In [2]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:
```
df = pd.read_csv(r"C:\Users\asus\Downloads\cancer patient data sets.csv")
df
```

Out[3]:

	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	...	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Sv
0	0	P1	33	Male	2	4	5	4	3	2	...	3	4	2	2	
1	1	P10	17	Male	3	1	5	3	4	2	...	1	3	7	8	
2	2	P100	35	Male	4	5	6	5	5	4	...	8	7	9	2	
3	3	P1000	37	Male	7	7	7	7	6	7	...	4	2	3	1	
4	4	P101	46	Male	6	8	7	7	7	6	...	3	2	4	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
995	995	P995	44	Male	6	7	7	7	7	6	...	5	3	2	7	
996	996	P996	37	Female	6	8	7	7	7	6	...	9	6	5	7	
997	997	P997	25	Female	4	5	6	5	5	4	...	8	7	9	2	
998	998	P998	18	Female	6	8	7	7	7	6	...	3	2	4	1	
999	999	P999	47	Male	6	5	6	5	5	4	...	8	7	9	2	

1000 rows × 26 columns

In [4]:
```
s = df.head(10)
s
```

Out[4]:

	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	...	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swal Di
0	0	P1	33	Male	2	4	5	4	3	2	...	3	4	2	2	
1	1	P10	17	Male	3	1	5	3	4	2	...	1	3	7	8	
2	2	P100	35	Male	4	5	6	5	5	4	...	8	7	9	2	
3	3	P1000	37	Male	7	7	7	7	6	7	...	4	2	3	1	
4	4	P101	46	Male	6	8	7	7	7	6	...	3	2	4	1	
5	5	P102	35	Male	4	5	6	5	5	4	...	8	7	9	2	
6	6	P103	52	Female	2	4	5	4	3	2	...	3	4	2	2	
7	7	P104	28	Female	3	1	4	3	2	3	...	3	2	2	4	
8	8	P105	35	Female	4	5	6	5	6	5	...	1	4	3	2	
9	9	P106	46	Male	2	3	4	2	4	3	...	1	2	4	6	

10 rows × 26 columns

In [5]:
```
df.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   object
 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(23), object(3)
memory usage: 203.2+ KB
```

In [6]:
```
s1 = df.tail(10)
s1
```

Out[6]:

	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	...	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Sv
990	990	P990	49	Male	6	5	6	5	5	4	...	8	7	9	2	
991	991	P991	37	Male	8	8	7	7	7	6	...	3	2	4	1	
992	992	P992	26	Female	7	7	7	7	7	6	...	2	7	6	7	
993	993	P993	37	Female	7	7	7	7	6	7	...	4	2	3	1	
994	994	P994	33	Male	6	7	7	7	7	7	...	8	5	7	6	
995	995	P995	44	Male	6	7	7	7	7	6	...	5	3	2	7	
996	996	P996	37	Female	6	8	7	7	7	6	...	9	6	5	7	
997	997	P997	25	Female	4	5	6	5	5	4	...	8	7	9	2	
998	998	P998	18	Female	6	8	7	7	7	6	...	3	2	4	1	
999	999	P999	47	Male	6	5	6	5	5	4	...	8	7	9	2	

10 rows × 26 columns

In [7]:
```
df.shape
```

Out[7]:
```
(1000, 26)
```

In [8]:
```
df.describe()
```

Out[8]:

	index	Age	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Obesity	...
count	1000.000000	1000.000000	1000.0000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	...
mean	499.500000	37.174000	3.8400	4.563000	5.165000	4.840000	4.580000	4.380000	4.491000	4.465000	...
std	288.819436	12.005493	2.0304	2.620477	1.980833	2.107805	2.126999	1.848518	2.135528	2.124921	...
min	0.000000	14.000000	1.0000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...
25%	249.750000	27.750000	2.0000	2.000000	4.000000	3.000000	2.000000	3.000000	2.000000	3.000000	...
50%	499.500000	36.000000	3.0000	5.000000	6.000000	5.000000	5.000000	4.000000	4.000000	4.000000	...
75%	749.250000	45.000000	6.0000	7.000000	7.000000	7.000000	7.000000	6.000000	7.000000	7.000000	...
max	999.000000	73.000000	8.0000	8.000000	8.000000	8.000000	7.000000	7.000000	7.000000	7.000000	...

8 rows × 23 columns

In [9]:
```
df.isnull().sum()
```

Out[9]:
```
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
```

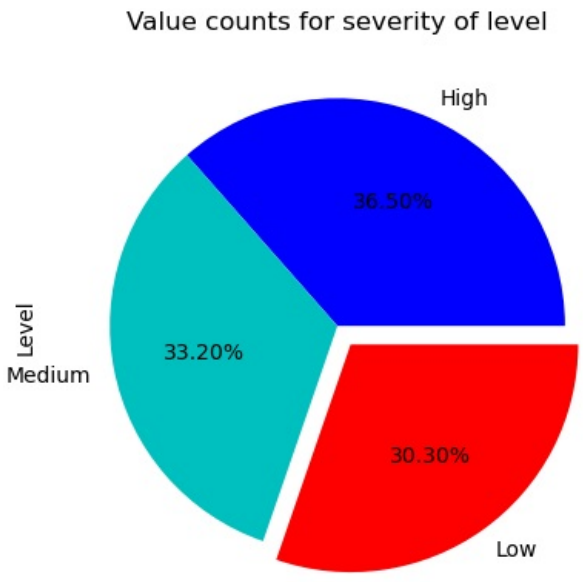
In [10]:
```
df['Gender'].value_counts()
```

Out[10]:
```
Male      598
Female    402
Name: Gender, dtype: int64
```

In [11]:
```
df['Level'].value_counts()
```

Out[11]:
```
High      365
Medium    332
Low       303
Name: Level, dtype: int64
```

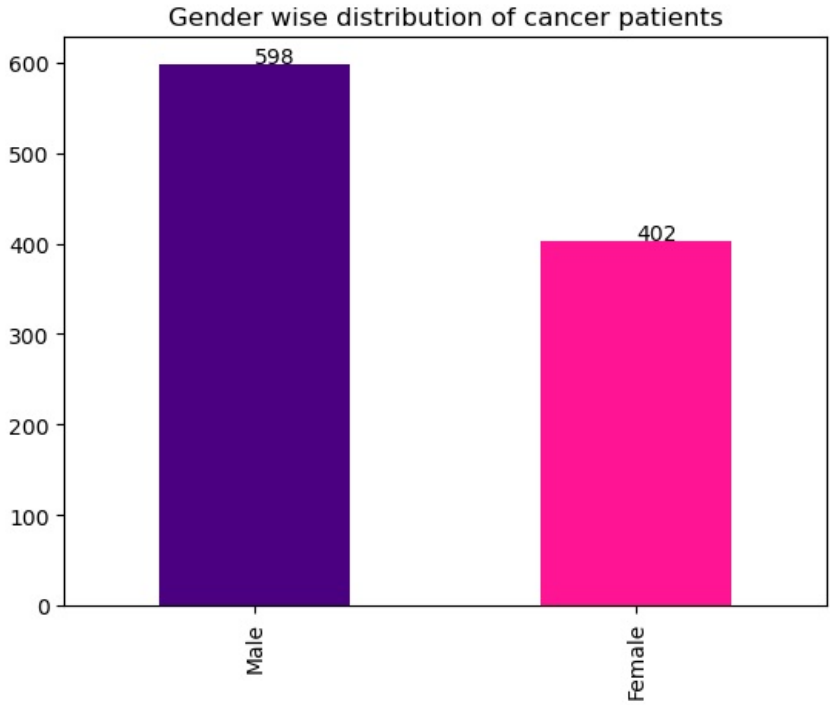
In [12]:
```
df['Level'].value_counts().plot(kind = 'pie',autopct = '%0.2f%%',explode = (0,0,0.1),colors = ('b','c','r'))
plt.title('Value counts for severity of level')
plt.show()
```



In [ ]:
```
This pie plot represents counts for severity of level. It shows that low level has minimum count.
```

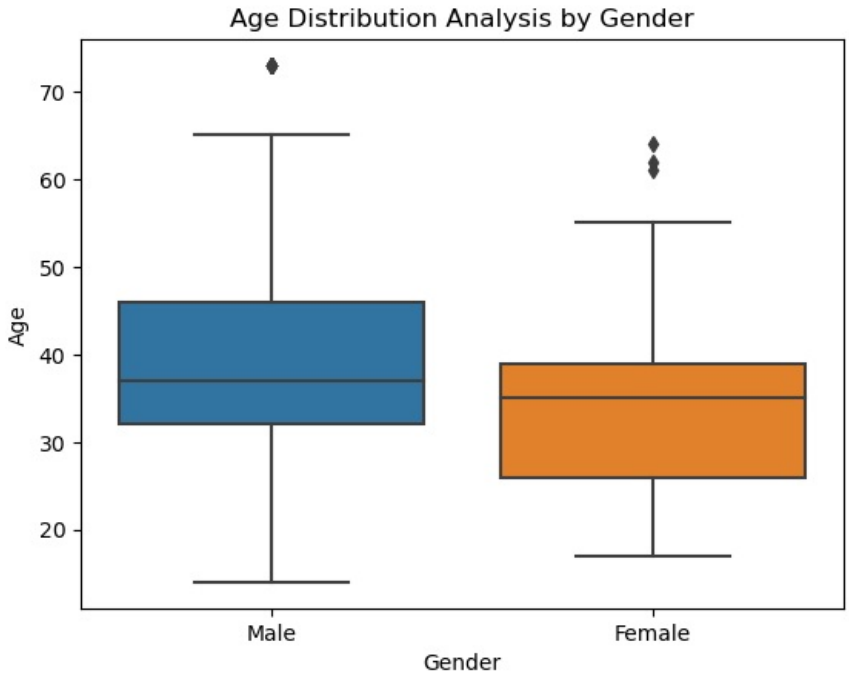
In [15]:
```
df['Gender'].value_counts().plot(kind = 'bar',color=['indigo','deeppink'])
plt.title('Gender wise distribution of cancer patients')
plt.text(0,600,'598',color = 'black')
plt.text(1,405,'402',color = 'black')
```

Out[15]:
```
Text(1, 405, '402')
```



In [ ]:
```
The above plot shows that there are 598 "male" cancer patients and 402 as "Female."
This means that there are more male patients who are affected by cancer than female
cancer patients.
```

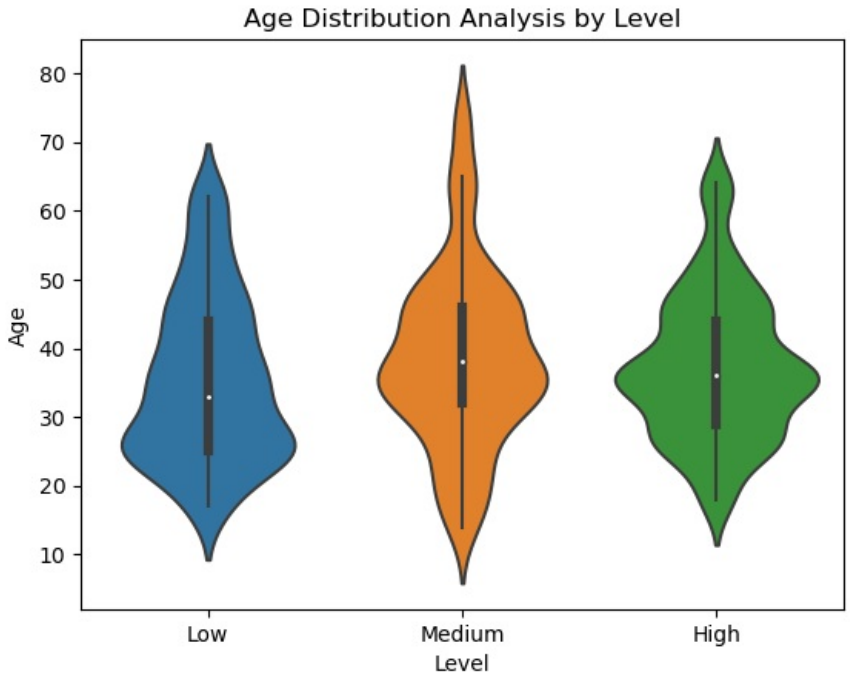
In [16]:
```
sns.boxplot(data = df,x = 'Gender',y = 'Age')
plt.title("Age Distribution Analysis by Gender")
plt.show()
```



In [ ]:
```
The box plot represents distribution of ages for cancer patients.
For both male and female cancer patients, the box plot shows the spread
and central tendency of ages.
The central box in each group represents the middle 50% of the age data.
The line within the box indicates the median age for each gender group.
```

In [17]:
```
sns.violinplot(data = df,x = 'Level',y = 'Age')
plt.title("Age Distribution Analysis by Level")
```

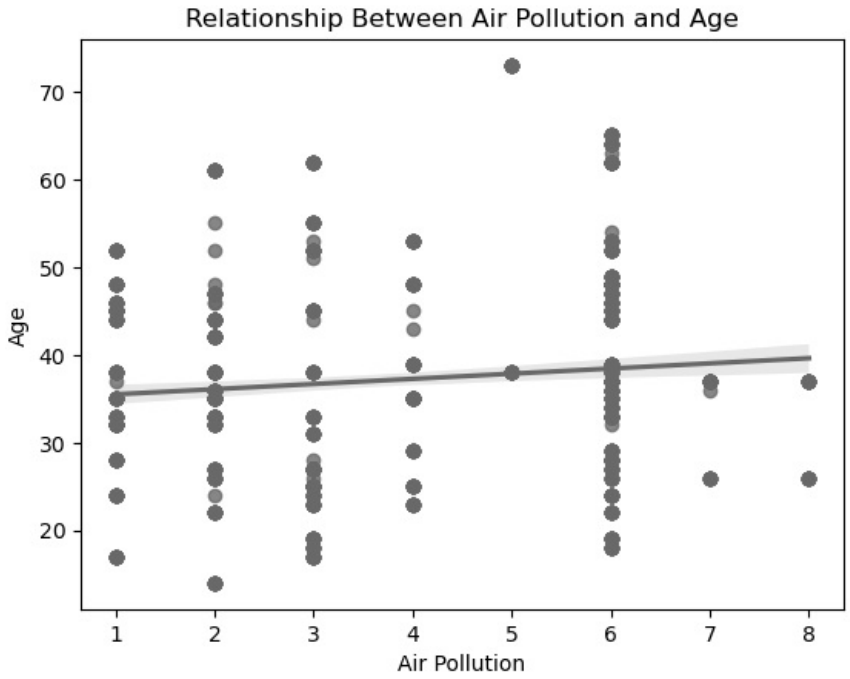
Out[17]:
```
Text(0.5, 1.0, 'Age Distribution Analysis by Level')
```



In [ ]:
```
The above violin plot help us to understand that there are significant age
differences among the severity levels.
It helps to identify the certain age groups are more prone to a
particular severity level.
```

In [18]:
```
sns.regplot(data = df,x = 'Air Pollution',y = 'Age',color="dimgrey");
plt.title("Relationship Between Air Pollution and Age")
```

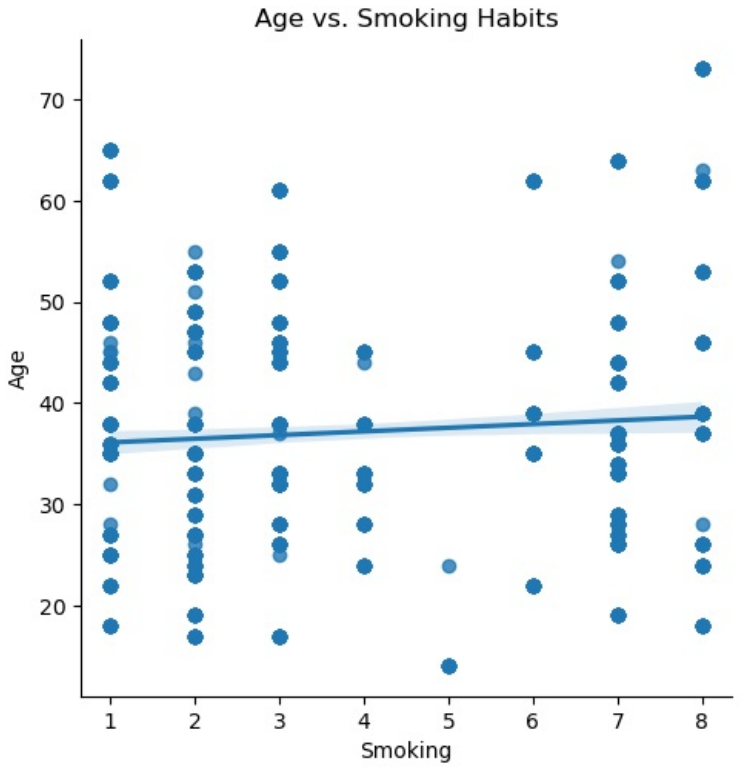
Out[18]:
```
Text(0.5, 1.0, 'Relationship Between Air Pollution and Age')
```



In [ ]:
```
The above reg plot represents The scatterplot displays individual data points,
where each point represents a cancer patient's age and the corresponding air
pollution level they were exposed to.
The regression line (the line of best fit) is drawn through the data points,
indicating the overall trend in the relationship between "Air Pollution" and "Age."
the regression line slopes upward,it suggests that as air pollution levels
increase, patient ages tend to increase (a positive correlation).
Conversely, if the line slopes downward, it suggests a negative correlation.
```

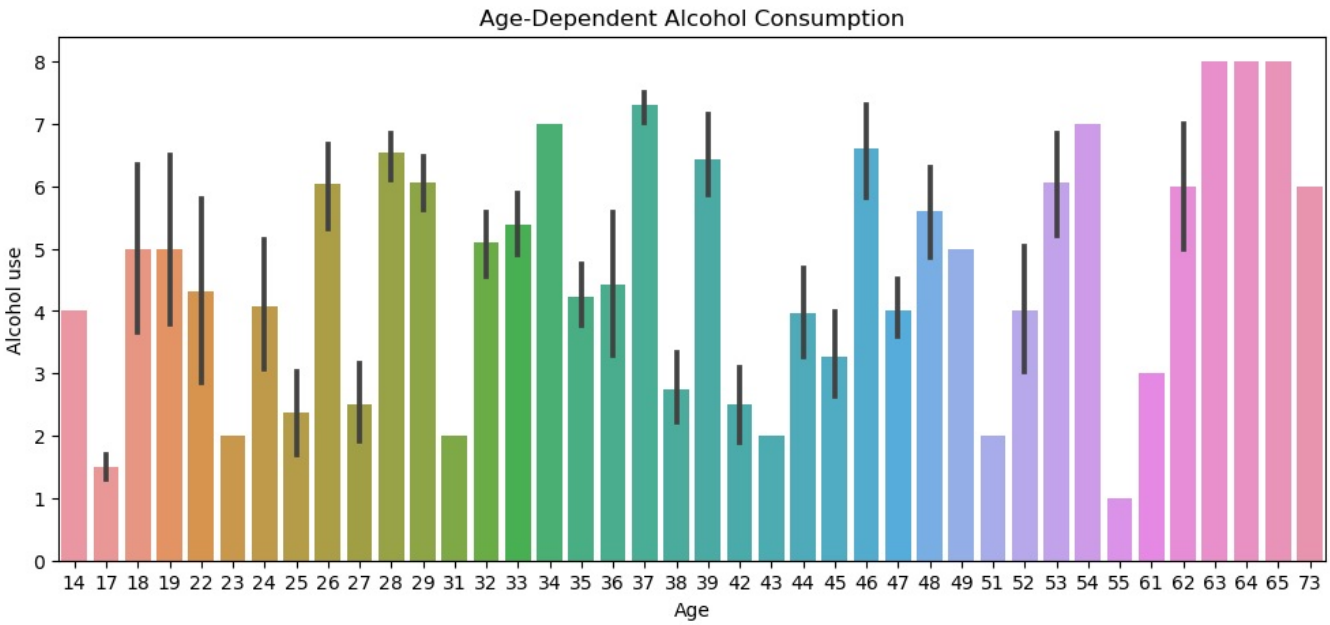
In [19]:
```
sns.lmplot(data = df,y = 'Age',x = 'Smoking')
plt.title("Age vs. Smoking Habits")
```

Out[19]:
```
Text(0.5, 1.0, 'Age vs. Smoking Habits')
```



In [ ]:
```
The scatterplot displays individual data points, where each point represents
a cancer patient's age and their smoking habits.
The "Smoking" variable could represent different smoking habits, such as
frequency or intensity.
The regression line (the line of best fit) is drawn through the data points,
indicating the overall trend in the relationship between "Age" and "Smoking."
The plot helps visualize that there is a significant relationship and
trend between a patient's age and their smoking habits.
```

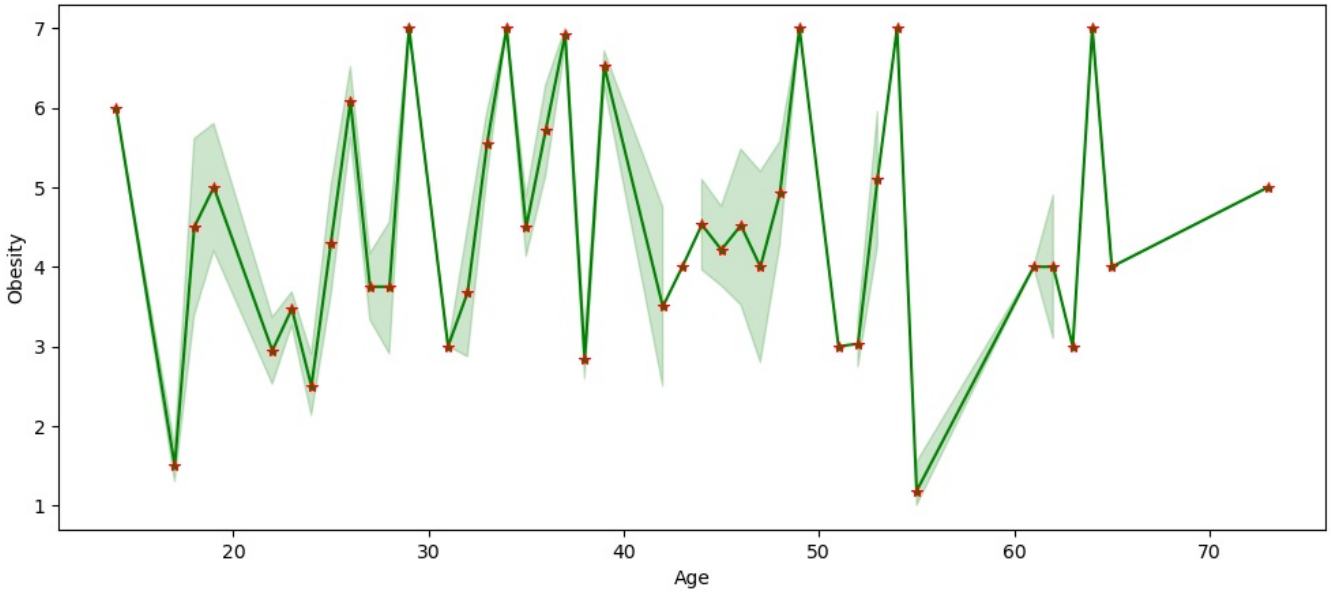
In [20]:
```
plt.figure(figsize = (12,5))
sns.barplot(data = df,x = 'Age',y = 'Alcohol use')
plt.title("Age-Dependent Alcohol Consumption")
plt.show()
```



In [ ]:
```
The bar plot visualizes the relationship between "Age" and "Alcohol use"
among cancer patients.
Each bar on the plot represents a specific age group, and the height of the bar indicates
the level of alcohol use for patientsin that age group.
The plot allows us to observe how alcohol use varies across
different age groups of cancer patients.
By examining the height of the bars, we can assess which age group has higher
or lower levels of alcohol consumption.
```

In [21]:
```
plt.figure(figsize = (12,5))
```

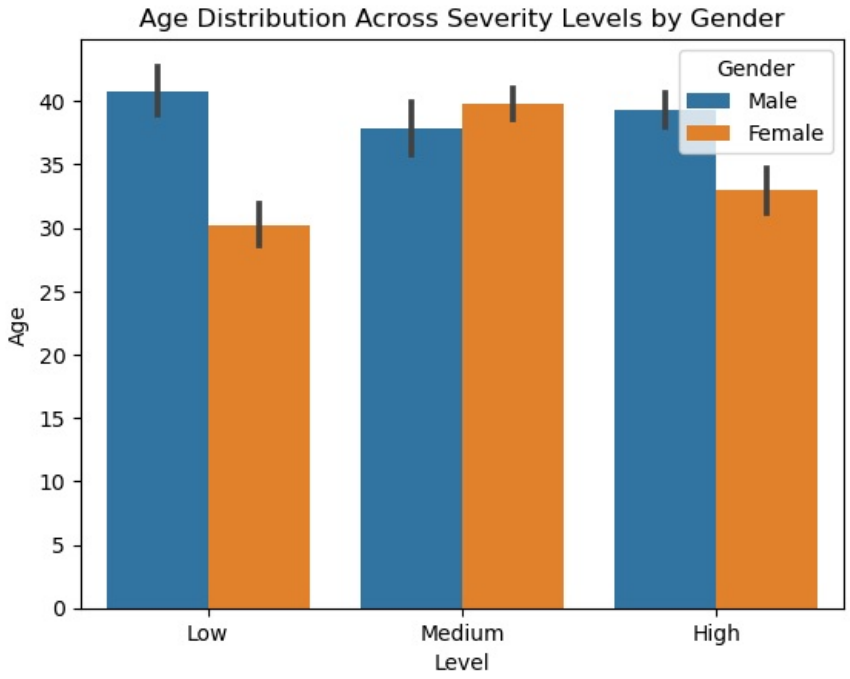
```
sns.lineplot(data = df,x = 'Age',y = 'Obesity',marker = '*',mec = 'red',color = 'green');
```



In [ ]:
```
The line plot show how "Obesity" levels vary with "Age"
for the given cancer patients.
the line has an upward trend, it suggests that, on average, obesity tends
to increase as patients get older.
This could indicate that older cancer patients are more likely to be obese.
the line has a downward trend, it suggests that obesity tends to decrease
as patients get older.
This could imply that younger cancer patients are more likely to be obese.
A relatively flat line suggests that there may not be a strong correlation
between age and obesity among the cancer patients.
```

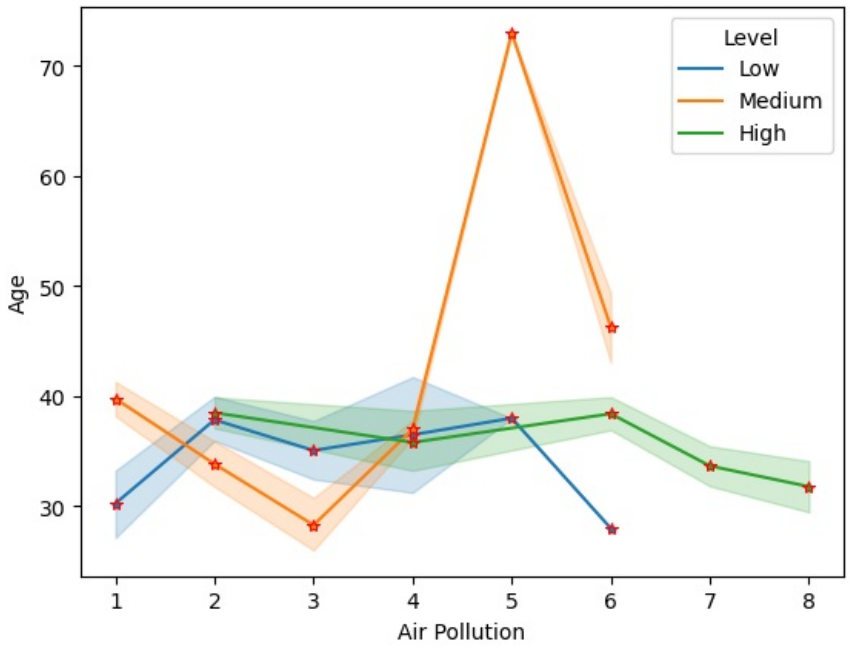
In [22]:
```
sns.barplot(data = df,x = 'Level',y = 'Age',hue = 'Gender')
plt.title("Age Distribution Across Severity Levels by Gender")
```

Out[22]:
```
Text(0.5, 1.0, 'Age Distribution Across Severity Levels by Gender')
```



In [ ]:
```
The above barplot shows
The "Low" severity level appears to have a mix of male and female patients with varying ages.
The "Medium" severity level shows the age distribution for both male and female patients.
The "High" severity level also exhibits age distribution with contributions from both genders.
```

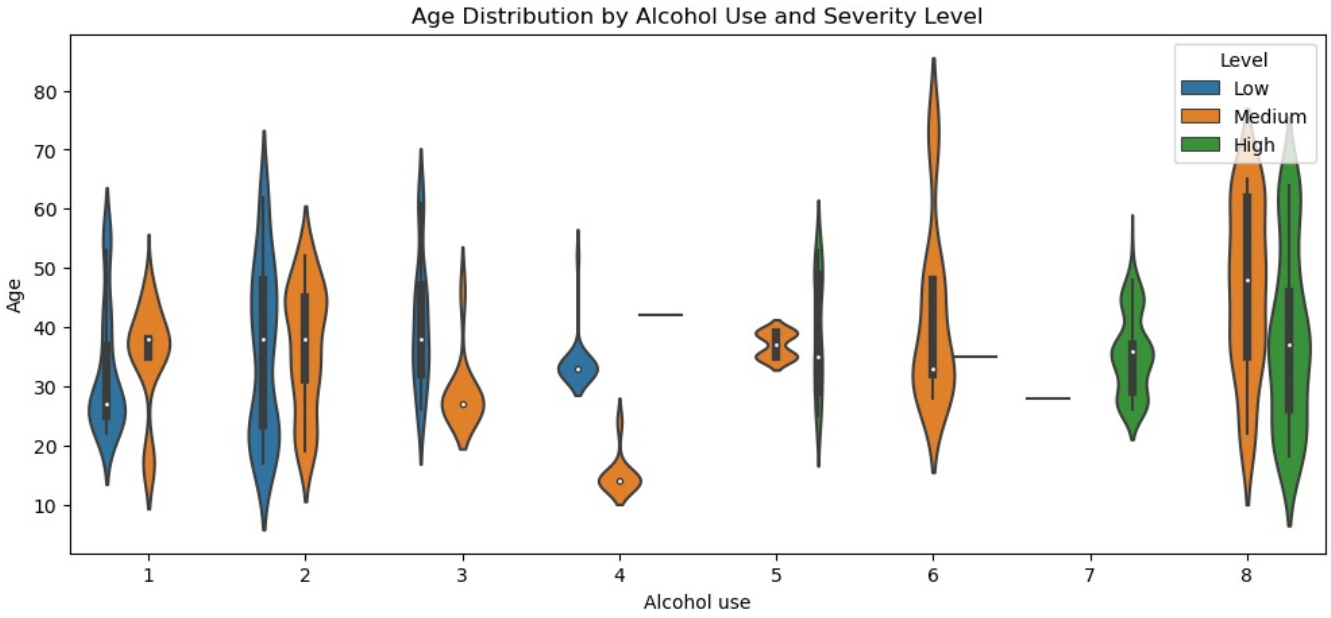
In [23]:
```
sns.lineplot(data = df,x = 'Air Pollution',y = 'Age',hue = 'Level',marker = '*',mec = 'red',color = 'green');
```



In [ ]:
```
Each line on the plot represents the relationship between "Air Pollution" and "Age"
for different severity levels of cancer.
I can observe how "Age" varies with different levels of "Air Pollution."
"High" severity level is trending upward, it may suggest that older patients are more
likely to be found in areas with higher air pollution levels.
The presence of different lines for different severity levels allows to compare how
"Age" is influenced by "Air Pollution" for patients with varying cancer severity.
```

In [24]:
```
plt.figure(figsize = (12,5))
sns.violinplot(data = df,x = 'Alcohol use',y = 'Age',hue = 'Level');
plt.title("Age Distribution by Alcohol Use and Severity Level")
```

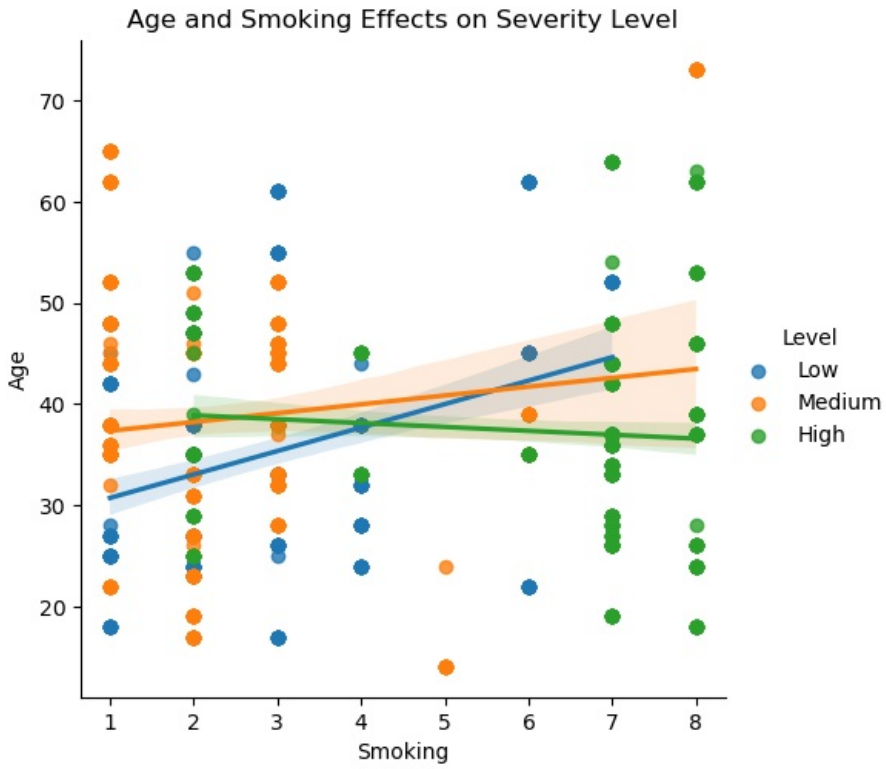
Out[24]:
```
Text(0.5, 1.0, 'Age Distribution by Alcohol Use and Severity Level')
```



In [ ]:
```
The the vioin plot shows The "Severity Level" parameter splits the violins for each level,
allowing to compare distributions across different severity levels.
The violin plot is wider at a specific level of alcohol use, it indicates that
there's a higher density of patients of age who have that particular level of alcohol use.
This analysis may provide insights into whether age is influenced by alcohol use for cancer patients
and whether this relationship differs for different severity levels.
```

In [25]:
```
sns.lmplot(data = df, x= 'Smoking',y = 'Age', hue = 'Level');
plt.title("Age and Smoking Effects on Severity Level")
```

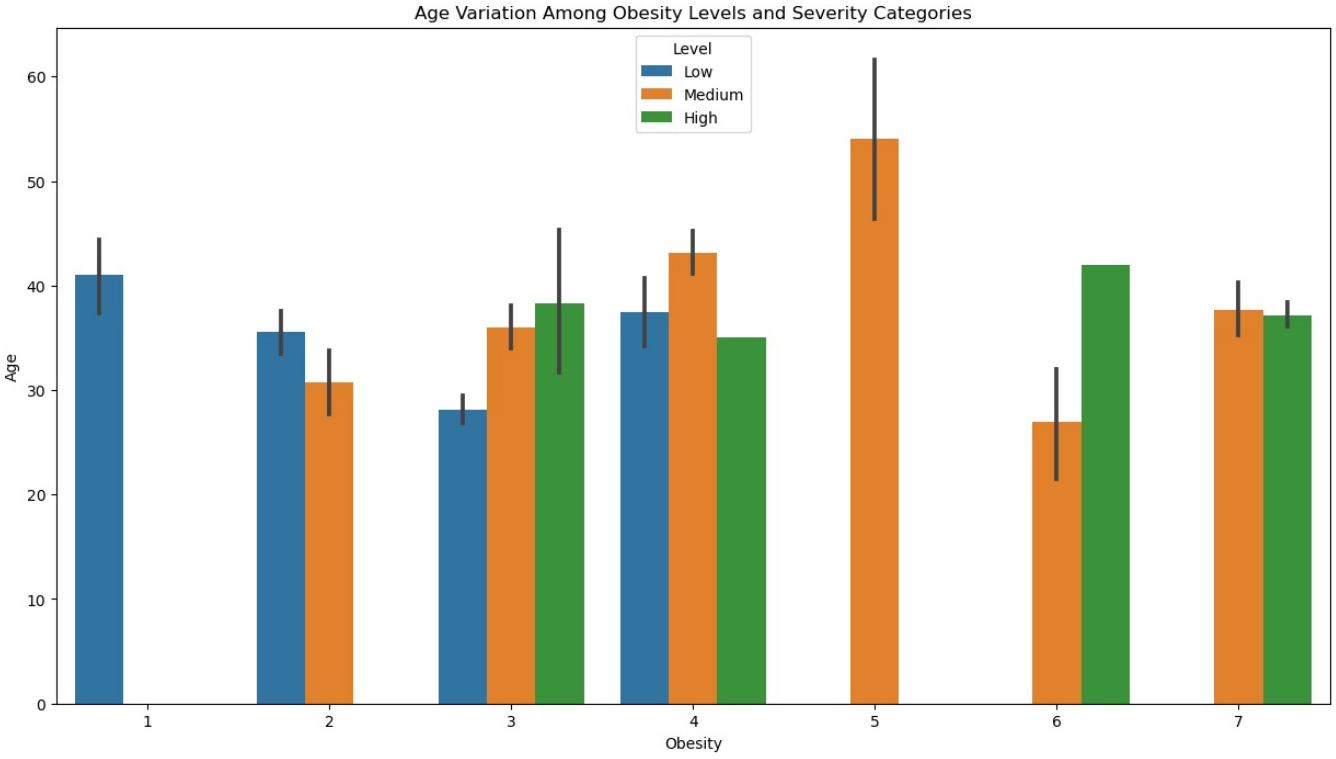
Out[25]:
```
Text(0.5, 1.0, 'Age and Smoking Effects on Severity Level')
```



In [ ]:
```
An lmplot shows the relationship between two continuous variables, "Age" and "Smoking,"
considering the impact of the "Severity Level."
The regression lines are positively sloped, it suggests that age tends to increase
with higher levels of smoking for a given severity level.
The color-coded points and regression lines help to observe how the relationship between
age, smoking, and severity level varies among patients.
```

In [39]:
```
plt.figure(figsize = (15,8))
sns.barplot(data = df, x= 'Obesity',y = 'Age', hue = 'Level');
plt.title("Age Variation Among Obesity Levels and Severity Categories")
```

Out[39]:
```
Text(0.5, 1.0, 'Age Variation Among Obesity Levels and Severity Categories')
```

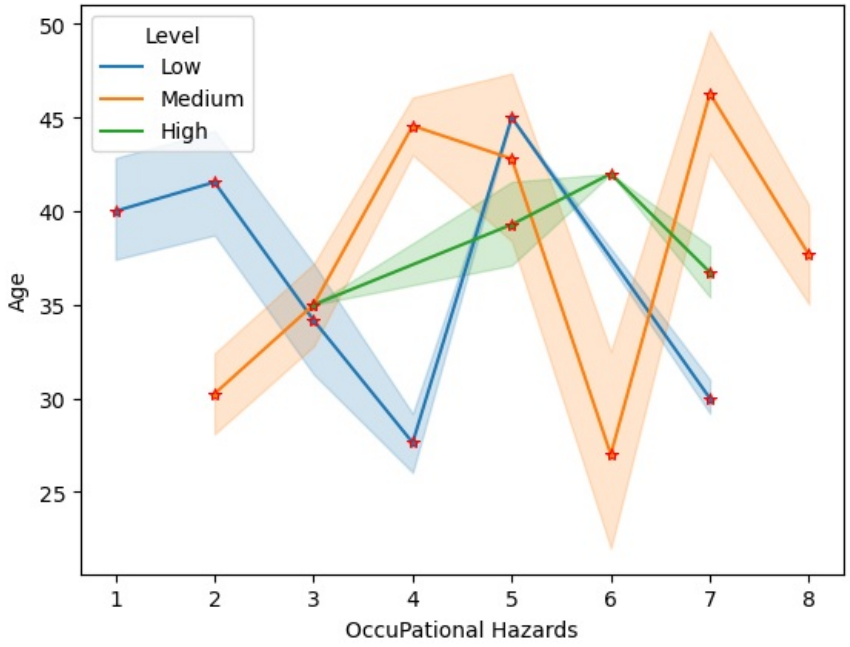


In [ ]:
```
The barplot show the age of patients varies across different levels of obesity
and how this variation differs for different cancer severity levels.
```

```
The bars represent the mean or central tendency of age for each obesity category
within each severity level.
the bars have different heights across obesity categories, it suggests that there
is age variations based on obesity.
```

In [26]:
```
sns.lineplot(data = df, x= 'OccuPational Hazards',y = 'Age', hue = 'Level',marker = '*',mec = 'red')
```

Out[26]:
```
<AxesSubplot:xlabel='OccuPational Hazards', ylabel='Age'>
```



In [ ]:
```
A lineplot is used to visualize the relationship between two continuous variables
(in this case, "Age" and "Occupational Hazards").
The hue parameter helps differentiate data points for each severity level,allowing you
to see how the relationship between age, occupational hazards, and severity level differs.
The lines represent the trend in age across different levels of occupational hazards.
```

In [27]:
```
df.corr()
```

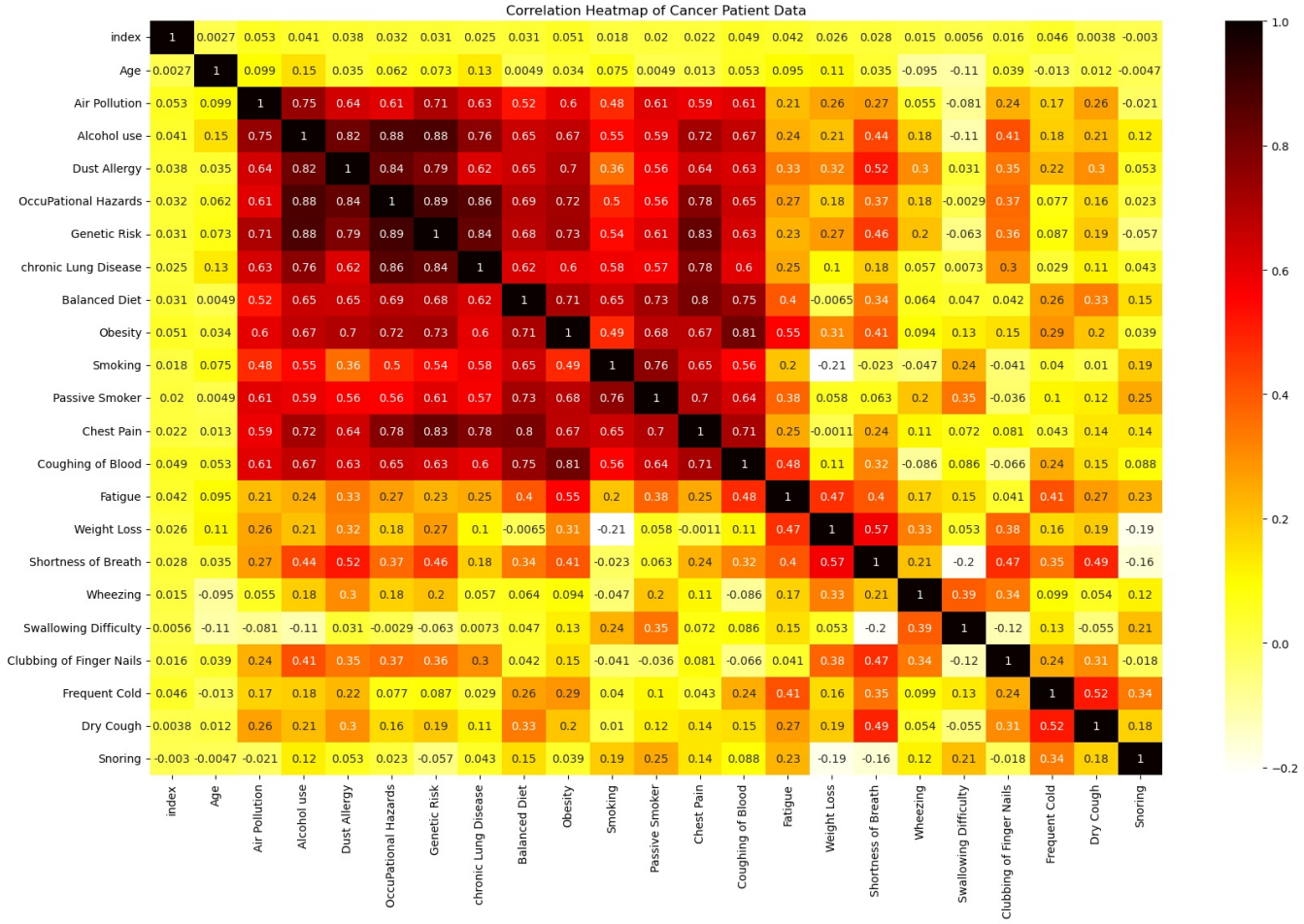
Out[27]:

	index	Age	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Obesity	...	Coughing of Blood	Fa
index	1.000000	0.002674	0.053307	0.041374	0.037960	0.032355	0.030725	0.025177	0.030743	0.050584	...	0.049401	0.04
Age	0.002674	1.000000	0.099494	0.151742	0.035202	0.062177	0.073151	0.128952	0.004863	0.034337	...	0.053006	0.09
Air Pollution	0.053307	0.099494	1.000000	0.747293	0.637503	0.608924	0.705276	0.626701	0.524873	0.601468	...	0.607829	0.21
Alcohol use	0.041374	0.151742	0.747293	1.000000	0.818644	0.878786	0.877210	0.763576	0.653352	0.669312	...	0.667612	0.23
Dust Allergy	0.037960	0.035202	0.637503	0.818644	1.000000	0.835860	0.787904	0.619556	0.647197	0.700676	...	0.625291	0.33
OccuPational Hazards	0.032355	0.062177	0.608924	0.878786	0.835860	1.000000	0.893049	0.858284	0.691509	0.722191	...	0.645947	0.26
Genetic Risk	0.030725	0.073151	0.705276	0.877210	0.787904	0.893049	1.000000	0.836231	0.679905	0.729826	...	0.632236	0.23
chronic Lung Disease	0.025177	0.128952	0.626701	0.763576	0.619556	0.858284	0.836231	1.000000	0.622632	0.601754	...	0.602987	0.24
Balanced Diet	0.030743	0.004863	0.524873	0.653352	0.647197	0.691509	0.679905	0.622632	1.000000	0.706922	...	0.745054	0.40
Obesity	0.050584	0.034337	0.601468	0.669312	0.700676	0.722191	0.729826	0.601754	0.706922	1.000000	...	0.814805	0.55
Smoking	0.018407	0.075333	0.481902	0.547035	0.358691	0.497693	0.543259	0.578585	0.645390	0.486795	...	0.555289	0.20
Passive Smoker	0.019517	0.004908	0.606764	0.592576	0.560002	0.555311	0.609071	0.572698	0.725123	0.681889	...	0.636223	0.37
Chest Pain	0.022210	0.012864	0.585734	0.717242	0.639983	0.775619	0.831751	0.782646	0.798207	0.673150	...	0.712158	0.25
Coughing of Blood	0.049401	0.053006	0.607829	0.667612	0.625291	0.645947	0.632236	0.602987	0.745054	0.814805	...	1.000000	0.48
Fatigue	0.042346	0.095059	0.211724	0.237245	0.332472	0.267844	0.230530	0.247697	0.400678	0.552788	...	0.481540	1.00
Weight Loss	0.026393	0.106946	0.258016	0.207851	0.321756	0.176226	0.271743	0.104080	-0.006544	0.313495	...	0.105857	0.46
Shortness of Breath	0.027950	0.035329	0.269558	0.435785	0.518682	0.366482	0.458200	0.182426	0.343623	0.406203	...	0.318777	0.39
Wheezing	0.015078	-0.095354	0.055368	0.180817	0.304850	0.178925	0.204973	0.057214	0.063930	0.094287	...	-0.085698	0.17
Swallowing Difficulty	0.005573	-0.105833	-0.080918	-0.114073	0.031141	-0.002853	-0.062948	0.007279	0.046807	0.127213	...	0.086289	0.14
Clubbing of Finger Nails	0.015706	0.039258	0.241065	0.414992	0.345714	0.366447	0.357815	0.298023	0.041967	0.149093	...	-0.066443	0.04
Frequent Cold	0.045687	-0.012706	0.174539	0.180778	0.219389	0.077166	0.087092	0.028759	0.263931	0.288368	...	0.244235	0.40
Dry Cough	0.003793	0.012128	0.261489	0.211277	0.300195	0.159887	0.194399	0.114161	0.331995	0.200618	...	0.147659	0.27
Snoring	-0.002957	-0.004700	-0.021343	0.122694	0.052844	0.022916	-0.056831	0.043375	0.152677	0.039422	...	0.087944	0.23

23 rows × 23 columns

In [28]:
```
plt.figure(figsize=(20, 12))
sns.heatmap(df.corr(),annot=True,cmap='hot_r',cbar=True)
plt.title("Correlation Heatmap of Cancer Patient Data")
```

Out[28]:
```
Text(0.5, 1.0, 'Correlation Heatmap of Cancer Patient Data')
```



In [ ]:
```
The heatmap help quickly to identify which variables have strong correlations with one another.
Positive correlations suggest that when one variable increases, the other tends to increase as well,
and vice versa for negative correlations.
This heat map shows cancer severity ("Level"), variables that have a strong positive
or negative correlation with "Level." Those variables important predictors.
```

In [ ]:
```
```