Individual Project: Lung Cancer Data Analysis

About the dataset: The data is collected from the website online lung cancer prediction system. Total number of attributes:16 Number of instances: 284

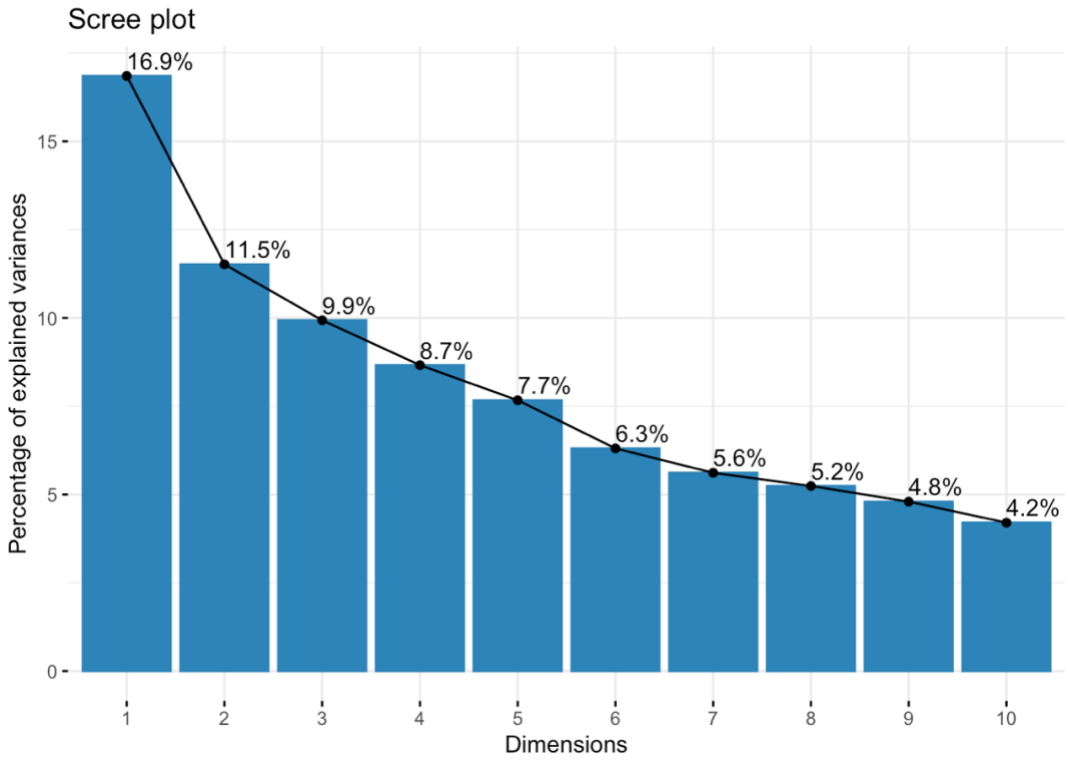
Data Dictionary:

1. Lung_Cancer – If the person is diagnosed with lung cancer or not (1 – Yes, 0 – No)
   Gender: 1(male), 0(female)
2. Age: Age of the person
3. Smoking: If the person smokes or not (1 – Yes, 0 – No).
4. Yellow fingers: If the person has yellow fingers or not (1 – Yes, 0 – No)
5. Anxiety: If the person has anxiety or not (1 – Yes, 0 – No)
6. Peer_pressure: If the person has peer pressure or not (1 – Yes, 0 – No)
7. Chronic_Disease: If the person is diagnosed with chronic disease or not (1 – Yes, 0 – No)8)
8. Fatigue: If the person has fatigue or not (1 – Yes, 0 – No)
9. Allergy: If the person has allergy or not (1 – Yes, 0 – No)
10. Wheezing: If the person is wheezing or not (1 – Yes, 0 – No)11)
11. Alcohol: If the person consumes alcoholic beverages or not (1 – Yes, 0 – No)
12. Coughing: If the person coughs too much or not (1 – Yes, 0 – No)
13. Shortness_of_Breath: If the person suffers from shortness of breath or not (1 – Yes, 0 – No)
14. Swallowing_Difficulty: If the person has swallowing difficulty or not (1 – Yes, 0 – No)
15. Chest_pain: If the person suffers from chest pain or not (1 – Yes, 0 – No)
16. Weight: Weight of the people in lb
17. Height_inch: Height of the people in inches

Questions:

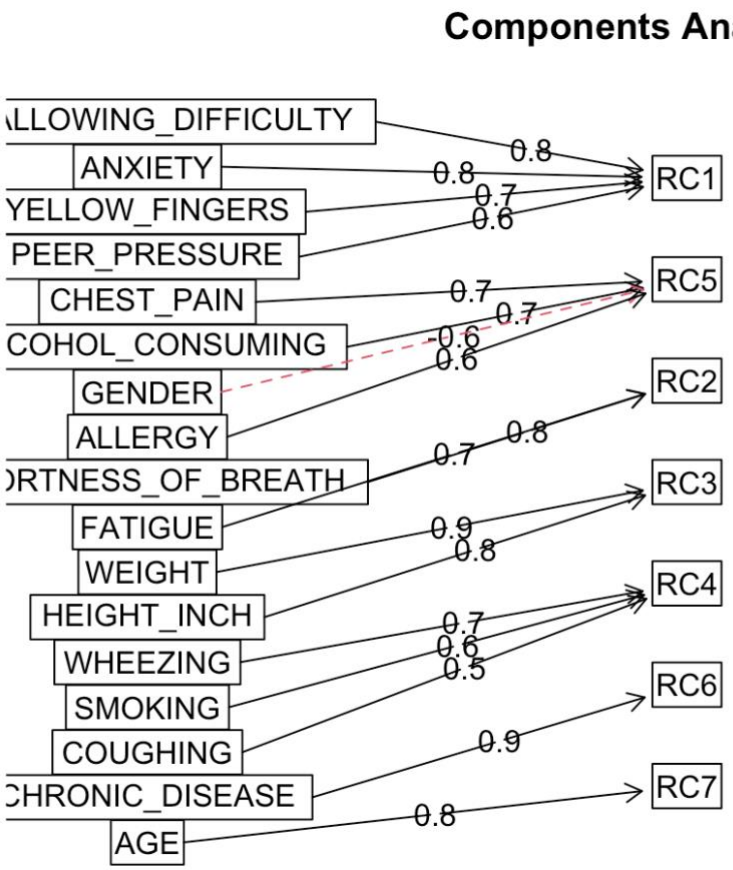
Q1) Which are the important factors contributing to lung cancer and relationship between the factors?
Q2) Predict if a person is can get lung cancer or not.

Principal Component Analysis: Q1) Which are the important factors contributing to lung cancer?



- The scree diagram shows us that sum of the first 2 principal components is 28.4% which is less than 70%.
- It means that these two components do not capture a significant amount of variation in the data.
- The data is more complex and requires more principal components to accurately represent the underlying patterns.

Exploratory Factor Analysis: Q1) Which are the important factors contributing to lung cancer and relationship between the factors?



- Factor 1 consists of yellow fingers, anxiety, peer pressure and difficulty in swallowing. Yellow fingers can be a sign of nicotine staining, which is often associated with smoking. Smoking is a habit that is often linked to anxiety, peer pressure, and other social factors. Swallowing difficulty can also be caused by health concerns, such as acid reflux or throat infections. Anxiety and peer pressure are both psychological factors that can impact a person's behavioural and physical symptoms. Someone who is anxious may experience physical symptoms such as difficulty in swallowing.
- Factor 2 consists of shortness of breath and fatigue. Shortness of breath can be caused by underlying respiratory issues which can lead to lung cancer in the future. Similarly, fatigue can also be a symptom of these conditions, as the body may have to work harder to breathe and deliver oxygen to the tissues.
- Factor 4 consists of wheezing, smoking and coughing. Smoking is a known cause of decreased lung function, which can lead to both wheezing and coughing.
- Factor 5 consists of chest pain, alcohol consumption and allergies. Allergies and alcohol consumption can both cause inflammation in the body. This inflammation can cause chest pain in some individuals. For example, alcohol consumption can cause inflammation of the oesophagus or stomach, which may lead to chest pain. Similarly, allergies can cause inflammation in the airways, which may also lead to chest pain. Both allergies and chest pain can be exacerbated by stress. Stress can also be a trigger for excessive alcohol consumption, which can lead to chest pain in some individuals.

Logistic Regression: Q2) Predict if a person has lung cancer or not given certain information about the person.

```
# Confusion matrix
conf_mat <- table(actual_lc, predicted_lc)
conf_mat
```

```
##           predicted_lc
## actual_lc No Yes
##       No   7   1
##       Yes  3  51
```

Here, there were 7 people that the model predicted don't have lung cancer and actually those 7 people don't have lung cancer. The model predicted that 3 people don't have lung cancer but they actually have lung cancer. The model predicted that 1 person has lung cancer and that person actually does not have lung cancer. The model predicted that 51 people have lung cancer and they actually have lung cancer.

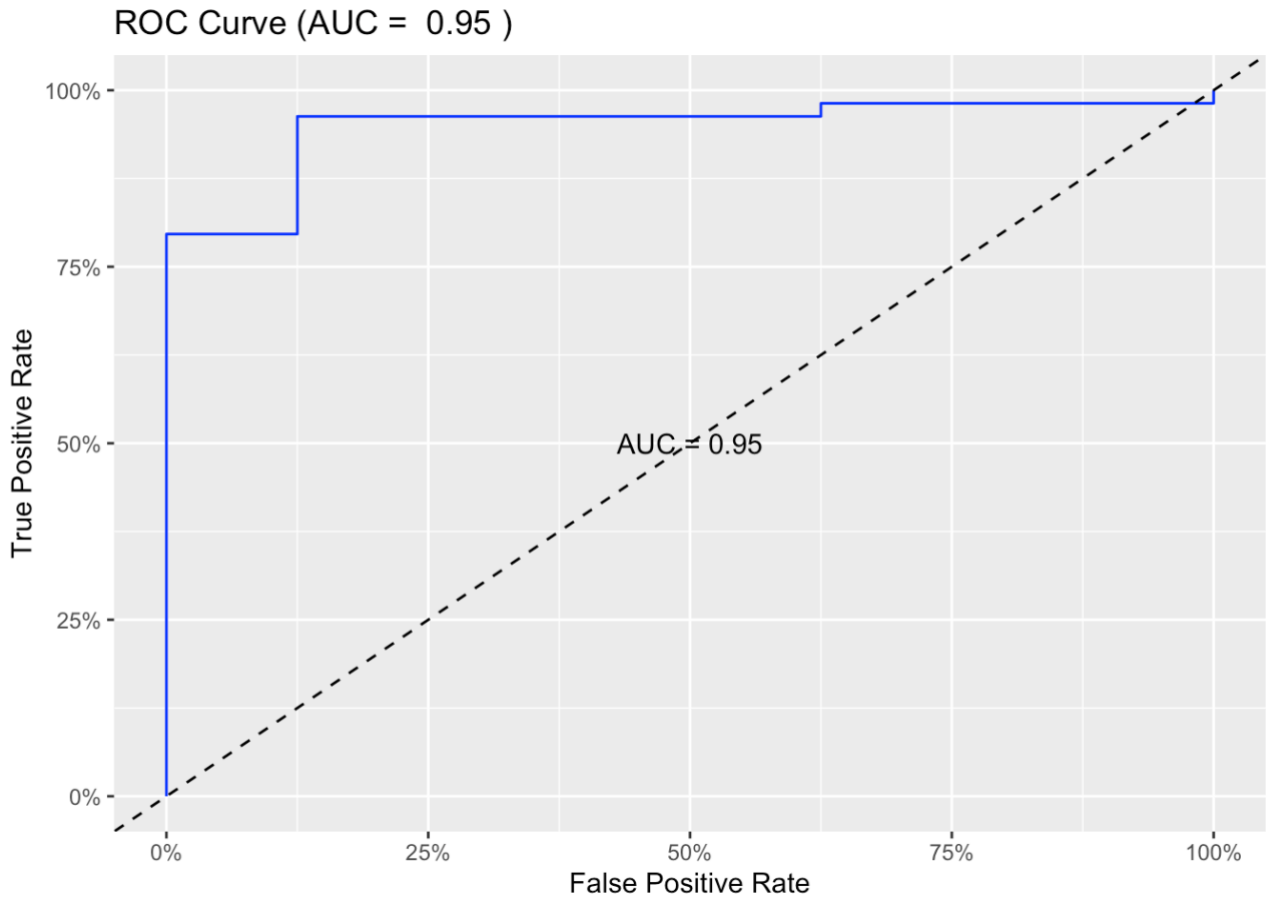
```
# Precision
precision <- conf_mat[2, 2] / sum(conf_mat[, 2])
precision
```

```
## [1] 0.9807692
```

```
# Recall
recall <- conf_mat[2, 2] / sum(conf_mat[2, ])
recall
```

```
## [1] 0.9444444
```

The model has 98% precision and recall rate of 94%.



The model has 95% of accuracy.

Conclusion: The model has 94% precision and 95% accuracy in predicting if a person has lung cancer or not given certain factors.