Branden Ho
Anna Nguyen

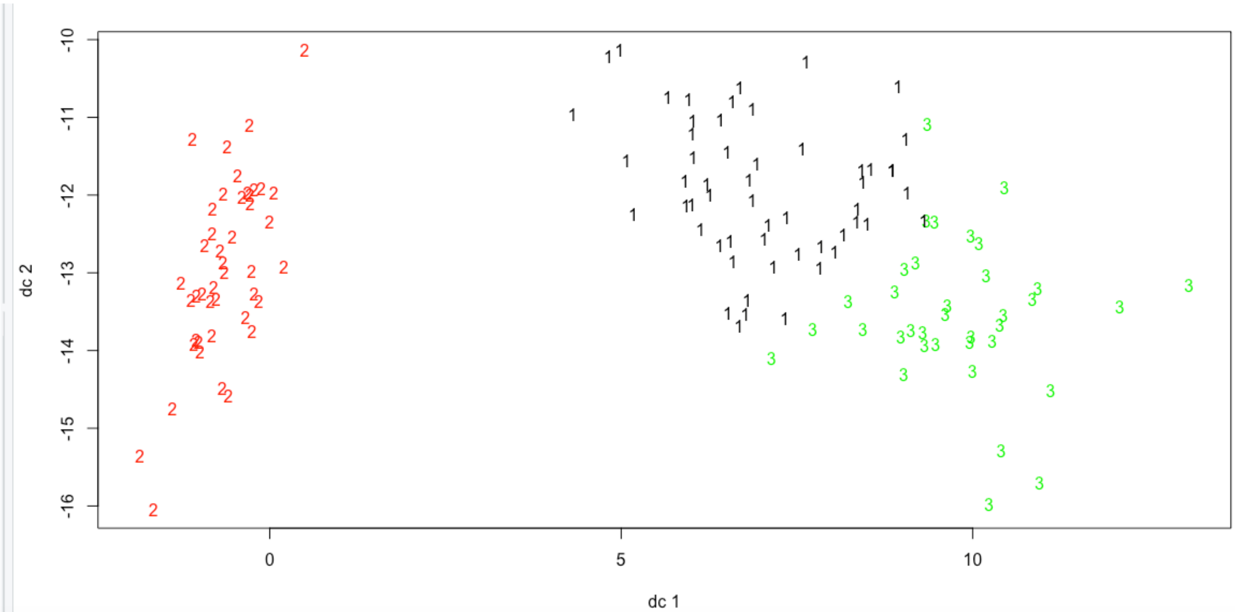
Case 5

Part 1 : Clustering Analysis

1. K-means clustering

1	2	3
55	44	36

=> with a three-cluster solution, this table displays the number of clusters in cluster 1, 2, and 3



=> this graph tells us that RStudio tries to group non-overlapping clusters in a way that clusters that are homogeneous will be put together in each of the three clusters while dissimilar clusters will be grouped separately (which is cluster 1, 2, and 3).

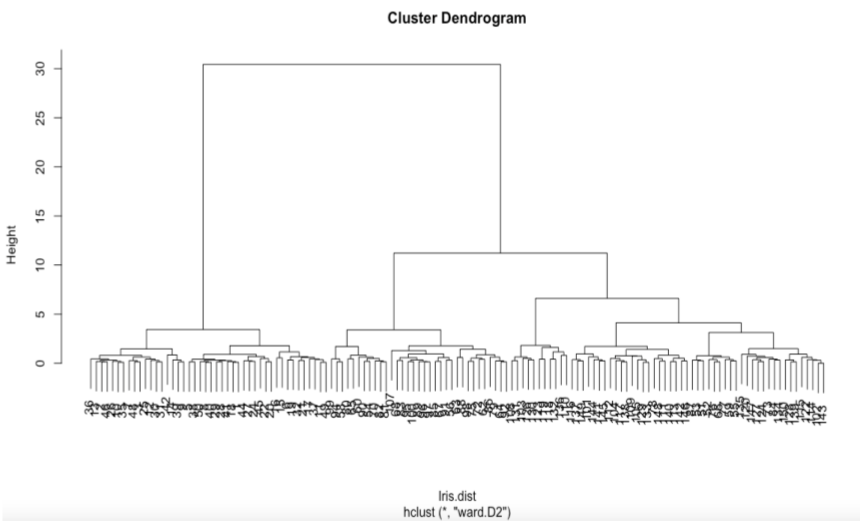
```
> iris$Species[fit$cluster == 1]
 [1] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
 [9] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
[17] setosa     setosa     setosa     versicolor versicolor versicolor versicolor versicolor
[25] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
[33] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
[41] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[49] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[57] virginica  virginica  virginica  virginica  virginica
Levels: setosa versicolor virginica
> iris$Species[fit$cluster == 2]
 [1] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
 [9] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
[17] setosa     setosa     setosa     versicolor versicolor versicolor versicolor versicolor
[25] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
[33] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[41] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[49] virginica
Levels: setosa versicolor virginica
> iris$Species[fit$cluster == 3]
 [1] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
 [9] setosa     setosa     setosa     setosa     versicolor versicolor versicolor versicolor
[17] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
[25] versicolor versicolor versicolor versicolor virginica  virginica  virginica  virginica
[33] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
Levels: setosa versicolor virginica
```

=> this output shows us which items are in cluster 1, 2, and 3

```
  Group.1 Sepal.Length Sepal.Width Petal.Length Petal.Width
1       1     5.883636    2.747273     4.390909   1.4400000
2       2     5.020455    3.395455     1.472727   0.2545455
3       3     6.836111    3.069444     5.691667   2.0444444
```

=> this output shows us the cluster mean for each variable within each cluster. These statistics can be used to see the what our averages are across each cluster and help us see why certain flowers were out into which clusters.

2. Hierarchical clustering



=> by using the Wards method to obtain clusters, we can see that the dendrogram groups similar observations into a branch and most similar pair of clusters will be merged into one single big cluster, pair of clusters are merged based on their distance

```
groupIris.3
 1  2  3
44 58 33
```

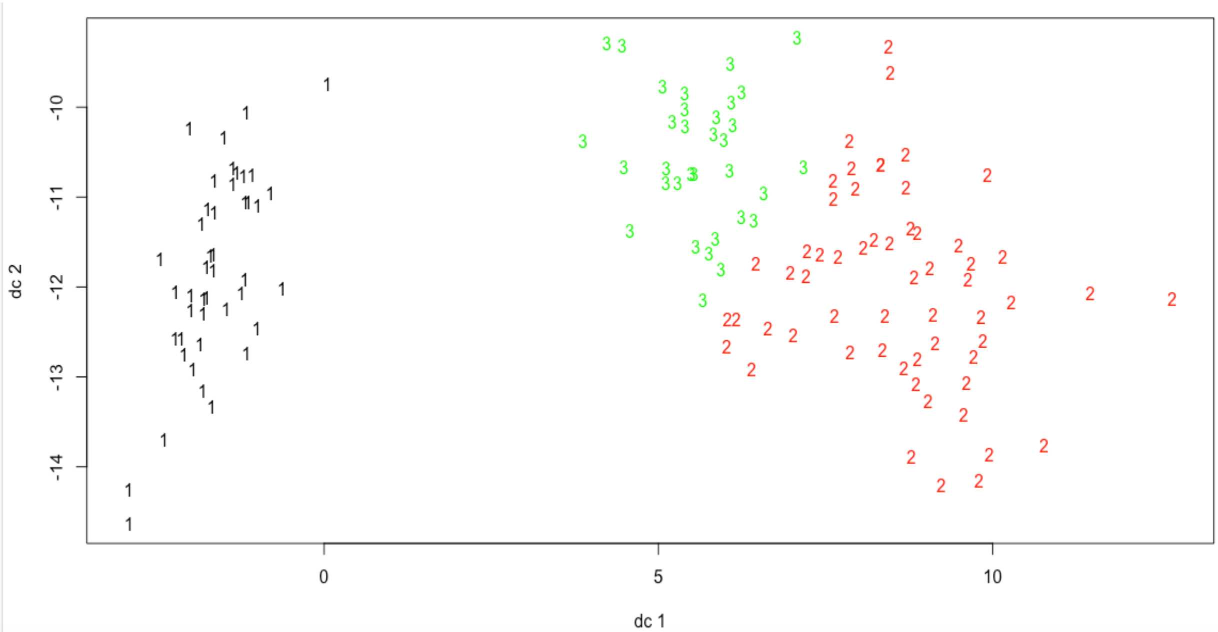
=> if we want to cut the dendrogram at the 3 cluster level, this output shows us the number of items in each cluster 1, 2, and 3. As you can see, with the hierarchical clustering method, cluster 2 has the most items while with the K-means method, cluster 1 has the most items.

```
> iris$Species[groupIris.3 == 3]
 [1] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
 [9] setosa     setosa     setosa     versicolor versicolor versicolor versicolor versicolor
[17] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
[25] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[33] virginica  virginica  virginica  virginica
Levels: setosa versicolor virginica
> iris$Species[groupIris.3 == 2]
 [1] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
 [9] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
[17] setosa     setosa     setosa     setosa     versicolor versicolor versicolor versicolor
[25] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
[33] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
[41] versicolor versicolor versicolor versicolor virginica  virginica  virginica  virginica
[49] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[57] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[65] virginica
Levels: setosa versicolor virginica
> iris$Species[groupIris.3 == 1]
 [1] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
 [9] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
[17] setosa     setosa     setosa     versicolor versicolor versicolor versicolor versicolor
[25] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
[33] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[41] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[49] virginica
Levels: setosa versicolor virginica
```

=> this output shows us which items are in cluster 1, 2, and 3

	Group.1	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	1	5.020455	3.395455	1.472727	0.2545455
2	2	6.594828	2.967241	5.374138	1.9103448
3	3	5.672727	2.712121	4.081818	1.2727273

=> this output shows us the cluster mean for each variable within each cluster.



=> This graph shows us that the hierarchical clustering method groups each cluster in a way that clusters are formed based on distance between objects and we cut the dendrogram at the three cluster level to result in this centroid graph.

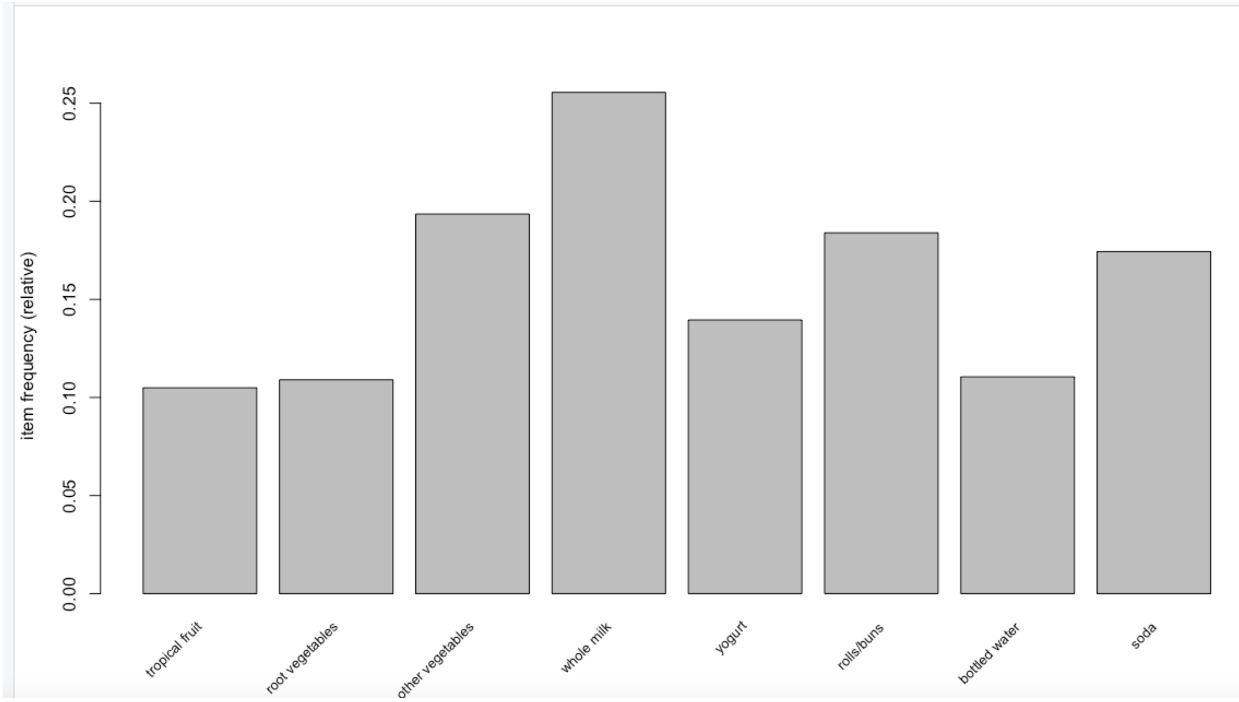
Part 2: Association Rules

```
transactions as itemMatrix in sparse format with
 9835 rows (elements/itemsets/transactions) and
 169 columns (items) and a density of 0.02609146

most frequent items:
      whole milk other vegetables       rolls/buns             soda           yogurt
            2513             1903             1809             1715             1372
         (Other)
           34055

element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18
2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46   29   14
  19   20   21   22   23   24   26   27   28   29   32
  14    9   11    4    6    1    1    1    1    3    1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   4.409   6.000  32.000

includes extended item information - examples:
       labels  level2           level1
1 frankfurter sausage meat and sausage
2     sausage sausage meat and sausage
3  liver loaf sausage meat and sausage
```

=> as you can see, we have 9835 rows (transactions) and 169 columns (items), most frequently purchased items are whole milk, and most transactions bought between 1 to 4 items with a mean number of items per transaction would be 4.41. And the largest (or maximum) number of items per transaction is 32 items.

```
transactions as itemMatrix in sparse format with
 10 rows (elements/itemsets/transactions) and
 169 columns (items) and a density of 0.01775148

most frequent items:
      whole milk           yogurt other vegetables       rolls/buns     citrus fruit
               4                3                2                2                1
         (Other)
              18

element (itemset/transaction) length distribution:
sizes
1 2 3 4 5
3 1 1 3 2

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    1.25    3.50    3.00    4.00    5.00

includes extended item information - examples:
       labels  level2           level1
1 frankfurter sausage meat and sausage
2     sausage sausage meat and sausage
3  liver loaf sausage meat and sausage
```

=> if we take a closer look at the first 10 transactions, we can see that whole milk was also purchased the most frequently with a frequency of 4 times, and we noticed that three transactions bought 1 item and three transactions bought 4 items with a mean number of items per transaction is 3 items. And the largest transaction involved 5 items.



=> the frequency plot with 10% support also shows us that whole milk was purchased the most frequently with a relative frequency of 25%.

```
set of 22 rules

rule length distribution (lhs + rhs):sizes
 3  4
13  9

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   3.000   3.000   3.409   4.000   4.000

summary of quality measures:
    support           confidence        coverage             lift           count
 Min.   :0.005084   Min.   :0.6022   Min.   :0.008134   Min.   :2.357   Min.   :50.00
 1st Qu.:0.005414   1st Qu.:0.6088   1st Qu.:0.008566   1st Qu.:2.427   1st Qu.:53.25
 Median :0.005745   Median :0.6222   Median :0.009253   Median :2.463   Median :56.50
 Mean   :0.006202   Mean   :0.6282   Mean   :0.009881   Mean   :2.599   Mean   :61.00
 3rd Qu.:0.006660   3rd Qu.:0.6368   3rd Qu.:0.010244   3rd Qu.:2.627   3rd Qu.:65.50
 Max.   :0.009354   Max.   :0.7000   Max.   :0.014642   Max.   :3.273   Max.   :92.00

mining info:
      data ntransactions support confidence
 Groceries          9835   0.005        0.6
                                                                                  call
 apriori(data = Groceries, parameter = list(support = 0.005, confidence = 0.6))
```

=> in this output it shows that I generated 22 association rules, 13 rules with 3 items and 9 rules with 4 items. It also shows that my average lift is 2.599.

```
    lhs                                            rhs                support     confidence coverage    lift     count
[1] {citrus fruit, root vegetables, whole milk} => {other vegetables} 0.005795628 0.6333333  0.009150991 3.273165 57
[2] {pip fruit, root vegetables, whole milk}    => {other vegetables} 0.005490595 0.6136364  0.008947636 3.171368 54
[3] {pip fruit, whipped/sour cream}             => {other vegetables} 0.005592272 0.6043956  0.009252669 3.123610 55
[4] {root vegetables, onions}                   => {other vegetables} 0.005693950 0.6021505  0.009456024 3.112008 56
[5] {tropical fruit, root vegetables, yogurt}   => {whole milk}       0.005693950 0.7000000  0.008134215 2.739554 56
```

=> if we take a closer look at the first 5 rules sorted by the highest lift, we can see that items in the RHS which is other vegetables has the highest lift which means that they are likely to be purchased 3 times more when being purchased with items in the LHS. The confidence level also tells us that there are higher chances of items in the RHS to be purchased together with items in the LHS the higher the confidence level is.