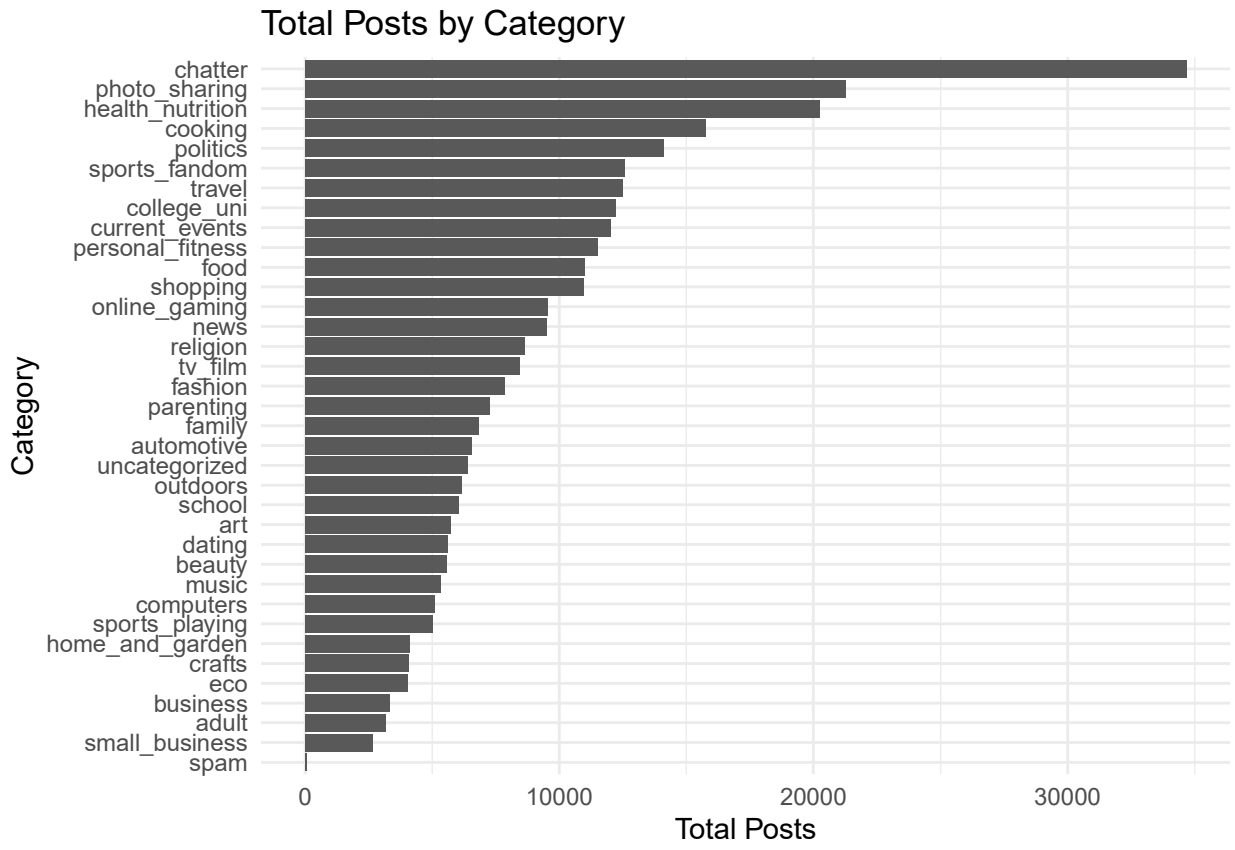
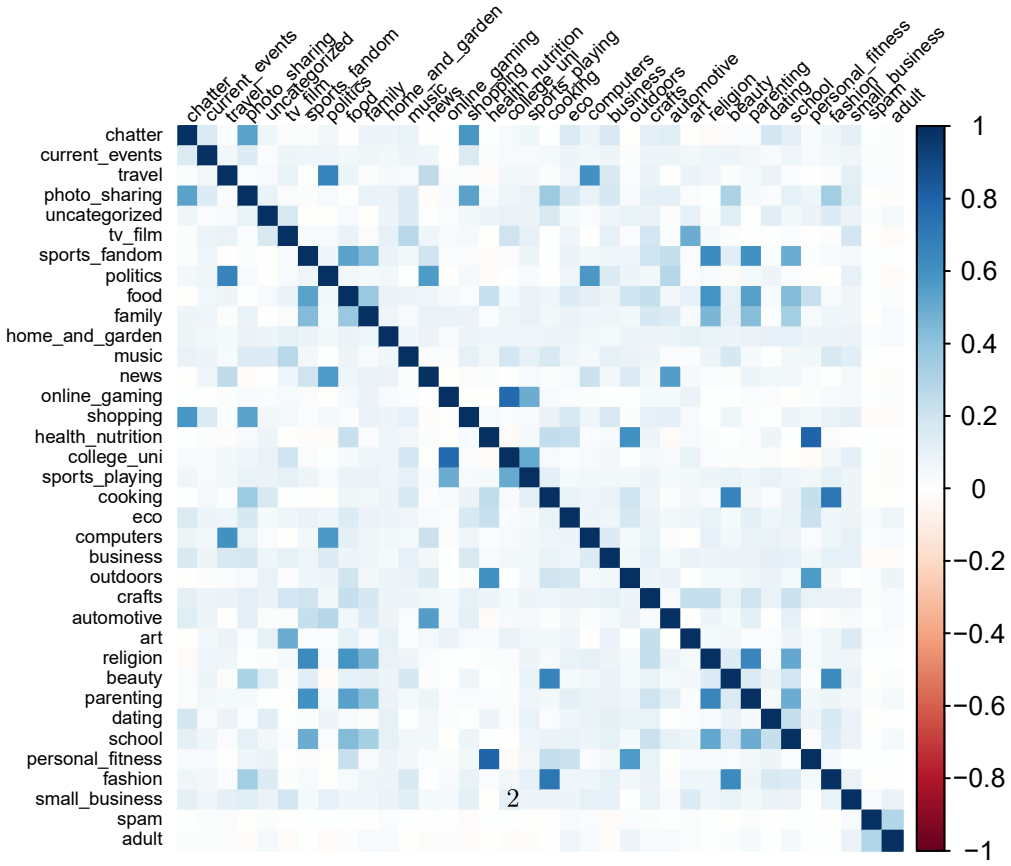
Market segmentation

EDA

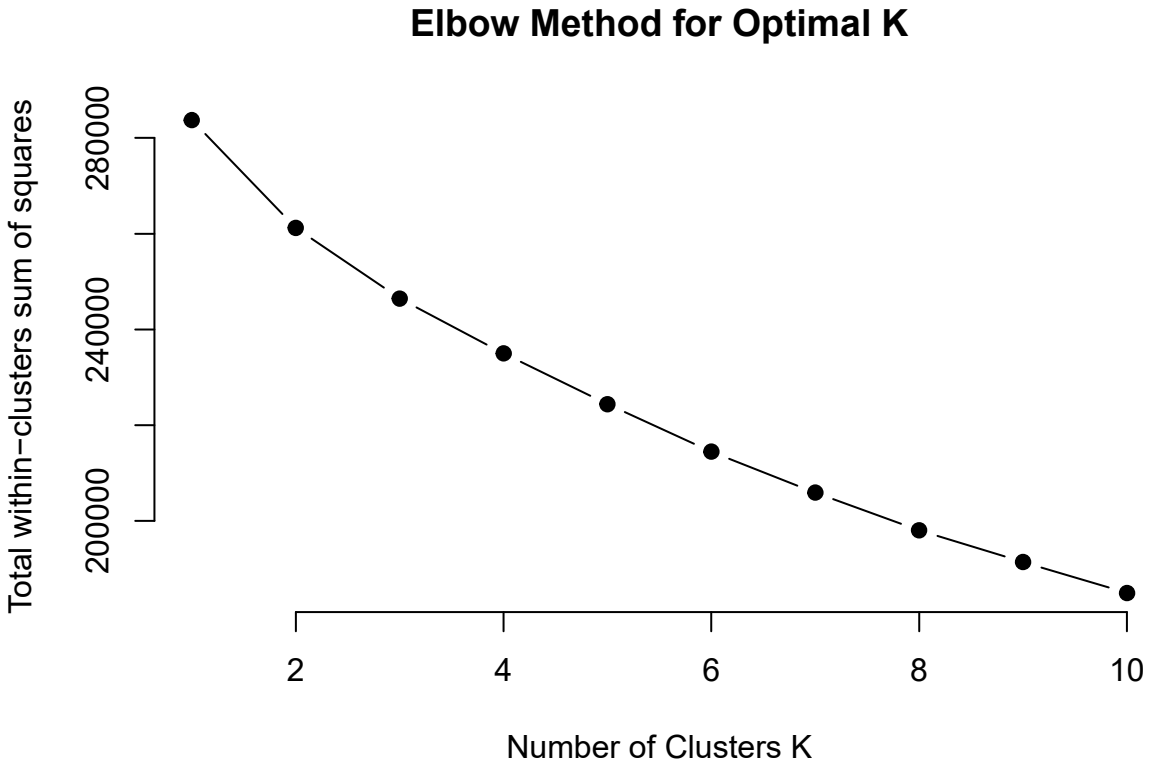


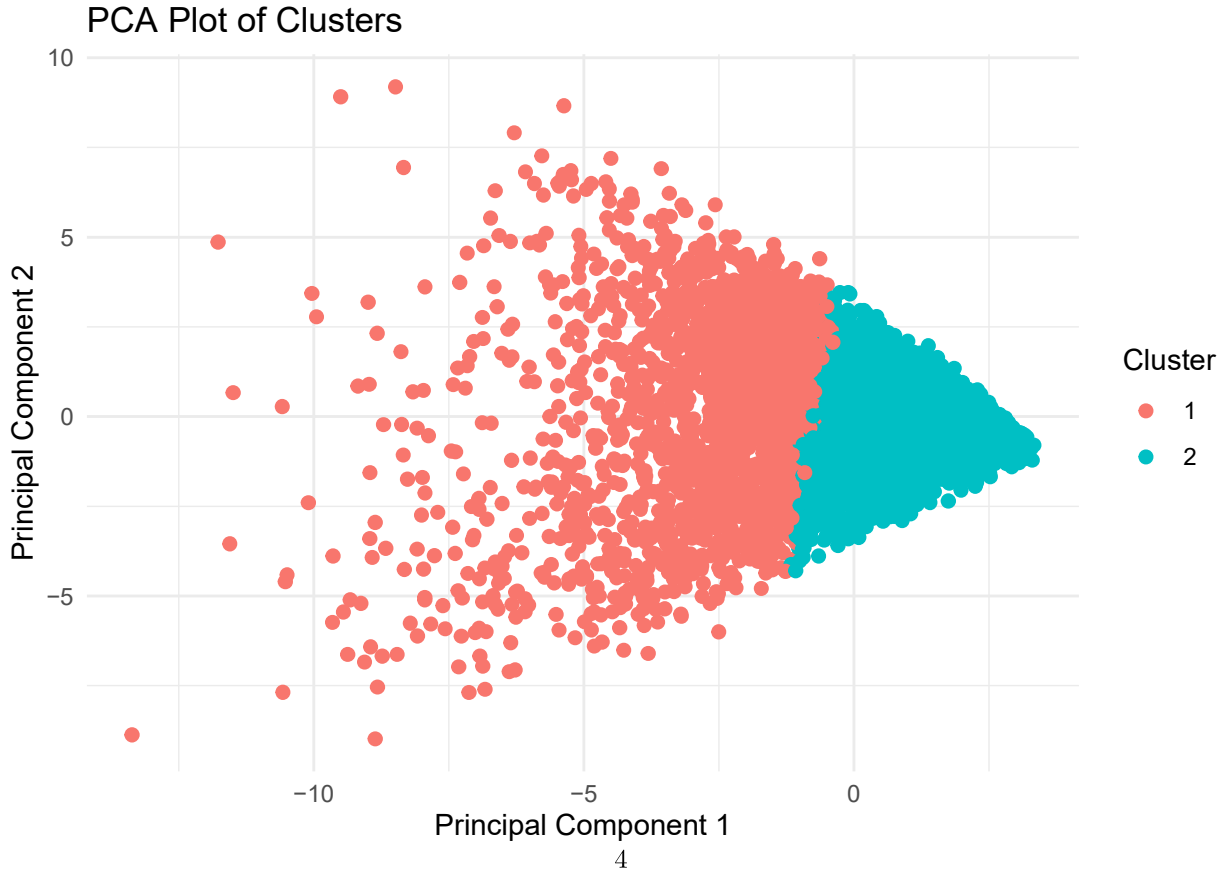


The chatter category is the most popular category, as deemed by the human annotators. This may be partially explained by the annotators possibly using the "chatter" category to capture posts that didn't clearly fit into the other categories.

Looking at the correlation matrix, some correlations make a lot of sense in the real world. For example, there is a high correlation between personal_fitness and health_nutrition. Another example is the higher correlation between fashion and beauty.

Clustering


Elbow Method for Optimal K
Total within−clusters sum of squares
200000
240000
280000
2
4
6
8
10
Number of Clusters K


PCA Plot of Clusters
Principal Component 2
10
5
0
−5
−10
−5
0
Principal Component 1
Cluster
1
2

```
## # A tibble: 2 x 37
##   cluster chatter_mean current_events_mean travel_mean photo_sharing_mean
##   <fct>          <dbl>               <dbl>       <dbl>               <dbl>
## 1 1               5.67                1.83        2.27                4.13
## 2 2               3.87                1.40        1.30                2.10
## # i 32 more variables: uncategorized_mean <dbl>, tv_film_mean <dbl>,
## #   sports_fandom_mean <dbl>, politics_mean <dbl>, food_mean <dbl>,
## #   family_mean <dbl>, home_and_garden_mean <dbl>, music_mean <dbl>,
## #   news_mean <dbl>, online_gaming_mean <dbl>, shopping_mean <dbl>,
## #   health_nutrition_mean <dbl>, college_uni_mean <dbl>,
## #   sports_playing_mean <dbl>, cooking_mean <dbl>, eco_mean <dbl>,
## #   computers_mean <dbl>, business_mean <dbl>, outdoors_mean <dbl>, ...
```

My approach

I first scaled the data to prepare it for clustering. Then I created an elbow plot to determine the optimal the number of clusters, which appeared to be k = 2. Afterwards, for visualization purposes, I reduced the dimensionality of the data using PCA and plotted the data. I also created a table to analyze the mean value of each category within each cluster to understand the characteristics of each segment.

Some Insights:

Cluster 1:

chatter_mean = 3.870221:

On average, users in Cluster 1 have about 3.87 posts classified under the "chatter" category. This suggests that users in this cluster tend to engage in general conversation or miscellaneous topics on social media.

current_events_mean = 1.399390:

Users in Cluster 1 post about 1.4 times on average about current events. This indicates a moderate interest in news or happenings in the world among users in this segment.

travel_mean = 1.299049:

With an average of 1.3 posts about travel, users in this cluster have some interest in travel, but it's not their primary focus.

photo_sharing_mean = 2.101598:

An average of 2.1 posts about photo sharing suggests that users in this cluster enjoy sharing photos, but it's not their dominant activity.

Overall for Cluster 1:

The users in this cluster seem to be fairly engaged in general conversation ("chatter") with moderate interest in current events, travel, and photo sharing. This cluster might represent a more casual or broad-interest segment of NutrientH20's audience, with no strong focus on any specific topic.

Cluster 2:

chatter_mean = 5.672869:

Users in Cluster 2 have a higher average number of posts in the "chatter" category (5.67), indicating that they are more active in general conversation compared to Cluster 1.

current_events_mean = 1.832107:

With an average of 1.83 posts about current events, this cluster has a slightly stronger interest in news and world events than Cluster 1.

travel_mean = 2.274340:

Users in Cluster 2 post more about travel (2.27 posts on average), indicating a higher interest in travel-related content.

photo_sharing_mean = 4.131545:

The average of 4.13 posts about photo sharing suggests that users in this cluster are significantly more engaged in sharing photos compared to Cluster 1.

Overall for Cluster 2:

Cluster 2 users appear to be more active and engaged across several categories, particularly in "chatter" and "photo sharing." This segment might represent a more socially active group, with a strong inclination towards sharing content and participating in discussions about various topics, including travel and current events.

Key Takeaways:

Cluster 1 could be characterized as a segment with moderate engagement, possibly representing users who are generalists or casual users of social media, without a strong focus on any particular category.

Cluster 2 appears to be more engaged and social, particularly in sharing photos and participating in general chatter. This segment might be more valuable for targeted social media campaigns that focus on visual content and interactive communication.