Summry:

Comprehensive EDA and Inferential Analysis on Large-Scale Cancer Patients Dataset

Dataset:
https://docs.google.com/spreadsheets/d/1xqoFuviCB9CFKFfuXOsXxCkoUvjQd-HblQ3KoA0rQLY/edit?usp=sharing

---

Project Vision

This project harnesses the power of **advanced analytics** and **global health data** to uncover actionable insights into cancer care, outcomes, and disparities. Using a robust dataset of **50,000 cancer patient records** collected from multiple countries between **2015 and 2024**, we aim to bridge the gap between raw data and life-saving decisions.

What the Data Covers

Our dataset offers a 360-degree view of cancer patient profiles, including:

- **Demographics:** Age, gender, country, year of diagnosis.
- **Genetic & Lifestyle Risks:** Genetic predisposition, smoking, alcohol use, obesity
- **Environmental Exposure:** Air pollution
- **Clinical & Economic Variables:** Cancer type, stage, treatment cost
- **Patient Outcomes:** Survival years, severity scores

Core Objectives

---

1. Exploratory Data Analysis (EDA) 🔎

- Identify key trends, hidden patterns, and relationships
- Visualize disparities in diagnosis, lifestyle, treatment, and outcomes
- Highlight variations across countries, age groups, and cancer stages

---

2. Inferential & Predictive Analytics

Using statistical methods, I have explore and answer critical healthcare questions:

- Determine the relationship between risk factors and cancer severity
- Analyze the proportion of early-stage diagnoses by cancer type
- Identify key predictors of cancer severity and survival years
- Explore the economic burden of cancer treatment across different demographics and countries
- Assess whether higher treatment cost is associated with longer survival
- Evaluate if higher cancer stages lead to greater treatment costs and reduced survival years
- Examine whether higher genetic risk amplifies the negative effects of smoking on cancer severity and survival outcomes

3. Insight Extraction and Documentation

- For each major analysis, I tried to draw **clear, evidence-backed inferences**
- All findings will be **systematically documented** for future clinical, academic, and operational use
- Each insight will be linked to **real-world implications** — to guide decision-making and policy formulation

Report :

Report :

Cancer Data EDA And Inferential Analysis

Prepared by: Naman Verma

1. Exploratory Data Analysis (EDA) 🔎

The Exploratory Data Analysis phase focused on uncovering trends, disparities, and structural patterns across patient demographics, risk factors, clinical indicators, and outcomes.

1.1 Identify key trends, hidden patterns, and relationships

Key Trends Identified

- **Smoking and Genetic Risk emerged as the strongest factors** associated with higher cancer severity.
- **Treatment cost shows a right-skewed distribution**, indicating that a small subset of patients incur extremely high medical expenses.
- **Severity scores cluster around moderate values**, with fewer extreme-severity cases.
- **Survival years vary widely**, but most patients fall within a mid-range of 3–7 years.

Hidden Patterns

- **Smoking ↔ Severity:** Strong positive correlation — severity increases with smoking frequency.
- **Genetic Risk ↔ Severity:** Nearly identical strength to smoking, confirming biological vulnerability.
- **Treatment Cost ↔ Severity:** Higher severity correlates with higher treatment cost.

- **Air Pollution ↔ Severity:** Moderate positive association — environmental exposure affects outcomes.

Relationships Identified (from correlation heatmap & regression plots)

- Lifestyle risks (smoking, alcohol) show a clearer relationship with severity than demographic factors.
- **Gender, obesity, and stage** show *weak or negligible* relationships with severity or survival.
- Age has *limited* effect on severity but a mild upward effect on treatment cost.

1.2 Visualize disparities in diagnosis, lifestyle, treatment, and outcomes

Diagnosis-Related Disparities

- **Cancer type distribution** varies substantially across the dataset, with lung cancer among the most common.
- **Stage distribution** is relatively uniform—no dominance of early or late-stage cases.

Lifestyle Disparities

- Smoking and alcohol use vary significantly by population group.
- Genetic predisposition levels (Low, Medium, High) differ by patient clusters, influencing severity.

Treatment Disparities

- **Developed countries** (USA, Australia, China) show significantly higher treatment costs.
- **Developing countries** (India, Pakistan) show lower average costs, likely due to reduced pricing structures or limited access to advanced care.

Outcome Disparities

- Survival years show moderate variation across demographics but **little variation across cancer stage**.
- Severity scores differ more by **risk factors** than by demographic variables.

1.3 Highlight variations across countries, age groups, and cancer stages

Country-Level Variations

- **USA > Australia > China** consistently show the highest treatment costs.
- Countries with **public healthcare systems** (Canada, Germany, UK) show more cost stability across age groups.
- Severity varies moderately between countries but not drastically.

Age Group Variations

- Treatment costs **rise progressively with age**, especially for patients aged 61+.
- Severity does not show a strong age-based pattern.
- Older patients show slightly lower survival years, but not significantly distinct.

Cancer Stage Variations

- Surprisingly **little difference** in:
    - Treatment cost across stages
    - Survival years across stages
- EDA hinted at this, and inferential tests later confirmed no statistically significant differences.

---

2. Inferential & Predictive Analytics

This section answers each inferential question using statistical tests and model-based insights applied in your project.

2.1 Analyze the proportion of early-stage diagnoses by cancer type

EDA visualizations showed:

- Many cancer types (e.g., breast, colorectal) have **higher proportions of Stage I and Stage II diagnoses**, indicating effective early detection programs.
- Cancers like lung cancer showed more mid-to-late stage diagnoses.

Conclusion:
Early-stage diagnosis varies by cancer type, with some cancers detected earlier, while others appear more in advanced stages.

2.2 Identify key predictors of cancer severity and survival years

Using feature importance scores:

Top Predictors of Severity

1. **Smoking** — strongest predictor
2. **Genetic Risk** — almost equally strong
3. **Treatment Cost** — higher severity → higher cost
4. **Alcohol Use**

5. **Air Pollution**

Weak predictors:

- Gender
- Obesity
- Age × Gender interaction

Top Predictors of Survival Years

- No strong predictors identified; most factors showed **weak correlations** with survival.
- Severity and stage were surprisingly **not major drivers** of survival years in this dataset.

2.3 Explore the economic burden of cancer treatment across different demographics and countries

Findings from EDA + grouped summaries:

Economic Burden Depends Strongly on Country

- Highest costs in **USA, Australia, China**
- Lowest in **India, Pakistan**

Age Impact

- Treatment costs rise sharply for patients **older than 60**, reflecting additional comorbidities and intensity of care.

Genetic & Lifestyle Factors

Patients with higher severity (linked to smoking & genetic risk) tend to require more intensive and expensive treatment.

Conclusion:
Economic burden is shaped by **country, age, and severity**, with country being the strongest determinant.

2.4 Assess whether higher treatment cost is associated with longer survival

Findings:

- Correlation analysis showed **weak or no relationship** between treatment cost and survival.
- Visual inspection of bar plots showed no clear increasing pattern.

Conclusion:
Your data **does not support** the idea that paying more results in longer survival.

2.5 Evaluate if higher cancer stages lead to greater treatment costs and reduced survival years

Statistical test used: **Kruskal–Wallis**

- **Treatment Cost vs Stage:** p = 0.4254 → no significant difference
- **Survival Years vs Stage:** p = 0.6033 → no significant difference

Conclusion:
Cancer stage **does not significantly influence** average treatment cost or survival years in this dataset.

2.6 Examine whether higher genetic risk amplifies the negative effects of smoking on cancer severity and survival outcomes

Statistical method: **Multiple Linear Regression with interaction term**

- Interaction coefficient: **–0.000228** (nearly zero)
- p-value: **0.628** (> 0.05)

Conclusion:
There is **no evidence** that genetic predisposition amplifies or reduces the effect of smoking on:

- Severity scores
- Survival years

Smoking and genetic risk act **independently**, not interactively.

📌 Key Findings Summary

- Smoking and genetic risk are the **dominant predictors of cancer severity**.
- Treatment cost varies strongly across countries, especially higher in developed nations.
- Age increases treatment cost but does **not** greatly impact severity or survival.
- Cancer type influences early vs late diagnosis patterns.
- **Cancer stage does NOT significantly affect treatment cost or survival years.**
- No interaction effect between genetic risk and smoking on severity/survival.
- Survival years show **weak associations** with all analyzed predictors.
- Treatment cost does **not** correlate meaningfully with longer survival.
- Environmental exposure (air pollution) moderately affects severity.
- Demographics like gender and obesity contribute minimally to clinical outcomes.