R Project

Ntsika Mahle Mdingi

2025-03-25

Introduction

This project analyzes the air quality dataset (`airquality`) in R. The goal is to clean the data, visualize trends, and summarize key statistics.

Data Cleaning

The dataset contains missing values. Below, I check for and remove missing values.

```r
# Load dataset
data("airquality")

# Check missing values
sum(is.na(airquality))
```

```
## [1] 44
```

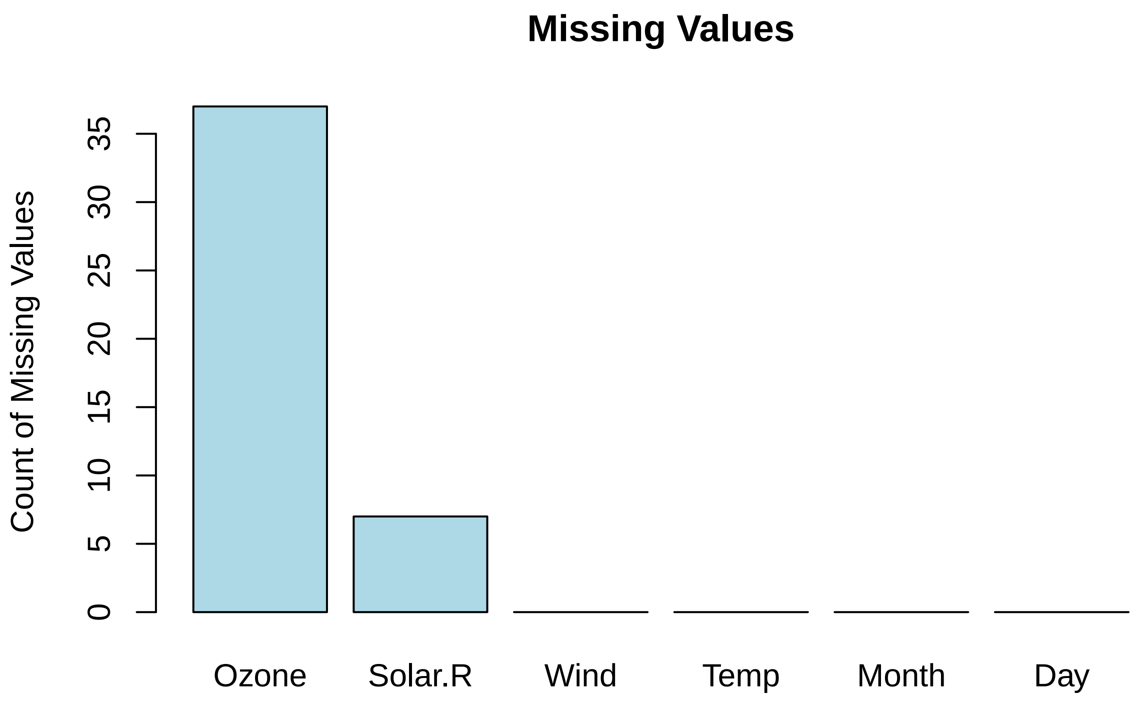
```r
# Remove missing values
Clean_AirQuality <- na.omit(airquality)
```

Data Visualization

Missing Values by Column

This bar chart shows the count of missing values per column.

```r
barplot(colSums(is.na(airquality)),
        main = "Missing Values",
        ylab = "Count of Missing Values",
        col = "lightblue")
```



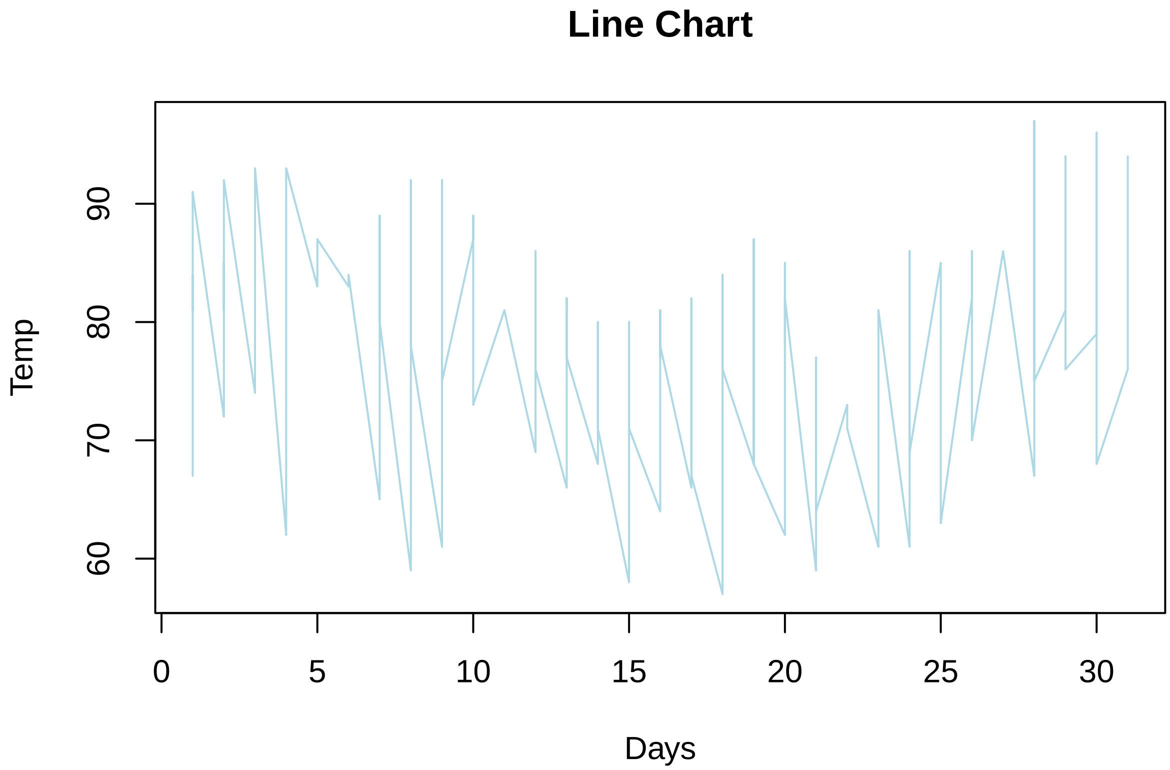
This bar chart shows the count of missing values in each column of the airquality dataset. We observe that Ozone and Solar.R have the highest number of missing values. This could impact the reliability of analysis involving these variables, so handling missing data is necessary before further analysis.

Temperature Trends Over Time

A line chart is used to visualize temperature fluctuations.

```
Clean_AirQuality <- Clean_AirQuality[order(Clean_AirQuality$Day), ]

plot(Clean_AirQuality$Day, Clean_AirQuality$Temp,
     type = "l",
     main = "Line Chart",
     xlab = "Days",
     ylab = "Temp",
     col = "lightblue")
```



This line chart displays the trend of temperature over the days recorded in the dataset. There is a noticeable fluctuation in temperature, with an increasing trend in certain periods. Understanding these patterns can help in predicting air quality conditions.

Summary Statistics

The summary statistics give insights into the distribution of key variables.

```r
summary(Clean_AirQuality[, c("Ozone", "Temp", "Wind")])
```

```
##      Ozone            Temp            Wind      
##  Min.   :  1.0   Min.   :57.00   Min.   : 2.30  
##  1st Qu.: 18.0   1st Qu.:71.00   1st Qu.: 7.40  
##  Median : 31.0   Median :79.00   Median : 9.70  
##  Mean   : 42.1   Mean   :77.79   Mean   : 9.94  
##  3rd Qu.: 62.0   3rd Qu.:84.50   3rd Qu.:11.50  
##  Max.   :168.0   Max.   :97.00   Max.   :20.70  
```

The summary statistics provide insights into the distribution of Ozone, Temperature, and Wind. The mean ozone level is 42.1, with a maximum of 168 and a minimum of 1, showing high variability. The average temperature is 77.9°F, with the highest recorded at 97°F. The wind speed varies between 2.3 and 20.7 mph, with a median value of 9.7 mph. These statistics help in understanding the central tendency and spread of the data, which can be useful for further analysis

Conclusion

From this analysis of the "airquality" dataset, we observed the following key insights:

Missing Data:

The dataset contained missing values, particularly in the Ozone and Solar.R columns. These were removed to ensure accurate analysis.

Temperature Trends:

Temperature fluctuates over the recorded days, with noticeable variations. This could be influenced by seasonal changes or external environmental factors.

Wind Speed:

The average wind speed is approximately 9.9 mph, with some variation. Understanding wind patterns could help explain changes in air quality.

Ozone Levels:

Ozone levels vary across different days, and further analysis could explore potential correlations between Ozone, temperature, and wind speed.

This initial analysis provides a basic understanding of air quality trends, but deeper insights could be gained by investigating external factors like pollution sources, weather conditions, or seasonal effects.