Futurion
UPSKILLING
INDIA

Data Analysis

Portfoliio Project Challenge

Lung Cancer Analytics: Insights for Early Detection & Prevention

By Abhijeet Gaykwad

2/21/2025

Tools Used : SQL & Power Bi

Futurion
UPSKILLING INDIA

Problem Statement & Objectives

Problem Statement:

Lung cancer is one of the leading causes of death worldwide. Early detection and analysis of patient data can help in better diagnosis and treatment.

Objectives:

- ☑ Identify key risk factors from the dataset
- ☑ Analyze patient trends (e.g., age, gender, smoking history)
- ☑ Provide insights for early detection

Futurion
UPSKILLING INDIA

Dataset Overview

Dataset Details:

📌 **Number of Records: 220632**

📌 **Number of Columns: 24**

📌 **Main Features (Columns):**

- **Age** (Patient's age)
- **Gender** (Male/Female)
- **Smoking History** (Yes/No)
- **Coughing, Fatigue, Chest Pain, Weight Loss** (Symptoms)
- **Lung Cancer Diagnosis** (Positive/Negative)

Futurion

UPSKILLING INDIA

SQL Report

Futurion
UPSKILLING INDIA

Basic Level Question & Answer

```sql
-- 1.Retrieve all records for individuals diagnosed with lung cancer
SELECT * FROM  lung_cancer_analysis
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	S
26	Pakistan	225	40	Female	Yes	11	17	No	No	Yes	S
32	Nigeria	206	55	Male	Yes	9	8	No	Yes	Yes	S
33	Turkey	85	33	Male	Yes	4	12	No	Yes	Yes	S
93	UK	67	61	Male	Yes	14	28	No	No	Yes	S
106	Ethiopia	120	70	Male	Yes	7	21	No	No	Yes	S
157	Germany	83	72	Male	Yes	25	26	No	No	Yes	S
168	Indonesia	273	47	Female	No	0	0	Yes	No	Yes	S
188	Egypt	102	71	Male	Yes	36	7	Yes	No	Yes	S

```sql
-- 2.Count the number of smokers and non-smokers
SELECT Smoker, COUNT(*) AS Count
FROM  lung_cancer_analysis
GROUP BY Smoker;
```

Smoker	Count
Yes	11993
No	18364

```sql
-- 3.List all unique cancer stages present in the dataset
SELECT DISTINCT Cancer_Stage
FROM  lung_cancer_analysis;
```

Cancer_Stage
None
Stage 1
Stage 2
Stage 3
Stage 4

```sql
-- 4. Retrieve the average number of cigarettes smoked per day by smokers
SELECT AVG(Cigarettes_per_Day) AS Avg_Cigarettes_Per_Day
FROM lung_cancer_analysis
WHERE Smoker = 'Yes';
```

Avg_Cigarettes_Per_Day
17.3966

```sql
-- 5.Count the number of people exposed to high air pollution
SELECT COUNT(*) AS Count
FROM lung_cancer_analysis
WHERE Air_Pollution_Exposure = 'High';
```

Count
7623

Futurion
UPSKILLING INDIA

```sql
-- 6. Find the top 5 countries with the highest lung cancer deaths

SELECT Country, SUM(Annual_Lung_Cancer_Deaths) AS Total_Deaths
FROM lung_cancer_analysis
GROUP BY Country
ORDER BY Total_Deaths DESC
LIMIT 5;
```

Country	Total_Deaths
China	841800000
USA	156260000
Japan	94500000
India	85330000
Russia	73800000

```sql
-- 7.Count the number of people diagnosed with lung cancer by gender

SELECT Gender, COUNT(*) AS Count
FROM lung_cancer_analysis
WHERE Lung_Cancer_Diagnosis = 'Yes'
GROUP BY Gender;
```

Gender	Count
Female	471
Male	755

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

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

Age	Gender	Smoker	Years_of_Smoking	Cigarettes_per_Day	Passive_Smoker	Family_History	Lung_Cancer_Diagnosis
61	Male	Yes	14	28	No	No	Yes
70	Male	Yes	7	21	No	No	Yes
72	Male	Yes	25	26	No	No	Yes
71	Male	Yes	36	7	Yes	No	Yes
85	Male	Yes	40	26	No	No	Yes
66	Male	Yes	25	6	No	No	Yes
73	Male	Yes	27	24	No	No	Yes
72	Male	Yes	21	27	No	No	Yes
81	Male	Yes	31	17	No	Yes	Yes
64	Male	Yes	1	23	No	No	Yes
73	Male	Yes	25	26	Yes	No	Yes
64	Male	Yes	13	14	No	No	Yes
65	Male	No	0	0	No	No	Yes
63	Male	No	0	0	No	No	Yes
66	Male	Yes	8	10	No	No	Yes
68	Male	Yes	18	21	No	No	Yes

Intermediate Level Question & Answer

```sql
-- 1.Find the percentage of smokers who developed lung cancer
SELECT
    (COUNT(CASE WHEN Lung_Cancer_Diagnosis = 'Yes' THEN 1 END) * 100.0 / COUNT(*)) AS Percentage_Smokers_Lung_Cancer
FROM lung_cancer_analysis
WHERE Smoker = 'Yes';
```

Percentage_Smokers_Lung_Cancer
7.12916

```sql
-- 2. Calculate the average survival years based on cancer stages
SELECT Cancer_Stage, AVG(Survival_Years) AS Avg_Survival_Years
FROM lung_cancer_analysis
GROUP BY Cancer_Stage;
```

Cancer_Stage	Avg_Survival_Years
None	0.0000
Stage 1	5.7526
Stage 2	5.4583
Stage 3	5.3533
Stage 4	4.9969

```sql
-- 3.Count the number of lung cancer patients based on passive smoking
SELECT Passive_Smoker, COUNT(*) AS Count
FROM lung_cancer_analysis
WHERE Lung_Cancer_Diagnosis = 'Yes'
GROUP BY Passive_Smoker;
```

Passive_Smoker	Count
No	870
Yes	356

```sql
-- 4. Find the country with the highest lung cancer prevalence rate
SELECT Country, MAX(Lung_Cancer_Prevalence_Rate) AS Max_Prevalence
FROM lung_cancer_analysis
GROUP BY Country;
```

Country	Max_Prevalence
China	2.5
Iran	2.5
Mexico	2.5
Indonesia	2.49
South Africa	2.5
India	2.5
Myanmar	2.5
Ethiopia	2.5
Nigeria	2.5
Egypt	2.5
Italy	2.5
France	2.5
Germany	2.5
Brazil	2.5
Turkey	2.5
Thailand	2.5
Japan	2.5
Pakistan	2.5
Philippines	2.5
Bangladesh	2.5
Russia	2.5
Vietnam	2.5
USA	2.5
DR Congo	2.5
UK	2.5

```sql
-- 5.Identify the smoking years' impact on lung cancer
SELECT Years_of_Smoking, COUNT(*) AS Lung_Cancer_Count
FROM lung_cancer_analysis
WHERE Lung_Cancer_Diagnosis = 'Yes'
GROUP BY Years_of_Smoking
ORDER BY Years_of_Smoking;
```

Years_of_Smoking	Lung_Cancer_Count
0	371
1	19
2	15
3	22
4	25
5	24
6	15
7	21
8	19
9	22
10	21
11	23
12	23
13	29
14	30
15	26
16	22
17	18
18	24
19	27
20	25
21	15
22	15
23	28
24	15
25	15
26	22
27	19
28	28
29	22

```
-- 6.Determine the mortality rate for patients with and without early detection

SELECT Early_Detection, AVG(Mortality_Rate) AS Avg_Mortality_Rate
FROM lung_cancer_analysis
GROUP BY Early_Detection;
```

Early_Detection	Avg_Mortality_Rate
No	3.042668321154468 7
Yes	3.016850430332636 4

```
-- 7.Group the lung cancer prevalence rate by developed vs. developing countries

SELECT Developed_or_Developing, AVG(Lung_Cancer_Prevalence_Rate) AS Avg_Prevalence_Rate
FROM lung_cancer_analysis
GROUP BY Developed_or_Developing;
```

Developed_or_Developing	Avg_Prevalence_Rate
Developing	1.5010629510477334
Developed	1.497745751189665 5

Futurion
UPSKILLING INDIA

Advanced Level Question & Answer

```sql
-- 1.Identify the correlation between lung cancer prevalence and air pollution levels
SELECT Air_Pollution_Exposure, AVG(Lung_Cancer_Prevalence_Rate) AS Avg_Prevalence
FROM lung_cancer_analysis
GROUP BY Air_Pollution_Exposure;

-- 2.Find the average age of lung cancer patients for each country
SELECT Country, AVG(Age) AS Avg_Age
FROM lung_cancer_analysis
WHERE Lung_Cancer_Diagnosis = 'Yes'
GROUP BY Country;

-- 3.Calculate the risk factor of lung cancer by smoker status, passive smoking, and family history
SELECT Smoker, Passive_Smoker, Family_History, COUNT(*) AS Risk_Factor
FROM lung_cancer_analysis
WHERE Lung_Cancer_Diagnosis = 'Yes'
GROUP BY Smoker, Passive_Smoker, Family_History;

-- 4. Rank countries based on their mortality rate
SELECT Country, AVG(Mortality_Rate) AS Avg_Mortality_Rate
FROM lung_cancer_analysis
GROUP BY Country
ORDER BY Avg_Mortality_Rate DESC;

-- 5.Determine if treatment type has a significant impact on survival years
SELECT Treatment_Type, AVG(Survival_Years) AS Avg_Survival_Years
FROM lung_cancer_analysis
GROUP BY Treatment_Type;
```

1.

Air_Pollution_Exposure	Avg_Prevalence
Low	1.5027923439958084
Medium	1.5013670064874856
High	1.4955293191656858

2.

Country	Avg_Age
Pakistan	49.0286
Nigeria	53.7959
Turkey	51.5000
UK	52.0784
Ethiopia	54.8627
Germany	53.1111
Indonesia	53.8333
Egypt	54.3023
Iran	53.7692
Russia	53.2619
DR Congo	52.6304
Mexico	52.4091
Philippines	48.3750
China	51.8140
Vietnam	51.9600
Thailand	50.9800
Brazil	52.6818
Japan	50.2045
India	49.8214
USA	49.5455
South Africa	49.3455
Bangladesh	53.3061
Myanmar	46.1154
France	50.8036
Italy	52.6667

3.

Smoker	Passive_Smoker	Family_History	Risk_Factor
Yes	No	No	524
Yes	No	Yes	92
No	Yes	No	97
Yes	Yes	No	210
No	No	Yes	39
No	No	No	215
No	Yes	Yes	20
Yes	Yes	Yes	29

4.

Country	Avg_Mortality_Rate
USA	3.5132196339434287
Indonesia	3.4622527944969903
India	3.4187858900738326
Philippines	3.392479806138934
Turkey	3.383904991948 47
Thailand	3.330347222222222
Iran	3.3184412265758088
South Africa	3.266661367249 6028
Myanmar	3.2632722731057453
Ethiopia	3.213679401993355
France	3.19890494296578
Italy	3.159695885509839
UK	3.1462077493816984
Vietnam	3.139504950495 05
Nigeria	3.138257956448 9115
Bangladesh	2.9482555282555283
DR Congo	2.7560225442834136
Brazil	2.7512187247780457
Mexico	2.726140939597315
Egypt	2.7248720066061107
Germany	2.7123656776263028
China	2.6555245901639 35
Japan	2.6114603174603177
Russia	2.548861788617886
Pakistan	2.149689075630253

5.

Treatment_Type	Avg_Survival_Years
None	0.0533
Chemotherapy	5.2525
Radiotherapy	5.4360
Surgery	5.4434

Futurion
UPSKILLING INDIA

```sql
-- 6.Compare lung cancer prevalence in men vs. women across countries

SELECT Country, Gender, AVG(Lung_Cancer_Prevalence_Rate) AS Avg_Prevalence
FROM lung_cancer_analysis
WHERE Lung_Cancer_Diagnosis = 'Yes'
GROUP BY Country, Gender;
```

Country	Gender	Avg_Prevalence
UK	Male	1.6391176470588238
Ethiopia	Male	1.5140625
Germany	Male	1.3253125
Indonesia	Female	1.5930769230769226
Egypt	Male	1.3578124999999999
Iran	Male	1.4803030303030307
Russia	Male	1.5654166666666667
DR Congo	Female	1.4223809523809525
Mexico	Female	1.5505555555555552
Indonesia	Male	1.4614285714285715
DR Congo	Male	1.6916
UK	Female	1.5117647058823533
Philippines	Female	1.518125
China	Male	1.5965517241379312
Vietnam	Female	1.6845833333333333
Thailand	Female	1.2742105263157892
Brazil	Male	1.7577272727272728
Japan	Male	1.6256666666666668
Turkey	Female	1.51
India	Female	1.456666666666667
Thailand	Male	1.4396774193548387
USA	Male	1.4491891891891886
South Africa	Male	1.5802857142857143
Brazil	Female	1.6209090909090909
Ethiopia	Female	1.4157894736842103
Philippines	Male	1.404
Nigeria	Female	1.5849999999999997

```sql
-- 7.Find how occupational exposure, smoking, and air pollution collectively impact lung cancer rates

SELECT Occupational_Exposure, Smoker, Air_Pollution_Exposure, COUNT(*) AS Lung_Cancer_Count
FROM lung_cancer_analysis
WHERE Lung_Cancer_Diagnosis = 'Yes'
GROUP BY Occupational_Exposure, Smoker, Air_Pollution_Exposure;
```

Occupational_Exposure	Smoker	Air_Pollution_Exposure	Lung_Cancer_Count
No	Yes	Low	144
Yes	Yes	Medium	144
No	Yes	Medium	313
Yes	No	High	30
Yes	Yes	Low	63
No	Yes	High	125
No	No	Medium	126
No	No	High	67
Yes	No	Low	36
Yes	Yes	High	66
No	No	Low	61
Yes	No	Medium	51

```sql
-- 8.Analyze the impact of early detection on survival years

SELECT Early_Detection, AVG(Survival_Years) AS Avg_Survival_Years
FROM  lung_cancer_analysis
WHERE Lung_Cancer_Diagnosis = 'Yes'
GROUP BY Early_Detection;
```

Early_Detection	Avg_Survival_Years
No	5.3943
Yes	5.3382

Futurion
UPSKILLING INDIA

Power BI

Visualization Reports

Futurion
UPSKILLING INDIA

Lung Cancer Overview

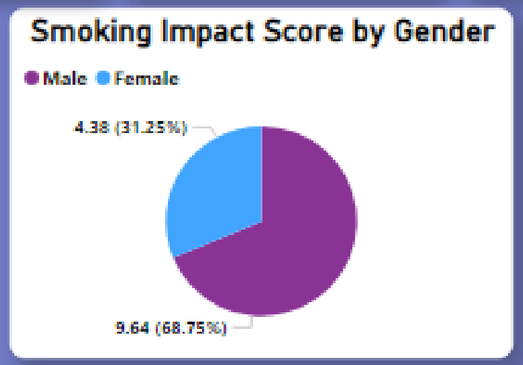
8961
Total_Lung_Cancer_Cases

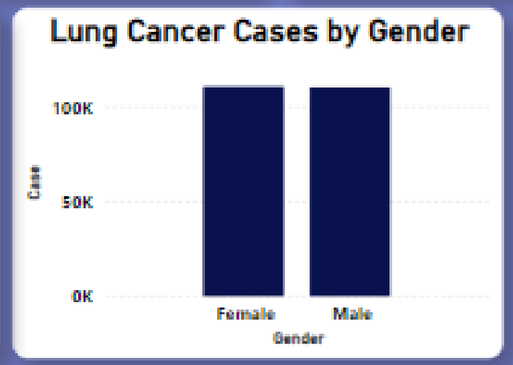
52.66
Avg_Age_Lung_Cancer

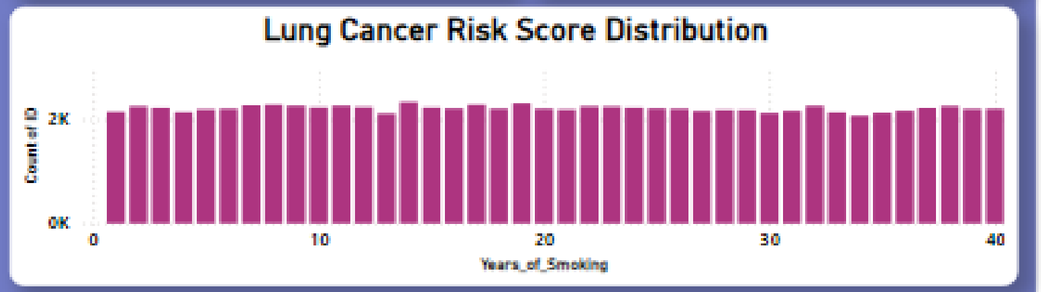
69.74
Percentage_Smokers_Cancer

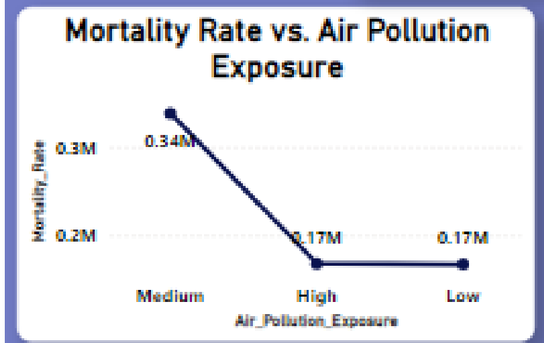
3.05
Mortality Rate for Lung Canc...

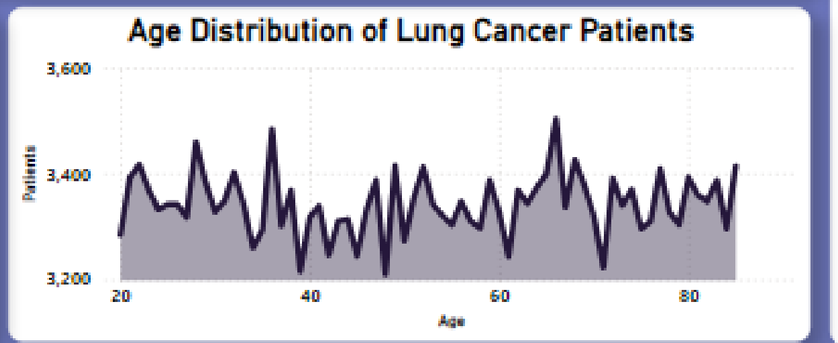
Cancer_Stage
- Stage 1
- Stage 2
- Stage 3
- Stage 4











Key insights

Lung Cancer Overview

Males exhibit a higher prevalence of lung cancer compared to females.

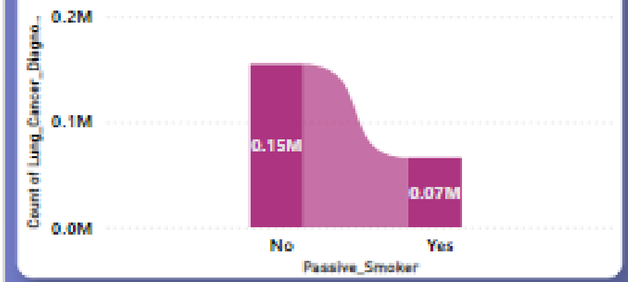
Smokers constitute the majority of diagnosed cases, reinforcing smoking as a major risk factor.

Countries with high air pollution levels report significantly higher lung cancer cases.

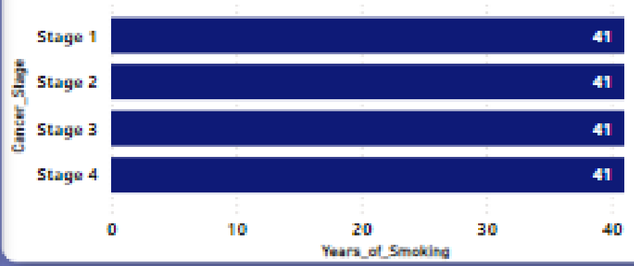
Smoking and Risk Factors Analysis

88K	8.18	34K	0.28
Total_Smokers	Avg_Years_of_Smoking	High_Risk_Patients	Early_Detection_Rate

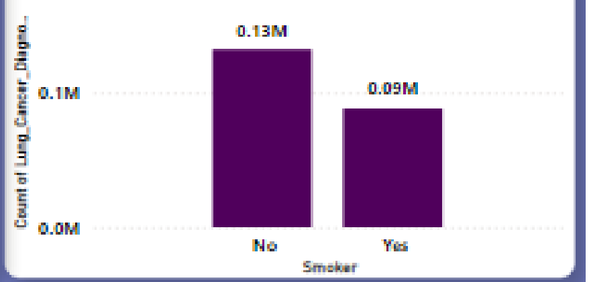
Passive Smoking Impact on Cancer Cases



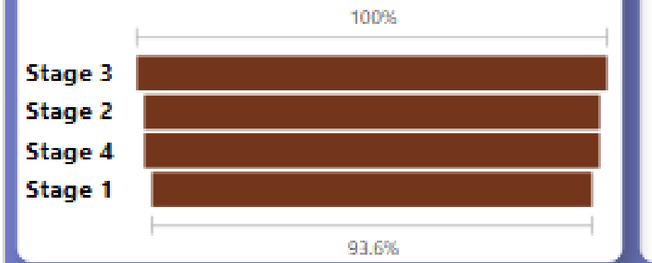
Years of Smoking vs. Cancer Stage



Smoking vs. Lung Cancer Cases

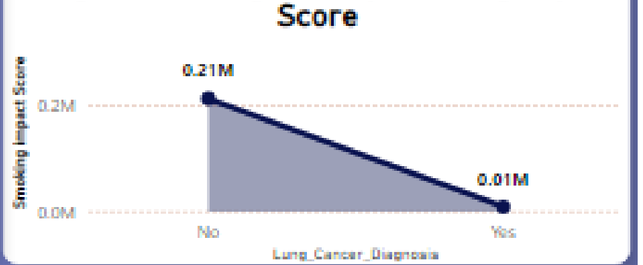


Early Detection Rate vs. Cancer Stage



Smoking and Air Pollution Exposure Relationship

Lung Cancer Diagnosis by Smoking Impact Score



Age

20 — 85

Survival_Years

1	2	3	4	5
6	7	8	9	10

key insights

Smoking & Risk Factors

Prolonged smoking duration directly correlates with advanced cancer stages.

Passive smokers also show a noticeable increase in lung cancer cases.

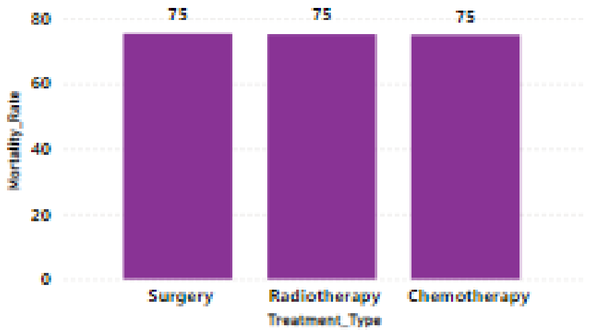
Air pollution and occupational exposure amplify lung cancer risk across regions.

Futurion
UPSKILLING INDIA

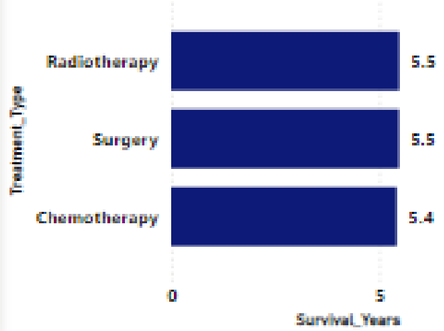
Treatment & Survival Analysis

0.22	8079	1.50	0.50
Avg_Survival_Years	Total Annual Lung Cancer Deaths	Lung Cancer Prevalence Rate	Survival_Rate_Early_Detection

Mortality Rate vs. Treatment Type

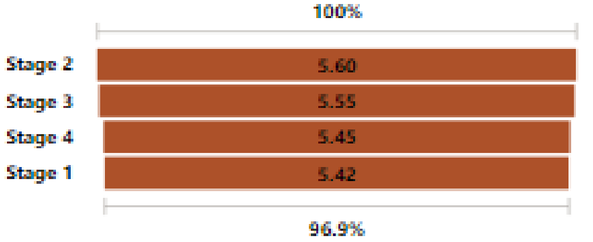


Survival Years by Treatment Type

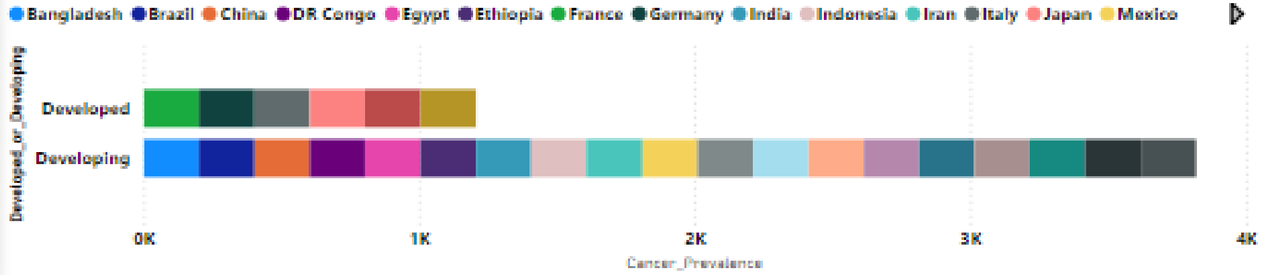


Lung Cancer Deaths by Country

Survival Rate Distribution by Cancer Stage



Developed vs. Developing Countries: Cancer Prevalence



Survival years over gender

Banglad...	Brazil	China	DR Congo	Egypt	Ethiopia	France	Germany
India	Indonesia	Iran	Italy	Japan	Mexico	Myanmar	Nigeria
Pakistan	Philippin...	Russia	South Af...	Thailand	Turkey	UK	USA

key insights

Treatment & Survival Analysis

Early detection significantly improves survival rates, with Stage 1 patients having the highest longevity.

Mortality rates are higher in developing nations, emphasizing the need for better healthcare infrastructure.

Patients receiving advanced treatments show improved survival outcomes compared to those with basic treatment plans.

Futurion
UPSKILLING INDIA

Conclusion & Recommendations

Key Findings & Insights:

1.**Smoking remains the leading cause of lung cancer**, with long-term smokers facing the highest risk.

2. **Early detection significantly increases survival rates**, as patients diagnosed at Stage 1 or 2 have better treatment outcomes.

3.**Environmental factors like air pollution and occupational hazards contribute to rising lung cancer cases**, especially in urban and industrial regions.

4.**Developing countries have a higher mortality rate**, mainly due to inadequate healthcare infrastructure and late diagnosis.

5.**Advanced treatment methods improve survival**, but accessibility remains a challenge for lower-income populations.

Actionable Business Recommendations:

- ☑ **Launch large-scale anti-smoking initiatives**, including higher taxation on tobacco products and public awareness campaigns.
- ☑ **Invest in AI-driven early screening programs** to detect lung cancer in high-risk individuals before it progresses.
- ☑ **Improve healthcare access in developing nations**, ensuring affordable treatment and early intervention
.
- ☑ **Strengthen industrial regulations and workplace safety policies** to limit exposure to harmful pollutants.
- ☑ **Encourage research on innovative cancer treatments**, making cutting-edge therapies more accessible and cost-effective.

By implementing these strategies, we can **reduce lung cancer cases, improve patient survival, and create a proactive approach to healthcare management**.

Futurion
UPSKILLING INDIA

References & Acknowledgments

Data Source:

•**Lung Cancer Dataset** – Includes patient demographics, smoking history, air pollution exposure, cancer diagnosis, treatment details, and survival outcomes.

Tools & Technologies Used:

•**SQL**– Used for querying and analyzing patient data.
•**Power BI** – Created interactive dashboards and visualizations.

Acknowledgments:

This project was successfully accomplished under the expert guidance and invaluable support of **Futurion Tech and Sakshi Mogal Ma’am**, whose insights and mentorship played a crucial role in shaping the analysis and outcomes.

This analysis provides **data-driven insights** to improve **early detection, risk assessment, and healthcare strategies** for lung cancer prevention and treatment.