Genetic Predisposition and Environmental Factors: Understanding the Roots of Lung Cancer

By

Bryan Wu – 2602087900

Yonathan Henry Christianto – 2602158683

Data Mining and Visualization's

Assurance of Learning

Binus University 2023/2024

ABSTRACT

Lung cancer remains a formidable challenge in public health, demanding a nuanced understanding of its multifaceted origins and risk factors. This research delves into the intricate interplay of genetic, environmental, and lifestyle contributors, aiming to unravel the complex tapestry that leads to lung carcinogenesis. Employing a diverse range of datasets and advanced analytical techniques, we explore the significance of factors beyond tobacco smoke, including environmental exposures, genetic predisposition, occupational hazards, and lifestyle choices. Our findings highlight the intricate network of interactions influencing lung cancer susceptibility, shedding light on lesser-known risk factors and their potential contribution to the disease. The impact of air quality, inflammatory pathways, and emerging biomarkers is meticulously examined, providing a comprehensive picture of the intricate molecular landscape associated with lung cancer. Furthermore, this study underscores the necessity of a holistic approach in understanding lung cancer causes, as we navigate through the amalgamation of intrinsic and extrinsic elements that shape an individual's risk profile. By synthesizing insights from diverse sources, we contribute to the ongoing discourse on precision medicine, emphasizing the need for tailored prevention and intervention strategies based on an individual's unique risk profile. In conclusion, our research seeks to advance the field's understanding of lung cancer etiology, moving beyond traditional paradigms and encouraging a more nuanced perspective on the intricate factors contributing to this devastating disease.

1. BACKGROUND

Lung cancer, a formidable and pervasive malignancy, represents a significant global health challenge. Characterized by uncontrolled cell growth in the lungs, this disease manifests in various forms, with non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) being the primary classifications. While smoking, notably tobacco use, has historically been a predominant risk factor, recent research has unveiled the intricate interplay of genetic predisposition and environmental exposures in shaping an individual's susceptibility. The symptoms of lung cancer, often subtle in the early stages, include persistent cough, chest pain, and difficulty breathing, contributing to delayed diagnosis and advanced disease at presentation.

With advancements in diagnostic technologies, such as computed tomography (CT) scans and molecular profiling, there is a growing emphasis on early detection and personalized treatment strategies. Despite progress in therapeutic modalities, challenges persist, underscoring the urgent need for continued research into the diverse molecular mechanisms and novel interventions that can improve outcomes for individuals affected by this complex and devastating disease. Lung cancer, with its multifaceted nature, demands a comprehensive approach encompassing prevention, early detection, and innovative treatment paradigms to confront its global impact on public health.

2. OBJECTIVES

The objectives of this article are to:

1. To determine the causes of lung cancer
2. To find out the most prominent aspects that effect the causes of lung cancer
3. To explore the datasets provided by kaggle about lung cancer and connect each attributes that available in the datasets
4. To find out what are the symptoms of lung cancer patients

3. METHODOLOGY

The method used in this paper about lung cancer datasets explanatory is a analysis and exploratory method using R language to execute the mining process with R Studio as our platform.

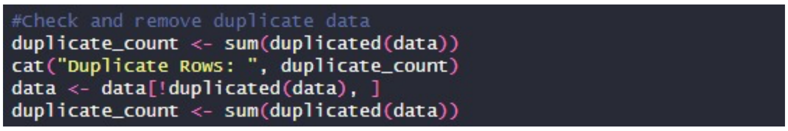
4. RESULTS AND DISCUSSION

4.1. DATA INTRODUCTION

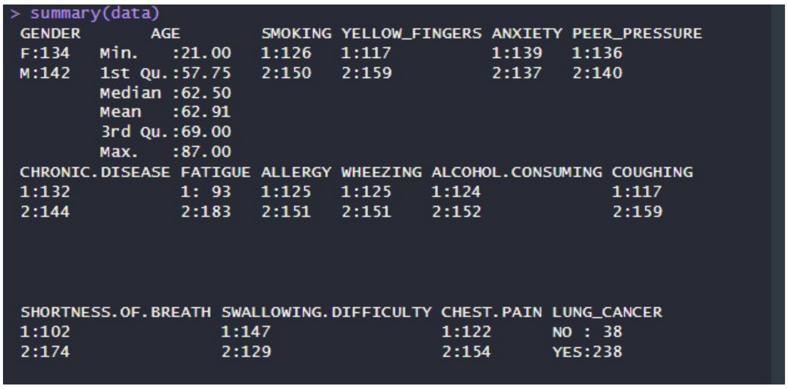
The dataset that will be used for this explanatory is called “Lung Cancer”. It is a dataset found from Kaggle that contains various attributes and variables about each aspects that could possibly effects the causes of lung cancer. The dataset itself contains 284 attributes and 16 variables.

The goals of this explanatory is to reach a conclusion that the dataset provides and create a full explanation, exploration, visualization, and discussion based on the data provided by the dataset.

4.2. DATA ANALYSIS AND EXPLORATORY

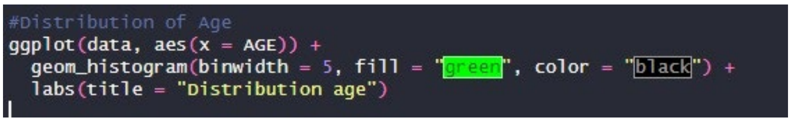


```
#Check and remove duplicate data
duplicate_count <- sum(duplicated(data))
cat("Duplicate Rows: ", duplicate_count)
data <- data[!duplicated(data), ]
duplicate_count <- sum(duplicated(data))
```



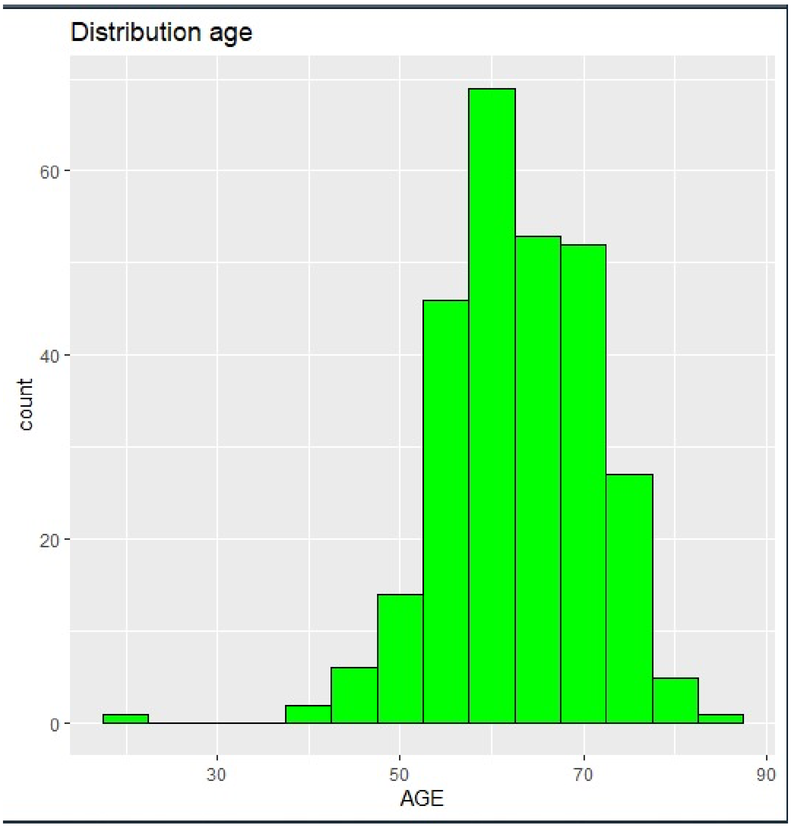
```
> summary(data)
 GENDER       AGE          SMOKING YELLOW_FINGERS ANXIETY PEER_PRESSURE
 F:134   Min.   :21.00   1:126   1:117          1:139   1:136
 M:142   1st Qu.:57.75   2:150   2:159          2:137   2:140
         Median :62.50
         Mean   :62.91
         3rd Qu.:69.00
         Max.   :87.00
 CHRONIC.DISEASE FATIGUE ALLERGY WHEEZING ALCOHOL.CONSUMING COUGHING
 1:132           1: 93   1:125   1:125    1:124             1:117
 2:144           2:183   2:151   2:151    2:152             2:159




 SHORTNESS.OF.BREATH SWALLOWING.DIFFICULTY CHEST.PAIN LUNG_CANCER
 1:102               1:147                 1:122      NO : 38
 2:174               2:129                 2:154      YES:238
```

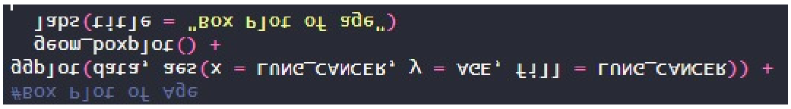
Above are the main exploratory of the dataset with the duplicated data already removed.



```
#Distribution of Age
ggplot(data, aes(x = AGE)) +
  geom_histogram(binwidth = 5, fill = "green", color = "black") +
  labs(title = "Distribution age")
```

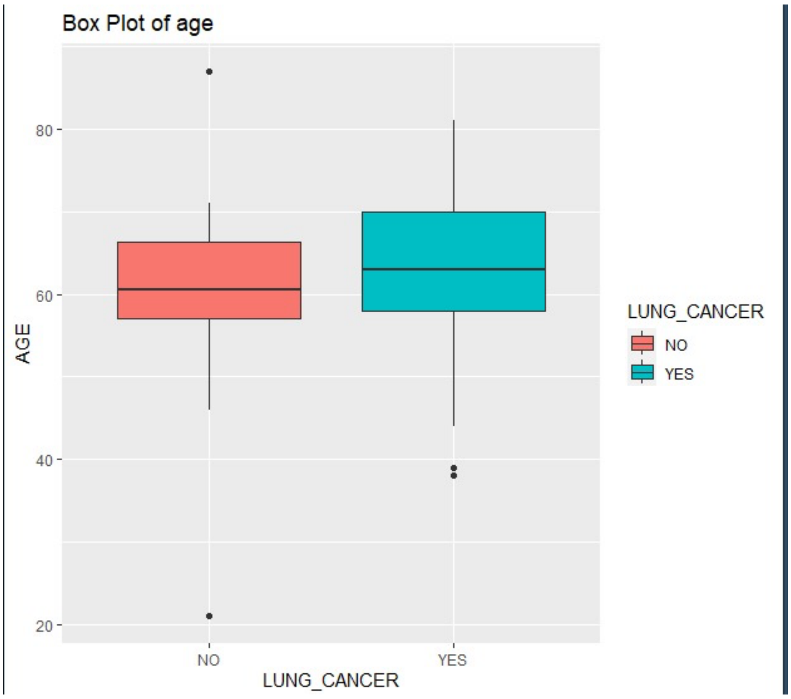
First is Distribution of Age. Below is the graph that shows a number of distribution of lung cancer based on their ages.

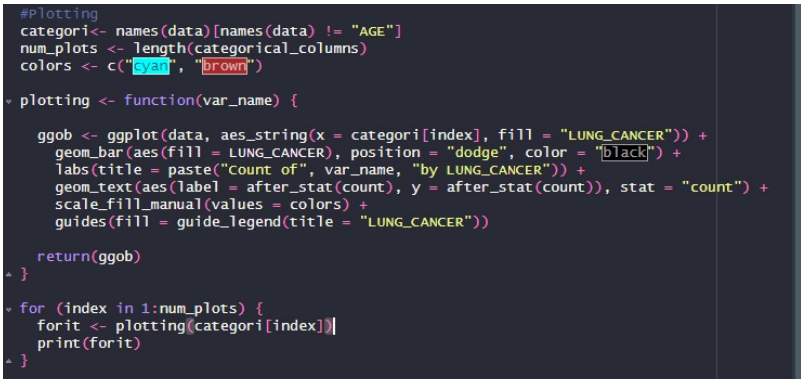




```
#Box plot of Age
ggplot(data, aes(x = LUNG_CANCER, y = AGE, fill = LUNG_CANCER)) +
  geom_boxplot() +
  labs(title = "Box plot of age")
```

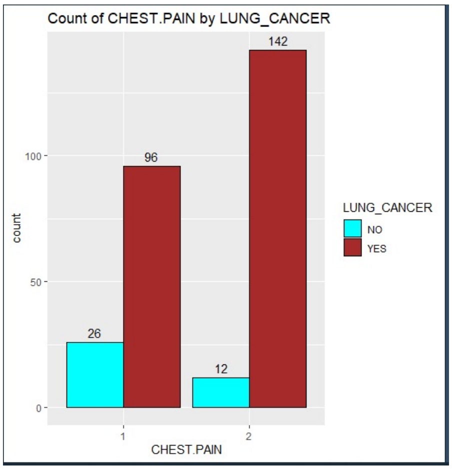
Below is the graph of age distribution using Box Plot on age distributions.

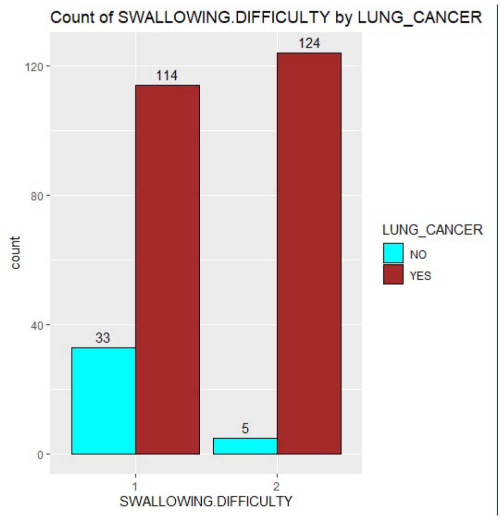


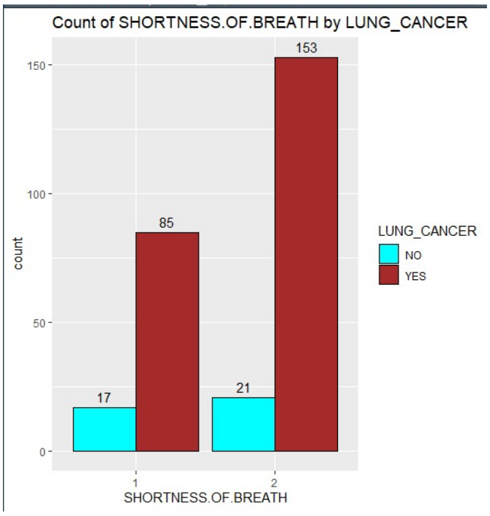


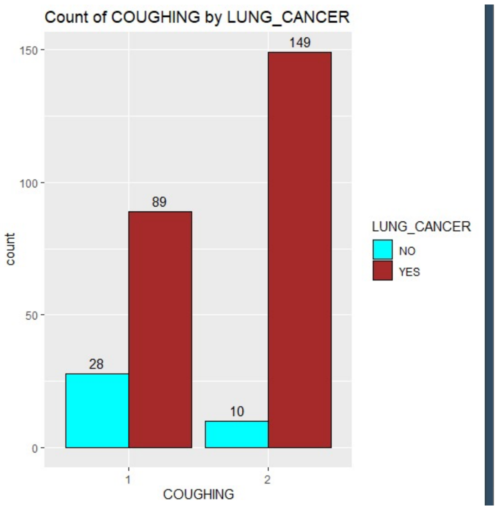
```
#Plotting
categori<- names(data)[names(data) != "AGE"]
num_plots <- length(categorical_columns)
colors <- c("cyan", "brown")

plotting <- function(var_name) {

  ggob <- ggplot(data, aes_string(x = categori[index], fill = "LUNG_CANCER")) +
    geom_bar(aes(fill = LUNG_CANCER), position = "dodge", color = "black") +
    labs(title = paste("Count of", var_name, "by LUNG_CANCER")) +
    geom_text(aes(label = after_stat(count), y = after_stat(count)), stat = "count") +
    scale_fill_manual(values = colors) +
    guides(fill = guide_legend(title = "LUNG_CANCER"))

  return(ggob)
}

for (index in 1:num_plots) {
  forit <- plotting(categori[index])
  print(forit)
}
```

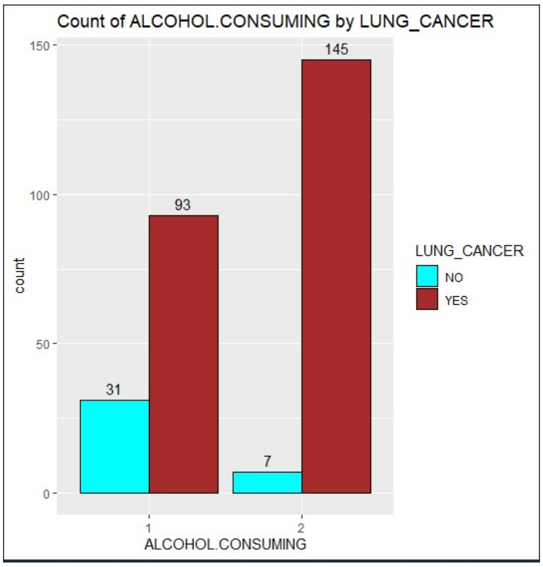
Above is a code on a multivariate graph based on each variables available in the dataset. The result will be shown below.

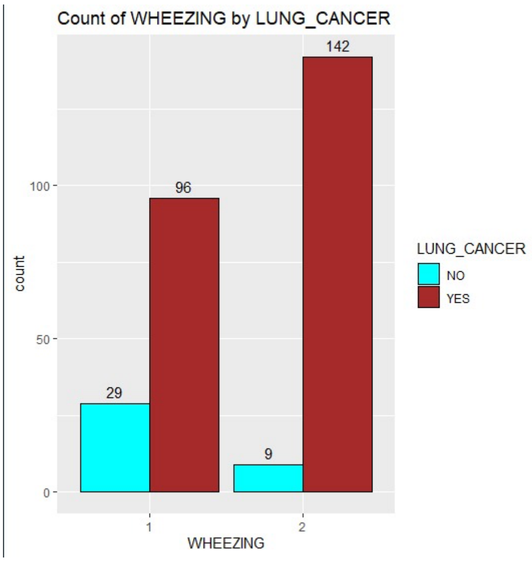

Count of CHEST.PAIN by LUNG_CANCER
count
100
50
0
26
96
12
142
1
2
CHEST.PAIN
LUNG_CANCER
NO
YES

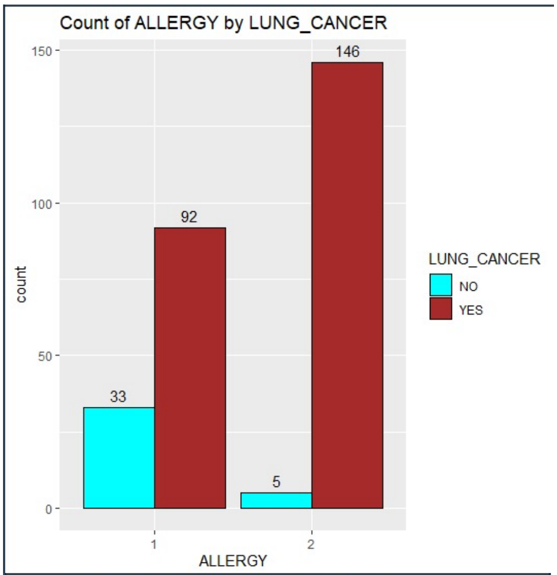

Count of SWALLOWING.DIFFICULTY by LUNG_CANCER
count
120
80
40
0
33
114
5
124
1
2
SWALLOWING.DIFFICULTY
LUNG_CANCER
NO
YES

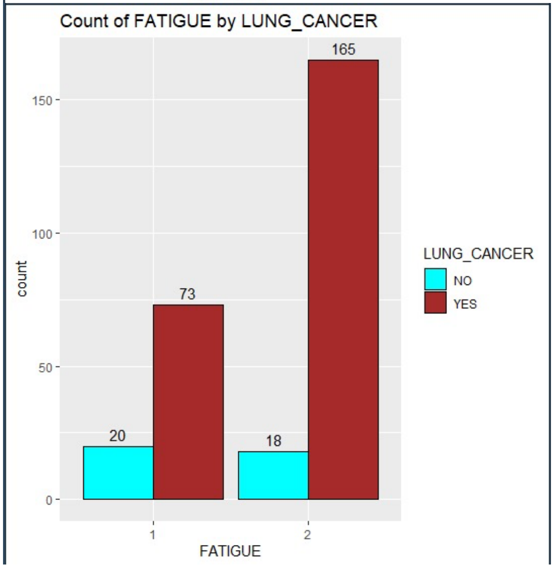

Count of SHORTNESS.OF.BREATH by LUNG_CANCER
count
150
100
50
0
17
85
21
153
1
2
SHORTNESS.OF.BREATH
LUNG_CANCER
NO
YES

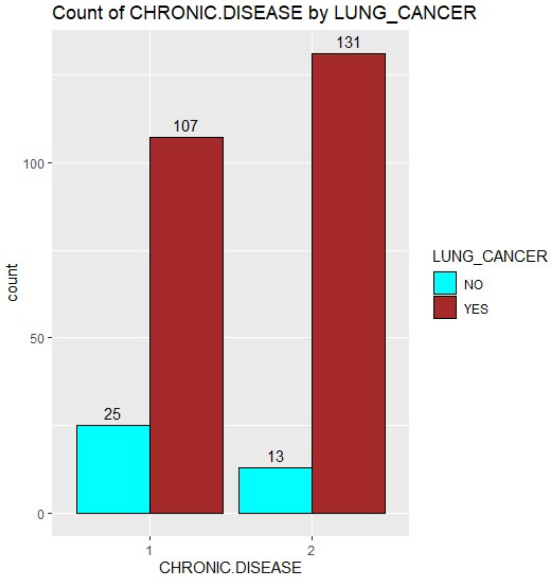

Count of COUGHING by LUNG_CANCER
count
150
100
50
0
28
89
10
149
1
2
COUGHING
LUNG_CANCER
NO
YES

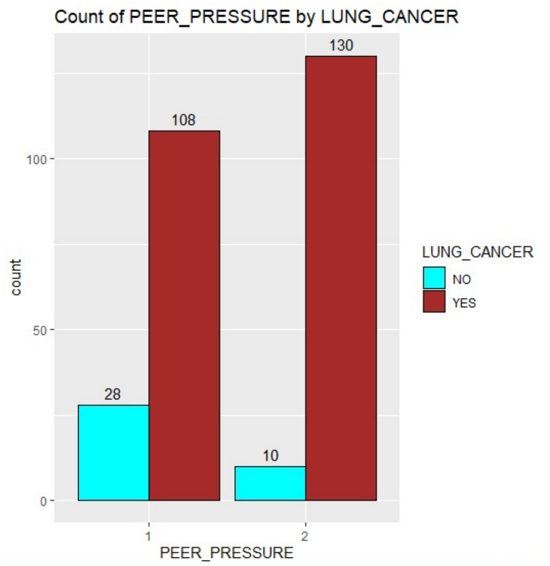

Count of ALCOHOL.CONSUMING by LUNG_CANCER
count
150
100
50
0
31
93
7
145
1
2
ALCOHOL.CONSUMING
LUNG_CANCER
NO
YES

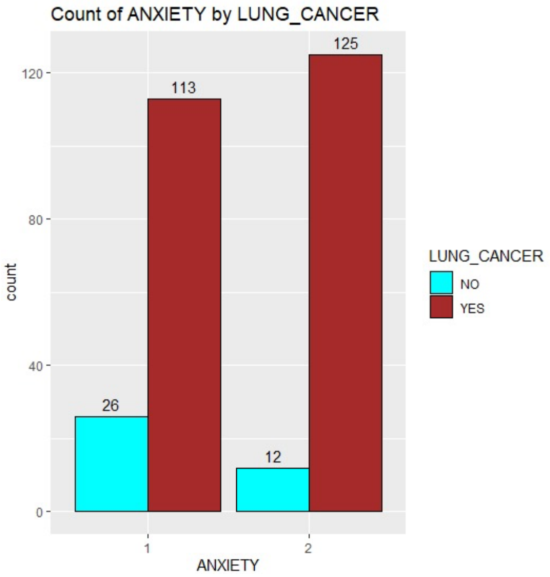

Count of WHEEZING by LUNG_CANCER
count
100
50
0
29
96
9
142
1
2
WHEEZING
LUNG_CANCER
NO
YES

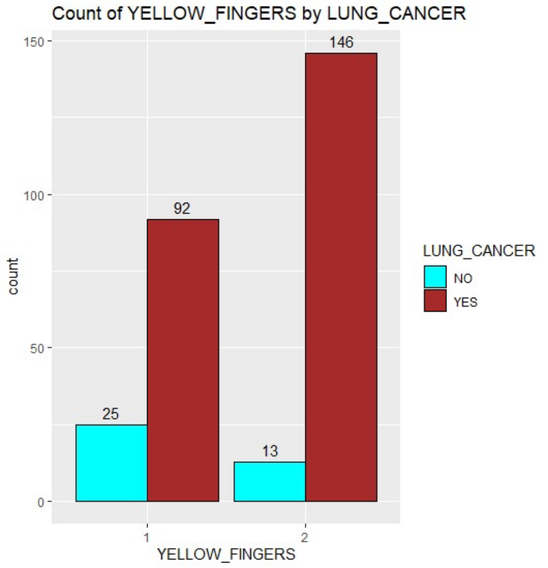

Count of ALLERGY by LUNG_CANCER
count
150
100
50
0
33
92
5
146
1
2
ALLERGY
LUNG_CANCER
NO
YES

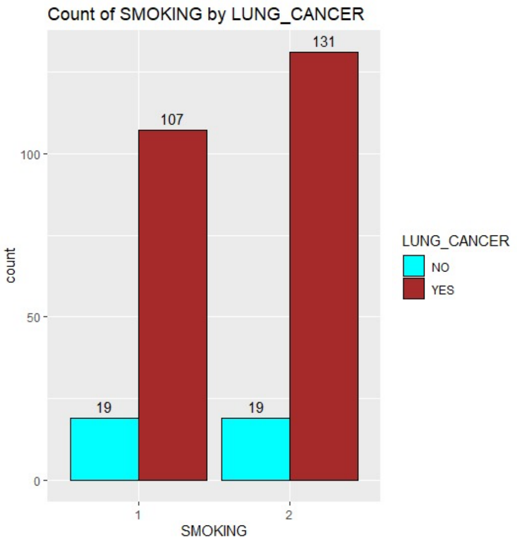

Count of FATIGUE by LUNG_CANCER
count
150
100
50
0
20
73
18
165
1
2
FATIGUE
LUNG_CANCER
NO
YES

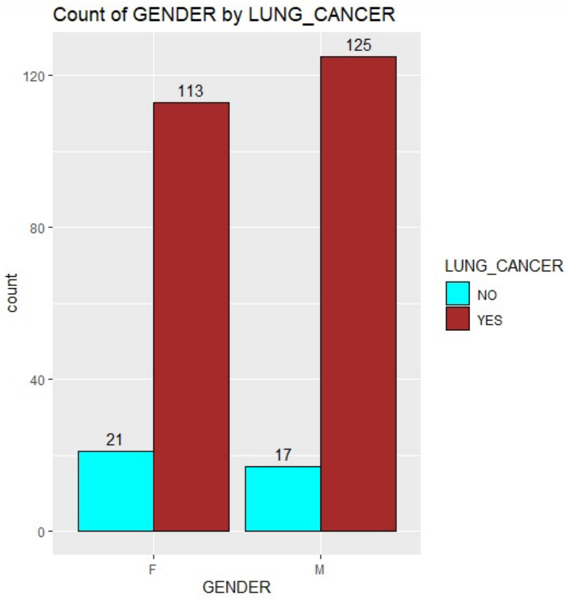

Count of CHRONIC.DISEASE by LUNG_CANCER
count
100
50
0
25
107
13
131
1
2
CHRONIC.DISEASE
LUNG_CANCER
NO
YES

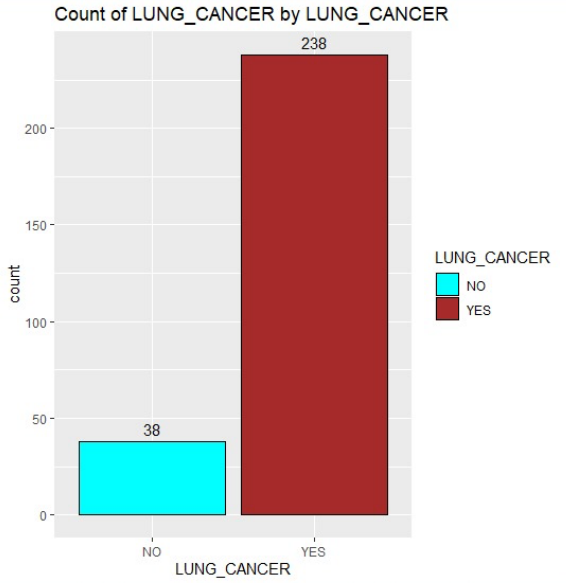

Count of PEER_PRESSURE by LUNG_CANCER
count
100
50
0
28
108
10
130
1
2
PEER_PRESSURE
LUNG_CANCER
NO
YES


Count of ANXIETY by LUNG_CANCER
count
120
80
40
0
26
113
12
125
1
2
ANXIETY
LUNG_CANCER
NO
YES


Count of YELLOW_FINGERS by LUNG_CANCER
count
150
100
50
0
25
92
13
146
1
2
YELLOW_FINGERS
LUNG_CANCER
NO
YES


Count of SMOKING by LUNG_CANCER
count
0
50
100
19
107
19
131
1
2
SMOKING
LUNG_CANCER
NO
YES


Count of GENDER by LUNG_CANCER
count
0
40
80
120
21
113
17
125
F
M
GENDER
LUNG_CANCER
NO
YES



As we can see above, there are multiple variables that held a huge responsible on causing a lung cancer. The graphs shown above covered all those multiple variables and connect that exact variable on the subjects based on their gender.

5. CONCLUSION

In conclusion, the analysis conducted in this paper has provided valuable insights into the multifaceted landscape of lung cancer, encompassing genetic, environmental, and clinical dimensions. The exploration of genetic predisposition revealed intricate relationships between inherited mutations, somatic alterations, and an individual's susceptibility to lung cancer. The interplay of environmental factors, extending beyond tobacco smoke to include pollutants and occupational exposures, underscored the complexity of risk factors contributing to the disease. Molecular heterogeneity emerged as a prominent theme, emphasizing the diverse genetic and molecular profiles inherent in lung cancer.

This diversity has implications for treatment response and underscores the need for personalized therapeutic strategies. The evaluation of existing diagnostic methods highlighted opportunities for enhancing early detection, a critical factor in improving patient outcomes. Immunotherapeutic approaches, particularly immune checkpoint inhibitors, demonstrated promising avenues for lung cancer treatment.

However, challenges, such as treatment resistance, were also elucidated, prompting further investigations into combination therapies and overcoming resistance mechanisms. The analysis also underscored the importance of addressing lifestyle factors in lung cancer prevention and management. Behavioral interventions and lifestyle modifications were identified as potential adjuncts to conventional treatments. Furthermore, the study shed light on disparities in lung cancer care, emphasizing the need for equitable access to resources, early detection programs, and personalized treatments across diverse populations. In summary, this analysis contributes to the ongoing discourse on lung cancer by providing a comprehensive understanding of its complexities.

The findings not only advance our knowledge of the disease but also offer practical recommendations for clinicians, researchers, and policymakers. Moving forward, continued research and a collaborative, multidisciplinary approach will be essential in tackling the challenges posed by lung cancer and improving outcomes for those affected by this pervasive and complex condition.