Homework N2

Part 1: Theoretical Questions

1. In a dataset with a non-normal distribution and potential extreme values, how are the whiskers in a boxplot determined, and what are the limitations of the standard IQR-based rule in such cases?

In a boxplot the whiskers are defined using the IQR, the left whisker is the smallest value bigger than Q_1 - 1.5 * IQR and the right one is the largest value smaller than Q_3 + 1.5 * IQR. If a dataset has extreme values they may misrepresent the plot and skew it heavily to on side and it may be the case that the 1.5 multiplier has to be adjusted in order to more accurately describe the data. In those cases plotting the boxplot of the log of the data may be more appropriate.

2. Given a dataset with heavy skewness and multiple peaks, how can a boxplot misrepresent outliers, and what alternative methods exist for identifying them more accurately?

If a boxplot is skewed with a lot of outliers it will not accurately convey the information about the data. It might be the case that by sorting the data we are losing a lot of information about it and if before it had a clear correlation after sorting and plotting it we will lose that information and the peaks will become outliers even though they are not. Instead if sorting the data is too destructive we can use scatter plot.

3. Explain the conceptual difference between median and mean in the context of non-symmetric distributions. Why does a boxplot prioritize the median, and in what cases could this choice obscure important data characteristics?

Both have their own issues. The median since it's the middle point of the data can misrepresent the other half of it. For example we can have very low values the first 51% and high the other 49% and the median will be a low value which does not help us much in this case. The average can have another issue where the outliers can heavily skew it to one direction thus misrepresenting the data. For example we can say that the average salary is 200,000 drams but it does not tell much about the salary of people since it can be the case where bottom 80% earn 80-100k and top 20% earn way more bringing the average up. The boxplot prioritizes the median since it is not prone to be skewed by outliers.

4. If a boxplot exhibits strong right skewness, what can you infer about the underlying probability distribution? How would this skewness affect statistical measures such as variance, skewness coefficient, and potential model assumptions?

Since the boxplot is right skewed we can infer that $|Q_2 - Q_1| > |Q_2 - Q_3|$ which tells us that the probability of the data is higher in the range (Q_1, Q_2), Hence we can tell that the underlying probability distribution is also skewed right. It will have more or less the same variance as seen in the boxplot (That being how spread out is the boxplot). The data will most probably not come from a normal distribution hence a lot of tests in frequentest statistics can not be applied to it.

5. Why are boxplots particularly useful for comparing multiple groups in high-dimensional data? What are the limitations of boxplots when dealing with overlapping distributions or categorical variables with small sample sizes?

Boxplots are more understandable when looked at since we can have multiple non overlapping boxplots in one graph to more easily compare them to each other. If data comes from overlapping distributions they will look similar, but from the graph we can not tell that they came from the same distribution.

6. What are the theoretical consequences of selecting an inappropriate number of bins in a histogram, particularly in datasets with varying density regions or multimodal distributions? How does bin width selection affect kernel density estimation (KDE)?

We can think of the bins size as the resolution of the graph. If we select too small bin size we would not get any good information from the plot as the varying data will not be visible. KDE helps us to understand if our bin selection is right as if the KDE is similar to the histogram we had a good bin choice.

7. Histograms and bar charts both use rectangular bars to display data. How does the interpretation of frequency differ in these two visualizations, and why is bin choice irrelevant in bar charts but crucial in histograms?

In bar charts we are not interested in ranges rather we are visualizing how many items there are in a given category. Hence bin choice is irrelevant in the bar charts. In histogram we would like to visualize how many elements there are in a range of numbers. That range is the bin size hence it's choice affects the shape of the plot and what information it coveys.

8. Under what conditions might a histogram distort the perception of a dataset's distribution? Provide an example where binning choices lead to misleading conclusions, and explain how alternative visualizations (e.g., KDE or violin plots) could address these distortions.

It might distort it if we chose the wrong bin size. For example take
1,1,1,2,2,2,3,4,5,6,7,7,8,9,9,9,9,9,9.
For 2 bins one [1,8] another [9,10] we will have, two bars first one significantly longer than the other.
For 5-6 equal length bins we will have normalish* looking graph which is far more accurate to the data.
Hence by choosing the wrong bins we can have completely different results.

9. How does a density plot differ from a histogram in terms of its mathematical foundation and interpretability? What challenges arise when choosing a kernel function and bandwidth for density estimation, particularly in sparse datasets?

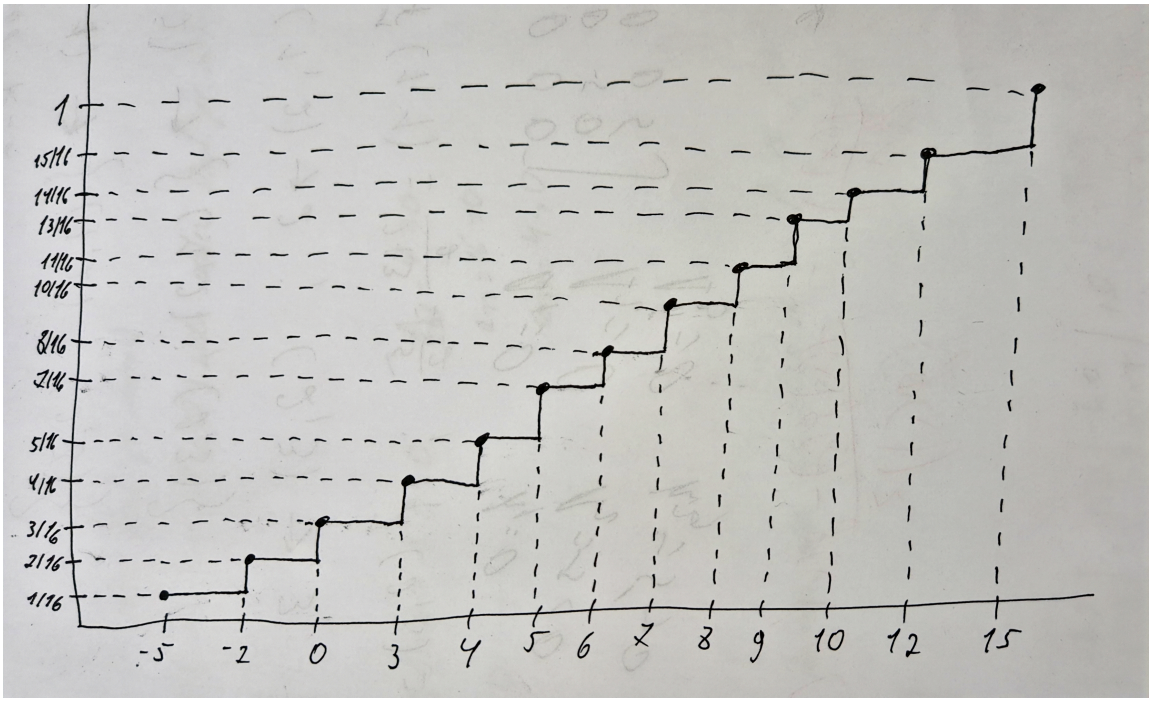
Density plot represents the data by counting frequencies of the data and dividing it by the bin length thus getting approximate distribution the data. KDE has the same idea behind it but uses different method to achieve it. It uses a kernel function and bandwidth in order to approximate the density. When data is sparse the bandwidth and the function are very important since they play a key role as to how the plot will come out. The bandwidth controls the smoothness of the curve while the kernel decides the algorithm of approximation.

10. Explain why the area under a density plot is always equal to 1. How does this property relate to probability theory, and what implications does it have for comparing distributions with different sample sizes?

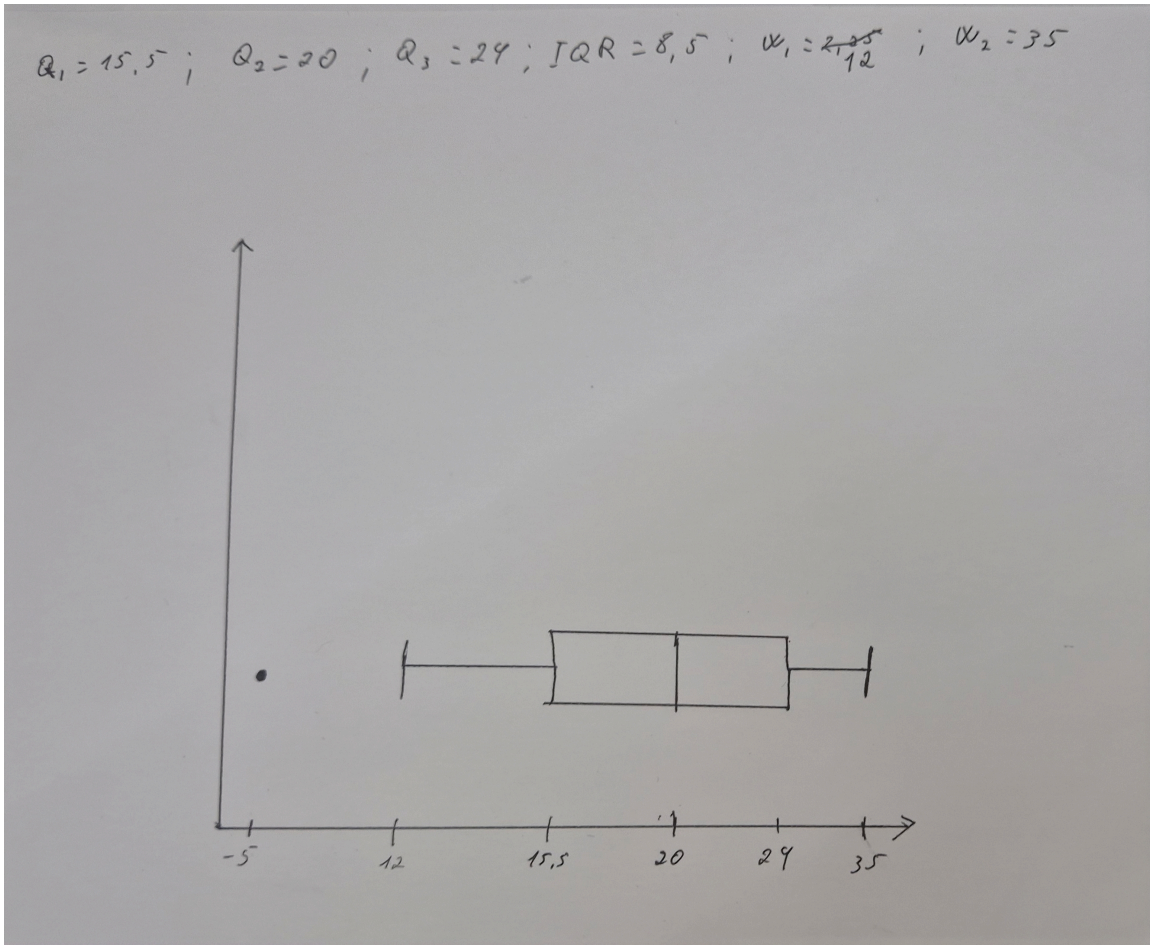
It is a density function of the data. Every density function is defined in a way that the area under their curve is 1. This is the case since the probability of every event summed is always 1. The sample size does not matter in this fact since the sum of the probabilities of every event is always 1. If the data came from the same distribution sample does not matter much after a certain amount. No matter the case the KDE-s will approximately be the same.

Part 2: Hand-Drawn Graphs Create graphs by hand using the provided datasets.

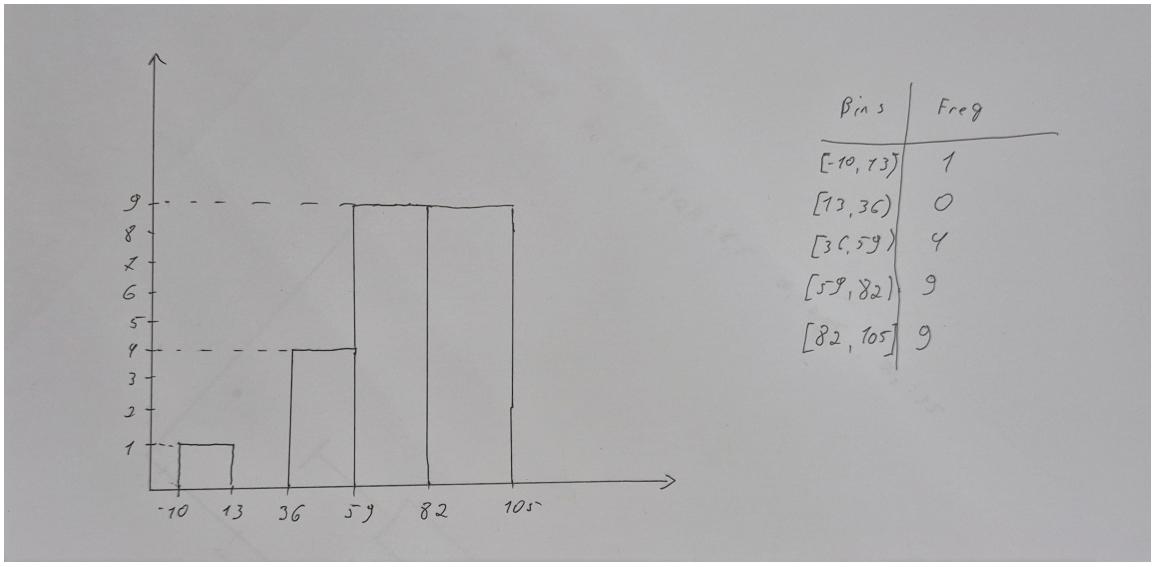
1. Given the numbers: -5, -2, 0, 3, 4, 5, 5, 6, 7, 7, 8, 9, 9, 10, 12, 15, draw an ECDF plot.



2. Given the dataset: -5, 12, 14, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 24, 25, 29, 30, 35, create a boxplot. Indicate the median, quartiles, and any potential outliers.



3. Given the test scores: -10, 45, 50, 55, 55, 60, 62, 65, 68, 70, 73, 74, 80, 80, 82, 85, 88, 90, 91, 92, 94, 97, 100, 105, create a histogram using 5 bins and label the axes.



Part 3: Use the datasets provided to create graphs

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df_lung = pd.read_csv("lung_cancer_prediction_dataset.csv")
df_lung.head()
```

Out[1]:

	ID	Country	Population_Size	Age	Gender	Smoker	Years_of_Smoking	Cigarettes_per_
0	0	China	1400	80	Male	Yes	30	
1	1	Iran	84	53	Male	No	0	
2	2	Mexico	128	47	Male	Yes	12	
3	3	Indonesia	273	39	Female	No	0	
4	4	South Africa	59	44	Female	No	0	

5 rows × 24 columns

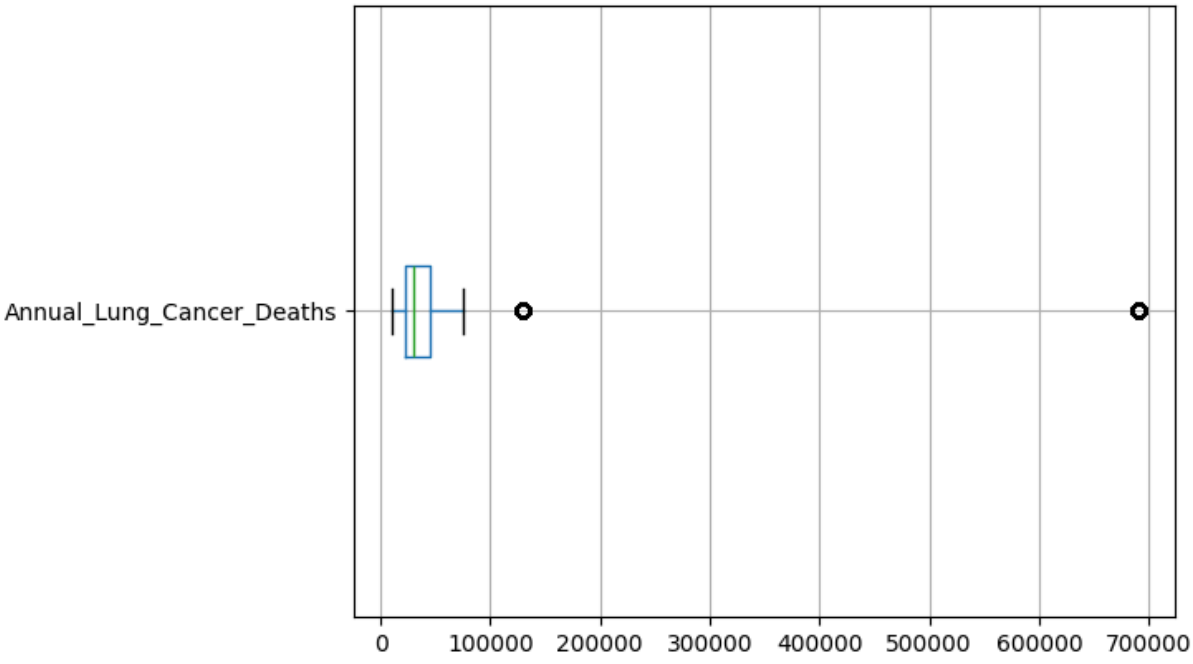
In [2]:
```
df_air = pd.read_csv("global_air_pollution_dataset.csv")
df_air.head()
```

Out[2]:

	Country	City	AQI_Value	AQI_Category	CO_AQI_Value	CO_AQI_Category	Ozor
0	Russian Federation	Praskoveya	51	Moderate	1	Good	
1	Brazil	Presidente Dutra	41	Good	1	Good	
2	Italy	Priolo Gargallo	66	Moderate	1	Good	
3	Poland	Przasnysz	34	Good	1	Good	
4	France	Punaauia	22	Good	0	Good	

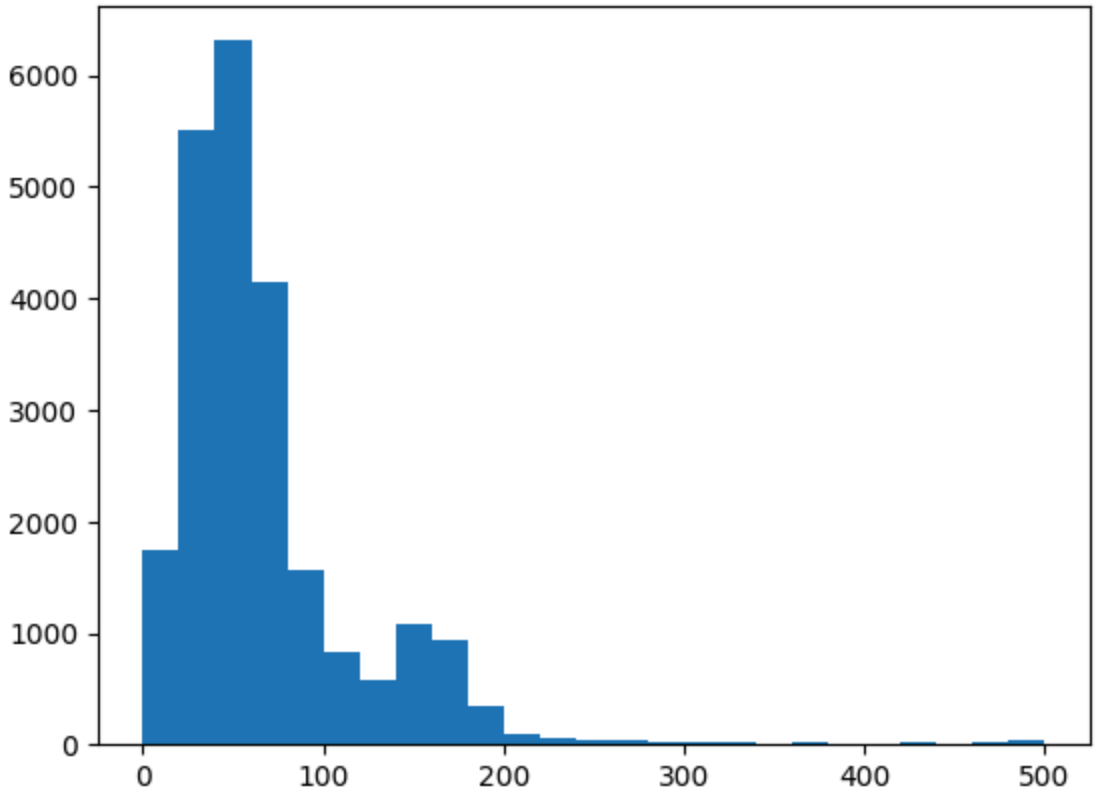
1. Create a Boxplot of Lung Cancer Deaths Distribution.

In [3]:
```
df_lung.boxplot("Annual_Lung_Cancer_Deaths", vert=False)
plt.show();
```



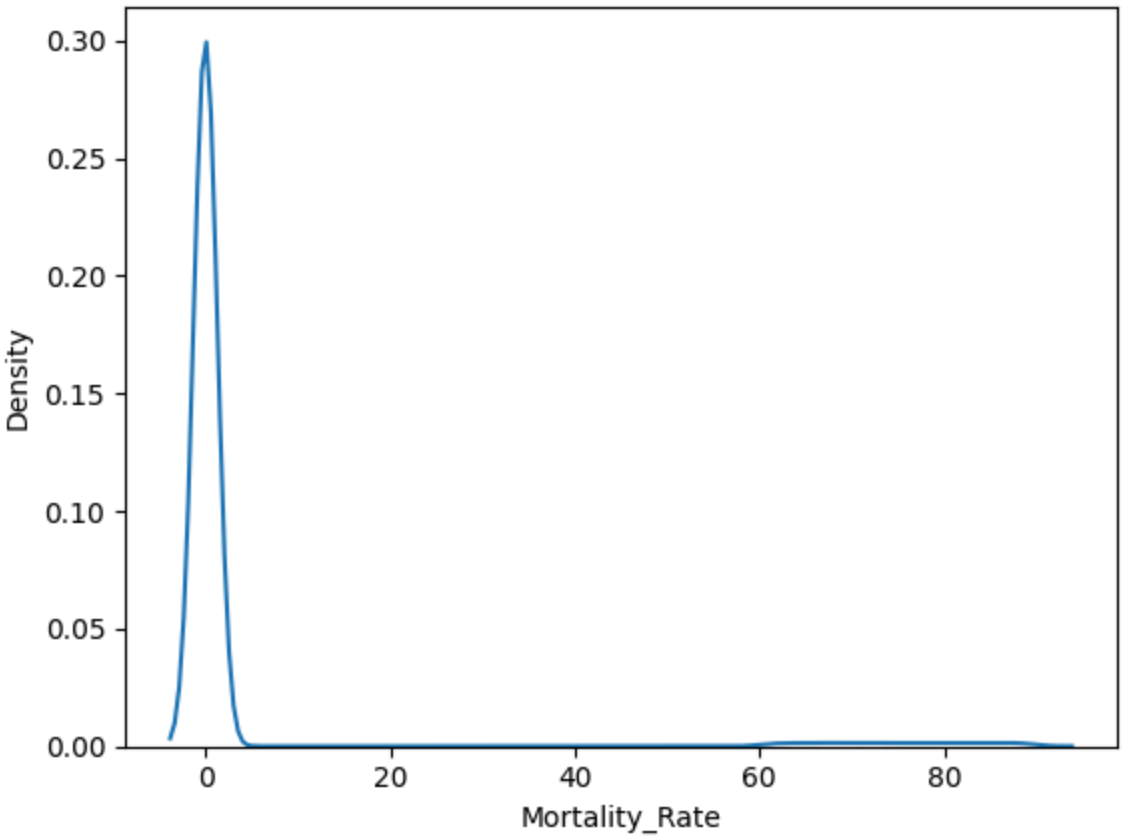
2. Create a Histogram of PM2.5 AQI Values.

In [4]:
```
plt.hist(x = df_air["PM2.5_AQI_Value"], bins=25)
plt.show();
```



3. Create a Density Plot of the Lung Cancer Mortality Rate.

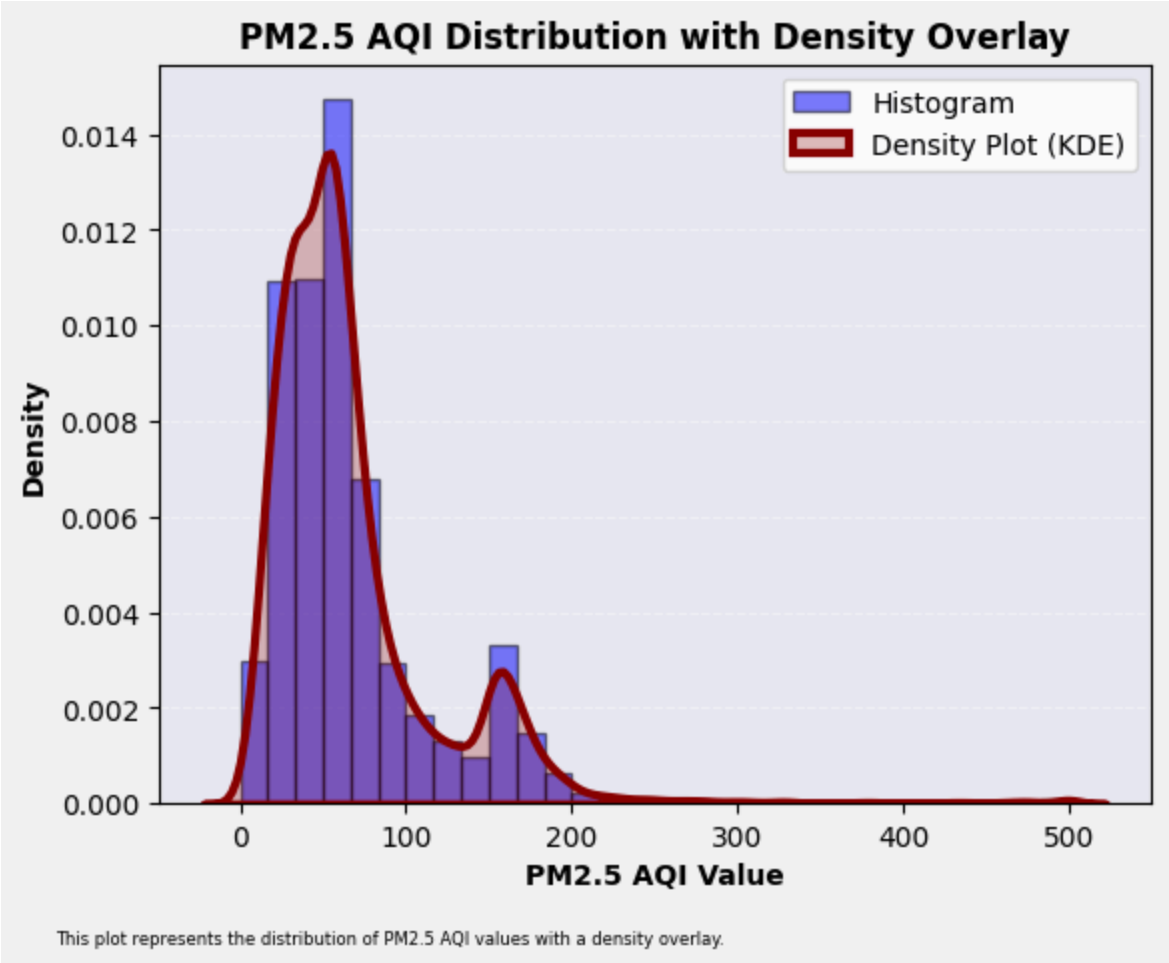
In [5]:
```
sns.kdeplot(data=df_lung, x = "Mortality_Rate")
plt.show();
```



Part 4: Recreate the following graphs

1. Use the matplotlib library for this graph.

In [6]:
```
plt.rc("axes", axisbelow=True) #Moves gridines to background
plt.figure(facecolor="#f4f4f4") #Background outer color
plt.gca().set_facecolor("#eaeaf2") #Background color of the grid
plt.grid(axis = "y", linestyle="--", alpha=0.5, color="white") #Adds the dashed y g
plt.hist(data=df_air, x = "PM2.5_AQI_Value", bins=30, edgecolor="black", color="blu
sns.kdeplot(data=df_air, x="PM2.5_AQI_Value", color="darkred", fill=True, linewidth
plt.xlabel("PM2.5 AQI Value", weight = "bold")
plt.ylabel("Density", weight="bold")
plt.title("PM2.5 AQI Distribution with Density Overlay", weight='bold')
plt.legend(["Histogram", "Density Plot (KDE)"])
plt.text(90, -0.003, "This plot represents the distribution of PM2.5 AQI values wit
            wrap=False, verticalalignment="bottom", horizontalalignment="center", f
plt.show();
```


PM2.5 AQI Distribution with Density Overlay
Histogram
Density Plot (KDE)
Density
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
100
200
300
400
500
PM2.5 AQI Value
This plot represents the distribution of PM2.5 AQI values with a density overlay.