

How do I create a scree plot in R step-by-step?

Authored by
stats writer

April 26, 2024

RECOMMENDED CITATION

stats writer (2024). *How do I create a scree plot in R step-by-step?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=139847>

Creating a scree plot in R involves the following steps:

1. Load the necessary packages: Before creating a scree plot, you need to load the "factoextra" package which contains the necessary functions.
2. Prepare your data: The data should be in the form of a matrix or data frame with numerical values.
3. Perform a principal component analysis (PCA): Use the "prcomp" function to perform a PCA on your data. This will generate a list containing the eigenvalues and eigenvectors.
4. Extract the eigenvalues: Use the "eigen" function to extract the eigenvalues from the list generated in the previous step.
5. Calculate the cumulative proportion of variance explained: Use the "cumsum" function to calculate the cumulative proportion of variance explained by each principal component.
6. Create the scree plot: Use the "fviz_eig" function from the "factoextra" package to create the scree plot. This function takes the eigenvalues and cumulative proportion of variance as inputs.
7. Customize the plot: You can add labels, change the color or size of the points, and add a title and axis labels using the various options available in the "fviz_eig" function.

Following these steps will allow you to create a scree plot in R, which can help you visually identify the number of significant principal components in your data.

Create a Scree Plot in R (Step-by-Step)

Principal components analysis (PCA) is an that seeks to find principal components - linear combinations of the predictor variables - that explain a large portion of the variation in a dataset.

When we perform PCA, we're often interested in understanding what percentage of the total variation in the dataset can be explained by each principal

component.

One of the easiest ways to visualize the percentage of variation explained by each principal component is to create a scree plot.

This tutorial provides a step-by-step example of how to create a scree plot in R.

Step 1: Load the Dataset

For this example we'll use a dataset called USArrests, which contains data on the number of arrests per 100,000 residents in each U.S. state in 1973 for various crimes.

The following code shows how to load and view the first few rows of this dataset:

```
#load data  
data("USArrests")
```

```
#view first six rows of data  
head(USArrests)
```

```
Murder Assault UrbanPop Rape  
Alabama 13.2 236 58 21.2
```

Alaska 10.0 263 48 44.5
Arizona 8.1 294 80 31.0
Arkansas 8.8 190 50 19.5
California 9.0 276 91 40.6
Colorado 7.9 204 78 38.7

Step 2: Perform PCA

Next, we'll use the `prcomp()` function built into R to perform principal components analysis.

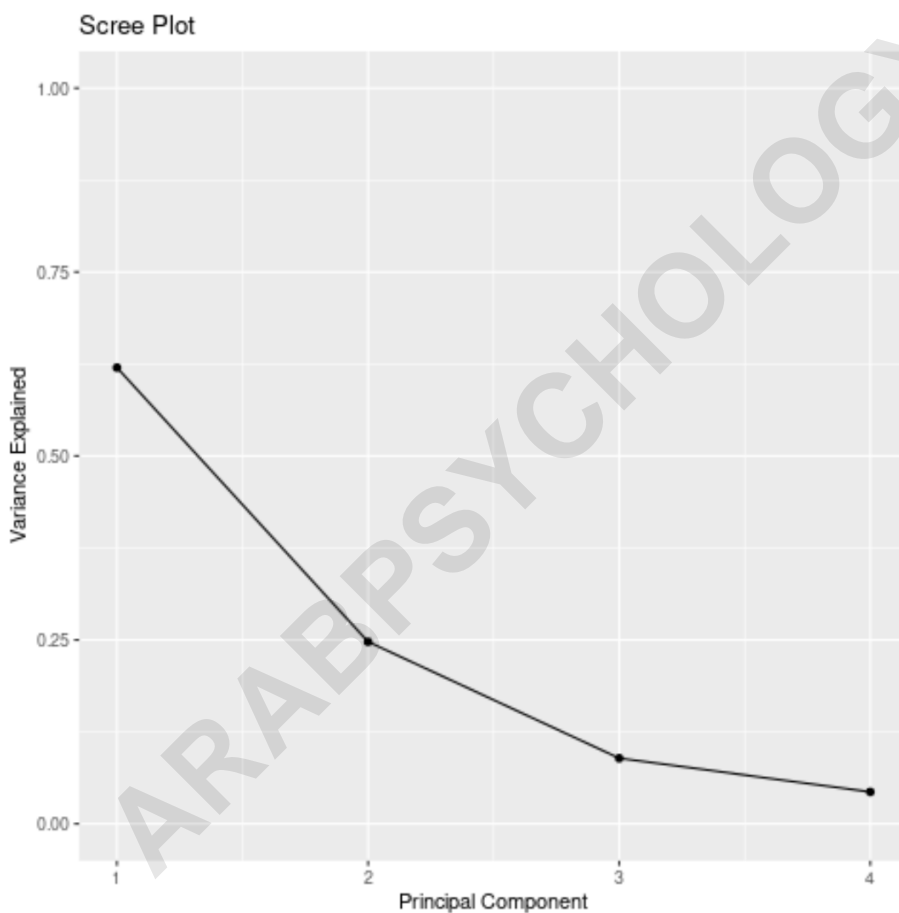
```
#perform PCA  
results <- prcomp(USArrests, scale = TRUE)
```

Step 3: Create the Scree Plot

Lastly, we'll calculate the percentage of total variance explained by each principal component and use `ggplot2` to create a scree plot:

```
#calculate total variance explained by each principal  
component  
var_explained = results$sdev^2 / sum(results$sdev^2)  
  
#create scree plot  
library(ggplot2)
```

```
qplot(c(1:4), var_explained) +  
geom_line() +  
xlab("Principal Component") +  
ylab("Variance Explained") +  
ggtitle("Scree Plot") +  
ylim(0, 1)
```



The x-axis displays the principal component and the y-axis displays the percentage of total variance explained by each individual principal component.

We can also use the following code to display the exact percentage of total variance explained by each principal component:

```
print(var_explained)
```

```
0.62006039 0.24744129 0.08914080 0.04335752
```

We can see:

The first principal component explains 62.01% of the total variation in the dataset. The second principal component explains 24.74% of the total variation in the dataset. The third principal component explains 8.91% of the total variation in the dataset. The fourth principal component explains 4.34% of the total variation in the dataset.

Notice that all of the percentages sum to 100%.

You can find more machine learning tutorials on .