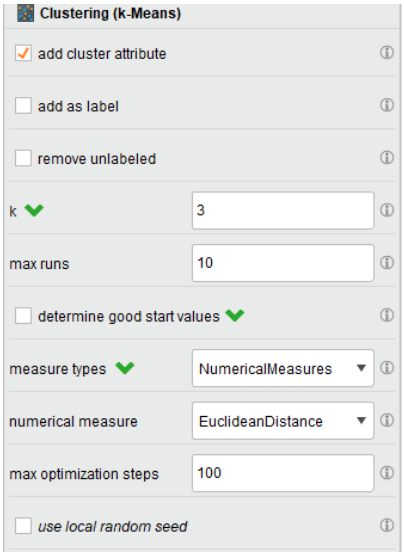
CIDM 6355 Week 9 RapidMiner Lab - Hierarchical Clustering

Purposes:

- Understand some advanced issues about K-Means algorithm
- Understand and Apply Hierarchical clustering algorithms
- Interpret Hierarchical Clustering results
- Conduct Post-Clustering analysis

Dataset: Please download the dataset titled "CarsCSV.csv" from WT Class to use it with this lab session.

Requirements: The same requirements with previous Lab sessions, especially screenshots must be taken with date and time there.



1. Review K-Means Algorithms

1.1. Import the data file to a new RapidMiner process using the Read CSV operator and run it.

1.2. Check if there are any data quality issues such as missing values under the Statistics view.

1.3. Select the following three attributes: “cubicinches”, “weightlbs” and "cylinders", using the Select Attributes operator.

1.4. Develop a k-means model using the Clustering (k-Means) operator with the parameters listed in the right screenshot. Save your process as Lab9_Step 1.4.

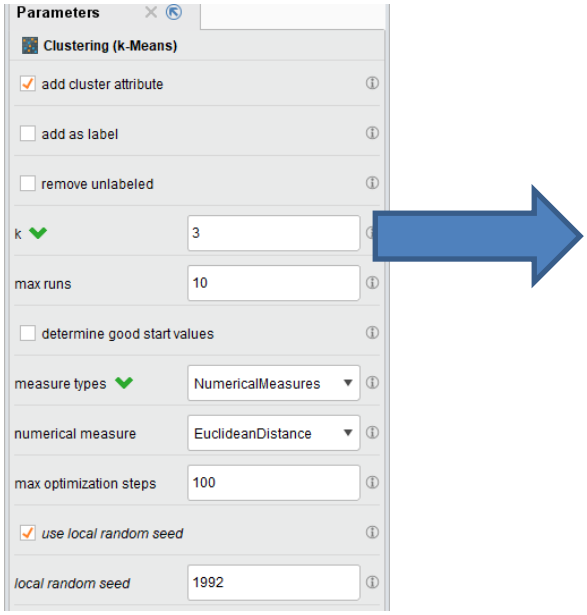
1.5. Run your process and take a look at the centroid table. Please record the cluster size under the Descriptive view.

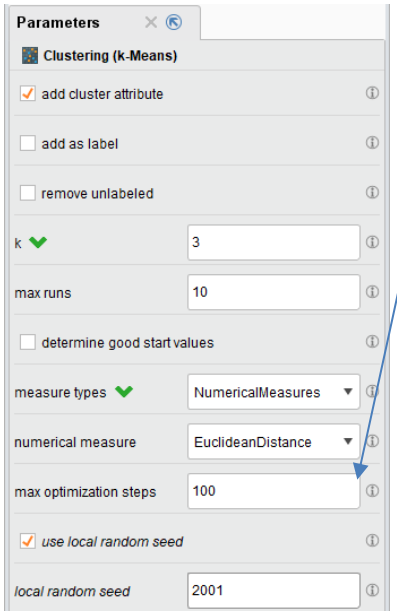
1.6. [Pause and Ponder] What could be a potential issue this model might encounter? You do not need to submit an answer to this question.

2. Selection of Initial Centroid in K-Means Algorithms

2.1. Continue your process in Step 1.4. Check “use local random seed” and set it as 1992 to run this process again. Record the cluster size and cluster name.

2.2. Set the local random seed as 2001 and run it. Record the cluster size and name. When comparing the cluster size and name with those at Step 2.1, you may find that the results might be slightly different.

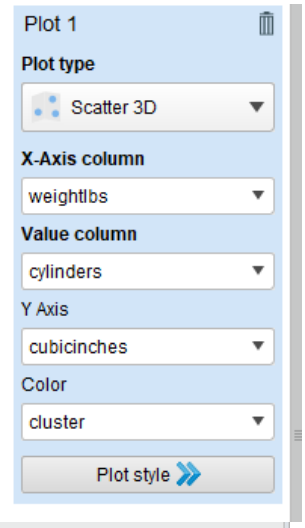




Note: This parameter indicates if a *local random seed* should be used for randomization of the *k* different **starting points** of the algorithm, which involves the selection of the initial centroid.

2.3. Note: With the same local random seed, you are supposed to get the same result each time (but not always). We will use 2001 as the default local random seed because it is also the global random seed for random generators (when you move your cursor to the blank area in your process and then left click it, you will see the parameters for the process, in which 2001 is specified as the random seed.

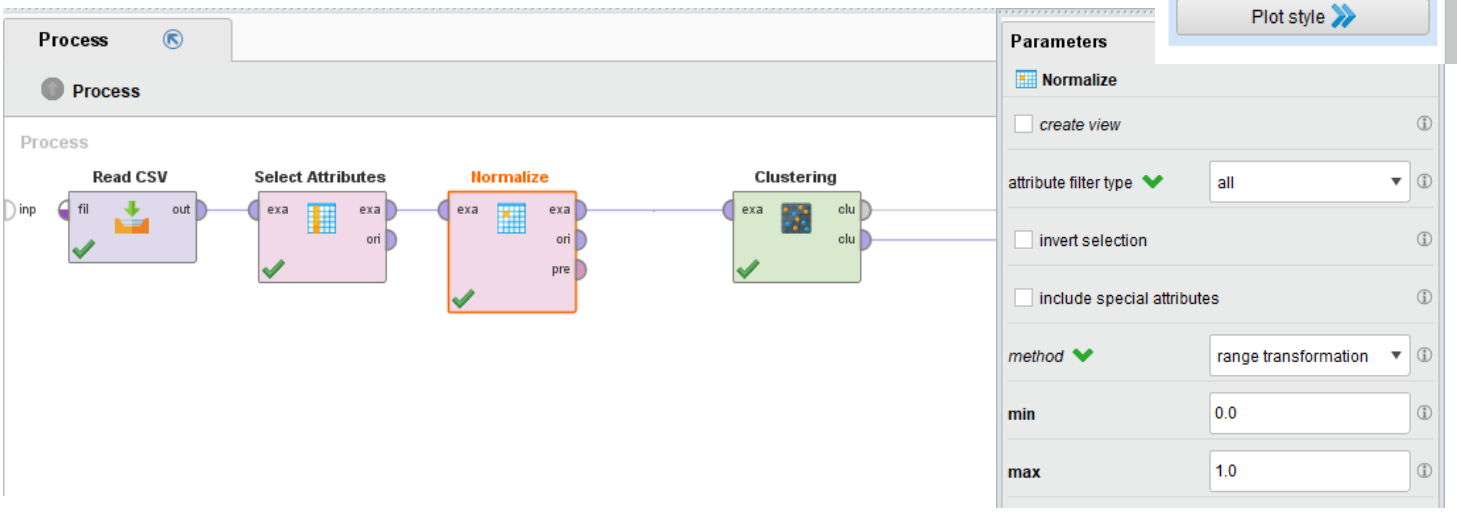
3. Interval-Scaled Attributes in K-Means Algorithms



3.1. Continue your process and clustering result at Step 2.2. Click the Visualization view under ExampleSet(Clustering), please generate a Scatter 3D plot with the settings specified at the right. Take a look at your Scatter 3D plot.

3.2. Rotate the 3D Scatter Plot using your cursor to view it from different angles. [Pause and Ponder] What issues or observations do you notice? Consider questions like: How are clusters distributed across each dimension? Do all attributes equally contribute to the clustering model, or is there one attribute that dominates the model?

3.3. Re-scale the three attributes by using the Normalize operator. Use range transformation as the method with min 0.0 and max 1.0 (see the parameters as below). Save the process as Lab9_Step 3.3.



3.4. Run this process and take a look at the Statistics view under ExampleSet(Clustering). You will find that the minimum and maximum values for the three attributes “cubicinches”, “weightlbs” and "cylinders" are 0 and 1, respectively.

3.5. Obtain the size for each cluster under Cluster Model (Clustering) and Take a look at the centroid table.

3.6. [Pause and Ponder] Compare this model with the one in Step 2.2. What differences can you find in terms of cluster size and centroid table? Which model makes more sense? Why? You do not need to submit an answer to this question.

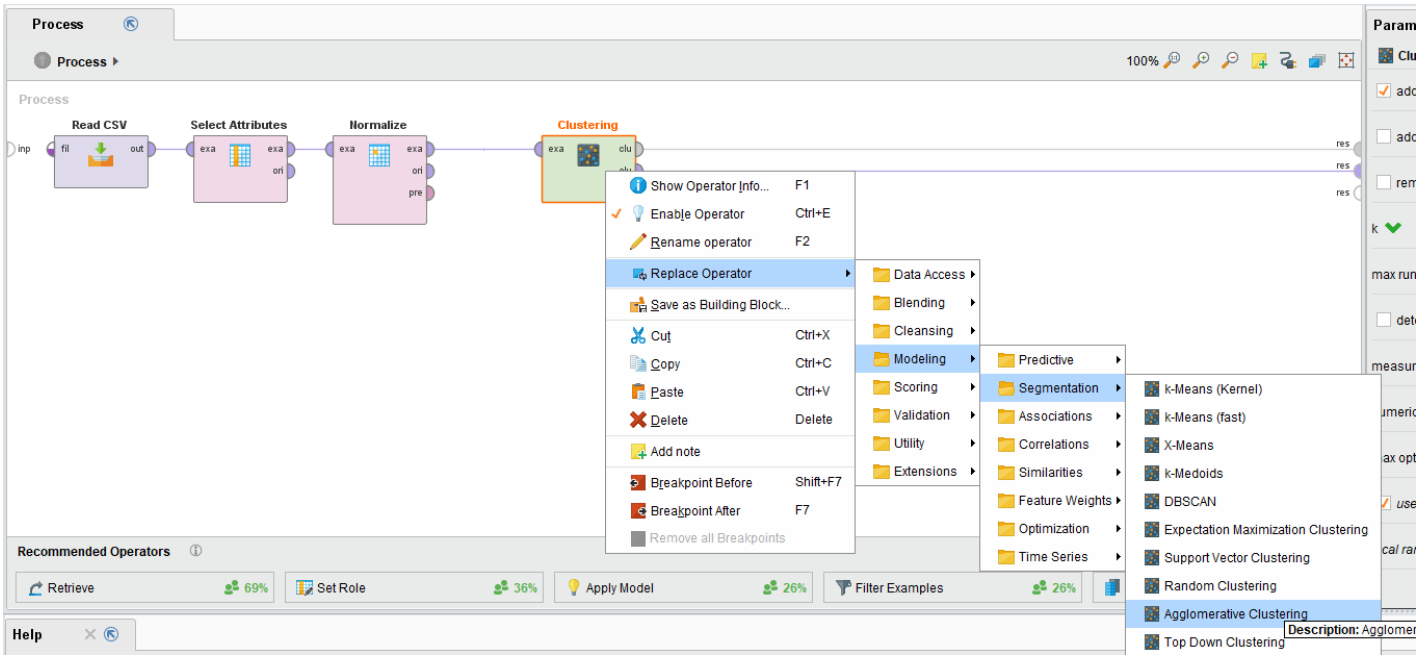
4. Agglomerative Clustering

Now, we are going to practice Agglomerative Clustering. Let’s start with a review of two strategies for hierarchical clustering:

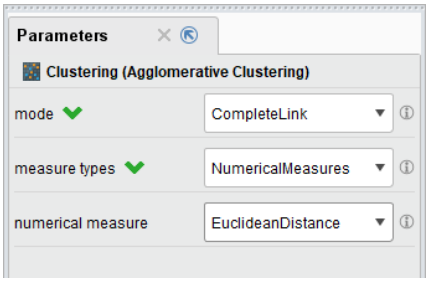
- Agglomerative: This is a bottom-up approach, in which each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. This type of clustering is implemented in RapidMiner as the Agglomerative Clustering operator.

- Divisive: This is a top-down approach, in which all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

4.1. Go back to your process, Lab9_Step 3.3. Right click the operator Clustering (K-Means) and then Replace it by the operator Agglomerative Clustering (Ignore any possible warning).



4.2. Use the paramters in the right for the operator Clustering (Agglomerative Clustering). Here, we use Complete Link. Attention: unlike K-Means, this operator does not have a random seed, which means that we are supposed to get the same results each time.



4.3. Run this process and take a look at the Description view and the Dendrogram View.

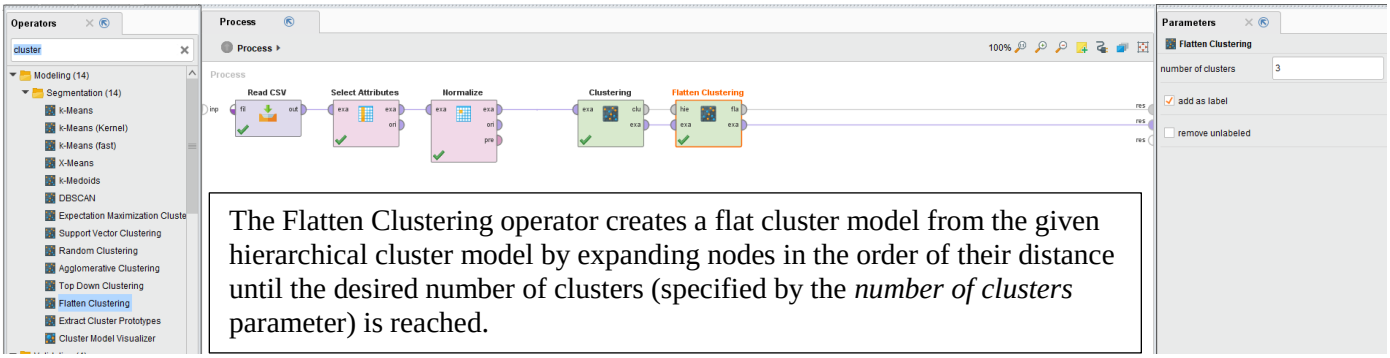
4.4. Agglomerative Clustering allows us to generate any number of clusters (no more than the number of items). In the lecture video, I show you how you can generate the required number of clusters from the dendrogram. If you are asked to generate five unique clusters from this dendrogram manually, how you are going to do? Please show it in your dendrogram and then take a screenshot of it with date and time (Screenshot 1). Which cluster has the largest number of records? Please label it in your dendrogram. You may refer to the Appendix at page 7.

4.5. Now, add the Flatten Clustering operator into your process, after the Clustering operator, as shown in the screenshot at the next page. Flatten Clustering helps you cut the dendrogram at a specified cluster number. This is useful for creating a certain number of clusters.

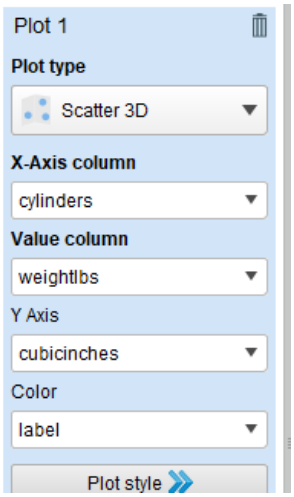
4.6. In the parameters of the Flatten Clustering operator, please type 3 as the number of clusters and Check the "add as label" configuration, which indicates that the cluster role is stored as a label in the ExampleSet. S

4.7. Save this process as Lab9_Step4.7 and then run it. Under Cluster Model, please record the size of each cluster.

4.8. Under the ExampleSet, please Copy and paste the data into an Excel spreadsheet and then save it as an Excel file. You will need this for Step 4.12 and Deliverable R3 in the R lab. Alternatively, you can add "Write Excel" operator to write the ExampleSet into your local computer (Please refer to the end of Week 8 RM Lab Instruction for more details).



4.9. Under the ExampleSet, please click the Visualization view and then generate a Scatter 3D Plot with the settings specified in the right. You can explore the plot by holding down the button and moving around.



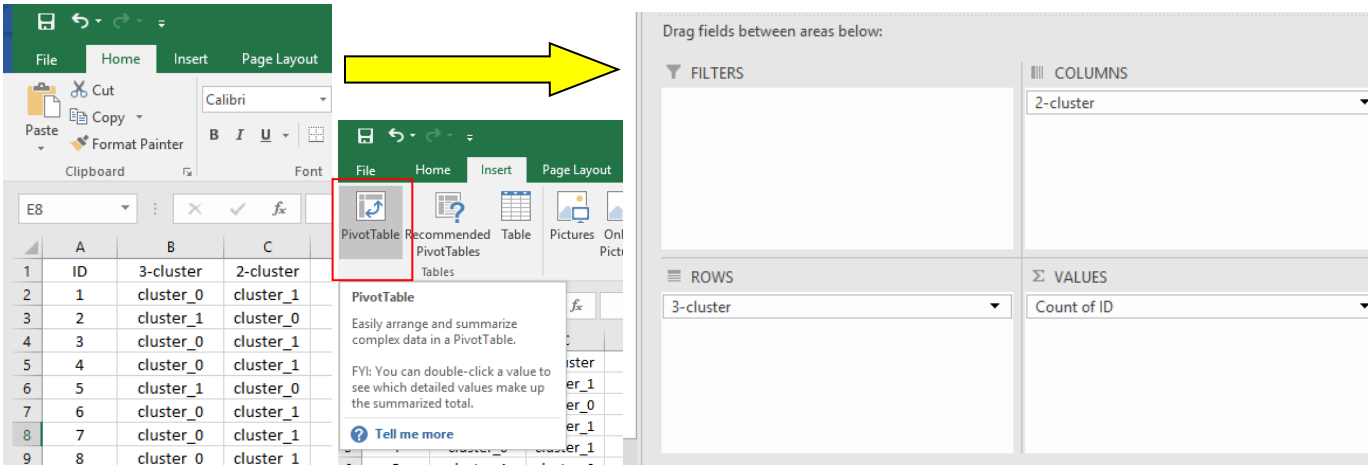
4.10. Change the number of clusters from 3 to 2 and keep other parameters unchanged. Run this process again. Record the cluster size and generate a same Scatter 3D plot.

4.11. Compare the two 3D Scatter Plots (Steps 4.9 and 4.10) and then think about how many clusters are better, 2 or 3. Why?

4.12. When you observe and compare the number of data objects in each cluster for the two models. You may find one cluster in the 2-cluster model has the exactly same number of data objects as one cluster in the 3-cluster model (cluster 1). Do the two clusters include the same data objects and why? This question will help you understand how the two clusters in the 2-cluster model evolve into the three clusters in the 3-cluster model. In order to answer this question, I provide you with theoretical and empirical examinations as below.

- Theoretical examination: go back to the dendrogram in your first deliverable. Imagine that you are going to cut the dendrogram to the two and three clusters respectively using a "scissor". When you move down your "scissor", you will find that one cluster keeps the same while the other one is divided into two clusters.

- Empirical examination (compare them in an Excel file): you can copy and paste the label column (sorted by the ID column) in the ExampleSet (Flatten Clustering) under each cluster model in an Excel file. Then, insert a PivotTable to show their relationship. See the screenshots below. Take a screenshot of your PivotTable for the empirical examination with date and time (Screenshot 2). What conclusion can you make based on the PivotTable?



	A	B	C
1	ID	3-cluster	2-cluster
2	1	cluster_0	cluster_1
3	2	cluster_1	cluster_0
4	3	cluster_0	cluster_1
5	4	cluster_0	cluster_1
6	5	cluster_1	cluster_0
7	6	cluster_0	cluster_1
8	7	cluster_0	cluster_1
9	8	cluster_0	cluster_1

4.13. Now, change the number of clusters back to 3 and then use the "SingleLink" as the mode under the Clustering operator parameters. Then, run this process.

4.14. [Pause and Ponder] Compare the clustering result with the one in Step 4.7 and then answer the following questions: what is the difference between single link and complete link? Does the use of SingleLink and CompleteLink generate different clustering results in this case? Why? You do not need to submit an answer to this question.

5. Post-Clustering Analysis

Post-clustering analysis is quite useful in practice. You are asked to test the difference among three clusters with respect to their means for the mpg attribute.

5.1. Similar to the previous step, import the "CarsCSV.csv" data file to a new RapidMiner process.

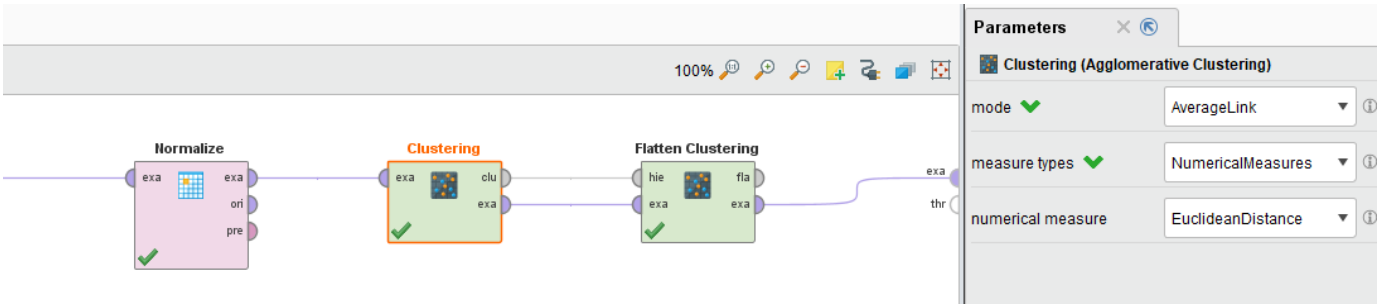
5.2. Add the Work on Subset operator to your model and connect it to the Read CSV operator. Work on Subset is a multi-layer operator that allows us to select specific attributes for analysis and then combine the results with the rest of the attributes.

5.3. Configure the operator Work on Subset as:

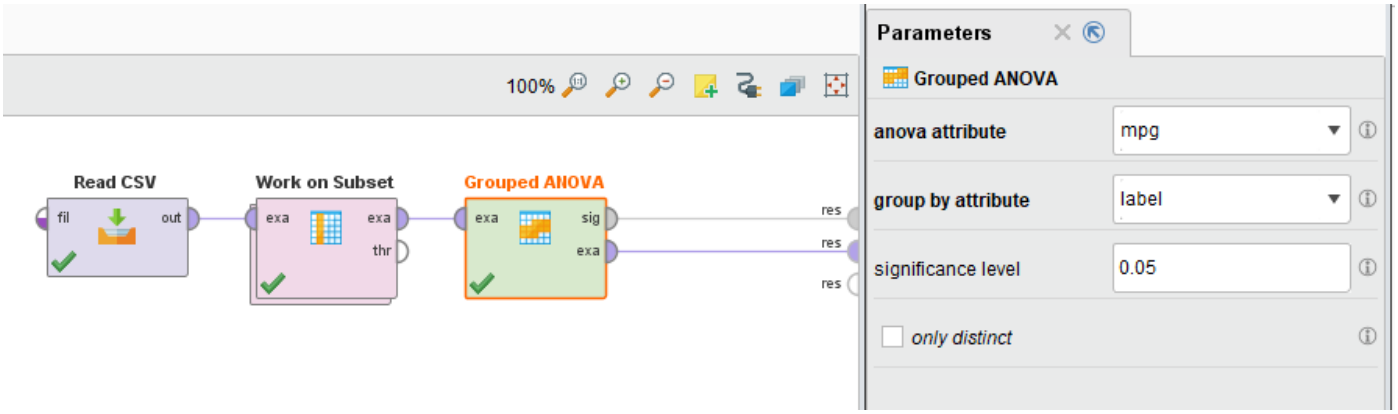
- Attribute filter type: subset
- Attributes (Selected attributes): "cubicinches", "weightlbs" and "cylinders".
- All the other boxes are **unchecked**.

5.4. Double click on the operator Work on Subset to go to the subset process layer. Create the clustering process here by using Normalize, Agglomerative Clustering, and Flatten Cluster operators, similar to those previous steps. Make sure you use the following parameters for each operator,

- Normalize method is range transformation with 0.0 as the minimum and 1.0 as the maximum.
- Clustering mode is "**AverageLink**" with "NumericalMeasures" type and "EuclideanDistance" measure.
- Number of clusters is "3" for Flatten Clustering and **"Add as label" is checked**.
- Below is the screenshot for this sub-process;

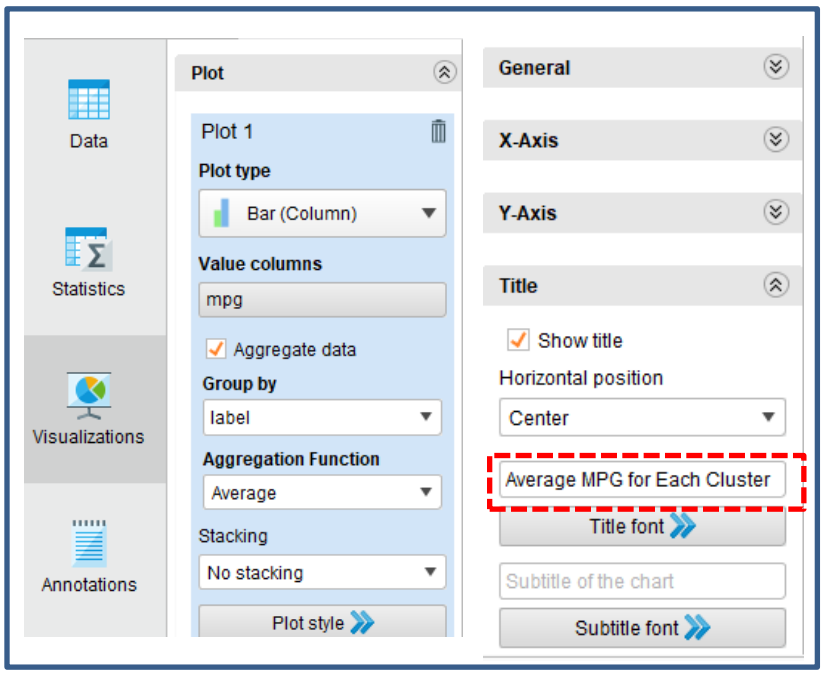


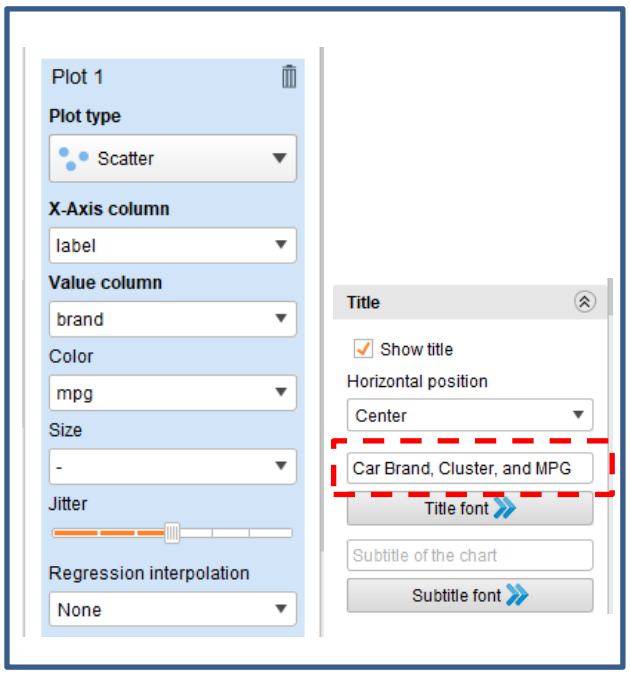
5.5. Now, go back to the main process and add the Grouped ANOVA operator after Work on Subset and link the "exa" ports. Configure parameters for Grouped ANOVA as below:



5.6. Save the process as Lab9_Step5.6 and run it.

5.7. Go to the Visualization view under ExampleSet and make a Bars chart (change the chart title to Average MPG for Each Cluster) and a scatter plot (change the chart title to Car Brand, Cluster, and MPG) using the following settings.





5.8. Take a screenshot of your column and scatter charts with date and time (Screenshots 3 and 4). What conclusion can you make from each of the two charts?

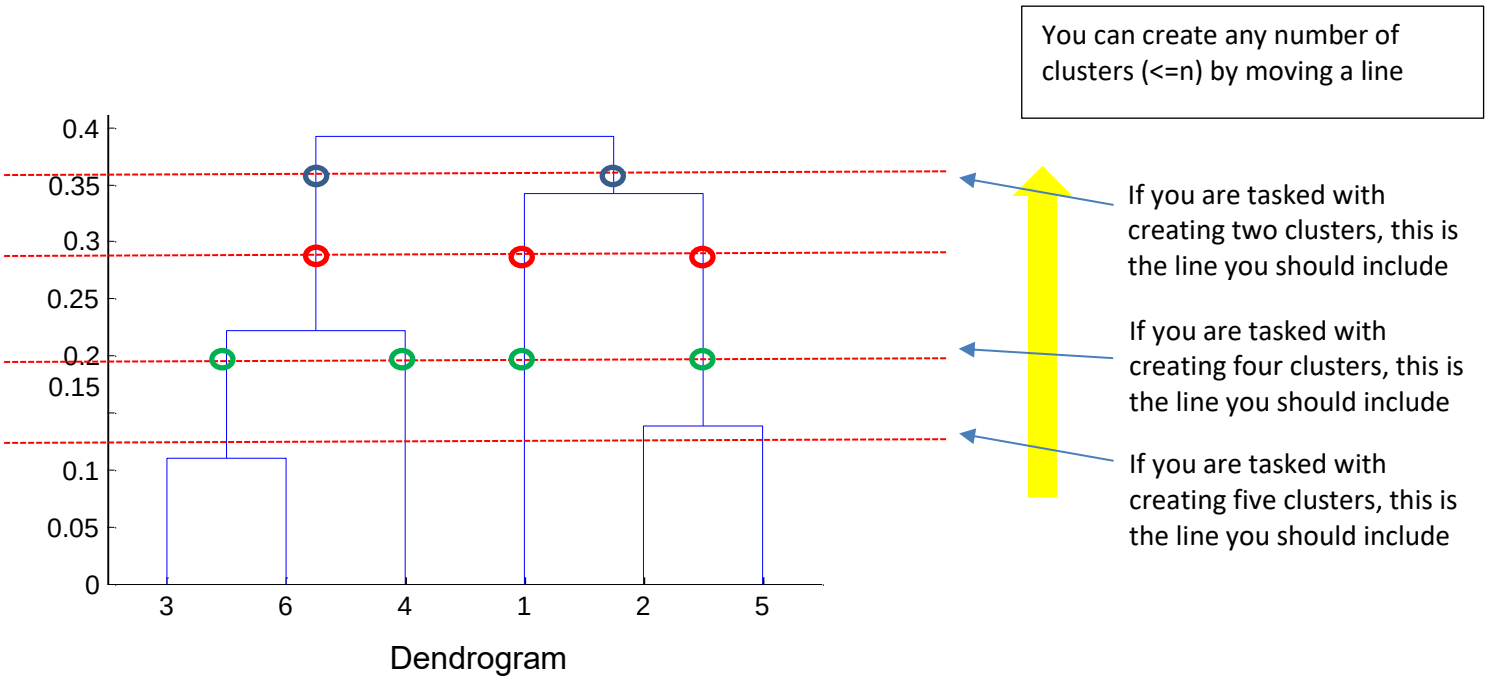
5.9. Take a screenshot of the ANOVA Test table with date and time (Screenshot 5). Interpretation: Suppose that you have the following competing hypotheses:

- H_0: All the three clusters of cars have the equal average mpg.
- H_1: At least two clusters of cars have different average mpg.

Based on the ANOVA table, do you think the mean mpg of the three clusters differ at the 95% confidence level? Why? Note, if you are not familiar with ANOVA, please check ANOVA slides at WTCLASS or other online resources.

Appendix: How to Cut Dendrogram into any number of clusters

First of all, take a screenshot of your dendrogram from RM.
Next copy and paste your dendrogram in an MS WORD file and then draw a line to show how you generate the requested number of clusters.



When you are asked to create a specific number of clusters, you just need to show a line, instead of multiple lines.

You may wonder what the numbers in the y-axis mean. Please refer to the final slide in our R Lab Instruction.