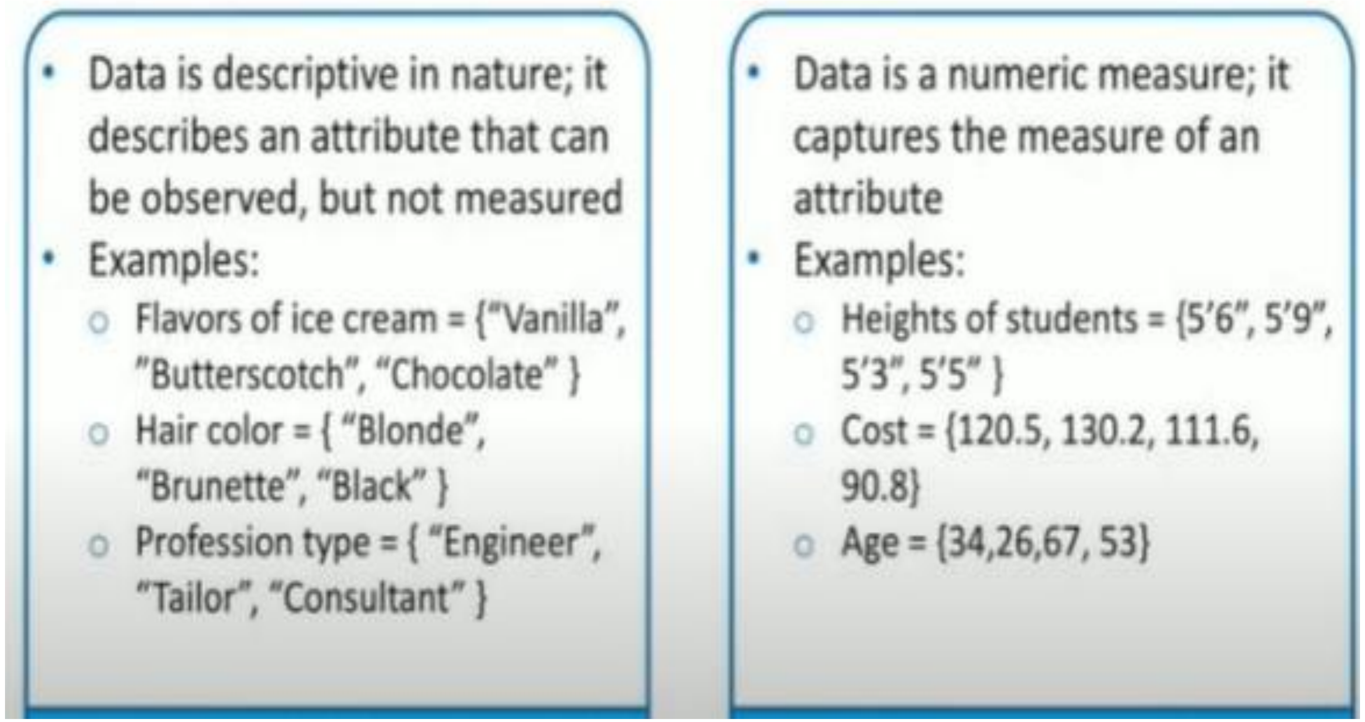
2. Managing Large Scale Data

Contents

- Types of Data and Data Representations,
- Acquire Data (E.G., Crawling),
- Process and Parse Data,
- Data Manipulation,
- Data Wrangling
- Data Cleaning.

Types of Data

- Data is a set of qualitative and quantitative



- Data is descriptive in nature; it describes an attribute that can be observed, but not measured
- Examples:
  - Flavors of ice cream = {"Vanilla", "Butterscotch", "Chocolate" }
  - Hair color = { "Blonde", "Brunette", "Black" }
  - Profession type = { "Engineer", "Tailor", "Consultant" }

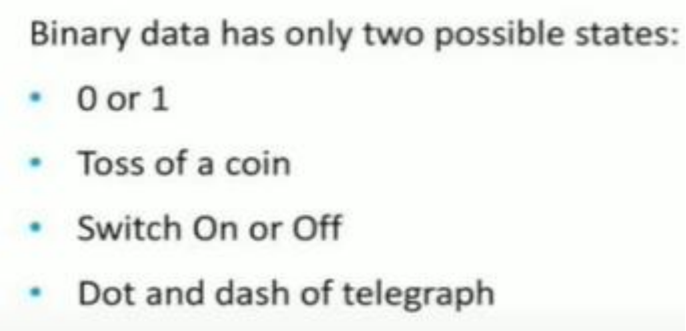
- Data is a numeric measure; it captures the measure of an attribute
- Examples:
  - Heights of students = {5'6", 5'9", 5'3", 5'5" }
  - Cost = {120.5, 130.2, 111.6, 90.8}
  - Age = {34,26,67, 53}

Types of Data

- Data is a set of qualitative and quantitative
- Quantitative Random Variable: Discrete and Continuous, Interval and Ratio
- Categorical Variable: Binary, Nominal, Ordinal

Types of Data

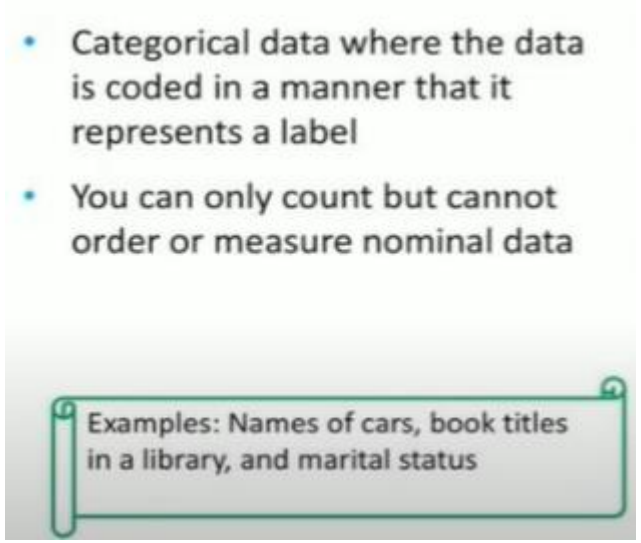
- Binary



Binary data has only two possible states:

- 0 or 1
- Toss of a coin
- Switch On or Off
- Dot and dash of telegraph

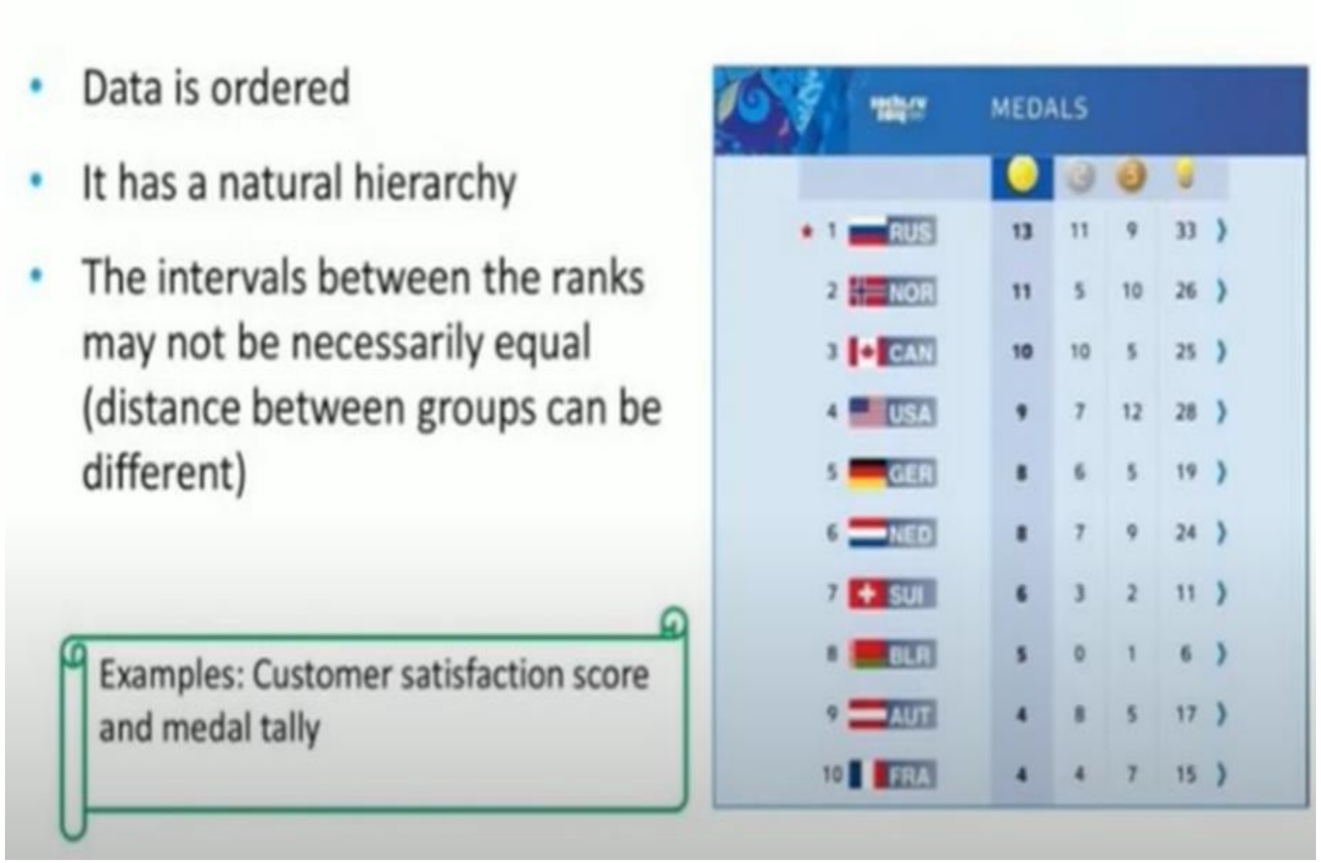
Nominal



- Categorical data where the data is coded in a manner that it represents a label
- You can only count but cannot order or measure nominal data

Examples: Names of cars, book titles in a library, and marital status

Ordinal Data



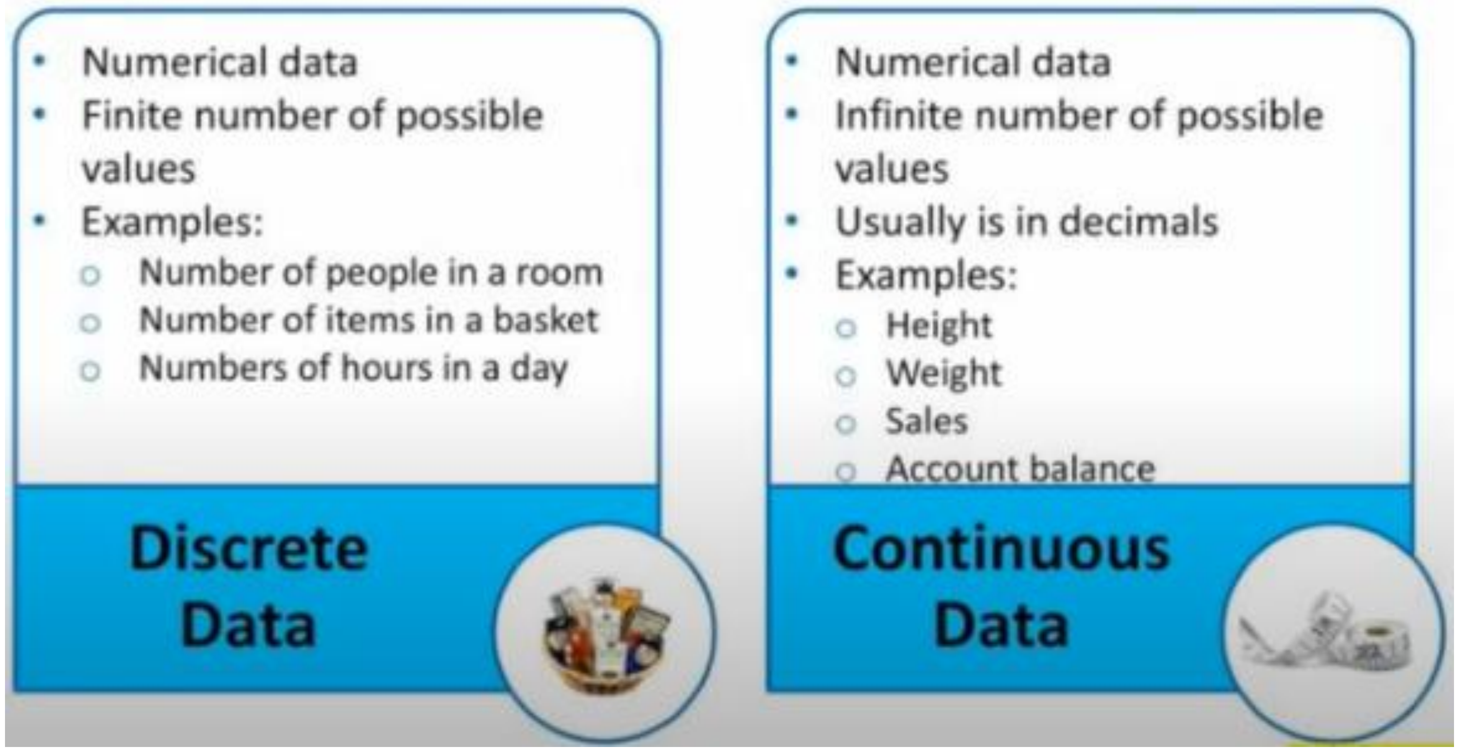
- Data is ordered
- It has a natural hierarchy
- The intervals between the ranks may not be necessarily equal (distance between groups can be different)

Examples: Customer satisfaction score and medal tally

MEDALS

★ 1	RUS	13	11	9	33	›
2	NOR	11	5	10	26	›
3	CAN	10	10	5	25	›
4	USA	9	7	12	28	›
5	GER	8	6	5	19	›
6	NED	8	7	9	24	›
7	SUI	6	3	2	11	›
8	BLR	5	0	1	6	›
9	AUT	4	8	5	17	›
10	FRA	4	4	7	15	›

Discrete and Continuous Data



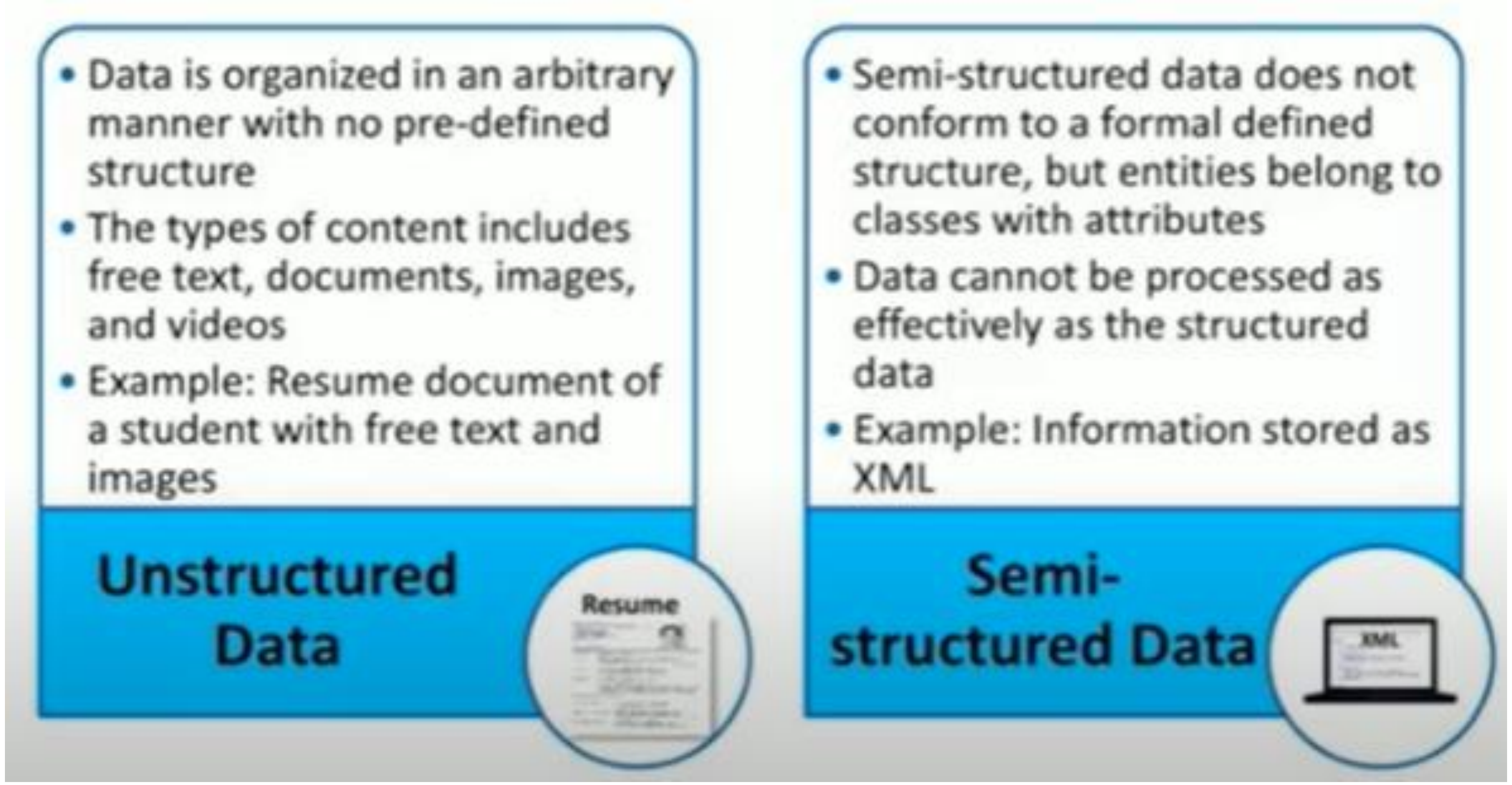
- Numerical data
- Finite number of possible values
- Examples:
  - Number of people in a room
  - Number of items in a basket
  - Numbers of hours in a day

Discrete Data

- Numerical data
- Infinite number of possible values
- Usually is in decimals
- Examples:
  - Height
  - Weight
  - Sales
  - Account balance

Continuous Data

Forms of Data: Structured, Semi structured, Unstructured



- Data is organized in an arbitrary manner with no pre-defined structure
- The types of content includes free text, documents, images, and videos
- Example: Resume document of a student with free text and images

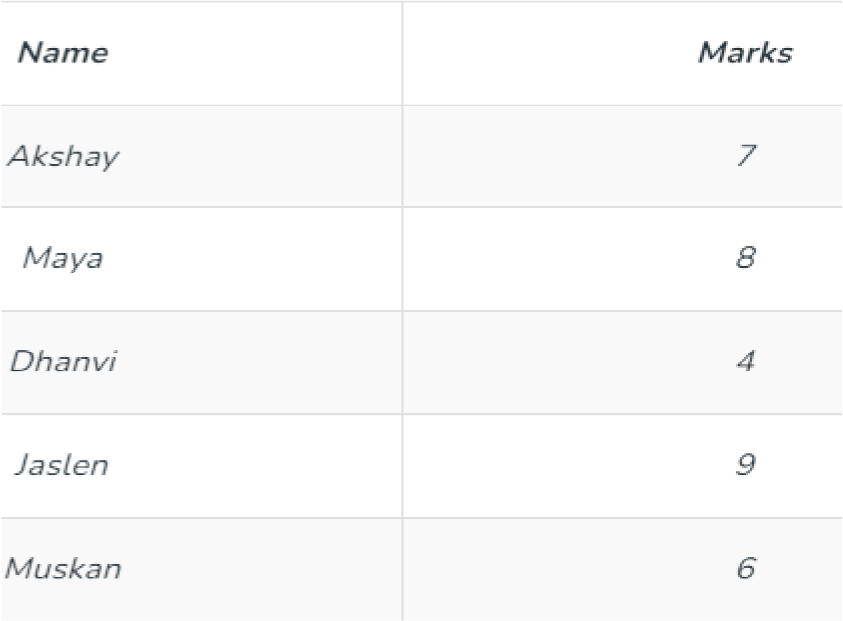
Unstructured Data

- Semi-structured data does not conform to a formal defined structure, but entities belong to classes with attributes
- Data cannot be processed as effectively as the structured data
- Example: Information stored as XML

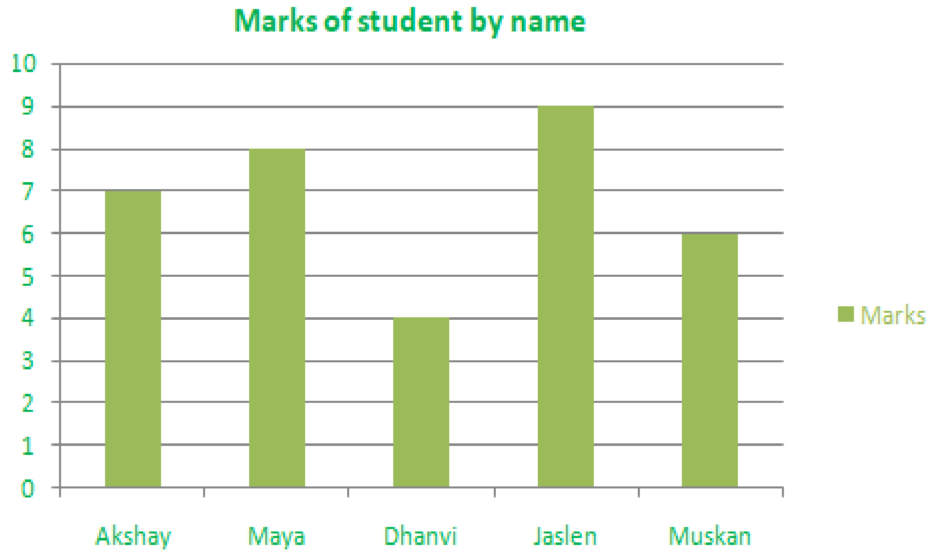
Semi-structured Data

Data Representations

- **Bar Chart**
- Bar chart helps us to represent the collected data visually.
- The collected data can be visualized horizontally or vertically in a bar chart like amounts and frequency. It can be grouped or single

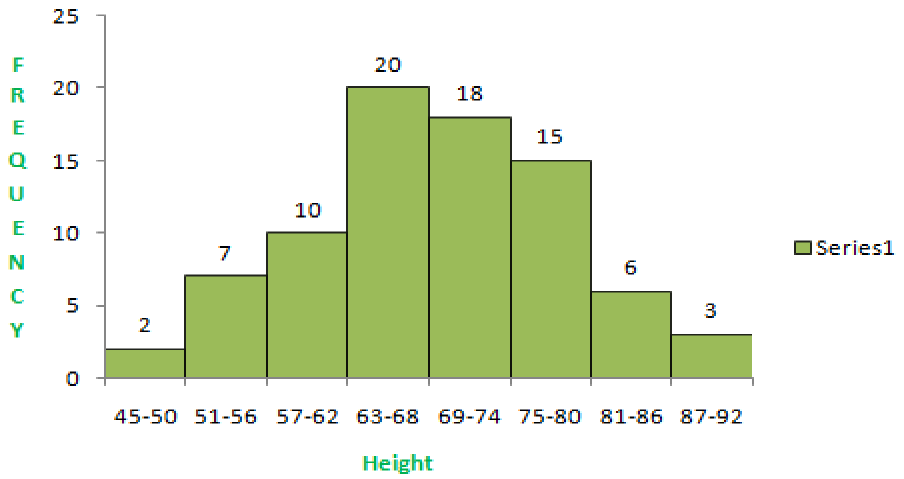


Name	*Marks*
Akshay	*7*
Maya	*8*
Dhanvi	*4*
Jaslen	*9*
Muskan	*6*



Histogram

- A histogram is the graphical representation of data. It is similar to the appearance of a bar graph but there is a lot of difference between histogram and bar graph because a bar graph helps to measure the frequency of categorical data.
- A categorical data means it is based on two or more categories like gender, months, etc. Whereas histogram is used for quantitative data.

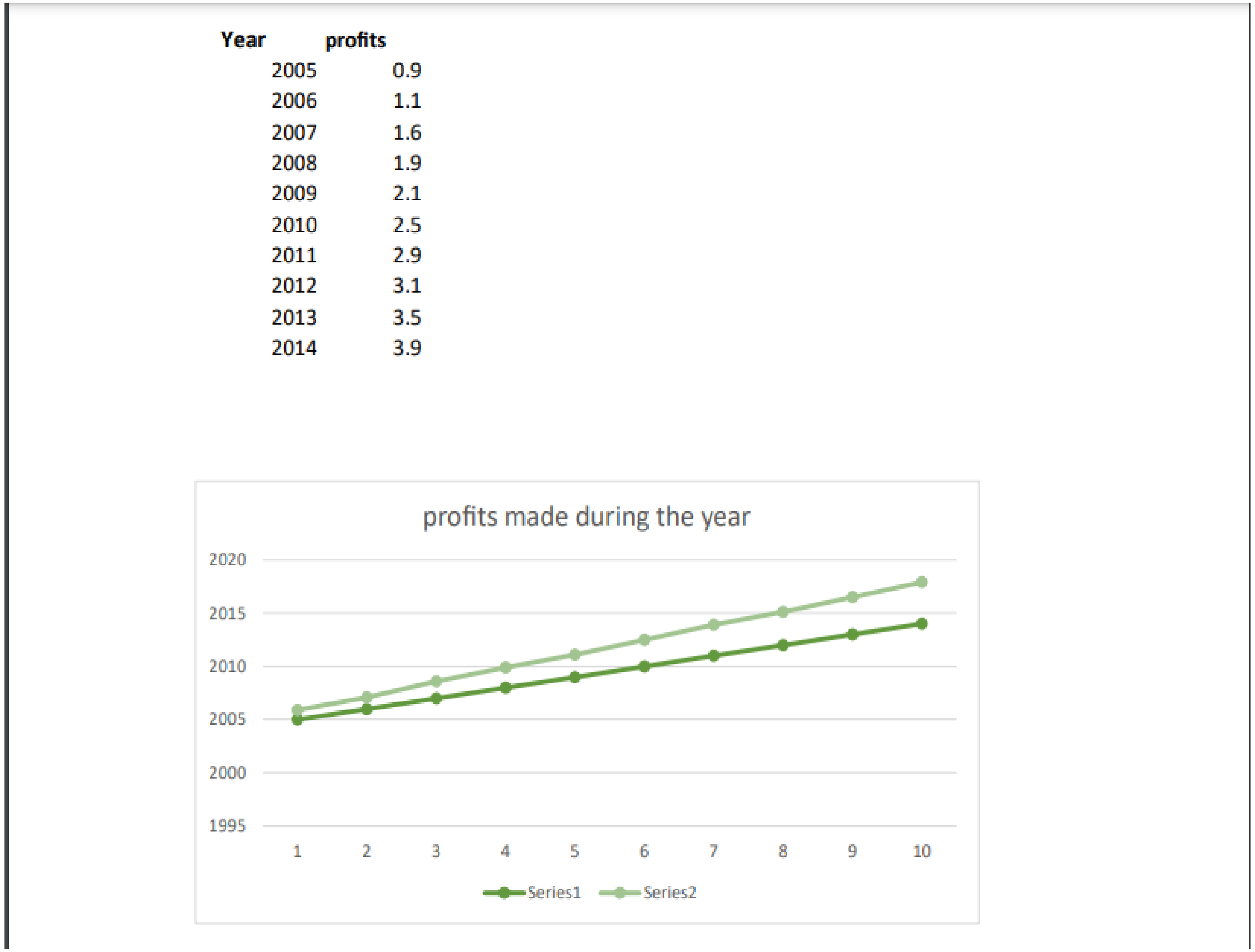


Line Graph

- The graph which uses lines and points to present the change in time is known as a line graph. Line graphs can be based on the number of animals left on earth, the increasing population of the world day by day, or the increasing or decreasing the number of bitcoins day by day, etc.
- The line graphs tell us about the changes occurring across the world over time. In a line graph, we can tell about two or more types of changes occurring around the world.

Line Graph

- **Line Graph**

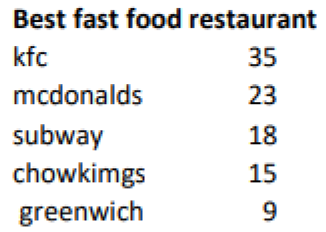


Year	profits
2005	0.9
2006	1.1
2007	1.6
2008	1.9
2009	2.1
2010	2.5
2011	2.9
2012	3.1
2013	3.5
2014	3.9

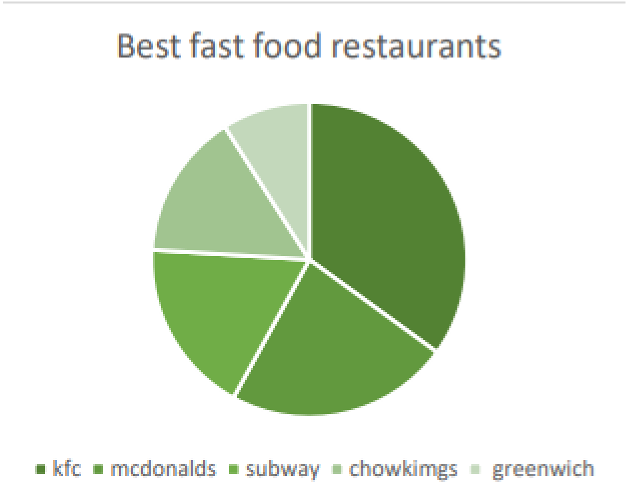
profits made during the year

Pie Chart

- Pie chart is a type of graph that involves a structural graphic representation of numerical proportion. It can be replaced in most cases by other plots like a bar chart, dot plot, etc.
- As per the research, it is shown that it is difficu different sections of a given pie chart, or if it is across different pie charts.



Best fast food restaurant	
kfc	35
mcdonalds	23
subway	18
chowkimgs	15
greenwich	9



Best fast food restaurants

Scatter Plot

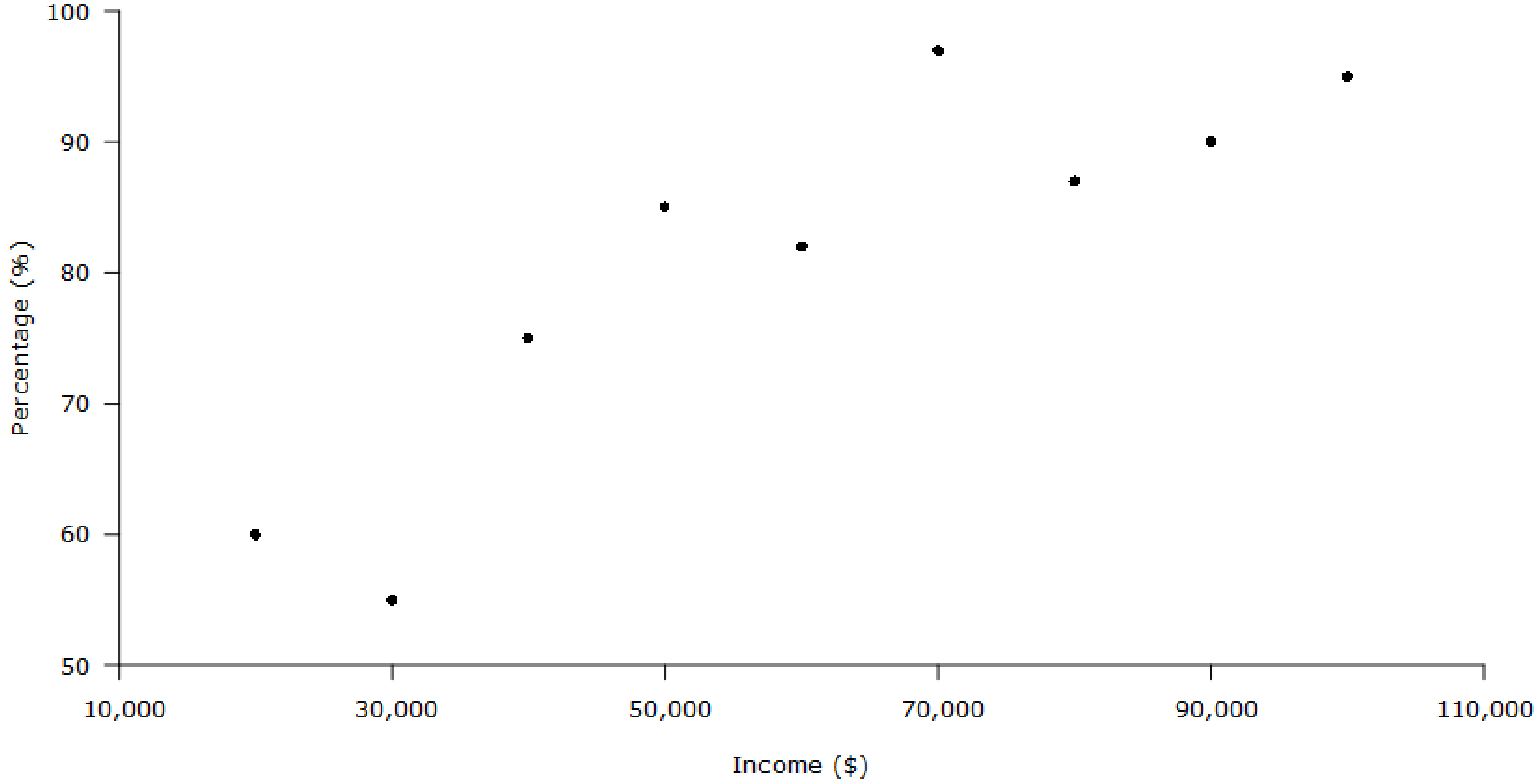
- In science, the scatterplot is widely used to present measurements of two or more related variables.

- It is particularly useful when the values of the variables of the y-axis are thought to be dependent upon the values of the variable of the x-axis.

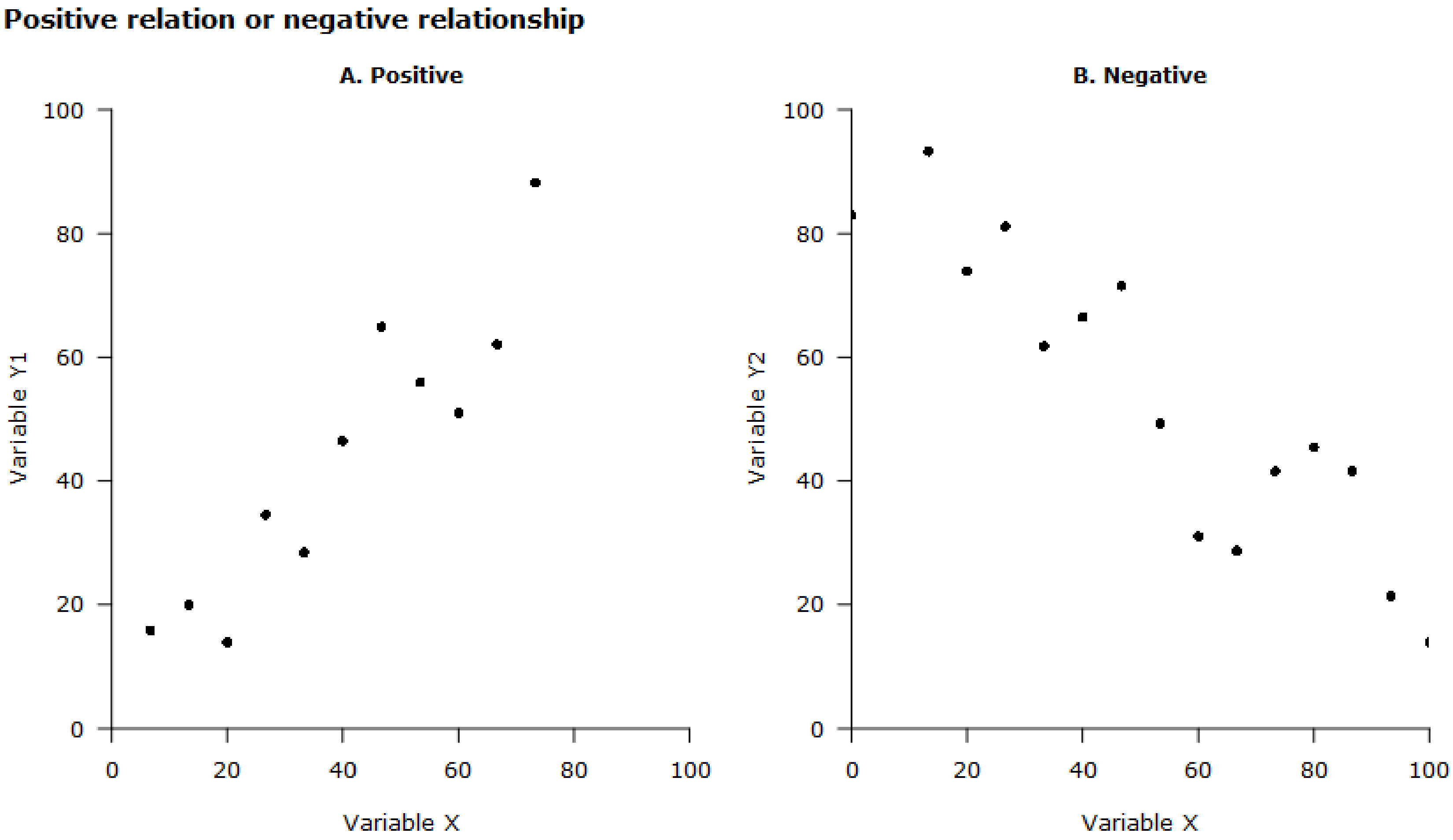
- Example: Car ownership increases as the household income increases, showing that there is a positive relationship between these two variables.

Scatter Plot

Income ($)	Percentage (%)
20,000	60
30,000	55
40,000	75
50,000	85
60,000	82
70,000	97
80,000	87
90,000	90
100,000	95

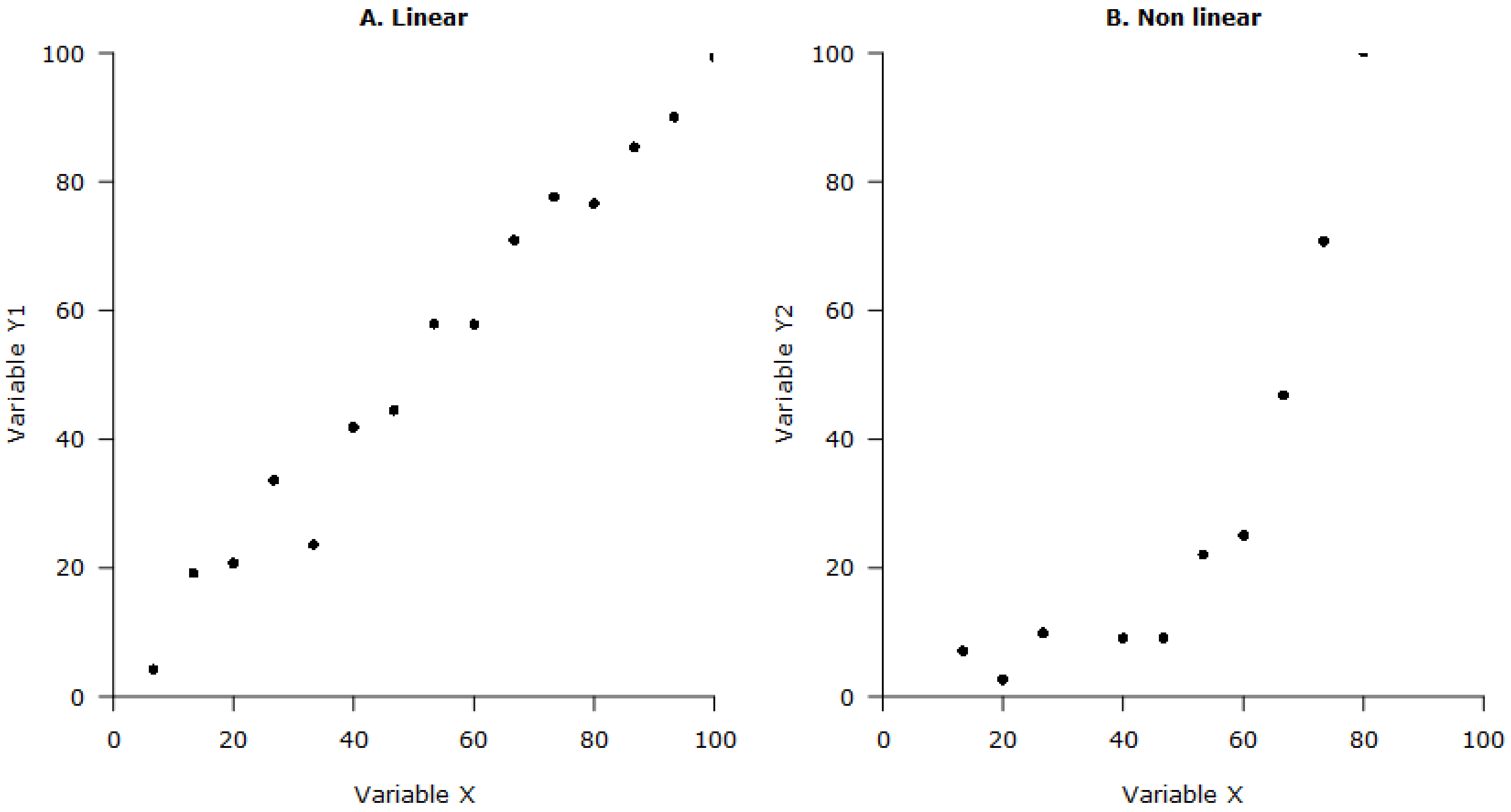


Scatter Plot



Scatter Plot

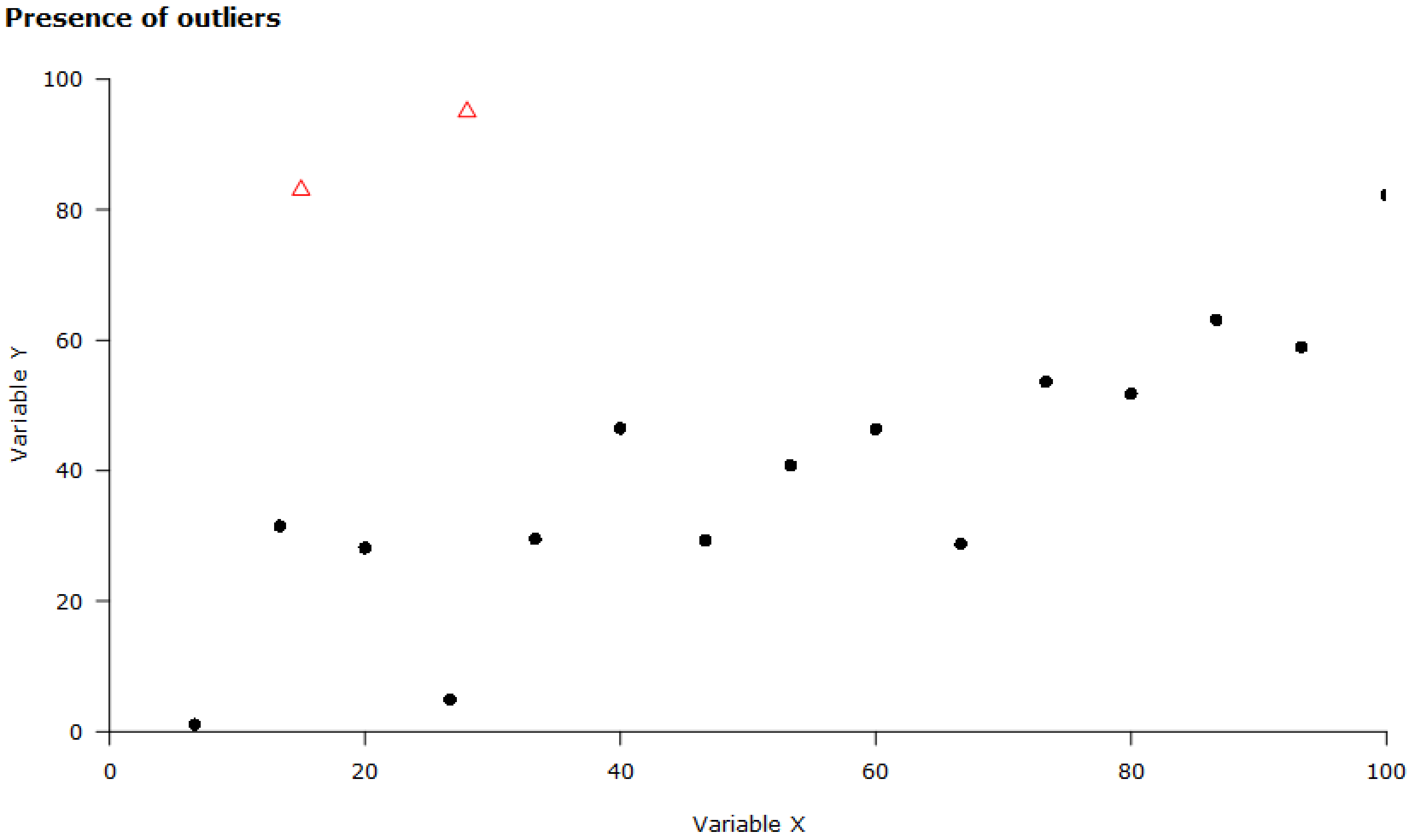
Linear relation or non linear relationship



A. Linear

B. Non linear

Scatter Plot

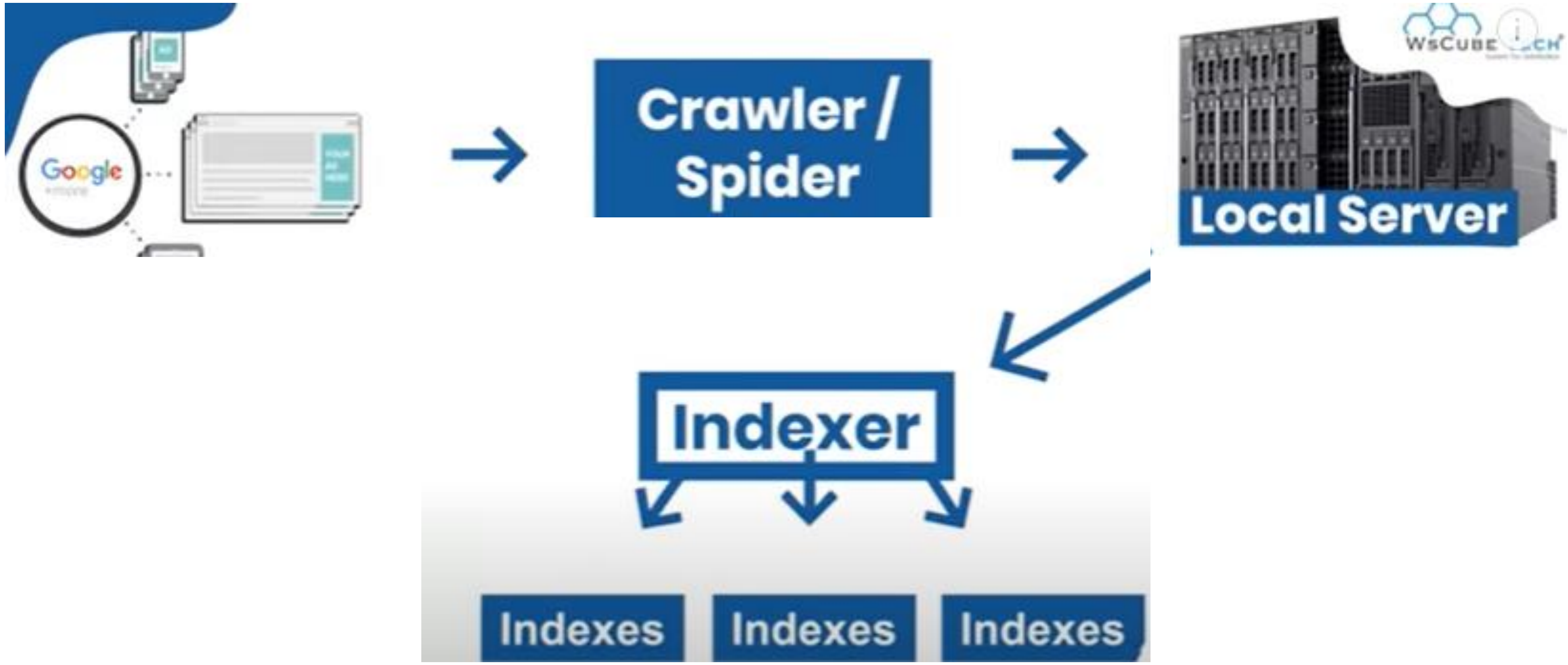


How Do Data Scientists Collect Data?

- **Use Existing Datasets**
  - **Use public datasets**: There are numerous datasets on the internet to be used as a benchmark for general computer science problems
  - **Purchase datasets**: There are various online platforms and marketplaces where you can buy datasets such as environmental data, political data, customer data, etc.
  - **Company's datasets**: Companies can easily access their own data stack.

- **Create a new Datasets:**
  - **Create data manually**: Data scientists can manually create online surveys to gather results. Or, they can use old surveys and their results or pay employees to perform manual tasks of data classification and data labeling.
  - **Convert existing data into a dataset**: Another great way to gather data from the internet is by crawling websites and downloading public data. This can be done via dedicated web crawlers or manually through RPA bots that are programmed for web crawling.

Web Crawler



- Crawler, Indexer and Page ranking algorithm.

Acquire Data: Web Crawling

- Web crawling is the technique used to collect a huge amount of data from different websites and learn what every webpage on the website is all about. The collected data can help you to retrieve specific information that you need.
- A web crawler is typically operated by search engines such as Google, Bing, and Yahoo. The goal is to index the content of different websites all over the internet so that they can appear on the search engine result whenever a person tries to find something on the web.
- Web crawlers can receive a search query and apply a search algorithm to search and provide relevant information in response to the search queries by using search engines.

Web Crawling

- Ever wondered how a giant search engine like Google collects data to display in the search engine results pages? Does it use a web crawler to retrieve data faster?
- A Web crawler, also known as a web robot, a web spider or a spider bot, is an automated script or program that logically browses the internet. **This automated process of indexing data on web pages is known as web crawling or spidering**.
- Search engines such as Bing, Google, and etc. use web crawlers to provide up-to-date information in SERPs.

Web Crawler- Use Cases

- Many companies rely on a web crawler to collect data about their customers, products, and services on the web.
- Data science project starts by formulating the business problem to solve and then followed by the second stage of collecting the right data to solve that problem.
- In this stage, you can use web crawlers to collect the data on the internet that you need for your data science project.

Use Cases

1. Collect Social Media Data for Sentiment Analysis

- Many companies use web crawling to collect posts and comments on various social media platforms such as Facebook, X and Instagram. Companies use the collected data to assess how their brand is performing and discover how their products or services are reviewed by their customers, it can be a positive review, negative review or neutral.

2. Collect Financial Data for Stock Prices Forecasting

- The stock market is full of uncertainty, therefore stock price forecasting is very important in business. Web crawling is used to collect stock prices data from different platforms for different periods (for example 54 weeks, 24 months e.t.c).

No-Code Web Crawling Tools

- Octoparse is a visual software tool that you can use to extract different types of data from the web without writing codes. It also has various features that make it **easier to collect data within a short period**.
- Parsehub is another easy-to-learn **visual web crawling tool** that is simple, friendly to use, powerful and flexible to extract data from the web. **It offers an easy-to-use interface to set your run and automatically extract millions of data points from any website in minutes**.
- Webscraper is a web crawling tool that does not require you to write code and it runs within the browser as an extension. **You can use this tool to collect data from the web on an hourly, daily, or weekly basis**. It can also automatically export data to Dropbox, Google sheets, or Amazon S3.

Process and Parse Data

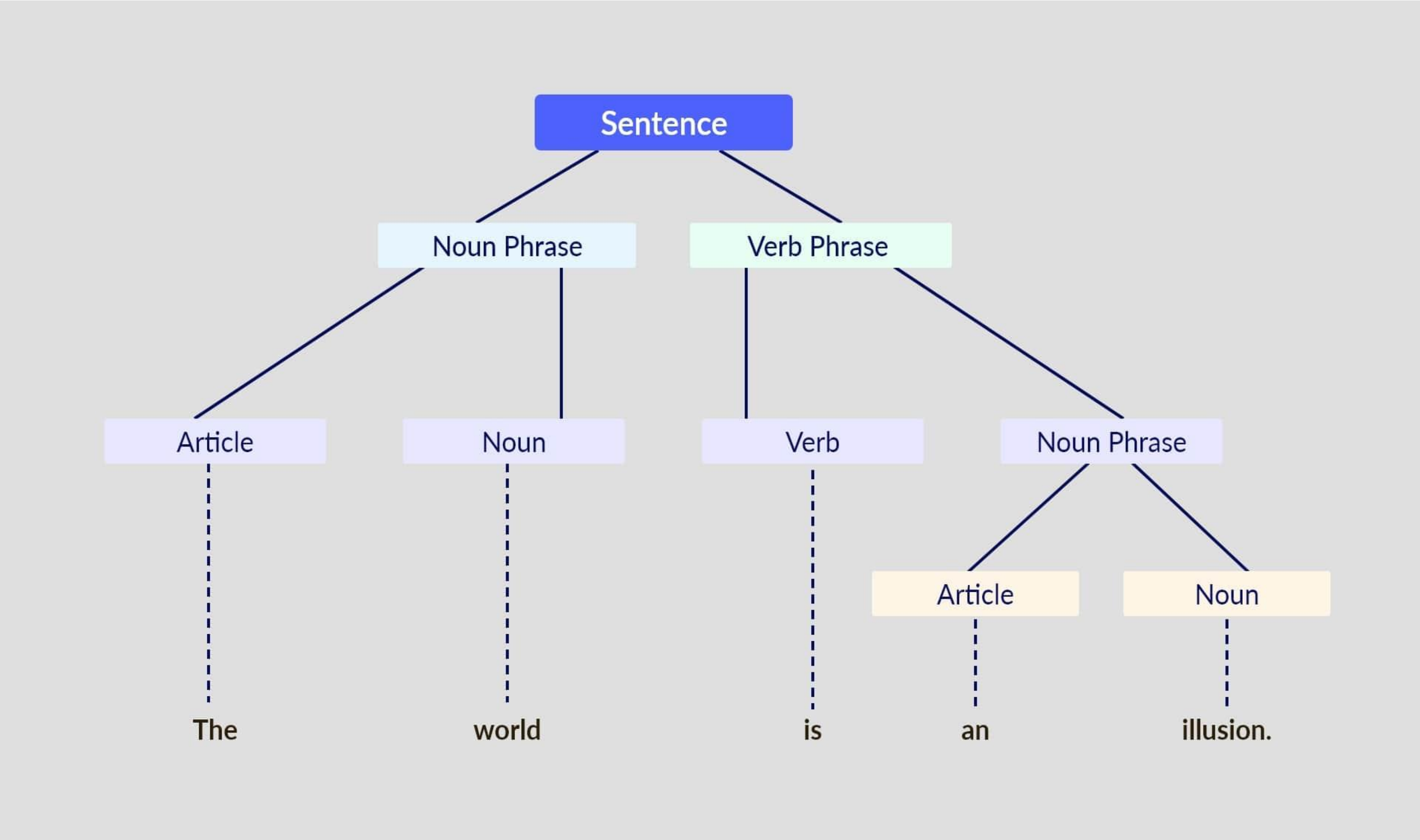
- An important aspect of parsing is to capture information from data in a way that it fits contextual structures.

- Data parsing is used for crawling information from large datasets and structuring it in a way humans can understand. Traditional data parsing is done on HTML files where the parser converts HTML text into readable data. Data parse program is used for converting unstructured data into JSON, CSV, and other file formats and adds structure to said information.

- However, not all parsers work the same and there are distinct differences in parsing technologies.

- There are numerous benefits of data parsing for businesses ranging from automated data extraction, improved visibility, cutting costs, and boosting employee productivity.

Parser


Sentence
Noun Phrase
Verb Phrase
Article
Noun
Verb
Noun Phrase
Article
Noun
The
world
is
an
illusion.

Data Manipulation

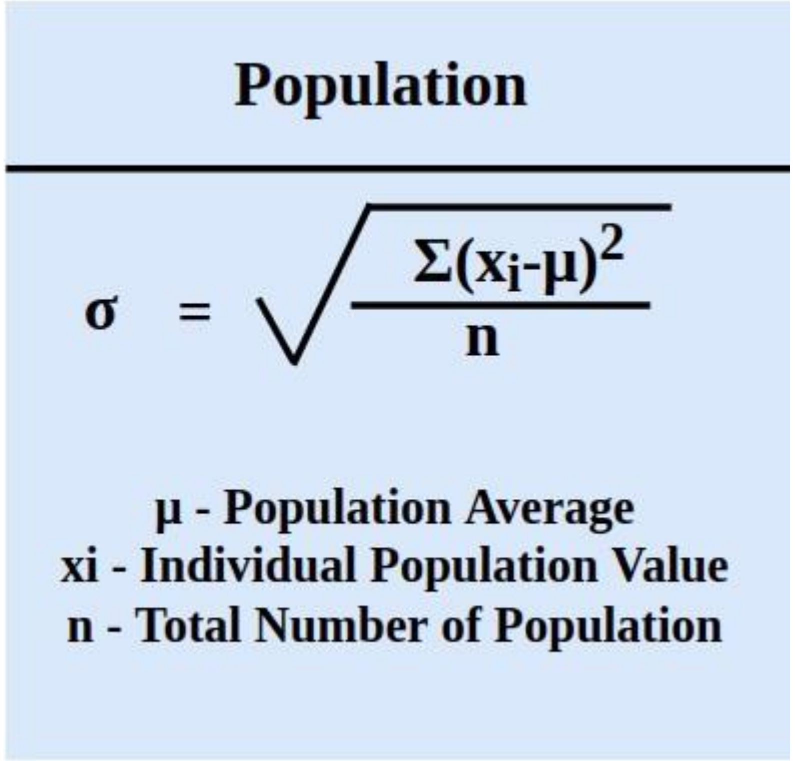
- Data manipulation refers to the process of adjusting data to make it organised and easier to read.

- Data manipulation language, or DML, is a programming language that adjusts data by inserting, deleting and modifying data in a database such as to cleanse or map the data. SQL, or Structured Query Language, is a language that communicates with databases. When using SQL- data change statements for data manipulation, four functions can occur, namely:
- Select
- Update
- Insert
- Delete

Data Manipulation

- These commands tell the database where to select data from and what to do with it.
- Here's how it works:
- **SELECT:** The select statement allows users to pull a selection from the database to work with. You tell the computer what to SELECT and FROM where.
- **UPDATE:** To change data that already exists, you will use the UPDATE statement. You can tell the database to update certain sets of information and the new information that should be input, either with single records or multiple records at a time.
- **INSERT:** You can move data from one location to another by using the INSERT statement.
- **DELETE:** To get rid of existing records within a table, you use the DELETE statement. You tell the system where to delete from and what files to get rid of.
- Since SQL does not allow you to import or export data from outside sources, some providers can store data and give you the tools to manipulate data for your business needs.

Standard Deviation

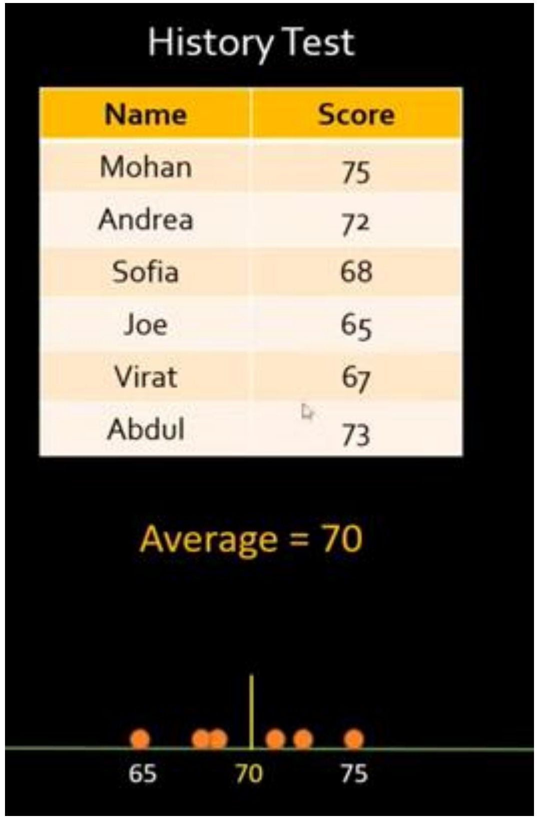


Population

$$\sigma = \sqrt{\frac{\Sigma(x_i-\mu)^2}{n}}$$

μ - Population Average
xi - Individual Population Value
n - Total Number of Population

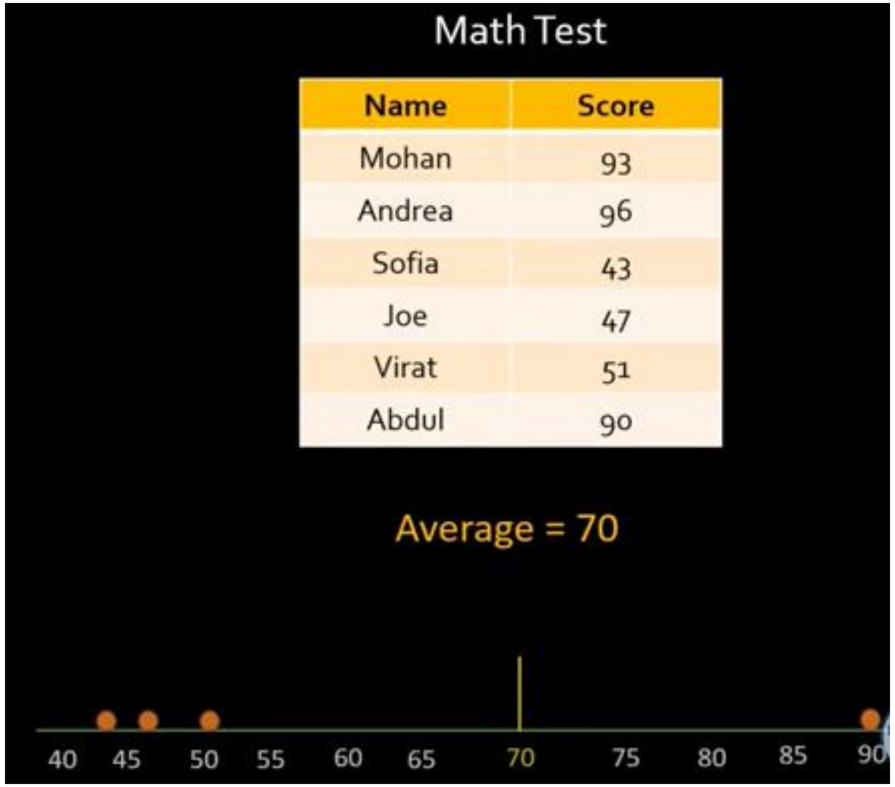
Standard Deviation



History Test

Name	Score
Mohan	75
Andrea	72
Sofia	68
Joe	65
Virat	67
Abdul	73

Average = 70

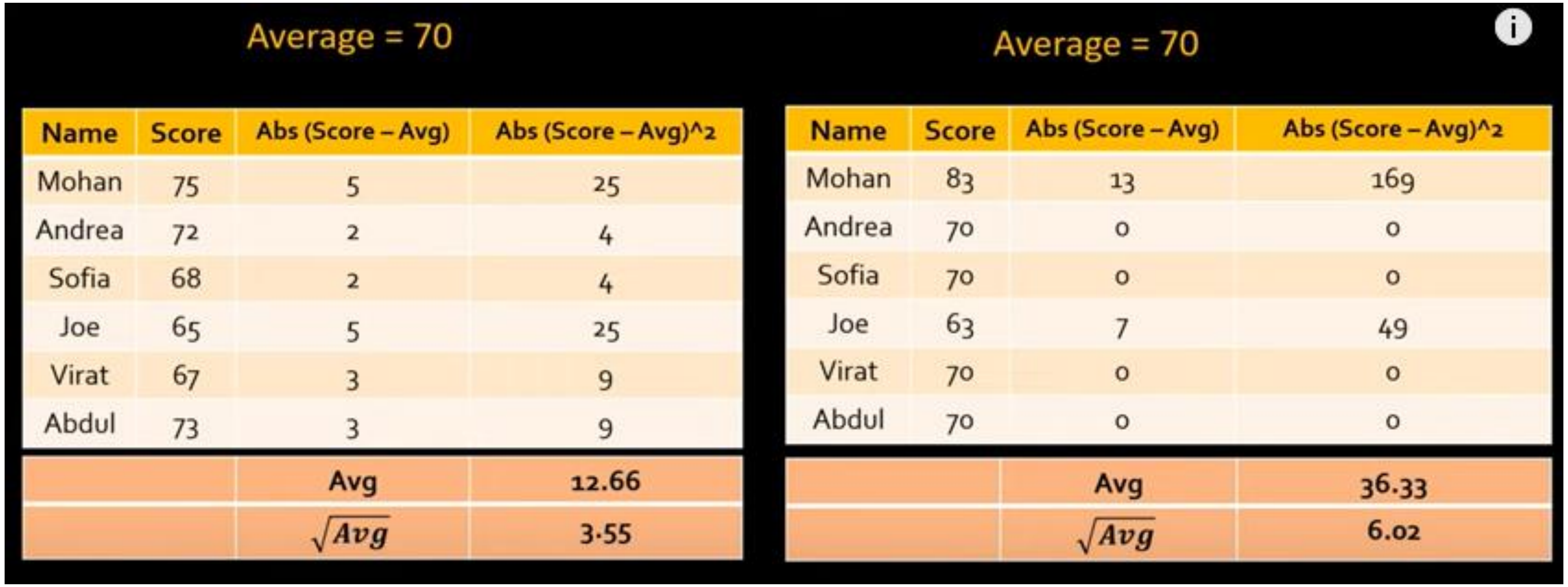


Math Test

Name	Score
Mohan	93
Andrea	96
Sofia	43
Joe	47
Virat	51
Abdul	90

Average = 70

Standard Deviations



Average = 70

Name	Score	Abs (Score – Avg)	Abs (Score – Avg)^2
Mohan	75	5	25
Andrea	72	2	4
Sofia	68	2	4
Joe	65	5	25
Virat	67	3	9
Abdul	73	3	9
		Avg	12.66
		$\sqrt{Avg}$	3.55

Average = 70

Name	Score	Abs (Score – Avg)	Abs (Score – Avg)^2
Mohan	83	13	169
Andrea	70	0	0
Sofia	70	0	0
Joe	63	7	49
Virat	70	0	0
Abdul	70	0	0
		Avg	36.33
		$\sqrt{Avg}$	6.02

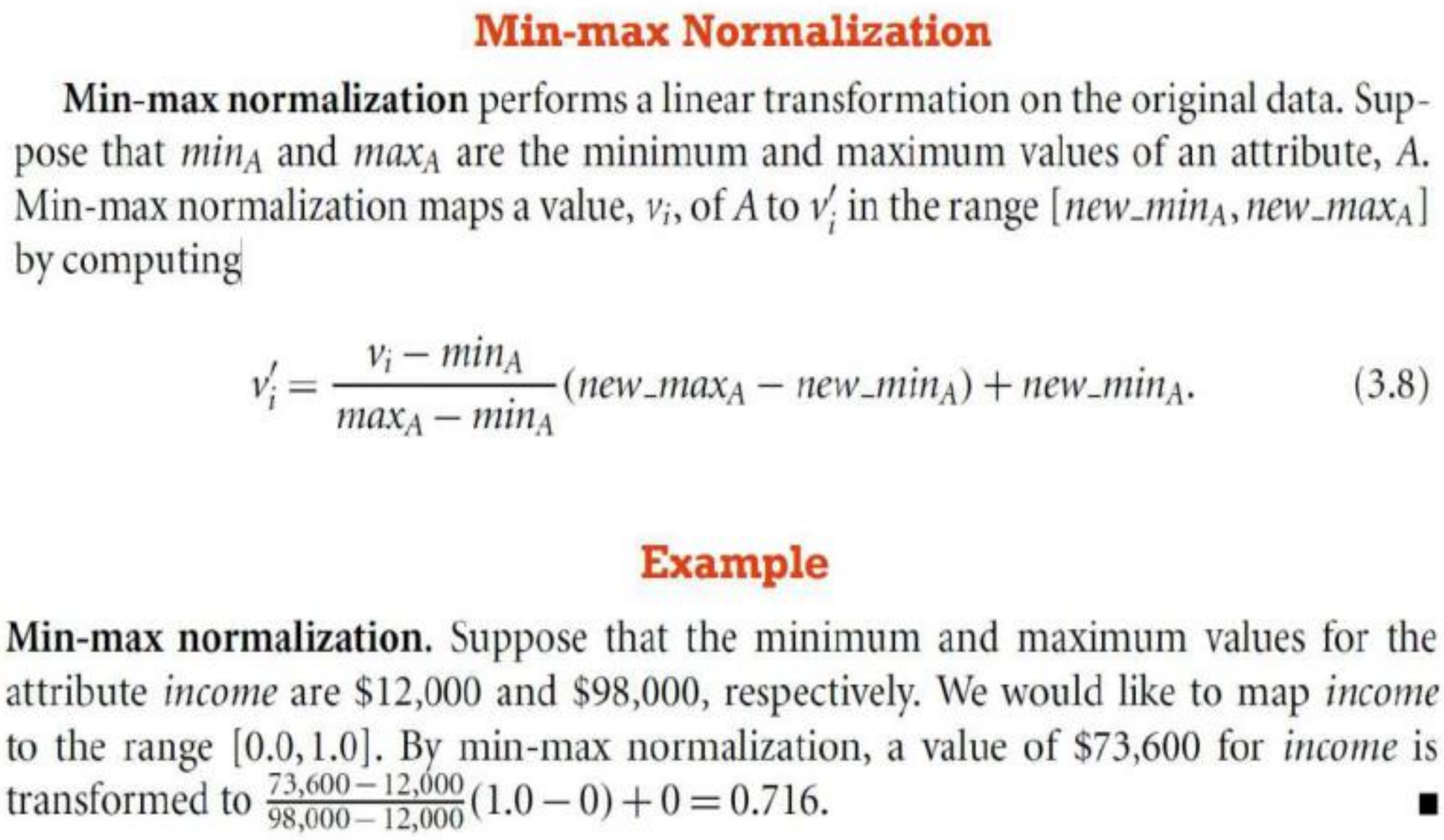
Data Transformation

- Normalization Methods
- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values
- Normalization is a technique often applied as part of data preparation for data science. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. Every dataset does not require normalization. It is required only when features have different ranges.

Normalization

- For example, consider a data set containing two features, age(x1), and income(x2). Where age ranges from 0–100, while income ranges from 0–20,000 and higher. Income is about 1,000 times larger than age and ranges from 20,000–500,000.
- So, these two features are in very different ranges. When we do further analysis, some time, for example, the attributed income will intrinsically influence the result more due to its larger value. But this doesn’t necessarily mean it is more important as a predictor.

Normalization Methods



Min-max Normalization

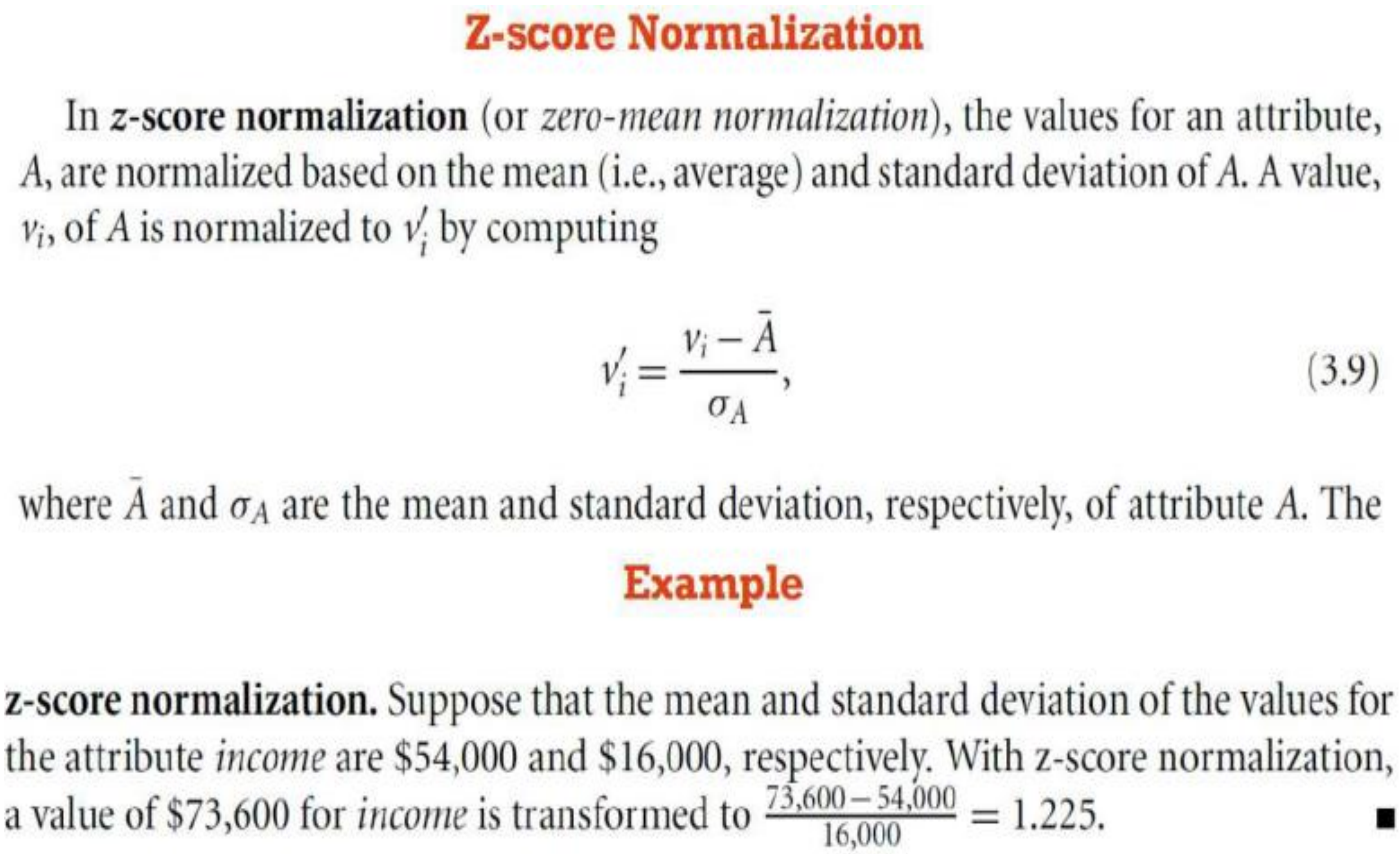
Min-max normalization performs a linear transformation on the original data. Suppose that min_A and max_A are the minimum and maximum values of an attribute, A. Min-max normalization maps a value, v_i, of A to v_i' in the range $[new\_min_A, new\_max_A]$ by computing

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A. \qquad (3.8)$$

Example

Min-max normalization. Suppose that the minimum and maximum values for the attribute *income* are \$12,000 and \$98,000, respectively. We would like to map *income* to the range $[0.0, 1.0]$. By min-max normalization, a value of \$73,600 for *income* is transformed to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$. ■

Normalization Methods



Z-score Normalization

In **z-score normalization** (or *zero-mean normalization*), the values for an attribute, *A*, are normalized based on the mean (i.e., average) and standard deviation of *A*. A value, v_i, of *A* is normalized to v_i' by computing

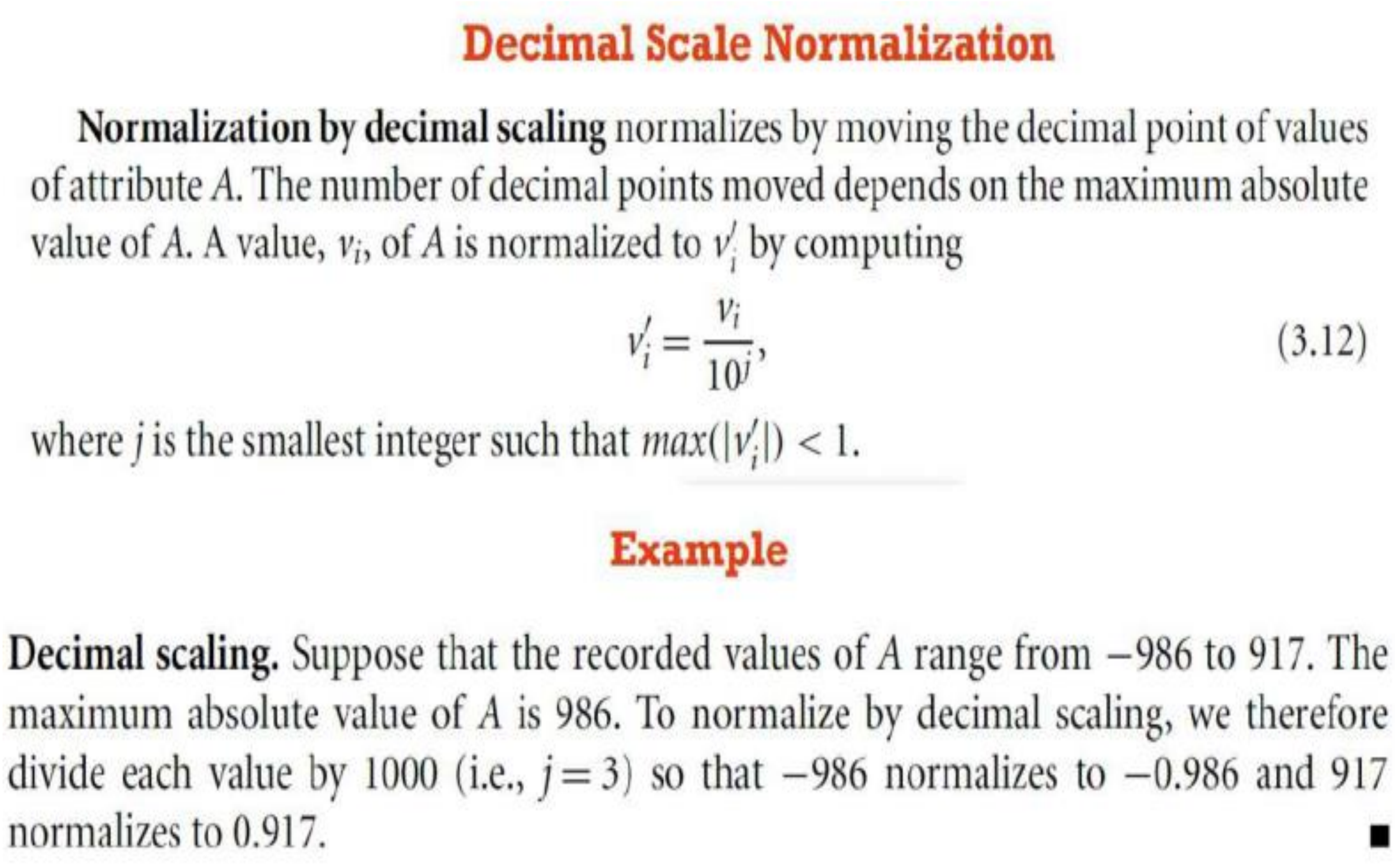
$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}, \tag{3.9}$$

where $\bar{A}$ and σ_A are the mean and standard deviation, respectively, of attribute *A*. The

Example

z-score normalization. Suppose that the mean and standard deviation of the values for the attribute *income* are \$54,000 and \$16,000, respectively. With z-score normalization, a value of \$73,600 for *income* is transformed to $\frac{73,600 - 54,000}{16,000} = 1.225$. ■

Normalization Methods



Decimal Scale Normalization

Normalization by decimal scaling normalizes by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A. A value, v_i, of A is normalized to v_i' by computing

$$v_i' = \frac{v_i}{10^j}, \tag{3.12}$$

where j is the smallest integer such that $max(|v_i'|) < 1$.

Example

Decimal scaling. Suppose that the recorded values of A range from -986 to 917. The maximum absolute value of A is 986. To normalize by decimal scaling, we therefore divide each value by 1000 (i.e., $j = 3$) so that -986 normalizes to -0.986 and 917 normalizes to 0.917. ■