## **🎯 Project Title:**

**"Demographic Insights from a Stage III Lung Cancer Cohort"**

## **🗂️ I. Project Overview**

* **Objective**:  
   To explore and visualize demographic patterns—specifically age, gender, and BMI—among patients with Stage III lung cancer.
* **Data Source**:  
   An anonymized dataset of patients diagnosed with Stage III lung cancer.
* **Tools Used**:  
  + Python (Pandas, Seaborn, Matplotlib)
  + Jupyter Notebook

## **🧹 II. Data Preparation**

* **Initial Steps**:  
  + Identified relevant demographic columns: gender, age, bmi.
  + Detected formatting issues (e.g., commas instead of decimals).
  + Cleaned and converted age and BMI to numeric format.
* **Missing Data**:  
  + Non-numeric BMI or age entries were converted to NaN for robustness in analysis.

## **📊 III. Exploratory Data Analysis (EDA)**

### **1. Summary Statistics**

* Presented central tendencies and spread for age and BMI.
* Counted gender distribution.

### **2. Visualizations**

* **Age Distribution**:  
  + Histogram + KDE to show skewness and spread.
* **BMI Distribution**:  
  + Identified central tendencies and outliers.
* **BMI by Gender**:  
  + Boxplot highlighted differences in distribution between genders.
* **Age vs. BMI Scatterplot**:  
  + Visualized correlation with Pearson r = *[insert value]*.
  + Color-coded by gender to show overlap and trends.

## **⚖️ IV. Additional Insights**

* **BMI Classification** (Optional Step):  
  + Categorized patients into Underweight, Normal, Overweight, and Obese based on WHO standards.
  + Prepared the dataset for future clinical interpretation or predictive modeling.

## **🧠 V. Key Takeaways**

* Clear variation in BMI by gender, though overlapping distributions.
* Mild correlation between age and BMI in this population.
* Demonstrated ability to clean, analyze, and communicate real-world health data effectively.

## **🚀 VI. Future Directions**

* Join this demographic data with clinical outcomes (e.g., survival, recurrence).
* Stratify by smoking status or treatment timing.
* Explore multivariate models to predict risk factors for poor outcomes.