Badge-3 Lab-2 [Clustering]

Out date: Aug 3, 2022
Due date: Aug 7, 2022 at 11:59PM

Submission

1. Prepare your solution in Orange and save the workspace for Problem 1 (e.g., Badge3_Lab-1_LastName.ows) **[20 points]**
2. Complete the tables given below and save the file (e.g., Badge3_Lab-1_LastName.docx). **[80 points]**
3. Upload the files to the Canvas.

Objective: To cluster the given data sets using Hierarchical and k-Means clustering methods and understand the clusters.

Problem 1 [100 points]

Data: For this lab, please download *1_GOMFields_Reserves_Processed.csv* file from Canvas to your folder.

(**Data Source:** https://www.data.bsee.gov/Main/FieldReserves.aspx#ascii)
Oil & BOE reserves and production in MMbbl
Gas reserves and production in Bcf
Field GOR in SCF/STB

Lab Instructions

1. Open Orange and load the *1_GOMFields_Reserves_Processed.csv* file using the **File** widget.

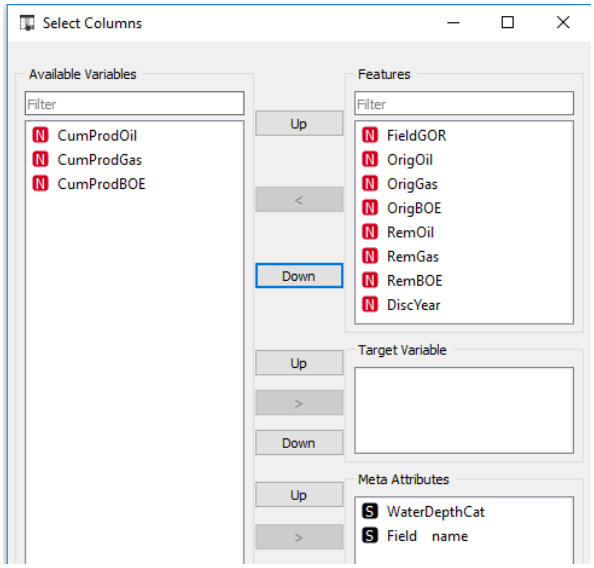
2. Open **File** widget and load the data. Complete the table below:

# of instances/rows	1319
# of features, attribute types and roles	13 features (11 numeric, 1 cate, 1 meta)

3. Add **Data Table, Feature Statistics and Distributions** widget and inspect features. Complete the table below:

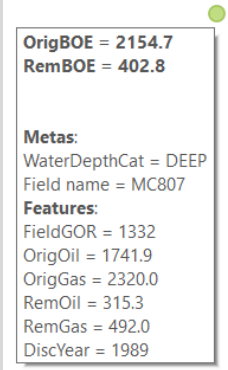
List 3 observations based on inspecting the features	No missing data, features are mostly skewed, the target has imbalances, target variable will appear to the left in data table

4. Add **Select Columns** widget and ensure selection of variables is as below:

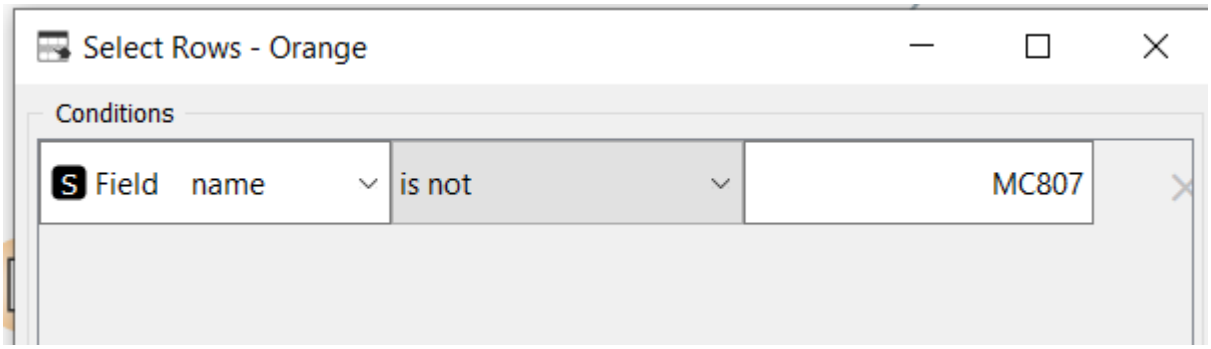


Add **Scatter Plot** widget to the pipeline and open this widget. Examine relationship between the variables OrigBOE and RemBOE. Add DiscYear to Color. Add Field name to the Label.

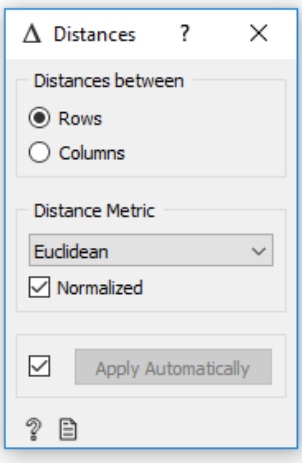
List 3 observations:	Each dot is a field, there is one outlier way to upper right, clustering sensitive to outliers, consider fields with row remaining oil as depleted fields

What is the name of the outlier field?	 MC807

5. Add **Select Rows** to the <u>pipeline and filter out the outlier</u>.



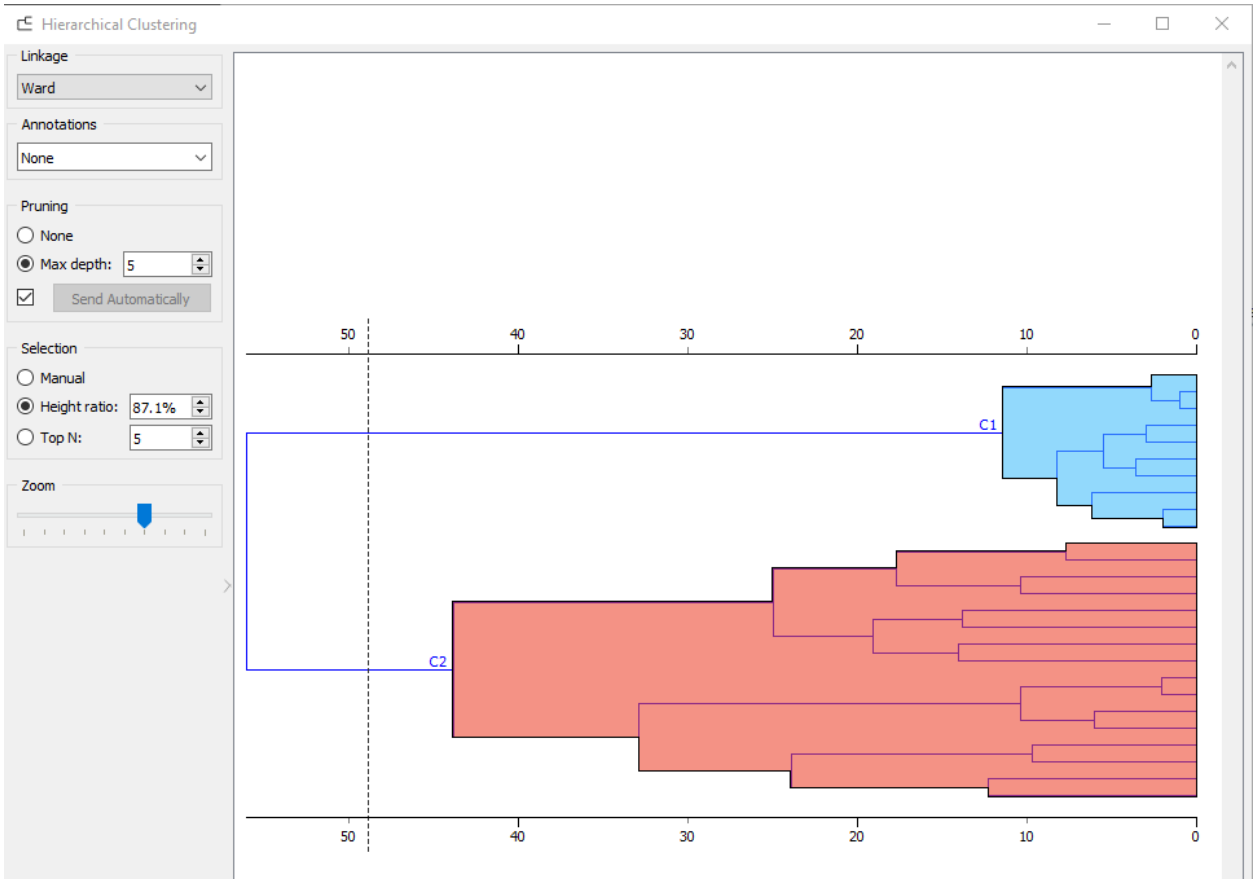
6. Add **Distances** widget to the pipeline. Open and set the parameters as below:

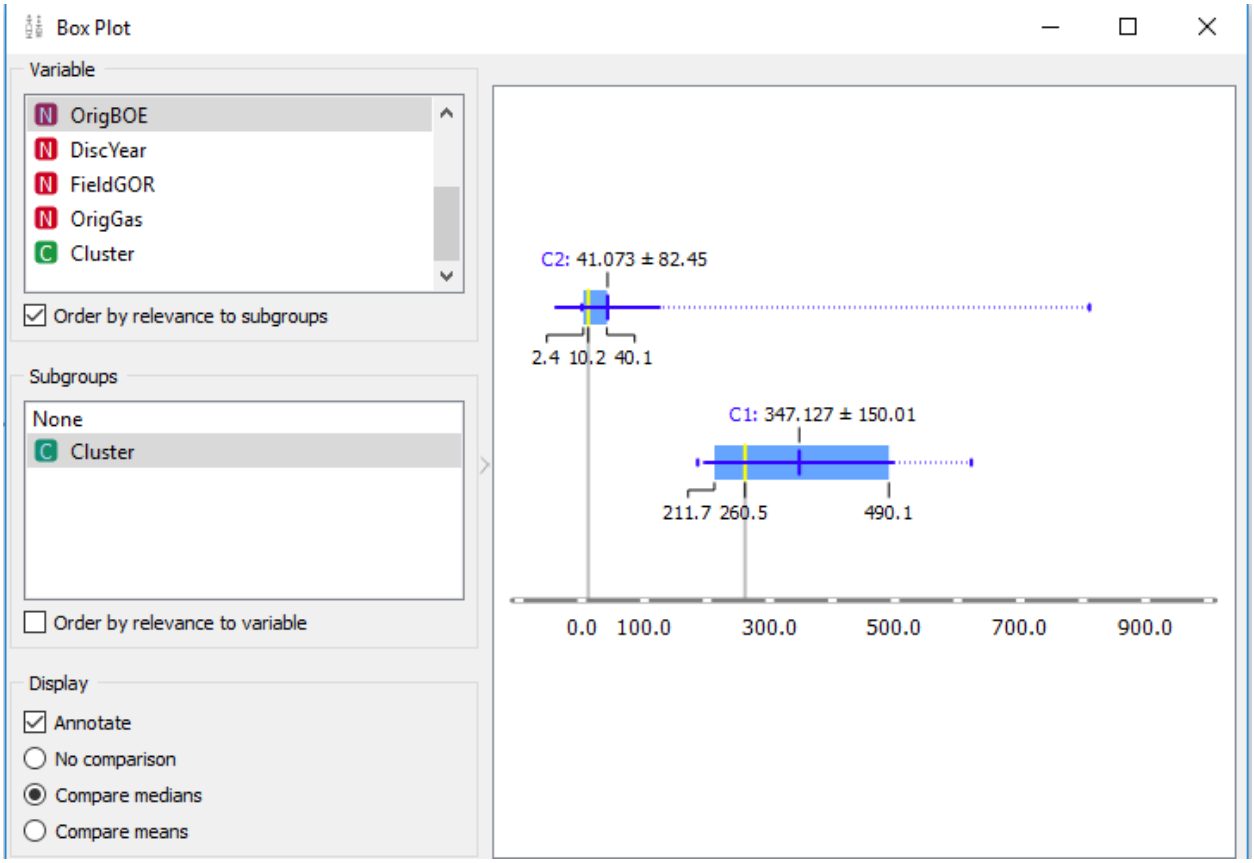


7. Add **Hierarchical Clustering** widget to the pipeline.

8. Add **Box plot** widget to the pipeline.

9. Open **Hierarchical Clustering** and **Box plot** widgets.

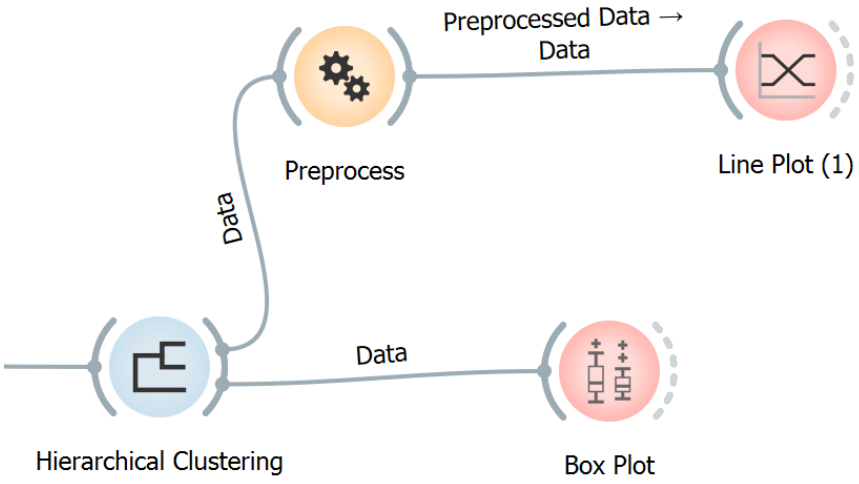




10. Tune clustering parameters and observe cluster results in the box plot. Complete the table below:

Linkage Type	Variable (Box Plot)	# Clusters	Cluster Means
Average	RemBOE	2	C1: 171.46, C2: 1.654
Average	OrigBOE	2	C1: 347.13, C2: 41.073
Average	OrigGas	2	C1: 193.745, C2: 146.221

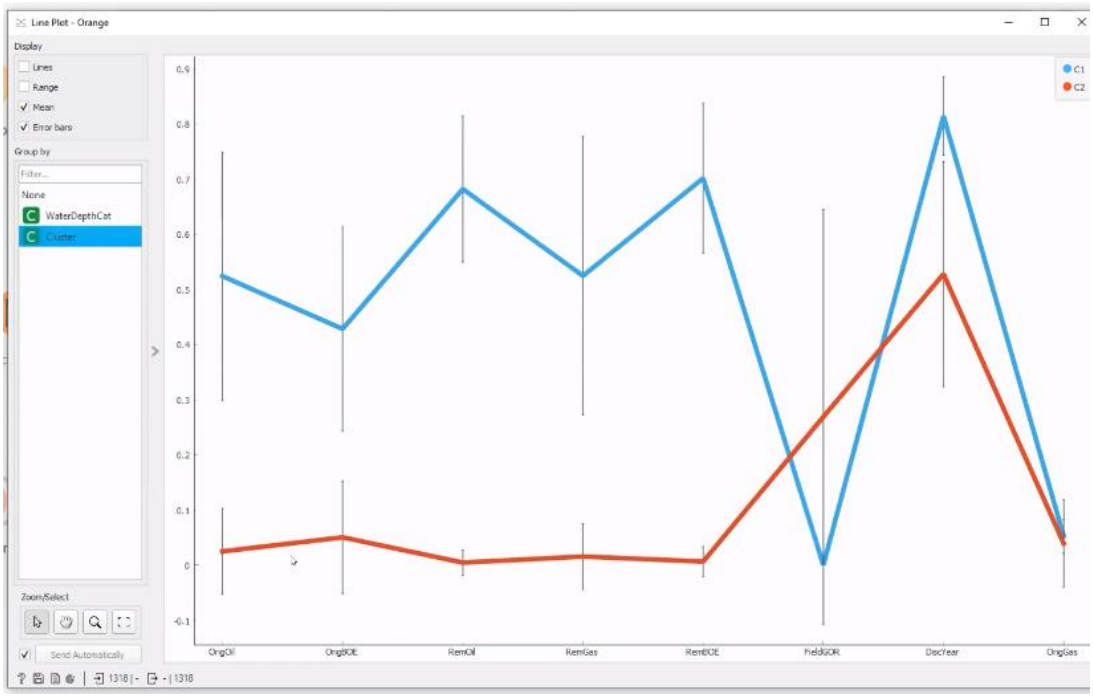
11. Add **Preprocess** and **Line Plot** widgets as shown below.



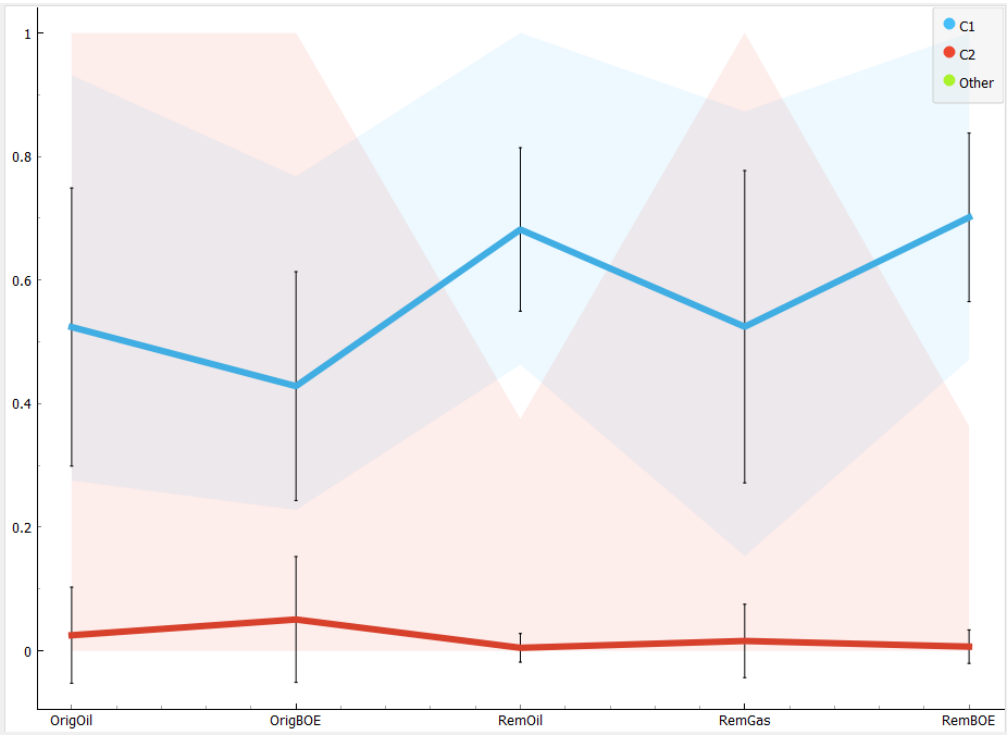
12. Do you see possibility of removing any features from the analysis to improve the cluster performance?

Yes, makes them nicely separated	FieldGOR, OrigGas, DiscYear

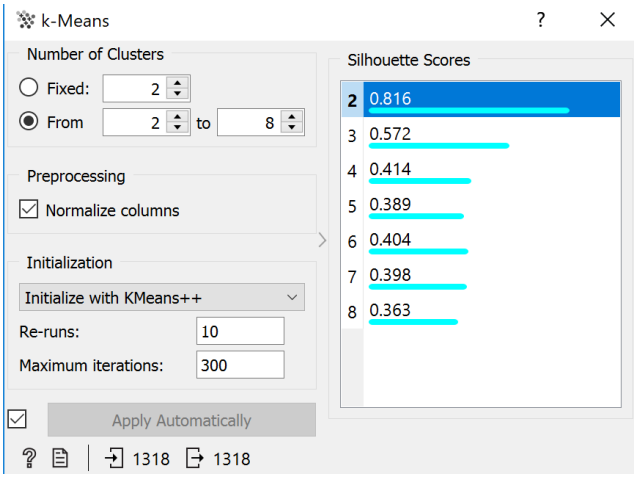
Before:



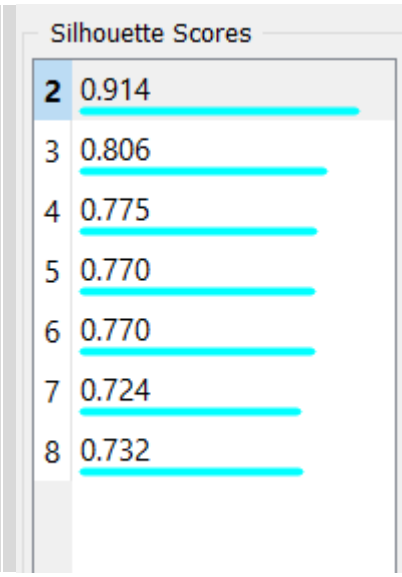
After:



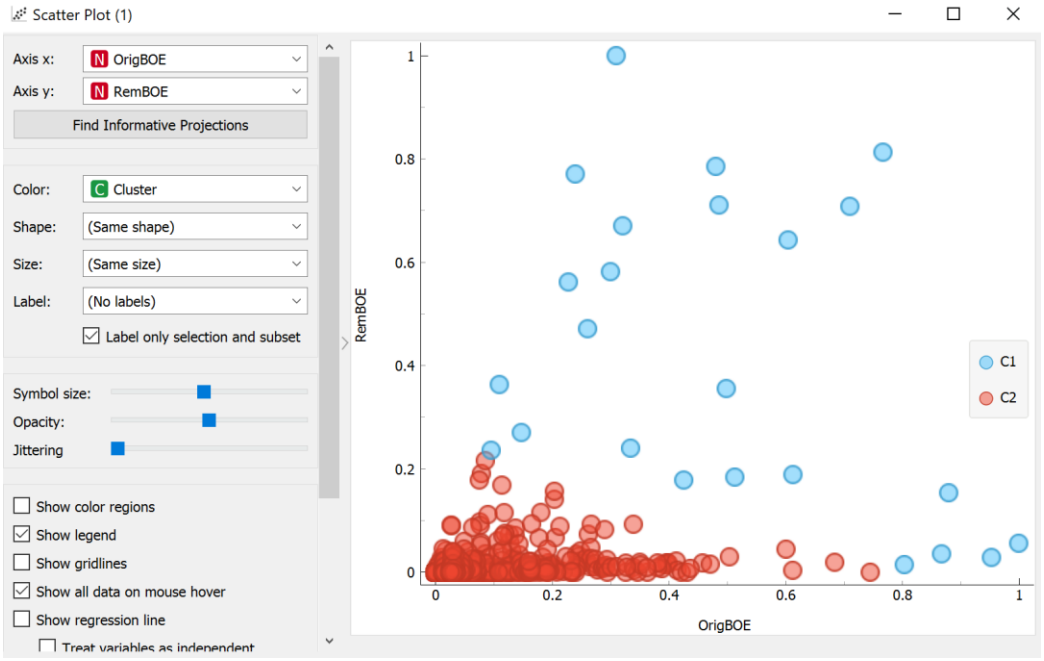
13. Add **k-Means** widget to the **Select Rows** widget in the pipeline.

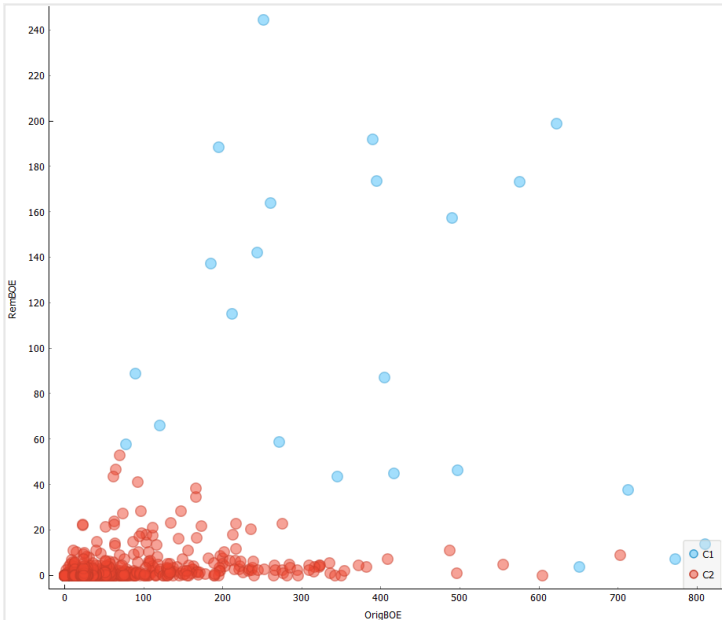


How many clusters are suggested based on Silhouette Scores?



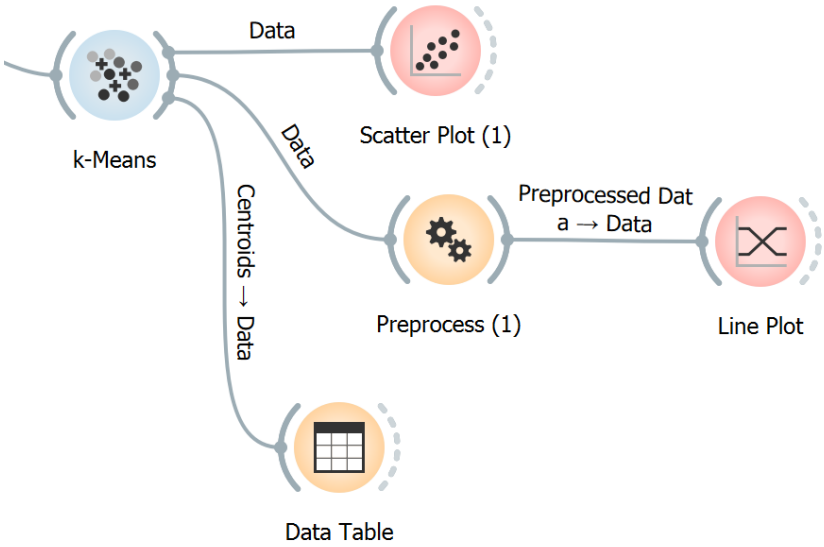
14. Add **Scatter Plot** widget to the **k-Means** widget. Confirm selections are as below:





What do you think about the clusters?	A lot of C2 is populated to the bottom left (smaller field indication), while we have more outliers with C1 (indicates bigger fields)

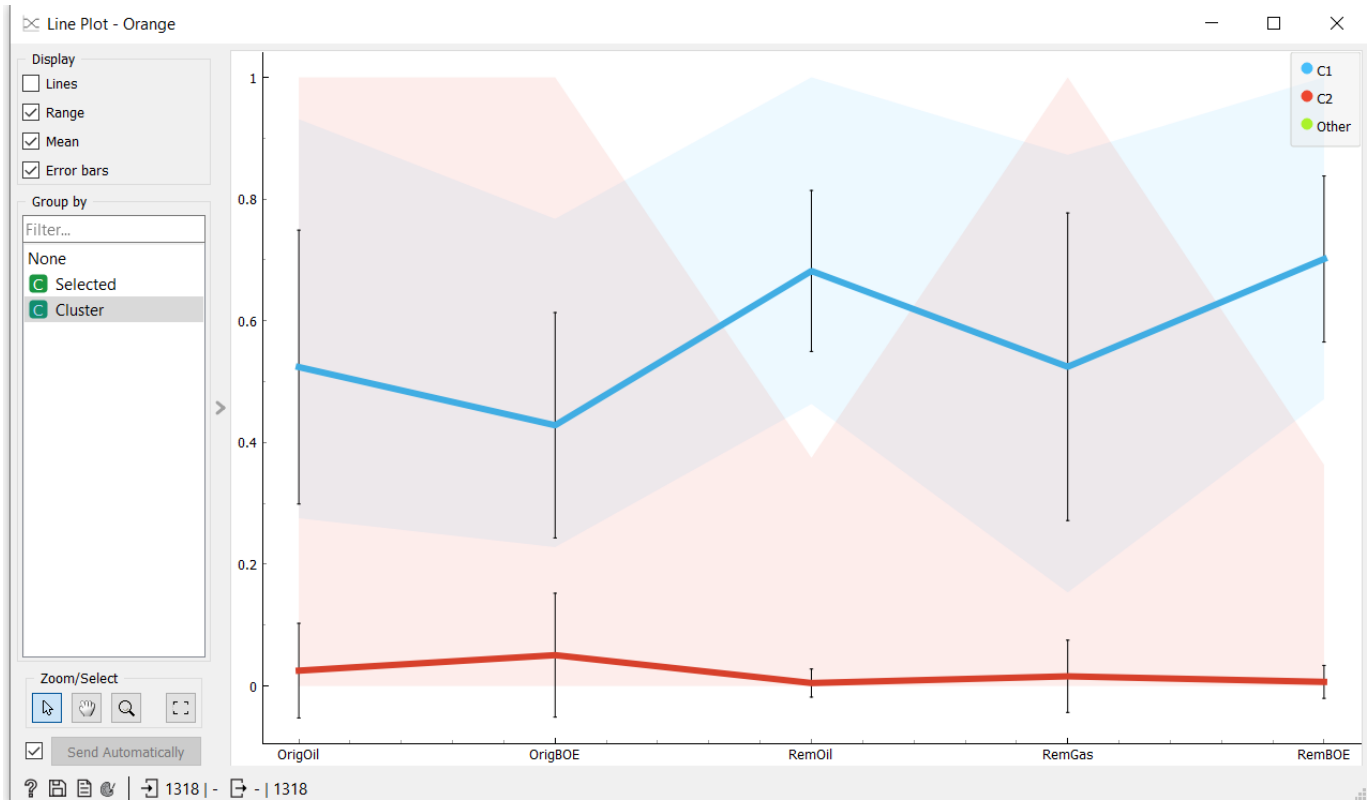
15. Add **Preprocess**, **Line Plot, and Data Table** widgets as shown below.



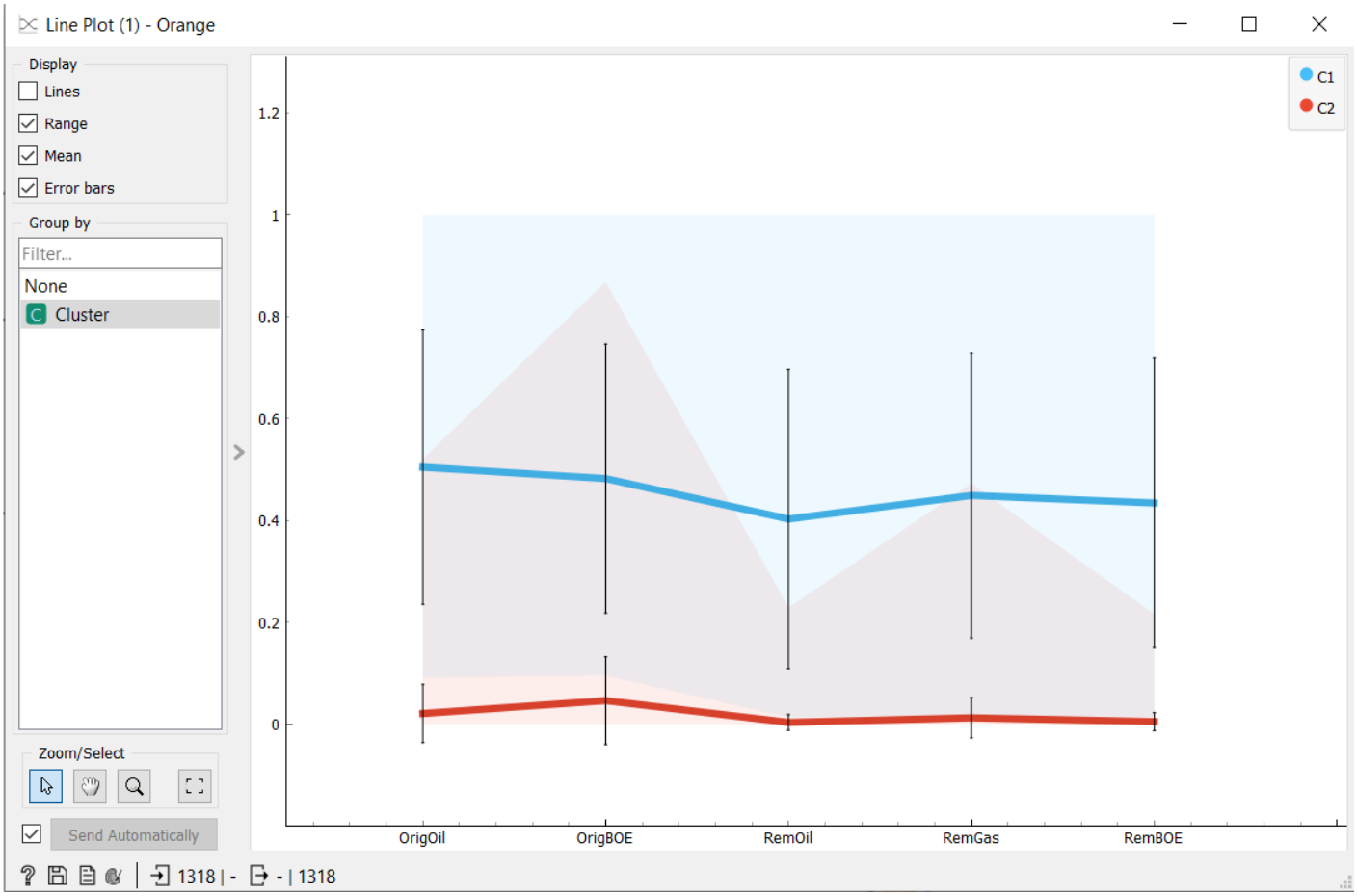
16. What can you observe from the line plot?

k-Means vs Hierarchical	The range is vastly different, k-means have less steeps and changes in line, error bars seem to expand more in k-Means

Hierarchical:



k-Means:


Line Plot (1) - Orange
Display
Lines
Range
Mean
Error bars
Group by
Filter...
None
Cluster
Zoom/Select
Send Automatically
1318 | -
- | 1318
C1
C2
1.2
1
0.8
0.6
0.4
0.2
0
OrigOil
OrigBOE
RemOil
RemGas
RemBOE