

Homework 4

Frank Chou, Tejaswi Pukkalla

April 23, 2019

Question 1: Clustering and PCA

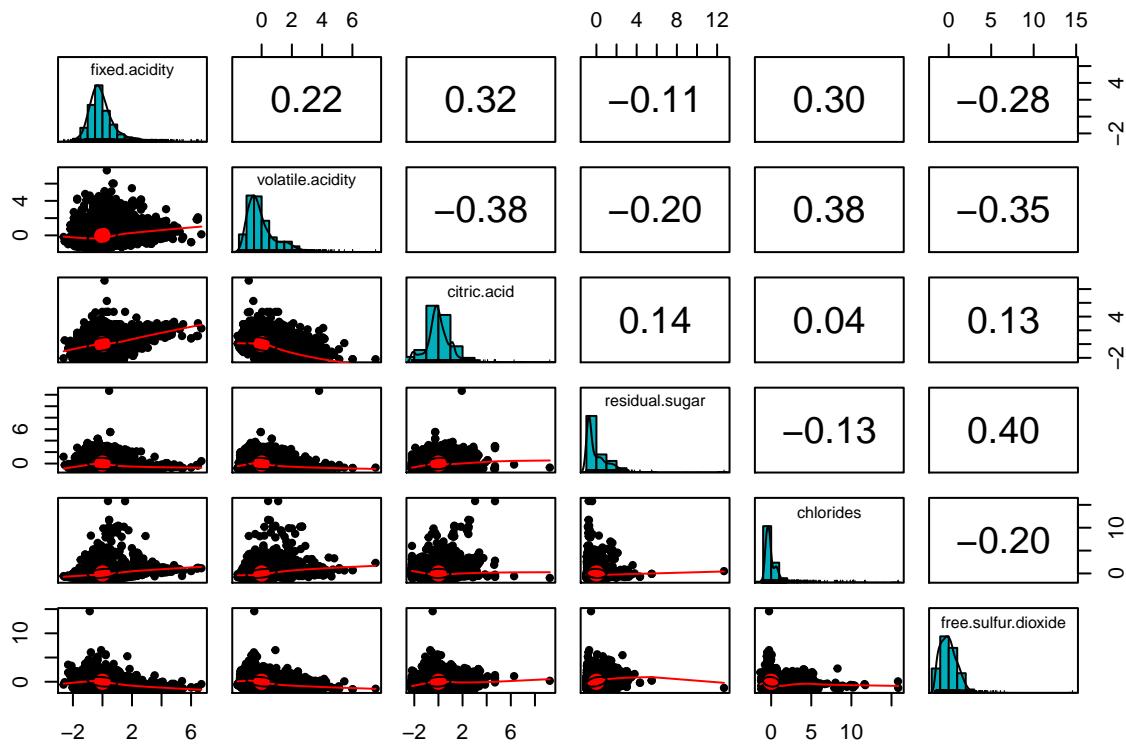
Given the fact that the **wines.csv** dataset consists of 11 chemical properties and 6,500 different bottles of *vinho verde* wine from northern Portugal, we have some difficulty determining which dimensionality reduction technique would be the better model for the task. With only 11 variables of the chemical properties of the wine, **Principal Component Analysis (PCA)** offers a easy-to-use tool to reduce the total number of variables down to a handful to work with, however at the same time, important information might be lost if the data is reduced to just a handful. On the other hand, **clustering** offers a method where all of the variables are at play, meaning that all 11 variables would be considered when determining which wine would be a part of which artificial group we are creating. However the drawback is that given the inherent randomization of cluster selection and our algorithm settings, we would have minute yet inconsistent results.

Principal Component Analysis

Our first dimensionality reduction technique would be Principal Component Analysis or PCA. In this case, the goal is to find low-dimensional summaries of high-dimensional data sets. As mentioned before, given the fact that our dataset only has 11 feature variables, the determination of what is high and low is a factor here. 11 variables by itself, is a rather low number of variables in play. However given its mathematical foundations, utilizing linear algebra to determine a vector subspace of the overall data, we capture the entirety of the data when determining our principal components.

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|----------|
| (Intercept) | 5.8184 | 0.0100 | 582.87 | 0.0000 |
| scoresPC1 | 0.0382 | 0.0057 | 6.66 | 0.0000 |
| scoresPC2 | -0.1742 | 0.0063 | -27.55 | 0.0000 |
| scoresPC3 | -0.1508 | 0.0080 | -18.85 | 0.0000 |

With the results of our PCA regression, we find that with just three PCA components, we are able to find a model that has **residual standard error** 0.8046135 that captures a significant portion of the observations within the data. In addition, each coefficient's p-values indicate that each are well within the confidence level needed to utilize each principal component accurately.



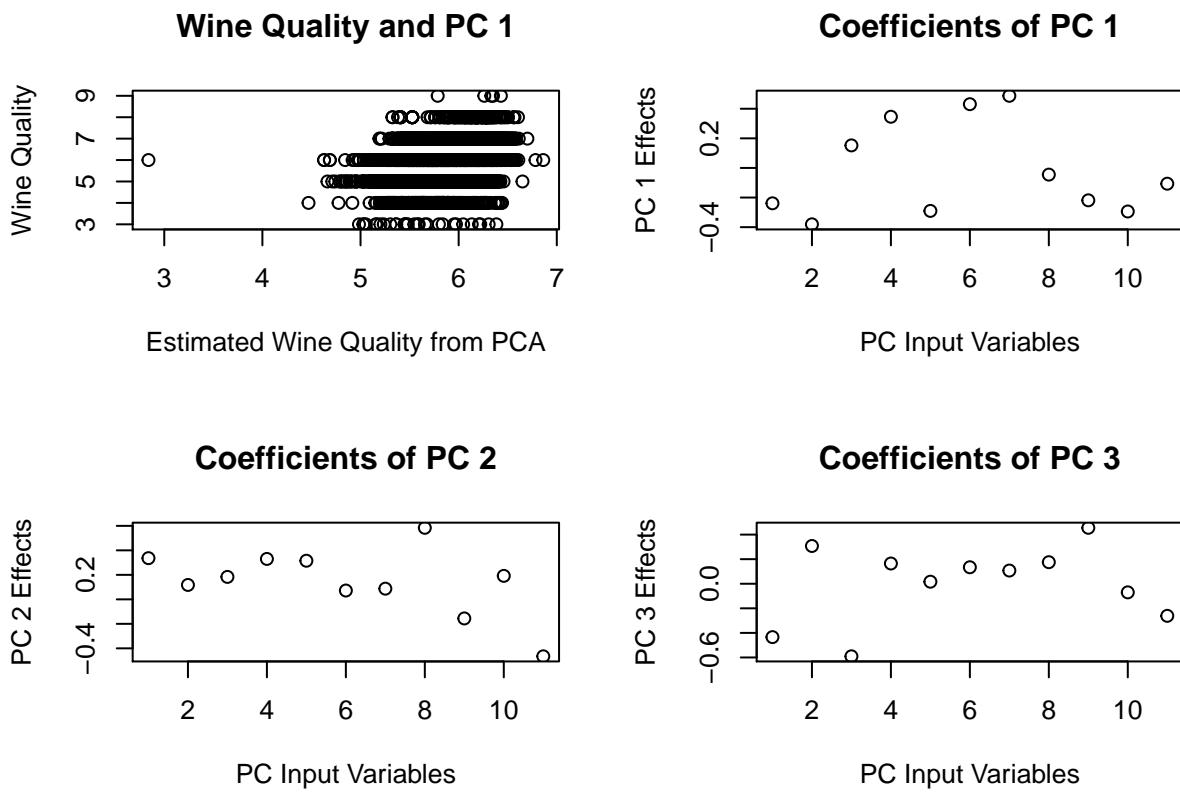
In this matrix of different plots, we take 6 from the 11 descriptor variables in the dataset and display the following:

- Bi-variate scatter plots below the diagonal
- Histograms on the diagonal
- Pearson correlation above the diagonal.

In the bi-variate scatter plots, we find a visual correlation between different variables. By utilizing PCA we can safely presume that linear combinations of similar variables would be a suitable approach in decomposing the data to a handful of variables.

While in the histograms on the diagonal, we see that once we normalize the underlying dataset, we would have a large concentration of values in the left hand side of the total range. Further supporting the notion that there is high correlation of similar wines.

Lastly, in the Pearson correlation, the rather low (relative to zero) values hints that there is little support for multicollinearity among the variables. Ultimately, PCA is one of two approaches we used to understand the data.



Before we move to our other approach, clustering, we present a 2x2 graph depicting the relationship of each PCA component among another. Given the scattered distribution of points, we find that there is considerable difficulty discerning the quality of wine based on the betas derived from each principal component.

Clustering

Our second method would be to apply a clustering algorithm on the data. Here we will apply a supervised learning algorithm to determine which wine type and quality level each of the wines in our dataset would be. Given our rather small categorical classifications: **red** or **white** this approach would be accurate enough to correctly classify the majority of the wines in question. Given our linear scale, be able to predict the wine quality within a close parameter of the underlying wine judged quality level. With only two clusters to predict, we have an easier time implementing the code to achieve better results.

```
## [1] "Cluster 1"

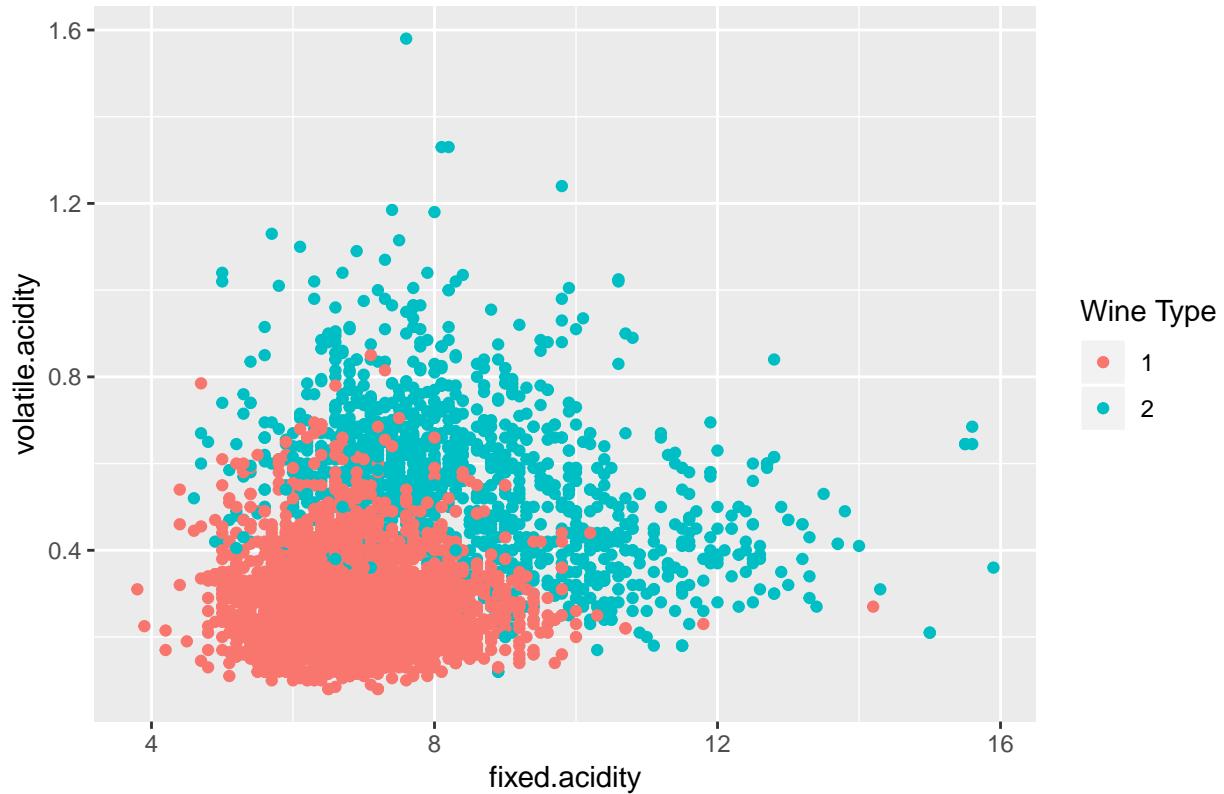
##      fixed.acidity    volatile.acidity      citric.acid
##      6.85066928        0.27395181        0.33576400
##      residual.sugar      chlorides free.sulfur.dioxide
##      6.40047364        0.04519007       35.62149918
## total.sulfur.dioxide      density          pH
##      138.66422982       0.99401827      3.18893740

## [1] "Cluster 2"

##      fixed.acidity    volatile.acidity      citric.acid
##      8.29433272        0.53412553        0.26794028
##      residual.sugar      chlorides free.sulfur.dioxide
##      2.61060329        0.08812249       15.44485070
## total.sulfur.dioxide      density          pH
##      47.92138940       0.99670402      3.30598416
```

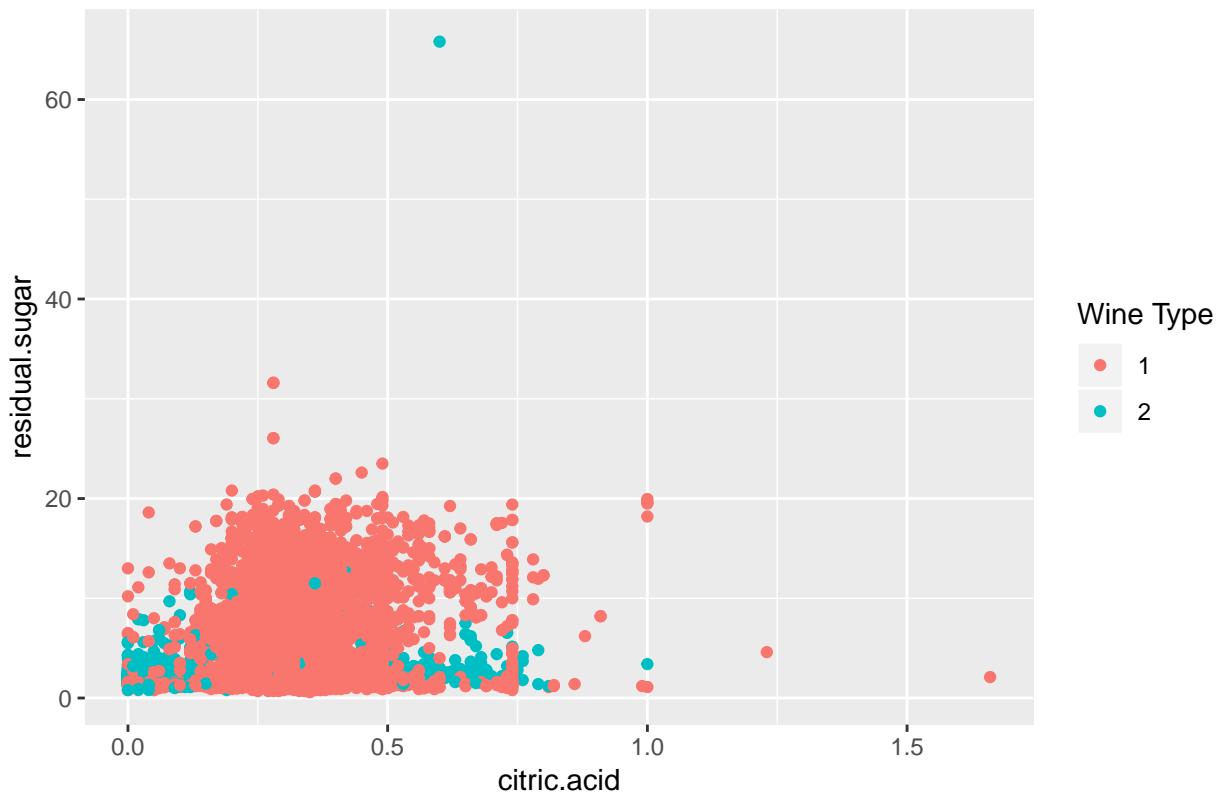
Here we have the criteria that the algorithm devised to determine which wine would be assigned to which cluster category. We see that given two clusters to work with, either one presumably to determine whether or not it is white or red wine, we can see distinct differences in all of the cutoff levels for each variable.

Volatile and Fixed Acidity by Wine Type



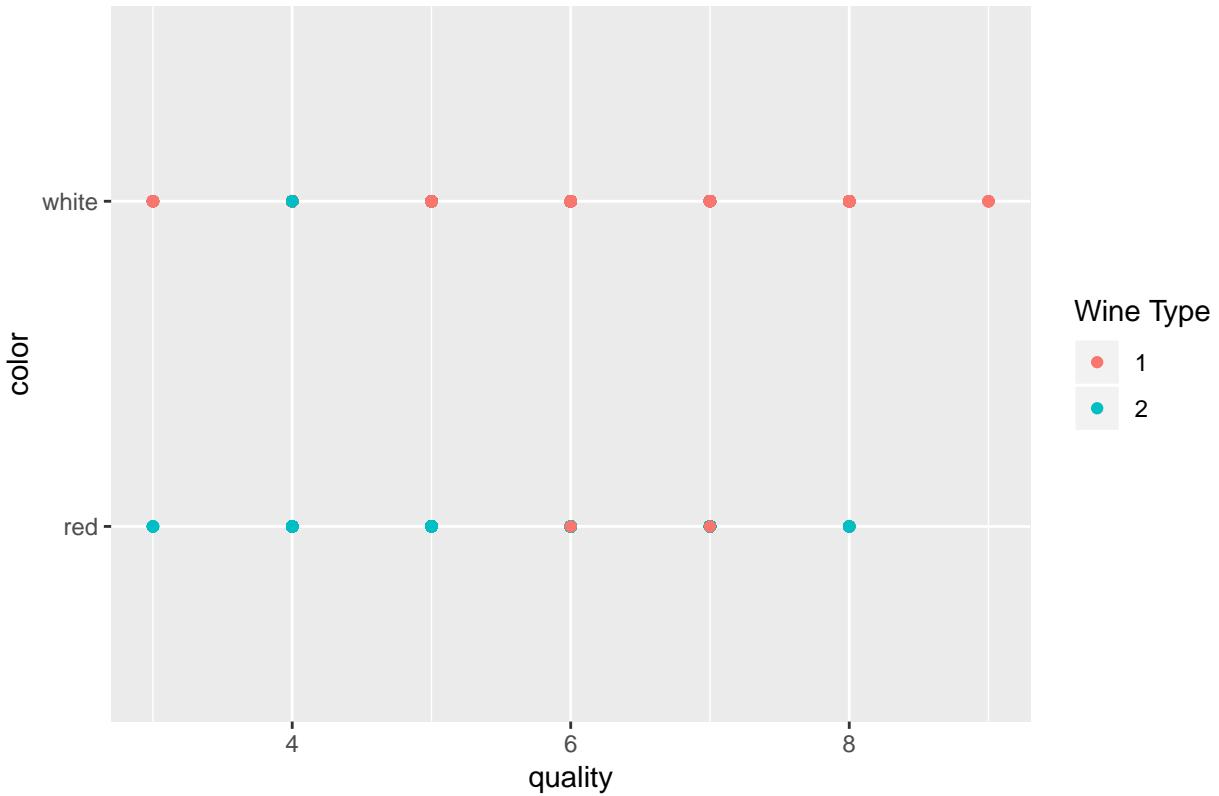
By comparing our two variable relating to acidity **Volatile Acidity** and **Fixed Acidity** we find that the clustering algorithm has captured the two main groups of wine with an intermixed amount near the middle diagonal. As we depict more comparisons, we will find that the cluster approach produces an easier to understand graph of the results.

Citric Acid and Residual Sugar by Wine Type by Wine Type



Here is an example where clustering failed to adequately determine the correct wine type based on the **Residual Sugar** and **Citric Acid** samples of the wine. With **Red** as 1 and **White** wines tagged as 2, we find that the clustering method incorrectly created a large centroid of wines near the center of the mass.

Wine Qualiy and Color by actual Wine Type

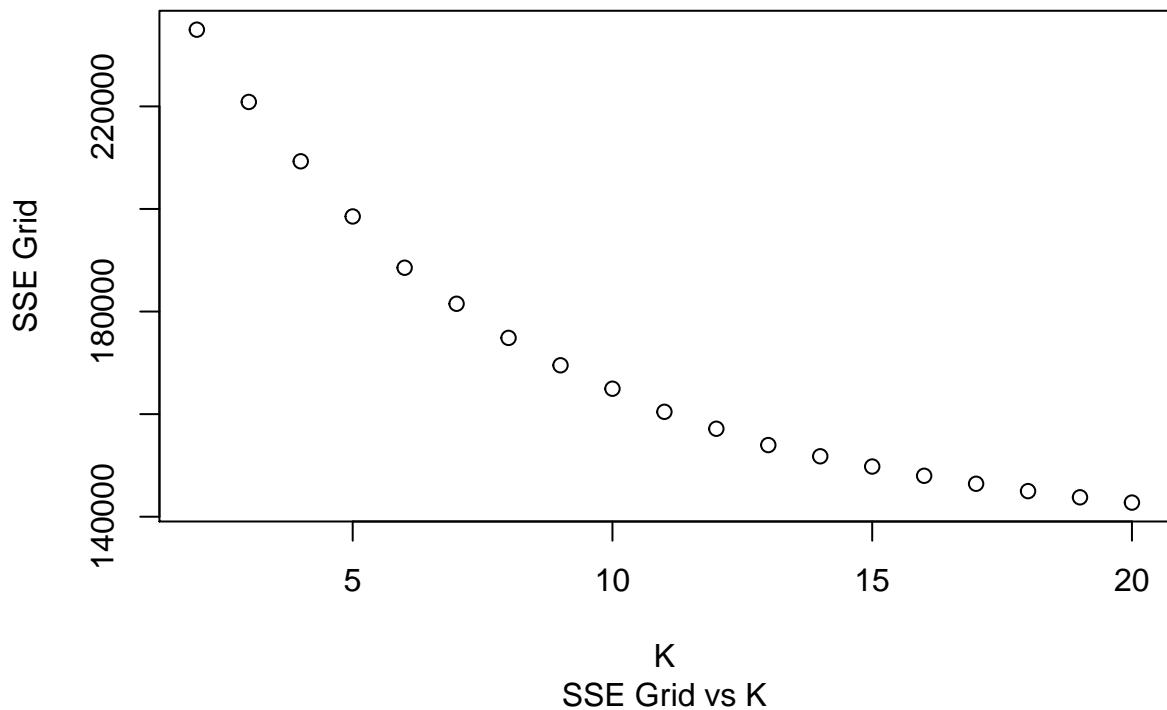


Lastly we have the final clustering of the wine **Color** and **Quality**. The clustering approach correctly labels all but one of the white wines while missing all but two of the reds. Given the fact that there are only colors of wines to work with, missing even one or two of the wine set would mean that although a cluster approach would be easier to implement, it will still miss 10% or more of the time when determining whether or not a wine is red or white.

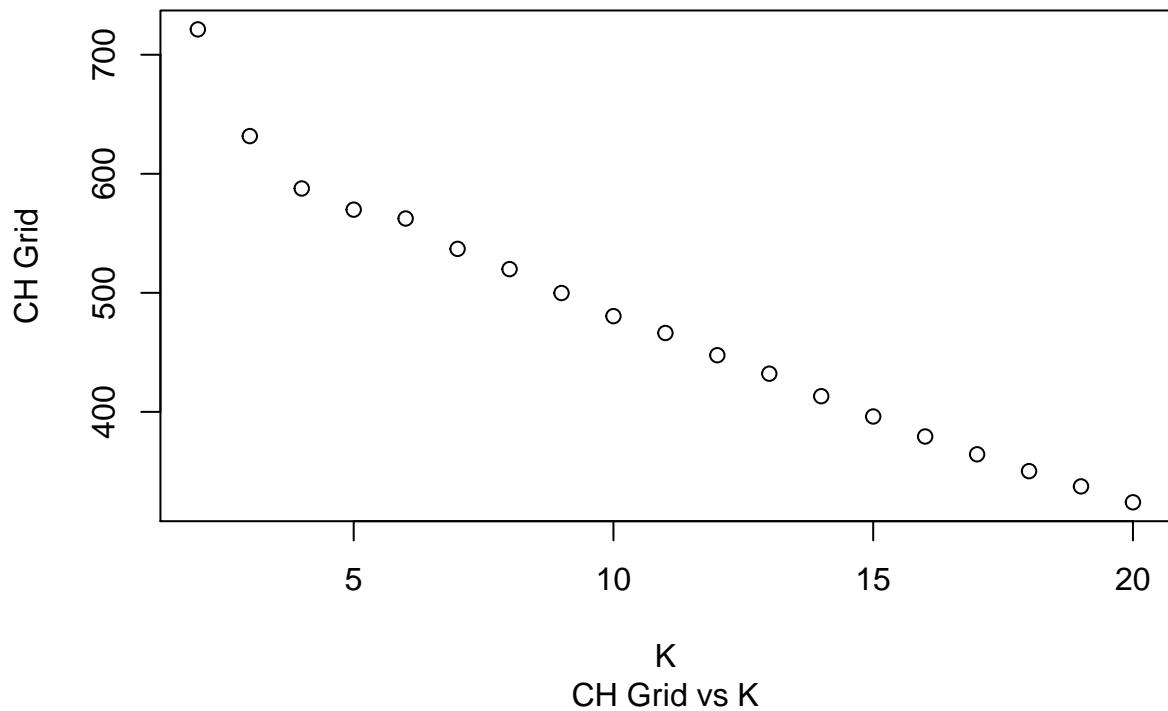
Question 2: Market Segmentation

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.

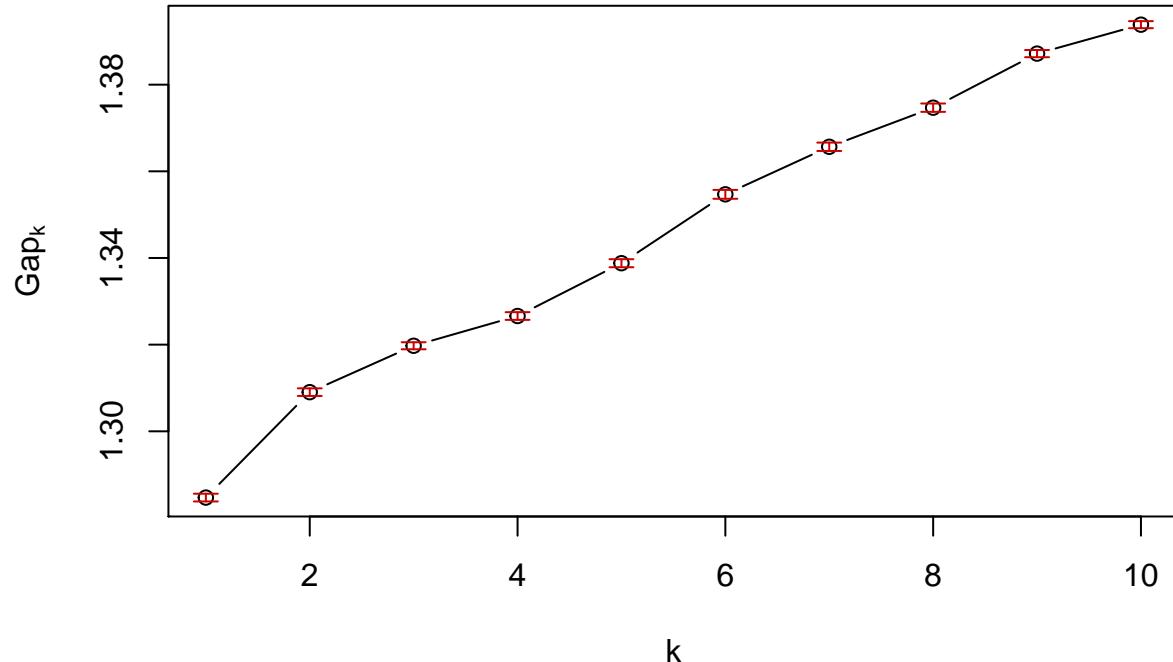


Next we use K-means clustering again, but this time we measure the CH grid values for their respective K values.



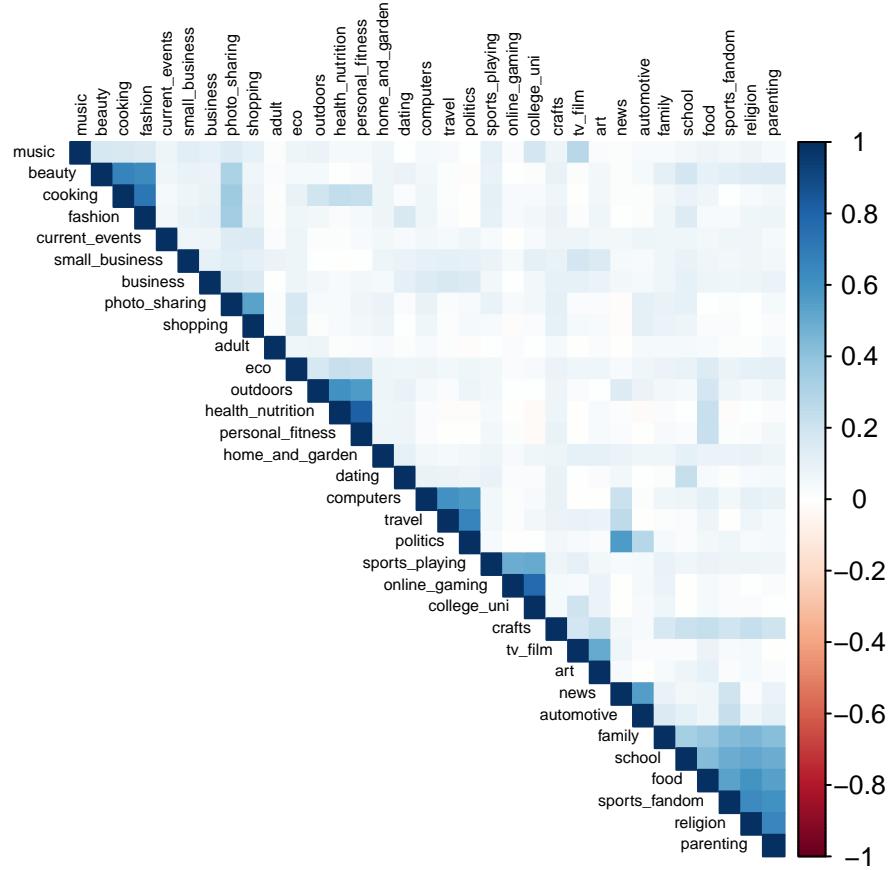
Next, we use the cluster gap function to see where the dip in cluster gap is.

```
clusGap(x = SocialData_scaled, FUNcluster = kmeans, K.max = 10, B = 10, nstart = 20)
```



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.

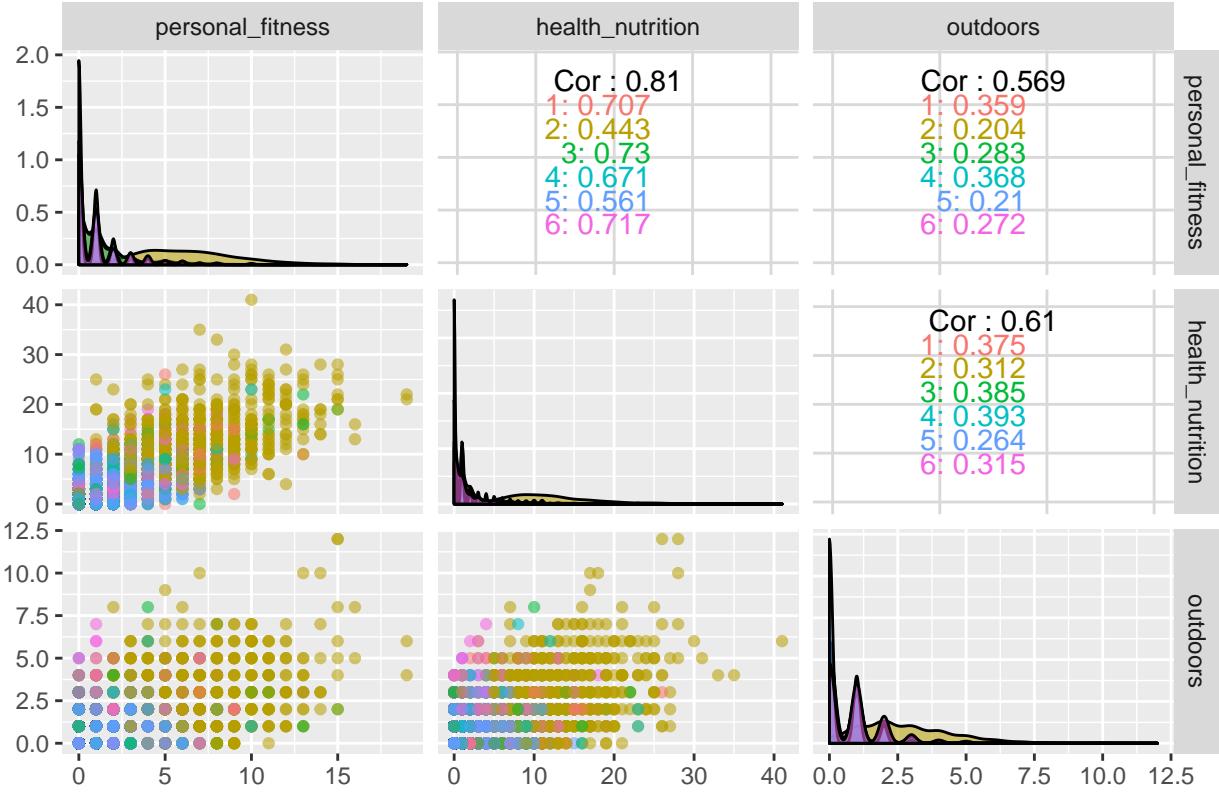
To analyse the K-means clustering, let's first take a look into the correlation between these various post indicators/factors.



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.

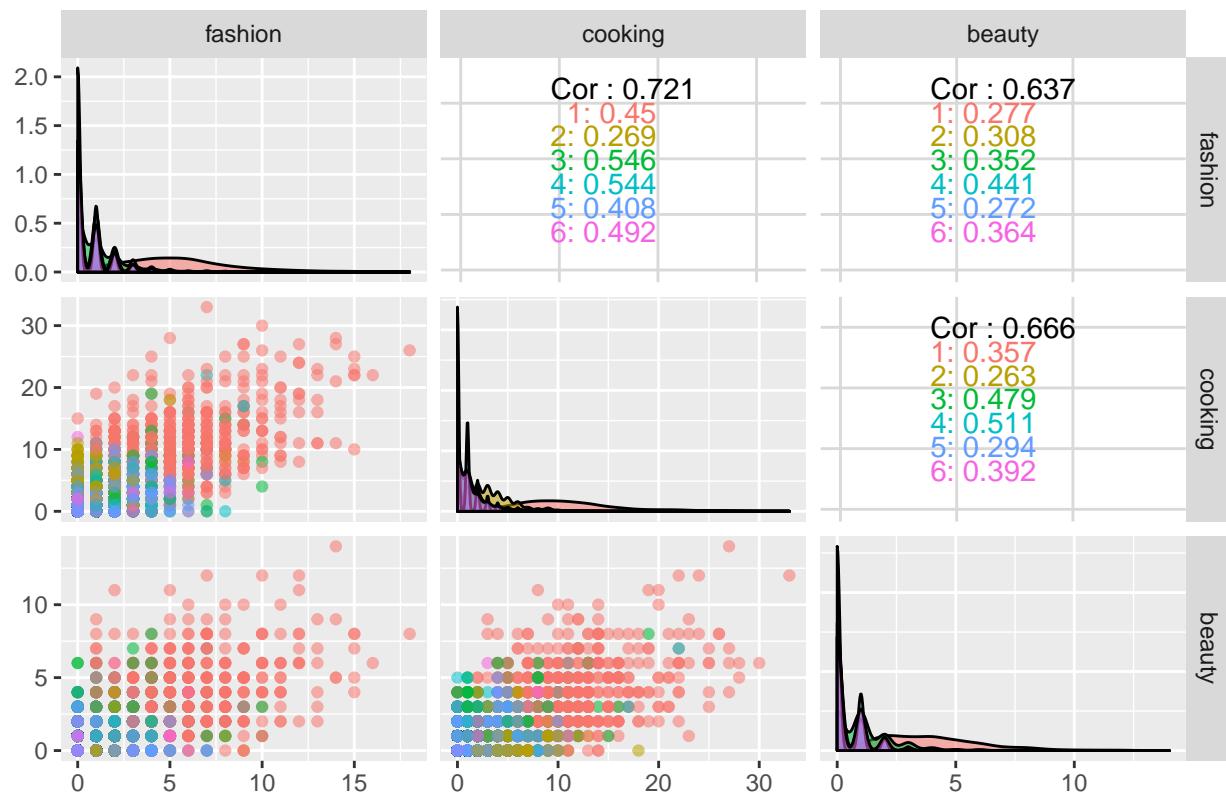
- i) Personal fitness, Health, Outdoors

Clusters on personal fitness, health and outdoors



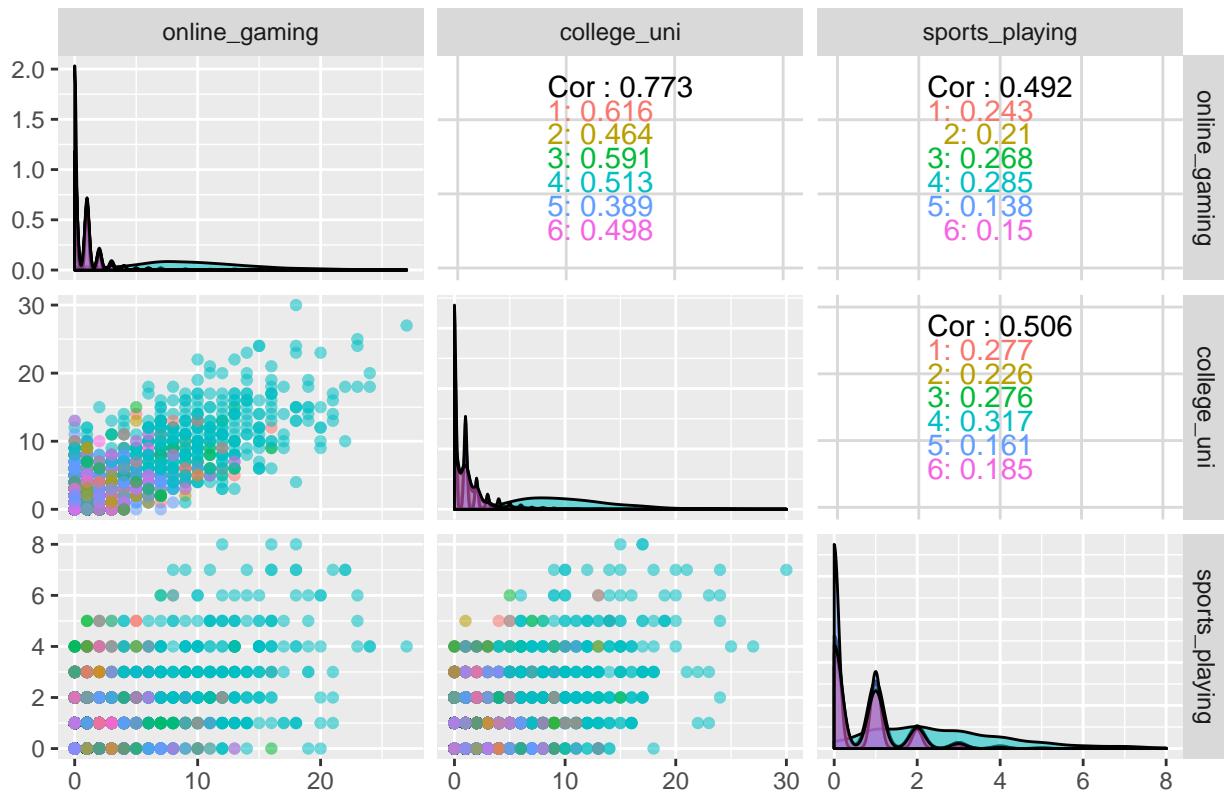
ii) Fashion, Cooking, Beauty, Shopping, Photo sharing

Clusters on fashion, cooking, beauty, shopping and photosharing



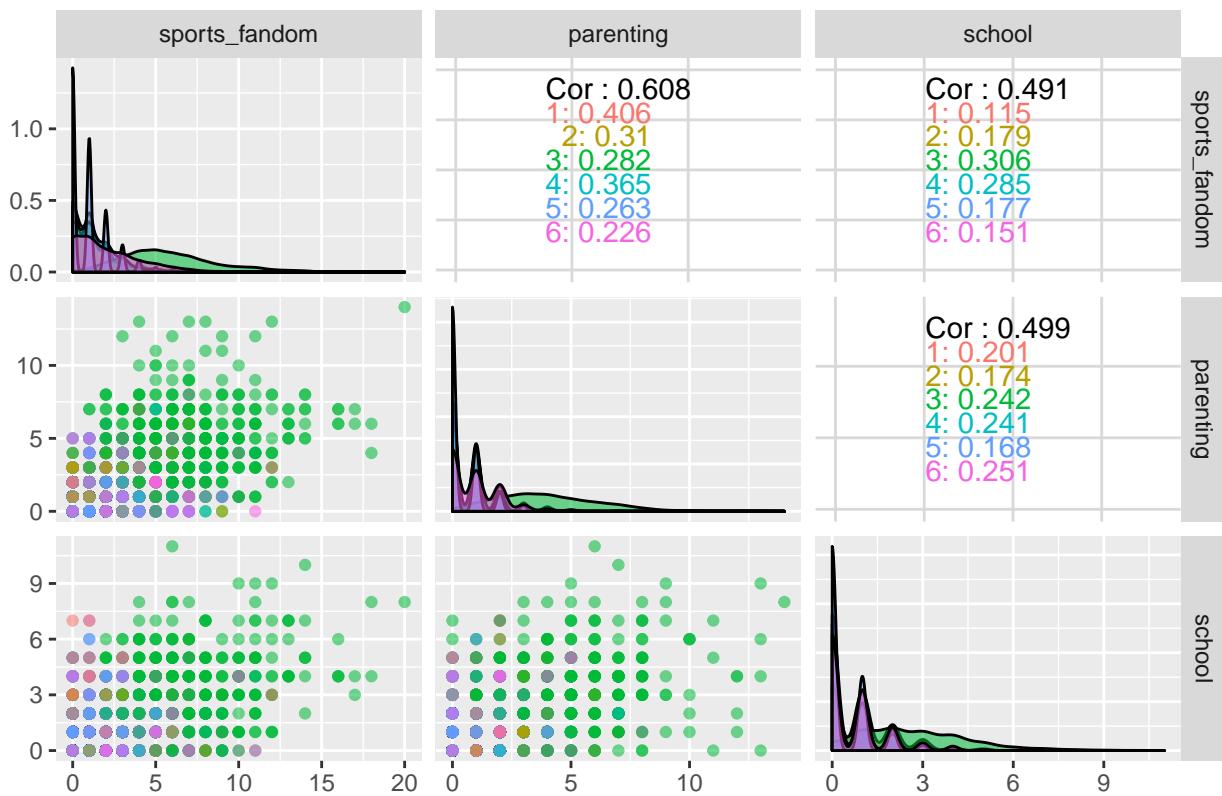
iii) Gaming, College, Sports

Clusters on gaming, university and sports



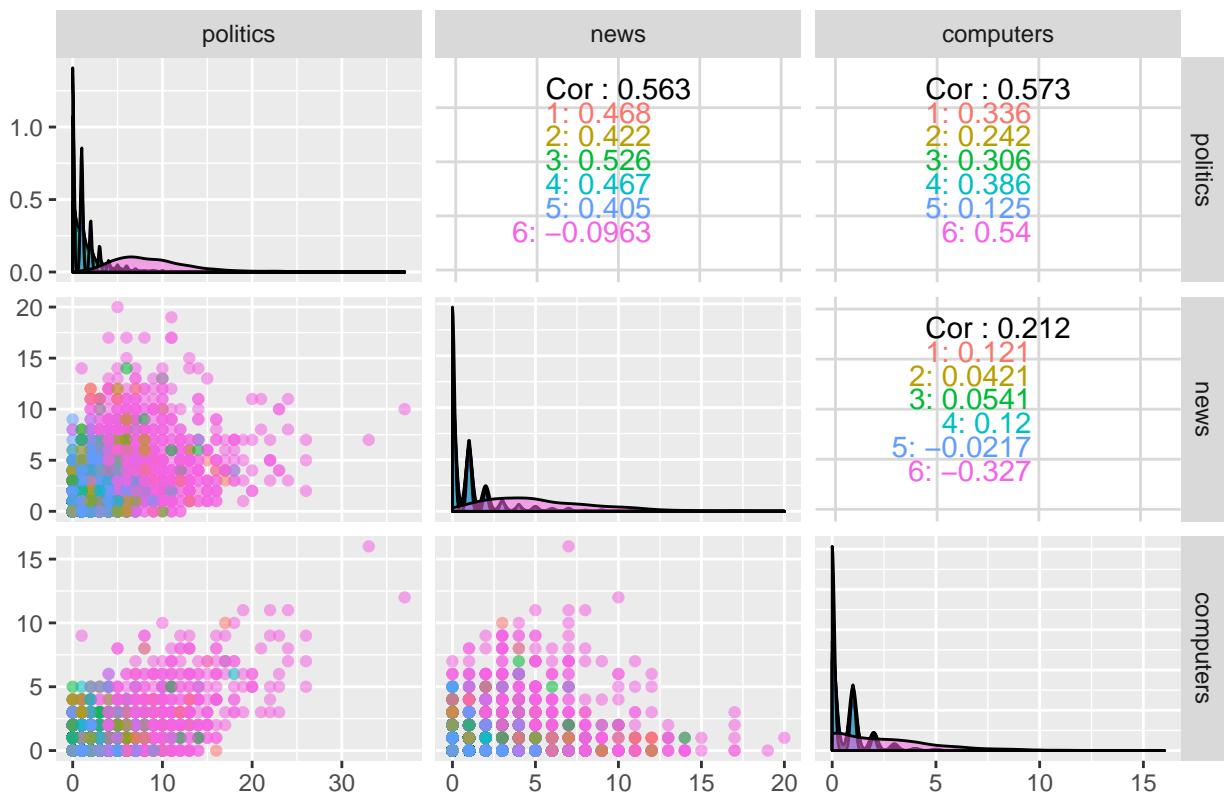
iv) Sports fan, Parenting, School, Food, Family

Clusters on sports, parenting, school, food, family



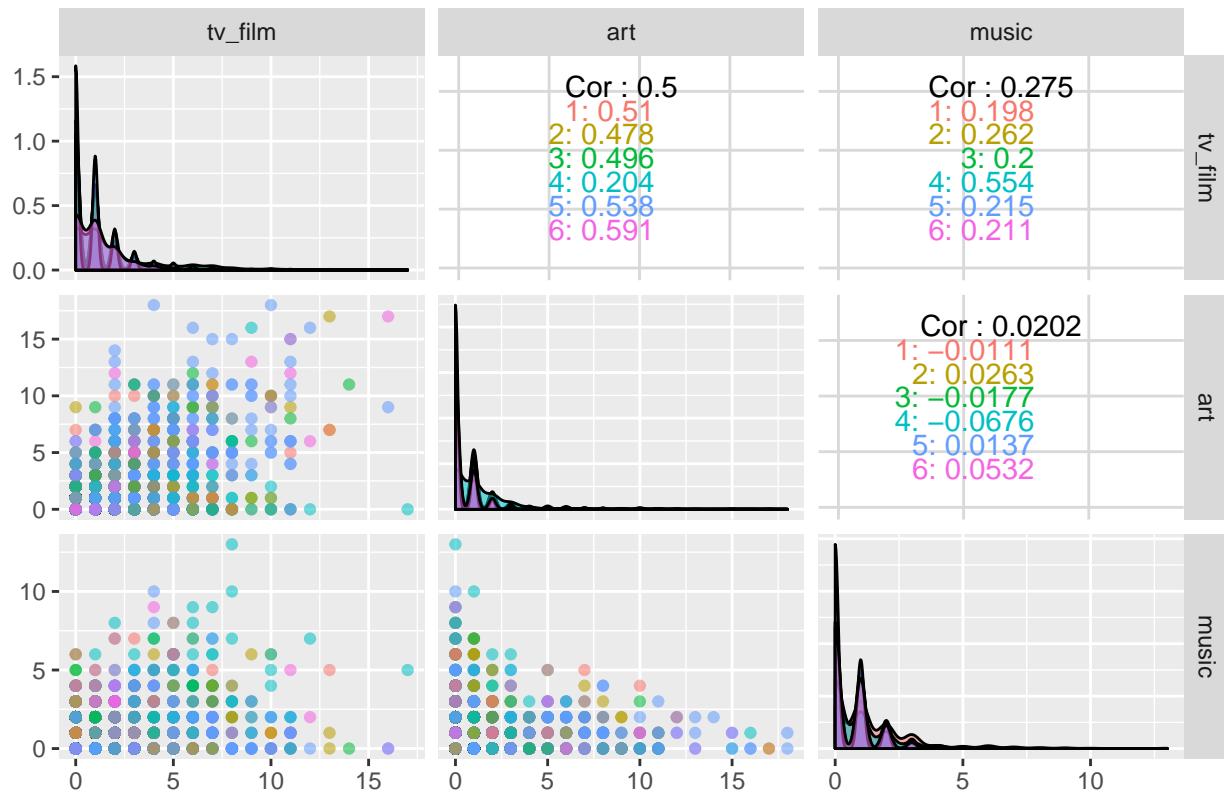
v) Politics, News, Computers, Travel, Automotive

Clusters on politics, news, computers, travel, automotive



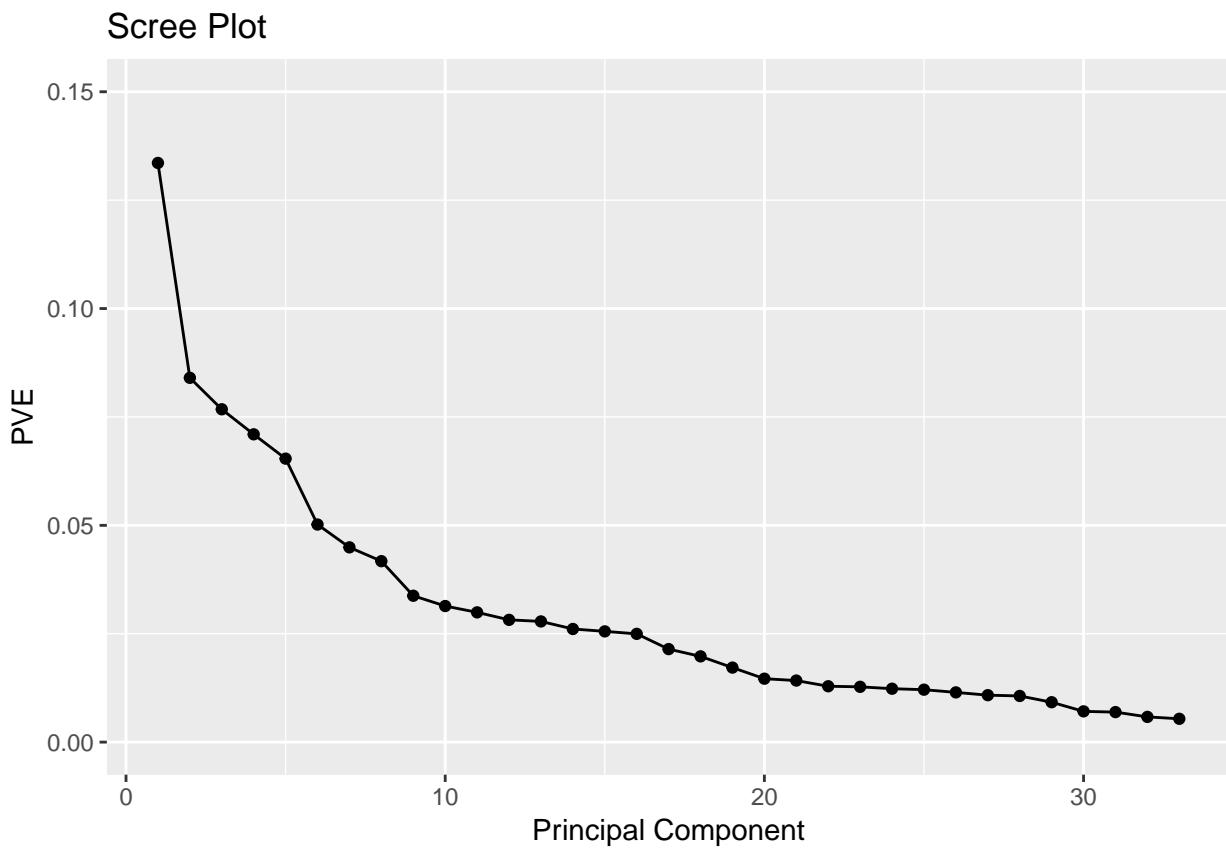
vi) TV, Art, Music

Clusters on tv, art, music

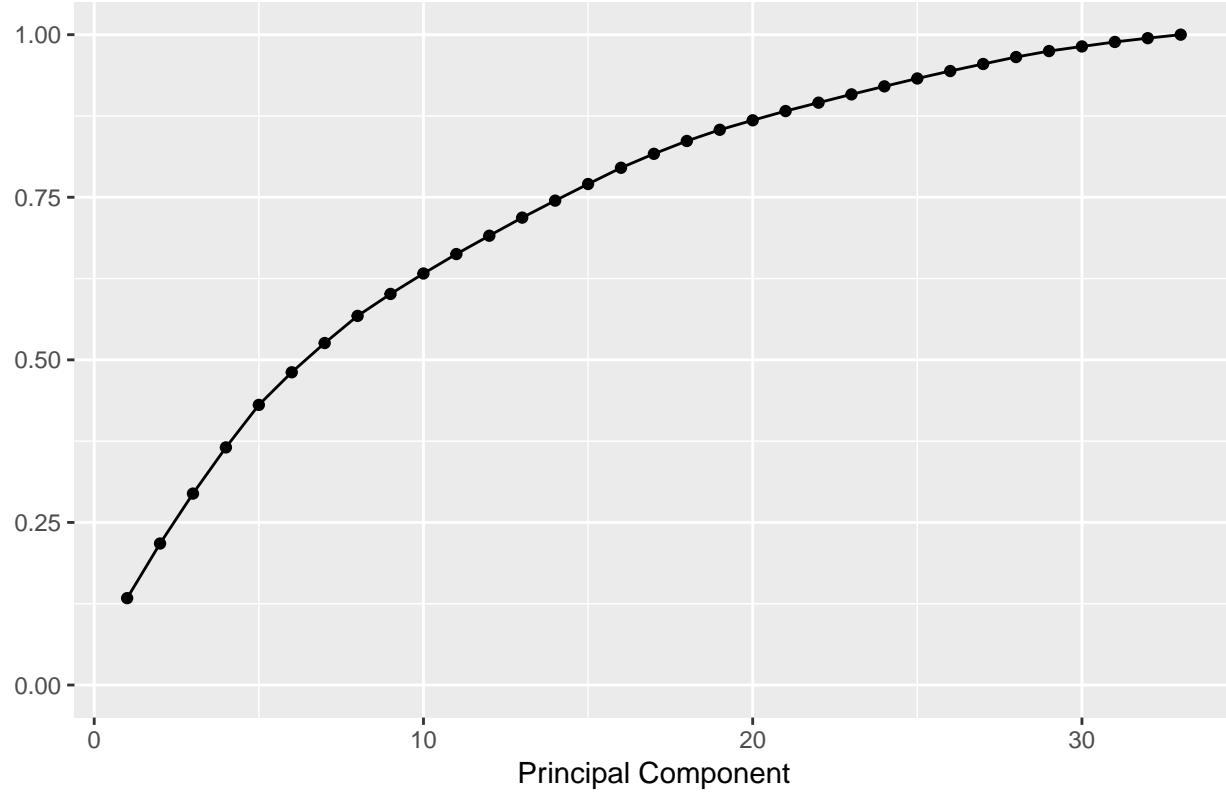


From the above graphs, we can see that there isn't much we can try to extract in terms of segments from the above data. Very few factors are highly explained by other factors. Hence, we decided to do the PCA analysis to extract them into different market segments so that we can see what are the factors 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.



Cumulative Scree Plot

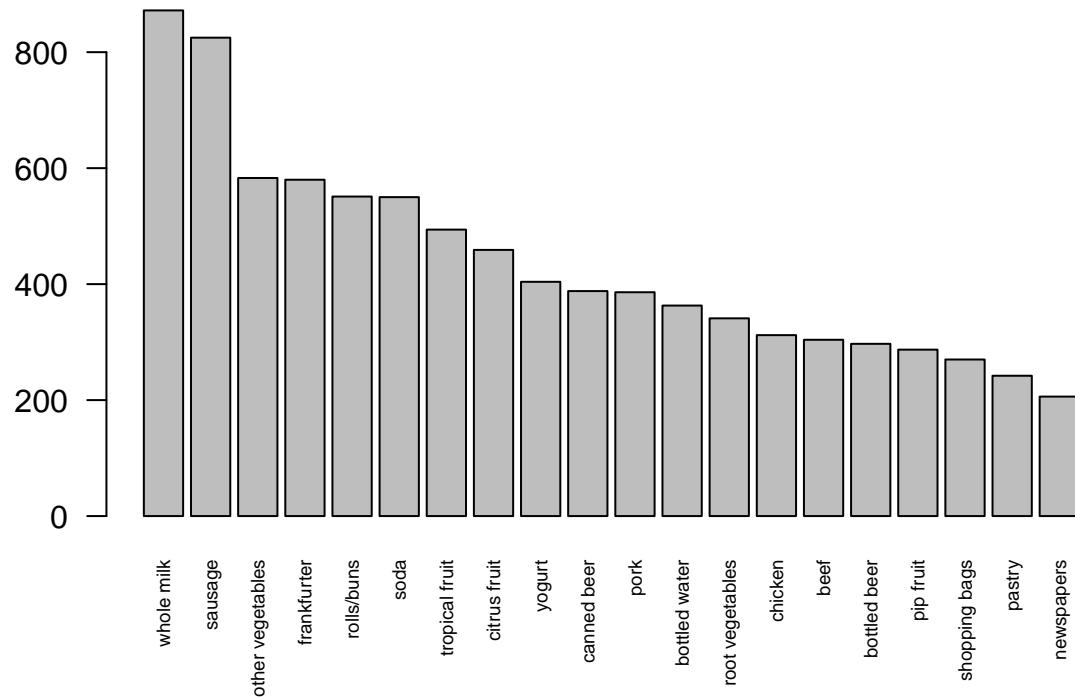


From the above graphs, we can see that almost 50% of the variance is explained by the 6 out of 33 factors. Hence, this reaffirms our initial value of K as well. We get the following market segments that NutrientH2O can use to segment its audience and target them with sharper messages.

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 could 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 be 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.

Question 3: Association Rules for Grocery Purchases

The problem of any grocery marketer is what items to stock and in what quantities. If there are certain groups of products, like peanut butter and jelly with bread, then it would make sense to stock items of group value. Market basket analysis has the purpose of quantifying items within the shelves in a store. Placing products is more of an art than a science - however it doesn't mean data analysis should not be a major component in its creation. The goal is to influence customer purchasing habits to chain-buy their groceries. With this in mind, we will delve into the data set at hand.



Here is a list of the top items within the data set. It is clear that the most popular item is milk. Being the mainstay of various food groups, milk and cereal, milk and baking, milk in .

With only 2160 rules created, this is a good start in understanding the associations of different goods within the data set. But the holy grail is to visualize the data using a network graph.

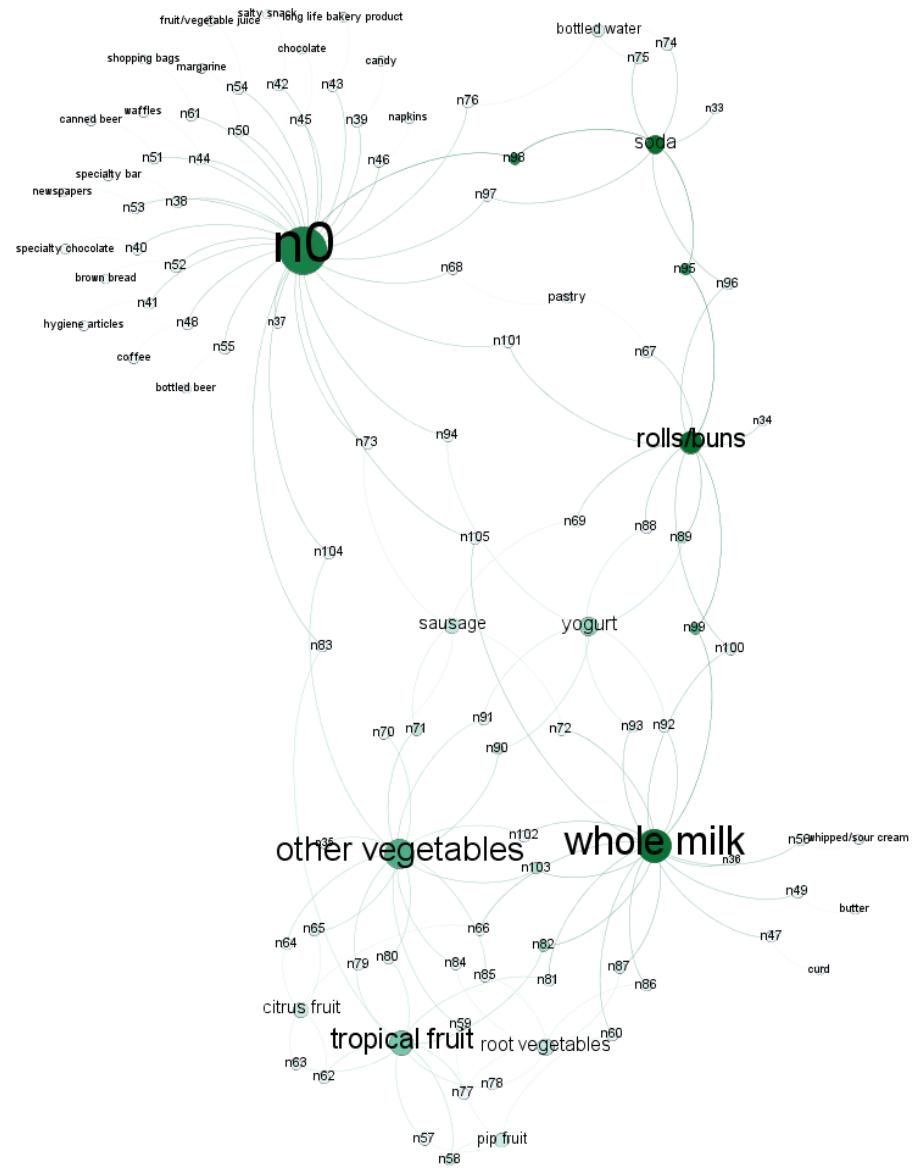


Figure 1: Market Basket Association Network