Duygu Abbasoğlu - 22102978

CTIS365 - Lab Project

In this project, I used the built-in airquality dataset. This dataset contains daily air quality measurements collected in New York between May and September 1973. The variables included in my study are Ozone (ppb), Solar.R (solar radiation), Wind (mph), Temp (°F), Month, and Day.
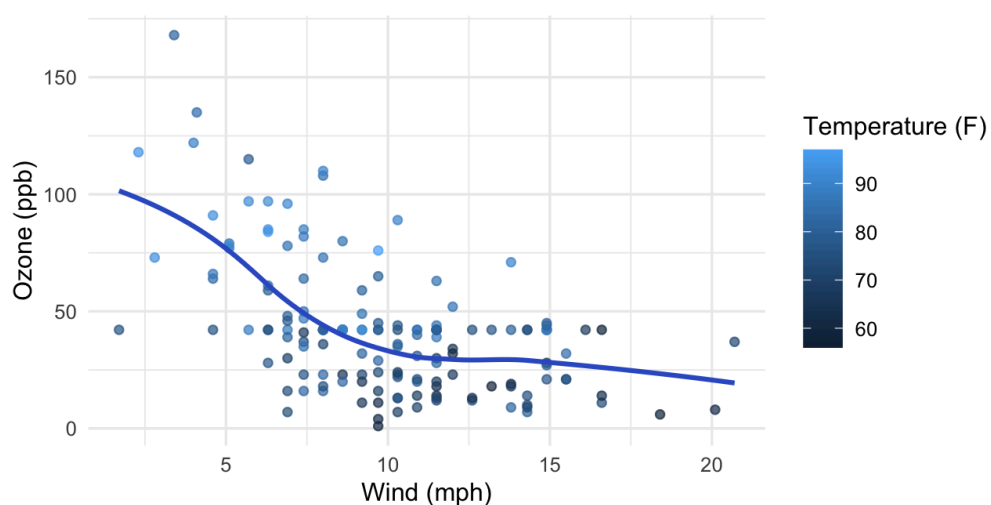
1. Dataset Preparation

I started by looking for missing values in the dataset. I realized that Solar.R and Ozone had NA values. I used the mutate() and across() functions from dplyr to clean the data by replacing all numerical NA values with their column means. Following this stage, the dataset contained no missing values.

Additionally, I used group_by() and summarise() utilities to create a brief summary table that determines the average ozone level for each month. For easier reading, the month numbers (5–9) were changed to labels (May–September).
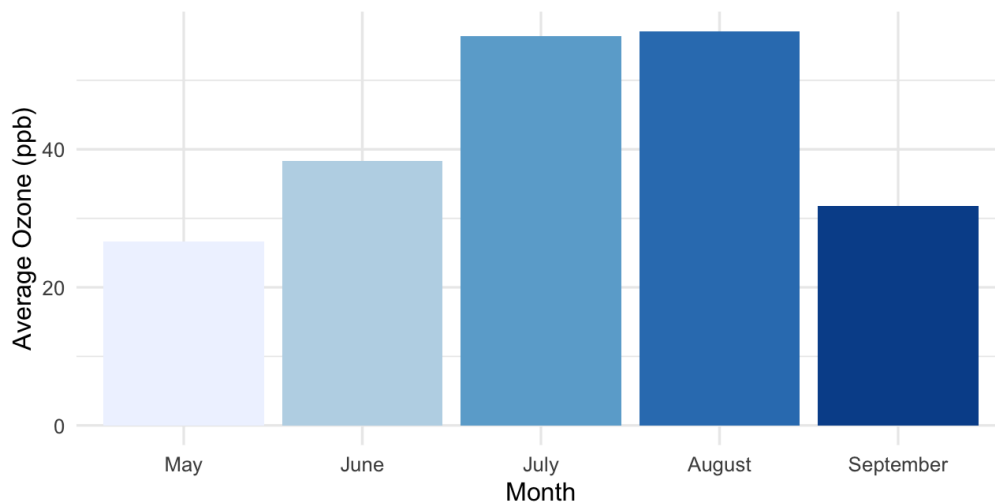
2. Graphs

2.1 - Wind Speed vs Ozone Levels

For the first plot, I visualized the relationship between Wind and Ozone, and added temperature as a color scale.

Results:

- Ozone levels tend to be higher on warmer days.

- There is a negative relationship between wind speed and ozone—higher wind generally corresponds to lower ozone levels, likely because wind helps disperse pollutants.

2.2 – Average Ozone Levels by Month

The second plot shows the average ozone level for each month using a bar chart.

Results:

- Ozone levels rise from May to July, with the highest values in July and August.

- In September, average ozone values decreased again.

- This pattern suggests that ozone formation is strongly influenced by summer heat and sunlight.

3. Conclusion

Overall, the data indicates that higher ozone levels are linked to higher temperatures. Wind also lowers the concentration of ozone. The dataset's highest ozone levels are found in the midsummer months of July and August. These findings show how simple data preparation and visualization methods can highlight significant trends and reflect typical seasonal air quality behavior.