

# Modeling in Revolution R Enterprise

## Module 9: K-Means Clustering Solutions

July 23, 2014







```
infile <- file.path("data", "BankXDF.xdf")  
BankDS <- RxXdfData(file = infile)
```



# Exercise: Solution



```
Kmeans2 <- rxKmeans(~age + day, data = BankDS, numClusters = 3)
```

```
## Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.002 seconds
## Rows Read: 10000, Total Rows Processed: 20000, Total Chunk Time: 0.001 seconds
## Rows Read: 10000, Total Rows Processed: 30000, Total Chunk Time: Less than .001 seconds
## Rows Read: 10000, Total Rows Processed: 40000, Total Chunk Time: 0.001 seconds
## Rows Read: 5211, Total Rows Processed: 45211, Total Chunk Time: Less than .001 seconds
## Rows Read: 10000, Total Rows Processed: 10000, Total Chunk Time: 0.002 seconds
...
```

```
Kmeans2
```

```
## Call:
## rxKmeans(formula = ~age + day, data = BankDS, numClusters = 3)
##
## Data: BankDS
## Number of valid observations: 45211
## Number of missing observations: 0
...
```

# Exercise: Solution



From the previous output, both variables have been divided into three distinct categories, as defined in the function call:

```
Kmeans2$centers
```

```
##      age      day  
## 1 35.09   8.263  
## 2 34.94  22.951  
## 3 53.73  15.813
```

Finally, the within cluster sum of squares (i.e. distance from the mean) for each cluster is:

```
Kmeans2$withinss
```

```
##      1      2      3  
## 772997 849332 1513565
```

# Thank you

Revolution Analytics is the leading commercial provider of software and support for the popular open source R statistics language.

[www.revolutionanalytics.com](http://www.revolutionanalytics.com), 1.855.GET.REVO, Twitter: @RevolutionR

