EDA - New York air quality analysis

Gómez-Alonso, I.

2025-06-13

Contents

1. Introduction

This project presents an Exploratory Data Analysis (EDA) of the airquality dataset, a dataset available in RStudio https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/airquality.html.

The general purpose of this study is to analyze the temporal distribution of the main air quality variables (ozone, solar radiation, wind and temperature) during the study period, identifying seasonal or monthly trends in ozone levels and other variables, as well as to evaluate the relationship and possible significant correlations between ozone concentration and environmental variables such as temperature, wind speed and solar radiation.

1.1. Dataset description The airquality dataset contains daily air quality measurements in New York City taken between May and September 1973. It includes the following key variables:

- Ozone: Mean ozone in parts per billion (ppb) from 1300 to 1500 hours at Roosevelt Island
- Solar.R: Solar radiation in Langleys in the frequency band 4000–7700 Angstroms from 0800 to 1200 hours at Central Park
- Wind: Average wind speed in miles per hour (mph) at 0700 and 1000 hours at LaGuardia Airport
- Temp: Maximum daily temperature in degrees Fahrenheit (°F) at LaGuardia Airport
- Month: The month of the observation (5 through 9, representing May through September)
- Day: The day of the month of the observation (1 -31)

The data were obtained from the New York State Department of Conservation (ozone data) and the National Weather Service (meteorological data).

2. Data preparation and exploration

Let's start by loading the dataset in our Rstudio environment and perform an initial inspection to familiarize ourselves with its dimensions, the type of variables it contains and the first rows of data.

2.1. Data loading and initial inspection First, we start by loading the required libraries.

```
library(dplyr)
```

```
##
## Adjuntando el paquete: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
```

Next, we load the airquality dataset, this is easily accomplished, since airquality is a built-in dataset in R, which avoids the need to import external files.

```
data_air <- airquality # assign the dataset in a variable
```

Once loaded, we will perform an initial inspection to get an overview of the information. We will start by using basic functions such as *dim*() to get the dimensions of the database.

```
dim(data_air) #dataset size
```

```
## [1] 153   6
```

The dataset contains 153 rows and 6 columns.

The *head*() function allows us to view the first few rows of the dataset, which will give us a quick look at the structure of the data and the type of values contained in each column.

```
head(airquality) #first rows in data
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

Next, The *summary*() function provides a quick overview of the distribution of the variables, helps to identify possible outliers (by looking at ranges and quartiles) and detects the presence of missing values.

```
summary(data_air) #summary of variables in data
```

```
##      Ozone           Solar.R           Wind             Temp      
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
##  NA’s   :37       NA’s   :7                                       
##      Month            Day      
##  Min.   :5.000   Min.   : 1.0  
##  1st Qu.:6.000   1st Qu.: 8.0  
##  Median :7.000   Median :16.0  
##  Mean   :6.993   Mean   :15.8  
##  3rd Qu.:8.000   3rd Qu.:23.0  
##  Max.   :9.000   Max.   :31.0  
## 
```

Key findings by variable:

- **Ozone (Ozone Concentration)**:
  - Ozone concentration shows a considerable range, from a low of 1 ppb to a high of 168 ppb.
  - The mean (42.13 ppb) is notably higher than the median (31.50 ppb), suggesting a positive asymmetric distribution (skewed to the right), indicating the presence of some exceptionally high ozone values.
  - A critical aspect is the presence of 37 missing values (NA’s), which represent a significant portion of the observations for this variable and will require careful consideration during the cleaning phase.
- **Solar.R (Solar Radiation)**:
  - Solar radiation fluctuates from 7 Langleys to 334 Langleys.
  - The mean (185.9 Langleys) and median (205.0 Langleys) are relatively close, suggesting a more symmetrical distribution than ozone, although the median is slightly higher, indicating a slight asymmetry towards lower values.
  - Seven missing values (NA’s) were identified in this variable, a smaller number than in Ozone, but still in need of attention.
- **Wind (Wind Speed)**:
  - Wind speed ranges from 1.7 mph to 20.7 mph.
  - The mean (9.958 mph) and median (9.700 mph) are very similar, indicating a fairly symmetrical distribution for wind speed, with most values concentrated around the mean.
  - No missing values were observed for this variable, which simplifies its handling.
- **Temp (Temperature)**:
  - The reported temperature ranges from 56°F to 97°F.
  - The mean (77.88°F) and median (79.00°F) are very close, suggesting a relatively symmetrical temperature distribution.
  - The quartiles indicate that most of the temperatures are between 72°F and 85°F.
  - There are no missing values for the temperature variable.

- **Month and Day**:

The variables **Month** and **Day** are of discrete type and act as temporal identifiers. Month spans from month 5 (May) to month 9 (September), confirming the study period. Day ranges from day 1 to day 31, as would be expected for a daily record. They have no missing values, and their descriptive statistics reflect their nature as time indices.

Finally, *str*() will provide us with the internal structure of the dataset, showing the column names, their data types (**numeric**, **integer**, etc.) and the number of observations.

```r
str(data_air) #structure of dataset
```

```
## 'data.frame':    153 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

An *int* (integer) is a numeric data type used to store whole numbers without any decimal component. A *num* (numeric),it's used to store numbers that can have decimal components.

2.2. Data cleaning Due to the nature of the dataset, it is possible that there are no duplicate rows, however, it is important to corroborate

```r
num_duplicated <- sum(duplicated(data_air)) #verify duplicated rows
print(num_duplicated)
```

```
## [1] 0
```

This dataset has no rows with duplicate data.

```r
colSums(is.na(data_air)) #Counting NA's by column
```

2.2.1 Missing value management strategy (NA):

```
##   Ozone Solar.R    Wind    Temp   Month     Day
##      37       7       0       0       0       0
```

Given the nature of this dataset (daily air quality records) and the number of NAs (37 in Ozone out of 153 observations, and 7 in Solar.R), there are several possible strategies:

a) Imputation: Replace the NAs with an estimated value (mean, median, mode, or using more complex models). This option is more sophisticated and seeks to preserve the size of the dataset, but introduces a "falsification" of data that could bias the results if the imputation is not adequate.

b) Removal of entire rows (na.omit()): This is the simplest and most straightforward option. It deletes any row containing at least one NA in any of its columns. While easy to implement, it may result in the loss of a significant amount of data if there are many NAs distributed in different rows.

For this EDA and given that the main objective is to explore clear patterns and relationships, we will opt for the *elimination of complete rows.* This decision is justified because:

- The number of NAs in Ozone is significant, and imputation could distort key distributions or relationships.
- For Solar.R, even though they are few, removing them simplifies the dataset without losing critical information.
- We want to ensure that any correlation or pattern analysis is based on complete and reliable observations.

In order to implement the elimination of rows with NAs, the *na.omit()* function is used.

```
data_air_clean <- na.omit(data_air) # Remove all rows containing at least one NA

dim(data_air_clean) # Check dimensions of new data frame
```

```
## [1] 111   6
```

A new dataframe called *data_air_clean* was created. Out of 153 original rows, we are left with 111 rows. That's a loss of 42 rows (37 from Ozone + 5 additional rows where Solar.R had NA and Ozone did not).

```
colSums(is.na(data_air_clean)) # Check again for presence of NA's to confirm cleanup
```

```
##   Ozone Solar.R    Wind    Temp   Month     Day 
##       0       0       0       0       0       0
```

```
summary(data_air_clean) # Verify again the summary
```

```
##      Ozone          Solar.R           Wind            Temp      
##  Min.   :  1.0   Min.   :  7.0   Min.   : 2.30   Min.   :57.00  
##  1st Qu.: 18.0   1st Qu.:113.5   1st Qu.: 7.40   1st Qu.:71.00  
##  Median : 31.0   Median :207.0   Median : 9.70   Median :79.00  
##  Mean   : 42.1   Mean   :184.8   Mean   : 9.94   Mean   :77.79  
##  3rd Qu.: 62.0   3rd Qu.:255.5   3rd Qu.:11.50   3rd Qu.:84.50  
##  Max.   :168.0   Max.   :334.0   Max.   :20.70   Max.   :97.00  
##      Month            Day       
##  Min.   :5.000   Min.   : 1.00  
##  1st Qu.:6.000   1st Qu.: 9.00  
##  Median :7.000   Median :16.00  
##  Mean   :7.216   Mean   :15.95  
##  3rd Qu.:9.000   3rd Qu.:22.50  
##  Max.   :9.000   Max.   :31.00
```

```


numeric_cols_clean <- c("Ozone", "Solar.R", "Wind", "Temp") # Define the numerical columns of interest
```

```


descriptive_stats_clean <- data_air_clean %>% # Calculate the descriptive statistics for each column of
  select(all_of(numeric_cols_clean)) %>% # Select only the numerical columns that interest us
  summarise( # Summarize each column
    # Ozone
    Ozone_Mean = mean(Ozone, na.rm = TRUE),
    Ozone_Median = median(Ozone, na.rm = TRUE),
    Ozone_SD = sd(Ozone, na.rm = TRUE),
    Ozone_N = n(),

    # Solar radiation
    SolarR_Mean = mean(Solar.R, na.rm = TRUE),
    SolarR_Median = median(Solar.R, na.rm = TRUE),
    SolarR_SD = sd(Solar.R, na.rm = TRUE),
    SolarR_N = n(),

    # Wind
    Wind_Mean = mean(Wind, na.rm = TRUE),
    Wind_Median = median(Wind, na.rm = TRUE),
    Wind_SD = sd(Wind, na.rm = TRUE),
    Wind_N = n(),

    # Temperature
    Temp_Mean = mean(Temp, na.rm = TRUE),
    Temp_Median = median(Temp, na.rm = TRUE),
    Temp_SD = sd(Temp, na.rm = TRUE),
    Temp_N = n()
  ) %>%
  # Use pivot_longer to transform the table from width to length
  pivot_longer(
    cols = everything(), # Select all columns
    names_to = c("Variable", ".value"), # Split the names in ‘Variable’ and the type of statistic
    names_pattern = "(.+)_(Mean|Median|SD|N)" # regex pattern to extract the variable and the statistic
  )

print(descriptive_stats_clean) # Show the resulting table
```

2.3. Descriptive statistics

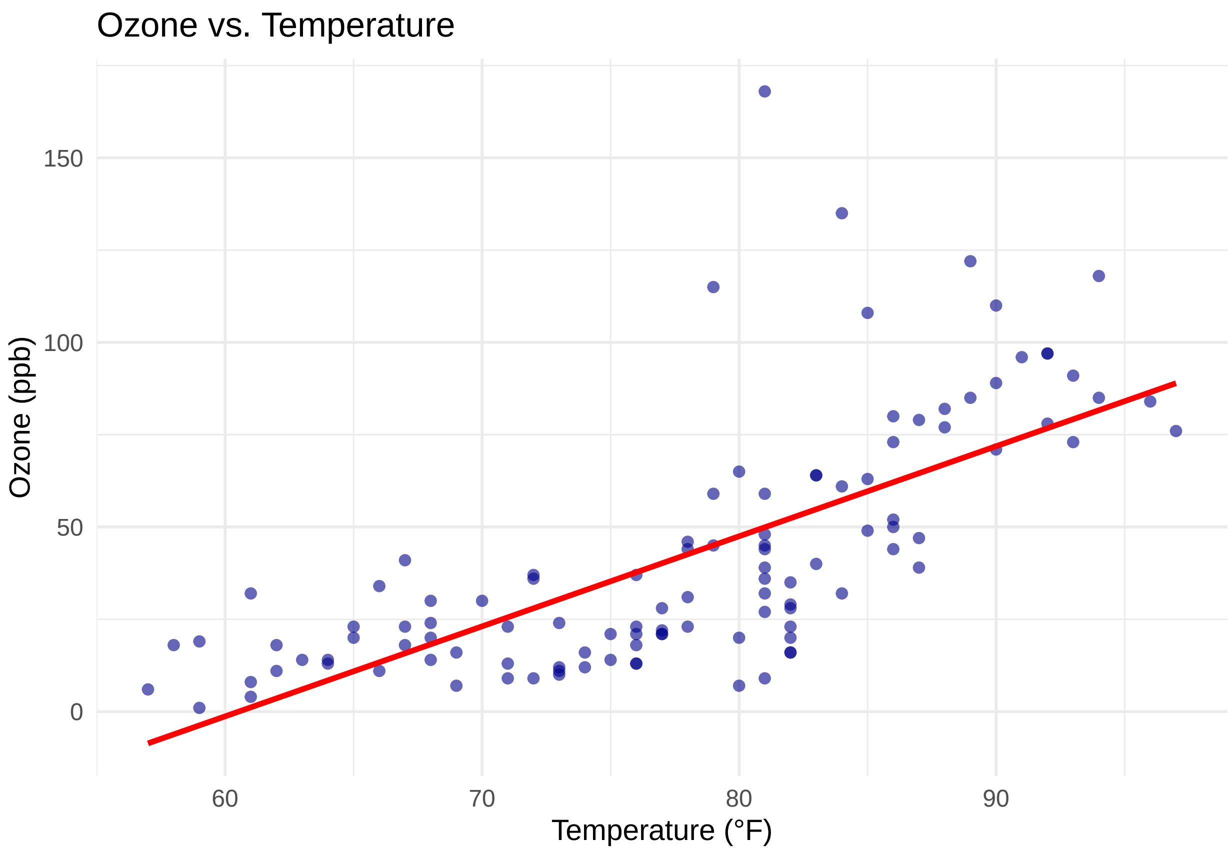
```
## # A tibble: 4 x 5
##   Variable   Mean Median    SD     N
##   <chr>     <dbl>  <dbl> <dbl> <int>
## 1 Ozone     42.1    31   33.3    111
## 2 SolarR   185.    207   91.2    111
## 3 Wind       9.94    9.7  3.56    111
## 4 Temp      77.8    79    9.53    111
```

3. Exploratory Data Analysis (EDA)

Ozone vs. temperature relationship:

```
ggplot(data_air_clean, aes(x = Temp, y = Ozone)) +
  geom_point(alpha = 0.6, color = "darkblue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") + # Add a linear regression line
  labs(title = "Ozone vs. Temperature",
       x = "Temperature (°F)",
       y = "Ozone (ppb)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



We observe that as the temperature increases, the ozone concentration generally tends to increase as well. This indicates a *positive correlation* between temperature and ozone levels.

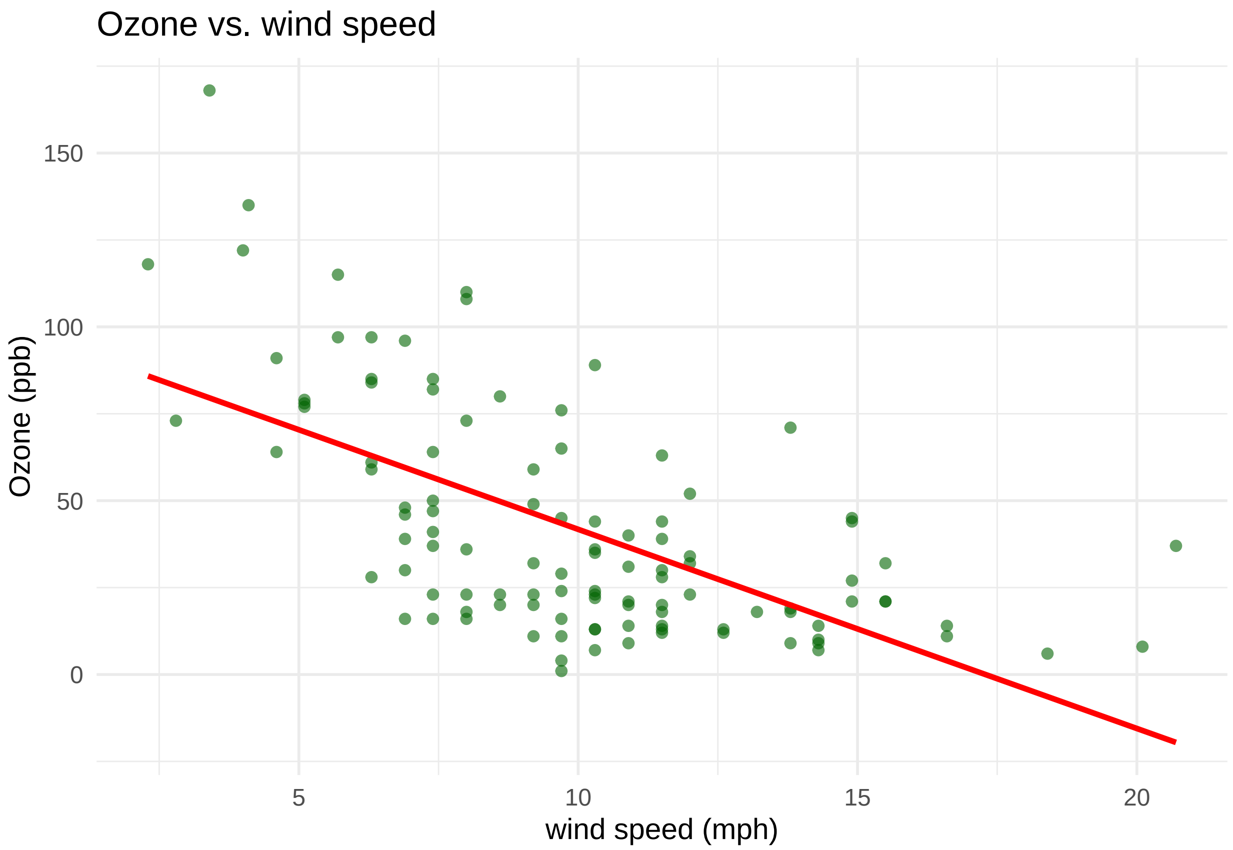
Although there is a clear positive trend, the data points show considerable scatter around the trend line, especially at higher temperatures. This means that temperature is not the only determinant of ozone levels and that other factors can have an influence. For example, around 80°F, ozone levels vary widely from near 0 ppb to over 150 ppb.

There are several cases where ozone levels are exceptionally high (e.g., above 150 ppb) at specific temperatures, especially in the 80°F-90°F range. These cases could be considered *outliers* that have not been accounted for.

Ozone vs. wind relationship:

```
ggplot(data_air_clean, aes(x = Wind, y = Ozone)) +
  geom_point(alpha = 0.6, color = "darkgreen") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Ozone vs. wind speed",
       x = "wind speed (mph)",
       y = "Ozone (ppb)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The higher the wind speed, the lower the ozone concentration generally decreases. This suggests that higher wind speed is associated with lower ozone levels denoting a *negative linear relationship*.

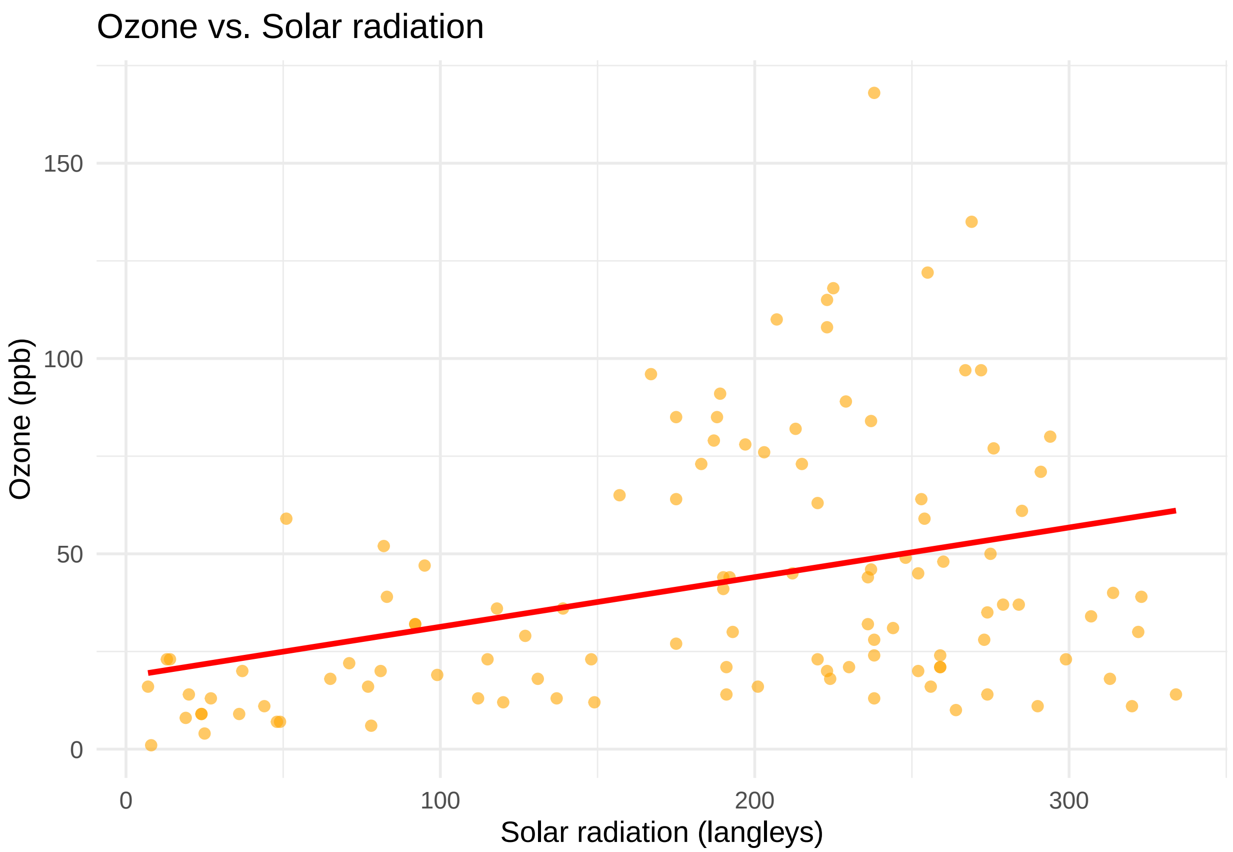
This inverse relationship is expected because stronger winds tend to disperse pollutants, including ozone, reducing their concentration in a given area. Conversely, stagnant air conditions (low wind speed) can lead to ozone accumulation.

There are some data points, particularly at very low wind speeds (e.g., around 2-4 mph), where ozone levels are exceptionally high (e.g., above 150 ppb), indicating that very still air can lead to extreme ozone accumulation.

Ozone vs. Solar radiation:

```
ggplot(data_air_clean, aes(x = Solar.R, y = Ozone)) +
  geom_point(alpha = 0.6, color = "orange") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Ozone vs. Solar radiation",
       x = "Solar radiation (langleys)",
       y = "Ozone (ppb)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



With increasing solar radiation, the ozone concentration tends to increase, indicating a *weak positive relationship*.

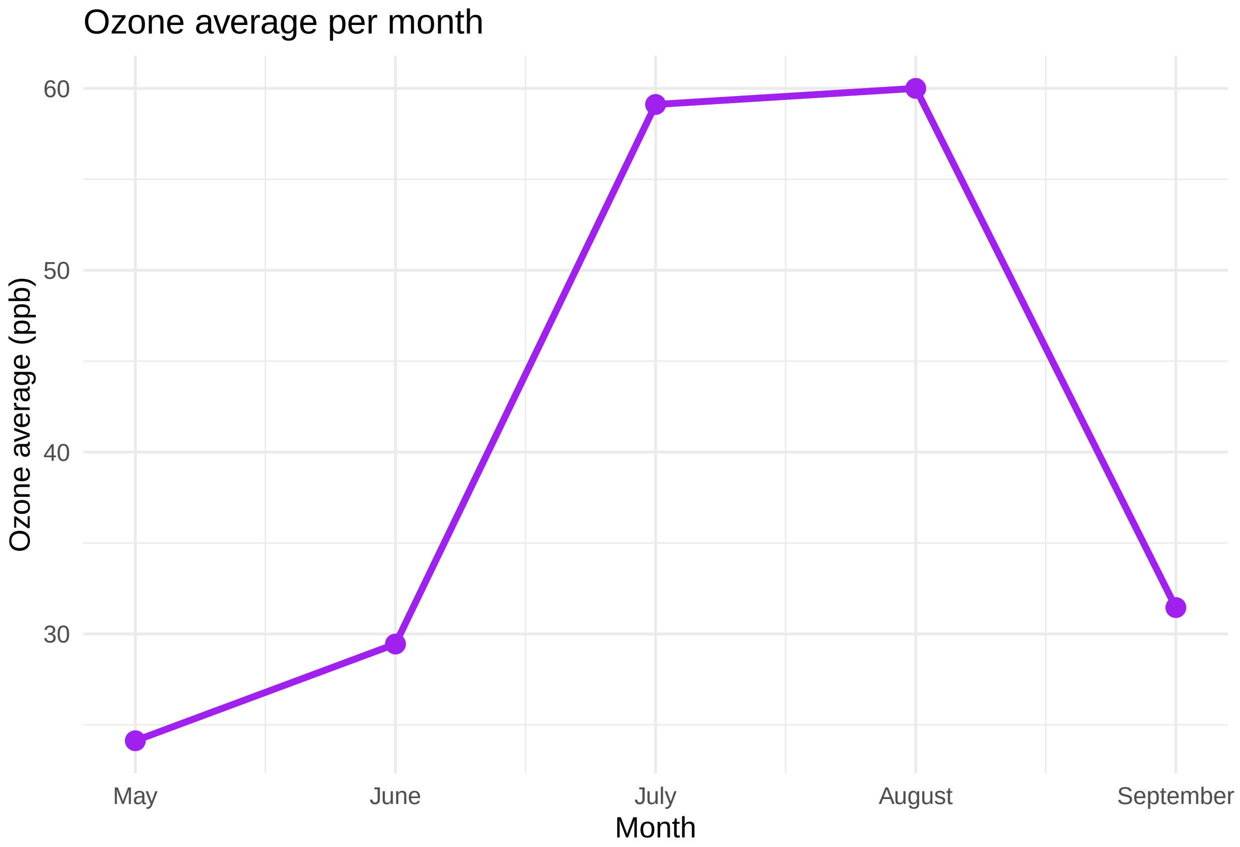
There is a higher density of data points at higher solar radiation ranges, especially between 200 and 300 langleys.

Several points with remarkably high ozone concentrations (above 100 ppb) are observed at high solar radiation ranges.

Average ozone per month

```
data_air_clean %>%
  group_by(Month) %>%
  summarise(Avg_Ozone = mean(Ozone, na.rm = TRUE)) %>%
  ggplot(aes(x = Month, y = Avg_Ozone)) +
  geom_line(color = "purple", size = 1.2) +
  geom_point(color = "purple", size = 3) +
  labs(title = "Ozone average per month",
       x = "Month",
       y = "Ozone average (ppb)") +
  scale_x_continuous(breaks = 5:9, labels = c("May", "June", "July", "August", "September")) +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

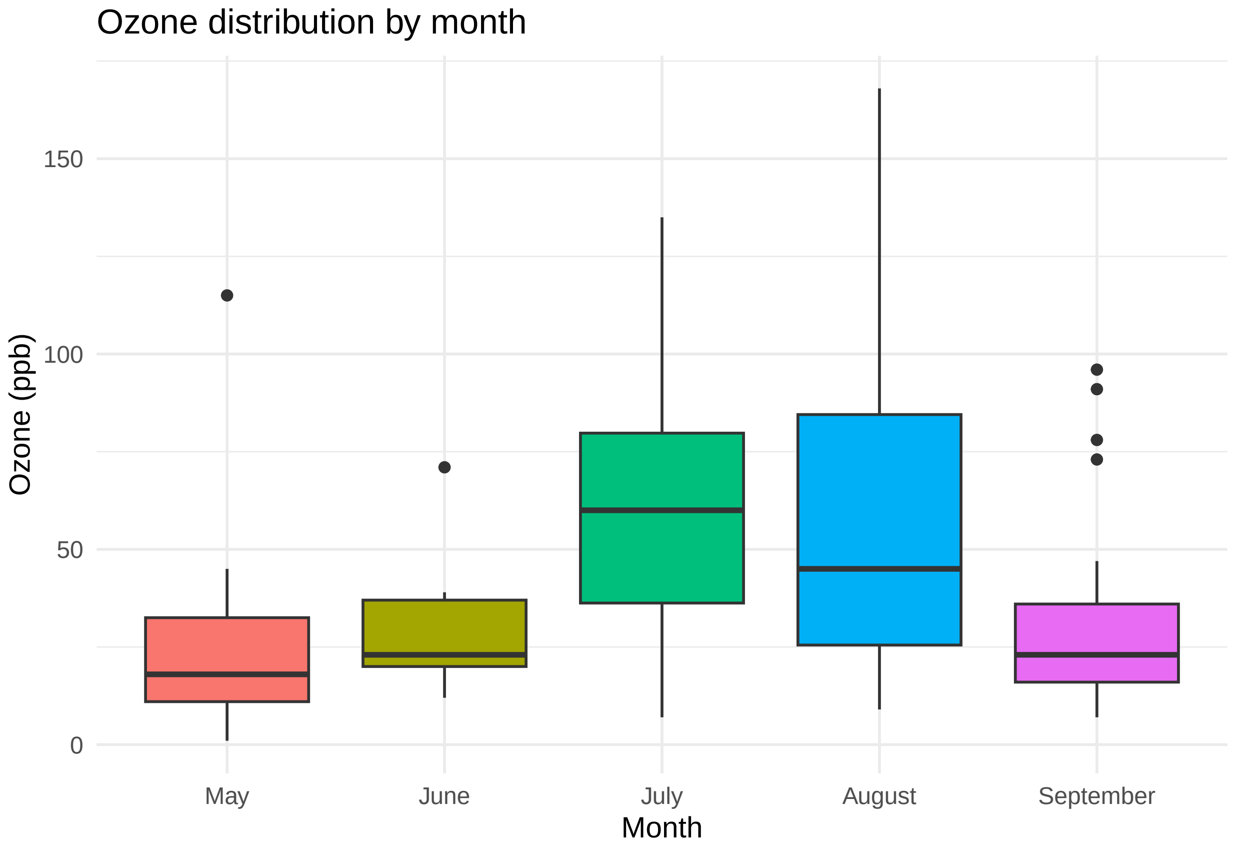


Some seasonal patterns are observed such as increasing ozone levels in summer; ozone concentrations reach a significant peak during July and August, reaching almost 60 ppb. This suggests that the summer months experience the highest average ozone levels.

Decline into the fall; after the August peak, there is a sharp drop in mean ozone in September, falling back to just over 30 ppb.

Boxplot of ozone distribution by month:

```
ggplot(data_air_clean, aes(x = factor(Month), y = Ozone, fill = factor(Month))) +
  geom_boxplot(na.rm = TRUE) +
  labs(title = "Ozone distribution by month",
       x = "Month",
       y = "Ozone (ppb)") +
  scale_x_discrete(labels = c("May", "June", "July", "August", "September")) +
  theme_minimal() +
  guides(fill = "none")
```



May and June (spring) have the lowest ozone concentrations. The boxes are shorter and closer to the bottom of the graph, indicating lower average ozone levels and less variability. May has some atypical elevated ozone readings.

July and August (summer peak) show a significant increase in ozone levels. The squares are much higher on the graph, which means that typical ozone concentrations are much higher. They also show wider squares and longer "whiskers," suggesting greater variability in ozone levels during these summer months.

In September (autumn decline), ozone levels begin to drop again, although they are still generally higher than in May. The box is lower than in July and August, showing a decrease in typical ozone values.

There are several individual points (dots) above the "whiskers" in the box plots, especially in May and September. They represent outlier ozone readings, i.e., unusually high concentrations in those particular months.

4. Questions

How are the main air quality variables (Ozone, Solar Radiation, Wind and Temperature) distributed over the study period?

- Ozone: It presents a positive asymmetric distribution (skewed to the right), with a considerable range from 1 ppb to 168 ppb. The mean (42.13 ppb) is greater than the median (31.50 ppb), indicating exceptionally high values.
- Solar Radiation: Fluctuations from 7 to 334 Langleys, with a more symmetrical distribution than ozone, as the mean (185.9 Langleys) and median (205.0 Langleys) are close. Wind: Range from 1.7 mph to 20.7 mph. The distribution is fairly symmetrical, with most values concentrated around the mean (9.958 mph) and median (9.700 mph).

- Temperature: Range 56°F to 97°F. Relatively symmetrical distribution, with most temperatures between 72°F and 85°F.

Are there seasonal or monthly trends in ozone levels and other variables?

- Ozone: Ozone levels show a clear seasonal pattern. They start relatively low in May (about 24 ppb on average) and increase in June (about 30 ppb). Ozone concentrations peak significantly during July and August, reaching nearly 60 ppb on average, indicating that the summer months experience the highest ozone levels. In September, there is a marked decrease in average ozone, returning to values near 30 ppb. The distribution by month also shows greater variability in July and August.
- Temperature and Ozone by Month: Temperature follows a similar seasonal trend, being higher in the summer months, which aligns with ozone peaks.

What is the relationship between ozone concentration and other environmental variables such as temperature, wind speed and solar radiation?

- Ozone vs. Temperature: There is a positive correlation. As temperature increases, ozone concentration tends to increase. However, there is considerable scatter in the data, especially at higher temperatures, suggesting that temperature is not the only factor. Exceptionally high ozone (over 150 ppb) is observed in the 80°F to 90°F range.
- Ozone vs. Wind Speed: A negative linear relationship is observed. The higher the wind speed, the lower the ozone concentration tends to be. This is explained by the fact that high winds disperse pollutants, including ozone, while stagnant air conditions (low wind speed) can lead to ozone accumulation. There are data points with exceptionally high ozone levels at very low wind speeds (2-4 mph).
- Ozone vs. Solar Radiation: There is a positive relationship. With increasing solar radiation, ozone concentration tends to increase. A higher density of data points is observed at higher solar radiation ranges, and several points with remarkably high ozone concentrations (greater than 100 ppb) occur at these higher radiation ranges.

5. References

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983) Graphical Methods for Data Analysis. Belmont, CA: Wadsworth.