Exploratory Data Analysis (EDA) with Pandas in Health Care

The purpose of this project is to explore and analyse a healthcare dataset using the Pandas framework to derive insights into patient behavior, treatment trends, and hospital performance.

Goals of the Project

- Explore the health care dataset using Pandas.
- Perform feature engineering to derive useful insights.
- Visualize data distributions and trends with various plot types.
- Summarize key findings that can aid in business decision-making

Materials and Methods

The data for this project is from a simulated healthcare platform, containing information about patients, treatments, diagnoses, and hospital stays. This dataset includes patient demographics, medical conditions, treatment costs, and more. The analysis aims to understand patient behavior, treatment effectiveness, and hospital performance.

General Part

- **Libraries Import**: Pandas, NumPy, Seaborn, Matplotlib
- **Dataset Exploration**: Initial exploration of the dataset, checking for missing values, duplicates, and generating summary statistics.
- **Feature Engineering**: Transformation of date columns and creation of new features like hospital stay duration and Billing Amount
- **Visualization in Pandas**: Distribution analysis, relationships between variables, and time-based trends.

Project Outcome & Insights

The project performs Exploratory Data Analysis (EDA) on a healthcaredataset to gain meaningful insights into patient behaviour, treatment effectiveness, and hospital performance. Below are the key outcomes:

1. Patient Performance

- **Patient Segment Wise Top Treatments:** The project groups treatments based on different patient segments to identify the most effective treatments**.**
- **Time Series Analysis**: It shows treatment trends over time, helping healthcare providers identify seasonal fluctuations and peak treatment periods.
- **Top Performing Treatments:** Identifies the treatments with the highest success rates and patient satisfaction.

2. Patient Behaviour Analysis

- **Hospital_stay_duration: Days between admission date and discharge date.**
- **Risk Level:** Categorized based on age and medical condition.
- **stay category:** Categorized hospital stays into Same Day, Short, Moderate, and Long.
- **Treatment Profitability Analysis:** The treatment profitability feature segments treatments into categories based on their profitability, allowing better resource management.

3. **Profitability & Healthcare Growth**

- **Profit Margin Analysis:** Helps understand profitability per treatment and identify areas for improving profit margins.
- **Year-over-Year Growth:** Tracks annual growth in treatment costs and patient numbers, enabling better financial planning

Feature Engineering

Created new columns such as:

- **Hospital stay duration:** Days between admission date and discharge
- **admission year, admission month, admission weekday** (Extracted from admission date).
- **Returning patient:** to the healthcare dataset. This column will be a Boolean flag indicating whether a patient is a returning patient (i.e., has made multiple visits).
- **age category** (Binned age into categories: Child, Adult, Middle, Age, Senior).

Key Questions and Insights to be Addressed:

```
total_billing_by_hospital =
df.groupby('Hospital')['Billing Amount'].sum()
 department_categories = pd.qcut(total_billing_by_hospital,
q=4, labels=['Low', 'Medium', 'High', 'Very High'])
 df['hospital_category'] =
df['Hospital'].map(department_categories)
```

Answer: billing percentage

```
25%     35.000000   13243.718641

50%     52.000000   25542.749145

75%     68.000000   37819.858159
```

- Which medication categories have the highest patient?

```
patients_by_medication_category = df.groupby('Medication
Category')['age_group].nunique().sort_values(ascending=Fals
e)

print(patients_by_medication_category)
```

- How to convert date columns to datetime format?

```
date_columns = ['Date of Admission', 'Discharge Date']

for col in date_columns:

    df[col] = pd.to_datetime(df[col], dayfirst=True, errors='coerce')
```

- To analyse the average hospital stay duration and how it varies by medical condition

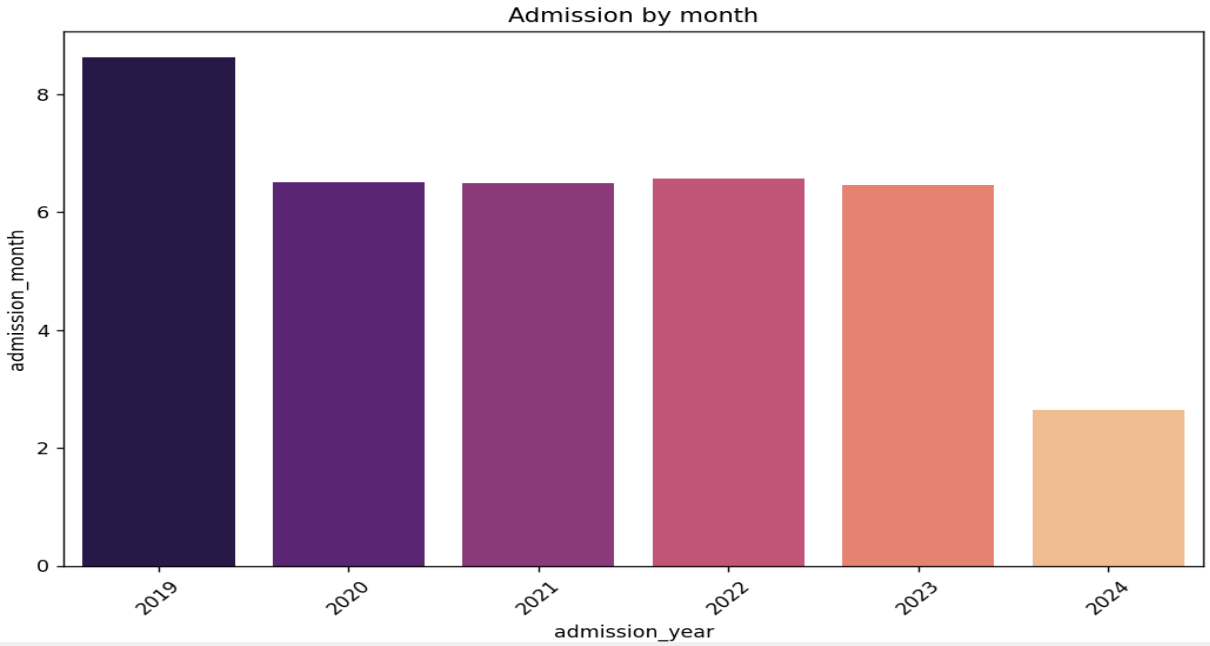
```
 avg_stay_duration_by_medical_condition = df.groupby('medical condition')['hospital_stay_duration'].mean().sort_values(ascending=False)

print (avg_stay_duration_by_ medical_condition)
```

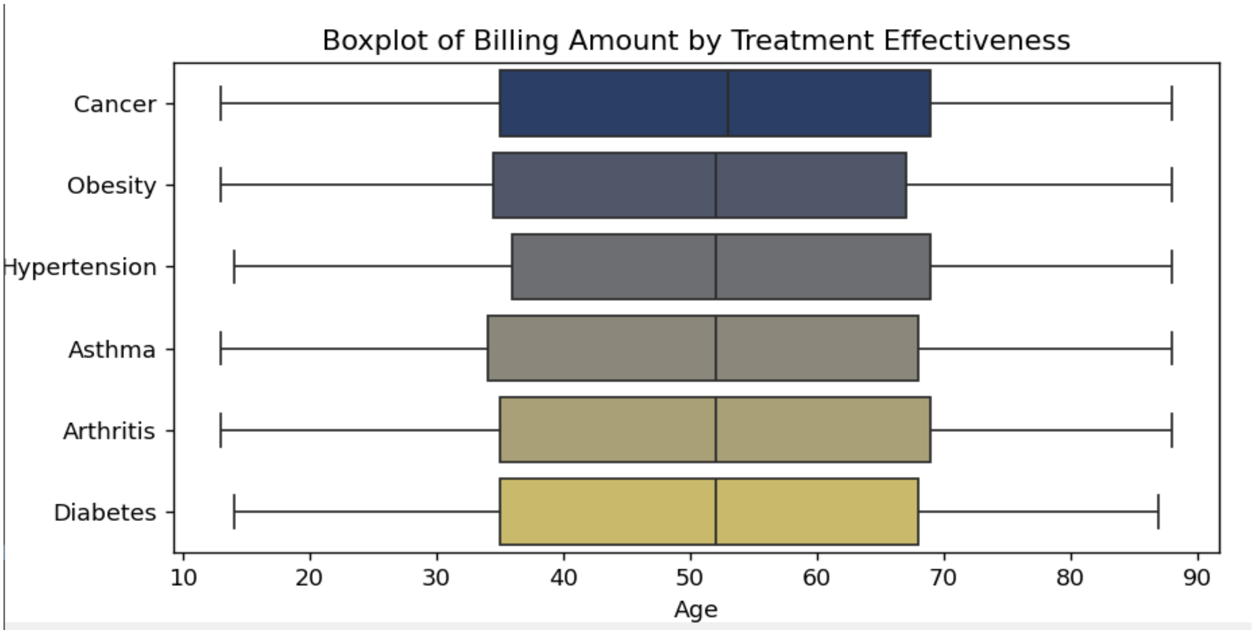
Visualization:

Several charts created to present inside including:

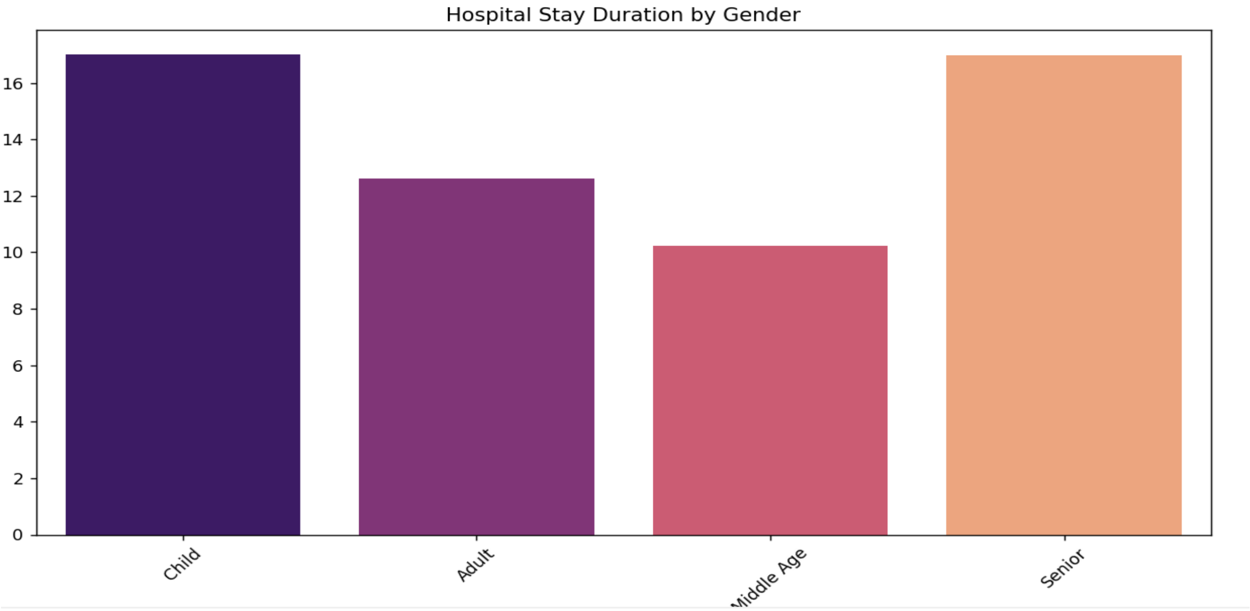
Admission by month and year:



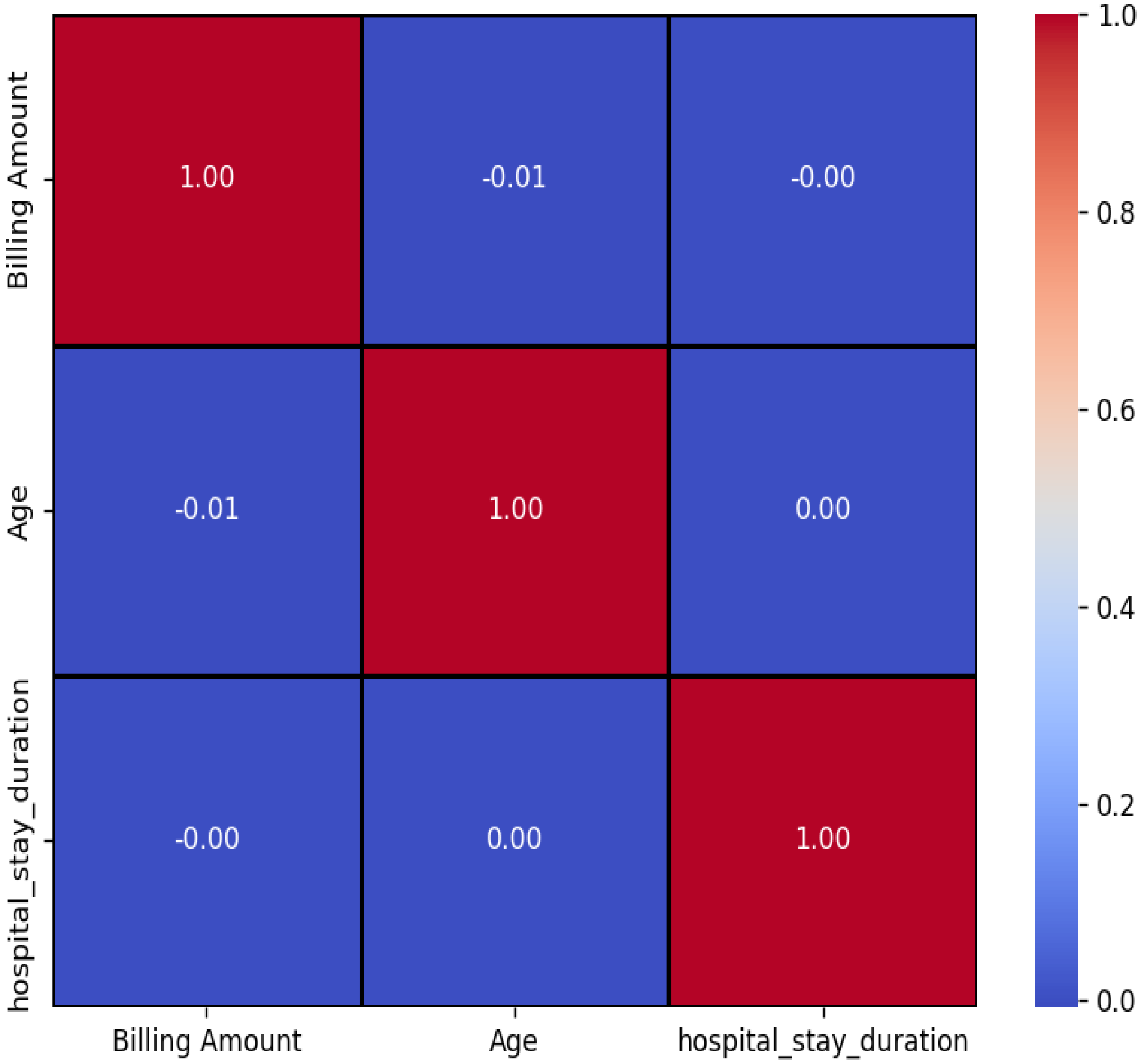
Billing amount by treatment effectiveness:



Stay duration by gender:

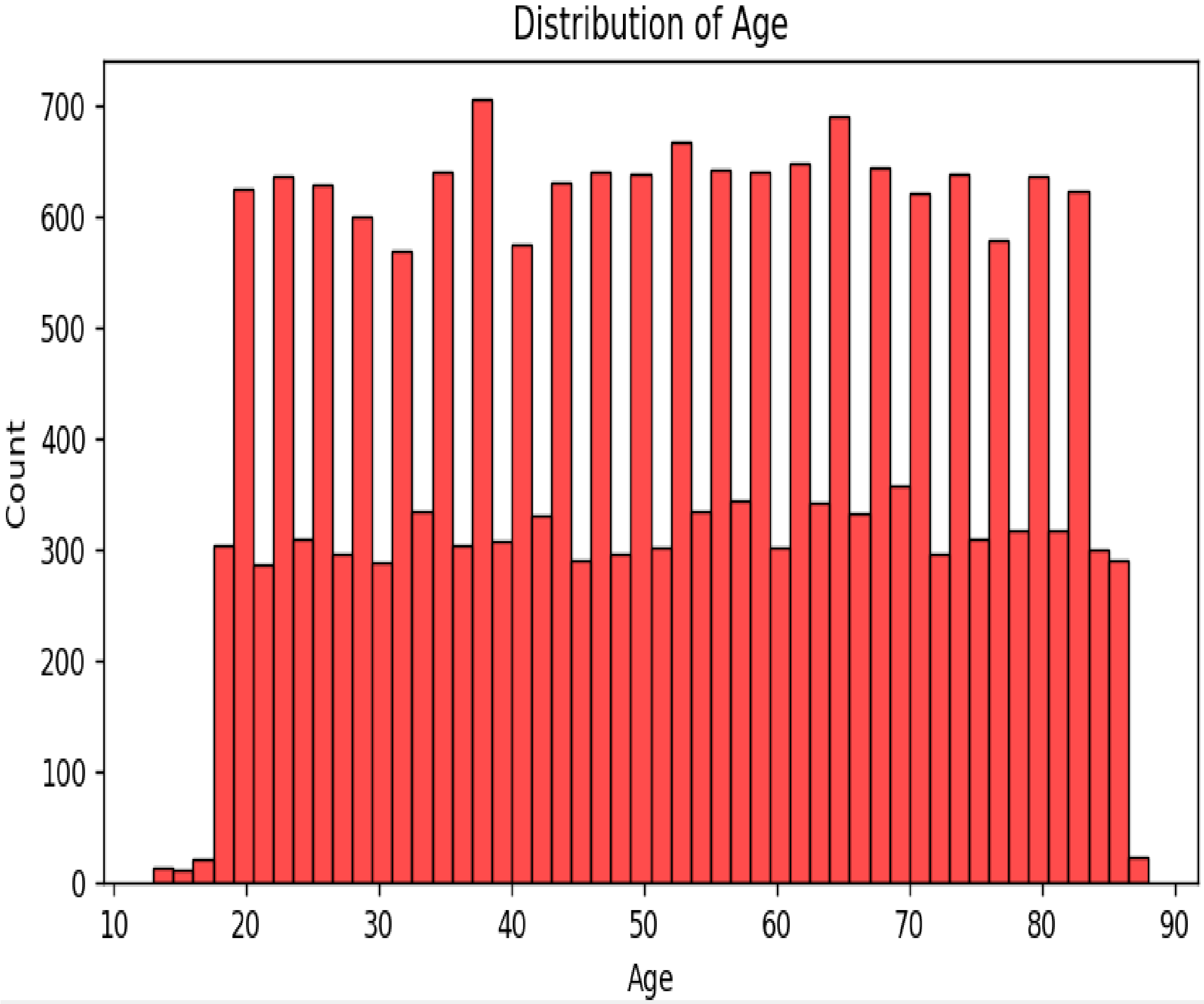


Correlation for healthcare data:



Correlation Matrix for Healthcare Data

Distribution of age:


Distribution of Age
Count
Age