R Notebook

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.4.2

## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.0     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0

## Warning: package 'ggplot2' was built under R version 3.4.4

## Warning: package 'tibble' was built under R version 3.4.4

## Warning: package 'tidyr' was built under R version 3.4.3

## Warning: package 'purrr' was built under R version 3.4.4

## Warning: package 'dplyr' was built under R version 3.4.4

## Warning: package 'forcats' was built under R version 3.4.3

## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.4.3
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.4.4

## Loading required package: lattice

## Warning in as.POSIXlt.POSIXct(Sys.time()): unknown timezone 'zone/tz/2018i.
## 1.0/zoneinfo/America/Chicago'

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

Market segmentation is a strategy that **divides a broad target market of customers into smaller, more similar groups, and then designs a marketing strategy specifically for each group**. Clustering is a common technique for market segmentation since it automatically finds similar groups given a data set.

In this problem, we'll see how clustering can be used to find similar groups of customers who belong to an airline's frequent flyer program. The airline is trying to learn more about its customers so that it can target different customer segments with different types of mileage offers.

The file AirlinesCluster.csv contains information on 3,999 members of the frequent flyer program. This data comes from the textbook "Data Mining for Business Intelligence," by Galit Shmueli, Nitin R. Patel, and Peter C. Bruce. For more information, see the website for the book.

There are seven different variables in the dataset, described below:

Balance = number of miles eligible for award travel

QualMiles = number of miles qualifying for TopFlight status

BonusMiles = number of miles earned from non-flight bonus transactions in the past 12 months

BonusTrans = number of non-flight bonus transactions in the past 12 months

FlightMiles = number of flight miles in the past 12 months

FlightTrans = number of flight transactions in the past 12 months

DaysSinceEnroll = number of days since enrolled in the frequent flyer program

1.1) Normalizing the Data

Read the dataset AirlinesCluster.csv into R and call it "airlines".

Looking at the summary of airlines, which TWO variables have (on average) the smallest values?

Which TWO variables have (on average) the largest values?

```
airlines = read.csv('AirlinesCluster.csv')

summary(airlines)
```

```
##     Balance          QualMiles         BonusMiles       BonusTrans
##  Min.   :      0   Min.   :    0.0   Min.   :     0   Min.   : 0.0
##  1st Qu.:  18528   1st Qu.:    0.0   1st Qu.:  1250   1st Qu.: 3.0
##  Median :  43097   Median :    0.0   Median :  7171   Median :12.0
##  Mean   :  73601   Mean   :  144.1   Mean   : 17145   Mean   :11.6
##  3rd Qu.:  92404   3rd Qu.:    0.0   3rd Qu.: 23800   3rd Qu.:17.0
##  Max.   :1704838   Max.   :11148.0   Max.   :263685   Max.   :86.0
##   FlightMiles       FlightTrans     DaysSinceEnroll
##  Min.   :    0.0   Min.   : 0.000   Min.   :   2
##  1st Qu.:    0.0   1st Qu.: 0.000   1st Qu.:2330
##  Median :    0.0   Median : 0.000   Median :4096
##  Mean   :  460.1   Mean   : 1.374   Mean   :4119
##  3rd Qu.:  311.0   3rd Qu.: 1.000   3rd Qu.:5790
##  Max.   :30817.0   Max.   :53.000   Max.   :8296
```

Explanation

You can read in the data and look at the summary with the following commands:

airlines = read.csv("AirlinesCluster.csv")

summary(airlines)

For the smallest values, BonusTrans and FlightTrans are on the scale of tens, whereas all other variables have values in the thousands.

For the largest values, Balance and BonusMiles have average values in the tens of thousands.

1.2) Normalizing the Data

In this problem, we will normalize our data before we run the clustering algorithms. Why is it important to normalize the data before clustering?

```
include_graphics('1.2.png')
```

- ○ If we don't normalize the data, the clustering algorit
- ○ If we don't normalize the data, it will be hard to inter
- ◉ If we don't normalize the data, the clustering will be
- ○ If we don't normalize the data, the clustering will be

1.3) Normalizing the Data

Let's go ahead and normalize our data. You can normalize the variables in a data frame by using the preProcess function in the "caret" package. You should already have this package installed from Week 4, but if not, go ahead and install it with install.packages("caret"). Then load the package with library(caret).

Now, create a normalized data frame called "airlinesNorm" by running the following commands:

```
pre_proc = preProcess(airlines)

airlines_norm = predict(pre_proc, airlines)

summary(airlines_norm)
```

```
##     Balance           QualMiles          BonusMiles         BonusTrans      
##  Min.   :-0.7303   Min.   :-0.1863   Min.   :-0.7099   Min.   :-1.20805  
##  1st Qu.:-0.5465   1st Qu.:-0.1863   1st Qu.:-0.6581   1st Qu.:-0.89568  
##  Median :-0.3027   Median :-0.1863   Median :-0.4130   Median : 0.04145  
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000  
##  3rd Qu.: 0.1866   3rd Qu.:-0.1863   3rd Qu.: 0.2756   3rd Qu.: 0.56208  
##  Max.   :16.1868   Max.   :14.2231   Max.   :10.2083   Max.   : 7.74673  
##   FlightMiles        FlightTrans        DaysSinceEnroll   
##  Min.   :-0.3286   Min.   :-0.36212   Min.   :-1.99336  
##  1st Qu.:-0.3286   1st Qu.:-0.36212   1st Qu.:-0.86607  
##  Median :-0.3286   Median :-0.36212   Median :-0.01092  
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000  
##  3rd Qu.:-0.1065   3rd Qu.:-0.09849   3rd Qu.: 0.80960  
##  Max.   :21.6803   Max.   :13.61035   Max.   : 2.02284  
```

The first command pre-processes the data, and the second command performs the normalization. If you look at the summary of airlinesNorm, you should see that all of the variables now have mean zero. You can also see that each of the variables has standard deviation 1 by using the sd() function.

In the normalized data, which variable has the largest maximum value?

In the normalized data, which variable has the smallest minimum value?

Explanation

After running the two lines of code to normalize the data, you can look at the summary of airlinesNorm with the command:

summary(airlinesNorm)

You can see from the output that FlightMiles now has the largest maximum value, and DaysSinceEnroll now has the smallest minimum value. Note that these were not the variables with the largest and smallest values in the original dataset airlines.

2.1) Hierarchical Clustering

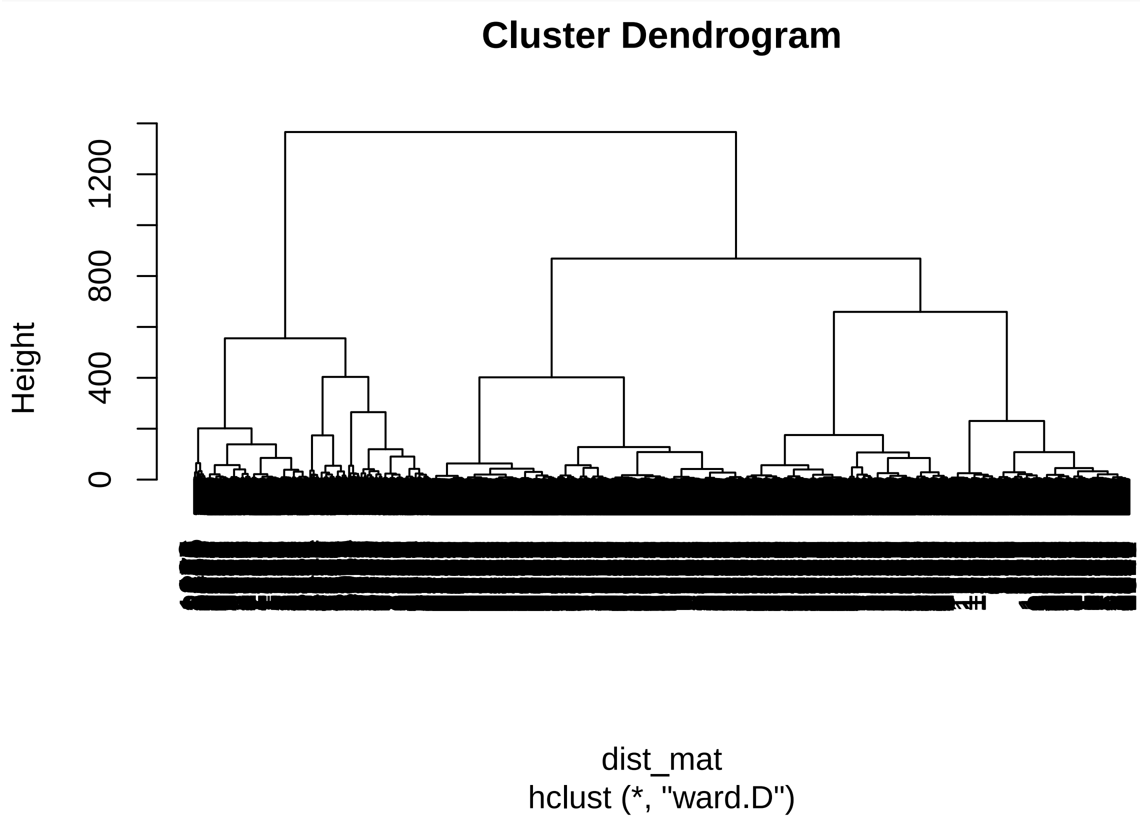
Compute the distances between data points (using euclidean distance) and then run the Hierarchical clustering algorithm (using method="ward.D") on the normalized data. It may take a few minutes for the commands to finish since the dataset has a large number of observations for hierarchical clustering.

```
dist_mat = dist(airlines_norm, method = 'euclidean')

clusters_hc = hclust(dist_mat, method = 'ward.D')
```

Then, plot the dendrogram of the hierarchical clustering process. Suppose the airline is looking for somewhere between 2 and 10 clusters. According to the dendrogram, which of the following is NOT a good choice for the number of clusters?

```
plot(clusters_hc)
```



```r
include_graphics('2.1.png')
```

- 2
- 3
- 6 ✔
- 7

Explanation

You can plot the dendrogram with the command:
plot(hierClust)
If you run a horizontal line down the dendrogram, you ca
clusters, 3 clusters, or 7 clusters. However, it it hard to se
clusters is probably not a good choice.

2.2) Hierarchical Clustering

Suppose that after looking at the dendrogram and discussing with the marketing department, the airline decides to proceed with 5 clusters. Divide the data points into 5 clusters by using the cutree function. How many data points are in Cluster 1?

```r
clusters = cutree(clusters_hc, 5)

table(clusters)
```

```
## clusters
```

```
##    1    2    3    4    5
##  776  519  494  868 1342
```

Explanation

You can divide the data points into 5 clusters with the following command:

clusterGroups = cutree(hierClust, k = 5)

If you type table(clusterGroups), you can see that there are 776 data points in the first cluster.

2.3) Hierarchical Clustering

Now, use tapply to compare the average values in each of the variables for the 5 clusters (the centroids of the clusters). You may want to compute the average values of the unnormalized data so that it is easier to interpret. You can do this for the variable "Balance" with the following command:

```r
tapply(airlines$Balance, clusters, mean)
```

```
##         1         2         3         4         5
##  57866.90 110669.27 198191.57  52335.91  36255.91
```

```r
tapply(airlines$QualMiles, clusters, mean)
```

```
##            1            2            3            4            5
##    0.6443299 1065.9826590   30.3461538    4.8479263    2.5111773
```

```r
tapply(airlines$BonusMiles, clusters, mean)
```

```
##         1         2         3         4         5
## 10360.124 22881.763 55795.860 20788.766  2264.788
```

```r
tapply(airlines$BonusTrans, clusters, mean)
```

```
##         1         2         3         4         5
## 10.823454 18.229287 19.663968 17.087558  2.973174
```

```r
tapply(airlines$FlightMiles, clusters, mean)
```

```
##          1          2          3          4          5
##   83.18428 2613.41811  327.67611  111.57373  119.32191
```

```r
tapply(airlines$FlightTrans, clusters, mean)
```

```
##         1         2         3         4         5
## 0.3028351 7.4026975 1.0688259 0.3444700 0.4388972
```

```r
tapply(airlines$DaysSinceEnroll, clusters, mean)
```

```
##        1        2        3        4        5
## 6235.365 4402.414 5615.709 2840.823 3060.081
```

Compared to the other clusters, Cluster 1 has the largest average values in which variables (if any)?

```r
include_graphics('2.3.a.png')
```

- [ ] Balance
- [ ] QualMiles
- [ ] BonusMiles
- [ ] BonusTrans
- [ ] FlightMiles
- [ ] FlightTrans
- [x] DaysSinceEnroll ✔
- [ ] None

How would you describe the customers in Cluster 1?

```
include_graphics('2.3.b.png')
```

- ○ Relatively new customers who don't use the airline v
- ● Infrequent but loyal customers. ✔
- ○ Customers who have accumulated a large amount o
- ○ Customers who have accumulated a large amount o
transactions.
- ○ Relatively new customers who seem to be accumula

2.4) Hierarchical Clustering

Compared to the other clusters, Cluster 2 has the largest average values in which variables (if any)? Select all that apply.

```
tapply(airlines$Balance, clusters, mean)
```

```
##         1         2         3         4         5
##  57866.90 110669.27 198191.57  52335.91  36255.91
```

```
tapply(airlines$QualMiles, clusters, mean)
```

```
##            1            2            3            4            5
##    0.6443299 1065.9826590   30.3461538    4.8479263    2.5111773
```

```
tapply(airlines$BonusMiles, clusters, mean)
```

```
##         1         2         3         4         5
## 10360.124 22881.763 55795.860 20788.766  2264.788
```

```
tapply(airlines$BonusTrans, clusters, mean)
```

```
##         1         2         3         4         5
## 10.823454 18.229287 19.663968 17.087558  2.973174
```

```
tapply(airlines$FlightMiles, clusters, mean)
```

```
##          1           2           3           4           5
```

```
##   83.18428 2613.41811  327.67611  111.57373  119.32191
```

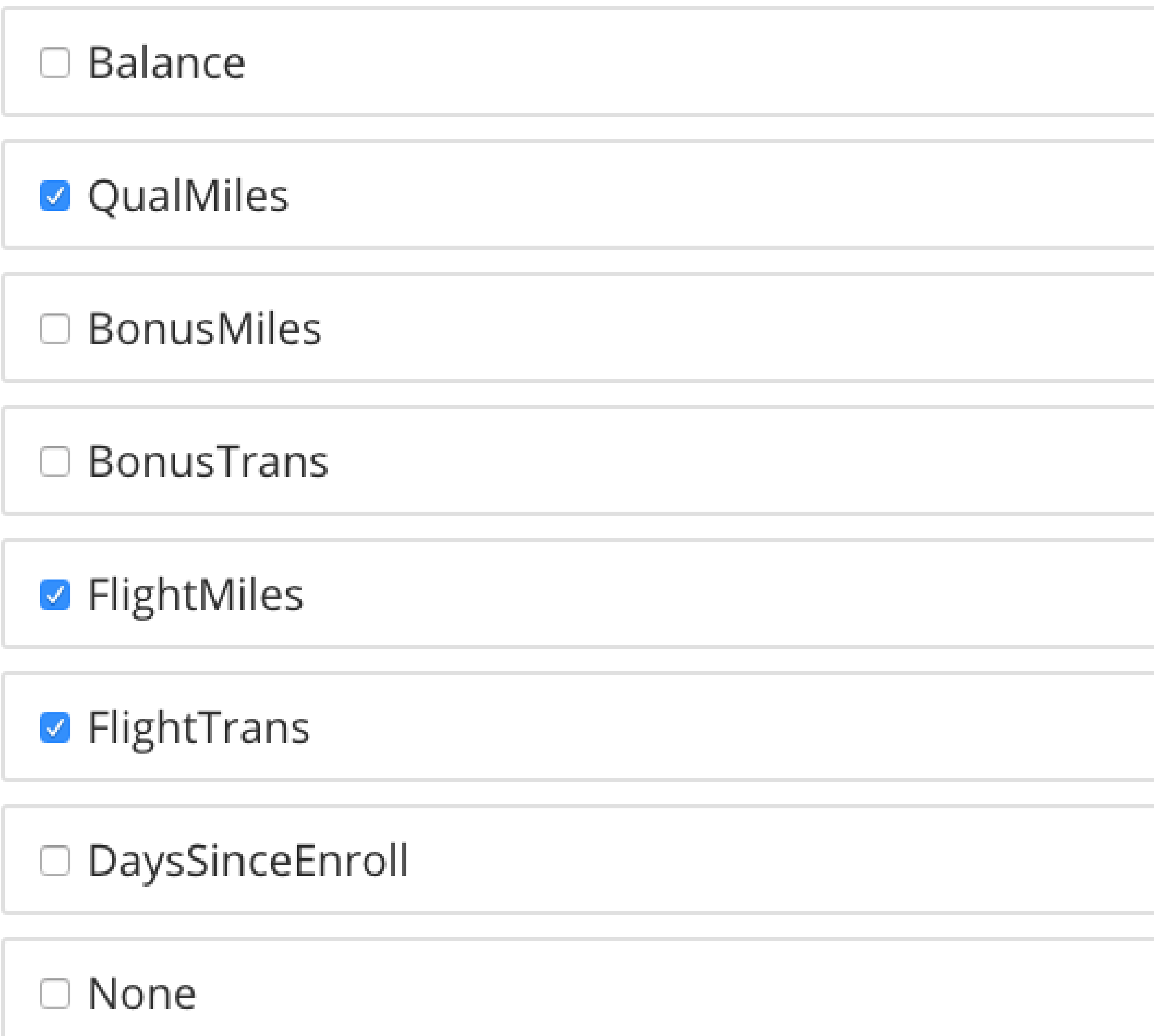
```r
tapply(airlines$FlightTrans, clusters, mean)
```

```
##         1         2         3         4         5 
## 0.3028351 7.4026975 1.0688259 0.3444700 0.4388972
```

```r
tapply(airlines$DaysSinceEnroll, clusters, mean)
```

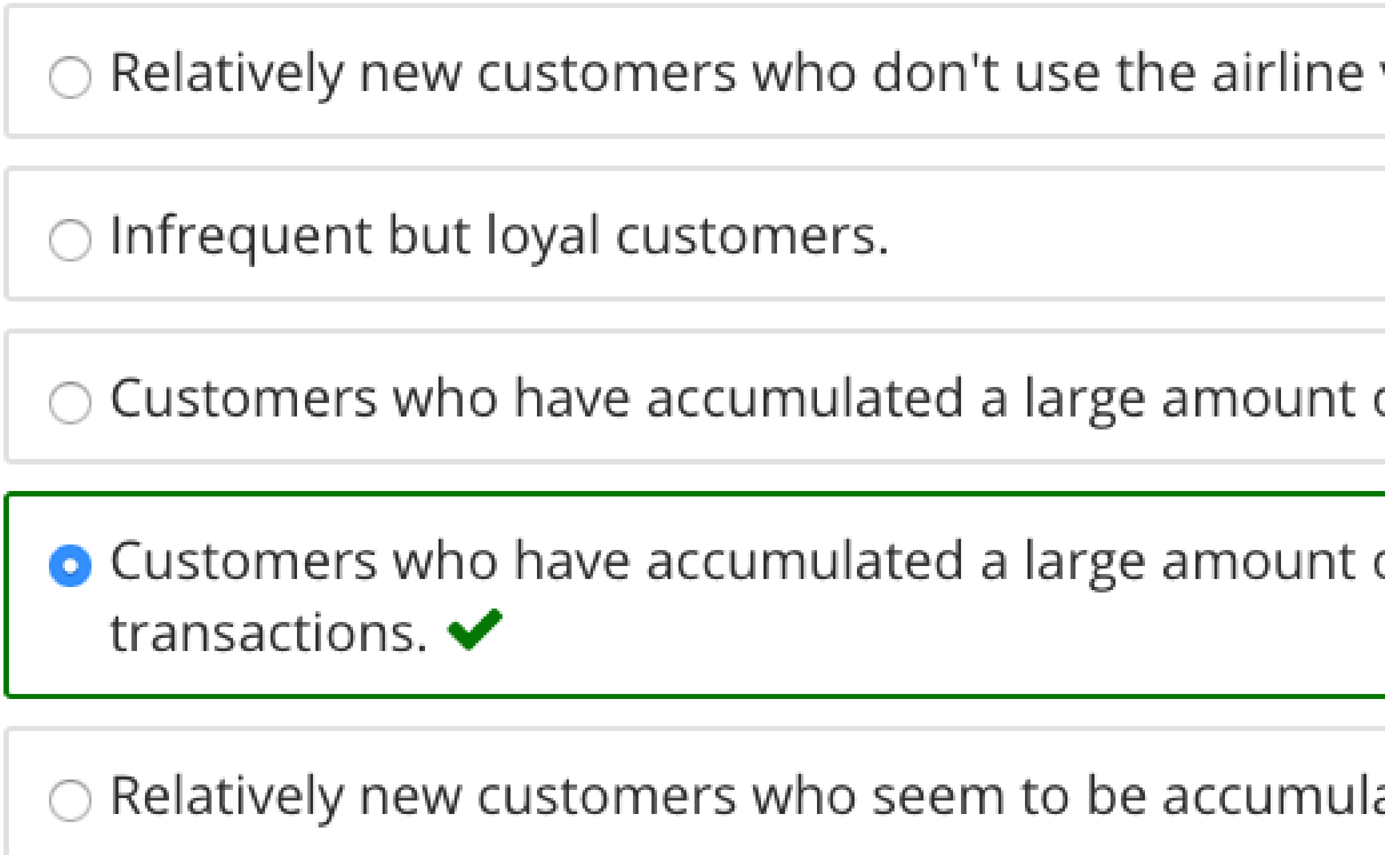
```
##        1        2        3        4        5 
## 6235.365 4402.414 5615.709 2840.823 3060.081
```

```r
include_graphics('2.4.a.png')
```



How would you describe the customers in Cluster 2?

```r
include_graphics('2.4.b.png')
```



2.5) Hierarchical Clustering

Compared to the other clusters, Cluster 3 has the largest average values in which variables (if any)? Select all that apply.

```r
tapply(airlines$Balance, clusters, mean)
```

```
##         1         2         3         4         5 
##  57866.90 110669.27 198191.57  52335.91  36255.91
```

```r
tapply(airlines$QualMiles, clusters, mean)
```

```
##            1            2            3            4            5 
##    0.6443299 1065.9826590   30.3461538    4.8479263    2.5111773
```

```r
tapply(airlines$BonusMiles, clusters, mean)
```

```
##         1         2         3         4         5 
## 10360.124 22881.763 55795.860 20788.766  2264.788
```

```r
tapply(airlines$BonusTrans, clusters, mean)
```

```
##         1         2         3         4         5 
## 10.823454 18.229287 19.663968 17.087558  2.973174
```

```r
tapply(airlines$FlightMiles, clusters, mean)
```

```
##          1          2          3          4          5 
##   83.18428 2613.41811  327.67611  111.57373  119.32191
```

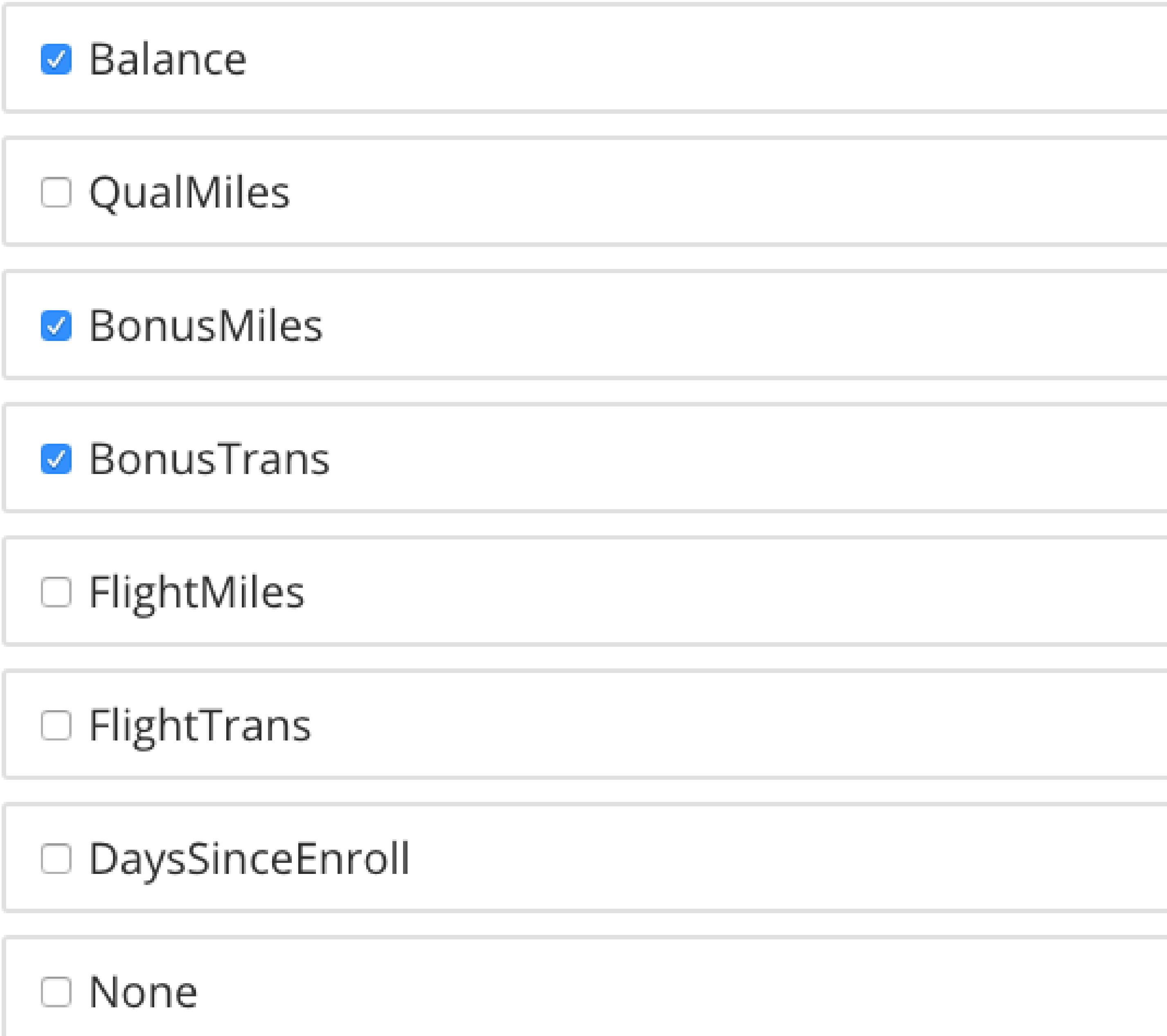
```r
tapply(airlines$FlightTrans, clusters, mean)
```

```
##         1         2         3         4         5 
## 0.3028351 7.4026975 1.0688259 0.3444700 0.4388972
```

```r
tapply(airlines$DaysSinceEnroll, clusters, mean)
```

```
##        1        2        3        4        5 
## 6235.365 4402.414 5615.709 2840.823 3060.081
```

```r
include_graphics('2.5.a.png')
```



- [x] Balance
- [ ] QualMiles
- [x] BonusMiles
- [x] BonusTrans
- [ ] FlightMiles
- [ ] FlightTrans
- [ ] DaysSinceEnroll
- [ ] None

How would you describe the customers in Cluster 3?

```
include_graphics('2.5.b.png')
```

- Relatively new customers who don't use the airline ve
- Infrequent but loyal customers.
- Customers who have accumulated a large amount of
- Customers who have accumulated a large amount of transactions.
- Relatively new customers who seem to be accumulat

2.6) Hierarchical Clustering

Compared to the other clusters, Cluster 4 has the largest average values in which variables (if any)? Select all that apply.

```r
tapply(airlines$Balance, clusters, mean)
```

```
##         1         2         3         4         5
##  57866.90 110669.27 198191.57  52335.91  36255.91
```

```r
tapply(airlines$QualMiles, clusters, mean)
```

```
##            1            2            3            4            5
##    0.6443299 1065.9826590   30.3461538    4.8479263    2.5111773
```

```r
tapply(airlines$BonusMiles, clusters, mean)
```

```
##         1         2         3         4         5
## 10360.124 22881.763 55795.860 20788.766  2264.788
```

```r
tapply(airlines$BonusTrans, clusters, mean)
```

```
##         1         2         3         4         5
## 10.823454 18.229287 19.663968 17.087558  2.973174
```

```r
tapply(airlines$FlightMiles, clusters, mean)
```

```
##          1          2          3          4          5
```

```
##   83.18428 2613.41811  327.67611  111.57373  119.32191
```

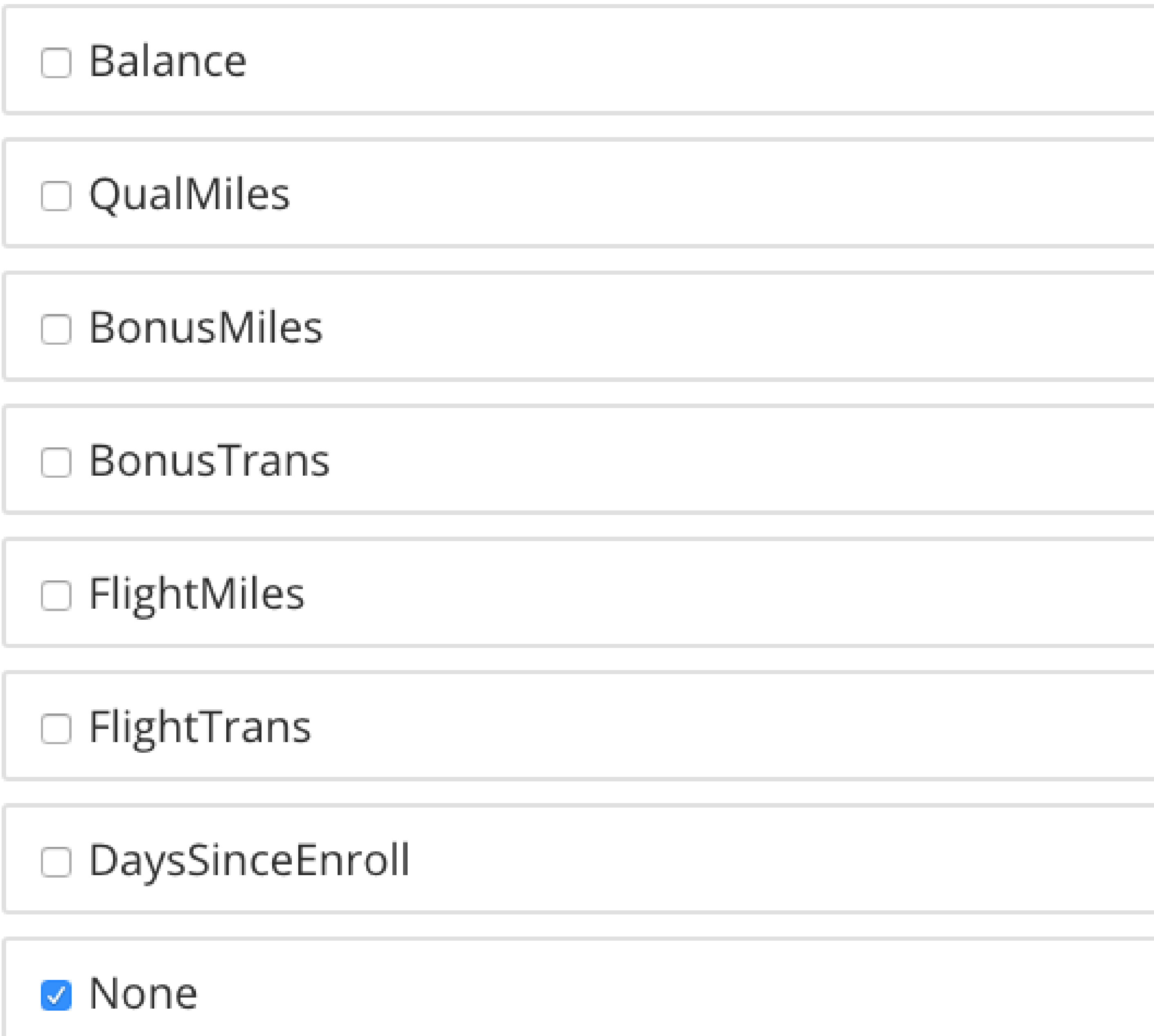
```r
tapply(airlines$FlightTrans, clusters, mean)
```

```
##         1         2         3         4         5
## 0.3028351 7.4026975 1.0688259 0.3444700 0.4388972
```

```r
tapply(airlines$DaysSinceEnroll, clusters, mean)
```

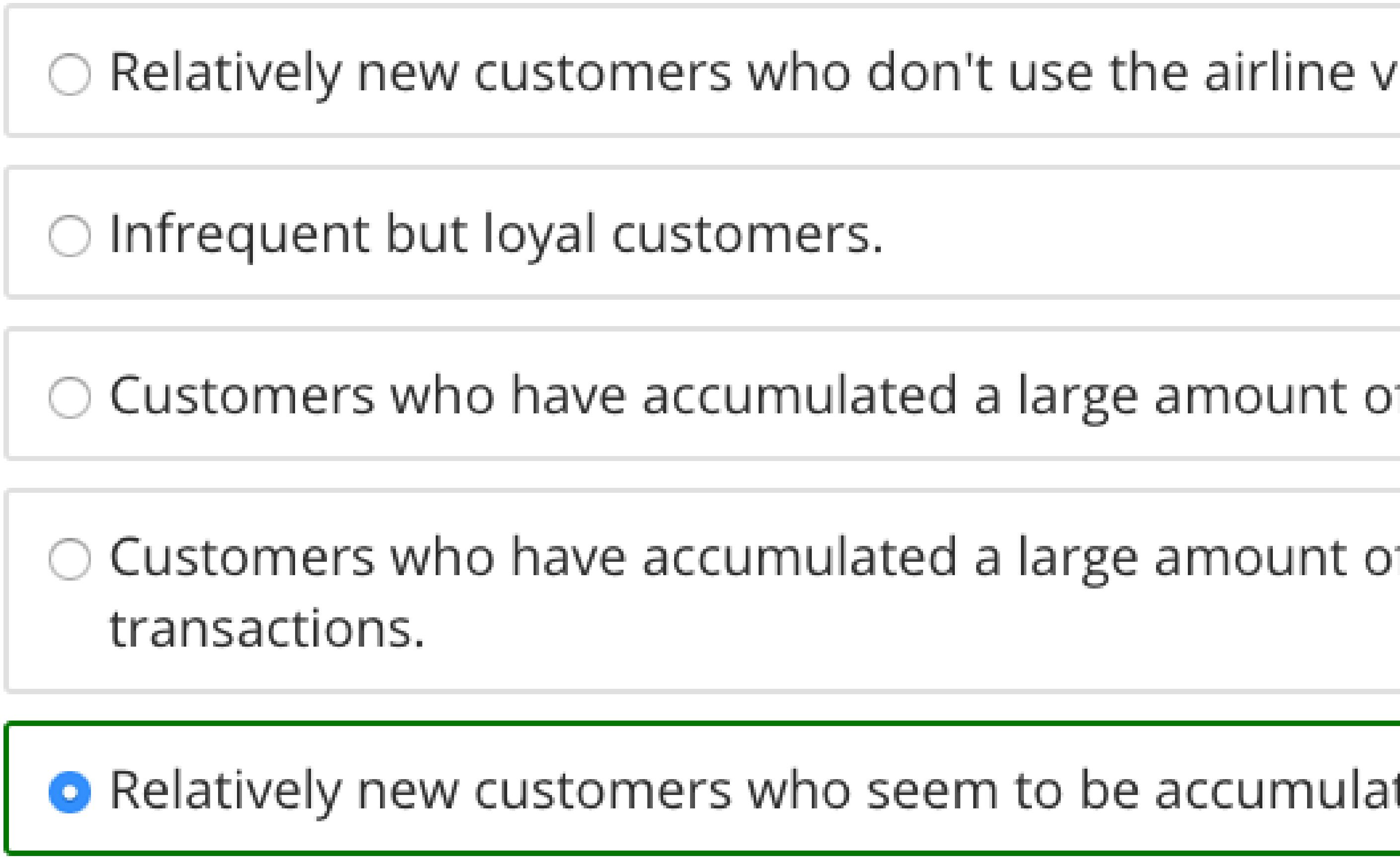
```
##        1        2        3        4        5
## 6235.365 4402.414 5615.709 2840.823 3060.081
```

```r
include_graphics('2.6.a.png')
```



How would you describe the customers in Cluster 4?

```r
include_graphics('2.6.b.png')
```



2.7) Hierarchical Clustering

Compared to the other clusters, Cluster 5 has the largest average values in which variables (if any)? Select all that apply.

```r
tapply(airlines$Balance, clusters, mean)
```

```
##         1         2         3         4         5 
##  57866.90 110669.27 198191.57  52335.91  36255.91
```

```r
tapply(airlines$QualMiles, clusters, mean)
```

```
##            1            2            3            4            5 
##    0.6443299 1065.9826590   30.3461538    4.8479263    2.5111773
```

```r
tapply(airlines$BonusMiles, clusters, mean)
```

```
##         1         2         3         4         5 
## 10360.124 22881.763 55795.860 20788.766  2264.788
```

```r
tapply(airlines$BonusTrans, clusters, mean)
```

```
##         1         2         3         4         5 
## 10.823454 18.229287 19.663968 17.087558  2.973174
```

```r
tapply(airlines$FlightMiles, clusters, mean)
```

```
##          1          2          3          4          5
##   83.18428 2613.41811  327.67611  111.57373  119.32191
```

```r
tapply(airlines$FlightTrans, clusters, mean)
```

```
##         1         2         3         4         5
## 0.3028351 7.4026975 1.0688259 0.3444700 0.4388972
```

```r
tapply(airlines$DaysSinceEnroll, clusters, mean)
```

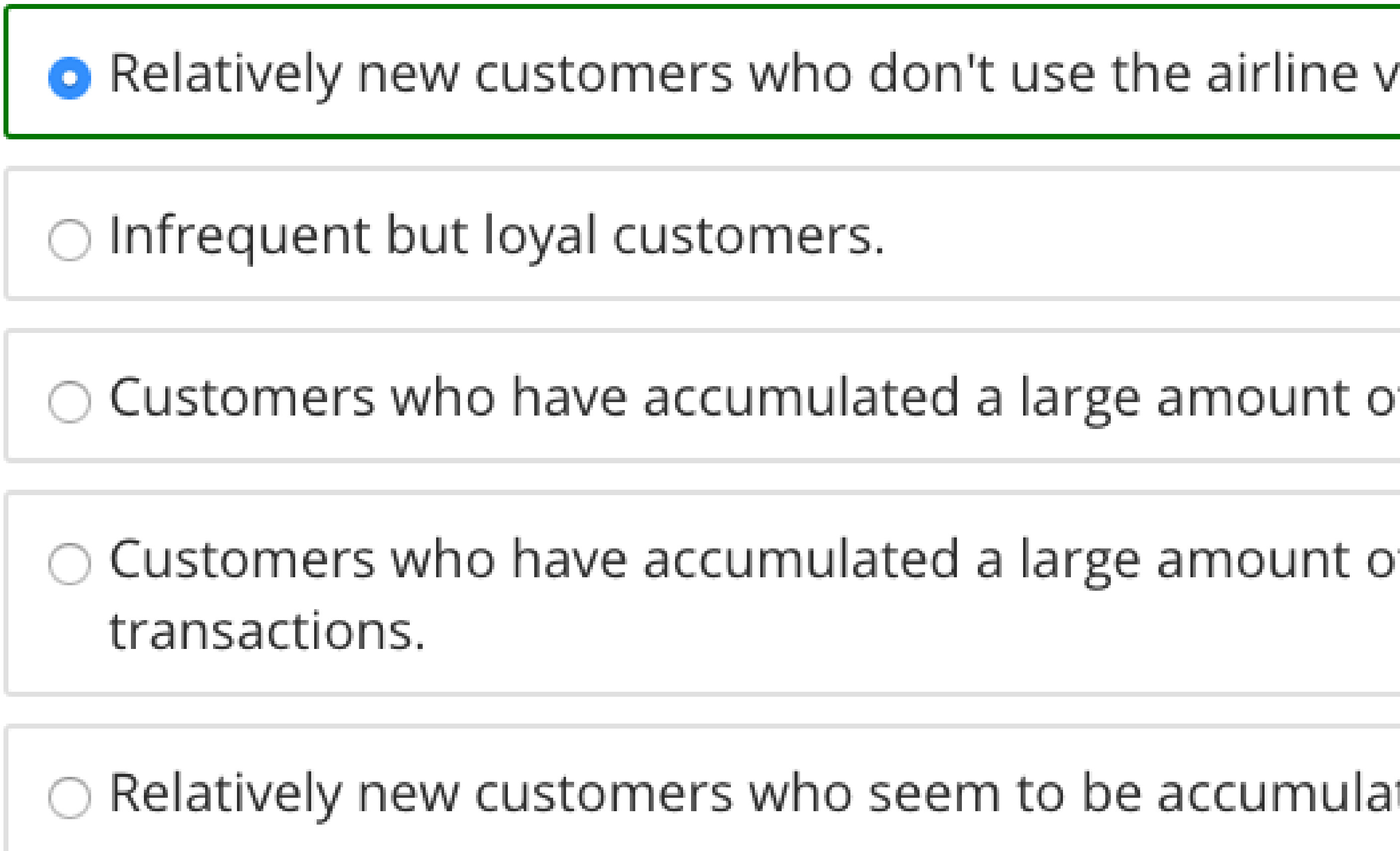
```
##        1        2        3        4        5
## 6235.365 4402.414 5615.709 2840.823 3060.081
```

```r
include_graphics('2.7.a.png')
```

- [ ] Balance
- [ ] QualMiles
- [ ] BonusMiles
- [ ] BonusTrans
- [ ] FlightMiles
- [ ] FlightTrans
- [ ] DaysSinceEnroll
- [x] None

How would you describe the customers in Cluster 5?

```
include_graphics('2.7.b.png')
```



3.1) K-Means Clustering

Now run the k-means clustering algorithm on the normalized data, again creating 5 clusters. Set the seed to 88 right before running the clustering algorithm, and set the argument iter.max to 1000.

```
set.seed(88)

clusters_k = kmeans(airlines_norm, centers = 5, iter.max = 1000)
```

How many clusters have more than 1,000 observations?

```
table(clusters_k$cluster)
```

```
##
##    1    2    3    4    5
##  408  141  993 1182 1275
```

Explanation

You can run the k-means clustering algorithm with the following commands:

set.seed(88)

kmeansClust = kmeans(airlinesNorm, centers=5, iter.max=1000)

And you can look at the number of observations in each cluster with the following command:

table(kmeansClust$cluster)

There are two clusters with more than 1000 observations.

3.2) K-Means Clustering

Now, compare the cluster centroids to each other either by dividing the data points into groups and then using tapply, or by looking at the output of kmeansClust$centers, where"kmeansClust"isthenameoftheoutputofthekmeansfunctio$ will be for the normalized data. If you want to look at the average values for the unnormalized data, you need to use tapply like we did for hierarchical clustering.)

```
clusters_no = cutree(clusters_hc, 5)
```

```
table(clusters_no, clusters_k$cluster)
```

```
##
## clusters_no    1    2    3    4    5
##           1    4    0   98  673    1
##           2   92  137  105   92   93
##           3  300    4  132   58    0
##           4   12    0  653   30  173
##           5    0    0    5  329 1008
```

```
clusters_k$centers
```

```
##       Balance   QualMiles BonusMiles BonusTrans FlightMiles FlightTrans
## 1  1.44439706  0.51115730  1.8769284  1.0331951   0.1169945   0.1444636
## 2  1.00054098  0.68382234  0.6144780  1.7214887   3.8559798   4.1196141
## 3 -0.05580605 -0.14104391  0.3041358  0.7108744  -0.1218278  -0.1287569
## 4 -0.13331742 -0.11491607 -0.3492669 -0.3373455  -0.1833989  -0.1961819
## 5 -0.40579897 -0.02281076 -0.5816482 -0.7619054  -0.1989602  -0.2196582
##   DaysSinceEnroll
## 1       0.7198040
## 2       0.2742394
## 3      -0.3398209
## 4       0.9640923
## 5      -0.8897747
```

Do you expect Cluster 1 of the K-Means clustering output to necessarily be similar to Cluster 1 of the Hierarchical clustering output?

```
include_graphics('3.2.png')
```

- ○ Yes, because the clusters are displayed in order of siz
- ○ Yes, because the clusters are displayed according to t
  similar.
- ◉ No, because cluster ordering is not meaningful in eith
- ○ No, because the clusters produced by the k-means al
  the Hierarchical algorithm.

Explanation

The clusters are not displayed in a meaningful order, so w
algorithm that is similar to Cluster 1 produced by the Hier