Abstract:

Communities in the US. Data combines socio-economic data from the '90 Census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats survey, and crime data from the 1995 FBI UCR

Data Set Characteristics:	Multivariate	**Number of Instances:**	2215
Attribute Characteristics:	Real	**Number of Attributes:**	147
Associated Tasks:	Regression	**Missing Values?**	Yes

Data Set Information:

Variables included in 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.

Crime attributes (N=18) that could be predicted are **8 crimes** considered 'Index Crimes' by the FBI, per capita (actually per 100,000 population) versions of each

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**.

Per capita **nonviolent** crime variable was calculated using the sum of crime variables considered non-violent crimes in the United States: **burglaries, larcenies, auto thefts and arsons**.

It **would not be interesting or appropriate to predict total crime** (e.g. violent crime). There are also identifying variables (community name, county code, community code) that are not predictive, and would get in the way of some algorithms.

Attribute Information:

(124 predictive, 5 non-predictive, 18 potential goal)

1	community name	Community name - not predictive - for information only (string)
2	state	US state (by 2 letter postal abbreviation)(nominal)
3	countyCode	numeric code for county - not predictive, and many missing values (numeric)
4	communityCode	numeric code for community - not predictive and many missing values (numeric)
5	fold	fold number for non-random 10 fold cross validation, potentially useful for debugging, paired tests - not predictive (numeric - integer)
1	population	population for community
2	householdsize	mean people per household (numeric - decimal)
3	racepctblack	percentage of population that is African American (numeric - decimal)

4	racePctWhite	percentage of population that is Caucasian (numeric - decimal)
5	racePctAsian	percentage of population that is of Asian heritage (numeric - decimal)
6	racePctHisp	percentage of population that is of Hispanic heritage (numeric - decimal)
7	agePct12t21	percentage of population that is 12-21 in age (numeric - decimal)
8	agePct12t29	percentage of population that is 12-29 in age (numeric - decimal)
9	agePct16t24	percentage of population that is 16-24 in age (numeric - decimal)
10	agePct65up	percentage of population that is 65 and over in age (numeric - decimal)
11	numbUrban	number of people living in areas classified as urban (numeric - expected to be integer)
12	pctUrban	percentage of people living in areas classified as urban (numeric - decimal)
13	medIncome	median household income (numeric - may be integer)
14	pctWWage	percentage of households with wage or salary income in 1989 (numeric - decimal)
15	pctWFarmSelf	percentage of households with farm or self-employment income in 1989 (numeric - decimal)
16	pctWInvInc	percentage of households with investment / rent income in 1989 (numeric - decimal)
17	pctWSocSec	percentage of households with social security income in 1989 (numeric - decimal)
18	pctWPubAsst	percentage of households with public assistance income in 1989 (numeric - decimal)
19	pctWRetire	percentage of households with retirement income in 1989 (numeric - decimal)
20	medFamInc	median family income (differs from household income for non-family households) (numeric - may be integer)
21	perCapInc	per capita income (numeric - decimal)
22	whitePerCap	per capita income for Caucasians (numeric - decimal)
23	blackPerCap	per capita income for African Americans (numeric - decimal)
24	indianPerCap	per capita income for native Americans (numeric - decimal)
25	AsianPerCap	per capita income for people with Asian heritage (numeric - decimal)
26	OtherPerCap	per capita income for people with 'other' heritage (numeric - decimal)
27	HispPerCap	per capita income for people with Hispanic heritage (numeric - decimal)
28	NumUnderPov	number of people under the poverty level (numeric - expected to be integer)
29	PctPopUnderPov	percentage of people under the poverty level (numeric - decimal)
30	PctLess9thGrade	percentage of people 25 and over with less than a 9th grade education (numeric - decimal)
31	PctNotHSGrad	percentage of people 25 and over that are not high school graduates (numeric - decimal)
32	PctBSorMore	percentage of people 25 and over with a bachelor's degree or higher education (numeric - decimal)
33	PctUnemployed	percentage of people 16 and over, in the labor force, and unemployed (numeric - decimal)
34	PctEmploy	percentage of people 16 and over who are employed (numeric - decimal)
35	PctEmplManu	percentage of people 16 and over who are employed in manufacturing (numeric - decimal)
36	PctEmplProfServ	percentage of people 16 and over who are employed in professional services (numeric - decimal)
37	PctOccupManu	percentage of people 16 and over who are employed in manufacturing (numeric - decimal)
38	PctOccupMgmtProf	percentage of people 16 and over who are employed in management or professional occupations (numeric - decimal)
39	MalePctDivorce	percentage of males who are divorced (numeric - decimal)
40	MalePctNevMarr	percentage of males who have never married (numeric - decimal)
41	FemalePctDiv	percentage of females who are divorced (numeric - decimal)
42	TotalPctDiv	percentage of population who are divorced (numeric - decimal)
43	PersPerFam	mean number of people per family (numeric - decimal)

44	PctFam2Par	percentage of families (with kids) that are headed by two parents (numeric - decimal)
45	PctKids2Par	percentage of kids in family housing with two parents (numeric - decimal)
46	PctYoungKids2Par	percent of kids 4 and under in two parent households (numeric - decimal)
47	PctTeen2Par	percent of kids age 12-17 in two parent households (numeric - decimal)
48	PctWorkMomYoungKids	percentage of moms of kids 6 and under in labor force (numeric - decimal)
49	PctWorkMom	percentage of moms of kids under 18 in labor force (numeric - decimal)
50	NumKidsBornNeverMar	number of kids born to never married (numeric - expected to be integer)
51	PctKidsBornNeverMar	percentage of kids born to never married (numeric - decimal)
52	NumImmig	total number of people known to be foreign born (numeric - expected to be integer)
53	PctImmigRecent	percentage of _immigrants_ who immigrated within last 3 years (numeric - decimal)
54	PctImmigRec5	percentage of _immigrants_ who immigrated within last 5 years (numeric - decimal)
55	PctImmigRec8	percentage of _immigrants_ who immigrated within last 8 years (numeric - decimal)
56	PctImmigRec10	percentage of _immigrants_ who immigrated within last 10 years (numeric - decimal)
57	PctRecentImmig	percent of _population_ who have immigrated within the last 3 years (numeric - decimal)
58	PctRecImmig5	percent of _population_ who have immigrated within the last 5 years (numeric - decimal)
59	PctRecImmig8	percent of _population_ who have immigrated within the last 8 years (numeric - decimal)
60	PctRecImmig10	percent of _population_ who have immigrated within the last 10 years (numeric - decimal)
61	PctSpeakEnglOnly	percent of people who speak only English (numeric - decimal)
62	PctNotSpeakEnglWell	percent of people who do not speak English well (numeric - decimal)
63	PctLargHouseFam	percent of family households that are large (6 or more) (numeric - decimal)
64	PctLargHouseOccup	percent of all occupied households that are large (6 or more people) (numeric - decimal)
65	PersPerOccupHous	mean persons per household (numeric - decimal)
66	PersPerOwnOccHous	mean persons per owner occupied household (numeric - decimal)
67	PersPerRentOccHous	mean persons per rental household (numeric - decimal)
68	PctPersOwnOccup	percent of people in owner occupied households (numeric - decimal)
69	PctPersDenseHous	percent of persons in dense housing (more than 1 person per room) (numeric - decimal)
70	PctHousLess3BR	percent of housing units with less than 3 bedrooms (numeric - decimal)
71	MedNumBR	median number of bedrooms (numeric - decimal)
72	HousVacant	number of vacant households (numeric - expected to be integer)
73	PctHousOccup	percent of housing occupied (numeric - decimal)
74	PctHousOwnOcc	percent of households owner occupied (numeric - decimal)
75	PctVacantBoarded	percent of vacant housing that is boarded up (numeric - decimal)
76	PctVacMore6Mos	percent of vacant housing that has been vacant more than 6 months (numeric - decimal)
77	MedYrHousBuilt	median year housing units built (numeric - may be integer)
78	PctHousNoPhone	percent of occupied housing units without phone (in 1990, this was rare!) (numeric - decimal)
79	PctWOFullPlumb	percent of housing without complete plumbing facilities (numeric - decimal)
80	OwnOccLowQuart	owner occupied housing - lower quartile value (numeric - decimal)
81	OwnOccMedVal	owner occupied housing - median value (numeric - decimal)
82	OwnOccHiQuart	owner occupied housing - upper quartile value (numeric - decimal)
83	OwnOccQrange	owner occupied housing - difference between upper quartile and lower quartile values (numeric - decimal)
84	RentLowQ	rental housing - lower quartile rent (numeric - decimal)

85	RentMedian	rental housing - median rent (Census variable H32B from file STF1A) (numeric - decimal)
86	RentHighQ	rental housing - upper quartile rent (numeric - decimal)
87	RentQrange	rental housing - difference between upper quartile and lower quartile rent (numeric - decimal)
88	MedRent	median gross rent (Census variable H43A from file STF3A - includes utilities) (numeric - decimal)
89	MedRentPctHousInc	median gross rent as a percentage of household income (numeric - decimal)
90	MedOwnCostPctInc	median owners cost as a percentage of household income - for owners with a mortgage (numeric - decimal)
91	MedOwnCostPctIncNoMtg	median owners cost as a percentage of household income - for owners without a mortgage (numeric - decimal)
92	NumInShelters	number of people in homeless shelters (numeric - expected to be integer)
93	NumStreet	number of homeless people counted in the street (numeric - expected to be integer)
94	PctForeignBorn	percent of people foreign born (numeric - decimal)
95	PctBornSameState	percent of people born in the same state as currently living (numeric - decimal)
96	PctSameHouse85	percent of people living in the same house as in 1985 (5 years before) (numeric - decimal)
97	PctSameCity85	percent of people living in the same city as in 1985 (5 years before) (numeric - decimal)
98	PctSameState85	percent of people living in the same state as in 1985 (5 years before) (numeric - decimal)
99	LemasSwornFT	number of sworn full time police officers (numeric - expected to be integer)
100	LemasSwFTPerPop	sworn full time police officers per 100K population (numeric - decimal)
101	LemasSwFTFieldOps	number of sworn full time police officers in field operations (on the street as opposed to administrative etc) (numeric - expected to be integer)
102	LemasSwFTFieldPerPop	sworn full time police officers in field operations (on the street as opposed to administrative etc) per 100K population (numeric - decimal)
103	LemasTotalReq	total requests for police (numeric - expected to be integer)
104	LemasTotReqPerPop	total requests for police per 100K population (numeric - decimal)
105	PolicReqPerOffic	total requests for police per police officer (numeric - decimal)
106	PolicPerPop	police officers per 100K population (numeric - decimal)
107	RacialMatchCommPol	a measure of the racial match between the community and the police force. High values indicate proportions in community and police force are similar (numeric - decimal)
108	PctPolicWhite	percent of police that are Caucasian (numeric - decimal)
109	PctPolicBlack	percent of police that are African American (numeric - decimal)
110	PctPolicHisp	percent of police that are Hispanic (numeric - decimal)
111	PctPolicAsian	percent of police that are Asian (numeric - decimal)
112	PctPolicMinor	percent of police that are minority of any kind (numeric - decimal)
113	OfficAssgnDrugUnits	number of officers assigned to special drug units (numeric - expected to be integer)
114	NumKindsDrugsSeiz	number of different kinds of drugs seized (numeric - expected to be integer)
115	PolicAveOTWorked	police average overtime worked (numeric - decimal)
116	LandArea	land area in square miles (numeric - decimal)
117	PopDens	population density in persons per square mile (numeric - decimal)
118	PctUsePubTrans	percent of people using public transit for commuting (numeric - decimal)
119	PolicCars	number of police cars (numeric - expected to be integer)
120	PolicOperBudg	police operating budget (numeric - may be integer)
121	LemasPctPolicOnPatr	percent of sworn full time police officers on patrol (numeric - decimal)
122	LemasGangUnitDeploy	gang unit deployed (numeric - integer - but really nominal - 0 means NO, 10 means YES, 5 means Part Time)

123	LemasPctOfficDrugUn	percent of officers assigned to drug units (numeric - decimal)
124	PolicBudgPerPop	police operating budget per population (numeric - decimal)
1	murders	number of murders in 1995 (numeric - expected to be integer) (to be predicted)
2	murdPerPop	number of murders per 100K population (numeric - decimal) (to be predicted)
3	rapes	number of rapes in 1995 (numeric - expected to be integer) (to be predicted)
4	rapesPerPop	number of rapes per 100K population (numeric - decimal) (to be predicted)
5	robberies	number of robberies in 1995 (numeric - expected to be integer) (to be predicted)
6	robbbPerPop	number of robberies per 100K population (numeric - decimal) (to be predicted)
7	assaults	number of assaults in 1995 (numeric - expected to be integer) (to be predicted)
8	assaultPerPop	number of assaults per 100K population (numeric - decimal) (to be predicted)
9	burglaries	number of burglaries in 1995 (numeric - expected to be integer) (to be predicted)
10	burglPerPop	number of burglaries per 100K population (numeric - decimal) (to be predicted)
11	larcenies	number of larcenies in 1995 (numeric - expected to be integer) (to be predicted)
12	larcPerPop	number of larcenies per 100K population (numeric - decimal) (to be predicted)
13	autoTheft	number of auto thefts in 1995 (numeric - expected to be integer) (to be predicted)
14	autoTheftPerPop	number of auto thefts per 100K population (numeric - decimal) (to be predicted)
15	arsons	number of arsons in 1995 (numeric - expected to be integer) (to be predicted)
16	arsonsPerPop	number of arsons per 100K population (numeric - decimal) (to be predicted)
17	ViolentCrimesPerPop	total number of violent crimes per 100K population (numeric - decimal) (to be predicted)
18	nonViolPerPop	total number of non-violent crimes per 100K population (numeric - decimal) (to be predicted)

Question:

1. A complete and concise data exploration exercise in terms of crime statistics, patterns, relations between different times of crimes vis-à-vis socio economic profiles/ demographics etc.
2. Relevant data preprocessing exercise whatever is needed
3. How can we build a predictive model against different type of crimes?
4. What are the relevant variables for this exercise? How do we identify them?
5. How do we validate the accuracy of model?
6. How can we increase the model accuracy?

Expected Outcome:

1. Executable Python / R code
2. A read out, explaining what you have done and why you have done. What is your conclusion?