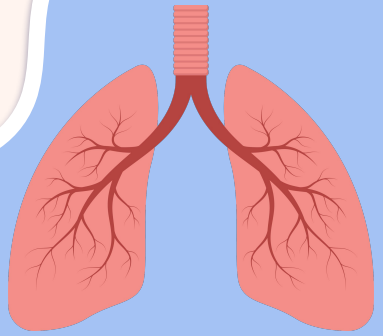
LUNG CANCER PATIENT PREDICTION

Data Analysis

GROUP 14



MOTIVATION

- Lung Cancer is the leading cause of cancer death in both men and women
- Data analysis on several factors can help people understand what habits and characteristics put them at risk
- Early detection and treatment are essential for remission and even curing cancer

OBJECTIVE

- Identify gaps in knowledge
- Illustrate trends that may be looked over
- Provide actionable information to those at risk of lung cancer

METHODOLOGY

Data Collection and Processing
Data Analysis & Model Training
Model Evaluation and Prediction

Data Collection and Processing

Handling Missing Data

Fill the missing data with mean or median values

Encoding Data

Convert categorical data into numerical numbers

Splitting Data

Splitting the data set into training and testing data sets

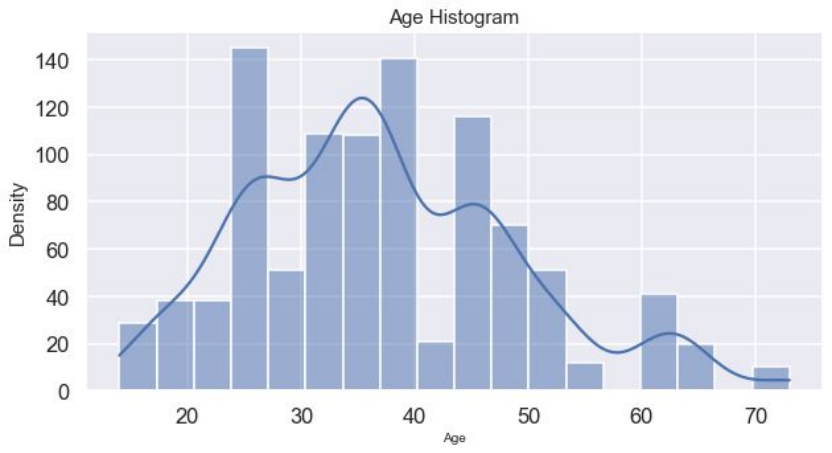
Data Set

- Kaggle dataset: https://www.kaggle.com/datasets/yusufdede/lung-cancer-dataset
- 1000 patients
- 23 categories rated 0-9 based on severity
- Example: A patient with moderate chest pain would rate it a 5
- Last column is the status of lung cancer in the patient.

Patient Info

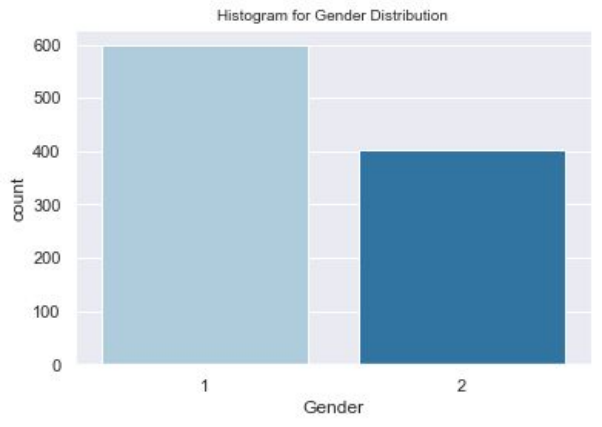
Age:

Mean age is 37.174, median is 36.0, standard deviation is 12.005.



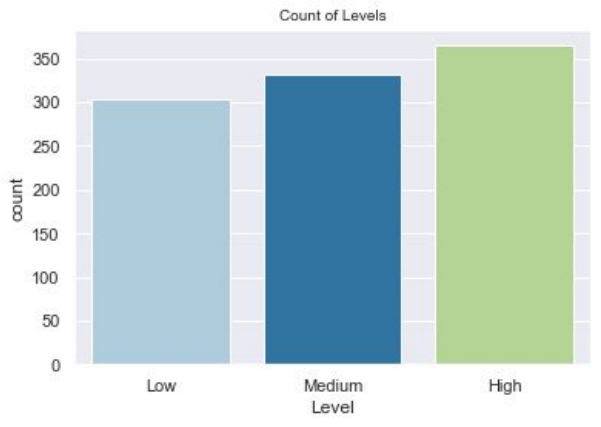
Gender:

More males than females

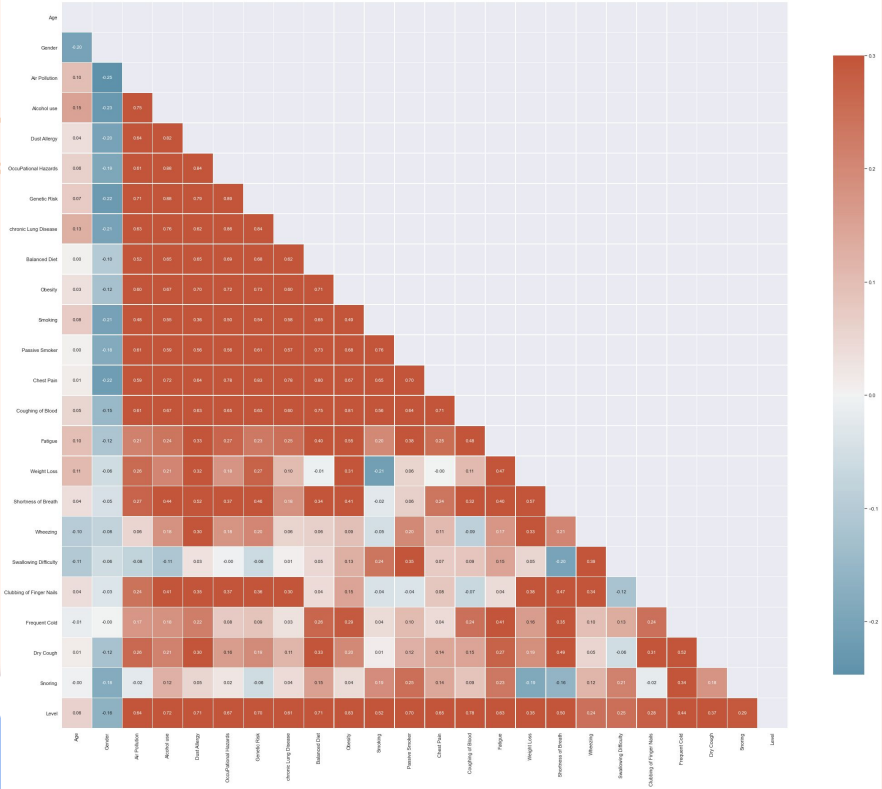


Lung cancer level:

High levels are slightly more than lower levels



Correlational Analysis



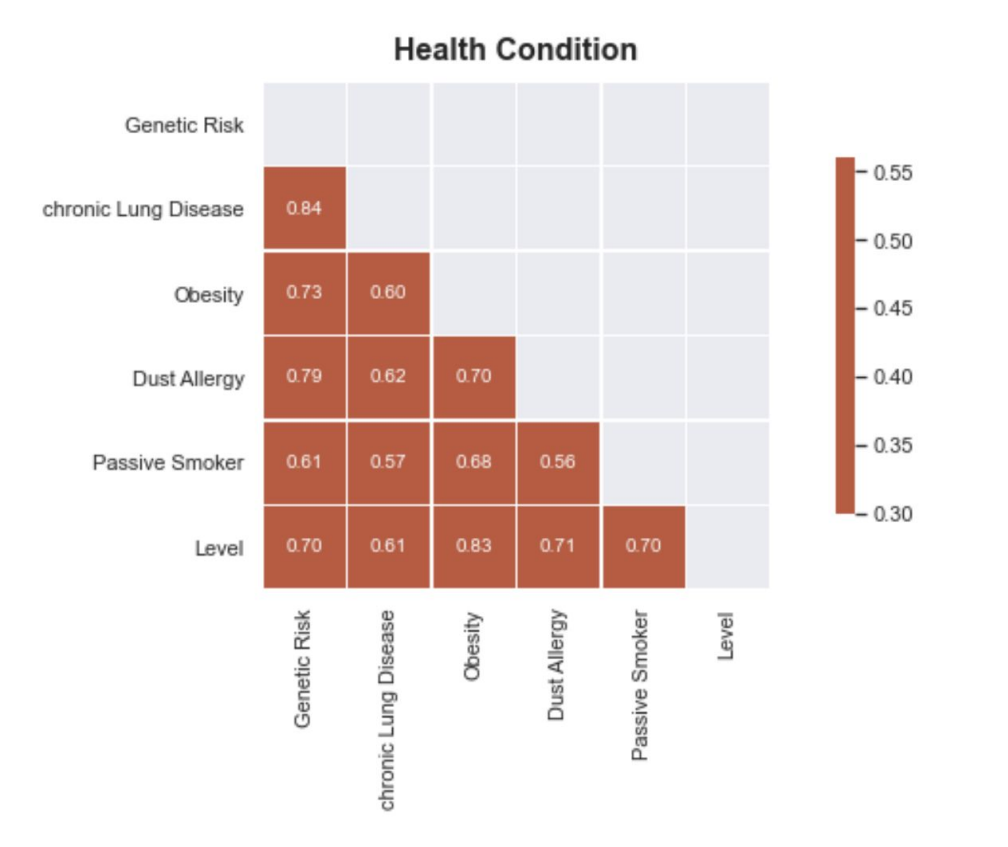
For the rest of variables, we perform a correlation analysis and it could be visualized by the heatmap shown.

Darker colors in the heatmap means higher correlation.

As we can notice, some variables are more closely related than others.

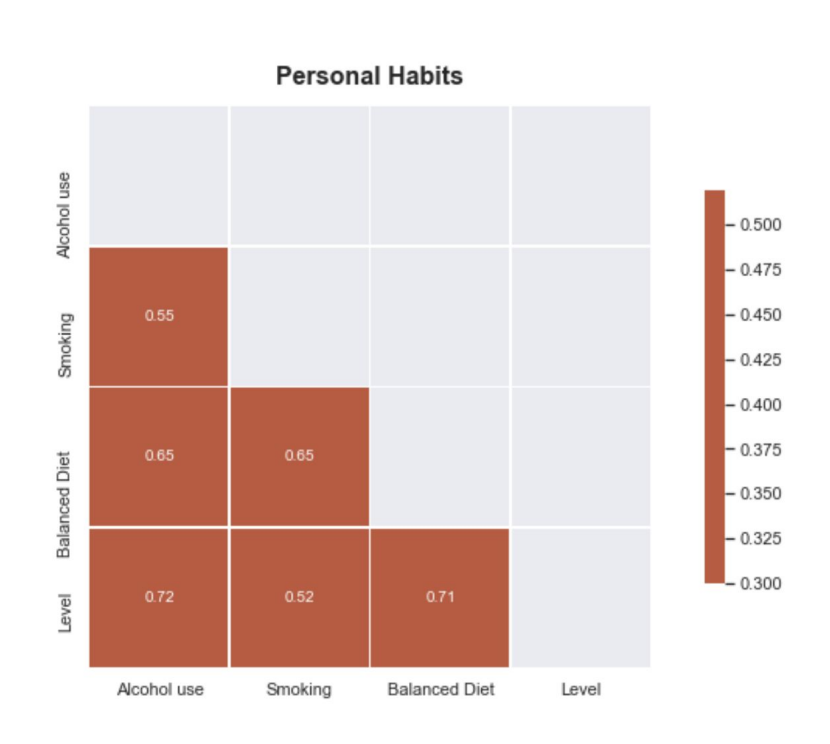
Correlational Analysis

- From the main correlation matrix, we break up characteristics based on context and their correlation with each other.
- Categorizing variables into one group if their correlation is higher than threshold.
- The first group is formed by health condition of patients; surprisingly, the level of lung cancer has strongest positive correlation with obesity.



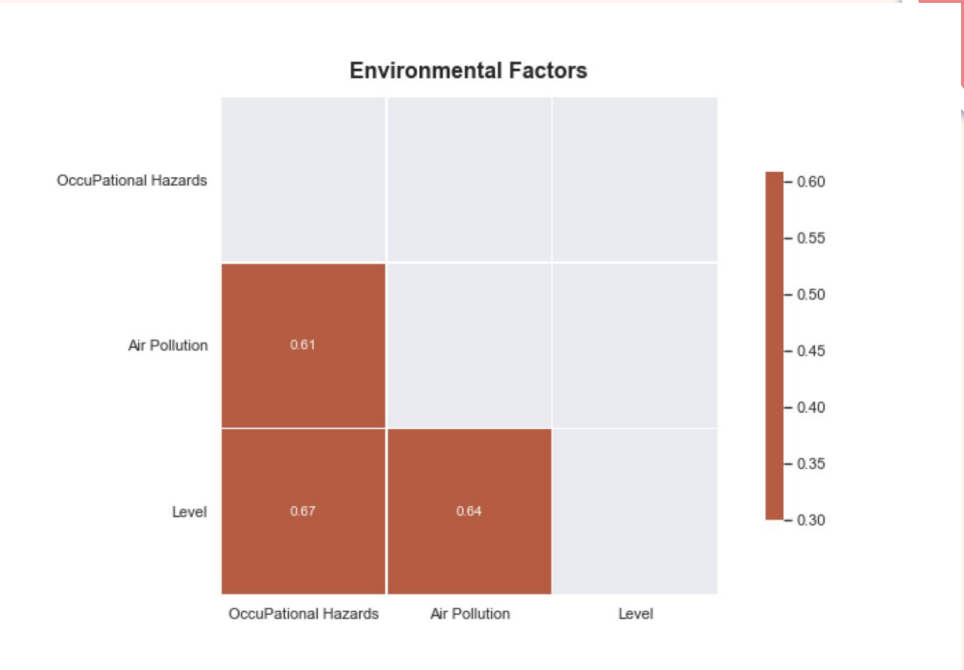
Correlational Analysis

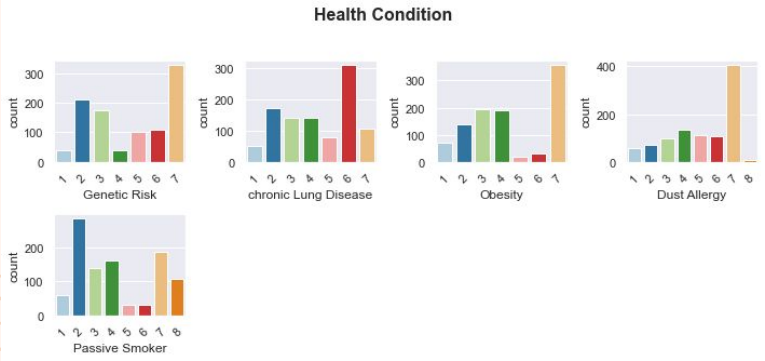
- For the personal habits category, alcohol use and diet shows the strongest correlation to the severity of lung cancer.
- Smoking was assumed to be the most significant factor correlated to lung cancer by a lot of people.

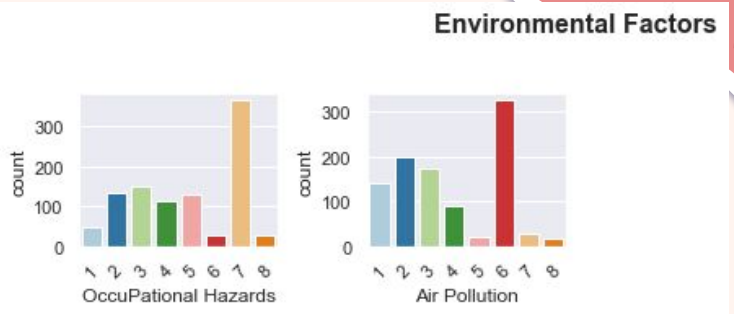


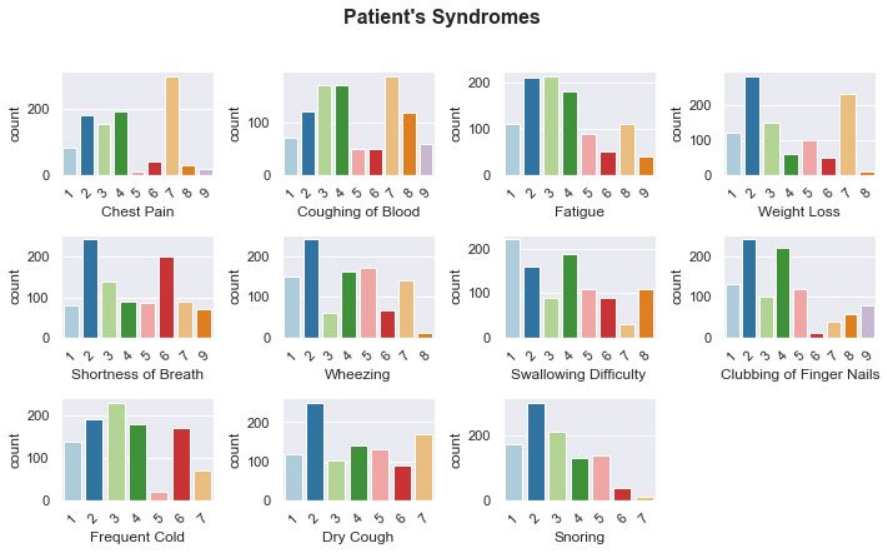
Correlational Analysis

- Air pollution and occupational hazards have a solid correlation to one another in occurrence
- The severity of lung cancer can be weakly attributed to the patient's occupation









Distribution of variables by category

The shapes are similar within category.

Model Training

Logistic Regression Model

Model was trained on data and found correlation coefficients of each of the categories.

XGBoost Model

Model found important categories that have the highest weight in influencing lung cancer.

Model Evaluation and Prediction

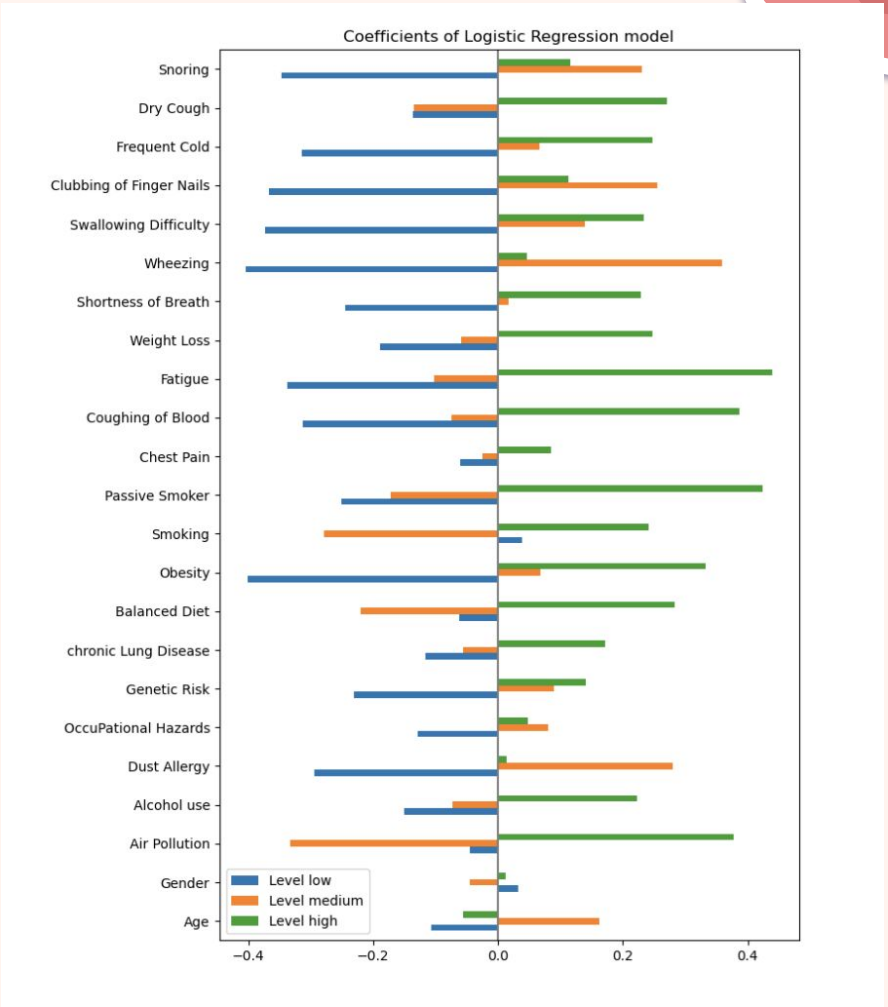
Illustrate the data graphically to easily understand the important data

Analyze Data and Draw Conclusions

Give suggestions on what people can do to help prevent and diagnose the disease.

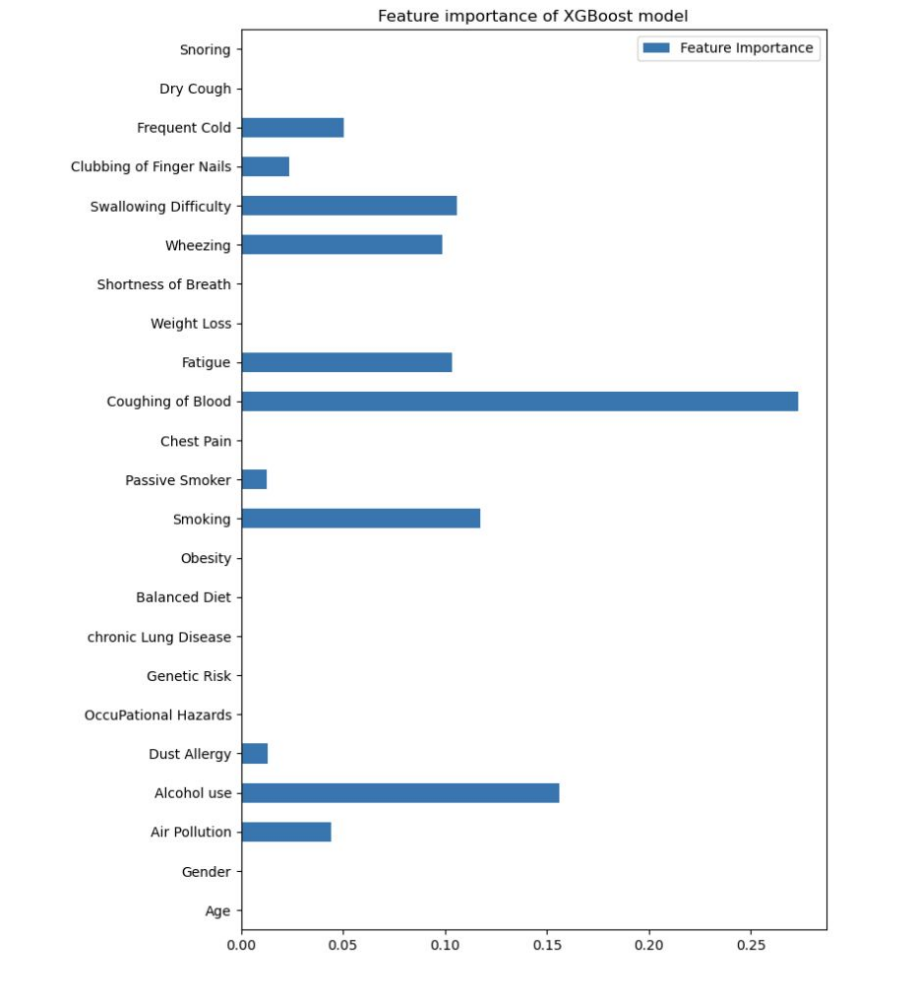
Findings

- Certain high level characteristics were more correlated with lung disease
- Fatigue, coughing blood, and a dry cough were some of the most important symptoms
- Passive smoking, obesity, and air pollution were some of the most important risk factors



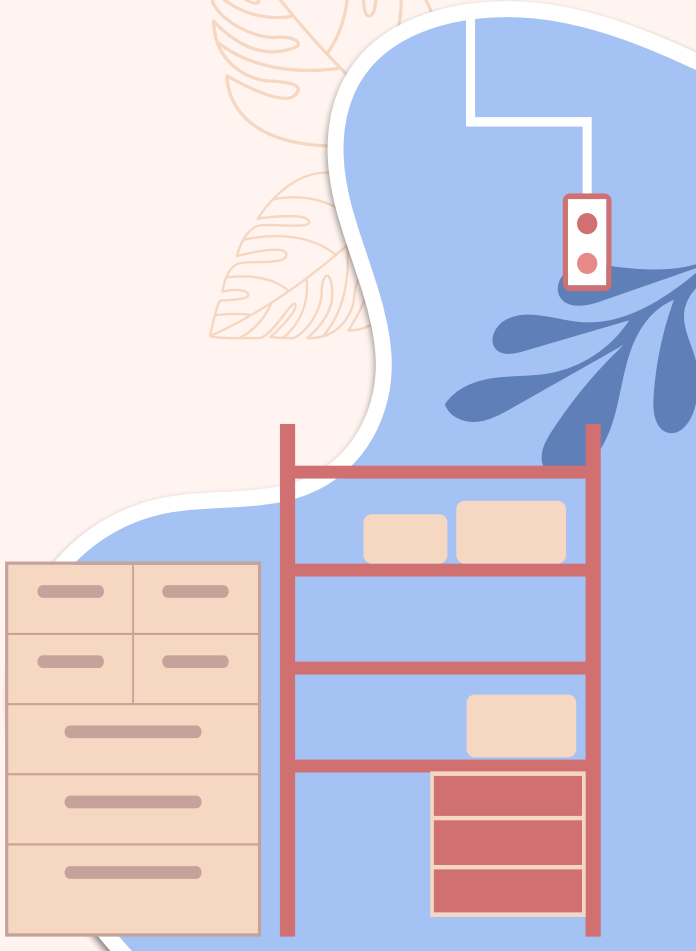
Findings

- Based on the XGBoost model only certain features were important to help rule out some of the other data
- Frequent cold, wheezing and alcohol use were some features not easily detected in the regression model



Conclusions

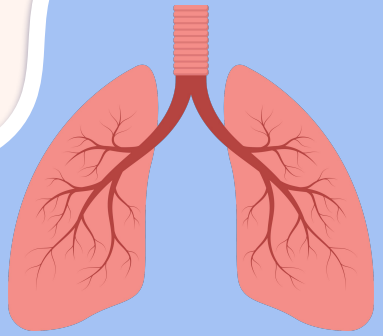
- Highest risk factors to address are smoking, obesity, and air pollution.
- Important symptoms to watch for are fatigue, coughing blood, and dry cough.
- More research needs to be done to answer questions such as:
    - What kinds of pollutants and at what concentrations?
    - How does obesity create an environment for lung cancer? What long term methods address obesity?



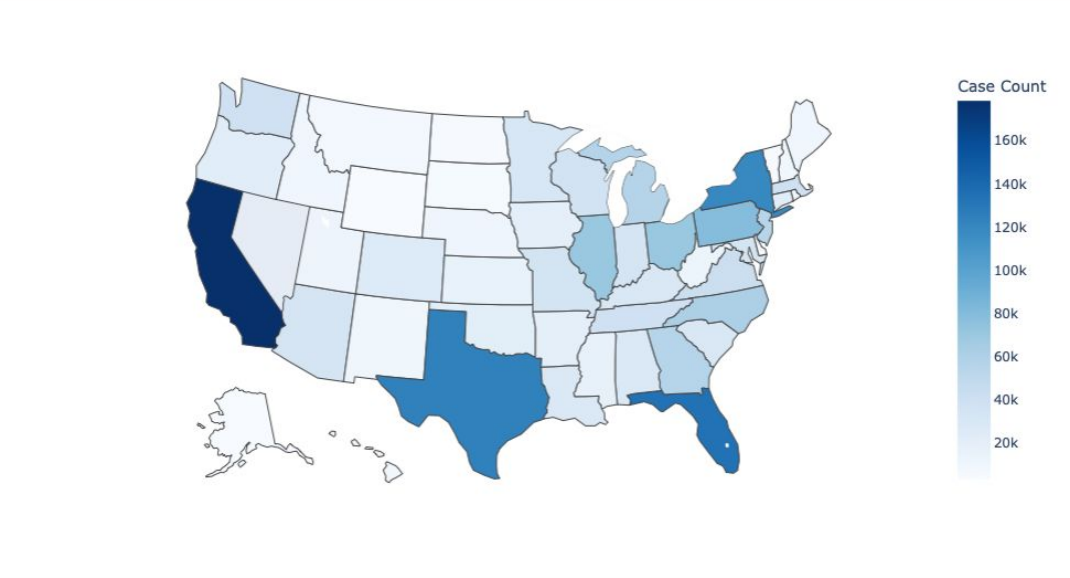
ANY QUESTIONS

Further information can be found in our github:

https://github.com/HongDeJheng/Cancer-Patient-Prediction



Some additional info



We utilized the data on new cancer cases in each state collected by the U.S. government to generate a choropleth map.

The results show that California has the highest number of cases, and interestingly, it also has the highest concentration of cities with air pollution, which aligns with our correlational analysis.