**CIDM 6355 Data Mining Methods HW3**

(60 points; Due 11:59 PM Central Time, October 15, 2023)

Requirements: This homework is open book, open slides, and open notes, but you are not allowed to collaborate nor discuss with anyone else before the due time. Any question about the homework should be addressed to the instructor. You are required to follow the instruction to complete all the questions and deliverables. This is an individual homework assignment, so sharing your RM processes, R scripts, screenshots, or answers with other students or parties is considered as cheating, which will be reported to the university authority. In addition, it is your responsibility to make your answers meet the required format; otherwise, you might lose points because of wrong format. Screenshots without date and time can only receive up to 50% of points. Please read, understand, and comply with these requirements in this homework assignment by typing your name as below.

Your name: Jordan Unfred

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 |

1. Deliverable 2 (Step 4.5): Take a screenshot of your Exampleset (Screenshot 1) [3 points]

A screenshot of a computer

Description automatically generated

1. 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 …….. [9 points: 3 points for each cluster]

Cluster\_0: has the highest scores in cost of living and climate, however it is the lowest in recreation and healthcare. These are cities such as El Paso, TX, Jacksonville, NC, and Amarillo, TX. We can name this group of cities Small Cities.

Cluster\_1: doesn’t have any high scores but rather median in Cost of living, healthcare, and recreation, but has the lowest scores in jobs and climate. Iowa City, IA, Lincoln, NE, and Lafayette, IN. We can name this group of cities Large Cities

Cluster\_2: has the highest scores in Jobs, Healthcare, and Recreation, but the lowest overall score on the cost of living. These are cities such as New Orleans, LA, Orlando, FL, and Dallas, TX. We can call these the Metropolitans.

1. Deliverable 4 (Step 6.2): Take a screenshot of your Result History page (Screenshot 2) [3 points]

A screenshot of a computer

Description automatically generated

|  |  |  |
| --- | --- | --- |
| (- RM) | Avg. within centroid distance | Davies Bouldin Index |
| **K=2** | 2878.835 | 1.467 |
| **K=3** | 2460.728 | 1.570 |
| **K=4** | 2119.961 | 1.429 |
| **K=5** | 1856.157 | 1.371 |
| **K=6** | 1685.531 | 1.320 |
| **K=7** | 1532.771 | 1.266 |
| **K=8** | 1429.280 | 1.303 |
| **K=9** | 1336.918 | 1.259 |
| **K=10** | 1256.453 | 1.292 |
| **K=11** | 1171.650 | 1.259 |
| **K=12** | 1123.722 | 1.242 |
| **K=13** | 1088.336 | 1.247 |
| **K=14** | 1039.819 | 1.232 |
| **K=15** | 1003.179 | 1.290 |

1. Deliverable 5 (Step 6.2): Please answer all the question in this deliverable [8 points]

* Based on the table above, when k increases, what happen to Avg. within centroid distance (increasing or decreasing)? [2 points]

When the k increases, the avg centroid decreases.

* What about Davies Bouldin Index when k increases? [2 points]

When the k increases, the Davies Bouldin Index also decreases.

* Imagine an extreme case, when k=325, what would Avg. within centroid distance be? [2 points]

The avg. within centroid distance would be zero

A screenshot of a computer

Description automatically generated

* What potential problem will we encounter if we only use Avg. within centroid distance as the main criterion for evaluating clustering models? [2 points]

When we evaluate clustering models, we need to use multiple performance measures instead of just one. Using only average within centroid distance as the main criterion can lead to a few potential problems. One of the main issues is that it does not take into account the number of clusters in the model. As a result, it may not be able to distinguish between models with different numbers of clusters. Additionally, it may not be able to identify clusters that are too small or too large, which can lead to suboptimal results. Other performance measures such as Davies-Bouldin index can be used in conjunction with average within centroid distance to provide a more comprehensive evaluation of clustering models

1. 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. [5 points: 3 points for screenshot and 2 points for your discussion]

A screen shot of a graph

Description automatically generated

I think K=6 is the best to choose from as it is the beginning of the straightest data, also known as the elbow on the graph.

1. Deliverable 7: Use k=3 to include all the four performance operators in your process. Take a screenshot of the description view of Cluster Density Performance and Item Distribution Performance (two screenshots in total: Screenshot 4 and Screenshot 5) and then briefly discuss each result. [7 points: 4 points for two screenshots and 3 points for your discussion]

A screenshot of a computer

Description automatically generated

My Cluster density avg within cluster distance was -7315.561

A screenshot of a computer

Description automatically generated

My Item distribution example distribution, or sum of squares, was equal to 0.338

1. Deliverable R1: take a screenshot of the result after running the script in Line 19 with date and time (Screenshot 6) and time briefly interpret the result. [5 points: 3 points for screenshot and 2 points for your interpretation]

A screenshot of a computer

Description automatically generated

The cluster shoes cluster 1, 2, and 3 and the different scores between cost of living, jobs, climate, health care, and recreation. The cluster also provides the sum of squares by cluster to 40.6%

1. Deliverable R2: take a screenshot of the result after running the script in Line 24 with date and time (Screenshot 7) and time briefly interpret the result. [5 points: 3 points for screenshot and 2 points for your interpretation]



The dataframe shows the size of cluster 1 through 3 and their respective average scores

1. Deliverable R3: take a screenshot of the result after running the script in Line 28 with date (Screenshot 8) and time briefly interpret the result. [5 points: 3 points for screenshot and 2 points for your interpretation]

A screenshot of a computer

Description automatically generated

CityRecords shows the top of the list organized descending from top recreation score of all of the clusters

1. 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 9). 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. [5 points: 3 points for screenshot and 2 points for your answer]

A screenshot of a computer

Description automatically generated

Comparing the clustering results between R and RM outputted the same between New Orleans and Norfolk Virginia Beach Newport News and so on. They are the same as they are in descending order of recreation score. The top results in both R and RM are found in metropolitan areas, and descend to large then smaller cities.