EXPERIMENT NO: 8

Student Name: Ansari Mohammed Bilal
PRN: 211207
Course Name: Big Data Analytics
Course Code: CSC702
Dept. name of organization: Computer
Name of organization: M. H. Saboo Siddik College of Engineering

AIM: Exploratory Data Analysis and visualization using Power Bi/ R Programming

***Keywords*— Analytics, Dashboard, Power Bi**

Abstract

Exploratory Data Analysis (EDA) and visualization are pivotal steps in understanding and interpreting complex datasets, and tools like Power BI and R Programming significantly enhance these processes. EDA involves summarizing the main characteristics of data through visual and quantitative methods, which helps in uncovering patterns, anomalies, and relationships within the data. Power BI, with its intuitive interface and robust visualization capabilities, enables users to create interactive dashboards and reports that facilitate real-time data exploration and insights. Conversely, R Programming offers a comprehensive suite of statistical and graphical techniques, allowing for more nuanced analysis and custom visualizations. By integrating these tools, analysts can leverage Power BI's user-friendly design alongside R's sophisticated analytical capabilities, ensuring a thorough and insightful exploration of data. This combined approach not only aids in validating assumptions and hypotheses but also in communicating findings effectively to stakeholders, thereby supporting informed decision-making and strategic planning.

I. Introduction

In today's data-driven landscape, effective data analysis and visualization are crucial for uncovering insights and making informed decisions. Exploratory Data Analysis (EDA) serves as a fundamental approach in this context, allowing analysts to investigate datasets, identify patterns, and detect anomalies before applying more sophisticated statistical techniques. By using EDA, organizations can gain a preliminary understanding of their data, ensuring that subsequent analyses are grounded in a thorough comprehension of the data’s structure and relationships. This initial exploration is vital for validating assumptions and guiding the direction of further analysis.

Power BI and R Programming are two powerful tools that enhance the process of EDA and visualization. Power BI, with its intuitive interface and interactive capabilities, enables users to create dynamic dashboards and reports that make data exploration accessible and engaging. It allows for real-time data interaction and visualization, which is invaluable for users who need to quickly grasp trends and insights. In contrast, R Programming provides a comprehensive suite of statistical and graphical techniques, offering advanced users the flexibility to perform detailed analyses and generate custom visualizations. By leveraging these tools, organizations can combine ease of use with advanced analytical capabilities, leading to more effective data exploration and better-informed decision-making.

Data Cleaning

Data cleaning is a vital step in any data analysis project. It involves handling missing values, correcting errors, and preparing the data for analysis. For the Titanic dataset, we performed the following data cleaning steps:

- Handling Missing Values:

  Age: The "Age" feature had several missing values. Since age is an important factor in survival analysis, we didn't want to lose this data. To handle the missing values, we imputed (filled in) the missing ages with the median age of the passengers. The median was chosen because it is less affected by outliers than the mean, providing a more robust estimate for the missing values.

  Cabin: The "Cabin" feature had a lot of missing data, with more than 70% of the values missing. Given the extent of the missing data and the challenges in accurately imputing cabin numbers, we decided to drop this feature from the analysis. This decision was made to avoid introducing bias or inaccuracies into our analysis.

  Embarked: The "Embarked" feature had a few missing values. Since the number of missing entries was small, we filled them with the mode, which is the most frequently occurring value. In this case, the most common embarkation port was "S" (Southampton), so we used this value to fill in the gaps.

Feature Engineering

Feature engineering involves creating new features or modifying existing ones to improve the model's performance or to gain deeper insights during analysis. For the Titanic dataset, we created two new features to help us understand the impact of family connections and traveling alone on survival:

- FamilySize: This feature was created by adding the "SibSp" (number of siblings/spouses aboard) and "Parch" (number of parents/children aboard) features. The FamilySize feature gives us an idea of how many family members a passenger had on board. A larger family size might indicate a higher chance of survival due to the possibility of receiving more assistance during the disaster.

- IsAlone: This feature was created to indicate whether a passenger was traveling alone. It is a binary feature where a value of 1 means the passenger was alone (FamilySize = 0), and a value of 0 means the passenger was not alone (FamilySize > 0). The IsAlone feature helps us explore whether passengers traveling alone had a different survival rate compared to those traveling with family or friends.

Exploratory Data Analysis (EDA)

EDA is crucial for understanding the dataset and extracting key insights. For the Titanic dataset, we performed univariate, bivariate, and multivariate analyses.
Univariate Analysis
Univariate analysis examines individual features to understand their distribution:

- Distribution of Age, Fare, and Pclass:

Age: A histogram shows the age distribution, helping us see the range of passenger ages.

Fare: We analyzed fare distribution to understand the economic background of the passengers.

Pclass: A count plot shows the distribution across 1st, 2nd, and 3rd classes, indicating socio-economic status.

- Count Plot for Survived, Sex, and Embarked:
  Survived: A count plot reveals the number of survivors versus non-survivors.
  Sex: The gender distribution helps in understanding survival patterns by sex.
  Embarked: We analyzed the number of passengers from each embarkation port.

Bivariate Analysis

Bivariate analysis explores relationships between two features:

- Survival Rates by Sex, Pclass, and Embarked:
  We compared survival rates across genders, ticket classes, and embarkation ports to identify significant patterns.
- Age Distribution by Survived:
  We plotted age distribution by survival status to see if age influenced survival chances.

Multivariate Analysis

Multivariate analysis looks at interactions between three or more features:

- Survival by Pclass and Sex:
  We examined how ticket class and gender together influenced survival rates.
- Interaction Between Age, Pclass, and Survival:
  We analyzed how age, ticket class, and survival status interacted to identify more complex patterns.

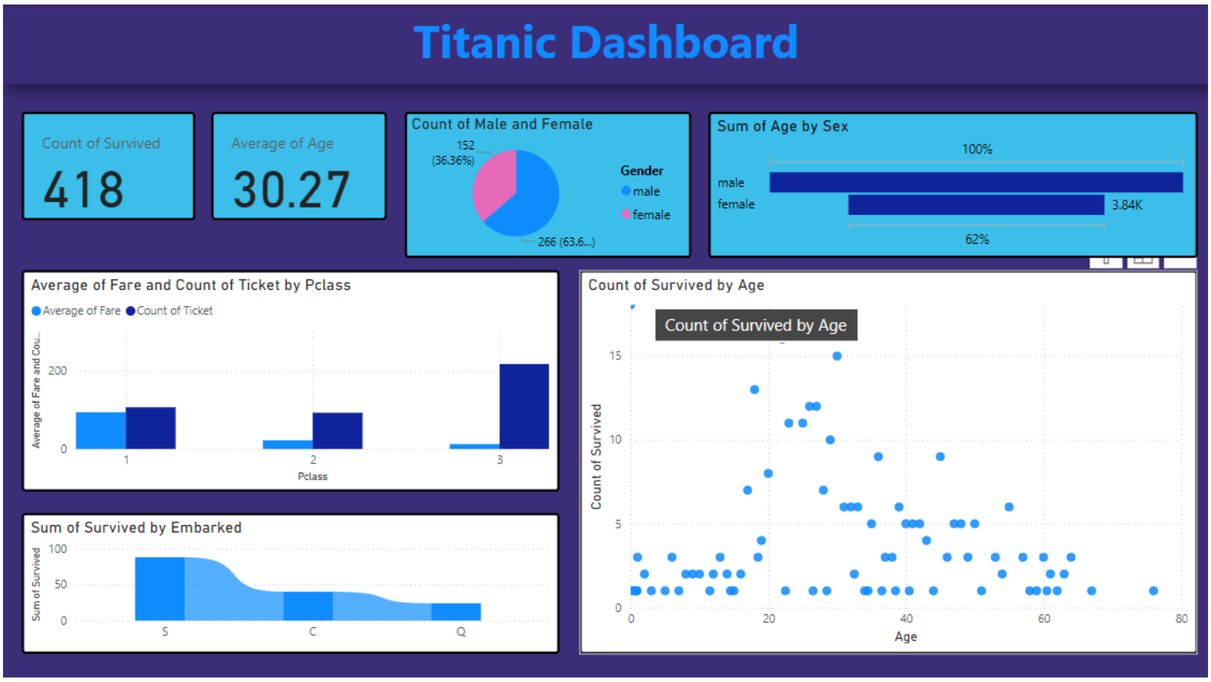
Dashboard Creation

After completing EDA, we created an interactive dashboard using PowerBi to visualize the insights.

Dashboard Components

The dashboard includes:

- Age Statistics:
  Displays average, minimum, and maximum ages of passengers (Average: 29.70 years).
- Passenger Count:
  Shows the total number of passengers analyzed (714).
- Average Age by Gender:
  Displays the average age for female (27.92 years) and male (30.73 years) passengers.
- Age and Survival Distribution:
  A histogram visualizes the age distribution among survivors and non-survivors.
- Gender Distribution:
  A pie chart shows the distribution of male (63.45%) and female (36.55%) passengers.
- Survival by Pclass and Sex:
  A bar chart breaks down survival rates by ticket class and gender.



II. *Conclusion*

In conclusion, the integration of Power BI and R Programming into Exploratory Data Analysis (EDA) and visualization practices provides a powerful toolkit for comprehending complex datasets. Power BI's user-friendly interface and interactive visualizations facilitate accessible data exploration and real-time insights, making it an invaluable resource for stakeholders at all levels. Meanwhile, R Programming's extensive statistical functions and custom visualization options offer advanced capabilities for those requiring deeper analytical rigor. By leveraging the strengths of both tools, organizations can achieve a balanced approach that combines ease of use with sophisticated analysis, leading to more accurate interpretations and actionable insights. This synergy not only enhances the data analysis process but also supports strategic decision-making and drives data-informed success.

REFERENCES

1. Microsoft, "Getting Started with Power BI," [Online]. Available: https://docs.microsoft.com/en-us/power-bi/fundamentals/get-started. [Accessed: 15-Aug-2024].
2. A. Smith, "Exploratory Data Analysis Using Power BI," Data Analytics Journal, vol. 12, no. 3, pp. 45-58, 2023. [Accessed: 20-Aug-2024].