# Data Visualization and Analysis

Mohammad Asif Irfan Khan

## 1 Introduction

Lung cancer, or lung carcinoma, is the uncontrolled division of epithelial cells which line the respiratory tract. There are some signs of lung cancer we need to pay attention.

## 2 Motivation

Unfortunately, lung cancer is one of those cancers where the causes are known, and the cause is something that we do willingly and get cancer. Smoke is the major causes of lung cancer if we take all the patient of lung cancer nearly 80 to 85 per cent of lung cancer is caused by heavy smoking not to say that every smoker gets lung cancer. Apart from smoking, there are some other causes like and which coughing, chest pain, alcohol consumption etc. If we detect those cause and try to avoid some, lung cancer rate can fall.

By the help of r language, I conduct some analysis on lung cancer data.

## 3 Data

To conduct my analysis I use a dataset of 12 symptoms of Lung cancer. survey lung cancer data

(source:https://data.world/sta427ceyin/survey-lung-cancer)

## 4 Data Clean

In this dataset, there were some duplicate values. That's why I use a distinct function to remove those values.When we provided factor variable as part of the visualization it will do attractive visualization. That's why I convert some variable as a factor and I use factor variable to colour code our visualization.

# 5 Analysis

First, we look at the lung cancer rate. Along the x-axis, we have lung cancer variable, "YES" for those folks who have lung cancer and "NO" for those people who have not this cancer.

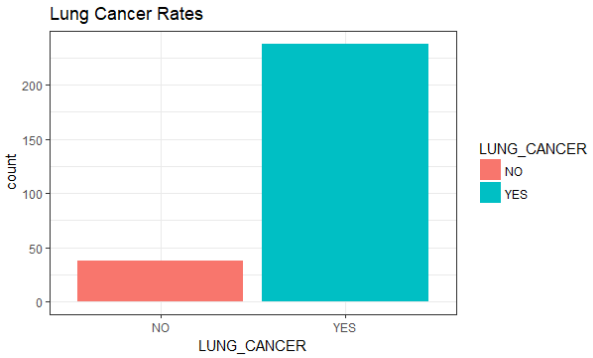


Figure 1: Lung cancer rates

From this figure 1, we saw that more people have lung cancer. This rate is alarming.

We have seen the lung cancer rates. Now we look at the lung cancer trends for different age people.

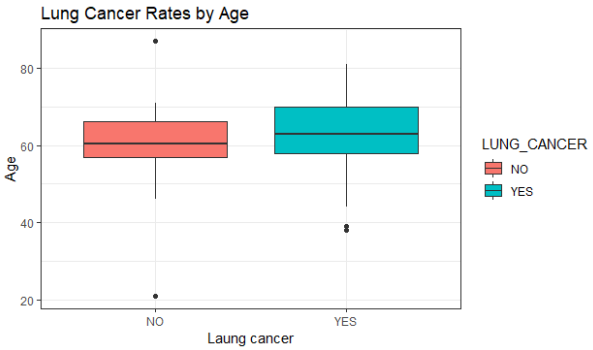


Figure 2: Lung cancer rates by Age

This box plot visualization tells us is in general people who have lung cancer is tended to be older than those who don't. It means older people have much more lung cancer probability.

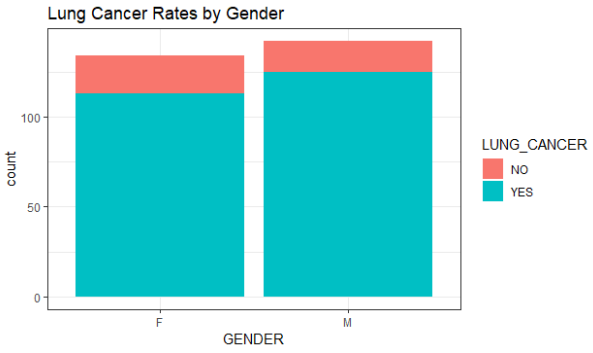


Figure 3: Lung cancer rates by Gender

In figure 3 we see that the lung cancer rate for a male is much more than female.

In figure number 4, we see the same thing that lung cancer rate is higher for older people than younger. And when we include the gender and smoke variable(Notice,1=who don't smoke,2=who do smoke), we see that female and male who doesn't smoke in middle age has the lower lung cancer probability.

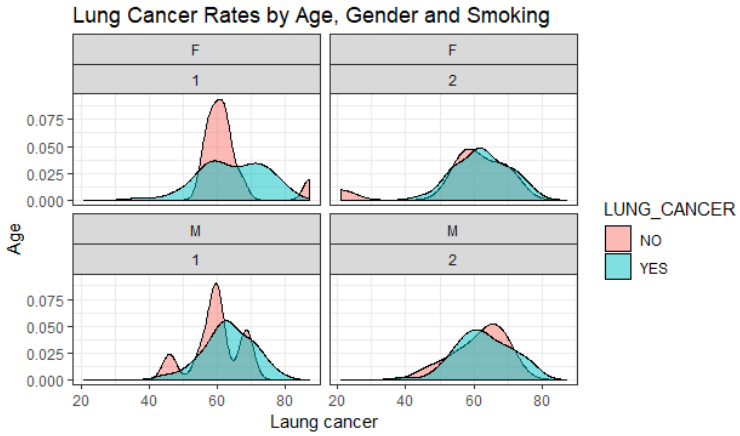


Figure 4: Lung Cancer Rates by Age, Gender and Smoking

Lung cancer depends on many causes. When these causes are bound together, then the probability of having lung cancer is increased.

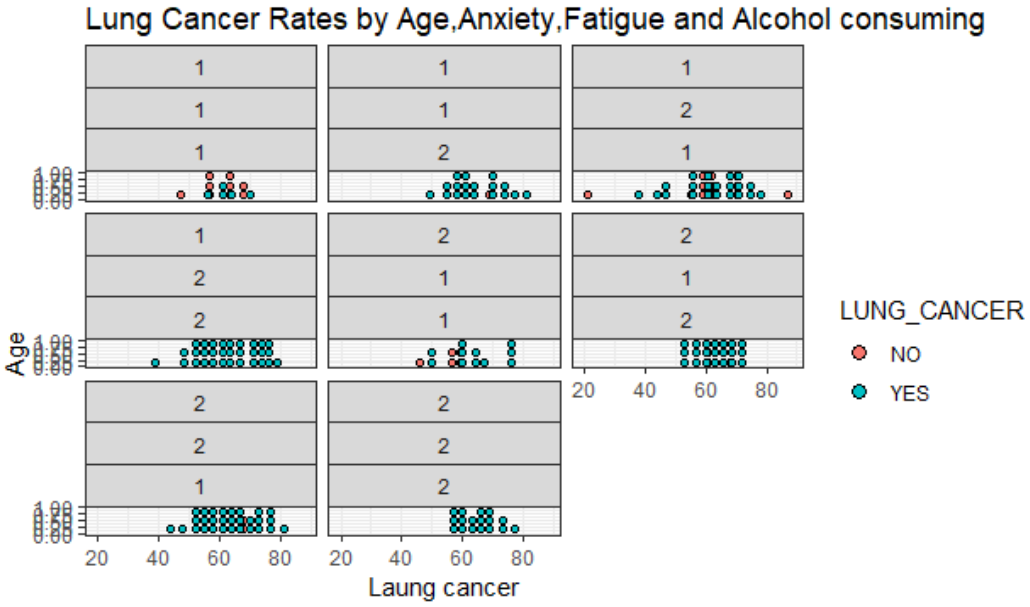


Figure 5: Lung Cancer Rates by Age, Anxiety, Fatigue and Alcohol consuming

There are some causes which dependent on man's behaviour, these causes lead to lung cancer. If we neglect these causes the cancer rate will drop.

In Figure 5, we saw that, a man who drinks alcohol, and feel anxiety and fatigue has almost 100 per cent lung cancer rate. On the other hand, the rate is different from those who don't have these issues. Once again, If we want to decrease lung cancer rate, we need to avoid these causes

lung cancer shows no symptom and it hard to notice it is in an early stage. Fortunately, there are some signs that indicate lung cancer and can help us in diagnosing before it too late. A persistent cough and swallowing difficulty is early stage sign of lung cancer. Now, we see the lung cancer rate of these sign.

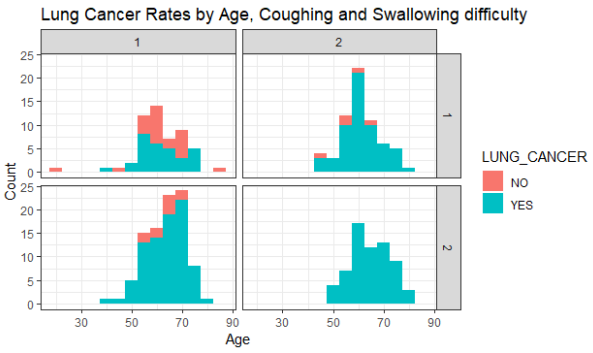


Figure 6: Lung Cancer Rates by Age, Coughing and Swallowing difficulty

In figure 6, we see that the people who have coughing and swallowing difficulty have 100 per cent lung caner probability.(Notice, 1=No, 2=Yes)

# 6 Statistical Test

Now we do some statistical test on lung cancer data.

First, we built a linear regression model. As I told earlier that lung cancer is not for one causes disease, there are several causes which are related to this cancer. That's why in our model, on an X axis we put all the causes and Y axis we put dependent variable LUNG CANCER.

```
Call:
lm(formula = LUNG_CANCER ~ +SMOKING + YELLOW_FINGERS + ANXIETY +
    PEER_PRESSURE + `CHRONIC DISEASE` + `SHORTNESS OF BREATH` +
    FATIGUE + ALLERGY + WHEEZING + `ALCOHOL CONSUMING` + COUGHING +
    `SHORTNESS OF BREATH` + `SWALLOWING DIFFICULTY` + `CHEST PAIN`,
    data = trainingdata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.69435 -0.11131  0.04909  0.14421  0.66588

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.168937   0.180337  -0.937  0.34997
SMOKING                  0.053557   0.037432   1.431  0.15401
YELLOW_FINGERS           0.098593   0.049275   2.001  0.04672 *
ANXIETY                  0.088917   0.052169   1.704  0.08982 .
PEER_PRESSURE            0.080184   0.041756   1.920  0.05620 .
`CHRONIC DISEASE`        0.115357   0.037237   3.098  0.00222 **
`SHORTNESS OF BREATH`    0.029640   0.044004   0.674  0.50134
FATIGUE                  0.226922   0.044384   5.113 7.24e-07 ***
ALLERGY                  0.122250   0.039822   3.070  0.00243 **
WHEEZING                 0.033940   0.041615   0.816  0.41569
`ALCOHOL CONSUMING`      0.267972   0.045207   5.928 1.28e-08 ***
COUGHING                 0.087132   0.043391   2.008  0.04594 *
`SWALLOWING DIFFICULTY`  0.112135   0.045278   2.477  0.01407 *
`CHEST PAIN`            -0.009079   0.039203  -0.232  0.81708
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2609 on 206 degrees of freedom
Multiple R-squared:  0.443,     Adjusted R-squared:  0.4079
F-statistic:  12.6 on 13 and 206 DF,  p-value: < 2.2e-16
```

Figure 7: Linear Regression Model

The summary statistics above tells us a number of things. First p-value. The p-value is a measure of the strength of the evidence against the null hypothesis. We can judge a linear model to be statistically significant only when both these p-Values are less than the pre-determined statistical significance level, here we saw that p-value is less than significant level 0.05.

We know that higher t-value is better, A larger t-value indicates coefficient is not equal to zero.In our model, we see that independent variable 'ALCOHOL CONSUMING', FATIGUE, CHRONIC DISEASE, ALLERGY, COUGHING, YELLOW FINGERS and SWALLOWING DIFFICULTY has higher t-value than other variables. It means lung cancer disease probability rate depends on these variable or causes.

To tell how this model will perform with new data I also build a predicting linear model. First, split the data where 80 per cent training data and 20 per cent testing data. Then build the model on the training data and use the model to predict the dependent variable on test data.

From predicting model review, the model p-value is less than the significant level for those variable. And also, the t-value is higher.

To evaluate our linear regression test I compare this with anova test. The result is quite similar for both tests.

# 7 Conclusion

The leading cause of lung cancer is alcohol consumption, fatigue, chronic disease, allergy coughing, yellow fingers and swallowing difficulty. Through the data exploratory data analysis, we also see that despite these causes other causes are also responsible for lung cancer.

Lung cancer is a kind of disease which we can easily overcome by our self. We just need to avoid those causes and lead a happy life.