Visualización con Knime

UNIVERSIDAD DE GRANADA

Pablo Morenilla Pinos
morenillapablo@correo.ugr.es
TID Prácticas Grupo 1

Índice

Índice de figuras

1. Ejercicio 1. Considera los datos de la tabla prestamo.xls.

a) Algunas variables son Booleanas (toman valor 0 o 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 and Replace para que cada variable esté correctamente tipificada.

Para hacer este apartado, primero hay que cambiar el tipo de las columnas Family y Education de tipo entero a tipo cadena, para ello se usa un conversor, indicando qué columnas se quieren cambiar y cuáles no.

Una vez hechas de tipo String, se van a reemplazar los valores 1,2,3 y 4 en el caso de la columna Family, por los valores: Monoparental, parental, estándar y numerosa (valores elegidos, ya que no aparece ninguna asociación previa). Para el caso Education, los valores 1, 2 y 3 tendrán los valores asociados al excel.

Todo esto se hace con la celda de 'String manipulation':

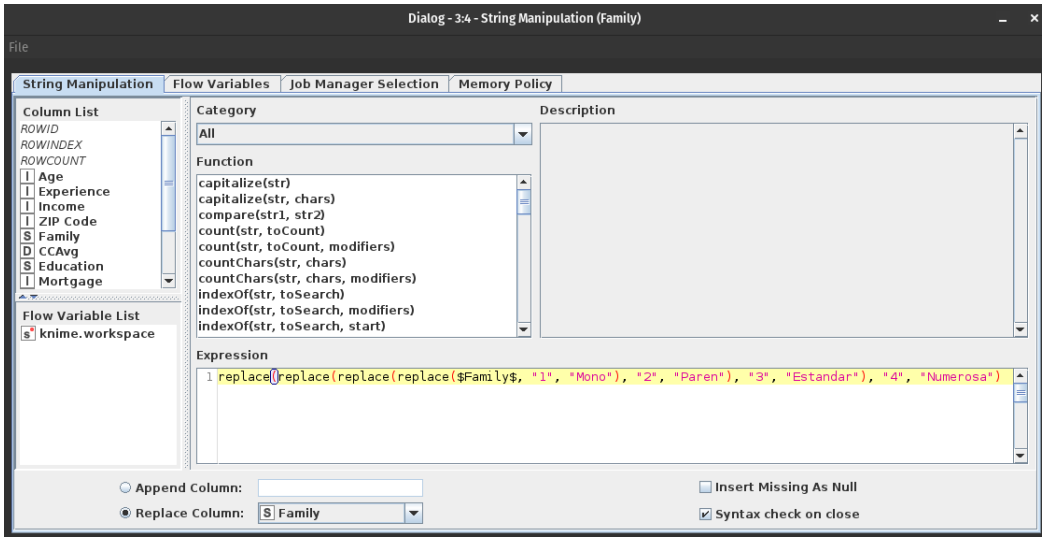


Figura 1: Cambio de valores columna Family.

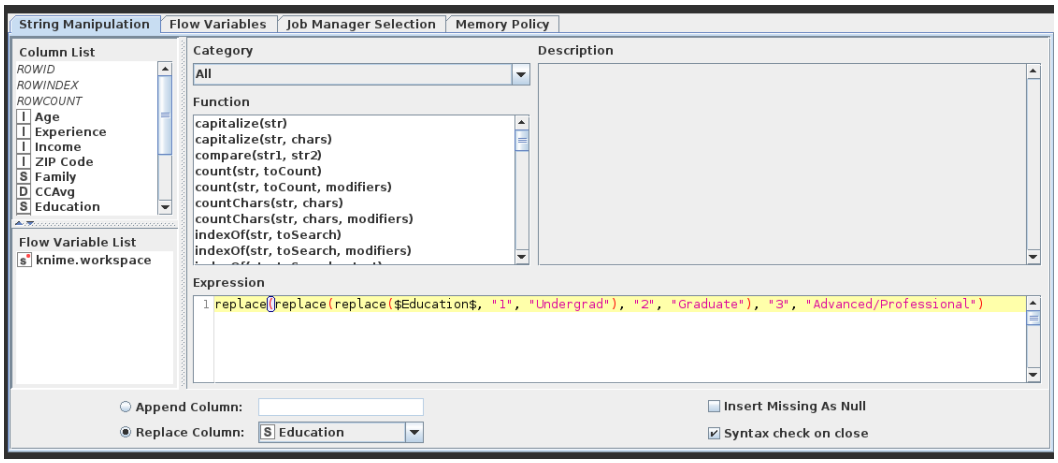
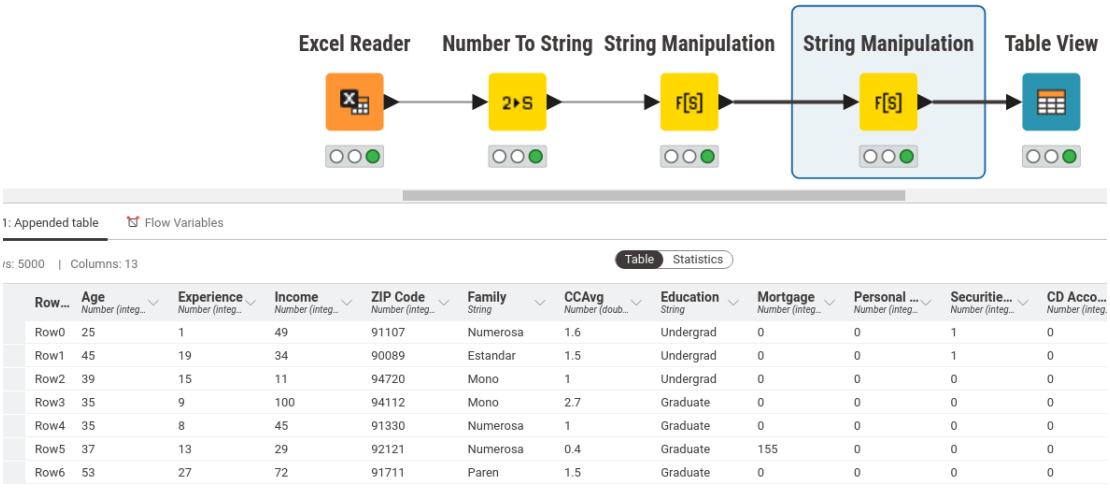


Figura 2: Cambio de valores columna Education.

Como se ha visto, en ambas capturas se usa la función replace concatenada, para evitar así crear una celda para cada uno de los cambios, en el replace indico la tabla sobre la que quiero cambiar, el valor a cambiar y el valor que se quiere añadir.

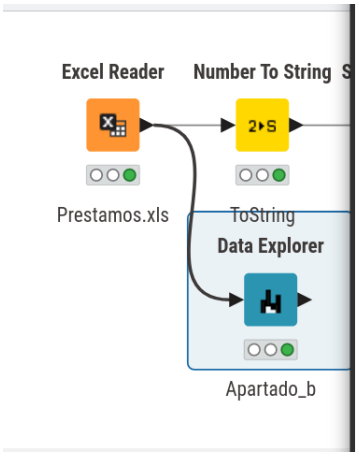


Row...	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal ...	Securitie...	CD Acco...
	Number (integ...	Number (integ...	Number (integ...	Number (integ...	String	Number (doub...	String	Number (integ...	Number (integ...	Number (integ...	Number (integ...
Row0	25	1	49	91107	Numerosa	1.6	Undergrad	0	0	1	0
Row1	45	19	34	90089	Estandar	1.5	Undergrad	0	0	1	0
Row2	39	15	11	94720	Mono	1	Undergrad	0	0	0	0
Row3	35	9	100	94112	Mono	2.7	Graduate	0	0	0	0
Row4	35	8	45	91330	Numerosa	1	Graduate	0	0	0	0
Row5	37	13	29	92121	Numerosa	0.4	Graduate	155	0	0	0
Row6	53	27	72	91711	Paren	1.5	Graduate	0	0	0	0

Figura 3: Visualización del resultado final.

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?

Para ver los datos con Data explorer primero hay que añadir este nodo, ya que se trata de una extensión, desde la página de knime se arrastra el módulo al programa y se instala, hecho esto se añade al IO del excel y visualizamos los datos:



Search:

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness
Age		23	67	45.338	11.463	131.404	-0.029
Experience		-3	43	20.105	11.468	131.514	-0.026
Income		8	224	73.774	46.034	2119.104	0.841
ZIP Code		90005	96651	93169.703	1759.714	3096594.472	-0.296
Family		1	4	2.396	1.148	1.317	0.155
CCAvg		0	10	1.938	1.748	3.054	1.598
Education		1	3	1.881	0.840	0.705	0.227
Mortgage		0	635	56.499	101.714	10345.698	2.104
Personal Loan		0	1	0.096	0.295	0.087	2.744

Figura 4: Visualización datos con nodo Data Explorer.

Destacan la variación que hay entre cada fila respecto a las columnas, es decir, por ejemplo la desviación estándar varía mucho entre la fila CD Acount con un valor de 0.238 y la fila ZIP Code con 1759.714. Esto ocurre en bastante más campos, también se debe que los cálculos que se realizan son sobre datos no bien definidos, por ejemplo, no tiene mucho sentido hacer una desviación estándar del código postal, el valor mínimo o máximo del código postal.

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.

Para realizar un diagrama de burbuja hacemos como en el caso anterior, importamos la extensión bubble chart, le añadimos los colores que queremos que tengan los distintos valores que tiene la columna 'Family' y configuramos el diagrama para que sea una relación 'Income/Age' agrupado por 'Family':

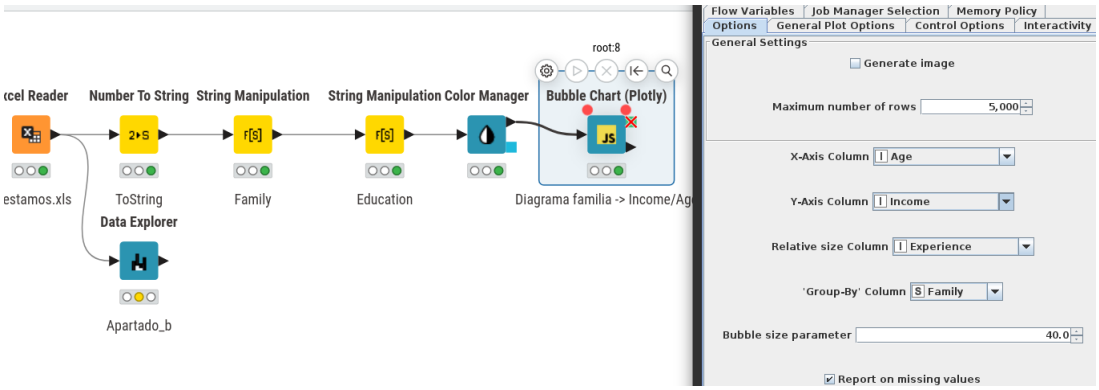


Figura 5: Configuración diagrama burbuja.

Hecho esto, se muestran dos imágenes, la primera es la vista general del diagrama y la segunda ya dentro de un rango más específico para poder ver de mejor manera los datos:

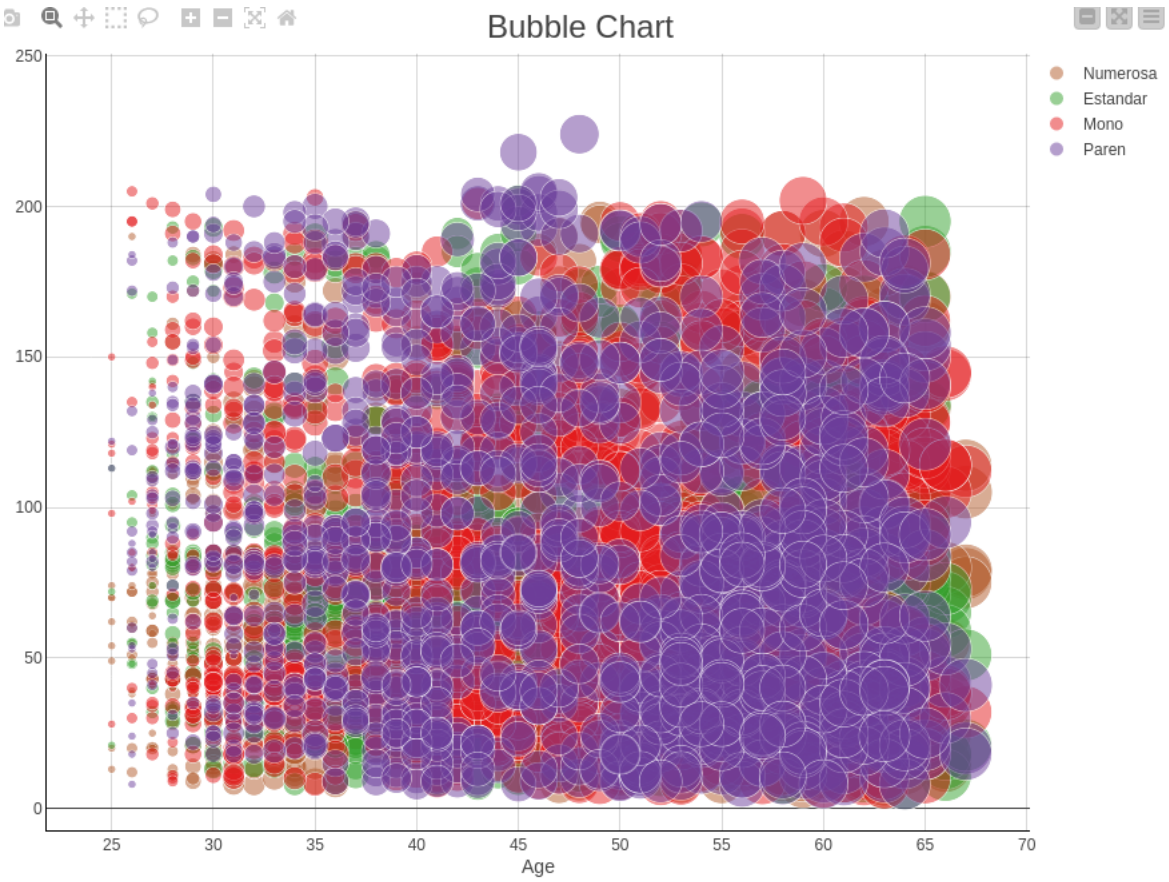


Figura 6: Diagrama burbuja.

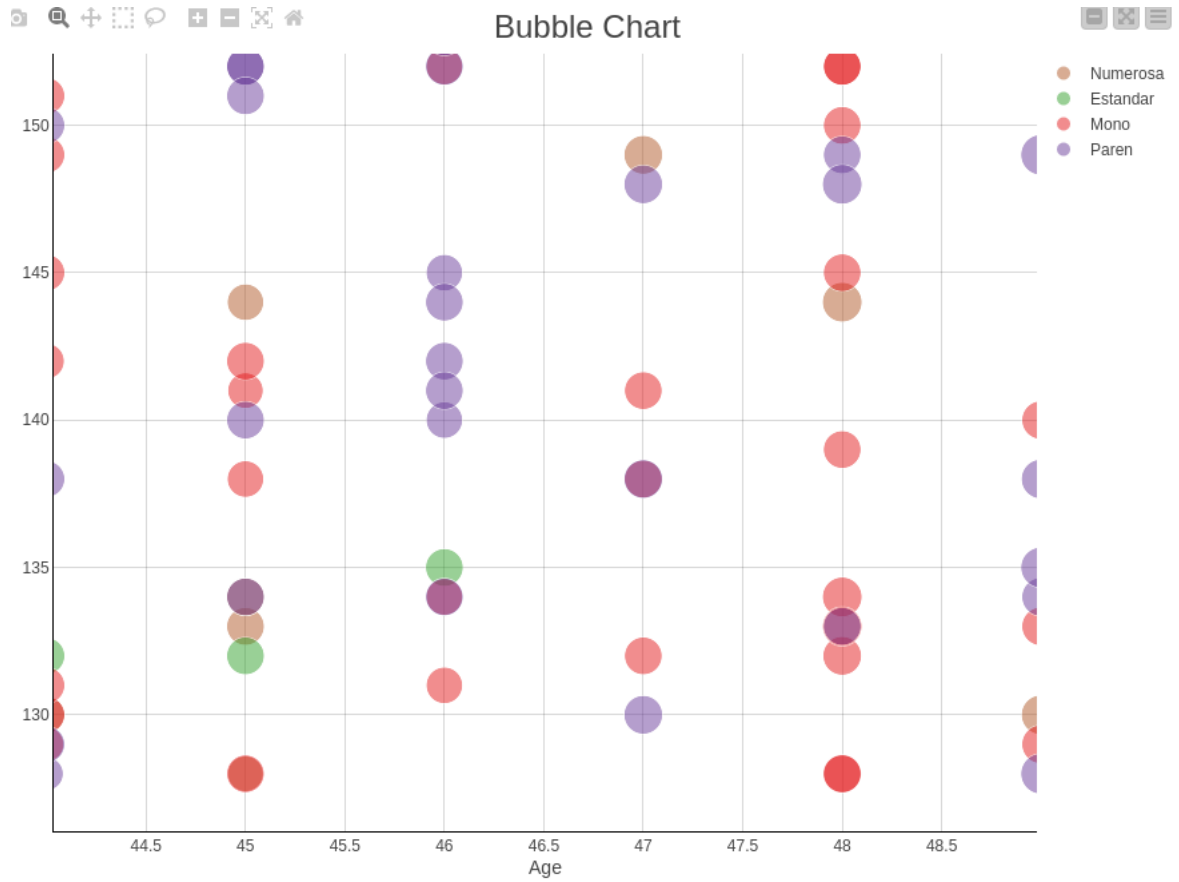


Figura 7: Diagrama ampliado.

Ahora, si intentamos visualizar los datos en un diagrama de barras, usaremos el tamaño familiar y será respecto a la edad, pero se hará con la edad media para poder visualizar de mejor manera los datos:

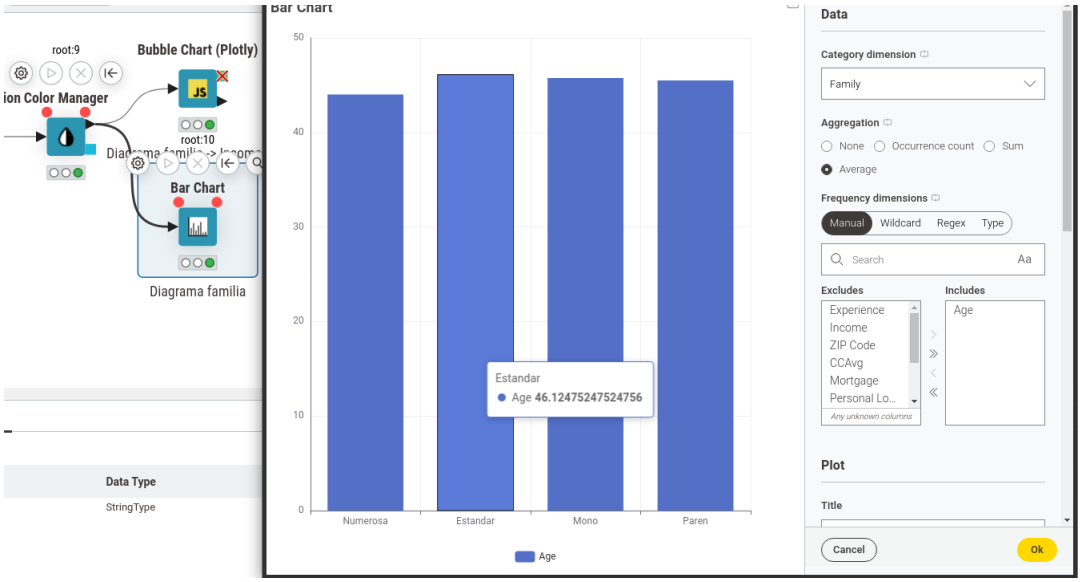


Figura 8: Diagrama de barras familia respecto a la edad media.

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

Visualizamos los datos con el diagrama de dispersión:

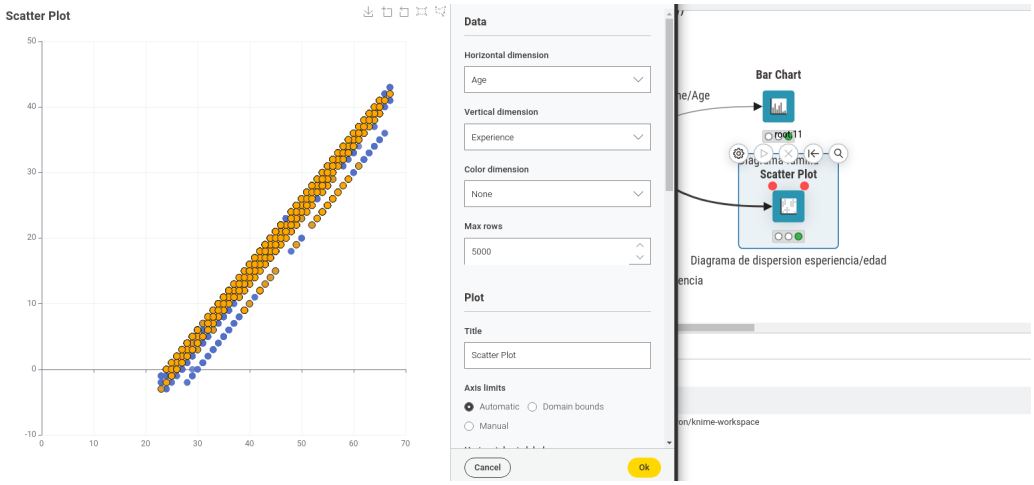


Figura 9: Diagrama de dispersión edad/experiencia.

Luego usamos la celda de correlación de Pearson, en la imagen se muestra como se ha configurado para que sea edad respecto a la experiencia. Se puede ver que tiene una correlación de 0.994, esto nos dice existe una correlación positiva. En este caso las variables estarían asociadas en sentido directo. Es casi 1, luego es casi una correlación perfecta. De este modo que la edad respecto a la experiencia es un buen indicador para evaluar los datos.

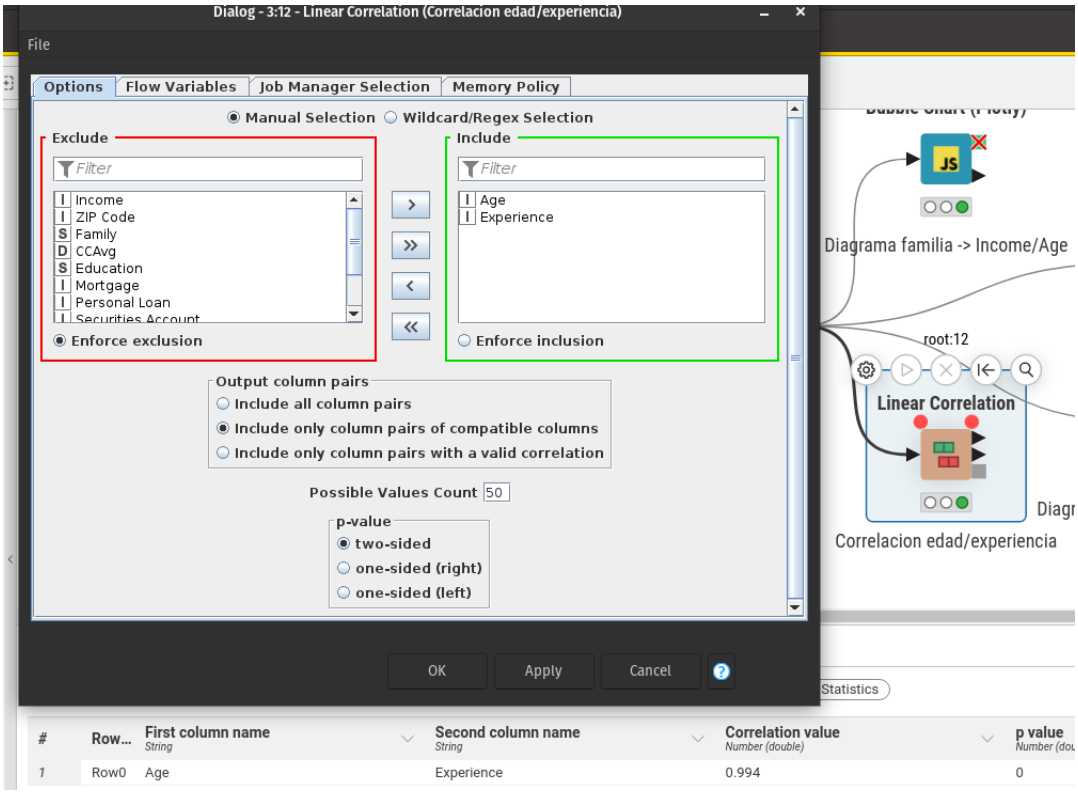


Figura 10: Correlación edad/experiencia y configuración.

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?

Si visualizamos con la matriz de dispersión todas las variables, se pueden ver que hay ciertos campos que pueden no ser importantes para el estudio, como se muestra en la siguiente imagen:

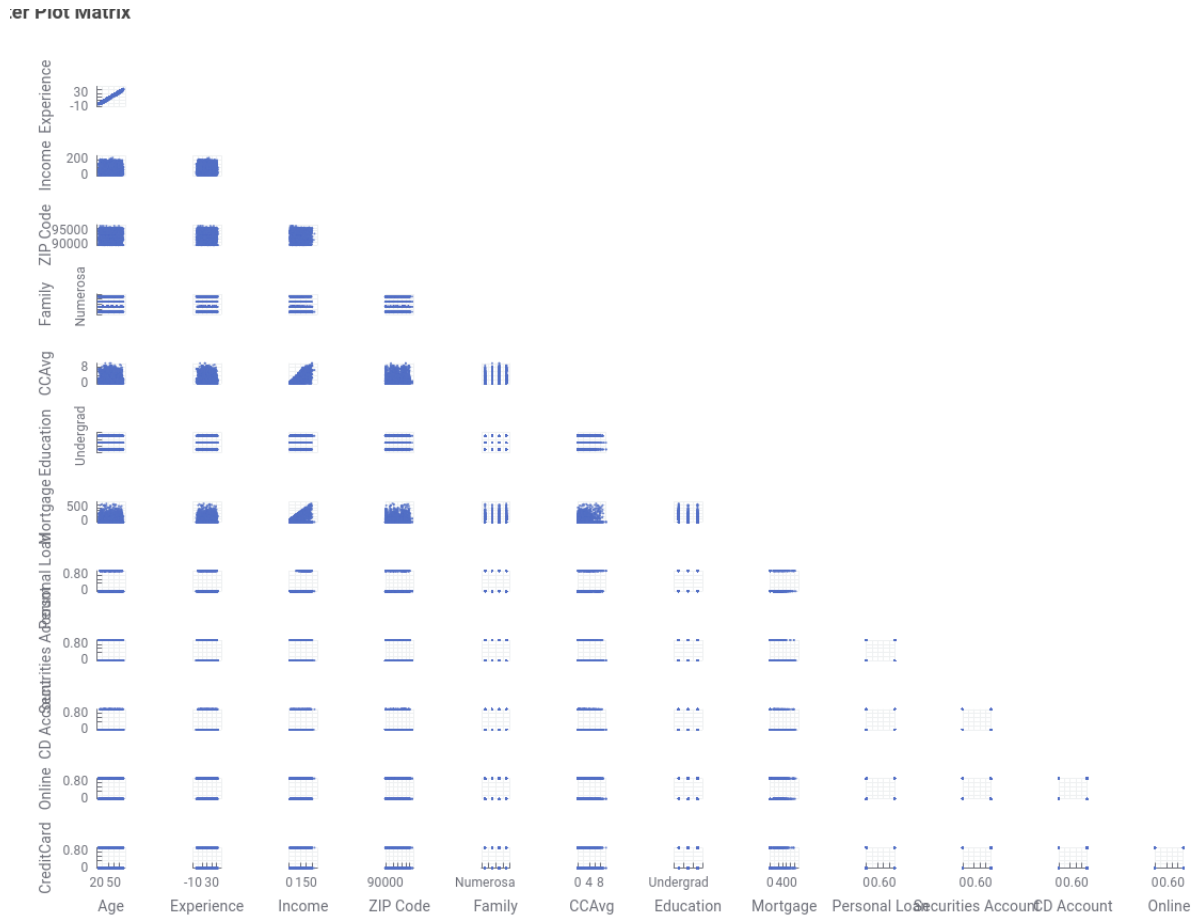


Figura 11: Matriz de dispersión.

En la imagen se puede ver que los campos credict card, online, CD account, Securities Account y personal loan podría ser algunos de los que podríamos quitar.

Para poder confirmar qué campos no serían adecuados, nos vamos a fijar también en la correlación que hay entre los campos, para ello hacemos el mismo proceso que el apartado anterior, pero con todos los campos y vemos cuáles son los que tienen más correlación entre ellos:

Row...	First column name *String*	Second column name *String*	Correlation value ↓ *Number (double)*
Row0	Age	Experience	0.994
Row20	Income	CCAvg	0.646
Row22	Income	Personal Loan	0.502
Row36	CCAvg	Personal Loan	0.367
Row50	Securities Account	CD Account	0.317
Row47	Personal Loan	CD Account	0.316
Row54	CD Account	CreditCard	0.279
Row21	Income	Mortgage	0.207
Row53	CD Account	Online	0.176

Figura 12: Mejores correlaciones.

Aquí vemos que destacan edad/experiencia e ingresos/CCAvg, Ingresos/Préstamo personal con una correlación superior a 0,5.

Confirmamos de esta manera que los atributos mencionados anteriormente, podría suprimirse para este estudio, ya que no existe una correlación buena entre los datos.

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?

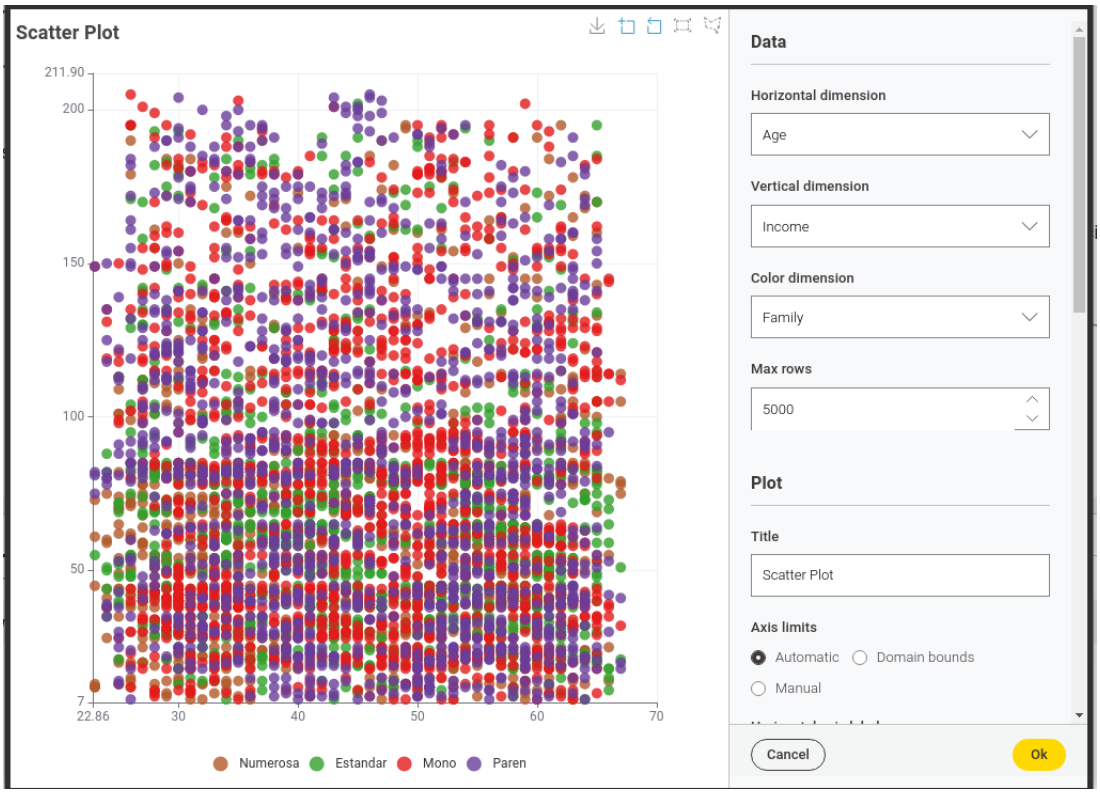


Figura 13: Diagrama dispersión familias respecto Edad/Ingresos.

A partir del diagrama anterior se pueden apreciar que las familias numerosas(4) y estándar(3) suelen aparecer cuando el salario es por debajo de los 100 y la edad oscila entre los 30 y 60. Al igual que las familias monoparentales(1), la edad oscila en ese rango, pero se dan también en más casos en los que el salario es superior a los otros dos grupos. El último grupo, parental(1), está repartido de forma más homogénea respecto a la edad y salario.

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?

Filtramos por el grupo de familia numerosa(4) y mostramos diagramas de dispersión en matrices y luego por experiencia/salario y salario/edad:

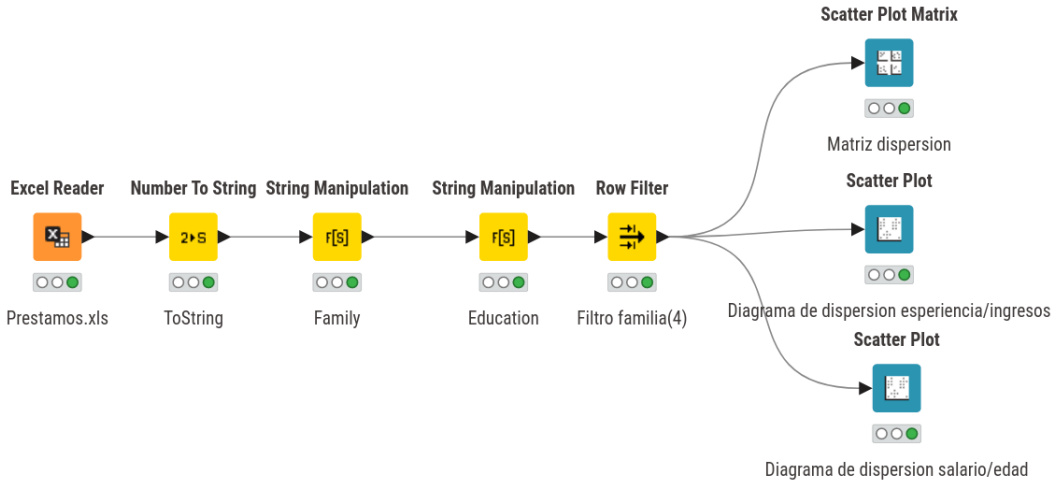


Figura 14: Configuración filtrado familia grupo 4.

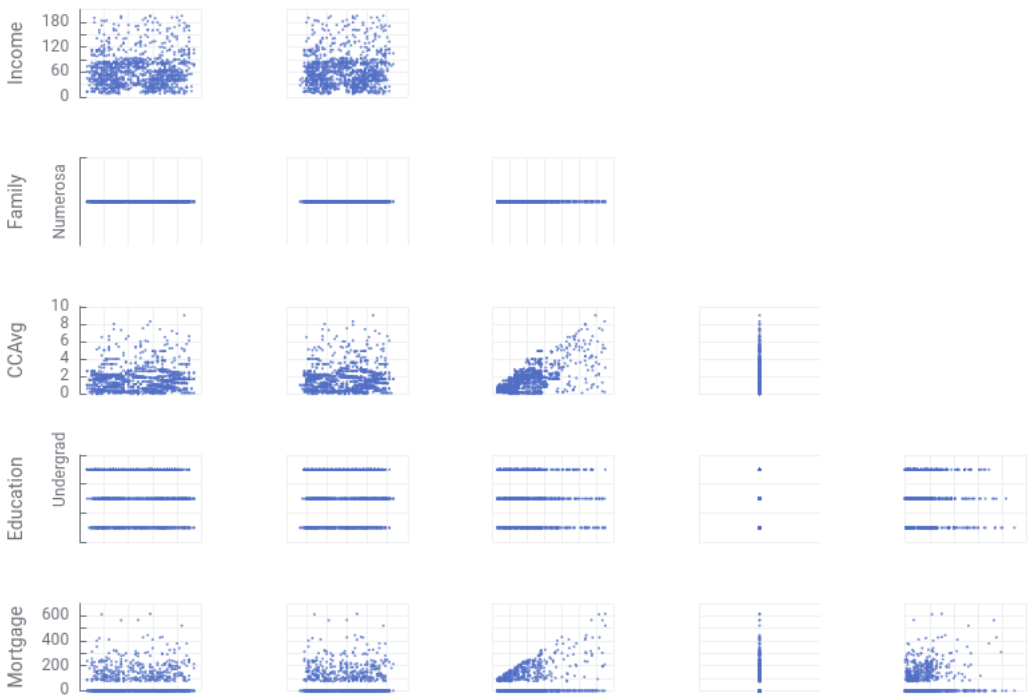


Figura 15: Matriz familia grupo 4.

Figura 16: Salario/edad familia grupo 4.

A partir de las configuraciones y gráficos mostrados previamente, se puede estimar que, para el grupo de familias numerosas(4), presenta una concentración de casi todos los datos cuando están cercanos a un salario por debajo de 100 y la edad oscila entre los 30 y 60. Dentro de este rango se da la mayor concentración de este subgrupo de familias.

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

Los grupos familiares Mono y Paren representan más del 55 % del total (29.71 % + 26.01 %). Estos dos grupos podrían ser vistos como los más "prometedores"para recibir una hipoteca, ya que representan una parte significativa de la población.

La diversidad de tipos de familias es importante, ya que cada una tiene necesidades y circunstancias diferentes. Los grupos Numerosa y Estándar también deben ser tenidos en cuenta, ya que representan alrededor del 44 % del total. Aunque su participación es menor, aún hay un número sustancial de familias en estas categorías.

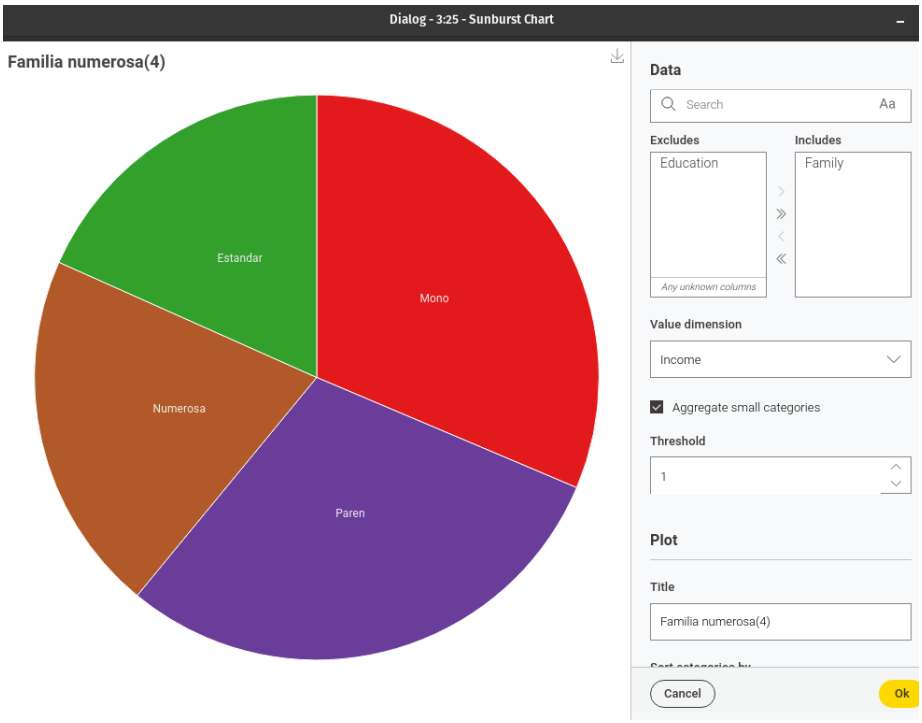


Figura 17: Ingresos por familia grupo 4.

2. Ejercicio 2 Fichero NBA.xlsx 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.

Se renombran todas las columnas que tienen nombres largos, en la captura se enseñan algunas de las renombradas:

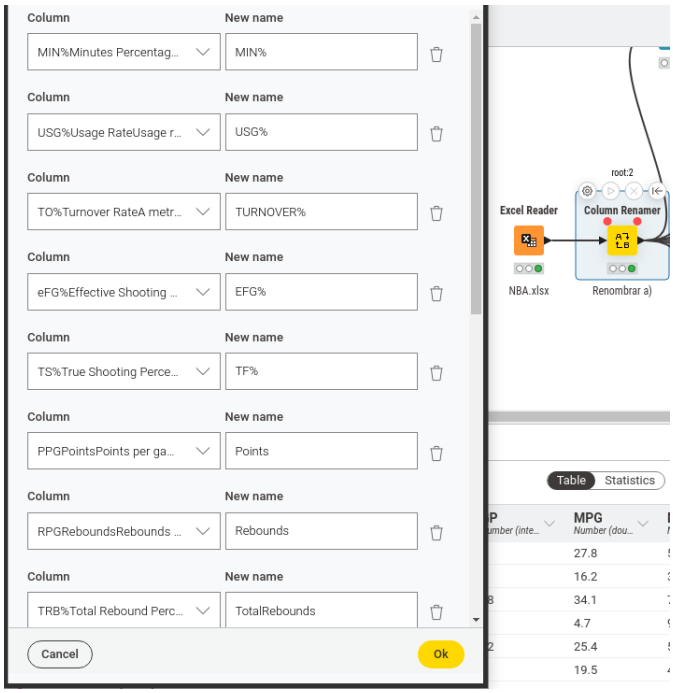
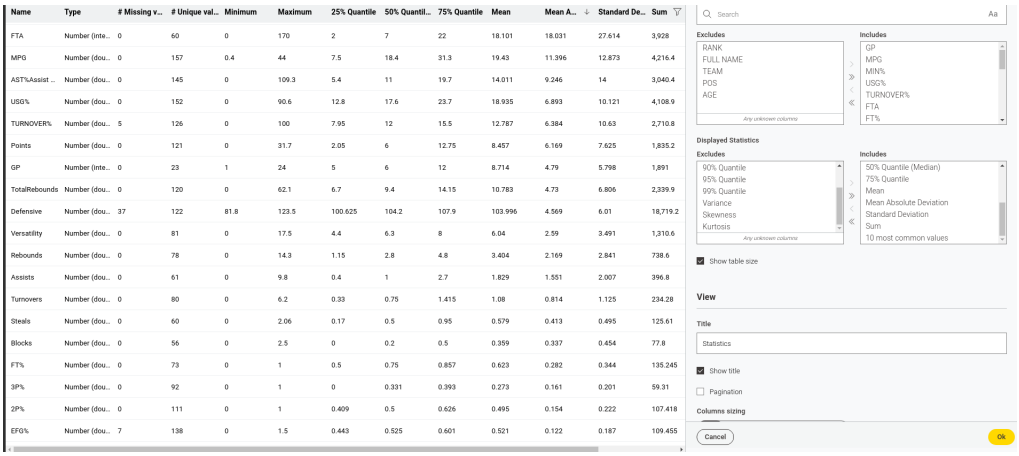


Figura 18: Cabeceras columnas NBA renombradas.

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

A partir de la imagen siguiente se puede ver que a la hora de hacer los cálculos, por ejemplo, la media respecto a la desviación media absoluta en algunas filas como el campo 'offensive' que tiene una media de 117.508, pero su desviación media absoluta es de 21.029, esto indica que al ser bastante más baja que la media, los datos están muy esparcidos. Principalmente, se debe a que se están haciendo los cálculos sobre todos los tipos de jugadores, sin importar la posición en la que se encuentre, por ende, existe tanta dispersión. No es lo mismo los datos defensivos de un defensa, que un atacante, de igual modo, los puntos que haga un pivote a los puntos de un defensor.

En resumen, deberían de calcularse las medias por equipo y por tipo de posición de los jugadores para tener mejores estimaciones.



Name	Type	# Missing v...	# Unique val...	Minimum	Maximum	25% Quantile	50% Quantil...	75% Quantile	Mean	Mean A...	Standard De...	Sum
FTA	Number (inte...	0	60	0	170	2	7	22	18.101	18.031	27.614	3,928
MPG	Number (dou...	0	157	0.4	44	7.5	18.4	31.3	19.43	11.396	12.873	4,216.4
AST%Assist ...	Number (dou...	0	145	0	109.3	5.4	11	19.7	14.011	9.246	14	3,040.4
USG%	Number (dou...	0	152	0	90.6	12.8	17.6	23.7	18.935	6.893	10.121	4,108.9
TURNOVER%	Number (dou...	5	126	0	100	7.95	12	15.5	12.787	6.384	10.63	2,710.8
Points	Number (dou...	0	121	0	31.7	2.05	6	12.75	8.457	6.169	7.625	1,835.2
GP	Number (inte...	0	23	1	24	5	6	12	8.714	4.79	5.798	1,891
TotalRebounds	Number (dou...	0	120	0	62.1	6.7	9.4	14.15	10.783	4.73	6.806	2,339.9
Defensive	Number (dou...	37	122	81.8	123.5	100.625	104.2	107.9	103.996	4.569	6.01	18,719.2
Versatility	Number (dou...	0	81	0	17.5	4.4	6.3	8	6.04	2.59	3.491	1,310.6
Rebounds	Number (dou...	0	78	0	14.3	1.15	2.8	4.8	3.404	2.169	2.841	738.6
Assists	Number (dou...	0	61	0	9.8	0.4	1	2.7	1.829	1.551	2.007	396.8
Turnovers	Number (dou...	0	80	0	6.2	0.33	0.75	1.415	1.08	0.814	1.125	234.28
Steals	Number (dou...	0	60	0	2.06	0.17	0.5	0.95	0.579	0.413	0.495	125.61
Blocks	Number (dou...	0	56	0	2.5	0	0.2	0.5	0.359	0.337	0.454	77.8
FT%	Number (dou...	0	73	0	1	0.5	0.75	0.857	0.623	0.282	0.344	135.245
3P%	Number (dou...	0	92	0	1	0	0.331	0.393	0.273	0.161	0.201	59.31
2P%	Number (dou...	0	111	0	1	0.409	0.5	0.626	0.495	0.154	0.222	107.418
EFG%	Number (dou...	7	138	0	1.5	0.443	0.525	0.601	0.521	0.122	0.187	109.455

Figura 19: Medías básicas.

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?

Tras usar la matriz de diagrama de dispersión, se pueden apreciar bastantes que tienen relación entre sí, las más destacables han sido las siguientes:

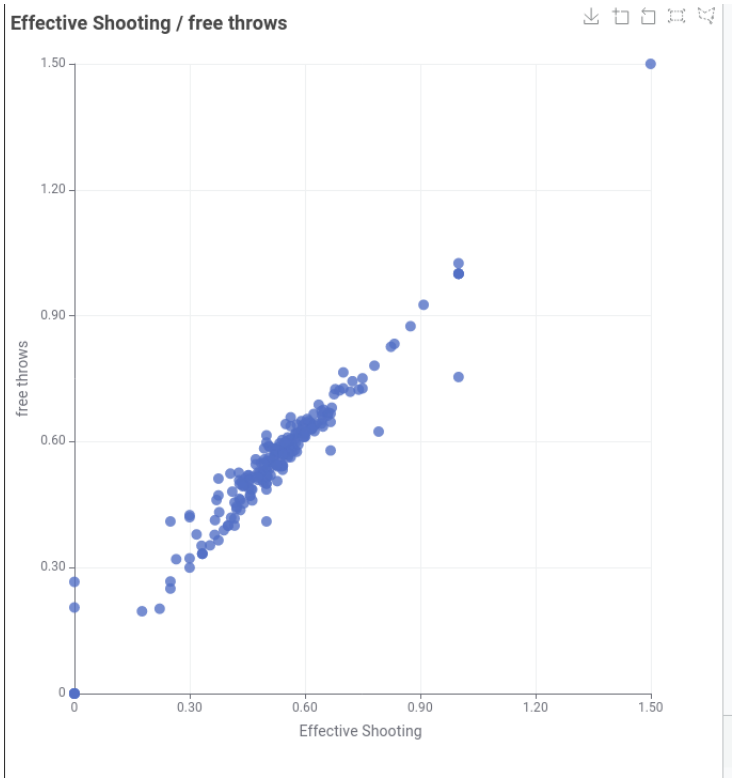


Figura 20: Effective shooting and free throws.

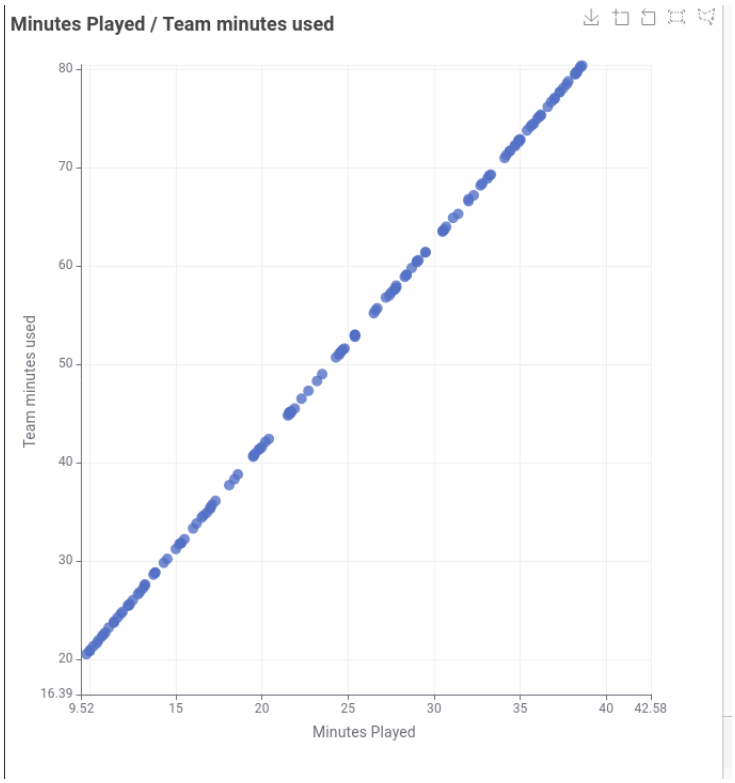


Figura 21: Minutes playes and team minutes used.

En estos se puede apreciar que es casi linear la relación entre ambos valores, por ejemplo en el caso de la primera imagen, se puede ver que a medida que el número de lanzamientos por parte del jugador es mayor, tiende a tener más puntuación, es decir, a mayor número de tiros tenga el jugador, mayor cantidad de puntos libres tendrá.

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

Row...	First column name String	Second column name String	Correlation value ↓ Number (double)
Row20	MPG	MIN%	1
Row...	EFG%	TF%	0.97
Row46	MIN%	Points	0.88
Row28	MPG	Points	0.88
Row...	Points	Turnovers	0.814
Row...	Assists	Turnovers	0.801
Row...	Points	Assists	0.77
Row31	MPG	Assists	0.755
Row49	MIN%	Assists	0.754
Row35	MPG	Turnovers	0.743
Row53	MIN%	Turnovers	0.743
Row29	MPG	Rebounds	0.73
Row47	MIN%	Rebounds	0.73
Row...	Points	Rebounds	0.716
Row...	AST%Assist PercentageAssist percentage i...	Versatility	0.697
Row33	MPG	Steals	0.694
Row51	MIN%	Steals	0.694
Row...	Assists	Steals	0.675

Figura 22: Mayores correlaciones.

Tras lo mencionado en el apartado anterior y lo visto con la correlación, se puede ver que si había relación entre el número de tiros libres realizados y la efectividad de tiro con un 0.97, siendo casi perfecta la correlación.

Por otro lado, existe una correlación perfecta de 1 entre el porcentaje de los minutos de equipo utilizados por un jugador mientras estaba en pista (MIN) y la media de minutos que un jugador ha jugado por partido (MPG).

Por otro lado, también hay una correlación bastante alta entre 'turnover' con puntos y asistencia. En el caso de pérdida de balón y puntos, quiere decir que a medida que un equipo comete más errores en la posesión de la pelota, también tiene una mayor probabilidad de anotar puntos en ese mismo período de tiempo.

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

Se ha optado por hacer un gráfico con la media de los jugadores en los campos EFG y TF por posición, si se hacía un gráfico de barras con muchos atributos saldrían demasiados campos por cada tipo posición, en este caso con solo dos por 7 tipos de posiciones, salen 14 barras. Por otro lado, se hace la media porque son muchos jugadores y saldría una gráfica enorme y sería muy complicado poder evaluarlo.

Figura 23: Diagrama de barras EFG, TF por POS .

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

Se realiza un diagrama de cajas con las asistencias por posición, vamos a analizar este diagrama teniendo en cuenta la posición FG y GF:

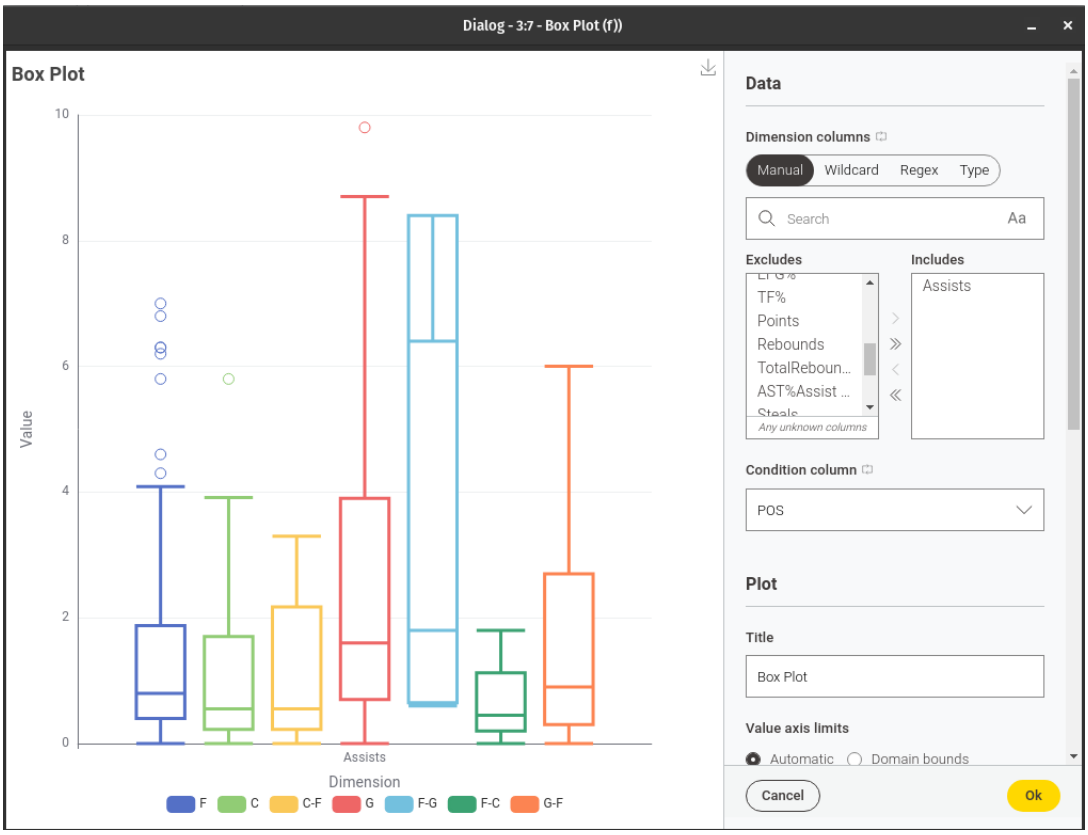


Figura 24: Diagrama de cajas.

Dos posiciones de baloncesto, FG (tirador) y FC (ala-pívot), en términos de asistencias. Estos datos se pueden analizar utilizando un gráfico de caja (box plot) para comprender mejor la distribución y la variabilidad de los valores. Aquí hay una interpretación de estos datos:

Posición FG (Tirador):

- Máximo (6.4): Esto indica que en el conjunto de datos, el tirador FG tuvo un máximo de 6.4 asistencias en un juego.
- Tercer Cuartil (Q3, 75 % de las asistencias por debajo, 8.4): El 75 % de las asistencias de FG están por debajo de 8.4, lo que sugiere que la mayoría de las veces registra un número relativamente alto de asistencias.
- Mediana (1.8): La mediana es el valor medio en el conjunto de datos, lo que significa que el 50 % de las asistencias de FG están por debajo de 1.8 y el 50 % están por encima.

- Primer Cuartil (Q1, 25 % de las asistencias por debajo, 0.64): El 25 % de las asistencias de FG están por debajo de 0.64, lo que sugiere que el 25 % más bajo de sus registros de asistencias está en este rango.

Posición FC (Ala-Pívot):

- Máximo (6): En el conjunto de datos, el ala-pívot FC tuvo un máximo de 6 asistencias en un juego.
- Tercer Cuartil (Q3, 75 % de las asistencias por debajo, 2.7): El 75 % de las asistencias de FC están por debajo de 2.7, lo que sugiere que la mayoría de las veces registra un número relativamente alto de asistencias, aunque menor en comparación con FG.
- Mediana (0.9): La mediana es el valor medio en el conjunto de datos, lo que significa que el 50 % de las asistencias de FC están por debajo de 0.9 y el 50 % están por encima.
- Primer Cuartil (Q1, 25 % de las asistencias por debajo, 0.3): El 25 % de las asistencias de FC están por debajo de 0.3, lo que sugiere que el 25 % más bajo de sus registros de asistencias está en este rango.

Comparación:

- En términos de la mediana, el tirador FG (1.8) tiene una mediana significativamente más alta que el ala-pívot FC (0.9), lo que sugiere que, en promedio, FG registra un mayor número de asistencias.
- El rango de asistencias de FG es más amplio, con un valor máximo de 6.4, en comparación con el ala-pívot FC, que tiene un valor máximo de 6.
- Ambos jugadores tienen la mayoría de sus asistencias en los cuartiles superiores (Q3), lo que indica que ambos son efectivos en proporcionar asistencias.

En resumen, el tirador FG tiende a registrar un mayor número de asistencias en promedio y tiene un rango más amplio de valores en comparación con el ala-pívot FC, aunque ambos jugadores son efectivos en proporcionar asistencias según los valores de Q3.

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

Realizamos el diagrama circular y lo filtramos por partidos jugados por equipos (GP):

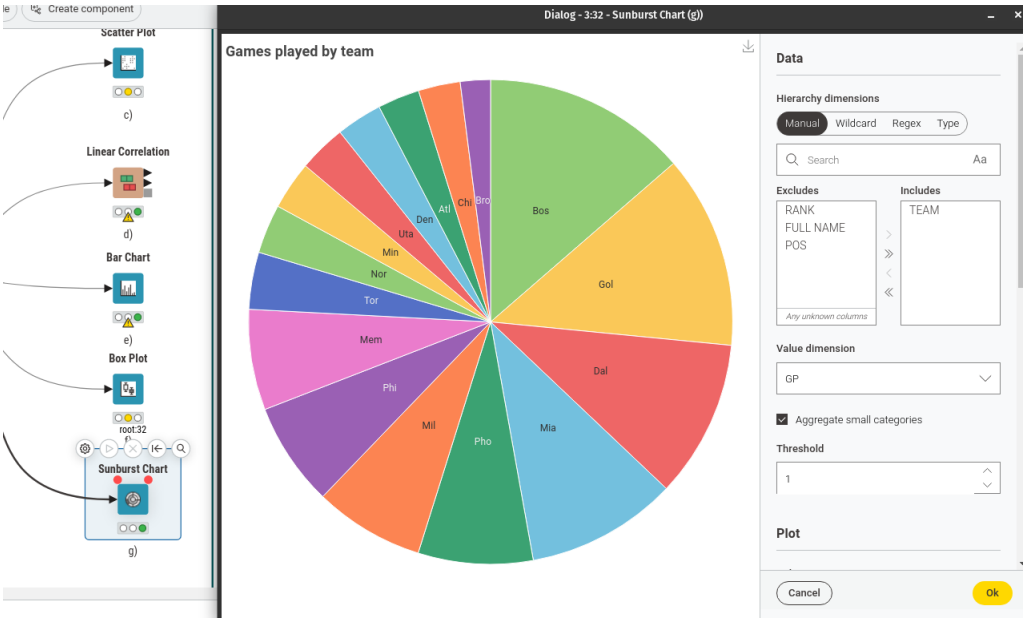


Figura 25: Diagrama circular partidos jugados por equipos.

A partir de esto se pueden deducir distintas posibilidades:

a) **Distribución de Equipos:** El hecho de que 10 de los 16 equipos ocupen el 37.81 % del total sugiere que estos equipos tienen una participación significativa en el conjunto, pero no mucha, ya que han jugado muchos menos partidos comparados con los 4 equipos que más han jugado.

b) **Importancia del Equipo BOS:** El equipo BOS, con un porcentaje del 13.64 %, tiene la mayor participación entre todos los equipos. Esto sugiere que el equipo BOS puede ser el más exitoso o el más seguido por los aficionados.

c) **Participación Relativa:** Se nota la diferencia en la participación entre los equipos. Los equipos que no ocupan un gran porcentaje del círculo tienen una representación relativamente baja en comparación con los equipos más destacados.

d) **Enfoque en los Equipos Principales:** Dado que 10 de los 16 equipos ocupan el 37.81 %, es posible que la atención, la inversión o el interés se concentren en estos equipos, ya sea debido a su historial de victorias, popularidad, ubicación o algún otro factor.

La distribución es desigual entre los equipos de baloncesto, con un grupo selecto que tiene una representación significativamente mayor. El equipo BOS se destaca como el líder en términos de porcentaje, lo que podría atribuirse a su éxito en el campo o a su popularidad entre los seguidores del baloncesto.

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 las líneas 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.

Vamos a realizar un filtrado de los datos con el equipo con más partidos jugados (visto en el anterior apartado BOS), hacemos la gráfica con 5 atributos: puntos, rebotes, asistencias, robos y bloqueos.

En la siguiente imagen se aprecia mucha variación, no es uniforme, hay bastantes jugadores que tienen valores casi nulos o por debajo de 1-2, puede deberse a distintas razones, la más probable es que sean reservas y no hayan participado en ningún partido.

Por otro lado, destaca mucho que el resto de jugadores tienen altas puntuaciones en rebotes suelen tener muchos puntos en bloqueos, por otro lado, los jugadores que tienen muchas asistencias, suelen tener también bastantes puntos.

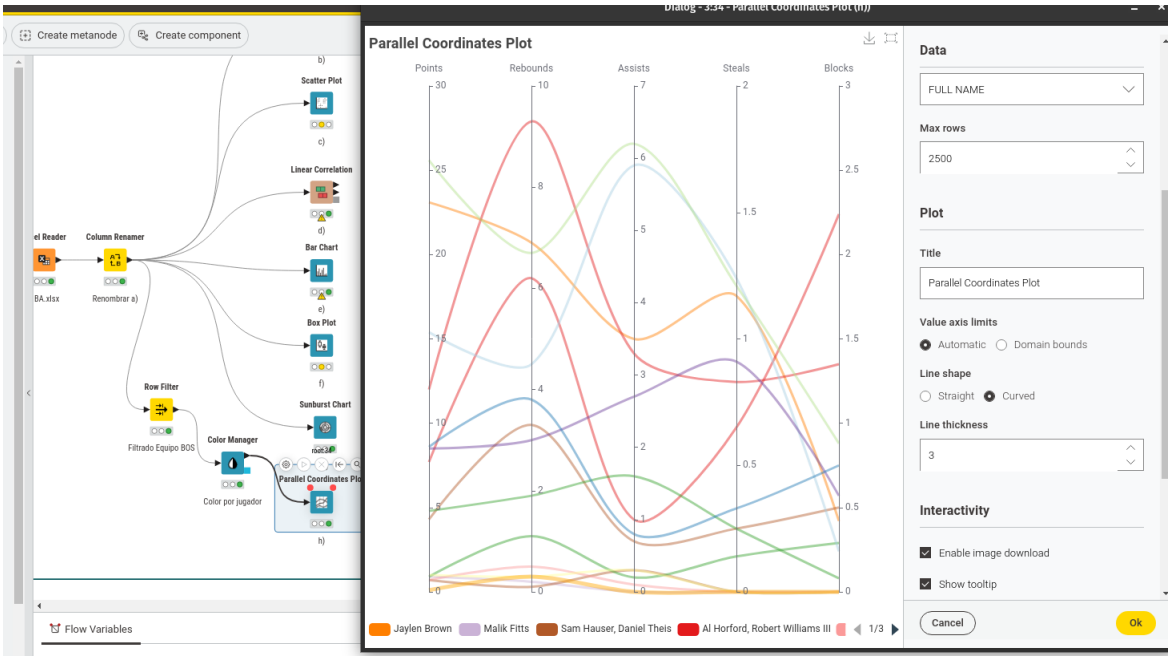


Figura 26: Diagrama de coordenadas paralelas del equipo BOS.

En esta imagen, aislamos a un jugador, en este caso uno con las mejores puntuaciones en casi todos los aspectos evaluados, excepto en bloqueos, llegando a tener un pico de asistencias por encima de 6, más de 25 puntos y cerca de 1.25 robos. Este jugador es bastante completo y destaca en casi todos los valores.

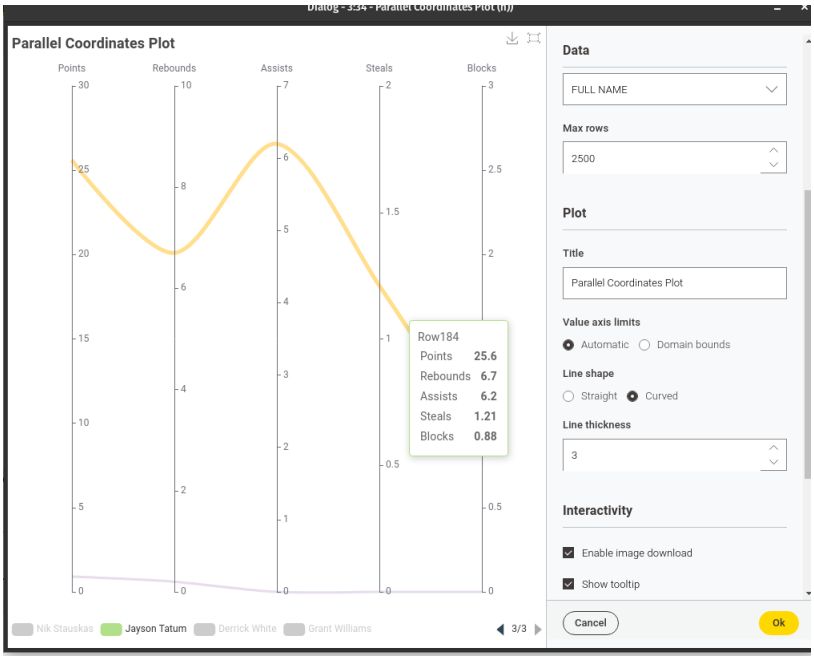


Figura 27: Jugador Jayson Tatum.

Por otro lado, al contrario que el otro jugador evaluado, este presenta valores prácticamente a 0, destacando solo que tiene 0.9 puntos. Esto puede deberse a que no juega en los partidos, es suplente, haya estado lesionado o por otra causa.

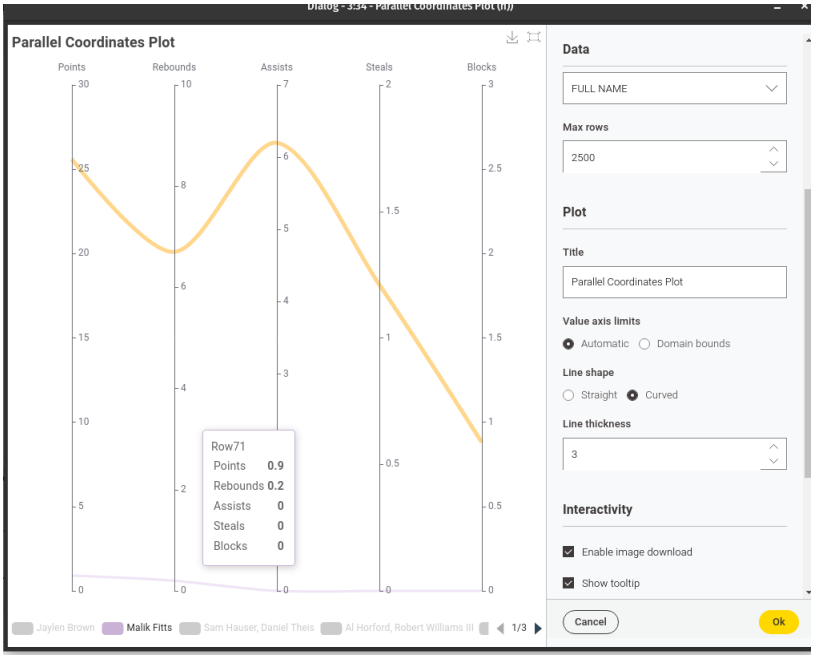


Figura 28: Jugador Malik Fitts.

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?

Tras lo visto en la gráfica anterior del equipo BOS, habría que filtrar los datos por jugadores que han tenido un tiempo de juego considerable, ya que si evaluamos todos los jugadores de un equipo por puntos y por ejemplo de 12 jugadores, sólo juegan 6, estos datos a evaluar no son realistas, puesto que se harían operaciones sobre 12 jugadores y no sobre los 6 que realmente son los jugadores activos. Por otro lado, se tendrían que hacer en grupos de posición o por equipos para tener datos fiables, ya que no se puede evaluar de igual manera los datos de un pivot del mejor equipo, que los datos de un pivot del peor equipo, tendría que ser evaluados por equipo y posición.