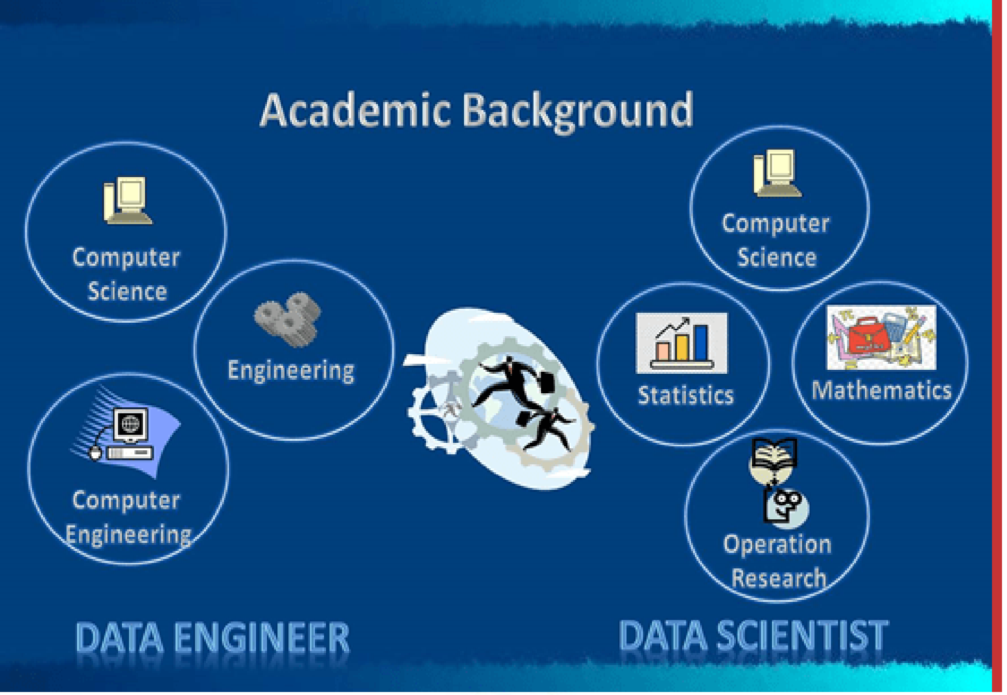
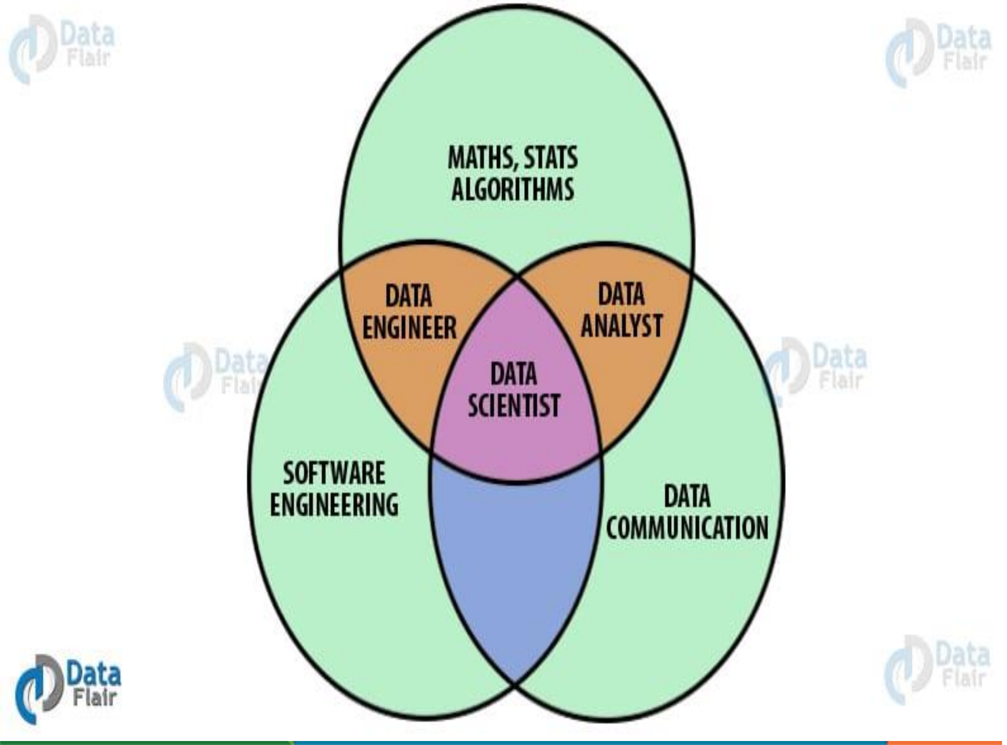
What is Data Engineering?

- Terminology used for collecting and validating quality data so that it can be used by data scientists.
- Data infrastructure, Data mining, Data crunching, Data acquisition, Data modelling, Data management
- Practice of designing and building systems for collecting, storing and analyzing data at scale
- Set of operations to make data available and usable to data scientists, data analyst, business intelligence (BI) developers and other specialists
- Data engineering requires dedicated experts i.e. Data Engineers to design and build systems for gathering and storing data at scale as well as preparing it for further analysis

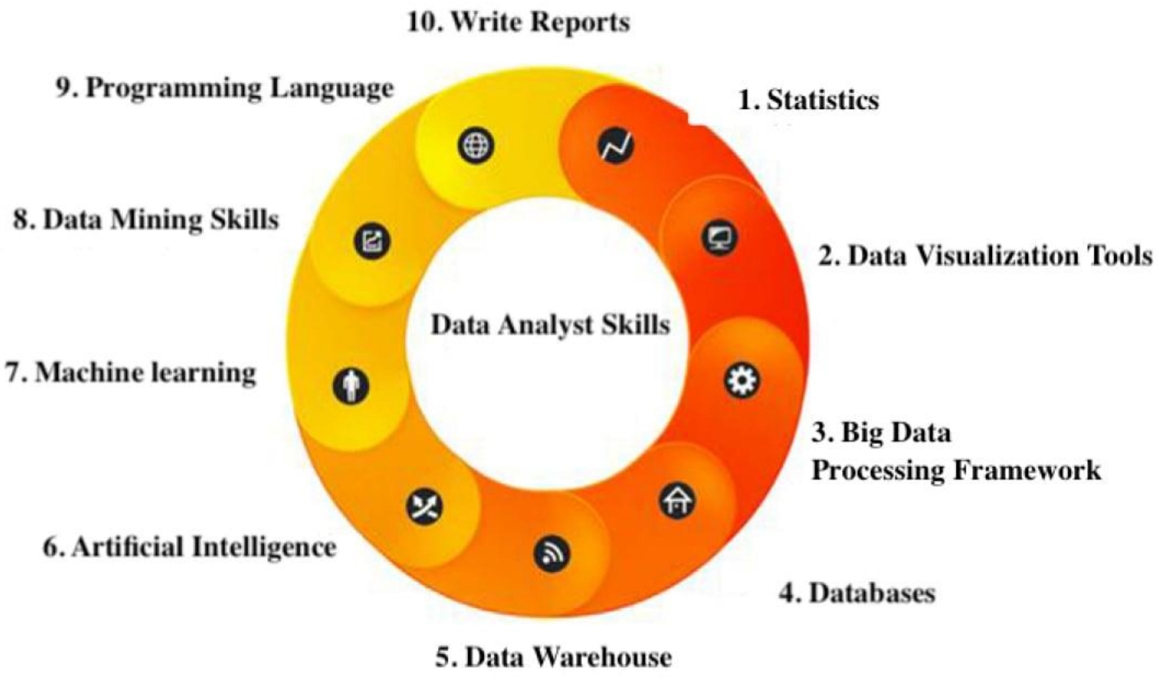
Data Engineer vs. Data Scientist

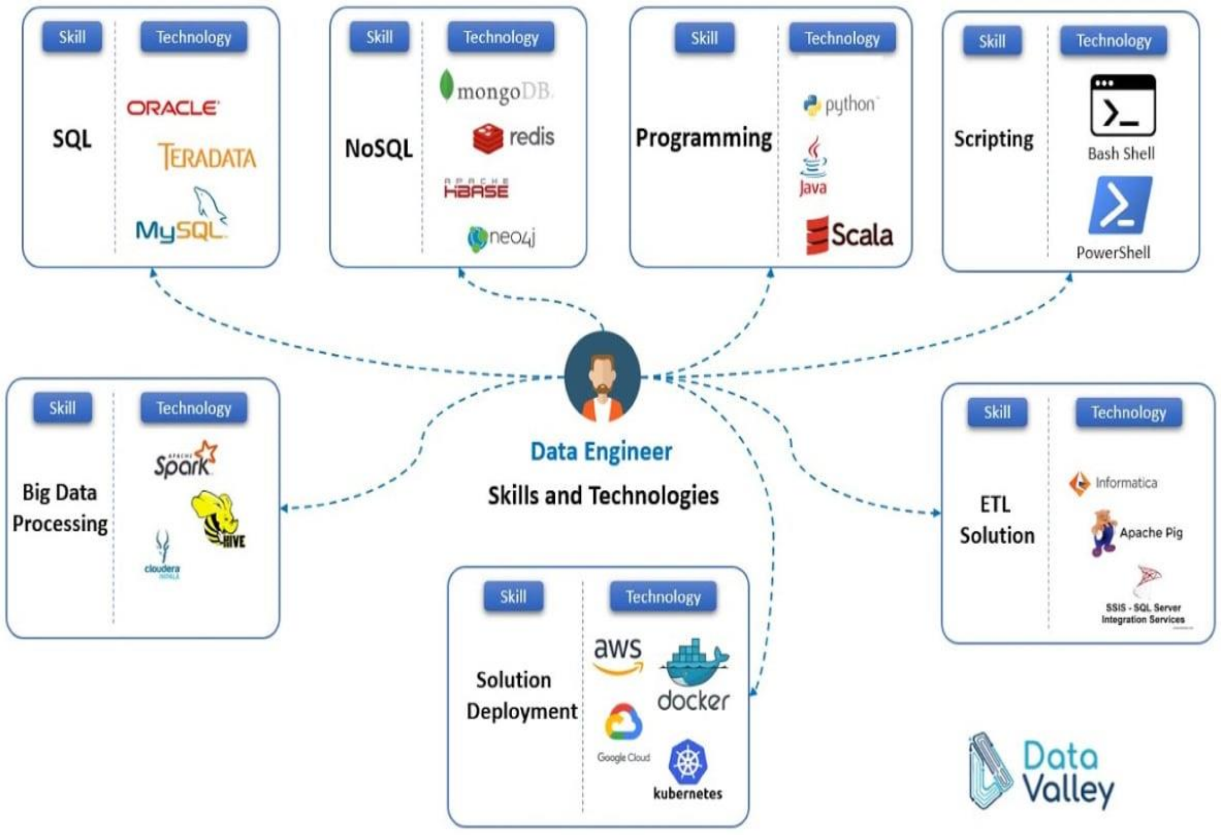




- Both work together
- The data that companies have in databases and other formats is prepared and organized by the data engineers. Data pipelines are also built by them that make data available to data scientists.
- This data is used by data scientists for analytics and other projects that improve business operations and outcomes.
- Data scientists and data engineers have a difference in their skill sets and focus.
- Data scientists tackle new and big-picture problems, while data engineers put the pieces in place to make that possible.

Data Engineer vs. Data Analyst





Data Analyst	Data Engineer	Data Scientist
Pre-processing and data gathering	Develop, test & maintain architectures	Responsible for developing Operational Models
Emphasis on representing data via reporting and visualization	Understand programming and its complexity	Carry out data analytics and optimization using machine learning & deep learning
Responsible for statistical analysis & data interpretation	Deploy ML & statistical models	Involved in strategic planning for data analytics
Ensures data acquisition & maintenance	Building pipelines for various ETL operations	Integrate data & perform ad-hoc analysis
Optimize Statistical Efficiency & Quality	Ensures data accuracy and flexibility	Fill in the gap between the stakeholders and customer

Importance of Data Engineering

1

Data engineering allows businesses to optimize data towards usability

2

In large organization, data is scattered in different formats prevents clear picture of business states and running analytics, data engineering solves this problem step by step

3

Data engineering simplifies data and makes it more reliable and useful for data scientist to work with

4

Data infrastructure built through data engineering allows organizations to leverage the valuable benefits of data analytics

5

Data engineers collect data from various sources then convert it into a format that can be analyzed and understood by data scientists

6

The goal of data engineering is to make data accessible, reliable and consistent for data scientist so they can focus on their analyses

Role of Data Engineering

1

Finding the best practices for refining your software development life cycle

2

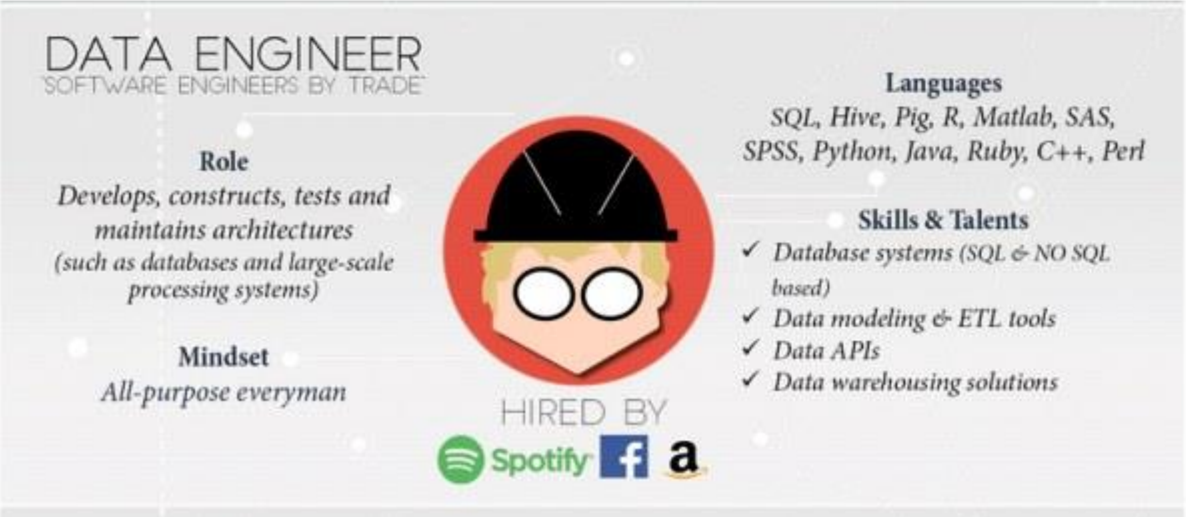
Tightening information security and protecting your business from cyber attacks

3

Increasing your knowledge about business domain

4

Bringing data together into one place



DATA ENGINEER

SOFTWARE ENGINEERS BY TRADE

Role

Develops, constructs, tests and maintains architectures (such as databases and large-scale processing systems)

Mindset

All-purpose everyman

HIRED BY

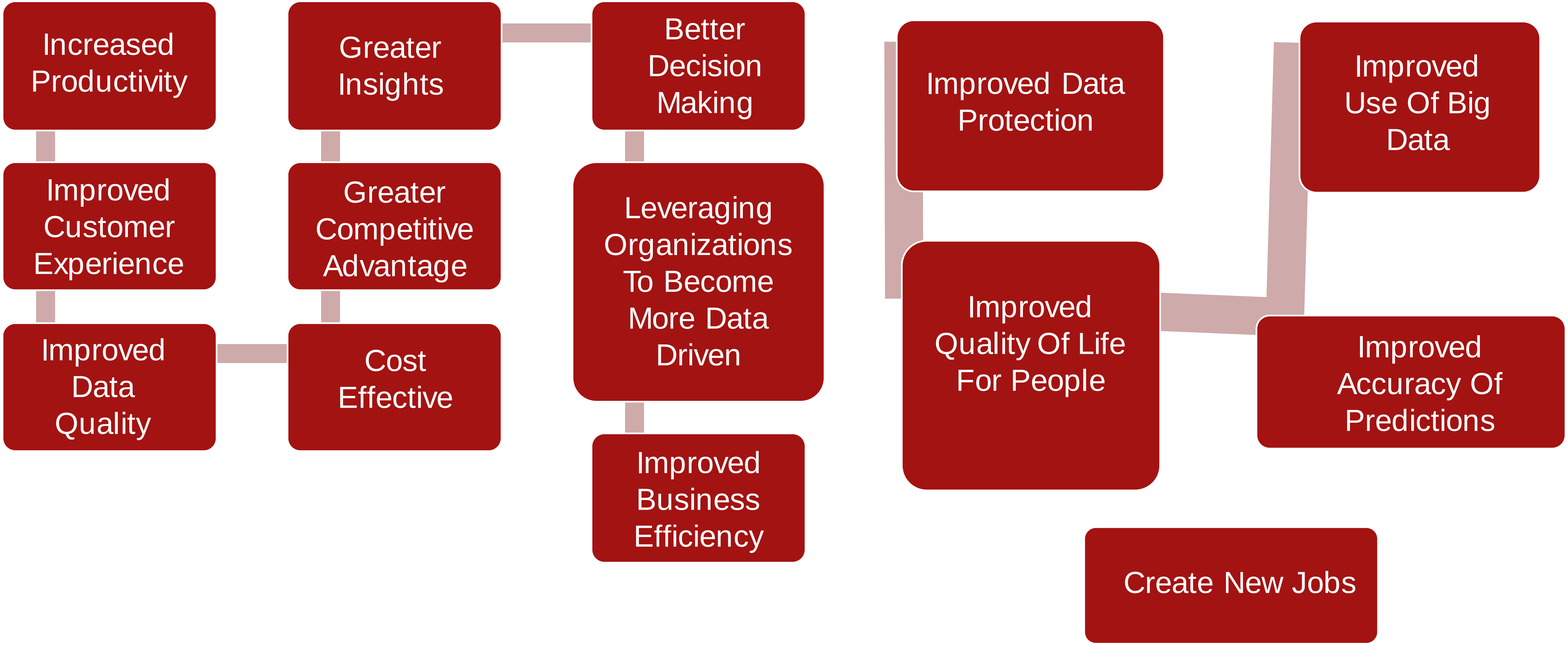
Languages

SQL, Hive, Pig, R, Matlab, SAS, SPSS, Python, Java, Ruby, C++, Perl

Skills & Talents

- ✓ *Database systems (SQL & NO SQL based)*
- ✓ *Data modeling & ETL tools*
- ✓ *Data APIs*
- ✓ *Data warehousing solutions*

Benefits of Data Engineering



Increased Productivity

Improved Customer Experience

Improved Data Quality

Greater Insights

Greater Competitive Advantage

Cost Effective

Better Decision Making

Leveraging Organizations To Become More Data Driven

Improved Business Efficiency

Improved Data Protection

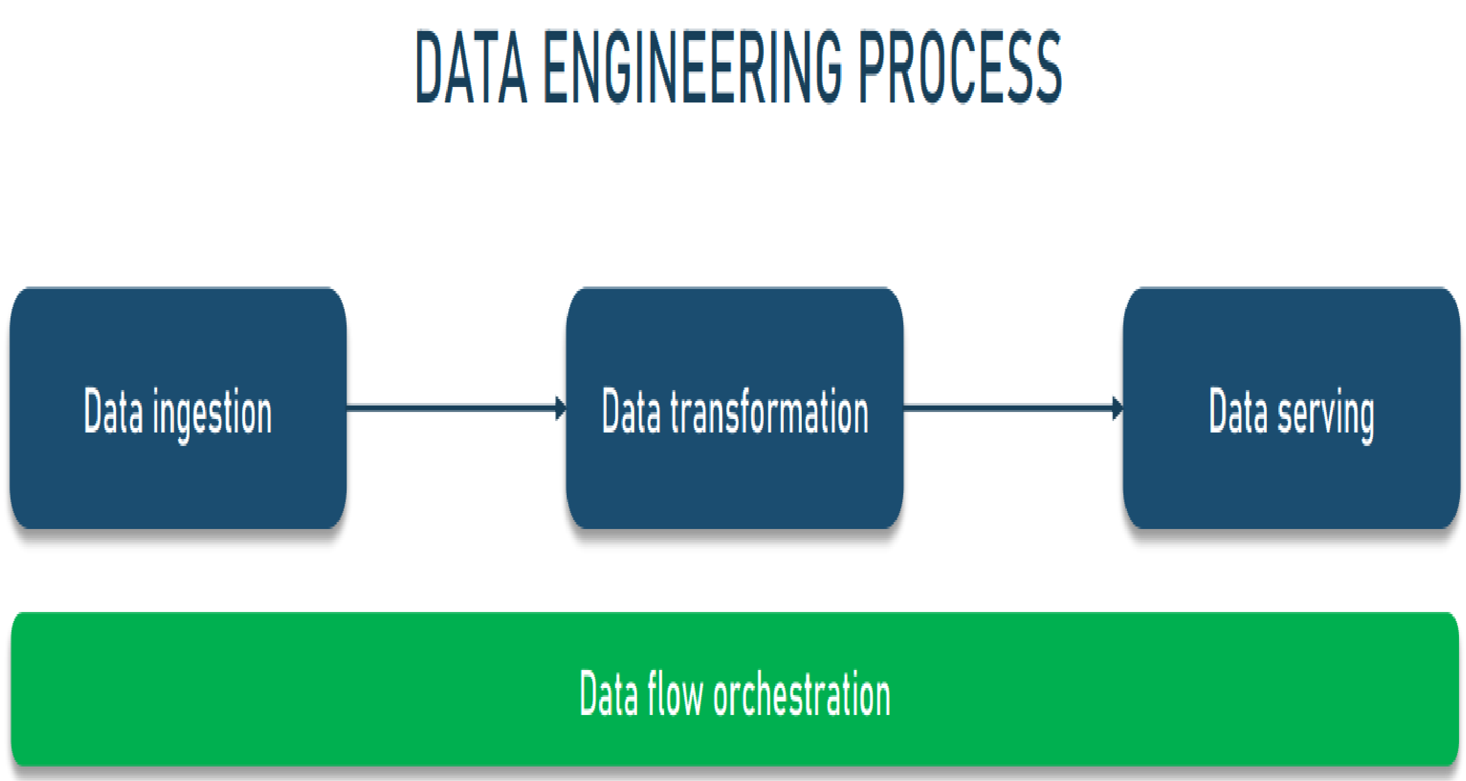
Improved Quality Of Life For People

Improved Use Of Big Data

Improved Accuracy Of Predictions

Create New Jobs

Data Engineering Process



DATA ENGINEERING PROCESS

Data ingestion → Data transformation → Data serving

Data flow orchestration

The mechanism that automates ingestion, transformation, and serving steps of the data engineering process is known as a data pipeline.

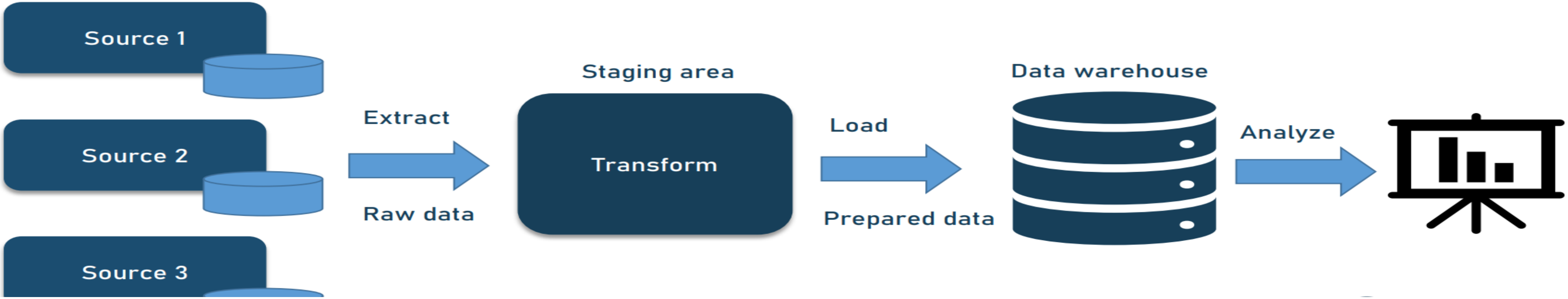
Data ingestion (acquisition) moves data from multiple sources

Data transformation adjusts disparate data to the needs of end users. It involves removing errors and duplicates from data, normalizing it, and converting it into the needed format.

Data serving delivers transformed data to end users — a BI platform, dashboard, or data science team.

Data flow orchestration provides visibility into the data engineering process, ensuring that all tasks are successfully completed

ETL PIPELINE



Source 1

Source 2

Source 3

Extract

Raw data

Staging area

Transform

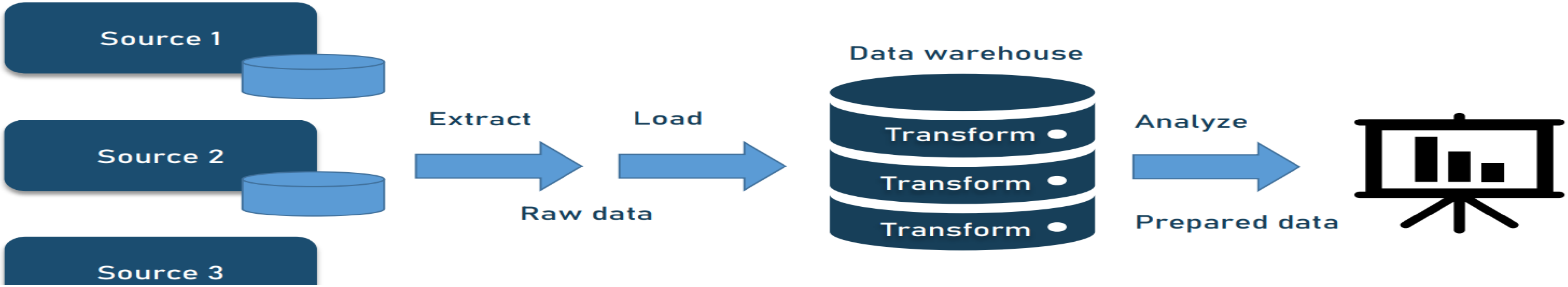
Load

Prepared data

Data warehouse

Analyze

ELT PIPELINE



Source 1

Source 2

Source 3

Extract

Raw data

Load

Data warehouse

Transform

Transform

Transform

Analyze

Prepared data

	ETL	ELT
Adoption of the technology and availability of tools and experts	ETL is a well-developed process used for over 20 years, and ETL experts are readily available.	ELT is a new technology, so it can be difficult to locate experts and more challenging to develop an ELT pipeline compared to an ETL pipeline.
Availability of data in the system	ETL only transforms and loads the data that you decide is necessary when creating the data warehouse and ETL process. Therefore, only this information will be available.	ELT can load all data immediately, and users can determine later which data to transform and analyze.
Can you add calculations?	Calculations will either replace existing columns, or you can append the dataset to push the calculation result to the target data system.	ELT adds calculated columns directly to the existing dataset.
Compatible with data lakes?	ETL is not normally a solution for data lakes. It transforms data for integration with a structured relational data warehouse system.	ELT offers a pipeline for data lakes to ingest unstructured data. Then it transforms the data on an as-needed basis for analysis.

Compliance	ETL can redact and remove sensitive information before putting it into the data warehouse or cloud server.	ELT requires you to upload the data before redacting/removing sensitive information.
Data size vs. complexity of transformations	ETL is best suited for dealing with smaller data sets that require complex transformations.	ELT is best when dealing with massive amounts of structured and unstructured data.
Data warehousing support	ETL works with cloud-based and onsite data warehouses. It requires a relational or structured data format.	ELT works with cloud-based data warehousing solutions to support structured, unstructured, semi-structured, and raw data types.
Hardware requirements	Cloud-based ETL platforms (like Integrate.io) don't require special hardware. Legacy, onsite ETL processes have extensive and costly hardware requirements, but they are not as popular today.	ELT processes are cloud-based and don't require special hardware.
How are aggregations different?	Aggregation becomes more complicated as the dataset increases in size.	As long as you have a powerful, cloud-based target data system, you can quickly process massive amounts of data.

Some of the top five critical differences between ETL vs. ELT are:

- ETL stands for Extract, Transform, and Load. ELT means Extract, Load, and Transform. Both are processes for data integration.
- Using the ETL method, data moves from the data source to staging, then into the data warehouse.
- ELT leverages the data warehouse to do basic transformations. There is no need for data staging.
- ETL can help with data privacy and compliance by cleaning sensitive and secure data before loading it into the data warehouse.
- ETL can perform sophisticated data transformations and can be more cost-effective than ELT

In what circumstances might you consider using ELT instead of ETL? Here are some of them:

Use Case #1:

A company with massive amounts of data. ELT works best with huge quantities of data, both structured and unstructured. As long as the target system is cloud-based, you will likely be able to process those huge amounts of data more quickly with an ELT solution. However, you may gain more accurate insights with ETL.

Use Case #2:

An organization with the resources to handle the processing power needed. With ETL, the majority of the processing takes place while the data is still in the pipeline before it gets to your warehouse. ELT does its work once the data has already arrived in the data lake.

Use Case #3:

A company that needs all its data in one place as soon as possible. Because the transformations take place at the end of the process, ELT prioritizes the speed of transfer over almost everything else, which means that all data—good, bad, and otherwise—ends up in the data lake for later transformation.

Data Source

A data source is the digital or physical location where data originates from or is stored, which influences how it is stored per its location (e.g., data table or data object) and its connectivity properties.

In many cases, a data source will refer to the first location where the data originated, though the movement and ingestion of data can change its source.

Additionally, a data source may be stored in the sole computer where it will be used (as is the case for desktop-based flat files or applications), or in an offsite location where it will be used by many different computers.

Types of Data Source

Machine data sources

- Unique to their machine
- Users can connect to the data using the information found within the machine data source
- query the data using the Data Source Name (DSN)
- DSN is a pointer to actual data in its respective databases or applications

File data sources

- Are Not Unique to their machine
- File data sources do not have a DSN since they are not registered to individual applications or systems

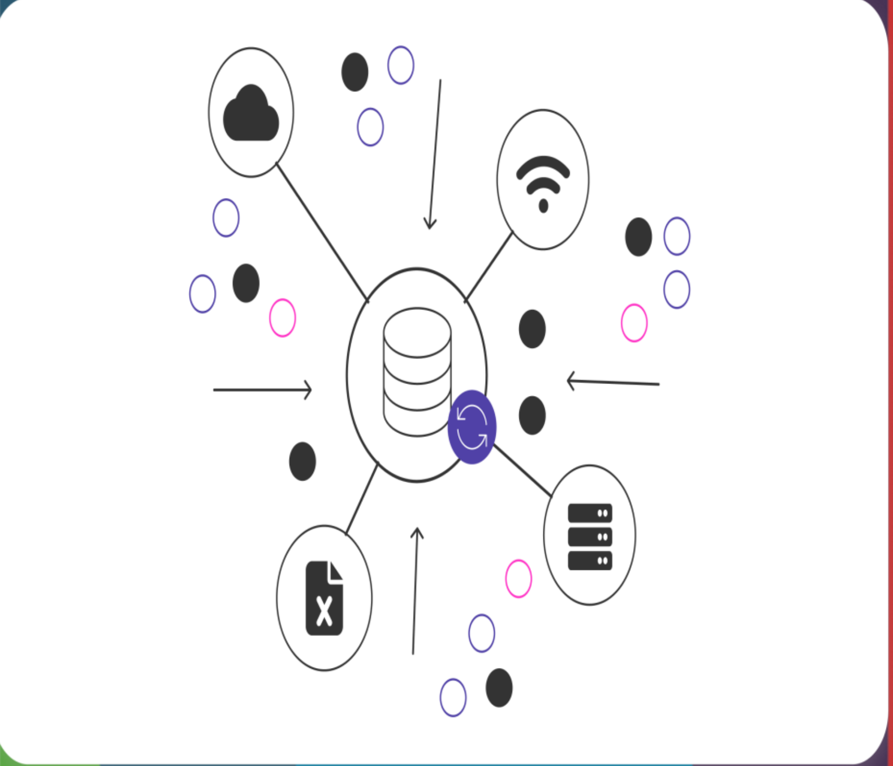
How Are Connections Established Between Data Sources?

File Transfer Protocol (FTP), HyperText Transfer Protocol (HTTP), or any of the many Application Programming Interfaces (APIs) provided by specific applications

Principally, an HTTP connection is used to access websites, while an FTP is used to transfer files between one host to another.

Additionally, HTTP only establishes data connection; FTP establishes both data as well as control connection. In general, HTTP excels at transferring smaller files like webpages, whereas FTP can quickly transfer large files

APIs allow applications to communicate with one another



What Is the Role of Data Engineering in Managing Data Sources?

Data engineers, responsible for the development, integrity, and maintenance of an organization’s data infrastructure

Data engineers must also be aware of who in the organization needs access to which data source and for what purpose

Traditionally, data engineers would work overtime to source, prepare, and join data from various sources in order to best serve the business

Now, data engineering work is best thought of as organization-wide work

Data engineers still handle the highly-technical work, such as migrating data from one database to another, but with the help of modern data engineering platforms, business analysts are now able to set up their own pipelines connecting to data sources, cleanse and enrich that data for use, and even set up schedules for repeat use.

Importance of Data Integration in Data Engineering

1

Data engineering allows businesses to optimize data towards usability

2

Data integration combines various types and formats of data from various sources into a single dataset that can be used to run applications or support business intelligence and analytics

3

- The volume, velocity, and variety of the data to be integrated.
- The characteristics of the sources & destinations of data.
- The time and resources available.
- The minimum performance standards.

4

1. Manual Data Integration
2. Application-based Data Integration
3. Common Data Storage
4. Data Virtualization
5. Middleware Data Integration

5

Early days of Data Engineering

6

Evolution of data

Data Integration Tools for Data Engineering

- What type of data will be in your data pipeline?
- How will that data be processed?
- Where will the data come from and go to?

Data Integration Techniques

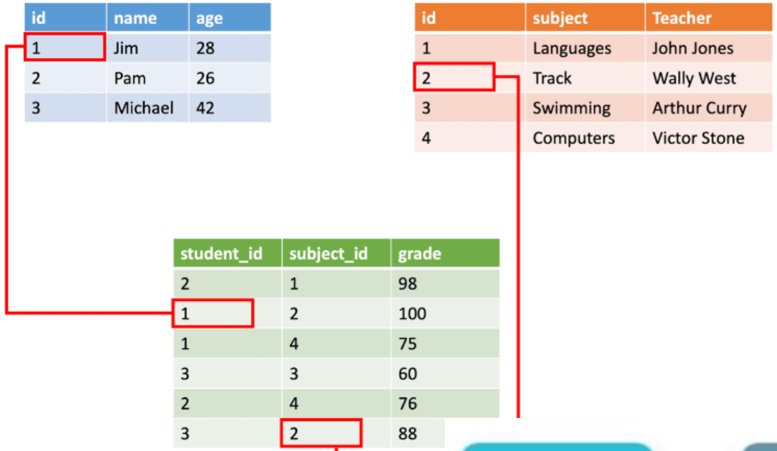
- *Extract, Transform, Load (ETL)*
- *Extract, Load, Transform (ELT)*
- *Change Data Capture (CDC)*
- *Enterprise Application Integration (EAI)*
- *Data Virtualization*
- *Master Data Management (MDM)*

Types of Data

Structured data

Unstructured data

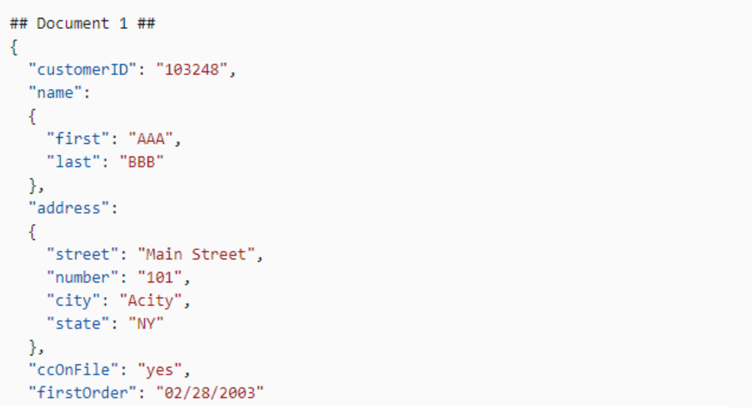
Semi structured data



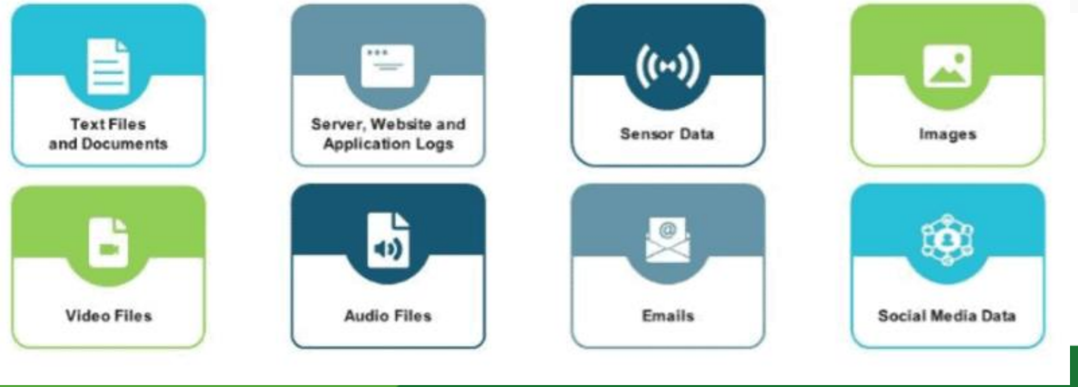
id	name	age
1	Jim	28
2	Pam	26
3	Michael	42

id	subject	Teacher
1	Languages	John Jones
2	Track	Wally West
3	Swimming	Arthur Curry
4	Computers	Victor Stone

student_id	subject_id	grade
2	1	98
1	2	100
1	4	75
3	3	60
2	4	76
3	2	88



```
## Document 1 ##
{
  "customerID": "103248",
  "name":
  {
    "first": "AAA",
    "last": "BBB"
  },
  "address":
  {
    "street": "Main Street",
    "number": "101",
    "city": "Acity",
    "state": "NY"
  },
  "ccOnFile": "yes",
  "firstOrder": "02/28/2003"
```

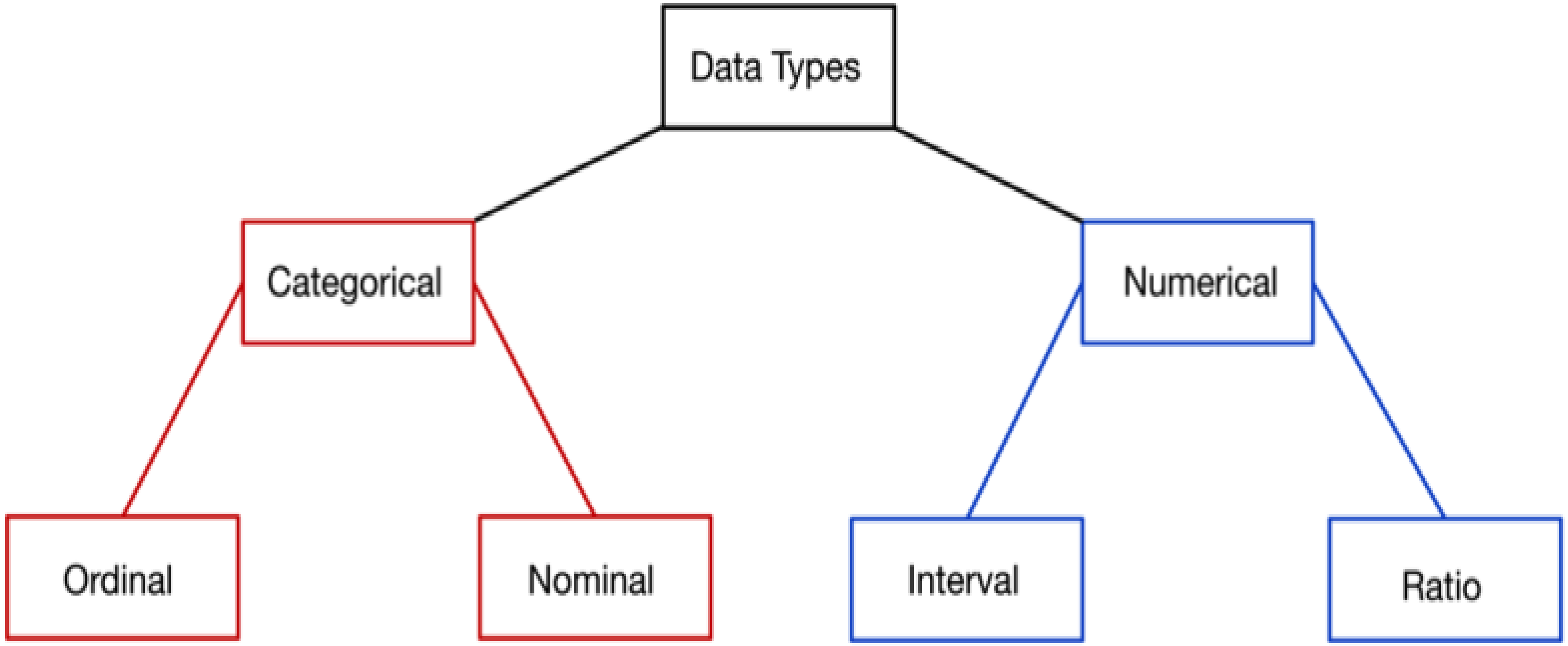


On the basis of	Structured data	Unstructured data
Technology	It is based on a relational database.	It is based on character and binary data.
Flexibility	Structured data is less flexible and schema-dependent.	There is an absence of schema, so it is more flexible.
Scalability	It is hard to scale database schema.	It is more scalable.
Robustness	It is very robust.	It is less robust.
Performance	Here, we can perform a structured query that allows complex joining, so the performance is higher.	While in unstructured data, textual queries are possible, the performance is lower than semi-structured and structured data.
Nature	Structured data is quantitative, i.e., it consists of hard numbers or things that can be counted.	It is qualitative, as it cannot be processed and analyzed using conventional tools.
Format	It has a predefined format.	It has a variety of formats, i.e., it comes in a variety of shapes and sizes.
Analysis	It is easy to search.	Searching for unstructured data is more difficult.

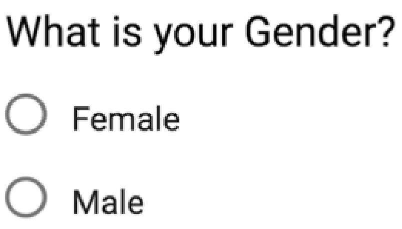
Data Types in Statistics

WHAT ARE THE MAIN TYPES OF DATA IN STATISTICS?

- **Nominal data**
- **Ordinal data**
- **Discrete data**
- **Continuous data**



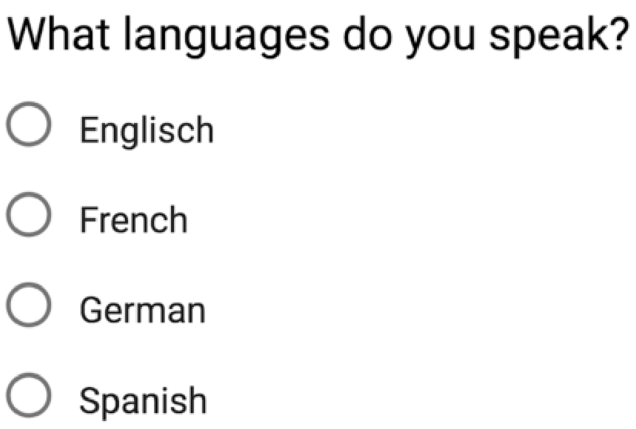
Nominal values represent discrete units and are used to label variables that have no quantitative value



What is your Gender?

○ Female

○ Male



What languages do you speak?

○ Englisch

○ French

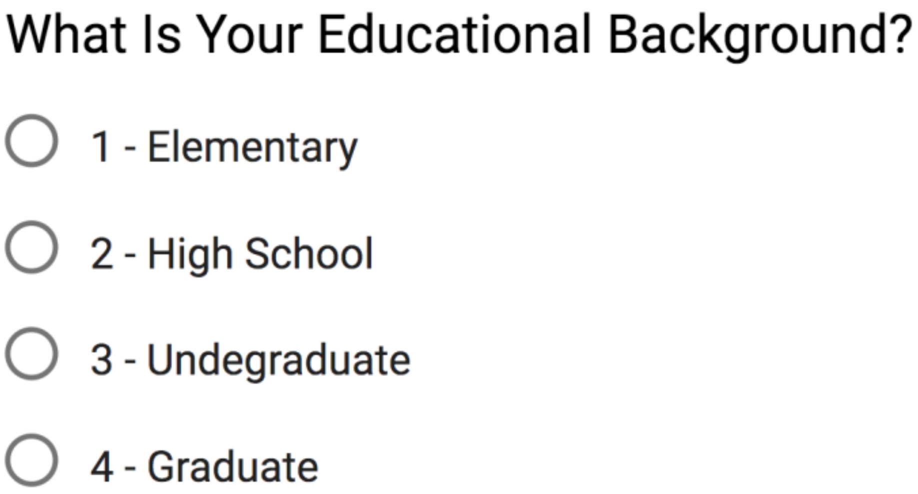
○ German

○ Spanish

Discrete data if the data can only take on certain values. This type of data can't be measured but it can be counted. It basically represents information that can be categorized into a classification. An example is the number of heads in 100-coin flips

Ratio values are also ordered units that have the same difference. Ratio values are the same as interval values, with the difference that they do have an absolute zero. Good examples are height, weight, length, etc.

Ordinal values represent discrete and ordered units



What Is Your Educational Background?

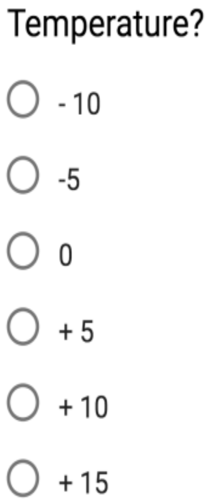
○ 1 - Elementary

○ 2 - High School

○ 3 - Undegraduate

○ 4 - Graduate

Interval values represent ordered units that have the same difference



Temperature?

○ - 10

○ -5

○ 0

○ + 5

○ + 10

○ + 15

Statistical Methods for Nominal, Ordinal and Continuous Data Types

SUMMARIZING NOMINAL DATA

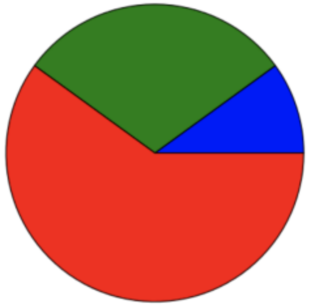
Frequencies: The frequency is the rate at which something occurs over a period of time or within a data set.

Proportion: You can easily calculate the proportion by dividing the frequency by the total number of events.

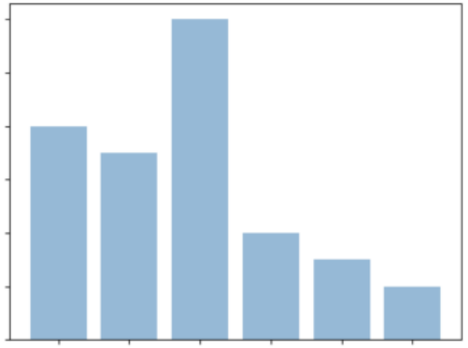
(e.g how often something happened divided by how often it could happen)

Visualization Methods: To visualize nominal data you can use a pie chart or a bar chart

Pie Chart



Bar Chart



Statistical Methods for Nominal, Ordinal and Continuous Data Types

SUMMARIZING ORDINAL DATA

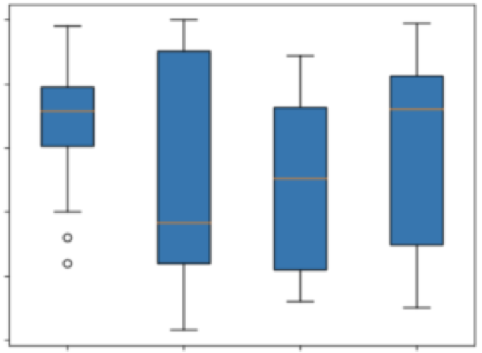
- When you are dealing with ordinal data, you can use the same methods as with nominal data, but you also have access to some additional tools
- You can summarize your ordinal data with frequencies, proportions, percentages. And you can visualize it with pie and bar charts.
- Additionally, you can use percentiles, median, mode and the interquartile range to summarize your data

Statistical Methods for Nominal, Ordinal and Continuous Data Types

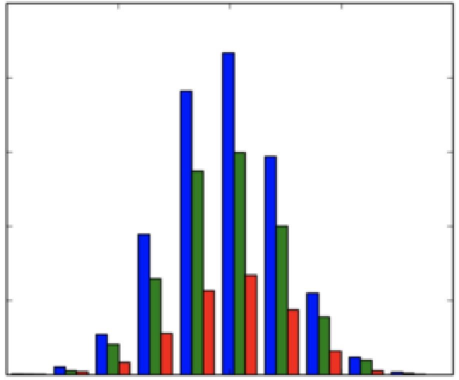
SUMMARIZING CONTINUOUS DATA

- When you are dealing with continuous data, you can use the most methods to describe your data.
- You can summarize your data using percentiles, median, interquartile range, mean, mode, standard deviation, and range
- To visualize continuous data, you can use a histogram or a boxplot

Boxplot



Histogram



Data Distribution

- Data distribution is a function that specifies all possible values for a variable and quantifies the relative frequency.
- Distributions are considered to be any population that has a scattering of data
- Data distributions are widely used in statistics.
- Data distribution methods organize the raw data into graphical methods (like histograms, box plots, run charts, etc.) and provide helpful information.
- A probability distribution is a mathematical model that calculates the probability of occurrence of different possible outcomes in a test or experiment.
- We use them to define different types of random variables (typically discreet or continuous) to make decisions, depending on the models.

Types of Data Distribution

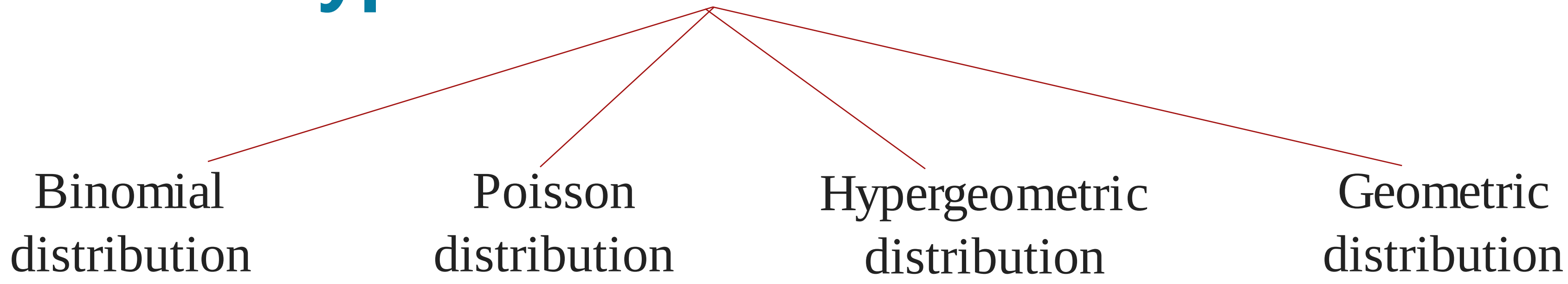
Discrete Distributions

Continuous Distributions

Discrete Distributions

- ✓ A discrete distribution results from countable data with a finite number of possible values.
- ✓ Ex: rolling dice, choosing several heads, etc.
- ✓ **Probability mass function (pmf):** A probability mass function is a frequency function that gives the probability for discrete random variables, also known as the discrete density function.
- ✓ Simply Discrete= counted

Types of Discrete Distributions



Binomial distribution

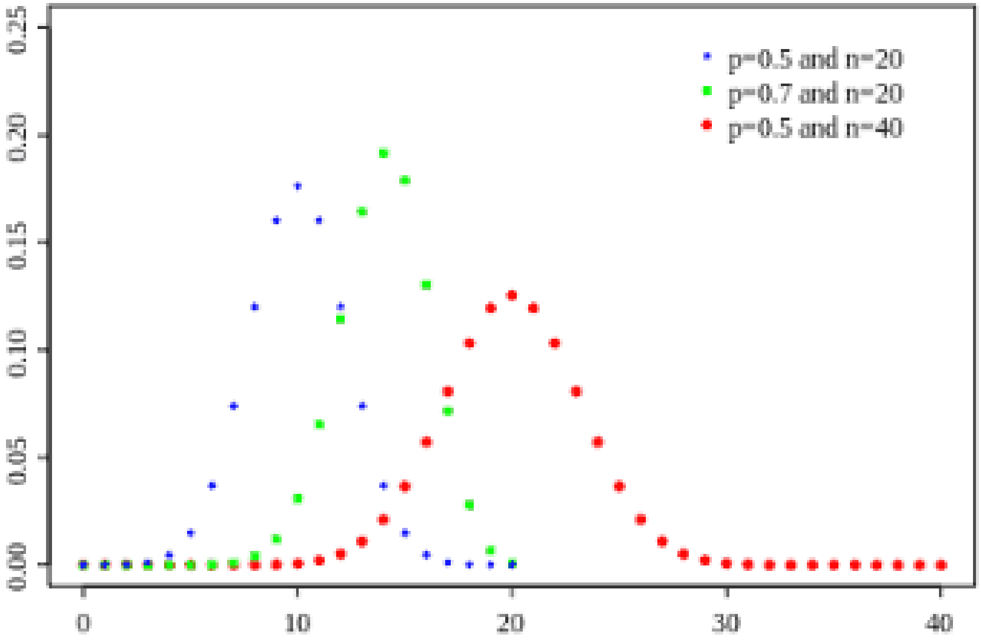
Poisson distribution

Hypergeometric distribution

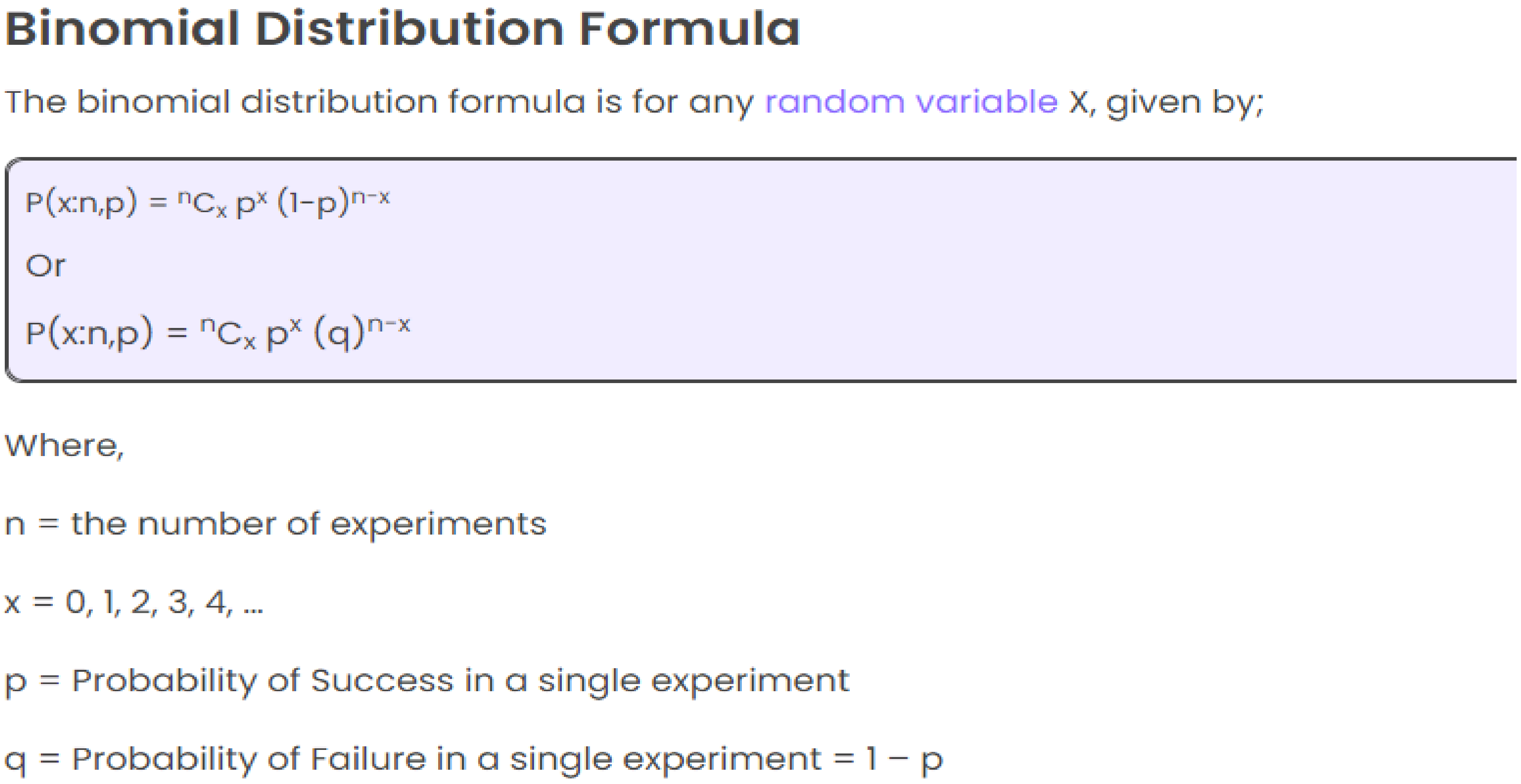
Geometric distribution

Binomial distributions

- The binomial distribution measures the probability of the number of successes or failure outcomes in an experiment on each try.
- Characteristics are classified into two mutually exclusive and exhaustive classes, such as the number of successes/failures and the number accepted/rejected that follow a binomial distribution.
- Ex: Tossing a coin: The probability of the coin landing Head is ½, and the probability of the coin landing tail is ½.



- In binomial probability distribution, the number of ‘Success’ in a sequence of n experiments, where each time a question is asked for yes-no, then the boolean-valued outcome is represented either with success/yes/true/one (probability p) or failure/no/false/zero (probability $q = 1 - p$).
- A single success/failure test is also called a Bernoulli trial or Bernoulli experiment, and a series of outcomes is called a Bernoulli process.
- For $n = 1$, i.e., a single experiment, the binomial distribution is a Bernoulli distribution.



Binomial Distribution Formula

The binomial distribution formula is for any random variable X, given by;

$P(x:n,p) = {}^{n}C_{x}\, p^{x}\, (1\text{-}p)^{n-x}$

Or

$P(x:n,p) = {}^{n}C_{x}\, p^{x}\, (q)^{n-x}$

Where,

n = the number of experiments

x = 0, 1, 2, 3, 4, …

p = Probability of Success in a single experiment

q = Probability of Failure in a single experiment = 1 – p

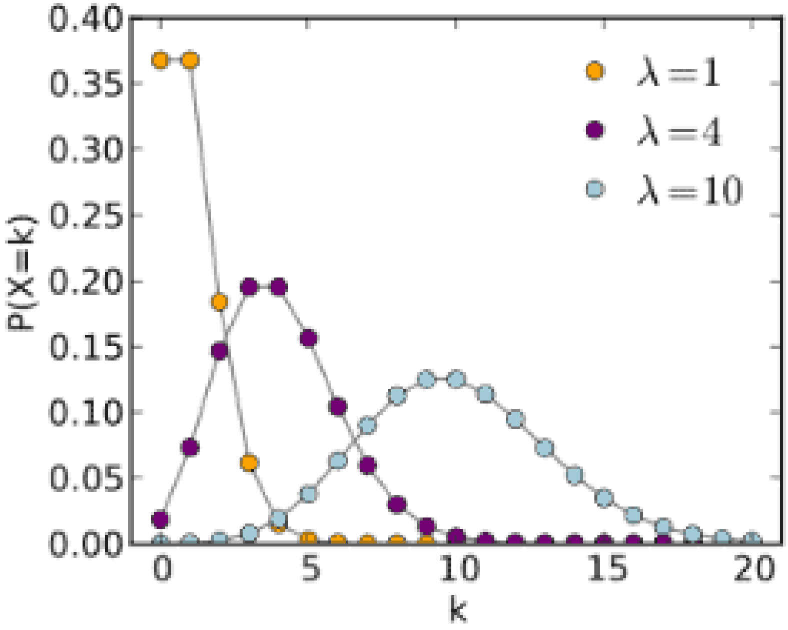
Binomial Distribution Examples

- Finding the quantity of raw and used materials while making a product.
- Taking a survey of positive and negative reviews from the public for any specific product or place.
- By using the YES/ NO survey, we can check whether the number of persons views the particular channel.
- To find the number of male and female employees in an organization.
- The number of votes collected by a candidate in an election is counted based on 0 or 1 probability.

Poisson distributions

- The Poisson distribution is the discrete probability distribution that measures the likelihood of a number of events occurring in a given period when the events occur one after another in a well-defined manner.
- Characteristics that can theoretically take large values but actually take small values have Poisson distribution.
- Ex: Number of defects, errors, accidents, absentees, etc.

1. The number of trials "n" tends to infinity
2. Probability of success "p" tends to zero
3. np = 1 is finite



- The formula for the Poisson distribution function is given by:
- $f(x) = (e^{-\lambda} \lambda^{x})/x!$

where,

e is the base of the logarithm

x is a Poisson random variable

λ is an average rate of value

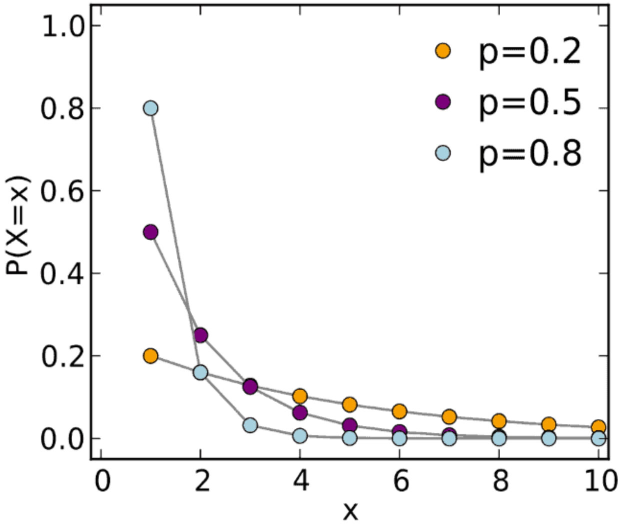
Telephone calls arrive at an exchange according to the Poisson process at a rate λ= 2/min. Calculate the probability that exactly two calls will be received during each of the first 5 minutes of the hour.

Hypergeometric distributions

- The hypergeometric distribution is a discrete distribution that measures the probability of a specified number of successes in (n) trials without replacement from a relatively large population (N).
- The basic difference of binomial distribution is that probability of success is the same for all trials, while it is not the same case for hypergeometric distribution.

Geometric distributions

- The geometric distribution is a discrete distribution that measures the likelihood of when the first success will occur.
- An extension of it may be considered a negative binomial distribution.
- Ex: Marketing representative from an advertising agency randomly selects hockey players from various universities until he finds a hockey player who attended the Olympics.

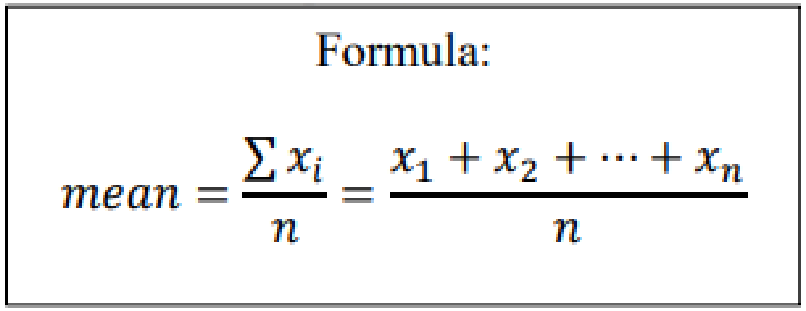


Measures of Central Tendency & Dispersion

Measures that indicate the approximate center of a distribution are called measures of central tendency .

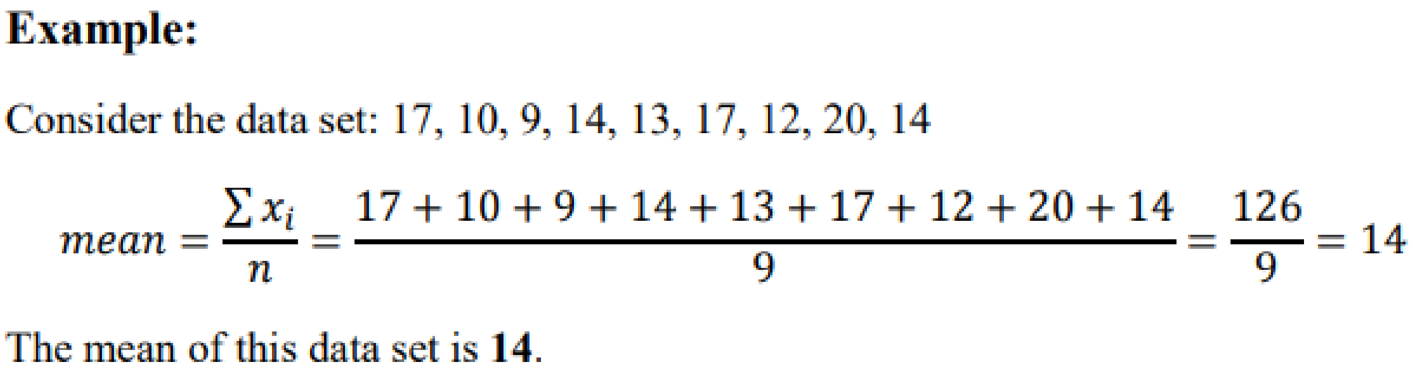
Measures that describe the spread of the data are measures of dispersion.

- The mean of a set of data is the sum of all values in a data set divided by the number of values in the set.
- It is also often referred to as an arithmetic average.
- To determine the mean of a data set:
- 1. Add together all data values.
- 2. Divide the sum from Step 1 by the number of data values in the set



Formula:

$$mean = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$



Example:

Consider the data set: 17, 10, 9, 14, 13, 17, 12, 20, 14

$$mean = \frac{\sum x_i}{n} = \frac{17 + 10 + 9 + 14 + 13 + 17 + 12 + 20 + 14}{9} = \frac{126}{9} = 14$$

The mean of this data set is **14**.

Mean of Ungrouped Data

Let $x_1, x_2, x_3 , . . . , x_n$ be n observations. We can find the arithmetic mean using the mean formula:

Mean, $\bar{x} = (x_1 + x_2 + ... + x_n)/n$

Example: If the heights of 5 people are 142 cm, 150 cm, 149 cm, 156 cm, and 153 cm.

Find the mean height.

Mean height, $\bar{x} = (142 + 150 + 149 + 156 + 153)/5$

$= 750/5$

$= 150$

Mean, $\bar{x} = 150$ cm

Thus, the mean height is 150 cm.

Mean = 9

Mean of Grouped Data

When the data is present in tabular form, we use the following formula:

Mean, $\bar{x} = (x1f1 + x2f2 + ... + xnfn)/(f1 + f2 + ... + fn)$

Consider the following example.

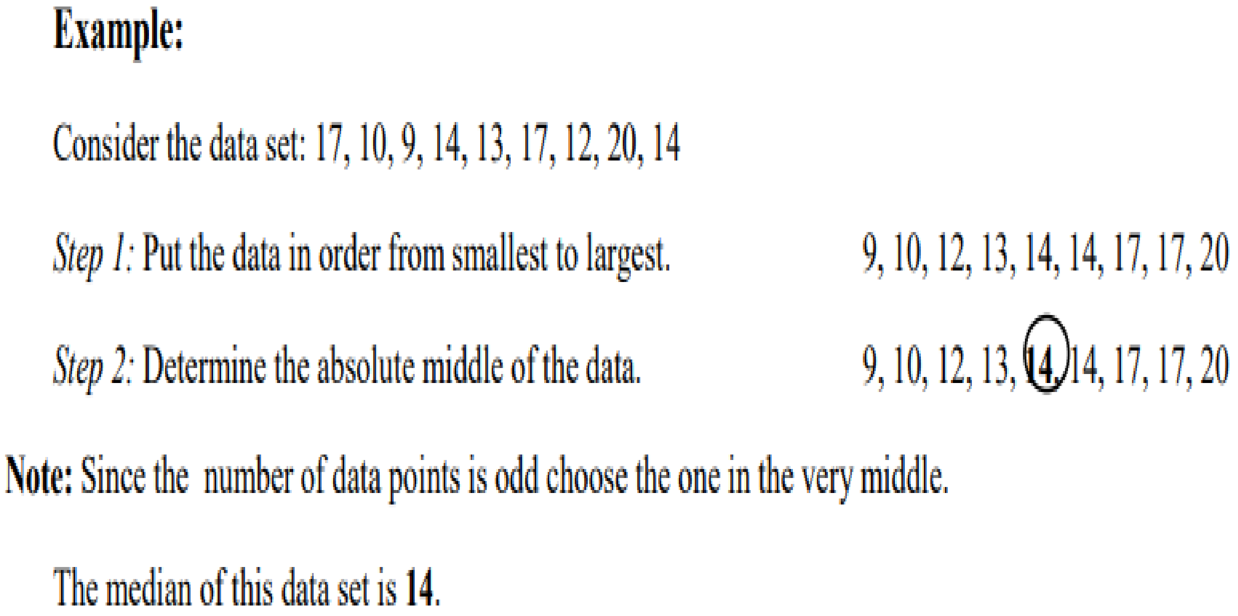
Example 1: Find the mean of the following distribution:

x	4	6	9	10	15
f	5	10	10	7	8

Finding the Median

- The median of a set of data is the “middle element” when the data is arranged in ascending order.
- To determine the median:

1. Put the data in order from smallest to largest.
2. Determine the number in the exact center.
    i. If there are an odd number of data points, the median will be the number in the absolute middle.
    ii. ii. If there is an even number of data points, the median is the mean of the two center data points, meaning the two center values should be added together and divided by 2.



Example:

Consider the data set: 17, 10, 9, 14, 13, 17, 12, 20, 14

Step 1: Put the data in order from smallest to largest. 9, 10, 12, 13, 14, 14, 17, 17, 20

Step 2: Determine the absolute middle of the data. 9, 10, 12, 13, $\boxed{14}$ 14, 17, 17, 20

Note: Since the number of data points is odd choose the one in the very middle.

The median of this data set is **14**.

Example 1: Let's consider the data: 56, 67, 54, 34, 78, 43, 23. What is the median?

Median = 54

Example 2: Let's consider the data: 50, 67, 24, 34, 78, 43. What is the median?

Median = 46.5

Example: Find the median marks for the following distribution:

Classes	0-10	10-20	20-30	30-40	40-50
Frequency	2	12	22	8	6

Solution:

N = 50

N/2 = 50/2 = 25

Median Class = (20 - 30)

l = 20, f = 22, c = 14, h = 10

Using Median formula:

Median = l+[((n/2)−c)/ f]×h

= 20 + (25 - 14)/22 × 10

= 20 + (11/22) × 10

= 20 + 5 = 25

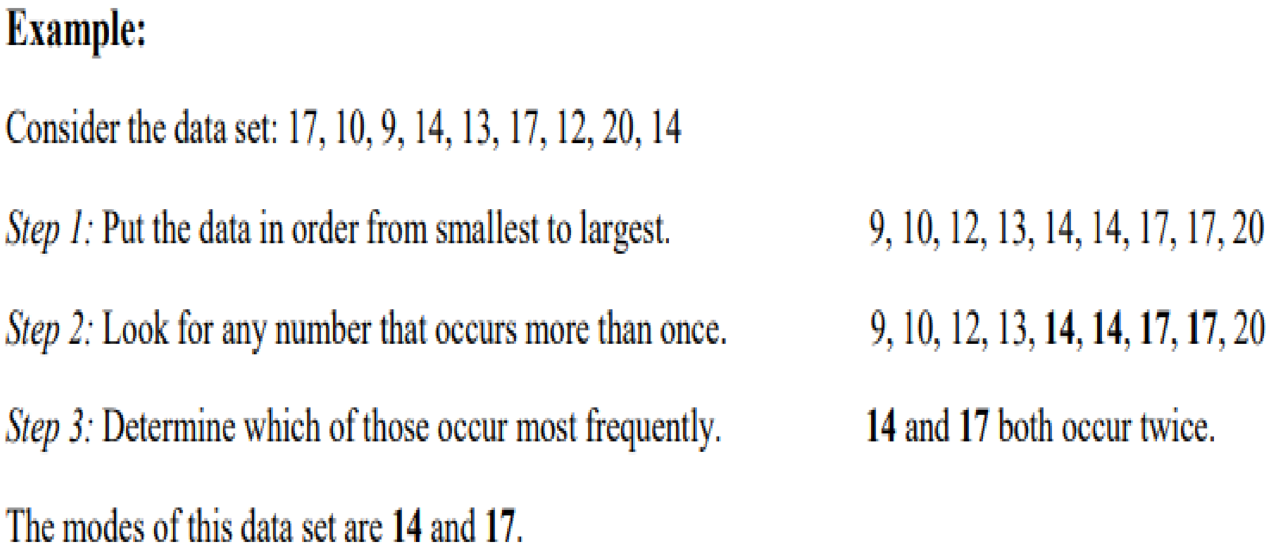
∴ Median = 25

Classes	Number of students	Cumulative frequency
0-10	2	2
10-20	12	2 + 12 = 14
20-30	22	14 + 22 = 36
30-40	8	36 + 8 = 44
40-50	6	44 + 6 = 50

Finding the Mode

The mode is the most frequently occurring measurement in a data set. There may be one mode; multiple modes, if more than one number occurs most frequently; or no mode at all, if every number occurs only once. To determine the mode:

1. Put the data in order from smallest to largest, as you did to find your median.
2. Look for any value that occurs more than once.
3. Determine which of the values from Step 2 occurs most frequently.



Example:

Consider the data set: 17, 10, 9, 14, 13, 17, 12, 20, 14

Step 1: Put the data in order from smallest to largest. 9, 10, 12, 13, 14, 14, 17, 17, 20

Step 2: Look for any number that occurs more than once. 9, 10, 12, 13, **14**, **14**, **17**, **17**, 20

Step 3: Determine which of those occur most frequently. **14** and **17** both occur twice.

The modes of this data set are **14** and **17**.

Data Cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset.

What is the difference between data cleaning and data transformation?

Data cleaning is the process that removes data that does not belong in your dataset. Data transformation is the process of converting data from one format or structure into another.

Transformation processes can also be referred to as data wrangling, or data munging, transforming and mapping data from one "raw" data form into another format for warehousing and analyzing.

Characteristics of Quality Data

1. Accuracy
2. Completeness
3. Reliability
4. Relevance
5. Timeliness

Data cleaning ensures that the quality of data is preserved and enhanced to meet the business needs. Insights are drawn from quality data help make the best business decisions.

What makes data messy?

It’s wrong. Sometimes—especially when data is manually entered by people—the data is just wrong.

It’s inconsistent. People express the same ideas in different ways

It’s built for machines. A lot of machine-generated data is created in a format that’s useful for a machine but not a human.

It’s scattered all over the place. Even if the data itself is clean, it may be scattered across a number of sources.

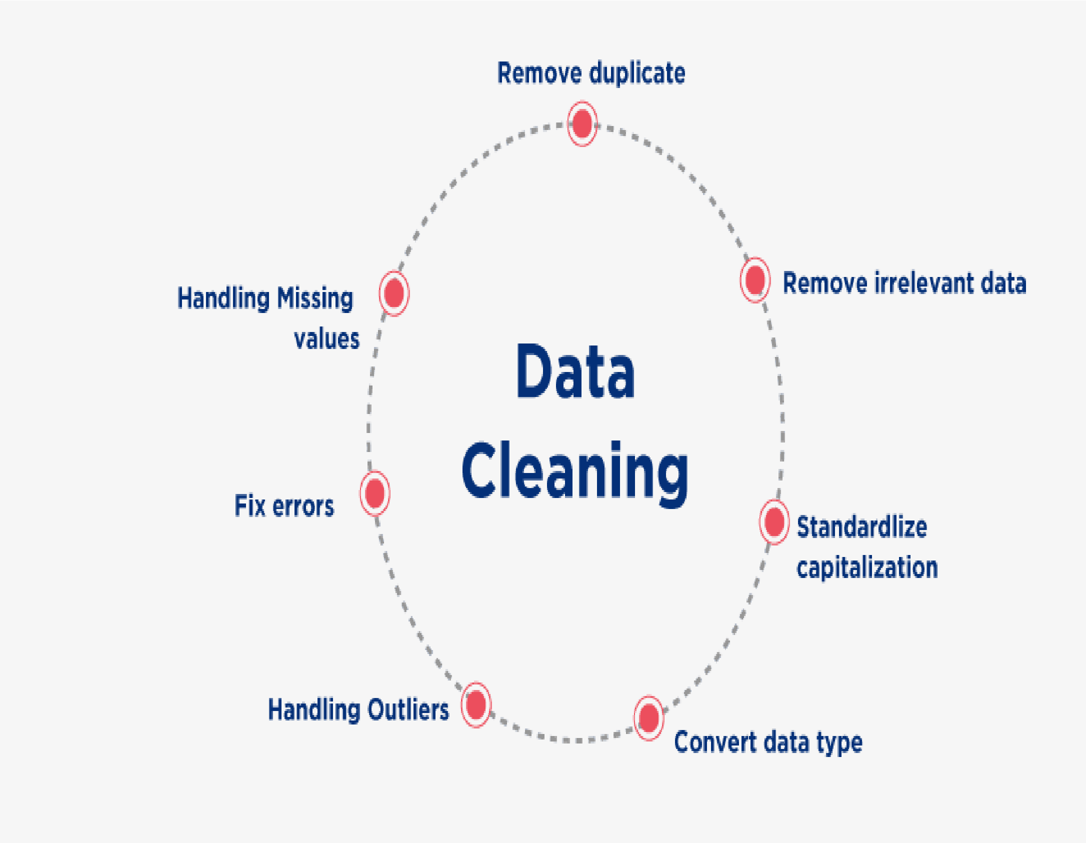
To solve these problems, companies typically first centralize their data into a single data warehouse before cleaning it up in any other way. If the source data is messy (e.g., if phone numbers are inconsistent in Salesforce), the data in the warehouse will be messy.

Benefits of Data Cleaning

It enhances the results one gets from their analysis.

Data cleaning removes errors from the dataset.

The data cleaning process can be used to help solve the incorrect or corrupted data collection.



Clean data makes it easy to make decisions.

How to clean data?

Step 1: Remove duplicate or irrelevant observations

When you combine data sets from multiple places, scrape data, or receive data from clients or multiple departments, there are opportunities to create duplicate data.

Step 2: Fix structural errors

Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalization.

Step 3: Filter unwanted outliers

This step is needed to determine the validity of that number. If an outlier proves to be irrelevant for analysis or is a mistake, consider removing it.

Step 4: Handle missing data

You can't ignore missing data because many algorithms will not accept missing values.

Step 5: Validate and QA

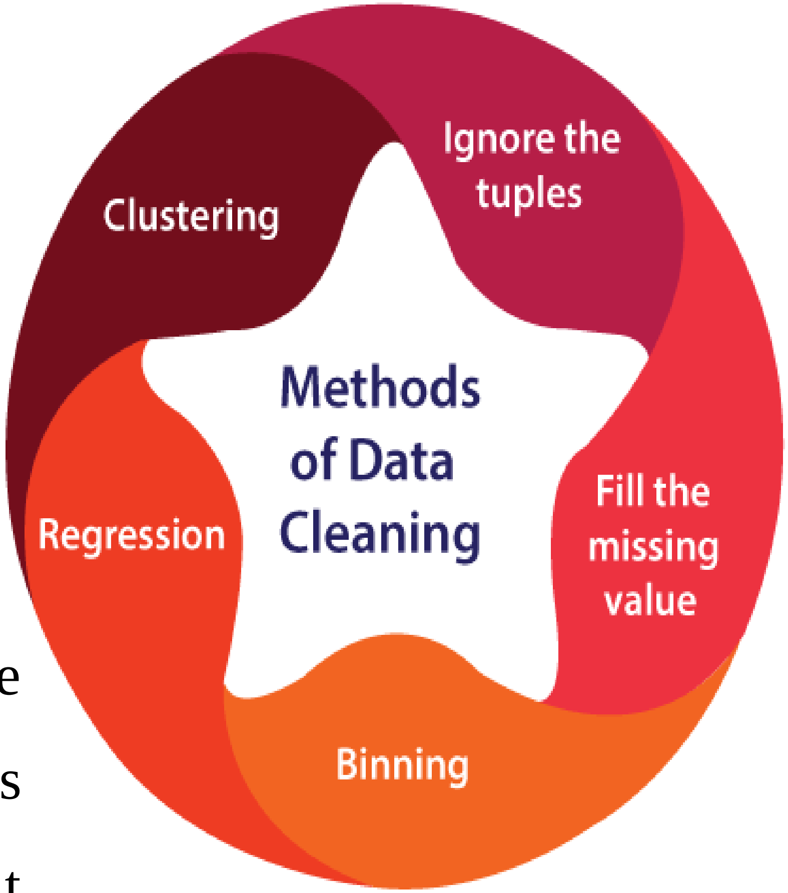
- Does the data make sense?
- Does the data follow the appropriate rules for its field?
- Does it prove or disprove your working theory, or bring any insight to light?
- Can you find trends in the data to help you form your next theory?
- If not, is that because of a data quality issue?

Methods of Data Cleaning

1.Ignore the tuples: This method is not very feasible, as it only comes to use when the tuple has several attributes is has missing values.

2.Fill the missing value: This approach is also not very effective or feasible. Moreover, it can be a time-consuming method. In the approach, one has to fill in the missing value.

3.Binning method: This approach is very simple to understand. The smoothing of sorted data is done using the values around it. The data is then divided into several segments of equal size. After that, the different methods are executed to complete the task.



Methods of Data Cleaning

4. Regression: The data is made smooth with the help of using the regression function. The regression can be linear or multiple. Linear regression has only one independent variable, and multiple regressions have more than one independent variable.

5. Clustering: This method mainly operates on the group. Clustering groups the data in a cluster. Then, the outliers are detected with the help of clustering. Next, the similar values are then arranged into a "group" or a "cluster".

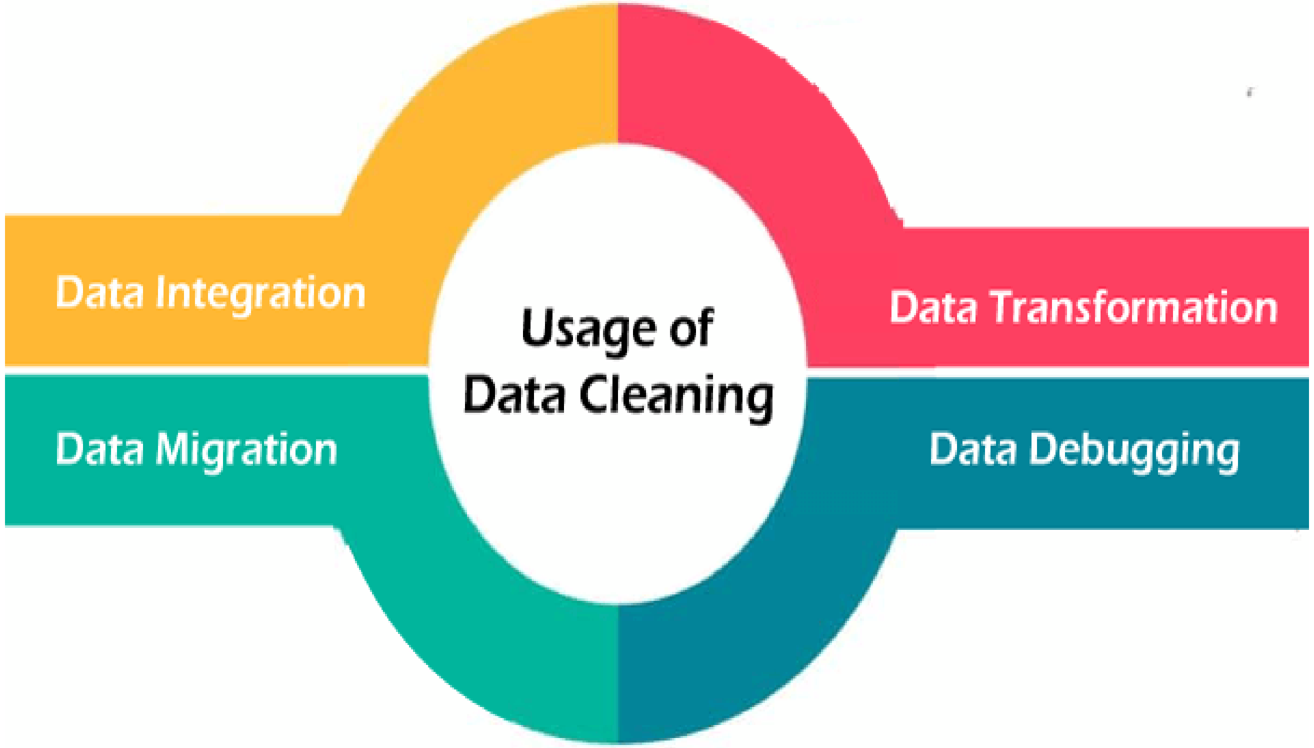
Usage of Data Cleaning

Data Integration: Since it is difficult to ensure quality in low-quality data, data integration has an important role in solving this problem.

Data Migration: Data migration is the process of moving one file from one system to another, one format to another, or one application to another.

Data Transformation: Before the data is uploaded to a destination, it needs to be transformed. This is only possible through data cleaning, which considers the system criteria of formatting, structuring, etc.

Data Debugging in ETL Processes: Data cleansing is crucial to preparing data during extract, transform, and load (ETL) for reporting and analysis.



Tools for Data Cleaning

1. OpenRefine
2. Trifacta Wrangler
3. Drake
4. Data Ladder
5. Data Cleaner
6. Cloudingo
7. Reifier
8. IBM Infosphere Quality Stage
9. TIBCO Clarity
10. Winpure

Descriptive Statistics

- Statistics used to describe and interpret sample data.
- Results are not really meant to apply to other samples or to the larger population

    - Frequency Distribution
    - Central Tendency (Mean, Median, Mode)
    - Percentile Values

Inferential Statistics

Statistics used to make inference about the population from which the sample was drawn.

- Correlation
- T-test
- ANOVA (Analysis of Variance)
- Regression

Population vs. Sample

- Population: A large group of people to which we are interested in generalizing.

  ‘parameter’

- Sample: A smaller group drawn from a population.

  ‘statistic’

Measures of Central Tendency

Statistics that identify where the center or middle of the set of scores are.

- **Mode** : Most frequently occurring scores.
- **Median** : the 50th percentile, the second quartile
- **Mean** : Arithmetic means, average, Add all the scores and divide by the number of scores.

Which central tendency to use?

Depends on :

1. The level of measurement of the data.

2. The shape of the score distribution. (Skewness)

Level of Measurement

- Nominal: Categorical scale
  - e.g. Male/Female, Blue eye/Brown eye/Green eye
- Ordinal: Ranking scale

  (Differences between the ranks need not be equal)
  - e.g. Scored highest (100 pts), middle (85 pts), lowest (20 pts)
- Interval: The distance between any two adjacent units of measurement (intervals) is the same but there is no meaningful zero point.
  - e.g. Fahrenheit temperature
- Ratio: The distance between any two adjacent units of measurement is the same and there is a true zero point. e.g. Height measurement, Weight measurement

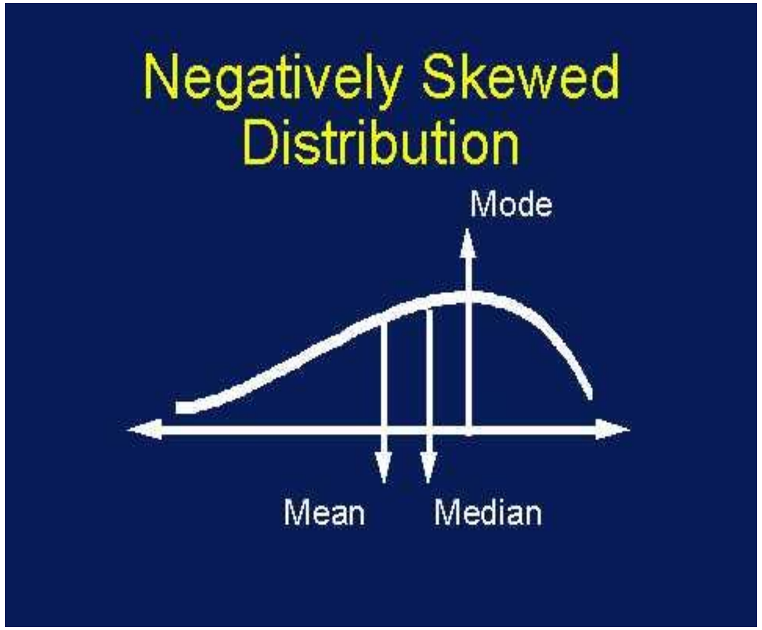
Which central tendency to use?

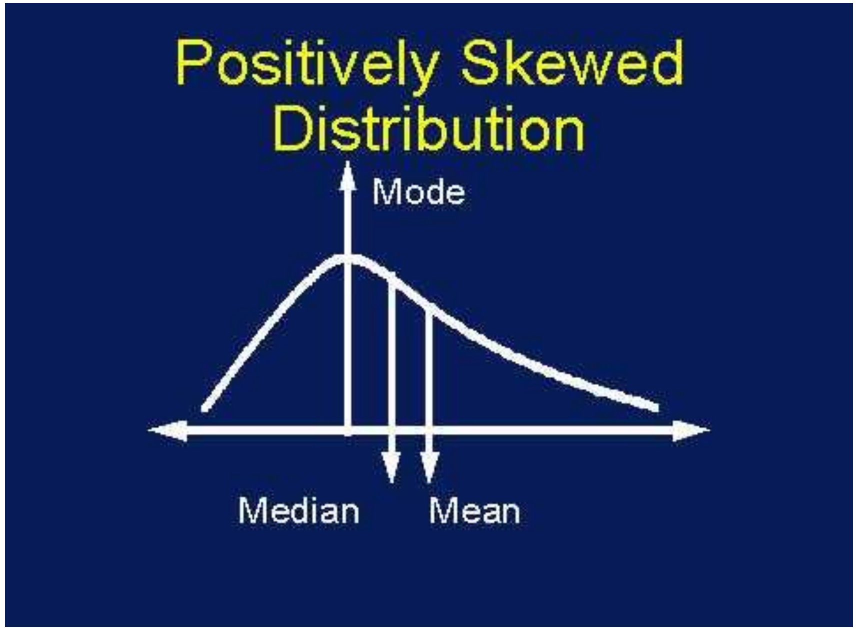
The level of measurement of the data.

- Mode---Nominal, Ordinal, Interval or Ratio
- Median--- Ordinal, Interval, or Ratio
- Mean---Interval or Ratio

Shape of the distribution: Skewness

- A measure of the lack of symmetry, or the lopsidedness of a distribution. (> or < 2)
- Use "median"

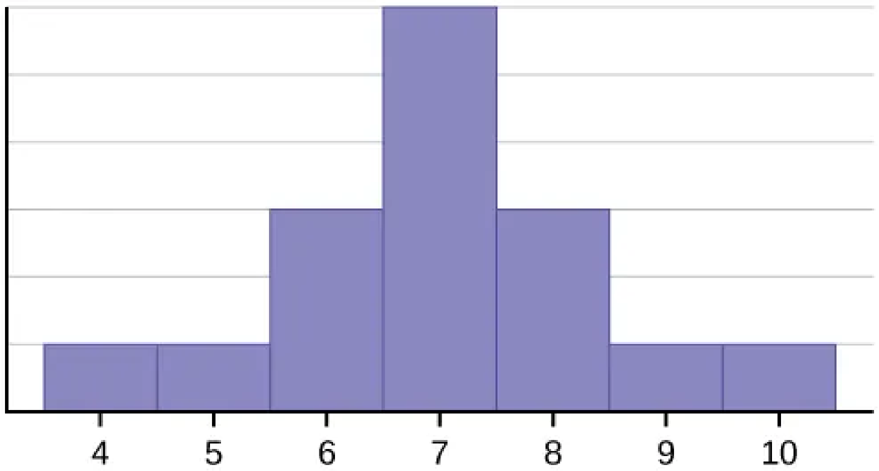




Consider the following data set.

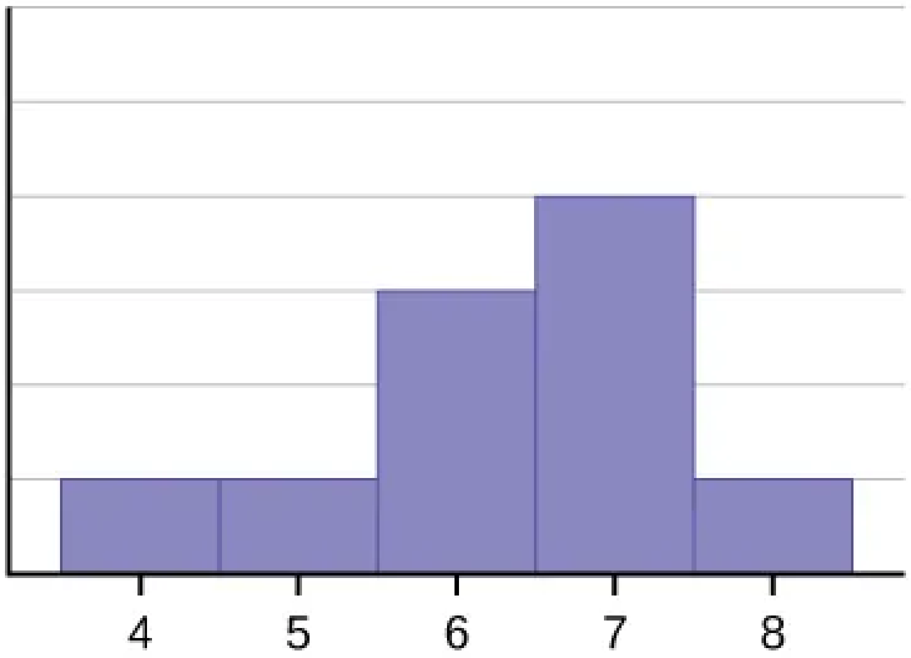
4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

This data set can be represented by following histogram. Each interval has width one, and each value is located in the middle of an interval.



- This histogram matches the supplied data. It consists of 7 adjacent bars with the x-axis split into intervals of 1 from 4 to 10. The heights of the bars peak in the middle and taper symmetrically to the right and left.
- The histogram displays a symmetrical distribution of data.
- A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data. In a perfectly symmetrical distribution, the mean and the median are the same.

The histogram for the data: 4; 5; 6; 6; 6; 7; 7; 7; 7; 8 shown below

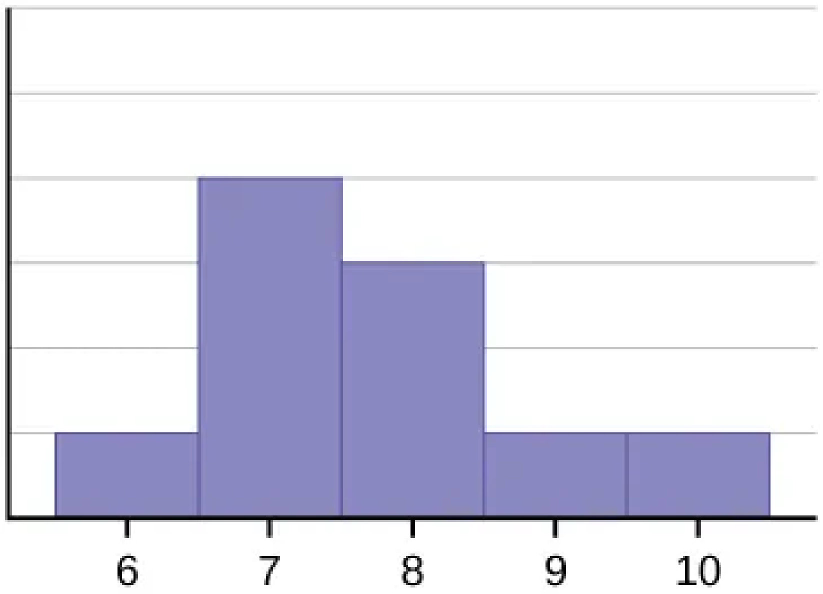


The mean is 6.3, the median is 6.5, and the mode is seven. Notice that the mean is less than the median, and they are both less than the mode. The mean and the median both reflect the skewing, but the mean reflects it more so.

A distribution of this type is called skewed to the left because it is pulled out to the left. We can formally measure the skewness of a distribution just as we can mathematically measure the center weight of the data or its general "speadness".

The greater the deviation from zero indicates a greater degree of skewness. If the skewness is negative then the distribution is skewed left as in above figure.

The histogram for the data: 6; 7; 7; 7; 7; 8; 8; 8; 9; 10 shown below, is also not symmetrical. It is skewed to the right.



This histogram matches the supplied data. It consists of 5 adjacent bars with the x-axis split into intervals of 1 from 6 to 10. The peak is to the left, and the heights of the bars taper down to the right.

The mean is 7.7, the median is 7.5, and the mode is seven. Of the three statistics, the mean is the largest, while the mode is the smallest. Again, the mean reflects the skewing the most.

The mean is affected by outliers that do not influence the mean. Therefore, when the distribution of data is skewed to the left, the mean is often less than the median. When the distribution is skewed to the right, the mean is often greater than the median.

$$a_3 = \sum \frac{(x_i - \bar{x})^3}{ns^3}$$

where,

s is the sample standard deviation of the data, Xi

$\bar{x}$ is the arithmetic mean

N is the sample size.

Measures of Variability

- Reflects how scores differ from one another.
  - spread
  - dispersion

- Example:

7, 6, 3, 3, 1

3, 4, 4, 5, 4,

4, 4, 4, 4, 4,

Measures of Variability

- Range

Highest score – lowest score

- Example:

7, 6, 3, 3, 1 ---- range = 6

3, 4, 4, 5, 4 ---- range = 2

4, 4, 4, 4, 4 ---- range = 0

- Variance
- Standard Deviation

Measures of Variability

- Range
- Standard Deviation
- Variance

Standard Deviation

- Standard Deviation: A measure of the spread of the scores around the mean.

- Average distance from the mean.

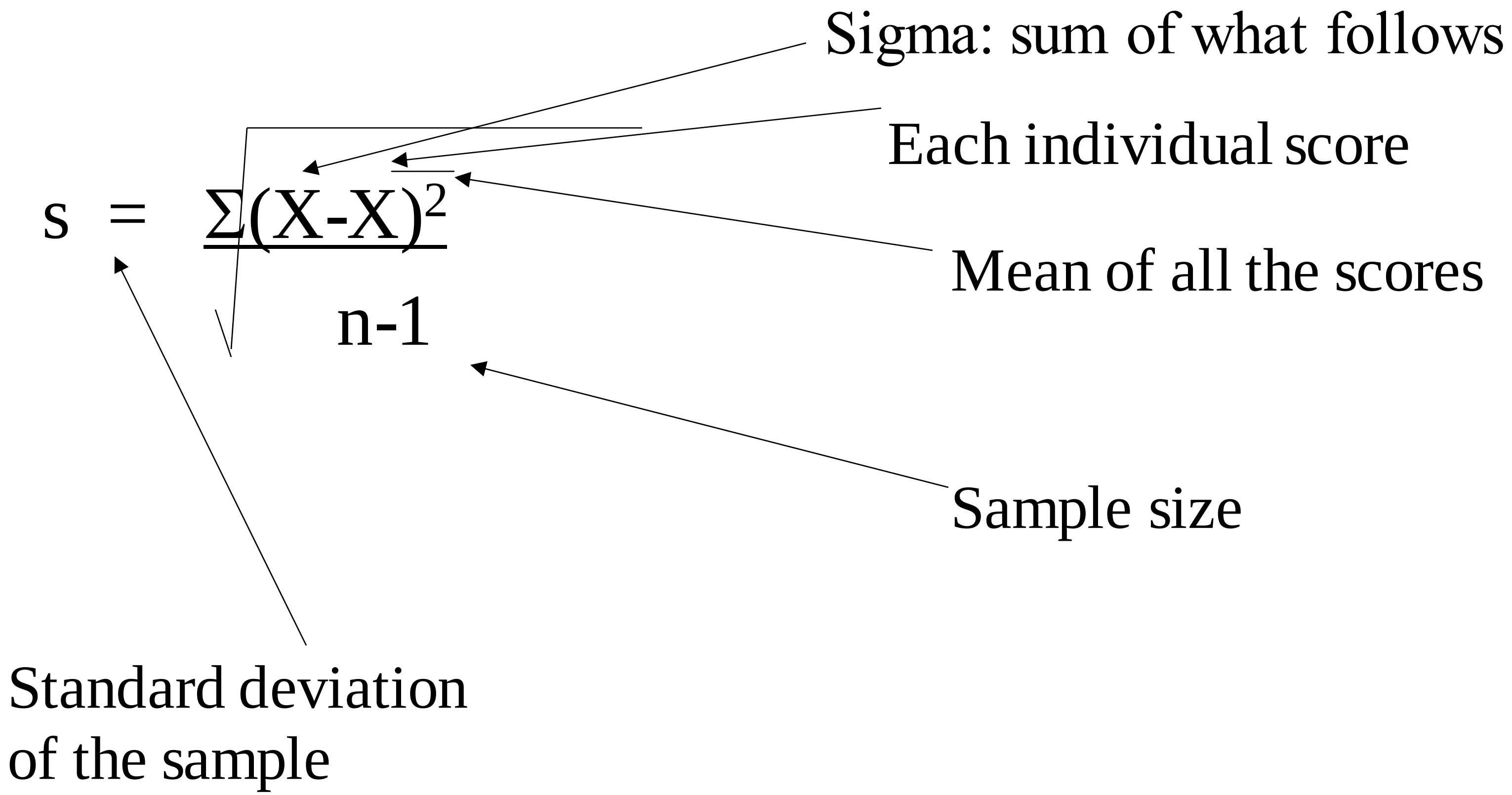
Example: Can you calculate the average distance of each score from the mean? (X=4)

7, 6, 3, 3, 1 (distance from the mean: 3,2,-1,-1,-3)

3, 4, 4, 5, 4, (distance from the mean: -1,0,0,1,0)

- You can't calculate the mean because the sum of the distance from the mean is always 0.

Formula for Standard Deviation



$$s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}$$

Sigma: sum of what follows

Each individual score

Mean of all the scores

Sample size

Standard deviation of the sample

In order to determine standard deviation:

1.Determine the mean (the average of all the numbers) by adding up all the data pieces (**xi**) and dividing by the number of pieces of data (**n**).

2.Subtract the mean (**x̄**) from each value.

3.Square each of those differences.

4.Determine the average of the squared numbers calculated in #3 to find the variance. (In sample sizes, subtract 1 from the total number of values when finding the average.)

5.Find the square root of the variance. That's the standard deviation!

Grading Tests

A class of students took a math test. Their teacher wants to know whether most students are performing at the same level, or if there is a high standard deviation.

1. The scores for the test were 85, 86, 100, 76, 81, 93, 84, 99, 71, 69, 93, 85, 81, 87, and 89. When the teacher adds them together, she gets 1279. She divides by the number of scores (15) to get the mean score.

$$1279 \div 15 = \mathbf{85.2\ (mean)}$$

2. 85.2 is a high score, but is everyone performing at that level? To find out, the teacher subtracts the mean from every test score.

85 - 85.2 = **-0.2**
86 - 85.2 = **0.8**
100 - 85.2 = **14.8**
76 - 85.2 = **-9.2**
81 - 85.2 = **-4.2**
93 - 85.2 = **7.8**
84 - 85.2 = **-1.2**
99 - 85.2 = **13.8**
71 - 85.2 = **-14.2**
69 - 85.2 = **-16.2**
93 - 85.2 = **7.8**
85 - 85.2 = **-0.2**
81 - 85.2 = **-4.2**
87 - 85.2 = **1.8**
89 - 85.2 = **3.8**

3. She squares each difference:

-0.2 x -0.2 = 0.04
0.8 x 0.8 = 0.64
14.8 14.8 = 219.04
-9.2 x -9.2 = 84.64
-4.2 x -4.2 = 17.64
7.8 x 7.8 = 60.84
-1.2 x -1.2 = 1.44
13.8 x 13.8 = 190.44
-14.2 x -14.2 = 201.64
-16.2 x -16.2 = 262.44
7.8 x 7.8 = 60.84
-0.2 x -0.2 = 0.04
-4.2 x -4.2 = 17.64
1.8 x 1.8 = 3.24
3.8 x 3.8 = 14.44

4. The teacher finds the variance, which is the average of the squares:

0.04 + 0.64 + 219.04 + 84.64 + 17.64 + 60.84 +1.44 +190.44 +201.64 +262.44 + 60.84 + 0.04 + 17.64 + 3.24 + 14.44 = 1135

830.64 ÷ 15 = 75.6 (variance)

5. Last, the teacher finds the square root of the variance:

Square root of 75.6 = 8.7 (standard deviation)

The standard deviation of these tests is 8.7 points out of 100. Since the variance is somewhat low, the teacher knows that most students are performing around the same level.

Variance

Variance: Standard deviation squared.

$S = \Sigma(X-\overline{X})^2 / n-1$

- Not likely to see the variance mentioned by itself in a report.
- Difficult to interpret.
- But it is important since it is used in many statistical formulas and techniques.

Example 1 : Find variance and standard deviation for dataset 17,10,9,14,13,17,12,20,14

Variance : 12.5

Standard Deviation : 3.536

Example 2 : Find variance and standard deviation for dataset 4, 2, 5, 8, and 6.

Variance : 20

Standard Deviation : 2.236

Data Processing

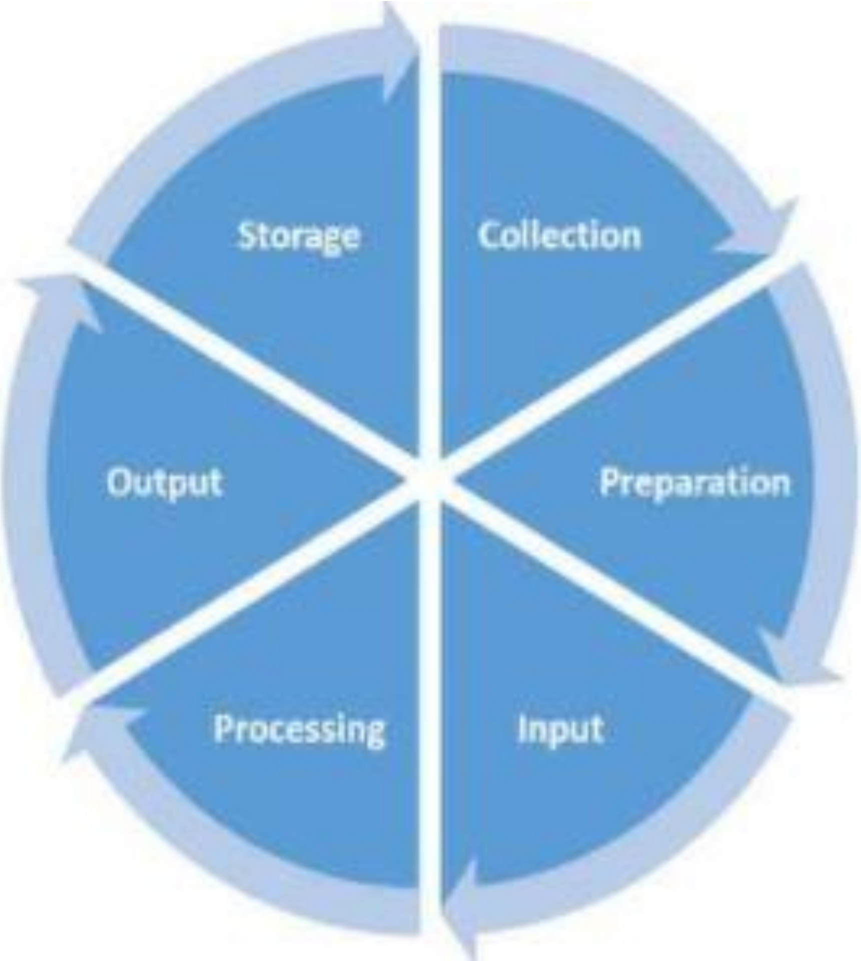
Data processing is the method of collecting raw data and translating it into usable information

It is usually performed in a step-by-step process by a team of data scientists and data engineers in an organization.

The raw data is collected, filtered, sorted, processed, analyzed, stored, and then presented in a readable format

By converting the data into readable formats like graphs, charts, and documents, employees throughout the organization can understand and use the data

Data processing is essential for organizations to create better business strategies and increase their competitive edge



Step 1 : Collection

The collection of raw data is the first step of the data processing cycle. The type of raw data collected has a huge impact on the output produced.

Step 2 : Preparation

Data preparation or data cleaning is the process of sorting and filtering the raw data to remove unnecessary and inaccurate data.

Step 3: Input

In this step, the raw data is converted into machine readable form and fed into the processing unit.

Step 4: Data Processing

In this step, the raw data is subjected to various data processing methods using machine learning and artificial intelligence algorithms to generate a desirable output.

Step 5: Output

The data is finally transmitted and displayed to the user in a readable form like graphs, tables, vector files, audio, video, documents, etc.

Step 6: Storage

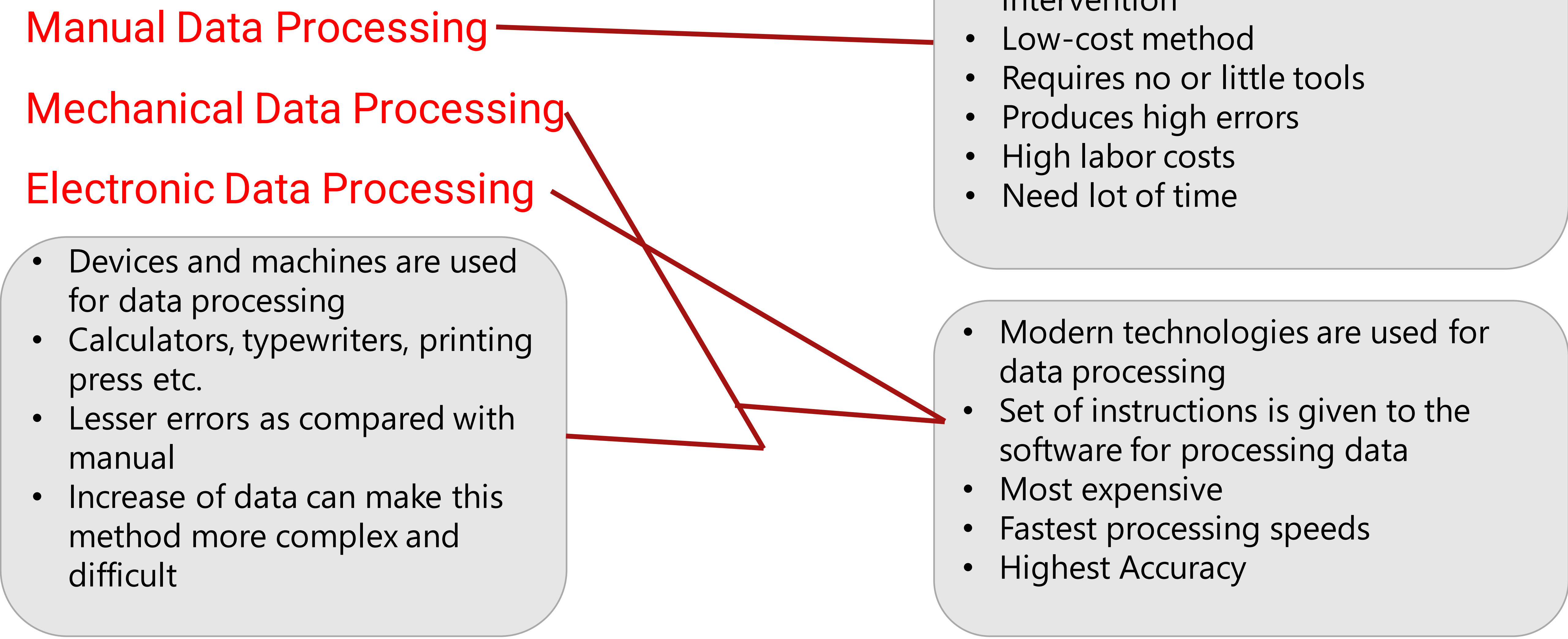
The last step of the data processing cycle is storage, where data and metadata are stored for further use.

Types of Data Processing

1. Batch Processing
2. Real-time Processing
3. Online Processing
4. Multiprocessing
5. Time-sharing

Type	Uses
Batch Processing	Data is collected and processed in batches. Used for large amounts of data. Eg: payroll system
Real-time Processing	Data is processed within seconds when the input is given. Used for small amounts of data. Eg: withdrawing money from ATM
Online Processing	Data is automatically fed into the CPU as soon as it becomes available. Used for continuous processing of data. Eg: barcode scanning
Multiprocessing	Data is broken down into frames and processed using two or more CPUs within a single computer system. Also known as parallel processing. Eg: weather forecasting
Time-sharing	Allocates computer resources and data in time slots to several users simultaneously.

Types of Data Processing



Manual Data Processing

Mechanical Data Processing

Electronic Data Processing

- Handled Manually
- Every process is done with human intervention
- Low-cost method
- Requires no or little tools
- Produces high errors
- High labor costs
- Need lot of time

- Devices and machines are used for data processing
- Calculators, typewriters, printing press etc.
- Lesser errors as compared with manual
- Increase of data can make this method more complex and difficult

- Modern technologies are used for data processing
- Set of instructions is given to the software for processing data
- Most expensive
- Fastest processing speeds
- Highest Accuracy

Examples of Data Processing

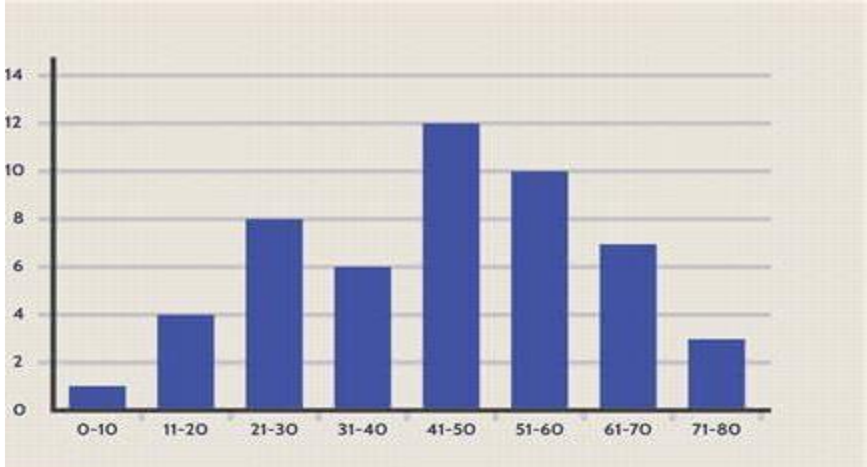
- A stock trading software that converts millions of stock data into a simple graph
- An e-commerce company uses the search history of customers to recommend similar products
- A digital marketing company uses demographic data of people to strategize location-specific campaigns
- A self-driving car uses real-time data from sensors to detect if there are pedestrians and other cars on the road

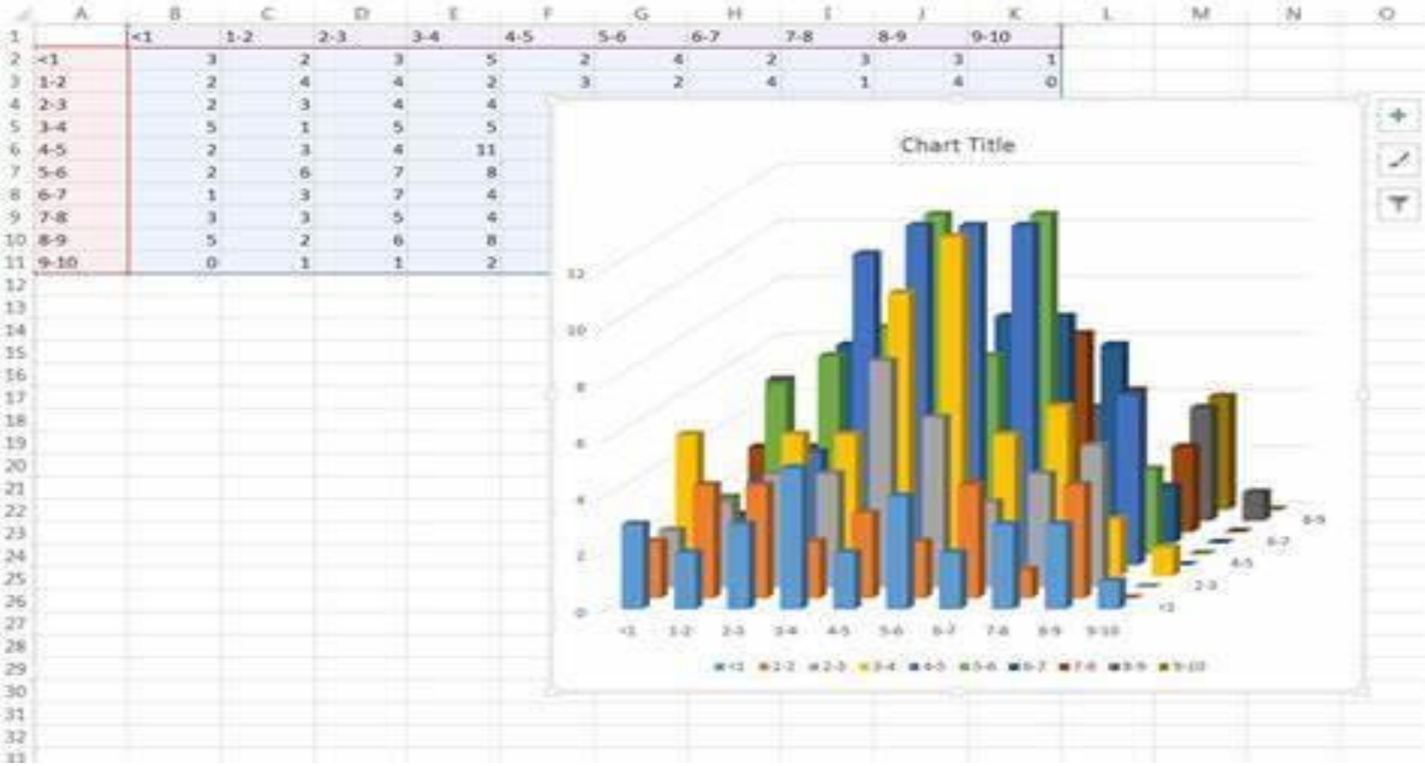
Histogram

A histogram is a common data analysis tool in the business world. It’s a column chart that shows the frequency of the occurrence of a variable in the specified range.

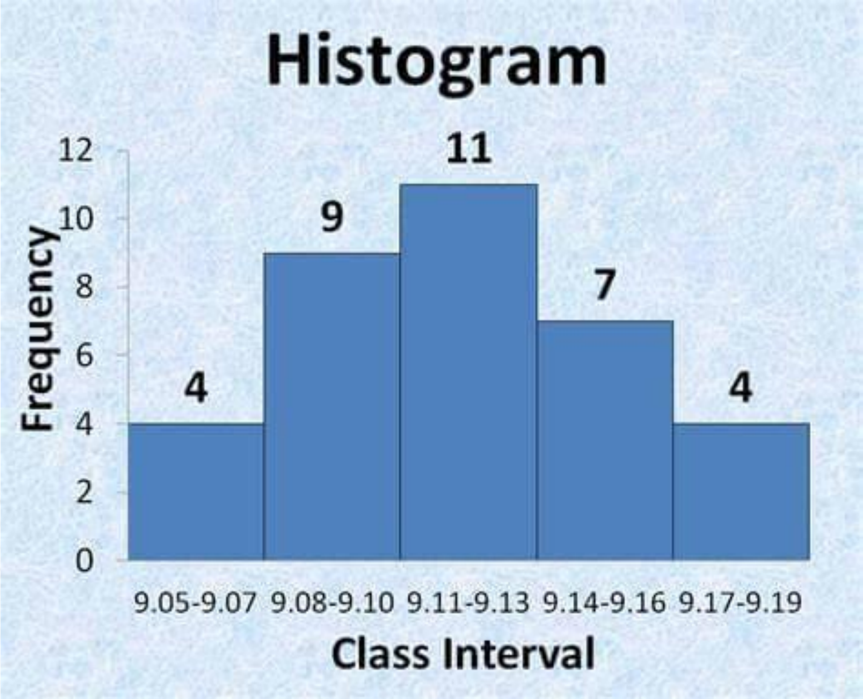
Histogram is a graphical representation, like a bar chart in structure, that organizes a group of data points into user-specified ranges.

Histogram graph is a type of graph that uses rectangular bars to represent the Frequency of discrete and continuous data. The rectangular bars represent the number of data points that fall within a certain class interval.



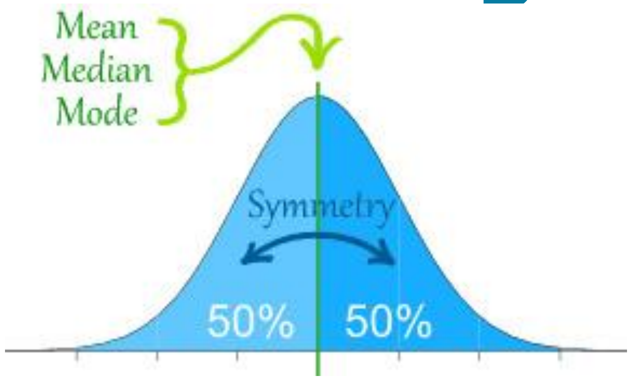


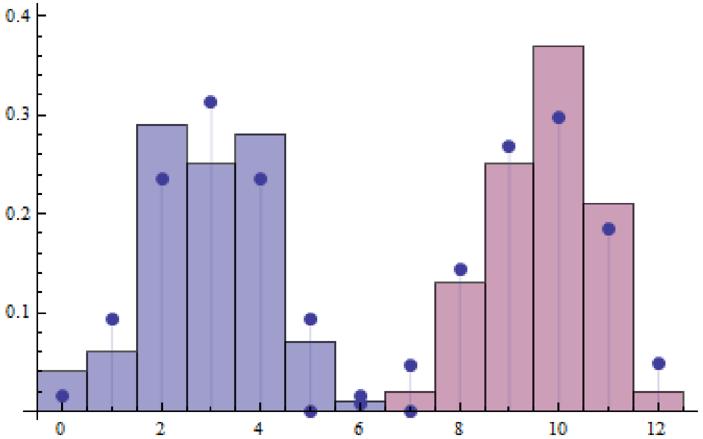
	<1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10
<1	3	2	3	5	2	4	2	3	3	1
1-2	2	4	4	2	3	2	4	1	4	0
2-3	2	3	4	4						
3-4	5	1	5	5						
4-5	2	3	4	11						
5-6	2	6	7	8						
6-7	1	3	7	4						
7-8	3	3	5	4						
8-9	5	2	6	8						
9-10	0	1	1	2						

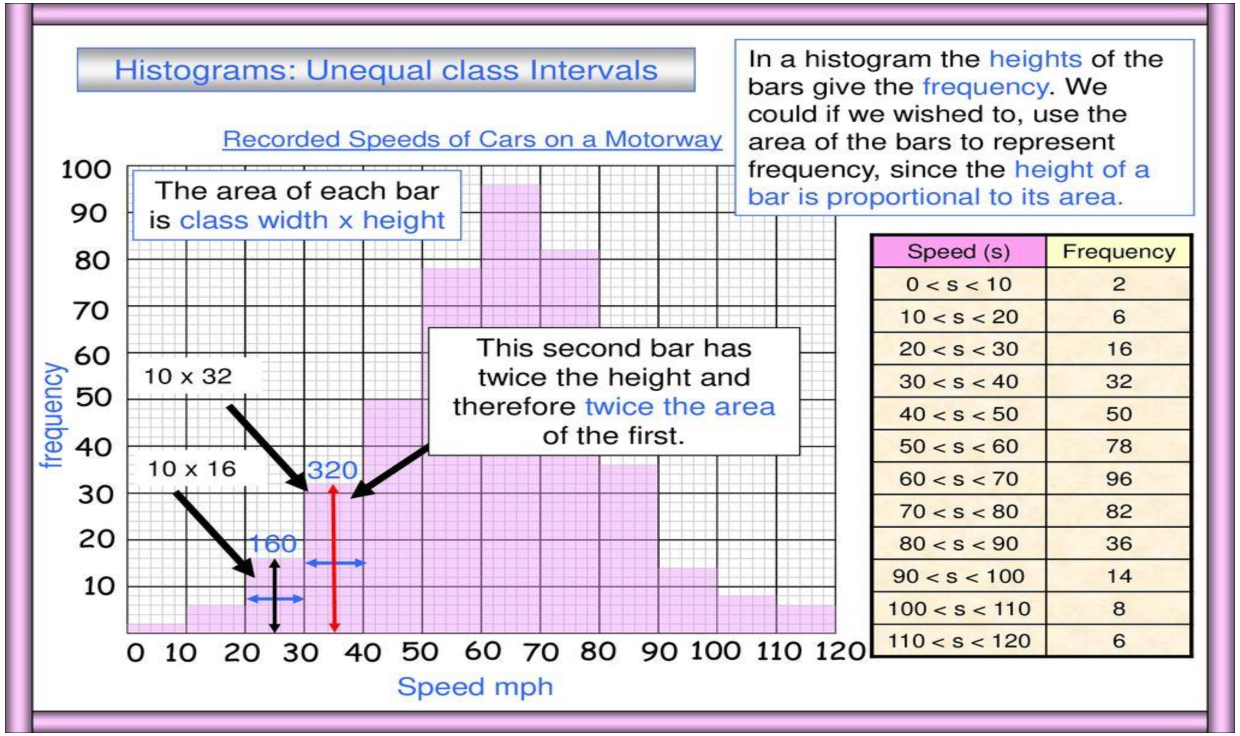


Types of Distribution in Histogram

1. Regular Distribution
2. Binary Distribution
3. Unbalanced Distribution
4. Random Distribution
5. Uniform Distribution







Histograms: Unequal class Intervals

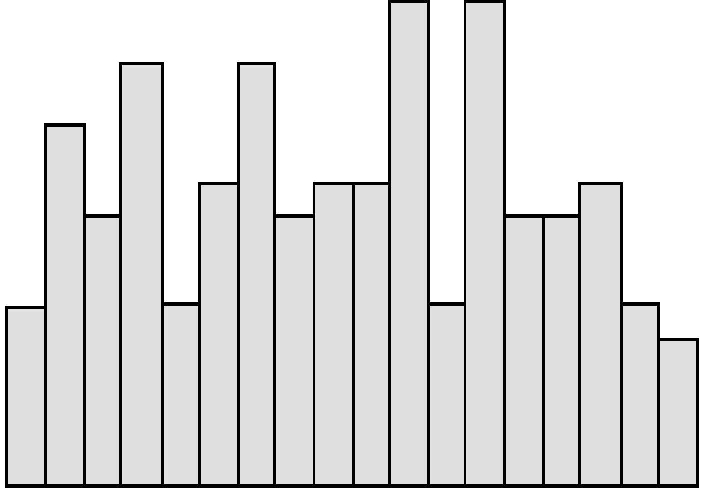
In a histogram the heights of the bars give the frequency. We could if we wished to, use the area of the bars to represent frequency, since the height of a bar is proportional to its area.

Recorded Speeds of Cars on a Motorway

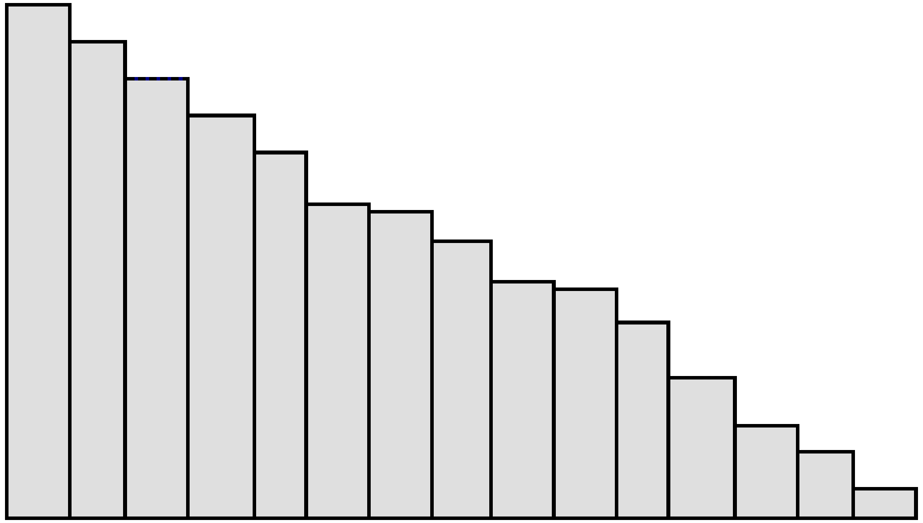
The area of each bar is class width x height

This second bar has twice the height and therefore twice the area of the first.

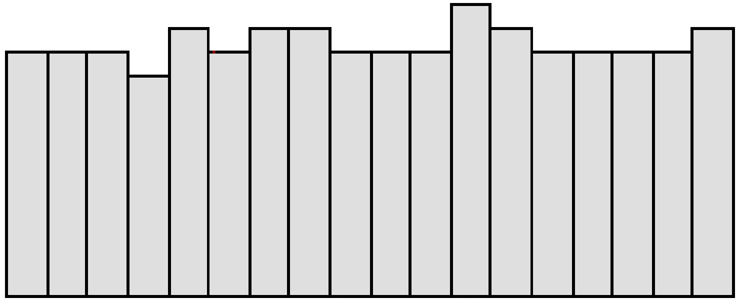
Speed (s)	Frequency
0 < s < 10	2
10 < s < 20	6
20 < s < 30	16
30 < s < 40	32
40 < s < 50	50
50 < s < 60	78
60 < s < 70	96
70 < s < 80	82
80 < s < 90	36
90 < s < 100	14
100 < s < 110	8
110 < s < 120	6



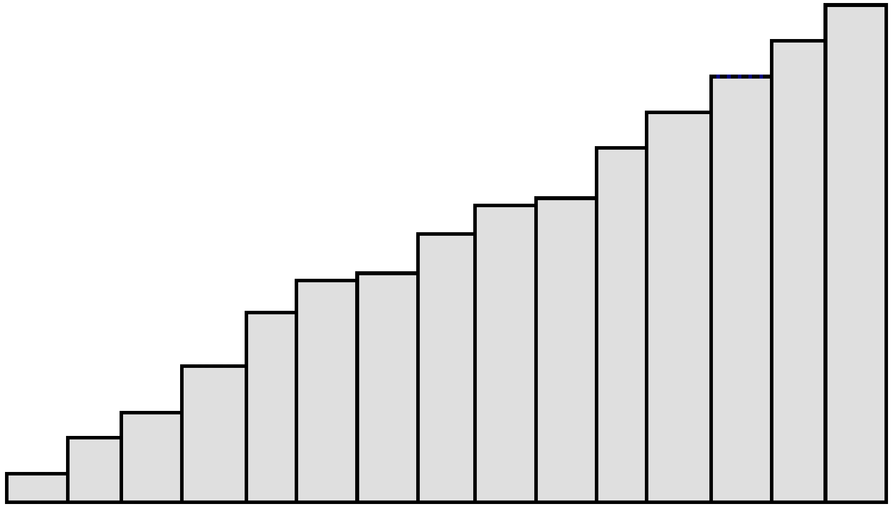
Random Distribution



Skewed Left



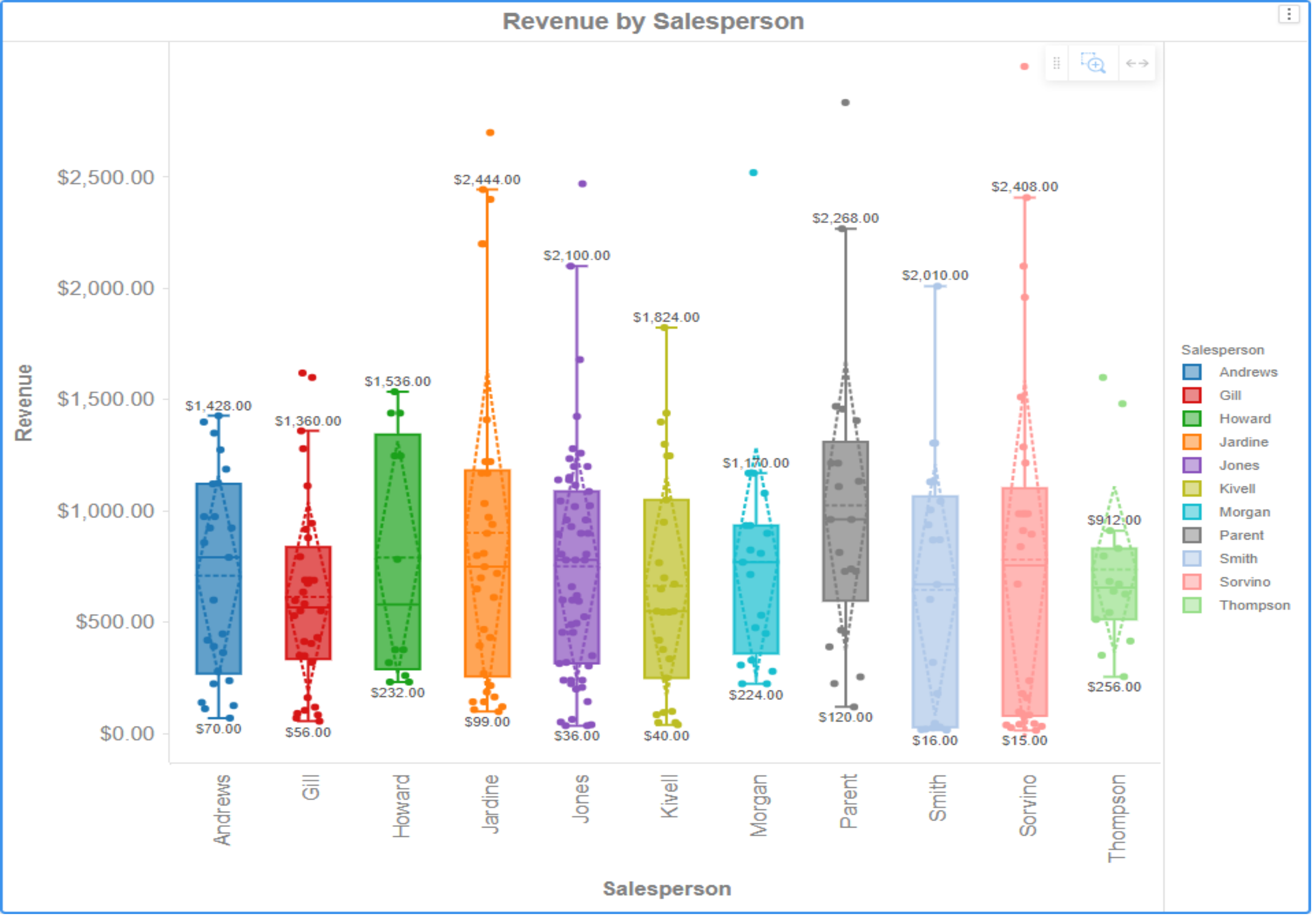
Uniform Distribution

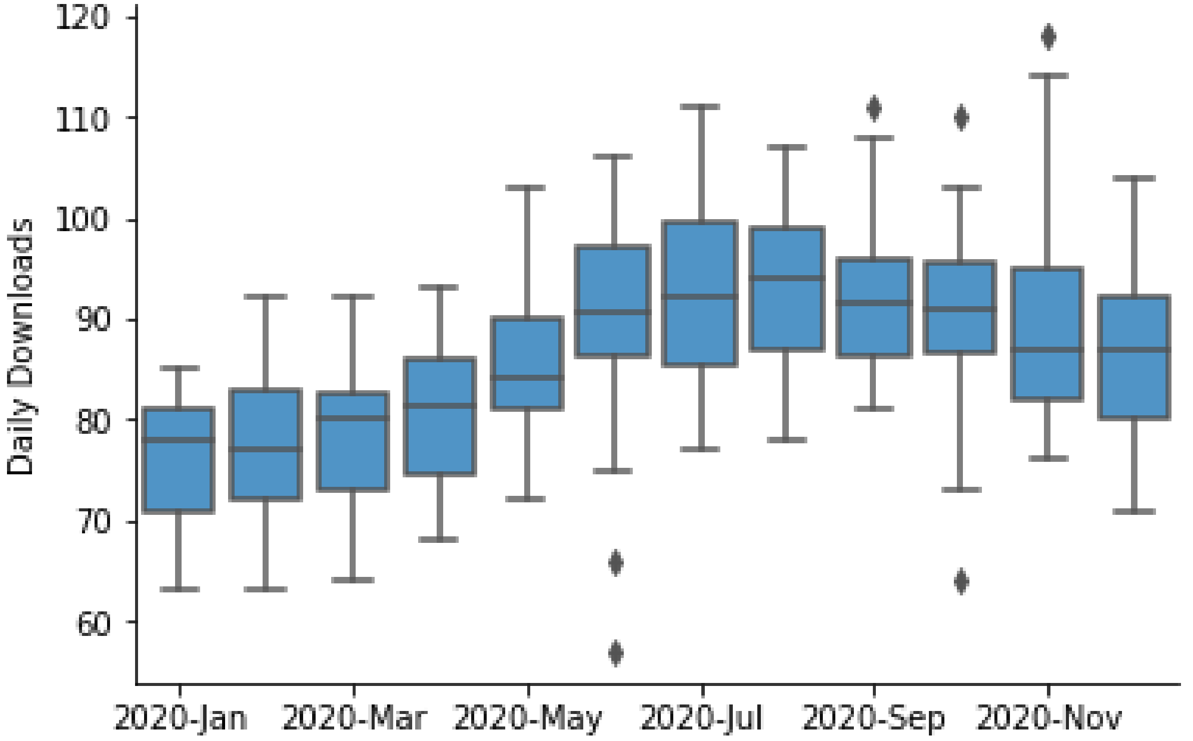


Skewed Right

Box Plots

- A box plot visualization allows you to examine the distribution of data.
- One box plot appears for each attribute element.
- Each box plot displays the minimum, first quartile, median, third quartile, and maximum values.
- In addition, you can choose to display the mean and standard deviation as dashed lines.
- Outliers appear as points in the visualization.
- You can adjust the spacing between points to avoid overlap.
- A box plot must include at least one metric and at least one attribute.

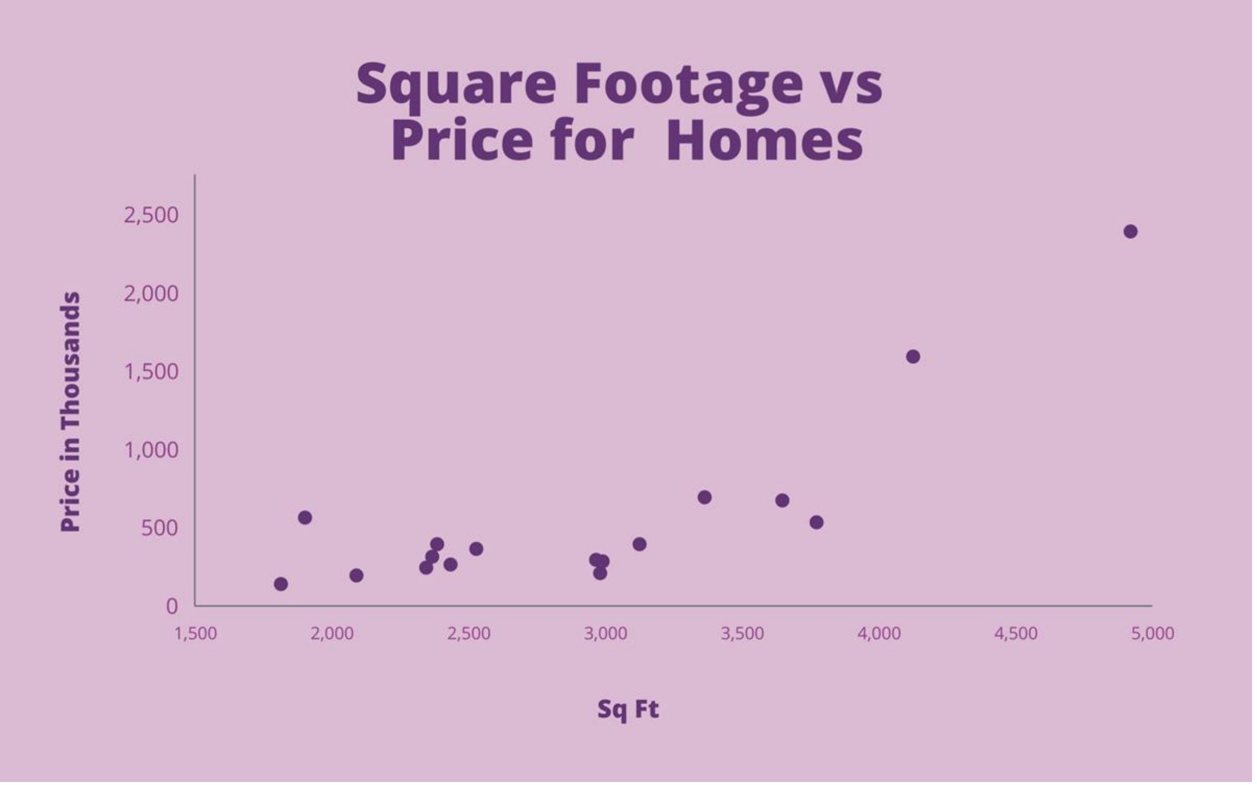

Revenue by Salesperson
Revenue
$2,500.00
$2,000.00
$1,500.00
$1,000.00
$500.00
$0.00
$1,428.00
$70.00
$1,360.00
$56.00
$1,536.00
$232.00
$2,444.00
$99.00
$2,100.00
$36.00
$1,824.00
$40.00
$1,170.00
$224.00
$2,268.00
$120.00
$2,010.00
$16.00
$2,408.00
$15.00
$912.00
$256.00
Andrews
Gill
Howard
Jardine
Jones
Kivell
Morgan
Parent
Smith
Sorvino
Thompson
Salesperson
Salesperson
Andrews
Gill
Howard
Jardine
Jones
Kivell
Morgan
Parent
Smith
Sorvino
Thompson


Daily Downloads
120
110
100
90
80
70
60
2020-Jan
2020-Mar
2020-May
2020-Jul
2020-Sep
2020-Nov

Scatter Plots

- A scatter plot is a type of data visualization that shows the relationship between different variables. This data is shown by placing various data points between an x- and y-axis.
- Essentially, each of these data points looks “scattered” around the graph, giving this type of data visualization its name.
- Scatter plots can also be known as scatter diagrams or x-y graphs, and the point of using one of these is to determine if there are patterns or correlations between two variables.

Use a scatter plot when your independent variable has multiple values for your dependent variable.



- The two variables are the square footage of a home versus its price.
- As the x-axis goes from the smallest size to the largest, we can see that there is a slight positive correlation showing that as square footage increases, so does the price.
- There could be other factors contributing to this, like location or recent renovations, but we can see from this scatter diagram that there is a correlation between the square footage and home cost.

Linear or Nonlinear: A linear correlation forms a straight line in its data points while a nonlinear correlation might have a curve or other form within the data points.

Strong or Weak: A strong correlation will have data points close together while a weak correlation will have data points that are further apart.

Positive or Negative: A positive correlation will point up (i.e., the x- and y-values are both increasing) while a negative correlation will point down (i.e., the x-values are increasing while the corresponding y-values are decreasing).

What is Exploratory Data Analysis(EDA)?

EDA is an approach for data analysis using variety of techniques to gain insights about the data.

Basic steps in any exploratory data analysis:

- Cleaning and preprocessing
- Statistical Analysis
- Visualization for trend analysis, anomaly detection, outlier detection (and removal).

Importance of EDA

Improve understanding of variables by extracting averages, mean, minimum, and maximum values, etc.

Discover errors, outliers, and missing values in the data.

Identify patterns by visualizing data in graphs such as bar graphs, scatter plots, heatmaps and histograms.

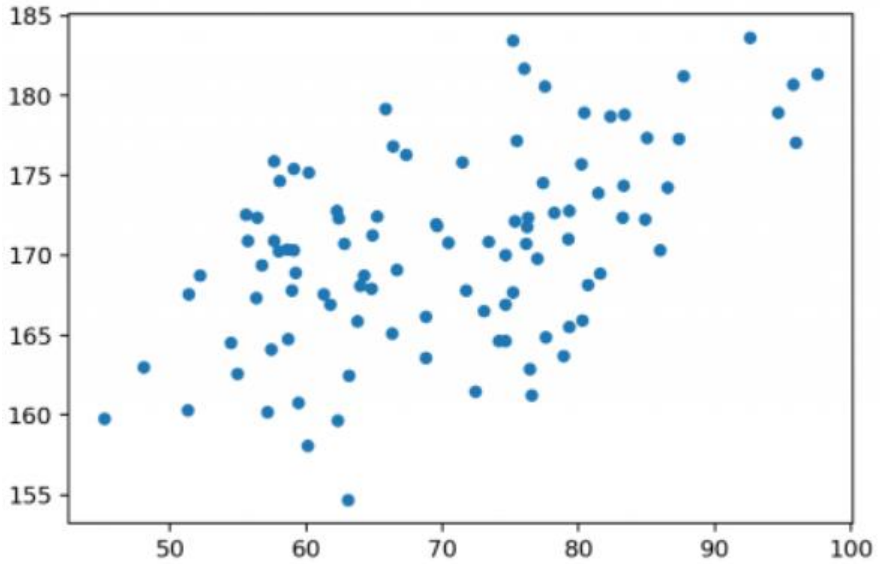
Visualization

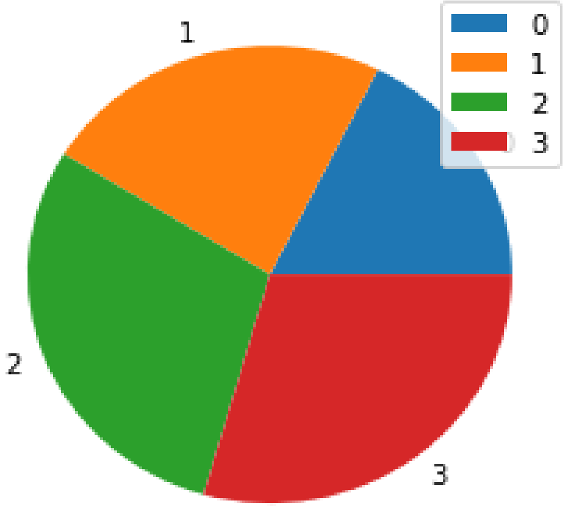
- Univariate: Looking at one variable/column at a time
  - Bar-graph
  - Histograms
  - Boxplot
- Multivariate : Looking at relationship between two or more variables
  - Scatter plots
  - Pie plots
  - Heatmaps(seaborn)

Scatterplot, Pieplot

- Scatterplot : Shows the data as a collection of points.
  - Syntax: dataframe.plot.scatter(x = 'x_column_name', y = 'y_columnn_name’)

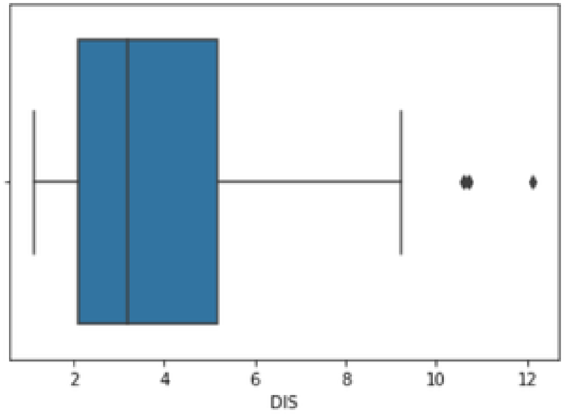
- Pie plot : Proportional representation of the numerical data in a column.
  - Syntax: dataframe.plot.pie(y=‘column_name’)

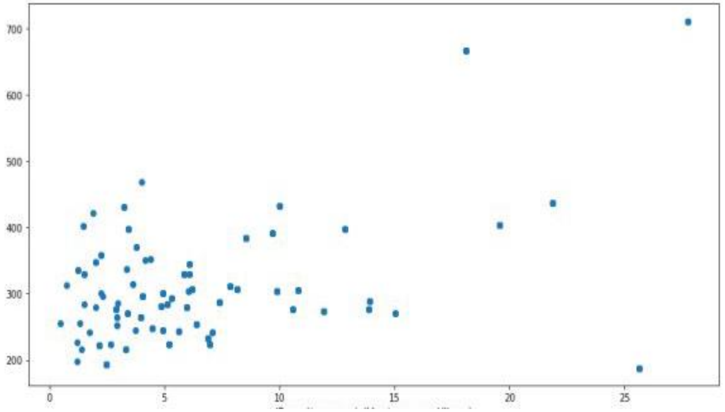




Outlier detection

- An outlier is a point or set of data points that lie away from the rest of the data values of the dataset..
- Outliers are easily identified by visualizing the data.
- For e.g.
  - In a boxplot, the data points which lie outside the upper and lower bound can be considered as outliers
  - In a scatterplot, the data points which lie outside the groups of datapoints can be considered as outliers





Outlier removal

Calculate the IQR as follows:

- ➢Calculate the first and third quartile (Q1 and Q3)
- ➢Calculate the interquartile range, **IQR = Q3-Q1**
- ➢Find the lower bound which is **Q1*1.5**
- ➢Find the upper bound which is **Q3*1.5**
- ➢Replace the data points which lie outside this range.
- ➢They can be replaced by mean or median.