HW9_Markdown

2024-04-11

Manay Divatia

md46245


```
library(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
```

```
library(ggplot2)
library(tidyr)
library(cluster)
social = read.csv("~/Downloads/social_marketing.csv")
summary(social)
```

```
##       X                chatter        current_events      travel
##  Length:7882        Min.   : 0.000   Min.   :0.000   Min.   : 0.000
##  Class :character   1st Qu.: 2.000   1st Qu.:1.000   1st Qu.: 0.000
##  Mode  :character   Median : 3.000   Median :1.000   Median : 1.000
##                     Mean   : 4.399   Mean   :1.526   Mean   : 1.585
##                     3rd Qu.: 6.000   3rd Qu.:2.000   3rd Qu.: 2.000
##                     Max.   :26.000   Max.   :8.000   Max.   :26.000
##  photo_sharing    uncategorized      tv_film       sports_fandom
##  Min.   : 0.000   Min.   :0.000   Min.   : 0.00   Min.   : 0.000
##  1st Qu.: 1.000   1st Qu.:0.000   1st Qu.: 0.00   1st Qu.: 0.000
##  Median : 2.000   Median :1.000   Median : 1.00   Median : 1.000
##  Mean   : 2.697   Mean   :0.813   Mean   : 1.07   Mean   : 1.594
##  3rd Qu.: 4.000   3rd Qu.:1.000   3rd Qu.: 1.00   3rd Qu.: 2.000
##  Max.   :21.000   Max.   :9.000   Max.   :17.00   Max.   :20.000
##     politics           food            family        home_and_garden
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.0000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.:0.0000
##  Median : 1.000   Median : 1.000   Median : 1.0000   Median :0.0000
##  Mean   : 1.789   Mean   : 1.397   Mean   : 0.8639   Mean   :0.5207
##  3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 1.0000   3rd Qu.:1.0000
##  Max.   :37.000   Max.   :16.000   Max.   :10.0000   Max.   :5.0000
##      music              news        online_gaming       shopping
##  Min.   : 0.0000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
```

```
##  1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 0.0000   Median : 0.000   Median : 0.000   Median : 1.000
##  Mean   : 0.6793   Mean   : 1.206   Mean   : 1.209   Mean   : 1.389
##  3rd Qu.: 1.0000   3rd Qu.: 1.000   3rd Qu.: 1.000   3rd Qu.: 2.000
##  Max.   :13.0000   Max.   :20.000   Max.   :27.000   Max.   :12.000
##  health_nutrition  college_uni     sports_playing      cooking
##  Min.   : 0.000   Min.   : 0.000   Min.   :0.0000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.: 0.000
##  Median : 1.000   Median : 1.000   Median :0.0000   Median : 1.000
##  Mean   : 2.567   Mean   : 1.549   Mean   :0.6392   Mean   : 1.998
##  3rd Qu.: 3.000   3rd Qu.: 2.000   3rd Qu.:1.0000   3rd Qu.: 2.000
##  Max.   :41.000   Max.   :30.000   Max.   :8.0000   Max.   :33.000
##       eco           computers          business         outdoors
##  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   :0.5123   Mean   : 0.6491   Mean   :0.4232   Mean   : 0.7827
##  3rd Qu.:1.0000   3rd Qu.: 1.0000   3rd Qu.:1.0000   3rd Qu.: 1.0000
##  Max.   :6.0000   Max.   :16.0000   Max.   :6.0000   Max.   :12.0000
##      crafts         automotive          art             religion
##  Min.   :0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.000
##  1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.000
##  Median :0.0000   Median : 0.0000   Median : 0.0000   Median : 0.000
##  Mean   :0.5159   Mean   : 0.8299   Mean   : 0.7248   Mean   : 1.095
##  3rd Qu.:1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.000
##  Max.   :7.0000   Max.   :13.0000   Max.   :18.0000   Max.   :20.000
##      beauty          parenting          dating           school
##  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   : 0.7052   Mean   : 0.9213   Mean   : 0.7109   Mean   : 0.7677
##  3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000
##  Max.   :14.0000   Max.   :14.0000   Max.   :24.0000   Max.   :11.0000
##  personal_fitness    fashion        small_business        spam
##  Min.   : 0.000   Min.   : 0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median : 0.000   Median : 0.0000   Median :0.0000   Median :0.00000
##  Mean   : 1.462   Mean   : 0.9966   Mean   :0.3363   Mean   :0.00647
##  3rd Qu.: 2.000   3rd Qu.: 1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :19.000   Max.   :18.0000   Max.   :6.0000   Max.   :2.00000
##      adult
##  Min.   : 0.0000
##  1st Qu.: 0.0000
##  Median : 0.0000
##  Mean   : 0.4033
##  3rd Qu.: 0.0000
##  Max.   :26.0000
```

From the summary, we can get some key metrics. What was most important to me was seeing if there were any null values. Based on this we know that there are not any which is good.

```
interest_distribution <- colSums(select(social, -X))
print(interest_distribution)
```

```
##          chatter   current_events           travel    photo_sharing
```

```
##            34671            12030            12493            21256
##    uncategorized          tv_film    sports_fandom         politics
##             6408             8436            12564            14098
##             food           family  home_and_garden            music
##            11015             6809             4104             5354
##             news    online_gaming         shopping health_nutrition
##             9502             9528            10951            20235
##      college_uni   sports_playing          cooking              eco
##            12213             5038            15750             4038
##        computers         business         outdoors           crafts
##             5116             3336             6169             4066
##       automotive              art         religion           beauty
##             6541             5713             8634             5558
##        parenting           dating           school personal_fitness
##             7262             5603             6051            11524
##          fashion   small_business             spam            adult
##             7855             2651               51             3179
```

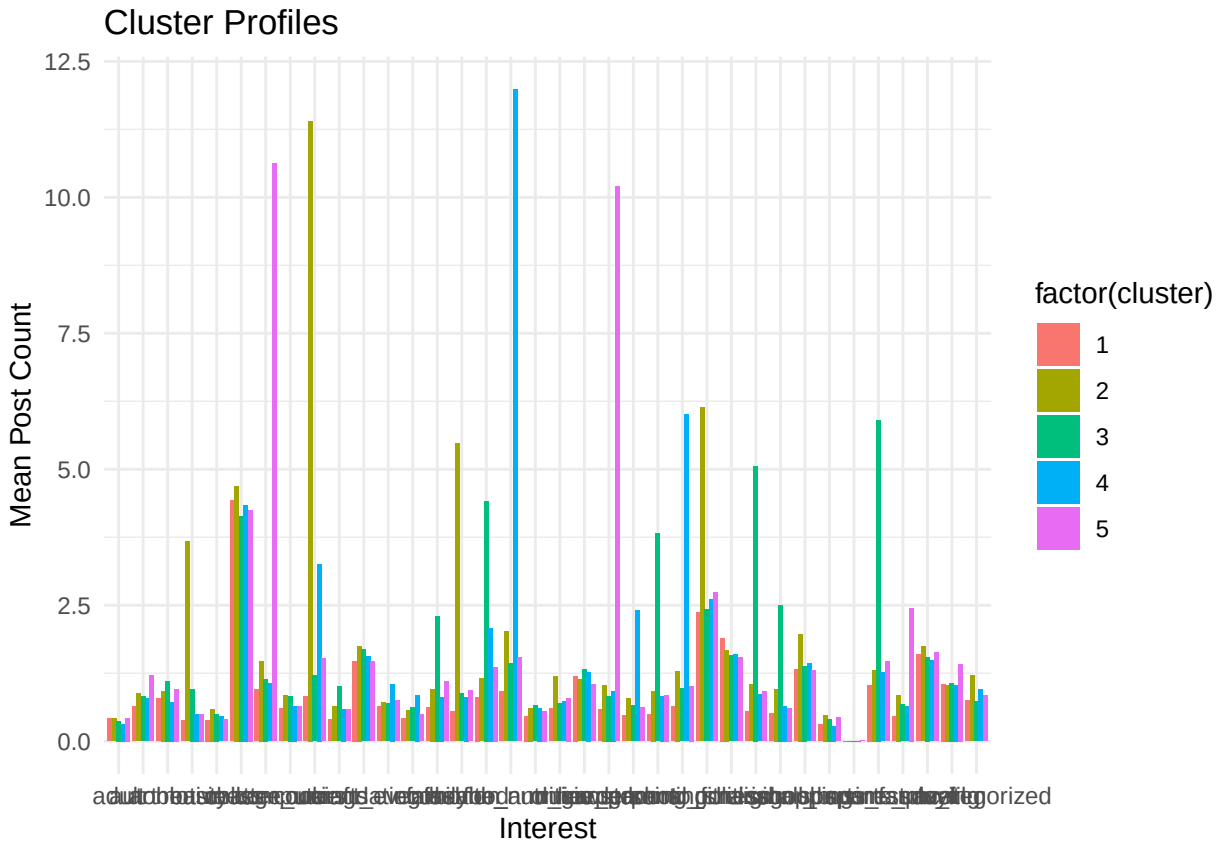
From this information, we are able to see how often each category appears. From this, we can see that chatter, health_nutrition, cooking, photo_sharing, college_uni, travel, and sports_fandom, and personal_fitness. Immediately, we can see how some of these go together. For example, personal fitness, health nutrition, and cooking are all similar topics. Additionally travel and photo sharing could also go together.

```
social = select(social, -X)
set.seed(123)
k <- 5
kmeans_model <- kmeans(social, centers = k)

social_clusters <- social %>%
  mutate(cluster = kmeans_model$cluster)

cluster_summary <- social_clusters %>%
  group_by(cluster) %>%
  summarize_all(mean)

print(cluster_summary)
```

```
## # A tibble: 5 x 37
##   cluster chatter current_events travel photo_sharing uncategorized tv_film
##     <int>   <dbl>          <dbl>  <dbl>         <dbl>         <dbl>   <dbl>
## 1       1    4.43           1.48   1.59          2.36         0.750    1.05
## 2       2    4.69           1.75   1.74          6.14         1.21     1.04
## 3       3    4.14           1.68   1.55          2.42         0.731    1.07
## 4       4    4.34           1.56   1.48          2.60         0.945    1.03
## 5       5    4.25           1.46   1.63          2.74         0.846    1.42
## # i 30 more variables: sports_fandom <dbl>, politics <dbl>, food <dbl>,
## #   family <dbl>, home_and_garden <dbl>, music <dbl>, news <dbl>,
## #   online_gaming <dbl>, shopping <dbl>, health_nutrition <dbl>,
## #   college_uni <dbl>, sports_playing <dbl>, cooking <dbl>, eco <dbl>,
## #   computers <dbl>, business <dbl>, outdoors <dbl>, crafts <dbl>,
## #   automotive <dbl>, art <dbl>, religion <dbl>, beauty <dbl>, parenting <dbl>,
## #   dating <dbl>, school <dbl>, personal_fitness <dbl>, fashion <dbl>, ...
```

```
cluster_profiles <- cluster_summary %>%
  gather(key = "interest", value = "mean_count", -cluster)
```

```
ggplot(cluster_profiles, aes(x = interest, y = mean_count, fill = factor(cluster))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Cluster Profiles",
       x = "Interest",
       y = "Mean Post Count") +
  theme_minimal()
```



This graph shows the mean post counts for each category based on the 5 clusters. We actually get good information from this. To start, we can see that the largest blue line in the middle is in cluster 4 and is the health_nutrition category. What we can alos see from this is that, unsurprisingly, the cooking and personal_fitness categories also have high mean post counts for this clusters which makes sense. We see a similar idea in cluster 5 which is interesting. Cluster 5 has 2 categories with large mean post counts which are college_uni and online_gaming. This also makes sense since a lot of college students do play video games and are likely to tweet about it.

Ultimately, by clustering and visualizing the data, we see some interesting trends in the data that we couldn't have seem if we just looked at the numbers.