Lung Cancer

Sarvesh Bhandary

2022-10-30

Introduction:

This projects aims to discuss the various characteristics of lung cancer patients. We observe different causes and symptoms of patients and analyze them using different graphs using ggplot and various other tools in R. We make use of different types of plots such as histograms, pie charts, bar graphs etc. for visualization.

Data set information:

The data set consists of 309 rows and 16 columns. Each row in the data set provides information about each patient. Out of the 16 columns, the first 15 columns provide information regarding the different features and the last column denotes whether the patient has lung cancer or not.

The different features taken into account are, or relate to, gender, age, smoking, yellow fingers, anxiety, peer pressure, chronic diseases, fatigue, allergy, wheezing, alcohol, coughing, shortness of breath, swallowing difficulty and chest pain.

Age is the only numeric feature and every other feature is categorical and binary in nature.

Out of the total of 309 patients, 270 had lung cancer.

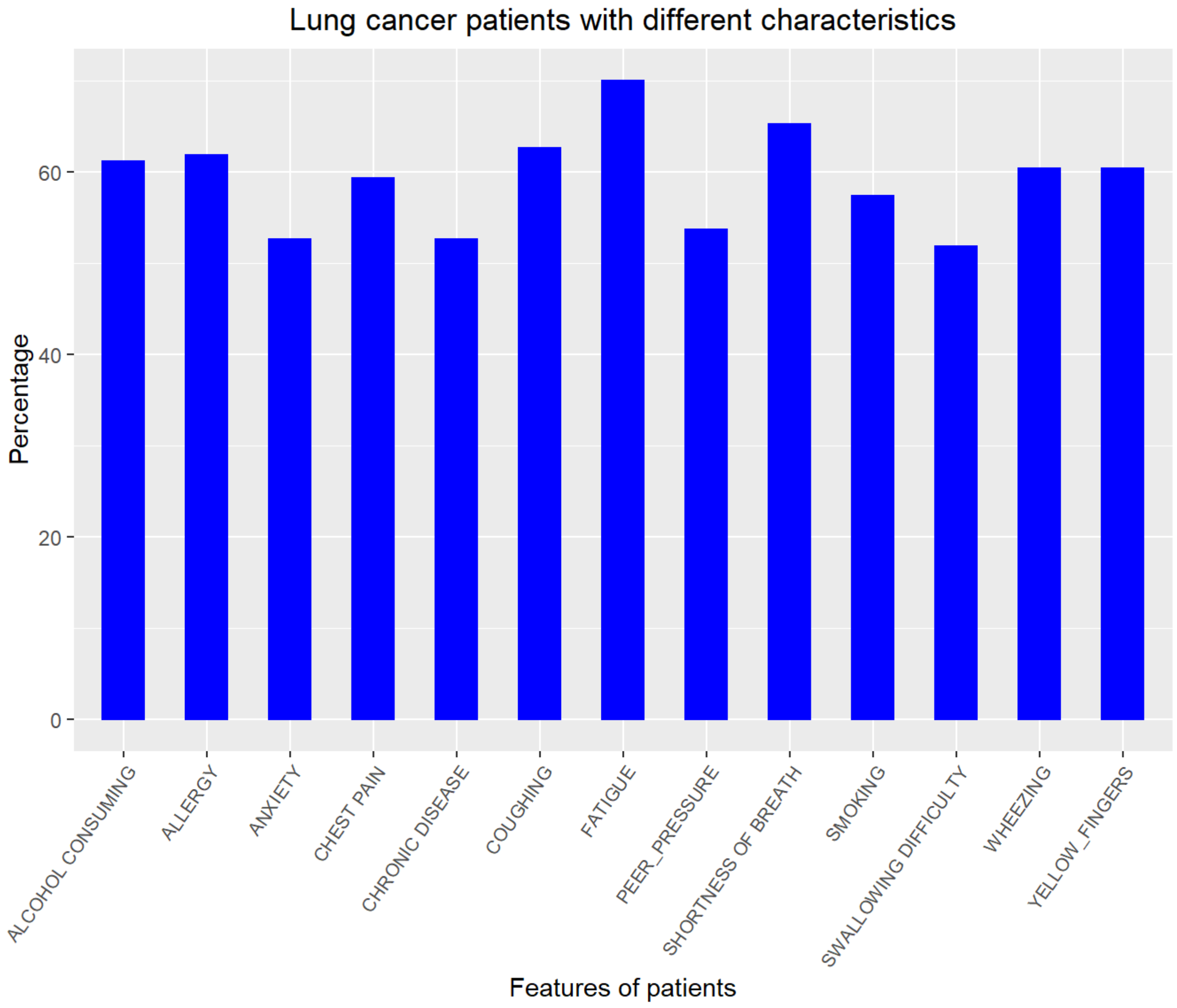
A subset of the data showing some of the features is displayed below.

	GENDER	A...	SMOKING	FATIGUE	ALLERGY	WHEEZING	LUNG_CANCER
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	M	69	1	2	1	2	YES
2	M	74	2	2	2	1	YES
3	F	59	1	2	1	2	NO
4	M	63	2	1	1	1	NO
5	F	63	1	1	1	2	NO
6	F	75	1	2	2	2	YES

6 rows

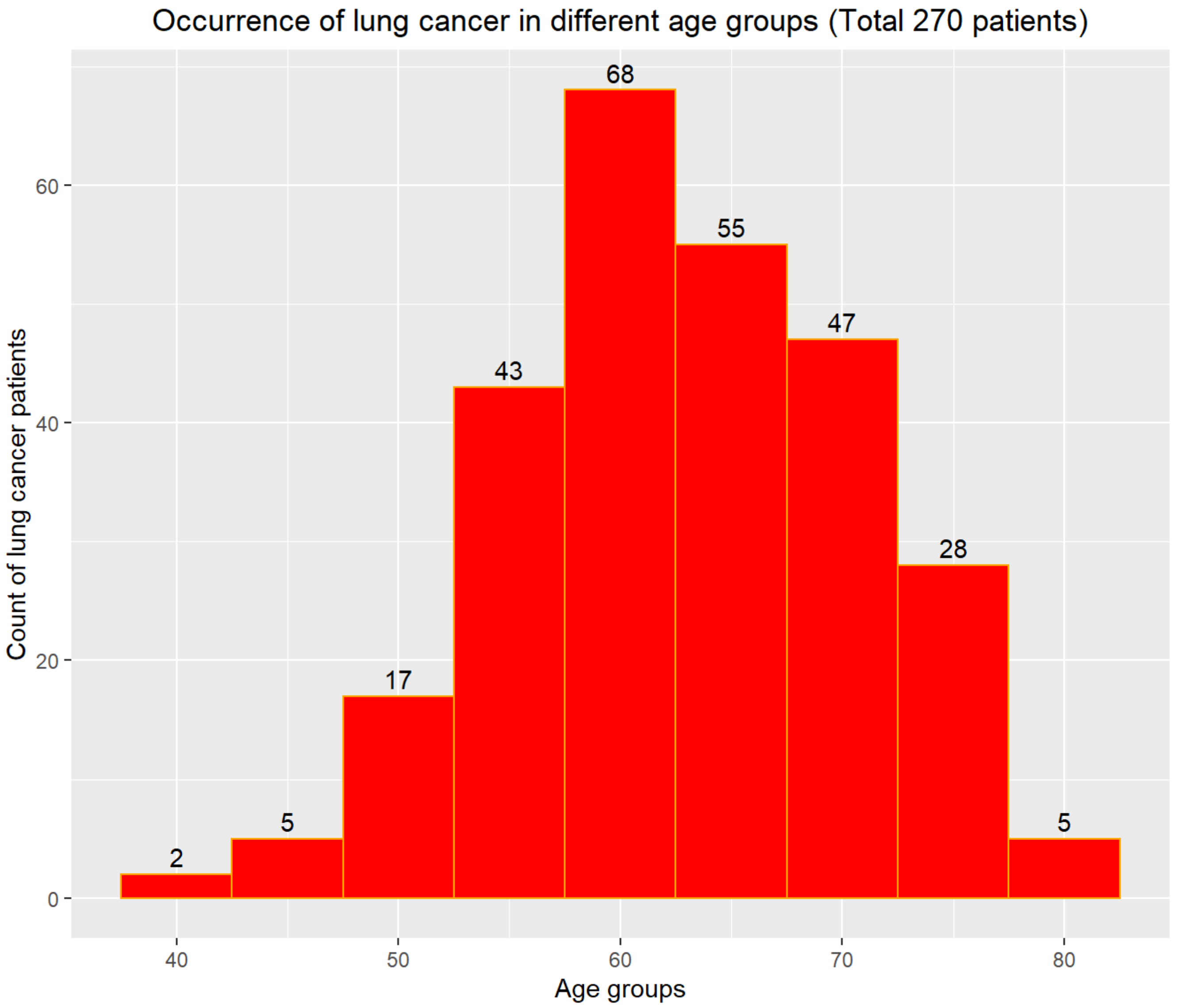
Graphical representation of key variables:

We look at the different features exhibited by the patients having lung cancer to understand which feature is seen to be present how often.

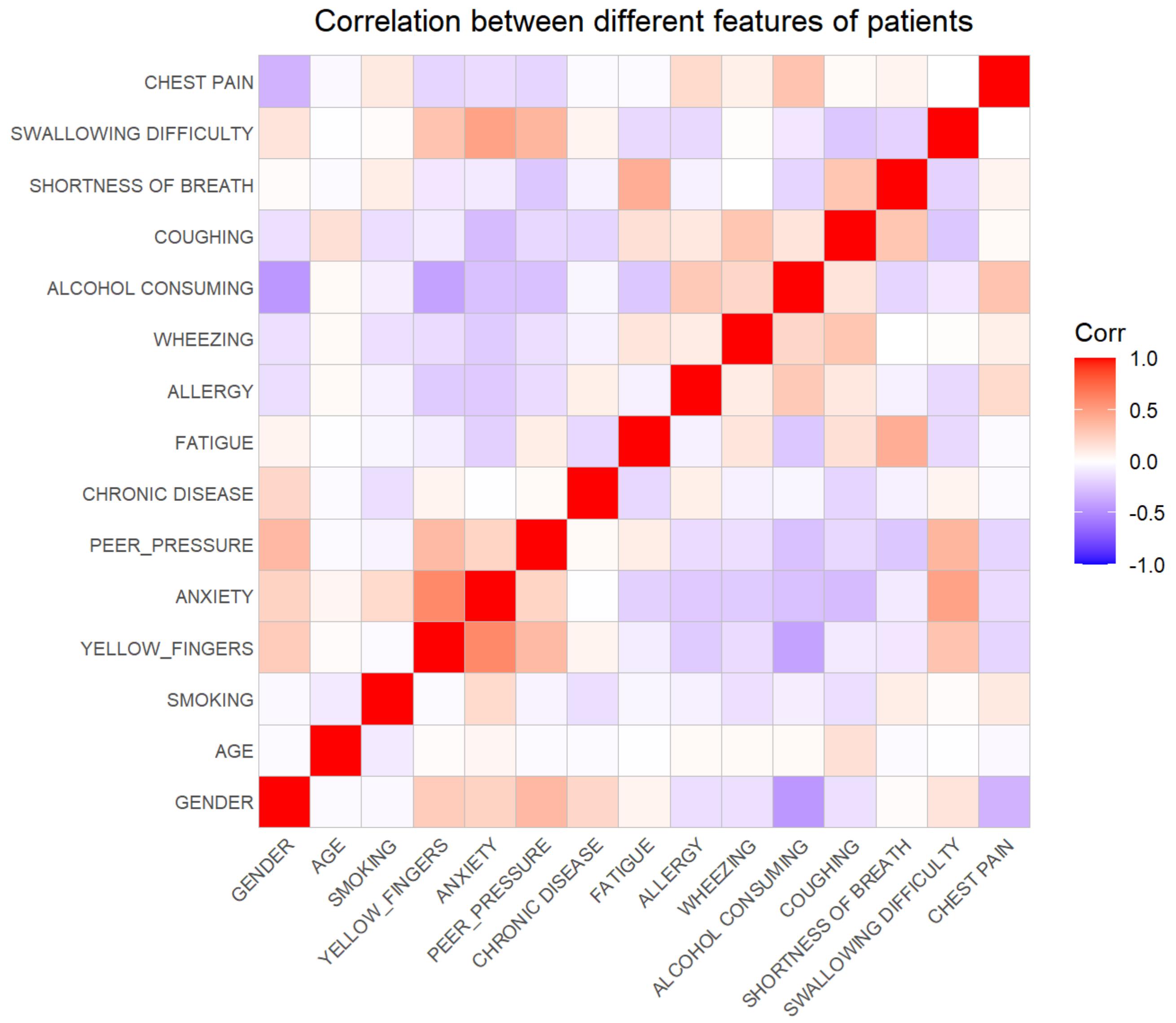


We observe that of all the analyzed features, fatigue is seen in most of the lung cancer positive patients.

Another important factor to consider while analyzing lung cancer patients is their age. We observe that the age group around 60 years is affected the most in terms of numbers. Older people are more prone to the disease when compared to younger ones.

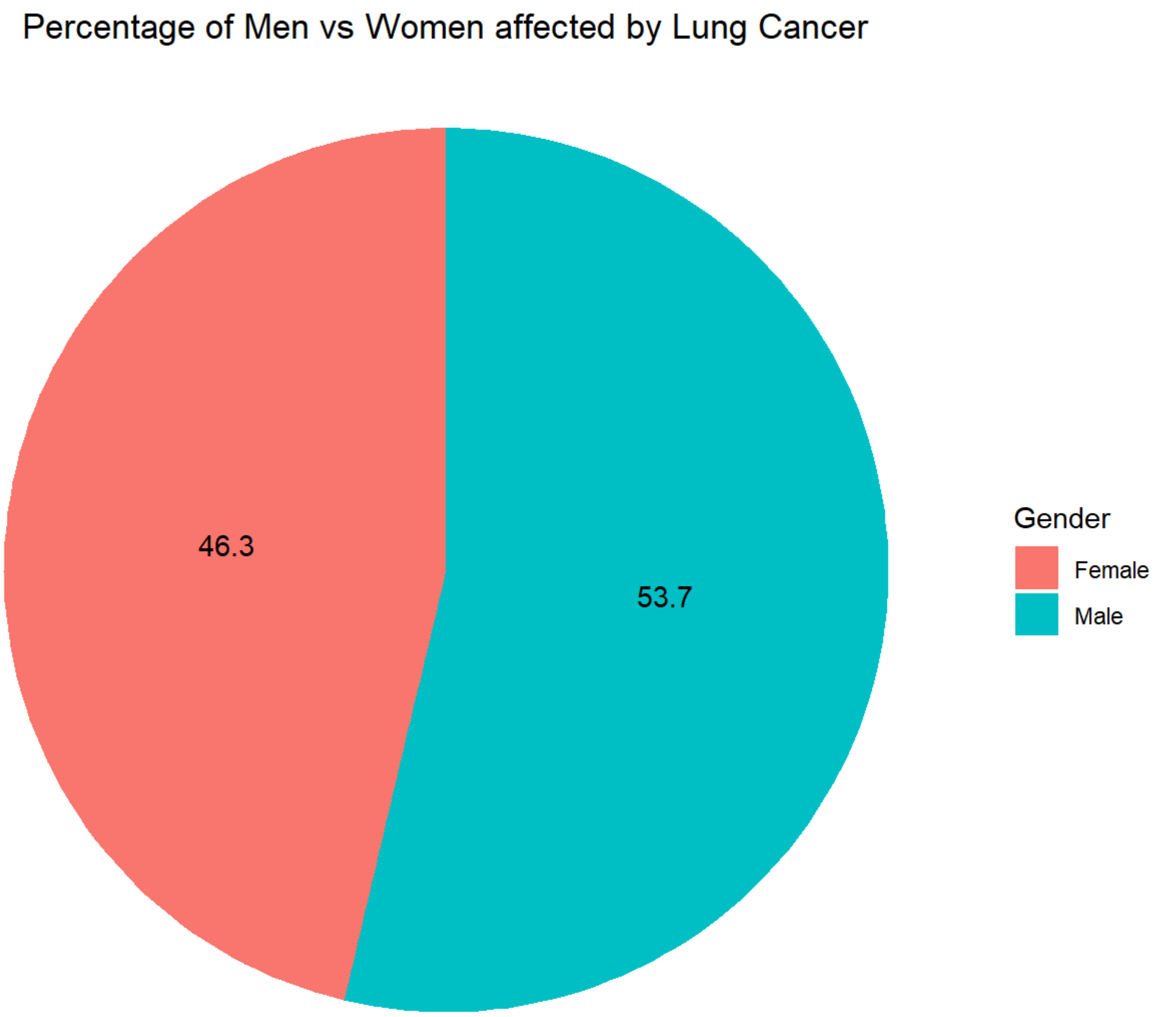


When analyzing different features, it is a good practice to understand how correlated the features are with each other. We do this with the help of a correlation matrix and plot it with the help of ggcorrplot. Darker colors indicate higher magnitude of correlation.



We notice that most features are not highly correlated with each other. In fact, the most highly correlated features that we observe are anxiety and having yellow fingers, whose correlation coefficient is 0.596. It is hence, safe to say that most features are not highly correlated with each other. This is important if we want to perform prediction using this data.

Finally, we also try to compare the number of men and women affected by lung cancer.



Here, we observe that among the 270 positive patients, 145 are men and 125 are women. The disease is observed relatively more in men than in women by a little margin.

Summary:

We analyzed the different features regarding lung cancer patients and made a couple of observations.

1. Fatigue is observed in most lung cancer patients.
2. Lung cancer is seen more in older people than young ones.
3. The features tabulated are not highly correlated to each other.
4. Lung cancer is seen slightly more in men than in women.

Conclusion:

We analyzed different features of lung cancer patients and some key graphs are plotted above. In the dashboard, a more detailed representation of the several features would be provided including how different features appear in men and women, visualizations of combinations of features seen in patients etc. It will provide a clearer understanding of the different features of the patients.

Lung cancer is among the leading causes of death among the different cancers and hence, visualizing different features of lung cancer patients helps us get an idea about the different causes and symptoms of the disease.