RDataMining

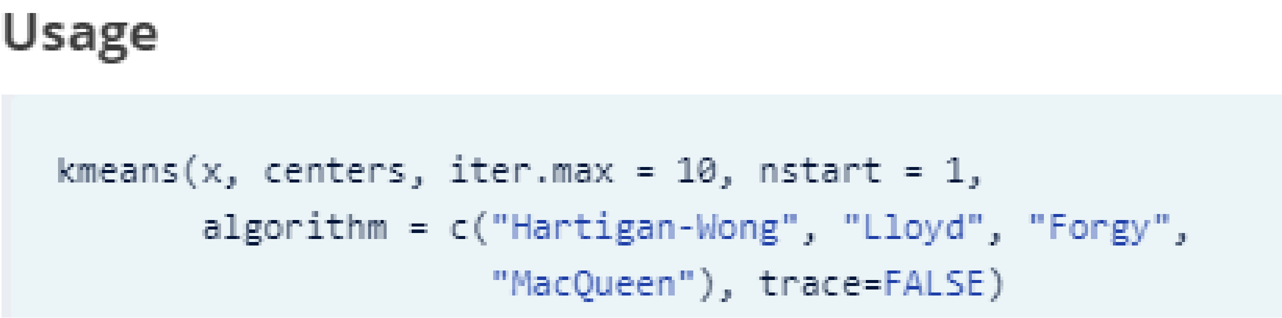
Week 8 Lab in R: K-Means Clustering

Dr. Liang (Leon) Chen

Instructions

- R script for K-Means Clustering is provided in the next slide, please follow it and complete the lab in R.
- You do not need to type notes (starting at #), but it's a good manner to have them in you script.
- In order to see codes and notes clearly, I show the script in RStudio.

- **The kmeans function will be used in this lab. For details, please visit this site.**

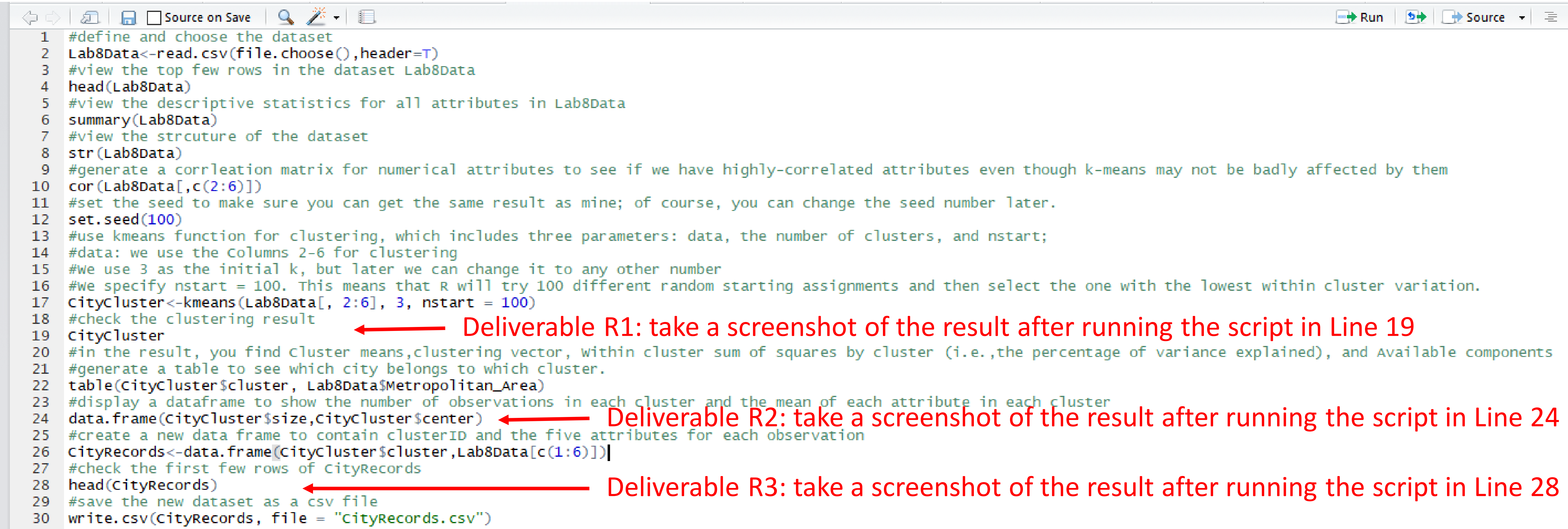


Usage

```
kmeans(x, centers, iter.max = 10, nstart = 1,
       algorithm = c("Hartigan-Wong", "Lloyd", "Forgy",
                     "MacQueen"), trace=FALSE)
```

The algorithm of Hartigan and Wong (1979) is used by default.

R Script for K-Means Clustering



```
1  #define and choose the dataset
2  Lab8Data<-read.csv(file.choose(),header=T)
3  #view the top few rows in the dataset Lab8Data
4  head(Lab8Data)
5  #view the descriptive statistics for all attributes in Lab8Data
6  summary(Lab8Data)
7  #view the strcuture of the dataset
8  str(Lab8Data)
9  #generate a corrleation matrix for numerical attributes to see if we have highly-correlated attributes even though k-means may not be badly affected by them
10 cor(Lab8Data[,c(2:6)])
11 #set the seed to make sure you can get the same result as mine; of course, you can change the seed number later.
12 set.seed(100)
13 #use kmeans function for clustering, which includes three parameters: data, the number of clusters, and nstart;
14 #data: we use the Columns 2-6 for clustering
15 #We use 3 as the initial k, but later we can change it to any other number
16 #We specify nstart = 100. This means that R will try 100 different random starting assignments and then select the one with the lowest within cluster variation.
17 CityCluster<-kmeans(Lab8Data[, 2:6], 3, nstart = 100)
18 #check the clustering result
19 CityCluster
20 #in the result, you find Cluster means,clustering vector, Within cluster sum of squares by cluster (i.e.,the percentage of variance explained), and Available components
21 #generate a table to see which city belongs to which cluster.
22 table(CityCluster$cluster, Lab8Data$Metropolitan_Area)
23 #display a dataframe to show the number of observations in each cluster and the mean of each attribute in each cluster
24 data.frame(CityCluster$size,CityCluster$center)
25 #create a new data frame to contain clusterID and the five attributes for each observation
26 CityRecords<-data.frame(CityCluster$cluster,Lab8Data[c(1:6)])
27 #check the first few rows of CityRecords
28 head(CityRecords)
29 #save the new dataset as a csv file
30 write.csv(CityRecords, file = "CityRecords.csv")
```

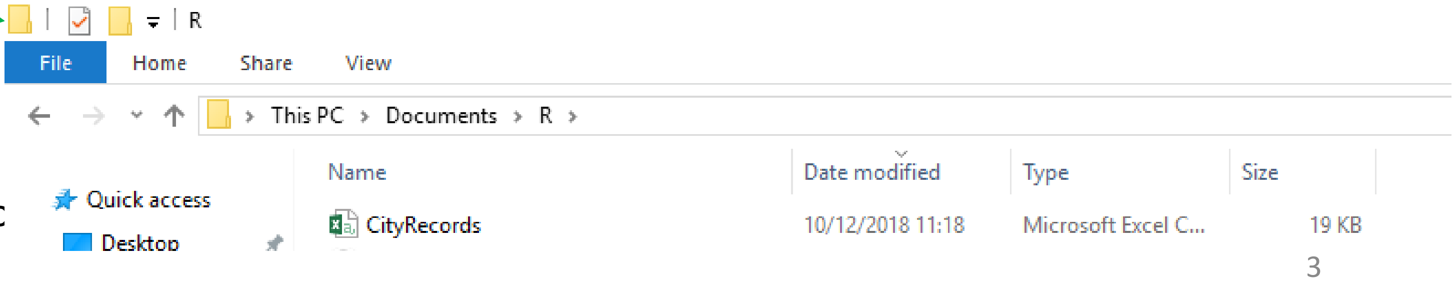
Deliverable R1: take a screenshot of the result after running the script in Line 19

Deliverable R2: take a screenshot of the result after running the script in Line 24

Deliverable R3: take a screenshot of the result after running the script in Line 28

Typically, the csv file is under the Documents folder or the folder you assign (I changed mine to Documents->R)

If you want to change the default working directory (folder) in R and RStudio, please check the first document in How-tc Files folder on WT Class.



Name	Date modified	Type	Size
CityRecords	10/12/2018 11:18	Microsoft Excel C...	19 KB

Deliverables

- Deliverable R1: take a screenshot of the result after running the script in Line 19 with date and time and briefly interpret the result.

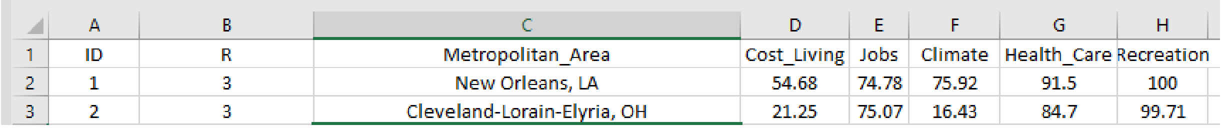
- Deliverable R2: take a screenshot of the result after running the script in Line 24 with date and time and briefly interpret the result.

- Deliverable R3: take a screenshot of the result after running the script in Line 28 with date and time and briefly interpret the result.

- Deliverable R4: Compare the clustering result for each observation in R (which is saved in CityRecords.csv) and that in RapidMiner (k=3 only). Compare the two clustering results and answer the question: Are the two clustering results in R and RM the same or not? Why? You may follow the instruction in the next slide and take a screenshot of your PivotTable with date and to support your answer. Attention: you cannot just simply compare the cluster name because R and RM may label each cluster differently. For example, New Orleans, LA is labeled as cluster_0 in RM, but Cluster 3 in R, but cluster_0 in RM might be the same with Cluster 3 in R.

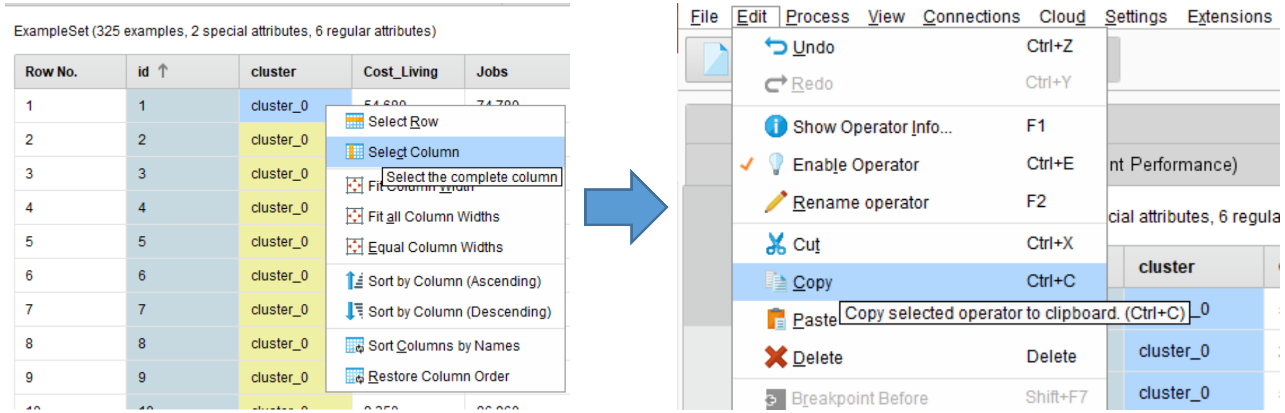
Deliverable R4 Instruction

1. Open CityRecords.csv in Excel (change the column names as I did as below)



	A	B	C	D	E	F	G	H
1	ID	R	Metropolitan_Area	Cost_Living	Jobs	Climate	Health_Care	Recreation
2	1	3	New Orleans, LA	54.68	74.78	75.92	91.5	100
3	2	3	Cleveland-Lorain-Elyria, OH	21.25	75.07	16.43	84.7	99.71

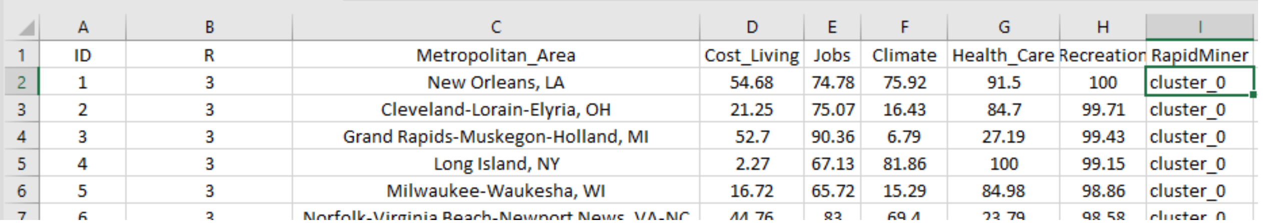
2. Select and copy the cluster column of ExampleSet in RapidMiner (Or you can use write CSV operator to export the clustering result). Attention: make sure the Row No. or ID column is at the ascending order; do not sort it by cluster)



ExampleSet (325 examples, 2 special attributes, 6 regular attributes)

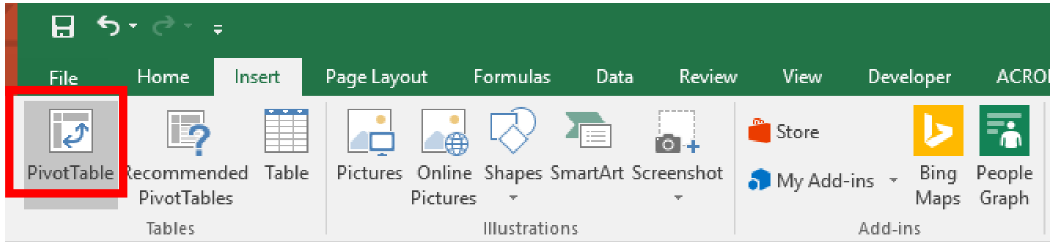
Row No.	id ↑	cluster	Cost_Living	Jobs
1	1	cluster_0		
2	2	cluster_0		
3	3	cluster_0		
4	4	cluster_0		
5	5	cluster_0		
6	6	cluster_0		
7	7	cluster_0		
8	8	cluster_0		
9	9	cluster_0		

3. Paste the cluster column to CityRecords.csv

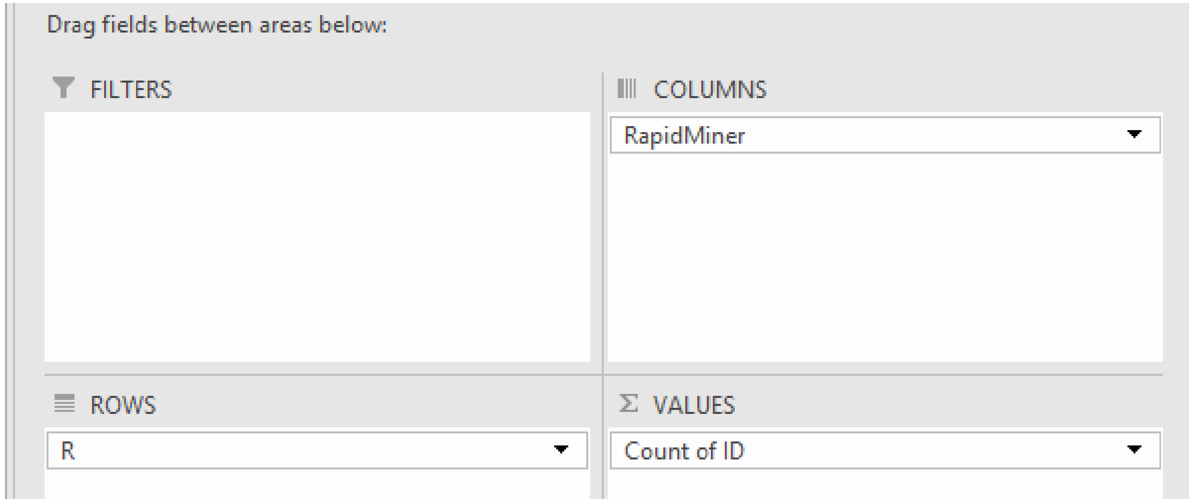


	A	B	C	D	E	F	G	H	I
1	ID	R	Metropolitan_Area	Cost_Living	Jobs	Climate	Health_Care	Recreation	RapidMiner
2	1	3	New Orleans, LA	54.68	74.78	75.92	91.5	100	cluster_0
3	2	3	Cleveland-Lorain-Elyria, OH	21.25	75.07	16.43	84.7	99.71	cluster_0
4	3	3	Grand Rapids-Muskegon-Holland, MI	52.7	90.36	6.79	27.19	99.43	cluster_0
5	4	3	Long Island, NY	2.27	67.13	81.86	100	99.15	cluster_0
6	5	3	Milwaukee-Waukesha, WI	16.72	65.72	15.29	84.98	98.86	cluster_0
7	6	3	Norfolk-Virginia Beach-Newport News, VA-NC	44.76	83	69.4	23.79	98.58	cluster_0

4. Insert a Pivottable



5. PivotTable: drag R to Rows and RapidMiner to Columns; count of ID ∑



Drag fields between areas below:

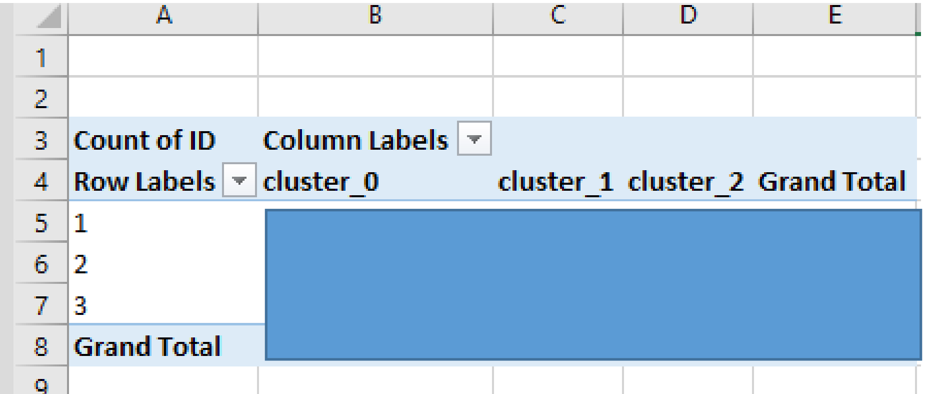
FILTERS

COLUMNS: RapidMiner

ROWS: R

Σ VALUES: Count of ID

6. Your PivotTable will be like the following. Make your conclusion based on this PivotTable



	A	B	C	D	E
1					
2					
3	Count of ID	Column Labels			
4	Row Labels	cluster_0	cluster_1	cluster_2	Grand Total
5	1				
6	2				
7	3				
8	Grand Total				
9					