



KJSCE/IT/SY/SEM-III/HO-AI-FDS/2022-2023

Experiment No.1

Title: Understanding of the Data

Batch:A2

Roll No.:16010421063

Experiment No.:1

Aim: Understanding of the Data

Resources needed: Any RDBMS, EXCEL, Data storage tool

Theory:

In order to make data ready for data mining process, data exploration is essential step to develop a high-level understanding of the data. Data exploration includes in detail analysis of attributes and their data values and visualization. It aimed at identifying possible relationship between two or more variables/objects.

Broadly classifying, there are two types of attributes, numeric and categorical.

Categorical Attribute:

In categorical, each value represents some kind of category, code, or state. Categorical variables are either nominal or ordinal, depending on the extent of information the numerical coding provides.

The values of a nominal attribute are symbols or names of things. Nominal means “relating to names.”

E.g. hair color and occupation are two attributes describing person objects. Possible values for hair color are black, brown, blond, red, auburn, gray, and white. For occupation, possible values are teacher, dentist, programmer, farmer etc.

An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known. For example, grade attribute with values A+, A, A-, B, C; Student_progress attribute with values Good, average, poor. The central tendency of an ordinal attribute can be represented by its mode and its median (the middle value in an ordered sequence), but the mean cannot be defined.

Nominal, binary, and ordinal attributes are qualitative. That is, they describe a feature of an object without giving an actual size or quantity. The values of such qualitative attributes are typically words representing categories.

Numeric Attributes:

A numeric attribute is quantitative; that is, it is a measurable quantity, represented in integer or real values. Numeric attributes can be interval-scaled or ratio-scaled.

Interval-Scaled Attributes:

Interval-scaled attributes are measured on a scale of equal-size units. The values of interval scaled attributes have order and can be positive, 0, or negative. Thus, in addition to providing a ranking of values, such attributes allow us to compare and quantify the difference between values.

For example, temperature, humidity attributes

Ratio-Scaled Attributes:

A ratio-scaled attribute is a numeric attribute with an inherent zero-point. That is, if a measurement is ratio-scaled, we can speak of a value as being a multiple (or ratio) of another value. In addition, the values are ordered, and we can also compute the difference between values, as well as the mean, median, and mode.

For example, *years of experience*

Procedure / Approach / Algorithm / Activity Diagram:

1. Download the large dataset for the purpose of exploration and ensure that dataset has variety of attributes; number of attributes must be at least 25.
2. Identify the category of each attribute from the dataset which you have created.
3. Identify the attributes which can provide any kind of useful information either collectively or as an individual. Also, discuss about the information provided by the attribute and how it will be computed?

Results: (Program printout with output / Document printout as per the format)

1. Download the large dataset for the purpose of exploration and ensure that the dataset has variety of attributes the number of attributes must be at least 25

top_200_youtubers																	🔍	📄	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	📁	
-------------------	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	--

2. Identify the category of each attribute from the dataset which you have created.

	A	B	
1	Attribute	Type	
2	Country	Nominal	
3	Channel Name	Nominal	
4	Category	Nominal	
5	Main Video Category	Nominal	
6	username	Nominal	
7	followers	ratio	
8	Main Topic	Nominal	
9	More Topics	Nominal	
10	Likes	Ratio	
11	Boost Index	Ratio	
12	Engagement Rate	Interval	
13	Engagement Rate 60days	Interval	
14	Views	Ratio	
15	Views Average	Ratio	
16	Avg. 1 Day	Ratio	
17	Avg. 7 Day	Ratio	
18	Avg. 14 Day	Ratio	
19	Avg. 30 Day	Ratio	
20	Avg. 60 Day	Ratio	
21	Comments Avg.	Ratio	
22			
23			
24			

3. Identify the attributes which can provide any kind of useful information either collectively or as an individual. Also, discuss about the information provided by the attribute and how it will be computed?

- User consistency and growth-** By seeing the change in average viewers over different time periods which would show the viewer growth pattern
Views Avg.
Avg. 1 Day
Avg. 3 Day
Avg. 7 Day
Avg. 14 Day
Avg. 30 day
Avg. 60 day
- Topic Popularity-** We would analyse the topics which are generating more engagement from the users.
Main Topic
More Topic
Likes
Engagement Rate
Engagement Rate 60days
- Channel popularity-** These metrics would help us analyse which YouTuber would be better for promotions as their videos will have more engagement from viewers
Likes
Boost Index

Questions:

1. Compare Discrete and Continuous Attributes. Give at least 5 examples of each.

Ans-

Data that can only take on certain values are discrete data. These values do not have to be complete numbers, but they are values that are fixed. It only contains finite values, the subdivision of which is not possible.

Example-

- Shoe size
- IQ tests
- Gender
- The number of pages in a book
- Population

Continuous data is the data that can be of any value. Over time, some continuous data can change. It may take any numeric value, within a potential value range of finite or infinite.

The continuous data can be broken down into fractions and decimals

Example-

- Time
- Temperature
- Length
- Weight
- Height

Outcomes:

CO1: Summarize the data

CO2: Comprehend descriptive and proximity measures of data

Conclusion: (Conclusion to be based on the objectives and outcomes achieved)

Found Dataset with more than 20 attributes and found the types of various attributes. Also

Understood the data to find relations which would be able to give us meaningful information.

Grade: AA / AB / BB / BC / CC / CD /DD

Signature of faculty in-charge with date

References:

Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining
Morgan Kaufmann 3nd

Concepts and Techniques",
Edition



(

(A Constituent College of Somaiya Vidyavihar University)