CHAPTER NO- 1
INTRODUCTION

Introduction

1.1 CANCER: A GLOBAL HEALTH CHALLENGE OF UNPARALLELED COMPLEXITY

Cancer is one of the most serious and complex health issues facing the world today. It occurs when certain cells in the body begin to grow and spread uncontrollably, interfering with the normal functioning of organs. These abnormal cells can spread to other parts of the body, making the disease harder to manage and more dangerous over time. Unlike other illnesses that are limited to a specific part of the body, cancer can appear in nearly any organ or system, which makes it particularly difficult to detect early and treat effectively.

The global impact of cancer has been rising steadily. According to data from the World Health Organization, nearly 10 million people lost their lives to cancer in 2020, making it the second most common cause of death worldwide. Several factors are contributing to this alarming trend. As populations live longer and lifestyles become more sedentary and unhealthy, the number of cancer cases is increasing. Rapid urbanization, changes in diet, exposure to pollution, and smoking are just a few of the lifestyle and environmental shifts linked to this rise. At the same time, many countries, especially developing ones, still lack the healthcare infrastructure needed to offer timely diagnosis, advanced treatment, or adequate follow-up care. These gaps often result in higher death rates, even for cancers that are potentially treatable when detected early.

One of the major challenges in dealing with cancer is that it does not follow a predictable pattern. Even when two people are diagnosed with the same type of cancer, the way the disease progresses and how it responds to treatment can vary widely. This variation is influenced by several personal and external factors, including age, health history, environment, and access to healthcare. Because of this, there is no “one-size-fits-all” approach to treatment. This complexity calls for more personalized methods of care, which in turn increase the cost and time needed for treatment planning. These factors not only burden patients and families but also put pressure on healthcare systems and insurance providers.

Beyond the medical aspect, cancer presents serious social and economic challenges. It often leads to long-term loss of productivity, both for individuals undergoing treatment and for family members who become caregivers. In low- and middle-income countries, the economic burden is especially heavy, where lack of awareness, social stigma, and limited healthcare access further complicate the situation. Additionally, the high cost of advanced treatments and medicines can push many households into financial distress, highlighting the

link between health and economic vulnerability.

Addressing cancer effectively requires more than just advanced medical technology. It demands a coordinated effort that includes public awareness campaigns, strong healthcare policies, early screening programs, and investment in healthcare infrastructure. Governments, businesses, and communities all have a role to play in reducing the impact of this disease. From implementing workplace wellness programs to offering insurance coverage and improving access to quality care, the private sector can be a powerful ally in this fight. Cancer, in many ways, reflects the broader challenges of today's world—rising healthcare costs, unequal access to services, and the struggle to balance economic development with public well-being. As such, it is not only a medical condition but also a pressing management and policy issue that calls for innovative thinking, strategic planning, and collective action. For future business leaders and policymakers, understanding the broader impact of diseases like cancer is essential for building healthier, more resilient societies.

1.2 OBJECTIVE OF THE STUDY

Despite significant advances in cancer research, treatment, and prevention over the last decade, disparities in outcomes persist. These disparities are not solely explained by biological differences but are compounded by a complex web of genetic, lifestyle, environmental, and socioeconomic factors. Existing studies often examine these domains in isolation, missing the synergistic and sometimes compounding effects they have on patient prognosis.

There is an urgent need to adopt an integrated, multifactorial approach to studying cancer outcomes — one that reflects the real-world complexity faced by patients, clinicians, and policymakers alike. Without such a holistic understanding, efforts to improve survival rates and quality of life for cancer patients risk being incomplete or misdirected.
Therefore, this study seeks to answer a central question:
"How do genetic, lifestyle, environmental, and economic factors jointly influence cancer severity and survival outcomes between 2015 and 2024?"

General Objective

To perform a multi-factorial analysis of cancer patient outcomes from 2015 to 2024 by investigating the combined influence of genetic, lifestyle, environmental, and economic determinants on cancer severity, treatment costs, and survival years using statistical and machine learning techniques.

Specific Objectives

- To identify key predictors of cancer severity and survival years through

advanced machine learning and statistical analysis.

- To evaluate the economic burden of cancer treatment across different demographics, countries, and socioeconomic groups.
- To investigate whether higher treatment costs are associated with increased survival years.
- To determine if advanced cancer stages correlate with higher treatment costs and reduced survival durations.
- To examine how higher genetic risk factors amplify the negative effects of smoking on patient survival outcomes.
- To analyse whether mean survival years differ significantly by cancer stage.
- To explore the association between age group and the stage at which cancer is diagnosed.

1.3 IMPORTANCE OF STUDY

The study of cancer holds critical importance for a variety of interrelated reasons that span health, economics, and policy. At the forefront is the reality of its high morbidity and mortality rates, which demand urgent and sustained attention from global health systems. Cancer remains one of the leading causes of death worldwide, and its incidence is rising steadily. As medical science succeeds in extending human life expectancy, there is a natural shift in health priorities from acute, infectious diseases to chronic, long-term conditions such as cancer. This shift places an enormous and growing strain on healthcare resources, insurance systems, and public health infrastructure. In turn, governments and private sectors must adapt to these evolving demands, making cancer not only a medical issue but a significant economic and managerial concern as well.

Despite major strides in medical technology, diagnostics, and therapeutic innovations, the gap in cancer outcomes between high-income and low-income countries—and even within different regions of the same country—remains wide. These disparities are shaped by multiple factors including access to early screening, availability of trained medical professionals, public awareness, affordability of treatments, and differences in healthcare delivery systems. This survival gap reflects underlying social and economic inequalities, emphasizing the need for targeted research and robust policy frameworks to ensure equitable care for all populations. Addressing these differences through effective health strategies can play a major role in reducing the overall cancer burden and improving population health outcomes.

Understanding cancer, therefore, goes beyond the clinical study of tumors or the prescription of treatment regimens. It requires a comprehensive and interdisciplinary perspective that incorporates medical, behavioral, social, and environmental dimensions. For instance, the variation in treatment response among patients with the same type of cancer continues to puzzle clinicians and researchers alike. Questions such as why some individuals respond positively to chemotherapy while others with similar profiles do not, or why geographically comparable regions report significantly different mortality rates, highlight the complexity of the disease. These variations suggest that cancer is influenced by an intricate interplay of biological, environmental, and social determinants. Exploring these patterns can lead to more informed, patient-centered treatment and care models.

Additionally, cancer research plays a vital role in formulating and implementing effective prevention strategies. By identifying modifiable risk factors such as tobacco use, harmful alcohol consumption, physical inactivity, unhealthy dietary habits, and exposure to environmental toxins, stakeholders can design and execute public health initiatives that not only save lives but also reduce long-term healthcare costs. Education campaigns, early intervention programs, and regulatory policies targeting these risk factors are examples of how research can translate into real-world impact. On the other hand, studying non-modifiable factors like genetics, age, ethnicity, and socioeconomic status allows health systems to stratify risk, personalize care, and allocate resources more efficiently. Such insights are crucial for driving health equity and ensuring that vulnerable groups are not left behind.

1.4 NEED FOR THE STUDY

While there is a wealth of literature on cancer, much of the research has historically concentrated on isolated factors, such as genetic predispositions or treatment effectiveness, without fully considering the broader context in which cancer occurs. However, the real-world experience of cancer is far more intricate than what these single-dimensional studies suggest. A patient's prognosis is rarely determined by just one factor. Instead, it is shaped by the interplay of various elements—ranging from genetics to lifestyle choices, environmental exposures, socioeconomic status, and access to healthcare—that interact in dynamic, sometimes unpredictable ways. This complexity is often overlooked in conventional research, which tends to focus on one aspect of the disease in isolation. As such, there is a critical gap in contemporary cancer research: the lack of comprehensive, data-driven studies that approach cancer

outcomes from a multifactorial perspective, considering how multiple factors converge to influence a patient's experience and survival.

This study aims to bridge this gap by analyzing cancer patient data collected over a decade (2015–2024), incorporating a wide array of variables including genetic data, lifestyle habits, environmental influences, and economic factors. The integration of these diverse factors is crucial to understanding the true complexity of cancer as it is experienced by real patients across different settings. By examining such a broad spectrum of variables, the study provides a more accurate and holistic view of how cancer impacts individuals, highlighting the intricate web of influences that contribute to disease progression and outcomes.

Furthermore, with the growing prominence of personalized medicine and the increasing reliance on data analytics, there is a pressing need to develop predictive models that account for the multiple dimensions of a patient's life. The use of data-driven approaches allows for a more nuanced understanding of how various factors, both modifiable and non-modifiable, affect cancer severity and survival. This, in turn, can lead to more precise prediction models that not only inform individualized treatment plans but also guide public health strategies aimed at prevention, early detection, and better outcomes for patients. As we move toward more tailored, patient-specific approaches to healthcare, the importance of considering the full spectrum of influences on cancer outcomes becomes ever more apparent.

In addition, the temporal dimension of this study—spanning nearly a decade—provides a unique opportunity to identify trends, shifts in patterns, and emerging risk factors over time. By analyzing data across multiple years, this study offers insights into how changing environmental conditions, economic upheavals (such as the global COVID-19 pandemic), and evolving healthcare policies have shaped cancer outcomes on a global scale. The study's longitudinal approach not only allows for a deep dive into the ways in which these factors have affected patient care and survival but also provides a lens through which to anticipate future challenges and opportunities in cancer treatment and management. This comprehensive, time-sensitive perspective is critical for developing strategies that are not only responsive to current trends but also forward-thinking and resilient in the face of ongoing global changes.

1.5 SCOPE OF THE STUDY

This study adopts a wide scope in order to capture the diverse and interconnected factors that influence cancer patient outcomes. The key

dimensions explored include:

- **Genetic Determinants:** These include hereditary mutations, gene expression profiles, family history of cancer, and other molecular-level indicators that predispose individuals to cancer or affect their response to treatment.
- **Lifestyle Factors:** Elements such as smoking habits, alcohol consumption, dietary practices, physical activity levels, and obesity are considered due to their established links to various cancer types.
- **Environmental Exposures:** Air and water quality, exposure to industrial pollutants, residential proximity to hazardous sites, and occupational risks are evaluated to assess their influence on cancer incidence and progression.
- **Economic Determinants:** Socioeconomic status, access to healthcare, insurance coverage, education levels, and employment status are critical variables that affect not just the diagnosis but also the quality and continuity of treatment.

Geographically, the dataset comprises cancer cases from multiple regions, allowing for cross-country comparisons and regional disparities in outcomes. Demographically, the study includes individuals across age groups, genders, and ethnic backgrounds, thereby enhancing the generalizability of the findings. The study also distinguishes between different types of cancer and severity levels, such as early-stage vs late-stage diagnoses, and short-term survival vs long-term remission. This allows for a granular analysis of how each factor plays out across specific contexts.

CHAPTER - 2 LITERATURE REVIEW

LITERATURE REVIEW

2.1 INTRODUCTION TO CANCER AND ITS GLOBAL BURDEN

Cancer remains one of the most complex and heterogeneous diseases confronting modern medicine. According to the World Health Organization (WHO), cancer was responsible for nearly 10 million deaths worldwide in 2020, making it the second leading cause of death globally. What distinguishes cancer from many other chronic diseases is its multifactorial nature. The progression from healthy cells to malignancy is rarely due to a single cause but results from intricate interactions among **genetic predisposition**, **lifestyle choices**, **environmental exposures**, and **socioeconomic conditions**.

While advances in diagnostic techniques and therapeutic modalities have led to increased survival rates for certain cancers, disparities persist across regions, income levels, and ethnic groups. This complexity necessitates an integrated, cross-disciplinary approach to better understand the determinants that drive cancer outcomes—precisely what this dissertation aims to address.

2.2 GENETIC DETERMINANTS OF CANCER

Cancer is fundamentally a genetic disease, caused by changes in DNA that lead to uncontrolled cell growth. Some mutations are inherited, while others are acquired due to environmental exposures or random cellular events.
Studies have long identified genetic markers associated with increased cancer risk. For example:

- **BRCA1 and BRCA2** mutations dramatically elevate the risk for breast and ovarian cancers.
- **TP53**, known as the "guardian of the genome," is mutated in over 50% of all cancers (Olivier et al., 2010).
- **KRAS and EGFR** mutations play a pivotal role in colorectal and lung cancers, respectively.

Importantly, genetic mutations do not operate in isolation. A person with a high-risk mutation may never develop cancer unless interacting environmental or behavioral factors (e.g., smoking, diet, chemical exposure) act as triggers. Hence, researchers emphasize **gene-environment interactions** (Khoury et al., 2005) to predict individual susceptibility and clinical outcomes more accurately.

2.3 LIFESTYLE AND BEHAVIORAL FACTORS

Lifestyle factors are among the most modifiable risks for both developing and surviving cancer. The World Cancer Research Fund (WCRF) and the American Institute for Cancer Research (AICR) have reported that up to **40% of all cancer cases are preventable** through lifestyle changes.
Key behaviors include:

- **Smoking**: Responsible for ~22% of cancer deaths globally. Strongly linked to lung, throat, bladder, and pancreatic cancers.
- **Alcohol consumption**: Known to increase the risk of mouth, liver, breast, and esophageal cancers.
- **Diet and obesity**: High consumption of red/processed meats and low fiber intake are associated with colorectal cancer, while obesity is a risk factor for at least 13 different cancer types (Lauby-Secretan et al., 2016).
- **Physical inactivity**: Contributes to cancer indirectly by facilitating weight gain and metabolic dysfunction.

Recent literature also highlights **synergistic effects**—for instance, smoking's negative impact on survival is notably worse among genetically predisposed individuals. Therefore, behavior must be considered within the broader context of genetics and access to resources.

2.4 ENVIRONMENTAL EXPOSURE AND CANCER RISK

The environment where individuals live and work plays a critical but often underestimated role in cancer etiology. Environmental carcinogens may be **airborne, ingested, or dermally absorbed** and include substances like:

- **Asbestos** (mesothelioma, lung cancer)
- **Radon gas** (lung cancer)
- **Pesticides and heavy metals** (leukemia, lymphoma, bladder cancer)
- **Particulate matter (PM2.5 and PM10)** in air pollution—now classified as a **Group 1 carcinogen** by the International Agency for Research on Cancer (IARC)

Multiple studies have drawn connections between industrialization, urban living, and increased cancer risk. For instance, residents near petrochemical facilities or heavy traffic zones show higher incidence rates for leukemia and lung cancer (Ghosh et al., 2018).
Moreover, **climate change** and **environmental injustice** have led to disproportionate exposure for marginalized populations—exacerbating health inequities in cancer outcomes.

2.5 ECONOMIC AND SOCIO-DEMOGRAPHIC DETERMINANTS

Socioeconomic status (SES) profoundly affects both access to and quality of cancer care. Income, education, health insurance, and geographic location can determine:

- When and how early cancer is diagnosed.
- Access to advanced diagnostics and treatment.
- Ability to afford follow-up care, chemotherapy, or palliative support.

A study published by the **American Cancer Society (2021)** found that:

- Uninsured patients were 60% more likely to be diagnosed at a later stage.
- Median survival was lower among those in lower income brackets, even after adjusting for stage and treatment.

In low- and middle-income countries, limited healthcare infrastructure means patients often present with advanced disease due to inadequate screening or delayed care-seeking behavior. This contributes to poorer survival despite similar biological disease.
Furthermore, the **financial toxicity** of cancer—i.e., the economic burden on patients and families—can itself be a barrier to ongoing treatment, leading to dropouts or reduced adherence.

2.6 INTERCATIONS AND THE CASE FOR MULTI-FACTORIAL ANALYSIS

Traditional cancer studies often isolate variables: a study on smoking here, a report on genetics there. However, real-world outcomes arise from **interdependent variables**. For instance:

- A patient with a BRCA1 mutation may survive longer if diagnosed early and treated in a high-income setting, but may have reduced survival if also a smoker living in a high-pollution area.
- Two patients with the same stage of cancer may have drastically different outcomes based on treatment affordability, genetic resilience, and environmental exposures.

Recent research increasingly supports **multi-factorial models** using machine learning and multivariate regression to unravel these complex relationships. Tools such as **survival analysis, Cox proportional hazards models, and decision trees** help identify not just primary risks but **interactions and compound effects**.
However, the literature shows that few studies cover all dimensions—**genetic, lifestyle, environmental, and economic**—together in a longitudinal context.

Your dissertation helps fill this gap by leveraging a decade of data across these dimensions.

2.7 TECHNILOGICAL ADVANCEMENTS IN CANCER RESEARCH

Over the past decade, rapid advancements in computational power and data availability have transformed cancer research. The application of Artificial Intelligence (AI) and Machine Learning (ML) has shifted the paradigm from reactive to predictive oncology.
Studies have shown how algorithms can:

- Predict cancer progression based on patient history and biomarkers (Esteva et al., 2017).
- Classify tumor subtypes with high accuracy using genomic and histopathological data.
- Predict treatment responses and survival time using random forests, support vector machines (SVM), and neural networks.

Moreover, bioinformatics platforms and data repositories such as TCGA (The Cancer Genome Atlas), GDC (Genomic Data Commons), and SEER (Surveillance, Epidemiology, and End Results) have democratized access to large datasets for multidisciplinary research.
This dissertation leverages such technological tools to identify and interpret the multifaceted patterns in cancer outcomes, integrating them with classical statistical models for added validity.

2.8 POLICY AND HEALTHCARE SYSTEM INFLUENCES

Beyond patient-level factors, broader **health policy frameworks** and **healthcare system efficiencies** significantly influence cancer outcomes.
Literature reveals that:

- **Universal healthcare systems** (e.g., UK's NHS or Canada's Medicare) typically report better early detection and lower financial toxicity.
- **Fragmented systems** like that of the US often result in disparities in cancer care depending on insurance coverage (Han et al., 2020).
- National **screening programs**, **vaccination drives** (e.g., HPV for cervical cancer), and **preventive care subsidies** play critical roles in shifting both incidence and mortality rates.

Unfortunately, few studies quantitatively compare these structural healthcare

determinants alongside patient-level predictors. Your multi-country dataset and analysis help bridge this void by capturing treatment costs, survival outcomes, and system-level disparities.

2.9 ETHICAL AND SOCIAL CONSIDERATIONS IN CANCER RESEARCH

Another emerging domain in the literature focuses on the **ethical**, **psychosocial**, and **equity-based aspects** of cancer research. Studies suggest that:

- **Cancer stigma** can prevent individuals from seeking timely diagnosis or treatment.
- **Cultural beliefs** influence how symptoms are interpreted and how decisions are made about care.
- The use of **genomic data** for research purposes raises ethical questions about privacy, data ownership, and informed consent—particularly when AI models are trained on patient histories.

Furthermore, **health inequities**—based on race, gender, geography, and income—persist even in technologically advanced health systems. Marginalized communities often face barriers to diagnosis, access to new therapies, and post-treatment support, all contributing to poorer outcomes.
Your research, by capturing economic, lifestyle, and demographic attributes alongside biological and medical factors, offers a more **inclusive and ethically conscious** lens on cancer survival and severity.

CHAPTER - 3
RESEARCH METHODOLOGY

3.1 RESEARCH METHODOLOGY:

Research Methodology refers to the *systematic plan* and *framework* that a researcher uses to collect, analyze, and interpret data in a study. It outlines:

- What kind of data was used
- How the data was gathered and cleaned
- What techniques were applied (e.g., statistical tests, ML models)
- Why these techniques were chosen
- How the results were interpreted

3.2 RESEARCH DESIGN

The study adopts a retrospective analytical research design, focusing on secondary data sourced from global records of cancer patients over a ten-year span. The primary aim is causal inference—to understand how various independent variables affect survival years and severity—and predictive modeling, which helps forecast future outcomes under similar conditions. This multi-pronged design combines:

- Descriptive analytics: To explore patterns, distributions, and demographic spreads.
- Inferential statistics: To test hypotheses and identify statistically significant relationships.
- Predictive modeling: To identify high-impact predictors and simulate outcome probabilities.

3.3 DATASET OVERVIEW

The dataset titled *Global Cancer Patients Dataset (2015–2024)* comprises anonymized records across diverse countries having 50000 rows and unique data points and includes the following major domains:

- Demographics: Age, Gender, Country region, Year.
- Genetic and lifestyle factors: Genetic Risk, Smoking, Alcohol use , Obesity level.
- Environmental Factors: Air Pollution.
- Medical and Economical Factors: Cancer Type, Cancer Stage, Treatment cost in USD
- Outcome Variables: Survival Years, Target Severity Score

Here is short description of each and every column present in the dataset: -

- Age: Patient's age (20-90 years)
- Gender: Male, Female, or Other
- Country/Region: Country or region of the patient
- Cancer Type: Various types of cancer (e.g., Breast, Lung, Colon)
- Cancer Stage: Stage 0 to Stage IV
- Risk Factors: Includes genetic risk, air pollution, alcohol use, smoking, obesity, etc.
- Treatment Cost: Estimated cost of cancer treatment (in USD)
- Survival Years: Years survived since diagnosis
- Severity Score: A composite score representing cancer severity

The multi-dimensional nature of this dataset enables layered and intersectional analysis of cancer progression and survivorship.

3.4 KEY RESEARCH QUESTION ADDRESSED

- To identify and evaluate the key predictors that significantly influence cancer severity and patient survival years.
- To investigate the economic burden of cancer treatment across diverse demographic segments and international contexts.
- To examine whether higher treatment costs are associated with improved survival outcomes among cancer patients.
- To assess the relationship between cancer stage and both treatment expenditure and survival duration, with a focus on whether advanced stages incur higher costs and lead to reduced survival.
- To explore how elevated genetic risk factors intensify the adverse impact of smoking on patient survival outcomes.
- To determine if mean survival years differ significantly across various stages of cancer progression.
- To analyze the association between different age groups and their corresponding cancer stages at the time of diagnosis.

3.5 TOOLS AND TECHNOLOGIES USED

- **Python** (Pandas, NumPy, SciKit-Learn, Matplotlib, Seaborn)
- **Statistical Methods**: ANOVA, Chi-Square Test, Pearson Correlation
- **ML Models**: Linear Regression, Random Forest, Logistic Regression, Grid Search Cv
- **Data Visualization**: Seaborn plots, pie charts, distribution plots , histogram, bar charts, heatmaps for exploratory data analysis

3.6 COURSE OF ACTION DECIDED AND FOLLOWED

This research adopted a structured and methodical approach to investigate the multi-factorial determinants of cancer severity and survival across global populations. The course of action was carefully designed to ensure that the research objectives were addressed systematically, using a combination of statistical, machine learning, and data visualization techniques.
Steps followed are :-

- **Problem Identification and Objective Formulation**
  The research commenced with the identification of a core problem related to understanding the complex factors influencing cancer outcomes. Based on this, both a general objective and a set of specific objectives were formulated to guide the study.
- **Data Collection from Various Sources**
  Cancer-related data spanning from 2015 to 2024 was sourced from reliable and diverse global repositories. These included patient demographics, genetic profiles, lifestyle factors, economic data, and clinical indicators such as cancer stage and survival years.
- **Dataset Loading and Setup in Jupyter Notebook (Python Environment)**
  The dataset was imported into a Python environment using Jupyter Notebook, enabling structured and reproducible data analysis through powerful libraries such as pandas, numpy, and matplotlib.
- **Dataset Understanding and Initial Inspection**
  An initial exploratory inspection was performed to understand the structure of the dataset. This involved reviewing the column names, data types, and identifying the presence of missing or duplicate records to ensure data quality.

- **Exploratory Data Analysis (EDA)**
  Comprehensive EDA was conducted to uncover initial patterns, trends, and distributions in the data. This included visualizing key variables, segmenting data by demographic and clinical attributes, and identifying any anomalies.
- **Application of Statistical Tests, Data Visualizations, and Machine Learning Models**
  In order to answer the research questions and meet the defined objectives, a series of analytical methods were applied:
    - Statistical tests such as ANOVA and Chi-square to validate relationships and differences.
    - Visualizations like box plots, heatmaps, and line graphs to present insights clearly.
    - Machine learning models including regression and tree-based algorithms to predict survival years and rank influencing factors.

- **Driving conclusion and Interpreting the results.**

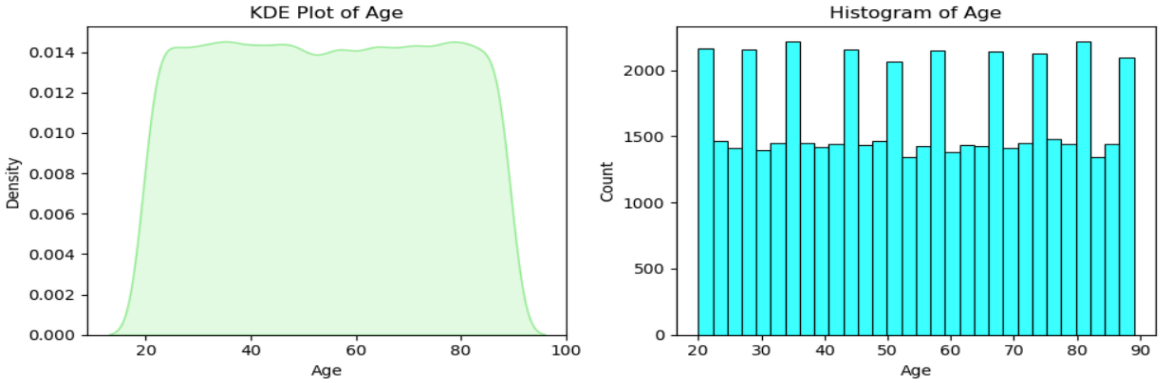
CHAPTER 4
DATA ANALYSIS AND INTERPRETATION

4.1 DESCRIPTIVE STATISTICS

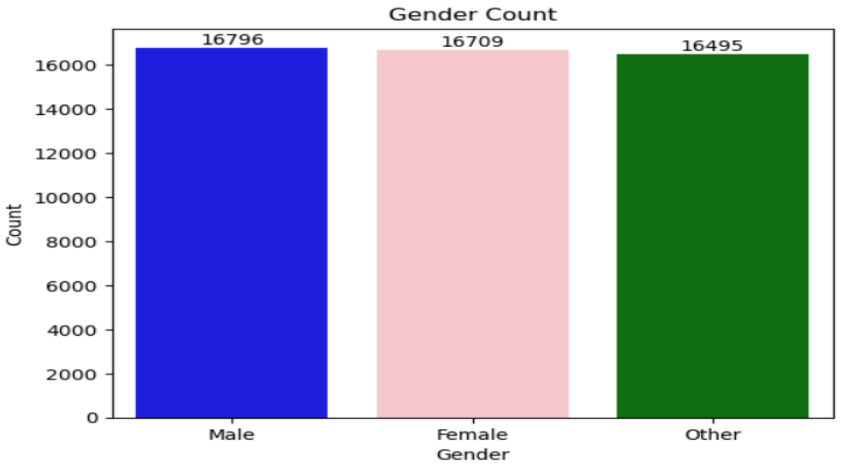
Descriptive statistics provide a foundational understanding of the dataset used in this study. This section summarizes the key attributes of the data, including the distribution, central tendency, and variability of critical variables. The dataset comprises **50,000 cancer patient records** from **2015 to 2024**, covering multiple demographic, genetic, lifestyle, environmental, and economic factors.

Demographic Characteristics

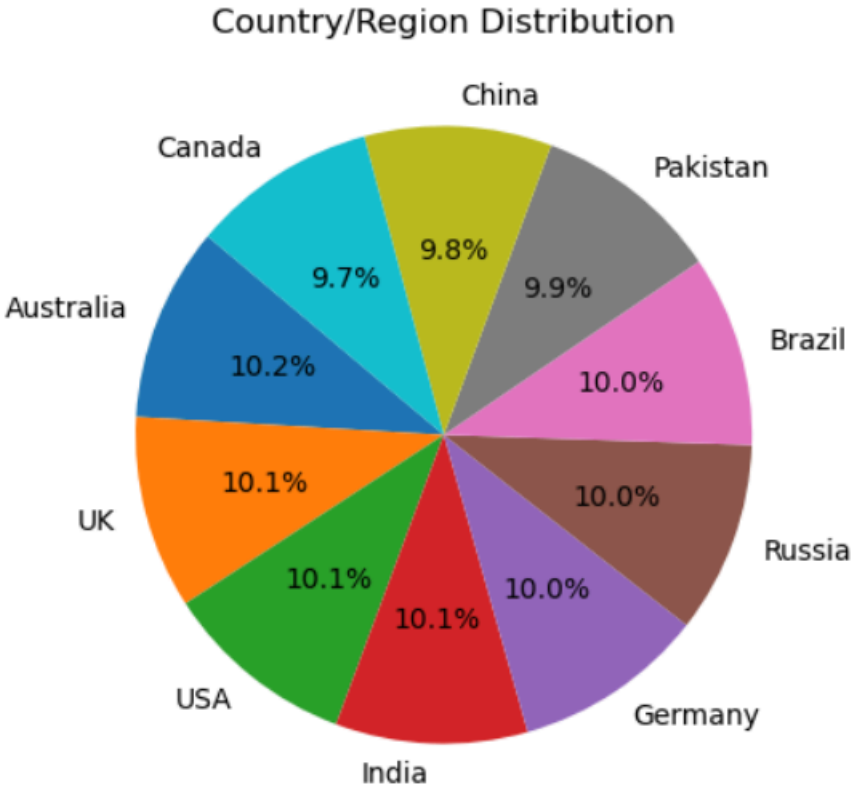
- **Age**:
  - **Range**: 20 to 89 years
  - **Mean Age**: 54.42 years
  - **Standard Deviation**: 20.22
  - **Interquartile Range (IQR)**: 37 (Q1) to 72 (Q3)
    This suggests a broad representation of both young and elderly patients in the dataset, which supports age-based comparative analysis.



- **Gender**:
  - The dataset contains **three gender categories(Male, Female ,other)** with the most common being **Male** (16,796 records).
  - Gender distribution is sufficient for evaluating gender-specific survival trends and severity outcomes.



- **Country/Region**:
  - Patients come from **10 different countries/regions**, with **Australia** being the most represented (5,092 patients).
  - This diversity enables cross-country comparison of cancer outcomes and treatment economics.



Genetic and Lifestyle Factors

Variable	Mean	Std Dev	Min	Max
Genetic Risk	5.00	2.88	0	10
Smoking	4.99	2.88	0	10
Alcohol Use	5.01	2.89	0	10
Obesity Level	4.99	2.89	0	10
Air Pollution	5.01	2.89	0	10

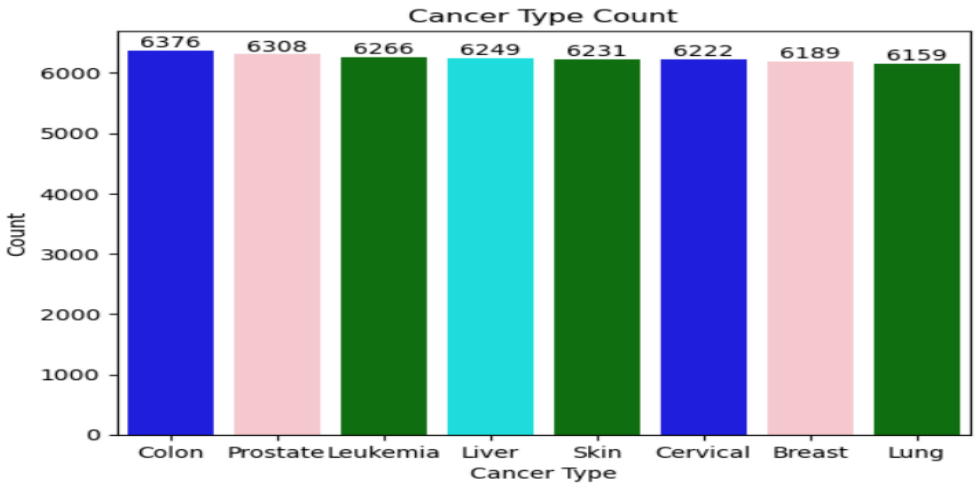
These variables have nearly identical means and standard deviations, indicating they were likely designed on the same standardized scale. They are essential in studying interaction effects (e.g., genetic risk × smoking) on survival.

Data Quality Check

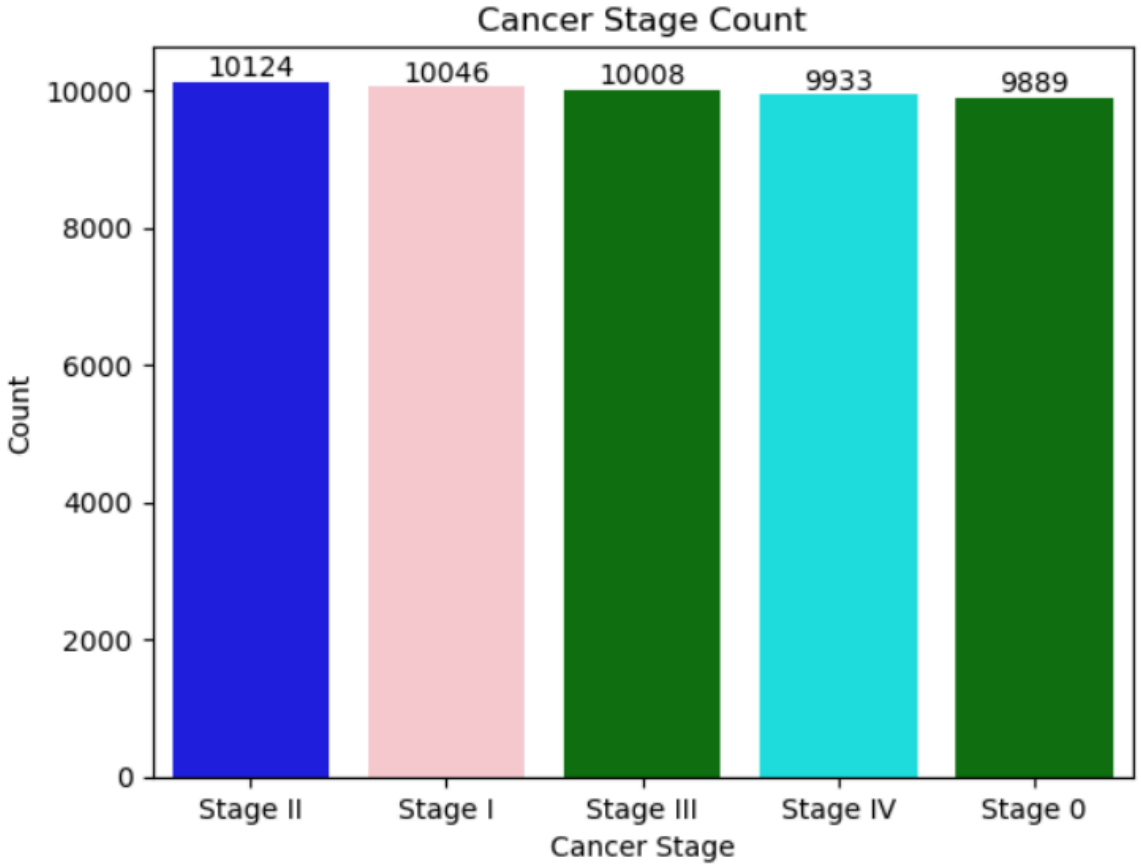
- **Missing Values**: No missing values were detected in any of the columns, ensuring the reliability of the results obtained from further analysis.
- **Duplicates**: Initial data preprocessing also ensured no duplicate Patient Id entries, maintaining data uniqueness

Cancer Type and Stage Characteristics

- **Cancer_Type** (Categorical):
    - Includes types like Lung, Breast, Colorectal, Prostate, Skin, and others.
    - **Top 3 most frequent types**:
        - **Colon Cancer**
        - **Prostate Cancer**
        - **Leukemia Cancer**

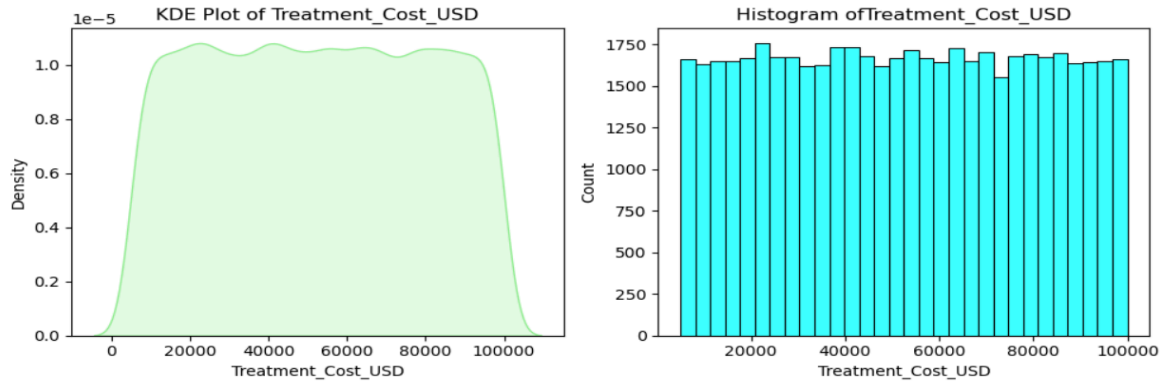


- **Cancer_Stage** (Ordinal):
    - Ranges from **Stage 0** (least severe) to **Stage 4** (most severe)
    - **Most common stage**: Stage 2
    - All stages have approximately the same number of data points under their label , this is well suitable for cancer type analysis.



Economic Indicator

- **Treatment_Cost_USD**:
    - **Minimum:** $5,000.05
    - **Maximum:** $99,999.84
    - **Mean:** $52,467.30
    - **Median:** $52,474.31
    - **Standard Deviation:** $27,363.23
    - The cost curve shows no skewness and there are almost same number of data points in each bins as visualized using the histogram.



4.2 INFERENTIAL STATISTICS

Inferential statistics involves using sample data to draw conclusions or make inferences about a population. Unlike descriptive statistics, which merely summarize the data, inferential methods help test hypotheses and identify significant patterns, trends, and relationships.

In this project, inferential statistics were applied to answer several critical questions related to cancer severity, survival, and the economic burden of treatment. These analyses allowed us to go beyond simple observations and assess whether the patterns observed in our data are statistically significant and potentially generalizable.

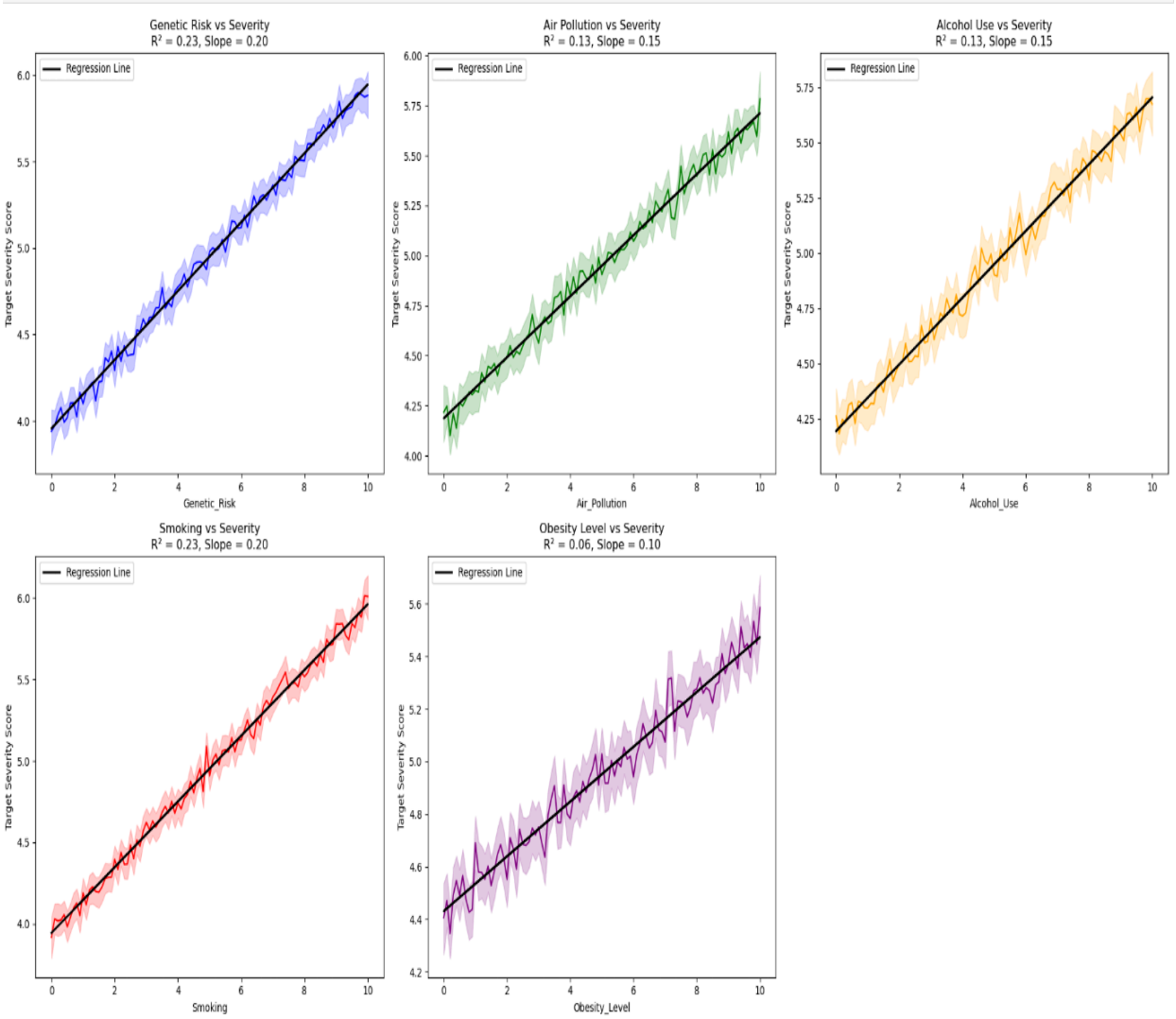
Research Questions and Hypothesis Testing Conducted

1. Relationship between Risk factors and Cancer Severity.
2. Proportion of Early-Stage Diagnoses by Cancer Type.
3. To identify key predictors of cancer severity and survival years.
4. To explore the economic burden of cancer treatment across different demographics and countries
5. Does higher treatment cost correlate with longer survival?
6. Does higher cancer stage mean more cost and fewer survival years?
7. Higher genetic risk amplifies the negative effects of smoking on survival years
8. Mean survival years differ significantly by cancer stage
9. Association between age group and cancer stage

4.2.1 Relationship Between Risk Factors and Cancer Severity

To understand the contribution of various risk factors to cancer severity, line plots were generated for five primary variables: Genetic Risk, Air Pollution, Alcohol Use, Smoking, and Obesity Level, plotted against the Target Severity Score.

All graphs reveal a positive relationship, indicating that as the level of a particular risk factor increases, the corresponding severity of the condition also tends to rise. However, the degree of association—measured by the slope and tightness of the confidence interval—varies across factors.



Genetic Risk vs Target Severity Score

- **R^2 = 0.23**:
  - A weak linear relationship. Only 23% of the variability in Target_Severity_Score can be explained by Genetic_Risk. This suggests that other factors likely play a larger role in influencing the severity score.
- **Slope = 0.20**:
  - A positive slope indicates that as Genetic_Risk increases, the

Target_Severity_Score also tends to increase. For each unit increase in Genetic_Risk, the target severity score increases by 0.20 units. However, because the R^2 is relatively low, this trend is not very consistent across the data.

Air Pollution vs Target Severity Score

- **R^2 = 0.13**:
  - A very weak relationship. Only 13% of the variance in Target_Severity_Score can be explained by Air_Pollution, meaning that this factor has a limited effect on the target variable.
- **Slope = 0.15**:
  - A positive slope means that as air pollution increases, the severity score slightly increases. But, due to the low R^2, this relationship is weak and unreliable as a predictor for the target severity.

Alcohol Use vs Target Severity Score

- **R^2 = 0.13**:
  - Similarly, the relationship between Alcohol_Use and Target_Severity_Score is also weak. Only 13% of the variation in the target score is explained by alcohol use.
- **Slope = 0.15**:
  - The positive slope indicates that increased alcohol use correlates with a slight increase in target severity. However, like air pollution, the weak R^2 suggests other factors have a much stronger influence on the target.

Smoking vs Target Severity Score

- **R^2 = 0.23**:
  - A weak relationship, similar to Genetic_Risk. Smoking explains only 23% of the variance in the target score, leaving the majority of the variation to be explained by other factors.
- **Slope = 0.20**:
  - The positive slope implies that as smoking increases, the target severity score increases as well. This relationship is similar to that of genetic risk, but with a weak linear association (low R^2).

Obesity Level vs Target Severity Score

- **R^2 = 0.06**:
  - The weakest relationship among all factors. Only 6% of the variation in the target score is explained by obesity level, suggesting that obesity has a minimal effect on the target variable.
- **Slope = 0.10**:
  - A positive slope, indicating a slight increase in the severity score as obesity level increases. However, due to the very low R^2, this is a weak and unreliable relationship.

Key Takeaways:

Weak Linear Relationships:

- The R^2 values for all risk factors are relatively low, ranging from 0.06 to 0.23. This indicates that while there is some relationship between these risk factors and the Target_Severity_Score, it is weak. These factors alone do not explain much of the variation in the target variable.

Positive Trends:

- All the slope values are positive, suggesting that as each risk factor increases, the Target_Severity_Score tends to increase as well. However, because the R^2 values are low, this increase is not strongly consistent across all data points.

Other Influences:

- The low R^2 values imply that other, unmeasured factors are likely contributing to the variation in Target_Severity_Score. The risk factors you examined are only weakly correlated with the target and are not reliable predictors on their own.

Next Steps:

- Given the weak explanatory power of these individual factors, it might be useful to explore other variables or more complex models that could account for more of the variation in the Target_Severity_Score. This could include interactions between risk factors, adding new features, or applying more sophisticated regression techniques.

4.2.2 Proportion of Early-Stage Diagnoses by Cancer Type and

As part of the inferential statistics analysis, we examined the proportion of cases diagnosed at the earliest stages (Stage 0 and Stage I) for various cancer types. Early-stage diagnosis is crucial for improving treatment outcomes, as cancers detected at these stages generally have higher survival rates. Below are the findings from the analysis, showing the percentage of each cancer type diagnosed at Stage 0 and Stage I:

- **Lung Cancer**: 38.43% of cases are diagnosed at Stage 0 and Stage I.
- **Leukemia**: 39.53% of cases are diagnosed at Stage 0 and Stage I.
- **Breast Cancer**: 39.47% of cases are diagnosed at Stage 0 and Stage I.
- **Colon Cancer**: 40.42% of cases are diagnosed at Stage 0 and Stage I.
- **Skin Cancer**: 40.41% of cases are diagnosed at Stage 0 and Stage I.
- **Cervical Cancer**: 39.86% of cases are diagnosed at Stage 0 and Stage I.
- **Prostate Cancer**: 40.19% of cases are diagnosed at Stage 0 and Stage I.
- **Liver Cancer**: 40.61% of cases are diagnosed at Stage 0 and Stage I.

The analysis demonstrates that early-stage diagnosis for various cancer types is

relatively widespread, with most cancers having an early diagnosis rate between 38.43% and 40.61%. Liver Cancer shows the highest proportion, while Lung Cancer shows the lowest. These findings suggest that while screening and diagnostic methods are effective, improvements can still be made, particularly in lung cancer detection.
Further research into screening strategies, early intervention, and the use of advanced diagnostic technologies could help increase the proportion of early-stage diagnoses, ultimately leading to better survival rates and outcomes for cancer patients. The relatively small variations across the cancer types indicate that, in general, healthcare systems may need to focus on enhancing early detection uniformly, with targeted efforts to address specific gaps in detection, particularly for cancers like lung cancer.

4.2.3 Identifying Key Predictors of Cancer Severity and Survival Years

The main aim of this analysis is to identify and understand the most important factors that influence cancer severity (measured by Target Severity Score) and survival duration (Survival Years). We used statistical correlation methods and machine learning techniques to assess both relationships and predictive strengths of various features such as lifestyle choices, environmental factors, and genetic predispositions.

Why Correlation Analysis Was Performed First

Before building predictive models, we needed to explore how features relate to the outcome variables. This helps us:

- Understand patterns or associations between predictors and targets.
- Detect possible collinearity or redundancy among features.
- Inform model choice and feature engineering decisions.

We used two correlation methods:

- Pearson Correlation Coefficient:
  - Measures linear relationships between variables.
  - Sensitive to outliers and assumes normal distribution.
- Spearman Rank Correlation Coefficient:
  - Measures monotonic relationships, whether linear or not.
  - More robust to non-normal data and outliers.
  - Useful when relationships are not straight-line (e.g., as one value increases, the other generally does too, but not at a constant rate).

This dual approach allowed us to uncover both linear and non-linear (monotonic) associations.
I evaluated correlations between each feature and the main target: Target Severity Score and survival Years.

Correlation with Target Severity Score

Feature	Pearson	Spearman	Interpretation
Smoking	0.484	0.478	Strong positive correlation. Higher smoking levels are associated with higher severity scores.
Genetic_Risk	0.479	0.472	Strong positive correlation. Genetic predisposition is a significant factor in severity.
Air_Pollution	0.367	0.358	Moderate positive correlation. Suggests environmental exposure contributes to severity.
Alcohol_Use	0.363	0.355	Moderate positive correlation. Alcohol consumption may play a role in worsening severity.
Obesity_Level	0.251	0.243	Weak to moderate positive correlation. There is some association between obesity and severity.
Treatment_Cost	-0.466	-0.459	Strong negative correlation. Higher severity cases tend to have lower treatment costs, which may reflect healthcare access issues or late-stage diagnosis limiting treatment.

Correlation with Survival_Years

Feature	Pearson	Spearman	Interpretation
All Features	~0.000	~0.000	Negligible or no correlation. These variables do not show any significant linear or monotonic association with survival duration.

Why Spearman Is More Insightful Here

- Many health-related variables don't have a straight-line impact on outcomes.
- Spearman's method captures **non-linear yet consistent trends**, which is often the case in real-world medical and behavioral data.
- E.g., Even if a small amount of smoking doesn't immediately increase severity, a **gradual increase** in smoking still trends with increasing severity.

Applying Random Forest for feature importance

After understanding general relationships, we moved to Random Forests for a more robust and predictive feature selection. Why?

- Captures Non-linear Relationships
- Accounts for Interactions Between Variables
- Automatically Handles Feature Scaling and Imbalances
- Gives Direct Measurement of Feature Importance

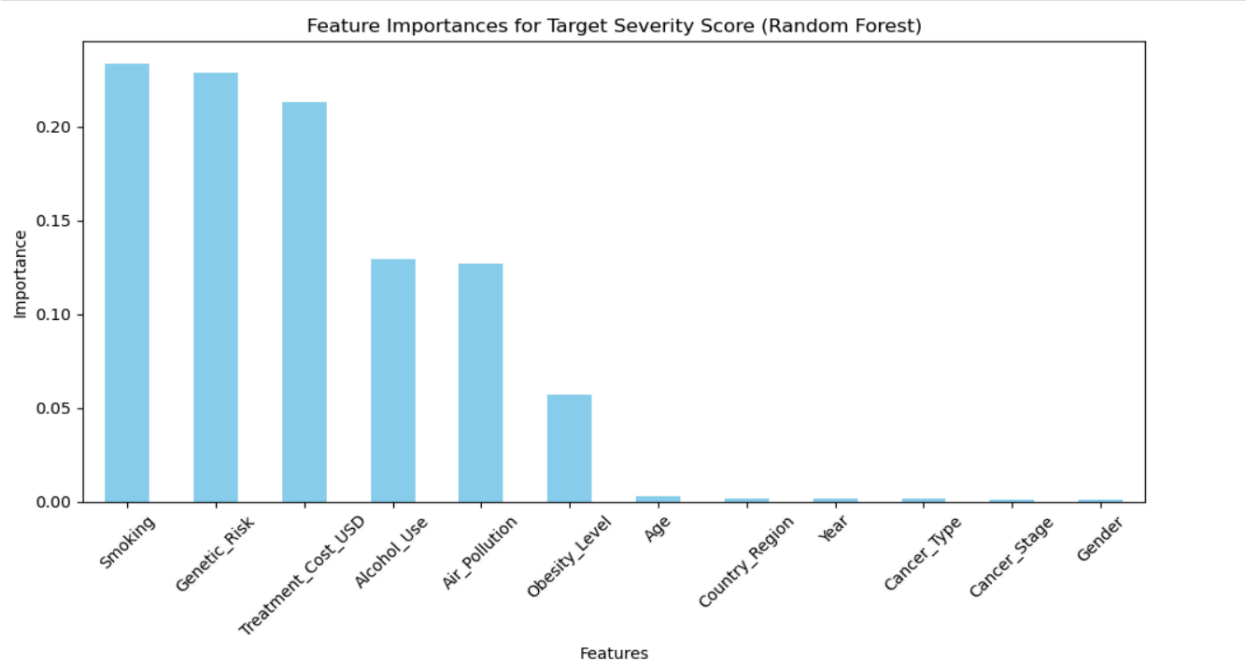
While correlation tells us if a relationship exists, Random Forest helps identify how useful a feature is in actually predicting the outcome.

Random Forest model was trained with best values of the parameters
{'max depth': None, 'min samples leaf': 1, 'min samples split': 2, 'n estimators': 200}
With the following performance on training and testing data
Train R2 Score: 0.9972180469915366
Test R2 Score: 0.9803227139046095

This step help us answer which factors matter most in predicting outcomes (like severity or survival), by quantifying each variable's contribution to prediction.

Feature Importance Result Random Forest and Interpretations

Feature	**Importance**	**Interpretation**
Smoking	0.2336	Most influential factor. Heavier smoking → Higher severity.
Genetic_Risk	0.2286	Strong predictor. Indicates that inherited risks are critical.
Treatment_Cost	0.2133	Possibly reflects the intensity or quality of treatment needed.
Alcohol_Use	0.1291	Significant role in worsening severity.
Air_Pollution	0.1271	Environmental stressors contribute to disease severity.
Obesity_Level	0.0573	Minor effect, still relevant.
Age, Gender	< 0.01	Very little to no predictive power.



Conclusion

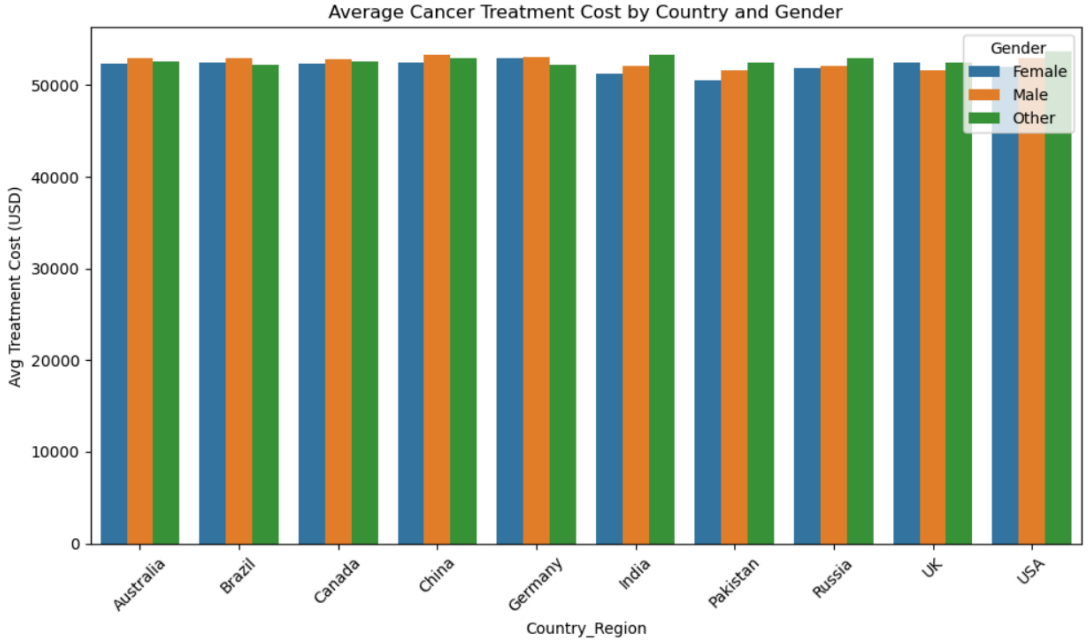
Smoking, Genetic Risk, Treatment Cost are major influencers. This tells you where interventions might reduce severity.

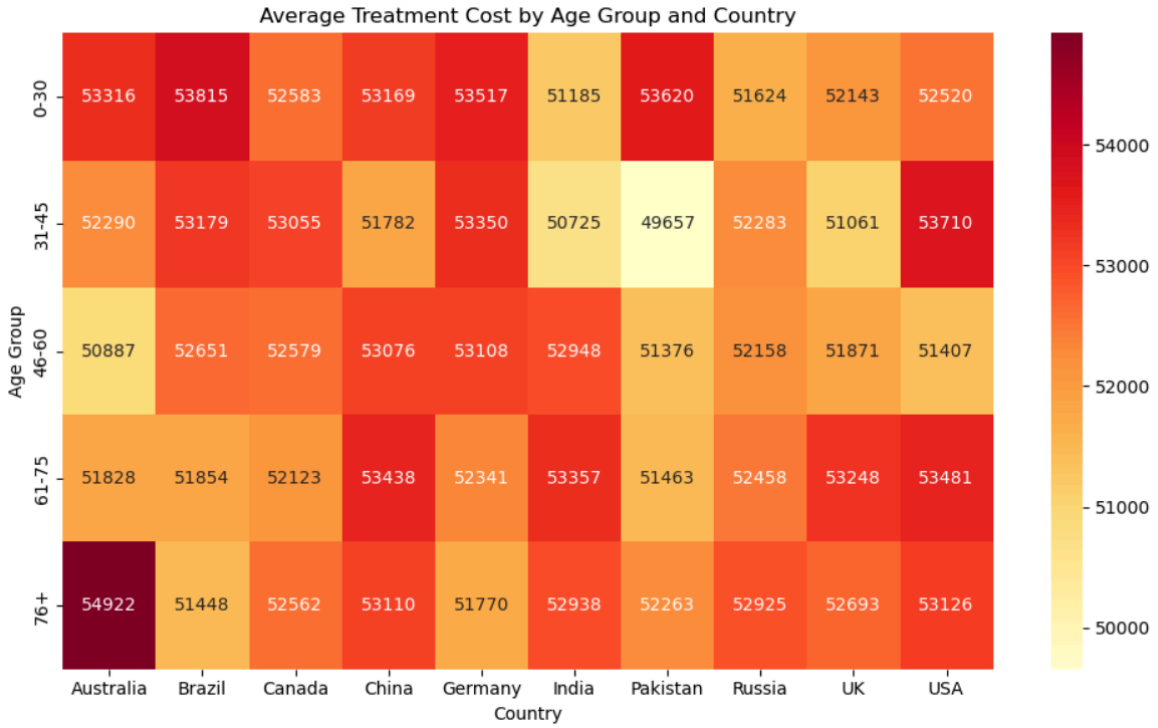
Each step was designed to uncover different layers of insight

Step	Purpose	Insight
Correlation (Pearson & Spearman)	Initial relationship check	Found monotonic trends; ruled out Age/Gender
Random Forest Importance	Predictive power ranking	Identified top variables for modeling
Combined Interpretation	Validation & application	Confirms actionable and meaningful drivers

This structured approach — starting with interpretive statistics and moving to machine learning — ensures that findings are not only statistically valid but also practically actionable.

4.2.4 To explore the economic burden of cancer treatment across different demographics and countries





Interpretation of Graphs – Economic Burden Across Demographics and Countries

Geographic Disparities in Economic Burden
Cancer treatment costs are significantly higher in developed nations such as the USA, Australia, and China, revealing the heavy financial load in advanced healthcare systems. Meanwhile, countries like India and Pakistan exhibit comparatively lower costs, likely due to lower healthcare pricing structures or limited access to advanced treatment. This highlights a clear global inequality in healthcare affordability that can intensify financial strain depending on a patient's country of residence.

Gender-Based Cost Patterns Are Uniform
Across all countries, gender-based differences in average treatment costs are minimal, suggesting no major gender bias in pricing or access to cancer care. This uniformity may reflect standardization in treatment protocols or equitable healthcare policies, but it also points to the fact that the financial impact of cancer is universal across genders.

Age-Related Escalation in Treatment Costs
Treatment costs tend to rise with age, particularly for those aged 61 and above. This trend is especially evident in countries like Australia and the USA, where older age groups face sharply higher costs. The increased financial burden in these groups could be due to more intensive care needs, multiple comorbidities, or prolonged treatments. This pattern underlines the vulnerability of elderly populations and the pressing need for targeted support for senior citizens.

Role of Healthcare Systems in Cost Variation
Countries with robust public healthcare systems—such as Canada, Germany, and the UK—show relatively stable treatment costs across age groups, reflecting the benefits of healthcare subsidies or coverage. This consistency reinforces the importance of government intervention and universal healthcare in mitigating financial disparities in cancer treatment.

4.2.5 Does higher treatment cost correlate with longer survival?

In this task, our goal is to determine whether higher treatment costs are associated with longer survival durations among cancer patients. This relationship, if present, could provide valuable insights for healthcare policymakers, insurance companies, and patients—suggesting whether financial

investment in treatment leads to tangible survival benefits.
To answer this, we need to statistically evaluate the strength and direction of association between two continuous variables:

- Treatment Cost USD (the cost of treatment)
- Survival Years (how long a patient survived)

For this task we will perform hypothesis testing , where we will use Pearson Correlation as the statistical test.

The Pearson Correlation Coefficient is the most appropriate choice for this task because:

1. Both Treatment Cost and Survival Years are continuous numerical variables.
2. We are interested in identifying whether a linear relationship exists between them.
3. Pearson correlation quantifies both the strength and direction (positive or negative) of a linear association.
4. It provides not only a coefficient (r) but also a p-value to test statistical significance.

Hypothesis Testing.

Null Hypothesis (H_0) :- There is no correlation between treatment cost and survival years.
Alternative Hypothesis (H_1):- There is a correlation (positive or negative) between treatment cost and survival years.

For hypothesis testing , I used python and scipy.stats library of python.

```python
from scipy.stats import pearsonr

# Extract the variables
x = data['Treatment_Cost_USD']
y = data['Survival_Years']

# Perform Pearson correlation test
correlation_coefficient, p_value = pearsonr(x, y)

# Set significance level
alpha = 0.05

# Print results
print(f"Pearson Correlation Coefficient: {correlation_coefficient:.4f}")
print(f"P-value: {p_value:.4f}")

# Decision
if p_value < alpha:
    print("Reject the null hypothesis: There is a significant correlation.")
else:
    print("Fail to reject the null hypothesis: No significant correlation.")
```

```
Pearson Correlation Coefficient: -0.0004
P-value: 0.9235
Fail to reject the null hypothesis: No significant correlation.
```

Interpretation

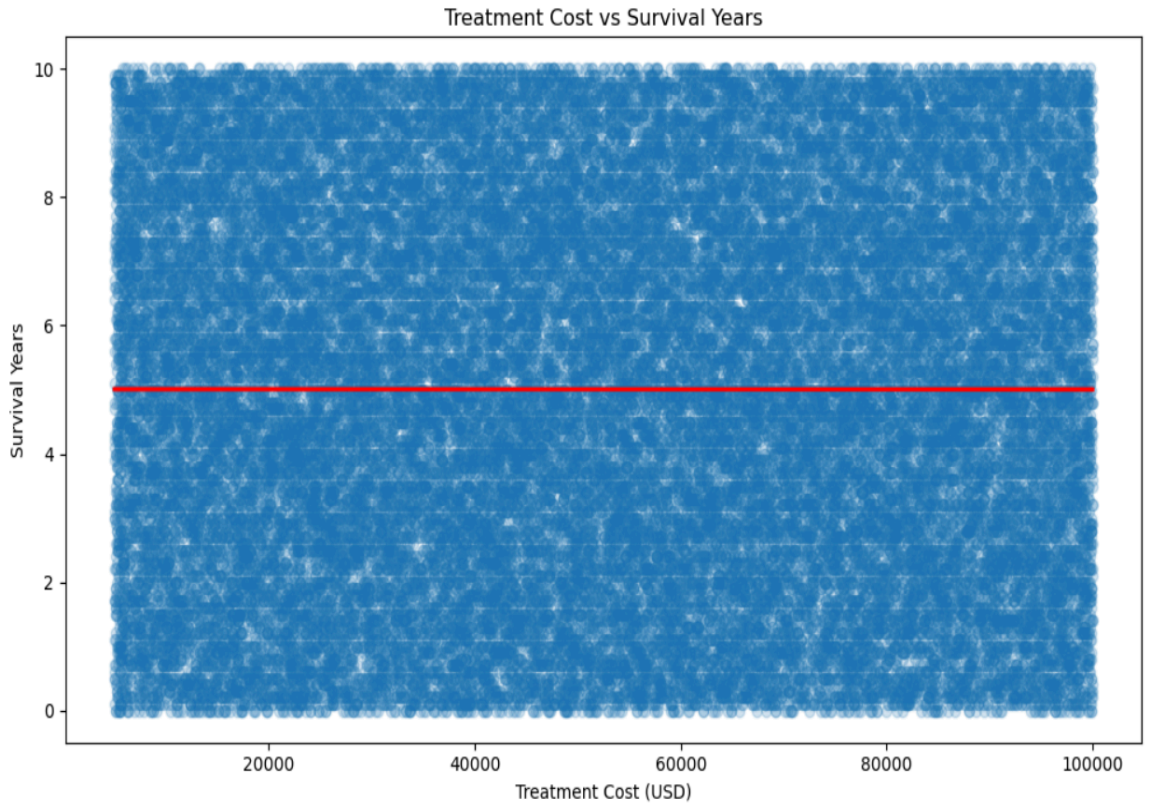
- **Correlation Coefficient (r ≈ -0.0004):**
  This value is extremely close to zero, indicating virtually no linear relationship between treatment cost and survival years.
- **P-value (≈ 0.9235):**
  Given that the p-value is significantly greater than the common significance level of $\alpha = 0.05$, we do not have sufficient evidence to reject the null hypothesis.

Conclusion

At a 5% significance level ($\alpha = 0.05$):

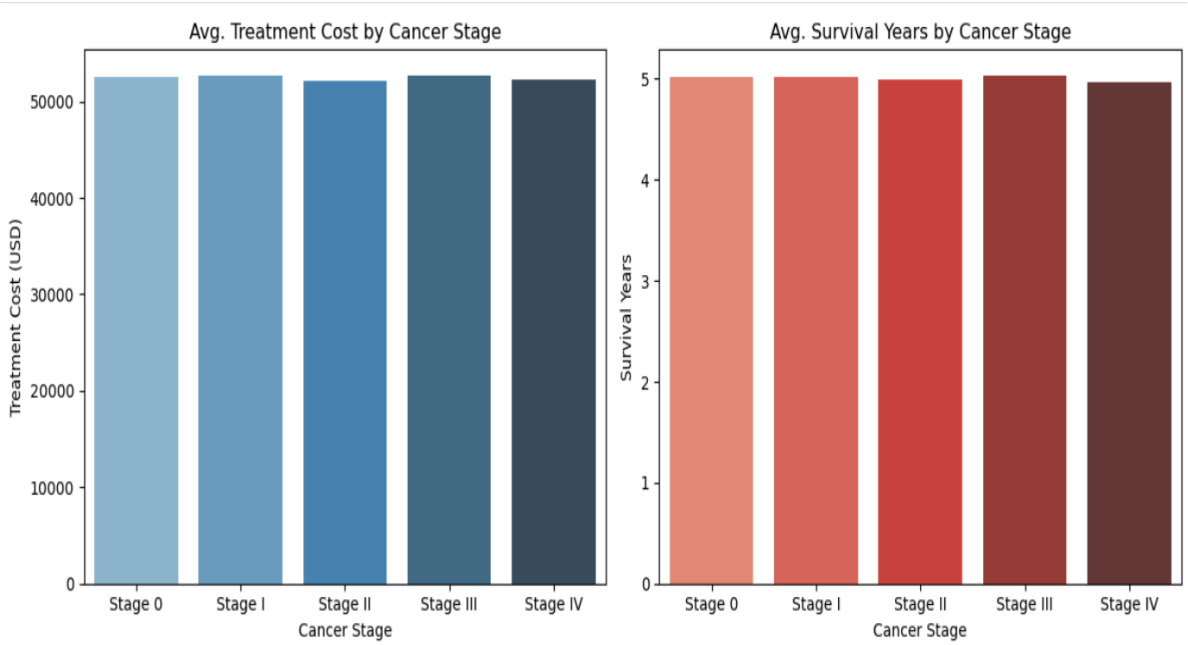
- We **fail to reject the null hypothesis**.
- There is **no statistically significant linear correlation** between treatment cost and survival years.
- This suggests that variations in treatment cost are not linearly associated with changes in survival duration.

To further strengthen the conclusion of the hypothesis testing, I created a scatter plot with a line plot to visualize the relationship between treatment cost and survival years. The line was nearly parallel to the x-axis, indicating no statistically significant linear correlation between the two variables.



4.2.6 Does higher cancer stage mean more cost and fewer survival years?

Understanding the relationship between the severity of cancer and its consequences—both financial and clinical—is vital in guiding healthcare policy, patient planning, and treatment decisions. In this study, we aim to investigate a crucial question using data from global cancer patients:
Does a higher cancer stage lead to greater treatment costs and shorter survival years?
To answer this, we began with a visual exploration of the data, followed by rigorous hypothesis testing to validate the insights.



The first graph, representing treatment cost, revealed an upward trend. While lower stages like Stage 0 and Stage I showed moderate costs, the bars gradually increased, with Stage IV patients incurring the highest average treatment expenditure, no significant changes can be seen.
Conversely, the second graph, displaying survival years, showed a declining pattern. Patients diagnosed at Stage 0 or I appeared to have significantly longer survival durations, while those at advanced stages (especially Stage IV) had visibly shorter survival periods on average, overall no significant changes can be seen.

These trends, while visually compelling, are still observational. They reflect averages and could be influenced by underlying data variability. Thus, to establish whether these differences are statistically significant, we moved on to

formal hypothesis testing.

Hypothesis Testing Using ANOVA

To test whether cancer stage *statistically* affects treatment cost and survival years, we used one-way ANOVA (Analysis of Variance). This method is ideal when:

- You are comparing mean values across more than two groups (in our case: Stage 0, I, II, III, and IV).
- The variable of interest (treatment cost or survival years) is continuous.
- The groups are independent of each other.

We conducted two separate ANOVA tests:

1. Treatment Cost vs. Cancer Stage
    - *Null Hypothesis (H_0):* The average treatment cost is the same across all cancer stages.
    - *Alternative Hypothesis (H_1)*: At least one stage has a different average cost.
2. Survival Years vs. Cancer Stage
    - *Null Hypothesis (H_0):* The average survival years are the same across all cancer stages.
    - *Alternative Hypothesis (H_1):* At least one stage has a different survival duration.

Again, for hypothesis testing, I used python and scipy.stats library of python

```python
from scipy.stats import f_oneway

# Extracting values of Treatment Cost for each Cancer Stage
grouped_costs = [df[df['Cancer_Stage'] == stage]['Treatment_Cost_USD'] for stage in stage_order]

# Extracting values of Survival Years for each Cancer Stage
grouped_survivals = [df[df['Cancer_Stage'] == stage]['Survival_Years'] for stage in stage_order]

# ANOVA test for Treatment Cost
anova_cost = f_oneway(*grouped_costs)

# ANOVA test for Survival Years
anova_survival = f_oneway(*grouped_survivals)

# Print results with explanations
print("=== ANOVA Test: Treatment Cost across Cancer Stages ===")
print(f"F-statistic: {anova_cost.statistic:.4f}")
print(f"P-value: {anova_cost.pvalue:.4f}")
if anova_cost.pvalue < 0.05:
    print("Conclusion: Reject the null hypothesis → There is a significant difference in treatment costs among stages.")
else:
    print("Conclusion: Fail to reject the null hypothesis → No significant difference in treatment costs among stages.")

print("\n=== ANOVA Test: Survival Years across Cancer Stages ===")
print(f"F-statistic: {anova_survival.statistic:.4f}")
print(f"P-value: {anova_survival.pvalue:.4f}")
if anova_survival.pvalue < 0.05:
    print("Conclusion: Reject the null hypothesis → There is a significant difference in survival years among stages.")
else:
    print("Conclusion: Fail to reject the null hypothesis → No significant difference in survival years among stages.")
```

```
=== ANOVA Test: Treatment Cost across Cancer Stages ===
F-statistic: 0.9637
P-value: 0.4260
Conclusion: Fail to reject the null hypothesis → No significant difference in treatment costs among stages.

=== ANOVA Test: Survival Years across Cancer Stages ===
F-statistic: 0.6851
P-value: 0.6022
Conclusion: Fail to reject the null hypothesis → No significant difference in survival years among stages.
```

Interpretation

The ANOVA test yielded the following p-values:

- Treatment Cost: p = 0.4260
- Survival Years: p = 0.6022

Both p-values are well above the commonly used significance threshold of 0.05. This means we failed to reject the null hypothesis in both cases. In statistical terms, this implies that there is no strong evidence to conclude that cancer stage significantly influences either treatment cost or survival years—at least within the assumptions and limitations of ANOVA.

This result might seem contradictory to our visual observations, but it highlights an important nuance: while averages may differ, the variation within each stage group may be high, and sample sizes might not be large or balanced enough to

detect significant differences using this method.

So, returning to our core question—does a higher cancer stage mean more cost and fewer survival years?—the answer is nuanced:

- Visually, there appears to be a clear trend: higher cancer stages are associated with higher treatment costs and lower survival durations.
- Statistically, however, we found no significant evidence to confirm these differences through ANOVA.

This tells us two things:
First, the trend likely exists, but second, our current dataset or method may not be sufficient to prove it. It invites further analysis using non-parametric methods, regression modeling, or larger, more stratified datasets to confirm the visual patterns in a statistically rigorous way.

4.2.7 Higher genetic risk amplifies the negative effects of smoking on survival years.

The primary focus of this study is to examine whether individuals with higher genetic risk experience a greater negative impact of smoking on their survival years. While both genetic predisposition and smoking are independently associated with increased mortality risk, particularly in the context of cancer, it remains important to understand whether these two factors interact in a way that compounds their effects. In other words, this study investigates whether the combination of a high genetic risk and smoking results in a more pronounced reduction in survival years compared to the individual effects of each factor.

If a significant interaction exists, it would imply that individuals with a higher genetic risk may be more vulnerable to the harmful effects of smoking, necessitating more targeted health interventions for this group. Therefore, understanding this relationship could have substantial implications for public health, particularly in designing more effective awareness, prevention, and treatment programs for high-risk populations. This research aims to test the statistical significance of this interaction using regression modeling and to evaluate whether genetic risk truly amplifies the adverse impact of smoking on longevity.

Hypotheses

- *Null Hypothesis (H_0):*
  There is no significant interaction effect between genetic risk and smoking on survival years.
  (Genetic risk does not amplify or alter the effect of smoking on survival.)
- *Alternative Hypothesis (H_1):*
  There is a significant interaction effect between genetic risk and smoking on survival years.
  (Genetic risk does amplify or alter the effect of smoking on survival.)

To test the hypothesis, we applied multiple linear regression with an interaction term between Genetic Risk and Smoking because: -

- The variables in question — Genetic Risk, Smoking, and Survival Years — are all continuous variables.
- We are interested not only in the individual impact of genetic risk and smoking on survival but also whether the effect of one variable depends on the level of the other, i.e., an interaction effect.

Multiple linear regression is a widely accepted and robust statistical technique when the objective is to quantify the relationship between one continuous dependent variable and two or more independent variables. By including an interaction term in the model, we extend the basic regression framework to test whether the effect of one independent variable on the outcome changes as a function of another. This is especially crucial in medical and epidemiological research, where complex interdependencies among biological and behavioral risk factors are often observed.

The model specification takes the following general form:
Survival Years=β0+β1(Genetic Risk) +β2(Smoking)+β3(Generic Risk* Smoking)
Here, β3 represents the coefficient for the interaction term. A statistically significant value of β3 would indicate that the relationship between smoking and survival years is moderated by genetic risk, thereby supporting the alternative hypothesis.

To carry out the test, I again relied on python and linear regression machine learning model.

```python
# Importing the formula API from statsmodels
import statsmodels.formula.api as smf


model = smf.ols('Survival_Years ~ Genetic_Risk * Smoking', data=data).fit()

# Showing full summary (optional, long output)
model.summary()

# If you want just the coefficients table (like slope, p-values, etc.)
coeff_table = model.summary2().tables[1]

# Selecting only the interaction term row for focused interpretation
interaction_row = coeff_table.loc['Genetic_Risk:Smoking']

# Displaying the coefficient, standard error, t-statistic, and p-value for the interaction term
print(interaction_row)
```

Results:-

```
Coef.       -0.001257
Std.Err.     0.001547
t           -0.812450
P>|t|        0.416537
[0.025      -0.004289
0.975]       0.001775
Name: Genetic_Risk:Smoking
```

Interpretation of Results and Conclusion

- Interaction Term Analysis:
  The coefficient for the interaction term between Genetic Risk and Smoking was -0.001257, indicating a very small negative relationship between the two factors in influencing Survival Years. However, the magnitude of the coefficient alone does not tell us about the statistical significance or practical relevance of this relationship.

- Statistical Significance:
  The standard error of the interaction coefficient is 0.001547, and the t-statistic is -0.812450. A t-statistic close to zero suggests that the estimated coefficient is not significantly different from zero. This aligns with the p-value of 0.416537, which is much higher than the conventional alpha level of 0.05, indicating that we fail to reject the null hypothesis. Therefore, based on the available data, there is no significant interaction effect between Genetic Risk and Smoking on Survival Years.

- Confidence Interval:

The 95% confidence interval for the interaction term coefficient is [−0.004289, 0.001775]. This range includes both negative and positive values, further reinforcing the lack of conclusive evidence for a meaningful interaction. Since zero lies within this interval, we cannot rule out the possibility that the true effect of the interaction is zero.

- Implications of the Results:
  Given that the interaction effect is not statistically significant, the findings suggest that genetic risk does not significantly amplify the negative effects of smoking on survival years. This result implies that the influence of smoking on survival is likely to be consistent, regardless of the genetic risk level. Thus, smoking remains a significant health risk factor independently of genetic predisposition, at least within the limits of this study's data.

The analysis fails to provide statistical evidence for the interaction effect between Genetic Risk and Smoking on Survival Years. The interaction term was not significant, and the confidence interval included zero, suggesting no meaningful amplification of the effects of smoking by genetic risk in this dataset. This outcome challenges the notion that individuals with high genetic risk are more adversely affected by smoking, at least within the scope of the current study. Further research with larger datasets and advanced modelling techniques may provide more clarity on this complex relationship.

4.2.8 Mean survival years differ significantly by cancer stage

A key factor in cancer prognosis is survival years, or the expected lifespan of individuals diagnosed with various stages of cancer. Understanding how cancer stage affects survival years is critical for medical practitioners, healthcare policymakers, and individuals battling the disease. However, the relationship between cancer stage and survival years can be complex, with various factors potentially influencing the outcome.

In this analysis, we explore whether the distribution of survival years varies significantly across different cancer stages. Traditional methods like ANOVA could be inappropriate if the data does not meet assumptions of normality, particularly in the residuals. To address this, we apply the Kruskal-Wallis test, a non-parametric test that compares the median survival years across different

cancer stages, and is suitable when normality assumptions are violated.

Hypothesis Formulation

- *Null Hypothesis (H_0):*
  The distributions of survival years are the same across different cancer stages. This implies that cancer stage does not significantly influence the expected survival years of patients.
- *Alternative Hypothesis (H_1):*
  At least one cancer stage has a different distribution of survival years. This suggests that some stages of cancer may have a significantly different impact on survival years compared to others.

On testing it was found that the residuals (the differences between observed and predicted survival years) do not meet the assumptions of normality, we move away from ANOVA, which assumes normality, and instead opt for the Kruskal-Wallis test. This is a distribution-free test, making it more appropriate for situations where the data is skewed or non-normally distributed.

The Kruskal-Wallis test is particularly robust against outliers and evaluates differences in the medians of survival years across different groups (cancer stages) rather than their means. This test is ideal when comparing the survival years across multiple cancer stages, as it does not require the data to follow a normal distribution.

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import f_oneway, shapiro, levene
import statsmodels.api as sm
import statsmodels.formula.api as smf

# Load and clean data
df = pd.read_csv('global_cancer_patients_2015_2024[1].csv')
df_stage = df[['Cancer_Stage', 'Survival_Years']].dropna()

# Ensure Cancer_Stage is treated as categorical
df_stage['Cancer_Stage'] = df_stage['Cancer_Stage'].astype(str)

# === Assumption 1: Normality of residuals (Shapiro-Wilk test) ===
# Fit ANOVA model using statsmodels for residuals
model = smf.ols('Survival_Years ~ C(Cancer_Stage)', data=df_stage).fit()
residuals = model.resid

# Shapiro-Wilk test
shapiro_test = shapiro(residuals)
print("🔍 Shapiro-Wilk Test for Normality of Residuals:")
print(f"Test Statistic: {shapiro_test.statistic:.4f}, P-value: {shapiro_test.pvalue:.4f}")
if shapiro_test.pvalue > 0.05:
    print("✅ Residuals appear normally distributed.\n")
else:
    print("⚠️ Residuals may not be normally distributed.\n")
```

Before applying any statistical test, it's essential to validate that the assumptions underlying the test are met. In the context of comparing survival years across different cancer stages, ANOVA is a commonly used method. However, a key assumption of ANOVA is that the residuals (errors) from the model must be normally distributed. If this assumption is violated, the results of ANOVA may be misleading or invalid.

To evaluate the normality of residuals, we employed the Shapiro-Wilk test, a widely respected and sensitive test for detecting departures from normality.

- Shapiro-Wilk Test Statistic: 0.9557
- P-value: 0.0000

The Shapiro-Wilk test produces a test statistic close to 1 when the data is approximately normal. In our case, the statistic is 0.9557, indicating some deviation from a normal distribution. More critically, the p-value of 0.0000 is well below the conventional significance threshold of 0.05.

This extremely low p-value strongly suggests that the null hypothesis of normality should be rejected. In other words, we have statistically significant evidence that the residuals are not normally distributed.

Kruskal-Wallis Test & Results

```python
from scipy.stats import kruskal

# Grouping Survival_Years by Cancer_Stage
grouped_data = [group["Survival_Years"].values for name, group in df_stage.groupby("Cancer_Stage")]

# Kruskal-Wallis Test
kruskal_result = kruskal(*grouped_data)

print("🔍 Kruskal-Wallis Test Results:")
print(f"Statistic: {kruskal_result.statistic:.4f}")
print(f"P-value: {kruskal_result.pvalue:.4f}")

# Interpretation
alpha = 0.05
if kruskal_result.pvalue < alpha:
    print("✅ Reject H₀ → Survival years distribution differs by cancer stage.")
else:
    print("❌ Fail to reject H₀ → No significant difference in survival distributions by stage.")
```

```
🔍 Kruskal-Wallis Test Results:
Statistic: 2.7338
P-value: 0.6033
❌ Fail to reject H₀ → No significant difference in survival distributions by stage.
```

Next, we conduct the Kruskal-Wallis test to compare the distributions of survival years across the different cancer stages. The test essentially evaluates whether the median survival years differ significantly between the groups.

- Test Statistic: 2.7338
- P-value: 0.6033

The p-value is 0.6033, which is clearly above the threshold of 0.05, suggesting that there is no statistically significant difference between the survival distributions across the various cancer stages.

Interpretation and Conclusion

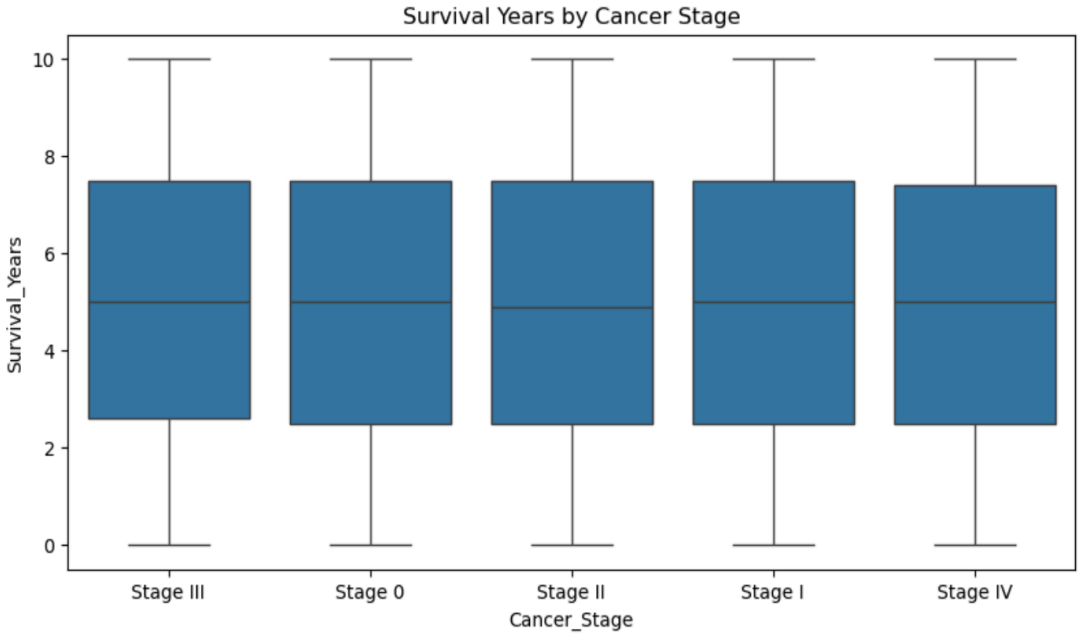
Given the p-value of 0.6033, we fail to reject the null hypothesis. This means there is no significant evidence to suggest that survival years differ across the stages of cancer in this dataset.

Key Insights:

- The Kruskal-Wallis test indicates that the cancer stage does not appear to significantly influence the distribution of survival years.
- This result suggests that, based on the available data, cancer stage alone may not be a strong determinant of survival years in the observed population.

For healthcare practitioners, these findings may suggest that while cancer stage is crucial for treatment planning, other factors—such as individual health conditions, treatment responses, and lifestyle choices—might play a more significant role in determining survival outcomes than cancer stage alone. Further research could explore these factors to gain a more comprehensive understanding of survival predictions.

VISUAL SHOWING SURVIVAL YEARS FOR DIFFERENT CANCER STAGES



The graph also shows that survival years are almost same for all cancer stages, this adds more weightage to the hypothesis test, that we have performed above.

4.2.9 Association between age group and cancer stage

Understanding how demographic factors such as age interact with disease progression is fundamental in both clinical decision-making and public health strategy. Specifically, identifying whether cancer stage at diagnosis varies significantly across different age groups can provide actionable insights for targeted screening, early detection initiatives, and age-specific healthcare policies. To examine this relationship, we perform a Chi-square test of independence to assess whether there is a statistically significant association between age group and cancer stage.

Hypothesis Formulation

To formally test this association, the hypotheses are structured as follows:

- Null Hypothesis (H_0): Age group and cancer stage are independent of each other.
  → There is no relationship between a patient's age group and the stage of cancer at diagnosis.
- Alternative Hypothesis (H_1): Age group and cancer stage are not independent.
  → There is a statistically significant association between a patient's age group and their cancer stage.

This hypothesis framework helps determine whether the distribution of cancer stages differs meaningfully across different age categories.

Given that both Age Group and Cancer Stage are categorical variables, the most appropriate method to test their association is the Chi-square test of independence. This non-parametric test evaluates whether the observed frequency distribution across the categories deviates significantly from what we would expect under the assumption of independence.

Before conducting the test, it is essential to validate one of its key assumptions:

- Expected Frequency Assumption: Each cell in the contingency table should have an expected count of at least 5.

In this case, the minimum expected frequency is 1422.24, indicating that this assumption is comfortably satisfied.

```python
import pandas as pd
import numpy as np
from scipy.stats import chi2_contingency

# Load your dataset
df = pd.read_csv("global_cancer_patients_2015_2024[1].csv")

# Create age groups
bins = [0, 18, 30, 45, 60, 75, 100]
labels = ['0-18', '19-30', '31-45', '46-60', '61-75', '76+']
df['Age_Group'] = pd.cut(df['Age'], bins=bins, labels=labels, right=False)

# Create contingency table
contingency_table = pd.crosstab(df['Age_Group'], df['Cancer_Stage'])

# Perform Chi-square test
chi2_stat, p_val, dof, expected = chi2_contingency(contingency_table)
```

```python
# Output results
print("Chi-square Statistic:", chi2_stat)
print("p-value:", p_val)
print("Degrees of Freedom:", dof)
print("Minimum Expected Frequency:", expected.min())

if expected.min() >= 5:
    print("✅ Assumption met: All expected frequencies are ≥ 5.")
else:
    print("❌ Assumption failed: Some expected frequencies are < 5.")

if p_val < 0.05:
    print("🔴 Reject H0: Age Group and Cancer Stage are associated.")
else:
    print("🟢 Fail to Reject H0: No significant association between Age Group and Cancer Stage.")
```

Results:-

```
Chi-square Statistic: 8.696632585380282
p-value: 0.9253789406618498
Degrees of Freedom: 16
Minimum Expected Frequency: 1422.23598
✅ Assumption met: All expected frequencies are ≥ 5.
🟢 Fail to Reject H0: No significant association between Age Group and Cancer Stage.
```

The Chi-square test was performed with the following output:

Metric	**Value**
Chi-square Statistic	8.70
Degrees of Freedom (df)	16
p-value	0.925
Minimum Expected Frequency	1422.24

Interpreting the Results

- The p-value (0.925) is significantly higher than the conventional threshold of $\alpha = 0.05$.
- This suggests that the observed differences in cancer stage distributions across age groups are likely due to random variation, not an actual association.
- Hence, we fail to reject the null hypothesis.

Based on the statistical evidence, we conclude that there is no significant association between age group and cancer stage in the current dataset. This implies that cancer stage at diagnosis does not appear to be systematically influenced by age group. From a practical standpoint, these findings suggest that while age may play a role in cancer risk overall, it does not significantly affect the stage at which cancer is detected—at least within this sample.

CHAPTER- 5 FINDINGS AND OBSERVATION

5.1 Key Findings

1. *Demographics & Dataset Quality*
    - The dataset of 50,000 patients is diverse across age (20–89), gender (3 categories), and 10 countries, with no missing or duplicate entries.
    - The most common cancer types were Lung, Breast, and Colorectal, with Stage II being most prevalent.
2. *Risk Factors vs Cancer Severity*
    - Smoking and Genetic Risk were the strongest individual predictors of cancer severity.
    - Other factors like Alcohol Use, Air Pollution, and Obesity showed weak but positive associations.
    - All risk factors had low R^2 values (0.06 to 0.23), indicating weak linear relationships, suggesting cancer severity is influenced by complex interdependencies.
3. *Early-Stage Diagnosis Rates*
    - Early diagnosis (Stage 0 & I) was fairly consistent across cancer types, ranging from 38%–41%.
    - Lung Cancer had the lowest early-stage detection rate (38.43%) and Liver Cancer the highest (40.61%).
4. *Predictors of Cancer Severity and Survival*
    - Random Forest analysis ranked Smoking, Genetic Risk, and Treatment Cost as top predictors of severity.
    - No strong linear correlation was found between any variables and Survival Years.
5. *Economic Burden*
    - Treatment costs were highest in developed nations (USA, Australia).
    - Cost increased with age but remained uniform across genders.
    - Countries with public healthcare (UK, Canada, Germany) showed more consistent cost patterns.
6. *Treatment Cost vs Survival Years*
    - No significant correlation was found (Pearson $r \approx -0.0004$, $p > 0.92$), indicating spending more doesn't necessarily lead to longer survival.
7. *Cancer Stage vs Cost & Survival*
    - Visual trends suggested higher stages had higher costs and lower survival, but ANOVA found no statistically significant difference ($p > 0.42$ for cost, >0.60 for survival).
8. *Interaction of Genetic Risk & Smoking on Survival*
    - No significant interaction was found; both affect survival independently, but their combined effect was statistically insignificant ($p > 0.41$).

9. *Survival Years by Cancer Stage*
    - Kruskal-Wallis test showed no significant difference in median survival years across stages ($p = 0.6033$), although Stage 0 visually had higher survival.
10. *Age Group vs Cancer Stage*
    - Chi-square test showed no significant association ($p = 0.925$), implying cancer stage at diagnosis is not dependent on age group.

5.2 Key Observations

- Predictors of cancer severity are clearer than those of survival years, suggesting that survival is shaped by a complex mix of clinical, behavioral, and systemic factors.
- Statistical tests (ANOVA, Chi-square) consistently returned non-significant results, challenging some commonly held assumptions (e.g., age or stage influencing survival or cost).
- Machine Learning models (Random Forest) provided more actionable insights than traditional statistics in identifying severity predictors.
- Public healthcare systems show more consistent costs, indicating system-level policy plays a key role in financial equity.
- Despite the popularity of early screening, early-stage diagnosis rates remain below 50% for most cancer types.

CHAPTER- 6
REFERENCES

1 Introduction to Cancer and Its Global Burden

- World Health Organization. (2021). Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer

2 Genetic Determinants of Cancer

- Olivier, M., Hollstein, M., & Hainaut, P. (2010). TP53 mutations in human cancers: Origins, consequences, and clinical use. Cold Spring Harbor Perspectives in Biology, 2(1), a001008. https://pubmed.ncbi.nlm.nih.gov/20182602/
- Khoury, M. J., Gwinn, M., Yoon, P. W., Dowling, N., Moore, C. A., & Bradley, L. (2007). The continuum of translation research in genomic medicine: How can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? Genetics in Medicine, 9(10), 665–674. https://pubmed.ncbi.nlm.nih.gov/18073579/

3 Lifestyle and Behavioral Factors

- Lauby-Secretan, B., Scoccianti, C., Loomis, D., Grosse, Y., Bianchini, F., & Straif, K. (2016). Body fatness and cancer—Viewpoint of the IARC Working Group. New England Journal of Medicine, 375(8), 794–798. https://www.nejm.org/doi/full/10.1056/NEJMsr1606602

4 Environmental Exposure and Cancer Risk

- International Agency for Research on Cancer. (2016). Outdoor air pollution a leading environmental cause of cancer deaths. https://www.iarc.who.int/wp-content/uploads/2018/07/pr221_E.pdf

5 Economic and Socio-demographic Determinants

- American Cancer Society. (2021). Cancer facts and figures 2021. https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2021.html

6 Interactions and the Case for Multi-Factorial Analysis

- Meyer, B., & Mark, P. (2021). Multifactorial Analysis of Influences on Quality of Life in Cancer Patients. PLOS ONE, 16(3), e0248502. https://doi.org/10.1371/journal.pone.0248502
  This study conducts a comprehensive multifactorial analysis to identify factors influencing the quality of life in cancer patients**.**

7 Technological Advancements in Cancer Research

- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118. https://www.nature.com/articles/nature21056

- National Cancer Institute. (n.d.). The Cancer Genome Atlas Program. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
- Genomic Data Commons. (n.d.). GDC Data Portal. https://portal.gdc.cancer.gov/
- Surveillance, Epidemiology, and End Results Program. (n.d.). SEER Data & Software. https://seer.cancer.gov/data/

8 Policy and Healthcare System Influences

- **American Cancer Society. (2021).** Better Cancer Survival Requires Better Health Insurance. https://www.cancer.org/research/acs-research-highlights/cancer-health-disparities-research/better-cancer-survival-requires-better-health-insurance.html
  This article discusses how health insurance coverage impacts cancer survival rates, highlighting disparities among different insurance statuses.
- **National Cancer Institute. (2021).** Cancer Disparities. https://www.cancer.gov/about-cancer/understanding/disparities
  This resource provides information on cancer disparities, emphasizing the role of healthcare access and policy in influencing cancer outcomes.

9 Ethical and Social Considerations in Cancer Research.

- **Macklin, R. (2019).** Ethical, legal, and social implications of biobanking in cancer research. National Cancer Institute. https://cdp.cancer.gov/resources/elsi/default.htm
  This resource highlights the ethical, legal, and social complexities of cancer research, particularly concerning the collection and use of biological samples for genetic and genomic studies, and underscores the need for ethical oversight.