Beverage Company | Segmentation Memo (V1)

Michael Spencer

2024-03-19

Assignment context

As our client (a company that produces healthy beverages) enters its next wave of growth, it needs data driven insights to inform where to fuel this expansion. First and foremost, [brand] wanted to understand the value-based mindsets and behaviors of its potential customer segments.

We fielded a survey to a representative sample of 1,000 U.S. adults interested in health and trying new things. The survey included a battery of 51 psychographic survey items to surface the segments of Americans most likely to be interested in kombucha according to several dimensions (e.g., openness and adventure, healthy lifestyles, values alignment). The segmentation is a useful heuristic for understanding and targeting [brand]'s current and potential future audiences based on values and behaviors that are key to activating kombucha drinkers.

For this assignment, we would like to:

- Create clear and concise segments from the psychograhic battery of questions
- Use the demographic variables as profiling / explanatory variables only (won't include these in the clustering process)

The goal is to group like-minded people together based on their mindsets and attitudes, not their demographic profile.

Imagine that we are the end client, and we want to know **what type of customers are out there, and which ones are more or less likely to try kombucha in particular, and why.**

Overview of Approach

Below I outline the approach taken to complete this analysis. This will include the following steps:

1. Import and clean the data
2. Prepare the data for use in NMF (Non-negative Matrix Factorization), which will provide us with clusters and their dominant features (i.e. values)
3. Cluster the data using NMF. This will include determining the optimal number of clusters to use
4. Assign the clusters to each respondent
5. Evaluate each cluster's demographic variables, including their likelihood to try kombucha
6. Analyze the values index for each cluster to understand how their values differ from the overall population

The steps above will allow us to understand the different segments of the population, their likelihood to try kombucha, and their values. This will allow us to make recommendations to the client on who these segments are and which to target.

Setup

```
# List of required packages
required_packages <- c("NMF", "labelled", "haven", "tidyverse")

# Loop through the list of required packages and install/load any that are missing
for (p in required_packages) {
    if (!require(p, character.only = TRUE)) {
    install.packages(p, dependencies = TRUE)
    }
    library(p, character.only = TRUE)
}
```

```
# Filepaths
path_assignment_data <- "assignment_data.rds"

# Load and clean data
assignment_data <-
    readRDS(path_assignment_data) %>%
    # Convert all demographic labelled variables to factors
    # Leaving the segmentation variables in numerical form for use in NMF algorithm
    mutate(
        across(
            c(starts_with("d_"), -"d_hh_adults", -"d_hh_children", "d_kom_aware"),
            as_factor
        )
    )
```

Data Manipulation & Cleaning

First, we glance at a summary of this data to identify any oddities such as outliers or missing values.

```
assignment_data %>%
    summary()
```

The only missing data in the *demographic* variables seems to be in response to "How often do you drink Kombucha" (`d_kom_drink`). This represents respondents who indicated they have never tried kombucha (see below). We'll leave these be for now as they will not be used for segmentation and can be handled later as needed.

```
assignment_data %>%
    filter(is.na(d_kom_drink)) %>%
    count(d_kom_tried)
```

```
## # A tibble: 3 x 2
##   d_kom_tried                      n
##   <fct>                        <int>
## 1 No                             476
## 2 Never heard of it              165
## 3 I don't remember/I'm not sure   16
```

We do see some missing data in the *segmentation* variables.

```
assignment_data %>%
    filter(if_any(starts_with("seg_"), is.na)) %>%
    count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    20
```

Given that our our segmentation methodology NMF requires a complete matrix (i.e. no missing values), I will remove rows with missing values. Since these only comprise 2% of the data (i.e. 20 of 1,000 respondents), this should not meaningfully impact the analysis.

```
assignment_data <-
    assignment_data %>%
    filter(!if_any(starts_with("seg_"), is.na))
```

To prepare the data for clustering analysis, we will:

- Create a dataframe with only the segmentation variables
- Create a dataframe with only the demographic variables, which will be used to characterize the clusters after we have identified them

```
# Select just the columns that start with seg_ and convert it to a matrix for use
# in NMF functions
nmf_data <-
    assignment_data %>%
    select(starts_with("seg_")) %>%
    as.matrix()

# Select demographic variables (including the repondent ID)
demo_data <-
    assignment_data %>%
    select(-starts_with("seg_")) %>%
    mutate(
        household_size = d_hh_adults + d_hh_children,
    region = case_when(
          d_state %in% c("Maine", "New Hampshire", "Vermont", "Massachusetts",
                                    "Rhode Island", "Connecticut", "New York",
                                    "New Jersey", "Pennsylvania") ~ "Northeast",
          d_state %in% c("Ohio", "Michigan", "Indiana", "Wisconsin", "Illinois",
                                    "Minnesota", "Iowa", "Missouri", "North Dakota",
                                    "South Dakota", "Nebraska", "Kansas") ~ "Midwest",
          d_state %in% c("Delaware", "Maryland", "District of Columbia", "Virginia",
                                    "West Virginia", "North Carolina", "South Carolina",
                                    "Georgia", "Florida", "Kentucky", "Tennessee",
                                    "Mississippi", "Alabama", "Oklahoma", "Texas",
                                    "Arkansas", "Louisiana") ~ "South",
          d_state %in% c("Idaho", "Montana", "Wyoming", "Nevada", "Utah",
                                    "Colorado", "Arizona", "New Mexico", "Alaska",
```

```
                                    "Washington", "Oregon", "California",
                                    "Hawaii") ~ "West"
    )
    )
```

Analysis

In this section, I:

- Determine the optimal number of clusters to use in NMF by examining the silhouette score and dispersion for each number of clusters, as well as their distribution. Ideally these will be kept high while creating an evenly distributed set of clusters.
- Run NMF using the optimal number of clusters. This will provide us with the clusters and their dominant features (i.e. values)
- Assign each respondent to their cluster
- Examine the demographic and values for each cluster, using these to characterize the clusters

Data Clustering

The estimation is run using between 2 and 8 clusters. Realistically, we would like more than 2-3 clusters to provide a more nuanced understanding of the population, but we also don't want to overcomplicate things.

```
rank_estimation <- nmfEstimateRank(nmf_data, 2:8, method = "KL", seed = 123)
```

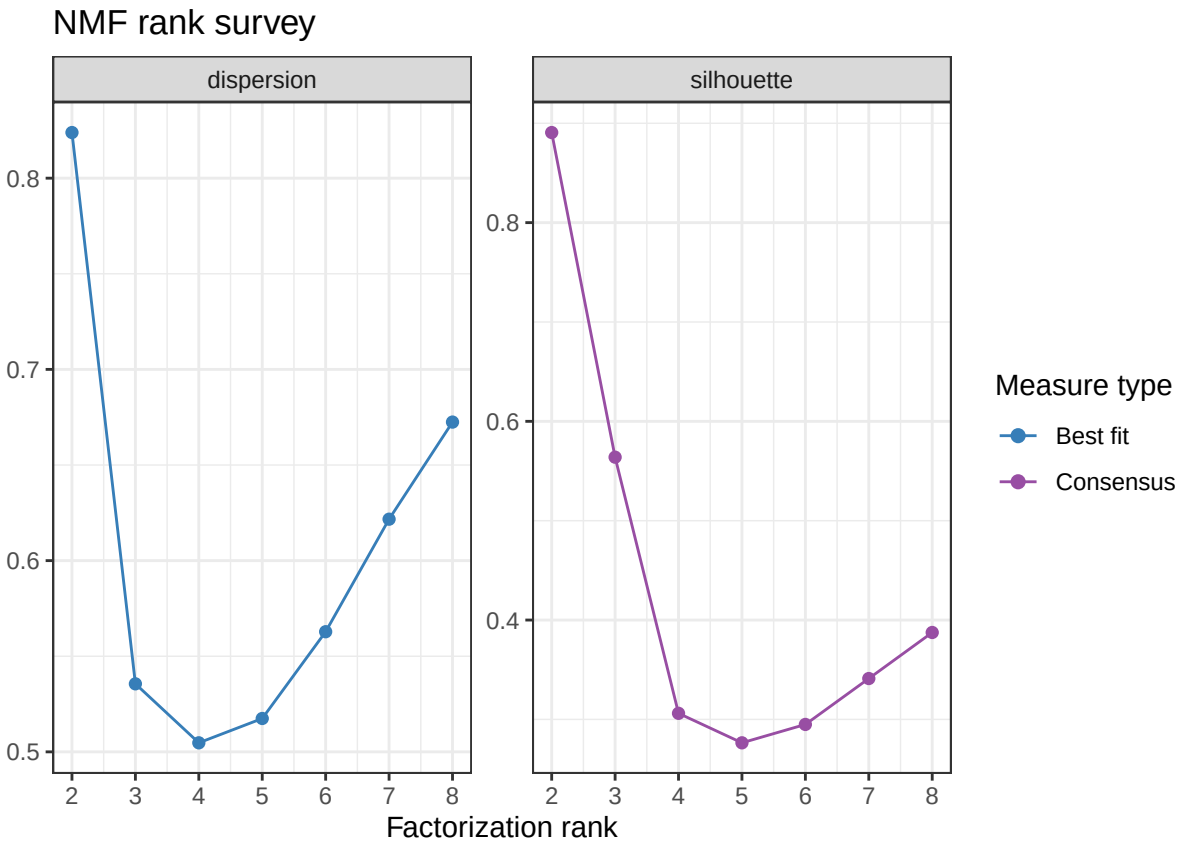
The following summary table shows us the silhouette score and dispersion for each number of clusters. The silhouette score is a measure of how similar an object is to its own cluster compared to other clusters. The dispersion is a measure of how consistently a given data point is assigned to the same cluster. In both cases, the max is 1 and higher is better.

```
model_summary <-
    summary(rank_estimation) %>%
    select(
        rank,
        silhouette.consensus,
        dispersion
    )

model_summary
```

```
##   rank silhouette.consensus dispersion
## 2    2            0.8906854  0.8238626
## 3    3            0.5639122  0.5355735
## 4    4            0.3061182  0.5047307
## 5    5            0.2764329  0.5174506
## 6    6            0.2949266  0.5628246
## 7    7            0.3411279  0.6216669
## 8    8            0.3874811  0.6724473
```

```
plot(
    rank_estimation,
    what = c("dispersion", "silhouette.consensus")
)
```



In addition to the error terms above, another important consideration is the distribution of respondents across our clusters and the coherence of those clusters. We want to avoid having a few large clusters and many small clusters, and we want to ensure that the clusters make logical sense. Let's investigate the first point by examining the relative size of each cluster for each number of clusters 2 through 8.

```
# Run an nmf model on each rank in the model_summary table and store
# metrics of interest
model_summary <-
    model_summary %>%
    mutate(
        model_info =
            map(
                rank,
                ~ nmf(nmf_data, rank = .x, method = "KL", seed = 123)
            ),
        smallest_segment_pct =
            map_dbl(
                model_info,
                ~ predict(., "features") %>%
                as_tibble() %>%
                count(value) %>%
```

```
                mutate(pct = round(n / sum(n), 2)) %>%
                pull(pct) %>%
                min()
            ),
        largest_segment_pct =
            map_dbl(
                model_info,
                ~ predict(., "features") %>%
                as_tibble() %>%
                count(value) %>%
                mutate(pct = round(n / sum(n), 2)) %>%
                pull(pct) %>%
                max()
            )
    )

model_summary %>%
    select(-model_info)
```

```
##   rank silhouette.consensus dispersion smallest_segment_pct largest_segment_pct
## 2    2            0.8906854  0.8238626                 0.45                0.55
## 3    3            0.5639122  0.5355735                 0.27                0.42
## 4    4            0.3061182  0.5047307                 0.24                0.28
## 5    5            0.2764329  0.5174506                 0.11                0.29
## 6    6            0.2949266  0.5628246                 0.05                0.35
## 7    7            0.3411279  0.6216669                 0.03                0.32
## 8    8            0.3874811  0.6724473                 0.05                0.23
```

These groupings are out:

- **2-3** clusters appear very good based solely on the error terms above (dispersion and the silhouette score are both high), however 2-3 clusters isn't the most useful in the real world. Ideally, we'd have more.
- **6-8** clusters also appear to perform well based on the error terms, but result in some exceedingly small clusters (i.e. <= 5% of the population) which we'd like to avoid

4-5 clusters appear to perform *okay* based on the error terms but they result in a more even distribution of clusters. Let's continue on with 5 clusters, so that we retain some additional nuance by having the additional cluster.

I'm mostly interested in seeing how evenly distributed the clusters are and whether they are coherent based on demographic variables (i.e. do the clusters actually make sense).

```
model_summary$model_info[[4]] %>%
    predict("features") %>%
    as_tibble() %>%
    count(value) %>%
    mutate(pct_of_whole = round(n / sum(n), 2)) %>%
    rename(cluster = value)
```

```
## # A tibble: 5 x 3
##   cluster     n pct_of_whole
```

```
##   <fct>   <int>        <dbl>
## 1 1         222         0.23
## 2 2         112         0.11
## 3 3         149         0.15
## 4 4         215         0.22
## 5 5         282         0.29
```

Based on this evaluation, 5 clusters could be a good choice. The clusters are relatively evenly distributed with the exception of one that is a bit smaller and one that is a bit larger, however that's okay. Let's examine some info about these clusters to see how usable they might be.

```
# Assign clusters to respondents based on the rank=4 nmf model.
# Row is indexed at 1, so rank 4=row 3
demo_data$cluster <- predict(model_summary$model_info[[4]], "features")

# Kombucha interest levels
demo_data %>%
  transmute(
    cluster,
    interested_in_kom = ifelse(
      d_interested %in% c("Very interested", "Extremeley interested"),
      "Interested in kombucha",
      "Not interested"
    ) %>%
      as.factor(),
    tried_kom = ifelse(
        d_kom_tried == "Yes", "Tried kombucha", "Has not tried"
    ) %>%
        as.factor(),
    drink_kom_frequency = ifelse(
      d_kom_drink %in% c(
        "Regularly (a few times a week)",
        "Often (a few times a month)"
      ),
      "Drinks kom multiple times per month",
      "Less frequently"
    ) %>%
      as.factor()
  ) %>%
  pivot_longer(-cluster) %>%
    group_by(name, cluster, value) %>%
    count() %>%
    group_by(name, cluster) %>%
    mutate(pct = round(n / sum(n), 2)) %>%
    ungroup() %>%
    select(cluster, value, pct) %>%
    filter(
        value %in% c(
            "Interested in kombucha",
            "Tried kombucha",
            "Drinks kom multiple times per month"
        )
    ) %>%
    pivot_wider(names_from = value, values_from = pct) %>%
```

```
    # Normalize % who drink kombucha based on those who have tried it
    mutate(
        `Drinks kom multiple times per month` = round(
            `Drinks kom multiple times per month` / `Tried kombucha`,
            2
        )
    )
```

```
## # A tibble: 5 x 4
##   cluster Drinks kom multiple times pe~1 Interested in kombuc~2 `Tried kombucha`
##   <fct>                           <dbl>                  <dbl>            <dbl>
## 1 1                                0.68                   0.34             0.5
## 2 2                                0.52                   0.31             0.48
## 3 3                                0.29                   0.42             0.38
## 4 4                                0.22                   0.27             0.18
## 5 5                                0.37                   0.29             0.27
## # i abbreviated names: 1: `Drinks kom multiple times per month`,
## #   2: `Interested in kombucha`
```

These clusters appear to have different levels of interest in kombucha and varied levels of experience trying it.

Interest in kombucha: 1, 2 and 3 appear more interested in kombucha and have more experience trying it than clusters 4 and 5. 1 Is the most engaged and 4 the least

1. 34% interested / 50% have tried / 68% of those who have tried it, drink multiple times per month
2. 31% interested / 48% have tried / 52% of those who have tried it, drink multiple times per month
3. 42% interested / 38% have tried / 29% of those who have tried it, drink multiple times per month
4. 27% interested / 18% have tried / 22% of those who have tried it, drink multiple times per month
5. 29% interested / 27% have tried / 37% of those who have tried it, drink multiple times per month

This could provide for good nuance in targeting and understanding the population. Let's continue examining each clusters' demographic composition.

Demographic Analysis

Note: Likely could have coded these variables differently to make analysis a bit easier, but wanted to keep the more granular levels to understand the nuances.

```
# Get all demographic variables
demo_vars <-
    demo_data %>%
    select(-response_id, -cluster, -d_kom_aware) %>%
    colnames()

for (demo_var in demo_vars) {
  print(
    demo_data %>%
          select(
            cluster,
            demo_var
          ) %>%
```

```
          pivot_longer(-cluster) %>%
            group_by(name, cluster, value) %>%
            count() %>%
            group_by(name, cluster) %>%
            mutate(pct = round(n / sum(n), 2)) %>%
            ungroup() %>%
            select(cluster, value, pct) %>%
            pivot_wider(names_from = value, values_from = pct)
  )
}
```

Living Location: 1 and 2 lean more urban compared to 4 and 5

1. 49% Urban / 40% Suburban / 11% Rural
2. 41% Urban / 47% Suburban / 12% Rural
3. 36% Urban / 48% Suburban / 16% Rural
4. 24% Urban / 56% Suburban / 20% Rural
5. 26% Urban / 54% Suburban / 21% Rural

Gender: Relatively even split across all, however cluster 3 is more heavily female (57%)

Age: Generally, 1 is younger than 2 is younger than 3, etc, with 5 being the oldest

1. 36% 18-30 / 46% 31-45
2. 39% 18-30 / 52% 31-64
3. 55% <= 45 / 45% 46+ (Pretty evenly distributed across all age groups)
4. 47% 46-64 / 25% 65+
5. 43% 46-64 / 28% 65+

Race: 1 and 2 are more likely to be non-white, while 4 and 5 are more likely to be white

1. 42% Non-White –> 20% AA / 10% Asian / 10% Hispanic
2. 45% Non-White –> 20% AA / 19% Hispanic
3. 31% Non-White
4. 21% Non-White
5. 26% Non-White

Education: Relatively even split across all, 2 is slightly less educated; 5 slightly more

1. 54% have some college degree
2. 51% have some college degree
3. 62% have some college degree
4. 59% have some college degree
5. 67% have some college degree

Income: Relatively even split here, 3 earns slightly more

1. 47% earn $50K or more
2. 44% earn $50K or more
3. 62% earn $50K or more
4. 49% earn $50K or more

5. 54% earn $50K or more

People in household: Households in 1 and 2 seem to be more composed of friends/families as opposed to 4 and 5, which are likely older retirees

1. 54% have 3+ people in house / 52% have children
2. 40% have 3+ people in house / 35% have children
3. 38% have 3+ people in house / 27% have children
4. 23% have 3+ people in house / 17% have children
5. 27% have 3+ people in house / 20% have children

Marital Status: No large differences here, 2 seems to contain more respondents who have never married

1. 49% married
2. 36% married
3. 46% married
4. 48% married
5. 56% married

Political View: 1 is more liberal than 2, which is more liberal than 3, etc. Very correlated with age and urbanity

1. 41% liberal / 40% moderate
2. 46% liberal / 34% moderate
3. 38% liberal / 34% moderate
4. 25% liberal / 40% moderate
5. 29% liberal / 38% moderate

Employment: Employment decreases and retirement increases as we go from 1 to 5

1. 50% working full-time / 8% retired
2. 44% working full-time / 13% retired
3. 43% working full-time / 25% retired
4. 33% working full-time / 29% retired
5. 35% working full-time / 32% retired

Religion: No large trends worth reporting

Region: Respondents primarily in American South and American West

1. 44% from South / 23% from West
2. 41% from South / 27% from West
3. 30% from South / 25% from Midwest
4. 37% from South / 25% from West
5. 36% from South / 23% from West

Healthy Attitude/Lifestyle: 1 is composed of more health conscious, younger individuals than 2; 5 is composed of more health conscious, older individuals than 4

1. 43% say healthy diet is extremely important / 44% always maintain one
2. 33% say healthy diet is extremely important / 35% always maintain one

3. 23% say healthy diet is extremely important / 22% always maintain one
4. 8% say healthy diet is extremely important / 9% always maintain one
5. 43% say healthy diet is extremely important / 41% always maintain one

Adventurous with new foods: Majority are interested (very or extremely) in trying new foods, except cluster 4

1. 38% extremely interested in trying new foods
2. 34% extremely interested in trying new foods
3. 30% extremely interested in trying new foods
4. 12% extremely interested in trying new foods
5. 30% extremely interested in trying new foods

We will use these to describe each cluster in an additional report. Next I'll examine the values index for each cluster to understand how the values of its members' differ from the overall population.

Values Index Analysis

```
# Assign clusters to respondents based on the rank=4 nmf model.
# Row is indexed at 1, so rank 4=row 3
assignment_data$cluster <- predict(model_summary$model_info[[4]], "features")

index_data <-
    assignment_data %>%
    select(
        cluster,
        contains("seg_")
    ) %>%
    set_names(var_label(., null = "fill")) %>%
    pivot_longer(-cluster) %>%
    mutate(name = str_extract(name, "(?<=\\?\\s-\\s).*")) %>%
    group_by(name) %>%
    mutate(
        overall_pct_agree = round(sum(value >= 3) / n(), 2)
    ) %>%
    ungroup() %>%
    group_by(name, overall_pct_agree, cluster) %>%
    summarize(
        cluster_pct_agree = round(sum(value >= 3) / n(), 2)
    ) %>%
    mutate(index_value = round(cluster_pct_agree / overall_pct_agree * 100, 0)) %>%
    select(
        cluster,
        name,
        overall_pct_agree,
        cluster_pct_agree,
        index_value
    )
```

```
# Top values for each cluster
index_data %>%
```

```
    group_by(cluster) %>%
    arrange(desc(index_value)) %>%
    slice_head(n = 5) %>%
    arrange(cluster, desc(index_value)) %>%
    split(.$cluster)
```

```
# Bottom values for each cluster
index_data %>%
    group_by(cluster) %>%
    arrange(desc(index_value)) %>%
    slice_tail(n = 5) %>%
    arrange(cluster, desc(index_value)) %>%
    split(.$cluster)
```

Results & Recommendations

The above tells us the questions for which a given cluster thinks most differently from the overall population. Understanding these values will help us characterize the clusters. I do this in the associated report rather than here, in the rmarkdown.