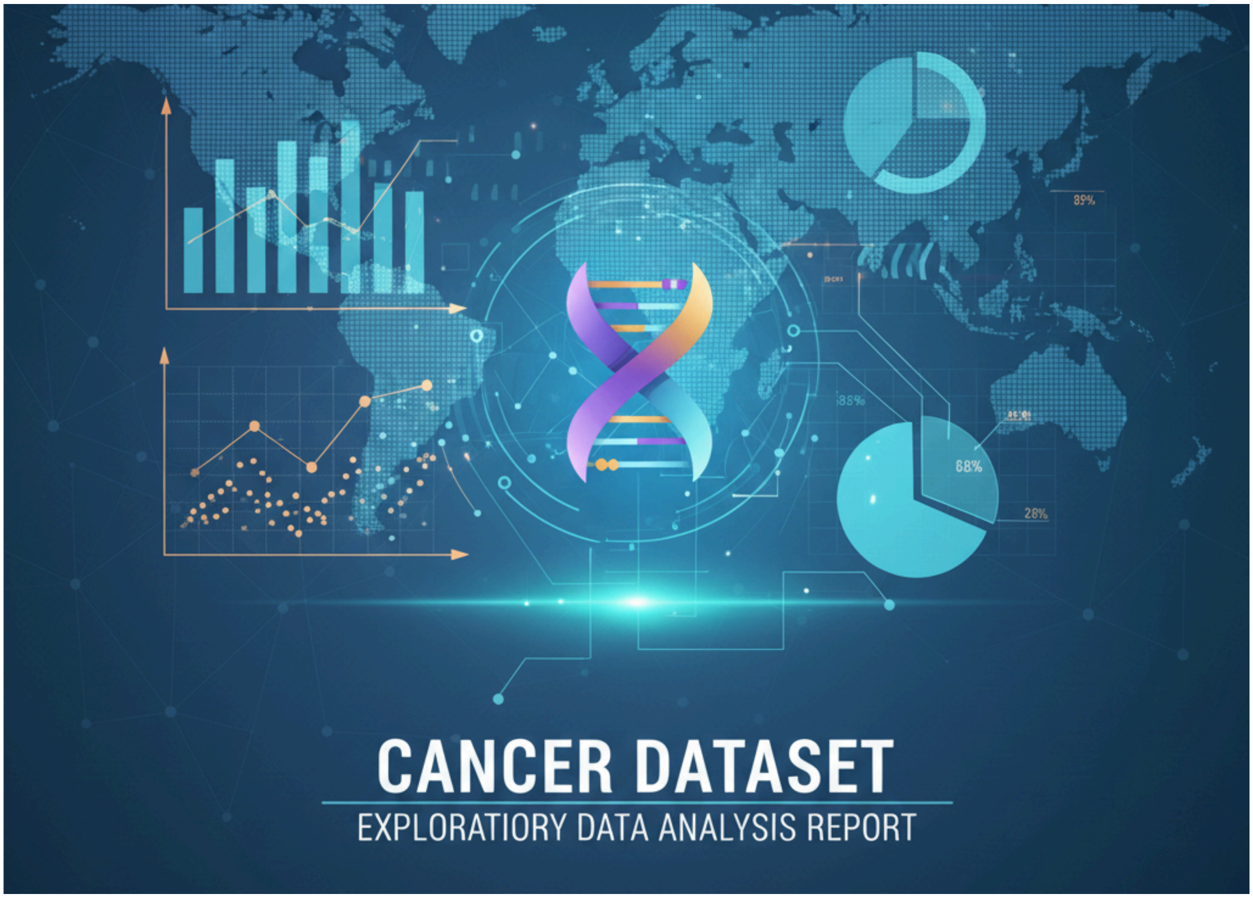
Comprehensive EDA on Global Cancer Patient Dataset

An Exploratory and Inferential Data Analysis of Global Cancer Patient Data, Examining Risk Factors, Cancer Severity, Survival Outcomes, and Treatment Costs Using Statistical Methods

Tools: Python | Pandas | Matplotlib | Seaborn | Statistical Analysis|Scikitlearn



1. Project Overview:

This project presents an exploratory and inferential analysis of a global cancer patient dataset containing **50,000 patient records** collected across **10 countries** between **2015 and 2024**.

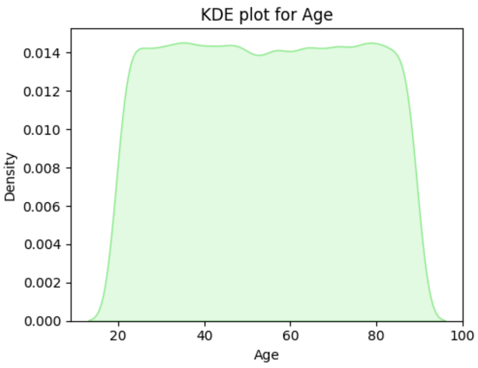
The objective of this analysis is to uncover patterns related to **demographics, risk factors, cancer severity, treatment cost, and survival outcomes**, and to evaluate whether commonly assumed relationships are supported by data.

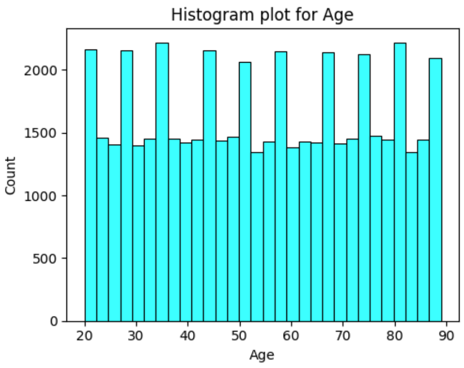
2. Dataset Description:

The dataset captures a wide range of patient-level information, including:

- **Demographics:** Age, gender, country, year of diagnosis
- **Risk Factors:** Smoking, alcohol use, obesity, genetic risk, air pollution exposure
- **Clinical Variables:** Cancer type, cancer stage, severity score
- **Economic & Outcome Variables:** Treatment cost (USD), survival years

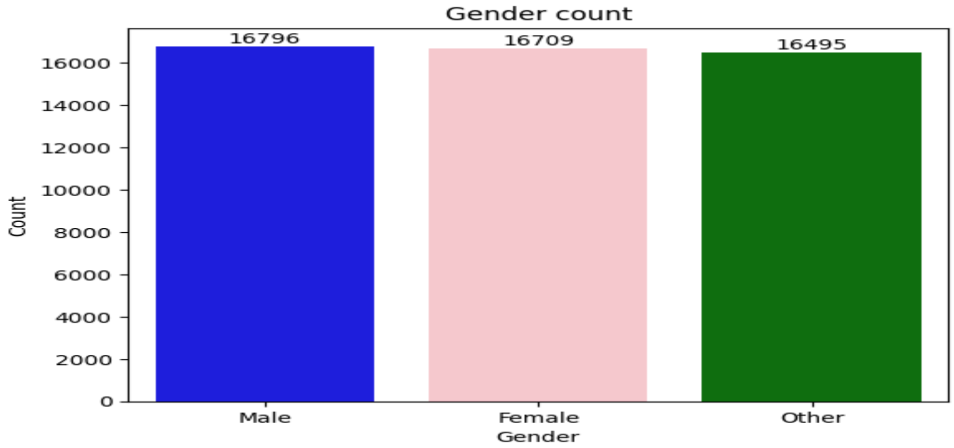
3. Demographic & Distribution Insights:





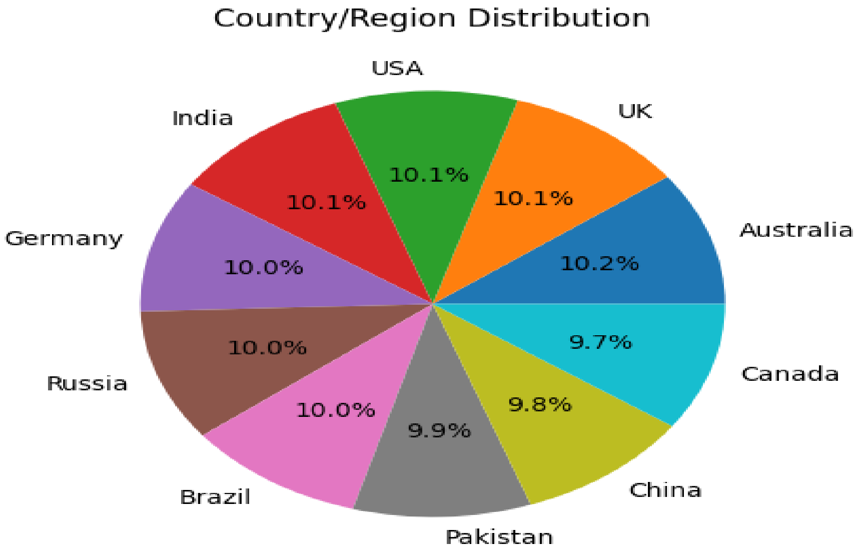
- Patients range from **20 to 89 years,** with a mean age of 54.4 years, ensuring representation across age groups.

- The dataset represents both younger and elderly patients well, making it suitable for age-based comparisons in severity, treatment cost, and survival analysis.
- Gender distribution supports meaningful gender-based analysis.
- Three gender categories: **Male, Female, Other**
- Most common: **Male (16,796 records)**



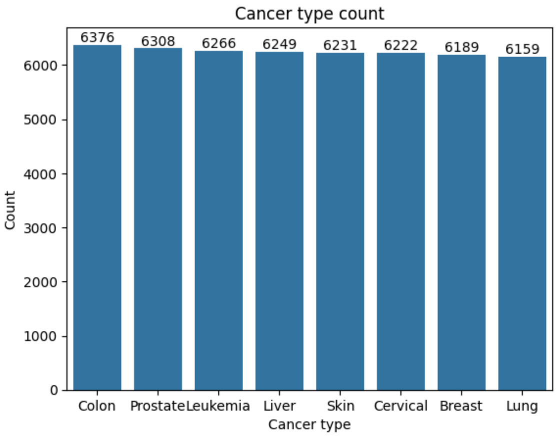
Country-Level Distribution

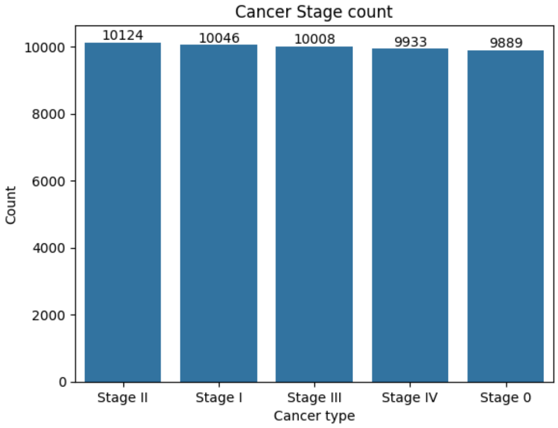
- Patients originate from **10 different countries**
- Australia has the highest count (5,092 patients)
- Overall, patient counts are **nearly uniform across countries**



Cancer Type & Stage Distribution

- **8 cancer types**, each with approximately similar sample sizes
- Most common cancers: **Colon cancer**, followed by Prostate cancer
- Cancer stages range from 0 to 4, with **Stage 2** being the most frequent



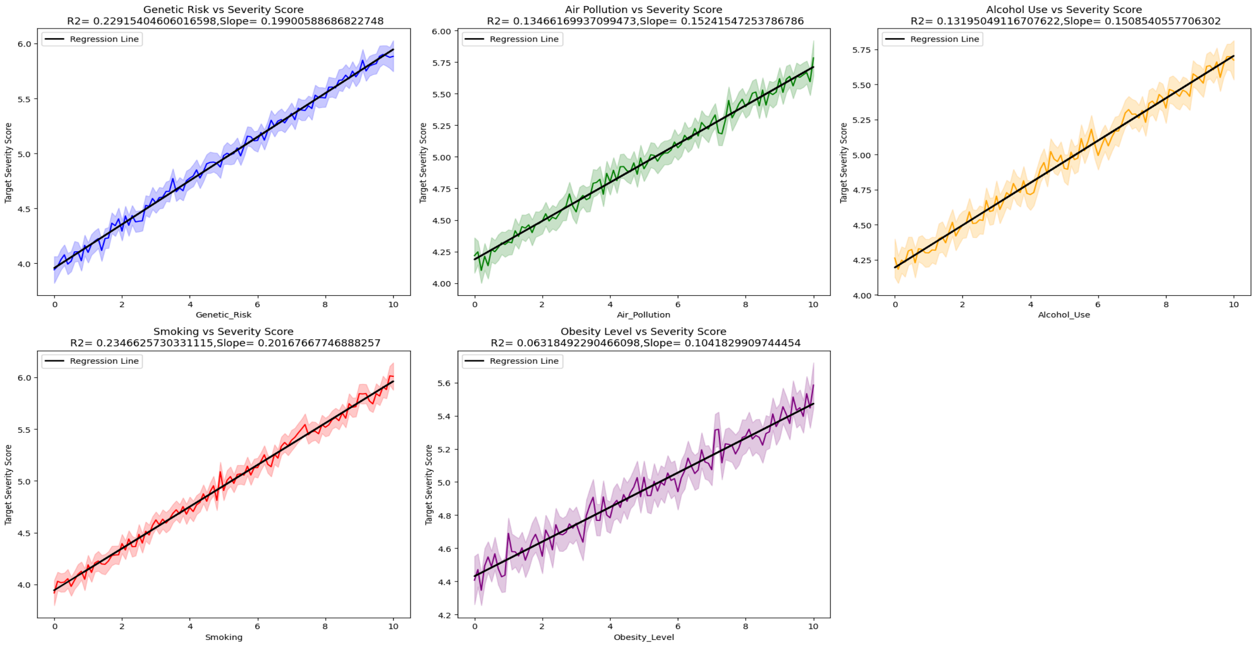


4. Relationship Between Risk Factors & Cancer Severity:

- Individual risk factors (smoking, alcohol use, obesity, pollution) show weak linear relationships with cancer severity (low R^2 values).
- While severity tends to increase as risk factors increase, no single factor strongly explains severity on its own, indicating a multi-factorial disease nature

Method Used

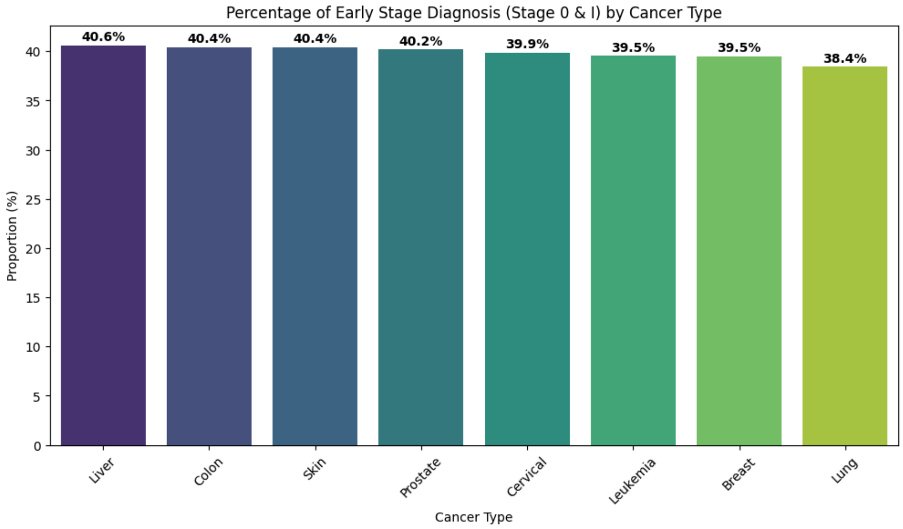
- **Simple linear regression** between individual risk factors and Target Severity Score
- **R^2 values range from 0.06 to 0.23**, indicating weak linear relationships
- All slopes are **positive**, suggesting severity increases as risk factors increase
- However, these increases are **not strongly consistent**



Risk Factors vs Cancer Severity

5. Early-Stage Diagnosis Across Cancer Types:

- Early-stage diagnosis rates range between 3**8.43% and 40.61%**
- **Liver cance**r shows the highest early detection rate
- Lung cancer shows the lowest

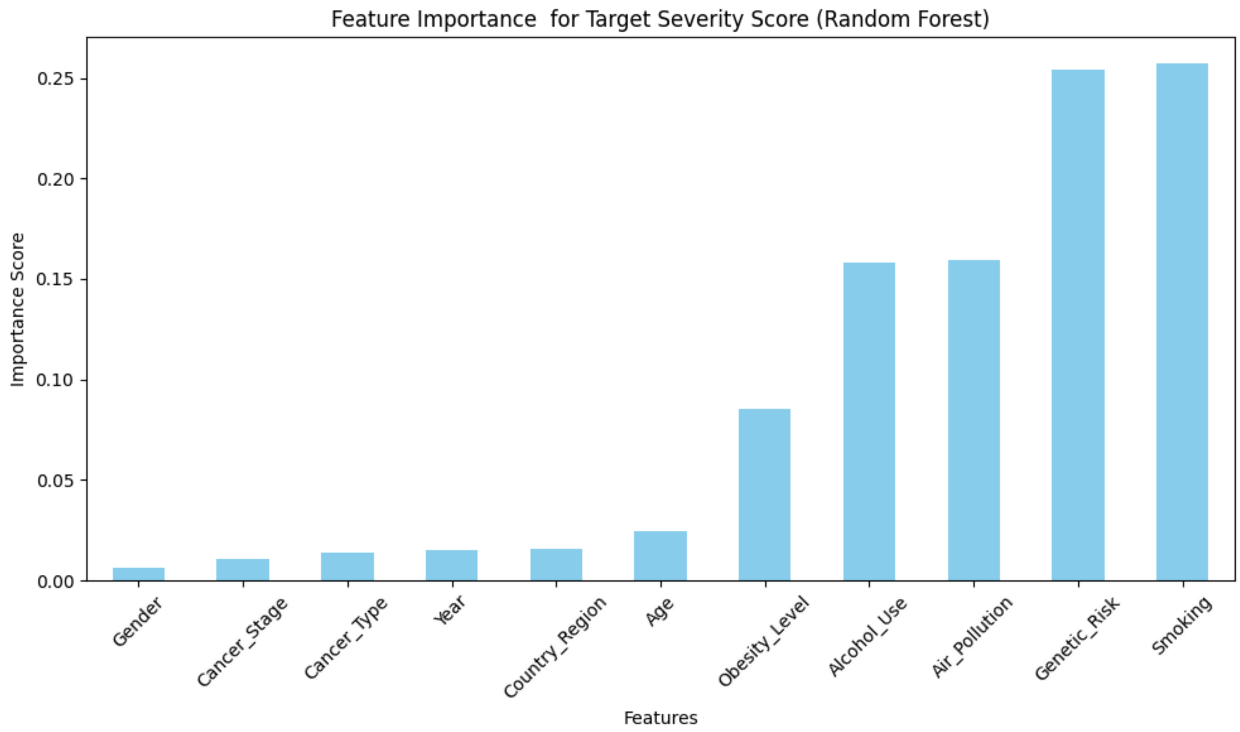


6. Feature Importance Analysis (Random Forest):

- **Smoking(0.2336) and genetic risk** are the strongest predictors of cancer severity
- Environmental factors like **air pollution** also play a significant role
- Age and gender contribute very little to severity prediction

Takeaway:

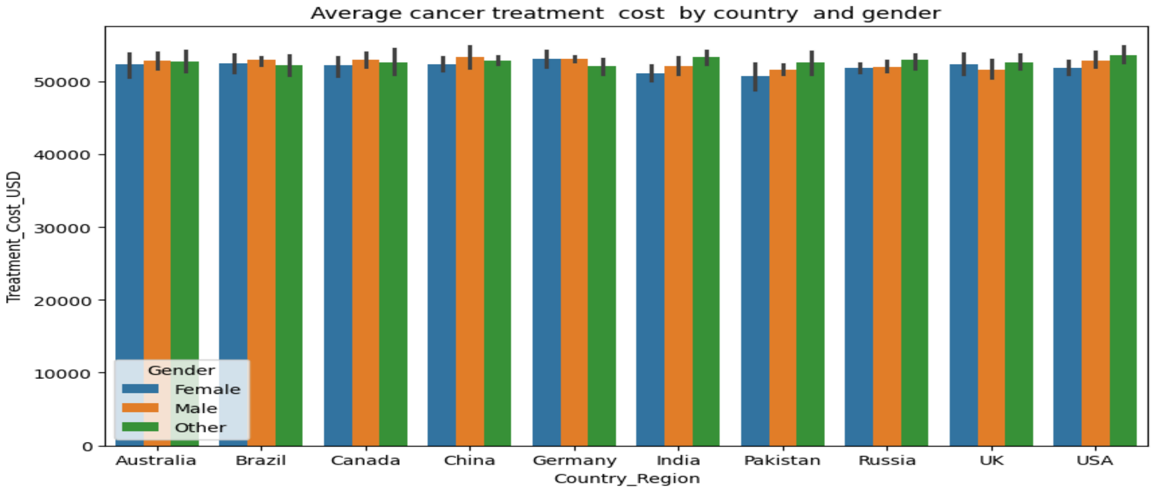
- Public health interventions targeting **smoking reduction, genetic screening, and environmental quality** may have the greatest impact on reducing cancer severity.



7. Economic Analysis – Treatment Cost Patterns:

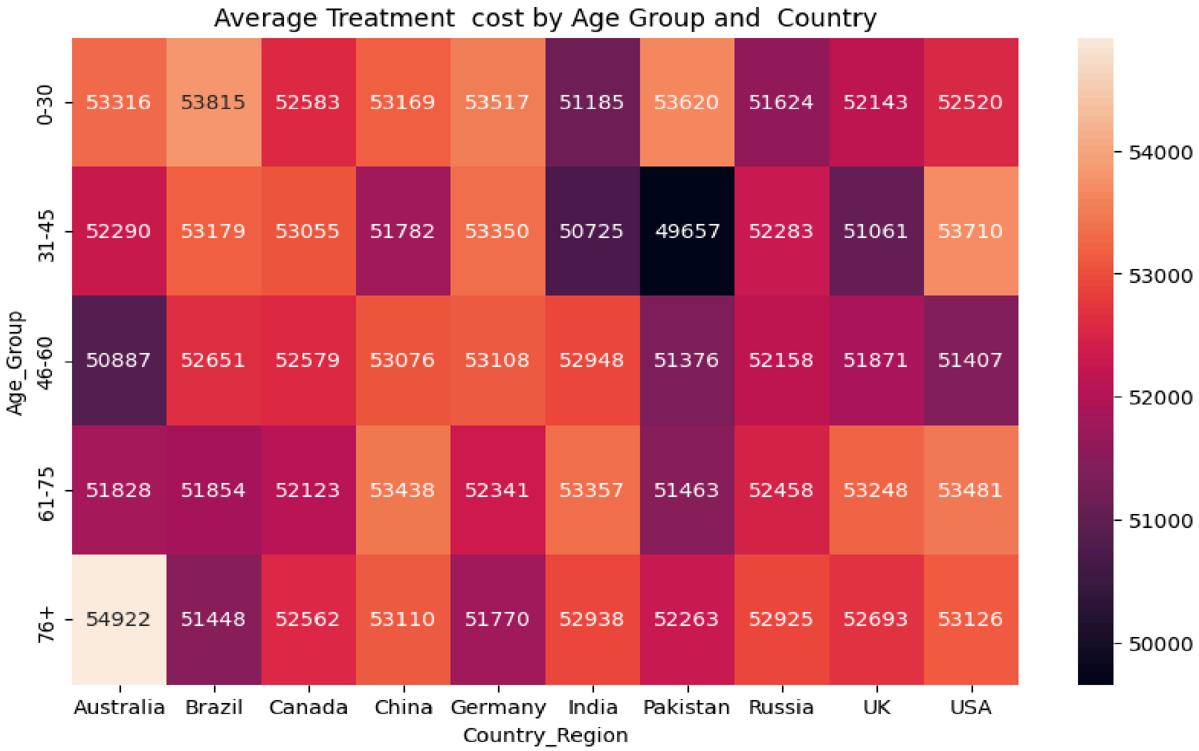
Country-Level Costs

- The highest treatment costs were observed in the **USA, Australia, and China**
- Significantly lower costs in **India and Pakistan**



Gender & Age Effects

- Treatment cost is **similar across genders**
- Costs increase sharply after **age 61**, especially in developed countries.



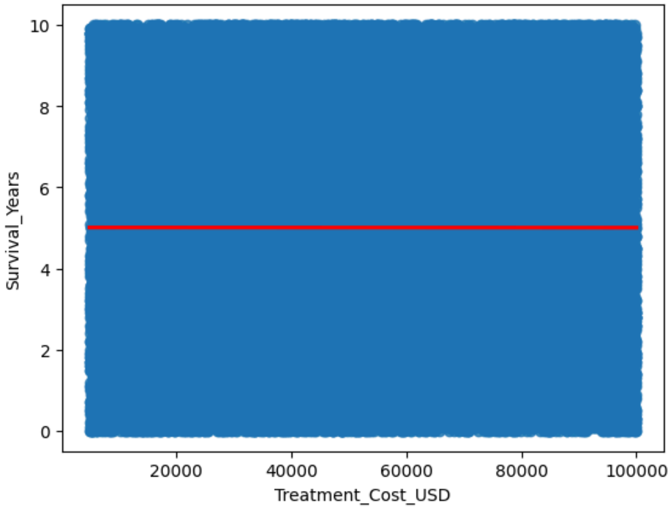
8. Inferential Statistical Tests:

To validate patterns, formal hypothesis testing was conducted using statistical methods.

Treatment Cost vs. Survival Years

A correlation test was performed to assess whether higher treatment costs are associated with longer survival.

- **Null Hypothesis (H_0):** No relationship between treatment cost and survival years.
- **Result:** No statistically significant relationship was found.



Treatment Cost vs. Cancer Stage

A **Kruskal–Wallis test** was conducted to compare average treatment costs across cancer stages (0–4).

- **H_0:** Treatment costs are equal across all cancer stages.
- **p-value = 0.4254**
- **Conclusion:** No statistically significant difference in costs across stages.

Survival Years vs. Cancer Stage

Another Kruskal–Wallis test evaluated whether survival duration differs by cancer stage.

- **H_0:** Survival years are equal across all cancer stages.
- **p-value = 0.6033**
- **Conclusion:** No statistically significant difference in survival years across stages.

Overall, these results indicate that in this dataset, cancer stage does not significantly explain variations in treatment cost or survival duration.

9. Interaction Between Genetic Risk and Smoking

A **Multiple Linear Regression model** was used to test whether genetic risk amplifies the effect of smoking on cancer severity.

- **H_0:** No interaction effect
- **Interaction coefficient:** -0.000228
- **p-value = 0.628 (> 0.05)**

Since the result is not statistically significant, we fail to reject the null hypothesis.

Genetic risk does not significantly modify the effect of smoking on cancer severity. Both may act independently, but there is **no strong interaction effect in this dataset**.

10. Final Takeaways:

- Cancer severity is influenced by **multiple weak factors**, not a single dominant variable
- Smoking, genetics, and environmental exposure are the strongest contributors

- Higher treatment cost does **not ensure better survival outcomes**
- Healthcare affordability and outcomes vary significantly across countries

---

11. Limitations & Future Scope

- **No Significant Stage-Based Differences**
  Although higher cancer stages are generally expected to increase treatment cost and reduce survival years, statistical tests did not show significant differences across stages. This may indicate limited variability in the dataset, balanced stage distribution, or the influence of other unmeasured variables.
- **Weak Individual Predictors**
  Most risk factors showed weak linear relationships with cancer severity (low R^2 values), suggesting that severity is influenced by multiple combined factors rather than any single variable.
- **No Significant Interaction Effect**
  The regression analysis found no significant interaction between genetic risk and smoking. While both may independently influence severity, the dataset does not show evidence that genetic risk amplifies the effect of smoking.

Future Scope

1. **Build Predictive Models**
   Develop regression or classification models to predict severity or survival years.
2. **Feature Engineering**
   Create interaction variables and composite risk scores to improve explanatory power.
3. **Country-Level Policy Analysis**
   Explore healthcare system efficiency and cost-effectiveness in more depth.
4. **Include More Medical Variables**
   Add treatment type, hospital type, genetic biomarkers, or longitudinal data.

Thank You

This report demonstrates my ability to perform structured data analysis, apply statistical testing, and derive actionable insights from complex datasets.

Prepared by:

Sonu Saroj
Data Analyst

Email: sonusaroj@02010@gmail.com
LinkedIn: Sonu Saroj
GitHub: github.com/sonusrj