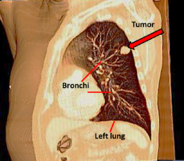
**Lung Cancer – Exploratory Data Analysis**

By- Subrahmanya Gaonkar

Lung cancer is the uncontrollable growth of abnormal cells in one or both of the lungs. Cigarette smoking causes most lung cancers when smoke gets in the lungs. Lung cancer kills 1.8 million people each year, more than any other cancer. It has an 80-90% death rate, and is the leading cause of cancer death in men, and the second leading cause of cancer death in women.

 **Global cancer burden -**

The global cancer burden is estimated to have risen to 18.1 million new cases and 9.6 million deaths in 2018. One in 5 men and one in 6 women worldwide develop cancer during their lifetime, and one in 8 men and one in 11 women die from the disease. Worldwide, the total number of people who are alive within 5 years of a cancer diagnosis, called the 5-year prevalence, is estimated to be 43.8 million.

**Global cancer patterns by sex -**

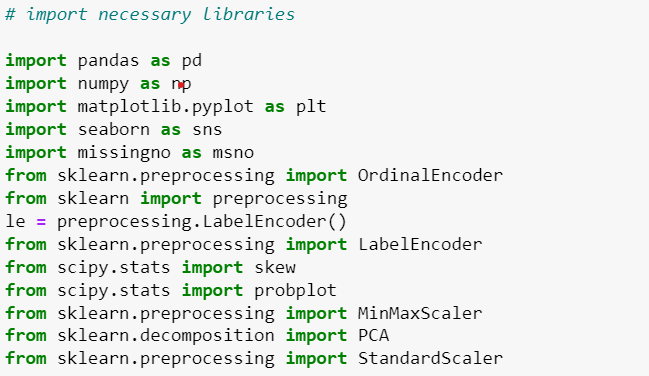
Lung cancer is the most commonly diagnosed cancer in men (14.5% of the total cases in men and 8.4% in women) and the leading cause of cancer death in men (22.0%, i.e., about one in 5 of all cancer deaths). In men, this is followed by prostate

Three-dimensional (3D) CT image, shows a tumor in the left lung.

cancer (13.5%) and colorectal cancer (10.9%) for incidence and liver cancer (10.2%) and stomach cancer (9.5%) for mortality. Breast cancer is the most commonly diagnosed cancer in women (24.2%, i.e., about one in 4 of all new cancer cases diagnosed in women worldwide are breast cancer), and the cancer is the most common in 154 of the 185 countries included in GLOBOCAN.

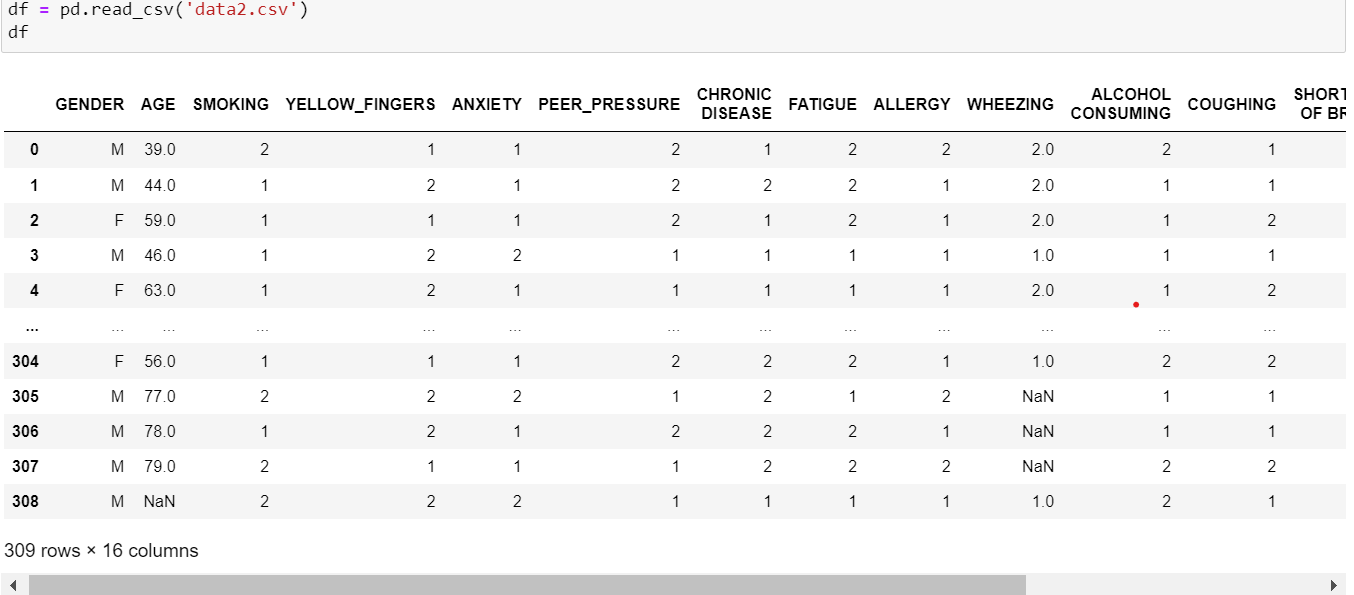
source - <https://www.iarc.who.int/wp-content/uploads/2018/09/pr263_E.pdf>

1. **Importing the library:**

****

libraries serve as indispensable tools for data exploration by providing efficiency, consistency, and a rich set of functionalities that cover data manipulation, visualization, statistical analysis, and integration with advanced analytical techniques. They empower data professionals to extract meaningful insights from data efficiently and effectively.

1. GENDER: Gender of the individual (M: Male, F: Female)
2. AGE: Age of the individual
3. SMOKING: Smoking status (2: Yes, 1: No)
4. YELLOW\_FINGERS: Presence of yellow fingers (2: Yes, 1: No)
5. ANXIETY: Anxiety level (2: High, 1: Low)
6. PEER\_PRESSURE: Peer pressure level (2: High, 1: Low)
7. CHRONIC DISEASE: Presence of chronic disease (2: Yes, 1: No)
8. FATIGUE: Fatigue level (2: High, 1: Low)
9. ALLERGY: Allergy status (2: Yes, 1: No)
10. WHEEZING: Wheezing condition (2: Yes, 1: No)
11. ALCOHOL CONSUMING: Alcohol consumption status (2: Yes, 1: No)
12. COUGHING: Presence of coughing (2: Yes, 1: No)
13. SHORTNESS OF BREATH: Shortness of breath condition (2: Yes, 1: No)
14. SWALLOWING DIFFICULTY: Difficulty in swallowing (2: Yes, 1: No)
15. CHEST PAIN: Presence of chest pain (2: Yes, 1: No)
16. LUNG\_CANCER: Lung cancer diagnosis (2: Yes, 1: No)
17. **Reading the data and saving it into variable df – Here 2 = YES and 1 = NO**

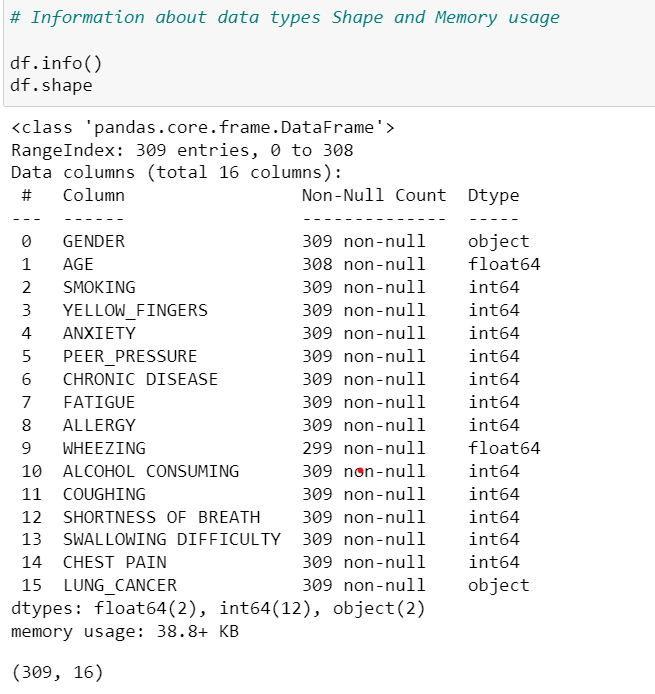
****

This code uses the Pandas library in Python to read a CSV (Comma Separated Values) file named 'data2.csv' and load its contents into a Pandas Data Frame**.**

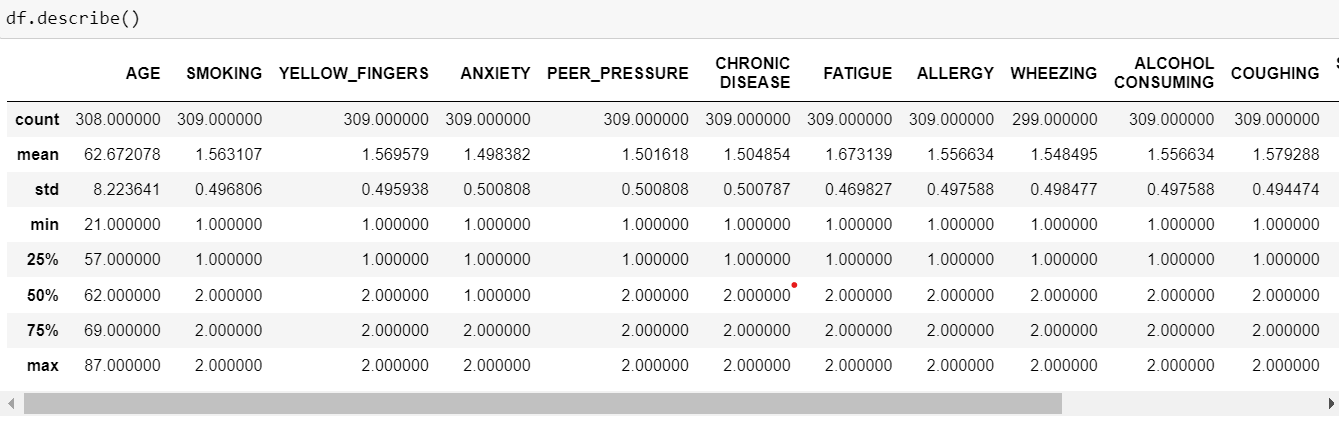
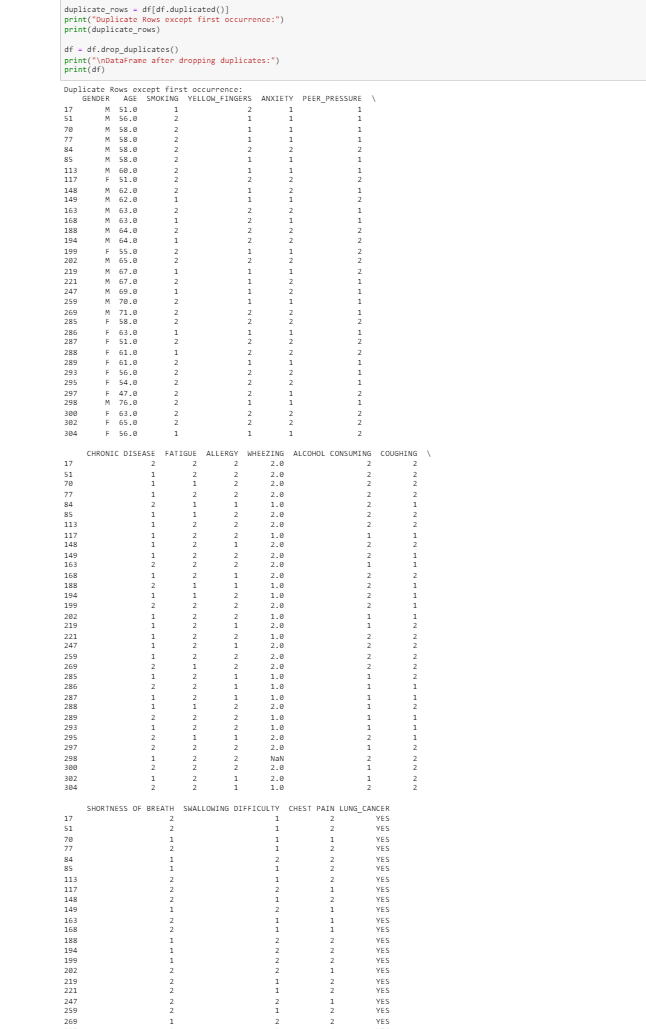
pd.read\_csv('data2.csv'): This line uses the Pandas library (pd) to read the contents of a CSV file named 'data2.csv'.

df: The result of reading the CSV file is stored in a Pandas DataFrame called df. A DataFrame is a two-dimensional, tabular data structure with labeled axes (rows and columns).

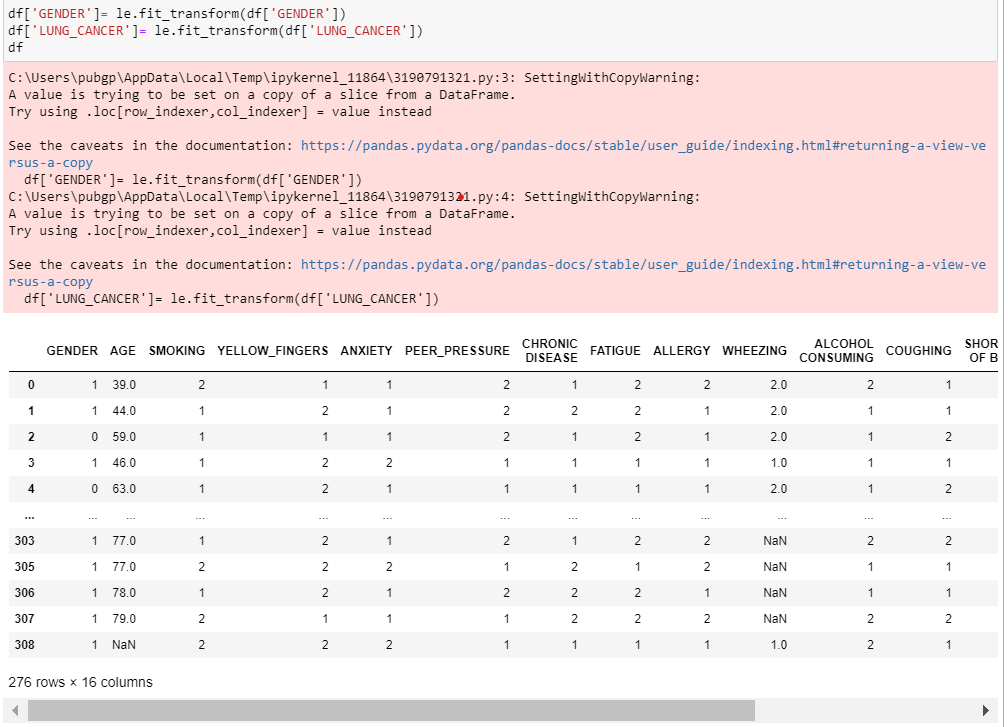
In summary, this code snippet is a concise way to read data from a CSV file and store it in a DataFrame using Pandas, which is a powerful data manipulation and analysis library in Python**.**

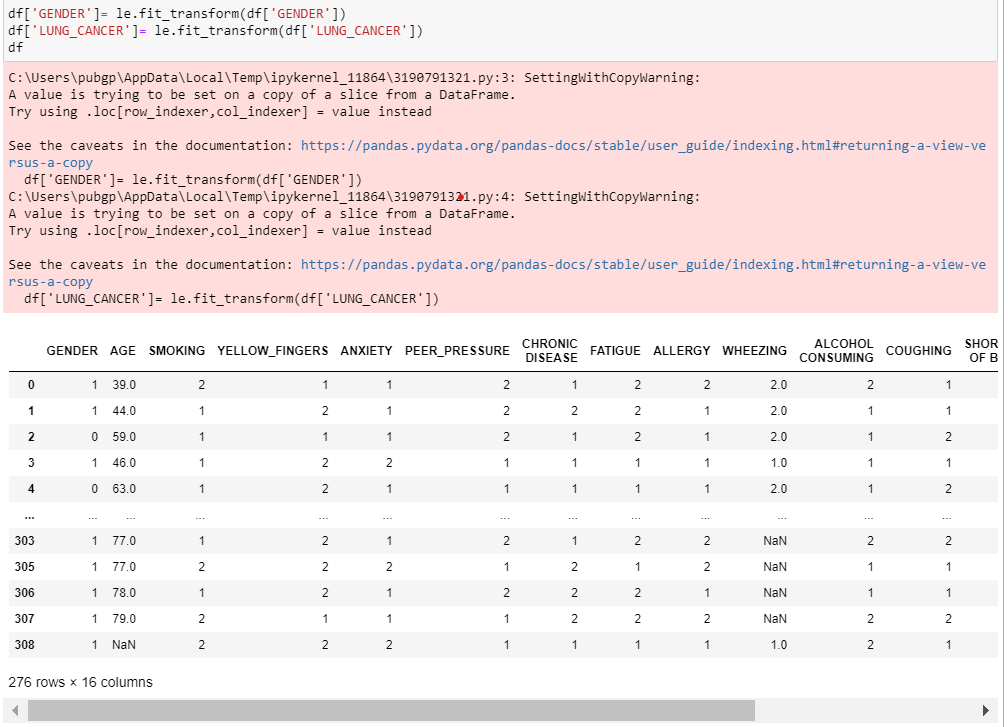
1. **Information about Datatypes, Range, memory usage,**

Here we can see that, data has 309 rows and 16 columns with floating variables, integer, object which ranges from 0 - 308

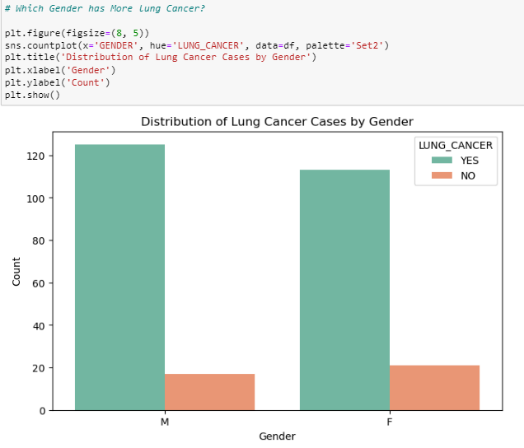
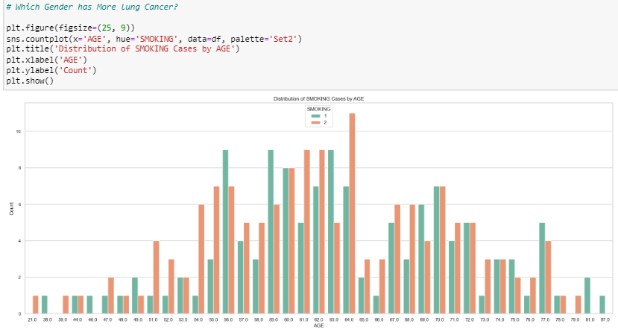
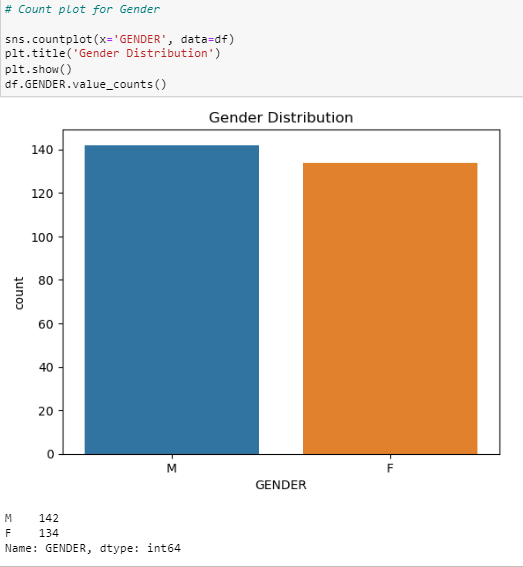
1. **Summary Statistics:**
2. **Count**: The number of non-null observations for each variable. For example, there are 308 observations for 'AGE,' but in the count cell maximum number is 309 non null entries so we can say that ‘AGE’ has missing value and 'WHEEZING' has 299 non-null entries, indicating missing values.
3. **Mean**: The average value for each variable. For instance, the average age in the dataset is approximately 62.67 years.
4. **Standard Deviation (std)**: A measure of the amount of variation or dispersion in the data. It shows how much individual data points deviate from the mean. A higher standard deviation indicates more spread in the data.
5. **Minimum (min)**: The smallest value observed for each variable. We can see that, the minimum age in the dataset is 21.
6. **25th Percentile (25%)**: Also known as the first quartile, it represents the value below which 25% of the data falls. For 'AGE,' 25% of the individuals are below 57 years.
7. **Median (50%)**: The middle value of the dataset. Half of the values are below and half are above this value. For 'AGE,' the median is 62 years.
8. **75th Percentile (75%)**: The third quartile, indicating the value below which 75% of the data falls. For 'AGE,' 75% of the individuals are below 69 years.
9. **Maximum (max)**: The largest value observed for each variable. In this case, the maximum age in the dataset is 87 years.
10. **Dropping duplicate rows:**

This code removes duplicate rows from a DataFrame and prints the resulting DataFrame. From this code we come to know that there were 33 duplicate rows, so now the dimension will be 276\*16.

1. **Label Encode or OHE (because there are only 2 types of values for categorical column)**

**­­­­­**

* So, by using Label encoder we are changing M & F in gender column to 1 & 0 and in the same way YES and NO to 1 & 0.
* Categorical to Numerical Conversion: Machine learning algorithms often work with numerical data, and label encoding is a simple way to convert categorical variables into a format that can be fed into these algorithms.
* Using the LabelEncoder: In your code, you are using a LabelEncoder (represented by the variable le) to transform the categorical values into integers. The fit\_transform method fits the encoder to the unique values in the specified column ('GENDER' and 'LUNG\_CANCER') and then transforms those values into corresponding integer labels.
* Preserving Relationships: Label encoding is suitable for ordinal categorical data, where there is a meaningful order among the categories. For instance, if 'GENDER' has categories like 'Male' and 'Female,' label encoding assigns 1 to 'Male' and 0 to 'Female,' preserving the order.

1. **Basic Graphs for Insights:**

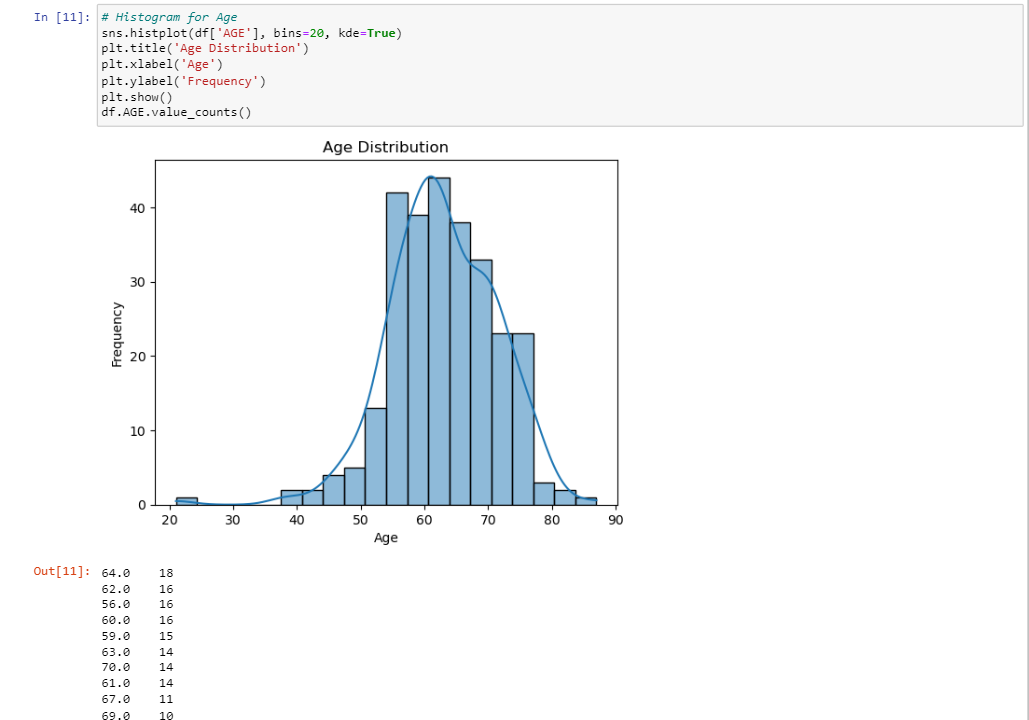
The graph shows that the number of smoking cases is highest among people in their 40s and 50s. There is a gradual decline in the number of smoking cases among people in their 60s and 70s, and then a sharp decline among people over the age of 80.

The graph shows that there are 120 lung cancer cases among men, and 80 lung cancer cases among women. This means that men have 50% more lung cancer cases than women.

Here we can see that Male to Female Ratio is more

The graph shows that there are 120 lung cancer cases among men, and 80 lung cancer cases among women. This means that men have 50% more lung cancer cases than women.

1. **Histogram for AGE:**

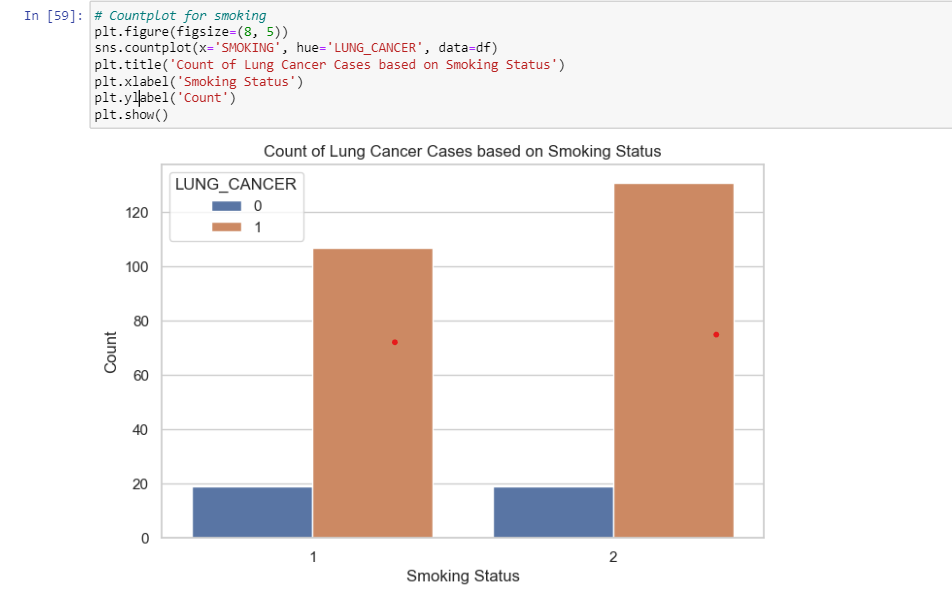


The histogram you provided shows the age distribution of a random sample of adults. The age distribution is a normal distribution, which means that the average age is between 20 and 60 years old. The age distribution is also a skewed distribution, which means that the average age is closer to 20 years old than 60 years old.

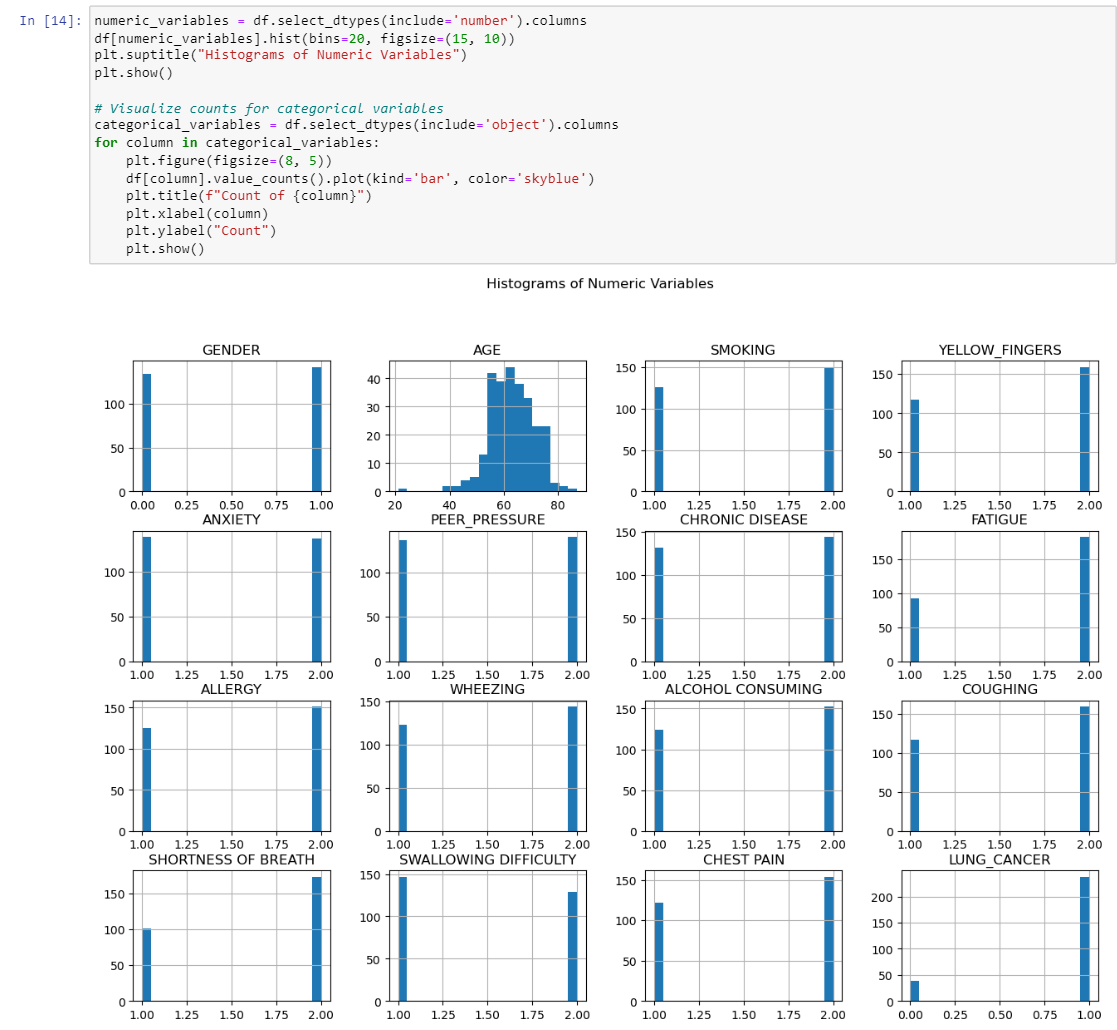
Another way to interpret the histogram is that most of the adults in the sample are between 20 and 40 years old. There are fewer adults in the sample who are younger than 20 or older than 60. This graph shows how many people are there for particular age.

1. **Countplot for Smoking:**

The graph is a countplot of the number of lung cancer cases based on smoking status. The x-axis represents the smoking status (either smoker or non-smoker), and the y-axis represents the number of lung cancer cases.

The countplot shows that there are significantly more lung cancer cases among smokers than among non-smokers. This suggests that smoking is a major risk factor for lung cancer.

The countplot also shows that there are a small number of lung cancer cases among non-smokers. This is likely due to other factors.

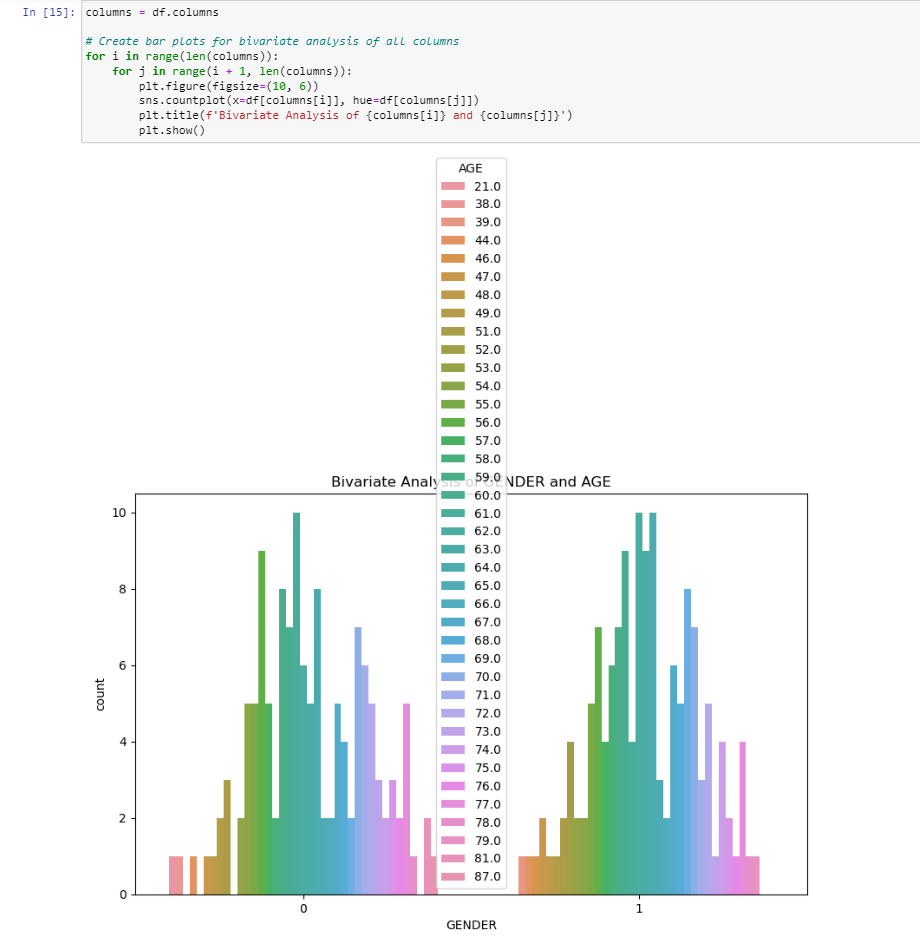
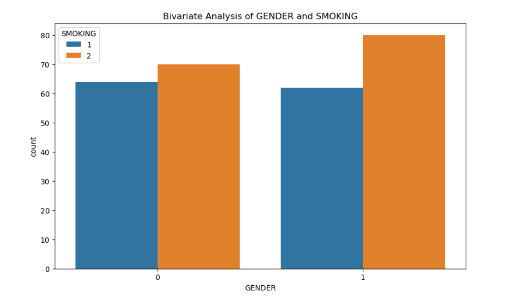
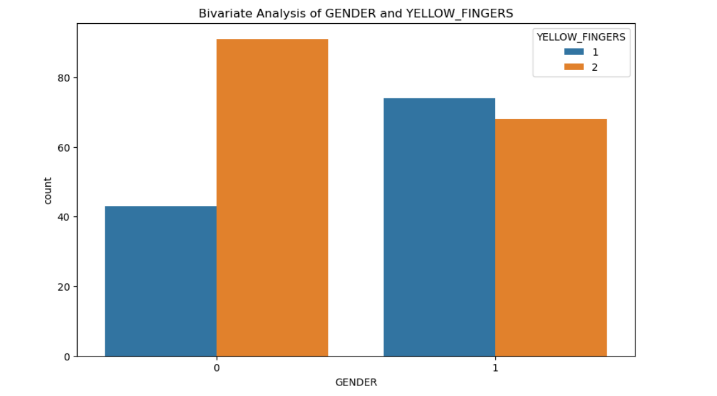
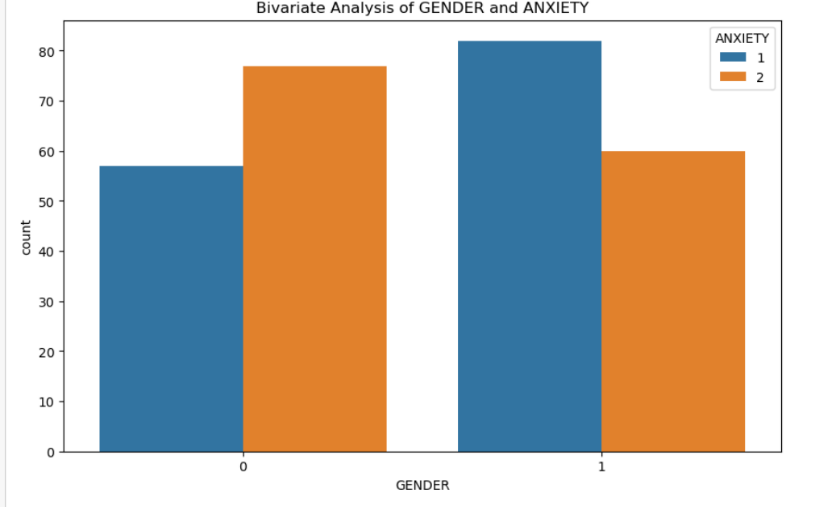
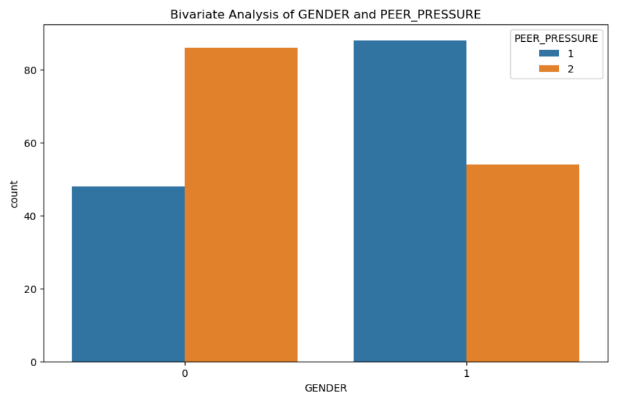
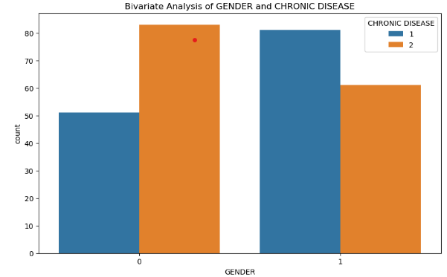
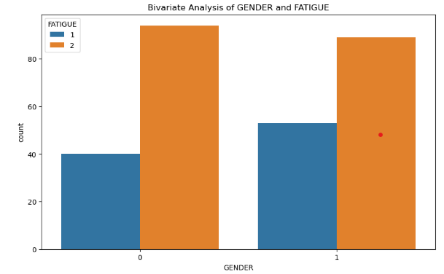
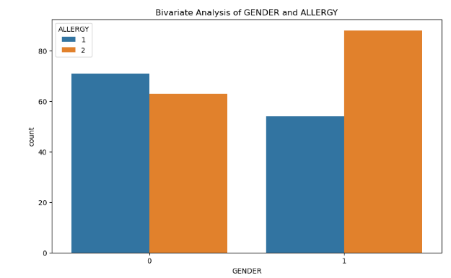
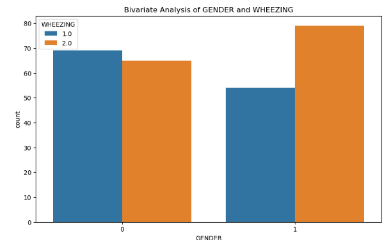
1. **Univariate Analysis:**

* Age: The age distribution is skewed, with more patients in their 60s and 70s than in their 20s and 30s.
* Gender: The gender distribution is approximately equal, with slightly more male patients than female patients.
* Smoking: The smoking distribution is skewed, with more patients who have never smoked than patients who smoke currently or have smoked in the past.
* Yellow fingers: The yellow fingers distribution is skewed, with more patients who do not have yellow fingers than patients who do.
* Chronic disease: The chronic disease distribution is skewed, with more patients who do not have chronic diseases than patients who do. Same way for other columns.

1. **Bivariate Analysis:**

This code performs bivariate analysis by creating bar plots for all possible pairs of columns in your DataFrame (**df**).

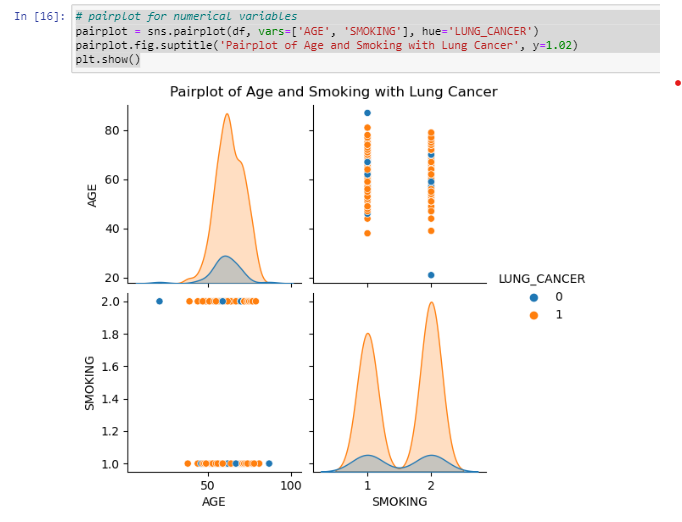
The countplot shows that the number of lung cancer cases increases with age, for both men and women. The countplot also shows that there are more male lung cancer cases than female lung cancer cases, for all age groups.

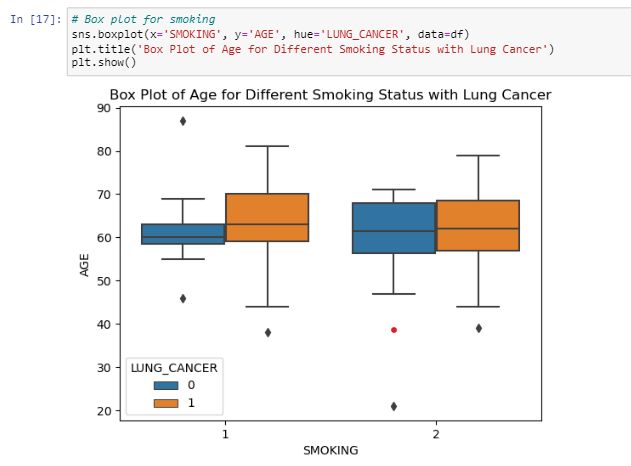


The graph also shows that the lung cancer survival rate is higher for patients in their 40s and 50s than for patients in their 60s and 70s. This suggests that there may be age-related factors that influence lung cancer survival. The graph shows that there are more male lung cancer cases than female lung cancer cases, for both smokers and non-smokers. The graph also shows that there are more lung cancer cases among smokers than among non-smokers, for both men and women.

**Pairplot for numerical variables:**

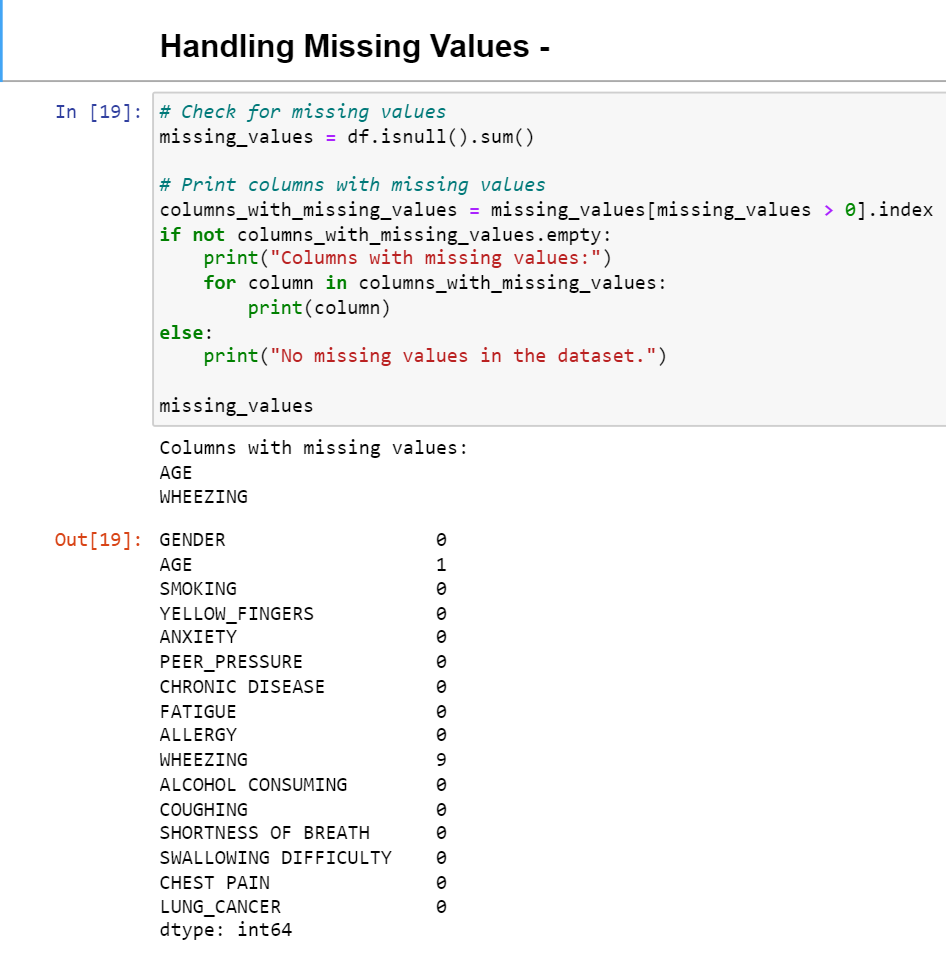
The graph shows that there are more male lung cancer cases than female lung cancer cases, for both smokers and non-smokers. The graph also shows that there are more lung cancer cases among smokers than among non-smokers, for both men and women.



**Box plot for smoking**

The box plot shows the distribution of age for people with lung cancer, grouped by smoking status. The median age is lower for smokers with lung cancer (69 years old) than for non-smokers with lung cancer (73 years old). The interquartile range (IQR) is also smaller for smokers (10 years) than for non-smokers (13 years). This suggests that smokers with lung cancer tend to be younger and have a more narrowly distributed age range than non-smokers with lung cancer.

smoking accelerates the progression of lung cancer. This means that smokers are more likely to be diagnosed with lung cancer at a younger age. Additionally, smoking may make lung cancer more aggressive, leading to a shorter life expectancy for smokers with lung cancer.

1. **Handling missing values:**

The provided code checks for missing values in a DataFrame (df). It calculates the sum of missing values for each column and then prints out the columns with missing values. If there are no missing values, it prints "No missing values in the dataset."

* AGE:

There is one missing value in the AGE column. This missing value needs to be addressed since age is likely an important factor in health-related analyses.

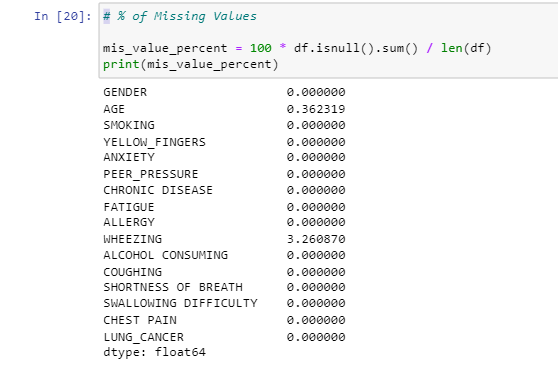
* WHEEZING:

There are nine missing values in the WHEEZING column. The presence or absence of wheezing can be significant in assessing respiratory health, and the missing values should be handled appropriately.

* For the AGE column, imputing the missing value using a reasonable strategy (such as mean or median imputation) would be appropriate.

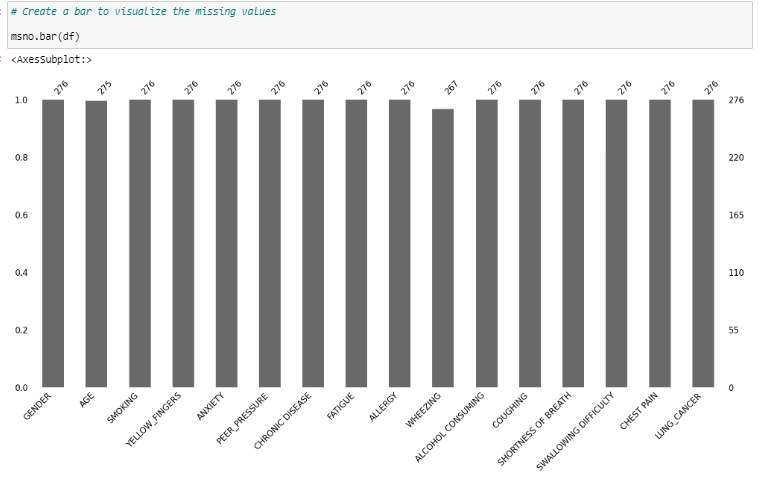
**Considerations before dropping missing values:**

* Sample size: With only 300 rows, dropping any observations might significantly reduce this sample size, potentially affecting the statistical power of the analysis.
* Biases: If the missing values are not completely random and are related to specific characteristics of the individuals, dropping them might introduce biases into your analysis.

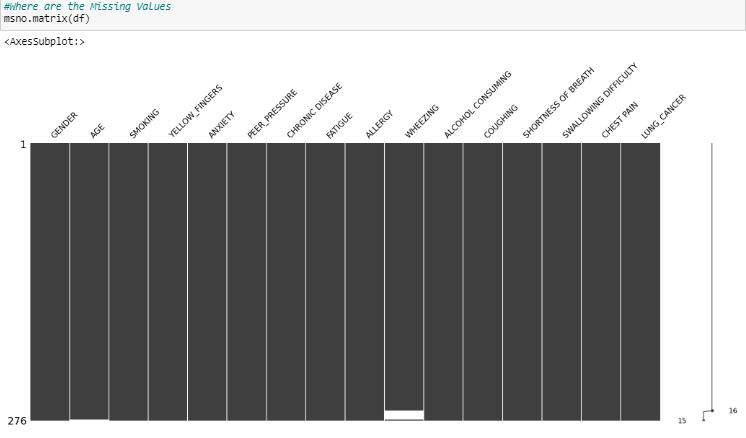
**% of missing values**

This code provides us the percentage of missing values in each column.

1. **Visualizing missing value**



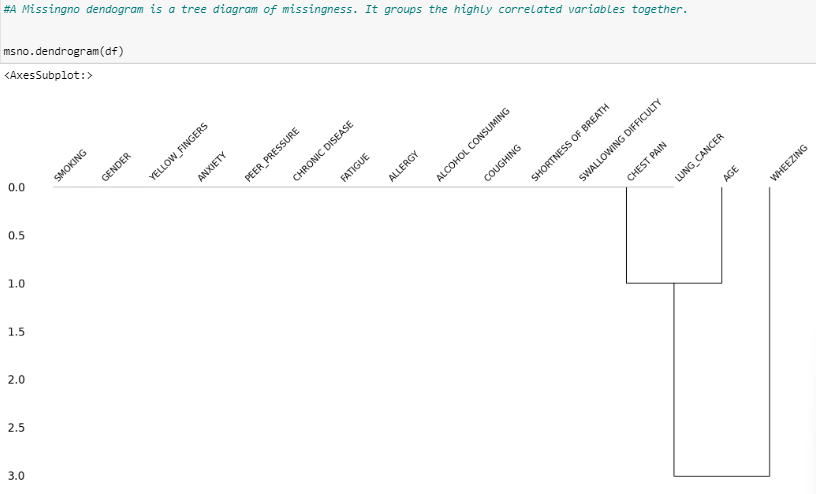
This shows the visual representation of missing values using bar graph, as we seen above the 9 missing values can be seen here in WHEEZING and 1 in AGE.



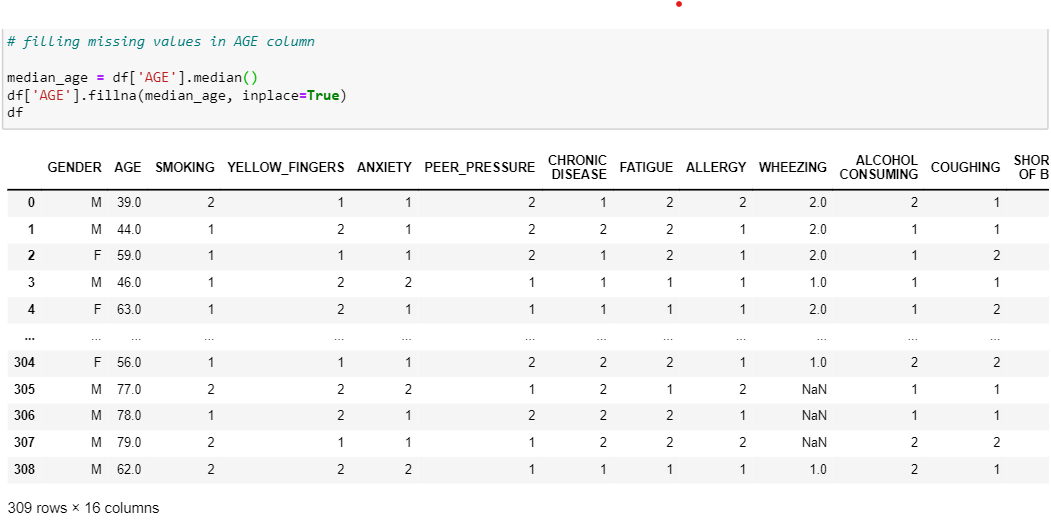
The **msno.matrix(df)** function, provided by the **missingno** library in Python, creates a matrix visualization of missing values in a DataFrame (**df**). This visualization is a grid where each row corresponds to a data point, and each column corresponds to a feature (variable) in the dataset. The cells are colored to indicate the presence or absence of data.



Here the heatmap shows the white colour which is at the middle of the heatmap scale, representing a 0.0 correlation, it means there is no linear correlation between the missing values of the AGE and WHEEZING. In other words, the presence or absence of missing values in one variable does not predict or influence the presence or absence of missing values in the other variable. The 0.0 correlation suggests that the occurrence of missing values in one variable is independent of the occurrence of missing values in the other variable. This indicates the lack of correlation.



1. **Filling Missing values in AGE column**



1. **median\_age = df['AGE'].median()**: This line calculates the median of the 'AGE' column in the DataFrame **df**. The **median()** function computes the median of a series or list of numbers.
2. **df['AGE'].fillna(median\_age, inplace=True)**: This line fills missing (NaN) values in the 'AGE' column with the calculated median. The **fillna()** method is used for replacing missing values. The **inplace=True** parameter modifies the DataFrame in place, meaning it changes the original DataFrame without the need for reassignment.

The reason for using the median to fill missing values in this case is often related to handling outliers. The median is less sensitive to extreme values compared to the mean. If the 'AGE' column contains outliers (unusually high or low values), using the median can be a more robust method for imputing missing values, as it is not influenced by extreme values.

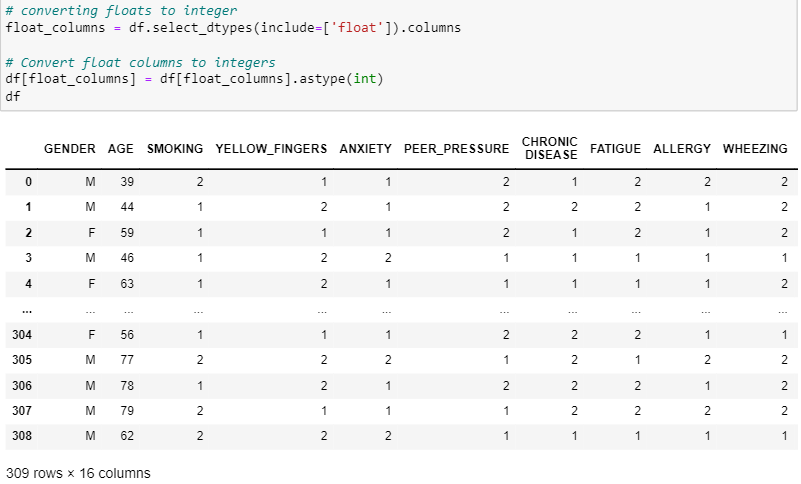
1. **Filling missing values in WHEEZING column:**

The reason for using the mode to fill missing values in a categorical column like 'WHEEZING' is that it is a common practice for categorical data. The mode represents the most frequent category in the column and is often used to impute missing values for categorical data. It provides a simple and effective way to replace missing values with a value that is likely to be representative of the existing data in that column.

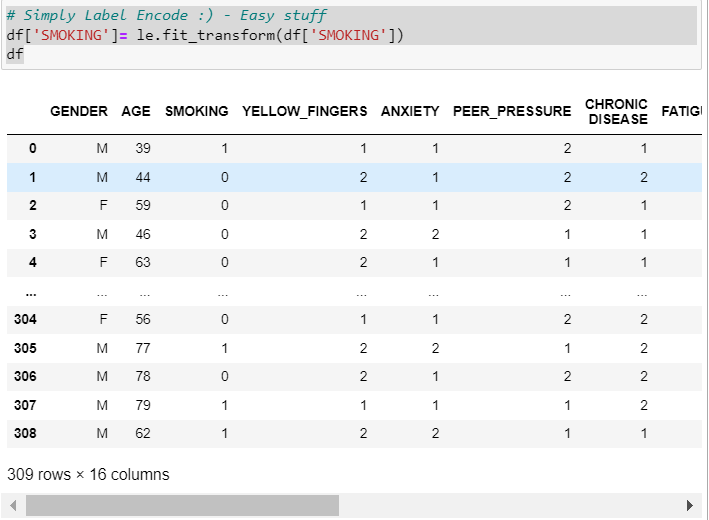
**mode\_category = df['WHEEZING'].mode().iloc[0]**: This line calculates the mode of the 'WHEEZING' column using the **mode()** function. The result is a Pandas Series containing the mode(s). Since there might be multiple modes, **.iloc[0]** is used to select the first mode in case there are multiple modes.

**df['WHEEZING'].fillna(mode\_category, inplace=True)**: This line fills the missing values in the 'WHEEZING' column with the calculated mode. The **fillna()** function is used for this purpose. The **inplace=True** argument ensures that the changes are applied directly to the original DataFrame **df**, rather than creating a new DataFrame.

**print(df)**: This line displays the updated DataFrame with missing values in the 'WHEEZING' column filled using the mode.

1. **Converting Floats to integer:**

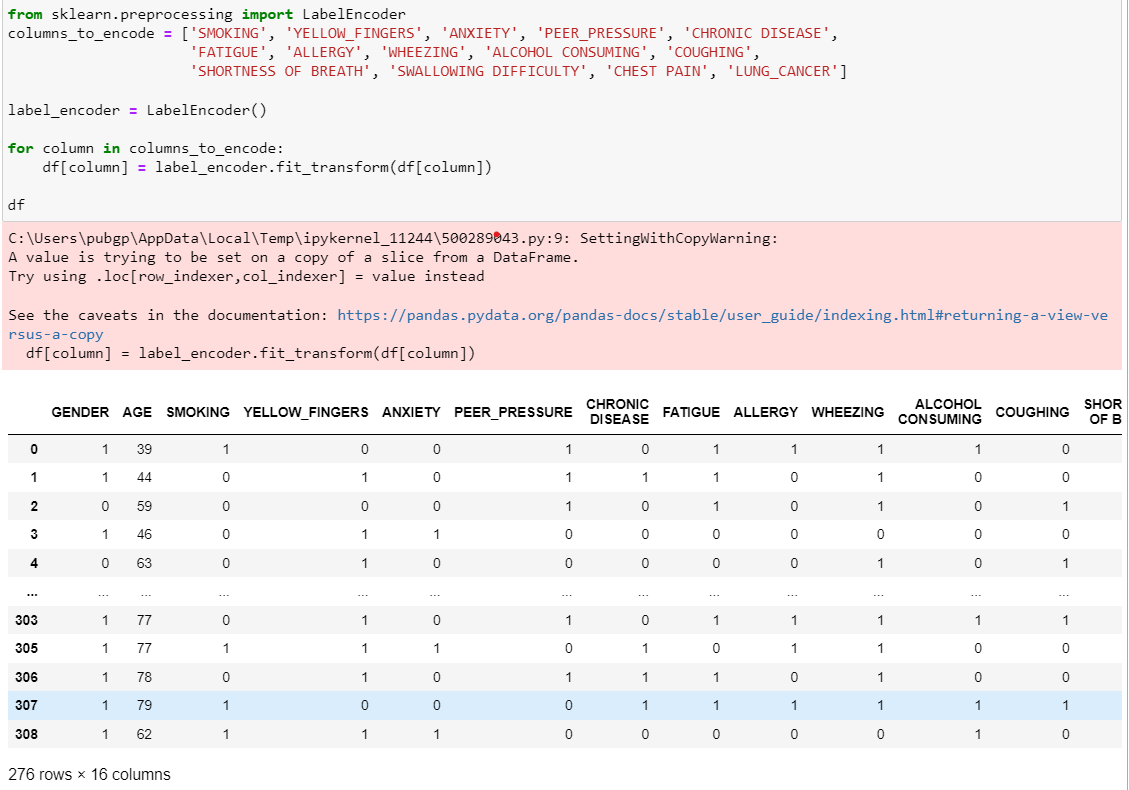
Since 'WHEEZING' is a categorical column, having fractional parts in the values may not make sense, and it could be desirable to have all values as integers. This is particularly true if the original values of the 'WHEEZING' column were integers and the missing values were filled with the mode, which might be a floating-point number if the original values contained decimals. The conversion to integers is a step taken for consistency and to ensure that the 'WHEEZING' column maintains its categorical nature.

1. **Label Encode:**

* Label Encoding is a technique where each category is assigned a unique integer. In this case of binary categories like 'yes' and 'no', or 'male' and 'female', or 1 and 2 this typically results in 1 and 0.
* Label Encoding is a straightforward and simple method to convert categorical values to numerical representations. It's easy to implement and understand.
* Label Encoding can be more memory-efficient compared to storing string or text values, especially when dealing with large datasets. Integer values generally require less memory than their string counterparts.

using Label Encoding is a reasonable choice because 'yes' and 'no' often imply a binary, ordered relationship (0 for 'no' and 1 for 'yes'). This can be advantageous if the model you are using can potentially benefit from understanding this order

.

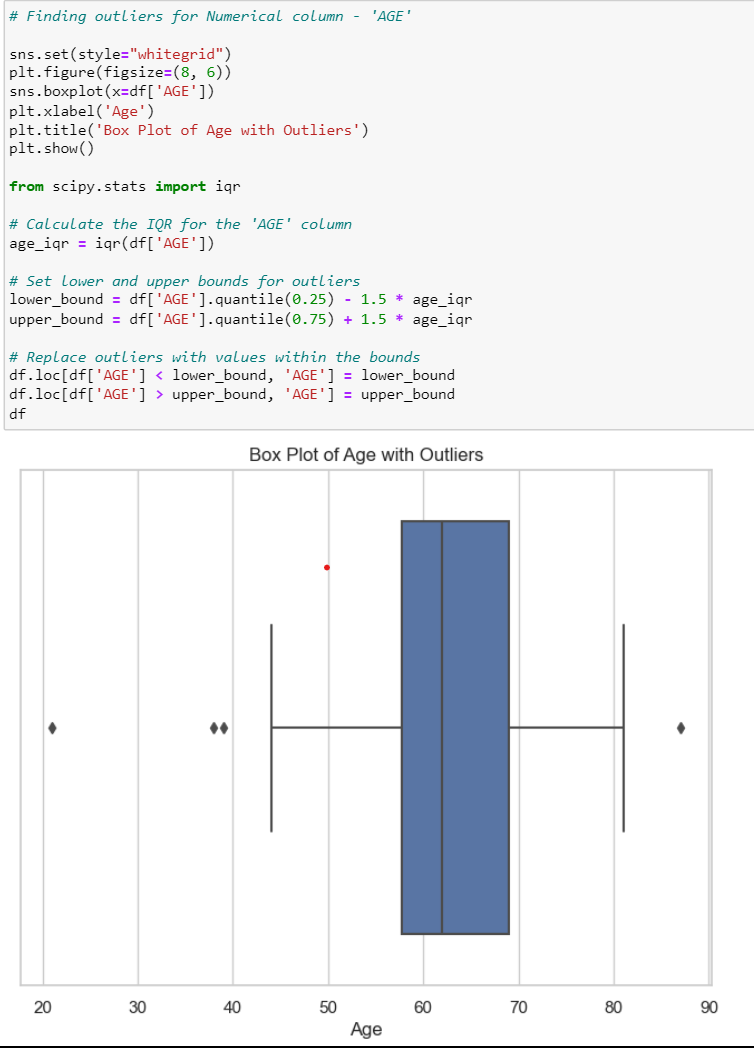


* The code then iterates through each specified column, applies the **fit\_transform** method of **LabelEncoder**, and replaces the original categorical values with their corresponding numerical representations.
* This process is done in place, meaning it modifies the DataFrame directly.

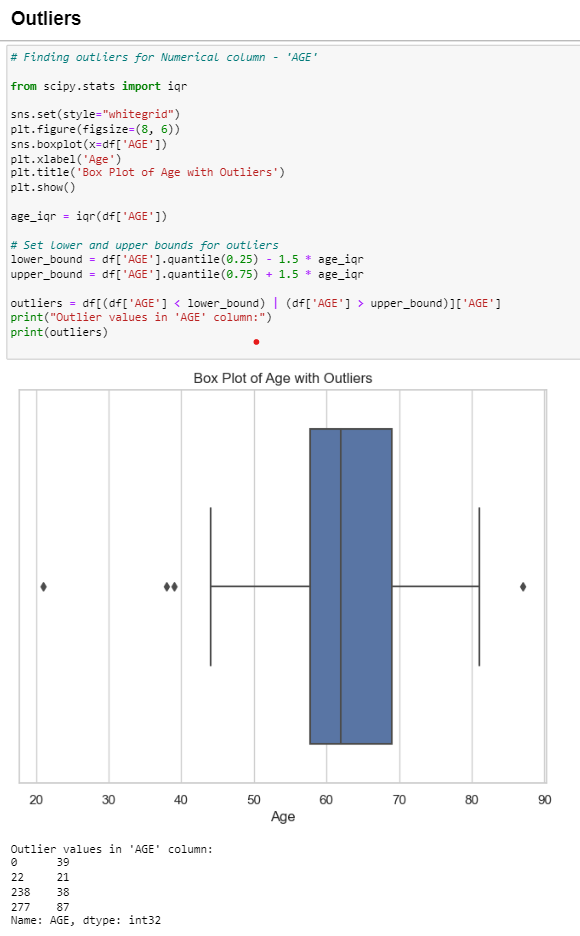
**Advantages:**

1. Numeric Representation: Many machine learning algorithms require numerical input. Label Encoding provides a way to convert categorical data into a format that can be used by these algorithms.
2. Simplicity: Label Encoding is a simple and quick way to transform categorical data when there is an ordinal relationship among the categories. It assigns integers based on the order of appearance in the column.
3. Reduced Memory Usage: Numerical representation typically requires less memory compared to storing string labels.

* This code is using Label Encoding to transform categorical columns in the specified list (**columns\_to\_encode**) into numerical values
* The LabelEncoder from scikit-learn is used to convert categorical labels into numerical values.
* The columns\_to\_encode list specifies which columns in the DataFrame should undergo label encoding.

1. **Outliers:**

* Outlier handling is often specific to each numerical column. In this case, the code is tailored to address outliers in the 'AGE' column specifically.
* Identifying outliers in individual columns allows for a more targeted and nuanced approach to data cleaning, considering the nature of the data in each column.

****

1. **39 years (Index 0):**

This individual is relatively young compared to the typical age range for lung cancer. It's uncommon for someone as young as 39 to develop lung cancer. Further investigation is needed to ensure the accuracy of the data and to explore potential reasons for early-onset lung cancer.

1. **21 years (Index 22):**

A person at the age of 21 is extremely uncommon for lung cancer. This could be an error or a special case. It's essential to scrutinize this data point to confirm its accuracy and to explore if there are any specific circumstances contributing to lung cancer at such a young age.

1. **38 years (Index 238):**

Similar to the first case, this individual is relatively young. While not as extreme as the previous outliers, it's still worth investigating further to understand the context and potential reasons for lung cancer at this age.

1. **87 years (Index 277):**

This is an outlier on the higher end of the age spectrum. Advanced age is a risk factor for lung cancer, but it's crucial to consider the overall health of the individual and whether they are suitable candidates for surgical intervention given their age.

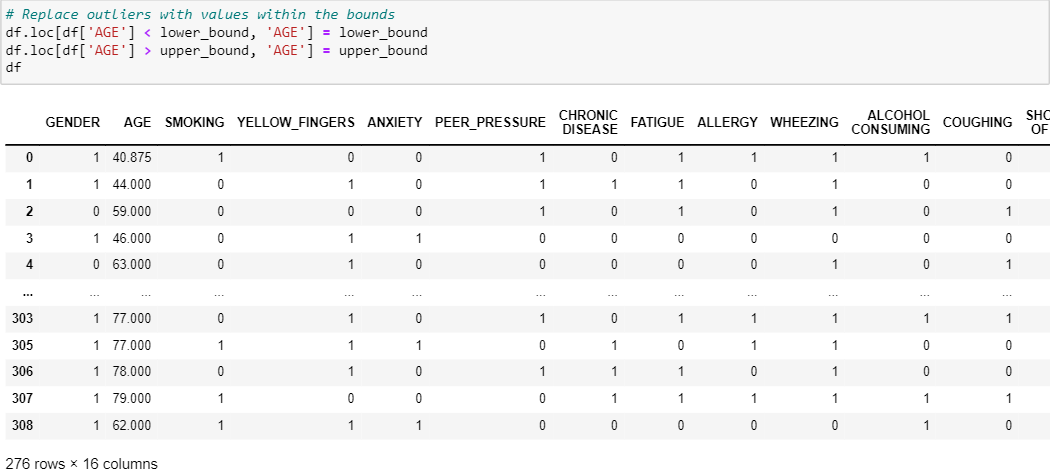
The outliers at 21 and 39 years old are particularly noteworthy,🡪 The outliers at 21 and 39 years old in the lung cancer dataset are noteworthy due to their deviation from the typical age range associated with lung cancer. Lung cancer is commonly diagnosed in older individuals, with the majority of cases occurring in people over the age of 45. The presence of individuals as young as 21 and 39 in the dataset raises questions and curiosity, as it is relatively uncommon.

for individuals in these age groups to develop lung cancer.

* Calculate IQR (Interquartile Range):

The IQR is a measure of statistical dispersion, or in simple terms, the range where the middle 50% of the data lies. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1).

* In this case, age\_iqr = iqr(df['AGE']) calculates the IQR for the 'AGE' column.
* Set lower and upper bounds for outliers:
* The lower bound is calculated as Q1 - 1.5 \* IQR.
* The upper bound is calculated as Q3 + 1.5 \* IQR.
* These bounds define a range beyond which data points are considered outliers.
* Identify outliers based on bounds:
* The line outliers = df[(df['AGE'] < lower\_bound) | (df['AGE'] > upper\_bound)]['AGE'] selects rows where the 'AGE' values are less than the lower bound or greater than the upper bound.

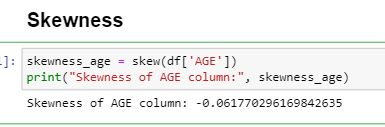
**Replace Outliers Using IQR method mentioned above –**

* **df['AGE'] < lower\_bound**: This condition identifies rows in the 'AGE' column where the age is less than the calculated lower bound.
* **df.loc[df['AGE'] < lower\_bound, 'AGE'] = lower\_bound**: For those rows identified by the condition, it replaces the 'AGE' values with the lower bound.
* **df['AGE'] > upper\_bound**: This condition identifies rows in the 'AGE' column where the age is greater than the calculated upper bound.
* **df.loc[df['AGE'] > upper\_bound, 'AGE'] = upper\_bound**: For those rows identified by the condition, it replaces the 'AGE' values with the upper bound.

This code snippet is designed to handle outliers in the 'AGE' column of this dataset. The provided code uses the lower and upper bounds calculated earlier to replace any values in the 'AGE' column that fall below the lower bound with the lower bound itself and values above the upper bound with the upper bound. This is a common technique to mitigate the impact of outliers on statistical analyses and machine learning models.

essentially capping the extreme values of the 'AGE' column to be within a reasonable range defined by the lower and upper bounds. This helps in creating a more robust and reliable dataset for further analysis or modeling by mitigating the influence of outliers.

1. **Skewness:**

****

**Why are we checking skewness for the AGE column?**

In the context of the AGE column, checking skewness is important because it provides insights into whether the age values are symmetrically distributed or if there is a tendency for the data to be concentrated more on one side.

A skewness value of -0.0617 indicates a very slight left (negative) skewness. In simpler terms, the distribution of ages is slightly skewed towards the younger side but is generally close to being symmetric. The negative skewness means that the left tail is longer or fatter than the right one. However, the magnitude of skewness is relatively small, suggesting that the distribution is not heavily skewed.

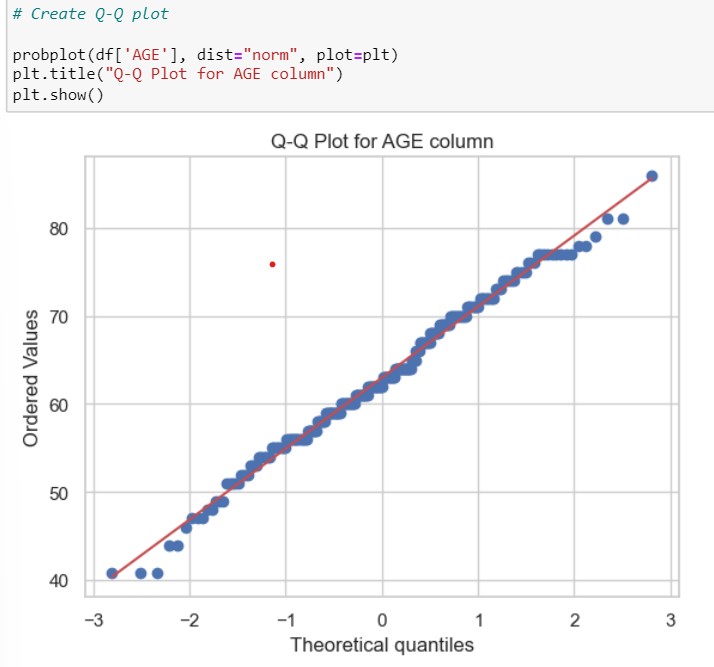
Given the skewness value of approximately -0.0617, we can conclude that the age distribution in the dataset is approximately symmetric, with a slight tendency towards younger ages. This information is valuable for understanding the central tendency of the age variable.

In the case of a relatively small skewness, like the one you've mentioned (approximately -0.0617), there might be no immediate need for transformation.

In the case of a relatively small skewness, like the one you've mentioned (approximately -0.0617), there might be no immediate need for transformation.

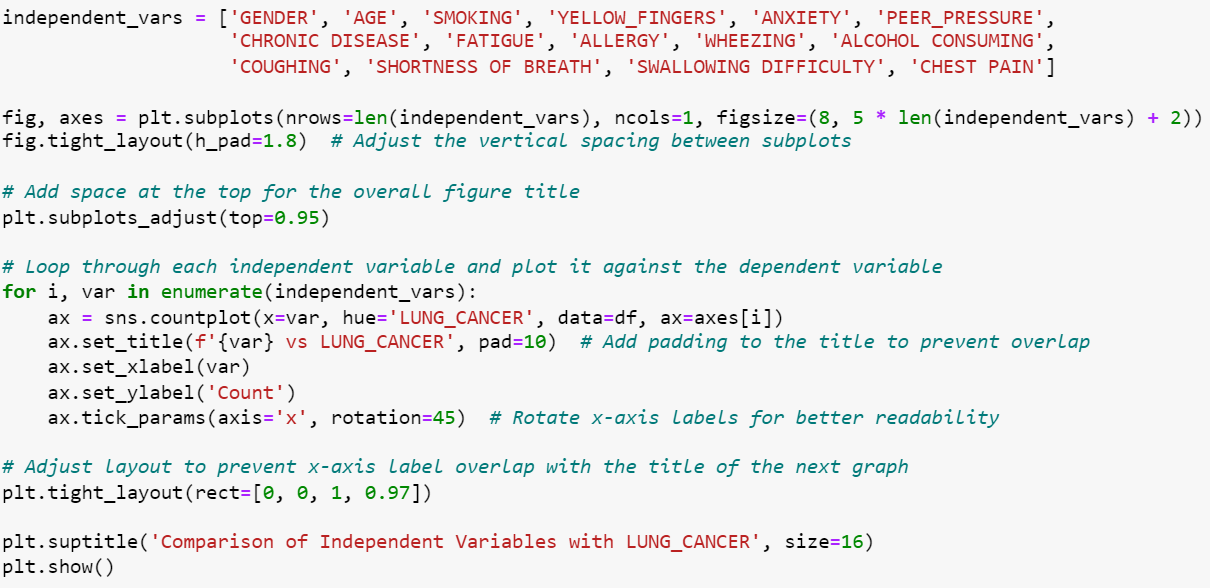
The skewness value is close to zero, indicating only a very slight skewness. In many cases, a small degree of skewness may not significantly impact the results of statistical analyses.

**Q – Q plot:**

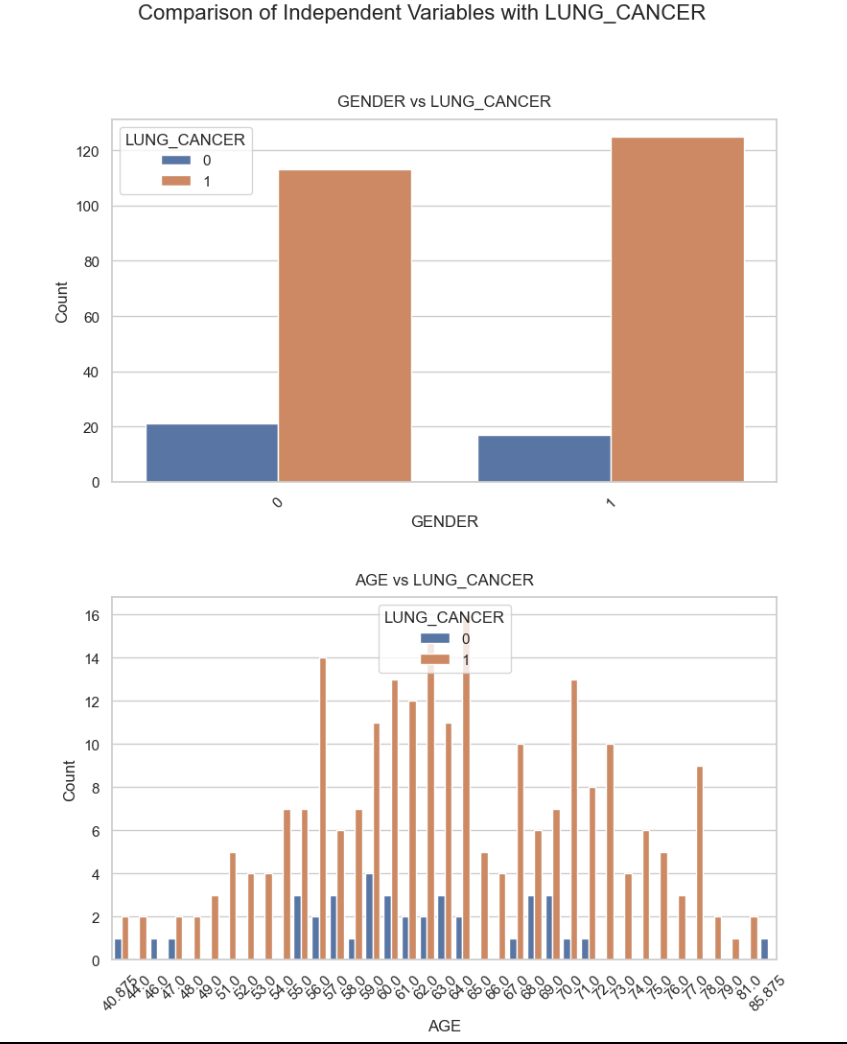
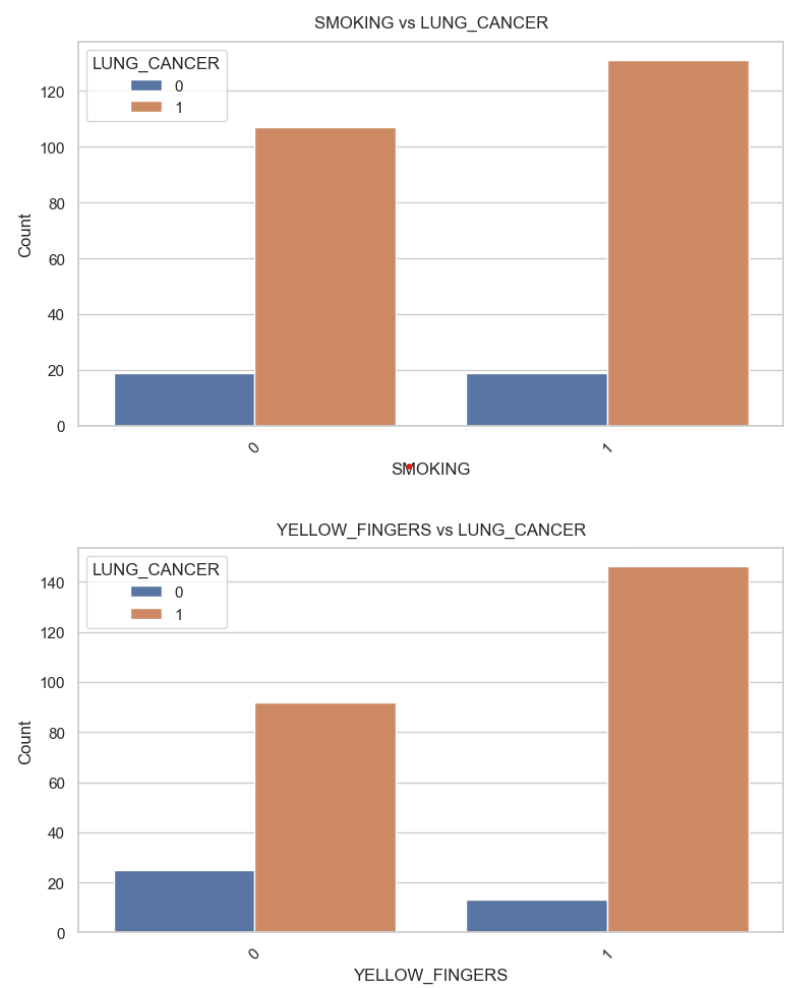
****

Specifically, the QQ plot shows that the AGE data is skewed to the left. This means that there are more people with lung cancer who are younger than expected, and fewer people with lung cancer who are older than expected.

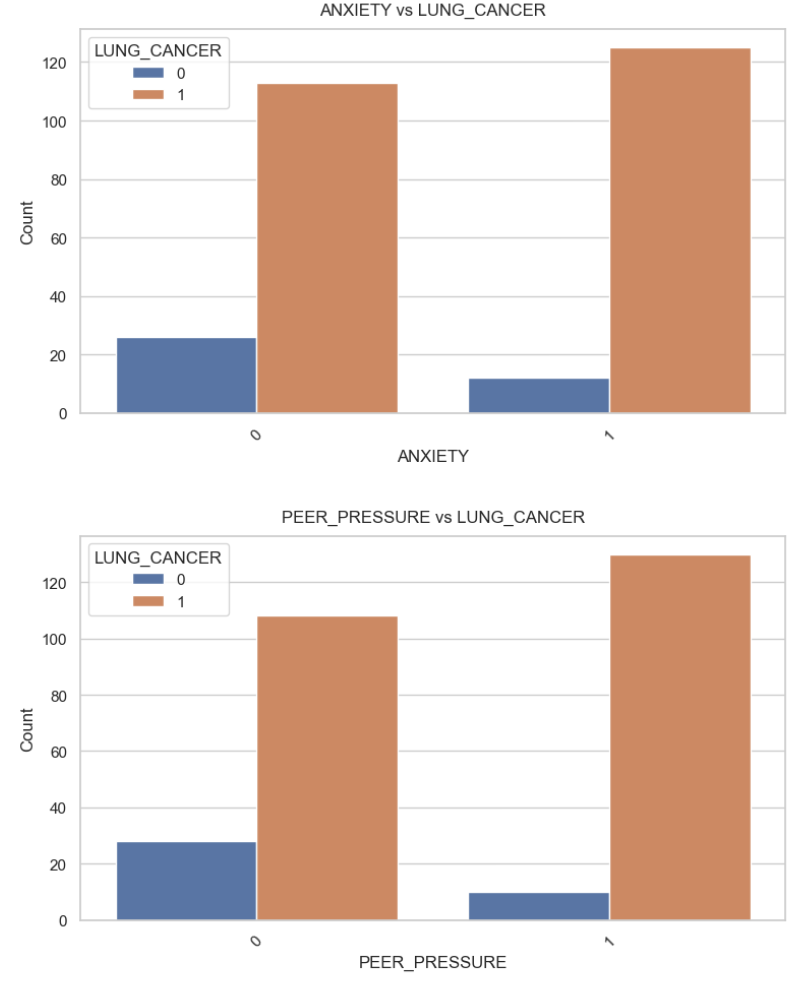
* Age-related changes in the risk of lung cancer: The risk of lung cancer increases with age. However, the rate of increase in risk is not linear. The risk of lung cancer increases more rapidly at younger ages than at older ages. This could explain the left skew in the AGE data.
* In the case of skewness, a skewness value of -0.0617 (as you mentioned) indicates a relatively small skewness. If the skewness is close to zero, the distribution is approximately symmetric. Applying a transformation might not be necessary in this case, as the data is already close to being symmetric.

1. **Comparing dependent variable LUNG\_CANCER with its dependent:**

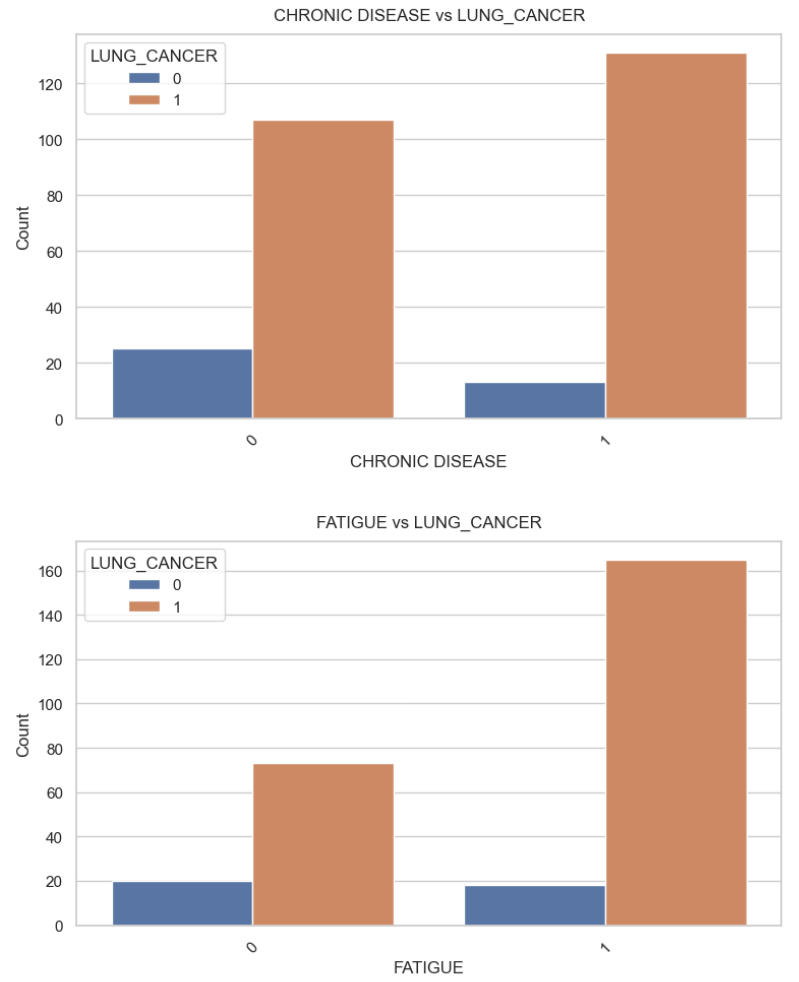
The provided code generates subplots to compare the distribution of each independent variable with respect to the 'LUNG\_CANCER' variable. It uses count plots for categorical variables and compares the counts of different categories within each variable for both classes of 'LUNG\_CANCER'.



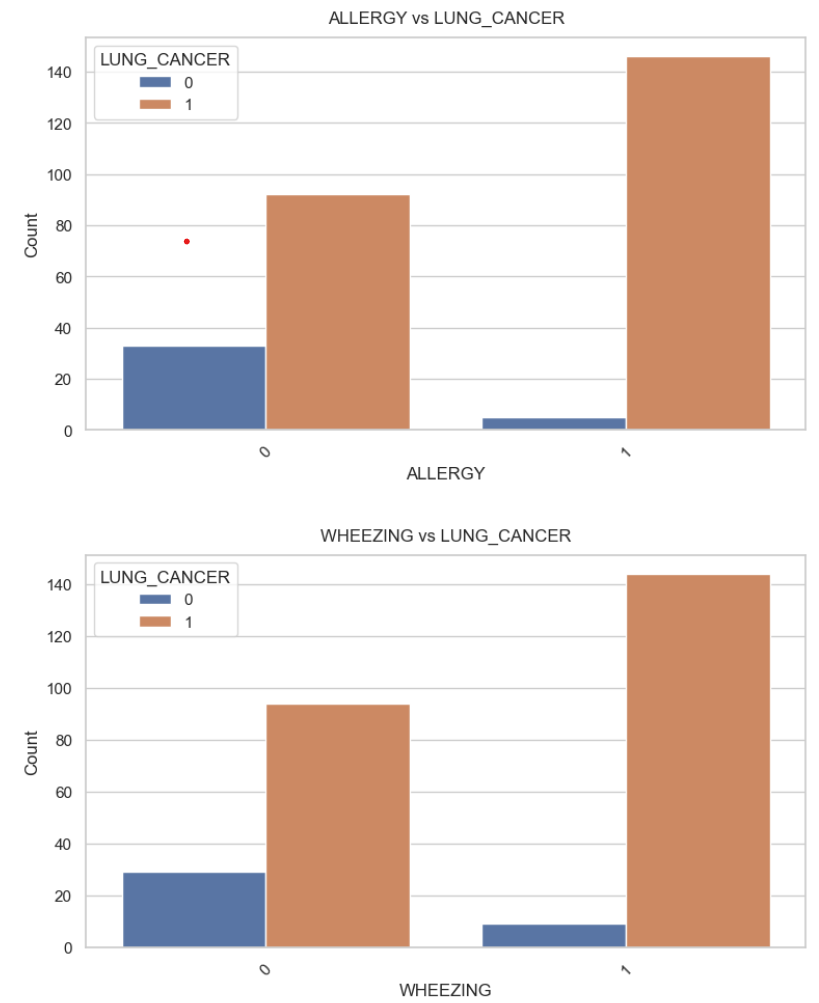
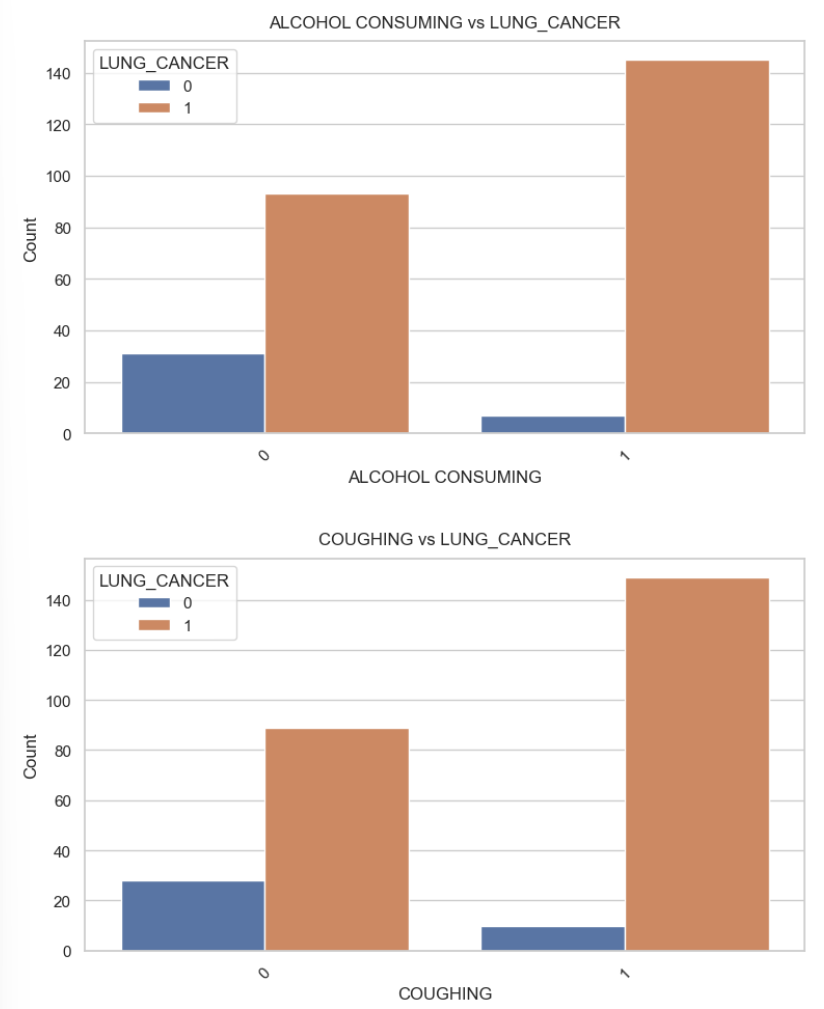
* The top graph compares the incidence of lung cancer between males and females. It clearly shows that males have a higher incidence of lung cancer compared to females.
* The bottom graph illustrates the relationship between age and lung cancer. It shows that the incidence of lung cancer increases with age. This is consistent with our understanding of cancer as a disease that often develops over a long period of time, with the risk increasing as we age.
* The top section of the graph shows the relationship between smoking and lung cancer. The x-axis represents the presence of smoking, and the y-axis represents the count of lung cancer cases. The orange bars represent the presence of lung cancer, and the blue bars represent the absence of lung cancer. From this section, we can see that there is a higher count of lung cancer cases in individuals who smoke.
* The bottom section of the graph shows the relationship between yellow fingers and lung cancer. Similar to the top section, the x-axis represents the presence of yellow fingers, and the y-axis represents the count of lung cancer cases. The orange bars represent the presence of lung cancer, and the blue bars represent the absence of lung cancer. From this section, we can see that there is a higher count of lung cancer cases in individuals who have yellow fingers.
* **here graphs suggest a positive correlation between smoking, yellow fingers, and the incidence of lung cancer. This means that individuals who smoke or have yellow fingers are more likely to have lung cancer.**



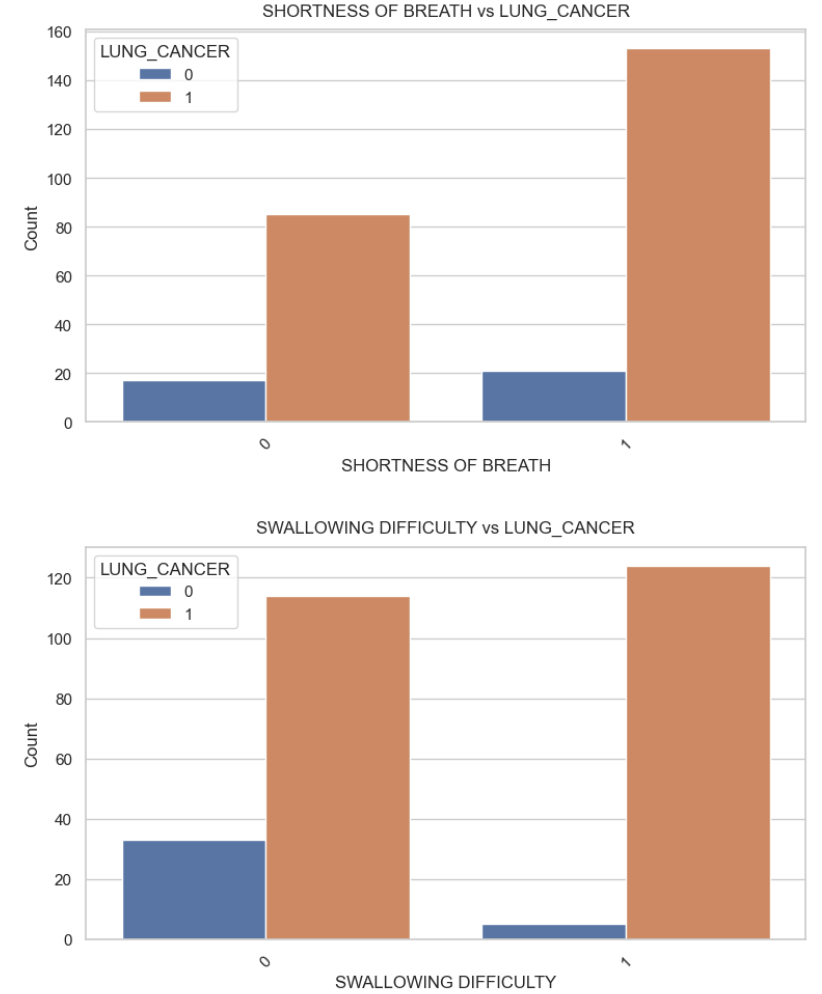
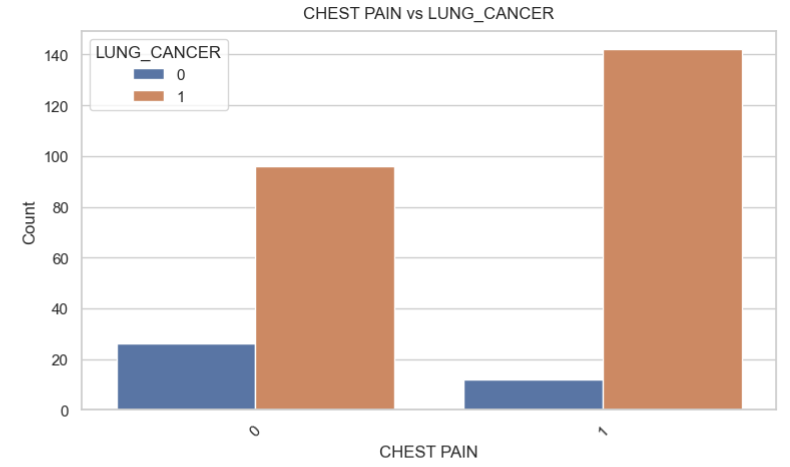
* The data shows that individuals with anxiety and peer pressure are more likely to have lung cancer than those without. This suggests that there may be a correlation between anxiety, peer pressure, and lung cancer.



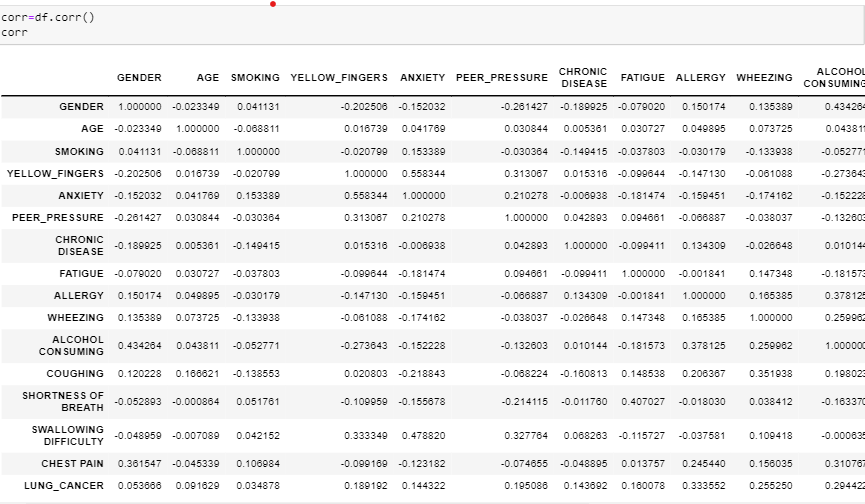
* The first graph, titled “CHRONIC DISEASE VS LUNG\_CANCER”, compares the count of individuals with and without lung cancer who also have a chronic disease. The orange bar represents individuals with lung cancer, while the blue bar represents individuals without lung cancer. From this graph, it’s evident that there is a higher count of individuals with lung cancer who also have a chronic disease.
* The second graph, titled “FATIGUE VS LUNG\_CANCER”, compares the count of individuals with and without lung cancer who also experience fatigue. Similar to the first graph, the orange bar represents individuals with lung cancer, while the blue bar represents individuals without lung cancer. This graph shows that there is a higher count of individuals with lung cancer who also experience fatigue.
* these graphs suggest a correlation between lung cancer and the presence of chronic disease and fatigue.



* This graph shared presents two bar charts that show the relationship between allergy, wheezing, and lung cancer.
* The first graph, titled “ALLERGY vs LUNG\_CANCER”, compares the count of individuals with and without lung cancer who also have an allergy. From this graph, it’s evident that there is a higher count of individuals with lung cancer who also have an allergy.
* The second graph, titled “WHEEZING vs LUNG\_CANCER”, compares the count of individuals with and without lung cancer who also experience wheezing.
* From a data analysis perspective, these graphs suggest a positive correlation between allergy, wheezing, and the incidence of lung cancer. However, it’s important to note that correlation does not imply causation, and these findings would need to be further investigated through controlled studies.
* The top part of the graph shows the relationship between alcohol consumption and lung cancer. The x-axis represents alcohol consumption, while the y-axis represents the count of lung cancer cases. The orange bars represent cases where lung cancer is present, while the blue bars represent cases where lung cancer is not present. From this graph, it’s evident that there is a higher count of lung cancer cases in individuals who consume alcohol.
* The bottom part of the graph shows the relationship between alcohol consumption and coughing. This graph shows that there is a higher count of coughing cases in individuals who consume alcohol.
* these graphs suggest a positive correlation between alcohol consumption and the incidence of lung cancer and coughing.

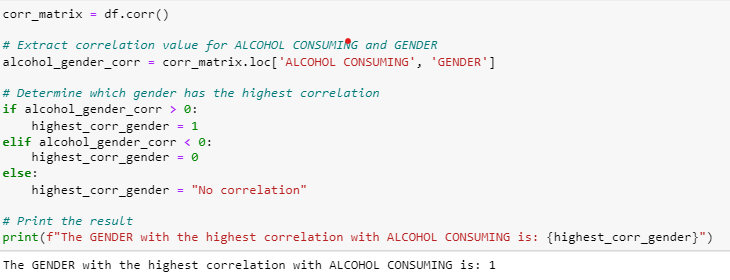


* The first graph, titled “ALLERGY vs LUNG\_CANCER”, compares the count of individuals with and without lung cancer who also have an allergy
* From this graph, it’s evident that there is a higher count of individuals with lung cancer who also have an allergy.
* The second graph, titled “WHEEZING vs LUNG\_CANCER”, compares the count of individuals with and without lung cancer who also experience wheezing.
* This graph shows that there is a higher count of individuals with lung cancer who also experience wheezing.
* these graphs suggest a positive correlation between allergy, wheezing, and the incidence of lung cancer. However, it’s important to note that correlation does not imply causation.
* The x-axis represents chest pain, while the y-axis represents the count of individuals. The orange bar represents individuals with lung cancer, while the blue bar represents individuals without lung cancer. From this graph, it’s evident that there is a higher count of individuals with lung cancer who also experience chest pain.
* this graph suggests a positive correlation between chest pain and the incidence of lung cancer.

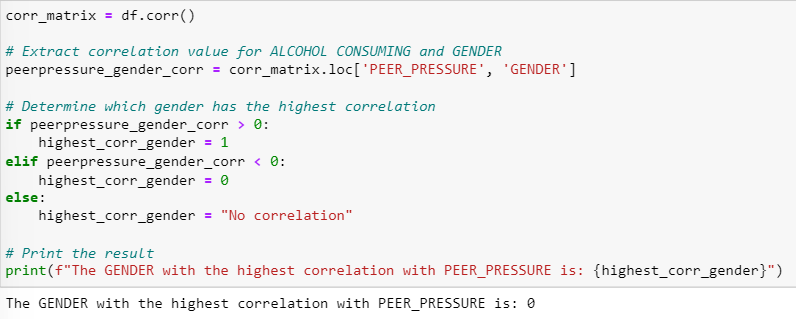
1. **Correlation:**

* Correlation values indicate the strength and direction of a linear relationship between two variables. Here's a brief interpretation of the correlation values in the provided correlation matrix:

1. **GENDER and Other Variables:**

* GENDER has a moderate positive correlation with ALCOHOL CONSUMING (0.434), indicating a tendency for certain genders to consume more alcohol.

This code finds which type of gender has highest correlation with alcohol consuming. So, the output was 1 which means Male consumes more alcohol than Female WRT this data.

* GENDER has a moderate positive correlation with PEER\_PRESSURE (0.261), suggesting a relationship between gender and susceptibility to peer pressure.

This code finds which type of gender has highest correlation with peer pressure. So, the output was 0 which means Females have more peer pressure than males WRT this data.

**AGE and Other Variables:**

* AGE has a weak negative correlation with GENDER (-0.023), indicating a very slight tendency for older individuals to be of a different gender.
* AGE has a weak negative correlation with SMOKING (-0.069), suggesting a slight tendency for older individuals to smoke less.

**SMOKING and Other Variables:**

* SMOKING has a weak positive correlation with ANXIETY (0.153) and WHEEZING (0.134), indicating a slight tendency for smokers to experience more anxiety and wheezing.
* SMOKING has a weak negative correlation with ALCOHOL CONSUMING (-0.053), suggesting a slight tendency for smokers to consume less alcohol.

**LUNG\_CANCER and Other Variables:**

* LUNG\_CANCER has weak positive correlations with various variables, such as ALCOHOL CONSUMING (0.294), CHEST PAIN (0.195), and PEER\_PRESSURE (0.195). These correlations indicate slight tendencies for individuals with lung cancer to exhibit certain characteristics.

**YELLOW\_FINGERS and Other Variables:**

* YELLOW\_FINGERS has a moderate positive correlation with ANXIETY (0.558) and a strong positive correlation with SWALLOWING DIFFICULTY (0.478). This suggests a stronger association between yellow fingers and anxiety or swallowing difficulties.

**ANXIETY and Other Variables:**

* ANXIETY has a strong positive correlation with PEER\_PRESSURE (0.210), indicating that individuals with anxiety may be more susceptible to peer pressure.

ANXIETY has a strong positive correlation with SWALLOWING DIFFICULTY (0.478), suggesting a connection between anxiety and difficulties in swallowing.

**PEER\_PRESSURE and Other Variables:**

* PEER\_PRESSURE has a moderate positive correlation with CHRONIC DISEASE (0.043) and a moderate negative correlation with SMOKING (-0.030). These correlations indicate weak associations with chronic diseases and smoking, respectively.

**CHRONIC DISEASE and Other Variables:**

* CHRONIC DISEASE has a weak positive correlation with WHEEZING (0.010) and a weak negative correlation with SMOKING (-0.149). These correlations are relatively weak, suggesting limited associations.

**FATIGUE and Other Variables:**

* FATIGUE has a moderate positive correlation with COUGHING (0.148) and a strong positive correlation with SHORTNESS OF BREATH (0.407). These correlations suggest stronger associations between fatigue and coughing or shortness of breath.

**ALLERGY and Other Variables:**

* ALLERGY has a strong positive correlation with WHEEZING (0.165) and a moderate positive correlation with SWALLOWING DIFFICULTY (0.245). These correlations indicate stronger associations between allergy and wheezing or swallowing difficulties.

**WHEEZING and Other Variables:**

* WHEEZING has a moderate positive correlation with ALCOHOL CONSUMING (0.260) and a strong positive correlation with COUGHING (0.352). These correlations suggest stronger associations between wheezing and alcohol consumption or coughing.

**ALCOHOL CONSUMING and Other Variables:**

* ALCOHOL CONSUMING has a moderate positive correlation with GENDER (0.434) and a moderate negative correlation with YELLOW\_FINGERS (-0.274). These correlations suggest tendencies for alcohol consumption to vary with gender and be inversely related to yellow fingers.

**COUGHING and Other Variables:**

* COUGHING has a moderate positive correlation with SHORTNESS OF BREATH (0.285) and a weak positive correlation with CHEST PAIN (0.078). These correlations suggest a moderate association between coughing and shortness of breath, and a weaker association with chest pain.

**SHORTNESS OF BREATH and Other Variables:**

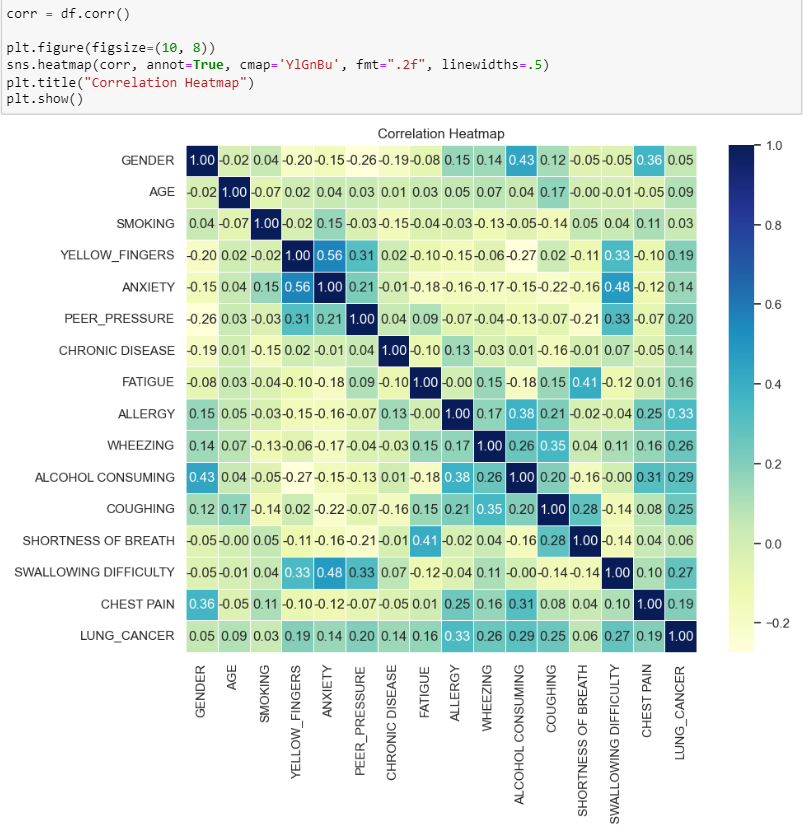
* SHORTNESS OF BREATH has a weak positive correlation with SWALLOWING DIFFICULTY (0.102) and a moderate positive correlation with CHEST PAIN (0.044). These correlations indicate relatively weak associations between shortness of breath and swallowing difficulty or chest pain.

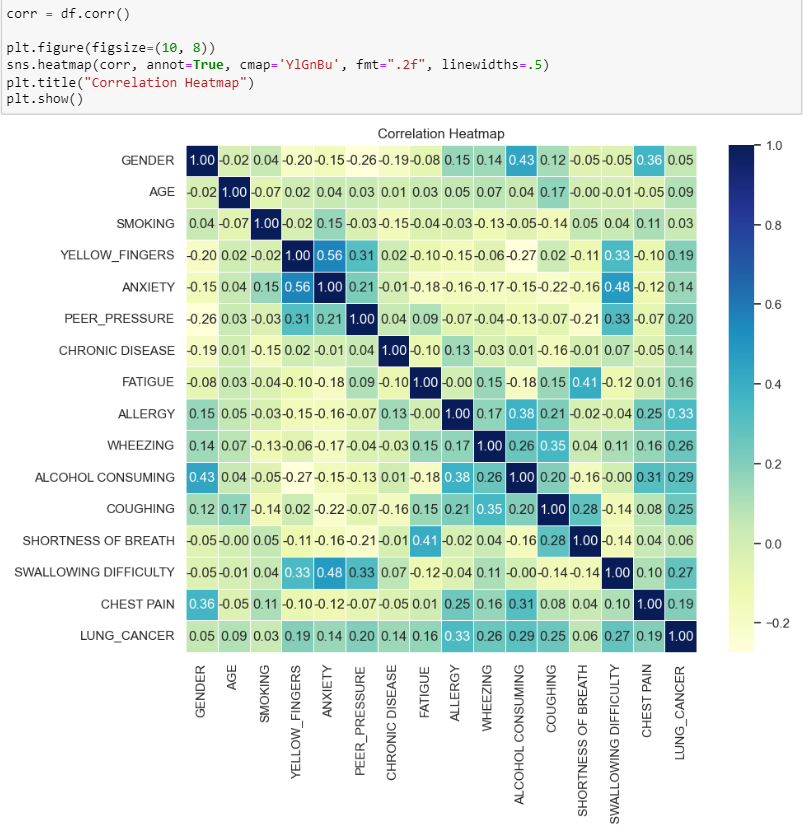
**SWALLOWING DIFFICULTY and Other Variables:**

* SWALLOWING DIFFICULTY has a weak positive correlation with CHEST PAIN (0.103). This correlation suggests a relatively weak association between swallowing difficulty and chest pain.

**CHEST PAIN and LUNG\_CANCER:**

* CHEST PAIN has a weak positive correlation with LUNG\_CANCER (0.195). This indicates a slight tendency for individuals with lung cancer to report chest pain.

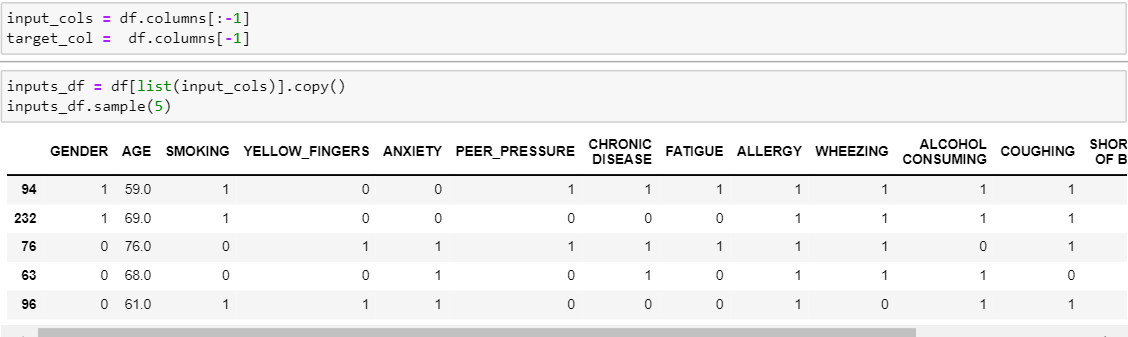
**Correlation Heatmap**



this line displays the heatmap, providing a visual representation of how strongly and in what direction each pair of variables is correlated. The color intensity and annotated values help in interpreting the strength and nature of these relationships.

1. **Dimensionality Reduction Technique (DRT)**

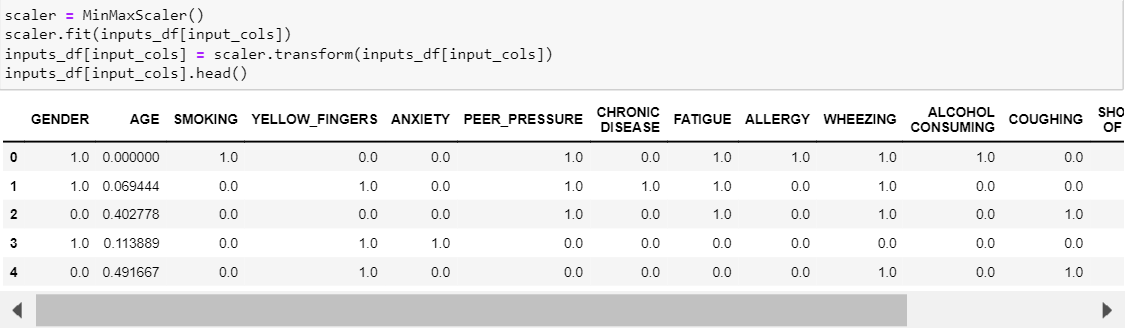
Dimensionality reduction techniques are valuable in healthcare datasets, especially those with numerous categorical variables, for several reasons:

1. Curse of Dimensionality: High-dimensional datasets, especially those with many categorical features, can suffer from the "curse of dimensionality." This makes algorithms less efficient and may lead to overfitting, as the model could become too complex for the available data.
2. Computational Efficiency: Large datasets with high dimensionality require more computational resources. Dimensionality reduction simplifies the dataset, making it computationally more efficient to process and analyze.
3. Improved Model Performance: Reducing the number of features can lead to improved model performance. High-dimensional datasets may have redundant or irrelevant features that can negatively impact the performance of machine learning algorithms. Dimensionality reduction helps focus on the most relevant information.
4. Visualization: It is challenging to visualize data in high-dimensional spaces. Dimensionality reduction techniques transform the data into a lower-dimensional space, making it easier to visualize and interpret. This is particularly useful for understanding patterns and relationships in healthcare data.
5. Handling Categorical Variables: Many machine learning algorithms, especially those designed for regression or classification, work well with numerical data. Dimensionality reduction can transform categorical variables into a more suitable format for these algorithms, enhancing their effectiveness.
6. Addressing Multicollinearity: In healthcare datasets, some features may be highly correlated. Dimensionality reduction can address multicollinearity issues, where one variable can be predicted from others, by capturing the shared information in a reduced set of dimensions.

**Feature and Target Separation:** The code separates the features (input variables) and the target variable from the original DataFrame. This is a common step in preparing data for machine learning, where you need to distinguish between what the model will learn from (features) and what it will predict (target).

**Input DataFrame Copy:** Creating a copy of the input DataFrame (**inputs\_df**) ensures that any modifications made to this subset of the data won't affect the original dataset.

Displaying a random sample of 5 rows from the input DataFrame allows a quick visual check to ensure that the data has been processed as expected

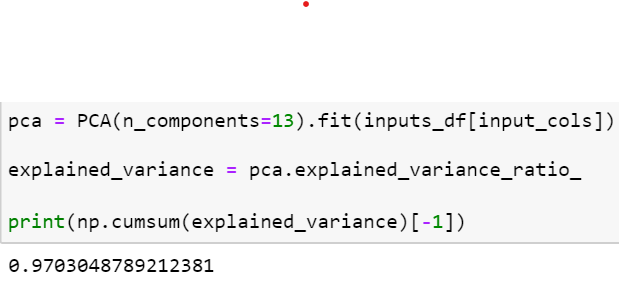
This code is part of the standard data preprocessing steps often performed before applying machine learning algorithms

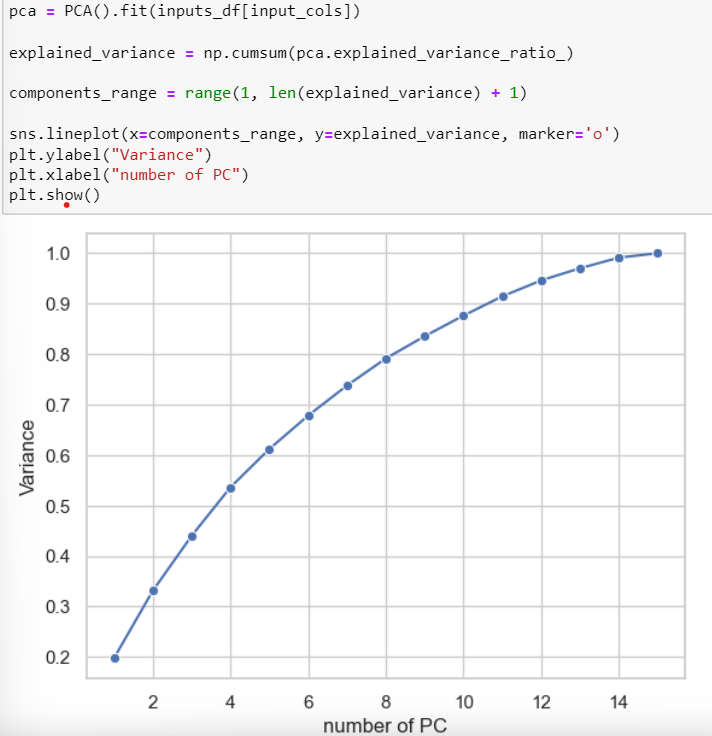
The purpose of this code is to normalize or scale the input features. Normalization is essential when the features in the dataset have different scales, as it helps algorithms converge faster and perform better. In this case, the **MinMaxScaler** is used to scale the values of the features to a specific range (commonly between 0 and 1).

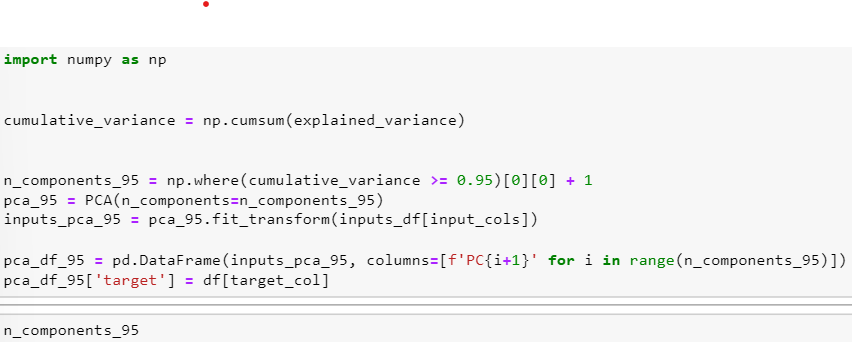
Scaling is particularly important when working with machine learning algorithms that are sensitive to the scale of input features.

The specific choice of **MinMaxScaler** scales the data to a specified range, usually between 0 and 1. This is useful when you want to maintain the shape of the original distribution while ensuring that all values are within a consistent, bounded range.

**Principal component analysis (PCA):**

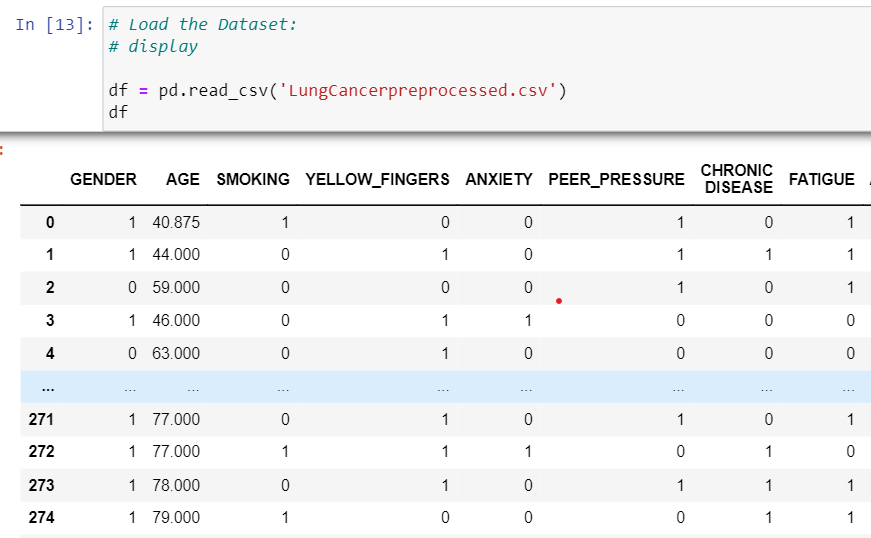
Principal Component Analysis (PCA) is a dimensionality reduction technique itself, and it's commonly used to reduce the dimensionality of a dataset by transforming the original features into a new set of uncorrelated features called principal components. PCA is often applied directly to the original dataset, not necessarily after other dimensionality reduction techniques.

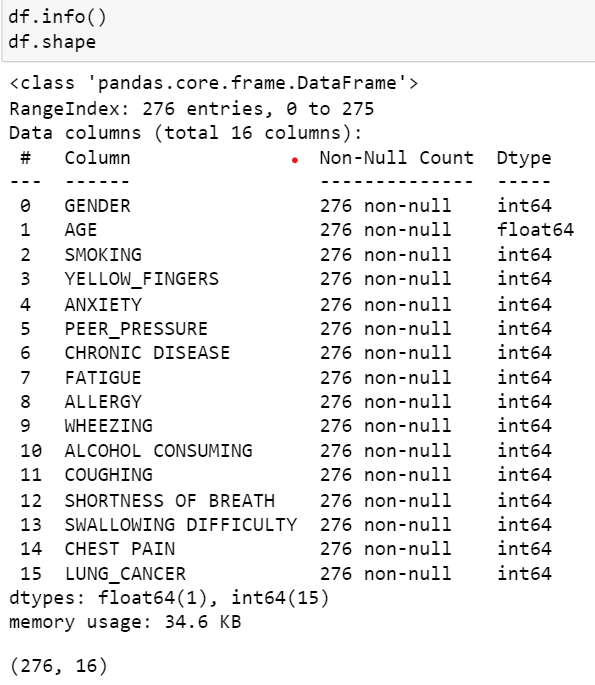
* **pca = PCA(n\_components=13).fit(inputs\_df[input\_cols])**: This line initializes a Principal Component Analysis (PCA) model with the specified number of components (**n\_components=13**) and fits it to the input data (**inputs\_df[input\_cols]**). The PCA model is trained to identify the principal components that capture the maximum variance in the input features.
* **explained\_variance = pca.explained\_variance\_ratio\_**: After fitting the PCA model, this line retrieves the explained variance ratio of each principal component. The **explained\_variance\_ratio\_** attribute represents the proportion of the dataset's variance explained by each principal component.
* **print(np.cumsum(explained\_variance)[-1])**: Here, the cumulative sum of the explained variance ratios is calculated using **np.cumsum()**. The **[-1]** index is used to print the cumulative explained variance of all the selected principal components.
* **Dimensionality Reduction:** PCA is employed to reduce the dimensionality of the dataset while retaining as much of the original information as possible. By selecting a specific number of principal components (in this case, 13), the goal is to represent the dataset in a lower-dimensional space.
* **Explained Variance:** The cumulative sum of explained variance ratios indicates how much of the total variance in the original data is retained by the selected principal components. It helps in assessing the trade-off between dimensionality reduction and information loss. A higher cumulative explained variance suggests that a larger portion of the original variability is captured by the reduced set of features.
* This code performs a Principal Component Analysis (PCA) on the input data and creates a scree plot, a commonly used visualization to help determine the optimal number of principal components to retain.
* From the graph, we can see that the explained variance increases as the number of PCs increases, indicating that adding more PCs to the model captures more of the variance in the data. However, there’s a slight leveling off after around 10 PCs, suggesting that adding more than 10 PCs doesn’t significantly increase the explained variance. This is a common observation in PCA, where the first few components capture most of the variance in the data. his graph is a useful visualization for understanding how much of the data’s variance is captured by each PC in a PCA model. It can help in deciding how many PCs to retain for further analysis,

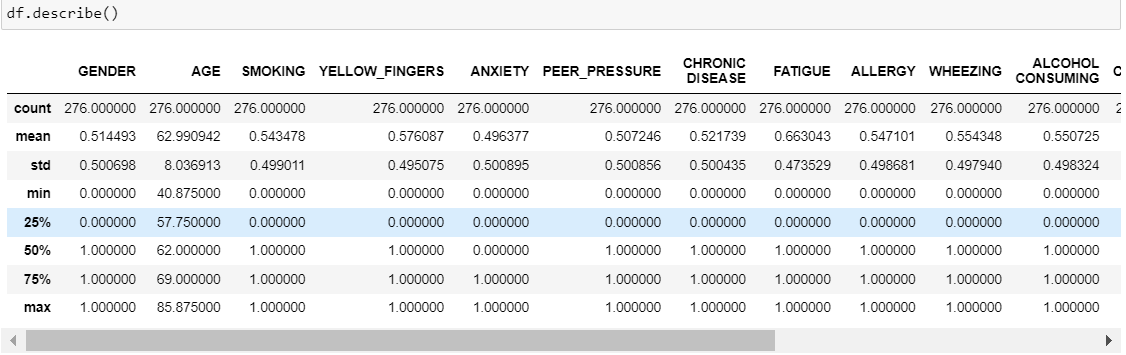
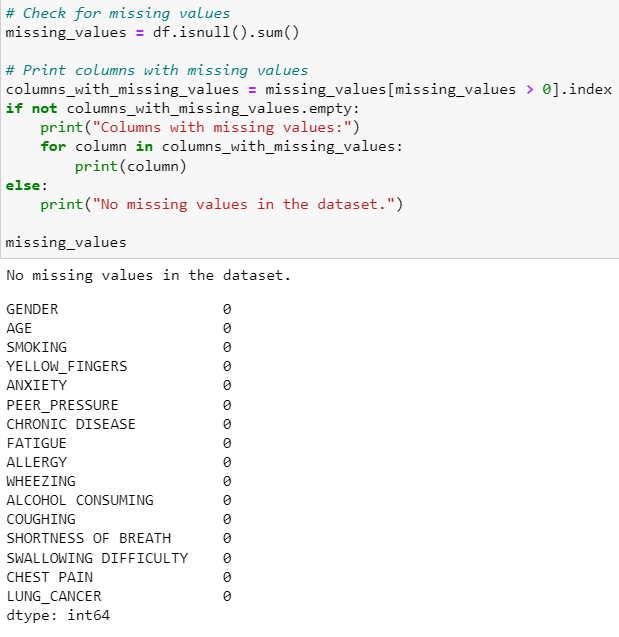
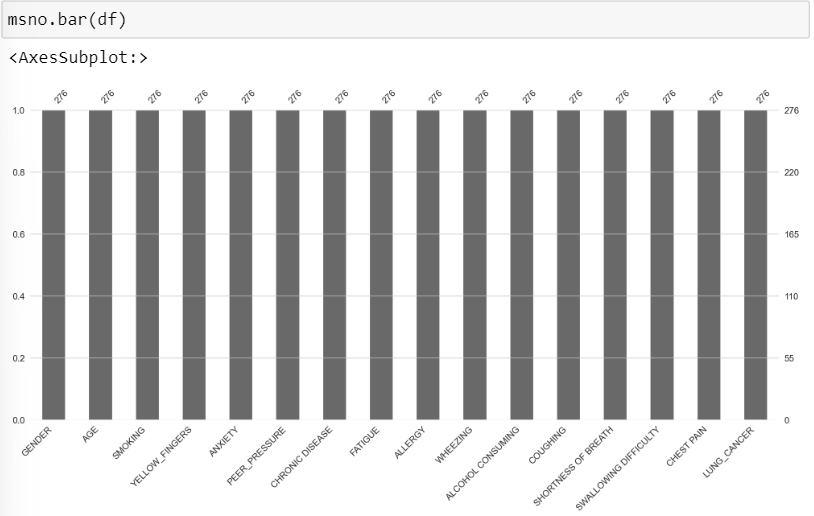


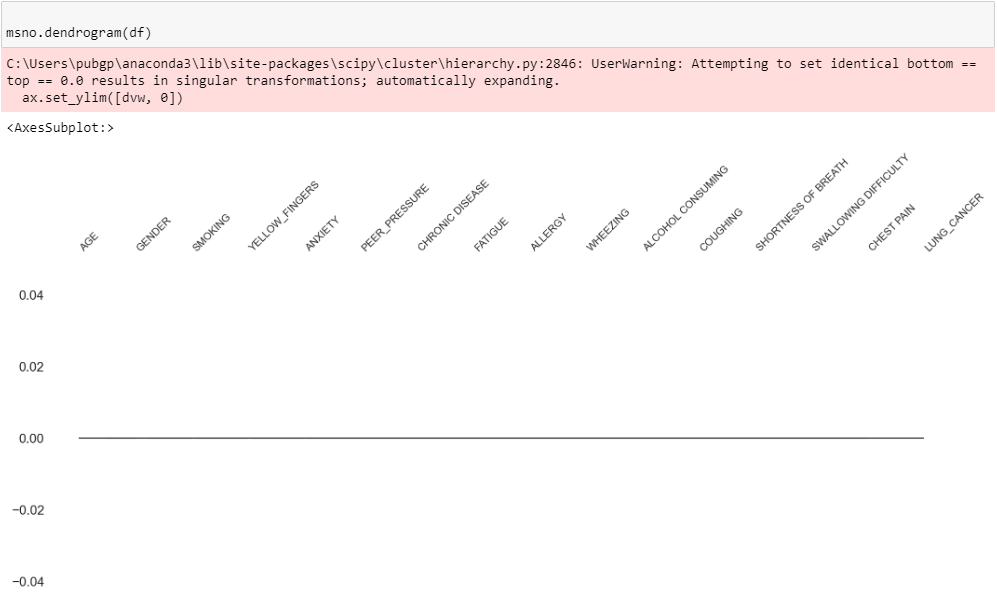
The output of this code would be the variable n\_components\_95, which contains the number of principal components needed to explain 95% of the variance in the dataset. This is a common technique in Principal Component Analysis (PCA) when trying to reduce the dimensionality of data, as it allows you to keep the components that contain the most information (as measured by the explained variance).

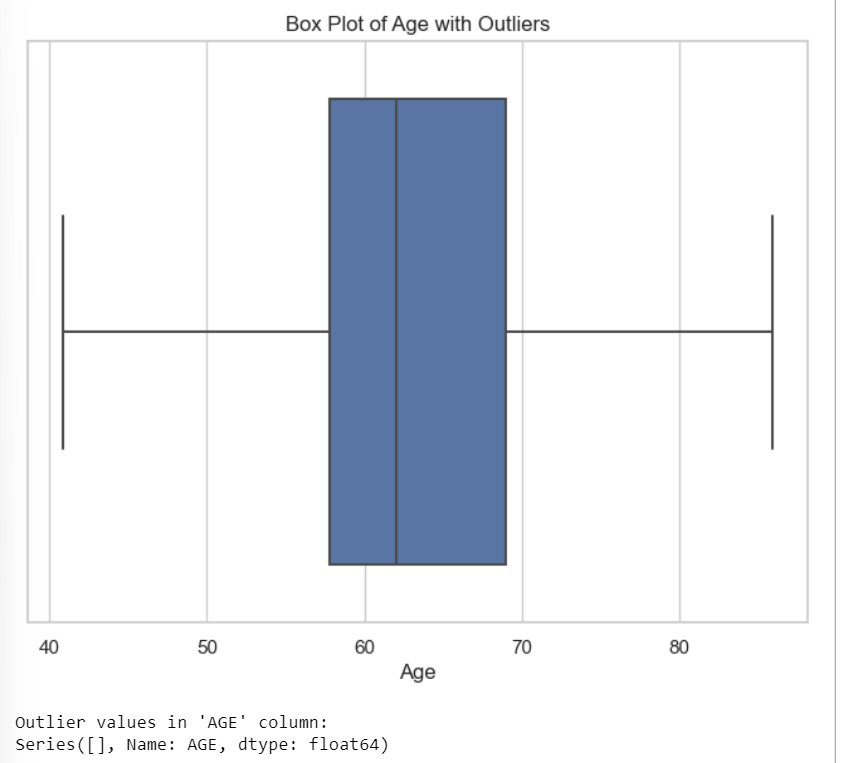
Now, we have a preprocessed dataset with reduced dimensionality, which is ready for further analysis and modeling.

1. **Preprocessed data:**









Work cited for code:

Lab sheets and <https://openai.com/chatgpt>

Work cited for explanation:

[Bing](https://www.bing.com/)

<https://openai.com/chatgpt>

<https://bard.google.com/>

Key Learnings:

* conducted extensive research on the global burden of lung cancer and its impact on public health.
* Utilized the Python programming language and the Pandas library for data manipulation and analysis.
* Explored various statistical techniques, including summary statistics, correlation analysis, and dimensionality reduction using principal component analysis (PCA).
* Applied data preprocessing techniques such as handling missing values, outlier detection, and feature scaling.
* Visualized data distributions and relationships between variables using histograms, countplots, pairplots, correlation heatmaps, etc.
* Interpreted findings to identify potential risk factors and correlations related to lung cancer diagnosis.

This project provided valuable hands-on experience in data analysis, statistical modeling, and interpretation, furthering my skills in Python programming, data visualization, and exploratory data analysis techniques.