Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)

LESSON 7: PRINCIPAL COMPONENT ANALYSIS (PCA)

Lesson 7: Principal Component Analysis

One technique for creating related groupings of variables is known as principal component analysis (PCA). The word principal means "main or important," and the word component can be thought of as meaning "related grouping." So PCA is an analysis that helps a data scientist mathematically find the most important related groupings of variables.

There are a variety of ways variables can be grouped, some of them being better groupings than others. In the previous grocery example, there are some ways to bag the groceries that are better than others. In the scenario in which the 100 grocery items are being delivered to several customers, grouping items by customer order might be better than grouping by household locations (such as refrigerated items, pantry items, etc.) as may have originally seemed best. The idea of PCA is to find the best way to group variables so that the new groupings give basically the same information as the individual variables.

The result from PCA is a group of columns less than or equal to the number of columns in the data set that are orthogonal to one another (meaning there is no statistical correlation between them). The columns give an indication of how the original data best groups together. You will never be selecting redundant information by using any gi

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)

LESSON 7: PRINCIPAL COMPONENT ANALYSIS (PCA)

below. In this example, the data set contains 44 variables that you should reduce.

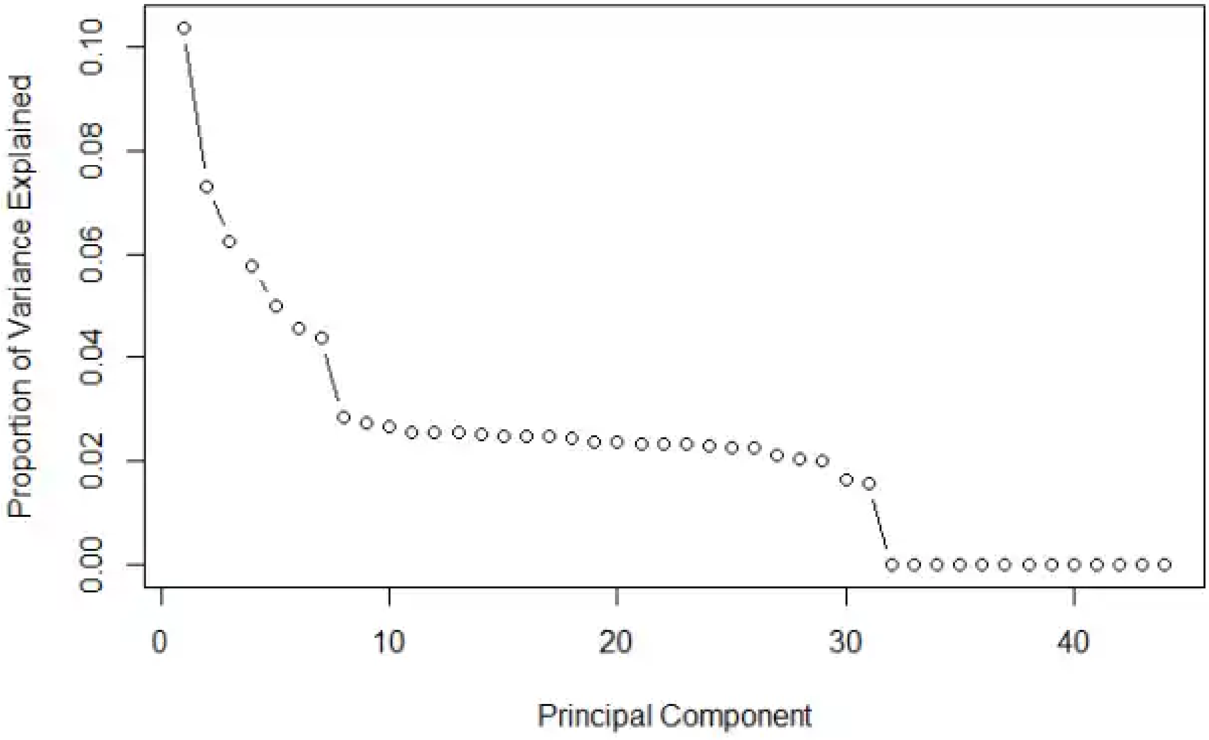

Figure 3: Sample scree plot graph