Communities and Crime Data Set

Abstract: The data combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR.

The dataset gathers information from different communities in the United States about several factors that can highly influence some common crimes such as robberies, murders or rapes. More precisely, at the beginning, there were 147 variables, four of which contained the names of the communities included in the sample and the state each one belonged to. 18 variables have the information about the crimes which are expected to be explained through 125 more variables with the main information about all the communities. The complete description of all these variables and the dataset can be obtained from the source repository "UCI Machine Learning" (https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime).

The data proposed for the project are a simplified version of the original dataset and contains 8 variables of interest which are explained below.

Qualitative or Categorical Variables

1. **Predominant ethnicity (Nominal)**
   - '1' African American
   - '2' Caucasian
   - '3' Asian heritage
   - '4' Hispanic heritage

   Variable is not explicitly defined in the dataset; it was obtained by summarizing four other parameters:
   - ✓ racepctblack: percentage of population that is African.
   - ✓ racePctWhite: percentage of population that is Caucasian.
   - ✓ racePctAsian: percentage of population that is of Asian heritage.
   - ✓ racePctHisp: percentage of population that is of Hispanic heritage.

2. **Predominant educational level (Ordinal)**
   - '1' Primary school
   - '2' Secondary school
   - '3' Bachelor's degree or higher education

   Variable is not explicitly defined in the dataset; it was obtained by summarizing three other parameters:
   - ✓ PctLess9thGrade: percentage of people 25 and over with less than a 9th grade education
   - ✓ PctNotHSGrad: percentage of people 25 and over that are not high school graduates
   - ✓ PctBSorMore: percentage of people 25 and over with a bachelor's degree or higher education

3. **Population Density (Ordinal)**
    - ‘1’ Low Density
    - ‘2’ Middle Density
    - ‘3’ High Density

   Variable was obtained by discretizing parameter:

    - ✓ PopDens: population density in persons per square mile

Quantitative Variables

4. **Population (Discrete)**
   Population for community

5. **Household Income (Continuous)**
   Median household income

6. **Social Security (Continuous)**
   Percentage of households with social security income in 1989

7. **Crime Rate (Continuous)**
   Variable obtained as the sum of Violent and Non-Violent crimes per 100K population.