**Title: Communities and Crime**

**Abstract: Communities within the United States. 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.**

**--------------------------------------------------------------------------------------**

**Source:**

**Creator: Michael Redmond (redmond 'at' lasalle.edu); Computer Science; La Salle**

**University; Philadelphia, PA, 19141, USA**

**-- culled from 1990 US Census, 1995 US FBI Uniform Crime Report, 1990 US Law**

**Enforcement Management and Administrative Statistics Survey, available from ICPSR at U**

**of Michigan.**

**-- Donor: Michael Redmond (redmond 'at' lasalle.edu); Computer Science; La Salle**

**University; Philadelphia, PA, 19141, USA**

**-- Date: July 2009**

**--------------------------------------------------------------------------------------**

**Data Set Information:**

**Many variables are included so that algorithms that select or learn weights for**

**attributes could be tested. However, clearly unrelated attributes were not included;**

**attributes were picked if there was any plausible connection to crime (N=122), plus**

**the attribute to be predicted (Per Capita Violent Crimes). The variables included in**

**the dataset involve the community, such as the percent of the population considered**

**urban, and the median family income, and involving law enforcement, such as per capita**

**number of police officers, and percent of officers assigned to drug units.**

**The per capita violent crimes variable was calculated using population and the sum of**

**crime variables considered violent crimes in the United States: murder, rape, robbery,**

**and assault. There was apparently some controversy in some states concerning the**

**counting of rapes. These resulted in missing values for rape, which resulted in**

**incorrect values for per capita violent crime. These cities are not included in the**

**dataset. Many of these omitted communities were from the midwestern USA.**

**Data is described below based on original values. All numeric data was normalized into**

**the decimal range 0.00-1.00 using an Unsupervised, equal-interval binning method.**

**Attributes retain their distribution and skew (hence for example the population**

**attribute has a mean value of 0.06 because most communities are small). E.g. An**

**attribute described as 'mean people per household' is actually the normalized (0-1)**

**version of that value.**

**The normalization preserves rough ratios of values WITHIN an attribute (e.g. double**

**the value for double the population within the available precision - except for**

**extreme values (all values more than 3 SD above the mean are normalized to 1.00; all**

**values more than 3 SD below the mean are nromalized to 0.00)).**

**However, the normalization does not preserve relationships between values BETWEEN**

**attributes (e.g. it would not be meaningful to compare the value for whitePerCap with**

**the value for blackPerCap for a community)**

**-----------------------------------------------------------------------------**

**Attribute Information:**

**-- state: US state (by number) - not counted as predictive above, but if considered, should be considerd nominal (nominal)**

**-- county: numeric code for county - not predictive, and many missing values (numeric)**

**-- population: population for community: (numeric - decimal)**

**-- householdsize: mean people per household (numeric - decimal)**

**-- racepctblack: percentage of population that is african american (numeric - decimal)**

**-- racePctWhite: percentage of population that is caucasian (numeric - decimal)**

**-- racePctAsian: percentage of population that is of asian heritage (numeric - decimal)**

**-- racePctHisp: percentage of population that is of hispanic heritage (numeric - decimal)**

**-- agePct12t21: percentage of population that is 12-21 in age (numeric - decimal)**

**-- agePct12t29: percentage of population that is 12-29 in age (numeric - decimal)**

**-- agePct16t24: percentage of population that is 16-24 in age (numeric - decimal)**

**-- agePct65up: percentage of population that is 65 and over in age (numeric - decimal)**

**-- numbUrban: number of people living in areas classified as urban (numeric - decimal)**

**-- pctUrban: percentage of people living in areas classified as urban (numeric - decimal)**

**-- medIncome: median household income (numeric - decimal)**

**-- pctWWage: percentage of households with wage or salary income in 1989 (numeric - decimal)**

**-- pctWFarmSelf: percentage of households with farm or self employment income in 1989 (numeric - decimal)**

**-- pctWInvInc: percentage of households with investment / rent income in 1989 (numeric - decimal)**

**-- pctWSocSec: percentage of households with social security income in 1989 (numeric - decimal)**

**-- pctWPubAsst: percentage of households with public assistance income in 1989 (numeric - decimal)**

**-- pctWRetire: percentage of households with retirement income in 1989 (numeric - decimal)**

**-- medFamInc: median family income (differs from household income for non-family households) (numeric - decimal)**

**-- perCapInc: per capita income (numeric - decimal)**

**-- whitePerCap: per capita income for caucasians (numeric - decimal)**

**-- blackPerCap: per capita income for african americans (numeric - decimal)**

**-- indianPerCap: per capita income for native americans (numeric - decimal)**

**-- AsianPerCap: per capita income for people with asian heritage (numeric - decimal)**

**-- OtherPerCap: per capita income for people with 'other' heritage (numeric - decimal)**

**-- HispPerCap: per capita income for people with hispanic heritage (numeric - decimal)**

**-- NumUnderPov: number of people under the poverty level (numeric - decimal)**

**-- PctPopUnderPov: percentage of people under the poverty level (numeric - decimal)**

**-- PctLess9thGrade: percentage of people 25 and over with less than a 9th grade education (numeric - decimal)**

**-- PctNotHSGrad: percentage of people 25 and over that are not high school graduates (numeric - decimal)**

**-- PctBSorMore: percentage of people 25 and over with a bachelors degree or higher education (numeric - decimal)**

**-- PctUnemployed: percentage of people 16 and over, in the labor force, and unemployed (numeric - decimal)**

**-- PctEmploy: percentage of people 16 and over who are employed (numeric - decimal)**

**-- PctEmplManu: percentage of people 16 and over who are employed in manufacturing (numeric - decimal)**

**-- PctEmplProfServ: percentage of people 16 and over who are employed in professional services (numeric - decimal)**

**-- MalePctDivorce: percentage of males who are divorced (numeric - decimal)**

**-- MalePctNevMarr: percentage of males who have never married (numeric - decimal)**

**-- FemalePctDiv: percentage of females who are divorced (numeric - decimal)**

**-- TotalPctDiv: percentage of population who are divorced (numeric - decimal)**

**-- PersPerFam: mean number of people per family (numeric - decimal)**

**-- PctFam2Par: percentage of families (with kids) that are headed by two parents (numeric - decimal)**

**-- PctKids2Par: percentage of kids in family housing with two parents (numeric - decimal)**

**-- PctYoungKids2Par: percent of kids 4 and under in two parent households (numeric - decimal)**

**-- PctTeen2Par: percent of kids age 12-17 in two parent households (numeric - decimal)**

**-- PctWorkMomYoungKids: percentage of moms of kids 6 and under in labor force (numeric - decimal)**

**-- PctWorkMom: percentage of moms of kids under 18 in labor force (numeric - decimal)**

**-- NumIlleg: number of kids born to never married (numeric - decimal)**

**-- PctIlleg: percentage of kids born to never married (numeric - decimal)**

**-- NumImmig: total number of people known to be foreign born (numeric - decimal)**

**-- PctImmigRecent: percentage of \_immigrants\_ who immigated within last 3 years (numeric - decimal)**

**-- PctImmigRec5: percentage of \_immigrants\_ who immigated within last 5 years (numeric - decimal)**

**-- PctImmigRec8: percentage of \_immigrants\_ who immigated within last 8 years (numeric - decimal)**

**-- PctImmigRec10: percentage of \_immigrants\_ who immigated within last 10 years (numeric - decimal)**

**-- PctRecentImmig: percent of \_population\_ who have immigrated within the last 3 years (numeric - decimal)**

**-- PctRecImmig5: percent of \_population\_ who have immigrated within the last 5 years (numeric - decimal)**

**-- PctRecImmig8: percent of \_population\_ who have immigrated within the last 8 years (numeric - decimal)**

**-- PctRecImmig10: percent of \_population\_ who have immigrated within the last 10 years (numeric - decimal)**

**-- PctSpeakEnglOnly: percent of people who speak only English (numeric - decimal)**

**-- PctNotSpeakEnglWell: percent of people who do not speak English well (numeric - decimal)**

**-- PctLargHouseFam: percent of family households that are large (6 or more) (numeric - decimal)**

**-- PctLargHouseOccup: percent of all occupied households that are large (6 or more people) (numeric - decimal)**

**-- PersPerOccupHous: mean persons per household (numeric - decimal)**

**-- PersPerOwnOccHous: mean persons per owner occupied household (numeric - decimal)**

**-- PersPerRentOccHous: mean persons per rental household (numeric - decimal)**

**-- PctPersOwnOccup: percent of people in owner occupied households (numeric - decimal)**

**-- PctPersDenseHous: percent of persons in dense housing (more than 1 person per room) (numeric - decimal)**

**-- PctHousLess3BR: percent of housing units with less than 3 bedrooms (numeric - decimal)**

**-- MedNumBR: median number of bedrooms (numeric - decimal)**

**-- HousVacant: number of vacant households (numeric - decimal)**

**-- PctHousOccup: percent of housing occupied (numeric - decimal)**

**-- PctHousOwnOcc: percent of households owner occupied (numeric - decimal)**

**-- PctVacantBoarded: percent of vacant housing that is boarded up (numeric - decimal)**

**-- PctVacMore6Mos: percent of vacant housing that has been vacant more than 6 months (numeric - decimal)**

**-- MedYrHousBuilt: median year housing units built (numeric - decimal)**

**-- PctHousNoPhone: percent of occupied housing units without phone (in 1990, this was rare!) (numeric - decimal)**

**-- PctWOFullPlumb: percent of housing without complete plumbing facilities (numeric - decimal)**

**-- OwnOccLowQuart: owner occupied housing - lower quartile value (numeric - decimal)**

**-- OwnOccMedVal: owner occupied housing - median value (numeric - decimal)**

**-- OwnOccHiQuart: owner occupied housing - upper quartile value (numeric - decimal)**

**-- RentLowQ: rental housing - lower quartile rent (numeric - decimal)**

**-- RentMedian: rental housing - median rent (Census variable H32B from file STF1A) (numeric - decimal)**

**-- RentHighQ: rental housing - upper quartile rent (numeric - decimal)**

**-- MedRent: median gross rent (Census variable H43A from file STF3A - includes utilities) (numeric - decimal)**

**-- MedRentPctHousInc: median gross rent as a percentage of household income (numeric - decimal)**

**-- MedOwnCostPctInc: median owners cost as a percentage of household income - for owners with a mortgage (numeric - decimal)**

**-- MedOwnCostPctIncNoMtg: median owners cost as a percentage of household income - for owners without a mortgage (numeric - decimal)**

**-- NumInShelters: number of people in homeless shelters (numeric - decimal)**

**-- NumStreet: number of homeless people counted in the street (numeric - decimal)**

**-- PctForeignBorn: percent of people foreign born (numeric - decimal)**

**-- PctBornSameState: percent of people born in the same state as currently living (numeric - decimal)**

**-- PctSameHouse85: percent of people living in the same house as in 1985 (5 years before) (numeric - decimal)**

**-- PctSameCity85: percent of people living in the same city as in 1985 (5 years before) (numeric - decimal)**

**-- PctSameState85: percent of people living in the same state as in 1985 (5 years before) (numeric - decimal)**

**-- LandArea: land area in square miles (numeric - decimal)**

**-- PopDens: population density in persons per square mile (numeric - decimal)**

**-- PctUsePubTrans: percent of people using public transit for commuting (numeric - decimal)**

**-- ViolentCrimesPerPop: total number of violent crimes per 100K popuation (numeric - decimal) GOAL attribute (to be predicted)**

**Distribution of the Goal Variable (Violent Crimes per Population):**

**Range Frequency**

**0.000-0.067 484**

**0.067-0.133 420**

**0.133-0.200 284**

**0.200-0.267 177**

**0.267-0.333 142**

**0.333-0.400 113**

**0.400-0.467 59**

**0.467-0.533 76**

**0.533-0.600 57**

**0.600-0.667 38**

**0.667-0.733 37**

**0.733-0.800 20**

**0.800-0.867 23**

**0.867-0.933 14**

**0.933-1.000 50**

**--------------------------------------------------------------------------------------**