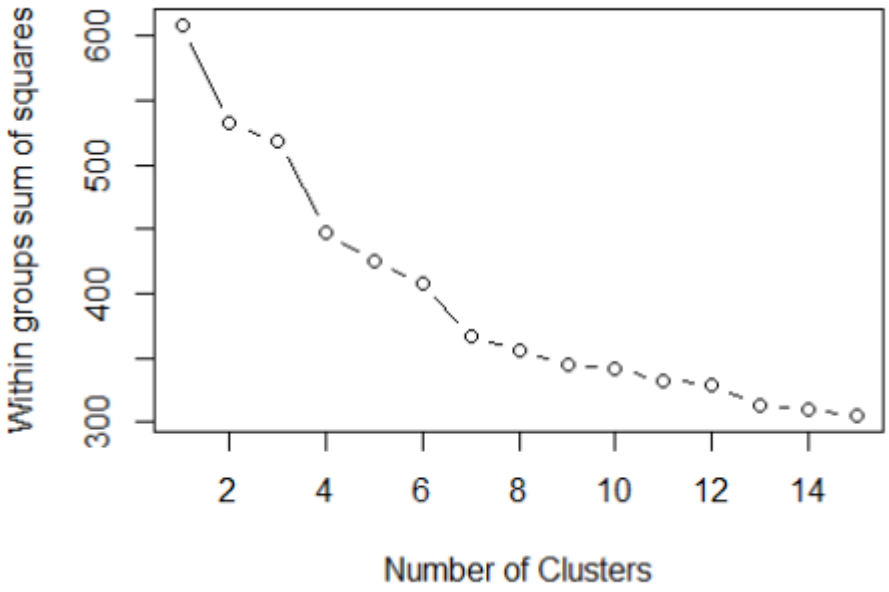
Different Approaches to Exercise 1, Problem 4

Approach 1

- First, eliminate the chatter, uncategorized, spam, and adult columns since they don't contribute much in the determination of segments
- Use k-means to cluster the data
- Identify the optimal number of clusters to use based on minimizing distances between each data point and its cluster center
-



(The optimal number of clusters is ~7)

- Look into each cluster and identify the top tweet categories by frequency %

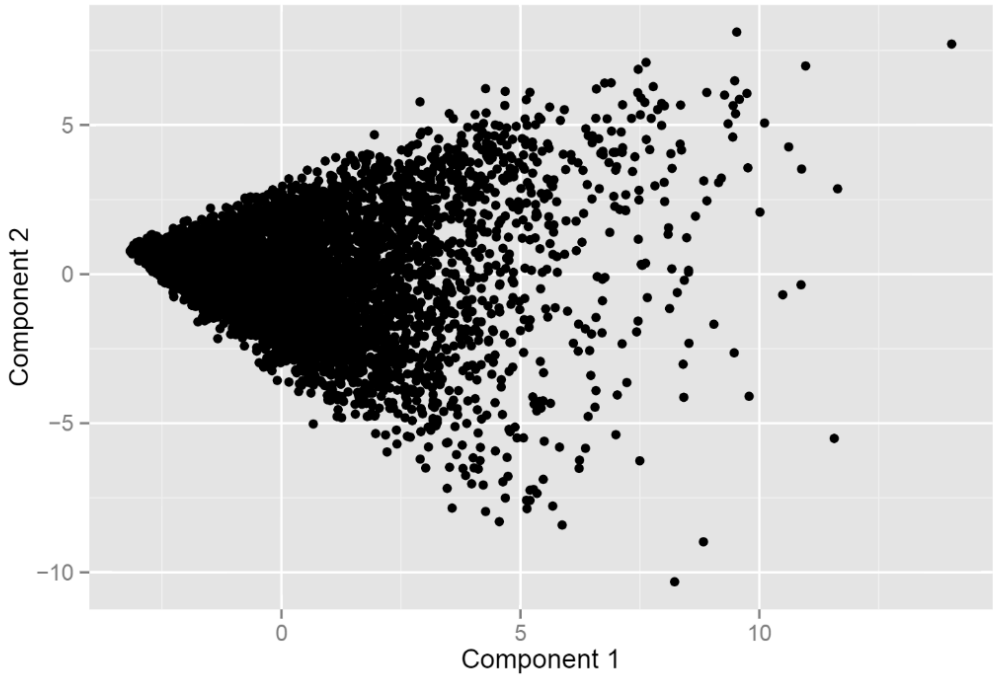
```
## [1] "College Male"

##    college_uni  online_gaming  photo_sharing sports_playing
##     0.20377040     0.19033973     0.05613751     0.04697395

## [1] "Father"

## sports_fandom      religion          food     parenting
##    0.12249805    0.09644103    0.08828672    0.07385373

## [1] "Health Enthusiast"

## health_nutrition personal_fitness          cooking    photo_sharing
##       0.23464577       0.11732524       0.05986224       0.04801825

## [1] "Post-College Arts & science Major"

##        tv_film current_events         travel            art
##     0.08894528     0.08557457     0.06205597     0.05434229
```

(We can attempt to identify each cluster based on their tweet category frequencies)

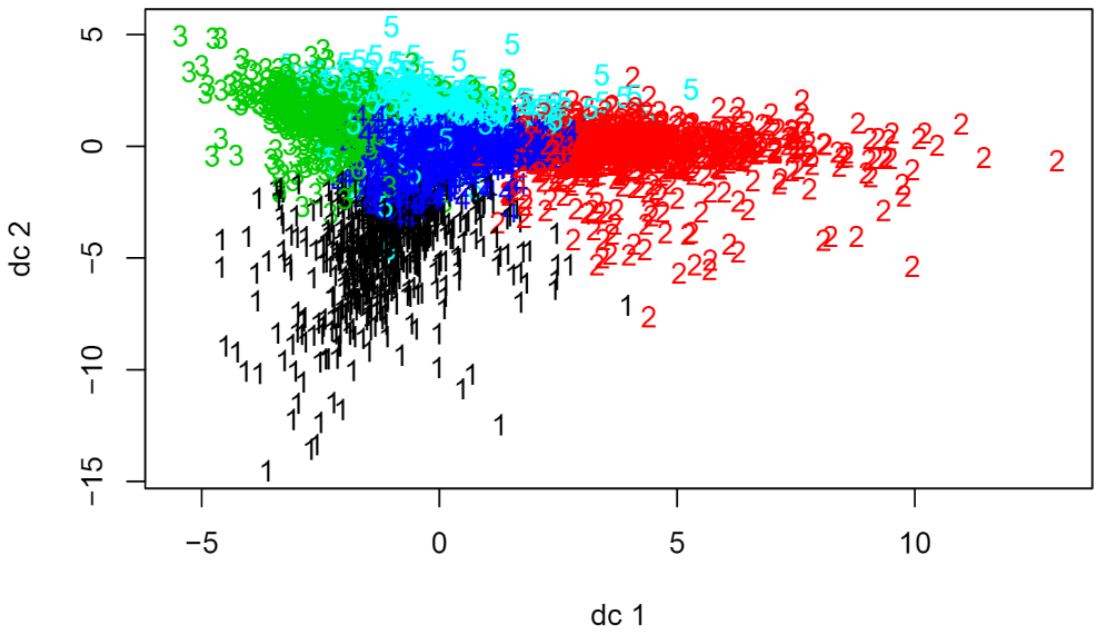
- Additionally, we can compute the distance of each person to the cluster in order to rank them by their affinity to a specific cluster

Approach 2

- Use principal component analysis(prcomp) to find principal components



- Using the principal components found in the previous step, apply k-means on the first few components to cluster data
-



- Check z-scores of each category for most common tweets (standard deviation above average)
    - This is equivalent to looking at the loadings vectors provided by prcomp

- Since clustering on all categories would introduce a lot of noise, we can first use PCA to reduce dimensions as well as noise
    - Then, we can cluster in this new dimensionally reduced space

$$For\ x_i \in \mathbb{R}^D\ , PC\ Vectors: v_1, v_2, \dots, v_5$$
$$For\ x_i:\ \ x_i^T v_1 = \alpha_{i1}\ , x_i^T v_2 = \alpha_{i2}, \dots\ ,$$
$$x_i^T v_5 = \alpha_{i5}\ \ where\ \ \alpha\ is\ the\ score\ of\ each\ x$$

- We can also look at the scores to see which users are best correlated to which each component

Key Differences Between Principal Component Analysis and Clustering

- Clustering is a single membership process:
    - Points are assigned to one —and only one —cluster
- Principal Component Analysis is a mixed membership process

What does mixed membership mean in the context of PCA?

- PCA is like a recipe
    - Factors/principal components are like ingredients
    - Scores are like amount of ingredient to be included
    - Mix of ingredients implies mix membership
- PCA Output:
    - Factors/principal components:
      $V_1, \ldots, V_k \in \mathbb{R}^D$
    - Scores:
      α_{ij} for i=1,…, N & j=1,…, K
    - Reconstruction:
      $X_i \approx \alpha_{i1}V_1 + \alpha_{i2}V_2 + \ldots + \alpha_{ik}V_k$
        - The vectors are the “ingredients” and the alphas tell us their quantity