Introduction to Data Science

DATAPLAY

- Concept of Data Science
- Need of Data Science
- Applications Of Data Science
- Solving Simple Data Science Use Cases
- Components of Data Science
- Data Science Life Cycle
- Market of Data Science
- Understanding AI /ML/DL/NLP/CV/Data Science
- Rise of Data Analytics
- Types of Data Analytics
- Data Analytics Lifecycle
- Need for Business Analytics

Concept of Data Science

- ❖ What is Data Science?
- ❖ Is Data Science a recent thing?

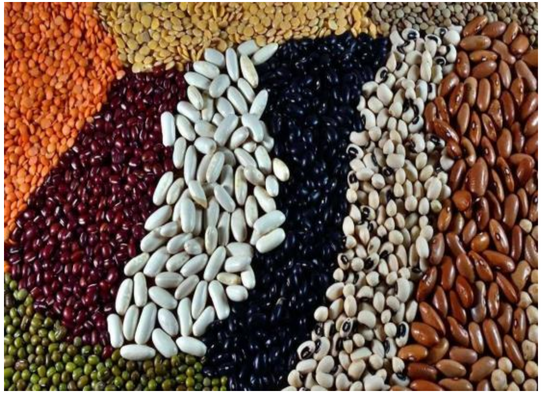
DATAPLAY

Decision Factors

1. Your past purchase experience
2. Proximity to store
3. Type of Purchase you want to make
4. Pricing and Discounts
5. Return and Exchange Policies
6. Shopping Experience

Predictive Analytics

A field of data analysis that uses historical data and statistical algorithms to make predictions about future events or outcomes.



Decision Factors

1. Shape
2. Color
3. Weight
4. Size
5. Texture

Clustering

Clustering is a machine learning and data analysis technique used to group similar data points together based on certain characteristics or features they share.

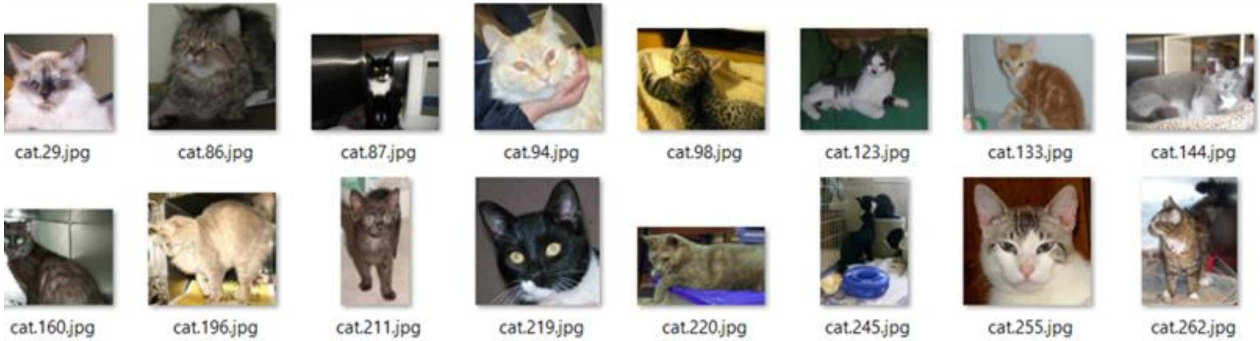
Decision Factors

1. Genre Preference
2. Ratings and Reviews
3. Movie Movie Similarity
4. Friend Friend Similarity
5. Mood

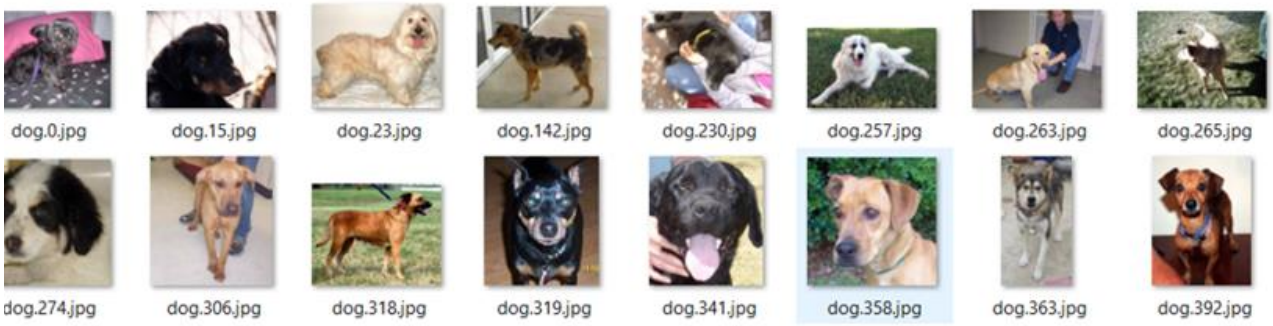
Recommendation System

It is a type of software or algorithm that provides personalized suggestions or recommendations to users. These recommendations are typically based on the users' preferences, past behavior, and other relevant data.

Dogs v/s Cats



Cats



Dogs

What is it? Cat or Dog?

Decision Factors

Animal	Fur Type	Tail Length	Ear Shape	Size
Dog	Short	Medium	Pointed	Medium
Cat	Long	Short	Pointed	Small

Is Data Science a new advancement?

No!

We have been doing this all our lives.

Need of data Science

- ❖ Then what is it about Data Science that is new?
- ❖ Why it became such a rage in the recent years?

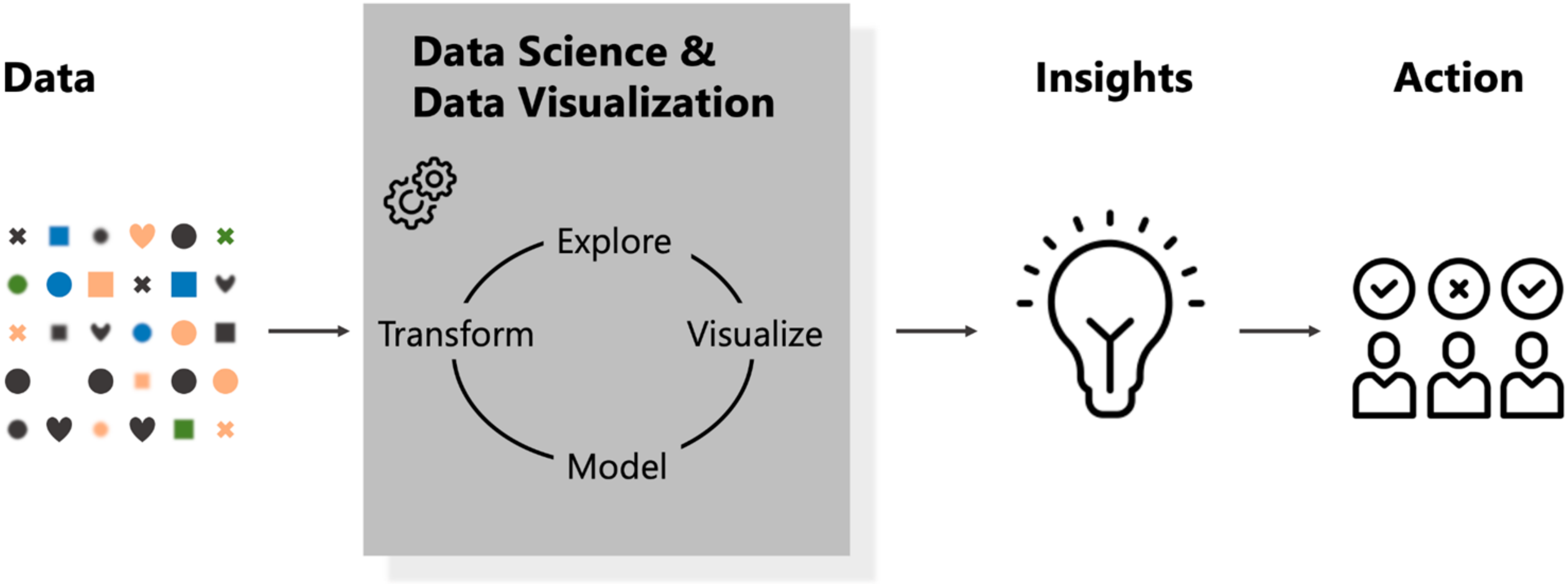
Big Data

Every two days now we create as much information as we did from the dawn of civilization up until 2003, according to Schmidt. That's something like five exabytes(10^{18}) of data, he says.
Good news today is that technology today helps us to utilise this data.

-Eric Schmidt (Former Google CEO), 2010

- **Technological Advancements:** The exponential growth in computing power, storage capacity, and data processing technologies has made it feasible to store, manage, and analyze massive volumes of data.
- **Proliferation of Digital Devices:** The widespread use of smartphones, IoT devices, and sensors has led to an explosion in data generation.
- **Internet and Social Media:** The internet and social media platforms have become integral parts of people's lives. These platforms generate enormous amounts of data through user interactions, content sharing, and online transactions.
- **E-commerce and Online Services:** The growth of e-commerce, online services, and digital platforms has resulted in vast amounts of transactional and user behaviour data.

Automation



Data

Data Science & Data Visualization

Explore

Visualize

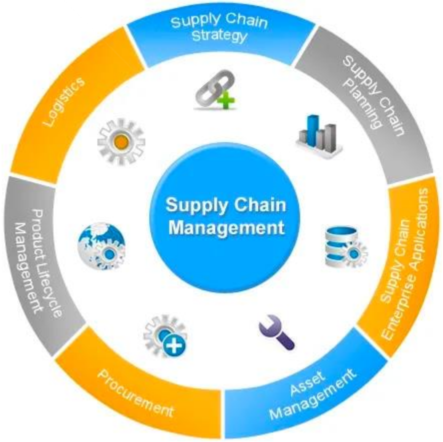
Model

Transform

Insights

Action

Business and Competitive Edge/ Scientific Research



Great Potential Unlocked

DATAPLAY

Areas of Application

1. **Prediction**
    1. **Classification**: These are algorithms that assign an input data point to one of several predefined categories. For example, email spam classification or image classification.
    2. **Regression:** These are algorithms that predict a continuous numerical value for a given input. For example, predicting the price of a house based on its features.
2. **Clustering:** These are algorithms that group similar data points into clusters. For example, segmenting customers into different groups based on their purchasing behavior.
3. **Anomaly Detection:** These are algorithms that identify data points that are significantly different from the norm. For example, detecting fraudulent transactions in a large dataset.
4. **Recommender Systems:** These are algorithms that suggest items to users based on their preferences or past behavior. For example, suggesting movies to watch or products to purchase.
5. **Natural Language Processing (NLP):** These are algorithms that process and analyze human language. For example, sentiment analysis of customer reviews or machine translation of written text.
6. **Computer Vision:** These are algorithms that process and analyze visual information. For example, object recognition in images or facial recognition in videos.

………..and many more.

DATAPLAY

Let's decode this by solving simple use cases !

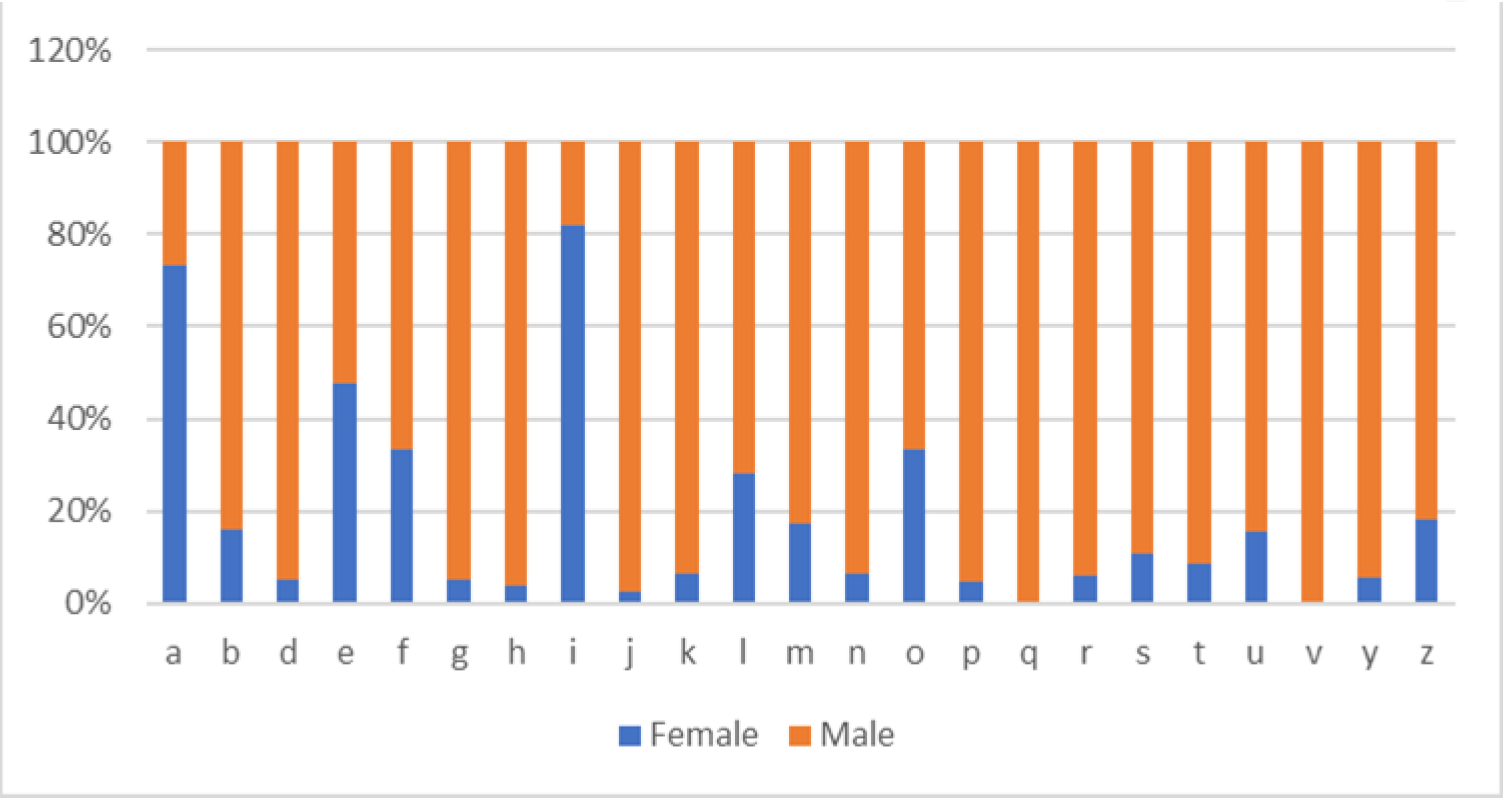
Identify gender of a person given Name

Think of a female name ending with n !

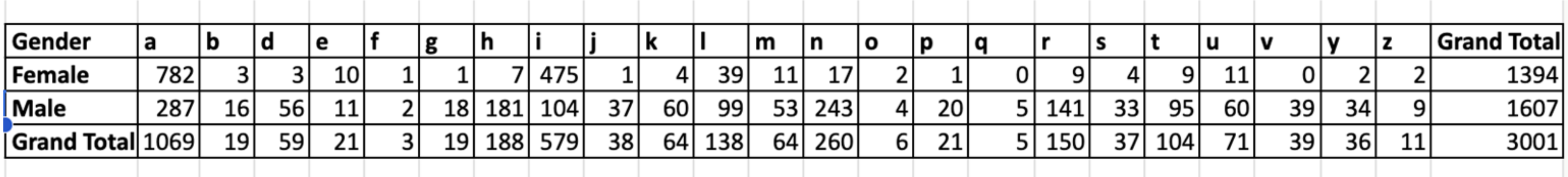
Think of a female name ending with i !

Think of a male name ending with a !

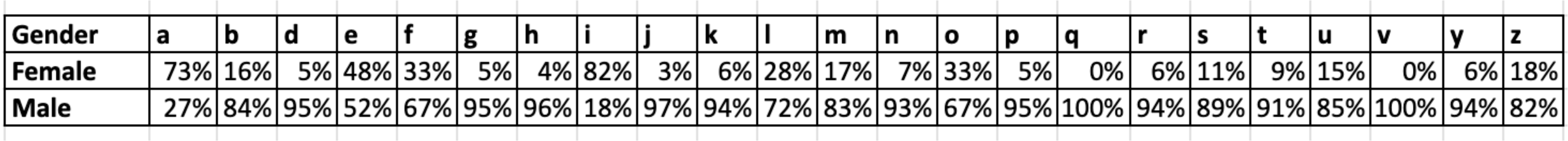
Note : Solve this problem using IF, String functions, Pivot Table, VLookup and other functions in Excel.



ATAPLAY

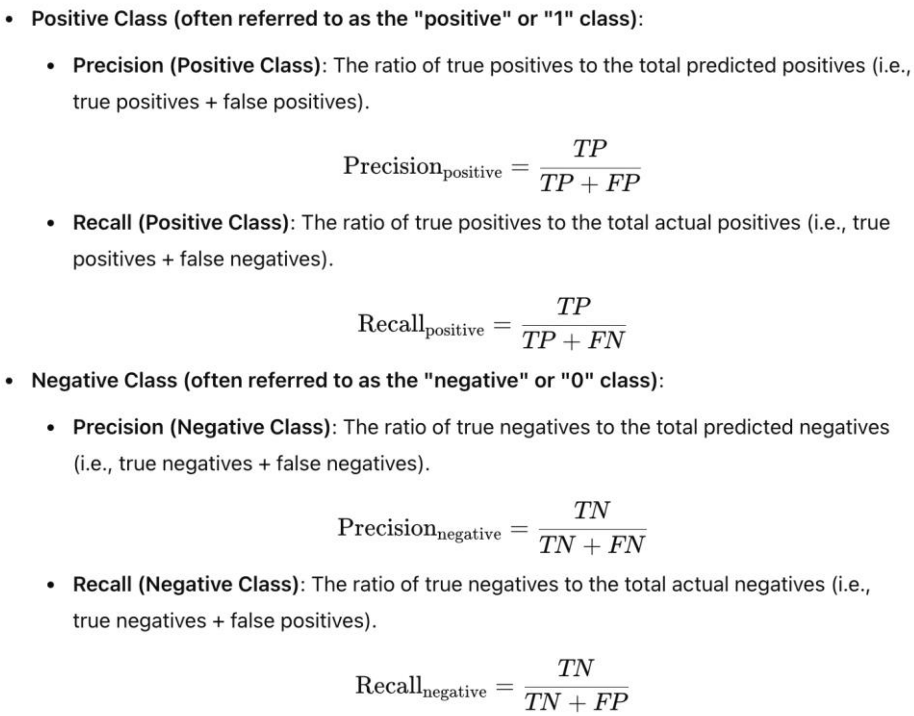


Gender	a	b	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	y	z	Grand Total
Female	782	3	3	10	1	1	7	475	1	4	39	11	17	2	1	0	9	4	9	11	0	2	2	1394
Male	287	16	56	11	2	18	181	104	37	60	99	53	243	4	20	5	141	33	95	60	39	34	9	1607
Grand Total	1069	19	59	21	3	19	188	579	38	64	138	64	260	6	21	5	150	37	104	71	39	36	11	3001



Gender	a	b	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	y	z
Female	73%	16%	5%	48%	33%	5%	4%	82%	3%	6%	28%	17%	7%	33%	5%	0%	6%	11%	9%	15%	0%	6%	18%
Male	27%	84%	95%	52%	67%	95%	96%	18%	97%	94%	72%	83%	93%	67%	95%	100%	94%	89%	91%	85%	100%	94%	82%

Calculating Performance of our model



- **Positive Class (often referred to as the "positive" or "1" class)**:
  - **Precision (Positive Class)**: The ratio of true positives to the total predicted positives (i.e., true positives + false positives).

$$\text{Precision}_{\text{positive}} = \frac{TP}{TP + FP}$$

  - **Recall (Positive Class)**: The ratio of true positives to the total actual positives (i.e., true positives + false negatives).

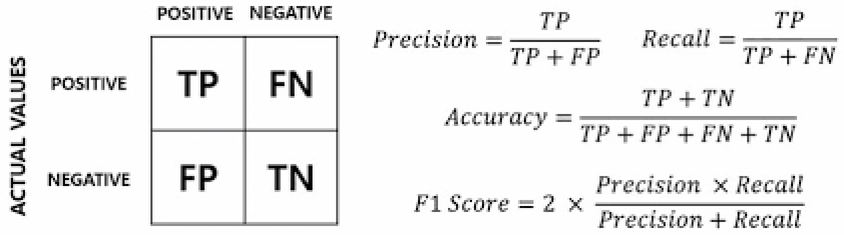
$$\text{Recall}_{\text{positive}} = \frac{TP}{TP + FN}$$

- **Negative Class (often referred to as the "negative" or "0" class)**:
  - **Precision (Negative Class)**: The ratio of true negatives to the total predicted negatives (i.e., true negatives + false negatives).

$$\text{Precision}_{\text{negative}} = \frac{TN}{TN + FN}$$

  - **Recall (Negative Class)**: The ratio of true negatives to the total actual negatives (i.e., true negatives + false positives).

$$\text{Recall}_{\text{negative}} = \frac{TN}{TN + FP}$$



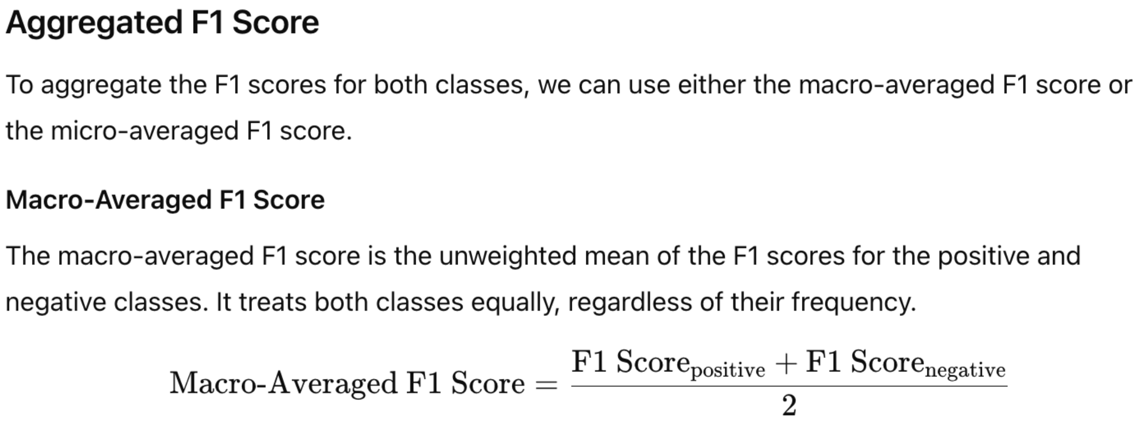
		POSITIVE	NEGATIVE
ACTUAL VALUES	POSITIVE	TP	FN
	NEGATIVE	FP	TN

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

F1 Score



Aggregated F1 Score

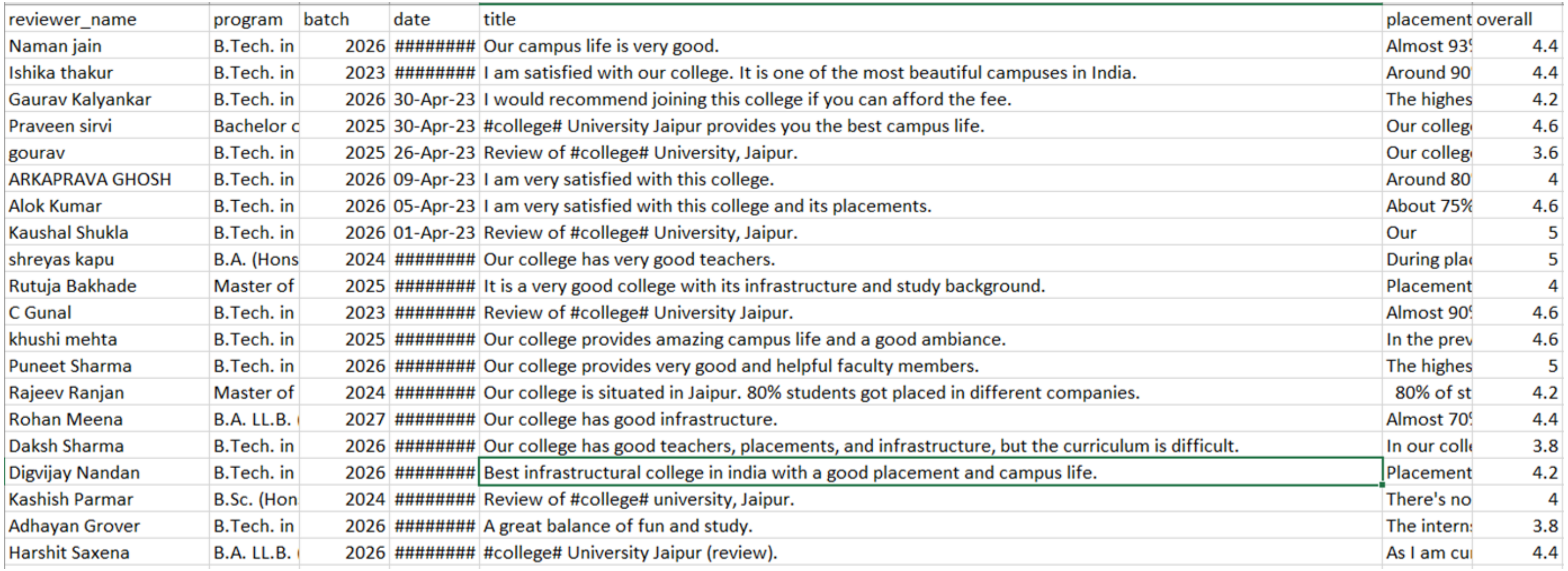
To aggregate the F1 scores for both classes, we can use either the macro-averaged F1 score or the micro-averaged F1 score.

Macro-Averaged F1 Score

The macro-averaged F1 score is the unweighted mean of the F1 scores for the positive and negative classes. It treats both classes equally, regardless of their frequency.

$$\text{Macro-Averaged F1 Score} = \frac{\text{F1 Score}_{\text{positive}} + \text{F1 Score}_{\text{negative}}}{2}$$

You want to choose a college for further studies. Now look at what people are saying about it.

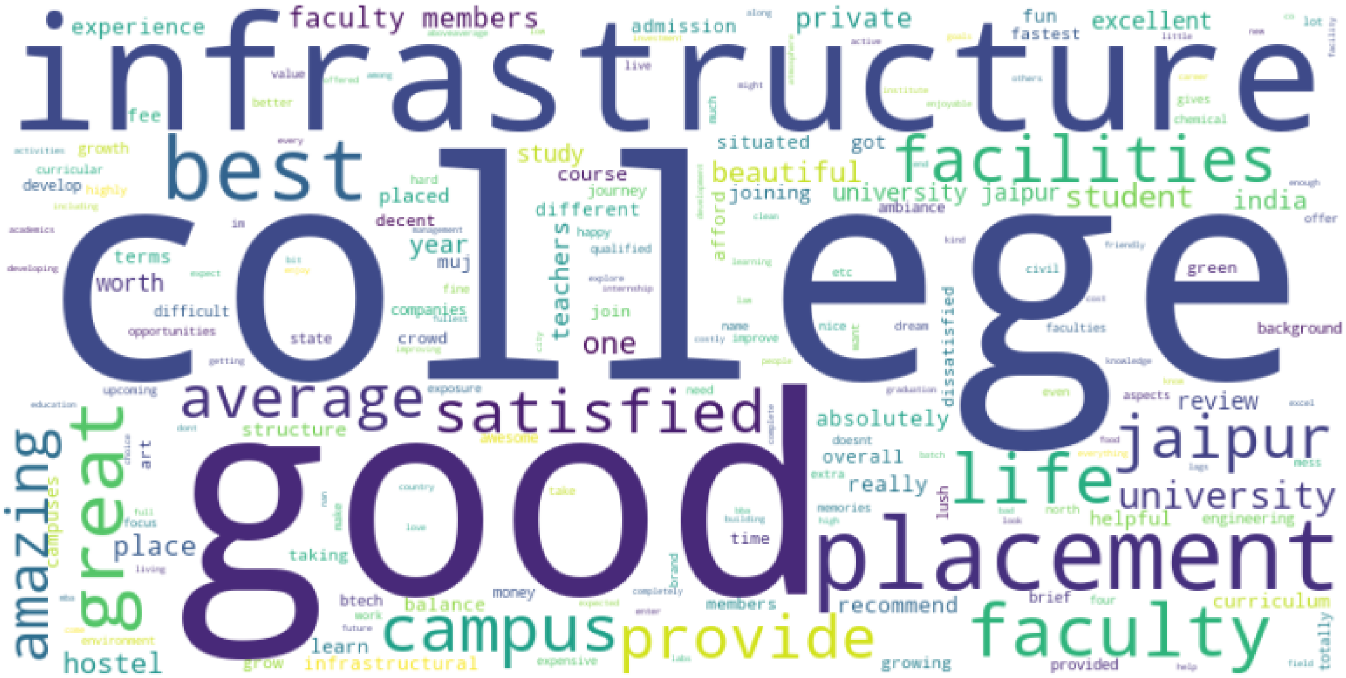


reviewer_name	program	batch	date	title	placement	overall
Naman jain	B.Tech. in	2026	########	Our campus life is very good.	Almost 93	4.4
Ishika thakur	B.Tech. in	2023	########	I am satisfied with our college. It is one of the most beautiful campuses in India.	Around 90	4.4
Gaurav Kalyankar	B.Tech. in	2026	30-Apr-23	I would recommend joining this college if you can afford the fee.	The highes	4.2
Praveen sirvi	Bachelor c	2025	30-Apr-23	#college# University Jaipur provides you the best campus life.	Our colleg	4.6
gourav	B.Tech. in	2025	26-Apr-23	Review of #college# University, Jaipur.	Our colleg	3.6
ARKAPRAVA GHOSH	B.Tech. in	2026	09-Apr-23	I am very satisfied with this college.	Around 80	4
Alok Kumar	B.Tech. in	2026	05-Apr-23	I am very satisfied with this college and its placements.	About 75%	4.6
Kaushal Shukla	B.Tech. in	2026	01-Apr-23	Review of #college# University, Jaipur.	Our	5
shreyas kapu	B.A. (Hons	2024	########	Our college has very good teachers.	During pla	5
Rutuja Bakhade	Master of	2025	########	It is a very good college with its infrastructure and study background.	Placement	4
C Gunal	B.Tech. in	2023	########	Review of #college# University Jaipur.	Almost 90	4.6
khushi mehta	B.Tech. in	2025	########	Our college provides amazing campus life and a good ambiance.	In the prev	4.6
Puneet Sharma	B.Tech. in	2026	########	Our college provides very good and helpful faculty members.	The highes	5
Rajeev Ranjan	Master of	2024	########	Our college is situated in Jaipur. 80% students got placed in different companies.	80% of st	4.2
Rohan Meena	B.A. LL.B.	2027	########	Our college has good infrastructure.	Almost 70	4.4
Daksh Sharma	B.Tech. in	2026	########	Our college has good teachers, placements, and infrastructure, but the curriculum is difficult.	In our coll	3.8
Digvijay Nandan	B.Tech. in	2026	########	Best infrastructural college in india with a good placement and campus life.	Placement	4.2
Kashish Parmar	B.Sc. (Hon	2024	########	Review of #college# university, Jaipur.	There's no	4
Adhayan Grover	B.Tech. in	2026	########	A great balance of fun and study.	The intern	3.8
Harshit Saxena	B.A. LL.B.	2026	########	#college# University Jaipur (review).	As I am cu	4.4

Note : Solve this problem using Unique, Count, CountIF and other functions in Excel. (Replicating Countvectorizer in Excel)

Solution using Word Cloud !

DATAPLAY



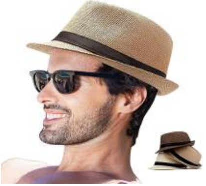
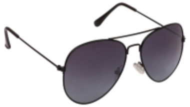
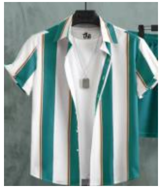
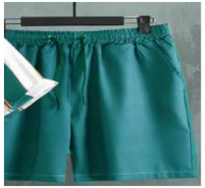
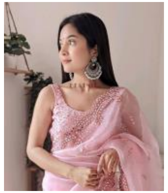
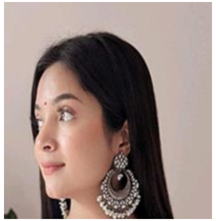
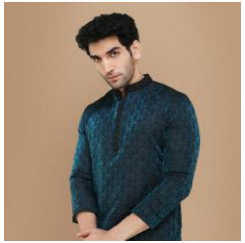
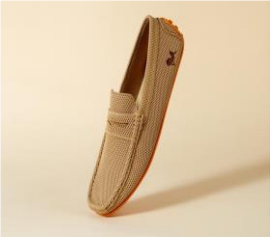
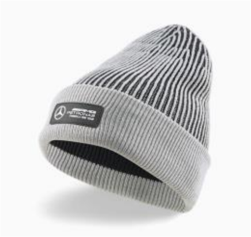
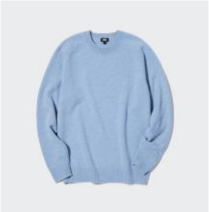
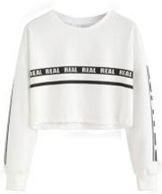
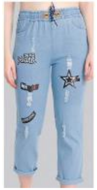
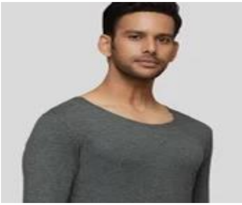
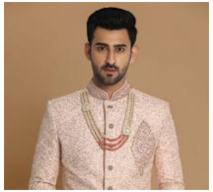
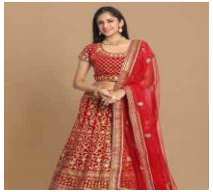
Mrs/Mr Khanna started an online clothing shopping store. She wants to increase the sales.

Lets help her out in increasing her numbers.

Solve this problem in Excel and generate insights for Mr/Mrs Khanna.

Product Catalogue

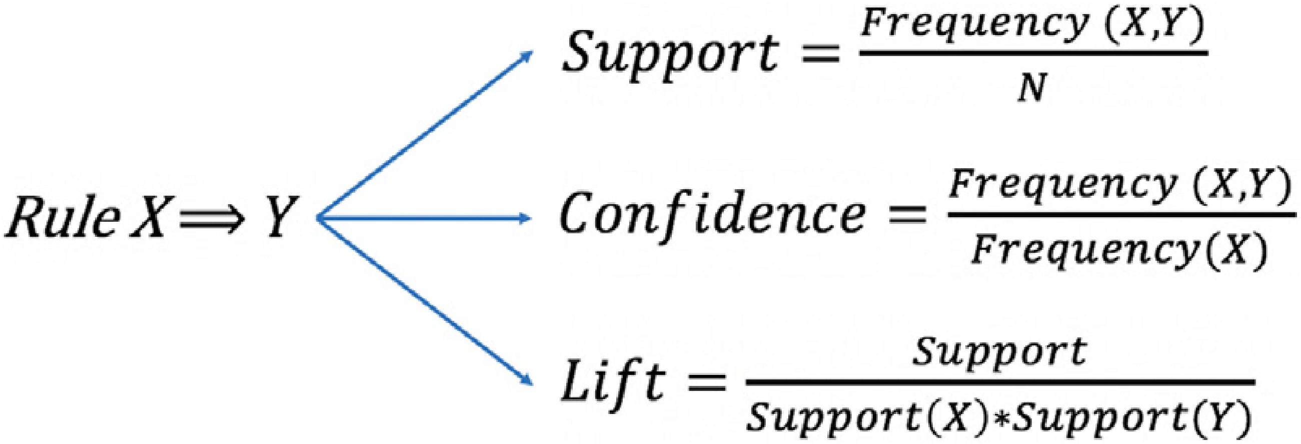
DATAPLAY

				
P001	P002	P003	P004	P005
				
P006	P007	P008	P009	P10
				
P10	P12	P13	P14	P15

OrderID	CustomerID	PurchaseDate (yy-mm-dd)	ProductID	Product	Quantity	UnitPrice
1001	101	27-06-2023	P01	Summer Cap	2	100
1001	101	27-06-2023	P02	Sunglasses	1	50
1002	102	28-07-2023	P07	Kurta		150
1003	103	29-07-2023	P03	Half Sleeve T-shirt	1	200
1003	103	29-07-2023	P04	Capri	2	350
1004	104	31-08-2023	P05	Saree	1	400
1004	104	31-08-2023	P06	Earrings	1	30
1005	105	01-09-2023	P03	Half Sleeve T-shirt	1	200
1005	105	01-09-2023	P04	Capri	2	350
1006	105	02-09-2023	P07	Kurta	2	150
1007	104	07-09-2023	P05	Saree	2	400
1007	104	07-09-2023	P06	Earrings	1	30
1008	106	05-10-2023	P15	Lahenga	1	4000
1009	107	16-10-2023	P14	Sherwani	1	2000
1010	108	26-10-2023	P15	Lahenga	1	4000
1011	103	27-10-2023	P03	Half Sleeve T-shirt	1	200
1011	103	29-07-2023	P04	Capri	2	350
1012	109	27-10-2023	P14	Sherwani	1	2000
1013	110	28-10-2023	P15	Lahenga	1	4000
1014	111	29-10-2023	P14	Sherwani	1	2000
1015	101	01-11-2023	P10	Sweatshirt	2	250
1016	103	02-11-2023	P11	Long Sleeve T-shirt	1	300
1016	103	02-11-2023	P12	Jeans	1	600
1017	103	04-11-2023	P13	Thermocoat	1	270
1018	101	05-11-2023	P10	Sweatshirt	2	250
1019	105	06-11-2023	P11	Long Sleeve T-shirt	1	300
1019	105	06-11-2023	P12	Jeans	1	600
1020	106	10-12-2023	P05	Saree	1	400
1020	106	10-12-2023	P06	Earrings	1	30
1021	107	01-01-2024	P11	Long Sleeve T-shirt	1	300
1021	107	01-01-2024	P12	Jeans	1	600
1022	101	01-02-2024	P09	Winter Cap	1	150
1022	101	01-02-2024	P02	Sunglasses	1	50
1023	111	03-02-2024	P09	Winter Cap	1	150
1023	111	03-02-2024	P02	Sunglasses	1	50

Association Rule Mining

DATAPLAY

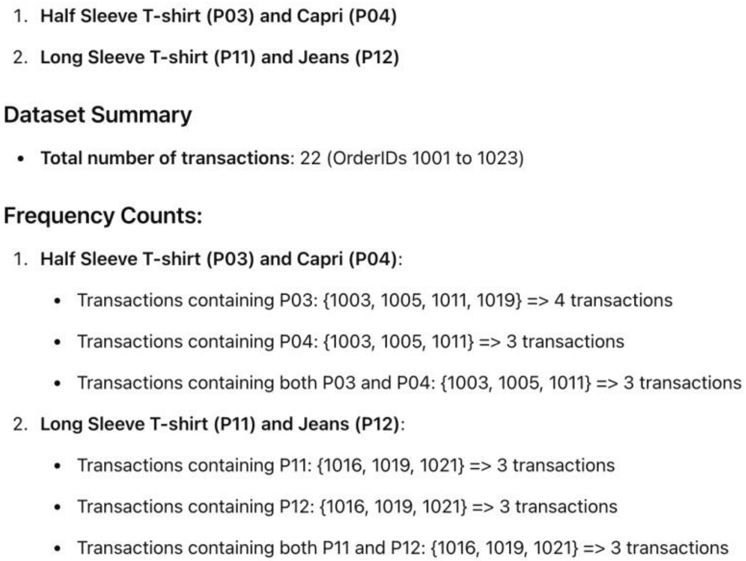


$$Rule\ X \Longrightarrow Y$$

$$Support = \frac{Frequency\ (X,Y)}{N}$$

$$Confidence = \frac{Frequency\ (X,Y)}{Frequency(X)}$$

$$Lift = \frac{Support}{Support(X) * Support(Y)}$$



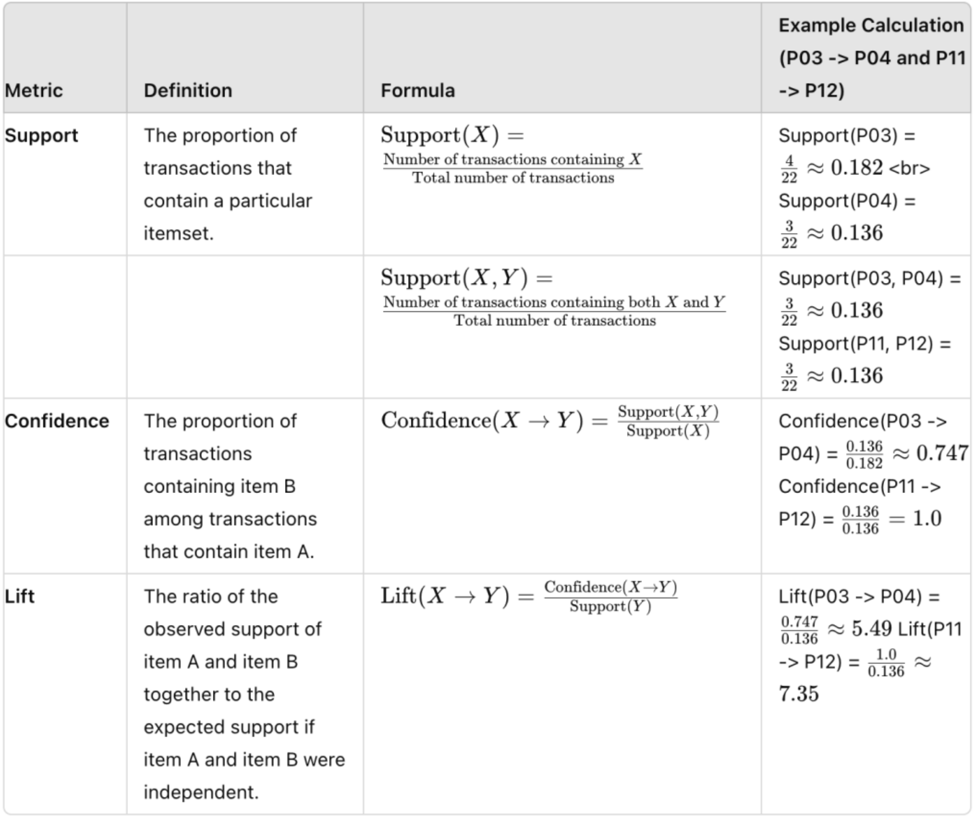
1. **Half Sleeve T-shirt (P03) and Capri (P04)**
2. **Long Sleeve T-shirt (P11) and Jeans (P12)**

Dataset Summary

- **Total number of transactions**: 22 (OrderIDs 1001 to 1023)

Frequency Counts:

1. **Half Sleeve T-shirt (P03) and Capri (P04)**:
   - Transactions containing P03: {1003, 1005, 1011, 1019} => 4 transactions
   - Transactions containing P04: {1003, 1005, 1011} => 3 transactions
   - Transactions containing both P03 and P04: {1003, 1005, 1011} => 3 transactions
2. **Long Sleeve T-shirt (P11) and Jeans (P12)**:
   - Transactions containing P11: {1016, 1019, 1021} => 3 transactions
   - Transactions containing P12: {1016, 1019, 1021} => 3 transactions
   - Transactions containing both P11 and P12: {1016, 1019, 1021} => 3 transactions



Metric	Definition	Formula	Example Calculation (P03 -> P04 and P11 -> P12)
Support	The proportion of transactions that contain a particular itemset.	$\text{Support}(X) = \frac{\text{Number of transactions containing } X}{\text{Total number of transactions}}$	Support(P03) = $\frac{4}{22} \approx 0.182$ Support(P04) = $\frac{3}{22} \approx 0.136$
		$\text{Support}(X, Y) = \frac{\text{Number of transactions containing both } X \text{ and } Y}{\text{Total number of transactions}}$	Support(P03, P04) = $\frac{3}{22} \approx 0.136$ Support(P11, P12) = $\frac{3}{22} \approx 0.136$
Confidence	The proportion of transactions containing item B among transactions that contain item A.	$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support}(X,Y)}{\text{Support}(X)}$	Confidence(P03 -> P04) = $\frac{0.136}{0.182} \approx 0.747$ Confidence(P11 -> P12) = $\frac{0.136}{0.136} = 1.0$
Lift	The ratio of the observed support of item A and item B together to the expected support if item A and item B were independent.	$\text{Lift}(X \rightarrow Y) = \frac{\text{Confidence}(X \rightarrow Y)}{\text{Support}(Y)}$	Lift(P03 -> P04) = $\frac{0.747}{0.136} \approx 5.49$ Lift(P11 -> P12) = $\frac{1.0}{0.136} \approx 7.35$

What all we can do?

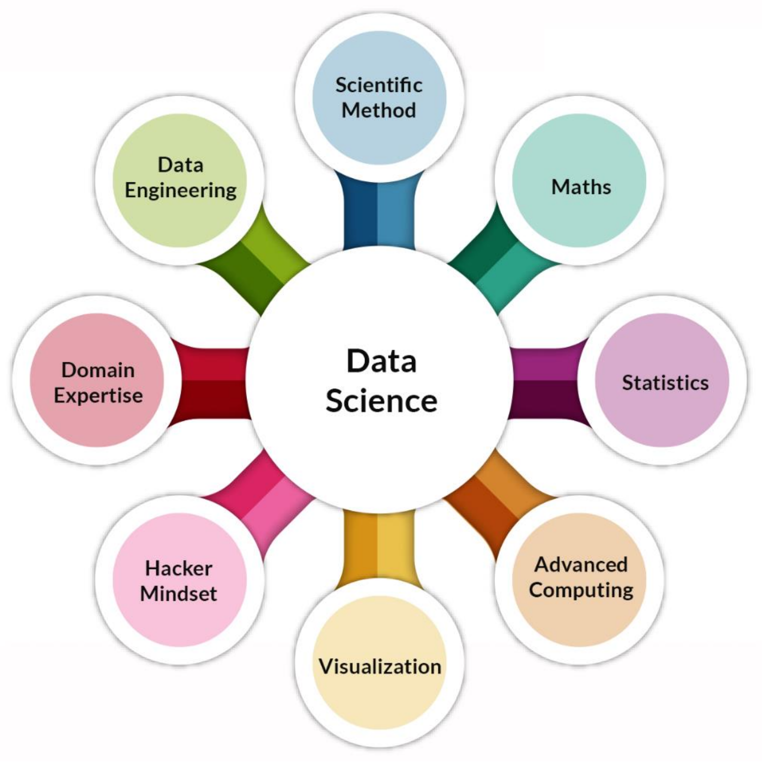
Product Bundling
Seasonal Promotions
Customer Loyalty Programme
Personalized Targeting

Market Basket Analysis

Market basket analysis (also known as association analysis or affinity analysis) is a data mining and analytics technique used by retailers and businesses to discover patterns and relationships in customer purchase data.

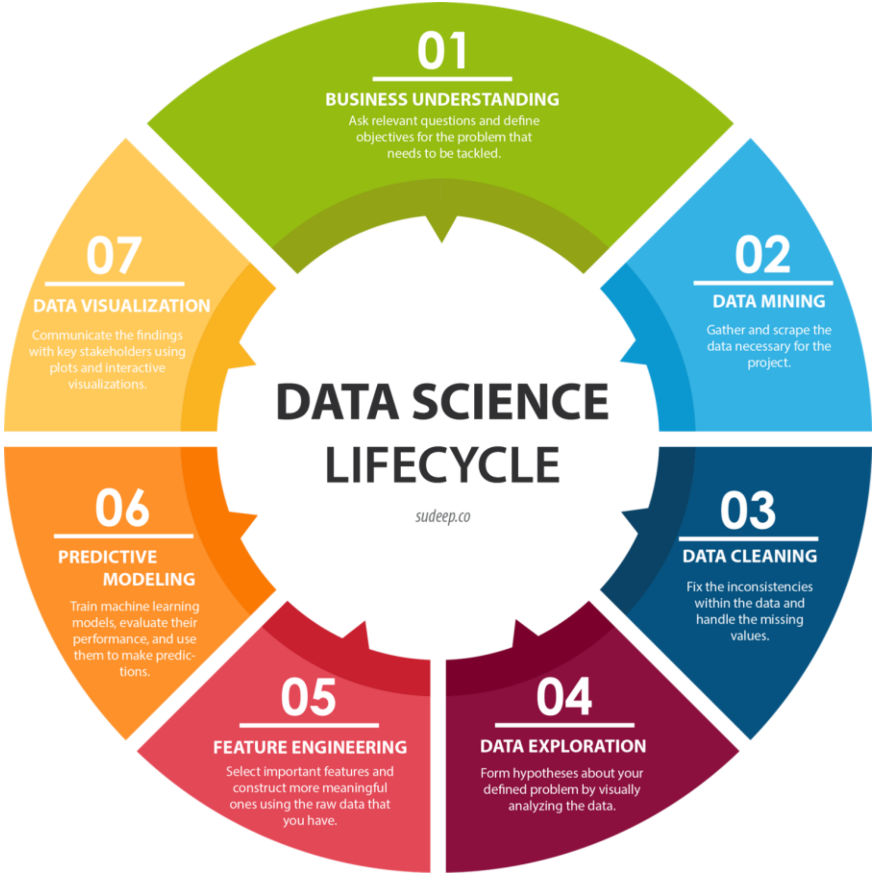
Components of Data Science

- Data
- Big Data
- Domain Expertise
- Data Engineering
- Mathematics
- Stats & Prob
- Machine Learning
- Programming Languages
- Visualization and Operationalization
- Advanced Computing
- Data Analysis and Models
- Development Tools



Data Science Life Cycle

DATAPLAY



DATA SCIENCE LIFECYCLE

sudeep.co

01 BUSINESS UNDERSTANDING

Ask relevant questions and define objectives for the problem that needs to be tackled.

02 DATA MINING

Gather and scrape the data necessary for the project.

03 DATA CLEANING

Fix the inconsistencies within the data and handle the missing values.

04 DATA EXPLORATION

Form hypotheses about your defined problem by visually analyzing the data.

05 FEATURE ENGINEERING

Select important features and construct more meaningful ones using the raw data that you have.

06 PREDICTIVE MODELING

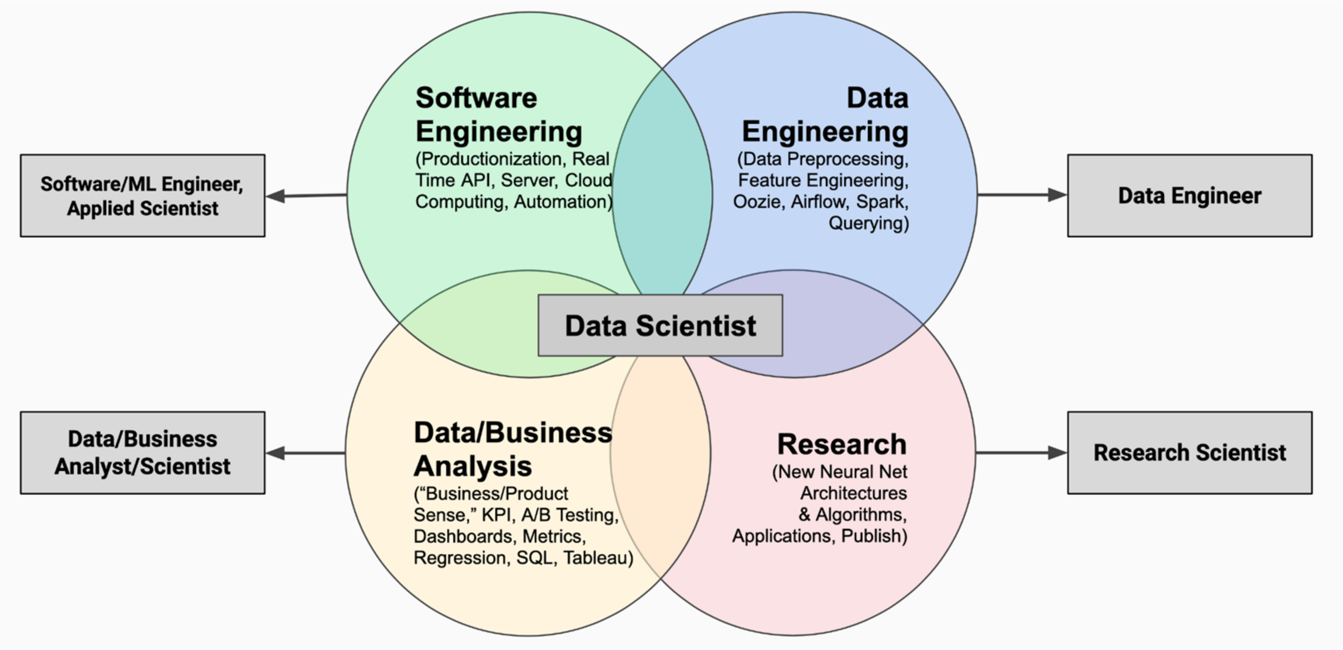
Train machine learning models, evaluate their performance, and use them to make predictions.

07 DATA VISUALIZATION

Communicate the findings with key stakeholders using plots and interactive visualizations.

Market of Data Science

DATAPLAY



Software Engineering
(Productionization, Real Time API, Server, Cloud Computing, Automation)

Data Engineering
(Data Preprocessing, Feature Engineering, Oozie, Airflow, Spark, Querying)

Data/Business Analysis
("Business/Product Sense," KPI, A/B Testing, Dashboards, Metrics, Regression, SQL, Tableau)

Research
(New Neural Net Architectures & Algorithms, Applications, Publish)

Data Scientist

Software/ML Engineer, Applied Scientist

Data Engineer

Data/Business Analyst/Scientist

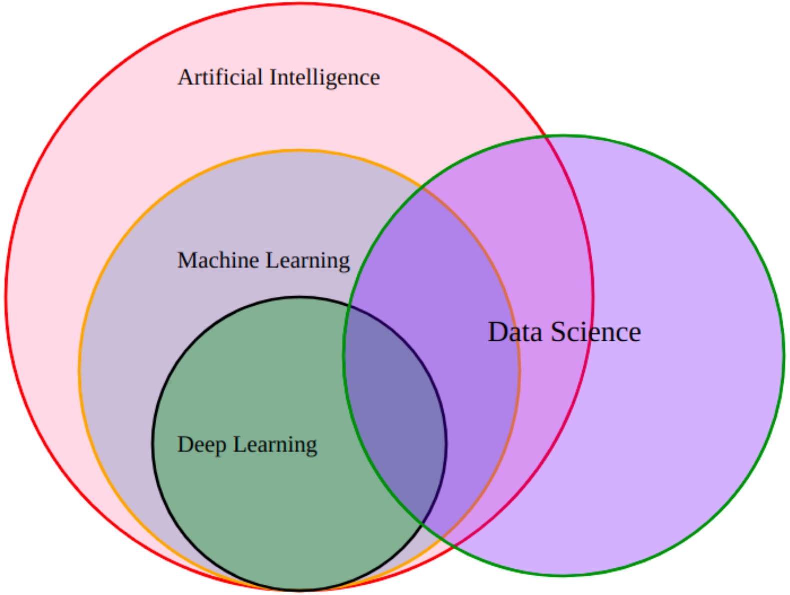
Research Scientist

DATAPLAY

Topics	Data Analyst / Business Analyst	Data Analyst / Business Analyst	Data Scientist	Data Scientist
	Entry Level	**Higher Level**	**Entry Level**	**Higher Level**
Excel	Expert	Expert	Expert	Expert
SQL	Intermediate - Expert	Expert - Proficient	Intermediate	Expert - Proficient
Python	No	Beginner	Expert	Proficient
Tableau / Power BI	Expert	Expert	Intermediate	No
Approximation Questions	Not mandatory	Expert	No	Yes
Machine Learning	No	Beginner	Expert	Proficient
Deep Learning	No	No	Depends on role	Depends on role

What do you understand by these Jargons ?

- Artificial Intelligence
- Data Science
- Machine Learning
- Deep Learning
    - Computer Vision
    - Natural Language Processing



DATAPLAY

- **Artificial Intelligence (AI):** AI is like teaching computers to think and learn like people do. Computers can do things, make choices, and solve problems all by themselves!

- **Data Science:** Data Science is when we use special computer tricks to look at a lot of information and find cool stuff that helps us understand things better. It's like being a detective for numbers!

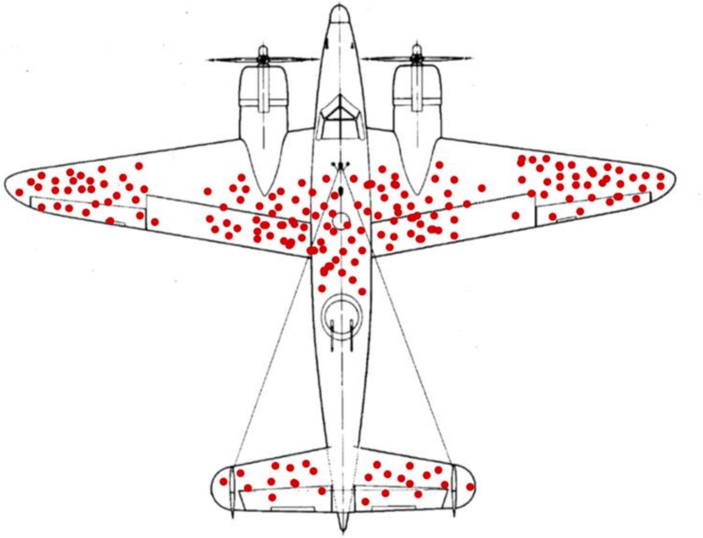
- **Machine Learning (ML):** Machine Learning is a type of computer learning where we show the computer many examples so it can learn and guess things on its own. It's like teaching a computer to guess if it will rain tomorrow.

- **Deep Learning:** Deep Learning is a bit like Machine Learning, but it's even smarter. It uses computer networks that are good at finding really tricky patterns, like telling the difference between cats and dogs in pictures.

Rise of Data Analytics

Abraham Wald & the missing Bullet holes (World War II)

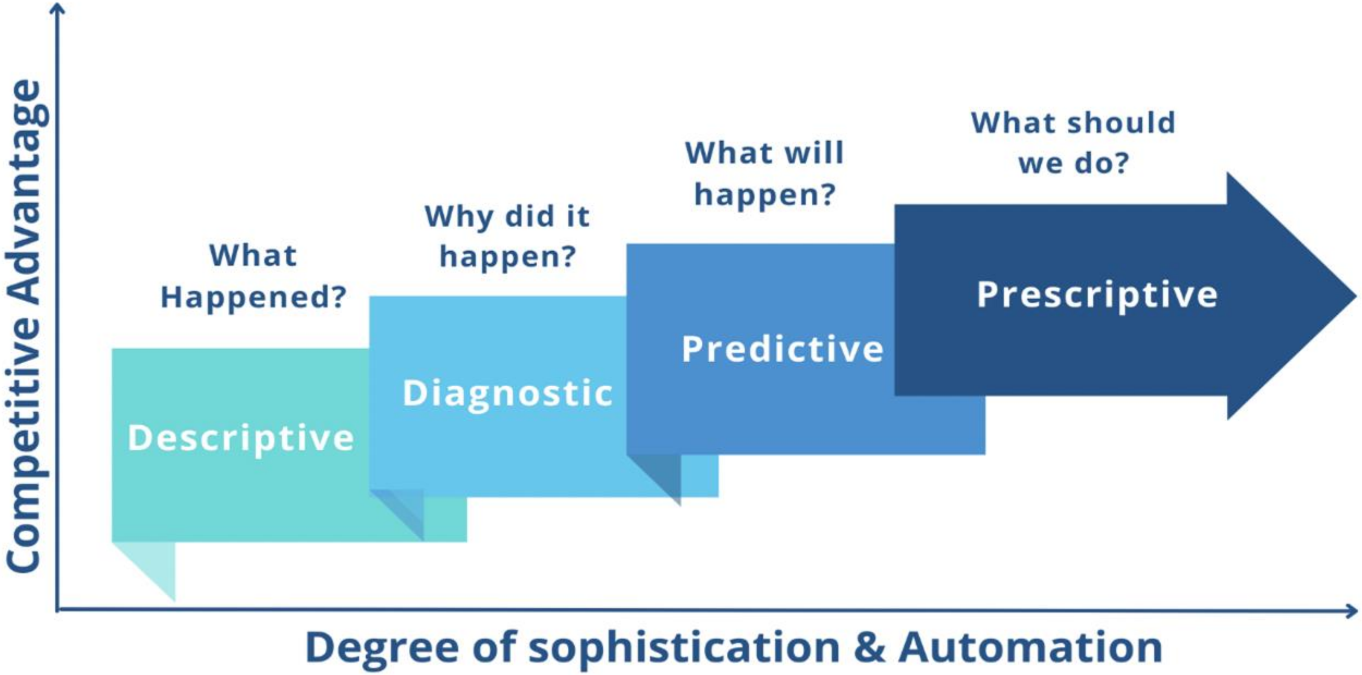
DATAPLAY



The missing bullet holes were on the missing planes.

Types of Data Analytics

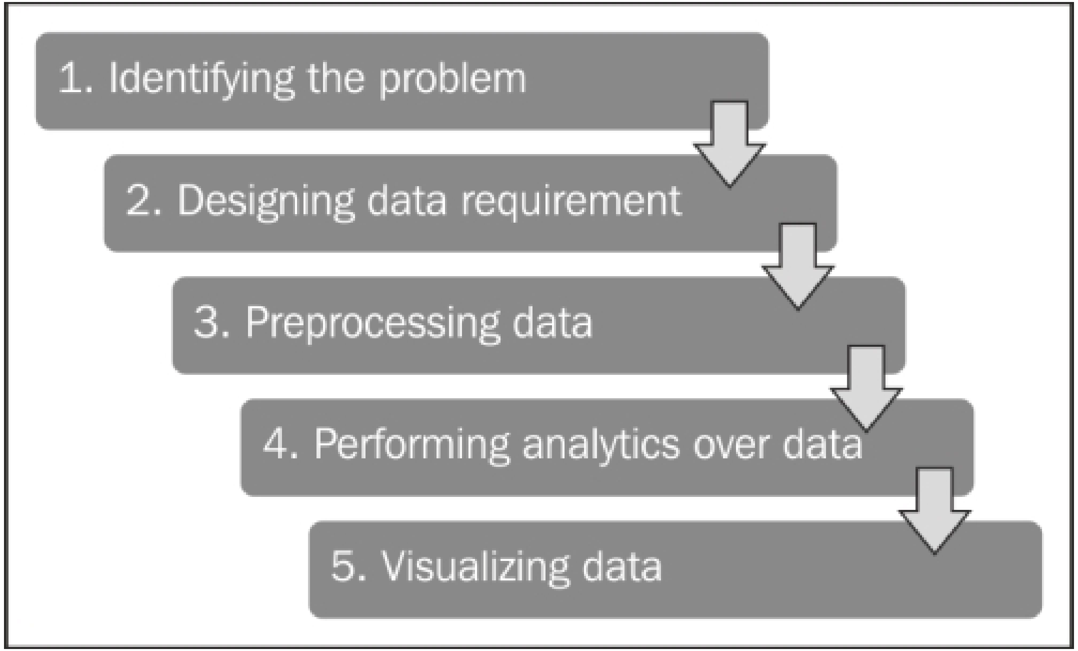
DATAPLAY



Data Analysis for Heart Disease

Descriptive : 70% People die of Heart Attack
Diagnostic : High Cholesterol Levels
Predictive : Adults below age of 30 are likely to witness Heart Problems
Prescriptive : Avoid Junk Food / Monitor Junk Food Outlets

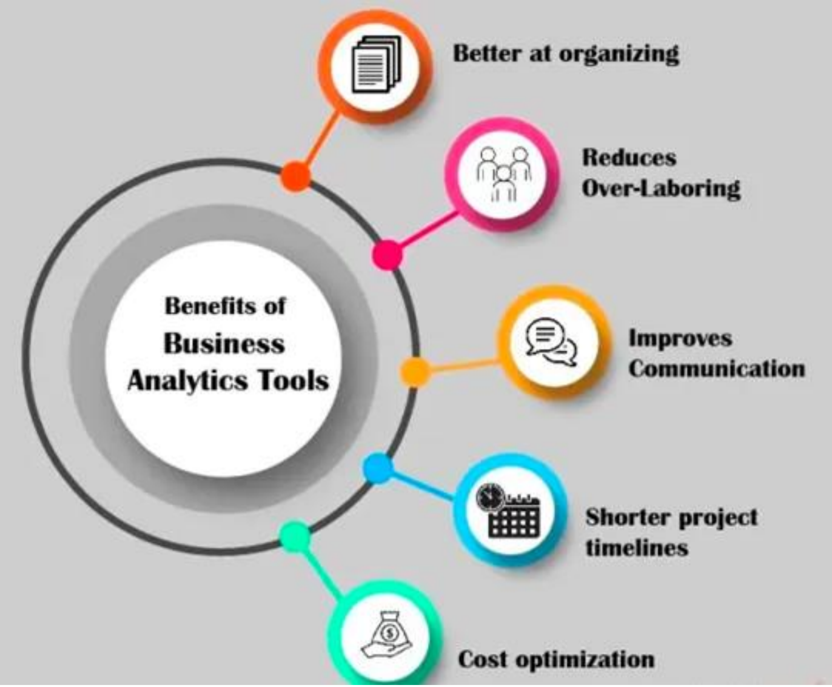
Data Analytics Lifecycle



1. Identifying the problem
2. Designing data requirement
3. Preprocessing data
4. Performing analytics over data
5. Visualizing data

Need for Business Analytics

DATAPLAY


Benefits of Business Analytics Tools
Better at organizing
Reduces Over-Laboring
Improves Communication
Shorter project timelines
Cost optimization