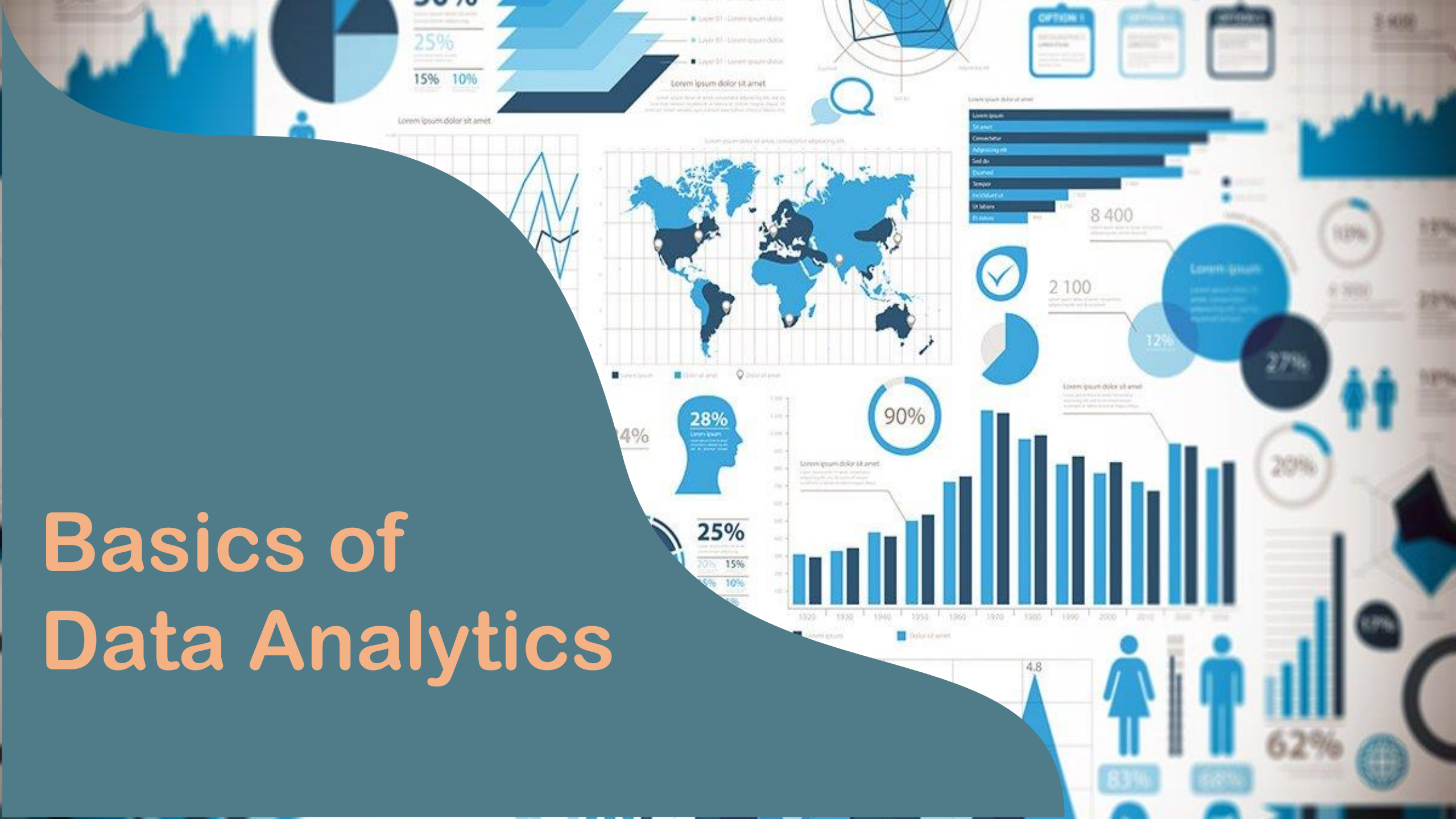
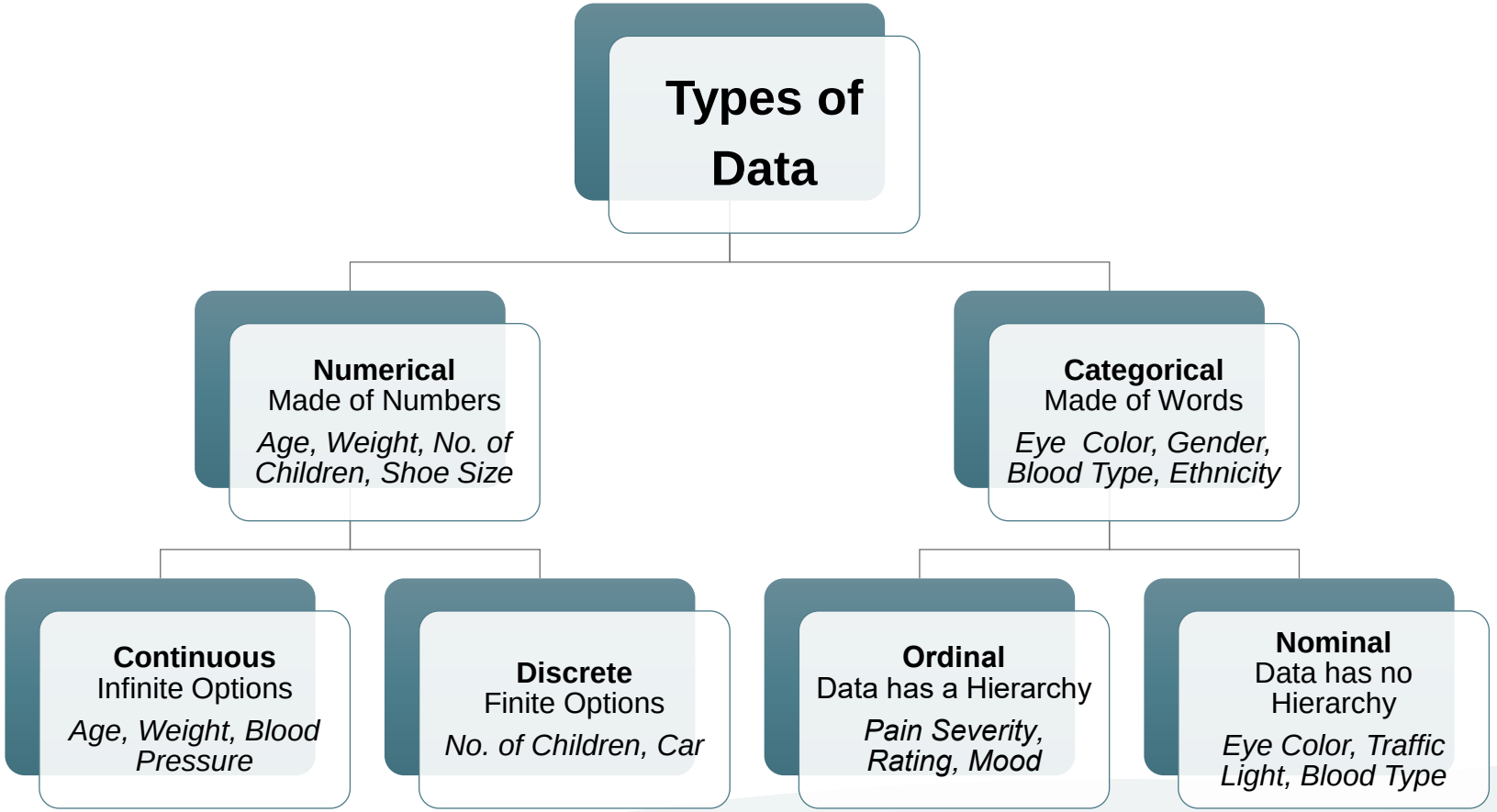
Basics of Data Analytics



Types of Data



Ratio variables → never fall below zero. E.g., Height and weight measure

Interval variable → indicates difference between 2 entities, No True Zero Value E.g., Temperature

1

Relationship within Variables

Dependent / Independent Variables

Age	Height	Smoke	Gender	LungCap
6	62.1	No	Male	6.48
18	74.7	Yes	Female	10.13
16	69.7	No	Female	9.55
14	71	No	Male	11.13
5	56.9	No	Male	4.80
11	58.7	No	Female	6.23
8	63.3	No	Male	4.95
11	70.4	No	Male	7.33
15	70.5	No	Male	8.88
11	59.2	No	Male	6.80
19	76.4	No	Male	11.50
17	71.7	No	Male	10.93

- Features vs Labels
- Independent vs Dependent Variables
- Explanatory vs Response Variable
- Predictors vs Target
- X vs y

- Here Lung Cap is taken as Dependent Variable / Response Variable / Target or Output Variable. Rest are Independent Variables….

Type of Analysis

1

Univariate Analysis

Focusing & Analyzing One Variable at a time

2

Bivariate Analysis

Comparing two variables at a time a) Numeric – Numeric b) Categorical – Numeric c) Numeric – Categorical d) Categorical - Categorical

3

Multivariate Analysis

Comparing More than two variables at a time which may be Categorical or Numerical

Univariate Analysis focuses & analyses one variable at a time…

Statistical Analysis

Measures of Central Tendency

- **Mean:** Totalling all dataset values & dividing by number of Values
- **Median:** Central Value in Dataset
- **Mode:** Most Frequently occurring value in Dataset

Measures of Dispersion

- Range
- Inter-Quartile Range
- **Variance:** Dispersion around mean
- **Standard Deviation:** Square Root of Variance

Measures of Shape

- **Skewness:** Symmetry in Distribution
- **Kurtosis:** Shape of Peaks in Distribution of Data

Visual Analysis

- Histogram
- Bar Chart
- Box Plot
- Pie Chart

Measures of Central Tendency / Location

- Estimate Central Point of Sample. Different ways of estimating central point of dataset are as follows: Above all help us in summarizing data

  - Mean: Most representative value in the data. Used with only numerical data
  - Median: Midpoint of a ranked distribution (sorted data in increasing order)
  - Mode: Most Common data value or highest frequency

Arithmetic Mean

- Arithmetic mean is a mathematical average, and it is the most popular measures of central Tendency. It is frequently referred to as ‘Mean’. It is denoted by $\bar{x}$.
- “It is obtained by dividing sum of all the values by the total number of observations”
- Say we measure the height of 10 students in class and calculate the average
- Mean is affected by extreme values

Median

- Median is a middlemost value of the distribution, or the value which divides the distribution in equal parts, when the values are arranged in order of magnitude
  - Median, Quartiles, Deciles, Percentiles
- Median for Raw data:
  - Arrange data in Ascending order.
  - Apply the formula.
- Median is not affected by extreme values

Mode

- Mode is the most frequent (most frequent) value in the distribution
- Mode is not affected by extreme values
- It is denoted by Z

Measures of Dispersion (Spread)

- Dispersion
  - More Similar the data points → Less Dispersion
  - Less Similar the data points → More Dispersion (More Spread-Out distribution = Larger Dispersion)
- Measures of Dispersion
  - Range
  - Variance / Standard Deviation
  - Interquartile Range

Range & Interquartile Range

- Range is the difference between the highest and the lowest value in the data.
  - R = H - L
  - where, H-Highest Value, L-Lowest Value
- A major drawback of Range is that, since it is based on extreme values, it is highly affected by abnormal values. (Sensitive to Outliers)

- Interquartile Range: Q3-Q1
  - Upper Limit = Q1-1.5(IQR)
  - Lower Limit =Q3+1.5(IQR)

Variance

- Variance is a measurement of the spread between numbers in a data set.
- It measures how far each number in the set is from the mean.
- If the data is a Sample (a selection taken from a bigger Population), then the calculation changes!
- When you have "N" data values:

✓ The Population: divide by N

✓ A Sample: divide by N-1

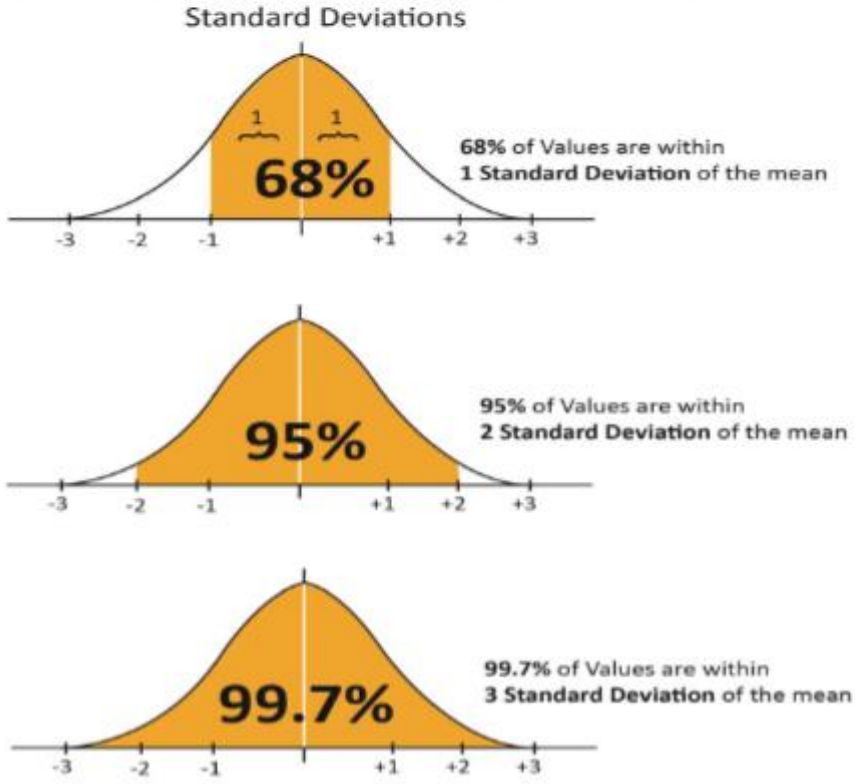
For Sample it is $$s^2 = \frac{1}{(N-1)} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

For Population it is $$\sigma^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i - \mu)^2$$

- To calculate the variance follow these steps:
  - Work out the Mean (Simple average of the numbers)
  - Then for each number: subtract the Mean and square the result (the squared difference).
  - Then work out the average of those squared differences.

Standard Deviation

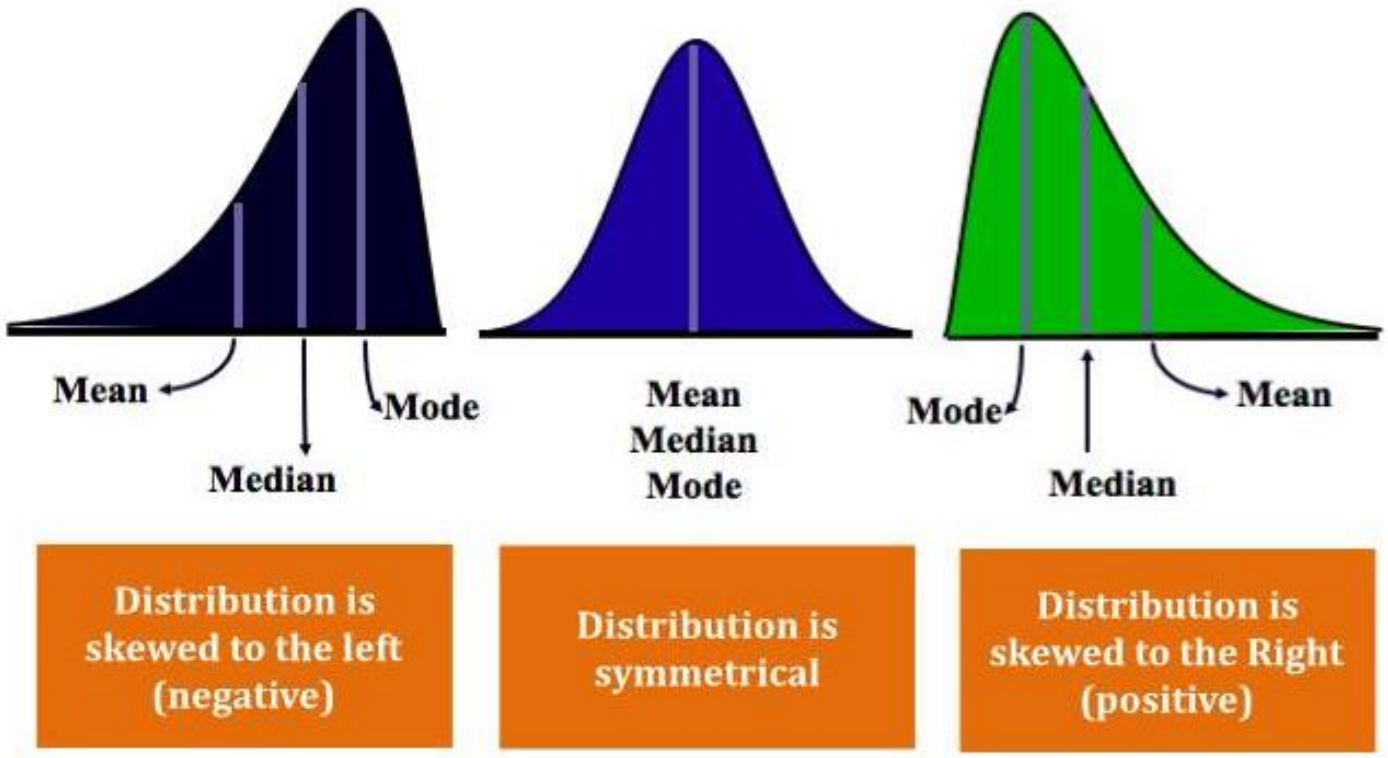
- **Standard deviation** is a measure of the dispersion of a set of data from its mean
- **Standard deviation** s (or σ) is just the square root of variance s^2 (or σ^2)
- When we calculate the standard deviation of normal distribution we find that (generally):



Measures of Shape

- Distribution of Data provides the shape of the data. Distribution of Data is visually represented by Histogram.
- Histogram has two properties
  - Skewness
    - Degree of Skewness → Deviation from Symmetry → Degree of Symmetry
    - Gives → Direction and Amount of Skewness
  - Kurtosis
    - How tall / sharp central peak is → Degree of Peaked Ness
- Normal Distribution = Skewness =0 , Kurtosis =0

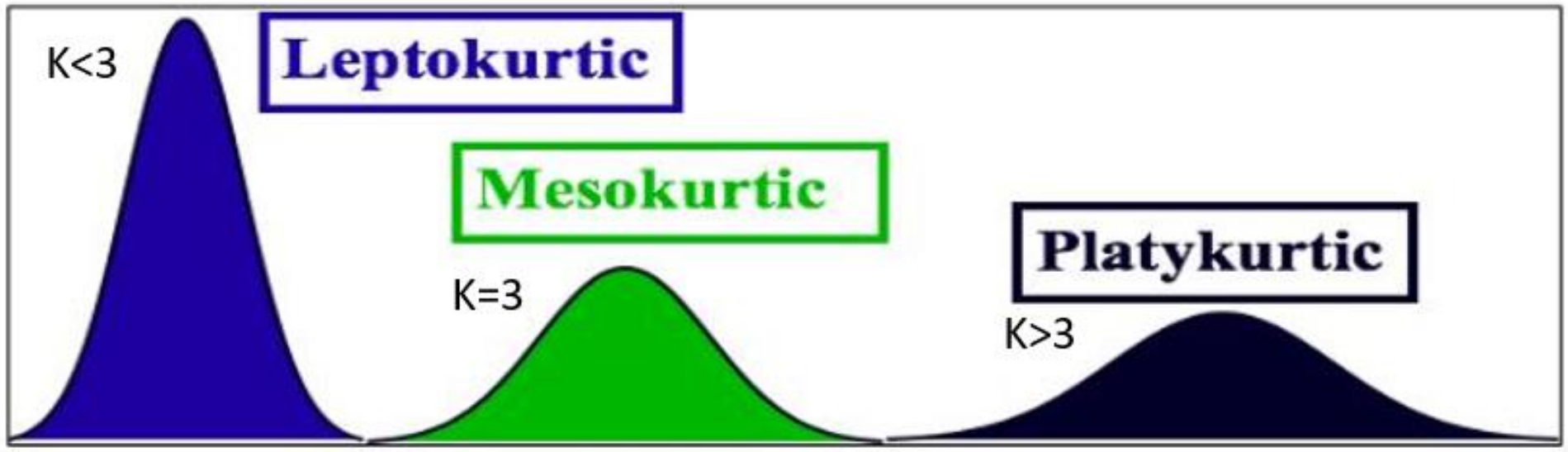
Skewness is the measure of symmetry of Distribution



Mathematical Value of Skewness	Degree of Skewness
0	No Skew
Between +0.5 and -0.5	Slight Skewed
Between 0.5 & 1 or -0.5 and -1	Moderately Skewed
More than +1 or -1	Highly Skewed

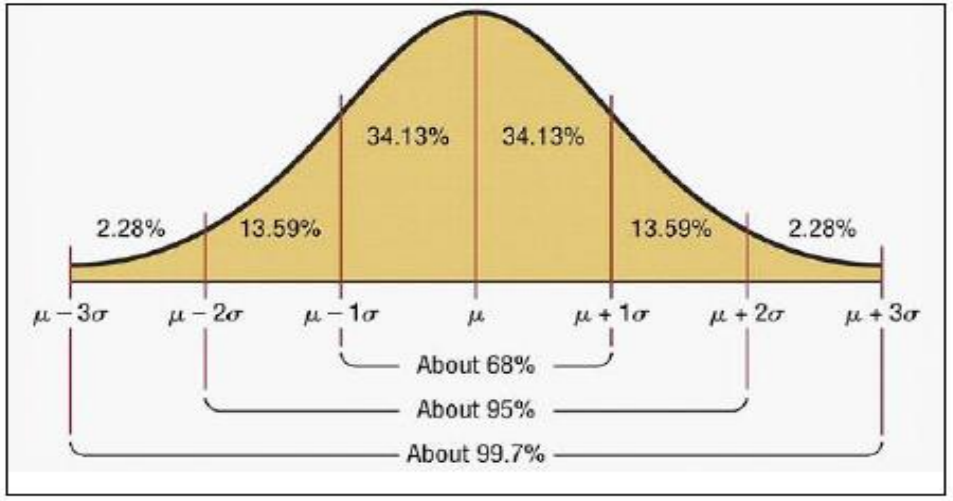
- Examples
  - Income of Individuals is mostly right skewed. There are fewer people with huge wealth.
  - Heights of people will be normal distributed
  - Class grades is usually left skewed (Marks scored in easy exam). Most students score CGPA between 7 & 10

Kurtosis considers the shape of the peaks in the probability distribution of data



Empirical Rule of Normal Distribution (3 sigma rule)

- Empirical rule can be applied for a symmetrical bell shaped frequency distribution
- Empirical rule is known as the three sigma rule
- The range within which approximate percentage of values of the distribution are likely to fall within a given number of standard deviation from the mean is determined below:
  - ✓ Approximately 68.26% of the data is within one standard deviation of the mean.
  - ✓ Approximately 95.44% of the data is within two standard deviations of the mean.
  - ✓ More than 99.72% of the data is within three standard deviations of the mean.



Visualizing Data

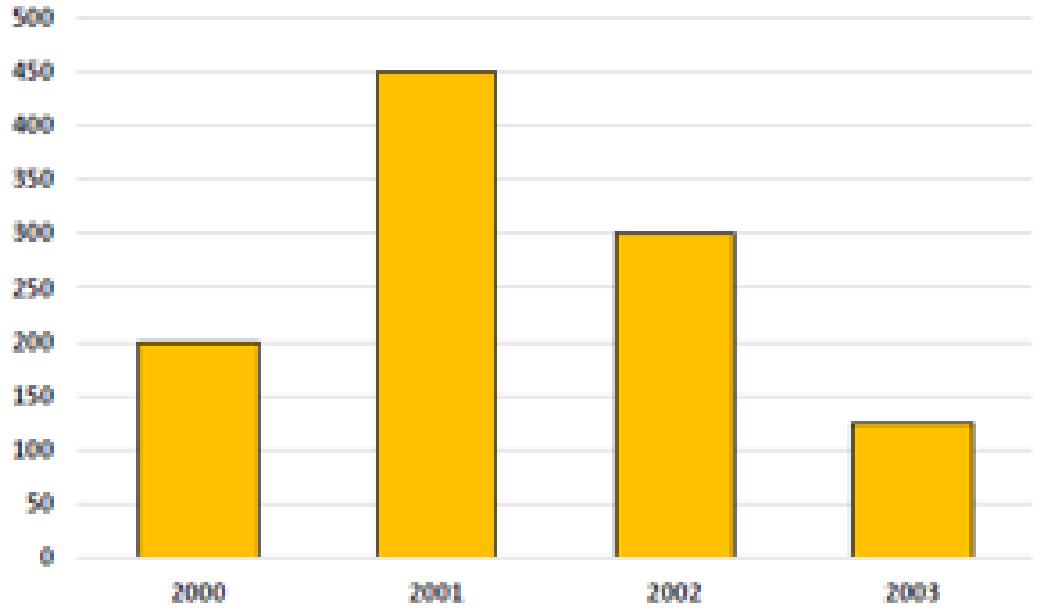
Graph: It is a mathematical diagram that shows the relationship between two or more sets of numbers or measurements. Say Line Graph or Frequency Table can be represented as graph

Chart: It is a type of representation of large sets of data (Datasets)., A chart can take the form of a diagram or a picture or a graph / Table

- Histogram
- Bar Chart
- Box Plot
- Side by Side Box Plot
- Pie Chart
- Line Chart
- Ogive
- Scatter Plot
- Side by Side Bar Chart
- Stacked Bar Chart

Bar Chart

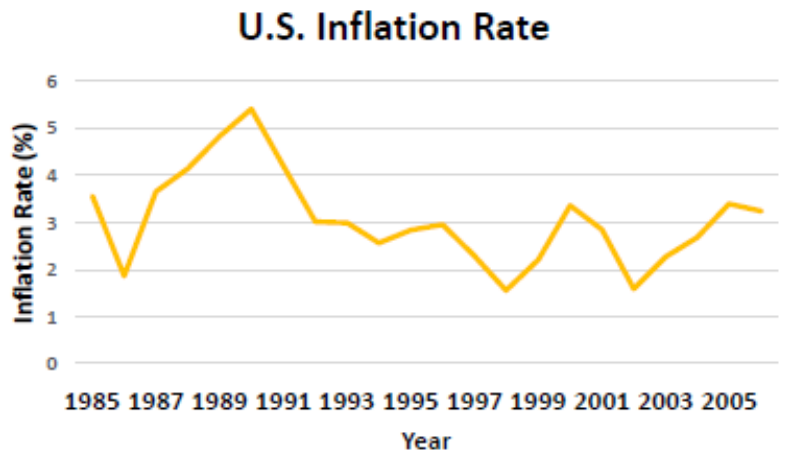
- Bar charts is often used for qualitative (category) data
- Example:



- (Note that bar charts can also be displayed with horizontal bars)

Line Chart

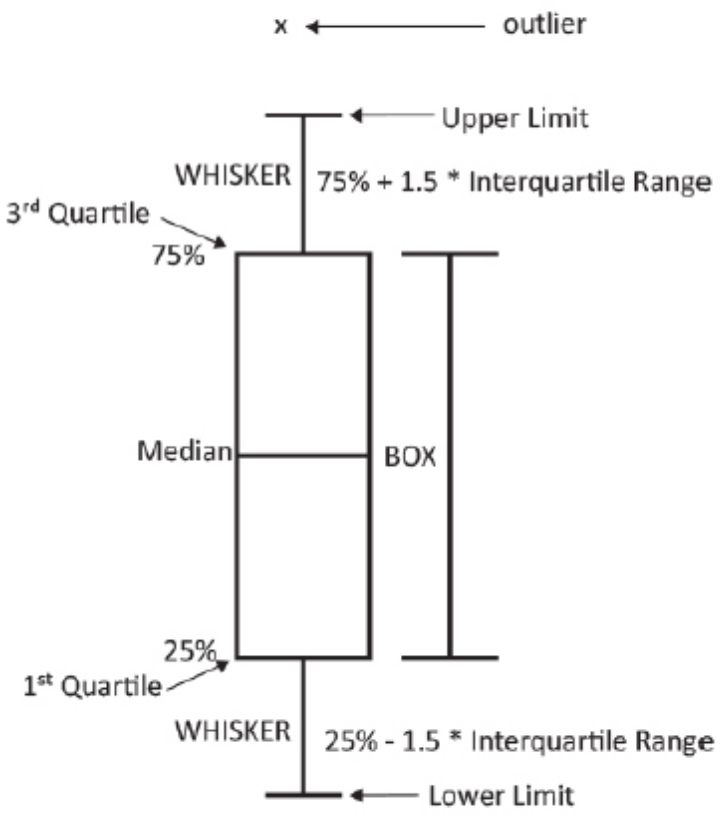
- A line chart or line graph is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments..
- Line charts show values of one variable vs. time
- Time is traditionally shown on the horizontal axis



Year	Inflation Rate
1985	3.56
1986	1.86
1987	3.65
1988	4.14
1989	4.82
1990	5.4
1991	4.21
1992	3.01
1993	2.99
1994	2.56
1995	2.83
1996	2.95
1997	2.29
1998	1.56
1999	2.21
2000	3.36
2001	2.85
2002	1.59
2003	2.27
2004	2.68
2005	3.39
2006	3.24

Box Plot

- **Box-and-whisker plots** are a handy way to display data broken into four quartiles, each with an equal number of data values. It shows where the middle of the data lies. It's a nice plot to use when analyzing how your data is skewed.
- The median is the middle value of the data where half of the points are above and half are below this value.
- The first quartile represents the point where 25% of the data is below it.
- The third quartile represents the point where 75% of the data is below it.
- The whisker extends up to the highest value of upper limit and down to the lowest value of the lower limit.
- The lowest point of the lower whisker is called the lower limit. It equals Q1 – 1.5 * (Q3-Q1 or interquartile range).
- The highest point of the upper whisker is the called the upper limit. It equals Q3 + 1.5 * (Q3-Q1).
- Outliers are points that fall outside the limits of the whiskers.
- The interquartile is represented by the distance between Q1 and Q3.



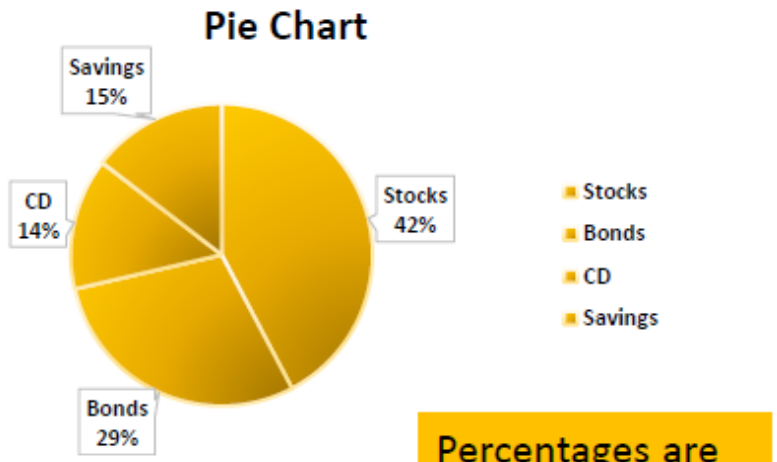
Pie Chart

- A pie chart is a circular statistical graphic which is divided into slices to illustrate numerical proportion.
- In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents.
- Size of pie slice shows the frequency or percentage for each category
- Example:

Current Investment Portfolio

Investment Type	Amount (in thousands $)	Percentage
Stocks	46.5	42.27
Bonds	32	29.09
CD	15.5	14.09
Savings	16	14.55
Total	110	100

(Variables are Qualitative)

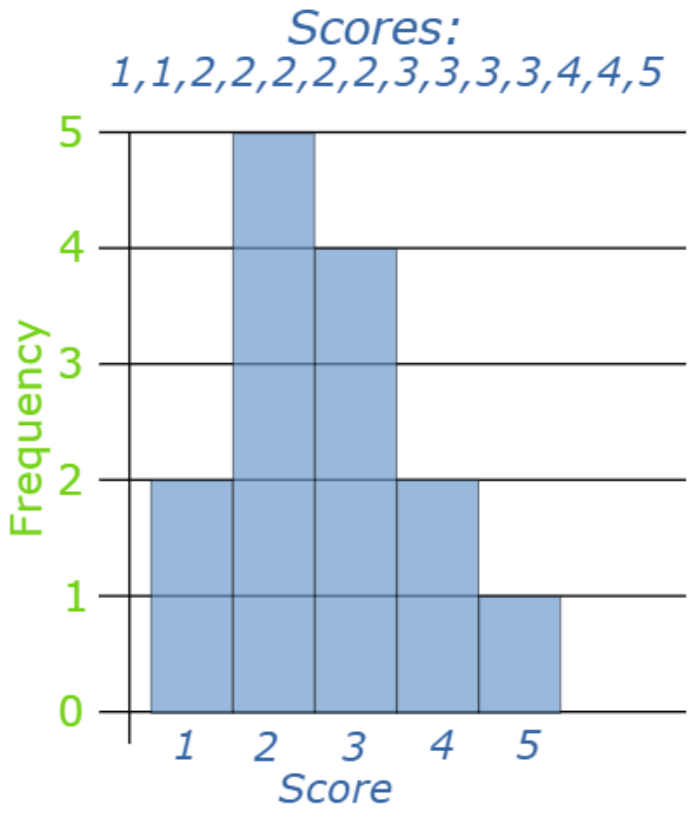


Percentages are rounded to the nearest percent

Frequency Histogram

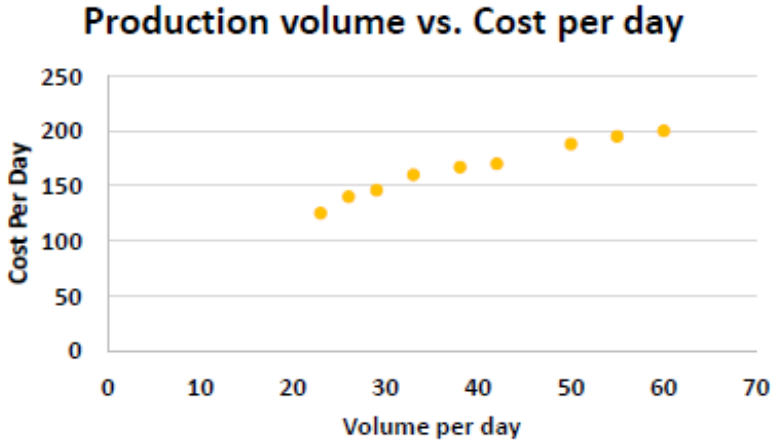
Frequency Histogram

A Frequency Histogram is a special graph that uses vertical columns to show frequencies (how many times each score occurs):



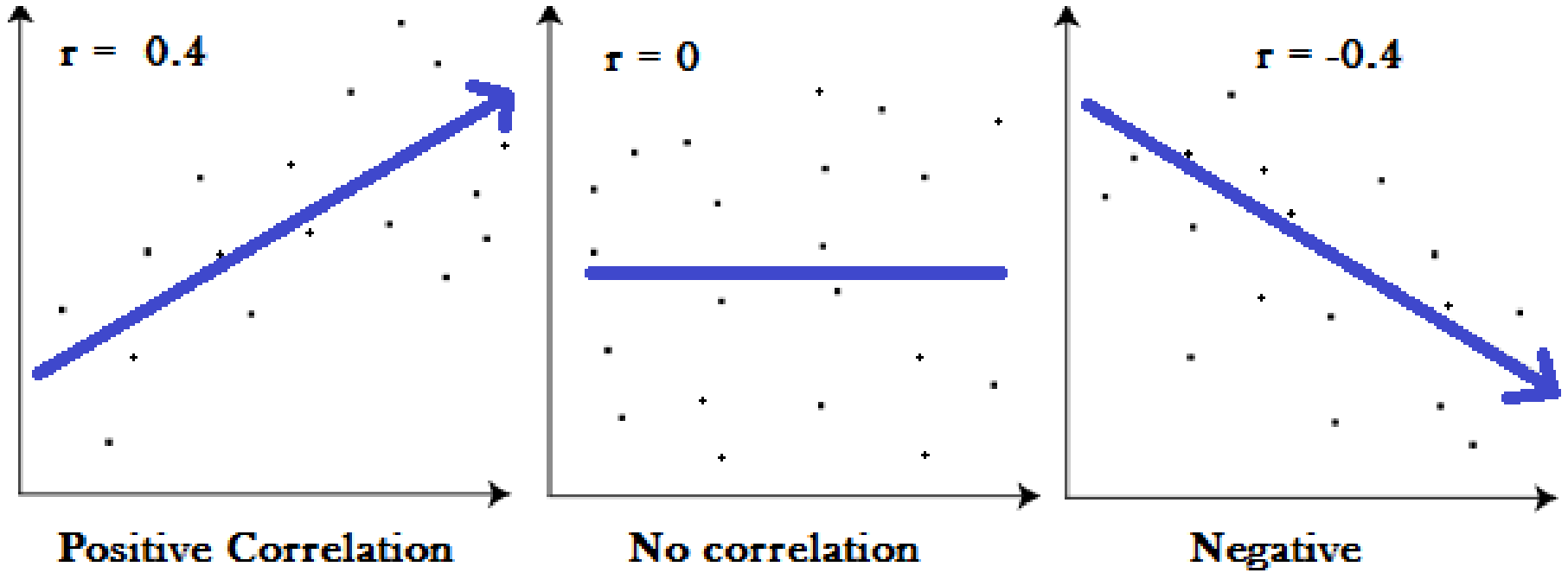
Scatter Diagram

- Scatter Diagrams show points for bivariate data. One variable is measured on the vertical axis and the other variable is measured on the horizontal axis.
- **Purpose:** Scatter plots shows the relationship between two variables.



Volume per day	Cost per day
23	125
26	140
29	146
33	160
38	167
42	170
50	188
55	195
60	200

Correlation (Direction & Strength of Relationship)



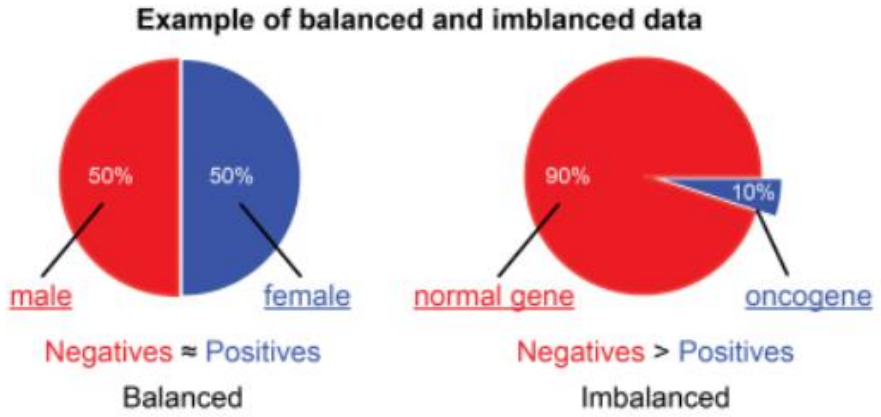
Calculating Correlation

- Pearson's Correlation (named for Karl Pearson)
  - The Pearson correlation coefficient can be used to summarize the strength of the linear relationship between two data samples.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[\, n\sum x^2 - (\sum x)^2 \,][\, n\sum y^2 - (\sum y)^2 \,]}}$$

- Spearman's Correlation (named for Charles Spearman)
  - Two variables may be related by a nonlinear relationship, such that the relationship is stronger or weaker across the distribution of the variables. (ie non-Gaussian distribution)
  - This test of relationship can also be used if there is a linear relationship between the variables, but will have slightly less power (e.g. may result in lower coefficient scores).
  - As with the Pearson correlation coefficient, the scores are between -1 and 1 for perfectly negatively correlated variables and perfectly positively correlated respectively.

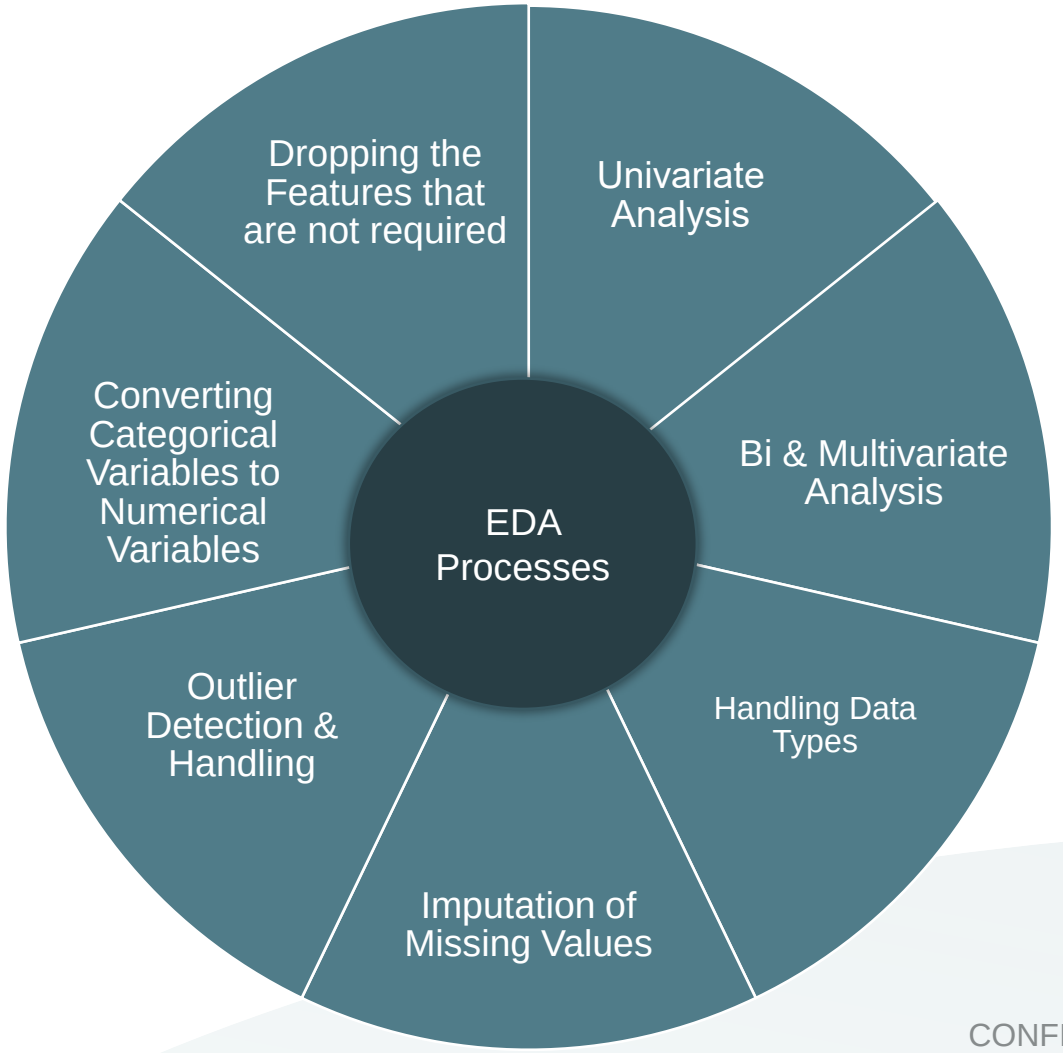
Examining Categorical Data- Class Imbalance



- **Class imbalance** refers to a problem in classification where the distribution of the classes are skewed. This can range from a slight to an extreme imbalance.

- **Resampling the unbalanced datasets.**
  - Resampling involves creating a new version of our imbalanced dataset.
  - There are 2 main approaches for resampling:
    - *Over sampling*: Randomly duplicating entries in the minority class. Appropriate for small datasets.
    - *Under sampling*: Randomly deleting entries from the majority class. Appropriate for large datasets.

Processes in EDA…



"There are no fixed steps in performing EDA. It is creative, based on data and we have all the liberty to choose how we do it"