In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.linear_model import Perceptron
from sklearn import linear_model
from sklearn import metrics
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

In [2]:
```
df=pd.read_excel("C:/Users/nikan/Downloads/cancer.xlsx")
```

In [3]:
```
df.head()
```

Out[3]:

	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	Occupational Hazards	Genetic Risk	Chronic Lung Disease	Balanced Diet	...	Fatigue	We
0	P1	33	1	2	4	5	4	3	2	2	...	3	
1	P10	17	1	3	1	5	3	4	2	2	...	1	
2	P100	35	1	4	5	6	5	5	4	6	...	8	
3	P1000	37	1	7	7	7	7	6	7	7	...	4	
4	P101	46	1	6	8	7	7	7	6	7	...	3	

5 rows × 25 columns

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 990 entries, 0 to 989
Data columns (total 25 columns):
 #   Column                    Non-Null Count  Dtype 
---  ------                    --------------  ----- 
 0   Patient Id                990 non-null    object
 1   Age                       990 non-null    int64 
 2   Gender                    990 non-null    int64 
 3   Air Pollution             990 non-null    int64 
 4   Alcohol use               990 non-null    int64 
 5   Dust Allergy              990 non-null    int64 
 6   Occupational Hazards      990 non-null    int64 
 7   Genetic Risk              990 non-null    int64 
 8   Chronic Lung Disease      990 non-null    int64 
 9   Balanced Diet             990 non-null    int64 
 10  Obesity                   990 non-null    int64 
 11  Smoking                   990 non-null    int64 
 12  Passive Smoker            990 non-null    int64 
 13  Chest Pain                990 non-null    int64 
 14  Coughing of Blood         990 non-null    int64 
 15  Fatigue                   990 non-null    int64 
 16  Weight Loss               990 non-null    int64 
 17  Shortness of Breath       990 non-null    int64 
 18  Wheezing                  990 non-null    int64 
 19  Swallowing Difficulty     990 non-null    int64 
 20  Clubbing of Finger Nails  990 non-null    int64 
 21  Frequent Cold             990 non-null    int64 
 22  Dry Cough                 990 non-null    int64 
```

```
 23  Snoring                  990 non-null    int64
 24  Level                    990 non-null    object
dtypes: int64(23), object(2)
memory usage: 193.5+ KB
```

In [5]:
```
df.describe()
```

Out[5]:

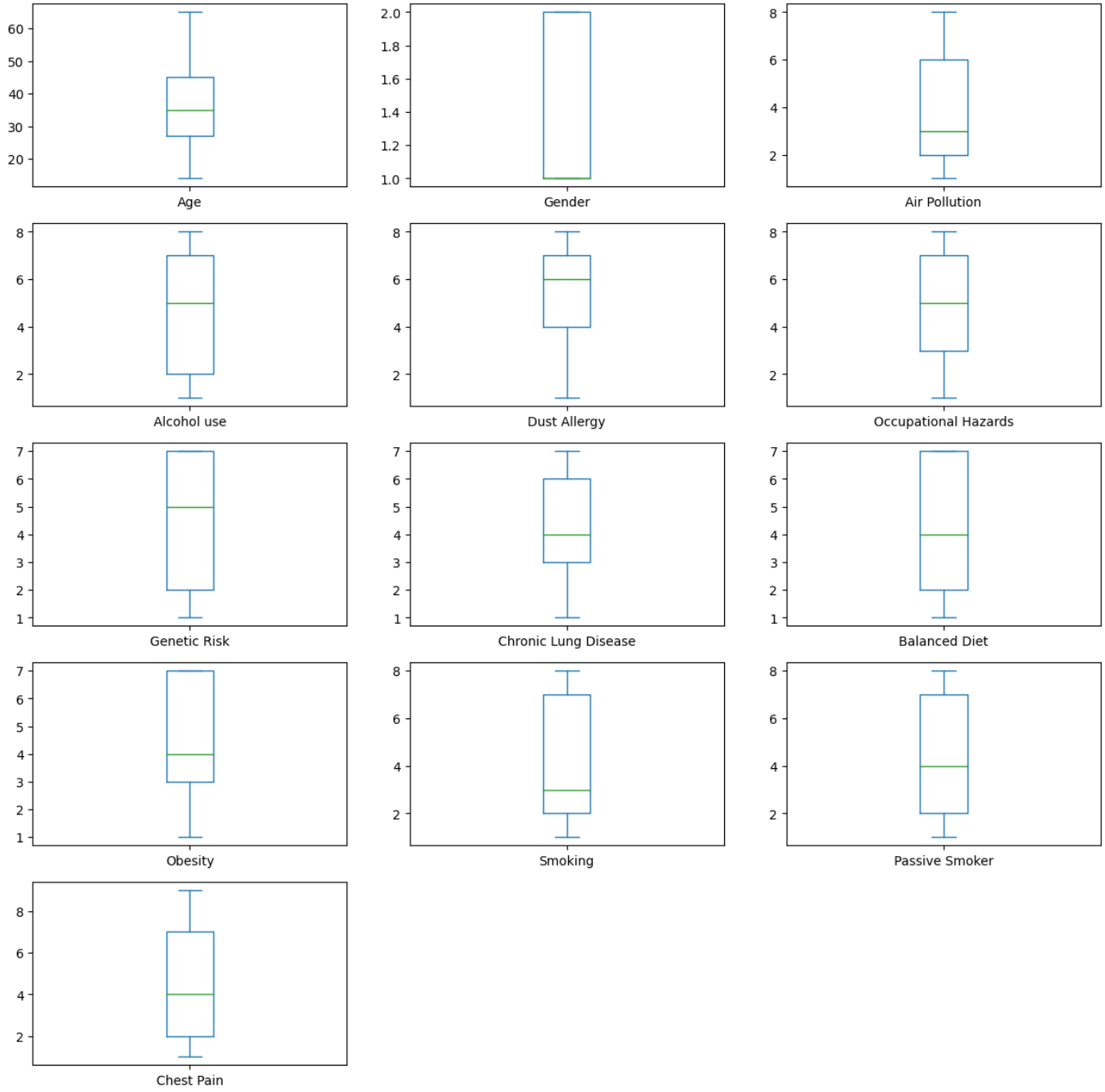
	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	Occupational Hazards	Genetic Risk	Chronic Lung Disease	Balanc D
count	990.000000	990.000000	990.000000	990.000000	990.000000	990.000000	990.000000	990.000000	990.0000
mean	36.812121	1.406061	3.828283	4.548485	5.156566	4.838384	4.565657	4.373737	4.4757
std	11.510010	0.491344	2.037269	2.629684	1.989033	2.118373	2.132903	1.856782	2.1408
min	14.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.0000
25%	27.000000	1.000000	2.000000	2.000000	4.000000	3.000000	2.000000	3.000000	2.0000
50%	35.000000	1.000000	3.000000	5.000000	6.000000	5.000000	5.000000	4.000000	4.0000
75%	45.000000	2.000000	6.000000	7.000000	7.000000	7.000000	7.000000	6.000000	7.0000
max	65.000000	2.000000	8.000000	8.000000	8.000000	8.000000	7.000000	7.000000	7.0000

8 rows × 23 columns

In [6]:
```
df=pd.read_excel("C:/Users/nikan/Downloads/cancer.xlsx")
```

In [7]:
```
boxplot=["Age","Gender","Air Pollution","Alcohol use","Dust Allergy","Occupational Hazar
    "Obesity","Smoking","Passive Smoker","Chest Pain"]
df[boxplot].plot(kind="box", subplots="True" , layout=(5, 3) ,figsize=(15,15))
```

Out[7]:
```
Age                    AxesSubplot(0.125,0.747241;0.227941x0.132759)
Gender                 AxesSubplot(0.398529,0.747241;0.227941x0.132759)
Air Pollution          AxesSubplot(0.672059,0.747241;0.227941x0.132759)
Alcohol use               AxesSubplot(0.125,0.587931;0.227941x0.132759)
Dust Allergy           AxesSubplot(0.398529,0.587931;0.227941x0.132759)
Occupational Hazards   AxesSubplot(0.672059,0.587931;0.227941x0.132759)
Genetic Risk              AxesSubplot(0.125,0.428621;0.227941x0.132759)
Chronic Lung Disease   AxesSubplot(0.398529,0.428621;0.227941x0.132759)
Balanced Diet          AxesSubplot(0.672059,0.428621;0.227941x0.132759)
Obesity                   AxesSubplot(0.125,0.26931;0.227941x0.132759)
Smoking                AxesSubplot(0.398529,0.26931;0.227941x0.132759)
Passive Smoker         AxesSubplot(0.672059,0.26931;0.227941x0.132759)
Chest Pain                AxesSubplot(0.125,0.11;0.227941x0.132759)
dtype: object
```

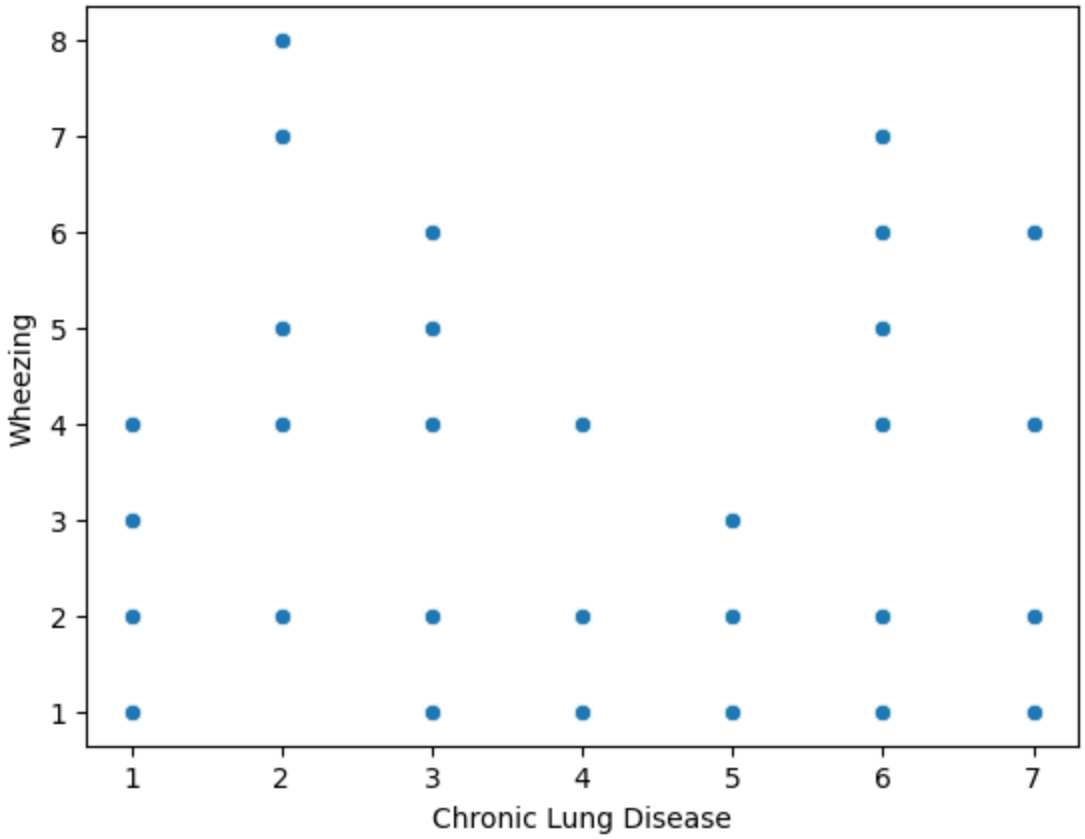


In [8]:
```
df.shape
```

Out[8]:
```
(990, 25)
```

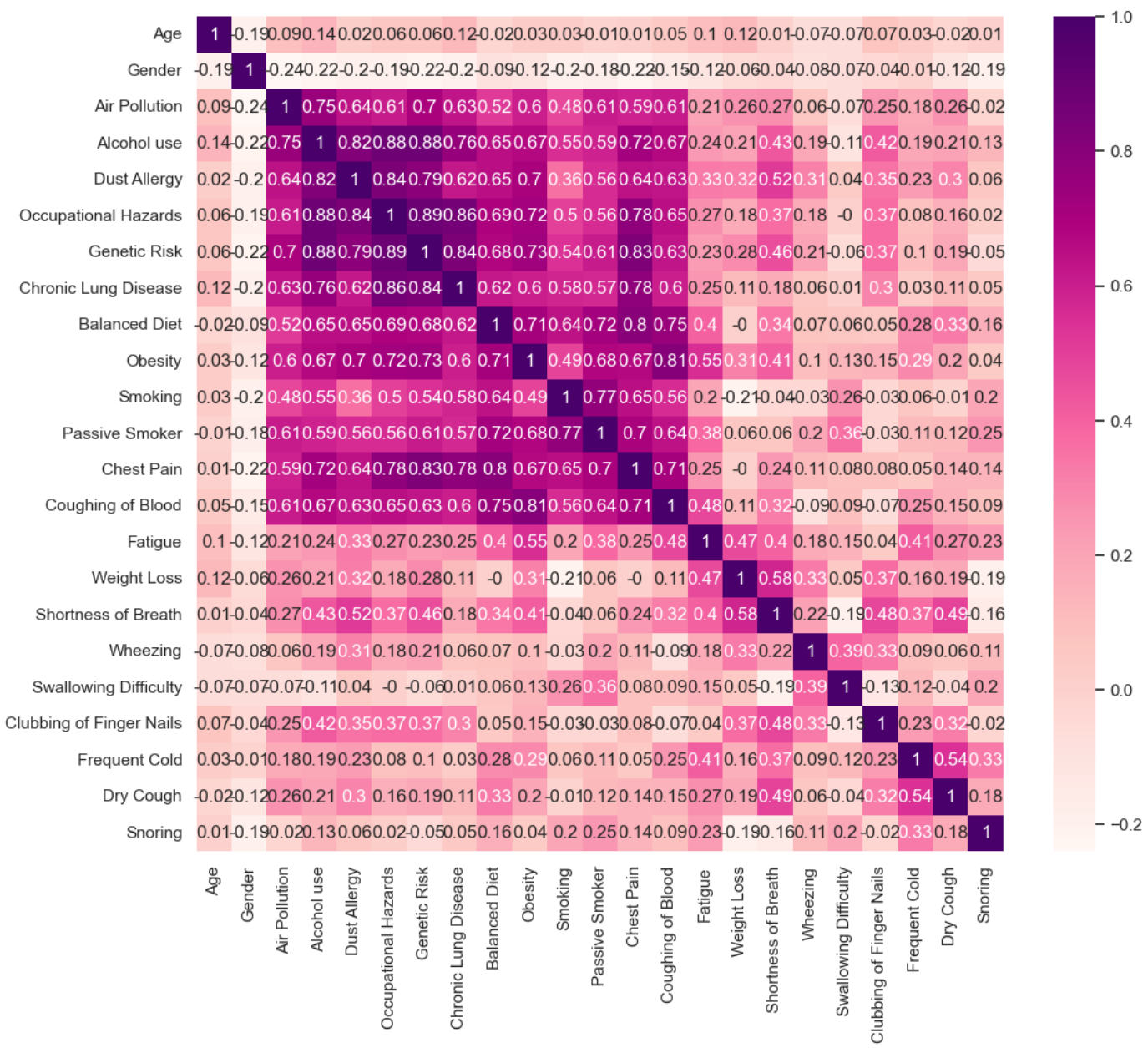
In [9]:
```
sns.scatterplot(data=df, x = "Chronic Lung Disease", y = "Wheezing")
```

Out[9]:
```
<AxesSubplot:xlabel='Chronic Lung Disease', ylabel='Wheezing'>
```



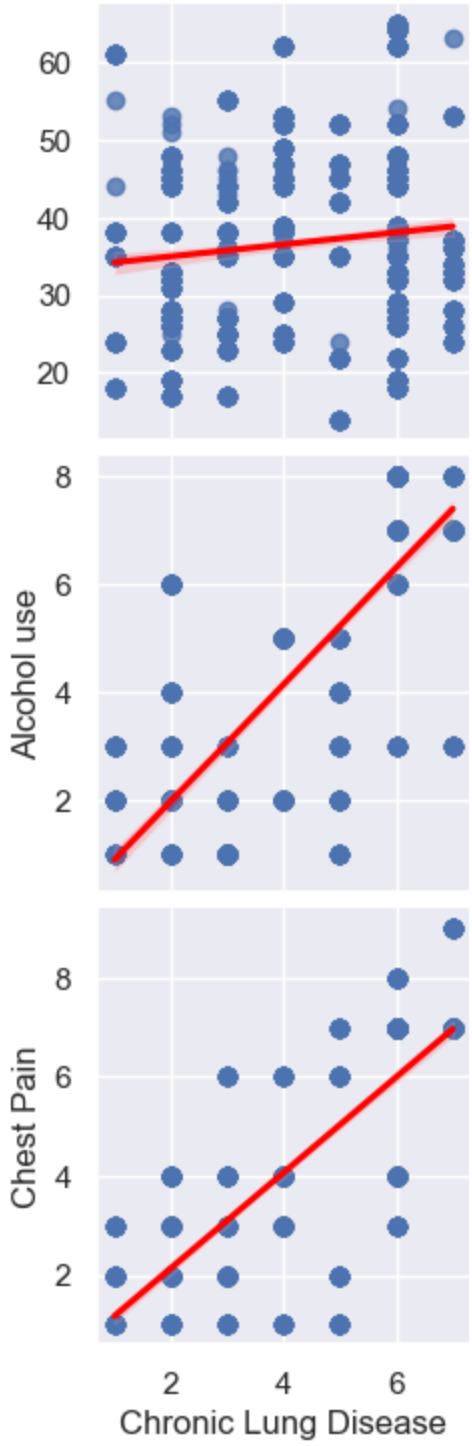
In [10]:
```
correlation_cancer = df.corr().round(2)

sns.set(rc={'figure.figsize':(12,10)})
sns.heatmap(data=correlation_cancer, cmap = 'RdPu',annot=True)
```

Out[10]:
```
<AxesSubplot:>
```



In [32]:
```
plt.figure()
cancer_graphic = sns.pairplot(df,x_vars=['Chronic Lung Disease'], y_vars=['Age', 'Alcoho
plt.show()
```

```
<Figure size 1200x1000 with 0 Axes>
```



In [34]:
```
correlations = df.corr()
print(correlations)
```

```
                            Age    Gender  Air Pollution  Alcohol use  \
Age                    1.000000 -0.186562       0.086372     0.141943
Gender                -0.186562  1.000000      -0.243406    -0.224195
Air Pollution          0.086372 -0.243406       1.000000     0.746492
Alcohol use            0.141943 -0.224195       0.746492     1.000000
Dust Allergy           0.023590 -0.201686       0.636691     0.818286
Occupational Hazards   0.062782 -0.192374       0.609510     0.879729
Genetic Risk           0.055693 -0.218430       0.704168     0.876817
Chronic Lung Disease   0.124643 -0.203087       0.626153     0.763311
Balanced Diet         -0.017296 -0.094445       0.522974     0.652072
Obesity                0.028041 -0.122181       0.601198     0.669148
Smoking                0.027996 -0.196772       0.479741     0.546180
Passive Smoker        -0.005873 -0.182674       0.606124     0.591909
Chest Pain             0.005691 -0.217189       0.585456     0.717187
Coughing of Blood      0.053731 -0.146525       0.608508     0.668318
```

Fatigue	0.097624	-0.116333	0.211707	0.237255
Weight Loss	0.124463	-0.061459	0.260883	0.210479
Shortness of Breath	0.012712	-0.039844	0.266347	0.433474
Wheezing	-0.072713	-0.084123	0.060728	0.186641
Swallowing Difficulty	-0.073232	-0.069092	-0.074607	-0.108335
Clubbing of Finger Nails	0.066845	-0.041166	0.246934	0.421480
Frequent Cold	0.030752	-0.012152	0.184632	0.190593
Dry Cough	-0.020709	-0.115314	0.257275	0.206915
Snoring	0.014965	-0.187834	-0.017780	0.126621

	Dust Allergy	Occupational Hazards	Genetic Risk	\
Age	0.023590	0.062782	0.055693	
Gender	-0.201686	-0.192374	-0.218430	
Air Pollution	0.636691	0.609510	0.704168	
Alcohol use	0.818286	0.879729	0.876817	
Dust Allergy	1.000000	0.836312	0.787541	
Occupational Hazards	0.836312	1.000000	0.894580	
Genetic Risk	0.787541	0.894580	1.000000	
Chronic Lung Disease	0.619035	0.858540	0.836328	
Balanced Diet	0.646398	0.692737	0.678376	
Obesity	0.700457	0.722250	0.730007	
Smoking	0.356875	0.503211	0.540758	
Passive Smoker	0.559364	0.555400	0.608466	
Chest Pain	0.639704	0.775690	0.832219	
Coughing of Blood	0.625617	0.645932	0.633284	
Fatigue	0.332504	0.267808	0.230622	
Weight Loss	0.323945	0.176662	0.275188	
Shortness of Breath	0.517417	0.367003	0.455393	
Wheezing	0.310026	0.180290	0.212138	
Swallowing Difficulty	0.036601	-0.001939	-0.055318	
Clubbing of Finger Nails	0.350612	0.368285	0.365272	
Frequent Cold	0.227706	0.078999	0.097601	
Dry Cough	0.297646	0.159983	0.188776	
Snoring	0.055682	0.023445	-0.052816	

	Chronic Lung Disease	Balanced Diet	Obesity	...	\
Age	0.124643	-0.017296	0.028041	...	
Gender	-0.203087	-0.094445	-0.122181	...	
Air Pollution	0.626153	0.522974	0.601198	...	
Alcohol use	0.763311	0.652072	0.669148	...	
Dust Allergy	0.619035	0.646398	0.700457	...	
Occupational Hazards	0.858540	0.692737	0.722250	...	
Genetic Risk	0.836328	0.678376	0.730007	...	
Chronic Lung Disease	1.000000	0.622161	0.601435	...	
Balanced Diet	0.622161	1.000000	0.707137	...	
Obesity	0.601435	0.707137	1.000000	...	
Smoking	0.581202	0.644036	0.489381	...	
Passive Smoker	0.572194	0.724910	0.681639	...	
Chest Pain	0.782496	0.798709	0.672945	...	
Coughing of Blood	0.603143	0.746538	0.814932	...	
Fatigue	0.247626	0.401242	0.552814	...	
Weight Loss	0.105534	-0.003788	0.314821	...	
Shortness of Breath	0.180458	0.339999	0.405590	...	
Wheezing	0.060431	0.070596	0.096905	...	
Swallowing Difficulty	0.011473	0.056004	0.131308	...	
Clubbing of Finger Nails	0.301916	0.047999	0.151690	...	
Frequent Cold	0.033800	0.277210	0.294854	...	
Dry Cough	0.111279	0.327133	0.199120	...	
Snoring	0.045621	0.157876	0.041116	...	

	Coughing of Blood	Fatigue	Weight Loss	\
Age	0.053731	0.097624	0.124463	
Gender	-0.146525	-0.116333	-0.061459	
Air Pollution	0.608508	0.211707	0.260883	
Alcohol use	0.668318	0.237255	0.210479	
Dust Allergy	0.625617	0.332504	0.323945	

Occupational Hazards	0.645932	0.267808	0.176662
Genetic Risk	0.633284	0.230622	0.275188
Chronic Lung Disease	0.603143	0.247626	0.105534
Balanced Diet	0.746538	0.401242	-0.003788
Obesity	0.814932	0.552814	0.314821
Smoking	0.561883	0.201686	-0.209544
Passive Smoker	0.636419	0.377932	0.059782
Chest Pain	0.712245	0.251057	-0.000127
Coughing of Blood	1.000000	0.481520	0.106167
Fatigue	0.481520	1.000000	0.470135
Weight Loss	0.106167	0.470135	1.000000
Shortness of Breath	0.319289	0.399333	0.579694
Wheezing	-0.085517	0.175720	0.329282
Swallowing Difficulty	0.087652	0.151474	0.049046
Clubbing of Finger Nails	-0.066189	0.041353	0.374843
Frequent Cold	0.247460	0.412838	0.156567
Dry Cough	0.147874	0.272018	0.193960
Snoring	0.088490	0.232625	-0.192096

	Shortness of Breath	Wheezing	\
Age	0.012712	-0.072713	
Gender	-0.039844	-0.084123	
Air Pollution	0.266347	0.060728	
Alcohol use	0.433474	0.186641	
Dust Allergy	0.517417	0.310026	
Occupational Hazards	0.367003	0.180290	
Genetic Risk	0.455393	0.212138	
Chronic Lung Disease	0.180458	0.060431	
Balanced Diet	0.339999	0.070596	
Obesity	0.405590	0.096905	
Smoking	-0.036500	-0.033348	
Passive Smoker	0.060455	0.204256	
Chest Pain	0.235906	0.109834	
Coughing of Blood	0.319289	-0.085517	
Fatigue	0.399333	0.175720	
Weight Loss	0.579694	0.329282	
Shortness of Breath	1.000000	0.215815	
Wheezing	0.215815	1.000000	
Swallowing Difficulty	-0.193067	0.387193	
Clubbing of Finger Nails	0.483581	0.333556	
Frequent Cold	0.366927	0.087866	
Dry Cough	0.489343	0.064260	
Snoring	-0.155176	0.111305	

	Swallowing Difficulty	Clubbing of Finger Nails	\
Age	-0.073232	0.066845	
Gender	-0.069092	-0.041166	
Air Pollution	-0.074607	0.246934	
Alcohol use	-0.108335	0.421480	
Dust Allergy	0.036601	0.350612	
Occupational Hazards	-0.001939	0.368285	
Genetic Risk	-0.055318	0.365272	
Chronic Lung Disease	0.011473	0.301916	
Balanced Diet	0.056004	0.047999	
Obesity	0.131308	0.151690	
Smoking	0.261416	-0.028402	
Passive Smoker	0.356040	-0.032829	
Chest Pain	0.075381	0.083694	
Coughing of Blood	0.087652	-0.066189	
Fatigue	0.151474	0.041353	
Weight Loss	0.049046	0.374843	
Shortness of Breath	-0.193067	0.483581	
Wheezing	0.387193	0.333556	
Swallowing Difficulty	1.000000	-0.130988	
Clubbing of Finger Nails	-0.130988	1.000000	
Frequent Cold	0.117445	0.234303	

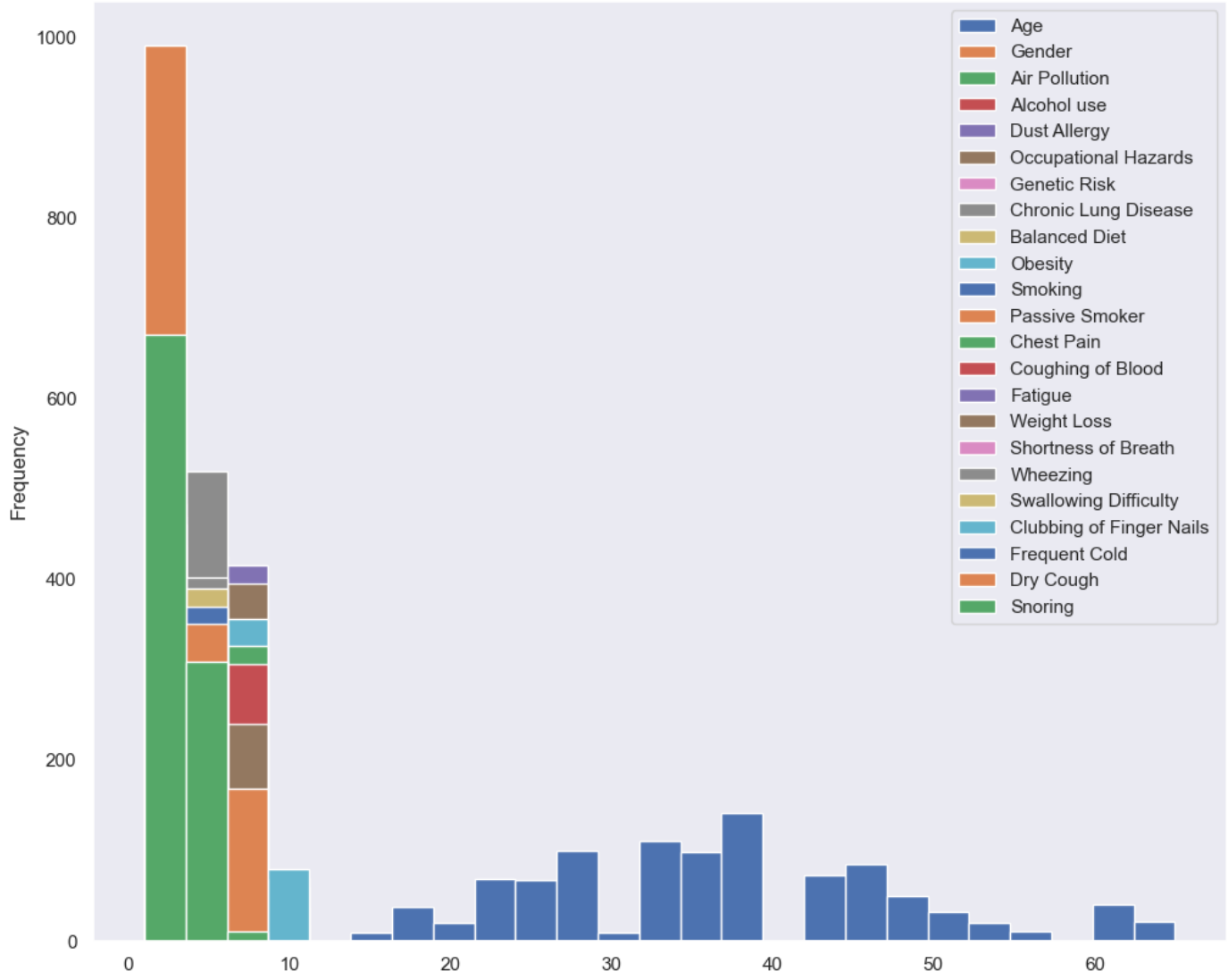
```
Dry Cough                          -0.043112                 0.318676
Snoring                             0.204786                -0.022770
```

	Frequent Cold	Dry Cough	Snoring
Age	0.030752	-0.020709	0.014965
Gender	-0.012152	-0.115314	-0.187834
Air Pollution	0.184632	0.257275	-0.017780
Alcohol use	0.190593	0.206915	0.126621
Dust Allergy	0.227706	0.297646	0.055682
Occupational Hazards	0.078999	0.159983	0.023445
Genetic Risk	0.097601	0.188776	-0.052816
Chronic Lung Disease	0.033800	0.111279	0.045621
Balanced Diet	0.277210	0.327133	0.157876
Obesity	0.294854	0.199120	0.041116
Smoking	0.063770	-0.007322	0.202477
Passive Smoker	0.110569	0.117787	0.250808
Chest Pain	0.046856	0.140388	0.141927
Coughing of Blood	0.247460	0.147874	0.088490
Fatigue	0.412838	0.272018	0.232625
Weight Loss	0.156567	0.193960	-0.192096
Shortness of Breath	0.366927	0.489343	-0.155176
Wheezing	0.087866	0.064260	0.111305
Swallowing Difficulty	0.117445	-0.043112	0.204786
Clubbing of Finger Nails	0.234303	0.318676	-0.022770
Frequent Cold	1.000000	0.538895	0.330931
Dry Cough	0.538895	1.000000	0.184232
Snoring	0.330931	0.184232	1.000000

```
[23 rows x 23 columns]
```

In [39]:
```
plt.figure()
df.plot(kind = 'hist', bins=25, grid=False)
```

Out[39]:
```
<AxesSubplot:ylabel='Frequency'>

<Figure size 1200x1000 with 0 Axes>
```



In [ ]:
```
*Per feature engineering, the comorbidity of symptoms present in Cancer diagnosi is to b
```