

UCM - Minería de Datos

CRM KNIME

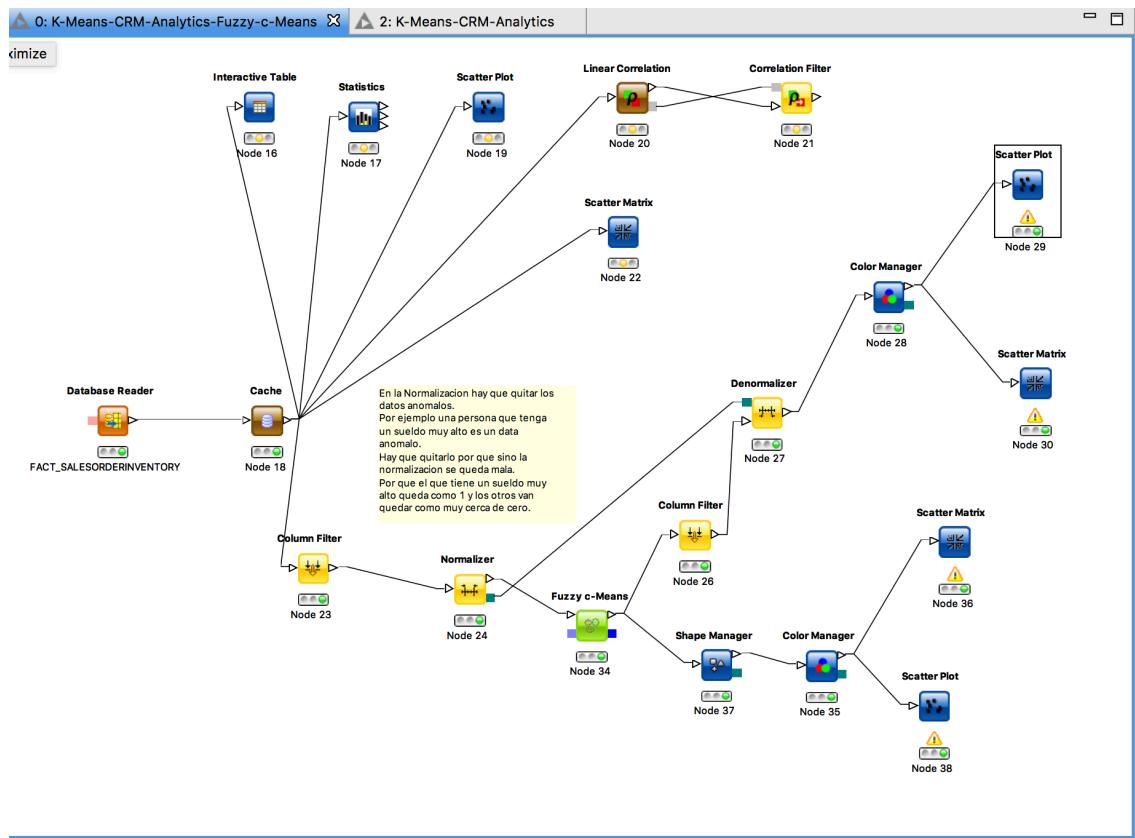
CRM

CAIO FERNANDES MORENO

1. Fuzzy c-Means con Knime

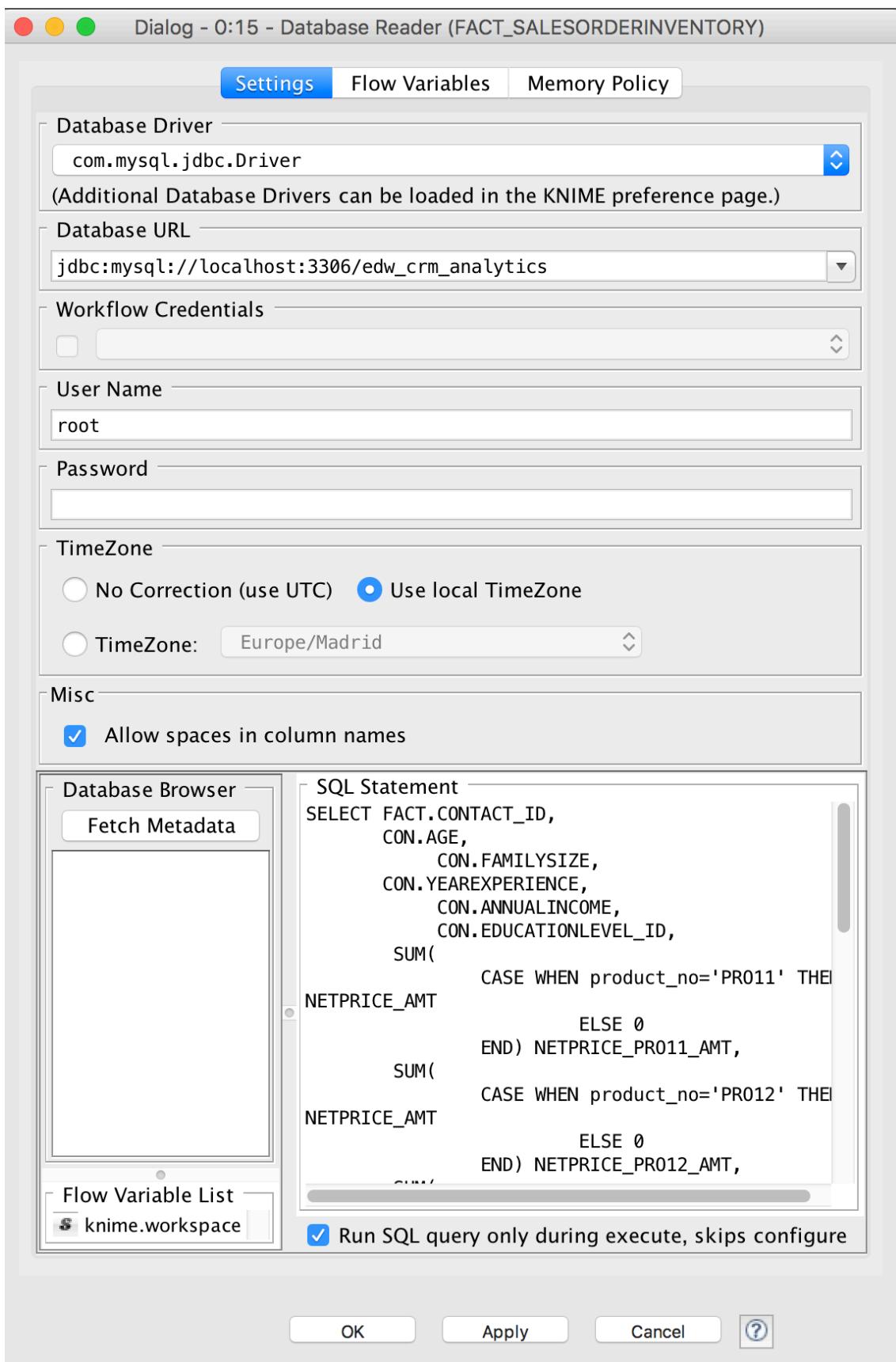
El objetivo del guión es explicar el proceso seguido para hacer una segmentación utilizando el algoritmo Fuzzy c -Means en la herramienta Knime.

La imagen a bajo es el resultado final de todo el proceso.



Los pasos seguidos son:

- 1) Leer los datos de la tabla de hecho utilizando el componente *Database Reader*. En la imagen abajo tenemos la configuración del componente *Database Reader*.



Consulta SQL utilizada:

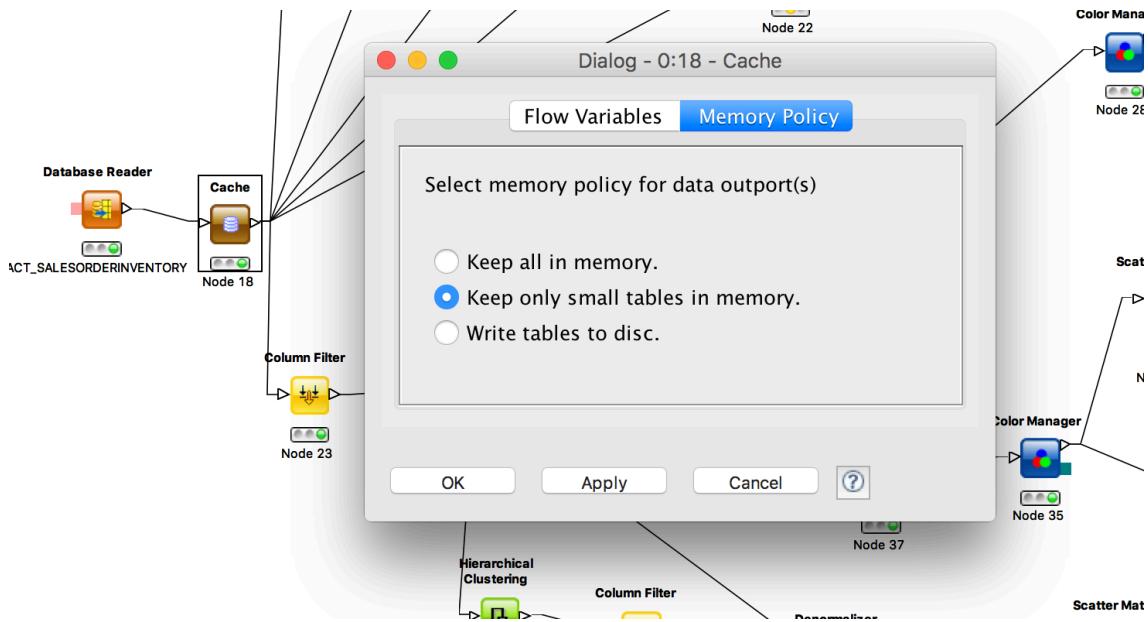
```
SELECT FACT.CONTACT_ID,  
      CON.AGE,  
      CON.FAMILYSIZE,  
      CON.YEAREXPERIENCE,  
      CON.ANUALINCOME,  
      CON.EDUCATIONLEVEL_ID,  
      SUM(  
          CASE WHEN product_no='PRO11' THEN NETPRICE_AMT  
                ELSE 0  
          END) NETPRICE_PRO11_AMT,  
      SUM(  
          CASE WHEN product_no='PRO12' THEN NETPRICE_AMT  
                ELSE 0  
          END) NETPRICE_PRO12_AMT,  
      SUM(  
          CASE WHEN product_no='PRO13' THEN NETPRICE_AMT  
                ELSE 0  
          END) NETPRICE_PRO13_AMT,  
      SUM(  
          CASE WHEN product_no='PRO14' THEN NETPRICE_AMT  
                ELSE 0  
          END) NETPRICE_PRO14_AMT,  
      SUM(  
          CASE WHEN product_no='PRO15' THEN NETPRICE_AMT  
                ELSE 0  
          END) NETPRICE_PRO15_AMT,  
      SUM(
```

```

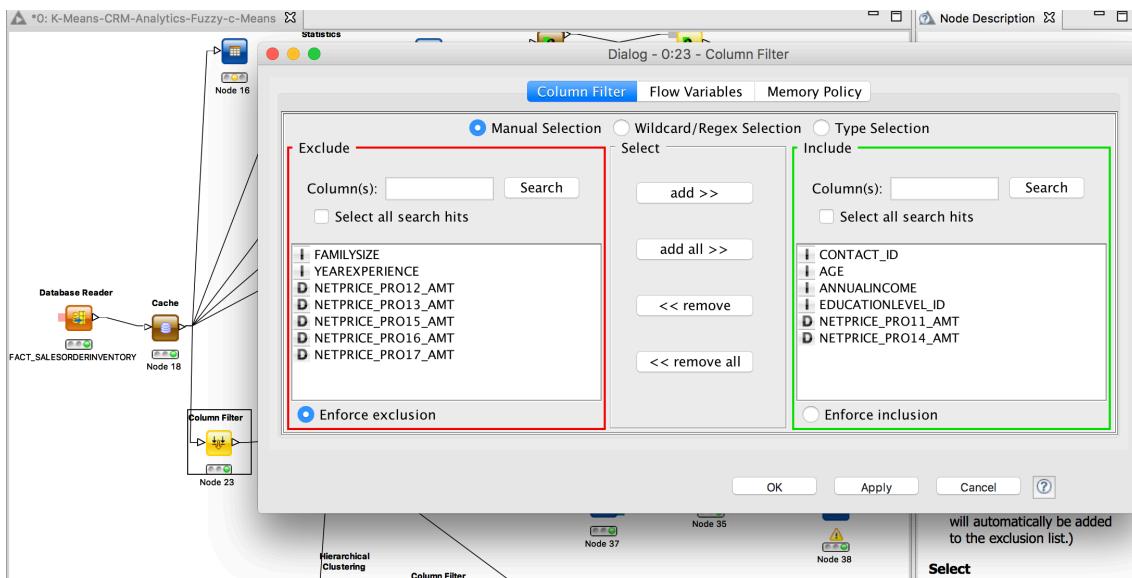
CASE WHEN product_no='PRO16' THEN NETPRICE_AMT
      ELSE 0
END) NETPRICE_PRO16_AMT,
SUM(
CASE WHEN product_no='PRO17' THEN NETPRICE_AMT
      ELSE 0
END) NETPRICE_PRO17_AMT
FROM FACT_SALESORDERINVENTORY FACT
JOIN DIM_CONTACTS CON ON CON.CONTACT_ID=FACT.CONTACT_ID
JOIN DIM_PRODUCTS PRO ON PRO.PRODUCT_ID=FACT.PRODUCT_ID
GROUP BY 1,2,3,4,5,6

```

2) Utilizamos el componente Cache para hacer el cache de los datos.



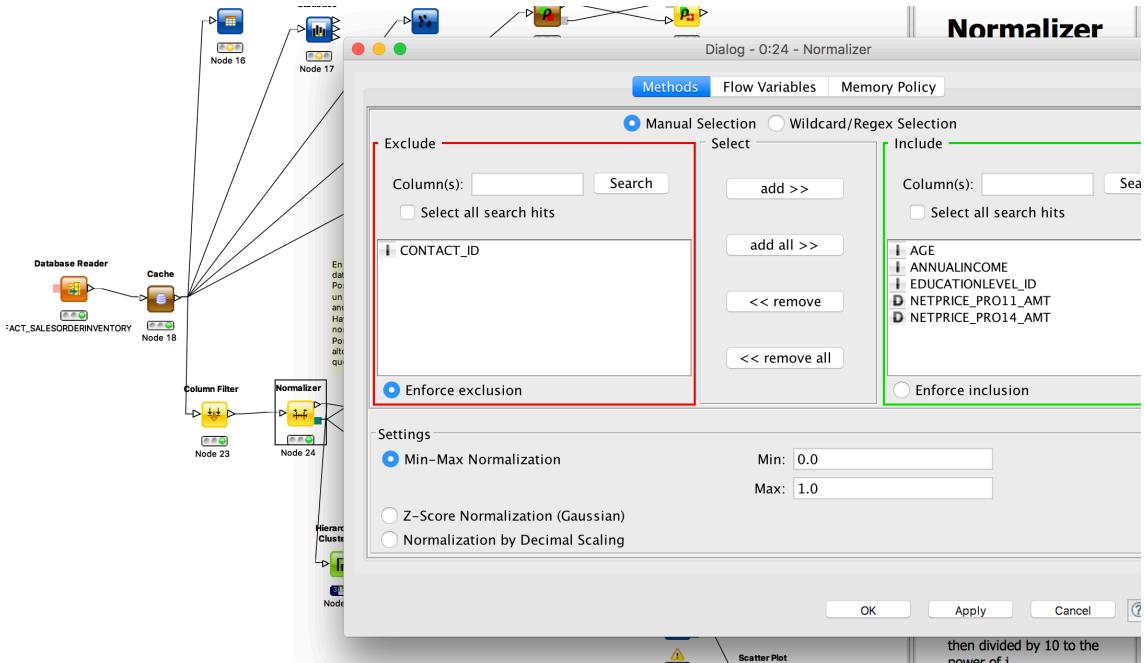
3) Utilizamos el Column Filter para traer apenas las columnas necesarias para hacer el K-means.



Quedamos apenas con las columnas CONTACT_ID, AGE, ANNUALINCOME, EDUCATIONLEVEL_ID, NETPRICE_PRO11_AMT, NETPRICE_PRO14_AMT.

Las columnas en rojo en la imagen arriba no vamos utilizar.

4) Normalizar los datos.



Para normalizar los datos quedamos apenas con la columna AGE, ANNUALINCOME, EDUCATIONLEVEL_ID, NETPRICE_PRO11_AMT, NETPRICE_PRO14_AMT y borramos el flujo de datos la columna CONTACT_ID.

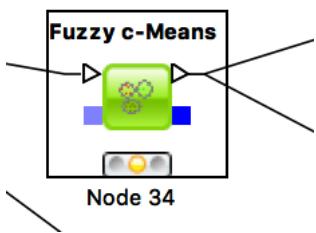
El proceso de normalizar es esencial para utilizar el algoritmo *Fuzzy c-Means* por que transforma los valores para una escala normalizada.

Se utiliza la opción *Min-Max Normalization* donde el Min es 0.0 y el máximo es 1.0

Importante:

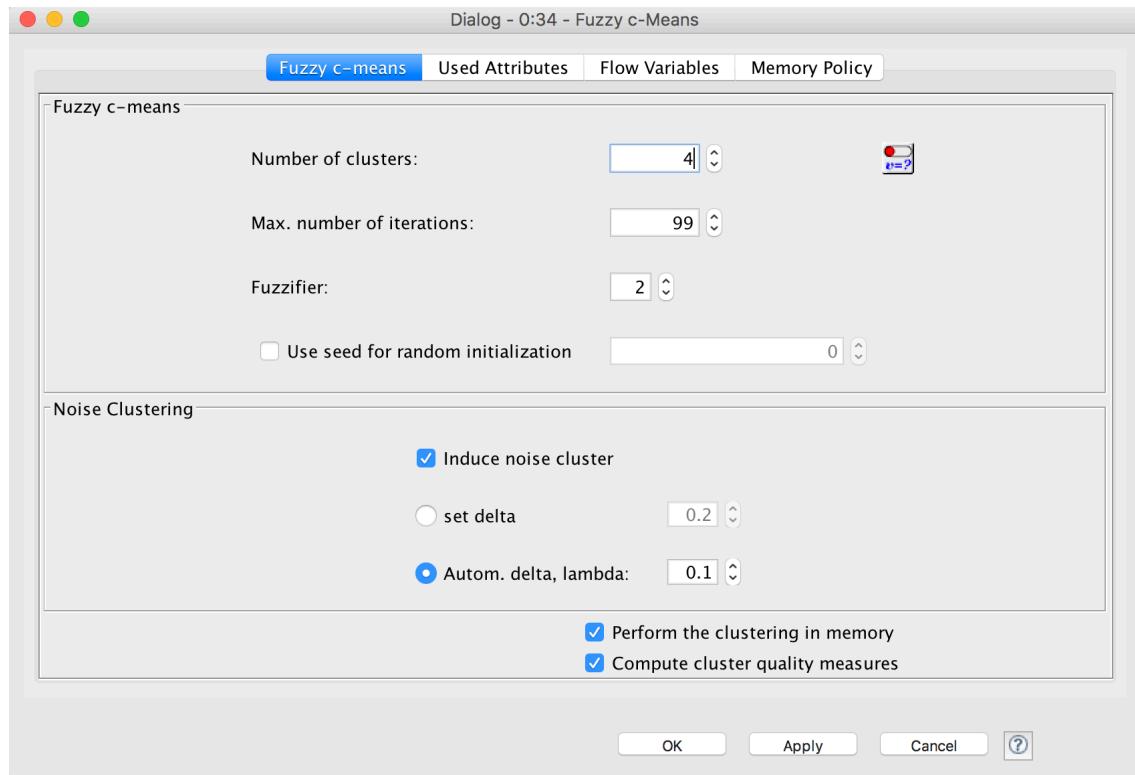
En la Normalización hay que quitar los datos anomalos. Por ejemplo una persona que tenga un sueldo muy alto es un dato anormal. Hay que quitarlo por que sino la normalización se queda mala. Por que el que tiene un sueldo muy alto queda como 1 y los otros van quedar como muy cerca de cero.

5) Fuzzy c-Means

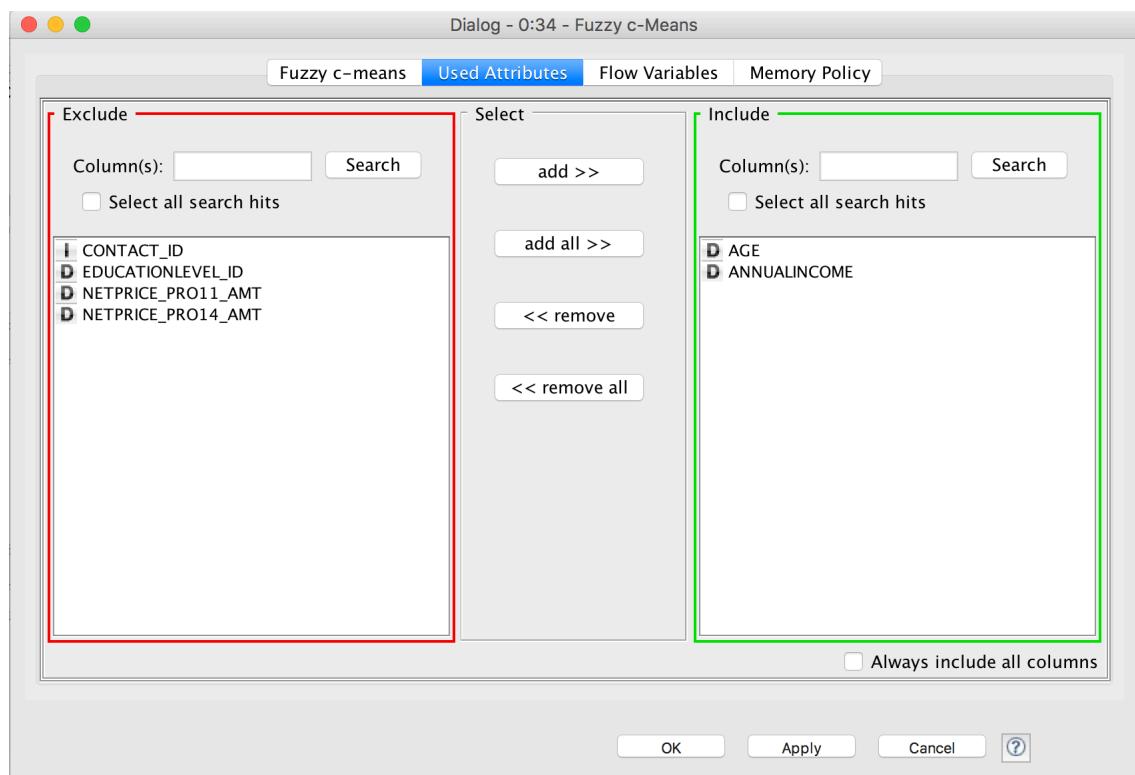


Para utilizar el Fuzzy c-Means hay que configurar el numero de clúster que hemos decidido dejar 4, hemos optado tener un clúster mas pequeño apenas de 4, pero se puede explicar las variables con mas clústeres.

La ventana abajo es donde se configura el numero de clústeres, el numero máximo de interacciones que hemos dejado 99 y otras variables como fuzzifier, noise clustering, etc.



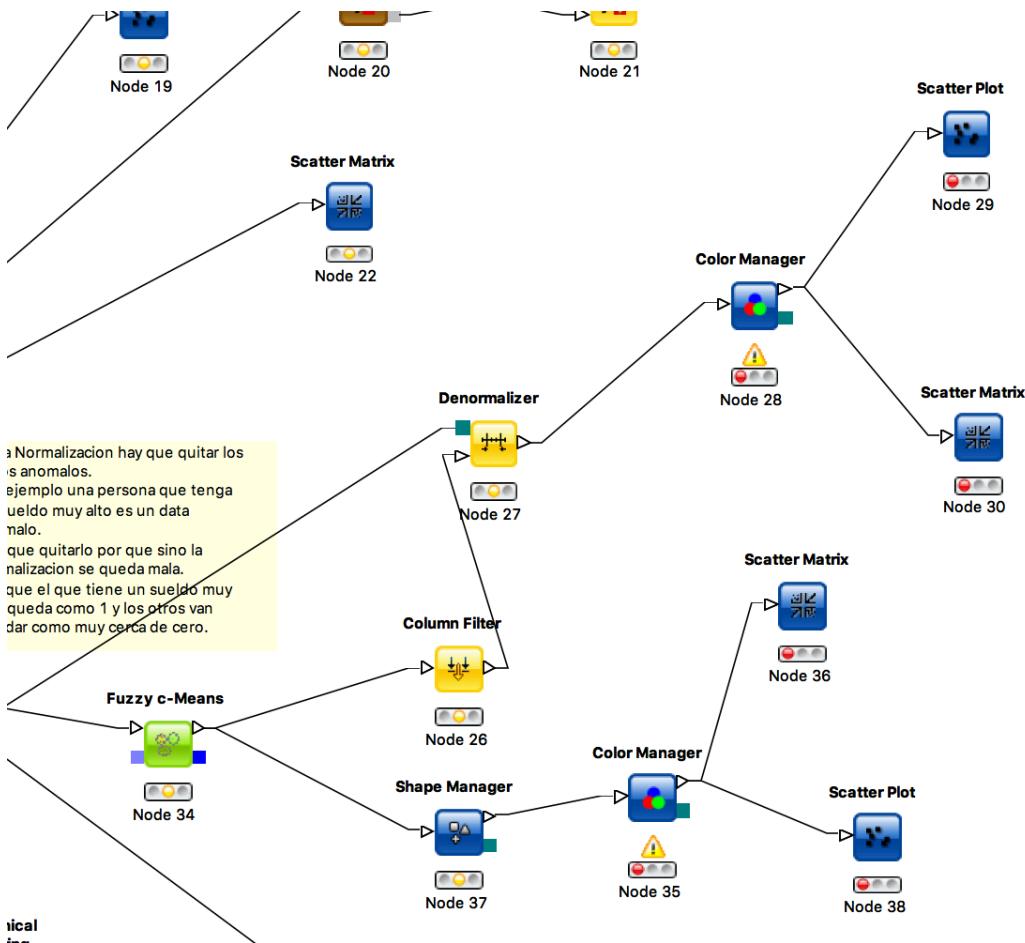
Hay que dejar solo las columnas AGE y ANNUALINCOME.



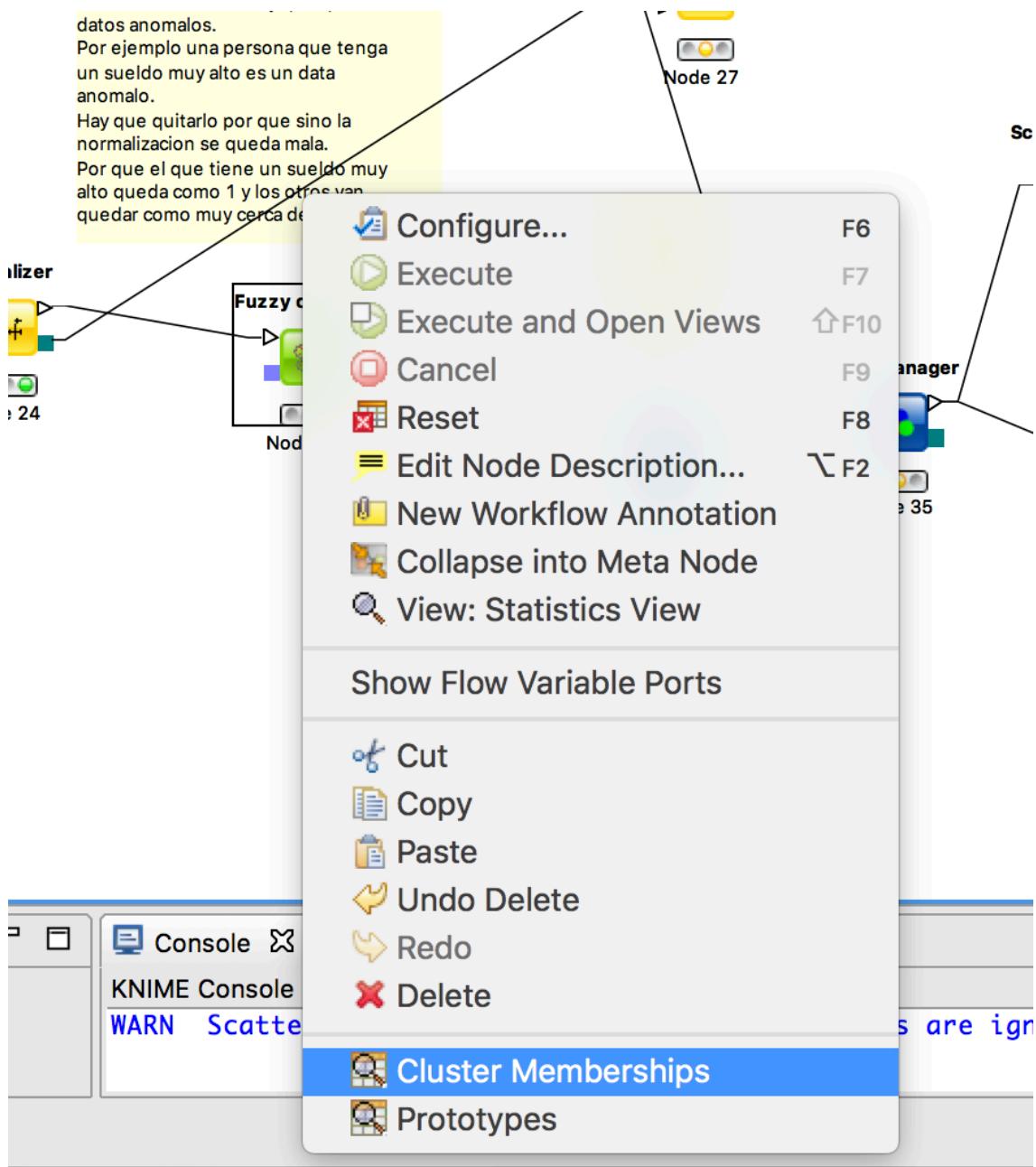
6) Visualización de los datos con Scatter Plot o Scatter Matrix

Hecho el Fuzzy c-Means los próximos pasos son visualizar los datos utilizando en este caso Scatter Plot o Scatter Matrix.

Se puede ver que hay dos flujos para la visualización: uno utilizando el componente Denormalizer y otro sin utilizar.



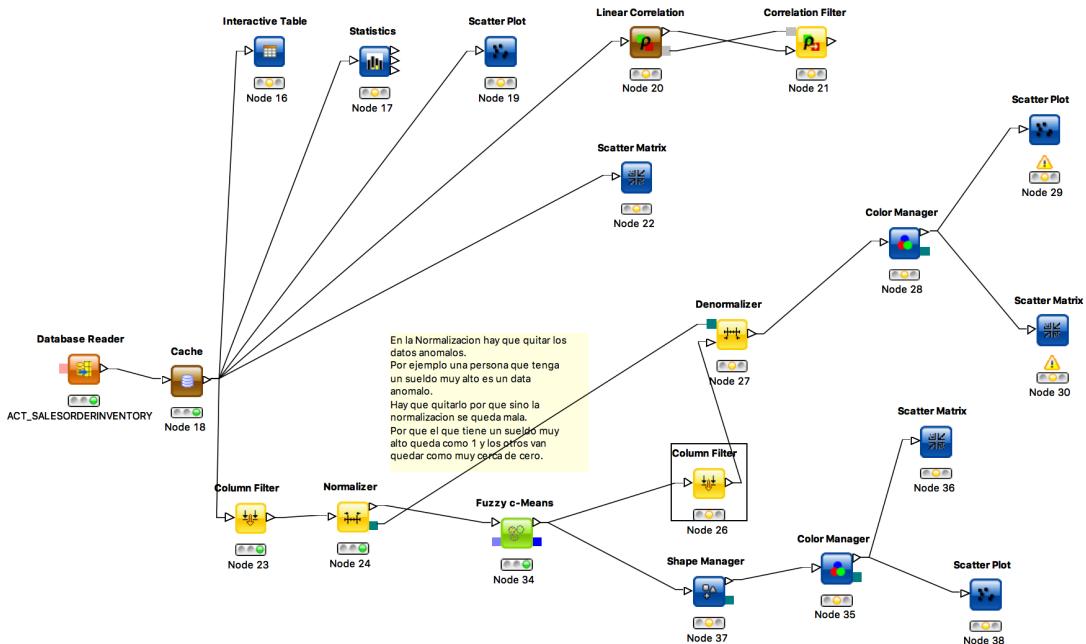
La opción *Cluster Memberships* cuando seleccionada nos enseña como se ha quedado cada clúster y también en Noise Cluster.



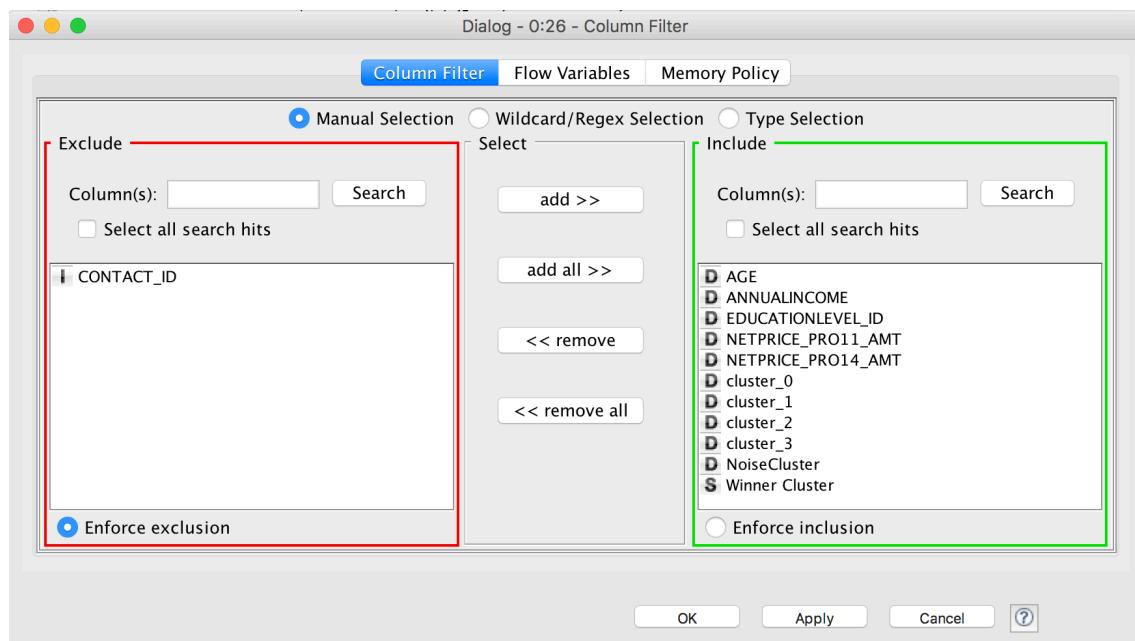
Se puede ver en la tabla Cluster Memberships de Fuzzy c-Means que la Row0 pertenece al cluster 0, pero hay ruido.

Cluster Memberships - 0:34 - Fuzzy c-Means											
Row ID	CONT...	AGE	ANNU...	EDUC...	NETP...	NETP...	cluste...	cluste...	cluste...	Noise...	Winne...
Row0	395	0.045	0.19	0	0	0.16	1	0	0	NoiseCluster	cluster_0
Row1	396	0.5	0.12	0	0	0.15	0	1	0	NoiseCluster	cluster_0
Row2	397	0.364	0.014	0	0	0.1	1	0	0	NoiseCluster	cluster_0
Row3	398	0.273	0.426	0.5	0	0.27	0	1	0	NoiseCluster	cluster_0
Row4	399	0.273	0.171	0.5	0	0.1	0	1	0	NoiseCluster	cluster_0
Row5	400	0.318	0.097	0.5	0.244	0.04	1	0	0	NoiseCluster	cluster_0
Row6	401	0.682	0.296	0.5	0	0.15	0	1	0	NoiseCluster	cluster_3
Row7	402	0.614	0.065	1	0	0.03	0	1	0	NoiseCluster	cluster_0
Row8	403	0.273	0.338	0.5	0.164	0.06	1	0	0	NoiseCluster	cluster_0
Row9	404	0.25	0.796	1	0	0.89	0	1	0	NoiseCluster	cluster_0
Row10	405	0.955	0.449	1	0	0.24	0	1	0	NoiseCluster	cluster_0
Row11	406	0.136	0.171	0.5	0	0.01	0	1	0	NoiseCluster	cluster_0
Row12	407	0.568	0.491	1	0	0.38	0	1	0	NoiseCluster	cluster_0
Row13	408	0.818	0.148	0.5	0	0.25	0	1	0	NoiseCluster	cluster_1
Row14	409	1	0.481	0	0	0.2	0	1	0	NoiseCluster	cluster_1
Row15	410	0.841	0.065	1	0	0.15	0	1	0	NoiseCluster	cluster_1
Row16	411	0.341	0.565	1	0.211	0.47	0	1	0	NoiseCluster	cluster_2
Row17	412	0.432	0.338	0	0	0.24	0	1	0	NoiseCluster	cluster_1
Row18	413	0.523	0.856	1	0	0.81	0	1	0	NoiseCluster	cluster_1
Row19	414	0.727	0.06	0.5	0	0.05	0	1	0	NoiseCluster	cluster_1
Row20	415	0.75	0.079	0.5	0.175	0.09	0	1	0	NoiseCluster	cluster_1
Row21	416	0.773	0.255	1	0	0.2	0	1	0	NoiseCluster	cluster_0
Row22	417	0.136	0.25	0	0.409	0.12	0	1	0	NoiseCluster	cluster_0
Row23	418	0.477	0.162	0	0.257	0.07	0	1	0	NoiseCluster	cluster_2
Row24	419	0.295	0.667	0	0.25	0.39	0	1	0	NoiseCluster	cluster_2
Row25	420	0.455	0.097	0	0.153	0.05	0	1	0	NoiseCluster	cluster_1
Row26	421	0.386	0.347	1	0	0.02	0	1	0	NoiseCluster	cluster_1
Row27	422	0.523	0.694	0	0	0.24	0	1	0	NoiseCluster	cluster_1
Row28	423	0.75	0.185	1	0	0.22	0	1	0	NoiseCluster	cluster_1
Row29	424	0.341	0.514	0.5	0	0.33	0	1	0	NoiseCluster	cluster_1
Row30	425	0.818	0.125	1	0.192	0.12	0	1	0	NoiseCluster	cluster_1
Row31	426	0.386	0.097	0.5	0	0.2	0	1	0	NoiseCluster	cluster_0
Row32	427	0.682	0.153	1	0.304	0.06	0	1	0	NoiseCluster	cluster_0
Row33	428	0.159	0.046	1	0	0.09	0	1	0	NoiseCluster	cluster_0
Row34	429	0.182	0.194	1	0	0.18	0	1	0	NoiseCluster	cluster_0
Row35	430	0.568	0.338	0	0	0.07	0	1	0	NoiseCluster	cluster_3
Row36	431	0.818	0.523	0	0	0.29	0	1	0	NoiseCluster	cluster_3
Row37	432	0.636	0.292	1	0.312	0.14	0	1	0	NoiseCluster	cluster_3

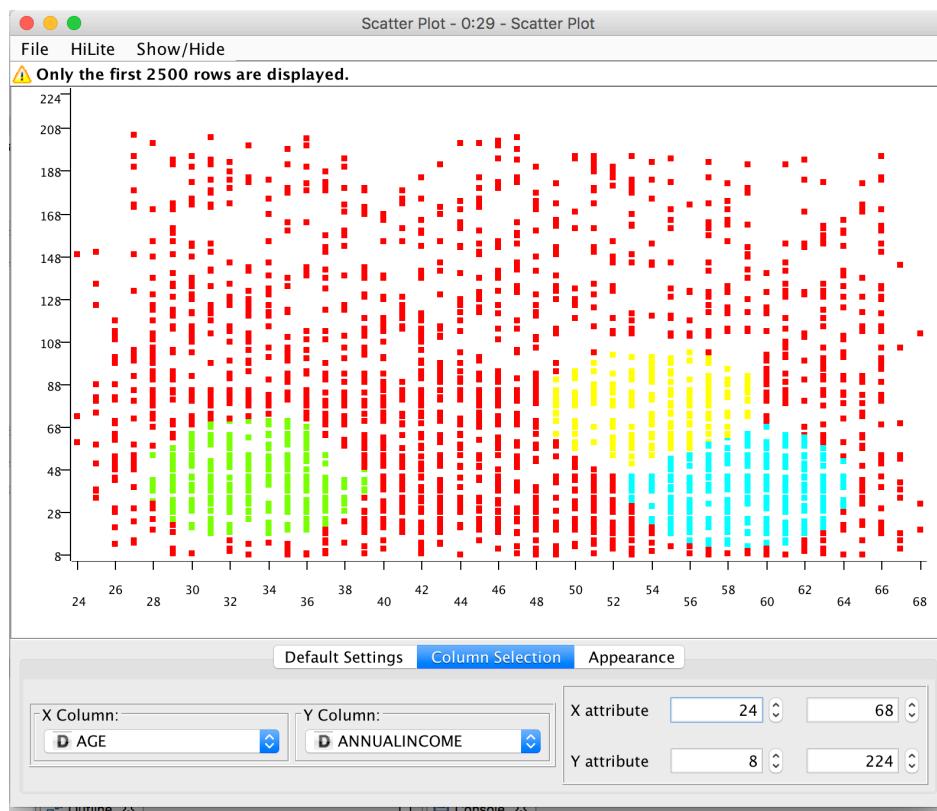
En la imagen abajo se puede ver muy bien todo el flujo de datos otra vez.



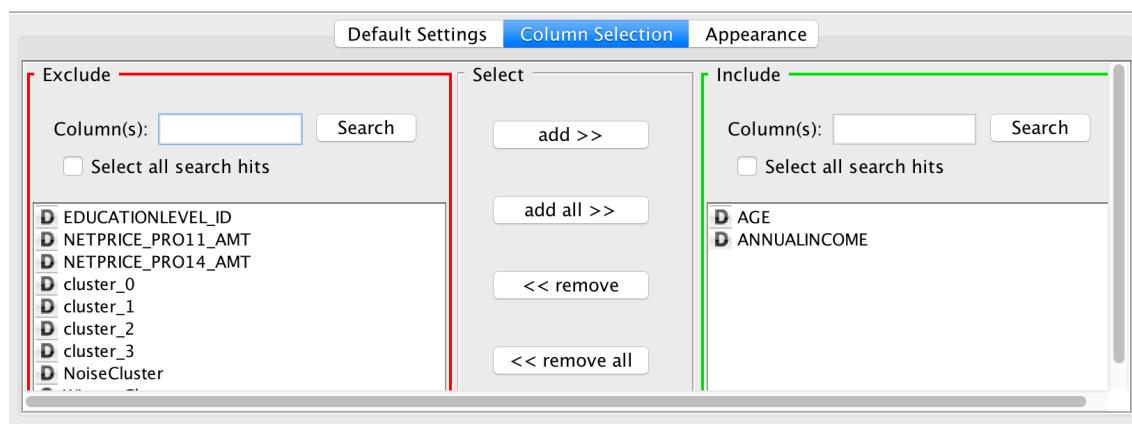
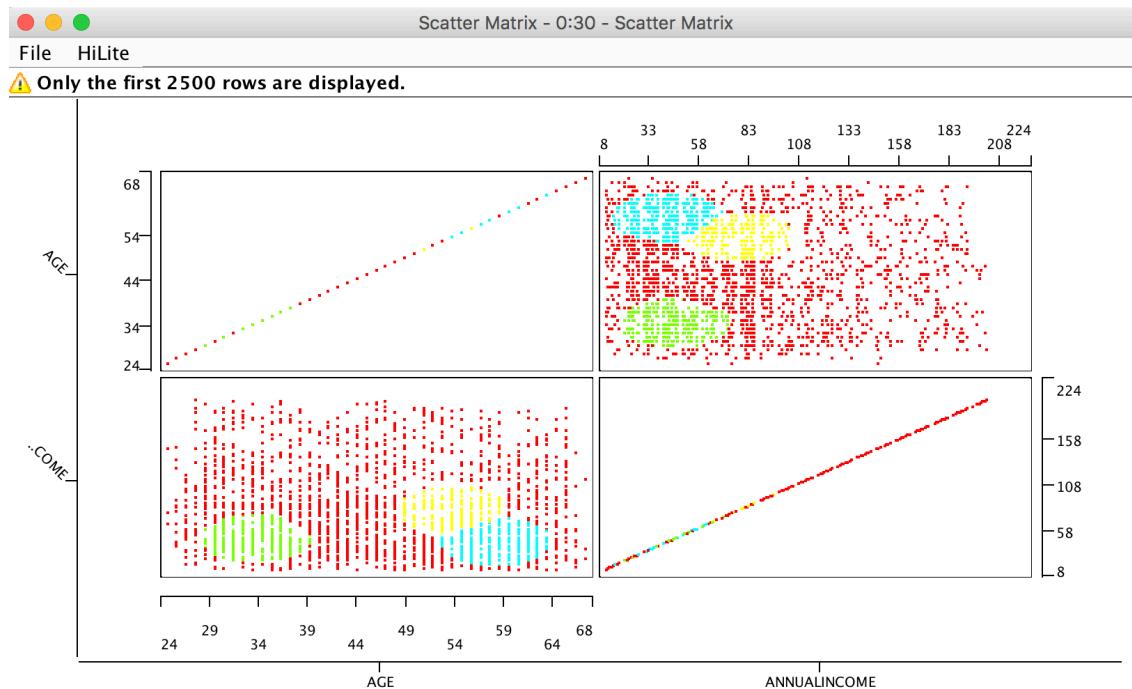
En el *Column Filter* se deja y quita las columnas como se puede ver la imagen abajo.



En el *Scatter Plot* se puede ver los clústeres por X = AGE (Edad) e Y = ANNUALINCOME (sueldo anual). Los clústeres en los colores Verde, Rojo, Amarillo y Azul.

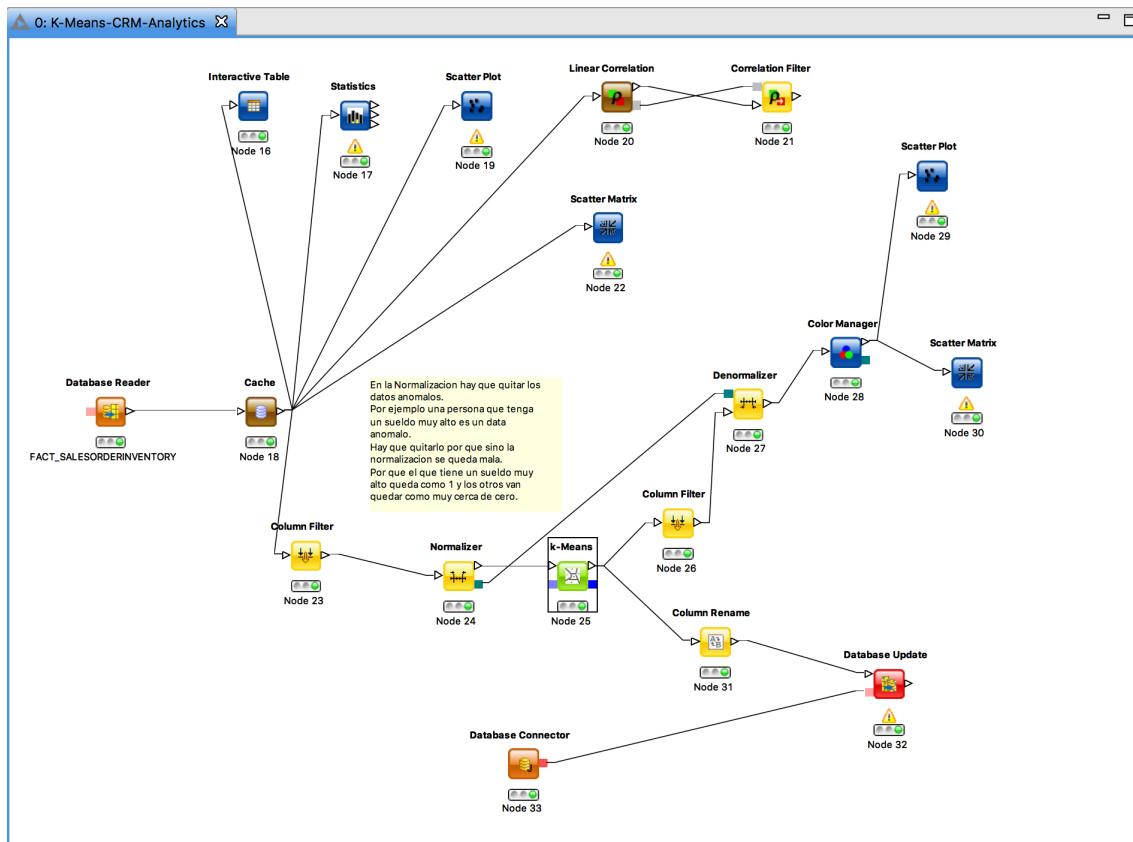


El *Scatter Matrix* se puede ver los mismo datos de una otra forma.

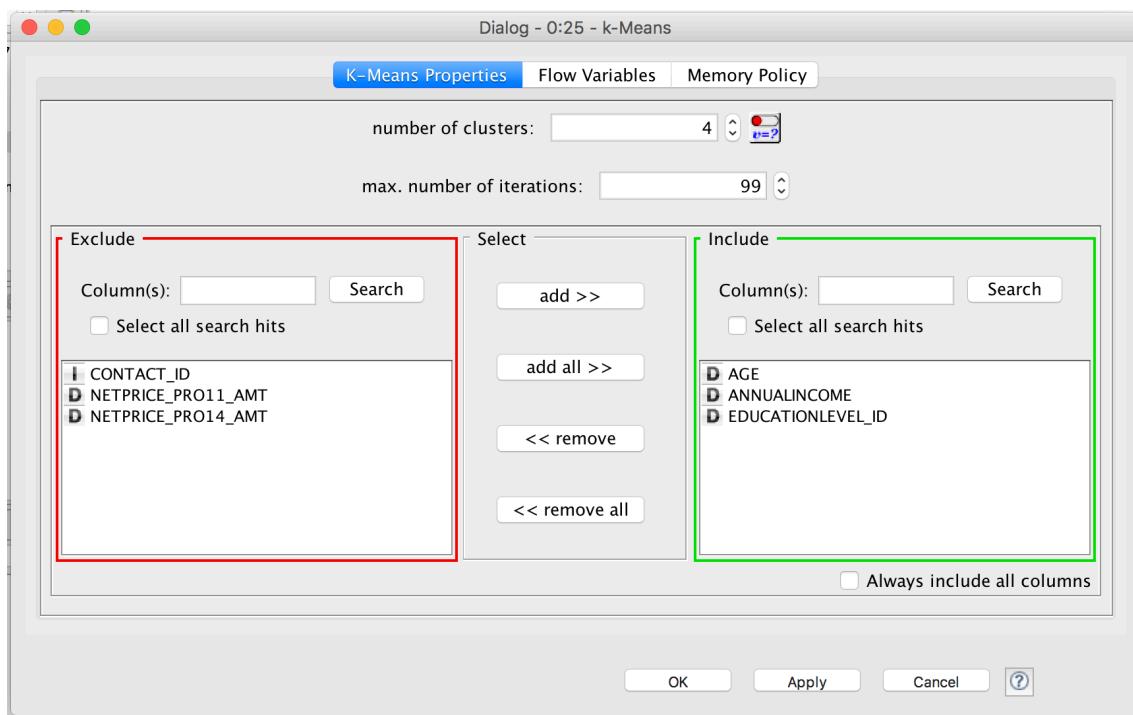


7) K-Means y Visualización de los datos con Scatter Plot o Scatter Matrix

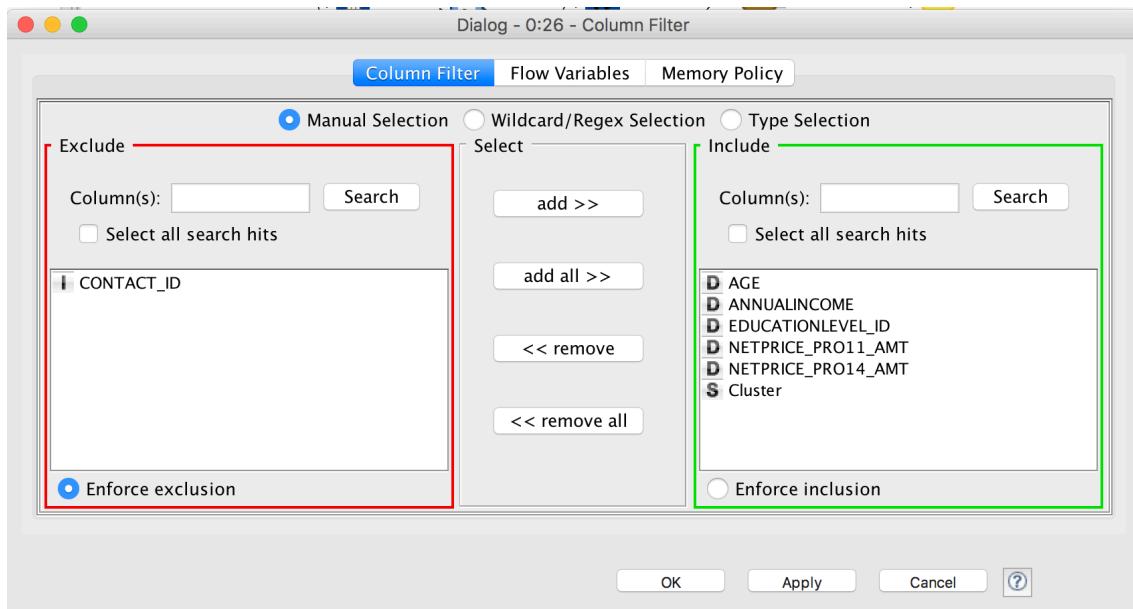
Lo mismo que hemos hecho con Fuzzy c-Means también hemos hecho con K-Means.



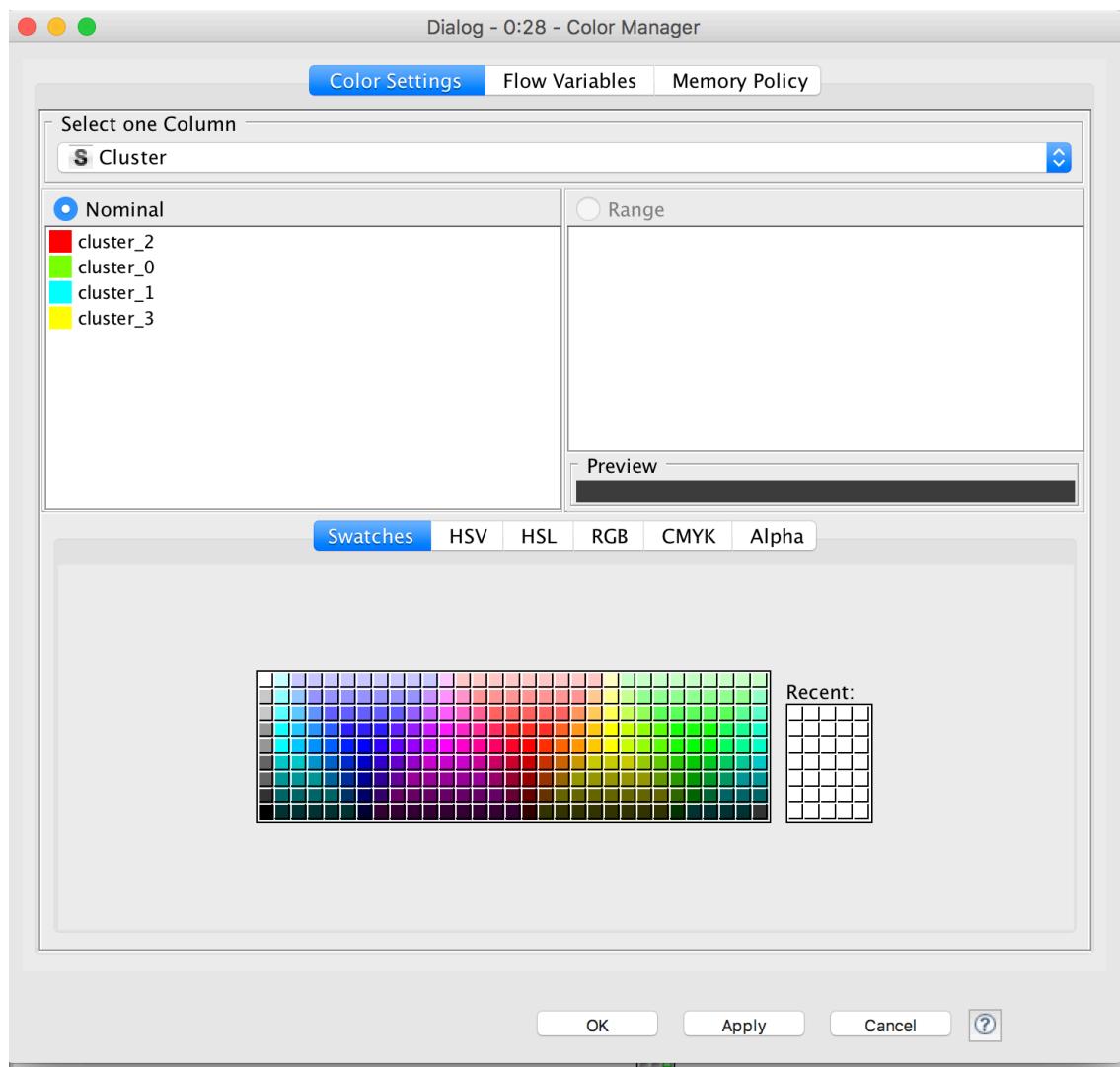
Configuración de K-Means.



Column Filter



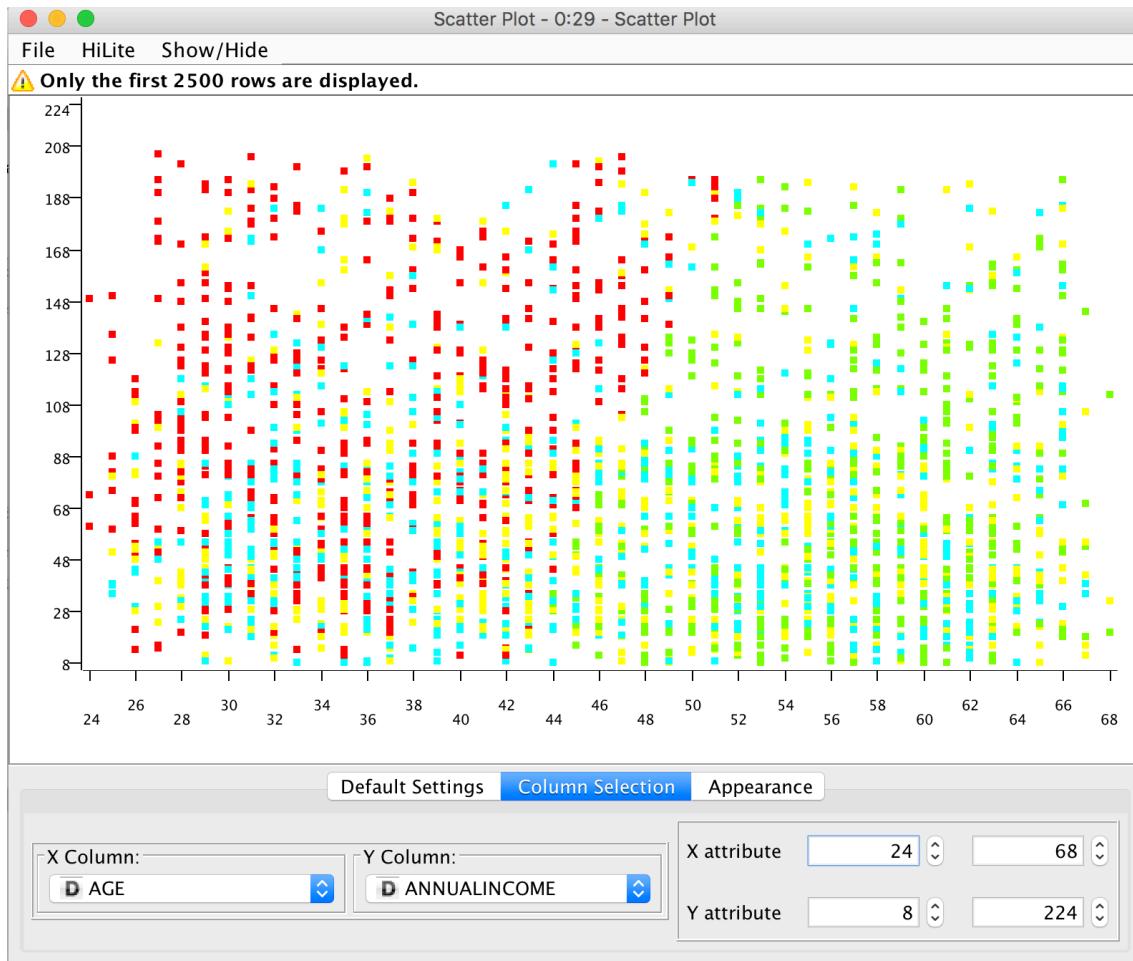
Color Manager



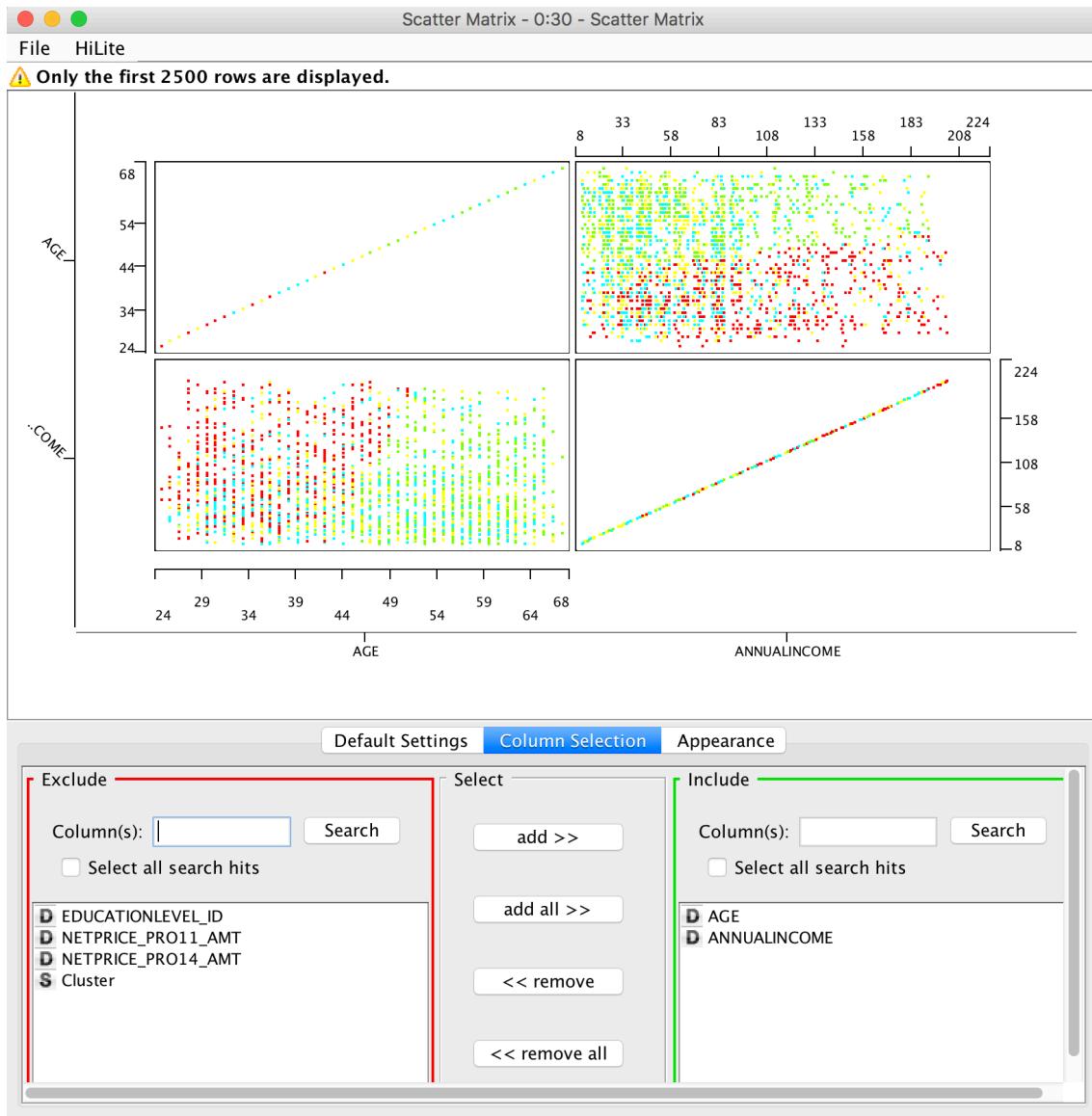
Scatter Plot de K-means donde X = AGE e Y = ANNUALINCOME.

Se puede ver que los rojos están más concentrados en las personas con edades de 24 a 50 años, siendo su renta anual de 8 a 208.

Las personas con edad de 50 a 68 años son parte del clúster verde, azul y amarillo.



Se puede analizar los mismos datos con la opción Scatter Matrix.



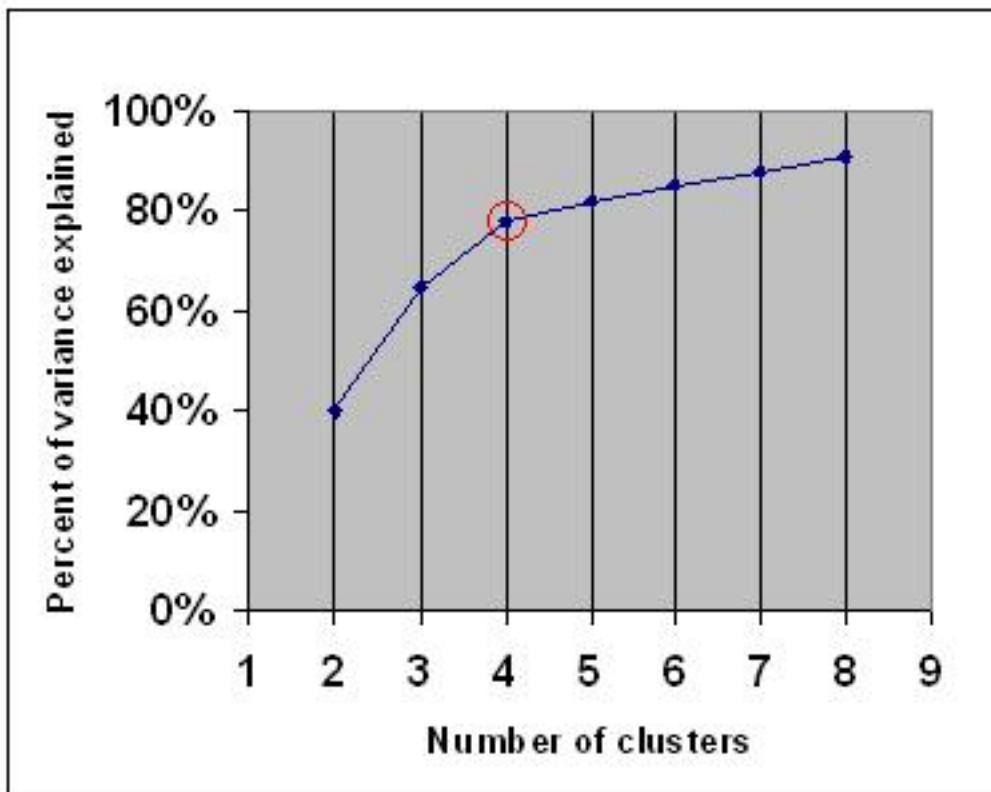
2. CONCLUSIONES

La herramienta KNIME es muy potente para trabajar con K-means y Fuzzy c-Means.

La tarea de hacer una buena segmentación es muy compleja siendo necesario saber cual son las variables que hay que utilizar en el algoritmo de segmentación (K-Means, Fuzzy c-Means), hacer una buena limpieza de los datos para quitar outliers, saber cuantos clústeres hay que escoger y después saber interpretar los datos.

Para la tarea de determinar con cuantos clústeres hay que quedarse recomiendo utilizar el método de Elbow.

La imagen abajo enseña un ejemplo de un grafico de Elbow, donde se puede ver que 4 clústeres explican 80% del porcentual de varianza explicada. Yo no he encontrado como se puede ver el grafico de elbow en KNIME, en la herramienta R se puede hacer esto mediante código.



Lo ideal es probar con 10 clústeres y mirar el grafico de Elbow y ver con cuantos clústeres quedamos.

Esto se puede hacer de forma mas sencilla con la herramienta R.

Se puede saber mas en el enlace:
https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set

Ejemplo con Python:

<https://datascienceLab.wordpress.com/2013/12/27/finding-the-k-in-k-means-clustering/>