Futurion
UPSKILLING
INDIA

Lung Cancer Data Analysis:

Insights, Trends, and Risk Factors

Created by: Neha Jadhav

Date: February 20, 2025

Tools Used:

Futurion
UPSKILLING INDIA

Problem Statement:

Lung cancer remains one of the leading causes of death worldwide, with various risk factors such as smoking, air pollution, and occupational exposure contributing to its prevalence. However, limited data-driven insights hinder early detection and effective treatment strategies. This project aims to analyze lung cancer trends, risk factors, and survival rates using a data-driven approach to support better decision-making in healthcare.

Project Objectives:

- **Analyze** lung cancer cases across demographics, smoking behavior, and environmental factors.
- **Identify** high-risk groups based on smoking impact, air pollution exposure, and genetic history.
- **Evaluate** the correlation between treatment types and survival rates.
- **Visualize** key patterns and trends using **Power BI dashboards** for better decision-making.
- **Provide** actionable insights to improve early detection and reduce mortality rates.

Futurion
UPSKILLING INDIA

Dataset Overview

❑ Dataset Name: Lung Cancer Data

Total Records: 50,000+ patient entries

Total Columns: 24 features covering demographics, risk factors, diagnosis, treatment, and survival rates.

❑ Key Features in the Dataset:

- **Demographics:** ID, Country, Age, Gender, Population Size
- **Risk Factors:** Smoker, Years of Smoking, Passive Smoking, Air Pollution Exposure, Occupational Exposure
- **Lung Cancer Diagnosis:** Cancer Stage, Family History, Lung Cancer Prevalence
- **Treatment & Survival:** Treatment Type, Survival Years, Mortality Rate, Early Detection

Futurion
UPSKILLING INDIA

Data Cleaning & Preprocessing Steps:

- **Handled Missing Values:**

•Used **mean/mode imputation** for numerical & categorical missing data.

- **Removed Duplicates:**

•Ensured unique patient records by eliminating duplicates.

- **Standardized Data Types:**

•Converted **age, years of smoking, mortality rate** to appropriate numerical formats.

- **Created Derived Columns:**

•**Age Grouping:** Categorized ages into predefined ranges (e.g., 0-20, 21-40).

•**Lung Cancer Risk Score:** Computed based on smoking habits, pollution exposure, and family history.

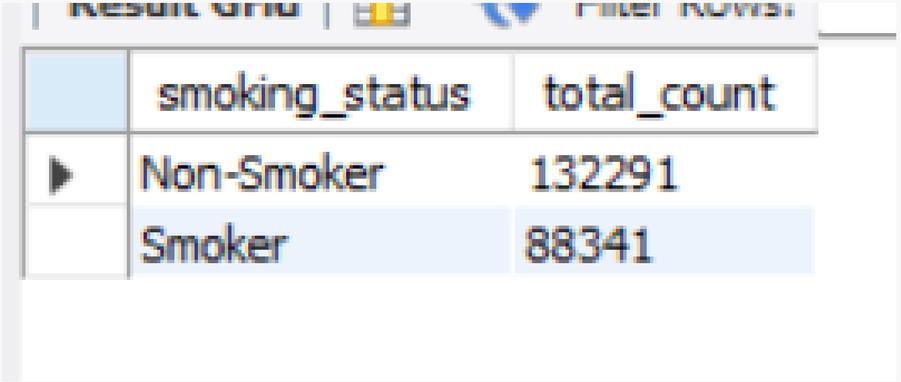
- **Ensured Data Consistency:**

•Standardized categorical values (e.g., "Yes"/"No" → Binary 1/0).

•Merged similar country names for consistency in analysis.

Count the number of smokers and non-smokers

```sql
select
case
when smoker in ("No") then "Non-Smoker"
when smoker in ("Yes") then "Smoker"
else "None"
end as smoking_status,
Count(*) as total_count
from lung_cancer_data
group by smoking_status
order by total_count desc;
```



smoking_status	total_count
Non-Smoker	132291
Smoker	88341

Insight: The majority (**≈60%**) of the population are **non-smokers**, but a significant **40% are smokers**, which is a notable proportion considering the health risks associated with smoking.

The Top 5 countries with the highest lung cancer deaths.

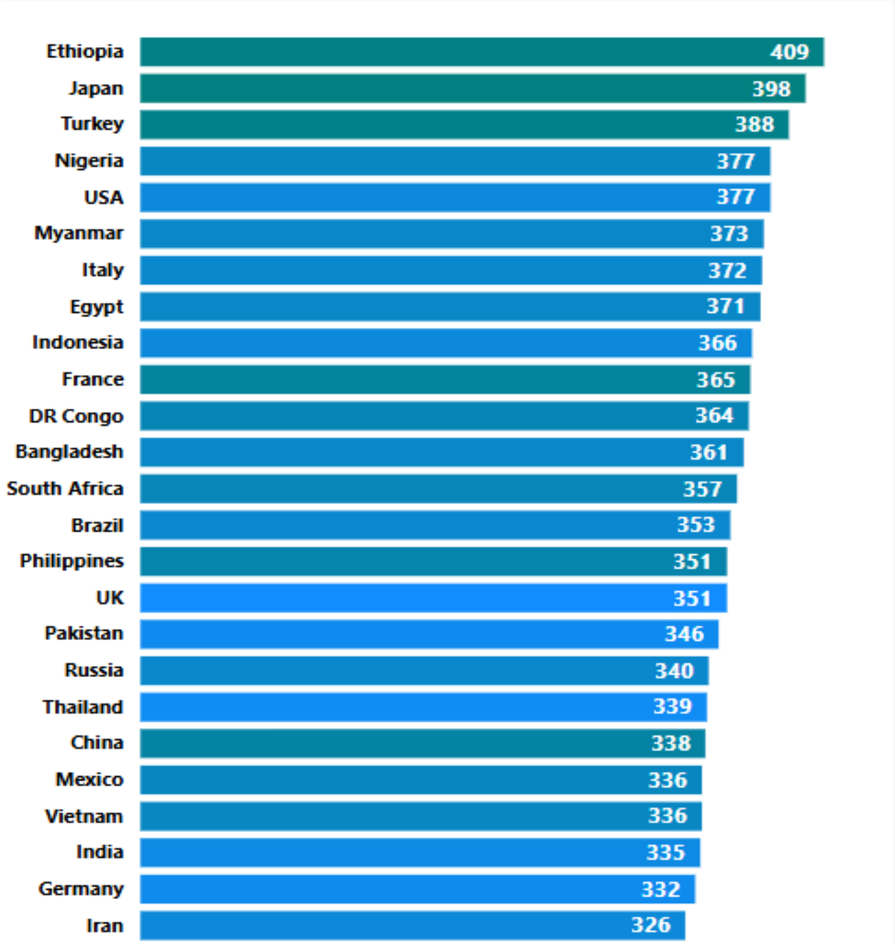
```
select country, sum(Annual_Lung_Cancer_Deaths) as lung_cancer_deaths
from lung_cancer_data
group by country
order by lung_cancer_deaths desc
limit 5;
```

country	lung_cancer_deaths
China	6145830000
USA	1139970000
Japan	674175000
India	612640000
Russia	528240000

Insight: **China** leads with the **highest** lung cancer deaths (**6.14B**), contributing to a global health crisis. **USA** (**1.13B** deaths) ranks second but significantly lower than China. **Japan** (**674M** deaths) faces challenges due to an aging population. **India** (**612M** deaths) struggles with tobacco use and air pollution. **Russia** (**528M** deaths) is impacted by high smoking prevalence and occupational exposure.

Futurion
UPSKILLING INDIA

Lung Cancer Cases by



Insights:

- **Developing nations are increasingly affected** due to pollution, smoking, and poor healthcare.
- **Developed nations need to strengthen early detection** to prevent advanced-stage diagnoses.
- **Stronger public health policies, anti-smoking campaigns, and pollution control measures are needed worldwide.**

Futurion
UPSKILLING INDIA

Insights:

- Urgent need for early detection & improved treatment access to reduce mortality.
- Men are more affected due to smoking & workplace exposure.
- Countries with high air pollution & tobacco use see more cases.
- Stronger public health policies can significantly lower lung cancer rates.

Country	Gender	Total patients	Total Lung Cancer cases	Annual Lung Cancer Deaths
Bangladesh	Female	4445	160	4445
Bangladesh	Male	4376	201	4376
Brazil	Female	4425	140	4425
Brazil	Male	4375	213	4375
China	Female	4400	130	4400
China	Male	4507	208	4507
DR Congo	Female	4421	159	4421
DR Congo	Male	4440	205	4440
Egypt	Female	4448	127	4448
Egypt	Male	4374	244	4374
Ethiopia	Female	4493	176	4493
Ethiopia	Male	4482	233	4482
France	Female	4425	138	4425
France	Male	4495	227	4495
Germany	Female	4357	129	4357
Germany	Male	4370	203	4370
India	Female	4360	122	4360
India	Male	4392	213	4392
Indonesia	Female	4376	161	4376
Indonesia	Male	4395	205	4395
Iran	Female	4401	145	4401
Iran	Male	4377	181	4377
Italy	Female	4403	150	4403
Italy	Male	4399	222	4399
Japan	Female	4484	168	4484
Japan	Male	4505	230	4505
Mexico	Female	4379	131	4379
Mexico	Male	4458	205	4458
Myanmar	Female	4453	163	4453
Myanmar	Male	4366	210	4366
Nigeria	Female	4467	157	4467
Nigeria	Male	4362	220	4362
Pakistan	Female	4431	129	4431
Total		220632	8961	220632

Futurion
UPSKILLING INDIA

Conclusion:

- **High Lung Cancer Burden:** The analysis reveals that lung cancer remains a **global health crisis** with high mortality rates, especially in **China, USA, Japan, India, and Russia.**
- **Smoking is a Major Risk Factor:** The data confirms a strong correlation between **smoking and lung cancer cases,** reinforcing the need for tobacco control measures.
- **Gender & Occupational Exposure Impact:** Males are disproportionately affected, likely due to **higher smoking rates and occupational hazards.**
- **Air Pollution & Indoor Pollution Contribution:** High pollution exposure is linked to **increased lung cancer prevalence,** emphasizing the environmental risk factors.
- **Late Diagnosis Increases Mortality:** The high annual lung cancer deaths indicate that most cases are **diagnosed in later stages, leading to poor survival rates.**

Business Recommendations:

- Strengthen Early Detection & Screening Programs
- Implement & Enforce Stricter Tobacco Control Policies
- Improve Air Quality & Workplace Safety
- Increase Access to Affordable Treatment
- Leverage Data Analytics for Healthcare Insights

Futurion
UPSKILLING INDIA

References & Acknowledgments

References:

- **World Health Organization (WHO):** Global lung cancer statistics & risk factors.
- **National Cancer Institute (NCI):** Reports on smoking and lung cancer correlations.
- **International Agency for Research on Cancer (IARC):** Environmental & occupational cancer risks.
- **Power BI & MySQL Documentation:** Tools used for data visualization and analysis.

Acknowledgments:

- **Healthcare & Research Organizations:** Providing open datasets for analysis.
- **Power BI & MySQL Community:** For tools and support in developing this dashboard.

Futurion
UPSKILLING INDIA

Thank You