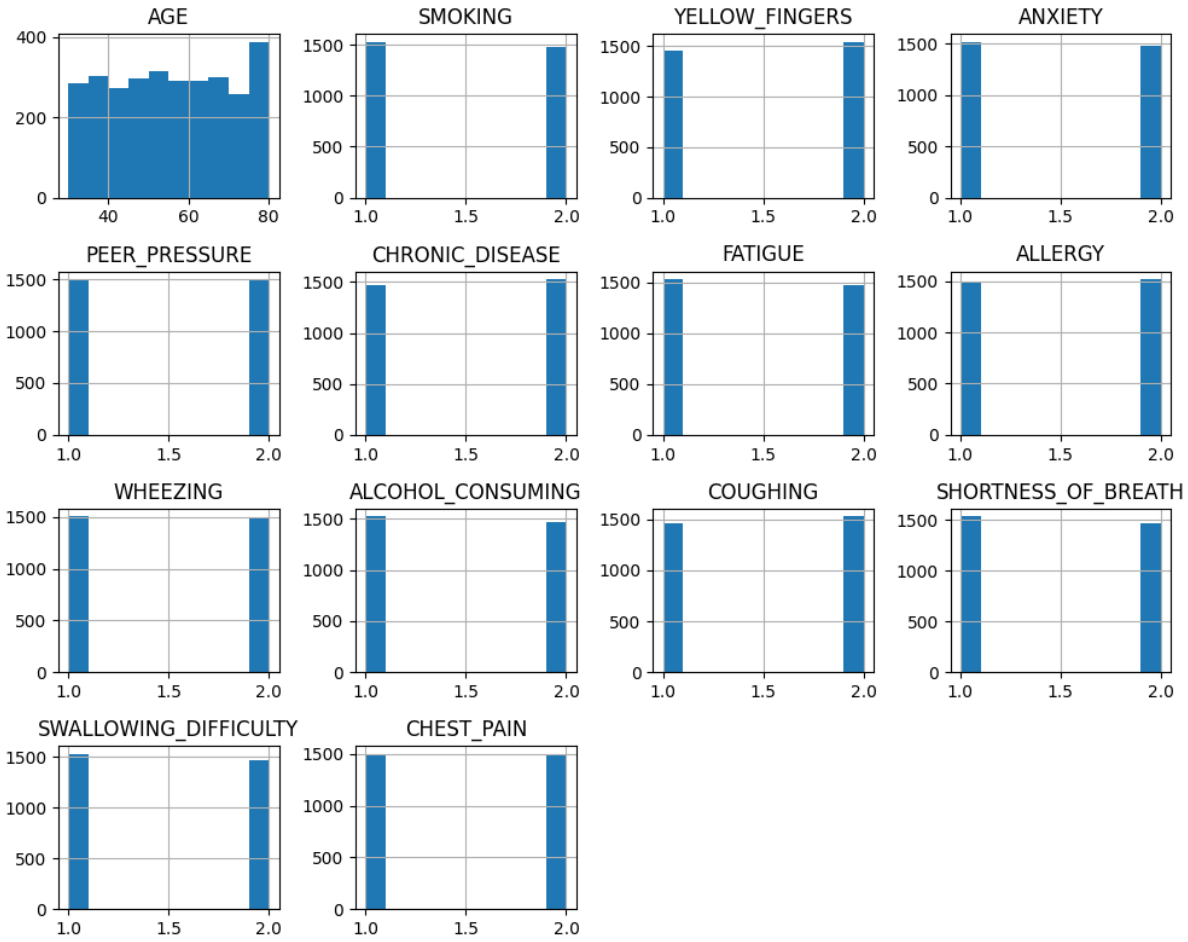
Histogram for all numeric columns



1. AGE

- Continuous variable.
- Most individuals are between **40 to 80 years old**.
- A slight peak is observed around **80**, suggesting many participants are in older age groups.

2. SMOKING

- Binary (1 = No, 2 = Yes).
- Tall bar at **2**: Most individuals are **smokers**.
- Fewer at **1**, indicating non-smokers are less common.

3. YELLOW_FINGERS

- Binary.
- Higher bar at **2**, meaning many have yellow fingers (a smoking-related symptom).

4. ANXIETY

- Binary.
- Higher bar at **2**, indicating **many participants report anxiety**.

5. PEER_PRESSURE

- Binary.
- Fairly balanced, but slightly more individuals experienced peer pressure (value 2).

6. CHRONIC_DISEASE

- Binary.
- Slightly more individuals **do not have chronic diseases** (value 1).

7. FATIGUE

- Binary.
- Higher count for **2**, showing fatigue is a **common symptom** in the dataset.

8. ALLERGY

- Binary.
- Slightly more individuals **do not have allergies** (value 1), but still fairly balanced.

9. WHEEZING

- Binary.
- Almost even distribution between **yes (2)** and **no (1)**.

10. ALCOHOL_CONSUMING

- Binary.
- Slightly more people **consume alcohol** (value 2) than not.

11. COUGHING

- Binary.
- Most individuals report **coughing** (value 2).

12. SHORTNESS_OF_BREATH

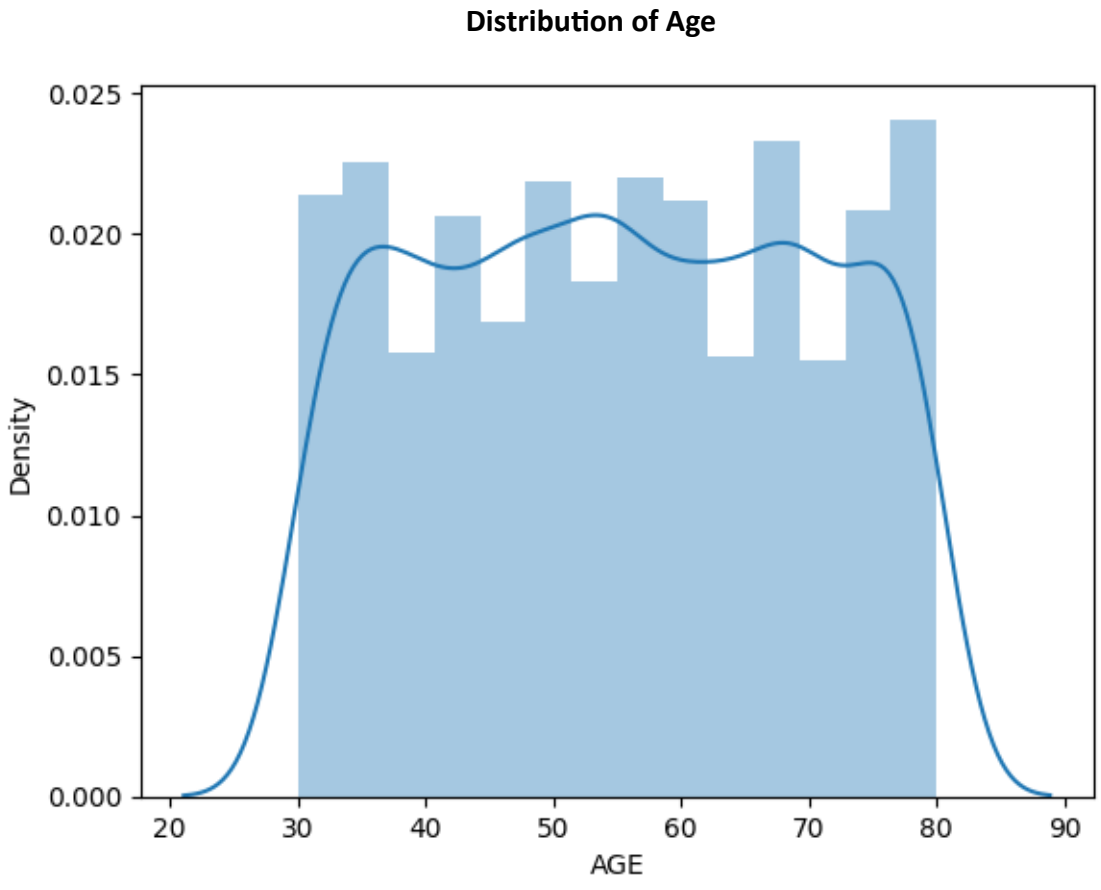
- Binary.
- More individuals experience **shortness of breath** (value 2).

13. SWALLOWING_DIFFICULTY

- Binary.
- Fairly balanced, but slightly more report **difficulty swallowing** (value 2).

14. CHEST_PAIN

- Binary.
- Very balanced distribution; chest pain is present in **about half** of the individuals.



Histogram (bars):

- Each bar represents a range of ages (like 30–40, 40–50, etc.).
- The **height** of the bar shows how many individuals fall into that age group.
- The bars are fairly **even in height**, meaning the **age distribution is uniform**—individuals are spread across all age groups from 30 to 80.

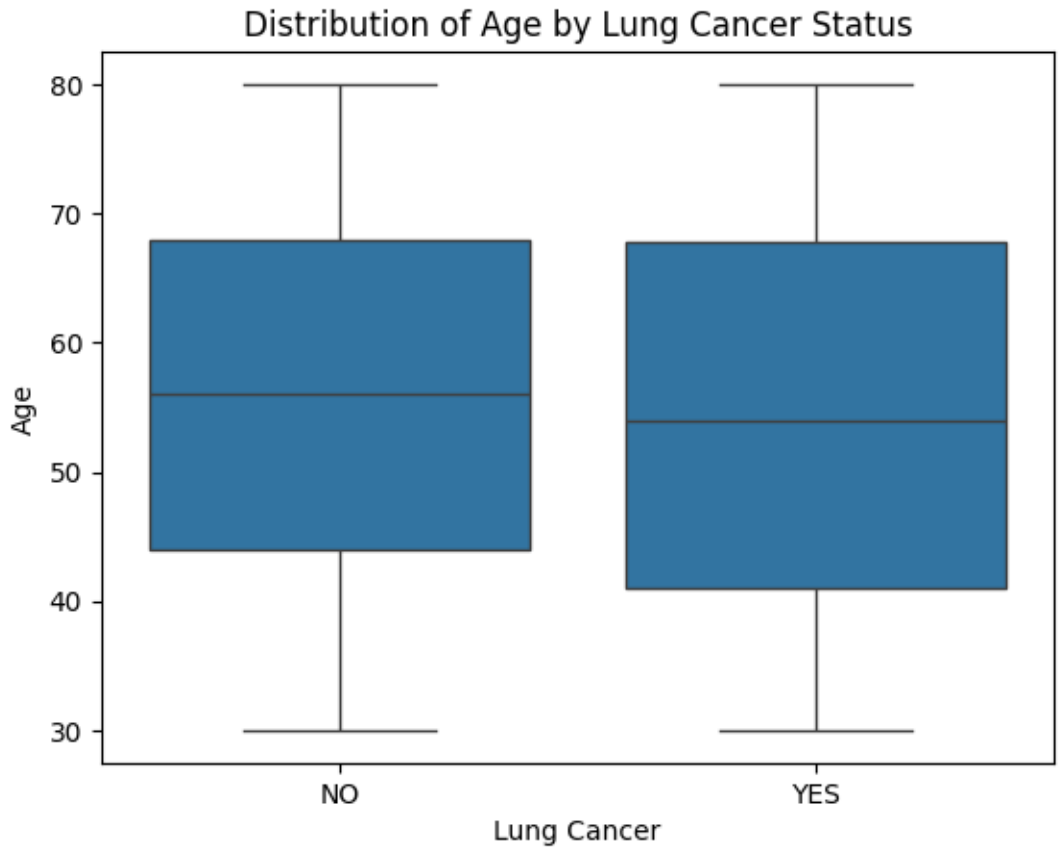
KDE Line (curve):

- The **blue line** represents a smooth estimate of the data distribution.
- It shows that the age values are **evenly distributed**, without sharp peaks.
- The curve is **flatter in the middle** and **tapers off at both ends**, indicating:
    - Few individuals are younger than **30** or older than **80**.
    - Most individuals fall between **30 and 80 years**.

Summary:

- The dataset includes a **balanced number of people across age groups**.
- The age distribution is **not skewed**—no particular age dominates.
- **Most common ages**: Between **30 to 80** years.

Box Plot



box plot showing the **distribution of Age** based on **Lung Cancer status** (Yes or No).

Detailed Explanation:

- The plot compares the **ages of people who have lung cancer (YES)** and those who **do not (NO)**.
- Each box shows the **middle 50%** of the data (from the 25th to 75th percentile).
- The **line inside the box** is the **median** (middle age).
- The **whiskers** extend to show the **range** of the data (excluding outliers).

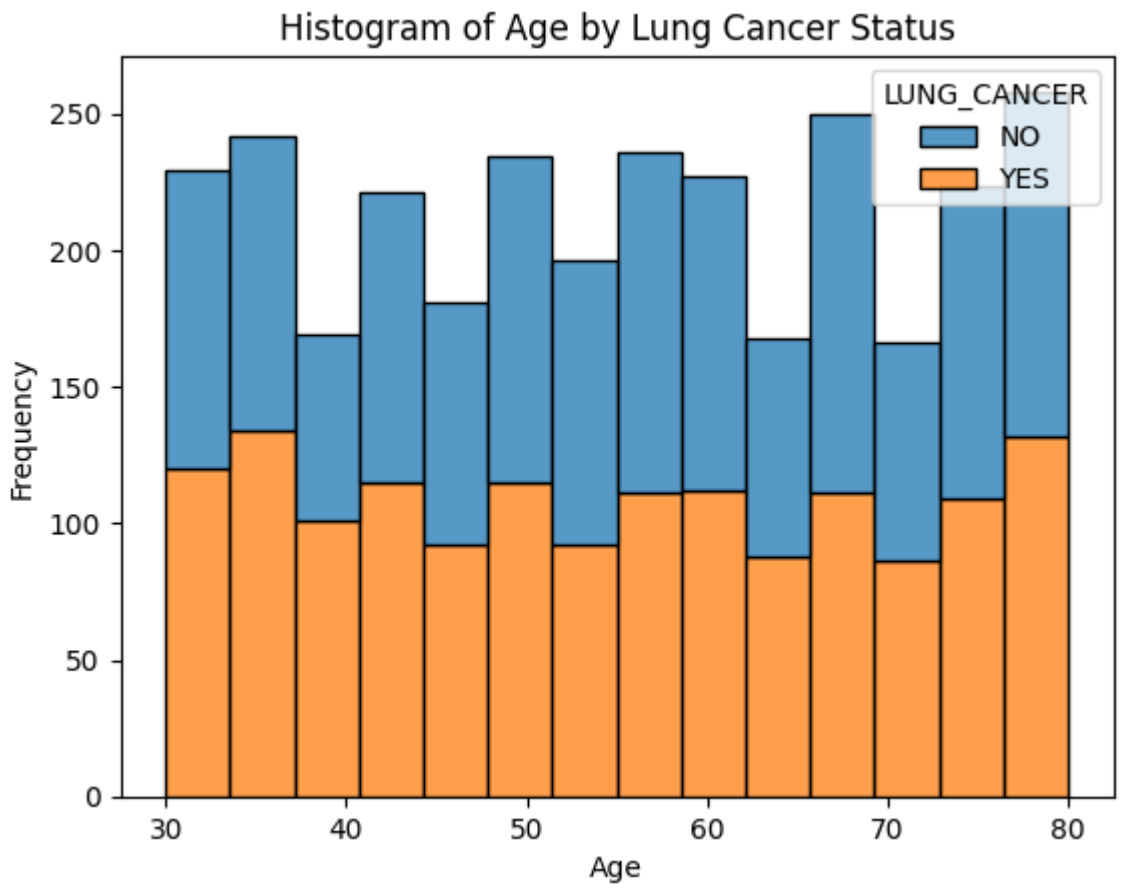
Insights:

- The **age ranges** for both groups are similar (about **30 to 80 years**).
- The **median age** of those **with lung cancer** (YES) is slightly **lower** than those **without** lung cancer (NO).
- Both groups have **similar spread** (variation) in ages.

Summary:

- People with and without lung cancer are spread across similar age ranges.
- Slight difference in **median age**, but not drastically different.
- **Age alone may not be a strong differentiator** for lung cancer in this dataset.

Histogram of lung cancer vs age

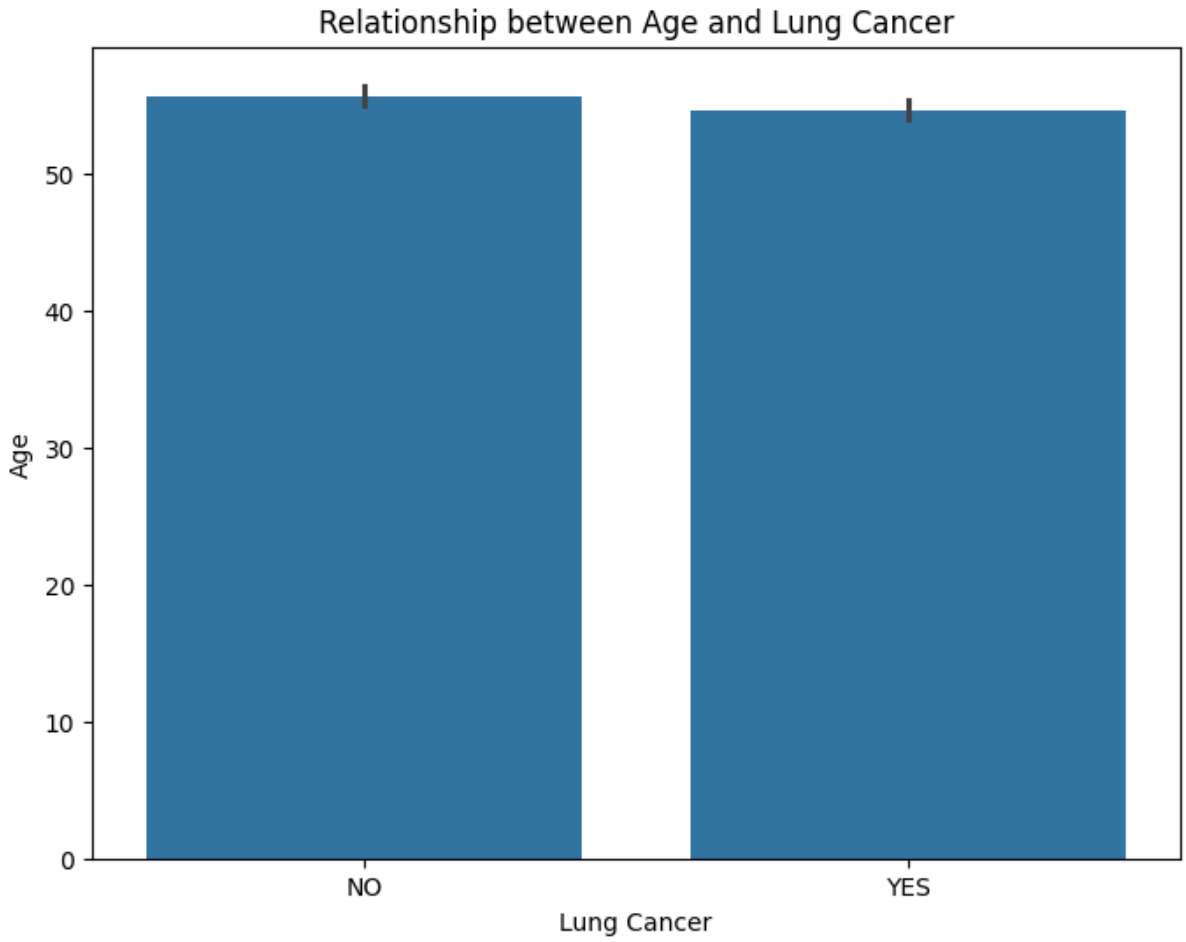


- **X-axis (Age)**: Age is divided into bins (e.g., 30–35, 35–40, etc.).
- **Y-axis (Frequency)**: Number of individuals in each age group.
- **Legend (LUNG_CANCER)**:
    - **Orange bars ("YES")**: Number of people with lung cancer.
    - **Blue bars ("NO")**: Number of people without lung cancer.

Interpretation:

- The chart helps visualize how **lung cancer cases vary with age**.
- Across all age groups, both "YES" and "NO" lung cancer cases are present.
- The total height of the bars shows the total number of people in each age group.
- The **height of the orange portion** indicates how many of those people had lung cancer.

Relationship between age and lung cancer



Type of Plot:

- This is a **bar chart** comparing the **mean age** of two groups:
    - People **without lung cancer** ("NO")
    - People **with lung cancer** ("YES")
- The bars are **vertical**, and their height represents the **average (mean) age** of individuals in each group.

Axes:

- **X-axis (Lung Cancer)**:
    - Two categories: NO and YES, indicating whether a person has lung cancer or not.
- **Y-axis (Age)**:
    - Represents the **average age** of individuals in each lung cancer group.
    - Scale goes from 0 to around 60.

Observations:

1. **Average Age Comparison**:
    - **NO** group (people without lung cancer) has a slightly higher average age (~56 years).
    - **YES** group (people with lung cancer) has a slightly lower average age (~55 years).
    - The difference is **very small**, suggesting that **age alone is not a strong distinguishing factor** between the two groups in this dataset.
2. **Error Bars**:
    - The small **black lines** on top of each bar represent the **confidence interval or standard error**.
    - These indicate the **variability** of the data within each group.
    - Since the error bars are short and overlap, it implies that the difference in average age between groups is **not statistically significant**.

Conclusion / Insight:

- The plot shows that the **mean age** of individuals with lung cancer is **almost the same** as that of individuals without lung cancer.
- There is **no substantial difference** in average age between the two groups.
- Therefore, while age may be a factor in lung cancer risk, **this plot alone suggests that other variables may be more influential** in this particular dataset.