

Universidade do Minho

Departamento de Informática

Sistemas Baseados em Similaridade

Trabalho Prático Individual 1

Gonçalo Almeida (A84610)

3 de Novembro de 2020

T1. Carregar, no Knime, o dataset descarregado. Aplicar nodos para exploração de dados, i.e., analisar os dados em relação à sua:

- a. Tendência central;
- b. Dispersão estatística;
- c. Correlação entre features.

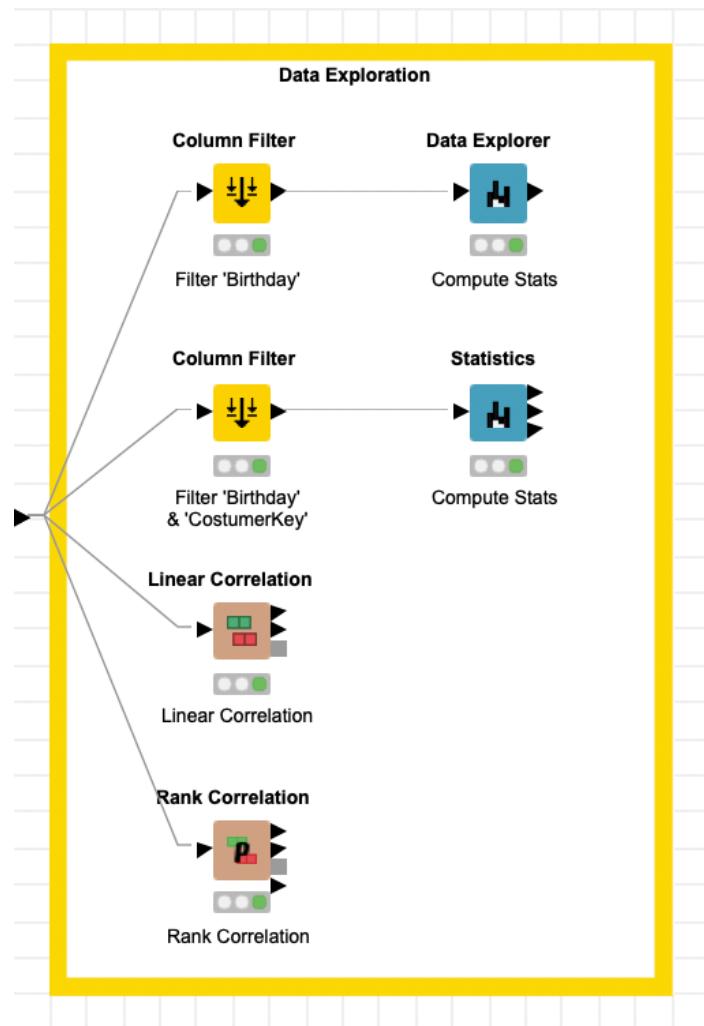


Figure 1 - Exploração de dados

Numeric Nominal Data Preview

Search:

Column	Exclude Column	Minimum	Maximum	Mean	Median	Standard Deviation	Variance
CustomerKey	□	11000	27336	17559.847	14967	5576.039	31092215.201
WebActivity	□	0	5	0.999	0	1.520	2.310
SentimentRating	□	0	5	1.851	2	1.620	2.624
EstimatedYearlyIncome	□	10000	170000	57718.072	60000	32091.910	1029890707.928
NumberOfContracts	□	0	4	1.465	1	1.145	1.311
Age	□	29	100	48.203	46	11.300	127.694
Target	□	0	1	0.487	0	0.500	0.250
Available401K	□	0	1	0.696	1	0.460	0.211
CustomerValueSegment	□	1	3	2.097	2	0.689	0.475
ChurnScore	□	0	1	0.269	0.100	0.332	0.110
CallActivity	□	1	5	3.237	3	1.262	1.594

Showing 1 to 11 of 11 entries

Figure 2 -View do nodo Data Explorer

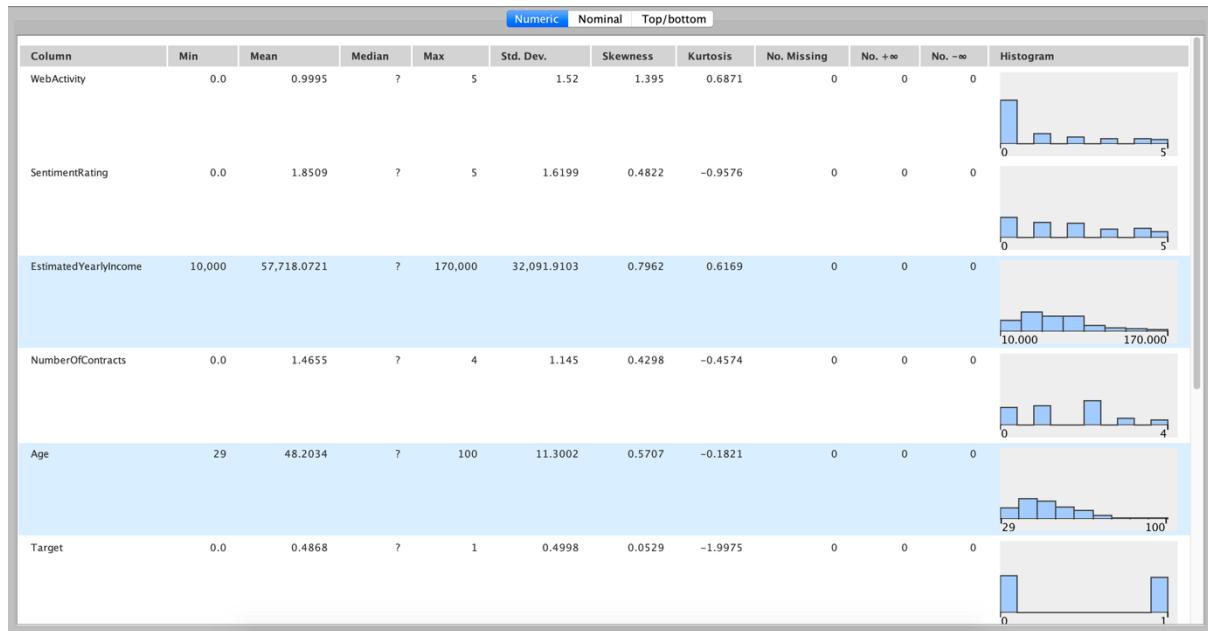


Figure 3 – View do nodo Statistics

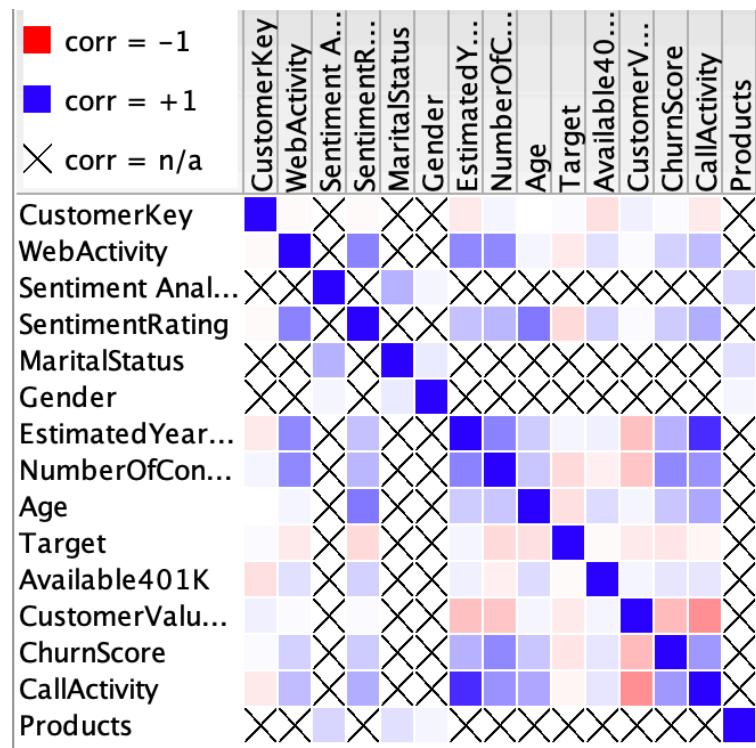


Figure 4 - Matriz de correlação linear

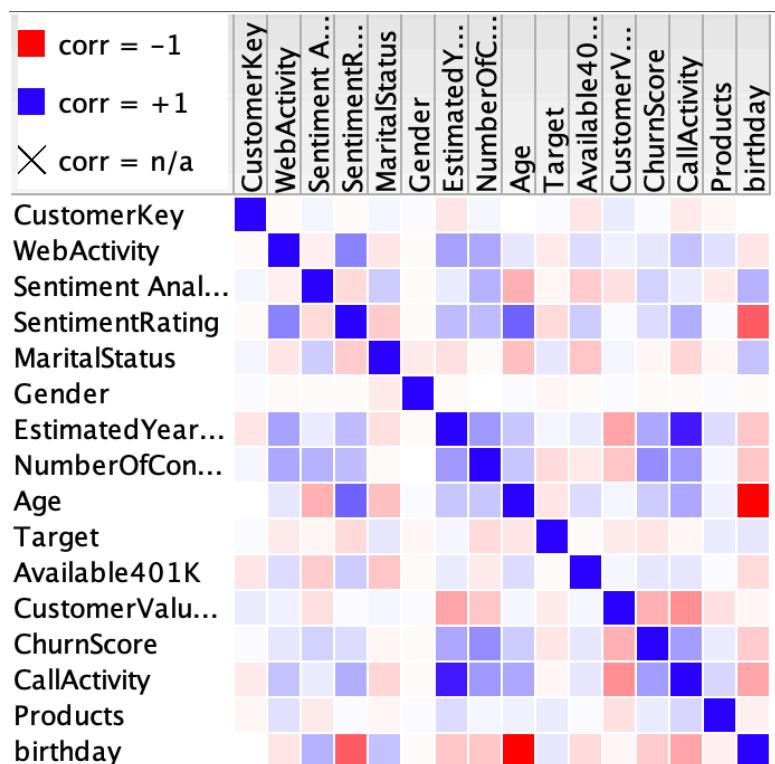
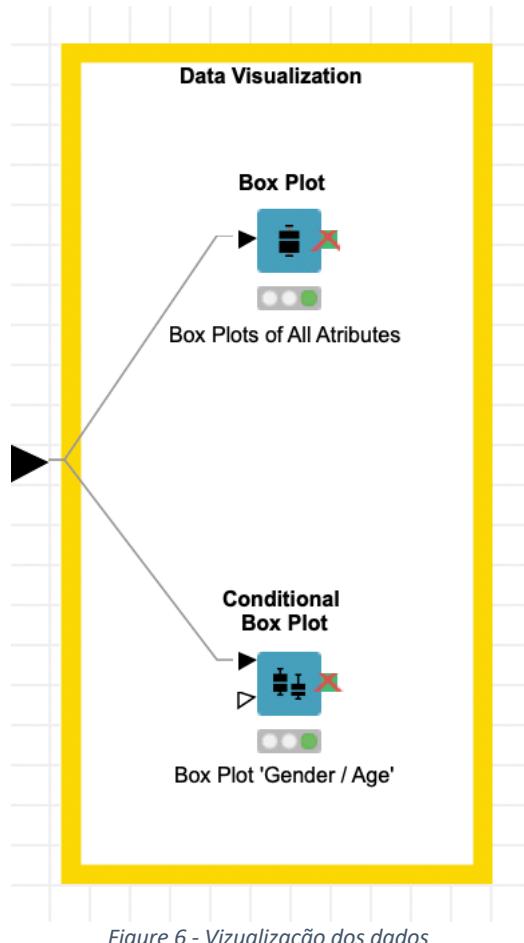


Figure 5 - Matriz de correlação de classificação

T2. Criar plots para visualização dos dados;



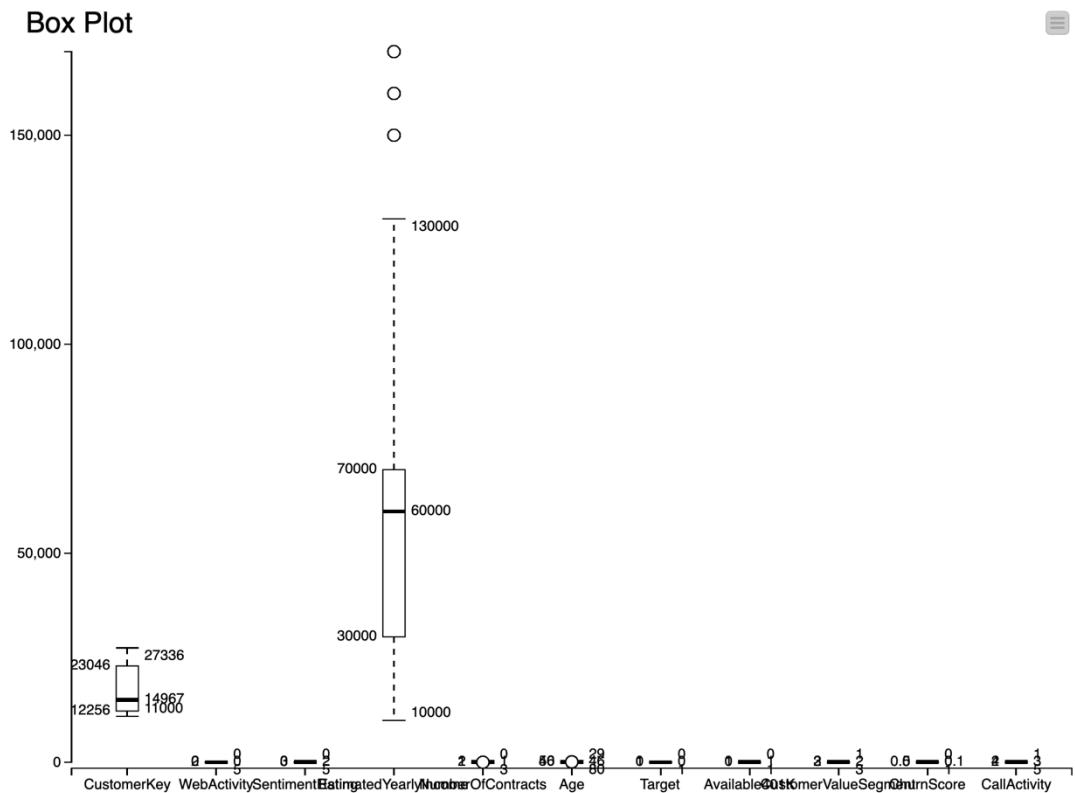


Figure 7 - Box Plot de todos os atributos

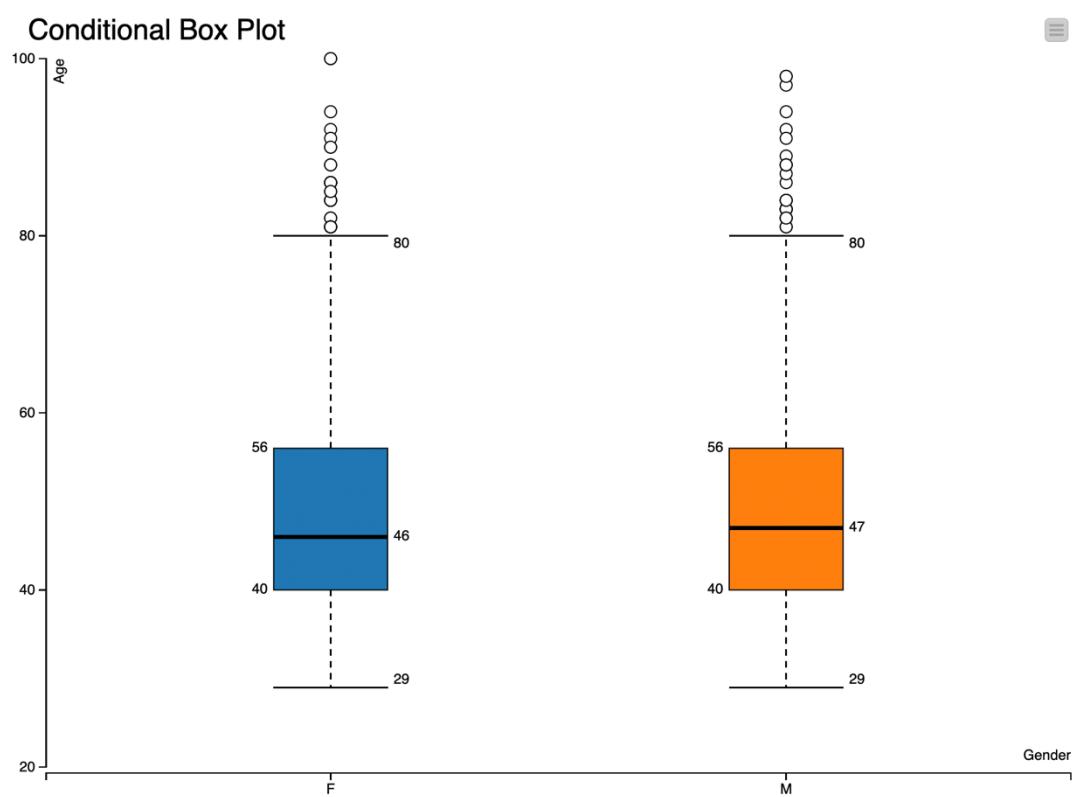


Figure 8 - Box Plot condicional entre os atributos Age e Gender

T3. Aplicar nodos para tratamento de dados de forma a:

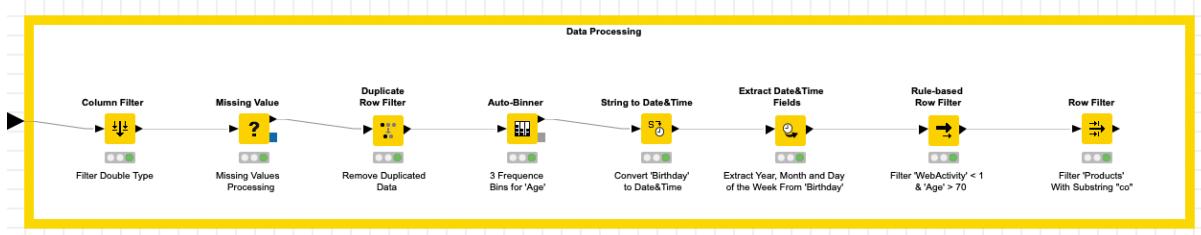


Figure 9 - Tratamento de dados

a. Excluir todas as colunas do tipo Double;

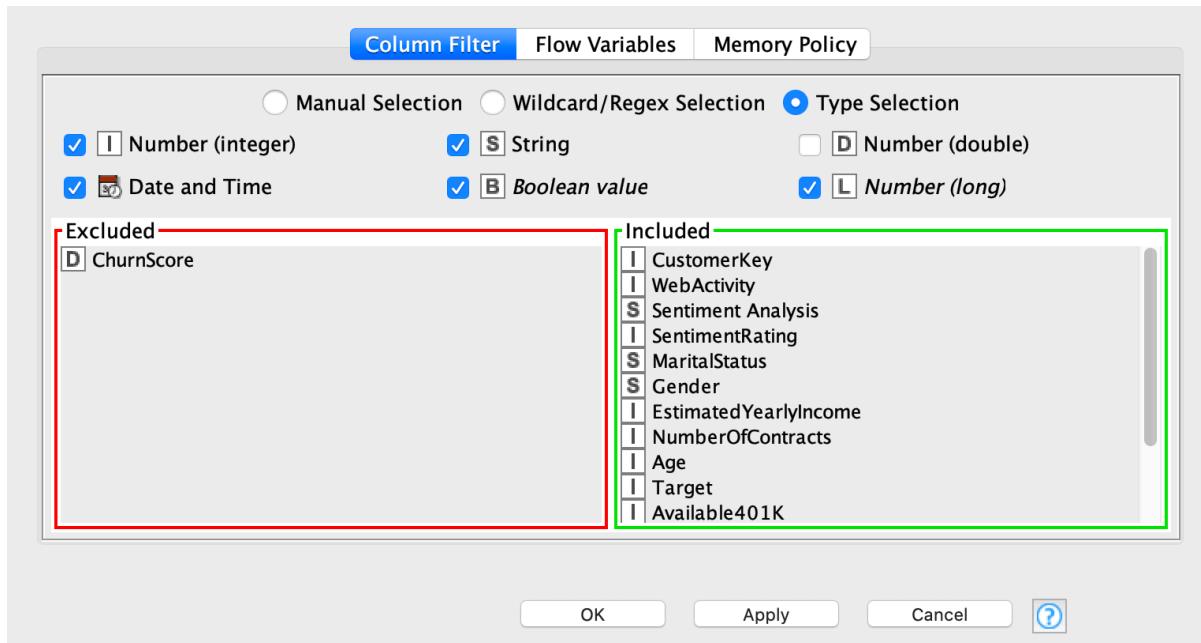


Figure 10 - Nodo Column Filter

b. Tratar valores em falta;

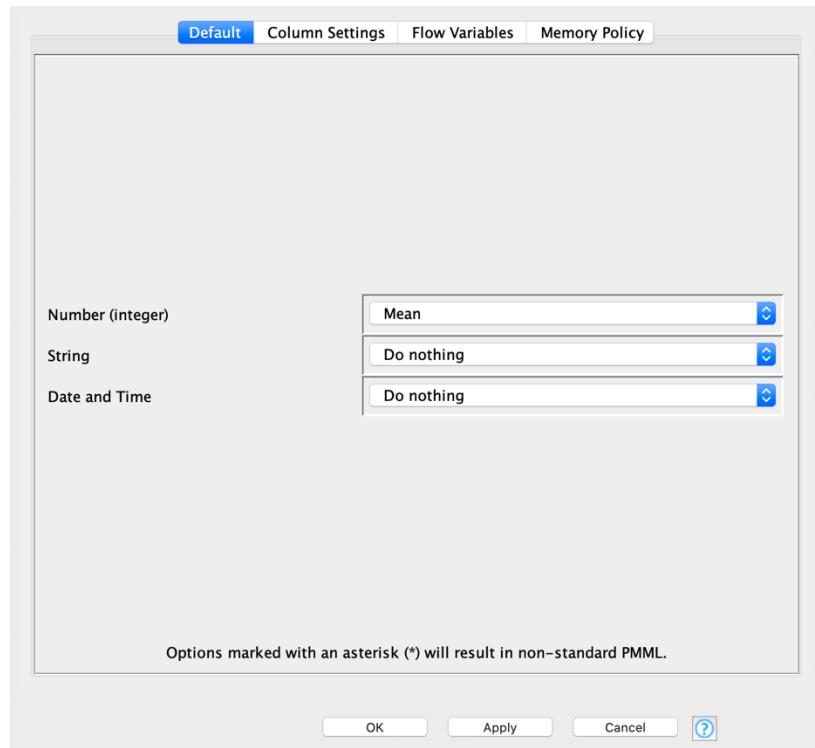


Figure 11 - Nodo Missing Value

c. Remover registos duplicados;

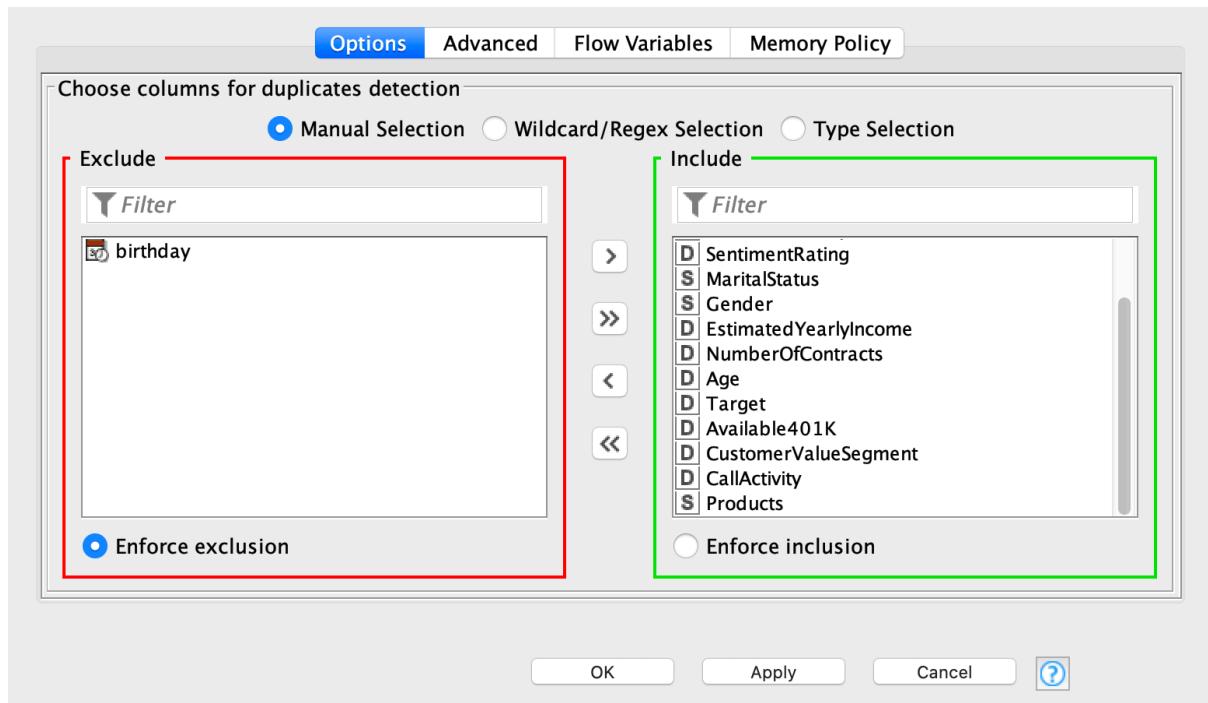


Figure 12 - Nodo Duplicate Row Filter

d. Criar 3 bins de igual frequência para a feature age;

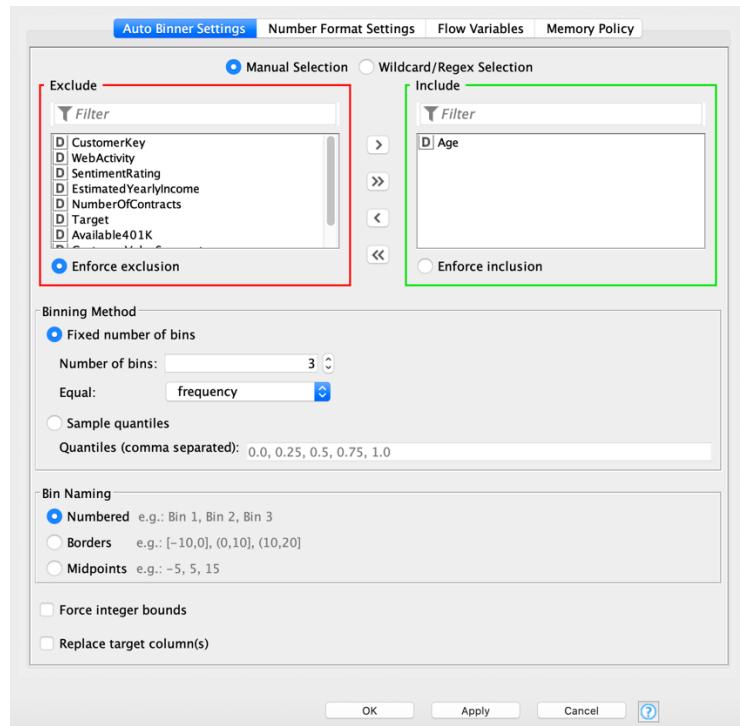


Figure 13 - Nodo Auto-Binner

e. Para cada registo, extraír o ano, mês e dia da semana da feature birthday;

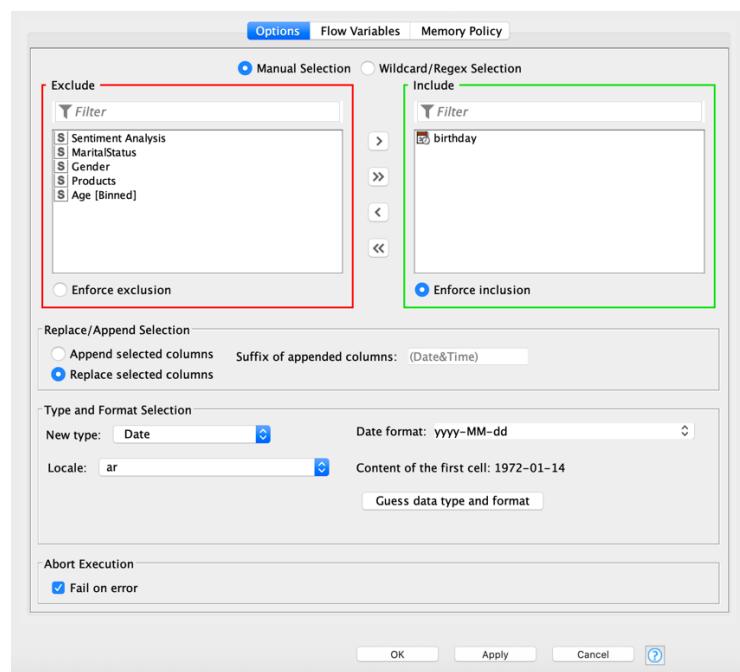


Figure 14 - Nodo String to Date&Time

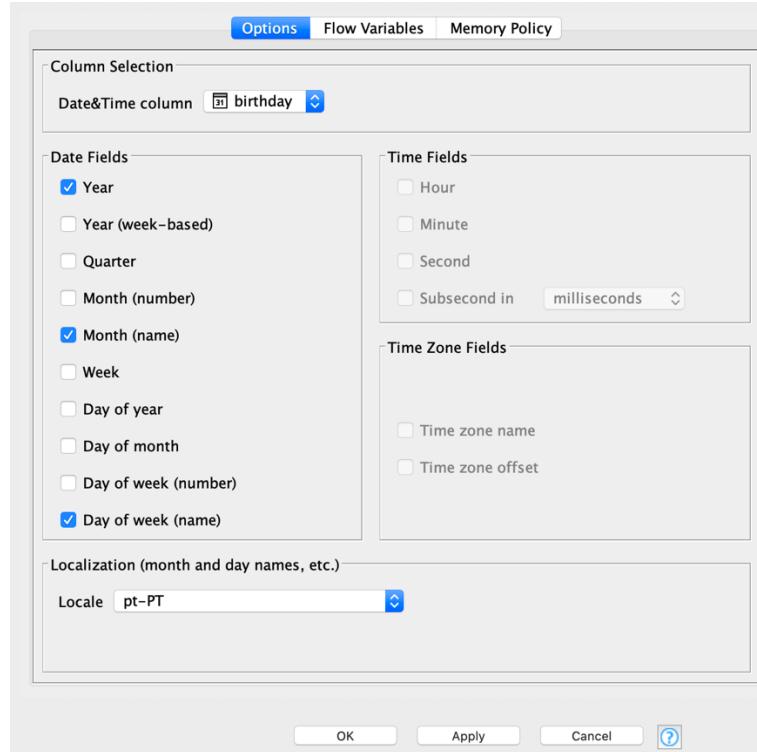


Figure 15 - Nodo Extract Date&Time Fields

f. Excluir utilizadores da plataforma que tenham uma atividade na plataforma (WebActivity) inferior a 1 hora e que tenham mais de 70 anos;

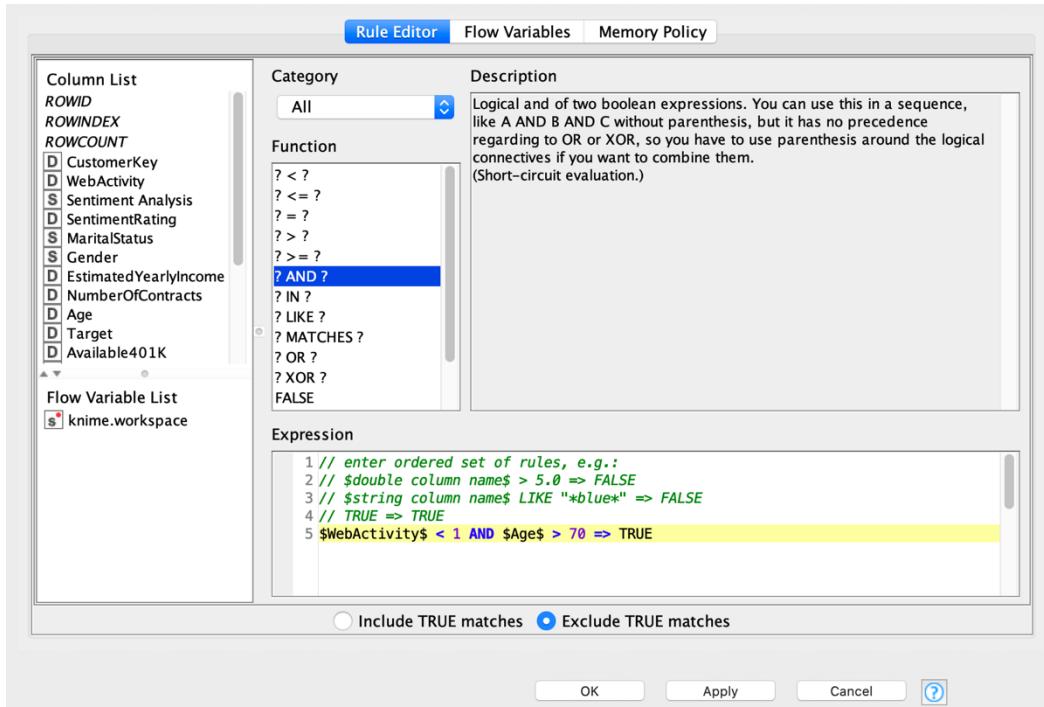


Figure 16 - Nodo Rule-based Row Filter

g. Excluir todos os registo que contenham a sub-string “co” no produto.

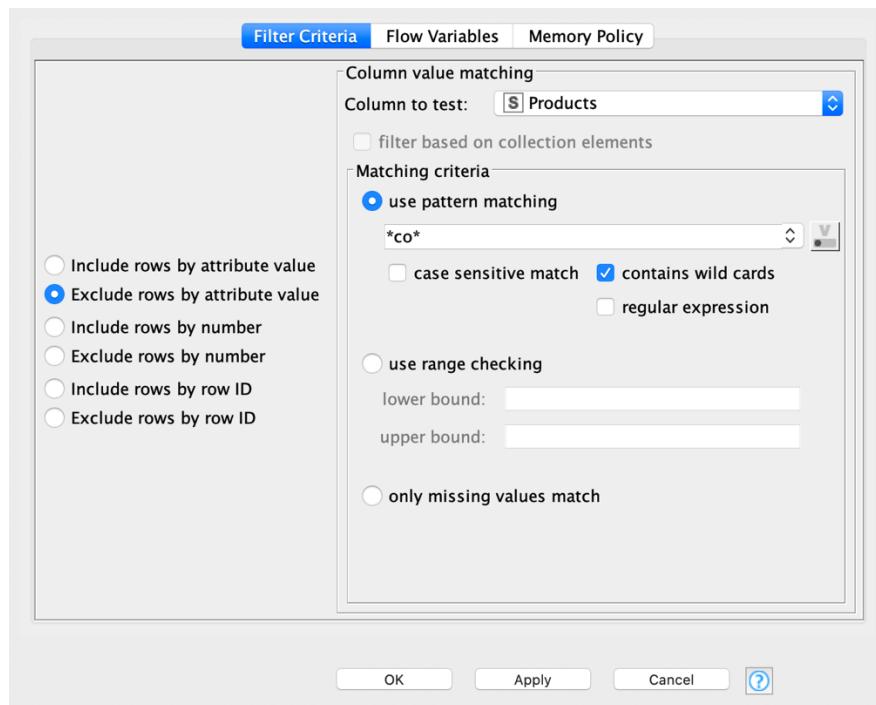


Figure 17 - Nodo Row Filter

T4. Aplicar nodos para agregação de dados de forma a:

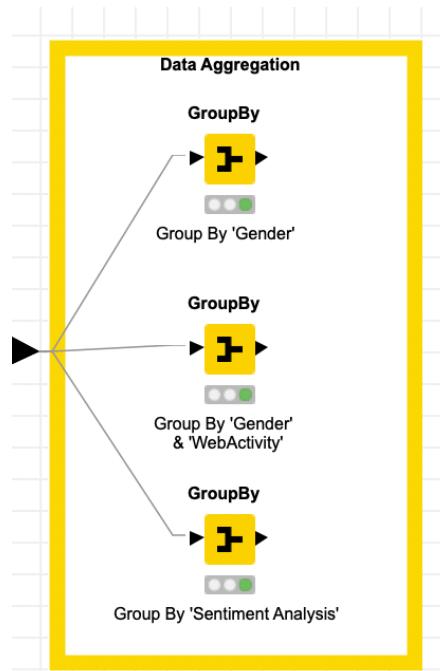


Figure 18 - Agregação de dados

- a. Por género, obter o número e a percentagem de registos, assim como a média da idade e da atividade na plataforma. Obter também o mínimo e máximo da idade;

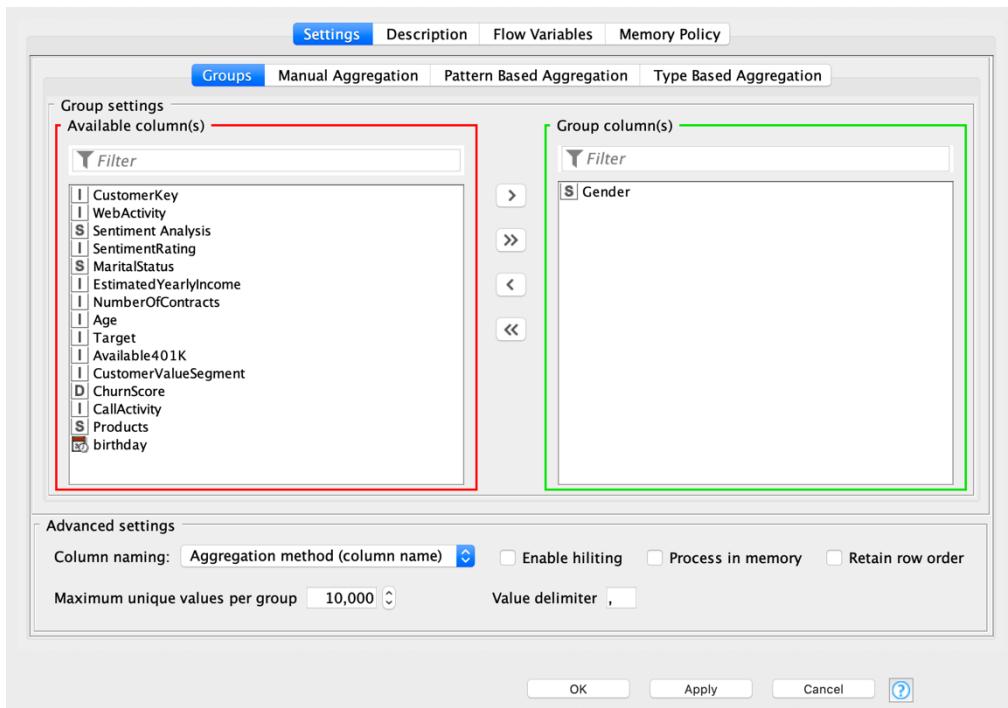


Figure 19 - Nodo GroupBy

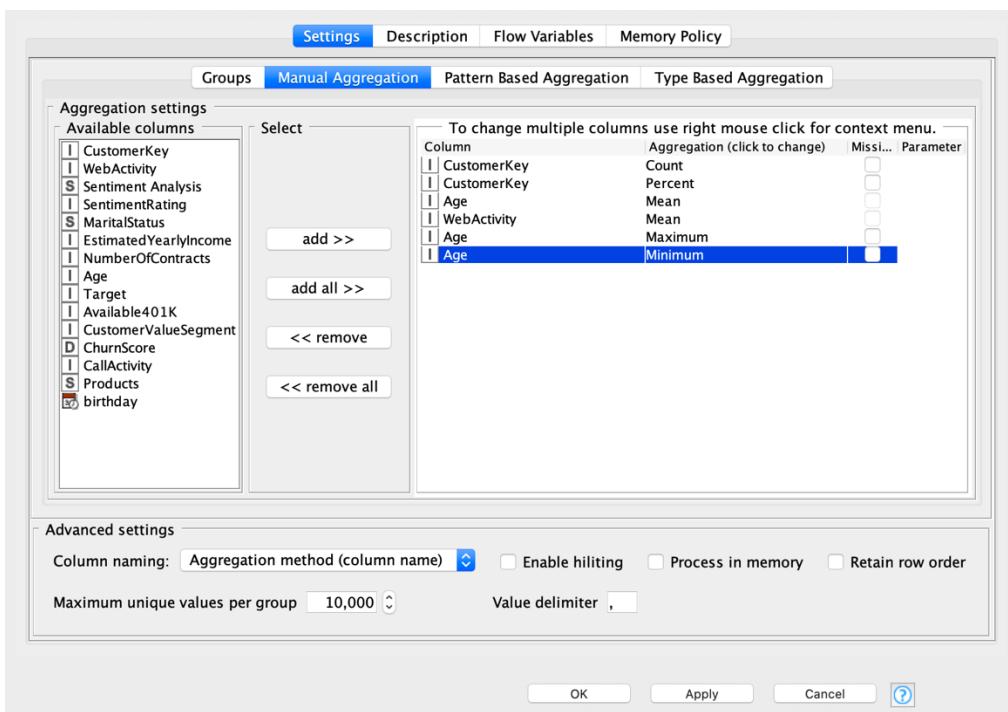


Figure 20 - Nodo GroupBy

- b. Por género e atividade na plataforma, obter a moda da análise do sentimento em relação à plataforma e a média da avaliação do sentimento;

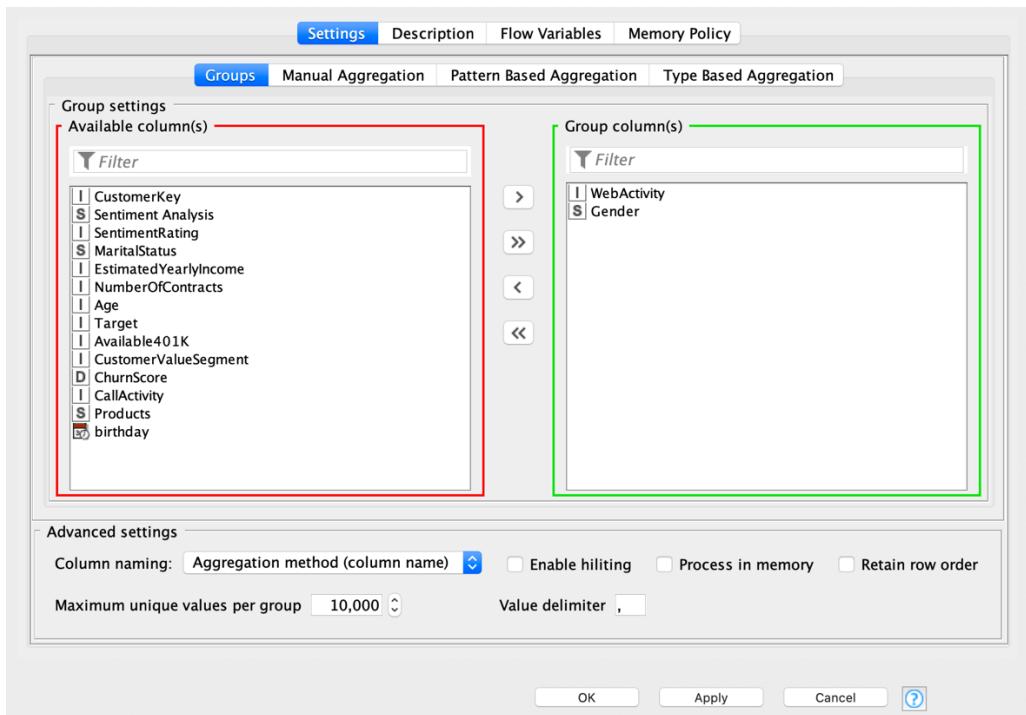


Figure 21 - Nodo GroupBy

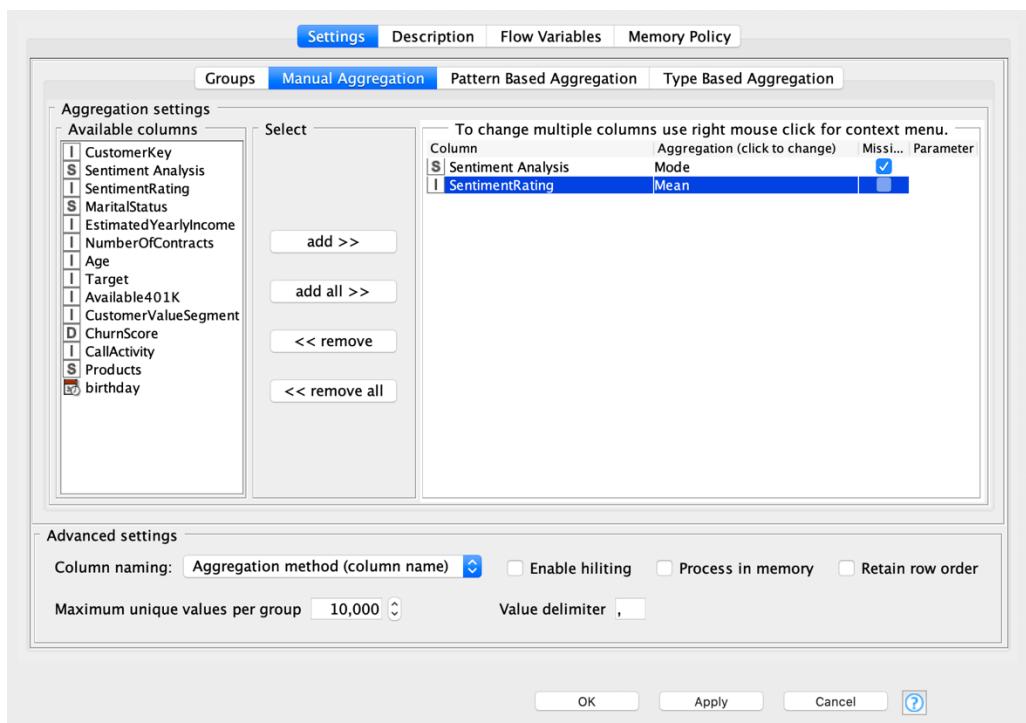


Figure 22 - Nodo GroupBy

- c. Por análise de sentimento, obter o número de registos, a média do salário anual estimado, o somatório do salário anual e a média do número de contratos.

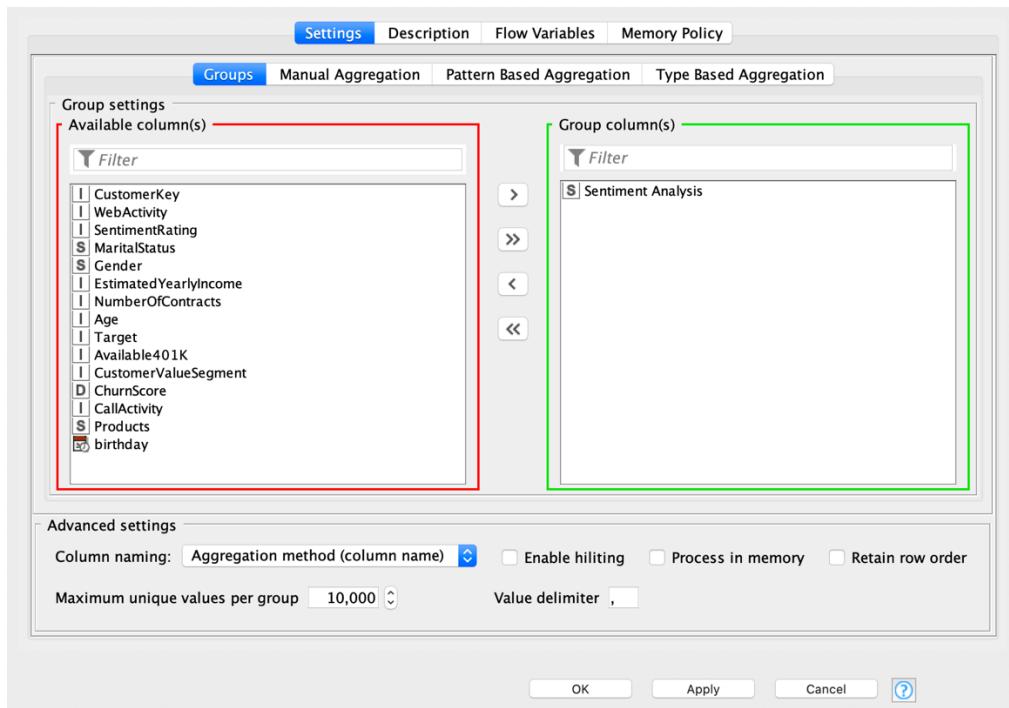


Figure 23 - Nodo GroupBy

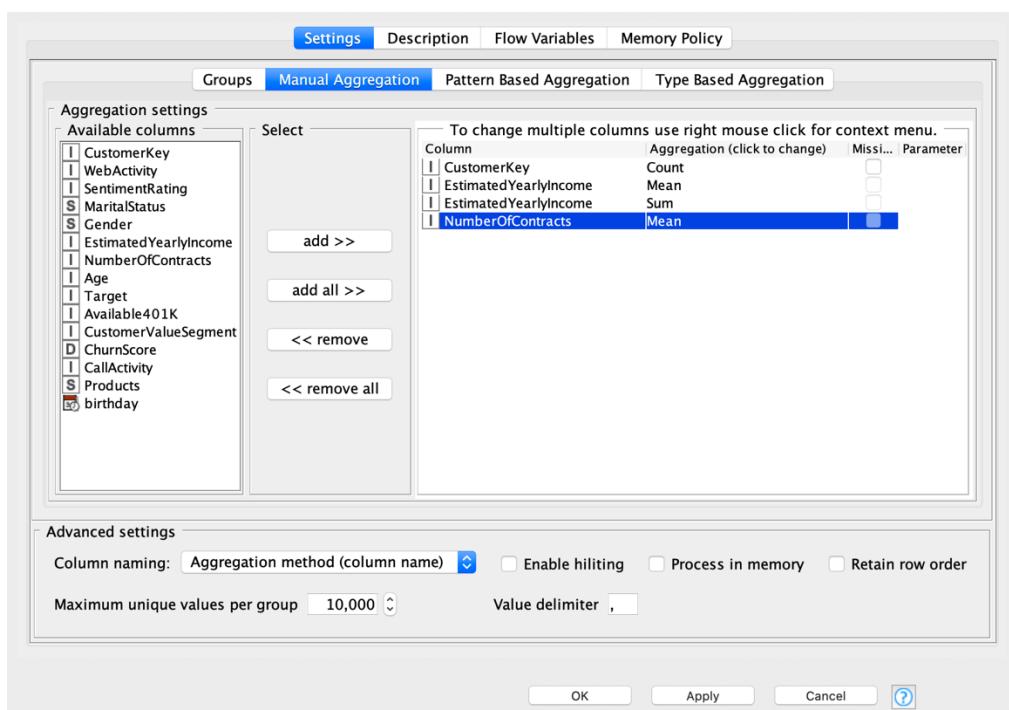


Figure 24 - Nodo GroupBy

T5. Análise crítica à informação extraída das agregações efetuadas na tarefa anterior. Que conclusões poderia a empresa tirar?

Analizando a informação extraída a partir das agregações efetuadas, algumas das conclusões são:

- O gênero masculino tem uma maior percentagem de regtos, apesar da diferença ser pouca;
- Os gêneros têm médias de idades muito próximas (aproximadamente 48 anos);
- O gênero feminino passa, em média, mais tempo na plataforma, apesar da diferença ser pouca;
- O gênero feminino tem o utilizador com maior idade (100 anos);
- A idade do utilizador mais novo é a mesma para ambos os gêneros (29 anos);
- A média do número de contratos de um utilizador aparenta ser proporcional à média do seu salário anual;
- As análises de sentimento extremas correspondem às médias de salário anual extremas, isto é, a pior análise corresponde à menor média salarial e a melhor análise à maior média salarial.

T6. Carregar, no Knime, o dataset descarregado. Explorar os dados, procurar informação relevante e mostrar essa mesma informação. P.e., qual a equipa mais indisciplinada? Qual o top-10 dos assistentes para golo? Qual o top-5 de nacionalidades na liga?

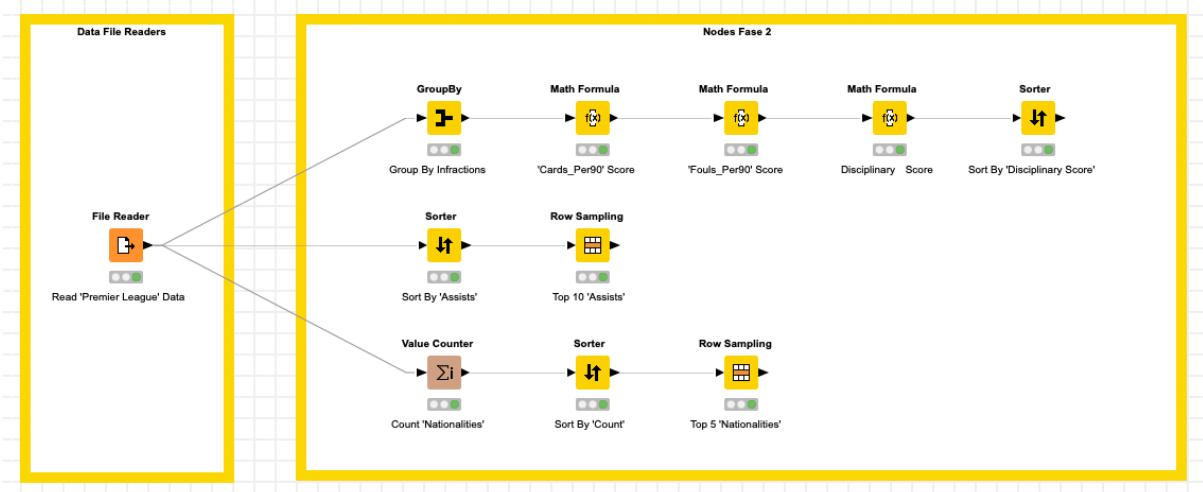


Figure 25 - Exploração dos dados

De modo a obter a equipa mais indisciplinada foi feito um agrupamento por equipas, realizando o somatório das faltas e cartões por 90 minutos. De seguida foram utilizados dois nodos *Math Formula* de modo a multiplicar os somatórios por 0.3 e 0.7 respetivamente de modo a dar maior relevância aos cartões. Por fim os resultados são somados para obter um score e a tabela é ordenada da equipa mais indisciplinada para a menos indisciplinada.

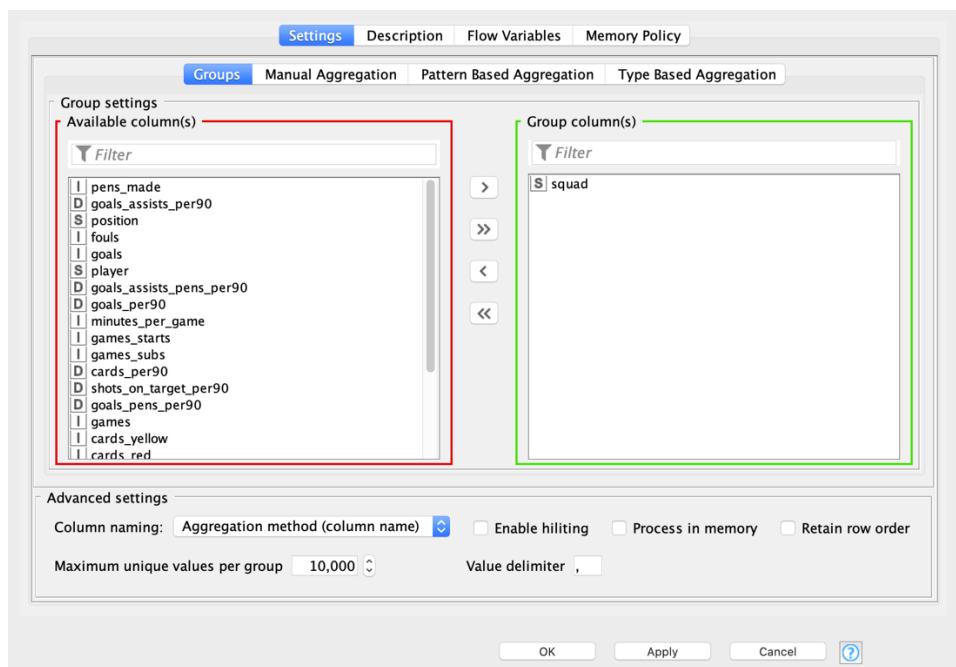


Figura 26 - Nod GroupBy

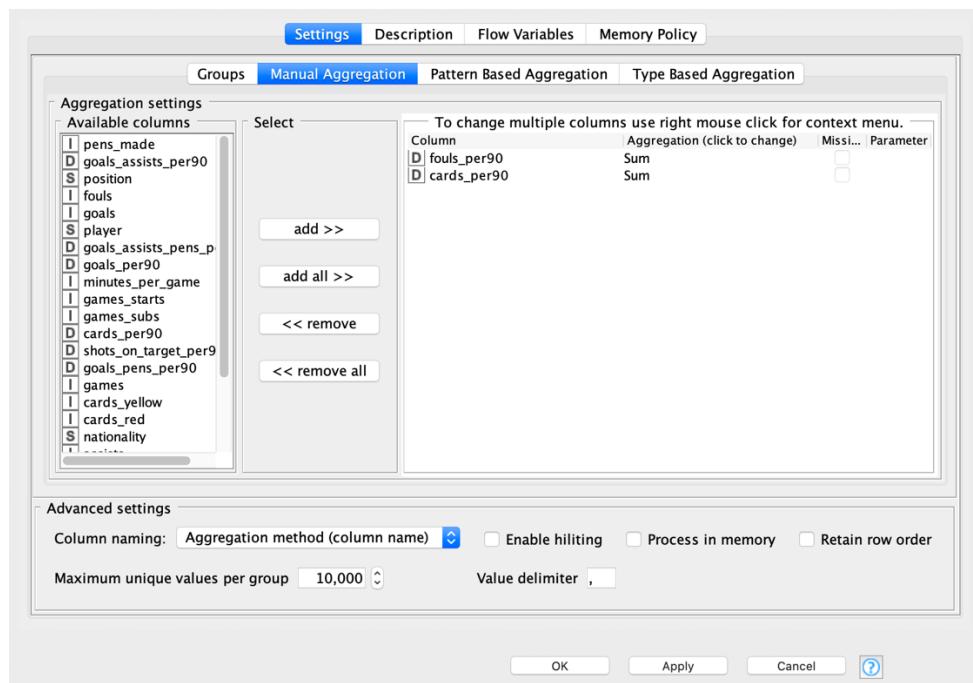


Figura 17 - Nodo GroupBy

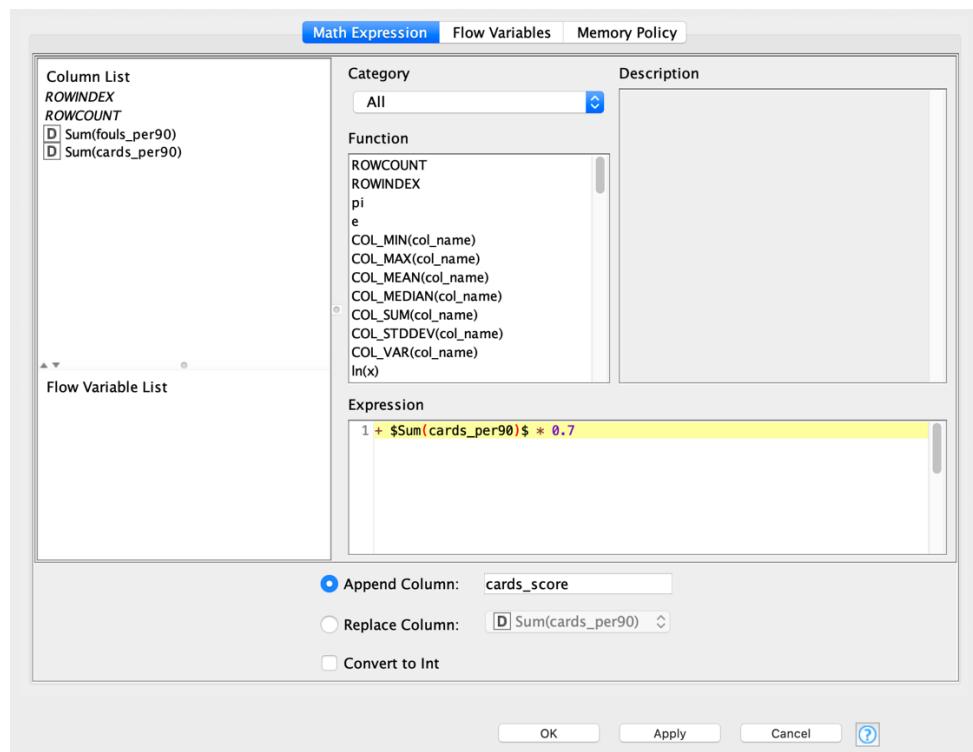


Figura 28 - Nodo Math Formula

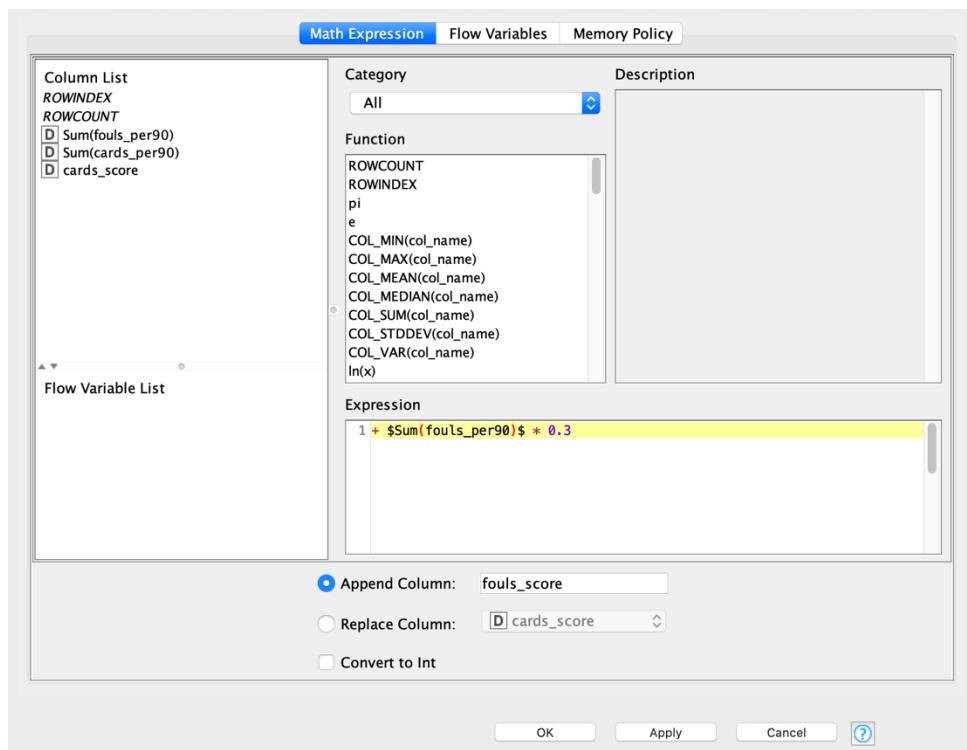


Figura 29 - Nodo Math Formula

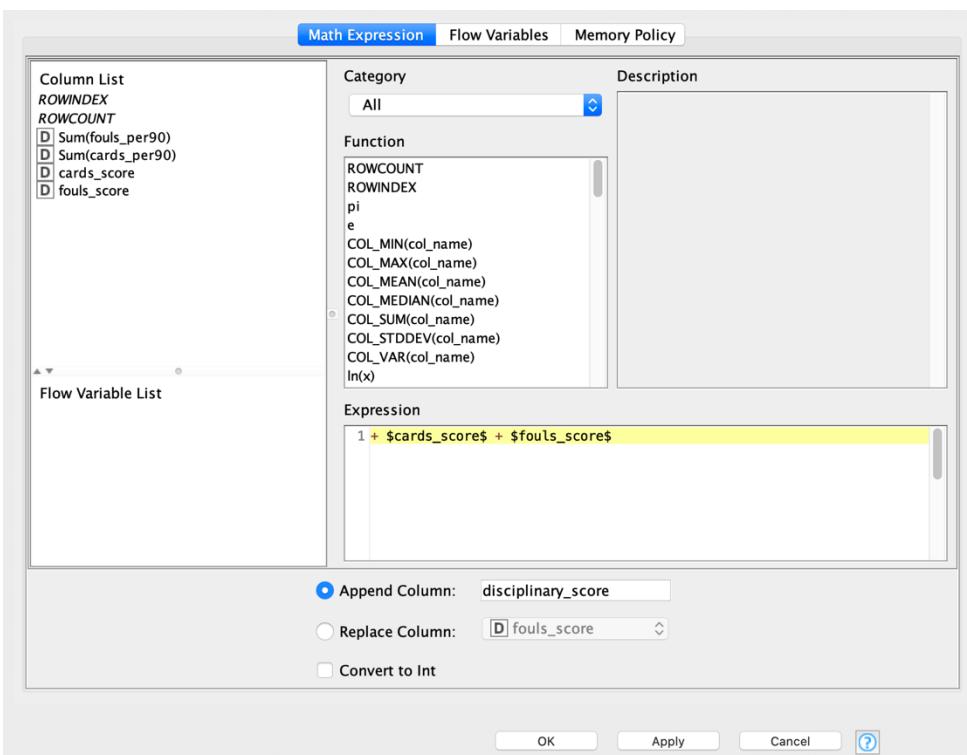


Figura 30 - Nodo Math Formula

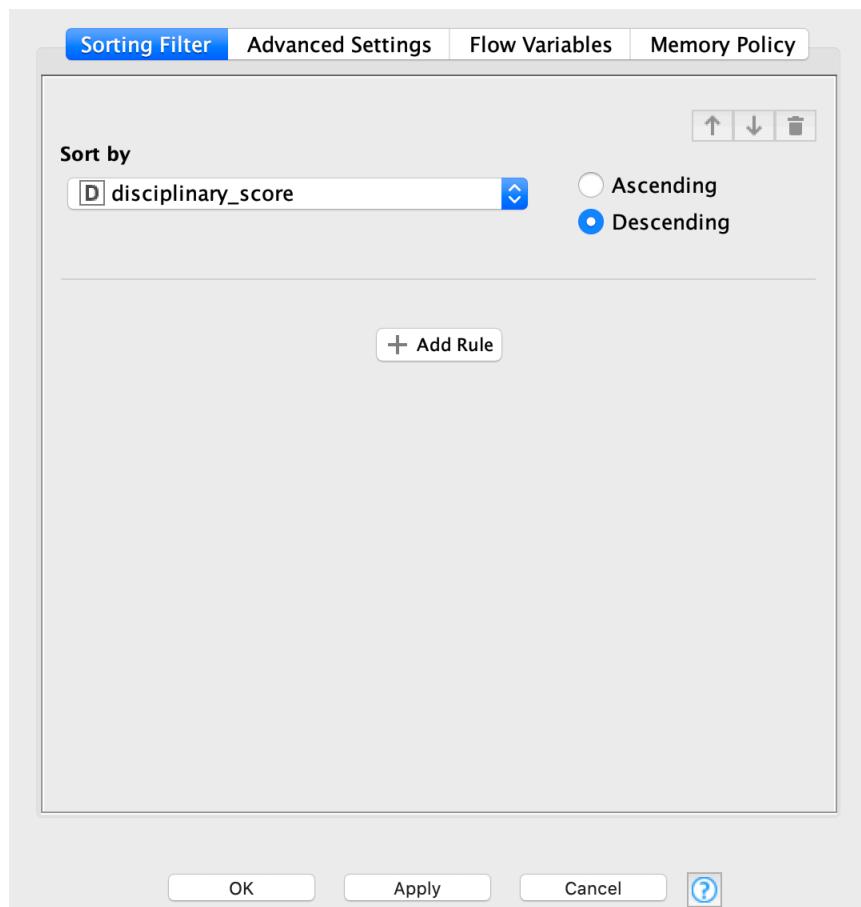


Figura 31 - Nodo Sorter

Table "default" – Rows: 20							Spec – Columns: 6	Properties	Flow Variables
Row ID	squad	Sum(f...)	Sum(c...)	cards...	fouls...	discipl...			
Row15	Swansea City	35.23	8.69	6.083	10.569	16.652			
Row7	Huddersfie...	29.54	7.58	5.306	8.862	14.168			
Row6	Everton	36.8	4.37	3.059	11.04	14.099			
Row17	Watford	34.25	4.2	2.94	10.275	13.215			
Row19	West Ham ...	33.22	4.64	3.248	9.966	13.214			
Row14	Stoke City	25.6	5.11	3.577	7.68	11.257			
Row8	Leicester City	28.64	3.41	2.387	8.592	10.979			
Row4	Chelsea	31.13	2.21	1.547	9.339	10.886			
Row18	West Brom...	23.86	5.18	3.626	7.158	10.784			
Row0	Arsenal	26.84	3.31	2.317	8.052	10.369			
Row11	Manchester...	24.12	3.8	2.66	7.236	9.896			
Row3	Burnley	23.76	3.11	2.177	7.128	9.305			
Row10	Manchester...	22.59	3.59	2.513	6.777	9.29			
Row5	Crystal Pal...	22.53	3.59	2.513	6.759	9.272			
Row16	Tottenham ...	22.86	3.24	2.268	6.858	9.126			
Row12	Newcastle ...	22.98	3.18	2.226	6.894	9.12			
Row13	Southampton	22.86	3.17	2.219	6.858	9.077			
Row2	Brighton & ...	22.37	2.62	1.834	6.711	8.545			
Row9	Liverpool	21.06	2.77	1.939	6.318	8.257			
Row1	Bournemouth	17.86	3.06	2.142	5.358	7.5			

Figura 32 - Tabela final

De modo a obter o top 10 dos assistentes para golo a tabela foi ordenada decrescentemente pelo atributo *assists* e de seguida foi recolhida uma amostra das 10 primeiras linhas.

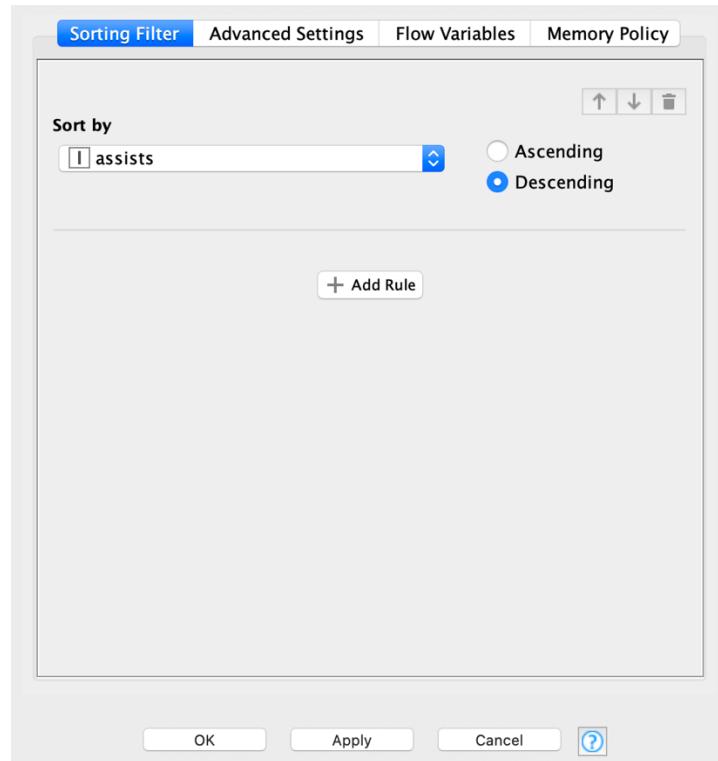


Figura 33 - Nodo Sorter

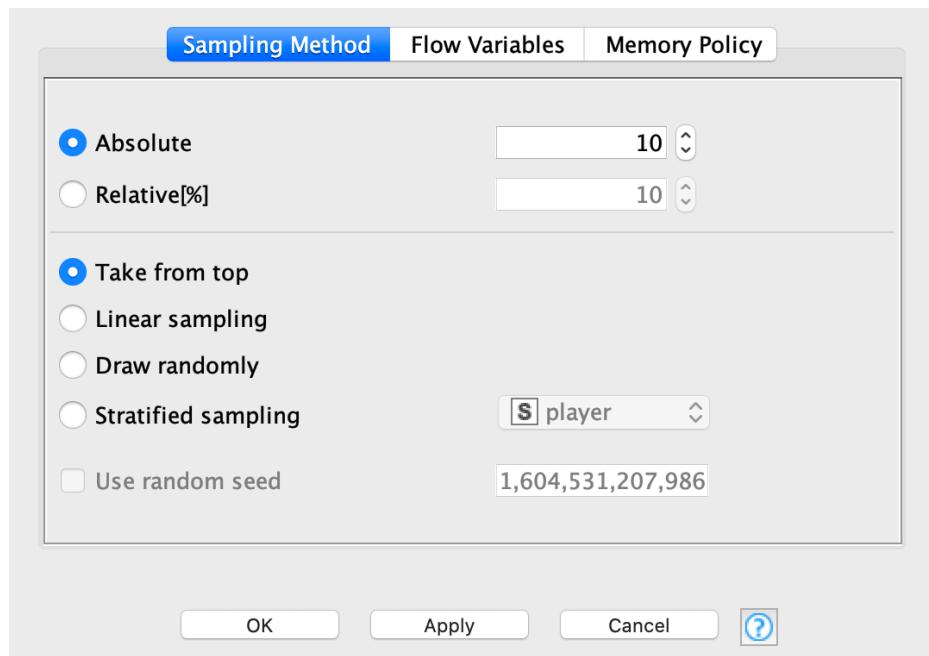


Figura 34 - Nodo Row Sampling

Table "EPL_Player_Stats_2017_2018.csv" – Rows: 10 Spec – Columns: 25 Properties Flow Variables

Row ID	I pens...	D goals...	S position	I fouls	I goals	S player	D goals...	S squad	D goals...	I minut...	I ga
37	0	0.7	MF	30	8	Kevin De Bruyne	0.7	Manchester City	0.23	83	36
104	0	0.93	FW,MF	24	10	Leroy Sané	0.93	Manchester City	0.37	76	27
85	1	1.01	MF,FW	39	18	Raheem Sterling	0.97	Manchester City	0.63	78	29
101	0	0.74	MF	21	9	David Silva	0.74	Manchester City	0.33	84	28
23	0	0.56	MF,DF	14	10	Christian Eriksen	0.56	Tottenham Ho...	0.28	87	37
48	0	0.58	MF	45	9	Dele Alli	0.58	Tottenham Ho...	0.27	82	34
49	0	0.67	MF,FW	23	12	Riyad Mahrez	0.67	Leicester City	0.37	82	34
52	1	1.3	MF,FW	16	32	Mohamed Salah	1.27	Liverpool	0.99	81	34
142	0	0.67	MF	44	6	Paul Pogba	0.67	Manchester U...	0.25	80	25
51	1	0.46	MF	33	7	Pascal Groß	0.43	Brighton & Ho...	0.22	77	35

Figura 35 - Tabela final

De modo a obter o top 5 nacionalidades na liga utilizou-se um nodo Value Counter para contar o número de ocorrências de cada nacionalidade, de seguida a tabela foi ordenada decrescentemente e foi recolhida uma amostra das 5 primeiras linhas.

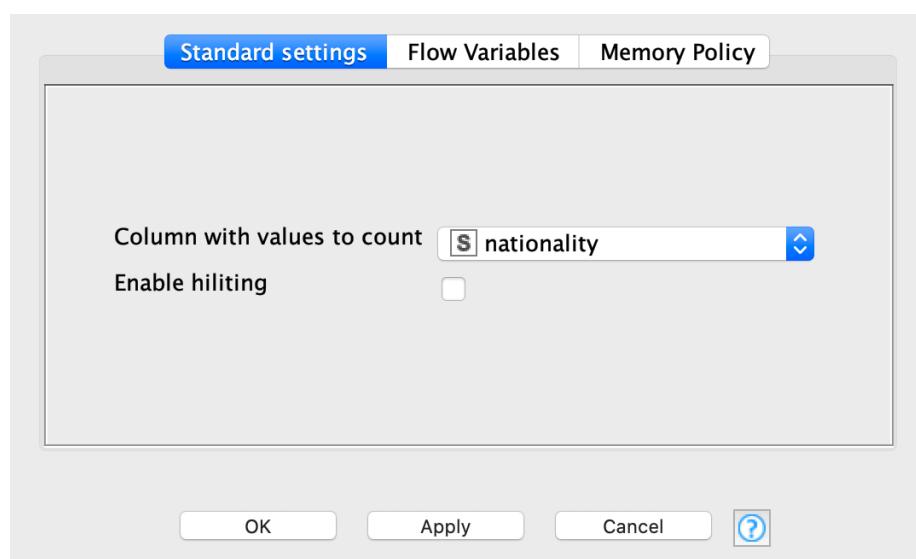


Figura 36 - Nodo Value Counter

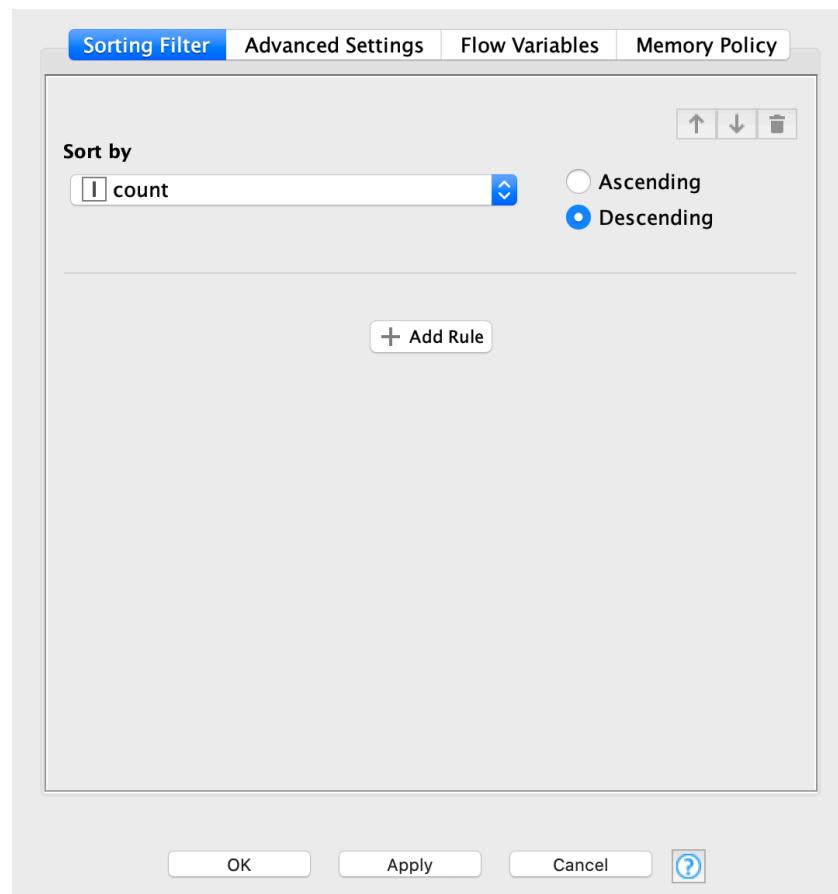


Figura 37 - Nodo Sorter

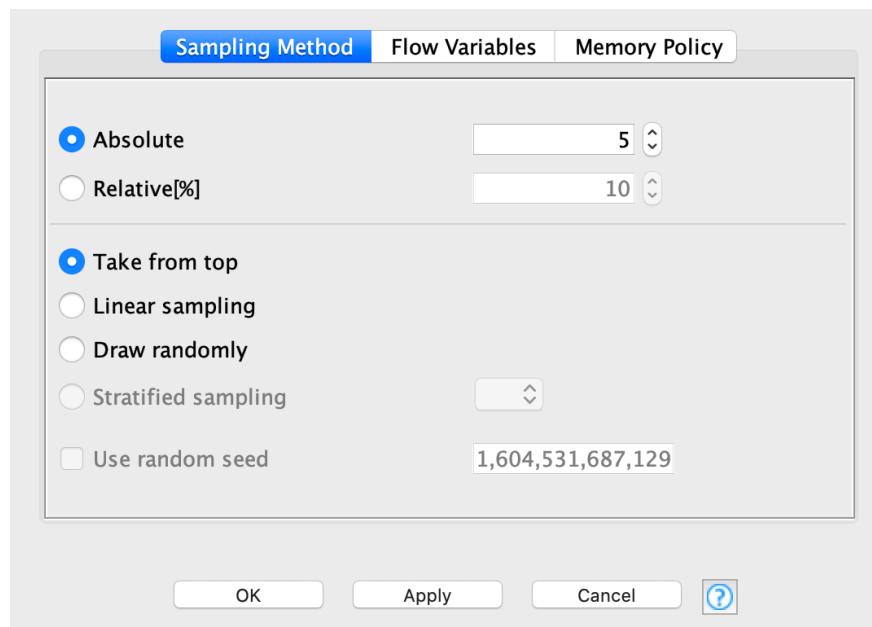


Figura 38 - Nodo Row Sampling

Row ID		count
eng ENG		174
es ESP		31
fr FRA		29
nl NED		23
be BEL		21

Figura 39 - Tabela final