CIDM 6355 Data Mining Methods HW3 Template

(100 points; Due 11:59 PM Central Time, October 20, 2024)

Requirements: This homework is open book, open slides, and open notes, but no collaboration or discussion is permitted before the due time. Any questions about the homework should be directed to the instructor. You must adhere to the instructions, completing all questions and deliverables. This is an individual assignment, so sharing your processes, scripts, screenshots, or answers with others constitutes cheating and will be reported. Additionally, ensure your answers meet the required format to avoid point deductions. Screenshots without date and time will receive a penalty of 50% of points. Identical screenshots will be considered academic dishonesty and will be reported to the university authority. Please acknowledge your understanding and agreement to these requirements by typing your name below.

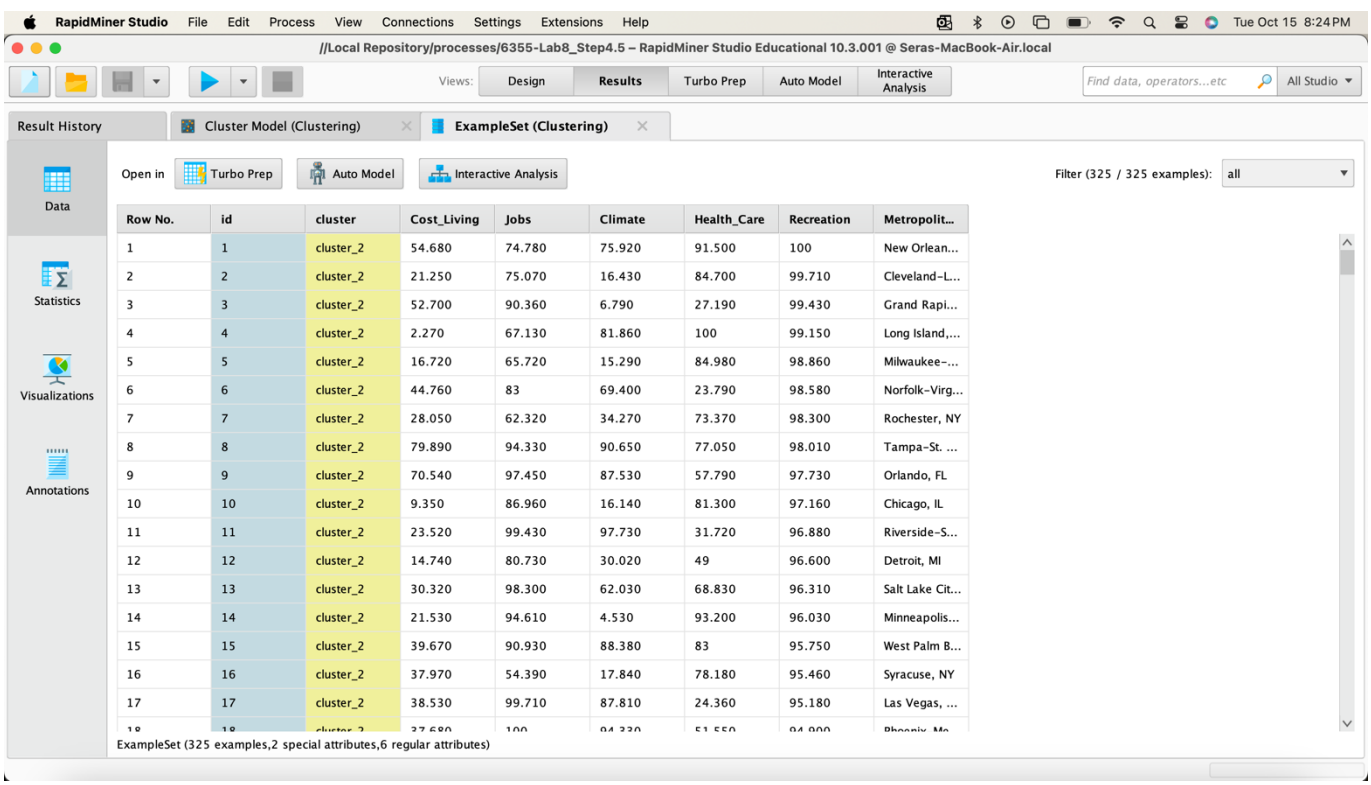
Type your name: Sera Hill

Instruction: Please compile all the deliverables with the required format as below.

1. Deliverable 1 (Step 1): Please write down the average for all the five attributes (round them the third decimal place). All these numbers below are the overall centroid for all 325 cities. [5 points]

Attributes	Cost_living	Jobs	Climate	Health_Care	Recreation
Average	51.910	51.023	52.035	47.865	50.227

2. Deliverable 2 (Step 4.5): Take a screenshot of your Exampleset (Screenshot 1) [5 points]

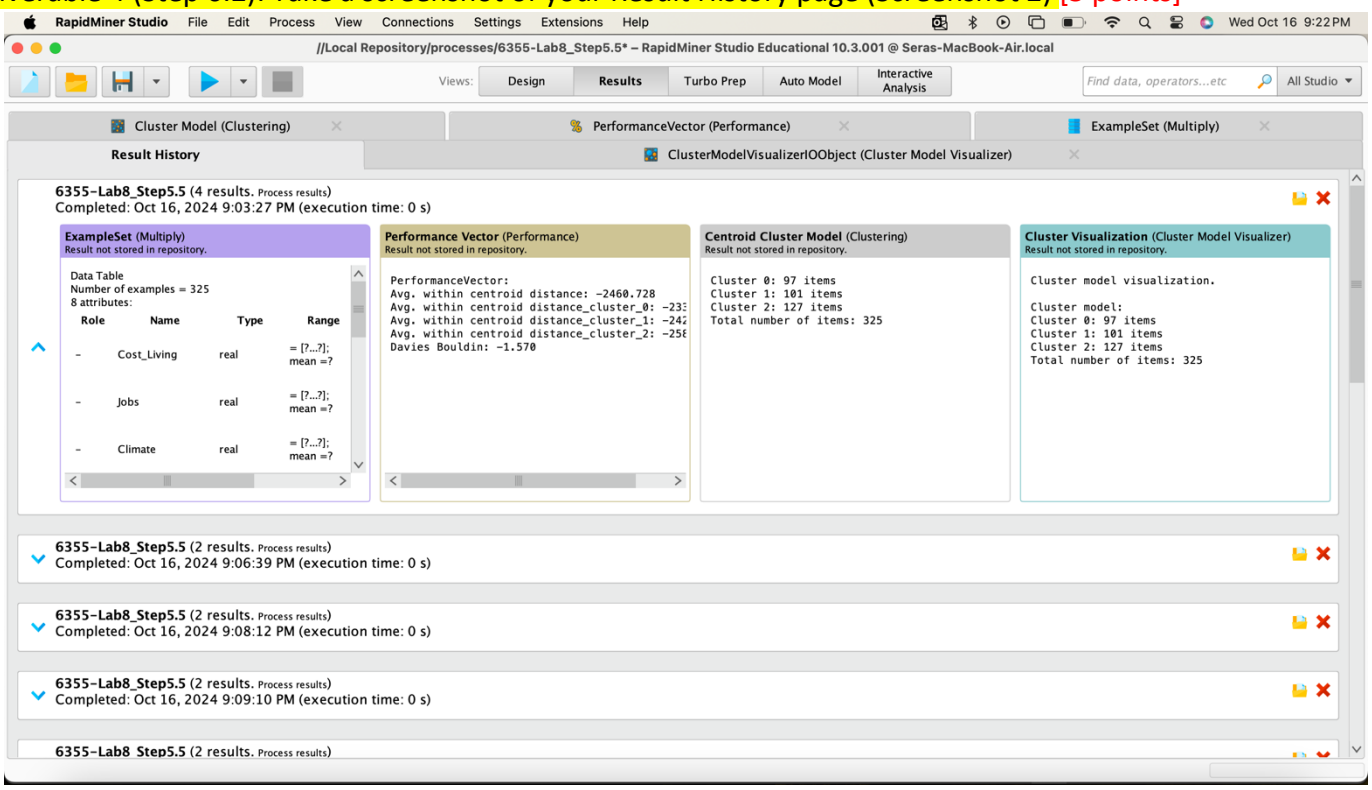


Row No.	id	cluster	Cost_Living	Jobs	Climate	Health_Care	Recreation	Metropolit...
1	1	cluster_2	54.680	74.780	75.920	91.500	100	New Orlean...
2	2	cluster_2	21.250	75.070	16.430	84.700	99.710	Cleveland-L...
3	3	cluster_2	52.700	90.360	6.790	27.190	99.430	Grand Rapi...
4	4	cluster_2	2.270	67.130	81.860	100	99.150	Long Island,...
5	5	cluster_2	16.720	65.720	15.290	84.980	98.860	Milwaukee-...
6	6	cluster_2	44.760	83	69.400	23.790	98.580	Norfolk-Virg...
7	7	cluster_2	28.050	62.320	34.270	73.370	98.300	Rochester, NY
8	8	cluster_2	79.890	94.330	90.650	77.050	98.010	Tampa-St. ...
9	9	cluster_2	70.540	97.450	87.530	57.790	97.730	Orlando, FL
10	10	cluster_2	9.350	86.960	16.140	81.300	97.160	Chicago, IL
11	11	cluster_2	23.520	99.430	97.730	31.720	96.880	Riverside-S...
12	12	cluster_2	14.740	80.730	30.020	49	96.600	Detroit, MI
13	13	cluster_2	30.320	98.300	62.030	68.830	96.310	Salt Lake Cit...
14	14	cluster_2	21.530	94.610	4.530	93.200	96.030	Minneapolis...
15	15	cluster_2	39.670	90.930	88.380	83	95.750	West Palm B...
16	16	cluster_2	37.970	54.390	17.840	78.180	95.460	Syracuse, NY
17	17	cluster_2	38.530	99.710	87.810	24.360	95.180	Las Vegas, ...

3. Deliverable 3 (Step 4.8): based on the results in 4.5-4.8, please discuss the characteristics in each cluster and find an appropriate name for each cluster. For example, Cluster 0 includes 128 cities such as New Orleans, LA and Long Island, NY have highest scores in job opportunities, climate, healthcare, and recreation. However, this group of cities have quite high living cost. We can name this group of cities Metropolitan Luxury…….. [21 points: 7 points for each cluster, including this cluster's sample size (1 pt.), sample cities (1 pt.), comparison on each dimension (4 pts), and name for this cluster (1 pt.)]

   Cluster 0 has 97 cities which include Houma, LA, Punta Gorda, FL, Panama City, FL. They have the highest score in climate, but it is the most expensive for cost of living. This group will be named Metropolitan Luxury.
   Cluster 1 has 101 cities which include Des Moines, IA, Fort Wayne, IN, Springfield, MA. They have the lowest score for jobs, and climate. This group will be called Metropolitan Austerity.
   Cluster 2 has 127 cities which include Orlando, FL, Chicago, IL, Detroit, MI. They have the highest score for jobs, health care, and recreation while having a low cost of living. This group will be called Metropolitan Exemplar.

4. Deliverable 4 (Step 6.2): Take a screenshot of your Result History page (Screenshot 2) [5 points]



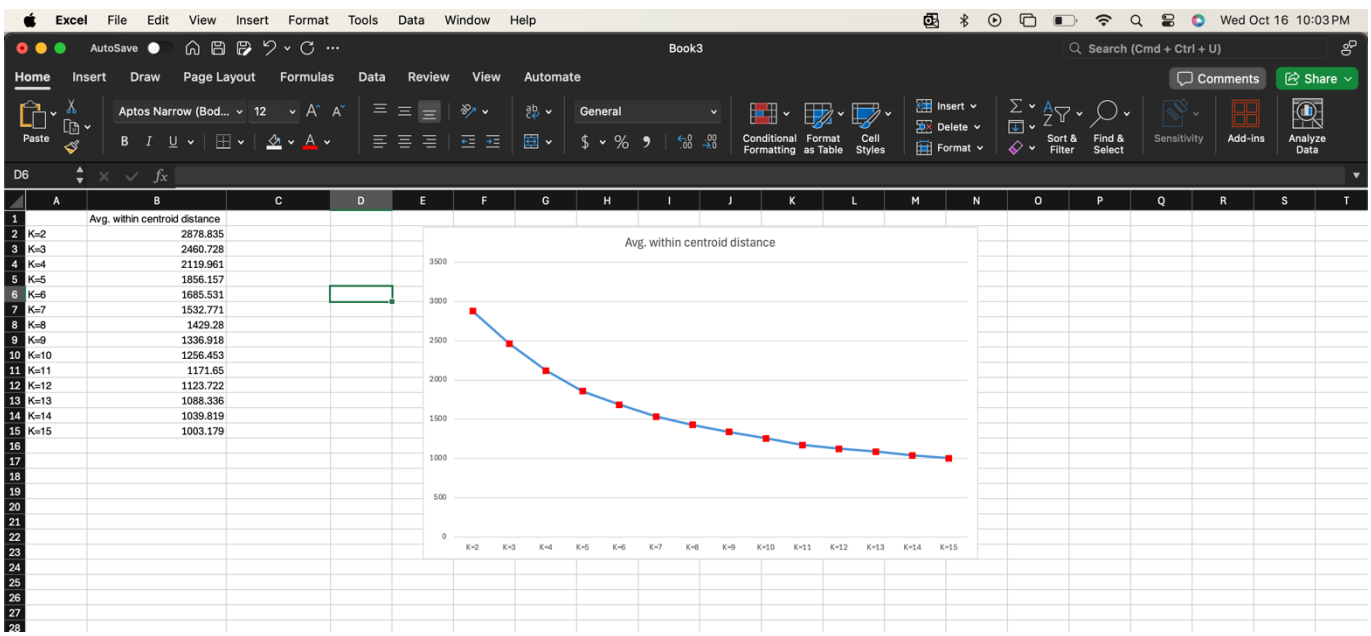
5. Deliverable 5 (Step 6.2): Please answer all the question in this deliverable [9 points]

   - Based on the table above, when k increases, what happen to Avg. within centroid distance

(increasing or decreasing)? [2 points] **It decreases.**

- What about Davies Bouldin Index when k increases? [2 points] **It alternates between decreasing and increasing.**
- Imagine an extreme case, when k=325, what would Avg. within centroid distance be? [2 points] **Zero.**
- What potential problem will we encounter if we only use Avg. within centroid distance as the main criterion for evaluating clustering models? [3 points] **We would be ignoring the size and shape of the cluster, and kind of ignores the possible presence of outliers.**

6. Deliverable 6 (Step 7.1): Draw an elbow chart using either average within centroid distance or DBI for k=2-15. Take a screenshot of your elbow chart with date and time (Screenshot 3). Observe your elbow chart and discuss which k is the best and why. [10 points: 5 points for screenshot and 5 points for your discussion]

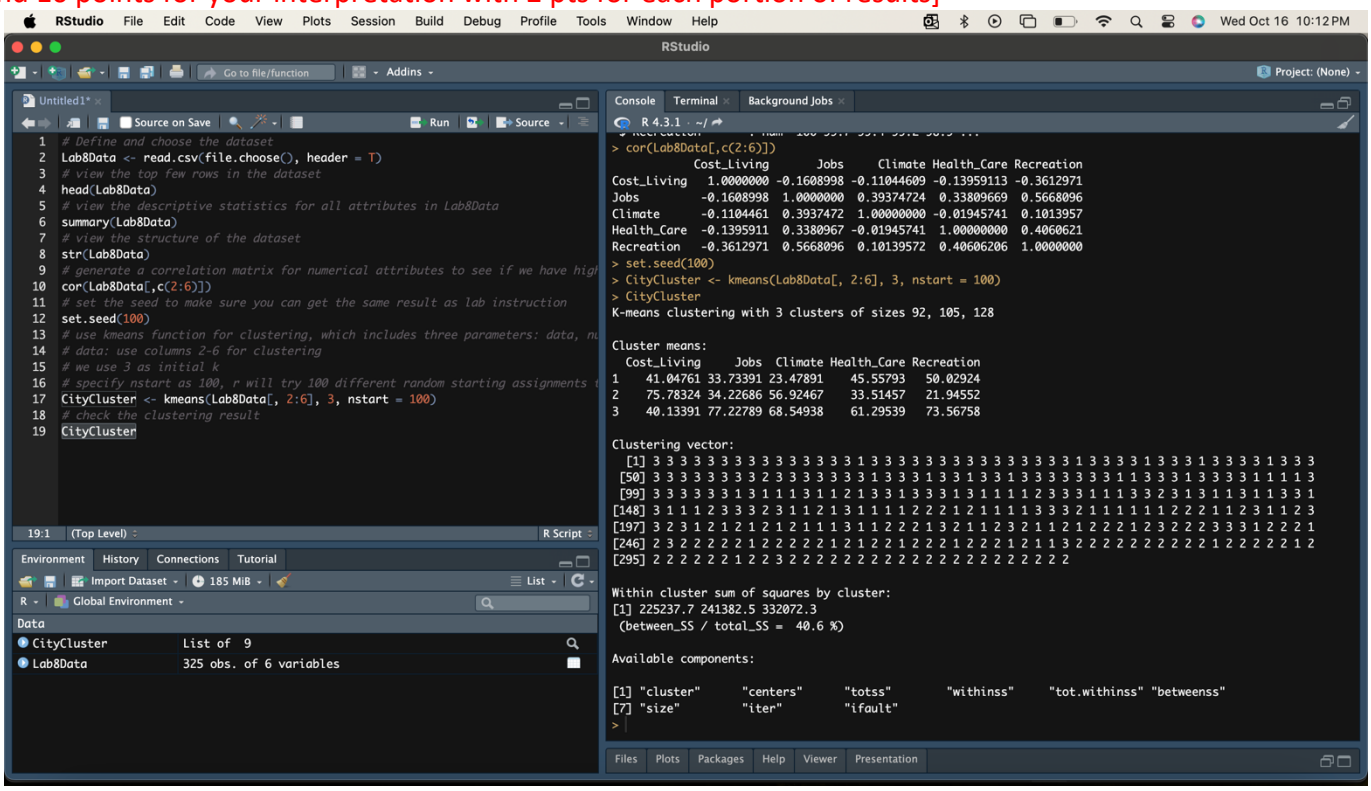


	Avg. within centroid distance
K=2	2878.835
K=3	2460.728
K=4	2119.961
K=5	1856.157
K=6	1685.531
K=7	1532.771
K=8	1429.28
K=9	1336.918
K=10	1256.453
K=11	1171.65
K=12	1123.722
K=13	1088.336
K=14	1039.819
K=15	1003.179

The best k would be one between 4 and 5 because this is when the elbow starts to flatten out.

7. Deliverable R1: take a screenshot of the result after running the script in Line 19 with date and time (Screenshot 4) and time and briefly interpret the result, explaining what each portion of results means. Your interpretation should cover the following five portions:
    - K-means clustering
    - Cluster means
    - Clustering vector
    - Within Cluster Sum of squares
    - Available components

Do some research if you do not know what each portion means. [15 points: 5 points for screenshot and 10 points for your interpretation with 2 pts for each portion of results]



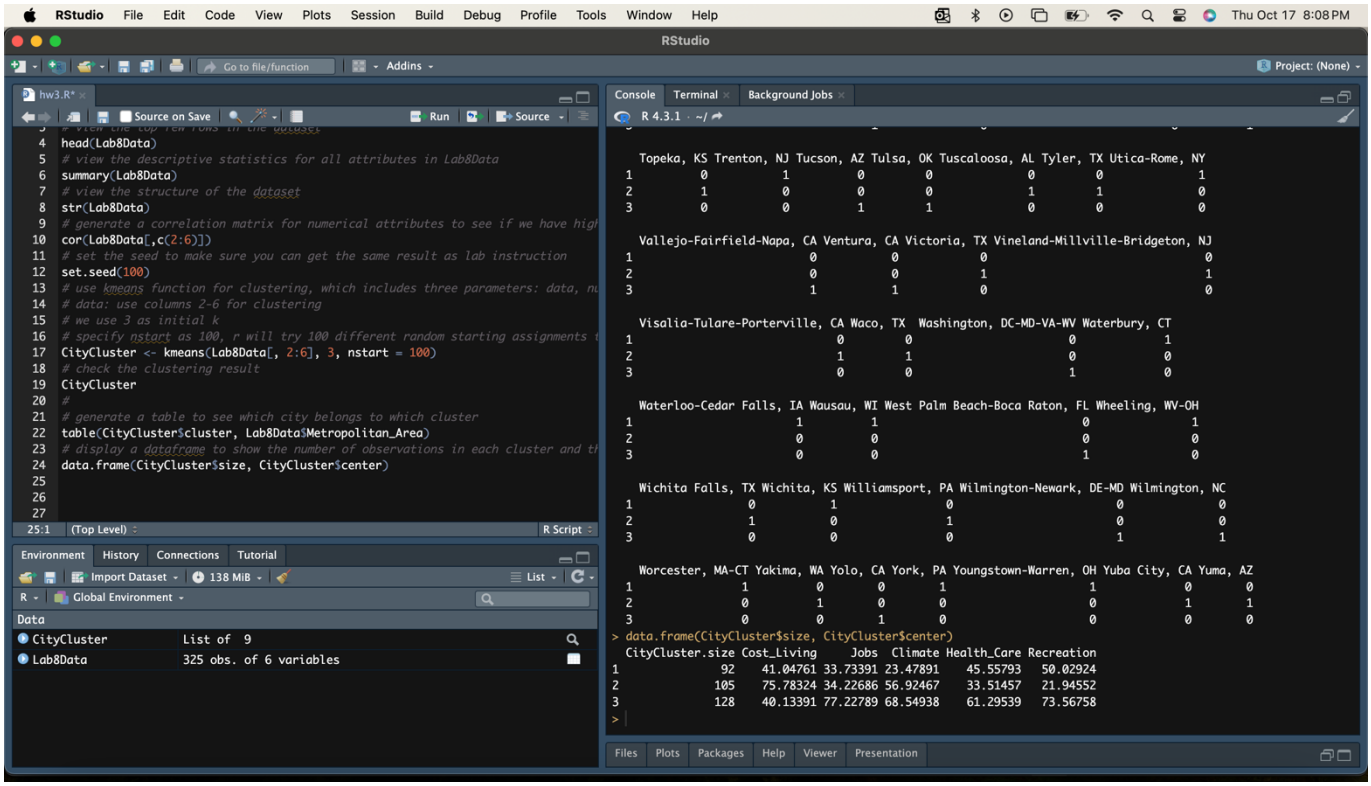
We defined k = 3 for our k-means clustering which created three groups; group 1 has 92 items, group 2 has 105 items, and group 3 has 128 items. The cluster means values shows us the center of each cluster for each group for each category. For group 1, cost of living is moderate compared to other groups, jobs are low, climate isn't as good as other groups, health care is better than others, and there are decent recreational opportunities. Group 2's cost of living is much higher, while the jobs are lower but still higher than group 1. Climate is favorable, but health care and recreation have lower rates than the other groups. Group 3 has moderate cost of living, with better job availability. Climate, health care and recreation are also very good compared to other groups.

The clustering vector shows us which group each datapoint belongs to. With a value of 1 belonging to group 1, value of 2 belonging to group 2 and value of 3 belonging to group 3.

The within cluster sum of squares by cluster, we can see that group 1's datapoints are closer to the centroid than the other groups, with group 3 having the furthest distance between from the centroid.

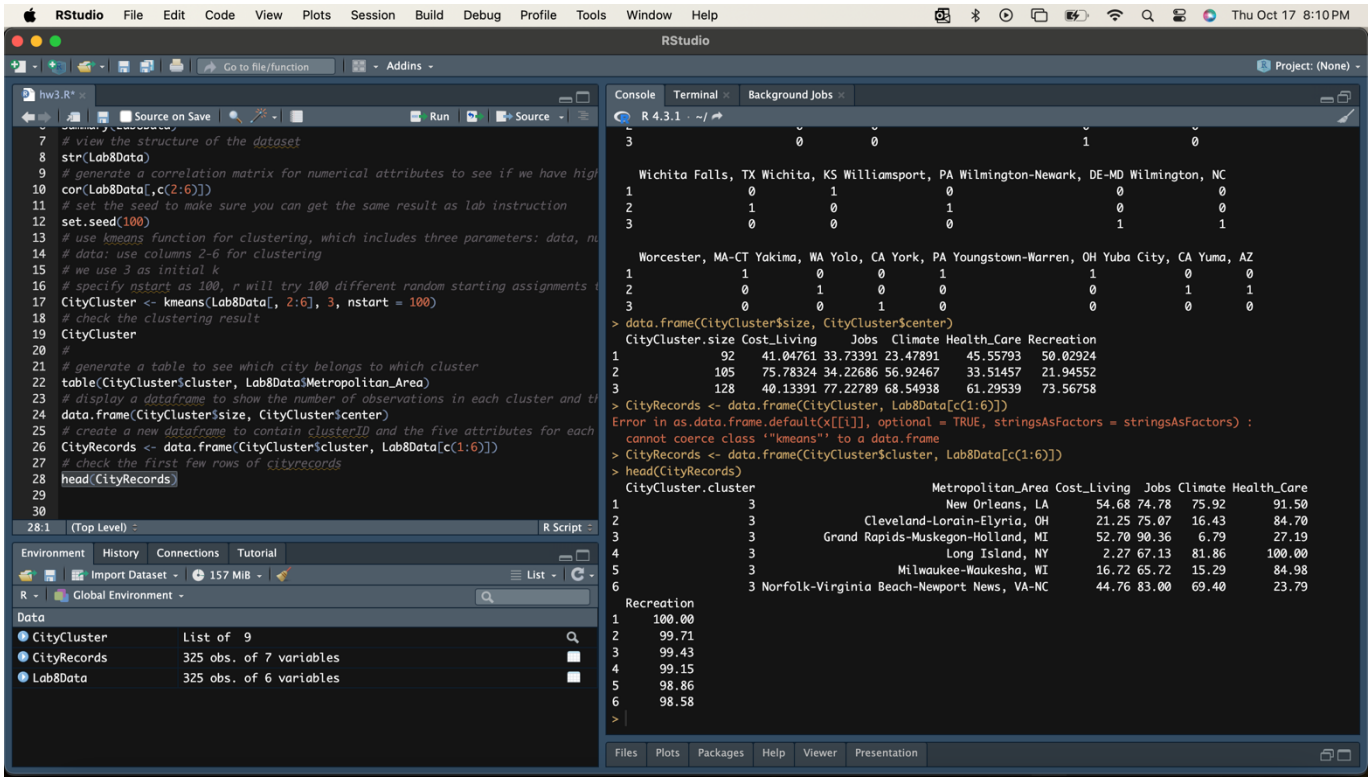
The available components show us which values that were used during our k-means analysis. Cluster tells us where each point belongs, centers show us the average position of each point in each cluster, totss shows us total variance, and withinss, and tot.withinss shows how well the points fit in each cluster. Betweenss shows how much separation each cluster has, and size tells us how many points are in each cluster. Iter is how many iterations the algorithm went through and ifault will let us know of any problems in the algorithm's execution.

8. Deliverable R2: take a screenshot of the result after running the script in Line 24 with date and time (Screenshot 5) and time and briefly interpret the result, explaining what the result is about and what each column means. [10 points: 5 points for screenshot and 5 points for your interpretation]



The result shows us how many observations are in each cluster as well as the mean of each attribute in the cluster. First column shows us cluster size, the following columns are the mean of the datapoints for each category: cost of living, jobs, climate, health care, and recreation.

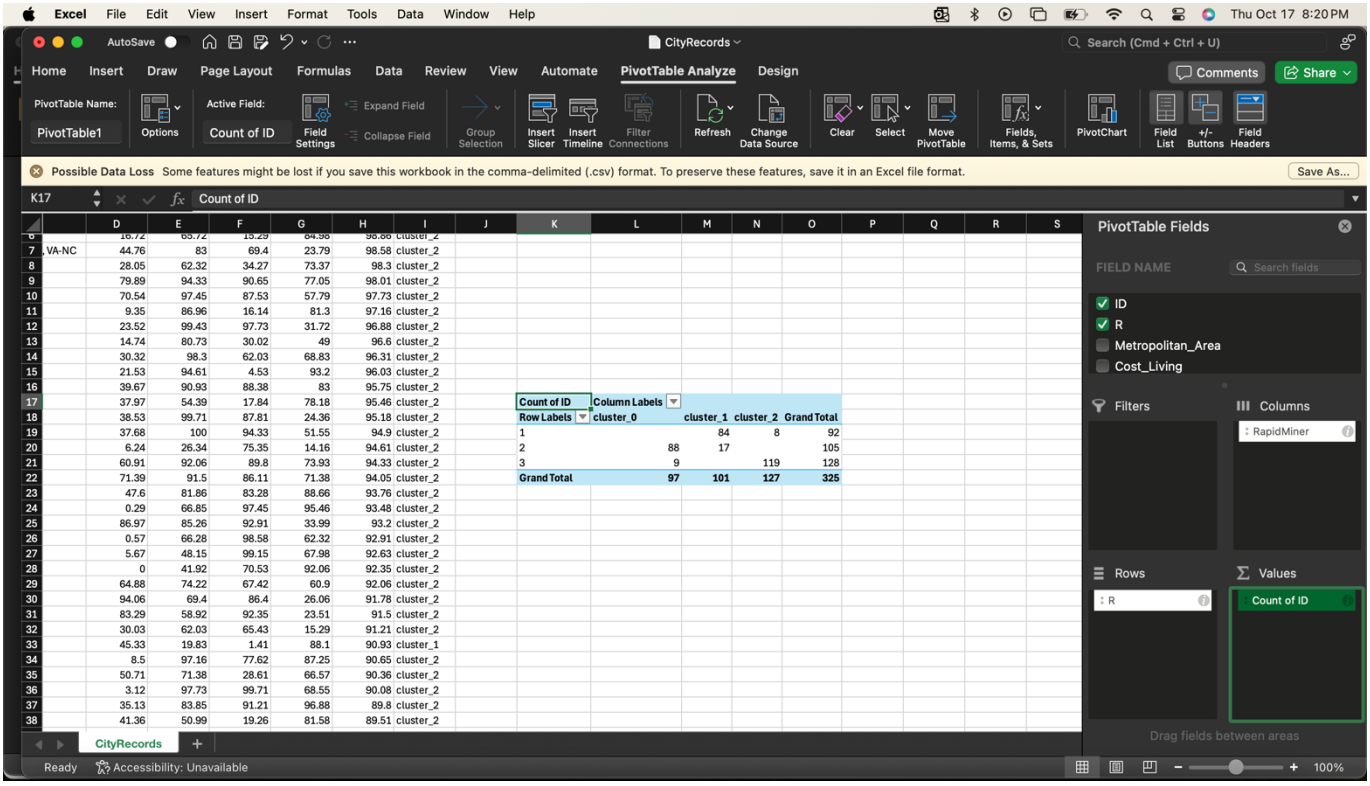
9. Deliverable R3: take a screenshot of the result after running the script in Line 28 with date (Screenshot 6) and time, and briefly interpret the result, explaining what the result is about and what each column means. [10 points: 5 points for screenshot and 5 points for your interpretation]



This result shows the first few rows of the CityCluster dataframe. This dataframe contains information about the different cities and what cluster they belong to (CityCluster.cluster and Metropolitan_Area columns), as well as the mean of each attribute in each cluster (Cost_Living, Jobs, Climate, Health_Care, Recreation columns).

10. 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 time to support your answer (Screenshot 7). 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. [10 points: 5 points

for screenshot and 5 points for your answer]



Count of ID	Column Labels			
Row Labels	cluster_0	cluster_1	cluster_2	Grand Total
1		84	8	92
2	88	17		105
3	9		119	128
Grand Total	97	101	127	325

The two clustering results are not the same because R and RapidMiner may be using different algorithms for the k-means analysis. They also may have different ways of handling preprocessing, how they handle equal distances, or how they scale the data.