

Sistemas Baseados em Similaridade - SBS 2019/2020

Mestrado Integrado em Engenharia Informática

Ficha 3



José Pinto

A84590

T1

a)

















Column 	Minimum 	Maximum 	Mean 	Median 
 CustomerKey	11000	27336	17559.847	14967
 WebActivity	0	5	0.999	0
 SentimentRating	0	5	1.851	2
 EstimatedYearlyIncome	10000	170000	57718.072	60000
 NumberOfContracts	0	4	1.465	1
 Age	29	100	48.203	46
 Target	0	1	0.487	0
 Available401K	0	1	0.696	1
 CustomerValueSegment	1	3	2.097	2
 ChurnScore	0	1	0.269	0.100
 CallActivity	1	5	3.237	3

Figura 1 - Tendência central

b)

Column	Standard Deviation
+ CustomerKey	5576.039
+ WebActivity	1.520
+ SentimentRating	1.620
+ EstimatedYearlyIncome	32091.910
+ NumberOfContracts	1.145
+ Age	11.300
+ Target	0.500
+ Available401K	0.460
+ CustomerValueSegment	0.689
+ ChurnScore	0.332
+ CallActivity	1.262

Figura 2 - Dispersão estatística

c)

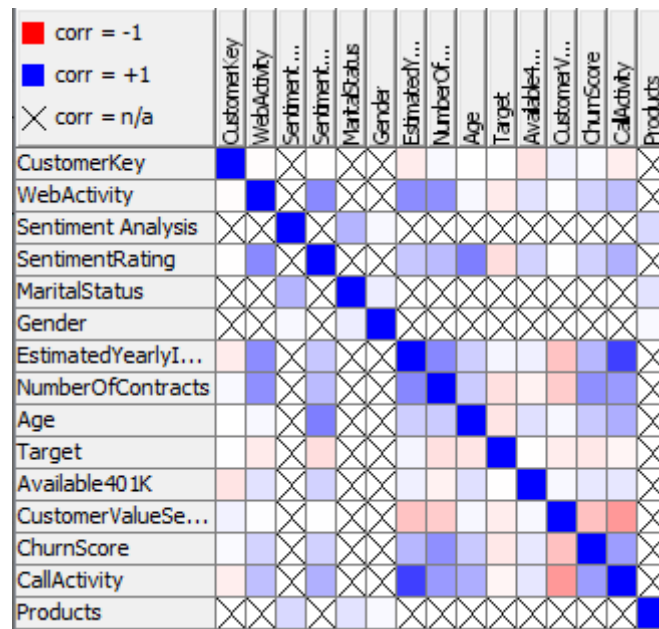


Figura 3 - Correlação entre features

T2

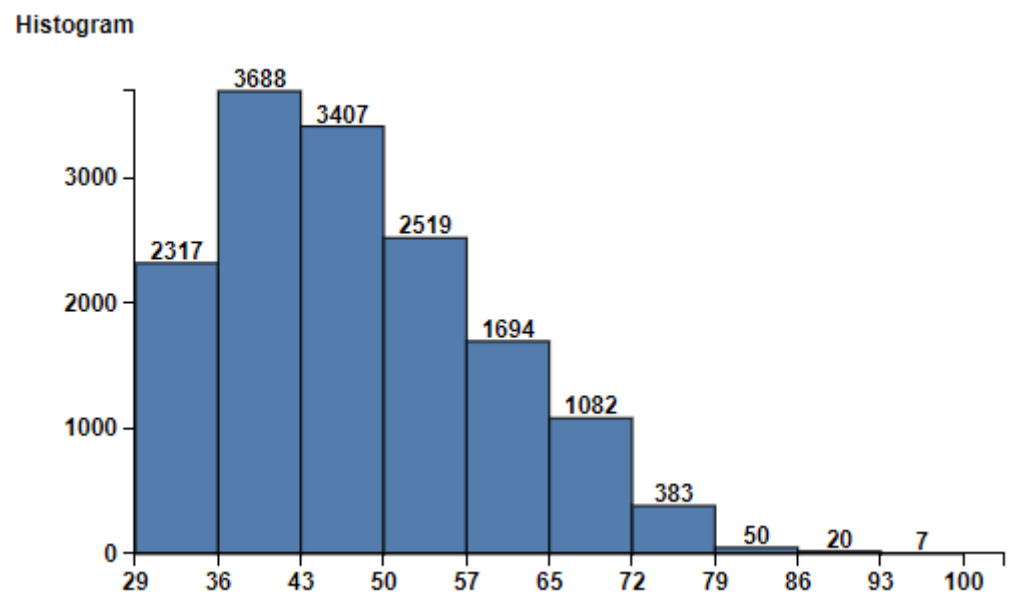


Figura 4 - Histograma Idades dos utilizadores

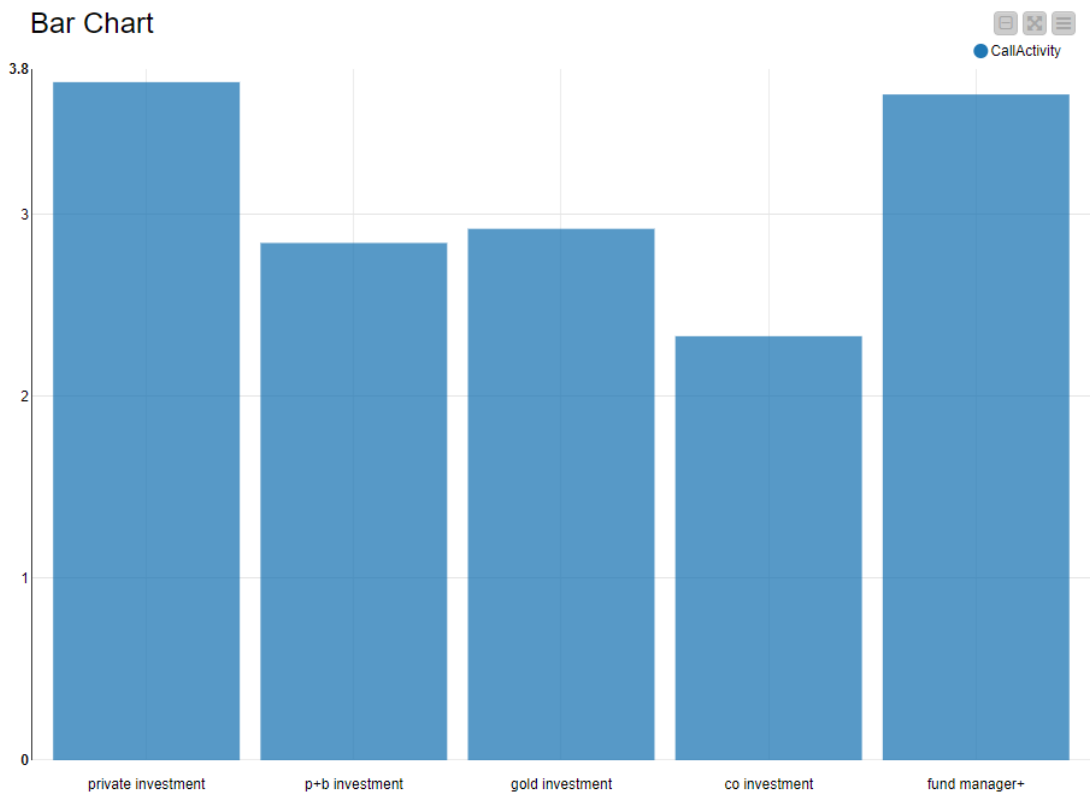


Figura 5 - Gráfico de barras Produtos e Atividade de chamadas

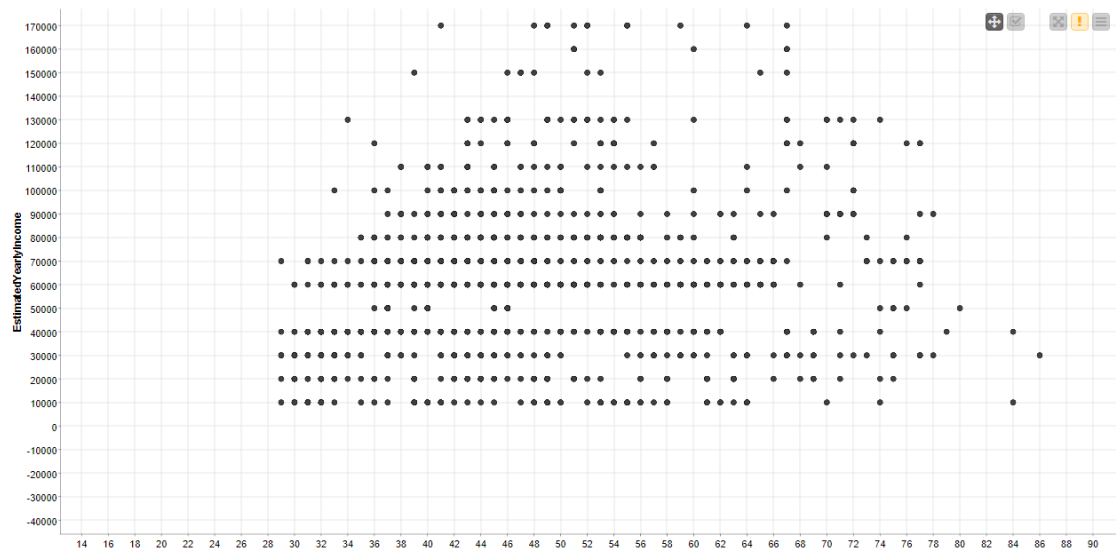


Figura 6 - Gráfico de dispersão Idade e Rendimento anual estimado

Pie Chart

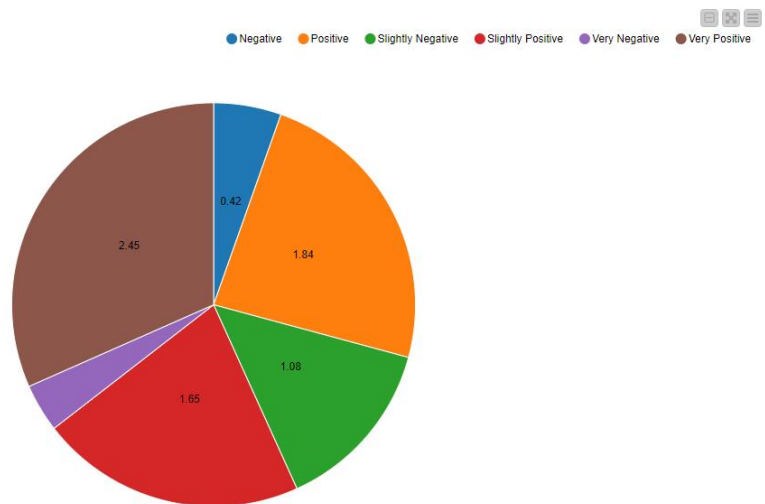


Figura 7 - Gráfico pizza entre o Sentimento e a Atividade Web dos clientes

T3

a)

The screenshot shows the 'Column Filter' dialog box with three tabs: 'Column Filter', 'Flow Variables', and 'Memory Policy'. The 'Manual Selection' radio button is selected. The 'Exclude' section on the left has a red border and contains a search filter 'Filter' and a list with 'D ChurnScore'. The 'Include' section on the right has a green border and contains a search filter 'Filter' and a list of variables: CustomerKey, WebActivity, Sentiment Analysis, SentimentRating, MaritalStatus, Gender, EstimatedYearlyIncome, NumberOfContracts, Age, Target, and AvailableOnWeb. Between the sections are navigation buttons: '>', '>>', '<', and '<<'. At the bottom, 'Enforce exclusion' is selected for the left and 'Enforce inclusion' is selected for the right.

Figura 8 - Filtragem das colunas com doubles

b)

The screenshot shows the 'Default' tab of the 'Column Settings' dialog box. It lists three data types: 'Number (integer)', 'String', and 'Date and Time'. To the right, there are three dropdown menus for missing value treatment: 'Mean' for 'Number (integer)', 'Most Frequent Value' for 'String', and 'Do nothing' for 'Date and Time'. A note at the bottom states: 'Options marked with an asterisk (*) will result in non-standard PMML.'

Figura 9 - Tratamento dos Missing Values

c)

Após a aplicação deste nodo a tabela passou de 15167 linhas para 15151 linhas. O que significa que haviam no total 16 linhas repetidas.

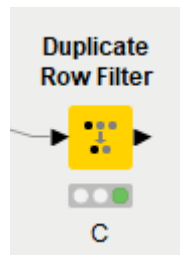


Figura 10 - Remoção dos registos duplicados

d)

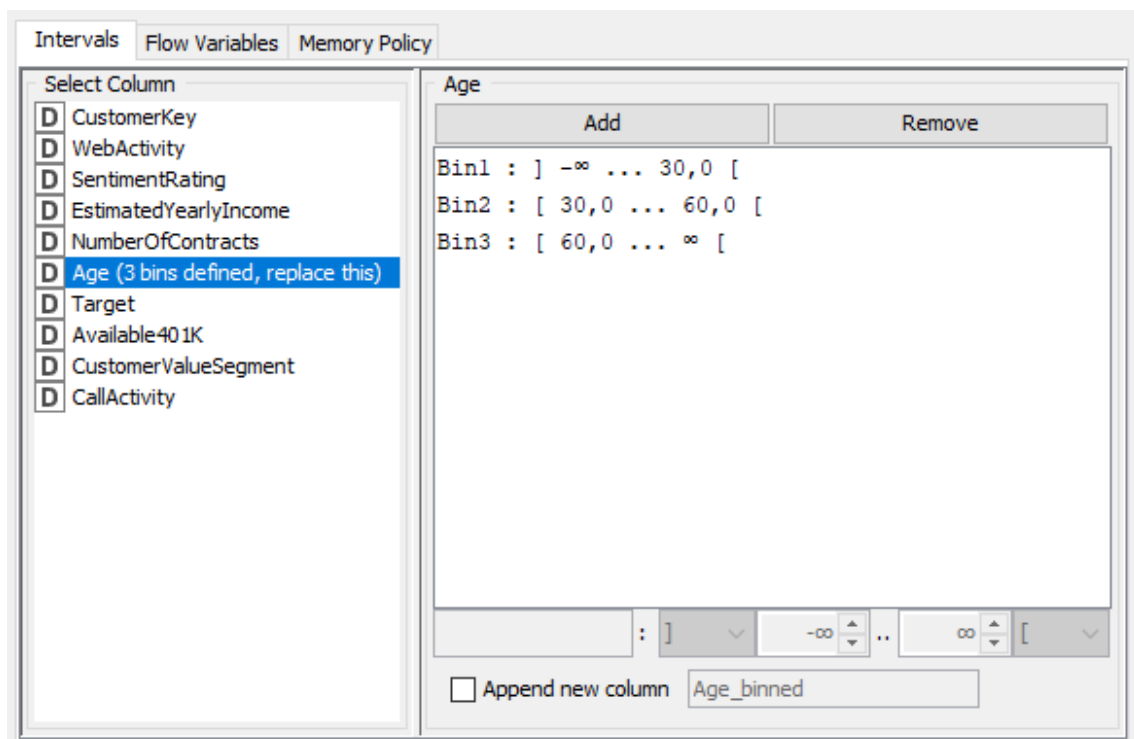


Figura 11 - Numeric Binner

e)

O obstáculo nesta alínea era o formato das datas que não permitia a extração direta dos dados. Por isso foi feita a conversão para String e depois novamente para data de forma a os meses data serem identificados por número em vez de nome. E depois foi feita a extração dos campos pedidos.

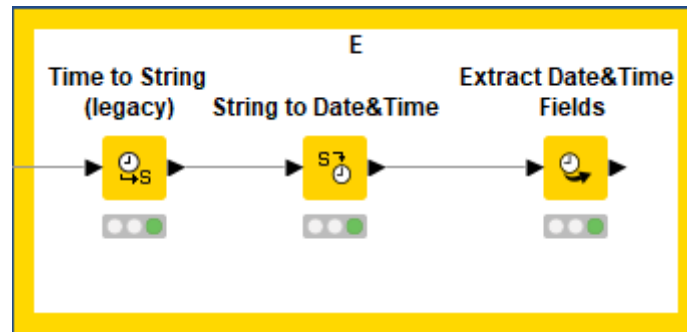


Figura 12 - Nodos alínea e

I Year	S Month (...)	S Day of ...
1972	Janeiro	Sexta-feira
1971	Agosto	Sábado
1970	Junho	Sexta-feira
1971	Fevereiro	Quinta-feira
1971	Janeiro	Quarta-feira
1971	Fevereiro	Quarta-feira
1973	Novembro	Quarta-feira
1974	Fevereiro	Quinta-feira
1973	Setembro	Sexta-feira
1974	Janeiro	Quarta-feira
1970	Junho	Sexta-feira
1970	Maio	Quarta-feira
1971	Julho	Quinta-feira
1972	Janeiro	Quarta-feira
1970	Janeiro	Sábado
1970	Fevereiro	Sábado
1970	Março	Quinta-feira
1969	Agosto	Segunda-feira
1969	Julho	Segunda-feira
1969	Outubro	Terça-feira
1969	Novembro	Quinta-feira
1969	Outubro	Domingo
1969	Novembro	Domingo
1970	Janeiro	Sábado

Figura 13 - Excerto das extraídas

f)

Expression

```

1 // enter ordered set of rules, e.g.:
2 // $double column name$ > 5.0 => FALSE
3 // $string column name$ LIKE "*blue*" => FALSE
4 // TRUE => TRUE
5 $Age$ >= 70 AND $WebActivity$ < 1 => TRUE

```

Figura 14 - Rule-based Row Filter

g)

Expression

```

1 // enter ordered set of rules, e.g.:
2 // $double column name$ > 5.0 => FALSE
3 // $string column name$ LIKE "*blue*" => FALSE
4 // TRUE => TRUE
5 $Products$ LIKE "*Co*" => TRUE

```

☐ Include TRUE matches
☒ Exclude TRUE matches

Figura 15 - Rule-based Row Filter

T4

a)

Row ID	S Gender	D Mean(...	D Mean(A...	I Min*(Age)	I Max*(A...	I First*(...	D Percent...
Row0	F	1.018	48.173	29	100	7581	50
Row1	M	0.981	48.233	29	98	7586	50

Figura 16 - Tabela resultante Gender-WebActivity-Age-Min(age)-Max(age)-Percent(registos)-Número(registos)

Column	Aggregation (click to change)	Missing
I WebActivity	Mean	<input type="checkbox"/>
I Age	Mean	<input type="checkbox"/>
I Age	Minimum	<input type="checkbox"/>
I Age	Maximum	<input type="checkbox"/>
I CustomerKey	Count	<input type="checkbox"/>

Figura 17 - Configuração Group By

I Count*(CustomerKey)	First	<input type="checkbox"/>
I Count*(CustomerKey)	Percent	<input type="checkbox"/>

Figura 18 - Configuração 2º Group By

b)

Row ID	I WebAc...	S Gender	S Mode(S...	D Mean(S...
Row0	0	F	Very Negative	1.331
Row1	0	M	Very Negative	1.308
Row2	1	F	Negative	1.869
Row3	1	M	Negative	1.914
Row4	2	F	Slightly Neg...	2.689
Row5	2	M	Slightly Neg...	2.605
Row6	3	F	Slightly Posit...	2.751
Row7	3	M	Slightly Posit...	2.801
Row8	4	F	Positive	3.665
Row9	4	M	Positive	3.709
Row10	5	F	Very Positive	3.307
Row11	5	M	Very Positive	3.266

Figura 19 - Tabela resultante

Column	Aggregation (click to change)	Missing
S Sentiment Analysis	Mode	<input checked="" type="checkbox"/>
I SentimentRating	Mean	<input type="checkbox"/>

Figura 20 - Configuração Group By

c)

Row ID	S Sentim...	I Count*...	D Mean(E...	I Sum(Es...	D Mean(N...
Row0	Negative	3122	51,287.636	160120000	0.714
Row1	Positive	1960	68,801.02	134850000	1.749
Row2	Slightly Neg...	3023	57,380.086	173460000	1.527
Row3	Slightly Posit...	1690	62,295.858	105280000	1.792
Row4	Very Negative	4173	51,603.163	215340000	1.448
Row5	Very Positive	1199	72,026.689	86360000	2.403

Figura 21 - Tabela resultante

Column	Aggregation (click to change)	Missing
I CustomerKey	Count	<input type="checkbox"/>
I EstimatedYearlyIncome	Mean	<input type="checkbox"/>
I EstimatedYearlyIncome	Sum	<input type="checkbox"/>
I NumberOfContracts	Mean	<input type="checkbox"/>

Figura 22 - Configuração Group By

T5

Na alínea a) retiramos que os géneros se encontram bem distribuídos, com uma média de horas de atividade web muito próximas, bem como média, máximo e mínimo de idades e o número de registos.

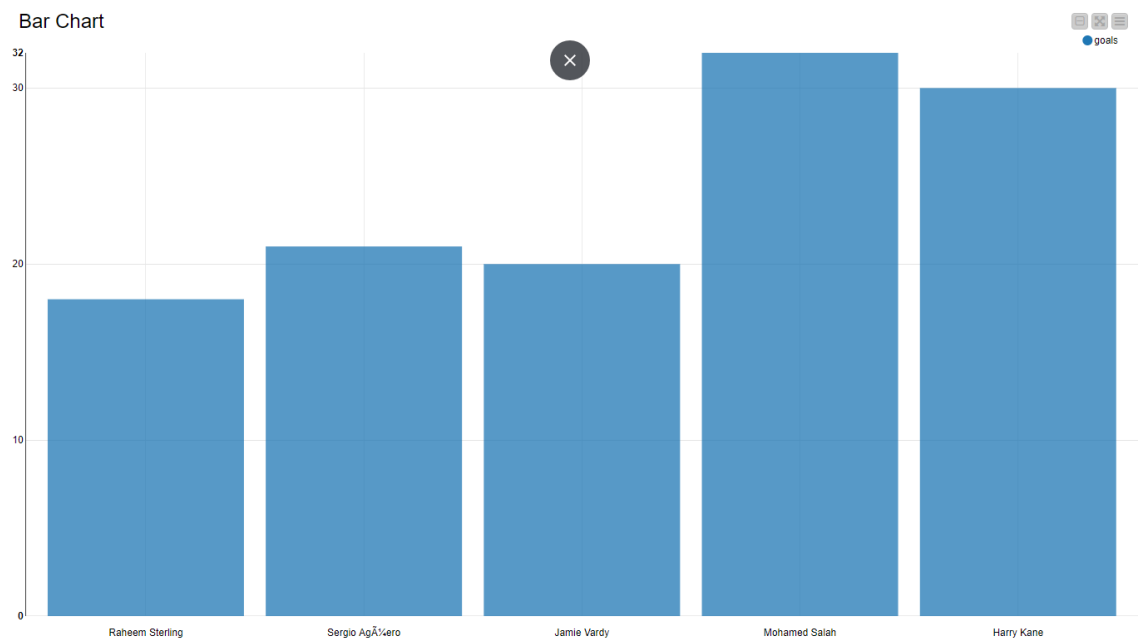
Os dados da alínea b) são mais valiosos uma vez que sabemos através da alínea a) que os géneros se encontram bem distribuídos. Olhando então para a tabela resultante cheguei à conclusão que existe uma razão de proporcionalidade direta entre o nível de satisfação dos clientes e as horas de atividade web. Sendo que quanto maior for a atividade maior vai ser a satisfação do cliente. Podemos concluir também que esta razão é completamente independente dos géneros uma vez que os valores da moda e média da satisfação sofrem praticamente as mesmas variações nos dois géneros.

Já da tabela da alínea c) podemos concluir que o rendimento dos clientes está diretamente relacionado com o número de contratos que fazem e com o seu nível de satisfação. Após a análise da tabela é conclusivo que maiores rendimentos levam a um maior número de contratos e a clientes mais satisfeitos, e por outro lado menores rendimentos levam a um menor número de contratos e a uma menor satisfação.

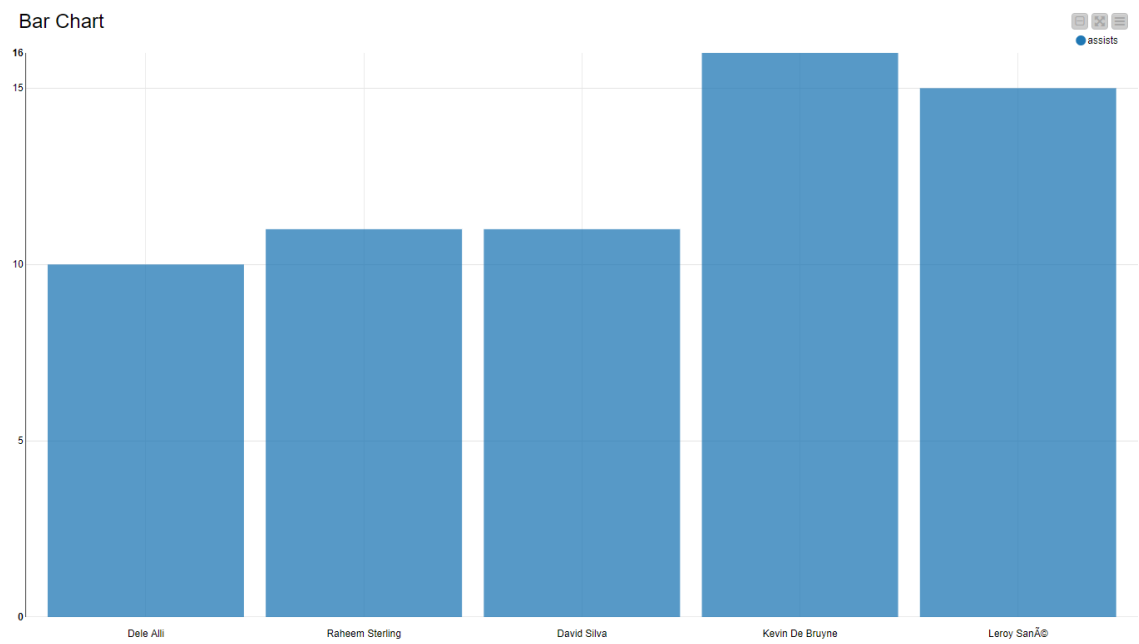
Após a análise da informação extraída das agregações a empresa pode mudar a sua estratégia de mercado tendo em conta não só os seus clientes alvo que são pessoas com altos rendimentos anuais, mas também como manter os seus clientes satisfeitos o que passa por aumentar as horas de atividade web dos mesmos.

T6

Melhores marcadores:

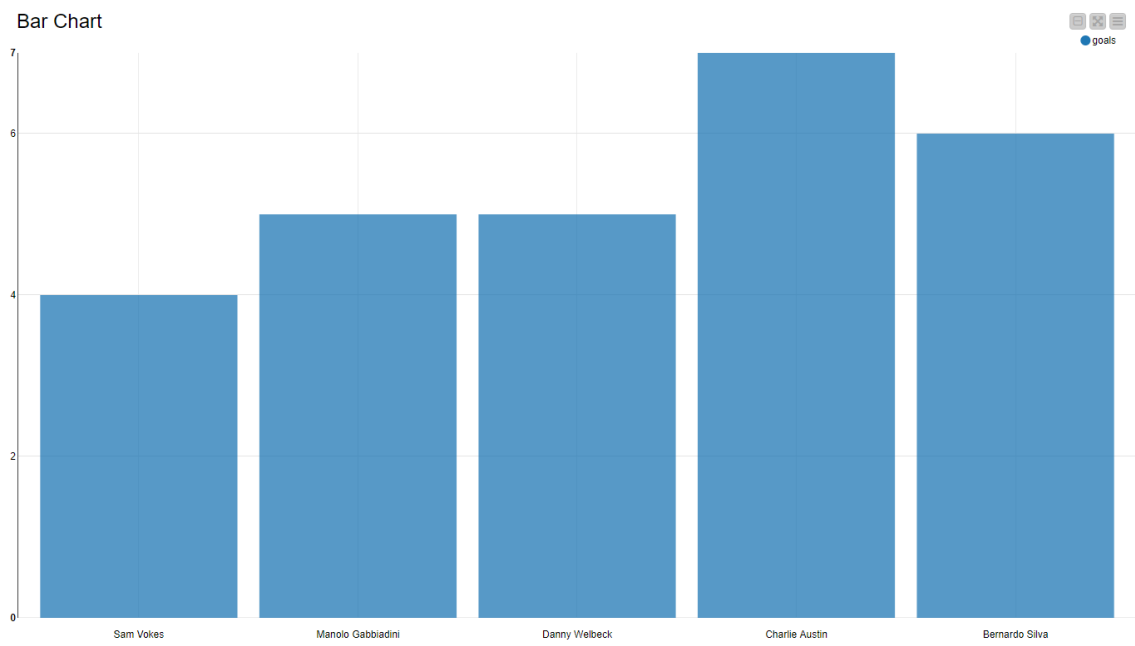


Mais assistências:

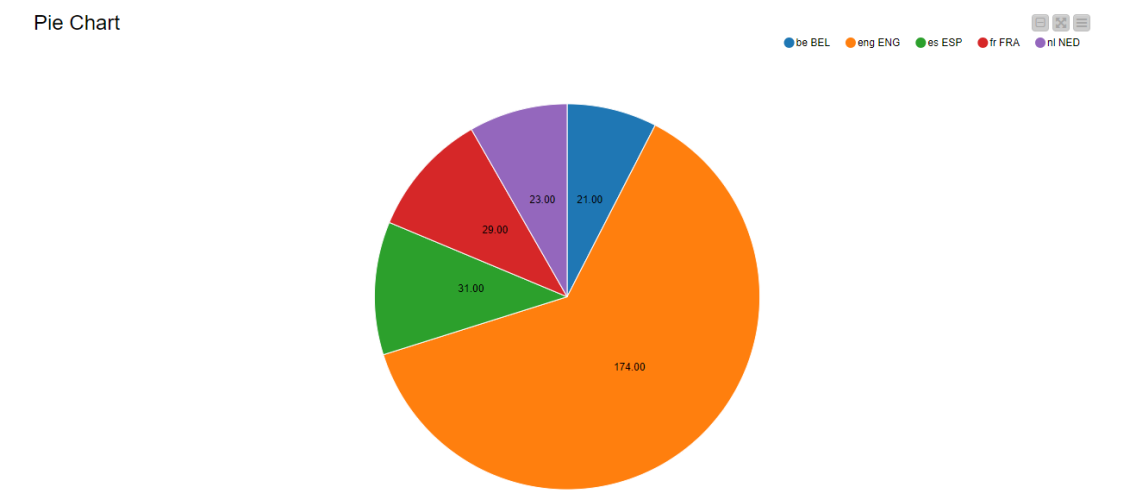


Super subs:

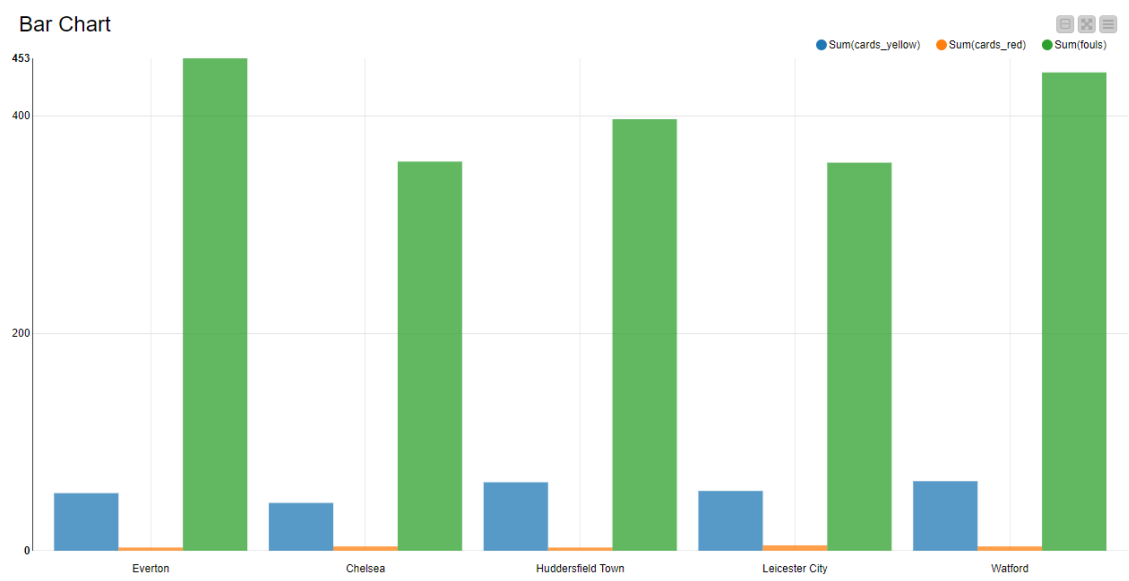
Os jogadores nesta tabela são os jogadores com mais golos marcados na liga e um tempo médio de jogo inferior a 45 minutos.



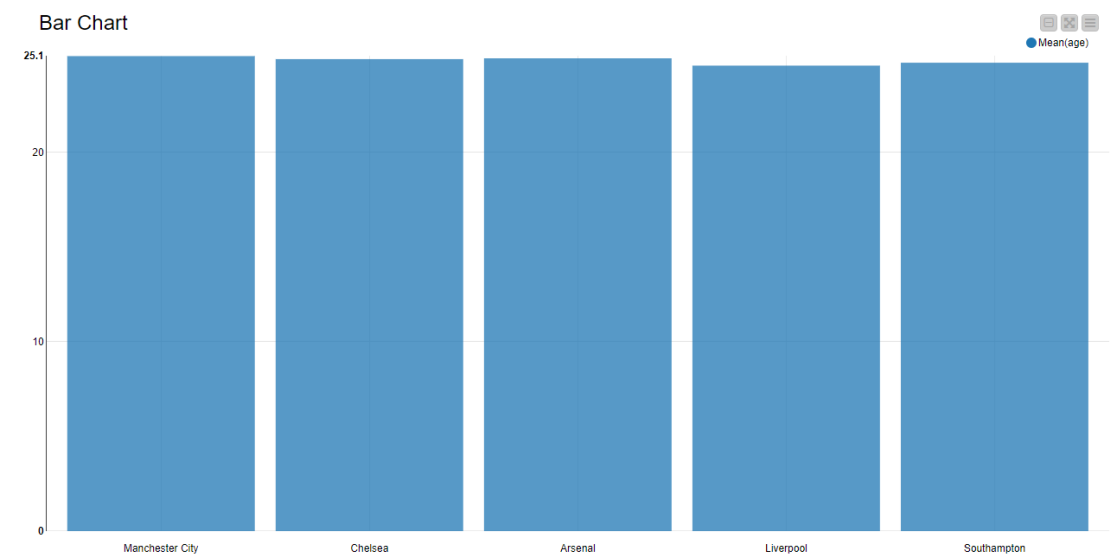
Top nacionalidades e a percentagem dos tops:



Equipas mais indisciplinadas:

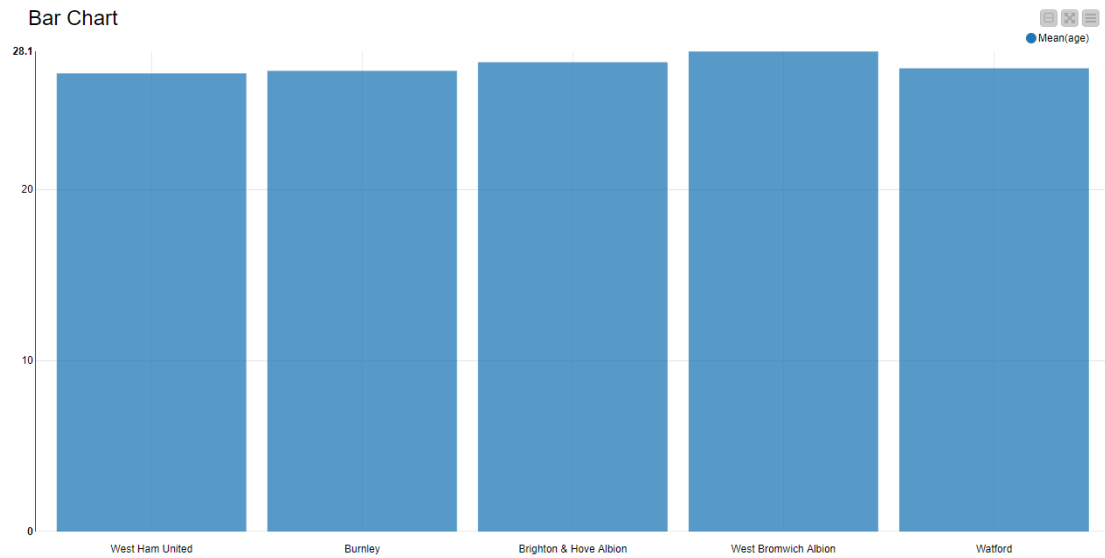


Planteis mais jovens:

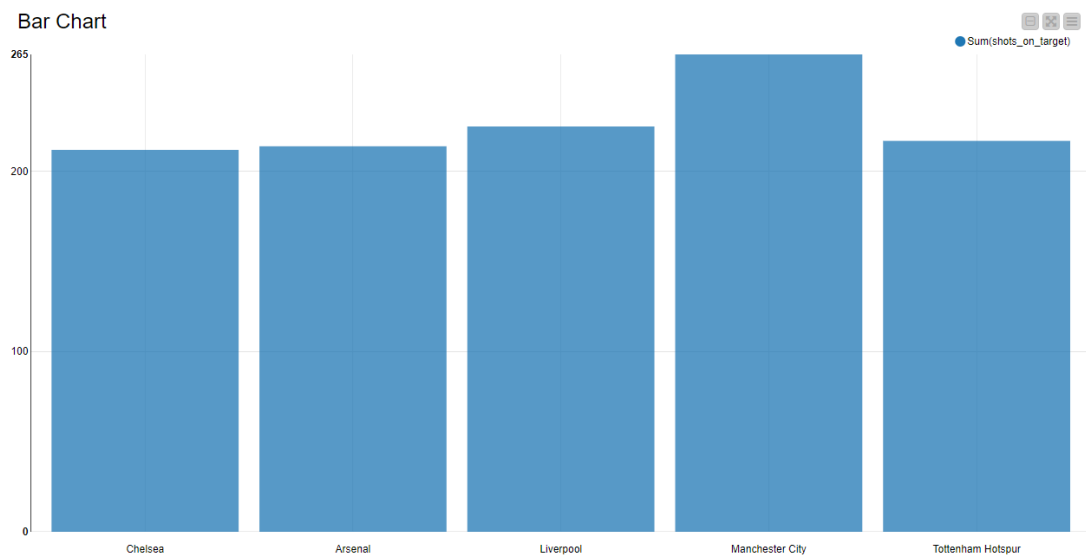


Planteis mais velhos:

Equipas com a idade média do plantel mais alta.



Mais remates à baliza:



Comparando às qualificações finais do campeonato é possível verificar que as 5 equipas mais rematadoras ficaram todas no top 6 da qualificação. Podemos por isso concluir que equipas com estratégias que resultem em vários remates à baliza têm bons resultados e por consequente a boas qualificações.