HW4

Jacob Bulzak

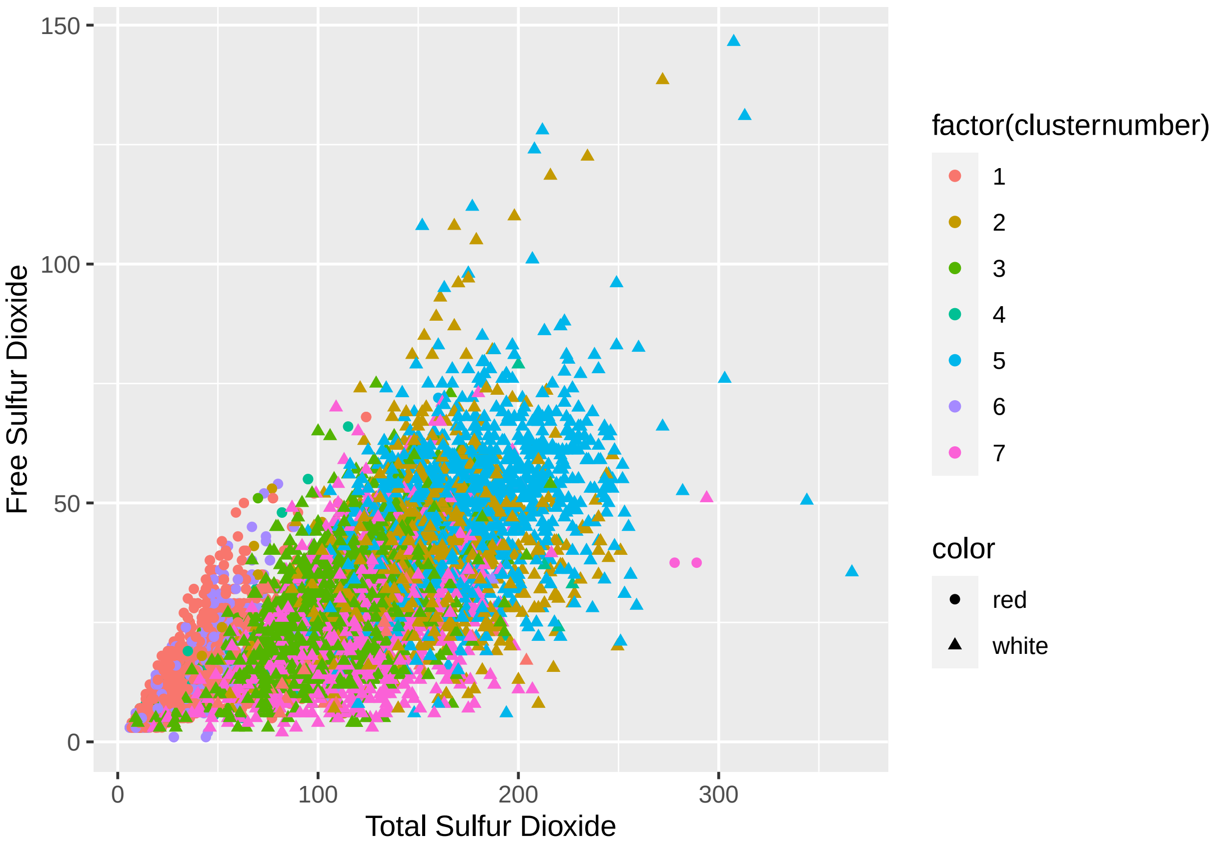
5/2/2022

Wine Clustering and PCA

I began by cleaning the data (removing outliers), followed by centering and re-scaling. After experimenting with several k-values, I decided to use k=7 which reflects the fact that there are 7 different levels of `quality` in the dataset.

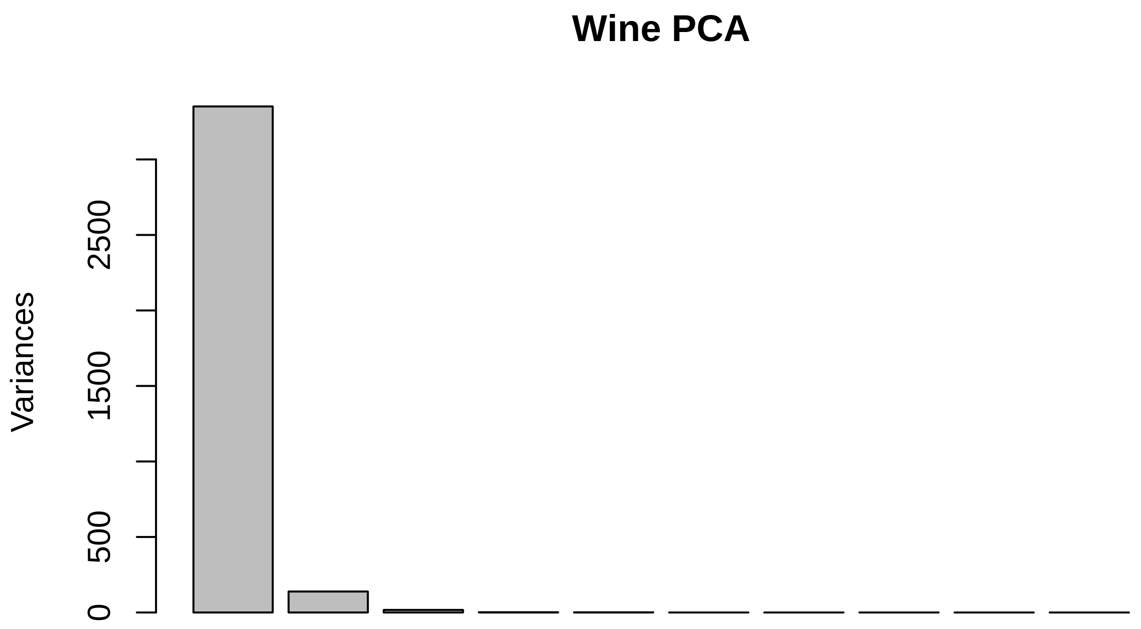
I then tried several k values and settled on 7 after comparing how well the clusters could designate red vs white wine. Next, I found that the greatest data point separation was achieved with `total.sulfur.dioxide` and `free.sulfur.dioxide` parameters, visualized below.

```r
ggplot(wine) +
  geom_point(aes(x=total.sulfur.dioxide,y=free.sulfur.dioxide,
                 color = factor(clusternumber),
                 shape = color)) + xlab("Total Sulfur Dioxide") + ylab("Free Sulfur Dioxide")
```



Next, we run the PCA for this data:

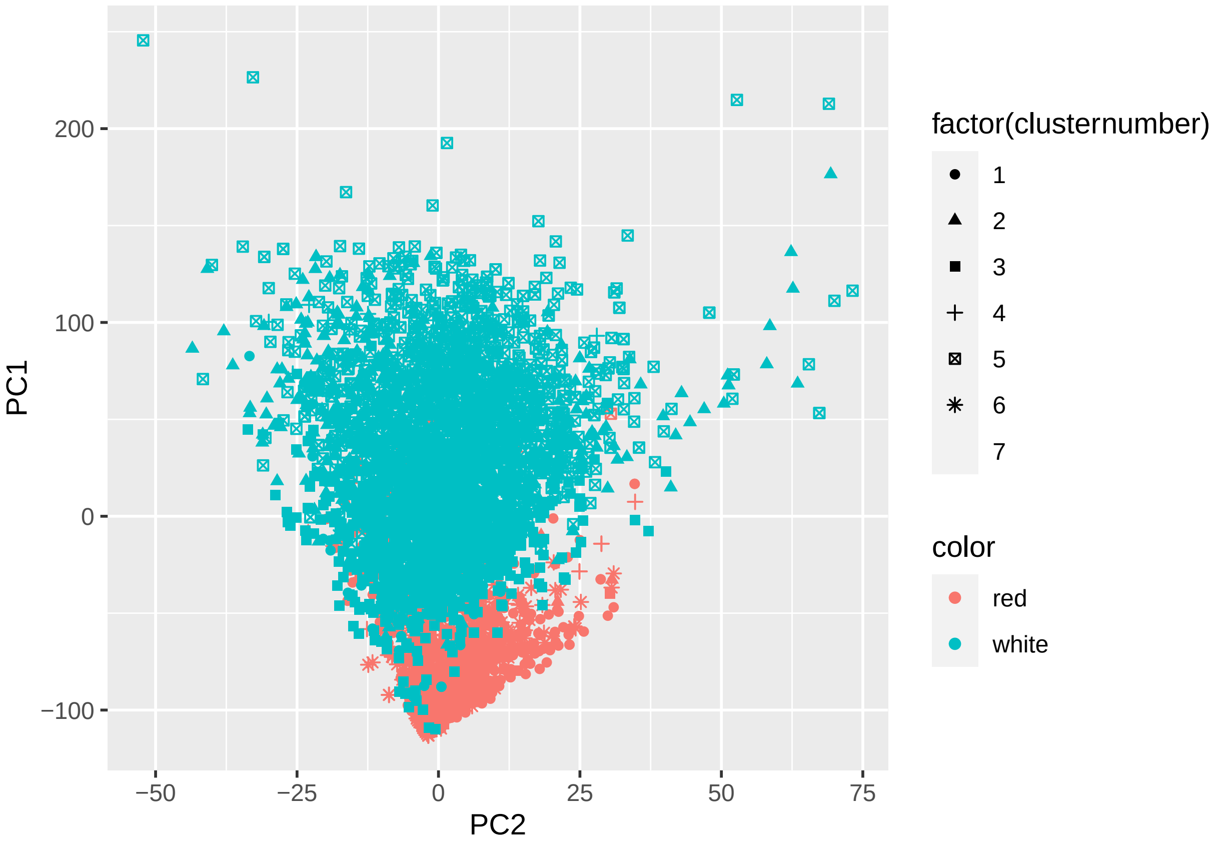
```
winevarplot = plot(wine_PCA , main = "Wine PCA")
```



As we can observe in the plot above, the PCA was able to explain a significant majority of the information with just one dimension. Despite this, I decided to instead use 2 dimensions to aid with visualization, and because there are two types of wine in the data set: red and white.

We proceed by plotting the two princinpal components:

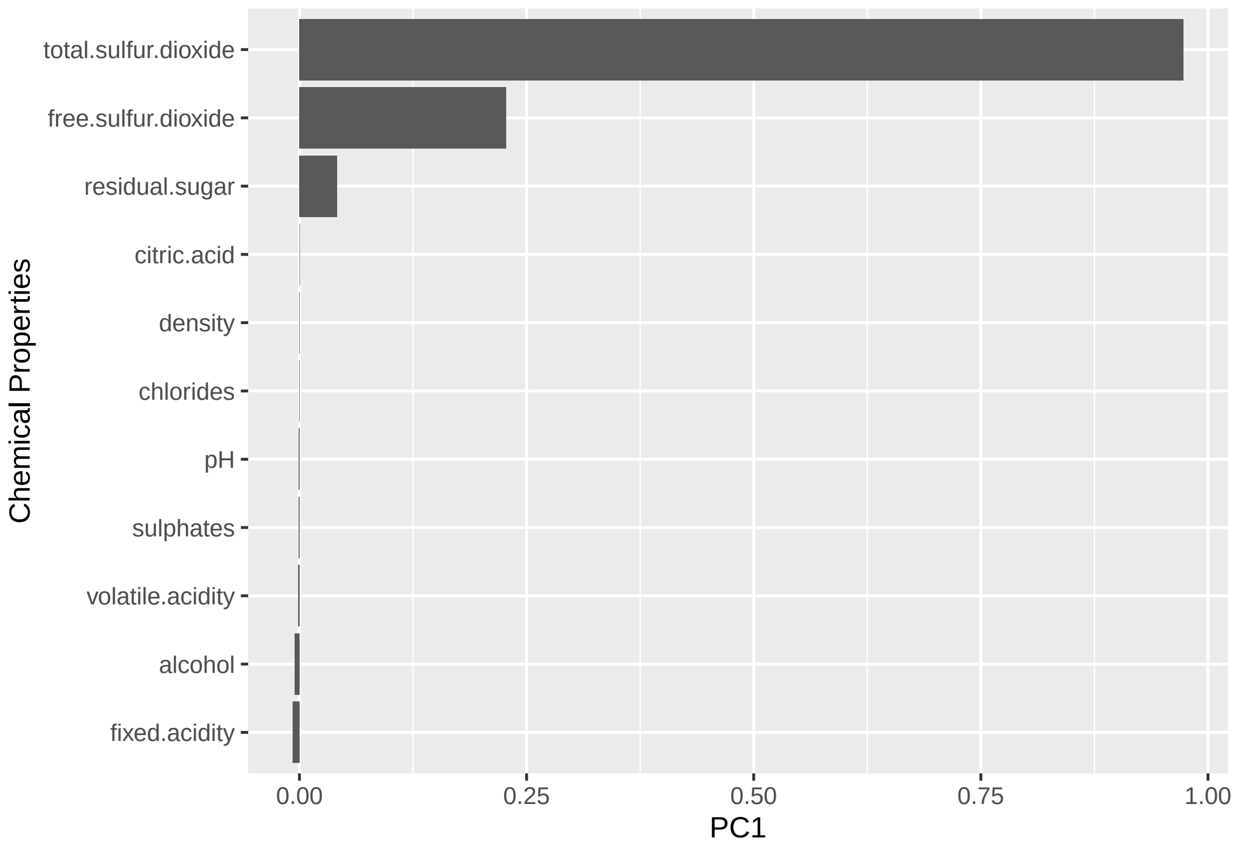
```
WineColorPlot
```



From the plot above we can observe a distinct separation in the two colors of wine. The red wines tend to have relatively high-magnitude (negative) PC1, and low positive magnitudes of PC2 (roughly between 0 and 25). The white wines on the other hand exhibit a greater dispersion in their magnitudes of PC1 and PC2.

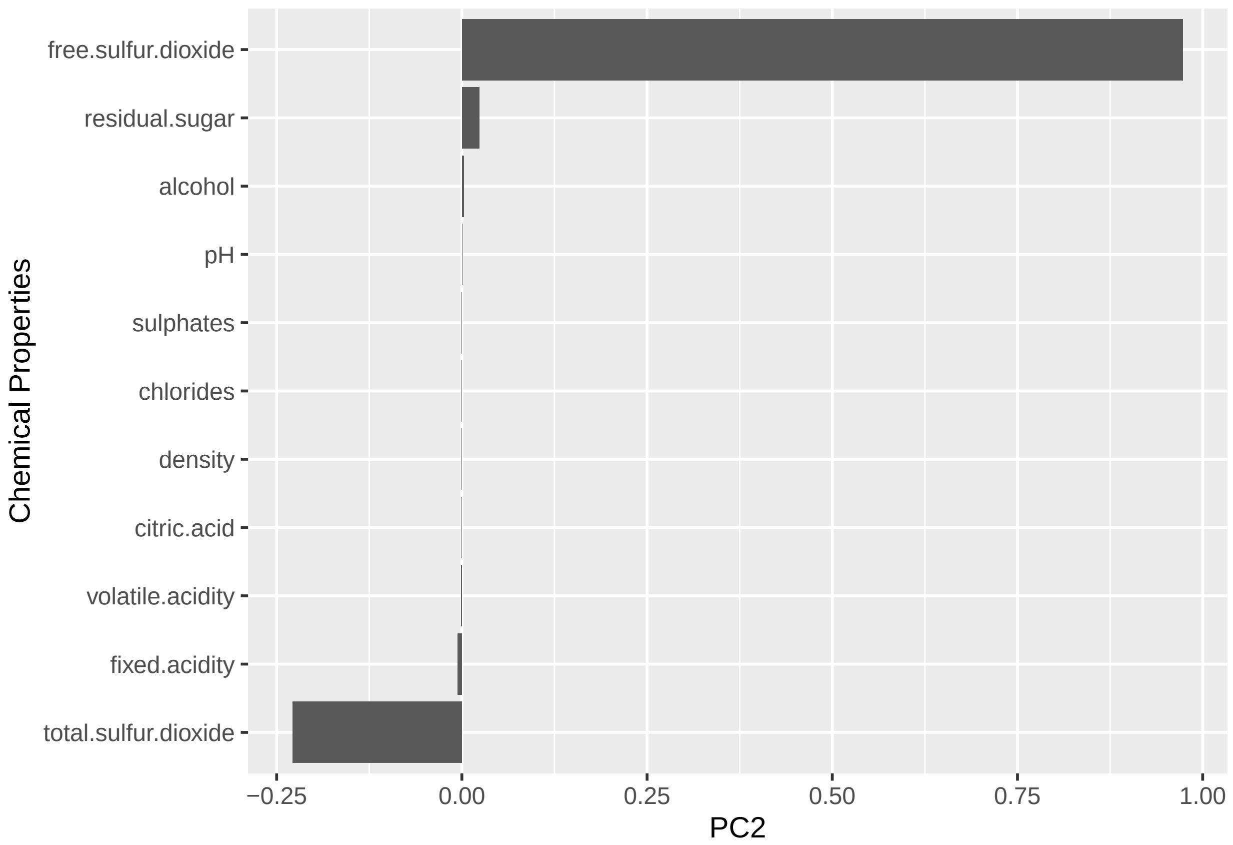
We introduce the following plots to determine which variables most significantly affect the principal components:

```
PC1barplot
```



From the figure above we see that the main feature in PC1 is `total.sulfur.dioxide`.

```
PC2barplot
```



We further observe that PC2 is primarily defined by `free.sulfur.dioxide`. We also see a notable lack of `total.sulfur.dioxide` which, as aforementioned, was the defining feature of PC1.

Finally, I introduce the table below to provide a summary of my findings.

```
final_wine_table
```

Table 1: Wine Color and Quality Table

clusternumber	color	Count	PC1	PC2	AvgQuality
1	red	898	-69.754206	1.7056150	5.433185
1	white	51	-19.580171	-8.5305425	4.725490
2	red	23	-17.521564	5.2985203	5.260870
2	white	970	33.260936	0.1273498	5.938144
3	red	32	-58.731518	3.1655819	6.500000
3	white	1168	-7.791770	-0.0294644	6.430651
4	red	45	-53.848661	1.5164399	5.422222
4	white	10	64.324142	-8.9407937	4.800000
5	red	2	14.245392	14.4058073	6.000000
5	white	1464	59.042969	3.4381232	5.622951
6	red	590	-78.330876	1.2096456	5.923729
6	white	14	16.507692	-11.1027121	4.928571
7	red	9	35.802692	-14.3143548	5.888889
7	white	1220	3.059332	-5.6388062	5.677049

Upon examining the table, several features become readily apparent. First, the clusters are quite apt at determining the color of the wines. In each cluster, one color of wine tends to dominate the other in quantity. For example, Cluster 7 contains 1220 whites and only 9 reds, whereas Cluster 1 is decidedly red, containing 898 reds and 51 whites.

Furthermore, by looking at the PC1 column we see that the reds predominantly have large negative or small positive values for PC1. Whites, on the other hand, display a somewhat inverse relationship with large positive values or small negative values for PC1. These patterns are indicative of the main differentiating factor between reds and whites being `total.sulfur.dioxide`.

Finally, an interesting result from the table is that the the clusters do not do a good job of clearly distinguishing across wine qualities. The average quality does not exhibit much variance across the colors and clusters, and indeed, most of the scores fall between 4 and 6. This seems to indicate that there does not exist a clear relationship between a wine's chemical content and its quality. A possible reason for this is that while wine sommeliers (snobs) likely have refined palates, it seems improbable that they would be able to distinguish variations in a wine's chemical profile. Instead, they likely judge wines on more subjective factors such as flavor, aroma, etc. which are emergent properties of a wine's chemical composition. A more prosaic explanation could be that whatever factors differentiate say, a quality-4 wine from a quality-5 wine are minute enough to "wash out" over the course of many wine tastings.

Market Segmentation

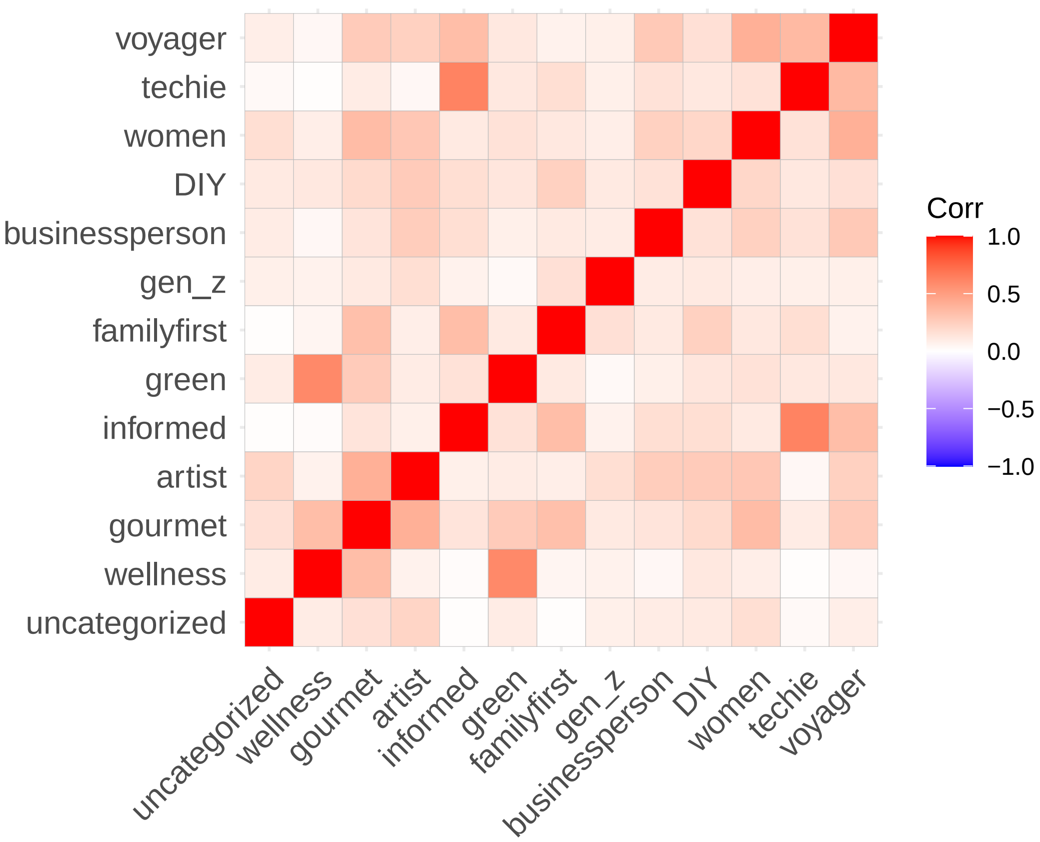
I began by cleaning the data and performing some manual compression before clustering to make analysis more straightforward down the road. Recall that the original data was grouped into 36 different categories such as `chatter`, `travel`, `politics` etc. Some of these categories were related and, it seemed most logical to group certain categories and map them onto a class of representative consumer who would likely be interested in a given set of topics. Thus sets of topics were aggregated into new columns and labeled to reflect the kind of consumer that would likely be interested in them. Some examples of the columns created are:

"artist": `art+music+fashion+tv_film` "gen_z": `online_gaming+college_uni+school` "DIY": `home_and_garden+crafts`

. . . and so on. I also decided to remove the `adult` and `spam` columns first, to get rid of any unsavory elements such as spam and pornography bots that slipped through the initial filter, and second because these categories did not seem very useful in helping NutrientH20 analyze market segments.

Finally, I also decided to drop the `chatter` column. This was done for two reasons: 1) To further combat any spam/bots and 2) To ensure the tweets in the data set were more definitively aligned with a category rather than belonging to "background noise" that could obscure results

In summary, I was left with 14 columns which I used to create a correlation matrix to help visualize any notable correlations between the generated "consumer types"



As can be seen in the plot above, two of the strongest correlations appear to be between the `wellness` and `green` categories, as well as between the `techie` and `informed`. We also note other, weaker, albeit intuitive correlations between types like `artist` and `gourmet`, or `familyfirst` and `informed`. These pairings are rather unsurprising, indicating that grouping the categories into "consumer types" is a reasonable way to approach the analysis of market segments.

Next, I scaled and centered the data in preparation for clustering. After some experimentation, I decided to go with 4 clusters in order to present only the most interesting groupings and keep the analysis straightforward.

By examining the clusters more closely we can obtain some surface-level information about H20Nutrients customers:

```
##                  clust1$center[1, ] * sigma + mu
## uncategorized                                0.6
## wellness                                     2.4
## gourmet                                      1.8
## artist                                       2.2
## informed                                     4.1
## green                                        0.8
## familyfirst                                  1.9
## gen_z                                        2.6
## businessperson                               0.5
## DIY                                          0.7
## women                                        1.8
## techie                                       0.9
## voyager                                      3.0
```

Cluster 1 exhibits the most tweets in the `wellness`, `gen_z` and `voyager` categories. This could be interpreted as a cluster of young, health-conscious users interested in travel.

```
##                clust1$center[2, ] * sigma + mu
## uncategorized                              0.7
## wellness                                   3.4
## gourmet                                    3.6
## artist                                     2.8
## informed                                  16.2
## green                                      1.5
## familyfirst                                5.7
## gen_z                                      3.8
## businessperson                             1.1
## DIY                                        1.3
## women                                      3.0
## techie                                     4.5
## voyager                                    7.0
```

In Cluster 2 the category `informed` stands out significantly, as well as `voyager` which makes sense as travel aficionados want to be up to date on current events. It is probable that these customers are more versed in world issues than the average user, and might be more socially conscious, thus placing more weight on NutrientH20's stance on global and social issues.

```
##                clust1$center[3, ] * sigma + mu
## uncategorized                              0.9
## wellness                                  19.5
## gourmet                                    6.0
## artist                                     3.3
## informed                                   5.2
## green                                      4.0
## familyfirst                                2.8
## gen_z                                      3.1
## businessperson                             0.7
## DIY                                        1.3
## women                                      2.9
## techie                                     1.2
## voyager                                    4.0
```

The category `wellness` stands out significantly in Cluster 3, as well as `gourmet` to a lesser extent. These health-conscious users likely place great importance on the quality and nutritional benefits of their food, hinting that NutrientH20 may want to address issues such as artificial colorings, natural ingredients, and sugar content when marketing their drinks.

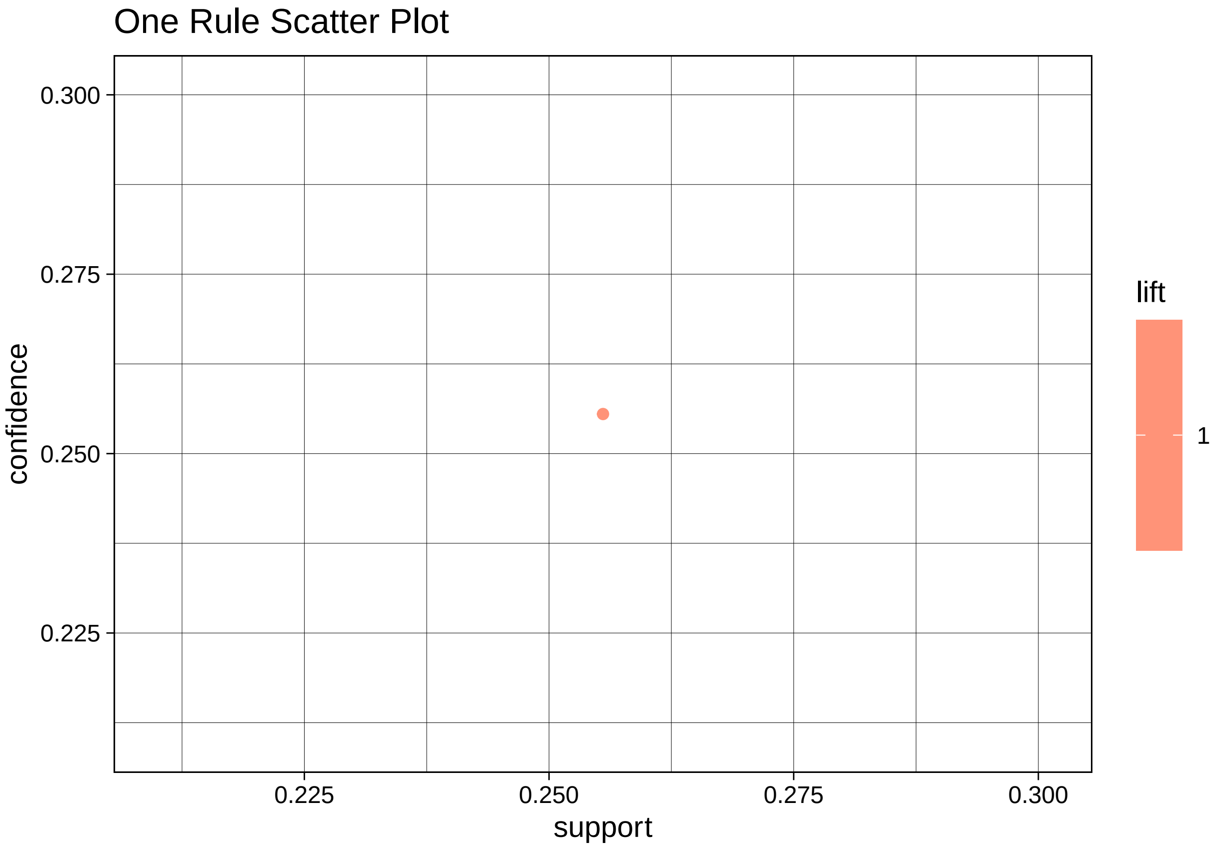
```
##                clust1$center[4, ] * sigma + mu
## uncategorized                              1.4
## wellness                                   4.1
## gourmet                                    7.0
## artist                                     8.3
## informed                                   6.0
## green                                      1.3
## familyfirst                                4.1
## gen_z                                      6.5
## businessperson                             1.4
## DIY                                        1.7
## women                                    5.9
## techie                                   1.4
## voyager                                  6.6
```

Cluster 4 has a high amount of the categories `artist`, `gourmet`, and `gen_z` among others. This can be conceptualized as a group of young, artistically-inclined customers who may be best marketed to by "cool" advertisements or packaging i.e. well-designed, aesthetically unique, and possibly even inspired by current artistic trends.

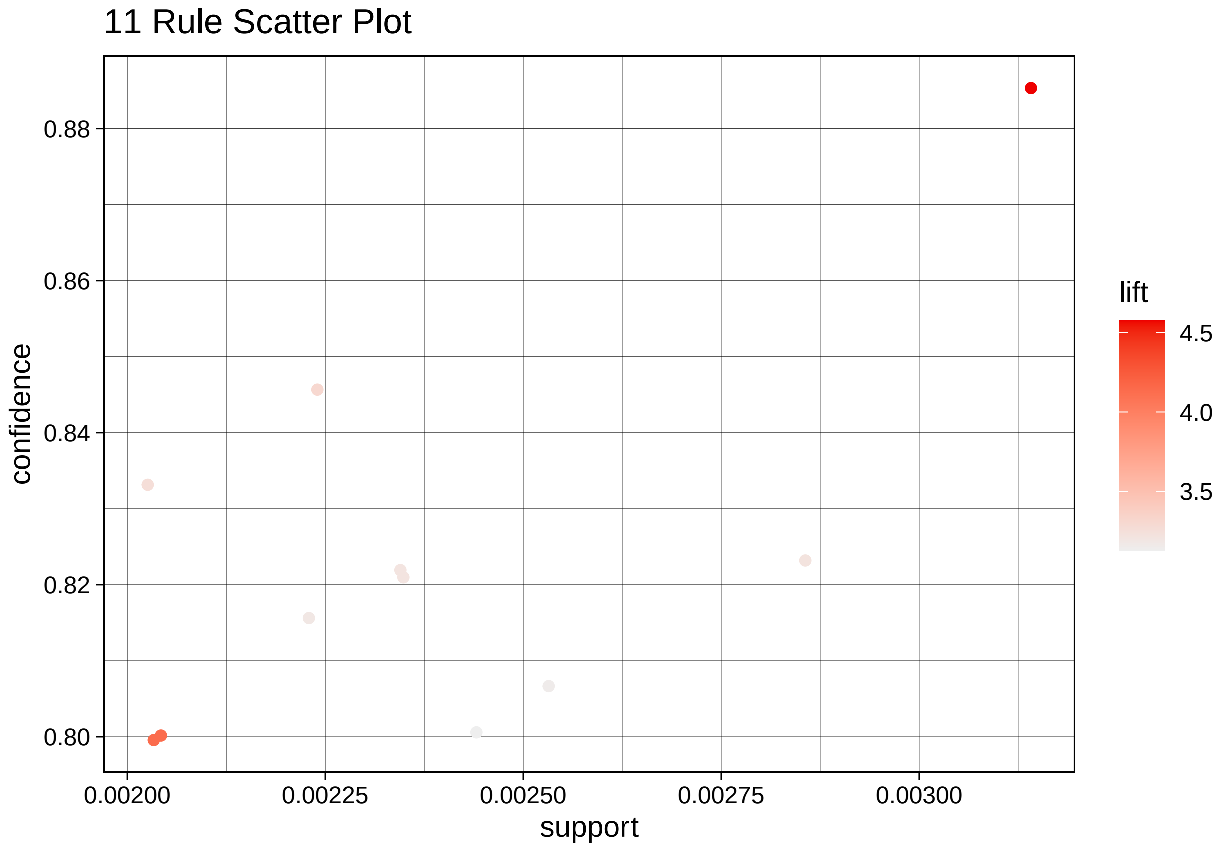
In conclusion, such market segment analysis are likely limited in their utility. While the segments provide some information about the characteristics of NutrientH20's customer base, they do not give the full story. At best, the company can make some inference about how to optimally market to their consumers. It should be noted however, that across clusters, categories such as health and wellness, current issues and Gen Z are rather prominent. This seems to indicate that NutrientH20's customers are young, socially-conscious people who place great emphasis on their health. Ergo, NutrientH20 could investigate shifting its marketing and advertisement to portary itself as a modern, socially aware company whose products promote wellness.

Groceries

I began by computing the associated rules, and distilled them into subsets informed byconfidence, lift and support. Specifically, I chose a confidence of 80%, a support of 2%, and a minimum lift of 1. I determined that these parameters were best suited for realizing the relationships between various baskets of groceries. The low level of support was chosen to reflect the fact that the variety of grocery items present in the dataset was not particularly high.The results are presented below:

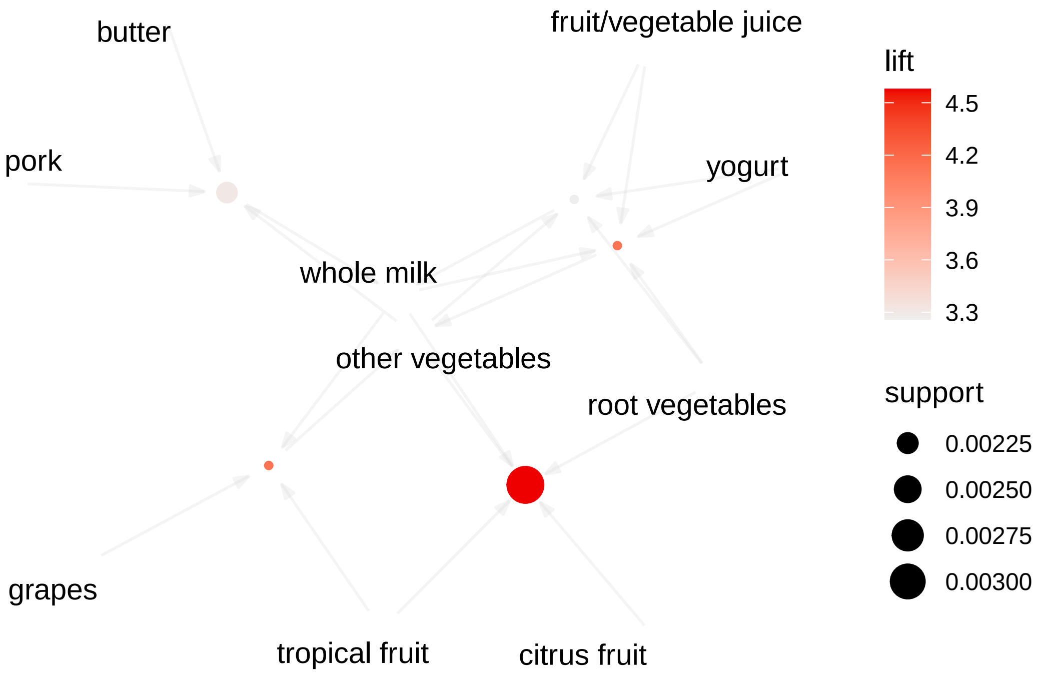


```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```



11 Rule Scatter Plot

Network Graph



The network graph above provides a convenient and intuitive way to visualuze associations between the products. We note that larger labels correspond to a higher frequency of transactions. First, a strong association between “whole milk” and “other vegetables” can be observed. A possible reason for this is that these goods are key to the average diet and are likely purchased together when grocery shopping. The relationship between “other vegetables”, “root vegetables” and fruits is also unsurprising. These goods are usually located in the same general area in most grocery stores. Thus, a customer shopping for say, tomatoes, may see zucchini or oranges on sale nearby and add them to their basket. It seems plausible that having some fruit/vegetable on your shopping list increases the probability of purchasing another fruit/vegetable. The relationship between the aforementioned goods exhibits a very high lift which hints at various vegetables being complementary goods. Indeed, it is somewhat unusual to consume one vegetable at a time. Instead, people tend to purchase many different kinds and use them as ingredients in stews, soups, or salads, hence the complementarity. Finally, the association between whole milk and butter is also unsurprising given that diary products are usually found next to each other in the same refrigerator in most grocery stores.

In summary, the rules I generated all hinged on the same basic intuition that the purchase of one good drives the purchase of another related good, given that similar foods are grouped together in stores. Finally, we note that lift might be the most important metrics vis-a-vis the aforementioned intuition as is reflects the conditional probability of purchasing a good from one group, given that a good from another group has already been bought. Ergo, unlike confidence and support, lift accounts for statistical.