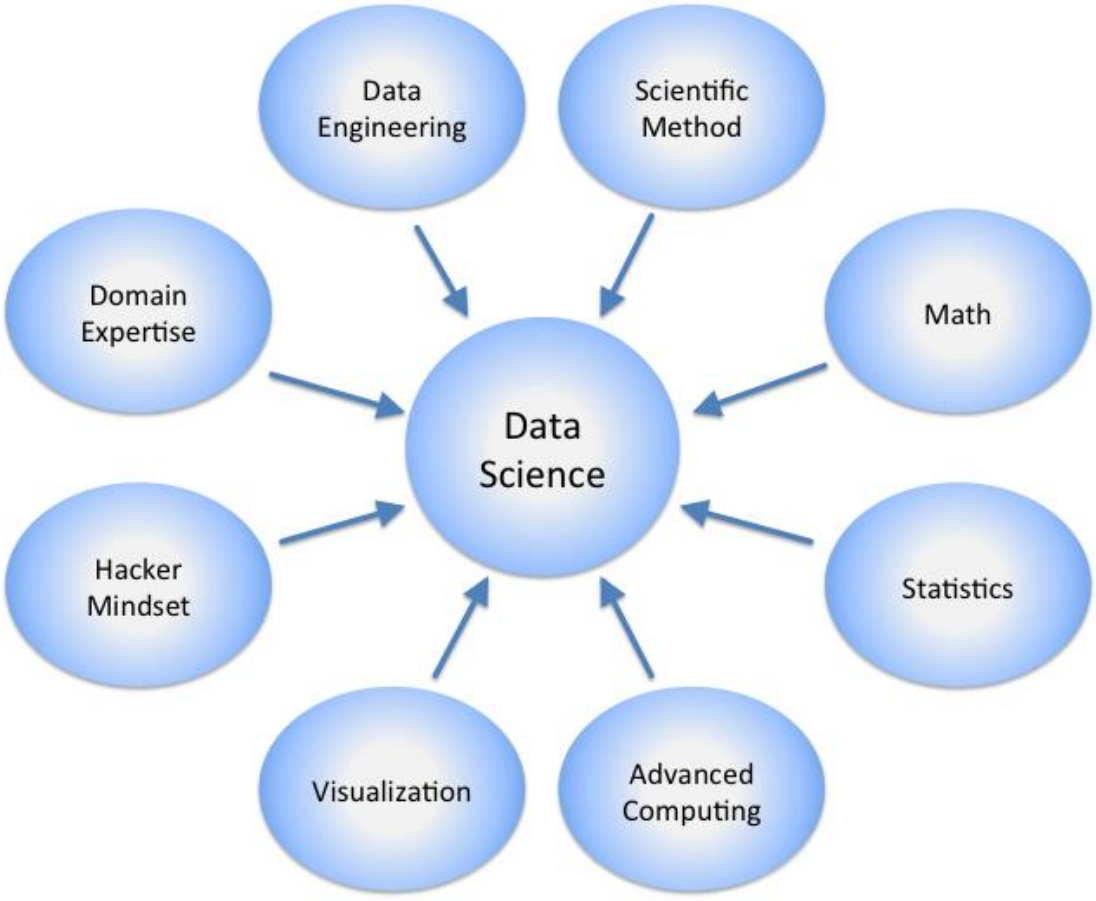
Data Science

Introduction to Data Science KD

Lesson-1

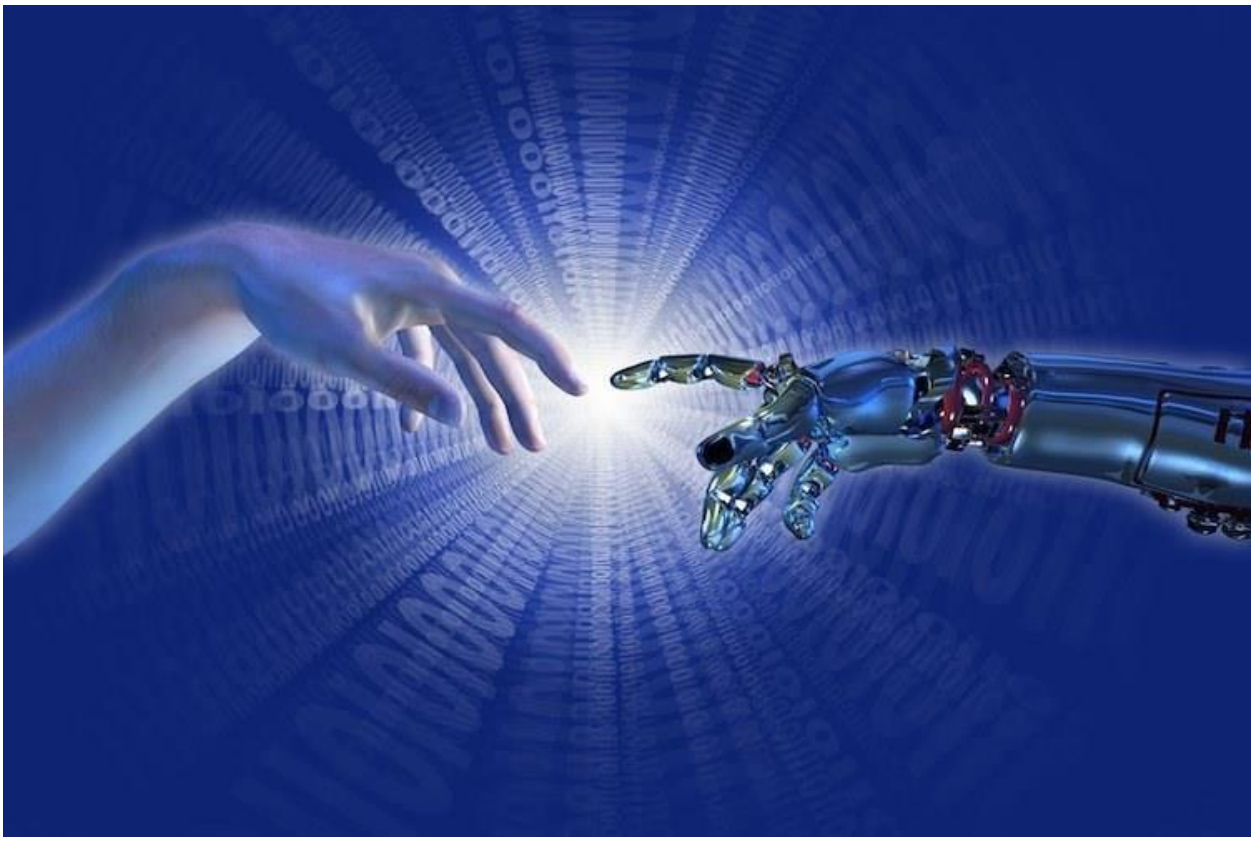


"Data Science couldn't have happened without..." : Tim Berners- Lee

Data Science couldn't have happened without : Internet

According to me data science couldn't have happened without "The invention of Internet". We could produce data/information,collect them using various resources (written form), could use some basic processing using Tabulating machine etc., could make predictions and take decisions as well to some extent using manual resources of collected data, but gathering relevant data/information from various sources, finding some insights from them collectively and making any valuable future prediction quite quickly wouldn't have been possible without internet.

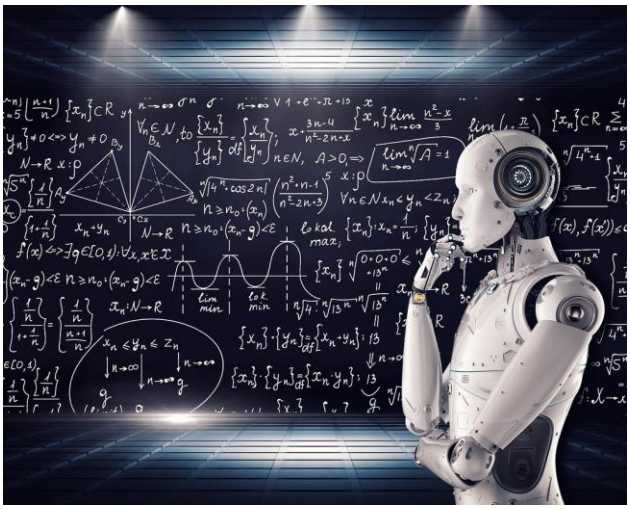
Lesson-2 Reasons For Growth in Data



1 - ***Data growth refers to the constant increase in the amount of data being produced by modern technological society as well as the increase in the amount of data an enterprise must store.***

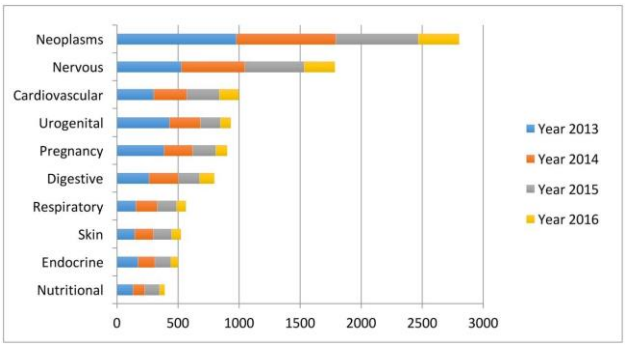
The amount of data generated annually has grown year-over-year since 2010. In fact, it is estimated that 90% of the data ever created has been created in the last 2 years . We can think about some reasons of data growth-

1-Technological -



Example of Technological Growth- Artificial Intelligence and Machine Learning.

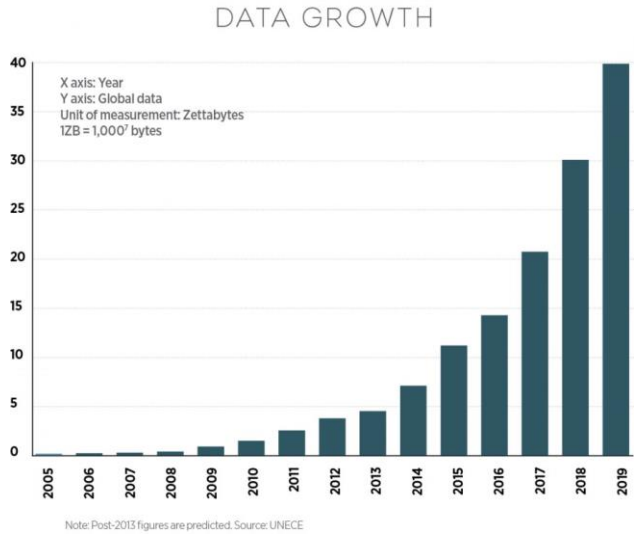
2- Societal -

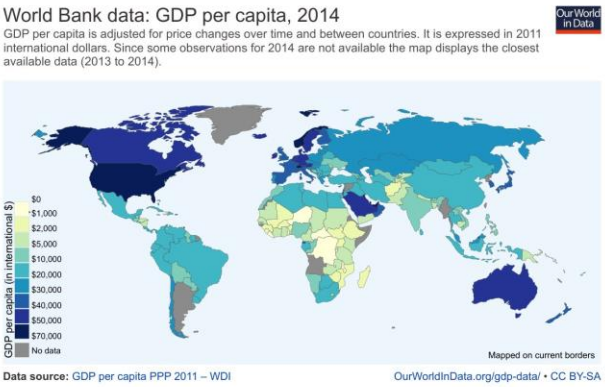


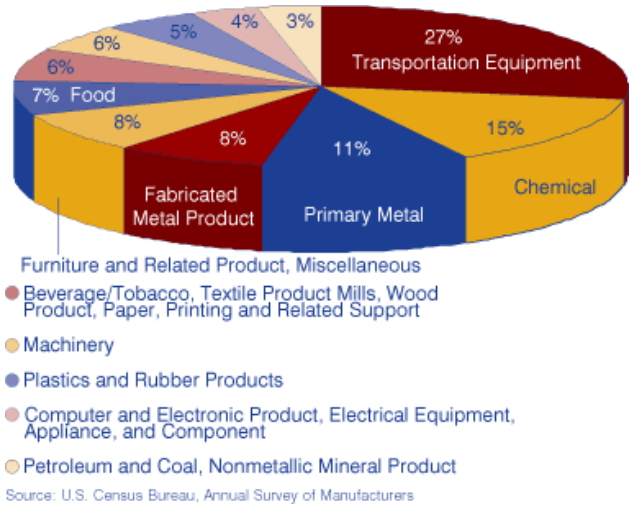
Example of Societal Growth- Analyzing health data and extracting the impact of particular medicine.

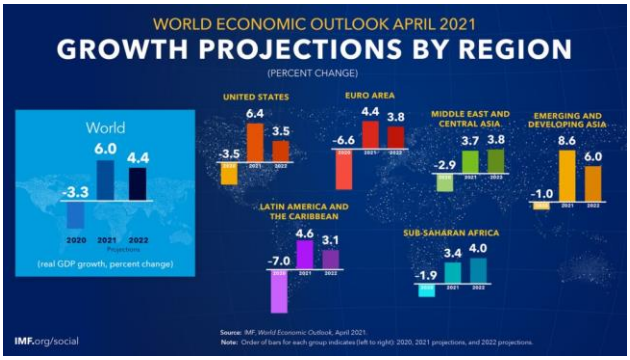
3- Economic-


ECONOMIC
GROWTH


DATA GROWTH
X axis: Year
Y axis: Global data
Unit of measurement: Zettabytes
1ZB = 1,000⁷ bytes
40
35
30
25
20
15
10
5
0
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Note: Post-2013 figures are predicted. Source: UNECE





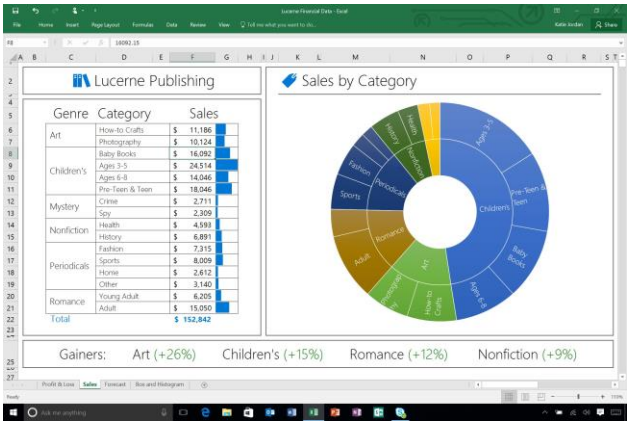


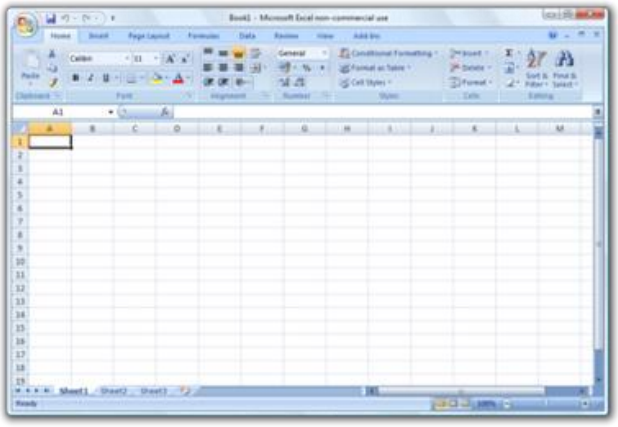
Example of Economic Growth-

- ***Fostering new products,***
- ***processes,***

- *organisational methods and market trends,*
- *enabling entirely new business models,*
- *making predictions*
- *taking decisions for future development.*

'Excel for Data Science'





Excel Functions-

All Excel functions have the same structure.

A function is a predefined formula that performs calculations using specific values in a particular order. This order is called structure of a function.

SUM FUNCTION-

The SUM function adds values. We can add individual values, cell references or ranges or a mix of all three. For example: =SUM(A2:A10) Adds the values in cells A2:10.

MAX FUNCTION-

The MAX function in Excel returns the highest value in a set of data that we specify. The syntax is as follows: MAX(number1, [number2], …) Where number can be represented by a numeric value, array, named range, a reference to a cell or range containing numbers.

MIN FUNCTION-

The MIN function is a premade function in Excel, which finds the lowest number in a range. The function ignores cells with text. It will only work for cells with numbers.

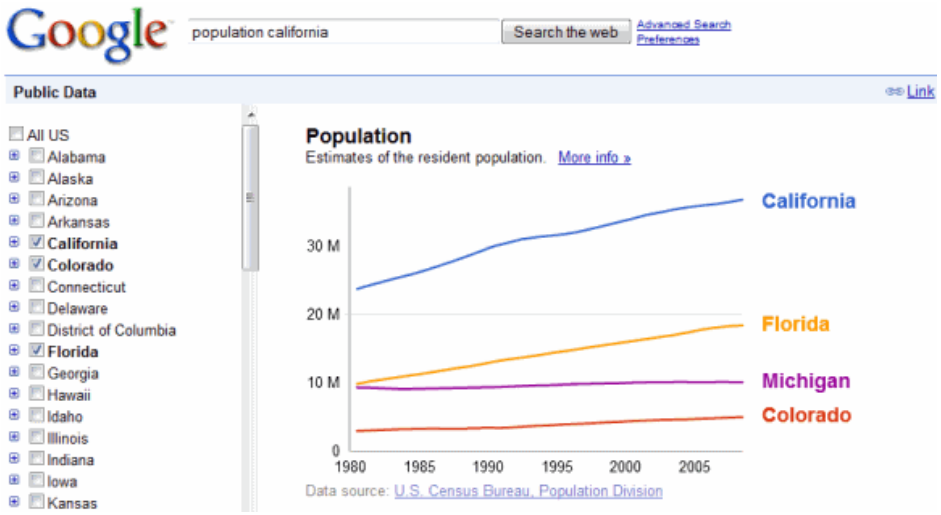
AVERAGE FUNCTION-

This function returns the average (arithmetic mean) of the arguments. For example, if the range A1:A20 contains numbers, the formula =AVERAGE(A1:A20) returns the average of those numbers.

COUNT FUNCTION-

Use the COUNT function to get the number of entries in a number field that is in a range or array of numbers. For example, you can enter the following formula to count the numbers in the range A1:A20: =COUNT(A1:A20). In this example, if five of the cells in the range contain numbers, the result is 5.

Lesson- 7 PUBLIC DATA


Google
population california
Search the web
Advanced Search
Preferences
Public Data
Link
All US
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Population
Estimates of the resident population. More info »
30 M
20 M
10 M
0
1980
1985
1990
1995
2000
2005
California
Florida
Michigan
Colorado
Data source: U.S. Census Bureau, Population Division


Creating
Value
through
Open Data

Scotland's Open Data Portals

Fork me on GitHub

This page provides a list of websites in Scotland that are partially or wholly focussed on making open data available. The list only includes sites that provide access to a **collection** of datasets, rather than just a single dataset. A further criterion for inclusion is that access to data is made available in at least one of the following modes:

- bulk download in a machine readable format (e.g., CSV, RSS, XML);
- via a documented API; or
- as a SPARQL endpoint.

This list is preliminary and additions are very welcome. Send suggestions by email to okfnscot [AT] gmail.com, or fork the CSV data file and send a pull request.

Open Glasgow Data Portal CSV JSON XML	Glasgow City Council data is now open by default. Datasets are categorized into Health, Living, Environment, Transportation, Energy, Geography, Demographics, Economy, Active Travel, and Public Safety.
City of Edinburgh Council Data CSV	The City of Edinburgh Council gathers many types of data to help deliver and inform its work, and will publish it where possible to add value for work in communities, organisations and businesses. The most recent versions are on Github.

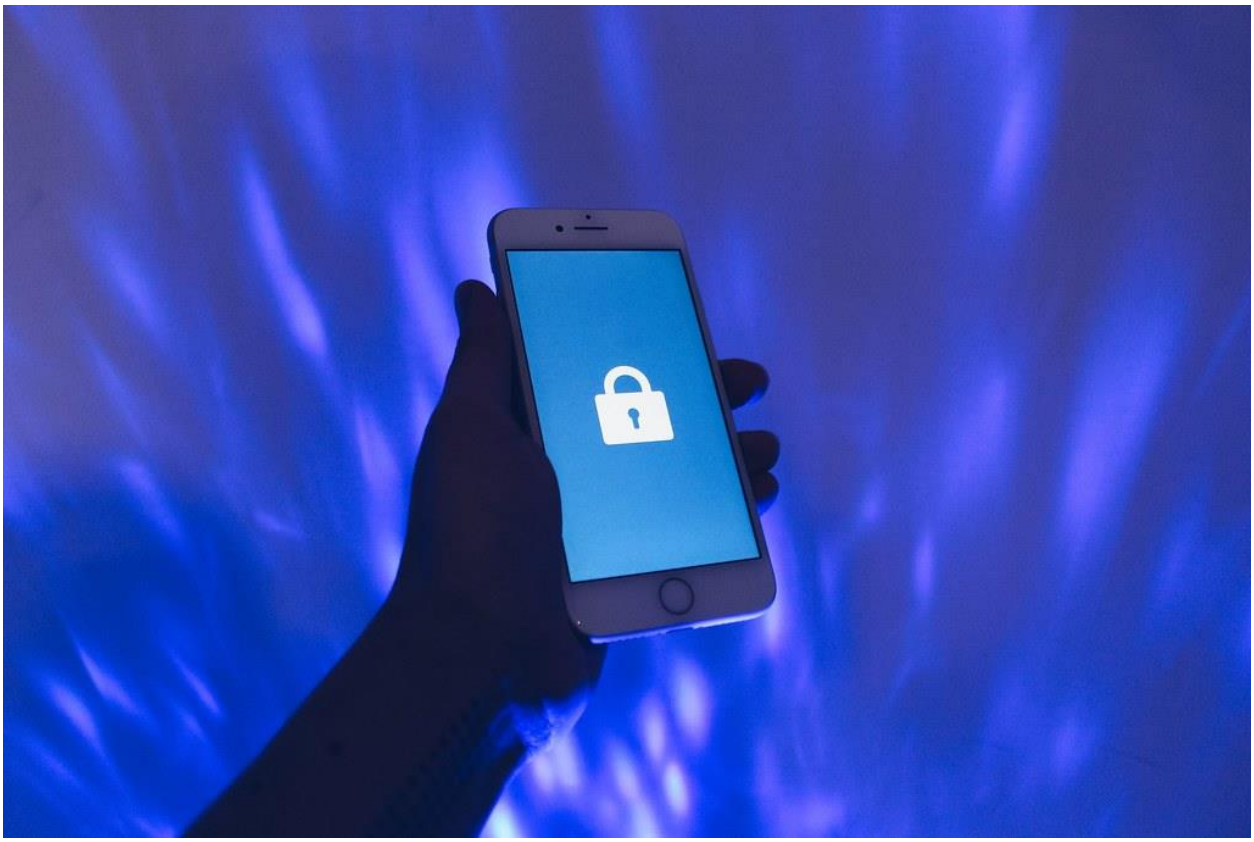
Definition of Public data:-

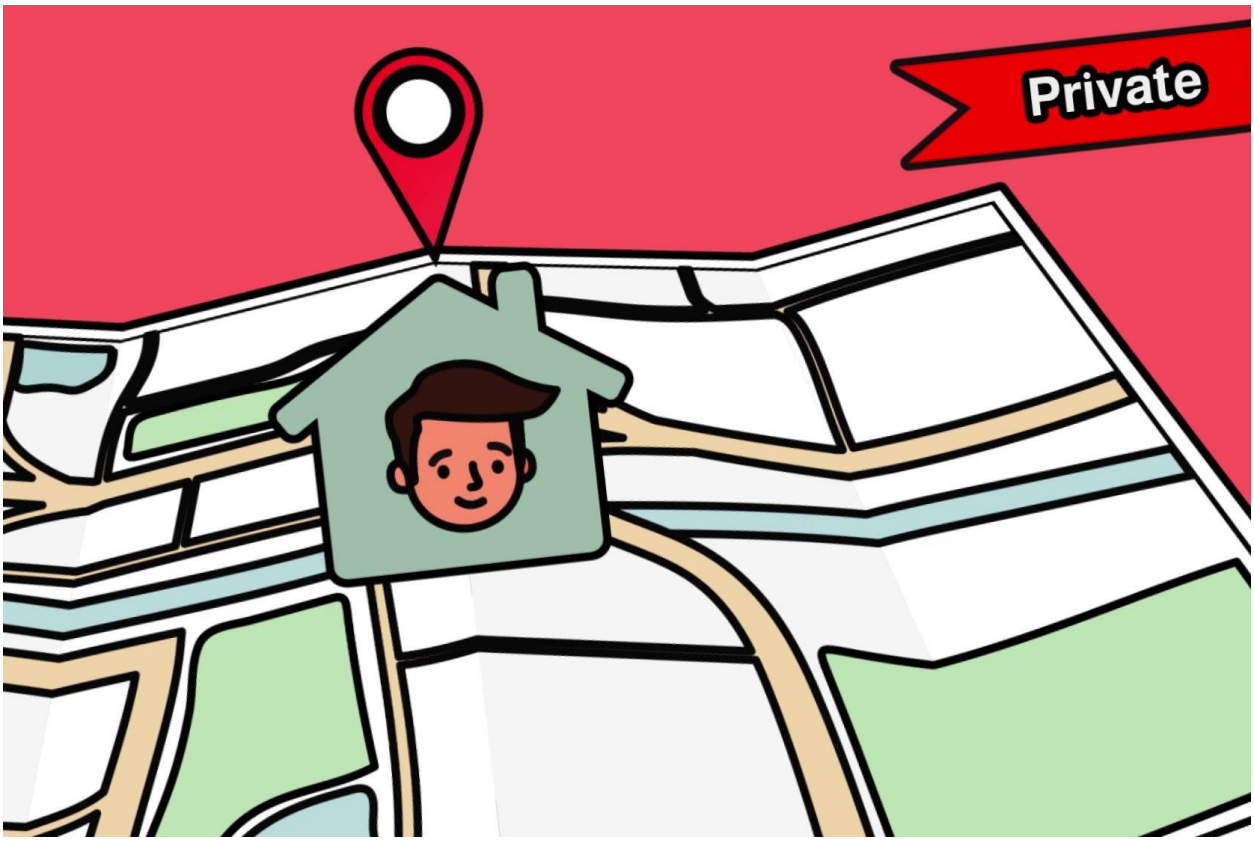
Public data is information that can be freely shared, used, reused and redistributed without restriction.

Examples:-

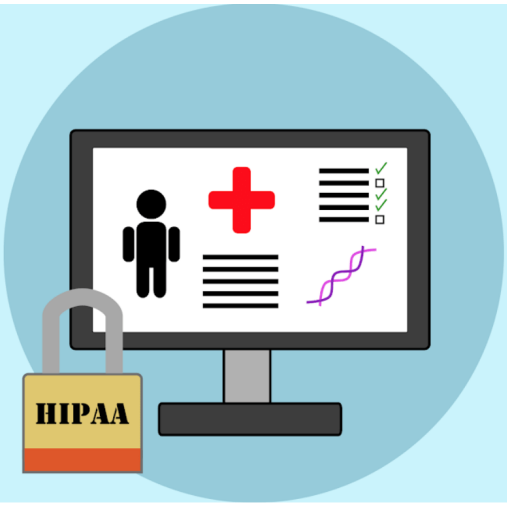
- ***Google Trends***
- ***National Climatic Data Center***
- ***Global Health Observatory Data***

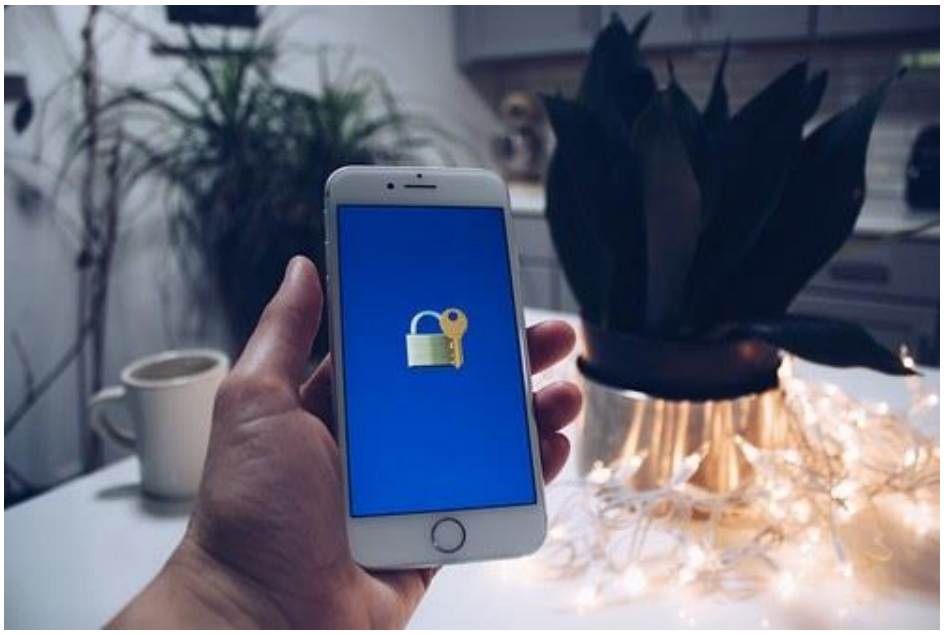
PRIVATE DATA





Health Data Privacy





Definition:-

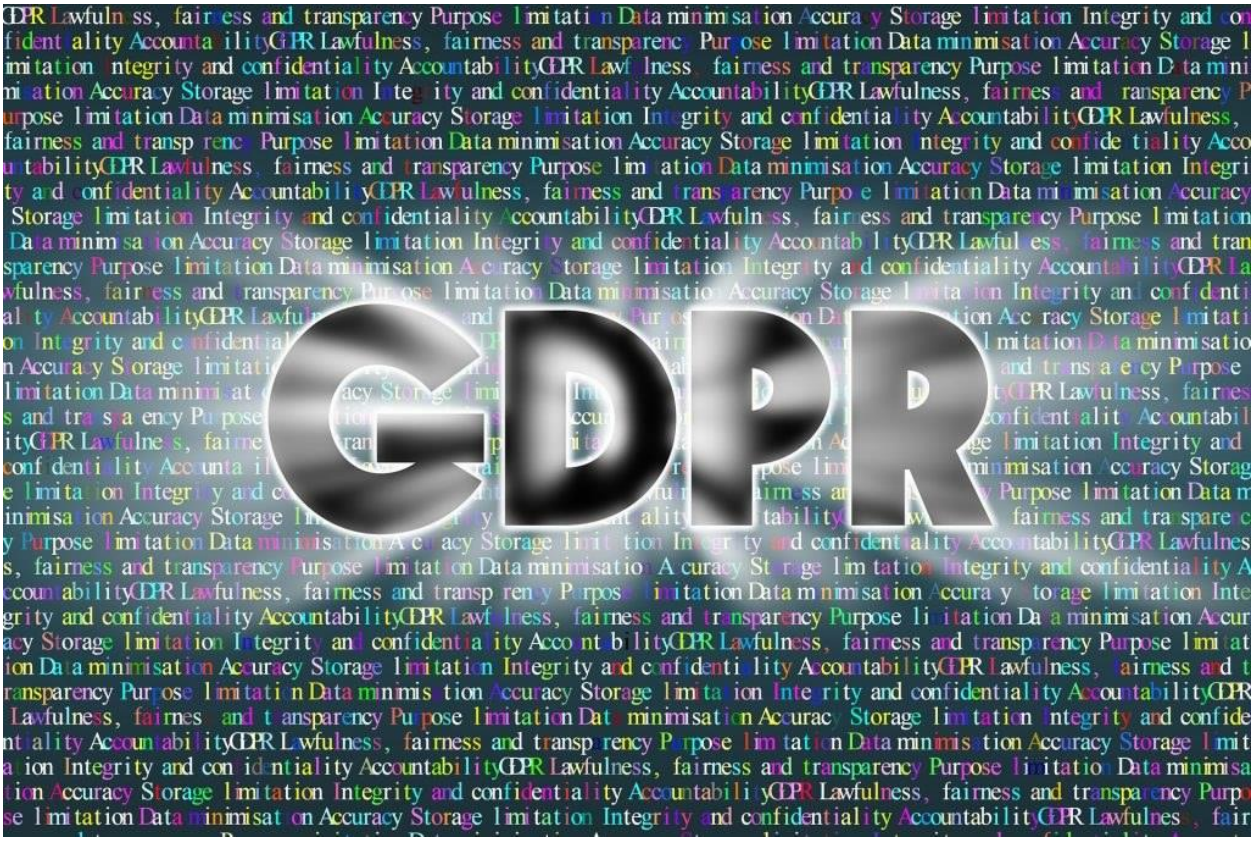
Private data is information concerning a person that can be reasonably expected to be secured from public view. Privacy is the freedom not to be observed. It is considered a basic right that is important to safety, security and quality of life. The following are illustrative examples of private data.

Example:-

- *a person's phone number and address,*
- *medical records,*
- *Financial records.*

GDPR


GDPR


GDPR

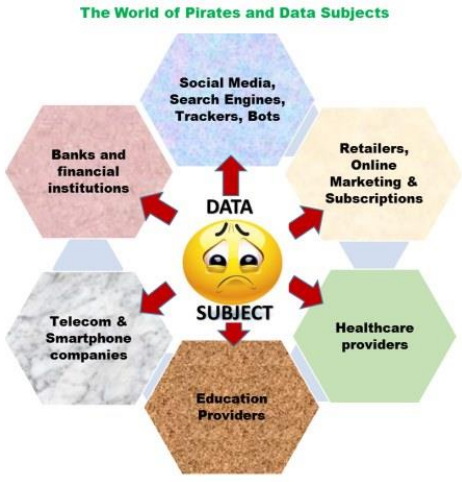
GDPR

General
Data
Protection
Regulation

The Data Protection Act 2018 (DPA) - aligns the UK's data protection laws with the EU's GDPR and updates the Data Protection Act 1998. This Act makes provision about the processing of personal data and has seven principles governing how data should be used:

1. ***Lawfulness, fairness and transparency***
2. ***Purpose limitation***
3. ***Data minimisation***
4. ***Accuracy***
5. ***Storage limitation***
6. ***Integrity and confidentiality (security)***
7. ***Accountability***

Data Subject



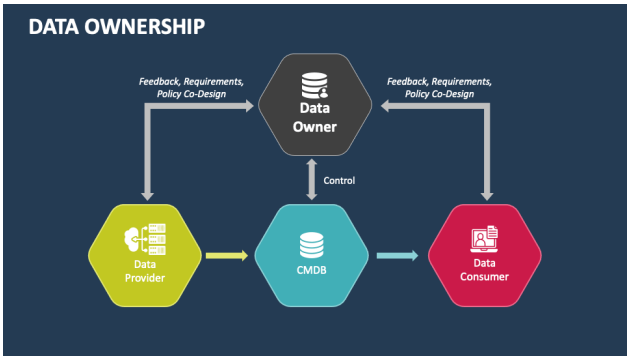
Eight rights of a data subject -

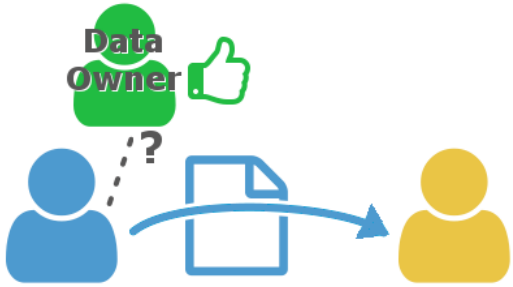
There are eight specific rights under the General Data Protection Regulation (GDPR) that data subjects have:

1. ***the right to be informed***
2. ***the right of access***
3. ***the right of rectification***
4. ***the right to erasure***
5. ***the right to restrict processing***
6. ***the right to data portability***
7. ***the right to object***
8. ***rights related to automated decision making including profiling.***

Data Owner


Ownership


DATA OWNERSHIP
Feedback, Requirements, Policy Co-Design
Data Owner
Feedback, Requirements, Policy Co-Design
Control
Data Provider
CMDB
Data Consumer

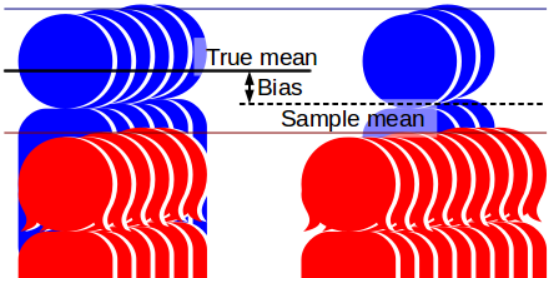

Data Owner
?

Responsibilities of Data Owner-

The data owner is responsible for ensuring that appropriate steps are taken to protect data and for the implementation of policies, guidelines and memorandums of understanding that define the appropriate use of the data.

In some cases this may refer to the data subject themselves especially in the legal context of GDPR, however with regards to ownership of data in organisations, this comes down to issues of accountability.

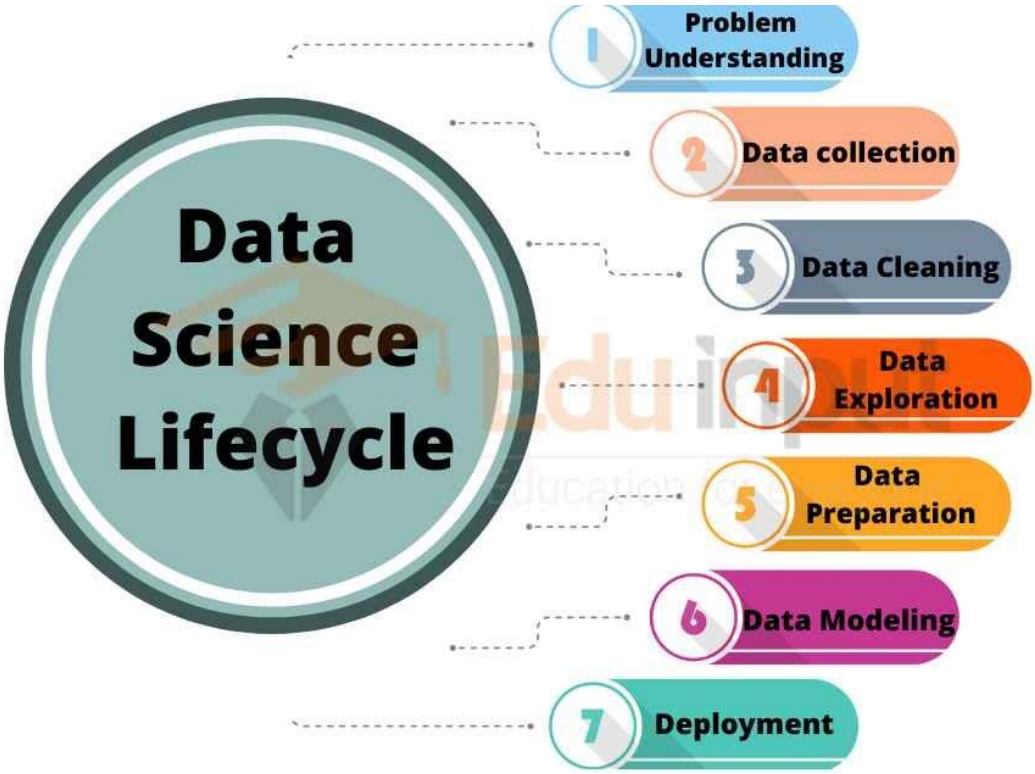
Lesson- 8 Data Bias in Data Science



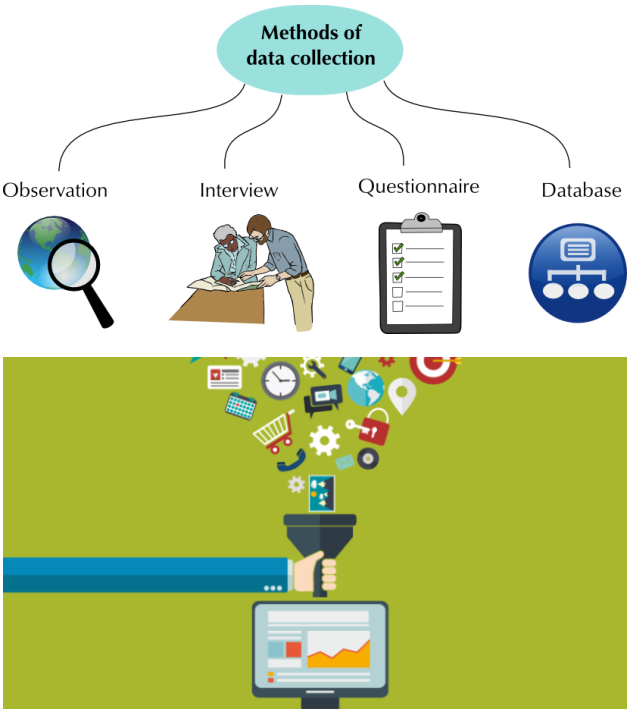
What is Bias:

Bias is a tendency to prefer one person or thing to another, and to favour that person or thing.

The Data Science Life Cycle



1- Gathering Data



The first stage in the data life cycle is collecting data from various internal and external sources. This may sound simple, but collecting large chunks of data accurately is quite the challenge as it might be biased at this stage.

Method to eliminate bias:-

To avoid bias in this stage the important practice that should be in consideration is that the gathered sample data must be representing the population. The sample should be representative of the population in terms of demographics, such as age, gender, and income.

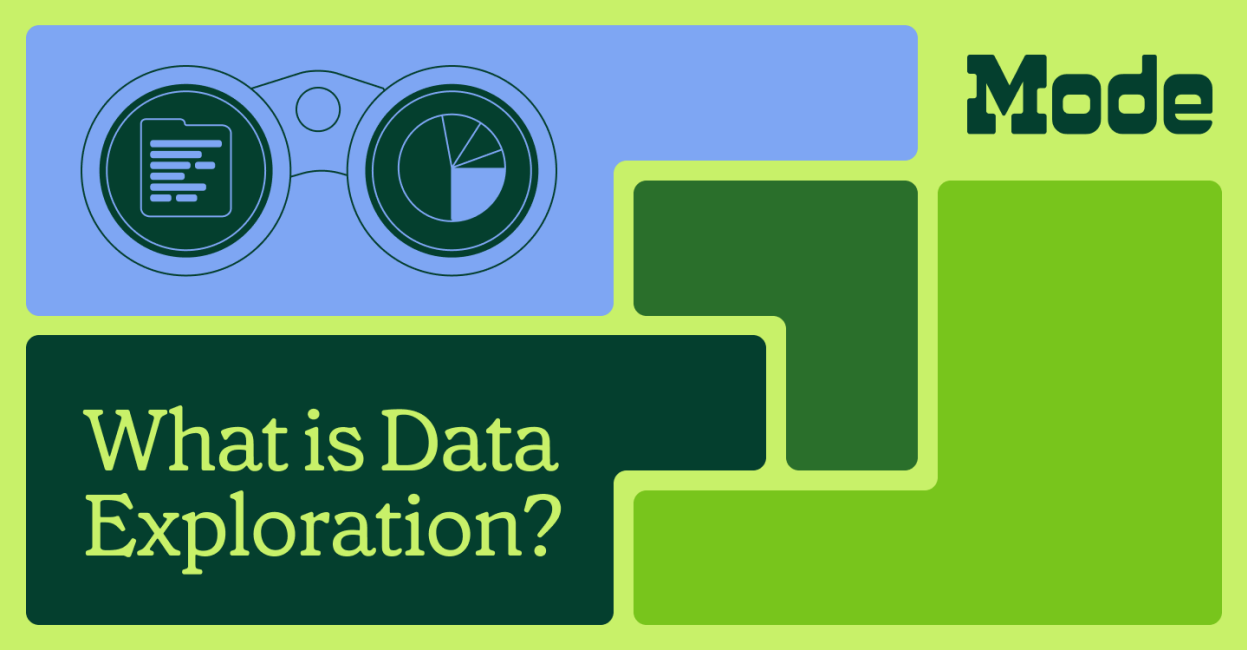
2- Cleaning Data

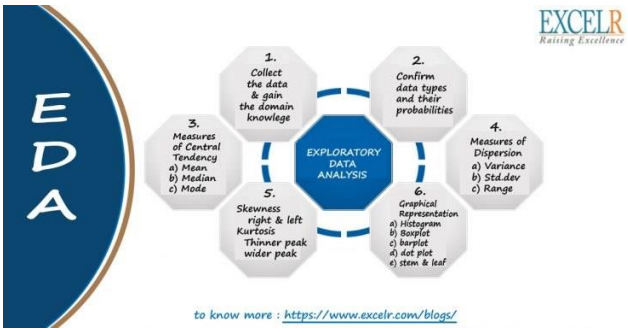
Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

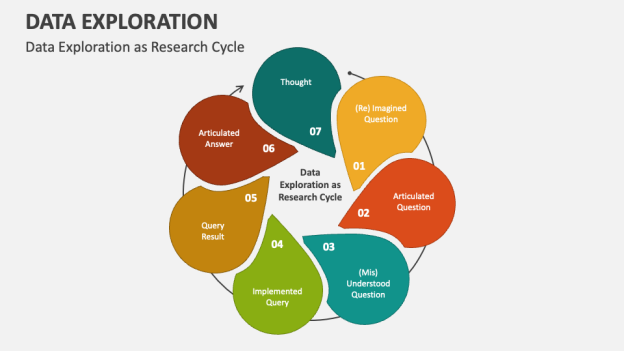
Method to eliminate bias at this stage:-

To avoid bias in data cleaning stage we have to apply various strategies, such as data augmentation, sampling, normalization, encoding, and validation.

3- Exploring Data







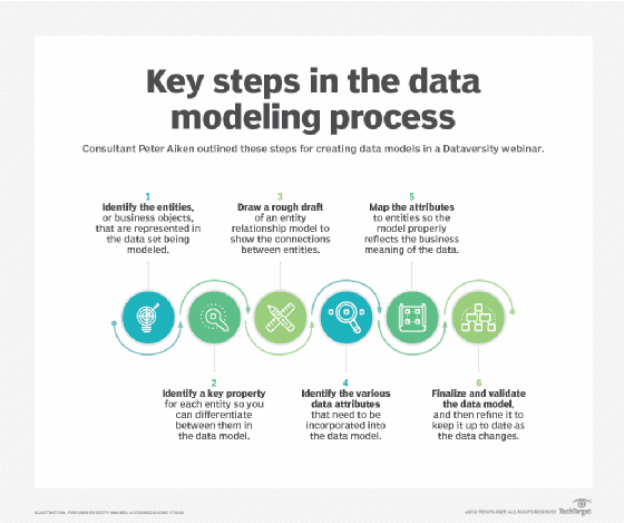
Data exploration:-

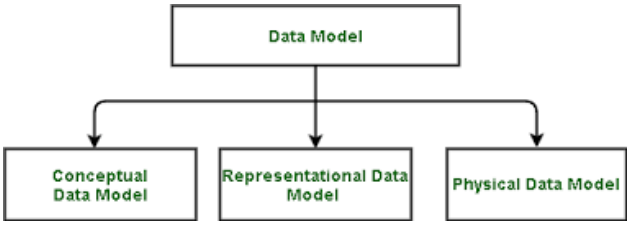
Data exploration refers to the initial step in data analysis in which data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data.

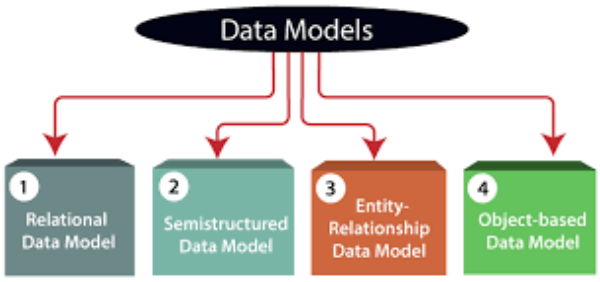
Method to eliminate bias at this stage:-

It is important to use diverse and representative data, to monitor the performance of models on various subgroups, and to use techniques such as fairness constraints to ensure that the models are making predictions that are fair and accurate.

4-Modelling Data







Data modelling:-

Data modelling is the process of creating a visual representation of either a whole information system or parts of it to communicate connections between data points and structures.

Method to eliminate bias at this stage:-

Data collection processes should be carefully designed to ensure a fair representation of all relevant data points.

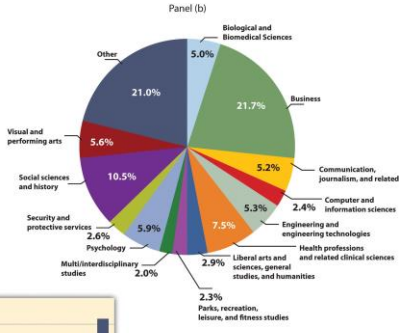
Diversity in the team is the best way to begin eliminating bias. Diversifying a team can have a major positive impact on machine learning models by producing well-rounded, representative datasets.

5- Interpreting Data

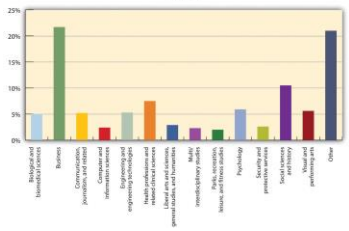
Panel (a)

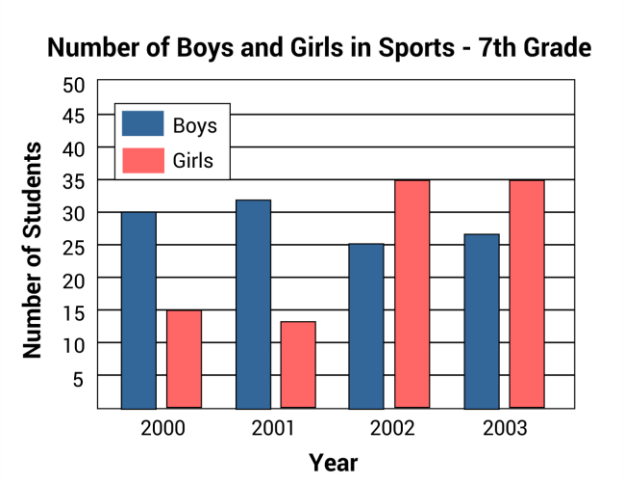
Field of Study	Percentage of Bachelor's Degrees
Biological and biomedical sciences	5.0%
Business	21.7%
Communication, journalism, and related	5.2%
Engineering and engineering technologies	5.3%
Health professions and related clinical sciences	7.5%
Liberal arts and sciences, general studies, and humanities	2.9%
Psychology	5.9%
Social sciences and history	10.5%
Visual and performing arts	5.6%
Security and protective services	2.6%
Computer and information sciences	2.4%
Multi/interdisciplinary studies	2.3%
Parks, recreation, leisure, and fitness studies	2.0%
Other	21.0%

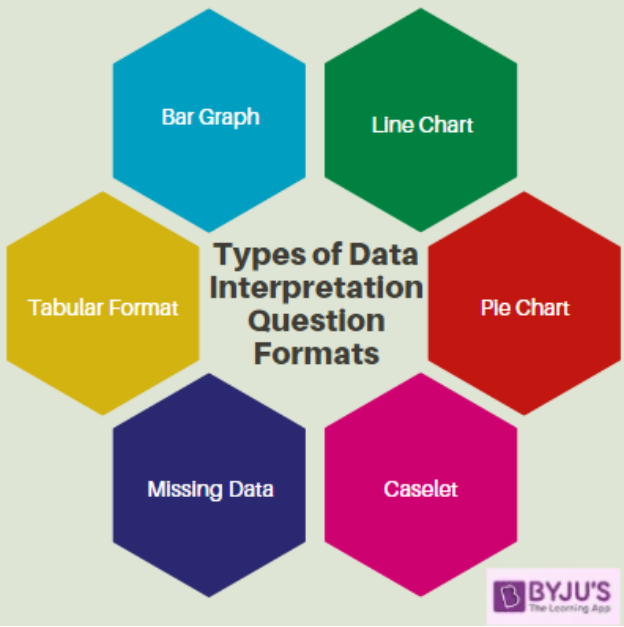
Panel (b)

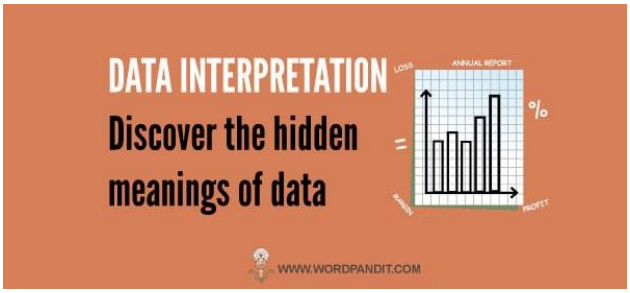


Panel (c)









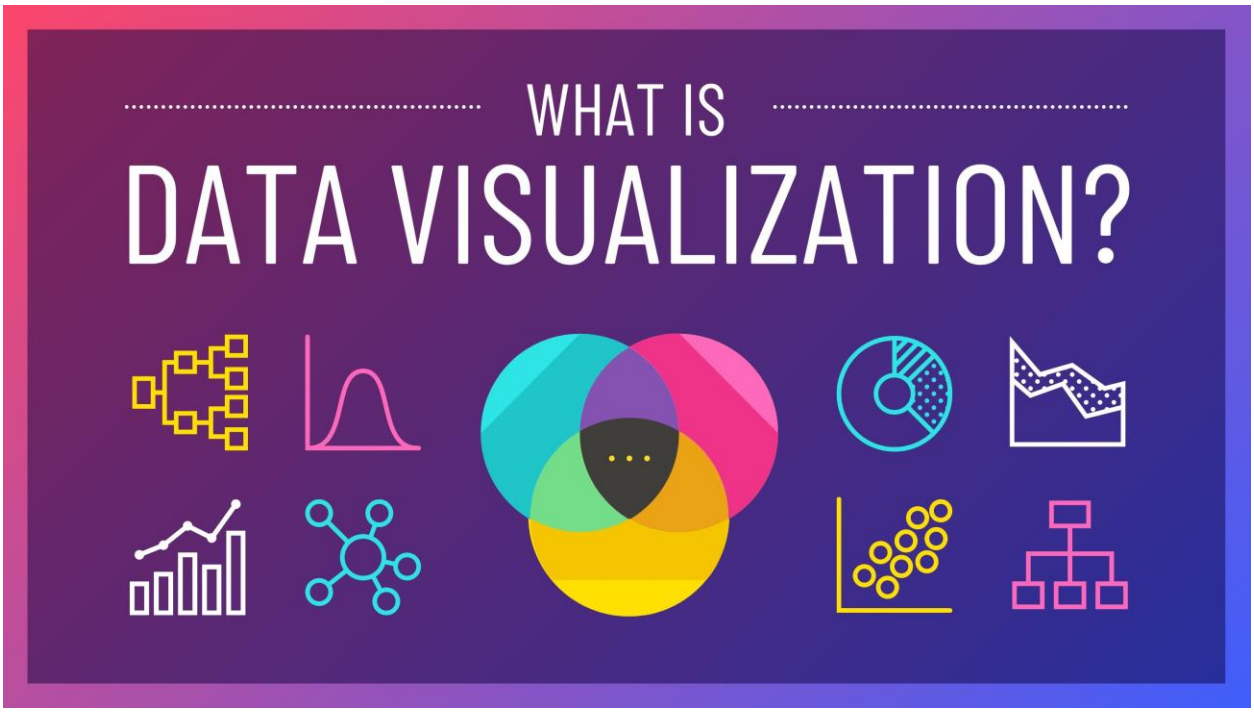
Interpreting Data:-

Data interpretation refers to the process of using diverse analytical methods to review data and arrive at relevant conclusions. The interpretation of data helps researchers to categorize, manipulate, and summarize the information in order to answer critical questions.

Method to eliminate bias at this stage:-

Incorporating fairness constraints: Fairness constraints can be incorporated into the model to ensure that it is not making biased predictions based on certain variables, such as race or gender.

Week- 9 Data Visualizations



Data Visualizations :-

" Data visualization helps to bridge the gap between numbers and words."

- Brie E. Anderson

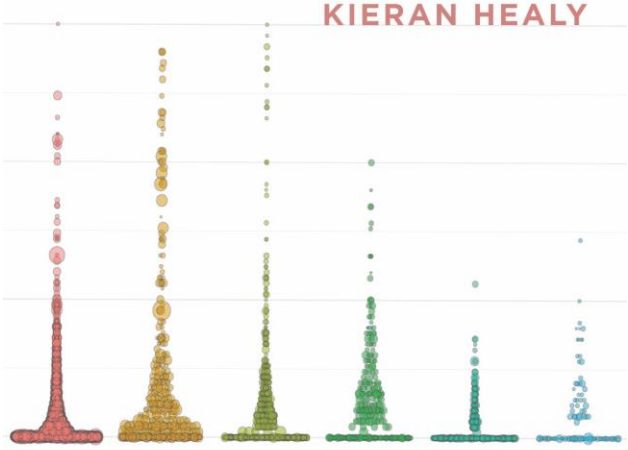

CHART 1
CHART 2
CHART 3
CHART 4
CHART 5
CHART 6
designed by freepik.com


可口可乐 (Coca-Cola)
Captive Audience
Mentions
Competition
Influencers
Person
Rank
Han Han
25153
Xiaoming Xu
21123
Chengpeng Li
18553
Xianping Lang
14213
Jinglei Xu
9153
Sentiment
Positive
Neutral
Demographics
20%
80%
Age Groups
18 and under 51%
19 - 30 23%
31 - 45 16 %
Geography
Issues
D/BM Online Monitor

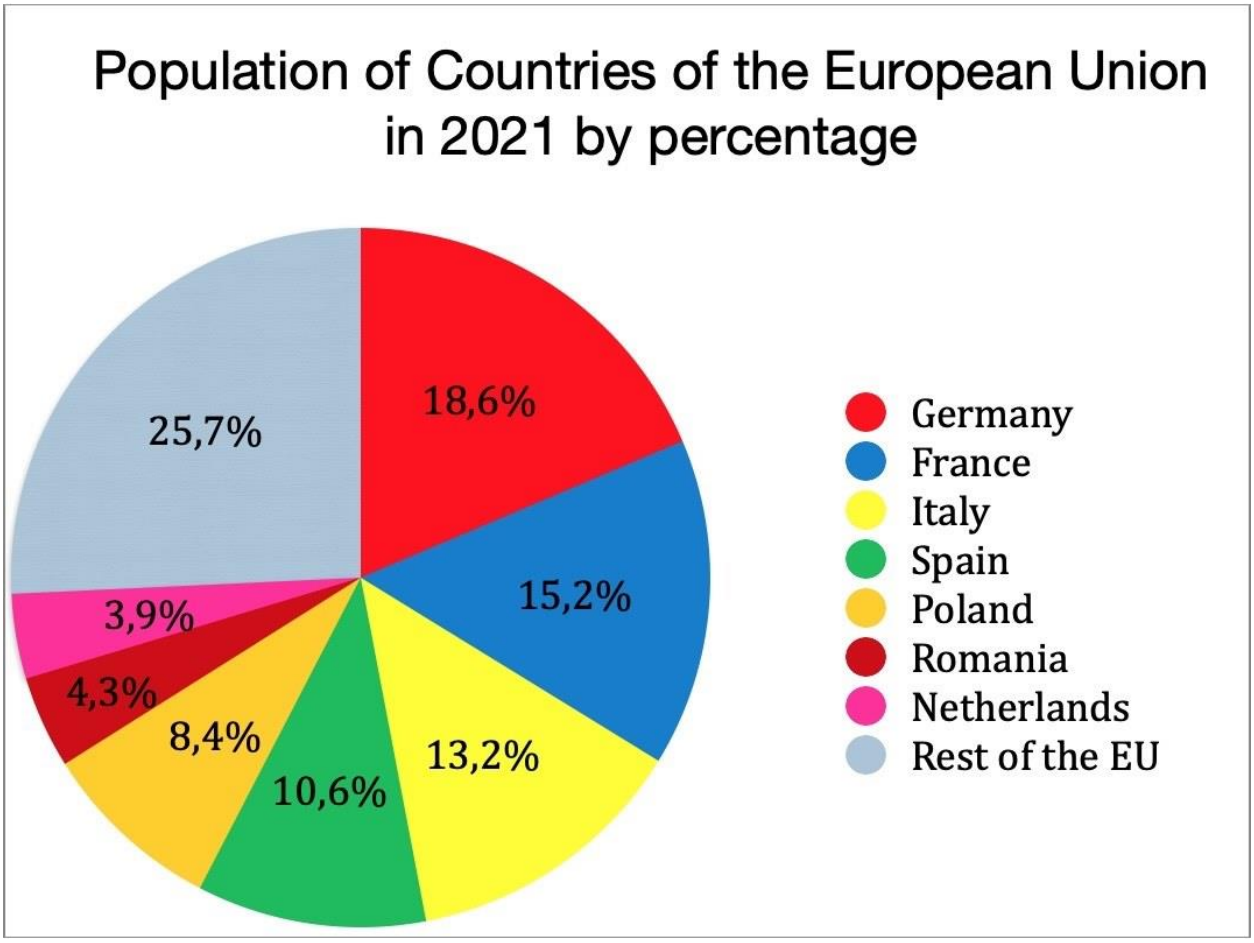
DATA VISUALIZATION

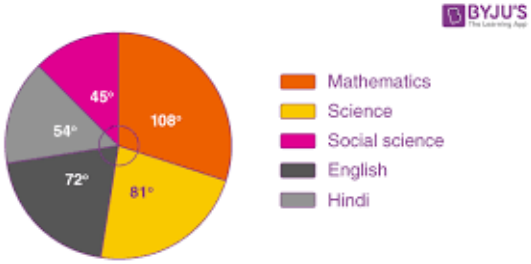
A PRACTICAL INTRODUCTION

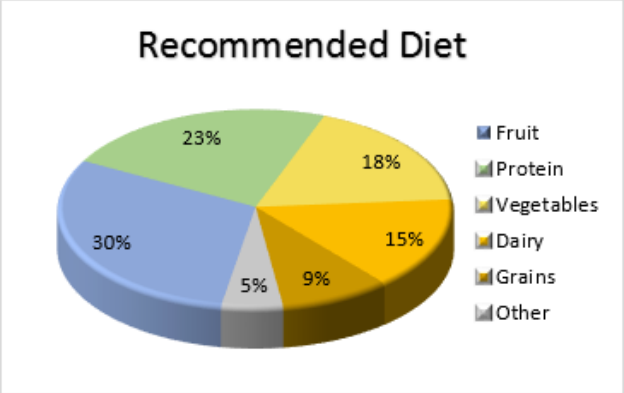
KIERAN HEALY



Pie charts


Population of Countries of the European Union in 2021 by percentage
18,6%
15,2%
13,2%
10,6%
8,4%
4,3%
3,9%
25,7%
Germany
France
Italy
Spain
Poland
Romania
Netherlands
Rest of the EU


BYJU'S
45°
108°
54°
72°
81°
Mathematics
Science
Social science
English
Hindi



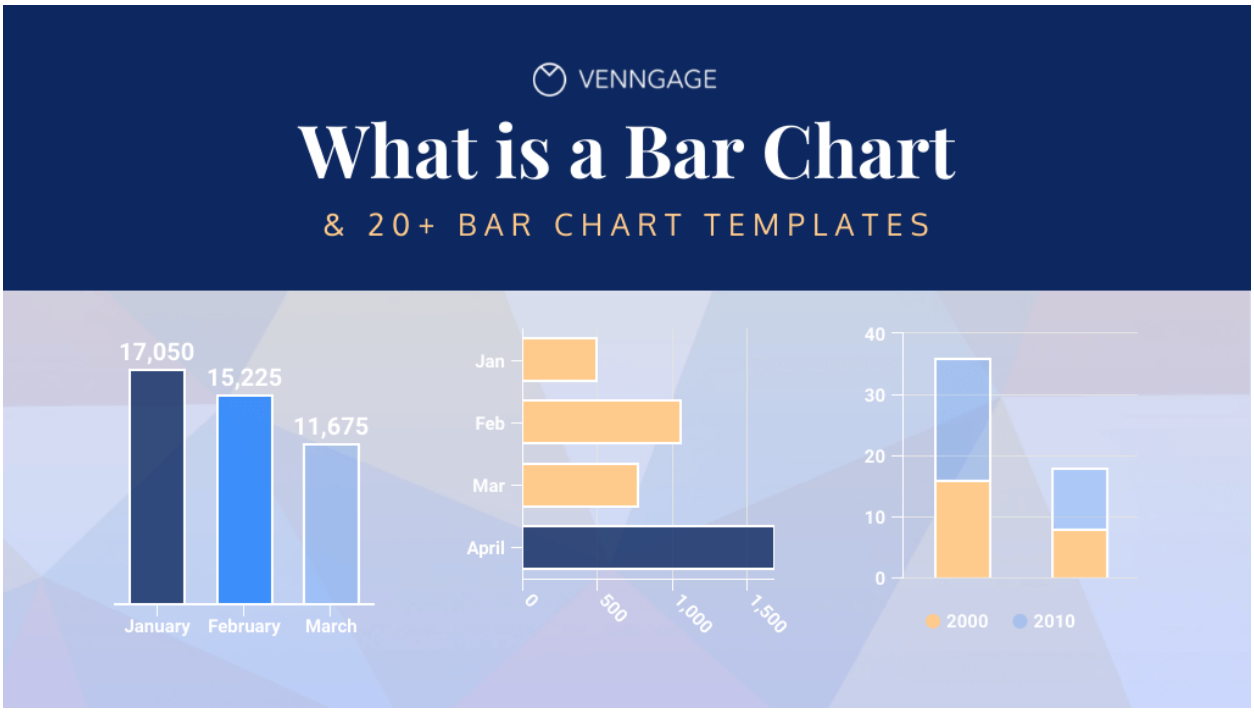
Pie Chart:-

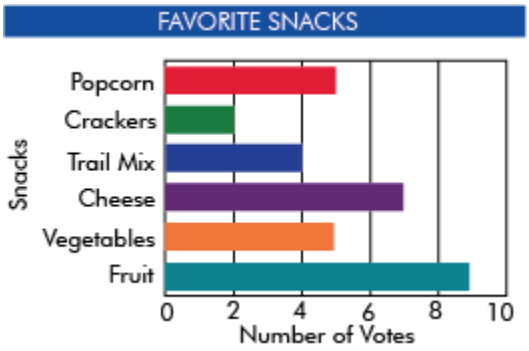
A pie chart is a type of graph that represents the data in the circular graph. The slices of pie show the relative size of the data, and it is a type of pictorial representation of data. A pie chart requires a list of categorical variables and numerical variables.

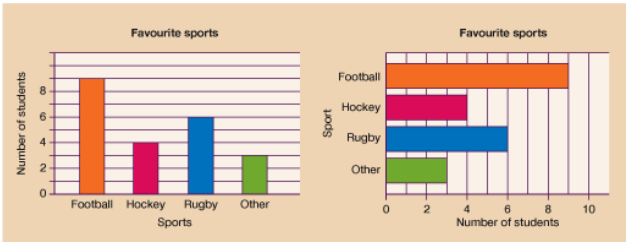
Uses of a Pie chart:-

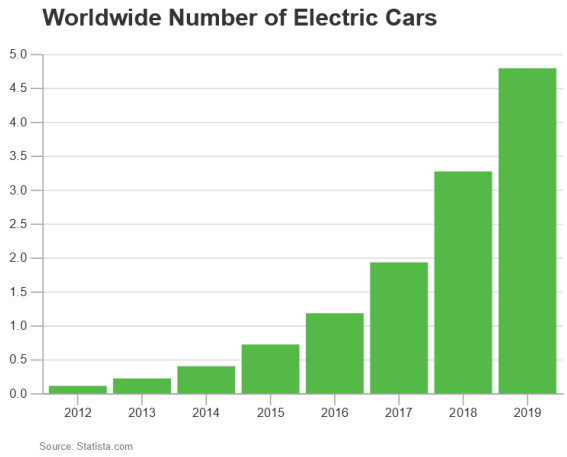
Pie charts can be used to show percentages of a whole, and represents percentages at a set point in time.

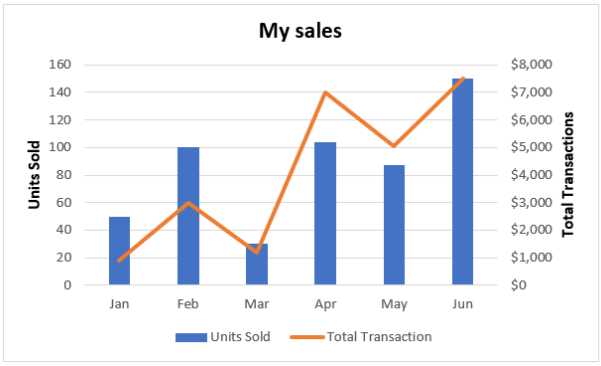
Bar charts


VENNGAGE
What is a Bar Chart
& 20+ BAR CHART TEMPLATES
17,050
15,225
11,675
January
February
March
Jan
Feb
Mar
April
0
500
1,000
1,500
40
30
20
10
0
2000
2010


FAVORITE SNACKS
Popcorn
Crackers
Trail Mix
Cheese
Vegetables
Fruit
Snacks
0
2
4
6
8
10
Number of Votes


Favourite sports
Number of students
0
2
4
6
8
Football
Hockey
Rugby
Other
Sports
Favourite sports
Sport
Football
Hockey
Rugby
Other
0
2
4
6
8
10
Number of students





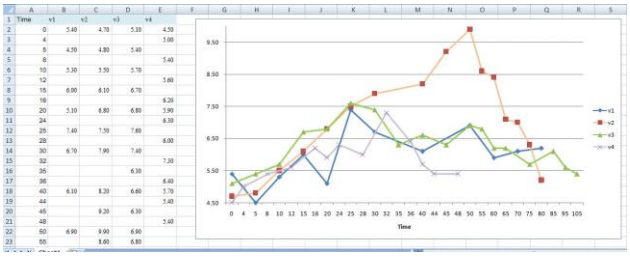
Bar Graph:-

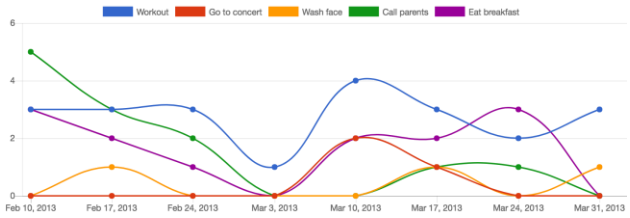
Bar graphs are the pictorial representation of data (generally grouped), in the form of vertical or horizontal rectangular bars, where the length of bars are proportional to the measure of data. They are also known as bar charts.

Uses of a Bar graph:-

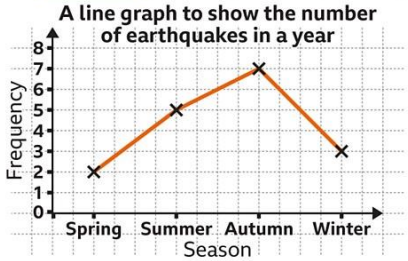
Bar charts enable us to compare numerical values like integers and percentages. They use the length of each bar to represent the value of each variable.

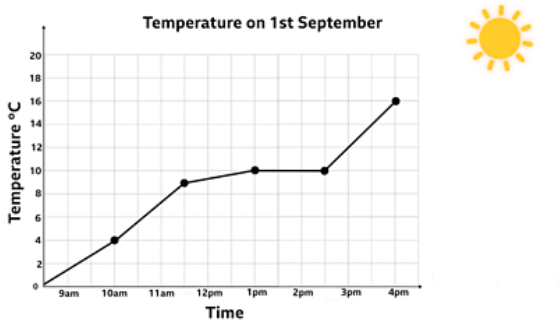
Line charts

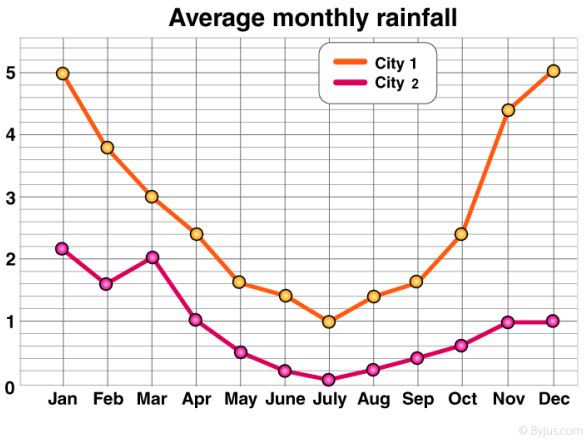




Season	Spring	Summer	Autumn	Winter
Frequency	2	6	7	3







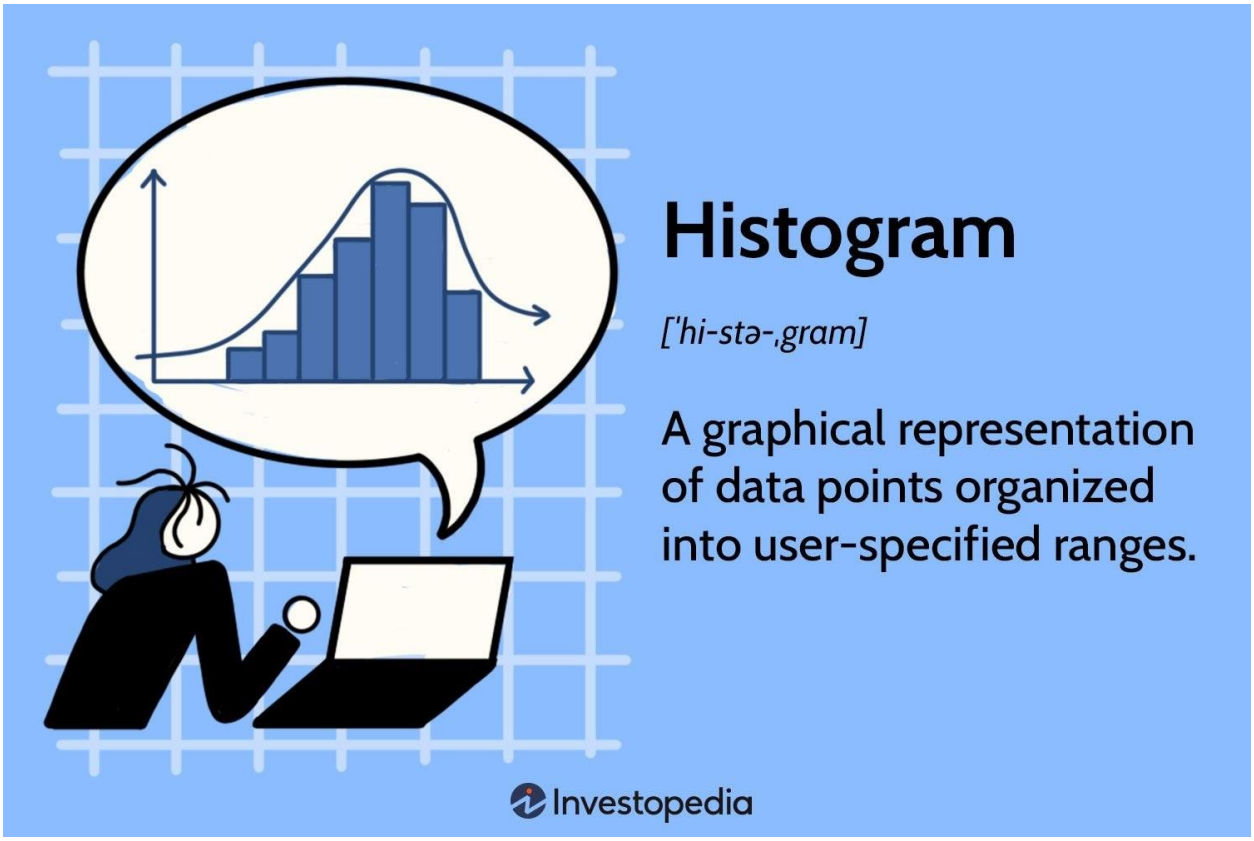
Line Chart:-

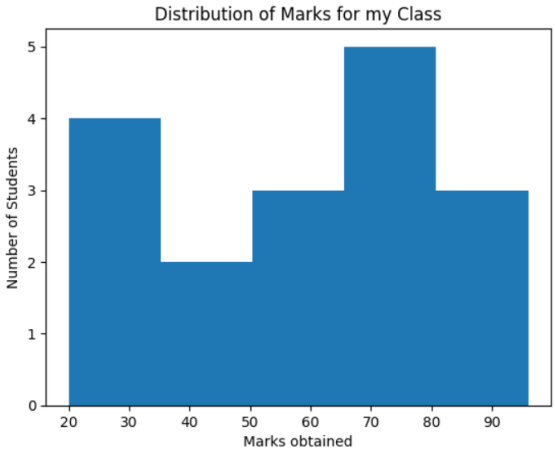
A line chart, also known as a line graph or curve chart, is a graphical representation used to display data points connected by straight lines. This type of chart is particularly useful for visualizing trends, changes, and relationships in data over a continuous interval, often time.

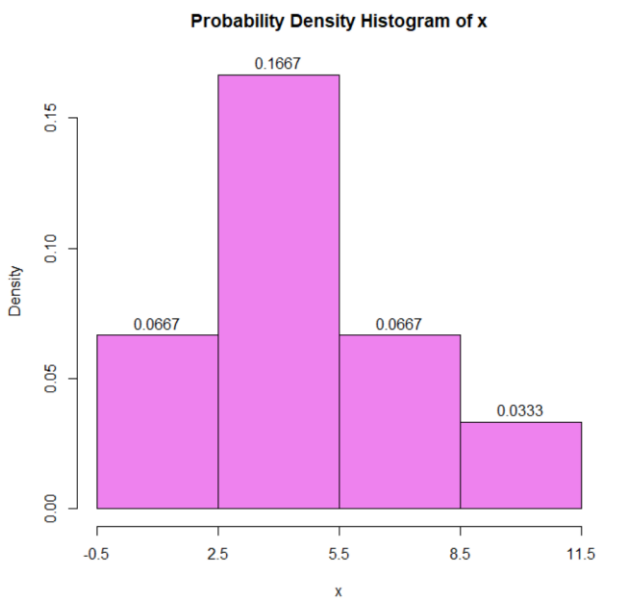
Uses of a Line chart:-

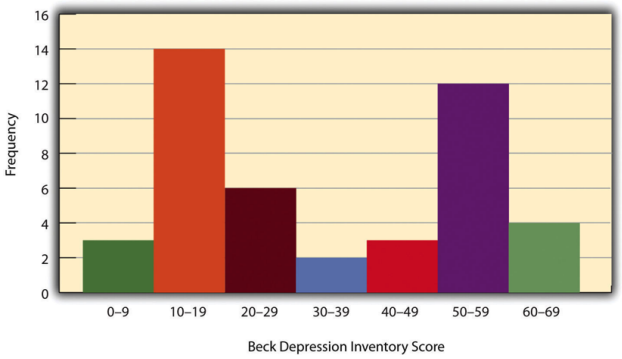
A line graph also known as a line plot or a line chart is a graph that uses lines to connect individual data points. A line graph displays quantitative values over a specified time interval. In finance, line graphs are commonly used to depict the historical price action of an asset or security.

Histogram


Histogram
['hi-stə-ˌgram]
A graphical representation of data points organized into user-specified ranges.
Investopedia


Distribution of Marks for my Class
Number of Students
Marks obtained
0
1
2
3
4
5
20
30
40
50
60
70
80
90





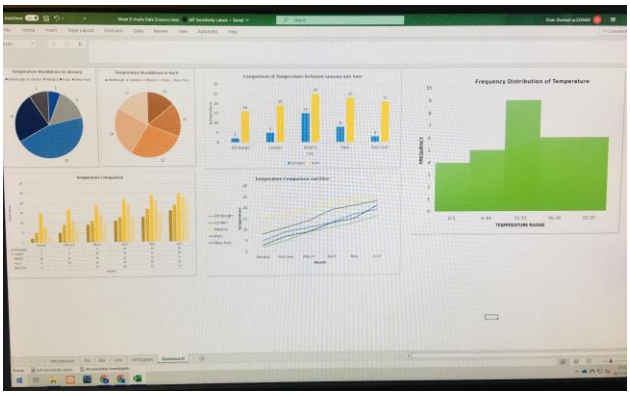
Histogram:-

A histogram is a graph that shows the frequency of numerical data using rectangles. The height of a rectangle (the vertical axis) represents the distribution frequency of a variable (the amount, or how often that variable appears).

Uses:-

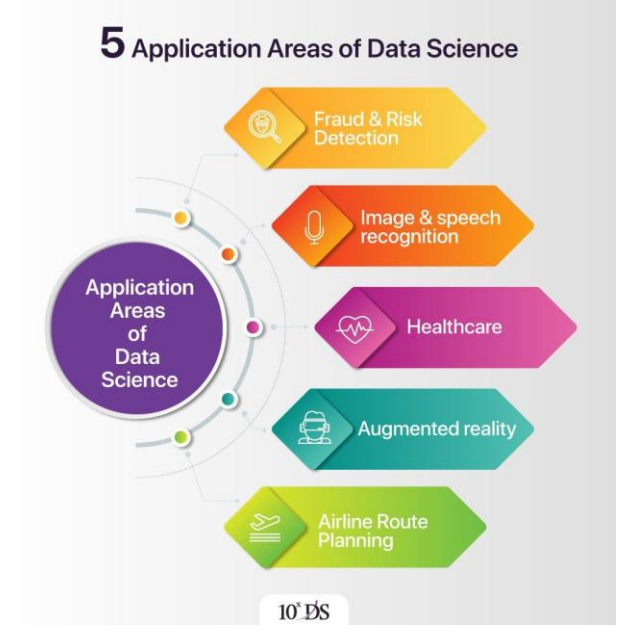
The histogram is a popular graphing tool. It is used to summarize discrete or continuous data that are measured on an interval scale. It is often used to illustrate the major features of the distribution of the data in a convenient form.

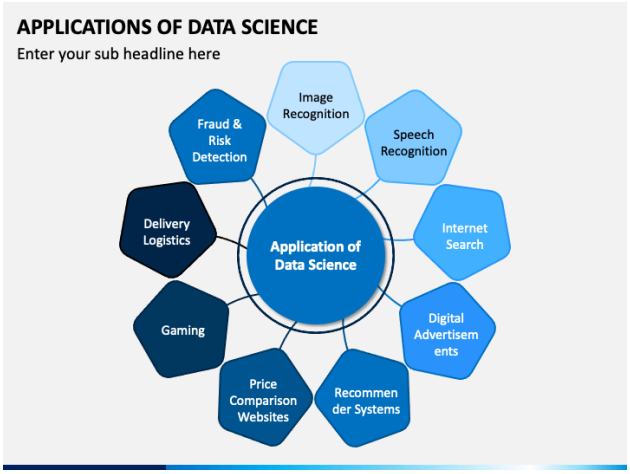
Temperature Dashboard



Lesson-10 Contemporary Application of Data Science

APPLICATIONS OF DATA SCIENCE





Free Beer for Cycling

FREE BEER for CYCLING-

For cyclists living in or visiting Bologna(Italy), that's a perk they can regularly enjoy. The Italian city wants to encourage sustainable transportation—namely biking, walking, and public transit—by rewarding bike commuters with things like free beer, ice cream, and movie tickets.

This is an anti-pollution scheme which is rewarding people who get out of their cars and use cycle, walk or take public transport instead. This scheme is called "Bella Mossa" that means Good Job. This idea is the brainchild of urban planner Macro Amadori.

This scheme aims to reduce car journeys and boost green travel with incentives like beer, ice cream, or cinema tickets for participants who complete eco friendly walks and rides.

WHERE IS THE APP and WHAT DOES IT DO-

For this people need to download the app called " Better Points". Then need to log their green journeys then need to swap their points.

This app uses GPS tracking to record their journeys. Each walk or trip by bike or bus earns points which can be built up and exchanged for goods and services at local businesses that have signed up to the scheme. Over 100 local businesses have signed up to give away discount vouchers.

The app allows a maximum of four trips to be logged each day, to encourage people to use the scheme over time and help make travelling around the city more sustainable.

The system of point collection is based not on the distance people travel but on a single trip, because it's important that even for short trips of 1 km people do it in a sustainable way.

ADVANTAGE-

Bella Mossa has proved popular and there is a talk of extending the initiative's four-month operating period. Vehicle emissions are damaging to the environment and our health and air pollution is the fourth biggest cause of premature deaths globally, with the developing world worst affected. So this scheme is quite wonderful which encourage people to have sustainable rides.

My problem

Description of my problem-

Updating new address everywhere individually where I have saved my current address including shopping sites, utility providers, bank accounts, GP and other healthcare, Insurance companies, school/education institution, Delivery of certificates or parcels etc. is quite stressful. Although I know there are some sites as HMRC where this can be done but not completely as I am aware.

Available data-

These days each and every site asks for our details including our name and address, contact information etc. It means our data has been stored on different sites/portals already. So there is no problem with finding the data of a person.

My idea-

I think there should be a centralized secure system where if we update our new address or any change in our personal details, it will automatically update to every place where we have saved our details earlier. It may send two/three step notification to confirm the changes. As per my understanding it can be done using Machine Learning algorithm or Artificial Intelligence.

Benefit-

This will reduce stress and time in changing details everywhere individually without missing out important sites/portals.

Next Step-

According to me with the help of data science this idea could be used in various areas like in business, finance, healthcare etc. with broader perspective.

Another Problem with Idea-

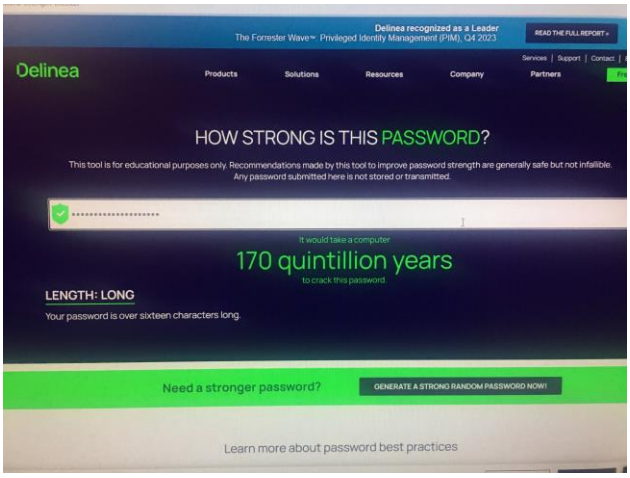
I have another problem and idea about finding the data of maximum number and required urgent attention for potholes in a road.

I think if vehicles have sensors underneath that could sense, collect and send data to a central system how many times the vehicle hit the potholes or jumped onto in a particular area/road. Sensors could have GPS as well to send exact location. If multiple sensors of multiple vehicles send the data for any particular area, then it can be analysed where are the maximum number of potholes present and how dangerous they are to be treated.

I am not sure if my problems and ideas are useful and practical but I am thinking about it as of now.

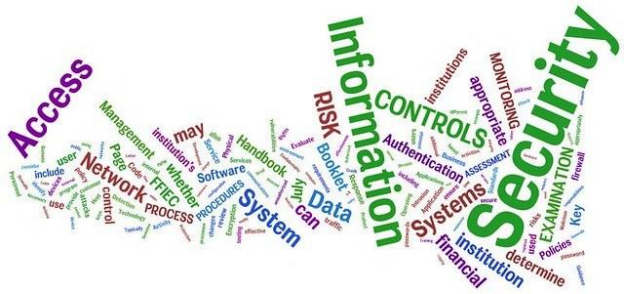
Lesson-11 Data Security

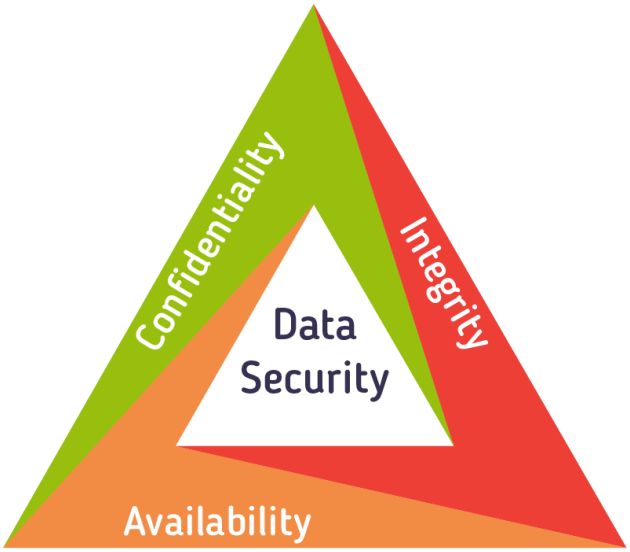
Password Checker

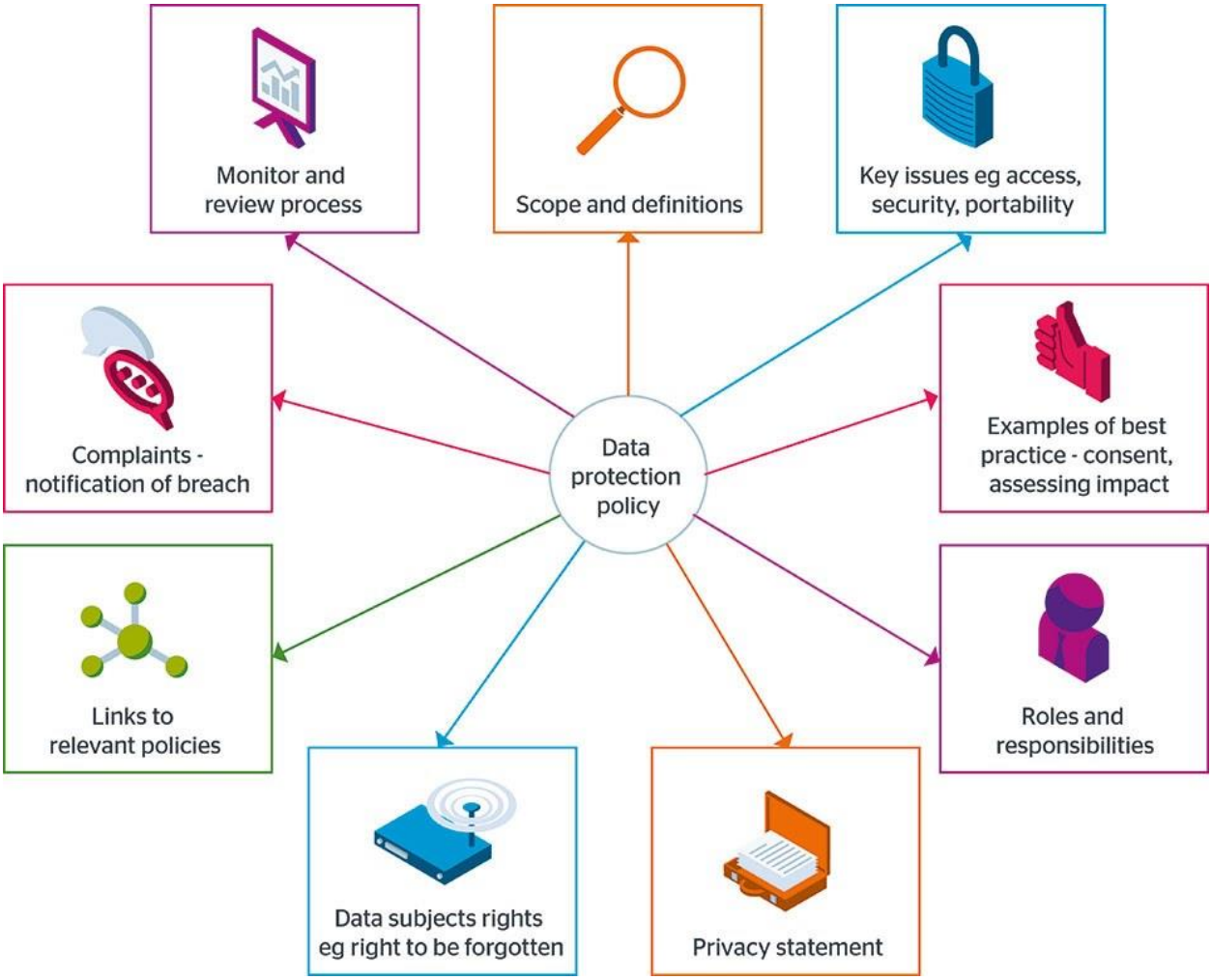


I have entered the password that I created with 'edinburghcollege', and changed 3 characters. The password checker says that it will take 170 quintillion years to break it.

Data Security Scenario




Confidentiality
Integrity
Data Security
Availability


Monitor and review process
Scope and definitions
Key issues eg access, security, portability
Complaints - notification of breach
Data protection policy
Examples of best practice - consent, assessing impact
Links to relevant policies
Roles and responsibilities
Data subjects rights eg right to be forgotten
Privacy statement

1- What is the potential risk to the business and its clients?

Ans- Risks associated with data breaches include exposing sensitive customer data, company secrets, eroding trust, damaging brand reputation, and potential financial losses. They can lead to financial loss, reputational damage, and loss of customer trust.

2- What should the business do to mitigate the risk?

The following are the most widely used mitigation strategies for business risks.

---

1. *Conduct a risk assessment*
2. *Develop a security policy*
3. *Use strong passwords and two-factor authentication*
4. *Regularly update software and devices*

5. *Use firewalls and antivirus software*

6. *Encrypt sensitive data*

7. *Limit employee access to data*

8. *Train employees on data security*

9. *Back up data regularly*

10. *Monitor and audit security*

3- What steps can the business take to prevent a similar incident from occurring in the future?

Ways To Mitigate Security Risks and Threats

- *Conduct a Cyber security Risk Assessment.*
- *Create an Incident Response (IR) Plan.*
- *Train Your Team.*
- *Monitor and Protect Your Network Traffic.*
- *Enforce the Use of Strong Passwords.*
- *Install Security Patches and Updates.*

4-How can the business ensure that all employees understand the importance of data security?

Ans- One of the best ways to involve your employees in corporate data security is to create a policy. Your policy should outline what is expected of employees in terms of keeping your organisation's data secure and the action they need to take.

5- What legal and regulatory requirements should the business be aware of regarding data security?

Businesses have a legal obligation to protect the information they have collected to ensure the confidentiality and integrity of the data. Companies need to follow security standards and practices to make sure sensitive information is protected from things like accidental loss, theft, and fraudulent activity.
