

Enunciado Prático nº 7

Maria José Borges Pires - A86268

16 de dezembro de 2020

1 Parte 1

1.1 Exercício 1

Para obter todas as cidades portuguesas atualmente disponibilizadas pela Open AQ utilizou-se o nodo *GET request*.

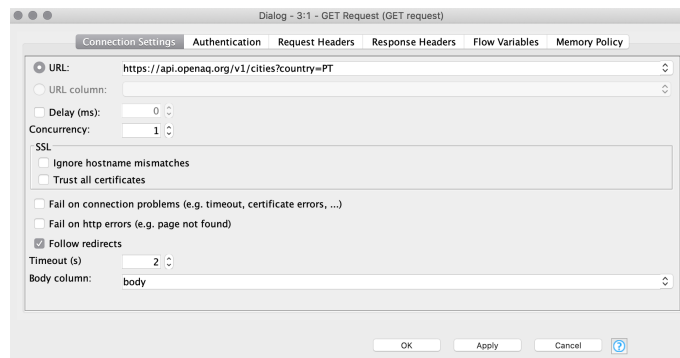


Figure 1: *GET Request*

1.2 Exercício 2

Transformação do JSON obtido no exercício anterior, garantindo a expansão dos arrays JSON para colunas, e o parâmetro *children expansion* até ao nível 2.

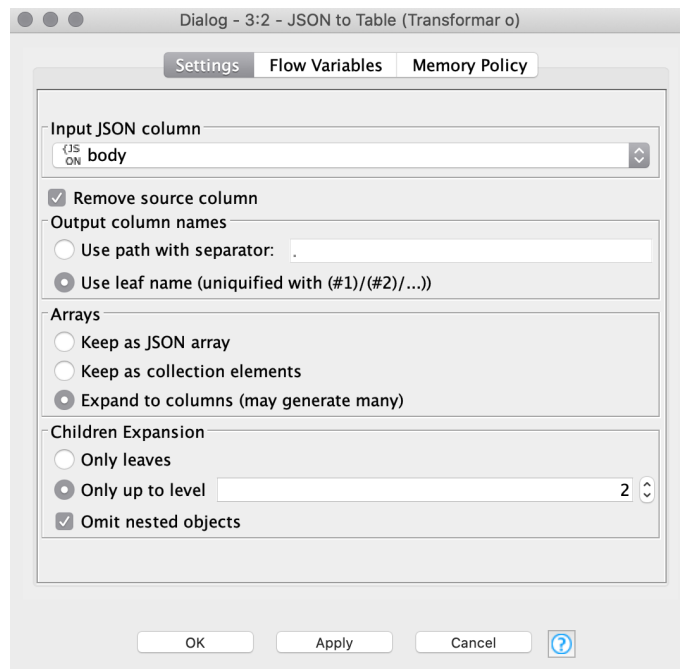


Figure 2: *JSON to Table*

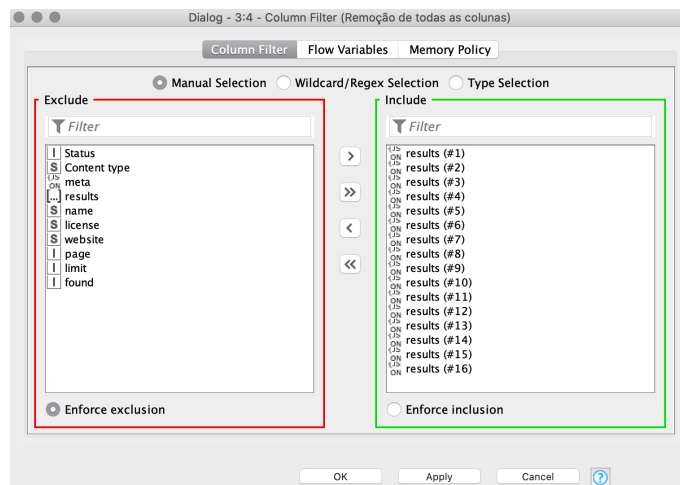


Figure 3: Remoção todas as colunas exceto resultados

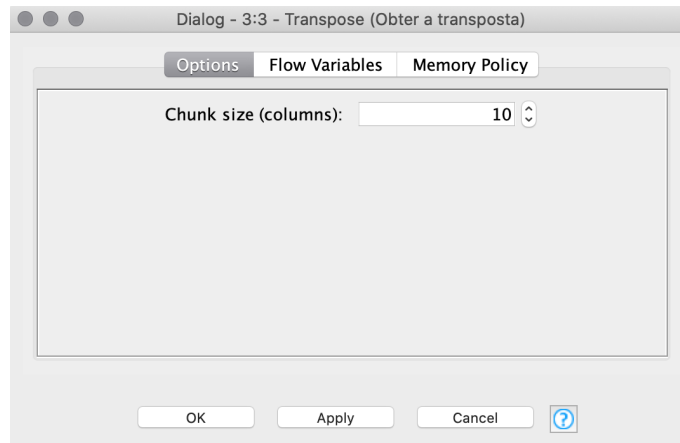


Figure 4: Transposta da tabela

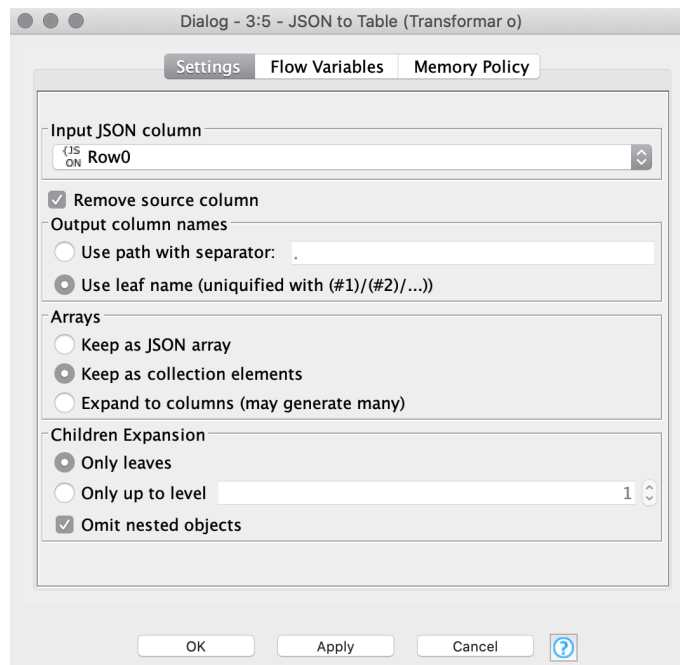


Figure 5: *JSON to Table*

Row ID	S country	S name	S city	I count	I locations
results (#1)	PT	Aveiro	Aveiro	374356	4
results (#2)	PT	Braga	Braga	146922	3
results (#3)	PT	Castelo Branco	Castelo Branco	118875	1
results (#4)	PT	Coimbra	Coimbra	173288	3
results (#5)	PT	Évora	Évora	126089	1
results (#6)	PT	Faro	Faro	367187	4
results (#7)	PT	Ilha da Madeira	Ilha da Madeira	310065	3
results (#8)	PT	Ilha do Faial	Ilha do Faial	139499	1
results (#9)	PT	Leiria	Leiria	132490	1
results (#10)	PT	Lisboa	Lisboa	1259717	15
results (#11)	PT	Porto	Porto	895499	15
results (#12)	PT	Santarém	Santarém	110873	1
results (#13)	PT	Setúbal	Setúbal	1187860	12
results (#14)	PT	Viana do Castelo	Viana do Castelo	71592	1
results (#15)	PT	Vila Real	Vila Real	106098	1
results (#16)	PT	Viseu	Viseu	80670	1

Figure 6: Tabela obtida

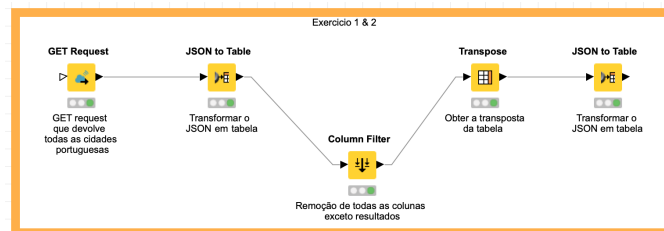


Figure 7: Fluxo após a resolução dos dois primeiros exercícios

1.3 Exercício 3

Para obter os dados mais recentes sobre o nível de ozono de cada cidade portuguesa recorreu-se ao nodo *Get Request*.

Aplicando-se o url: <https://api.openaq.org/v1/latest?country=PT¶meter=o3>

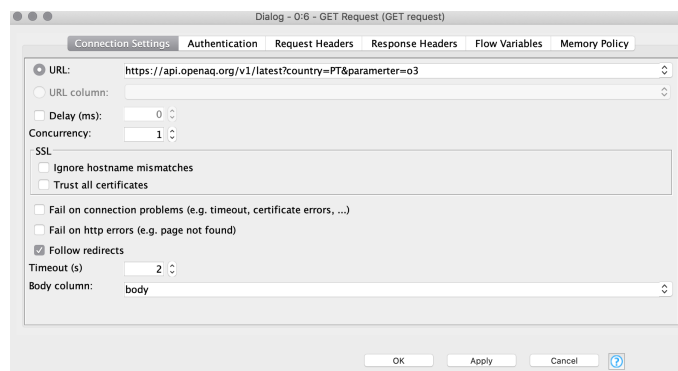


Figure 8: *Get Request*

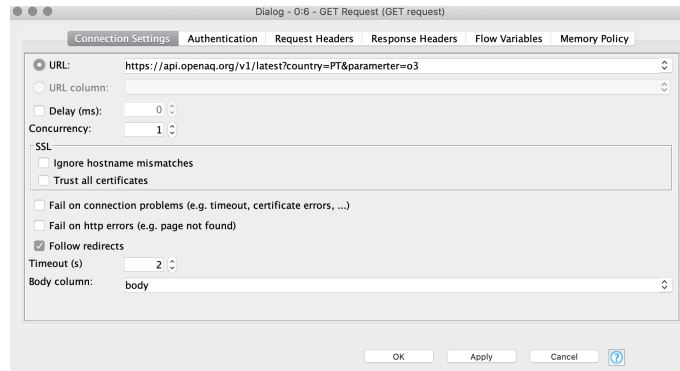


Figure 9: Fluxo de tratamento do JSON

De seguida foram filtradas as filas cujo *parameter* não é referente ao nível de ozono através de um *Row Filter*. Com o intuito de obter apenas um registo para cada cidade aplicou-se o nodo *Sorter* para organizar os dados em função da *feature lastUpdated* e de seguida foi aplicado um *Column Filter*.

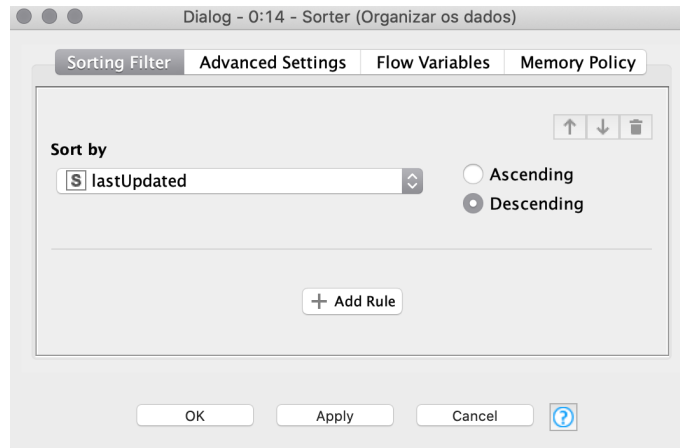


Figure 10: *Sorter*

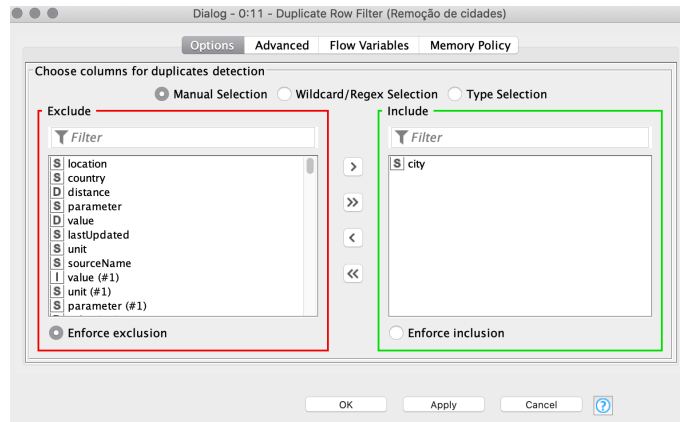


Figure 11: *Column Filter*

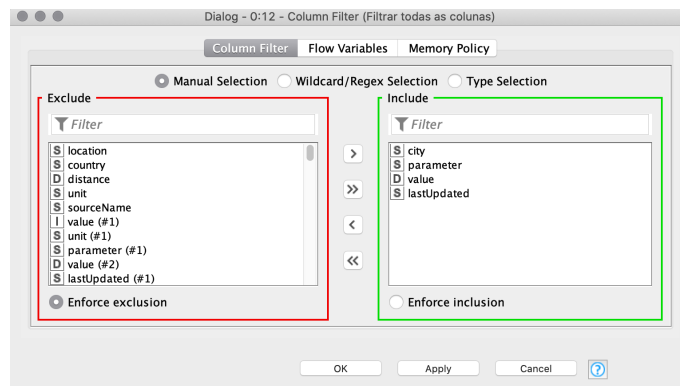


Figure 12: Filtrar todas as colunas exceto a city, parameter, value e lastUpdated

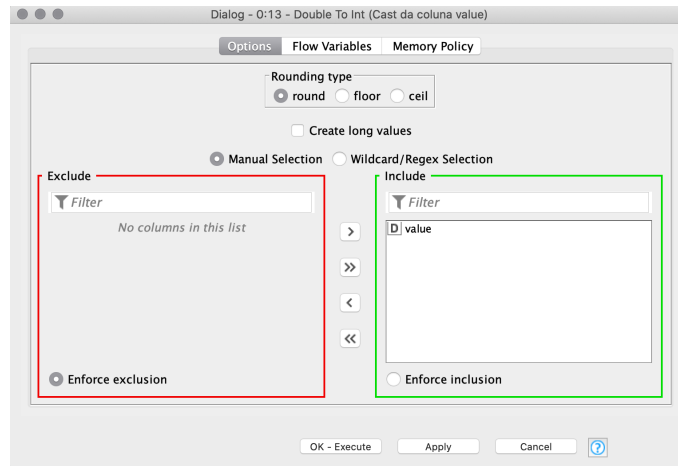


Figure 13: Cast da coluna value de double para inteiro

Row ID	S city	S parameter	D value	S lastUpdated
results (#29)	Viseu	o3	87	2020-12-03T03:00:00.000Z
results (#40)	Lisboa	o3	43	2020-12-03T03:00:00.000Z
results (#47)	Santarém	o3	66	2020-12-03T03:00:00.000Z
results (#8)	Porto	o3	16	2020-12-03T02:00:00.000Z
results (#55)	Setúbal	o3	61	2020-02-19T15:00:00.000Z

Figure 14: Tabela obtida no final do exercício

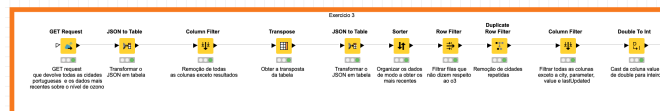


Figure 15: Fluxo após a resolução do terceiro exercício

1.4 Exercício 4

Apresenta-se na seguinte figura as definições aplicadas ao nodo *Sorter* com o objetivo de ordenar os registos de cada cidade por nível de ozono.

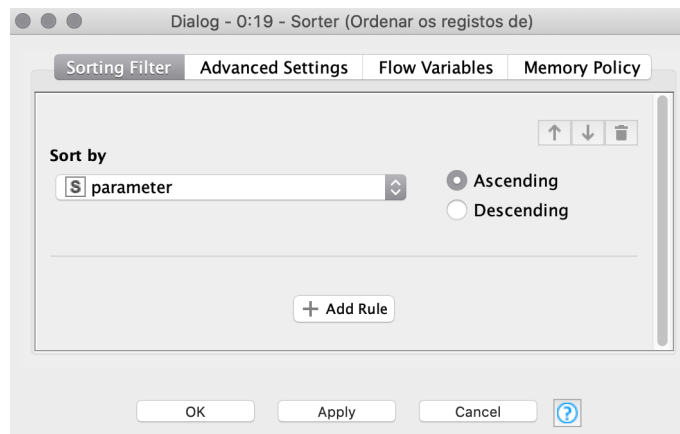


Figure 16: Ordenar os registos de cada cidade por nível de ozono

As técnicas de visualização de dados numa vista web aplicadas apresentam-se de seguida:

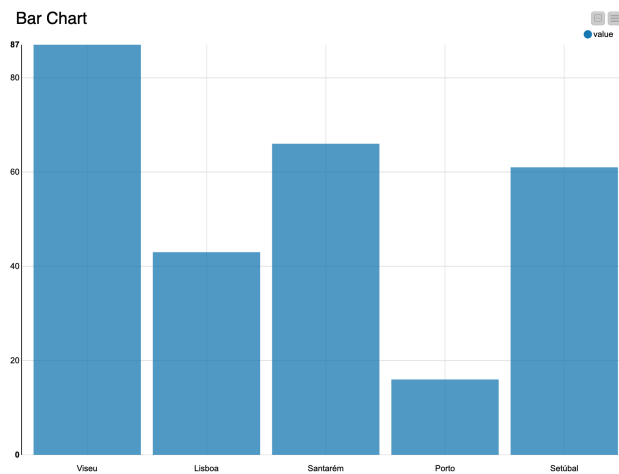


Figure 17: *Barchart* da média do *value* por cidade

Visualização de dados

Search:

<input type="checkbox"/>	RowID	city	parameter	value
<input type="checkbox"/>	results (#29)	Viseu	o3	87
<input type="checkbox"/>	results (#40)	Lisboa	o3	43
<input type="checkbox"/>	results (#47)	Santarém	o3	66
<input type="checkbox"/>	results (#8)	Porto	o3	16
<input type="checkbox"/>	results (#55)	Setúbal	o3	61

Showing 1 to 5 of 5 entries

Figure 18: *Table View*

Visualização de dados

Viseu parameter: o3 value: 87 lastUpdated: 2020-12-03T03:00:00.000Z <input type="checkbox"/>	Lisboa parameter: o3 value: 43 lastUpdated: 2020-12-03T03:00:00.000Z <input type="checkbox"/>	Santarém parameter: o3 value: 66 lastUpdated: 2020-12-03T03:00:00.000Z <input type="checkbox"/>
Porto parameter: o3 value: 16 lastUpdated: 2020-12-03T02:00:00.000Z <input type="checkbox"/>	Setúbal parameter: o3 value: 61 lastUpdated: 2020-02-19T15:00:00.000Z <input type="checkbox"/>	

Showing 1 to 5 of 5 entries

Previous 1 Next

Figure 19: *Tile View*

Para observar dados referentes a outros parâmetros ambientais, após obter os dados em forma de tabela ordenados através da *feature lastUpdated*, no exercício 3, aplicou-se o nodo *Duplicate Row Filter* de modo a remover registos de cidades repetidos para o mesmo parâmetro.

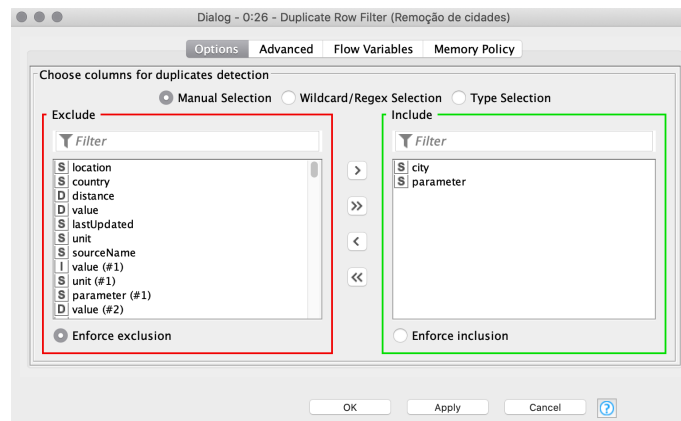


Figure 20: *Duplicate Row Filter*

Após a filtragem de algumas colunas cujos valores não são tão interessantes

e aplicação do nodo *Tile View* pode observar-se na figura seguinte os dados de vários parâmetros para as várias cidades portuguesas.

Dados referentes a outros parâmetros ambientais

Porto location: PT01030 parameter: pm10 value: 14 lastUpdated: 2020-12-03T03:00:00.000Z	Braga location: PT01046 parameter: pm10 value: 12 lastUpdated: 2020-12-03T03:00:00.000Z	Coimbra location: PT02006 parameter: co value: 140 lastUpdated: 2020-12-03T03:00:00.000Z	Viseu location: PT02021 parameter: o3 value: 87 lastUpdated: 2020-12-03T03:00:00.000Z
Setúbal location: PT03055 parameter: no2 value: 3.7 lastUpdated: 2020-12-03T03:00:00.000Z	Lisboa location: PT03070 parameter: no2 value: 4.9 lastUpdated: 2020-12-03T03:00:00.000Z	Santarém location: PT03096 parameter: o3 value: 66 lastUpdated: 2020-12-03T03:00:00.000Z	Évora location: PT04006 parameter: no2 value: 4 lastUpdated: 2020-12-03T03:00:00.000Z
Faro location: PT05008 parameter: no2 value: 3.7 lastUpdated: 2020-12-03T03:00:00.000Z	Ilha da Madeira location: PT06004 parameter: pm25 value: 12.6 lastUpdated: 2020-12-03T03:00:00.000Z		

Showing 1 to 10 of 16 entries

Previous 1 2 Next

Figure 21: *Tile View* de todos os parâmetros ambientais existentes

1.5 Exercício 5

Para utilizar a *OpenWeatherMaps* é necessário aplicar a API key ao url a inserir no nodo *GET Request*. O url inserido permite obter o estado meteorológico atual da cidade de Braga.

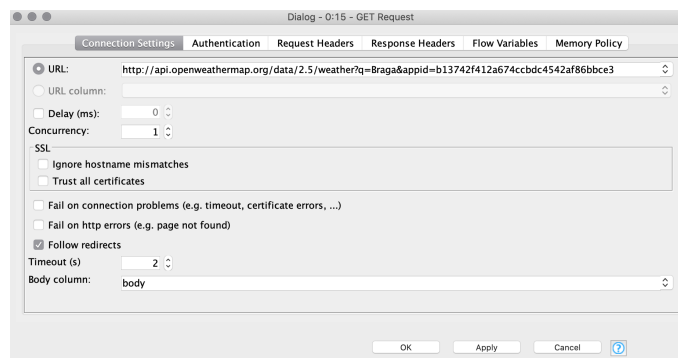


Figure 22: *GET Request*

Row ID	Row0
name	Braga
cod	200
lon	-8.42
lat	41.55
temp	280.93
feels_like	279.78
temp_min	280.93
temp_max	280.93
pressure	1013
humidity	91
speed	0.45
deg	244
gust	3.58
all	100
type	3
country	PT

Figure 23: Condições meteorológicas na cidade de Braga

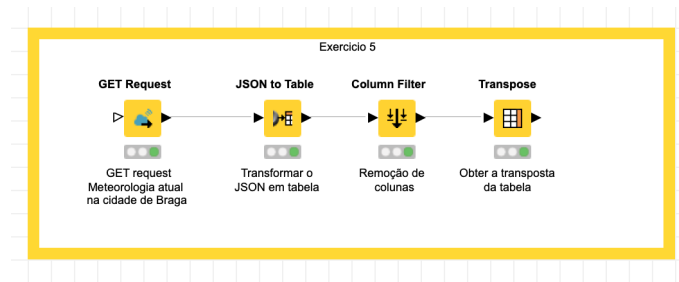


Figure 24: Fluxo do exercício 5

2 Parte 2

2.1 Exercício 6

O conjunto de dados escolhido contém dados sobre mais de 1700 barras de chocolate diferentes, tal como a sua classificação, a sua região de origem, percentagem de cacão, entre outros valores que podem ser observados na figura 25.

[S] Company ...	[S] Specific Bean Origin or Bar Name	[I] REF	[I] Review Date	[S] Cocoa Percent	[D] Rating	[S] Company Location	[S] Bean Type	[S] Broad Bean Origin
A. Morin	Agua Grande	1876	2016	63%	3.75	France		Sao Tome
A. Morin	Kpime	1676	2015	70%	2.75	France		Togo
A. Morin	Panama	1011	2013	70%	2.75	France		Panama
Arete	Kokoa Kamili	1724	2016	70%	3.75	U.S.A.		Tanzania
Madre	Criollo, Hawaii	995	2012	70%	3.25	U.S.A.	Criollo	Hawaii
Madre	Kaua'i	995	2012	70%	3.5	U.S.A.		Hawaii
Madre	Dominican	672	2011	70%	2.5	U.S.A.		Dominican Republic
Madre	Upala	693	2011	70%	2.75	U.S.A.		Costa Rica
Madre	Chiapas, Triple Cacao	607	2010	72%	2.75	U.S.A.		Mexico
Maglio	Africa	300	2008	75%	2	Italy		Tanzania
Maglio	Ecuador	308	2008	70%	3	Italy	Forastero (Na...	Ecuador
Maglio	Cuba	308	2008	70%	3.25	Italy	Criollo	Cuba
Maglio	Santo Domingo	308	2008	70%	3.75	Italy	Blend-Foraste...	Dominican Republic

Figure 25: Excerto do conjunto de dados a tratar

Verificou-se, através do nodo *Statistics* que as colunas *Bean Type* e *Broad Bean Origin* apresentam alguns valores em falta, posto isto, aplicou-se o nodo *Missing Values* com o objetivo de preencher como desconhecido (*Unknown*) as células com valores em falta.

2.2 Exercício 7

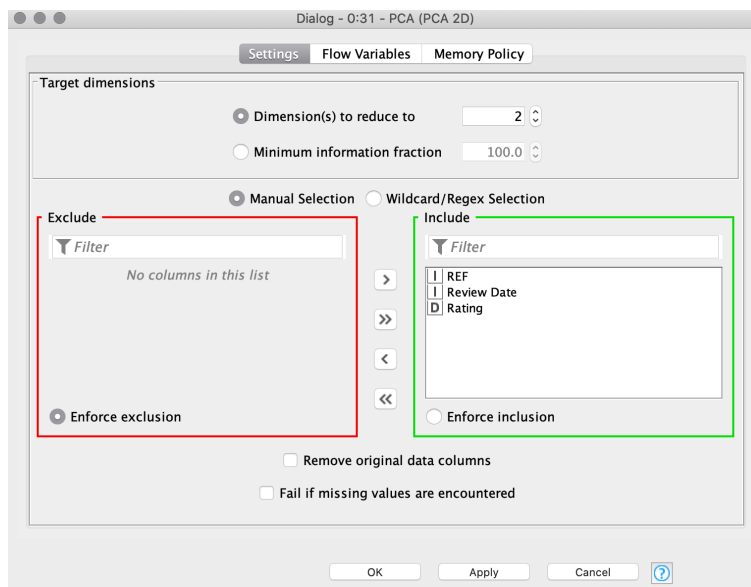


Figure 26: Análise de Componentes Principais (PCA)

☐ Create image at output

Maximum number of rows: 2,500

Selection column name: Selected (Scatter Plot)

Choose column for x axis
PCA dimension 0

Choose column for y axis
PCA dimension 1

☒ Report on missing values

Figure 27: Definições aplicadas ao nodo *Scatter plot*

Através do *Scatter plot* conseguimos identificar (figura 28), 12 clusters presentes no *dataset*.

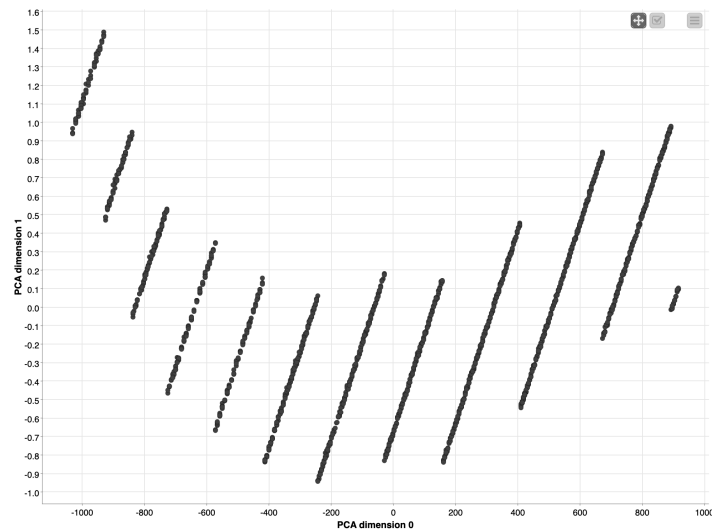


Figure 28: *Scatter plot*

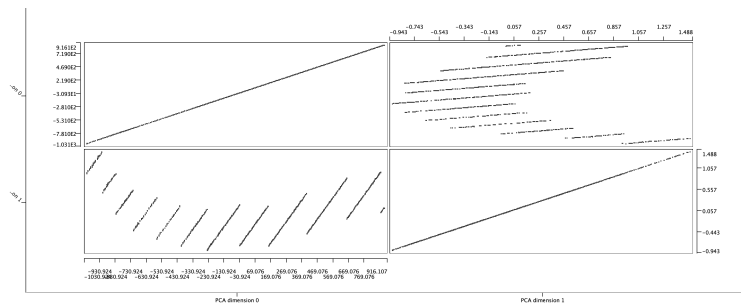


Figure 29: *Scatter Matrix*

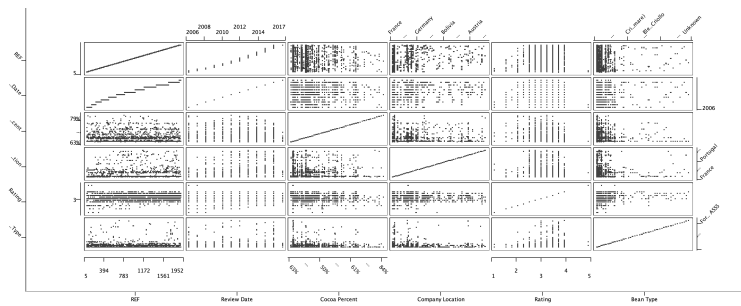


Figure 30: *Scatter Matrix*

2.3 Exercício 8

Distance Selection:

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

No columns in this list

☒ Enforce exclusion

Include

☒ REF

☒ Review Date

☒ Rating

☐ Enforce inclusion

Append Column Name

Chunk Size

Figure 31: Definições aplicadas à matriz de distancia

Dialog - 0:36 - k-Medoids

Default Flow Variables Memory Policy

Distance Matrix Column ↔ Distance ↕

Partition count (k)

Chunk size

☐ Constrain no. iterations

☒ Use static seed

☐ Output relative distances to medoids

☒ Choke on asymmetric distances

The "k" parameter is controlled by a variable.

Figure 32: Definições aplicadas ao nodo *K-medoids*

Medoids and Size - 0:36 - k-Medoids

Table "dcData" Rows: 10 Spec - Columns: 11 Properties Flow Variables

Row ID	Company (Maker-if known)	Specific Bean Origin or Bar Name	REF	Rev.	Cocoa Percent	Comp.	Rating	Bean Type	Broad...	Δs Dista...	partitionSize
Row1266	Parliament	Kilombero Valley	1856	2016	70%	U.S.A.	3.25			1266 [2...	195
Row1281	Pierre Marcolini	Haut Perig., w/ nibs	1658	2015	70%	Belgium	3.25	Forastero		1281 [2...	197
Row76	Amedei	Pococana	111	2007	70%	Italy	4	Criollo (Perlatino)		76 [17...	152
Row997	Madre	Chobua, Kona	1089	2013	70%	U.S.A.	2.75			997 [78...	184
Row1417	Salgado	Bahia Superior	2188	2008	70%	Argentina	3.5	Forastero		1417 [1...	145
Row637	Ethelal	Ecuador	1275	2014	80%	U.S.A.	3.5			637 [60...	183
Row1221	Onasheni	Chana	693	2011	80%	China	2.75	Forastero		1221 [1...	193
Row1283	Pierre Marcolini	Trinit	478	2010	75%	Belgium	3.25	Trinitario		1283 [1...	154
Row1154	Muchomas (Mesocacao)	Nicaragua	1462	2015	70%	U.S.A.	3.5			1154 [4...	211
Row712	Fris Holm (Borlat)	Red Mayan, Xoco	899	2012	70%	Denmark	3.25	Criollo, Trinitario		712 [97...	181

Figure 33: *Medoids and size*



Figure 34: Metanodo *MAE*

Group table - 0:55:51 - GroupBy (MAE)	
File Edit Hilite Navigation View	
Table "default" - Rows: 1 Spec - Column: 1 Properties	
Row ID	D Mean(MAE)
Row0	0.457

Figure 35: Mean Average Error (MAE)

Row ID	D MAE	I k	I Iteration
Row0#0	0.376	1	0
Row0#1	0.376	2	1
Row0#2	0.391	3	2
Row0#3	0.392	4	3
Row0#4	0.379	5	4
Row0#5	0.388	6	5
Row0#6	0.388	7	6
Row0#7	0.376	8	7
Row0#8	0.401	9	8
Row0#9	0.457	10	9

Figure 36: MAE de cada cluster

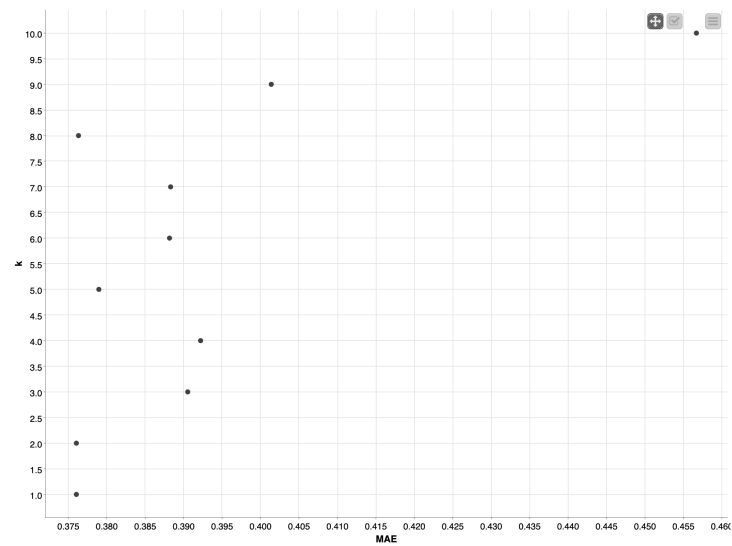


Figure 37: *Scatter Plot*

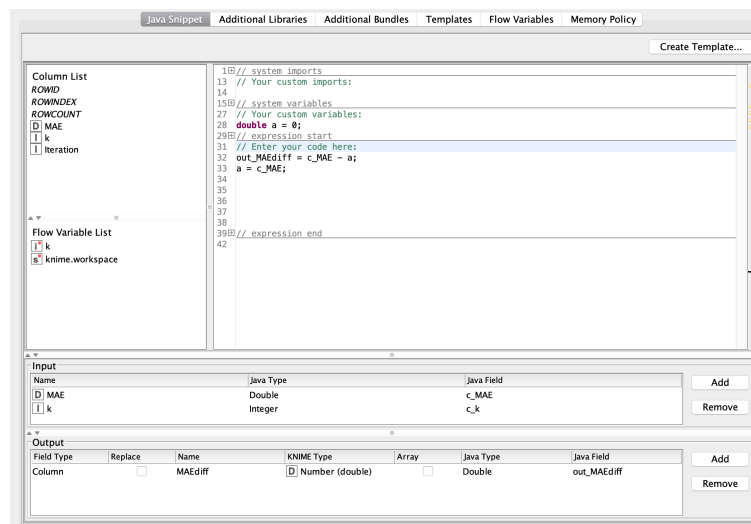


Figure 38: *Java Snippet para o calculo do MAE de cada cluster*

Row ID	MAE	k	Iteration	MAEdiff
Row0#0	0.376	1	0	0.376
Row0#9	0.457	10	9	0.055
Row0#8	0.401	9	8	0.025
Row0#2	0.391	3	2	0.014
Row0#5	0.388	6	5	0.009
Row0#3	0.392	4	3	0.002
Row0#6	0.388	7	6	0
Row0#1	0.376	2	1	0
Row0#7	0.376	8	7	-0.012
Row0#4	0.379	5	4	-0.013

Figure 39

Row ID	MAE	k	Iteration	MAEdiff
Row0#0	0.376	1	0	0.376

Figure 40: MAE

Após esta análise podemos concluir que o número ótimo de clusters é 3.

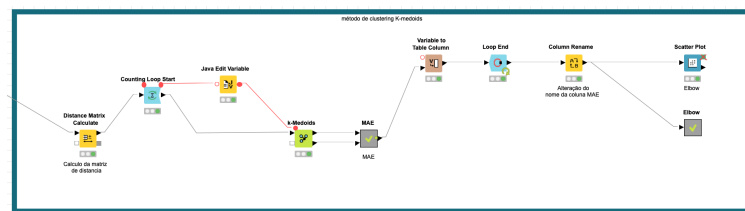


Figure 41: Fluxo final do exercício 8

2.4 Exercício 9

Para que um utilizador consiga analisar os gráficos gerados pelo método do cotovelo aplicou-se um *Scatter plot*.

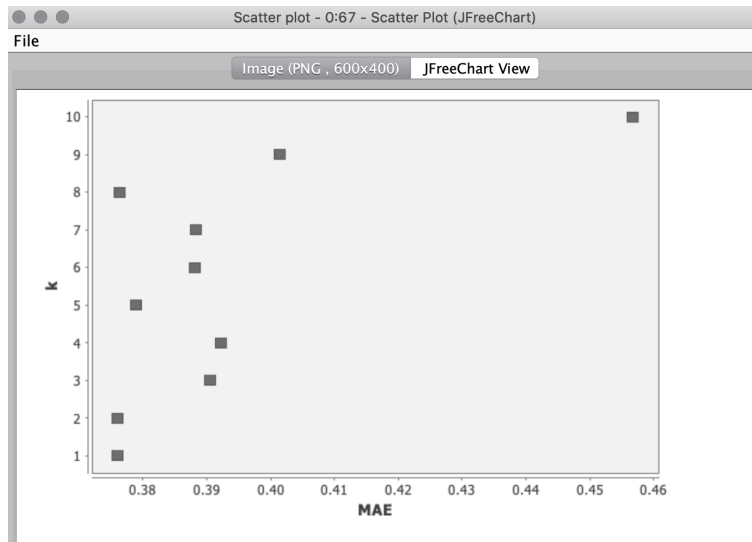


Figure 42: Scatter Plot

Dialog - 0:69 - Integer Configuration

Control Flow Variables

Label: n_clusters

Description: number of clusters

Parameter/Variable Name: n_clusters

Minimum: ☐ 0

Maximum: ☐ 100

Default Value: 0

OK Apply Cancel ?

Figure 43: Widget para definir o número de clusters a utilizar

2.5 Exercício 10

Row ID	D REF	D Review Date	D Rating
cluster_0	1,735.114	2,015.671	3.154
cluster_1	1,428.822	2,014.454	3.224
cluster_2	1,063.978	2,012.801	3.218
cluster_3	138.144	2,006.906	3.084
cluster_4	1,584	2,015	3.283
cluster_5	389.569	2,009.005	3.105
cluster_6	862.408	2,011.827	3.175
cluster_7	1,250.326	2,013.807	3.157
cluster_8	1,881.51	2,016.168	3.306
cluster_9	645.817	2,010.653	3.205

Figure 44: Clusters obtidos após a aplicação do nodo *K-means*

Row ID	D Mean(MAE)
Row0	0.375

Figure 45: Mean Average Error (MAE)

Row ID	D MAE	I k	I Iteration
Row0#0	0.384	1	0
Row0#1	0.38	2	1
Row0#2	0.377	3	2
Row0#3	0.376	4	3
Row0#4	0.376	5	4
Row0#5	0.375	6	5
Row0#6	0.376	7	6
Row0#7	0.374	8	7
Row0#8	0.376	9	8
Row0#9	0.375	10	9

Figure 46: MAE para cada cluster no final do ciclo

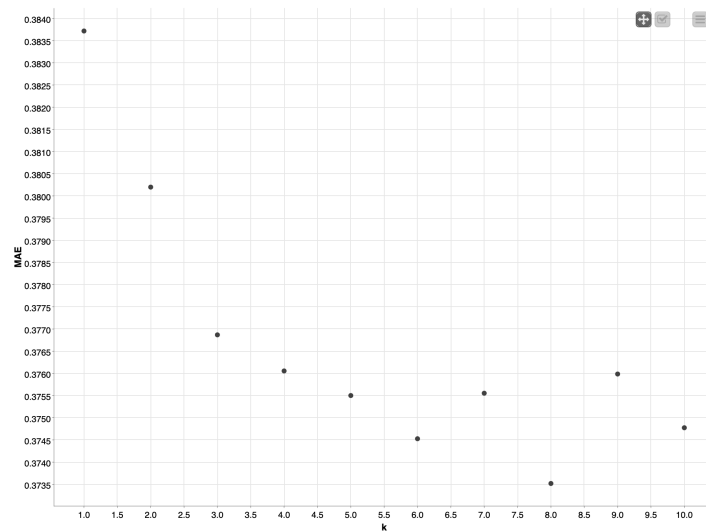


Figure 47: *Scatter Plot*

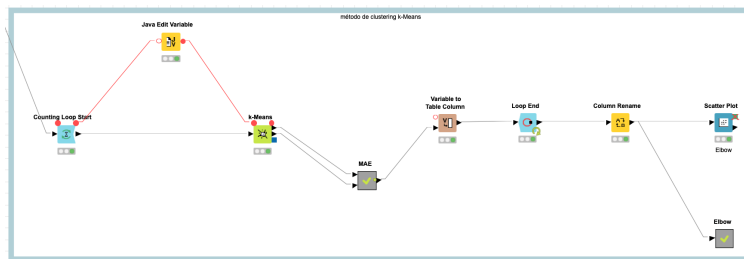


Figure 48: Fluxo final do exercício 10

Após a análise dos dois métodos conclui-se que ambos têm um número de clusters ótimo igual a 3, contudo a métrica utilizada como medida de qualidade, MAE, apresenta um valor ligeiramente mais baixo utilizando k-means.