CSIS734-01 Data Mining & Predictive Analytics
Garth Mortensen, mort0052@stthomas.edu

Graduate Program in Software

CSIS 734-01: Data Mining & Predictive Analytics

Assignment #8 (100 points)

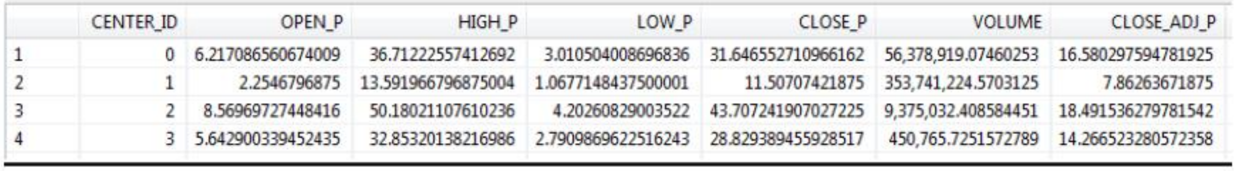
Due Date: May 12th, 2018

Perform k-means clustering of the NYSE dataset that has more than 9,211,031 NYSE trade data. The dataset will be placed on the Blackboard. Columns from 1 to 7 are:

1	ID	INTEGER,	record ID
2	OPEN_P	DOUBLE,	open price
3	HIGH_P	DOUBLE,	highest price
4	LOW_P	DOUBLE,	lowest price
5	CLOSE_P	DOUBLE,	close price
6	VOLUME	DOUBLE,	volume
7	CLOSE_ADJ_P	DOUBLE	close adjusted price

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.

__Question 1__: Carefully check the data before you perform the clustering task 1. Output the final four centers that were generated from this clustering process. You should output those four centers in the format shown in the following table.

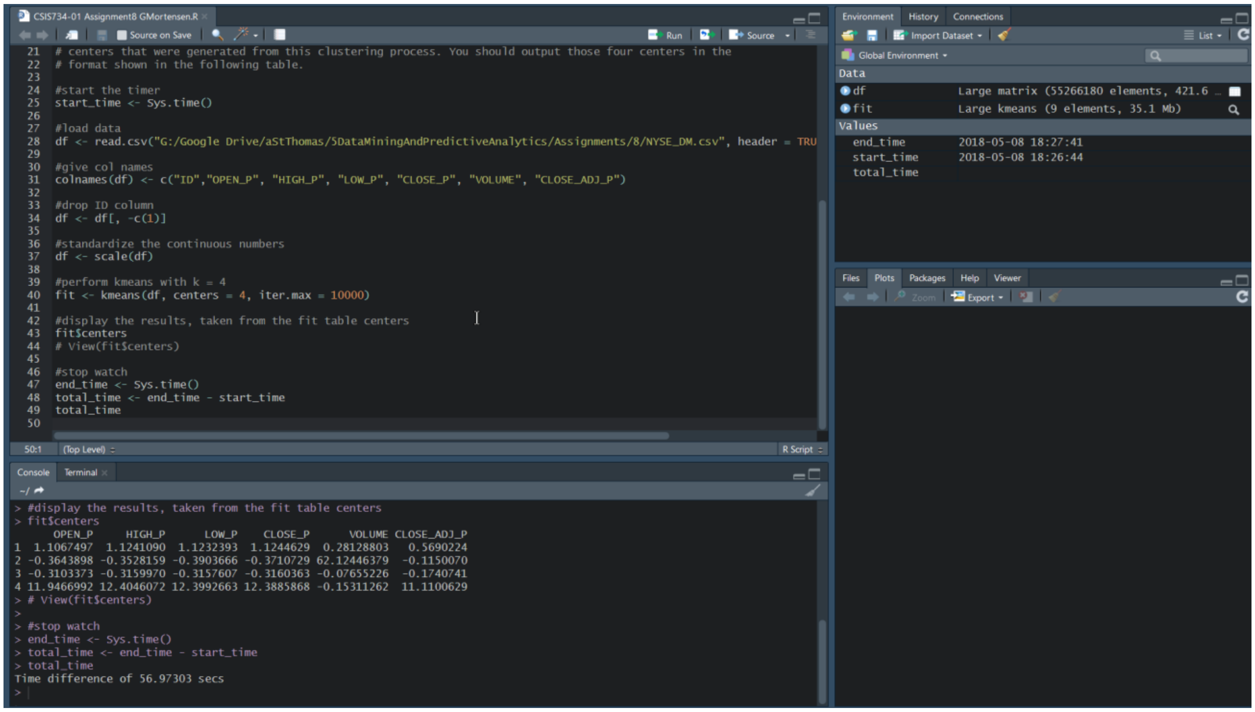


	CENTER_ID	OPEN_P	HIGH_P	LOW_P	CLOSE_P	VOLUME	CLOSE_ADJ_P
1	0	6.217086560674009	36.71222557412692	3.010504008696836	31.646552710966162	56,378,919.07460253	16.580297594781925
2	1	2.2546796875	13.591966796875004	1.0677148437500001	11.50707421875	353,741,224.5703125	7.86263671875
3	2	8.56969727448416	50.18021107610236	4.20260829003522	43.707241907027225	9,375,032.408584451	18.491536279781542
4	3	5.642900339452435	32.85320138216986	2.7909869622516243	28.829389455928517	450,765.7251572789	14.266523280572358

__Question 2__: Take a screenshot of your output and the execution time of the above clustering task. Put your screenshot and your code in a WORD document.

1. Please submit your WORD document and your code to clai@stthomas.edu.

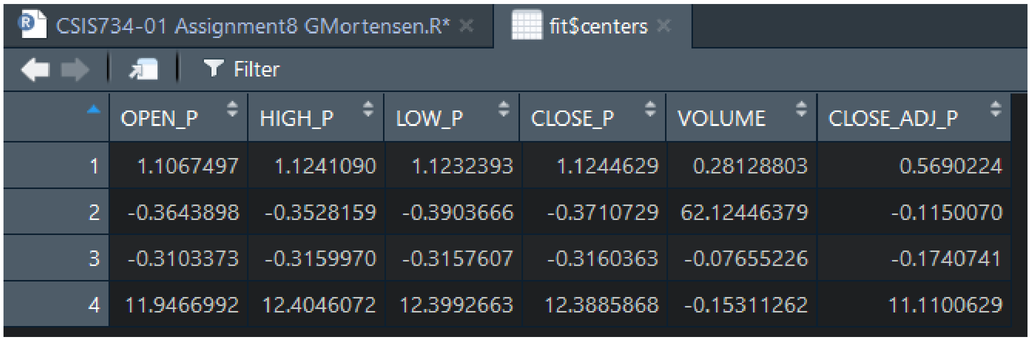
2. Print and submit the hardcopy of your WORD document in the class on the due date.



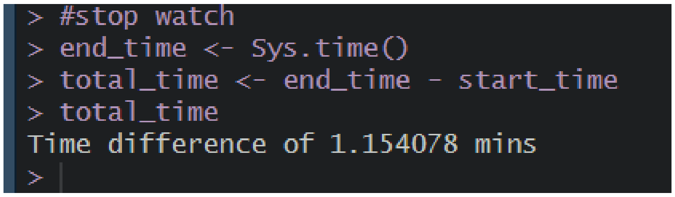
```
# centers that were generated from this clustering process. You should output those four centers in the
# format shown in the following table.

#start the timer
start_time <- Sys.time()

#load data
df <- read.csv("G:/Google Drive/aStThomas/5DataMiningAndPredictiveAnalytics/Assignments/8/NYSE_DM.csv", header = TRU

#give col names
colnames(df) <- c("ID","OPEN_P", "HIGH_P", "LOW_P", "CLOSE_P", "VOLUME", "CLOSE_ADJ_P")

#drop ID column
df <- df[, -c(1)]

#standardize the continuous numbers
df <- scale(df)

#perform kmeans with k = 4
fit <- kmeans(df, centers = 4, iter.max = 10000)

#display the results, taken from the fit table centers
fit$centers
# View(fit$centers)

#stop watch
end_time <- Sys.time()
total_time <- end_time - start_time
total_time
```

```
> #display the results, taken from the fit table centers
> fit$centers
      OPEN_P     HIGH_P      LOW_P    CLOSE_P      VOLUME CLOSE_ADJ_P
1  1.1067497  1.1241090  1.1232393  1.1244629  0.28128803   0.5690224
2 -0.3643898 -0.3528159 -0.3903666 -0.3710729 62.12446379  -0.1150070
3 -0.3103373 -0.3159970 -0.3157607 -0.3160363 -0.07655226  -0.1740741
4 11.9466992 12.4046072 12.3992663 12.3885868 -0.15311262  11.1100629
> # View(fit$centers)
>
> #stop watch
> end_time <- Sys.time()
> total_time <- end_time - start_time
> total_time
Time difference of 56.97303 secs
>
```

```
Data
df          Large matrix (55266180 elements, 421.6 …
fit         Large kmeans (9 elements, 35.1 Mb)
Values
end_time    2018-05-08 18:27:41
start_time  2018-05-08 18:26:44
total_time
```



	OPEN_P	HIGH_P	LOW_P	CLOSE_P	VOLUME	CLOSE_ADJ_P
1	1.1067497	1.1241090	1.1232393	1.1244629	0.28128803	0.5690224
2	-0.3643898	-0.3528159	-0.3903666	-0.3710729	62.12446379	-0.1150070
3	-0.3103373	-0.3159970	-0.3157607	-0.3160363	-0.07655226	-0.1740741
4	11.9466992	12.4046072	12.3992663	12.3885868	-0.15311262	11.1100629



```
> #stop watch
> end_time <- Sys.time()
> total_time <- end_time - start_time
> total_time
Time difference of 1.154078 mins
>
```

```
#CSIS734-01 Data Mining & Predictive Analytics

#Garth Mortensen, mort0052@stthomas.edu

#Assignment 8, large dataset


# Perform k-means clustering of the NYSE dataset that has more than 9,211,031 NYSE trade data. The
# dataset will be placed on the Blackboard. Columns from 1 to 7 are:


# 1 ID          INTEGER   record ID
# 2 OPEN_P      DOUBLE    open price
# 3 HIGH_P      DOUBLE    highest price
# 4 LOW_P       DOUBLE    lowest price
# 5 CLOSE_P     DOUBLE    close price
# 6 VOLUME      DOUBLE    volume
# 7 CLOSE_ADJ_P DOUBLE    close adjusted price


# 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.


#start the timer
start_time <- Sys.time()


#load data
df <- read.csv("G:/Google Drive/aStThomas/5DataMiningAndPredictiveAnalytics/Assignments/8/NYSE_DM.csv", header = TRUE, stringsAsFactors = TRUE)


#give col names
colnames(df) <- c("ID","OPEN_P", "HIGH_P", "LOW_P", "CLOSE_P", "VOLUME", "CLOSE_ADJ_P")
```

```
#drop ID column
df <- df[, -c(1)]

#standardize the continuous numbers
df <- scale(df)

#perform kmeans with k = 4
fit <- kmeans(df, centers = 4, iter.max = 10000)

#display the results, taken from the fit table centers
fit$centers
# View(fit$centers)

#stop watch
end_time <- Sys.time()
total_time <- end_time - start_time
total_time
```