In [16]:
```
import pandas as pd
```

In [17]:
```
import numpy as np
```

In [18]:
```
import matplotlib.pyplot as plt
```

In [19]:
```
data=pd.read_csv(r"C:\Users\EL BODA\Downloads\cancer patient data sets.csv")
```

In [20]:
```
data
```

Out[20]:

	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chr L Dis
0	0	P1	33	1	2	4	5	4	3	
1	1	P10	17	1	3	1	5	3	4	
2	2	P100	35	1	4	5	6	5	5	
3	3	P1000	37	1	7	7	7	7	6	
4	4	P101	46	1	6	8	7	7	7	
...	...	...	...	...	...	...	...	...	...	
995	995	P995	44	1	6	7	7	7	7	
996	996	P996	37	2	6	8	7	7	7	
997	997	P997	25	2	4	5	6	5	5	
998	998	P998	18	2	6	8	7	7	7	
999	999	P999	47	1	6	5	6	5	5	

1000 rows × 26 columns

In [21]:
```
data.head(10)
```

Out[21]:

	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chron Lun Diseas
0	0	P1	33	1	2	4	5	4	3	
1	1	P10	17	1	3	1	5	3	4	
2	2	P100	35	1	4	5	6	5	5	
3	3	P1000	37	1	7	7	7	7	6	
4	4	P101	46	1	6	8	7	7	7	
5	5	P102	35	1	4	5	6	5	5	
6	6	P103	52	2	2	4	5	4	3	
7	7	P104	28	2	3	1	4	3	2	
8	8	P105	35	2	4	5	6	5	6	
9	9	P106	46	1	2	3	4	2	4	

10 rows × 26 columns

In [22]:
```
data.tail(10)
```

Out[22]:

	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chr L Dis
990	990	P990	49	1	6	5	6	5	5	
991	991	P991	37	1	8	8	7	7	7	
992	992	P992	26	2	7	7	7	7	7	
993	993	P993	37	2	7	7	7	7	6	
994	994	P994	33	1	6	7	7	7	7	
995	995	P995	44	1	6	7	7	7	7	
996	996	P996	37	2	6	8	7	7	7	
997	997	P997	25	2	4	5	6	5	5	
998	998	P998	18	2	6	8	7	7	7	
999	999	P999	47	1	6	5	6	5	5	

10 rows × 26 columns

In [23]:
```
data.shape
```

Out[23]: (1000, 26)

In [24]:
```
data.sample()
```

Out[24]:

	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chr L Dis
802	802	P820	65	1	6	8	7	7	7	

1 rows × 26 columns

In [25]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 26 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   index                     1000 non-null   int64
 1   Patient Id                1000 non-null   object
 2   Age                       1000 non-null   int64
 3   Gender                    1000 non-null   int64
 4   Air Pollution             1000 non-null   int64
 5   Alcohol use               1000 non-null   int64
 6   Dust Allergy              1000 non-null   int64
 7   OccuPational Hazards      1000 non-null   int64
 8   Genetic Risk              1000 non-null   int64
 9   chronic Lung Disease      1000 non-null   int64
 10  Balanced Diet             1000 non-null   int64
 11  Obesity                   1000 non-null   int64
 12  Smoking                   1000 non-null   int64
 13  Passive Smoker            1000 non-null   int64
 14  Chest Pain                1000 non-null   int64
 15  Coughing of Blood         1000 non-null   int64
 16  Fatigue                   1000 non-null   int64
 17  Weight Loss               1000 non-null   int64
 18  Shortness of Breath       1000 non-null   int64
 19  Wheezing                  1000 non-null   int64
 20  Swallowing Difficulty     1000 non-null   int64
 21  Clubbing of Finger Nails  1000 non-null   int64
 22  Frequent Cold             1000 non-null   int64
 23  Dry Cough                 1000 non-null   int64
 24  Snoring                   1000 non-null   int64
 25  Level                     1000 non-null   object
dtypes: int64(24), object(2)
memory usage: 203.3+ KB
```

In [26]:
```
data.describe()
```

Out[26]:

	index	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPa H
count	1000.000000	1000.000000	1000.000000	1000.0000	1000.000000	1000.000000	1000.0
mean	499.500000	37.174000	1.402000	3.8400	4.563000	5.165000	4.8
std	288.819436	12.005493	0.490547	2.0304	2.620477	1.980833	2.1
min	0.000000	14.000000	1.000000	1.0000	1.000000	1.000000	1.0
25%	249.750000	27.750000	1.000000	2.0000	2.000000	4.000000	3.0
50%	499.500000	36.000000	1.000000	3.0000	5.000000	6.000000	5.0
75%	749.250000	45.000000	2.000000	6.0000	7.000000	7.000000	7.0
max	999.000000	73.000000	2.000000	8.0000	8.000000	8.000000	8.0

8 rows × 24 columns

In [27]:
```
data.isnull()
```

Out[27]:

	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	ch I Dis
0	False	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	...	
995	False	False	False	False	False	False	False	False	False	
996	False	False	False	False	False	False	False	False	False	
997	False	False	False	False	False	False	False	False	False	
998	False	False	False	False	False	False	False	False	False	
999	False	False	False	False	False	False	False	False	False	

1000 rows × 26 columns

In [28]:
```
data.isnull().sum()
```

Out[28]:
```
index                       0
Patient Id                  0
Age                         0
Gender                      0
Air Pollution               0
Alcohol use                 0
Dust Allergy                0
OccuPational Hazards        0
Genetic Risk                0
chronic Lung Disease        0
Balanced Diet               0
Obesity                     0
Smoking                     0
Passive Smoker              0
Chest Pain                  0
Coughing of Blood           0
Fatigue                     0
Weight Loss                 0
Shortness of Breath         0
Wheezing                    0
Swallowing Difficulty       0
Clubbing of Finger Nails    0
Frequent Cold               0
Dry Cough                   0
Snoring                     0
Level                       0
dtype: int64
```

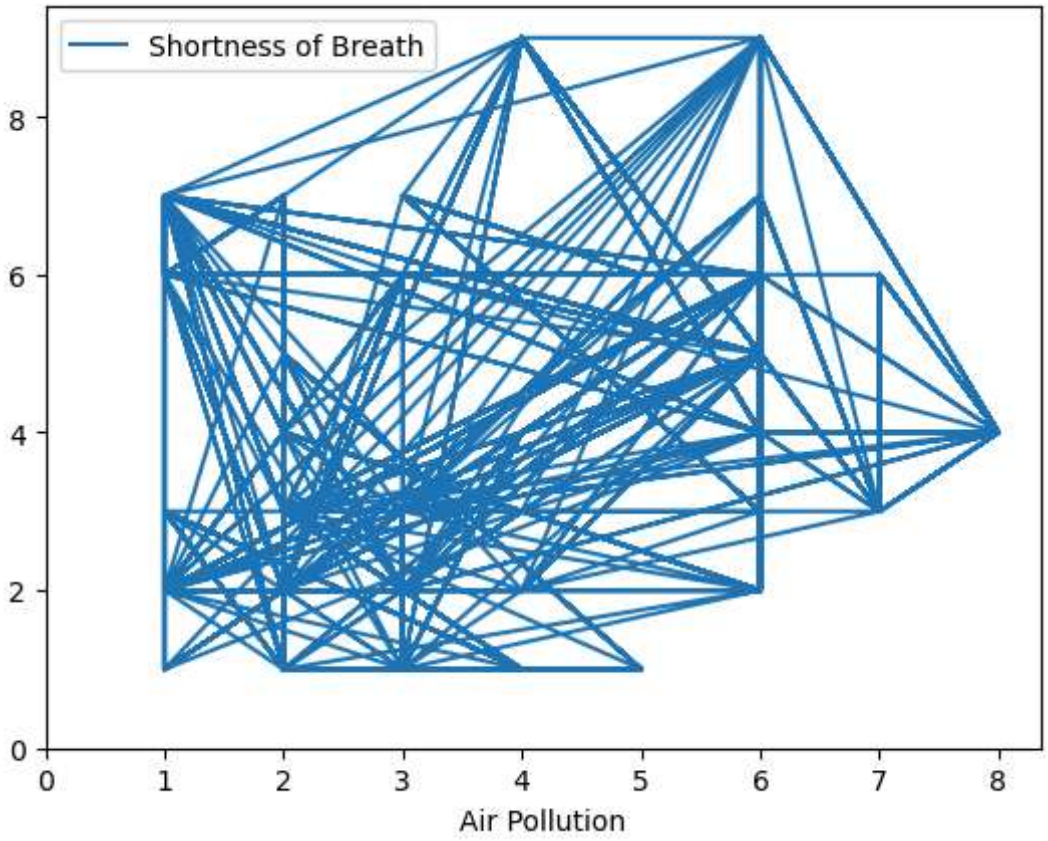
In [29]:
```
data.dropna(axis=0,inplace=True)
data
```

Out[29]:

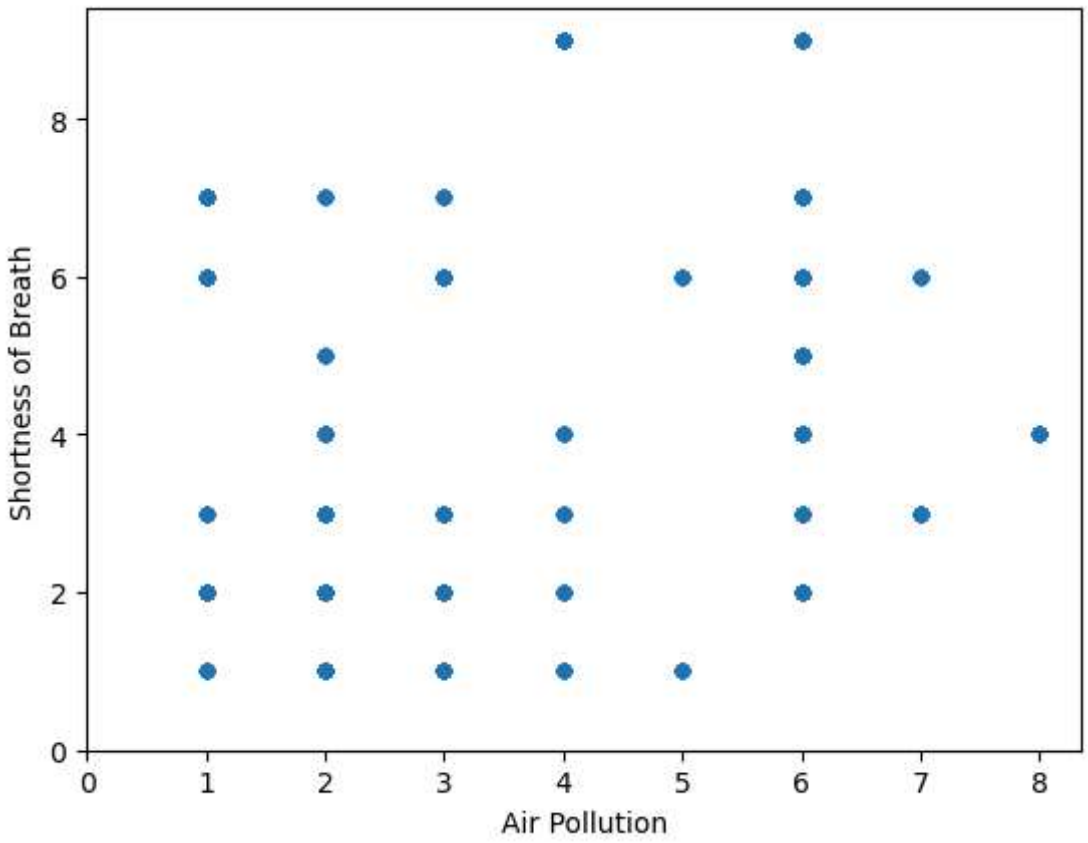
	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chr L Dis
0	0	P1	33	1	2	4	5	4	3	
1	1	P10	17	1	3	1	5	3	4	
2	2	P100	35	1	4	5	6	5	5	
3	3	P1000	37	1	7	7	7	7	6	
4	4	P101	46	1	6	8	7	7	7	
...	...	...	...	...	...	...	...	...	...	
995	995	P995	44	1	6	7	7	7	7	
996	996	P996	37	2	6	8	7	7	7	
997	997	P997	25	2	4	5	6	5	5	
998	998	P998	18	2	6	8	7	7	7	
999	999	P999	47	1	6	5	6	5	5	

1000 rows × 26 columns

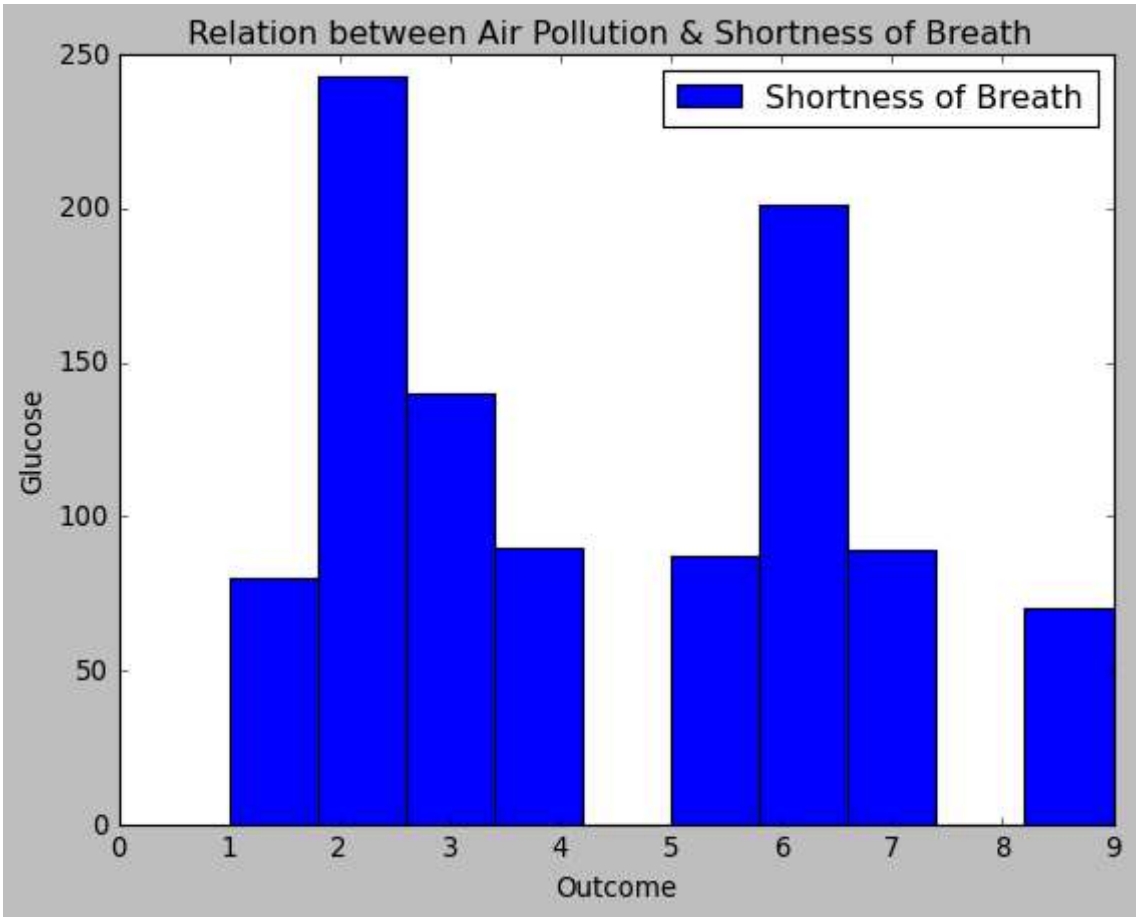
In [30]:
```
data.plot(x='Air Pollution',y='Shortness of Breath',kind='line'),
plt.ylim(ymin=0)
plt.xlim(xmin=0)
plt.show()
```



In [31]:
```
data.plot(x='Air Pollution',y='Shortness of Breath',kind='scatter'),
plt.ylim(ymin=0)
plt.xlim(xmin=0)
plt.style.use('classic')
plt.show()
```



In [32]:
```
data.plot(x='Air Pollution',y='Shortness of Breath',kind='hist'),
plt.ylim(ymin=0)
plt.xlim(xmin=0)
plt.ylabel('Glucose')
plt.xlabel('Outcome')
plt.title("Relation between Air Pollution & Shortness of Breath ")
plt.show()
```



In [33]:
```
x=data["Air Pollution"]
y=data["Shortness of Breath"]
slope_intercept=np.polyfit(x,y,1)
print(slope_intercept)
```

```
[0.30337024 3.07505828]
```

In [34]:
```
import sys
```

In [35]:
```
import matplotlib
```

In [36]:
```
matplotlib.use('Agg')
```

In [37]:
```
from scipy import stats
```

In [38]:
```
x=data["Air Pollution"]
y=data["Shortness of Breath"]
slope,intercept,r,p,std_err=stats.linregress(x,y)
def myfunc(x):
    return slope * x+ intercept
mymodel=list(map(myfunc,x))
plt.scatter(x,y)
plt.plot(x,mymodel)
plt.ylim(ymin=0)
plt.xlim(xmin=0)
plt.ylabel('Air Pollution')
```

```
plt.xlabel('Shortness of Breath')
plt.savefig("output_plot.png")
```