HW3_Q4

Mimis Jimenez

4/21/2020

load the data and packages

```
dataraw<- read.csv("~/GitHub/SDS323_Spring2020/hw3/q4/social_marketing.csv")
View(dataraw)

library(LICORS)
library(foreach)
library(mosaic)
```

```
## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Loading required package: lattice

## Loading required package: ggformula

## Loading required package: ggplot2

## Loading required package: ggstance

##
## Attaching package: 'ggstance'

## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh

##
## New to ggformula?  Try the tutorials:
##  learnr::run_tutorial("introduction", package = "ggformula")
##  learnr::run_tutorial("refining", package = "ggformula")

## Loading required package: mosaicData

## Loading required package: Matrix

## Registered S3 method overwritten by 'mosaic':
##   method                           from
```

```
##   fortify.SpatialPolygonsDataFrame ggplot2

##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
##
## Note: If you use the Matrix package, be sure to load it BEFORE loading mosaic.

##
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':
##
##     mean

## The following object is masked from 'package:ggplot2':
##
##     stat

## The following objects are masked from 'package:dplyr':
##
##     count, do, tally

## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var

## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```r
library(ggcorrplot)
```

Clean the data

```r
#Let's take out spam and adult content, since these are most likely bots
datatest <- subset(dataraw, spam == 0)
View(datatest)
dataclean <- subset(datatest, adult == 0)
View(dataclean)
#Remove "chatter" and "uncategorized" column from data. They will most likely not add anything useful t
dataclean$chatter <- NULL
dataclean$uncategorized <- NULL
dataclean$spam <- NULL
dataclean$adult <- NULL
View(dataclean)
```

I choose to remove columns such as "chatter" and "uncategorized" because these categories will most likely not contribute anything significant to finding a market segment. Even from a marketing point of view, there wouldn't be a meaningful way to market to people who are interested in "uncategorized." So these variables, along with "spam" and "adult" were removed from my analysis.

Let's run clusters to see how the followers can be grouped

```
#Center and scale the data
z <- dataclean[,-(1)]
z <- scale(z, center = TRUE, scale = TRUE)

#Extract centers and scales
mu = attr(z,"scaled:center")
sigma = attr(z,"scaled:scale")

#Let's run K-means clustering with K = 6
clust_6 <- kmeans(z, 6, nstart = 25)
clust_6$center
```

```
##   current_events      travel photo_sharing       tv_film sports_fandom
## 1    0.111766797 -0.12390948  -0.017969812 -0.0079869653     1.9589989
## 2    0.205880417 -0.02957148   1.223640197 -0.0004943682    -0.2216949
## 3   -0.061247827 -0.21808014  -0.145524390 -0.0437643306    -0.2893028
## 4    0.002669436 -0.15254297  -0.002610625 -0.0578234036    -0.2090516
## 5    0.103102623  1.73437257  -0.056113144  0.0737811086     0.1962970
## 6   -0.009443331 -0.01170284   0.044011788  0.4612026075    -0.1266301
##     politics        food      family home_and_garden       music        news
## 1 -0.2276967  1.77182915  1.43017195       0.1895999  0.03744181 -0.09442182
## 2 -0.1309429 -0.20595065  0.01697101       0.1282810  0.57486259 -0.09167878
## 3 -0.2586450 -0.36012001 -0.25958435      -0.1139404 -0.11394039 -0.24953645
## 4 -0.1856996  0.39954466 -0.08714553       0.1428898  0.03807886 -0.04845616
## 5  2.3319283  0.03708618  0.06581170       0.1401002 -0.03890610  1.93427538
## 6 -0.1833194 -0.07911038  0.16398906       0.1435188  0.34500389 -0.20005801
##   online_gaming     shopping health_nutrition   college_uni sports_playing
## 1   -0.06826536  0.055816154      -0.15745503 -0.1220458698    0.100289838
## 2   -0.04508575  0.317236842      -0.07044424 -0.0003889571    0.179366125
## 3   -0.23490579 -0.057149013      -0.33089913 -0.2231586618   -0.223997923
## 4   -0.13426255  0.053157710       2.08834455 -0.2157370373   -0.058016156
## 5   -0.13975571 -0.004254324      -0.20797073 -0.0837567212   -0.005117182
## 6    3.08311383 -0.024426120      -0.18804466  3.0772311160    2.009150207
##      cooking         eco   computers    business     outdoors      crafts
## 1 -0.1099012  0.19204739  0.05724093  0.10367034 -0.06795444  0.68970292
## 2  2.6283162  0.05630249  0.06286667  0.25376278  0.01047693  0.11941633
## 3 -0.3360670 -0.16717484 -0.24102130 -0.12505724 -0.32057854 -0.18822240
## 4  0.3719219  0.54051917 -0.06918821  0.06707358  1.61114846  0.10898538
## 5 -0.2161606  0.12347442  1.54180029  0.36271135  0.11673212  0.14300137
## 6 -0.1571421 -0.01179577 -0.07621386  0.04296754 -0.10526165  0.09090169
##    automotive          art    religion     beauty   parenting      dating
## 1  0.15502525  0.095913677  2.14540030  0.3027976  2.05574474  0.09625492
## 2  0.02412927  0.119293467 -0.14310414  2.4379926 -0.10580099  0.06183434
## 3 -0.18276464 -0.065745371 -0.29861888 -0.2701637 -0.30801290 -0.09245447
## 4 -0.13898166  0.024233305 -0.19281498 -0.2062151 -0.11456320  0.17452693
## 5  1.13017760  0.002646013 -0.03064009 -0.1736757  0.01918615  0.20090250
## 6  0.01463648  0.294690950 -0.12874447 -0.2038379 -0.16249871  0.01157918
##        school personal_fitness     fashion small_business
## 1  1.62264198      -0.11834977  0.02522115     0.10441124
## 2  0.17684266      -0.04314241  2.51061174     0.23314627
## 3 -0.24814873      -0.33631608 -0.26281447    -0.10125179
## 4 -0.16409371       2.05432243 -0.12109641    -0.05956371
```

```
## 5 -0.01312053      -0.18909829 -0.17353873     0.26006766
## 6 -0.21461952      -0.19808886 -0.05386000     0.25065507
```

```r
clust_6$center[1,]*sigma + mu
```

```
##   current_events           travel    photo_sharing          tv_film
##        1.6625173        1.2890733        2.6652835        1.0677732
##    sports_fandom         politics             food           family
##        5.8520055        1.1175657        4.5242047        2.4647303
##  home_and_garden            music             news    online_gaming
##        0.6528354        0.7247580        1.0152144        1.0179806
##         shopping health_nutrition      college_uni   sports_playing
##        1.5048409        1.8699862        1.2074689        0.7427386
##          cooking              eco        computers         business
##        1.6334716        0.6459198        0.7123098        0.4965422
##         outdoors           crafts       automotive              art
##        0.6860304        1.0733057        1.0304288        0.8686030
##         religion           beauty        parenting           dating
##        5.1867220        1.1092669        4.0331950        0.8838174
##           school personal_fitness          fashion   small_business
##        2.6804979        1.1756570        1.0470263        0.3900415
```

```r
clust_6$center[2,]*sigma + mu
```

```
##   current_events           travel    photo_sharing          tv_film
##        1.7820268        1.5066922        6.0726577        1.0803059
##    sports_fandom         politics             food           family
##        1.1128107        1.4130019        1.0210325        0.8738050
##  home_and_garden            music             news    online_gaming
##        0.6080306        1.2829828        1.0210325        1.0803059
##         shopping health_nutrition      college_uni   sports_playing
##        1.9808795        2.2619503        1.5640535        0.8202677
##          cooking              eco        computers         business
##       11.0076482        0.5430210        0.7189293        0.6003824
##         outdoors           crafts       automotive              art
##        0.7801147        0.6080306        0.8508604        0.9063098
##         religion           beauty        parenting           dating
##        0.8202677        3.9617591        0.7495220        0.8221797
##           school personal_fitness          fashion   small_business
##        0.9674952        1.3575526        5.6137667        0.4684512
```

```r
clust_6$center[3,]*sigma + mu
```

```
##   current_events           travel    photo_sharing          tv_film
##        1.4428160        1.0718405        2.3152331        1.0079289
##    sports_fandom         politics             food           family
##        0.9658818        1.0230658        0.7479577        0.5624700
##  home_and_garden            music             news    online_gaming
##        0.4310428        0.5675156        0.6862086        0.5699183
##         shopping health_nutrition      college_uni   sports_playing
##        1.2991350        1.0886593        0.9111004        0.4247958
##          cooking              eco        computers         business
##        0.8592023        0.3736185        0.3613647        0.3382989
##         outdoors           crafts       automotive              art
##        0.3829889        0.3570399        0.5670351        0.6078808
##         religion           beauty        parenting           dating
```

```
##       0.5235464       0.3438251       0.4423354       0.5458914
##          school personal_fitness         fashion   small_business
##       0.4639596       0.6484863       0.5177799       0.2647765
```

```
clust_6$center[4,]*sigma + mu
```

```
##   current_events           travel    photo_sharing          tv_film
##        1.5239808        1.2230216        2.7074341        0.9844125
##    sports_fandom         politics             food           family
##        1.1402878        1.2458034        2.0935252        0.7565947
##  home_and_garden            music             news    online_gaming
##        0.6187050        0.7254197        1.1127098        0.8405276
##         shopping health_nutrition      college_uni   sports_playing
##        1.5000000       11.9868106        0.9328537        0.5875300
##          cooking              eco        computers         business
##        3.2829736        0.9100719        0.5635492        0.4712230
##         outdoors           crafts       automotive              art
##        2.7002398        0.5995204        0.6270983        0.7529976
##         religion           beauty        parenting           dating
##        0.7254197        0.4292566        0.7362110        1.0239808
##           school personal_fitness          fashion   small_business
##        0.5635492        6.4304556        0.7781775        0.2901679
```

```
clust_6$center[5,]*sigma + mu
```

```
##   current_events           travel    photo_sharing          tv_film
##        1.6515152        5.5757576        2.5606061        1.2045455
##    sports_fandom         politics             food           family
##        2.0212121        8.9333333        1.4515152        0.9287879
##  home_and_garden            music             news    online_gaming
##        0.6166667        0.6454545        5.3181818        0.8257576
##         shopping health_nutrition      college_uni   sports_playing
##        1.3954545        1.6424242        1.3196970        0.6393939
##          cooking              eco        computers         business
##        1.2696970        0.5939394        2.4590909        0.6757576
##         outdoors           crafts       automotive              art
##        0.9075758        0.6272727        2.3681818        0.7181818
##         religion           beauty        parenting           dating
##        1.0348485        0.4727273        0.9393939        1.0712121
##           school personal_fitness          fashion   small_business
##        0.7424242        1.0045455        0.6818182        0.4848485
```

```
clust_6$center[6,]*sigma + mu
```

```
##   current_events           travel    photo_sharing          tv_film
##        1.5085995        1.5479115        2.8353808        1.8525799
##    sports_fandom         politics             food           family
##        1.3194103        1.2530713        1.2457002        1.0393120
##  home_and_garden            music             news    online_gaming
##        0.6191646        1.0442260        0.7911548        9.4914005
##         shopping health_nutrition      college_uni   sports_playing
##        1.3587224        1.7321867       10.5847666        2.6142506
##          cooking              eco        computers         business
##        1.4717445        0.4914005        0.5552826        0.4545455
##         outdoors           crafts       automotive              art
##        0.6412776        0.5847666        0.8378378        1.1891892
```

```
##         religion          beauty       parenting          dating
##        0.8476658       0.4324324       0.6633907       0.7321867
##           school personal_fitness        fashion   small_business
##        0.5036855       0.9828010       0.9017199       0.4791155
```

```r
religion_sports_followers <- clust_6$center[1,]*sigma + mu
health_enthusiasts <- clust_6$center[2,]*sigma + mu
college_gamers <- clust_6$center[3,]*sigma + mu
informed_travelers <- clust_6$center[5,]*sigma + mu
cooking_influencers <- clust_6$center[6,]*sigma + mu
```

I then proceed to create clusters from the cleaned data. I chose to do six clusters because in marketing, you don't want to spread out your audience into many segments. From here I looked at the centers for each cluster to see what categories were of highest interest. Most of these clusters had a few distinctive categories of interest. A small number of clusters, however, did not seem to have categories that were particularly high. Most likely, these clusters captured groups of followers that talked about broad topics, but did not have a great affinity for any of the categories.

From here, it is best to take the clusters that did have an affinity in at least two of the categories. These consisted of clusters 1, 2, 3, 5, and 6. From these clusters, we can now define five different market segments. The first segment is called "Religion and Sports Followers" because these followers tend to be really interested in both religion and a sports fandom. Next, we have "Health Enthusiasts" who are people mainly interested in health nutrition and personal fitness and somewhat interested in cooking. The next segment, "College Gamers," are people who are heavily interested in colleges and online gaming. These are most likely college students with an affinity for gaming. Then we have the "Informed Travelers," who are people interested in traveling, politics, and the news. Finally, we have "Cooking Influencers." This is an interesting group that is greatly interested in cooking, but also things such as fashion, beauty, and photo sharing. Most likely this is a group that enjoys the lifestyle of a social media persona (influencer) interested in cooking. In addition to a particular hobby/talent/interest social media influencers are known for photo sharing, which usually encompasses looking good in the photos, hence beauty and fashion.

```r
length(which(clust_6$cluster == 1))
```

```
## [1] 723
```

```r
length(which(clust_6$cluster == 2))
```

```
## [1] 523
```

```r
length(which(clust_6$cluster == 3))
```

```
## [1] 4162
```

```r
length(which(clust_6$cluster == 5))
```

```
## [1] 660
```

```r
length(which(clust_6$cluster == 6))
```

```
## [1] 407
```

Five segments is a lot for one brand to try to target and brand after. So it is my recommendation that NutrientH20 focus on no more than three segments. These three segments can be selected on the size of how many followers fall within these segments. Thus, the three largest segments, and the segments that NutrientH20 should target, are the "Health Enthusiasts," "Religion and Sports Followers," and the "Informed Travelers."