RAMCO INSTITUTE OF TECHNOLOGY, RAJAPALAYAM
DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
AD3301- Data Exploration and Visualization

UNIT-1 09.12.2024

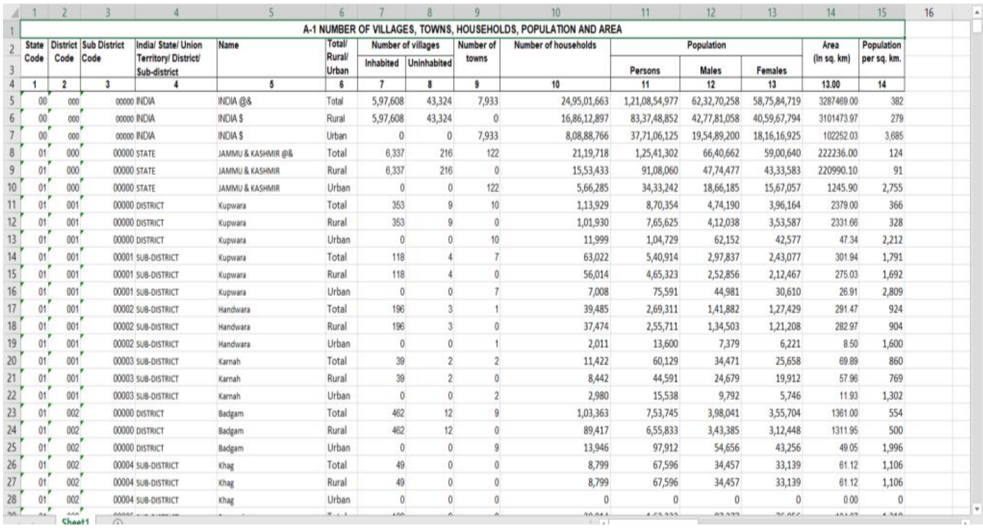
Prepared by

Dr.M.Klaippan, Professor and Head, Dept. of AI & DS

----

1. Data science (8 marks) OR (2 MARKS)

- Data science involves cross-disciplinary knowledge from computer science, data, statistics, and mathematics
- Data **(2 MARKS)**
    - It encompasses a collection of discrete objects, numbers, words, events, facts, measurements, observations, or even descriptions of things.
    - Such data is collected and stored by event occurring in several disciplines, including biology, economics, engineering, marketing, and others. Ex: Senses table of India
    - Processing such data elicits useful information and processing such information generates useful knowledge.

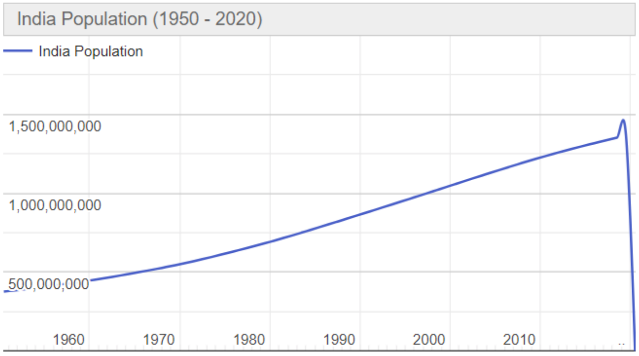


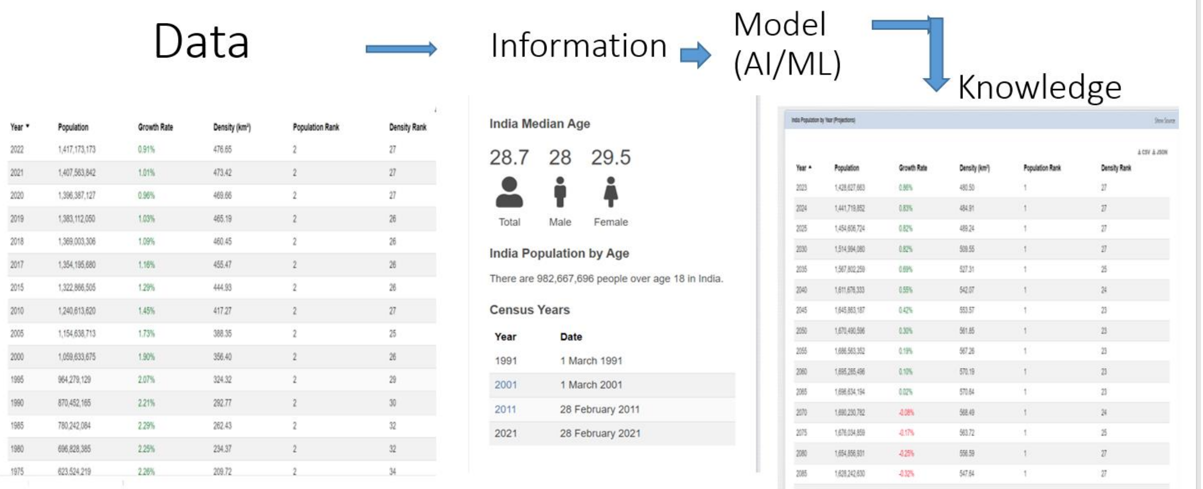
A-1 NUMBER OF VILLAGES, TOWNS, HOUSEHOLDS, POPULATION AND AREA

State Code	District Code	Sub District Code	India/ State/ Union Territory/ District/ Sub-district	Name	Total/ Rural/ Urban	Number of villages		Number of towns	Number of households	Population			Area (In sq. km)	Population per sq. km.
						Inhabited	Uninhabited			Persons	Males	Females		
1	2	3	4	5	6	7	8	9	10	11	12	13	13.00	14
00	000	00000	INDIA	INDIA @&	Total	5,97,608	43,324	7,933	24,95,01,663	1,21,08,54,977	62,32,70,258	58,75,84,719	3287469.00	382
00	000	00000	INDIA	INDIA $	Rural	5,97,608	43,324	0	16,86,12,897	83,37,48,852	42,77,81,058	40,59,67,794	3101473.97	279
00	000	00000	INDIA	INDIA $	Urban	0	0	7,933	8,08,88,766	37,71,06,125	19,54,89,200	18,16,16,925	102252.03	3,685
01	000	00000	STATE	JAMMU & KASHMIR @&	Total	6,337	216	122	21,19,718	1,25,41,302	66,40,662	59,00,640	222236.00	124
01	000	00000	STATE	JAMMU & KASHMIR	Rural	6,337	216	0	15,53,433	91,08,060	47,74,477	43,33,583	220990.10	91
01	000	00000	STATE	JAMMU & KASHMIR	Urban	0	0	122	5,66,285	34,33,242	18,66,185	15,67,057	1245.90	2,755
01	001	00000	DISTRICT	Kupwara	Total	353	9	10	1,13,929	8,70,354	4,74,190	3,96,164	2379.00	366
01	001	00000	DISTRICT	Kupwara	Rural	353	9	0	1,01,930	7,65,625	4,12,038	3,53,587	2331.66	328
01	001	00000	DISTRICT	Kupwara	Urban	0	0	10	11,999	1,04,729	62,152	42,577	47.34	2,212
01	001	00001	SUB-DISTRICT	Kupwara	Total	118	4	7	63,022	5,40,914	2,97,837	2,43,077	301.94	1,791
01	001	00001	SUB-DISTRICT	Kupwara	Rural	118	4	0	56,014	4,65,323	2,52,856	2,12,467	275.03	1,692
01	001	00001	SUB-DISTRICT	Kupwara	Urban	0	0	7	7,008	75,591	44,981	30,610	26.91	2,809
01	001	00002	SUB-DISTRICT	Handwara	Total	196	3	1	39,485	2,69,311	1,41,882	1,27,429	291.47	924
01	001	00002	SUB-DISTRICT	Handwara	Rural	196	3	0	37,474	2,55,711	1,34,503	1,21,208	282.97	904
01	001	00002	SUB-DISTRICT	Handwara	Urban	0	0	1	2,011	13,600	7,379	6,221	8.50	1,600
01	001	00003	SUB-DISTRICT	Karnah	Total	39	2	2	11,422	60,129	34,471	25,658	69.89	860
01	001	00003	SUB-DISTRICT	Karnah	Rural	39	2	0	8,442	44,591	24,679	19,912	57.96	769
01	001	00003	SUB-DISTRICT	Karnah	Urban	0	0	2	2,980	15,538	9,792	5,746	11.93	1,302
01	002	00000	DISTRICT	Badgam	Total	462	12	9	1,03,363	7,53,745	3,98,041	3,55,704	1361.00	554
01	002	00000	DISTRICT	Badgam	Rural	462	12	0	89,417	6,55,833	3,43,385	3,12,448	1311.95	500
01	002	00000	DISTRICT	Badgam	Urban	0	0	9	13,946	97,912	54,656	43,256	49.05	1,996
01	002	00004	SUB-DISTRICT	Khag	Total	49	0	0	8,799	67,596	34,457	33,139	61.12	1,106
01	002	00004	SUB-DISTRICT	Khag	Rural	49	0	0	8,799	67,596	34,457	33,139	61.12	1,106
01	002	00004	SUB-DISTRICT	Khag	Urban	0	0	0	0	0	0	0	0.00	0

Information

1,409,738,760





Year ▾	Population	Growth Rate	Density (km²)	Population Rank	Density Rank
2022	1,417,173,173	0.91%	476.65	2	27
2021	1,407,563,842	1.01%	473.42	2	27
2020	1,396,387,127	0.96%	469.66	2	27
2019	1,383,112,050	1.03%	465.19	2	26
2018	1,369,003,306	1.09%	460.45	2	26
2017	1,354,195,680	1.16%	455.47	2	26
2015	1,322,866,505	1.29%	444.93	2	26
2010	1,240,613,620	1.45%	417.27	2	27
2005	1,154,638,713	1.73%	388.35	2	25
2000	1,059,633,675	1.90%	356.40	2	26
1995	964,279,129	2.07%	324.32	2	29
1990	870,452,165	2.21%	292.77	2	30
1985	780,242,084	2.29%	262.43	2	32
1980	696,828,385	2.25%	234.37	2	32
1975	623,524,219	2.26%	209.72	2	34

India Population by Year (Historical)

India Median Age

28.7	28	29.5
Total	Male	Female

India Population by Age

There are 982,667,696 people over age 18 in India.

Census Years

Year	Date
1991	1 March 1991
2001	1 March 2001
2011	28 February 2011
2021	28 February 2021

India Population by Year (Projections) Show Source

CSV JSON

Year ▴	Population	Growth Rate	Density (km²)	Population Rank	Density Rank
2023	1,428,627,663	0.86%	480.50	1	27
2024	1,441,719,852	0.83%	484.91	1	27
2025	1,454,606,724	0.82%	489.24	1	27
2030	1,514,994,080	0.82%	509.55	1	27
2035	1,567,802,259	0.69%	527.31	1	25
2040	1,611,676,333	0.55%	542.07	1	24
2045	1,645,863,187	0.42%	553.57	1	23
2050	1,670,490,596	0.30%	561.85	1	23
2055	1,686,563,352	0.19%	567.26	1	23
2060	1,695,285,496	0.10%	570.19	1	23
2065	1,696,634,194	0.02%	570.64	1	23
2070	1,690,230,782	-0.08%	568.49	1	24
2075	1,676,034,859	-0.17%	563.72	1	25
2080	1,654,856,931	-0.25%	556.59	1	27
2085	1,628,242,630	-0.32%	547.64	1	27

India Population by Year (Projections)

2. EDA (16 marks)

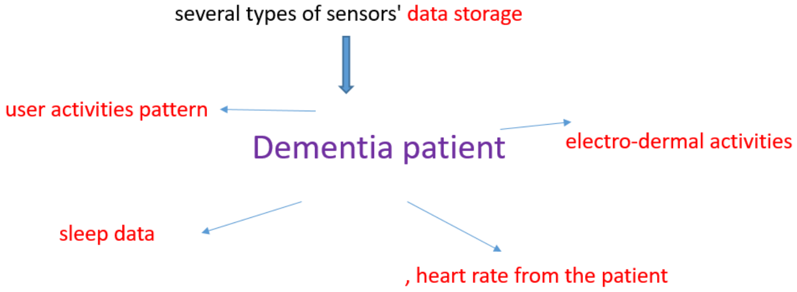
- EDA is a process of examining the available dataset to discover patterns, spot anomalies, test hypotheses, and check assumptions using statistical measures

2.1 Several phases of data analysis

- data requirements, data collection, data processing, data cleaning, exploratory data analysis, modeling and algorithms, and data product and communication. These phases are similar to the CRoss-Industry Standard Process for data mining (CRISP) framework in data mining.
- **Data requirements: (2 MARKS)**
  - There can be various sources of data for an organization. It is important to comprehend what type of data is required for the organization to be collected, curated, and stored.
    - For example, an application tracking the sleeping pattern of patients suffering from dementia requires several types of sensors' data storage, such as sleep data, heart rate from the patient, electro-dermal activities, and user activities pattern. All of these data points are required to correctly diagnose the mental state of the person. Hence, these are mandatory requirements for the application.
    - In addition to this, it is required to categorize the data, numerical or categorical, and the format of storage and dissemination.
- **Data collection: (2 MARKS)**
  - Data collected from several sources must be stored in the correct format and transferred to the right information technology personnel within a company. As mentioned previously, data can be collected from several objects on several events using different types of sensors and storage tools.
- **Data processing: (2 MARKS)**
  - Preprocessing involves the process of pre-curating the dataset before actual analysis. Common tasks involve correctly exporting the dataset, placing them under the right tables, structuring them, and exporting them in the correct format.

Data requirements

Application tracking the sleeping pattern of patients suffering from dementia



several types of sensors' data storage

Dementia patient

user activities pattern

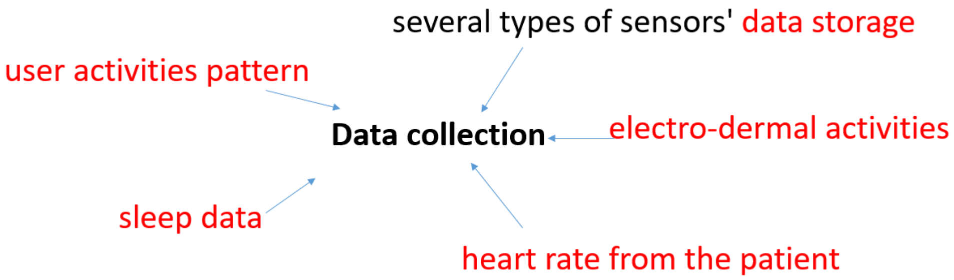
electro-dermal activities

sleep data

, heart rate from the patient

Data collection

Application tracking the sleeping pattern of patients suffering from dementia



several types of sensors' data storage

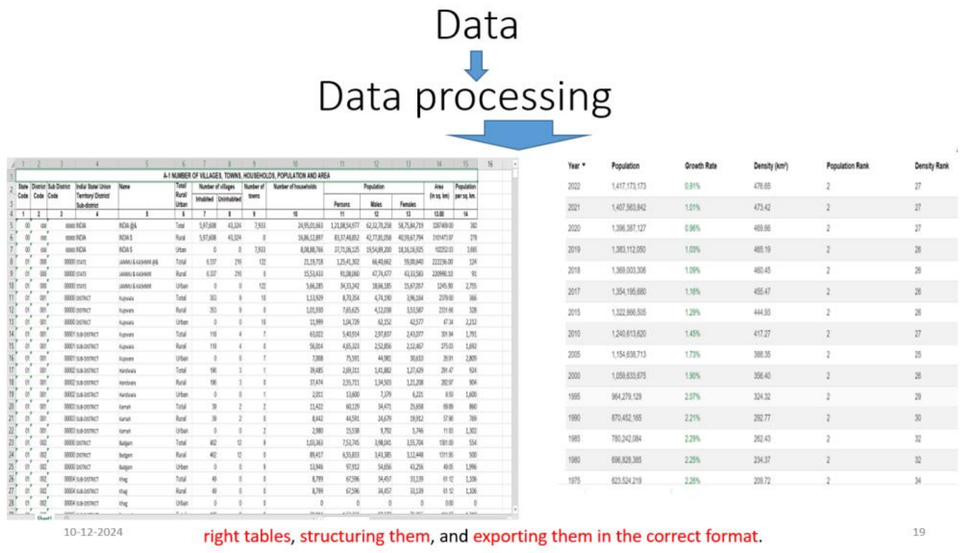
user activities pattern

Data collection

electro-dermal activities

sleep data

heart rate from the patient



Data

Data processing

A-1 NUMBER OF VILLAGES, TOWNS, HOUSEHOLDS, POPULATION AND AREA

State Code	District Code	Sub District Code	India/ State/ Union Territory/ District/ Sub-district	Name	Total/ Rural/ Urban	Number of villages		Number of towns	Number of households	Population			Area (in sq. km)	Population per sq. km.
						Inhabited	Uninhabited			Persons	Males	Females		
1	2	3	4	5	6	7	8	9	10	11	12	13	13.00	14
00	000	00000	INDIA	INDIA @&	Total	5,97,608	43,324	7,933	24,95,01,663	1,21,08,54,977	62,32,70,258	58,75,84,719	3287469.00	382
00	000	00000	INDIA	INDIA $	Rural	5,97,608	43,324	0	16,86,12,897	83,37,48,852	42,77,81,058	40,59,67,794	3101473.97	279
00	000	00000	INDIA	INDIA $	Urban	0	0	7,933	8,08,88,766	37,71,06,125	19,54,89,200	18,16,16,925	102252.03	3685
01	000	00000	STATE	JAMMU & KASHMIR @&	Total	6,337	216	122	21,19,718	1,25,41,302	66,40,662	59,00,640	222236.00	124
01	000	00000	STATE	JAMMU & KASHMIR	Rural	6,337	216	0	15,53,433	91,08,060	47,74,477	43,33,583	220990.10	91
01	000	00000	STATE	JAMMU & KASHMIR	Urban	0	0	122	5,66,285	34,33,242	18,66,185	15,67,057	1245.90	2,755
01	001	00000	DISTRICT	Kupwara	Total	361	9	10	1,13,929	8,70,354	4,74,190	3,96,164	2379.00	366
01	001	00000	DISTRICT	Kupwara	Rural	361	9	0	1,01,930	7,65,625	4,12,038	3,53,587	2331.66	328
01	001	00000	DISTRICT	Kupwara	Urban	0	0	10	11,999	1,04,729	62,152	42,577	47.34	2,212
01	001	00001	SUB-DISTRICT	Kupwara	Total	118	4	7	63,022	5,40,954	2,97,837	2,43,077	301.94	1,791
01	001	00001	SUB-DISTRICT	Kupwara	Rural	118	4	0	56,014	4,65,323	2,52,856	2,12,467	275.03	1,692
01	001	00001	SUB-DISTRICT	Kupwara	Urban	0	0	7	7,008	75,591	44,981	30,610	26.91	2,809
01	001	00002	SUB-DISTRICT	Handwara	Total	196	3	1	39,485	2,69,311	1,41,882	1,27,429	291.47	924
01	001	00002	SUB-DISTRICT	Handwara	Rural	196	3	0	37,474	2,55,711	1,34,503	1,21,208	282.97	904
01	001	00002	SUB-DISTRICT	Handwara	Urban	0	0	1	2,011	13,600	7,379	6,221	8.50	1,600
01	001	00003	SUB-DISTRICT	Karnah	Total	39	2	2	11,422	60,129	34,471	25,658	69.89	860
01	001	00003	SUB-DISTRICT	Karnah	Rural	39	2	0	8,442	44,591	24,679	19,912	57.96	769
01	001	00003	SUB-DISTRICT	Karnah	Urban	0	0	2	2,980	15,538	9,792	5,746	11.93	1,302
01	002	00000	DISTRICT	Badgam	Total	482	12	9	1,03,363	7,53,745	3,98,041	3,55,704	1361.00	554
01	002	00000	DISTRICT	Badgam	Rural	482	12	0	89,417	6,55,833	3,43,385	3,12,448	1311.95	500
01	002	00000	DISTRICT	Badgam	Urban	0	0	9	13,946	97,912	54,656	43,256	49.05	1,996
01	002	00004	SUB-DISTRICT	Khag	Total	49	0	0	8,799	67,596	34,457	33,139	61.12	1,106
01	002	00004	SUB-DISTRICT	Khag	Rural	49	0	0	8,799	67,596	34,457	33,139	61.12	1,106
01	002	00004	SUB-DISTRICT	Khag	Urban	0	0	0	0	0	0	0	0.00	0

Year	Population	Growth Rate	Density (km²)	Population Rank	Density Rank
2022	1,417,173,173	0.91%	476.65	2	27
2021	1,407,563,842	1.01%	473.42	2	27
2020	1,396,387,127	0.96%	469.66	2	27
2019	1,383,112,050	1.03%	465.19	2	26
2018	1,369,003,306	1.09%	460.45	2	26
2017	1,354,195,680	1.16%	455.47	2	26
2015	1,322,866,505	1.29%	444.93	2	26
2010	1,240,613,620	1.45%	417.27	2	27
2005	1,154,638,713	1.73%	388.35	2	25
2000	1,059,633,675	1.90%	356.40	2	26
1995	964,279,129	2.07%	324.32	2	29
1990	870,452,165	2.21%	292.77	2	30
1985	780,242,084	2.29%	262.43	2	32
1980	696,828,385	2.25%	234.37	2	32
1975	623,524,219	2.26%	209.72	2	34

10-12-2024

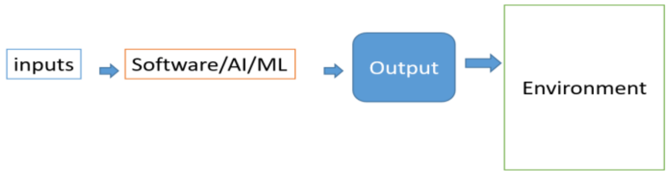
right tables, structuring them, and exporting them in the correct format.

19

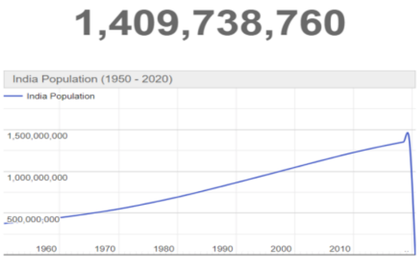
- **Data cleaning: (2 MARKS)**
    - Preprocessed data is still not ready for detailed analysis. It must be correctly transformed for an incompleteness check, duplicates check, error check, and missing value check. These tasks are performed in the data cleaning stage, which involves responsibilities such as matching the correct record, finding in accuracies in the dataset, understanding the overall data quality, removing duplicate items, and filling in the missing values. An example of data cleaning technique would be using outlier detection methods for quantitative data cleaning.

- **Modeling and algorithm:**
  - From a data science perspective, generalized models or mathematical formulas can represent relationships among different variables such as correlation.
  - These models or equations involve one or more variables that depend on other variables to cause an event.
  - For example, when buying, say, pens,
    - the total price of pens(Total) = price for one pen(UnitPrice) * the number of pens bought (Quantity).
    - Hence, our model would be Total = UnitPrice * Quantity.
    - Here, the total price is dependent on the unit price. and the unit price is referred to as an independent variable.
    - In general, a model always describes the relationship between independent and dependent variables.
  - **Inferential statistics** deals with quantifying relationships between particular variables.
    - The Judd model for describing the relationship between data, model, and error still holds true: Data = Model + Error.
- **Data Product: (2 MARKS)**
  - Any computer software that uses data as inputs, produces outputs, and provides feedback based on the output to control the environment is referred to as a data product.
  - A data product is generally based on a model developed during data analysis, for example, a recommendation model that inputs user purchase history and recommends a related item that the user is highly likely to buy.
- **Communication: (2 MARKS)**
  - This stage deals with disseminating the results to end stakeholders to use the result for business intelligence. One of the most notable steps in this stage is **data visualization**.
  - Visualization deals with information relay techniques such as tables, charts, summary diagrams, and bar charts to show the analyzed result.

Data Product



communication



data visualization

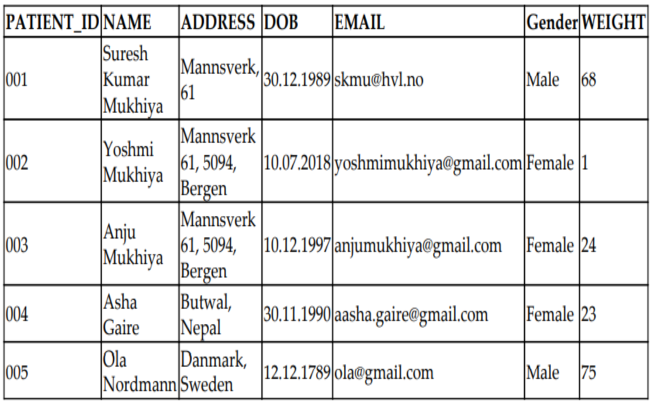
3. Steps in EDA (16 marks)

- **Problem definition:**
  - Before trying to extract useful insight from the data, it is essential to define the business problem to be solved. The main tasks involved in problem definition are defining the main objective of the analysis, defining the main deliverables, outlining the main roles and responsibilities, obtaining the current status of the data, defining the timetable, and performing cost/benefit analysis.
- **Data preparation: (2 MARKS)**
  - This step involves methods for preparing the dataset before actual analysis. In this step define the sources of data, define data schemas and tables, understand the characteristics of the data, clean the dataset, delete non-relevant datasets, transform the data, and divide the data into required chunks for analysis.
- **Data analysis: (2 MARKS)**
  - This is one of the most crucial steps that deals with descriptive statistics and analysis of the data. The main tasks involve summarizing the data, finding the hidden correlation and relationships among the data, developing predictive models, evaluating the models, and calculating the accuracies.
  - **Some of the techniques used for data summarization are**
    - summary tables, graphs, descriptive statistics, inferential statistics, correlation statistics, searching, grouping, and mathematical models.
- **Development and representation of the results:**
  - This step involves presenting the dataset to the target audience in the form of graphs, summary tables, maps, and diagrams. The result analyzed from the dataset should be interpretable by the business stakeholders.
  - **Most of the graphical analysis techniques include**
    - scattering plots, character plots, histograms, box plots, residual plots, mean plots, and others

4. Making sense of data (16 marks)

- It deals different types of data during analysis.
- Different disciplines store different kinds of data for different purposes.
  - For example, medical researchers store patients' data, universities store students' and teachers' data, and real estate industries storehouse and building datasets.
- A dataset contains many observations about a particular object.
  - For instance, a dataset about patients in a hospital can contain many observations. A patient can be described by a patient identifier (ID), name, address, weight, date of birth, address, email, and gender. Each of these features that describes a patient is a variable. Each observation can have a specific value for each of these variables.
  - PATIENT_ID = 1001
  - Name = Yoshmi Mukhiya

- Address = Mannsverk 61, 5094, Bergen, Norway
- Date of birth = 10th July 2018
- Email = yoshmimukhiya@gmail.com
- Weight = 10
- Gender = Female



PATIENT_ID	NAME	ADDRESS	DOB	EMAIL	Gender	WEIGHT
001	Suresh Kumar Mukhiya	Mannsverk, 61	30.12.1989	skmu@hvl.no	Male	68
002	Yoshmi Mukhiya	Mannsverk 61, 5094, Bergen	10.07.2018	yoshmimukhiya@gmail.com	Female	1
003	Anju Mukhiya	Mannsverk 61, 5094, Bergen	10.12.1997	anjumukhiya@gmail.com	Female	24
004	Asha Gaire	Butwal, Nepal	30.11.1990	aasha.gaire@gmail.com	Female	23
005	Ola Nordmann	Danmark, Sweden	12.12.1789	ola@gmail.com	Male	75

observations (001, 002, 003, 004, 005)

variables (PatientID, name, address, dob, email, gender, and weight).

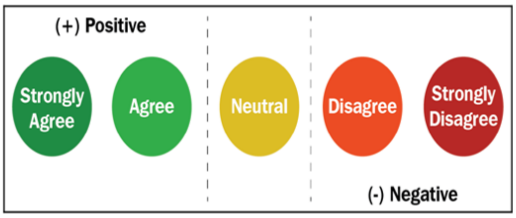
- Numerical data or quantitative data **(2 MARKS)**
  - This data has a sense of measurement involved in it; for example, a person's age, height, weight, blood pressure, heart rate, temperature, number of teeth, number of bones, and the number of family members. This data is often referred to as quantitative data in statistics.
  - The numerical dataset types
    - discrete or continuous types.
- Discrete data **(2 MARKS)**
  - This is data that is countable and its values can be listed out. For example, if we flip a coin, the number of heads in 200 coin flips can take values from 0 to 200 (finite) cases.
  - **discrete variable**
    - A variable that represents a discrete dataset is referred to as a discrete variable. The discrete variable takes a fixed number of distinct values.
    - For example, the Country variable can have values such as Nepal, India, Norway, and Japan. It is fixed. The Rank variable of a student in a classroom can take values from 1, 2, 3, 4, 5, and so on.
- Continuous data **(2 MARKS)**
  - A variable that can have an infinite number of numerical values within a specific range is classified as continuous data.
  - **continuous variable (2 MARKS)**

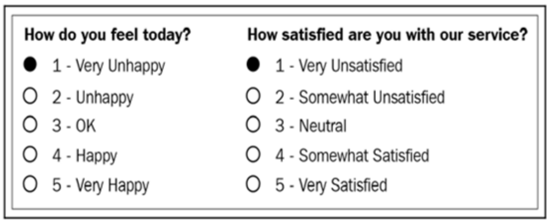
- A variable describing continuous data is a continuous variable. For example, what is the temperature of your city today? Can we be finite?

- **Categorical data or qualitative datasets (2 MARKS)**
  - This type of data represents the **characteristics of an object**; for example, gender, marital status, type of address, or categories of the movies. This data is often referred to as qualitative datasets in statistics.
  - **common types of categorical data :**
    - Gender (Male, Female, Other, or Unknown)
    - Marital Status (Annulled, Divorced, Interlocutory, Legally Separated, Married, Polygamous, Never Married, Domestic Partner, Unmarried, Widowed, or Unknown)
    - Movie genres (Action, Adventure, Comedy, Crime, Drama, Fantasy, Historical, Horror, Mystery, Philosophical, Political, Romance, Saga, Satire, Science Fiction, Social, Thriller, Urban, or Western)
    - Blood type (A, B, AB, or O)
    - Types of drugs (Stimulants, Depressants, Hallucinogens, Dissociative, Opioids, Inhalants, or Cannabis)
  - **categorical variable (2 MARKS)**
    - A variable describing categorical data is referred to as a categorical variable.
- Types of categorical variables: **(2 MARKS)**
  - binary categorical variable **(2 MARKS)**
    - A binary categorical variable can take **exactly two values** and is also referred to as a **dichotomous variable**. For example, when you create an experiment, the result is either success or failure. Hence, results can be understood as a binary categorical variable.
  - Polytomous variables **(2 MARKS)**
    - It can take more than two possible values.
    - For example, marital status can have several values, such as annulled, divorced, interlocutory, legally separated, married, polygamous, never married, domestic partners, unmarried, widowed, domestic partner, and unknown.

- **Measurement scales**
  - There are four different types of measurement scales described in statistics: nominal, ordinal, interval, and ratio.
  - **Nominal (2 MARKS)**
    - These are practiced for labeling variables without any quantitative value. The scales are generally referred to as labels. It do not carry any numerical importance. examples:
    - What is your gender?

      - Male
      - Female
      - Third gender/Non-binary
      - I prefer not to answer
      - The languages that are spoken in a particular country
      - Biological species
      - Parts of speech in grammar (noun, pronoun, adjective, and so on)
      - Taxonomic ranks in biology (Archea, Bacteria, and Eukarya)
  - **Nominal scales or qualitative data**
    - It is considered qualitative scales and the measurements that are taken using qualitative scales called qualitative data in the case of a nominal dataset:
    - Frequency is the rate at which a label occurs over a period of time within the dataset.
    - Proportion can be calculated by dividing the frequency by the total number of events.
    - visualize the nominal dataset using either a pie chart or a bar chart.
- **Ordinal (2 MARKS)**
  - In ordinal, the order of the values is a significant factor The main difference in the ordinal and nominal scale is **the order**.
  - **Ordinal scale (2 MARKS)**
    - Ordinal scales as an order of ranking (1st, 2nd, 3rd, 4th, and so on). The median item is allowed as the measure of central tendency, the average is not permitted
  - **Likert scale (2 MARKS)**
    - Ordinal Scales are referred to as the Likert scale
  - Example: Likert scale uses a variation of an ordinal scale?
    - WordPress is making content managers' lives easier. How do you feel about this statement?
    - The answer to the question is scaled down to five different ordinal values, Strongly Agree, Agree, Neutral, Disagree, and Strongly Disagree.



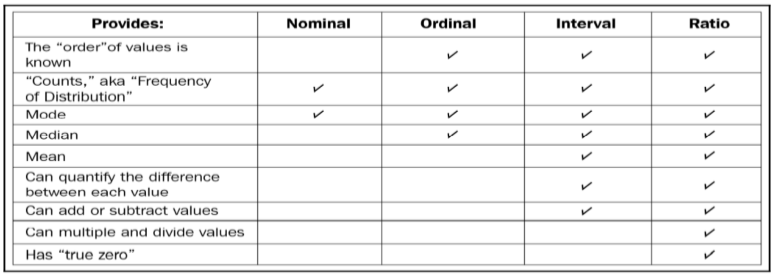


How do you feel today?	How satisfied are you with our service?
● 1 - Very Unhappy	● 1 - Very Unsatisfied
○ 2 - Unhappy	○ 2 - Somewhat Unsatisfied
○ 3 - OK	○ 3 - Neutral
○ 4 - Happy	○ 4 - Somewhat Satisfied
○ 5 - Very Happy	○ 5 - Very Satisfied

- **Interval**
  - In interval scales, both the order and exact differences between the values are significant.

- Interval scales are widely used in statistics. The measure of central tendencies, mean, median, standard deviations and mode are allowed on interval.
- Examples: location in Cartesian coordinates and direction measured in degrees from magnetic north..

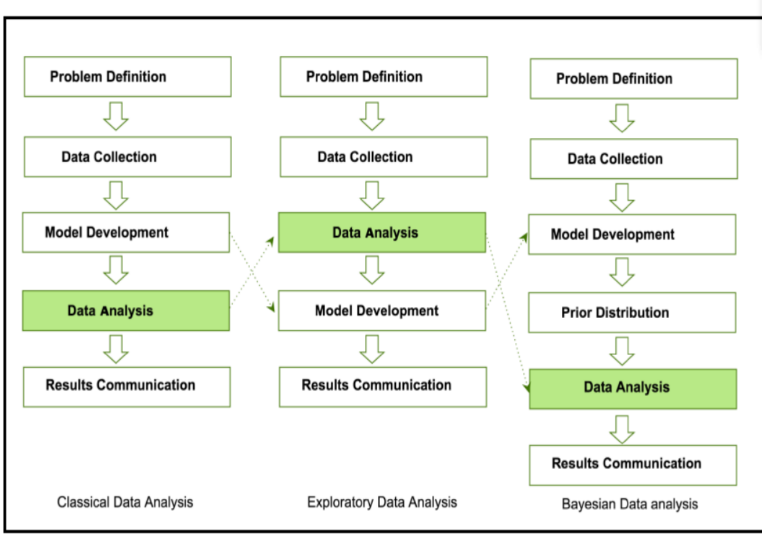
- **Ratio**
  - Ratio scales contain order, exact values, and absolute zero, which to be used in descriptive and inferential statistics.
  - These scales provide numerous possibilities for statistical analysis.
  - Mathematical operations, the measure of central tendencies, and the measure of dispersion and coefficient of variation can also be computed from these scales.



Provides:	Nominal	Ordinal	Interval	Ratio
The "order"of values is known		✓	✓	✓
"Counts," aka "Frequency of Distribution"	✓	✓	✓	✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓
Can quantify the difference between each value			✓	✓
Can add or subtract values			✓	✓
Can multiple and divide values				✓
Has "true zero"				✓

4. Comparing EDA with classical and Bayesian analysis

- **Classical data analysis:**
  - The problem definition and data collection step are followed by model development, which is followed by analysis and result communication.
- **Exploratory data analysis approach:**
  - It follows the same approach as classical data analysis except the model imposition and the data analysis steps are swapped. The main focus is on the data, its structure, outliers, models, and visualizations. Generally, in EDA, we do not impose any deterministic or probabilistic models on the data.
- **Bayesian data analysis approach:**
  - The Bayesian approach incorporates prior probability distribution knowledge. In this, prior probability distribution of any quantity expresses the belief about that particular quantity before considering some evidence.



5. Software tools available for EDA

Open source tools

- **Python**
    - This is an open source programming language widely used in data analysis, data mining, and data science (https:/ / www. python. org/).
- **R programming language**
    - R is an open source programming language that is widely utilized in statistical computation and graphical data analysis (https:/ / www. r- project. org).
- Tableau Public
    - Free Data Visualization Software. It connect to a spreadsheet or file and create interactive data visualizations for the web. (https://public.tableau.com › en-us).
- Power BI
    - It is an interactive data visualization software product developed by Microsoft with a primary focus on business intelligence
- Weka
    - This is an open source data mining package that involves several EDA tools and algorithms (https:/ / www. cs. waikato. ac. nz/ ml/ weka/).
- **KNIME**

This is an open source tool for data analysis and is based on Eclipse (https:/ / www. knime. com/).