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

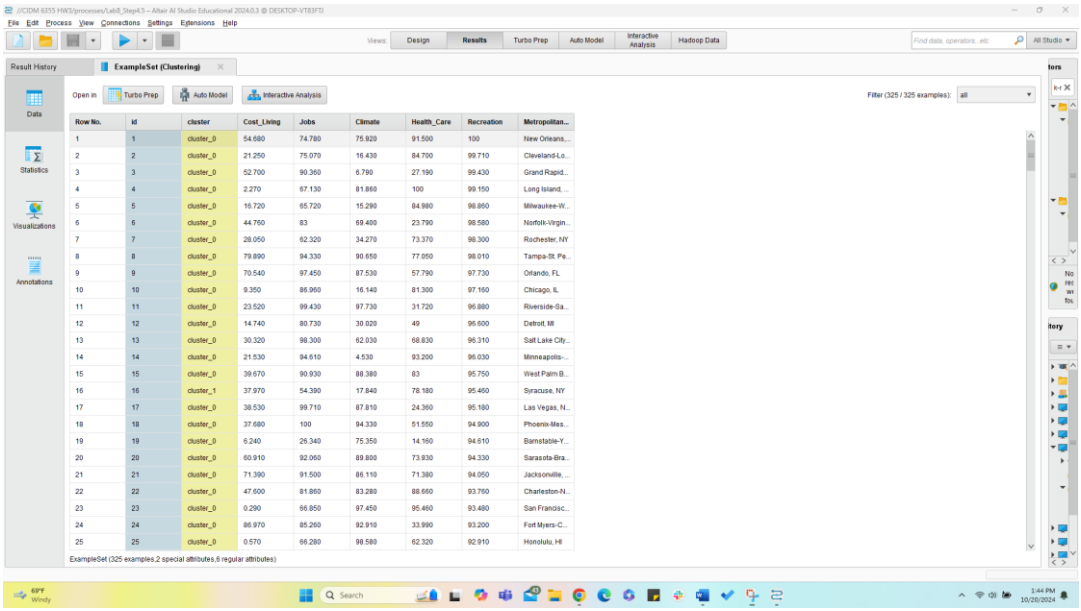
1. Deliverable 1 (Step 1): Please write down the average for all the five attributes (round them the third decimal place). All these numbers below are the overall centroid for all 325 cities. [5 points]

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]

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 includes 128 cities. Some of these cities are West Palm Beach-Boca Raton, FL, Las Vegas NV-AZ, and Sarasota-Bradenton, FL. This cluster of cities has the highest scores on Recreation, Health Care, Climate, and Jobs out of the three clusters. This cluster also has the lowest score when it comes to cost of living. We will name this group of cities 'Highest Score: Jobs'.
   - Cluster 1 includes 92 cities. Some of these cities are Syracuse, NY, Duluth-Superior, MN-WI,



     and Buffalo-Niagra Falls, NY. This cluster of cities has the lowest scores in Jobs and Climate. But in the middle of the three clusters when it comes to Cost of Living, Health Care, and Recreation. We will name this group of cities 'Highest Score: Recreation'.
   - Cluster 2 includes 105 cities. Some of these cities are Houma, LA, Panama City, FL, and Brownsville-Harlingen-San Benito, TX. This cluster of cities has the lowest scores in Health Care

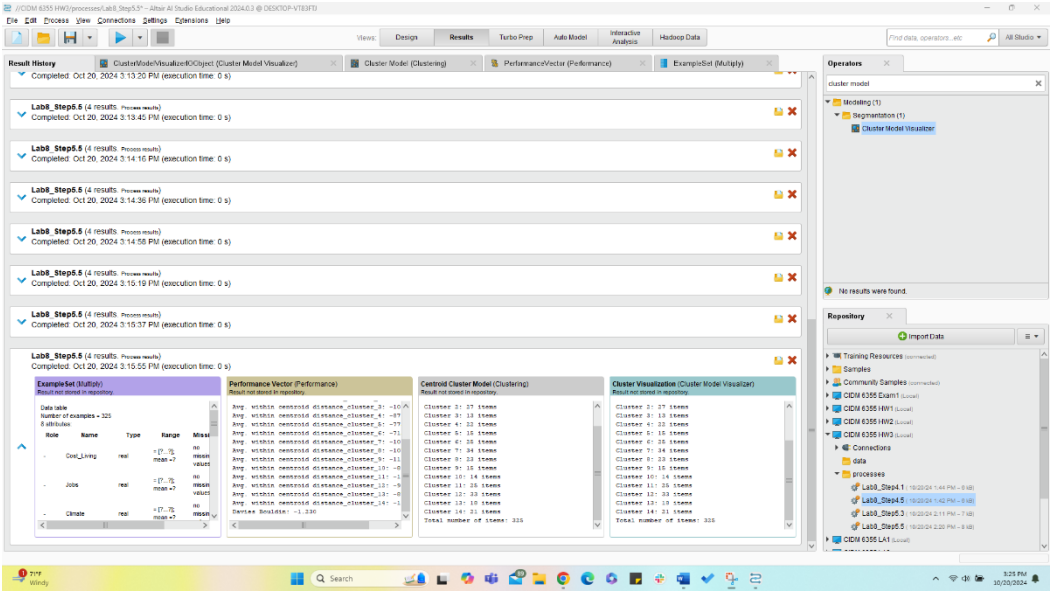
and Recreation, but the highest score in Cost of Living. We will name this group of cities 'Highest Score: Cost of Living'.

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]
- What about Davies Bouldin Index when k increases? [2 points]
- Imagine an extreme case, when k=325, what would Avg. within centroid distance be? [2 points]
- What potential problem will we encounter if we only use Avg. within centroid distance as the main criterion for evaluating clustering models? [3 points]

When k increases, the average within centroid distance decreases. Most of the time, the Davies Bouldin Index also decreased, but there were some instances where it increased instead of decreased when k was increased. From k=2(1.467) to k=3(1.567), there was an increase. From k=7(1.254) to k=8(1.271) there was a slight increase. From k=8(1.271) to k=9(1.298) there was also



a slight increase. From k=13(1.203) to k=14(1.253) was the last increase that I observed.

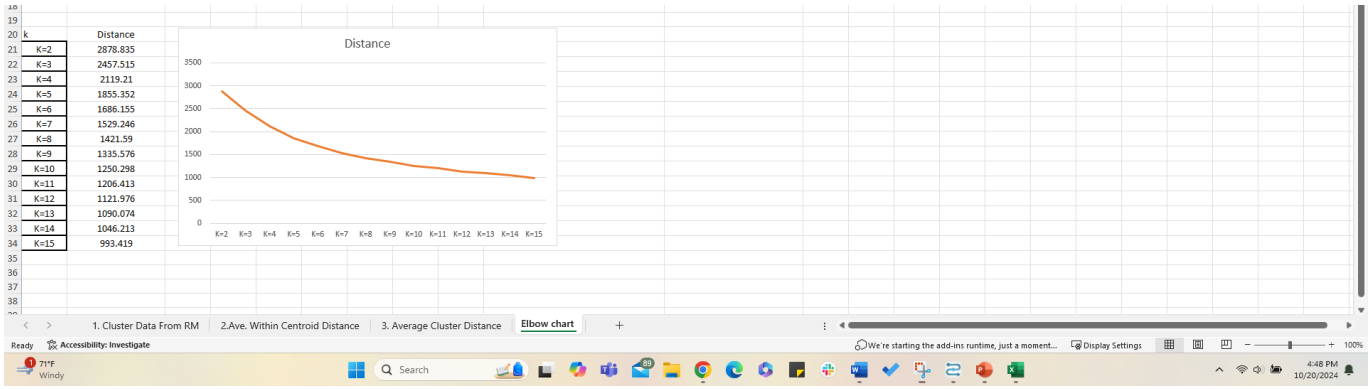
In the extreme case that k=325, the avg within centroid distance is showing to be 0.000.

A potential problem that we will encounter if only using the average within centroid distance is only looking at intra-cluster distance and not evaluating the distance within the clusters that gives us important information about the inter-cluster distance.

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]

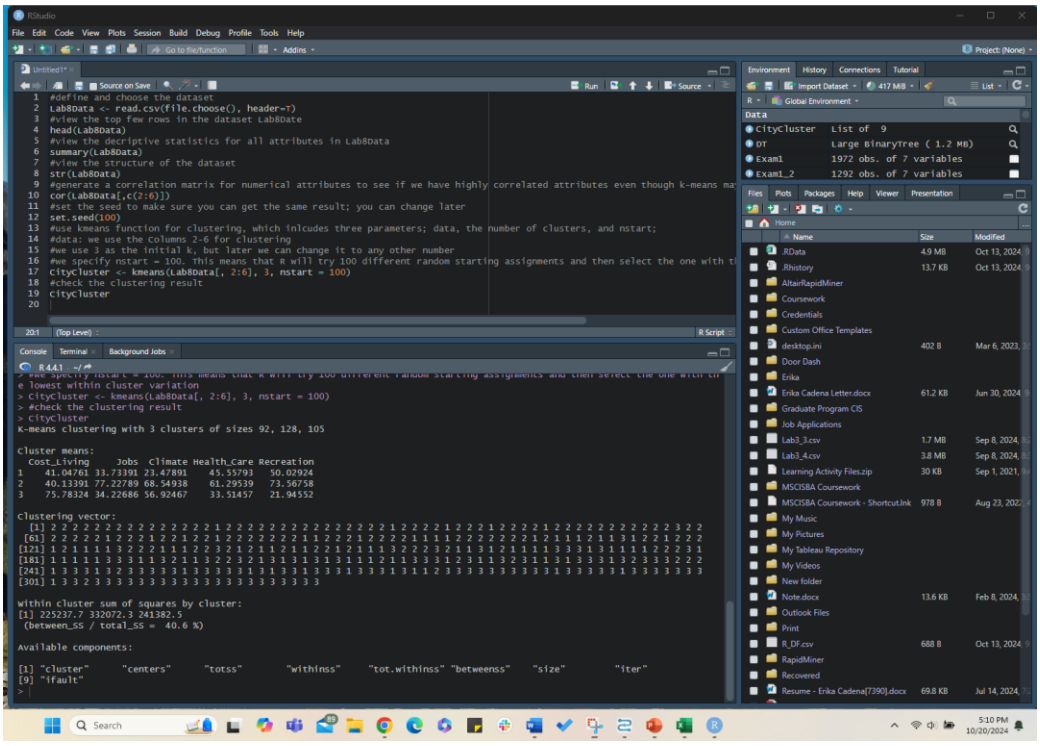
I used the average within centroid distance for this elbow chart. From this chart, you can see that k=5 is the best k to use because of the changes in the line afterwords. There is a slower decrease from 6 to 15 compared to 2 to 5.

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



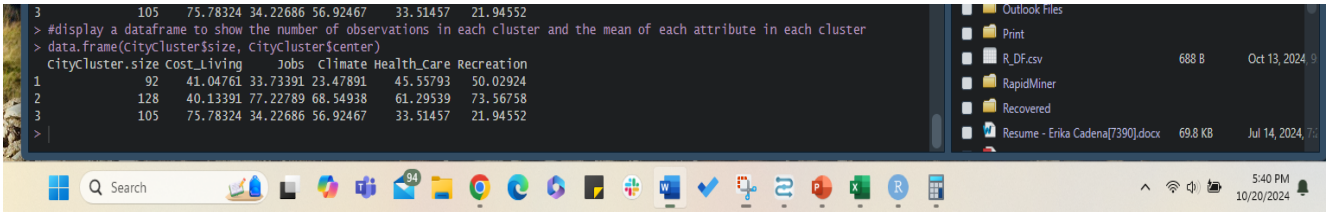
    - Within Cluster Sum of squares
    - Available components

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]



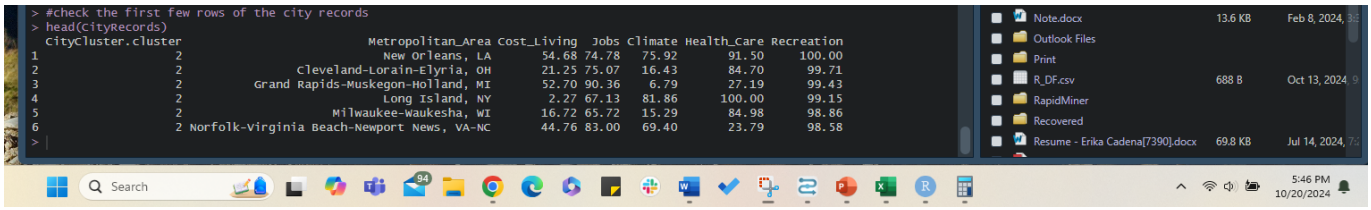
K-means clustering: States that there are three clusters with sizes 92, 128, and 105.
Cluster means: this portion displays the mean/average of each column. For example, the average in cluster one for Cost_Living is 41.04761.
Clustering vector: displays which cluster is assigned to each city. The first row starts at ID=1 and the second row starts at ID=61, and so on.
Within Cluster Sum of squares: this shows how compact each cluster is by taking the sum of squared distances of each point and centroid. Cluster 1 being 225237.7, Cluster 2 being 332072.3, and Cluster 3 being 241382.5 to show similarity between points. Per assignment notes: the percentage of variance explained.
Available components: these are available arguments and values that can be used in the kmeans function or objects that are returned

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 results display each cluster size and the average score across each column. For example, Cluster 3 with the size of 105 has the highest score when it comes to cost of living, but the lowest score when it comes to health care.

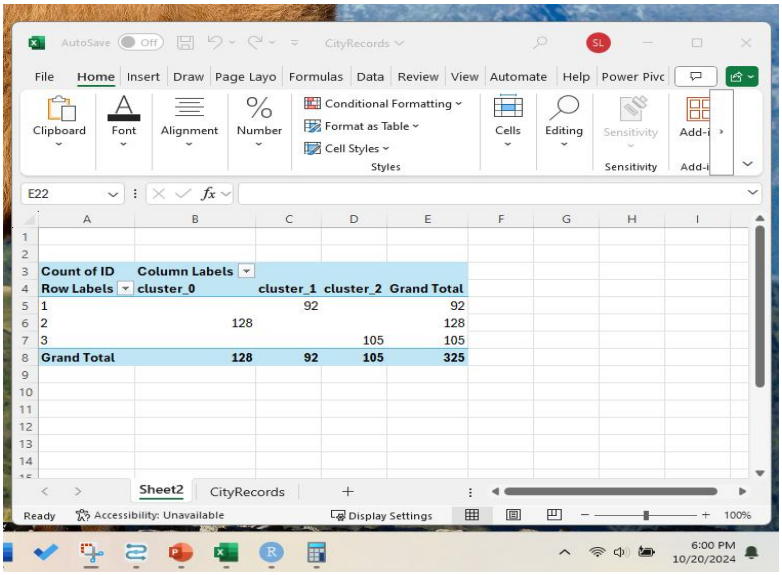
9. Deliverable R3: take a screenshot of the result after running the script in Line 28 with date (Screenshot 6) and time, and briefly interpret the result, explaining what the result is about and what each column means. [10 points: 5 points for screenshot and 5 points for your interpretation]



```
> #check the first few rows of the city records
> head(CityRecords)
  CityCluster.cluster                      Metropolitan_Area Cost_Living  Jobs Climate Health_Care Recreation
1                   2                          New Orleans, LA       54.68 74.78   75.92      91.50     100.00
2                   2                 Cleveland-Lorain-Elyria, OH       21.25 75.07   16.43      84.70      99.71
3                   2          Grand Rapids-Muskegon-Holland, MI       52.70 90.36    6.79      27.19      99.43
4                   2                            Long Island, NY        2.27 67.13   81.86     100.00      99.15
5                   2                      Milwaukee-Waukesha, WI       16.72 65.72   15.29      84.98      98.86
6                   2 Norfolk-Virginia Beach-Newport News, VA-NC       44.76 83.00   69.40      23.79      98.58
>
```

The results are showing the avg of scores in each city in the dataset along with the city name and which cluster it is assigned to. Citycluster.cluster is the cluster the city is assigned to, Metropolitan_Area is the city name, Cost_Living, Jobs, Climate, Health_Care, Recreation are attributes of the cities. For New Orleans, LA, this city is assigned to cluster 2, cost of living average score is 54.68, climate is 75.92, health care is 91.50, and recreation is 100.00.

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		92		92
2	128			128
3			105	105
Grand Total	128	92	105	325

Although the clusters are not named the same, as stated in the homework instructions, they are creating the same cluster sizes. The clustering results in R and RM are the same. As shown above in the pivot table, there were three clusters created with sizes 92, 128, and 105. Each city labeled as 1 in R, was labeled as cluster_1, labeled as 2 in R was labeled as cluster_0 in RM, and labeled as 3 in R was labeled as cluster_2 in RM. Each city was in the same group (cluster) in R as they were in RM.

References for research:

https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/kmeans

https://uc-r.github.io/kmeans_clustering