University of Padova
Computer Science
a.y. 2021/2022

Data Mining

Teacher: Annamaria Guolo

Dataset `pollution`

Dataset included in `pollution.RData` refers to a study about the relationship between mortality in 60 US areas and air pollution. Some environmental and demographical information are collected.

- `mortality`: mortality rate (annual deaths for 100000 persons)
- `precipitation`: mean annual precipitation (inches)
- `humidity`: percent relative humidity
- `Jan.temp`: mean January temperature (Farenheit)
- `July.temp`: mean July temperature (Farenheit)
- `over65`: percentage of the population aged 65 years or over
- `house`: population per household
- `education`: median number of school years completed for persons 25 years or older
- `comfort`: percentage of the housing that is sound with all facilities
- `density`: population density (in persons per square mile)
- `office`: percentage of office workers
- `poor`: percentage of households with annual income under 3000 dollars
- `HC`: level of hydrocarbons
- `NOX`: dangerous level of oxides of nitrogen?: Yes ($> 30\ \mu g/mc$), No ($\leq 30\ \mu g/mc$)
- `SO2`: dangerous level of sulfur dioxide?: Yes ($> 125\ \mu g/mc$), No ($\leq 125\ \mu g/mc$)

FIRST QUESTION.
Consider the dataset composed by `mortality`, `precipitation`, `humidity`, `HC`, `NOX`, `SO2`. Construct the most appropriate model for the purpose of the analysis. Which variables are associated to the mortality rate?

SECOND QUESTION.
Consider all the variables in the dataset. Construct the most appropriate model for the purpose of the analysis. Which variables are associated to the mortality rate?