PROJECT TITLE

Healthcare Data Analysis for Disease Risk Prediction and Patient Outcome Improvement

Internship Role: Data Analyst
Name: Keerti Neeramanigar
Week: Week 3 – Exploratory Data Analysis (EDA) and Visualization

Project Title

Healthcare Data Analysis for Disease Risk Prediction and Patient Outcome Improvement

---

1. Introduction

Exploratory Data Analysis (EDA) is an important step in healthcare data analytics. It helps in understanding the data, finding patterns, and identifying useful insights before applying advanced analysis. In healthcare projects, EDA helps doctors and decision-makers understand patient trends, disease patterns, and treatment outcomes.

This document explains how EDA and visualization would be performed on a hypothetical healthcare dataset using Python tools. The focus is on using simple statistical summaries and visual charts to tell a clear data story.

2. Data Understanding

The first step in EDA is to understand the healthcare dataset.

Steps include:

- Viewing number of rows and columns
- Understanding column names
- Checking data types (numerical or categorical)
- Identifying missing values

Healthcare datasets usually contain patient age, gender, disease type, test results, and treatment outcomes.

3. Identification of Key Variables

Key variables in a healthcare dataset may include:

- Age (numerical)

- Gender (categorical)
- Disease type (categorical)
- Blood pressure or sugar level (numerical)
- Hospital stay duration (numerical)
- Recovery status (categorical)

Identifying these variables helps in choosing the correct analysis and visualization.

4. Descriptive Statistics

Descriptive statistics summarize the data in a simple way.

Common statistics used:

- Mean
- Median
- Minimum and maximum values
- Standard deviation

For example:

- Average patient age
- Average blood sugar level
- Range of hospital stay days

These summaries help in understanding the overall data distribution.

5. Exploratory Data Analysis Techniques

The following EDA techniques would be used:

- Frequency counts for categorical data
- Distribution analysis for numerical data

- Correlation analysis between variables
- Outlier detection using box plots

These techniques help in identifying trends and unusual values.

6. Visualization Strategy

Visualization plays a very important role in healthcare analytics. Charts make data easy to understand.

Planned visualizations include:

- Bar charts for categorical variables
- Histograms for numerical distributions
- Box plots to detect outliers
- Scatter plots to show relationships
- Correlation heatmaps to identify strong relations

Each visualization is chosen to answer a specific healthcare question.

7. Simulated Visualizations (Description)

Examples of simulated charts:

- **Histogram:** Shows age distribution of patients
- **Bar Chart:** Shows number of patients by disease type
- **Box Plot:** Shows variation in blood pressure values
- **Scatter Plot:** Shows relation between age and sugar level
- **Heatmap:** Shows correlation between health indicators

These visuals help in identifying patterns and trends clearly.

8. Insight Generation and Healthcare Impact

Insights gained from EDA may include:

- Which age group is more affected by a disease
- Relationship between lifestyle indicators and disease risk
- Patients with longer hospital stays
- Common health risk factors

These insights help hospitals improve treatment planning and preventive care.

9. Tools and Libraries Used

Python libraries used for EDA and visualization include:

- **Pandas** – data handling and summaries
- **Matplotlib** – basic plotting
- **Seaborn** – advanced and attractive visualizations

These tools help in creating clear and meaningful charts.

10. Conclusion

This document explains a structured approach to exploratory data analysis and visualization for healthcare data. By using descriptive statistics and visual tools, important patterns and trends can be identified. Proper EDA helps in making data-driven healthcare decisions and improves understanding of patient data. This methodology can be applied to real-world healthcare datasets to support better clinical and business decisions.