

Enunciado Prático nº4

Maria José Borges Pires - A86268

11 de novembro de 2020

1 Exercício 1

Para a resolução do exercício 1, após a leitura do *dataset* de teste através do nodo *File Reader* é feita a exploração dos dados recorrendo ao nodo *Data Explorer*, onde se podem observar os valores da tendência central e dispersão estatística dos dados carregados.

Row #	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
Row0	7.6	0.7	0	1.9	0.076	11	34	0.998	3.51	0.56	9.4	-5
Row1	7.8	0.88	0	2.6	0.098	25	67	0.997	3.2	0.68	9.8	-5
Row2	7.8	0.79	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	-5
Row3	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	-6
Row4	7.4	0.7	0	1.9	0.076	11	34	0.998	3.51	0.56	9.4	-5
Row5	7.4	0.66	0	1.8	0.075	13	40	0.998	3.51	0.56	9.4	-5
Row6	7.8	0.6	0.06	1.6	0.069	15	59	0.996	3.3	0.46	9.4	-5
Row7	7.3	0.65	0	1.2	0.065	15	21	0.995	3.39	0.47	10	-7
Row8	7.8	0.58	0.02	2	0.073	9	18	0.997	3.36	0.57	9.5	-7
Row9	7.5	0.5	0.36	6.1	0.071	17	102	0.998	3.35	0.8	10.5	-5
Row10	6.7	0.53	0.08	1.8	0.067	15	65	0.996	3.28	0.54	9.2	-5
Row11	7.5	0.5	0.36	6.1	0.071	17	102	0.998	3.35	0.8	10.5	-5
Row12	5.6	0.615	0	1.6	0.069	16	59	0.994	3.58	0.52	9.9	-5
Row13	7.8	0.61	0.29	1.6	0.114	9	29	0.997	3.26	1.56	9.1	-5
Row14	8.9	0.62	0.18	3.8	0.176	52	145	0.999	3.16	0.88	9.2	-5
Row15	8.9	0.62	0.19	3.9	0.17	51	148	0.999	3.17	0.93	9.2	-5
Row16	8.5	0.28	0.56	1.8	0.092	35	103	0.997	3.3	0.75	10.5	-7
Row17	8.1	0.56	0.28	1.7	0.368	16	56	0.997	3.11	1.26	9.3	-5
Row18	7.4	0.59	0.08	4.4	0.086	6	29	0.997	3.38	0.5	9	-4
Row19	7.9	0.32	0.51	1.8	0.341	17	55	0.997	3.04	1.08	9.2	-6
Row20	8.9	0.22	0.48	1.8	0.077	29	60	0.997	3.39	0.53	9.4	-6
Row21	7.6	0.39	0.31	2.3	0.082	23	71	0.998	3.52	0.65	9.7	-5
Row22	7.9	0.43	0.21	1.6	0.106	10	37	0.997	3.17	0.91	9.5	-5

Figure 1: Excerto da tabela de dados carregados

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness	Kurtosis
fixed acidity	<input type="checkbox"/>	4.600	15.900	8.416	1.742	3.035	0.940	1.084
volatile acidity	<input type="checkbox"/>	0.120	1.580	0.526	0.180	0.032	0.720	1.346
citric acid	<input type="checkbox"/>	0	1	0.275	0.195	0.038	0.275	-0.811
residual sugar	<input type="checkbox"/>	0.900	15.500	2.544	1.405	1.975	4.471	28.063
chlorides	<input type="checkbox"/>	0.012	0.611	0.088	0.048	0.002	5.622	40.506
free sulfur dioxide	<input type="checkbox"/>	1	72	15.614	10.464	109.503	1.274	2.005
total sulfur dioxide	<input type="checkbox"/>	6	289	46.806	33.252	1105.702	1.497	3.771
density	<input type="checkbox"/>	0.990	1.004	0.997	0.002	0.000	0.015	0.958
pH	<input type="checkbox"/>	2.740	4.010	3.306	0.155	0.024	0.243	0.900
sulphates	<input type="checkbox"/>	0.330	2	0.659	0.173	0.030	2.427	11.470
alcohol	<input type="checkbox"/>	8.400	14.900	10.412	1.078	1.161	0.891	0.229

Figure 2: Valores de tendência central e dispersão estatística do *dataset*

2 Exercício 2

2.1 Fazer cast do atributo *quality* para inteiro

Para transformar o atributo *quality* do *dataset* num inteiro foi aplicado o nodo *String Replacer* com os settings que se podem observar na figura seguinte:

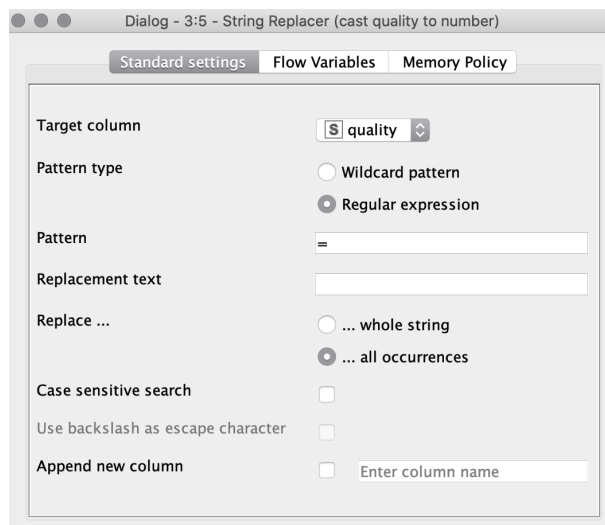


Figure 3: Settings aplicados ao nodo *String Replacer*

quality
5
5
5
6
5
5
5
7
7

Figure 4: Excerto da coluna *quality* após a aplicação do nodo em questão

2.2 Normalizar todos os atributos numéricos

Para normalizar os atributos numéricos foi utilizado o nodo *Normalizer*.

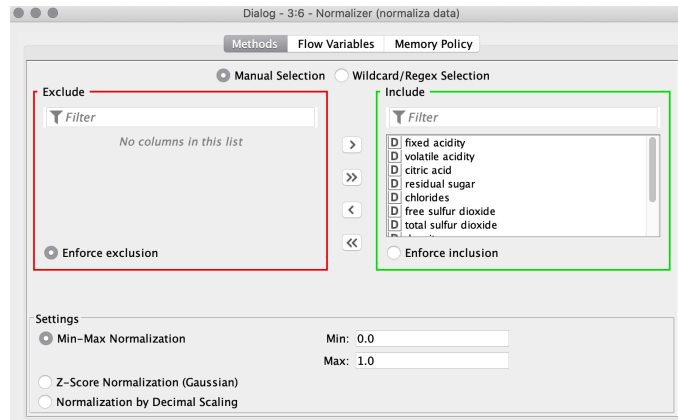


Figure 5: Settings aplicados ao nodo *Normalizer*

2.3 Criar 4 bins de igual frequência para a feature *citric acid*, substituindo a feature original

Para criar quatro bins de igual frequência utilizou-se o nodo *Auto-Binner*

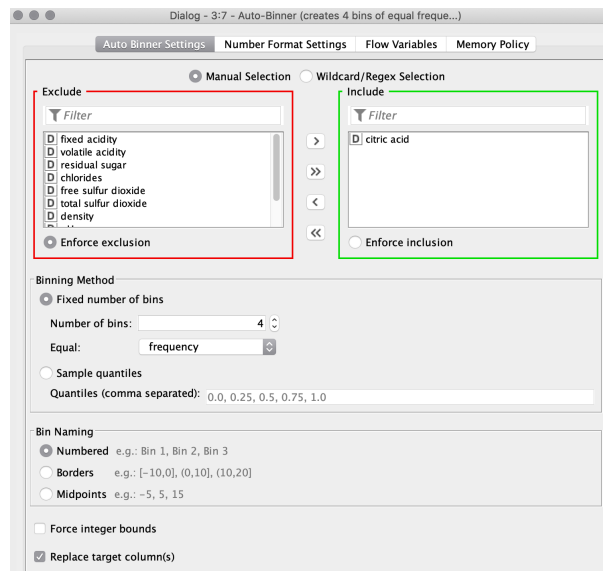


Figure 6: Settings aplicados ao nodo *Auto-Binner*

- 2.4 Renomear cada bin de forma a que o primeiro corresponda a Low, o segundo a Medium, o terceiro a High e o quarto a Very High.



Figure 7: Settings aplicados ao nodo *Cell Replacer*

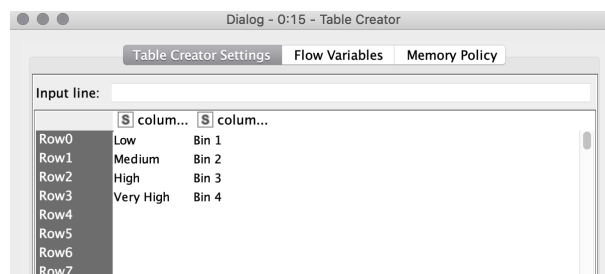


Figure 8: Settings aplicados ao nodo *Table Creator*

Row ID	D fixed ...	D volatil...	S citric acid	D residu...
Row0	0.248	0.397	Low	0.068
Row1	0.283	0.521	Low	0.116
Row2	0.283	0.438	Low	0.096
Row3	0.584	0.11	Very High	0.068
Row4	0.248	0.397	Low	0.068
Row5	0.248	0.37	Low	0.062
Row6	0.292	0.329	Low	0.048
Row7	0.239	0.363	Low	0.021
Row8	0.283	0.315	Low	0.075
Row9	0.257	0.26	High	0.356
Row10	0.186	0.315	Low	0.062
Row11	0.257	0.26	High	0.356
Row12	0.088	0.339	Low	0.048
Row13	0.283	0.336	High	0.048
Row14	0.381	0.342	Medium	0.199

Figure 9: Excerto da tabela obtida

3 Exercício 3

3.1 Análise de Componentes Principais (PCA) de forma a projetar os dados em apenas duas dimensões

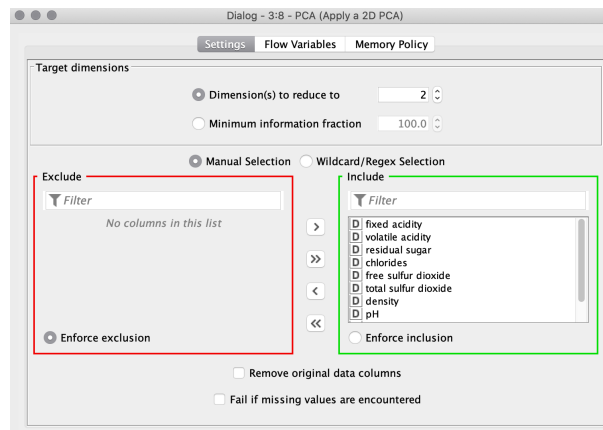
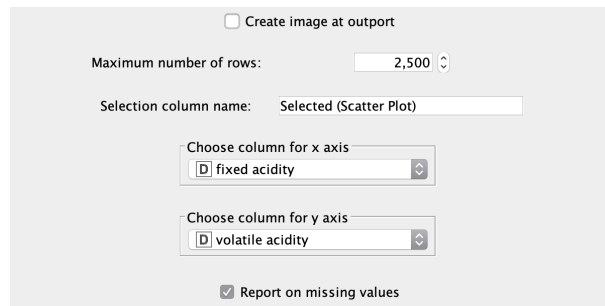


Figure 10: Settings aplicados ao nodo *PCA*

3.2 Utilizar um scatter plot para visualização dos resultados obtidos pelo PCA



The image shows the configuration interface for a 'Scatter Plot' node. At the top, there is an unchecked checkbox labeled 'Create image at output'. Below this, the 'Maximum number of rows' is set to 2,500. The 'Selection column name' is 'Selected (Scatter Plot)'. There are two dropdown menus: 'Choose column for x axis' is set to 'fixed acidity' and 'Choose column for y axis' is set to 'volatile acidity'. At the bottom, the 'Report on missing values' checkbox is checked.

Figure 11: Settings aplicados ao nodo *Scatter Plot*

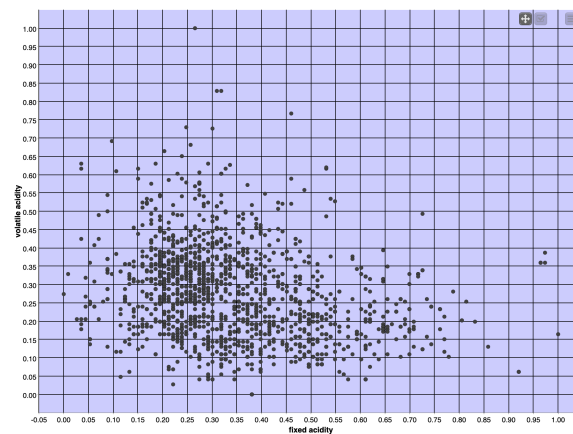


Figure 12: *Scatter Plot* das feautres *Fixed Acidity* e *Volatile Acidity*

4 Exercício 4

4.1 Segmentar o dataset aplicando o método k-means

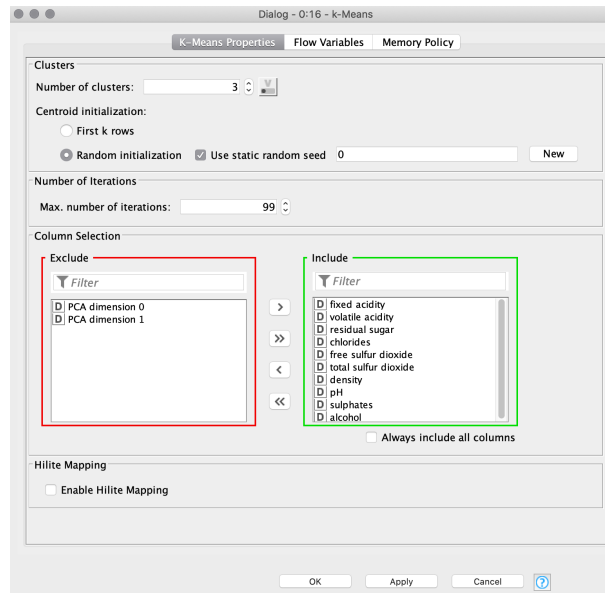


Figure 13: Settings aplicados ao nodo *K-means*

Row ID	D fixed acidity	D volatile acidity	D residual sugar	D chlorides	D free sulfur dioxide	D total sulfur dioxide	D density	D pH	D sulphates	D alcohol
cluster_0	0.285	0.324	0.104	0.132	0.234	0.181	0.511	0.466	0.174	0.209
cluster_1	0.549	0.213	0.14	0.147	0.154	0.102	0.625	0.332	0.249	0.319
cluster_2	0.246	0.245	0.104	0.099	0.198	0.11	0.339	0.511	0.195	0.509

Figure 14: Tabela com os 3 clusters obtidos

4.2 Atribuir diferentes cores por qualidade do vinho e diferentes formas aos clusters

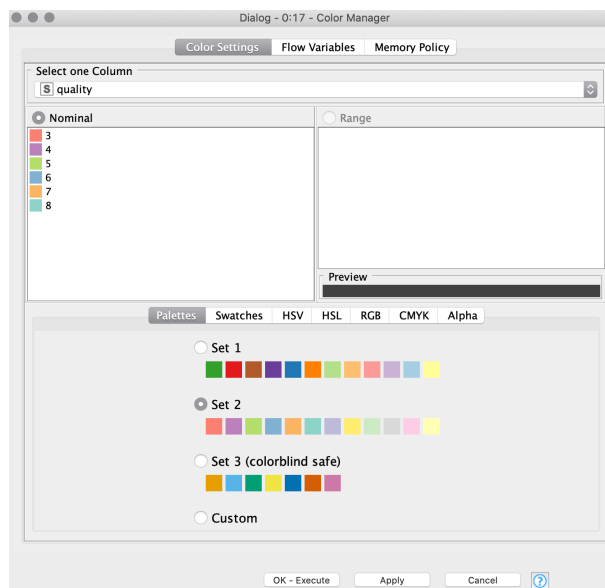


Figure 15: Settings de cor aplicados à qualidade do vinho

Row ID	D fixed ...	D volatil...	S citric ...	D residu...	D chlori...	D free s...	D total s...	D density	D pH	D sulph...	D alcohol
Row0	0.248	0.397	Low	0.068	0.107	0.141	0.099	0.568	0.606	0.138	0.154
Row1	0.283	0.521	Low	0.116	0.144	0.338	0.216	0.494	0.382	0.21	0.215
Row2	0.283	0.438	Low	0.096	0.134	0.197	0.17	0.509	0.409	0.192	0.215
Row3	0.584	0.11	Very High	0.068	0.105	0.225	0.191	0.582	0.331	0.15	0.215
Row4	0.248	0.397	Low	0.068	0.107	0.141	0.099	0.568	0.606	0.138	0.154
Row5	0.248	0.37	Low	0.062	0.105	0.169	0.12	0.568	0.606	0.138	0.154
Row6	0.292	0.329	Low	0.048	0.095	0.197	0.187	0.465	0.441	0.078	0.154
Row7	0.239	0.363	Low	0.021	0.088	0.197	0.053	0.333	0.512	0.084	0.246
Row8	0.283	0.315	Low	0.075	0.102	0.113	0.042	0.494	0.488	0.144	0.169
Row9	0.257	0.26	High	0.356	0.098	0.225	0.339	0.568	0.48	0.281	0.323
Row10	0.186	0.315	Low	0.062	0.142	0.197	0.208	0.428	0.425	0.126	0.123
Row11	0.257	0.26	High	0.356	0.098	0.225	0.339	0.568	0.48	0.281	0.323
Row12	0.088	0.339	Low	0.048	0.129	0.211	0.187	0.311	0.661	0.114	0.231
Row13	0.283	0.336	High	0.048	0.17	0.113	0.081	0.538	0.409	0.737	0.108
Row14	0.381	0.342	Medium	0.199	0.274	0.718	0.491	0.626	0.331	0.329	0.123
Row15	0.381	0.342	Medium	0.205	0.264	0.704	0.502	0.626	0.339	0.359	0.123
Row16	0.345	0.11	Very High	0.062	0.134	0.479	0.343	0.501	0.441	0.251	0.323
Row17	0.31	0.301	High	0.055	0.594	0.211	0.177	0.494	0.291	0.569	0.138
Row18	0.248	0.322	Low	0.24	0.124	0.07	0.081	0.538	0.504	0.102	0.092
Row19	0.292	0.137	Very High	0.062	0.549	0.225	0.177	0.501	0.236	0.449	0.123
Row20	0.381	0.068	Very High	0.062	0.109	0.394	0.191	0.494	0.512	0.12	0.154
Row21	0.265	0.185	High	0.096	0.117	0.31	0.23	0.597	0.614	0.192	0.2
Row22	0.292	0.212	Medium	0.048	0.157	0.127	0.11	0.479	0.339	0.347	0.169
Row23	0.345	0.253	Medium	0.096	0.12	0.113	0.216	0.494	0.339	0.12	0.154
Row24	0.204	0.192	Medium	0.103	0.122	0.282	0.12	0.494	0.543	0.18	0.2
Row25	0.15	0.185	Medium	0.034	0.114	0.141	0.06	0.399	0.472	0.138	0.138

Figure 16: Excerto da tabela após a aplicação do nodo *Color Manager*

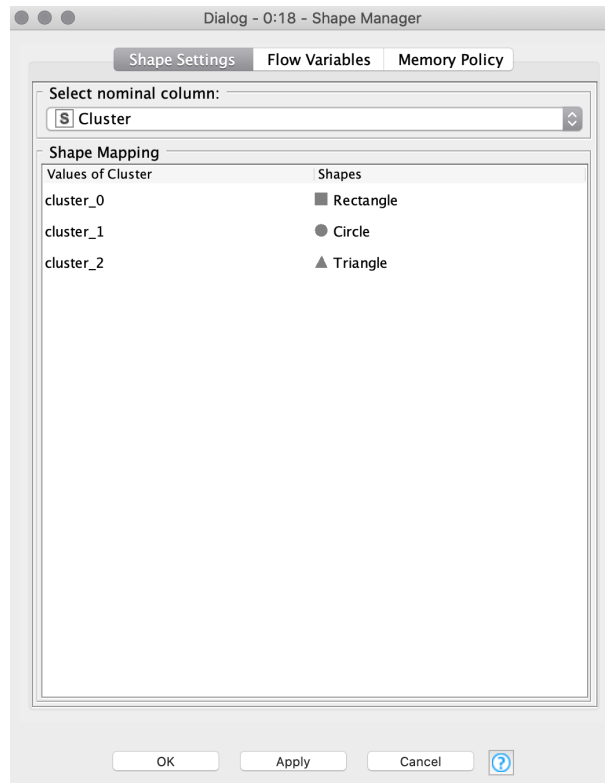


Figure 17: Settings aplicados ao nodo *Shape Manager*

4.3 Criar scatter plots e scatter matrixes que permitam ter uma noção gráfica, em duas dimensões, dos atributos e dos clusters criados

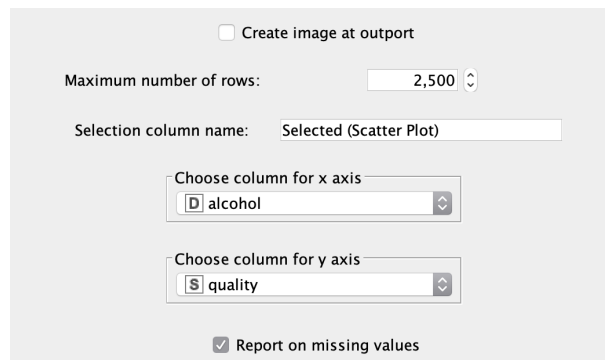


Figure 18: Settings aplicados ao nodo *Scatter Plot*

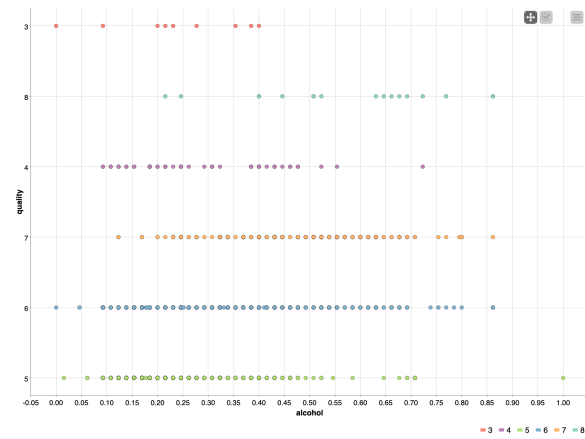


Figure 19: *Scatter Plot* das features *quality* e *alcohol*

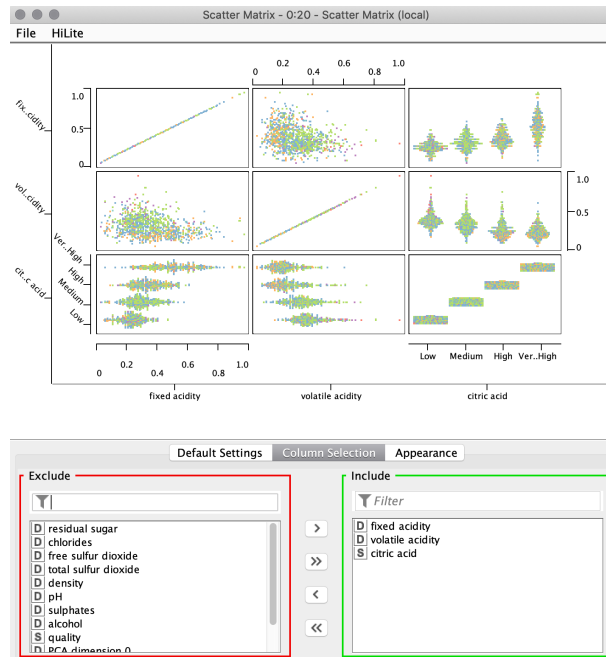


Figure 20: *Scatter Matrix*

4.4 Ler e tratar os dados de teste de forma a que, com base no modelo desenvolvido nos passos anteriores, seja atribuído um cluster a cada registo deste ficheiro

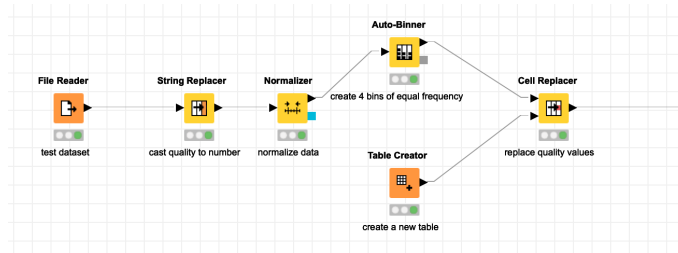


Figure 21: Fluxo de tratamento do *dataset* de teste

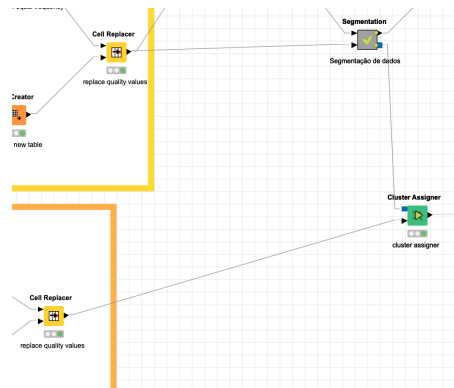


Figure 22: Fluxo para a atribuição de um cluster a cada registo do ficheiro

Row ID	critic	D_restdu	D_chort	D_free s	D_total s	D_density	D_pH	D_sulph	D_alcohol	S_quality	S_critic	D_PCA d	D_PCA d	S_Cluster
Row0	0.065	0.081	0.033	0.034	0.391	0.614	0.37	0.658	6	Bin 1	0.325	0.293	cluster_2	
Row1	0	0.193	0.849	0.025	0.391	0.281	0.185	0.158	5	Bin 1	-0.106	0.284	cluster_0	
Row2	0.226	0.228	0.262	0.437	0.635	0.105	0.185	0.105	5	Bin 3	-0.588	0.154	cluster_0	
Row3	0.073	0.203	0.164	0.405	0.628	0.368	0.204	0.211	5	Bin 3	-0.335	0.056	cluster_0	
Row4	0.081	0.198	0.016	0.017	0.476	0.719	0.481	0.526	6	Bin 4	0.335	0.059	cluster_2	
Row5	0.065	0.345	0	0	0.568	0.456	0.222	0.632	6	Bin 4	0.192	-0.049	cluster_2	
Row6	0.024	0.203	0.016	0	0.518	0.709	0.37	0.25	5	Bin 1	0.072	0.488	cluster_0	
Row7	0.04	0.162	0.016	0.017	0.489	0.719	0.426	0.474	6	Bin 1	0.208	0.432	cluster_2	
Row8	0.065	0.345	0	0	0.568	0.456	0.222	0.632	6	Bin 4	0.192	-0.049	cluster_2	
Row9	0.048	0.183	0.197	0.109	0.949	0.491	0.137	0.421	6	Bin 4	-0.085	-0.214	cluster_1	
Row10	0.024	0.122	0.033	0.067	0.212	0.316	0.5	0.553	5	Bin 4	0.357	-0.45	cluster_2	
Row11	0.056	0.046	0.344	0.303	0.67	0.649	0.63	0.184	6	Bin 3	-0.116	-0.172	cluster_0	
Row12	0.048	0.152	0.23	0.126	0.763	0.596	0.444	0.211	5	Bin 1	-0.19	0.253	cluster_0	

Figure 23: Excerto do output gerado pelo nodo *Cluster Assigner*

4.5 Guardar o resultado da atribuição num ficheiro csv

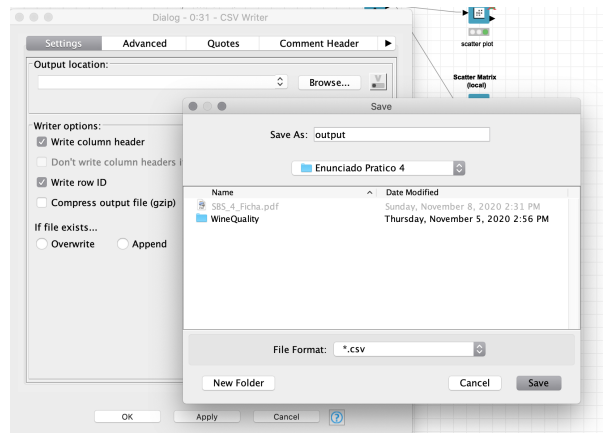


Figure 24: Settings aplicados ao nodo *CSV Writer*

5 Exercício 5

Para parametrizar o *workflow* recorreu-se ao nodo *Integer input* para criar uma variável de fluxo **local** para o número de bins e clusters.

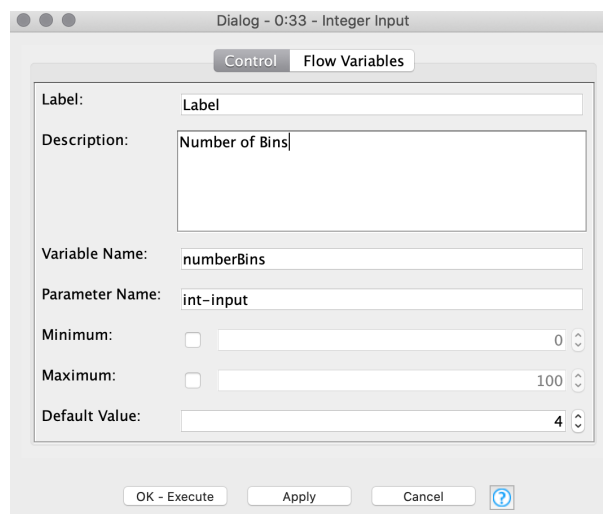


Figure 25: Settings aplicados ao nodo *Integer Input*

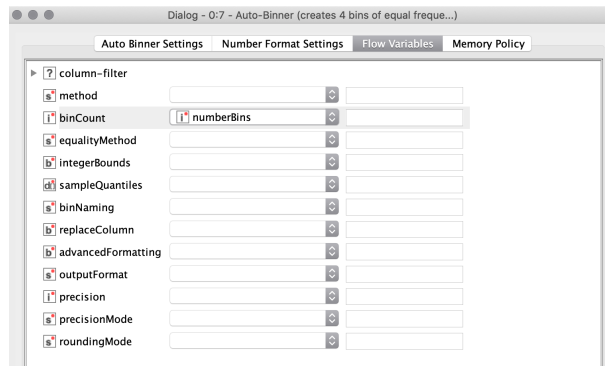


Figure 26: Configuração das variáveis de fluxo associadas ao nodo *Auto-Binner*

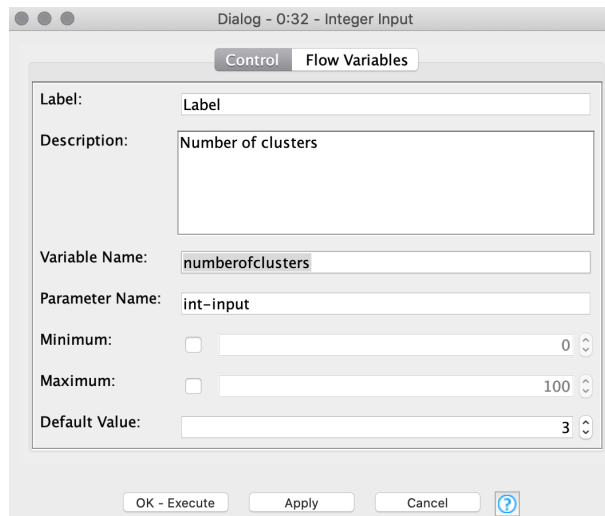


Figure 27: Settings aplicados ao nodo *Integer Input*

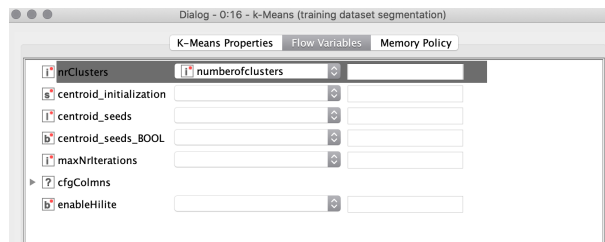


Figure 28: Configuração das variáveis de fluxo associadas ao nodo *K-Means*

Quanto aos títulos dos gráficos, criou-se também um variável local através

do nodo *String Widget*.

The image shows a software configuration window titled "Dialog - 0:36:21 - String Widget". It has two tabs: "Control" and "Flow Variables", with "Flow Variables" currently selected. The window contains several input fields and options for configuring a string widget's flow variables. The fields are as follows:

- Label:** A text field containing "changeTitle".
- Description:** A multi-line text area containing "Change title of graphs".
- Variable Name:** A text field containing "changeTitle".
- Editor type:** Two radio buttons: "Single-line" (selected) and "Multi-line".
- Multi-line editor width:** A numeric input field set to "60".
- Multi-line editor height:** A numeric input field set to "5".
- Regular Expression:** A dropdown menu.
- Validation Error Message:** An empty text field.
- Common Regular Expressions:** A dropdown menu with an "Assign" button next to it.
- Default Value:** An empty text field.

At the bottom of the dialog are four buttons: "OK", "Apply", "Cancel", and a help icon (a question mark inside a circle).

Figure 29: Configuração das variáveis de fluxo associadas ao títulos dos gráficos

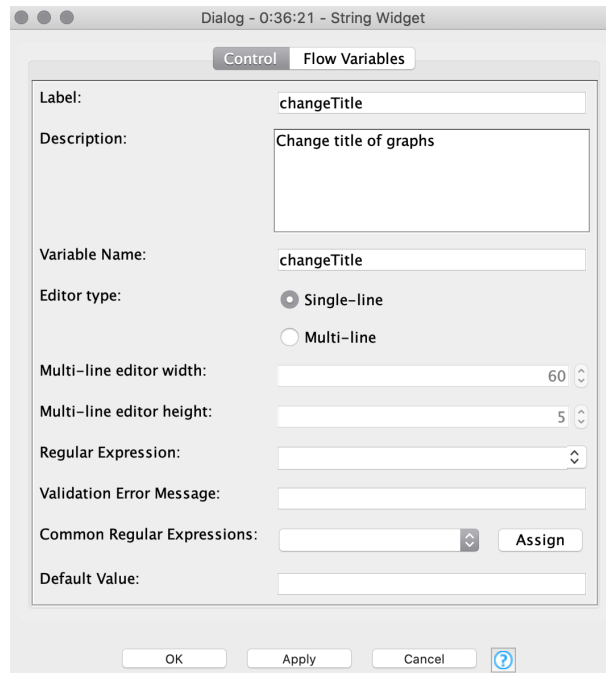


Figure 30: Configuração do nodo *Scatter Plot* para utilização da variavel local definida

6 Exercício 6

Para poder visualizar facilmente todo o *workflow* foram utilizados 2 meta-nodos para agrupar a segmentação de dados e os nodos de visualização. Para as diferentes partes do tratamento de dados foram utilizadas anotações. Apresenta-se de seguida o fluxo global final obtido e os meta-nodos.

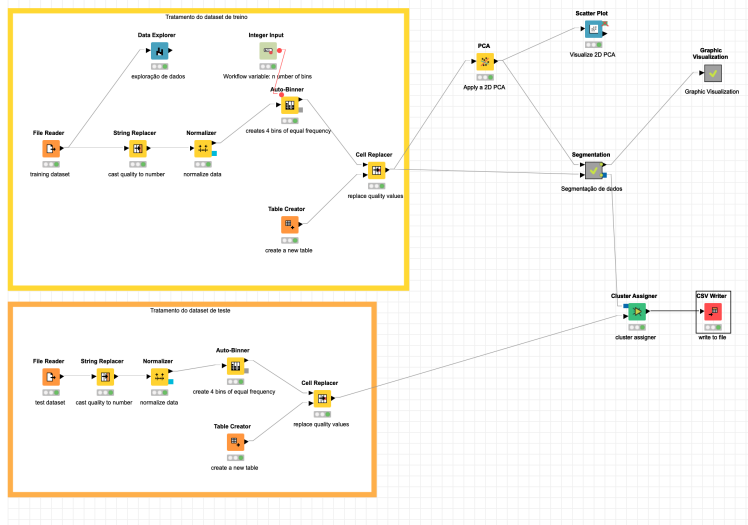


Figure 31: Workflow final

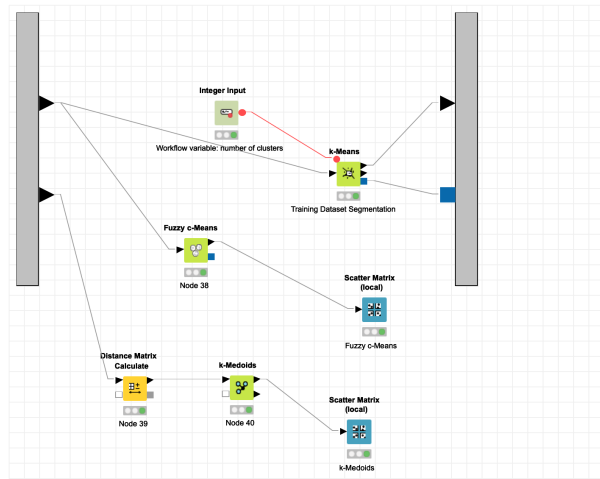


Figure 32: Metanodo *Segmentação*

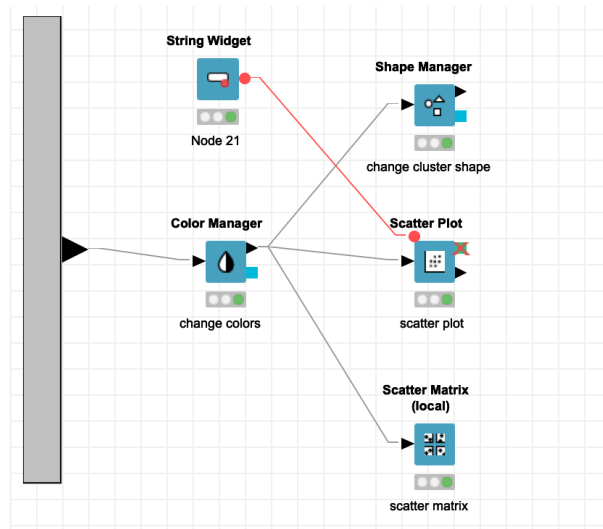


Figure 33: Metanodo *Graphic Visualization*

7 Exercício 7

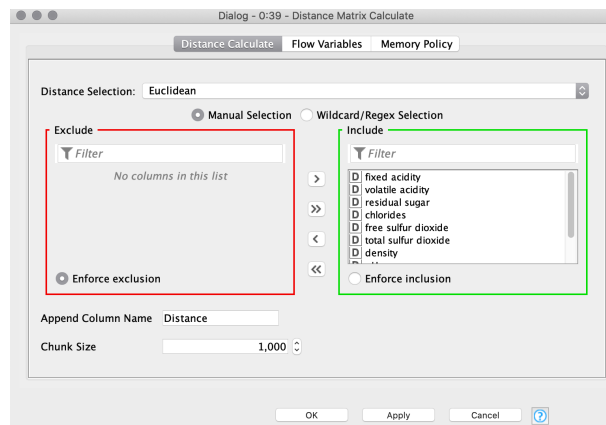


Figure 34: Calculo de vetor de distancia através do vetor *Distance Matrix Calculator*

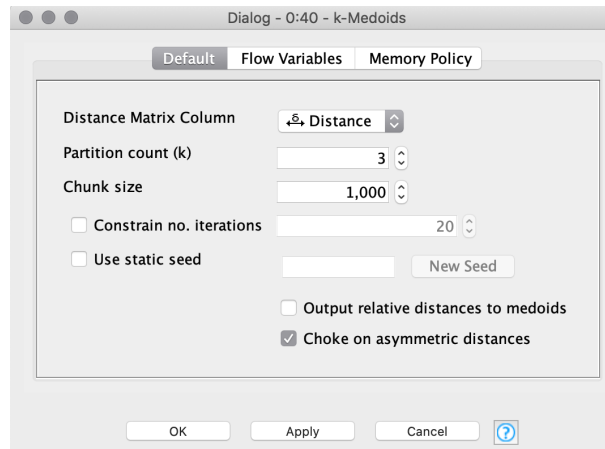


Figure 35: Settings aplicados ao nodo *K-Medoids*

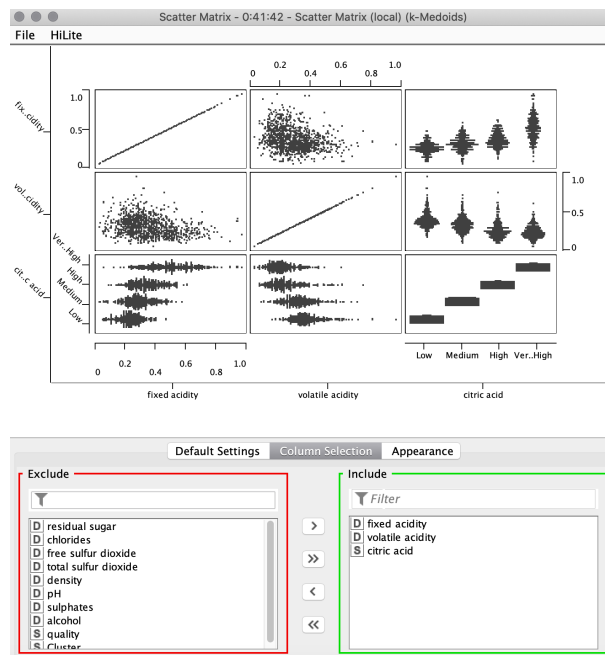


Figure 36: Scatter Matrix após segmentar os dados com o nodo *K-Medoids*

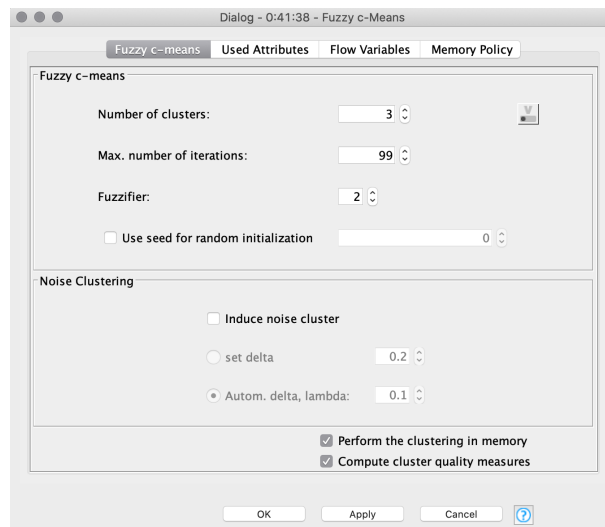


Figure 37: Settings aplicados ao nodo *Fuzzy*

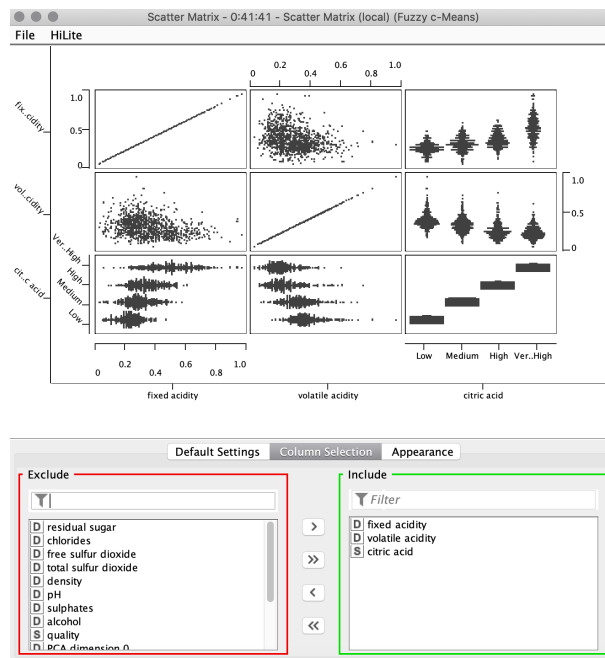


Figure 38: Scatter Matrix após segmentar os dados com o nodo *Fuzzy C-Means*