**Lung Cancer Analysis**

Name - Venkata Durga Avinash Bobbala

Student id: 110292912

**Introduction:**

The leading cause of high mortality in the modern world is considered to be lung cancer, which is also the worst illness. Lung cancer has a bigger impact on people than other types of cancer, and as predicted, it currently occupies the seventh spot in the death rate index, accounting for 1.5% of global mortality. Lung cancer begins in the lung and spreads to the brain.

Lung cancer is divided into two main categories. Small cell lung cancer and non-small cell lung cancer are the two types. Patients may have symptoms including acute chest discomfort, a dry cough, shortness of breath, weight loss, etc. When considering the development of cancer and its causes, physicians place more emphasis on smoking and second-hand smoke as though they were the main culprits.

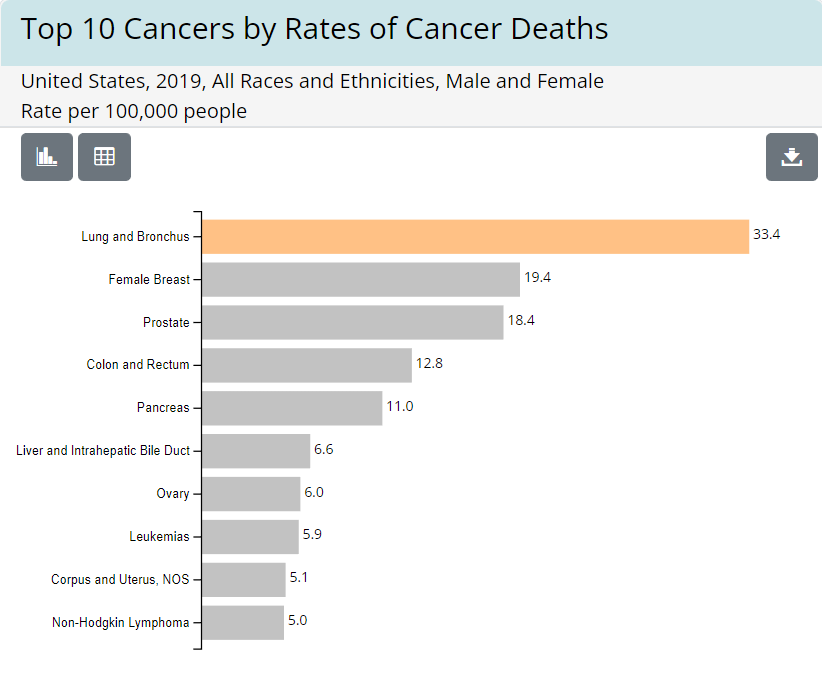
Lung cancer typically affects both men and women due to the lungs' unchecked cell proliferation. This leads to a major breathing issue in the chest's inhale and exhale regions. Compared to other malignancies, the death rate from lung cancer is rising steadily in both young and old people. The mortality rate is still not well under control despite the availability of high-tech medical facilities for thorough diagnosis and efficient medical treatment. Therefore, it is important to adopt preventative measures at the outset so that the disease's symptoms and effects may be identified early on for better diagnosis.

Fig 1: Top 10 Cancers by rates of cancer deaths

From fig.1 we can observe that Lung and Bronchus cancer is ranked no.1 when it comes to cancer deaths.

Surgery, chemotherapy, radiation therapy, immune therapy, and other treatments are used to treat lung cancer. Despite this, the diagnosis of lung cancer is quite poor since doctors can only detect the disease when it has advanced. To easily reduce the mortality rate with efficient control, early prevention before the last stage is crucial.

**Question definition:**

The main aim of this project is to better understand the things or habits to avoid to prevent suffering from lung cancer. Also, know which factors tend to have a higher risk of suffering from lung cancer. As said the treatment for lung cancer is pretty costly and time taking and cannot guarantee to save one’s life. So, preventing it by changing a few habits would be great. **An ounce of prevention is worth a pound of cure**.

**Methodology:**

**Data description:**

The data is borrowed from Kaggle. It contains 25 columns and 1000s of patient records. It consists of features like patient id, Age, Gender, Air Pollution that the patient is exposed to Alcohol use, Dust Allery, Genetic risk, Smoker, Obesity, Chest pain, coughing of blood, etc., and these are rated with a scale 1 – 9 and we had the output column of the level of cancer – Low, Medium, High. I would say it is a balanced dataset with 37% high, 33% of medium, and 30% of Low cancer patient data.

**Exploration:**

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Patient Id 1000 non-null object

1 Age 1000 non-null object

2 Gender 1000 non-null object

3 Air Pollution 1000 non-null object

4 Alcohol use 1000 non-null object

5 Dust Allergy 1000 non-null object

6 OccuPational Hazards 1000 non-null object

7 Genetic Risk 1000 non-null object

8 chronic Lung Disease 1000 non-null object

9 Balanced Diet 1000 non-null object

10 Obesity 1000 non-null object

11 Smoking 1000 non-null object

12 Passive Smoker 1000 non-null object

13 Chest Pain 1000 non-null object

14 Coughing of Blood 1000 non-null object

15 Fatigue 1000 non-null object

16 Weight Loss 1000 non-null object

17 Shortness of Breath 1000 non-null object

18 Wheezing 1000 non-null object

19 Swallowing Difficulty 1000 non-null object

20 Clubbing of Finger Nails 1000 non-null object

21 Frequent Cold 1000 non-null object

22 Dry Cough 1000 non-null object

23 Snoring 1000 non-null object

24 Level 1000 non-null object

There are no missing data or duplicate data in this dataset. This dataset is pretty clean and has the required columns for the analysis of what might majorly cause lung cancer.

**Unique values of each column:**

1. Age labels: [33 17 35 37 46 52 28 44 64 39 34 27 73 36 14 24 53 62 29 65 38 19 42 32 25 45 26 48 22 18 23 47 61 55 31 49 43 51 54 63] - Ages range from 14 – 73
2. Gender labels: [1 2] – 1 represents Male and 2 represents Female

**Below columns, values range from 1 – 7 or 8 or 9 (levels of habit or symptoms)**

1. Air Pollution labels: [2 3 4 7 6 5 1 8]
2. Alcohol use labels: [4 1 5 7 8 3 6 2]
3. Dust Allergy labels: [5 6 7 4 2 8 1 3]
4. Occupational Hazards labels: [4 3 5 7 2 6 8 1]
5. Genetic Risk labels: [3 4 5 6 7 2 1]
6. chronic lung disease labels: [2 4 7 6 3 5 1]
7. Balanced Diet labels: [2 6 7 4 5 3 1]
8. Obesity labels: [4 2 7 3 5 6 1]
9. Smoking labels: [3 2 7 8 1 6 5 4]
10. Passive Smoker labels: [2 4 3 7 6 8 5 1]
11. Chest Pain labels: [2 4 7 3 6 5 9 8 1]
12. Coughing of Blood labels: [4 3 8 9 1 5 7 6 2]
13. Fatigue labels: [3 1 8 4 5 9 2 6]
14. Weight Loss labels: [4 3 7 2 6 5 1 8]
15. Shortness of Breath labels: [2 7 9 3 4 5 6 1]
16. Wheezing labels: [2 8 1 4 6 7 5 3]
17. Swallowing Difficulty labels: [3 6 1 4 2 5 8 7]
18. Clubbing of Finger Nails labels: [1 2 4 5 6 8 7 9 3]
19. Frequent Cold labels: [2 1 6 4 3 7 5]
20. Dry Cough labels: [3 7 2 4 1 5 6]
21. Snoring labels: [4 2 5 3 1 6 7]
22. Level labels: ['Low' 'Medium' 'High'] – level of lung cancer

The above columns are generally the habits or symptoms of a person which could be related to or symptoms of lung cancer. For example, we assume smoking is the major cause of lung cancer but many factors make a person prone to lung cancer. Many factors need to be considered such as pollution, genetics, passive smoking, alcohol consumption, shortness of breath, obesity, diet, and many more.

The story of which habits were dangerous and many more insights are going to be described below with the visualizations comparing different columns with each other and with the level of lung cancer.

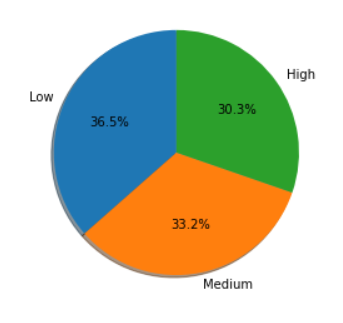
**Insights:**

Fig 1: Distribution of Cancer Level

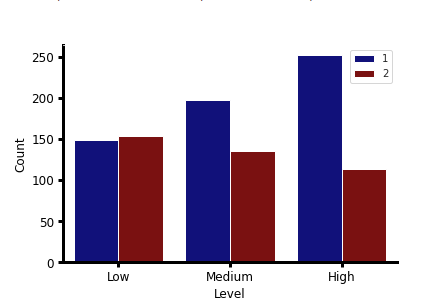
The above figure describes that the cancer levels are well-balanced. There are 36.5% of low-level cancer patients, 33.2% of medium-level cancer patients, and 30.3% of high-level cancer patients.

Fig 2: Cancer levels and gender

In the above figure, we can observe that Male (1) has more chance than female (2) to get affected by a medium or high level of cancer.

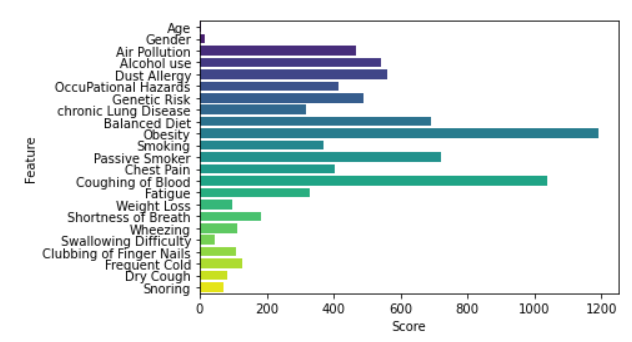


Fig 3: Scores for the features

From the above figure 3, it is inferred that Age, Gender, Weight loss, Shortness of breath, Wheezing, Swallowing Difficulty, Clubbing of Fingernails, Frequent colds, Dry cough, and Snoring are not the most important features or symptoms to determine the risk of lung cancer.

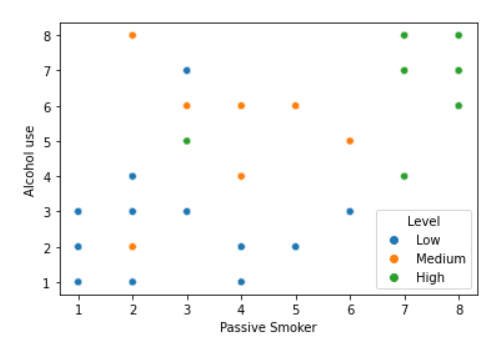


Fig 4: Passive smoker vs Alcohol use with Level of cancer

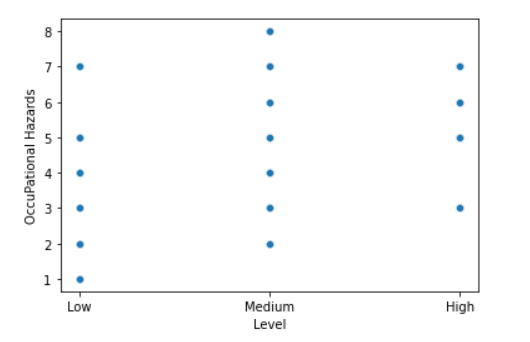
From the above figure, we can observe that people who are highly passive smokers and drinks too much alcohol are having a high chance of getting lung cancer.

Fig 5: Occupational Hazards vs Cancer level

Occupational hazards don’t directly relate to or cause lung cancer. This is what we can understand by seeing the above figure.

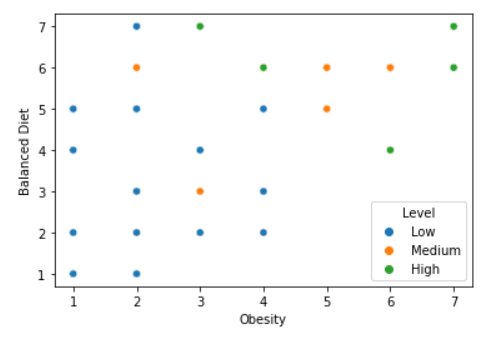


Fig 6: Obesity vs Balanced diet broke into levels of cancer.

From the above figure, we can observe that eating a balanced diet and lower obesity value tends to lower the risk of lung cancer.

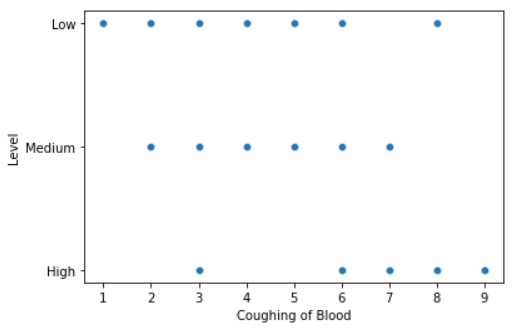


Fig 7: Level of cancer vs Coughing Blood

In general, people coughing blood seems to have cancer so thought to visualize and see whether it is true. A high value of coughing blood may result in a high risk of lung cancer but low to medium levels have almost the same values as coughing blood.

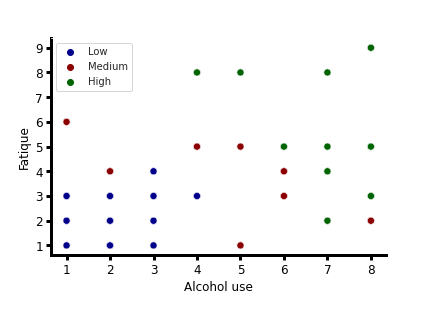


Fig 8: Alcohol use vs Fatigue broke into levels of cancer

We can observe that lower values of Alcohol use, Fatigue tends to lower the risk of cancer, medium values of Alcohol use and fatigue tend to medium risk of cancer and higher values tends to a higher risk of lung cancer. These two would be the main causes of lung cancer.

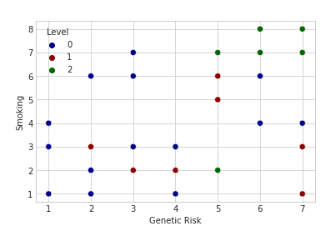


Fig 9: Genetic risk vs Smoking broke into levels of cancer.

From the above figure, we can infer that the lower and medium values of smoking and genetic risk are randomly associated with lower and medium risk of cancer but the extremely high values of both features tend to a higher risk of lung cancer.

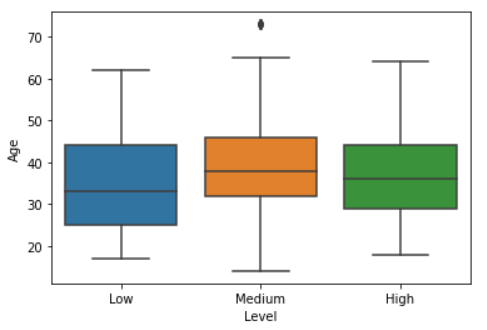


Fig 10: Box plot for Age and levels

From the above figure, we can infer that age is not important to determine whether a person has a low or medium, or high risk of cancer. Also, we can see that an average aged 30-40 has a low, medium, and high risk of cancer.

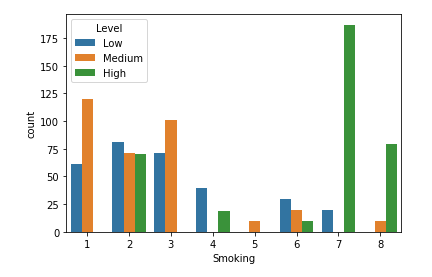


Fig 11: Smokers count broke on level of cancer.

From figure 11 we can observe that people smoking more has a greater chance of falling a higher risk of cancer. Whereas people who smoke less may be prone to lower or medium levels of cancer.

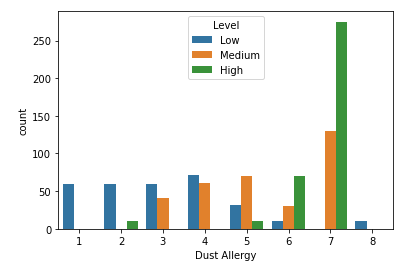


Fig 12: Dust Allergy count broke on level of cancer

It is inferred that lower levels of dust allergy tend to lower the risk of lung cancer, and Dust allergy 7 is most probable of having a higher risk of lung cancer.

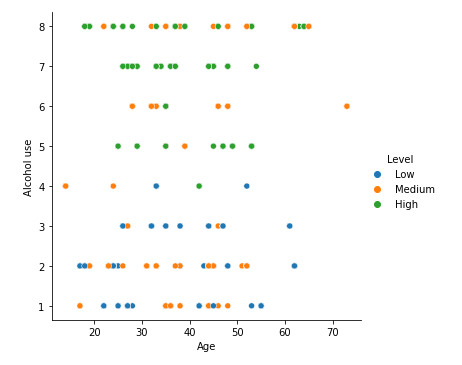


Fig 13: Alcohol use and Age broke into levels of cancer

From fig 13 we can infer that Alcohol use high irrespective of age tends to suffer a high risk of lung cancer. Less consumption of alcohol is represented by a medium or lower risk of lung cancer.

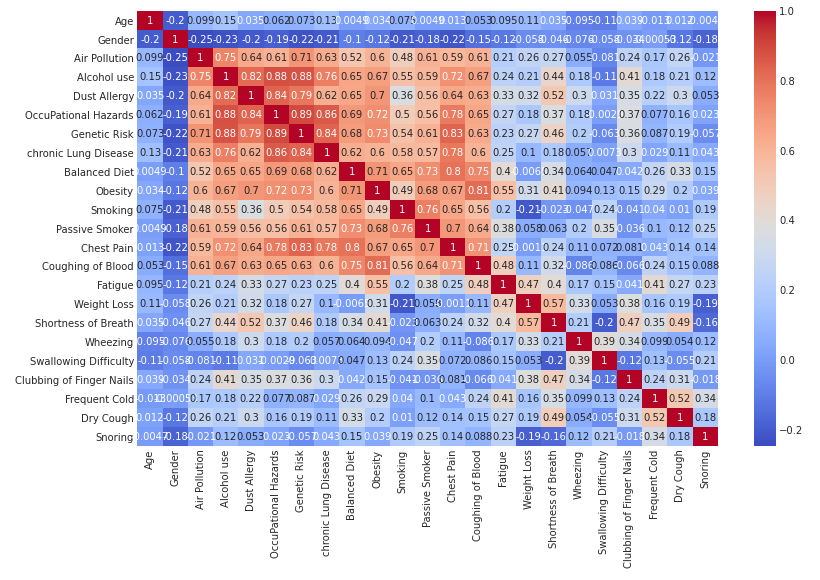


Fig 14: Correlation Matrix

**Conclusion:**

Lastly, we can conclude that different age groups and even youth were attracted to habits like overconsumption of alcohol, smoking, bad diet habits, and no proper exercise tends to a higher risk of health problems majorly very dangerous lung cancer and 1.5% of deaths are due to lung cancer today which is way greater than contagious virus covid-19. We feared and got scared a lot about covid-19 but cancer is way more dangerous than covid-19.

Cancer can be prevented by following healthy habits like eating on time. Consuming healthy food, reducing consumption of alcohol, preventing smoking, and avoiding being a passive smoker. Passive smoking is very dangerous compared with the normal habit of smoking.