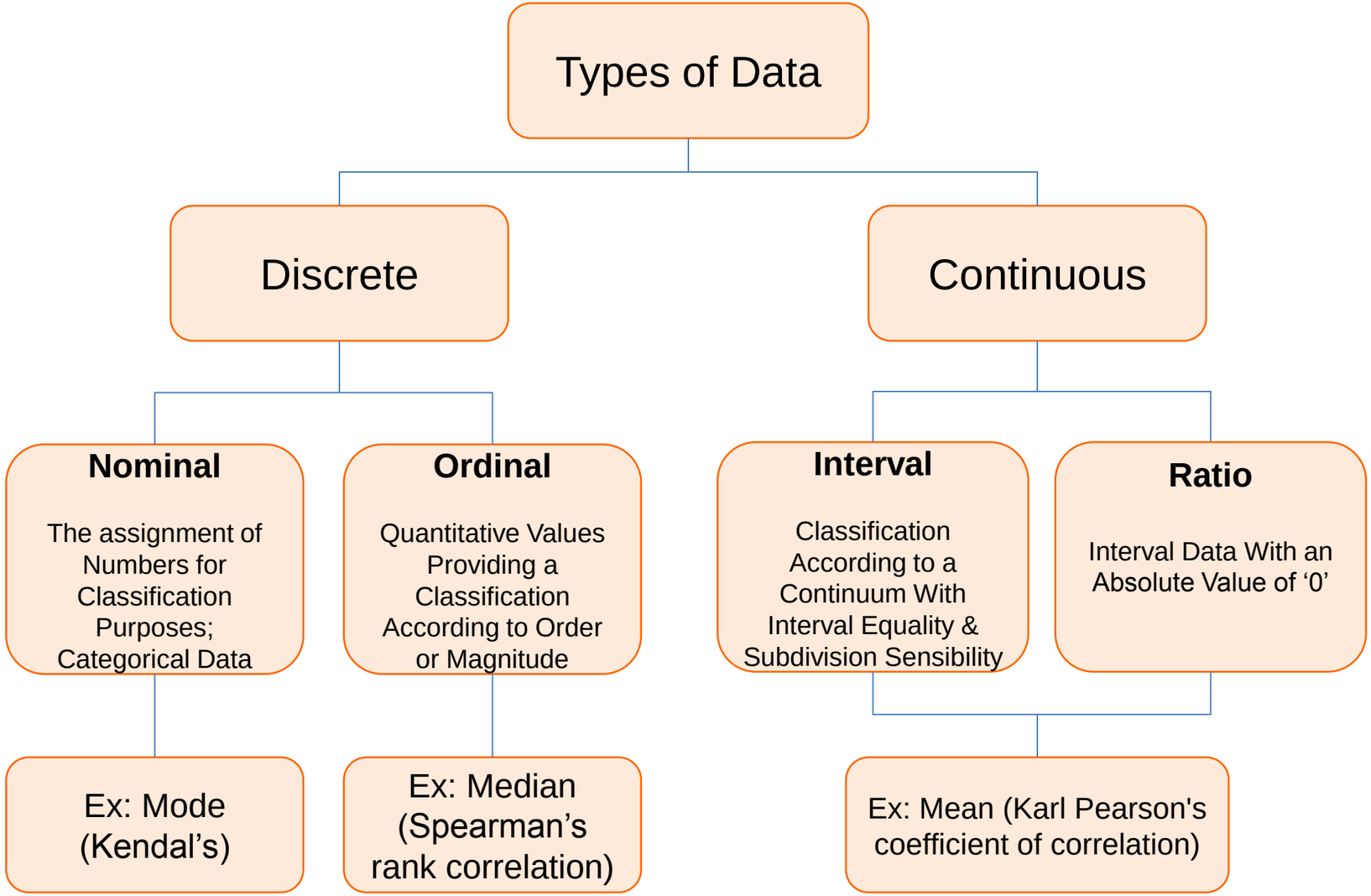
1. What are we trying to achieve, business wise? Why it is important?
2. What are the inputs and outputs for the task that we are trying to solve?
3. Given a hypothetical solution to the task, how would it affect our operations? (another way to ask this question: assuming that I have perfect solution to your machine learning task, how will you use it?)
4. Do we already have the ability to act based on such solution, or do we also need to develop that ability? (if the ability is there, learn it carefully. If not keep close contact with the team that is responsible for developing it)
5. How are we going to measure a suggested solution? (KPIs)
6. What would make it a success?
7. Do we have the input data available? How hard it is to extract it? Are we allowed to use it?
8. Are we experience in building similar solutions? Do we understand what it takes?
9. Do we have hard budget and timelines constraints?
10. Who will develop the solution? Do we have the required skills in house?

juxt smart mandate

# The Execution Steps

| | | |
|---|---|---|
| 1) | Understand Data with Descriptive Statistics | (Data Understanding) |
| 2) | Understand Data with Visualization | (Data Understanding) |
| 3) | Pre-Process Data/ Feature Engineering | (Data Preparation) |
| 4) | Spot-Check Algorithms | (Modeling) |
| 5) | Resampling Method (cross validation) | (Evaluation) |
| 6) | Algorithm Parameter Tuning | (Evaluation) |
| 7) | Evaluation Metrics | (Evaluation) |
| 8) | Finalize Model & Apply | (Deployment) |

# Measurement Scales & Types of Data



The problem with interval scales is that they don't have a "true zero." For example, there is no such thing as "no temperature." Without a true zero, it is impossible to compute ratios. With interval data, we can add and subtract, but cannot multiply or divide. For ex., consider this: 10 degrees + 10 degrees = 20 degrees. But, 20 degrees is not twice as hot as 10 degrees.

**Ratio** – Ratio scales are numerical measurements where the distance between numbers is of a known constant size, in addition, there is a also an absolute zero. Good examples of ratio variables include Height, Weight, Income, Age

# Primary Scales of Measurement

Table 8.1

| Scale | Basic Characteristics | Common Examples | Marketing Examples | Permissible Statistics | |
|---|---|---|---|---|---|
| | | | | Descriptive | Inferential |
| **Nominal** | Numbers identify & classify objects | Social Security nos., numbering of football players | Brand nos., store types | Percentages, mode | Chi-square, binomial test |
| **Ordinal** | Nos. indicate the relative positions of objects but not the magnitude of differences between them | Quality rankings, rankings of teams in a tournament | Preference rankings, market position, social class | Percentile, median | Rank-order correlation, Friedman ANOVA |
| **Interval** | Differences between objects | Temperature (Fahrenheit) | Attitudes, opinions, index | Range, mean, standard | Product-moment |
| **Ratio** | Zero point is fixed, ratios of scale values can be compared | Length, weight | Age, sales, income, costs | Geometric mean, harmonic mean | Coefficient of variation |

| Type | Definition | Examples |
|---|---|---|
| **Nominal** | Provides a name. If numeric, then no scale is implied | Male, Female<br>1 (Republican), 2 (Democratic), 3 (Independent) |
| **Ordinal** | Provides an ordered scale | 1 (Excellent), 2 (Good), 3 (Fair), 4 (Poor) |
| **Interval** | Can be manipulated mathematically.<br>Scale in equal increments. | Temperature in centigrade ($80^0$ is $20^0$ hotter than $60^0$, which is $20^0$ hotter than $40^0$, but $80^0$ is not twice as hot as $40^0$) |
| **Ratio** | Interval scale with a meaningful zero | Temperature in Kelvin ($80^0$ is twice as hot as $40^0$)<br>Weight, length, age |

# Numerical scale of measurement

**Qualitative data (Non Metric) :**

- **Nominal** – consist of categories in each of which the number of respective observations is recorded. The categories are in no logical order and have no particular relationship. The categories are said to be ***mutually exclusive*** since an individual, object, or measurement can be included in only one of them.
- **Ordinal** – contain more information. Consists of distinct categories in which order is implied. Values in one category are larger or smaller than values in other categories (e.g. rating- excelent, good, fair, poor)

**Quantitative data (Metric) :**

- **Interval** – is a set of numerical measurements in which the distance between numbers is of a known, constant size.
- **Ratio** – consists of numerical measurements where the distance between numbers is of a known, constant size, in addition, there is a nonarbitrary zero point.

# Variable Types Identify the correct type(s):

| Variable | Scoring | Nominal | Ordinal | Contin. |
|---|---|---|---|---|
| Quality of life | 1 = Poor<br>2 = Fair<br>3 = Average<br>4 = Good<br>5 = Very Good | | | |
| Ethnicity | 1 = Non-Hispanic<br>2 = Hispanic | | | |
| Race | 1 = African American<br>2 = Caucasian<br>3 = Other | | | |
| Diabetes | 1 = Absent<br>2 = Present | | | |
| Systolic BP | Ranges from 95 to 190 mmHg | | | |

# Variable Types Identify the correct type(s):

| Variable | Scoring | Nominal | Ordinal | Contin. |
|---|---|---|---|---|
| Quality of life | 1 = Poor<br>2 = Fair<br>3 = Average<br>4 = Good<br>5 = Very Good | | ● | |
| Ethnicity | 1 = Non-Hispanic<br>2 = Hispanic | ● | | |
| Race | 1 = African American<br>2 = Caucasian<br>3 = Other | ● | | |
| Diabetes | 1 = Absent<br>2 = Present | ● | | |
| Systolic BP | Ranges from 95 to 190 mmHg | | | ● |

# Frequency distributions – numerical presentation of quantitative data

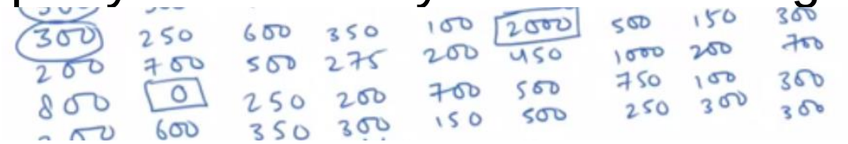
- **Individual** - Raw Data considered for any variable under study .

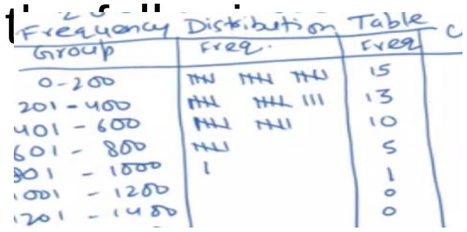
  Example : Marks of **ALL** students in a training class .

  Practically it is not possible to review each case . Hence , frequency distribution is considered .

- **Frequency distribution** – shows the frequency, or number of occurences, in each of several categories. Frequency distributions are used to summarize large volumes of data values.

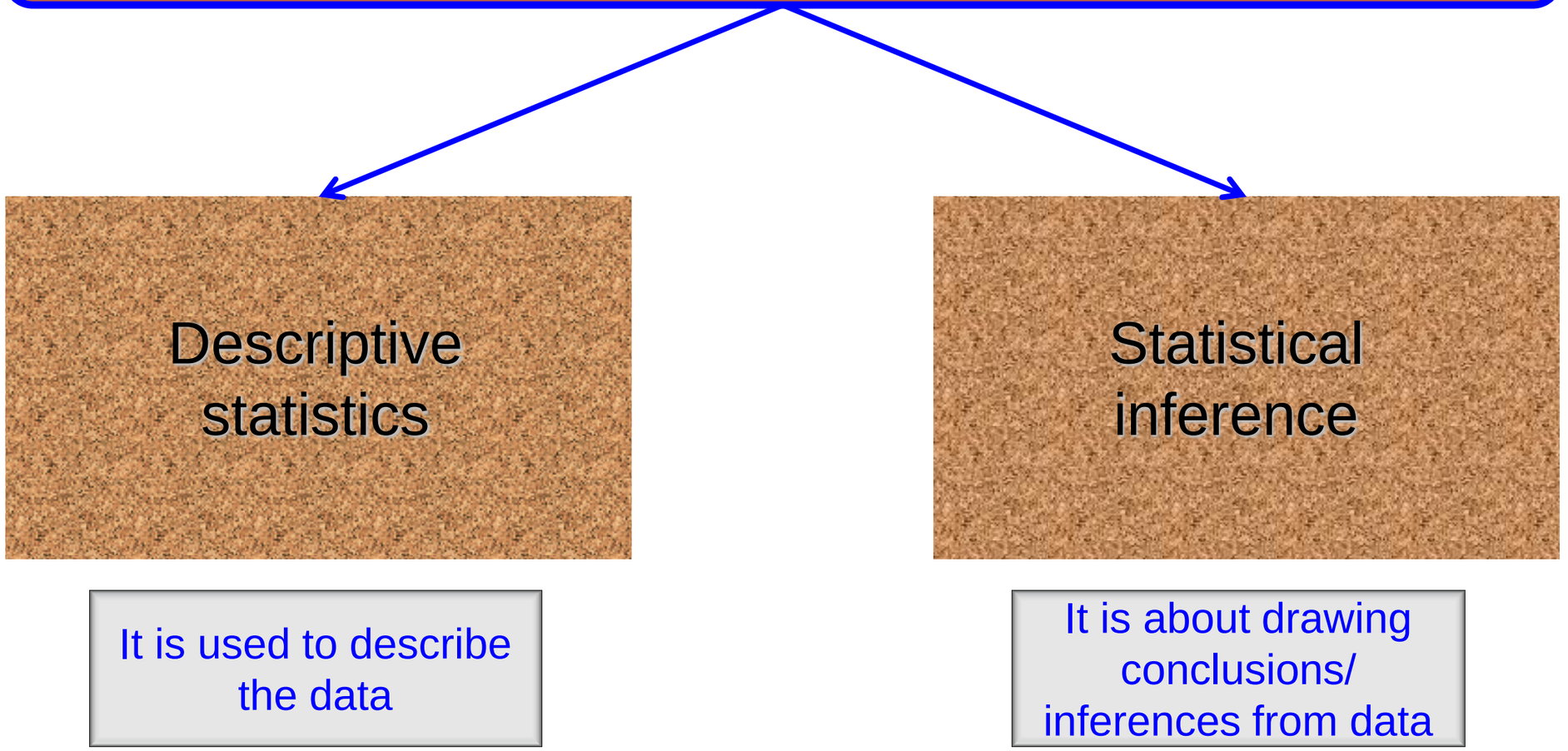
  Equally efficient way to summarizing data is in t




- **Continous Distribution** - When the raw data are measured on a quantitative scale, either interval, categories or classes must be designed for the data values before a frequency distribution can be formulated.

Now, this is the moment to see ***the branches in which statistics is divided:***



# Measures of Central Tendency

Measures of central tendency gives us an idea about the concentration of the values in the central part of the distribution. Plainly speaking, an average of a statistical series is the value of the variable which is representative of entire distribution.

The following are the measures of central tendency that are in common use:

- Mean
- Median
- Mode

# Measures of Central Location

- **Mean**: Average of a set of values ($\bar{x}$)

2, 1, 6

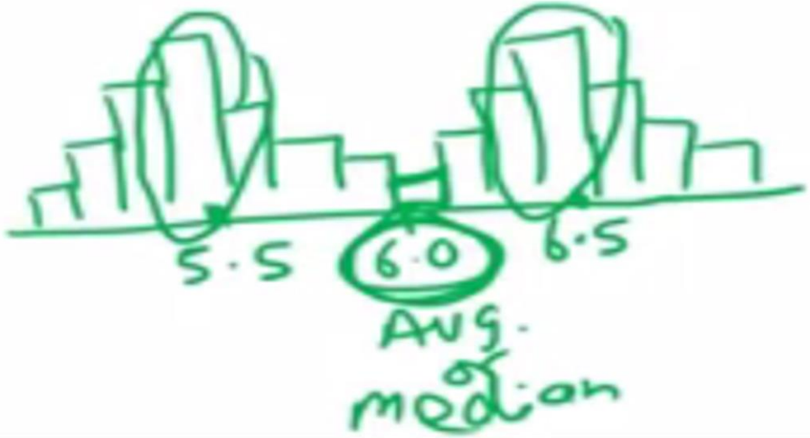
$\bar{x} = \frac{2+1+6}{3} = 3$

- **Median**: Midpoint in a string of sorted data, where 50% of the observations, or values, are below and 50% are above

1, 2, 6000

- **Mode**: The most frequently occurring value

Garment



Measures of central tendency: to understand the data through a single value (location).

## Mean (μ)

- The arithmetic average (add all of the scores together, then divide by the number of scores)
  .
- Example : To calculate average weight of a person .

| Child | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Age (years) | 8 | 10 | 9 | 9 | 10 | 9 | 11 | 11 | 11 | 11 |
| Weight (lbs.) | 52 | 64 | 65 | 70 | 72 | 76 | 80 | 84 | 88 | 94 |

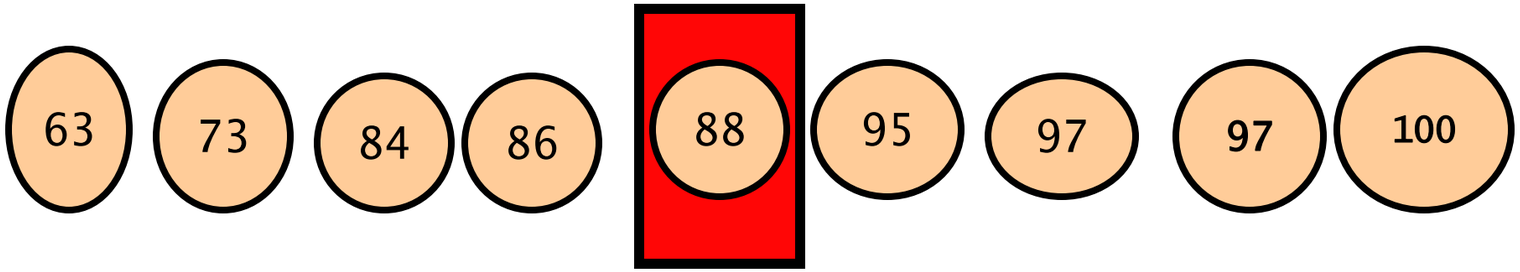
- *Mean* **= (52+64+65+70+72+76+80+84+88+94) / 10 = 745 / 10 = 74.5**

## Measures of central tendency

### Median

- The middle number (just like the median strip that divides a highway down the middle; 50/50)
- Often hear about the median price of housing or Median national income of a country .

Arrange values from least to greatest




Find the number that is in the middle.

**Half the numbers are**

**less than the median.**

The median is **88.**

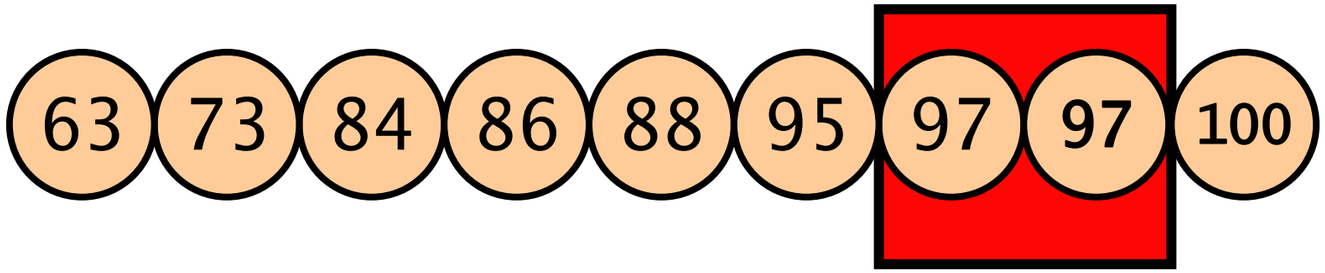
**Half the numbers are**

**greater than the median.**

## Measures of central tendency

### Mode

- The most frequently occurring number (score, measurement, value, cost)
- On a frequency distribution, it's the highest point .




Find the number that appears more or most frequently.

**The value 97 appears twice.**

All other numbers appear just once.

**97 is the MODE**

# Measures of Depression

In generally we can use measure of central tendency (i.e. Mean, Median, Mode) are for calculating the averages. By using that we can give an idea about the dataset. But it not revel the complete information about datasets. So by using the measures of depression we can get the proper idea about the dataset.

The following are the measures of depression:

- Range
- Interquartile Range
- Variance
- Standard Deviation
- Coefficient of variation

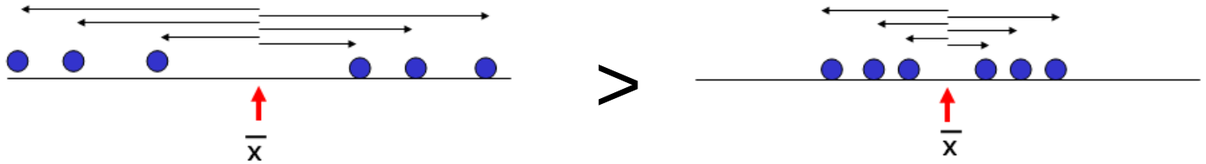
# Measures of Depression

## Range

- the sample **Range** is the difference between the largest and smallest observations in the sample .
- Example : easy to calculate;
  - Blood pressure example: min=113 and max=170, so the range=57 mmHg

## Variance

- The **sample variance**, $s^2$, is the arithmetic mean of the squared deviations from the sample mean:



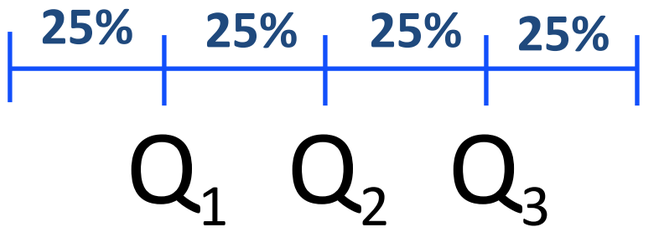

## Standard Deviation

- The **sample standard deviation, s,** is the square-root of the variance .
- **s** has the advantage of being in the same units as the original variable x .

# Understanding/ Exploring Data further using Quartiles, Deciles, Percentiles

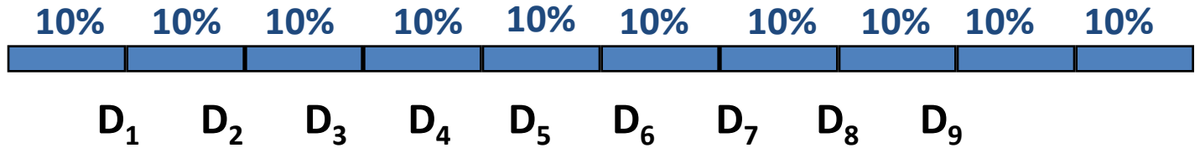
**Quartiles**

- Q1, Q2, Q3 divides ranked scores into four equal parts




**Deciles**

D1, D2, D3, D4, D5, D6, D7, D8, D9 divides ranked data into ten equal **parts .**




**Percentiles**

- Value of a variable below which a certain percent of observations fall.
- **Example:** The 20th percentile is the value (or score) below which 20 percent of the observations are found .

# Coefficient of Variation

A researcher is comparing two multiple-choice tests with different conditions.
The results from the two tests are:

Regular test : Mean 59.9 ,SD 10.2
Randomized test : Mean 44.8 , SD 12.7

Trying to compare the two test results is challenging.
Comparing standard deviations doesn't really work,
because the *means* are also different.

Calculation using the formula CV=(SD/Mean)*100 helps to make sense of the data:

Looking at the standard deviations of 10.2 and 12.7, you might think that the tests have similar results.
However, when you adjust for the difference in the means, the results have more significance:
Regular test: CV = 17.03
Randomized answers: CV = 28.35

# Coefficient of Variation (contd..)

The coefficient of variation can also be used to compare **variability** between different measures.

The coefficient of variation (CV) is a measure of relative variability.
It is the ratio of the standard deviation to the mean (average).
For example, the expression “The standard deviation is 17% of the mean” is a CV.

The CV is particularly useful when you want to compare results from two
different surveys or tests that have different measures or values.

In our example , Regular test has a CV of 17% and Randomized test has a CV of 28%,
We can say , Randomized test has more variation, relative to its mean.
So , the consistent results is from regular test .

| AGE | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 45 | 34 | 1.70 | 34 | 1.70 |
| 46 | 39 | 1.95 | 73 | 3.65 |
| 47 | 42 | 2.10 | 115 | 5.75 |
| 48 | 40 | 2.00 | 155 | 7.75 |
| 49 | 54 | 2.70 | 209 | 10.45 |
| 50 | 57 | 2.85 | 266 | 13.30 |
| 51 | 96 | 4.80 | 362 | 18.10 |
| 52 | 82 | 4.10 | 444 | 22.20 |
| 53 | 87 | 4.35 | 531 | 26.55 |
| 54 | 76 | 3.80 | 607 | 30.35 |
| 55 | 107 | 5.35 | 714 | 35.70 |
| 56 | 96 | 4.80 | 810 | 40.50 |
| 57 | 94 | 4.70 | 904 | 45.20 |
| 58 | 108 | 5.40 | 1012 | 50.60 |
| 59 | 85 | 4.25 | 1097 | 54.85 |
| 60 | 81 | 4.05 | 1178 | 58.90 |
| 61 | 60 | 3.00 | 1238 | 61.90 |
| 62 | 84 | 4.20 | 1322 | 66.10 |
| 63 | 75 | 3.75 | 1397 | 69.85 |
| 64 | 73 | 3.65 | 1470 | 73.50 |
| 65 | 75 | 3.75 | 1545 | 77.25 |
| 66 | 62 | 3.10 | 1607 | 80.35 |
| 67 | 60 | 3.00 | 1667 | 83.35 |
| 68 | 51 | 2.55 | 1718 | 85.90 |
| 69 | 66 | 3.30 | 1784 | 89.20 |
| 70 | 50 | 2.50 | 1834 | 91.70 |
| 71 | 49 | 2.45 | 1883 | 94.15 |
| 72 | 45 | 2.25 | 1928 | 96.40 |
| 73 | 38 | 1.90 | 1966 | 98.30 |
| 74 | 32 | 1.60 | 1998 | 99.90 |
| 75 | 2 | 0.10 | 2000 | 100.00 |

## Percentile Points/Groups

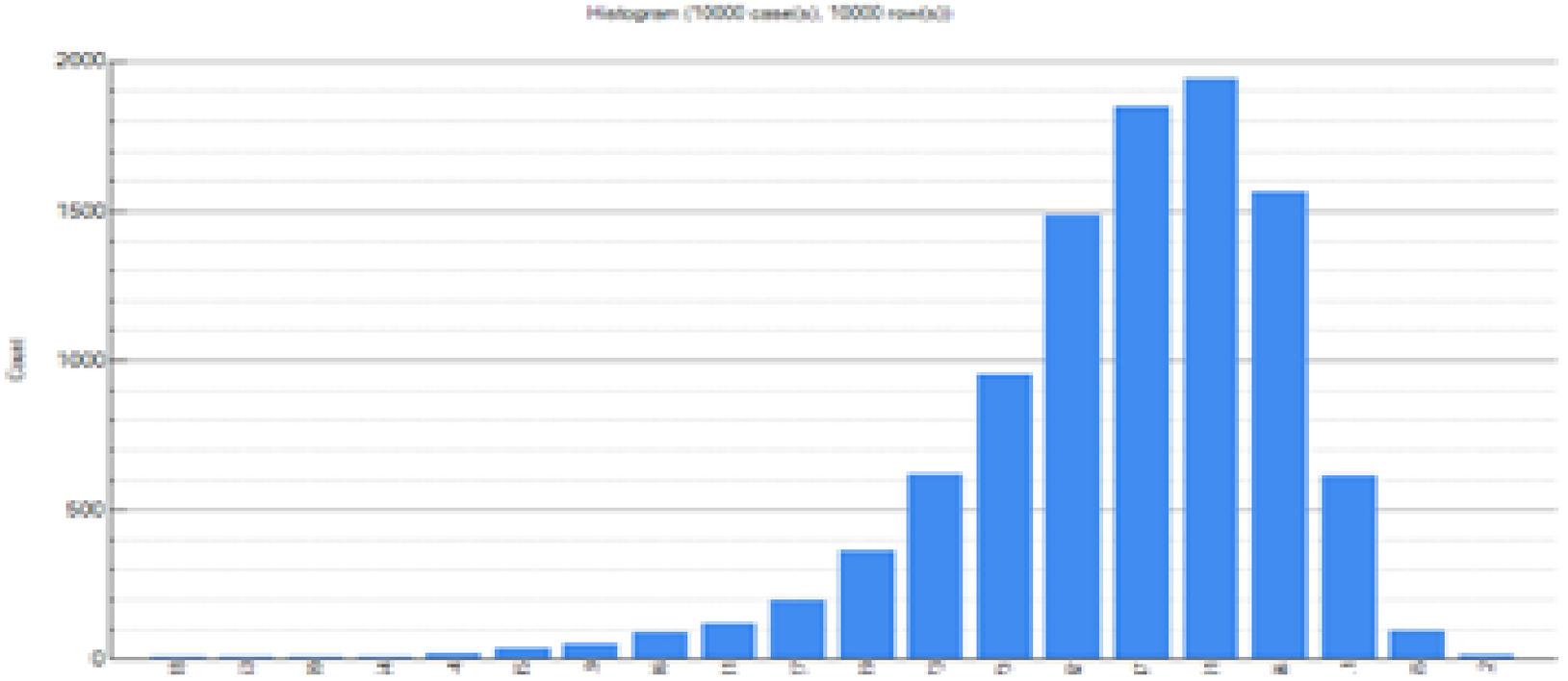
| Tertiles: | Age Groups |
|---|---|
| 0 – 33.3% | 45 to 54 |
| >33.3 – 66.7% | 55 to 62 |
| >66.7 to 100% | 63 to 75 |
| **Quartiles:** | |
| 0 - 25% | 45 to 52 |
| >25 – 50% | 53 to 57 |
| >50 – 75% | 58 to 64 |
| >75 to 100% | 65 to 75 |
| **Quintiles:** | |
| 0 - 20% | 45 to 51 |
| >20 – 40% | 52 to 55 |
| >40 – 60% | 56 to 60 |
| >60 to 80% | 61 to 65 |
| >80 – 100% | 66 to 75 |

juxt smart mandate

# Visualizations to understand Univariate data:

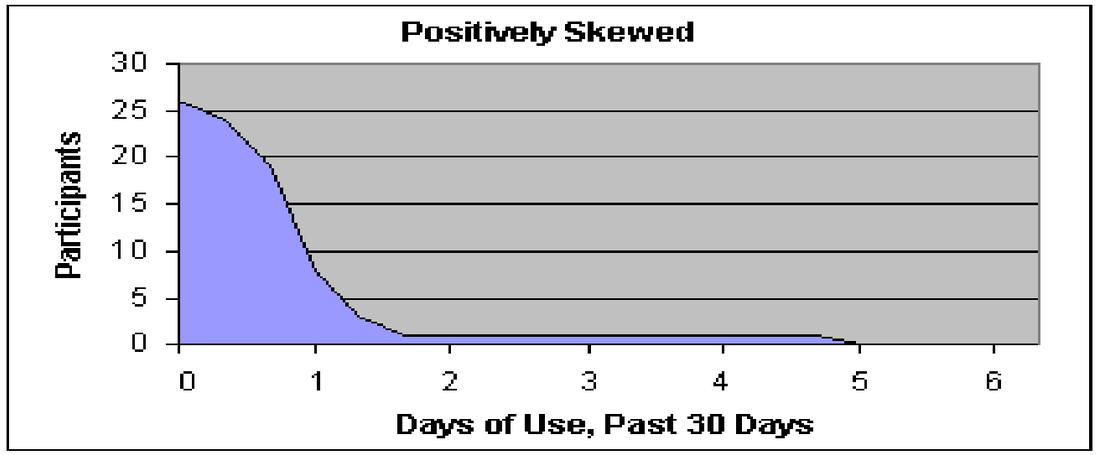
- ## Histogram (continuous):

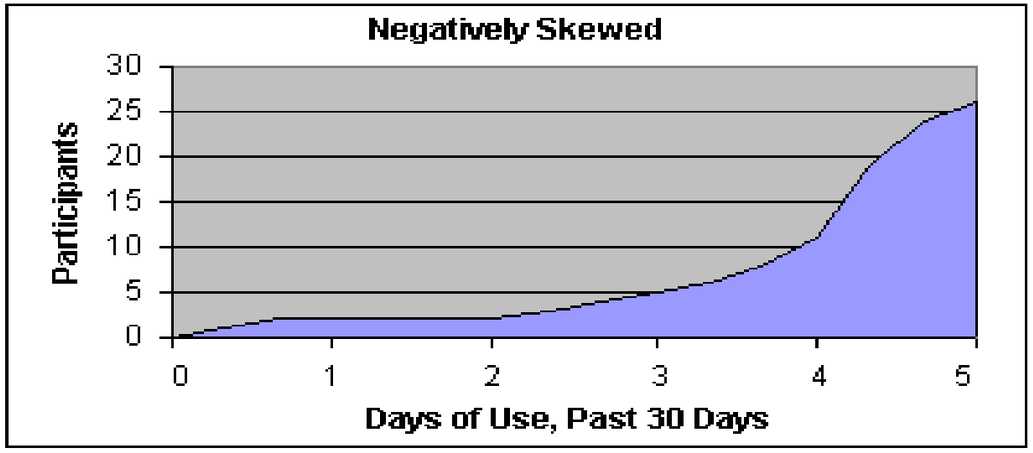
  Histogram is basically a plot that breaks the data into bins (or breaks) and shows frequency distribution of these bins.



# Understanding Shape of distribution - Skewness

It describes the shape of distribution .
Distributions that trail away to the left are negatively skewed and those that trail away to the right are positively skewed .

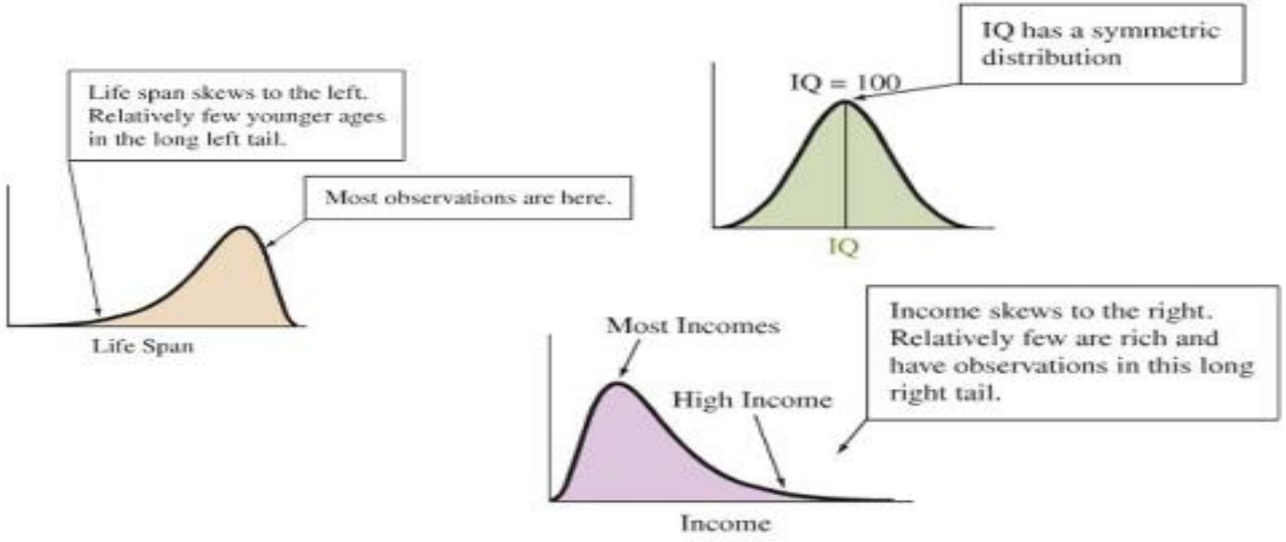



- The formula for the skewness of sample data is

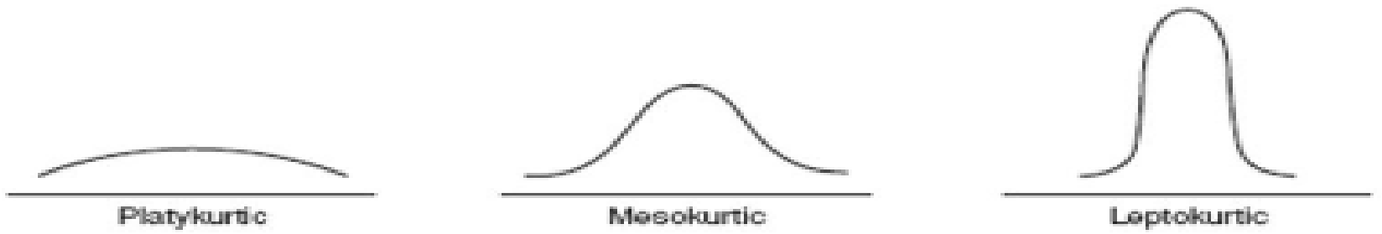
$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum \left( \frac{x_i - \bar{x}}{s} \right)^3$$

Distribution may be **positively or negatively skewed**. Limits for coefficient of skewness is ± 3.

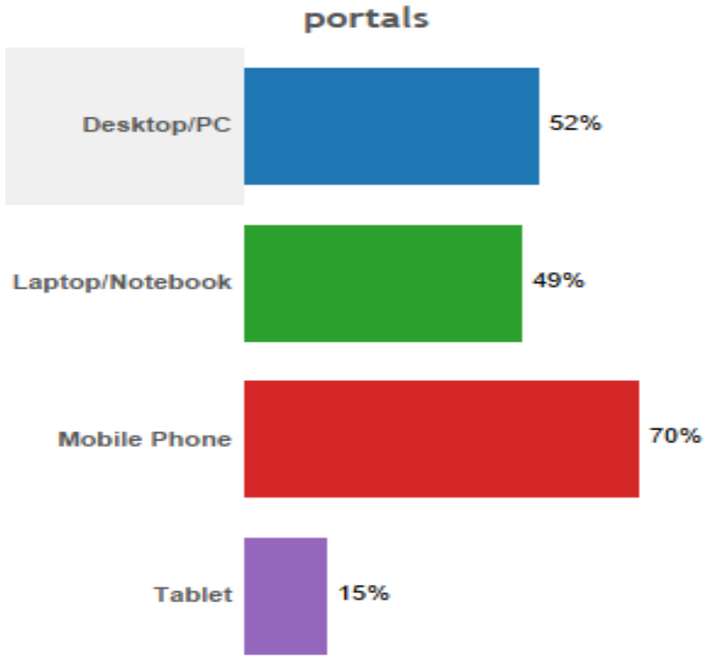


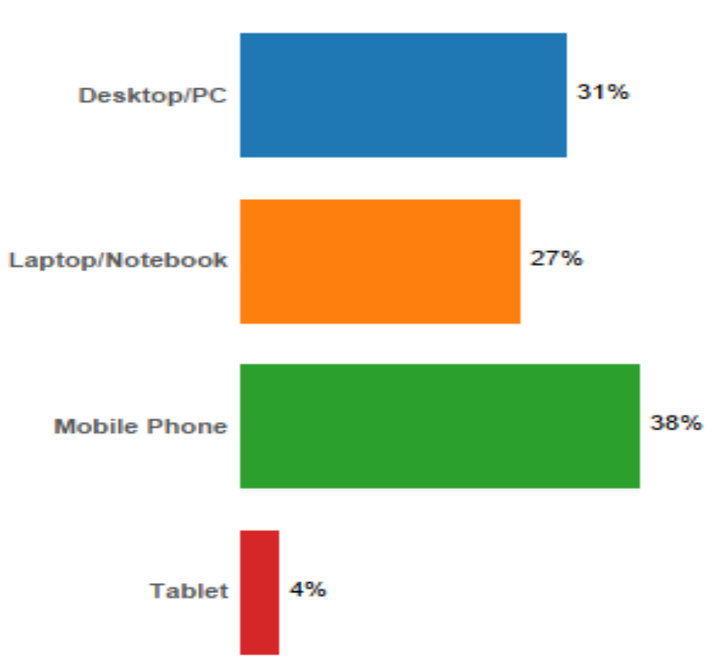
# Kurtosis

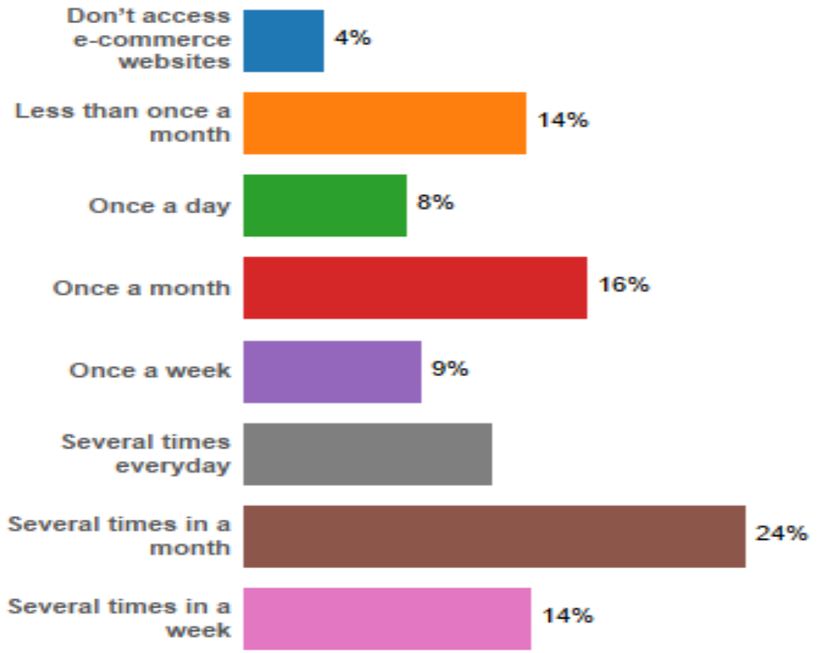
1. In statistics Kurtosis is the **degree of flatness** or **'peakedness'** in the region of mode of a frequency curve
2. It is measured **relative to the 'peakedness' of the normal curve**
3. It tells us the extent to which a distribution is more peaked or flat-topped than the normal curve
4. If the curve is more peaked than a normal curve it is called **Lepto Kurtic**. In this case items are more clustered about the mode.
5. If the curve is more flat-toped than the more normal curve, it is **Platy-Kurtic**.
6. The normal curve itself is known as **Meso Kurtic.**



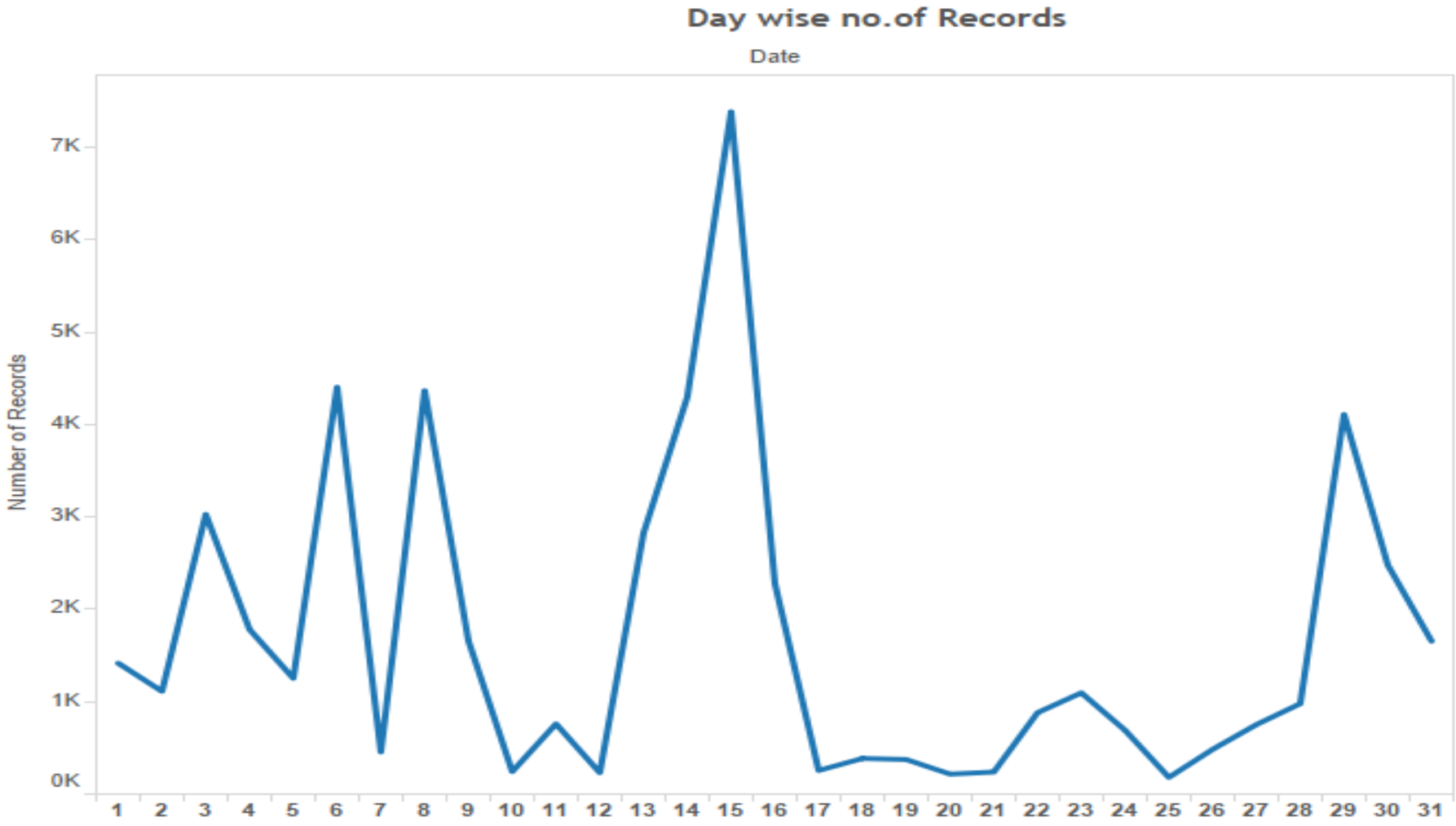
# Univariate Analysis: Simple Bar Charts



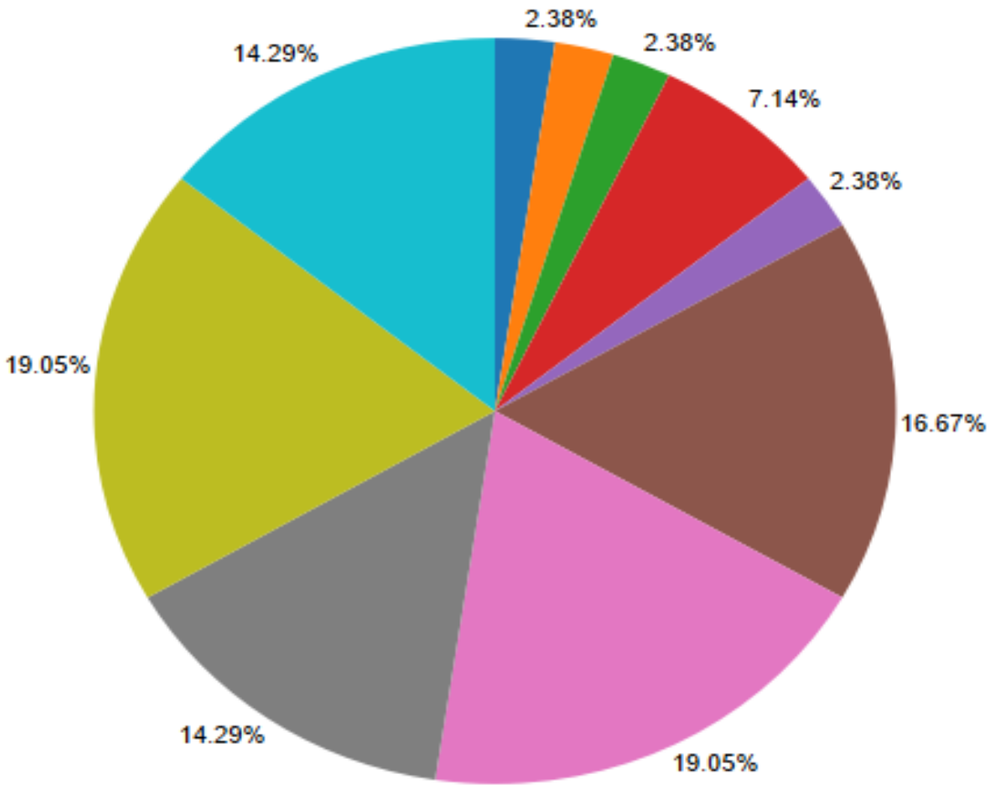



# Univariate analysis: Line Chart

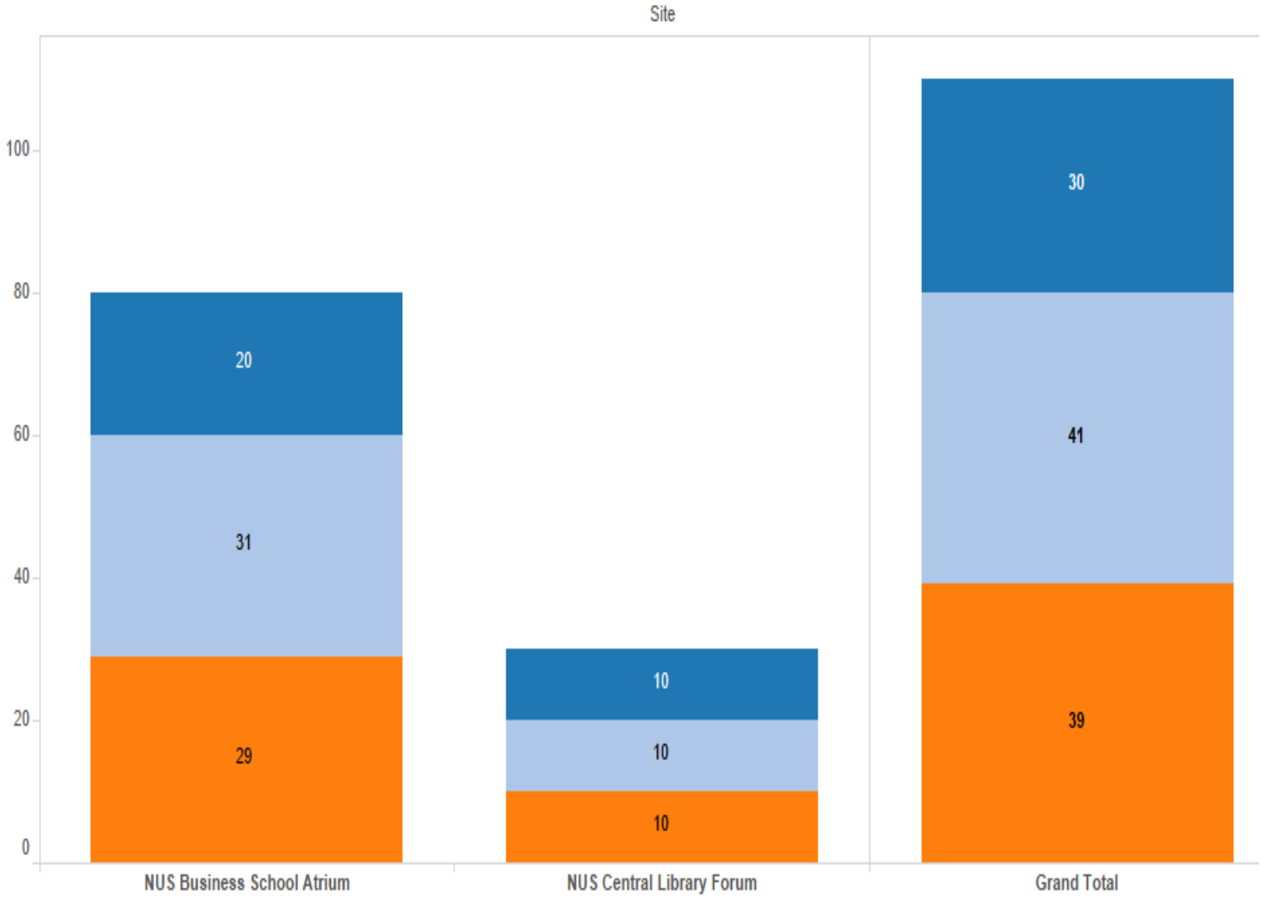


# Univariate Analysis: Pie and Stacked Chart




## Pie Chart

- Circular graph divided into sectors. Adds to 100%.
- Used to compare sectors within a particular segment




## Stacked Bar Chart

- Used to compare sectors across segments.
- Each breakup adds to 100%

juxt smart mandate

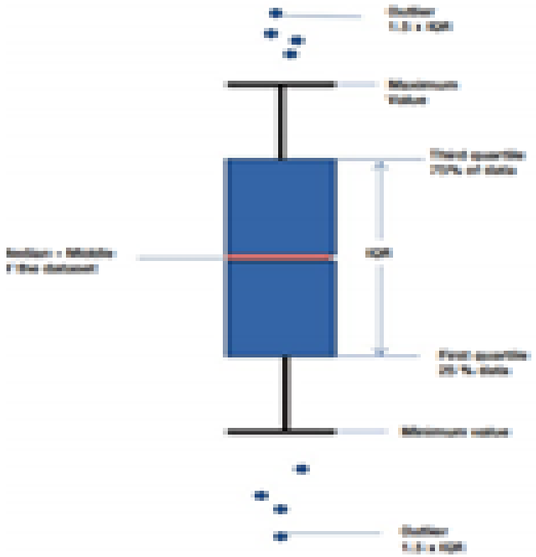
# Visualizations to understand Univariate data:

## ➢ Box Plot (continuous):

Box plots are a compact way to represent a distribution. The central rectangle spans the first and third quartile (interquartile, or IQR).

The line inside the rectangle shows the median.

The lines, also called whiskers, that are above and below the rectangle show the maximum and minimum of the data set.



# Visualizations to understand Univariate data:

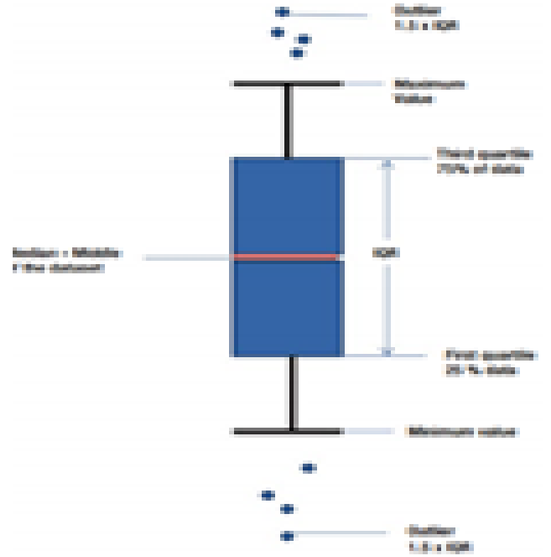
## ➢ Box Plot (continuous):

Normal data sets do not have a surprisingly high maximum value or low minimum value.

Outliers are generally outside the two whisker lines.

Tukey has provided following definitions for outliers:

- Outliers – 2/3 IQR above the third quartile or 2/3 IQR below the first quartile
- Suspected Outliers – 1.5 IQR above the third quartile or 1.5 IQR below the first quartile



# Understanding data for Bivariate variables:

1) For Categorical Vs Categorical variables – Proportion tables

| | Month of Date | | | | | |
|---|---|---|---|---|---|---|
| Weekday of Date | May 2015 | June 2015 | July 2015 | August 2015 | September 2015 | October 2015 |
| Sunday | 46 | 1,486 | 2,346 | 1,794 | 23,982 | 9,032 |
| Monday | 180 | 3,118 | 8,202 | 18,648 | 46,539 | 24,141 |
| Tuesday | 285 | 10,084 | 5,975 | 8,659 | 20,905 | 27,581 |
| Wednesday | | 2,111 | 18,169 | 8,548 | 54,978 | 3,293 |
| Thursday | | 955 | 8,632 | 20,083 | 26,165 | 23,036 |
| Friday | 474 | 2,607 | 5,877 | 8,015 | 22,333 | 29,878 |
| Saturday | 724 | 3,692 | 3,111 | 11,336 | 71,369 | 14,159 |
| Grand Total | 1,709 | 24,053 | 52,312 | 77,083 | 266,271 | 131,120 |

# Understanding data for Bivariate variables:

2) For Categorical Vs Continuous variables – Mean summary/ Percentages tables by Groups

| | | Gender | |
|---|---|---|---|
| | All Patients | Female | Male |
| Age | | | |
| Base:All Respondents | 490 | 149 | 341 |
| Upto 13 | 11.4% | 8.1% | 12.9% |
| 14 - 21 | 0.4% | - | 0.6% |
| 22 - 35 | 3.1% | 3.4% | 2.9% |
| 36 - 50 | 17.6% | 17.5% | 17.6% |
| 51 - 65 | 41.6% | 46.3% | 39.6% |
| 66+ | 25.9% | 24.8% | 26.4% |

| On a scale of 1 to 7, where 1 is “Not at all Important” and 7 is “Very Important,” how important are each of the following attributes to you in selecting a treatment? | Overall | Segment A | Segment B | Segment C | Segment D |
|---|---|---|---|---|---|
| Mean Summary | | | | | |
| N= | 360 | 134 | 124 | 63 | 39 |
| Resolves symptoms of psychosis within a few weeks | 5.7 | 5.4 | 6.2 | 5.0 | 6.5 |
| Resolves symptoms of psychosis fully | 5.8 | 5.5 | 6.2 | 5.1 | 6.6 |
| Controls symptoms of psychosis | 6.4 | 6.4 | 6.6 | 5.8 | 7.0 |
| Improves nighttime sleep | 5.6 | 5.4 | 6.0 | 4.8 | 6.6 |
| Increases daytime wakefulness | 5.2 | 4.9 | 5.6 | 4.0 | 6.5 |
| Neutral metabolic profile | 5.3 | 5.1 | 5.6 | 4.1 | 6.6 |

# Understanding data for Bivariate variables:

3) For Continuous Vs Continuous variables – Cross table by converting into different groups

| | | Age-Group | | | | | |
|---|---|---|---|---|---|---|---|
| | All Patients | Upto 13 | 14 - 21 | 22 - 35 | 36 - 50 | 51 - 65 | 66+ |
| # of days in hospital - groupings | | | | | | | |
| Base:All Respondents | 490 | 56 | 2 | 15 | 86 | 204 | 127 |
| 1 day | 37.5% | 44.8% | - | 40.0% | 35.9% | 40.3% | 31.1% |
| 2 days | 19.1% | 15.5% | - | 13.3% | 22.8% | 18.1% | 20.7% |
| 3 - 4days | 16.0% | 12.1% | - | 33.3% | 17.4% | 14.4% | 17.8% |
| 5 - 6days | 9.5% | 6.9% | - | 6.7% | 6.5% | 10.2% | 11.9% |
| 7 days or more | 18.0% | 20.7% | 100.0% | 6.7% | 17.4% | 17.1% | 18.5% |

# Multi variate analysis: 2x2 Scatter Chart

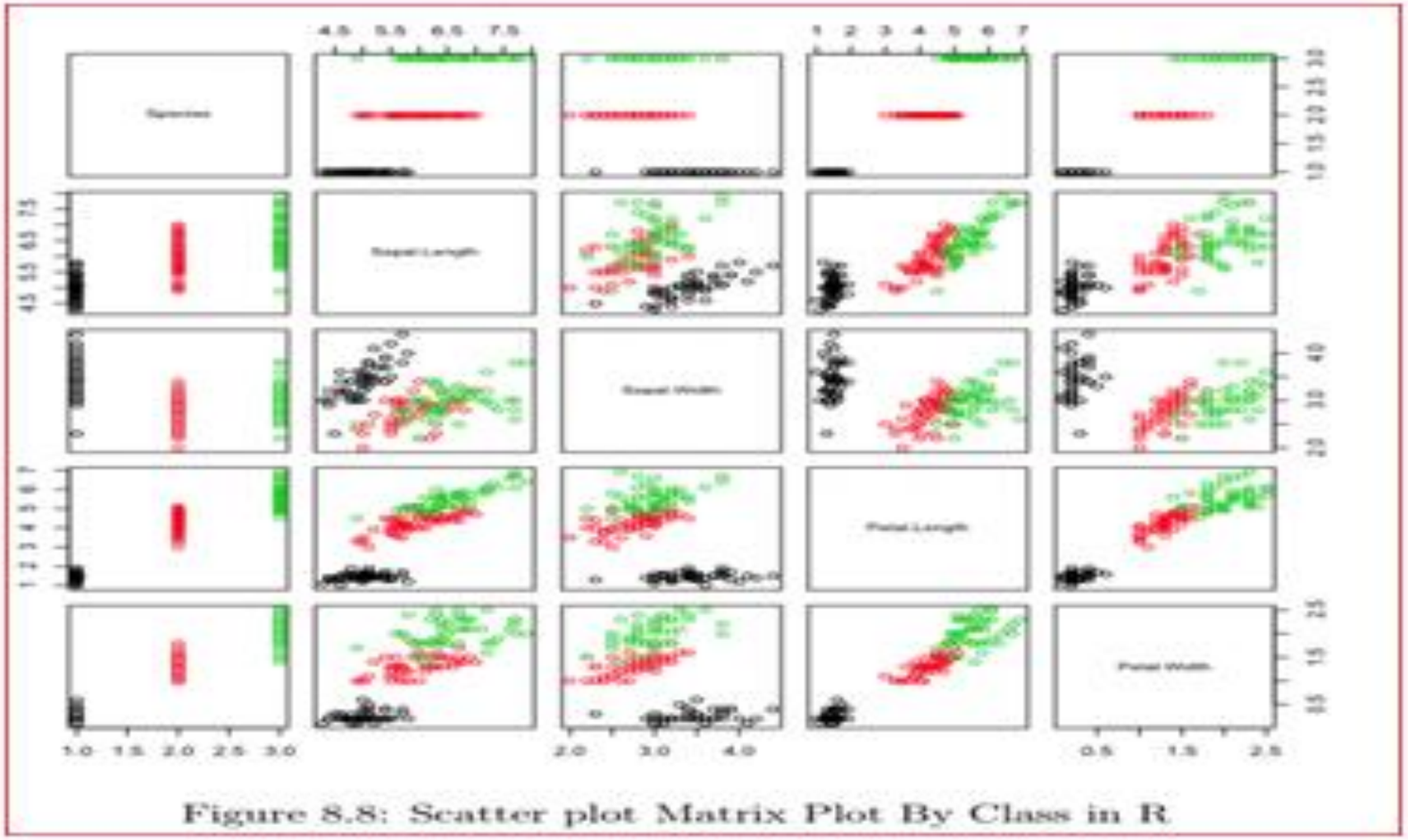
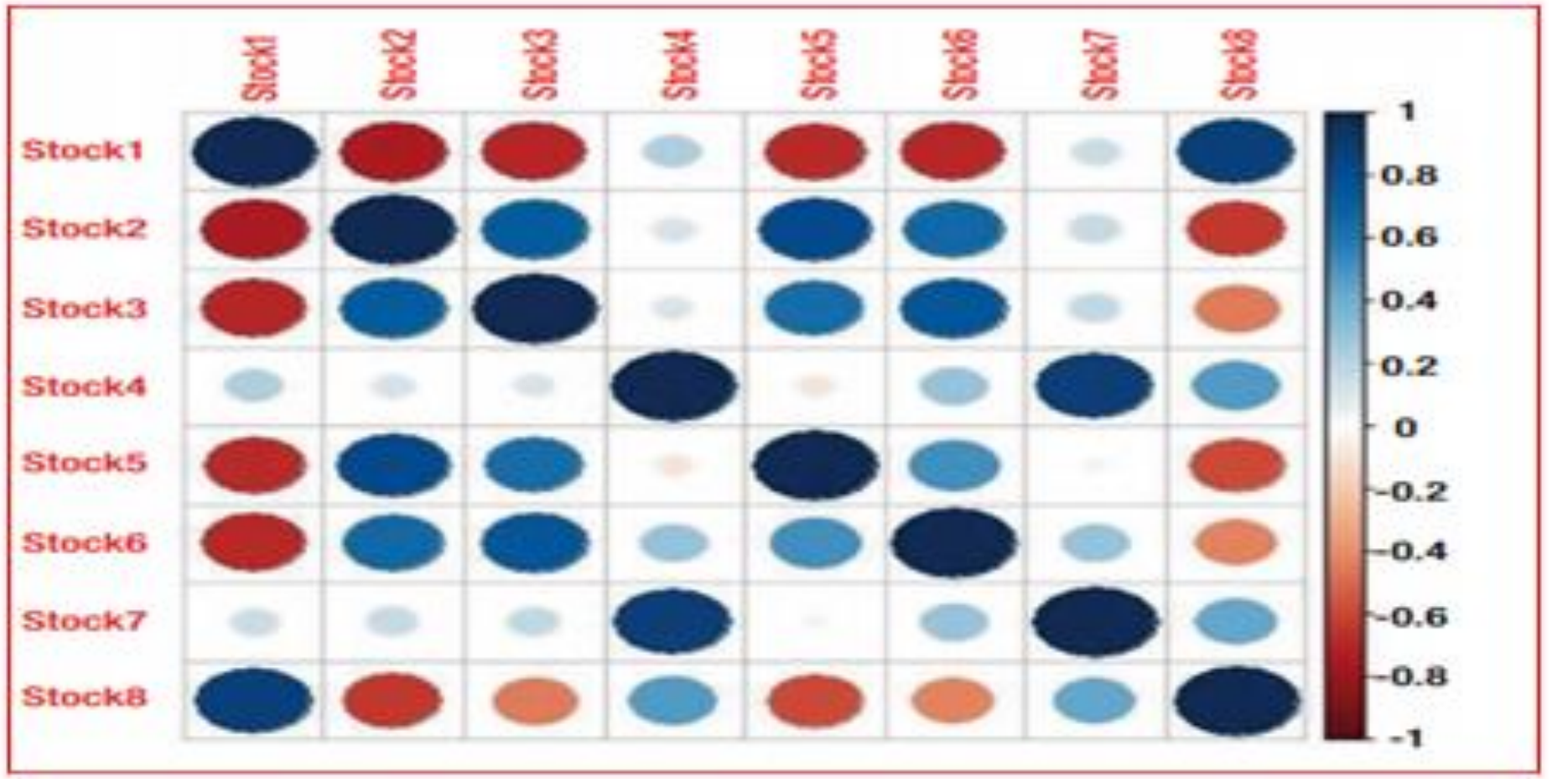


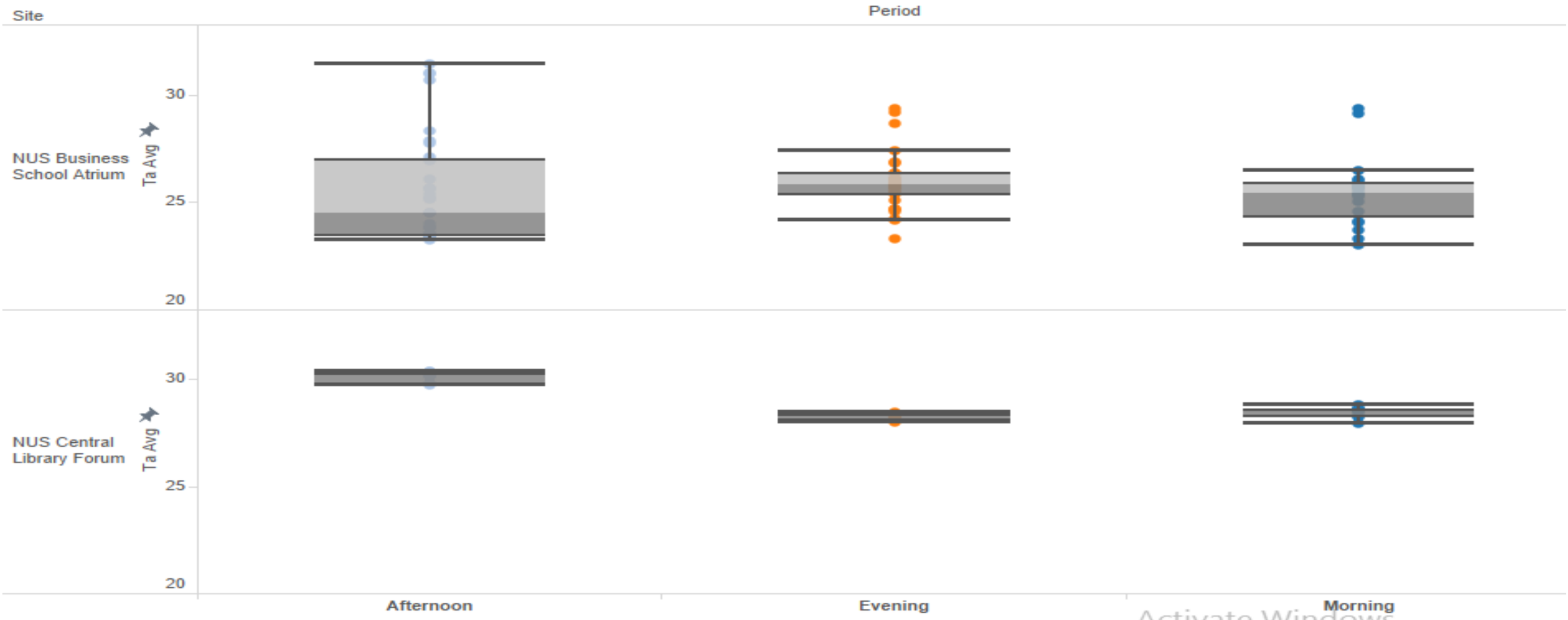
Figure 8.8: Scatter plot Matrix Plot By Class in R

# Multi variate analysis: 2x2 Scatter Chart

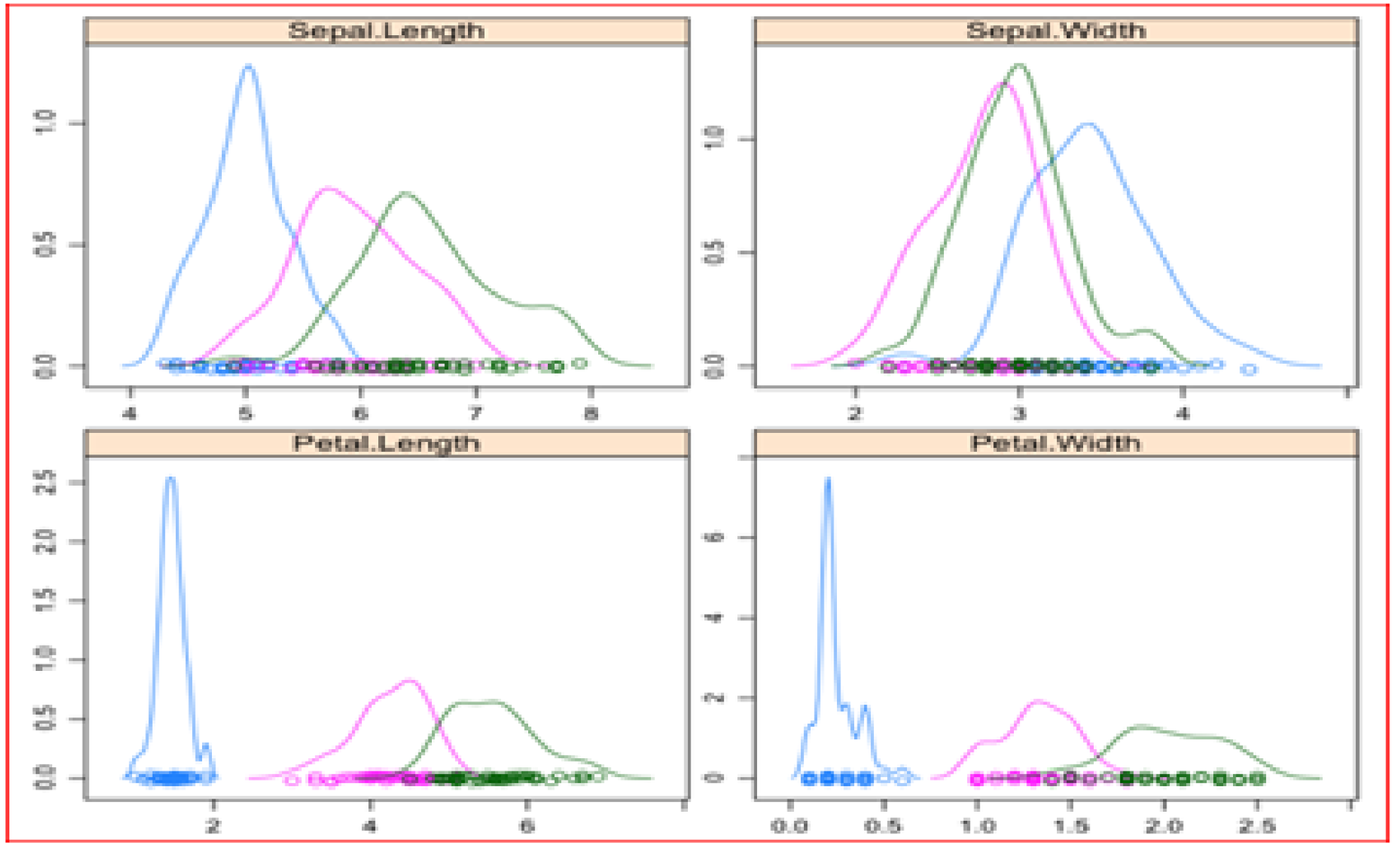
For Continuous Vs Continuous variables – Correlation Plot



# Multi variate analysis: Boxplot Chart



# Density Plots by Class



# Inferential Statistics

**Inference is the process of drawing conclusions or making decisions about a population based on sample results**

- Estimation
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 70 kg

# While doing so , we make mistakes ! :

Two types of decision errors:

Type I error = erroneous rejection of true $H_0$

Type II error = erroneous retention of false $H_0$

| Decision | Truth: $H_0$ true | Truth: $H_0$ false |
|---|---|---|
| Retain $H_0$ | Correct retention | Type II error |
| Reject $H_0$ | Type I error | Correct rejection |

α ≡ probability of a Type I error

β ≡ Probability of a Type II error

# Hypothesis Testing

- Is also called *significance testing*
- Tests a claim about a parameter using evidence (data in a sample )
- The procedure is broken into four steps

## Hypothesis Testing Steps

A. Null and alternative hypotheses
B. Test statistic
C. P-value and interpretation
D. Significance level (optional)

# Bi – Variate Analysis

| Features/ Response | Continuous | Categorical |
|---|---|---|
| Continuous | Person’s Correlation | T test/ ANOVA |
| Categorical | ANOVA | Chi-Square |

# Continuous Vs Continuous

- Correlation: Measure the association between two or more continuous variables.

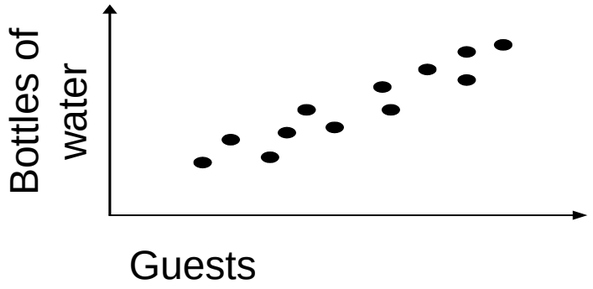
Correlation Close to 1: Strong +ve relationship

Correlation Close to -1: Strong –ve relationship

Correlation near to 0: Less or no relationship

## Positive linear relation

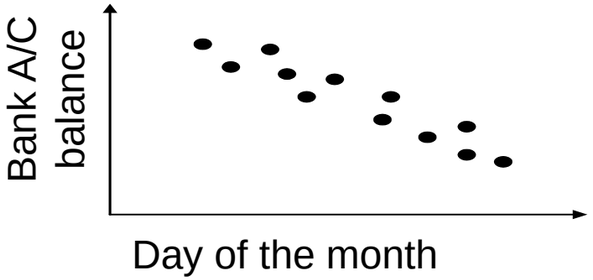
The two variables increase at the same time




The higher the number of guests, the higher number of bottles of water needed

## Negative linear relation

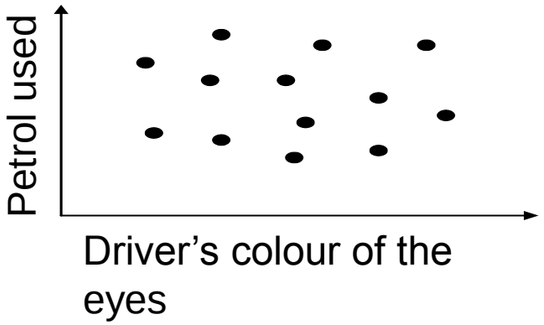
When one of the variables increases, the other decreases




The higher day of the month, the lower balance of my bank account

## No relation

The dots do not reflect any relation between data




There is no relation between the use of petrol of a car and the colour of the eyes of its driver

juxt smart mandate

# Categorical Vs Categorical:

## Chi-Square Test

This test is used to derive the statistical significance of relationship between the variables. Also, it tests whether the evidence in the sample is strong enough to generalize that the relationship for a larger population as well.

Chi-square is based on the difference between the expected and observed.
frequencies in one or more categories in the two-way table

It returns probability for the computed chi-square distribution with the degree of freedom.

## Case 3: Categorical versus categorical

### Chi-sq statistic

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Chi ^ 2 = chi-squared statistic

$O_i$ = the number of observation of type i.

$E_i$ = the expected frequency of type i.

n = the number of cells in the table.

# Categorical Vs Continuous:

While exploring relation between categorical and continuous variables, we can draw box plots for each level of categorical variables.

If levels are small in number, it will not show the statistical significance.

To look at the statistical significance we can perform Z-test, T-test or ANOVA.

Z-Test/ T-Test: Either test assess whether mean of two groups are statistically different from each other or not. If the probability of Z is small then the difference of two averages is more significant.

The T-test is very similar to Z-test but it is used when number of observation for both categories is less than 30.

$$Z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

$$S^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2}$$

Where:

- $\bar{x}_1 - \bar{x}_2$: Averages
- $s_1^2 - s_2^2$ : Variances
- $N_1, N_2$ :counts

t:has distribution with $N_1+N_2 - 2$ degree of freedom

juxt smart mandate

# ANOVA:

It assesses whether the average of more than two groups is statistically different.

## Case 2 : Continuous versus categorical

## ANOVA

Null Hypothesis. $H_o$: All population means are equal

Nternate Hypothesis $H_2$ : At least one population mean is different from the res

$$\text{F-ratio} = \frac{\text{Mean between group sum of squares}}{\text{Mean within group sum of squares}}$$

# More on T tests:

## One Sample T test:

A single-sample t-test compare a sample a known figure, for example where Measures of a manufactured item are compared against the required standard.

The general steps of testing hypothesis must be followed.

- HO : sample mean = population mean.
- Degrees of freedom = n - 1

$$t = \frac{X - \mu}{SE}$$

# More on T tests (contd..):

## 2 Sample t test for Proportions:

| | Sample 1 | Sample 2 |
|---|---|---|
| Sample size | 750 | 500 |
| Percentage | 28.0% | 18.5 % |
| 95% Lower confidence limit : | 22.9% | 15.1% |
| 95% Upper confidence limit : | 29.1% | 21.9% |
| t-value : | 3.087 | |
| Probaldility : | 0.002 | |
| Significance level : | 99% | |

where $P = \frac{n_1p_1 + n_2p_2}{n_1+n_2}$

The significance level is determined by Calculating a t- statistic using.

$$t = \frac{p_1 - p_2}{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}$$

$P_1$ = percentage for sample 1

$P_2$ = percentage for sample 2

$n_1$ = sample size for sample 1

$n_2$ = sample size for sample 2

juxt smart mandate

# More on T tests (contd..):

## Paired t test:

Matched-pair t-test: When samples appear in pairs(eg.: before-and-after).

To compare between the values (readings) of one sample but in 2 occasions.

$$t = \frac{\bar{d}}{\frac{sd}{\sqrt{n}}}$$

$$sd = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}}$$

# Summary of t test:

| Test | Purpose | Example |
|---|---|---|
| 1-sample 1 | Tests whether the mean of a single population is equal to a target value | Is the mean height of female college students greater than 5.5 feet? |
| 2-sample 1 | Tests whether the difference between the mans of two independent populations is equal to target value | Does the mean height of female college students significantly differ from the mean height of male college students? |
| Paired 1 | Tests whether the mean of the difference between dependent or paired observations is equal to a target value | If you measure the weight of male college students before and after each subject takes a weight loss significant enough to conclude that the pill works? |
| t-test regression output | Tests whether the value of coefficients In the regression equation differ Significantly from zero | Are high school SAT test scores Significant predictors of college GPA? |

# Non-Parametric test Alternatives:

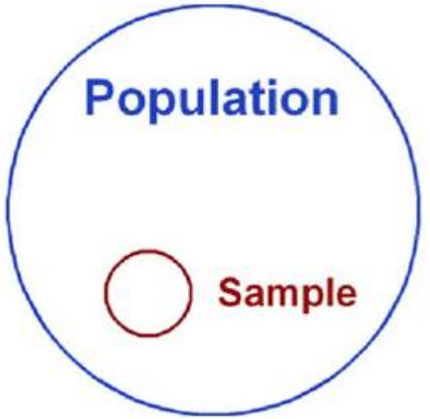
| Parametric Test | Non-Parametric Test |
|---|---|
| Two independent sample t-test | Mann-Whitney U test |
| Two dependent sample t-test | Wilcoxon Signed rank test |
| One way ANOVA | Kruskal-Walls test |
| Two way ANOVA | Friedman’s |
| Pearson’s correlation | Spearman correlation |

# Population

- A population can be defined as including all people or items with the characteristic one wishes to understand.
- Because there is very rarely enough time or money to gather information from everyone or everything in a population, the goal becomes finding a representative sample (or subset) of that population.

# Sampling

- Process of choosing a representative portion of the entire population
- Involve selecting a group of people, events, behaviors



# Why Sample?

Get information about large populations

- Lower cost
- More accuracy of results
- High speed of data collection
- Availability of Population elements
- Less field time
- When it is impossible to study the whole population

# What is Good Sample?

The sample must be:

- *Representative* of the population
- Appropriately sized
- Unbiased
- Random

# Types of Sampling

- ***Probability Sample****:* a method of sampling that uses of random selection so that all units/cases in the population have an equal probability of being chosen.
- ***Non-Probability Sample****:* does not involve random selection and methods are not based on the rationale of probability theory.

# Probability Sample

- Simple random sample (with and without replacement)
- Systematic random sample
- Stratified random sample
- Cluster sample

# Non-Probability Sample

- Convenience samples (ease of access): sample is selected from elements of a population that are easily accessible.
- Purposive sample (Judgmental Sampling): chose who we think should be in the study

# Simple Random Sampling

- Applicable when population is small, homogeneous & readily available
- All subsets of the frame are given an equal probability. Each element of the frame thus has an equal probability of selection. A table of random number or lottery system is used to determine which units are to be selected.

**Advantage:**

- Easy method to use
- No need prior information of population
- Equal chance of selection to every element

**Disadvantage:**

- If sampling frame large, this method impracticable.

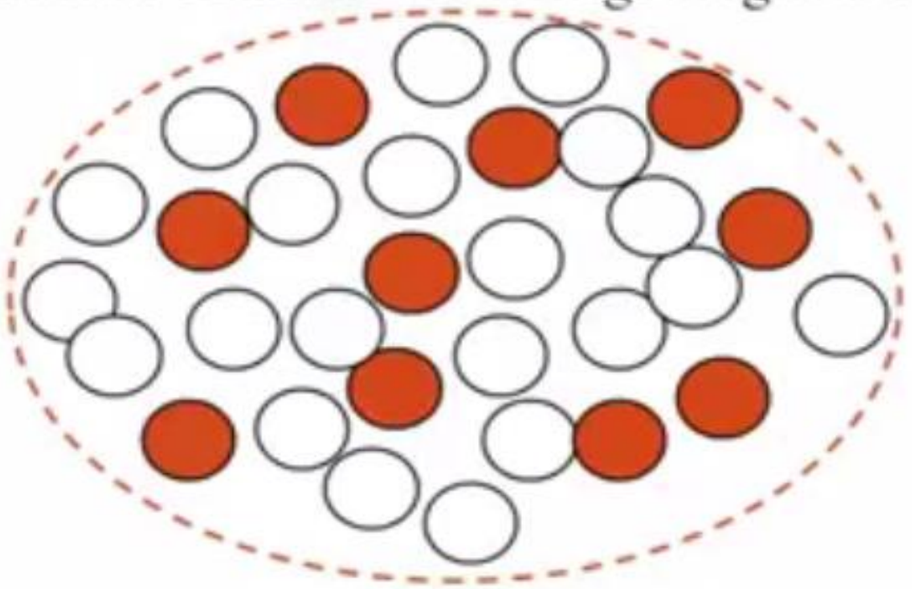
**Example:**

This is similar to the lottery. If the “population” is everyone who has bought a lottery ticket, then each person has an equal chance of winning the lottery (assuming they all have one ticket each).

# Simple Random Sampling

## Simple Random Sampling

Example: To estimate the average height of the class, select 10 students at random. Calculate the average height of the sample



Each item has equal probability of being selected

# Systematic Random Sampling

- Similar to simple random sample. No table of random numbers – select directly from sampling frame. Ratio between sample size and population size.

**Advantage:**

- Easy to select
- Suitable sampling frame can be identified easily
- Sample evenly spread over entire reference population
- Cost effective

**Disadvantage:**

- Each element does not get equal chance

**Example:**

If you take every nth name, you will get a systematic sample of the correct size. If, for example, you wanted to sample 150 children from a school of 1,500, you would take every 10th name.

# Systematic Random Sampling

## Systematic Random Sampling

Example: I ask every 10th person their opinion on state of the economy

Every "nth" item is sampled for study

# Stratified Random Sampling

- The population is divided into two or more groups called strata, according to some criterion, such as geographical location, grade level, age or income and subsample are randomly selected from each strata

**Advantage:**

- Enhancement of representativeness to each sample
- Higher statistical efficiency
- Easy to carry out

**Disadvantage:**

- Time consuming and expensive
- Classification Error

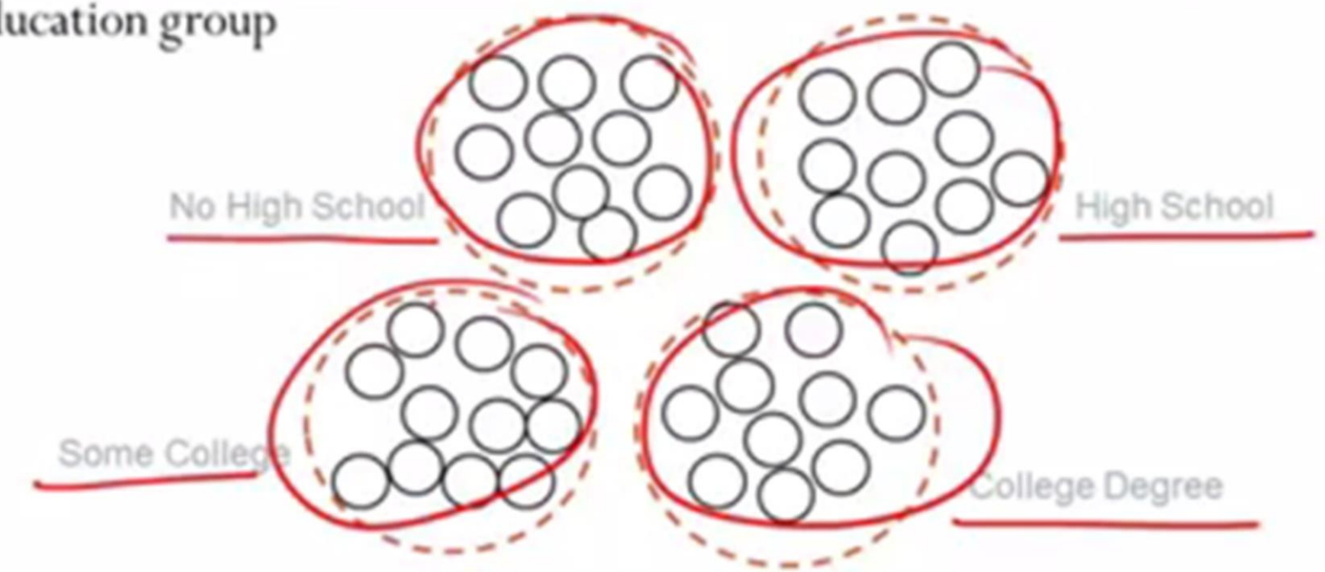
**Example:**

Students studying English Literature may spend more money on books than engineering students so if we use a very large percentage of English students or engineering students then our results will not be accurate.

# Stratified Random Sampling

## Stratified Random Sampling

Example: To estimate the average income of people in the US, break the population of the US into levels of education. Then sample randomly within each education group




Population is "stratified" into groups with random selection within each group

# Cluster Sampling

- Cluster sampling is an example of 'two-stage sampling' .
  - First stage a sample of areas is chosen;
  - Second stage a sample of respondents *within* those areas is selected.
- Population divided into clusters of homogeneous units, usually based on geographical contiguity.
- Sampling units are groups rather than individuals.
- A sample of such clusters is then selected.
- All units from the selected clusters are studied.
- The population is divided into subgroups (clusters) like families. A simple random sample is taken of the subgroups and then all members of the cluster selected are surveyed

**Advantage:**

- Cuts down on the cost of preparing a sampling frame. This can reduce travel and other administrative costs.

**Disadvantage:**

- sampling error is higher for a simple random sample of same size.

# Cluster Sampling contd..

- **Cluster sampling**: selecting a sample based on specific, naturally occurring groups (clusters) within a population.

  - Example: randomly selecting 20 hospitals from a list of all hospitals in India.

- **Multi-stage sampling**: cluster sampling repeated at a number of levels.

  -Example: randomly selecting hospitals by county and then a sample of patients from each selected hospital.

- Complex form of cluster sampling in which two or more levels of units are embedded one in the other.
- First stage, random number of districts chosen in all states.
- Followed by random number of talukas, villages.
- Then third stage units will be houses.
- All ultimate units (houses, for instance) selected at last step are surveyed.

# Non-Probability Sampling

- **Convenience Sampling:** as the name suggests, this involves collecting a sample from somewhere convenient to you: the mall, your local school, your church. Sometimes called accidental sampling, opportunity sampling or grab sampling.

- **Purposive Sampling:** where the researcher chooses a sample based on their knowledge about the population and the study itself. The study participants are chosen based on the study's purpose. There are several types of purposive sampling.
  - **Critical Case Sampling:** collecting cases that are likely to give you the most information about the phenomenon you are studying.
  - **Expert Sampling:** Sampling to include only those with expertise in a certain area.
  - **Extreme Case Sampling**: this technique focuses on participants with unique or special characteristics.
  - **Homogeneous Sampling**: collecting a very specific set of participants. For example, age 20-24, college educated, male.
  - **Maximum Variation Sampling**: collecting a wide range of participants with different viewpoints to study a certain phenomenon. Can uncover common themes.

# Feature Engineering

- Data Cleaning
- Scaling
- Outlier identification
- Missing value identification
- Data transformations
- Feature selection

# Data Cleaning

The process of insuring that all responses fall within allowable ranges

- Out-of-range checks
- Frequency table
- Consistency checks insuring
- Skip patterns
- Missing data
- Responses fall within allowable ranges

# Variable Transformation

In data modelling, transformation refers to the replacement of a variable by a function. For instance, replacing a variable x by the square / cube root or logarithm x is a transformation.

In other words, transformation is a process that changes the distribution or relationship of a variable with others.

When should we use Variable Transformation?

Below are the situations where variable transformation is a requisite:
When we want to **change the scale** of a variable or standardize the values of a variable for better understanding. While this transformation is a must if you have data in different scales, this transformation does not change the shape of the variable distribution

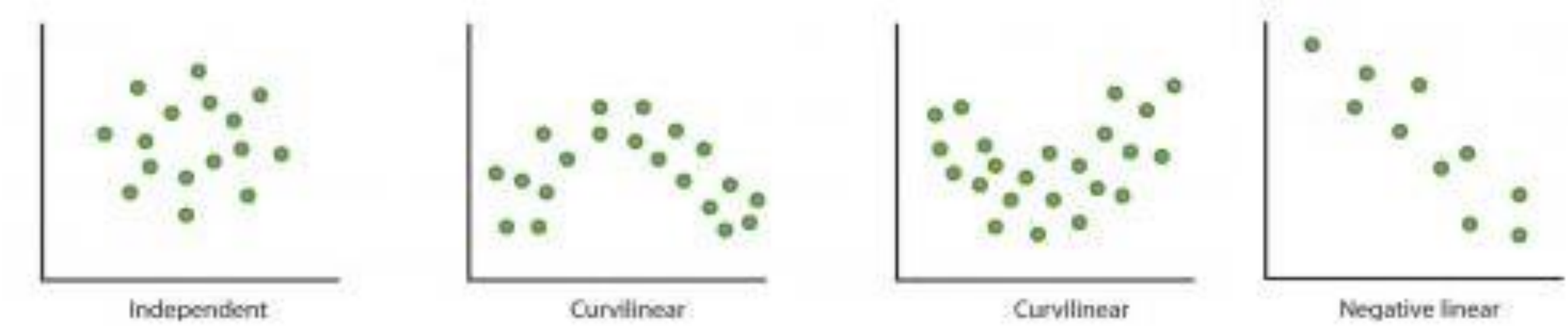
## Scaling

- Confining numerical variables into a certain range
- Commonly used methods are
  - Standard (Z) scaling
  - Min-max scaling
  - Log scaling
- Not necessarily needed for tree based models
- Must for models like Neural nets, kNN

# Transformation – Non Linear to Linear

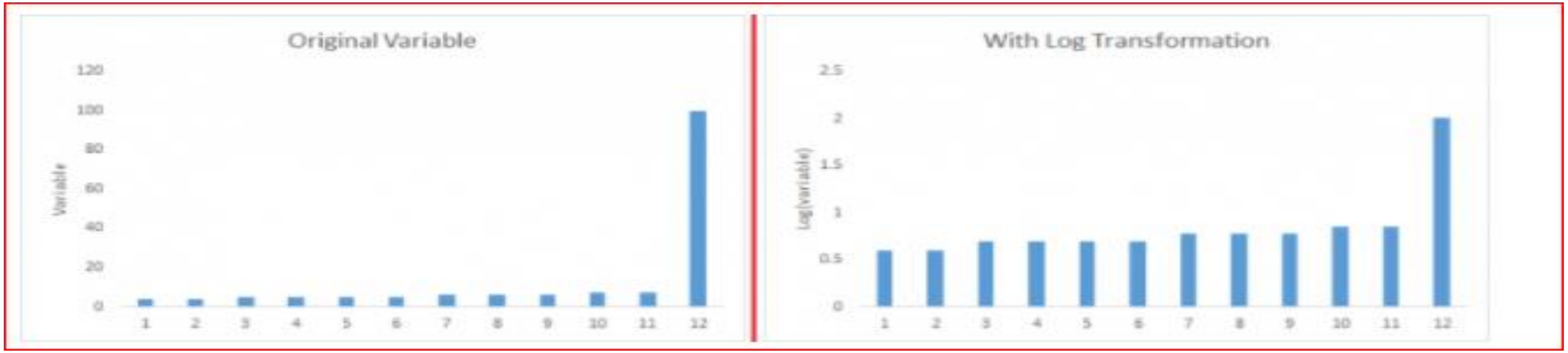
When we can transform complex non-linear relationships into linear relationships. Existence of a linear relationship between variables is easier to comprehend compared to a non-linear or curved relation.

- Transformation helps us to convert a non-linear relation into linear relation.
- Scatter plot can be used to find the relationship between two continuous variables.
- These transformations also improve the prediction.
- Log transformation is one of the commonly used transformation technique in these situations.



# Symmetric distribution is preferred over skewed distribution

Symmetric distribution is preferred over skewed distribution as it is easier to interpret and generate inferences.
Some modelling techniques requires normal distribution of variables. So whenever we have a skewed distribution, we can use transformations which reduce skewness.
For right skewed distribution, we take square/cube root or logarithm of variable and for left skewed, we take square cube or exponential of variables.




Logarithm/ Square / Cube root/ Binning/ Creating derived variable/ Creating dummy variables

# Outlier identification

Outliers are extreme values that deviate from other observations on data, they may indicate a variability in a measurement, experimental errors or a novelty.

In other words, an outlier is an observation that diverges from an overall pattern on a sample.

## Types of outliers

Outliers can be of two kinds: Univariate and Multivariate.

- Univariate outliers can be found when looking at a distribution of values in a single feature space.
- Multivariate outliers can be found in a n-dimensional space (of n-features).
- Looking at distributions in n-dimensional spaces can be very difficult for the human brain, that is why we need to train a model to do it for us.

# Most common causes of outliers on a data set

- Data entry errors (human errors)
- Measurement errors (instrument errors)
- Experimental errors (data extraction or experiment planning / executing errors)
- Intentional (dummy outliers made to test detection methods)
- Data processing errors (data manipulation or data set unintended mutations)
- Sampling errors (extracting or mixing data from wrong or various sources)
- Natural (not an error, novelties in data)

# What is the impact of Outliers on a dataset?

Outliers can drastically change the results of the data analysis and statistical modelling.
There are numerous unfavourable impacts of outliers in the data set:

- It increases the error variance and reduces the power of statistical tests
- If the outliers are non-randomly distributed, they can decrease normality
- They can bias or influence estimates that may be of substantive interest
- They can also impact the basic assumption of Regression, ANOVA and other

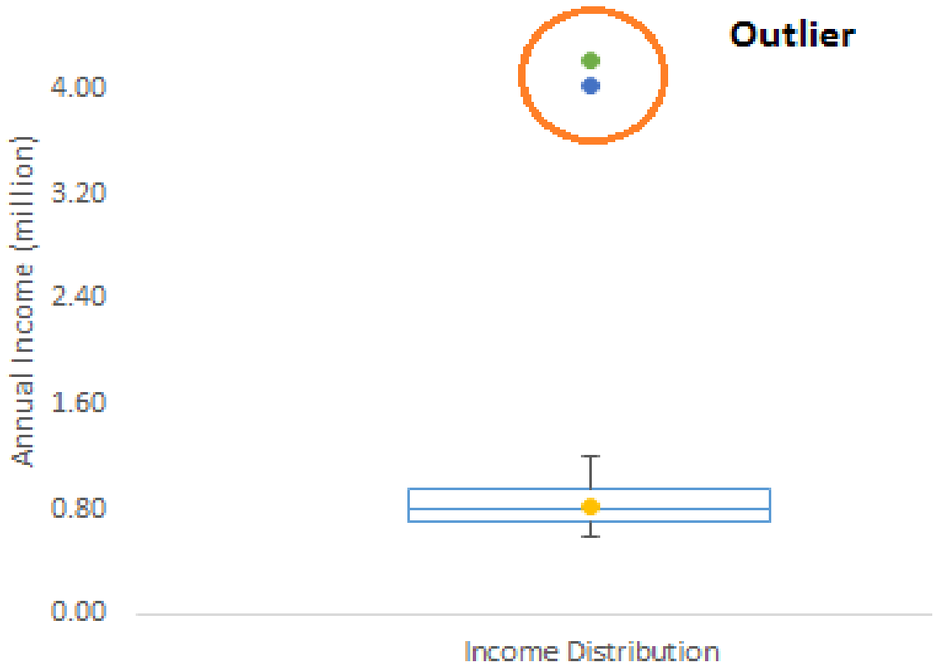
statistical model assumptions.

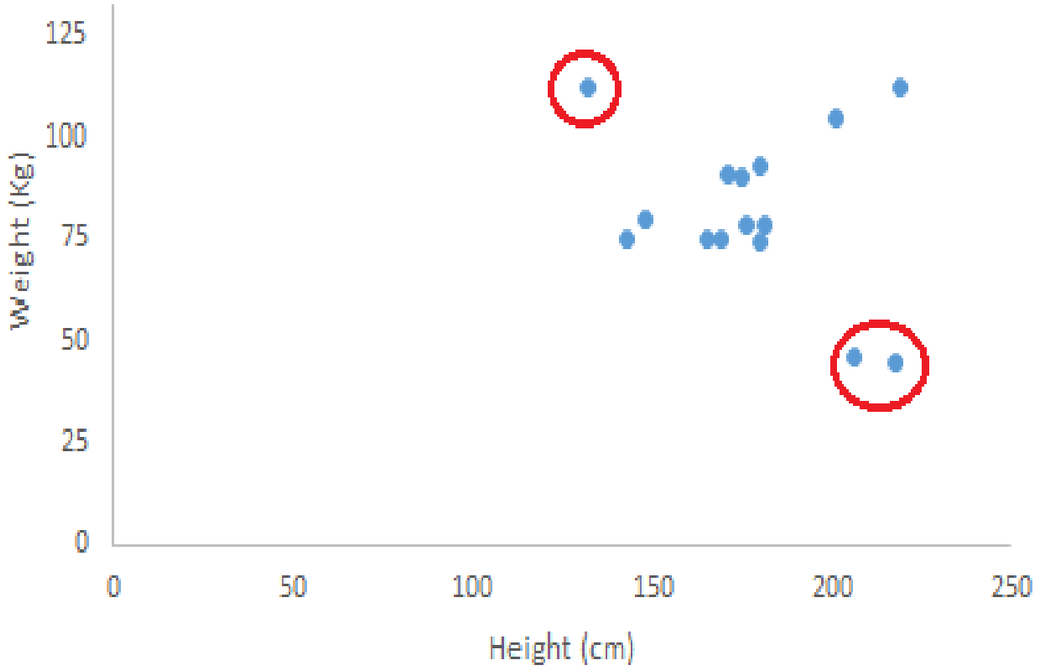
**Without outlier**
4,4,5,5,5,5,6,6,6,7,7
Mean = 5.45
Median = 5.00
Mode = 5.00
Standard Deviation = 1.04

**With outlier**
4,4,5,5,5,5,6,6,6,7,7,300
Mean = 30.00
Median = 5.5
Mode = 5.00
Standard Deviation = 85.03

# How to detect Outliers?

Most commonly used method to detect outliers is visualization.
We use various visualization methods, like Box-plot, Histogram, Scatter Plot
(Below, we have used box plot and scatter plot for visualization).



# How to detect Outliers?

Outlier detection is merely a special case of the examination of data for influential data points and it also depends on the business understanding. Some analysts also use various thumb rules to detect outliers.

## Some of them are:

- Any value, which is beyond the range of -1.5 □ IQR  to 1.5 x IQR
- Use capping methods. Any value which out of range of 5$^{th}$ and 95$^{th}$ percentile can be considered as outlier
- Data points, three or more standard deviation away from mean are considered outlier
- Bivariate and multivariate outliers are typically measured using either an index of influence or leverage, or distance. Popular indices such as Mahalanobis' distance and Cook's D are frequently used to detect outliers.

# How to remove Outliers?

- Deleting observations: We delete outlier values if it is due to data entry error, data processing error or outlier observations are very small in numbers. We can also use trimming at both ends to remove outliers.

- Transforming and binning values: Transforming variables can also eliminate outliers. Natural log of a value reduces the variation caused by extreme values. Binning is also a form of variable transformation. Decision Tree algorithm allows to deal with outliers well due to binning of variable. We can also use the process of assigning weights to different observations.

- Imputing: Like imputation of missing values, we can also impute outliers. We can use mean, median, mode imputation methods. Before imputing values, we should analyse if it is natural outlier or artificial. If it is artificial, we can go with imputing values. We can alos use statistical model to predict values of outlier observation and after that we can impute it with predicted values.

# Why missing values treatment is required?

Missing data in the training data set can reduce the power / fit of a model or can lead to a biased model because we have not analysed the behaviour and relationship with other variables correctly. It can lead to wrong prediction or classification.

| Name | Weight | Gender | Play Cricket/ Not |
|---|---|---|---|
| Mr. Amit | 58 | M | Y |
| Mr. Anil | 61 | M | Y |
| Miss Swati | 58 | F | N |
| Miss Richa | 55 | | Y |
| Mr. Steve | 55 | M | N |
| Miss Reena | 64 | F | Y |
| Miss Rashm | 57 | | Y |
| Mr. Kunal | 57 | M | N |

| Gender | #Students | #Play Cricket | %Play Cricket |
|---|---|---|---|
| F | 2 | 1 | 50% |
| M | 4 | 2 | 50% |
| Missing | 2 | 2 | 100% |

| Name | Weight | Gender | Play Cricket/ Not |
|---|---|---|---|
| Mr. Amit | 58 | M | Y |
| Mr. Anil | 61 | M | Y |
| Miss Swati | 58 | F | N |
| Miss Richa | 55 | F | Y |
| Mr. Steve | 55 | M | N |
| Miss Reena | 64 | F | Y |
| Miss Rashmi | 57 | F | Y |
| Mr. Kunal | 57 | M | N |

| Gender | #Students | #Play Cricket | %Play Cricket |
|---|---|---|---|
| F | 4 | 3 | 75% |
| M | 4 | 2 | 50% |

Notice the missing values in the image shown above: In the left scenario, we have not treated missing values. The inference from this data set is that the chances of playing cricket by males is higher than females.
On the other hand, if you look at the second table, which shows data after treatment of missing values (based on gender), we can see that females have higher chances of playing cricket compared to males.

# Why my data has missing values?

- Data Extraction
- Data collection
- Missing at random
- Missing completely at random

## Which are the methods to treat missing values ?

Deletion**:** It is of two types: List Wise Deletion and Pair Wise Deletion.

Mean/ Mode/ Median Imputation: Imputation is a method to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set to assist in estimating the missing values. Mean / Mode / Median imputation is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable. It can be of two types:-

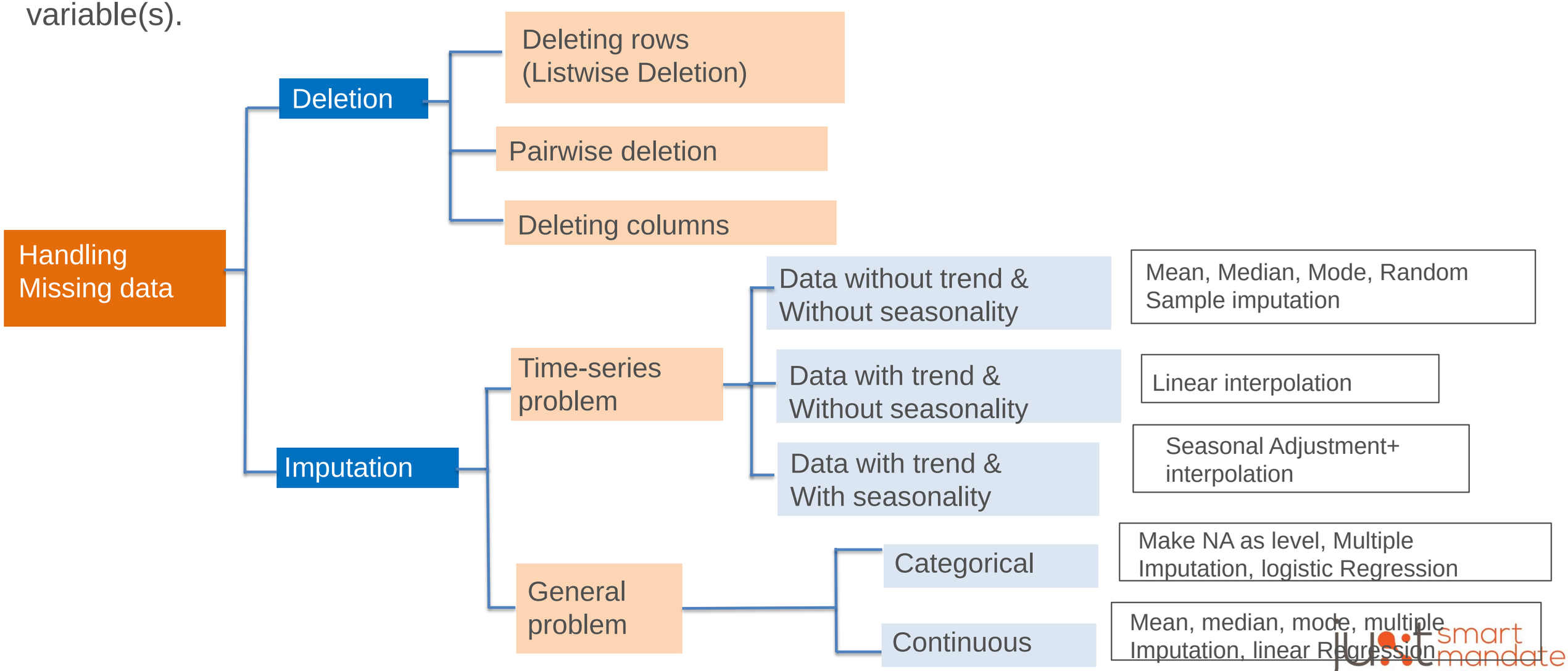
Prediction Model:
Prediction model is one of the sophisticated method for handling missing data. Here, we create a predictive model to estimate values that will substitute the missing data.

KNN Imputation: In this method of imputation, the missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing.

juxt smart mandate

# Feature/ Variable creation

Feature / Variable creation is a process to generate a new variables / features based on existing variable(s).



# Feature/ Variable Selection

- Variable subset selection
- Dimensionality reduction

## Top reasons to use feature selection are:

- It enables the machine learning algorithm to train faster.
- It reduces the complexity of a model and makes it easier to interpret.
- It improves the accuracy of a model if the right subset is chosen.
- It reduces overfitting

# Filter Methods (Univariate)

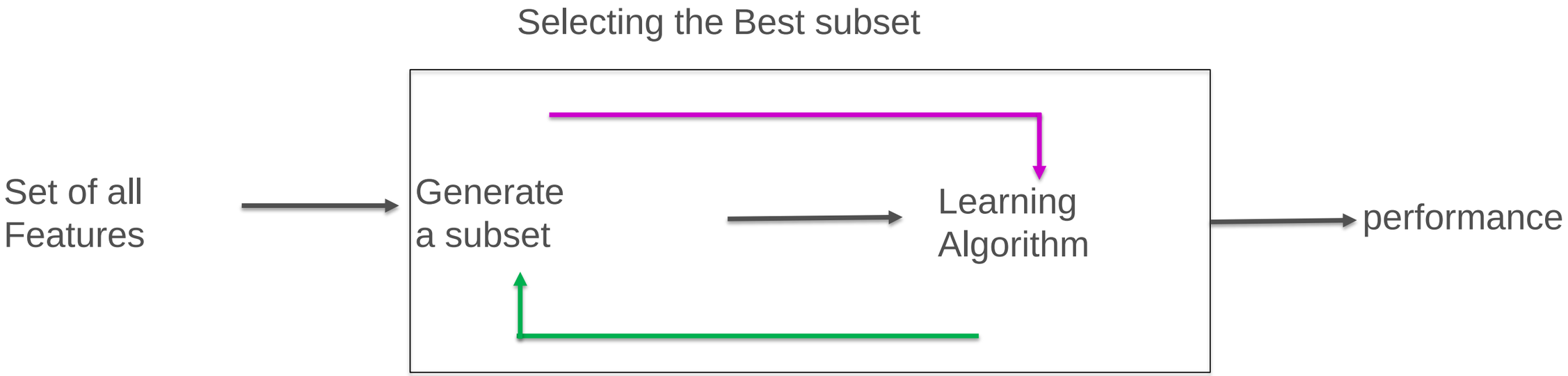
## 2. Filter Methods

Set of all Features → Selecting the Best subset → Learning Algorithm → Performance

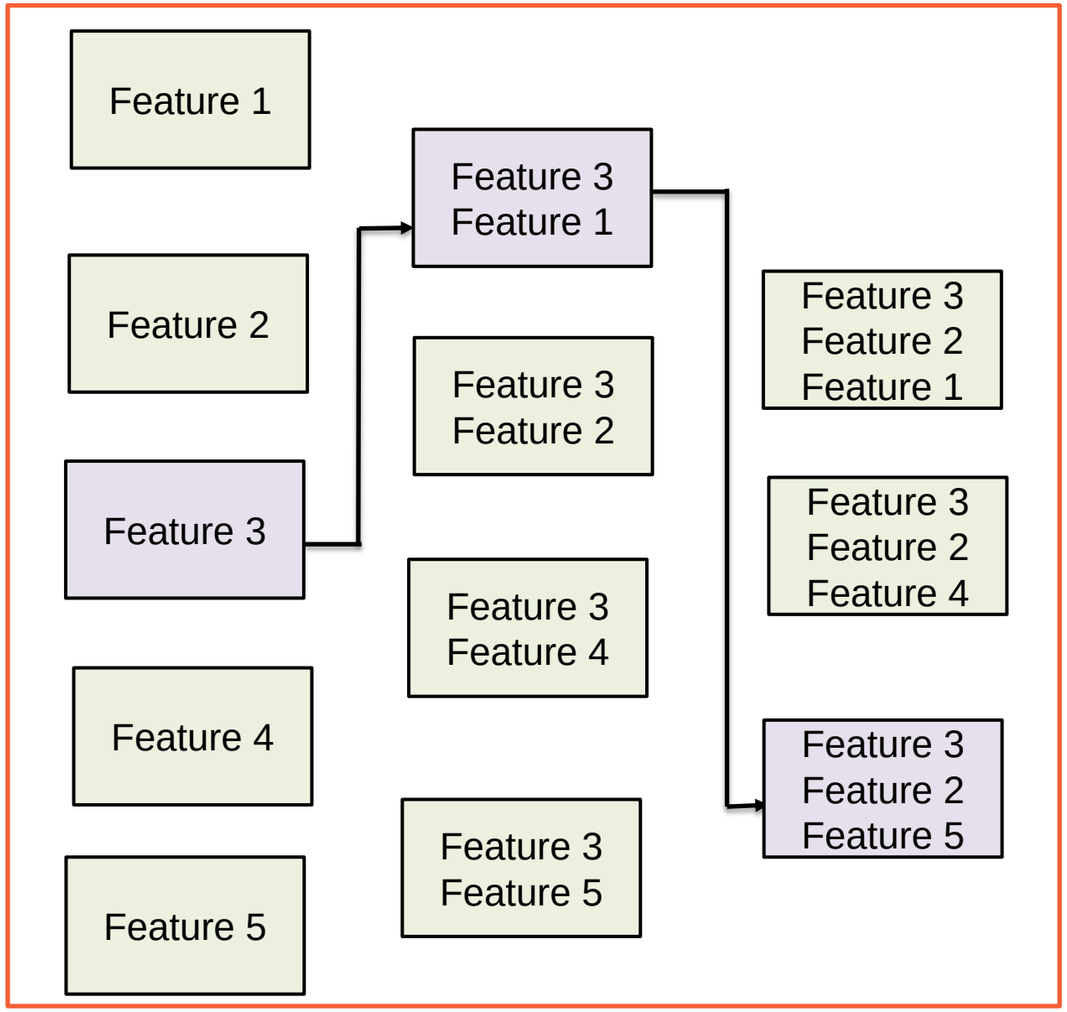
| Feature/Response | Continuous | Categorical |
|---|---|---|
| Continuous | Pearson's correlation | LDA |
| Categorical | Anova | Chi-square |

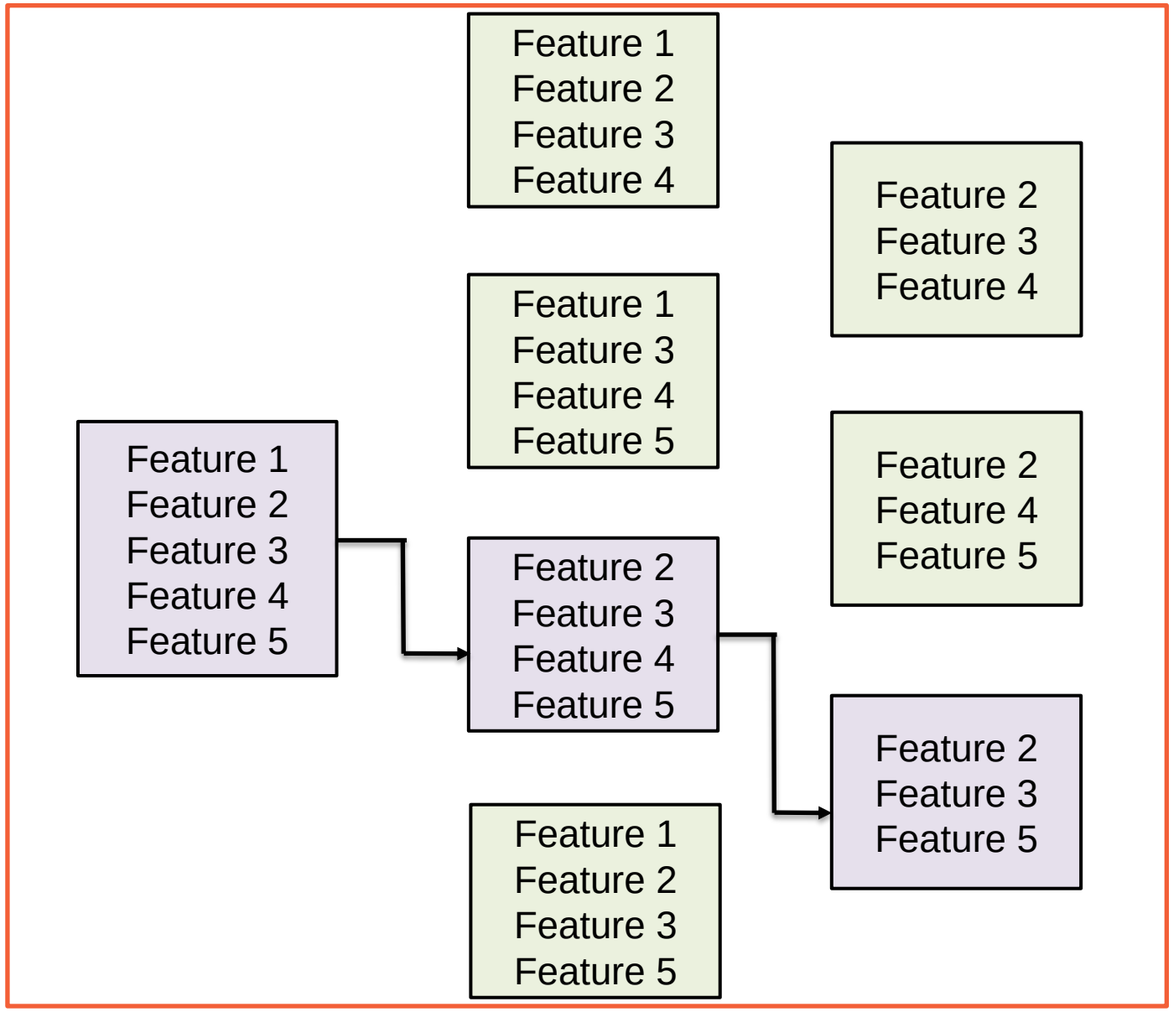
# Wrapper Methods (Selecting subset of variables)

## 3. Wrapper methods



# Forward Selection/Backward Elimination/Recursive Feature elimination (step wise regression)

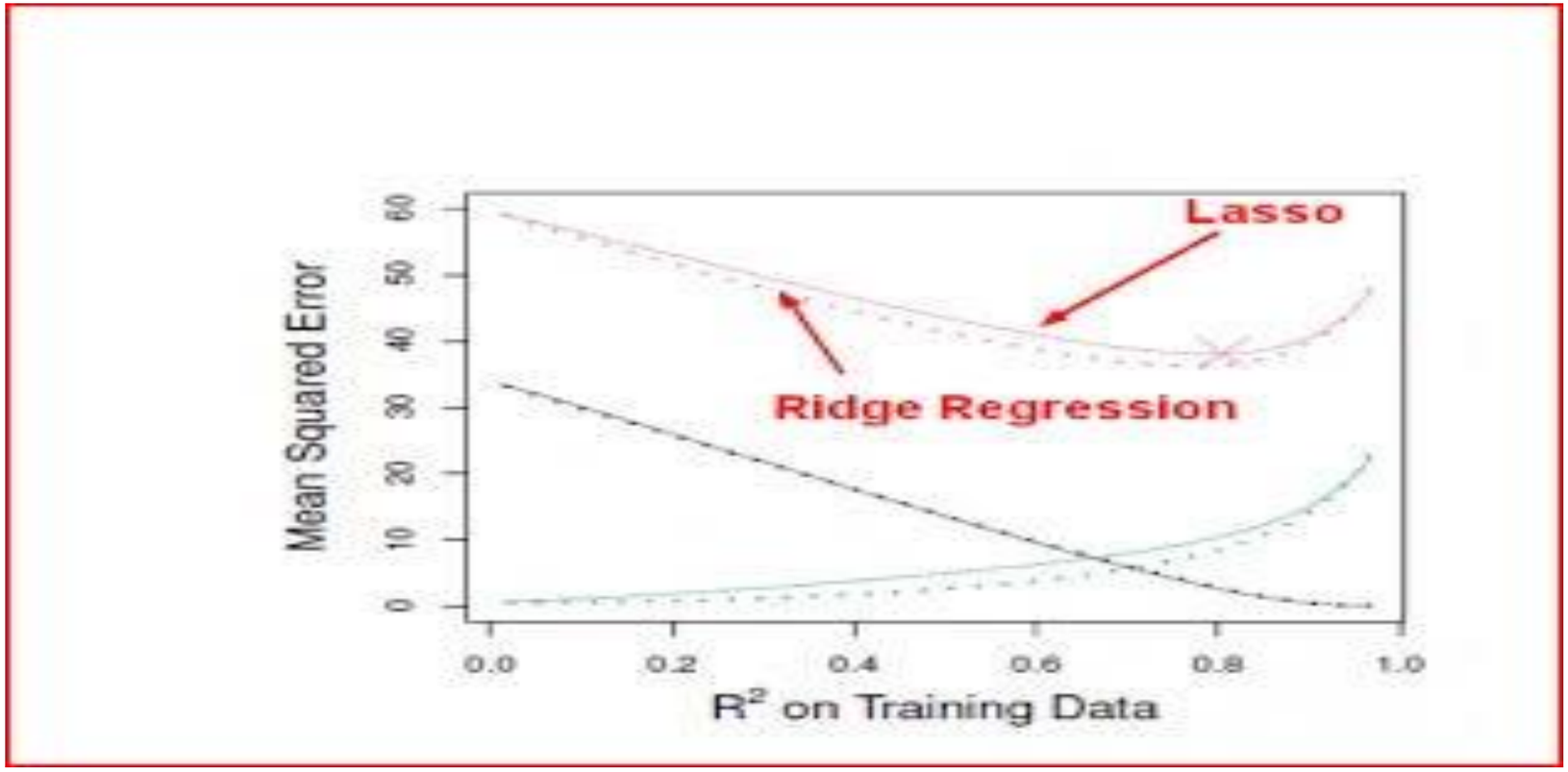


# EMBEDDED Methods

Embedded methods combine the qualities' of filter and wrapper methods. It's implemented by algorithms that have their own built-in feature selection methods.

Some of the most popular examples of these methods are LASSO and RIDGE regression which have inbuilt penalization functions to reduce over fitting.



# Ridge and Lasso regression

Ridge regression is similar to least squares except that the coefficients are estimated by minimizing a slightly different quantity.

Ridge regression, like OLS, seeks coefficient estimates that reduce RSS, however they also have a shrinkage penalty when the coefficients come closer to zero.

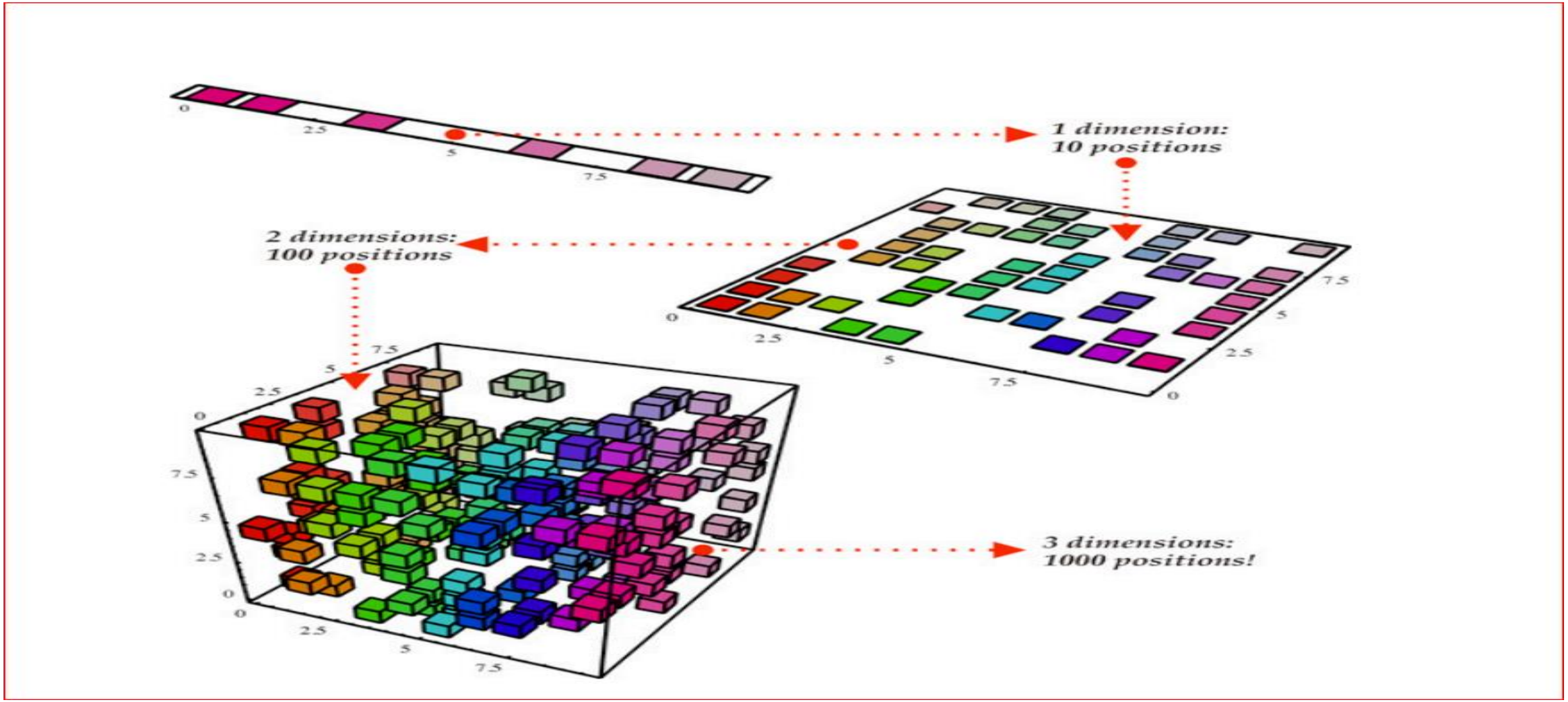
This penalty has the effect of shrinking the coefficient estimates towards zero.
Without going into the math, it is useful to know that ridge regression shrinks the features with the smallest column space variance.

Like in principal component analysis, ridge regression projects the data into directional space and then shrinks the coefficients of the low-variance components more than the high variance components, which are equivalent to the largest and smallest principal components.

Lasso regression performs L1 regularization which adds penalty equivalent to absolute value of the magnitude of coefficients.

Ridge regression performs L2 regularization which adds penalty equivalent to square of the magnitude of coefficients.

# Dimensionality Reduction



# Principal Components Regression

One can describe Principal Components Regression as an approach for deriving a low-dimensional set of features from a large set of variables.

The first principal component direction of the data is along which the observations vary the most. In other words, the first PC is a line that fits as close as possible to the data. One can fit p distinct principal components.

The second PC is a linear combination of the variables that is uncorrelated with the first PC, and has the largest variance subject to this constraint.

The idea is that the principal components capture the most variance in the data using linear combinations of the data in subsequently orthogonal directions. In this way, we can also combine the effects of correlated variables to get more information out of the available data, whereas in regular least squares we would have to discard one of the correlated variables.

# Machine Learning Algorithms

# Broadly, there are 3 types of Machine Learning Algorithms.

1. Supervised Learning

This algorithm consist of a target / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these set of variables, we generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data. Examples of Supervised Learning: Regression, Decision Tree, Random Forest, KNN, Logistic Regression etc.

2. Unsupervised Learning

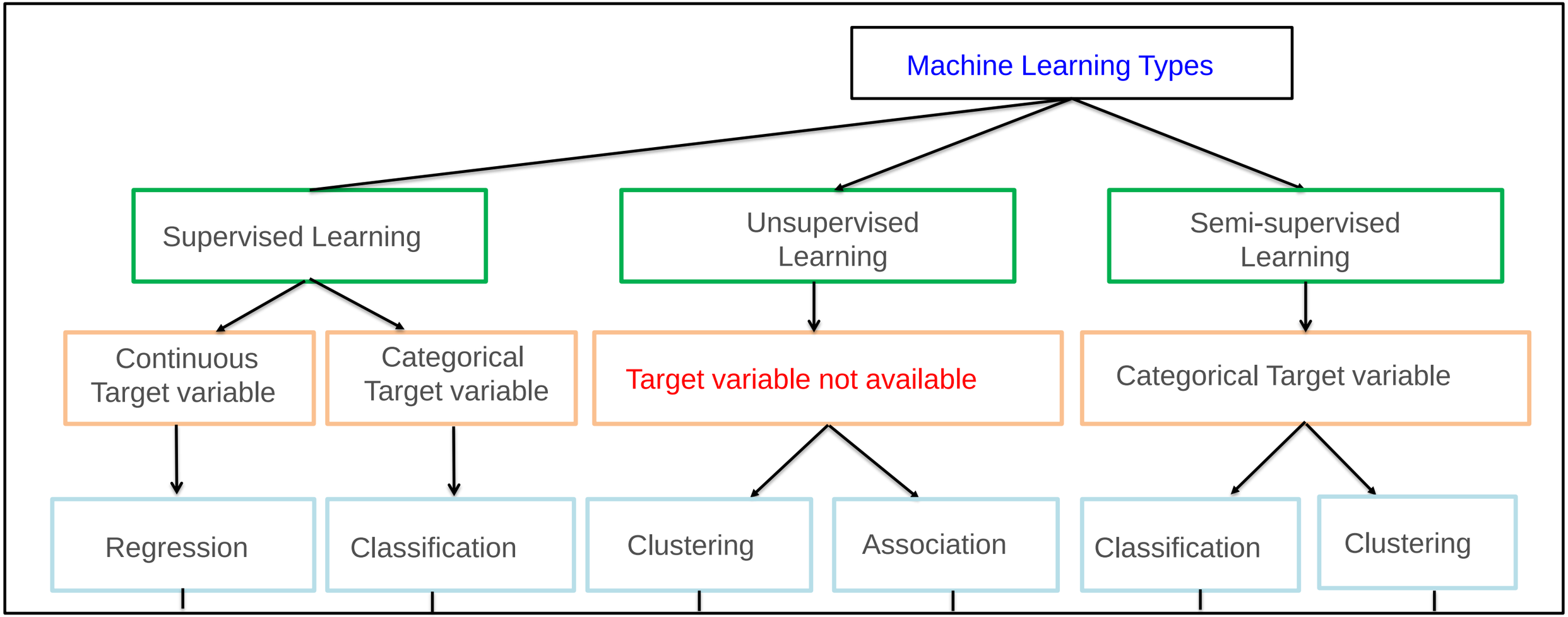
In this algorithm, we do not have any target or outcome variable to predict / estimate. It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention. Examples of Unsupervised Learning: Apriori algorithm, K-means.

3. Reinforcement Learning:

Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions. Example of Reinforcement Learning: Markov Decision Process.

juxt smart mandate

# Machine Learning Algorithms…




juxt smart mandate

# Machine Learning Algorithms…

| | Response Variable Continuous | Response Variable categorical | No Response Variable |
|---|---|---|---|
| Predictor Variable- continuous | Linear regression<br>Neural network K-nearest Neighbor (KNN) | Logistic regression<br>KNN<br>Neural network | Cluster analysis<br>Principal Component Analysis |
| Predictor Variable- categorical | Linear regression<br>Neural network | Decision/ classification Trees<br>logistic regression<br>Naïve bayes | Association rule |

# Linear Regression

# Linear Regression

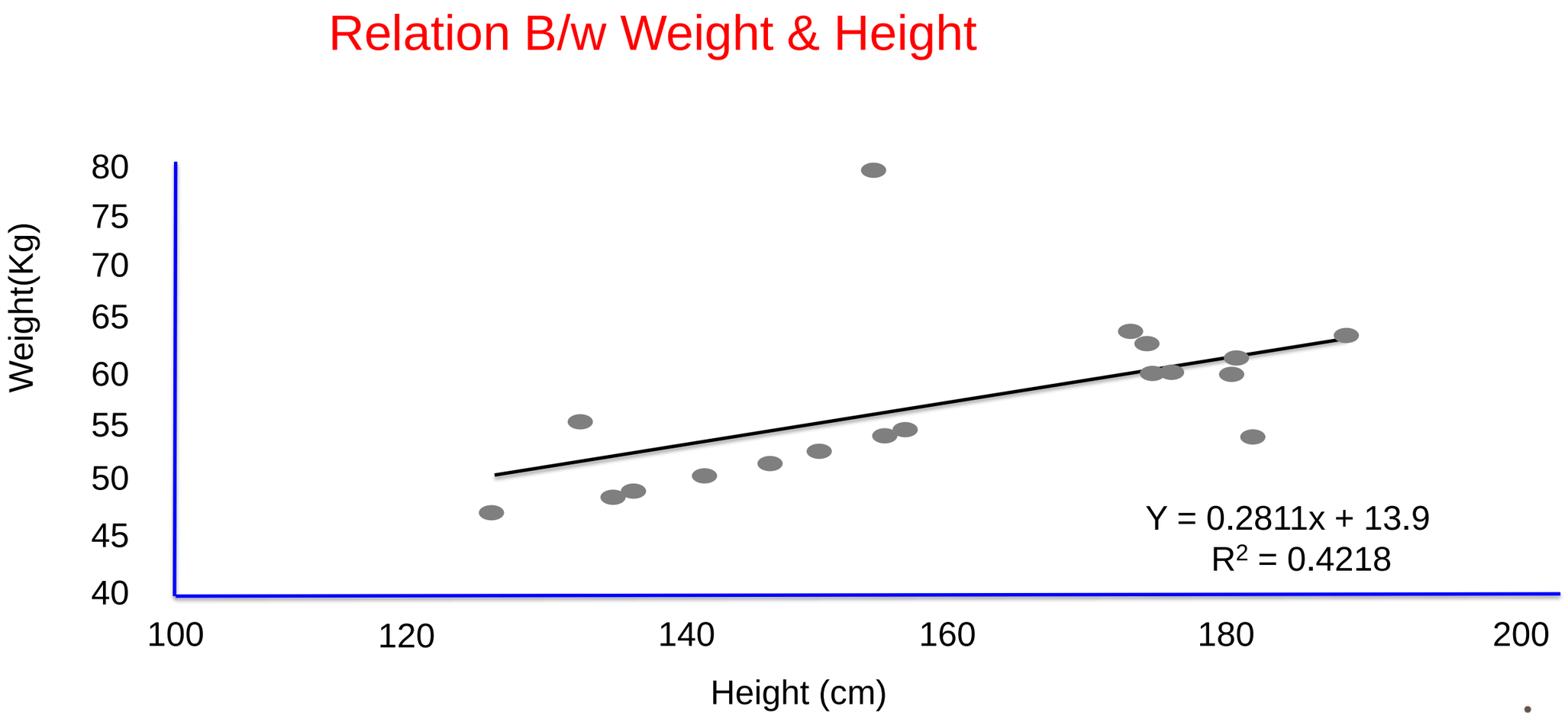
Linear Regression is used when we want to predict an outcome variable that is interval / continuous with a set of predictors that are also interval / continuous. While categorical / nominal data can also be included, The representation of linear regression is an equation that describes a line that best fits the relationship between the input variables (x) and the output variables (y), byfinding specific weightings for the input variables called coefficients (B)

For example: $y = B0 + B1 * x$ → Simple Regression

($Y = B1*X1 + B2*X2 + B3*X3 + \ldots + A$ -→ Multiple Regression

These coefficients B0 and B1 are derived based on minimizing the sum of Squared difference of distance between data points and regression line.

# Linear Regression



# Linear Regression

Linear Regression is mainly of two types:

- ## Simple Linear Regression

Simple Linear Regression is characterized by one independent variable.

- ## Multiple Linear Regression

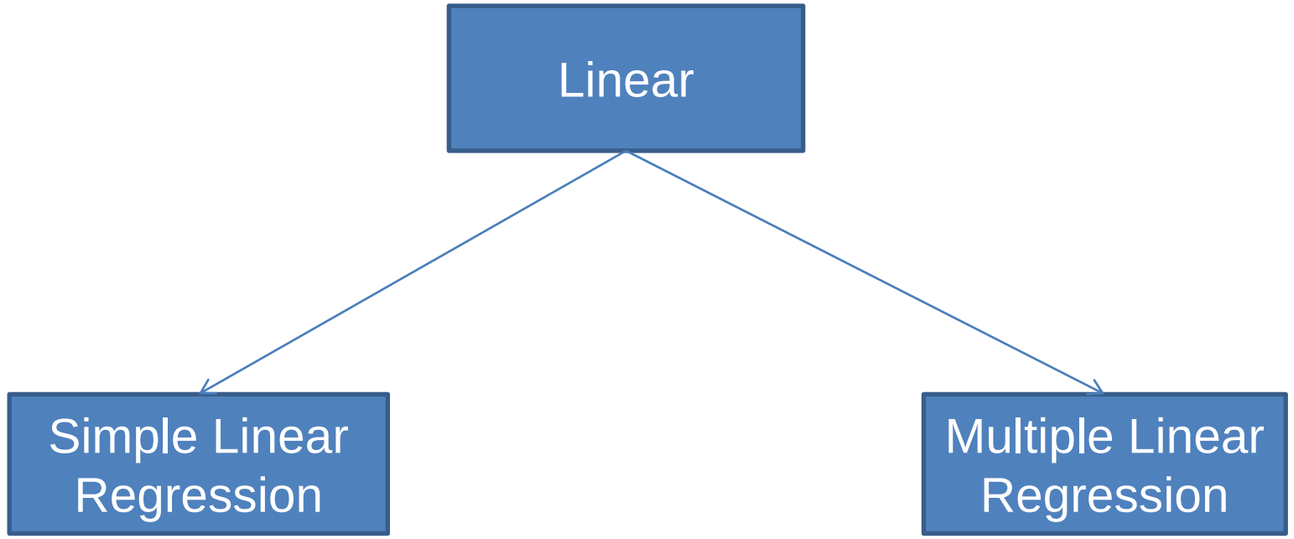
Multiple Linear Regression(as the name suggests) is characterized by Multiple(more than1) independent variables.
Some good rules of thumb when using this technique are to remove variables that are very similar (correlated) and to remove noise from your data, if possible. It is a fast and simple technique and good first algorithm to try.

# Regression

- In statistics, regression analysis is a statistical process for estimating the **relationship among variables.**
- The focus is on the relationship between a dependent variable and **one or more independent variables.**
- Linear Regression is a linear model, e.g. a model that assumes a linear relationship between the input variables(x) and single the output variable(y). More specifically, that y can be calculated from a **linear combination of the input variables(x).**
- Different techniques can be used to prepare or train the linear regression equation from data, the most common prepared this way as **ordinary Least Squares Linear Regression or just Least Squares Regression.**

# Regression Contd..

- When there is a **single input variable(x),** the method is referred to as **simple linear regression**. when there are **multiple input variables**, the method is referred as **multiple linear regression**.

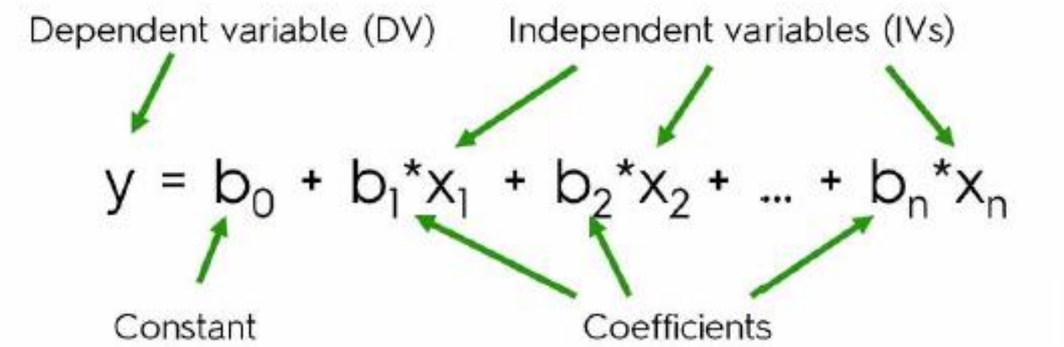


# Regression Contd..

Simple Linear Regression

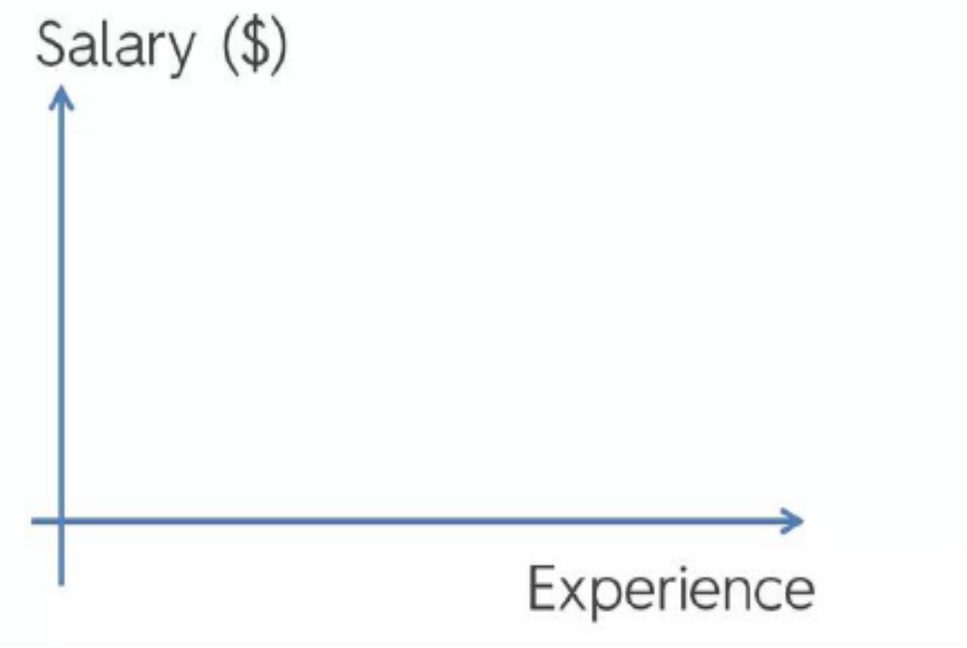
$$y = b_0 + b_1 * x_1$$

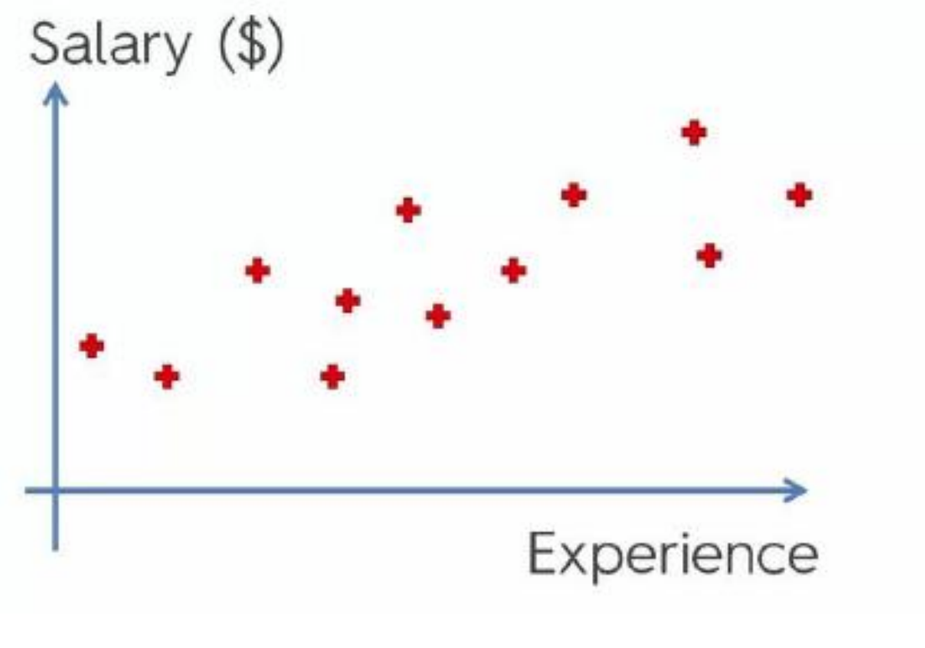
Multiple Linear Regression



# Simple Regression

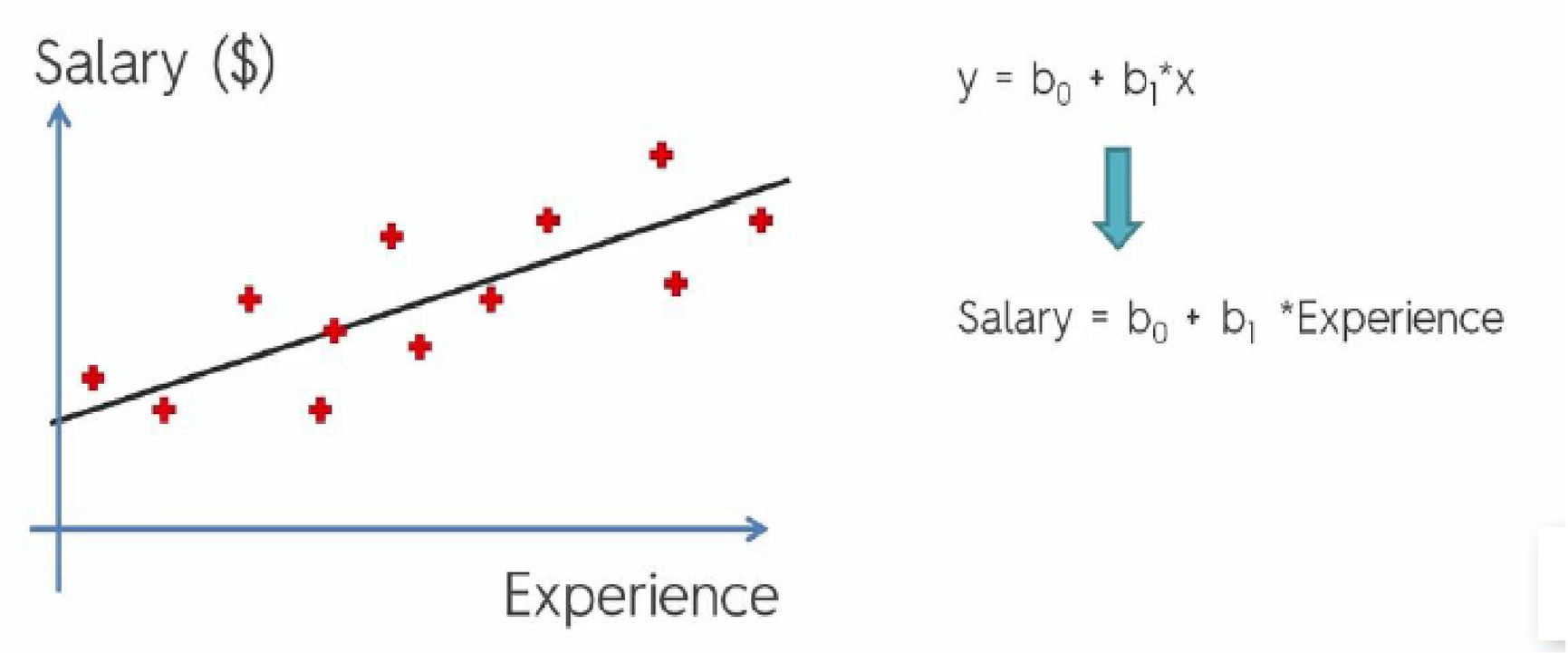
## Simple Linear Regression:



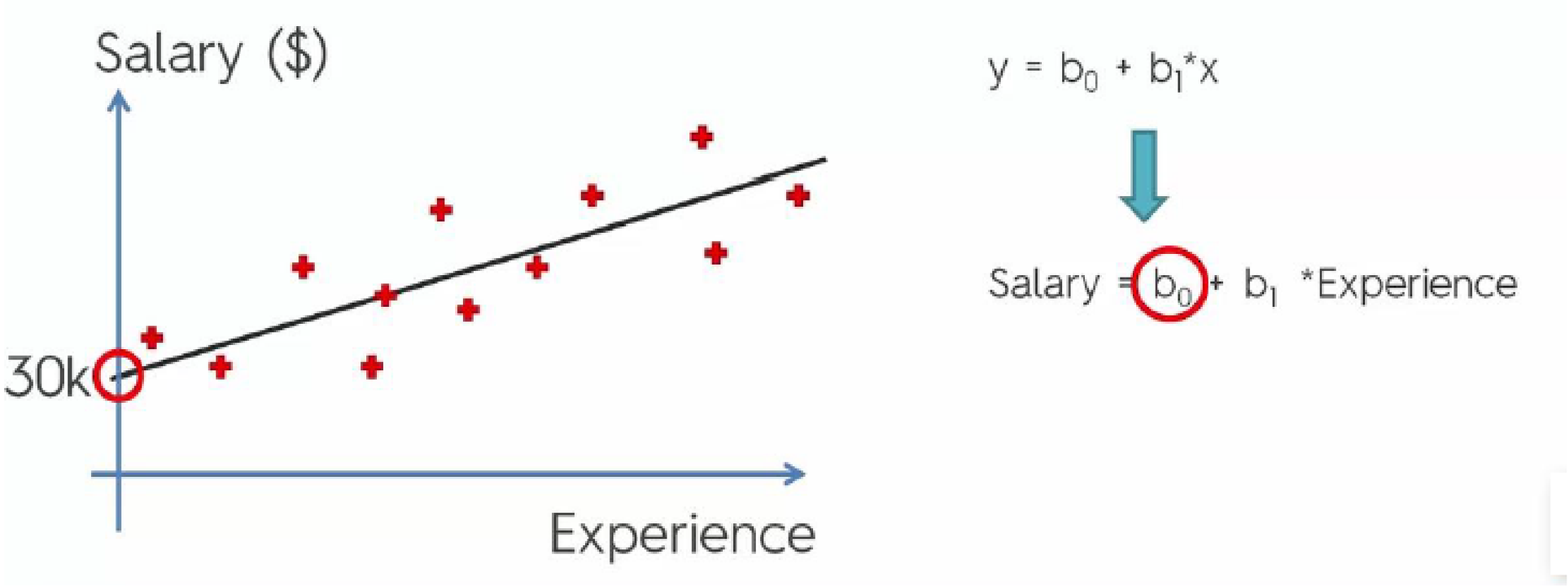
# Simple Regression Contd..

## Simple Linear Regression:



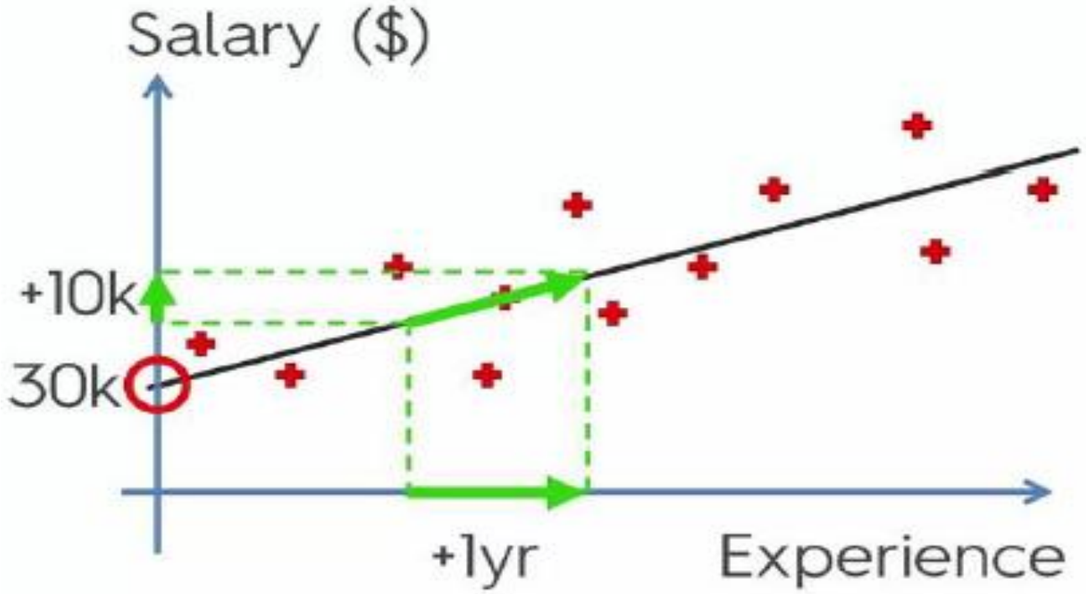
# Simple Regression Contd..

## Simple Linear Regression:



# Simple Regression Contd..

## Simple Linear Regression:

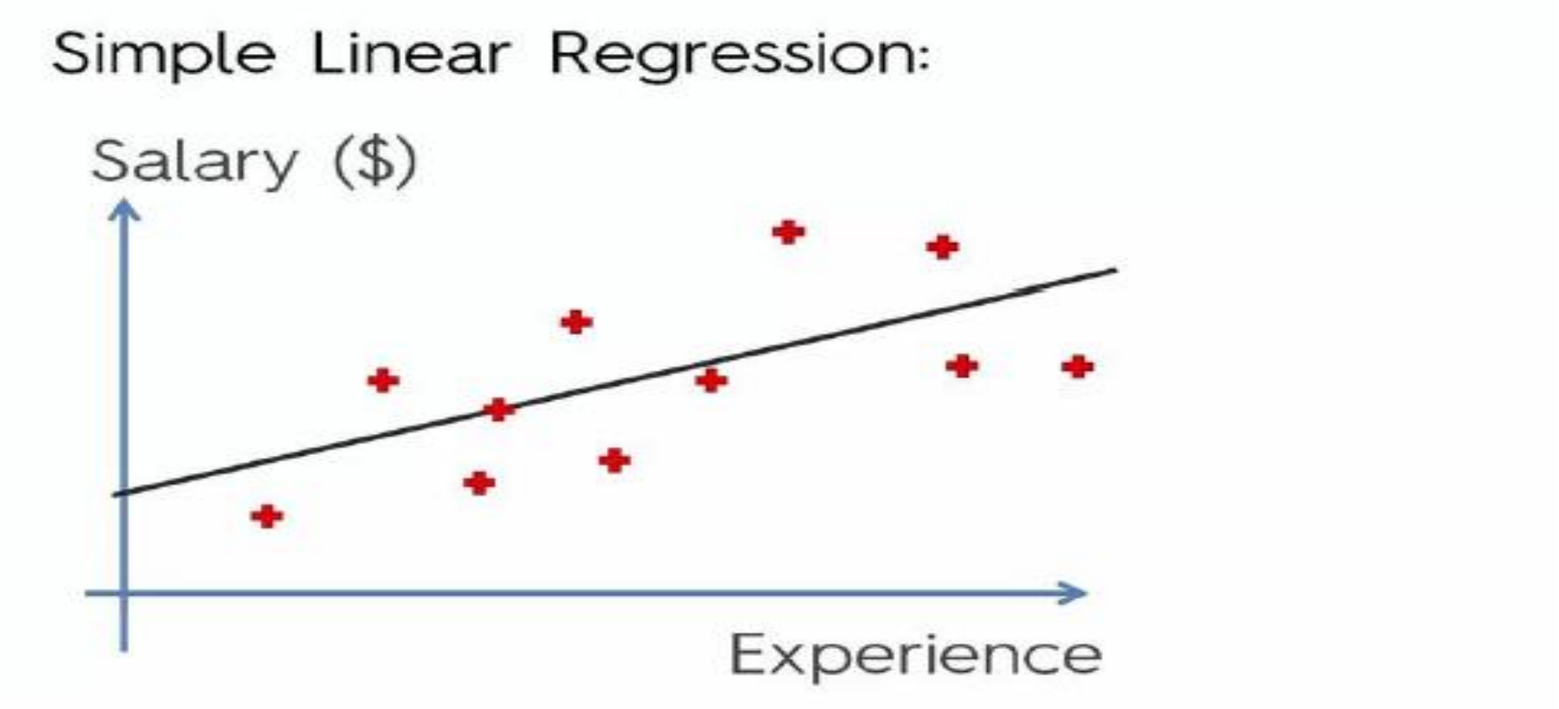



$$y = b_0 + b_1 * x$$

Salary = $b_0$ + $b_1$ *Experience

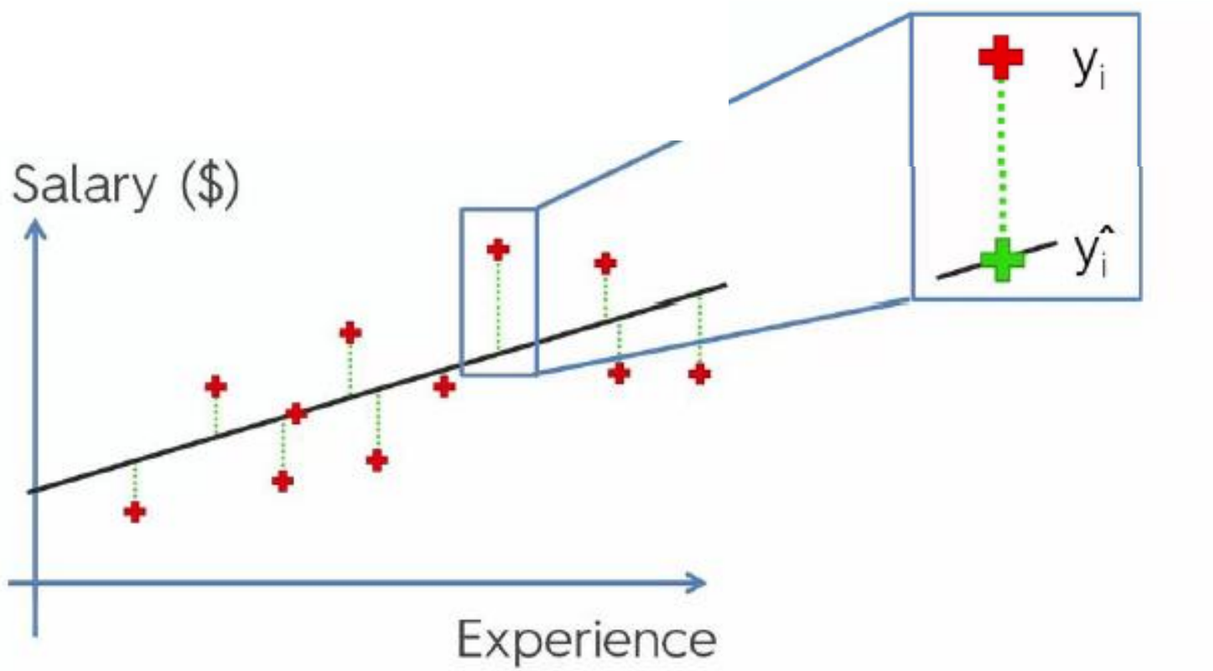
# Ordinary Least Squares

## Simple Linear Regression:



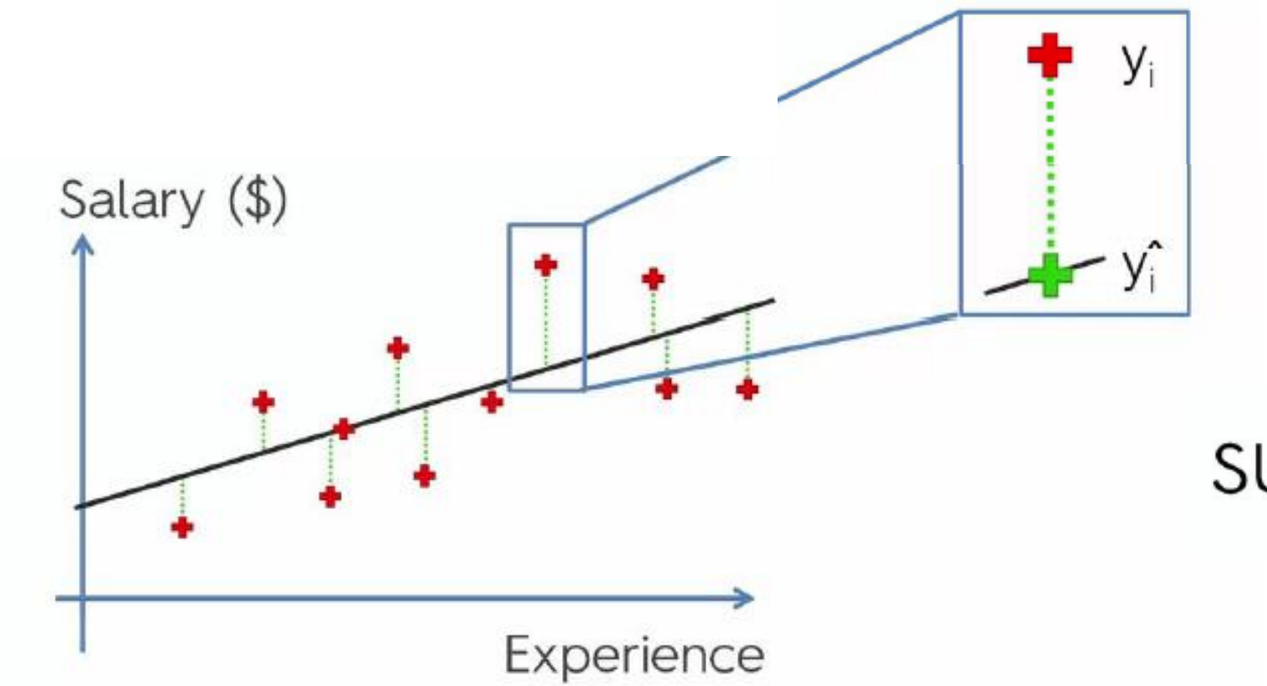
# Ordinary Least Squares

## Simple Linear Regression:



# R Squared

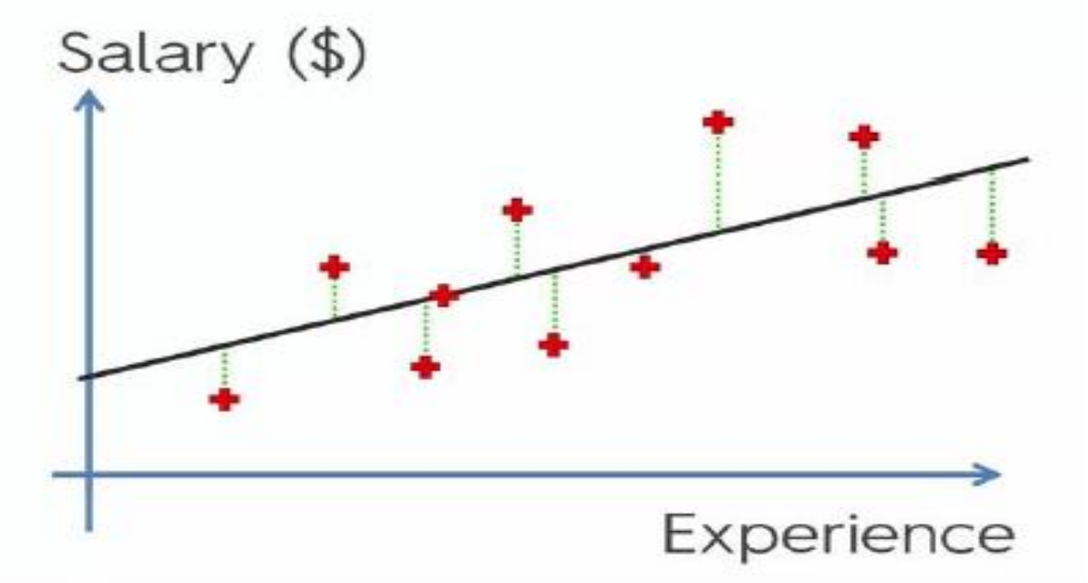
## Simple Linear Regression:




$$\text{SUM}\ (y_i - \hat{y_i})^2 \rightarrow \min$$

# R Squared

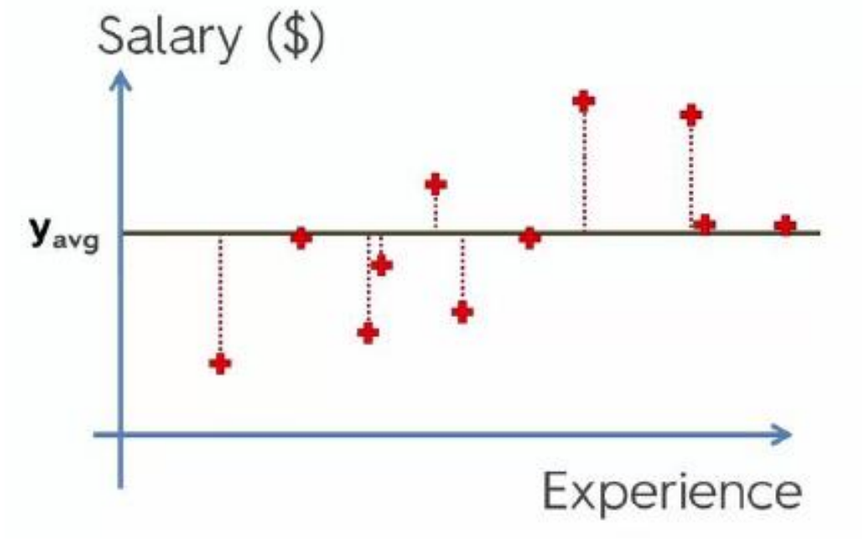
## Simple Linear Regression:




$$SS_{res} = SUM\ (y_i - \hat{y_i})^2$$

# R Squared

## Simple Linear Regression:




$$SS_{res} = SUM\ (y_i - \hat{y_i})^2$$

$$SS_{tot} = SUM\ (y_i - y_{avg})^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

# Adjusted R2

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$R^2$ – Goodness of fit
(greater is better)

$y = b_0 + b_1 * x_1$

$y = b_0 + b_1 * x_1 + b_2 * x_2$

$+ b_3 * x_3$

**Problem**:

$SS_{res}$ -> Min

$R^2$ will never decrease

# Adjusted R2

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$$\text{Adj } R^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1}$$

p - number of regressors

n – sample size

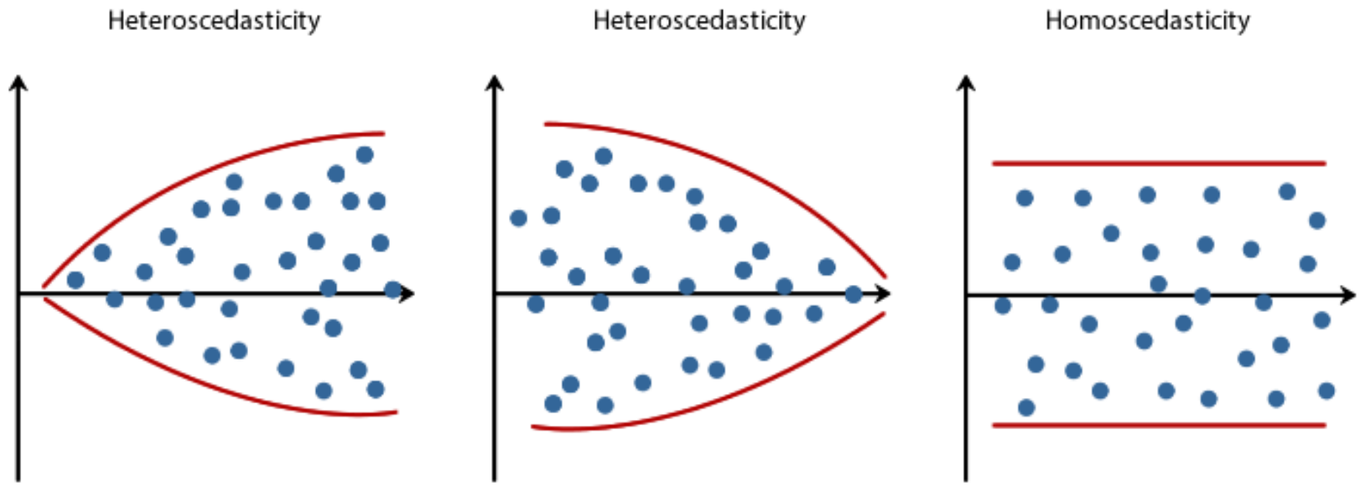
# Assumptions of Regression

Important assumptions in regression Analysis:

- There should be a **linear and additive relationship between dependent (response) variable and independent (predictor) variable(s).** A linear relationship suggests that a change in response Y due to one unit change in $X^1$ is constant, regardless of the value of $X^1$. An additive relationship suggests that the effect of $X^1$ on Y is independent of other variables.
- There should be **no correlation between the residual (error) terms**. Absence of this phenomenon is known as Autocorrelation.
- The **independent variables should not be correlated**. Absence of this phenomenon is known as multicollinearity.
- The **error terms must have constant variance**. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to heteroskedasticity.
- The error terms must be **normally distributed**.

# Assumptions of Regression

- We should have **independence of observations**
- Data needs to show **homoscedasticity**, which is where the variances along the line of best fit remain similar as you move along the line. take a look at the three scatterplots below, which provide three simple examples: two of data that fail the assumption (called heteroscedasticity) and one of data that meets this assumption (called homoscedasticity):




- Finally, we need to check that the **residuals (errors)** of the regression line are **approximately normally distributed**

# Heteroscedasticity

**Heteroscedasticity** means unequal scatter. In regression analysis, we talk about heteroscedasticity in the context of the residuals or error term. Specifically, heteroscedasticity is a systematic change in the spread of the residuals over the range of measured values. Heteroscedasticity is a problem because ordinary least squares(OLS) regression assumes that all residuals are drawn from a population that has a constant variance (homoscedasticity).

**Example:**

If you model household consumption based on income, you'll find that the variability in consumption increases as income increases. Lower income households are less variable in absolute terms because they need to focus on necessities and there is less room for different spending habits. Higher income households can purchase a wide variety of luxury items, or not, which results in a broader spread of spending habits.

# Autocorrelation

**Autocorrelation** is a characteristic of data in which the correlation between the values of the same variables is based on related objects.

- Autocorrelation represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals.
- Autocorrelation measures the relationship between a variable's current value and its past values.
- An autocorrelation of +1 represents a perfect positive correlation, while an autocorrelation of negative 1 represents a perfect negative correlation.
- Technical analysts can use autocorrelation to see how much of an impact past prices for a security have on its future price.

**Example**:

Emma is looking to determine if a stock's returns in her portfolio exhibit autocorrelation; the stock's returns relate to its returns in previous trading sessions. If the returns do exhibit autocorrelation, Emma could characterize it as a momentum stock because past returns seem to influence future returns. Emma runs a regression with two prior trading sessions returns as the independent variables and the current return as the dependent variable.

# Multicollinearity

Multicollinearity occurs when independent variables in a regression  model are correlated. This correlation is a problem because independent variables should be *independent*. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.

Why is Multicollinearity a Potential Problem?

when independent variables are correlated, it indicates that changes in one variable are associated with shifts in another variable. The stronger the correlation, the more difficult it is to change one variable without changing another. It becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable *independently* because the independent variables tend to change in unison.

# An Example

A salesperson for a large car brand wants to determine whether there is a relationship between an individual's income and the price they pay for a car. As such, the individual's "income" is the independent variable and the "price" they pay for a car is the dependent variable. The salesperson wants to use this information to determine which cars to offer potential customers in new areas where average income is known.

# Output interpretation – Model summary

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .873[a] | .762 | .749 | 874.779 |

a. Predictors: (Constant), Income

- The *R* value represents the simple correlation and is 0.873, which indicates a high degree of correlation.

- The $R^2$ value (the "**R Square**" column) indicates how much of the total variation in the dependent variable, Price, can be explained by the independent variable, Income. In this case, 76.2% can be explained, which is very large.

# Output interpretation – Regression model statistically significantly predicts the outcome variable

**ANOVA**[a]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 44182633.37 | 1 | 44182633.37 | 57.737 | .000[b] |
| | Residual | 13774291.07 | 18 | 765238.393 | | |
| | Total | 57956924.44 | 19 | | | |

a. Dependent Variable: Price

b. Predictors: (Constant), Income

- **ANOVA** table, which reports how well the regression equation fits the data (i.e., predicts the dependent variable)

- Here, *p* = 0.000, which is less than 0.05, and indicates that - overall, the regression model statistically significantly predicts the outcome variable (i.e., it is a good fit for the data).

# Output interpretation – Coefficient table

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 8286.786 | 1852.256 | | 4.474 | .000 |
| | Income | .564 | .074 | .873 | 7.598 | .000 |

a. Dependent Variable: Price

- **Coefficients** table provides us with the necessary information to predict price from income, as well as determine whether income contributes statistically significantly to the model (by looking at the "**Sig.**" column).

- Here the regression equation is:
  - Price = 8287 + 0.564(Income)

- Hands On

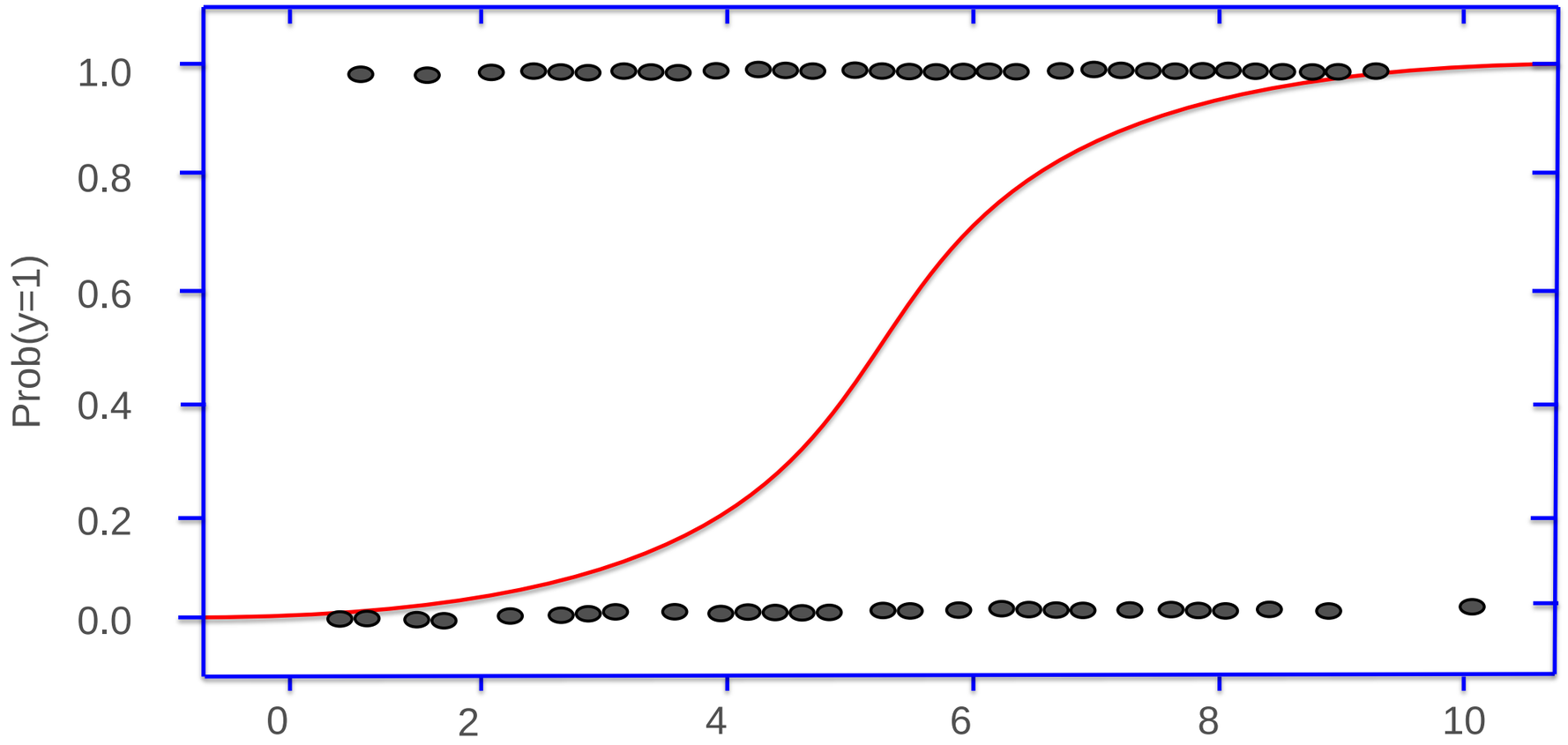
# Logistic Regression

Logistic regression is a classification algorithm and it is used to estimate discrete values ( Binary values like 0/1, yes/no, true/false) based on given set of independent variable(s).

In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function. Hence, it is also known as logit regression.

Since, it predicts the probability, its output values lies between 0 and 1 (as expected).

# Logistic Regression



# Logistic Regression

Logistic regression does not directly model Y (dependent variable).
Logistic regression transforms the dependent into a logit variable (natural log of the odds of Y occurring or not occurring, which is ln(p/1-p)) and uses maximum likelihood estimation (MLE) to estimate the coefficients.

$$\text{prob}(\text{event}) = \frac{\exp^{(B_1 * X_1 + B_2 * X_2 + A)}}{\left(1 + \exp^{(B_1 * X_1 + B_2 * X_2 + A)}\right)}$$

Above, p is the probability of presence of the characteristic of interest. It chooses parameters that maximize the likelihood of observing the sample values (maximum likelihood estimation) rather than that minimizes the sum of squared errors (like in ordinary regression).

Like linear regression, logistic regression does work better when you remove attributes that are unrelated to the output variable as well as attributes that are very similar (correlated) to each other. It's a fast model to learn and effective on binary classification problems.

# What is Logistic Regression

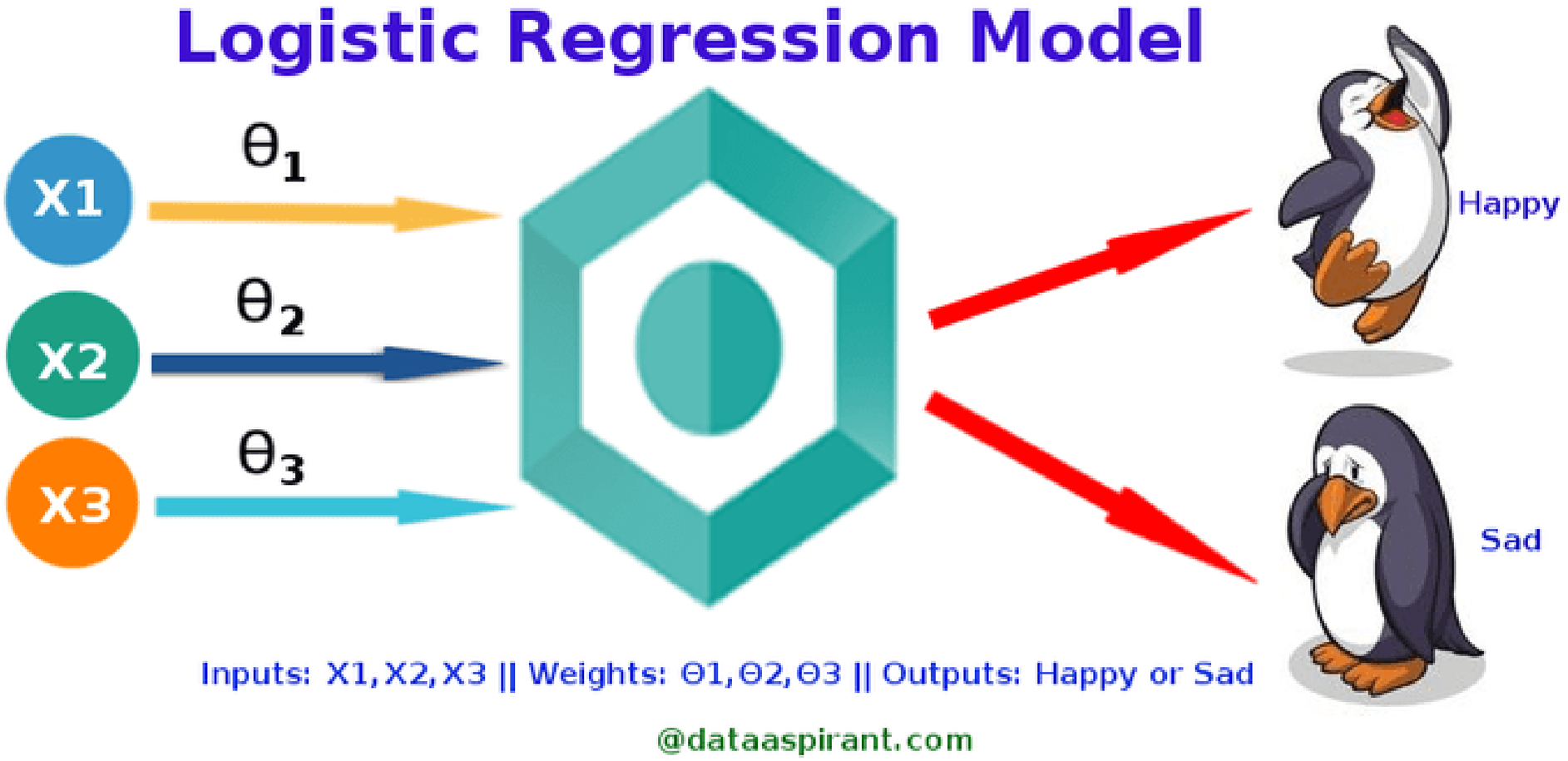
- Logistic Regression is a classification algorithm

- It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.

**Examples:**

- A patient has heart attack or not (Yes/No) based on Age, gender etc.,
- Predict whether an American voter will vote Democratic or Republican, based on age, income, sex, race, state of residence, votes in previous elections, etc.
- How likely a customer will buy iPod having iPhone in his/her pocket.
- How likely India cricket team will win when Virat kohli is in rest.
- What is the probability to get into best university by scoring decent marks in mathematics, physics?
- What is the probability to get a kiss from your girlfriend when you gifted her favorite dress on behalf of your birthday?

# Logistic Regression Model



# Logistic Regression Model

The "logit" model solves these problems:

$$\ln[p/(1-p)] = \alpha + \beta X + e$$

- p is the probability that the event Y occurs, p(Y=1)
- p/(1-p) is the "odds ratio"
- ln[p/(1-p)] is the log odds ratio, or "logit"

# Logistic Regression Model

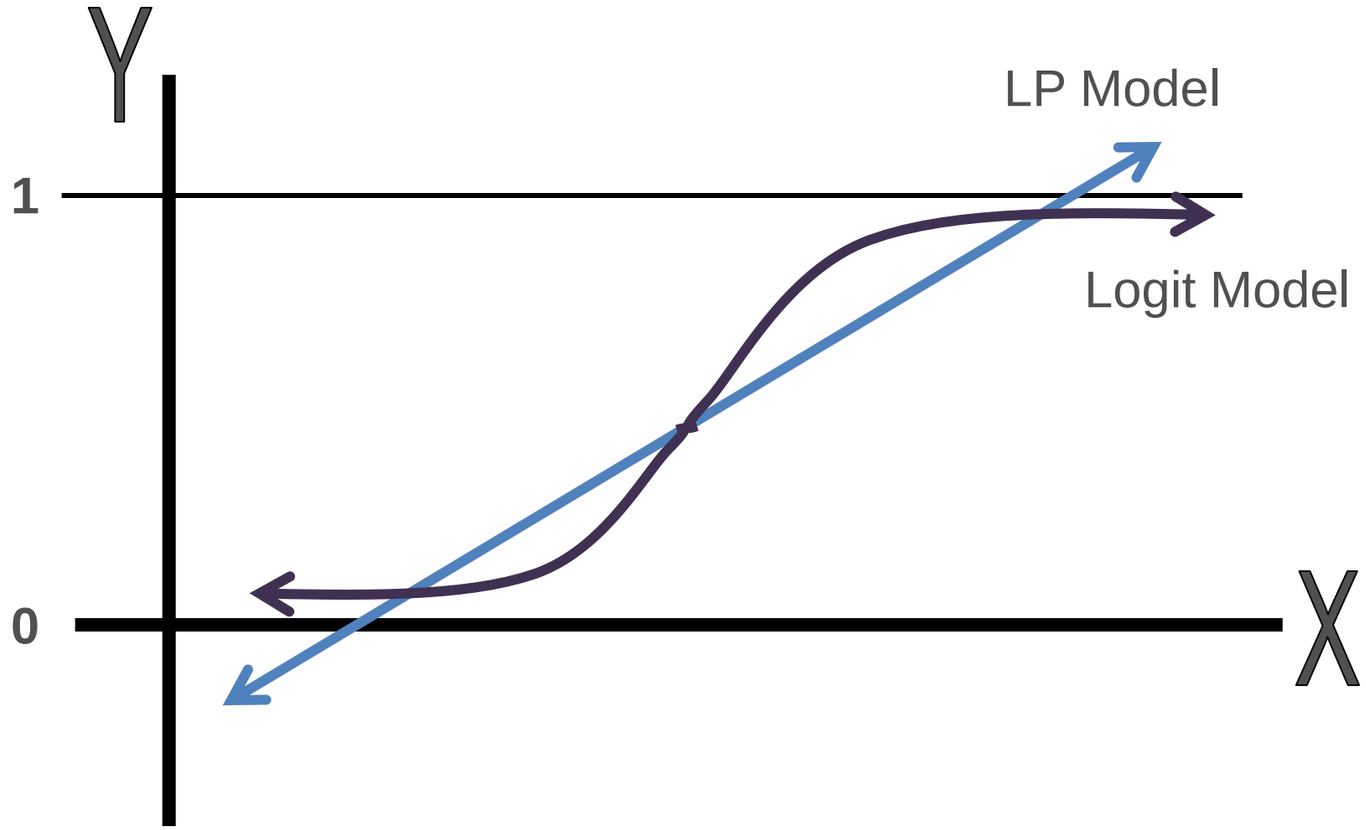
**More:**

- The logistic distribution constrains the estimated probabilities to lie between 0 and 1.
- The estimated probability is:

$$p = 1/[1 + \exp(-\alpha - \beta X)]$$

- if you let $\alpha + \beta X = 0$, then p = .50
- as $\alpha + \beta X$ gets really big, p approaches 1
- as $\alpha + \beta X$ gets really small, p approaches 0

# Comparing the LP and logistic Regression



# How to estimate β coefficients – using Maximum Likelihood Estimation (MLE)

- MLE is a statistical method for estimating the coefficients of a model.
- The likelihood function (L) measures the probability of observing the particular set of dependent variable values ($p_1$, $p_2$, ..., $p_n$) that occur in the sample:

$$L = \text{Prob}\,(p_1 * p_2 * * * p_n)$$

- The higher the L, the higher the probability of observing the ps in the sample.

# Maximum Likelihood Estimation (MLE)

**More:**

- MLE involves finding the coefficients ($\alpha$, $\beta$) that makes the log of the likelihood function (LL < 0) as large as possible
- Or, finds the coefficients that make -2 times the log of the likelihood function (-2LL) as small as possible
- The maximum likelihood estimates solve the following condition:

$$\{Y - p(Y=1)\}X_i = 0$$

summed over all observations, i = 1,…,n

# An Example

- A group of 20 students spend between 0 and 6 hours studying for an exam. How does the number of hours spent studying affect the probability that the student will pass the exam?

The reason for using logistic regression for this problem is that the values of the dependent variable, pass and fail, while represented by "1" and "0“

If the problem was changed so that pass/fail was replaced with the grade 0–100 (cardinal numbers), then simple regression analysis could be used.

| **Hours** | 0.5 | 0.75 | 1 | 1.25 | 1.5 | 1.75 | 1.75 | 2 | 2.25 | 2.5 | 2.75 | 3 | 3.25 | 3.5 | 4 | 4.25 | 4.5 | 4.75 | 5 | 5.5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Pass** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |

# Evaluating Classification Models

## Confusion Matrix

Confusion matrix is one of the most popular ways to evaluate a classification model. Although the matrix by itself is not a metric, the matrix representation can be used to define a variety of metrics, all of which become important in some specific case or scenario.

A confusion matrix can be created for a binary classification as well as a multi-class classification model.
A confusion matrix is created by comparing the predicted class label of a data point with its actual class label. This comparison is repeated for the whole dataset and the results. of this comparison are compiled in a matrix or tabular format

Accuracy: This is a ratio indicating the number of predictions that were correct.
Precision: The ratio of positive cases that were correctly identified.
Recall (sensitivity):
The ratio of actual positive cases that are identified correctly Specificity: The ratio of actual negative cases that are identified correctly

# Evaluating Classification Models – Confusion Matrix

| | | Predicted classed: Positive ($C_0$) | Predicted classed: Negative ($C_1$) | |
|---|---|---|---|---|
| Actual class | Positive ($C_0$) | a = number of correctly Classified c0 cases | c = number of $c_0$ cases Incorrectly classified as $c_1$ | Precision = a/(a + c) |
| | Negative ($C_0$) | b = number of $c_1$ cases Incorrectly classified as $c_0$ | d = number of correctly classified $c_1$ cases | |
| | | Sensitivity (Recall) = a/(a+b) | Specificity = d/c+d | Accuracy = (a+b)(a+b+c+d) |

Specificity : The ratio of actual negative cases that are identified correctly.

Table 5-3 shows an example confusion matrix.

Table 5-3. Example of classifications Accuracy measurement

| | | Predicted classed: Positive ($C_0$) | Predicted classed: Negative ($C_1$) | |
|---|---|---|---|---|
| Actual class | Positive ($C_0$) | 80 | 30 | Precision = 70/110=0.63 |
| | Negative ($C_1$) | 40 | 90 | |
| | | Recall=80/120=0.67 | Specificity = 90/240=0.75 | Accuracy = 80+90/240=0.71 |

# Evaluating Classification Models

| Metric | Description | Formula |
|---|---|---|
| Accuracy | What% of predictions were correct? | (TP + TN)/(TP + TN + EP + FN) |
| Misclassification rate | What % of prediction is wrong? | (FP + FN)/(TP + TN + FP + FN) |
| True positive rate OR Sensitivity or recall (completeness) | What % of positive cases did Model catch? | TP/(FN + TP) |
| False positive Rate | What % ‘NO’ were predicted as ‘Yes’? | FP/FP+TN) |
| Specificity | What % ‘NO’ were predicted as ‘NO’? | TN/(TN + FP) |
| Precision(exactness) | What % of positive predictions were correct? | TP(TP + FP) |
| FI score | Weighted average of precision and recall | 2*((precision*recall)/ (precision + recall)) |

# Output interpretation

| Variables in the Equation | | | | | | |
|---|---|---|---|---|---|---|
| Variables | B | S.E. | Wald | df | Sig. | Exp(B) |
| Hours | 1.505 | .629 | 5.727 | 1 | .017 | 4.503 |
| Constant | -4.078 | 1.761 | 5.362 | 1 | .021 | .017 |

- From the above output hours studying is significantly associated with the probability of passing the exam

- Below are the coefficients:
  **Intercept=-4.0777 and Hours =1.5046**

- Probability of passing exam i.e., p= 1/1+exp{-(-4.078+1.505*Hours)}

# Output interpretation

## More:

- Probability of passing exam i.e., p= 1/1+exp{-(-4.078+1.505*Hours)}

- For example, for a student who studies 2 hours, entering the value Hours=2 in the equation gives the estimated probability of passing the exam of **0.26**:

  p= 1/1+exp{-(-4.078+1.505*2)}

- Similarly, for a student who studies 4 hours, entering the value Hours=4 in the equation gives the estimated probability of passing the exam of **0.87**:

  p= 1/1+exp{-(-4.078+1.505*4)}

This table shows the probability of passing the exam for several values of hours studying.

| Hours of study | Probability of passing exam |
|---|---|
| 1 | 0.07 |
| 2 | 0.26 |
| 3 | 0.61 |
| 4 | 0.87 |
| 5 | 0.97 |

# Output interpretation - Classification summary

- **More:**

| Observed | | Predicted | | |
|---|---|---|---|---|
| | | Pass | | Percentage Correct |
| | | .00 | 1.00 | |
| Pass | .00 | 8 | 2 | 80.0 |
| | 1.00 | 2 | 8 | 80.0 |
| Overall Percentage | | | | 80.0 |

**Overall Percentage** – This gives the overall percent of cases that are correctly predicted by the model. By using this model we correctly predicted 80% of the cases.

- Hands On

# Author Info

Thank you

## Venugopala Rao Manneni

A doctor in statistics from Osmania University. I have been working in the fields of data analysis and research for the last 14 years. My expertise is in data mining and machine learning – in these fields I've also published papers. I love to play cricket and badminton.