UNIVERSIDAD DE GRANADA

Visualización con Knime

Tratamiento Inteligente de Datos
Master Universitario en Ingeniería Informática

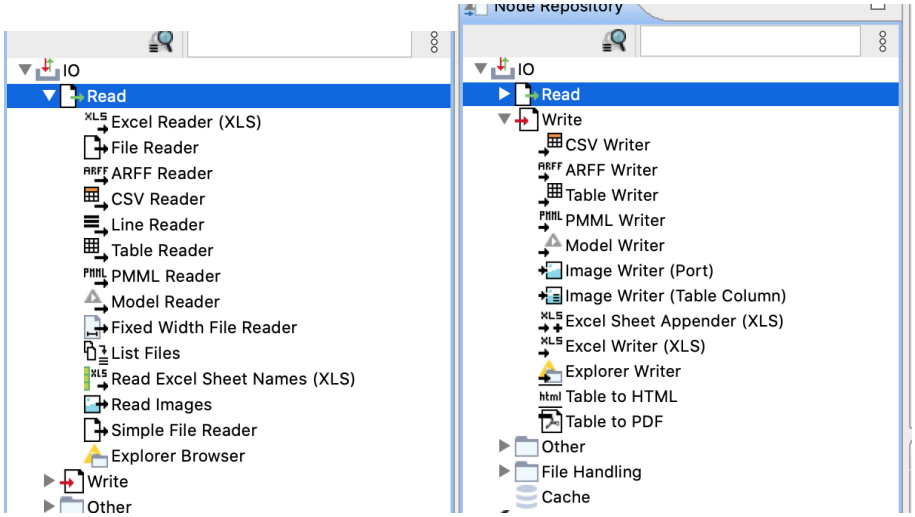
1. Objetivos

Uno de los primeros pasos en el proceso de análisis de datos es la visualización. En muchos casos, la visualización de datos es útil para explorarlos y crear gráficos. Por ejemplo, aquellos que se utilizan en los informes para describir los datos y el sistema subyacente. Knime proporciona muchos nodos para la visualización de datos, incluyendo diagramas de dispersión, gráficos circulares, gráficos de caja, histogramas, así como las nubes de etiquetas y visualizaciones de redes.

En esta práctica veremos el uso básico de Knime mediante la construcción de diferentes gráficas a partir de un conjunto de datos. El objetivo es familiarizarse con el programa y conocer algunos métodos simples de viasualización de datos.

2. Nodos para estrada/salida de datos

Knime cuenta con una serie de nodos para entrada y salidas en la carpeta `IO` del repositorio de nodos.



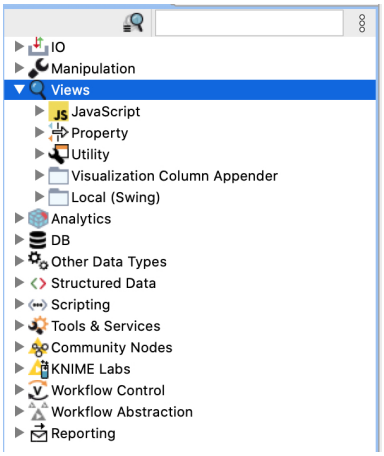
En general, para utilizarlos, simplemente se debe configurar el nodo añadiendo la ruta del archivo que contiene los datos; aunque dependiendo del tipo/formato de archivo se puede necesitar alguna configuración extra. Algunos de los nodos más comunes son:

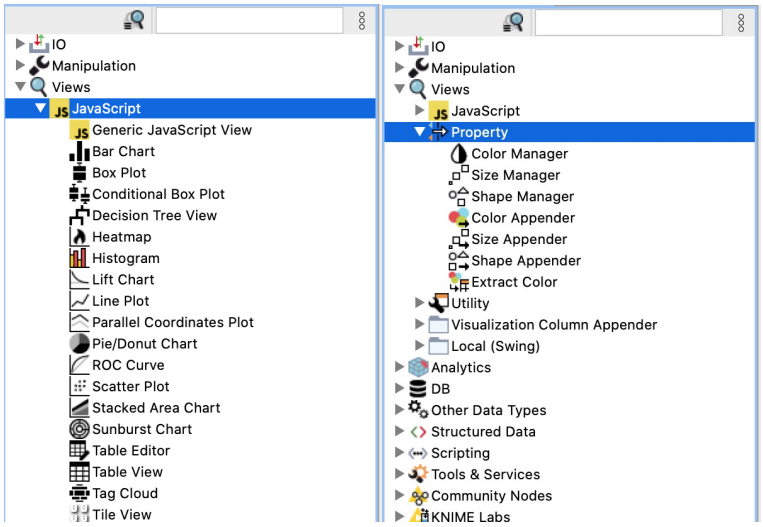
- `File Reader`. Este nodo puede ser utilizado para leer datos de un archivo ASCII o de una ubicación URL. Se puede configurar para leer varios formatos. Cuando se abre el diálogo de configuración del nodo y se proporciona un nombre de archivo, este trata de adivinar el formato mediante el análisis del contenido del archivo. Una versión más sencilla se encuentra en el nodo `Simple File Reader`.

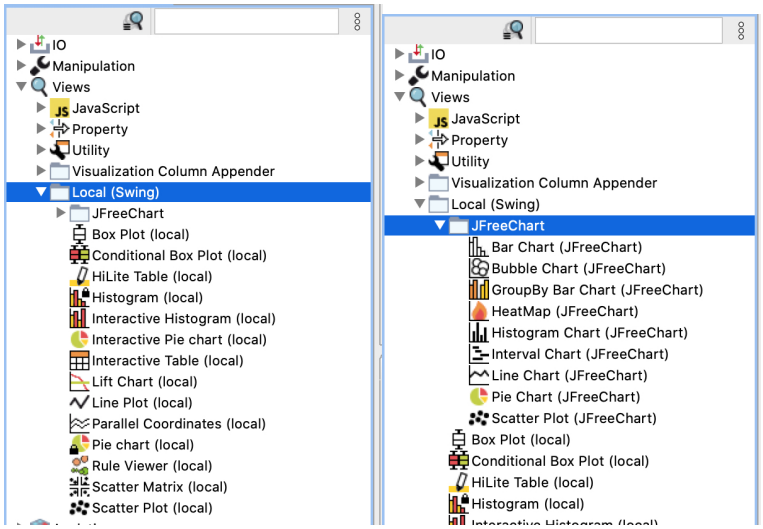
- `Excel Reader (XLS)`. Este nodo lee una hoja de cálculo y la proporciona en su puerto de salida. Lee sólo los datos de una hoja (por defecto la primera). Puede leer sólo los datos numéricos (o strings), pero no diagramas, imágenes, u otros artículos. Knime soporta actualmente datos de tipo String, Double, e Int. La lectura de archivos de gran tamaño necesita mucho tiempo y utiliza una gran cantidad de memoria (especialmente los archivos en formato .xlsx).

3. Nodos de visualización

Knime proporciona muchos nodos para la visualización de datos, incluyendo diagramas de dispersión, gráficos circulares, gráficos de caja, histogramas, así como las nubes de etiquetas y visualizaciones de redes. Los nodos clásicos están disponibles en la carpeta `Views`.

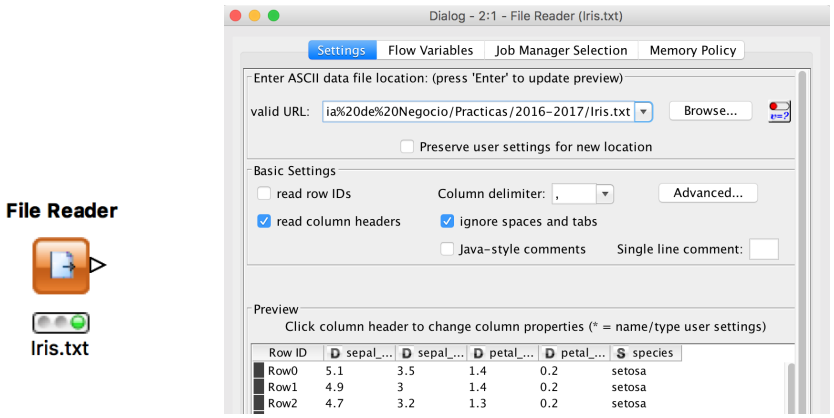




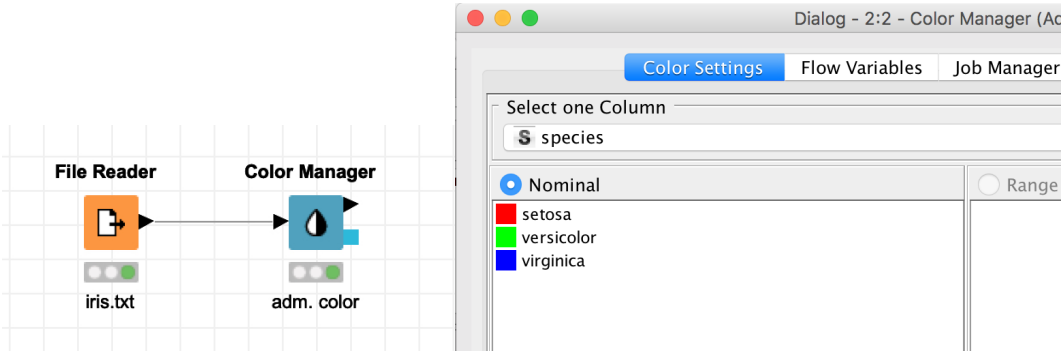


Un ejemplo de uso:

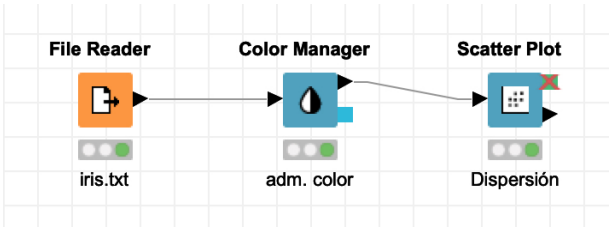
1. Crea un nuevo flujo de trabajo y carga el conjunto de datos del archivo `iris.txt` .

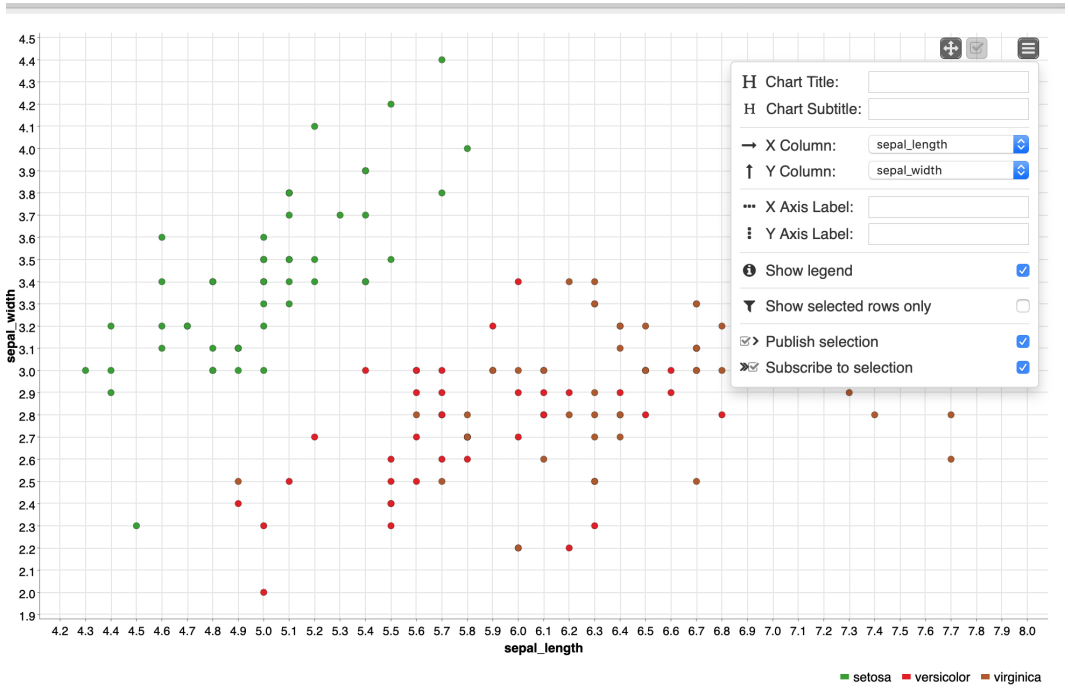


2. Anexar colores basados en la columna de la especie mediante el Administrador de color (hay que ejecutar antes el nodo de lectura del fichero).



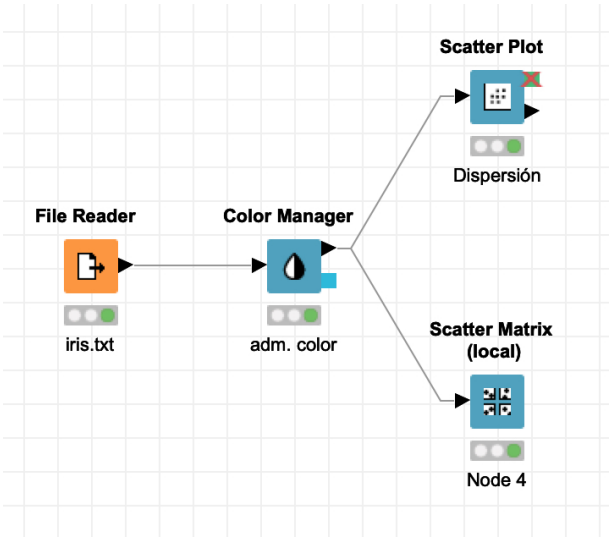
3. Visualizar los datos utilizando un gráfico de dispersión `Scatter Plot` en `Views/JavaScript`.

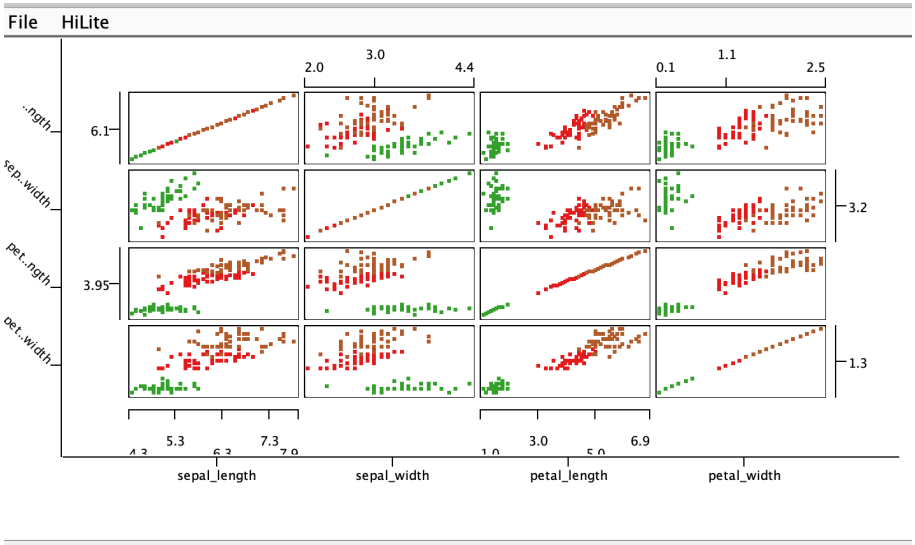




Con el botón en la esquina superior izquierda del diagrama de dispersión podemos cambiar la configuración del diagrama.

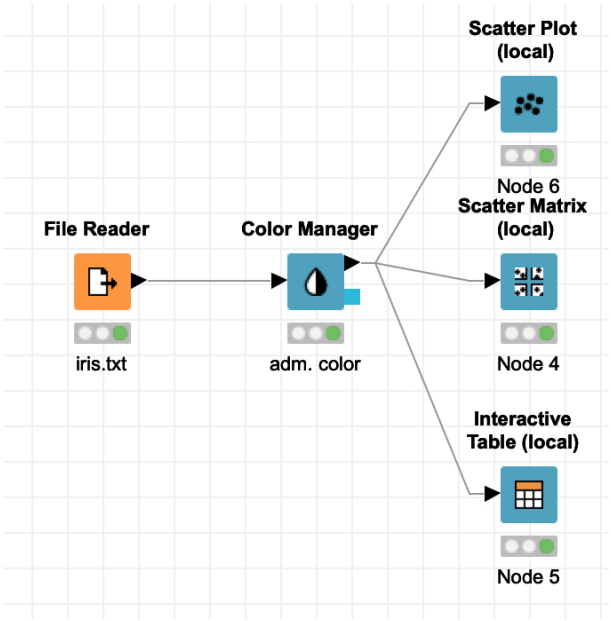
4. Visualizar los datos utilizando una matriz de dispersión `Scatter Matrix`.



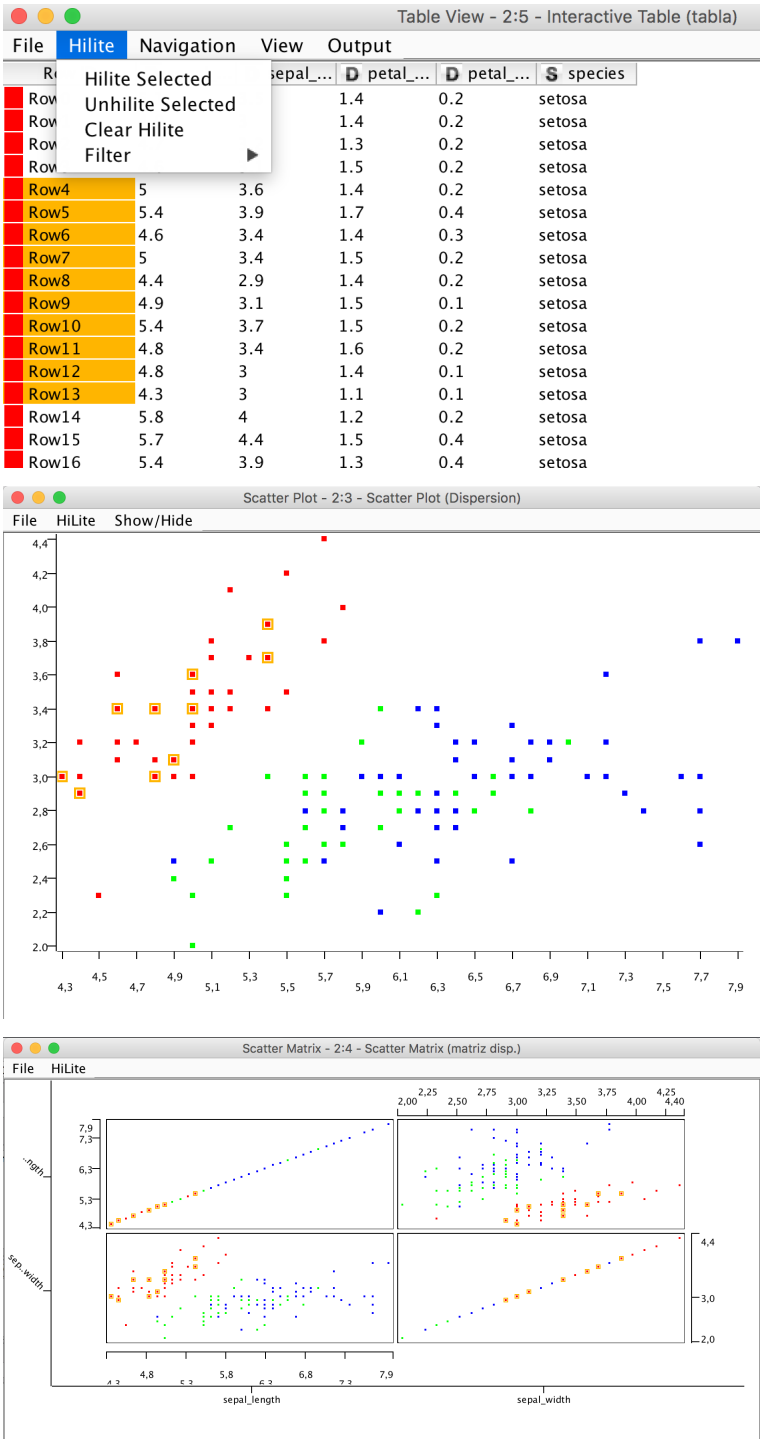


Añade todos los atributos en la matrix, ¿puedes deducir algo de la visualización de las matrices?

5. Visualizar los datos utilizando el nodo `Interactive Table`, utilizamos también `Scatter Plot` en `Views/Local`.



6. Selecciona algunos puntos en la tabla interactiva (Hilite), observa como se muestran en el gráfico de dispersión y en la matriz de dispersión.



7. Realiza el proceso contrario, selecciona puntos en el gráfico de dispersión y observa qué puntos son en la tabla interactiva.

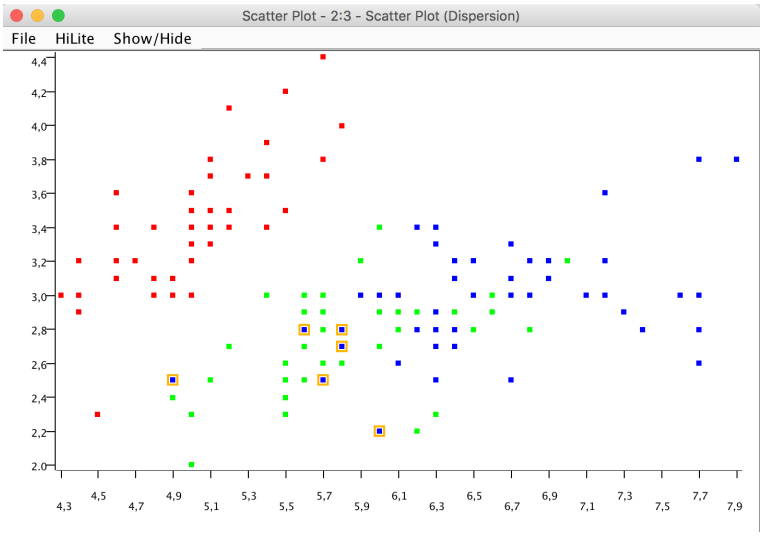


Table View - 2:5 - Interactive Table (tabla)

File Hilite Navigation View Output

Row ID	sepal_...	sepal_...	petal_...	petal_...	species
Row99	5.7	2.8	4.1	1.3	versicolor
Row100	6.3	3.3	6	2.5	virginica
Row101	5.8	2.7	5.1	1.9	virginica
Row102	7.1	3	5.9	2.1	virginica
Row103	6.3	2.9	5.6	1.8	virginica
Row104	6.5	3	5.8	2.2	virginica
Row105	7.6	3	6.6	2.1	virginica
Row106	4.9	2.5	4.5	1.7	virginica
Row107	7.3	2.9	6.3	1.8	virginica
Row108	6.7	2.5	5.8	1.8	virginica
Row109	7.2	3.6	6.1	2.5	virginica
Row110	6.5	3.2	5.1	2	virginica
Row111	6.4	2.7	5.3	1.9	virginica
Row112	6.8	3	5.5	2.1	virginica
Row113	5.7	2.5	5	2	virginica
Row114	5.8	2.8	5.1	2.4	virginica
Row115	6.4	3.2	5.3	2.3	virginica
Row116	6.5	3	5.5	1.8	virginica

También puedes encontrar más nodos en la carpeta `KNIME Labs/JavaScript Views (Labs)` Para realizar estadísticas básicas (media, desviación típica, máximo, mínimo, etc) puedes utilizar los nodos en la carpeta `Analytics/Statistics` y `KNIME Labs/Statistics`.

4. Tareas a realizar

1. (6 puntos) Considera los datos de la tabla `prestamo.xls`.

   a) Algunas variables son Booleanas (toman valor 0 ó 1, verdadero o falso), sin embargo el lector de ficheros Excel de KNIME los reconoce como de tipo entero. Lo mismo ocurre con la variable `Family` y `Education`. Utiliza los nodos de cambio de tipo de dato en la carpeta `Manipulation/Column/Convert & Replace` para que cada variable esté correctamente tipificada.

   b) Utiliza el nodo `Data Explorer` para una visualización y estudio básico de las variables de la base de datos. ¿Existe alguna característica que te llame la atención?¿Por qué?

   c) Realiza un diagrama de burbujas para analizar el número de miembros de la familia de las personas recogidas en la tabla. Dibuja un diagrama de barras de esta característica.

   d) Realiza un diagrama de dispersión para analizar visualmente los campos edad y años de experiencia. ¿Que opinas sobre la gráfica? Estudia el coeficiente de correlación entre dichas variables.

e) Estudia las matrices de dispersión del resto variables. Calcula la correlación lineal para las variables en las que eso tenga sentido. ¿Sería razonable eliminar alguna característica en base a este estudio?

f) Asigna colores a las filas en al número de componentes de la familia. A partir de esto, crea un diagrama de dispersión entre la edad y el salario donde los puntos tengan un color dependiendo del número de componentes de la familia. ¿Se puede observar algo en este diagrama?

g) Filtra la base de datos según la variable `Family` (nodo `Row Filter`). Por ejemplo, considera los clientes con familias de 4 miembros (o cualquier otro número, si lo prefieres). Realiza un estudio parecido a los puntos anteriores. ¿Se puede observar algo destacable para estos clientes?

h) De la base de datos filtrada anterior, calcula diagramas circulares para las variables nominales.

2. (4 puntos) Descarga el fichero `NBA.xlsx` desde la página de la asignatura de Google Classroom. Contiene datos sobre los jugadores de la NBA que han jugado los playoffs durante la temporada 2021-2022.

a) El nombre de las variables es demasiado largo. Cámbialos por otros más cortos utilizando el nodo `Column Rename`.

b) Calcula las medidas básicas utilizando el nodo `Statistics`. ¿Existe algún dato que te llame la atención?

c) Realiza diagramas de dispersión de las variables numéricas de tipo real (double). Al observar las gráficas, ¿consideras que algunas variables están relacionadas? ¿Cuáles? ¿Existe alguna explicación razonable para esto?

d) Calcula matriz de correlación de las variables numéricas utilizando el nodo `Linear correlation`. Compara con el punto anterior.

e) Realiza un diagrama de barras de la columna `POS` (posición del jugador).

f) Realiza diagramas de cajas de las variables numéricas. ¿Existen jugadores que consideras que sobresalen del resto en alguna característica (outliers)?

g) Calcula un diagrama circular de la columna `GP` (partidos jugados).

h) Selecciona un equipo y filtra la tabla para obtener los jugadores que pertenecen a dicho equipo, usando el nodo `Row Filter`. Selecciona 5 características numéricas y filtra la tabla para quedarse únicamente con esas columnas (además de la columna con el nombre). Con la tabla resultante dibuja un diagrama de coordenadas paralelas (dibuja la lineas paralelas en colores a partir de la columna del nombre). ¿Es posible observar algún jugador con unas características diferentes del resto? Prueba con otros equipos y otras características, si no ha sido posible.

i) Si tuvieras que dividir los jugadores en varios grupos, basándote en los estudios realizados, ¿cómo lo harías? ¿En cuántos grupos? ¿Cómo denominarías/describirías a cada grupo?