LUNG CANCER - SURVEY ANALYSIS

Uncontrolled growth of cells in the body is cancer. It is called lung cancer when it begins in the lungs. The most common form of cancer is lung cancer. In the United States, lung cancer ranks second among both men and women, excluding some kinds of skin cancer. As fewer people smoke cigarettes and as lung cancer treatments improve, the national lung cancer rate is decreasing after increasing for decades. Lung cancer symptoms may vary from person to person. The majority of people with lung cancer do not experience symptoms until the cancer is advanced.

About The Project

My project contains a dataset listing the main symptoms of lung cancer in the world. In this project, we will analyze the relationship between the disease, sex and the symptoms of the disease. Moreover, we will also analyze some graphical representations to obtain more information. With this, we will be able help doctors to improve their treatment on cancer and also draw useful insights from the data.

```
!pip install jovian opendatasets --upgrade --quiet
```

Let's begin by downloading the data, and listing the files within the dataset.

```
# Change this
dataset_url = 'https://www.kaggle.com/datasets/nancyalaswad90/lung-cancer'
```

```
import opendatasets as od
od.download(dataset_url)
```

```
Skipping, found downloaded files in "./lung-cancer" (use force=True to force download)
```

The dataset has been downloaded and extracted.

```
# Change this
data_dir = './lung-cancer'
```

```
import os
os.listdir(data_dir)
```

```
['survey lung cancer.csv']
```

Let us save and upload our work to Jovian before continuing.

```
project_name = "lung cancer" # change this (use lowercase letters and hyphens only)
```

```
!pip install jovian --upgrade -q
```

```
import jovian
```

```
jovian.commit(project=project_name)
```

```
[jovian] Updating notebook "sainihaldiddi2002/lung-cancer" on https://jovian.ai
[jovian] Committed successfully! https://jovian.ai/sainihaldiddi2002/lung-cancer
```

```
'https://jovian.ai/sainihaldiddi2002/lung-cancer'
```

Data Preparation and Cleaning

TODO - Write some explanation here.

> Instructions (delete this cell):
>
> - Load the dataset into a data frame using Pandas
> - Explore the number of rows & columns, ranges of values etc.
> - Handle missing, incorrect and invalid data
> - Perform any additional steps (parsing dates, creating additional columns, merging multiple dataset etc.)

```
import pandas as pd
import numpy as np
```

```
lung_df = pd.read_csv('survey lung cancer.csv')
lung_df
```

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEE
0	M	69	1	2	2	1	1	2	1	
1	M	74	2	1	1	1	2	2	2	
2	F	59	1	1	1	2	1	2	1	
3	M	63	2	2	2	1	1	1	1	
4	F	63	1	2	1	1	1	1	1	
...	...	...	...	...	...	...	...	...	...	
304	F	56	1	1	1	2	2	2	1	
305	M	70	2	1	1	1	1	2	2	
306	M	58	2	1	1	1	1	1	2	
307	M	67	2	1	2	1	1	2	2	
308	M	62	1	1	1	2	1	2	2	

309 rows × 16 columns

The dataset contains gender, age and the causes of lung cancer. Around 309 people have accepted that they have faced each of the issues given in the dataset. The numbers 1 and 2 represent no and yes respectively. We can replace that by using numpy library.

```
lung_df = pd.read_csv('survey lung cancer.csv')
# replacing all the 1s and 2s with NOs and YESs respectively
lung_df['SMOKING'] = np.where(lung_df['SMOKING']==1, 'NO','YES')
lung_df["YELLOW_FINGERS"] = np.where(lung_df["YELLOW_FINGERS"]==1, 'NO', 'YES')
lung_df["ANXIETY"] = np.where(lung_df["ANXIETY"]==1, 'NO', 'YES')
lung_df["PEER_PRESSURE"] = np.where(lung_df["PEER_PRESSURE"]==1, 'NO', 'YES')
lung_df["CHRONIC DISEASE"] = np.where(lung_df["CHRONIC DISEASE"]==1, 'NO', 'YES')
lung_df["FATIGUE "] = np.where(lung_df["FATIGUE "]==1, 'NO', 'YES')
lung_df["ALLERGY "] = np.where(lung_df["ALLERGY "]==1, 'NO', 'YES')
lung_df["WHEEZING"] = np.where(lung_df["WHEEZING"]==1, 'NO', 'YES')
lung_df["ALCOHOL CONSUMING"] = np.where(lung_df["ALCOHOL CONSUMING"]==1, 'NO', 'YES')
lung_df["COUGHING"] = np.where(lung_df["COUGHING"]==1, 'NO', 'YES')
lung_df["SHORTNESS OF BREATH"] = np.where(lung_df["SHORTNESS OF BREATH"]==1, 'NO', 'YES
lung_df["SWALLOWING DIFFICULTY"] = np.where(lung_df["SWALLOWING DIFFICULTY"]==1, 'NO',
lung_df["CHEST PAIN"] = np.where(lung_df["CHEST PAIN"]==1, 'NO', 'YES')
lung_df
```

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZ
0	M	69	NO	YES	YES	NO	NO	YES	NO	
1	M	74	YES	NO	NO	NO	YES	YES	YES	
2	F	59	NO	NO	NO	YES	NO	YES	NO	
3	M	63	YES	YES	YES	NO	NO	NO	NO	
4	F	63	NO	YES	NO	NO	NO	NO	NO	
...	...	...	...	...	...	...	...	...	...	
304	F	56	NO	NO	NO	YES	YES	YES	NO	
305	M	70	YES	NO	NO	NO	NO	YES	YES	
306	M	58	YES	NO	NO	NO	NO	NO	YES	
307	M	67	YES	NO	YES	NO	NO	YES	YES	
308	M	62	NO	NO	NO	YES	NO	YES	YES	

309 rows × 16 columns

```
lung_df.drop(["SMOKING","CHRONIC DISEASE","ALLERGY ","PEER_PRESSURE",'ALCOHOL CONSUMING
lung_df
```

	GENDER	AGE	YELLOW_FINGERS	ANXIETY	FATIGUE	WHEEZING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CH F
0	M	69	YES	YES	YES	YES	YES	YES	YES	
1	M	74	NO	NO	YES	NO	NO	YES	YES	
2	F	59	NO	NO	YES	YES	YES	YES	NO	
3	M	63	YES	YES	NO	NO	NO	NO	YES	
4	F	63	YES	NO	NO	YES	YES	YES	NO	
...	...	...	...	...	...	...	...	...	...	
304	F	56	NO	NO	YES	NO	YES	YES	YES	

	GENDER	AGE	YELLOW_FINGERS	ANXIETY	FATIGUE	WHEEZING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CH F
305	M	70	NO	NO	YES	YES	YES	YES	NO	
306	M	58	NO	NO	NO	YES	YES	NO	NO	
307	M	67	NO	YES	YES	NO	YES	YES	NO	
308	M	62	NO	NO	YES	YES	NO	NO	YES	

309 rows × 11 columns

As our main goal of this project is to analyze the symptoms of lung cancer, it becomes necessary to drop columns like smoking, peer pressure, alcohol consumption and chronic disease because they are causes of lung cancer, not symptoms. Moreover, we also need to change some of the column names for suitability.

```
lung_df.rename(columns = {'SHORTNESS OF BREATH':'SHORTNESS_OF_BREATH', 'SWALLOWING DIFF
lung_df
```

	GENDER	AGE	YELLOW_FINGERS	ANXIETY	FATIGUE	WHEEZING	COUGHING	SHORTNESS_OF_BREATH	SWALL
0	M	69	YES	YES	YES	YES	YES	YES	
1	M	74	NO	NO	YES	NO	NO	YES	
2	F	59	NO	NO	YES	YES	YES	YES	
3	M	63	YES	YES	NO	NO	NO	NO	
4	F	63	YES	NO	NO	YES	YES	YES	
...	...	...	...	...	...	...	...	...	
304	F	56	NO	NO	YES	NO	YES	YES	
305	M	70	NO	NO	YES	YES	YES	YES	
306	M	58	NO	NO	NO	YES	YES	NO	
307	M	67	NO	YES	YES	NO	YES	YES	
308	M	62	NO	NO	YES	YES	NO	NO	

309 rows × 11 columns

Now the dataset looks more comfortable to work with. First, let's check for any NaN values.

```
lung_df.isnull().any()
```

```
GENDER                 False
AGE                    False
YELLOW_FINGERS         False
ANXIETY                False
FATIGUE                False
WHEEZING               False
COUGHING               False
SHORTNESS_OF_BREATH    False
SWALLOWING_DIFFICULTY  False
CHEST_PAIN             False
LUNG_CANCER            False
dtype: bool
```

As you can see, there are no NaN values in the dataset to get rid of.

```
lung_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 309 entries, 0 to 308
Data columns (total 11 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   GENDER                 309 non-null    object
 1   AGE                    309 non-null    int64
 2   YELLOW_FINGERS         309 non-null    object
 3   ANXIETY                309 non-null    object
 4   FATIGUE                309 non-null    object
 5   WHEEZING               309 non-null    object
 6   COUGHING               309 non-null    object
 7   SHORTNESS_OF_BREATH    309 non-null    object
 8   SWALLOWING_DIFFICULTY  309 non-null    object
 9   CHEST_PAIN             309 non-null    object
 10  LUNG_CANCER            309 non-null    object
dtypes: int64(1), object(10)
memory usage: 26.7+ KB
```

All the values in the dataset are of string type except the age which is an int type

```
lung_df.describe()
```

	AGE
count	309.000000
mean	62.673139
std	8.210301
min	21.000000
25%	57.000000
50%	62.000000
75%	69.000000
max	87.000000

The average of person in this dataset is around 63 where the min and max are 21 and 87 respectively. Now, let's calculate the number of people who faced common symptoms of cancer.

Number of male and female patients

```
lung_df.GENDER.value_counts()
```

```
M    162
```

```
F    147
Name: GENDER, dtype: int64
```

Number Of Patients Who Smoke

```
lung_df.ANXIETY.value_counts()
```

```
NO     155
YES    154
Name: ANXIETY, dtype: int64
```

Number Of Patients Facing Shortness Of Breath

```
lung_df.SHORTNESS_OF_BREATH.value_counts()
```

```
YES    198
NO     111
Name: SHORTNESS_OF_BREATH, dtype: int64
```

Number Of Patients Facing Wheezing Problems

```
lung_df.LUNG_CANCER.value_counts()
```

```
YES    270
NO      39
Name: LUNG_CANCER, dtype: int64
```

```
lung_df.WHEEZING.value_counts()
```

```
YES    172
NO     137
Name: WHEEZING, dtype: int64
```

By looking at the results, most of the patients face shortness of breath when they have lung cancer and there are around 39 people who have disagreed to have lung cancer, however our project does not focus on them as there is a possibility of them suffering from an other disease. We need to get a graphical representation of the situation to get a better insight of the problem. As the dataset is clear and ready for us to work with, we don't need to do any more cleaning or parsing.

```
import jovian
```

```
jovian.commit()
```

```
[jovian] Updating notebook "sainihaldiddi2002/lung-cancer" on https://jovian.ai
[jovian] Committed successfully! https://jovian.ai/sainihaldiddi2002/lung-cancer
```

```
'https://jovian.ai/sainihaldiddi2002/lung-cancer'
```

Exploratory Analysis and Visualization

TODO - write some explanation here.

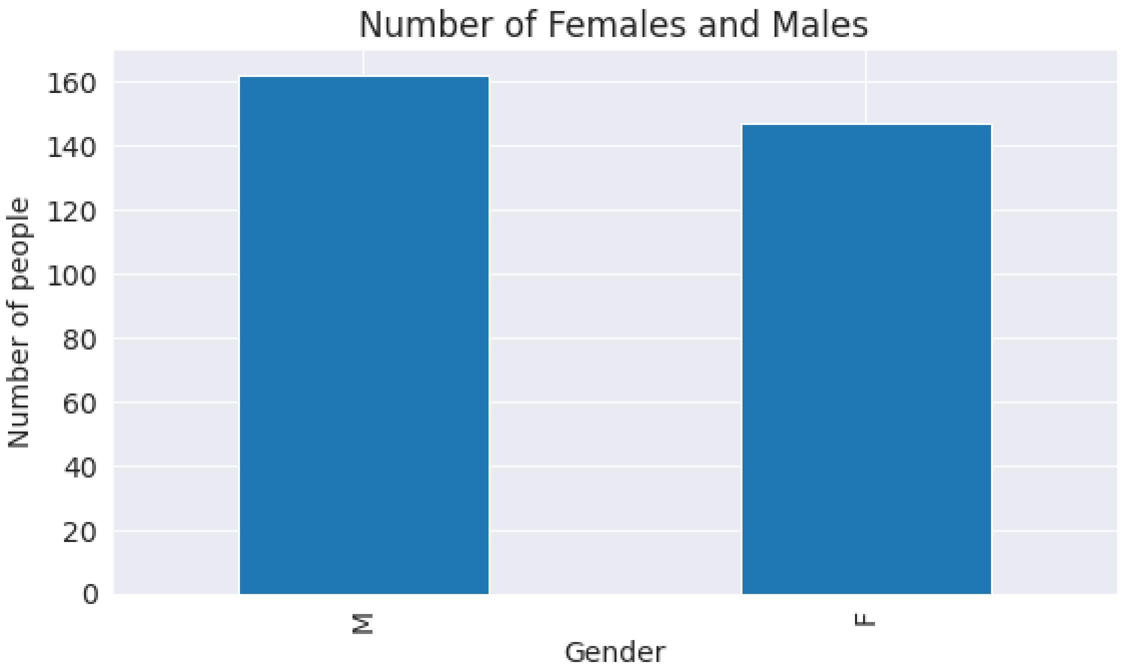
> Instructions (delete this cell)
>
> - Compute the mean, sum, range and other interesting statistics for numeric columns
> - Explore distributions of numeric columns using histograms etc.
> - Explore relationship between columns using scatter plots, bar charts etc.
> - Make a note of interesting insights from the exploratory analysis

Let's begin by importing `matplotlib.pyplot` and `seaborn` .

```
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline

sns.set_style('darkgrid')
matplotlib.rcParams['font.size'] = 14
matplotlib.rcParams['figure.figsize'] = (9, 5)
matplotlib.rcParams['figure.facecolor'] = '#00000000'
```

TODO - Explore one or more columns by plotting a graph below, and add some explanation about it

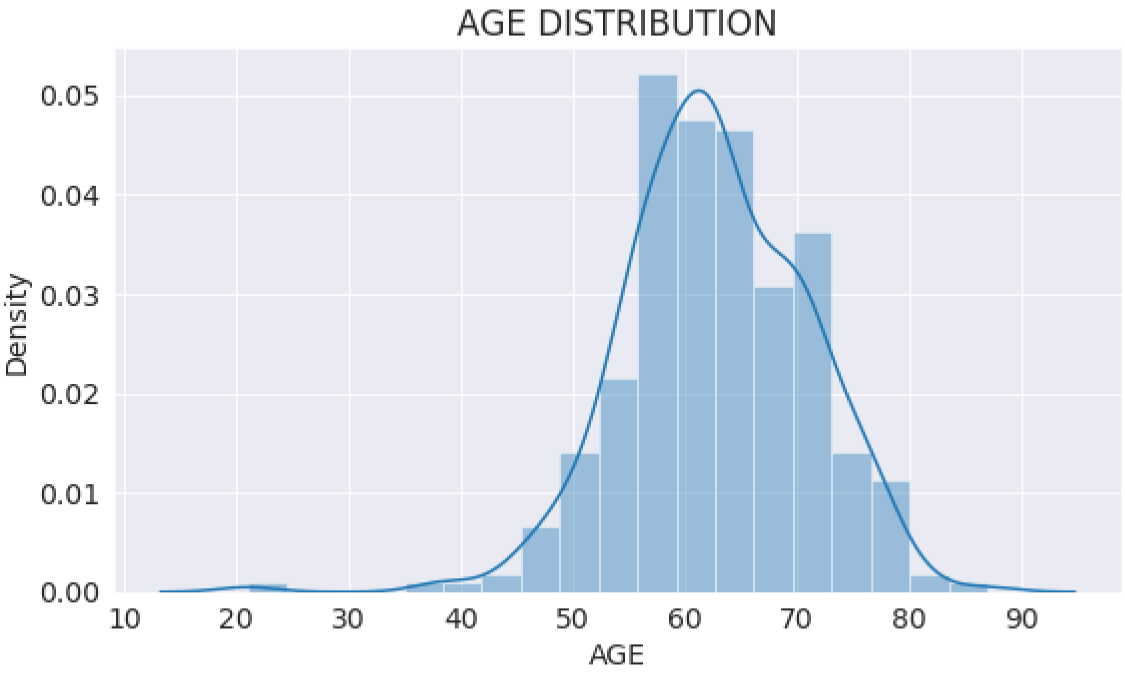
```
graph = lung_df['GENDER'].value_counts()
graph.plot(kind='bar');
plt.xlabel("Gender");
plt.ylabel("Number of people");
plt.title("Number of Females and Males");
```



As you can see, more number of males are there in this dataset than females. It shows around 163 males and 143 females took part in this survey dataset

```
sns.distplot(lung_df["AGE"], kde=True);
plt.title("AGE DISTRIBUTION");
```

```
/opt/conda/lib/python3.9/site-packages/seaborn/distributions.py:2619: FutureWarning:
`distplot` is a deprecated function and will be removed in a future version. Please
adapt your code to use either `displot` (a figure-level function with similar
flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

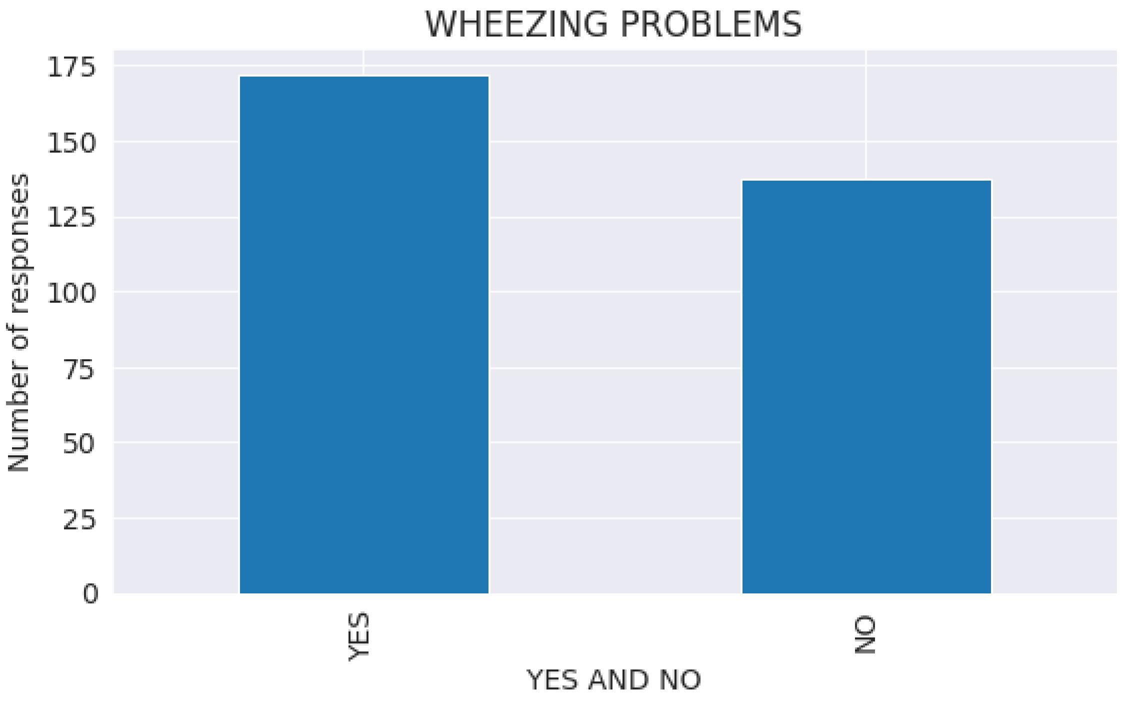


Most of the people's age is concentrated in the interval 40-85 years of age.

TODO - Explore one or more columns by plotting a graph below, and add some explanation about it

```
df = lung_df["WHEEZING"].value_counts()
df.plot(kind='bar');
plt.title("WHEEZING PROBLEMS")
plt.xlabel("YES AND NO")
plt.ylabel("Number of responses")
```

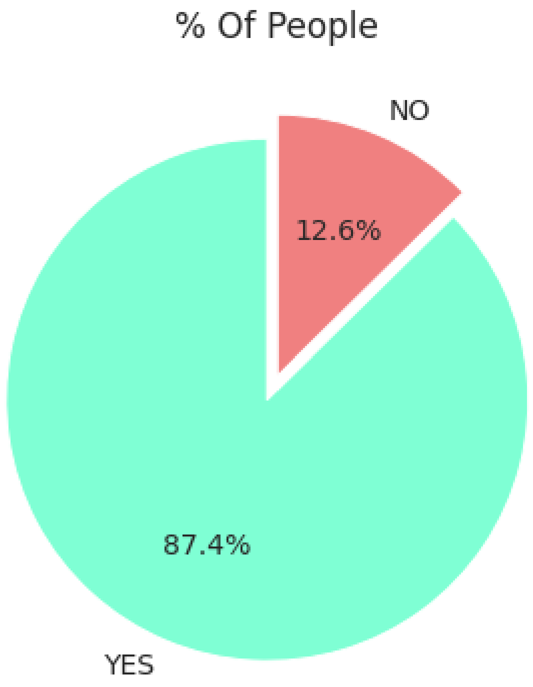
```
Text(0, 0.5, 'Number of responses')
```



According to the data, most of the people have faced wheezing problems when they had lung cancer. Generally, this problem occurs due to the blockages caused by large tumors from lung cancer.

TODO - Explore one or more columns by plotting a graph below, and add some explanation about it

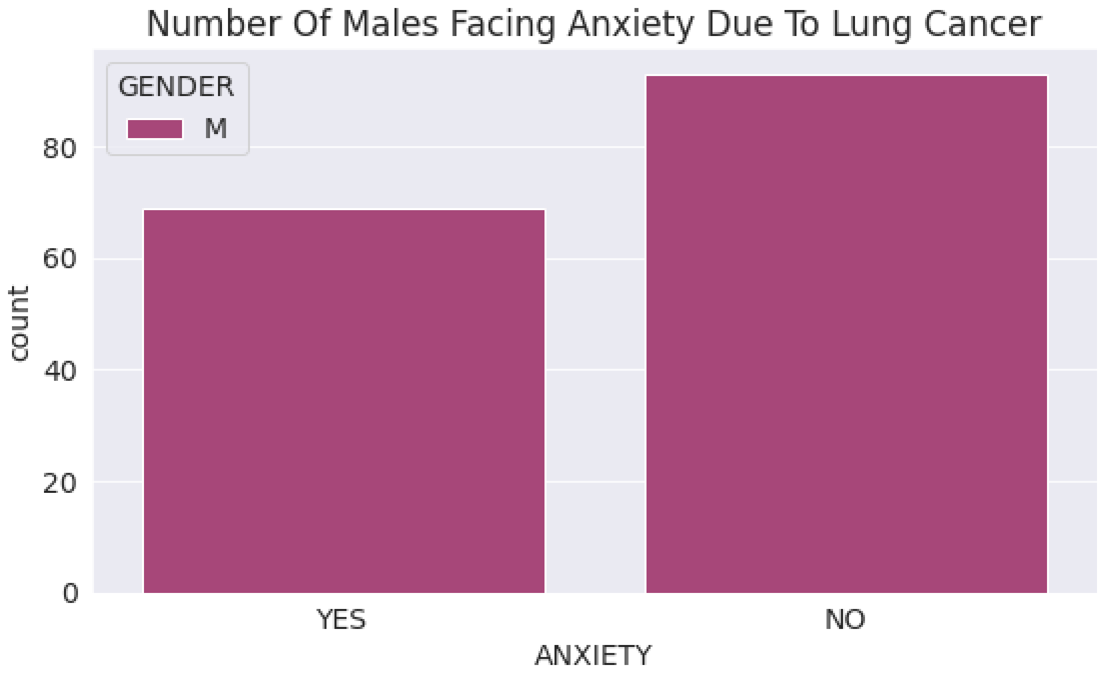
```
lung = lung_df.LUNG_CANCER.value_counts()
plt.figure(figsize=(12, 6))
plt.pie(lung, labels = ['YES', 'NO'] , autopct ='%.1f%%', startangle = 90, explode=[0.1
plt.title("% Of People");
```



As per the pie-chart, many people have responded that they have lung cancer. The main problem is smoking. Smoking has been everyone's enemy for many decades and people have become very addicted towards it.

TODO - Explore one or more columns by plotting a graph below, and add some explanation about it

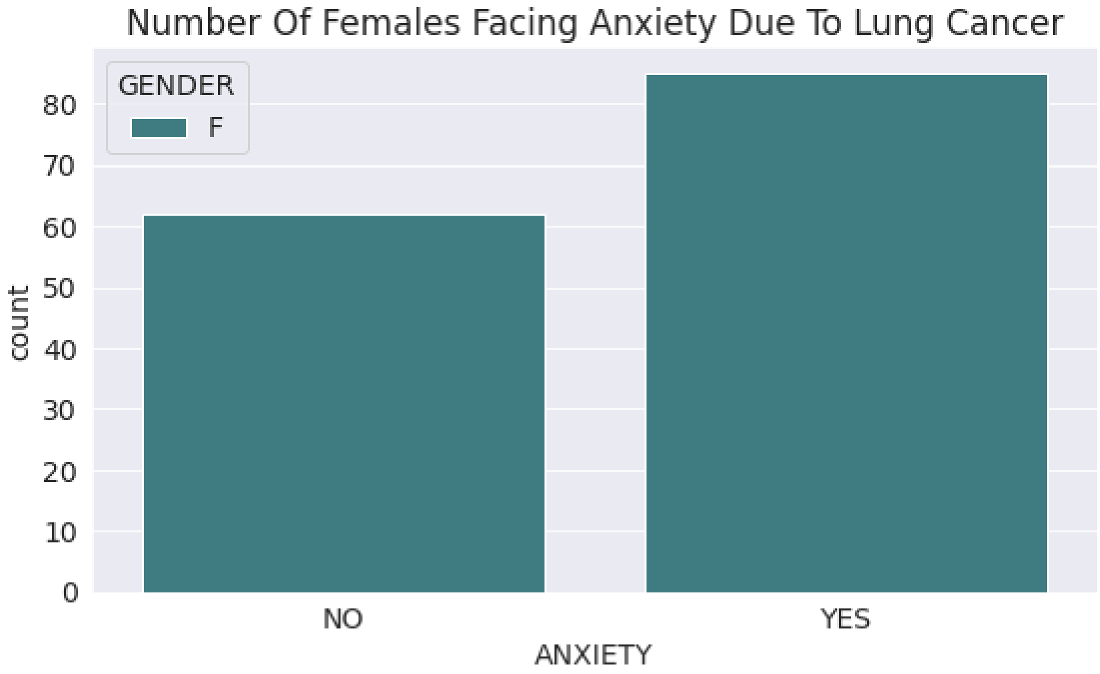
```
df2_males = lung_df[lung_df.GENDER=="M"]
sns.countplot(x = 'ANXIETY', hue = 'GENDER', data =df2_males, palette = 'magma');
plt.title("Number Of Males Facing Anxiety Due To Lung Cancer");
```



Around 69 adult males feel anxious about their disease(lung_cancer) and 93 of them do not face anxiety. A lung cancer diagnosis is a life-changing event. There is a possibility that you will feel overwhelmed, scared, stressed, or anxious as a result. Anxiety and stress are common during the journey.

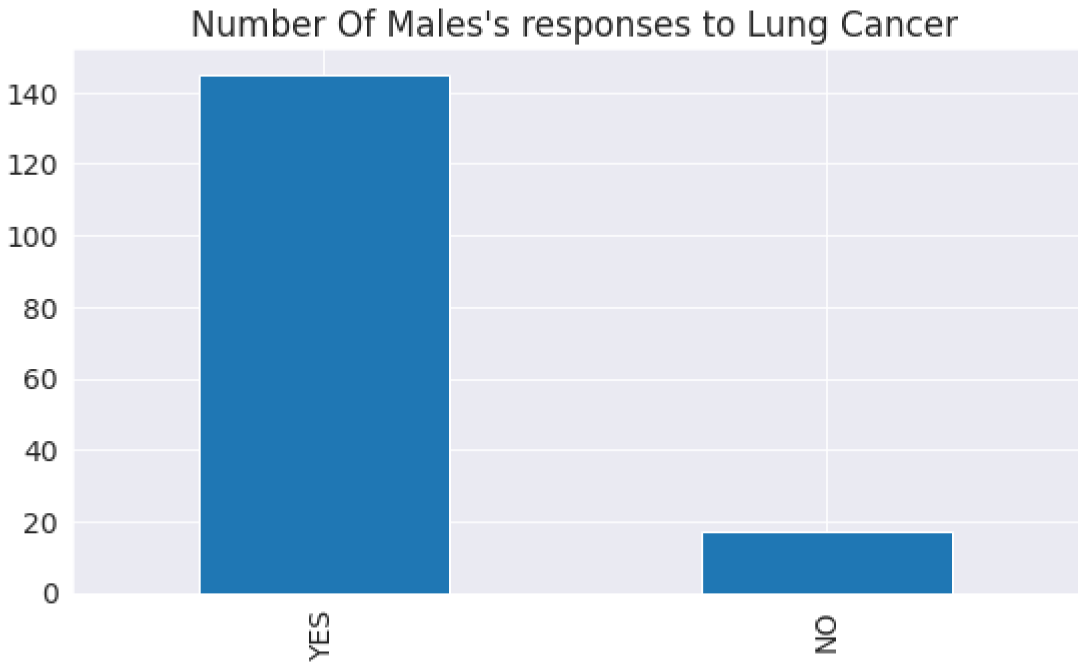
TODO - Explore one or more columns by plotting a graph below, and add some explanation about it

```
df2 = lung_df[lung_df.GENDER=="F"]
sns.countplot(x = 'ANXIETY', hue = 'GENDER', data =df2, palette = 'crest');
plt.title("Number Of Females Facing Anxiety Due To Lung Cancer");
```

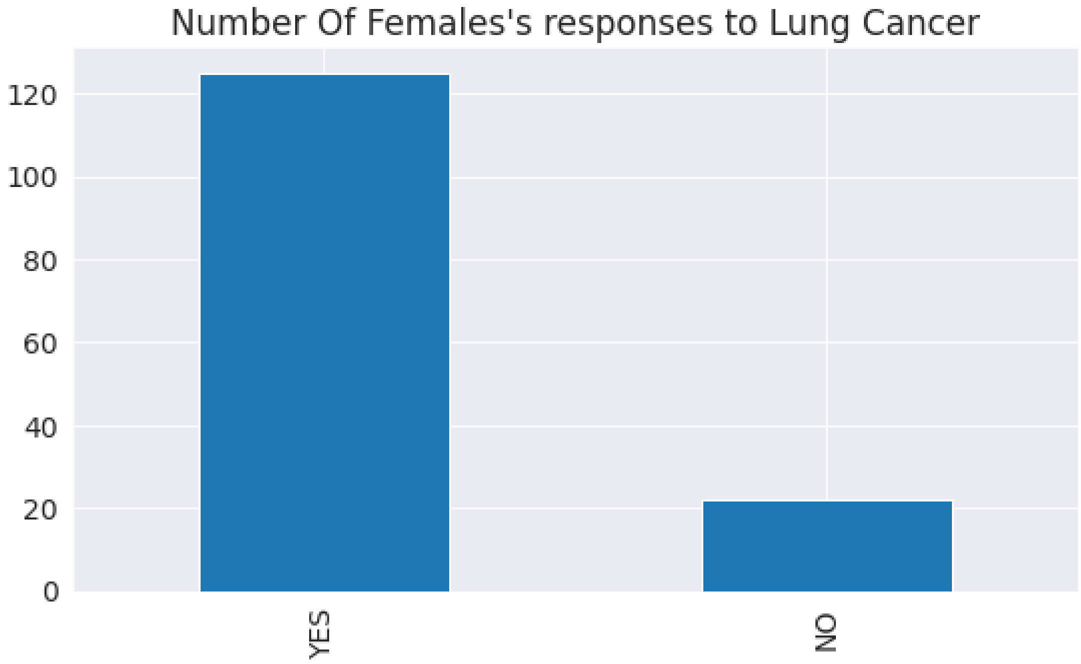


From the above 2 charts, it is clear that women feel more anxious than men. The reason why this happens is still inconclusive, but according to the survey dataset, the main problems are emotional distress and the fear that the disease may return.

```
df2_males = lung_df[lung_df.GENDER=="M"].LUNG_CANCER.value_counts()
df2_females = lung_df[lung_df.GENDER=="F"].LUNG_CANCER.value_counts()
df2_males.plot(kind='bar')
plt.title("Number Of Males's responses to Lung Cancer");
```



```
df2_females.plot(kind='bar');
plt.title("Number Of Females's responses to Lung Cancer");
```



the number of males having lung cancer is more than that of females as the smoking patterns in males is higher than that of females according to a study. Males have known to consume tobacco more than females.

Let us save and upload our work to Jovian before continuing

```
import jovian
```

```
jovian.commit()
```

```
[jovian] Updating notebook "sainihaldiddi2002/lung-cancer" on https://jovian.ai
[jovian] Committed successfully! https://jovian.ai/sainihaldiddi2002/lung-cancer
```

```
'https://jovian.ai/sainihaldiddi2002/lung-cancer'
```

Asking and Answering Questions

So far, we have made a good perception about the dataset by using the python libraries, but there are some interesting questions to be asked about the data. Here are some 5 questions which I feel is interesting.

> Instructions (delete this cell)
>
> - Ask at least 5 interesting questions about your dataset
> - Answer the questions either by computing the results using Numpy/Pandas or by plotting graphs using Matplotlib/Seaborn
> - Create new columns, merge multiple dataset and perform grouping/aggregation wherever necessary
> - Wherever you're using a library function from Pandas/Numpy/Matplotlib etc. explain briefly what it does

Q1: TODO - How many people in the age group 65-70 have wheezing problems?

```
def series(number):
    condition = [
    (number >= 0) & (number < 5),
    (number>= 5) & (number < 10),
    (number>= 10) & (number< 15),
    (number>= 15) & (number< 20),
    (number>= 20) & (number< 25),
    (number>= 25) & (number< 30),
    (number>= 30) & (number< 35),
    (number>= 35) & (number< 40),
    (number>= 40) & (number< 45),
    (number>= 45) & (number< 50),
    (number>= 50) & (number< 55),
    (number>= 55) & (number< 60),
    (number>= 60) & (number< 65),
    (number>= 65) & (number< 70),
    (number>= 70) & (number< 75),
    (number>= 75) & (number< 80),
    (number>= 80) & (number< 85),
    (number>= 85) & (number< 90),
    ]

    output = ['0-5', '5-10', '10-15', '15-20', '20-25', '25-30', '30-35', '35-40', '40-
```

```
              '60-65', '65-70', '70-75', '75-80', '80-85', '85-90']
    result = np.select(condition, output, '>90')
    return pd.Series(result)

lung_df['age_group'] = series(lung_df.AGE)
lung_df
```

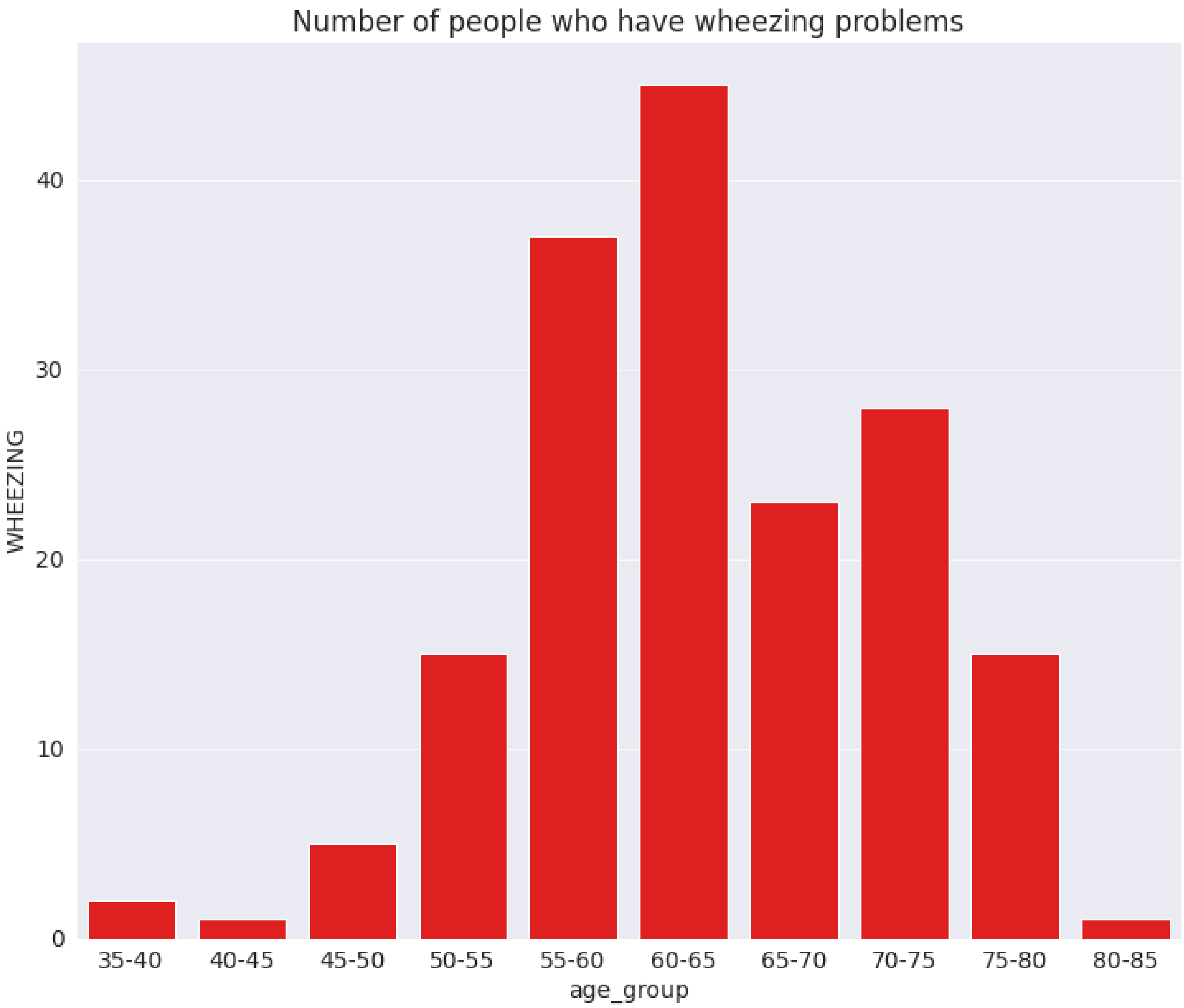
	GENDER	AGE	YELLOW_FINGERS	ANXIETY	FATIGUE	WHEEZING	COUGHING	SHORTNESS_OF_BREATH	SWALL
0	M	69	YES	YES	YES	YES	YES	YES	
1	M	74	NO	NO	YES	NO	NO	YES	
2	F	59	NO	NO	YES	YES	YES	YES	
3	M	63	YES	YES	NO	NO	NO	NO	
4	F	63	YES	NO	NO	YES	YES	YES	
...	...	...	...	...	...	...	...	...	
304	F	56	NO	NO	YES	NO	YES	YES	
305	M	70	NO	NO	YES	YES	YES	YES	
306	M	58	NO	NO	NO	YES	YES	NO	
307	M	67	NO	YES	YES	NO	YES	YES	
308	M	62	NO	NO	YES	YES	NO	NO	

309 rows × 12 columns

```
lung = lung_df[lung_df.WHEEZING=='YES'].groupby('age_group')[["WHEEZING"]].count()
lung
```

	WHEEZING
age_group	
35-40	2
40-45	1
45-50	5
50-55	15
55-60	37
60-65	45
65-70	23
70-75	28
75-80	15
80-85	1

```
plt.figure(figsize=(12,10));
sns.barplot(x = lung.index , y = lung.WHEEZING, alpha=1, palette =['red']);
plt.title('Number of people who have wheezing problems');
```

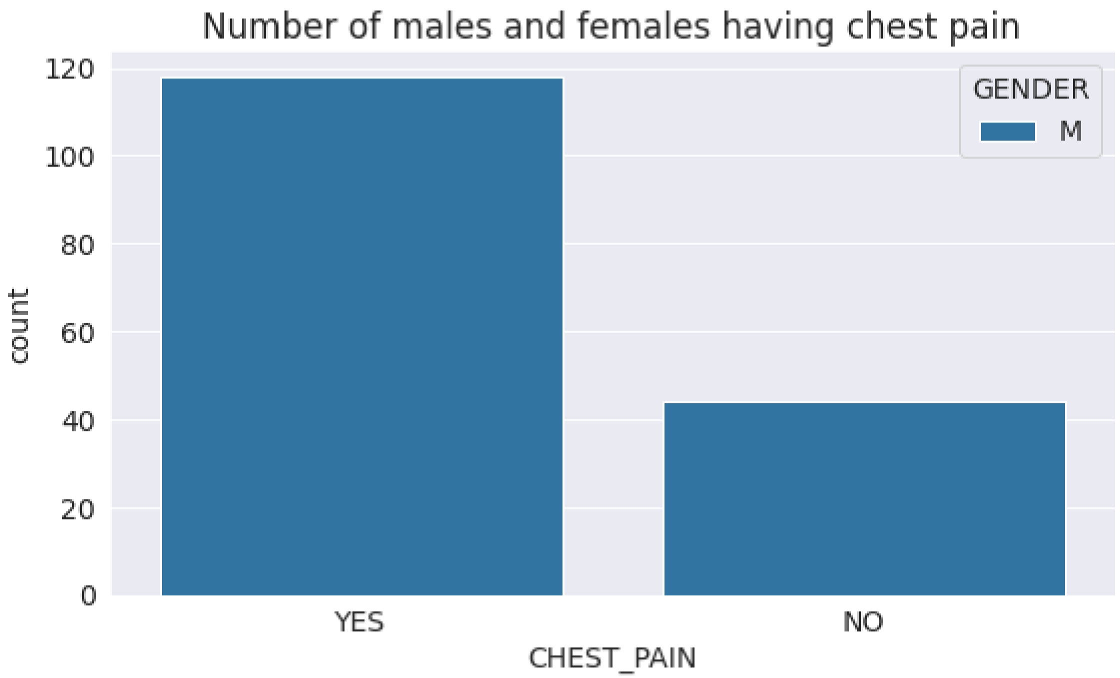


It seems people in the age group of 60-65 face wheezing problems than others. Generally, people above 60 are considered old and facing lung cancer at this age can be detrimental to health. Moreover, other problems like allergies, exercise and air pollution can increase the effect of this problem.

Q2: TODO - Do males have chest pain more than females?if yes, why?

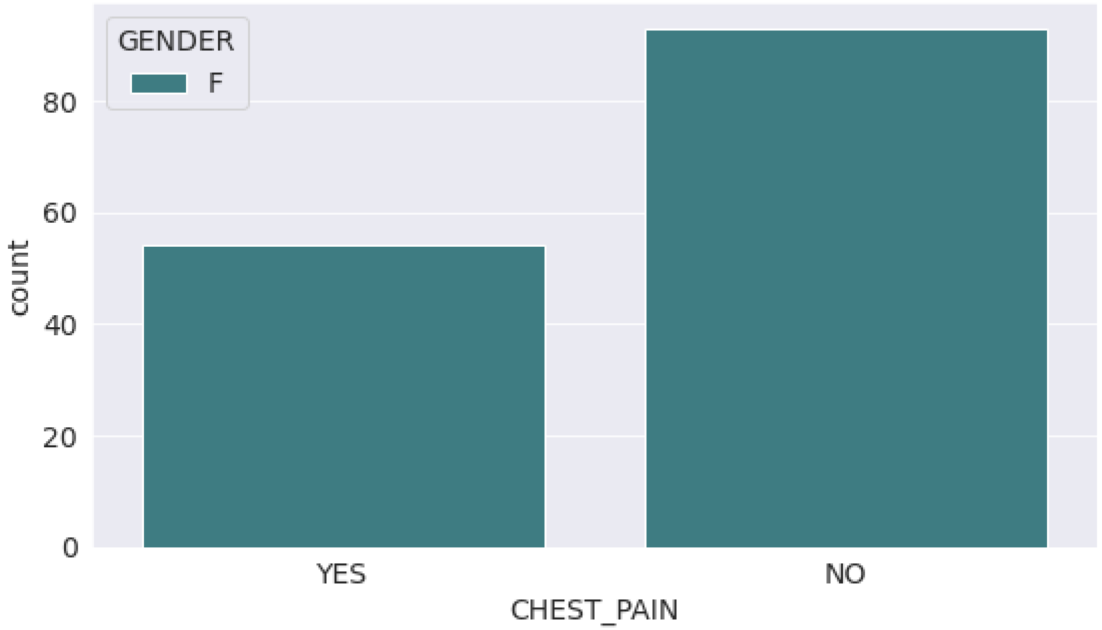
```
df_males = lung_df[lung_df.GENDER=="M"]
df_females = lung_df[lung_df.GENDER=="F"]
sns.countplot(x="CHEST_PAIN",hue='GENDER',data=df_males)
plt.title("Number of males and females having chest pain")
```

```
Text(0.5, 1.0, 'Number of males and females having chest pain')
```



```
sns.countplot(x="CHEST_PAIN",hue='GENDER',data=df_females, palette='crest')
```

```
<AxesSubplot:xlabel='CHEST_PAIN', ylabel='count'>
```



Yes, more than females, males have responded yes to having chest pain. This is because the lung cancer in women is different from the one in men. Moreover, a recent study showed that men smoke at a higher rate than women in most of the countries and smoking is the main cause of lung cancer. Since, men smoke more than women, they face chest pains more then women

Q3: TODO - How many females took part in this survey ? How many of them faced shortness of breath ?

```
lung_df.GENDER.value_counts()
```

```
M    162
F    147
Name: GENDER, dtype: int64
```

147 females took part in this survey

```
df2 = lung_df[lung_df.GENDER=="F"]
df2.SHORTNESS_OF_BREATH.value_counts()
```

```
YES    99
NO     48
Name: SHORTNESS_OF_BREATH, dtype: int64
```

99 females responded "YES" to shortness of breath

Q4: TODO - What fraction of women face coughing during lung cancer?

```
number_of_women = lung_df.GENDER.value_counts()["F"]
women = lung_df[lung_df.GENDER=="F"]
frac = women.COUGHING.value_counts()["YES"]
result = frac/number_of_women
```

The fraction of women who face coughing is around 0.5

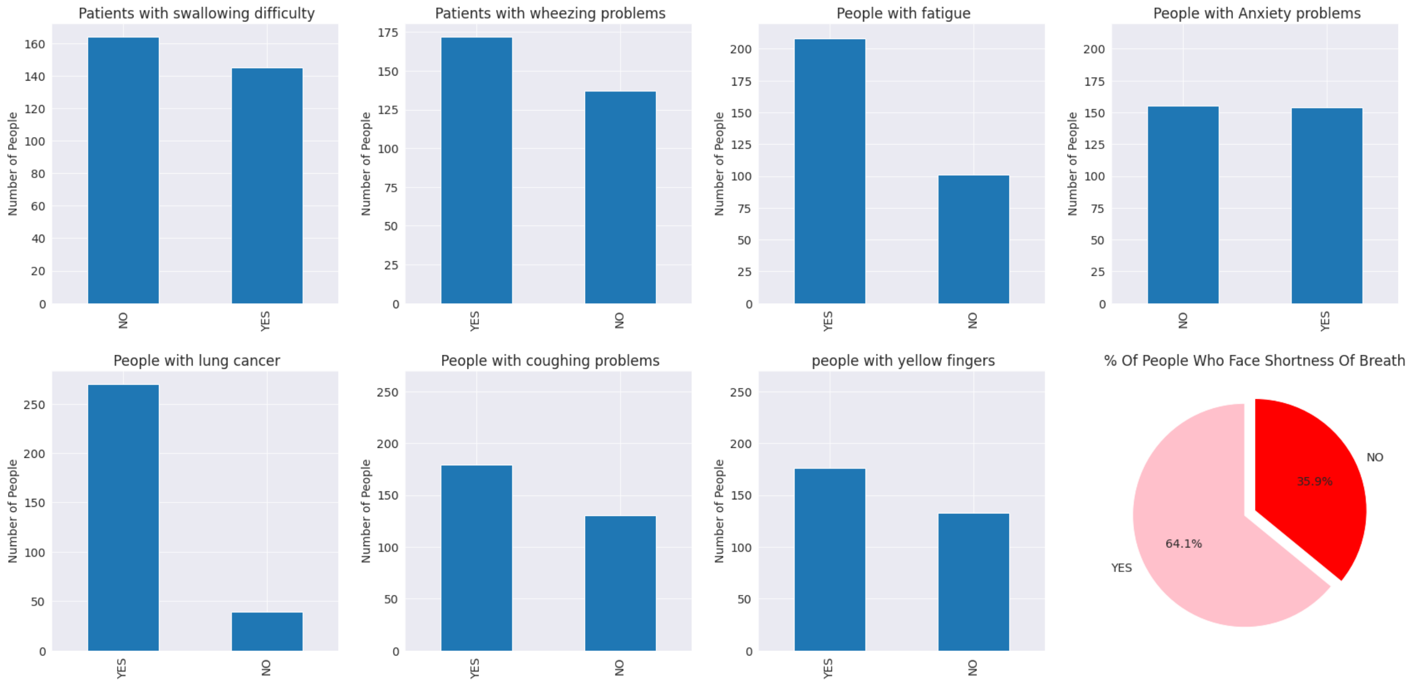
Q5: TODO - Print subplots showing important information about people's responses to the symptoms

```
swal_pat = lung_df.SWALLOWING_DIFFICULTY.value_counts()
wheez_pat = lung_df.WHEEZING.value_counts()
short_br = lung_df.SHORTNESS_OF_BREATH.value_counts()
Canc_data = lung_df.LUNG_CANCER.value_counts()
anx_pat = lung_df.ANXIETY.value_counts()
fat_br = lung_df.FATIGUE.value_counts()
cou_data = lung_df.COUGHING.value_counts()
yell_data = lung_df.YELLOW_FINGERS.value_counts()

swal_pat.index
```

```
Index(['NO', 'YES'], dtype='object')
```

```
fig, axes = plt.subplots(2,4, figsize =(24, 12))
#axis(0,0) use this axis for Anaemia
axes[0,0].set_title(" Patients with swallowing difficulty")
swal_pat.plot(kind='bar', ax=axes[0,0])
axes[0,0].set_ylabel("Number of People")

# axis(0,1) use this axis for Diabetes
axes[0,1].set_title(" Patients with wheezing problems")
wheez_pat.plot(kind='bar',ax=axes[0,1])
axes[0,1].set_xlabel("")
axes[0,1].set_ylabel("Number of People")
# axes[0,1].set_ylim([0, 220])
```

```
#axis(0,2) use this axis for High blood pressure
plt.pie(short_br, labels = ['YES', 'NO'] , autopct ='%.1f%%', startangle = 90, explode=
plt.title("% Of People Who Face Shortness Of Breath");
axes[1,1].set_xlabel("")

#axis(0,3) use this axis for smoking habit
axes[0,3].set_title(" People with Anxiety problems")
anx_pat.plot(kind='bar',ax=axes[0,3])
axes[0,3].set_xlabel("")
axes[0,3].set_ylabel("Number of People")
axes[0,3].set_ylim([0, 220])

axes[0,2].set_title(" People with fatigue")
fat_br.plot(kind='bar',ax=axes[0,2])
axes[0,2].set_xlabel("")
axes[0,2].set_ylabel("Number of People")
axes[0,2].set_ylim([0, 220])


#axis(1,0) use this axis for Ejection Fraction
axes[1,0].set_title(" People with lung cancer")
Canc_data.plot(kind='bar',ax=axes[1,0])
axes[1,0].set_xlabel("")
axes[1,0].set_ylabel("Number of People")

#axis(1,1) use this axis for Platelets count
axes[1,1].set_title(" People with coughing problems")
cou_data.plot(kind='bar',ax=axes[1,1])
axes[1,1].set_xlabel("")
axes[1,1].set_ylabel("Number of People")
axes[1,1].set_ylim([0, 270])

#axis(1,2) use this axis for Creatinine level
axes[1,2].set_title(" people with yellow fingers")
yell_data.plot(kind='bar',ax=axes[1,2])
axes[1,2].set_xlabel("")
axes[1,2].set_ylabel("Number of People")
axes[1,2].set_ylim([0, 270])


plt.tight_layout(pad=2)
```



Since the main symptoms of lung cancer are swallowing difficulty, wheezing and shortness of breath,we have plotted 3 graphs about them and also the people with lung cancer

number of women in the age group of 60-65 facing anxiety problems is more than the others

Let us save and upload our work to Jovian before continuing.

```
import jovian
```

```
jovian.commit()
```

```
[jovian] Updating notebook "sainihaldiddi2002/lung-cancer" on https://jovian.ai
[jovian] Committed successfully! https://jovian.ai/sainihaldiddi2002/lung-cancer
```

```
'https://jovian.ai/sainihaldiddi2002/lung-cancer'
```

Inferences and Conclusion

1. Around 309 people took part in this survey(Male=162, Females=147).Most of the people in the data are above 40 years of age
2. 270 people have agreed that they have lung cancer whereas 39 of them have disagreed.To be specific, more males have lung cancer than females.
3. According to the dataset, around 172 people have wheezing problems which is one of the most common symptoms of lung cancer. Most of them lie in the age group 60-65.
4. A high proportion of people seem to have agreed to most of the symptoms of lung cancer (Wheezing, yellow fingers and fatigue). There's a possibility that new patients would face the same symptoms.
5. Even though there are few people who have disagreed to have the disease, there's a chance that they maybe affected by any other disease.

Lung cancer has been doctor's worst enemy since generations. Till now, there's not a perfect cure for this disease. Around 140,000 people died due to this disease. According to scientists, the curable rate of lung cancer is 80%-90% in its earliest stage, the rate decreases as the stages increase.Overall, with the help of this data, we can get a better knowledge of people's issues during lung cancer and thereby approach to them with the current treatment.

```
import jovian
```

```
jovian.commit()
```

```
[jovian] Updating notebook "sainihaldiddi2002/lung-cancer" on https://jovian.ai
[jovian] Committed successfully! https://jovian.ai/sainihaldiddi2002/lung-cancer
```

```
'https://jovian.ai/sainihaldiddi2002/lung-cancer'
```

References and Future Work

Working on this project really helped to me to know where I stand and how I need to improve. This is my first time analyzing a real-world dataset and it has been a great experience. in the future, I decide to work for companies analyzing their sales, handle risks and personalize customer relations.

REFERENCES:

1. https://www.cancer.net/blog/2018-06/just-diagnosed-with-lung-cancer-answers-expert#:~:text=As with many other cancers,as 80%25 to 90%25.
2. https://www.mayoclinic.org/diseases-conditions/lung-cancer/symptoms-causes/syc-20374620#:~:text=Lung cancer can cause complications,expand fully when you inhale.
3. https://doi.org/10.1016/S1535-6108(02)00027-2
4. https://medlineplus.gov/lungcancer.html

```
import jovian
```

```
jovian.commit()
```

```
[jovian] Updating notebook "sainihaldiddi2002/lung-cancer" on https://jovian.ai
[jovian] Committed successfully! https://jovian.ai/sainihaldiddi2002/lung-cancer
```

```
'https://jovian.ai/sainihaldiddi2002/lung-cancer'
```