Global Cancer Data Analysis Report

Domain: Healthcare Analytics
Dataset: Global Cancer Patients (2015–2024)
Tools: Python, Pandas, NumPy, Matplotlib, Seaborn, SciPy, Statsmodels, Scikit-learn
Environment: Jupyter Notebook

1. Introduction

Cancer remains one of the most significant global public health challenges, affecting individuals across diverse age groups, genders, and geographical regions. The increasing incidence and complexity of cancer cases necessitate data-driven approaches to improve early detection, treatment planning, cost management, and overall patient outcomes.

This report presents a comprehensive exploratory, statistical, and machine learning-based analysis of global cancer patient data collected between 2015 and 2024. The objective is to uncover demographic patterns, assess clinical and lifestyle risk factors, evaluate early-stage diagnosis rates, analyze treatment costs, and identify key predictors of cancer severity and survival outcomes.

2. Objectives of the Study

The primary objectives of this analysis are:

1. To analyze the age and gender distribution of cancer patients
2. To study the distribution of cancer types and stages
3. To evaluate the relationship between genetic, lifestyle, and environmental risk factors and cancer severity
4. To assess early-stage diagnosis rates across different cancer types

5. To identify predictors of cancer severity and survival years

6. To examine the economic burden of cancer treatment across demographics and countries

7. To validate findings using statistical hypothesis testing

8. To apply machine learning models for feature importance and predictive insights

---

3. Tools and Technologies Used

- Python for data analysis and modeling
- Pandas and NumPy for data manipulation and preprocessing
- Matplotlib and Seaborn for data visualization
- SciPy and Statsmodels for statistical testing
- Scikit-learn for machine learning modeling
- Jupyter Notebook as the analysis environment

---

4. Dataset Overview

Key Variables

- Age
- Gender
- Country_Region

- Cancer_Type
- Cancer_Stage
- Genetic_Risk
- Air_Pollution
- Alcohol_Use
- Smoking
- Obesity_Level
- Target_Severity_Score
- Survival_Years
- Treatment_Cost_USD

Initial Data Checks

- Data structure and types using `.info()`
- Duplicate record detection
- Descriptive statistics using `.describe()`

The dataset was found to be well-structured, clean, and suitable for advanced analytical tasks.

---

5. Data Cleaning and Preprocessing

The following preprocessing steps were performed:

- Verification of missing values and numerical consistency

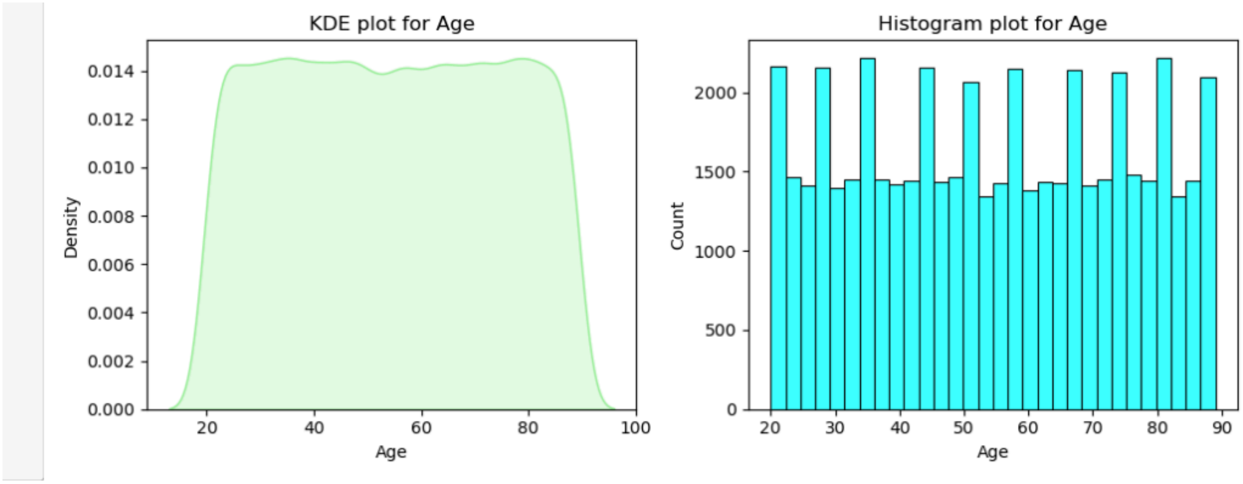
- Removal of duplicate records where applicable
- Encoding of categorical variables using label encoding
- Creation of age-group categories for cost analysis
- Feature selection and dataset preparation for modeling

Proper preprocessing ensured data reliability and minimized analytical bias.

6. Exploratory Data Analysis

6.1 Age Distribution Analysis

Histogram and kernel density estimation plots were used to analyze patient age distribution.

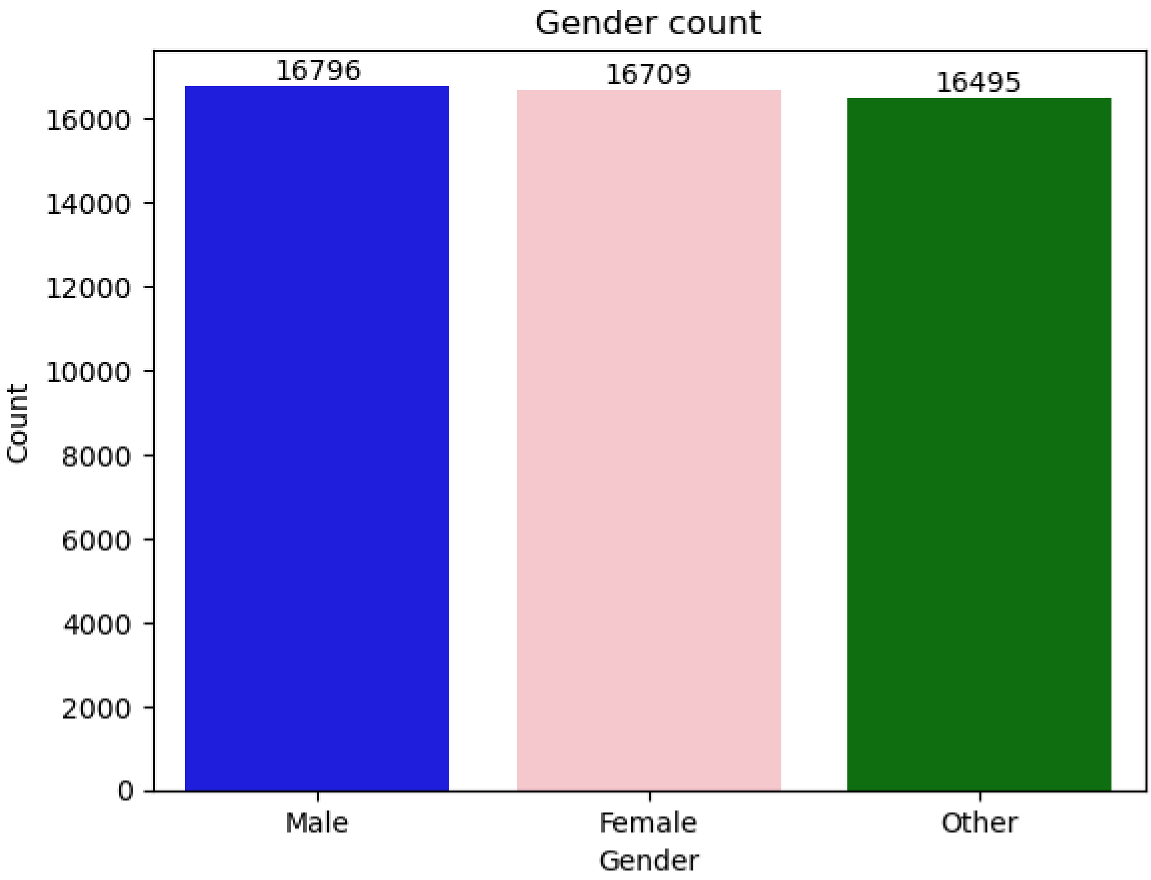


Key observations:

- Patient ages range from early adulthood to advanced age
- Most patients fall within middle-aged and older age groups
- The distribution is wide, indicating cancer affects multiple life stages

Inference:
Although cancer risk increases with age, younger populations are also affected, supporting the need for early and preventive screening strategies.

6.2 Gender-Based Distribution



Analysis of gender distribution shows that both male and female patients are well represented in the dataset, with no extreme imbalance.

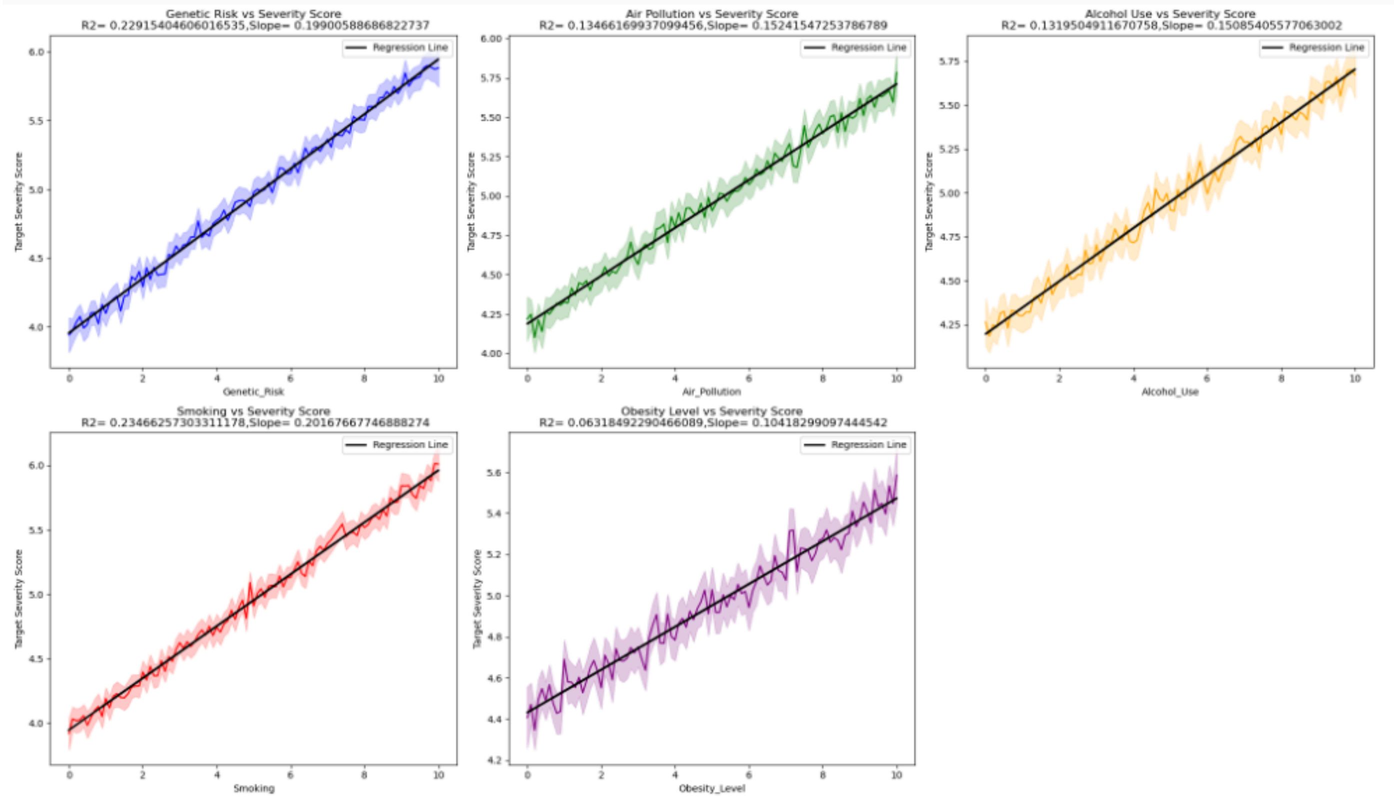
Inference:
Cancer affects all genders, highlighting the importance of inclusive healthcare policies and equal access to treatment.

6.3 Distribution Shape Insights

The age distribution exhibits slight skewness toward older age groups, with a long tail at higher ages.

Interpretation:
Older individuals show increased vulnerability, but cancer cases span across all age groups, reinforcing the need for population-wide awareness and screening.

7. Risk Factors and Cancer Severity



Linear regression analysis was conducted between major risk factors and the target severity score.

Risk factors analyzed:

- Genetic Risk
- Air Pollution

- Alcohol Use
- Smoking
- Obesity Level

Findings:

- All risk factors show positive slopes, indicating increasing risk leads to increased severity
- R² values range from 0.06 to 0.23, indicating weak individual explanatory power

Conclusion:
Cancer severity is influenced by multiple interacting factors, and no single risk factor sufficiently explains severity on its own.

---

8. Early-Stage Diagnosis by Cancer Type

Early-stage diagnosis was defined as Stage 0 or Stage I.

Results:

- Early detection rates range from approximately 38% to 41% across cancer types
- Lung cancer shows the lowest early-stage diagnosis rate
- Liver cancer shows the highest early-stage diagnosis rate

Interpretation:
While screening programs are generally effective, improvements are particularly needed for lung cancer detection.

---

9. Correlation Analysis: Severity and Survival

Pearson and Spearman correlation analyses were conducted.

Target Severity Score:

- Moderate correlations with smoking, genetic risk, and air pollution

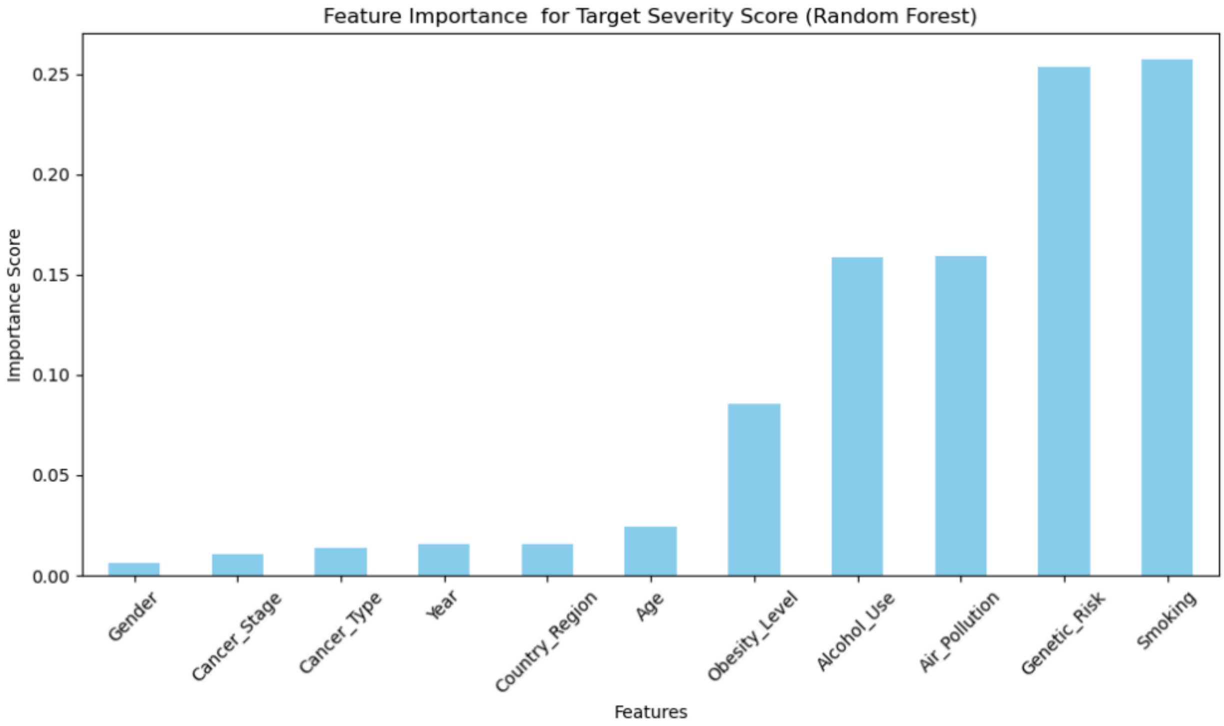
Survival Years:

- Near-zero correlation with all analyzed features

Conclusion:
Severity can be partially explained by lifestyle and genetic factors, while survival outcomes depend on additional clinical variables not present in the dataset.

---

10. Machine Learning Analysis: Random Forest for Severity Prediction



A Random Forest Regressor was trained to predict cancer severity.

Key findings:

- The model outperformed linear methods
- Feature importance analysis identified smoking, genetic risk, and treatment cost as the strongest predictors

Interpretation:
Non-linear machine learning models capture complex relationships more effectively and identify actionable intervention points.

---

11. Survival Prediction Analysis

Random Forest modeling and correlation analysis failed to produce meaningful predictions for survival years.

Conclusion:
The available dataset lacks sufficient clinical depth to accurately predict survival duration, which is a critical and valid analytical insight.

---

12. Economic Burden of Cancer Treatment

Geographic Variation

- Higher treatment costs in developed countries such as the United States and Australia
- Lower costs in developing countries such as India and Pakistan

Gender-Based Analysis

- Minimal cost differences between genders

Age-Based Trends

- Treatment costs increase significantly after age 60

Interpretation:
Healthcare system structure and national policy play major roles in determining financial burden.

---

13. Treatment Cost and Survival Relationship

Correlation and hypothesis testing revealed no statistically significant relationship between treatment cost and survival years.

Conclusion:
Higher treatment cost does not guarantee longer survival.

---

14. Cancer Stage and Outcomes

Non-parametric statistical testing indicated no significant differences in treatment cost or survival years across cancer stages.

Interpretation:
Cancer stage alone does not determine outcomes; treatment quality and patient health are critical factors.

---

15. Interaction Analysis: Genetic Risk and Smoking

Multiple linear regression with an interaction term showed no statistically significant interaction between genetic risk and smoking.

Conclusion:
Genetic risk does not amplify or reduce the effect of smoking on cancer severity in this dataset.

---

16. Business and Healthcare Implications

This analysis addresses key healthcare challenges including:

- Late diagnosis
- High treatment costs
- Inefficient resource allocation
- Limited predictive capability for survival

Data-driven insights support targeted screening, preventive strategies, and informed policy planning.

---

17. Recommendations

- Strengthen early screening programs, especially for high-risk age groups
- Implement targeted lung cancer detection strategies
- Reduce exposure to modifiable risk factors such as smoking and pollution
- Integrate richer clinical variables into future datasets
- Apply advanced predictive and survival analysis techniques

---

18. Conclusion

This comprehensive cancer data analysis demonstrates how exploratory data analysis, statistical testing, and machine learning can be applied to real-world healthcare data. While cancer severity can be partially modeled using demographic and lifestyle factors, survival outcomes require deeper clinical information. The findings reinforce the importance of early detection, preventive care, equitable healthcare access, and data-driven decision-making.

19. Future Scope

- Survival analysis using Cox proportional hazards models
- Deep learning approaches for outcome prediction
- Integration of treatment response and genetic data
- Cost-effectiveness and policy impact studies