Market Segmentation

2024-08-11

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(dplyr)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(cluster)
```

```
# Import Data
social_marketing <- read.csv("~/Downloads/social_marketing.csv", row.names = 1)
```

After looking at the initial data, we decided to drop all users who had more than 1 spam tweet and more than 1 adult tweet. This seemed like an appropriate way to remove followers that may be a bot.

After removing those users, we decided to drop the spam and adult column because it is not going to provide any valuable imformation about NutritionH20's market segments.

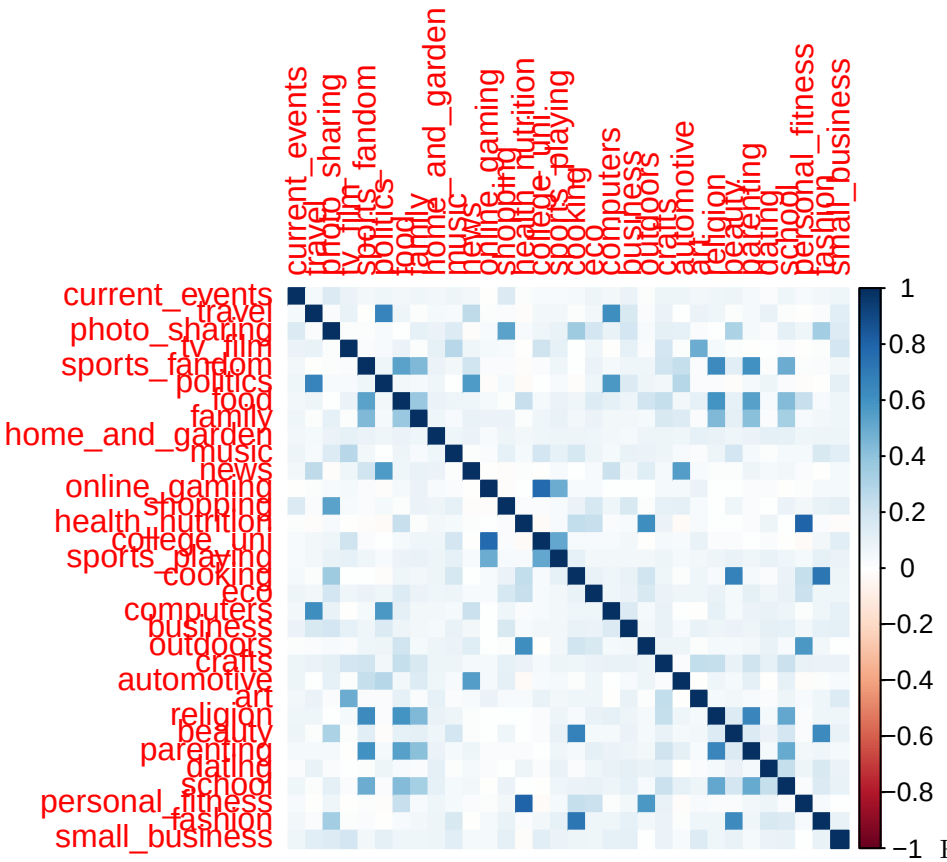
```
# Drop all users who have and span and adult content more than once (Potential Spa)
social_marketing <- social_marketing %>%
  filter(spam <= 1)

social_marketing <- social_marketing %>%
  filter(adult <= 1)

social_marketing <- social_marketing %>% select(-spam, -adult)
```

Because the uncategorized column was used to label all miscellaneous tweets, we decided to drop this column as well. These columns does not provide any unique information that would be valuable for identifying market segments. Out of fear that the chatter column was used for a similar purchase, we ultimately decided to drop this column as well.

```
# Drop chatter column because it provides value
social_marketing <- social_marketing %>% select(-chatter, -uncategorized)
```

As a starting point, we created a correlation matrix to begin to understand the potential market segments that may exist.

```
numeric_df <- social_marketing %>% select_if(is.numeric)
corr_matrix <-cor(numeric_df)
corrplot(corr_matrix, method="color")
```



From the correlation matrix we see that there is strong correlation between:

1) Personal Fitness and Health/Nutrition
2) College/University and Online Gaming
3) Cooking and Fashion
4) Fashion and Beauty

These associations are a great starting point for understanding what our market segments might be, but to understand multi-dimensional relationships, we can perform k-means clustering.

```
# Center/scale the data for the sake of interpretability
social_marketing = scale(social_marketing, center=TRUE, scale=TRUE)
```

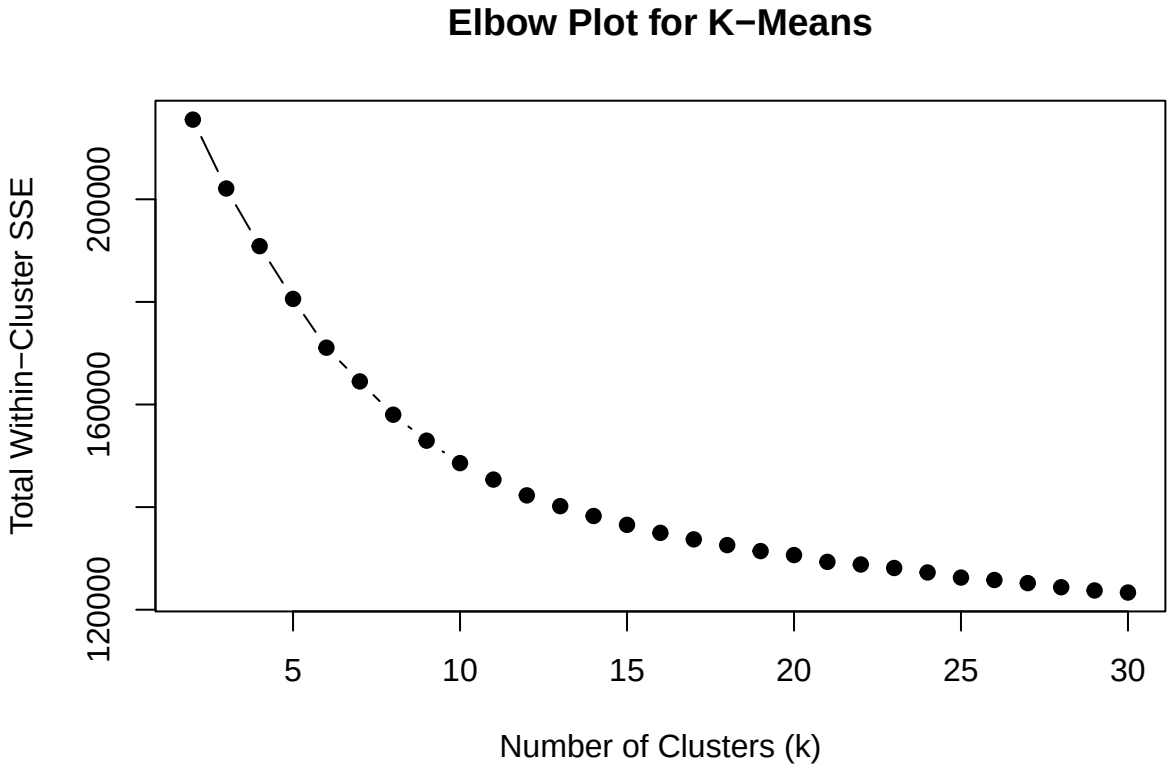
```
library(foreach)
```

```
##
## Attaching package: 'foreach'
```

```
## The following objects are masked from 'package:purrr':
##
##     accumulate, when
```

```
k_grid = seq(2, 30, by=1)
SSE_grid = foreach(k = k_grid, .combine='c') %do% {
cluster_k = kmeans(social_marketing, k, nstart=50)
cluster_k$tot.withinss
```

```
}

plot(k_grid, SSE_grid, type='b', pch=19, xlab="Number of Clusters (k)", ylab="Total Within-Cluster SSE"
```



Creating the elbow plot for this problem proved to be extremely unuseful. For the sake of interpretability of the market segments, we decided to use 5 centers in our kmeans model.

```
# Perform K-means clustering
set.seed(123)
social_marketing_clustering <- kmeans(social_marketing, centers=5, nstart=25)
```

Below we can see information about the centers of each cluster. Keep in mind that the data has been scaled so all values are in terms of the number of standard deviations away from the mean.

```
cluster_centers <- data.frame(social_marketing_clustering$centers)
row.names(cluster_centers) <- c("Cluster 1", "Cluster 2", "Cluster 3", "Cluster 4","Cluster 5")
cluster_centers
```

```
##           current_events      travel photo_sharing     tv_film sports_fandom
## Cluster 1    0.103312936 -0.12080996  -0.016805884  0.02509295     1.9611928
## Cluster 2   -0.057300904 -0.21064174  -0.140666320 -0.02089455    -0.2871731
## Cluster 3    0.006628872 -0.15025405   0.002447883 -0.03192213    -0.2102113
## Cluster 4    0.091189663  1.73265819  -0.076238475  0.08576580     0.1931905
## Cluster 5    0.194976779 -0.03382451   1.209474875  0.07700655    -0.2157764
##             politics        food      family home_and_garden       music
## Cluster 1 -0.2255696  1.75867631  1.42195473       0.1959753  0.06074508
## Cluster 2 -0.2606533 -0.35381032 -0.23566487      -0.1029169 -0.09240814
## Cluster 3 -0.1895447  0.40633060 -0.08314284       0.1509395  0.06166833
## Cluster 4  2.3135934  0.04416153  0.06730083       0.1414083 -0.04841534
## Cluster 5 -0.1396843 -0.19039791  0.03982076       0.1520062  0.60787706
##                  news online_gaming    shopping health_nutrition  college_uni
```

```
## Cluster 1 -0.09342610    0.03040995  0.05232404      -0.15965160 -0.008926010
## Cluster 2 -0.25300833   -0.01455222 -0.06357530      -0.33358025 -0.008953394
## Cluster 3 -0.05223230   -0.01489848  0.04625676       2.07232121 -0.093862396
## Cluster 4  1.90528860   -0.01838501 -0.01980441      -0.20573084  0.035462995
## Cluster 5 -0.09297868    0.11885677  0.38257514      -0.08732869  0.182012421
##           sports_playing    cooking         eco   computers    business
## Cluster 1     0.15727861 -0.1126275  0.18567230  0.06172256  0.09211735
## Cluster 2    -0.08169206 -0.3374552 -0.16024872 -0.24034569 -0.11947855
## Cluster 3     0.03172271  0.3720485  0.52410964 -0.07961805  0.05660958
## Cluster 4     0.07446459 -0.2178031  0.11259072  1.54623028  0.36428749
## Cluster 5     0.29738807  2.4806282  0.08105529  0.07613677  0.29414389
##              outdoors     crafts  automotive         art    religion     beauty
## Cluster 1 -0.06110606  0.6813153  0.15086237  0.11350826  2.14850250  0.2892479
## Cluster 2 -0.32029515 -0.1803748 -0.17460406 -0.05090849 -0.29792925 -0.2770801
## Cluster 3  1.59097566  0.1251638 -0.13177902  0.03977800 -0.18674468 -0.2096542
## Cluster 4  0.11893832  0.1521294  1.10655771  0.02264328 -0.03446354 -0.1838535
## Cluster 5  0.02485303  0.1529535  0.04973965  0.16310076 -0.13936915  2.3292348
##             parenting      dating      school personal_fitness     fashion
## Cluster 1  2.03034821  0.06515056  1.60675972      -0.11580190  0.02188234
## Cluster 2 -0.30258300 -0.08883919 -0.25319643      -0.34028885 -0.26042083
## Cluster 3 -0.12014741  0.16785247 -0.16915057       2.03543277 -0.12093662
## Cluster 4  0.01504147  0.19402609 -0.02879161      -0.18734962 -0.17813147
## Cluster 5 -0.10955084  0.12358450  0.18193028      -0.05775191  2.40504647
##           small_business
## Cluster 1     0.11387158
## Cluster 2    -0.08216203
## Cluster 3    -0.06999735
## Cluster 4     0.24202259
## Cluster 5     0.31224183
```

We used the highest values in each cluster to identify characteristics of markets segments for NutritionH20. The results are summarized below:

1. **Suburban Dad** - These users often tweet about sports-fandom, family, religion, parenting, school, and food. They probably attend community events and most-likely prioritize products that support a balanced-lifestyle for their family and community.

2. **The Generalist** - Users in this market segment engage and speak about a wide array of topics and interests online. They are a less predictable group, but they may respond well to a wide variety of marketing strategies from NutrientH20.

3. **The Gym Bro** - These people often tweet about personal fitness and health/nutrition. We imagine that this is who Nutrient H20 markets their products toward. This market segment cares about quality, and how this product will impact their overall health and wellness.

4. **The Know-It-All** - These users often tweet about traveling, politics, computers and automobiles. This market segment probably spends too much time on reddit. They are educated, and probably opinionated. It will be important to target this market segment with detailed product information that appeals to their intellectual curiousity.

5. **The Brand Lover** - The Brand Lover - These people often tweet about cooking, beauty, photo-sharing and fashion. A demographoc that cares more about what is "popular" than what is good. These are the people who will buy NutrientH20 because they saw it on instagram or because their favorite actor drinks it.