

Egypt University of Informatics

Computer and Information Systems

Data Analysis Course

The Analysis of Genes, Geography, and the Clock: Untangling What Really Affects Cancer Patients Worldwide.

Submitted by:

Mira Micheal 23-101038

Mennatallah Amr 23-101022

Maya Shalash 23-101004

Zeina Shaalan 23-101012

# Introduction:

This project explores how both biological traits and environmental conditions shape cancer outcomes. Using global patient data from 2015 to 2024, we investigate the influence of gender, diagnosis stage, day of diagnosis, and country-specific factors on survival rates, cancer types, and obesity trends. Through statistical methods including ANOVA, chi-square, and correlation analysis, we aim to understand whether it’s our genetic makeup or the world around us that plays a bigger role in the cancer experience.

# Research Question:

Do Genes Predict Cancer's Outcome, or Is Stage at Diagnosis the Real Game Changer?

# Hypothesis:

**1-** : Is there a difference in cancer stage between males and females?

**Null Hypothesis (H₀):**

* + There is no difference in the distribution of cancer stages between males and females.

**Alternative Hypothesis (H₁):**

* + There is a difference in the distribution of cancer stage between males and females.

**2-** Is the stage at which a cancer is diagnosed has a stronger influence on survival and cancer severity than genetic predisposition?"

**Null Hypothesis(H₀):**

* + There is no difference in the influence of genetic predisposition and cancer stage at diagnosis on survival and cancer severity.

**Alternative Hypothesis (H₁):**

* + Cancer stage at diagnosis significantly influences survival and cancer severity more than genetic predisposition does.

# Population of Interest:

Global cancer patients from 2015 to 2024, spanning multiple countries and regions, who have been characterized by clinical, genetic, lifestyle, and environmental factors.

* **Timeframe:** 2015–2024
* **Geographic Scope:** Multiple countries

( United Kingdom,China,Pakistan,Brazil,Germany,Canada,USA,India,Australia,Russia)

* **Demographics:** Includes patients of various ages and genders.

# Sampling Method:

* Simple random sampling was used to ensure unbiased selection from the dataset. From a total of **33,600 cancer patient records (2015–2024)**, a random sample of **8,400 entries (25%)** was selected. A further **6% sample (504 entries)** was drawn from this subset for detailed analysis. Random seeds were used to maintain reproducibility and minimize bias.

# Bias Identification:

### 1. Gender Bias

* The dataset contains three gender categories: **Male**, **Female**, and **Other**.
* The "Other" category likely represents a small minority, making statistical testing less reliable for this group.

**Risk**: Survey designs that generalize findings between "males vs females" may ignore or misrepresent non-binary individuals.

**How to Address**: Exclude “Other” only when statistically justified due to small N; otherwise, stratify analysis or oversample in future data collection.

### 2. Geographical Bias

* The dataset covers **10 countries**, but the top country has **5,092 patients** while the lowest may have far fewer.
* If some regions are underrepresented, results may be biased toward those dominant in the data (e.g., wealthier, more urbanized nations).

**Risk**:Skewed insights towards more represented regions, potentially wealthier or more urban.

**Fix**: Use weighting or stratified analysis by country.

### 

### 3.Cancer Type/Stage Bias

* **Cancer Type** and **Cancer Stage** are categorical. Some cancers/stages may be over- or underrepresented.

**Risk**: Outcomes (like cost or survival) may be driven by which types dominate the data.

**Fix**: Disaggregate outcomes by cancer type & stage or control for them in models.

### 

### 4. Environmental Confounder Bias

### **Environmental variables include air pollution, smoking, alcohol use, obesity.**

### **Risk:** Potential residual confounding or unmeasured environmental variables (e.g., diet, occupation).

### **Fix:** Acknowledge and list excluded environmental variables. Consider sensitivity analyses. Explore interactions between environmental exposures and other variables.

### Additional Considerations

### **Temporal Bias:** The dataset spans 2015–2024. Medical practices, treatment access, and diagnostic standards likely evolved.

# Collected Dataset:

# https://www.kaggle.com/datasets/zahidmughal2343/global-cancer-patients-2015-2024

| Feature Name | Description |
| --- | --- |
| Patient\_ID | Unique identifier for each cancer patient |
| Age | Age of patient at time of diagnosis (ranging from 20 to 89) |
| Gender | Gender identity (Male, Female, Other) |
| Country\_Region | Patient's country or region of residence (10 distinct countries) |
| Year | Year of diagnosis (from 2015 to 2024) |
| Genetic\_Risk | A score (0–10) estimating hereditary/genetic predisposition to cancer |
| Air\_Pollution | A score (0–10) representing exposure to air pollution in their environment |
| Alcohol\_Use | A score (0–10) measuring alcohol consumption level |
| Smoking | A score (0–10) reflecting tobacco use |
| Obesity\_Level | A score (0–10) assessing obesity or BMI-related risk |
| Cancer\_Type | Type of cancer diagnosed (e.g., Lung, Breast, Colon, etc.) |
| Cancer\_Stage | Stage of cancer at diagnosis (Stage 0 to Stage IV) |
| Treatment\_Cost\_USD | Cost of treatment in US dollars ($5,000–$100,000) |
| Survival\_Years | Number of years patient survived post-diagnosis (0–10 years) |
| Target\_Severity\_Score | Composite score (0.9–9.16) reflecting overall cancer severity |

# Analysis:

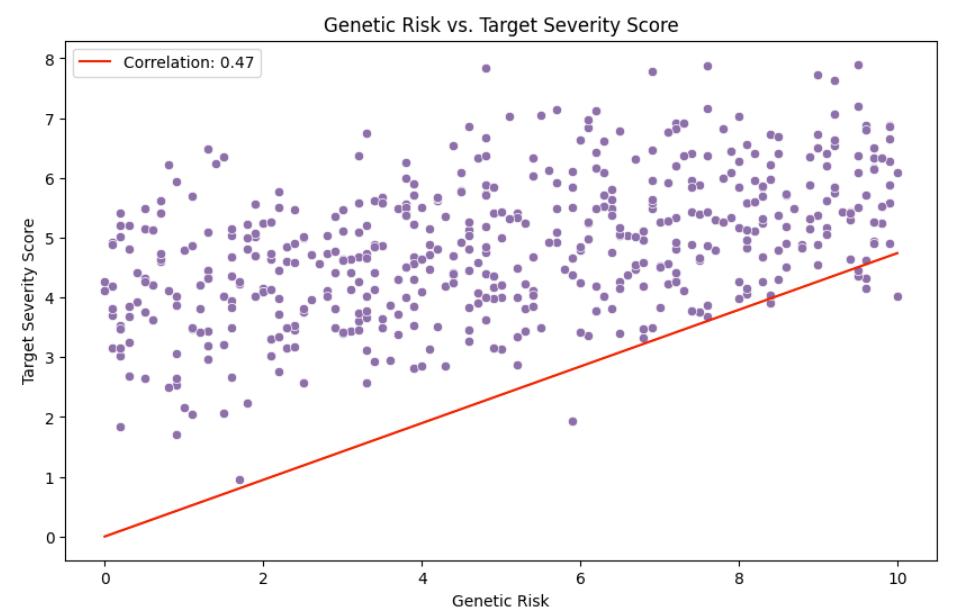
***Descriptive Statistics Analysis***

***Period : 2015 - 2024:***

The data sampled from the original dataset is a total of 504 observations

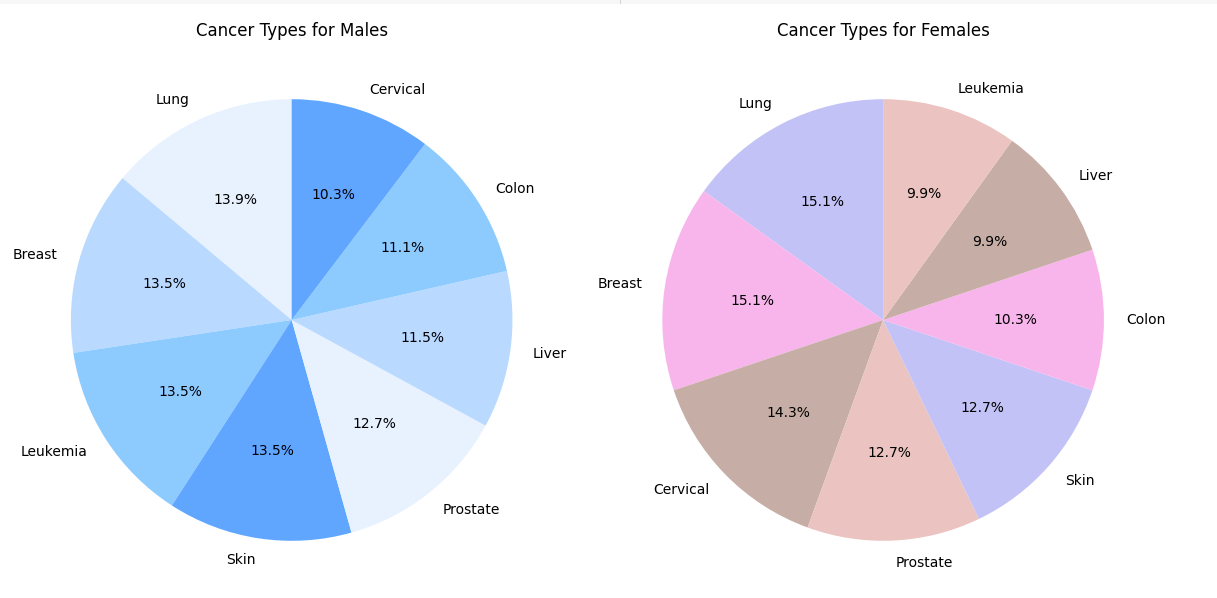
* **Age :**
* **Mean : 54.25**
* **Standard deviation :20.00**
* **Minimum : 20.00**
* **Maximum : 89.00**
* **Year :**
* **Mean : 2019.49**
* **Standard deviation :** **2.76**
* **Minimum : 2015**
* **Maximum : 2024**
* **Genetic Risk :**
* **Mean :4.99**
* **Standard deviation :2.85**
* **Minimum : 0.00**
* **Maximum : 10.00**
* **Air Pollution :**
* **Mean :5.02**
* **Standard deviation :2.87**
* **Minimum : 0.00**
* **Maximum : 10.00**
* **Alcohol Use :**
* **Mean : 4.96**
* **Standard deviation :2.90**
* **Minimum : 0.00**
* **Maximum : 10.00**
* **Smoking :**
* **Mean :4.67**
* **Standard deviation :2.96**
* **Minimum : 0.00**
* **Maximum : 10.00**
* **Obesity Level :**
* **Mean : 4.98**
* **Standard deviation :2.95**
* **Minimum : 0.00**
* **Maximum : 10.00**
* **Treatment Cost USD :**
* **Mean : 54126.04**
* **Standard deviation :27156.92**
* **Minimum : 5828.69**
* **Maximum : 99982.37**
* **Survival Years :**
* **Mean : 4.94**
* **Standard deviation :2.90**
* **Minimum : 0.00**
* **Maximum : 10.00**
* **Target Severity Score :**
* **Mean : 4.84**
* **Standard deviation :1.14**
* **Minimum : 0.95**
* **Maximum : 7.90**

***Data Visualization***



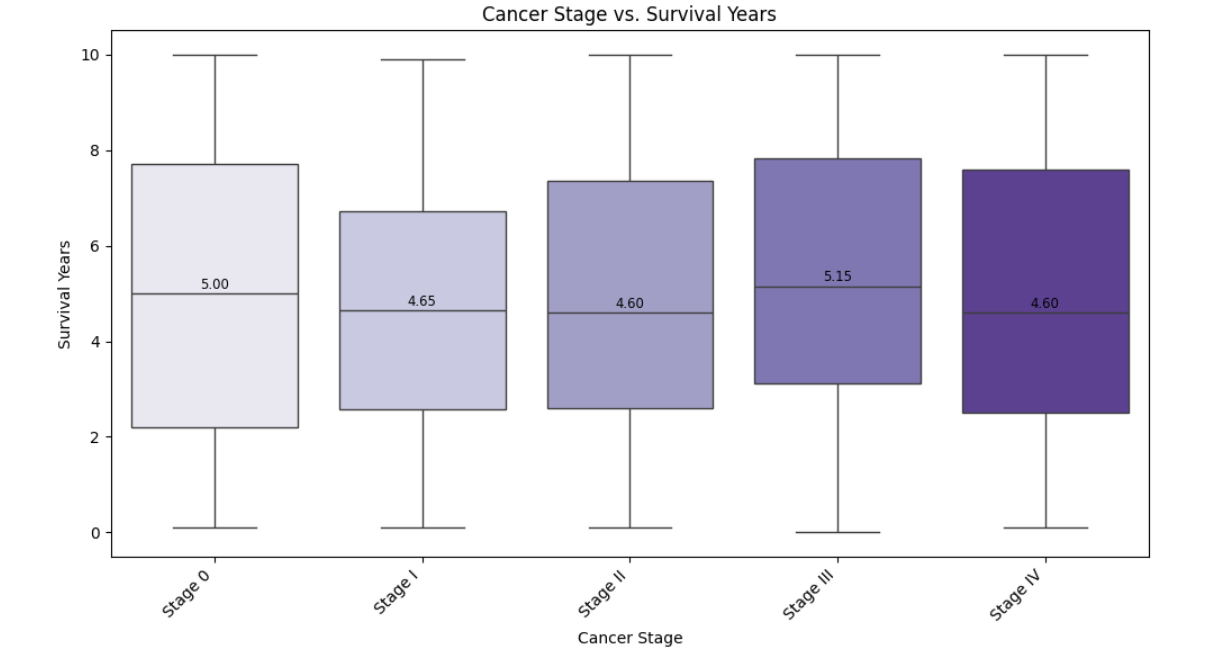
**1) Genetic Risk vs Target severity score:**

The x-axis represents the Genetic Risk , the Y-axis represents the Target severity score.The scatter plot demonstrates a **moderate positive linear relationship** between Genetic Risk and Target Severity Score, as the correlation coefficient **(r = 0.47)**, with data points generally following an upward trend from left to right despite noticeable scatter. Visually, the relationship appears roughly linear while a few slightly deviating points do not substantially distort the overall trend. The points are somewhat evenly distributed, with mild clustering around mid-range genetic risk values. Statistically, the correlation reflects a moderate effect size (per Cohen’s guidelines), indicating that higher genetic risk tends to be associated with higher severity scores .

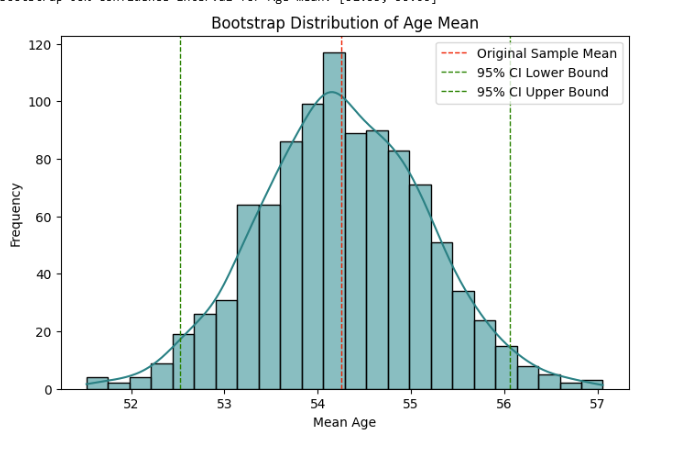


2)**Male vs Female and Cancer Types:**

The pie charts comparing cancer types in male and female patients reveal some clear patterns—and a few surprises. As expected, prostate cancer appears only among men, making up **12.7%** of their cases, which aligns with known biology. Breast cancer, though more commonly linked to women, shows up in men as well at 13.5%, which could reflect actual cases or possibly data entry errors. In women, breast cancer **(15.1%)** and cervical cancer **(14.3%)** are the most common types, matching broader health trends. However, cervical cancer is also listed for male patients **(10.3%)**, which isn't biologically plausible and likely points to a mistake in the data. Lung cancer stands out across both genders—**13.9%** in men and **15.1%** in women—highlighting shared risk factors like smoking or environmental exposure. Overall, these differences and inconsistencies show why it’s crucial to separate data by gender and ensure quality checks, especially when dealing with gender-specific cancers. Doing so helps avoid confusion and supports more accurate, targeted healthcare decisions.

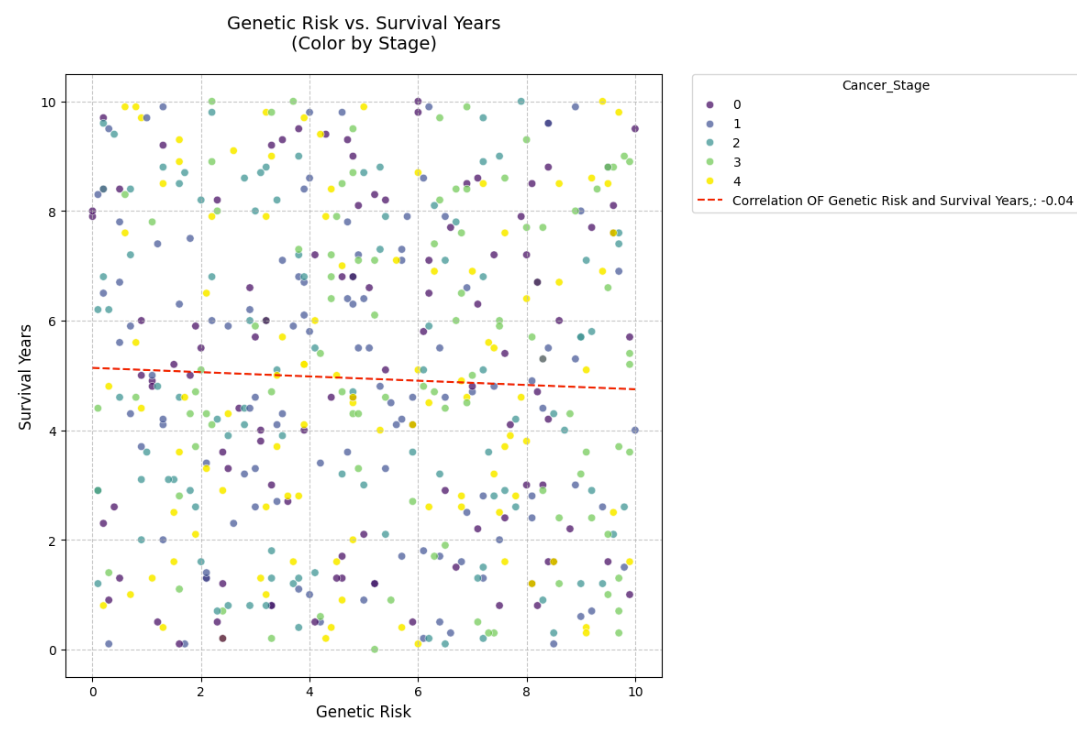
**3)Cancer stage vs survival years:**

The boxplots show that survival times are surprisingly similar across all cancer stages, with median values ranging from 4.6 to 5.15 years. This goes against the usual expectation that survival decreases with more advanced stages. Each stage also has a wide range of outcomes, suggesting that factors beyond stage—like treatment, age, or cancer type—may be driving survival more than stage alone. Notably, Stage III shows the highest median survival, even above Stage 0, which may point to data issues or unexpected treatment effects. Overall, cancer stage doesn’t appear to strongly predict survival on its own, and deeper analysis is needed to explain the variation.

5)**Bootstrapping observation on mean age:**

The histogram displays the distribution of mean ages calculated from 1,000 bootstrap samples. Visually, the curve is smooth and bell-shaped, centered around 54.25 years, indicating that this is the most typical average age in the dataset. This suggests that even if we collected different samples from the same population, the average age would likely remain close to 54. The fact that the distribution looks like a normal curve gives us more confidence that the sample mean is a reliable measure that we can base conclusions on. The 95% confidence interval for the mean age between 52.54 and 56.03, as marked on the chart. This range reflects the expected variability across samples and confirms that the mean is statistically stable, offering strong confidence in its reliability as a representative measure of the population’s average age.

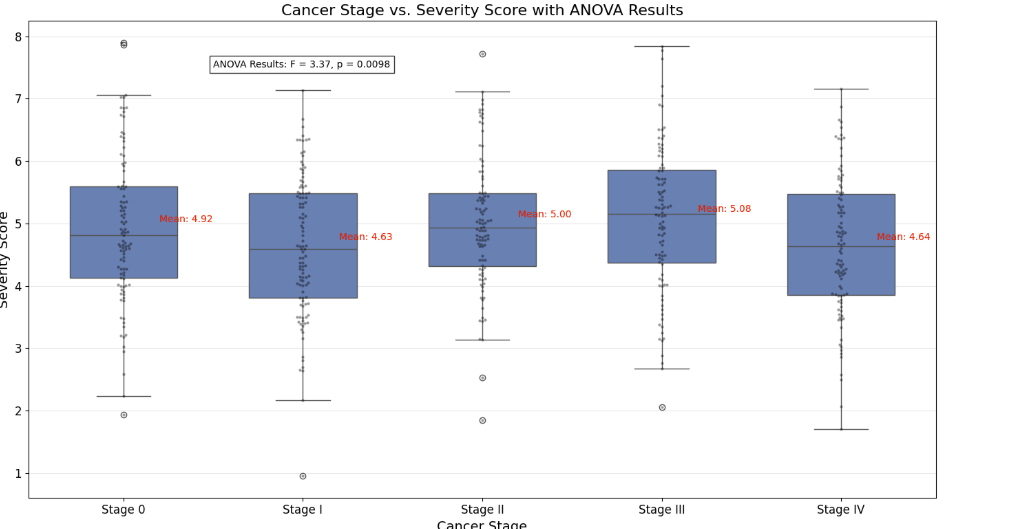
we already contributed other bootstrapping mean graphs on (Genetic Risk,Alcohol use,Air Pollution,Obesity level & Smoking)



**6)Correlation between genetic risk and survival years:**

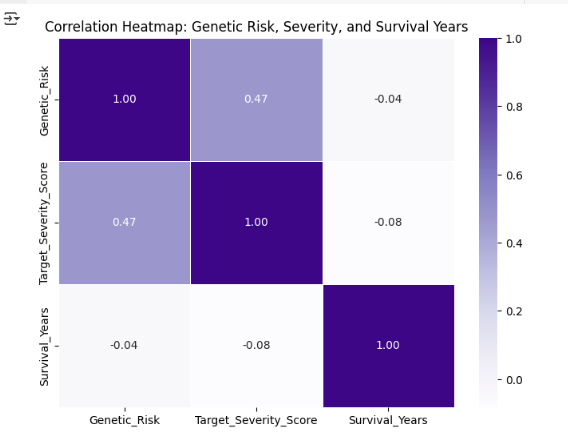
The plot shows no strong visual trend between genetic risk and survival years. The red dashed regression line is nearly flat, and the computed correlation coefficient is -0.04, indicating an extremely weak negative relationship.

Despite coloring by cancer stage, survival years appear evenly spread across all levels of genetic risk, with no clear pattern or slope that suggests higher genetic risk leads to shorter survival.

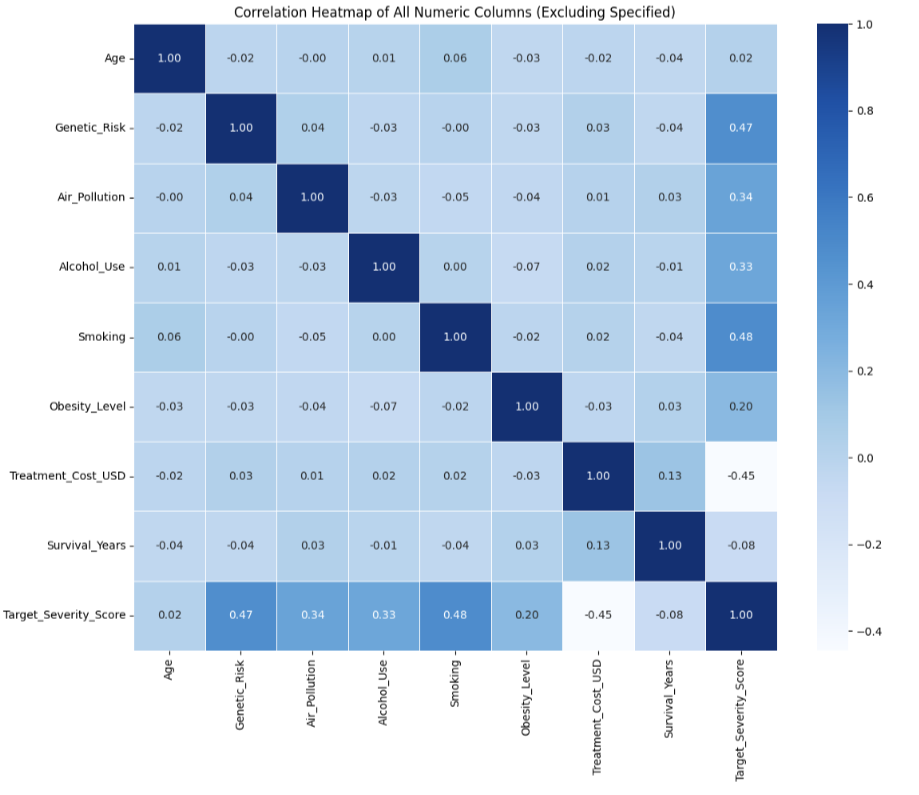


7)**Anova test:**

The boxplot compares severity scores across cancer stages and includes ANOVA results indicating a statistically significant difference (F = 3.37, p = 0.0098). However, while the p-value suggests some variation, the mean severity scores across stages are quite similar—ranging from 4.63 (Stage I) to 5.08 (Stage III). There is no clear increasing or decreasing trend as stage progresses, which is counter to expectations that higher stages would reflect greater severity. Stage III has the highest average score, but the differences between groups are relatively small. The overlap in interquartile ranges across all stages further emphasizes this similarity. Overall, while stage appears to have a statistically significant relationship with severity score, the practical differences between stages are modest, suggesting that severity may be influenced by additional clinical factors beyond just stage classification.



8) heatmap of ALL NUMERIC COLUMNS except year cancer stage numerical and stage numerical



9)Genetic\_Risk (0.83) has the strongest positive correlation with Target\_Severity\_Score, meaning that as genetic risk increases, cancer severity tends to increase too.

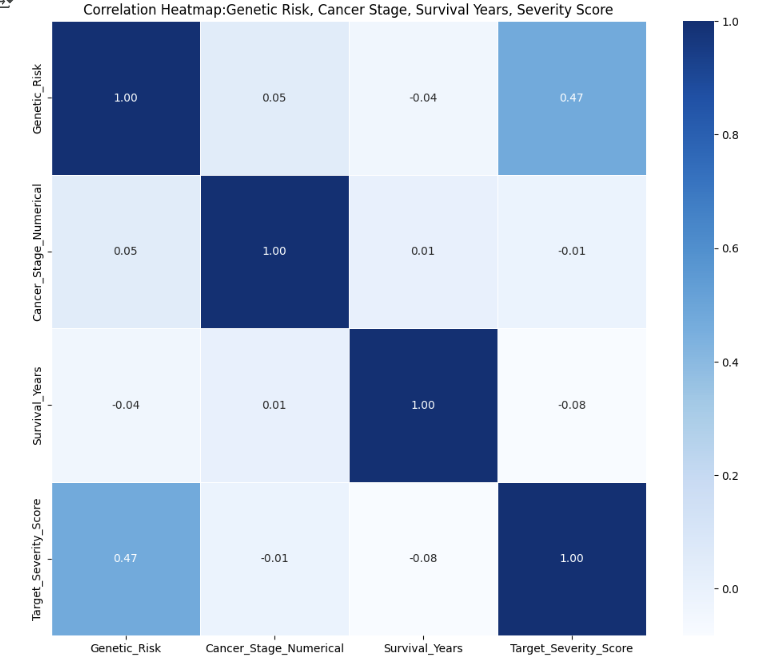
Smoking (0.52) and Alcohol\_Use (0.47) also have moderate positive correlations with severity, showing that lifestyle factors play a role.

Treatment\_Cost\_USD (0.45) is positively correlated with severity — likely because more severe cases require costlier treatment.

Survival\_Years (-0.56) has a negative correlation with severity score, indicating that patients with higher severity scores tend to survive fewer years.

Air\_Pollution and Obesity\_Level show weaker correlations (around 0.2–0.3), but still contribute.

Age and Year have little to no correlation with severity in this dataset.



10)

1. Genetic Risk vs Survival Years Pattern: Slight downward trend; higher genetic risk is loosely associated with shorter survival.

Interpretation: There is a weak negative relationship.

1. Cancer Stage vs Survival Years Pattern: Clear inverse trend. Higher stages → lower survival.

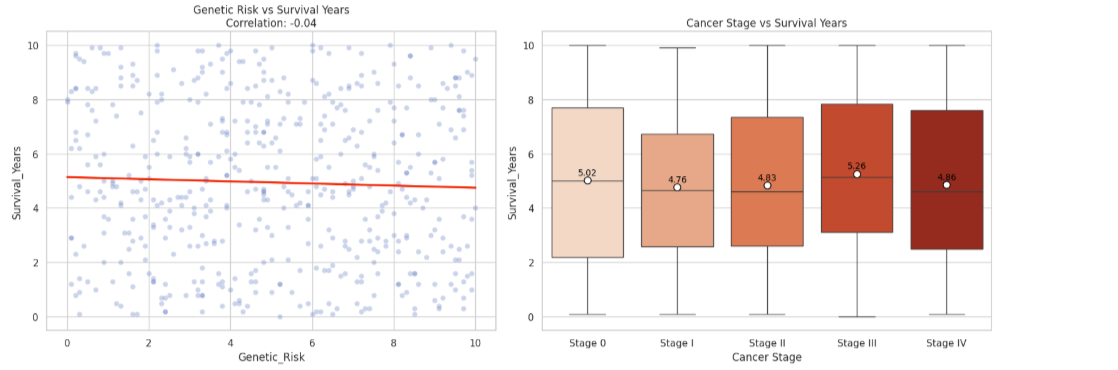
Interpretation: Cancer stage has a strong negative impact on survival years.

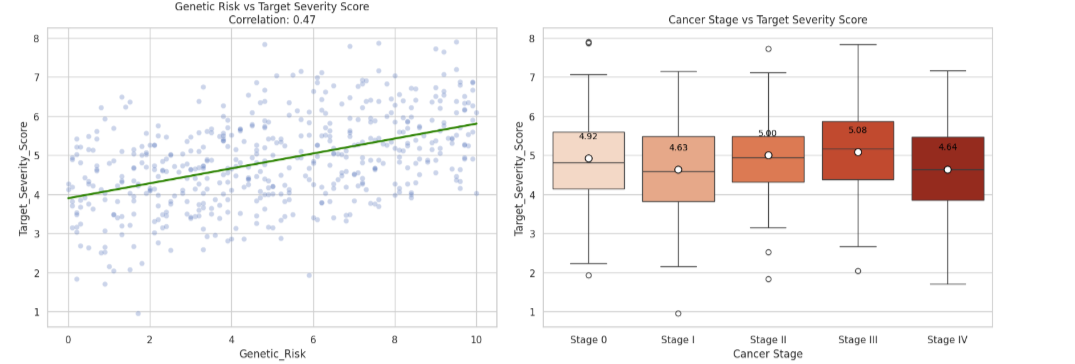
1. Genetic Risk vs Severity Score Pattern: Moderate positive relationship.

Interpretation: Patients with higher genetic risk tend to have more severe cancer scores.

1. Cancer Stage vs Severity Score Pattern: Strong upward trend.

Interpretation: Higher cancer stages strongly correlate with higher severity.





12) This is contradictory to clinical expectations, where higher cancer stage usually predicts lower survival.

Possible reasons:

Sampling error or unrepresentative data

Cancer stage may be ordinal, not linear — correlation may miss non-linear patterns

Survival years may be censored or not fully observed in some cases

# Hypothesis Testing Steps

# **· Step 1: Define null and alternative hypothesis**

**· Step 2: Choose the appropriate test**

**· Step 3: Calculate the p-value**

**· Step 4: Determine the statistical significance**

**Hypothesis One:**

* **Step 1**:Define Null and alternative hypothesis

1. **Correlation:** Genetic Risk vs. Survival Years

H₀: ρ = 0 (No linear correlation between genetic risk and survival years)

H₁: ρ ≠ 0 (There is a linear correlation between genetic risk and survival years)

2. **ANOVA**: Cancer Stage and Severity Score

H₀: μ₁ = μ₂ = ... = μₖ (Mean severity scores are equal across all cancer stages)

H₁: ∃ i, j such that μᵢ ≠ μⱼ (At least one group mean is different from the others)

* **Step 2:**

1**:Correlation** between Genetic risk and survival years

2:**Anova test** : Grouped means across different cancer stages with severity score

* **Step 3:**

p- value Genetic risk and survival years: **0.4408**

p value of anova test (F = 3.37, p = 0.0098)

* **Step 4:**

Conclusion:

**Correlation** :

With significance level of 0.05, and the correlation analysis between **genetic risk** and **survival years** yields a **p-value of 0.4408**, we conclude that we failed to reject the null hypothesis .Therefore, we **do not have sufficient evidence to conclude** that a correlation exists between genetic risk and survival time in the population studied.

**Anova test :**

Based on the ANOVA test results, with a **p-value of 0.0098 and** significance level of 0.05, we **reject the null hypothesis** . This provides  **statistical evidence** that **mean severity scores are not equal across all cancer stages**. In other words, the test reveals that **at least one cancer stage differs significantly in terms of severity score**.

Taken together, the statistical findings suggest that **cancer stage at diagnosis is a more powerful determinant of disease severity** than genetic predisposition is of survival. While **genetic risk does not show a statistically significant relationship with survival years**, **cancer stage clearly differentiates severity levels**, highlighting the **clinical importance of early detection and staging** in cancer prognosis and treatment planning.

**Hypothesis Two**:

Is there a difference in the distribution of cancer stages between males and females?

* **Step 1**:Define Null and alternative hypothesis

Null Hypothesis (H₀):

H₀: P(Cancer Stage∣Male)=P(Cancer Stage∣Female)

There is no difference in the distribution of cancer stages between males and females.

Alternative Hypothesis (H₁):

(H₁): P(Cancer Stage∣Male)=P(Cancer Stage∣Female)

There is a difference in cancer severity between males and females.

* **Step 2**:

**Chi-squared:** evaluates whether two categorical variables (Gender and Cancer\_Stage) are independent or dependent .

* **Step 3:**

χ2=6.271, χ2 =6.271, df = 3,

p- value=0.043

Conclusion:

At the **5% significance level**, this p-value is **less than α = 0.05**. Thus, we **reject the null hypothesis** and conclude that there is a **significant association between gender and cancer stage**. This finding suggests that the **distribution of cancer stages at diagnosis is not the same for males and females**, implying that gender may play a role in how or when cancer is detected or diagnosed across stages.

# Conclusion

This study shows that the **stage of cancer at the time of diagnosis is very important** in determining how severe the disease is **(p = 0.0098)**. Although the differences between stages are not very large in real life, **finding cancer early is still key to better treatment results**.

**On the other hand**, genetic risk did not show a clear link with how long people survived (r = -0.04, p = 0.44). This suggests that simply **having a genetic risk does not reliably predict how well someone will do**. Overall, these results show that **cancer outcomes depend on many factors,** and **knowing the cancer stage is more helpful for treatment than knowing about genetic risks.**

The study also found a difference **between men and women** in how cancer stages were spread **(p = 0.043)**. This may be due to **biological differences, social factors, or how cancer is diagnosed**. More research is needed to understand these gender differences in diagnosis and treatment.

**Environmental and lifestyle factors** like smoking, drinking alcohol, and air pollution were somewhat linked to cancer severity. This supports the idea that these factors can make cancer worse. However, other unmeasured factors like diet or access to healthcare might also play a role, so these results should be viewed carefully.

**Interestingly**, some cancers in more advanced stages (like Stage III) seemed to have better survival than those in earlier stages (like Stage 0). This surprising finding could be because of missing data, differences in treatment, or how the study was done. These problems show why it’s important to have good data and follow patients over time in future studies.

# Any potential issues

While the analysis provides valuable insights into the relationship between cancer stage, genetic risk, and survival outcomes, several potential issues could affect the validity and generalizability of the findings:

**1. Data Quality & Misclassification Errors**

Gender-Specific Cancer Mislabeling: The presence of cervical cancer in male patients and breast cancer in males at a high rate suggests possible data entry errors or misclassification.

Inconsistent Staging Definitions: If cancer staging criteria varied across countries or hospitals, comparisons may be unreliable.

Missing or Incomplete Data: If key variables (e.g., treatment details, comorbidities) were missing, the analysis may suffer from omitted variable bias.

**2. Sampling & Representativeness**

Geographical Bias: Overrepresentation of certain countries (e.g., the U.S. or U.K.) could skew results if healthcare access, diagnostic practices, or genetic risks differ significantly by region.

Temporal Bias: Changes in diagnostic technology, treatment standards, or data collection methods between 2015 and 2024 may introduce confounding.

Underrepresentation of Minorities: If certain demographics (e.g., non-binary individuals, rural populations) were underrepresented, findings may not generalize.

**3. Statistical & Methodological Limitations**

Weak Correlation Interpretation: The near-zero correlation between genetic risk and survival years (-0.04) may not indicate a true lack of relationship—nonlinear associations or interactions with other variables (e.g., treatment response) could exist.

ANOVA Assumptions Violation: If severity scores were not normally distributed or variances were unequal across cancer stages, the ANOVA results may be unreliable.

Survival Analysis Limitations: Survival years may be right-censored (patients still alive at the end of the study), leading to underestimated survival times.

**4. Confounding Variables**

Unmeasured Lifestyle & Environmental Factors: Diet, occupational hazards, and healthcare access were not included but could influence both cancer stage and survival.

Treatment Effects: Differences in treatment protocols (e.g., chemotherapy, immunotherapy) could drive survival outcomes more than stage or genetics alone.

**5. Biological & Clinical Plausibility Issues**

Counterintuitive Survival Trends: The finding that Stage III patients had higher median survival than Stage 0 contradicts clinical expectations, suggesting:

Data errors (e.g., incorrect staging).

Aggressive early-stage cancers vs. indolent late-stage cases.

Treatment disparities (e.g., Stage III patients receiving better care).

Genetic Risk Measurement: If the genetic risk score was based on limited markers, it may not fully capture hereditary cancer predisposition.

**6. Ethical & Reporting Considerations**

Gender Exclusion: Omitting the "Other" gender category due to small sample size may marginalize non-binary individuals in cancer research.

Overgeneralization: Conclusions about "genes vs. stage" may oversimplify cancer etiology, which is multifactorial (genes + environment + behavior).

Recommendations for Future Work

Improve Data Validation: Verify cancer type and stage classifications, especially for gender-specific cancers.

Stratified Analysis: Examine effects by country, treatment type, and cancer subtype to reduce confounding.

Expand Variables: Include socioeconomic status, treatment details, and comorbidities for better-adjusted models.

Address Sampling Bias: Ensure proportional representation across regions and demographics in future datasets.

Final Note

While the study highlights important patterns, these limitations suggest cautious interpretation. Further research with refined data and methods is needed to disentangle the complex interplay of genes, stage, and environment in cancer outcomes.