AirQuality Data set

Anahi Quezada

University of Kansas

November 2024

1 Data information

Data: Air quality

link : https://archive.ics.uci.edu/dataset/360/air+quality

Contains the responses of a gas multi-sensor device deployed on the field in an Italian city. Hourly response averages are recorded along with gas concentrations references from a certified analyzer.

Additional Information

The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level,within an Italian city. Data were recorded from March 2004 to February 2005 (one year)representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) were provided by a co-located reference certified analyzer [1].

2 Data process

I cleaned the data to obtain a new dataset without NAs values using a source file `na_delete.R` with the function `na.omit`. The raw dataset contained some columns that I will not use for these analyses. This way, I selected the following columns: `CO.GT.`, `C6H6.GT.`, `NOx.GT.`, and Contamination column. All these process was made using R 4.0.5. After this execution, as part of the data visualization, I obtained (Figure 1) where I can show the frequencies of the pollutants in the air.

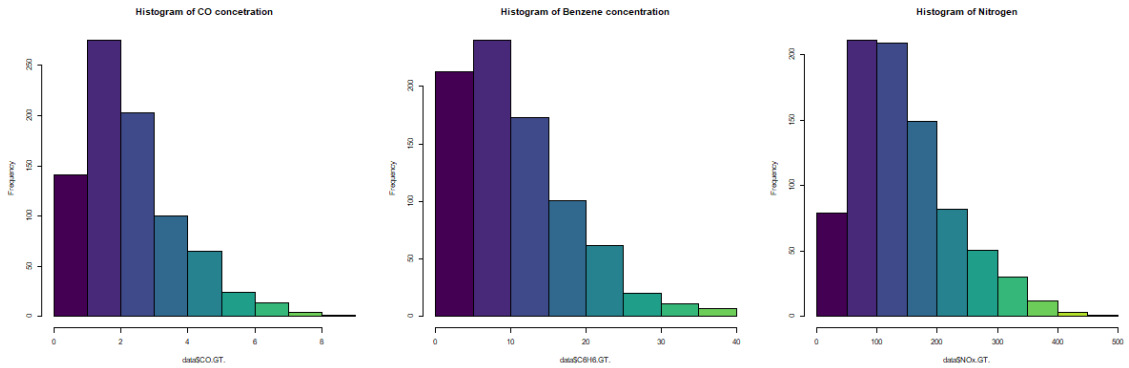

Figure 1: Histogram of pollutants

The palette color of this graphic was made with the source colores.R which contains colorblind-friendly colors.

Permissible threshold

```
data$contaminacion <- ifelse(data$CO.GT. > threshold_CO |          (1)
                               data$C6H6.GT. > threshold_C6H6 |    (2)
                             data$NOx.GT. > threshold_Nox.GT, 1, 0) (3)
```

Equation 3, shows the process in which, according to Italian regulations (maximum permissible limit), the contamination column was created. This is the column that will be used for the following analyses. After this process, the data contains the contaminacion column which has values from 0 to 1. O represents the absence of pollution and 1 the presence of pollution. The execution of this part of the code was made as indicated:

```
data$contaminacion <- factor(data$contaminacion,   (4)
                             levels = c(0, 1),     (5)
                           labels = c("L", "H"))   (6)
```

To visualize these primary results, figure 2, shows the presence of contamination.

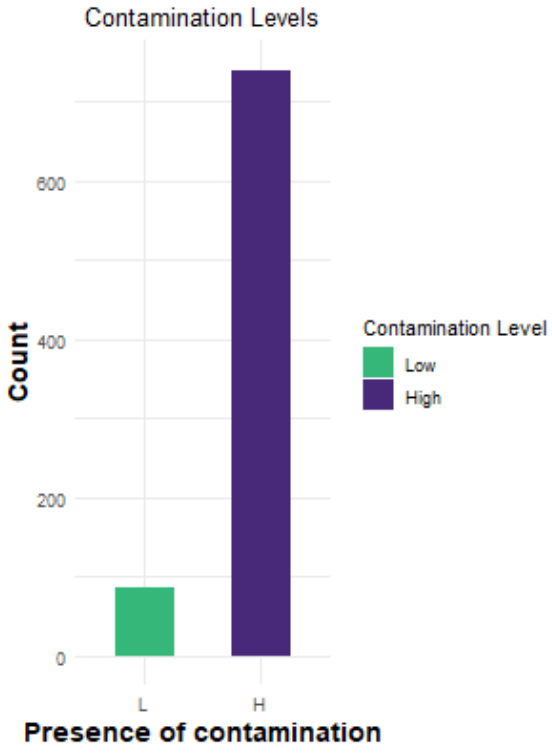

Figure 2: Contamination levels

3 Methods

4 Results

5 Discussion

References

[1] Saverio Vito. Air Quality, 2008.