UCM - Mineria de Datos

CRM KNIME

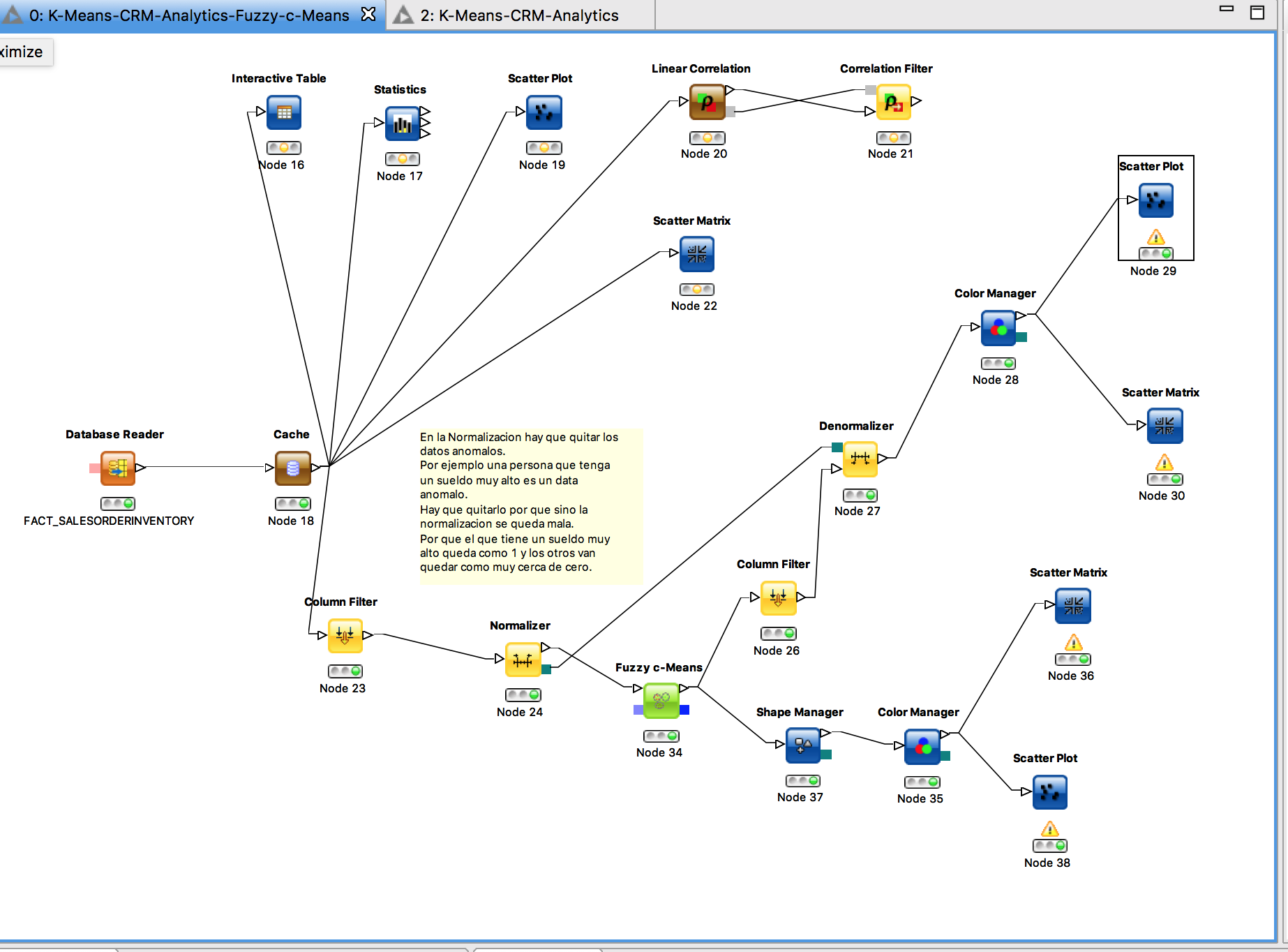
CRM

CAIO FERNANDES MORENO

# 1. Fuzzy c-Means con Knime

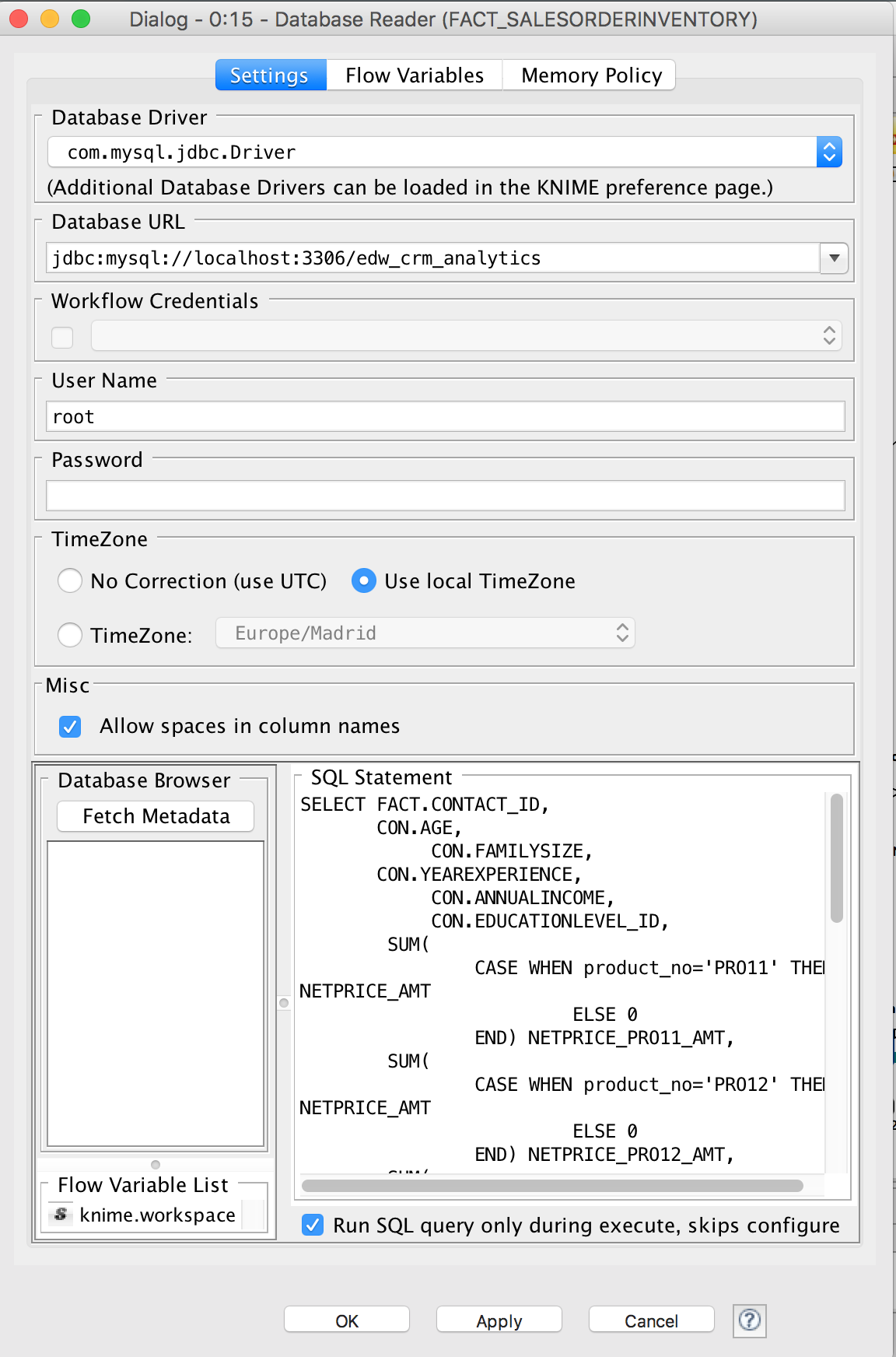
El objetivo del guión es explicar el proceso seguido para hacer una segmentación utilizando el algoritmo Fuzzy *c-Means* en la herramienta *Knime*.

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.ANNUALINCOME,

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

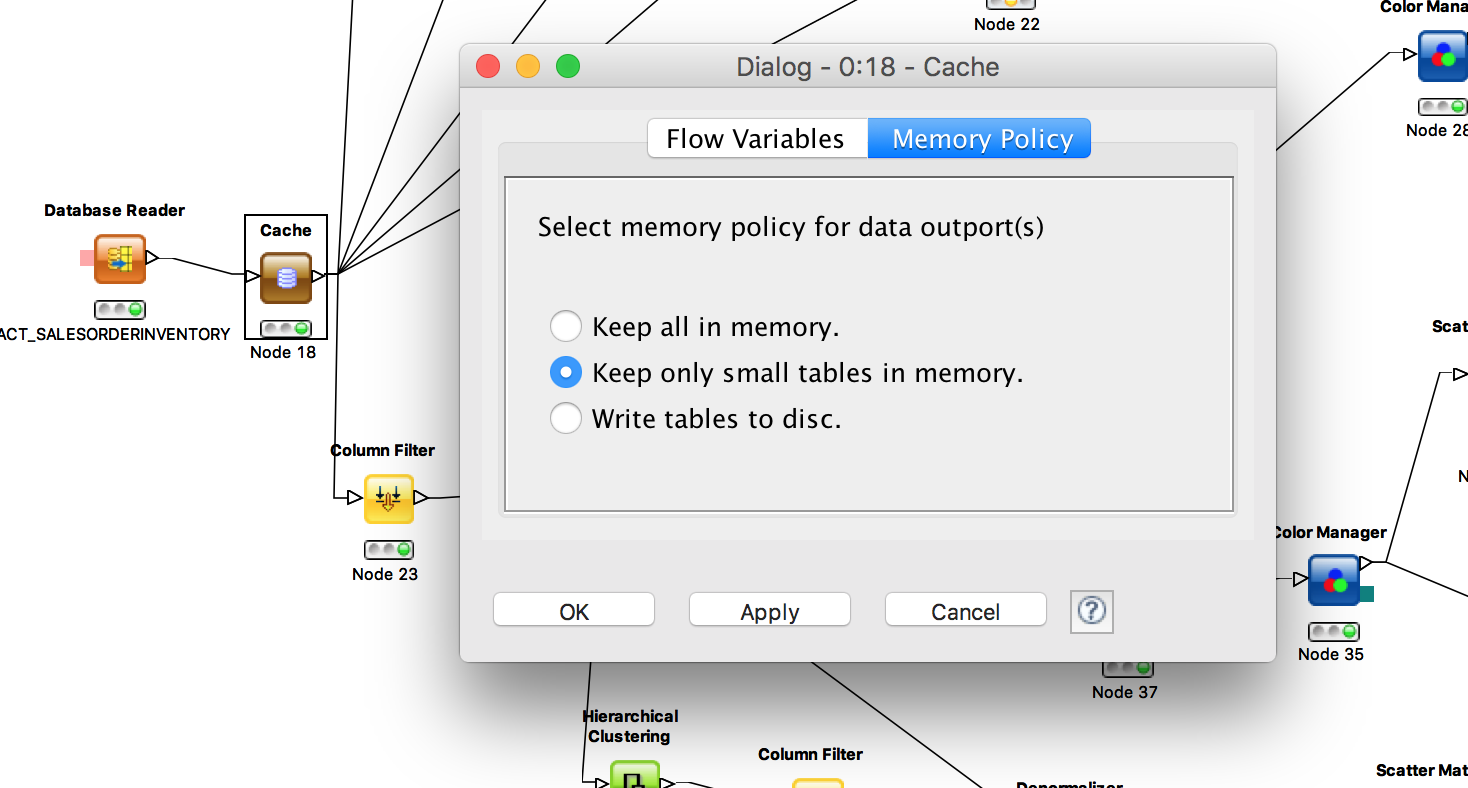
FROM FACT\_SALESORDERINVENTORY FACT

JOIN DIM\_CONTACTS CON ON CON.CONTACT\_ID=FACT.CONTACT\_ID

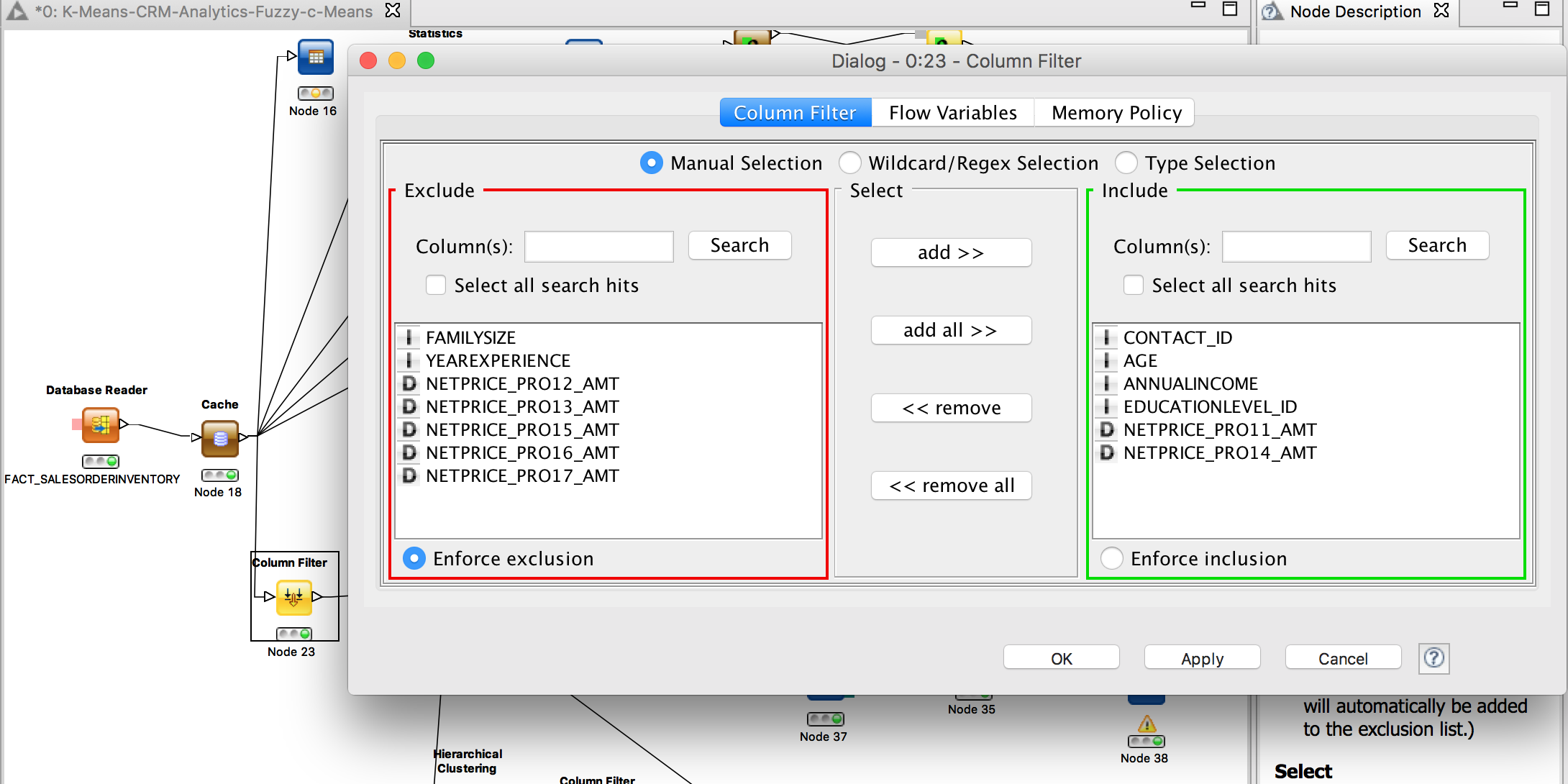
JOIN DIM\_PRODUCTS PRO ON PRO.PRODUCT\_ID=FACT.PRODUCT\_ID

GROUP BY 1,2,3,4,5,6

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



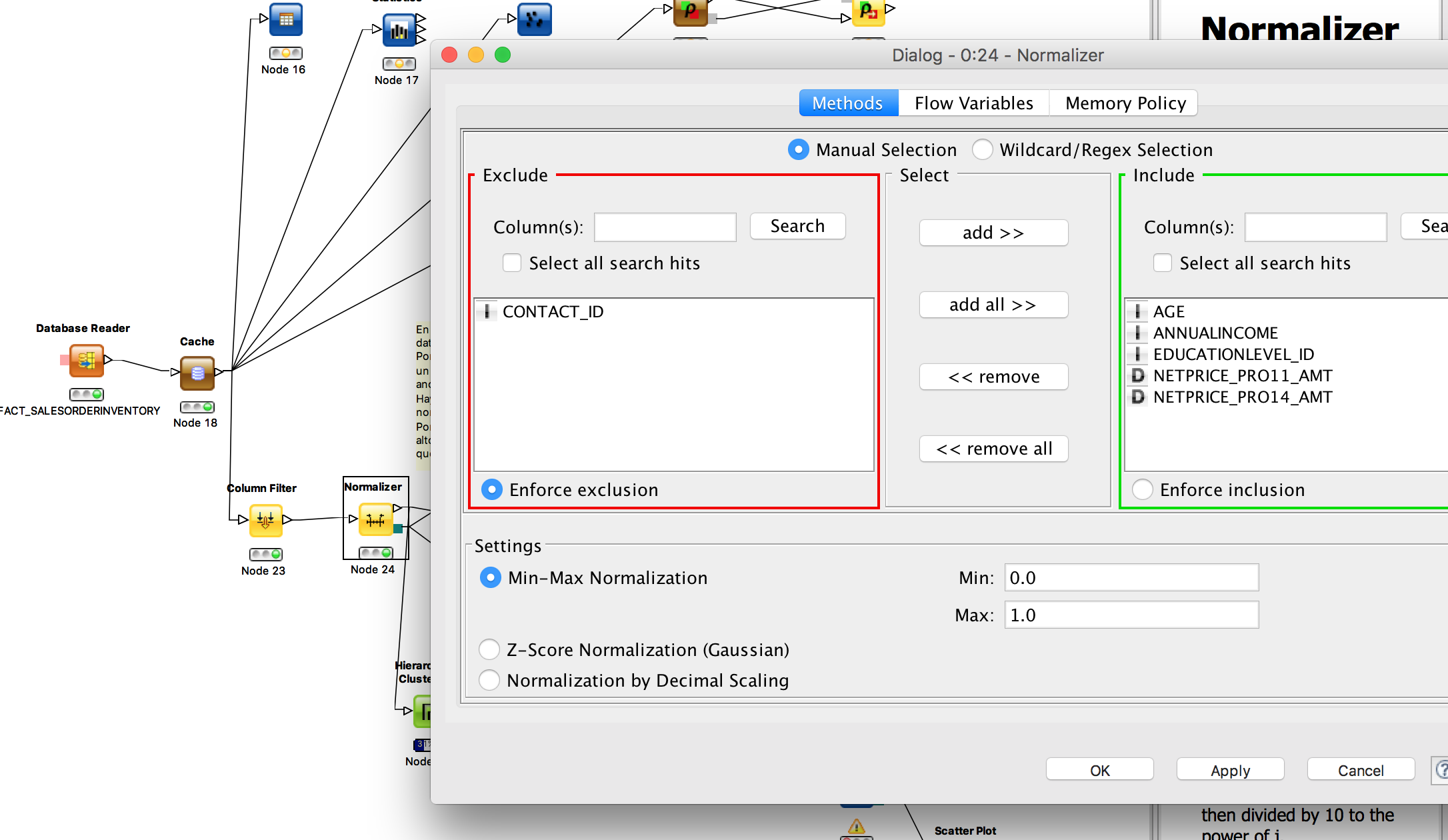
**3) Utilizamos el Column Filter para traer apenas las colunas necesarias para hacer el K-means.**



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.

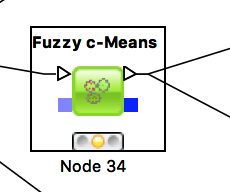
El proceso de normalizar es esencial para utilizar el algoritmo *Fuzzy c-Means* por que transforma los valores para una escala normalizada.

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

Importante:

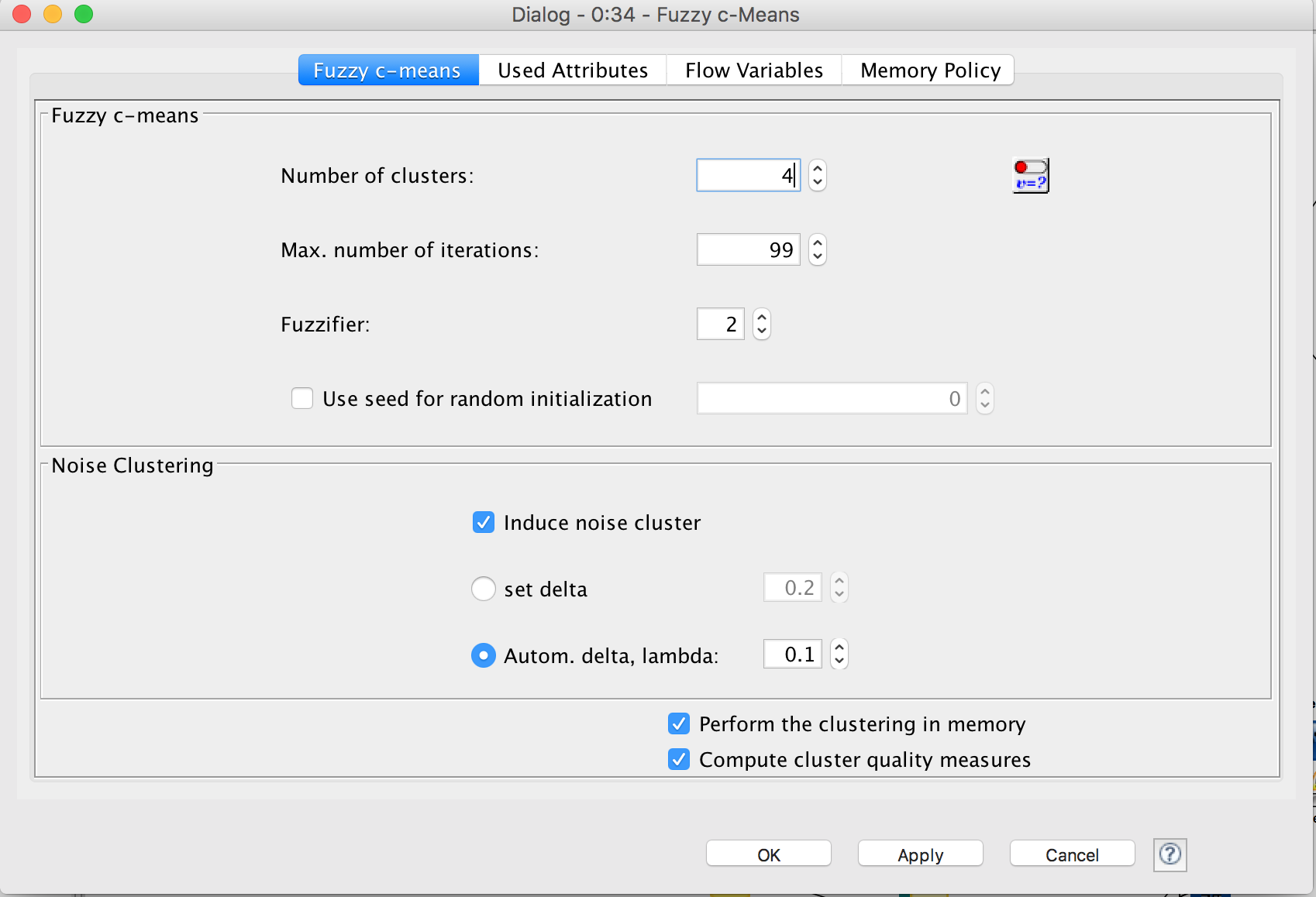
En la Normalizacion hay que quitar los datos anomalos. Por ejemplo una persona que tenga un sueldo muy alto es un data anomalo. Hay que quitarlo por que sino la normalizacion se queda mala. Por que el que tiene un sueldo muy alto queda como 1 y los otros van quedar como muy cerca de cero.

**5) Fuzzy c-Means**

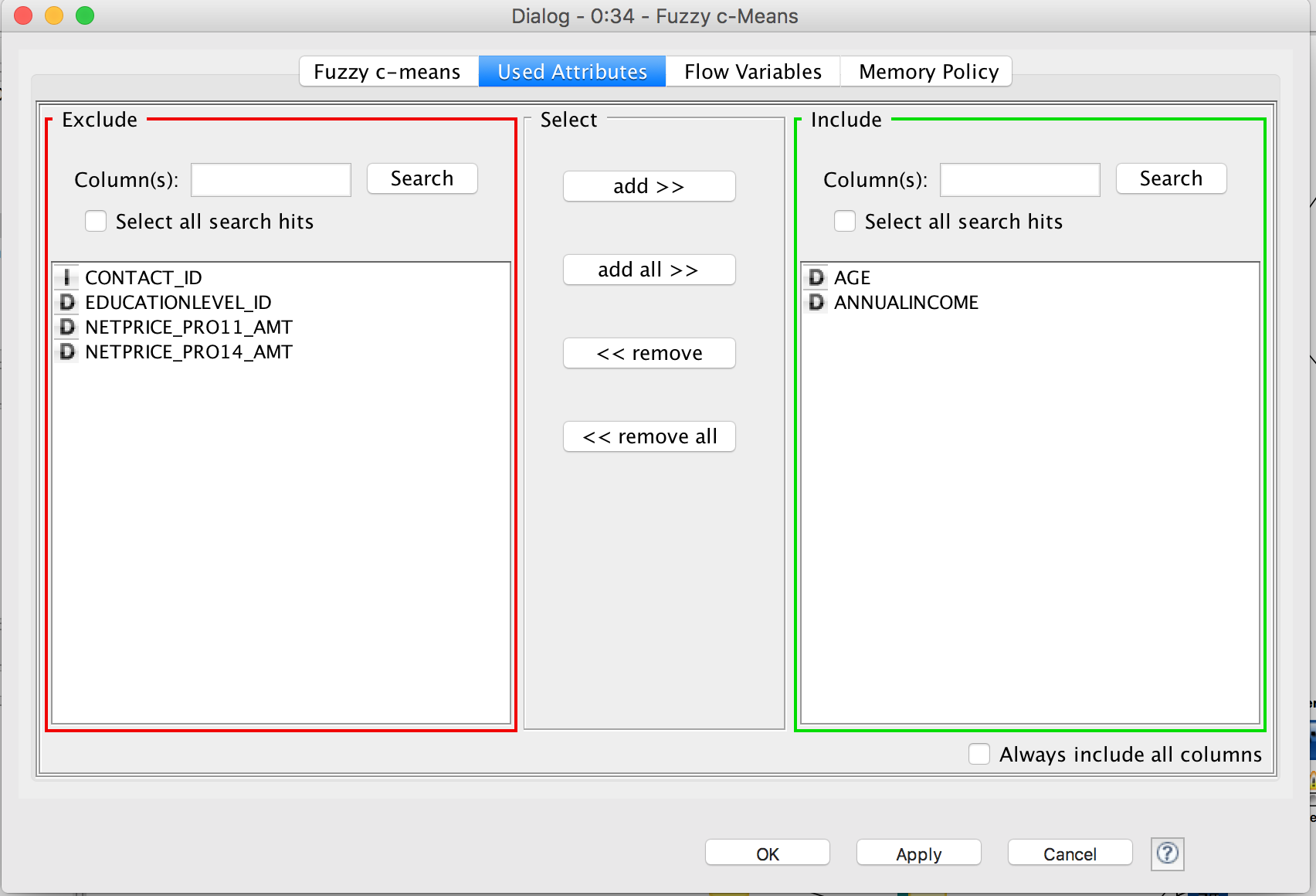


Para utilizar el Fuzzy c-Means hay que configurar el numero de clúster que hemos decidido dejar 4, hemos optado tener un clúster mas pequeño apenas de 4, pero se puede explicar las variables con mas clústeres.

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.



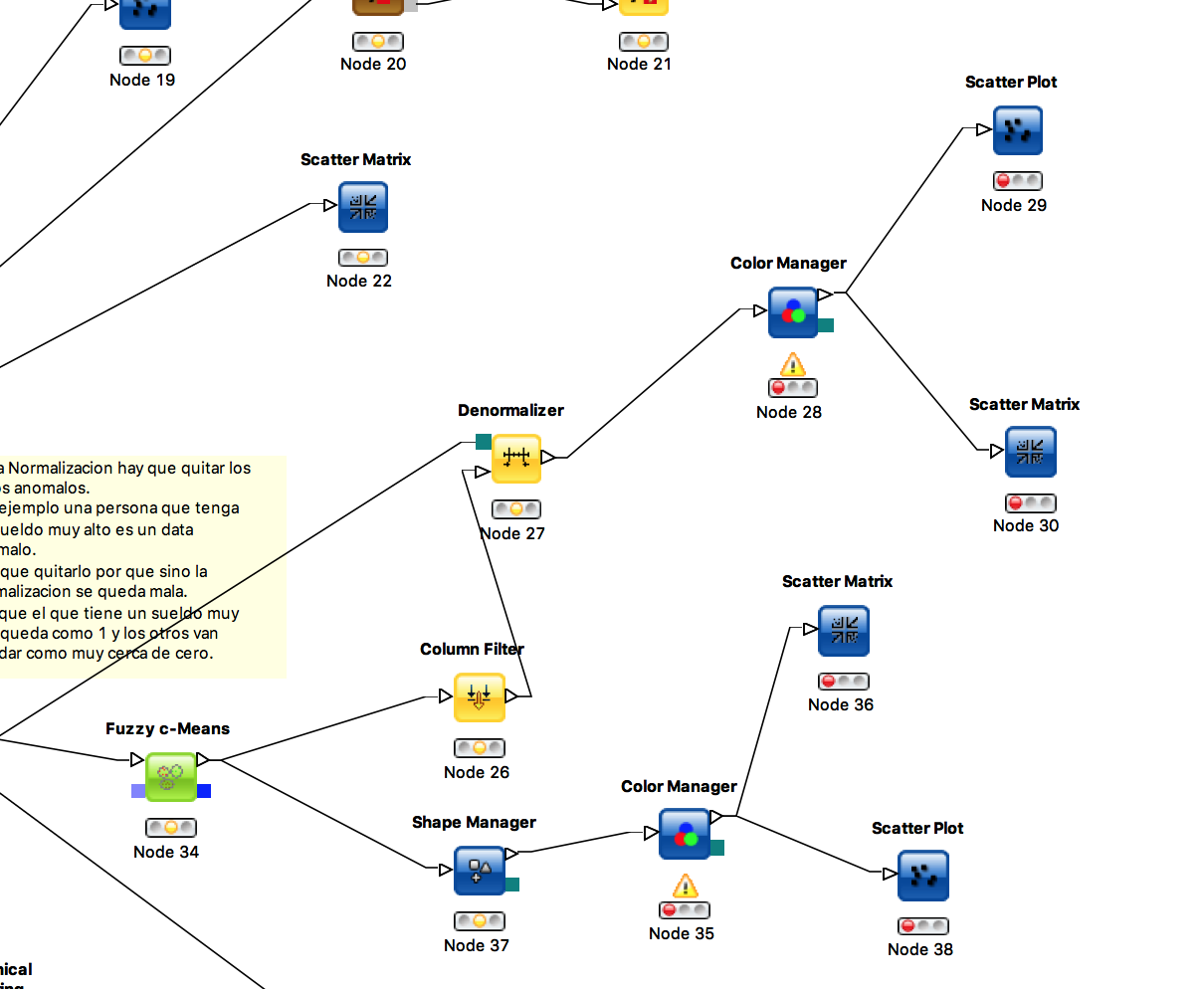
Hay que dejar solo las columnas AGE y ANNUALINCOME.



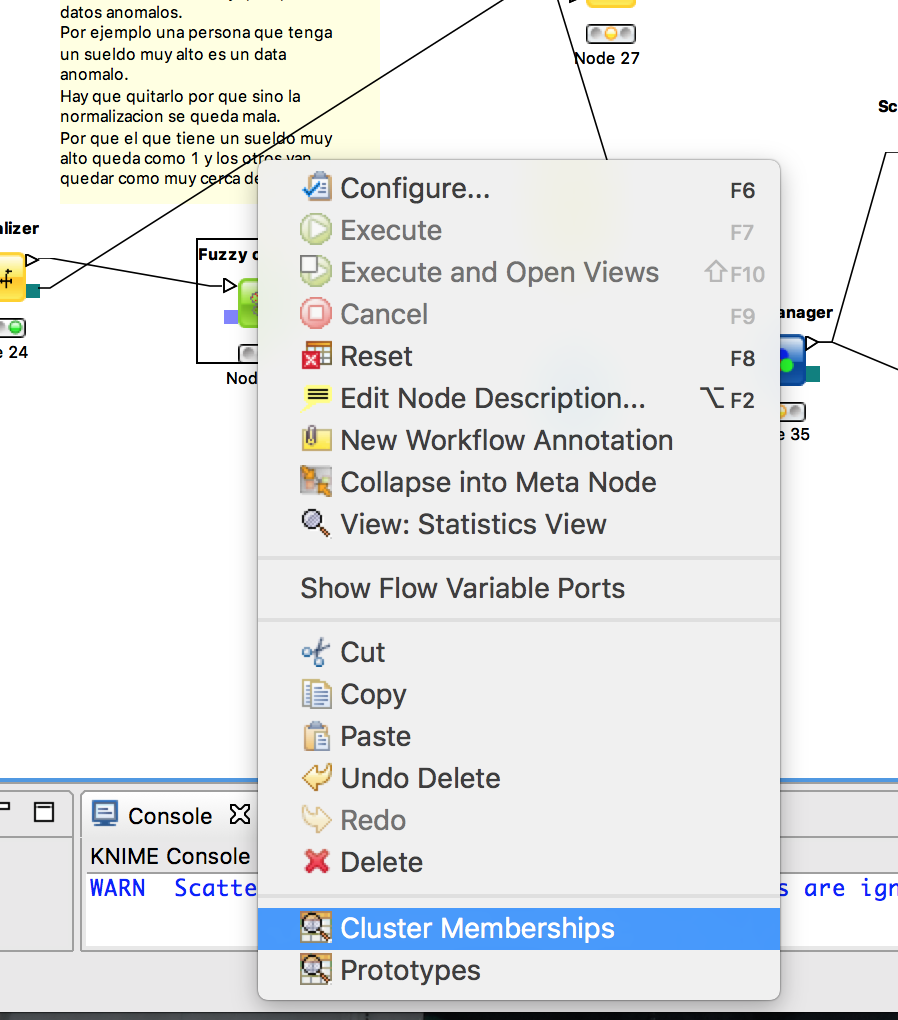
**6) Visualización de los datos con Scatter Plot o Scatter Matrix**

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

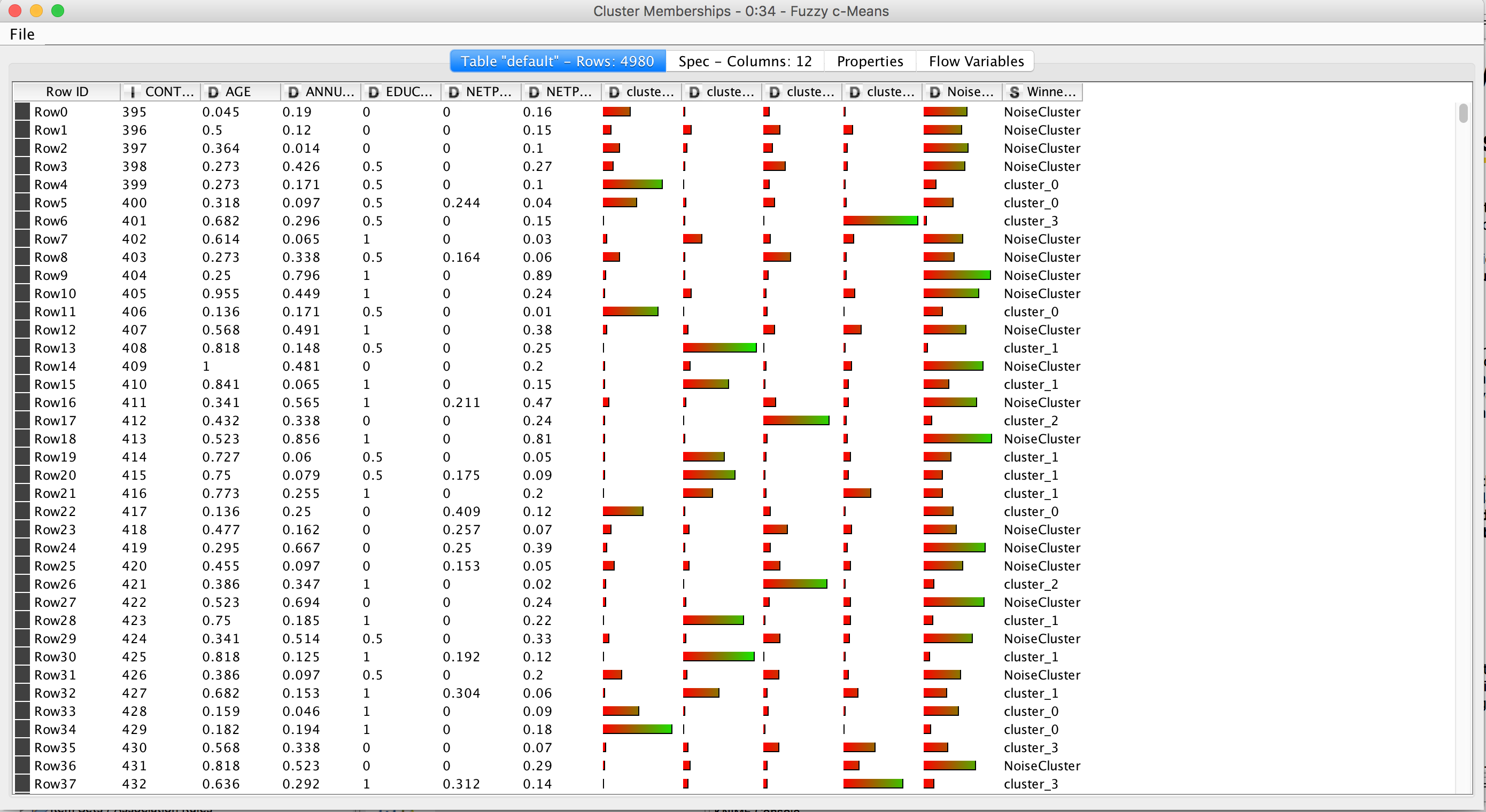
Se puede ver que hay dos flujos para la visualización: uno utilizando el componente Denormalizer y otro sin utilizar.



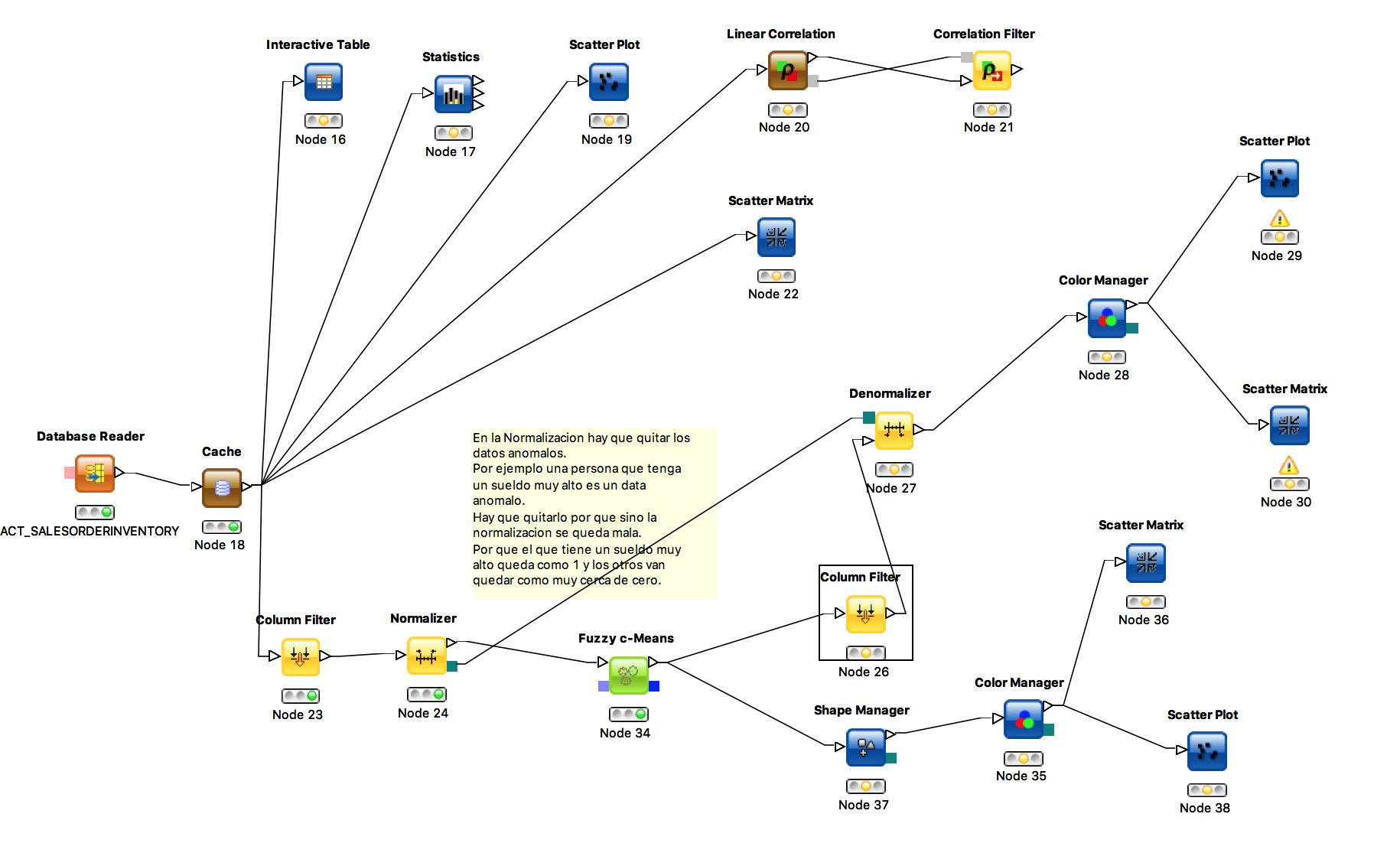
La opción *Cluster Memberships* cuando seleccionada nos enseña como se ha quedado cada clúster y también en Noise Cluster.



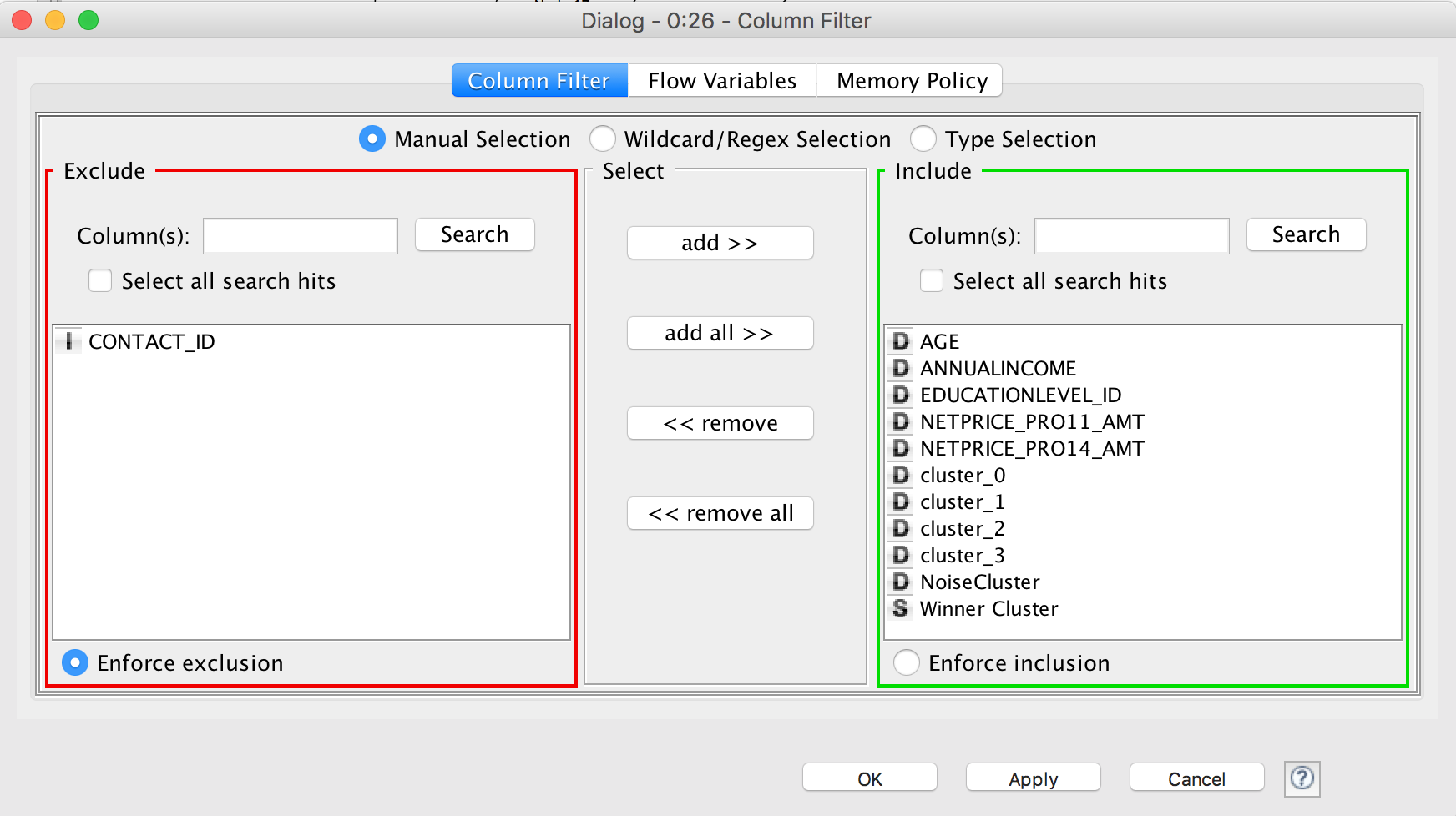
Se puede ver en la tabla Cluster Memberships de Fuzzy c-Means que la Row0 pertenece al cluster 0, pero hay ruido.



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



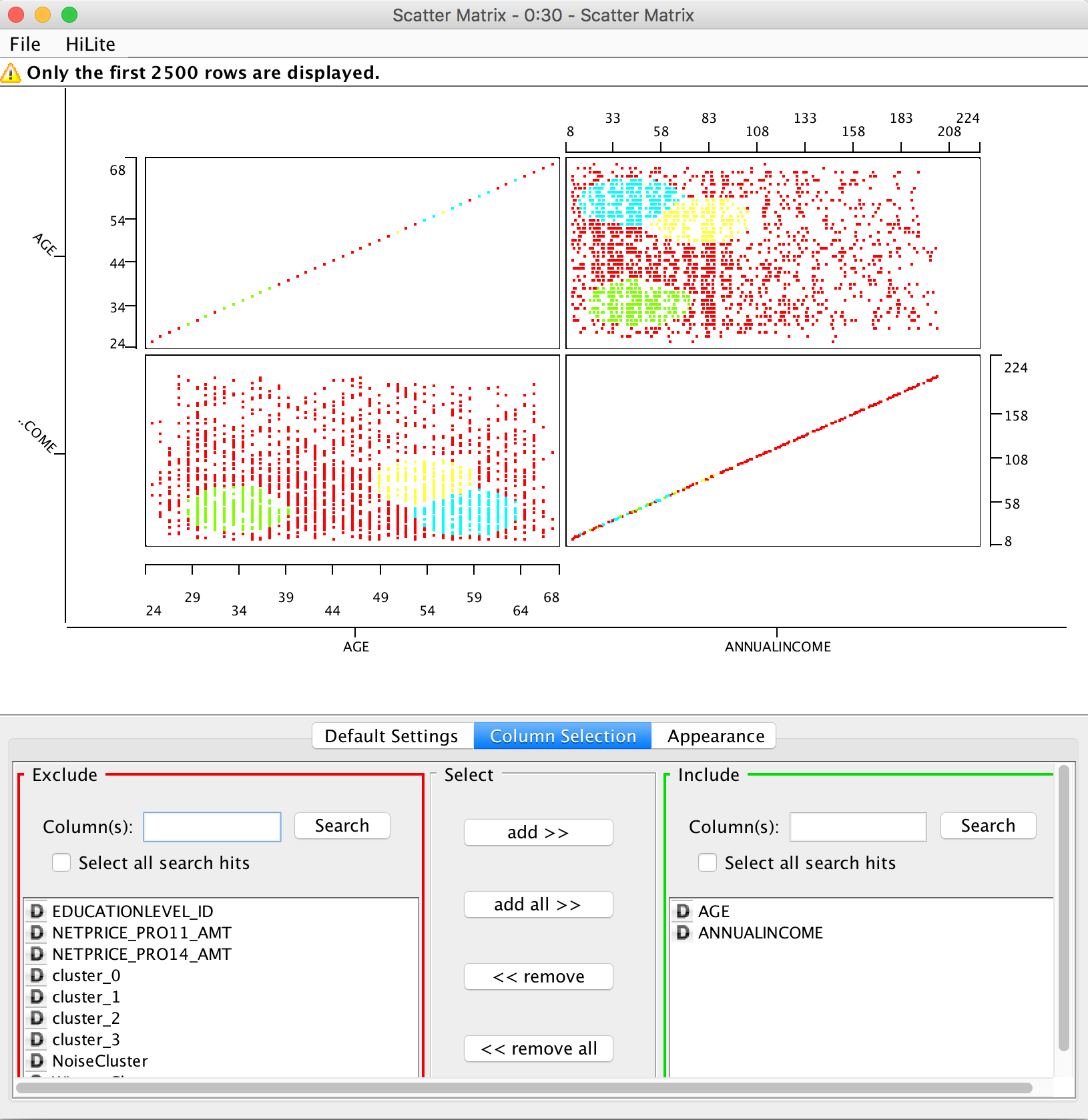
En el *Column Filter* se deja y quita las columnas como se puede ver la imagen abajo.



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

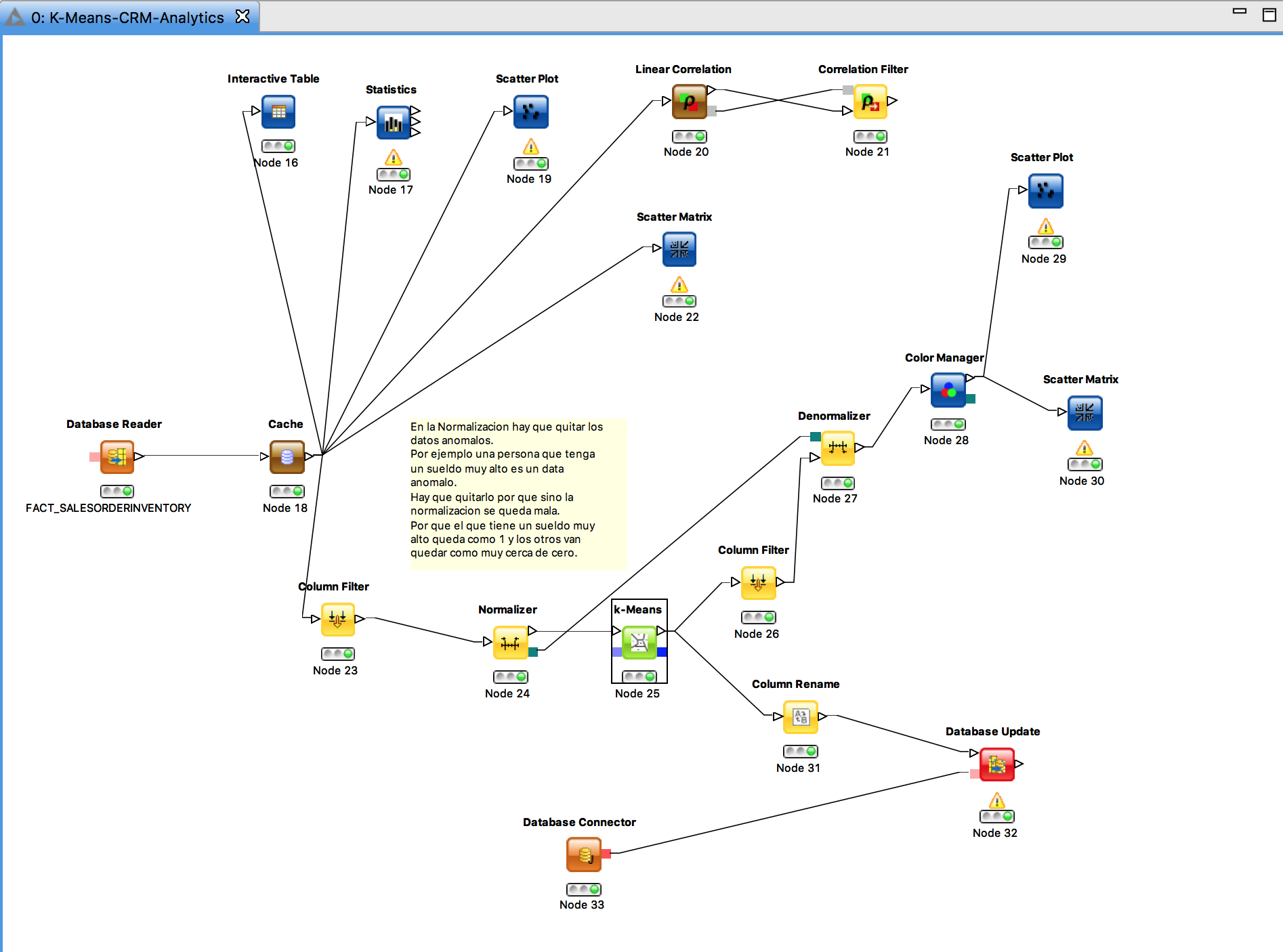


El *Scatter Matrix* se puede ver los mismo datos de una otra forma.

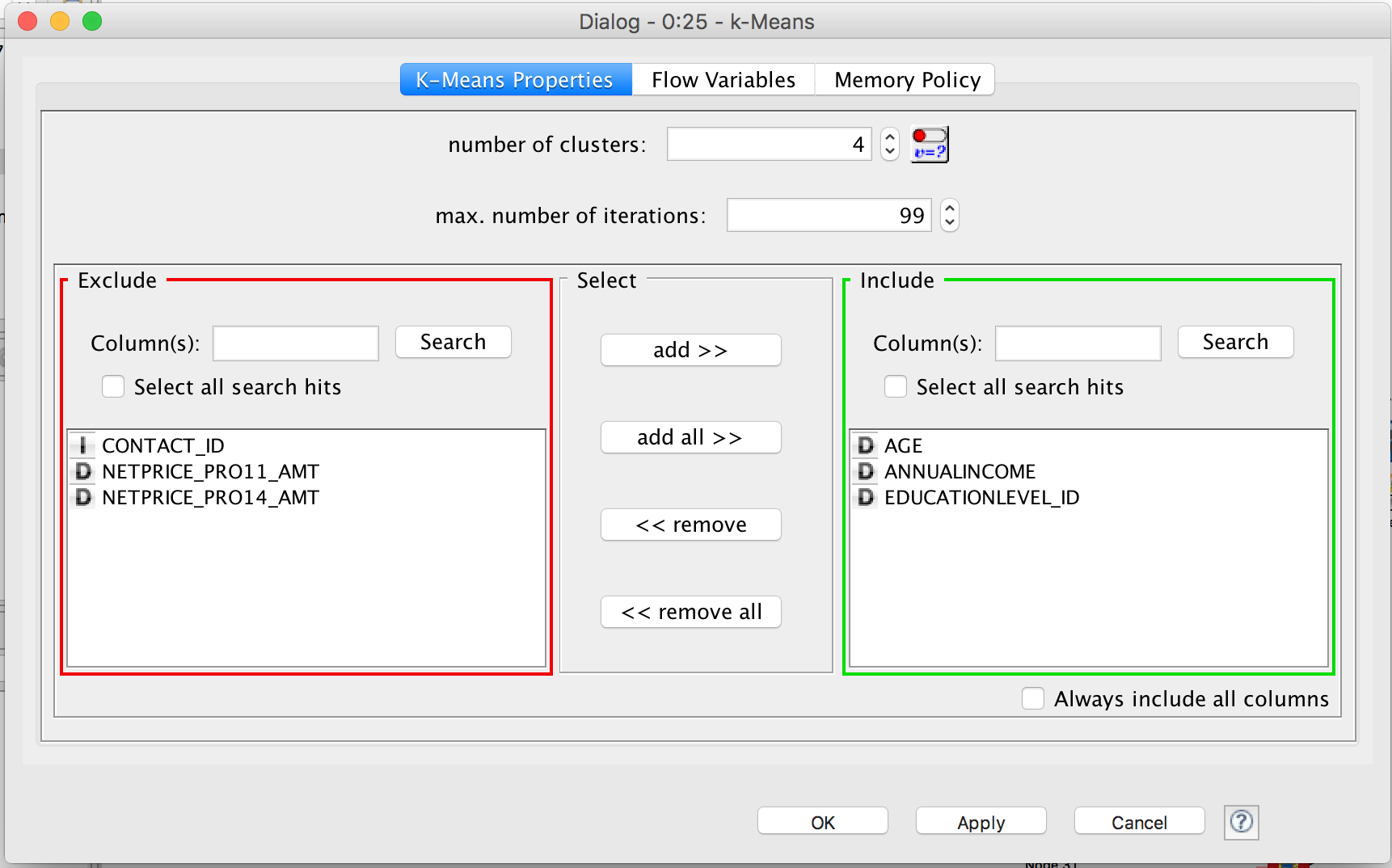


**7) K-Means y Visualización de los datos con Scatter Plot o Scatter Matrix**

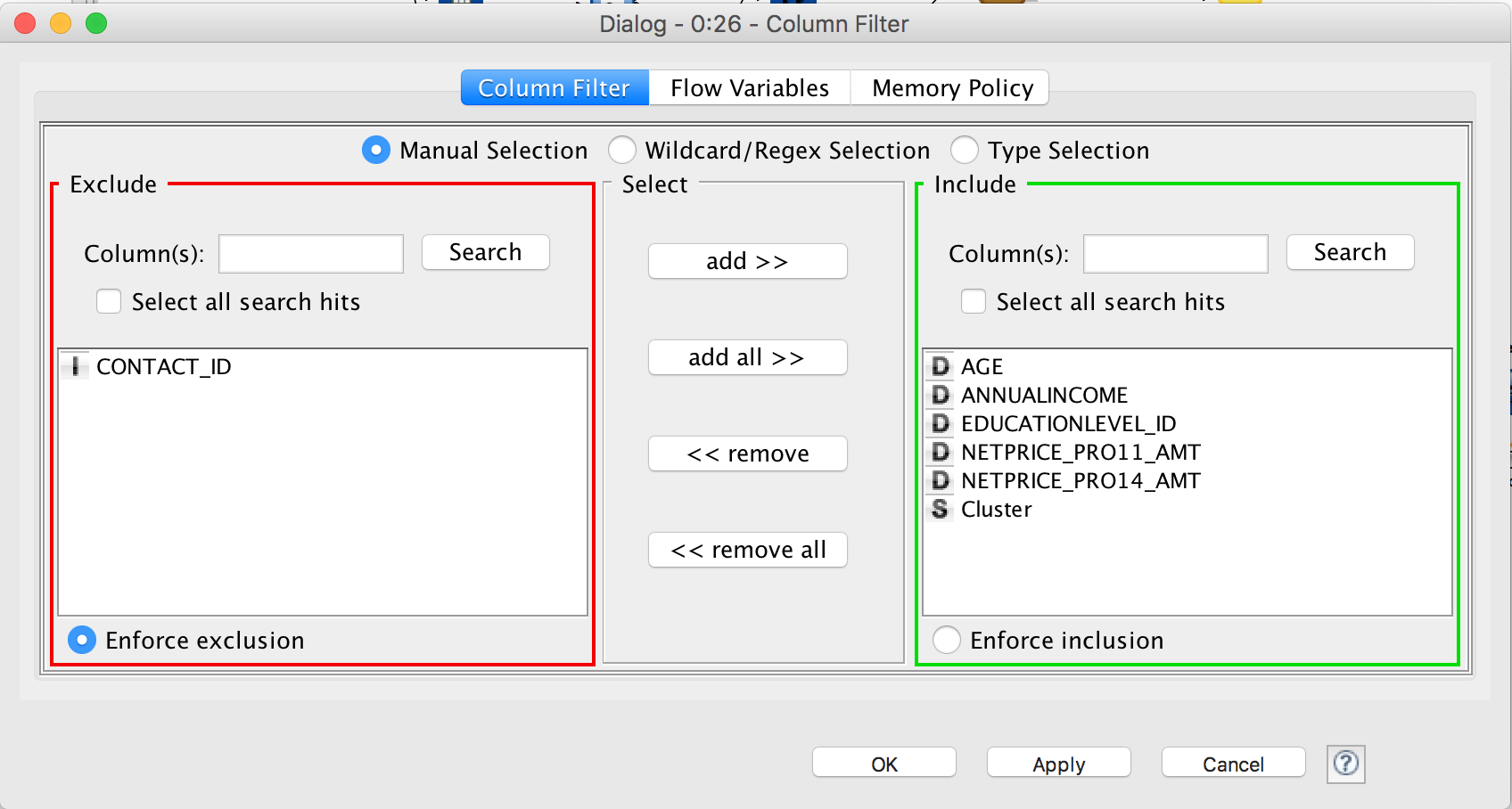
Lo mismo que hemos hecho con Fuzzy c-Means también hemos hecho con K-Means.



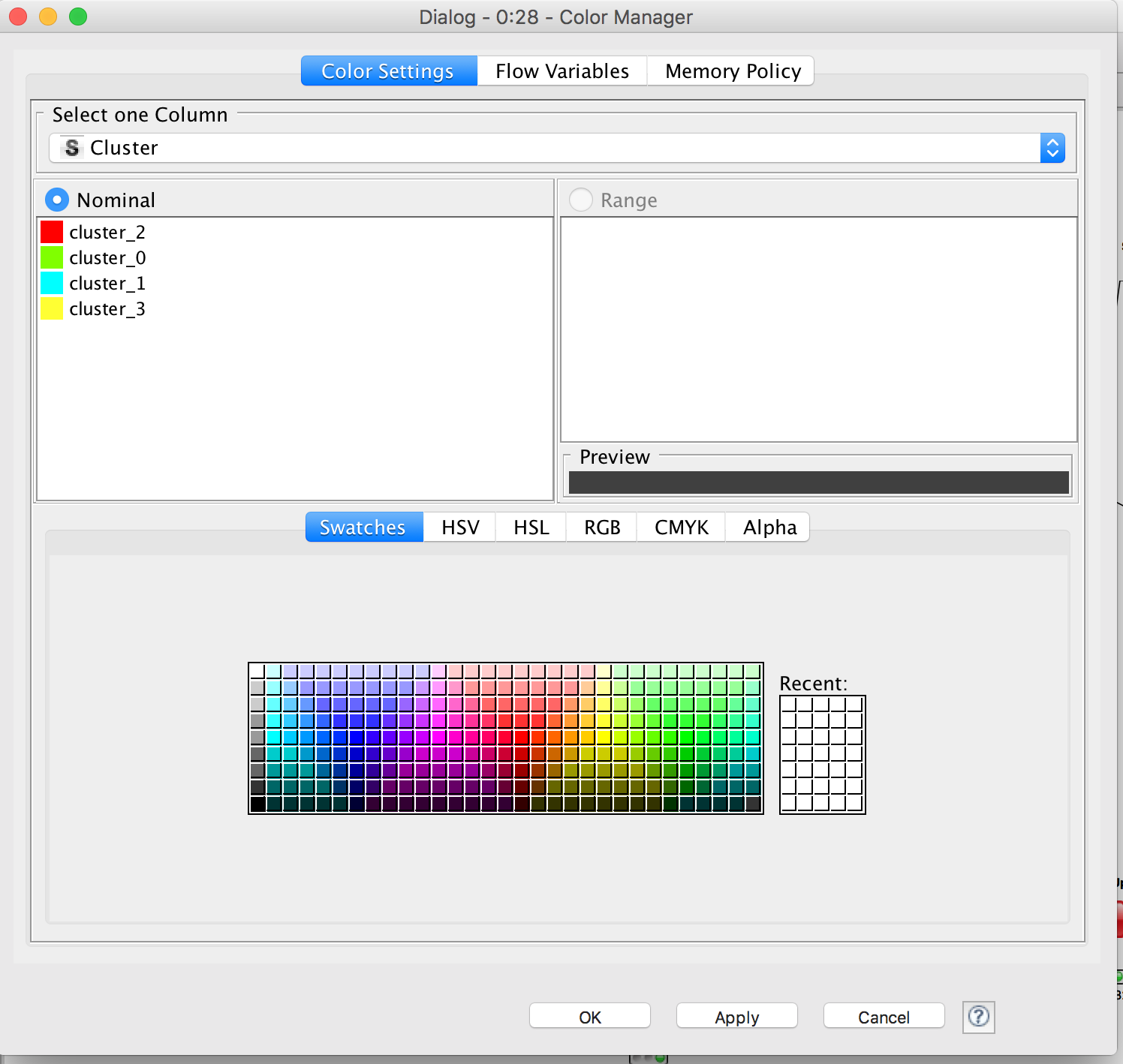
Configuración de K-Means.



Column Filter



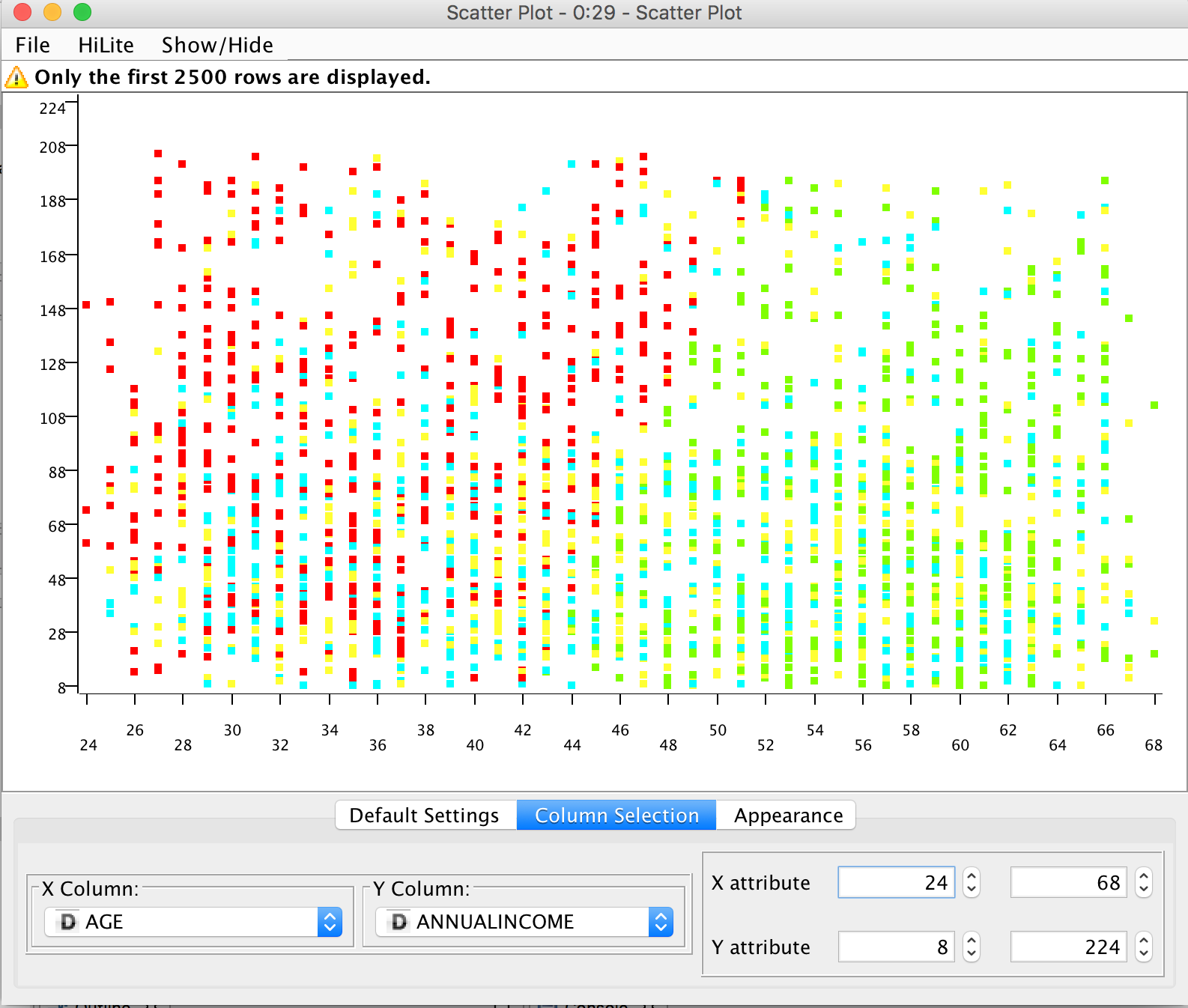
Color Manager



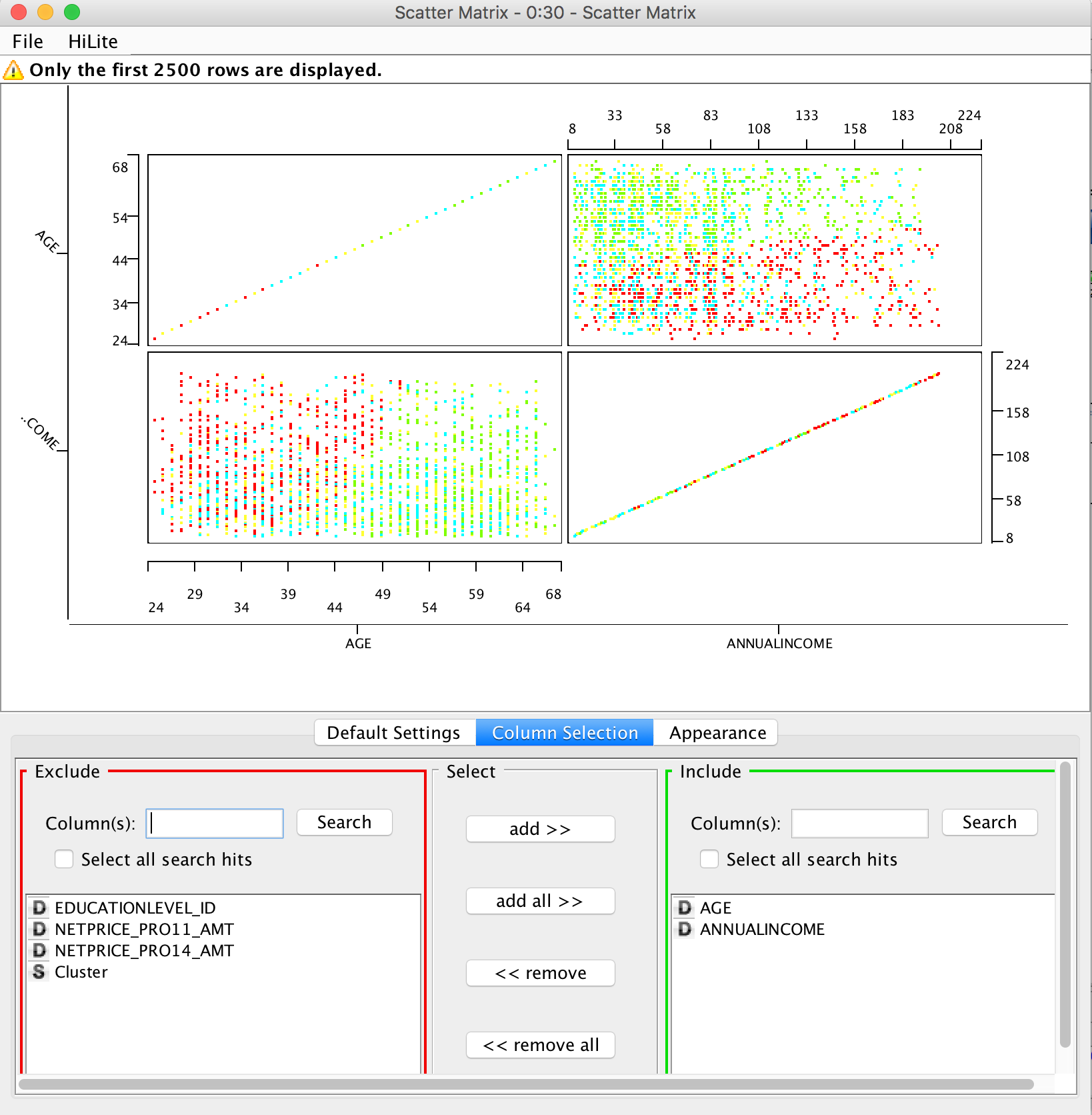
Scatter Plot de K-means donde X = AGE e Y = ANNUALINCOME.

Se puede ver que los rojos están mas concentrado 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.

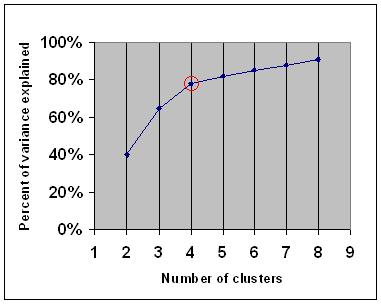


# 2. CONCLUSIONES

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 quedar se 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/>