**📌 Problem Statement**

Lung cancer is one of the leading causes of mortality worldwide, and early detection plays a crucial role in improving survival rates. However, identifying high-risk patients and understanding key contributing factors (such as smoking habits, passive smoking, and cancer stage) remain critical challenges.

This project aims to analyze a lung cancer dataset to uncover patterns in patient demographics, risk factors, and disease progression. By leveraging data analysis techniques, we seek to provide insights that can help in early detection, risk assessment, and potential prevention strategies all over world.

**🎯 Objectives**

✅ **Analyzedthe distribution of lung cancer stages** across different patient groups.  
✅ **Identify key risk factors** (e.g., smoking, passive smoking) associated with lung cancer severity.  
✅ **Find correlations between lifestyle factors and cancer stage** using data visualization techniques.  
✅ **Use SQL, Power BI, Tableau, or Excel** to transform raw data into meaningful insights.  
✅ **Provide actionable recommendations** for healthcare professionals and policymakers to improve early detection efforts.

**About Data**

This project leverages two key datasets: **‘Lung Cancer Patient Records’** and **‘Environmental & Healthcare Factors’** to conduct a comprehensive analysis of lung cancer risk, survival rates, and healthcare accessibility. These datasets contain **206,918 rows and 24 columns**, providing detailed insights into patient demographics, lifestyle factors, smoking habits, environmental exposures, medical diagnoses, treatment types, and country-level lung cancer statistics.

This dataset contains individual patient details, including medical history, cancer diagnosis, and treatment progress. It helps in identifying risk factors and survival trends among lung cancer patients.

* **Patient ID** – Unique identifier for each patient.
* **Age** – Age of the patient.
* **Gender** – Male or Female.
* **Smoker** – Whether the patient smokes.
* **Years of Smoking** – Duration of smoking in years.
* **Cigarettes per Day** – Number of cigarettes smoked daily.
* **Passive Smoker** – Exposure to passive smoking.
* **Family History** – Presence of lung cancer in family history.
* **Lung Cancer Diagnosis** – Whether diagnosed with lung cancer.
* **Cancer Stage** – Stage I to IV of cancer at diagnosis.
* **Survival Years** – Years the patient survived after diagnosis.
* **Adenocarcinoma Type** – Type of lung cancer.
* **Country** – Country of residence.
* **Developed or Developing** – Country classification.
* **Population Size** – Country's population size.
* **Air Pollution Exposure** – Level of air pollution (Low, Medium, High).
* **Occupational Exposure** – Exposure to hazardous substances at work.
* **Indoor Pollution** – Exposure to indoor pollutants.
* **Healthcare Access** – Availability of healthcare for treatment.
* **Early Detection** – Whether cancer was detected early.
* **Treatment Type** – Type of cancer treatment received.
* **Annual Lung Cancer Deaths** – Number of annual lung cancer deaths in the country.
* **Lung Cancer Prevalence Rate** – Prevalence rate of lung cancer in the population.
* **Mortality Rate** – Percentage of non-surviving lung cancer patients.

📊 **Data Cleaning & Handling Missing Values**

**The dataset initially contained a significant number of missing values, particularly in two key columns:**

* Cancer Stage – Over 211,671 missing values out of 220,632, replaced with the placeholder "Unknown" to ensure consistency.
* Treatment Type – Contained a large proportion of missing values, which were replaced with "No Treatment" to indicate patients who may not have received treatment or whose treatment data was unavailable.

Aside from these two columns, the dataset was otherwise clean with no additional missing values. This ensures that the analysis remains reliable and minimizes bias due to missing data

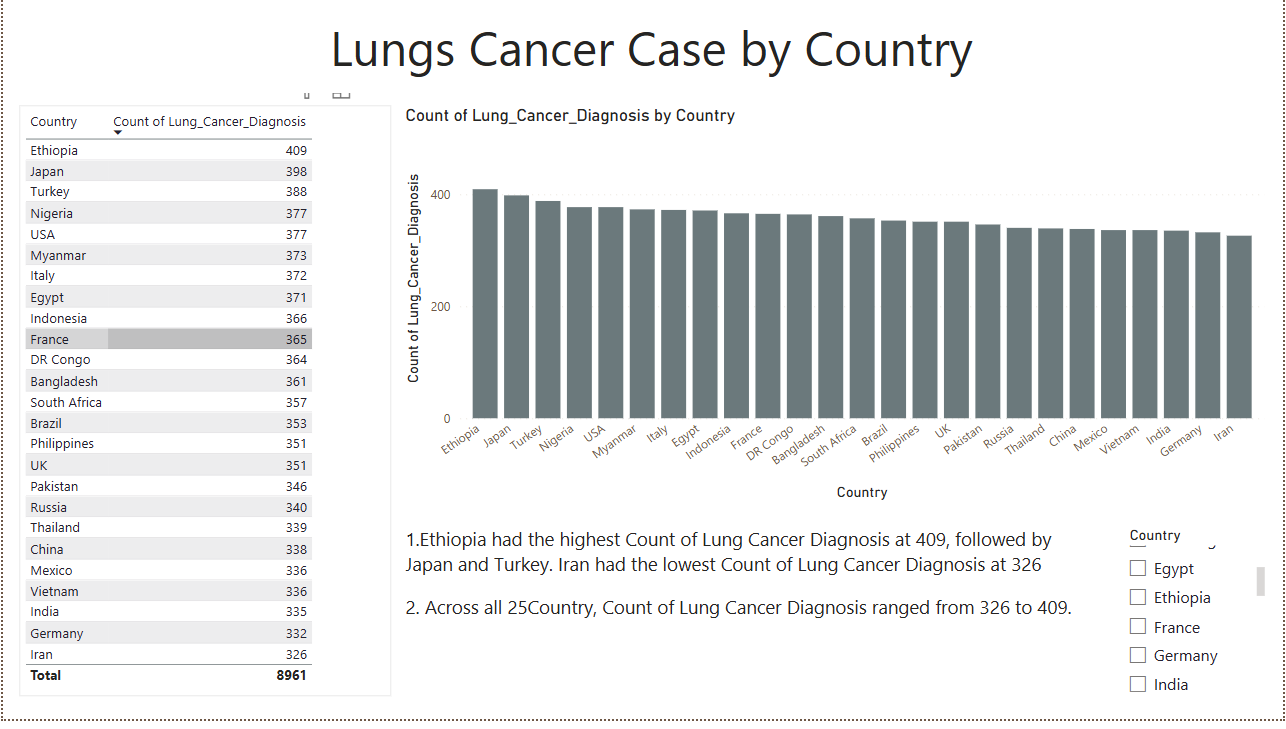
**🔍 Key Insights**

**1. Lung Cancer Diagnosis by Country**

1️. Ethiopia recorded the highest number of lung cancer diagnoses at 409 cases, followed by Japan and Turkey.

2️. Iran had the lowest number of lung cancer diagnoses at 326 cases.

3️. Along of 25 countries, the number of lung cancer diagnoses ranged between 326 and 409 cases, indicating a moderate variation in reported cases across different regions.



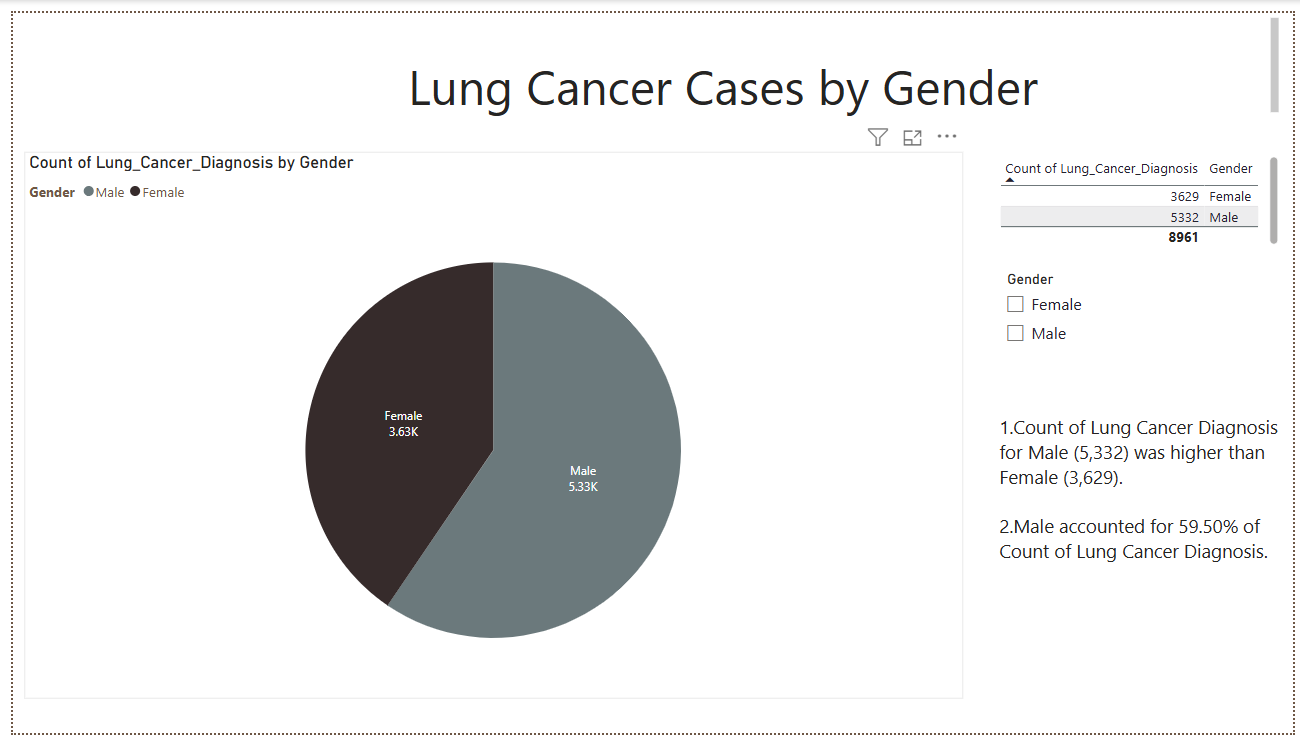
**2.Lung Cancer Cases by Gender**

1️. Higher Lung Cancer Cases in Males

* Males (5,332 cases) had a higher count of lung cancer diagnoses compared to females (3,629 cases).

2️. Gender Distribution of Lung Cancer Cases

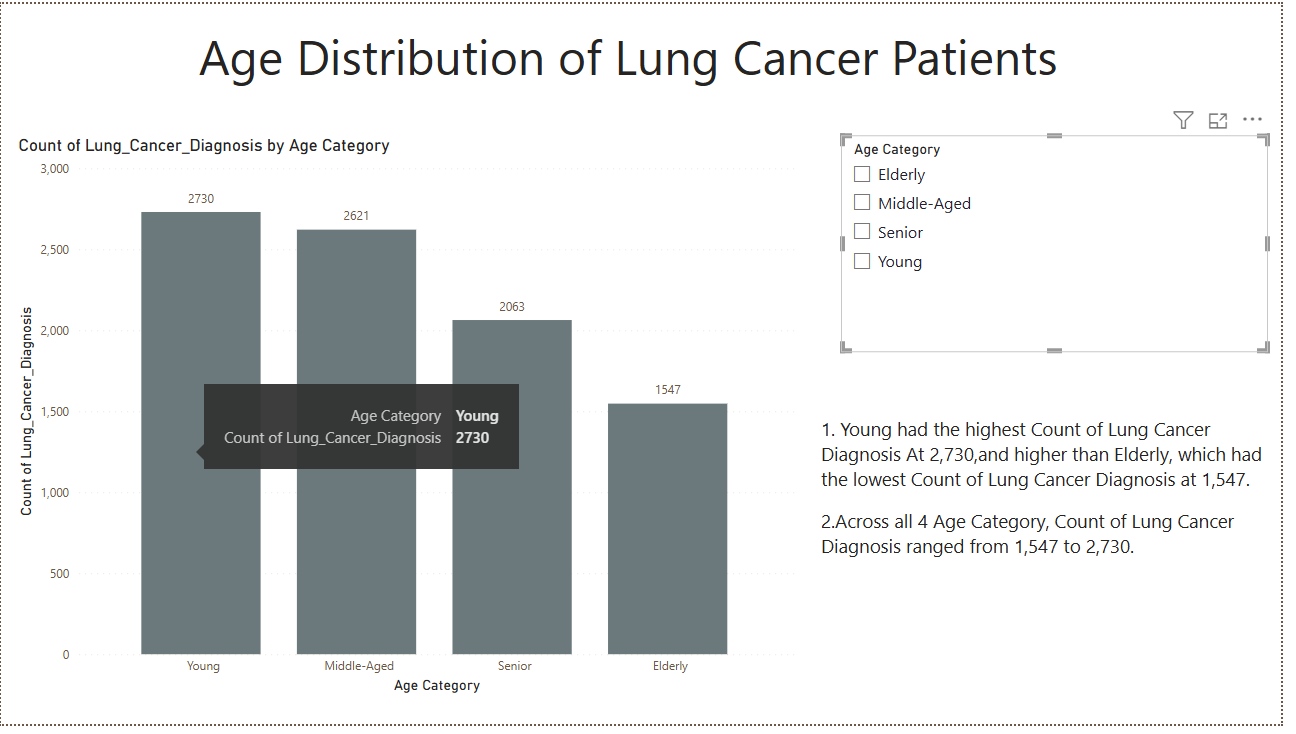
* Males accounted for 59.5% of total lung cancer diagnoses, indicating a higher risk factor compared to females.



**3. Age Distribution of Lung Cancer Patients**

1.Across all four age categories (Young, Young Senior, Middle-Aged, Elderly), Young (2,730 cases) recorded the highest lung cancer diagnoses, indicating a rising trend in early-age cases.

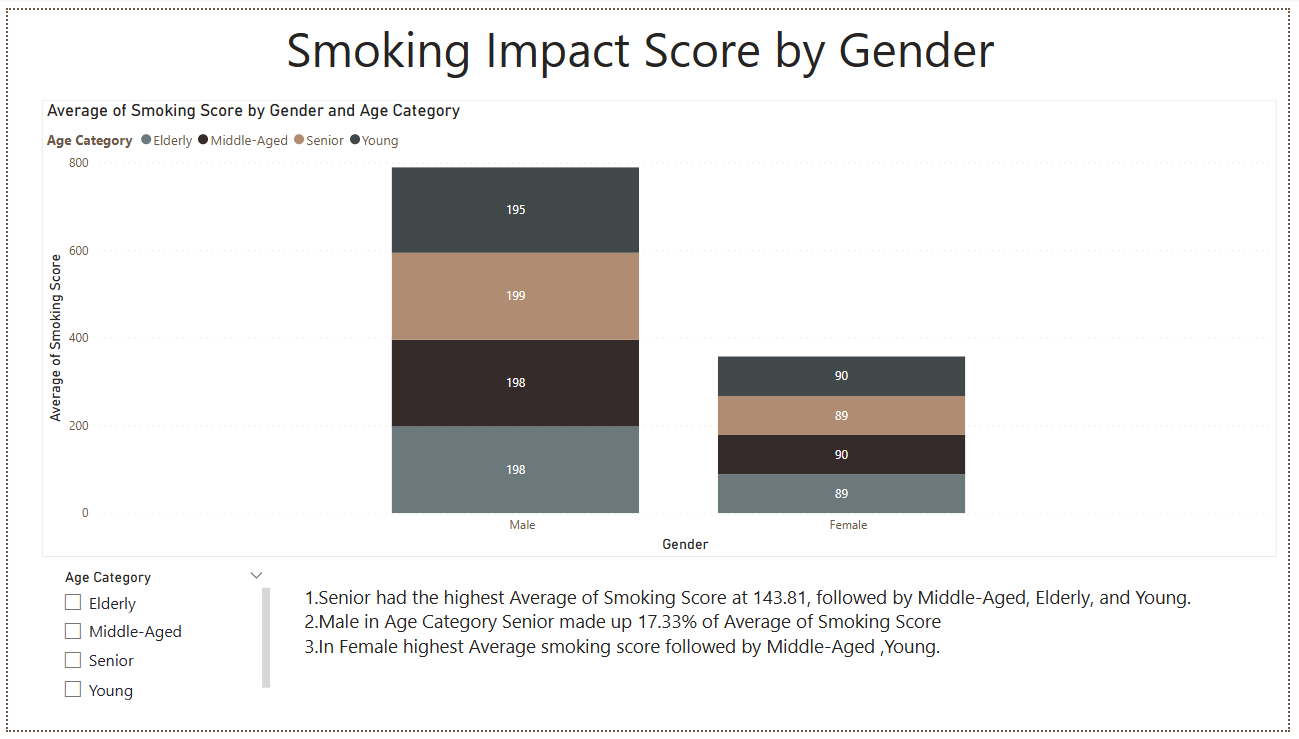
2.Elderly had the lowest count (1,547 cases), possibly due to lower survival rates or fewer diagnoses.



**4.Smoking Impact Score By Gender**

1.Seniors recorded the highest average smoking score (143.81), followed by Middle-Aged, Elderly, and Young

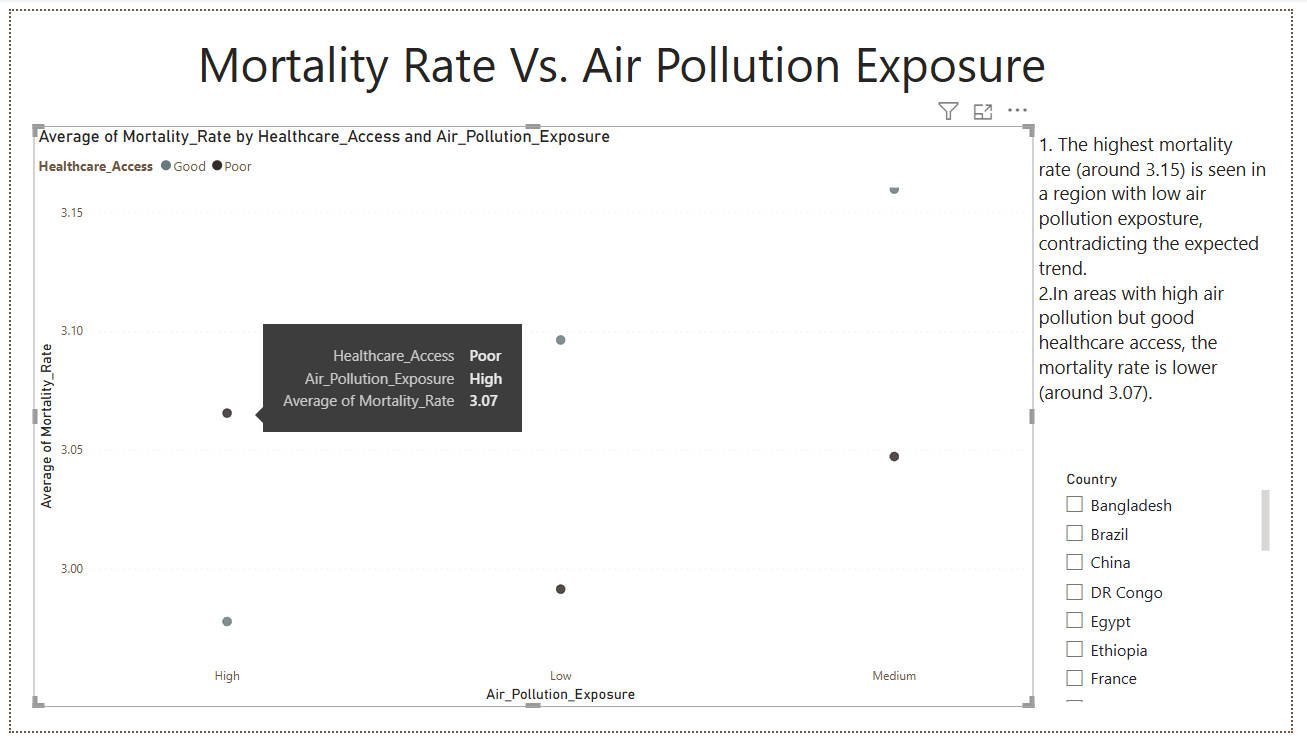
2. Among females, Middle-Aged had the highest smoking score, followed by Young, suggesting shifting smoking patterns among women.



**5. Mortality Rate vs. Air Pollution Exposure**

**Higher Air Pollution Doesn't Always Mean Higher Mortality**

* The **highest mortality rate** (around **3.15**) is seen in a region with **low air pollution exposure**, contradicting the expected trend.
* Countries with **poor healthcare access** show **higher mortality rates** across all air pollution levels.
* In areas with **high air pollution but good healthcare access**, the **mortality rate is lower** (around **3.00**).

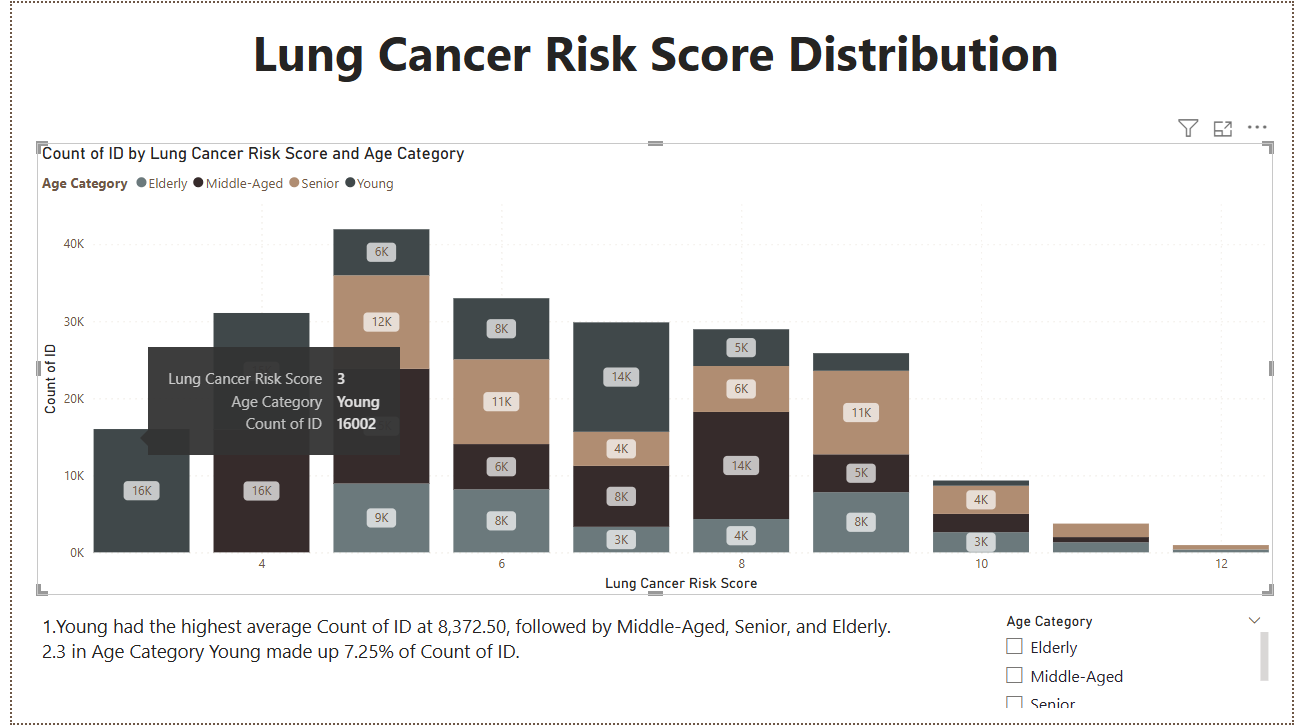


**6. Lung Cancer Risk Score Distribution**

**1.**The **Young** group has the **highest average Count of ID** at **8,372.50**.

2.**Young** individuals, even with a lower age-based score, could have higher risk scores if they smoke or have family histories of cancer.

3.**Middle-Aged, Senior, and Elderly** individuals start with a higher score due to their age and might have additional risk factors (smoking, pollution, family history) contributing to an even higher total score.

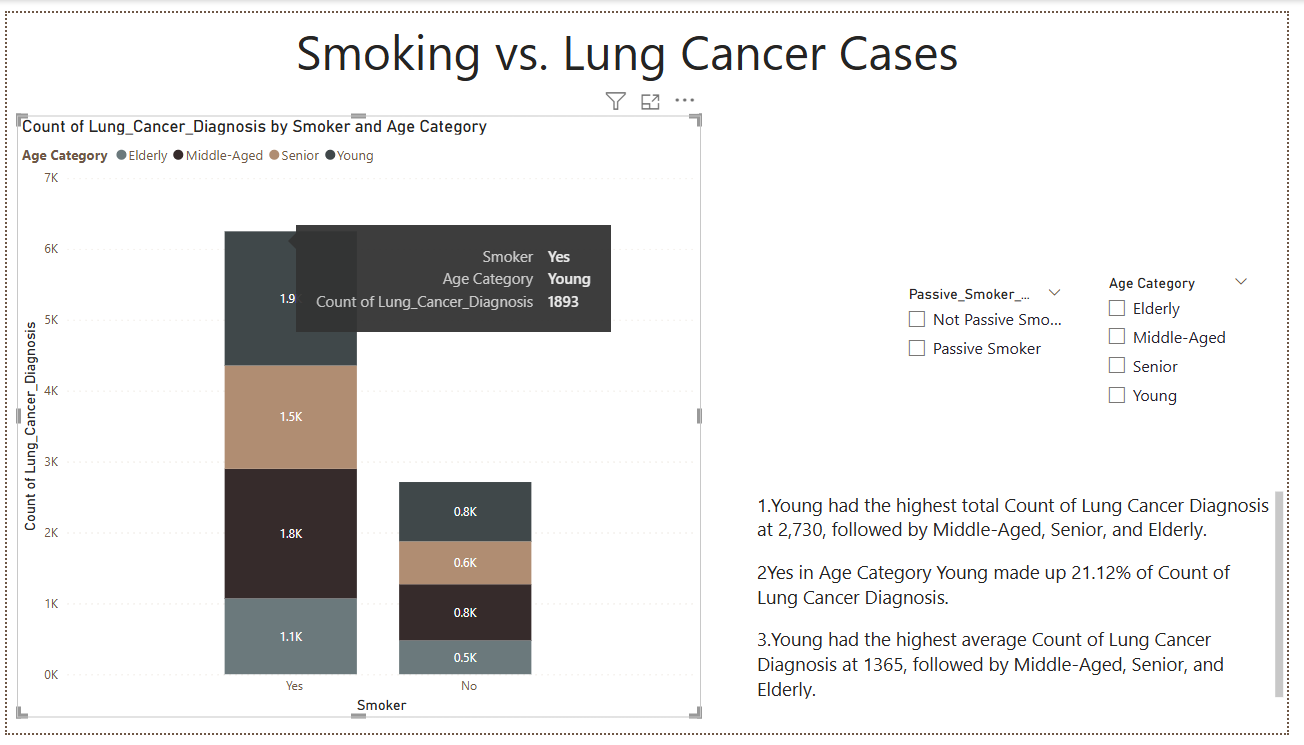


**7. Smoking vs. Lung Cancer Cases**

1.The **Young** group has an unexpectedly high count of lung cancer diagnoses among active smokers.

2.**Elderly** and **Middle-Aged** groups have the highest lung cancer diagnoses among passive smokers, indicating long-term exposure effects.

3.The **Young** category has the highest average count of lung cancer diagnoses (1,365), followed by older age groups.

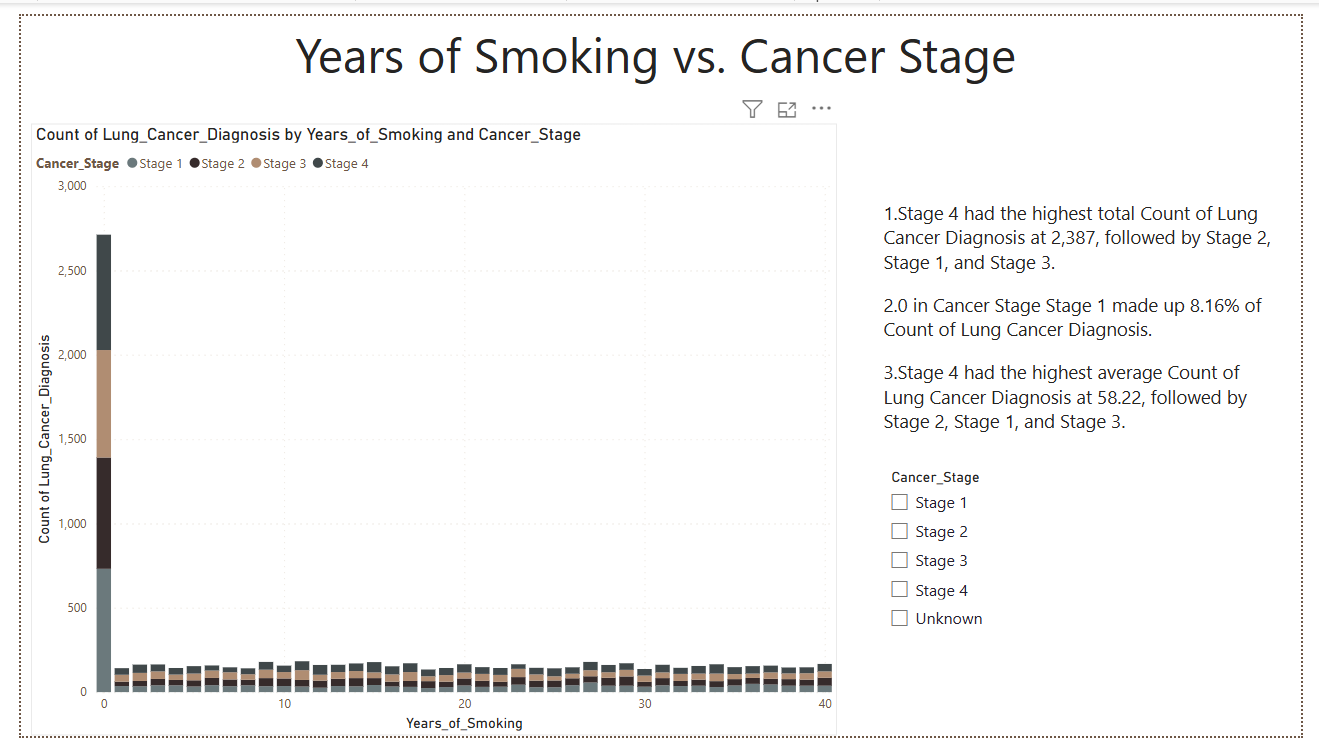


8. **Lung Cancer Diagnosis by cancer stages**:

1.In individuals with **0 years of smoking**, lung cancer is diagnosed at **all stages** (1-4), highlighting that smoking is not the sole cause of the disease. Other factors like **passive smoking**, **environmental carcinogens**, and **genetic predisposition** play a major role.

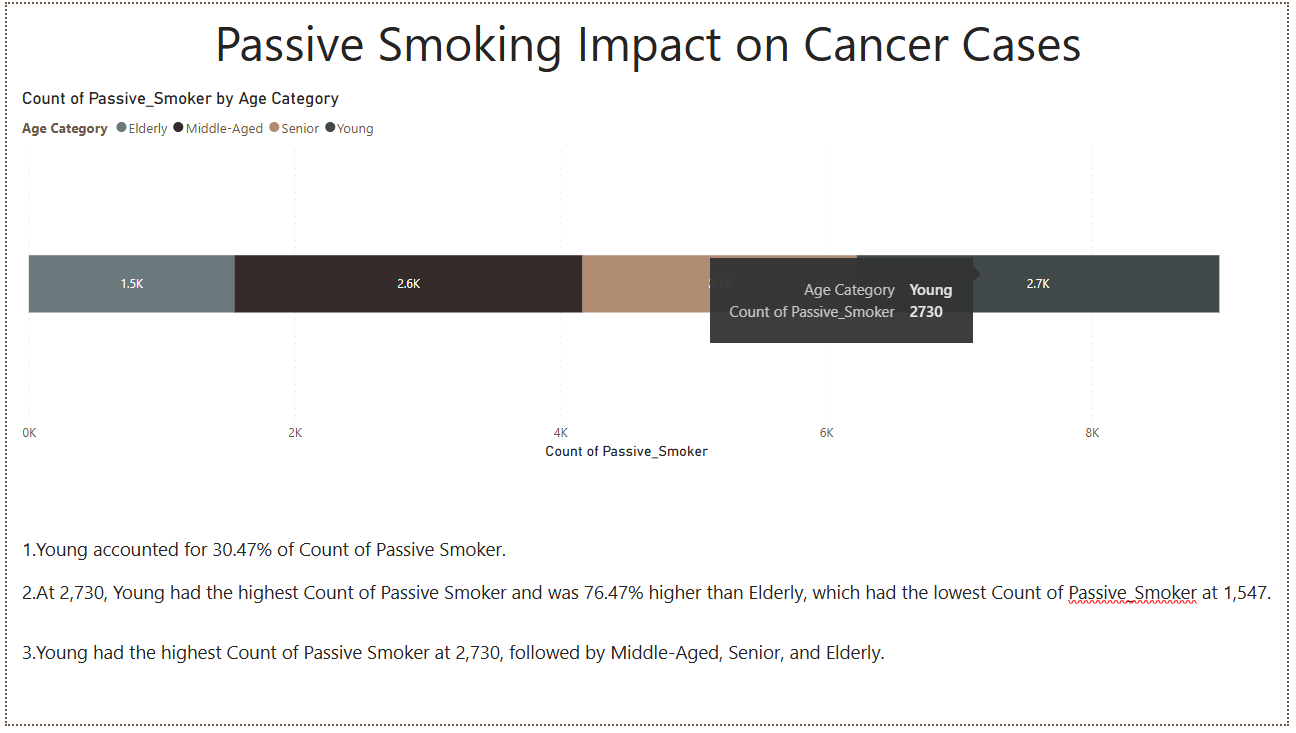
2.The highest percentage of diagnoses (33.36%) occur at **Stage 1**, indicating that many non-smokers are diagnosed with cancer at an early stage.

**3.Stage 2** (30.03%), **Stage 3** (29.15%), and **Stage 4** (28.66%) are also significant, showing that lung cancer in non-smokers progresses through various stages, with some reaching more advanced stages despite no smoking history.



**9.Passive Smoking Impact on Cancer Cases**

1. **Young** accounts for **30.47%** of passive smokers.
2. **Young** had the highest count of passive smokers (**2,730**), **76.47% higher** than the **Elderly**.
3. The order of passive smokers by count: **Young** > **Middle-Aged** > **Senior** > **Elderly**.



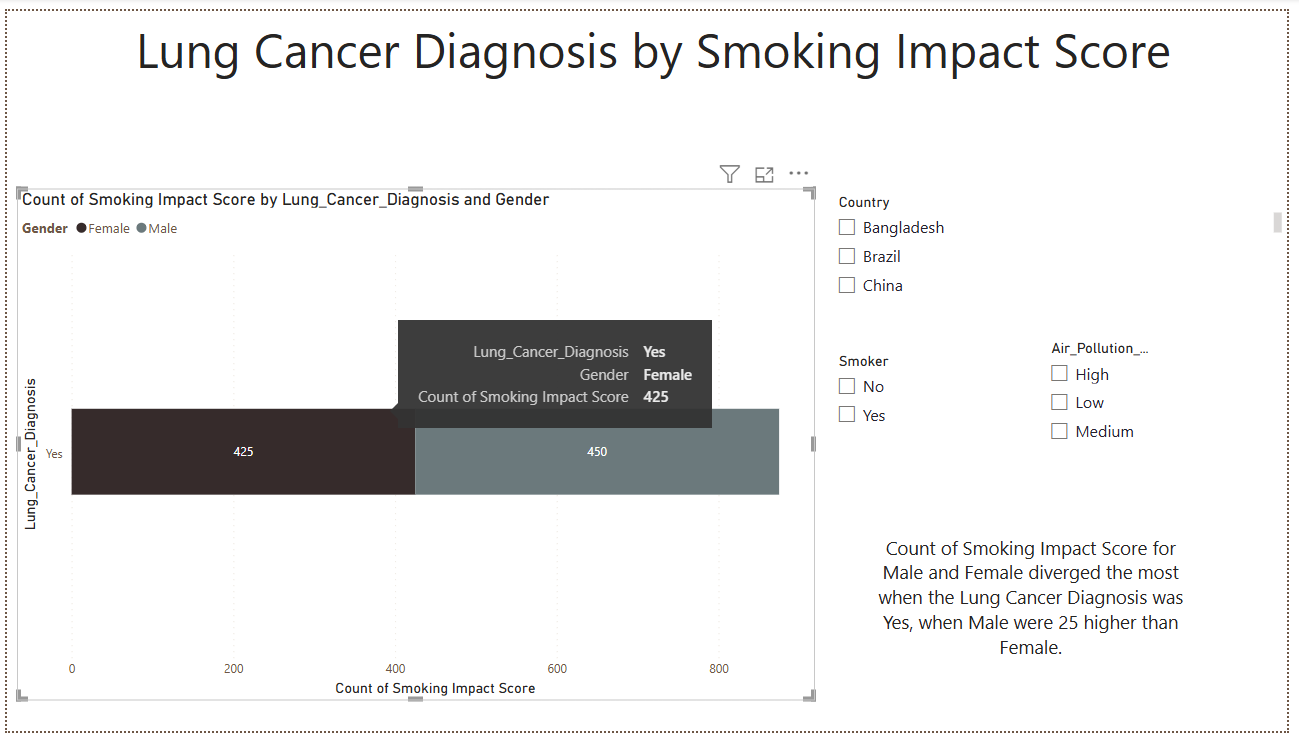
**10. Lung Cancer Diagnosis by Smoking Impact Score**

1.**Male Dominates Smoking Impact**:

* **Male** made up **51.43%** of the total count of smoking impact scores, indicating a higher proportion of male individuals affected by smoking in the data.

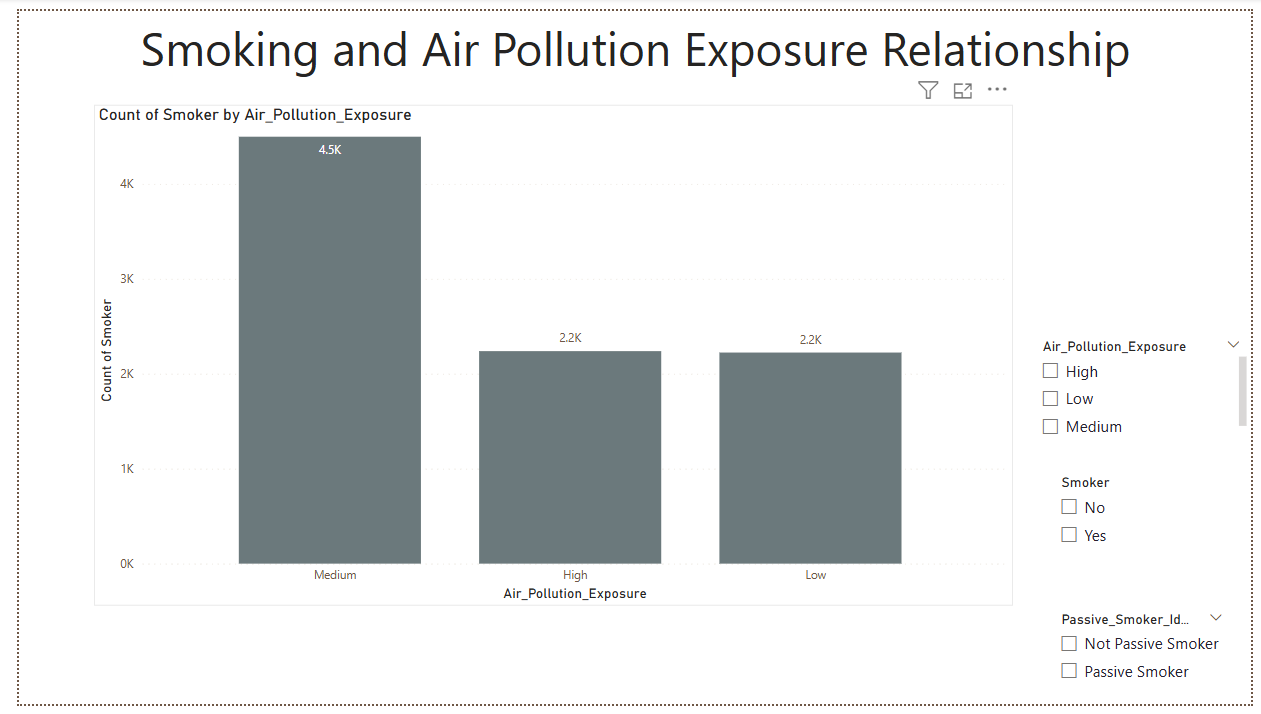
2.**Gender Divergence in Lung Cancer Diagnosis**:

* The **difference in smoking impact scores** between **Male** and **Female** was most pronounced when **Lung Cancer Diagnosis was Yes**. In this case, **Males** had **25** more diagnoses than **Females**, showing a higher smoking-related cancer impact among men.



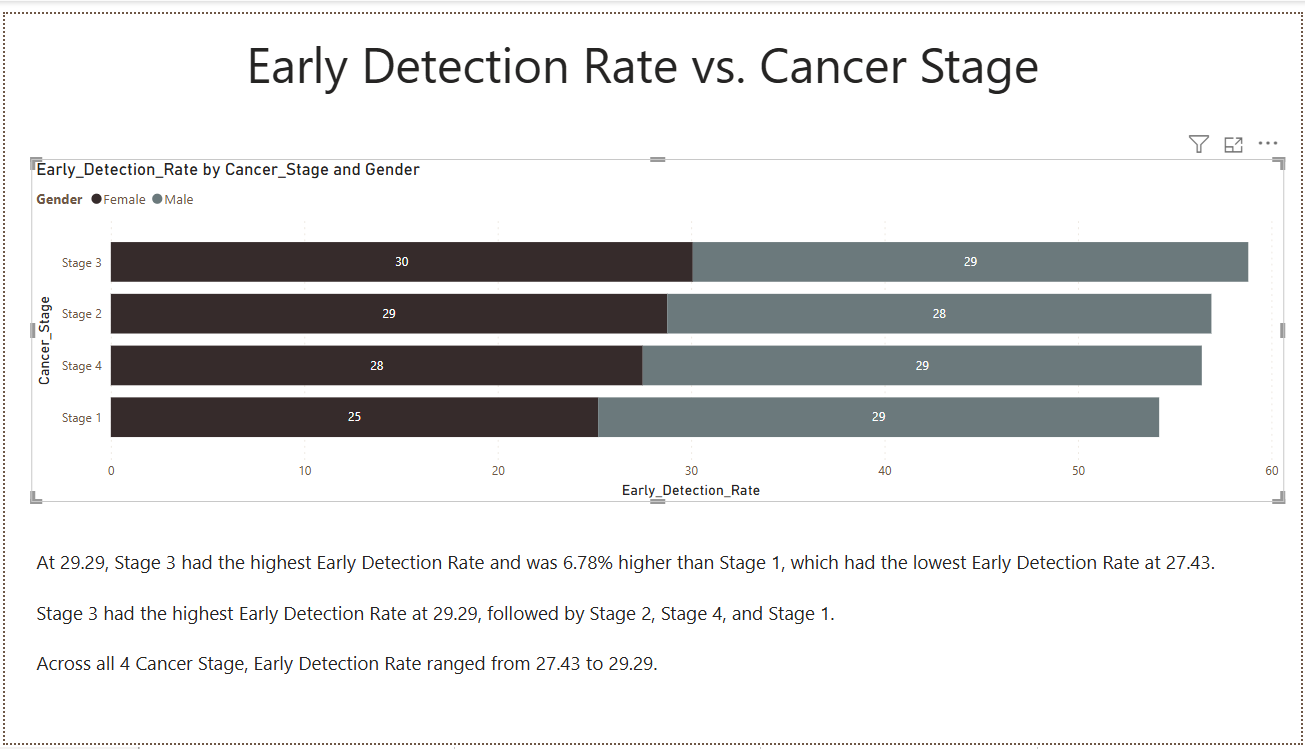
**11.Smoking and Air Pollution Exposure Relationship**

1️. **Medium Air Exposure** has the highest smokers (4,498, 50.20%).  
2️.**Low Air Exposure** has the least smokers (2,224, 102.25% lower than medium).  
3️.**High Air Exposure** (2,239 smokers) is slightly above low but far below medium.  
4️.**Smoking is most common in medium air exposure**, suggesting other factors may influence the trend.



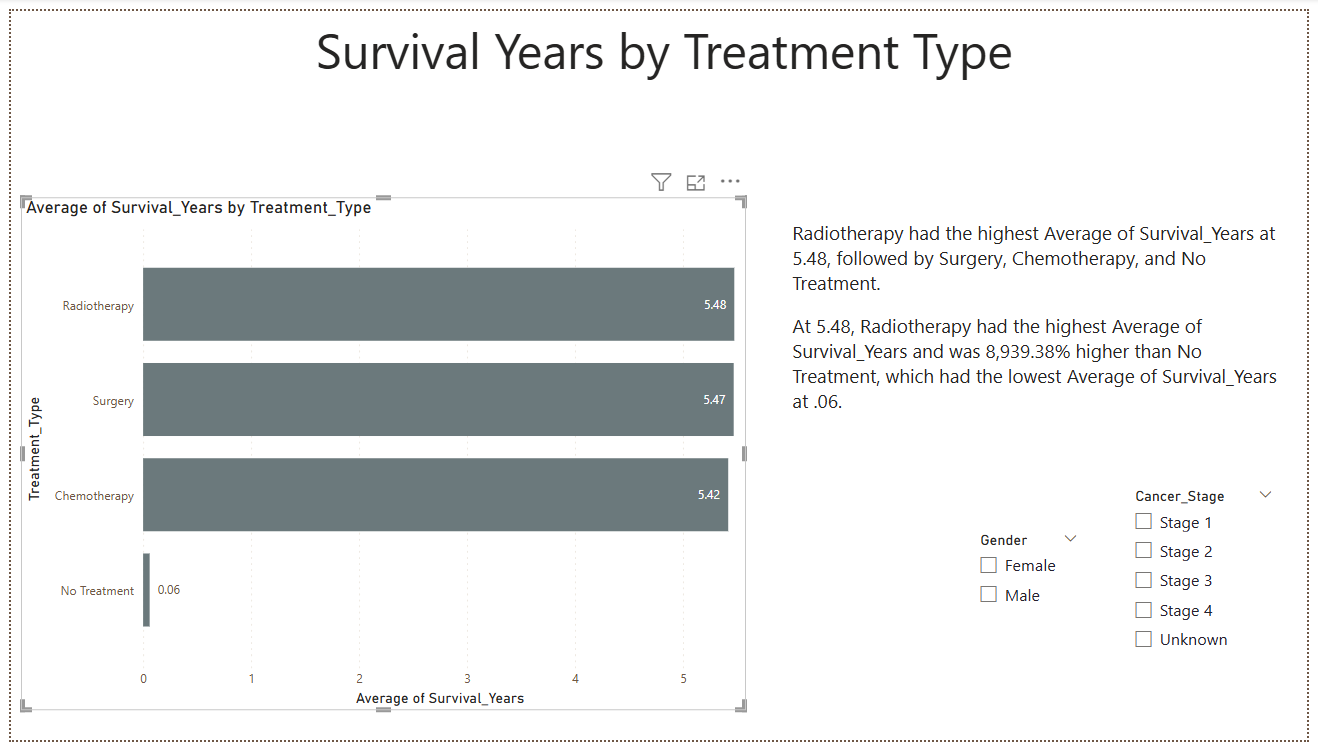
**12.Early Detection Rate Vs. Cancer stage**

1. **Stage 3** has the **highest** early detection rate at **29.29**, leading all stages.
2. **Stage 1** has the **lowest** early detection rate at **27.43**.
3. **Stage 3 is 6.78% higher than Stage 1 in early detection**.



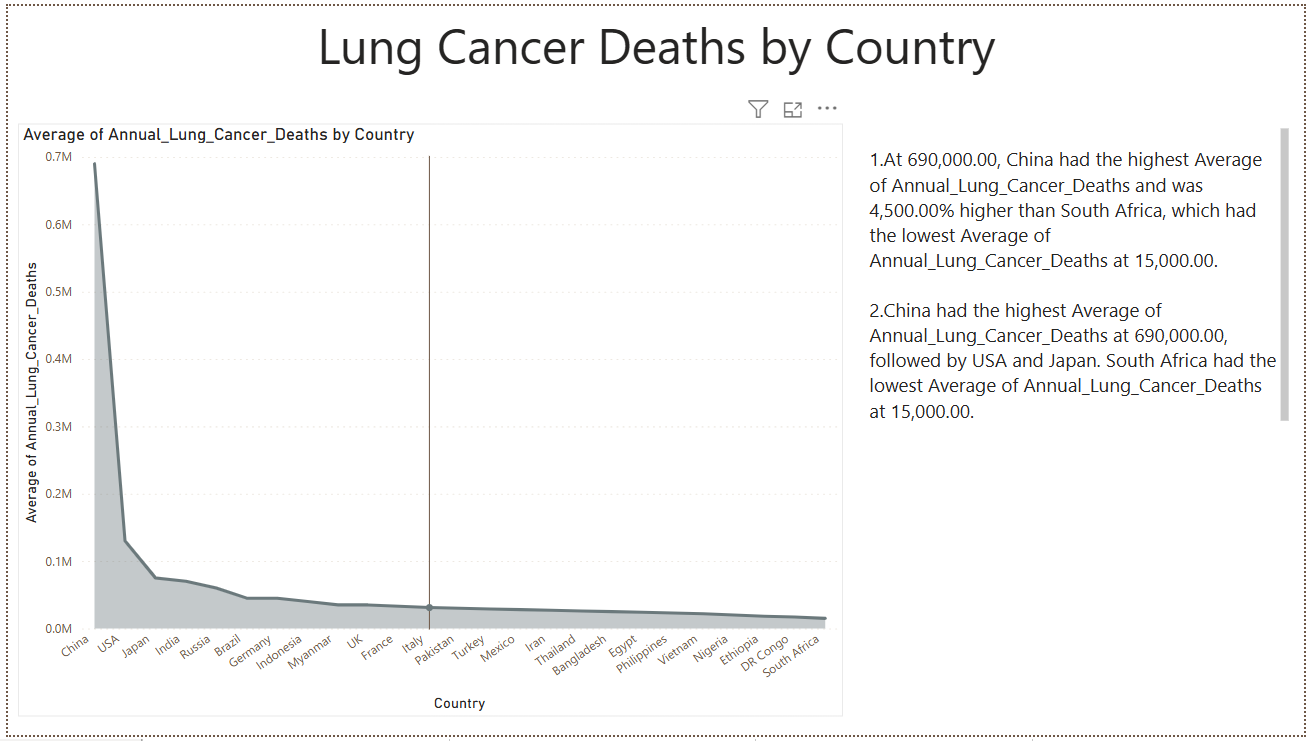
**13.Survival Years by Treatment Type**

1. **Radiotherapy has the highest** average survival years at 5.48, indicating it is the most effective treatment option.
2. Radiotherapy is significantly more effective than No Treatment, with an 8,939.38% higher average survival rate.
3. **No Treatment** has the lowest average survival years at 0.06, suggesting minimal benefit.
4. Surgery and Chemotherapy have lower survival rates than Radiotherapy, but they still offer better survival outcomes than No Treatment.



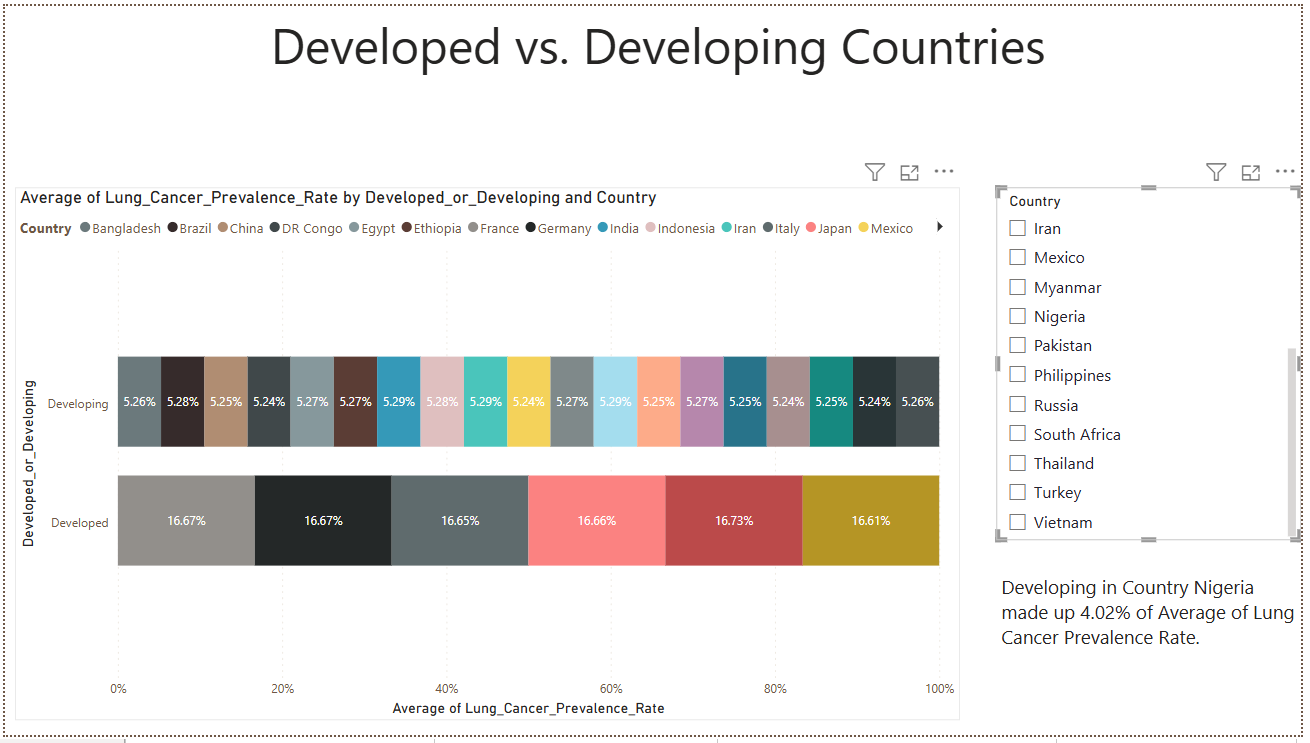
**14. Lung Cancer Deaths by Country**

1. **China has the highest average annual lung cancer deaths** at 690,000, which is 4,500% higher than **South Africa,** with the lowest at 15,000.
2. China leads in annual lung cancer deaths, followed by the **USA and Japan**. South Africa reports the lowest average annual lung cancer deaths.



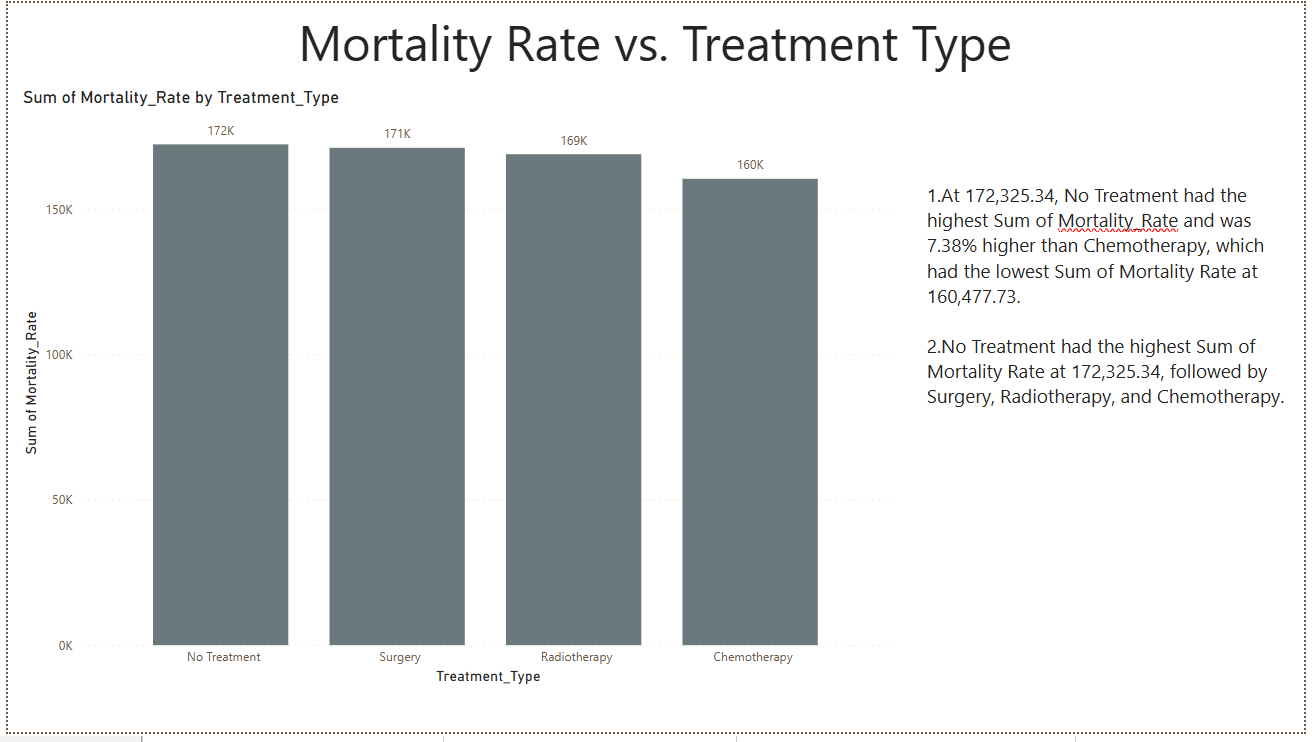
**15.Developed vs. Developing Countries**

**Nigeria** accounted for **4.02%** of the average lung cancer prevalence rate in developing countries, highlighting its contribution to the overall lung cancer burden within that category.



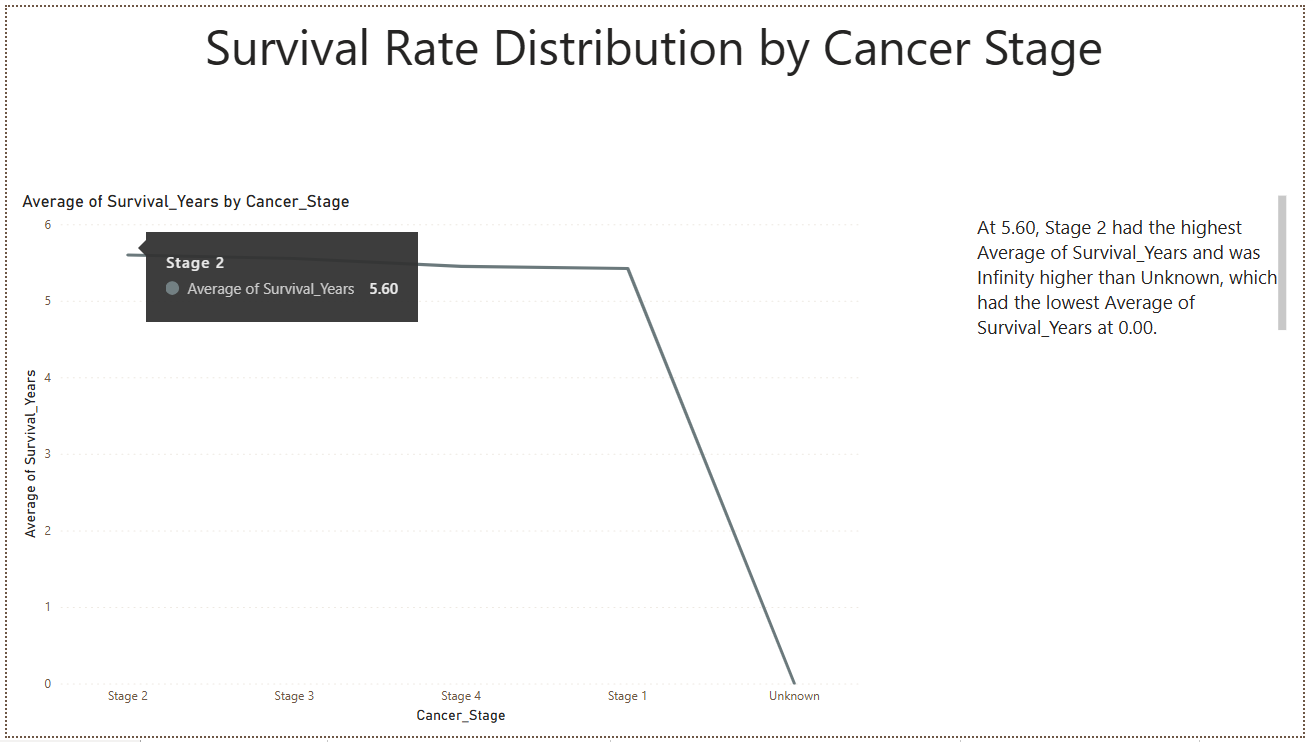
**16.Mortality rate Vs. Treatment Type**

1. **No Treatment** leads to the highest mortality rate, with a sum of **172,325.34**, surpassing other treatments by **7.38%**.
2. **Chemotherapy** shows the lowest mortality rate at **160,477.73**, indicating its effectiveness in reducing fatalities.
3. The treatment order, from highest to lowest mortality rate, is: **No Treatment**, **Surgery**, **Radiotherapy**, and **Chemotherapy**, emphasizing the crucial role of medical interventions in reducing cancer mortality.



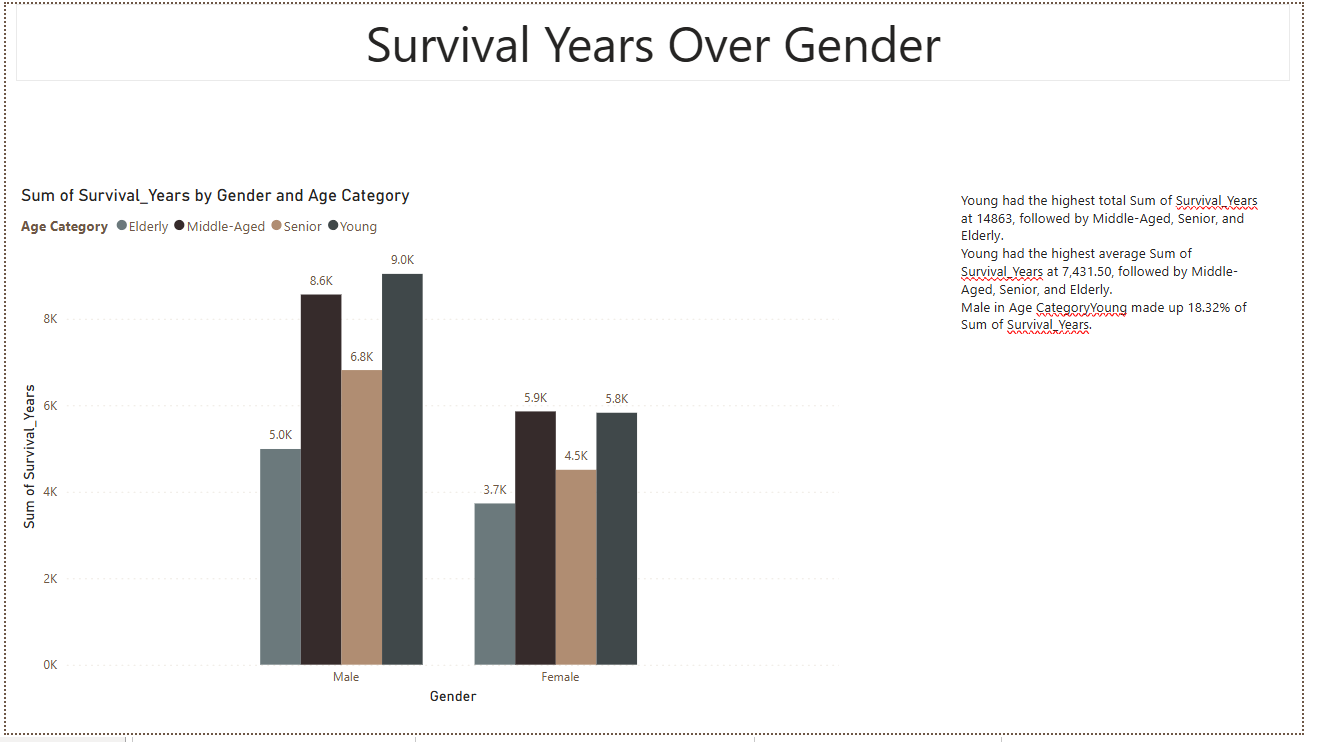
**17.Survival Rate Distribution by Cancer Stage**

1. **Stage 2** has the highest average survival years at **5.60**, significantly outperforming the **Unknown** category, which has an average of **0.00** years.
2. The contrast between **Stage 2** and **Unknown** underscores the importance of early diagnosis in improving survival outcomes.



**18.Survival Years Over Gender**

1. **Male Young** contributes 18.32% to the total **Survival Years** for males.
2. **Young** has the highest **Sum of Survival Years** at 7,431.50, followed by **Middle-Aged**, **Senior**, and **Elderly**.



**All Dashboards**

