Tranformations to Normality

Principal Component Analysis (PCA)

Assignments:

1) **Find the best Box-Cox Transformation for each variable.**
2) **Construct the Normal probability plots for the original and transformed variables. Discuss the quality and appropriateness of the transformations.**

Many statistical test and intervals are based on assumption of normality ie. they assume the data sets are normally distributed. this assumptions often leads to tests that are simple, mathematically tractable, and powerful compared to tests that do not make normality assumption. However, many real data are not normally distributed. However, to fit our needs, an appropriate transformation of a dataset can often yield a data set that does follow approximately a normal distribution . This improves the applicability and usefulness of statistical techniques based on normality assumption. [1]

The Box-Cox transformation is a particularly useful family of transformations. It is defined as

$$T(Y)= (Y^{\lambda} -1)/\lambda......................... \quad 1$$

where Y is the response variable and λ is the transformation parameter.

Given a particular transformation such as the Box-Cos transformation defined above, it is helpful to define a measure of the normality of the resulting transformation. One measure is to compute the correlation coefficient of a normal probability plot. The correlation is compute between the vertical and horizontal axis of variables of the probability plot and is a convenient measure of the linearity of the probability plot (the more linear the probability plot, the better a normal distribution fits the data)

The value of λ corresponds to the maximum corelation on the Box-Cox Normality plot.

We then transform the variable using the equation (1) above.

Sometimes the tranformation might not result a normal distribution due to various reasons. Hence, not every data can be transformed to normality using Box-Cox transformation.

We want to try to change USArrests dataset to Normality. It has four data, Murder, Assault, UrbanPop and Rape.

The analysis of all those variables have been studied before and after applying Box-Cox Transformation.

Murder Data:

Before Transformation

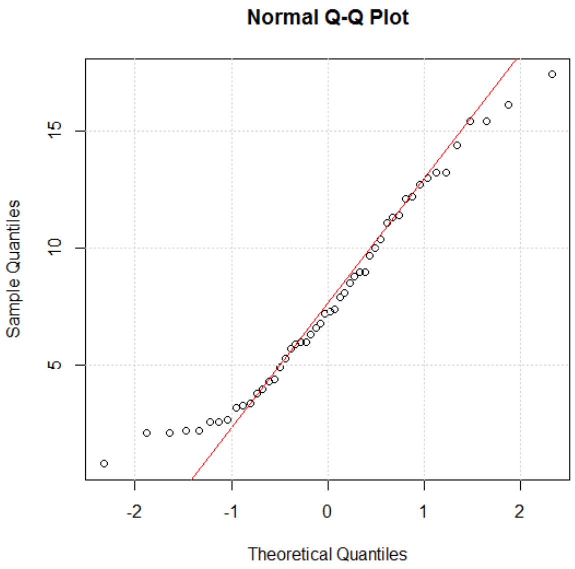
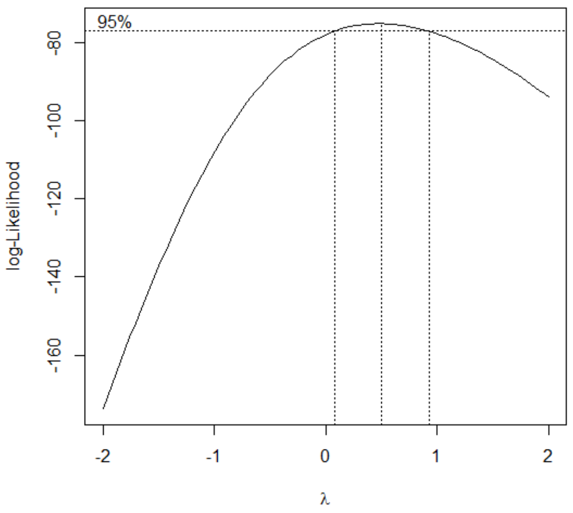


Fig 1.1 Nomal Q-Q Plot



After Box-Cox Transformation:

Fig.1.2 Log-likelihood Profile

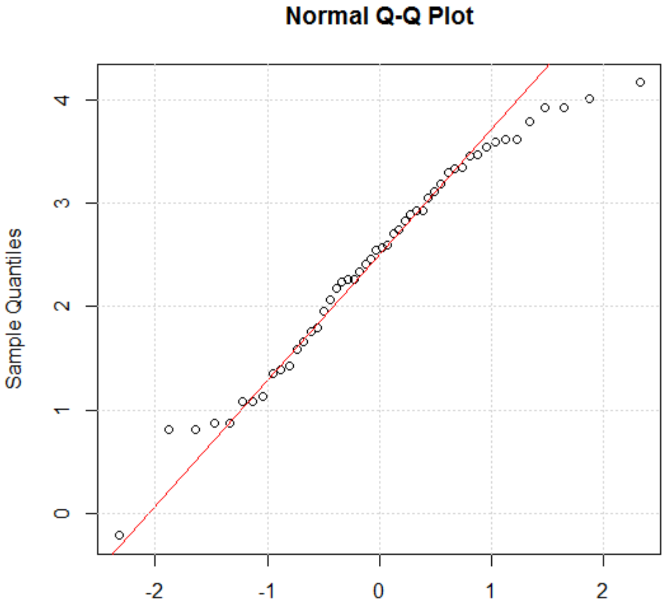


Fig 1.3 Normal Q -Q Plot after transformation

For Murder Data, we can see that the overall distribution of the data has been changed a bit, which can be seen from the elements distributed on two halves of red line in figure 1 and figure 3.

The third figure seems more normal than first, but there is nothing drastic. It is not much appropriate i.e quality of transformation is not so good.

Also, the scale of quantiles have been changed indicating the change in scale after the transformation.

Fig 2 illustrate the selection range for the best λ for Box-Cox Transformation.

For **Assault** data:

Before Transformation:

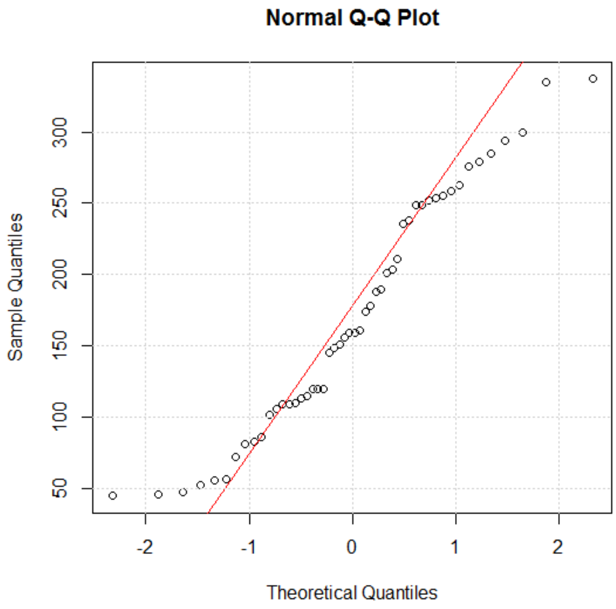


Fig 2.1. Normal Q Q Plot

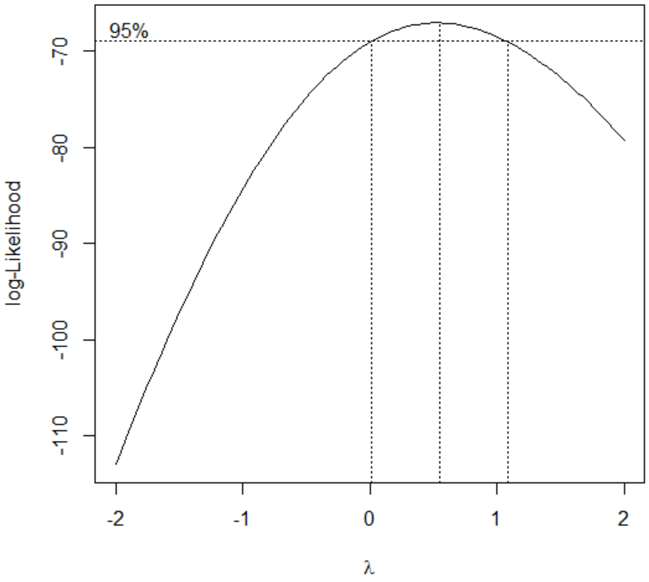


Fig 2.2. log Likelihood Profile

After Transformation:

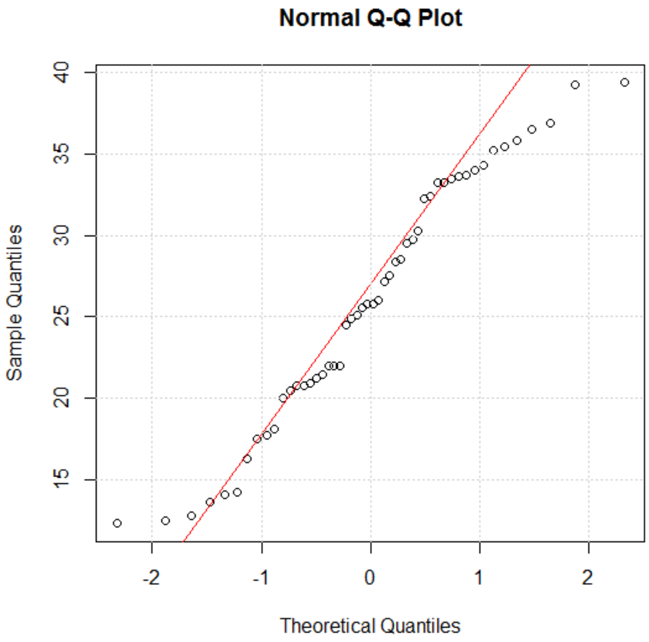


Fig 2.3. Normal Q Q Plot after transformation

In the adjacent figure, we can see that the QQ plot after transformation looks more normal than the original QQ Plot, and the normal line is closer to the data points in the figure.

It brings in some improvement in the normality of the data.

We should also make a note of the change of scales in the sample Quantiles.

For **UrbanPop** data:

Before Transformation:

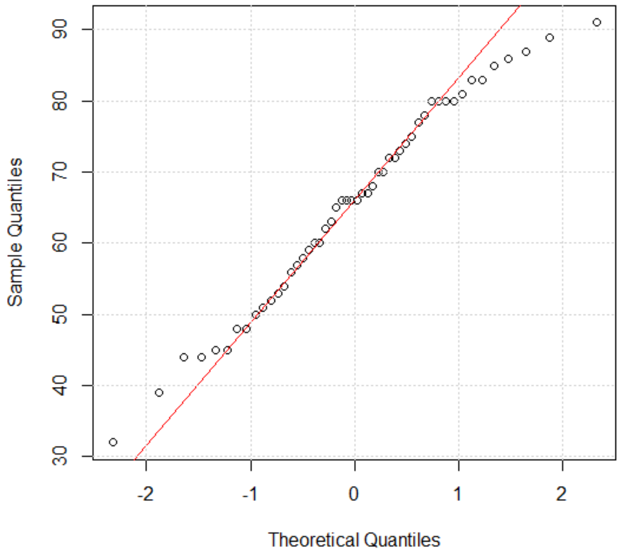


Fig 3.1. Normal Q Q Plot

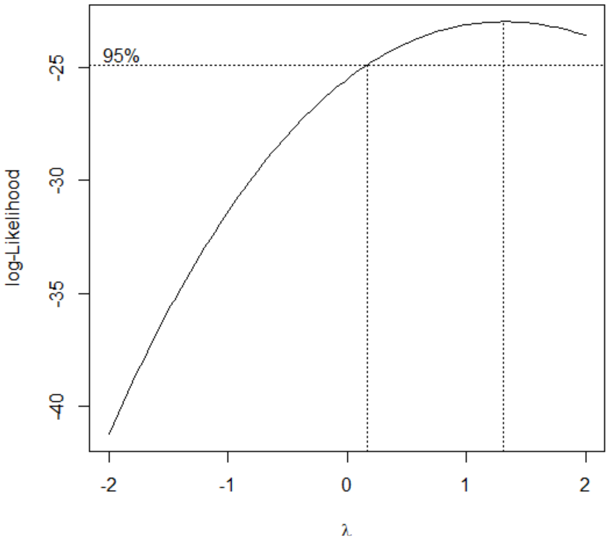


Fig 3.2. Log Likelihood Profile

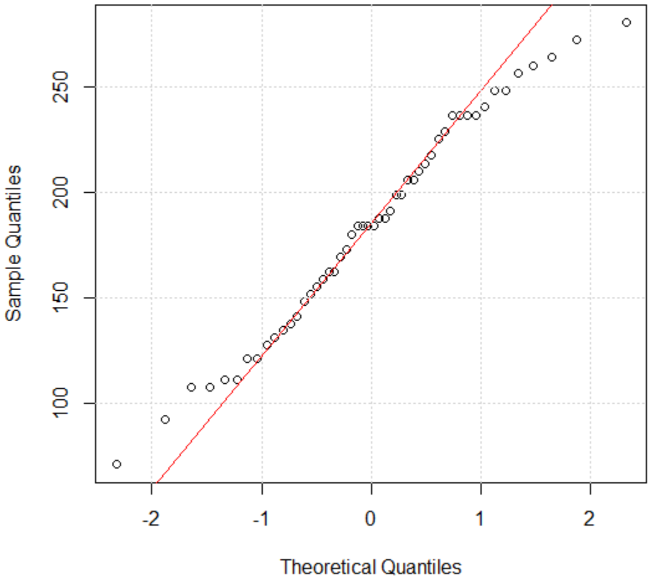


Fig 3.3 Normal Q Q Plot after transformation

The three figure show QQ Plot before transformation, the log likelihood profile and the QQ plot after transformation.

Unlike previous data, there is almost no change in the data after transformation; only the scales seems to have changed.

The λ is 1.31 in this case.

So, the operation of conversion to normality is quite unsuccessful in this case, and can be avoided.

For **Rape** data:

Before Transformation:

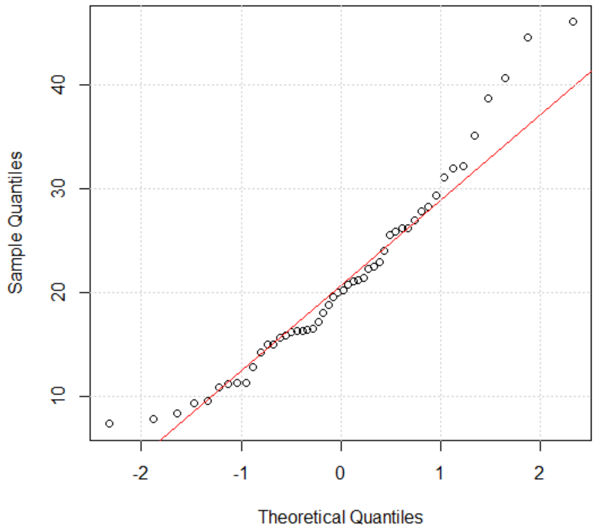


Fig 4.1 Normal Q Q Plot

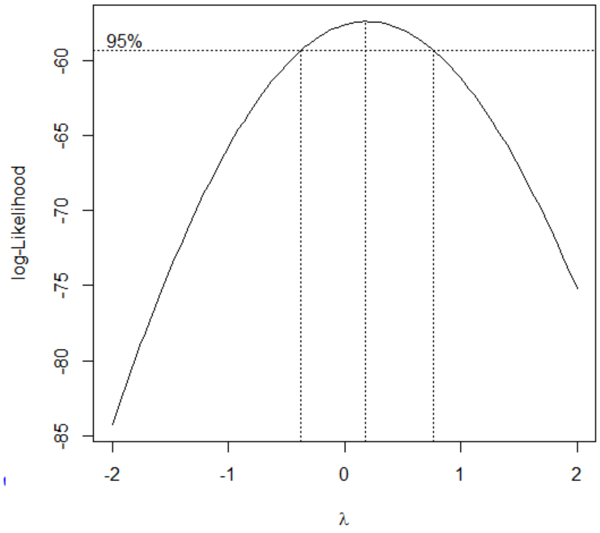


Fig 4.2 Log Likelihood Profile

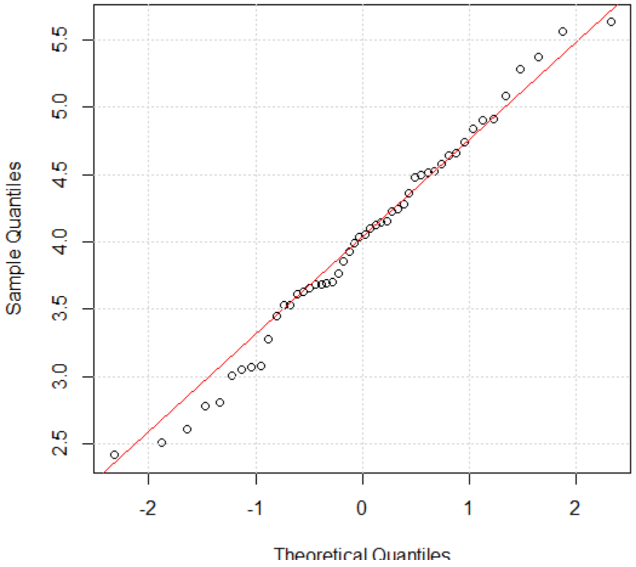


Fig 4.3 Normal Q Q Plot after transformation

There can be seen significant transformation in this case of **Rape** data before and after the transformation.

The λ in this case is taken to be 0.19.

This is the best quality transformation in the four cases as can be seen.

3) Perform PCA and find the loadings.

4) Compute principal components and find the correlations between the first PC and the original variables.

Principal Component Analysis (PCA) is a dimension reduction tool that can be used to reduce large set of variables to a small set that still contains most of the information in the large set. It is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components.

The first component accounts for as much of the variability in the data as possible and each succeeding component accounts for as much of the remaining variability as possible.

PCA can be considered as a rotation of the axes of the original variable coordinate system to new orthogonal axes, called principal axes, such that the new axes coincide with directions of maximum variation of original observations. Consider the axis passing though the ends of the elliptical cluster of points in Figure 1. Project the original data points onto this axis. The point y_{1m} is the projection of the point $(X_{1m,} X_{2m)}$ onto the axis defined by the direction $Y_{1.}$ This axis has the property that the variance of the projected points y_{1m}, m=1,…n, is greater than the variance of the points when projected onto any other line or axis passing through (x_1, x_2). Any line parallel to Y_1 also has the same property.

The property of maximum variation of the projected points defines the first principal axis; it is the line or direction with maximum variation of the projected values of the original data points. The projected values corresponding to this direction of maximum variation are the principal component scores. The first principal axis is also called the line of best fit since the sum of squares of the perpendicular deviations of the original data points from the line is minimum. Successive principal axes are determined with the property that they are orthogonal to the previous principal axes and that they maximize the variation of the projected points subject to these constraints.[2]

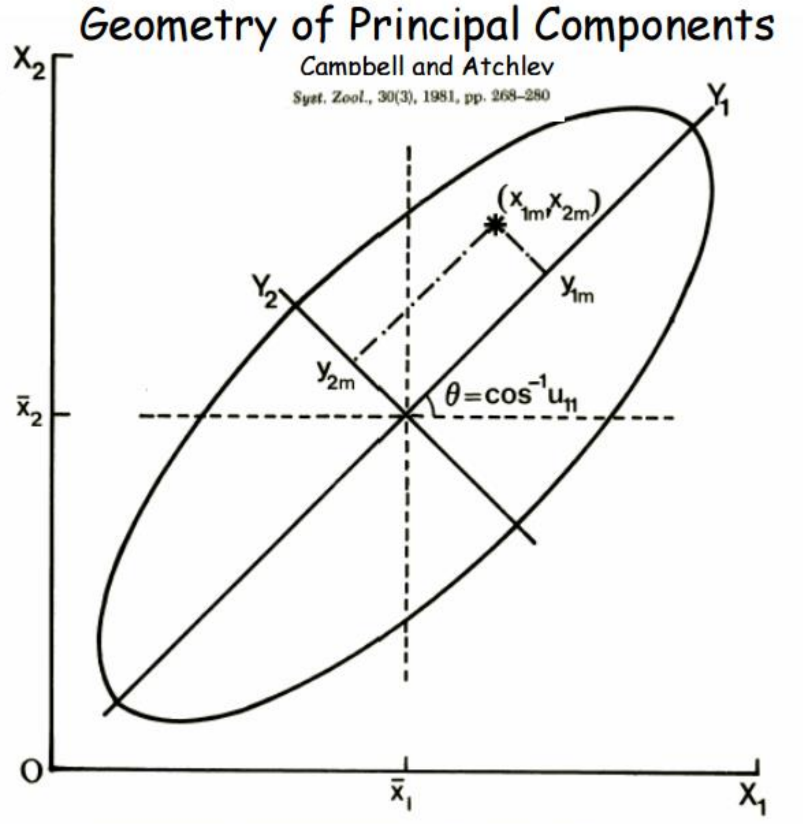


FIG. 1.—Idealized representation of scatter diagram for two variables, showing the mean for each variable ($\bar{x}_1$ and $\bar{x}_2$), 95% concentration ellipse, and principal axes Y_1 and Y_2. The points y_{1m} and y_{2m} give the principal component scores for the observation $\mathbf{x}_1 = (x_{1m}, x_{2m})^T$. The cosine of the angle θ between Y_1 and X_1 gives the first component u_{11} of the eigenvector corresponding to Y_1.

Objective of PCA:

1) Reduce the attribute space from a larger number of variables to a small number of factors and as such is a "non-dependent" procedure.
2) PCA is dimensionality reduction or data compression method, but there is no guarantee that the dimensions are interpretable.

3) To select a subset of variable from a larger set, based on which original variables have the highest corelations with the principal component.

The principal components are linear combinations of the original variables weighted by their contribution to explaining the variance of a particular orthogonal dimension. It then removes this variance and seeks a second linear combination which explains the maximum proportion of remaining variance, and so on. This is called principal axis method.

Eigen Vectors: Principal components reflect both common and unique variance of the variables and may be seen as a variance-focused approach seeking to reproduce both the total variable variance with all components and to reproduce the correlations.

Eigen Values: Also called characteristic roots. The eigenvalue for a given factor measures the variance in all the variables which is accounted for by that factor.
A factor's eigenvalue may be computed as the sum of its squared factor loadings for all the variables.

Factor Loadings: The factor loadings, also called coefficient loadings in PCA, are the correlation coefficients between the variable(rows) and factors(column).

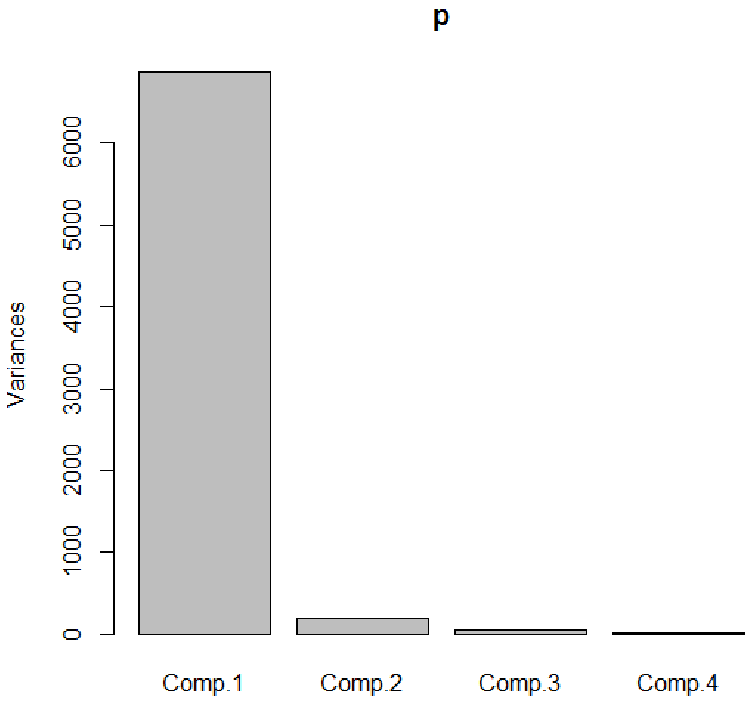
PC Scores: Also called component scores in PCA, these scores are the scores of each case(row) on each factor(column). To compute the factor score for a given case for a given factor, one takes the case's standardized score on each variable, multiplies by the corresponding factor loading of the variable for the given factor, and sum these products.

We computed the principal components, and found the loadings.

Observation:

In the statistical analysis that we did with USArrests Data, we obtained the following results.

Component 1 accounts for the maximum variability(96.5%) in the dataset, and is the principal component. Therefore, we can say that component 1 can represent the original data without losing much of the information.



First Principal Component Analysis:
The first Principal component only related with X2 (Assault data) negatively, i.e. decreasing assault would increase this component. It would follow that places with low assault would have high principal component, places with high assault would have low value of this principal component.
First component was not related with any other original variables.

Second Principal Component Analysis:

Second PC is related highly with Urban Population, and quite related to Rape as well. We can derive that places with low urban population would have high second principal component.

Similar observations can be made on third and fourth component.

Reference

1. http://www.itl.nist.gov/div898/handbook/eda/section3/eda336.htm
2. ftp://statgen.ncsu.edu/pub/thorne/molevoclass/AtchleyOct19.pdf