



**Batch: A3 Roll No.:16010421075 Experiment No.:1**

**Aim**: Understanding of the Data

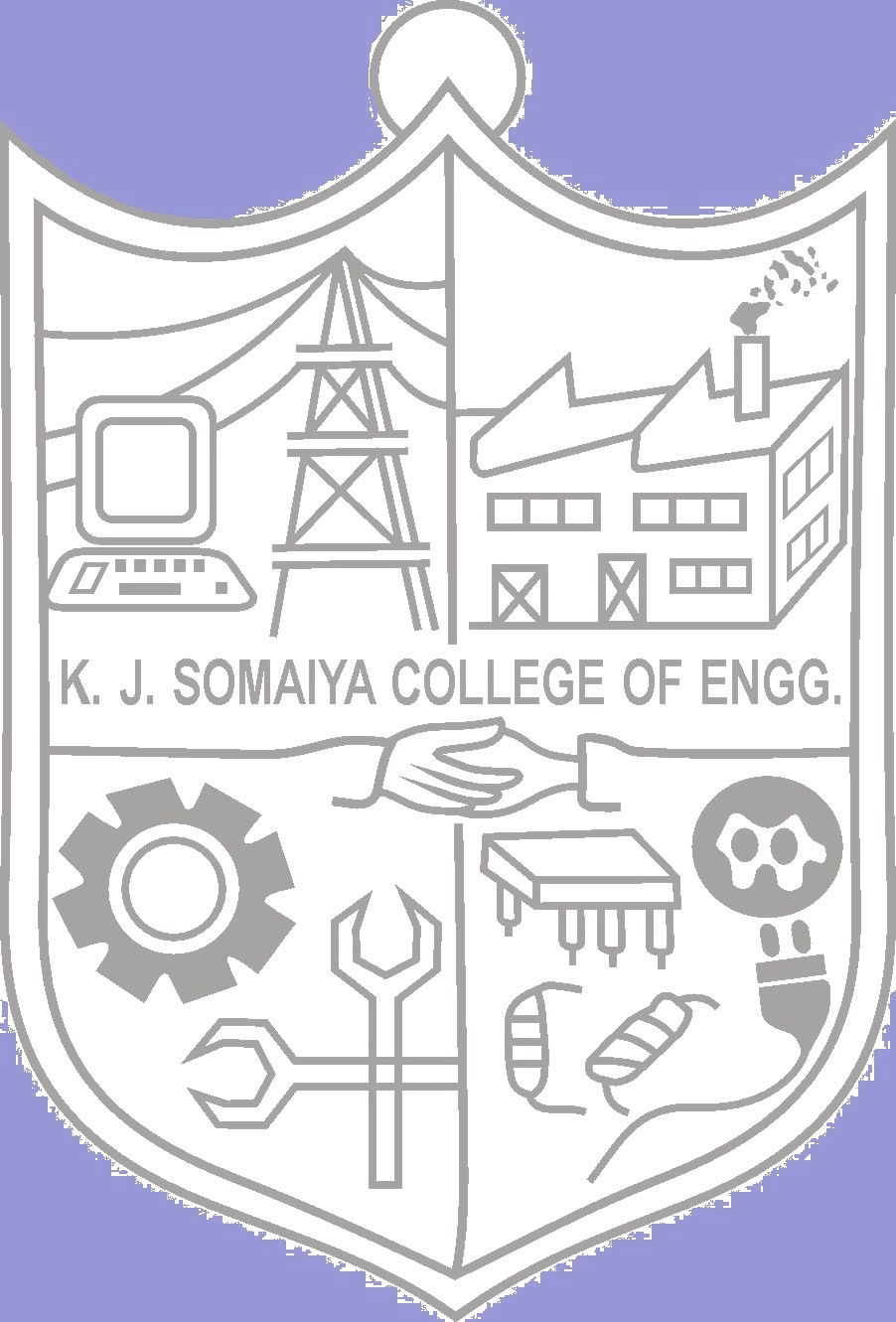
**Resources needed:** Any RDBMS, EXCEL, Data storage tool

# 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.

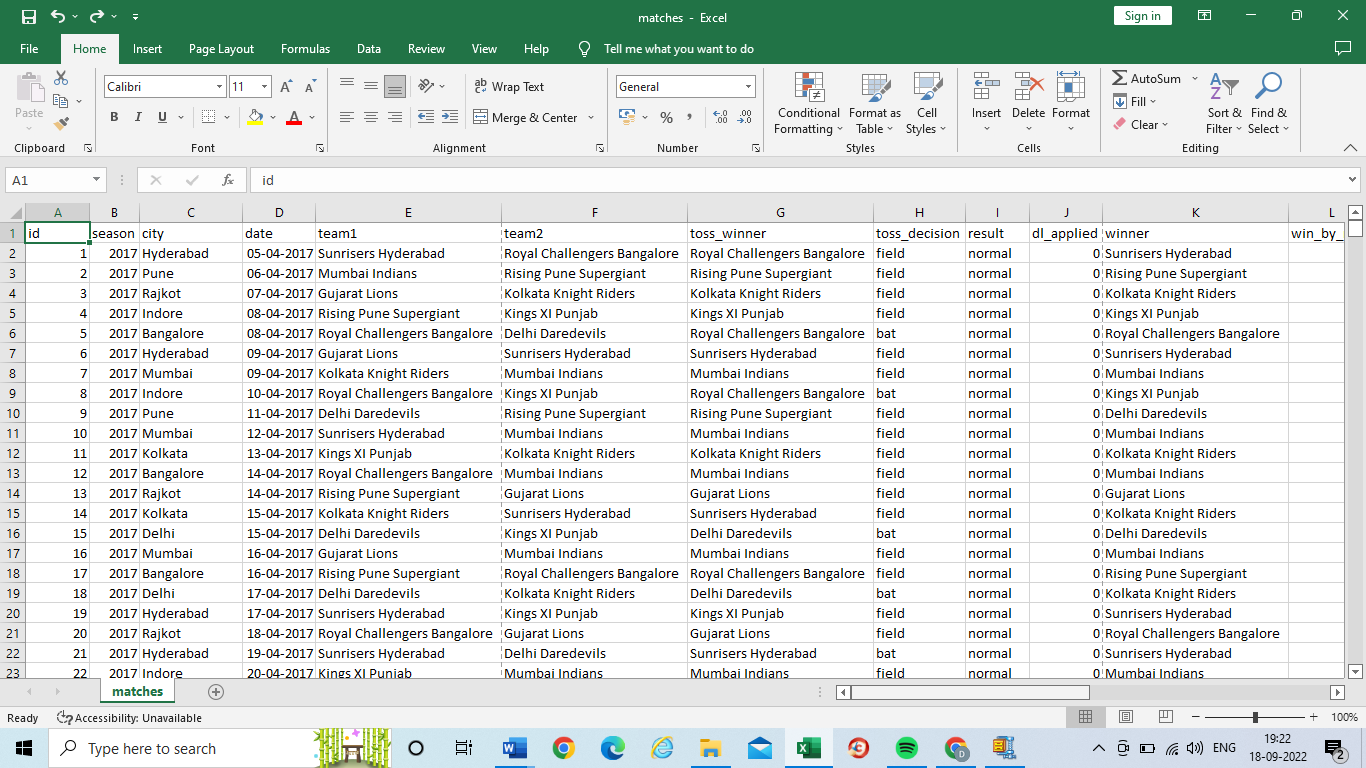
For example, *years of experience*

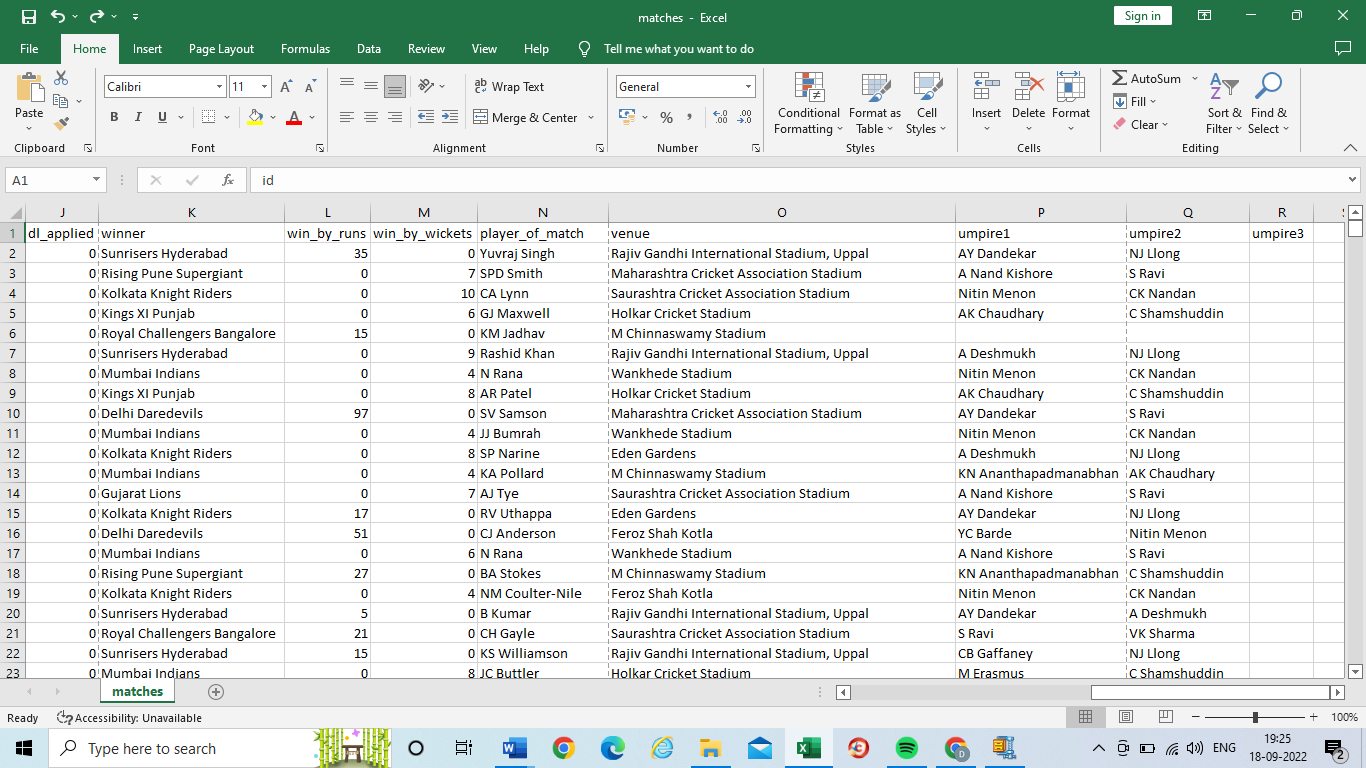
# Procedure / Approach /Algorithm / Activity Diagram:

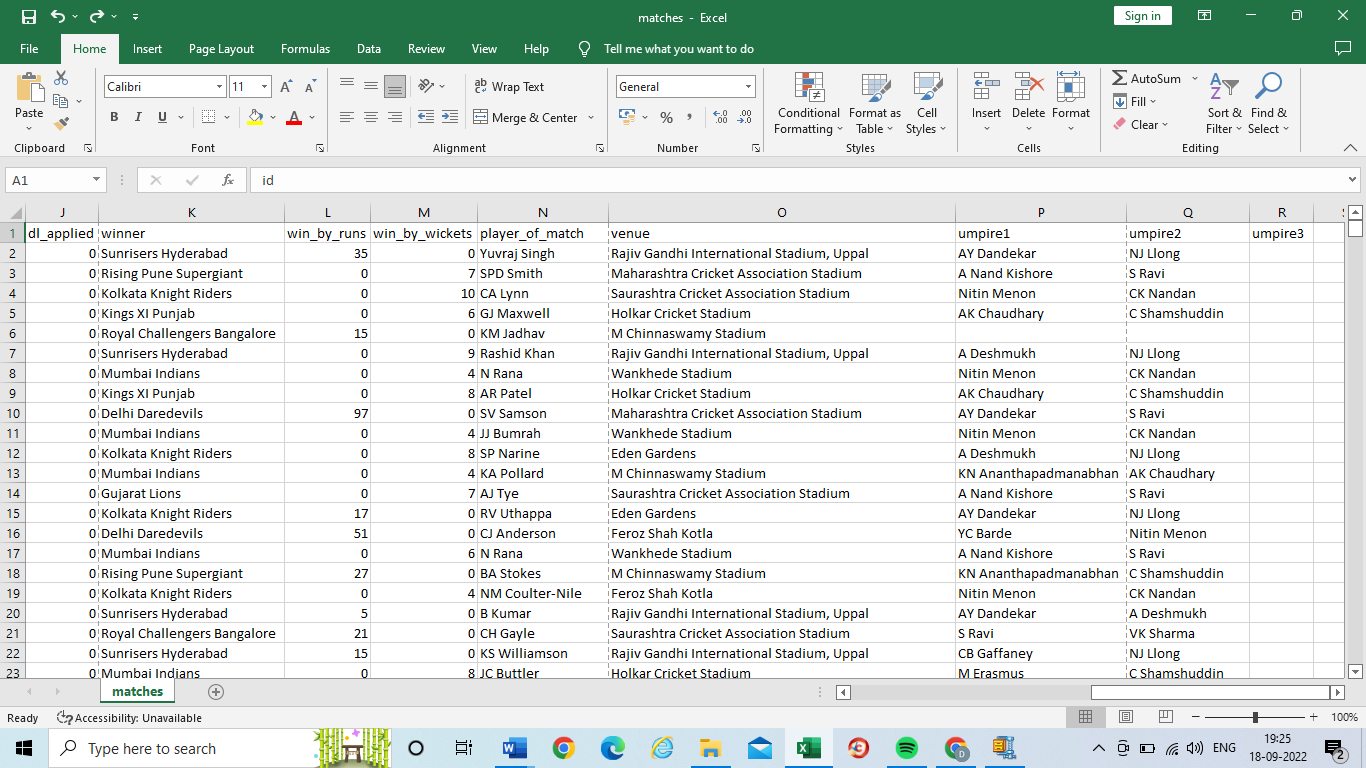
1. Download the large dataset for the purpose of exploration and ensure that dataset has variety of attributes; number of attributes must be at least 25.
2. Identify the category of each attribute from the dataset which you have created.
3. Identify the attributes which can provide any kind of useful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

**Results:**

Topic: Dataset of IPL







Types of Attributes:

Column A (ID): Categorical (ordinal) attribute

Column B (Season): Numeric (interval) attribute

Column C (city): Categorical (nominal) attribute

Column D (Date): Numeric (interval) attribute

Column E,F,G,H: Categorical (nominal) attribute

Column I (result): Categorical (ordinal) attribute

Column J (dl applied): Numeric (interval) attribute

Column K (winner): Categorical (nominal) attribute

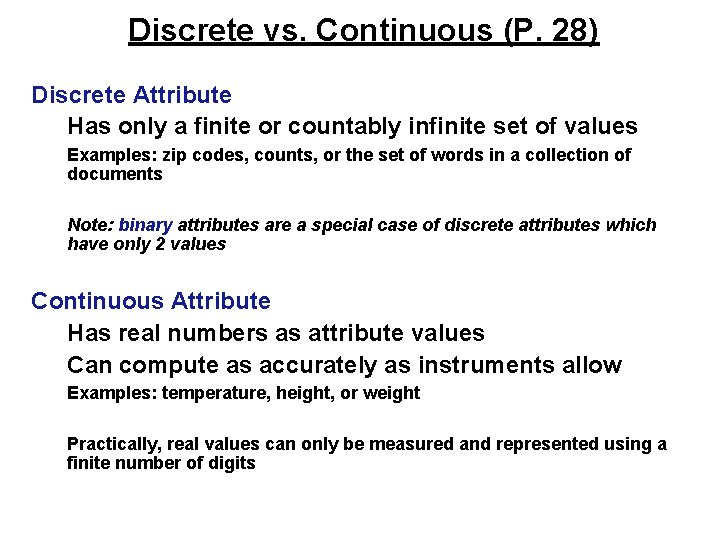
Column L,M (win by runs and wickets): Numeric (ratio scaled) attribute

Column N,O,P,Q : Categorical (nominal) attribute

# Post Lab Questions:

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

Ans:



# Outcomes: CO1: Summarize the data

**Conclusion:** We successful learnt and implemented the types of data in a data set and bifurcated them accordingly. We also learnt the diﬀerence between discrete and continuous attributes.

**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