Lungs Cancer Analysis

EDA USING R

Problem Questions

- What is the distribution of lung cancer cases?
- How does age vary among individuals with and without lung cancer?
- What is the age distribution of patients with lung cancer?
- Are there any strong correlations between any of the predictor variables and Lung Cancer cases?

Data distribution in our dataset

Summary of Dataset

```
> summary(Cancer_df)
    GENDER                 AGE            SMOKING       YELLOW_FINGERS     ANXIETY       PEER_PRESSURE   CHRONIC.DISEASE     FATIGUE
 Length:309         Min.   :21.00   Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 Class :character   1st Qu.:57.00   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000
 Mode  :character   Median :62.00   Median :2.000   Median :2.00   Median :1.000   Median :2.000   Median :2.000   Median :2.000
                    Mean   :62.67   Mean   :1.563   Mean   :1.57   Mean   :1.498   Mean   :1.502   Mean   :1.505   Mean   :1.673
                    3rd Qu.:69.00   3rd Qu.:2.000   3rd Qu.:2.00   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
                    Max.   :87.00   Max.   :2.000   Max.   :2.00   Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000
    ALLERGY         WHEEZING      ALCOHOL.CONSUMING    COUGHING     SHORTNESS.OF.BREATH SWALLOWING.DIFFICULTY   CHEST.PAIN    LUNG_CANCER
 Min.   :1.000   Min.   :1.000   Min.   :1.000     Min.   :1.000   Min.   :1.000       Min.   :1.000          Min.   :1.000   Length:309
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000     1st Qu.:1.000   1st Qu.:1.000       1st Qu.:1.000          1st Qu.:1.000   Class :character
 Median :2.000   Median :2.000   Median :2.000     Median :2.000   Median :2.000       Median :1.000          Median :2.000   Mode  :character
 Mean   :1.557   Mean   :1.557   Mean   :1.557     Mean   :1.579   Mean   :1.641       Mean   :1.469          Mean   :1.557
 3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000     3rd Qu.:2.000   3rd Qu.:2.000       3rd Qu.:2.000          3rd Qu.:2.000
 Max.   :2.000   Max.   :2.000   Max.   :2.000     Max.   :2.000   Max.   :2.000       Max.   :2.000          Max.   :2.000
```

Structure of Dataset

```
'data.frame':   309 obs. of  17 variables:
 $ GENDER                : chr  "Male" "Male" "Female" "Male"
...
 $ AGE                   : int  69 74 59 63 63 75 52 51 68 53
...
 $ SMOKING               : int  1 2 1 2 1 1 2 2 2 2 ...
 $ YELLOW_FINGERS        : chr  "Yes" "No" "No" "Yes" ...
 $ ANXIETY               : chr  "Yes" "No" "No" "Yes" ...
 $ PEER_PRESSURE         : chr  "No" "No" "Yes" "No" ...
 $ CHRONIC.DISEASE       : chr  "No" "Yes" "No" "No" ...
 $ FATIGUE               : chr  "Yes" "Yes" "Yes" "No" ...
 $ ALLERGY               : chr  "No" "Yes" "No" "No" ...
 $ WHEEZING              : int  2 1 2 1 2 2 2 1 1 1 ...
 $ ALCOHOL.CONSUMING     : chr  "Yes" "No" "No" "Yes" ...
 $ COUGHING              : chr  "Yes" "No" "Yes" "No" ...
 $ SHORTNESS.OF.BREATH   : chr  "Yes" "Yes" "Yes" "No" ...
 $ SWALLOWING.DIFFICULTY: chr  "Yes" "Yes" "No" "Yes" ...
 $ CHEST.PAIN            : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ LUNG_CANCER           : chr  "YES" "YES" "NO" "NO" ...
```

Top and Bottom 5 Values

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC.DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL.CONSUMING	COUGHING
1	M	69	1	2	2	1	1	2	1	2	2	2
2	M	74	2	1	1	1	2	2	2	1	1	1
3	F	59	1	1	1	2	1	2	1	2	1	2
4	M	63	2	2	2	1	1	1	1	1	2	1
5	F	63	1	2	1	1	1	1	1	2	1	2

	SHORTNESS.OF.BREATH	SWALLOWING.DIFFICULTY	CHEST.PAIN	LUNG_CANCER
1	2	2	2	YES
2	2	2	2	YES
3	2	1	2	NO
4	1	2	2	NO
5	2	1	1	NO

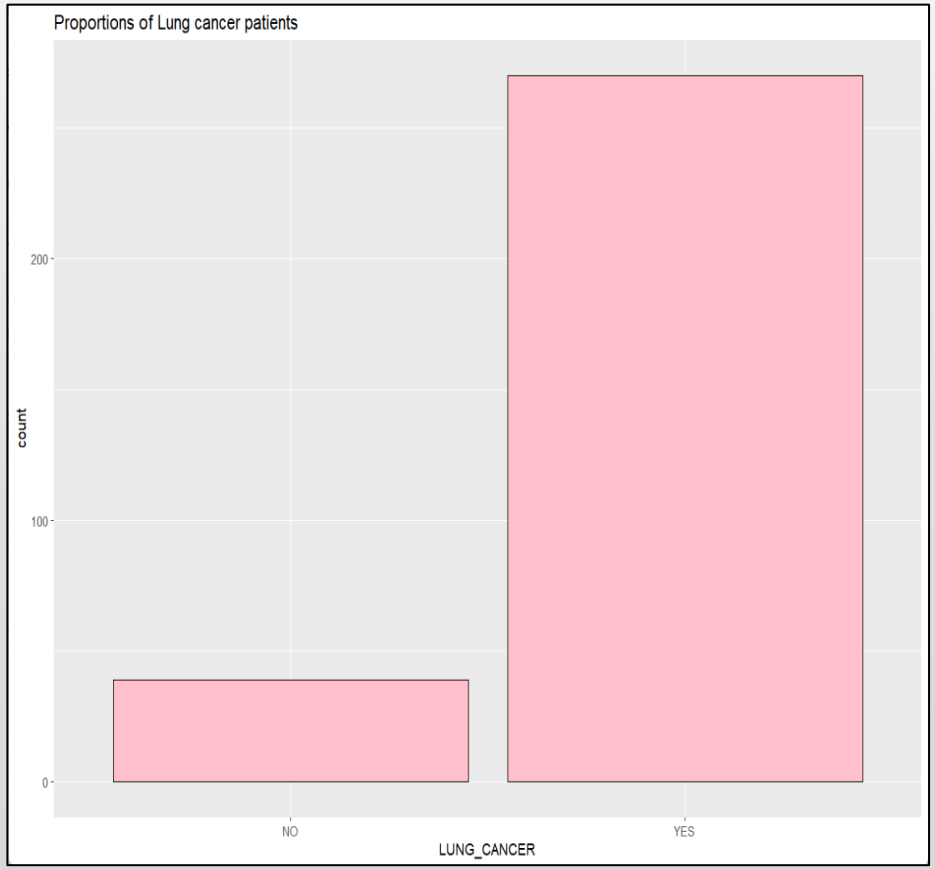
	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC.DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL.CONSUMING	COUGHING
305	F	56	1	1	1	2	2	2	1	1	2	2
306	M	70	2	1	1	1	1	2	2	2	2	2
307	M	58	2	1	1	1	1	1	2	2	2	2
308	M	67	2	1	2	1	1	2	2	1	2	2
309	M	62	1	1	1	2	1	2	2	2	2	1

	SHORTNESS.OF.BREATH	SWALLOWING.DIFFICULTY	CHEST.PAIN	LUNG_CANCER
305	2	2	1	YES
306	2	1	2	YES
307	1	1	2	YES
308	2	1	2	YES
309	1	2	1	YES

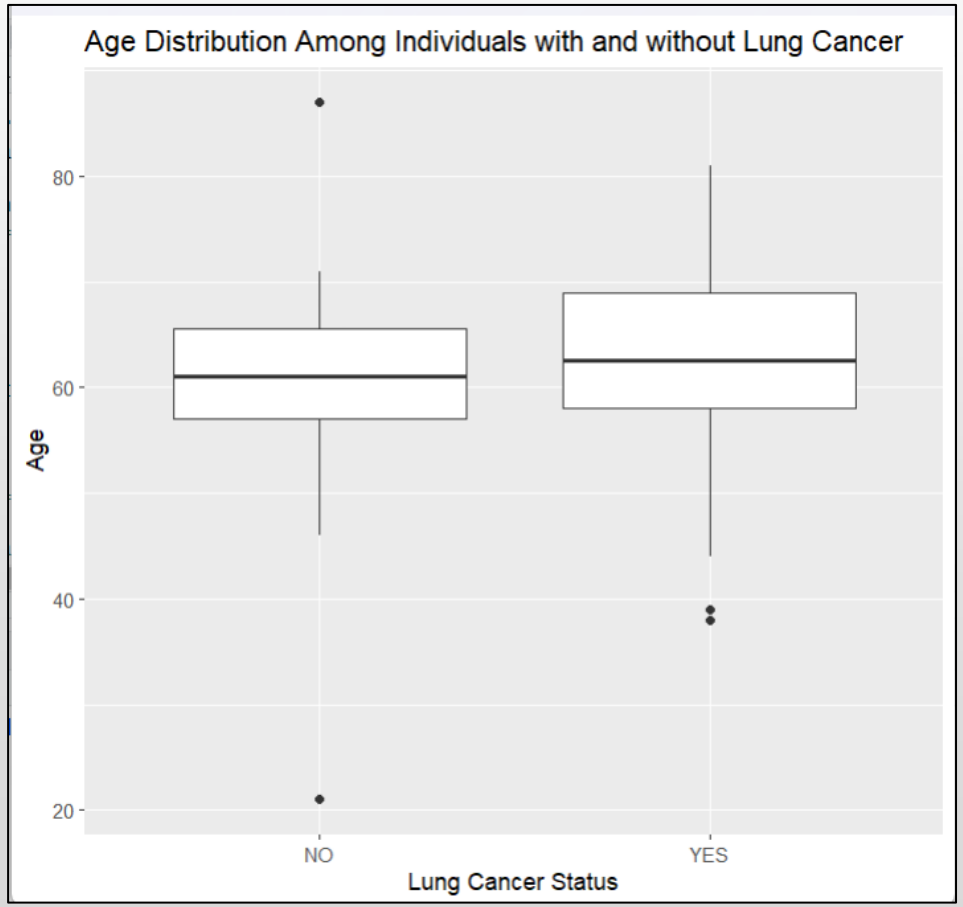
EDA – Exploratory Data Analysis

- Proportions of Lung cancer

NO	YES
0.13	0.87

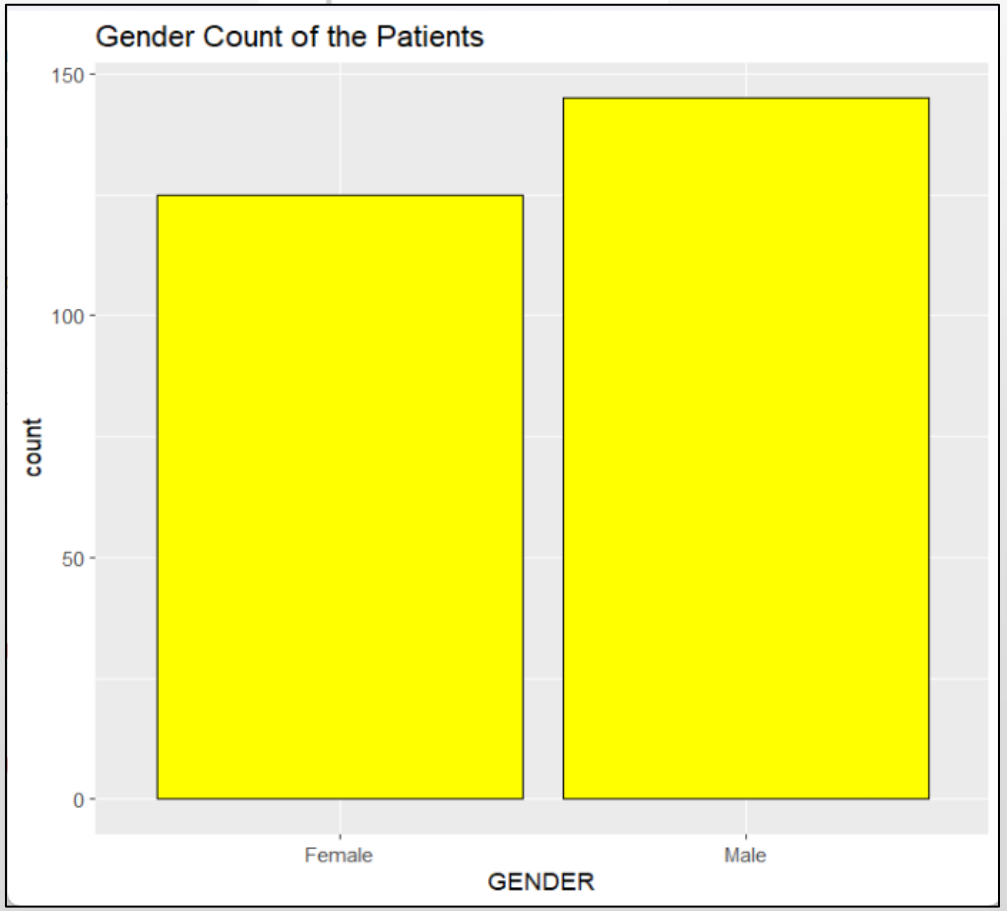


Age Distribution Among Individuals with and without Lung Cancer

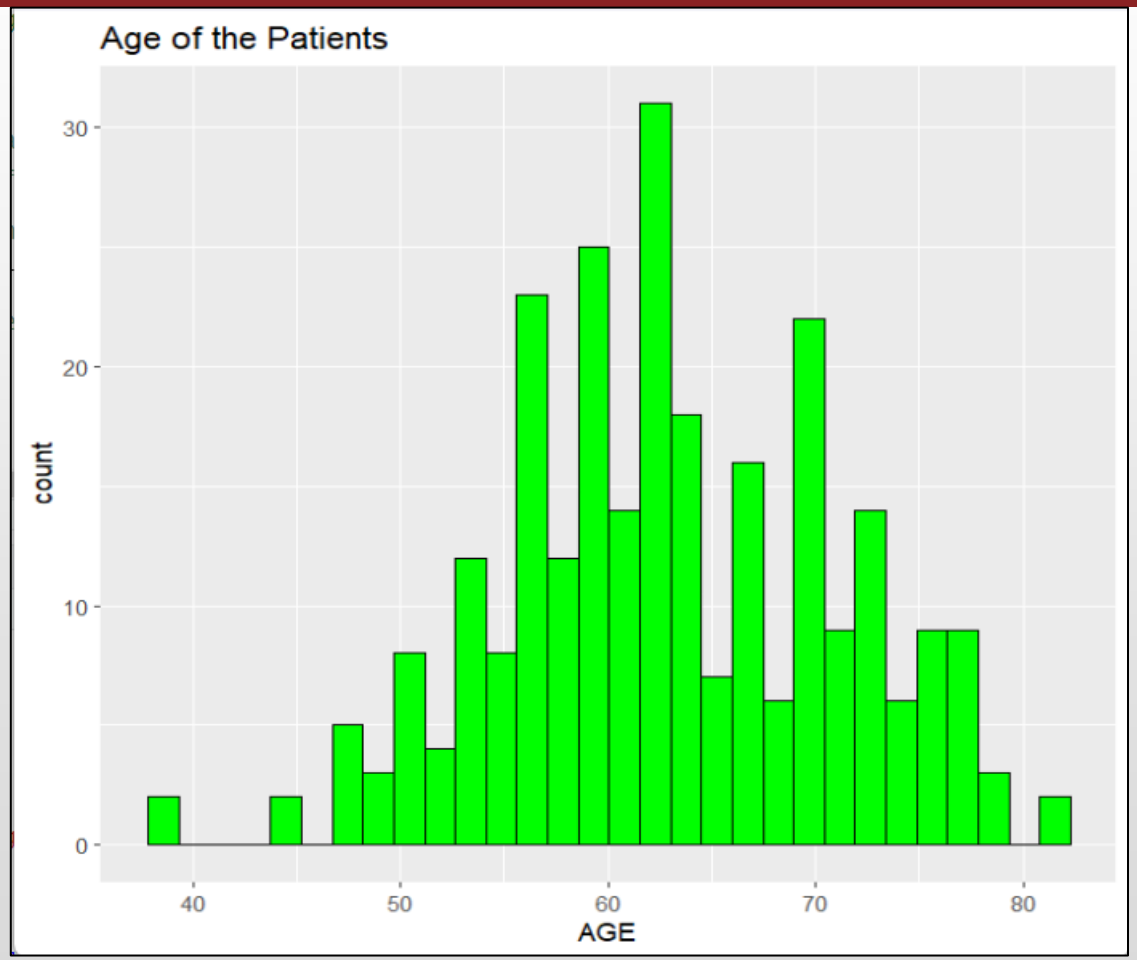


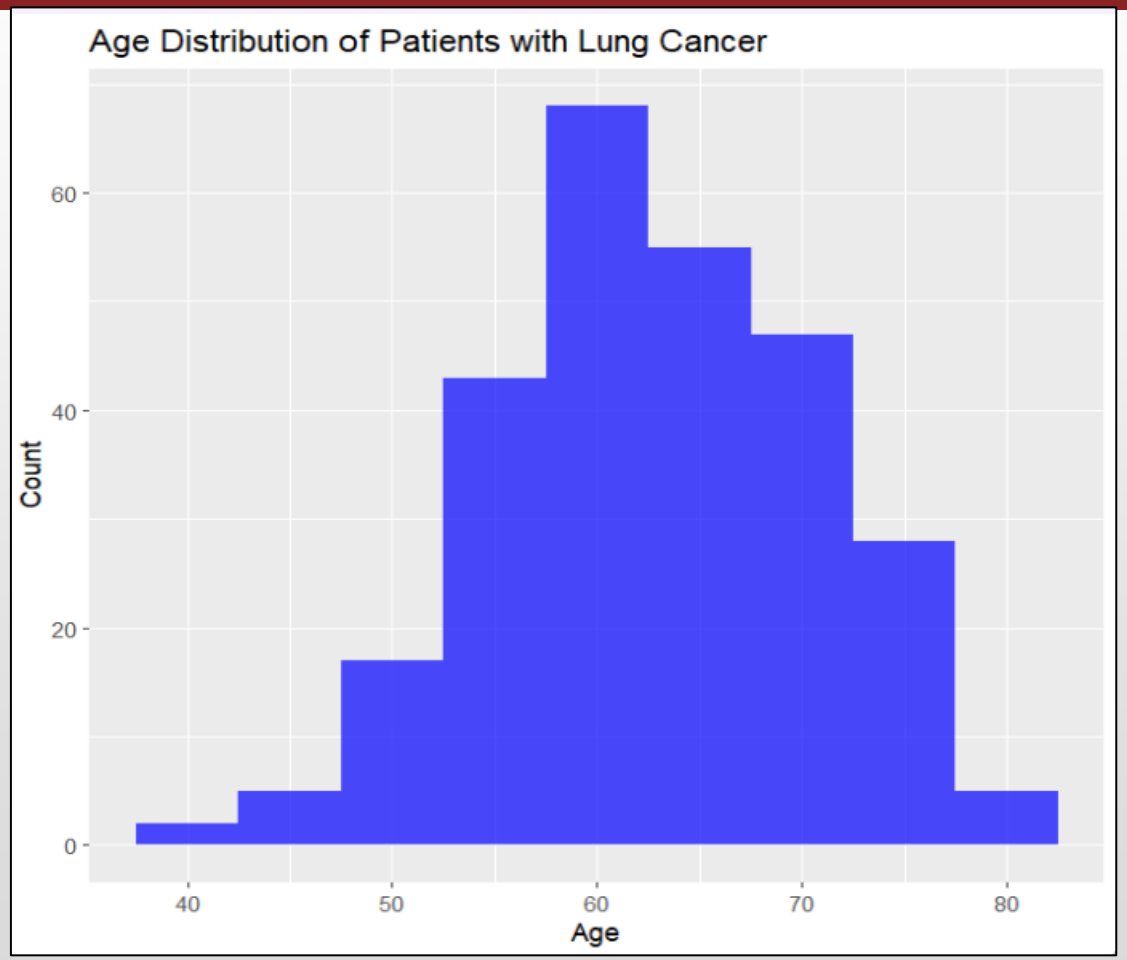
Proportions of gender

```
Female   Male
  0.46   0.54
```



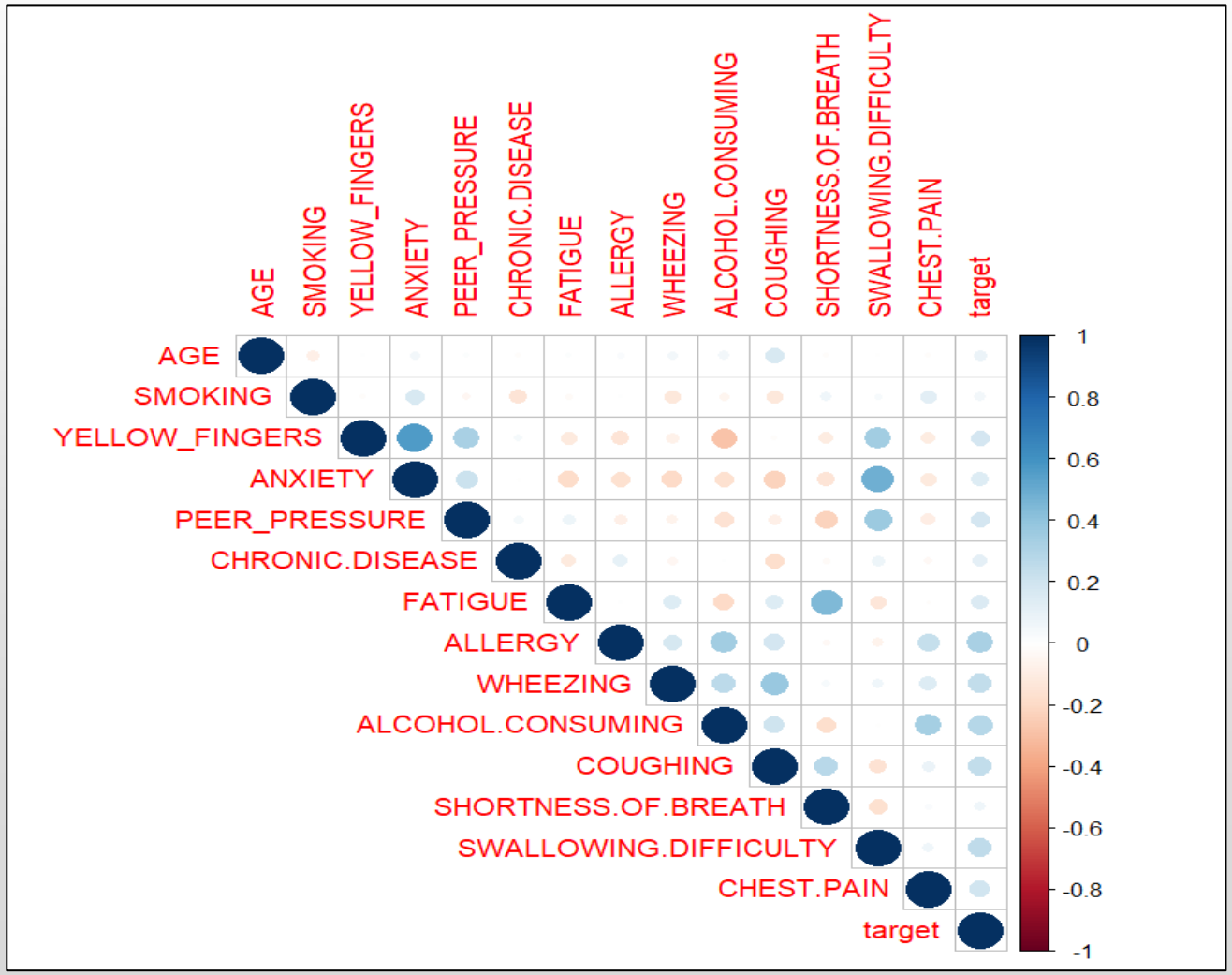
Age distribution of patients with lung cancer





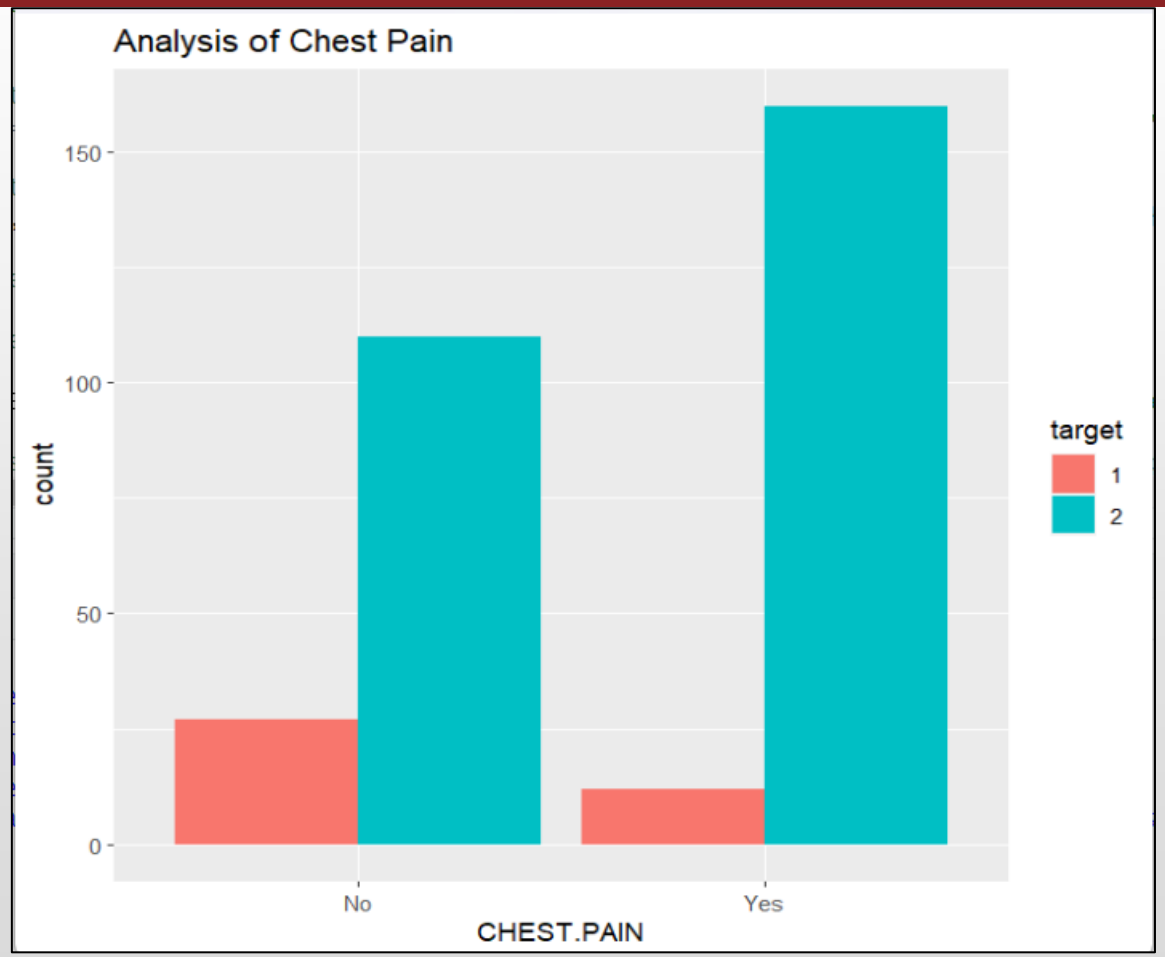
By looking at the histogram, we can determine the peak age group is 55 to 77 where the prevalence of lung cancer is highest.

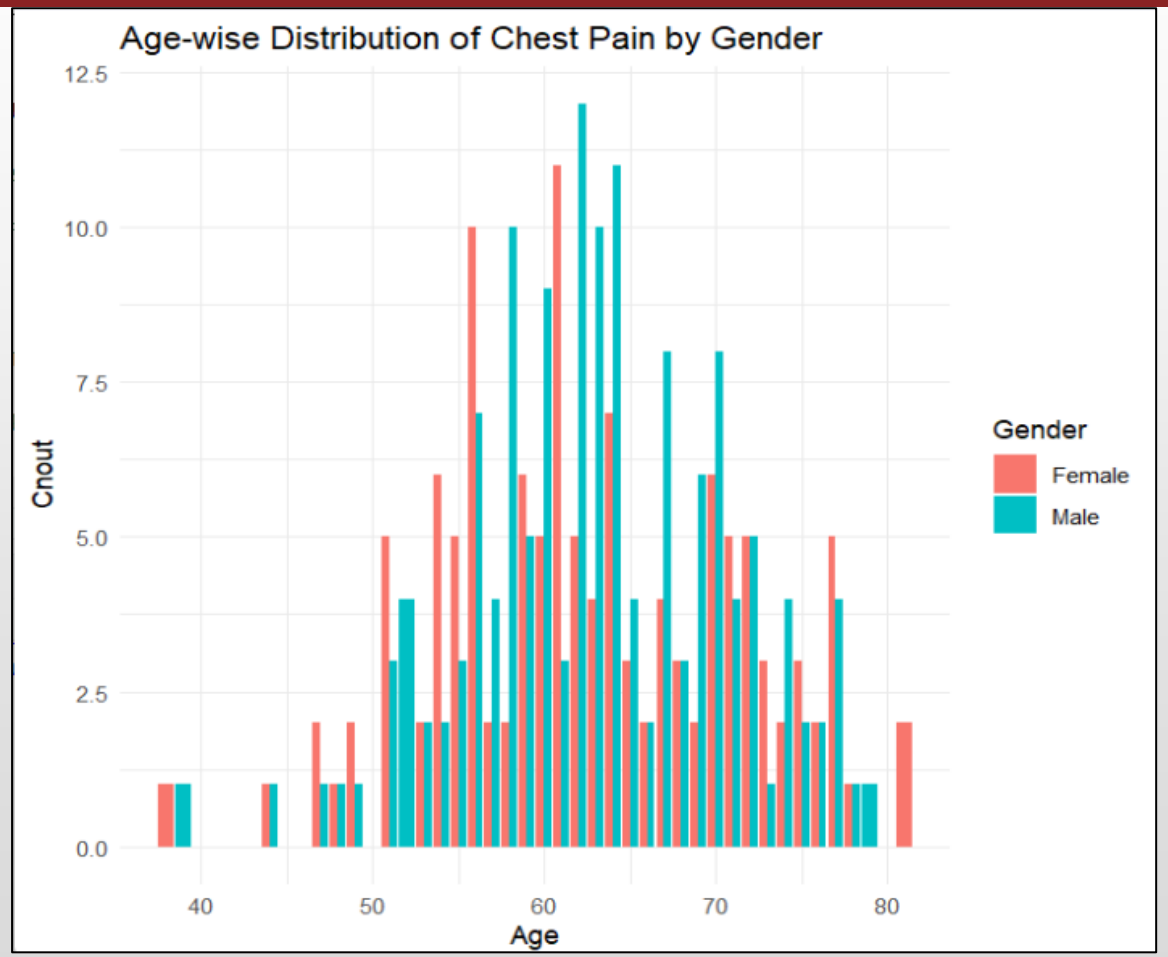
Correlation of variables with each other



A robust and statistically significant association exists between Lung Cancer and both Allergy and Alcohol Consumption, indicating a higher risk of Lung Cancer among individuals with allergies and those who consume alcohol. Further research is needed to explore the causal mechanisms underlying these strong relationships.

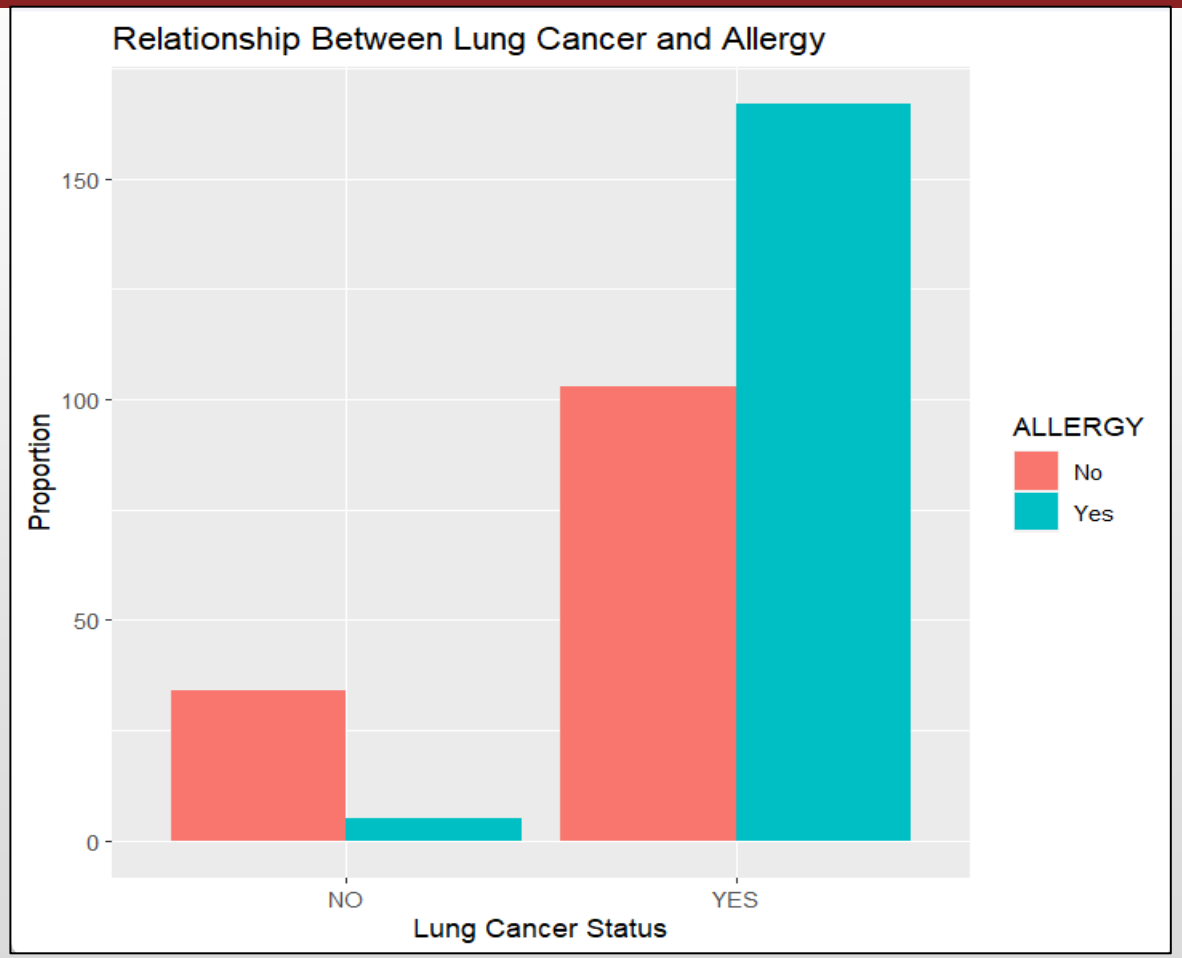
Relation between the Lung Cancer and chest pain





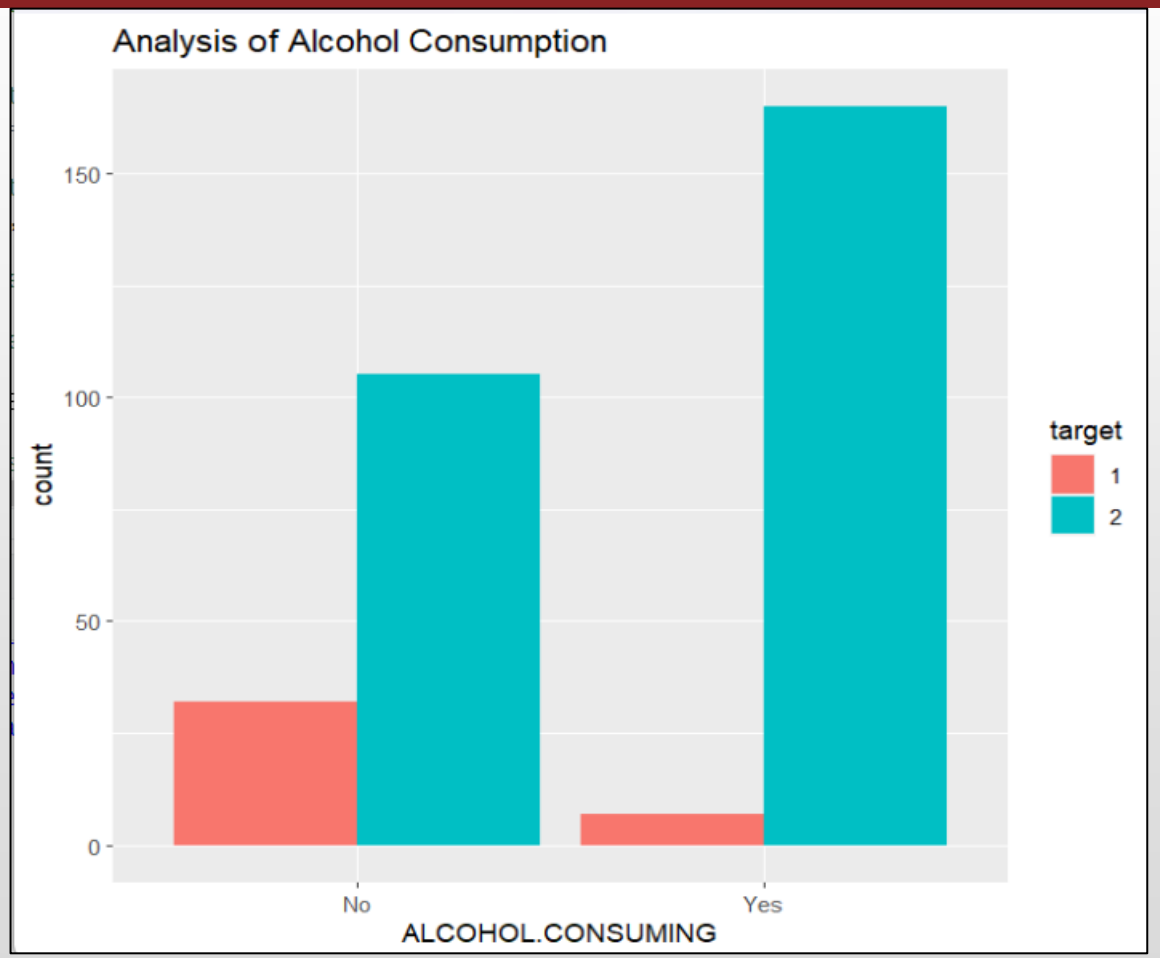
The conclusion that can be drawn from this analysis is that there appears to be an increased incidence of chest pain in the age group of 57 to 65 in the dataset I examined.

Relationship Between Lung Cancer and Allergy



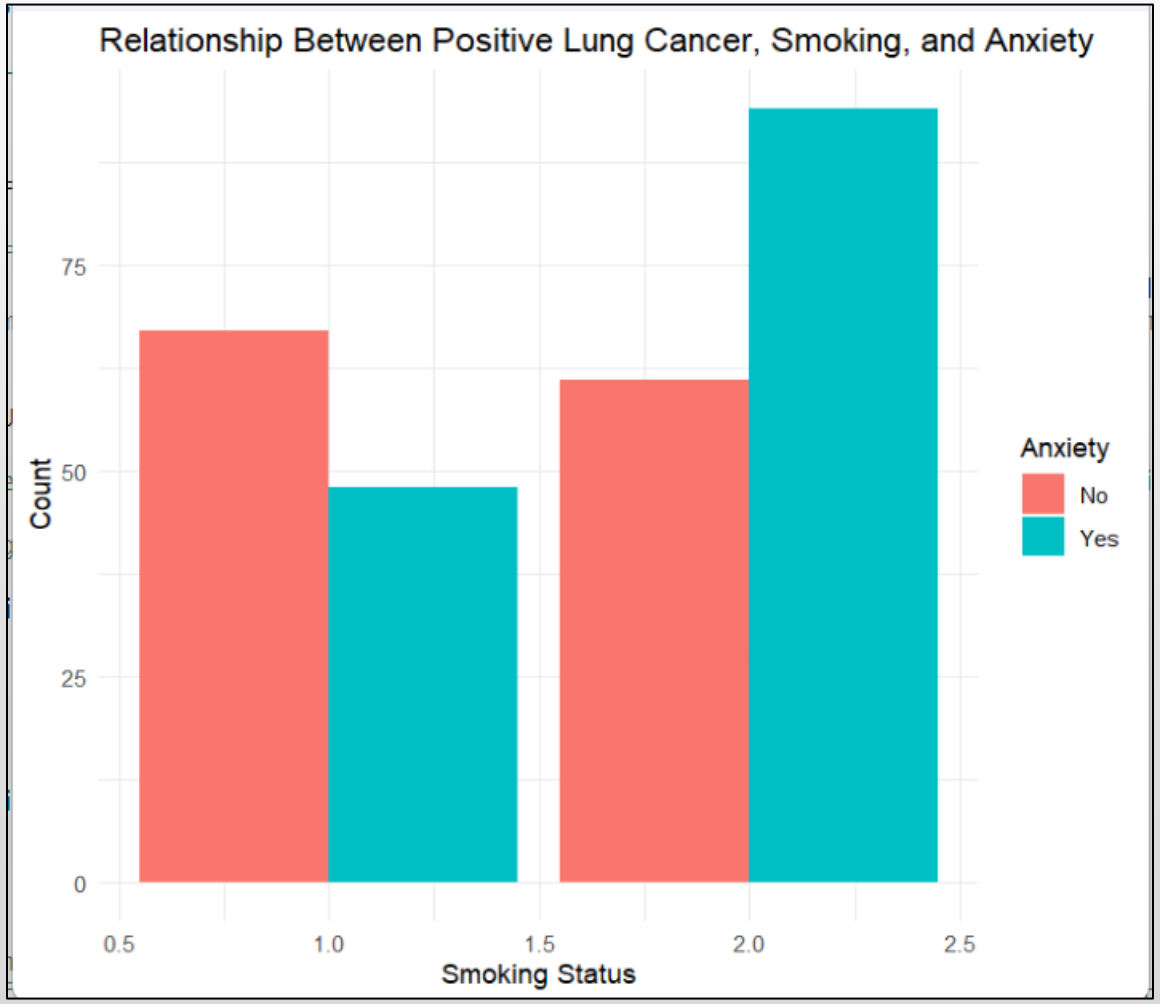
The presence of allergies appears to have a limited impact on individuals without lung cancer, suggesting a potential link between allergies and the risk of developing lung cancer.

Relationship Between Lung Cancer and Alcohol Consumption



The data suggests a minimal influence of alcohol consumption on the prevalence of lung cancer, suggesting that its relationship to the disease is not strongly pronounced

Relationship between lung cancer, smoking, and anxiety



Lung cancer can occur in individuals with both smoking and anxiety.
Over 60 cases of lung cancer are found in individuals who neither smoke nor have anxiety (genetics).
This highlights the complex nature of lung cancer, involving multiple risk factors and the influence of factors beyond smoking and anxiety.

- The data reveals a higher incidence of lung cancer in males compared to females, indicating a significant gender disparity in the prevalence of this disease
- The peak prevalence of lung cancer occurs in the age group of 55 to 80, signifying a prominent age-related pattern in disease occurrence.
- Upon further in-depth analysis, it becomes evident that allergies exhibit a discernible impact on lung cancer, while alcohol consumption demonstrates a relatively limited influence on the prevalence of this disease.
- The analysis suggests an elevated occurrence of chest pain within the age range of 57 to 65, yet it does not appear to have a substantial impact on the prevalence of lung cancer in the examined dataset.
- Lung cancer happens for different reasons, like smoking and anxiety, but it's also influenced by your genes, making it a complex condition.

Findings and Insight

According to our Analysis

- Implement comprehensive public health campaigns to reduce smoking rates and increase awareness about the risks associated with smoking.
- Promote regular health check-ups and lung cancer screenings, especially for individuals in high-risk groups, to diagnose the disease at an earlier, more treatable stage.
- Offer genetic counseling and testing for individuals with a family history of lung cancer to assess their risk and provide personalized recommendations.
- Develop programs and resources for managing anxiety and stress, potentially reducing one of the contributing factors to lung cancer.
- Invest in research to better understand the complex causes of lung cancer and educate the public about the multiple risk factors involved.
- Provide support and resources for individuals with lung cancer who do not smoke but have genetic predisposition, addressing the unique needs of this group.

Points of Solution