DM_PS4_all

Fan Ye, Jinming Li, Xiangmeng Qin

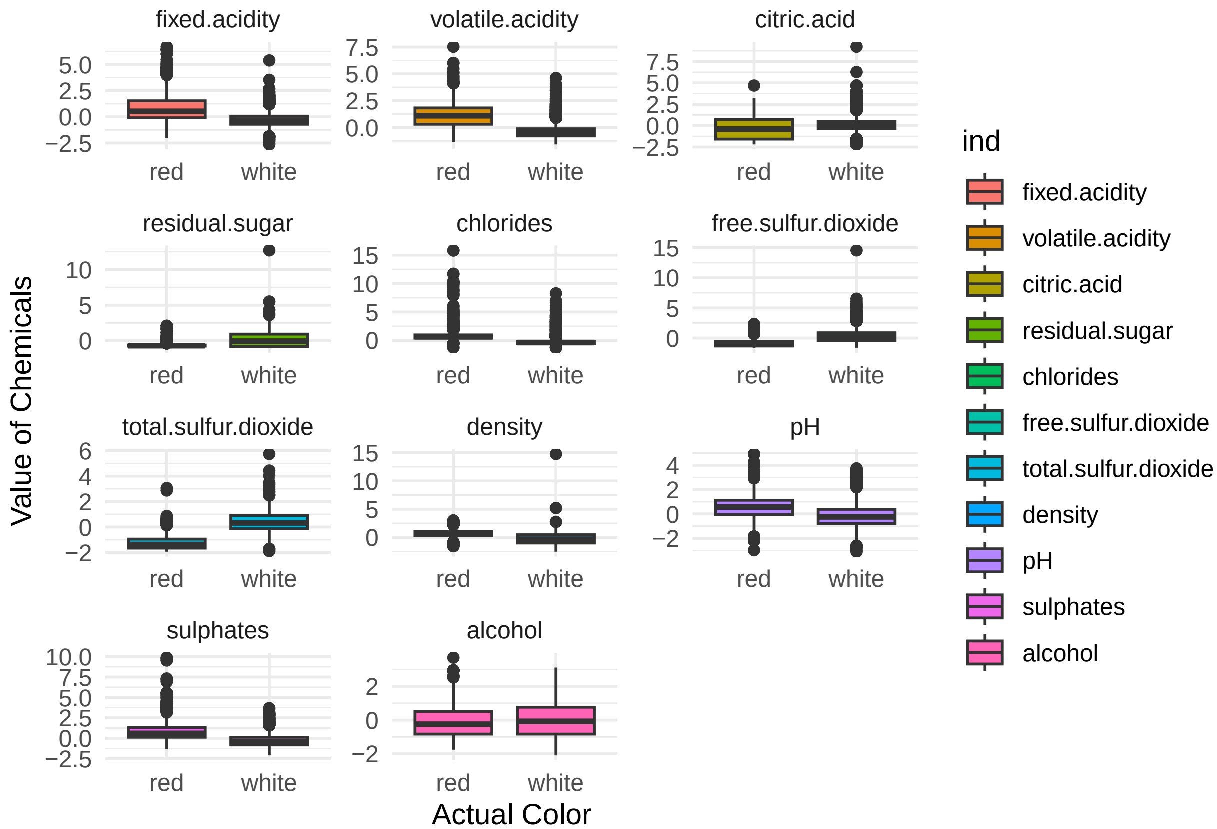
2024-04-22

Question 1: Clustering and PCA

Clustering

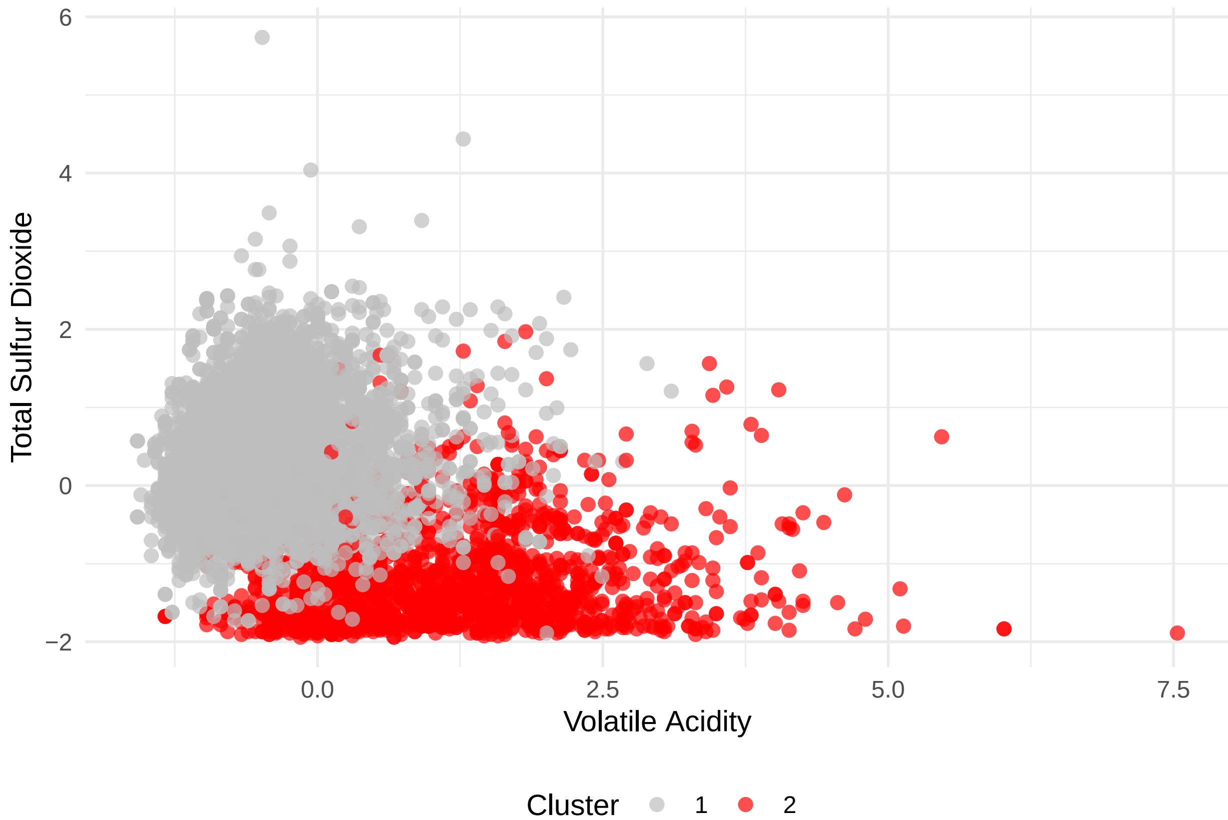
Color of wines

We standardizes the features of a wine dataset, excluding `quality` and `color`, and performs K-means clustering with two different numbers of centers (2 and 7). Then we visualizes the distribution of various chemical properties across actual wine colors using a box plot created with `ggplot2`.



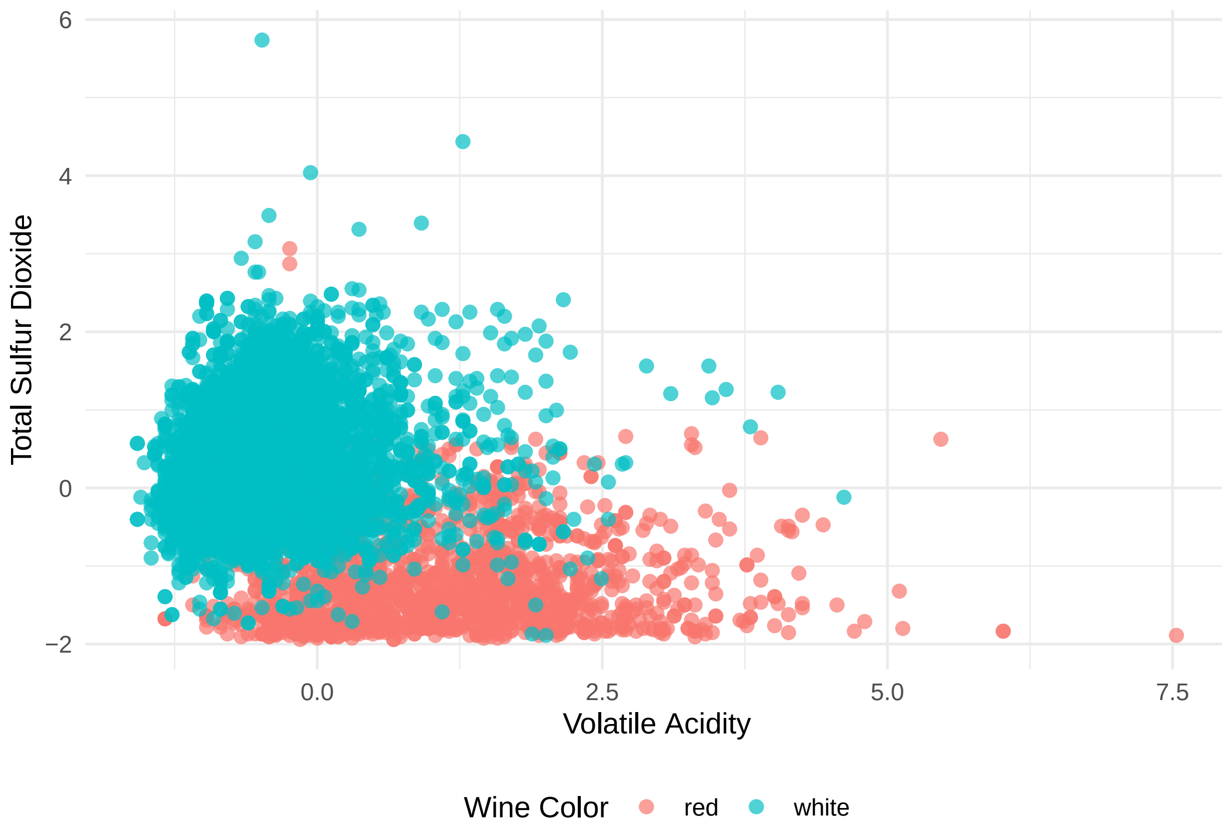
From the above results, it can be concluded that "colors can be most distinctly differentiated by volatile acidity and total sulfur dioxide." In the box plots, it is observed that these two chemical substances show significant differences in median values among different colors of wine.

`volatile.acidity` and `total.sulfur.dioxide` We attempt to show in the scatter plot how two chemical components of wine data are clustered according to color, marked by different colors for different cluster groups (volatile acidity, total sulfur dioxide). This visually demonstrates how these chemical components are distributed across different clusters.



The code is used to analyze and visualize how well the clustering algorithm classifies wines based on their chemical properties. Through clustering analysis, the first group has been identified as white wine, while the second group has been identified as red wine.

The code below show the distribution of existing wine color classifications and chemical properties (volatile acidity and total sulfide) in the data set.



confusion matrix

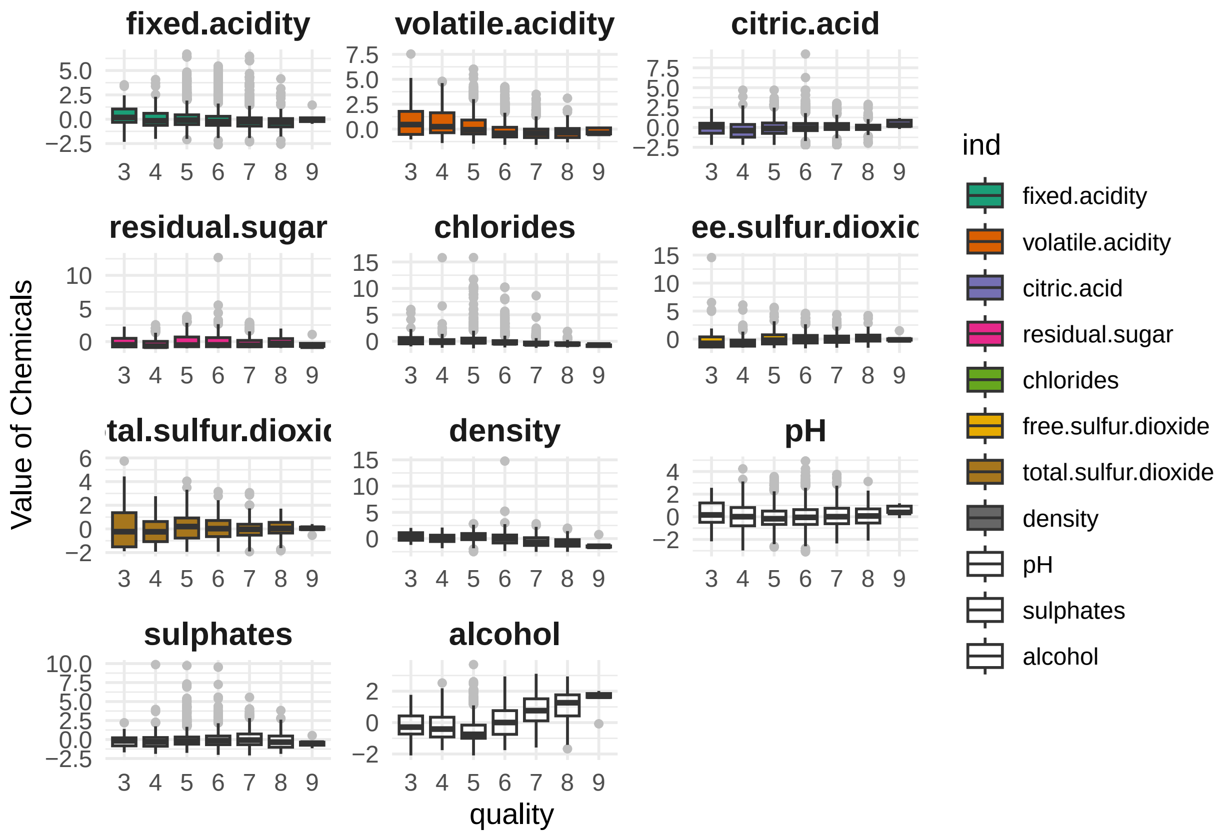
Using the confusion matrix to evaluate K-means can verify how the clusters align with the actual labels, especially in scenarios where color clusters are clustered, such as red wine versus white wine.

```
##         Predicted
## Actual      1    2
##   white 4830   68
##   red     24 1575
```

The clustering accuracy of category 1 (white wine) is very high because the vast majority of white wines are correctly grouped into this category. Category 2 (red wine) also showed high clustering accuracy, with most red wines correctly identified. The relatively small number of misclassifications suggests that the K-means clustering algorithm's ability to distinguish between red and white wines on this dataset is fairly accurate.

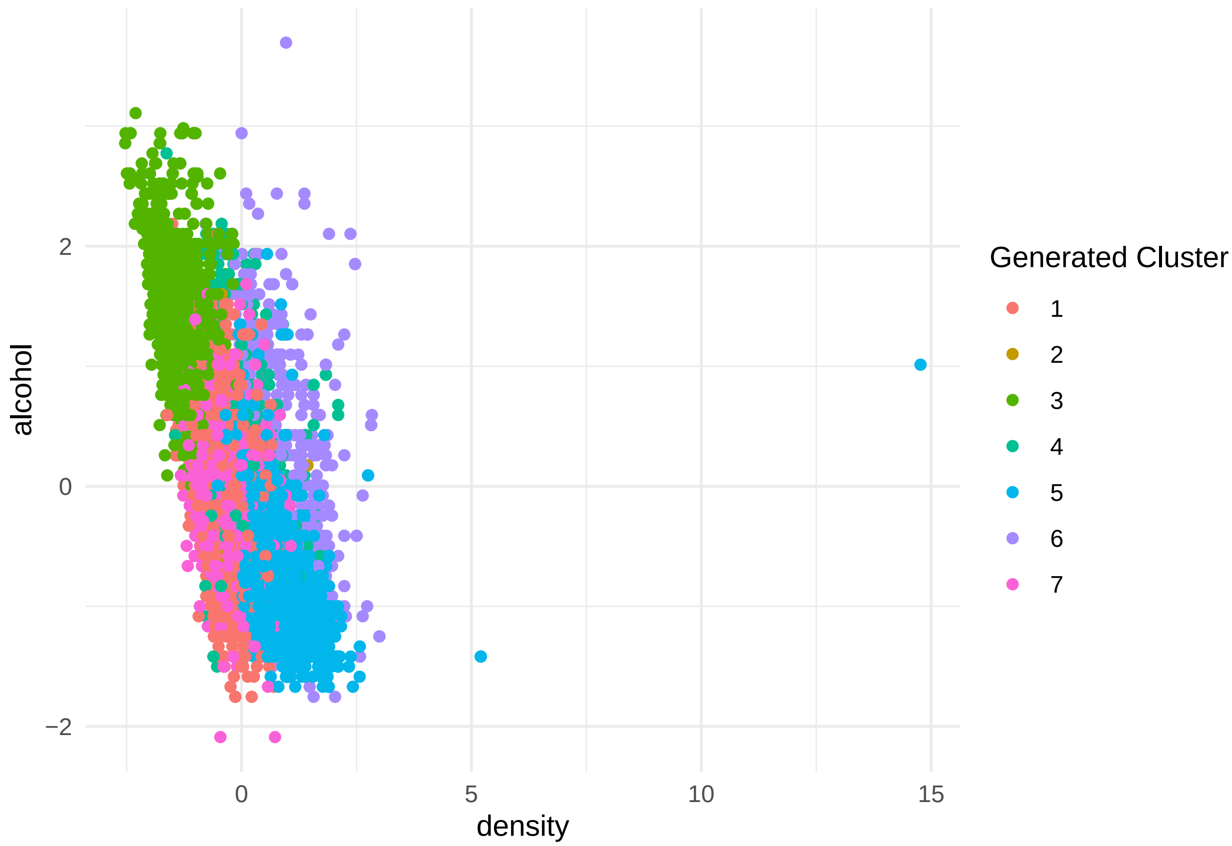
Quality of wines

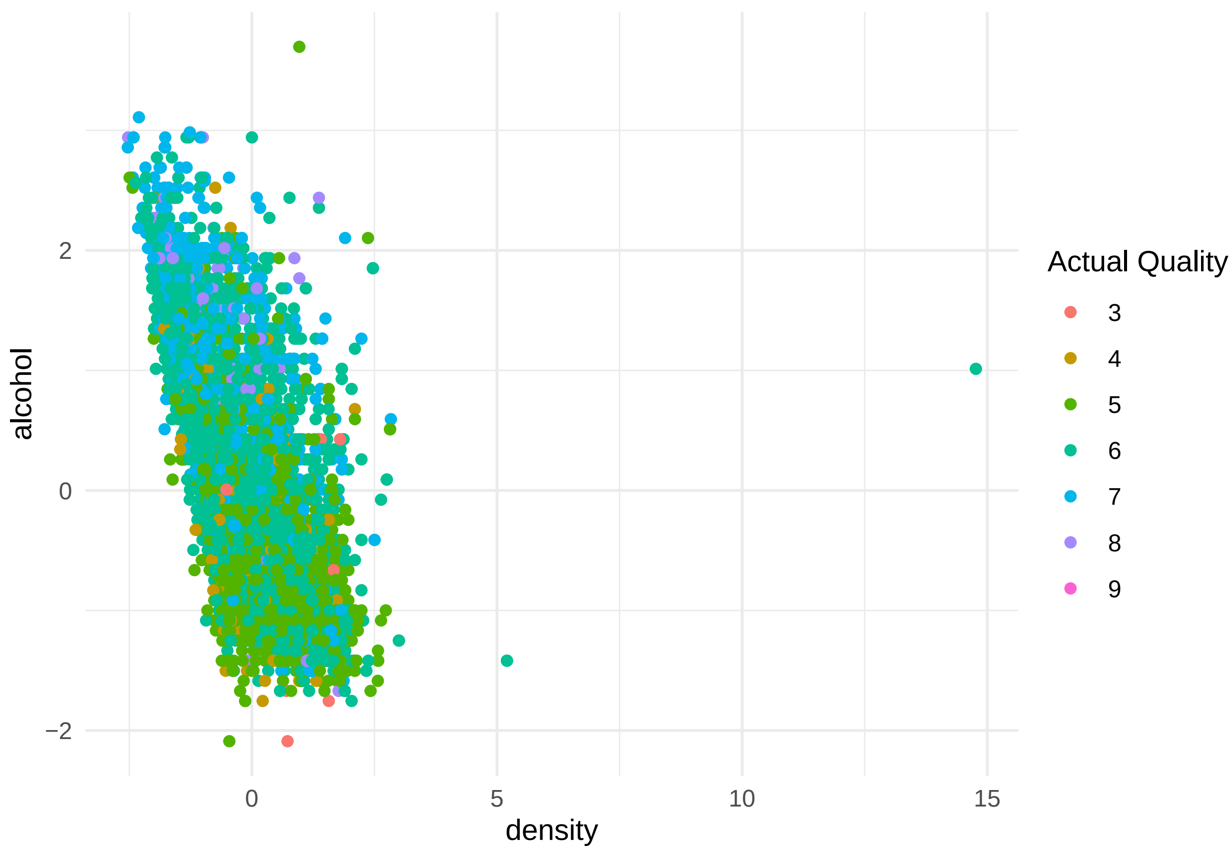
This code is used to create a boxplot to visualize the distribution of chemical composition values for wines of different qualities, thereby helping the observer understand the chemical differences between wine qualities.



Based on the median values of these characteristics, we predict that at least density and alcohol can distinguish between high quality wines.

density and alcohol





The distribution of the amount of wine in the cluster should be similar to the distribution in the real quality group, so that the cluster can classify the seven levels of quality. The lower the density, the higher the mass. The higher the alcohol, the higher the quality.

Principal component analysis

Color of wines

load matrix of PCA

We try to generate the load matrix of PCA

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
fixed.acidity	-0.2387989	-0.3363545	-0.4343013	0.1643462	-0.1474800	-0.2045537	-0.2830794	0.4012356	-0.3440567	0.2812677	0.3346793
volatile.acidity	-0.3807575	-0.1175497	0.3072594	0.2127849	0.1514560	-0.4921431	-0.3891598	-0.0874351	0.4969327	-0.1521767	0.0847718
citric.acid	0.1523884	-0.1832994	-0.5905697	-0.2643008	-0.1553487	0.2276338	-0.3812850	-0.2934123	0.4026887	-0.2344633	-0.0011090
residual.sugar	0.3459199	-0.3299142	0.1646884	0.1674430	-0.3533619	-0.2334778	0.2179755	-0.5248729	-0.1080032	0.0013728	0.4497651
chlorides	-0.2901126	-0.3152580	0.0166791	-0.2447439	0.6143911	0.1609764	-0.0460682	-0.4715168	-0.2964437	0.1966302	0.0434376
free.sulfur.dioxide	0.4309140	-0.0719326	0.1342239	-0.3572789	0.2235323	-0.3400514	-0.2993632	0.2078076	-0.3666563	-0.4802433	-0.0002125

This load matrix tell us how much each variable contributes to the construction of each principal component. It can explain what aspects of the data each principal component represents. In general, the greater the absolute value of the weight, the greater the influence of the variable on the corresponding principal component.

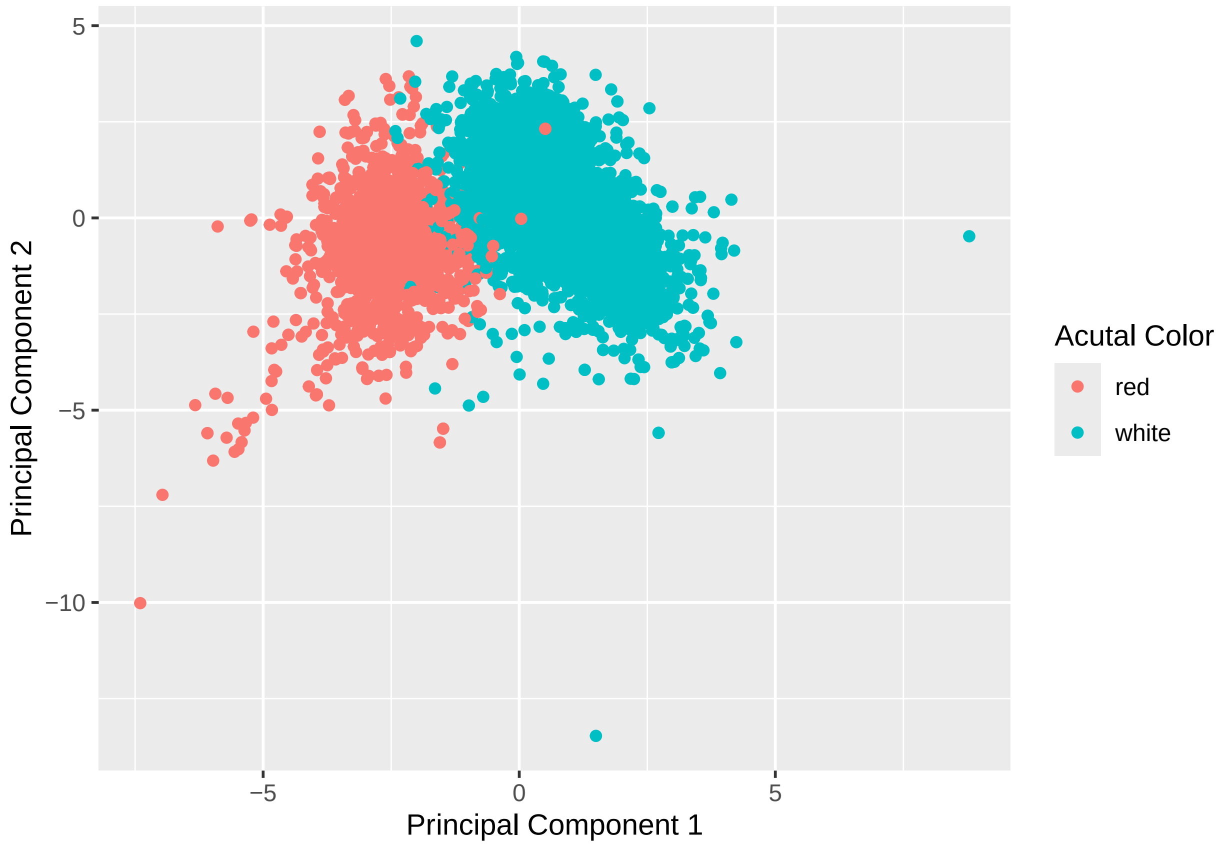
Statistical overview of the importance of principal components in PCA results

```
## 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
```

From this output, we typically focus on those points where the cumulative variance ratio approaches 1 to determine how many principal components need to be retained. In many cases, it is only when the cumulative variance ratio reaches a high value (such as 80% or 90%) that we believe we have captured most of the information in the data set. In this example, the first seven principal components already explain more than 90% of the variance in the data, so all 11 principal components may not be needed to capture most of the information in the data set.

The first two components, which have the biggest variance, seem to distinguish the color of the wine well;

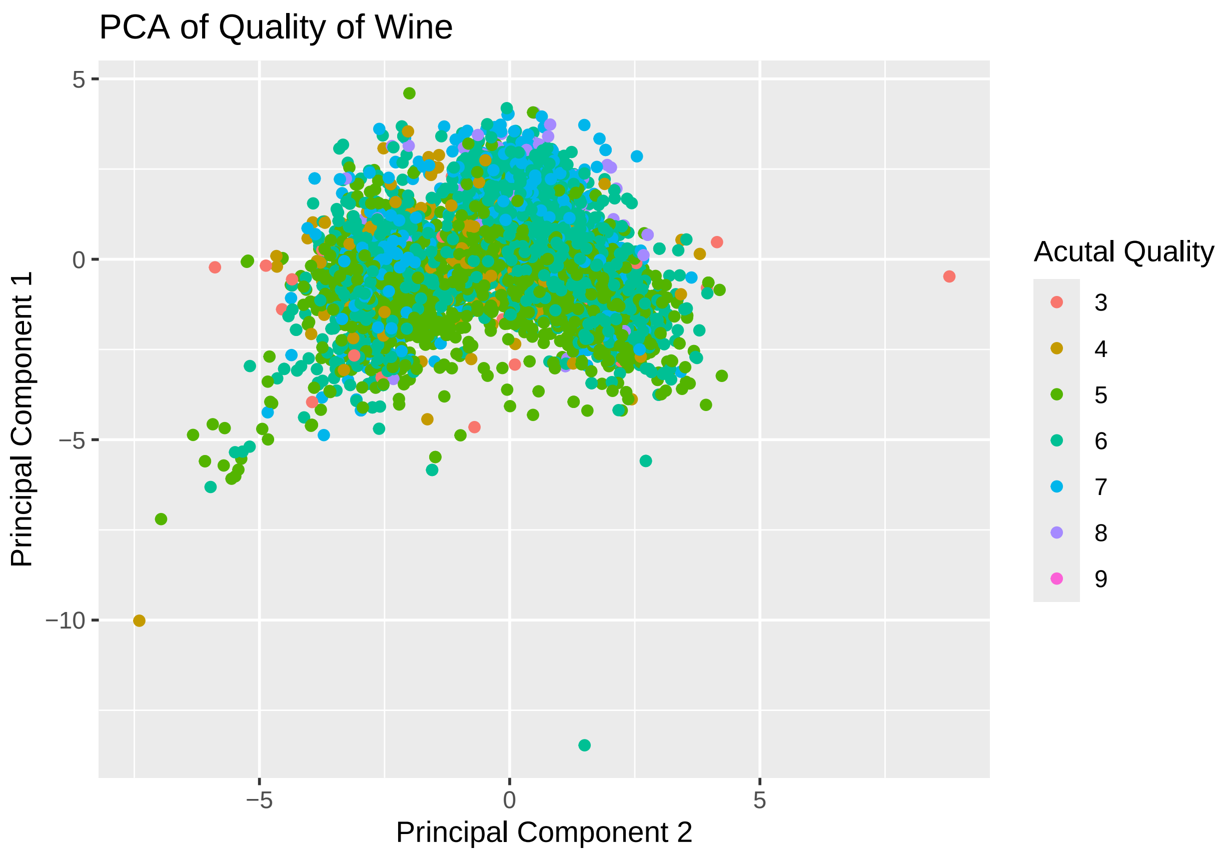
Principal component analysis (PCA) results in the first and second principal components The first two components, which have the biggest variance, seem to distinguish the color of the wine well.



We confirmed that red and white wines can be distinguished using principal component 1 (PC1) : white wines tend to have higher PC1 scores than red wines.

Quality of wines

Exploring the relationship between wine quality and principal components



When different colors are used to signify the quality of wines, the clusters overlap significantly, rendering the PCA output inconclusive. It appears that PCA does not effectively differentiate between wines of higher and lower quality.

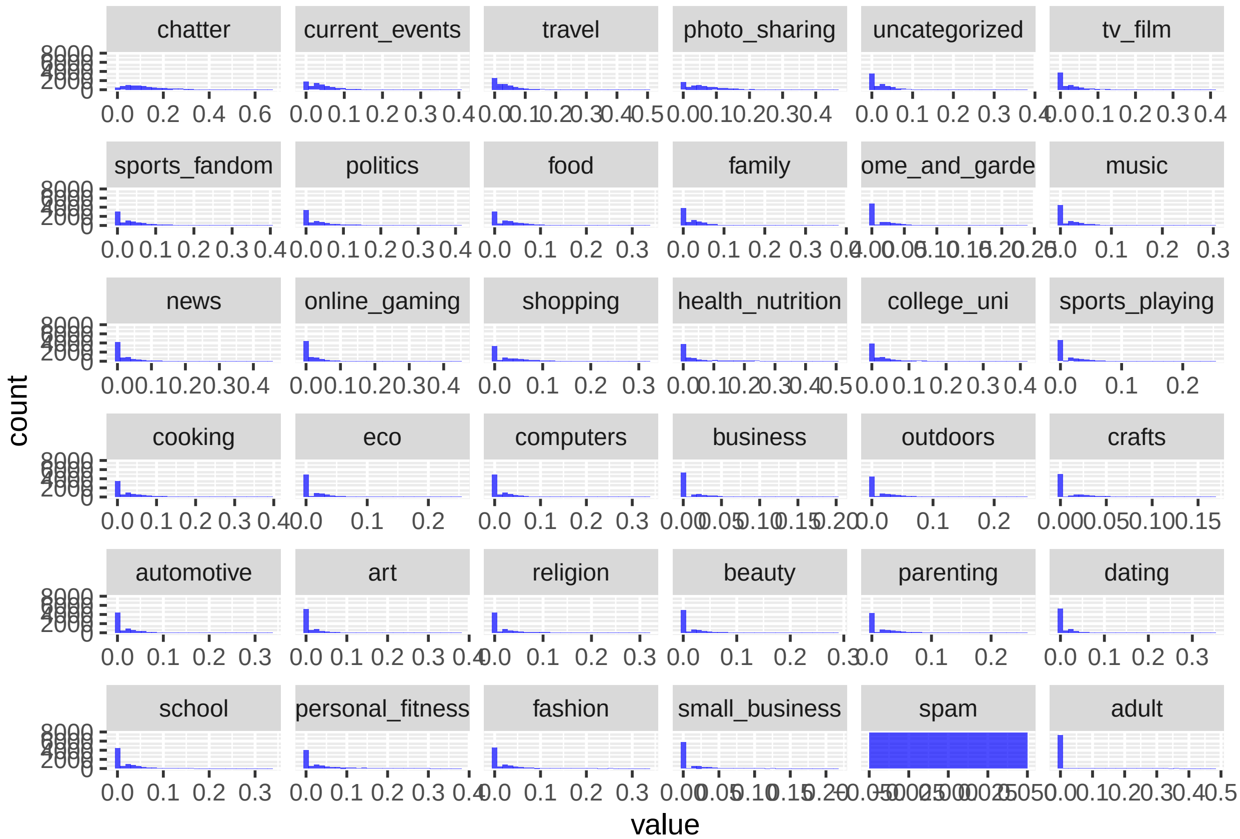
Conclusion

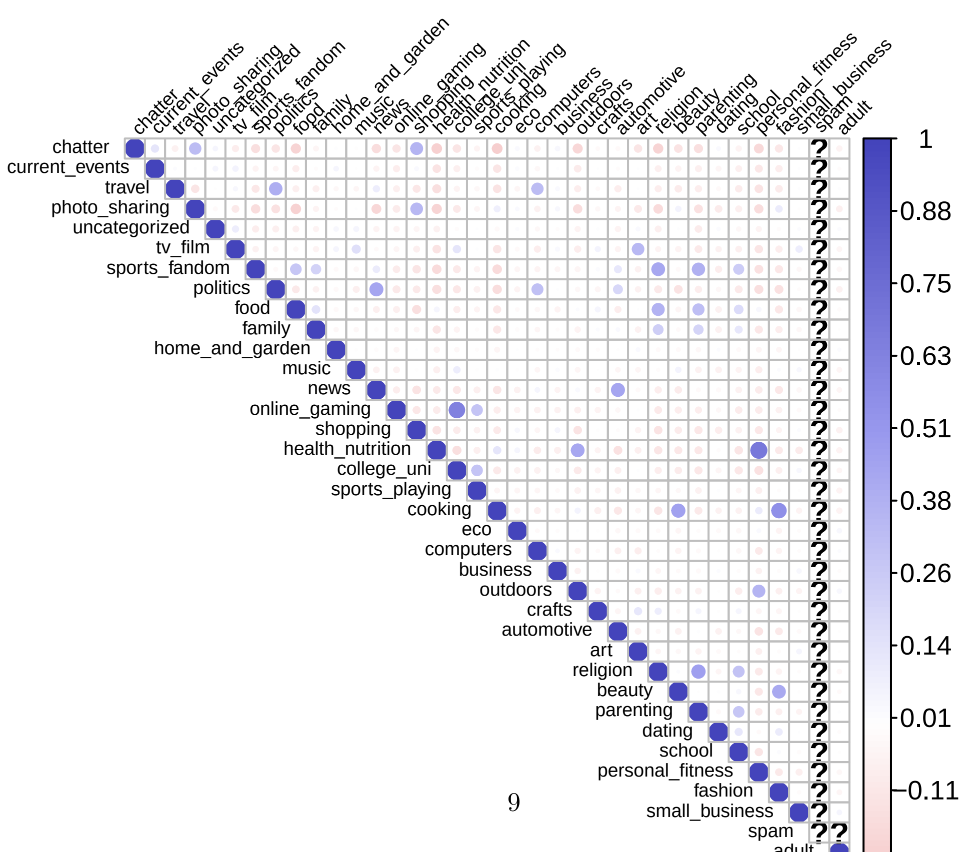
To sum up, while PCA and clustering algorithms can differentiate red from white wines, it appears that neither method is effective at discerning wines of higher quality from those of lower quality.

However, in k-means algorithm, two characteristics of density and alcohol content can be used to identify high-quality wine to a certain extent.

Question 2: Market segmentation

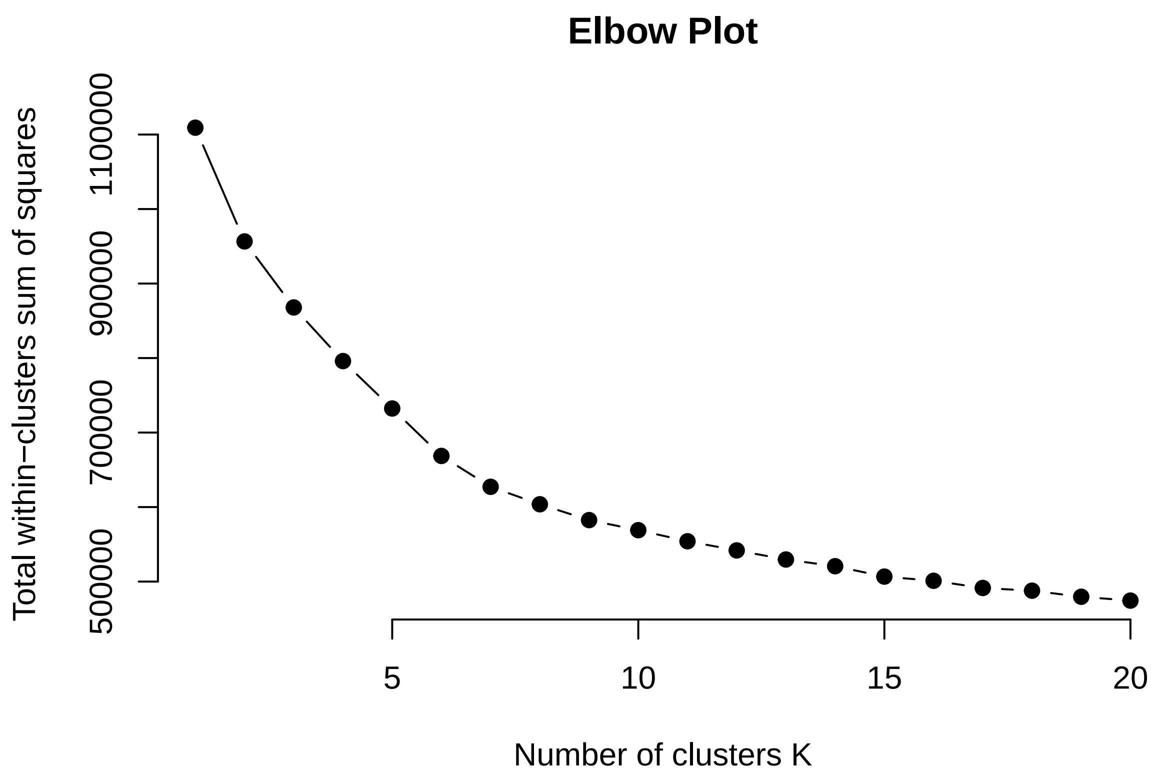
Step1: Data Preprocessing and Exploratory Data Analysis (EDA)





This graph shows how closely related different topics are based on social media data. Categories that are often talked about together, like 'home_and_garden' and 'family', show a strong positive connection with big, dark blue circles. Other topics like 'sports_playing' and 'health_nutrition' are somewhat related, with smaller, lighter blue circles. There are also topics that don't seem to be talked about together much at all; they have very light blue or no circles connecting them.

Step2: k-means and Elbow Method

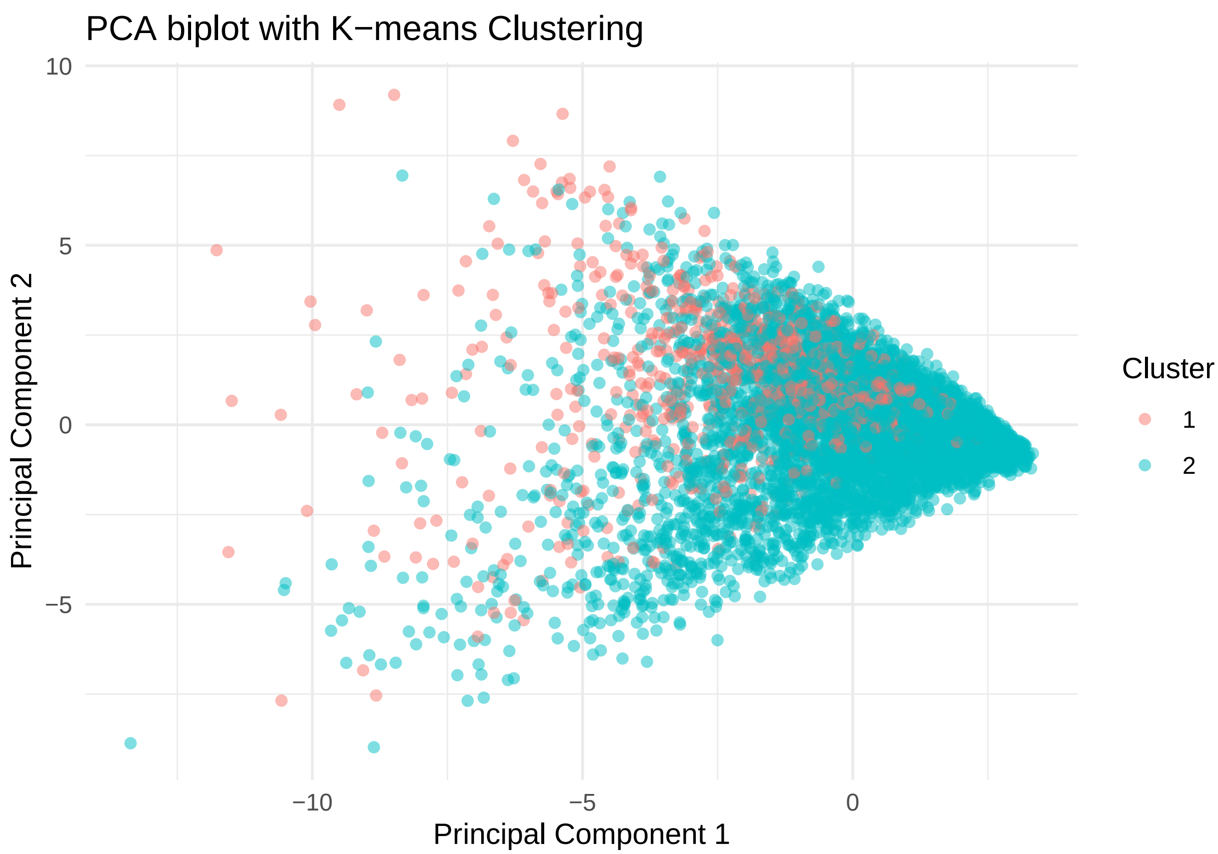


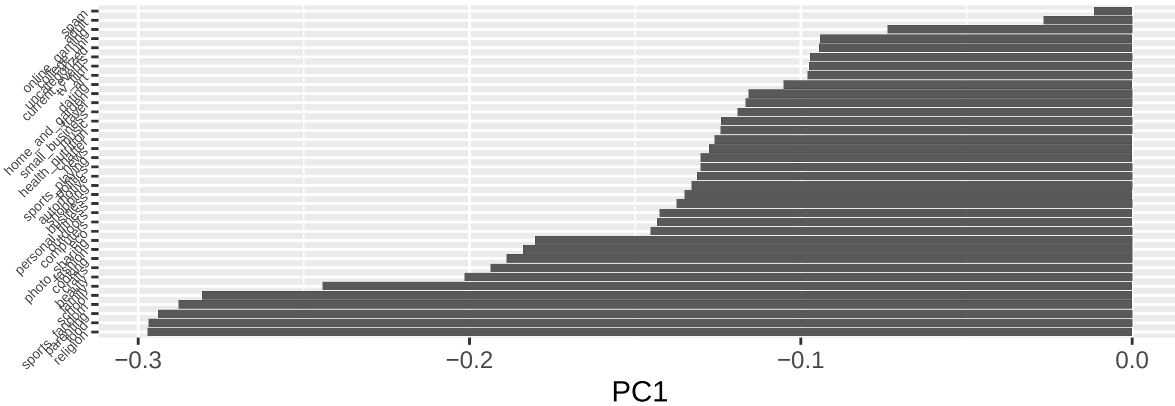

```
## [1] 2
```

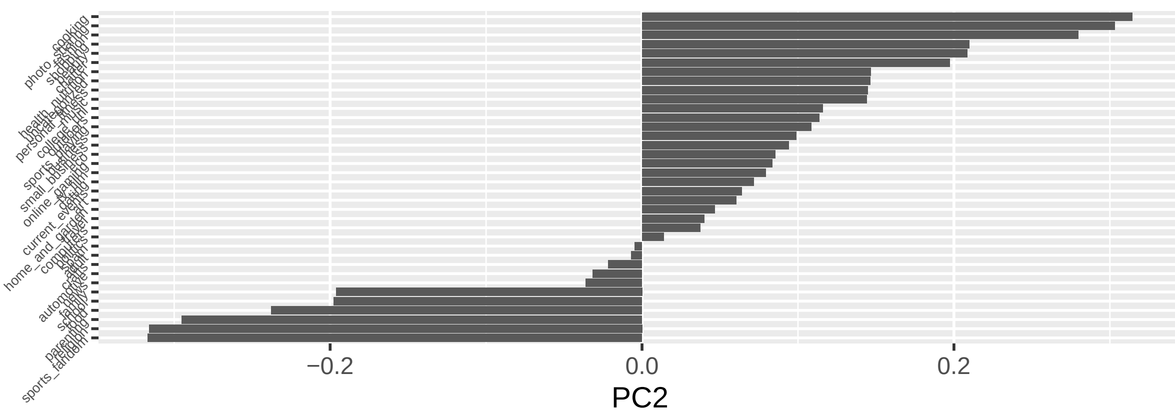
After calculating, we can find out the the optimal number of clusters is 2.

Step3: PCA_Clustering

First, we run a k-means clustering algorithm with a predetermined number of two clusters to categorize the social data into groups. We then perform Principal Component Analysis (PCA) on the same dataset to reduce its dimensions while retaining the essence of the original data.


PCA biplot with K−means Clustering
Principal Component 2
Principal Component 1
10
5
0
−5
−10
−5
0
Cluster
1
2





Discussion and Conclusion

Principal Component 1 (PC1) suggests a significant negative relationship with variables such as 'sports_fan', 'photo_sharing', and 'personal_care'. This implies a group within the audience that is less likely to be engaged with sports-related content, photo sharing, or personal care topics. This segment might represent consumers more interested in areas less correlated with an active lifestyle or personal image.

Principal Component 2 (PC2) shows a contrasting trend, where the same variables ('sports_fan', 'photo_sharing', 'personal_care') present a positive relationship. This indicates a segment of the audience that is more aligned with an active, healthy lifestyle, which may also be interested in sharing their experiences and tips online, suggesting a higher engagement with social media.

Key Segments Defined

Discreet Consumers: Fitted with PC1, this segment is less vocal on social media about personal activities and interests. They may value privacy and be interested in products without public endorsement. Marketing strategies here may focus on direct benefits instead of the need for social sharing.

Active Lifestyle Consumers: This segment fits with PC2, where interest in sports, health, and sharing content about personal care on social media is high. NutrientH20 can target this group with content and products related to fitness, health supplements, and active living.

Recommendations for NutrientH20

1.Develop different campaigns for each segment. For Active Lifestyle Consumers, focus on social media engagement, sponsorships with sports influencers, and sharing success stories. For Discreet Consumers, focus on the quality of products and privacy-respecting marketing channels.

2.Tailor the positioning of products to align with each segment's values—highlighting the social aspect and community for Active Lifestyle Consumers and emphasizing product efficacy for Discreet Consumers.

3.Create content that appeals to the varied interests within each segment. Engage the Active Lifestyle Consumers with interactive and visually appealing posts, while providing informative and detailed content for Discreet Consumers.

Question 3: Association rules for grocery purchases

We employed the arules to discover association rules that reveal the relationships between items purchased together to analyze the data from the grocery transactions. By setting the thresholds for support, confidence, and lift, we aimed to uncover meaningful patterns in shopping behavior.

Parameters and Methodology

Support: A minimum support threshold of 0.001 was chosen to ensure we capture frequent enough patterns without focusing solely on the most common items. Confidence: We set a moderate confidence level of 0.25 to strike a balance between the reliability of the rule and the inclusion of less obvious associations. Lift: we focused on rules with the highest lift values after the generation of rules to spotlight the most significant associations.

Associations Found

Alcoholic Beverages Combination: The rule involving {bottled beer, red/blush wine} => {liquor} with a lift of 35.71 and confidence of 39.58% is indicative of a strong association between these beverages. It suggests that customers who purchase beer and wine are also likely to purchase liquor, pointing towards a trend in buying multiple types of alcoholic beverages together, possibly for events or gatherings.

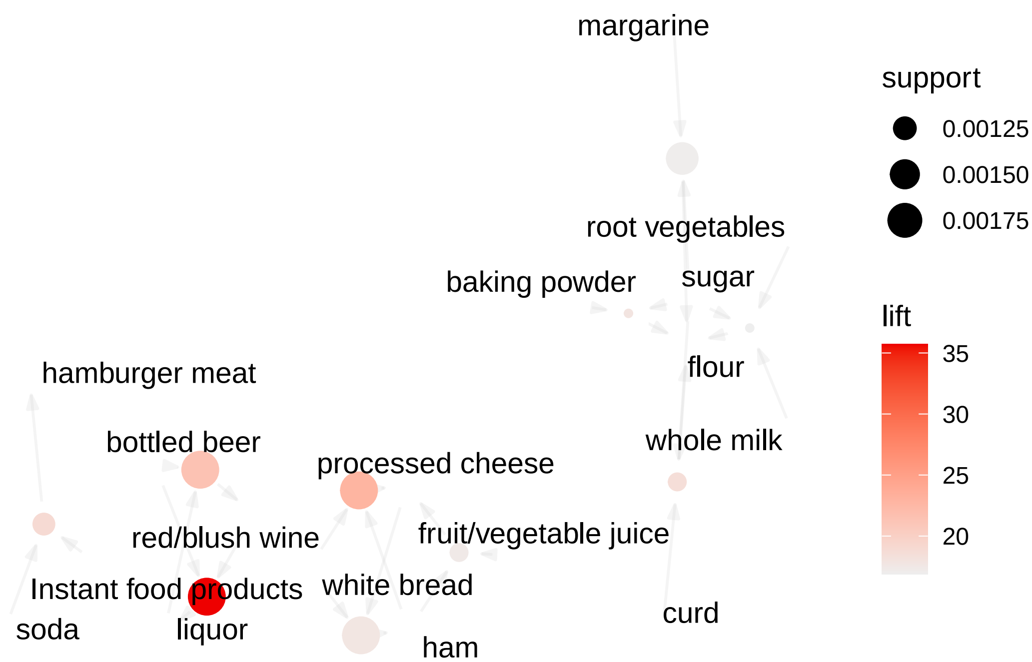
Baking Combination: Several rules indicate that customers purchasing certain baking ingredients like {curd, sugar} and {baking powder, sugar} are also likely to buy {flour}. The lifts of these rules range from 16.92 to 18.61, which could inform stores to place these items in close proximity to encourage baking-related purchases.

Meal Combination: The rule {Instant food products, soda} => {hamburger meat} with a lift of nearly 19 shows that instant food products and soda are often purchased along with hamburger meat. This could be useful for stores to bundle these items for promotions or place them near each other to increase sales.

Visual Analysis

The scatter plot visualizes the strength and reliability of these associations. The plot shows a clustering of rules with higher lift values at the lower levels of support, which is expected since high-lift rules are often less common.

The two-key plot shows that as the size of the itemset increases, the support generally decreases, which is again typical in market basket analysis.


margarine
support
0.00125
0.00150
0.00175
root vegetables
baking powder
sugar
lift
35
30
25
20
flour
hamburger meat
whole milk
bottled beer
processed cheese
red/blush wine
fruit/vegetable juice
Instant food products
white bread
curd
soda
liquor
ham

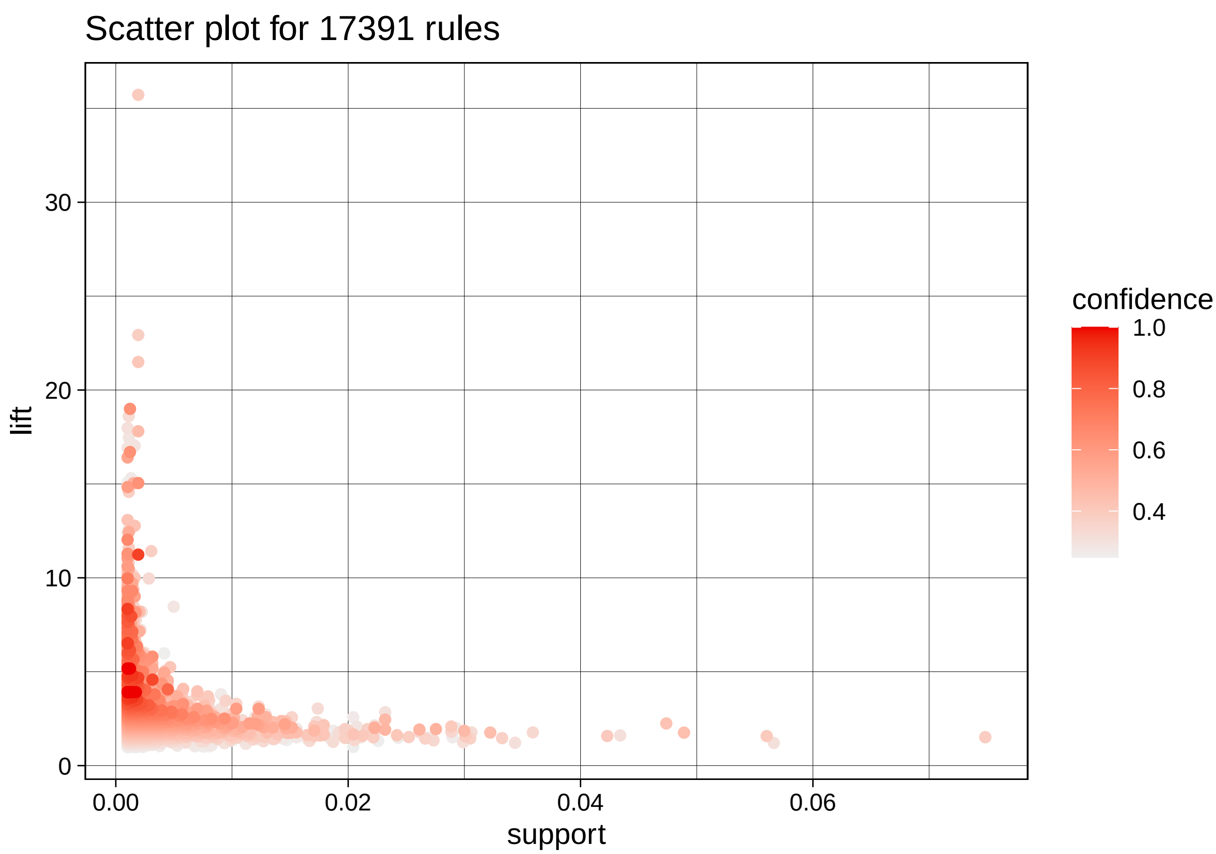

Scatter plot for 17391 rules
lift
30
20
10
0
0.00
0.02
0.04
0.06
support
confidence
1.0
0.8
0.6
0.4

Table 2: Top 10 Rules by Lift

	rules	support	confidence	coverage	lift	count
380	{bottled beer,red/blush wine} => {liquor}	0.0019319	0.3958333	0.0048805	35.71579	19
1147	{ham,white bread} => {processed cheese}	0.0019319	0.3800000	0.0050839	22.92822	19
379	{bottled beer,liquor} => {red/blush wine}	0.0019319	0.4130435	0.0046772	21.49356	19
442	{Instant food products,soda} => {hamburger meat}	0.0012201	0.6315789	0.0019319	18.99565	12
1585	{curd,sugar} => {flour}	0.0011185	0.3235294	0.0034570	18.60767	11
1497	{baking powder,sugar} => {flour}	0.0010168	0.3125000	0.0032537	17.97332	10
1146	{processed cheese,white bread} => {ham}	0.0019319	0.4634146	0.0041688	17.80345	19
1150	{fruit/vegetable juice,ham} => {processed cheese}	0.0011185	0.2894737	0.0038638	17.46610	11
1588	{margarine,sugar} => {flour}	0.0016268	0.2962963	0.0054906	17.04137	16
7495	{root vegetables,sugar,whole milk} => {flour}	0.0010168	0.2941176	0.0034570	16.91606	10

Conclusions

The identified rules make sense in the context of shopping behavior. Items that are frequently purchased together, such as combinations of alcoholic beverages or components for baking, can be targeted for cross-promotions or placed together in-store to increase basket size. The rules with high lift, especially those involving complementary items, validate the utility of the association rule mining in uncovering shopping basket patterns.