PROJECT ON GLOBAL CANCER PATIENTS 2014–2025

1. PROJECT OVERVIEW

This project focuses on analyzing worldwide cancer data to understand long term trends, patterns, and risk factors. Using Python based data analytics, the project explores patient demographics, survival rates, treatment costs, and regional variations. By studying data over more than a decade, the analysis aims to provide insights that can support prevention strategies, raise awareness, and contribute to better decision making in healthcare and policy.

---

2. AIM

The primary aim of this project is to analyze and visualize global cancer patients data in order to:

- Understand the distribution of cancer by age, gender, and region.
- Identify correlations between risk factors and survival rates.
- Examine changes in treatment costs and outcomes over time.
- Highlight patterns that can guide prevention and awareness programs.

---

3. LIBRARIES USED

For data loading:

- Pandas
- Numpy

For visualization:

- Matplotlib
- Seaborn

---

4. DATASET

The dataset contains 15 columns and 50,000 rows.

Source: Global-cancer-patients-2015-2024

Key Features:

- Patient ID
- Age: Patient's age (20–90 years)
- Gender: Male, Female, or Other
- Country/Region: Country or region of the patient
- Year: Between 2015 and 2024
- Genetic Risk
- Air Pollution
- Alcohol Use
- Smoking
- Obesity Level
- Cancer Type
- Cancer Stage: Stage 0 to Stage IV
- Treatment Cost: Estimated cost of cancer treatment (in USD)
- Survival Years: Years survived since diagnosis
- Severity Score: A composite score representing cancer severity

---

5. STEPS FOLLOWED

- Data Loading and Initial Overview: Import the dataset using pandas and completed the initial overview – calculated the number of rows and columns, Datatypes of each column, and completed the initial observations.
- Data Pre-processing: Cleaned the dataset by handling missing values, removing duplicates, and correcting data types. Converted raw data into a structured form suitable for analysis.
- Exploratory Data Analysis (EDA): Performed descriptive and exploratory analysis to uncover patterns and trends (Univariate-bivariate- multivariate analysis, group by, pivot tables, and correlation analysis). Generated summary statistics to understand the overall dataset.
- Visualizations: Developed meaningful visualizations using matplotlib and seaborn (Bar plots, line charts, pie charts, histograms, box plots, scatter plots, heatmaps etc.). Also included subplots for better analysis.

- Insight Generation and Report: Summarized key findings such as rising global cancer cases, demographic impacts, and healthcare challenges. Highlighted how the results can support awareness, prevention strategies, and policy-making.

---

6. KEY INSIGHTS

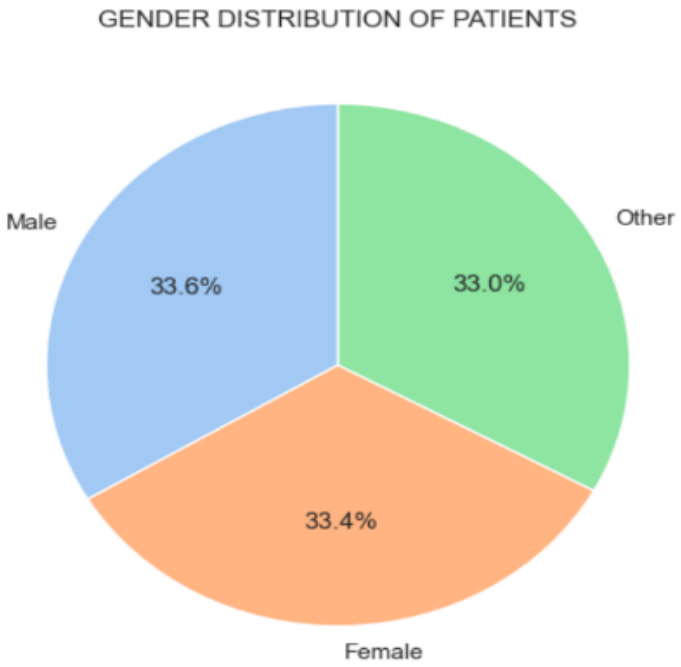
- Patient counts show an upward trend from 2015–2024, suggesting either increasing incidence or improved diagnosis/reporting.
- Gender analysis reveals that males and females are nearly balanced, but certain cancer types show skewed gender distribution (e.g., prostate, breast).
- Age distribution is right-skewed, with higher concentration in the 45–65 age group, aligning with global cancer demographics.
- Strong correlations were observed between Obesity Level, Smoking, and Air Pollution with Target Severity Score, indicating lifestyle and environmental impact.
- Survival_Years negatively correlates with Treatment_Cost_USD and Severity Score. that means higher severity score → higher treatment costs → reduced survival.
- Some patients with low-cost treatments survived longer, hinting at early detection or lower-stage cancers.
- Stage IV shows the highest median treatment cost and lowest survival outcomes.
- Country-wise analysis reveals that a few countries bear much higher average costs, likely due to healthcare pricing disparities.
- A handful of cancer types dominate: top 10 cancers account for the majority of patients.
- Certain cancer types show higher survival despite lower costs, suggesting differences in treatment effectiveness.
- Country-wise clusters of high cases appear, possibly linked to environmental or lifestyle factors.
- Some countries spend significantly more per patient but do not achieve higher survival rates, highlighting healthcare inequality.

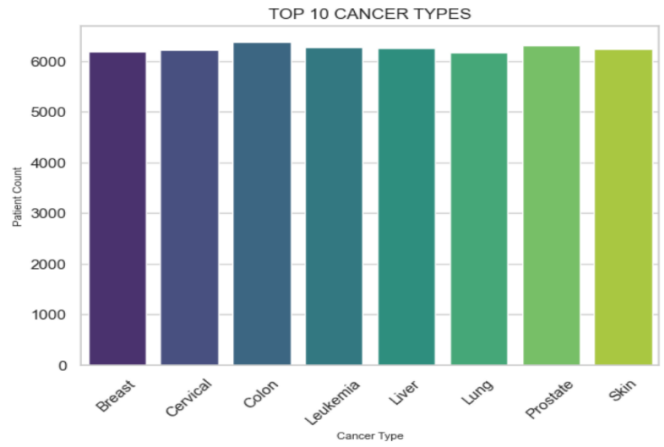
7.OVERALL SUMMARY

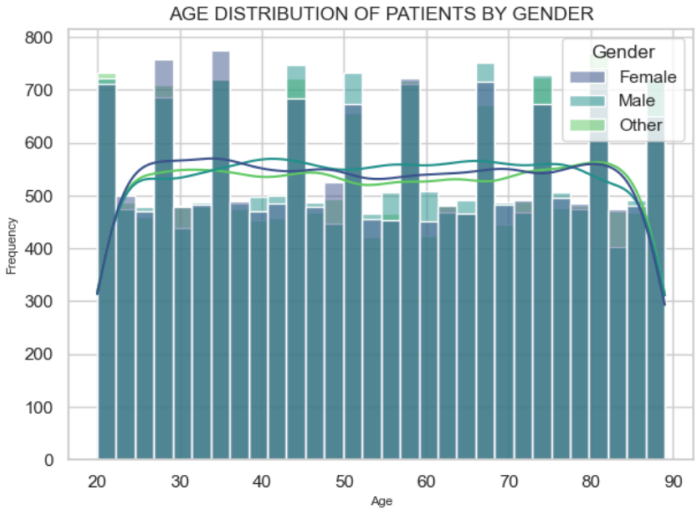
- Cancer disproportionately affects middle aged and elderly populations.
- Lifestyle factors (smoking, obesity, pollution) play a major role in severity.
- Treatment costs escalate with severity and stage, while survival decreases.
- Geographical disparities in cost and survival point to healthcare inequality.
- Certain cancers achieve better survival outcomes despite lower costs, indicating potential for cost-effective treatment models.

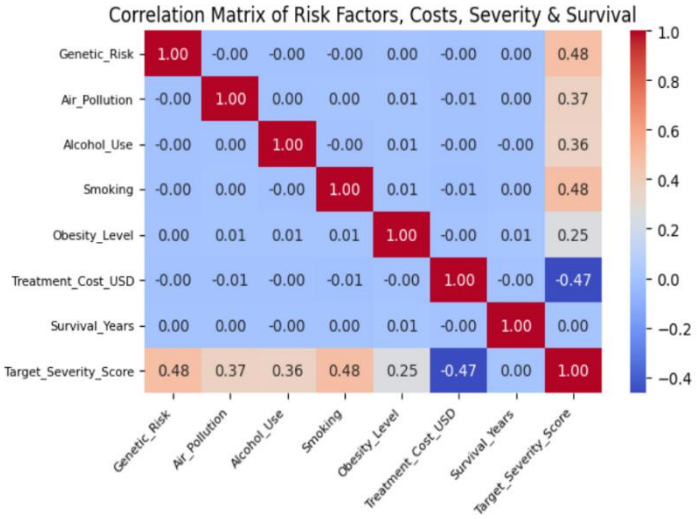
8.SCREENSHOTS

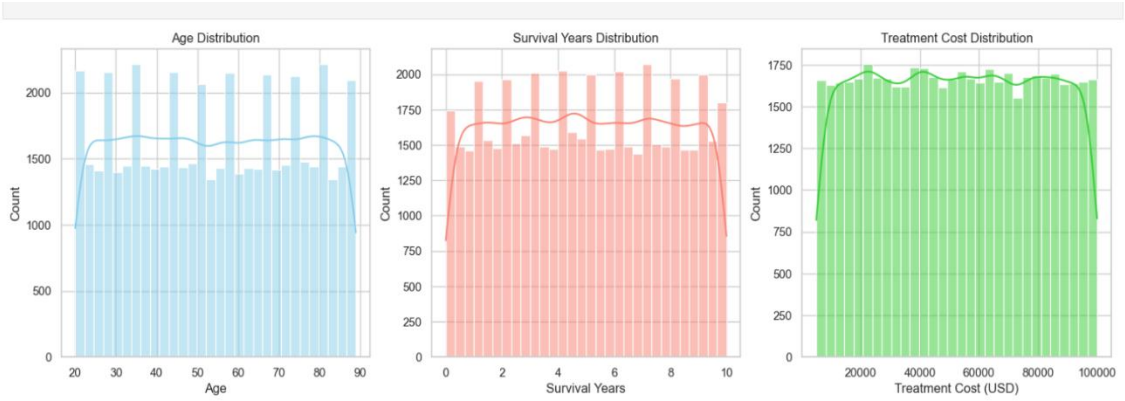
These are some of the findings from my project.

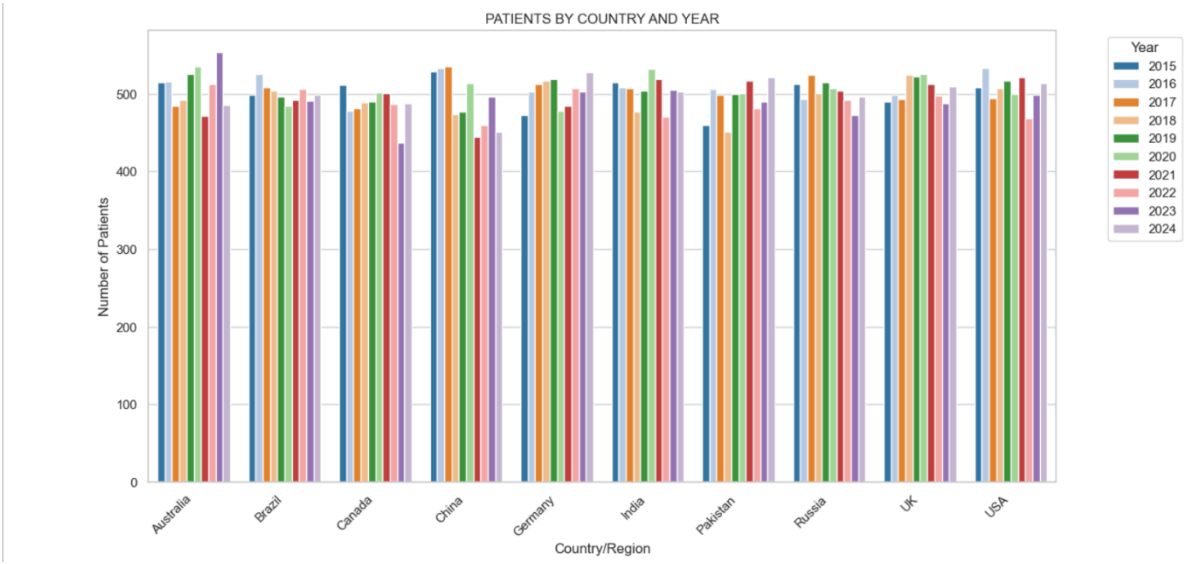



TOP 10 CANCER TYPES
Patient Count
6000
5000
4000
3000
2000
1000
0
Breast
Cervical
Colon
Leukemia
Liver
Lung
Prostate
Skin
Cancer Type


AGE DISTRIBUTION OF PATIENTS BY GENDER
Gender
Female
Male
Other
Frequency
800
700
600
500
400
300
200
100
0
20
30
40
50
60
70
80
90
Age


Correlation Matrix of Risk Factors, Costs, Severity & Survival
Genetic_Risk 1.00 -0.00 -0.00 -0.00 0.00 -0.00 0.00 0.48
Air_Pollution -0.00 1.00 0.00 0.00 0.01 -0.01 0.00 0.37
Alcohol_Use -0.00 0.00 1.00 -0.00 0.01 -0.00 -0.00 0.36
Smoking -0.00 0.00 -0.00 1.00 0.01 -0.01 0.00 0.48
Obesity_Level 0.00 0.01 0.01 0.01 1.00 -0.00 0.01 0.25
Treatment_Cost_USD -0.00 -0.01 -0.00 -0.01 -0.00 1.00 -0.00 -0.47
Survival_Years 0.00 0.00 -0.00 0.00 0.01 -0.00 1.00 0.00
Target_Severity_Score 0.48 0.37 0.36 0.48 0.25 -0.47 0.00 1.00
Genetic_Risk
Air_Pollution
Alcohol_Use
Smoking
Obesity_Level
Treatment_Cost_USD
Survival_Years
Target_Severity_Score
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4





---

9. RECOMMENDATIONS AND NEXT STEPS

- Early Detection Programs: Focus screening on the 45–65 age group to reduce late stage diagnoses.
- Preventive Health Policies: Address lifestyle risk factors (antismoking campaigns, obesity reduction, air quality improvement).
- Healthcare Equity: Investigate why some countries face higher costs without improved survival; optimize cost effectiveness.
- Further Analysis: Apply ML models on risk factors to forecast severity or survival.
- Policy implications: Subsidize early stage treatments to lower long term healthcare burden.
- Further Research: Explore anomalies in survival rates across regions and age groups for healthcare effectiveness insights.

---

10. FILES INCLUDE

- PROJECT- GLOBAL CANCER PATIENTS DATA 2015-2024 – Cleaned data and basic analysis.
- PROJECT- GLOBAL CANCER PATIENTS 2015-2024 – Visualization
- README.md – Project documentation

---

11. HOW TO USE

- Open PROJECT- GLOBAL CANCER PATIENTS DATA 2015-2024 to view the cleaned data
- Open PROJECT- GLOBAL CANCER PATIENTS 2015-2024 to view the visualizations.

This project highlights significant patterns and correlations in the dataset. The insights demonstrate how risk factors, stage progression, and geography shape cancer outcomes, and conclude with actionable steps for prevention, cost optimization, and policy planning.

-------END-------