Introduction:

In this laboratory experiment, we will be working with a dataset called USArrests. This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas. There are 50 observations of the four variables – Murder, Assault, UrbanPop and Rape. Since this data is not normally distributed, we will find box-cox transformation of each variable to find the best estimate of transformation parameter that transforms these variables into normal variables. After transformation to normality, we will be performing Principal Component Analysis on these variables to find a new set of variables as a linear combination of our existing four variables and some coefficients of linear transformation which we call the loadings. These new set of variables are called the principal components of our data which are independent and the variance of our dataset will be concentrated among the first few of these principal components.

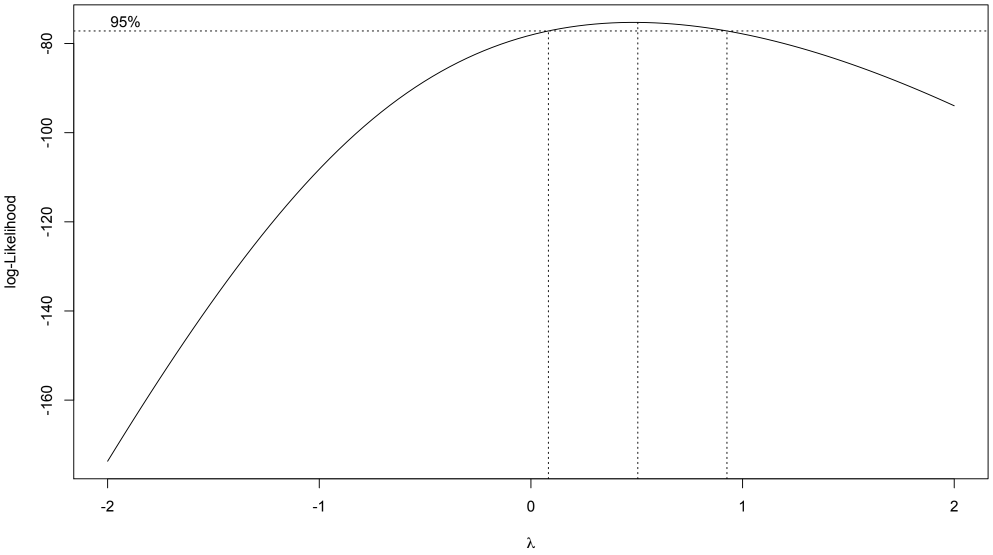
Perform PCA for data set USArrests from the data collection.

1. **Find the best Box-Cox transformation for each variable;**

Solution:

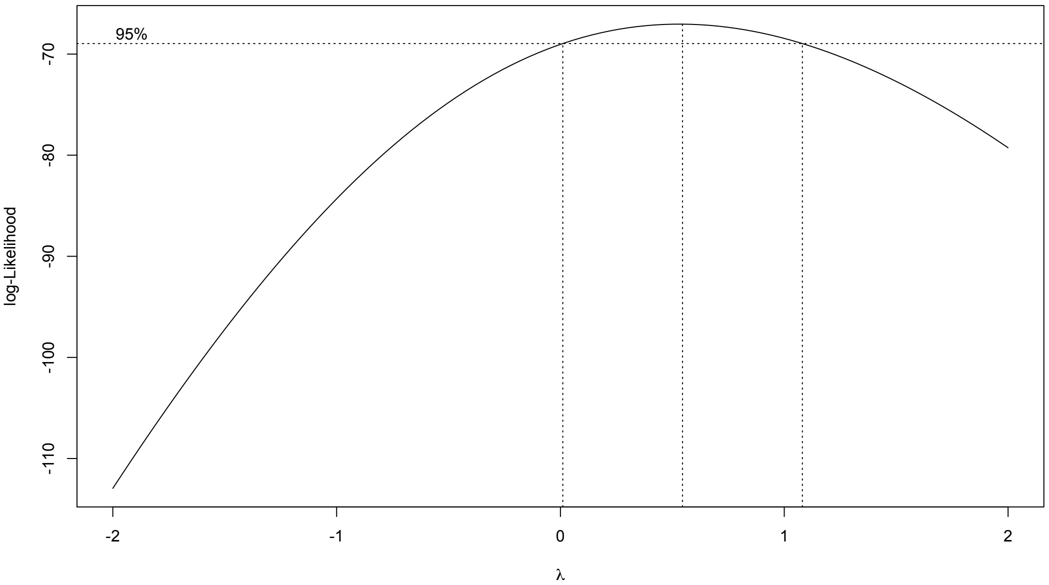
The box-cox transformation for each variable from USArrests with the best estimate for the transformation parameters (λ) in each case as shown below:

a. Murder



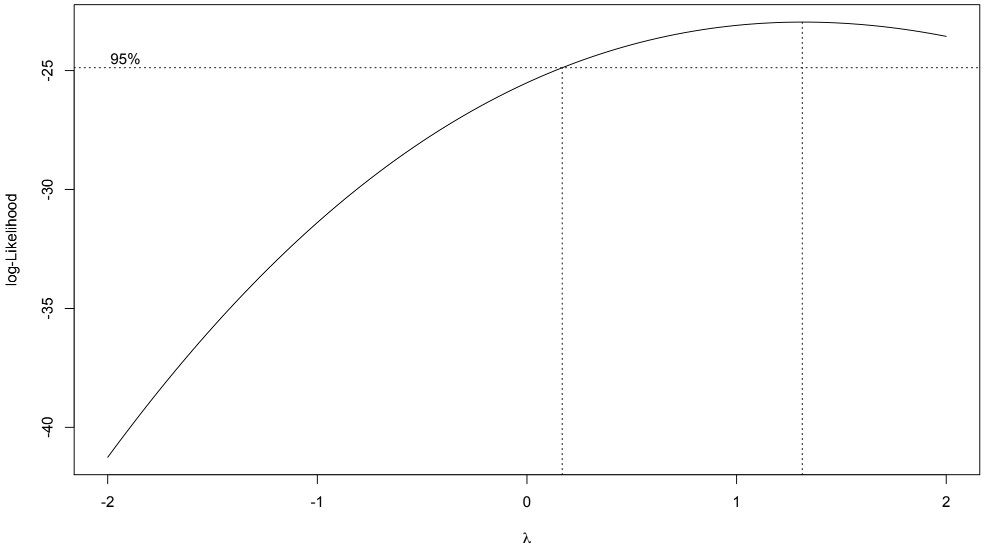
Best estimate of λ = 0.458268

b. Assault



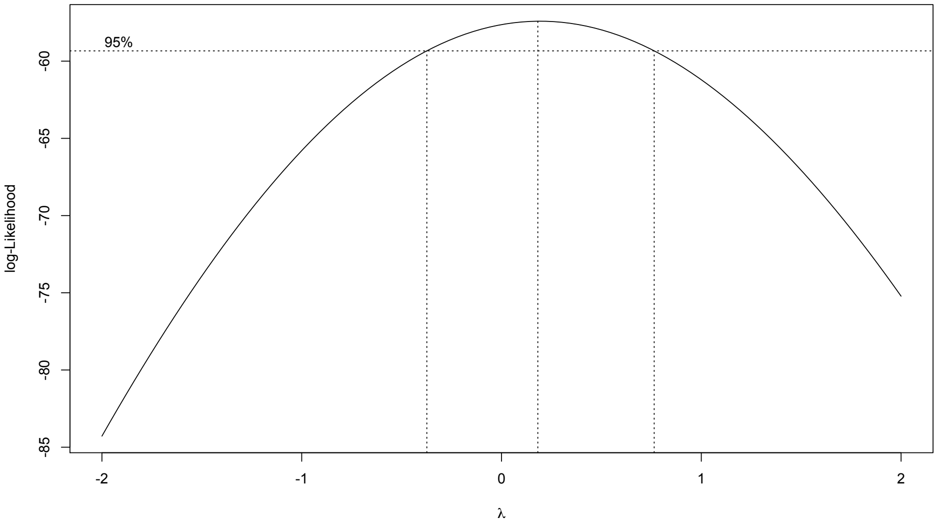
Best estimate of λ = 0.5311643

c. UrbanPop



Best estimate of λ = 1.316366

d. Rape



Best estimate for λ = 0.1928036

2. Construct the Normal probability plots for the original and transformed variables. Discuss the quality and appropriateness of the transformations;

Solution:

Normality plots for each of the original variables, and after transforming them with their best estimate of transformation parameters are shown below:

a. Murder

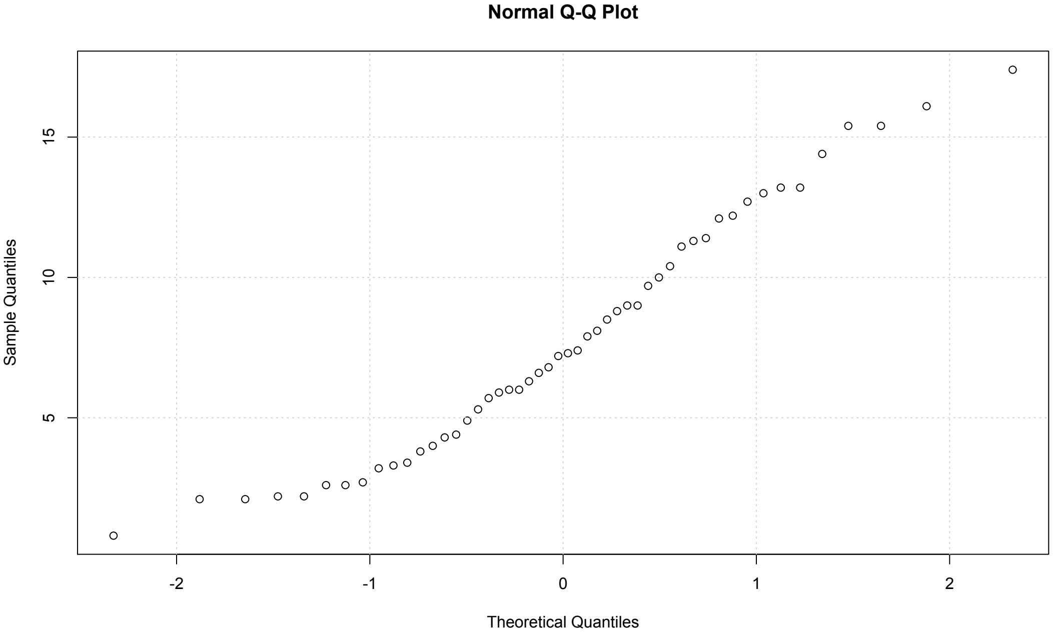


Fig 1: Normality plot for original variable Murder

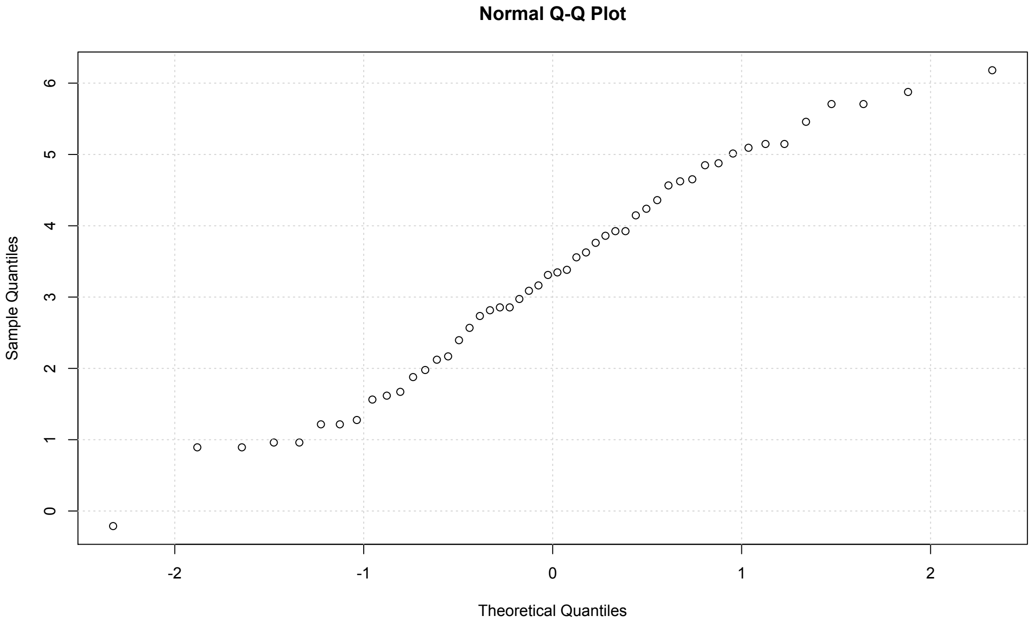


Fig 2: Normality plot for transformed variable Murder

Using our transformation parameter λ = 0.458268, we were able to transform the variable Murder to a new transformed variable that follows normality, as shown from in Fig 2.

b. Assault

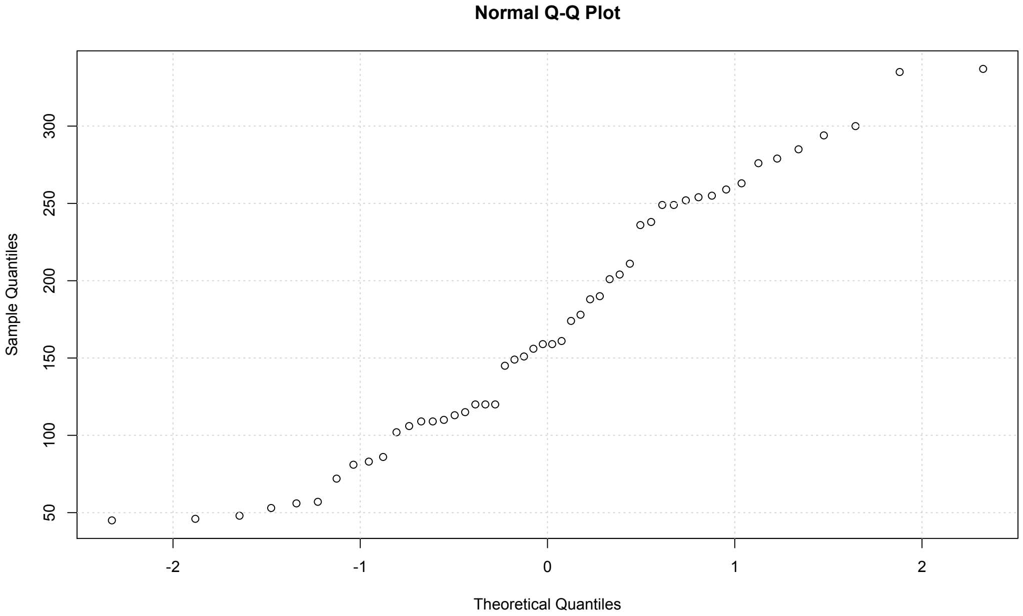


Fig 3: Normality plot for original variable Assault

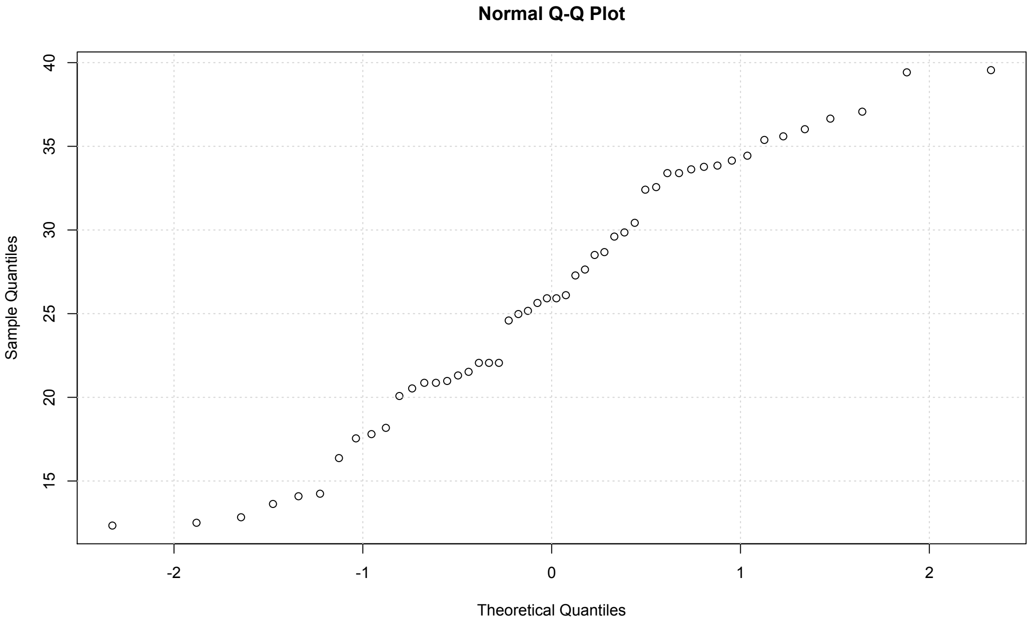


Fig 4: Normality plot for transformed variable Assault

The transformed variable Assault also follows the normal distribution as shown in Fig 4, after it has been transformed with parameter λ = 0.5311643.

c. UrbanPop

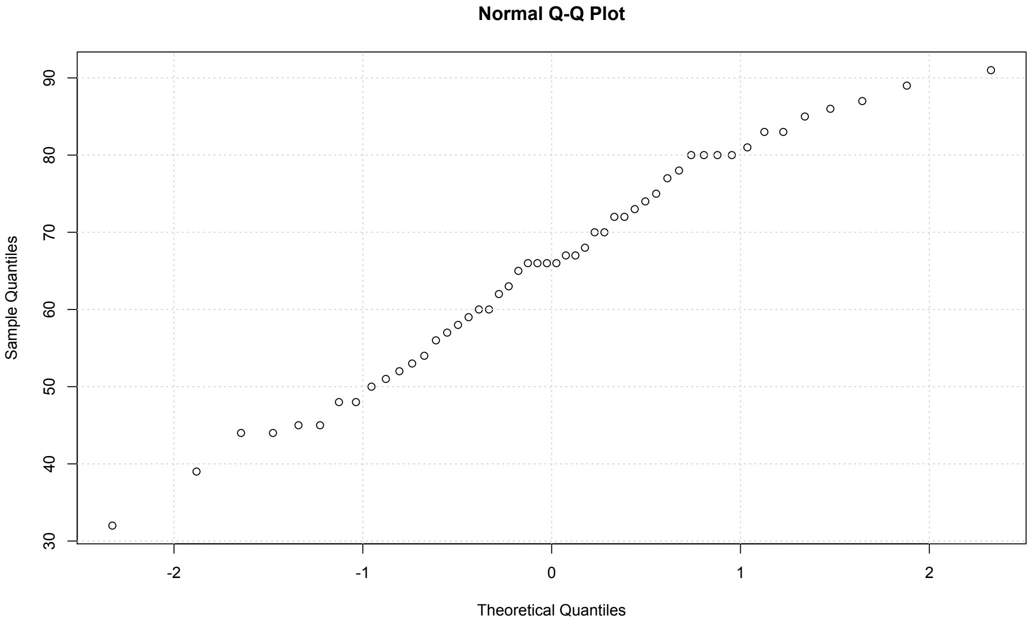


Fig 5: Normality plot for original variable UrbanPop

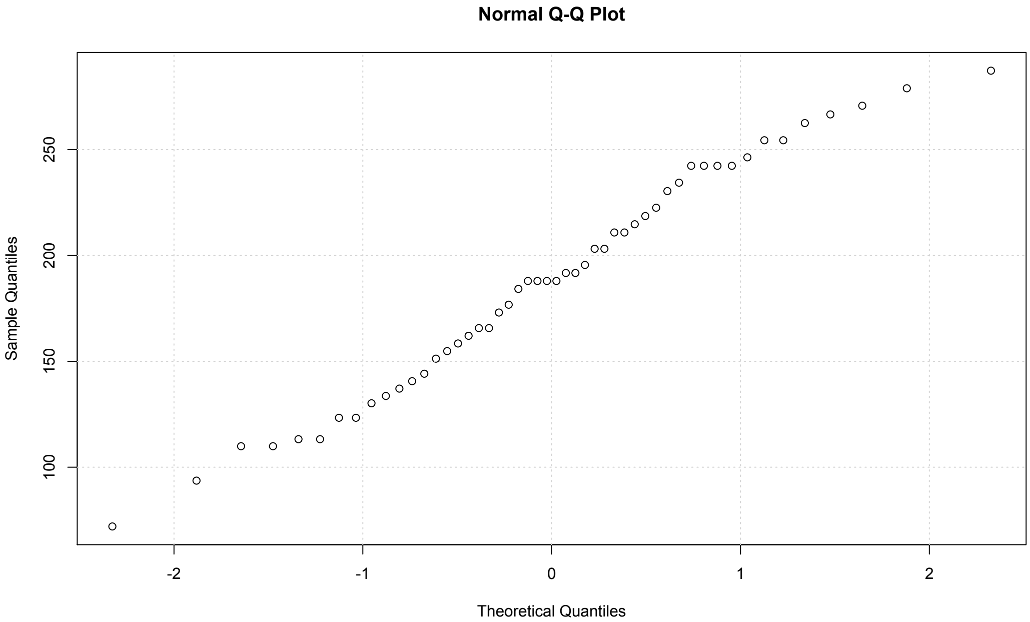


Fig 6: Normality plot for transformed variable UrbanPop

As seen from Fig 5, the original variable UrbanPop already has a normal distribution. It is not necessary to transform it into normality. As seen from Fig 6, even after applying the transformation parameter λ = 1.316366, the change in normality plot is not significant.

d. Rape

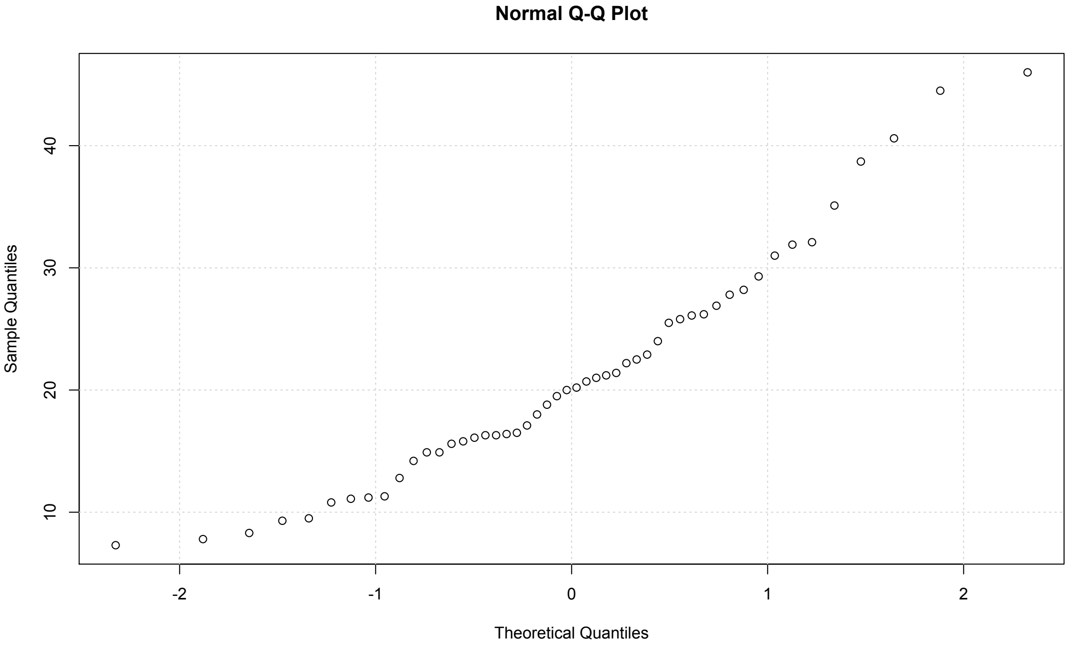


Fig 7: Normality plot for original variable Rape

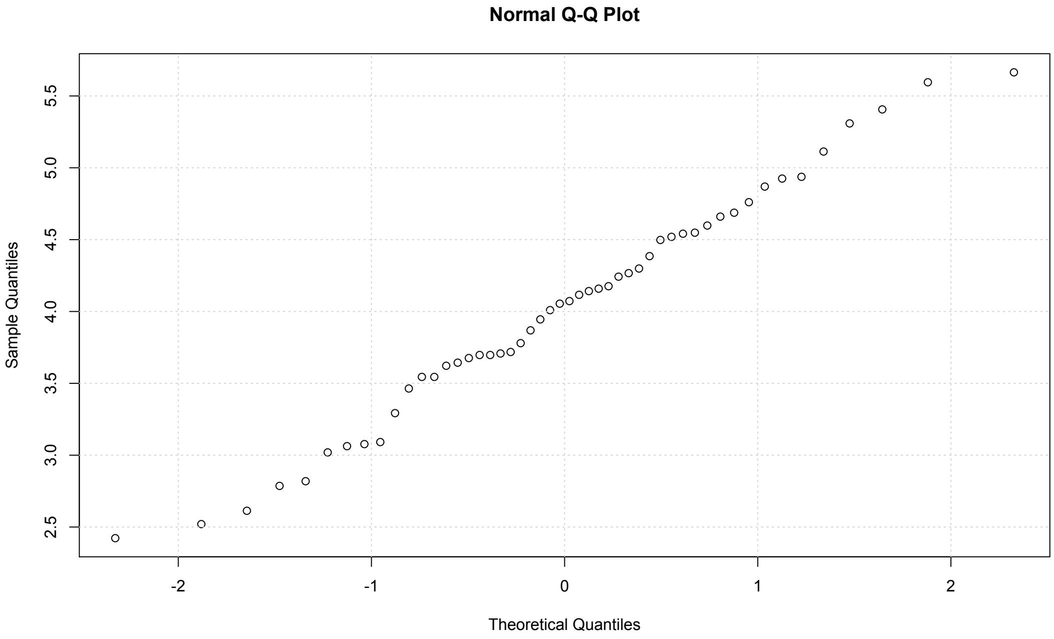


Fig 8: Normality plot for transformed variable Rape

The original variable Rape is far from Normality as shown in fig 7. However, after applying the transformation parameter λ = 0.1928036, we obtain somewhat successful normal transformation as shown in fig 8.

3. Perform PCA and find the loadings;

Solution:

The Principal Component Analysis was performed on the transformed-normalized data USArrests and the loadings were found to be:

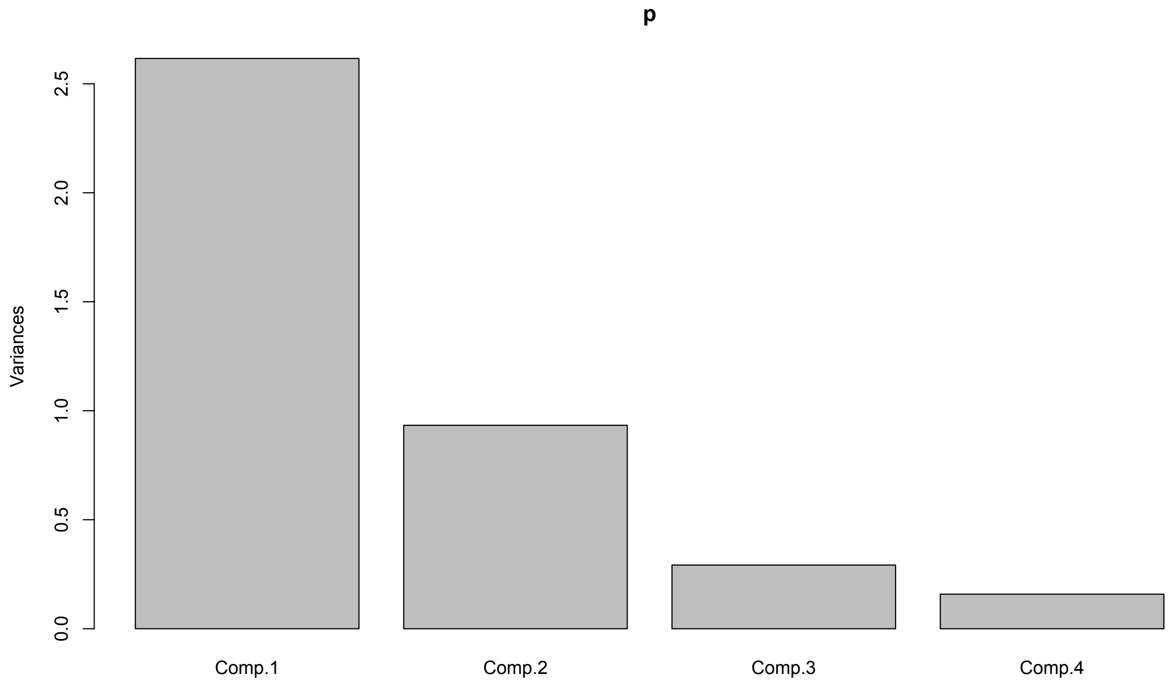
	Comp1	Comp2	Comp3	Comp4
Murder	-0.541	0.383	-0.237	0.710
Assault	-0.567	0.189	-0.430	-0.677
UrbanPop	-0.289	-0.899	-0.282	0.170
Rape	-0.550		0.824	

From this table, we can see that the first principal component will be slightly highly correlated Murder, Assault and Rape. So, these variables are essential to describe the first principal component. For the second principal component, it is highly correlated to UrbanPop meaning that this variable is important to define second principal component.

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

Solution:

The scree plot for the principal components of the normally transformed data USArrests is shown below:



As desired and expected, the variance of the first principal component is much higher than the rest of the components and it's in decreasing order.

Table below shows the correlation between the first principal component and the original variables:

Correlation

Variables	**PC 1**
Murder	-0.835
Assault	-0.899
UrbanPop	-0.47
Rape	-0.844

The first principal component has high correlations with murder, assault and rape.

5. Discuss and interpret your findings;

Solution:

A summary of the Principal Component Analysis on our dataset tells us that after the variables are transformed, the first principal component accounts for 65.4% of the total variance of the data, second describes about 23.33%, third for around 7.3% and fourth for around 3.9%.

Conclusion:

From this lab, we learnt that box-cox transformation gives us a very good estimate to find the transformation parameter that helps to transform a variable into normal variable. Principal component analysis on a normal dataset gives us more significant results in terms of analysis and interpretation of the data. Through principal component analysis, we could find that the variance of our dataset is highly concentrated on the first component. We also learnt that among the four variables, three of them are highly correlated – murder, assault and rape which fall into the category of crime. Thus, using PCA, we were able to find independent components from our dependent variables that best describes our data and the variability within it.