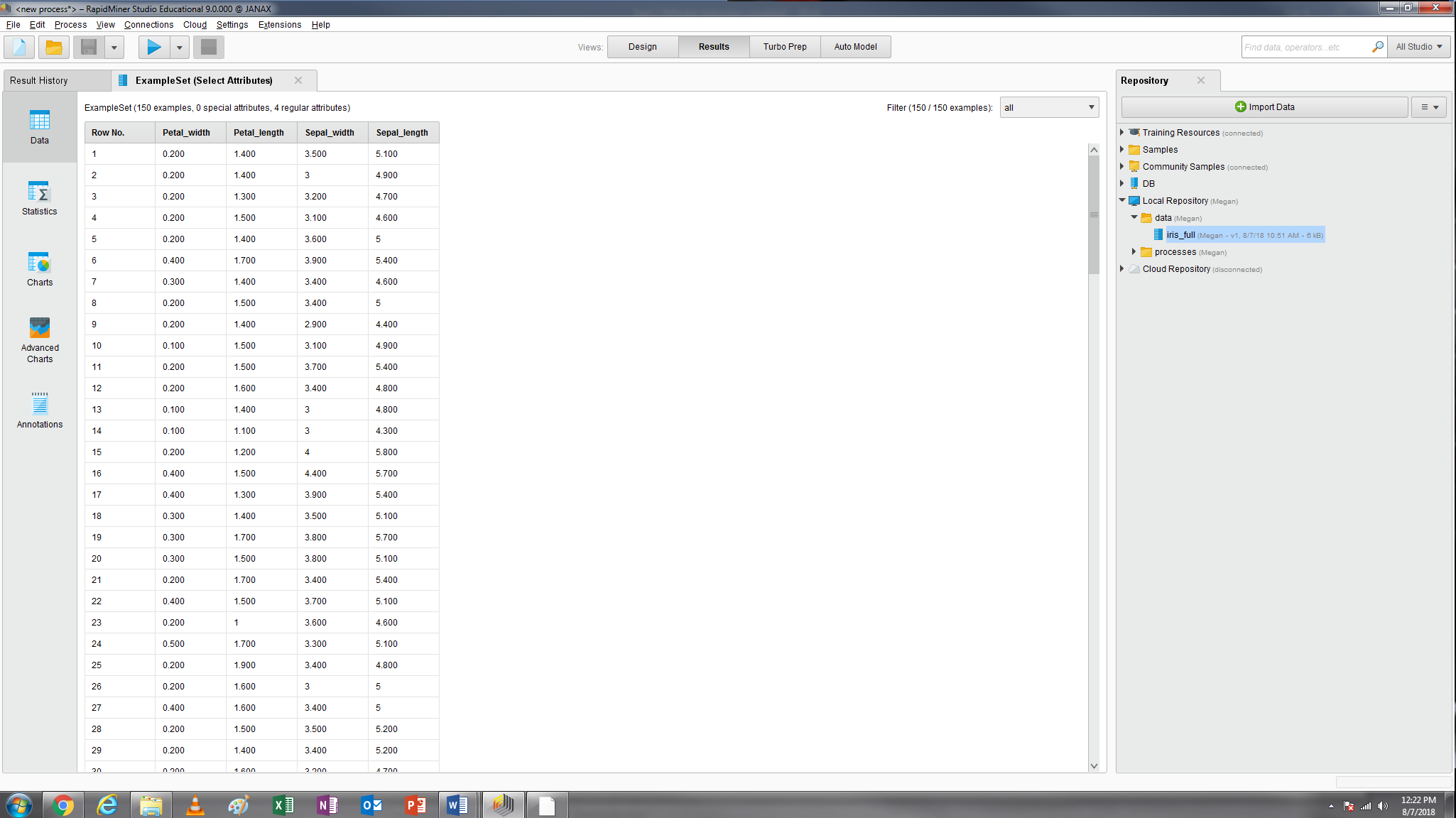
Class 8 Lab & HW3 Clustering with K-means Algorithm

3.4 Answer the question: how many records or examples left in your result?

150

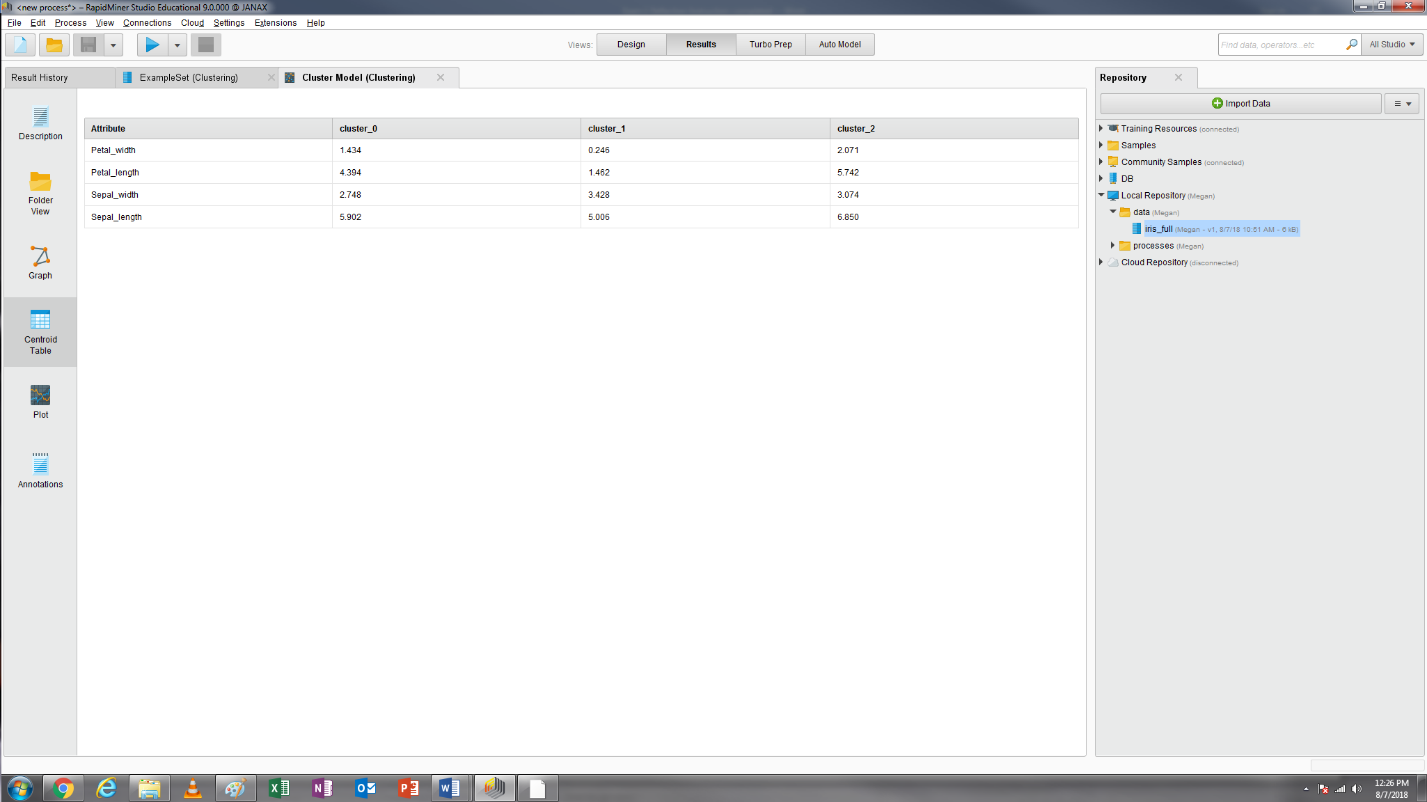
4.7 Answer the question: How many regular attributes left in your result? Take a screenshot of your result page (Screenshot 1)

Four regular attributes



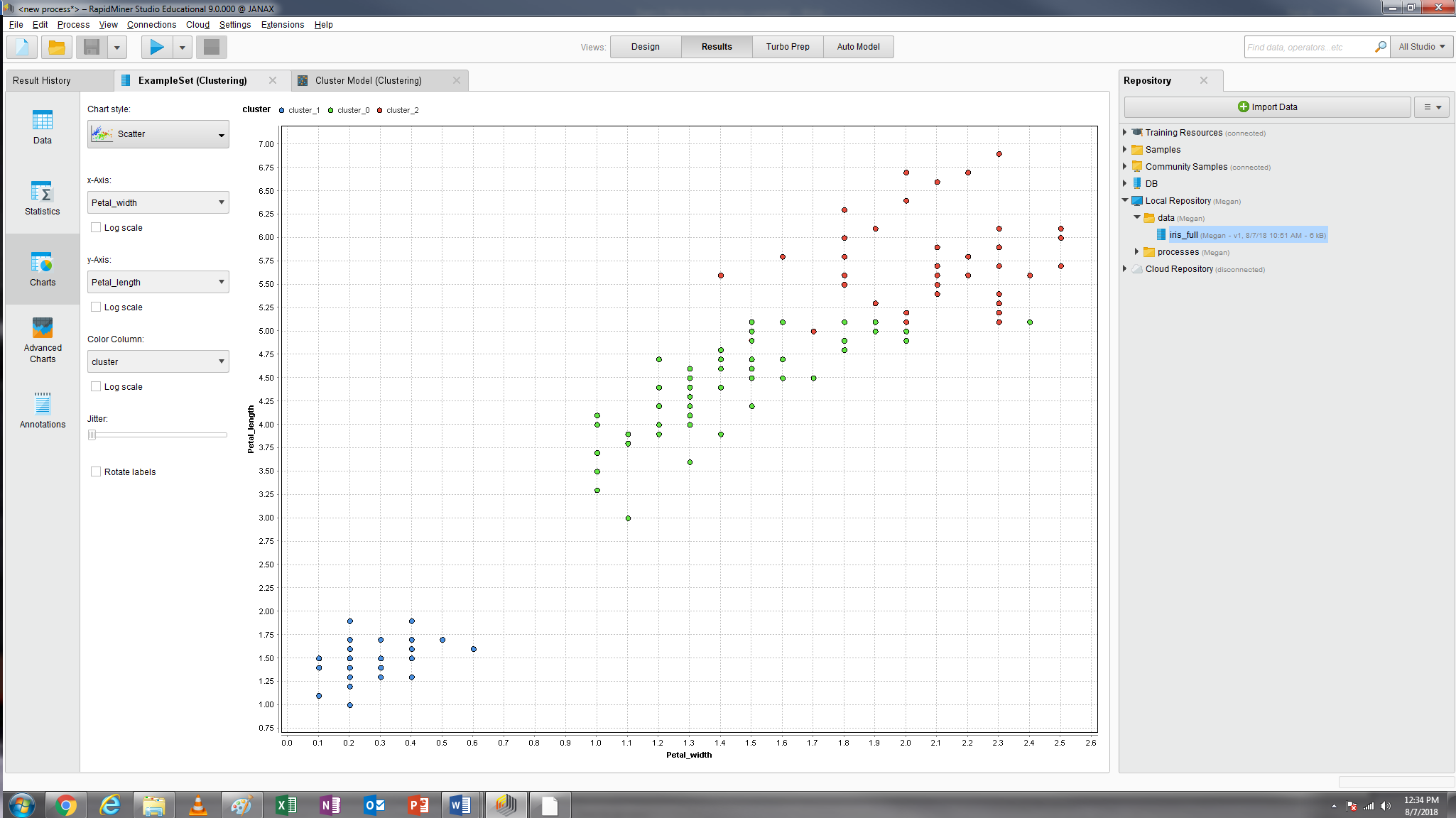
5.7 Under the Cluster Model results, in the Description view, you can see the numbers of individuals in each clusters. In the centroid table view, you can see the centroid values for each cluster. Take a screenshot of your centroid table view (Screenshot 2). Observe the results in Description (at the left sidebar) and Centroid Table (at the left sidebar) to complete the following table (round the centroid to the second decimal place).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Cluster\_0 | Cluster\_1 | Cluster\_2 |
| The number of items in each cluster | | 62 | 50 | 38 |
| Final Centroids in each cluster | Petal\_width | 1.43 | 0.25 | 2.07 |
| Petal\_length | 4.39 | 1.46 | 5.74 |
| Sepal\_width | 2.75 | 3.43 | 3.07 |
| Sepal\_length | 5.90 | 5.00 | 6.85 |

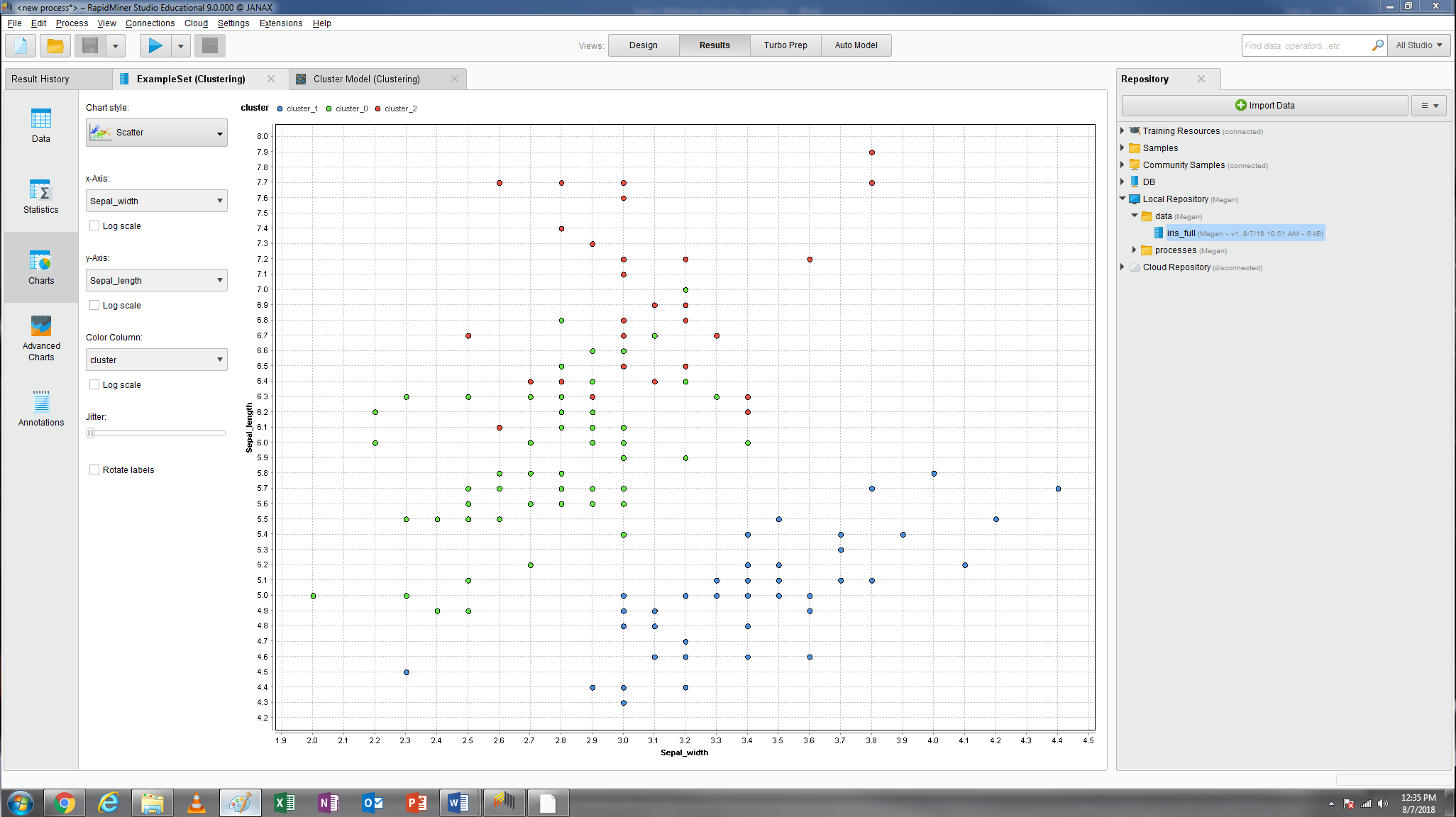


5.8 Please use the Charts View under the ExampleSet results and get the Scatter with each cluster colored, as shown below. Generate three scatterplots using the following three pairs of attributes. The first scatterplot is included as below, you need to generate the other two by your own and take a screenshot of each scatterplot (Screenshots 3 and 4):

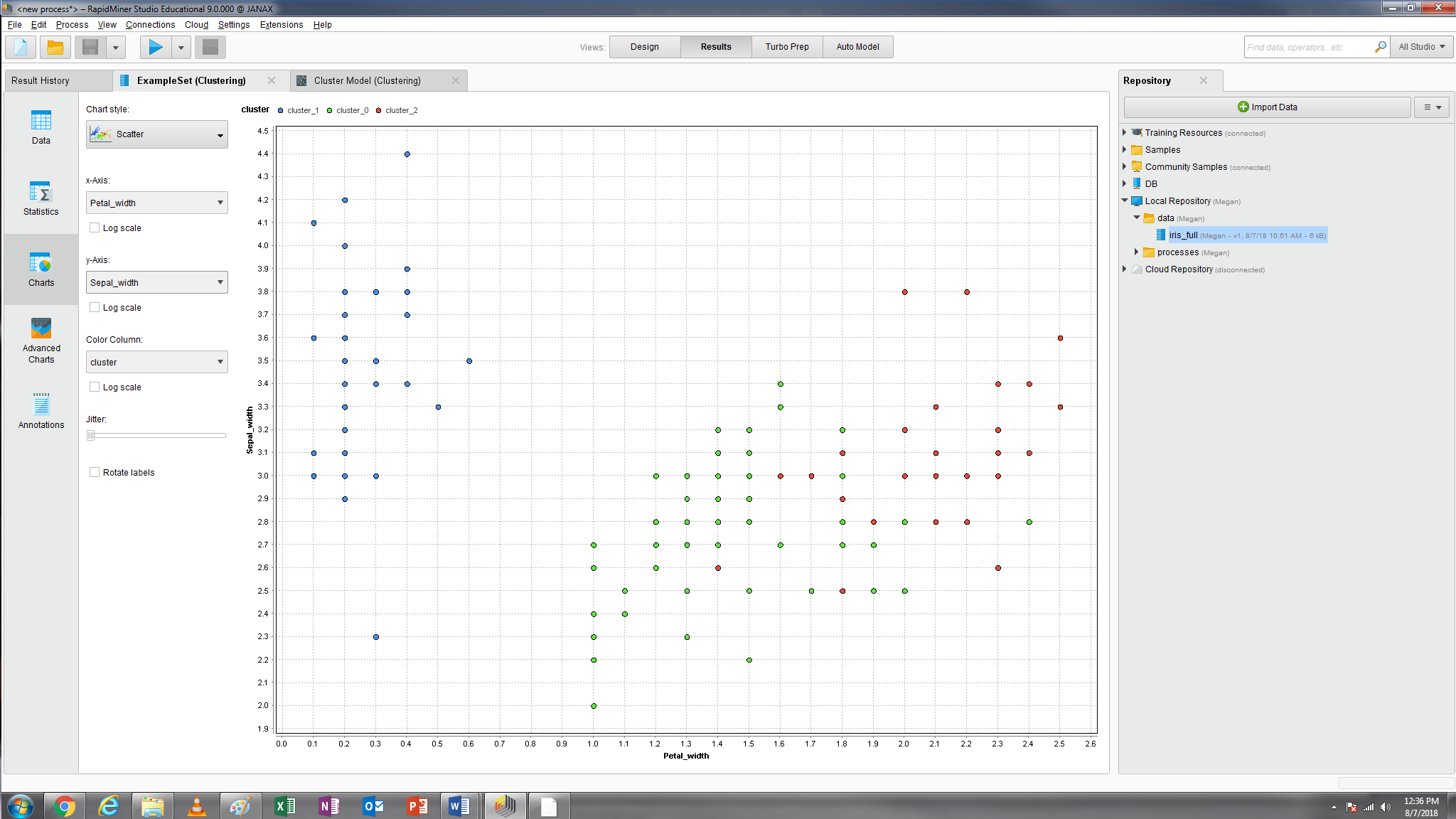
Pair 1: Petal\_width and Petal\_length



Pair 2: Sepal\_width and Sepal\_length



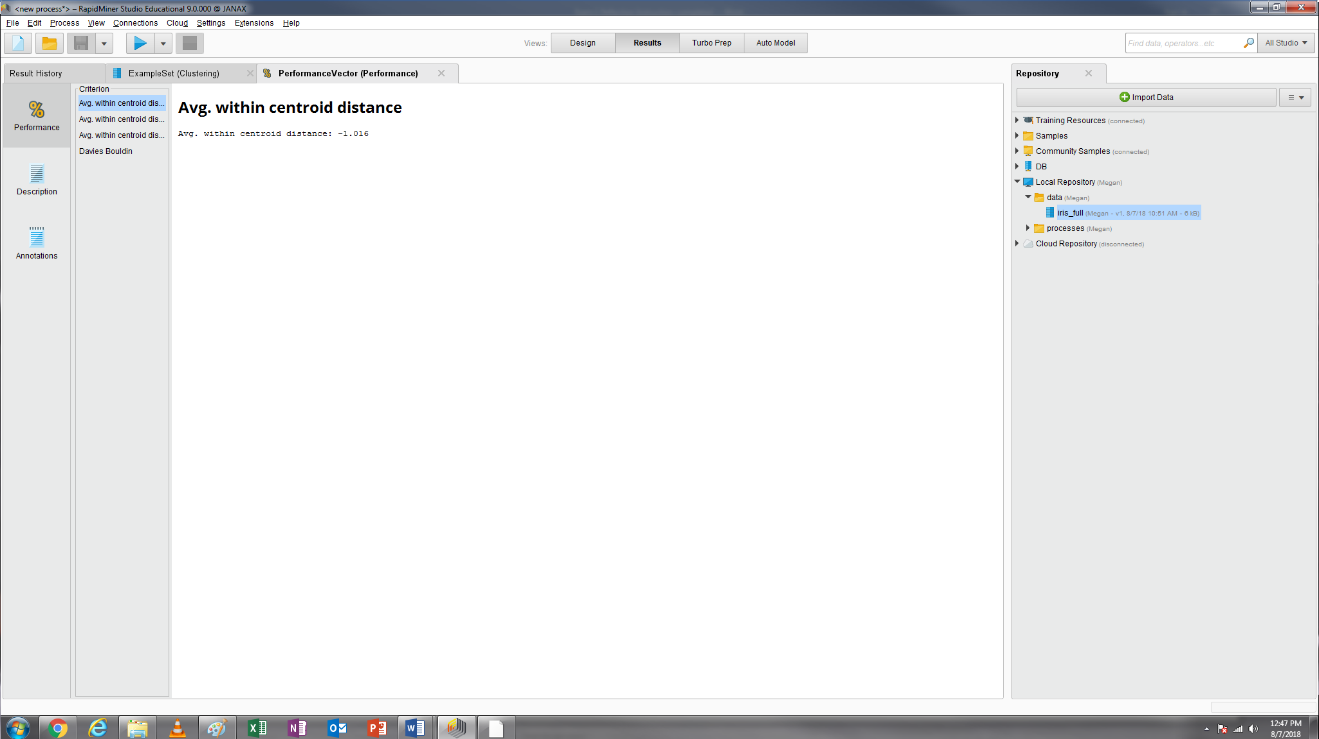
Pair 3: Petal\_width and Sepal\_width



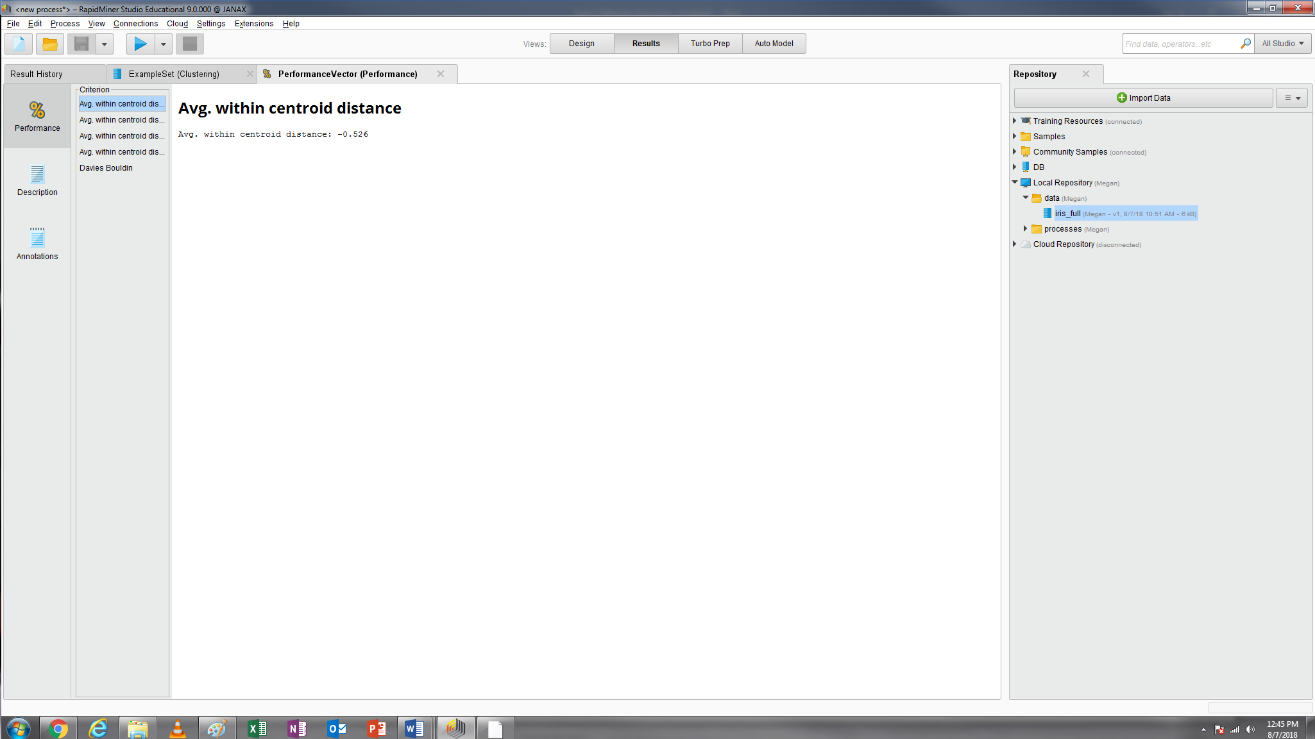
Observe the three scatterplots you generate and then answer the questions: Which pair of attributes does a better job at revealing the distinct clusters? Pair 1: Petal\_width and Petal\_length

6.4.2 Change the k to 2, 5, 8, 10 (after you change the number of clusters, you’d better to press **Enter** on your keyboard to make it really work) while keeping other parameters the same. Take a screenshot of your Avg. within centroid distance (the same the screen as I provide above) for each model (k=2, 5, 8, 10). (Screenshots 5, 6, 7, and 8).

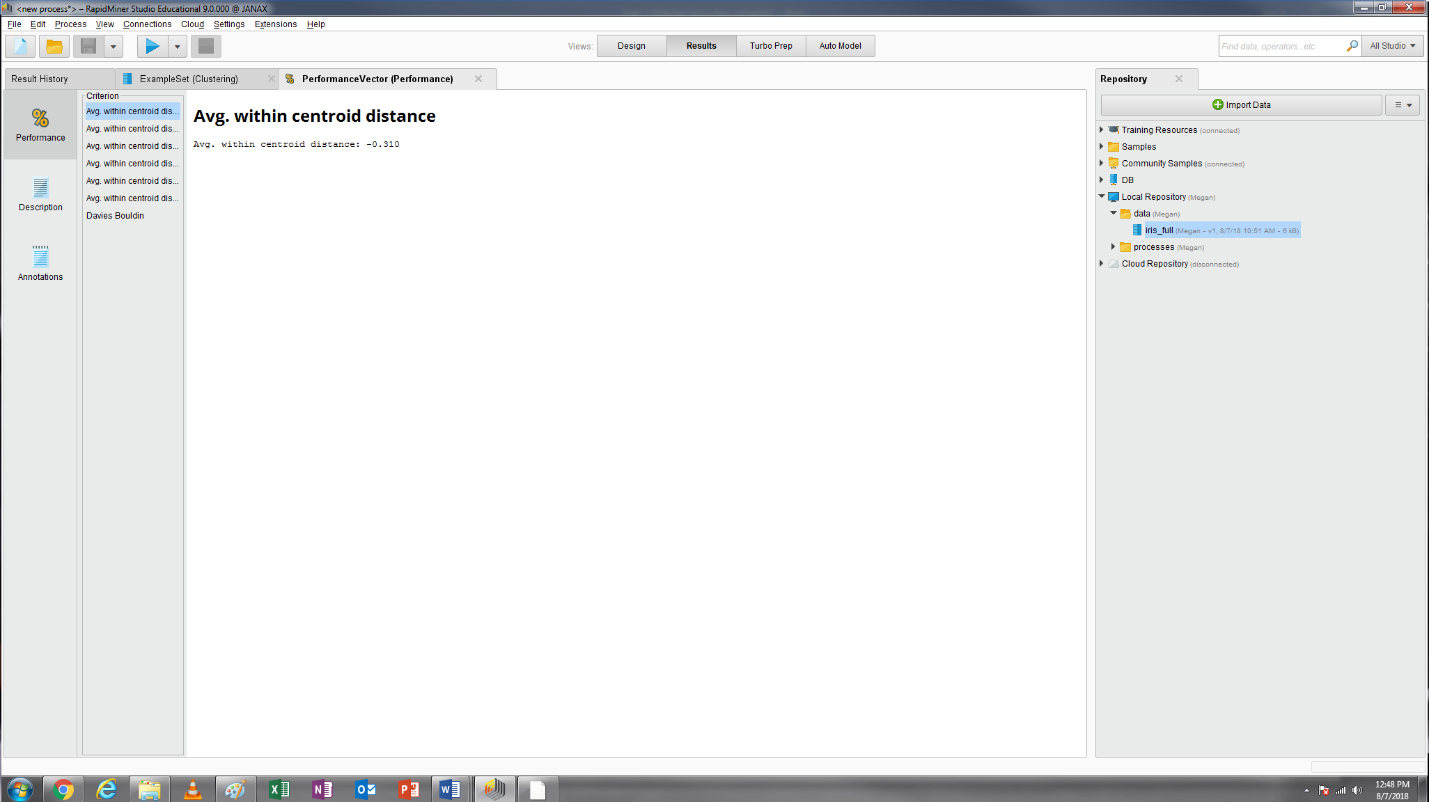
K=2



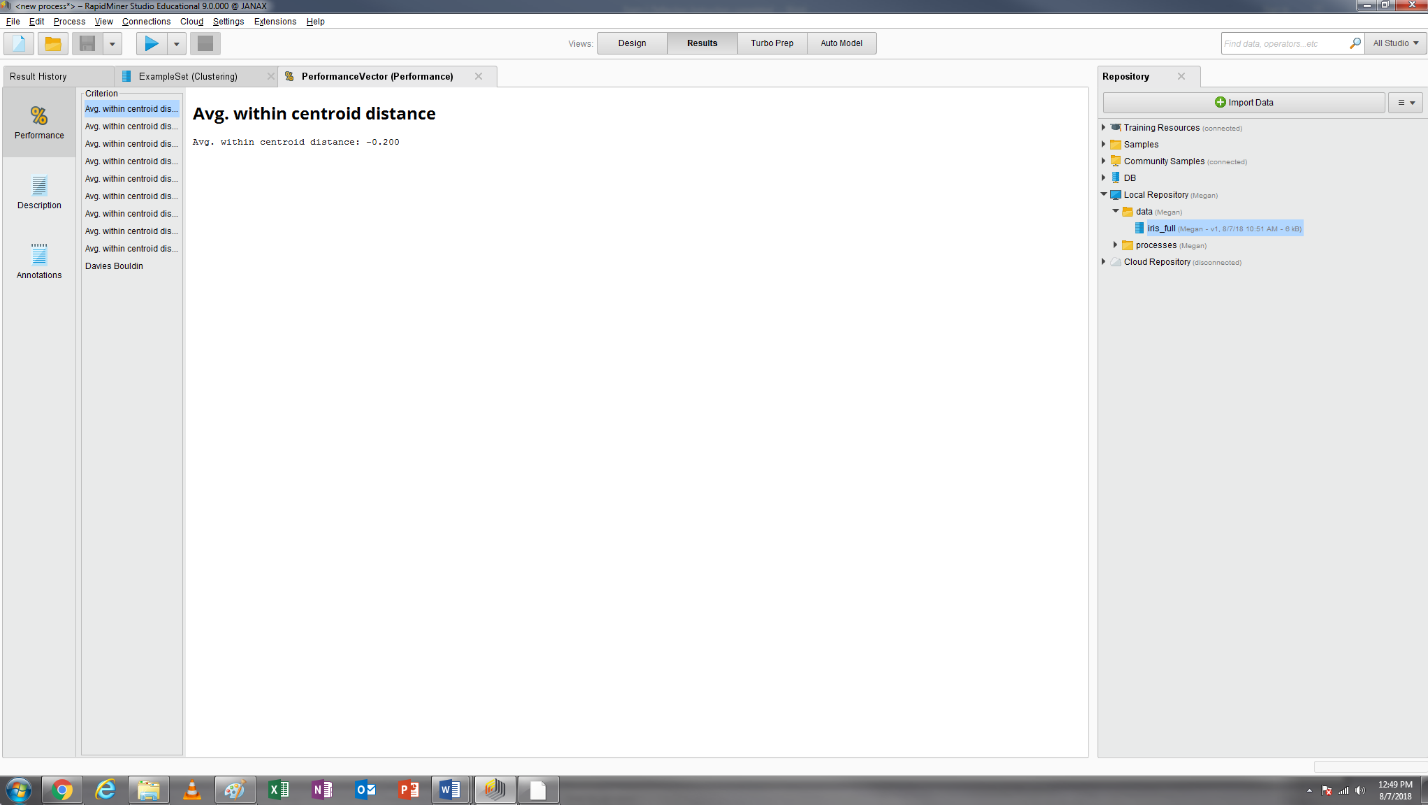
K=3



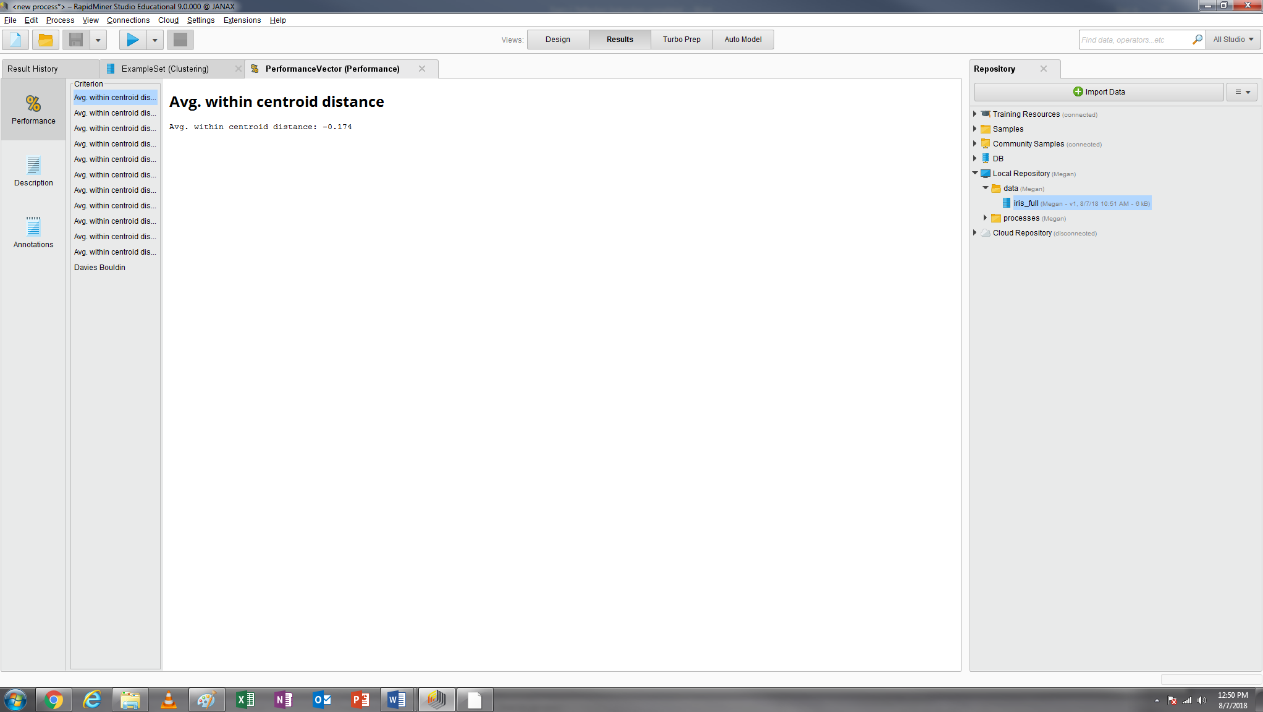
K=5



K=8



K=10



6.4.3 Now, you have five clustering models (k=2, 3, 5, 8, 10). For each k value, write down its Avg. within centroid distance and Davies Bouldin index in the following table (round your answer to the third decimal place).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Evaluation criteria | K=2 | K=3 | K=5 | K=8 | K=10 |
| Avg. within centroid distance | -1.016 | -0.526 | -0.310 | -0.200 | -0.174 |
| Davies Bouldin index | -0.404 | -0.662 | -0.806 | -0.972 | -1.038 |

6.4.4 Please answer the following questions:

a. If you only use Avg. within centroid distance to evaluate the five models, please indicate that which one is the best: k=10

b. If you only use Davies Bouldin index to evaluate the five models, please indicate that which one is the best: k=2

c. Are Avg. within centroid distance and Davies Bouldin index consistent in determining which model is best in this case? Yes or No? No

d. Your boss asked you to combine both measures to generate a new measure, which is calculated by this formula: Avg. within centroid distance\*36%+ Davies Bouldin index\*64%.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Evaluation criteria | K=2 | K=3 | K=5 | K=8 | K=10 |
| New Measure | -0.624 | -0.613 | -0.627 | -0.694 | -0.727 |

Based on this new measure, please indicate that which one is the best: k=3

e. In the original dataset, there is an attribute “Species\_name”, which was assigned by the biologists. How many species did the biologists generate? three

f. Based on the information given in question e, which model do you think is the best? K=3