Futurion
UPSKILLING
INDIA

# LUNG CANCER ANALYSIS

## Portfolio Project Challenge

### *Data Analysis Project*

By Sahil Gupta

Tools used SQL and Power BI

Futurion
UPSKILLING INDIA

# Instructions

**Problem Statement: Lung Cancer Analysis**

- Lung cancer remains one of the leading causes of cancer-related deaths worldwide, with late diagnosis significantly reducing survival rates. Key risk factors such as **smoking, passive smoking, air pollution, and genetic predisposition** contribute to its prevalence. Early identification of high-risk individuals and understanding survival patterns are essential for improving patient outcomes.

- This project leverages **SQL and Power BI** for data analysis and visualization, utilizing patient records to uncover insights into **risk factors, diagnosis trends, survival rates, and treatment effectiveness**. The goal is to identify key patterns, assess the impact of different factors on lung cancer progression, and provide data-driven insights to aid early detection and treatment planning.

Futurion
UPSKILLING INDIA

# Instructions

## PROJECT OBJECTIVE:

- **Patient Data Segmentation:** Retrieve and categorize records based on **lung cancer diagnosis, smoking status, age, gender, and geography**.
- **Risk Factor Analysis:** Assess the impact of **smoking, passive smoking, and air pollution** on lung cancer prevalence.
- **Cancer Progression Insights:** Identify **unique cancer stages** and analyze **survival years** based on disease progression.
- **Mortality Rate Evaluation:** Determine **death rates based on early detection and treatment effectiveness**.
- **Global Prevalence Ranking:** Identify **countries with the highest lung cancer rates and mortality statistics**.
- **Environmental & Occupational Risk Assessment:** Establish **correlations between air pollution, occupational exposure, and lung cancer risk**.
- **Treatment Effectiveness Analysis:** Assess the **impact of treatment types and early detection on survival rates**.
- **Gender-Based Analysis:** Compare **lung cancer prevalence across men and women in different regions.**

Futurion
UPSKILLING INDIA

# Instructions

## Dataset Overview

- **Demographic Information:** ID, Country, Population_Size, Age, Gender
- **Lifestyle & Environmental Factors:** Smoker, Years_of_Smoking, Cigarettes_per_Day, Passive_Smoker, Air_Pollution_Exposure, Occupational_Exposure, Indoor_Pollution
- **Medical History & Diagnosis:** Family_History, Lung_Cancer_Diagnosis, Cancer_Stage, Adenocarcinoma_Type
- **Healthcare & Treatment:** Healthcare_Access, Early_Detection, Treatment_Type
- **Patient Outcomes & Statistics:** Survival_Years, Developed_or_Developing, Annual_Lung_Cancer_Deaths, Lung_Cancer_Prevalence_Rate, Mortality_Rate

Nr of columns: 23
Nr of Records: 2206332

**Data Cleaning for Lung Cancer Analysis**

**1.Data Validation and Standardization** – Used SQL queries to identify and correct inconsistencies in data types, column formats, and categorical values to ensure uniformity.

**2.Handling Missing and Duplicate Values** – Checked for null entries and duplicate records, applying appropriate cleaning techniques like imputation or removal to maintain data integrity.

**3.Ensuring Data Consistency** – Standardized naming conventions, date formats, and categorical labels to create a structured and reliable dataset for analysis.

Futurion
UPSKILLING INDIA

# Instructions

**Business Problems solved**

1. Retrieve all records for individuals diagnosed with lung cancer.

```
SELECT * FROM Lung_Cancer_Data
WHERE Lung_Cancer_Diagnosis = 'Yes’
```

| ID | Country | Population_Size | Age | Gender | Smoker | Years_of_Smoking | Cigarettes_per_Day | Passive_Smoker | Family_History | Lung_Cancer_Diagnosis | Canc |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 26 | Pakistan | 225 | 40 | Female | Yes | 11 | 17 | No | No | Yes | Stag |
| 32 | Nigeria | 206 | 55 | Male | Yes | 9 | 8 | No | Yes | Yes | Stag |
| 33 | Turkey | 85 | 33 | Male | Yes | 4 | 12 | No | Yes | Yes | Stag |
| 93 | UK | 67 | 61 | Male | Yes | 14 | 28 | No | No | Yes | Stag |
| 106 | Ethiopia | 120 | 70 | Male | Yes | 7 | 21 | No | No | Yes | Stag |
| 157 | Germany | 83 | 72 | Male | Yes | 25 | 26 | No | No | Yes | Stag |
| 168 | Indonesia | 273 | 47 | Female | No | 0 | 0 | Yes | No | Yes | Stag |
| 188 | Egypt | 102 | 71 | Male | Yes | 36 | 7 | Yes | No | Yes | Stag |
| 207 | Iran | 84 | 51 | Male | Yes | 28 | 26 | No | No | Yes | Stag |
| 227 | Russia | 145 | 85 | Male | Yes | 40 | 26 | No | No | Yes | Stag |
| 229 | Turkey | 85 | 66 | Male | Yes | 25 | 6 | No | No | Yes | Stag |
| 289 | DR Con... | 95 | 32 | Female | Yes | 15 | 8 | No | No | Yes | Stag |
| 298 | Nigeria | 206 | 60 | Male | Yes | 19 | 21 | No | No | Yes | Stag |
| 300 | Mexico | 128 | 28 | Female | No | 0 | 0 | No | Yes | Yes | Stag |
| 335 | Indonesia | 273 | 73 | Male | Yes | 27 | 24 | No | No | Yes | Stag |
| 346 | Indonesia | 273 | 20 | Male | Yes | 3 | 6 | No | No | Yes | Stag |
| 438 | DR Con... | 95 | 51 | Male | Yes | 26 | 26 | No | Yes | Yes | Stag |
| 454 | Egypt | 102 | 48 | Male | Yes | 34 | 18 | No | No | Yes | Stag |
| 482 | Turkey | 85 | 32 | Male | Yes | 14 | 11 | Yes | No | Yes | Stag |
| 500 | UK | 67 | 28 | Female | No | 0 | 0 | No | No | Yes | Stag |
| 570 | Philippi... | 113 | 21 | Female | No | 0 | 0 | No | Yes | Yes | Stag |
| 618 | China | 1400 | 48 | Male | Yes | 10 | 8 | Yes | No | Yes | Stag |

y executed successfully. | LAPTOP-MQQSV87C\SQLEXPRESS ... | LAPTOP-MQQSV87C\sahil ... | Lung_Caner_Dataset | 00:00:00 | 8,961 rows

Futurion
UPSKILLING INDIA

# Instructions

**2. Count the number of smokers and non-smokers.**

```sql
SELECT
    CASE
        WHEN Smoker = 'Yes' THEN 'Smoker'
        WHEN Smoker = 'No' THEN 'Non-Smoker'
        ELSE 'Unknown'
    END AS Smoking_Status,
    FORMAT(CAST(COUNT(ID) AS BIGINT), 'N0') AS Total_Count
FROM Lung_Cancer_Data
GROUP BY
     Smoker;
```

| Smoking_Status | Total_Count |
|---|---|
| Smoker | 88,341 |
| Non-Smoker | 132,291 |

**3. List all unique cancer stages present in the dataset.**

```sql
SELECT
   DISTINCT Cancer_Stage
FROM Lung_Cancer_Data
WHERE
     Cancer_Stage <> 'None'
ORDER BY
     Cancer_Stage
```

| Cancer_Stage |
|---|
| Stage 1 |
| Stage 2 |
| Stage 3 |
| Stage 4 |

Futurion
UPSKILLING INDIA

# Instructions

4. Retrieve the average number of cigarettes smoked per day by smokers.

```
SELECT
      AVG(Cigarettes_per_Day* 1.00) Avg_Nr_of_Cigarret_Smoked_by_Smokers
FROM Lung_Cancer_Data
WHERE
      Smoker = 'Yes’
```

| Avg_Nr_of_Cigarret_Smoked_by_Smokers |
| --- |
| 17.501296 |

5. Count the number of people exposed to high air pollution.

```
SELECT
      FORMAT(COUNT(ID), 'N0') Nr_of_People
FROM Lung_Cancer_Data
WHERE
      Air_Pollution_Exposure = 'High'
```

| Nr_of_People |
| --- |
| 55,108 |

Futurion
UPSKILLING INDIA

# Instructions

**6. Find the top 5 countries with the highest lung cancer deaths.**

```
SELECT * FROM
(SELECT
   DISTINCT Country,
   Annual_Lung_Cancer_Deaths,
   DENSE_RANK()OVER(ORDER BY Annual_Lung_Cancer_Deaths DESC) Rank
FROM Lung_Cancer_Data
)t
WHERE
      Rank<=5
ORDER BY
       Rank
```

| Country | Annual_Lung_Cancer_Deaths | Rank |
|---|---|---|
| China | 690000 | 1 |
| USA | 130000 | 2 |
| Japan | 75000 | 3 |
| India | 70000 | 4 |
| Russia | 60000 | 5 |

**7. Count the number of people diagnosed with lung cancer by gender.**

```
SELECT
   Gender,
   COUNT(Lung_Cancer_Diagnosis) Nr_of_People_Diagnosed_by_Cancer
FROM Lung_Cancer_Data
WHERE
      Lung_Cancer_Diagnosis = 'Yes'
GROUP BY
      Gender
```

Futurion
UPSKILLING INDIA

# Instructions

8. Retrieve records of individuals older than 60 who are diagnosed with lung cancer.

```
SELECT * FROM Lung_Cancer_Data
WHERE Age > 60 AND Lung_Cancer_Diagnosis = 'Yes'
```

| ID | Country | Population_Size | Age | Gender | Smoker | Years_of_Smoking | Cigarettes_per_Day | Passive_Smoker | Family_History | Lung_Canc... | Cancer_Stag |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 93 | UK | 67 | 61 | Male | Yes | 14 | 28 | No | No | Yes | Stage 1 |
| 106 | Ethiopia | 120 | 70 | Male | Yes | 7 | 21 | No | No | Yes | Stage 2 |
| 157 | Germany | 83 | 72 | Male | Yes | 25 | 26 | No | No | Yes | Stage 3 |
| 188 | Egypt | 102 | 71 | Male | Yes | 36 | 7 | Yes | No | Yes | Stage 1 |
| 227 | Russia | 145 | 85 | Male | Yes | 40 | 26 | No | No | Yes | Stage 2 |
| 229 | Turkey | 85 | 66 | Male | Yes | 25 | 6 | No | No | Yes | Stage 1 |
| 335 | Indonesia | 273 | 73 | Male | Yes | 27 | 24 | No | No | Yes | Stage 2 |
| 848 | Nigeria | 206 | 72 | Male | Yes | 21 | 27 | No | No | Yes | Stage 1 |
| 879 | Thailand | 70 | 81 | Male | Yes | 31 | 17 | No | Yes | Yes | Stage 4 |
| 907 | USA | 331 | 64 | Male | Yes | 1 | 23 | No | No | Yes | Stage 4 |
| 973 | Japan | 125 | 73 | Male | Yes | 25 | 26 | Yes | No | Yes | Stage 4 |
| 1044 | South Africa | 59 | 64 | Male | Yes | 13 | 14 | No | No | Yes | Stage 2 |
| 1116 | Germany | 83 | 65 | Male | No | 0 | 0 | No | No | Yes | Stage 4 |
| 1194 | UK | 67 | 63 | Male | No | 0 | 0 | No | No | Yes | Stage 4 |
| 1230 | Thailand | 70 | 66 | Male | Yes | 8 | 10 | No | No | Yes | Stage 1 |
| 1291 | UK | 67 | 68 | Male | Yes | 18 | 21 | No | No | Yes | Stage 2 |
| 1320 | DR Congo | 95 | 73 | Female | No | 0 | 0 | Yes | No | Yes | Stage 2 |
| 1397 | Ethiopia | 120 | 67 | Female | Yes | 7 | 24 | No | No | Yes | Stage 2 |
| 1406 | Philippines | 113 | 62 | Male | Yes | 17 | 26 | Yes | No | Yes | Stage 2 |
| 1458 | UK | 67 | 63 | Female | Yes | 31 | 13 | No | Yes | Yes | Stage 4 |

ry executed successfully. | LAPTOP-MQQSV87C\SQLEXPRESS ... | LAPTOP-MQQSV87C\sahil ... | Lung_Caner_Dataset | 00:00:00 | 3,476 rows

Futurion
UPSKILLING INDIA

# Instructions

**9. Find the percentage of smokers who developed lung cancer.**

```
SELECT
ROUND(
CAST(
   SUM(
   CASE
      WHEN Smoker = 'Yes' AND Lung_Cancer_Diagnosis = 'Yes' THEN 1
   END ) AS float) /
   SUM(
   CASE
      WHEN Smoker = 'Yes' THEN 1
   END ) * 100,
2) Percentage_of_Smokers_with_developed_Lung_Cancer
FROM Lung_Cancer_Data
```

| Percentage_of_Smokers_with_developed_Lung_Cancer |
|---|
| 7.07 |

Futurion
UPSKILLING INDIA

# Instructions

**10. Calculate the average survival years based on cancer stages.**

```
SELECT
    Cancer_Stage,
    AVG(Survival_Years * 1.0) AS Avg_Survival_Years
FROM Lung_Cancer_Data
WHERE
      Cancer_Stage <> 'None'
GROUP BY
      Cancer_Stage
ORDER BY
      AVG(Survival_Years * 1.0) DESC;
```

| Cancer_Stage | Avg_Survival_Years |
|---|---|
| Stage 2 | 5.596906 |
| Stage 3 | 5.551487 |
| Stage 4 | 5.448680 |
| Stage 1 | 5.421725 |

**11. Count the number of lung cancer patients based on passive smoking.**

```
SELECT
   Passive_Smoker,
   COUNT(ID) Nr_of_Lung_Cancer_Patient
FROM Lung_Cancer_Data
WHERE
      Lung_Cancer_Diagnosis = 'Yes'
GROUP BY
      Passive_Smoker
```

| Passive_Smoker | Nr_of_Lung_Cancer_Patient |
|---|---|
| Yes | 2735 |
| No | 6226 |

Futurion
UPSKILLING INDIA

# Instructions

**12. Find the country with the highest lung cancer prevalence rate.**

```
SELECT * FROM
(
SELECT
   DISTINCT Country,
   Lung_Cancer_Prevalence_Rate,
   DENSE_RANK() OVER(ORDER BY Lung_Cancer_Prevalence_Rate DESC)
    Rank
FROM Lung_Cancer_Data
)t
WHERE RANK = 1
```

| Country | Lung_Cancer_Prevalence_Rate | Rank |
|---|---|---|
| Philippines | 2.5 | 1 |
| Germany | 2.5 | 1 |
| India | 2.5 | 1 |
| France | 2.5 | 1 |
| UK | 2.5 | 1 |
| Vietnam | 2.5 | 1 |
| South Africa | 2.5 | 1 |
| Pakistan | 2.5 | 1 |
| Thailand | 2.5 | 1 |
| DR Congo | 2.5 | 1 |
| Ethiopia | 2.5 | 1 |
| Brazil | 2.5 | 1 |
| Iran | 2.5 | 1 |
| Russia | 2.5 | 1 |
| USA | 2.5 | 1 |
| Egypt | 2.5 | 1 |
| Indonesia | 2.5 | 1 |
| Italy | 2.5 | 1 |
| China | 2.5 | 1 |
| Nigeria | 2.5 | 1 |
| Japan | 2.5 | 1 |
| Mexico | 2.5 | 1 |
| Bangladesh | 2.5 | 1 |
| Myanmar | 2.5 | 1 |
| Turkey | 2.5 | 1 |

Futurion
UPSKILLING INDIA

# Instructions

**13.(A) Identify the smoking years' impact on lung cancer**

**Impact of Smoking Duration on Lung Cancer Stages:**
**A Case Count Analysis**

```
SELECT
   Years_of_Smoking,
   Cancer_Stage,
   COUNT(*) AS Cases
FROM Lung_Cancer_Data
WHERE
      Lung_Cancer_Diagnosis = 'Yes’
      AND
      Smoker = 'Yes'
GROUP BY
   Years_of_Smoking,
   Cancer_Stage
ORDER BY
    Cancer_Stage,
    COUNT(*) DESC
```

| Years_of_Smoking | Cancer_Stage | Cases |
|---|---|---|
| 27 | Stage 1 | 57 |
| 36 | Stage 1 | 49 |
| 37 | Stage 1 | 45 |
| 23 | Stage 1 | 44 |
| 15 | Stage 1 | 42 |
| 6 | Stage 1 | 41 |
| 26 | Stage 1 | 41 |
| 4 | Stage 1 | 40 |
| 31 | Stage 1 | 40 |
| 3 | Stage 1 | 40 |
| 8 | Stage 1 | 40 |
| 20 | Stage 1 | 39 |
| 35 | Stage 1 | 39 |
| 40 | Stage 1 | 38 |
| 29 | Stage 1 | 38 |
| 38 | Stage 1 | 38 |
| 33 | Stage 1 | 38 |
| 39 | Stage 1 | 38 |
| 28 | Stage 1 | 38 |
| 1 | Stage 1 | 36 |
| 32 | Stage 1 | 36 |
| 13 | Stage 1 | 36 |
| 10 | Stage 1 | 36 |

y executed successfully.

Futurion
UPSKILLING INDIA

# Instructions

```sql
13.(B)Average Years of Smoking Across Lung Cancer Stages
SELECT
   Cancer_Stage,
   AVG(Years_of_Smoking * 1.00) AS Avg_Smoking_Years
FROM Lung_Cancer_Data
WHERE
   Lung_Cancer_Diagnosis = 'Yes'
   AND
   Smoker = 'Yes'
GROUP BY
      Cancer_Stage
ORDER BY
      Cancer_Stage;
```

| Cancer_Stage | Avg_Smoking_Years |
|---|---|
| Stage 1 | 20.921917 |
| Stage 2 | 20.633940 |
| Stage 3 | 19.722868 |
| Stage 4 | 20.418085 |

Futurion
UPSKILLING INDIA

**14. Determine the mortality rate for patients with and without early detection.**

```
SELECT
   Early_Detection,
   COUNT(*) AS Total_Patients,
   ROUND(AVG(Mortality_Rate), 2) AS Avg_Mortality_Rate,
   ROUND(MAX(Mortality_Rate),2) AS Max_Mortality_Rate,
   MIN(Mortality_Rate) AS Min_Mortality_Rate
FROM Lung_Cancer_Data
GROUP BY
      Early_Detection;
```

| Early_Detection | Total_Patients | Avg_Mortality_Rate | Max_Mortality_Rate | Min_Mortality_Rate |
|---|---|---|---|---|
| Yes | 61719 | 3.08 | 89.97 | 0 |
| No | 158913 | 3.04 | 90 | 0 |

Futurion
UPSKILLING INDIA

# Instructions

15. Group the lung cancer prevalence rate by developed vs. developing countries.

```
SELECT
   Developed_or_Developing as Country_Status,
   COUNT(ID) Nr_of_Patient,
   ROUND(AVG(Lung_Cancer_Prevalence_Rate),4) Avg_LCPR,
   MAX(Lung_Cancer_Prevalence_Rate) Max_LCPR,
   MIN(Lung_Cancer_Prevalence_Rate) Min_LCPR
FROM Lung_Cancer_Data
GROUP BY
      Developed_or_Developing
```

| Country_Status | Nr_of_Patient | Avg_LCPR | Max_LCPR | Min_LCPR |
|---|---|---|---|---|
| Developing | 167741 | 1.5022 | 2.5 | 0.5 |
| Developed | 52891 | 1.5018 | 2.5 | 0.5 |

Futurion
UPSKILLING INDIA

# Instructions

**16. Identify the correlation between lung cancer prevalence and air pollution levels.**

```
SELECT
   Air_Pollution_Exposure,
   SUM(
   CASE
   WHEN Lung_Cancer_Diagnosis ='Yes' THEN 1
   ELSE 0
   END) Nr_of_Lung_Cancer_Patient,
   ROUND(AVG(Lung_Cancer_Prevalence_Rate)* 1.00,3) Avg_LCP_rate,
   ROUND(MAX(Lung_Cancer_Prevalence_Rate)* 1.00,3) Max_LCP_Rate,
   ROUND(MIN(Lung_Cancer_Prevalence_Rate)* 1.00,3) Max_LCP_Rate
FROM Lung_Cancer_Data
GROUP BY
      Air_Pollution_Exposure
```

| Air_Pollution_Exposure | Nr_of_Lung_Cancer_Patient | Avg_LCP_rate | Max_LCP_Rate | Max_LCP_Rate |
|---|---|---|---|---|
| High | 2239 | 1.503 | 2.5 | 0.5 |
| Low | 2224 | 1.503 | 2.5 | 0.5 |
| Medium | 4498 | 1.501 | 2.5 | 0.5 |

Futurion
UPSKILLING INDIA

**17. Find the average age of lung cancer patients for each country.**

```
SELECT
   Country,
   AVG(Age*1.00) Avg_Age
FROM Lung_Cancer_Data
WHERE
      Lung_Cancer_Diagnosis = 'Yes'
GROUP BY
      Country
ORDER BY
      AVG(Age*1.00) DESC
```

| Country | Avg_Age |
|---|---|
| Germany | 54.177710 |
| Egypt | 54.040431 |
| South Africa | 53.731092 |
| Russia | 53.420588 |
| Brazil | 53.376770 |
| Italy | 53.231182 |
| Philippines | 53.011396 |
| Mexico | 52.913690 |
| China | 52.884615 |
| Ethiopia | 52.772616 |
| Thailand | 52.705014 |
| UK | 52.675213 |
| Vietnam | 52.657738 |
| Indonesia | 52.631147 |
| France | 52.591780 |
| Bangladesh | 52.432132 |
| DR Congo | 52.351648 |
| Pakistan | 52.320809 |
| Japan | 52.183417 |
| Nigeria | 52.148541 |
| Turkey | 52.128865 |
| Myanmar | 51.852546 |
| USA | 51.766578 |
| Iran | 51.668711 |
| India | 51.065671 |

Futurion
UPSKILLING INDIA

# Instructions

**18. Calculate the risk factor of lung cancer by smoker status, passive smoking, and family history.**

```
SELECT
CASE
WHEN Lung_Cancer_Diagnosis = 'Yes' THEN 'Diagnosed'
WHEN Lung_Cancer_Diagnosis = 'No' THEN 'Not Diagnosed'
END AS Diagnosis_Status,
(SUM(CASE
        WHEN Smoker = 'Yes' THEN 1
        ELSE 0
        END) *100.00)/ COUNT(*) Smoker_Risk_Percent,
(SUM(CASE
        WHEN Passive_Smoker = 'Yes' THEN 1
        ELSE 0
        END )*100.00)/COUNT(*) Passive_Smoker_Risk_Percent,
(SUM(CASE
        WHEN Family_History = 'Yes' THEN 1
        ELSE 0
        END )*100.00)/COUNT(*) Family_History_Risk_Percent
FROM
Lung_Cancer_Data
GROUP BY
    CASE
    WHEN Lung_Cancer_Diagnosis = 'Yes' THEN 'Diagnosed'
    WHEN Lung_Cancer_Diagnosis = 'No' THEN 'Not Diagnosed'
    END
```

| Diagnosis_Status | Smoker_Risk_Percent | Passive_Smoker_Risk_Percent | Family_History_Risk_Percent |
|---|---|---|---|
| Diagnosed | 69.7355205892199 | 30.5211471933935 | 14.5296283896886 |
| Not Diagnosed | 38.7828280680867 | 29.8704121017994 | 14.8924510206877 |

Futurion
UPSKILLING INDIA

# Instructions

**19. Rank countries based on their mortality rate.**

```
SELECT
       Country,
       ROUND(AVG(Mortality_Rate* 1.00),2) Avg_Mortality_Rate,
       DENSE_RANK() OVER(ORDER BY ROUND(AVG(Mortality_Rate* 1.00),2) DESC)
Rank_by_Mortality
FROM Lung_Cancer_Data
GROUP BY
        Country
```

| Country | Avg_Mortality_Rate | Rank_by_Mortality |
|---|---|---|
| Ethiopia | 3.43 | 1 |
| Japan | 3.32 | 2 |
| Turkey | 3.26 | 3 |
| USA | 3.23 | 4 |
| Myanmar | 3.21 | 5 |
| Nigeria | 3.21 | 5 |
| Egypt | 3.16 | 6 |
| Italy | 3.15 | 7 |
| Indonesia | 3.13 | 8 |
| Banglad... | 3.08 | 9 |
| DR Con... | 3.08 | 9 |
| Brazil | 3.06 | 10 |
| UK | 3.06 | 10 |
| France | 3.03 | 11 |
| South A... | 3.02 | 12 |
| Pakistan | 2.97 | 13 |
| Philippi... | 2.95 | 14 |
| Thailand | 2.93 | 15 |
| Russia | 2.92 | 16 |
| Vietnam | 2.87 | 17 |
| India | 2.87 | 17 |
| Germany | 2.84 | 18 |
| China | 2.84 | 18 |

**20. Determine if treatment type has a significant impact on survival years.**

```
SELECT
    Treatment_Type,
    AVG(Survival_Years* 1.00) Avg_Survival_Years
FROM Lung_Cancer_Data
GROUP BY
       Treatment_Type
ORDER BY
       AVG(Survival_Years* 1.00) DESC
```

| Treatment_Type | Avg_Survival_Years |
|---|---|
| Radiotherapy | 5.475555 |
| Surgery | 5.470070 |
| Chemotherapy | 5.419234 |
| None | 0.060574 |

Futurion
UPSKILLING INDIA

```sql
21. Compare lung cancer prevalence in men vs. women across countrie
WITH LungCancerCTE
AS
(
SELECT
   Country,
    CASE
        WHEN Gender = 'Male' THEN 'Men'
    ELSE 'Women'
    END AS Gender,
Lung_Cancer_Prevalence_Rate
FROM Lung_Cancer_Data
)
SELECT
     Country,
     Gender,
     ROUND(AVG(Lung_Cancer_Prevalence_Rate), 2) AS
Avg_Lung_Cancer_Prevalence_Rate
FROM LungCancerCTE
GROUP BY
    Country,
    Gender
ORDER BY
    Country,
    Gender;
```

| Country | Gender | Avg_Lung_Cancer_Prevalence_Rate |
|---|---|---|
| Bangladesh | Men | 1.5 |
| Bangladesh | Women | 1.5 |
| Brazil | Men | 1.52 |
| Brazil | Women | 1.49 |
| China | Men | 1.51 |
| China | Women | 1.49 |
| DR Congo | Men | 1.5 |
| DR Congo | Women | 1.49 |
| Egypt | Men | 1.51 |
| Egypt | Women | 1.49 |
| Ethiopia | Men | 1.5 |
| Ethiopia | Women | 1.51 |
| France | Men | 1.51 |
| France | Women | 1.49 |
| Germany | Men | 1.49 |
| Germany | Women | 1.51 |
| India | Men | 1.5 |
| India | Women | 1.52 |
| Indonesia | Men | 1.5 |
| Indonesia | Women | 1.51 |
| Iran | Men | 1.51 |
| Iran | Women | 1.51 |
| Italy | Men | 1.5 |
| Italy | Women | 1.5 |
| Japan | Men | 1.5 |

Futurion
UPSKILLING INDIA

**22. Find how occupational exposure, smoking, and air pollution collectively impact lung cancer rates.**

```
WITH
ExposureImpact
AS (
SELECT
CASE
WHEN Lung_Cancer_Diagnosis = 'Yes' THEN 'Diagnosed'
WHEN Lung_Cancer_Diagnosis = 'No' THEN 'Not Diagnosed'
END AS Diagnosis_Status,
COUNT(*) AS Nr_of_Cases,
SUM(CASE WHEN Occupational_Exposure = 'Yes' THEN 1 ELSE 0 END) AS Occupational_Exposure_Cases,
SUM(CASE WHEN Smoker = 'Yes' THEN 1 ELSE 0 END) AS Smoking_Cases,
SUM(CASE WHEN Air_Pollution_Exposure = 'High' THEN 1 ELSE 0 END) AS High_Air_Pollution_Cases,
SUM(CASE WHEN Air_Pollution_Exposure = 'Medium' THEN 1 ELSE 0 END) AS Medium_Air_Pollution_Cases,
SUM(CASE WHEN Air_Pollution_Exposure = 'Low' THEN 1 ELSE 0 END) AS Low_Air_Pollution_Cases
FROM Lung_Cancer_Data
GROUP BY
CASE
WHEN Lung_Cancer_Diagnosis = 'Yes' THEN 'Diagnosed'
WHEN Lung_Cancer_Diagnosis = 'No' THEN 'Not Diagnosed'
END
)
SELECT
    Diagnosis_Status,
    Nr_of_Cases,
    (Occupational_Exposure_Cases * 100.0) / Nr_of_Cases AS Occupational_Exposure_Percent,
    (Smoking_Cases * 100.0) / Nr_of_Cases AS Smoking_Percentage,
    (High_Air_Pollution_Cases * 100.0) / Nr_of_Cases AS High_Air_Pollution_Percent,
    (Medium_Air_Pollution_Cases * 100.0) / Nr_of_Cases AS Medium_Air_Pollution_Percent,
    (Low_Air_Pollution_Cases * 100.0) /Nr_of_Cases AS Low_Air_Pollution_Percent
FROM ExposureImpact;
```

Futurion
UPSKILLING INDIA

| Diagnosis_Status | Nr_of_Cases | Occupational_Exposure_Percent | Smoking_Percenta... | High_Air_Pollution_Per... | Medium_Air_Pollution_Percent | Low_Air_Pollution_Percent |
|---|---|---|---|---|---|---|
| Diagnosed | 8961 | 30.867090726481 | 69.735520589219 | 24.986050663988 | 50.195290704162 | 24.818658631849 |
| Not Diagnosed | 211671 | 30.117966088883 | 38.782828068086 | 24.976968975438 | 49.916615880304 | 25.106415144256 |

**23. Analyze the impact of early detection**

```
SELECT
Cancer_Stage,
Early_Detection,
AVG(Survival_Years * 1.00) Avg_Survival_ye
FROM Lung_Cancer_Data
WHERE
Lung_Cancer_Diagnosis = 'Yes'
GROUP BY
Cancer_Stage,
Early_Detection
ORDER BY
Cancer_Stage
```

| Cancer_Stage | Early_Detection | Avg_Survival_years |
|---|---|---|
| Stage 1 | No | 5.370440 |
| Stage 1 | Yes | 5.557404 |
| Stage 2 | No | 5.604828 |
| Stage 2 | Yes | 5.576923 |
| Stage 3 | No | 5.535922 |
| Stage 3 | Yes | 5.589062 |
| Stage 4 | Yes | 5.511078 |
| Stage 4 | No | 5.423976 |

# Lung Cancer Overview

Click on Below Button to Navigate Through Dashboard

Clear all slicers

Adenocarcinoma_Type
All

Country Status
All

Cancer_Stage
All

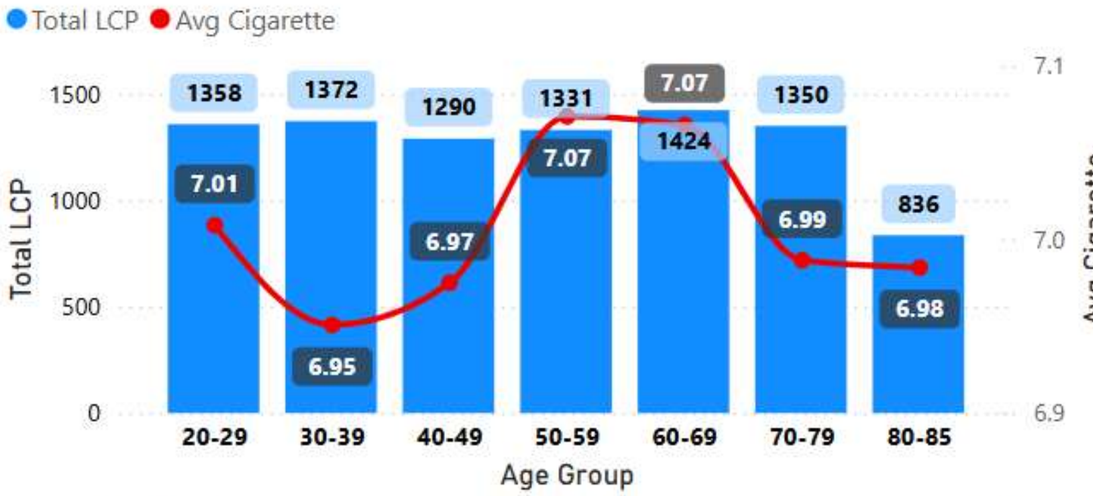
Early_Detection
All

Air_Pollution_Exposure
All

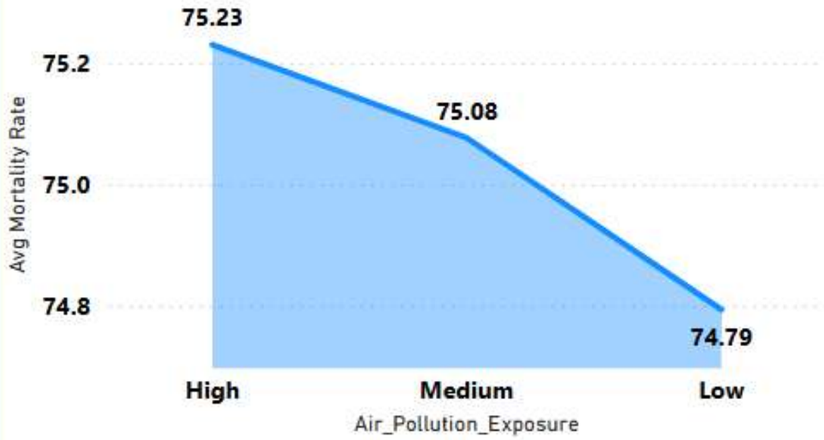
| Total Lung Cancer Cases | Avg Age of LC Patients | Smokers with LC % | Avg Mortality Rate for LCP |
|---|---|---|---|
| 8961 | 52.66 | 7.07% | 75.09 |

## Global Lung Cancer Incidence by Country

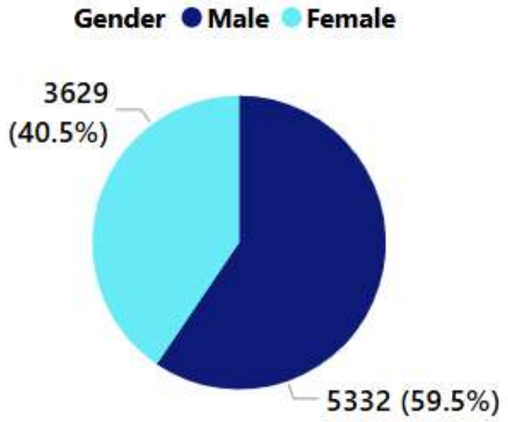

## Distribution of Lung Cancer Cases by Age Group & Avg Nr of Cigarrete



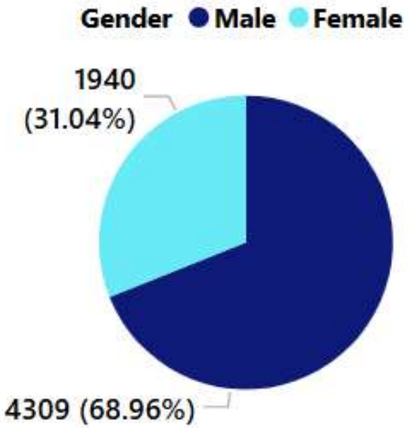

## Avg Mortality Rate by Air_Pollution_Exposure




## Lung Cancer Cases by Gender




## Smoker's Lung Cancer Cases by Gender



Click on Below Button to Navigate Through Dashboard

# Smoking & Risk Factors

| Total Smokers | Avg Years of Smoking | High Env Risk Patient | Early Detection Rate |
|---|---|---|---|
| 88,341 | 20.42 | 2542 | 28.37% |

Clear all slicers

Adenocarcinoma_Type: All

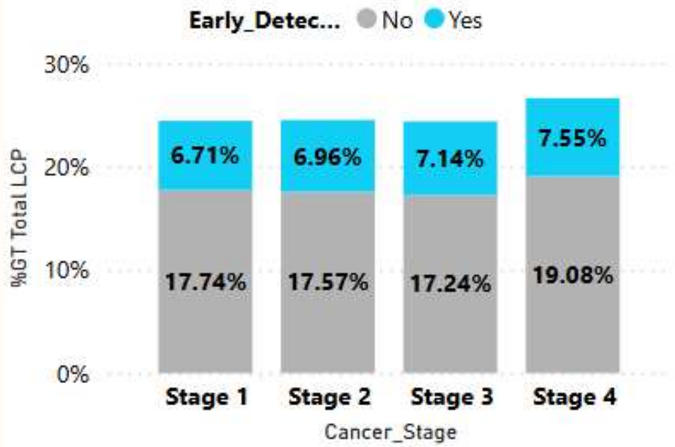
Country Status: All

Cancer_Stage: All

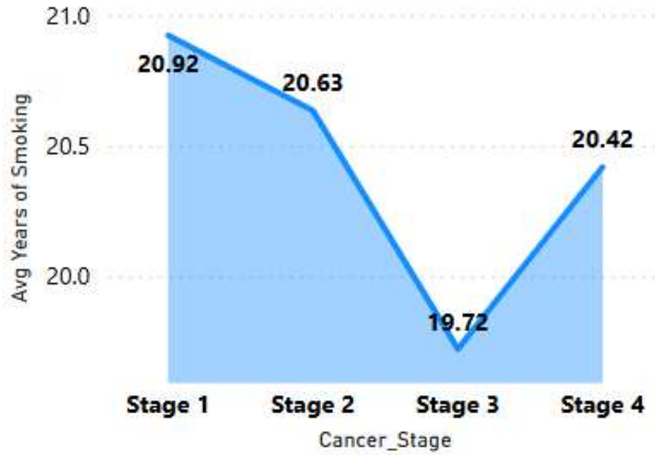
Early_Detection: All

Air_Pollution_Exposure: All

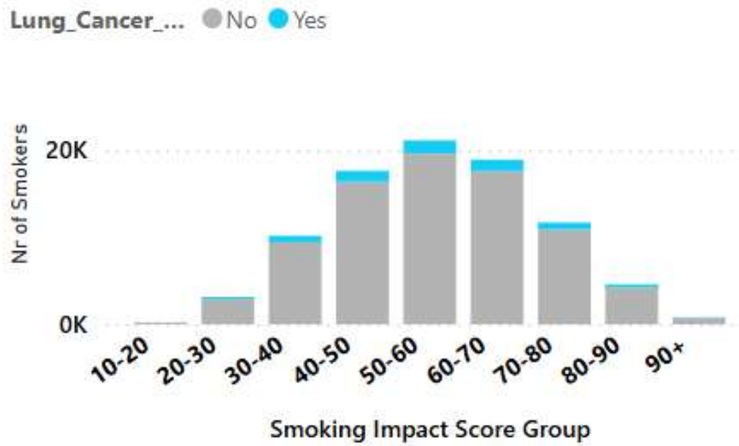
## Distribution Lung Cancer Cases(%) by Cancer Stage & Early Detection



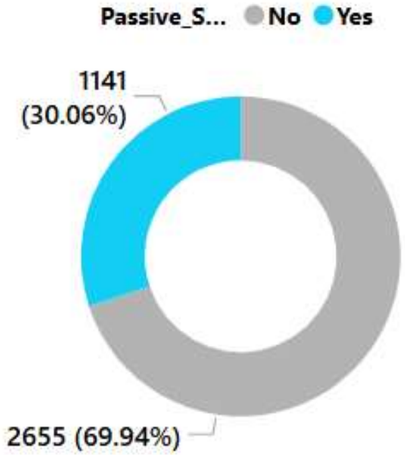

## Distribution Lung Cancer Cases(%) by Cancer Stage & Early Detection



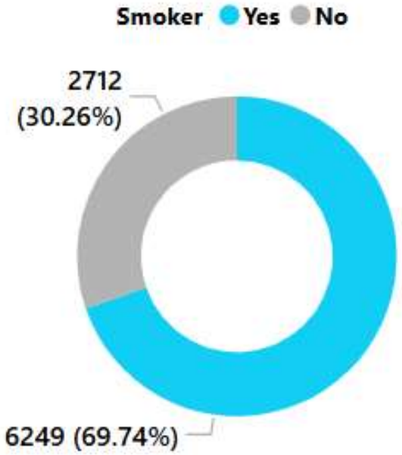

## Distribution of Lung Cancer Patient by Lung Cancer Diagnosis and Smoking Impact Score Score



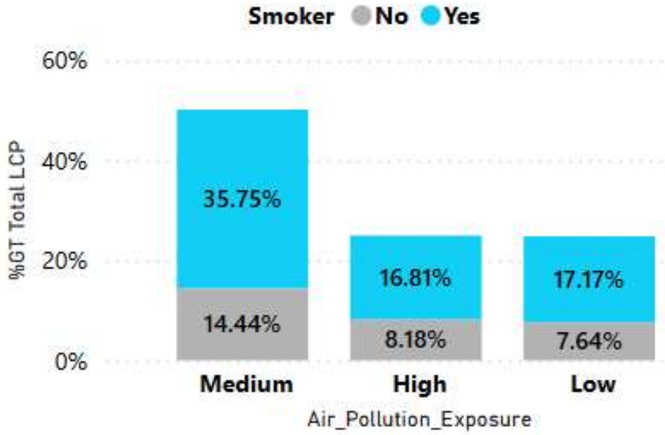

## Lung Cancer Cases by Passive Smoking




## Distribution of Lung Cancer Cases by Smokers




## Lung Cancer cases by Air pollution Exposure & Smokers



Click on Below Button to Navigate Through Dashboard

# Treatment & Survival Analysis



Clear all slicers

Adenocarcinoma_Type
All

Country Status
All

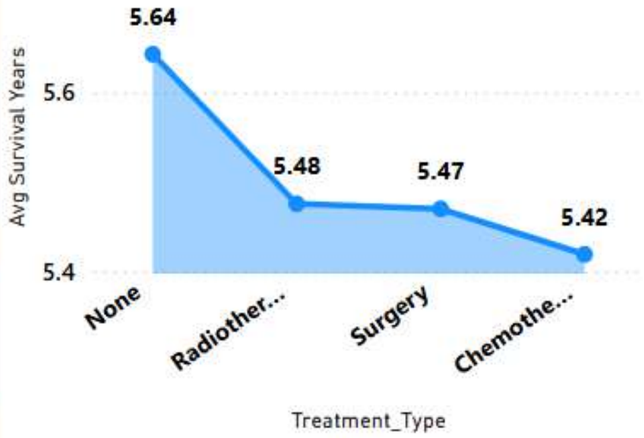
Cancer_Stage
All

Early_Detection
All

Air_Pollution_Exposure
All

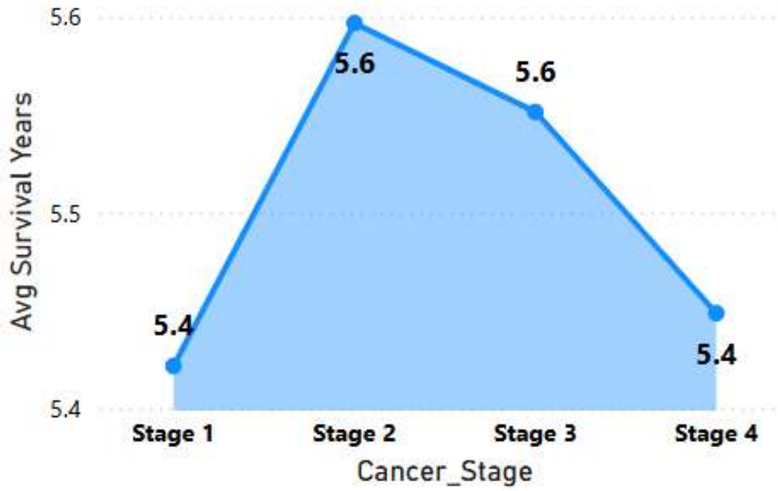
| Avg Survival Years | Total LC Deaths | Avg LC Prevalence Rate | Survival Rate by Early Detection |
|---|---|---|---|
| 5.5 | 14bn | 1.50 | 28.65% |

## Avg Survival Years by Treatment Type

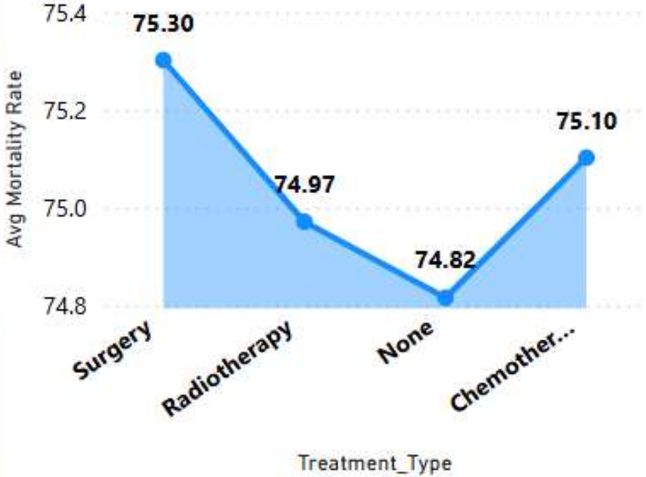



| Country | Average of Annual_LC Deaths |
|---|---|
| China | 690000 |
| USA | 130000 |
| Japan | 75000 |
| India | 70000 |
| Russia | 60000 |
| Myanmar | 59999 |
| Myanmar | 59989 |
| Myanmar | 59986 |
| Myanmar | 59980 |
| Myanmar | 59963 |
| Myanmar | 59961 |
| Myanmar | 59959 |

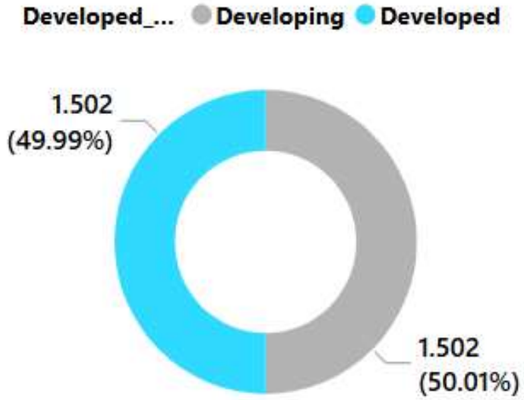
## Avg Survival Years by Cancer Stage



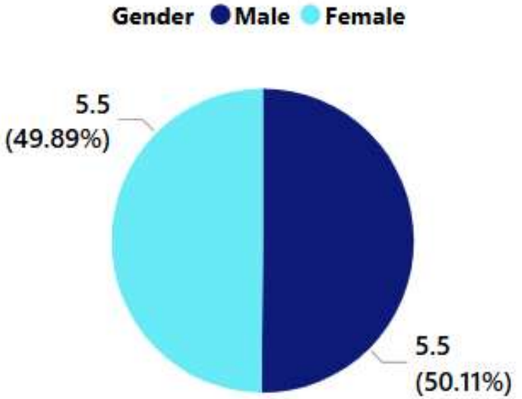

## Avg Mortality Rate by Treatment Type




## Avg Lung Cancer Prevalence by Country Status




## Avg Survival Years by Gender



Futurion
UPSKILLING INDIA

# Instructions

## Insights

### High-Level Observations on Lung Cancer

- **Total Lung Cancer Cases:** 8,961 – Highlights the disease's prevalence.
- **Average Age of Patients:** 52.66 years – Most cases occur in middle-aged or older individuals.
- **Smokers with Lung Cancer:** 7.07% – Indicates other factors (environmental, genetic) also play a role.
- **Average Mortality Rate:** 75.09% – High mortality underscores the need for early detection.

### Geographic Distribution

High cases in **North America, Europe, and Asia**, possibly due to industrialization, air pollution, and smoking.

### Age-Wise Distribution

- Cases remain **consistent from ages 20-79 (1,300–1,400 per group)**.
- **Steep decline in cases (836) at 80-85 age group**, likely due to mortality before reaching this age.
- Cigarette consumption is constant (~33K) until it declines sharply in **80+ age group (20.2K).**

### Air Pollution & Mortality Impact

**Mortality Rate by Air Pollution Levels:**

- **High:** 75.23%
- **Medium:** 75.08%
- **Low:** 74.79%

- **Insight:** Pollution worsens outcomes, but other factors like smoking and late detection play bigger roles.

Futurion
UPSKILLING INDIA

# Instructions

## Insights

**Smoking & Lung Cancer Risk**

- **Total Smokers:** 88,341 – Indicates a large high-risk population.
- **Early Detection Rate:** 28.37% – Over **70% of lung cancer cases are detected late**, reducing survival chances.

**Passive Smoking Impact:**

- **1 in 3 lung cancer cases (30.06%) are from passive smokers.**
- **Non-smokers (69.74%) also develop lung cancer, proving other risk factors matter.**

**Air Pollution vs. Smoking Risk**

- **Highest lung cancer cases** occur in **medium & high pollution exposure areas.**
- **Even in low-pollution areas, smokers still have a high risk, proving smoking is a dominant risk factor.**

**Gender-Based Analysis**

**Men (5,332 cases, 59.5%) are more affected than women (3,629 cases, 40.5%)**.

**Smoker's Lung Cancer Cases:**

- **Males:** 4,309 cases (68.96%)
- **Females:** 1,940 cases (31.04%)

**Insight:** Smoking has a **stronger impact on men**, likely due to **historical smoking trends, occupational hazards, and lifestyle.**

Futurion
UPSKILLING INDIA

# Instructions

## Recommendations

**Age-Based Screening**

- Since cases remain **consistent across ages 20-79**, early detection should start **in the 20s or 30s**, not just for people aged 50+.

**Air Pollution Awareness & Policies**

- Although mortality is **high across all pollution levels**, **air pollution control should be prioritized** to reduce compounding risks.

**Targeted Male-Focused Anti-Smoking Campaigns**

- **68.96% of male lung cancer cases are smoking-related** → Anti-smoking efforts should be **specifically aggressive toward men.**

**Stricter Passive Smoking Regulations**

- **30.06% of passive smokers develop lung cancer → Public smoking bans, stricter home & office rules needed** to reduce second-hand smoke exposure.

**Stronger Early Detection Strategies**

- **Since only 28.37% of lung cancer cases are detected early, routine screenings (CT scans) should be mandatory for high-risk groups (smokers, passive smokers, and pollution-exposed individuals).**

**Occupational Risk Prevention**

- **Males are disproportionately affected**, likely due to workplace exposure (factories, mining, construction).
- **Stronger safety measures, improved ventilation, and regular health checkups should be implemented in high-risk jobs.**

**Pollution & Smoking Control Together**

- **Air pollution increases cancer risk for smokers & non-smokers alike → Governments should control industrial emissions, promote clean energy, and plant more green spaces.**