

## Story: Understanding Lung Cancer Patients through Data Analysis

### Introduction

Lung cancer remains the leading cause of cancer-related deaths worldwide. In this analysis, I leveraged clinical and genomics data to understand the characteristics of lung cancer patients, uncover patterns, and explore potential correlations that could aid in prognosis and treatment strategies. My ultimate goal is to contribute to the development of prognostic models and decision support systems for improved clinical practice.

#### 1: Unveiling the Data - Exploratory Data Analysis (EDA)

**Objective:** To understand the main characteristics and distributions of the dataset.

##### Analysis & Findings:

- **Age Distribution:** The patient ages ranged from 56 to 84, with a predominant age group of 68-78, forming a bell curve. This suggests that lung cancer is more common in older adults.
- **Tumor Size:** A bimodal distribution was observed, indicating two prevalent tumor sizes among patients. This could suggest different subtypes or stages of lung cancer.
- **Number of Mutated Genes:** The data skewed to the left, showing that most patients had a lower number of mutated genes.

#### 2: Digging Deeper - Bivariate and Multivariate Analysis

**Objective:** To explore relationships between different variables and understand how they interact with each other.

##### Analysis & Findings:

- **Tumor Size vs. Survival Months:** No clear trend was observed, indicating that tumor size alone may not be a sufficient predictor of survival.
- **Age vs. Number of Mutated Genes:** The scatter plot showed two distinct groups, possibly indicating different subtypes or stages of the disease.
- **Multivariate Analysis:** By adding another dimension like "Survival.months" to the color shelf, I could visualize how survival months vary with tumor size and age, adding depth to my analysis.

#### 3: Identifying Trends and Patterns

**Objective:** To find trends and patterns that could lead to more profound insights.

##### Analysis & Findings:

- **Comparative Analysis:** By comparing different groups based on characteristics like tumor stage and age, I aimed to uncover patterns. However, the box plots did not reveal any consistent pattern.
- **Correlation Analysis:** I attempted to find correlations between variables like age, tumor size, and number of mutated genes. However, the scatter plots did not show strong correlations.

#### 4: Genomic Data Exploration (relationship between genes and their mutations to see which gene has most mutations in Lung cancer)

**Objective:** To explore the genomic data and understand its relationship with clinical variables.

##### Analysis & Findings:

- **Gene Mutation Analysis:** I created heatmaps to visualize the relationship between gene mutations and numerical variables like tumor size. This could potentially uncover specific genes that are frequently mutated in larger tumors or more aggressive cancer stages.

#### 5: Compiling Insights - Dashboard and Reporting

**Objective:** To consolidate all my findings into an interactive dashboard for easy exploration and interpretation.

##### Analysis & Findings:

- **Interactive Dashboard:** I created a Tableau dashboard that combines all my visualizations, allowing stakeholders to interactively explore the data and draw their own conclusions.

#### Conclusion and Recommendations

While my analysis provided valuable insights into the characteristics and distributions of lung cancer patients, the lack of clear trends and correlations indicates the complexity and heterogeneity of the disease. Lung cancer is a multifaceted disease, and a comprehensive approach considering demographic, clinical, genomic, and possibly imaging data is crucial for accurate prognosis and treatment planning.

#### Recommendations:

- **Further Data Collection:** More comprehensive data, including imaging and more detailed genomic data, may provide additional insights.
- **Advanced Modeling:** Consider using advanced machine learning models that can handle high-dimensional data and capture complex relationships.
- **Clinical Collaboration:** Work closely with clinicians to validate findings and ensure that the models and insights are clinically relevant.

By following these recommendations, I aim to enhance our understanding of lung cancer, contribute to the development of prognostic models, and ultimately improve patient care and outcomes.