CPSC 4300/6300: Applied Data Science

Exploratory Data Analysis and Visualization Report

Name: Shubham Narale
Course: CPSC 4300/6300
Instructor: Nina Hubig
Spring 2024

Problem Statement

The goal of this project was to use this program to conduct Exploratory Data Analysis aka EDA on datasets usuch as the the titanic passenger dataset as well as a dataset on heart disease. The primary objective was to explore the data and what the analysts looked for were general trends which are then captured by plots and statistical methods. Tasks included managing missing values, understanding categorical and continuous data and visualizing feature correlation with Python data science libraries.

Tools and Libraries Used

- **Programming Language**: Python
- **Libraries**: Matplotlib, Seaborn, Pandas, Numpy, Scipy, Scikit-learn

Key Steps and Methodology

1. Data Loading and Preprocessing:

- Loaded datasets such as the Titanic dataset and heart disease dataset.
- Performed initial inspection of the data using .info() and .describe() to understand the data types, missing values, and summary statistics.
- Handled missing values by identifying and counting null entries (e.g., missing deck information in Titanic data).
- Dropped unnecessary features like embarked, alive, and embark_town in the Titanic dataset for more focused analysis.

2. Univariate Analysis:

- **Histograms**: Visualized the distribution of continuous variables like age and fare using Seaborn's distplot().
- **Boxplots**: Used boxplots to highlight the distribution and outliers for age across different classes in the Titanic dataset.
- **Kernel Density Estimates (KDE)**: Plotted smoothed distributions to visualize the density of variables like age and fare.

3. Bivariate and Multivariate Analysis:

- **Scatter Plots**: Explored relationships between variables such as age and fare and incorporated a third variable (e.g., survival status) using color encoding.
- **Boxplots by Group**: Compared the age distributions across ticket classes to infer trends such as whether younger passengers were more likely to be in first class.

- **Pie Charts**: Visualized proportions of categorical variables like gender to explore the percentage of men and women seeking cardiological exams.
- **Pairplots**: Generated pairplots for multiple variables to explore correlations and patterns within subsets of the data.

4. Categorical Analysis:

- Created count plots to analyze the frequency of categories in features like class and deck.
- Used dummy variables to convert categorical data into numerical representations, enabling further statistical and machine learning applications.

5. Additional Exercises:

- Addressed questions about age distributions of Titanic passengers and trends in cardiology visits based on age, gender, and resting blood pressure.
- For the heart disease dataset, replaced the response variable with a binary classification to simplify the analysis (e.g., num converted to hd).

Results and Insights

- **Titanic Dataset**:
    - Age distribution indicated a concentration of younger passengers, with noticeable patterns for specific classes.
    - First-class passengers tended to be older compared to second and third-class passengers.
    - Survival rates varied significantly based on class and gender.
- **Heart Disease Dataset**:
    - Men were more likely to seek cardiological exams compared to women.
    - Resting blood pressure showed a positive correlation with age.
    - Conversion of categorical features to numerical dummies expanded the feature space to 18 dimensions, enabling detailed analysis.

Challenges and Solutions

- **Missing Data**: Addressed missing values by visualizing and dropping irrelevant features, ensuring a clean dataset for analysis.
- **Feature Engineering**: Converted categorical variables into dummy variables to enhance interpretability and usability for further modeling tasks.
- **Visualization Choices**: Selected appropriate visualizations (e.g., histograms for distributions, scatter plots for relationships) to ensure clarity and meaningful insights.

Conclusion

This project demonstrated the ability to apply advanced EDA and visualization techniques to uncover patterns and insights in real-world datasets. By integrating tools like Seaborn and

Matplotlib, I was able to create clear and informative plots that enhanced the understanding of the data. These techniques will serve as a foundation for building predictive models and further data-driven decision-making.