**EDA LUNG CANCER**

**1. Dataset Overview:**

* **Total Records:** 3,000 entries
* **Total Features:** 16 columns, including both categorical and numerical data.

**2. Missing Values and Data Types:**

* **Missing Values:** There are no missing values in the dataset, ensuring the integrity of the data.
* **Data Types:**
  + 14 features are numerical (int64), mainly representing binary data.
  + 2 features are categorical (object): GENDER and LUNG\_CANCER.

**3. Univariate Analysis:**

* **Age Distribution:**
  + **Range:** 30 to 80 years
  + **Mean:** 55 years
  + **Median:** 58 years
  + **Standard Deviation:** 12.8 years
  + **Skewness:** Slightly left-skewed, indicating that a larger proportion of the population is older.
  + **Insights:** The age distribution suggests that the dataset covers a wide age range, with a concentration in the 50-60 year age bracket, which aligns with the demographic most at risk for lung cancer.
* **Gender Distribution:**
  + **Male:** 1,550 (51.7%)
  + **Female:** 1,450 (48.3%)
  + **Insights:** The gender distribution is relatively balanced, ensuring that gender-related insights into lung cancer risk can be adequately analyzed.
* **Lung Cancer Distribution:**
  + **Positive (YES):** 1,485 (49.5%)
  + **Negative (NO):** 1,515 (50.5%)
  + **Insights:** The lung cancer variable is almost perfectly balanced, which is advantageous for binary classification tasks and reduces concerns of class imbalance.

**4. Bivariate Analysis:**

* **Correlation Matrix:**
  + The correlation matrix revealed that most features are weakly correlated with each other, with values ranging mostly between -0.1 to 0.3.
  + **Strongest Correlation:** AGE and CHRONIC\_DISEASE had a correlation of approximately 0.30, suggesting older individuals are more likely to have chronic diseases.
  + **Insights:** The lack of high correlations (>0.7) between features suggests minimal multicollinearity, which is beneficial for model interpretability.
* **Gender vs. Lung Cancer:**
  + **Male Lung Cancer Rate:** 52.3% (811 out of 1,550 males)
  + **Female Lung Cancer Rate:** 46.6% (674 out of 1,450 females)
  + **Insights:** Males show a slightly higher incidence of lung cancer compared to females, which might be explored further in relation to lifestyle factors such as smoking.
* **Age vs. Lung Cancer:**
  + **Median Age of Positive Cases:** 63 years
  + **Median Age of Negative Cases:** 51 years
  + **Boxplot Insights:** Individuals diagnosed with lung cancer tend to be older, supporting the hypothesis that age is a significant risk factor for lung cancer.

**5. Multivariate Analysis:**

* **Categorical Features vs. Lung Cancer:**
  + **Smoking:**
    - **Lung Cancer Rate Among Smokers:** 68% (significantly higher than the overall rate)
    - **Non-Smokers:** Much lower prevalence, confirming smoking as a key risk factor.
  + **Chronic Disease:**
    - **With Chronic Disease:** 61% have lung cancer, reinforcing the importance of chronic illness as a co-morbidity.
  + **Alcohol Consumption:**
    - **Regular Alcohol Consumers:** 55% lung cancer rate, slightly above average, suggesting potential risk or correlation.
  + **Insights:** Lifestyle factors such as smoking, chronic disease, and alcohol consumption show clear associations with lung cancer prevalence, highlighting areas for targeted interventions.

**Key Insights:**

1. **Age and Smoking:**
   * **Age:** Older age is strongly associated with lung cancer, particularly in those aged 60 and above.
   * **Smoking:** Strongest lifestyle-related risk factor, with a nearly 70% lung cancer rate among smokers.
2. **Gender Differences:**
   * Males have a marginally higher risk of lung cancer than females, which may relate to historically higher smoking rates among males.
3. **Lifestyle Factors:**
   * **Chronic diseases and alcohol consumption** are important factors that, when combined with smoking, could elevate the risk further.