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



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*



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.

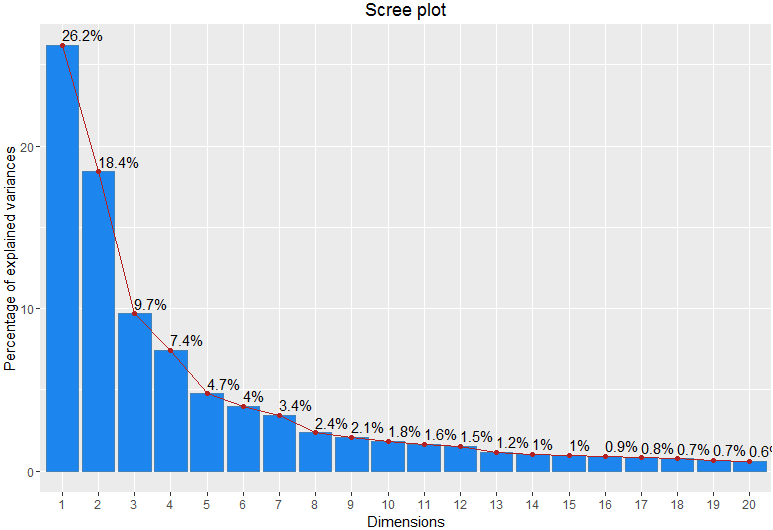
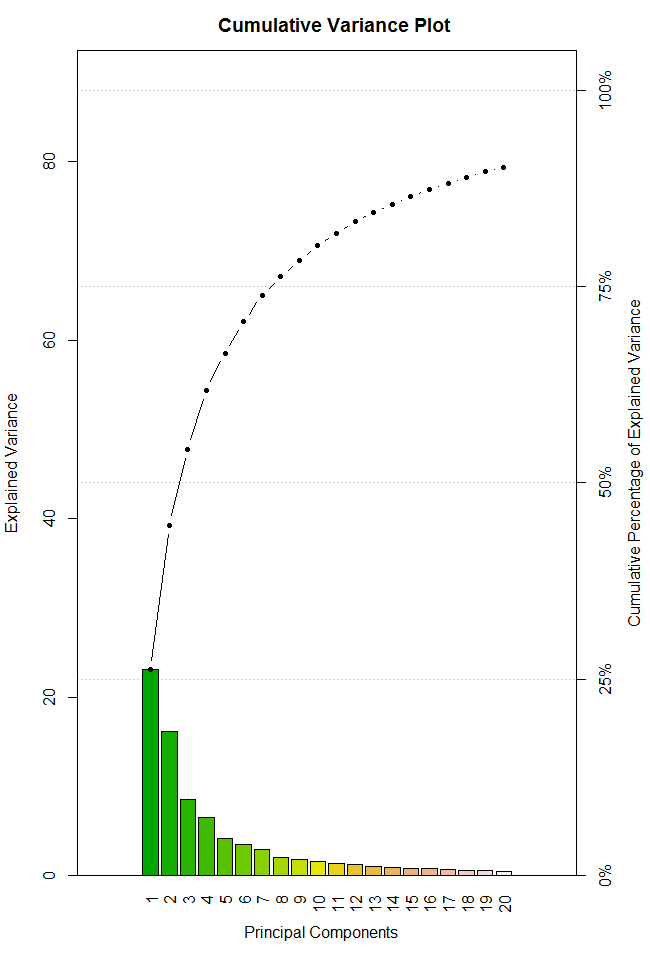
 

Figure 1. Scree Plot (left) and Cumulative Variance Plot (right) for First 20 PCs

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 R2 Statistics for first 20 PCs



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, R2,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-R2 values are 0.6814 and 0.7073 for seven and fifteen PCs respectively. This, in combination with the other PC adjusted-R2 values, clearly expresses that not much is gained by including fifteen components vs. seven components and that the rate of increasing R2 values decreases substantially after seven PCs are included (See Figure 2), further validating this selection.

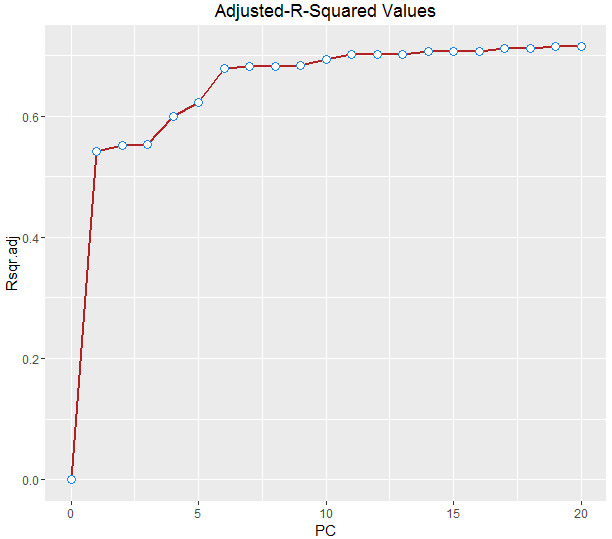


Figure 2. Adjusted-R2 Values for First 20 PCs

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.

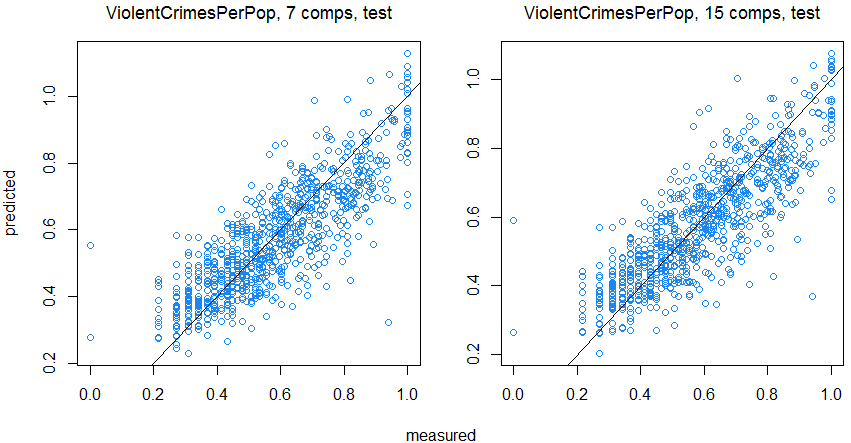
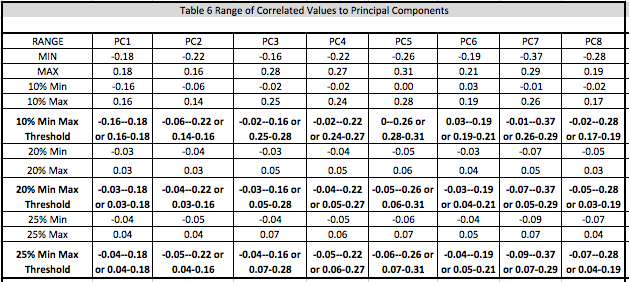


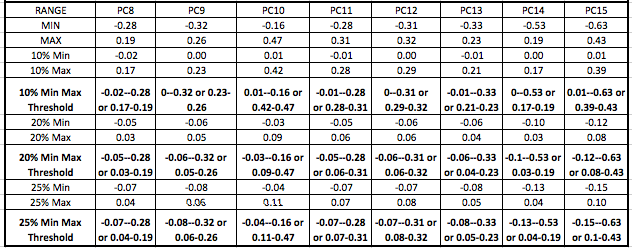
Figure 3. Predicted vs. Measured Response Values for the Test Data Set

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





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 R2 in our previous analysis sections.

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



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



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



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



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



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



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



PCR Interpretation and Conclusion:

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



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

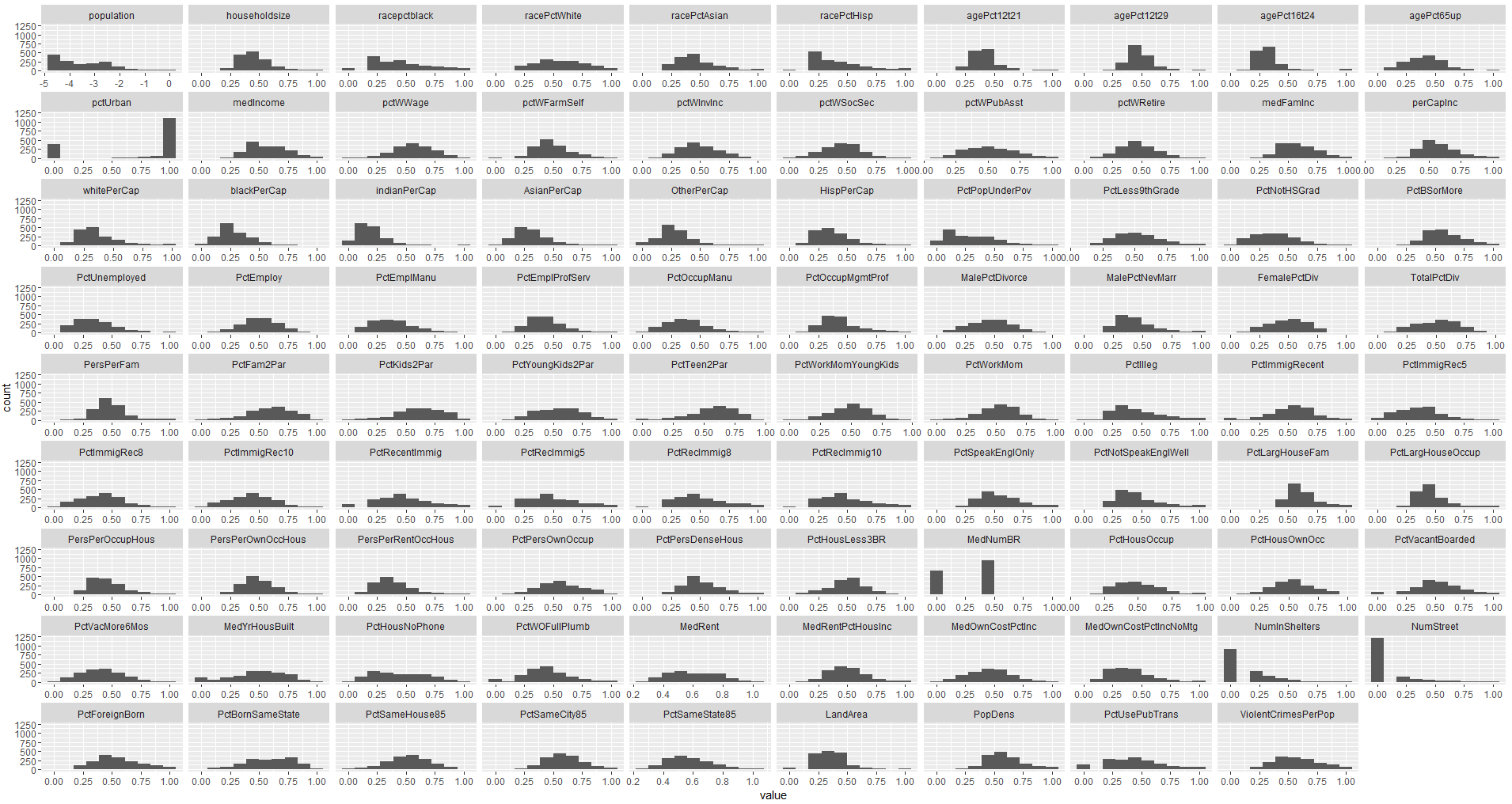


Figure 5. Variable Distributions After Transformation

Table 8. Principal Component Eigenvalues and Variance Explained



Table 9. Variable Transformation Summary



*Table 10. Correlation Matrix Summary*

