**CIDM 6355 Data Mining Methods HW3**

(60 points; Due 11:59 PM Central Time, March 18, 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 maximum of 50% of points. Please acknowledge your understanding and agreement to these requirements by typing your name below.

Type Your name: Mariam Adegbindin

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** includes 64 cities, such as Cleveland-Lorain-Elyria, OH and Grand Rapids-Muskegon-Holland, MI. These cities offer balanced scores in job opportunities, recreation, and health care facilities, with a moderate climate. However, this group of cities has a slightly higher cost of living. We can name this group of cities **"**Balanced Opportunity Cities**."**

**Cluster 1** includes 56 cities, such as Barnstable-Yarmouth, MA and Burlington, VT. These cities are characterized by their very low cost of living and moderate climate but with lower job opportunities and health care quality. Recreation scores are average. This cluster is ideal for those prioritizing affordability and decent climate over job markets or health care excellence. We can name this group of cities "Affordable Lifestyle Cities."

**Cluster 2** includes 58 cities, such as Daytona Beach, FL and Melbourne-Titusville-Palm Bay, FL. These cities have the highest climate scores, indicating possibly warmer or more desirable weather conditions, combined with moderate to high cost of living and job opportunity scores. However, they score lower in recreation and health care. This cluster might attract those who value climate and job opportunities over access to recreation or top-tier health care. We can name this group of cities "Warm Climate Work Cities."

**Cluster 3** includes 75 cities, such as Beaumont-Port Arthur, TX and Huntington-Ashland, WV-KY-OH. These cities are characterized by the highest cost of living and the lowest scores in jobs, recreation, and relatively low health care quality, with a moderate climate. Cities in this cluster are likely expensive to live in with fewer job opportunities and lower recreation and health care amenities. We can name this group of cities "High-Cost Living Cities."

**Cluster 4** includes 72 cities, such as New Orleans, LA and Long Island, NY. These cities stand out for having high job opportunities, excellent recreation scores, and a very favorable climate, combined with a lower cost of living and good health care scores. They are attractive for those seeking a blend of career opportunities, lifestyle quality, and affordability. We can name this group of cities "Ideal Work-Life Balance Cities."

**Top of Form**

**Bottom of Form**

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

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]

The Avg. within centroid distance decreases because data points are likely to be closer to the centroids of their clusters. With more clusters, each cluster tends to be smaller, and the centroids can more accurately represent the points within each cluster.

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

Increasing k will lower the Davies-Bouldin Index at first if it helps to split the groups better. But after the best number of clusters, the index may start to rise because the clusters start to break up natural groups of data or add noise and outliers that aren't well separated.

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

As an extreme case, if k was the same as the number of data points (325 in this case), each data point would be its own cluster, with the center of each cluster being the data point itself. It follows that the average distance between places inside each centroid would be zero, since each point is its own centroid. The distance between a point and itself is zero.

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

One problem could be Overfitting, if you only minimize the average within-cluster distance, the optimal solution is to increase the number of clusters to the point where each data point becomes its own cluster (as with k=325 in the previous example). This obviously defeats the purpose of clustering, which is to find patterns and groupings within the data.

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

k=4 is the best choice for the number of clusters because it balances within-cluster variance reduction with the complexity of the model

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

A screenshot of a computer

Description automatically generated

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 three clusters contain 92, 105, and 128 items (cities) respectively. These are the numbers of cities that belong to each cluster. The within-cluster sum of squares is a measure of the variance within each cluster. Lower values generally indicate that the points within each cluster are closer to their respective centroids. The sum of squares within each cluster is provided, for example, Cluster 1 has a sum of squares of 225237.7. The ratio of the between-cluster sum of squares to the total sum of squares is 40.6%. This value indicates the proportion of the total variance in the dataset that is explained by the clustering.

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]

A screenshot of a computer

Description automatically generated

A data.frame is printed out showing the size (number of observations) and mean of each attribute (Cost\_Living, Jobs, Climate, Health\_Care, Recreation) within each cluster.

Cluster 1 has the lowest average Cost of Living and Health Care scores, but has relatively higher Jobs and Recreation scores compared to Cluster 2. The Climate score is moderate.

Cluster 2 has moderate means across all features, indicating a balanced profile, but with the lowest Jobs and Recreation scores among the three clusters.

Cluster 3 has the highest means across all features indicating that this cluster may consist of cities with higher Cost of Living, better Job opportunities, a warmer Climate, better Health Care, and Recreation facilities.

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

Cluster 3 consists of metropolitan areas that are diverse in terms of Cost of Living and Climate but generally offer good to excellent Health Care and Recreation opportunities, with robust job markets. There is some variability within the cluster, which could be due to regional differences or specific local factors that influence these scores. For example, Grand Rapids-Muskegon-Holland has moderate Cost of Living and the highest Jobs score among the six cities. The Climate score is low, suggesting possibly harsher weather conditions, but Recreation opportunities are among the highest, despite the Health Care score being relatively low.

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

We see that not all observations from R in cluster 3 (marked as "3") are assigned to RM's cluster\_2 (which could be considered the equivalent of R's cluster 3). This mismatch suggests that the clustering results from R and RM are not the same

RM Cluster Result

|  |
| --- |
| Row No ID. Cluster. Cost\_Living |
| 1 1 cluster\_2 54.68 74.78 75.92 91.5 100 New Orleans, LA |
| 2 2 cluster\_2 21.25 75.07 16.43 84.7 99.71 Cleveland-Lorain-Elyria, OH |
| 3 3 cluster\_2 52.7 90.36 6.79 27.19 99.43 Grand Rapids-Muskegon-Holland, MI |
| 4 4 cluster\_2 2.27 67.13 81.86 100 99.15 Long Island, NY |
| 5 5 cluster\_2 16.72 65.72 15.29 84.98 98.86 Milwaukee-Waukesha, WI |
| 6 6 cluster\_2 44.76 83 69.4 23.79 98.58 Norfolk-Virginia Beach-Newport News, VA-NC |
| 7 7 cluster\_2 28.05 62.32 34.27 73.37 98.3 Rochester, NY |
| 8 8 cluster\_2 79.89 94.33 90.65 77.05 98.01 Tampa-St. Petersburg-Clearwater, FL |
| 9 9 cluster\_2 70.54 97.45 87.53 57.79 97.73 Orlando, FL |
| 10 10 cluster\_2 9.35 86.96 16.14 81.3 97.16 Chicago, IL |

R Cluster Result

A screenshot of a graph

Description automatically generated