



UNIVERSIDAD DE GRANADA

Práctica 1: Visualización con Knime

Tratamiento Inteligente de Datos

Máster Universitario en Ingeniería Informática

Autor: Pablo Valenzuela Álvarez (pvalenzuela@correo.ugr.es)

Índice

Tarea 1	2
a).....	2
b)	2
c).....	3
d)	4
e).....	5
f)	7
g).....	8
h)	9
 Tarea 2	 11
a).....	11
b)	11
c).....	13
d)	14
e).....	14
f)	14
g).....	16
h)	16
i).....	17

Tarea 1

Cargando *prestamo.xls* con el nodo “Excel Reader” no salían todos los datos. Hay que ajustar la fila donde empiezan los valores ajustando la opción *By position* a 1, porque en este fichero la primera fila se usa para ordenar.

a) Hay varias columnas con valores 0 y 1 que corresponden a los valores true o false. Como dice el enunciado, el lector los toma como enteros cuando son valores categóricos.



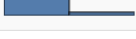
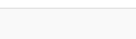



Cambiamos las columnas: “Family”, “education”, “personal loan”, “securities account”, “cd account”, “online” y “credit card”.

b) El valor mínimo de experiencia es -3. Supongo que esta característica se corresponde a los años trabajados, por lo que no tiene sentido que empiece con ese valor.

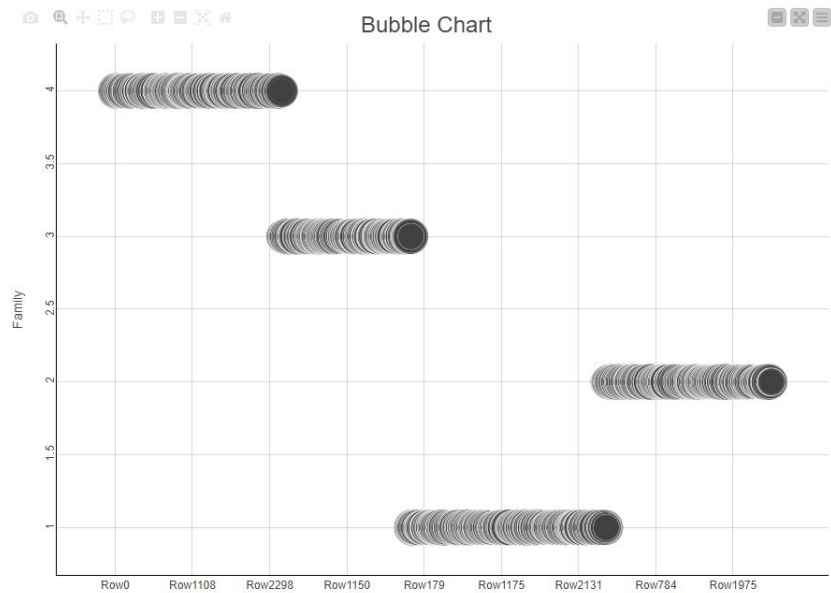
Según parece, no ofrecen préstamos a gente con más de 67 años.

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness
+ Age	<input type="checkbox"/>	23	67	45.338	11.463	131.404	-0.029
+ Experience	<input type="checkbox"/>	-3	43	20.105	11.468	131.514	-0.026
+ Income	<input type="checkbox"/>	8	224	73.774	46.034	2119.104	0.841
+ ZIP Code	<input type="checkbox"/>	90005	96651	93169.703	1759.714	3096594.472	-0.296
+ CCAvg	<input type="checkbox"/>	0	10	1.938	1.748	3.054	1.598
+ Mortgage	<input type="checkbox"/>	0	635	56.499	101.714	10345.698	2.104

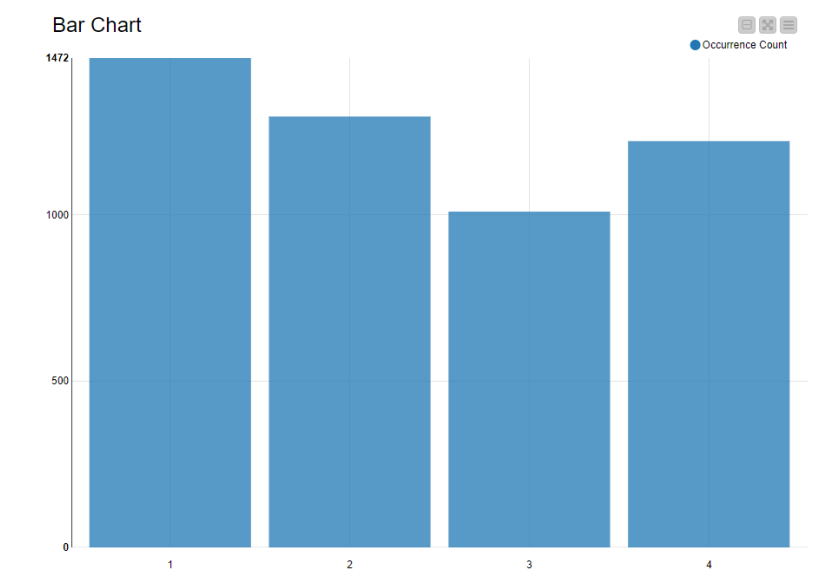
Sobre los datos nominales, quitando los que tienen solo dos valores, se puede ver que familia tiene cuatro categorías y educación tres.

Column	Exclude Column	No. missings	Unique values	All nominal values	Frequency Bar Chart
Family	<input type="checkbox"/>	0	4	1, 2, 4, 3	
Education	<input type="checkbox"/>	0	3	1, 3, 2	
Personal Loan	<input type="checkbox"/>	0	2	0, 1	
Securities Account	<input type="checkbox"/>	0	2	0, 1	
CD Account	<input type="checkbox"/>	0	2	0, 1	
Online	<input type="checkbox"/>	0	2	1, 0	
CreditCard	<input type="checkbox"/>	0	2	0, 1	

c) He agrupado el resultado por la columna “Family” para que se vea bien en la tabla de burbujas y no esté tan disperso.

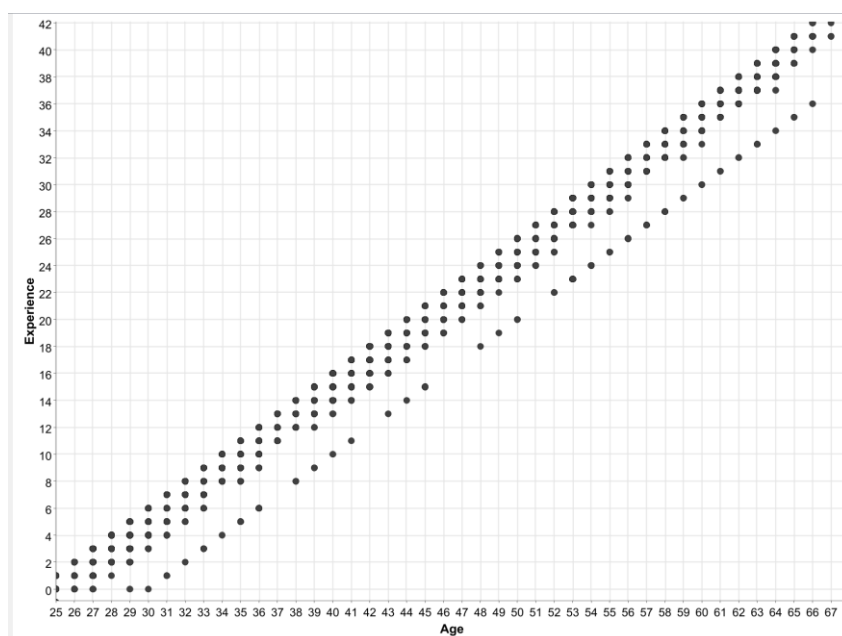


El diagrama de barras muestra la distribución la gente sobre las distintas categorías de la columna “Family”.



La categoría 1 es la que más integrantes tiene y la 3 la que menos.

d) La gráfica de dispersión de las columnas “Age” y “Experience” muestra la relación directa que hay entre estos dos atributos.



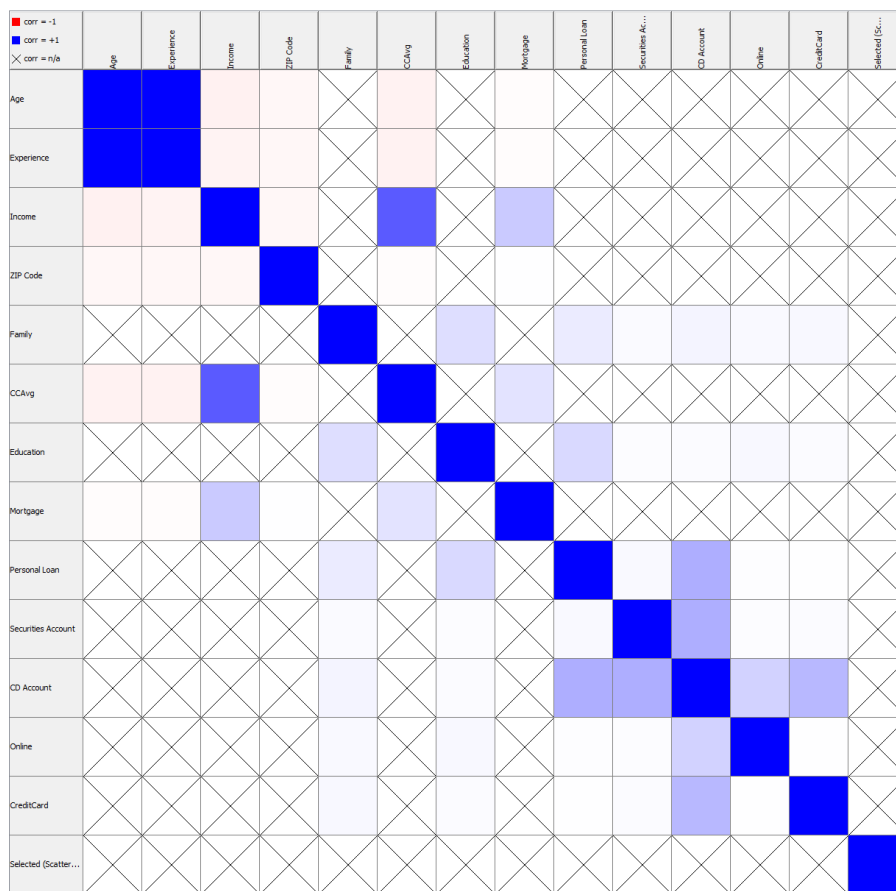
Suponiendo que “Experience” sean los años trabajados, es normal que suban respecto a la edad. Se puede observar perfectamente en la gráfica anterior.

Y para confirmarlo, ejecutamos el nodo “Linear Correlation” en Knime.

Row ID	D Age	D Experience
Age	1.0	0.9942148569683138
Experience	0.9942148569683138	1.0

Lo valores para esta interpretación son [-1,1] siendo 1 la relación directa entre dos atributos. La imagen anterior muestra un resultado casi igual a 1 por lo que podríamos confirma que tienen una relación directa.

e) Estudiando la matriz de la siguiente imagen podemos establecer si puede haber más relaciones entre las variables de la tabla.



Ya sabemos que hay una relación entre “Age” y “Experience”, pero parece haber otra entre “CAvg” e “Income”. Y otras entre “CD Account” - “Personal loan” - “Security Account”, y “CD Account” - “Online” - “CreditCard”.

Coefficiente entre “CAvg” e “Income”.

Row ID	D Income	D CCAvg
Income	1.0	0.645983669624...
CAvg	0.64598366...	1.0

Coeficiente entre “CD Account” - “Personal loan” - “Security Account”.

Row ID	D Person...	D Securities Account	D CD Account
Personal Loan	1.0	0.021953882216311...	0.31635482941440...
Securities Acc...	0.02195388...	1.0	0.3170344156806614
CD Account	0.31635482...	0.3170344156806614	1.0

Hay casi el mismo coeficiente entre “CD Account” - “Personal loan” y “CD Account” - “Security Account”.

Coeficiente entre “CD Account” - “Online” - “CreditCard”.

Row ID	D CD Acc...	D Online	D CreditCard
CD Account	1.0	0.17588001598343...	0.2786443646147...
Online	0.17588001...	1.0	0.0042096561546...
CreditCard	0.27864436...	0.00420965615468...	1.0

Viendo los resultados, si tendría sentido eliminar algunas columnas.

Las relaciones directas son columnas que tienen los mismos valores por lo que podemos eliminar una a elegir entre “Age” o “Experience”.

El otro caso, en el que dos relaciones tengan el mismo resultado, puede que también convenga eliminar una de estas dos columnas: “Personal loan” o “Security Account”.

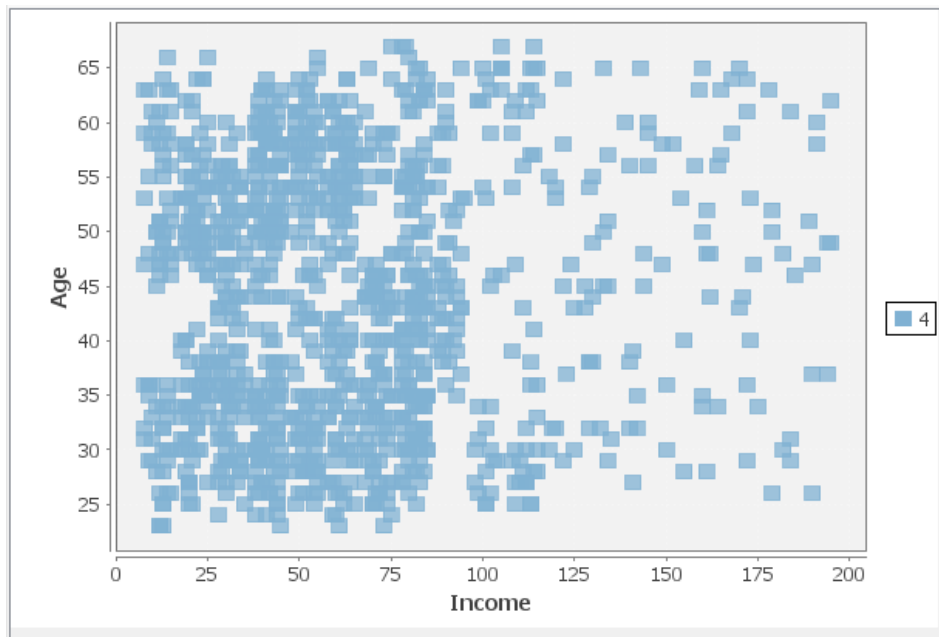
f) Según el gráfico, se puede observar que las familias tipo '3' y '4' se “acumulan” mayoritariamente en valores menores de 100 respecto al salario. Suponiendo que este tipo de familias sean las familias con hijos, son las que tienen menos ingresos respecto a las otras.



g) Se puede observar lo que comentaba en el punto anterior, la mayoría de los puntos se agrupan con los valores inferiores a 100 de salario sin importar la edad.

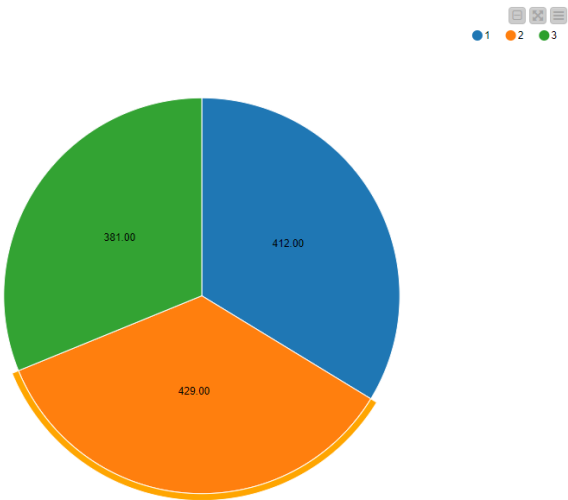
Pero ahora además se pueden apreciar dos pequeños grupos entre los siguientes valores:

- Income (100-125) y Age (25-35)
- Income (100-125) y Age (60-65)

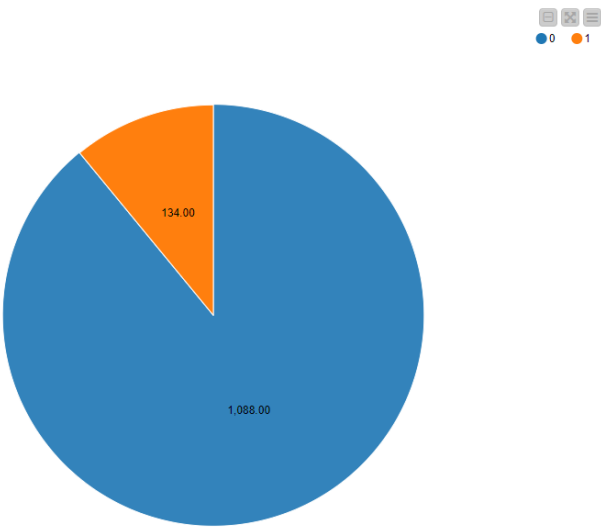


h) Diagramas circulares

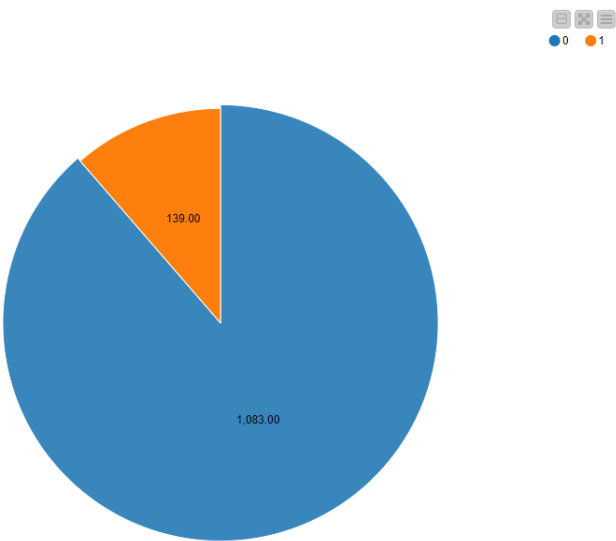
Education



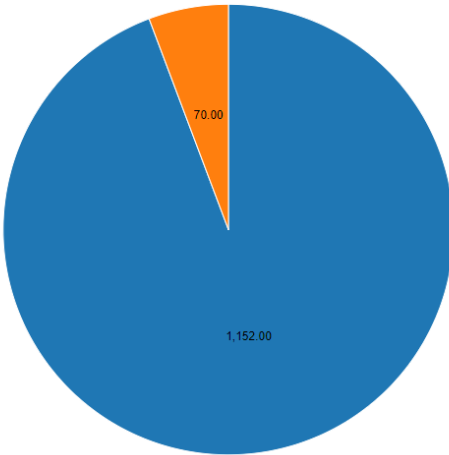
Personal loan



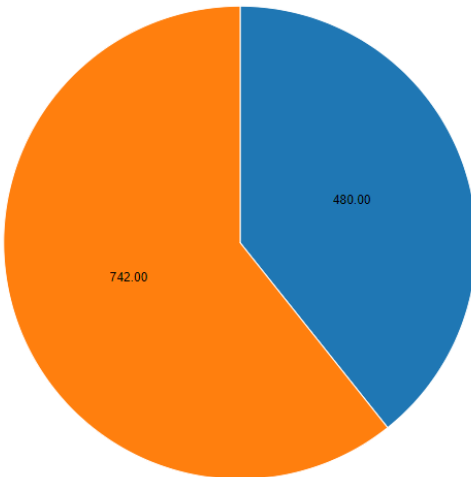
Security account



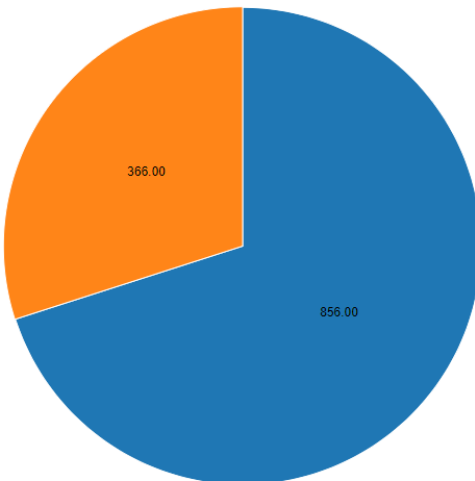
CD account



Online



CreditCard











Tarea 2

Al cargar el fichero me he dado cuenta de que la primera fila no tiene valores, y el lector de Excel no me dejaba saltármela. Por lo que he tenido que eliminarla a mano.

Después de poder cargar el conjunto, he filtrado la columna “RANK” dado a que no contiene valores.

a) Cambiado el nombre de las columnas a bastantes columnas. La imagen siguiente muestra algunos, no todos, los cambios.

Column	New name
MIN%Minutes Percenta... ▾	MIN% 
USG%Usage RateUsage ... ▾	USG% 
TO%Turnover RateA met... ▾	TO% 
eFG%Effective Shooting ... ▾	eFG% 
TS%True Shooting Perc... ▾	TS% 
PPGPointsPoints per ga... ▾	PPG 
RPGReboundsRebound... ▾	RPG 
TRB%Total Rebound Per... ▾	TRB% 

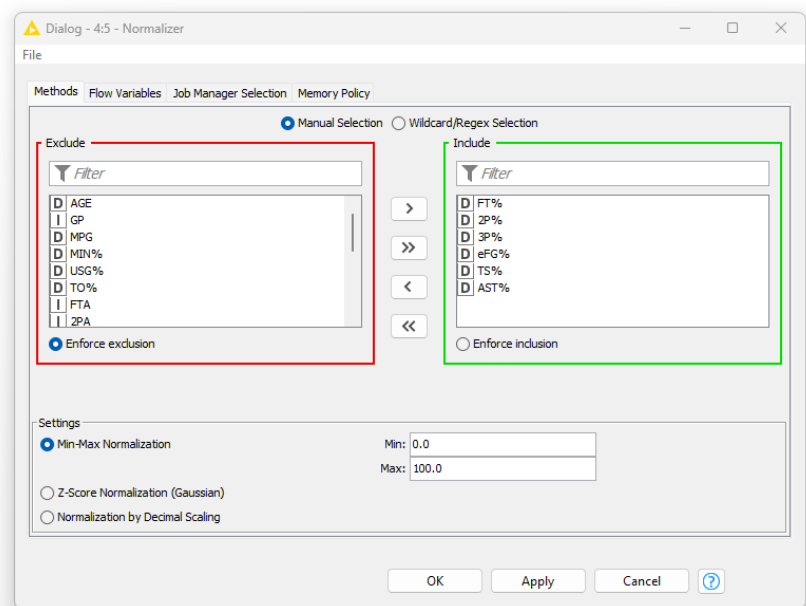
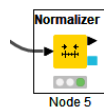
b) Los porcentajes de tiro libre (FT%), tiro de 2 (2P%) y de 3 (3P%) tienen valores entre 0 y 1, en vez de ir de 0 a 100 como otros.

Lo mismo pasa con los valores para tiro verdadero (TS%) y tiro efectivo (eFG%), que van de 0 a 1’5.

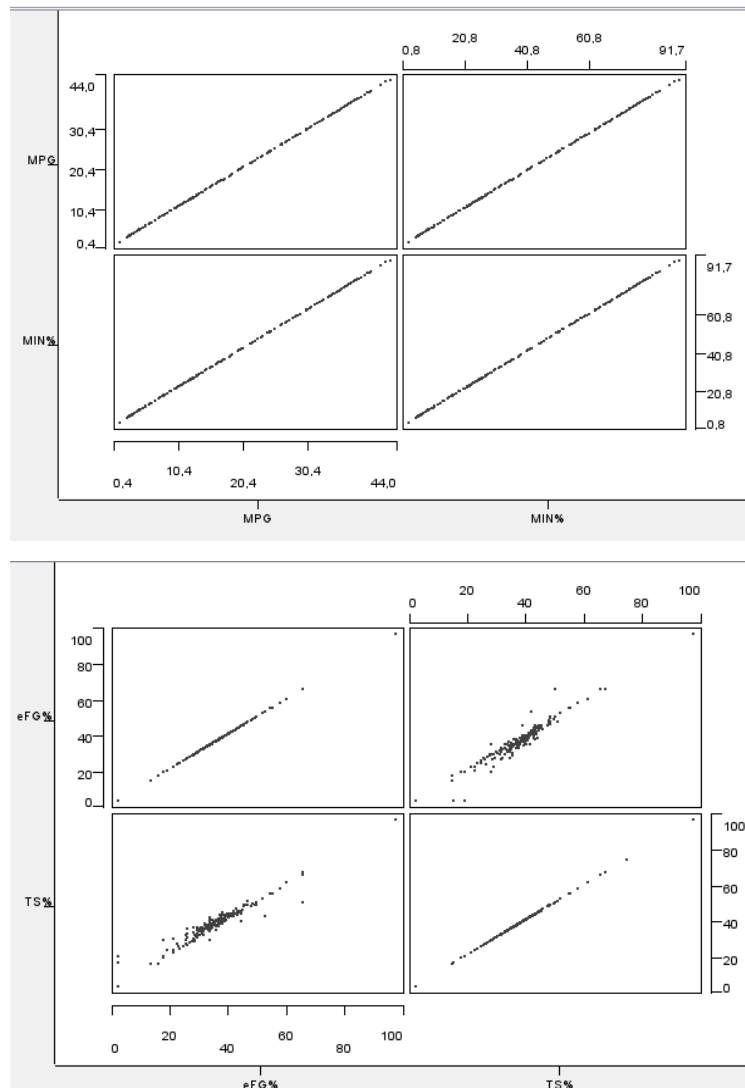
También el valor máximo del porcentaje de asistencias estimadas pasa del 100%, siendo 109'3.

FTA	Number (integer)	0	60	0	170
FT%	Number (double)	0	73	0	1
2PA	Number (integer)	0	90	0	276
2P%	Number (double)	0	111	0	1
3PA	Number (integer)	0	78	0	229
3P%	Number (double)	0	92	0	1
eFG%	Number (double)	7	138	0	1.5
TS%	Number (double)	6	155	0	1.5
PPG	Number (double)	0	121	0	31.7
RPG	Number (double)	0	78	0	14.3
TRB%	Number (double)	0	120	0	62.1
APG	Number (double)	0	61	0	9.8
AST%	Number (double)	0	145	0	109.3
SPG	Number (double)	0	60	0	2.06

Voy a normalizar estos campos entre 0 y 100 usando el nodo “Normalizer”.



c) Al observar las gráficas veo que puede haber una relación directa entre las variables “MPG” y “MIN” y puede que otra, aunque no está tan clara entre “TS%” y “eFG”.



La primera relación, como se puede ver en la imagen posterior, es una relación directa. Minutos por partido (MPG) muestra el mismo valor que la variable MIN%, es decir, los minutos de media que esta un jugador en pista.

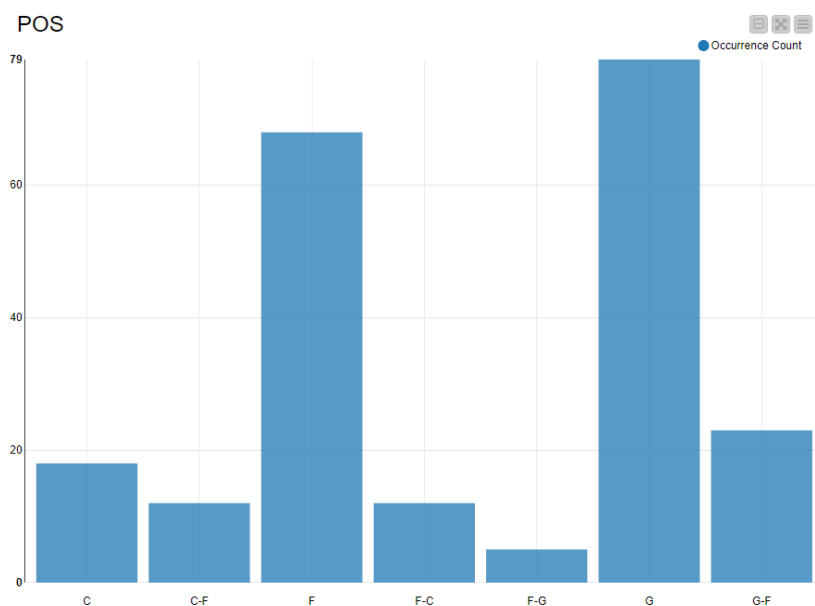
Row ID	D MPG	D MIN%
MPG	1.0	0.9999973041323...
MIN%	0.99999730...	1.0

La otra relación que comentaba es directa también según el coeficiente de correlación. Las dos se refieren a la eficiencia de los lanzamientos realizados por un jugador, por lo que tienen valores similares.

Row ID	D eFG%	D TS%
eFG%	1.0	0.9699092588110...
TS%	0.96990925...	1.0

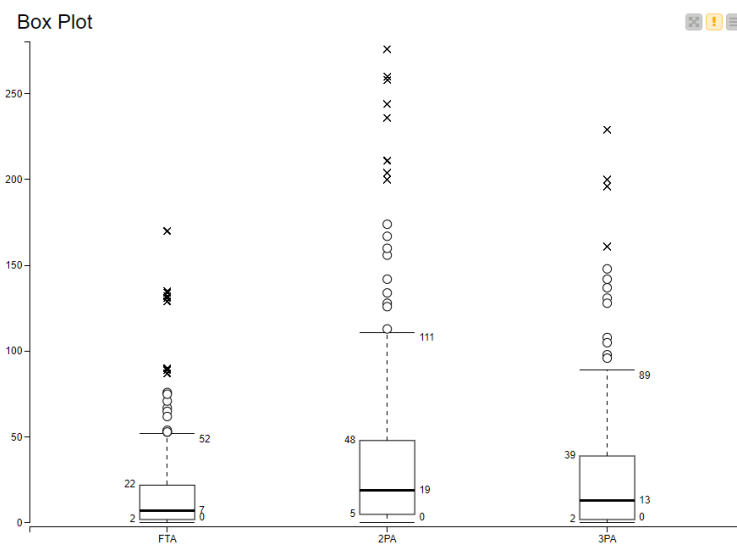
d) Comentado en el punto anterior.

e)



f) He detectado varias variables que contienen “outliers”.

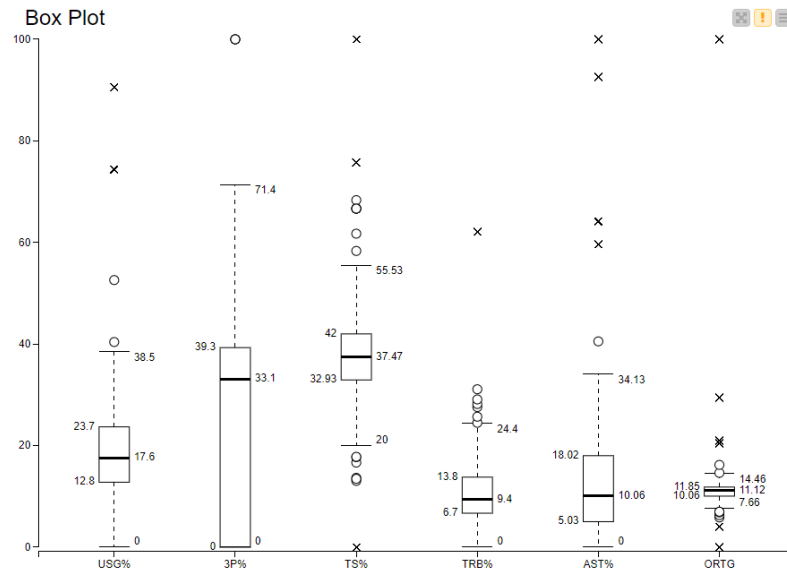
Entre ellas las que contienen los tiros de campo.



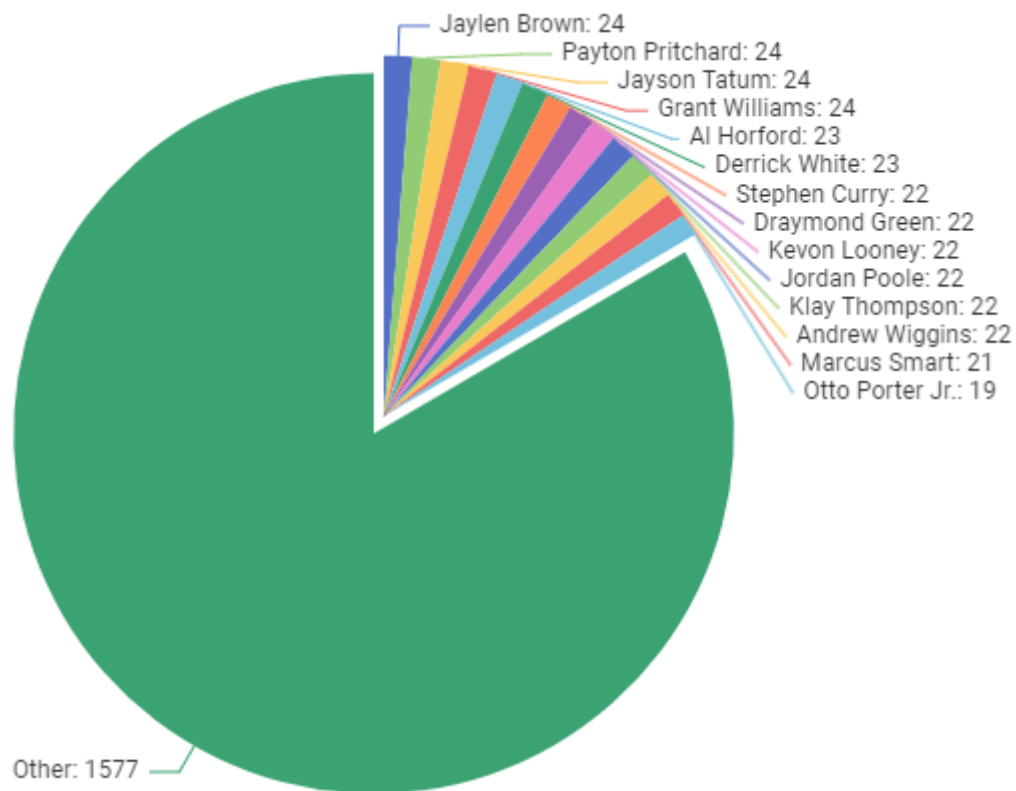
Como se puede observar en la imagen anterior, hay varios jugadores que lanzan más tiros de lo normal.

Entre los porcentajes destacan (imagen siguiente):

- Unos jugadores que juegan más minutos de lo normal. Puede ser por que haya prórrogas en sus partidos.
- Destaca un jugador con un 100% o cerca de acierto en tiros de 3. Puede ser porque solo haya tirado uno y lo hubiese encestado.
- Destacan por lo bien y lo mal que lanzan varios jugadores como se puede ver en la columna TS%.
- Hay también varios casos de jugadores que rebotean y asisten mejor.

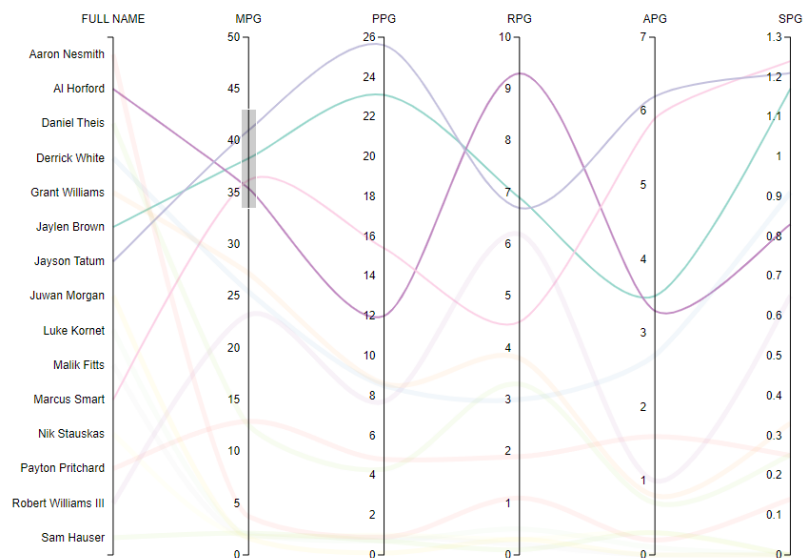


g) He agrupado los jugadores que sólo han jugado un partido para que sea más visible el gráfico.

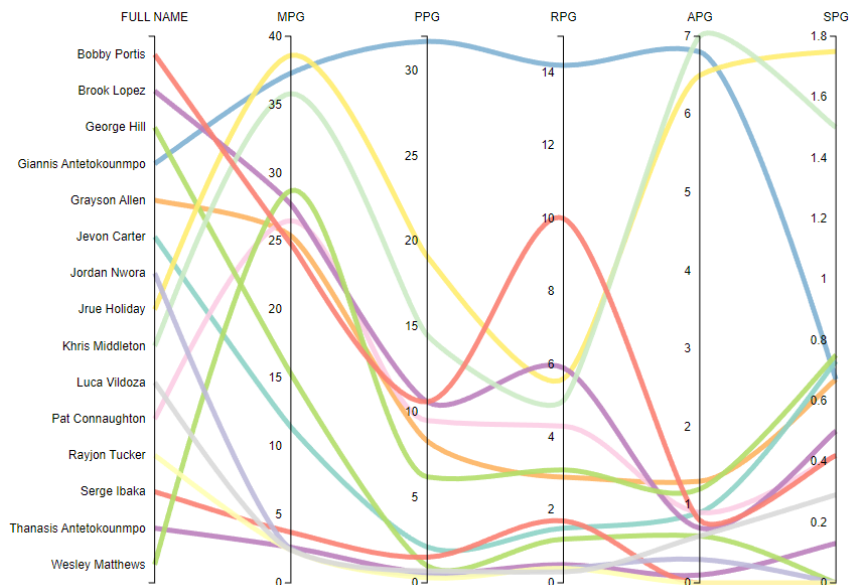


h) Voy a seleccionar como a equipo a los Celtics ("Bos" en la tabla). Y como variables numéricas los minutos, puntos, rebotes, asistencias y robos por partido (MPG, PPG, RPG, APG, SPG).

Según el diagrama de coordenadas paralelas, se puede observar que los jugadores con más minutos obtienen o suelen obtener los mejores números en las siguientes estadísticas.



Boston tiene 4 jugadores con estadísticas parecidas, pero si nos movemos a Milwaukee vemos como el jugador “Giannis Antetokounmpo” (línea azul) destaca sobre el resto en puntos y rebotes.



i) En el baloncesto en general se busca tener un equipo equilibrado con:

1. Un quinteto inicial con varios jugadores que dominen alguna estadística medible (puntos, rebotes, asistencias, etc). Estos son los que suelen jugar más minutos.
2. Jugadores de rotación con buenas estadísticas para dar descanso al quinteto inicial. Tienen 20-25 minutos de juego.
3. Jugadores de banquillo anoten o defiendan efectivamente en los pocos minutos que juegan (unos minutos o segundos por partido).

Para cumplir los puntos anteriores:

1. Buscar jugadores que tengan las variables de puntos, rebotes, asistencias, etc, más alta de la media.
2. Buscar jugadores que jueguen entre 20-25 minutos con buenos porcentajes de acierto e intentos de tiro sobre la media.
3. Buscar jugadores con buenos porcentajes de acierto y pocos minutos.

En los puntos 2 y 3, según las necesidades del equipo se pueden usar otras variables como la media de asistencias o rebotes.