Week 3

R Packages

The packages you will need to install for the week are VIM, lessR, ggplot2, cluster, fpc, wbstats, and NbClust.

Cluster Analysis

In the previous two weeks, we have looked at two supervised learning methods: regression and decision tree. For Week 3, we are going to examine our first unsupervised learning method: clustering. Unsupervised learning is 'unsupervised' because we do not have a target (outcome variable).

Clustering is meant to be used for "knowledge discovery" instead of "prediction." The basis of clustering is what sociologists call "homophily"-or birds of the same feather flock together. The goal of clustering is to find groups, or clusters, in a data set. We want to partition our dataset so that observations within each group are similar to each other while observations in different groups are different from each other.

There are many clustering algorithms, which are based on many different approaches of grouping data points. We will examine the two most common approaches in this class: 1) partitioning and 2) hierarchical. The partitioning approach divides the dataset into multiple partitions. The hierarchical approach disaggregates the dataset into a tree structure (similar to decision trees). We will look at two partitioning methods: k-means and k-medoids. We will talk about k-means in class and briefly discuss k-medoids. One of the tasks for this week's homework assignment is for you to research the k-medoids methods.

Learning Goal for the Week

What interesting things can we learn from online postings of 30,000 teenagers on a social media site? (Notice that we have no target/outcome variable. We am simply looking for interesting patterns.)

The Dataset

We will use the Teen Market Segmentation dataset from Chapter 9 in the Lantz textbook. According to Lantz, the dataset is a random sample of 30,000 U.S. high school students who had profiles on a social networking service (SNS) in 2006. The full text of the SNS profiles were downloaded. Each teen's gender, age, and number of SNS friends were recorded. From the top 500 words that appeared across all SNS profiles, a smaller list of 36 words were chosen to represent five categories of interest: extracurricular activities, fashion, religion, romance, and antisocial behavior (Lantz 2013, p. 279).

Getting Started

```
setwd("C:/Users/corylowe/OneDrive/Code/R Practice Code/Applied Data Mining_Portfolio/Week 3")
teens<-read.csv("snsdata.csv", header=TRUE, sep=",")
str(teens)</pre>
```

```
## 'data.frame':
                   30000 obs. of 40 variables:
##
                 $ gradyear
                 : Factor w/ 2 levels "F", "M": 2 1 2 1 NA 1 1 2 1 1 ...
   $ gender
## $ age
                       19 18.8 18.3 18.9 19 ...
   $ friends
                 : int
                       7 0 69 0 10 142 72 17 52 39 ...
##
   $ basketball : int 0000000000...
   $ football
                        0 1 1 0 0 0 0 0 0 0 ...
                 : int
##
   $ soccer
                 : int
                        0 0 0 0 0 0 0 0 0 0 ...
##
   $ softball
                 : int
                        0 0 0 0 0 0 0 1 0 0 ...
##
   $ volleyball : int
                        0 0 0 0 0 0 0 0 0 0 ...
   $ swimming
                 : int
                        0 0 0 0 0 0 0 0 0 0 ...
##
                        0 0 0 0 0 0 0 0 0 0 ...
   $ cheerleading: int
##
   $ baseball
                 : int
                        0 0 0 0 0 0 0 0 0 0 ...
##
   $ tennis
                        0 0 0 0 0 0 0 0 0 0 ...
                 : int
##
   $ sports
                 : int
                        0 0 0 0 0 0 0 0 0 0 ...
##
   $ cute
                 : int
                        0 1 0 1 0 0 0 0 0 1 ...
##
                        0 0 0 0 1 1 0 2 0 0 ...
   $ sex
                 : int
##
                 : int
                        0 0 0 0 0 0 0 1 0 0 ...
   $ sexy
                        0 0 0 0 0 0 0 0 0 1 ...
##
   $ hot
                 : int
##
   $ kissed
                 : int
                        0 0 0 0 5 0 0 0 0 0 ...
##
   $ dance
                 : int
                        1 0 0 0 1 0 0 0 0 0 ...
   $ band
                        0 0 2 0 1 0 1 0 0 0 ...
                 : int
                 : int
##
                        0 0 0 0 0 1 1 0 0 0 ...
   $ marching
                        0 2 1 0 3 2 0 1 0 1 ...
##
   $ music
                 : int
##
                        0 2 0 1 0 0 0 1 0 1 ...
   $ rock
                 : int
   $ god
                 : int
                        0 1 0 0 1 0 0 0 0 6 ...
##
   $ church
                        0 0 0 0 0 0 0 0 0 0 ...
                 : int
                 : int
                        0 0 0 0 0 0 0 0 0 2 ...
##
   $ jesus
##
  $ bible
                        0 0 0 0 0 0 0 0 0 0 ...
                 : int
   $ hair
                 : int
                        0600100001...
##
   $ dress
                 : int
                        0 4 0 0 0 1 0 0 0 0 ...
##
   $ blonde
                 : int
                        0 0 0 0 0 0 0 0 0 0 ...
##
   $ mall
                 : int
                        0 1 0 0 0 0 2 0 0 0 ...
                        0 0 0 0 2 1 0 0 0 1 ...
##
   $ shopping
                 : int
##
   $ clothes
                 : int
                        0 0 0 0 0 0 0 0 0 0 ...
   $ hollister
##
                        0000002000...
                 : int
  $ abercrombie : int
                        0 0 0 0 0 0 0 0 0 0 ...
##
  $ die
                 : int
                        0 0 0 0 0 0 0 0 0 0 ...
##
   $ death
                 : int
                        0 0 1 0 0 0 0 0 0 0 ...
##
                        0 0 0 0 1 1 0 0 0 0 ...
   $ drunk
                 : int
                        0 0 0 0 1 0 0 0 0 0 ...
   $ drugs
                 : int
```

Exploratory Data Analysis

```
summary(teens)
```

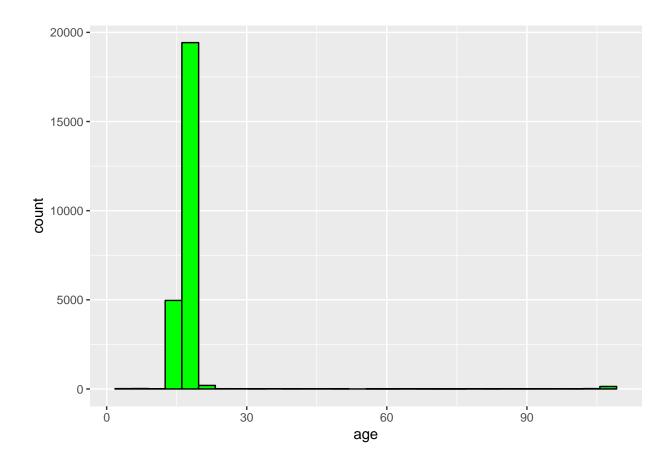
```
##
       gradyear
                    gender
                                                       friends
                                      age
##
    Min.
           :2006
                   F
                        :22054
                                        : 3.086
                                                           : 0.00
                                 Min.
                                                   Min.
                                 1st Qu.: 16.312
    1st Qu.:2007
                   Μ
                        : 5222
                                                    1st Qu.: 3.00
    Median:2008
                   NA's: 2724
                                 Median : 17.287
                                                   Median : 20.00
##
    Mean
           :2008
                                        : 17.994
                                                           : 30.18
                                 Mean
                                                   Mean
```

```
3rd Qu.:2008
                               3rd Qu.: 18.259
                                                 3rd Qu.: 44.00
##
   Max. :2009
                               Max.
                                      :106.927
                                                Max.
                                                        :830.00
                               NA's
                                      :5086
##
##
     basketball
                        football
                                                           softball
                                           soccer
##
   Min. : 0.0000
                     Min. : 0.0000
                                       Min. : 0.0000
                                                        Min. : 0.0000
##
   1st Qu.: 0.0000
                     1st Qu.: 0.0000
                                       1st Qu.: 0.0000
                                                        1st Qu.: 0.0000
   Median: 0.0000
                     Median: 0.0000
                                       Median : 0.0000
                                                        Median: 0.0000
   Mean : 0.2673
                                                        Mean : 0.1612
                     Mean : 0.2523
                                       Mean : 0.2228
##
    3rd Qu.: 0.0000
                     3rd Qu.: 0.0000
                                       3rd Qu.: 0.0000
                                                         3rd Qu.: 0.0000
##
   Max. :24.0000
                     Max. :15.0000
                                       Max. :27.0000
                                                        Max. :17.0000
##
##
     volleyball
                                                          baseball
                        swimming
                                       cheerleading
   Min. : 0.0000
                     Min. : 0.0000
##
                                       Min. :0.0000
                                                       Min. : 0.0000
                                       1st Qu.:0.0000
##
   1st Qu.: 0.0000
                     1st Qu.: 0.0000
                                                        1st Qu.: 0.0000
   Median : 0.0000
                     Median : 0.0000
                                       Median :0.0000
                                                       Median : 0.0000
##
   Mean : 0.1431
                     Mean : 0.1344
                                       Mean :0.1066
                                                       Mean : 0.1049
##
   3rd Qu.: 0.0000
                     3rd Qu.: 0.0000
                                       3rd Qu.:0.0000
                                                        3rd Qu.: 0.0000
##
   Max. :14.0000
                     Max. :31.0000
                                       Max. :9.0000
                                                       Max. :16.0000
##
##
       tennis
                          sports
                                          cute
                                                            sex
                                                       Min. : 0.0000
##
   Min. : 0.00000
                      Min. : 0.00
                                      Min. : 0.0000
    1st Qu.: 0.00000
                      1st Qu.: 0.00
                                      1st Qu.: 0.0000
                                                        1st Qu.: 0.0000
   Median : 0.00000
                      Median: 0.00
                                      Median : 0.0000
                                                       Median: 0.0000
##
   Mean : 0.08733
                      Mean : 0.14
                                      Mean : 0.3229
                                                       Mean : 0.2094
   3rd Qu.: 0.00000
##
                      3rd Qu.: 0.00
                                      3rd Qu.: 0.0000
                                                        3rd Qu.: 0.0000
   Max. :15.00000
                      Max. :12.00
                                      Max. :18.0000
                                                       Max. :114.0000
##
##
       sexy
                          hot
                                          kissed
                                                            dance
##
   Min. : 0.0000
                     Min. : 0.0000
                                       Min. : 0.0000
                                                        Min. : 0.0000
                                                         1st Qu.: 0.0000
   1st Qu.: 0.0000
                                       1st Qu.: 0.0000
                     1st Qu.: 0.0000
##
   Median : 0.0000
                     Median : 0.0000
                                       Median : 0.0000
                                                         Median: 0.0000
   Mean : 0.1412
                     Mean : 0.1266
                                       Mean : 0.1032
                                                         Mean : 0.4252
                                       3rd Qu.: 0.0000
                                                         3rd Qu.: 0.0000
   3rd Qu.: 0.0000
                     3rd Qu.: 0.0000
   Max. :18.0000
                     Max. :10.0000
                                       Max. :26.0000
                                                         Max. :30.0000
##
##
##
        band
                        marching
                                          music
                                                             rock
                                       Min. : 0.0000
   Min. : 0.0000
                     Min. : 0.0000
                                                        Min. : 0.0000
##
   1st Qu.: 0.0000
                     1st Qu.: 0.0000
                                       1st Qu.: 0.0000
                                                         1st Qu.: 0.0000
##
   Median : 0.0000
                     Median : 0.0000
                                       Median : 0.0000
                                                         Median: 0.0000
   Mean : 0.2996
##
                     Mean : 0.0406
                                       Mean : 0.7378
                                                        Mean : 0.2433
    3rd Qu.: 0.0000
                     3rd Qu.: 0.0000
                                       3rd Qu.: 1.0000
                                                         3rd Qu.: 0.0000
                                       Max. :64.0000
##
   Max. :66.0000
                     Max. :11.0000
                                                        Max. :21.0000
##
##
                         church
                                          jesus
                                                           bible
        god
                     Min. : 0.0000
                                       Min. : 0.0000
                                                        Min. : 0.00000
   Min. : 0.0000
   1st Qu.: 0.0000
                     1st Qu.: 0.0000
                                       1st Qu.: 0.0000
                                                         1st Qu.: 0.00000
##
   Median : 0.0000
                     Median : 0.0000
                                       Median : 0.0000
                                                         Median: 0.00000
##
   Mean : 0.4653
                     Mean : 0.2482
                                       Mean : 0.1121
                                                         Mean : 0.02133
   3rd Qu.: 1.0000
                     3rd Qu.: 0.0000
                                       3rd Qu.: 0.0000
                                                         3rd Qu.: 0.00000
##
   Max. :79.0000
                     Max. :44.0000
                                       Max. :30.0000
                                                         Max. :11.00000
##
##
        hair
                         dress
                                         blonde
                                                            mall
   Min. : 0.0000
                     Min. :0.000
                                     Min. : 0.0000
                                                       Min. : 0.0000
                                     1st Qu.: 0.0000
   1st Qu.: 0.0000
                     1st Qu.:0.000
                                                       1st Qu.: 0.0000
```

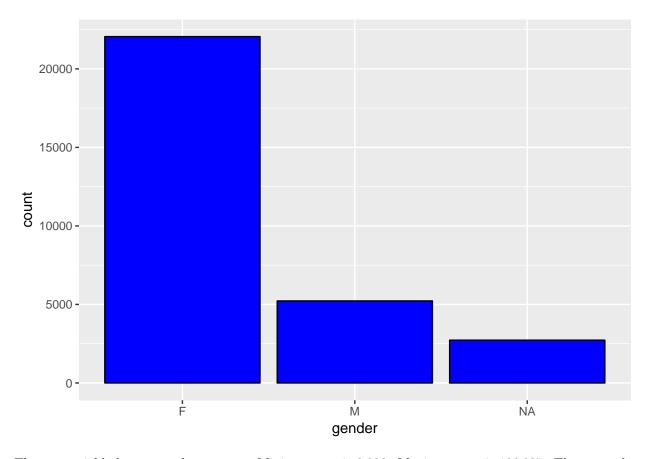
```
## Median : 0.0000
                   Median :0.000
                                   Median : 0.0000
                                                     Median : 0.0000
## Mean
                          :0.111
                                   Mean
         : 0.4226 Mean
                                          : 0.0989
                                                     Mean
                                                           : 0.2574
                                   3rd Qu.: 0.0000
   3rd Qu.: 0.0000
                    3rd Qu.:0.000
                                                     3rd Qu.: 0.0000
## Max.
          :37.0000
                           :9.000
                                          :327.0000
                                                     Max.
                                                            :12.0000
                    Max.
                                   Max.
##
##
                      clothes
                                     hollister
                                                     abercrombie
      shopping
   Min. : 0.000
                          :0.0000
                                   Min.
                                          :0.00000
                                                    Min.
                                                           :0.00000
                   Min.
   1st Qu.: 0.000
                   1st Qu.:0.0000
                                   1st Qu.:0.00000
                                                     1st Qu.:0.00000
##
##
   Median : 0.000
                   Median :0.0000
                                   Median :0.00000
                                                    Median :0.00000
##
  Mean
         : 0.353
                         :0.1485
                                          :0.06987
                                                     Mean
                                                           :0.05117
                   Mean
                                   Mean
   3rd Qu.: 1.000
                   3rd Qu.:0.0000
                                    3rd Qu.:0.00000
                                                     3rd Qu.:0.00000
##
  Max. :11.000
                   Max.
                         :8.0000
                                   Max.
                                          :9.00000
                                                     Max.
                                                           :8.00000
##
##
                                                          drugs
        die
                        death
                                         drunk
##
        : 0.0000
                    Min. : 0.0000
                                     Min.
                                            :0.00000
                                                      Min. : 0.00000
  Min.
##
   1st Qu.: 0.0000
                    1st Qu.: 0.0000
                                     1st Qu.:0.00000
                                                       1st Qu.: 0.00000
## Median : 0.0000
                    Median : 0.0000
                                     Median :0.00000
                                                      Median : 0.00000
## Mean
         : 0.1841
                    Mean : 0.1142
                                     Mean
                                           :0.08797
                                                      Mean : 0.06043
## 3rd Qu.: 0.0000
                    3rd Qu.: 0.0000
                                     3rd Qu.:0.00000
                                                      3rd Qu.: 0.00000
                                                      Max. :16.00000
## Max. :22.0000
                    Max. :14.0000
                                     Max. :8.00000
##
```

Problematic Data Values

```
library (ggplot2)
ggplot(data=teens) + geom_histogram(aes(x=age), fill="green", color="black")
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 5086 rows containing non-finite values (stat_bin).
```



ggplot(data=teens) + geom_bar(aes(x=gender), fill="blue", color="black")



The age variable has a very large range. Minimum age is 3.086. Maximum age is 106.927. There are also $5{,}086$ missing values.

The gender variable has 2,724 missing values. We should also note the gender distribution: 22,054 females and 5,222 males.

Let's see the percentage of missing values for our variables:

```
pMiss <- function(x){sum(is.na(x))/length(x)*100}
apply(teens,2,pMiss)</pre>
```

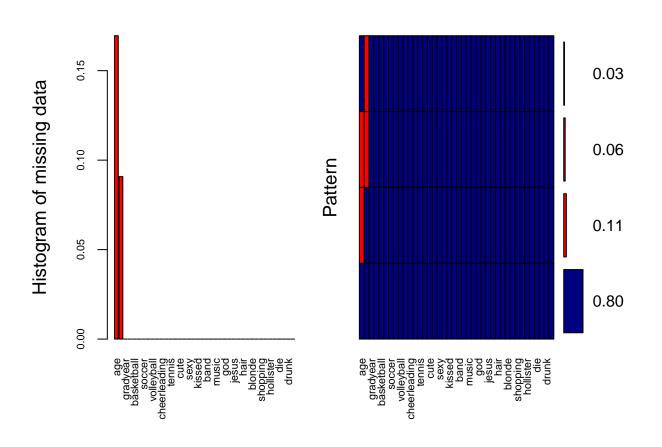
##	gradyear	gender	age	friends	basketball
##	0.00000	9.08000	16.95333	0.00000	0.00000
##	football	soccer	softball	volleyball	swimming
##	0.00000	0.00000	0.00000	0.00000	0.00000
##	cheerleading	baseball	tennis	sports	cute
##	0.00000	0.00000	0.00000	0.00000	0.00000
##	sex	sexy	hot	kissed	dance
##	0.00000	0.00000	0.00000	0.00000	0.00000
##	band	marching	music	rock	god
##	0.00000	0.00000	0.00000	0.00000	0.00000
##	church	jesus	bible	hair	dress
##	0.00000	0.00000	0.00000	0.00000	0.00000
##	blonde	mall	shopping	clothes	hollister
##	0.00000	0.00000	0.00000	0.00000	0.00000
##	abercrombie	die	death	drunk	drugs
##	0.00000	0.00000	0.00000	0.00000	0.00000

Let's visualize what we just found above:

library(VIM)

```
## Loading required package: colorspace
## Loading required package: grid
## Loading required package: data.table
## VIM is ready to use.
## Since version 4.0.0 the GUI is in its own package VIMGUI.
##
## Please use the package to use the new (and old) GUI.
## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues
##
## Attaching package: 'VIM'
## The following object is masked from 'package:datasets':
##
## sleep
```

aggr_plot <- aggr(teens, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(teens), cex</pre>



```
##
##
    Variables sorted by number of missings:
                      Count
##
        Variable
             age 0.1695333
##
##
          gender 0.0908000
##
        gradyear 0.0000000
         friends 0.0000000
##
      basketball 0.0000000
##
##
        football 0.0000000
##
          soccer 0.0000000
##
        softball 0.0000000
      volleyball 0.0000000
##
##
        swimming 0.0000000
    cheerleading 0.0000000
##
##
        baseball 0.0000000
##
          tennis 0.0000000
##
          sports 0.0000000
##
            cute 0.0000000
##
             sex 0.0000000
##
            sexy 0.0000000
##
             hot 0.0000000
          kissed 0.0000000
##
           dance 0.0000000
##
            band 0.0000000
##
##
        marching 0.0000000
##
           music 0.0000000
            rock 0.0000000
##
             god 0.0000000
##
##
          church 0.0000000
##
           jesus 0.0000000
##
           bible 0.0000000
##
            hair 0.0000000
##
           dress 0.0000000
          blonde 0.0000000
##
##
            mall 0.0000000
##
        shopping 0.0000000
##
         clothes 0.0000000
##
       hollister 0.0000000
     abercrombie 0.0000000
##
##
             die 0.0000000
           death 0.0000000
##
           drunk 0.0000000
##
           drugs 0.0000000
```

Source: The two code chunks above are from this entry from Rblogger.

16% of data values for age is missing. 9% of data values for gender is missing. If we compound the fact that some people did not report their true age, this variable is our "bigger" problem. Let's tackle it first.

One Proble at a Time: Recoding Age via Imputation

First, we need to make an assumption: **Teenagers are between the age of 13 and 20.** Anyone who does not have a reported age in this assumed range will be recoded as "NA."

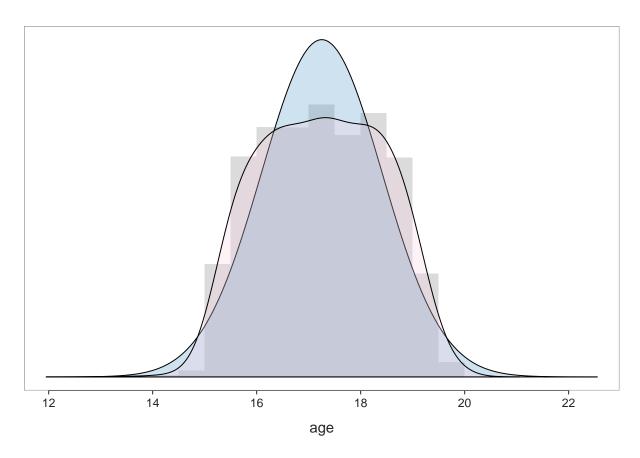
To handle the missing age values, we will use imputation. It is common to impute missing values with expected values (i.e. what we expect those values to be). Mean and median imputations are common techniques. If the distribution is normal, we use mean imputation. If the distribution is skewed, we use median imputation.

We will use a package called lessR to draw a histogram of the age and then superimposes a normal curve on top for comparison purpose.

library(lessR)

Density(age, data=teens)

```
##
## lessR 3.7.9
                   feedback: gerbing@pdx.edu
                                                  web: lessRstats.com/new
## 1. mydata <- Read("")
                              Read text, Excel, SPSS, SAS or R data file
## 2. Help()
                              Get help
## 3. hs(), bc(), or ca()
                              All histograms, all bar charts, or both
## 4. Plot(X) or Plot(X,Y)
                              For continuous and categorical variables
                              numerical X: Violin, Box, Scatter plot
## 5. by1= , by2=
                              Trellis graphics, a plot for each by1, by2
## 6. reg(Y ~ X, Rmd="eg")
                              Regression + R markdown file that, when
##
                              knit, provides full interpretative output
## 7. style("lightbronze")
                              Return to previous, more neutral theme
      style(show=TRUE)
                              all color/style options and current values
##
## 8. getColors()
                              create many types of color palettes
##
## Attaching package: 'lessR'
## The following object is masked from 'package:data.table':
##
##
       set
```



```
##
##
## Sample Size: 24477
## Missing Values: 5523
##
## Density bandwidth for general curve: 0.3580
\ensuremath{\mbox{\#\#}} For a smoother curve, increase bandwidth with option: bw
##
## Sample size out of range for Shapiro-Wilk normality test.
##
##
   (Box plot) Outliers: 5
##
##
## Small
                Large
## ----
   13.0
##
   13.1
##
##
   13.1
##
    13.1
    13.4
```

The distribution looks normal. Let's proceed with mean imputation.

```
# Finding the mean age by cohort

mean(teens$age) # Doesn't work b/c of NA
```

```
## [1] NA
mean(teens$age, na.rm = TRUE) #This tells R to ignore NA in calculating the mean.
## [1] 17.25243
# Review age by cohort
aggregate(data = teens, age ~ gradyear, mean, na.rm = TRUE)
##
     gradyear
                   age
## 1
         2006 18.65586
## 2
         2007 17.70617
## 3
         2008 16.76770
## 4
         2009 15.81957
# Calculating the expected age for each person
# This creates a new variable called ave_age
ave_age <- ave(teens$age, teens$gradyear,</pre>
                 FUN = function(x) mean(x, na.rm = TRUE))
#print(ave_age) #To view average age table created above.
teens$age <- ifelse(is.na(teens$age), ave_age, teens$age)</pre>
#Removes the missing values and replaces with mean age.
# Check to make sure missing values are eliminated
summary(teens$age)
##
                    Median
                               Mean 3rd Qu.
      Min. 1st Qu.
                                               Max.
```

Second Problem: Missing Gender Values

17.24

17.24

13.03

F

22054 5222 2724

M <NA>

16.28

We have three possible levels: female, male, and NA (no reported gender). We will create two dummy variables to handle the gender missing values: 1) female and 2) no_gender.

20.00

18.21

```
table(teens$female, useNA = "ifany")

##

## 0 1

## 7946 22054

table(teens$no_gender, useNA = "ifany") #We have 2,724 cases of unknown gender. This matches up with ou

##

## 0 1

## 27276 2724
```

What Do We Want to Examine?

We want to cluster what these 30,000 teenagers talked about on their SNS profiles with regards to the five categories of interests: extracurricular activities, fashion, religion, romance, and antisocial behavior.

```
interests <- teens[5:40] #Take the 5th through the 40th variables into the model.
```

Let's Talk Cluster Analysis

Partitioning Approach

General process:

- 1. Choose the number of clusters (k)
- 2. Partition the dataset into k clusters so that the sum of squared distances is minimized between the data points (p) and some center point [c(i)] in each cluster.

Two questions naturally arise from above:

Question 1: How do we determine the center points?

Answer: We select a clustering algorithm. We will examine k-means and k-medoids.

Question 2: How do you measure the distance between the data points and center points?

Answer: We use either Euclidean (straight line) or Manhattan distance (city block).

K-Means Clustering

We will begin by building a cluster model with five clusters. There's no right place to start. Just pick a k value that you think is most suitable and start.

Remember that in k-means, the starting centroids are randomly chosen.

nstart is the number of times the starting points are re-sampled. Think of it this way: R does clustering assignment for each data point 25 times and picks the center that have the lowest within cluster variation. The "best" centroids become the starting point by which kmeans will continue to iterate. Typically you can set nstart to between 20 and 25 to find the best overall random start. See Morissette & Chartier (2013)

paper for explanations of the different kmeans algorithms. We recommen reviewing Table 5 in the paper for additional information on the various kmeans algorithm.

iter.max = maximum number of iterations before stopping (unless convergence is already achieved before max iterations).

The default algorithm is Hartigan-Wong, which minimizes the within-cluster sum of squares.

```
set.seed(123)
teen_clusters_5 <- kmeans(interests, centers=5)</pre>
```

Let's see what are the outputs from kmeans:

```
names(teen_clusters_5)

## [1] "cluster"    "centers"    "withinss"

## [5] "tot.withinss" "betweenss"    "size"    "iter"

## [9] "ifault"
```

Size: Number of people in each cluster. Cluster 3 has the most number of people. Follows by Clusters 5 & 1.

```
teen_clusters_5$size
```

773

4500

Let's see each row and its assigned cluster.

```
teen_clusters_5$cluster
```

4216 1538 18973

[1]

```
##
   ##
  [35] 3 3 3 3 3 4 3 3 5 1 3 1 5 2 3 1 3 1 5 3 3 3 1 4 3 3 3 3 4 3 1 3 3 1
##
  ##
  [103] 1 5 1 3 3 3 5 1 3 3 3 1 3 3 4 3 5 3 3 2 3 3 3 3 2 3 1 3 3 2 5 2 3 3
  [137] 5 3 5 3 1 2 1 3 3 3 2 3 2 3 3 3 3 3 3 3 3 3 2 3 3 3 5 5 3 3 3 5 5
##
  [171] 5 3 3 3 3 1 3 5 4 2 3 3 1 3 3 1 2 5 5 3 1 3 3 3 4 1 1 3 3 3 3 3 1 3
##
##
  ##
##
  ##
  [307] 3 2 3 3 3 1 2 1 3 5 3 3 2 3 3 3 3 2 3 3 1 1 1 3 3 3 3 5 1 3 1 3 2 3
  ##
##
  [375] 1 3 3 3 3 3 3 1 3 1 2 3 3 3 3 3 2 3 3 1 4 2 3 3 5 3 3 3 3 3 3 3 3 3 3
##
  ##
  ##
  [511] 3 3 3 3 3 3 3 3 5 5 3 3 3 5 5 3 3 3 1 3 4 3 1 3 3 1 5 1 3 1 3 1 5 3 3
##
##
  [545] 3 3 5 1 3 3 3 3 3 3 3 2 3 3 3 1 5 2 5 5 1 3 3 3 3 3 3 3 3 3 1 5 3
  ##
  ##
##
  [681] 3 2 3 3 3 3 3 3 3 1 3 3 3 2 3 3 3 5 5 1 3 2 3 3 1 3 3 3 1 1 3 3 2 3
##
  ##
```

```
[9929] 3 3 3 5 1 3 3 5 2 3 3 3 5 4 3 3 3 5 1 3 3 3 4 3 3 1 3 3 5 5 1 3 3 1
  [9963] 2 1 3 3 3 3 3 1 3 5 5 3 3 1 3 3 3 1 5 3 3 3 5 1 3 5 3 3 3 5 3 3 3 1
##
  ## [10065] 3 3 3 3 2 5 3 3 5 1 3 3 3 3 5 3 3 4 3 1 1 3 3 5 3 3 3 3 3 3 3 3 5 3
## [10201] 3 3 3 3 3 3 3 3 3 5 3 3 3 3 3 5 3 3 3 5 5 3 3 5 5 3 3 3 5 5
## [10269] 3 3 3 5 3 5 3 1 3 3 5 3 3 3 3 1 3 1 3 5 5 3 3 4 1 3 5 3 3 4 3 3 1 5
## [10303] 2 1 5 3 3 2 1 3 3 5 1 3 3 3 1 3 3 3 3 1 1 1 5 3 2 3 5 1 3 3 1 2 2 3
## [10337] 3 2 3 3 1 5 2 2 3 3 3 1 3 5 2 5 5 1 5 1 2 3 3 5 3 3 3 3 3 1 3 3 1 5
## [10405] 1 3 3 3 3 3 3 1 1 2 5 5 3 5 3 2 3 3 3 3 1 5 3 4 3 3 5 3 5 3 3 2 3 5 3
## [10473] 5 3 5 5 2 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 1 4 3 1 2 3 2 3 1 3 3
## [10541] 1 2 1 5 3 5 1 3 3 3 5 5 3 5 3 3 1 3 3 3 3 5 3 3 3 5 3 3 5 4 3 3 3
## [10609] 1 1 3 3 3 5 3 3 1 3 3 2 3 3 5 1 2 2 1 3 3 1 3 1 1 3 1 3 3 5 1 3 1 5
## [10643] 1 3 1 1 3 3 5 3 3 3 3 3 3 3 3 4 1 3 1 3 3 1 3 3 5 3 3 5 3 3 3 3 3
## [10677] 3 3 3 3 3 3 3 3 3 2 3 1 3 5 3 3 3 3 3 3 3 3 3 3 3 3 2 3 5 3 3 1 3 3 3
## [10711] 3 3 3 1 3 3 5 3 3 3 5 3 5 4 3 5 3 3 5 2 3 3 3 5 5 3 3 3 3 5 2 3 3
## [10745] 1 3 3 5 3 3 1 3 3 3 1 3 3 5 3 3 3 1 3 3 3 1 5 2 3 3 3 5 1 2 3 3
## [10779] 3 1 3 2 1 3 5 3 3 3 3 3 3 3 3 3 3 3 3 3 5 2 1 3 2 3 3 5 3 1 5
## [10813] 3 3 3 3 3 1 3 5 3 3 3 4 3 5 4 3 1 1 3 3 3 3 3 3 3 3 3 3 4 3 1 1 3 5
## [10949] 5 5 1 3 3 3 2 4 3 3 1 3 1 3 3 3 2 3 5 1 3 3 1 3 3 5 3 3 3 5 3 5
## [10983] 3 3 1 5 1 3 3 5 3 5 3 5 5 3 1 5 3 5 3 3 5 3 3 5 3 3 5 3 3 5 3 1 3 5 3
## [11017] 3 3 3 1 5 3 3 5 5 3 3 5 5 3 3 3 2 3 1 3 1 3 5 3 3 1 1 5 3 5 3 3 5 1
## [11051] 5 2 1 3 3 3 3 3 5 3 3 3 1 3 3 2 3 3 3 3 1 1 3 3 3 1 5 3 3 3 3 5 3
## [11085] 1 3 1 3 5 1 3 2 1 5 3 3 3 3 3 3 3 3 3 3 3 5 3 3 3 3 3 2 3 5 3 3 3
## [11119] 3 5 3 5 3 3 1 3 1 3 3 3 1 5 2 1 3 5 3 3 3 5 1 3 5 3 4 3 3 3 5 3 5 3
## [11153] 5 3 3 2 3 3 3 1 5 2 3 4 5 1 5 3 1 3 5 1 3 2 3 4 1 3 5 3 2 1 3 2 5 3
## [11187] 1 3 3 3 3 3 3 4 3 1 3 1 1 3 3 3 3 3 5 3 3 3 3 5 3 3 4 5 1 3 3 3 1
## [11221] 3 1 3 3 3 3 1 3 1 3 3 3 4 3 1 3 3 3 5 5 4 5 5 3 3 3 3 3 3 3 3 3 3
## [11255] 3 3 3 5 1 3 2 3 3 4 3 3 5 3 3 3 1 3 5 3 3 3 3 3 3 5 2 3 1 1 3 3 3 3
## [11289] 3 3 3 3 3 3 3 3 3 3 3 3 5 5 3 3 5 3 2 3 3 3 1 3 5 3 5 1 1 3 3 3 5 3 3
## [11323] 3 3 5 3 3 3 3 1 5 3 5 3 3 3 5 3 3 3 2 3 3 3 3 3 3 4 3 3 3 5 3 1 1 3
## [11391] 1 3 3 5 3 1 3 2 3 1 3 3 5 3 3 3 3 3 5 3 3 5 3 1 3 3 3 1 2 2 5
## [11425] 3 5 5 3 3 5 3 3 3 3 3 3 1 3 3 3 2 1 3 3 3 3 5 3 3 3 2 3 3 3 3 3 3
## [11459] 3 2 1 3 1 3 4 3 3 3 3 3 3 3 3 5 5 3 5 1 3 3 5 4 5 3 2 3 3 3 3 3 3 3
## [11527] 5 2 3 3 3 1 3 5 1 3 5 2 2 3 3 3 3 2 4 3 1 5 3 5 3 3 3 1 3 3 3 3 3 5
## [11629] 2 3 3 3 3 3 3 1 3 3 3 3 2 2 3 3 3 3 5 3 3 3 3 1 3 1 4 3 3 1 1 3 3 3 3
## [11663] 3 3 3 3 3 3 3 3 3 1 5 2 3 1 3 3 3 3 5 5 3 5 3 3 3 3 1 3 3 3 5 1
## [11697] 3 3 2 3 1 3 5 1 3 3 2 3 1 5 1 1 3 5 5 1 3 3 5 3 1 3 1 1 1 4 1 1 3 1
## [11731] 1 1 3 3 5 3 3 3 3 5 3 3 3 5 2 3 3 1 3 3 1 3 5 1 2 2 1 5 3 3 2 3 2 3
```

```
## [20979] 3 1 3 3 1 3 2 1 3 3 3 3 3 3 3 2 3 5 3 3 5 3 3 1 1 3 3 1 5 3 3 3 5 1 1
## [21013] 3 3 1 3 5 3 3 3 2 3 3 3 5 3 5 3 4 4 5 3 3 3 3 3 1 3 1 3 3 3 3 3 3 3 1
## [21047] 3 5 3 1 3 3 3 3 3 3 3 5 1 1 3 3 2 3 3 3 3 5 5 3 5 5 5 3 3 1 1 3 3
## [21081] 3 3 1 3 3 5 1 3 3 1 5 1 3 3 5 5 5 2 4 3 3 3 3 3 1 3 3 3 5 1 3 2 3
## [21149] 3 5 5 1 3 3 1 3 4 3 3 3 3 3 5 3 3 3 1 3 3 2 2 3 3 3 3 1 3 3 3 2 1
## [21285] 4 3 5 5 1 3 3 3 3 3 3 3 4 3 3 3 3 3 5 2 3 1 3 3 3 3 3 3 3 5 3 3 3
## [21319] 3 3 3 3 3 3 5 3 3 3 3 5 3 4 3 2 3 3 3 3 3 3 3 1 5 5 3 1 5 1 3 3 1
## [21387] 3 3 3 5 1 3 3 3 3 3 5 1 3 1 5 4 5 2 3 3 4 1 3 2 3 5 3 3 5 3 2 5 5
## [21523] 3 3 3 4 3 3 3 3 5 3 2 1 3 3 3 3 1 5 3 3 5 3 3 5 3 5 3 1 3 3 5
## [21557] 3 3 5 1 1 2 3 3 3 5 3 3 3 1 3 1 3 5 3 3 3 5 1 3 3 3 5 3 3 5 3 1 3 3
## [21591] 1 3 4 3 5 3 5 3 3 2 5 3 3 3 3 1 3 1 3 3 5 3 3 2 3 3 3 2 3 5 3 1 1 5
## [21625] 3 1 3 3 5 3 3 3 1 3 5 1 3 3 3 3 3 2 2 1 1 3 4 2 3 3 3 3 3 3
## [21659] 3 3 4 3 3 3 2 5 3 4 5 5 3 3 5 5 3 1 5 1 5 3 3 3 3 5 5 3 3 5 5 3 2 5 5
## [21727] 3 3 3 3 5 3 2 3 3 3 3 3 5 5 3 3 5 5 5 3 3 1 5 3 3 5 5 3 3 3 5 5 3 1
## [21761] 2 3 3 3 3 2 1 5 3 3 3 3 3 3 3 3 3 3 3 3 3 5 5 3 3 3 5 5 3 3 5 1 3 3
## [21795] 5 2 3 3 3 4 3 3 3 3 3 3 2 5 3 1 3 3 3 3 3 3 2 3 3 5 3 3 3 3 2
## [21829] 5 3 3 5 4 3 3 3 4 3 4 1 3 3 5 3 3 3 1 3 3 3 3 3 3 1 3 2 3 5 3 3 3
## [21897] 4 3 3 1 3 3 4 5 5 5 3 2 3 4 2 3 3 5 3 3 3 3 1 5 3 3 3 5 3 3 5 1 3
## [21931] 3 1 3 3 3 3 3 4 3 3 1 1 4 5 5 4 3 2 3 3 2 3 1 3 2 3 3 5 3 3 5 3 3 5
## [21965] 3 5 3 3 3 3 3 4 3 3 3 1 5 3 1 5 3 5 3 3 3 2 3 3 5 1 2 3 3 3 5 5 3 3
## [21999] 3 3 3 5 5 5 3 3 5 5 5 4 3 3 2 3 5 1 4 3 3 3 5 3 5 3 5 3 5 3 2 3 3 3 3 3
## [22033] 3 3 3 5 3 5 3 3 3 1 3 5 3 3 3 3 4 3 5 3 3 3 1 1 3 5 1 3 3 5 3 4 3
## [22101] 3 3 5 4 5 5 3 3 5 3 3 3 4 5 3 1 5 3 2 5 1 3 3 3 3 3 3 5 5 2 5 3 5 3
## [22135] 3 2 3 3 3 3 1 5 5 1 3 1 5 1 3 3 3 3 5 1 5 3 3 2 3 3 3 3 3 3 3 3 3 3
## [22169] 5 3 3 3 5 5 3 3 5 5 3 5 3 5 3 3 3 1 3 3 2 3 3 1 3 3 3 1 1 3 5 1 2 3 3 3 3
## [22203] 3 3 5 3 3 3 3 3 5 1 3 5 4 1 3 3 3 3 3 3 5 2 2 3 3 3 3 5 4 3 2 3 3
## [22237] 3 5 3 5 1 3 3 3 1 3 3 1 3 5 1 4 3 3 3 5 3 3 5 4 3 1 3 5 5 5 3 1 3 3
## [22305] 1 3 3 1 3 3 3 4 3 3 3 3 1 1 3 3 1 5 3 3 3 2 3 3 3 1 4 3 5 5 2 3 3
## [22441] 3 3 3 3 3 3 5 5 3 1 3 5 3 3 3 1 3 5 1 2 1 5 1 3 4 3 3 3 1 5 1 4 5 3
## [22509] 3 3 3 3 3 3 3 3 5 5 4 5 4 1 3 5 2 3 3 3 3 3 3 4 3 3 3 5 3 3 1 5 5
## [22543] 1 2 3 3 3 3 3 2 3 3 4 3 3 5 2 5 3 3 3 3 3 5 1 3 3 3 5 3 5 5 5 5 3 3 3
## [22577] 3 3 3 5 3 3 1 3 1 3 3 3 5 3 3 4 5 3 3 3 1 3 2 3 5 5 4 5 3 3 1 3 3
## [22611] 3 1 2 5 2 5 4 5 5 3 3 3 3 3 3 3 3 3 3 3 3 5 3 3 1 3 3 5 3 3 3 5 3
## [22645] 5 3 5 5 3 5 1 1 3 3 3 3 3 3 5 1 3 3 5 5 3 3 3 5 4 3 3 1 3 3 3 5 3
## [22679] 3 3 5 1 3 3 1 4 3 3 3 3 1 5 1 3 1 3 3 3 5 3 3 1 1 3 3 5 1 5 3 1 3 3
## [22713] 3 3 4 5 1 3 3 3 3 1 5 3 3 3 1 1 3 5 3 1 3 1 1 3 3 1 3 3 4 3 3 1 1 1
```

```
## [26453] 3 4 3 3 3 5 3 3 5 3 3 3 2 3 2 2 2 4 5 5 3 1 3 3 3 3 1 5 2 1 5 3 3 2
## [26521] 3 5 3 5 3 3 3 5 5 3 3 5 5 3 3 5 3 3 3 5 3 3 3 3 3 3 3 5 5 5 1 1 5
## [26555] 3 3 1 3 3 2 3 3 3 5 1 3 3 5 3 3 2 3 3 5 3 3 3 3 5 2 3 3 3 1 3 2 2 3
## [26589] 1 1 3 5 2 3 4 5 5 3 3 5 3 3 5 1 5 5 3 2 3 3 3 4 3 5 3 3 3 3 3 3 5 3
## [26623] 5 1 3 3 5 3 3 3 5 5 1 3 4 3 3 1 5 3 5 1 3 5 5 3 3 1 4 3 1 3 1 1 3 3
## [26657] 3 5 3 3 5 3 3 3 1 1 3 4 2 2 3 3 3 3 3 5 3 4 5 3 2 1 1 3 1 3 3 3 3 3
## [26691] 2 3 5 4 1 3 5 3 5 1 5 3 4 3 4 5 5 5 3 3 3 2 5 3 3 1 3 1 1 5 5 5 3 3
## [26725] 3 3 3 3 3 1 1 3 3 3 1 5 3 1 3 5 3 3 3 2 3 1 5 5 3 1 5 5 5 3 3 3 3 3
## [26793] 3 3 3 3 4 3 5 3 3 5 5 5 3 3 1 3 1 2 3 3 3 3 3 3 4 2 3 5 3 3 3 5 5 1
## [26827] 5 3 4 1 3 3 1 3 3 3 2 3 3 5 5 1 2 5 3 2 3 3 3 3 3 3 3 1 5 3 1 3 3 3
## [26861] 3 3 3 3 3 3 1 3 3 3 5 3 3 3 3 3 3 1 2 3 5 3 5 2 3 5 3 3 3 3 3 4 1 5 3
## [26963] 5 3 5 3 5 5 5 5 5 3 3 3 2 3 5 3 3 3 1 3 5 1 3 5 5 3 3 5 3 3 5 3 3 5
## [26997] 3 3 3 3 5 3 5 2 4 2 3 3 3 3 1 5 3 3 3 3 3 3 5 3 5 5 5 3 2 4 4 3 3 3 1
## [27031] 3 5 3 3 5 3 5 4 3 1 3 3 3 5 3 3 5 5 3 1 3 4 3 3 5 5 5 3 2 3 3 3 3 1
## [27065] 3 4 3 1 1 3 3 5 3 3 3 3 3 3 1 2 3 5 3 3 2 3 3 3 1 3 5 3 3 2 2 3 3 1
## [27099] 5 5 2 3 3 1 3 3 2 3 3 3 3 3 3 4 3 1 5 3 3 5 5 3 3 3 3 5 5 3 3 3 1 1
## [27133] 1 3 3 3 5 4 3 3 1 3 1 3 3 3 3 2 3 3 5 3 3 1 3 3 3 3 3 3 3 3 1 3 5 2
## [27201] 2 3 3 3 5 5 3 3 1 3 3 3 1 3 3 3 2 3 3 3 4 3 3 3 2 3 3 1 3 3 3 3 1 3
## [27235] 1 1 1 5 3 5 3 3 3 1 3 2 3 5 5 5 5 5 5 5 1 3 3 5 5 1 4 1 5 3 4 3 5 5 1
## [27405] 3 1 3 3 3 2 3 5 3 5 3 3 3 3 5 3 1 1 2 1 3 3 3 3 3 3 2 3 3 5 3 1 1 3
## [27473] 3 5 3 3 1 1 1 1 3 5 2 2 5 3 5 5 1 5 3 1 4 1 5 1 3 1 3 2 3 3 1 1 5 1
## [27507] 3 5 2 3 3 5 1 3 3 3 3 3 3 3 3 5 1 2 3 5 3 5 3 3 3 3 3 5 3
## [27541] 3 3 1 3 3 3 5 5 5 3 2 5 5 1 3 3 1 3 3 5 5 1 3 3 3 3 2 1 5 1 3 5 3
## [27575] 3 3 3 1 5 3 3 5 5 3 3 2 5 5 3 5 3 3 2 5 3 3 5 5 3 3 1 3 3 5 3 1 5
## [27609] 1 3 5 3 5 4 3 3 5 2 1 3 3 3 5 3 5 3 5 1 3 5 3 3 3 3 3 3 3 1 4 3 3
## [27643] 3 3 3 1 3 5 3 3 3 3 5 3 5 3 1 5 3 5 3 3 1 2 3 3 3 5 3 3 3 3 3 5 3 5
## [27677] 3 3 5 3 3 3 3 5 5 5 5 3 3 3 3 3 3 1 2 1 3 3 5 5 3 5 5 3 5 3 3 3 3 2
## [27711] 3 3 3 1 3 4 3 4 1 3 3 3 3 3 3 5 2 3 5 3 3 2 3 3 5 5 3 3 2 4 3 3 2
## [27745] 5 5 1 5 3 3 1 3 4 5 3 1 3 4 3 3 3 1 3 3 5 3 5 1 5 3 3 3 4 5 1 3 2
## [27779] 5 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 5 3 5 1 3 3 3 3 2 3 5 5 3 3 3 5 3 1 1 3
## [27813] 3 2 3 1 3 4 3 3 3 3 3 3 3 3 2 1 5 3 5 3 1 3 3 3 5 3 5 3 5 5 5 3 2 3
## [27847] 3 3 3 1 3 3 1 3 3 3 5 3 3 3 1 3 3 5 2 3 1 2 3 3 3 5 3 1 3 3 3 1 3 5
## [27881] 5 3 3 3 3 1 1 3 1 1 3 3 3 3 3 3 3 5 1 3 3 5 1 3 3 1 3 5 3 1 3 2 3
## [27983] 5 2 3 5 5 3 3 3 5 5 5 3 5 1 5 5 3 5 4 3 3 5 5 2 3 5 5 5 1 2 5 5 5 3
## [28017] 5 5 3 3 5 3 3 3 4 2 3 3 3 3 5 3 1 5 3 1 1 3 5 3 5 3 3 1 3 3 2 3 3 1
## [28051] 3 1 3 3 1 3 3 3 3 5 3 3 3 3 1 3 2 3 4 3 3 5 3 3 3 3 5 2 3 3 2 3 1 3
## [28085] 2 5 3 3 3 3 3 5 1 3 4 3 5 5 1 5 3 3 1 4 3 5 3 3 1 1 3 3 3 5 3 5 3 3
## [28119] 3 1 1 1 3 5 3 3 3 3 3 3 5 3 2 3 1 5 3 3 5 3 3 2 3 3 2 3 2 3 5 3 3 3
## [28187] 5 5 5 5 3 3 3 3 3 3 3 3 3 5 1 3 3 3 4 5 3 3 1 5 3 3 1 3 3 3 3 3 3 3 3
## [28221] 3 4 3 1 3 3 1 5 5 3 3 3 5 5 2 5 1 2 3 3 1 4 4 3 1 1 3 1 5 3 1 3 3 3
## [28255] 5 3 1 3 3 2 5 3 4 1 3 5 3 5 3 5 4 3 3 1 3 3 3 3 5 1 3 3 3 3 5 3 3 5
```

```
## [28357] 5 2 5 5 5 2 3 3 2 3 3 5 5 5 5 3 3 3 3 1 1 5 3 3 1 3 3 3 5 3 1 4 3 5
## [28391] 3 3 3 3 3 3 3 5 3 3 5 5 5 2 4 4 3 3 3 4 1 5 5 1 3 3 4 3 2 2 3 3 5 5
## [28425] 1 3 5 3 3 3 3 1 3 1 5 3 1 3 3 5 3 1 3 3 2 3 3 5 2 5 1 1 5 3 5 3
## [28493] 1 3 3 5 5 3 3 3 1 5 3 2 1 5 3 3 3 5 2 3 3 5 2 3 3 1 3 3 3 3 3 1 3
## [28527] 3 3 3 3 3 3 2 3 1 5 3 3 3 3 2 5 1 3 1 5 3 1 3 2 3 1 2 4 3 5 5 3 3 3 3
## [28561] 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 5 3 3 1 2 3 5 3 3 1 3 3 1 3 4 3 3 5 3
## [28629] 5 3 1 5 5 3 5 3 3 3 3 3 3 3 3 3 1 5 3 2 5 3 3 2 3 5 3 3 3 2 1 3 3 1 3
## [28731] 3 1 3 5 3 2 3 3 2 3 5 3 3 5 3 1 3 5 3 3 3 5 5 5 3 2 3 3 1 3 5 5 2 3
## [28799] 3 1 5 3 3 3 1 4 3 3 3 3 3 3 2 3 5 3 5 1 1 1 3 3 3 3 5 3 3 1 3 3 5 3
## [28833] 5 3 3 5 5 3 1 1 5 3 3 3 5 4 3 3 3 1 2 5 5 4 3 3 3 3 3 3 3 3 5 3 3 3 1
## [28867] 3 3 5 3 1 5 3 1 3 5 5 1 3 3 3 3 3 3 3 5 4 3 3 2 3 3 3 1 2 3 5 5
## [28901] 5 3 3 3 3 3 5 3 4 2 3 5 3 5 5 5 3 1 5 1 3 3 3 3 3 5 3 5 3 3 1 3 3 5
## [28935] 3 1 3 3 5 5 3 3 1 3 3 5 3 5 3 4 3 3 3 3 4 3 4 5 3 5 3 1 2 3 1 5 4 3
## [28969] 3 3 1 3 5 5 2 3 4 5 5 3 3 5 2 2 1 2 3 3 3 3 3 3 5 5 5 3 3 5 5 5 3 3 5
## [29003] 3 5 1 3 1 3 4 2 4 3 3 3 3 3 5 5 3 3 3 5 5 3 3 1 3 5 2 4 3 5 5 5 3 3 3
## [29037] 5 3 3 1 3 1 3 3 3 5 3 1 2 5 3 3 5 5 5 5 5 3 3 3 3 3 2 5 5 5 5 3 5 5 3 3
## [29071] 3 3 2 3 3 3 5 3 5 3 5 3 3 4 3 1 5 4 5 3 3 3 4 1 3 1 3 5 2 1 3 5 3 5 5 3
## [29105] 1 3 3 3 5 3 5 3 3 5 5 3 3 5 3 3 5 3 3 5 3 5 5 3 5 3 5 4 5 5 3 3 2 1
## [29139] 3 3 3 5 5 3 3 3 3 3 3 1 3 3 5 3 3 3 1 3 5 2 1 4 5 3 1 2 3 3 5 3
## [29173] 3 3 1 3 3 3 3 5 3 2 3 3 3 3 4 4 3 5 3 3 3 1 1 1 5 5 3 5 5 3 2 3 3 4
## [29275] 5 3 3 3 3 3 5 5 5 3 3 3 2 3 5 3 5 1 3 3 3 5 3 1 5 3 5 3 3 3 1 3 3
## [29309] 5 3 5 3 5 3 4 1 3 3 3 3 3 5 1 3 3 3 3 3 3 3 3 1 3 3 1 3 5 5 3 3 1 5
## [29343] 3 3 5 5 5 2 3 3 3 4 3 3 3 5 5 3 3 3 4 3 3 5 2 3 3 3 1 2 2 4 1 3 2
## [29377] 2 3 2 3 3 5 3 1 5 5 5 2 3 3 3 3 3 3 3 2 1 5 2 3 1 5 1 2 3 3 5 3 3
## [29411] 3 5 3 3 2 2 3 3 3 3 5 3 5 3 3 3 1 3 3 3 3 3 1 3 3 2 3 5 3 3 5 5 5 3
## [29445] 3 5 2 3 5 1 1 3 1 3 3 5 4 5 3 3 5 3 3 1 4 3 5 3 5 2 3 1 3 3 3 3 1 5
## [29479] 3 5 5 3 3 5 3 5 4 2 3 3 2 1 3 5 3 4 3 3 3 3 4 3 3 3 3 3 3 5 3 3 3 5
## [29513] 3 3 5 3 5 3 3 2 3 2 1 1 3 3 3 3 3 3 3 3 5 3 3 3 5 3 3 5 3 3 5 3 3 5 3 3 5 5
## [29547] 3 5 5 5 1 5 3 3 3 4 3 5 5 5 3 2 3 3 3 5 3 3 3 5 5 3 5 5 3 3 5 5 1 3
## [29581] 3 3 3 4 5 5 3 3 5 5 3 5 5 3 5 3 5 1 3 3 3 1 5 2 3 3 1 3 3 5 3 5 3
## [29615] 3 3 1 1 3 3 5 3 3 3 3 1 3 3 5 3 3 3 1 2 5 4 3 3 5 5 1 5 3 3 1 3 3
## [29649] 5 4 3 5 1 5 4 3 3 5 5 5 3 5 2 3 3 5 3 3 3 5 1 5 5 3 1 3 2 3 3 3 3 3
## [29683] 3 3 1 3 4 4 5 3 3 1 5 5 5 3 2 5 2 3 3 1 5 5 3 3 5 4 3 5 3 3 3 3 3 5
## [29717] 3 3 4 3 3 3 3 3 3 3 3 3 5 5 3 3 3 1 3 3 5 5 3 3 3 4 1 3 3 3 5 5 3 3
## [29751] 3 3 3 3 3 3 3 3 3 5 3 5 3 5 3 3 3 3 1 4 3 3 3 3 1 5 3 1 1 3 2 3 5
## [29785] 3 3 5 1 5 5 5 5 3 3 5 3 3 3 3 5 1 3 3 5 5 3 1 3 3 1 1 1 2 3 3 5 1
## [29819] 1 1 3 3 3 2 5 5 3 4 4 3 1 3 5 3 3 5 3 3 5 1 3 3 3 5 1 3 3 5 3 3 5 3 3
## [29853] 3 3 3 3 3 3 1 3 1 3 4 4 3 3 3 3 5 3 3 2 3 3 3 3 1 3 4 3 3 5 3 1 3 3 3
## [29887] 3 3 5 5 5 2 3 3 3 3 3 2 5 3 3 3 3 3 3 5 5 1 3 3 5 3 3 2 1 3 3 1 3
## [29955] 3 3 3 3 3 3 1 1 3 3 1 5 5 3 3 3 3 5 1 2 3 3 3 1 4 5 3 3 3 5 3 1 3
## [29989] 3 5 3 2 1 3 3 3 3 3 3 1
```

Let's show the coordinates of the cluster centroids for the interest variables.

```
##
     basketball football
                             soccer softball volleyball
     0.1883302 0.2326850 0.1209677 0.1034156 0.09582543 0.15417457
## 2 0.3185956 0.2964889 0.1462939 0.1579974 0.13524057 0.15669701
     0.1282876 0.1536394 0.0829600 0.0764771 0.08090444 0.09007537
     0.7968952 0.7102199 0.4954722 0.3635188 0.26908150 0.32341527
     0.8191111 0.5928889 0.8868889 0.5388889 0.43088889 0.26266667
     cheerleading
                    baseball
                                 tennis
                                            sports
## 1
       0.06783681 0.09369070 0.10578748 0.14302657 0.2599620 0.18477230
## 2
       0.15734720 0.08972692 0.08777633 0.15149545 0.4785436 0.18985696
       0.06920361 0.07046856 0.06967796 0.08996996 0.1638117 0.08970642
## 3
## 4
       0.29236740 0.38809832 0.16429495 0.62483829 0.9495472 3.40491591
## 5
       0.25155556 \ 0.21733333 \ 0.13111111 \ 0.26066667 \ 0.8915556 \ 0.19488889
##
          sexy
                      hot.
                              kissed
                                         dance
                                                    band
                                                            marching
## 1 0.1425522 0.11835863 0.11480076 0.3394213 1.1892789 0.17314991 2.4305028
## 2 0.2067620 0.20221066 0.09947984 3.1105332 0.3498049 0.03706112 0.9174252
## 3 0.1078902 0.08164233 0.03895009 0.1891635 0.1044115 0.01644442 0.2829284
## 4 0.5148771 0.34152652 1.43596378 1.0944373 0.8576973 0.06597671 2.7024580
## 5 0.1937778 0.26111111 0.13555556 0.4677778 0.1760000 0.01511111 0.6711111
                     god
          rock
                            church
                                       jesus
                                                  bible
                                                             hair
## 1 0.4295541 0.4070209 0.2277040 0.1159867 0.02371917 0.3275617 0.12286528
## 2 0.3250975 3.0988296 0.7711313 0.5845254 0.12418726 0.5156047 0.24187256
## 3 0.1369841 0.2381279 0.1186423 0.0675697 0.01054130 0.1504243 0.06177199
## 4 1.4100906 0.9456662 0.4036223 0.1461837 0.03492885 4.4773609 0.37645537
## 5 0.2888889 0.4951111 0.6080000 0.1286667 0.02711111 0.9306667 0.21688889
##
                                       clothes hollister abercrombie
         blonde
                     mall shopping
## 1 0.06499051 0.2020873 0.3244782 0.15085389 0.04625237 0.03130930
## 2 0.10078023 0.3114434 0.4902471 0.23732120 0.07932380 0.07022107
## 3 0.03868655 0.1360354 0.1999157 0.07452696 0.04095293 0.03025352
## 4 1.28978008 0.7839586 0.6727038 0.84346701 0.28201811 0.24450194
## 5 0.17955556 0.7117778 0.9233333 0.30844444 0.17422222 0.11822222
                    death
           die
                               drunk
                                          drugs
## 1 0.2151328 0.14136622 0.08870968 0.07068311
## 2 0.3582575 0.22041612 0.12288687 0.06306892
## 3 0.1227007 0.07884889 0.05760818 0.02529911
## 4 0.9573092 0.42432083 0.57956016 0.67917206
## 5 0.2215556 0.14844444 0.11888889 0.09177778
```

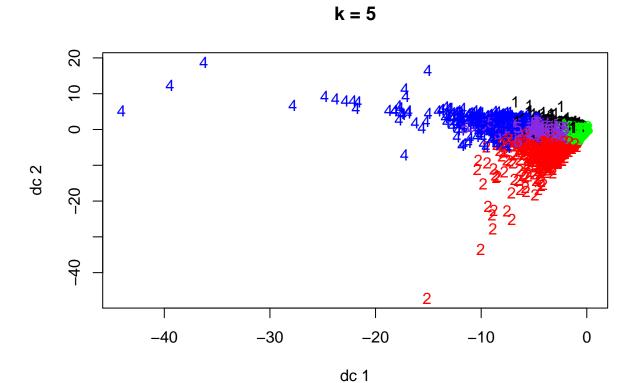
t(teen_clusters_5\$centers) #transpose for ease of reading purpose

```
5
##
                                    2
                         1
                                                3
                0.18833017 0.31859558 0.12828757 0.79689521 0.81911111
## basketball
## football
                0.23268501 0.29648895 0.15363938 0.71021992 0.59288889
                0.12096774 0.14629389 0.08296000 0.49547219 0.88688889
## soccer
## softball
                0.10341556 0.15799740 0.07647710 0.36351876 0.53888889
                0.09582543 0.13524057 0.08090444 0.26908150 0.43088889
## volleyball
                0.15417457 0.15669701 0.09007537 0.32341527 0.26266667
## swimming
## cheerleading 0.06783681 0.15734720 0.06920361 0.29236740 0.25155556
## baseball
                0.09369070 0.08972692 0.07046856 0.38809832 0.21733333
## tennis
                0.10578748 0.08777633 0.06967796 0.16429495 0.13111111
                0.14302657 0.15149545 0.08996996 0.62483829 0.26066667
## sports
```

```
0.25996205 0.47854356 0.16381173 0.94954722 0.89155556
## sex
               0.18477230 0.18985696 0.08970642 3.40491591 0.19488889
## sexy
               0.14255218 0.20676203 0.10789016 0.51487710 0.19377778
                0.11835863 0.20221066 0.08164233 0.34152652 0.26111111
## hot
## kissed
                0.11480076 0.09947984 0.03895009 1.43596378 0.13555556
               0.33942125 3.11053316 0.18916355 1.09443726 0.46777778
## dance
               1.18927894 0.34980494 0.10441153 0.85769728 0.17600000
## band
               0.17314991 0.03706112 0.01644442 0.06597671 0.01511111
## marching
## music
                2.43050285 0.91742523 0.28292837 2.70245796 0.67111111
                0.42955408 0.32509753 0.13698414 1.41009056 0.28888889
## rock
## god
                0.40702087 3.09882965 0.23812787 0.94566624 0.49511111
                0.22770398 0.77113134 0.11864228 0.40362225 0.60800000
## church
## jesus
                0.11598672 0.58452536 0.06756970 0.14618370 0.12866667
## bible
                0.02371917 0.12418726 0.01054130 0.03492885 0.02711111
               0.32756167 0.51560468 0.15042429 4.47736093 0.93066667
## hair
## dress
               0.12286528 0.24187256 0.06177199 0.37645537 0.21688889
               0.06499051 0.10078023 0.03868655 1.28978008 0.17955556
## blonde
## mall
                0.20208729 0.31144343 0.13603542 0.78395860 0.71177778
               0.32447818 0.49024707 0.19991567 0.67270375 0.92333333
## shopping
## clothes
                0.15085389 0.23732120 0.07452696 0.84346701 0.30844444
## hollister
               0.04625237 0.07932380 0.04095293 0.28201811 0.17422222
## abercrombie 0.03130930 0.07022107 0.03025352 0.24450194 0.11822222
## die
                0.21513283\ 0.35825748\ 0.12270068\ 0.95730918\ 0.22155556
## death
                0.14136622 0.22041612 0.07884889 0.42432083 0.14844444
               0.08870968 0.12288687 0.05760818 0.57956016 0.11888889
## drunk
## drugs
               0.07068311 0.06306892 0.02529911 0.67917206 0.09177778
```

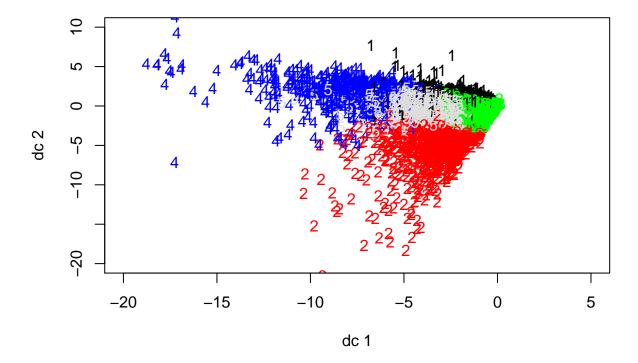
Visualizing the Clusters

```
library(fpc) #load this
plotcluster(interests, teen_clusters_5$cluster, main="k = 5") #creates a visualization of the K=5 clust
```



#If all your data ends up in a corner and hard to read- change the lim for y and x: #sometime you need to run it first with out the lims and then add them in and run again. plotcluster(interests, teen_clusters_5cluster, main="k=5", xlim=c(-20,5), ylim=c(-20,10))

k=5

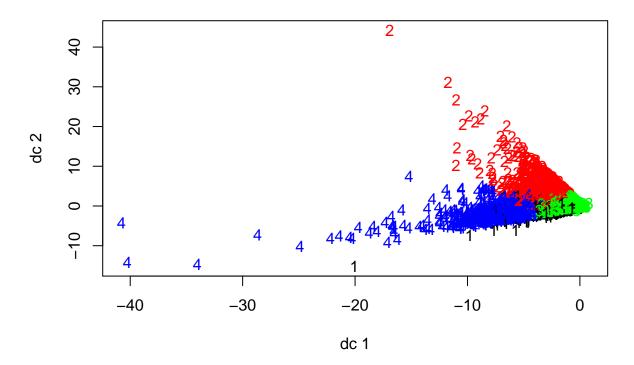


The plot you see here is two dimensional whereas your dataset has 36 dimensions (because of the 36 "interest" variables). In another word, it is a simplified version. The coordinates on the x and y axes are called "usual discriminant coordinates." Here's more on the topic: see here and see here

What about k=4?

```
set.seed(123)
teen_clusters_4 <- kmeans(interests, centers=4)
plotcluster(interests, teen_clusters_4$cluster, main="k=4")</pre>
```

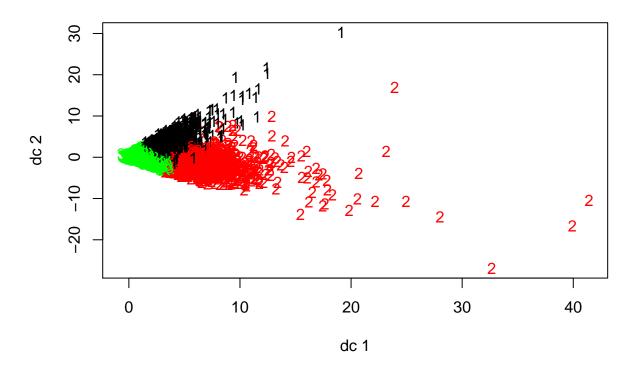




What about k=3?

```
set.seed(123)
teen_clusters_3 <- kmeans(interests, centers=3)
plotcluster(interests, teen_clusters_3$cluster, main="k=3")</pre>
```





Picking Among the K's

A Digression on Sum of Squares

Within Sum of Squares (withinss)

We want our clusters to be "unique." In another word, we want the sum of squares within each cluster to be small because it means the cluster is cohesive. As we stated earlier, the default algorithm in kmeans is Hartigan & Wong, which minimizes the withinss. What are the withinss for each cluster? Look at Clusters 3, 5, and 1 in particular. Which cluster has the largest withinss?

teen_clusters_5\$withinss

[1] 78835.33 80153.80 102071.44 179006.65 128069.21

Between Sum of Squares (betweenss)

We want each cluster to be different from its neighboring clusters. The betweenss is the most useful when we want to compare among multiple kmeans models.

teen_clusters_5\$betweenss

[1] 92856.37

Total Sum of Squares (totss)

```
totss = betweenss + withinss
```

```
teen_clusters_5$totss
```

[1] 660992.8

Method 1: Use the visualizations

Look at your cluster plots. Can you make a determination this way?

Method 2: Examine the betweenss and withinss ratios!

We want the clusters to demonstrate both cohesion and separation. Cohesion is measured by minimizing the ratio of withinss/totalss. Separation is measured by maximizing the ratio of betweenss/totalss.

Cluster Separation

```
clusters3<- teen_clusters_3$betweenss/teen_clusters_3$totss
clusters4<- teen_clusters_4$betweenss/teen_clusters_4$totss
clusters5<- teen_clusters_5$betweenss/teen_clusters_5$totss

betweenss.metric <- c(clusters3, clusters4, clusters5)
print(betweenss.metric) #Look for a ratio that is closer to 1.</pre>
```

```
## [1] 0.09368819 0.12094187 0.14048016
```

k=5 has the most separation.

Cluster Cohesion

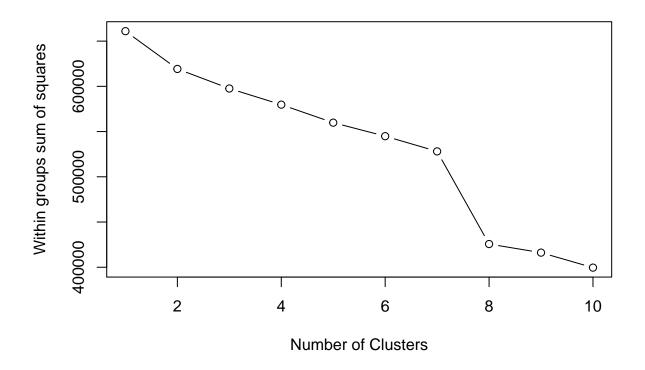
```
clusters3<- teen_clusters_3$tot.withinss/teen_clusters_3$totss
clusters4<- teen_clusters_4$tot.withinss/teen_clusters_4$totss
clusters5<- teen_clusters_5$tot.withinss/teen_clusters_5$totss

totwithinss.metric <- c(clusters3, clusters4, clusters5)
print(totwithinss.metric) #Looking for a ratio that is closer to 0.</pre>
```

```
## [1] 0.9063118 0.8790581 0.8595198
```

k=5 also has the most cluster cohesion.

Method 3: Using the "Elbow Method"



Source: The above code chunk is from here

Method 4: Use Your Business Knowledge!

What is actionable? What is not? what do you know about your customers? Your data?

A Side Note: Trying an Automatic Pick

```
\label{lineary} \begin{tabular}{ll} $\# teen\_clusters\_optimal <-kmeansruns(interests, krange=2:10) $\# finds the "best"" $K$ between 2 and 10 $\# teen\_clusters\_optimal $$bestk$ \end{tabular}
```

Creating an Aggregate Profile for Our Clusters

To create "meaning" for our clusters, we need to give each cluster an "identity."

```
teen_clusters_5$size #Get the size of each cluster
```

```
## [1] 4216 1538 18973 773 4500
```

 $\hbox{Clusters_5$<-data.frame(teen_clusters_5$$centers)} \ \textit{\#Put the cluster centroids into a data frame Clusters_5$<-data.frame(t(teen_clusters_5$$centers))} \ \textit{\#Transpose for easier reading }$

We can sort the centroids for each cluster to see what the teens were writing on their profiles.

Clusters_5[order(-Clusters_5\$X1),]

```
X5
##
                        Х1
                                   X2
                                               ХЗ
                                                          Х4
## music
                2.43050285 0.91742523 0.28292837 2.70245796 0.67111111
## band
                1.18927894 0.34980494 0.10441153 0.85769728 0.17600000
                0.42955408 0.32509753 0.13698414 1.41009056 0.28888889
## rock
                0.40702087 3.09882965 0.23812787 0.94566624 0.49511111
## god
                0.33942125 3.11053316 0.18916355 1.09443726 0.46777778
## dance
## hair
                0.32756167 0.51560468 0.15042429 4.47736093 0.93066667
                0.32447818 0.49024707 0.19991567 0.67270375 0.92333333
## shopping
                0.25996205 0.47854356 0.16381173 0.94954722 0.89155556
## cute
## football
                0.23268501 \ 0.29648895 \ 0.15363938 \ 0.71021992 \ 0.59288889
                0.22770398 0.77113134 0.11864228 0.40362225 0.60800000
## church
                0.21513283 0.35825748 0.12270068 0.95730918 0.22155556
## die
## mall
                0.20208729 0.31144343 0.13603542 0.78395860 0.71177778
                0.18833017 0.31859558 0.12828757 0.79689521 0.81911111
## basketball
## sex
                0.18477230 0.18985696 0.08970642 3.40491591 0.19488889
## marching
                0.17314991 0.03706112 0.01644442 0.06597671 0.01511111
                0.15417457 0.15669701 0.09007537 0.32341527 0.26266667
## swimming
## clothes
                0.15085389 0.23732120 0.07452696 0.84346701 0.30844444
## sports
                0.14302657 0.15149545 0.08996996 0.62483829 0.26066667
## sexy
                0.14255218 0.20676203 0.10789016 0.51487710 0.19377778
## death
                0.14136622 0.22041612 0.07884889 0.42432083 0.14844444
## dress
                0.12286528 0.24187256 0.06177199 0.37645537 0.21688889
                0.12096774 0.14629389 0.08296000 0.49547219 0.88688889
## soccer
                0.11835863 0.20221066 0.08164233 0.34152652 0.26111111
## hot
## jesus
                0.11598672 0.58452536 0.06756970 0.14618370 0.12866667
## kissed
                0.11480076 0.09947984 0.03895009 1.43596378 0.13555556
                0.10578748 0.08777633 0.06967796 0.16429495 0.13111111
## tennis
                0.10341556 0.15799740 0.07647710 0.36351876 0.53888889
## softball
                0.09582543 0.13524057 0.08090444 0.26908150 0.43088889
## volleyball
                0.09369070 0.08972692 0.07046856 0.38809832 0.21733333
## baseball
## drunk
                0.08870968 0.12288687 0.05760818 0.57956016 0.11888889
                0.07068311 0.06306892 0.02529911 0.67917206 0.09177778
## drugs
## cheerleading 0.06783681 0.15734720 0.06920361 0.29236740 0.25155556
                0.06499051 0.10078023 0.03868655 1.28978008 0.17955556
## blonde
                0.04625237 0.07932380 0.04095293 0.28201811 0.17422222
## hollister
## abercrombie 0.03130930 0.07022107 0.03025352 0.24450194 0.11822222
## bible
                0.02371917 0.12418726 0.01054130 0.03492885 0.02711111
```

Clusters_5[order(-Clusters_5\$X2),]

```
##
                                    X2
                                               ХЗ
                                                          Х4
                                                                      X5
                        X 1
                0.33942125 3.11053316 0.18916355 1.09443726 0.46777778
## dance
                0.40702087 3.09882965 0.23812787 0.94566624 0.49511111
## god
                2.43050285 0.91742523 0.28292837 2.70245796 0.67111111
## music
                0.22770398 0.77113134 0.11864228 0.40362225 0.60800000
## church
                0.11598672 0.58452536 0.06756970 0.14618370 0.12866667
## jesus
                0.32756167\ 0.51560468\ 0.15042429\ 4.47736093\ 0.93066667
## hair
                0.32447818 0.49024707 0.19991567 0.67270375 0.92333333
## shopping
## cute
                0.25996205 0.47854356 0.16381173 0.94954722 0.89155556
                0.21513283 0.35825748 0.12270068 0.95730918 0.22155556
## die
## band
                1.18927894 0.34980494 0.10441153 0.85769728 0.17600000
```

```
## rock
                0.42955408 0.32509753 0.13698414 1.41009056 0.28888889
                0.18833017 0.31859558 0.12828757 0.79689521 0.81911111
## basketball
## mall
                0.20208729 0.31144343 0.13603542 0.78395860 0.71177778
## football
                0.23268501 0.29648895 0.15363938 0.71021992 0.59288889
## dress
                0.12286528 0.24187256 0.06177199 0.37645537 0.21688889
                0.15085389 0.23732120 0.07452696 0.84346701 0.30844444
## clothes
                0.14136622 0.22041612 0.07884889 0.42432083 0.14844444
## death
                0.14255218 0.20676203 0.10789016 0.51487710 0.19377778
## sexy
## hot
                0.11835863 0.20221066 0.08164233 0.34152652 0.26111111
                0.18477230 0.18985696 0.08970642 3.40491591 0.19488889
## sex
## softball
                0.10341556 0.15799740 0.07647710 0.36351876 0.53888889
## cheerleading 0.06783681 0.15734720 0.06920361 0.29236740 0.25155556
                0.15417457 0.15669701 0.09007537 0.32341527 0.26266667
  swimming
## sports
                0.14302657 0.15149545 0.08996996 0.62483829 0.26066667
                0.12096774 0.14629389 0.08296000 0.49547219 0.88688889
## soccer
## volleyball
                0.09582543 0.13524057 0.08090444 0.26908150 0.43088889
                0.02371917 0.12418726 0.01054130 0.03492885 0.02711111
## bible
## drunk
                0.08870968 0.12288687 0.05760818 0.57956016 0.11888889
                0.06499051 0.10078023 0.03868655 1.28978008 0.17955556
## blonde
## kissed
                0.11480076 0.09947984 0.03895009 1.43596378 0.13555556
## baseball
                0.09369070 0.08972692 0.07046856 0.38809832 0.21733333
                0.10578748 0.08777633 0.06967796 0.16429495 0.13111111
## tennis
                0.04625237\ 0.07932380\ 0.04095293\ 0.28201811\ 0.17422222
## hollister
## abercrombie 0.03130930 0.07022107 0.03025352 0.24450194 0.11822222
                0.07068311 0.06306892 0.02529911 0.67917206 0.09177778
## drugs
## marching
                0.17314991 0.03706112 0.01644442 0.06597671 0.01511111
```

Clusters_5[order(-Clusters_5\$X3),]

```
##
                        X1
                                   X2
                                               ХЗ
                                                          X4
                                                                     Х5
## music
                2.43050285 0.91742523 0.28292837 2.70245796 0.67111111
                0.40702087 3.09882965 0.23812787 0.94566624 0.49511111
## god
                0.32447818 0.49024707 0.19991567 0.67270375 0.92333333
## shopping
                0.33942125 3.11053316 0.18916355 1.09443726 0.46777778
## dance
                0.25996205 0.47854356 0.16381173 0.94954722 0.89155556
## cute
## football
                0.23268501 0.29648895 0.15363938 0.71021992 0.59288889
## hair
                0.32756167 0.51560468 0.15042429 4.47736093 0.93066667
                0.42955408 0.32509753 0.13698414 1.41009056 0.28888889
## rock
                0.20208729 0.31144343 0.13603542 0.78395860 0.71177778
## mall
## basketball
                0.18833017 0.31859558 0.12828757 0.79689521 0.81911111
## die
                0.21513283 0.35825748 0.12270068 0.95730918 0.22155556
## church
                0.22770398 0.77113134 0.11864228 0.40362225 0.60800000
                0.14255218 0.20676203 0.10789016 0.51487710 0.19377778
## sexy
## band
                1.18927894 0.34980494 0.10441153 0.85769728 0.17600000
                0.15417457 0.15669701 0.09007537 0.32341527 0.26266667
## swimming
                0.14302657 0.15149545 0.08996996 0.62483829 0.26066667
## sports
                0.18477230 0.18985696 0.08970642 3.40491591 0.19488889
## sex
                0.12096774 0.14629389 0.08296000 0.49547219 0.88688889
## soccer
                0.11835863 0.20221066 0.08164233 0.34152652 0.26111111
## hot
                0.09582543\ 0.13524057\ 0.08090444\ 0.26908150\ 0.43088889
## volleyball
                0.14136622 0.22041612 0.07884889 0.42432083 0.14844444
## death
## softball
                0.10341556 0.15799740 0.07647710 0.36351876 0.53888889
                0.15085389 0.23732120 0.07452696 0.84346701 0.30844444
## clothes
                0.09369070 0.08972692 0.07046856 0.38809832 0.21733333
## baseball
```

```
0.10578748 0.08777633 0.06967796 0.16429495 0.13111111
## tennis
## cheerleading 0.06783681 0.15734720 0.06920361 0.29236740 0.25155556
## jesus
                0.11598672 0.58452536 0.06756970 0.14618370 0.12866667
                0.12286528 0.24187256 0.06177199 0.37645537 0.21688889
## dress
## drunk
                0.08870968 0.12288687 0.05760818 0.57956016 0.11888889
                0.04625237 0.07932380 0.04095293 0.28201811 0.17422222
## hollister
## kissed
                0.11480076 0.09947984 0.03895009 1.43596378 0.13555556
                0.06499051 0.10078023 0.03868655 1.28978008 0.17955556
## blonde
## abercrombie 0.03130930 0.07022107 0.03025352 0.24450194 0.11822222
                0.07068311 0.06306892 0.02529911 0.67917206 0.09177778
## drugs
## marching
                0.17314991 0.03706112 0.01644442 0.06597671 0.01511111
                0.02371917 0.12418726 0.01054130 0.03492885 0.02711111
## bible
```

Clusters_5[order(-Clusters_5\$X4),]

```
##
                        Х1
                                   Х2
                                              ХЗ
                                                          X4
                                                                     Х5
## hair
                0.32756167 0.51560468 0.15042429 4.47736093 0.93066667
## sex
                0.18477230 0.18985696 0.08970642 3.40491591 0.19488889
                2.43050285 0.91742523 0.28292837 2.70245796 0.67111111
## music
## kissed
                0.11480076 0.09947984 0.03895009 1.43596378 0.13555556
## rock
                0.42955408 0.32509753 0.13698414 1.41009056 0.28888889
## blonde
                0.06499051 0.10078023 0.03868655 1.28978008 0.17955556
## dance
                0.33942125 3.11053316 0.18916355 1.09443726 0.46777778
## die
                0.21513283 0.35825748 0.12270068 0.95730918 0.22155556
                0.25996205 0.47854356 0.16381173 0.94954722 0.89155556
## cute
                0.40702087 3.09882965 0.23812787 0.94566624 0.49511111
## god
## band
                1.18927894 0.34980494 0.10441153 0.85769728 0.17600000
## clothes
                0.15085389 0.23732120 0.07452696 0.84346701 0.30844444
## basketball
                0.18833017  0.31859558  0.12828757  0.79689521  0.81911111
                0.20208729 0.31144343 0.13603542 0.78395860 0.71177778
## mall
                0.23268501 0.29648895 0.15363938 0.71021992 0.59288889
## football
## drugs
                0.07068311 0.06306892 0.02529911 0.67917206 0.09177778
## shopping
                0.32447818 0.49024707 0.19991567 0.67270375 0.92333333
                0.14302657 0.15149545 0.08996996 0.62483829 0.26066667
## sports
## drunk
                0.08870968 0.12288687 0.05760818 0.57956016 0.11888889
                0.14255218 0.20676203 0.10789016 0.51487710 0.19377778
## sexy
## soccer
                0.12096774 0.14629389 0.08296000 0.49547219 0.88688889
## death
                0.14136622 0.22041612 0.07884889 0.42432083 0.14844444
## church
                0.22770398 0.77113134 0.11864228 0.40362225 0.60800000
                0.09369070 0.08972692 0.07046856 0.38809832 0.21733333
## baseball
                0.12286528 0.24187256 0.06177199 0.37645537 0.21688889
## dress
                0.10341556 0.15799740 0.07647710 0.36351876 0.53888889
## softball
## hot
                0.11835863 0.20221066 0.08164233 0.34152652 0.26111111
                0.15417457 0.15669701 0.09007537 0.32341527 0.26266667
## swimming
## cheerleading 0.06783681 0.15734720 0.06920361 0.29236740 0.25155556
                0.04625237 0.07932380 0.04095293 0.28201811 0.17422222
## hollister
                0.09582543 0.13524057 0.08090444 0.26908150 0.43088889
## volleyball
## abercrombie 0.03130930 0.07022107 0.03025352 0.24450194 0.11822222
## tennis
                0.10578748 0.08777633 0.06967796 0.16429495 0.13111111
                0.11598672 0.58452536 0.06756970 0.14618370 0.12866667
## jesus
                0.17314991 0.03706112 0.01644442 0.06597671 0.01511111
## marching
                0.02371917 0.12418726 0.01054130 0.03492885 0.02711111
## bible
```

Clusters_5[order(-Clusters_5\$X5),]

```
##
                        Х1
                                   Х2
                                               ХЗ
                                                          X4
                                                                     Х5
                0.32756167 0.51560468 0.15042429 4.47736093 0.93066667
## hair
## shopping
                0.32447818 0.49024707 0.19991567 0.67270375 0.92333333
                0.25996205 0.47854356 0.16381173 0.94954722 0.89155556
## cute
                0.12096774 0.14629389 0.08296000 0.49547219 0.88688889
## soccer
## basketball
                0.18833017 0.31859558 0.12828757 0.79689521 0.81911111
## mall
                0.20208729 0.31144343 0.13603542 0.78395860 0.71177778
## music
                2.43050285 0.91742523 0.28292837 2.70245796 0.67111111
                0.22770398 0.77113134 0.11864228 0.40362225 0.60800000
## church
## football
                0.23268501 0.29648895 0.15363938 0.71021992 0.59288889
                0.10341556 0.15799740 0.07647710 0.36351876 0.53888889
## softball
## god
                0.40702087 3.09882965 0.23812787 0.94566624 0.49511111
## dance
                0.33942125 3.11053316 0.18916355 1.09443726 0.46777778
                0.09582543 0.13524057 0.08090444 0.26908150 0.43088889
## volleyball
                0.15085389 0.23732120 0.07452696 0.84346701 0.30844444
## clothes
## rock
                0.42955408 0.32509753 0.13698414 1.41009056 0.28888889
## swimming
                0.15417457 0.15669701 0.09007537 0.32341527 0.26266667
                0.11835863 0.20221066 0.08164233 0.34152652 0.26111111
## hot
## sports
                0.14302657 0.15149545 0.08996996 0.62483829 0.26066667
## cheerleading 0.06783681 0.15734720 0.06920361 0.29236740 0.25155556
## die
                0.21513283 0.35825748 0.12270068 0.95730918 0.22155556
                0.09369070 0.08972692 0.07046856 0.38809832 0.21733333
## baseball
## dress
                0.12286528 0.24187256 0.06177199 0.37645537 0.21688889
                0.18477230 0.18985696 0.08970642 3.40491591 0.19488889
## sex
                0.14255218 0.20676203 0.10789016 0.51487710 0.19377778
## sexy
## blonde
                0.06499051 0.10078023 0.03868655 1.28978008 0.17955556
                1.18927894 0.34980494 0.10441153 0.85769728 0.17600000
## band
## hollister
                0.04625237 \ 0.07932380 \ 0.04095293 \ 0.28201811 \ 0.17422222
                0.14136622 0.22041612 0.07884889 0.42432083 0.14844444
## death
## kissed
                0.11480076 0.09947984 0.03895009 1.43596378 0.13555556
                0.10578748 0.08777633 0.06967796 0.16429495 0.13111111
## tennis
                0.11598672 0.58452536 0.06756970 0.14618370 0.12866667
## jesus
                0.08870968 0.12288687 0.05760818 0.57956016 0.11888889
## drunk
                0.03130930 0.07022107 0.03025352 0.24450194 0.11822222
## abercrombie
                0.07068311 0.06306892 0.02529911 0.67917206 0.09177778
## drugs
                0.02371917 0.12418726 0.01054130 0.03492885 0.02711111
## bible
                0.17314991 0.03706112 0.01644442 0.06597671 0.01511111
## marching
```

Cluster 1 (4,216 teens): music, band. Other words with smaller centroids: rock, god, dance, hair, shopping, cute, football, church.

Are these the "band kids"?

Cluster 2 (1,538 teens): dance, god. Other words with smaller centroids: music, church, jesus, hair, shopping, cute, die, band.

Are these the "religious/church" kids?

Cluster 3 (18,973 teens): All very low centroid values: music, god, shopping, dance, cute, football, hair, rock, mall, basketball

Who are these "kids"? The "basket cases"?

Cluster 4 (773 teens): hair, sex, music, kissed, rock, blonde. Other words with smaller centroids: dance, die, cute, god

Are these "princesses"?

Cluster 5 (4,500 teens): Moderate centroid values: hair, shopping, cute, soccer, basketball, mall, music, church, football, softball

Who are these "kids"? Another "basket cases"?

Let's add back the demographic information.

```
# apply the cluster IDs to the original data frame
teens$cluster <- teen_clusters_5$cluster #adds the cluster number to each recond
# mean age by cluster
aggregate(data = teens, age ~ cluster, mean)
##
     cluster
                  age
## 1
           1 17.27793
           2 17.20026
## 2
## 3
           3 17.29521
## 4
           4 17.04603
## 5
           5 17.00078
# proportion of females by cluster
aggregate(data = teens, female ~ cluster, mean)
##
     cluster
                female
## 1
           1 0.7352941
## 2
           2 0.8146944
## 3
           3 0.6954093
## 4
           4 0.8680466
## 5
           5 0.8524444
# mean number of friends by cluster
aggregate(data = teens, friends ~ cluster, mean)
##
     cluster friends
## 1
           1 30.74763
## 2
           2 35.92133
## 3
           3 27.57076
## 4
           4 30.31824
## 5
           5 38.65978
```

K-Medoid Clustering

The problem with k-means is that it is sensitive to outliers. A workaround to this issue is k-medoids clustering. Instead of finding centroids, we find medoids. What is a medoid? Medoid is just basically the most "central" data point in a cluster. Instead of finding the mean point in a cluster, we just choose one of the existing data points in each cluster to make it the "center."