Documentation for SMSA Dataset: Air Pollution and Mortality

Amy Willis

BACKGROUND

Researchers at General Motors collected data on 60 U.S. Standard Metropolitan Statistical Areas (SMSA's) in a study of whether air pollution contributes to mortality. The data include variables measuring demographic characteristics of the cities, variables measuring climate characteristics, and variables recording the pollution potential of three different air pollutants.

The data was published in 1973 in *Technometrics* by McDonald & Schwing (1973).

AVAILABLE DATA

The data are stored in the file `smsa.csv`. Each line corresponds to the observations on one of the 60 Standard Metropolitan Statistical Areas (a standard Census Bureau designation of the region around a city) in the United States, collected from a variety of sources. To my knowledge, the dataset dates to the 1960's.

The descriptions of the variables are as follows:

- `city`: City name
- `state`: State that the city is located in
- `JanTemp`: Mean January temperature (degrees Farenheit)
- `JulyTemp`: Mean July temperature (degrees Farenheit)
- `RelHum`: Relative Humidity (percentage)
- `Rain`: Mean annual rainfall (inches)
- `Mortality`: Total age-adjusted all-cause mortality rate, expressed as deaths per 100,000 residents
- `Education`: Median school years completed for those over 25 in 1960 (Years)
- `PopDensity`: Population density: Population per square mile in urbanised area
- `pNonWhite`: Percentage of 1960 population in urbanised areas who are not White
- `pWC`: Percentage of employed workers in white-collar occupations in 1960
- `pop`: Population
- `pophouse`: Population per household in 1960
- `income`: Median income
- `HCPot`: HC pollution potential
- `NOxPot`: Nitrous Oxide pollution potential
- `SO2Pot`: Sulfur Dioxide pollution potential

According to the original paper, "The pollution potential is determined as the product of the tons emitted per day per square kilometre of each pollutant and a dispersion factor which accounts for mixing height, wind speed, number of episodes days and dimension of each SMSA." Therefore, the units of the pollution measurements could be "adjusted tonnes per day per square kilometre," or more simply, "pollution potential units".

You can read in the data as follows:

```r
library(tidyverse)
read_csv("smsa.csv")
```