Statistical Concepts

Data exploration

Data exploration is the first step in data analysis involving the use of data visualization tools and statistical techniques to uncover data set characteristics and initial patterns.

This is also sometimes referred to as exploratory data analysis, which is a statistical technique employed to analyze raw data sets in search of their broad characteristics.

Why is data exploration important?

It's challenging for data scientists to review thousands of rows of data points and infer meaning without assistance.

Data visualization tools and elements like colors, shapes, lines, graphs and angles aid in effective data exploration of metadata, enabling relationships or anomalies to be detected.

What industries use data exploration?

Any business or industry that collects or utilizes data can benefit from data exploration. A few common industries include software development, healthcare and education.

By visualizing patterns and finding commonalities in complex data flows, data exploration can help enterprises make data-driven decisions to streamline processes, better target their ideal audience, increase productivity and achieve greater returns.

Data exploration tools

Data exploration tools make data analysis easier to present and understand through interactive, visual elements, making it easier to share and communicate key insights.

Data exploration tools include data visualization software and business intelligence platforms, such as Microsoft Excel, Microsoft Power BI, Qlik and Tableau.

Exploratory Data Analysis

Raw data are not very informative. Exploratory Data Analysis (EDA) is how we make sense of the data by converting them from their raw.

In particular, EDA consists of:

- **organizing and summarizing** the raw data,
- **discovering important features and patterns** in the data and any striking deviations from those patterns, and then
- **interpreting our findings** in the context of the problem

There are two important features to the structure of the EDA

Examining Distributions — exploring data one variable at a time (univariate).

Some univariate data examples are salaries of employees in a company, the number of pets in different households, heights of students in a certain age group.

Examining Relationships — exploring data two variables at a time (bivariate).

Sale of Ice cream compared to the temperature of that day. Traffic accidents along with the weather on a particular day

In Exploratory Data Analysis, our exploration of data will always consist of the following two elements: visual displays, supplemented by numerical measures.

Population

It includes all the elements from the data set and measurable characteristics of the population such as mean and standard deviation are known as a **parameter**. For example, all people living in India indicates the population of India.

There are different types of population. They are:

- Finite Population
- Infinite Population
- Existent Population
- Hypothetical Population

Finite Population

The finite population is also known as a countable population in which the population can be counted. In other words, it is defined as the population of all the individuals or objects that are finite

Infinite Population

The infinite population is also known as an uncountable population in which the counting of units in the population is not possible. Example of an infinite population is the number of germs in the patient's body is uncountable.

Existent Population

The existing population is defined as the population of concrete individuals. In other words, the population whose unit is available in solid form is known as existent population. Examples are books, students etc.

Hypothetical Population

The population in which whose unit is not available in solid form is known as the hypothetical population. A population consists of sets of observations, objects etc that are all something in common. In some situations, the populations are only hypothetical. Examples are an outcome of rolling the dice, the outcome of tossing a coin.

Population and Sample Examples

- All the people who have the ID proofs is the population and a group of people who only have voter id with them is the sample.
- All the students in the class are population whereas the top 10 students in the class are the sample.
- All the members of the parliament are population and the female candidates present there is the sample.

Definition of Sample

A part of population chosen at random for participation in the study.

The sample so selected should be such that it represents the population in all its characteristics, and it should be

In other words, the respondents selected out of population constitutes a ‘sample’.

• The process of selecting respondents is known as ‘**sampling.’**

• The units under study are called **sampling units.**

• The number of units in a sample is called **sample size.**