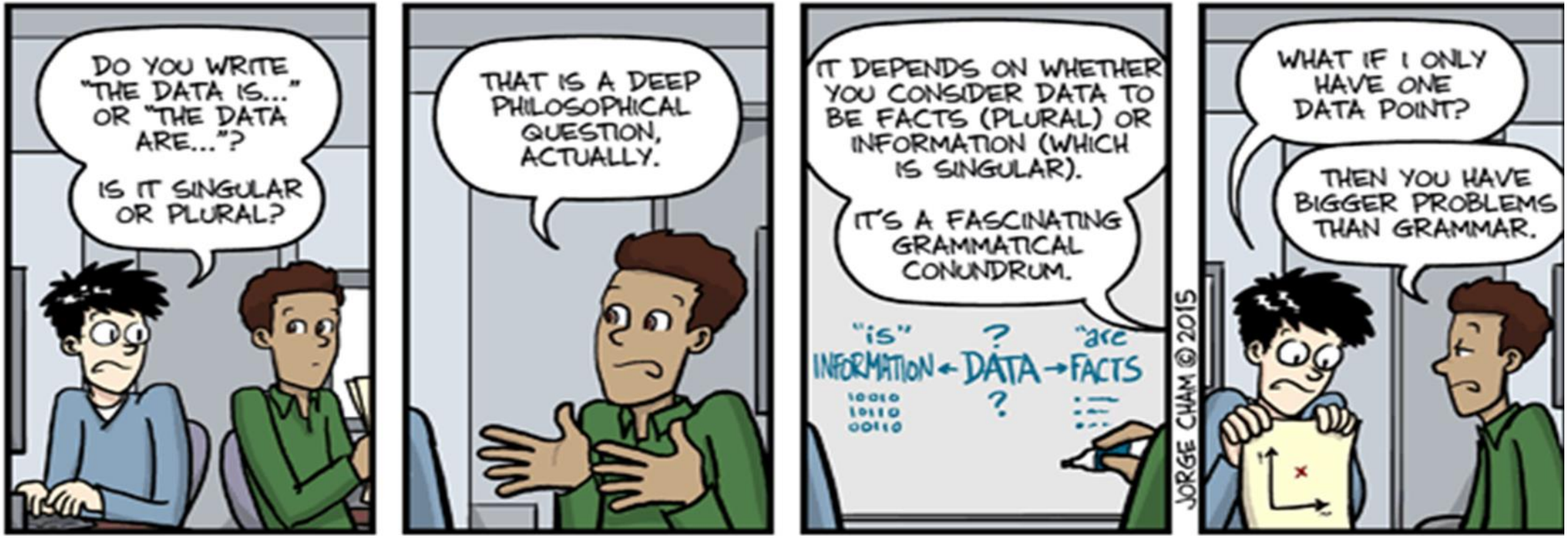
# 5CS037-Concepts and Technologies of AI. Week02, Lecture02

# Data Meets Statistics.

**Introduction to Data, Descriptive Statistics and Data Exploration.**

**Siman Giri**

# 1. Data




# What is Data?

# 1.1 Data-Definitions.

- **"Data"** :a **collection of facts** about any **objects or phenomenon**.
  - Facts/Measurements can be of quantitative(numeric) or qualitative(descriptive) in nature.
  - **Variables** and **Measurements**
- Some similar definitions:
  - Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation
  - Information in digital form that can be transmitted or processed
  - Information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

**Cautions!!!!**

**Datum**

A single piece of information, which can be treated as an observation

**Data**

The plural of datum; multiple observations

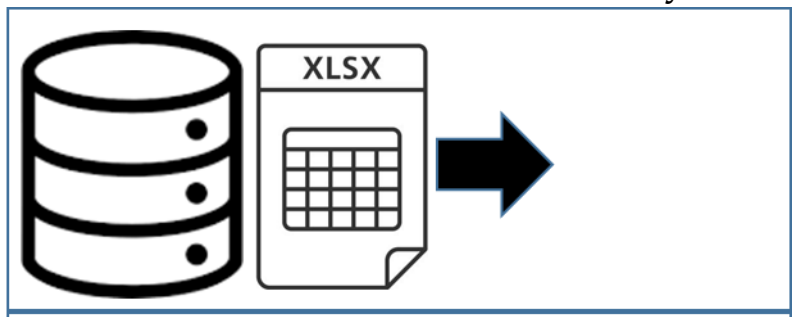
**Dataset**

A homogenous collection of data (each datum must have the same focus)

# 1.2 Data – Types.

## Structured Data

- Structured data refers to data that is organized in a well-defined format, with a clear and consistent structure that can be easily processed and analyzed by computers.
- This type of data is typically stored in databases or spreadsheets, and can be easily queried using SQL (Structured Query Language).
- Examples of structured data include financial records, customer data, and inventory data.




Databases. Spreadsheets. Tools.

## Unstructured Data

- Unstructured data refers to data that does not have a clear and consistent structure, and is not organized in a predefined format.
- This type of data is typically text-heavy, and can include things like emails, social media posts, and other types of unstructured text data.
- Unstructured data is often more difficult to analyze and process than structured data, as it requires more complex algorithms and natural language processing techniques to make sense of the information.

Social Media. e-Commerce Photos Measurements.

# 1.3 Quantitative vs. Qualitative Data.

## Quantitative Data

- It can be expressed as a number, so it can be **quantified**. In simple words, it can be measured by **numerical variables**.
- It tries to find the answers to questions such as
  - **"how many,**
  - **"how much" and**
  - **"how often".**
- Easy manipulation and representation using various **statistical** tools.
- There are two major types:
  - **Discrete Data.**
    - A numerical variable results in discrete data if the possible values of the variable correspond to isolated points on the number line.
  - **Continuous Data.**
    - A numerical variable results in continuous data id the set of possible values forms an entire interval on the number line.

## Qualitative Data

- Sometimes also called **Categorical**
  - This type of data can't be counted or measured easily using numbers and therefore divided into categories.
  - The gender of a person (male, female, or others) is a good example of this data type.
- There are two major types:
  - **Nominal Data**
    - These are the set of values that **don't possess a natural ordering**.
    - The **color** of a smartphone can be considered as a nominal data type as we can't compare one color with others.
  - **Ordinal Data**
    - These types of values have a natural ordering while maintaining their class of values.
    - If we consider the size of a clothing brand then we can easily sort them according to their name tag in the order of **small < medium < large**.

# 1.4 Aspects of data: formats.

**Plain Text**:

Ends in .txt (generally)

No formatting, font type, font size, color, etc.

Text position is provided by whitespace characters (space, tab, return)

```
ALICE'S ADVENTURES IN WONDERLAND

Lewis Carroll

THE MILLENNIUM FULCRUM EDITION 3.0

CHAPTER I. Down the Rabbit-Hole

Alice was beginning to get very tired
of sitting by her sister on the bank,
and of having nothing to do: once or
twice she had peeped into the book her
sister was reading, but it had no
pictures or conversations in it, 'and
what is the use of a book,' thought
Alice 'without pictures or
conversations?'
```

JSON

```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 27,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    },
    {
      "type": "mobile",
      "number": "123 456-7890"
    }
  ],
  "children": [],
  "spouse": null
}
```

***.csv format***

Tab-separated (.tsv/.csv)

**Delimiter**: character that seperates.

```
Bill #, Jack Reed (RI), Elizabeth Warren (MA)
Bill 27, Yay, Yay
Bill 28, Yay, Nay
Bill 30,, Nay
Bill 47, Nay, Nay
Bill 91, Nay, Nay
Bill 105, Yay,Yay
```

XML

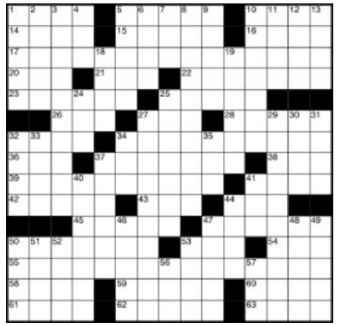
```
<studentsList>
    <student id="1">
        <firstName>Greg</firstName>
        <lastName>Dean</lastName>
        <certificate>True</certificate>
        <scores>
            <module1>70</module1>
            <module12>80</module12>
            <module3>90</module3>
        </scores>
    </student>
    <student ind="2">
        <firstName>Wirt</firstName>
        <lastName>Wood</lastName>
        <certificate>True</certificate>
        <scores>
            <module1>80</module1>
            <module12>80.2</module12>
            <module3>80</module3>
        </scores>
    </student>
</studentsList>
```

# 1.4 Aspects of data: formats.

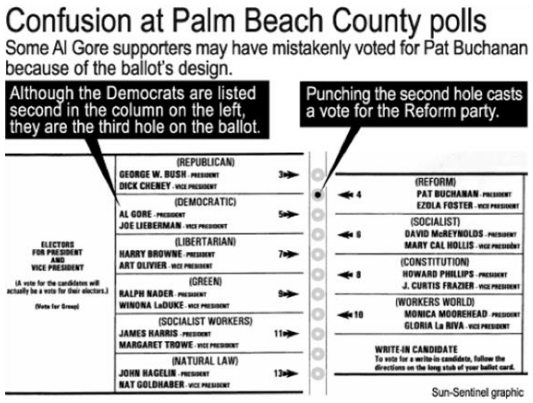
- How difficult is it to analyze a dataset?

hard for computers

easy for computers



| | A | B | C |
|---|---|---|---|
| 1 | name | age | height |
| 2 | Michael | 46 | 5'9" |
| 3 | Jim | 31 | 6'0" |
| 4 | Pam | 29 | 5'7" |
| 5 | Meredith | 53 | 5'6" |
| 6 | Dwight | 35 | 5'10" |




easy for people ⟷ hard for people

# 1.5 Aspects of data: Data Science Process.

Ask an interesting Question.

Collect the Data.

Explore the Data.

Build the Model.

Result/Decision.

**Fig: Elements of Data Science Process.**

Disclaimer!!! This is not the Data Science Course, We may not discuss all the elements mentioned above.

# 1.5 Aspects of data: Data Science Process.

Ask an interesting Question.

- What are the goals?
- What would you do if you had all the data?
- What do you want to predict or estimate?

Collect the Data.

Explore the Data.

Build the Model.

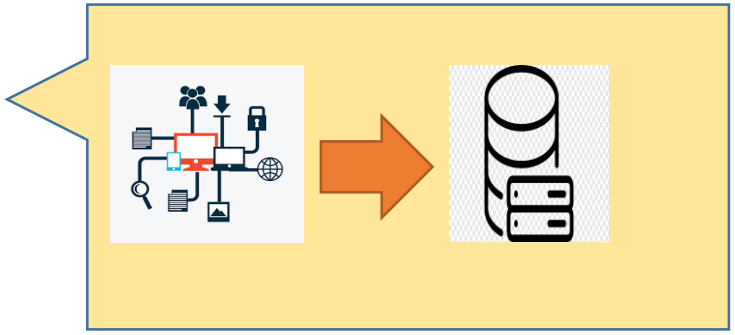
Result/Decision.

Disclaimer!!! This is not the Data Science Course, We may not discuss all the elements mentioned above.

# 1.5 Aspects of data: Data Science Process.

Ask an interesting Question.

Collect the Data.



Explore the Data.

Build the Model.

Result/Decision.

Disclaimer!!! This is not the Data Science Course, We may not discuss all the elements mentioned above.

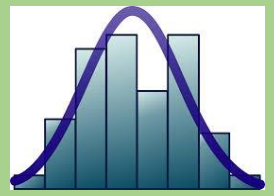
# 1.5 Aspects of data: Data Science Process.

Ask an interesting Question.

Collect the Data.

Explore the Data.

**Statistical analysis**



**Information/Data Visualization**

Build the Model.

Result/Decision.

Disclaimer!!! This is not the Data Science Course, We may not discuss all the elements mentioned above.

# 3.3 Aspects of data: Data Science Process.

Ask an interesting Question.

Collect the Data.

Explore the Data.

Build the Model.

- Build/Train and Validate a Machine Learning Model.

Result/Decision.

Disclaimer!!! This is not the Data Science Course, We may not discuss all the elements mentioned above.

# 1.5 Aspects of data: Data Science Process.

Ask an interesting Question.

Collect the Data.

Explore the Data.

Build the Model.

Result/Decision.

Did it answered the question?

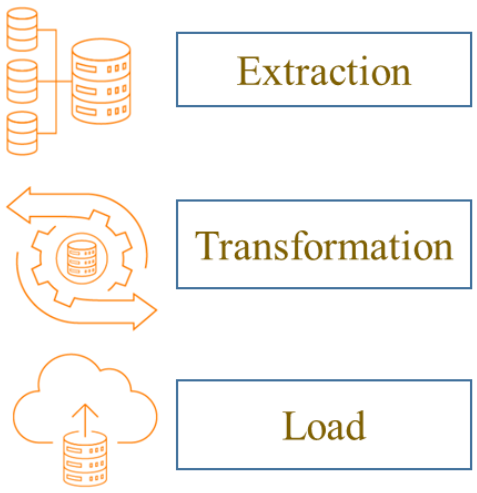
Disclaimer!!! This is not the Data Science Course, We may not discuss all the elements mentioned above.

# 2.Collect the Data.

**Extraction, Transformation and Load**

# 2.1 What is ETL?

- **ETL**, which stands for **extract, transform and load**, is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse or other target system.
  - ETL provides the **foundation** for data analytics and machine learning work streams/pipeline.
  - ETL cleanses and organizes data in a way which addresses specific intelligence needs, which can improve back-end processes or end user experiences.
- ETL is often used by an organization to:
  - Extract data from legacy systems
  - Cleanse the data to improve data quality and establish consistency
  - Load data into a target database




**Disclaimer!! The Definition of ETL have been modified as per the module requirements and may have boarder meaning depending on the field of Data Engineering or Big Data.**

# 2.2 How ETL Works?

Extraction

Transformation

Load

- raw data is copied or exported from source locations to a staging area
- can extract data from a variety of data sources, which can be structured or unstructured
- Those sources include but are not limited to
  - SQL or NoSQL servers
  - CRM and ERP systems
  - Flat files
  - Email
  - Web pages

# 2.2 How ETL Works?

Extraction

Transformation

Load

- the raw data undergoes data (pre)-processing.
- Here, the data is transformed and consolidated for its intended analytical use case.
- This phase (may) involve the following tasks:
  - Filtering, cleansing, de-duplicating, validating, and authenticating the data.
  - Performing calculations, translations, or summarizations based on the raw data. This can include changing row and column headers for consistency, converting currencies or other units of measurement, editing text strings, and more.
  - Conducting audits to ensure data quality and compliance
  - Removing, encrypting, or protecting data governed by industry or governmental regulators
  - Formatting the data into tables or joined tables to match the schema of the target data warehouse.

# 2.2 How ETL Works?

Extraction

Transformation

Load

- In this last step, the transformed data is moved from the staging area into a target data warehouse/system.
- Typically, this involves an
  - initial loading of all data,
  - followed by periodic loading of incremental data changes and,
  - less often, full refreshes to erase and replace data in the warehouse.

# 2.3 Data Collection: How to choose data?

- Considerations when choosing a dataset:
  - What data is necessary to answer our question?
  - **How difficult is it to analyze a dataset?**
  - Is the source authoritative? (.com, .NET, .org, .gov, .name)
  - **Comprehensive data vs. sampled data?**
  - **Biases**
  - What is the allowed usage of data under its license (Copyright issues)?
  - Who collected the data?
  - When was the data collected?
  - How was the data collected?
  - How is the data formatted?
  - Does your data collection procedures need to be approved by an IRB(Review Board)?
  - Confidentiality/Privacy Concerns

# 2.3 Data Collection: Data induced Biases.

- Data induced biases.
- Common biases in selecting the source of data:
  - **Omission**: Using only arguments from one side
  - **Source selection**: Include more sources or more authoritative sources for one side over the other
  - **Story selection**: Regularly including stories that agree or reinforce the arguments of one side
  - **Placement**: Using the benefit of the perceived importance of position to highlight certain stories
  - **Labelling (two types)**:
    - Using only arguments from one side
    - Labeling people on one side of the argument with labels and not the other
  - **Spin**: Story provides only one interpretation of the events

# 2.3 Data Collection: Data induced Biases.

- Data induced biases.
- **Common biases in data (sampled dataset):**
  - **A bias in sampled data occurs when a procedure causes the sample to over-represent a subpopulation**
  - Biases may not necessarily be intentional.
  - Even if you don't *think* your over-/ under-representation of a subpopulation will impact your results, it's still a bias.
  - Always strive to minimize any biases in your data collection procedures.
- **Moral of the story:**
  - Nearly all datasets involve a human in some way or another, and our world is far from being uniform and equal. This is not an excuse but evidence that your dataset probably has bias. The goal is to minimize it as much as possible.

# 2.3 Data Collection: Data(Pre)-processing.

- **Major Tasks in Data (Pre)-processing:**
  - **Data cleaning**
    - Fill in missing values, smooth noisy data, **identify** or remove **outliers**, and resolve inconsistencies
  - **Data integration**
    - Integration of multiple databases, or with in same data set.
  - **Data transformation**
    - Normalization/Scaling (scaling to a specific range)
  - **Data reduction**
    - Feature Selection/Extraction.

# 2.4 Data pre-processing: Data Cleaning-Missing Data.

- **Data is not always available.**
  - E.g., many variables have no recorded value for several attributes, such as customer income in sales data
- **Missing data may be due to :**
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- **Missing data may need to be inferred.**

# 2.4 Data pre-processing: Data Cleaning-Missing Data.

- How to handle **missing Data**?
  - The process of handling missing data is known as Data Imputation. Some of the common approaches are:
    - Fill in the missing value **manually**: tedious + infeasible?
    - Use a **global constant** to fill in the missing value: e.g., "unknown", a new class?!
    - Use the **attribute mean** to fill in the missing value
    - Use the **attribute mean for all samples of the same class** to fill in the missing value: smarter
    - Use the most probable value to fill in the missing value
    - **inference**-based such as regression, Bayesian formula, decision tree

# 2.5 Data pre-processing: Data Transformation/Integration.

- Some common task in Data Transformation are:
  - **Scaling**.
  - **Encoding**.
  - Feature Selection and Feature Engineering (Dimensionality Reduction)
  - Learned Embedding (often for text data).

# 2.5 Data pre-processing: Data Transformation – Scaling.

- **Goal: bring them all with the same range-scale.**
  - Use when different numeric features have different scales (different range of values).
  - Features with much higher values may overpower the others.
- Different methods exist. Most common techniques are:
  - Standard Scaling
  - Min-max Scaling

# 2.5 Data pre-processing: Data Transformation – Scaling.

- **Standard Scaling(Standardization)**:
  - Generally most useful and used, assumes data is normally distributed.
  - For every feature or attribute subtract the mean value and scale by standard deviation.
  - New feature has mean 0 and standard deviation 1.
    - $X_{new} = \frac{X-\mu}{\sigma}$
- **Min-Max Scaling**:
  - Scales all features between a given min and max value.
  - Only used **when min and max has some sense** in data.
  - Sensitive to outliers.
    - $X_{new} = \frac{X-x_{min}}{x_{max}-x_{min}} . (\max - \min) + min.$

# 2.6 Data pre-processing: Data Transformation – Encoding.

- Many algorithms can only handle numeric features, so we need to encode the categorical ones.
  - **Ordinal/Integer Encoding**:
    - Assigns an integer value to each category in the order they are encountered.
    - Only useful for ordinal data types.

| | boro | boro_ordinal | salary |
|---|---|---|---|
| 0 | Manhattan | 2 | 103 |
| 1 | Queens | 3 | 89 |
| 2 | Manhattan | 2 | 142 |
| 3 | Brooklyn | 1 | 54 |
| 4 | Brooklyn | 1 | 63 |
| 5 | Bronx | 0 | 219 |

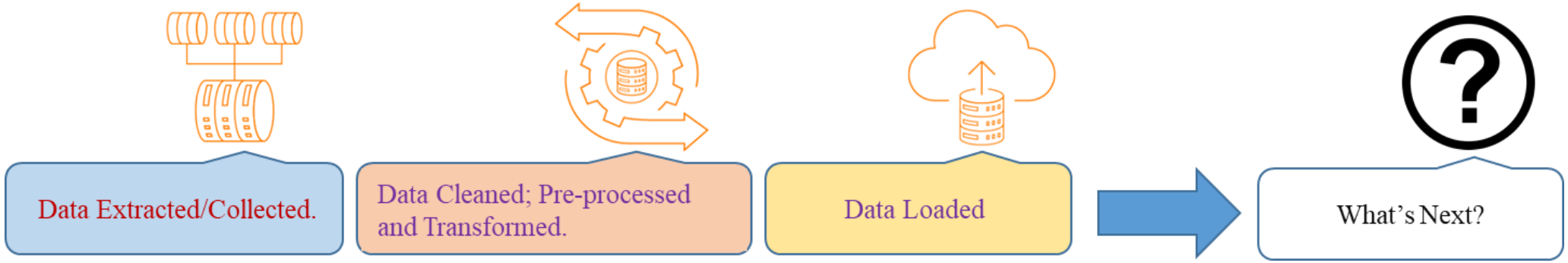
Ordinal Encoding

  - **One Hot Encoding**:
    - Also known as dummy encoding.
    - Adds a new features/attributes for every category in data.

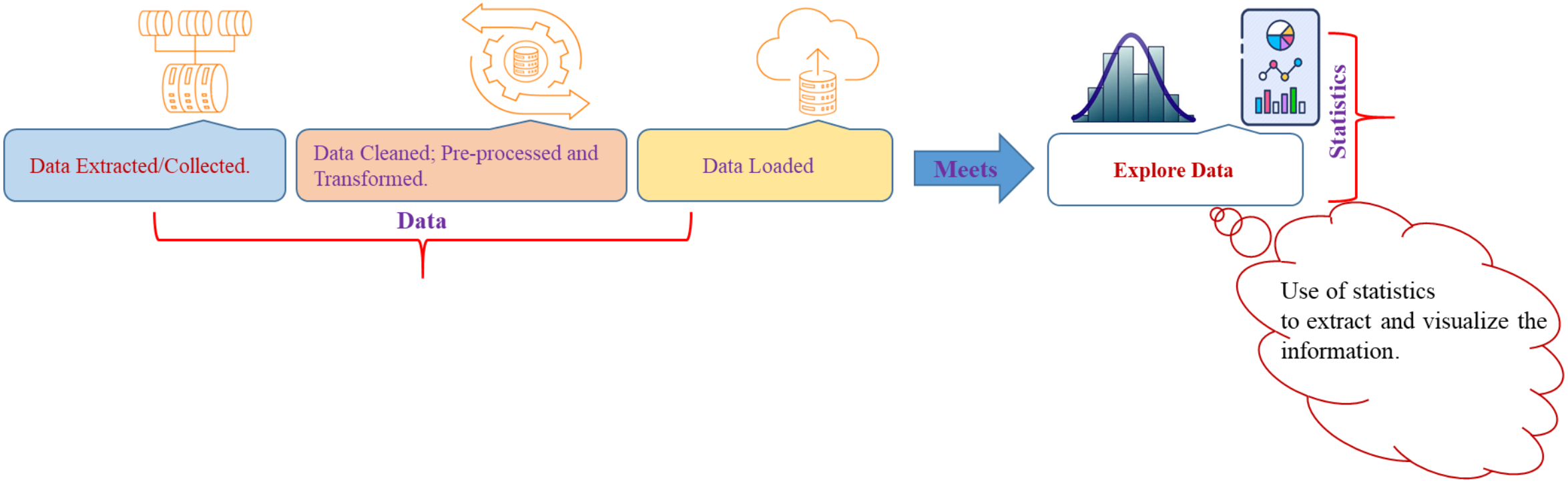
One Hot Encoding

| | boro | boro_Bronx | boro_Brooklyn | boro_Manhattan | boro_Queens | salary |
|---|---|---|---|---|---|---|
| 0 | Manhattan | 0 | 0 | 1 | 0 | 103 |
| 1 | Queens | 0 | 0 | 0 | 1 | 89 |
| 2 | Manhattan | 0 | 0 | 1 | 0 | 142 |
| 3 | Brooklyn | 0 | 1 | 0 | 0 | 54 |
| 4 | Brooklyn | 0 | 1 | 0 | 0 | 63 |
| 5 | Bronx | 1 | 0 | 0 | 0 | 219 |

# Story So far……


Data Extracted/Collected.
Data Cleaned; Pre-processed and Transformed.
Data Loaded
What's Next?

# Story So far……


Data Extracted/Collected.
Data Cleaned; Pre-processed and Transformed.
Data Loaded
Data
Meets
Explore Data
Statistics
Use of statistics to extract and visualize the information.

# 3.Statistics.

*Background and Motivation for statistics.*

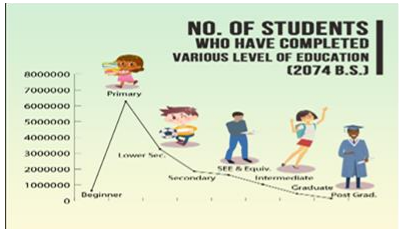
# 3.1 Statistics-Definition.

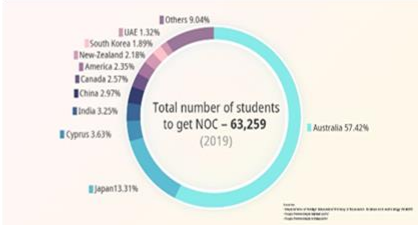
- "a branch of **mathematics** dealing with the **collection, analysis, interpretation, and presentation** of masses of numerical data"
- " a collection of (quantitative) data"
  - ***-Merriam Webster Dictionary***
- *"The objective of statistics is to make an **inference** about a **population** based on information contained in a **sample** from that population and to provide an associated measure of goodness for the inference."*
  - ***- Wackerly et.al. -Mathematical Statistics with Applications***
- It is the discipline that provides tools, techniques and methods to help us make sense of data.

## *Use case of statistics:*

### News

### Policy Making




### Key Social Indicator

NEPAL — Social Statistics: Key Indicators — STATISTICS

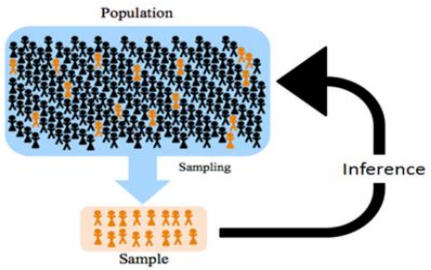
SOCIAL INDICATORS (Latest Available Year)

| Poverty Gap Ratio at USD 1.90 (PPP) a day (%) | Literacy Rate of 15–24 Year-olds (%) | | | Ratio of Girls to Boys in Education | | |
|---|---|---|---|---|---|---|
| | 2015 | | | 2014 | | |
| 2012 | Total | Female | Male | Primary | Secondary | Tertiary |
| 3.04 | 89.88 | 87.39 | 92.59 | 1.08 | 1.07 | 0.82 |

| Proportion of Population below Poverty Line (in % of pop.) | Total Net Enrolment Ratio in Primary Education (%) | | | Proportion of Pupils Starting Grade 1 Who Reach Last Grade (%) | | |
|---|---|---|---|---|---|---|
| | 2014 | | | 2013 | | |
| 2014 | Total | Female | Male | Total | Female | Male |
| 25.20 | 97.13 | 96.36 | 97.86 | 70.10 | 71.74 | 68.44 |

Source: Asian Development Bank ADB, "Basic Statistics 2016"

# Terminology Alert!!!




- **Population**
  - Collection of all possible measurements/outcomes for any given context/objectives is called the population.
- **Sample**
  - Subset of **population** is **sample**.
  - Randomly sampled from the population, sample should be representative of the whole population.
  - Sample (may) induces biasness.
- **Inference**
  - is the process of using a **sample** to infer the properties of a **population**.

# 3.3 Statistics-Branches.

- **Descriptive Statistics:**
  - describe the features or characteristics of data.
  - Summarize and delivers quantitative insights about the data through **numerical** or **graphical representations**.
- **Inferential Statistics:**
  - used to make **conclusions or inferences** of **entire populations** from the available **sample data**.

# 3. Statistics -Summary.

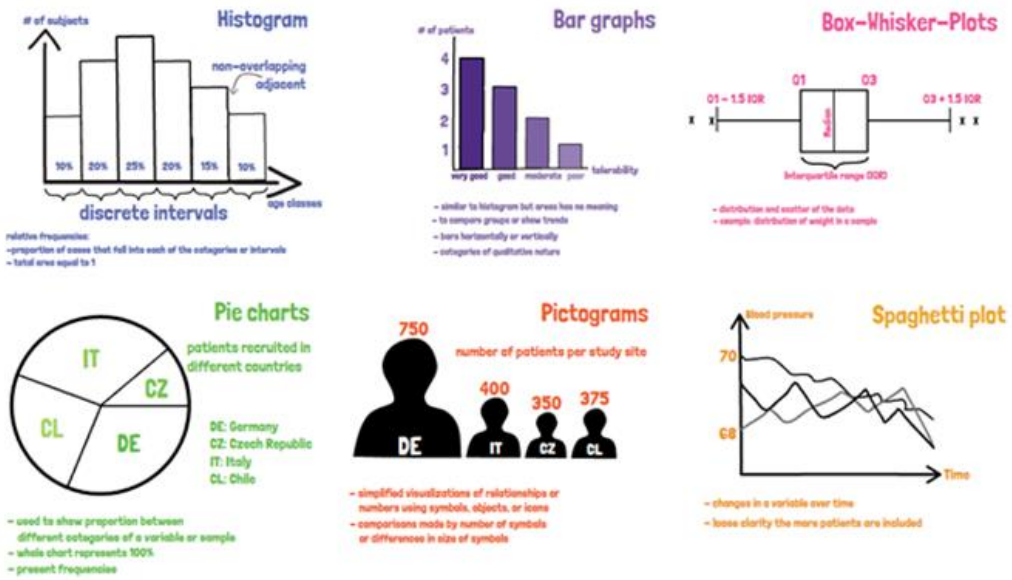
- Statistics is a science whose focus is on collecting, analyzing and drawing conclusion from data.
- **Statistics measures the variability in data**.

# 4.1 Descriptive Statistics: Describing Data.

- describes the data i.e. provides **quantitative** insights about the data through **numerical** or **graphical** representation.
- It only **reflects the data to which they are applied**.
- Types of **Descriptive(numerical) Statistics**:
  - **Graphical**
    - Tables
    - Graphs
  - **Numerical**
    - **Measure of Frequency: Count, Percent, Frequency**
    - **Measure of Central Tendency: Mean, Median, and Mode**
    - Measure of Dispersion: Range, Variance, Standard Deviation
    - Measure of Position: Percentile, Quartile

# 4.Descriptive Statistics: Graphical Methods.

## Tabular and Pictorial Methods.



# 4.1 Descriptive Statistics: Tabular Methods.

- **Frequency Distributions for Categorical Data:**
  - A tabular display of data can be an effective way to **summarize and communicate information**.
  - A **Frequency distribution table**(categorical data) is a table that displays the possible categories along with associated frequencies or relative frequencies.
    - frequency: number of time particular category appear in the dataset.
    - relative frequencies: For a particular category relative frequency is the fraction or proportion of the observations resulting in the category: given by-
      - $\boldsymbol{Relative\ Frequency} = \frac{\boldsymbol{frequency}}{\boldsymbol{number\ of\ observations\ in\ the\ dataset}}$.
    - cumulative relative frequency: is the accumulation of the previous relative frequencies.
      - To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row.

# Tabular Methods: Example.

| Present the information in the frequency table | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Example:1: pets owned by tenants in a building | | | | | | | | | |
| Dog | Cat | Dog | Hamster | Snake | Cat | Dog | Parrot | Hamster | Dog |

# Tabular Methods: Example.

| Present the information in the frequency table | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Example:1: pets owned by tenants in a building | | | | | | | | | |
| Dog | Cat | Dog | Hamster | Snake | Cat | Dog | Parrot | Hamster | Dog |

| Objects(Data) | frequency | relative frequency | cumulative frequency |
|---|---|---|---|
| Dog | 4 | 4/10 = 0.4 | 0.4 |
| Cat | 2 | 2/10 = 0.2 | 0.4 + 0.2 = 0.6 |
| Hamster | 2 | 2/10 = 0.2 | 0.6 + 0.2 = 0.8 |
| Snake | 1 | 1/10 = 0.1 | 0.8 + 0.1 = 0.9 |
| Parrot | 1 | 1/10 = 0.1 | 0.9 + 0.1 = 1.0 |

# 4.2 Descriptive Statistics : Graphical Methods-Bar Chart.

**Bar chart**:

When to use: Categorical Data.

How to construct?

Hope You Know it.

What to look for?

Frequently and infrequently occurring categories.

**Comparative Bar Chart**:

What to look for?

Visual comparison of two or more categories.

**Pie-chart**:

When to use:

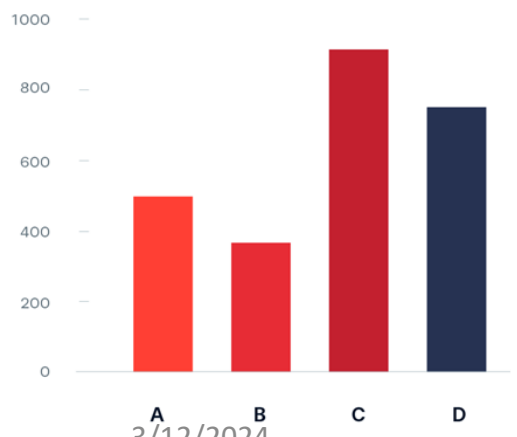
Categorical Data with relatively small number of possible categories.

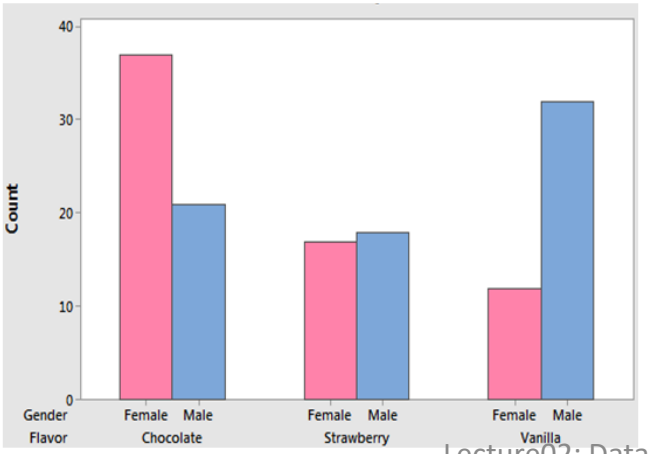
Pie charts are most useful for illustrating proportions of the whole data set for various categories.

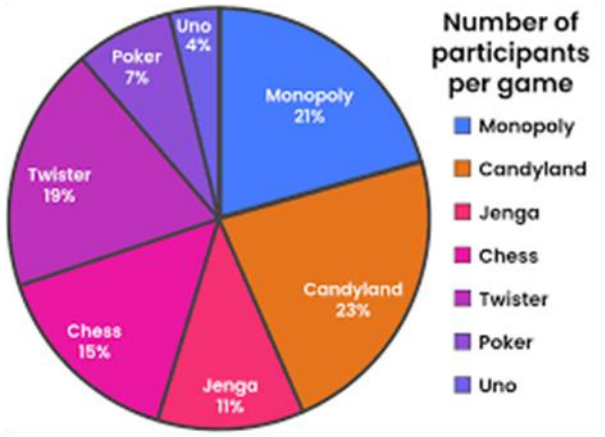
How to construct?

What to look for?

Categories that form large and small proportions of the dataset.



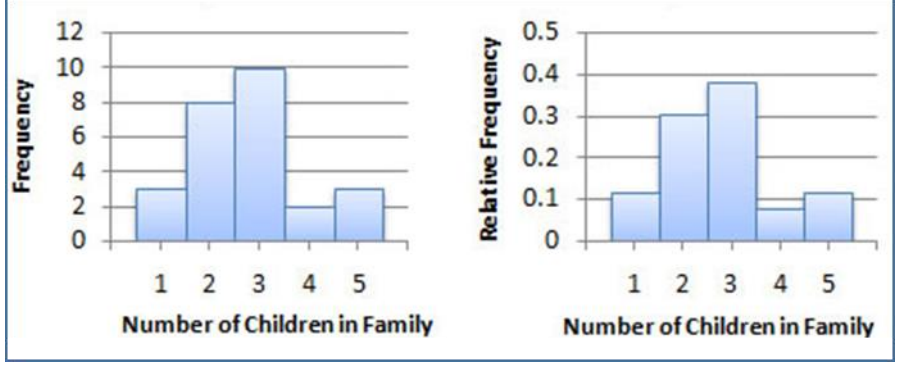



# 4.2 Descriptive Statistics : Graphical Methods-Histograms.

- **Histogram for Discrete Numerical Data:**
  - When to use: Discrete numerical data. Works well even for large datasets.
  - How to construct?
    - Draw a horizontal scale, and mark the possible values of the variable.
    - Draw a vertical scale, and mark it with either frequency or relative frequency.
    - Above each possible value on vertical scale, draw a rectangle centered at that value.(so that the rectangle for 1 is centered at 1, the rectangle for 5 is centered at 5, and so on.
    - The height of each rectangle is determined by the corresponding frequency or relative frequency.
  - What to look for?
    - Central or typical value.
    - Extent of spread or variation.
    - General shape.
    - Location and number of peaks.
    - Presence of gaps and outliers.

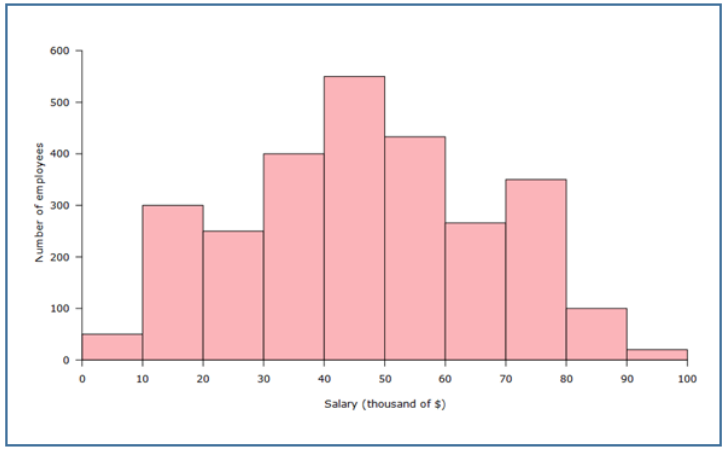
| Children | frequency | relative frequency |
|---|---|---|
| 1 | 3 | 3/26 ≈ 0.12 |
| 2 | 8 | 8/26 ≈ 0.31 |
| 3 | 10 | 10/26 ≈ 0.38 |
| 4 | 2 | 2/26 ≈ 0.08 |
| 5 | 3 | 3/26 ≈ 0.12 |



# 4.2 Descriptive Statistics : Graphical Methods-Histograms.

- **Histogram for Continuous Numerical Data when class interval has equal widths**:
  - When to use: Continuous numerical data. Works well even for large datasets.
  - How to construct?
    - Mark the boundaries of the class intervals on a horizontal axis.
    - Mark either with frequency or relative frequency on a vertical axis.
    - Above each possible value on vertical scale, draw a rectangle on the corresponding interval such that edges are at the class boundaries.
  - What to look for?
    - Central or typical value.
    - Extent of spread or variation.
    - General shape.
    - Location and number of peaks.
    - Presence of gaps and outliers.

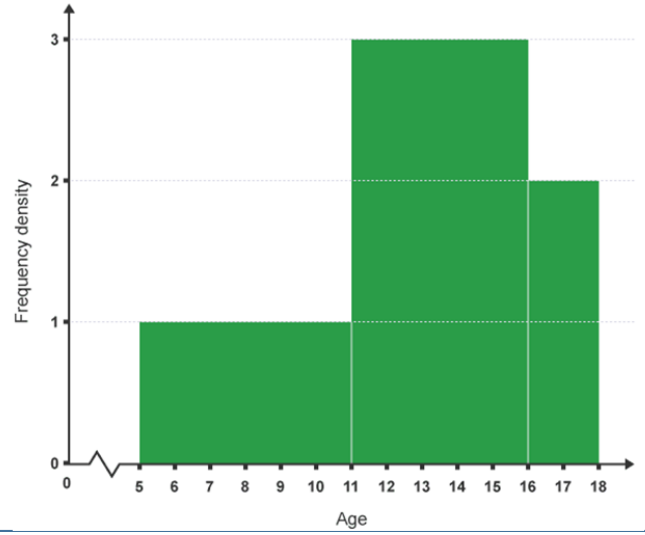
| Salary (thousands) | No of Employees |
|---|---|
| 0-10 | 50 |
| 11-20 | 300 |
| 21-30 | 250 |
| …….. | …… |
| 91- | 20 |



# 4.2 Descriptive Statistics : Graphical Methods-Histograms.

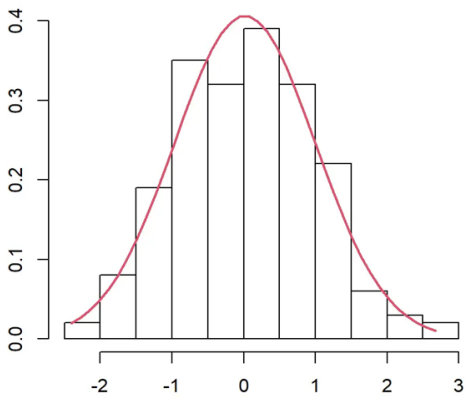
- **Histogram for Continuous Numerical Data when class interval has un-equal widths:**
  - When to use: Continuous numerical data. Works well even for large datasets.
  - How to construct?
  - In this case, frequencies or relative frequencies should not be used on the vertical axis.
  - Instead, the height of each rectangle, called the density for the class, is given by:
    - Density = rectangle height = $\frac{\textit{relative frequency of class interval}}{\textit{class interval width}}$

| Age | frequency | Class width | Frequency density |
|---|---|---|---|
| 5-10 | 6 | 6 | 6/6 = 1 |
| 11-15 | 15 | 5 | 15/5 = 3 |
| 16-17 | 4 | 2 | 4/2 = 2 |

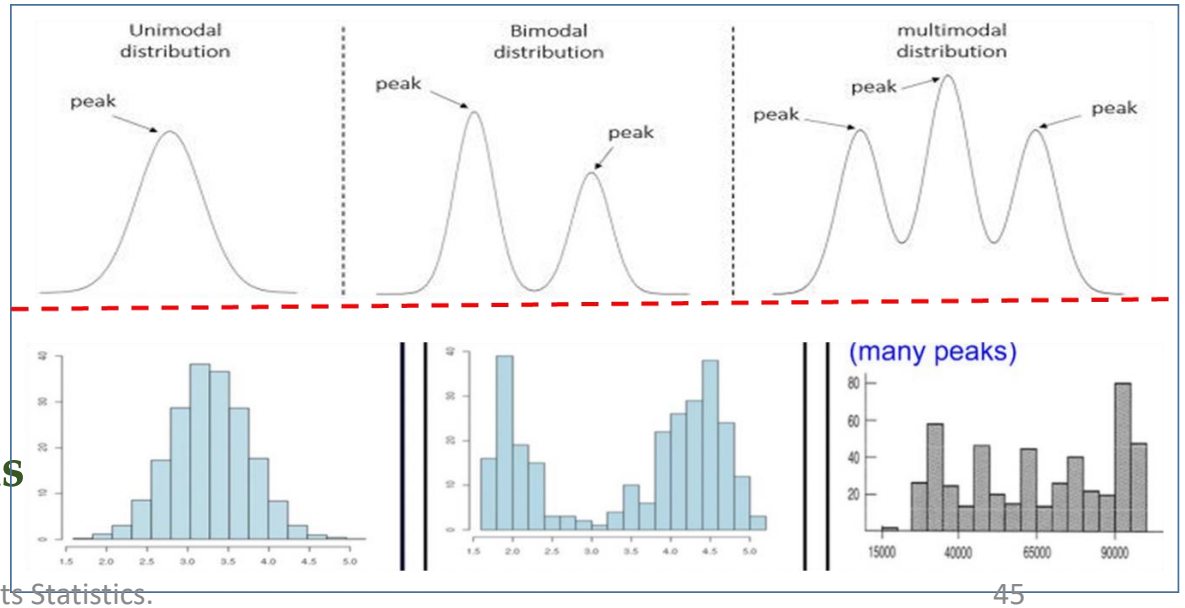


# 4.3 Interpreting Histogram: Shapes.

- **General Shapes** is an important characteristics of a histogram.
  - While describing various **shapes** it is convenient to approximate the histogram itself with a smooth curve (called a **smoothed histogram**: obtained by superimposing a smooth curve on the rectangles, that illustrate the various possibilities.).
- Shapes and interpretation: A histogram is said to be
  - **Unimodal: if it has a single peak**
  - **Bimodal: if it has two peaks.**
  - **Multimodal: if it has more than two peaks**

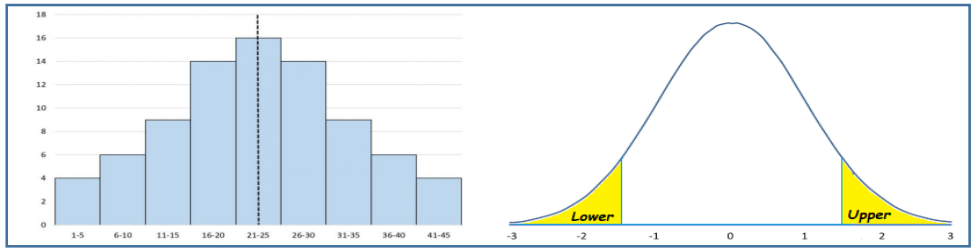



**Smoothed Histogram**

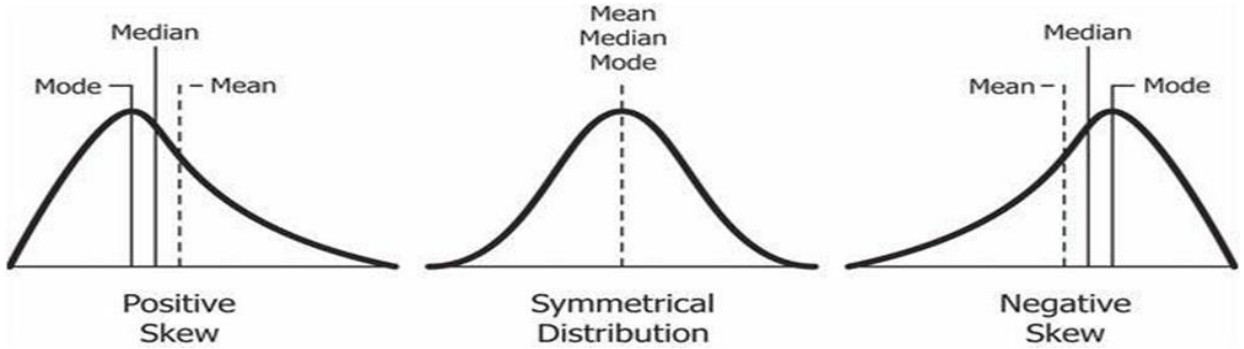


# 4.4 Interpreting Histogram: Unimodal Shapes.

- *Unimodal Symmetric Histogram: type of histogram that has perfectly identical two halves.*

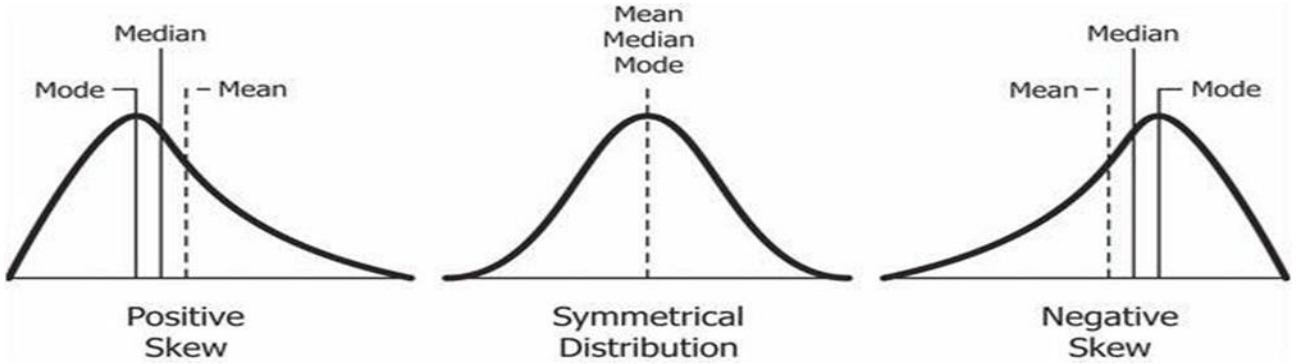



- A Unimodal Histogram which is not symmetric are called skewed.
  - Positively Skewed: If the upper tail of the histogram stretches out much further than lower tail.
  - Negatively Skewed: If lower tail is much longer than upper tail.



# Terminology Alert!!!

- Skewness
  - Measure the symmetry of the distribution
  - If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
  - If the skewness is between -1 and -0.5(negatively skewed) or between 0.5 and 1(positively skewed), the data are moderately skewed.
  - If the skewness is less than -1(negatively skewed) or greater than 1(positively skewed), the data are highly skewed.



# 5. Descriptive Statistics: Numerical methods.

## *Measurement of Central Tendency.*

# 5.1 Measure of Central Tendency: The Mean.

- Arithmetic Mean/average:
  - **Sample mean**:
    - Average of all the observation in sample data i.e.
    - The mean of a sample of "n" measured responses given by
      - $x = \{x_1, .. x_n\}$
    - is given by:
      - $\bar{x} = \frac{1}{n}\sum_i^n x_i.$
  - The corresponding **population mean** is denoted by $\mu$.

- Good to know:
  - Geometric Mean:
    - Given by: $GM = \sqrt[n]{\pi x}$
  - Harmonic Mean:
    - Given by: $HM = n/\sum\left(\frac{1}{x_i}\right)$
  - Quadratic Mean:
    - Given by: $QM = \sqrt{\sum x_i^2/n}$

# 5.2 Measure of Central Tendency: The Median. And The Mode.

## The Median

- The **Middle** value when ordered….
  - Once the data values have been listed in **order from smallest to largest**, **the median** is the **middle value** in the list, and it divides the list into two equal parts.

$$Median(X) = \begin{cases} X\left[\frac{N+1}{2}\right]^{th} \text{ term ; When N is odd.} \\ \frac{X\left[\frac{N}{2}\right]^{th} \text{term} + X\left[\frac{N}{2}+1\right]^{th} \text{term}}{2} \text{ ; When N is even.} \end{cases}$$

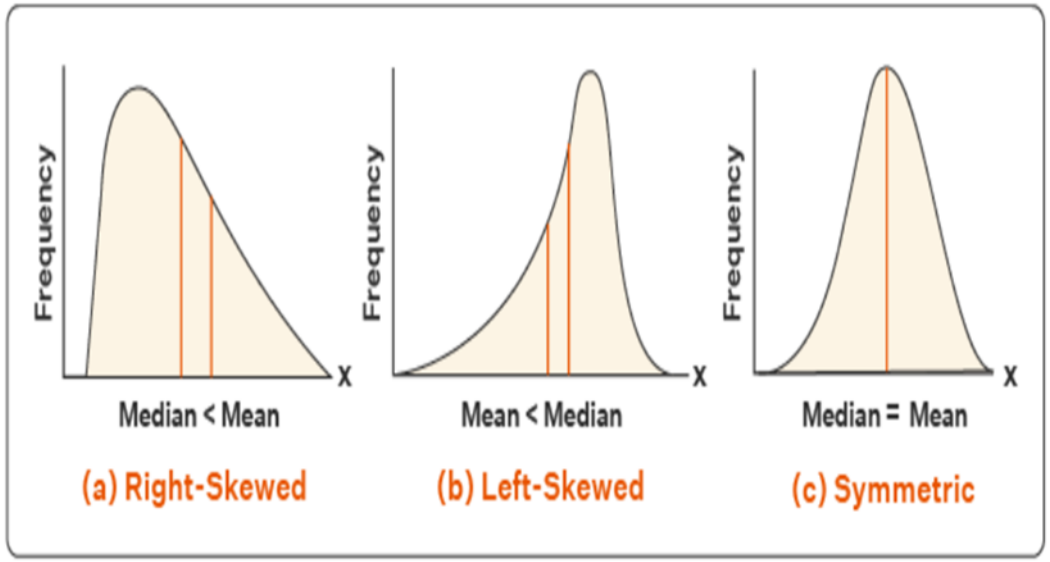
$X$ = Ordered list of values in dataset.
$N$ = number of values in dataset.

## The Mode

- The most **common data** point is called the mode.
- It may give you the **most likely** experience rather than the "typical" or "central" experience.
- **In symmetric distributions, the mean, median, and mode are the same.**
- In skewed data, the mean and median lie further toward the skew than the mode.

# 5.2 Measure of Central Tendency: Mean vs. Median vs. Mode.

- For symmetric distributions, **mean = median**.
- For **skewed distributions**, mean is drawn in direction of longer tail, relative to median.
- Mean sensitive to "outliers" (**median often preferred for highly skewed distributions**).
- When **distribution symmetric or mildly skewed or discrete with few values, mean preferred** because uses numerical values of observations.



# 6. Descriptive Statistics: Numerical methods.

## Measurement of Dispersion.

# 6.1 Measure of Dispersion: The Range and Interquartile Range.

- **The Range:**
  - The spread, or the distance, between the lowest and highest values of a variable.
  - To get the range for a variable, you subtract its lowest value from its highest value.
- **The Interquartile Range:**
  - The interquartile range is a measure of variability that is resistant to the effects of outliers. It is based on quantities called quartiles.
  - The lower quartile separates the bottom 25% of the data set from the upper 75%, and the upper quartile separates the top 25% from the bottom 75%.
  - The middle quartile is the median, and it separates the bottom 50% from the top 50%.

# 6.1 Measure of Dispersion: The Variance and Standard Deviation.

- *The Variance:*
  - A (average) measure of the spread of the data points from the mean.
  - The larger the variance, the further the individual datapoints are from the mean.
  - The smaller the variance, the closer the individual datapoints are to the mean.
  - The population variance is given by: $\sigma^2 = \sum_i \frac{(x_i - \mu)^2}{N}$
  - The sample variance then will $s^2 = \sum_i^n \frac{(x_i - \bar{x})^2}{n-1}$.

# 6.1 Measure of Dispersion: The Variance and Standard Deviation.

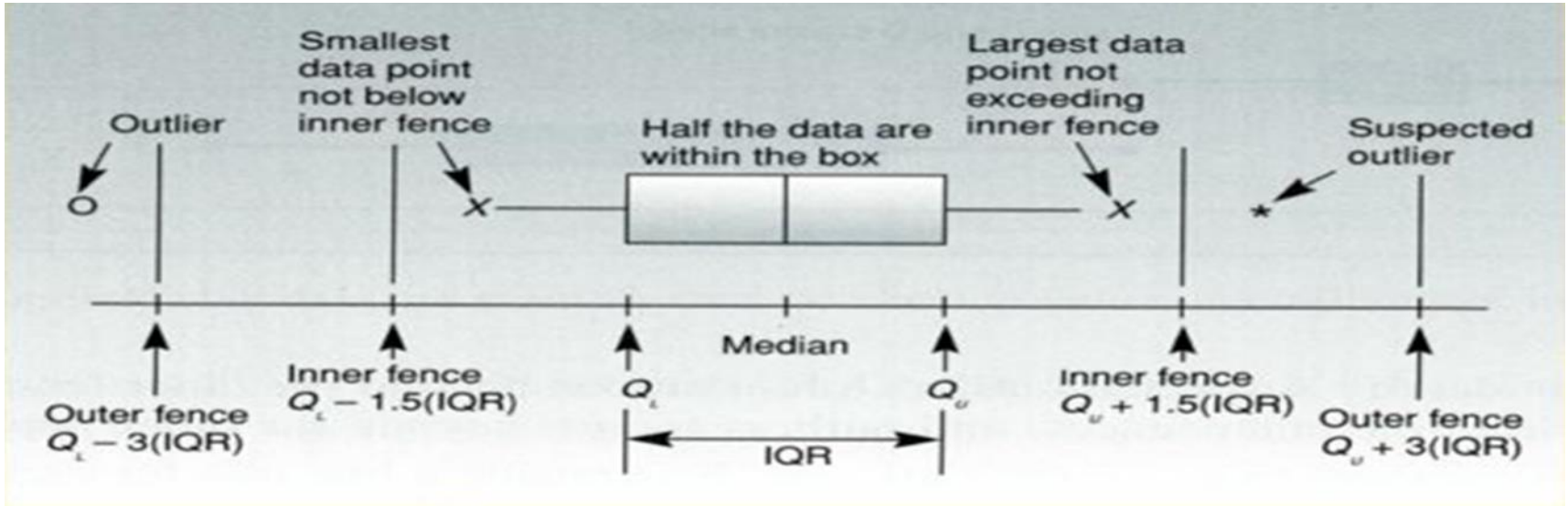
- **The Standard Deviation:**
  - Square root of variance i.e.
  - For population: $\sigma = \sqrt{\frac{\sum_i^N (x_i - \mu)^2}{N}}$
  - For sample: $s = \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{n-1}}$
- **Coefficient of Variance**:
  - $cv = \frac{s}{\bar{x}}$.

# 6.1 Measure of Dispersion: The Variance vs. Standard Deviation.

- Standard deviation is the square root of the variance and is expressed in the same units as the data set. Variance can be expressed in squared units.
- Standard deviation measures how far apart numbers are in a data set. Variance, on the other hand, gives an actual value to how much the numbers in a data set vary from the mean.
- CV can be used to measure and compare the variability between more than one dataset.(unit independent).

# 7. Descriptive Statistics: Summarizing Data.

# 7.1 Summarizing Data: Box Plot.



# 8.Bivariate Data.

# Motivation!!! What is Bivariate Data?

- Experiment Study-Graduate Rate and Student Related Expenditure.

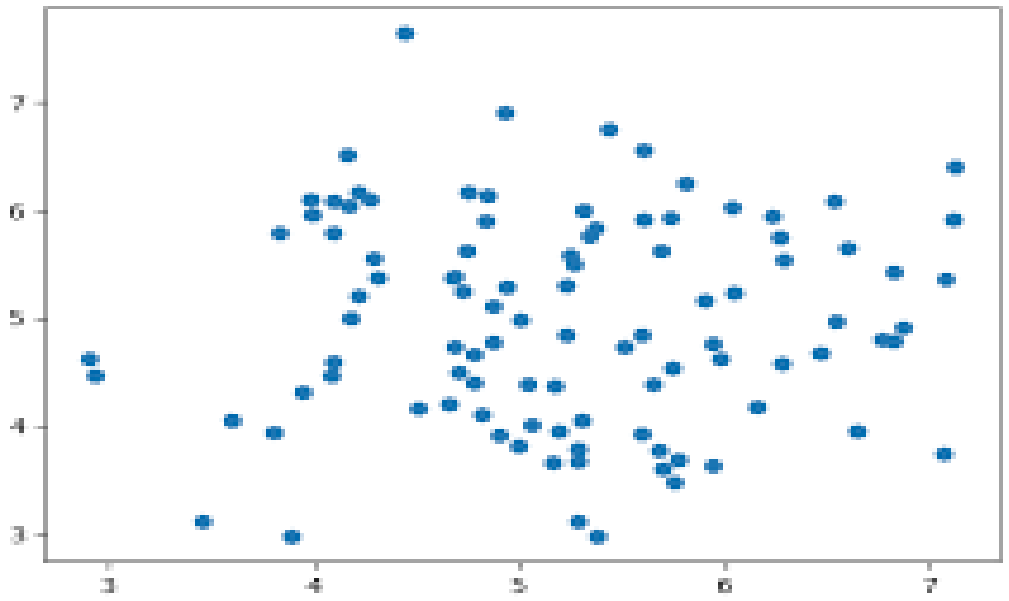
| Graduation Rate | Student Related Expenditure |
|---|---|
| 64.6 | 8011 |
| 53.0 | 7323 |
| 46.3 | 8735 |
| 38.5 | 7071 |

- Explain, Visualize and Summarize above data!!!

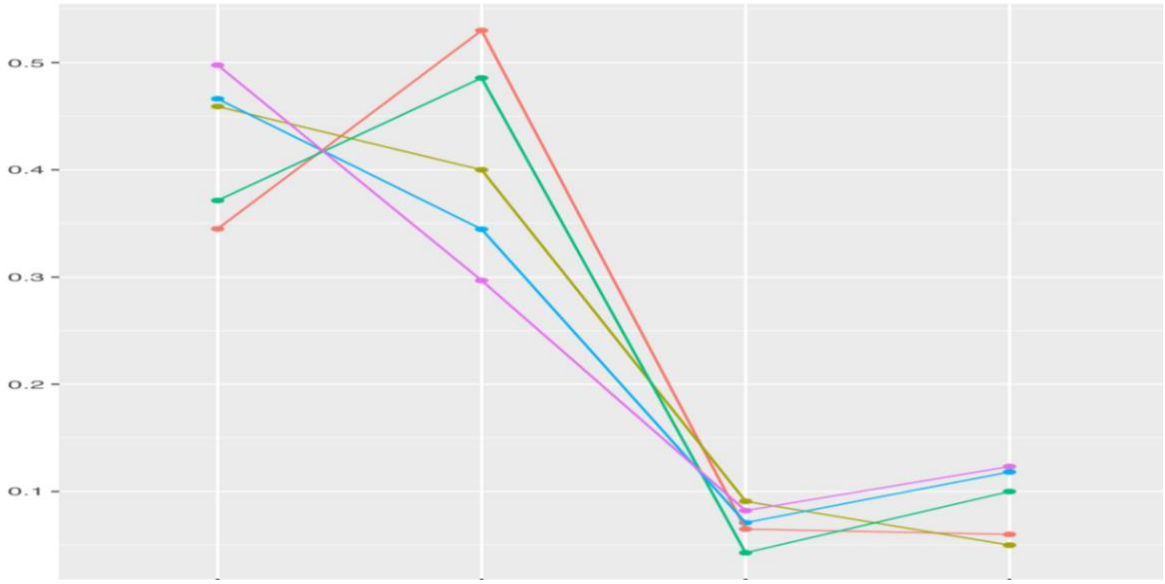
# 8.1 What is Bivariate Data?

- *"Bivariate data refers to a type of **statistical data** that involves the observation or measurement of two variables for each individual or element within a dataset."*
  - *chatgpt*
- These two variables are typically analyzed together to understand the ***relationship*** or ***association*** between them. {*cause-relationship*}
- bivariate data examines *how changes* in *one variable* are related to changes in *another variable*.

# 8.2 Graphical Displays : Bivariate Data.




Scatter Plot




Lines and Curves Plot

# 8.3 Graphical Displays : Common Relationships.

Scatter plot of bivariate numerical data gives a visual impression of how strongly x values and y values are related.

How to make precise statements and draw conclusions from data? Numerical assessment of the strength of relationship between the (x,y) pairs is required.

## Relationships:

a) Positive Linear Relationships. b) Another positive linear Relationships.

c) Negative Linear Relationships. d) No Relationships. e) Curved Relationships.

# 8.4 Correlation!!!

- A correlation coefficient (from co- and relation) is a numerical assessment of the strength of relationship between the x and y values in a set of (x, y) pairs.
- Some popular techniques for calculating correlation coefficients:
  - **Pearson Correlation Coefficient**
  - Spearman Rank Correlation Coefficient
  - Kendall's Tau

# 8.5 Pearson Correlation Coefficient

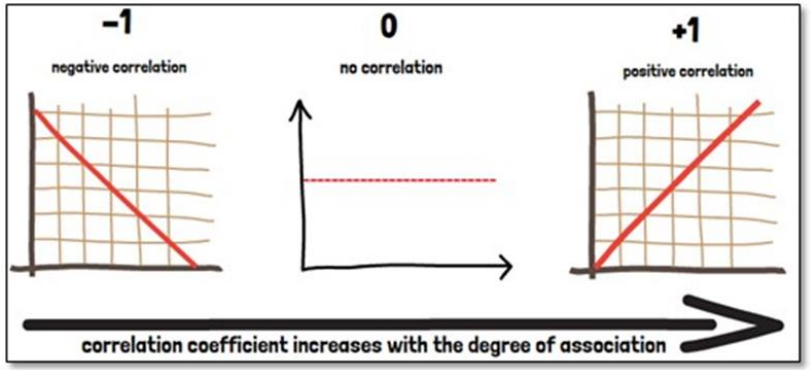
- For a pair of **sample Data** $\{(x_1, y_1), \dots (x_n, y_n)\} \in (X, Y)$ pairs, Pearson correlation coefficient "***r***" is given by:
  - $r_{x,y} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$
- An equivalent representation is :
  - $r_{x,y} = \frac{\sum z_x z_y}{n-1}$

Here:

- $z_{x,} = \frac{X_i - \bar{X}}{s_x}$ ; $z_{y,} = \frac{Y_i - \bar{Y}}{s_y}$

# 8.5 Pearson Correlation Coefficient: Properties.

- The value of "r" does not depend on the unit of measurement for either variable.
- The value of "r" does not depend on which of the two variables is considered x.
- The value of "r" is between **-1 and +1** . A value near the **upper limit: +1** indicates a **substantial positive relationship** and value **near lower limit -1** suggests a **substantial negative relationship**.
- The value of "r" **= 1** only when all the points in scatterplot of the data lie exactly on a **straight line** that **slopes upward**. Similarly, **"r" = -1** only when all the points lie exactly on a **downward** sloping line.
- The value of "r" is a **measure** of the **extent to which x and y are linearly related**. A value close to 0 does not rule out any strong relationship between x and y, there may exist a non linear relationship.



# 8.6 Caution: Correlation and Causation.

- Correlation **measures the extent of association**, but **association does not imply causation**.
- Example: A study suggest, among all elementary school children, the relationship between the number of cavities in child's teeth and the size of there vocabulary is strong and positive, this does not imply more cavities -> increase vocabulary size or vice versa.

# Thank You.

- Question!!!