Exploratory Data Analysis (EDA) of Cancer Treatment Outcomes and Patient Demographics using Python

Aim

The aim of this project is to conduct a comprehensive Exploratory Data Analysis (EDA) on a chaotic clinical dataset of cancer patients from the UAE using Python. The primary goal is to clean, standardize, and analyze this raw data to uncover significant patterns and relationships between patient demographics, cancer types, treatment methodologies, and final outcomes. By doing so, the project seeks to identify key factors influencing patient recovery, reveal potential disparities in healthcare access and success rates across different demographics and emirates, and ultimately generate actionable insights that can inform and improve future oncology care strategies and resource allocation within the region.

Phase 1:

Problem Definition

This project addresses the challenge of transforming a raw, unstructured, and "chaotic" clinical dataset into a source of actionable intelligence through a rigorous, end-to-end Exploratory Data Analysis (EDA) process using Python. The core problem is not merely the presence of data quality issues—such as missing values, inconsistent formatting, invalid entries, and potential duplicates—but the consequent inability to perform a reliable analysis to answer critical healthcare questions. The project's success is defined by the ability to demonstrate comprehensive data wrangling skills to clean and preprocess the data, followed by the application of statistical summaries and advanced visualizations to uncover hidden patterns, trends, and relationships within the dataset, ultimately generating a clear and insightful narrative about cancer patient demographics, treatment modalities, and outcomes in the UAE.

Dataset Selection

For this project, I worked with a comprehensive clinical dataset of cancer patients from the UAE:

Dataset: UAE Cancer Patients Clinical Records
Total Rows: 10,000 patients after cleaning
Total Columns: 20 clinical and demographic features

Phase 2:

Data Loading and Initial Overview

● Import the dataset using Pandas and provide an overview:

○ Number of rows and columns ○ Data types of each column

○ Initial observations (e.g., head(), info(), describe())

Data Pre-processing

● Perform all necessary cleaning steps such as:

○ Handling missing values ○ Removing duplicates ○ Correcting data types ○ Creating derived columns ○ Filtering or aggregating data

Tasks Completed In Phase 2

- 🗑️ Removed duplicate rows (5%)
- 🔠 Fixed mixed casing issues
- 📅 Standardized date formats
- 🏷️ Cleaned column names
- 🔢 Fixed numeric values

Phase 3:

Exploratory Data Analysis (EDA)

● Conduct descriptive and exploratory analysis to uncover patterns and trends:

○ Univariate, bivariate, and multivariate analysis

○ Use groupby, pivot tables, and correlation analysis

○ Include statistical summaries to support findings

Visualizations

● Use Matplotlib / Seaborn / Plotly to generate meaningful visualizations:

○ Bar plots, line charts, pie charts, histograms, box plots, scatter plots, heatmaps, etc
○ Ensure visuals should have proper titles, labels, legends, and color schemes

○ Use subplots where applicable for better layout

Phase 3 Findings

Treatment Effectiveness

- Immunotherapy demonstrates superior success (50.3%) compared to other treatments, emerging as the most effective cancer treatment strategy in the dataset

Demographic Patterns

- Young expatriates show exceptional recovery rates (85.7%), suggesting potential advantages in early diagnosis or treatment response among this demographic
- Minimal outcome disparities observed between Emirati (49.0%) and Expatriate (49.8%) patients, indicating equitable healthcare access

Cancer Stage Insights

- Surprisingly consistent recovery rates across all cancer stages (48.7%-50.4%), challenging conventional expectations that later stages would show significantly poorer outcomes

Patient Profile

- Middle-aged patient population with average age of 53.5 years, representing a broad demographic spectrum from young adults (18) to elderly (89)

Smoking Impact

- Clear correlation patterns identified between smoking status and cancer types, with former smokers showing distinct risk profiles worthy of further investigation

Conclusion and Recommendations

The comprehensive analysis reveals that the UAE's oncology care system demonstrates remarkable consistency in treatment outcomes across diverse demographics, with an overall recovery rate of 49.3% showing minimal variation between Emirati (49.0%) and Expatriate (49.8%) patients, indicating equitable healthcare access. The ecosystem is characterized by middle-aged patient predominance (average age 53.5 years) and leukemia as the most prevalent cancer type, while immunotherapy emerges as the most effective treatment modality.