ECO395M_Exercise4

Youngseok Yim (EID: yy9739)

2023-04-17

1. Clustering and PCA

First, the data is cleaned by centering and scaling.

I utilized k-means for clustering with k=2, as there were two wine varieties by color: red and white, with 25 observations. To verify the efficacy of k-means in separating the data points into the correct wine color groups, I compared the chemical property averages of the original white and red wine data with those in the clustered data.

```
## # A tibble: 2 x 13
##   color fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
##   <chr>         <dbl>            <dbl>       <dbl>          <dbl>     <dbl>
## 1 red            8.32            0.528       0.271           2.54    0.0875
## 2 white          6.85            0.278       0.334           6.39    0.0458
##   free.sulfur.dioxide total.sulfur.dioxide density    pH sulphates alcohol
##                 <dbl>                <dbl>   <dbl> <dbl>     <dbl>   <dbl>
## 1                15.9                 46.5   0.997  3.31     0.658    10.4
## 2                35.3                138.    0.994  3.19     0.490    10.5
##   quality
##     <dbl>
## 1    5.64
## 2    5.88
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##            8.2895922            0.5319416            0.2695435
##       residual.sugar            chlorides  free.sulfur.dioxide
##            2.6342666            0.0883238           15.7647596
## total.sulfur.dioxide              density                   pH
##           48.6396835            0.9967404            3.3097200
##            sulphates              alcohol
##            0.6567194           10.4015216
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##           6.85167903           0.27458385           0.33524928
##       residual.sugar            chlorides  free.sulfur.dioxide
##           6.39402555           0.04510424          35.52152864
## total.sulfur.dioxide              density                   pH
##         138.45848785           0.99400486           3.18762464
##            sulphates              alcohol
##           0.48880511          10.52235888
```

By comparing the chemical property averages of red and white wine in both the original and clustered data, it is evident that k-means effectively separates red and white wines. The averages of chemical properties are almost the same in both the original and k-means clustered data for red wine, as well as for white wine.

To validate this, I also created a confusion matrix. The results show that k-means accurately clustered the wine data by color, with an accuracy rate of 98.6%. This confirms that k-means clustering achieved excellent dimension reduction in this instance.

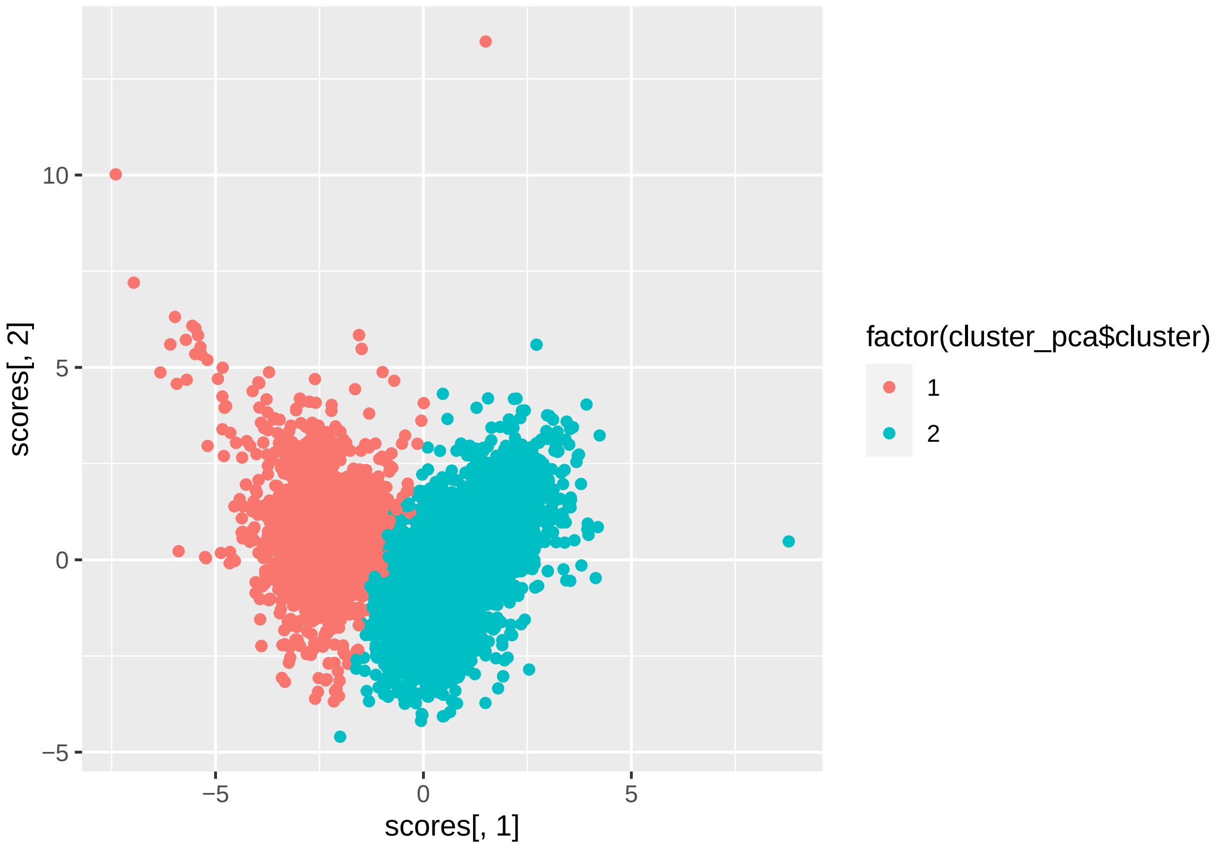
```
##
##         red_hat white_hat
##   red      1575        24
##   white      68      4830
```

```
## [1] 0.9858396
```

After implementing k-means, I moved on to Principal Component Analysis (PCA). The below table shows that the first three principal components account for 64.4% of the total variance in the data set, which is a significant amount. Consequently, I utilized the first three components for clustering.

```
## 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
```

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




```
##        cluster
## color   red_hat white_hat
##   red      1575        24
##   white      82      4816
```

```
## [1] 0.9836848
```

The clustering based on the scores of the three principal components also showed good results with an accuracy of 98.4%. However, PCA is not as straightforward as k-means. In this case, I utilized the scores from the principal components to form clusters. Given the higher accuracy of k-means and its straightforwardness, it is more practical to use k-means for this data set.

The wine quality was rated on a scale of 1 to 10 in the data set, with the absence of ratings 1, 2, and 10. As a result, the wine in the data set was rated between 2 and 9. I applied k-means with k=7 and 25 observations.

```
##             cluster2$cluster
## wine$quality   1   2   3   4   5   6   7
##            3   5   7   6   2   4   4   2
##            4  65  24  64   2  14  21  26
##            5 446 652 479  30 183  79 269
##            6 549 640 346  19 259 551 472
##            7 137 122  43   2 138 446 191
##            8  27  22   4   0  12  98  30
##            9   1   0   0   0   0   4   0
```

The confusion matrix reveals that k-means clustering failed to differentiate between the various wine quality ratings. For instance, all of the clusters have a substantial number of wines rated 5, 6, and 7, lacking clear differentiation.

2. Market Segmentation

I began by cleaning the dataset, which originally had 7,882 data points and 36 variables.

To eliminate spam and pornographic content, I filtered out all users whose tweets were classified as "spam" or "adult". I then removed the "spam" and "adult" variables from the dataset. Since they did not offer any valuable insights, I also excluded the "uncategorized" and "chatter" variables. The final dataset consisted of 7,309 data points and 32 variables.

In order to identify market segments, I employed cluster analysis. Since the data lacked any hierarchical structure, I chose to use K-means clustering. I utilized the K-means++ algorithm for this analysis.

To determine the optimal number of clusters (K) for the analysis, I utilized both the Elbow plot and CH index methods.

Figure 2.1 Elbow Plot

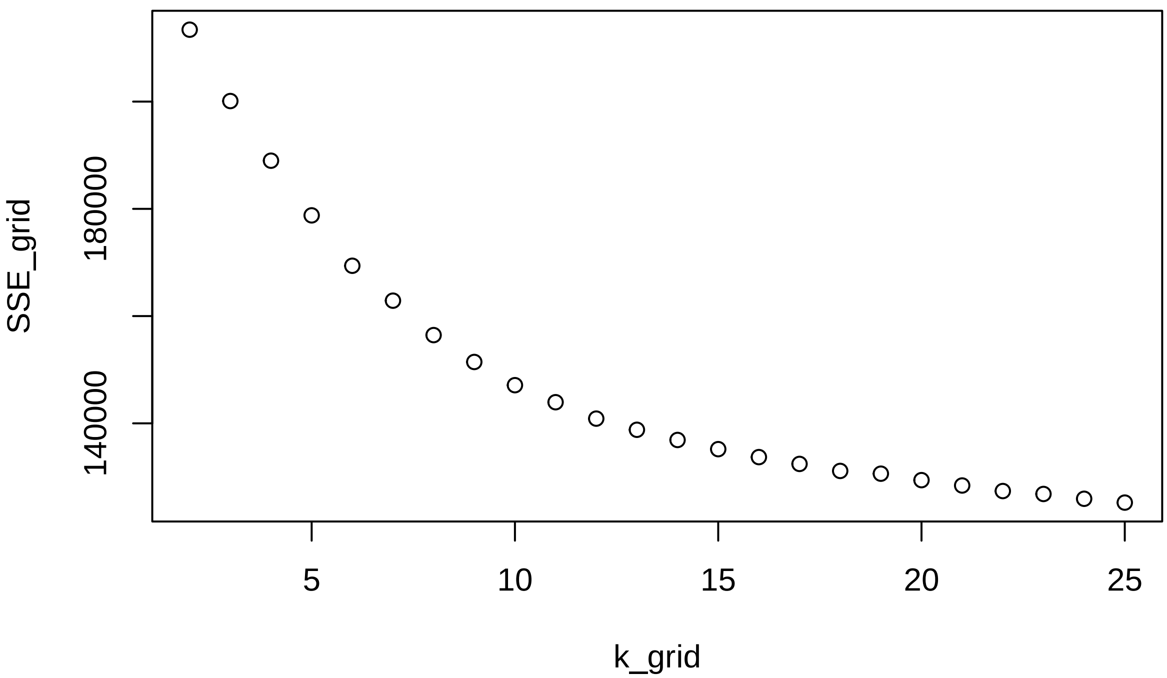
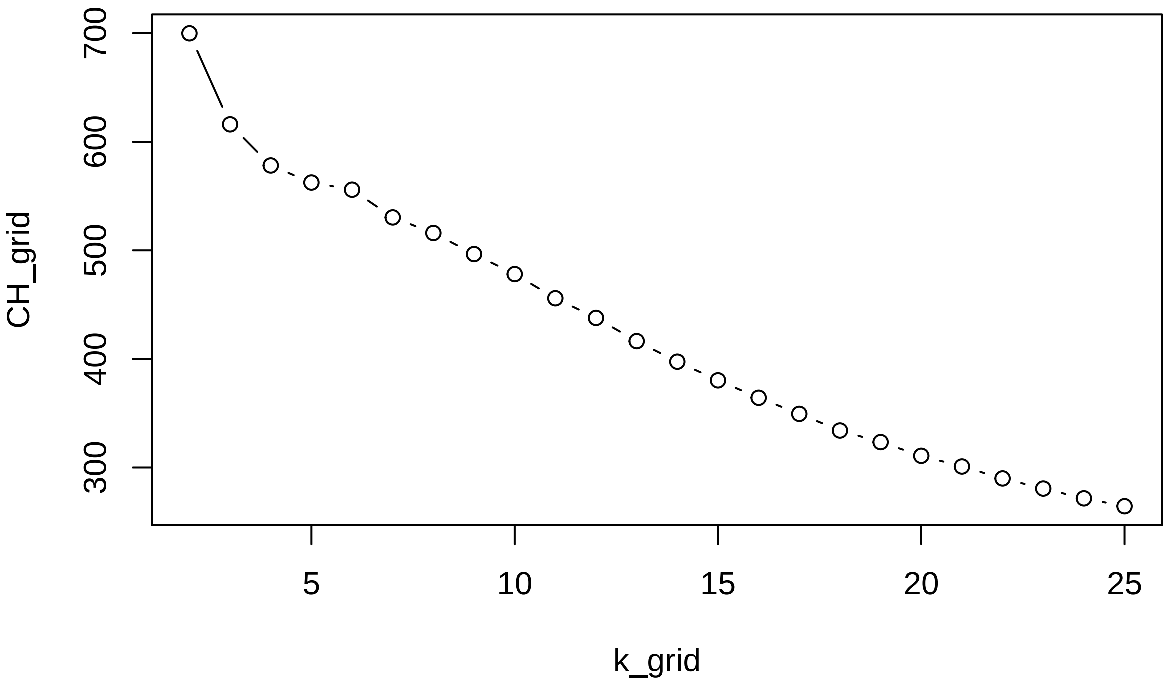


Figure 2.2 CH index plot



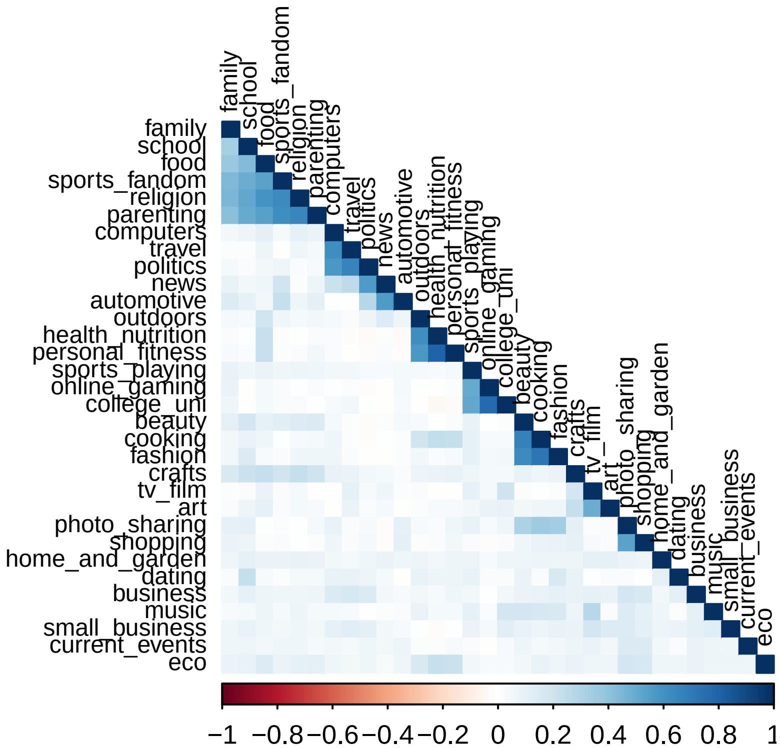
The optimal value of K is not immediately obvious from the graph. However, the plots suggest that K=5 may be a potential candidate. To validate this, I have also plotted a correlogram to identify any singularities among the variables.

Figure 2.3 Correlogram

```
## Warning in text.default(pos.xlabel[, 1], pos.xlabel[, 2], newcolnames, srt =
## tl.srt, : "nstart" is not a graphical parameter
```

```
## Warning in text.default(pos.ylabel[, 1], pos.ylabel[, 2], newrownames, col =
## tl.col, : "nstart" is not a graphical parameter
```

```
## Warning in title(title, ...): "nstart" is not a graphical parameter
```



The correlogram reveals the existence of subgroups of variables that exhibit high levels of correlation. The variables 'family', 'school', 'food', 'sports_fandom', and 'religion' appear to have a strong relationship. Similarly, 'computers', 'travel', 'politics', 'news', and 'automotive' show correlations. There is also a correlation among 'outdoors', 'health_nutrition', and 'personal_fitness'. 'Sports_playing', 'online_gaming', and 'college_uni' also display a relationship. Lastly, 'beauty', 'cooking', and 'fashion' seem to have a substantial correlation. Thus, the correlogram supports the finding that the optimal value of K is 5.

Summary of cluster 1

```
##    Mode   FALSE    TRUE
## logical    5268    1229
```

Summary of Cluster 2

```
##    Mode   FALSE    TRUE
## logical    5027    1470
```

Summary of Cluster 3

```
##    Mode   FALSE    TRUE
## logical    5561     936
```

Summary of Cluster 4

```
##    Mode   FALSE    TRUE
## logical    6464      33
```

Summary of Cluster 5

```
##    Mode   FALSE    TRUE
## logical    5858     639
```

What are the clusters?

```
##    current_events            travel    photo_sharing           tv_film
##       1.638853082       0.553979552      3.367347083       0.641621115
##     sports_fandom          politics             food            family
##       1.143536925       0.987436873      1.527047251       0.383900184
##   home_and_garden             music             news    online_gaming
##      -0.014126594       0.160561810      1.020497302       1.451943420
##          shopping health_nutrition      college_uni    sports_playing
##       0.597289743       3.650767039      0.795031758       0.440915459
##           cooking               eco        computers          business
##       1.084307424       0.560779197      0.152883574      -0.075513607
##          outdoors            crafts       automotive               art
##       0.160888645       0.435600277      0.945520110       0.000136094
##          religion            beauty        parenting            dating
##       1.547134719       0.353718541      0.594941871       0.227395720
##            school  personal_fitness          fashion    small_business
##       0.852433877       0.450428657     -0.328127674       0.018418911
```

```
##    current_events            travel    photo_sharing           tv_film
##        1.30163770        0.75772402       3.57404531        3.52388670
##     sports_fandom          politics             food            family
##        1.26879613        4.69529578       3.16045578        1.88876785
##   home_and_garden             music             news    online_gaming
##        0.14843616        0.39799137      -0.65082054        0.73791489
##          shopping health_nutrition      college_uni    sports_playing
##        0.75812386        3.99006039       5.84566132        0.49742454
##           cooking               eco        computers          business
##        5.24143277        1.25981243       1.72575546        0.07840710
##          outdoors            crafts       automotive               art
##        0.43508861       -0.20727223       0.58121860        0.14258331
##          religion            beauty        parenting            dating
##        1.69084217        2.65572509       0.68249975        2.40187886
##            school  personal_fitness          fashion    small_business
##        1.94503415        3.68349270       0.08066392        0.15764269
```

```
##    current_events            travel    photo_sharing           tv_film
##        1.61861591        5.49224653      -0.80777967        0.02779664
##     sports_fandom          politics             food            family
##        3.07945304       -0.63058544      -0.68550673        1.40953463
##   home_and_garden             music             news    online_gaming
##        1.21986471        1.09046282       0.68576804        1.40909271
##          shopping health_nutrition      college_uni    sports_playing
##        4.49551489       -3.20266726      -0.28058045        1.31427912
##           cooking               eco        computers          business
##       -0.72976418       -0.38610502       1.22486593        1.09287711
##          outdoors            crafts       automotive               art
##        1.23480220        0.30684665       0.92365743        3.45270794
##          religion            beauty        parenting            dating
##       -1.35563244       -0.13653051       1.94817350       -0.72149752
##            school  personal_fitness          fashion    small_business
##       -0.62756151        2.65390602       2.77495253        0.56369361
```

```
##    current_events            travel    photo_sharing           tv_film
##        2.46727803        4.00187403       6.05958143        0.29994333
##     sports_fandom          politics             food            family
```

```
##      20.70441373      -0.28458971       0.15931462       1.70668663
##  home_and_garden            music             news    online_gaming
##      -0.09996170       4.30042849      -0.44588411       3.20607171
##         shopping health_nutrition      college_uni   sports_playing
##       3.31902775       8.07004633       0.19630555       9.26555352
##          cooking              eco        computers         business
##      -0.34184553      -0.02455543       1.53544608      -0.15678560
##         outdoors           crafts       automotive              art
##       4.94189105      -0.12844342       1.84048658       2.41071929
##         religion           beauty        parenting           dating
##       3.41891727       0.08082748      14.26826011      -0.51858745
##           school personal_fitness          fashion   small_business
##      -0.06245363       3.29230864      -0.54397458       2.44595103
```

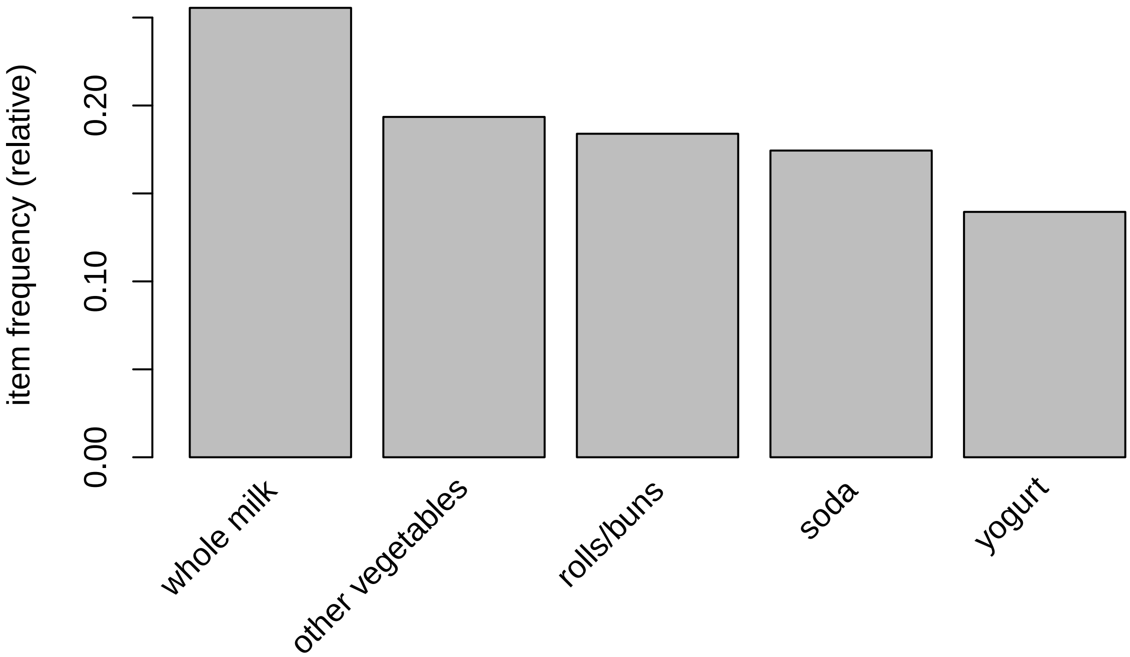
```
##   current_events           travel    photo_sharing          tv_film
##       4.07403569       2.63669724       5.22293746       0.13906739
##    sports_fandom         politics             food           family
##       3.41419182      -0.90891582      -0.86974257       1.94445068
##  home_and_garden            music             news    online_gaming
##       0.51399846       2.03680847       1.45205996       6.60826170
##         shopping health_nutrition      college_uni   sports_playing
##       2.24136903       6.69670253      -0.08559933       1.46529222
##          cooking              eco        computers         business
##      -1.04181874      -0.46495056       1.78395457       0.42453563
##         outdoors           crafts       automotive              art
##       2.32770716       0.60160088       3.57630374       1.45626078
##         religion           beauty        parenting           dating
##       2.83723850      -0.04766084       2.18216044      -0.88472507
##           school personal_fitness          fashion   small_business
##      -0.75079837       3.80312192       0.99993224       1.11861111
```

After conducting a K-means++ clustering analysis with K=5, I evaluated the distribution of data points among the clusters. The cluster with the largest number of data points, accounting for approximately 60% of the total sample, consisted of individuals who had tweeted an average of less than 2 times across all categories. This could indicate that most followers of "NutrientH20" are inactive on Twitter or social media platforms. Despite their inactivity, they continue to follow "NutrientH20", suggesting that the company's current social media marketing strategy is effective.

The cluster with the smallest number of data points, on the other hand, comprised individuals who tweeted more frequently about topics such as photo sharing, cooking, and fashion. To reach and appeal to these individuals, who have a higher interest in these topics, the company should position their brand as relevant to photo sharing, cooking, or fashion.

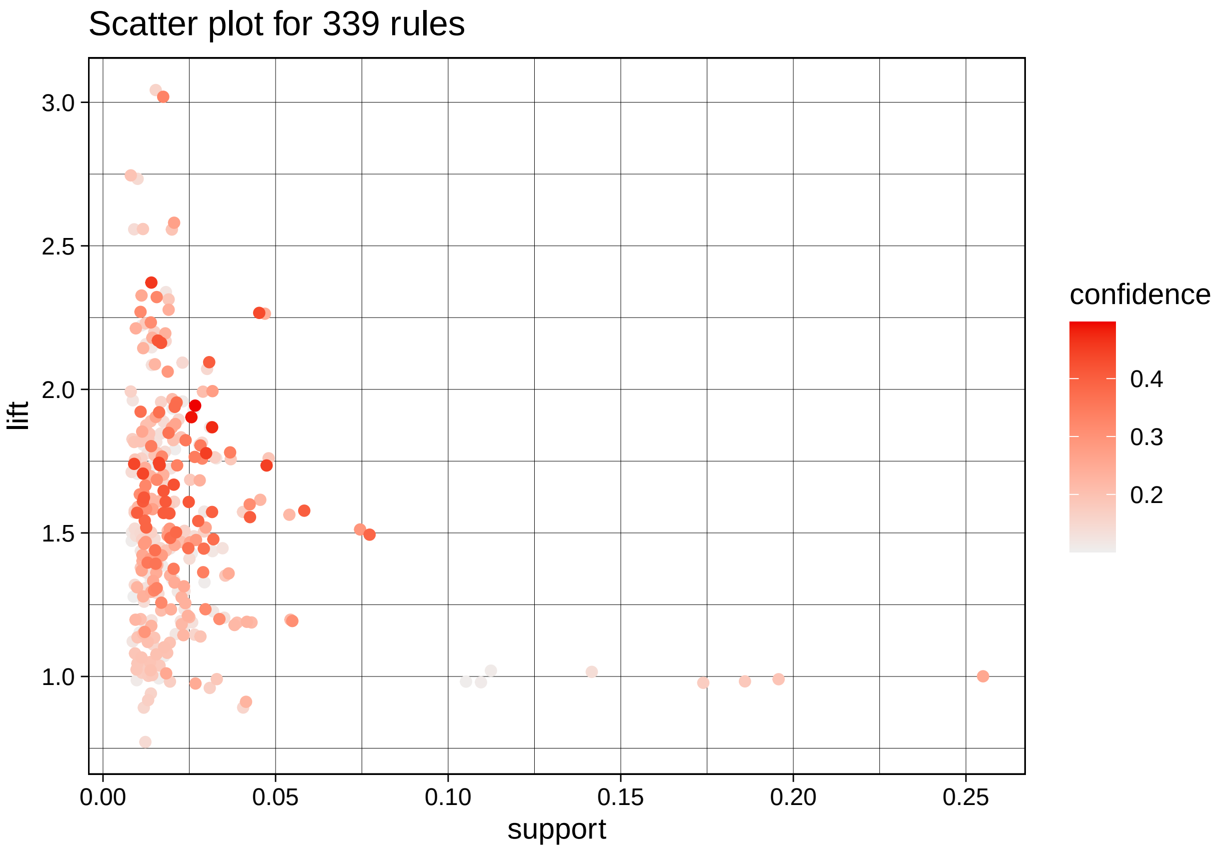
3. Association rules for grocery purchases

Figure 3.1 Top 5 items with highest support



I utilized the 'apriori' function to identify various association rules with a support of 0.01, confidence of 0.1, and a maximum length of 2, resulting in a set of 339 rules. Upon examining the items with the highest support, the top 5 items identified were whole milk, other vegetables, rolls/buns, soda, and yogurt, as indicated in the accompanying figure.

Figure 3.2 Plot of Association rules



To uncover strong associations, I applied a threshold of 0.3 for confidence and 2 for lift to the association rules generated from the apriori function with support set at 0.01 and a maximum length of 2. The result was a subset of 9 association rules with high lift and confidence. The threshold selection was based on the visualization of the association rule plot, where the majority of the points did not exceed a confidence of 0.3 or a lift of 2.

```
##     lhs                     rhs                support    confidence coverage
## [1] {onions}             => {other vegetables} 0.01423488 0.4590164  0.03101169
## [2] {berries}            => {yogurt}           0.01057448 0.3180428  0.03324860
## [3] {hamburger meat}     => {other vegetables} 0.01382816 0.4159021  0.03324860
## [4] {cream cheese}       => {yogurt}           0.01240468 0.3128205  0.03965430
## [5] {chicken}            => {other vegetables} 0.01789527 0.4170616  0.04290798
## [6] {beef}               => {root vegetables}  0.01738688 0.3313953  0.05246568
## [7] {curd}               => {yogurt}           0.01728521 0.3244275  0.05327911
## [8] {whipped/sour cream} => {other vegetables} 0.02887646 0.4028369  0.07168277
## [9] {root vegetables}    => {other vegetables} 0.04738180 0.4347015  0.10899847
##     lift     count
## [1] 2.372268 140
## [2] 2.279848 104
## [3] 2.149447 136
## [4] 2.242412 122
## [5] 2.155439 176
## [6] 3.040367 171
## [7] 2.325615 170
## [8] 2.081924 284
## [9] 2.246605 466
```

The results of the association rule analysis show that the majority of the rules make logical sense. The first rule in the table highlights that the presence of onions implies the presence of other vegetables, which is a common combination. Additionally, the association between beef and root vegetables, as well as hamburger meat and other vegetables, are also plausible.