MSDS 6372 Project 2

Communities and Crime PCA and PCR

Team members:

Brian Kruse, Christopher Boomhower, Andrew Abbott, Johnny Quick

Date: 11/12/2016

Introduction:

Throughout time, there has been a desire to determine factors in a community that produce a proclivity toward higher violent crime rates. One of the primary benefits of knowing these factors in advance is that it allows for more efficient allocation of law enforcement resources. Another benefit of knowing these factors is that it can help dictate policy that will hopefully aid in reducing the violent crime rate. As an example, if it is determined that there is a correlation between the violent crime rate and the percentage of houses that are vacant and boarded (or vacant longer than 6 months), then policy can be proposed to alleviate or reduce the number of vacant houses.

For this study, we will utilize data from the 1990 US Census Bureau and the 1995 FBI Crime Report. There is a vast array of variables available in this dataset to evaluate a community's makeup. Some of these variables include percentage of various races in the community, percentage of immigrants and their length of time in the community, household size, median income, percentage of people unemployed, percentage of houses vacant and boarded, and many others which will be described shortly.

Specifically, we will use this dataset to look at the violent crime rate per population and what factors contributed to the violent crime rate per population. We have two main objectives in our study: 1) to reduce the number of variables to a more manageable number of variables to define a model for multiple linear regression, and 2) to reduce the number of variables to define a model for predicting the violent crime rate using principal component regression (PCR). To accomplish this, we will apply principal component analysis (PCA) to determine which variables may have the most influence on the violent crime rate per population, which will reduce the number of variables utilized for prediction purposes. After determining which variables may have the most influence on the violent crime rate per population, PCR is applied to reduce the number of variables to predict the violent crime rate, and then also to determine the most efficient variables to predict the violent crime rate per population using multiple linear regression. Although in this study we will not be performing multiple linear regression using these identified variables, this research will allow our findings to be utilized in a future multiple linear regression model to predict the violent crime rate.

Descriptive Statistics:

This Communities and Crime dataset is found in the UCI Machine Learning Repository. A related dataset was used in a 2002 paper by Redmond and Baveja. This dataset initially includes 128 variables including 122 with possible correlation to crime, 5 that are not predictive and one response variable. Because we are doing a PCA and PCR, to prevent all the variance in the response variable from being explained by the variable with the largest variance, all numeric data has already been standardized into the range 0.00 – 1.00. Standardization preserves attribute distributions and any skewness.

Before continuing with the PCA and PCR analysis the data still needs to be cleaned up. As a first step in cleaning the data, we remove the following categorical data and police data which is found to be 84% missing (See Table 1). The categorical variables removed all relation to geographic location or name of place. Our goal is to compare general communities without regard to specific communities.

Table 1. Categorical and Sparse Police Variables

state	county	community	communityname
fold	LemasSwornFT	LemasSwFTPerPop	LemasSwFTFieldOps
LemasSwFTFieldPerPop	LemasTotalReq	LemasTotReqPerPop	PolicReqPerOffic
PolicPerPop	RacialMatchCommPol	PctPolicWhite	PctPolicBlack
PctPolicHisp	PctPolicAsian	PctPolicMinor	OfficAssgnDrugUnits
NumKindsDrugsSeiz	PolicAveOTWorked	PolicCars	PolicOperBudg
LemasPctPolicOnPatr	LemasGangUnitDeploy	LemasPctOfficDrugUn	PolicBudgPerPop

Next we remove observation 131 which has a population value of zero. There are eleven variables that we determine to be redundant. These are variables that represent the same characteristic in different forms, such as number of people living in areas classified as urban (*numbUrban*) and percentage of people living in areas classified as urban (*Pcturban*) (See Table 2).

Table 2. Redundancy Among Variables: Removed vs. Retained Variables

Removed	numbUrban	NumUnderPov	NumIlleg	HousVacant	NumImmig	
Retained	Pcturban	PctPopUnderPov	PctIlleg	PctHousOccup	Pctimmig	
Removed	OwnOccLowQuart	OwnOccMedVal	OwnOccHiQuart	RentLowQ	RentMedian	RentHighQ
Retained	PctHousOwnOcc	PctHousOwnOcc	PctHousOwnOcc		MedRentPctHousInc	

The new, reduced dataset contains 88 explanatory variables and one response variable displayed with their distributions in the Appendix, Figure 4. It is clear from their distributions that many of the variables are not normally distributed as assumed in PCA analysis. The non-normality of variables will be addressed in the analysis section below.

Again, the goal of this study is to find out which of the community characteristics are most predictive of the number of violent crimes per 100k population.

Analysis – Part I (Principal Components Analysis):

With the data set described and reasoning provided for manual variable termination complete, it is almost time to proceed with the Principal Components Analysis. As indicated in the previous section, however, some variables, including the response variable – *ViolentCrimesPerPop*, portray non-normal distributions. Being that one of the primary assumptions of PCA is that the variables are normally distributed, these violations should be addressed before continuing with the analysis. With some exploratory effort, it was determined that aside from population data, which was best transformed via log transformation, all remaining violations were best transformed via square-root and cube-root transformation (See Appendix, Table 9). In cases where distributions were left-skewed, such as with *racePctWhite*, additional pre-conditioning was required before applying cube- or square-root transformation to make the distributions right-skewed, first by subtracting 1 from the data values and second, computing absolute value. The final transformed distributions are displayed in the Appendix, Figure 5.

Since the end-goal of this analysis is to be able to efficiently predict the amount of violent crimes per population for U.S. communities between 1990 and 1995, there is desire to not only generate a working prediction model but to test it as well. For this reason, half the observations within the data set are randomly sampled as training data and the remaining half are assigned as test data in preparation for PCR. This leaves 804 observations for training purposes and 804 observations for validating the model.

It is finally time to run a full PCA using explanatory variable training data. Since prior knowledge does not lend itself to judging whether standardization should be avoided, transforming the data de-standardized some variables, and the 88 explanatory variables are comprised of many different measures, the decision is made to standardize all data. When performing the PCA, each variable's mean is subtracted from each variable and a correlation matrix is computed. A correlation matrix summary is provided in the Appendix, *Table 10*, indicating moderate correlation among most of the community variables. Such relationships among our community attributes further solidify Principal Components Analysis as being a suitable approach to analyzing community data. Therefore, linear combinations of the data are created next, the number of which is equivalent to the number of variables present in the data set (88 in this case). Each of these linear combinations makes up a component and comprises an eigenvector. These eigenvectors are ordered by largest eigenvalues, representing the extent of variability explained by each vector. The vectors with largest eigenvalues make up the principal components of the analysis.

The full PCA produces eigenvectors with the eigenvalues and variance proportions displayed in Table 8 within the Appendix. This output indicates that, for example, the first principal component accounts for the most variance in the data, explaining 26.21% of the variance. According to the cumulative variance, the number of variables may be reduced from 88 to 10 if the desire is to explain 80% of variance. Though this table is helpful, the use of Scree and Cumulative Proportion of Variance plots will help more effectively select the appropriate number of principal components

As mentioned above, 80% of the variance in this data set is explained by 10 principal components. However, 80% explained may not necessarily be an appropriate target depending on the rate at which explained variance decreases in significance from one principal component to the next. By plotting the Scree Plot in Figure 1, it is easier to judge where this rate of decreasing explained variance occurs (Only first 20 principal components shown in Figure 1 to enhance granularity during analysis). Note the rate of change in explained variance among the first seven principal components – the change is rather steep through the seventh component. After the 1% drop between component seven and component eight, the rate of decreasing explained variance begins to somewhat flatten out, reducing to a 0.3% change or less. By now referring to the Cumulative Variance Plot in Figure 1 (once again, only first 20 principle components plotted), it may be seen that the cumulative variance arguably begins to plateau around the seventh principal component and that the first seven components together explain about 74% of variance in the data set. For this reason, seven principal components may be selected as being the most appropriate given the variables among these data.

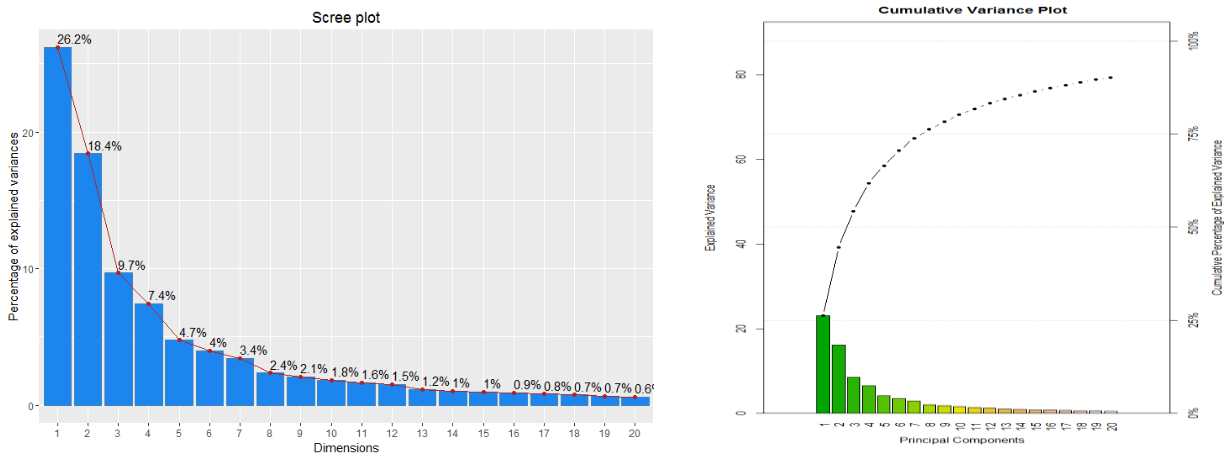


Figure 1. Scree Plot (left) and Cumulative Variance Plot (right) for First 20 PCs

Each principal component within the full PCA is comprised of variable loadings (similar to variable coefficients) for all 88 variables used in the analysis. These loadings provide insight into underlying factors among the variables and describe those demonstrating the most variance. Loadings for the first 7 principal components will be interpreted in detail in the *Loadings Interpretation* section following *Analysis – Part II*.

Analysis – Part II (Principal Components Regression):

The next step toward predicting the amount of community violent crimes is to perform the actual Principal Components Regression. This will facilitate insertion of the same principal components derived during PCA into a linear model for response prediction. Before applying the model, however, further analysis may be made around principal component selection via cross-validation.

Leave-one-out (LOO) cross-validation may be performed in conjunction with PCR to enhance principal component selection. Both Root-Mean-Square-Error Coefficient of Variance (CV) and PRESS statistics are calculated during LOO cross-validation, and their values are displayed in Table 3. The threshold for principal component selection when reviewing PRESS and CV values is the number of components at which the variance value drops to its lowest initial point and then begins to climb again. This takes place at fifteen principal components as highlighted in green in Table 3 (Yellow indicating values before and after fifteen components). In other words, LOO cross-validation identifies fifteen components as being appropriate for community violent crime predictions since this is the point of least variance before added variance is introduced again.

Table 3. LOO Cross-Validation Coefficient of Variance, PRESS, and R^2 Statistics for first 20 PCs

Response.Var	PC	PRESS	CV	R^2	R^2-adj
ViolentCrimesPerPop	1 comps	13.6431	0.1303	0.5416	0.5411
	2 comps	13.3674	0.1289	0.5522	0.5511
	3 comps	13.3661	0.1289	0.5544	0.5527
	4 comps	11.9618	0.122	0.6021	0.6001
	5 comps	11.31	0.1186	0.6243	0.6219
	6 comps	9.601	0.1093	0.6813	0.6789
	7 comps	9.5421	0.1089	0.6842	0.6814
	8 comps	9.5217	0.1088	0.6857	0.6825
	9 comps	9.503	0.1087	0.6872	0.6837
	10 comps	9.191	0.1069	0.6983	0.6945
	11 comps	8.9885	0.1057	0.7052	0.7012
	12 comps	8.9755	0.1057	0.7066	0.7022
	13 comps	8.9985	0.1058	0.7066	0.7018
	14 comps	8.8733	0.1051	0.7117	0.7066
	15 comps	8.8494	0.1049	0.7128	0.7073
	16 comps	8.8749	0.1051	0.713	0.7072
	17 comps	8.7461	0.1043	0.7178	0.7117
	18 comps	8.7777	0.1045	0.7178	0.7113
	19 comps	8.679	0.1039	0.7216	0.7148
	20 comps	8.6995	0.104	0.7216	0.7145

Recall the selected components in the previous section were the first seven principal components. Fifteen principal components are more than double the previous selection! While cross-validation has accounted for changes in variance, it has not accounted for decreasing rates of diminishing explained variance as is done while reviewing Scree and Cumulative Variance plots. Therefore, both seven and fifteen principal components will be considered going forward as valid options.

Being that predictions on community violent crimes will be made, it is helpful to note the Coefficient of Determination, R^2, values for both principal component model scenarios. These values are highlighted in purple within Table 3 for both seven and fifteen components. When accounting for the number of PCs included in each model, the adjusted-R^2 values are 0.6814 and 0.7073 for seven and fifteen PCs respectively. This, in combination with the other PC adjusted-R^2 values, clearly expresses that not much is gained by including fifteen components vs. seven components and that the rate of increasing R^2 values decreases substantially after seven PCs are included (See Figure 2), further validating this selection.

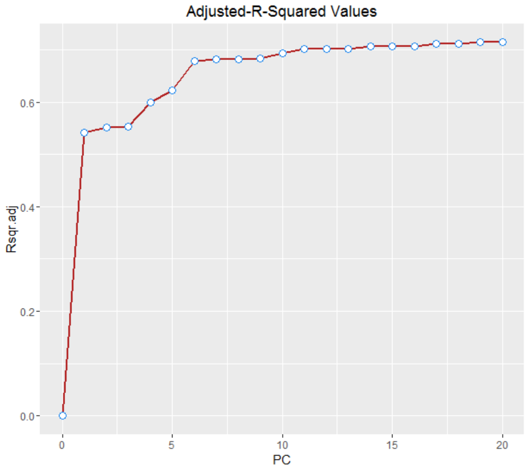


Figure 2. Adjusted-R^2 Values for First 20 PCs

With both models selected, it is now time to apply them to the test data sidelined previously. Doing so will further validate the strength of each model. The first step in doing so shall be to compare the predicted Root-Mean-Square-Error Coefficient of Variance against the CV obtained during LOO cross-validation using the training set. Both values are provided in Table 4 and their difference calculated. These results indicate that the difference in variance between the test and training data is less than 0.5% for fifteen PCs and less than 0.2% for seven PCs. Based on these differences, the models appear to be good fits.

Table 4. CV Comparison Between Training and Test Data Sets for 7 and 15 PCs

Response.Var	PC	Training CV	Test CV	CV Delta
ViolentCrimesPerPop	7 comps	0.1089	0.1108	0.0019
	15 comps	0.1049	0.1092	0.0043

The final step is to apply the models for prediction of community violent crime rates. The plots of Figure 3 display predicted *ViolentCrimesPerPop* values vs. measured values for every observation within the test data set, for both the seven PC model and fifteen PC model. Portraying predictions in this way facilitates a high-level comparison of model accuracy. Further interpretations follow.

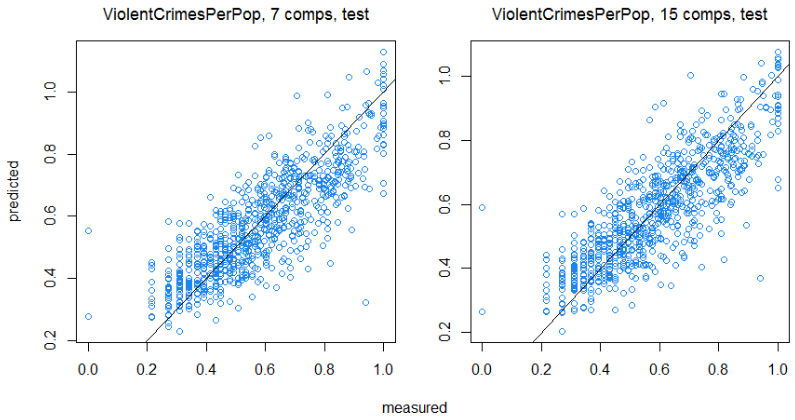
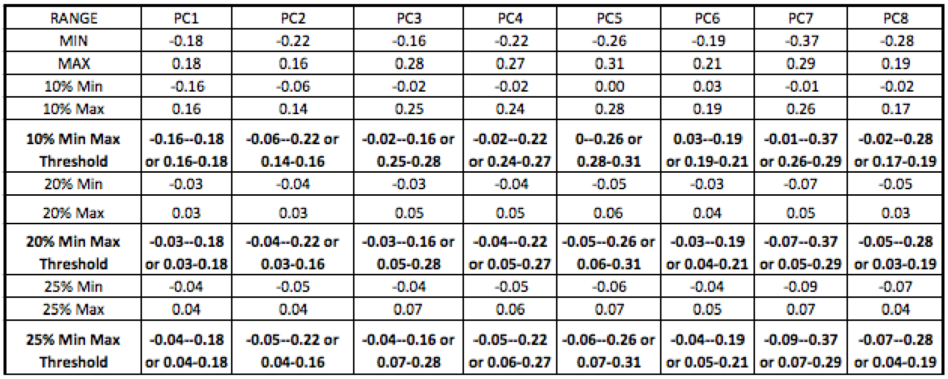


Figure 3. Predicted vs. Measured Response Values for the Test Data Set

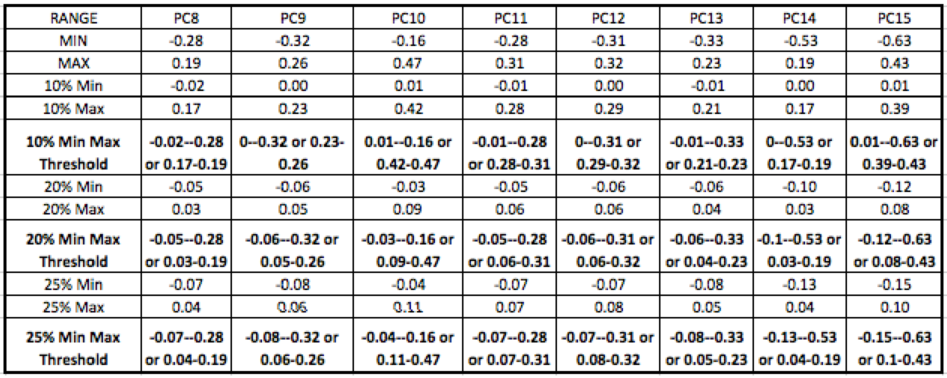
Loadings Interpretation:

As part of the decision-making process for identifying four thresholds (min & max, 10% of the min max, 20% of the min max, 25% of the min max loading values), we reviewed 88 variables across the fifteen principal components. Due to the wide variability in minimum and maximum values between principal components, we evaluated different threshold ranges (from minimum & maximum loading values) and thresholds of values that were within 10%, 20% and 25% of the maximum and minimum loading value. This evaluation was repeated across fifteen principal components and 88 variables as depicted in Table 6.

Table 5. Range of Correlated Values to Principal Components



RANGE	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
MIN	-0.18	-0.22	-0.16	-0.22	-0.26	-0.19	-0.37	-0.28
MAX	0.18	0.16	0.28	0.27	0.31	0.21	0.29	0.19
10% Min	-0.16	-0.06	-0.02	-0.02	0.00	0.03	-0.01	-0.02
10% Max	0.16	0.14	0.25	0.24	0.28	0.19	0.26	0.17
10% Min Max Threshold	**-0.16--0.18 or 0.16-0.18**	**-0.06--0.22 or 0.14-0.16**	**-0.02--0.16 or 0.25-0.28**	**-0.02--0.22 or 0.24-0.27**	**0--0.26 or 0.28-0.31**	**0.03--0.19 or 0.19-0.21**	**-0.01--0.37 or 0.26-0.29**	**-0.02--0.28 or 0.17-0.19**
20% Min	-0.03	-0.04	-0.03	-0.04	-0.05	-0.03	-0.07	-0.05
20% Max	0.03	0.03	0.05	0.05	0.06	0.04	0.05	0.03
20% Min Max Threshold	**-0.03--0.18 or 0.03-0.18**	**-0.04--0.22 or 0.03-0.16**	**-0.03--0.16 or 0.05-0.28**	**-0.04--0.22 or 0.05-0.27**	**-0.05--0.26 or 0.06-0.31**	**-0.03--0.19 or 0.04-0.21**	**-0.07--0.37 or 0.05-0.29**	**-0.05--0.28 or 0.03-0.19**
25% Min	-0.04	-0.05	-0.04	-0.05	-0.06	-0.04	-0.09	-0.07
25% Max	0.04	0.04	0.07	0.06	0.07	0.05	0.07	0.04
25% Min Max Threshold	**-0.04--0.18 or 0.04-0.18**	**-0.05--0.22 or 0.04-0.16**	**-0.04--0.16 or 0.07-0.28**	**-0.05--0.22 or 0.06-0.27**	**-0.06--0.26 or 0.07-0.31**	**-0.04--0.19 or 0.05-0.21**	**-0.09--0.37 or 0.07-0.29**	**-0.07--0.28 or 0.04-0.19**



RANGE	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15
MIN	-0.28	-0.32	-0.16	-0.28	-0.31	-0.33	-0.53	-0.63
MAX	0.19	0.26	0.47	0.31	0.32	0.23	0.19	0.43
10% Min	-0.02	0.00	0.01	-0.01	0.00	-0.01	0.00	0.01
10% Max	0.17	0.23	0.42	0.28	0.29	0.21	0.17	0.39
10% Min Max Threshold	**-0.02--0.28 or 0.17-0.19**	**0--0.32 or 0.23-0.26**	**0.01--0.16 or 0.42-0.47**	**-0.01--0.28 or 0.28-0.31**	**0--0.31 or 0.29-0.32**	**-0.01--0.33 or 0.21-0.23**	**0--0.53 or 0.17-0.19**	**0.01--0.63 or 0.39-0.43**
20% Min	-0.05	-0.06	-0.03	-0.05	-0.06	-0.06	-0.10	-0.12
20% Max	0.03	0.05	0.09	0.06	0.06	0.04	0.03	0.08
20% Min Max Threshold	**-0.05--0.28 or 0.03-0.19**	**-0.06--0.32 or 0.05-0.26**	**-0.03--0.16 or 0.09-0.47**	**-0.05--0.28 or 0.06-0.31**	**-0.06--0.31 or 0.06-0.32**	**-0.06--0.33 or 0.04-0.23**	**-0.1--0.53 or 0.03-0.19**	**-0.12--0.63 or 0.08-0.43**
25% Min	-0.07	-0.08	-0.04	-0.07	-0.07	-0.08	-0.13	-0.15
25% Max	0.04	0.06	0.11	0.07	0.08	0.05	0.04	0.10
25% Min Max Threshold	**-0.07--0.28 or 0.04-0.19**	**-0.08--0.32 or 0.06-0.26**	**-0.04--0.16 or 0.11-0.47**	**-0.07--0.28 or 0.07-0.31**	**-0.07--0.31 or 0.08-0.32**	**-0.08--0.33 or 0.05-0.23**	**-0.13--0.53 or 0.04-0.19**	**-0.15--0.63 or 0.1-0.43**

Before proceeding, our team evaluated whether to interpret 7 or 15 principal components. As part of this process, we counted all variables that would be removed if their loading values are not within a specified range of the minimum or maximum value for each of the 7 and 15 principal components. As we lower the tolerances for retention, based on the thresholds in Table 5, we identify how many variables are not associated with either 7 or 15 principal components and may be eliminated from further analysis. We notice that the drop off in removing explanatory variables is more extreme with fifteen vs. seven principal components. Therefore, we recommended limiting our interpretation to seven principal components and loading values that are within ten percent of the minimum and maximum loading values for a principal component because of the drop off in variable removal with fifteen principal components per the table below. This notion corresponds with our discussion of R^2 in our previous analysis sections.

Table 6. Observation that 15 Principal Components Retain Far More than 7 Principal Components

Loading	MIN MAX	10% of MIN MAX	20% of MIN MAX	25% of MIN MAX
PC7 Variables Removed	74	47	33	18
PC15 Variables Removed	64	33	14	4

Based on being 10% within the min/max loading for each principal component, several variables are identified as influencing each component. In lieu of describing each component in detail, we've chosen to summarize our findings via the following tables. Each table represents a principal component and a brief summary is provided describing the underlying factor we believe the principal component is describing. As mentioned previously, our intent is to extract only the most impactful variables for application during a future MLR. Therefore, only the variables within 10% of min/max loadings are shown.

Principal Component #1 – variable loadings are focused on a lack of education, family structure, income, and residence

Principal Component 1 Variable Selection, Relationship as Increase in Value, Variable Categorization

Relationship	Category	Variable	Loading	Descritption
-	Education	PctNotHSGrad	-0.17	percentage of people 25 and over that are not high school graduates
-	Family Structure	PctIlleg	-0.18	percentage of kids born to never married
-	Family Structure	PctYoungKids2Par	-0.18	percent of kids 4 and under in two parent households
+	Family Structure	PctFam2Par	0.18	percentage of families
+	Family Structure	PctKids2Par	0.18	percentage of kids in family housing with two parents
-	Income	PctPopUnderPov	-0.18	percentage of people under the poverty level
-	Income	PctUnemployed	-0.17	percentage of people 16 and over, in the labor force, and unemployed
-	Income	pctWPubAsst	-0.18	percentage of households with public assistance income in 1989
+	Income	medFamInc	0.17	median family income
+	Income	pctWInvInc	0.18	percentage of households with investment / rent income in 1989
-	Residence	PctHousNoPhone	-0.17	percent of occupied housing units without phone
+	Residence	PctPersOwnOccup	0.17	percent of people in owner occupied households

Principal Component #2 – variable loadings are focused on immigration, age, and race(Asian)

Principal Component 2 Variable Selection, Relationship as Increase in Value, Variable Categorization

Relationship	Category	Variable	Loading	Descritption
-	Immigration	PctForeignBorn	-0.21	percent of people foreign born
-	Immigration	PctRecImmig10	-0.22	percent of _population_ who have immigrated within the last 10 years
+	Age	pctWSocSec	0.17	percentage of households with social security income in 1989
-	Immigration	PctRecImmig8	-0.22	percent of _population_ who have immigrated within the last 8 years
-	Immigration	PctRecImmig5	-0.22	percent of _population_ who have immigrated within the last 5 years
-	Race	racePctAsian	-0.20	percentage of population that is of asian heritage
-	Immigration	PctRecentImmig	-0.21	percent of _population_ who have immigrated within the last 3 years

Principal Component #3 – variable loadings are focused on increase based on family structure and decreasing based on vocation, residence and education

Principal Component 3 Variable Selection, Relationship as Increase in Value, Variable Categorization

Relationship	Category	Variable	Loading	Descritption
+	Family Structure	PersPerOccupHous	0.28	mean persons per household
+	Family Structure	PersPerOwnOccHous	0.26	mean persons per owner occupied household
+	Family Structure	PersPerFam	0.26	mean number of people per family
-	Vocation	PctEmplProfServ	-0.15	percentage of people 16 and over who are employed in professional services
-	Residence	PctSameState85	-0.15	percent of people living in the same state as in 1985
-	Education	PctBSorMore	-0.16	percentage of people 25 and over with a bachelors degree or higher education

Principal Component #4 – variable loadings are focused on age and negatively on residency duration

Principal Component 4 Variable Selection, Relationship as Increase in Value, Variable Categorization

Relationship	Category	Variable	Loading	Descritption
-	Residence	PctSameHouse85	-0.22	percent of people living in the same house as in 1985
+	Age	agePct12t21	0.28	percentage of population that is 12-21 in age
+	Age	agePct12t29	0.26	percentage of population that is 12-29 in age
+	Age	agePct16t24	0.25	percentage of population that is 16-24 in age
-	Age	agePct65up	-0.22	percentage of population that is 65 and over in age

Principal component #5 – variable loadings are negatively related to marital status and positive related to vocation

Principal Component 5 Variable Selection, Relationship as Increase in Value, Variable Categorization				
Relationship	Category	Variable	Loading	Descritption
-	Marital Status	FemalePctDiv	-0.26	percentage of females who are divorced
-	Marital Status	TotalPctDiv	-0.27	percentage of population who are divorced
-	Marital Status	MalePctDivorce	-0.27	percentage of males who are divorced
+	Vocation	PctEmplProfServ	0.31	percentage of people 16 and over who are employed in professional services

Principal component #6 – variable loadings are positively related to race, and related to residency

Principal Component 6 Variable Selection, Relationship as Increase in Value, Variable Categorization				
Relationship	Category	Variable	Loading	Descritption
+	Residence	PctBornSameState	0.21	percent of people born in the same state as currently living
-	Residence	MedYrHousBuilt	-0.20	median year housing units built
+	Residence	PctUsePubTrans	0.22	percent of people using public transit for commuting
+	Race	racepctblack	0.20	percentage of population that is african american
-	Residence	PctSameCity85	-0.19	percent of people living in the same city as in 1985

Principal component #7 – variable loadings are negatively associated with vocation and positively associated with residence

Principal Component 7 Variable Selection, Relationship as Increase in Value, Variable Categorization				
Relationship	Category	Variable	Loading	Descritption
+	Vocation	PctEmplManu	0.29	percentage of people 16 and over who are employed in manufacturing
-	Residence	LandArea	-0.37	land area in square miles

<u>PCR Interpretation and Conclusion:</u>

The overall results of testing our dataset against the PCR model utilizing the first 7 and first 15 principal components can be seen in Figure 3; in this figure, we can see the measured values are very close to the values predicted by the PCR model for both the first 7 and first 15 principal components. A few individual results in testing our dataset against the PCR model can be seen in Table 7.

Table 7. Test Data Community Violent Crime Per Population

Observation	ViolentCrimesPerPop - Measured -	ViolentCrimesPerPop - 7 PCs -	ViolentCrimesPerPop - 15 PCs -
1	0.5848	0.5749	0.5713
4	0.4932	0.5694	0.7096
6	0.5192	0.5342	0.5336
7	0.3107	0.3779	0.3748
8	0.8193	0.6953	0.7076

In this table, we see the predicted violent crimes per population for both the first 7 and the first 15 principal components are very close to the actual violent crime rate per population for observations 1, 4, 6, 7, and 8. With these results, we conclude that this model is a good fit. For further testing using a multiple linear regression model, we determined through the principal component loadings that the variables to be included in the model will include the following:

pctWInvInc, pctWPubAsst, medFamInc, PctPopUnderPov, PctFam2Par, PctKids2Par, PctYoungKids2Par, PctIlleg, PctHousNoPhone, PctRecentImmig, PctRecImmig5, PctRecImmig8, PctRecImmig10, PctForeignBorn, PersPerFam, PersPerOwnOccHous, MedYrHousBuilt, PctSameCity85, agePct12t21, agePct16t24, agePct12t29, PctEmplProfServ, MalePctDivorce, FemalePctDiv, TotalPctDiv, racepctblack, agePct65up, PctBornSameState, PctSameHouse85, PctUsePubTrans, PctEmplManu, LandArea, PctNotHSGrad, PctUnemployed, pctWSocSec, racePctAsian, PersPerOccupHous, PctPerOwnOccup, PctSameCity85, PctSameState85, and *PctBSorMore*.

After our data cleansing, PCA, and PCR analysis, PCR reduced our prediction model from 88 principal components down to 7 principal components, and we reduced the number of variables to be utilized in a future MLR prediction model from 122 variables to a much more manageable 41 variables.

Appendix:

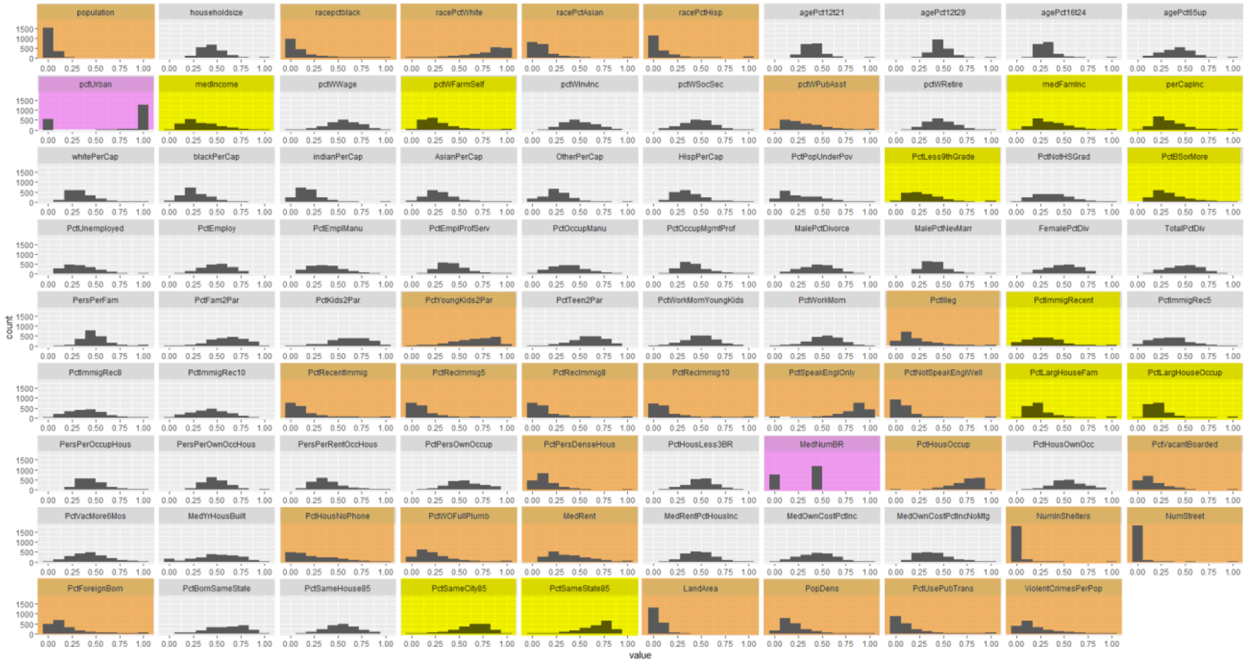


Figure 4. Variable Distributions Before Transformation

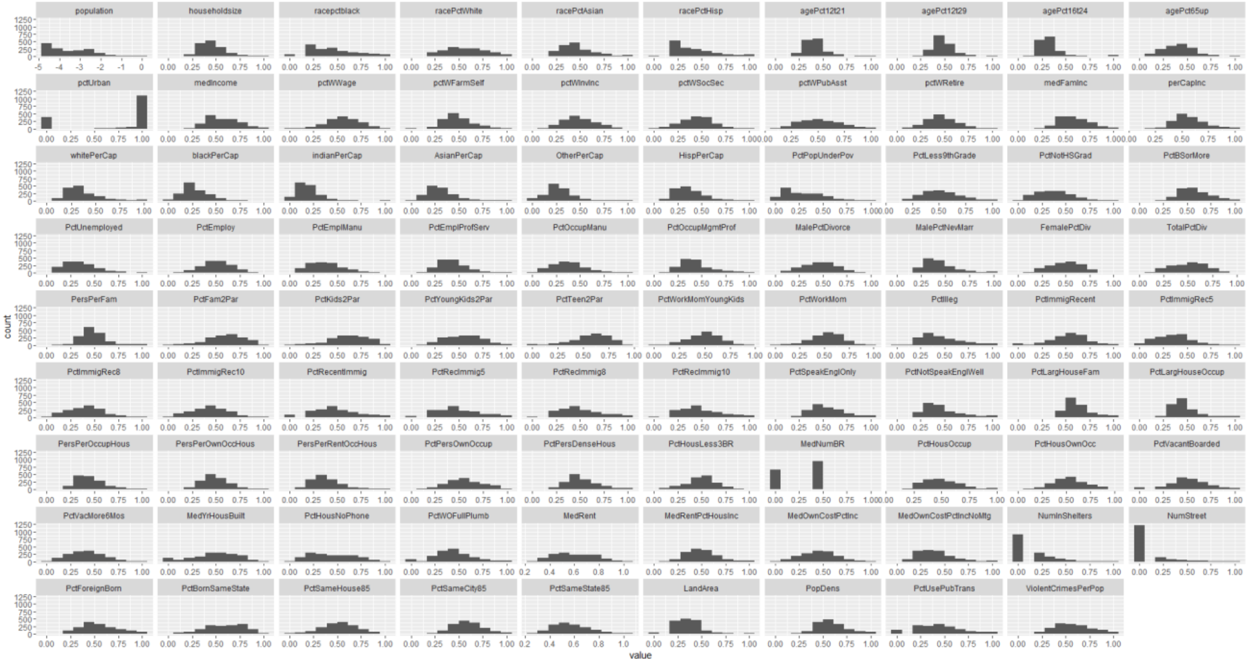


Figure 5. Variable Distributions After Transformation

Table 8. Principal Component Eigenvalues and Variance Explained

	Eigenvector	Variance. Proportion	Variance. CumProportion		Eigenvector	Variance. Proportion	Variance. CumProportion
PC1	23.06155	0.262063	0.262063	PC45	0.1201181	0.00136498	0.9825061
PC2	16.2108	0.1842136	0.4462766	PC46	0.113379	0.0012884	0.9837944
PC3	8.518739	0.09680385	0.5430805	PC47	0.1058267	0.00120258	0.984997
PC4	6.552144	0.07445619	0.6175367	PC48	0.09934603	0.00112893	0.986126
PC5	4.17888	0.04748727	0.6650239	PC49	0.08889468	0.00101017	0.9871361
PC6	3.509785	0.03988392	0.7049079	PC50	0.08067279	0.00091674	0.9880529
PC7	3.005185	0.03414984	0.7390577	PC51	0.07724849	0.00087782	0.9889307
PC8	2.072638	0.02355271	0.7626104	PC52	0.06940795	0.00078873	0.9897194
PC9	1.830682	0.0208032	0.7834136	PC53	0.06342719	0.00072076	0.9904402
PC10	1.593952	0.01811309	0.8015267	PC54	0.06047979	0.00068727	0.9911274
PC11	1.433136	0.01628564	0.8178123	PC55	0.05975235	0.000679	0.9918064
PC12	1.325446	0.01506189	0.8328742	PC56	0.05462066	0.00062069	0.9924271
PC13	1.020168	0.01159281	0.844467	PC57	0.05240185	0.00059548	0.9930226
PC14	0.8876897	0.01008738	0.8545544	PC58	0.05044703	0.00057326	0.9935959
PC15	0.8566376	0.00973452	0.8642889	PC59	0.04393341	0.00049924	0.9940951
PC16	0.8027904	0.00912262	0.8734116	PC60	0.0432712	0.00049172	0.9945868
PC17	0.716864	0.00814618	0.8815577	PC61	0.04197151	0.00047695	0.9950638
PC18	0.6579654	0.00747688	0.8890346	PC62	0.04040844	0.00045919	0.995523
PC19	0.5878844	0.0066805	0.8957151	PC63	0.03704013	0.00042091	0.9959439
PC20	0.5452262	0.00619575	0.9019109	PC64	0.03261188	0.00037059	0.9963145
PC21	0.5107496	0.00580397	0.9077148	PC65	0.03084828	0.00035055	0.996665
PC22	0.4839226	0.00549912	0.913214	PC66	0.02932695	0.00033326	0.9969983
PC23	0.4376481	0.00497327	0.9181872	PC67	0.02808608	0.00031916	0.9973174
PC24	0.435707	0.00495122	0.9231385	PC68	0.02789	0.00031693	0.9976344
PC25	0.4153049	0.00471937	0.9278578	PC69	0.02444554	0.00027779	0.9979122
PC26	0.3990482	0.00453464	0.9323925	PC70	0.02272911	0.00025829	0.9981704
PC27	0.3889517	0.00441991	0.9368124	PC71	0.02164478	0.00024596	0.9984164
PC28	0.3722223	0.0042298	0.9410422	PC72	0.01951833	0.0002218	0.9986382
PC29	0.3359008	0.00381705	0.9448592	PC73	0.01807379	0.00020538	0.9988436
PC30	0.3119081	0.00354441	0.9484036	PC74	0.01549631	0.0001761	0.9990197
PC31	0.2989307	0.00339694	0.9518006	PC75	0.01304393	0.00014823	0.9991679
PC32	0.2888835	0.00328277	0.9550833	PC76	0.01238431	0.00014073	0.9993086
PC33	0.2671271	0.00303554	0.9581189	PC77	0.01119455	0.00012721	0.9994359
PC34	0.2447813	0.00278161	0.9609005	PC78	0.00905598	0.00010291	0.9995388
PC35	0.2350593	0.00267113	0.9635716	PC79	0.00785741	8.93E-05	0.9996281
PC36	0.2259813	0.00256797	0.9661396	PC80	0.00622501	7.07E-05	0.9996988
PC37	0.2035015	0.00231252	0.9684521	PC81	0.00558529	6.35E-05	0.9997623
PC38	0.1971301	0.00224012	0.9706922	PC82	0.00498142	5.66E-05	0.9998189
PC39	0.1826707	0.0020758	0.972768	PC83	0.00464921	5.28E-05	0.9998717
PC40	0.1613267	0.00183326	0.9746013	PC84	0.00395828	4.50E-05	0.9999167
PC41	0.1596613	0.00181433	0.9764156	PC85	0.00347548	3.95E-05	0.9999562
PC42	0.1485225	0.00168776	0.9781034	PC86	0.00248792	2.83E-05	0.9999844
PC43	0.138494	0.0015738	0.9796772	PC87	0.0009191	1.04E-05	0.9999949
PC44	0.128824	0.00146391	0.9811411	PC88	0.00044937	5.11E-06	1.0000000

Table 9. Variable Transformation Summary

Variable	Transformation	Variable	Transformation
population	log	PctSpeakEnglOnly	cubed root
racepctblack	cubed root	PctNotSpeakEnglWell	cubed root
racePctWhite	abs cubed root	PctLargHouseFam	cubed root
racePctAsian	cubed root	PctLargHouseOccup	sqrt
racePctHisp	cubed root	PctPersDenseHous	cubed root
medIncome	sqrt	PctHousOccup	abs sqrt
pctWFarmSelf	sqrt	PctVacantBoarded	cubed root
pctWPubAsst	sqrt	PctHousNoPhone	sqrt
medFamInc	sqrt	PctWOFullPlumb	sqrt
perCapInc	sqrt	MedRent	sqrt
PctLess9thGrade	sqrt	NumInShelters	cubed root
PctBSorMore	sqrt	NumStreet	cubed root
PctYoungKids2Par	abs sqrt	PctForeignBorn	cubed root
PctIlleg	sqrt	PctSameCity85	abs sqrt
PctImmigRecent	sqrt	PctSameState85	abs sqrt
PctRecentImmig	cubed root	LandArea	cubed root
PctRecImmig5	cubed root	PopDens	cubed root
PctRecImmig8	cubed root	PctUsePubTrans	cubed root
PctRecImmig10	cubed root	ViolentCrimesPerPop	cubed root

Table 10. Correlation Matrix Summary

>33% variables correlated				
PctPopUnderPov	PctPersDenseHous	PctFam2Par	PctKids2Par	medIncome
PctYoungKids2Par	PctHousNoPhone	pctWInvInc	medFamInc	PctPersOwnOccup
pctWPubAsst	perCapInc	PctNotHSGrad	PctUnemployed	PctIlleg
MedRent	PctLess9thGrade	ViolentCrimesPerPop	racePctWhite	PctHousOwnOcc
PctTeen2Par	PctBSorMore	PctOccupMgmtProf		
<33% and > 20% variables correlated				
TotalPctDiv	pctWWage	PctHousLess3BR	MalePctDivorce	FemalePctDiv
PctLargHouseFam	whitePerCap	blackPerCap	PctOccupManu	PctRecImmig5
PctRecImmig8	PctWOFullPlumb	HispPerCap	PctEmploy	MalePctNevMarr
PctRecentImmig	PctRecImmig10	PctVacantBoarded	racePctAsian	PctLargHouseOccup
racePctHisp	PctSpeakEnglOnly	PctNotSpeakEnglWell	PctImmigRec10	PersPerRentOccHous
<20% and > 12% variables correlated				
PctImmigRec8	PopDens	agePct12t29	MedNumBR	PctForeignBorn
PctSameHouse85	racepctblack	pctWSocSec	PctImmigRec5	NumInShelters
PctBornSameState	householdsize	agePct65up	PctImmigRecent	PersPerOccupHous
PctRecentImmig	PctRecImmig10	PctVacantBoarded	racePctAsian	PctLargHouseOccup
agePct16t24	PersPerFam	PersPerOwnOccHous	PctHousOccup	MedOwnCostPctInc
<12% variables correlated				
PctSameCity85	PctUsePubTrans	PctVacMore6Mos	agePct12t21	pctWRetire
OtherPerCap	AsianPerCap	MedYrHousBuilt	population	MedRentPctHousInc
NumStreet	PctEmplProfServ	LandArea	pctUrban	PctSameState85
PctEmplManu	PctWorkMomYoung	PctWorkMom	MedOwnCostPct	pctWFarmSelf
indianPerCap				