Lung Cancer Prediction Exploratory Data Analysis

Name: smahmood@bellarmine.edu

ABSTRACT
Lung cancer is a type of cancer that begins in the lungs and spreads to other organs of the body. It is the second most common form of cancer worldwide with at least 2.2 million cases reported per year. Thus, it is crucial to identify factors that can possibly predict the likelihood of developing lung cancer. This exploratory data analysis will look into the possible predictors of lung cancer given various factors and symptoms collected from patients. Additionally, we will analyze what factors and symptoms are most likely associated with lung cancer in hopes of identifying whether a patient has a higher or lower likelihood of developing lung cancer.

I. INTRODUCTION

This dataset was found on Kaggle. I'm trying to predict the likelihood of a patient developing lung cancer given different factors and symptoms provided by the patient. Additionally, I will be identifying what factors and symptoms are highly correlated with lung cancer. Moreover, I hypothesize that smoking could be a leading contributing factor to developing lung cancer so I would want to see how well it correlates with the level.

The various factors collected by patients include their age and gender. As well as their level of air pollution exposure, level of alcohol use, level of dust allergy, level of occupational hazards, level of genetic risk, level of chronic lung disease, level of a balanced diet, level of obesity, level of smoking, level of passive smoking. The various symptoms collected from patients include chest pain, coughing of blood, fatigue, weight loss, shortness of breath, wheezing, swallowing difficulty, clubbing of fingernails, and snoring.

II. BACKGROUND

Lung cancer comes in two forms, one is non-small cell lung cancer while the other form is small-cell lung cancer. According to the American Lung Association, the leading causes of lung cancer have been found to be smoking, second-hand smoke, hazardous chemicals, particle pollution, and genetics. However, there are countless studies which have looked into other possible causes of lung cancer. Moreover, the American Lung Association has found that the most common symptoms of lung cancer include fatigue, shortness of breath, coughing of blood, weight loss, wheezing etc. (*American Lung Association. 2022*). The World Cancer Research Fund Institute has found that lung cancer is the most common type of cancer for men and the second most common type for women (*World Cancer Research Fund International*). Thus, studying this disease and its potential predictors is still critical.

The authors of this dataset referenced a new study which found that air pollution could possibly be linked to developing lung cancer even if the individual is a nonsmoker. This study was published by the Nature Medicine Journal which looked into the data collected from 462,000 people in China over a six-year period. The population was divided into two groups. One group was comprised of individuals who lived in areas with high air pollution while the other group was made up of individuals living in areas with low air pollution.

The researchers of this study found that the high air pollution group were more likely to develop lung cancer compared to the low air pollution group. Additionally, the likelihood of developing lung cancer increased with age. Overall, the study implied there could possibly be a link between air pollution and lung cancer though it is important to note that the study is not suggesting that air pollution may cause lung cancer. The author wanted to confirm the results of this study and identify the effect of different levels of air pollution in relation to the risk of lung cancer.

III. EXPLORATORY ANALYSIS

The dataset contains 1000 rows and 26 columns with mostly integer data types and two object data types. There were no missing values within the dataset which can be referenced in **Table 1**.

The Y variable for this dataset was the "Level" variable which refers to the likelihood or chance of developing lung cancer. In terms of the level of lung cancer, 30.3% were low, 33.2% were medium, and 36.5% were high. This can be visualized in the figures labeled **Figure 1** and **Figure 2** respectively.

In terms of data distribution, the distribution of the possible factors, it seems there is a higher frequency of data within a higher-level range for chronic lung cancer, dust allergy, balanced, air pollution, alcohol use, occupational hazards, genetic risk, obesity, smoking, and passive smoking. This can be referenced in **Figure 3.** The distributions of the possible symptoms for lung cancer be seen in **Figure 4**.

Table 1: Data Types

Variable Name	*Data Type*	*Missing Data (%)*
index	int64, ratio	0%
Patient Id	Object, nominal	0%
Age	int64, ratio	0%
Gender	int64, ratio	0%
Air Pollution	int64, ordinal	0%
Alcohol use	int64, ordinal	0%
Dust Allergy	int64, ordinal	0%
OccuPational Hazards	int64, ordinal	0%
Genetic Risk	int64, ordinal	0%
chronic Lung Disease	int64, ordinal	0%
Balanced Diet	int64, ordinal	0%
Obesity	int64, ordinal	0%
Smoking	int64, ordinal	0%
Passive Smoker	int64, ordinal	0%
Chest Pain	int64, ordinal	0%
Coughing of Blood	int64, ordinal	0%
Fatigue	int64, ordinal	0%
Weight Loss	int64, ordinal	0%
Shortness of Breath	int64, ordinal	0%
Wheezing	int64, ordinal	0%
Swallowing Difficulty	int64, ordinal	0%
Clubbing of Finger Nails	int64, ordinal	0%
Frequent Cold	int64, ordinal	0%
Dry Cough	int64, ordinal	0%
Snoring	int64, ordinal	0%
Level	Object, ordinal	0%

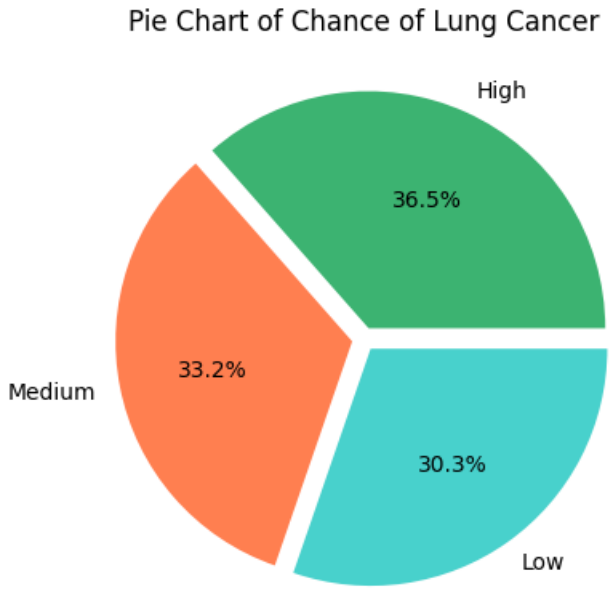

Figure 1. shows the Pie chart of the different levels for the chance of developing lung cancer.

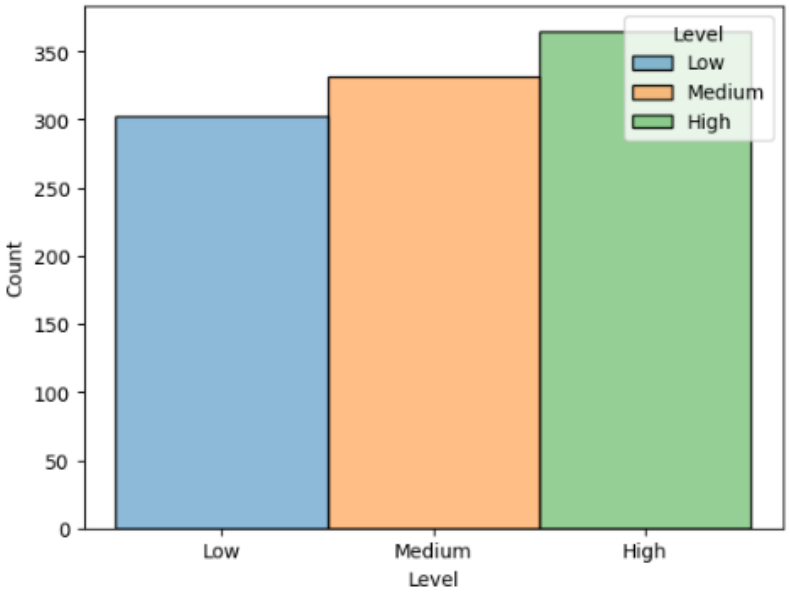

Figure 2. Shows the histogram of the different levels for the chance of developing lung cancer.

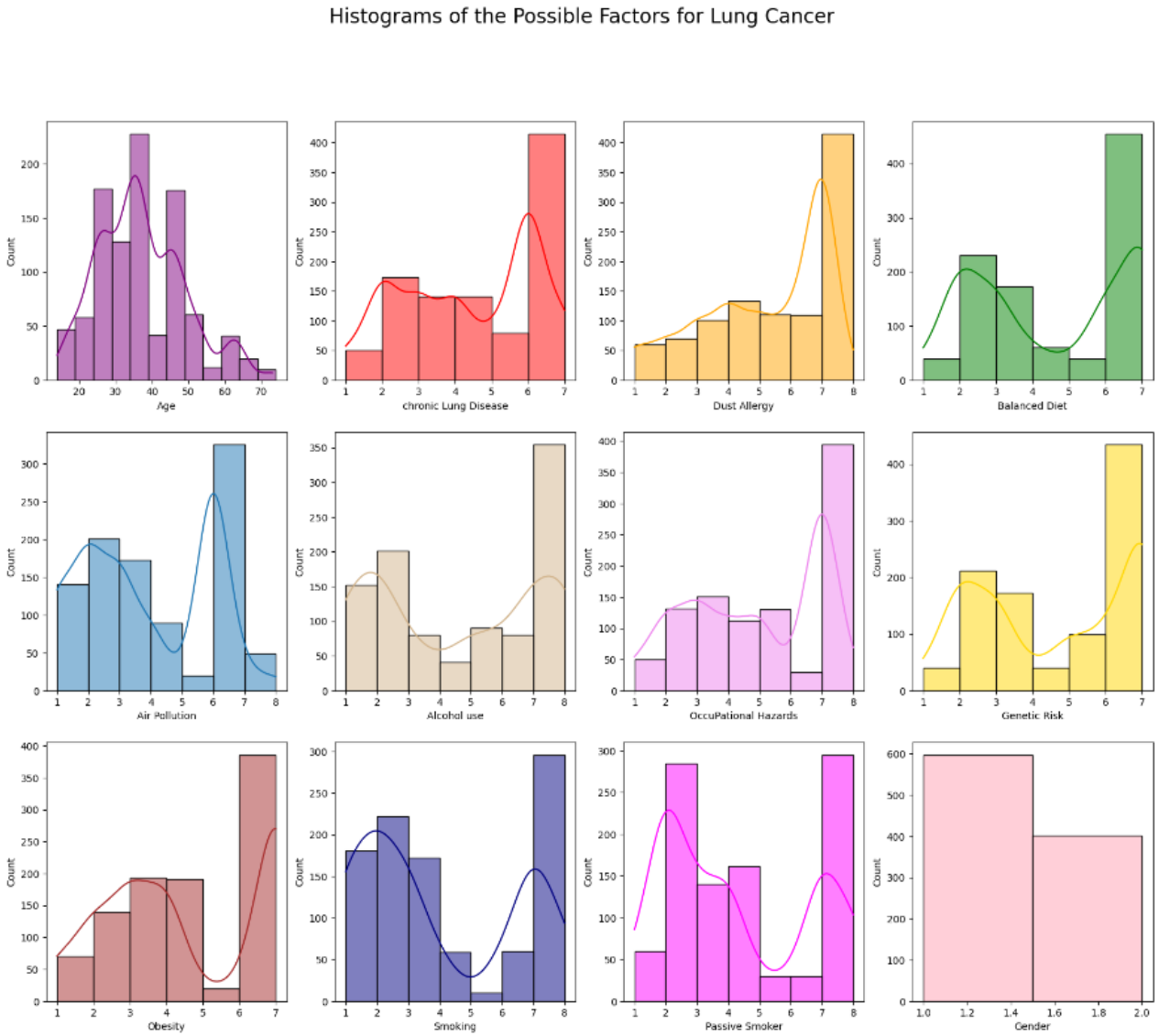

Figure 3. Shows the histograms of the possible factors for lung cancer.

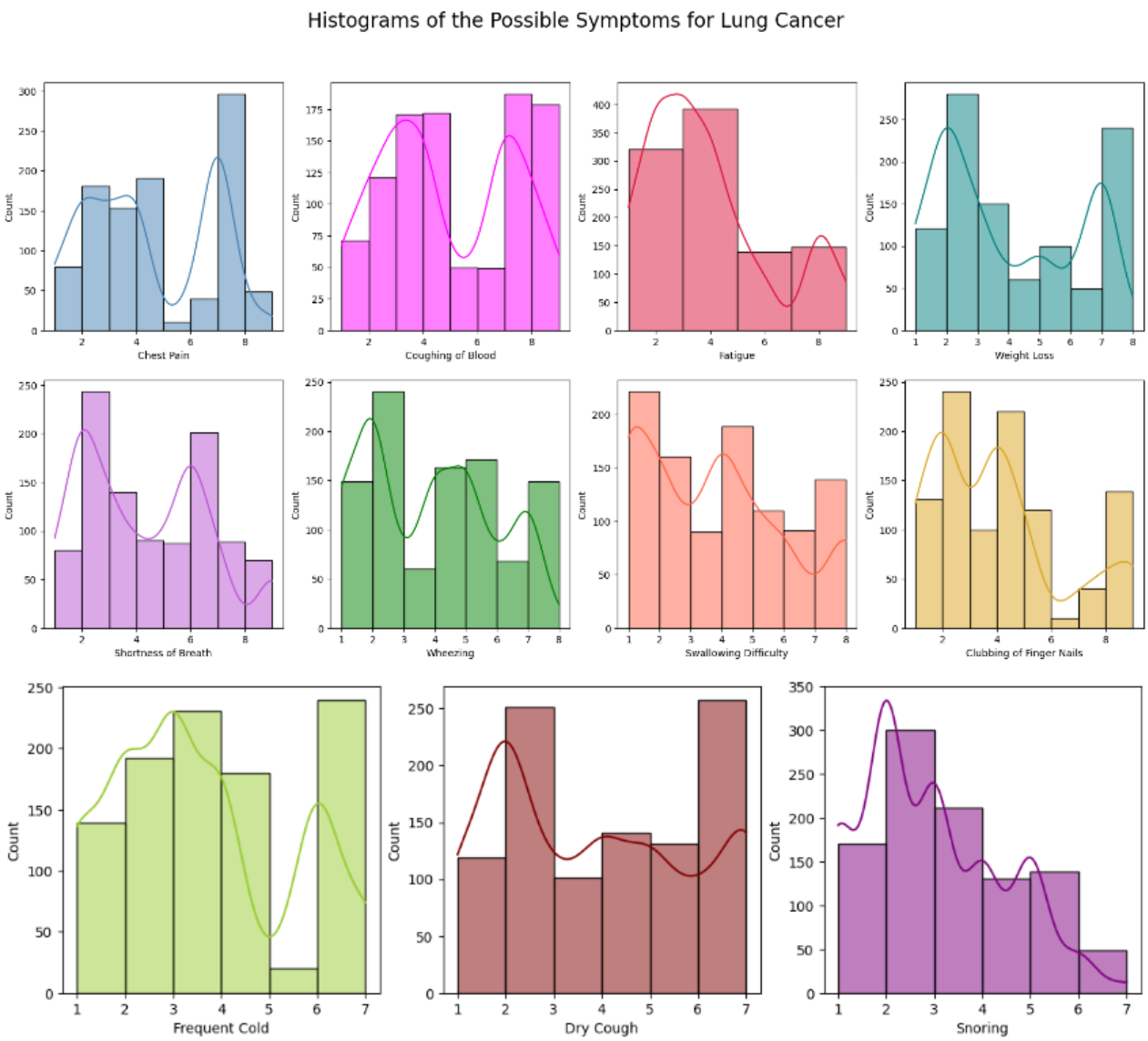

Figure 4. Shows the histograms of the possible symptoms for lung cancer.

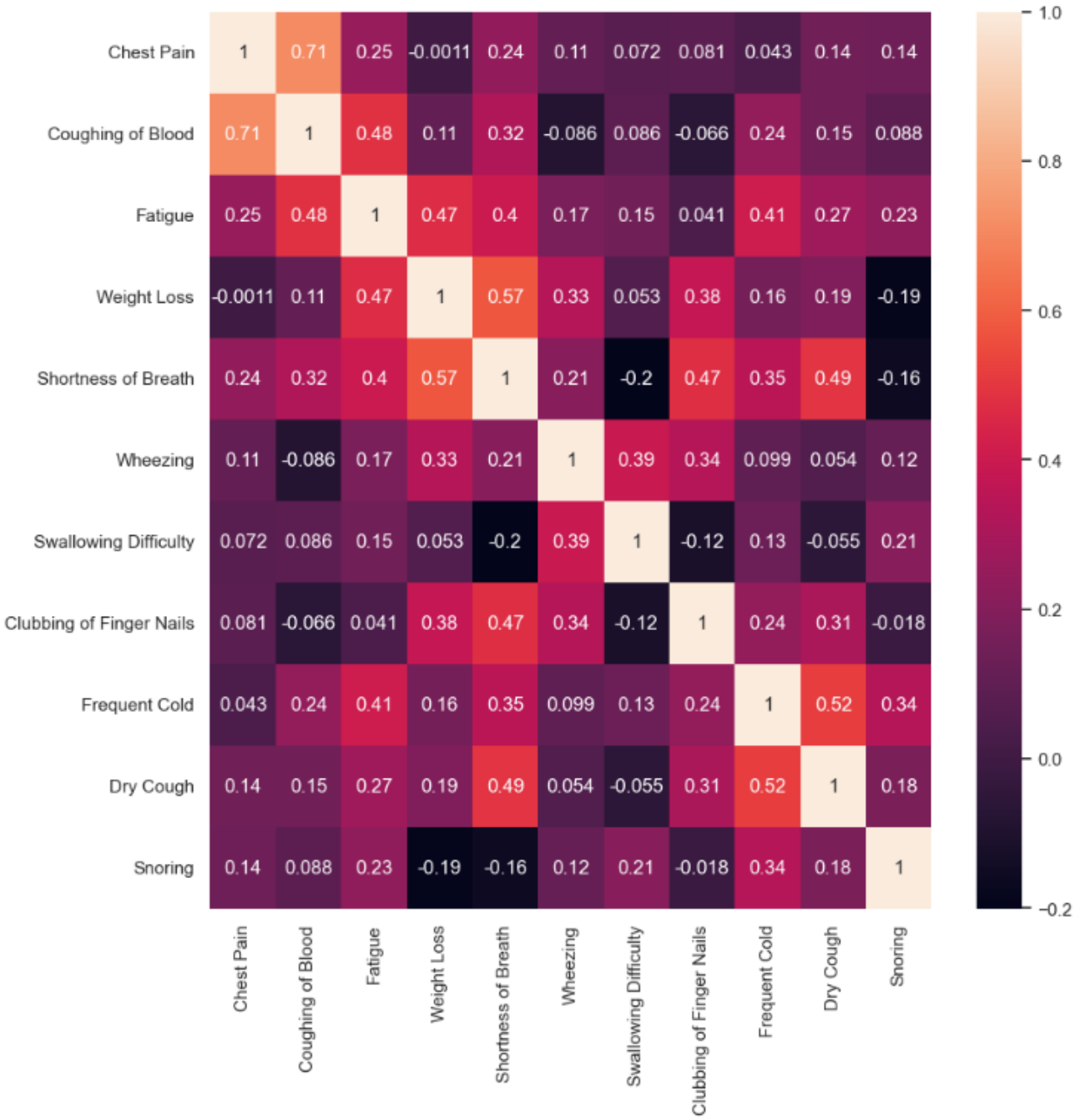

Figure 5. Shows the heatmap for the possible symptoms for lung cancer.

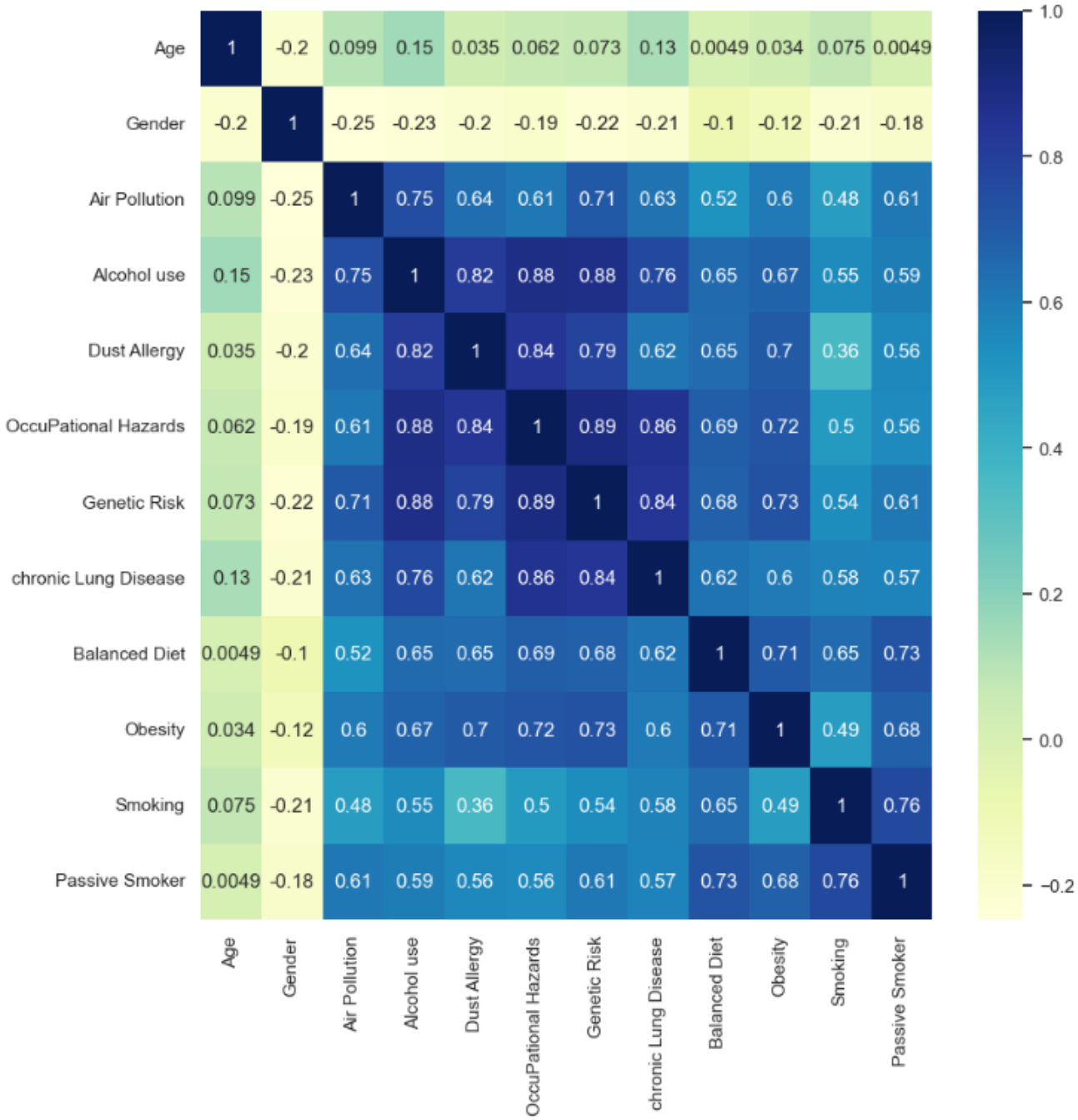

Figure 6. Shows the heatmap for the possible factors for lung cancer.

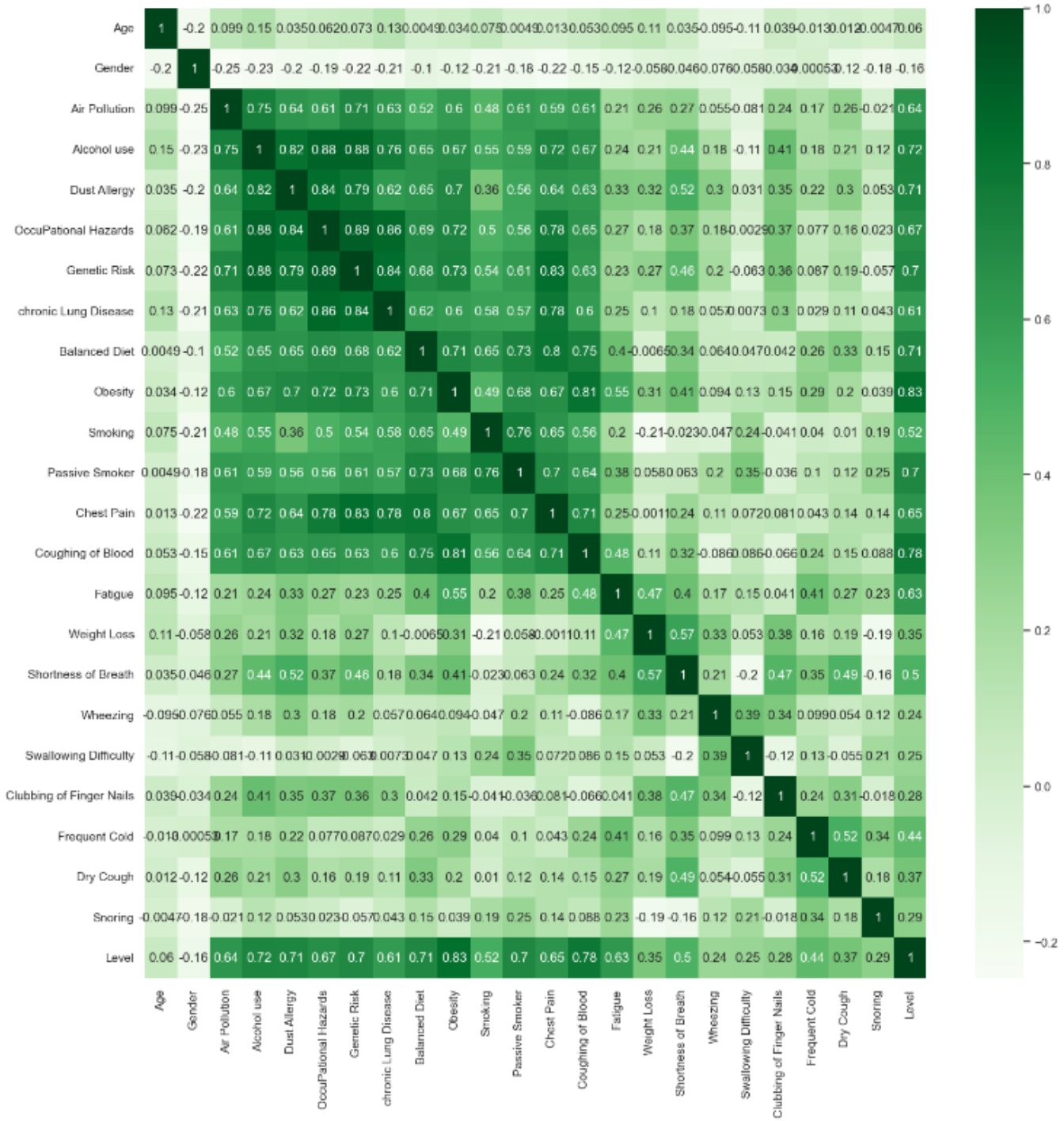

Figure 7. Shows the correlation matrix for the entire dataset including the level.

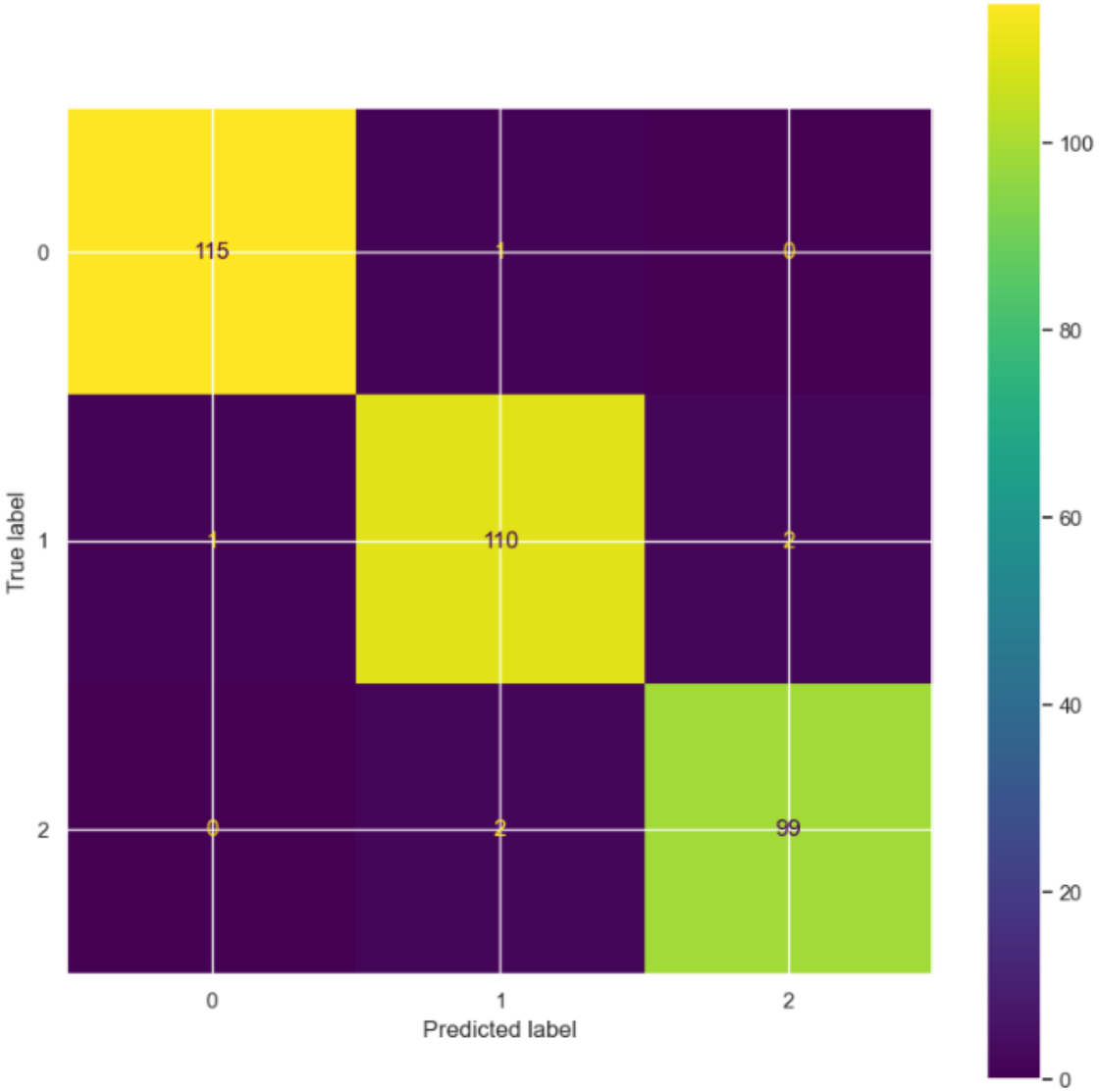

Figure 7. Shows the confusion matrix for the logistic regression model for experiment 1.

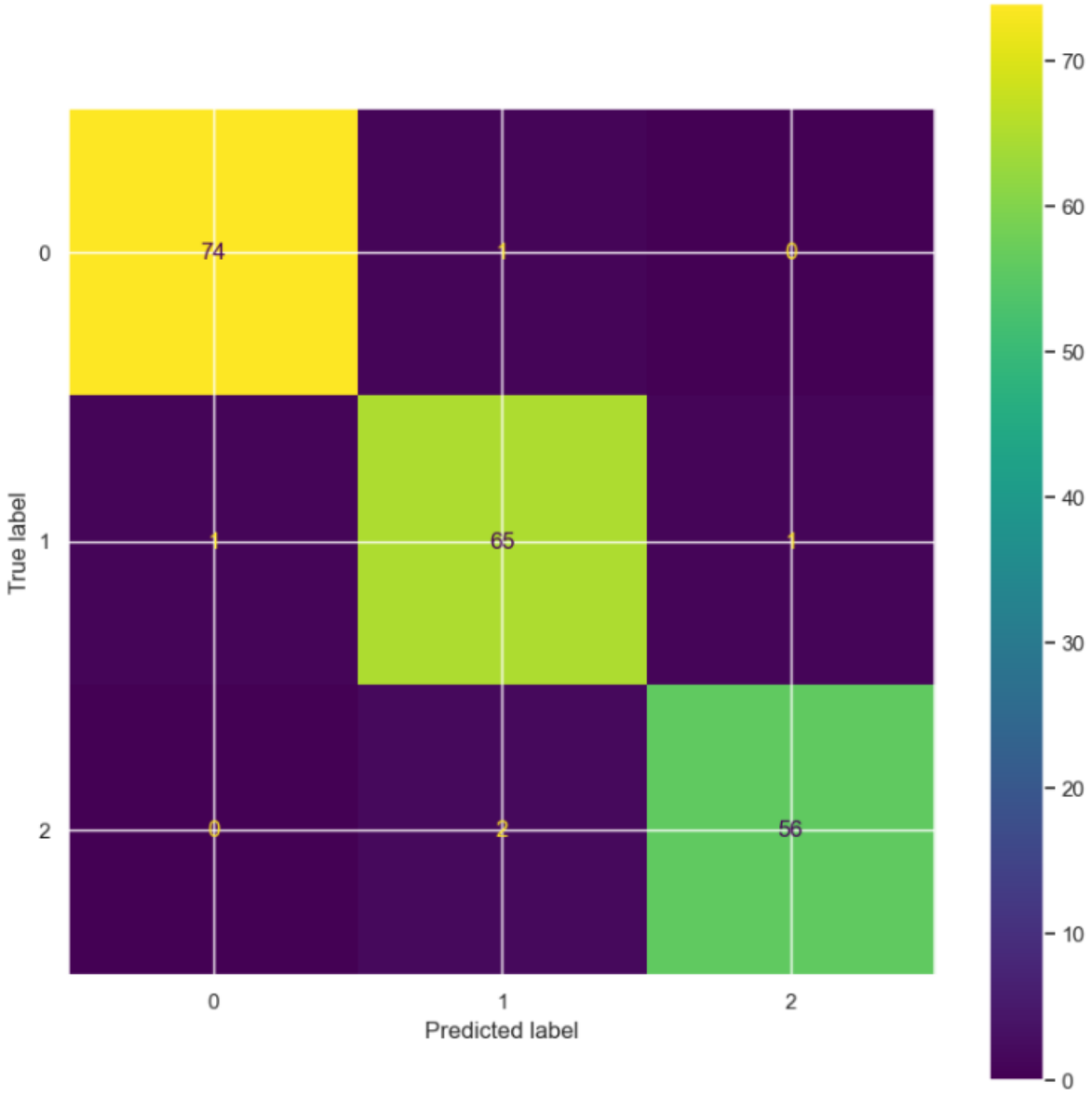

Figure 8. Shows the confusion matrix for the logistic regression model for experiment 2.

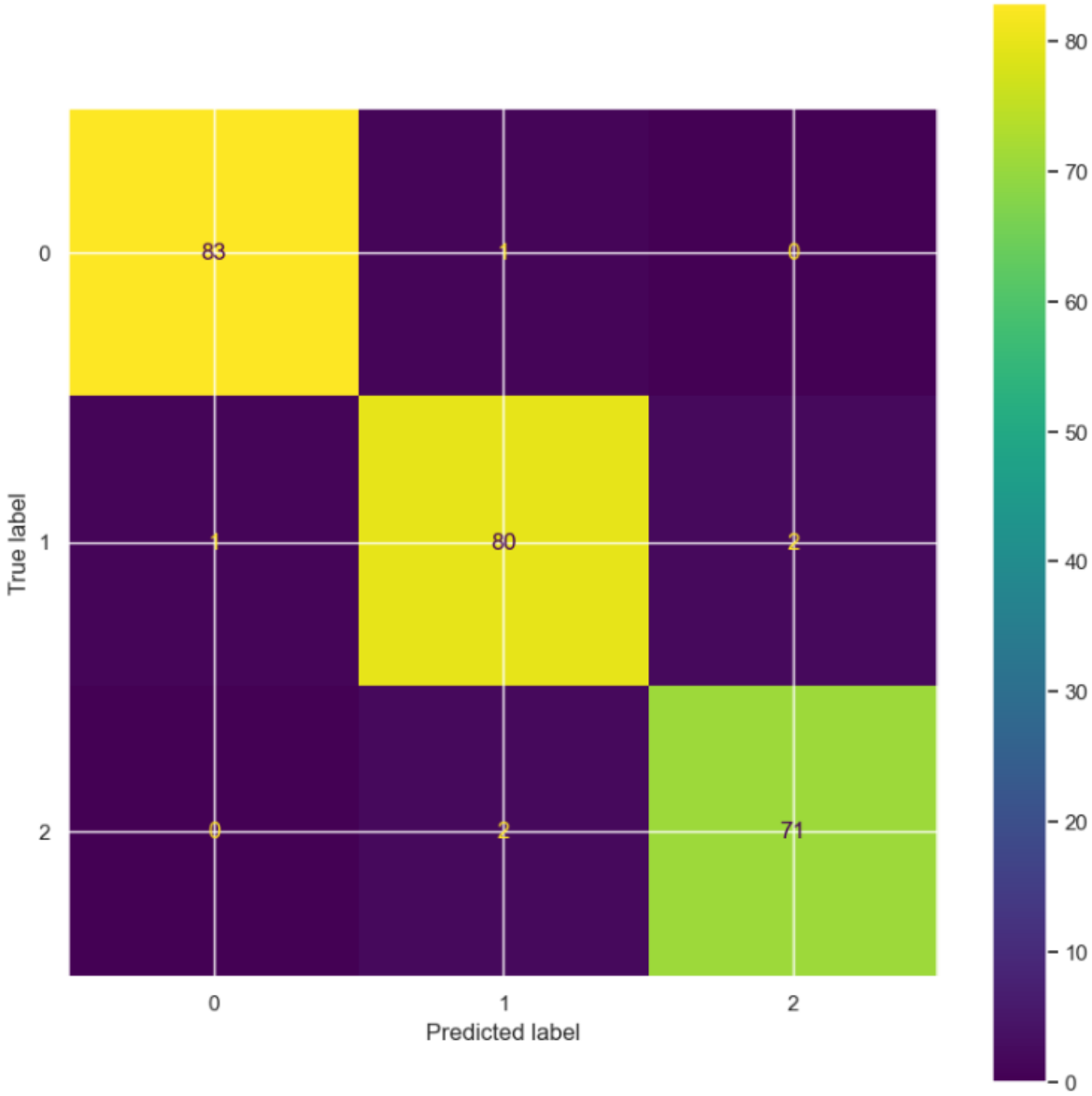

Figure 9. Shows the confusion matrix for the logistic regression model for experiment 3.

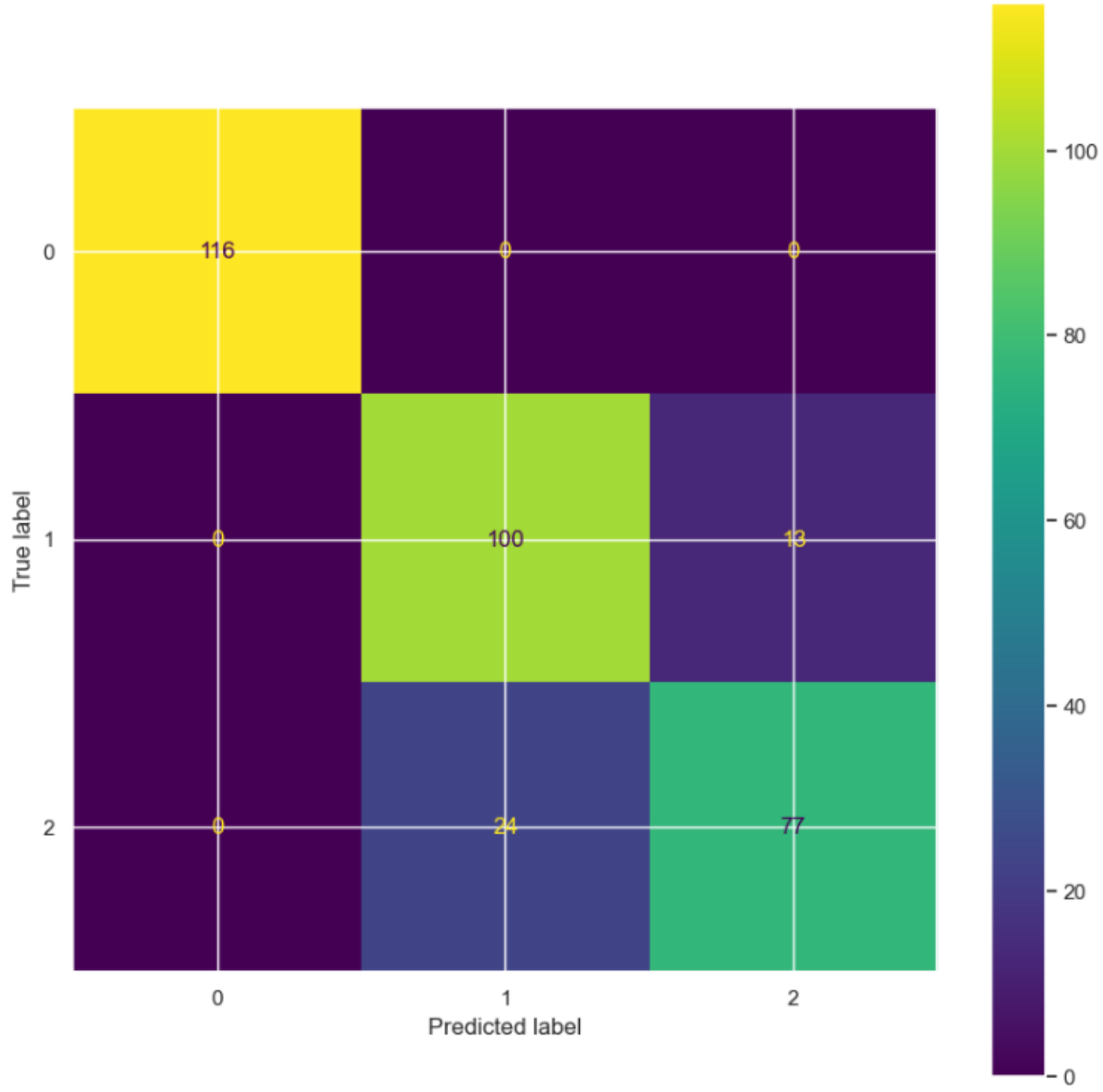

Figure 10. Shows the confusion matrix for the logistic regression model for experiment 4.

IV. METHODS

A. Data Preparation

In terms of preparing the data, first, I loaded and viewed the dataset. Then I checked for null values. From there I deleted two columns. “Index” and “Patient Id” since it was not needed for this data analysis.

For looking at the correlations between different variables in the dataset, I created a third heatmap which served as a correlation matrix. For this, I wanted to see which variables correlated the most with my Y variable “Level.” Through this, I could identify which factors and symptoms would be the most associated with my Y variable “Level.” To do this, I made a copy of my original dataset and gave integer values 0, 1, and 2 to the various levels low, medium, and high respectively.

For my logistic regression models, I set my dependent variable as “Level” and my independent variable as all the other variables within the dataset for my first three experiments. Since the “Level” variable was categorical, I used a label encoder which encoded the various levels into integer values 0, 1, and 2 to correspond with low, medium, and

high respectively. This was applied to the original dataset. From there, I split the dataset into the training set and test set with varying test sizes for each experiment. The random states were kept constant at 0 for all experiments. From there I trained the logistic regression model on the testing set before predicting the test set to verify the accuracy for each experiment.

For my fourth experiment, I filtered the data so only the most highly correlated variables with my Y variable "Level" were used. This was done to see whether using the most correlated variables would make the regression model better. The variables that correlated the most with my Y variable were based on the results from the third heatmap which was the correlation matrix for the entire dataset. My previous three experiments used all the variables of the dataset.

B. Experimental Design

Table 2: Experiment Parameters

Experiment Number	Parameters
1	All the raw features with a 67/16.5/16.5 split for train, validate, and test
2	All the raw features with an 80/10/10 split for train, validate, and test
3	All the raw features with a 70/12/12 split for train, validate, and test
4	Filtered data with the most highly correlated variables with Level with a 67/16.5/16.5 split for train, validate, and test.

C. Tools Used

The following toolers were used for this data analysis: Python v6.5.4 running on the Anaconda environment. An HP ENVY x360 laptop was used for this analysis and the implementations. In addition to the bade Python, the following libraries were used: Pandas 2.1.0, NumPy 1.25.2, Seaborn 0.12.2. Additionally, I utilized Matplotlib, Warnings, and SKLearn. I used Pandas and NumPy for data manipulation and analysis purposes and SKLearn for creating my regression model. I used Seaborn and Matplotlib for data visualization and I used Warnings to hide warning messages.

V. RESULTS

A. Classification Measures/Accuracy measure

The following table, **Table** 3, shows the contingency table for each model to measure how well it classified data and how accurate the data was.

For my accuracy and classification measures I created classification reports for each experiment. This provided the accuracy percentage as well as the precision and f1-score. Additionally, I predicted my test set result after creating each confusion matrix as an additional accuracy measure.

Table 3: Contingency Table

Experiment Number	Accuracy
1	98%
2	97%
3	97%
4	89%

B. Discussion of Results

The confusion matrix for experiment 1 can be referenced in **Figure 7**. In terms of correct predictions, the chance of lung cancer was predicted to be low for ***115 patients*** who actually had a low chance of getting lung cancer. The chance of lung cancer was predicted medium for ***110 patients*** who actually had a medium chance of getting lung cancer. The chance of lung cancer was predicted high for ***99 patients*** who actually had a high chance of getting lung cancer. For this model, only ***6 patients*** were incorrectly predicted which implies that this model performed very well.

The confusion matrix for experiment 2 can be referenced in **Figure 8**. In terms of correct predictions, the chance of lung cancer was predicted to be low for ***74 patients*** who actually had a low chance of getting lung cancer. The chance of lung cancer was predicted medium for ***65 patients*** who actually had a medium chance of getting lung

cancer. The chance of lung cancer was predicted to be high for ***56 patients*** who actually had a high chance of getting lung cancer. For this model, only ***5 patients*** were incorrectly predicted which implies that this model also performed very well.

The confusion matrix for experiment 3 can be referenced in **Figure 9**. When looking at the correct predictions, the chance of lung cancer was predicted to be low for ***83 patients*** who actually had a low chance of getting lung cancer. The chance of lung cancer was predicted to be medium for ***80 patients*** who actually had a medium chance of getting lung cancer. The chance of lung cancer was predicted to be high for ***71 patients*** who actually had a high chance of getting lung cancer. Only ***6 patients*** were incorrectly predicted which implies that this model also performed well.

The confusion matrix for experiment 4 can be referenced in **Figure 10**. When looking at the correct predictions, the chance of lung cancer was predicted to be low for ***116 patients*** who actually had a low chance of getting lung cancer. The chance of lung cancer was predicted to be medium for ***100 patients*** who actually had a medium chance of getting lung cancer. The chance of lung cancer was predicted to be high for ***77 patients*** who actually had a high chance of getting lung cancer. However, ***37 patients*** were incorrectly predicted which implies this model performed okay but it did not perform as well compared to the three previous models.

Based on all four confusion matrixes, experiment 1, experiment 2, and experiment 3 all performed very well with most of the predicted values matching the actual values. Experiment 4, which was only made up of the most highly correlated variables, surprisingly performed worse than the previous three experiments. Though experiment 4 still performed okay. This result implies that each variable of the entire dataset may be important contributors to the effectiveness of the regression model as a whole. Thus, moving forward, experiments for this dataset should include all the raw features for its regression model.

When looking at the classification reports for each experiment, it is clear that experiment 1 performed the best given it had the highest accuracy of 98%. experiment 2 and experiment 3 had an accuracy of 97% each. This can be referenced in **Table 3**. Thus, the first three experiments performed extremely well. experiment 4 had an accuracy of 89%. Though it performed relatively well, compared to the other experiments, it performed the worst. The reason the accuracy scores were so high for each experiment could be because the dataset was very balanced. The reason the first experiment performed the best may be because of the specific test size and the fact that it encompasses all the variables within the dataset.

C. Problems Encountered

I initially ran into some problems when creating a histogram to show the distribution of all the possible symptoms. Since there were 11 variables for the symptoms, this uneven number caused the alignment of each individual subplot to be off. Eventually I decided to create a plot with two rows containing 4 subplots. The remaining three subplots were added separately in a third row. This allowed me to create a more cohesive subplot.

Additionally, I was unable to create a heatmap for only the level and its most correlated variables in a stacked manner from most correlated to least correlated. This was due to the limitations within my coding so this will be something that I would need to work on for the future.

D. Limitations of Implementation

Given that my model is a logistic regression model, it comes with some limitations. Firstly, a logistic regression model cannot predict a continuous outcome since it only works for a discontinuous outcome. Additionally, logistic regression is not always accurate when working with small datasets or sample sizes. This can lead to overfitting which can be a difficult problem to deal with in modelling. Overall, given that my dataset had discrete variables, a logistic regression model worked best. However, if my dataset was not discrete, then a linear regression model would work better.

E. Improvements/Future Work

In the future I plan on performing more experiments. I would be interested in seeing how an experiment would perform using the Y variable “Level” and its most correlated factor “Obesity” or symptom “Coughing of Blood” would perform. Additionally, I would want to try performing more experiments using all the variables but with different test sizes that I had not previously used before. Moreover, the author of this dataset had emphasized the

possible correlation with “air pollution” so that may be a variable I would want to try focusing on for my next experiment.

VI. CONCLUSION

In terms of the correlations, for the possible symptoms for lung cancer (**Figure 5**). Through this, I found that there is a highly positive correlation between coughing blood and chest pain. There are somewhat high positive correlations between fatigue and coughing up blood, fatigue and weight loss, shortness of breath and weight loss, shortness of breath and clubbing of fingers, shortness of breath and dry cough, and between dry cough and frequent cold. For the possible factors for lung cancer (**Figure 6**), it’s clear that age and gender have no correlation with the other factors. However, all the other factors seem to have a somewhat strong to very strong positive correlation with one another except for the correlation of smoking and dust allergy which seems to be a relatively weak positive correlation.

The correlation matrix for the entire dataset (**Figure 7**), showed that the chance of getting lung cancer has a very high positive correlation with levels of the following factors: air pollution, alcohol use, dust allergy, occupational hazards, genetic risk, chronic lung disease, balanced diet, obesity, passive smoker. Moreover, the chance of getting lung cancer has a very high positive correlation with levels of the following symptoms: chest pain, coughing up blood, fatigue. Additionally, the chance of getting lung cancer had a relatively high positive correlation with smoking and shortness of breath. These findings can be supported by American Cancer Society which listed symptoms that matched with what I found.

The factor that had the highest correlation with the chance of getting lung cancer was obesity at 0.83. The symptom that had the highest correlation with the chance of getting lung cancer was coughing up blood at 0.78. Surprisingly, though smoking was somewhat highly correlated at 0.52, it was not as much of a leading factor for high level of lung cancer as I had initially assumed. Therefore, it is not worth testing alone for my regression model. Overall, I think I was able to answer my initial hypothesis by identifying the possible factors and symptoms of lung cancer.

In terms of the results for my models, overall, they performed well. Experiment 1 performed extremely well with an accuracy of 98% and only 6 incorrect predictions, which was very impressive. Therefore, making experiment 1 the best performing model. Experiment 2 performed very well with an accuracy of 97% and only 5 incorrect predictions. Experiment 3 also performed well with an accuracy of 97% and 6 incorrect predictions. Experiment 4 performed decently with an accuracy of 89% and 37 inaccurate predictions. These findings indicate that the dataset was balanced and that all the variables played an important role in the accuracy of the model itself. Thus, using these models can help predict the likelihood of a patient developing lung cancer.

REFERENCES

1. (2022, November 7). Lung Cancer Causes & Risk Factors. *American Lung Association.* https://www.lung.org/lung-health-diseases/lung-disease-lookup/lung-cancer/basics/what-causes-lung-cancer#:~:text=Smoking%20is%20the%20number%20one,do%20for%20your%20lung%20health.
2. (n.d.). Lung Cancer Statistics. *World Cancer Research Fund International*. https://www.wcrf.org/cancer-trends/lung-cancer-statistics/#:~:text=It%20is%20the%20most%20common,shown%20in%20the%20tables%20below.
3. Lung Cancer Prediction. *Kaggle*. https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link

APPENDIX

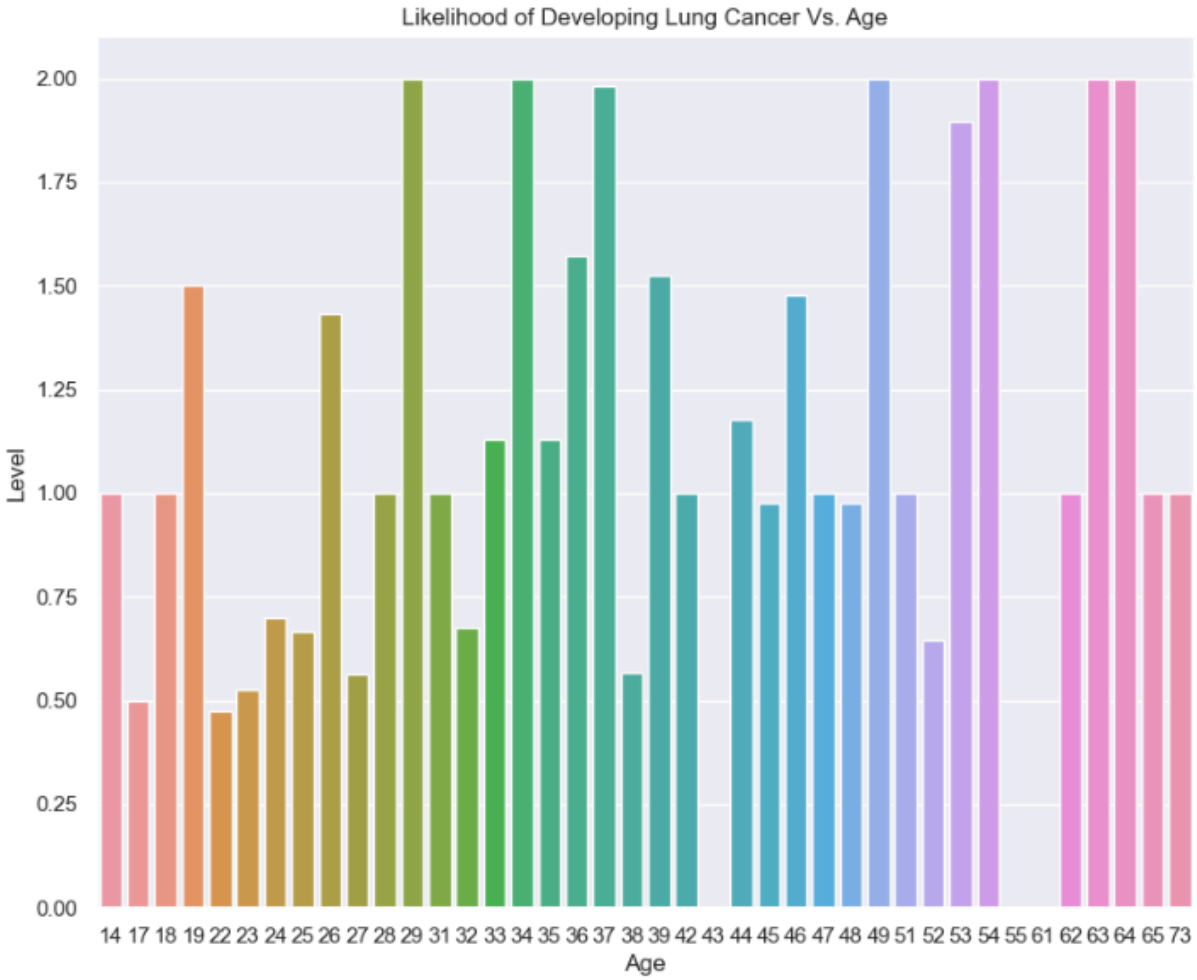

Figure 11. Shows the bar graph of the likelihood of developing lung cancer with age.

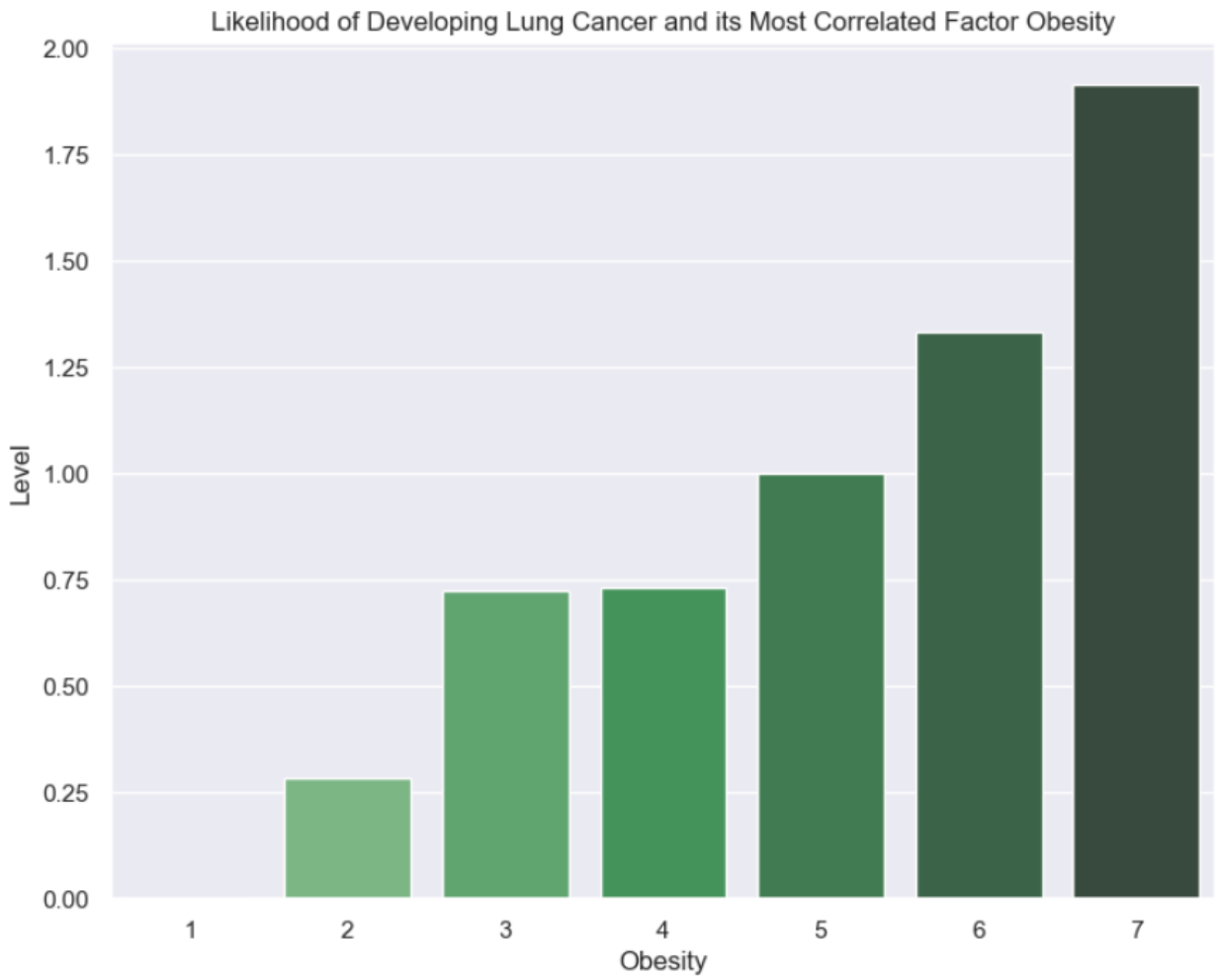

Figure 12. Shows the bar graph of the likelihood of developing lung cancer and obesity.

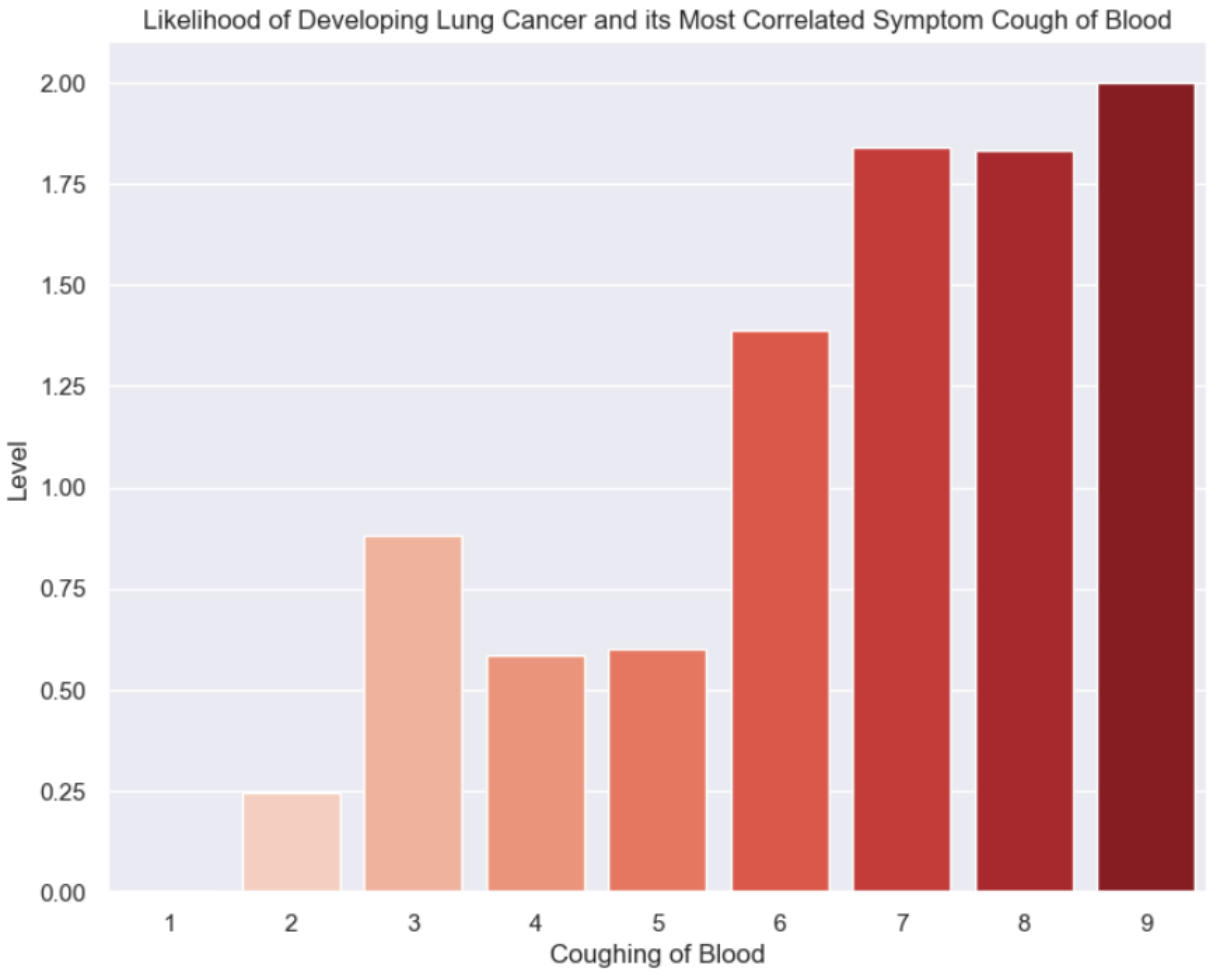

Figure 13. Shows the bar graph of the likelihood of developing lung cancer and coughing of blood.

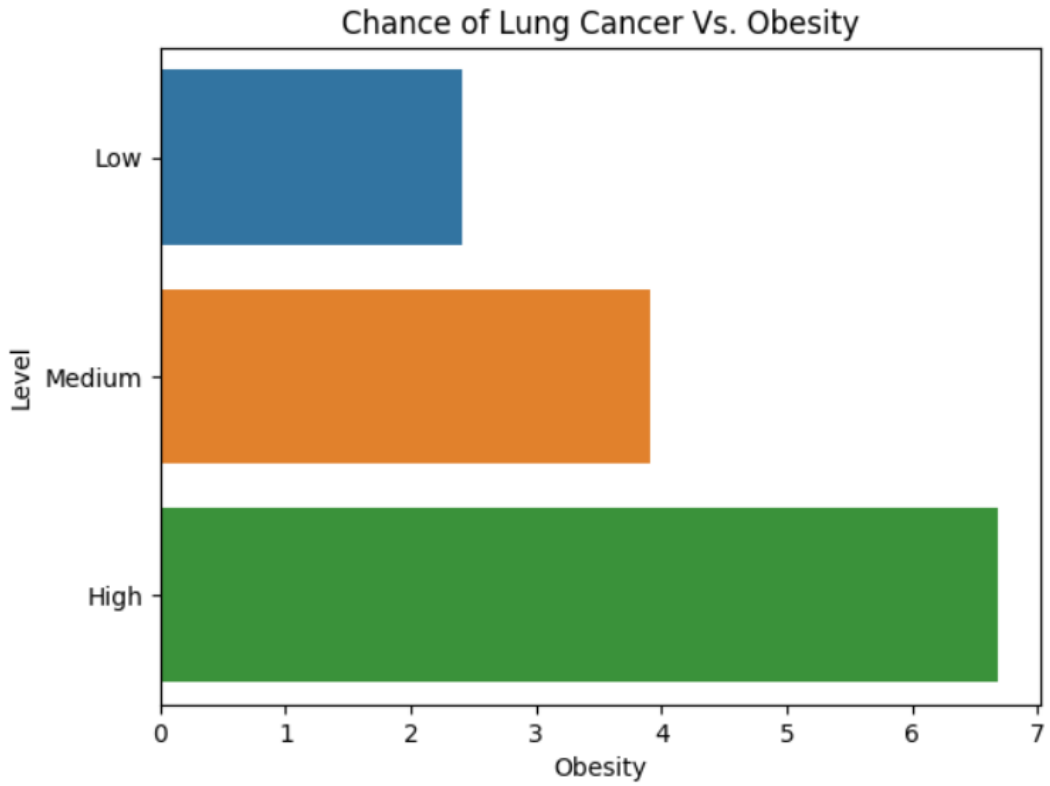

Figure 14. Bar graph which shows higher levels of obesity correspond with higher level of lung cancer.

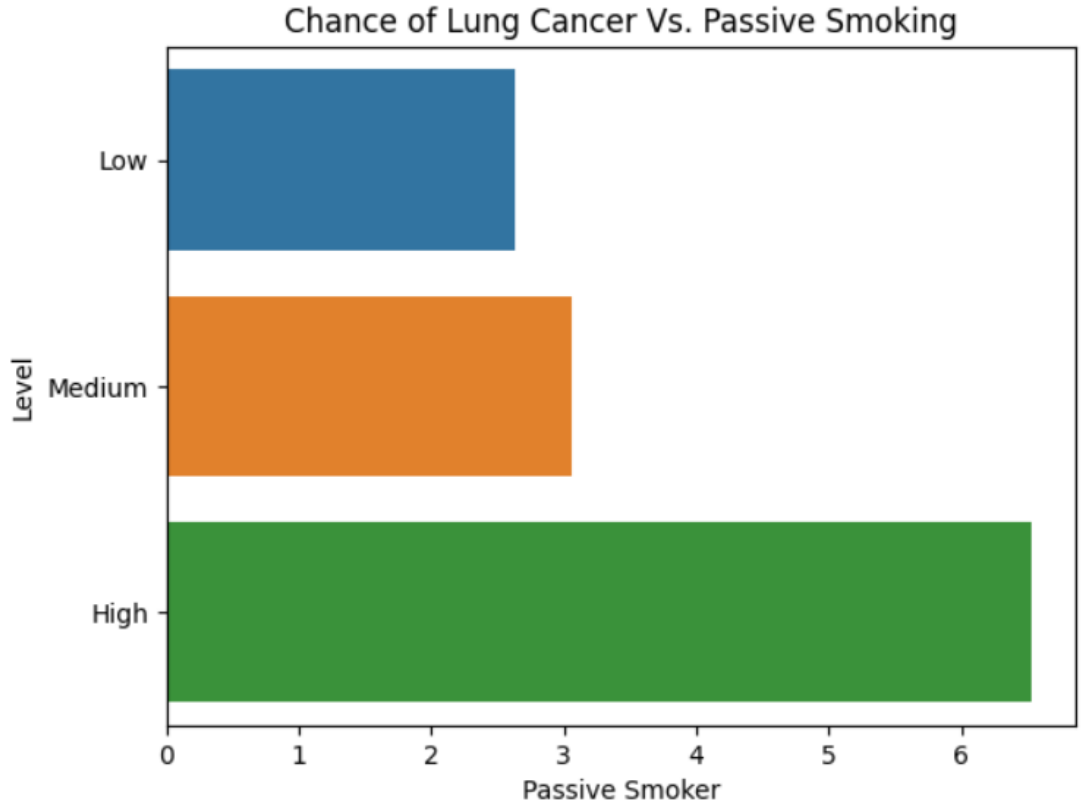

Figure 15. Bar graph which shows higher levels of passive smoking correspond with higher level of lung cancer.

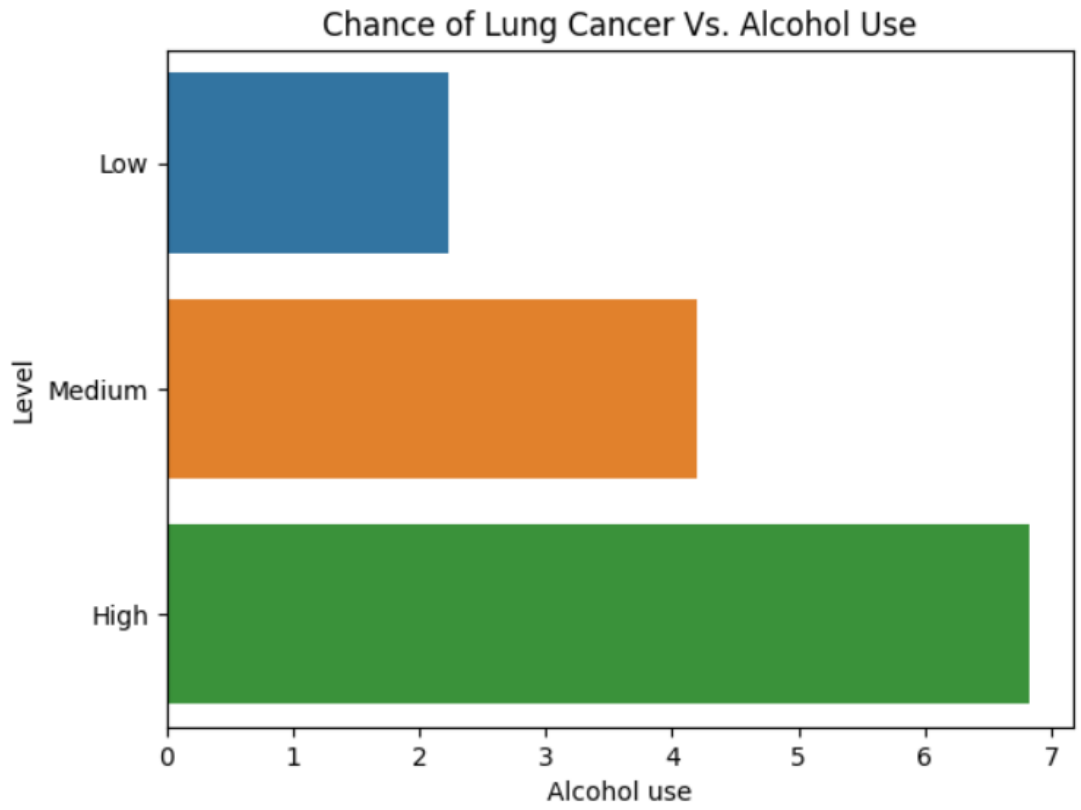

Figure 16. Bar graph which shows higher levels of alcohol use correspond with higher level of lung cancer.

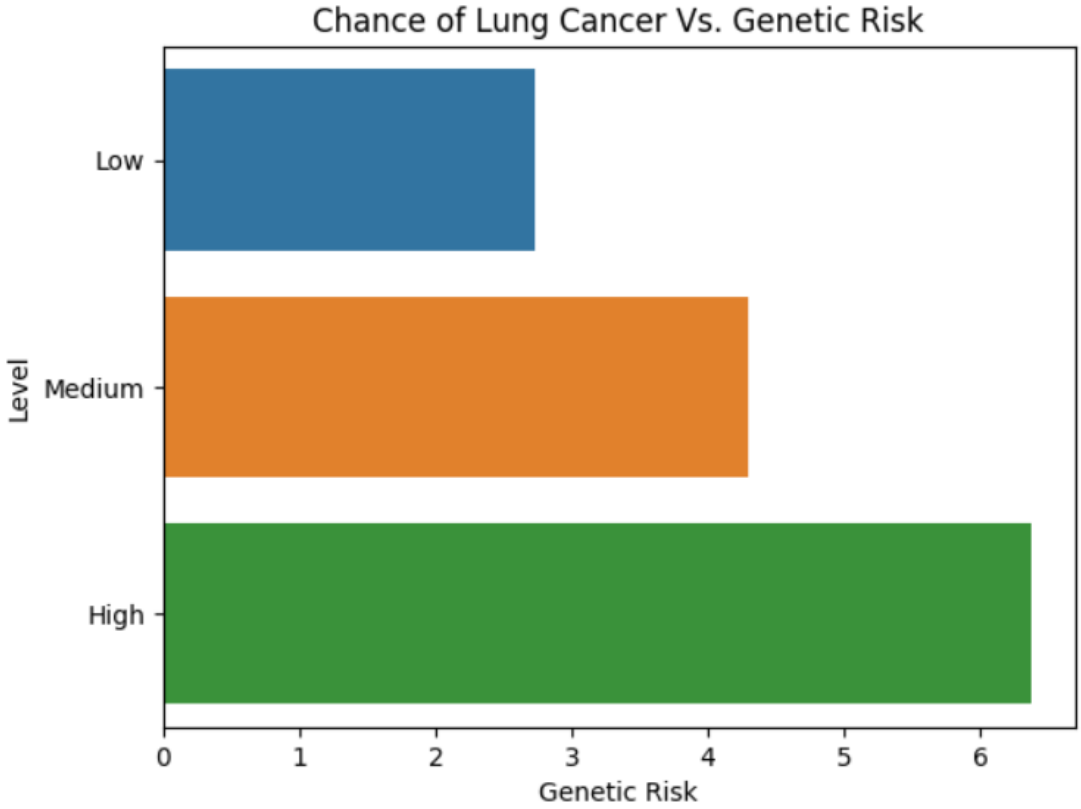

Figure 17. Bar graph which shows higher levels of genetic risk correspond with higher level of lung cancer.

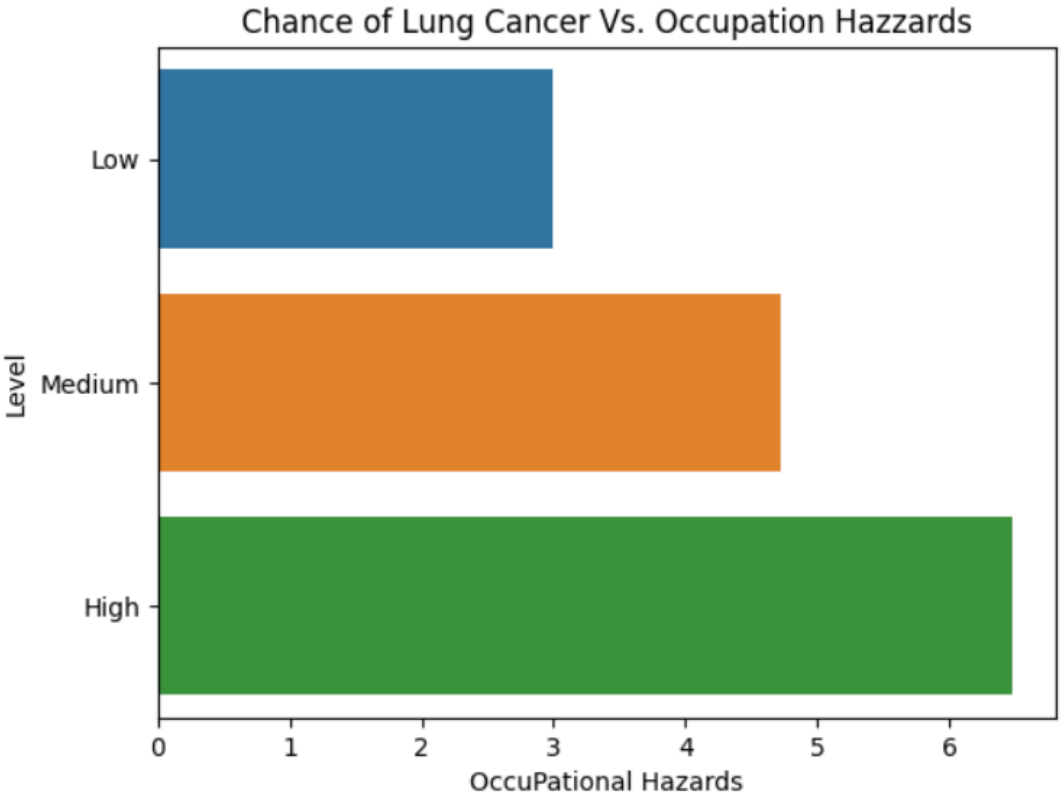

Figure 18. Bar graph which shows higher levels of occupational hazards correspond with higher level of lung cancer.

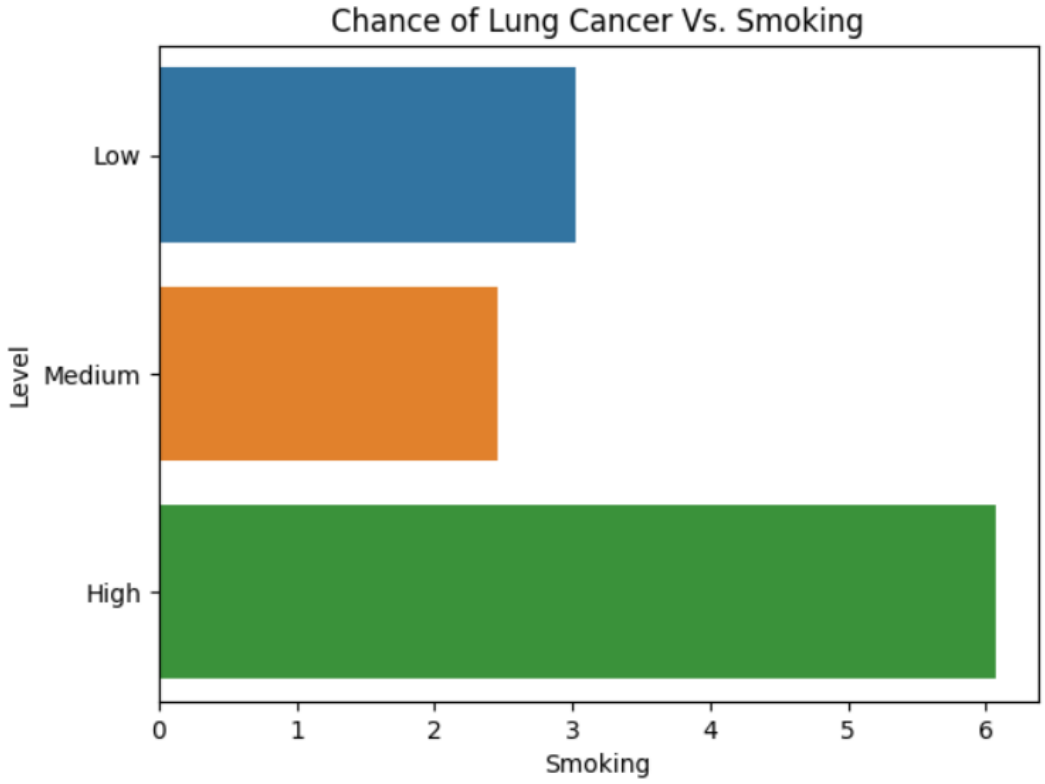

Figure 19. Bar graph which shows higher levels of smoking correspond with higher level of lung cancer.

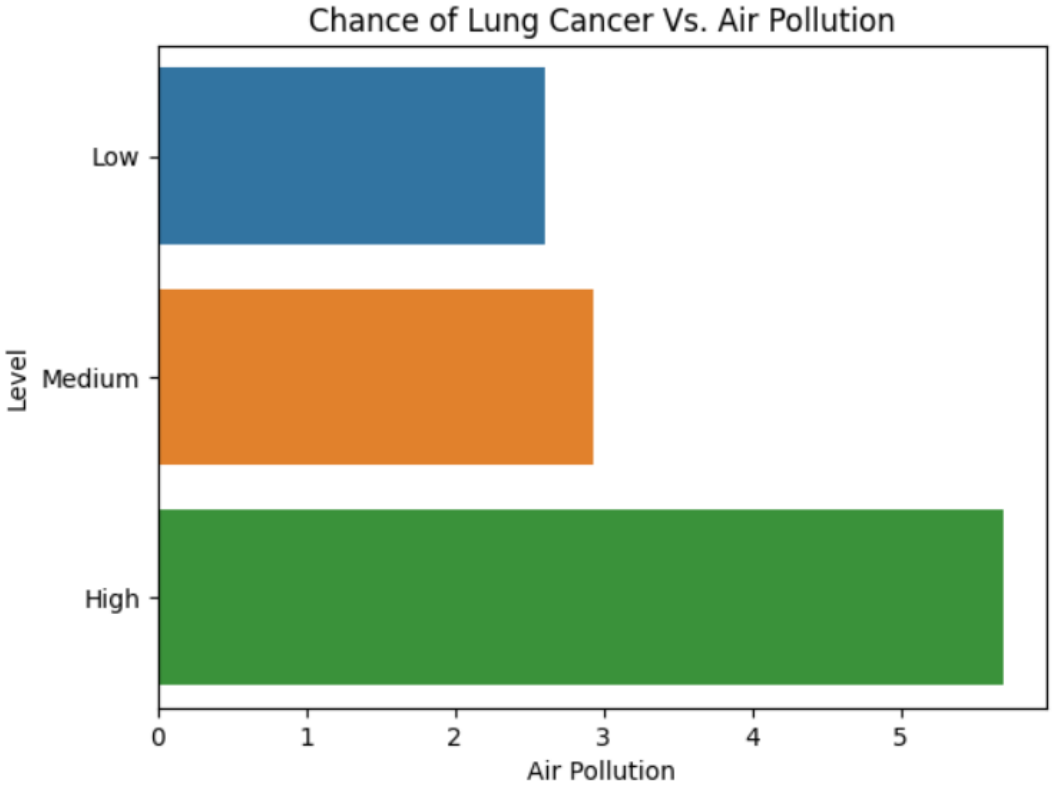

Figure 20. Bar graph which shows higher levels of air pollution correspond with higher level of lung cancer.

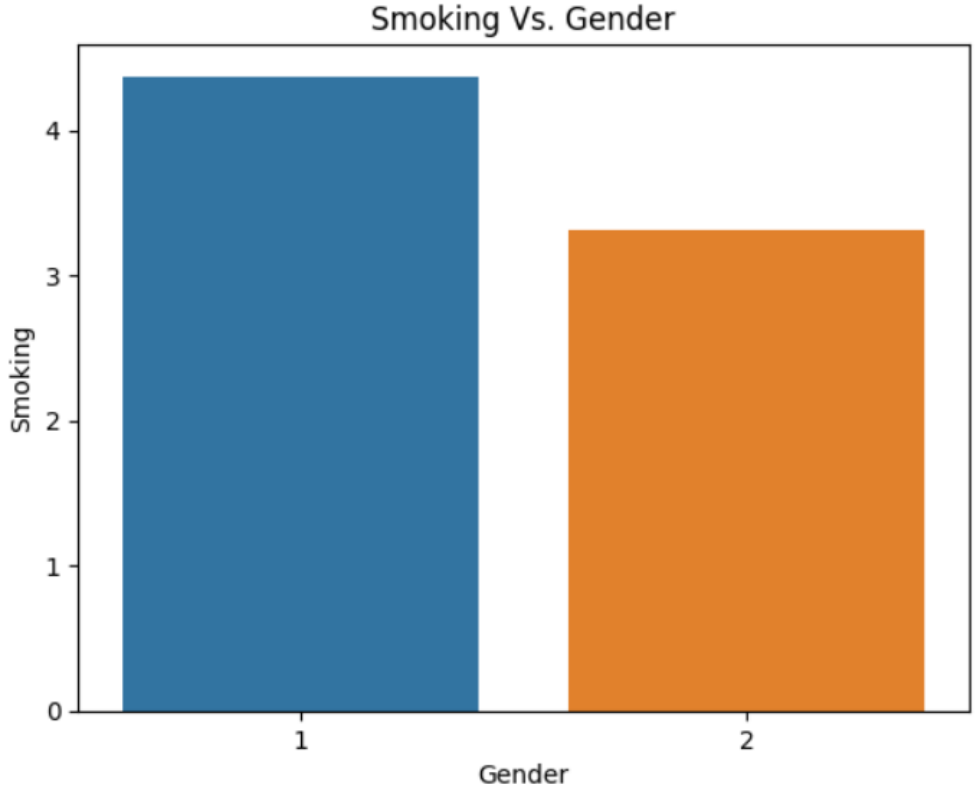

Figure 21. Shows the bar graph of Smoking Vs. Gender where 1 represents male and 2 represents female.