

# **TECHNICAL REPORT**

## **Analysis of Lung Cancer**

### **Outlines**

Introduction

Story of data

Data splitting and Preprocessing

Pre Analysis

Post Analysis

Data Visualization and Charts

Recommendation and Observation

Conclusion

### **Introduction**

Lung cancer remains one of the most prevalent and deadliest forms of cancer globally, often diagnosed at later stages when treatment options are limited. With lifestyle, environmental, and genetic factors playing significant roles in its development, early detection and analysis of contributing variables are critical to improving patient outcomes.

This project aims to explore and analyze a structured dataset related to lung cancer patients using data analytics and visualization techniques. The dataset contains patient-level information, including demographic attributes (such as age and gender), lifestyle factors (like smoking and alcohol consumption), symptoms (e.g., coughing, fatigue), and health conditions (e.g., chronic disease, anxiety, allergy). The objective of the analysis is to uncover patterns, identify risk factors, and generate actionable insights that can support public health interventions and clinical decision-making.

The report outlines the data preprocessing steps, analytical procedures, and visualization strategies applied to extract meaningful interpretations. Ultimately, it seeks to provide evidence-based recommendations to enhance awareness, preventive care, and patient management in lung cancer cases.

## Dataset Overview

The dataset used in this project focuses on patient records related to lung cancer diagnoses. It contains 309 observations (rows), each representing an individual patient, and includes multiple categorical and numerical variables capturing key medical, demographic, and behavioral characteristics.

### Key Features of the Dataset

- Demographics:
  - GENDER: Indicates the patient's gender (Male or Female)
  - AGE: The age of the patient (numeric)
- Lifestyle Factors:
  - SMOKING: Whether the patient smokes
  - ALCOHOL CONSUMING: Alcohol consumption status
  - PEER\_PRESSURE: Influence of peer behavior on the patient
- Medical History and Symptoms:
  - CHRONIC DISEASE: Presence of chronic illness
  - FATIGUE, COUGHING, WHEEZING, SHORTNESS OF BREATH: Common respiratory symptoms

- ANXIETY, ALLERGY: Additional health-related indicators
- YELLOW\_FINGERS, SWALLOWING DIFFICULTY, CHEST PAIN: Clinical signs often associated with respiratory or lung issues
- Target Variable:
  - LUNG\_CANCER: Indicates whether the patient was diagnosed with lung cancer (Yes or No)

## **Story of the Data**

The data used in this project represents a simulated collection of medical records aimed at understanding factors associated with lung cancer diagnoses. Each record captures a patient's demographic profile, lifestyle habits, clinical symptoms, and diagnosis status. The dataset serves as a simplified yet insightful snapshot of real-world scenarios where multiple factors interplay in the development of lung cancer.

The story embedded within the data revolves around identifying which characteristics are more frequently observed among patients diagnosed with lung cancer versus those who are not. It highlights how common behaviors such as smoking, alcohol consumption, and exposure to peer pressure might increase risk. At the same time, it draws attention to physiological signs such as fatigue, chronic disease, and respiratory symptoms that could indicate early or advanced stages of illness.

This dataset also enables us to observe gender-based and age-related trends, offering a holistic view of how lung cancer risk may vary across different segments of the population. For example, it helps us examine whether smoking is more prevalent in males than females, or whether fatigue and anxiety are more common in patients with positive diagnoses.

By analyzing this data, we gain an opportunity to translate raw numbers into meaningful narratives—narratives that can inform public health strategies, guide clinical practices,

and foster early interventions. The story of this dataset is not just about identifying who has lung cancer, but understanding why and how certain patterns emerge, helping us bridge the gap between data science and healthcare decision-making.



## **Data Splitting and Preprocessing**

After importing the dataset into Power BI, the first step was to open the data in Power Query Editor for cleaning and transformation. The dataset initially included binary-encoded values (1 and 2) for categorical fields such as symptoms and risk factors.

- **Gender Normalization:**  
The GENDER column was originally labeled as "M" and "F". These were replaced with the full labels "Male" and "Female" for clarity.
- **Binary Value Replacement:**  
All binary columns representing Yes/No responses (e.g., SMOKING, WHEEZING, COUGHING, ALCOHOL CONSUMING) were cleaned by replacing:
  - 1 with "Yes"
  - 2 with "No"
- **Age Grouping:**  
A new column named Age Group was created to categorize patients into four age brackets:
  - Under 20
  - 20–39

- 40–59
- 60+
- Country Column Addition:  
A Country column was added with the value set as "Nigeria" for all records, enabling the use of geographic visuals such as maps.
- Patient ID Generation:  
An index column was added and formatted as a custom ID (e.g., PAT001, PAT002...) to uniquely identify each patient in the dataset.
- Null & Data Type Handling:  
Columns were checked for null values, and appropriate data types were assigned (e.g., Text, Whole Number) to ensure consistency and accuracy during analysis.

### **DAX Measures Created**

- Total Patients  
Calculates the total number of patients in the dataset.  
 *Result: 309* (Displayed using a card visual)
- Total Lung Cancer Cases  
Calculates the number of patients diagnosed with lung cancer.  
 *Result: 270 cases*
- % of Lung Cancer Cases  
Computes the percentage of patients diagnosed with lung cancer relative to the total number of patients.

- **Total Smokers**  
Sums the number of patients identified as smokers.
- **Average Age**  
Calculates the average age of all patients in the dataset.

## **Objectives**

The primary objective of this project is to leverage data analytics and visualization techniques to uncover key patterns and risk factors associated with lung cancer diagnosis. Using Power BI, the project aims to:

1. Identify common symptoms and behavioral patterns among patients diagnosed with lung cancer.
2. Compare lung cancer prevalence across demographic segments such as gender and age groups.
3. Analyze the impact of lifestyle factors (e.g., smoking, alcohol consumption, peer pressure) on lung cancer cases.
4. Visualize the distribution of lung cancer patients geographically, with a focus on Nigeria.
5. Develop interactive and insightful dashboards using DAX measures and Power BI visuals to support data-driven decision-making and public health awareness.

## **Limitations within the Dataset**

While the dataset provides valuable insights into lung cancer risk factors and patient characteristics, there are a few limitations that may impact the depth and scope of analysis:

## 1. Binary Encoding with Limited Detail

- Many important features (e.g., COUGHING, WHEEZING, SMOKING) are recorded as binary Yes/No values, which do not reflect severity, frequency, or duration of the symptom or behavior.

## 2. No Time-Based Information

- The dataset lacks time variables such as diagnosis date, symptom onset, or duration, which restricts trend or time-series analysis.

## 3. No Clinical or Diagnostic Details

- Key medical data such as test results, tumor stages, treatment types, or outcomes are not included, which limits clinical insight.

## 4. Imbalanced Target Variable

- The dataset is heavily skewed toward lung cancer-positive cases, which may affect the generalizability of predictive analysis.

## 5. Small Sample Size

- With only 309 records, the sample size is relatively small for population-wide inferences or model training.

## Pre Analysis

Before diving into detailed analysis and building visualizations, an initial exploration of the dataset was conducted to understand the structure, distribution, and overall quality of the data. This helped shape the direction of the analysis and informed the design of key visuals and DAX measures.

## Observations

### 1. Total Records

- The dataset contains 309 patient records, each representing an individual with various attributes related to lung cancer.

### 2. Demographic Breakdown

- Gender: The dataset includes both male and female patients, with male patients forming the majority.
- Age: Patients range widely in age, with most falling between 40 and 60 years.

### 3. Lung Cancer Diagnosis

- Out of 309 records, 270 patients were diagnosed with lung cancer, indicating a high prevalence within the dataset.

### 4. Lifestyle Risk Factors

- A significant portion of patients reported smoking, alcohol use, and peer pressure, all of which are known contributors to lung cancer.



## 5. Common Symptoms

- Symptoms such as coughing, wheezing, fatigue, and chest pain were frequently observed, especially among those diagnosed with lung cancer.

## 6. Data Quality

- The dataset had no missing values.
- All columns were successfully cleaned and transformed into readable formats (e.g., Yes/No, Male/Female, Age Groups).

## Post Analysis

After building visuals and applying DAX measures in Power BI, deeper insights were uncovered regarding lung cancer risk factors, symptoms, and demographic trends. The post-analysis stage focused on interpreting these patterns and drawing conclusions based on the visualizations and metrics.

### Key Findings:

#### 1. High Lung Cancer Prevalence

- Out of 309 patients, 270 (≈87%) were diagnosed with lung cancer, highlighting a significantly high occurrence within the dataset.

#### 2. Impact of Smoking

- A large proportion of patients diagnosed with lung cancer were active smokers, confirming smoking as one of the most dominant contributing factors.

### 3. Gender Disparity

- Male patients had a higher representation in both the total patient population and in smoking-related lung cancer cases.
- Males also recorded higher average age and total age distribution.

### 4. Symptom Frequency

- Coughing, wheezing, and fatigue were among the most common symptoms reported by patients with lung cancer.
- A total of 101 patients reported fatigue, many of whom were also cancer-positive.

### 5. Chronic Disease Association

- 142 lung cancer patients also had an existing chronic disease, suggesting a potential link between chronic conditions and increased cancer risk.

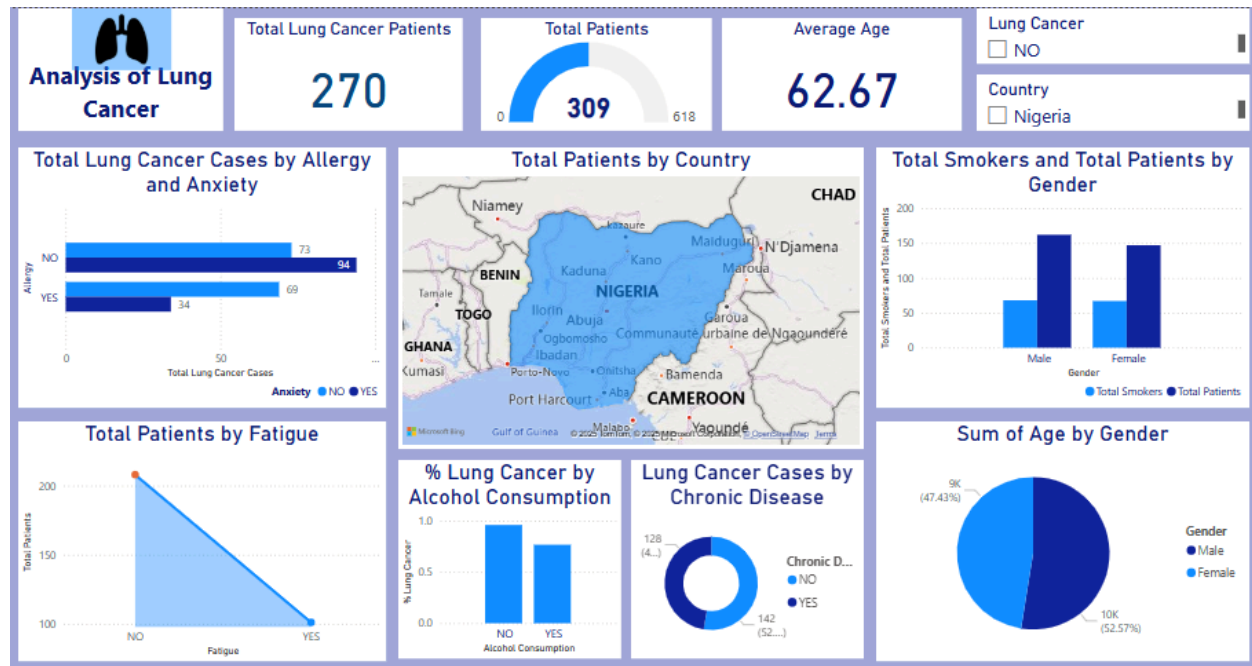
### 6. Psychological and Environmental Influences

- Factors such as anxiety, peer pressure, and alcohol consumption were present in several diagnosed patients, although with lower frequency compared to physical symptoms.

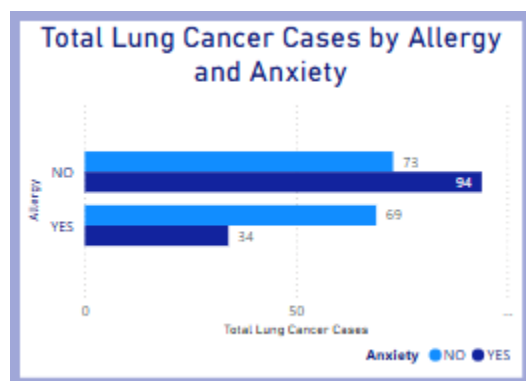
### 7. Age Distribution

- Patients aged 40–59 formed the largest group affected by lung cancer, reinforcing the relevance of age as a non-modifiable risk factor.

## Data Visualization and Charts



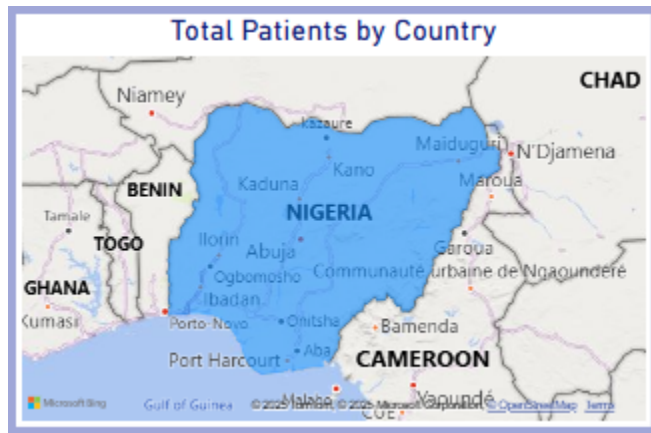
A range of visuals were created in Power BI to effectively communicate insights from the lung cancer dataset. These visuals allowed for interactive exploration of patterns across symptoms, demographics, and risk factors.



### Clustered Bar Chart – Allergy & Anxiety in Lung Cancer Patients

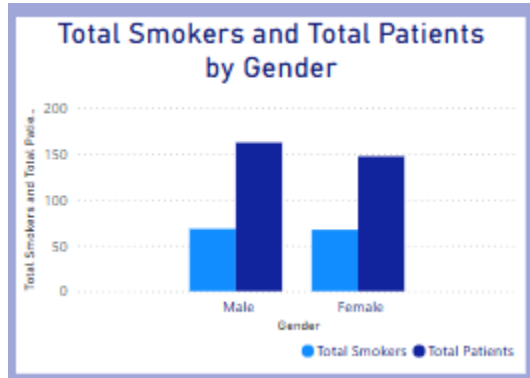
- This visual compares the number of lung cancer patients with allergy and anxiety symptoms.

- Key Insight: 69 patients with lung cancer reported allergies, while fewer experienced anxiety.



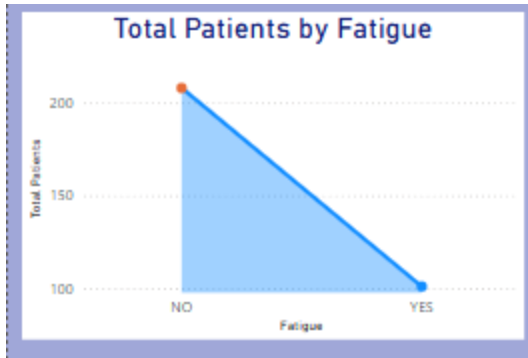
Filled Map – Patient Distribution by Country

- Since the dataset was focused on Nigeria, a filled map was used to highlight Nigeria as the geographic region of analysis.



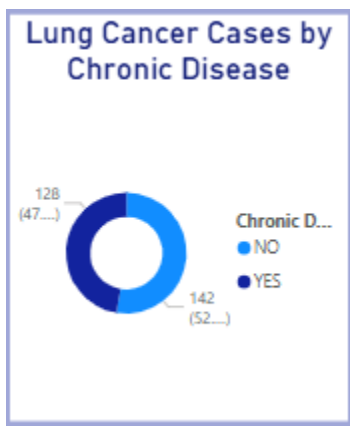
Clustered Column Chart – Smokers by Gender

- Used to compare total smokers and total patients by gender.
- Key Insight: Male smokers were significantly more than female smokers, and they also had a higher count among lung cancer cases.



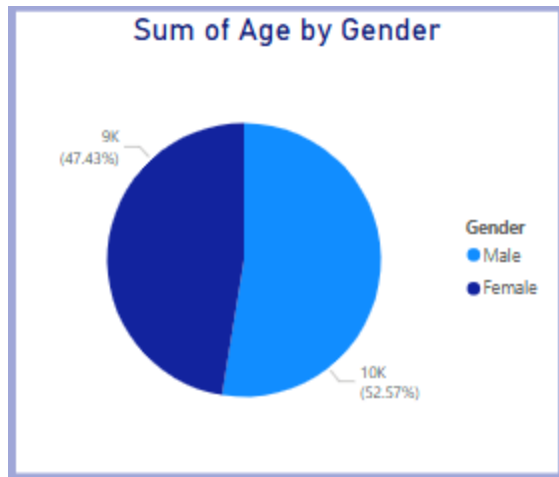
Area Chart – Fatigue Among Patients

- Visualized how many patients experienced fatigue, a common lung cancer symptom.
- Key Insight: 101 out of 309 patients reported fatigue.



Doughnut Chart – Lung Cancer by Chronic Disease

- Showed the number of lung cancer cases in patients with chronic diseases.
- Key Insight: 142 patients with lung cancer also had chronic diseases.



Pie Chart – Sum of Age by Gender

- Displayed the total age distribution by gender.
- Key Insight: The male gender accounted for a higher sum of age, reflecting a larger or older male population.

#### Card Visuals – Key Performance Indicators (KPIs)

- Highlighted key figures at a glance:
  - Total Patients: 309
  - Total Lung Cancer Cases: 270
  - Total Smokers: [DAX-driven count]
  - Average Age: [Calculated from dataset]

#### Slicers

- Implemented across fields such as Gender, Smoking, Age Group, and Lung Cancer to allow for interactive filtering.

## Observations

Following data cleaning, transformation, and visualization, several key observations emerged from the analysis of the lung cancer dataset:

### 1. High Prevalence of Lung Cancer

- Out of 309 patients, 270 were diagnosed with lung cancer, indicating a very high proportion ( $\approx 87\%$ ) of lung cancer-positive cases in the dataset.

### 2. Smoking as a Major Risk Factor

- A large number of lung cancer patients were active smokers, reinforcing the well-established link between smoking and lung cancer development.

### 3. Gender Disparity

- Male patients not only made up the majority of total cases but also had higher smoking rates and higher age distribution compared to females.

### 4. Symptoms Linked to Lung Cancer

- Symptoms such as coughing, wheezing, chest pain, and especially fatigue were commonly reported among patients with lung cancer.

### 5. Chronic Disease Co-Occurrence

- 142 lung cancer patients also suffered from a chronic disease, suggesting a potential compounding effect of pre-existing health conditions.

### 6. Age Group Patterns

- The majority of patients affected by lung cancer fell within the 40–59 and 60+ age groups, confirming age as a significant risk factor.

#### 7. Low Reporting of Mental Health Factors

- Fewer patients reported anxiety or peer pressure, which could indicate underreporting or lower awareness of the psychological dimensions of lung health.

#### 8. Geographic Limitation

- All data points were focused on Nigeria, so geographic insights are localized and cannot be generalized to other countries.

### **Recommendations**

Based on the patterns, relationships, and risk factors revealed in the analysis, the following recommendations are proposed to support lung cancer prevention, early detection, and public health planning:

#### 1. Implement Targeted Anti-Smoking Campaigns

- Smoking is the most prominent behavioral risk factor observed. Stronger awareness campaigns, smoking cessation programs, and stricter tobacco regulations should be targeted, especially toward male adults.

#### 2. Encourage Early Screening for High-Risk Age Groups

- Individuals aged 40 and above should be prioritized for regular lung health screening, particularly if they have a history of smoking or display early



symptoms such as coughing and wheezing.

### 3. Strengthen Chronic Disease Management

- Since many lung cancer patients also suffer from chronic diseases, integrated care approaches should be implemented to monitor and support these individuals more closely.

### 4. Promote Public Education on Early Symptoms

- Educating the public about early warning signs like fatigue, shortness of breath, and chest pain can drive early medical attention and potentially improve outcomes through early diagnosis.

### 5. Integrate Mental Health Awareness into Preventive Care

- Although less frequently reported, anxiety and peer pressure were observed. Healthcare systems should include mental health assessments and support in routine care, especially among smokers and younger populations.

### 6. Develop Gender-Specific Health Interventions

- Since men are more likely to smoke and develop lung cancer, gender-focused education and intervention programs can help reduce risk levels in this demographic.

### 7. Expand and Diversify Data Collection

- Future data collection should include patients from multiple countries and regions, with more detailed clinical, environmental, and temporal variables for more robust and generalizable analysis.

## Conclusion

This project successfully utilized Power BI to analyze a lung cancer dataset containing demographic details, lifestyle habits, symptoms, and medical history of patients.

Through data cleaning, transformation, and visualization, key insights were uncovered about the major risk factors and symptom patterns associated with lung cancer.

The analysis revealed that **smoking**, **chronic disease**, and age above 40 were strongly associated with lung cancer diagnosis. Male patients were more likely to be smokers and represented a larger share of lung cancer cases. Symptoms such as **fatigue**, **wheezing**, and **coughing** were frequently present among affected individuals.

The dataset provided valuable insights that can inform public health strategies. By applying DAX measures, calculated columns, and interactive visualizations, the dashboard enabled meaningful interpretation and exploration of lung cancer trends.

Overall, this project demonstrates the power of data analytics and visualization in healthcare—highlighting how even a basic dataset can be transformed into actionable intelligence to support early detection, prevention, and awareness efforts.

