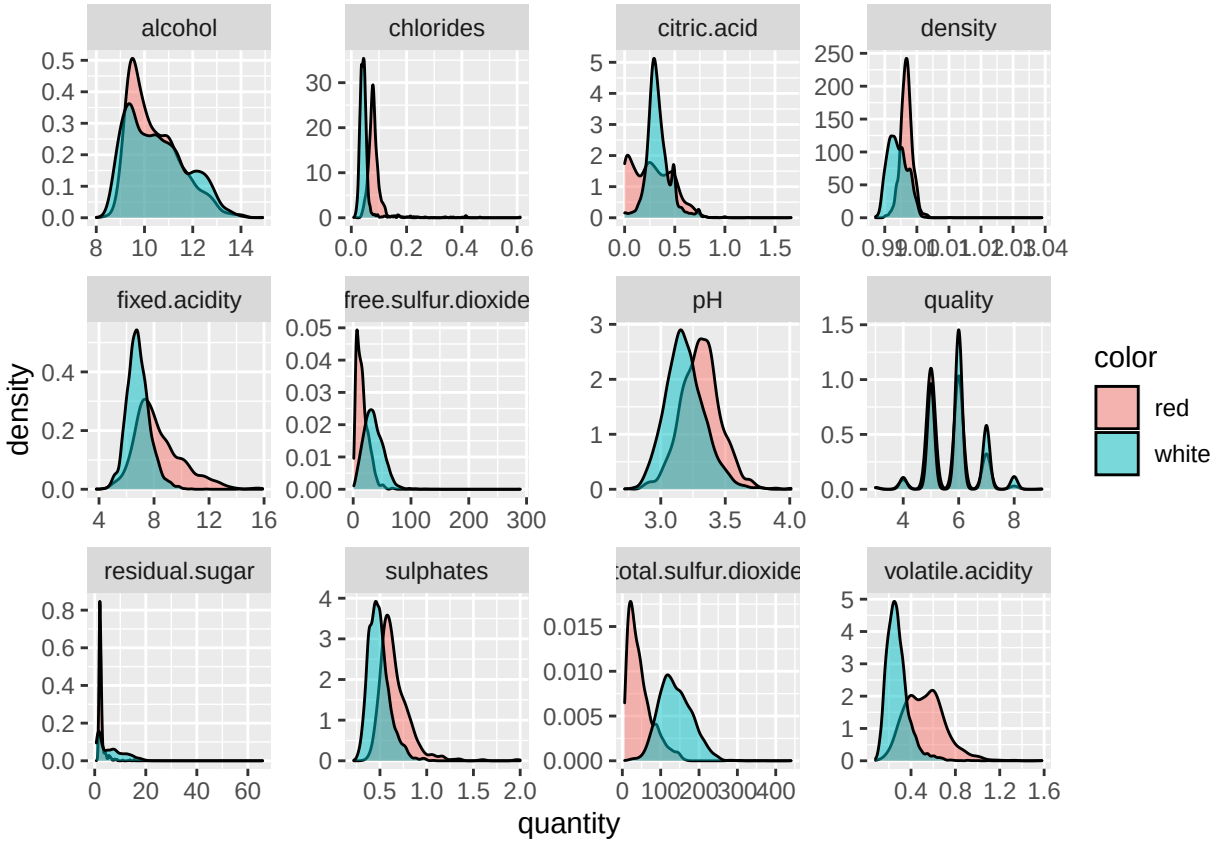
Exercises #4: Data Mining and Statistical Learning

#1) Clustering and PCA with Wine Data

We're asked to compare the results between clustering and PCA using a dataset on the chemical composition of over 6,000 wines.

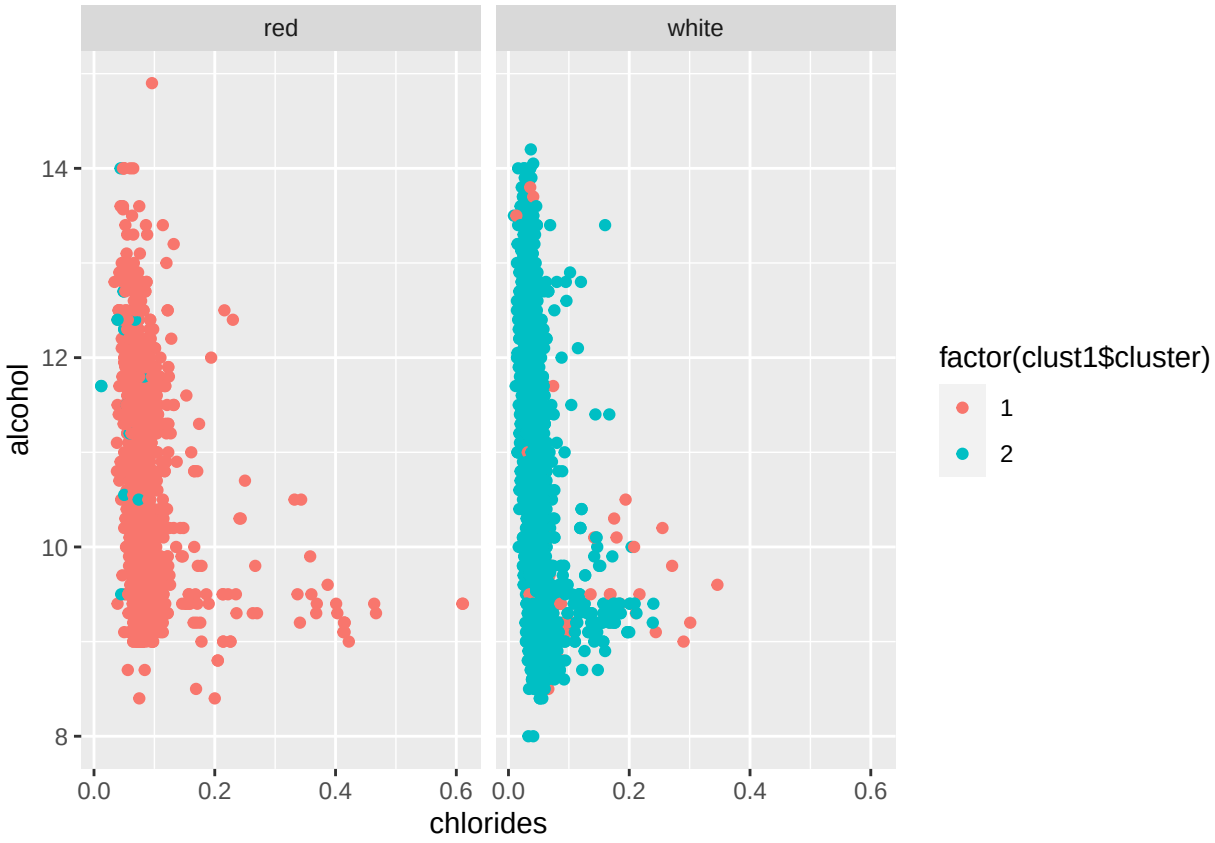


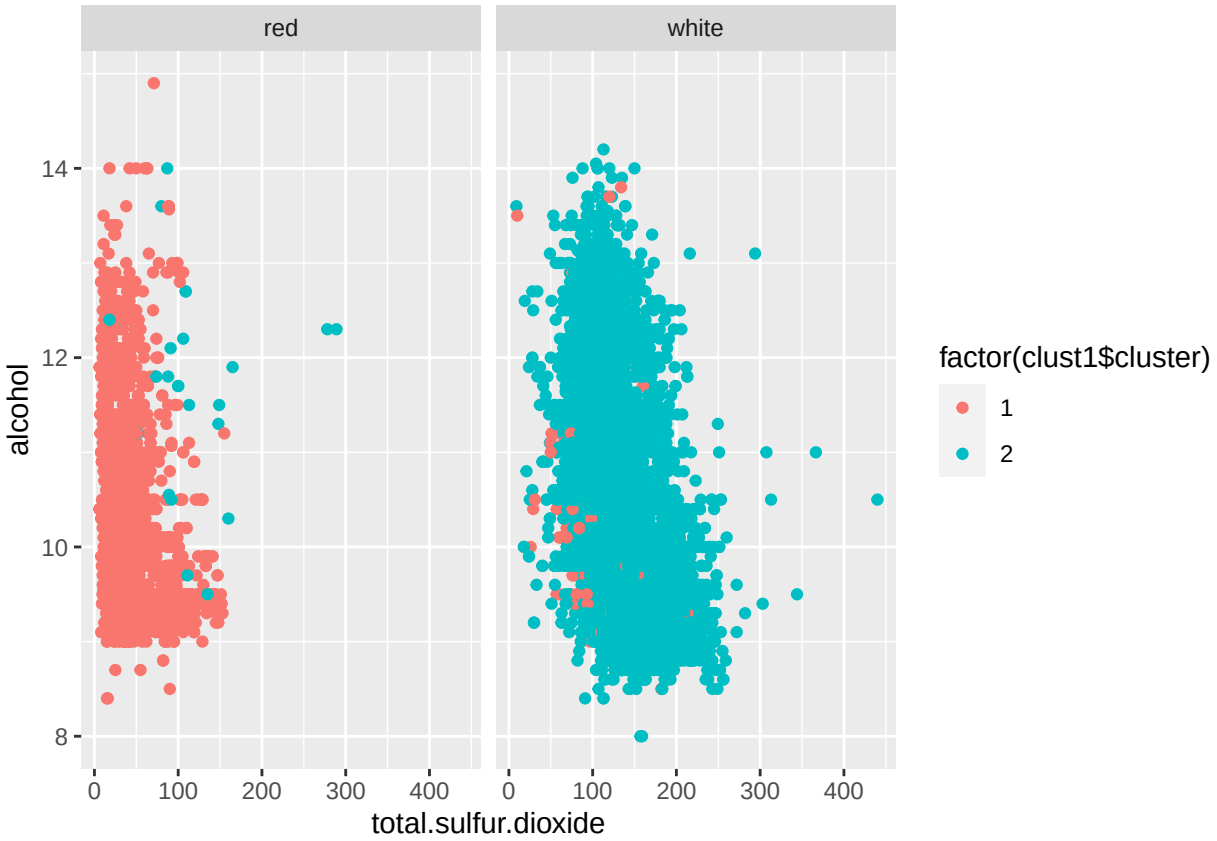
From what we can see, there are some similarities in the chemical composition between the two types of wine. However, there are also some differences in the chemical composition. White Wine tends to contain less chlorides than red wine, but has a higher quantity of total sulfur dioxide. What is interesting is that the pH levels between the two wines are not that distinct; The way that people normally talk about would suggest that red wine is significantly more acidic than white wine, but the explanatory analysis of these wines suggest otherwise.

To start our analysis, we will first see if we are able to distinguish the reds from the white using the k-means clustering method and comparing it to the principal component analysis method.

Clustering

First, we drop the 13th column, as it is a non-numeric column, and we want to re-scale the data. We then perform a k-means clustering on our data where we create two clusters to represent each wine.


red
white
alcohol
14
12
10
8
0.0
0.2
0.4
0.6
chlorides
factor(clust1$cluster)
1
2



As we can see, the two chemicals that might be best to use to differentiate red and white wine are the quantity of chlorides and total sulfur dioxide. Looking at the above figures, it would seem that regular clustering method works relatively well as it correctly clustered most of wines according to the color.

Principal Component Analysis

Now, we'll use PCA to see if it can differentiate between different types of wines given our data. Before our analysis, we drop columns 12 and 13 because we want this to be an unsupervised task, using only the eleven chemical components.

```
##                              PC1         PC2         PC3         PC4         PC5
## fixed.acidity         -0.23879890  0.33635454 -0.43430130  0.16434621 -0.1474804
## volatile.acidity      -0.38075750  0.11754972  0.30725942  0.21278489  0.1514560
## citric.acid            0.15238844  0.18329940 -0.59056967 -0.26430031 -0.1553487
## residual.sugar         0.34591993  0.32991418  0.16468843  0.16744301 -0.3533619
## chlorides             -0.29011259  0.31525799  0.01667910 -0.24474386  0.6143911
## free.sulfur.dioxide    0.43091401  0.07193260  0.13422395 -0.35727894  0.2235323
## total.sulfur.dioxide   0.48741806  0.08726628  0.10746230 -0.20842014  0.1581336
## density               -0.04493664  0.58403734  0.17560555  0.07272496 -0.3065613
## pH                    -0.21868644 -0.15586900  0.45532412 -0.41455110 -0.4533764
## sulphates             -0.29413517  0.19171577 -0.07004248 -0.64053571 -0.1365769
## alcohol               -0.10643712 -0.46505769 -0.26110053 -0.10680270 -0.1888920
##                              PC6         PC7
## fixed.acidity         -0.20455371 -0.28307944
## volatile.acidity      -0.49214307 -0.38915976
## citric.acid            0.22763380 -0.38128504
```

```
## residual.sugar       -0.23347775  0.21797554
## chlorides             0.16097639 -0.04606816
## free.sulfur.dioxide  -0.34005140 -0.29936325
## total.sulfur.dioxide -0.15127722 -0.13891032
## density               0.01874307 -0.04675897
## pH                    0.29657890 -0.41890702
## sulphates            -0.29692579  0.52534311
## alcohol              -0.51837780 -0.10410343
```

```
## Importance of first k=7 (out of 11) 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
```

We can see that 7 principal components seem to explain about 90% of the variation in our data. However, it's not clear whether these components represent white or red wines.

In the end, we defer to clustering for classifying red and white wines using only their chemical properties.

2) Identifying Market Segments: NutrientH20's Twitter Data

We're given data set made up of the total number of times that a particular user (row) tweeted about a particular subject (column) over a week-long period in June of 2014, and are asked to identify any market segments that stand out. There are over 30 topics listed, and include chatter, current events, travel, photo sharing, tv/film, sports, food, politics family, home/garden, music, news, gaming, shopping, and so on.

At a glance, there are several seemingly correlated topics. One might expect cooking and sports fandom to be associated, or even family and religion. Additionally, one could reason that someone who is into television/film might also be into art, or politics. Similarly, someone into personal fitness might be interested in health and nutrition, or sports.

For our analysis, we decide to use principal component analysis. This will be useful because there are many columns, and many seemingly related interests; if these interests are as associated as we believe, then there should be several potential market segments that we can identify by collapsing these 30+ columns into just a handful.

The results are printed below.

Principal Component Analysis, Rank = 4

```
##                        PC1         PC2          PC3           PC4
## chatter          0.074883359  0.67159460 -0.283213309  0.1325371902
## current_events   0.015506521  0.05963470  0.007228135  0.0137387183
## travel           0.006193471  0.11441416  0.383379907  0.1968648401
## photo_sharing    0.116776018  0.48264126 -0.177063419  0.0021574306
## uncategorized    0.023899990  0.02493758  0.002985219 -0.0173835221
## tv_film          0.005109371  0.03401576  0.064107588 -0.0523582859
## sports_fandom    0.020170811  0.06296468  0.153918179  0.0209304880
## politics         0.006526645  0.19545248  0.605067050  0.3736867657
## food             0.095919447  0.01836725  0.122347518  0.0012762032
## family           0.018901971  0.04515902  0.049332937 -0.0085365424
## home_and_garden  0.014736298  0.01970403  0.014560844 -0.0015802188
## music            0.022797491  0.04313340  0.010900348 -0.0311710789
## news             0.017935306  0.07485343  0.307780698  0.1776763888
## online_gaming    0.012999833  0.12286469  0.259424928 -0.5726255240
## shopping         0.047927389  0.26708624 -0.106992580  0.0437324564
## health_nutrition 0.825055601 -0.21471066  0.012035804  0.0313457483
## college_uni      0.001604936  0.16917431  0.294908715 -0.6245427298
## sports_playing   0.017904066  0.04793512  0.064047646 -0.1190323999
## cooking          0.315430810  0.20623443 -0.051587171 -0.1358752289
## eco              0.039030172  0.02853638  0.007089299  0.0060314484
## computers        0.014797895  0.06702828  0.153405950  0.0860829790
## business         0.012189367  0.03852813  0.017008494  0.0107870672
## outdoors         0.146996128 -0.02641590  0.034263947  0.0151607207
## crafts           0.020651057  0.03132161  0.022894696  0.0008188389
## automotive       0.005200126  0.07744024  0.094823797  0.0502708968
## art              0.018217445  0.02388688  0.051325515 -0.0392906027
## religion         0.026407163  0.03707328  0.123017954  0.0057116583
## beauty           0.050625562  0.08727072 -0.011961660 -0.0352147880
## parenting        0.028032726  0.04122222  0.087909809  0.0105873613
## dating           0.037454685  0.06656071  0.025729580  0.0251088119
## school           0.019640750  0.05526384  0.038321044  0.0123905564
## personal_fitness 0.394578868 -0.07748228  0.008252702  0.0151991663
```

```
## Importance of first k=4 (out of 32) components:
##                            PC1    PC2    PC3    PC4
## Standard deviation     5.1834 4.2848 3.8189 3.7611
## Proportion of Variance 0.2009 0.1373 0.1091 0.1058
## Cumulative Proportion  0.2009 0.3382 0.4473 0.5531
```

When we collapse our data into four principal components, we can identify four disparate groups based on their interests. These alone explain over 55% of the variation in the original 30+ columns of data. While additional components can increase the explained variation, for the most part these increases are marginal (2-3%) and aren't large enough to constitute what we'd like to call "market segments".

In the next section, we're going to try and isolate the interests of these market segments, and use PCA further to see how much information/variance we can preserve by collapsing these them into a single principal component. If these principal components are indeed separate market segments, then much of the information should be preserved when we collapse the interests that make them up into a single component. If this is the case, we report the identified segments below.

Fitness Buffs

The first principal component appears to be made up of people who tweet about photo sharing, cooking, nutrition, outdoors, and personal fitness. We dub this group the "fitness gurus".

Now, we'll collapse these interests into one group to see how much of the information we lose.

```
##                         PC1        PC2
## photo_sharing    0.08615616  0.5199104
## health_nutrition 0.85534353 -0.2861710
## cooking          0.31028699  0.7964327
## personal_fitness 0.40581589 -0.1161651
```

```
## Importance of first k=2 (out of 4) components:
##                           PC1    PC2
## Standard deviation     5.0800 3.5085
## Proportion of Variance 0.5706 0.2722
## Cumulative Proportion  0.5706 0.8428
```

If we were to collapse these four components into a single component, we can see that we'd lose about 43% of the information in the process. This indicates that there may be a second market segment within these four interest. We can see that the first component is made up of people who primarily tweet about health/nutrition, personal fitness, and cooking, whereas the second component is made up of photo sharing and cooking. These might represent two distinct groups within our market segment: those whose interests are focused around fitness and wellness, and those who enjoy cooking and sharing their meals on social media.

College Gamers

The next segment that stands out as a potential market segment are people who tweet about college and online gaming from the second principal component. This is not very surprising, considering that most online gaming tends to skew towards educated groups. A Pew survey from 2003 noted that over 70% of college students reported playing a game on a PC or online, and we could reasonably expect this number to have grown from 2003 to 2014.

```
##                      PC1
## college_uni   0.7405565
## online_gaming 0.6719941

## Importance of first k=1 (out of 2) components:
##                            PC1
## Standard deviation     3.7221
## Proportion of Variance 0.8871
## Cumulative Proportion  0.8871
```

The results from summary(pc_z) tell us that by collapsing tweets about college/university and online gaming into one component, that we would only lose about 12% of the data in that process. We argue that this is a good second candidate for a market segment, which we will characterize as *college gamers*.

Urban Professionals

Next, it appears that people who tweet about politics also tweet about traveling and computers. Let's see how much information is preserved when we collapse this into two principal components:

```
##                   PC1        PC2
## travel    -0.5391785  0.8040721
## computers -0.2227266  0.1507210
## politics  -0.8122065 -0.5751098

## Importance of first k=2 (out of 3) components:
##                             PC1    PC2
## Standard deviation     3.576 1.5023
## Proportion of Variance 0.809 0.1428
## Cumulative Proportion  0.809 0.9519
```

We can see that by collapsing tweets about computers, politics, and traveling into two components, we're able to preserve 95% of the information we had before. The first principal component preserves about 80% of the variation from all three columns. We argue that the users tweeting about these subjects might be aptly described as urban professionals. These may be highly-educated, politically-engaged, tech-savvy people.

Media Socialites

Next, based on the third principal component, it looks like chatter and photo sharing are pretty strongly associated. We also include two other seemingly correlated subjects, beauty and shopping into the analysis.

```
##                      PC1
## chatter       0.8015406
## photo_sharing 0.5104294
## shopping      0.3114394

## Importance of first k=1 (out of 3) components:
##                            PC1
## Standard deviation     4.1465
## Proportion of Variance 0.7415
## Cumulative Proportion  0.7415
```

By collapsing these four into one principal component, it looks like we're able to maintain over 70% of the variation in those four columns. We argue that this is a pretty strong indication of a market segment.

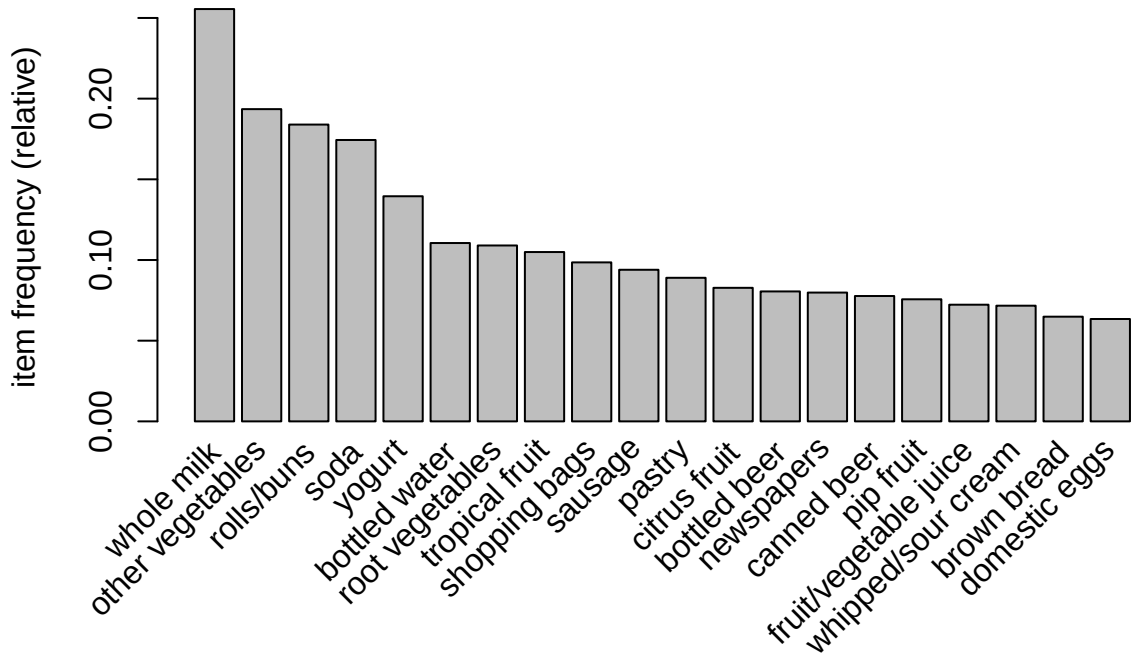
Conclusion

In total, we were able to identify 4 sizeable market segments. These were the *Fitness Buffs*, individuals whose tweets fell into the photo sharing, cooking, nutrition, and fitness categories. Next, we identified the *College Gamers*, whose posts were centered around university life and online gaming. Then, we uncovered the *Urban Professionals*, who mainly discussed politics, traveling, and computers. Lastly, we found the *Media Socialites*, who were people that tweeted about chatter, photo sharing, beauty and shopping.

#3) Text Association and Groceries

Next, we're given a dataset on several thousand grocery shopping bundles. We're asked to find any interesting associations between purchases. First, we'll do a little visual inspection.

Relative Frequency of Items in Baskets



Here, we can see that the five most frequently bought items are whole milk, other vegetables, rolls/buns, soda, and yogurt. Whole milk appears in over 1/4 of grocery baskets, whereas vegetables and rolls appear in about 1/5 baskets. Yogurt, soda, and bottled water are in just over 10% of baskets, too.

Next, we'll look at a Sparse Matrix to see if there are any visually apparent associations between bundles.

Sparse Matrix: Identifying Popular Items Visually

Although it's not clear which items are in what columns, it's clear that certain items appear in baskets more frequently than others. Some columns are populated enough to almost take on a piecemeal line; these are items likely items like whole milk and vegetables, of which at least one has a 1/2 chance of showing up in any given shopping cart going through the checkout.

10 Prominent Associations

```
## Apriori
##
## Parameter specification:
```

```
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##        0.25    0.1    1 none FALSE            TRUE       5   0.007      2
##  maxlen target  ext
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 68
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
## sorting and recoding items ... [104 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [363 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
##      lhs                     rhs                support     confidence
## [1]  {herbs}              => {root vegetables}  0.007015760 0.4312500
## [2]  {herbs}              => {other vegetables} 0.007727504 0.4750000
## [3]  {herbs}              => {whole milk}       0.007727504 0.4750000
## [4]  {processed cheese}   => {whole milk}       0.007015760 0.4233129
## [5]  {semi-finished bread} => {whole milk}      0.007117438 0.4022989
## [6]  {detergent}          => {whole milk}       0.008947636 0.4656085
## [7]  {pickled vegetables}  => {whole milk}      0.007117438 0.3977273
## [8]  {baking powder}      => {other vegetables} 0.007320793 0.4137931
## [9]  {baking powder}      => {whole milk}       0.009252669 0.5229885
## [10] {flour}              => {whole milk}       0.008439248 0.4853801
##      coverage   lift     count
## [1]  0.01626843 3.956477 69
## [2]  0.01626843 2.454874 76
## [3]  0.01626843 1.858983 76
## [4]  0.01657346 1.656698 69
## [5]  0.01769192 1.574457 70
## [6]  0.01921708 1.822228 88
## [7]  0.01789527 1.556565 70
## [8]  0.01769192 2.138547 72
## [9]  0.01769192 2.046793 91
## [10] 0.01738688 1.899607 83
```

Here, we see 10 product associations that we were able to uncover. Some that make sense are herbs and other/root vegetables, baking powder/flour and whole milk, specialty bars and soda, and grapes and vegetables. Next, we'll inspect the associations with the top five lifts.

Identifying Likely Complements

```
##     lhs                   rhs                      support confidence   coverage     lift count
## [1] {herbs}            => {root vegetables}    0.007015760  0.4312500 0.01626843 3.956477    69
## [2] {berries}          => {whipped/sour cream} 0.009049314  0.2721713 0.03324860 3.796886    89
## [3] {other vegetables,
##      tropical fruit,
```

```
##      whole milk}       => {root vegetables}    0.007015760  0.4107143 0.01708185 3.768074    69
## [4] {beef,
##      other vegetables} => {root vegetables}    0.007930859  0.4020619 0.01972547 3.688692    78
## [5] {other vegetables,
##      tropical fruit}   => {pip fruit}          0.009456024  0.2634561 0.03589222 3.482649    93
```

We can see that customers who buy herbs are about four times as likely to buy roots than other customers. Customers who buy berries are almost three times as likely to buy whipped cream, and those who buy tropical fruit and whole milk tend to be four times as likely to buy root vegetables. These may represent people with sweet tooths and vegetarians, respectively.

Most Associated Purchases

```
##     lhs                   rhs                support    confidence coverage
## [1] {other vegetables} => {whole milk}       0.07483477 0.3867578  0.1934926
## [2] {whole milk}       => {other vegetables} 0.07483477 0.2928770  0.2555160
## [3] {rolls/buns}       => {whole milk}       0.05663447 0.3079049  0.1839349
## [4] {yogurt}           => {whole milk}       0.05602440 0.4016035  0.1395018
## [5] {root vegetables}  => {whole milk}       0.04890696 0.4486940  0.1089985
##     lift     count
## [1] 1.513634 736
## [2] 1.513634 736
## [3] 1.205032 557
## [4] 1.571735 551
## [5] 1.756031 481
```

When we limit the associations to the those that appear the most, what appears are associations between the most commonly bought items, namely whole milk, other vegetables, rolls/buns, yogurt, and root vegetables. It makes sense that whole milk appears in many of these associations, given its prominence in the human diet (breakfast cereal, afternoon snacks, coffee, etc...). Its association with vegetables is also unsurprising, given that meals tend to feature a side dish; the same applies to rolls/buns.

#4) Text Analysis (Incomplete)

This problem was not able to be finished. However, we still describe the process we would have taken. The code we would have used is included in the .Rmd file, and while most of it runs in R Markdown chunks, these are all set to eval=FALSE and echo = FALSE due to errors and to allow the document to knit.

The process we would have like to have taken would have been as follows. After initializing a corpus for each author, we would have then then removed punctuation, numbers, removed white space, converted our text to lowercase, and removed stop words.

Next, we would have created term-document matrices for each of our authors. Next, we would have created weighted term-document matrices, with the intent of isolating words that appeared most frequently. From there, we would have split these into test/train splits.

At that point, we could have then used a number of techniques to draw comparisons between texts or attribute authorship given a particular text. We could use KNN to compare how one document compares to others most like it, or cluster based on the frequency of a set of words particular to an author.