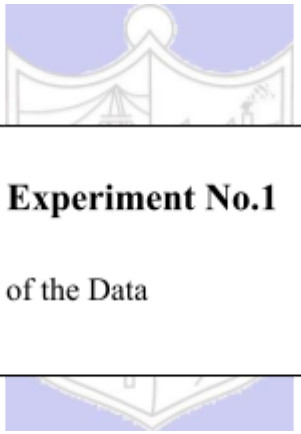


Experiment No.1

Title: Understanding of the Data

Batch: A1 **Roll No.: 16010421015** **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.

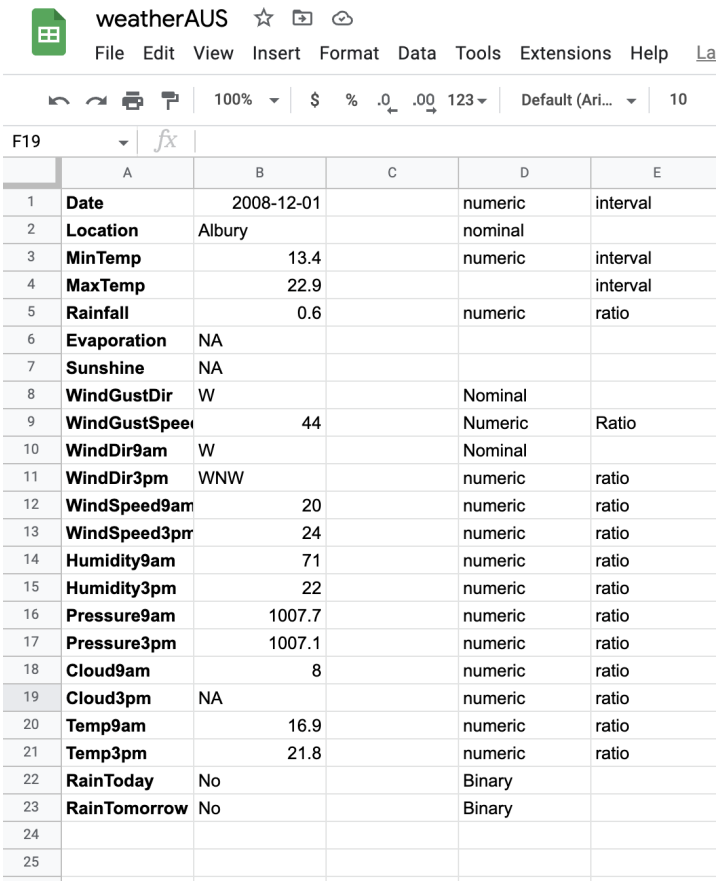
Interval-Scaled Attributes:
Interval-scaled attributes are measured on a scale of equal-size units. The values of interval-scaled attributes have order and can be positive, 0, or negative. Thus, in addition to providing a ranking of values, such attributes allow us to compare and quantify the difference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



weatherAUS

File Edit View Insert Format Data Tools Extensions Help La

100% $ % .0 .00 123 Default (Ari... 10

F19 fx

	A	B	C	D	E
1	**Date**	2008-12-01		numeric	interval
2	**Location**	Albury		nominal	
3	**MinTemp**	13.4		numeric	interval
4	**MaxTemp**	22.9			interval
5	**Rainfall**	0.6		numeric	ratio
6	**Evaporation**	NA			
7	**Sunshine**	NA			
8	**WindGustDir**	W		Nominal	
9	**WindGustSpee**	44		Numeric	Ratio
10	**WindDir9am**	W		Nominal	
11	**WindDir3pm**	WNW		numeric	ratio
12	**WindSpeed9am**	20		numeric	ratio
13	**WindSpeed3pm**	24		numeric	ratio
14	**Humidity9am**	71		numeric	ratio
15	**Humidity3pm**	22		numeric	ratio
16	**Pressure9am**	1007.7		numeric	ratio
17	**Pressure3pm**	1007.1		numeric	ratio
18	**Cloud9am**	8		numeric	ratio
19	**Cloud3pm**	NA		numeric	ratio
20	**Temp9am**	16.9		numeric	ratio
21	**Temp3pm**	21.8		numeric	ratio
22	**RainToday**	No		Binary	
23	**RainTomorrow**	No		Binary	
24					
25					

<u>Attributes which provide useful information about :-</u>

1) **Seasons in every region of the country:**
   - **Date**
   - **Location**

   These two attributes help us understand when and where the rainfall occurs in Australia and accordingly helps us understand the Rainfall Season in each region every year.

2) **Regions with high possibility of Natural Disasters:**
   - **Pressure**
   - **WindDir**
   - **WindSpeed**

   The study of atmospheric pressure is important to us as **it is an important component of changing weather and climatic conditions**.
   b. Since atmospheric pressure varies on Earth , and is the cause of various phenomenon like Jet streams, tropical cyclones, it is important to study the rate of atmospheric pressure.

3) **Reanalysis:**
   - **Majority Attributes in the Dataset**
     Reanalysis has benefits for Numerical Weather Prediction (NWP) as well as climate studies. It has an important role in providing high quality and detailed data on the climate of the past and present which are required to support decisions for adaptation.

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

BASIS FOR COMPARISON	DISCRETE VARIABLE	CONTINUOUS VARIABLE
Meaning	Discrete variable refers to the variable that assumes a finite number of isolated values.	Continuous variable alludes to the a variable which assumes infinite number of different values.
Range of specified number	Complete	Incomplete
Values	Values are obtained by counting.	Values are obtained by measuring.
Classification	Non-overlapping	Overlapping
Assumes	Distinct or separate values.	Any value between the two values.
Represented by	Isolated points	Connected points

Outcomes:

CO1: Summarize the Data

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

From the following Experiment we have learnt how to understand the data from the given datasets on the basis of their types which will help us in further studying and analysis of the datasets.

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

Signature of faculty in-charge with date

References:

Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition