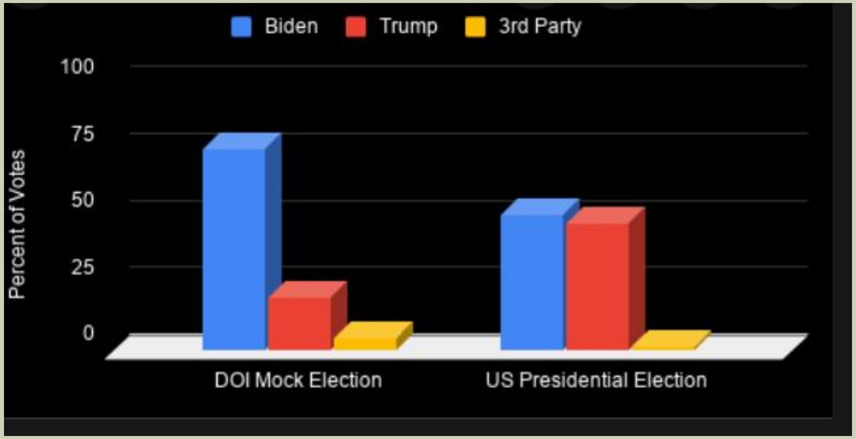
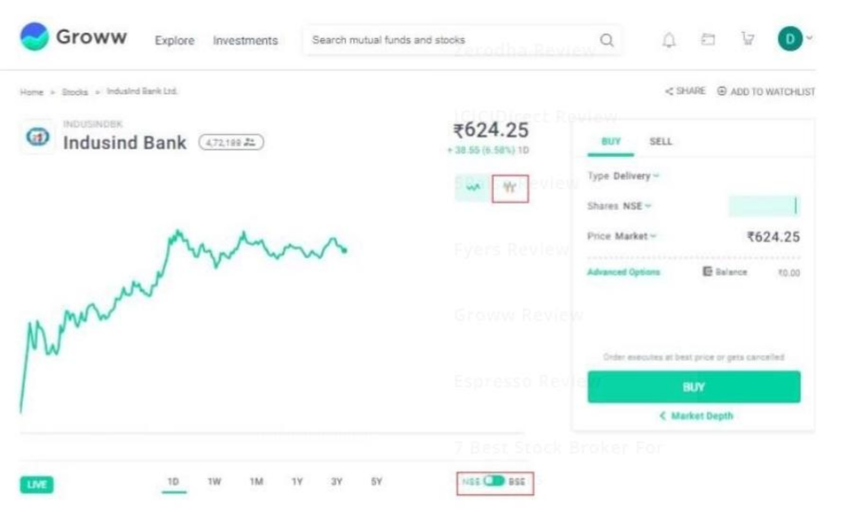
Data Analytics Techniques, Applications & Use Cases

Presented By:
Dr. Jagannath Singh
Associate Professor, School of Comp. Engg.
Kalinga Institute of Industrial Technology, BBSR


Biden
Trump
3rd Party
Percent of Votes
100
75
50
25
0
DOI Mock Election
US Presidential Election

Seasonal

Network Optimization
Customer Experience
Fraud Detection


Groww
Explore
Investments
Search mutual funds and stocks
Home > Stocks > IndusInd Bank Ltd.
SHARE
ADD TO WATCHLIST
INDUSINDBK
Indusind Bank
₹624.25
BUY
SELL
Type Delivery
Shares NSE
Price Market
₹624.25
Advanced Options
Balance
₹0.00
Order executes at best price or gets cancelled
BUY
Market Depth
LIVE
1D
1W
1M
1Y
3Y
5Y
NSE
BSE

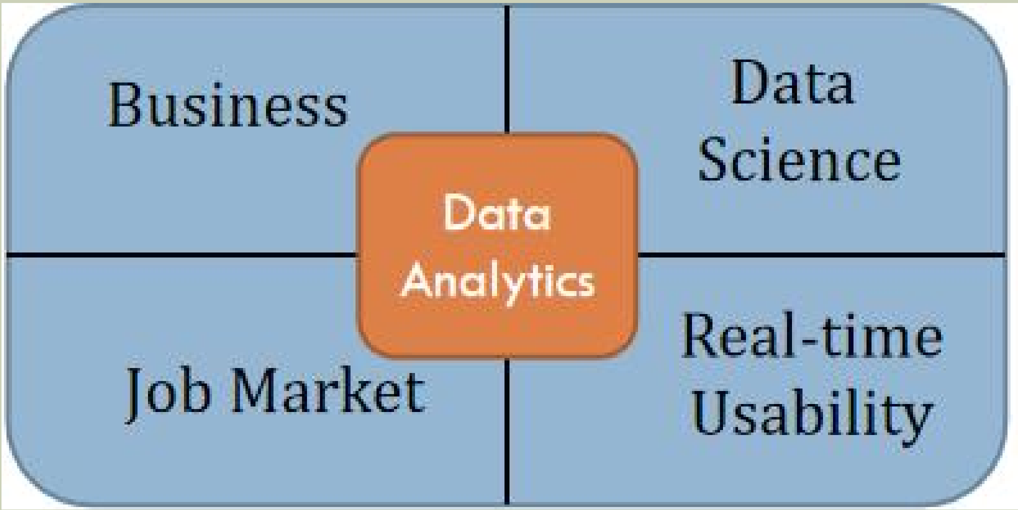
Data
Analysis
Idea

Outline

1. Need of the Topic
2. Introdction to Data
2. What is of Big Data?
3. Data Analytics & types
4. Regression Techniques
5. Market Basket Analysis
6. Conculsion

Importance of the topic

- The data analytics is indeed a revolution in the field of information technology.
- The use of data analytics by the companies is enhancing.
- Many organizations are actively looking out for the right talent to analyze vast amounts of data.



Why Learn Data Analytics?

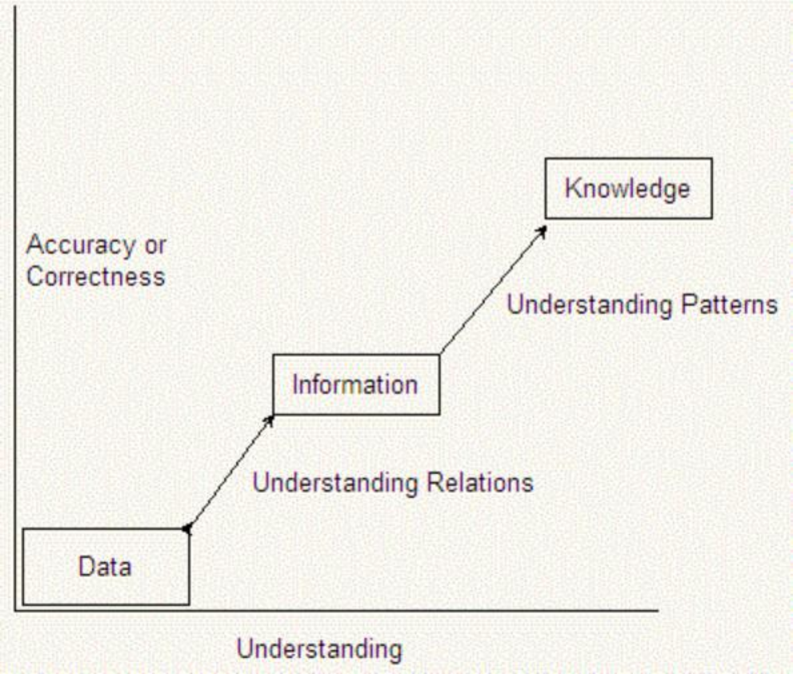
- A priority for top organizations.
- Gain problem solving skills.
- High demand: Increasing job opportunities and pay.
- Analytics is everywhere.
- It's only becoming more important.
- It represents perfect freelancing opportunities.
- Develop new revenue streams.

Data

Definition: A representation of information, knowledge, facts, concepts or instructions which are being prepared or have been prepared in a formalized manner.

- Data is the plural of datum
- It must be interpreted, by a human or machine to derive meaning.

Data

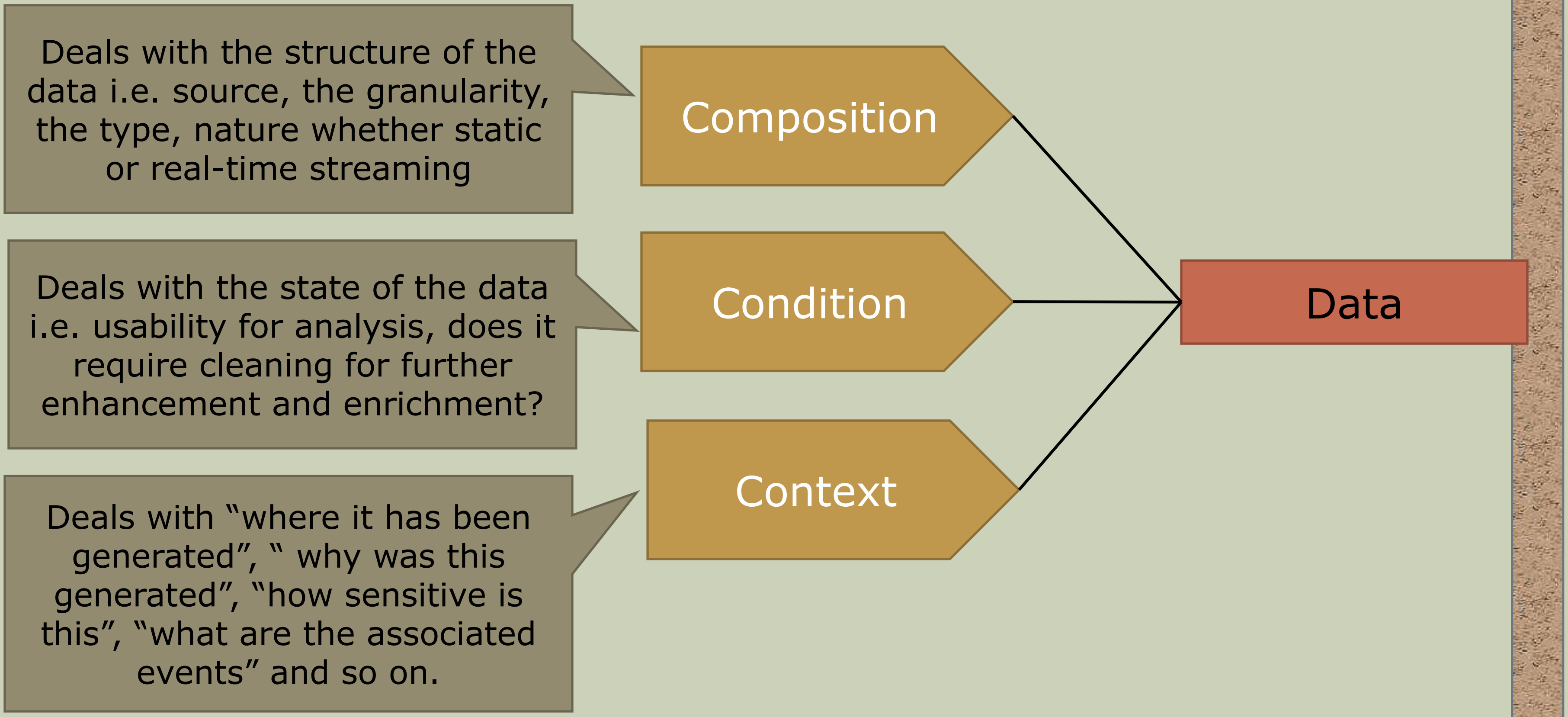


Data → Information → Knowledge → Actionable Insights

Importance of Data

- ➢ Increasingly important to businesses, cure a disease, boost a company's revenue, understand and interpret market trends, study customer behavior and take financial decisions.
- ➢ Managers may need to understand high volumes of data before they can make the necessary decisions.
- ➢ Relevant data creates strong strategies
- ➢ It helps in identifying real problems.
- ➢ Data improves quality of life.

Characteristics of Data



Deals with the structure of the data i.e. source, the granularity, the type, nature whether static or real-time streaming

Composition

Deals with the state of the data i.e. usability for analysis, does it require cleaning for further enhancement and enrichment?

Condition

Deals with “where it has been generated”, “ why was this generated”, “how sensitive is this”, “what are the associated events” and so on.

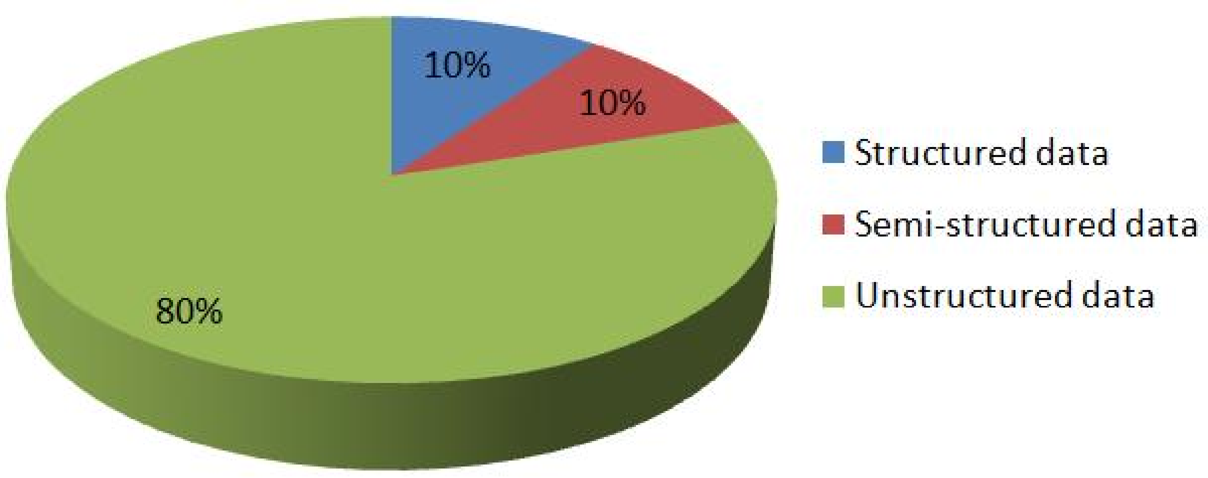
Context

Data

Classification of Digital Data

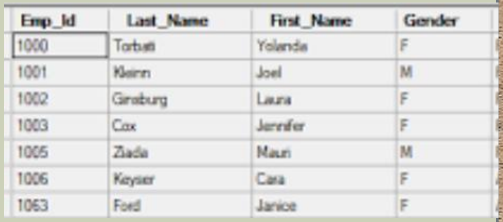
- Digital data is classified into the following categories:

1. Structured data
2. Semi-structured data
3. Unstructured data



Classification of Digital Data…

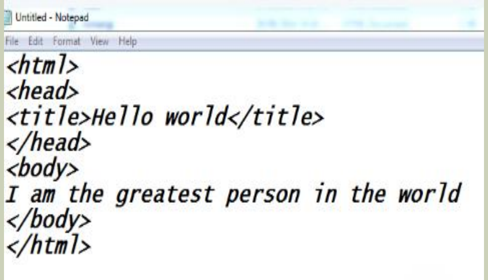
Structured: It is defined as the data that has a well-defined repeating pattern and this pattern makes it easier for any program to sort, read, and process the data.



Emp_Id	Last_Name	First_Name	Gender
1000	Torbati	Yolanda	F
1001	Kleinn	Joel	M
1002	Ginsburg	Laura	F
1003	Cox	Jennifer	F
1005	Ziada	Mauri	M
1006	Keyser	Cara	F
1063	Ford	Janice	F

Semi-structured: These data known to have a schema-less or self-describing structure.

Example, emails, XML, markup languages like HTML.

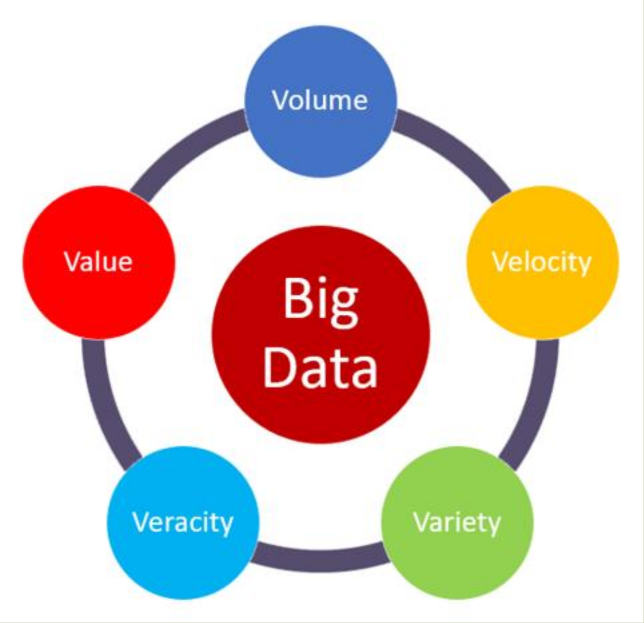


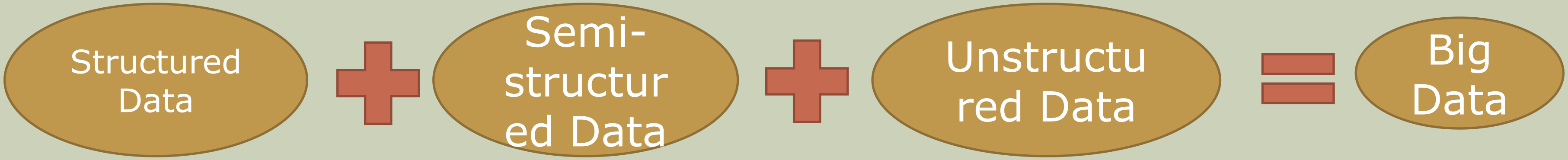
```
<html>
<head>
<title>Hello world</title>
</head>
<body>
I am the greatest person in the world
</body>
</html>
```

Unstructured: It is a set of data that might or might not have any logical or repeating patterns.

Big Data

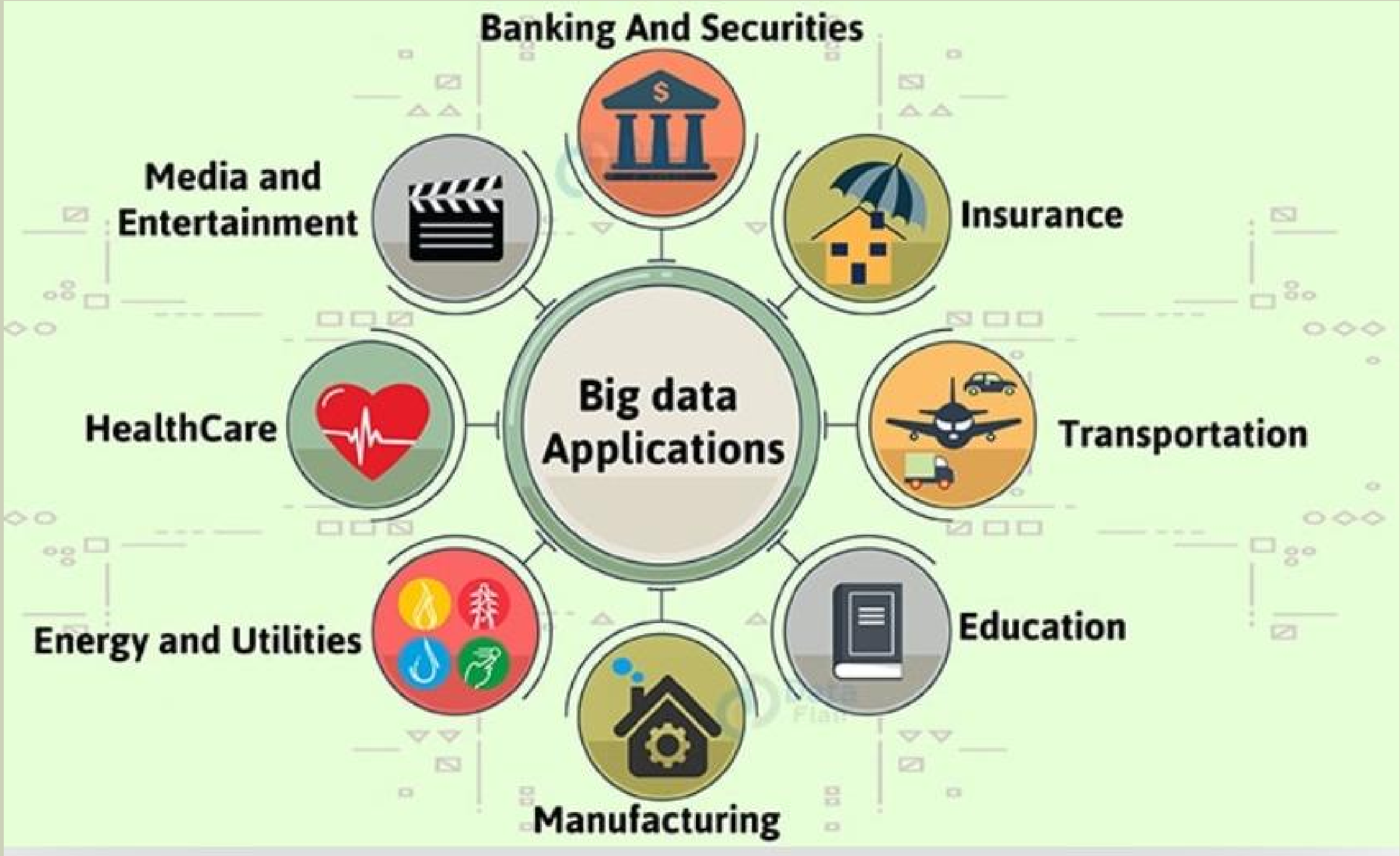
Big Data is high-volume, high-velocity, and high-variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making.





Structured Data + Semi-structured Data + Unstructured Data = Big Data

Big Data Usage



Data Analytics

- Rapid advances in computing, data storage, networks etc have dramatically increased the ability to access, store, and process huge amount of data.
- It is the need of the hour to extract relevant information from the huge amounts of data from heterogeneous data sources such as sensors, text achieves, images, videos, audio etc.
- In such voluminous data, general patterns, structures, regularities go undetected. But, such patterns are very useful.

Data Analytics

- In the data-rich age, understanding how to analyze and extract true meaning from the insights is one of the primary drivers of success.
- With so much data and so little time, knowing how to collect, clean, organize, and make sense of all of this potentially business-boosting information can be a game changer...

Data Analysis is the solution.

Analytical questions

- **Retailers** use it to understand their customer needs and buying habits to predict trends, recommend new products, and boost their business.
- **Healthcare** industries analyze patient data to provide lifesaving treatment options. They also deal with healthcare plans, insurance information to derive key insights.
- **Manufacturing** industries can discover new cost-saving and revenue opportunities. They can solve complex supply chain issues, labor constraints, and equipment breakdowns.
- **Banking** institutions gather and access large volumes of data to derive analytical insights and make sound financial decisions. They find out probable loan defaulters, customer churn out rate, and detect frauds in transactions.
- **Logistics** companies use data analytics to develop new business models, optimize routes, improve productivity, and order processing capabilities as well as performance management.

Data Analysis Vs Data Analytics

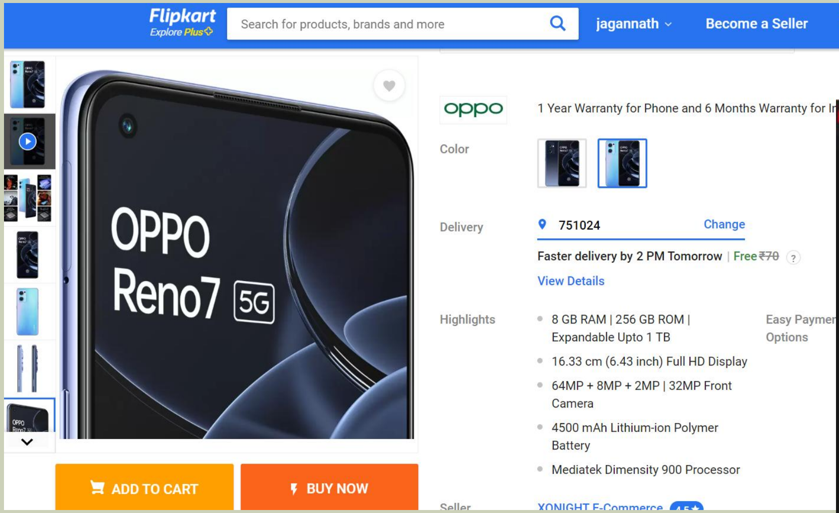
- Data Analysis is a process of inspecting, cleaning, transforming and modeling data with the goal of discovering useful information.

- Based on study of past data.

- Data analytics refers to the process of examining datasets to uncover patterns to take next step in decision making.

- It is to detemine the future planing.

Data Analytics



oppo 1 Year Warranty for Phone and 6 Months Warranty for In

Color

Delivery 751024 Change

Faster delivery by 2 PM Tomorrow | Free ~~₹70~~ ?

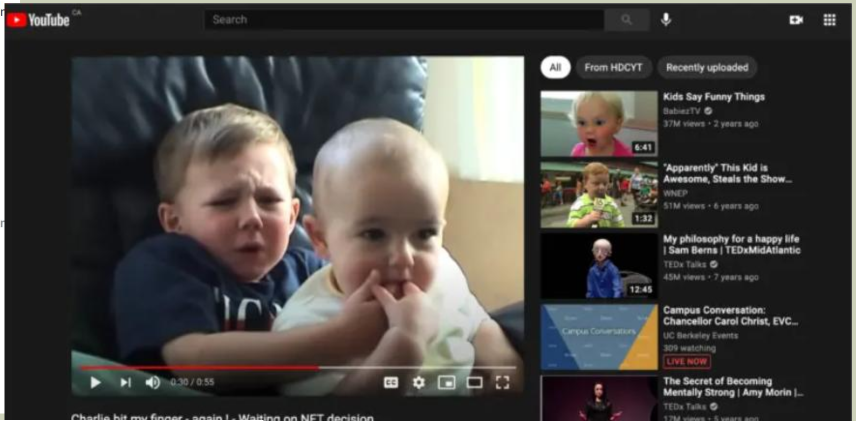
View Details

Highlights

- 8 GB RAM | 256 GB ROM | Expandable Upto 1 TB
- 16.33 cm (6.43 inch) Full HD Display
- 64MP + 8MP + 2MP | 32MP Front Camera
- 4500 mAh Lithium-ion Polymer Battery
- Mediatek Dimensity 900 Processor

Easy Paymen Options

Seller XONIGHT E-Commerce



Charlie bit my finger - again ! - Waiting on NFT decision

Kids Say Funny Things
BabiezTV
37M views · 2 years ago

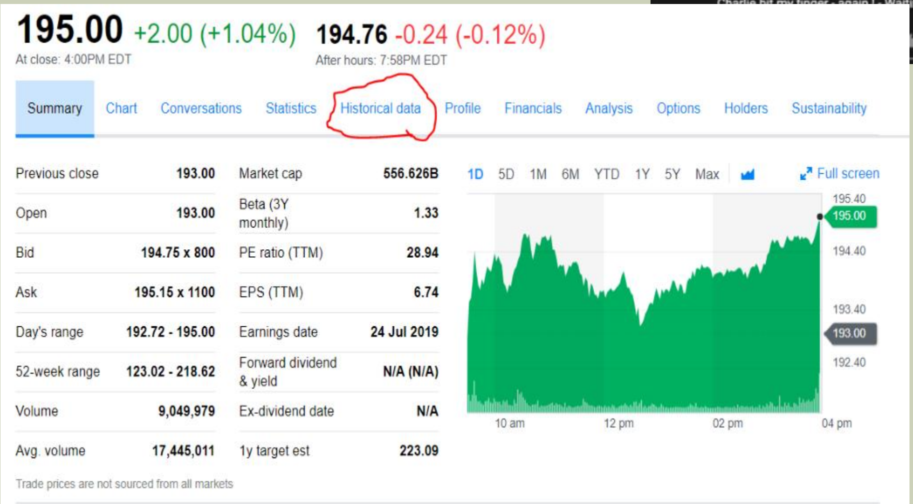
"Apparently" This Kid is Awesome, Steals the Show...
WNEP
51M views · 6 years ago

My philosophy for a happy life | Sam Berns | TEDxMidAtlantic
TEDx Talks
45M views · 7 years ago

Campus Conversation: Chancellor Carol Christ, EVC...
UC Berkeley Events
309 watching

The Secret of Becoming Mentally Strong | Amy Morin |...
TEDx Talks
17M views · 5 years ago

'Friends' Reunion Exclusive: Cast Reflects on Beloved Sho...
People



195.00 +2.00 (+1.04%)
At close: 4:00PM EDT

194.76 -0.24 (-0.12%)
After hours: 7:58PM EDT

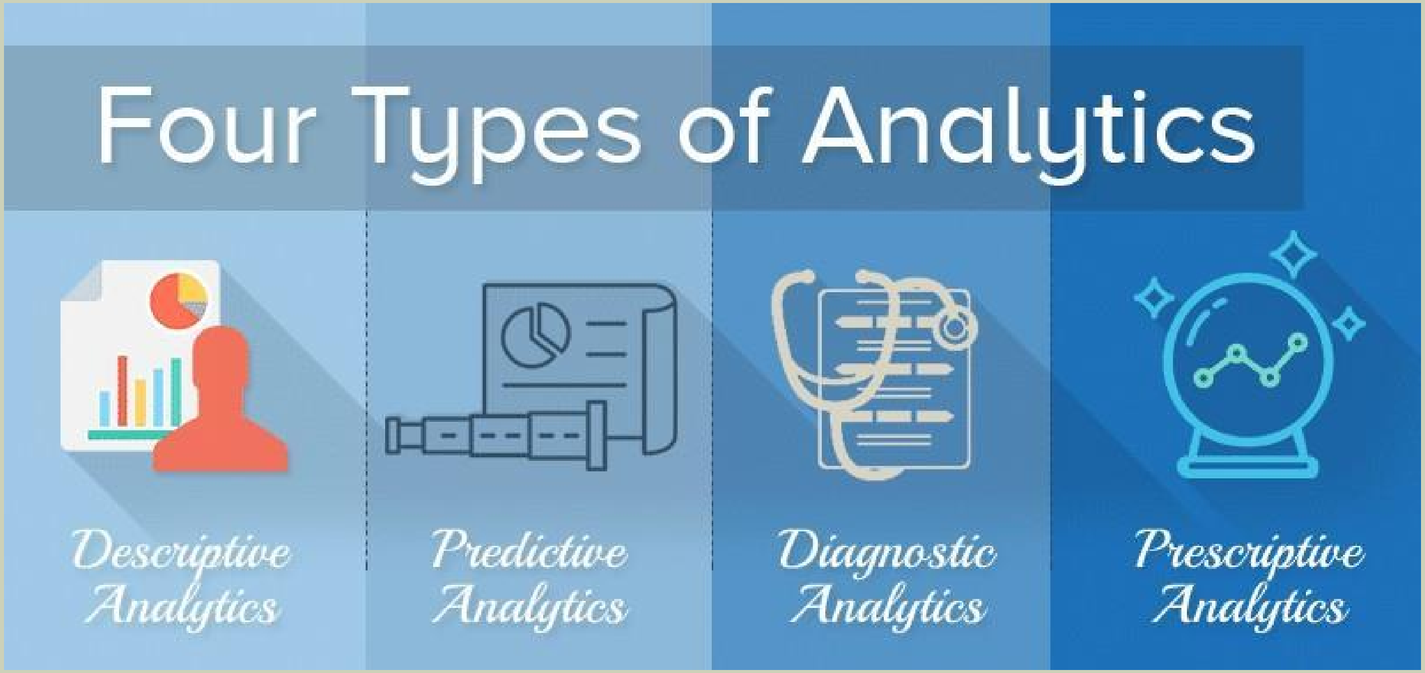
Previous close	193.00	Market cap	556.626B
Open	193.00	Beta (3Y monthly)	1.33
Bid	194.75 x 800	PE ratio (TTM)	28.94
Ask	195.15 x 1100	EPS (TTM)	6.74
Day's range	192.72 - 195.00	Earnings date	24 Jul 2019
52-week range	123.02 - 218.62	Forward dividend & yield	N/A (N/A)
Volume	9,049,979	Ex-dividend date	N/A
Avg. volume	17,445,011	1y target est	223.09

Trade prices are not sourced from all markets

Dr. Jagannath Singh
School of Computer Engineering

KIIT
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
Deemed to be University U/S 3 of the UGC Act, 1956

Types of Data Analytics



What happened?

What might happen in the future?

Why did this happen?

What should we do next?

Importance of Data Analytics

- Data analysis is important to businesses. In fact, no business can survive without analyzing available data. Use Cases:

1. A pharmacy company is performing trials on number of patients to test its new drug to fight cancer. The number of patients under the trial is well over 500.

2. A company wants to launch new variant of its existing line of fruit juice. It wants to carry out the survey analysis and arrive at some meaningful conclusion.

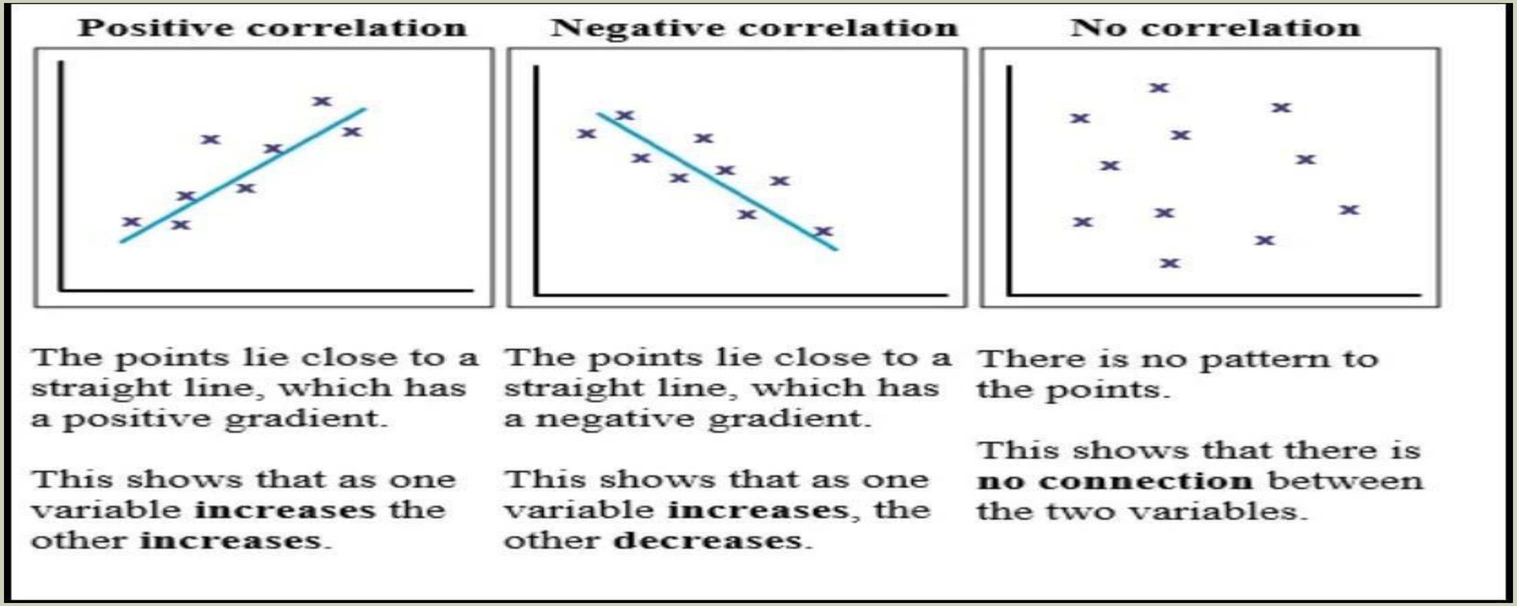
3. Sales director of a company knows that there is something wrong with one of its successful products, however hasn't yet carried out any market research data analysis. How and what does heconclude?

Data Analytics Applications

1. Understanding and targeting customers.
2. Understanding and optimizing busineess processes.
3. Improving Sports Performance.
4. Improving Science and Research.
5. Improving Security and law enforcement.
6. Improving and optimizing cities and countries.

Data Analytics Technique

Correlation: Correlation means association - more precisely it is a measure of the extent to which two variables are related. There are three possible results of a correlational study: a positive correlation, a negative correlation, and no correlation.



Positive correlation

The points lie close to a straight line, which has a positive gradient.

This shows that as one variable **increases** the other **increases**.

Negative correlation

The points lie close to a straight line, which has a negative gradient.

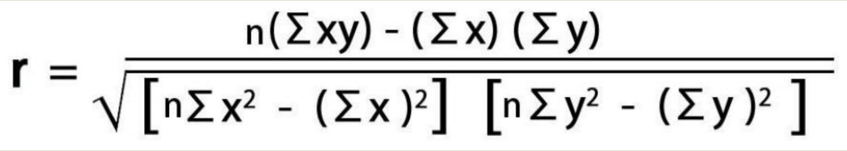
This shows that as one variable **increases**, the other **decreases**.

No correlation

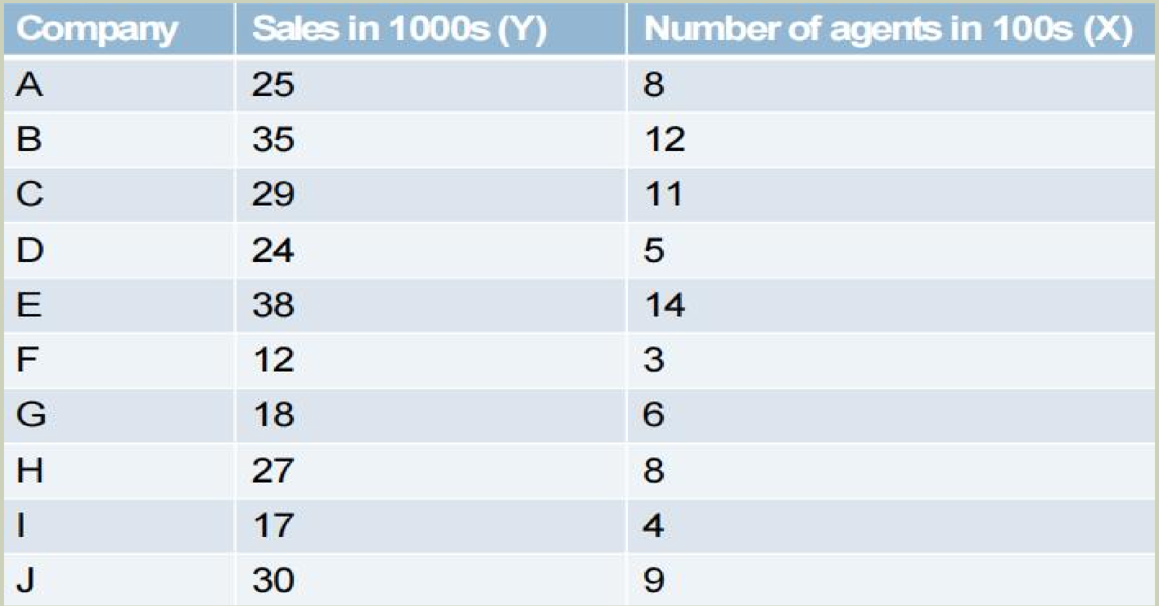
There is no pattern to the points.

This shows that there is **no connection** between the two variables.

Correlation Coefficients Calculation



$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$



Company	Sales in 1000s (Y)	Number of agents in 100s (X)
A	25	8
B	35	12
C	29	11
D	24	5
E	38	14
F	12	3
G	18	6
H	27	8
I	17	4
J	30	9

- n = 10, ∑X = 80, ∑Y = 255, ∑XY =2289
- ∑X2 = 756, ∑Y2 = 7097, (∑X)2 = 6400, (∑Y)2 = 65025, **r = 0.95**

Regression Modelling Technique

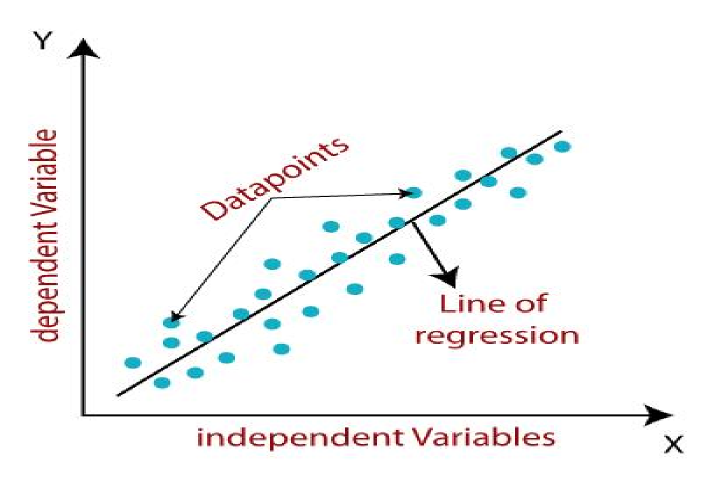
- One of the fundamental task in data analysis is to find how different variables are related to each other and one of the central tool for learning about such relationships is regression.
- Example: Predict annual sales.

  - competitive pricing,
  - product quality,
  - shipping time & cost,
  - online reviews,

In this case, sales is your dependent variable. Factors affecting sales are independent variables.

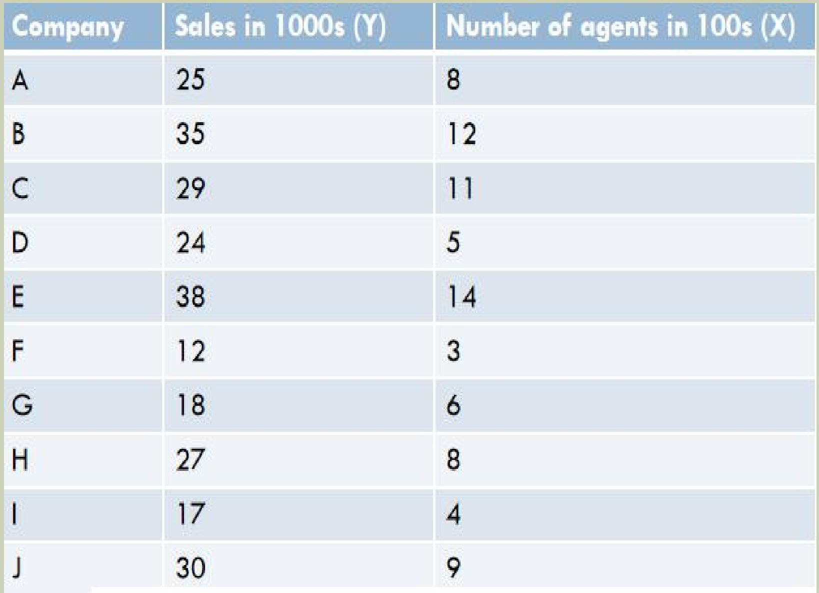
Linear Regression

- Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data.
- One variable is considered to be an explanatory (independent) variable, and the other is considered to be a dependent variable.
- For example, we want to relate the weights of individuals to their heights using a linear regression model.



Linear Regression

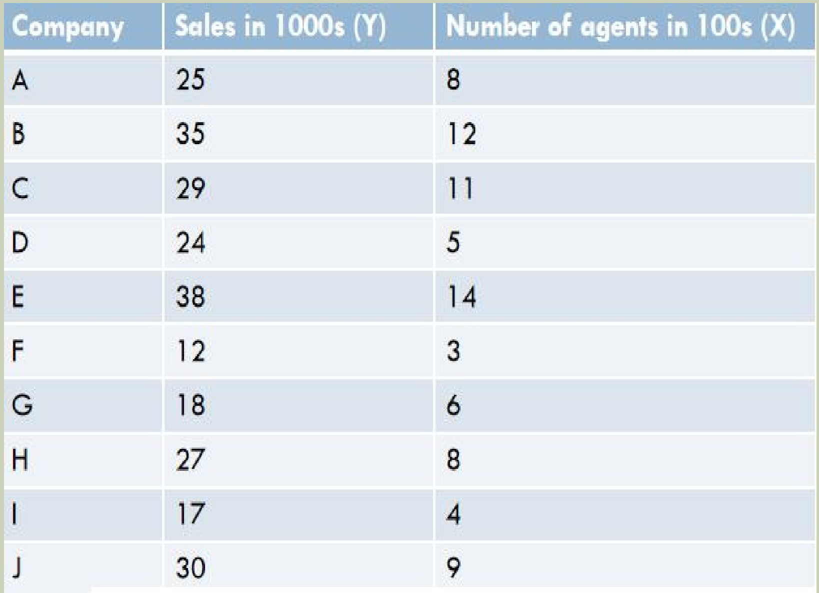
- **Y = a + bX + e**



Company	Sales in 1000s (Y)	Number of agents in 100s (X)
A	25	8
B	35	12
C	29	11
D	24	5
E	38	14
F	12	3
G	18	6
H	27	8
I	17	4
J	30	9

Linear Regression

- **Y = a + bX + e**



Company	Sales in 1000s (Y)	Number of agents in 100s (X)
A	25	8
B	35	12
C	29	11
D	24	5
E	38	14
F	12	3
G	18	6
H	27	8
I	17	4
J	30	9

$$b = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} \qquad a = \frac{\Sigma Y}{n} - b\,\frac{\Sigma X}{n}$$

❑ **Predicted (Y) = 8.3272 + 2.1466 X**

Multiple Linear Regression

- Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables.

Example:

- ❑ Do age and intelligence quotient (IQ) scores predict grade point average (GPA)?
- ❑ Do weight, height, and age explain the variance in cholesterol levels?
- ❑ Do height, weight, age, and hours of exercise per week predict blood pressure?

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots \ldots \ldots \ldots \ldots \ldots + \beta_n x_n + e$

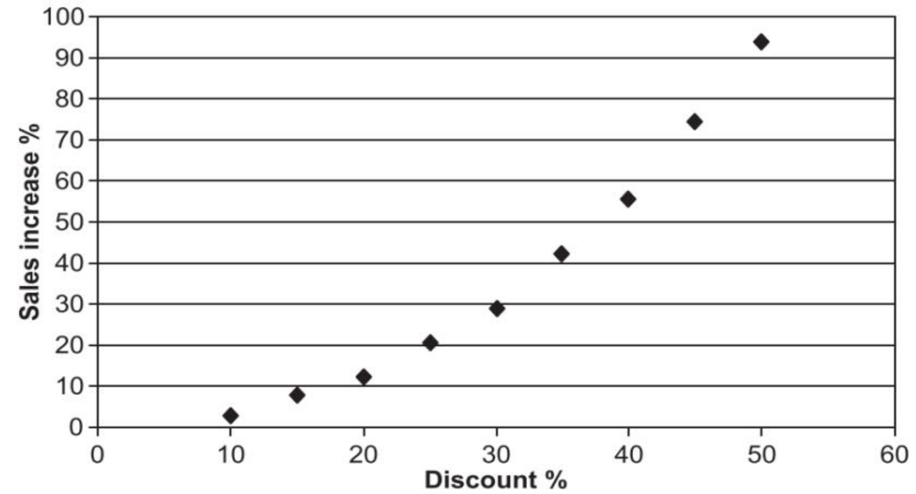
Non-Linear Regression

- There may be the case where the results from the correlation analysis show no linear relationship but these variables might still be closely related.
- If the result of the data analysis show that there is a non-linear (also known as curvilinear) association between the two variables, then the need is to develop a non-linear regression model.
- The non-linear data can be handled in 2 ways:

1. Use of polynomial rather than linear regression model
2. Transform the data and then use linear regression model.

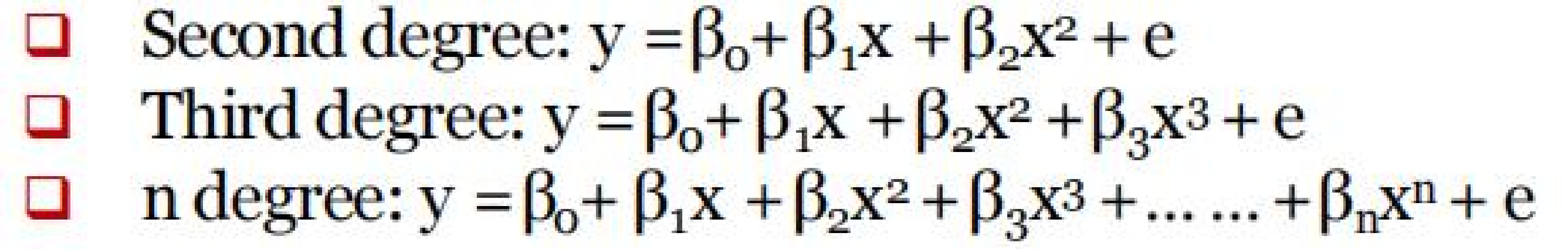
Non-Linear Regression

Product	Increase in sale in% (Y)	Discount in %(X)
A	3.05	10
B	7.62	15
C	12.19	20
D	20.42	25
E	28.65	30
F	42.06	35
G	55.47	40
H	74.68	45
I	93.88	50

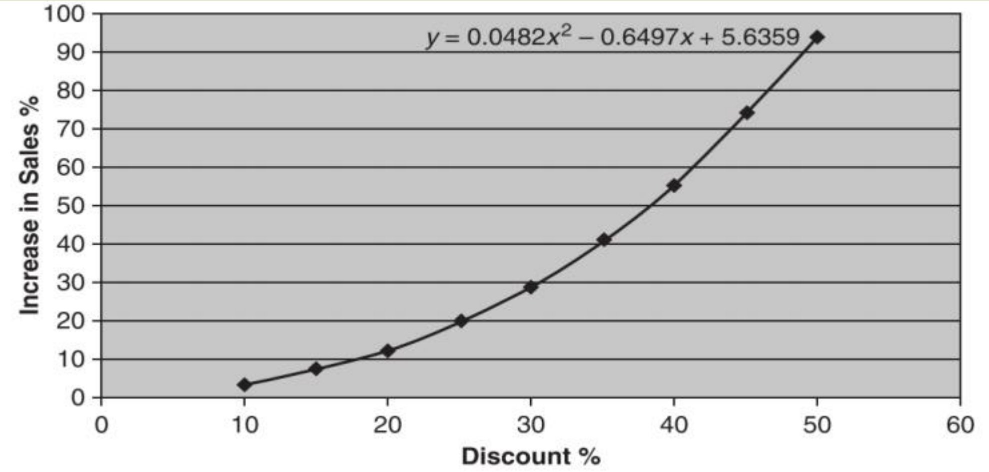


Non-Linear Regression

➢Polynomial Solution:



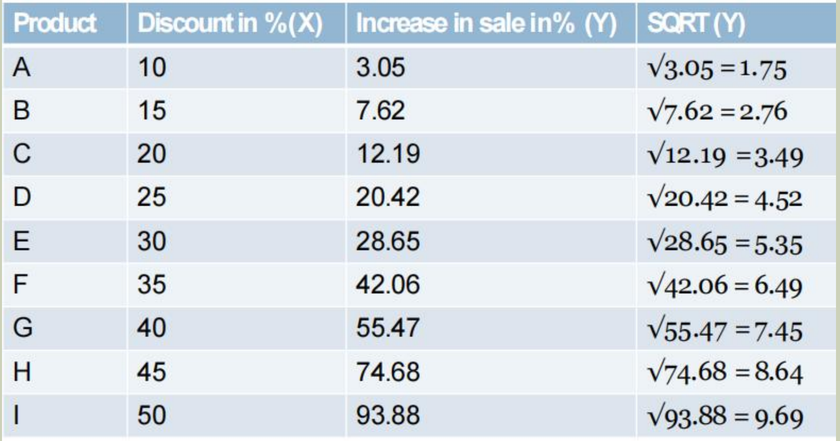
- ❑ Second degree: $y = \beta_o + \beta_1 x + \beta_2 x^2 + e$
- ❑ Third degree: $y = \beta_o + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + e$
- ❑ n degree: $y = \beta_o + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots\ldots + \beta_n x^n + e$



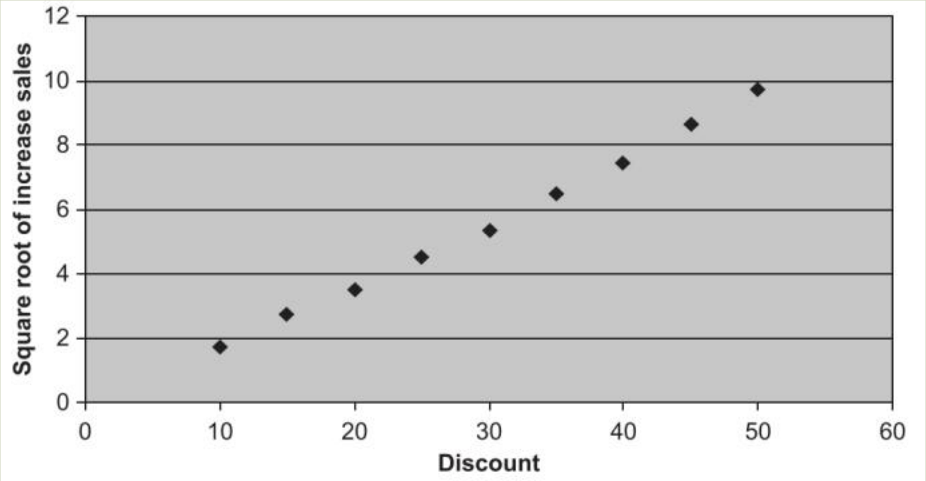
Non-Linear Regression

➢Transform solution

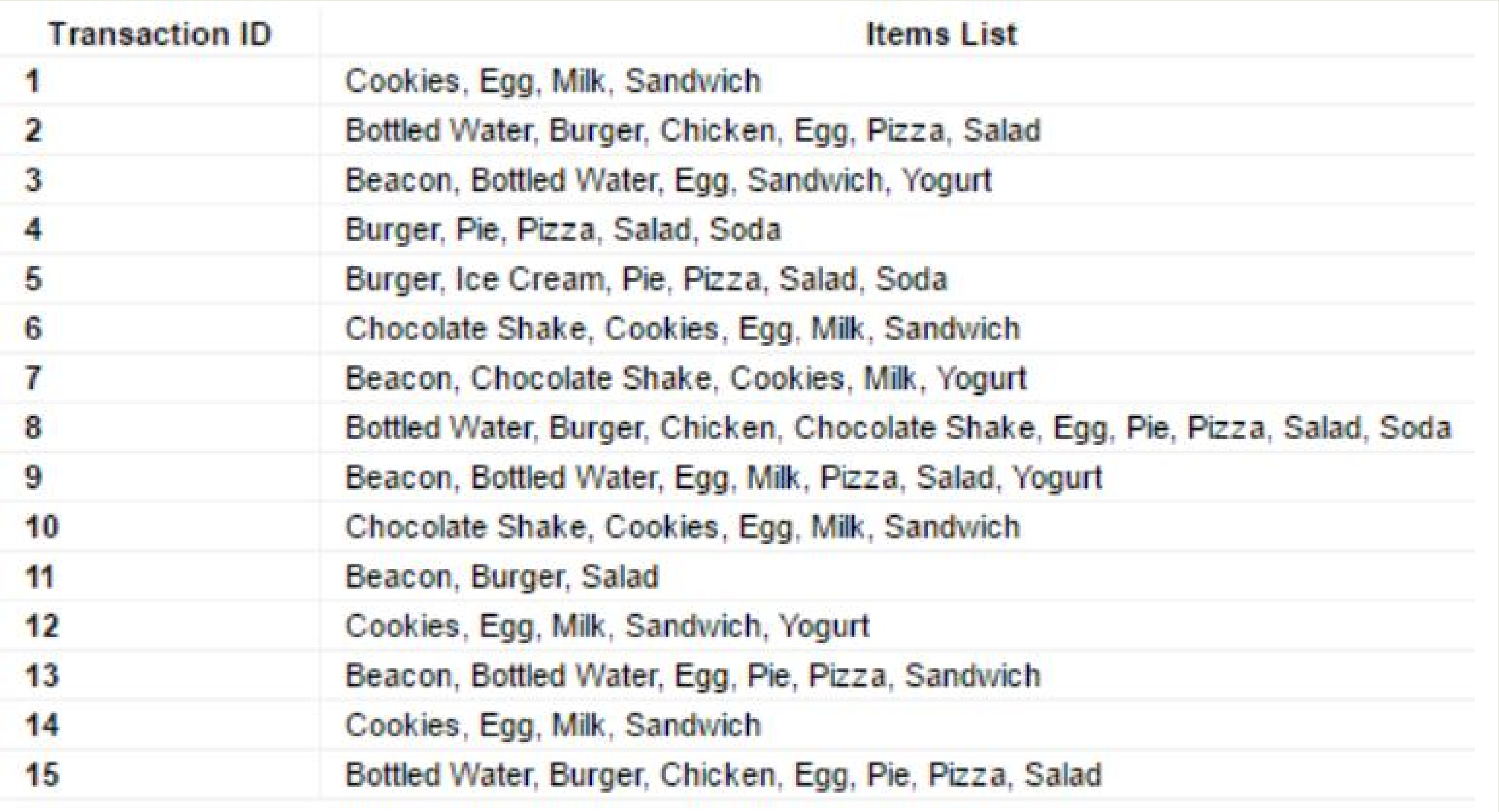
- ❑ Square root(√X)
- ❑ Logarithm (log X)
- ❑ Negative reciprocal (- 1/ X)



Product	Discount in %(X)	Increase in sale in% (Y)	SQRT (Y)
A	10	3.05	√3.05 =1.75
B	15	7.62	√7.62 =2.76
C	20	12.19	√12.19 =3.49
D	25	20.42	√20.42 = 4.52
E	30	28.65	√28.65 = 5.35
F	35	42.06	√42.06 = 6.49
G	40	55.47	√55.47 =7.45
H	45	74.68	√74.68 =8.64
I	50	93.88	√93.88 = 9.69



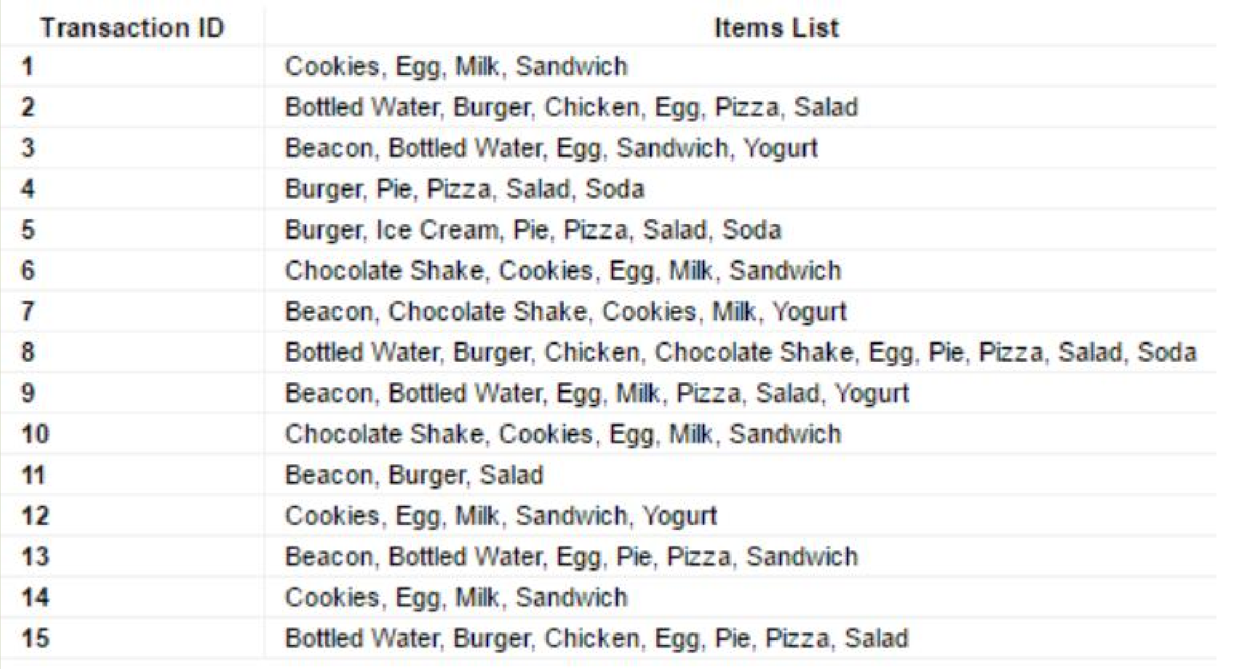
...



Transaction ID	Items List
1	Cookies, Egg, Milk, Sandwich
2	Bottled Water, Burger, Chicken, Egg, Pizza, Salad
3	Beacon, Bottled Water, Egg, Sandwich, Yogurt
4	Burger, Pie, Pizza, Salad, Soda
5	Burger, Ice Cream, Pie, Pizza, Salad, Soda
6	Chocolate Shake, Cookies, Egg, Milk, Sandwich
7	Beacon, Chocolate Shake, Cookies, Milk, Yogurt
8	Bottled Water, Burger, Chicken, Chocolate Shake, Egg, Pie, Pizza, Salad, Soda
9	Beacon, Bottled Water, Egg, Milk, Pizza, Salad, Yogurt
10	Chocolate Shake, Cookies, Egg, Milk, Sandwich
11	Beacon, Burger, Salad
12	Cookies, Egg, Milk, Sandwich, Yogurt
13	Beacon, Bottled Water, Egg, Pie, Pizza, Sandwich
14	Cookies, Egg, Milk, Sandwich
15	Bottled Water, Burger, Chicken, Egg, Pie, Pizza, Salad

Market-Basket Model

- It is a technique which identifies the strength of association between pairs of products purchased together and identify patterns of co-occurrence.



Transaction ID	Items List
1	Cookies, Egg, Milk, Sandwich
2	Bottled Water, Burger, Chicken, Egg, Pizza, Salad
3	Beacon, Bottled Water, Egg, Sandwich, Yogurt
4	Burger, Pie, Pizza, Salad, Soda
5	Burger, Ice Cream, Pie, Pizza, Salad, Soda
6	Chocolate Shake, Cookies, Egg, Milk, Sandwich
7	Beacon, Chocolate Shake, Cookies, Milk, Yogurt
8	Bottled Water, Burger, Chicken, Chocolate Shake, Egg, Pie, Pizza, Salad, Soda
9	Beacon, Bottled Water, Egg, Milk, Pizza, Salad, Yogurt
10	Chocolate Shake, Cookies, Egg, Milk, Sandwich
11	Beacon, Burger, Salad
12	Cookies, Egg, Milk, Sandwich, Yogurt
13	Beacon, Bottled Water, Egg, Pie, Pizza, Sandwich
14	Cookies, Egg, Milk, Sandwich
15	Bottled Water, Burger, Chicken, Egg, Pie, Pizza, Salad

Market-Basket Model

- Applications:

Mobile Service Providers: Mobile service providers use data mining to design their marketing campaigns and to retain customers from moving to other vendors. churn Customer

Retail Sector: Data Mining helps the supermarket and retail sector owners to know the choices of the customers.

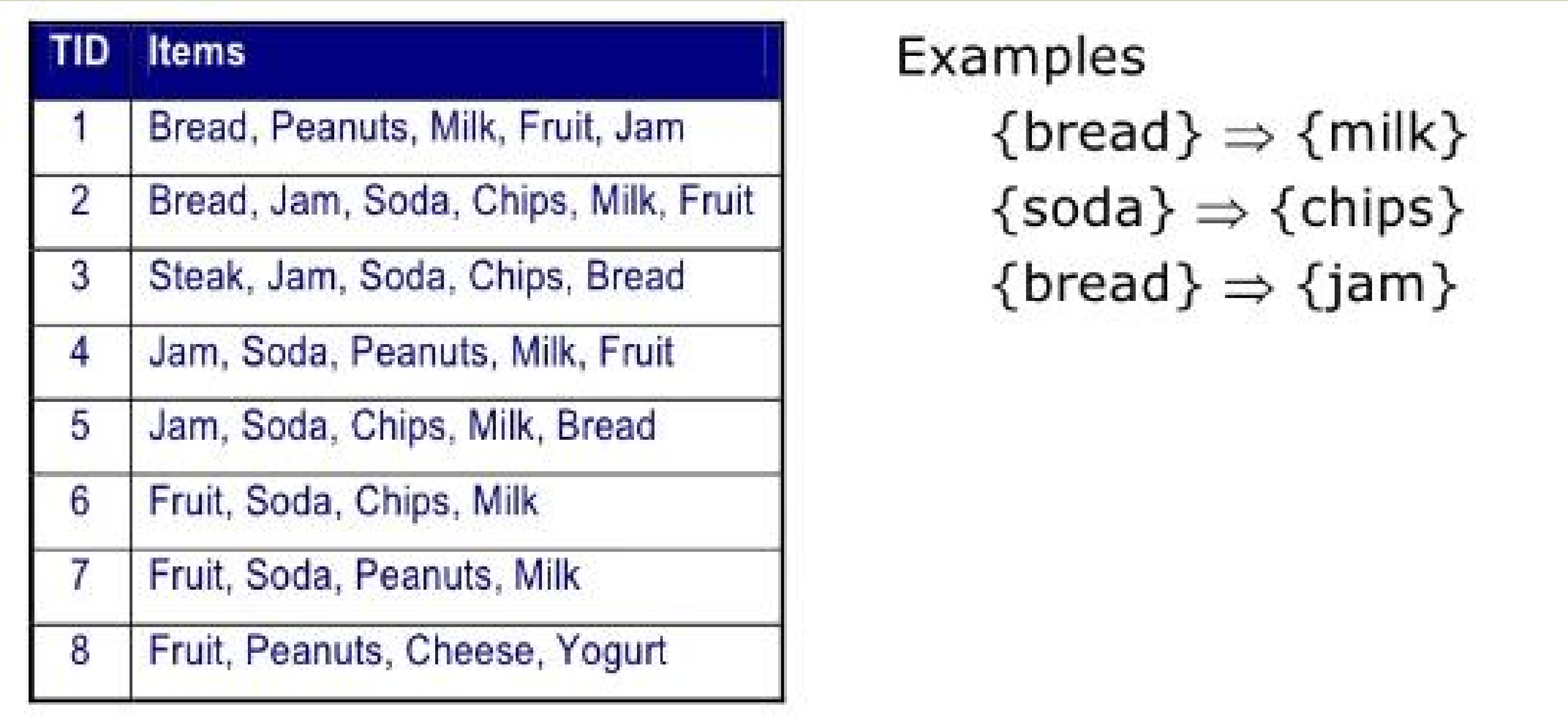
Ecommerce: Many E-commerce sites use data mining to offer cross-selling and upselling of their products. The shopping sites such as Amazon, Flipkart show “People also viewed”, “Frequently bought together” to the customers who are interacting with the site.

Market-Basket Model

- Association rule mining finds interesting associations and relationships among large sets of data items.
- It creates If-Then scenario rules, for example, if item A is purchased then item B is likely to be purchased.
- The rules could be written as:

{A} => {B}. The If part of the rule (the {A}) is known as the **antecedent** and the THEN part of the rule is known as the **consequent** (the {B}).

Example



TID	Items
1	Bread, Peanuts, Milk, Fruit, Jam
2	Bread, Jam, Soda, Chips, Milk, Fruit
3	Steak, Jam, Soda, Chips, Bread
4	Jam, Soda, Peanuts, Milk, Fruit
5	Jam, Soda, Chips, Milk, Bread
6	Fruit, Soda, Chips, Milk
7	Fruit, Soda, Peanuts, Milk
8	Fruit, Peanuts, Cheese, Yogurt

Examples

$$\{\text{bread}\} \Rightarrow \{\text{milk}\}$$
$$\{\text{soda}\} \Rightarrow \{\text{chips}\}$$
$$\{\text{bread}\} \Rightarrow \{\text{jam}\}$$

Conclusion

➢In this section you have learned

- What are the importance of data and Big Data Challenges are
- What exactly is Data Analytics and what do Data Scientists do with data
- Some data mining usages
- Case Study & Use Cases

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
Deemed to be University U/S 3 of the UGC Act, 1956