Contents

1 Introduction **2**

2 Detailed description of the data **2**

3 Visualisation **2**

4 Conclusion **10**

5 Appendix **10**

1 Introduction

The data combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS (Law Enforcement Management And Administrative Statistics) survey, and crime data from the 1995 FBI UCR (Uniform Crime Reporting). The main problem of interest intended with the data is the prediction of crimes. The data was obtained from http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized. We also looked at a normalized version of this dataset from http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime that has been normalized with a unsupervised, equal-interval binning method. The data has previously been used to predict violent crimes per population. The primary machine learning task is regression, as the data is a collection of attributes that are thought to be related to the number of crimes. We would like to predict the autotheft attribute with a regression model using the data as training set. As a classification problem we could predict whether new data objects belong to a rich/poor community or a colored/white community. We would also like to find groups of communities that are similar to eachother based on socio-economic attributes. We could also detect communities with deviating from the normal of communities with anomaly detection.

2 Detailed description of the data

There was apparently a limitation to this dataset, which was that the LEMAS survey was of the police departments with at least 100 officers, plus a random sample of smaller departments, so many communities are missing values in police related attributes, as we can also see in the summary statistic in the appendix. There was also 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. Our data consists of 147 attributes, most of which are discrete or continous ratios. For example there is a population attribute for each community, and a householdsize attribute which is the mean number of people per household in the community. There are also nominal attributes, one with the names of the community and the state's name. LemanGangUnitDeploy is a nominal discrete attribute where 0 means no gangunit deployed and 1 means there is a gangunit, and 0.5 means a part-time gang unit is deployed. Many of the attributes have missing values which are represented by a questionmark in the dataset. These we will try to deal with in various ways. We will try to estimate the mean, delete the data objects and delete the attributes with the missing values. The statistical summary also shows that it seems pctblack and pctwhite have high correlation with crimes per population as well as the attributes concerning families with two parents.

3 Visualisation

Based on a principal component analysis of the data (excluding the first columns including only identifier information like state names) there appear to be some

potential problems with outliers. In all PCA illustrations, the colors represent amount of auto theft per population, black being no auto theft and white being the maximum number of thefts per population.

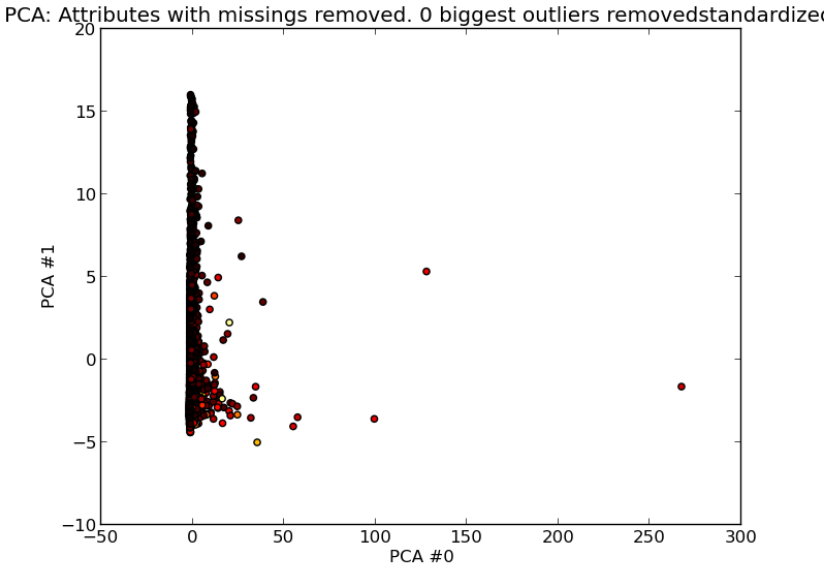

Figure 1: PCA, where attributes with missing values have been removed removed.

We are not yet sure how best to deal with these, but in order to explore our data we also performed principal componenet analysis on our data set with some of the biggest outliers removed.

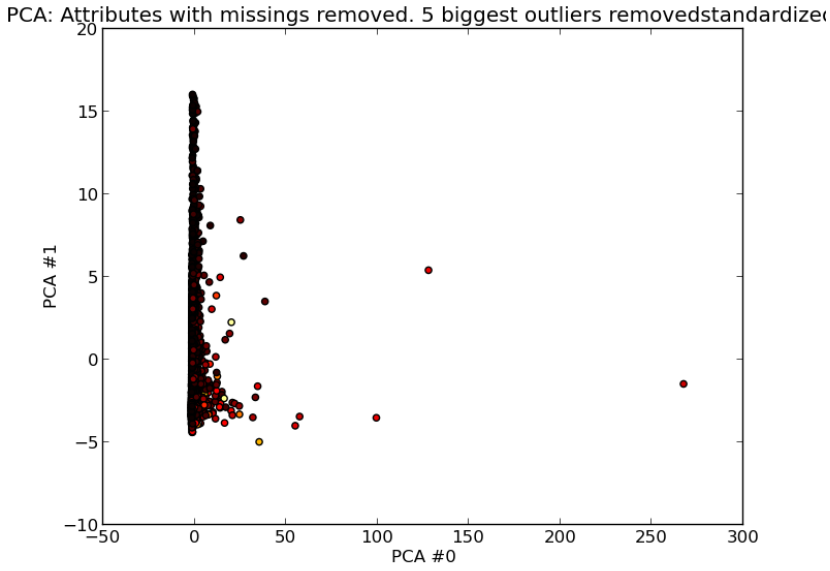

Figure 2: PCA, where attributes with missing values have been removed removed. Also the 5 biggest outliers from before have been removed prior to PCA.

The above pictures illustrate our results, when we dealt with missing values by removing the corresponding attribute completely. We also performed PCA after removing data objects with missing values instead, see below

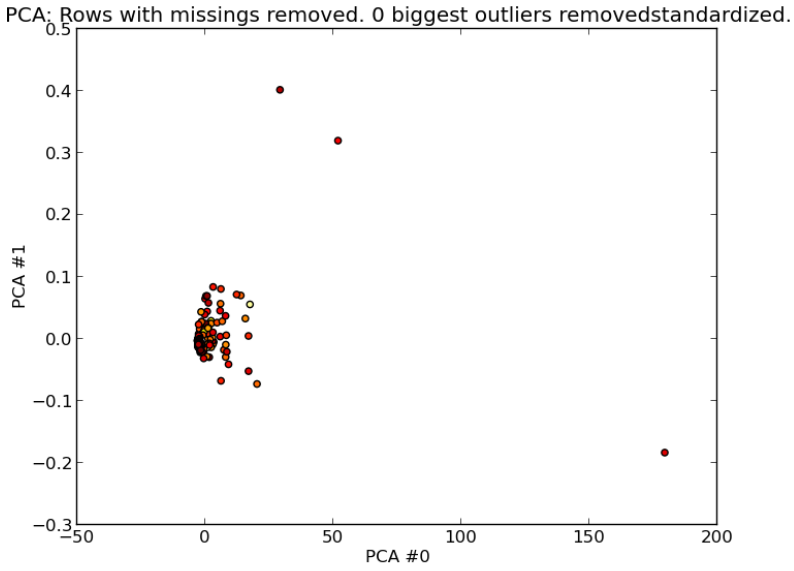

Figure 3: PCA, where data objects with missing values have been removed removed.

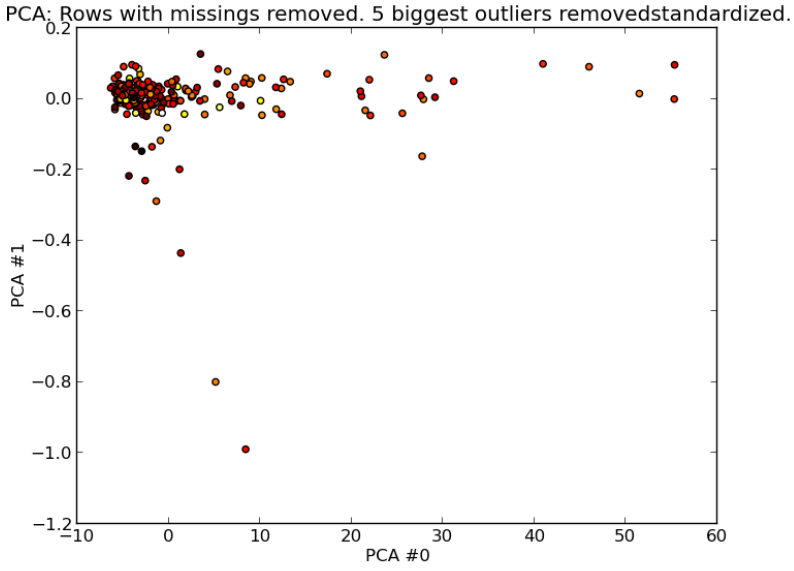

Figure 4: PCA, where data objects with missing values have been removed removed. Also the 5 biggest outliers from before have been removed prior to PCA.

We also did PCA where we simply put the means along attributes instead

of missing values, producing the following principal components.

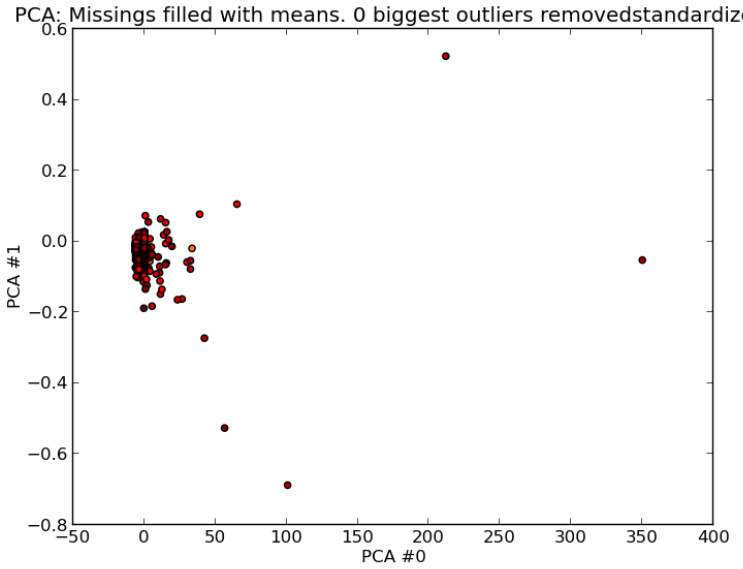

Figure 5: PCA, where missing values have been replaced with means.

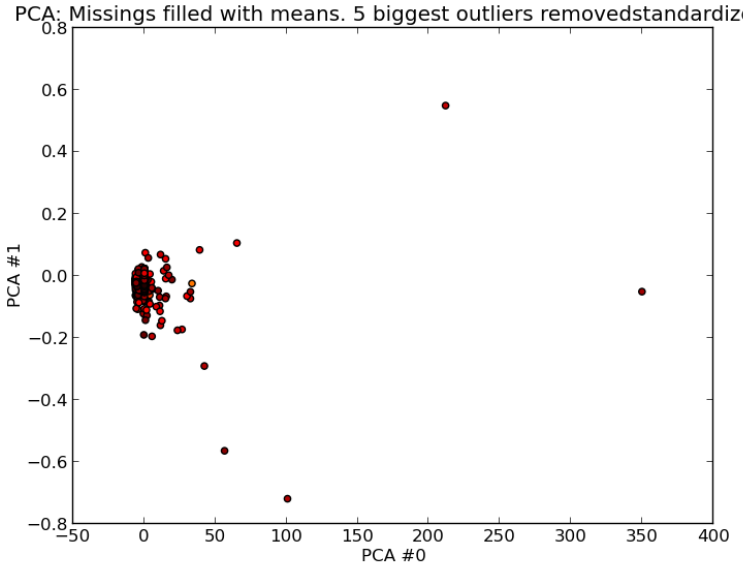

Figure 6: PCA, where missing values have been replaced with means. Also the 5 biggest outliers from before have been removed prior to PCA.

Investigating normalized data, leads us to believe that considerably better

preprocessing can be done to the data. As an example, we have performed principal component analysis also on the normalized data.

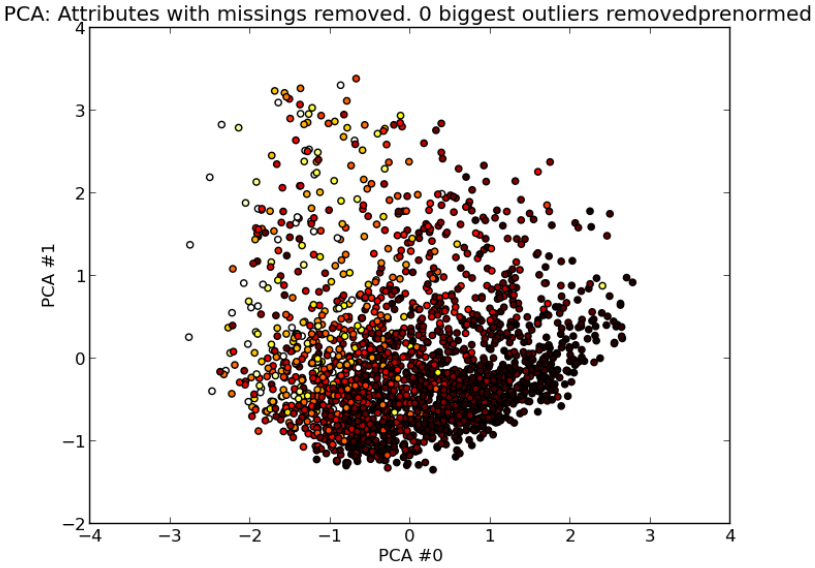

Figure 7: Using normalized data from other source, attributes with missing values are removed.

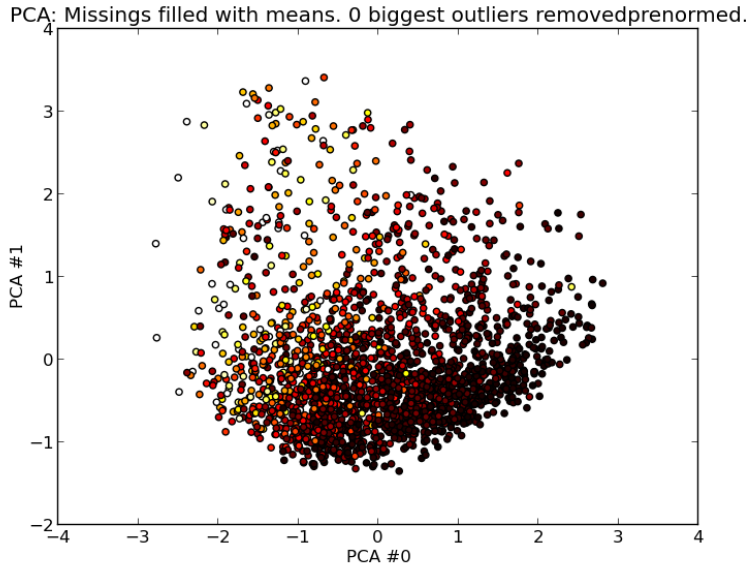

Figure 8: Using normalized data from other source, missing values replaced with means.

The above data also appears to be normally distributed, in contrast to our own PCA.

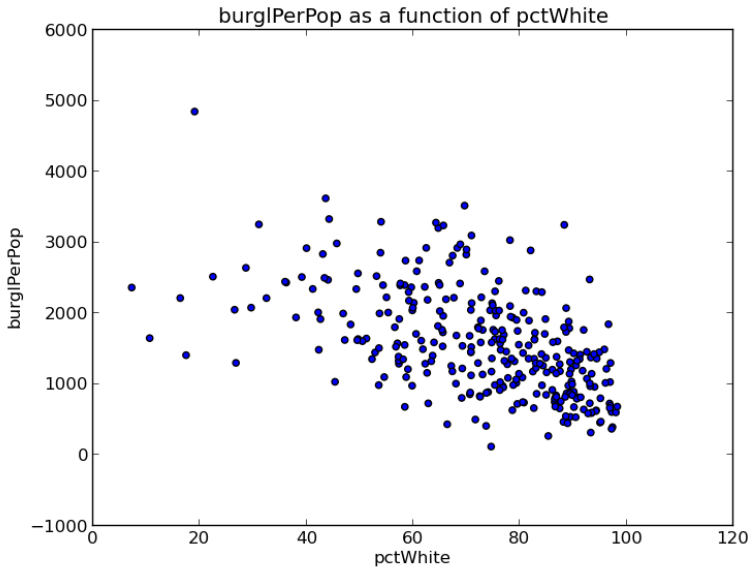

Figure 9: Burgleries per population (1K people) as function of percentage of white population.

Many of our attributes have little correlation, as seen above, which is why regression is an obvious machine learning task. Looking at the normalized data, furthermore leads us to believe that the task is feasable.

There are, however, some obvious attributes with high correlation, like population and auto theft, as seen below.

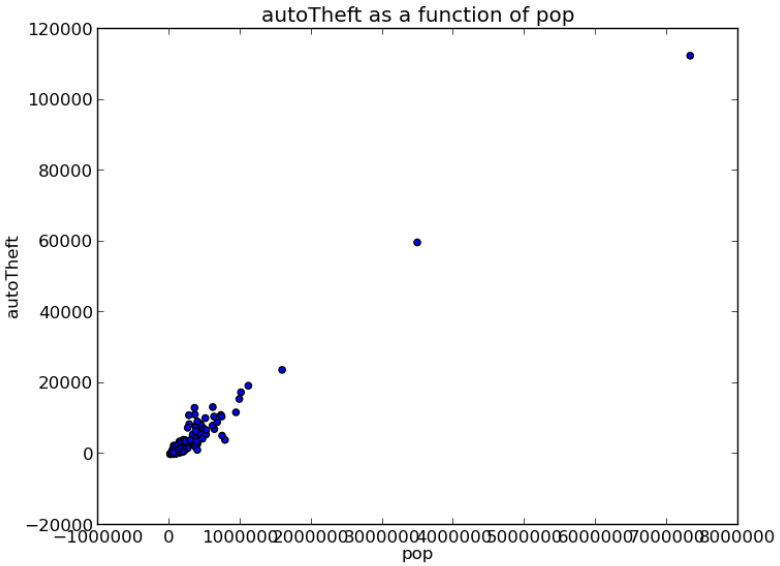

Figure 10: Auto theft as function of population.

The variance explained is shown below for our data.

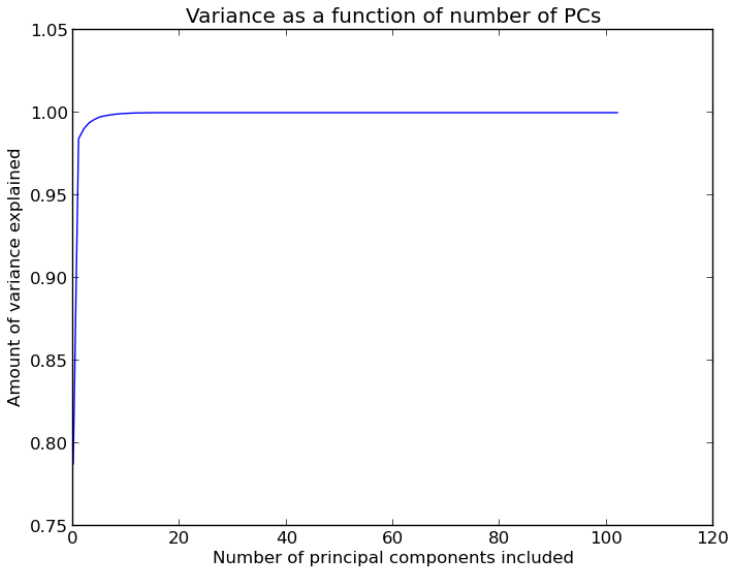

Figure 11: Variance explained as a function of number of principal components

The first component explains 99.9 percent of the variance, which is really high. We think this has to do with us including the various crime attributes in the PCA. I.e number of burglaries per population will have very high correlation

with number of violent crimes. As well as the number of nonviolent crimes. We think this may be why our data looks so different to the normalized data, as all the crime attributes are removed there except the violent crimes per population. Here is the variance explained for the normalisedzdata:

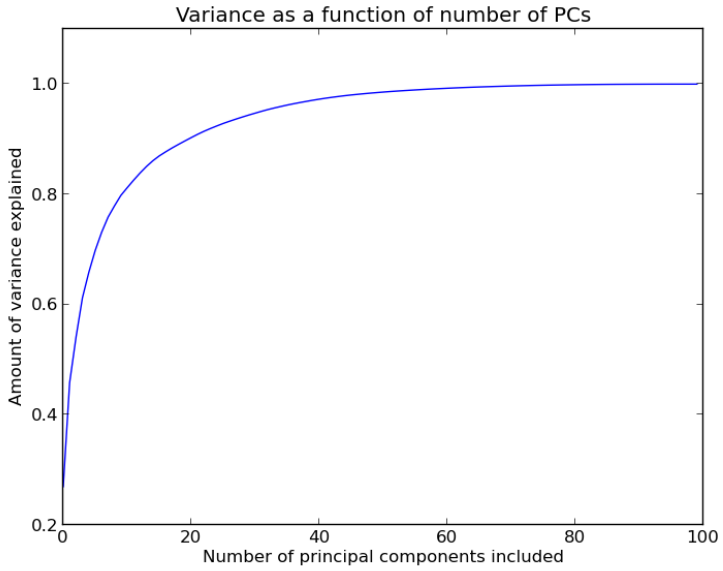

Figure 12: Variance explained as a function of number of principal components on the normalized data

4 Conclusion

From looking at the normalized data, we have learned that better or additional preprocessing should be possible. We think the handling of outliers with this dataset has much impact on the outcome of the analysis. We think that performing regression on the data to predict auto theft should be very much possible when looking at the normalized data.

5 Appendix

Variable	Minimum	Maximum	Mean	Standard Deviation	Correlation w/ ViolPerPop	Median	Mode	Missing
pop	10005.00	7322564.00	53117.98	204620.25	0.21	22792.00	12361.00	0.00
perHoush	1.60	5.28	2.71	0.33	-0.02	2.66	2.60	0.00
pctBlack	0.00	96.67	9.34	14.25	0.63	2.87	0.24	0.00
pctWhite	2.68	99.63	83.98	16.42	-0.68	90.35	98.04	0.00
pctAsian	0.03	57.46	2.67	4.47	0.03	1.23	0.41	0.00
pctHisp	0.12	95.29	7.95	14.59	0.25	2.18	0.78	0.00
pct12-21	4.58	54.40	14.45	4.52	0.02	13.62	13.62	0.00
pct12-29	9.38	70.51	27.64	6.18	0.11	26.78	26.78	0.00
pct16-24	4.64	63.62	13.98	5.97	0.05	12.54	11.61	0.00
pct65up	1.66	52.77	11.84	4.78	0.06	11.73	11.06	0.00
persUrban	0.00	7322564.00	47734.72	205606.69	0.21	18041.00	0.00	0.00
pctUrban	0.00	100.00	70.47	44.08	0.09	100.00	100.00	0.00
medIncome	8866.00	123625.00	33984.70	13424.68	-0.40	31441.00	27095.00	0.00
pctWwage	31.68	96.76	78.31	7.95	-0.29	78.61	85.12	0.00
pctWfarm	0.00	6.53	0.88	0.69	-0.15	0.69	0.54	0.00
pctWdiv	5.81	89.04	43.75	12.79	-0.56	42.88	41.65	0.00
pctWsocsec	4.81	76.39	26.41	8.30	0.11	26.59	21.51	0.00
pctPubAsst	0.18	44.82	6.80	4.70	0.56	5.61	2.27	0.00
pctRetire	3.46	45.51	15.97	4.62	-0.10	15.65	13.14	0.00
medFamIncome	10447.00	139008.00	39857.06	14251.21	-0.41	36678.00	30546.00	0.00
perCapInc	5237.00	63302.00	15603.52	6281.56	-0.32	14101.00	11252.00	0.00
whitePerCap	5472.00	68850.00	16567.70	6346.84	-0.19	15073.00	12735.00	0.00
blackPerCap	0.00	212120.00	11541.75	9232.10	-0.21	9777.00	0.00	0.00
NAperCap	0.00	480000.00	12229.19	14853.84	-0.06	9895.00	0.00	0.00
asianPerCap	0.00	106165.00	14227.99	9881.27	-0.13	12250.00	0.00	0.00
otherPerCap	0.00	137000.00	9442.77	7926.47	-0.10	8186.00	0.00	1.00
hispPerCap	0.00	54648.00	11019.00	5884.06	-0.22	9721.00	0.00	0.00
persPoverty	78.00	1384994.00	7590.85	39361.46	0.24	2142.00	470.00	0.00
pctPoverty	0.64	58.00	11.62	8.60	0.51	9.33	3.26	0.00
pctLowEdu	0.20	49.89	9.19	6.67	0.37	7.74	5.78	0.00
pctNotHSgrad	1.46	73.66	22.31	10.99	0.47	21.38	11.27	0.00
pctCollGrad	1.63	79.18	23.06	12.69	-0.30	19.65	14.20	0.00
pctUnemploy	1.32	31.23	6.05	2.90	0.48	5.45	4.36	0.00
pctEmploy	24.82	84.67	62.02	8.31	-0.32	62.44	62.60	0.00
pctEmployMfg	2.05	50.03	18.23	8.10	-0.05	17.30	25.38	0.00
pctEmployProfServ	8.69	62.67	24.53	6.66	-0.06	23.39	21.52	0.00
pctOccupManu	1.37	44.27	13.82	6.43	0.28	13.15	16.52	0.00
pctOccupMgmt	6.48	64.97	28.21	9.33	-0.32	26.24	28.31	0.00
pctMaleDivorc	2.13	20.08	9.13	2.80	0.51	9.15	10.82	0.00
pctMaleNevMar	12.06	76.60	30.68	8.13	0.27	29.00	26.78	0.00
pctFemDivorc	3.35	23.92	12.33	3.26	0.54	12.52	14.36	0.00
pctAllDivorc	2.83	22.23	10.81	3.00	0.54	10.90	11.77	0.00
persPerFam	2.29	4.64	3.13	0.24	0.15	3.10	3.13	0.00
pct2Par	22.97	93.60	74.06	10.53	-0.70	75.03	72.16	0.00
pctKids2Par	18.30	92.58	71.23	12.05	-0.73	72.53	63.25	0.00
pctKids-4w2Par	8.70	100.00	81.87	12.26	-0.66	83.99	100.00	0.00
pct12-17w2Par	20.20	97.34	75.52	10.37	-0.66	76.92	77.49	0.00
pctWorkMom-6	24.42	87.97	60.54	8.01	-0.02	60.71	63.48	0.00
pctWorkMom-18	41.95	89.37	68.85	6.68	-0.15	69.23	65.64	0.00
kidsBornNevrMarr	0.00	527557.00	2141.42	14692.58	0.24	352.00	139.00	0.00
pctKidsBornNevrMarr	0.00	27.35	3.12	3.13	0.74	2.04	0.97	0.00
numForeignBorn	20.00	2082931.00	6277.27	55419.65	0.14	1024.00	147.00	0.00
pctFgnImmig-3	0.00	64.29	13.53	9.78	0.16	12.26	0.00	0.00
pctFgnImmig-5	0.00	76.16	20.42	12.41	0.20	19.08	0.00	0.00
pctFgnImmig-8	0.00	80.81	27.54	14.37	0.24	26.72	0.00	0.00
pctFgnImmig-10	0.00	88.00	34.73	16.33	0.28	34.79	0.00	0.00
pctImmig-3	0.00	13.71	1.10	1.60	0.22	0.50	0.00	0.00
pctImmig-5	0.00	19.93	1.70	2.46	0.23	0.75	0.00	0.00
pctImmig-8	0.00	25.34	2.31	3.29	0.24	1.04	0.00	0.00
pctImmig-10	0.00	32.63	2.94	4.25	0.25	1.31	0.00	0.00
pctSpeakOnlyEng	6.15	98.98	87.07	14.08	-0.22	92.18	93.57	0.00
pctNotSpeakEng	0.00	38.33	2.41	4.21	0.27	0.92	0.44	0.00
pctLargHousFam	0.96	34.87	5.39	3.79	0.34	4.28	3.71	0.00
pctLargHous	0.44	30.87	3.92	3.18	0.26	3.05	2.98	0.00
persPerOccupHous	1.58	4.52	2.62	0.32	-0.02	2.57	2.44	0.00
persPerOwnOccup	1.61	4.48	2.74	0.30	-0.10	2.71	2.65	0.00
persPerRenterOccup	1.55	4.73	2.37	0.39	0.24	2.29	2.17	0.00
pctPersOwnOccup	13.93	97.24	66.37	14.18	-0.51	65.91	63.79	0.00
pctPopDenseHous	0.05	59.49	4.13	5.60	0.40	2.34	1.31	0.00
pctSmallHousUnits	3.06	95.34	45.41	13.78	0.45	46.39	53.15	0.00

medNumBedrm	1.00	4.00	2.64	0.51	-0.35	3.00	3.00	0.00
houseVacant	36.00	172768.00	1748.37	6503.87	0.29	558.00	246.00	0.00
pctHousOccup	37.47	99.00	92.93	5.04	-0.26	94.21	95.38	0.00
pctHousOwnerOccup	16.86	96.49	63.37	13.97	-0.46	62.83	56.17	0.00
pctVacantBoarded	0.00	39.89	2.78	3.59	0.48	1.66	0.00	0.00
pctVacant6up	3.12	82.13	34.77	13.91	0.03	34.10	37.50	0.00
medYrHousBuilt	1939.00	1987.00	1962.62	11.17	-0.11	1964.00	1939.00	0.00
pctHousWOphone	0.00	23.88	4.29	4.09	0.47	2.85	0.00	0.00
pctHousWOplumb	0.00	5.33	0.43	0.43	0.31	0.32	0.00	0.00
ownHousLowQ	14999.00	500001.00	88695.80	66670.78	-0.19	65500.00	34000.00	0.00
ownHousMed	19500.00	500001.00	113097.52	81906.36	-0.18	82800.00	500001.00	0.00
ownHousUperQ	28200.00	500001.00	145318.26	99030.91	-0.17	106700.00	500001.00	0.00
ownHousQrange	0.00	331000.00	56622.46	39106.50	-0.09	43400.00	28100.00	0.00
rentLowQ	99.00	1001.00	329.97	144.14	-0.25	307.00	252.00	0.00
rentMed	120.00	1001.00	428.54	170.71	-0.23	397.00	316.00	0.00
rentUpperQ	182.00	1001.00	527.25	199.29	-0.22	486.00	1001.00	0.00
rentQrange	0.00	803.00	197.29	85.21	-0.11	171.00	139.00	0.00
medGrossRent	192.00	1001.00	501.47	169.27	-0.23	467.00	1001.00	0.00
medRentpctHousInc	14.90	35.10	26.30	2.98	0.32	26.10	24.70	0.00
medOwnCostpct	14.00	32.70	20.99	2.99	0.06	21.00	22.60	0.00
medOwnCostPctWO	10.10	23.40	13.01	1.42	0.06	12.80	11.80	0.00
persEmergShelt	0.00	23383.00	66.95	564.25	0.19	0.00	0.00	0.00
persHomeless	0.00	10447.00	17.82	245.45	0.14	0.00	0.00	0.00
pctForeignBorn	0.18	60.40	7.34	8.42	0.19	4.31	2.97	0.00
pctBornStateResid	6.75	93.14	61.54	16.75	-0.07	64.49	74.45	0.00
pctSameHouse-5	11.83	78.56	51.54	10.52	-0.14	52.17	54.85	0.00
pctSameCounty-5	27.95	96.59	77.41	10.88	0.08	79.49	81.47	0.00
pctSameState-5	32.83	99.90	88.11	7.29	-0.01	90.03	92.69	0.00
numPolice	65.00	25655.00	499.20	1681.47	0.19	173.00	100.00	1872.00
policePerPop	29.40	3437.23	246.49	273.80	0.07	196.01	#N/A	1872.00
policeField	14.00	22496.00	432.56	1493.71	0.19	152.00	94.00	1872.00
policeFieldPerPop	19.21	3290.62	210.84	235.48	0.07	170.27	183.22	1872.00
policeCalls	2100.00	8328470.00	252404.99	689449.78	0.23	90000.00	50000.00	1872.00
policCallPerPop	2704.80	1926281.50	120651.72	148211.34	0.15	91034.60	#N/A	1872.00
policCallPerOffic	20.80	2162.50	523.66	307.84	0.15	443.20	422.60	1872.00
policePerPop2	29.40	3437.20	246.49	273.80	0.07	196.00	171.50	1872.00
racialMatch	42.15	100.00	85.50	10.94	-0.47	87.93	100.00	1872.00
pctPolicWhite	1.60	100.00	82.52	15.33	-0.39	86.18	100.00	1872.00
pctPolicBlack	0.00	67.31	9.26	11.02	0.51	5.00	0.00	1872.00
pctPolicHisp	0.00	98.40	5.46	10.60	0.06	2.04	0.00	1872.00
pctPolicAsian	0.00	18.57	0.68	1.71	0.07	0.00	0.00	1872.00
pctPolicMinority	0.00	98.40	15.24	14.83	0.42	11.37	0.00	1872.00
officDrugUnits	0.00	1773.00	26.29	100.82	0.17	12.00	6.00	1872.00
numDiffDrugsSeiz	1.00	15.00	8.82	2.84	0.12	9.00	9.00	1872.00
policAveOT	0.00	634.70	119.11	92.50	0.01	98.70	0.00	1872.00
landArea	0.90	3569.80	27.42	109.82	0.08	13.70	4.90	0.00
popDensity	10.00	44229.90	2783.84	2828.99	0.26	2027.30	3217.70	0.00
pctUsePubTrans	0.00	54.33	3.04	4.91	0.19	1.22	0.00	0.00
policCarsAvail	20.00	3187.00	185.48	318.54	0.31	86.00	55.00	1872.00
policOperBudget	2380215.00	1617293056.00	32176019.34	110456627.50	0.20	11164110.00	8000000.00	1872.00
pctPolicPatrol	10.85	99.94	87.13	10.35	-0.09	89.58	93.07	1872.00
gangUnit	0.00	10.00	4.29	4.06	0.11	5.00	0.00	1872.00
pctOfficDrugUnit	0.00	48.44	0.98	2.88	0.32	0.00	0.00	0.00
policBudgetPerPop	15260.40	2422367.00	153577.87	203040.89	0.06	114582.00	#N/A	1872.00
murders	0.00	1946.00	7.76	58.17	0.25	1.00	0.00	0.00
murdPerPop	0.00	91.09	5.86	9.16	0.67	2.17	0.00	0.00
rapes	0.00	2818.00	28.05	105.62	0.34	7.00	0.00	208.00
rapesPerPop	0.00	401.35	36.26	34.24	0.58	26.92	0.00	208.00
robberies	0.00	86001.00	237.95	2250.72	0.21	19.00	1.00	1.00
robbbPerPop	0.00	2264.13	162.61	234.49	0.83	74.80	0.00	1.00
assaults	0.00	62778.00	326.53	1987.95	0.30	56.00	12.00	13.00
assaultPerPop	0.00	4932.50	378.00	438.24	0.95	226.53	0.00	13.00
burglaries	2.00	99207.00	761.24	3111.70	0.32	205.00	79.00	3.00
burglPerPop	16.92	11881.02	1033.43	763.35	0.70	822.72	728.93	3.00
larcenies	10.00	235132.00	2137.63	7600.57	0.30	747.00	547.00	3.00
larcPerPop	77.86	25910.55	3372.98	1901.32	0.51	3079.51	4631.10	3.00
autoTheft	1.00	112464.00	516.69	3258.16	0.24	75.00	16.00	3.00
autoTheftPerPop	6.55	4968.59	473.97	504.67	0.64	302.36	213.62	3.00
arsons	0.00	5119.00	30.91	180.13	0.23	5.00	0.00	91.00
arsonsPerPop	0.00	436.37	32.15	39.24	0.42	21.08	0.00	91.00
violentPerPop	0.00	4877.06	589.08	614.78	1.00	374.06	223.06	221.00
nonViolPerPop	116.79	27119.76	4908.24	2739.71	0.68	4425.45	4295.96	97.00