Data Mining - Exercise 4

Chen & Yangxi & Shuyan

Clustering and PCA

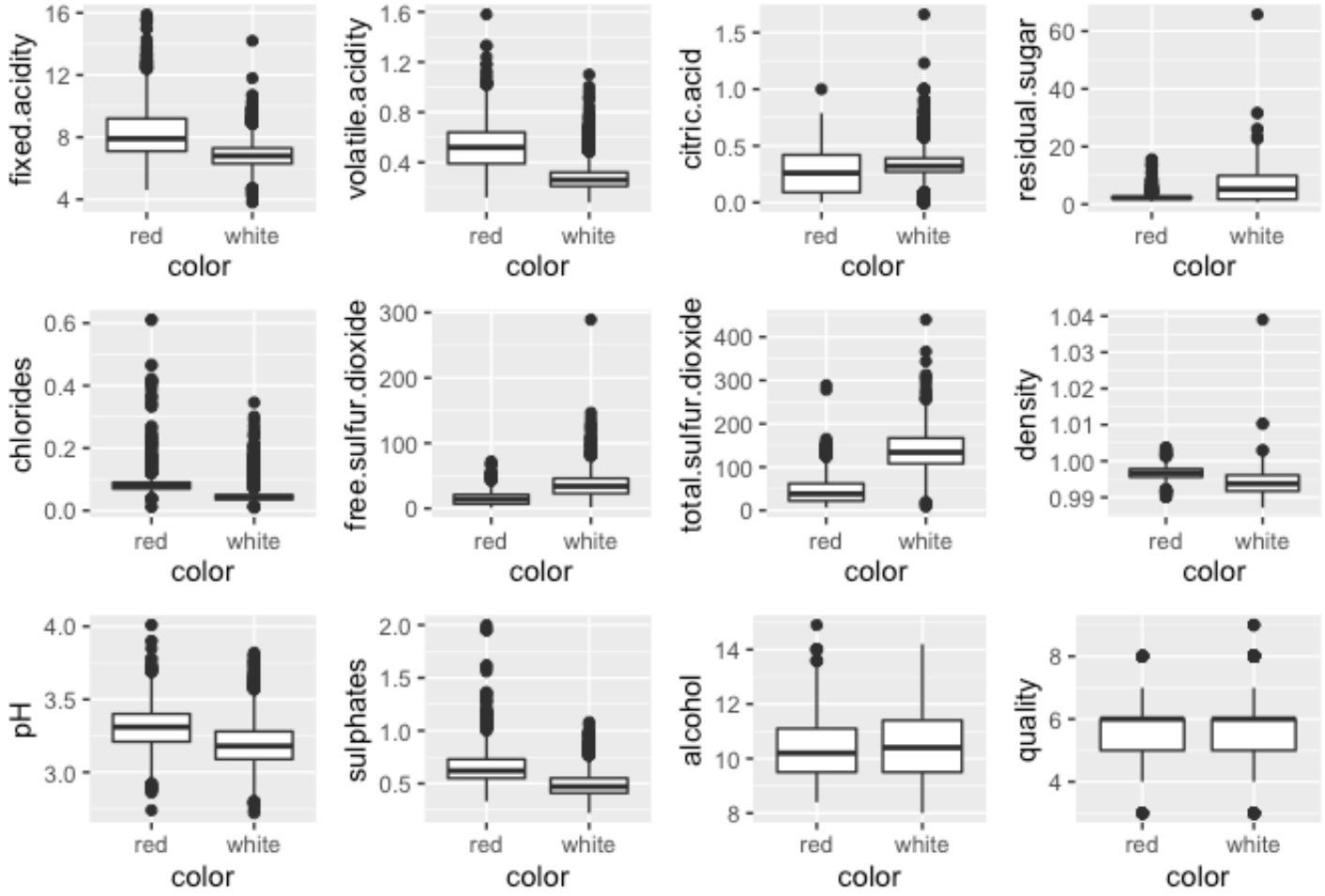
Overview

The goal of this problem is to figure out whether unsupervised learning methods such as clustering and PCA can different the red wine and the white wine, also distinguish the higher from the lower quality wine.

Red Vs White Wine

Firstly, we want to see which chemical properties have significant distinction between white and red wine.

Figure 1.1 chemical properties among white and red wines

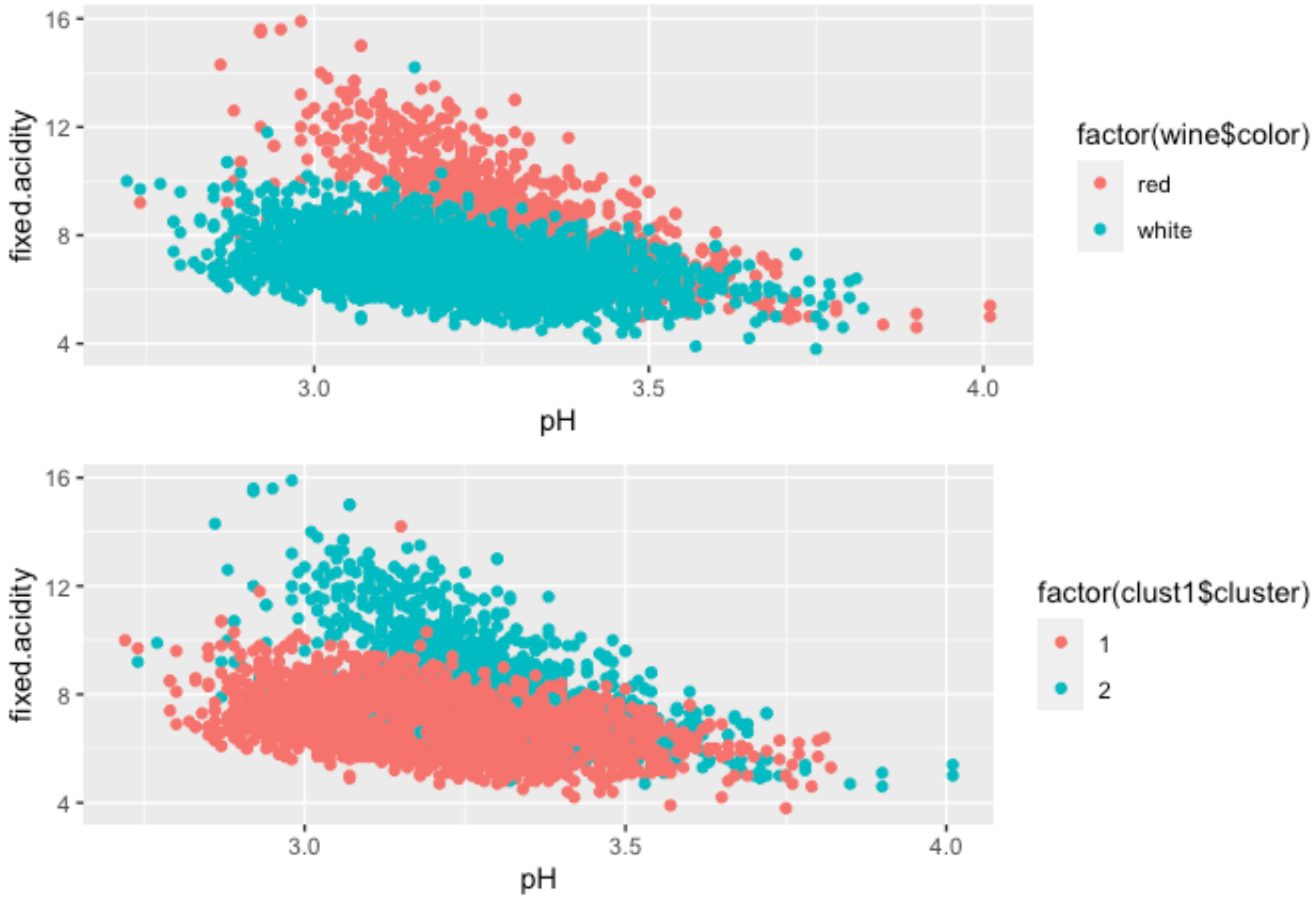


From the boxplots we can see that the red and white wine differs mostly from fixed.acidity, total sulfur dioxide, volatile acidity, and pH.

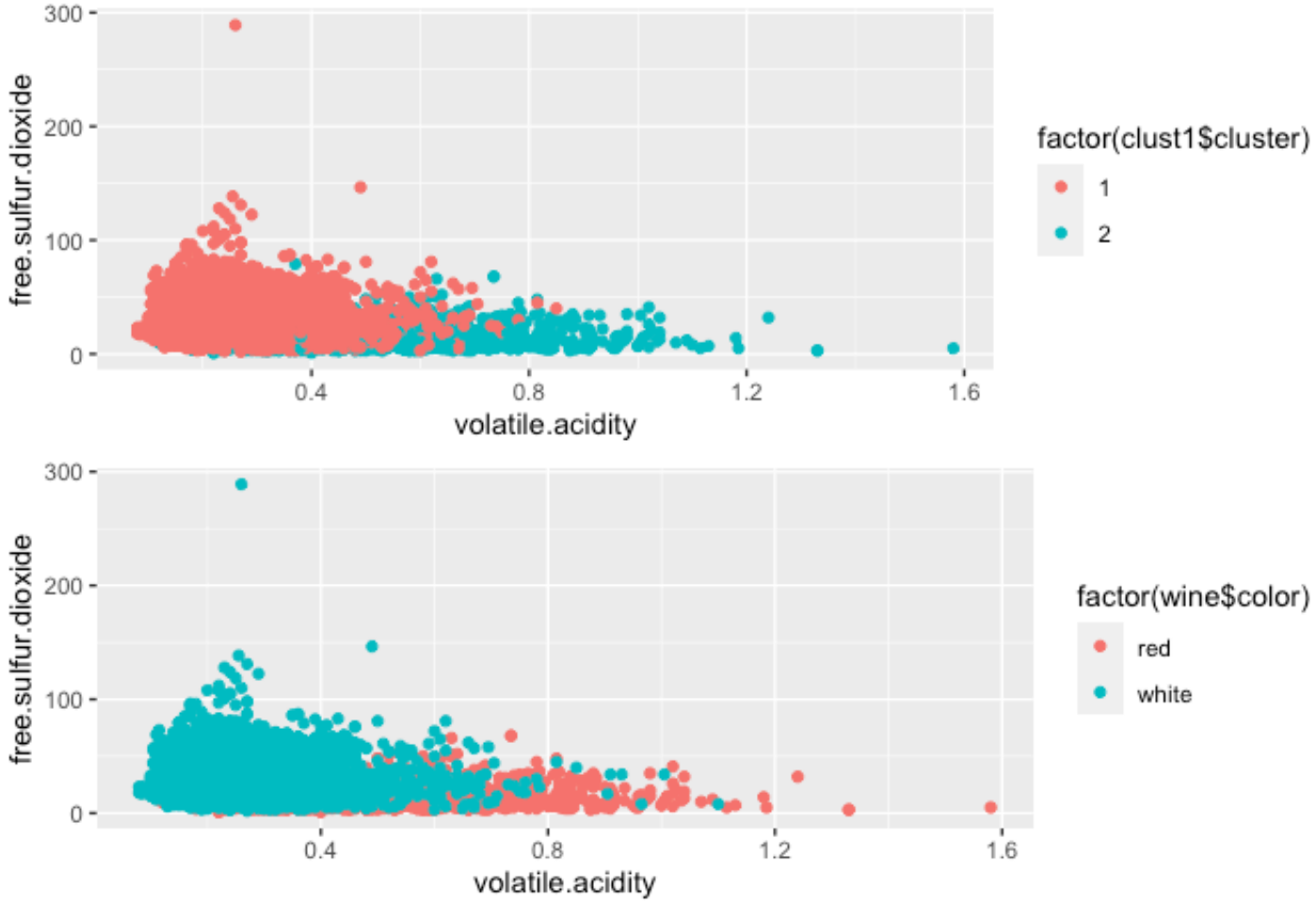
k-means Clustering

Now we use the k-means clustering to different the red and white wine by color, so we set the cluster of 2, and run the
k-means clustering on 11 chemical properties.

'igure 1.2 Results of K-means clustering for wine colors in the dimensions of fixed.acidity and pH



3 Results of K-means clustering for wine colors in the dimension of volatile.acidity and free.sulfu



From the graphs above, we can see that the k-means clustering successfully depart the two color of wine according to their chemical properties. Because in the figure 1.1, the two kinds of wine differs remarkably in four chemical properties, so we only show the clustering results of these four dimensions.

PCA

Next we use PCA to distinguish the reds from the whites. From the results below we can conclude that the first 5 PC explain nearly 80% of the variation.

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.7407 1.5792 1.2475 0.98517 0.84845 0.77930 0.72330
## Proportion of Variance 0.2754 0.2267 0.1415 0.08823 0.06544 0.05521 0.04756
## Cumulative Proportion  0.2754 0.5021 0.6436 0.73187 0.79732 0.85253 0.90009
##                           PC8     PC9    PC10    PC11
## Standard deviation     0.70817 0.58054 0.4772 0.18119
## Proportion of Variance 0.04559 0.03064 0.0207 0.00298
## Cumulative Proportion  0.94568 0.97632 0.9970 1.00000
```

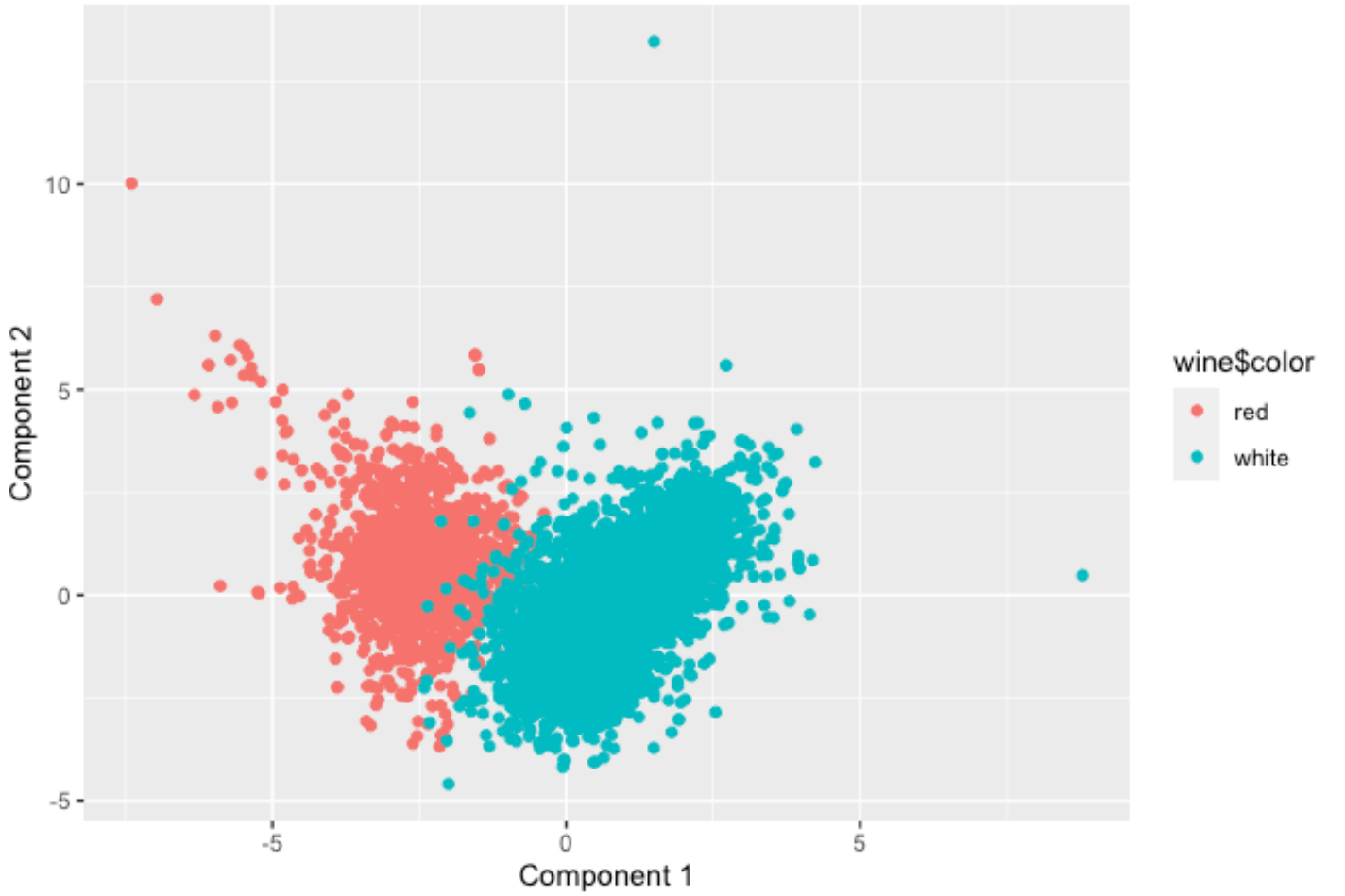
Next we examine the linear relationships between the first 5 PC and the original chemical properties.

Table 1.1. The first five principal components

	PC1	PC2	PC3	PC4	PC5
fixed.acidity	-0.24	0.34	-0.43	0.16	-0.15
volatile.acidity	-0.38	0.12	0.31	0.21	0.15
citric.acid	0.15	0.18	-0.59	-0.26	-0.16
residual.sugar	0.35	0.33	0.16	0.17	-0.35
chlorides	-0.29	0.32	0.02	-0.24	0.61
free.sulfur.dioxide	0.43	0.07	0.13	-0.36	0.22
total.sulfur.dioxide	0.49	0.09	0.11	-0.21	0.16
density	-0.04	0.58	0.18	0.07	-0.31
pH	-0.22	-0.16	0.46	-0.41	-0.45
sulphates	-0.29	0.19	-0.07	-0.64	-0.14
alcohol	-0.11	-0.47	-0.26	-0.11	-0.19

In the following graph we can see the PCA can cluster two types of wine by the first two principal components, where the white wine tends to be in the positive dimension of first principle component.

Figure 1.4 Results of PCA for wine colors in the dimensions of first two principal c



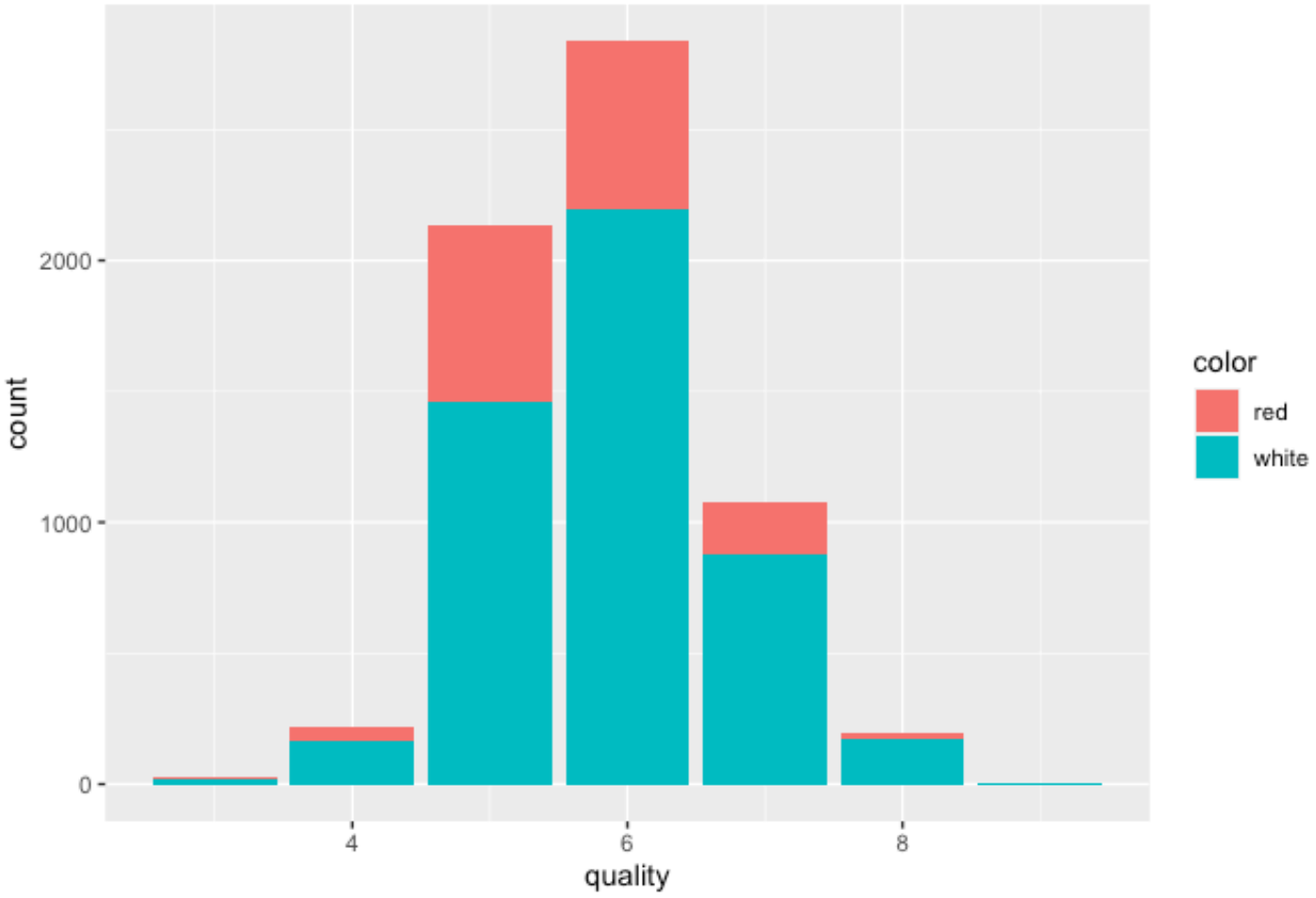
In conclusion, both k-means clustering and PCA can successfully distinguish the red and white wine by only using unsupervised information.

Wine Quality

The bar plot shows the the distribution of of the wine quality, which has 7 degrees from 3 to 9, and We manually divided them into 3 levels low, medium, and high. According to the boxplot, we can see that wines with different quality may differ in volatile acidity, alcohol, and free sulfur dioxide.

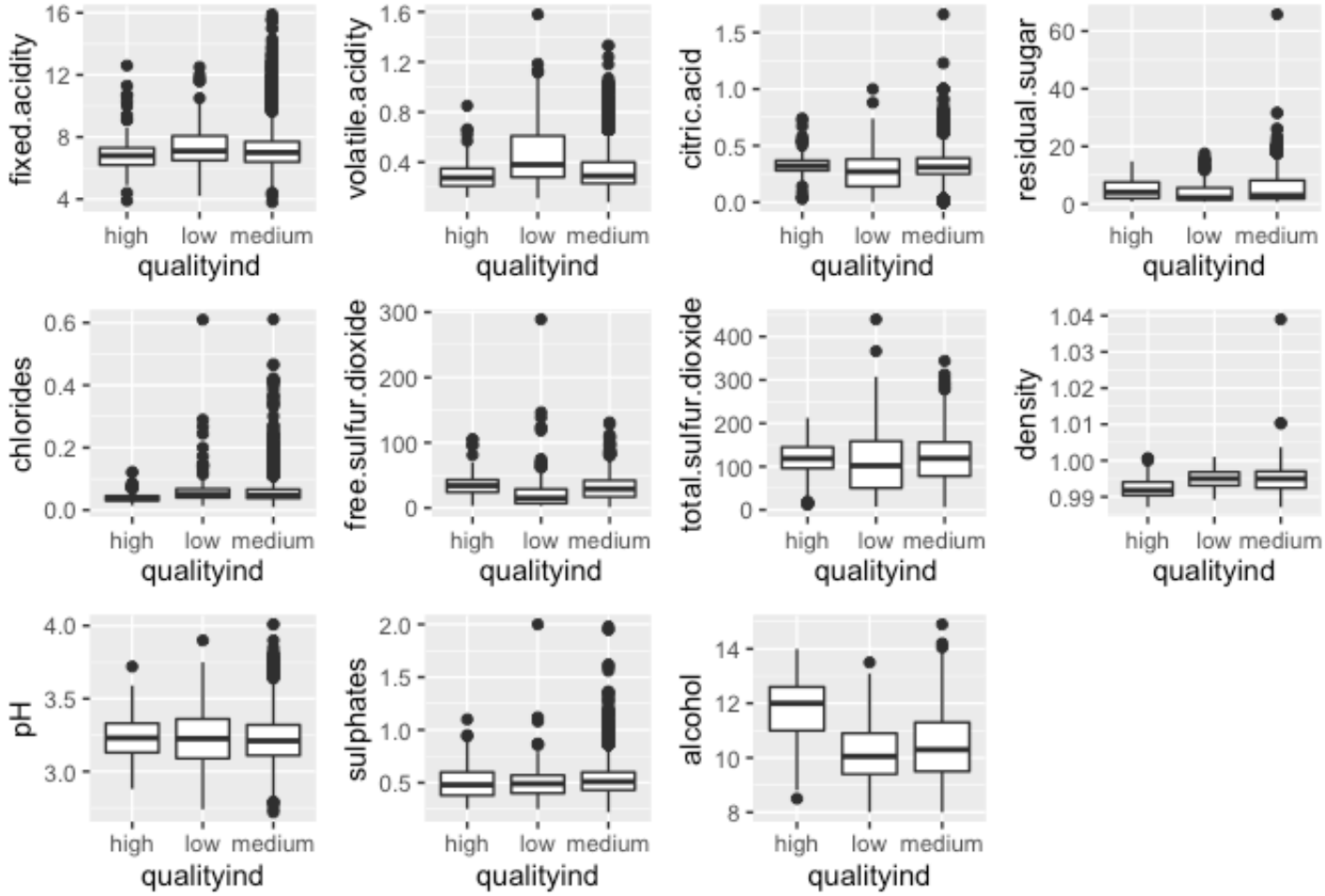
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.818   6.000   9.000
```

Figure 1.5 Histogram of wine qualities



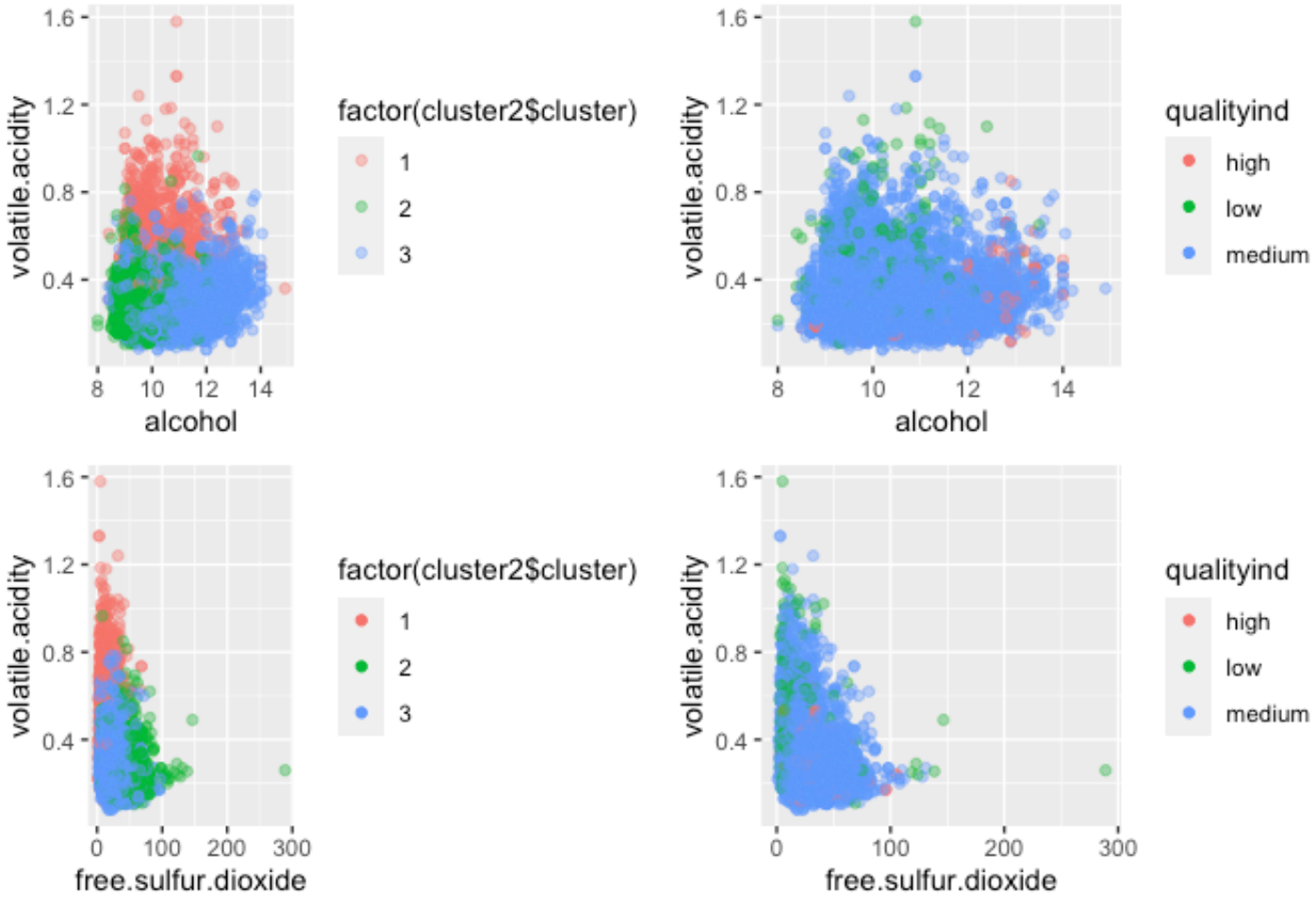
Clustering

Figure 1.6 Boxplots of chemical properties of three wine quality levels



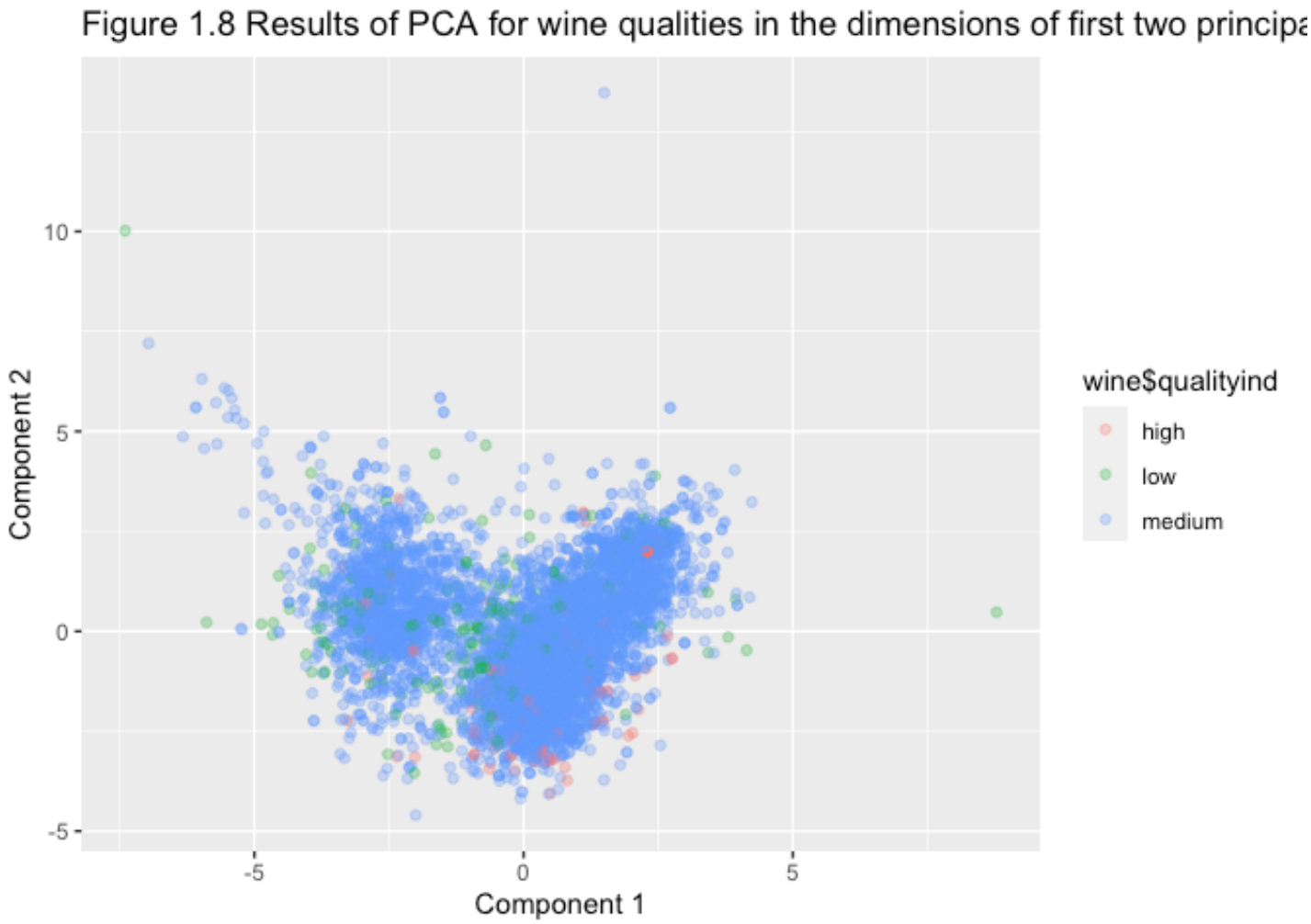
The following figure shows the results of k-means clustering to differentiate the quality of wines according to the chemical properties. However, it's difficult to see how wines with different qualities differs with each other.

Figure 1.7 Results of K-means clustering for wine qualities.



PCA

Now we try PCA to distinguish the wine quality. From previous results we conclude that the first 5 PC explain closely 80% of the
variation.



The graph above also shows that PCA couldn't perform the clustering of wine quality.

Market Segmentation

Introduction

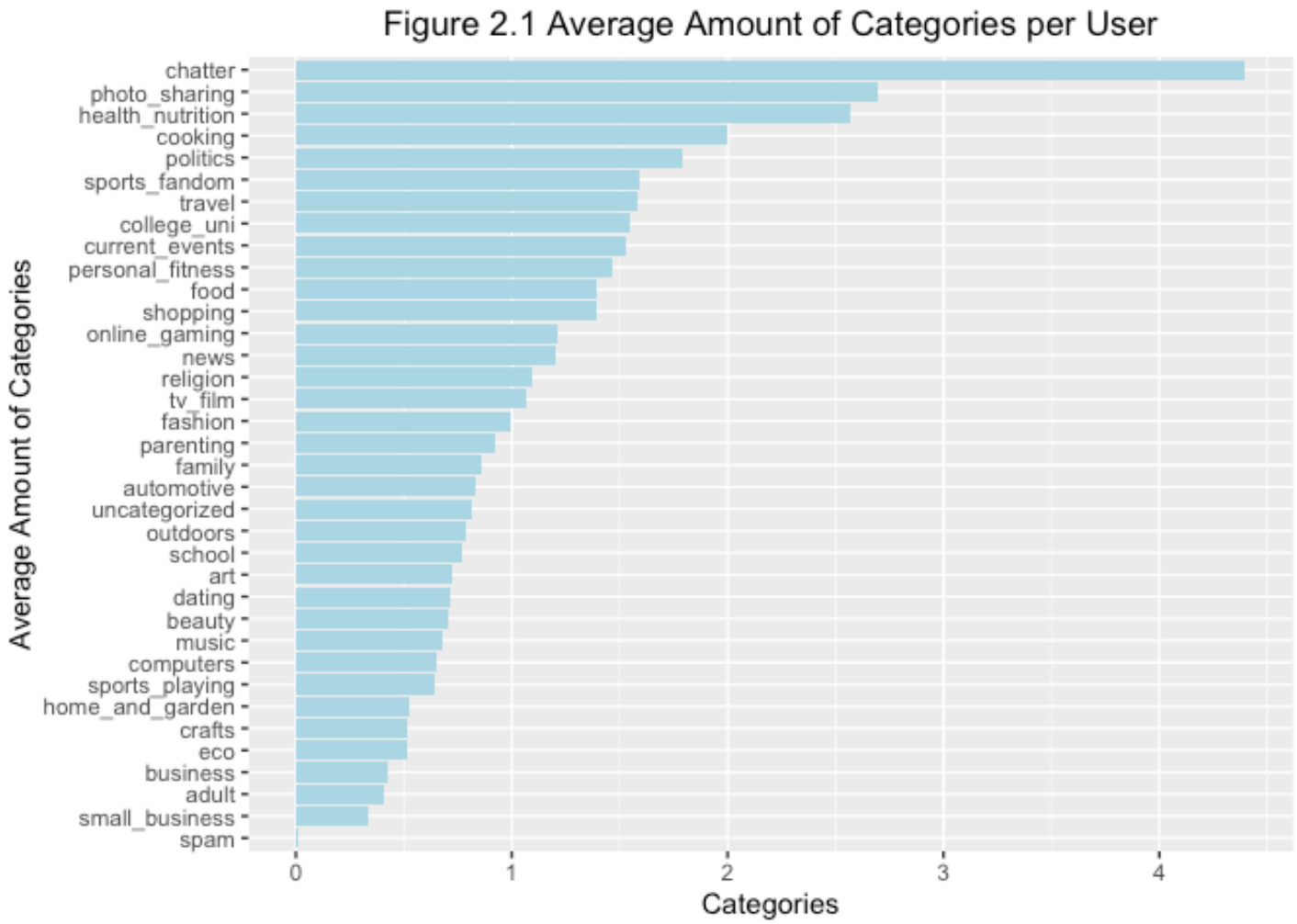
"NutrientH20", a large consumer drinks brand(hypothetically), wants to better understand the market based on the users tweet's content in Twitter post. Each tweet was categorized based on its content using a pre-specified scheme of 36 different categories. The goal for this analysis is to achieve market segmentation or user portrait on the basis of the a great bunch of tweets'contents.

EDA

Initially, we take an overview of dataset by summing up the amount of average amount of categories per user.

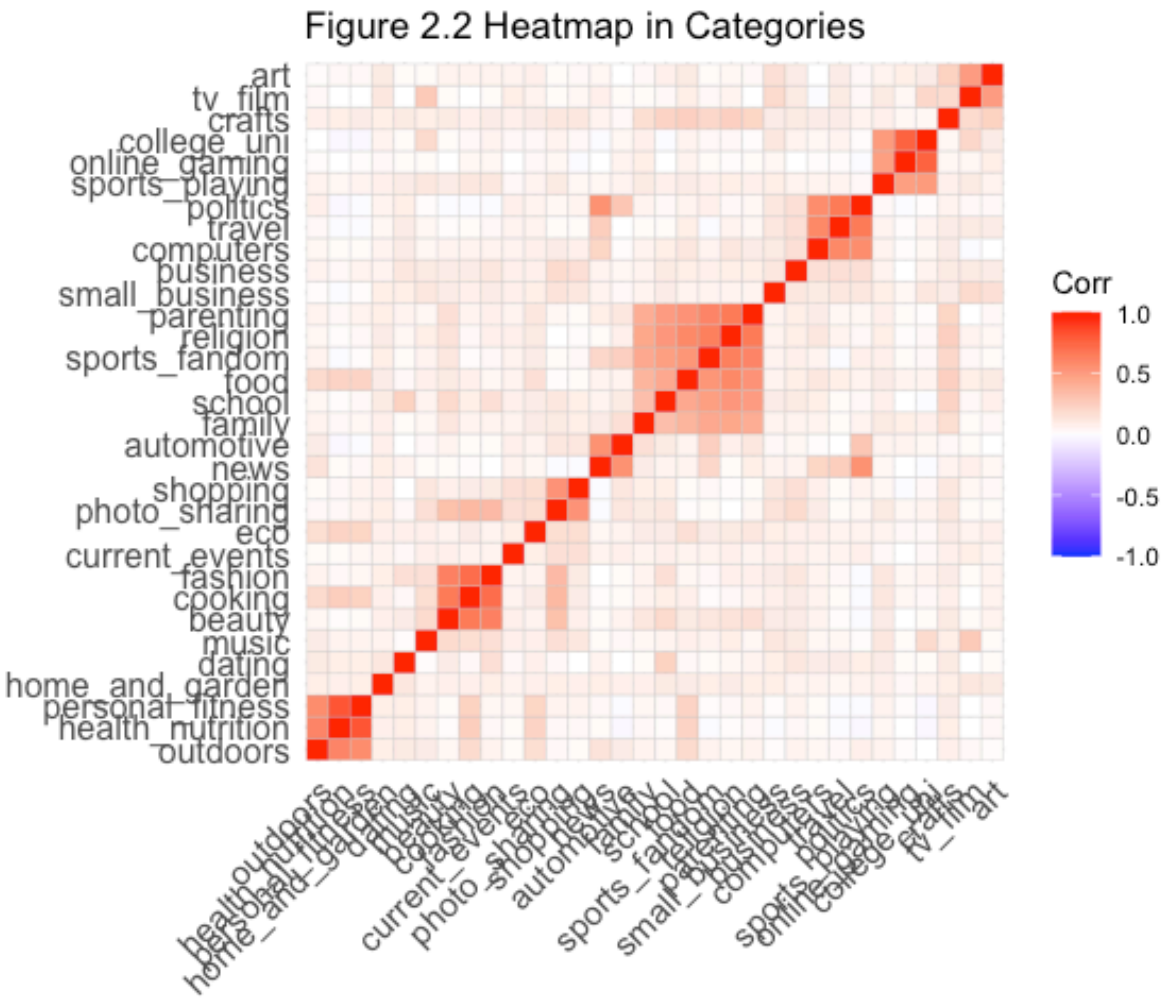
```
##           X chatter current_events travel photo_sharing uncategorized tv_film
## 1 hmjoe4g3k       2              0      2             2             2       1
## 2 clk1m5w8s       3              3      2             1             1       1
## 3 jcsovtak3       6              3      4             3             1       5
## 4 3oeb4hiln       1              5      2             2             0       1
```

```
## 5 fd75x1vgk       5              2      0             6             1       0
## 6 h6nvj91yp       6              4      2             7             0       1
##   sports_fandom politics food family home_and_garden music news online_gaming
## 1             1        0    4      1               2     0    0             0
## 2             4        1    2      2               1     0    0             0
## 3             0        2    1      1               1     1    1             0
## 4             0        1    0      1               0     0    0             0
## 5             0        2    0      1               0     0    0             3
## 6             1        0    2      1               1     1    0             0
##   shopping health_nutrition college_uni sports_playing cooking eco computers
## 1        1               17           0              2       5   1         1
## 2        0                0           0              1       0   0         0
## 3        2                0           0              0       2   1         0
## 4        0                0           1              0       0   0         0
## 5        2                0           4              0       1   0         1
## 6        5                0           0              0       0   0         1
##   business outdoors crafts automotive art religion beauty parenting dating
## 1        0        2      1          0   0        1      0         1      1
## 2        1        0      2          0   0        0      0         0      1
## 3        0        0      2          0   8        0      1         0      1
## 4        1        0      3          0   2        0      1         0      0
## 5        0        1      0          0   0        0      0         0      0
## 6        1        0      0          1   0        0      0         0      0
##   school personal_fitness fashion small_business spam adult
## 1      0               11       0              0    0     0
## 2      4                0       0              0    0     0
## 3      0                0       1              0    0     0
## 4      0                0       0              0    0     0
## 5      0                0       0              1    0     0
## 6      0                0       0              0    0     0
```



As shown in the Figure 2.1, the bar chart shows the average frequency of the topic. Chatter(No category) appears the most. “Photo_sharing”, “health_nutrition”, “cooking”, “politics” and “sports_fandom” are top 5 meaningful topics in the users interests.

We will remove “chatter”, “spam”, “adult”, “uncategorized”, which are meaningless for this topic. Plot the **heatmap** to show the relationship between different categories.

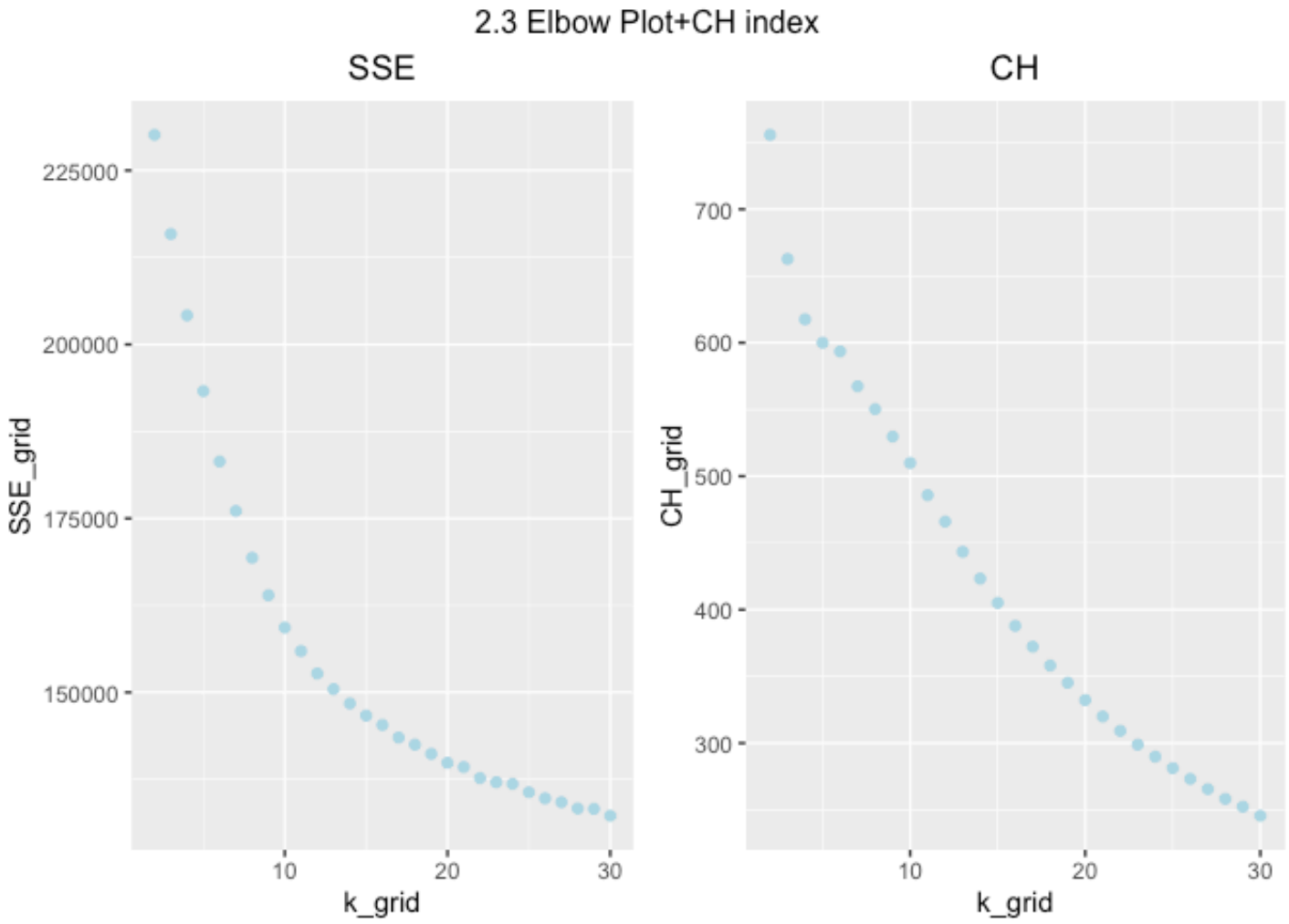


In Figure 2.2 **heatmap**, the interesting thing is all the variables are mutually positive correlative because it starts from 0.

Methodology

To achieve market segmentation, the methodology is unsupervised learning, including **Kmeans** and **PCA**.

(1) Kmeans



By eyeballing observation, we find 9 is the optimal number for group numbers of clustering.

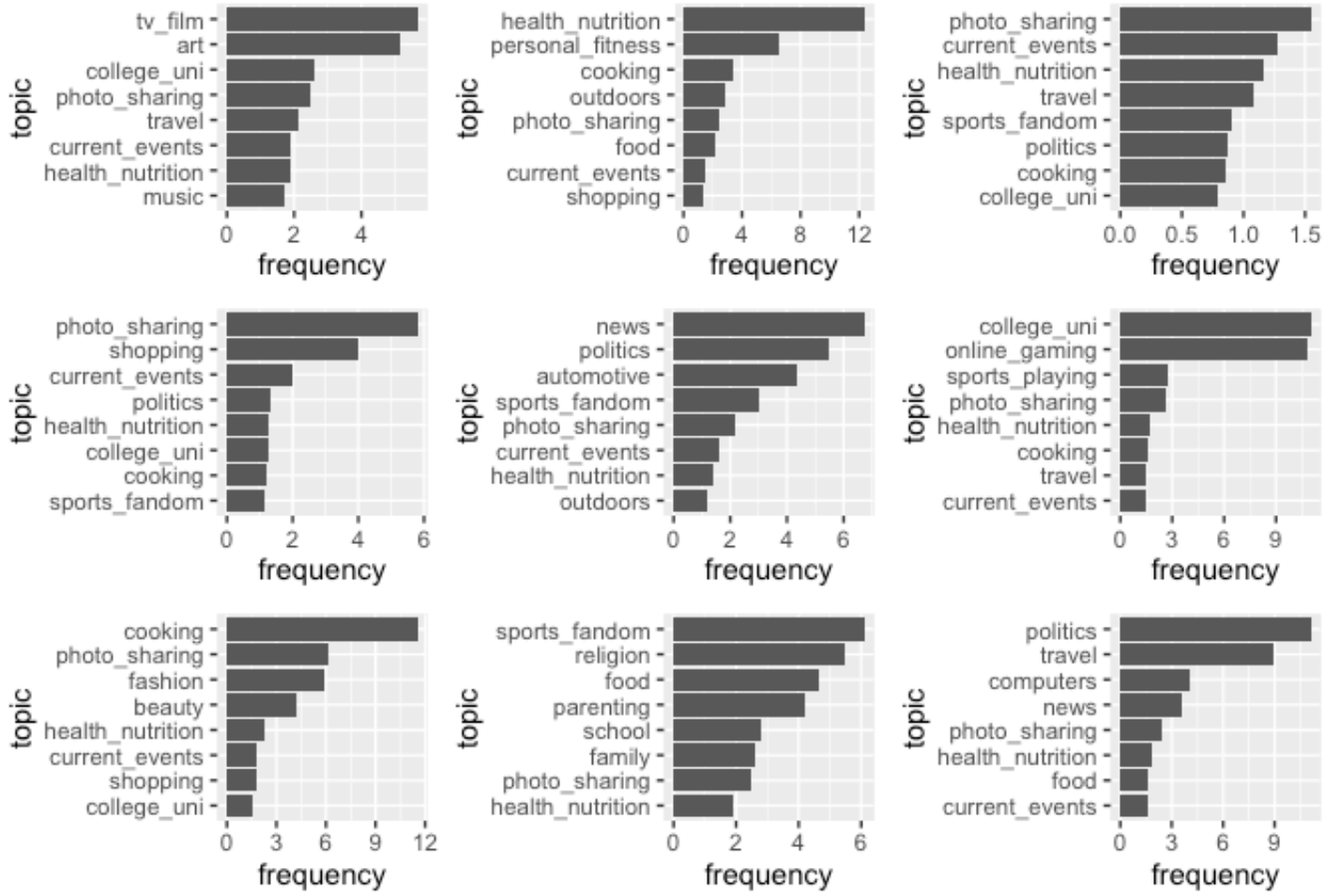
**Table 2.1 : The Number of Cluster Group **

Label	Numbers
1	405
2	804
3	3378
4	960
5	437
6	363
7	491
8	685
9	359

Table 2.2: Top 5 Categories in Cluster

	life_moment	health-life	lifestyle	college_life	man-topic	lady-topic	art_style	News	Complex
no.1	tv_film	health_nutrition	photo_sharing	photo_sharing	news	college_uni	cooking	sports_fandom	politics
no.2	art	personal_fitness	current_events	shopping	politics	online_gaming	photo_sharing	religion	travel
no.3	college_uni	cooking	health_nutrition	current_events	automotive	sports_playing	fashion	food	computers
no.4	photo_sharing	outdoors	travel	politics	sports_fandom	photo_sharing	beauty	parenting	news
no.5	travel	photo_sharing	sports_fandom	health_nutrition	photo_sharing	health_nutrition	health_nutrition	school	photo_sharing

Figure 2.4 Top 8 Categories in 9 Cluster



Using **Kmeans**, we create 9 clusters. The number of cluster group are illustrated in the table 2.1. We add the label into original data and Top 8 categories are shown in the Figure 4.4. Based on these high-frequency categories, we empirically build up these topics for different labels. The Clusters are (Table 2.2) as follows:

For Cluster1, “Photosharing”, “shopping” and “current events” are most topics. (life-moment)

For Cluster2, “health_nutrition”, “personal_fitness” and “cooking” are most topics. (health-life)

For Cluster3, “sports_fandom”, “religion” and “food” are most topics. (lifestyle)

For Cluster4, ”college_uni“,”online_gaming" and “sports_playing” are most topics. (college_life)

For Cluster5, “news”, “politics” and “automotive” are most topics. (man-topic)

For Cluster6, “cooking”, “photo-sharing”, “fashion”, “beauty” are most topics. (lady-topic)

For Cluster7, “TV_film”, “arts” are most topics. (art_style)

For Cluster8, “Politics”, “travel”, “computer”, “news” are most topics.(News)

For Cluster9, “Photosharing”, “current events” and “health_nutrition” are most topics. (Complex) (hard to explain)

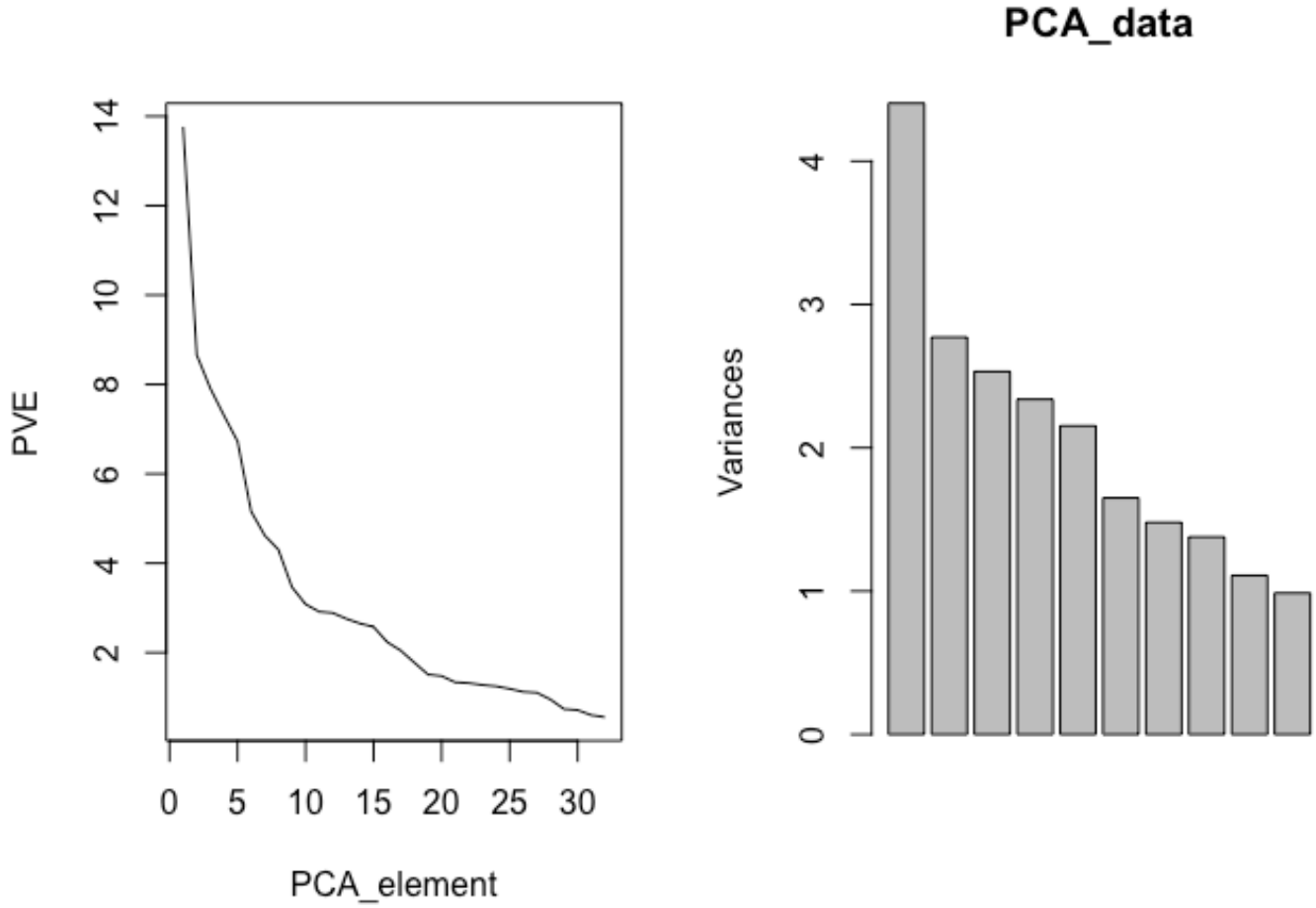
(2) PCA

PCA, focus on reducing the feature space, allowing most of the information or variability in the data set to be explained using fewer features.

Table 2.3 : Summary of PCA Variance

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.098691	1.664833	1.591187	1.529121	1.467282	1.284772
Proportion of Variance	0.137640	0.086610	0.079120	0.073070	0.067280	0.051580
Cumulative Proportion	0.137640	0.224260	0.303380	0.376450	0.443720	0.495310

Figure 2.4 Variance in PCA



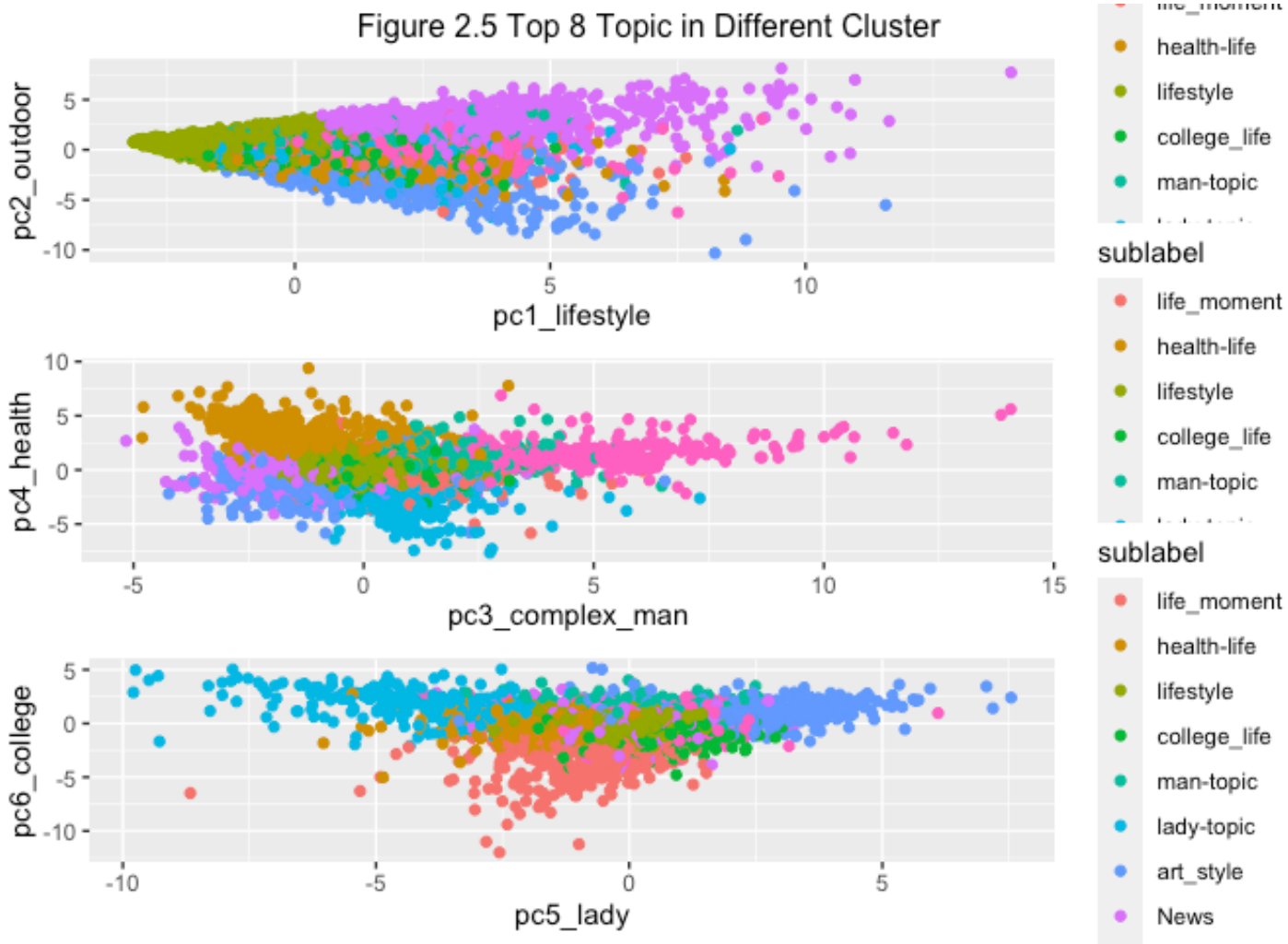
Based on We use 6 components because these components could explain 50% variance and decreasing rate of variance becomes much slower after that. Hence, we choose 6 components to analysis.

Table 2.4: Top 5 Categories in PCA

categories	pc1_lifestype	pc2_outdoor	pc3_complex_man	pc4_health	pc5_lady	pc6_college
no.1	religion	sports_fandom	politics	health_nutrition	beauty	online_gaming
no.2	food	religion	travel	personal_fitness	fashion	sports_playing
no.3	parenting	parenting	computers	outdoors	cooking	college_uni
no.4	sports_fandom	food	news	politics	photo_sharing	cooking
no.5	school	school	automotive	news	shopping	automotive

As illustrated in the **Table 2.4**, we empirically make labels for different principle component. To some extent, these components are similar to clusters labels. In the next, we will contrast these two unsupervised methodology.

Figure 2.5 Top 8 Topic in Different Cluster



From the Figure 2.5, we find some strong link between PCA and Cluster, which proves that our cluster are meaningful.

For the first graph,

Cluster lifestyle has high point in pca1_lifestyle and pca2_outdoor

For the second graph,

Cluster health life has high point in pca4_health

Cluster college life has low point in pca4_health.

Cluster News has high point in pc3_complex_man

Cluster lay_topic has low point in pc3_complex

For the third graph,

Cluster lady topic has high point in pc5_lady

Cluster college life has low point in pc5_lady

Cluster college life has high point in pc6_college

Cluster art_style has low point in pc6_college

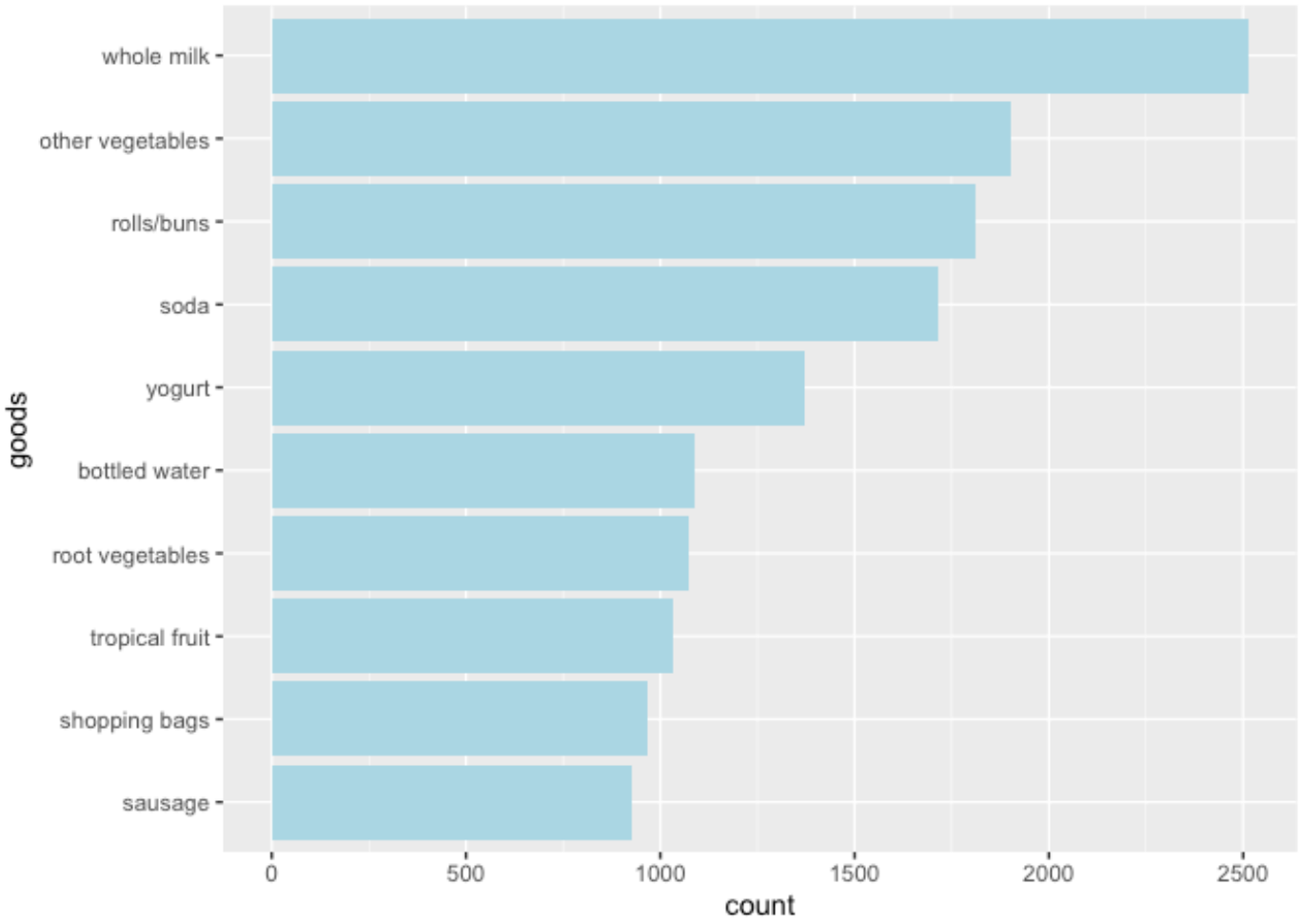
Conclusion

We use k-means and PCA to make interesting market segmentation for NutrientH20. The interesting group we found is health, outdoor group focusing on outdoor life and health nutrition. Besides, lady-topic group focusing on beauty and fashion need more customized lady-like product. In addition, college-life group also clearly identified by PCA and Kmeans. NutrientH20 could design some activity or sub-brand target for college student.

Association rules for grocery purchases

Top Ten Popular Goods

We show the most 10 popular goods among the cousumers. From the plot, we can see that the most popular goods are whole milk, other vegetables, rolls and buns, soda and yogurt.



Total Association Rules

We use the rules with *support >= 0.01, confidence >= 0.1, length (maxlen) <= 2.* We find 45 rules in total, and the following table show the summary of all the association rules.

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support
minlen
```

```
##         0.1    0.1    1 none FALSE            TRUE       5    0.01
1
##  maxlen target  ext
##       2  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 152
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 15296 transaction(s)] done [0.00s].
## sorting and recoding items ... [71 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 done [0.00s].
## writing ... [45 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].

## set of 45 rules
##
## rule length distribution (lhs + rhs):sizes
##  1  2
##  4 41
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   1.911   2.000   2.000
##
## summary of quality measures:
##     support          confidence        coverage             lift
##  Min.   :0.01020   Min.   :0.1006   Min.   :0.03426   Min.   :0.9421
##  1st Qu.:0.01255   1st Qu.:0.1277   1st Qu.:0.06041   1st Qu.:1.4236
##  Median :0.01438   Median :0.1786   Median :0.08970   Median :1.6789
##  Mean   :0.02654   Mean   :0.1959   Mean   :0.17099   Mean   :1.8383
##  3rd Qu.:0.02262   3rd Qu.:0.2408   3rd Qu.:0.12441   3rd Qu.:1.9991
##  Max.   :0.16429   Max.   :0.4037   Max.   :1.00000   Max.   :3.8648
##      count
##  Min.   : 156
##  1st Qu.: 192
##  Median : 220
##  Mean   : 406
##  3rd Qu.: 346
```

```
##  Max.   :2513
##
## mining info:
##        data ntransactions support confidence
##  goodstrans         15296    0.01        0.1
##
               call
##  apriori(data = goodstrans, parameter = list(support = 0.01, confidence
= 0.1, maxlen = 2))
```

Subset

We check the subset to see less rules.

First, we check the rule with *lift >= 3*, which indicating strong connections. We can see four rules in the following table. From the table, we can see that, for example, the first rules shows if a consumer buy pip fruit, the probability that he also but tropical fruit is 3.9 times higher than if the consumer doesn't buy pip fruit. Given this information, grocery can put pip fruit by the tropical fruit.

```
##     lhs                rhs              support    confidence coverage

## [1] {pip fruit}      => {tropical fruit} 0.01268305 0.2607527
0.04864017
## [2] {tropical fruit} => {pip fruit}      0.01268305 0.1879845
0.06746862
## [3] {citrus fruit}   => {tropical fruit} 0.01248692 0.2346437
0.05321653
## [4] {tropical fruit} => {citrus fruit}   0.01248692 0.1850775
0.06746862
##     lift    count
## [1] 3.86480 194
## [2] 3.86480 194
## [3] 3.47782 191
## [4] 3.47782 191
```

Secondly, there are 5 rules satisfying *confidence > 0.3.* For example, we can see that, the first rule shows that if a consumer buy curd, then we are 37% positive that he will also buy whole milk. This type of consumer may want to make desserts by using milk and curd. Given this information, grocery can put curd by the whole milk.

```
##     lhs                   rhs                support    confidence
coverage
## [1] {curd}             => {whole milk}       0.01261768 0.3683206
0.03425732
## [2] {butter}           => {whole milk}       0.01438285 0.4036697
0.03563023
## [3] {root vegetables}  => {other vegetables} 0.02536611 0.3619403
0.07008368
## [4] {root vegetables}  => {whole milk}       0.02262029 0.3227612
0.07008368
## [5] {other vegetables} => {whole milk}       0.04086036 0.3284288
0.12441161
##     lift     count
## [1] 2.241875 193
## [2] 2.457036 220
## [3] 2.909216 388
## [4] 1.964566 346
## [5] 1.999064 625
```

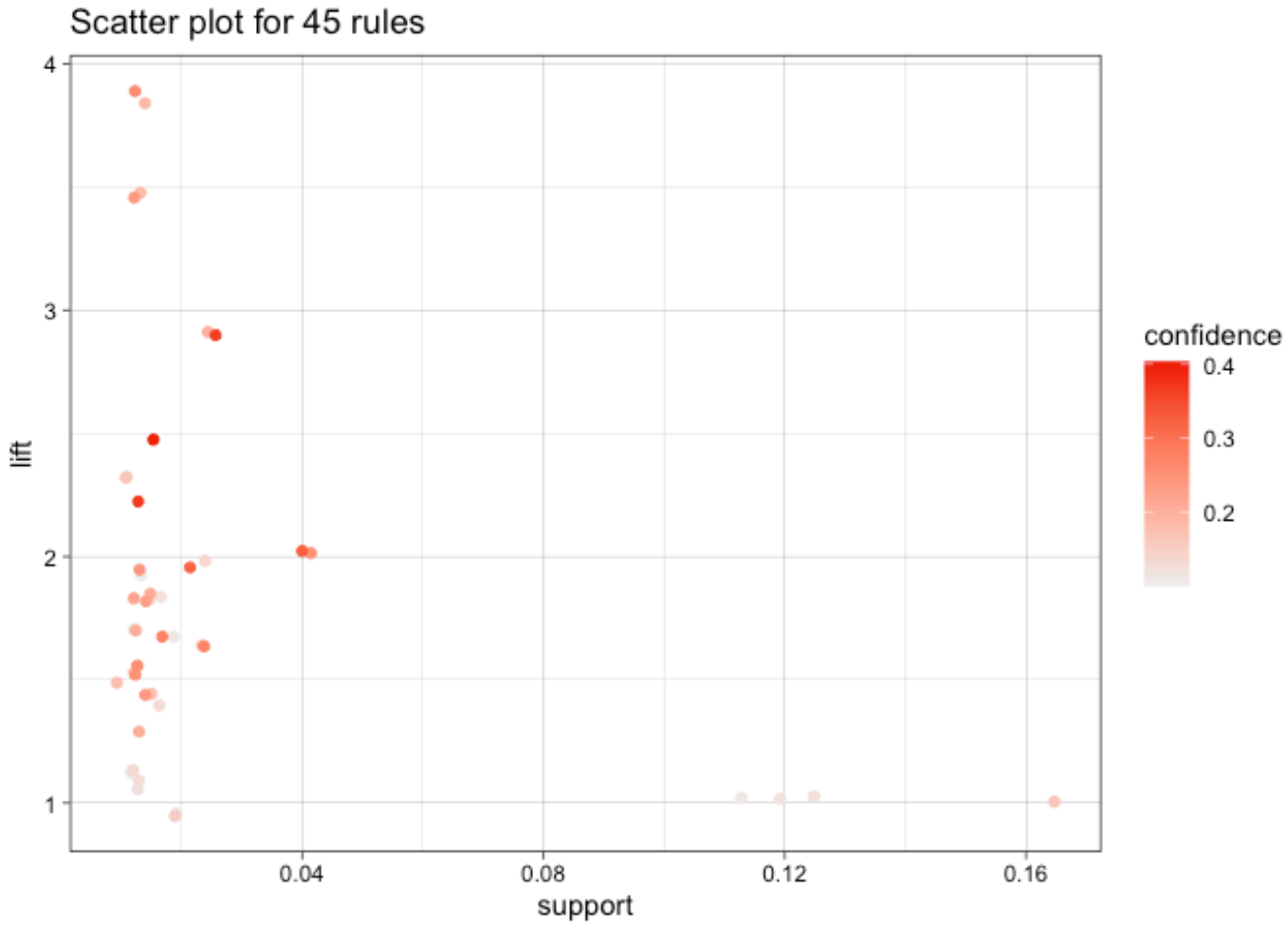
Thirdly, as the thresholds we set above are too strict that once we combine them, there would be too little association rules satisfied, we purposely relaxed the constraints. Considering visualization and associations, we set *lift > 2 & confidence > 0.2,* and there are 20 rules satisfied. For example, if a consumer buy butter, then we are 40% positive that he will also buy whole milk, and this is 2.25% times higher than if he don't buy butter. Also, if a consumer buy whipped / sour cream, then we are 25% positive that he will also buy whole milk, and this is 1.5% times higher than if he don't buy whipped / sour cream. So given these information, it make sense to put the foods that need to make dessert together in grocery store.

```
##      lhs                    rhs                support    confidence
## [1]  {curd}                 => {whole milk}       0.01261768 0.3683206
## [2]  {butter}               => {whole milk}       0.01438285 0.4036697
## [3]  {whipped/sour cream}   => {whole milk}       0.01144090 0.2482270
## [4]  {pip fruit}            => {tropical fruit}   0.01268305 0.2607527
```

```
## [5]  {pip fruit}          => {other vegetables} 0.01091789 0.2244624
## [6]  {pip fruit}          => {whole milk}       0.01255230 0.2580645
## [7]  {citrus fruit}       => {tropical fruit}   0.01248692 0.2346437
## [8]  {citrus fruit}       => {other vegetables} 0.01281381 0.2407862
## [9]  {citrus fruit}       => {whole milk}       0.01281381 0.2407862
## [10] {sausage}            => {other vegetables} 0.01261768 0.2088745
## [11] {sausage}            => {whole milk}       0.01255230 0.2077922
## [12] {bottled water}      => {soda}             0.01464435 0.2060718
## [13] {tropical fruit}     => {other vegetables} 0.01549425 0.2296512
## [14] {tropical fruit}     => {whole milk}       0.01830544 0.2713178
## [15] {root vegetables}    => {other vegetables} 0.02536611 0.3619403
## [16] {other vegetables}   => {root vegetables}  0.02536611 0.2038886
## [17] {root vegetables}    => {whole milk}       0.02262029 0.3227612
## [18] {yogurt}             => {whole milk}       0.02425471 0.2704082
## [19] {other vegetables}   => {whole milk}       0.04086036 0.3284288
## [20] {whole milk}         => {other vegetables} 0.04086036 0.2487067
##      coverage   lift     count
## [1]  0.03425732 2.241875 193
## [2]  0.03563023 2.457036 220
## [3]  0.04609048 1.510895 175
## [4]  0.04864017 3.864800 194
## [5]  0.04864017 1.804191 167
## [6]  0.04864017 1.570774 192
## [7]  0.05321653 3.477820 191
## [8]  0.05321653 1.935400 196
## [9]  0.05321653 1.465605 196
## [10] 0.06040795 1.678898 193
## [11] 0.06040795 1.264779 192
## [12] 0.07106433 1.837944 224
## [13] 0.06746862 1.845898 237
## [14] 0.06746862 1.651444 280
## [15] 0.07008368 2.909216 388
## [16] 0.12441161 2.909216 388
## [17] 0.07008368 1.964566 346
## [18] 0.08969665 1.645907 371
## [19] 0.12441161 1.999064 625
## [20] 0.16429132 1.999064 625
```

Visualization

The following is the scatter plot of all 45 rules.



We use network graph to visualize connections. We use the rules with *support > 0.005, confidence > 0.2,* there are 20 rules satisfied.

```
## set of 20 rules
##
## rule length distribution (lhs + rhs):sizes
##  2
## 20
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       2       2       2       2       2       2
##
## summary of quality measures:
##     support         confidence        coverage            lift
##  Min.   :0.01092   Min.   :0.2039   Min.   :0.03426   Min.   :1.265
##  1st Qu.:0.01260   1st Qu.:0.2284   1st Qu.:0.04864   1st Qu.:1.650
##  Median :0.01360   Median :0.2485   Median :0.06041   Median :1.891
```

```
##  Mean   :0.01828   Mean   :0.2670   Mean   :0.06957   Mean   :2.102
##  3rd Qu.:0.02303   3rd Qu.:0.2842   3rd Qu.:0.07033   3rd Qu.:2.296
##  Max.   :0.04086   Max.   :0.4037   Max.   :0.16429   Max.   :3.865
##      count
##  Min.   :167.0
##  1st Qu.:192.8
##  Median :208.0
##  Mean   :279.6
##  3rd Qu.:352.2
##  Max.   :625.0
##
## mining info:
##        data ntransactions support confidence
##  goodstrans         15296    0.01        0.1
##
               call
##  apriori(data = goodstrans, parameter = list(support = 0.01, confidence
= 0.1, maxlen = 2))
```