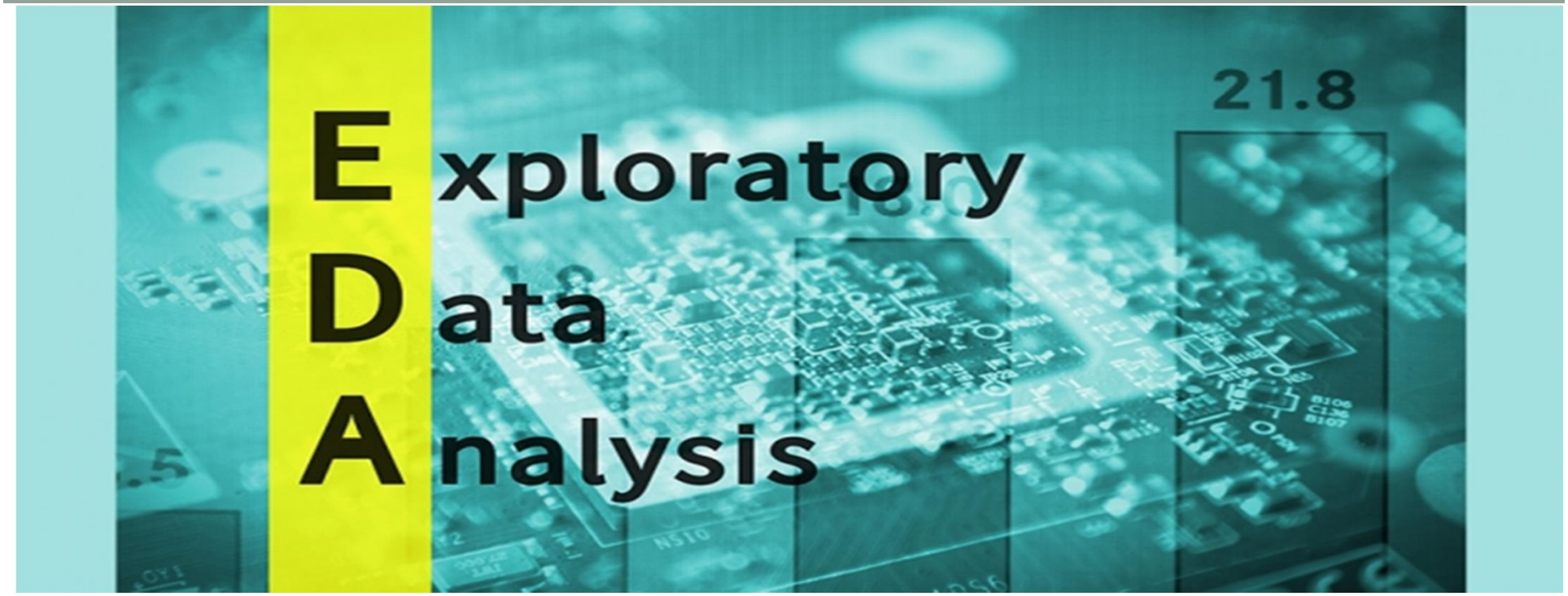
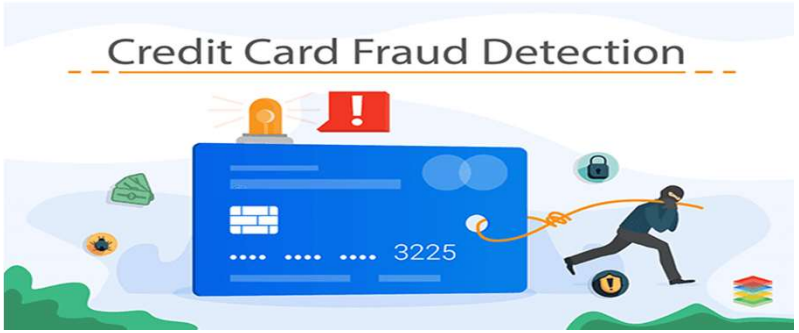




Md Firoz Alam

ML & DL Intern

Email:-firozalam1160006@gmail.com

Exploratory Data Analysis(EDA)

Table of Contents:-

Introduction

- Exploratory Data Analysis (EDA) is a crucial step in the data analysis process, where analysts summarize main characteristics, often with visual methods, to better understand the dataset. It helps in uncovering patterns, spotting anomalies, and testing hypotheses.

- Exploratory Data Analysis (EDA) is a process of describing the data by means of statistical and visualization techniques in order to bring important aspects of that data.

- EDA is a crucial initial step in data science projects. Its normally carried out as a preliminary step before undertaking extra formal statistical analyses or modeling. Skipping EDA may result in includes generating inaccurate models, generating accurate models but on the wrong data, not creating the right types of variables in data preparation, and using resources inefficiently.

Cont..

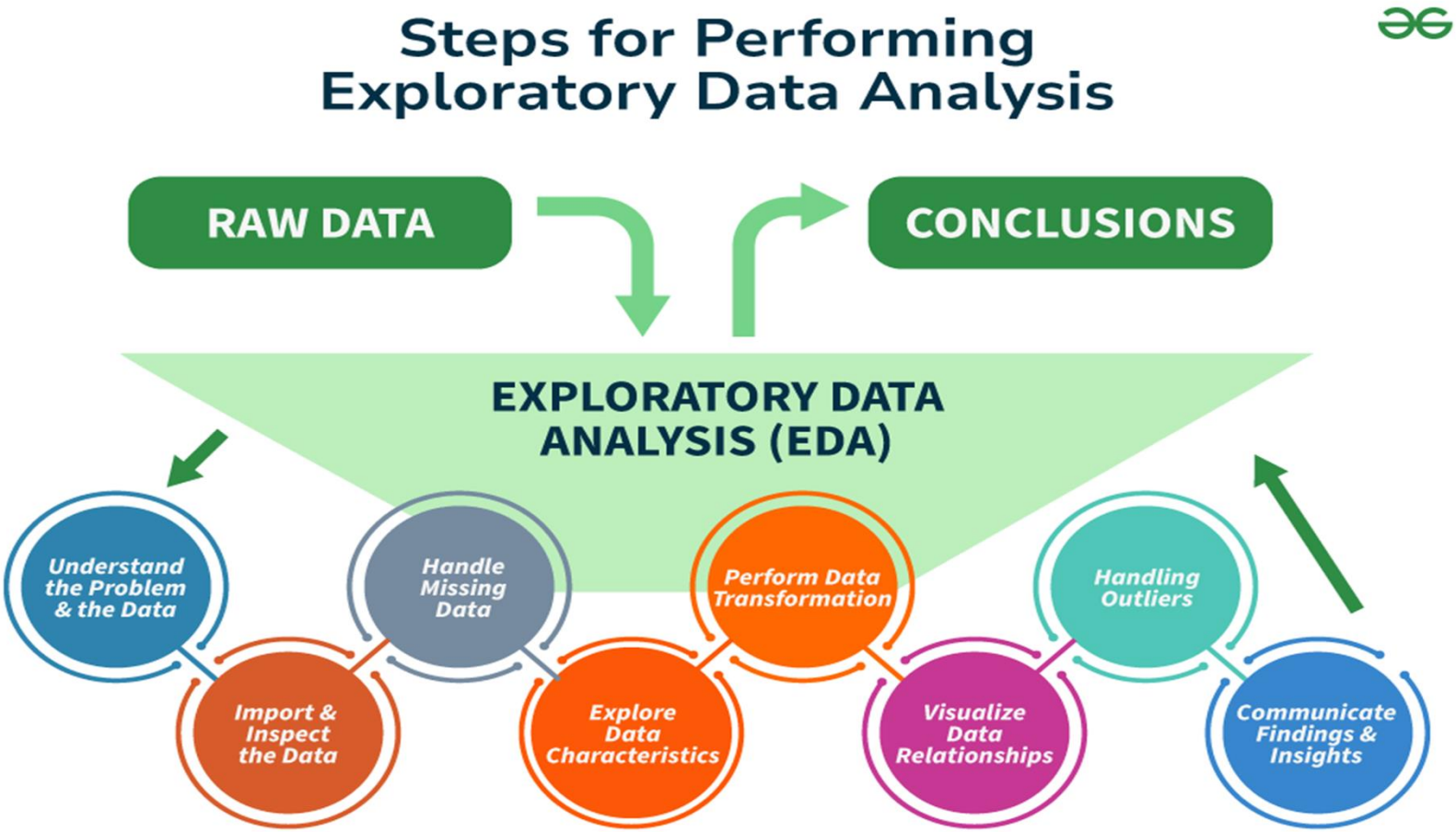
- Suppose you are an employee at a credit card company. Often times you receive complaints about fraud transactions. You want to solve this problem so that customers are not charged for items that they did not purchase. You have loads and loads of transaction data available.

- You want to make use of this data to identify patterns and derive insights from it. This would help convert the raw data into useful information and understand the characteristics of a fraud transaction.

Types of EDA

- **Univarate Analysis:** In EDA Analysis, univariate analysis exam individual variables to understand their distributions and summary statistics.
- **Bivariate Analysis:** This aspect of EDA explores the relationship between two variables, uncovering patterns through techniques like scatter plots and correlation analysis.
- **Visualization Techniques:** EDA relies heavily on visualization methods to depict data distributions, trends, and associations using various charts and graphs.
- **Outlier Detection:** Outlier detection is a method used to find unusual or abnormal data points in a set of information. Imagine you have a group of friends, and you're all about the same age, but one person is much older or younger than the rest. That person would be considered an outlier because they stand out from the usual pattern.

Different Key Steps in EDA



Problem Statement

Problem Statement:-

- The main problem statement of EDA is to explore and understand the data comprehensively, ensuring that any subsequent analysis or modeling is based on accurate, well-understood, and properly prepared data

Objective of EDA

- **Summarization:** Provide concise summaries of the data through statistics and visualizations.
- **Visualization:** Create visual representations to easily observe patterns, trends, and relationships.
- **Pattern Detection:** Identify significant patterns and trends within the data.
- **Anomaly Detection:** Spot any anomalies or outliers that could affect subsequent analyses.
- **Assumption Validation:** Check the assumptions underlying statistical methods and models.
- **Preparation for Modeling:** Clean and transform the data to prepare it for modeling and analysis.

Cont..

- The objective of this report is to evaluate the importance of EDA in data analysis, highlight its methodologies, and demonstrate its application through practical examples.

Technology Stack

Technology Stack:-

- For conducting EDA, various tools and programming languages can be employed. Commonly used technologies include:
- Programming Languages: Python, R
- Libraries: Pandas, NumPy, Matplotlib, Seaborn (for Python); ggplot2 (for R)
- Visualization Tools: Tableau, Power BI
- Statistical Software: SPSS(Statistical Package for the Social Sciences), SAS(Statistical Analysis System)

Project Demonstration

- **Project Demonstration**
- **Data Collection**: Gather the dataset relevant to the analysis.
- **Data Cleaning**: Handle missing values, outliers, and inconsistencies in the dataset.
- **Descriptive Statistics**: Calculate measures like mean, median, mode, standard deviation, etc., to understand the central tendency and spread of data.
- **Univariate Analysis**: Analyze individual variables to identify patterns and distributions.
- **Bivariate Analysis**: Explore relationships between pairs of variables using scatter plots, correlation matrices, etc.
- **Multivariate Analysis**: Investigate interactions between multiple variables through techniques like cluster analysis, principal component analysis (PCA), etc.
- **Visualization**: Create visual representations such as histograms, box plots, heatmaps, etc., to gain insights into the data.
- **Hypothesis Testing**: Conduct statistical tests to validate assumptions or hypotheses about the data.
- **Insights Generation**: Summarize findings and draw actionable insights from the analysis.

Conclusion

- Exploratory Data Analysis is a fundamental step in the data analysis pipeline, providing valuable insights into the underlying structure and characteristics of the dataset. By employing various techniques and methodologies, analysts can uncover hidden patterns, identify trends, and make informed decisions based on data-driven evidence.

Reference

- McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media.
- Wickham, H., & Grolemund, G. (2017). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media.
- Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.


Q&A


Thank You
For Your
Attention