

## 1 Overview

Today's adventure will involve a segmentation analysis of teenagers. The data is derived from status updates that the teens posted on a popular social networking platform. Our marketers honed in on 37 terms that they hoped would aid in the segmentation of the sample into groups with different interests and different ideas of what was "cool." By getting to know the different types of teenage consumers, our brand hopes to find novel ways to connect with each group. Your job is to create and profile these groups in a way that lets us get to know the "prototype" of each cluster.

## 2 First things First: Explore your Data

No matter where you are going with it, the first step on your clustering adventure should be a deep exploratory dive into the data. This typically involves examining distributions - looking for signs that normalization or standardization might help space individuals out in a way that is amenable to clustering.

## 3 Processing your Data

Now that you've had a chance to explore your variables and their distributions, it's time to decide whether or not you want to use any form of standardization or normalization on your input features. Here is a list of some common approaches for data of this variety for you to choose from. There are other actions that you can take that are just as appropriate, so feel free to take liberty here if there is something else you'd like to try.

1. Do nothing. Let the data live on in its current form.
2. Transform variables, for example log transformation.
3. Employ statistical standardization (z-scores) by subtracting the mean of each column and dividing by the standard deviation of each column
4. Employ range standardization, putting each variable on a scale from 0 to 1, by transforming each column,  $x$ .
5. Standardize *each observation* so that the rows sum to 1. This would represent each individual's usage of each word as a proportion of the total number of words they used from this list.

## 4 Visualize the processed data

Once we've processed the data for input to a clustering method, we should explore once more to make sure that nothing strange happened. For this, we can repeat the steps taken in section 2 with our new data frame. We can also consider a projection onto principal components. For all but option 3 above, we can still choose between two different versions of PCA - covariance or correlation (for option 3, these two versions would be identical). If you choose to observe the data projected onto principal components, you can first plot the first two PCs of each version.

## 5 Experiment with different clustering algorithms

Experiment with different clustering algorithms. Try your hand at hierarchical clustering, k-means clustering, etc.

## 6 Determine a Final Clustering

Now that we've finished our exploration into the data and the potential number of clusters it could contain, it's time to *create the clusters*. To determine a final clustering for this adventure, choose from the algorithms that you've tried (choose clusters that makes the most sense for the problem at hand).

## 7 Create the Deliverable

At the end of the day, you want to sell your segmentation solution to the marketing team by describing your cluster profiles in a way that is honest and memorable. This means you'll have to give each cluster a name that would help any lay person soak up the profiled information, and probably create a table that highlights how each cluster stands out (or doesn't stand out) across each variable. Submit your information [here](#)

