Q2 RMD

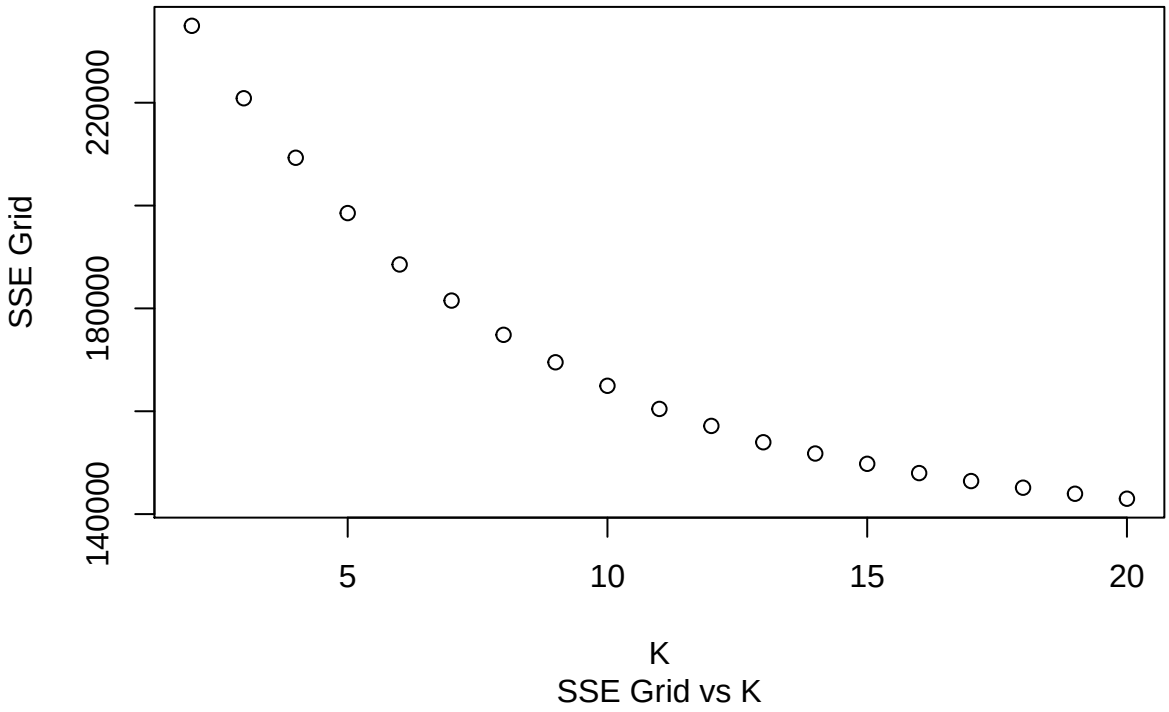
Tejaswi Pukkalla

April 26, 2019

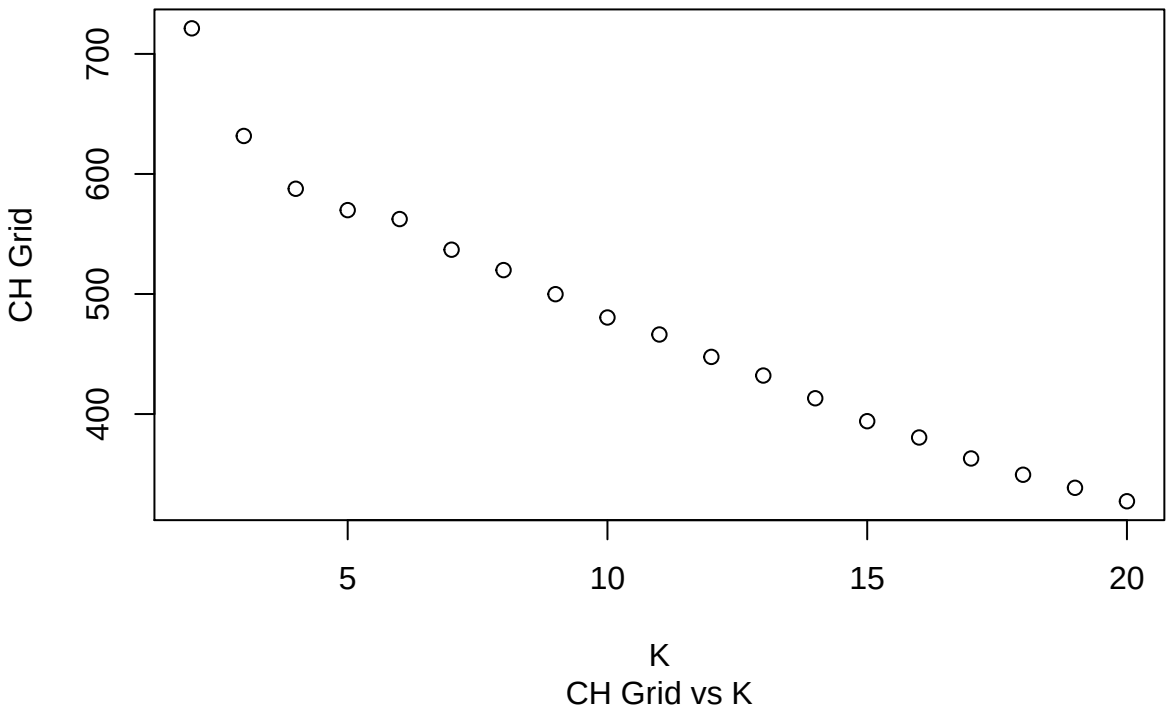
Question 2: Market Segmentation

Our objective is to help NutrientH2O understand its social-media audience a little better, so it could work on targeting its market segment more aptly. We start off the same way we start any data analysis, cleaning up the data first. We try our best to clear out out the bots that might have slipped through the initial filtering. We delete users marked as spam. We measure the amount of adult content in the user's messages and delete users that have more than a quarter of their messages flagged as adult content. After removing the unused attributes now that we have deleted users with non zero values in those attributes, we move on to centering and scaling the data.

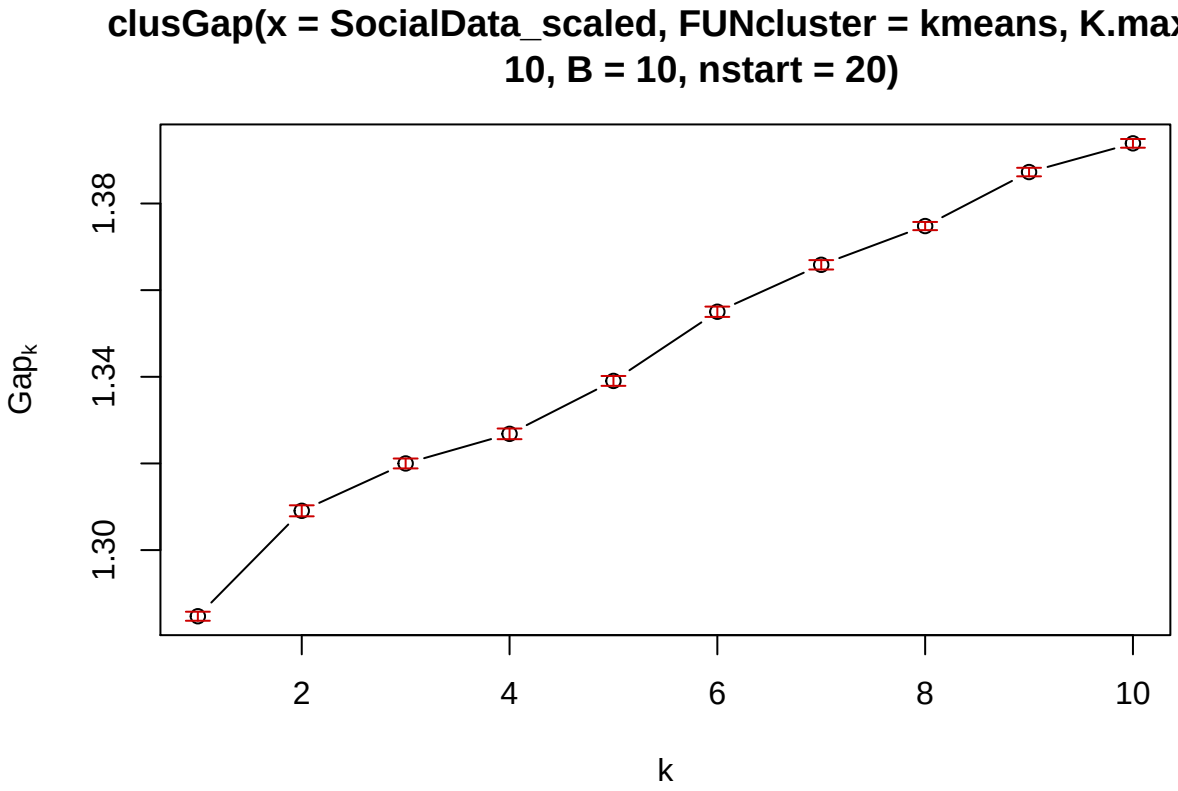
We first use K-means clustering to measure the SSE error and check the various values of SSE error for different values of K.



Next we use K-means clustering again, but this time we measure the CH grid values for their respective K values.

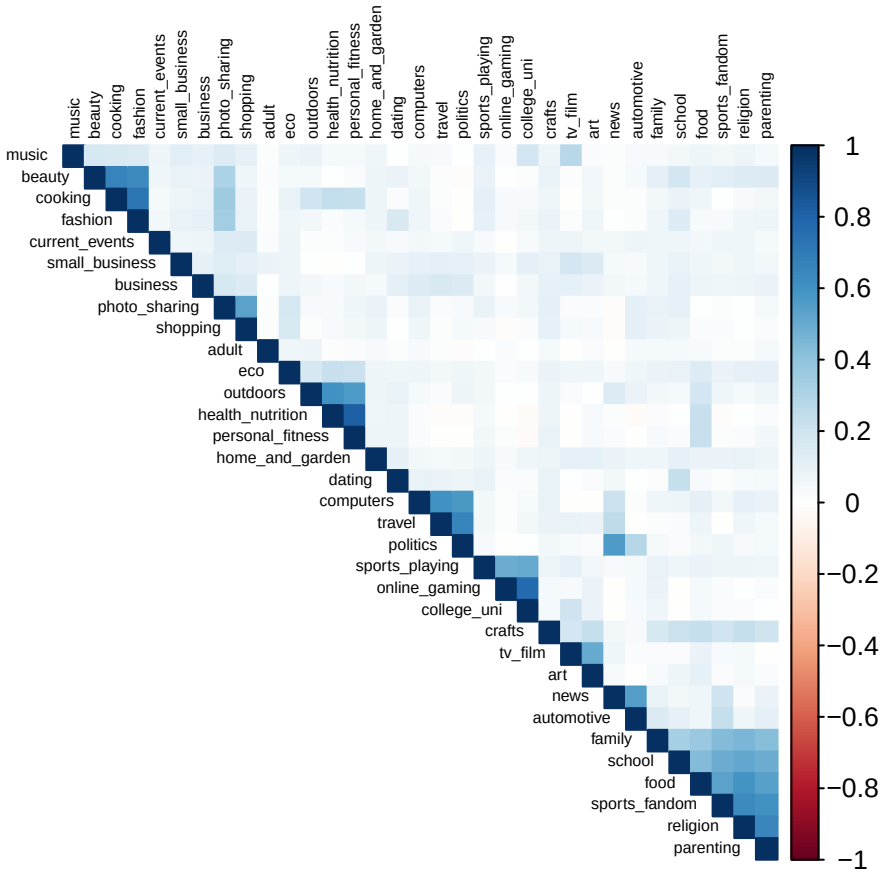


Next, we use the cluster gap function to see where the dip in cluster gap is.



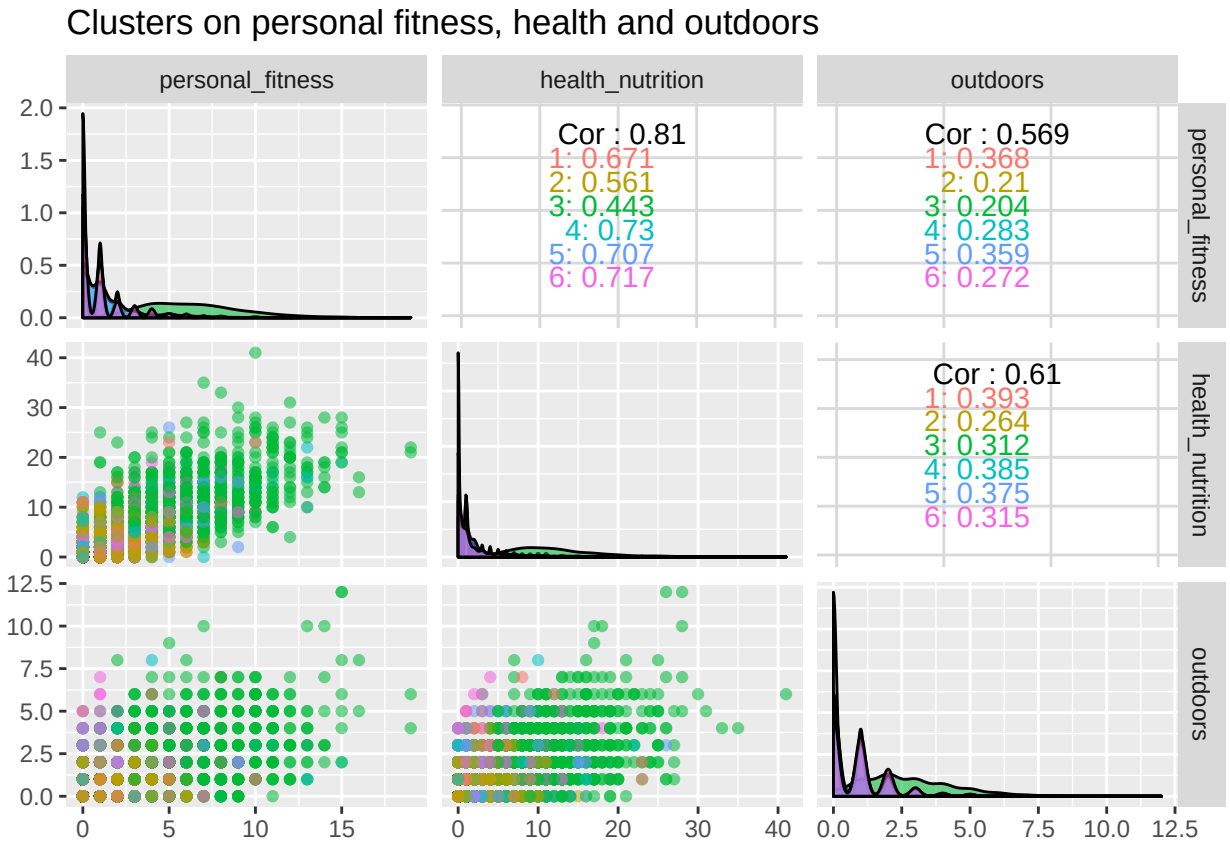
From the above plots, we chose K=6 intuitively as the optimum K value. There is a slight dip in the CH Grid Vs K graph at K=6. However, there is no clear dip in Cluster gaop at all showing this isn't a convex cluster.

To analyse the K-means clustering, let's first take a look into the correlation between these various post indicators/factors.

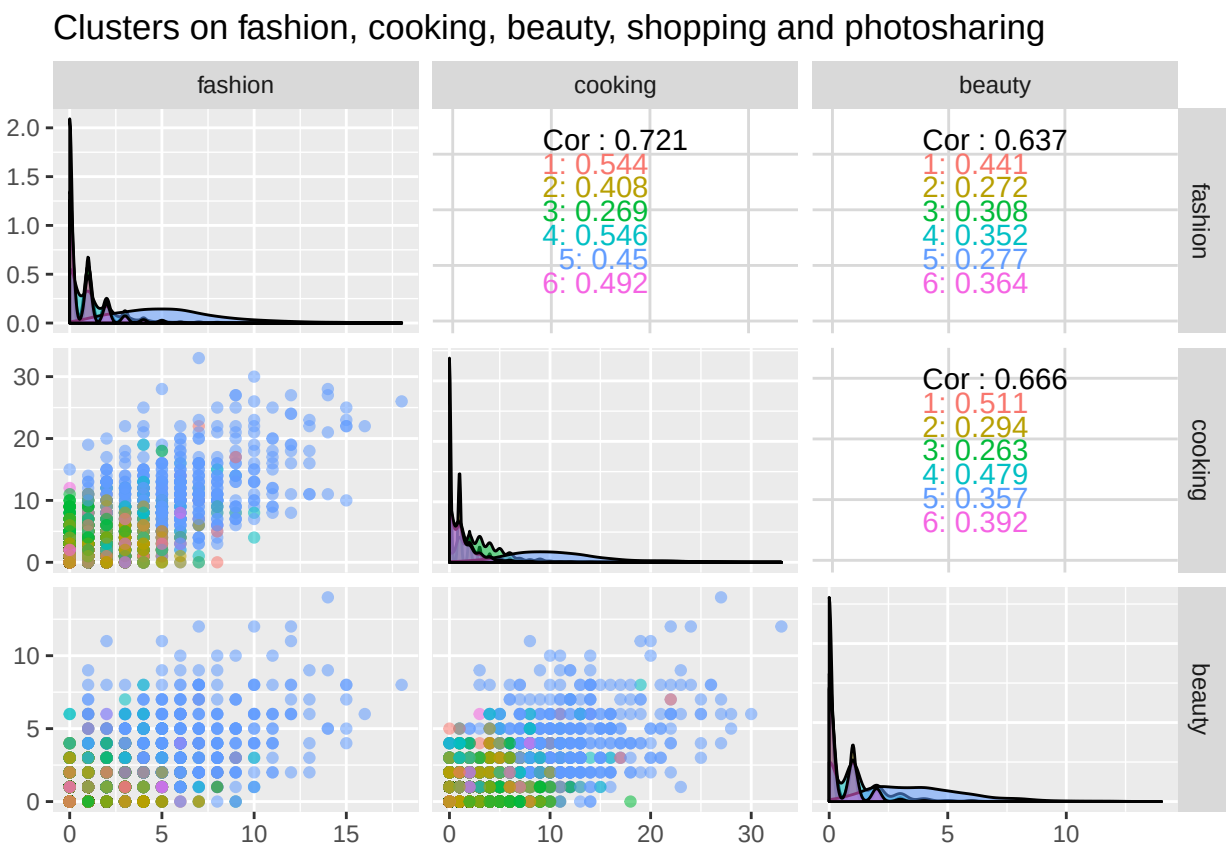

music
beauty
cooking
fashion
current_events
small_business
business
photo_sharing
shopping
adult
eco
outdoors
health_nutrition
personal_fitness
home_and_garden
dating
computers
travel
politics
sports_playing
online_gaming
college_uni
crafts
tv_film
art
news
automotive
family
school
food
sports_fandom
religion
parenting
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

From the above graph, we pick the factors that have highest correlation with each other and try to analyze the K-means clustering results of these factors.

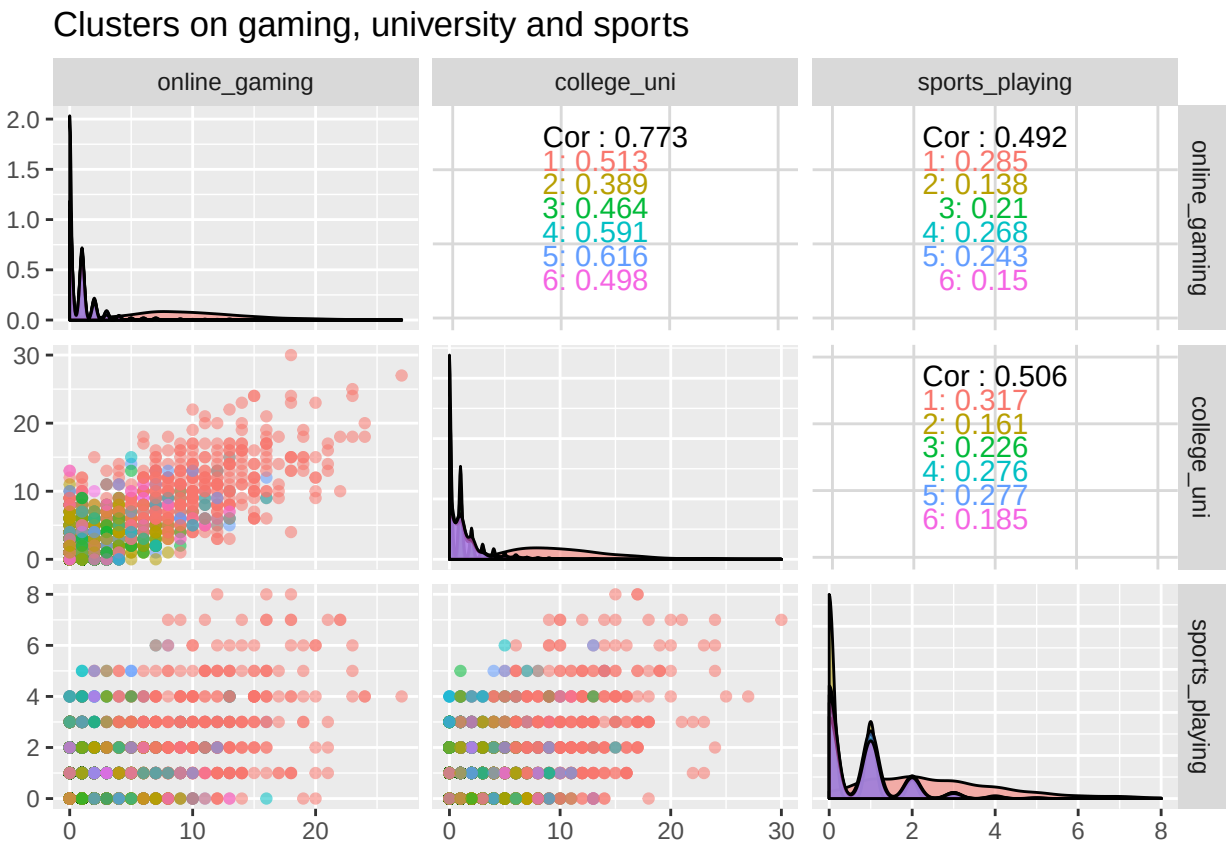
i) Personal fitness, Health, Outdoors



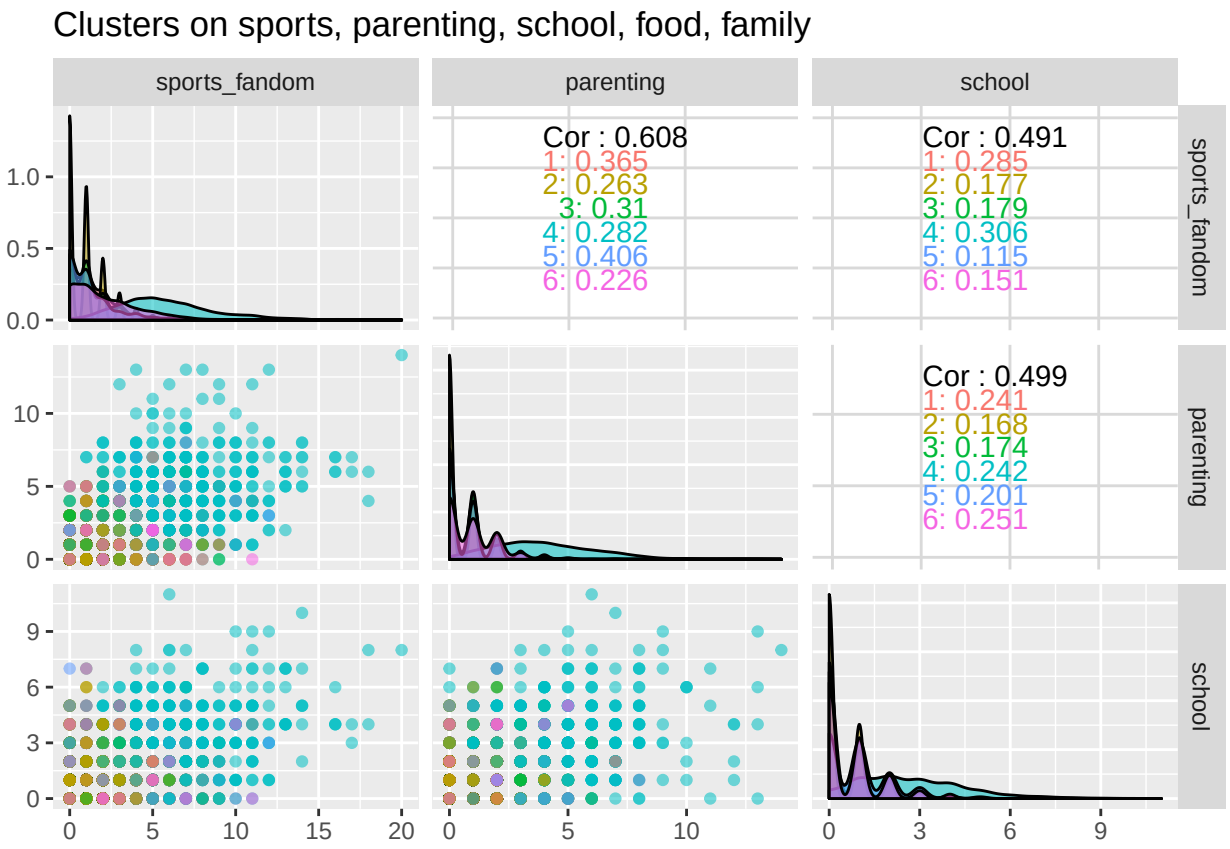
ii) Fashion, Cooking, Beauty, Shopping, Photo sharing



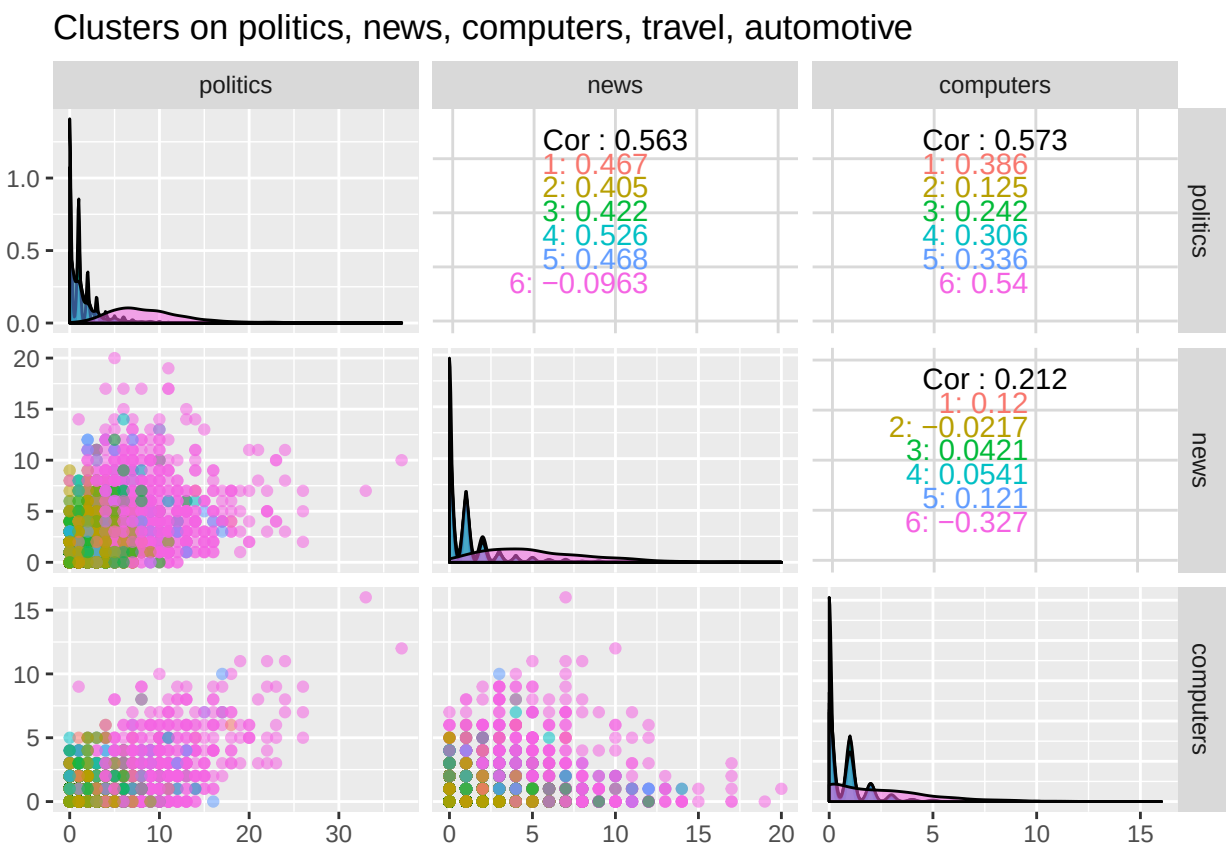
iii) Gaming, College, Sports



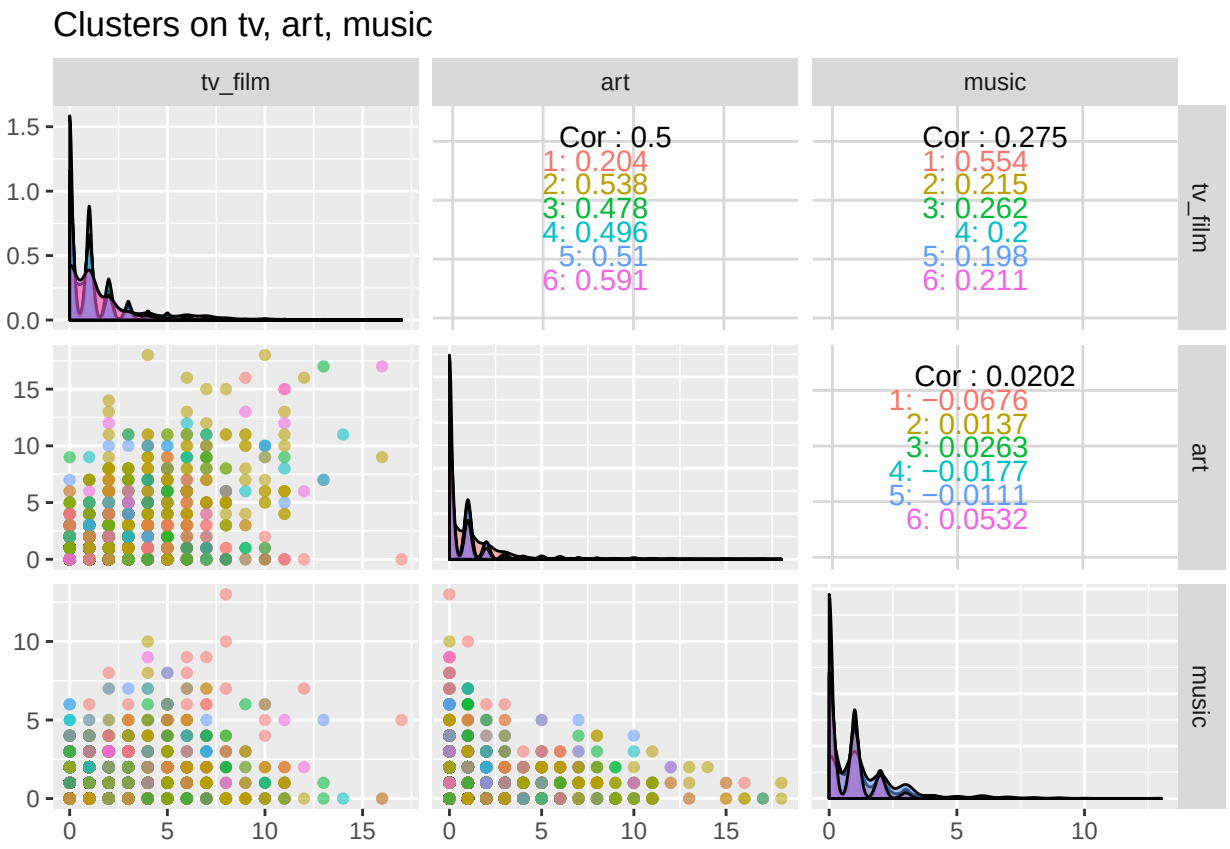
iv) Sports fan, Parenting, School, Food, Family



v) Politics, News, Computers, Travel, Automotive

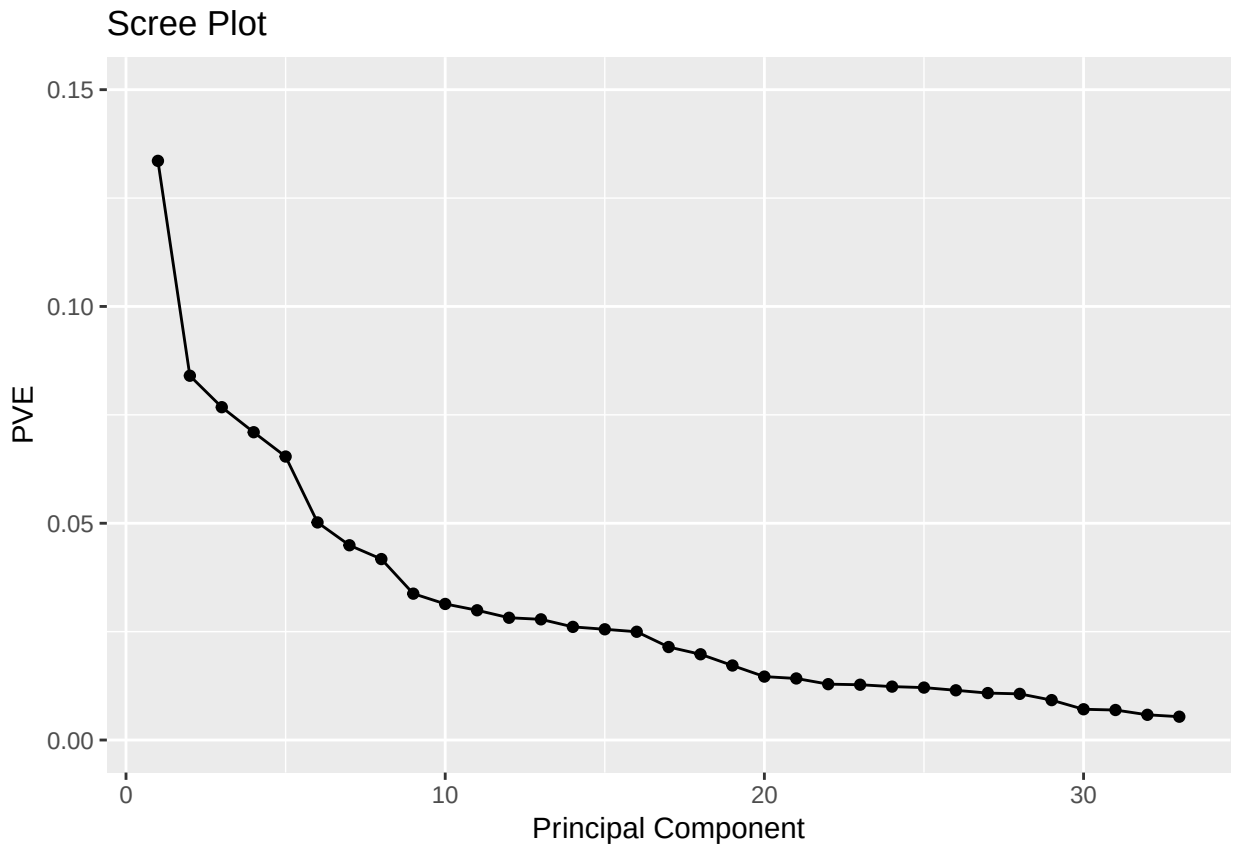


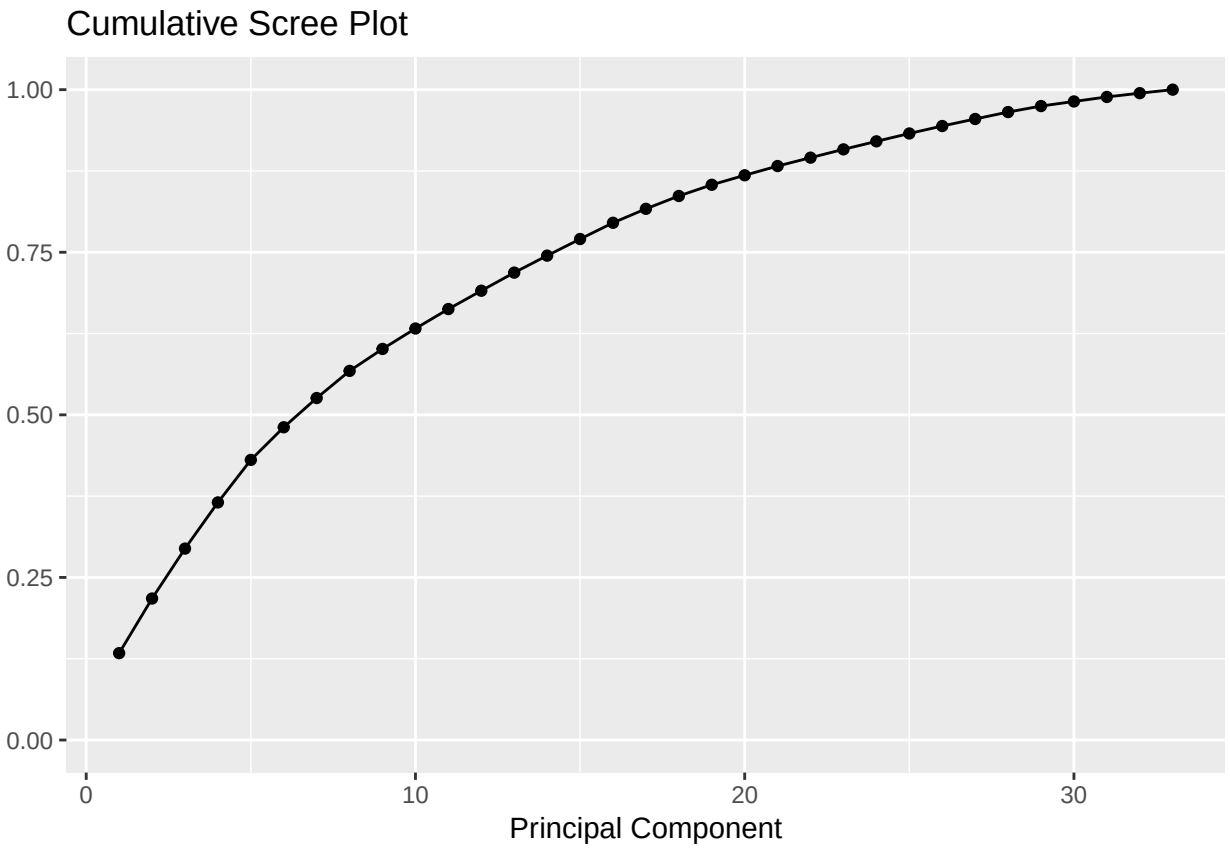
vi) TV, Art, Music



From the above graphs, we can see that there isn't much we can try to extract in terms of segments from the above data. Very few factors are highly explained by other factors. Hence, we decided to do the PCA analysis to extract them into different market segments so that we can see what are the factors explaining the most amount of the variance in their messages. As we have done multiple edits to the original dataset, we again extract the original dataset from the website and clean it up again using the same criterion as before.

After doing the PCA analysis of the data and calculating both variance and the proportion of variance explained by these low dimensional summaries, we plot the Scree plot and the cumulative scree plot to get an idea on the different segments explaining these differences.


Scree Plot
0.15
0.10
0.05
0.00
PVE
0
10
20
30
Principal Component



From the above graphs, we can see that almost 50% of the variance is explained by the 6 out of 33 factors. Hence, this reaffirms our intiial value of K as well. We get the following market segments that NutrientH2O can use to segment its audience and target them with sharper messages.

The first market segment is parents with children who go to school, are religious, have interest in food and are sports fans. These are your average parents who can be targeted better by advertising the product as a sports drink for their kids. The second market segment is the exact same as the first. Hence, we remove the repetition. The third market segment is people into politics, travel, computers, news and automotive. These are your typical white collar employees who can be targeted better by advertising the drink as a health drink. The fourth segment is people who are heavily into health, fitness, outdoors, politics and news. These are your regular outdoorsy health conscious people who coud be of any age or any employment. They can again be targeted better by advertising NutrientH2O as a health drink. The fifth segment is people who post about beauty, fashion, cooking, photosharing and shopping. This segment would easily represent both college students as well as married women. They can be targeted better by offering better prices and packaging. The sixth segment is people who post about gaming, playing sports, their colleges, automotives and cooking. This is a very typical yet accurate description of most college students. They can targeted better by advertising the drink as a sports drink and offering better prices as this is a highly price sensitive group. The seventh segment is people who post about automotives, shopping, photo sharing, news and current events. This seems to be quite a generic group combining different kinds of interests. Hence, removing the repeated segment, we get a total of six different market segments as mentioned above that can be targeted better and advertised to in different ways.