Miss Misso Production Internship Report

Data Analytics

-Khushi Bhati

4th Year

IIT ROORKEE

GitHub Link- https://github.com/khushi796/MMP_Intern

Objective –

This internship aims to understand the dataset, perform Data Cleaning, Exploratory Data Analysis (EDA), and generate insights from a lung cancer survey dataset. The goal is to understand patterns and potential risk factors related to lung cancer and support decision-making using visualizations and statistics. Various visualizations were used to explore relationships, and the data was grouped into age intervals for clearer insights. All steps were completed and documented in a Google Collab using the pandas, seaborn, and matplotlib libraries.

Tools & Libraries Used -

- **Python (Google Collab)**
- **Libraries:**

```
[1]  # Libraries
     import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns
```

  - pandas – data handling
  - NumPy – numerical operations
  - seaborn, matplotlib – data visualization

Dataset Overview -

- **Dataset -**

```
# Uploading Dataset
df = pd.read_csv('survey lung cancer.csv')
```

- **Dataset Preview-**

The above table shows the first five rows of the lung cancer survey dataset. Each row represents a participant and their responses to various health-related and lifestyle questions. The dataset contains both categorical and numerical features. Some key observations:

- **GENDER** and **AGE** columns provide demographic information.
- Features like **SMOKING**, **FATIGUE**, **WHEEZING**, and **COUGHING** represent symptom and behavior-related responses, with values coded as 1 or 2, where:
  Likely 1 = No and 2 = Yes
- The **LUNG_CANCER** column is the target variable, with values "YES" or "NO" indicating diagnosis status.

This preview helps confirm the structure and content of the dataset before proceeding with cleaning and analysis.

```
# Quick check how data looks like
df.head()
```

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
0	M	69	1	2	2	1	1	2	1	2	2	2	2	2	2	YES
1	M	74	2	1	1	1	2	2	2	1	1	1	2	2	2	YES
2	F	59	1	1	1	2	1	2	1	2	1	2	2	1	2	NO
3	M	63	2	2	2	1	1	1	1	1	2	1	1	2	2	NO
4	F	63	1	2	1	1	1	1	1	2	1	2	2	1	1	NO

Data Cleaning-

- **Null Value Check-**

The output confirms that **there are no missing (null) values** in any of the columns of the dataset. Every feature—whether demographic (like GENDER, AGE) or symptom-related (like COUGHING, WHEEZING, CHEST PAIN)—has complete data for all 309 entries.
This means:

- No imputation or handling of missing data is required.
- The dataset is already clean in terms of completeness and ready for further analysis.

This step ensures the integrity of the data before moving on to exploratory data analysis (EDA).

```
# Check null values
df.isnull().sum()
```

	0
GENDER	0
AGE	0
SMOKING	0
YELLOW_FINGERS	0
ANXIETY	0
PEER_PRESSURE	0
CHRONIC DISEASE	0
FATIGUE	0
ALLERGY	0
WHEEZING	0
ALCOHOL CONSUMING	0
COUGHING	0
SHORTNESS OF BREATH	0
SWALLOWING DIFFICULTY	0
CHEST PAIN	0
LUNG_CANCER	0

dtype: int64

- **Dataset Dimensions -**

The output of df.shape shows that the dataset contains **309 rows** and **16 columns**. This means there are 309 individual survey responses, and each response includes 16 attributes (features), such as demographic information, symptoms, and the lung cancer diagnosis status.

This provides a moderately sized dataset that is suitable for exploratory data analysis without requiring heavy computational resources.

```
# No. of rows and columns in Dataset
df.shape
```

```
(309, 16)
```

- **Dataset Structure & Data Types -**

The df.info() output gives a comprehensive overview of the dataset's structure. It confirms that the dataset contains **309 entries (rows)** and **16 columns**, with **no missing values** in any column.

- **14 columns** are of data type int64, which includes survey features such as AGE, SMOKING, FATIGUE, WHEEZING, and others. These are likely encoded as 1/2 to represent Yes/No type answers.
- **2 columns** are of type object:
  - GENDER: Contains string values like 'M' and 'F'
  - LUNG_CANCER: Target column with values 'YES' and 'NO'

This summary helps identify which columns may need encoding (e.g., GENDER, LUNG_CANCER) and confirms that the dataset is clean in terms of structure and completeness.

```
# Check Datatype and null data in columns
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 309 entries, 0 to 308
Data columns (total 16 columns):
 #   Column                 Non-Null Count  Dtype 
---  ------                 --------------  ----- 
 0   GENDER                 309 non-null    object
 1   AGE                    309 non-null    int64 
 2   SMOKING                309 non-null    int64 
 3   YELLOW_FINGERS         309 non-null    int64 
 4   ANXIETY                309 non-null    int64 
 5   PEER_PRESSURE          309 non-null    int64 
 6   CHRONIC DISEASE        309 non-null    int64 
 7   FATIGUE                309 non-null    int64 
 8   ALLERGY                309 non-null    int64 
 9   WHEEZING               309 non-null    int64 
 10  ALCOHOL CONSUMING      309 non-null    int64 
 11  COUGHING               309 non-null    int64 
 12  SHORTNESS OF BREATH    309 non-null    int64 
 13  SWALLOWING DIFFICULTY  309 non-null    int64 
 14  CHEST PAIN             309 non-null    int64 
 15  LUNG_CANCER            309 non-null    object
dtypes: int64(14), object(2)
memory usage: 38.8+ KB
```

- Duplicate Records Check -

The code output shows that the dataset initially contained **33 duplicate rows**. These rows had identical values across all columns and could introduce bias or redundancy in the analysis.
To ensure data quality, the duplicates were removed using:
df.drop_duplicates(inplace=True)
After this operation, the dataset was cleaned and reduced to 276 unique records, ensuring that all further analysis is based on distinct survey responses.

```
# Check for duplicates
print("Duplicate rows:", df.duplicated().sum())

# Drop duplicates if any
df.drop_duplicates(inplace=True)
```

```
Duplicate rows: 33
```

- **Unique Value Inspection -**

This step helps us understand the distinct values present in each column. The results show:

- **GENDER** has two categories: 'M' and 'F', representing Male and Female.
- **AGE** contains a range of values between 21 and 87, indicating the ages of the survey participants.
- Most other columns (like SMOKING, ANXIETY, WHEEZING, CHEST_PAIN) have values 1 and 2, which likely represent **binary responses**:
  - 1 = No
  - 2 = Yes

  (Note: This encoding will be flipped to 0 and 1 later for analysis and modeling.)
- The target column LUNG_CANCER contains two values: 'YES' and 'NO'.

  This step helped identify categorical columns that required conversion for plotting and model training, ensuring proper preprocessing in later stages.

```
# Preview unique values in each column
for col in df.columns:
    print(f"{col}: {df[col].unique()}")
```

```
GENDER: ['M' 'F']
AGE: [69 74 59 63 75 52 51 68 53 61 72 60 58 48 57 44 64 21 65 55 62 56 67 77
 70 54 49 73 47 71 66 76 78 81 79 38 39 87 46]
SMOKING: [1 2]
YELLOW_FINGERS: [2 1]
ANXIETY: [2 1]
PEER_PRESSURE: [1 2]
CHRONIC DISEASE: [1 2]
FATIGUE : [2 1]
ALLERGY : [1 2]
WHEEZING: [2 1]
ALCOHOL CONSUMING: [2 1]
COUGHING: [2 1]
SHORTNESS OF BREATH: [2 1]
SWALLOWING DIFFICULTY: [2 1]
CHEST PAIN: [2 1]
LUNG_CANCER: ['YES' 'NO']
```

- **Final Dataset Shape After Cleaning -**

After removing the 33 duplicate rows, the dataset now contains **276 unique records** and **16 columns**. This ensures that all the entries used for analysis are distinct, which improves the reliability of the insights and avoids duplication bias in results or modeling.
This is a crucial preprocessing step before performing any kind of exploratory data analysis (EDA) or building predictive models.

```
# No of Rows and Columns after removing Duplicate values
df.shape
```

```
(276, 16)
```

- **Column Names Overview -**

Using df.columns, we listed all the column names present in the dataset. There are **16 columns**, each representing a feature collected from survey participants. These include:

- **Demographic feature:** GENDER, AGE
- **Lifestyle habits:** SMOKING, ALCOHOL CONSUMING, PEER_PRESSURE
- **Medical conditions & symptoms:** FATIGUE, COUGHING, CHEST PAIN, SHORTNESS OF BREATH, WHEEZING, etc.
- **Target variable:** LUNG_CANCER indicates whether the individual has been diagnosed with lung cancer.

This step helped confirm the structure and naming of all fields before starting detailed analysis.

```
# Column Name
df.columns
```

```
Index(['GENDER', 'AGE', 'SMOKING', 'YELLOW_FINGERS', 'ANXIETY',
       'PEER_PRESSURE', 'CHRONIC DISEASE', 'FATIGUE ', 'ALLERGY ', 'WHEEZING',
       'ALCOHOL CONSUMING', 'COUGHING', 'SHORTNESS OF BREATH',
       'SWALLOWING DIFFICULTY', 'CHEST PAIN', 'LUNG_CANCER'],
      dtype='object')
```

Exploratory Data Analysis (EDA) –

Exploratory Data Analysis (EDA) was performed to understand the underlying structure, patterns, and relationships within the dataset. Various visualizations, such as bar plots and heatmaps, were used to explore the distribution of key features like gender, age, smoking status, and lung cancer diagnosis.

Through EDA, we identified the following trends:

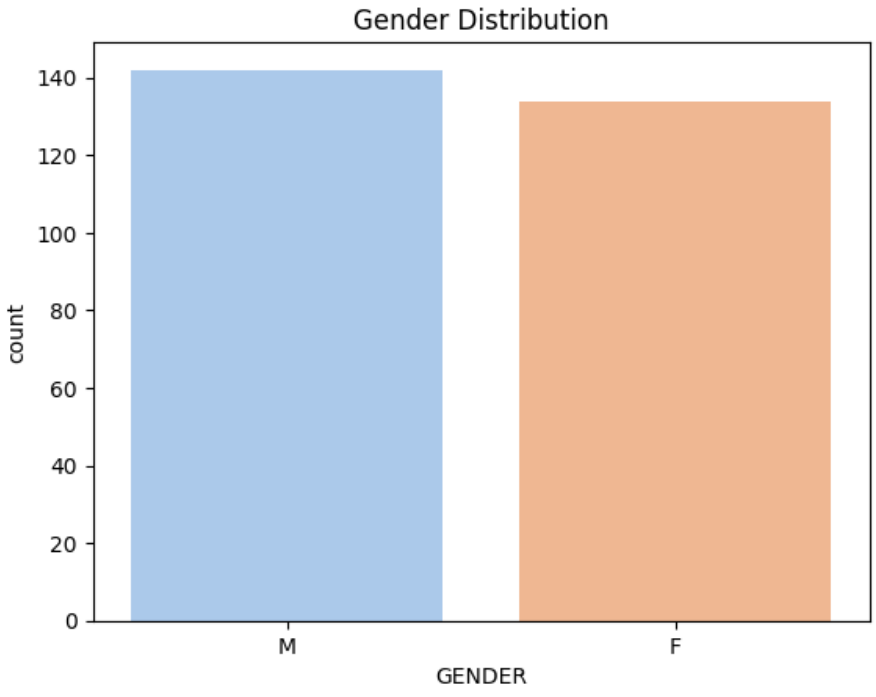
- **Gender Distribution -**

The bar chart above illustrates the distribution of gender in the dataset. The GENDER column consists of two categories: Male (M) and Female (F).

- Male (M): Approximately 142 individuals
- Female (F): Approximately 133 individuals

The gender distribution is relatively balanced, with a slightly higher number of male participants compared to female participants. This indicates that the dataset does not suffer from severe gender imbalance, which is beneficial for any analysis or model that may be sensitive to demographic bias

```
# Plot Gender Distribution
sns.countplot(data=df, x='GENDER', palette='pastel')
plt.title('Gender Distribution')
plt.show()
```



- ## Lung Cancer Diagnosis Distribution -

The bar chart displays the distribution of individuals diagnosed with lung cancer in the dataset.

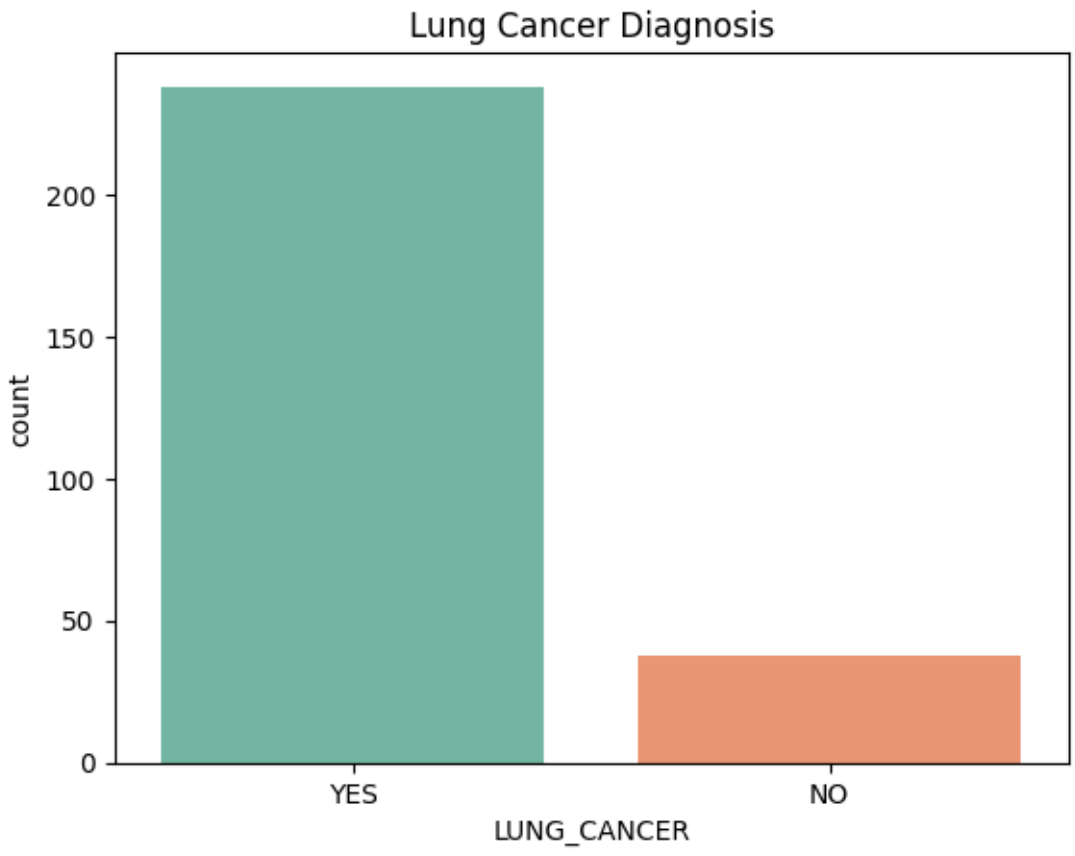
- **YES**: A significantly higher number of individuals have been diagnosed with lung cancer (approximately 235).
- **NO**: A comparatively smaller group (around 37 individuals) have not been diagnosed.

This indicates that a **majority of the dataset comprises individuals with a positive diagnosis of lung cancer**, suggesting either:

- The dataset is biased toward affected individuals (perhaps due to medical record sampling), or
- The target variable is imbalanced.

Such imbalance should be carefully considered during model training or any statistical analysis, as it can lead to biased predictions favoring the majority class.

```
# Plot Lung Cancer Diagnosis
sns.countplot(data=df, x='LUNG_CANCER', palette='Set2')
plt.title('Lung Cancer Diagnosis')
plt.show()
```



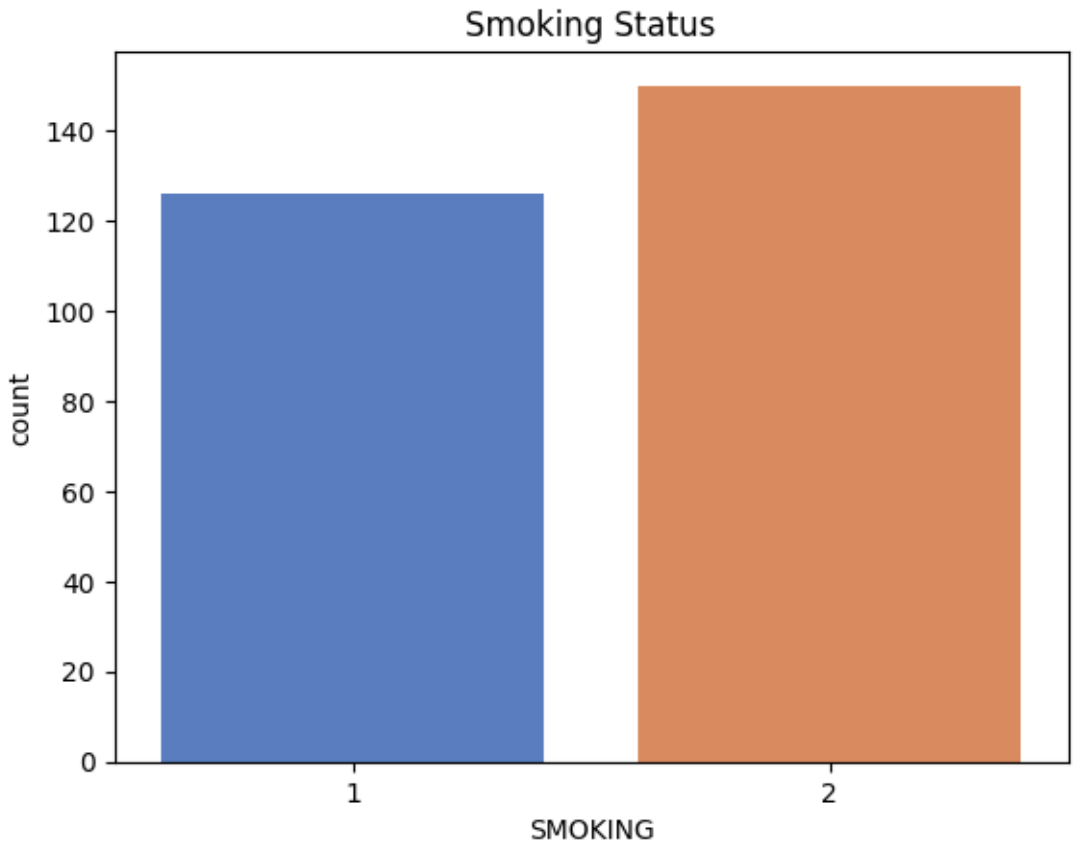
- ## Smoking Status Distribution

The bar chart presents the distribution of individuals based on their smoking habits.

- **Label 1**: Represents non-smokers (approximately 127 individuals)
- **Label 2**: Represents smokers (approximately 150 individuals)

The dataset contains **a slightly higher number of smokers than non-smokers**, indicating a fairly balanced but slightly skewed distribution toward smokers. If smoking is a potential risk factor for lung cancer in the dataset, this variable may play an important role in further analysis or predictive modeling.

```
# Smoking vs Non-smoking
sns.countplot(data=df, x='SMOKING', palette='muted')
plt.title('Smoking Status')
```



- ## Age Group Distribution

The bar chart shows the distribution of individuals across different age groups, divided into 10-year intervals from 0 to 100.

Key observations:

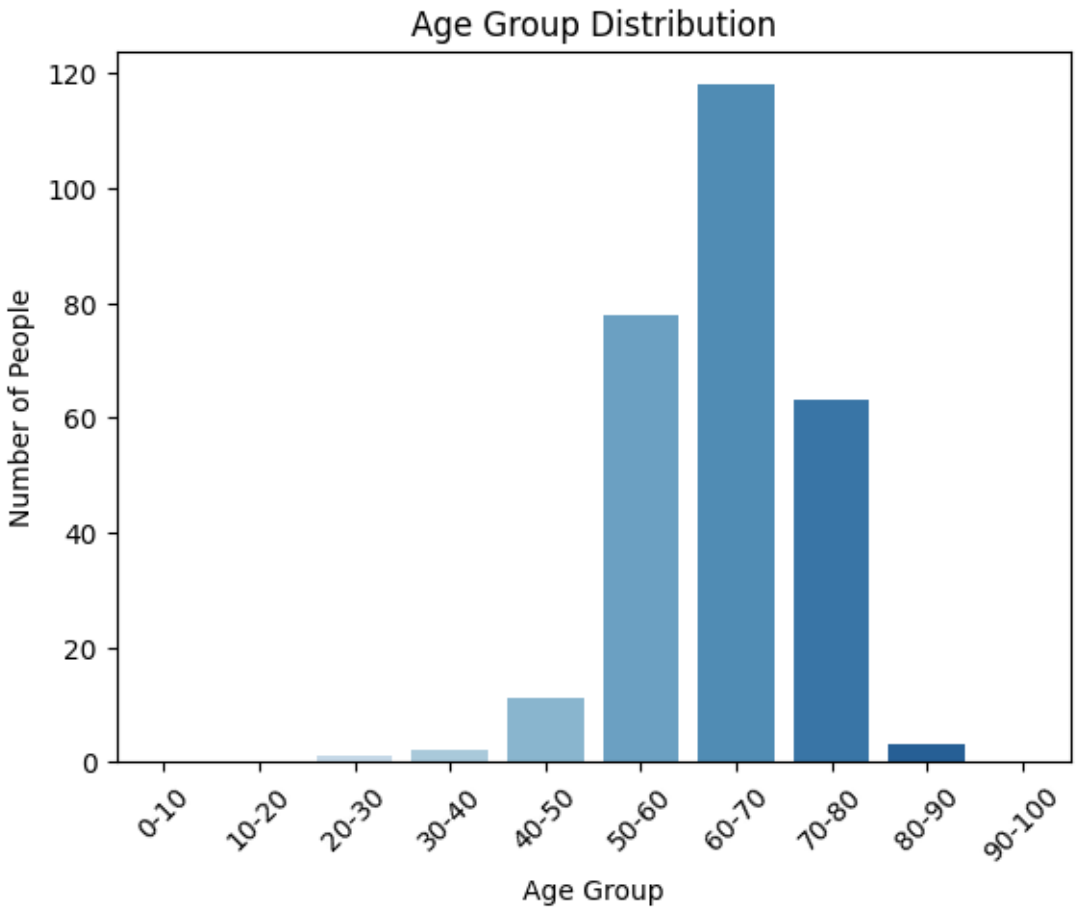
- The **60–70** age group has the highest number of individuals (approximately 119), indicating that the dataset is heavily concentrated around elderly individuals.
- The **50–60** and **70–80** age groups also have significant representation, with around 78 and 63 individuals respectively.
- Age groups **below 40** and **above 80** have very few participants, suggesting that lung cancer diagnoses in this dataset are more prevalent among the **middle-aged and elderly population**.

This distribution is important for understanding the **age-related trends and risks** associated with lung cancer in the dataset.

```
# Define bins (0–100 in steps of 10)
bins = list(range(0, 101, 10))  # [0, 10, 20, ..., 100]
labels = [f'{i}-{i+10}' for i in bins[:-1]]

# Create a new column with age groups
df['AGE_GROUP'] = pd.cut(df['AGE'], bins=bins, labels=labels, right=False)

# Countplot of age groups
sns.countplot(data=df, x='AGE_GROUP', palette='Blues')
plt.title('Age Group Distribution')
plt.xlabel('Age Group')
plt.ylabel('Number of People')
plt.xticks(rotation=45)
plt.show()
```



- **Lung Cancer Diagnosis by Gender -**

The grouped bar chart displays the number of male and female individuals diagnosed with or without lung cancer.

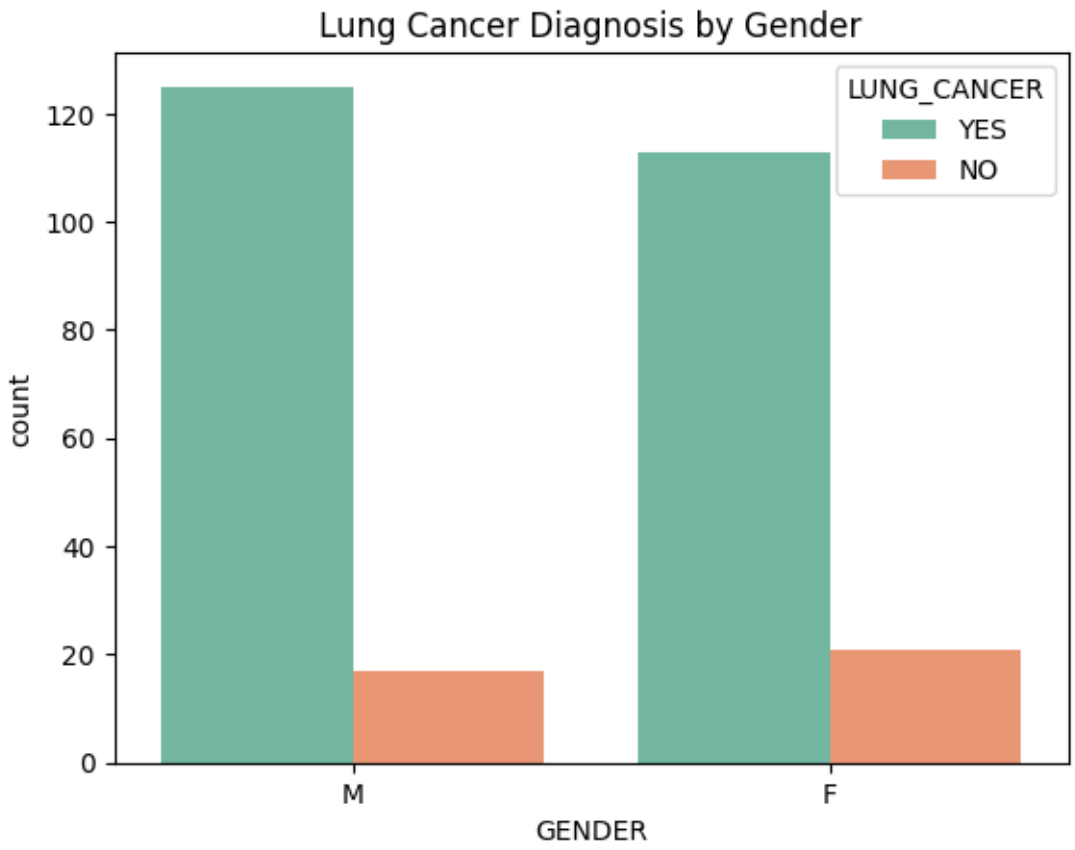
Key observations:

- Both **males and females** show a significantly higher count in the **"YES"** (diagnosed with lung cancer) category compared to the **"NO"** category.
- Among **males (M)**, about **125** have been diagnosed, while around **17** have not.
- Among **females (F)**, approximately **113** have been diagnosed, and **21** have not.

This indicates:

- Lung cancer affects both genders at **similar rates**, with **slightly more male cases**.
- The pattern suggests a relatively **balanced gender distribution** in lung cancer diagnosis in this dataset.

```
sns.countplot(data=df, x='GENDER', hue='LUNG_CANCER', palette='Set2')
plt.title('Lung Cancer Diagnosis by Gender')
plt.show()
```



- **Lung Cancer by Smoking Status -**

The chart above shows the relationship between **smoking status** and **lung cancer diagnosis**:

- **Smoking status 1** = Non-smokers
- **Smoking status 2** = Smokers

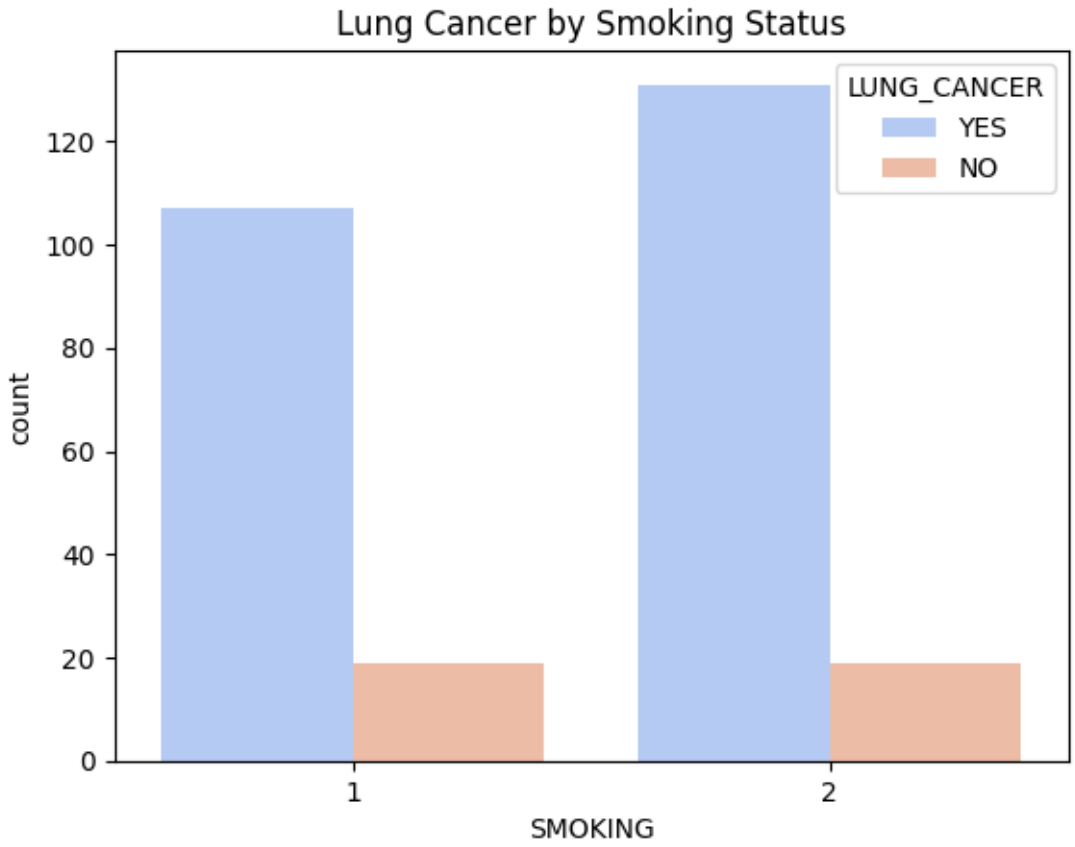
Observations:

- Among **non-smokers**, over **100 individuals** have been diagnosed with lung cancer, and around **20** have not.
- Among **smokers**, about **130 individuals** have lung cancer, with a similar number of non-diagnosed cases (also around 20).

Key Insight:

- Both smokers and non-smokers have a **high number of lung cancer cases**, which suggests that while smoking may be a contributing factor, it is **not the sole determinant** in this dataset.
- The number of lung cancer diagnoses is **slightly higher among smokers**, which could support a weak positive correlation.

```
sns.countplot(data=df, x='SMOKING', hue='LUNG_CANCER', palette='coolwarm')
plt.title('Lung Cancer by Smoking Status')
plt.show()
```



- ## Correlation Heatmap Analysis

The heatmap above visualizes the pairwise **correlation coefficients** between numeric variables in the dataset. Correlation values range from -1 (perfect negative) to +1 (perfect positive), with values close to 0 indicating no linear relationship.

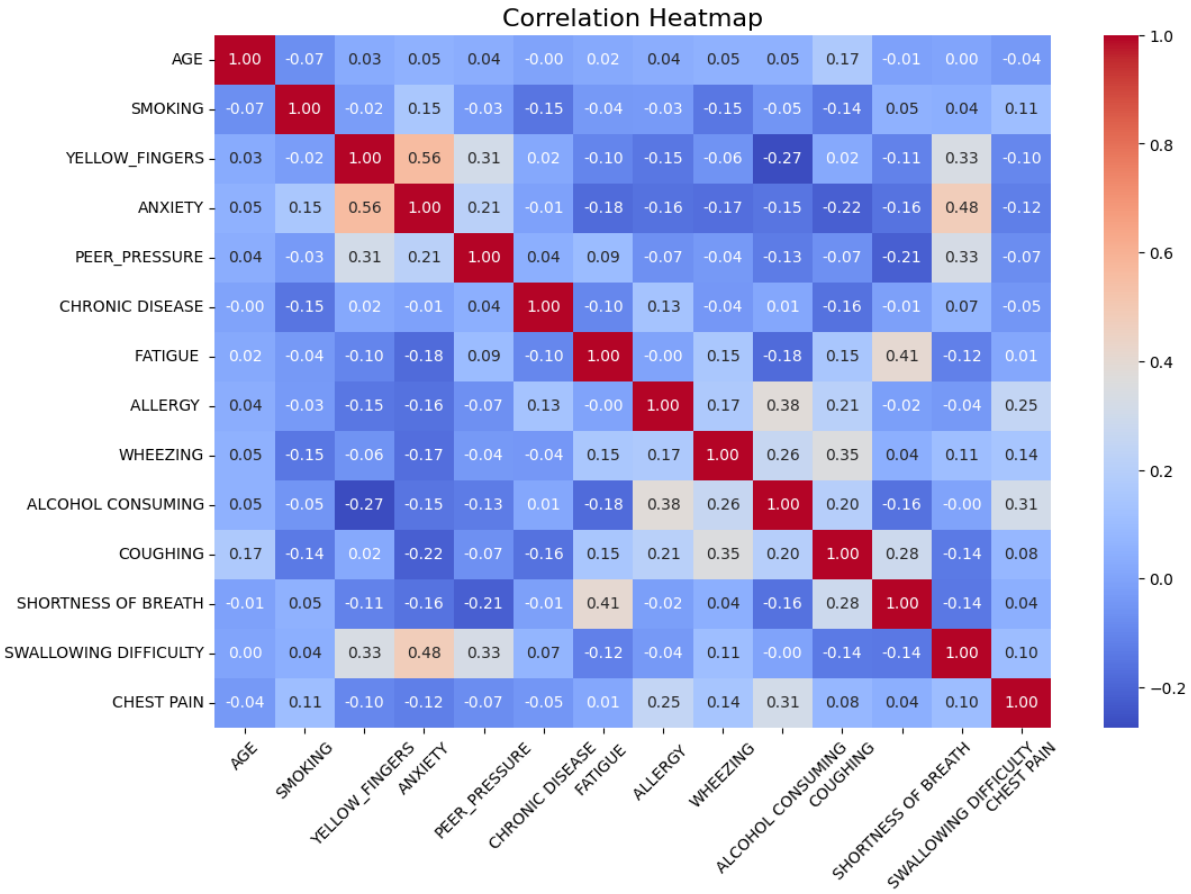
Key Observations:

1. **Strongest Positive Correlations:**

- YELLOW_FINGERS and ANXIETY (**0.56**): Suggests individuals with yellow fingers (often linked to smoking) tend to report higher anxiety levels.
- ANXIETY and SWALLOWING_DIFFICULTY (**0.48**): Could reflect a psychosomatic link between anxiety and swallowing issues.
- YELLOW_FINGERS and PEER_PRESSURE (**0.31**): Peer influence may contribute to smoking-related behaviors.
- WHEEZING shows a moderate positive correlation with COUGHING (**0.35**) and SHORTNESS OF BREATH (**0.26**), which is medically consistent.

2. **Negative/Weak Correlations:**
   - Most variables are **weakly correlated**, with many values between **-0.2 and 0.2**, suggesting limited linear relationships.
   - ALCOHOL_CONSUMING has a slight negative correlation with variables like YELLOW_FINGERS (**-0.27**) and PEER_PRESSURE (**-0.13**), but not strong enough to draw firm conclusions.
3. **AGE** shows **almost no significant correlation** with any variable, suggesting that the symptoms and risk factors in the dataset are not strongly age-dependent in a linear way.

```
# Correlation heatmap
plt.figure(figsize=(12, 8))
corr = df.corr(numeric_only=True)
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap', fontsize=16)
plt.xticks(rotation=45)
plt.yticks(rotation=0)
plt.show()
```



Conclusion -

This internship project provided hands-on experience in data analytics using a real-world health dataset focused on lung cancer. Through systematic data cleaning, transformation, and exploratory data analysis, several meaningful patterns were uncovered, including trends related to age, gender, smoking habits, and symptom correlations.

Key takeaways include:

- A strong presence of lung cancer cases in the dataset, highlighting class imbalance.
- Middle-aged and elderly individuals showed higher diagnosis rates.
- Smoking is slightly associated with lung cancer, but not the only contributing factor.
- Behavioral and symptomatic features like yellow fingers, anxiety, and wheezing showed stronger interrelationships.

Overall, the project enhanced my ability to draw insights from data using Python and laid the groundwork for more in-depth analysis. The structured approach to EDA enabled a clear understanding of the data, supporting evidence-based health analysis.

Thank You