SNS COLLEGE OF TECHNOLOGY

Coimbatore-37.

An Autonomous Institution

COURSE NAME : 19CSE301 INTRODUCTION TO DATA SCIENCE

III YEAR/ V SEMESTER

UNIT – II

Topic: Exploratory Data Analysis

Ms.B.Sumathi

Assistant Professor

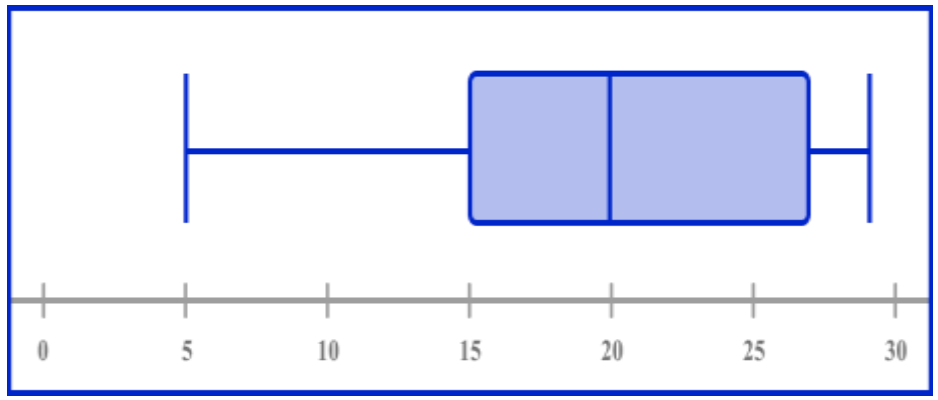
Department of Computer Science and Engineering

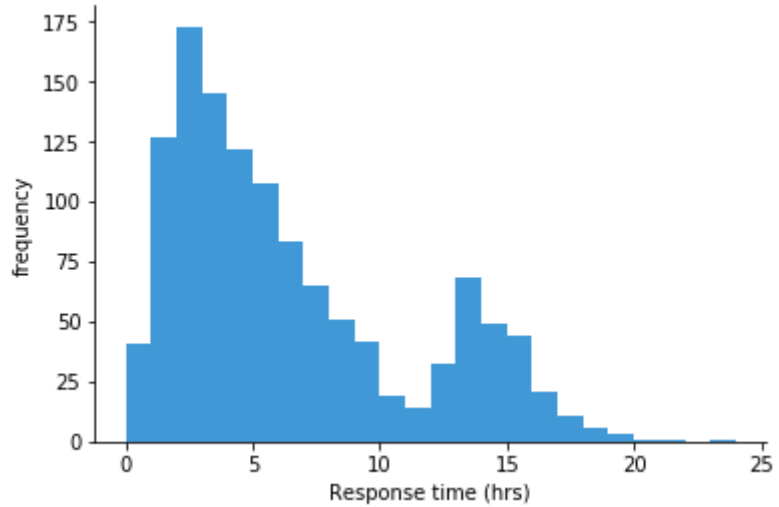
Exploratory Data Analysis

- Exploratory Data Analysis is a data analytics process to understand the data in depth and learn the different data characteristics, often with visual means.
- This allows you to get a better feel of your data and find useful patterns in it.
- Exploratory data analysis is generally cross-classified in two ways.
- First, each method is either non-graphical or graphical.
- second, each method is either univariate or multivariate (usually just bivariate).

Univariate Analysis

- Univariate analysis is the simplest form of data analysis, where the data being analyzed consists of only one variable.
- Since it's a single variable, it doesn't deal with causes or relationships.
- The main purpose of univariate analysis is to describe the data and find patterns that exist within it.
- Box Plot
- Histogram



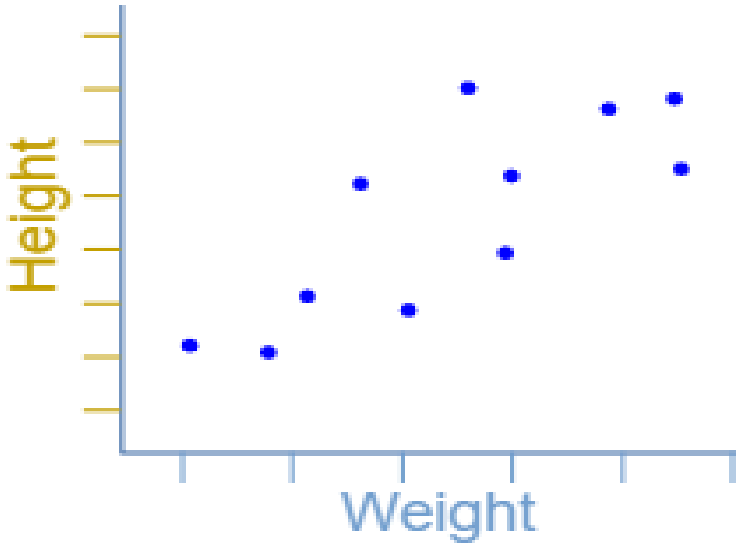


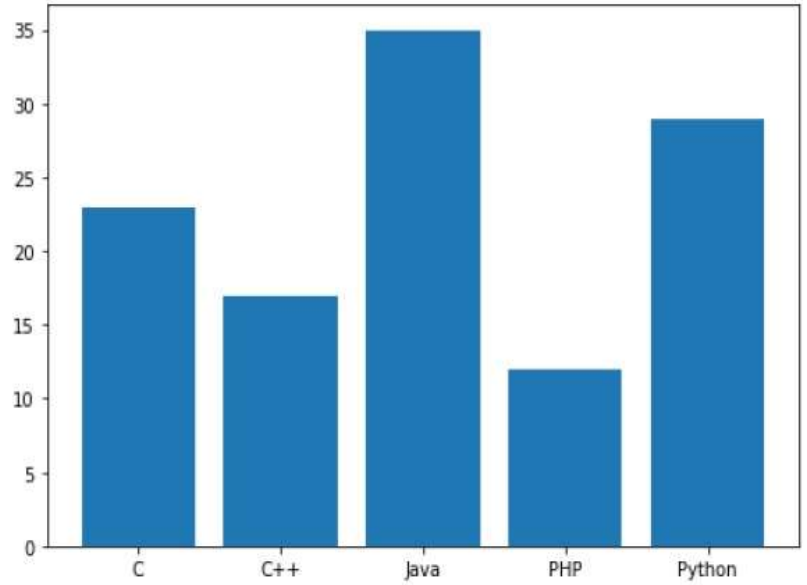
Bivariate Analysis

- The aim is to find patterns/relationships within the dataset using two attributes.
- It is useful in testing simple associations.
- One plot which can be used for the analysis is the **pair plot**.
- Pair plots are an easy way to visualize relationships within your data.
- A matrix of each variable associated with another variable is produced for our analysis.

Multivariate Analysis

- Multivariate data analysis refers to any statistical technique used to analyze data that arises from more than one variable. This models more realistic applications, where each situation, product, or decision involves more than a single variable.
- Scatter Plot
- Bar Plot





Steps in Exploratory Data Analysis

Data collection:

- Data collection is an essential part of exploratory data analysis. It refers to the process of finding and loading data into our system. Good, reliable data can be found on various public sites or bought from private organizations

Data Cleaning:

- Data Cleaning refers to the process of removing unwanted variables and values from your dataset and getting rid of any irregularities in it. Such anomalies can disproportionately skew the data and hence adversely affect the results. Some steps that can be done to clean data are:
- Removing missing values, outliers, and unnecessary rows/ columns.
- Re-indexing and reformatting our data.

References

- Tom M. Mitchell, “Machine Learning”, McGraw-Hill Education (India) Private Limited, 2013.
- 2Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, “An Introduction to Statistical Learning: with Applications in R”, Springer; First Edition 2013.


Thank
You