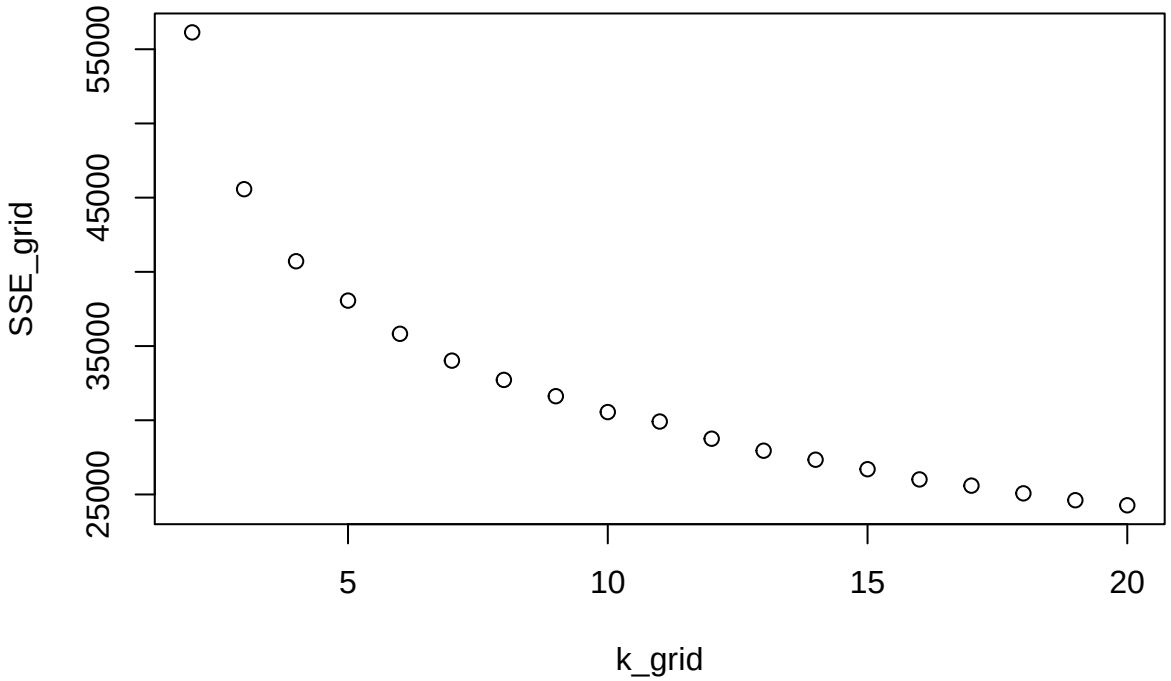
HW4

Akanksha Arora, Chad Rudolph, Sam Stripling

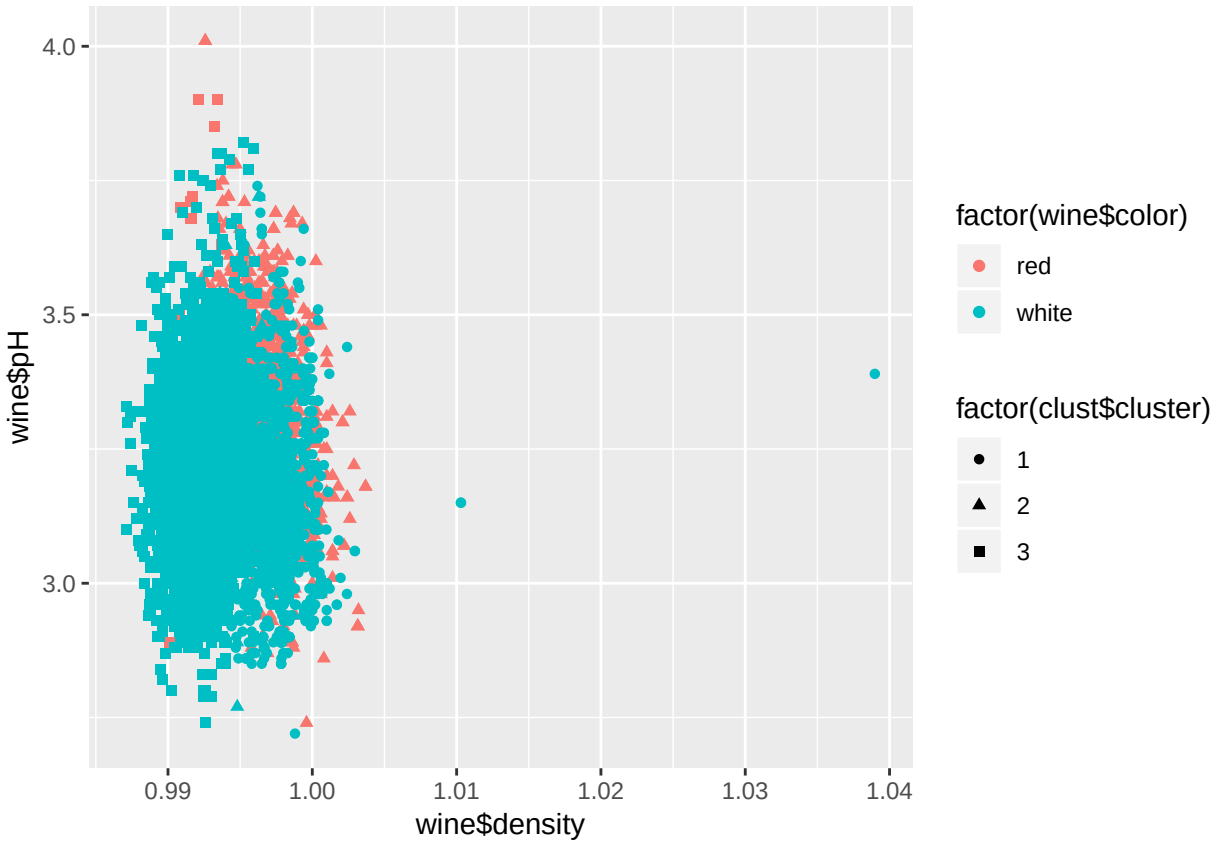
4/26/2019

Question 1

We begin with clustering. In this instance, we felt that the elbow plot was the best method for identifying an optimal K. The elbow plot below identifies K to be a value of 3:



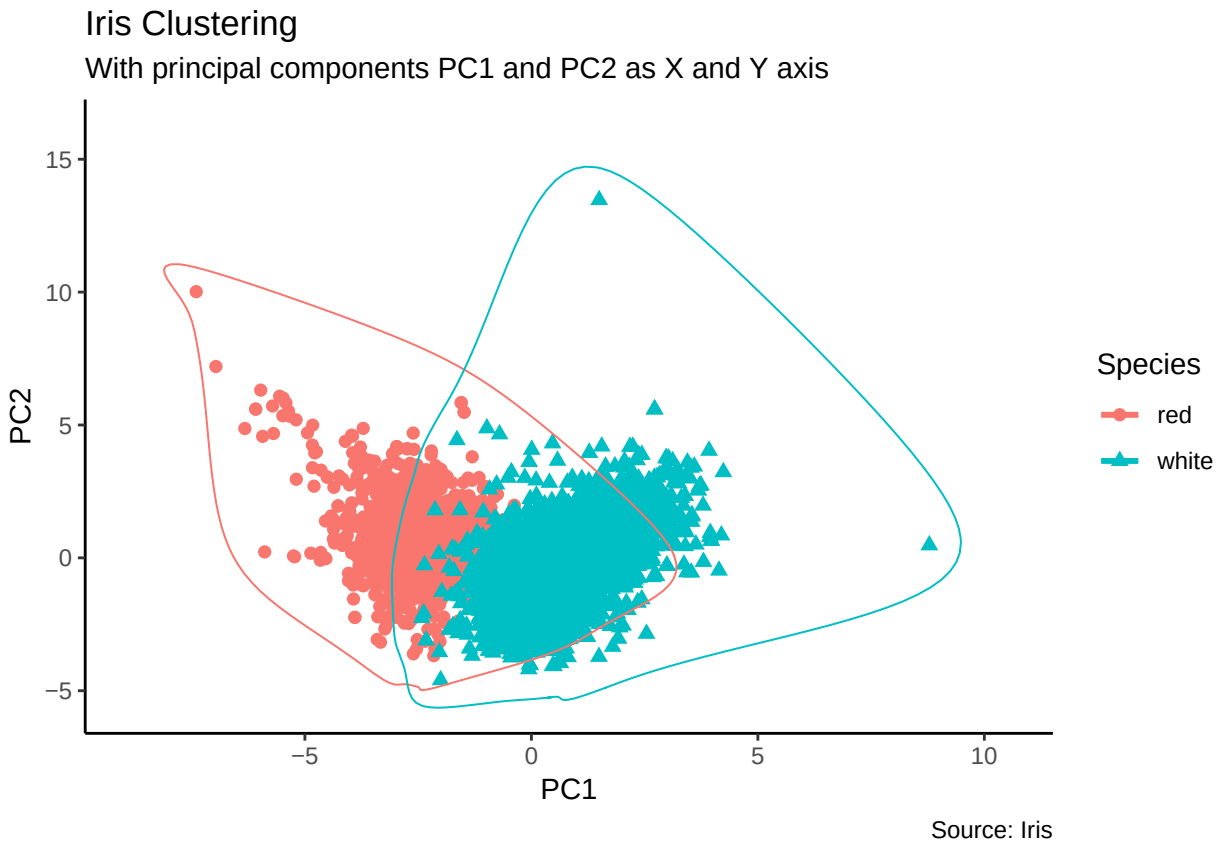
Now that the optimal K is known, we can obtain clusters using this value and seperate red and white wine. Using the density and pH levels of the wines as references, we observe the breakdown of red vs. white wine.



From this chart, it appears that red wines tend to have higher densities on average compared to white wine, but similar pH levels. Finally, we observe a table with the breakdown of wine color per cluster:

```
##              wine$color
## clust$cluster  red white
##             1    4  1891
##             2 1538    55
##             3   57  2952
```

White wine seems to dominate the first two clusters, while red wine dwarfs its counterpart in the 3rd cluster. Next, we move to a PCA analysis.



Via higher-order PC analysis we can easily see the difference between red and white wine. Running PC2 on the residual matrix from PC1 allows for a clearer picture than clustering, clearly separating the two wines. It is feasible that this method could sort the high quality wine from the low quality, specifically if the qualities were given within a range, such as 0-5 equals low quality and 6-10 equals high quality.

Question 2

NutrientH20 Customer Report: We have identified several groups that could be a good target for our marketing campaigns. We outline our process in selecting these groups. Given that our advertising efforts are expensive, we narrowed down our search to the key group that we should target. We examined a table of the pairwise correlations between Twitter chat categories. There was a .81 correlation between health_nutrition and personal_fitness Tweets. This was the largest correlation observed for any given pair of variables. Given that we are a nutrition company, this group should be our prime marketing target. Every individual who had more than three of both health_nutrition and personal_fitness Tweets is considered to be our "Bronze" marketing base. Our "Silver" marketing base is those individuals with more than 5 of both nutrition and personal fitness Tweets. Gold is for individuals with over 10 of both. We list the Gold individuals below.

Note that we considered running K-means clustering to partition the Twitter users into groups. However, the strong correlation that exists between health_nutrition and personal_fitness Tweets suggests a simpler method of finding our key marketing targets (we are a nutrition company) that can avoid the pitfalls of K-means clustering. For instance, it is not clear how many clusters to use, nor which distance measure (ex. Euclidean, Taxi Cab) to use in determining our clusters. We could very easily end up with clusters too broad (that don't allow us to pinpoint our group of interst) or too narrow (and suffer from overfitting).

Gold Targets (Over 10 Tweets in health_nutrition and personal_fitness categories):

```
print(Gold)
```

```
##              X
```

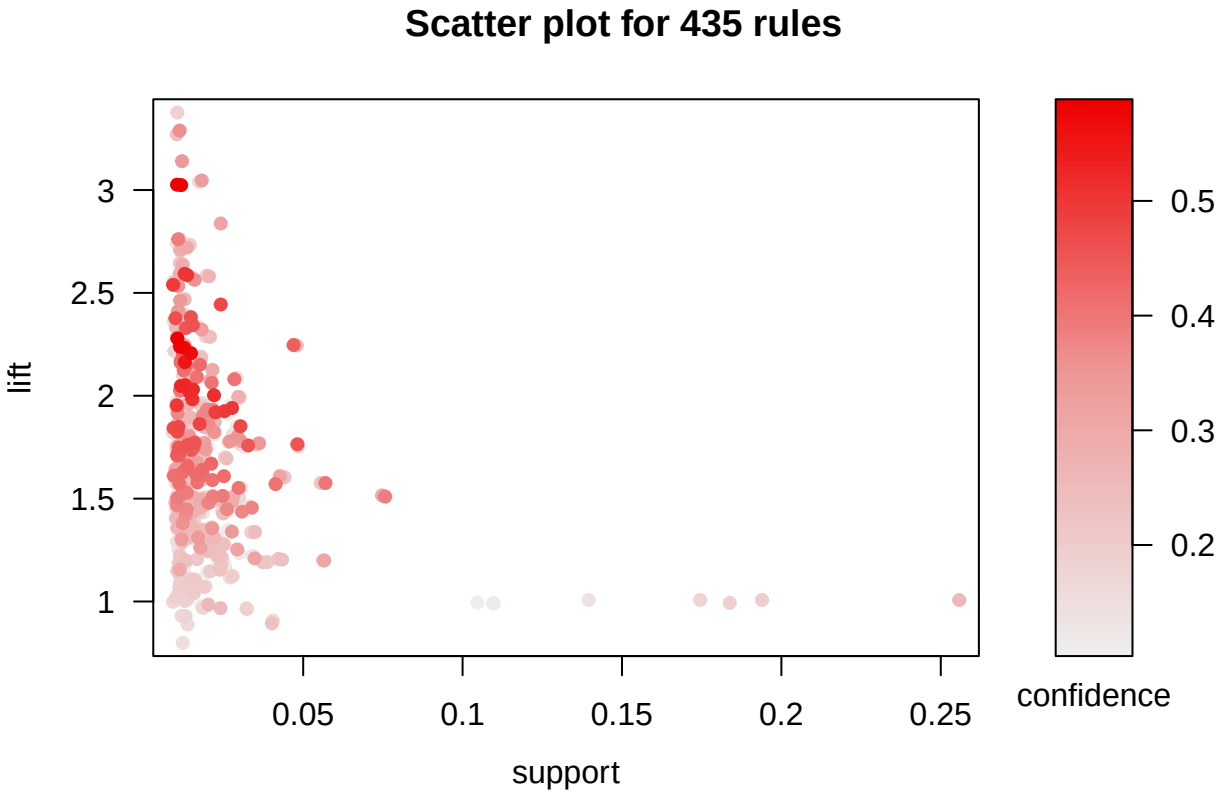
```
## 1    hmjoe4g3k
## 9    y2g68vhkf
## 86   p94myqeo6
## 105  k3ury7nvg
## 121  47nchz8gb
## 173  7kbre6zo3
## 227  ceokhdut7
## 233  zedtqi9y8
## 277  dpe1m5j7f
## 550  w9e5aisrh
## 710  b937ti6yq
## 818  hsul8baz2
## 880  vopldhf12
## 1006 rkelthwvn
## 1080 oenvhs172
## 1099 7gijla8x2
## 1147 unflz26qx
## 1217 oprvg2bzw
## 1311 sbm1o46hg
## 1316 ru1bfxqpi
## 1342 txyump81e
## 1365 mhkdow2li
## 1420 d2ukegmpn
## 1459 klagj7dew
## 1518 l3otmq7uh
## 1533 i7fcoa38b
## 1544 xjsfm12tq
## 1852 7tvl324na
## 1995 dlev61xku
## 2086 zfwgkmvay
## 2270 p3orj15vs
## 2437 8k2fe14zy
## 2630 ewzrilmdo
## 2916 kdbxczi63
## 3148 fxthw8rga
## 3250 e8xavrhym
## 3350 tjr9ixg1n
## 3369 o1l2f6ckn
## 3570 nwryhmtuj
## 3576 d6zpfsmve
## 3679 v9c1olyfa
## 3683 7ilmze639
## 3715 89qiyfdhw
## 3960 6uwxfp5os
## 3990 isb5dm3zy
## 4014 62xml1gnt
## 4170 smdtan8gu
## 4209 oiqj31dt7
## 4229 y79twkv82
## 4369 2feyhmsrt
## 4458 zp2y6ieou
## 4533 vpx1f3ygn
## 4572 ytmj1eiqz
## 4694 14wuvr5ph
```

```
## 4754 w5vqp3d9f
## 5053 ldz59peax
## 5130 s6kzn17qi
## 5624 j5orpadby
## 5740 dlfzukcs2
## 5745 q7hfrlktj
## 5749 lhixoq27s
## 5865 afquj4vs6
## 5912 6w3lohqdm
## 5920 49pq2x8ng
## 5955 t13vjxpbl
## 6041 sbo8lrgy2
## 6077 jmh569bz7
## 6118 hgwblyq4o
## 6132 obqzy8vnd
## 6427 v1o65a3yl
## 7107 l6xtj81mg
## 7118 tnov74ic3
## 7228 xs5ulvge7
## 7581 dp9e43q6w
## 7586 bay897li2
## 7599 qznx1fta8
## 7660 drujonq46
## 7725 aqgxscu9p
```

Question 3

After reorganizing the data in a user-friendly way, we began our analysis. We limited the data to fit the following categories: 1% support, 10% confidence, and a max length of 10. Below is a graphical summary of the 435 rules this produces.

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



In the graph above, confidence is displayed by a shade of red, which darkens with higher confidence. The support is displayed upon the x-axis, and the lift is displayed upon the y-axis. From the graph, it is easy to tell that the majority of the high confidence, high lift data resides the support range of 1% to 5%. Specifically, it appears that a lift of 2 yielded many high-confidence rules. So, we decided to inspect the data at a lift of 2 and a confidence of 50%. The results are as follows:

```
##      lhs                     rhs                   support confidence     lift count
## [1]  {curd,
##       yogurt}             => {whole milk}       0.01006609  0.5823529 2.279125    99
## [2]  {butter,
##       other vegetables}   => {whole milk}       0.01148958  0.5736041 2.244885   113
## [3]  {domestic eggs,
##       other vegetables}   => {whole milk}       0.01230300  0.5525114 2.162336   121
## [4]  {whipped/sour cream,
##       yogurt}             => {whole milk}       0.01087951  0.5245098 2.052747   107
## [5]  {other vegetables,
##       pip fruit}          => {whole milk}       0.01352313  0.5175097 2.025351   133
## [6]  {citrus fruit,
##       root vegetables}    => {other vegetables} 0.01037112  0.5862069 3.029608   102
## [7]  {root vegetables,
##       tropical fruit}     => {other vegetables} 0.01230300  0.5845411 3.020999   121
## [8]  {root vegetables,
##       tropical fruit}     => {whole milk}       0.01199797  0.5700483 2.230969   118
## [9]  {tropical fruit,
##       yogurt}             => {whole milk}       0.01514997  0.5173611 2.024770   149
## [10] {root vegetables,
##       yogurt}             => {whole milk}       0.01453991  0.5629921 2.203354   143
## [11] {rolls/buns,
##       root vegetables}    => {other vegetables} 0.01220132  0.5020921 2.594890   120
## [12] {rolls/buns,
```

```
##       root vegetables}    => {whole milk}       0.01270971  0.5230126 2.046888   125
## [13] {other vegetables,
##       yogurt}             => {whole milk}       0.02226741  0.5128806 2.007235   219
```

In general, it appears whole milk and "other vegetables" go with a variety of bundles. The items within these bundles make logical sense. For example, butter and other vegetables match with whole milk. Interestingly, root vegetables always match with other vegetables at these levels. We obtain the highest lifts (above 3) when matching fruits and root vegetables with other vegetables. Again, this should be expected, as many people will buy their fruits and vegetables at the same time. In conclusion, the most common items that make up a bundle are vegetables and milk, which is unsurprising given their common use in cooking meals.