B. Ektha Charishma

Healthcare Data Analysis Report

1. Dataset Description

1.1 Source:

The dataset used for this project is hospital data analysis.csv, containing information about 1,000 patients. It captures key details required for healthcare data analysis.

1.2 Columns:

- Patient_ID: Unique identifier for each patient
- Age: Age of the patient
- Gender: Male or Female
- Blood_Pressure: Blood pressure levels
- Cholesterol: Cholesterol levels
- Diagnosis: Diagnosed disease or condition
- Treatment_Cost: Cost of the treatment
- Length_of_Stay: Number of days admitted
- Outcome: Result of treatment (Recovered / Not Recovered / Deceased)

1.3 Data Quality:

- The dataset is mostly clean and structured.
- Some missing values were handled during preprocessing.
- Data has good variation across age groups, diseases, and outcomes.

2. Operations Performed

2.1 Data Loading and Exploration

- Created a Spark session and loaded the dataset.
- Verified the schema and displayed the first few rows using .show().
- Counted total rows and confirmed ~1000 records.

2.2 Data Inspection

- Selected important columns such as Age, Diagnosis, Treatment_Cost, and Outcome.
- Checked descriptive statistics and identified distribution patterns.

2.3 Filtering Operations

- Filtered patients with treatment cost greater than 50,000 to identify high-cost cases.
- Filtered patients with long hospital stays (>10 days) to study severe cases.

2.4 Aggregation and Summary Statistics

- Calculated average, minimum, and maximum treatment cost using PySpark aggregate functions.
- Computed average length of stay by diagnosis to find critical diseases.
- Calculated recovery rate by disease category.

3. Key Insights

3.1 Disease Trends

- Certain diseases showed high admission rates, indicating common health issues.
- Some chronic diseases required longer stays and higher treatment costs.

3.2 Cost Analysis

- Average treatment costs varied by diagnosis and patient age.
- Chronic diseases like diabetes and heart conditions showed higher average costs.

3.3 Outcome Patterns

- Recovered patients usually had shorter hospital stays.
- Non-recovered or deceased patients had higher costs and longer treatment durations.

3.4 Demographic Insights

- Most patients were aged between 40–60 years.
- Gender distribution was balanced across most disease types.

4. Recommendations

4.1 Operational Efficiency

- Monitor high-cost treatments and explore cost optimization strategies.
- Improve early diagnosis programs for chronic diseases to reduce long hospital stays.

4.2 Patient Care Strategy

- Focus on preventive healthcare for the 40–60 age group.

- Introduce awareness campaigns for diseases with high admission rates.

4.3 Future Analytics Opportunities

- Build predictive models to forecast patient recovery chances and expected stay duration.
- Use clustering to group patients by risk levels for targeted healthcare interventions.

5. Conclusion

The analysis of the hospital dataset revealed critical healthcare patterns related to patient demographics, treatment costs, and recovery outcomes. Diseases with higher costs and longer hospital stays indicate areas for improvement in preventive care and treatment efficiency. The dataset provided a strong foundation for understanding hospital performance and planning data-driven strategies to enhance patient care and resource utilization.