ABSTRACT

This project presents an Exploratory Data Analysis (EDA) of responses collected from participants of a Data Science Workshop. The analysis utilizes Python libraries such as pandas, numpy, matplotlib, and seaborn to systematically clean, transform, and visualize the dataset. Key steps include renaming columns for clarity, handling missing and inconsistent data, and converting data types for accurate analysis. The project focuses on understanding participant demographics and characteristics, such as age distribution, height, weight, and travel details. Through data visualization and summary statistics, the project uncovers patterns and insights that can inform future workshops and participant engagement strategies. This work demonstrates essential data preprocessing and analysis techniques, providing a foundation for more advanced data science tasks.

INTRODUCTION

This project focuses on Exploratory Data Analysis (EDA) of responses collected from participants of a Data Science Workshop. Using Python and popular data analysis libraries such as pandas, numpy, matplotlib, and seaborn, the project aims to clean, transform, and visualize the dataset. The analysis provides insights into various attributes of the participants, such as age, weight, height, and travel details, helping to understand the demographic and other characteristics of the workshop attendees. Through systematic data preprocessing and visualization, the project demonstrates essential steps in preparing and analyzing real-world data for meaningful interpretation.

RATIONALE BEHIND THIS PROJECT

The rationale behind this project is to develop practical data analysis skills by working with real-world survey data collected from a Data Science Workshop. Handling such data often involves challenges like inconsistent column names, missing values, and the need for data transformation before meaningful analysis can be performed. By systematically cleaning, transforming, and visualizing the dataset, this project demonstrates essential steps in the data analytics workflow. The project not only helps in understanding the demographic and behavioral patterns of workshop participants but also provides hands-on experience with Python data analysis libraries. This foundational work is crucial for making informed decisions, identifying trends, and preparing data for more advanced analytics or machine learning tasks.

OBJECTIVES

1. **To perform exploratory data analysis (EDA) on the Data Science Workshop responses dataset.**
2. **To clean and preprocess the data** by renaming columns, handling missing values, and correcting data types for accurate analysis.
3. **To analyze demographic and physical characteristics** of participants, such as age, weight, height, and travel details.
4. **To visualize key attributes** (e.g., age distribution) using appropriate plots for better understanding and communication of insights.
5. **To develop practical skills in data handling and visualization** using Python libraries like pandas, numpy, matplotlib, and seaborn.
6. **To extract meaningful patterns and trends** that can inform future workshops and participant engagement strategies.

METHODOLOGY

The methodology for this project involves the following systematic steps:

1. **Data Import**
    - The dataset is loaded from an Excel file using pandas.read_excel() for further analysis.
2. **Data Inspection**
    - The structure and basic information of the dataset are examined using df.info() and by displaying the columns.
3. **Data Cleaning**
    - Columns with lengthy or unclear names are renamed for clarity and ease of use.
    - Data types are corrected (e.g., converting the "Age" column to numeric).
    - Invalid or unrealistic values (such as negative ages or ages above 100) are filtered out.
    - Missing values are identified and rows containing any NaN values are removed to ensure data quality.
4. **Data Transformation**
    - Additional transformations are performed as needed, such as standardizing column names and ensuring consistency across the dataset.

5. **Exploratory Data Analysis (EDA)**
    - Key attributes, such as age, are analyzed to understand their distribution.
    - Visualizations (e.g., histograms) are created using matplotlib to represent the data graphically.
6. **Interpretation and Insights**
    - The results from the visualizations and summary statistics are interpreted to extract meaningful insights about the workshop participants.

RESULTS

1. **Data Cleaning and Preparation**
    - All columns were renamed for clarity (e.g., "Age", "Weight", "Height", etc.).
    - The "Age" column was converted to numeric, and only valid ages (0–100) were retained.
    - All rows with missing values were removed, resulting in a clean dataset for analysis.
2. **Data Overview**
    - The cleaned dataset contains only complete and valid records, ensuring reliable analysis.
    - The structure and columns of the dataset were confirmed using df.info() and df.columns.
3. **Age Distribution Analysis**
    - A histogram was plotted to visualize the distribution of participant ages.
    - The histogram provides a clear view of the frequency of different age groups among workshop participants.
4. **Missing Data**
    - After cleaning, the dataset had no missing values, as confirmed by df.isnull().sum().

CONCLUSIONS AND KEY LEARNINGS

CONCLUSIONS

This project successfully demonstrated the process of exploratory data analysis (EDA) on real-world survey data from a Data Science Workshop. Through systematic data cleaning, transformation, and visualization, the dataset was prepared for meaningful analysis. The age distribution and other participant characteristics were explored, providing valuable insights into the workshop attendees.

KEY LEARNINGS

1. **Data Cleaning is Essential:**
   Renaming columns, handling missing values, and correcting data types are crucial steps to ensure data quality and reliability.

2. **Data Transformation Improves Usability:**
   Standardizing and filtering data makes analysis easier and results more accurate.

3. **Visualization Aids Understanding:**
   Graphical representations like histograms help in quickly identifying patterns and trends in the data.

4. **Practical Use of Python Libraries:**
   Hands-on experience with pandas, numpy, matplotlib, and seaborn is invaluable for any data analyst.

5. **Insight Generation:**
   EDA helps uncover important information about participants, which can inform future workshops and decision-making.