```
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt
```

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

```
df.head(3)
```

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH
0	M	69	1	2	2	1	1	2	1	2	2	2	2
1	M	74	2	1	1	1	2	2	2	1	1	1	2
2	F	59	1	1	1	2	1	2	1	2	1	2	2

Next steps: Generate code with df New interactive sheet

```
df.tail()
```

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNE OF BREA
304	F	56	1	1	1	2	2	2	1	1	2	2	
305	M	70	2	1	1	1	1	2	2	2	2	2	
306	M	58	2	1	1	1	1	1	2	2	2	2	
307	M	67	2	1	2	1	1	2	2	1	2	2	
308	M	62	1	1	1	2	1	2	2	2	2	1	

```
df.sample()
```

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNES OF BREAT
73	F	60	2	1	1	1	2	2	1	1	1	1	

```
df.shape
```

```
(309, 16)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 309 entries, 0 to 308
Data columns (total 16 columns):
 #   Column                 Non-Null Count  Dtype 
---  ------                 --------------  ----- 
 0   GENDER                 309 non-null    object
 1   AGE                    309 non-null    int64 
 2   SMOKING                309 non-null    int64 
 3   YELLOW_FINGERS         309 non-null    int64 
 4   ANXIETY                309 non-null    int64 
 5   PEER_PRESSURE          309 non-null    int64 
 6   CHRONIC DISEASE        309 non-null    int64 
 7   FATIGUE                309 non-null    int64 
 8   ALLERGY                309 non-null    int64 
 9   WHEEZING               309 non-null    int64 
 10  ALCOHOL CONSUMING      309 non-null    int64 
 11  COUGHING               309 non-null    int64 
 12  SHORTNESS OF BREATH    309 non-null    int64 
 13  SWALLOWING DIFFICULTY  309 non-null    int64 
 14  CHEST PAIN             309 non-null    int64 
 15  LUNG_CANCER            309 non-null    object
dtypes: int64(14), object(2)
memory usage: 38.8+ KB
```

```
df.describe()
```

	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	
count	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000	3
mean	62.673139	1.563107	1.569579	1.498382	1.501618	1.504854	1.673139	1.556634	1.556634	1.556634	
std	8.210301	0.496806	0.495938	0.500808	0.500808	0.500787	0.469827	0.497588	0.497588	0.497588	
min	21.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	
25%	57.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	
50%	62.000000	2.000000	2.000000	1.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	
75%	69.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	
max	87.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	

```
df.dtypes
```

	0
GENDER	object
AGE	int64
SMOKING	int64
YELLOW_FINGERS	int64
ANXIETY	int64
PEER_PRESSURE	int64
CHRONIC DISEASE	int64
FATIGUE	int64
ALLERGY	int64
WHEEZING	int64
ALCOHOL CONSUMING	int64
COUGHING	int64
SHORTNESS OF BREATH	int64
SWALLOWING DIFFICULTY	int64
CHEST PAIN	int64
LUNG_CANCER	object

dtype: object

```
df.isnull().sum()
```

```
                        0
GENDER                  0
AGE                     0
SMOKING                 0
YELLOW_FINGERS          0
ANXIETY                 0
PEER_PRESSURE           0
CHRONIC DISEASE         0
FATIGUE                 0
ALLERGY                 0
WHEEZING                0
ALCOHOL CONSUMING       0
COUGHING                0
SHORTNESS OF BREATH     0
SWALLOWING DIFFICULTY   0
CHEST PAIN              0
LUNG_CANCER             0

dtype: int64
```

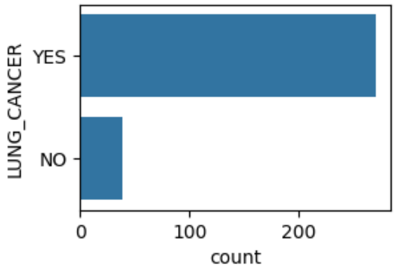
```
df['LUNG_CANCER'].value_counts()
```

```
              count
LUNG_CANCER
YES             270
NO               39

dtype: int64
```

```
plt.figure(figsize=(3,2))
sns.countplot(df['LUNG_CANCER'])
```

```
<Axes: xlabel='count', ylabel='LUNG_CANCER'>
```



encoding

x and y me break

train test split

standard scaler

mode train

encoding

Double-click (or enter) to edit

```
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
df['GENDER']=le.fit_transform(df['GENDER'])
df['LUNG_CANCER']=le.fit_transform(df['LUNG_CANCER'])
```

```
df
```

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNE OF BREA
0	1	69	1	2	2	1	1	2	1	2	2	2	
1	1	74	2	1	1	1	2	2	2	1	1	1	
2	0	59	1	1	1	2	1	2	1	2	1	2	
3	1	63	2	2	2	1	1	1	1	1	2	1	
4	0	63	1	2	1	1	1	1	1	2	1	2	
...	...	...	...	...	...	...	...	...	...	...	...	...	
304	0	56	1	1	1	2	2	2	1	1	2	2	
305	1	70	2	1	1	1	1	2	2	2	2	2	
306	1	58	2	1	1	1	1	1	2	2	2	2	
307	1	67	2	1	2	1	1	2	2	1	2	2	
308	1	62	1	1	1	2	1	2	2	2	2	1	

309 rows × 16 columns

Next steps: Generate code with df New interactive sheet

```
x = df.drop('LUNG_CANCER', axis=1)
y = df['LUNG_CANCER']
```

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=42)
```

```
from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
x_train=sc.fit_transform(x_train)
x_test=sc.transform(x_test)
```

```
from sklearn.svm import SVC
model=SVC()
model.fit(x_train,y_train)
```

```
▾ SVC ⓘ ?
SVC()
```

```
model.score(x_train,y_train)*100,model.score(x_test,y_test)*100
```

```
(94.73684210526315, 96.7741935483871)
```

```
y_pred=model.predict(x_test)
```

```
from sklearn.metrics import confusion_matrix
cm=confusion_matrix(y_test,y_pred)
```

```
cm
```

```
array([[ 1,  1],
       [ 1, 59]])
```

```
from sklearn.metrics import accuracy_score, classification_report
print("Accuracy:",accuracy_score(y_test,y_pred)*100)
print(classification_report(y_test,y_pred))
```

```
Accuracy: 96.7741935483871
              precision    recall  f1-score   support

           0       0.50      0.50      0.50         2
           1       0.98      0.98      0.98        60

    accuracy                           0.97        62
   macro avg       0.74      0.74      0.74        62
weighted avg       0.97      0.97      0.97        62
```

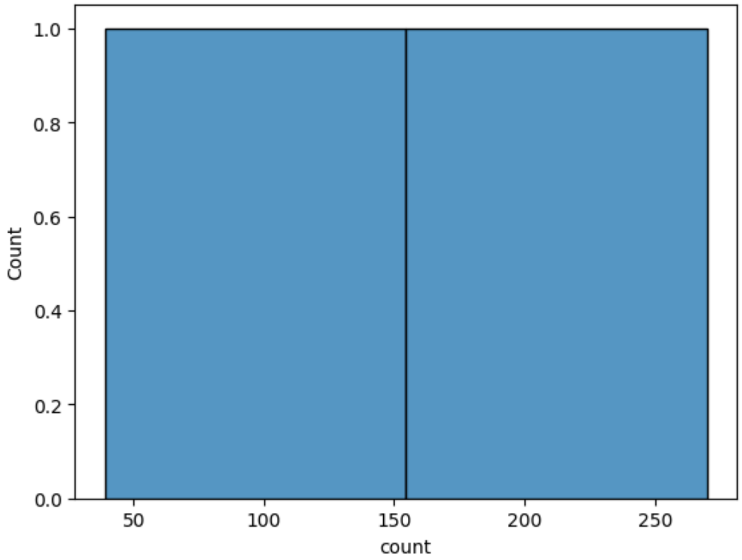
```
a=df['LUNG_CANCER'].value_counts()
a
```

	count
LUNG_CANCER	
1	270
0	39

dtype: int64

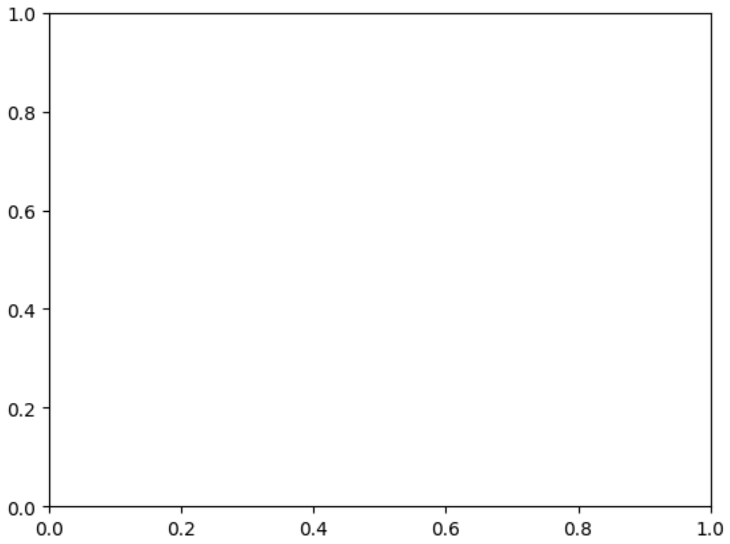
```
sns.histplot(a)
```

```
<Axes: xlabel='count', ylabel='Count'>
```



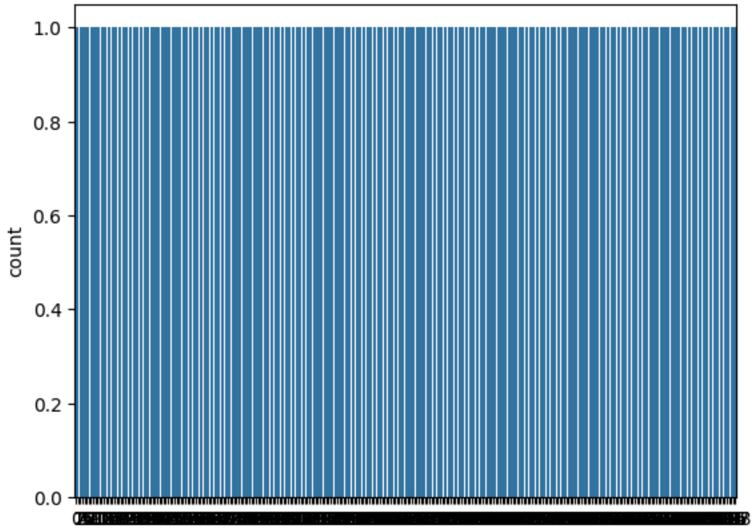
```
sns.boxplot(a=[10,20,30,200])
```

```
<Axes: >
```



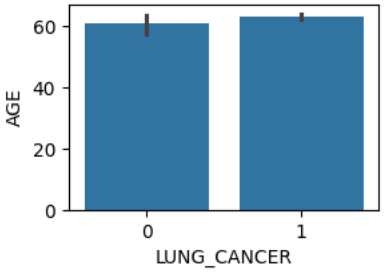
```
sns.countplot(df['LUNG_CANCER'])
```

```
<Axes: ylabel='count'>
```



```
plt.figure(figsize=(3, 2))
sns.barplot(x='LUNG_CANCER',y='AGE',data=df)
```

```
<Axes: xlabel='LUNG_CANCER', ylabel='AGE'>
```

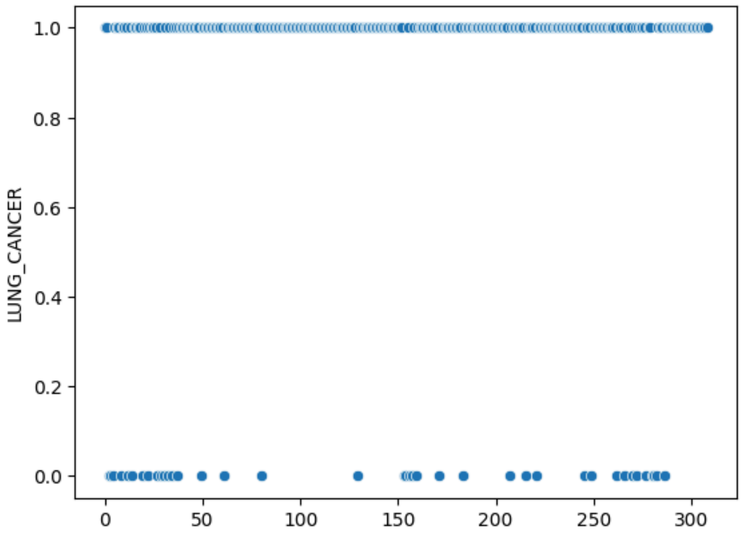


```
df.corr()
```

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING
GENDER	1.000000	0.021306	0.036277	-0.212959	-0.152127	-0.275564	-0.204606	-0.083560	0.154251	0.141207
AGE	0.021306	1.000000	-0.084475	0.005205	0.053170	0.018685	-0.012642	0.012614	0.027990	0.055011
SMOKING	0.036277	-0.084475	1.000000	-0.014585	0.160267	-0.042822	-0.141522	-0.029575	0.001913	-0.129426
YELLOW_FINGERS	-0.212959	0.005205	-0.014585	1.000000	0.565829	0.323083	0.041122	-0.118058	-0.144300	-0.078515
ANXIETY	-0.152127	0.053170	0.160267	0.565829	1.000000	0.216841	-0.009678	-0.188538	-0.165750	-0.191807
PEER_PRESSURE	-0.275564	0.018685	-0.042822	0.323083	0.216841	1.000000	0.048515	0.078148	-0.081800	-0.068771
CHRONIC DISEASE	-0.204606	-0.012642	-0.141522	0.041122	-0.009678	0.048515	1.000000	-0.110529	0.106386	-0.049967
FATIGUE	-0.083560	0.012614	-0.029575	-0.118058	-0.188538	0.078148	-0.110529	1.000000	0.003056	0.141937
ALLERGY	0.154251	0.027990	0.001913	-0.144300	-0.165750	-0.081800	0.106386	0.003056	1.000000	0.173867
WHEEZING	0.141207	0.055011	-0.129426	-0.078515	-0.191807	-0.068771	-0.049967	0.141937	0.173867	1.000000
ALCOHOL CONSUMING	0.454268	0.058985	-0.050623	-0.289025	-0.165750	-0.159973	0.002150	-0.191377	0.344339	0.265659
COUGHING	0.133303	0.169950	-0.129471	-0.012640	-0.225644	-0.089019	-0.175287	0.146856	0.189524	0.374265
SHORTNESS OF BREATH	-0.064911	-0.017513	0.061264	-0.105944	-0.144077	-0.220175	-0.026459	0.441745	-0.030056	0.037834
SWALLOWING DIFFICULTY	-0.078161	-0.001270	0.030718	0.345904	0.489403	0.366590	0.075176	-0.132790	-0.061508	0.069027
CHEST PAIN	0.362958	-0.018104	0.120117	-0.104829	-0.113634	-0.094828	-0.036938	-0.010832	0.239433	0.147640
LUNG_CANCER	0.067254	0.089465	0.058179	0.181339	0.144947	0.186388	0.110891	0.150673	0.327766	0.249300

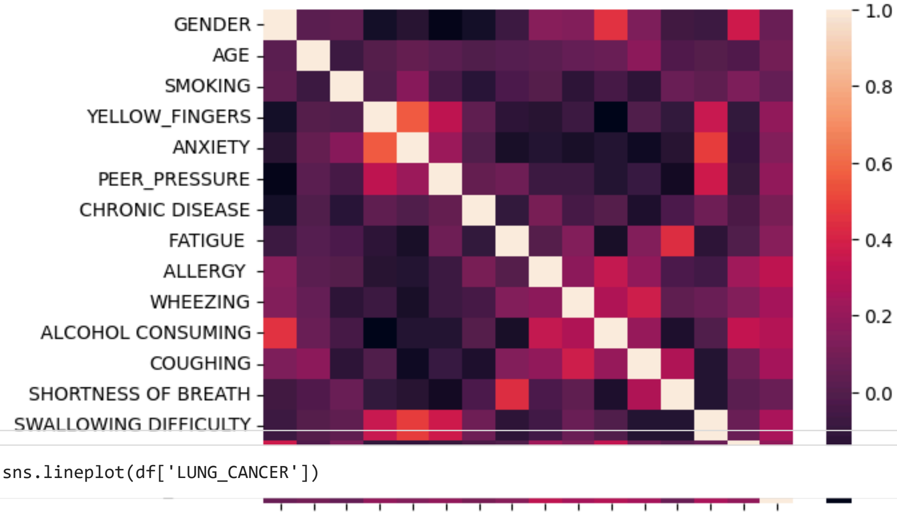
```
sns.scatterplot(df['LUNG_CANCER'])
```

```
<Axes: ylabel='LUNG_CANCER'>
```



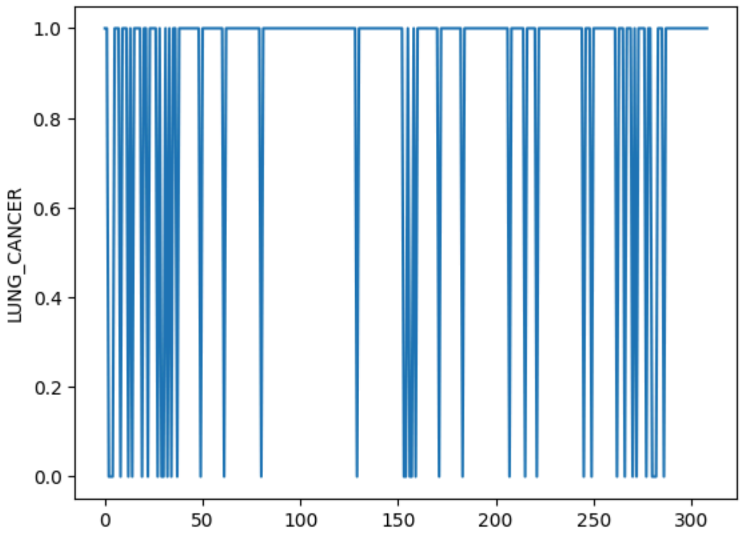
```
sns.heatmap(df.corr())
```

```
<Axes: >
```



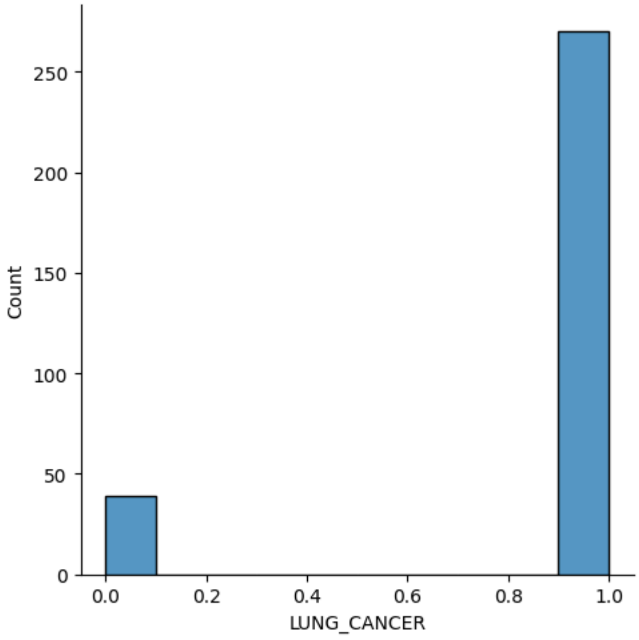
```
sns.lineplot(df['LUNG_CANCER'])
```

```
<Axes: ylabel='LUNG_CANCER'>
```



```
sns.displot(df['LUNG_CANCER'])
```

```
<seaborn.axisgrid.FacetGrid at 0x7ff126313530>
```



```
sns.relplot(df['LUNG_CANCER'])
```