K Means Clustering

Davis Nyabuti

November 17, 2016

SEIS734-02

HW - 8

1 Load Data

The dataset contains more than 9,211,031 NYSE trade data.

The first step was to extract `NYSE_DM.7z` and save the extracted file in the same directory with this script.

```
In [1]: data <- read.csv("NYSE_DM.csv", header = FALSE)
        head(data)
```

	V1	V2	V3	V4	V5	V6	V7
157801	25	25	25	25	0	1.32
279752	25	25	25	25	0	6.34
346856	25	25	25	25	0	4.96
347167	25	25	25	25	0	4.62
347169	25	25	25	25	0	4.62
347170	25	25	25	25	0	4.62

Add names to the dataframe

```
In [2]: names(data) <- c("ID","OPEN_P","HIGH_P","LOW_P"
                         ,"CLOSE_P","VOLUME","CLOSE_ADJ_P")
        head(data)
```

	ID	OPEN_P	HIGH_P	LOW_P	CLOSE_P	VOLUME	CLOSE_ADJ_P
157801	25	25	25	25	0	1.32
279752	25	25	25	25	0	6.34
346856	25	25	25	25	0	4.96
347167	25	25	25	25	0	4.62
347169	25	25	25	25	0	4.62
347170	25	25	25	25	0	4.62

Use columns 2 to 7 from the input data and perform the k-means clustering with k = 4. If your tool allows you to control the maximum number of iterations, set the maximum number of iterations to 10,000.

2 Generate Models

2.1 k = 4

```
In [3]: # Set seed for reproducibility
        set.seed(20)
        start.time <- Sys.time()
        nyse4Cluster <- kmeans(x = data[, 2:7]
                               ,centers = 4
                               ,iter.max = 10000
                               ,nstart = 1)
        end.time = Sys.time()
        time.taken = end.time - start.time
        cat("Duration ", time.taken, " seconds")
```

```
Duration  13.65237  seconds
```

1. Output the final four centers that were generated from this clustering process.

```
In [4]: nyse4Cluster$centers
```

	OPEN_P	HIGH_P	LOW_P	CLOSE_P	VOLUME	CLOSE_ADJ_P
1	31.09257	31.93051	30.11115	31.03209	62711915.7	22.10730
2	28.21482	28.56833	27.91018	28.26443	500931.3	19.02216
3	42.84503	43.63152	42.02259	42.84677	10420542.3	24.65419
4	11.22353	11.76984	10.63157	11.23396	393693260.4	10.44012

2.2 k = 200

2. Perform the same clustering task with the same parameters except setting k= 200.

```
In [5]: # Set seed for reproducibility
        set.seed(20)
        start.time <- Sys.time()
        nyse200Cluster <- kmeans(x = data[, 2:7]
                               ,centers = 200
                               ,iter.max = 10000
                               ,nstart = 1)
        end.time = Sys.time()
        time.taken = end.time - start.time
        cat("Duration ", time.taken, " seconds")
```

```
Warning message:
"Quick-TRANSfer stage steps exceeded maximum (= 460551550)"

Duration  26.09723  seconds
```