DS 116 / CS 343 Data Visualization
Spring 2025

Homework N2

Instructions:

Part 1: Answer the questions in written form. Provide 2-3 sentences for each question, bring examples when needed. Collect your answers in a PDF file. (20 points)
Part 2: Draw the graphs by hand. Provide clear pictures of each graph. You can attach these pictures in the same PDF file as Part 1. (30 points)
Part 3: Complete each task in both Python and R and provide your solutions in .ipynb and .rmd formats. (20 points)
Part 4: Recreate each plot as indicated in the problem. Store the *matplotlib* graph in your .ipynb file and the *ggplot* graphs in the same .rmd file as Part 3. (30 points)

Submit your homework on GitHub and provide the link in Moodle.

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?
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?
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?
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?
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?
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)?
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?

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.
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?
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?

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

Overview of the datasets.

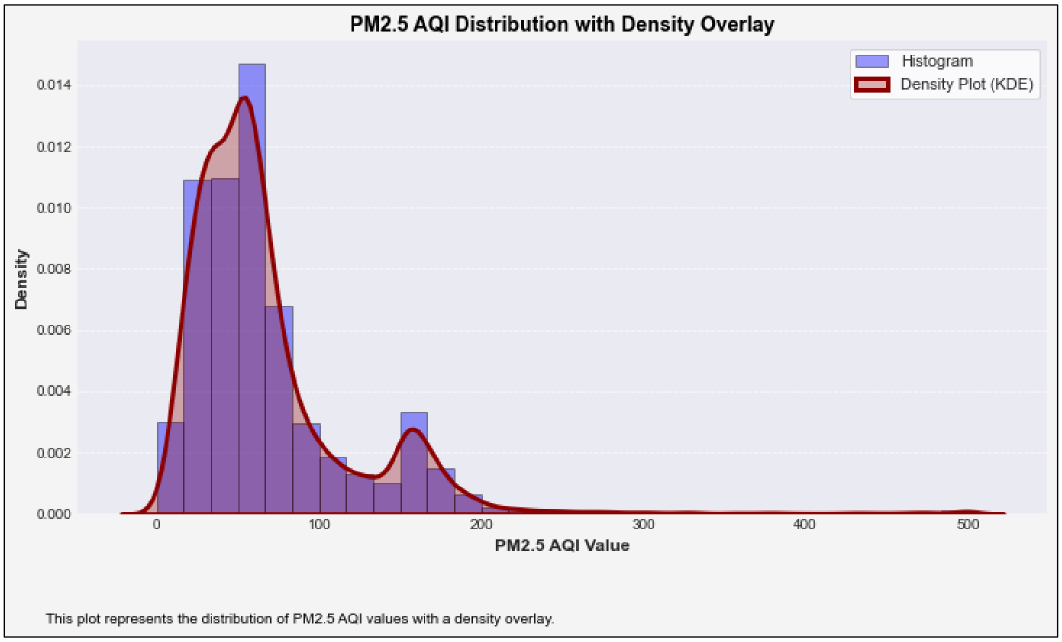
The "lung_cancer_prediction_dataset.csv" provides valuable insights into lung cancer cases, risk factors, smoking trends, and healthcare access across 25 of the world's most populated countries. It includes 220,632 individuals with details on their age, gender, smoking history, cancer diagnosis, environmental exposure, and survival rates. The dataset is useful for medical research, predictive modeling, and policymaking to understand lung cancer patterns globally.
source

The "global_air_pollution_dataset.csv" provides information about the contamination of the indoor or outdoor environment by any chemical, physical or biological agent that modifies the natural characteristics of the atmosphere. It includes 23,463 rows with data presented in the following columns: Country, City, AQI Value, AQI Category, CO AQI Category, Ozone AQI Value, Ozone AQI Category, NO2 AQI Value, NO2 AQI Category, PM2.5 AQI Value, PM2.5 AQI Category.
source

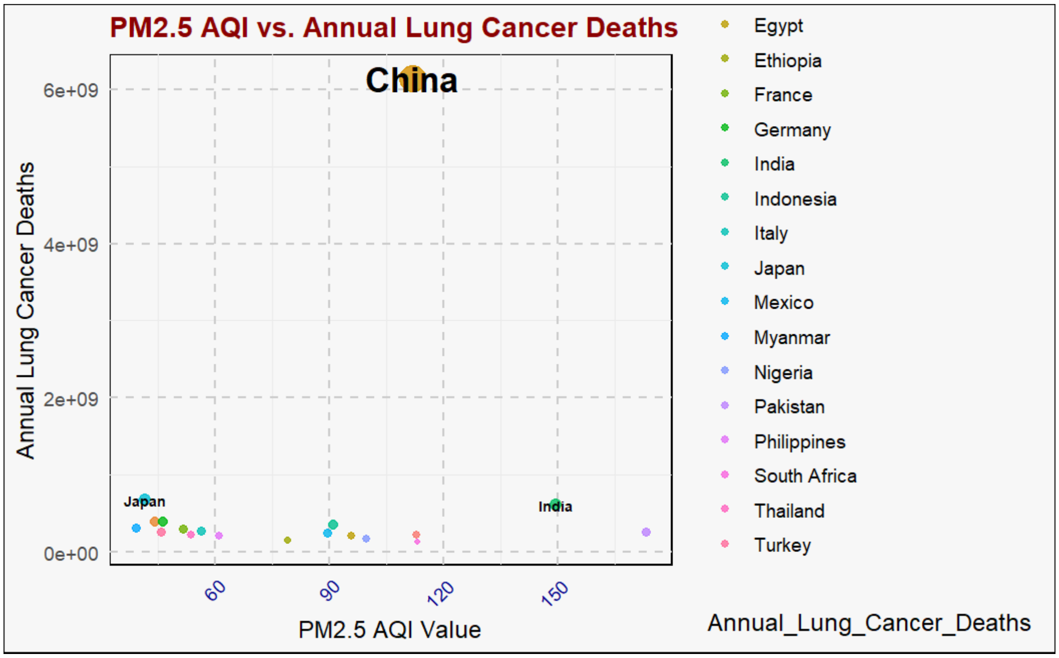
1. Create a Boxplot of Lung Cancer Deaths Distribution. (Python & R)
2. Create a Histogram of PM2.5 AQI Values. (Python & R)
3. Create a Density Plot of the Lung Cancer Mortality Rate. (Python & R)
4. Create a Scatter Plot by generating 100 random values from both the normal and logistic distributions. The points should be brown and use theme_solarized with argument light set to false. (R, *not related to the datasets provided*)

Part 4: Recreate the following graphs

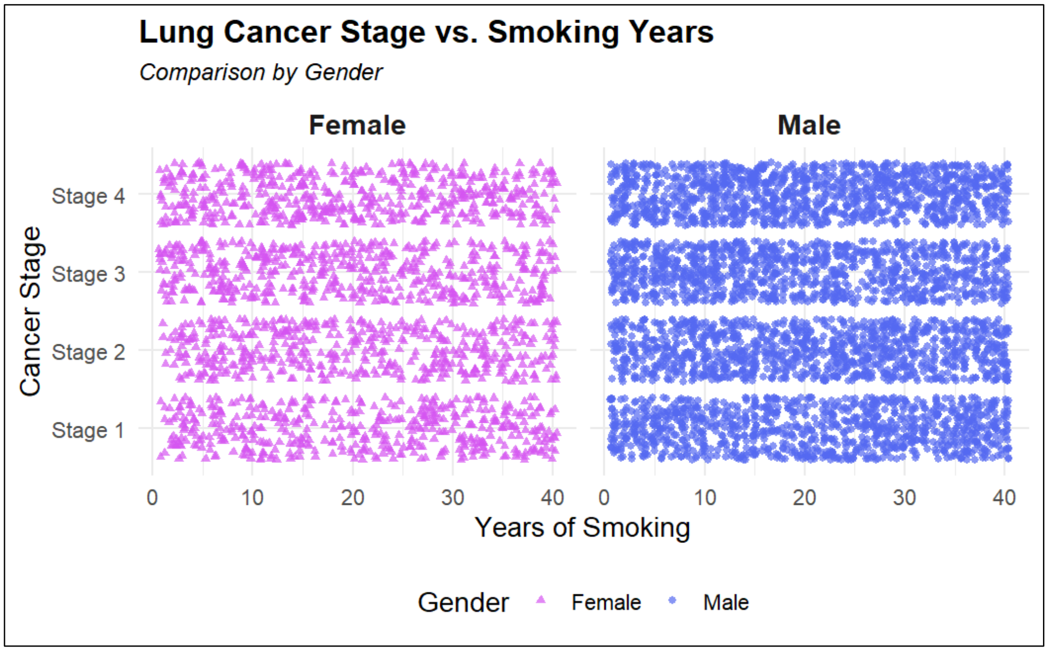
1. Use the matplotlib library for this graph.



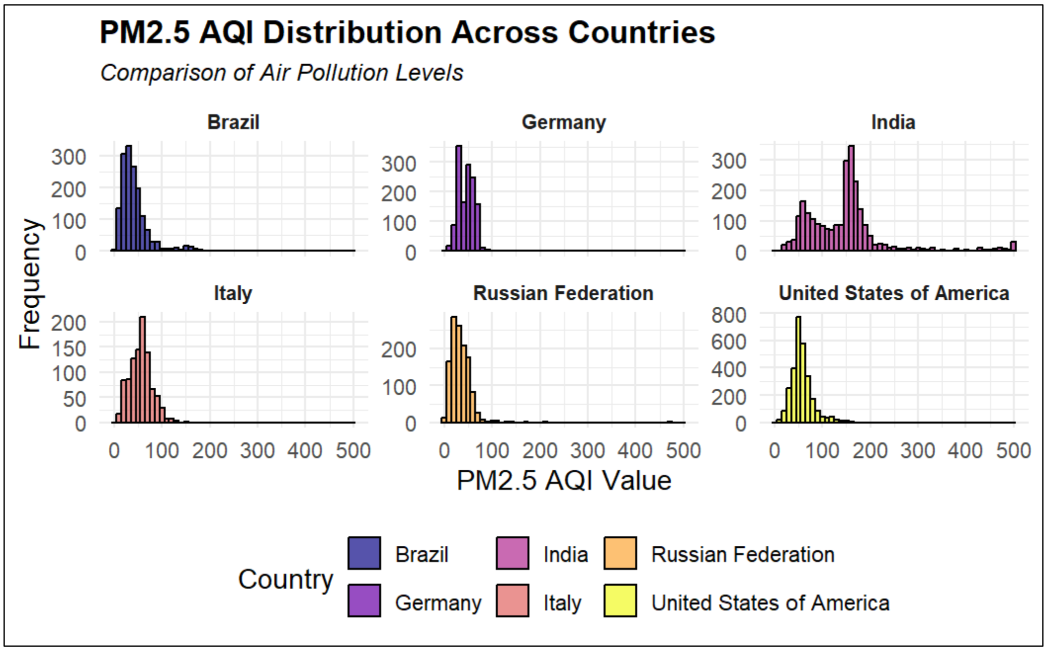
2. Use the gpplot2 package for this graph. *(Hint: Aggregate the data then merge the two datasets. Use only the necessary columns.)*



3. Use the ggplot2 package for this graph. (*Hint: use geom_jitter since y axis contains categorical data, also use the following colors: #5469f1 , #d554f1*)



4. Use the ggplot2 package for this graph. (*Hint: use scale_fill_viridis_d(option = "plasma" to get the same colors*)


PM2.5 AQI Distribution Across Countries
Comparison of Air Pollution Levels
Brazil
Germany
India
Italy
Russian Federation
United States of America
Frequency
PM2.5 AQI Value
Country
Brazil
Germany
India
Italy
Russian Federation
United States of America