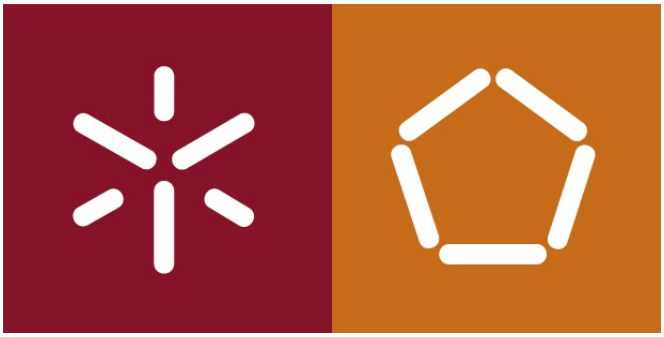

Universidade do Minho
Departamento de Informática

Sistemas Baseados em Similaridade

Trabalho Prático Individual 4

Gonçalo Almeida (A84610)

Novembro 2020

1 Tarefa 1

1.1 Carregar, no Knime, o dataset descarregado e explorar os dados;

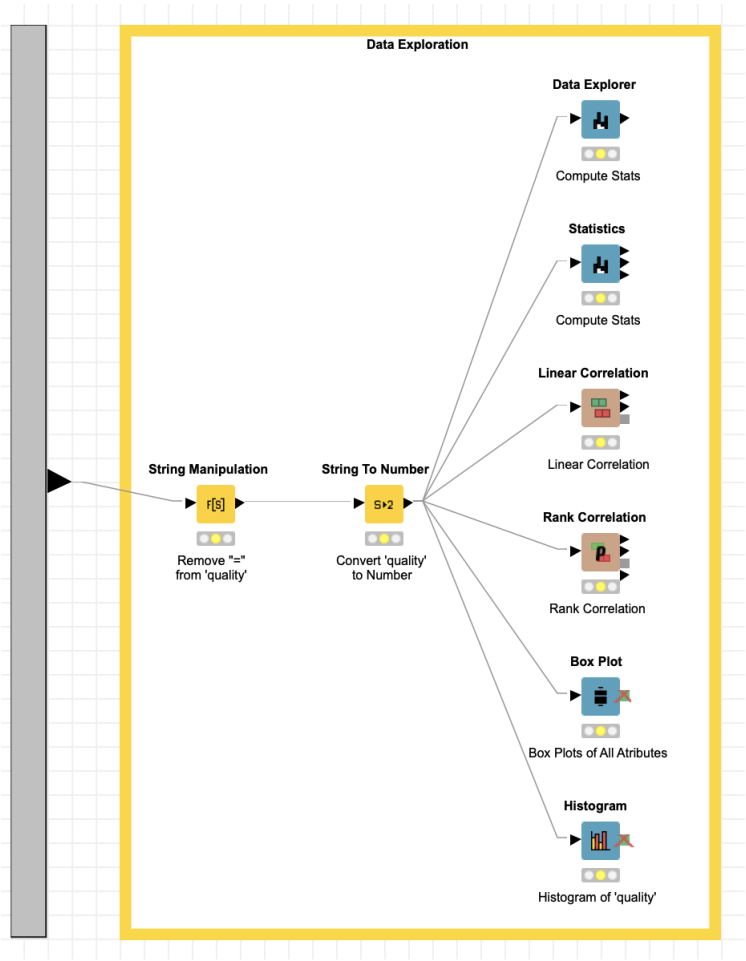

Figure 1: Metanodo Data Exploration

2 Tarefa 2

2.1 Tratar os dados, i.e.:

2.1.1 Fazer cast do atributo “quality” para inteiro;

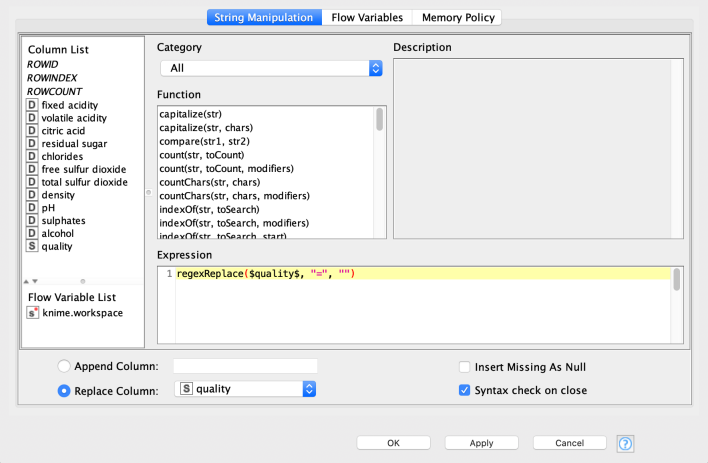

Figure 2: Remover o caracter ”=” do atributo ’quality’

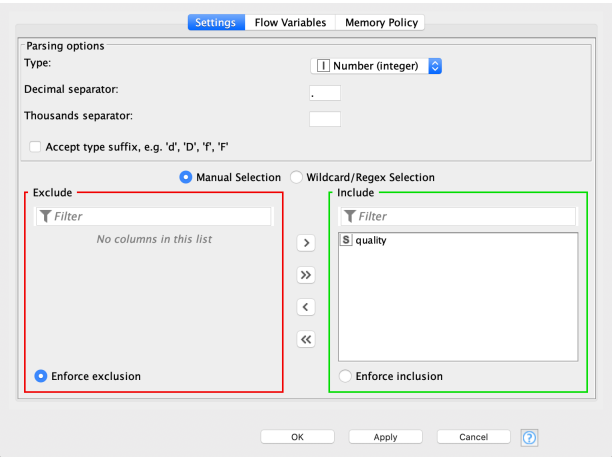

Figure 3: Cast do atributo ’quality’

2.1.2 Normalizar todos os atributos numéricos utilizando a transformação linear Min- max de forma a produzir um input normalizado entre 0 e 1;

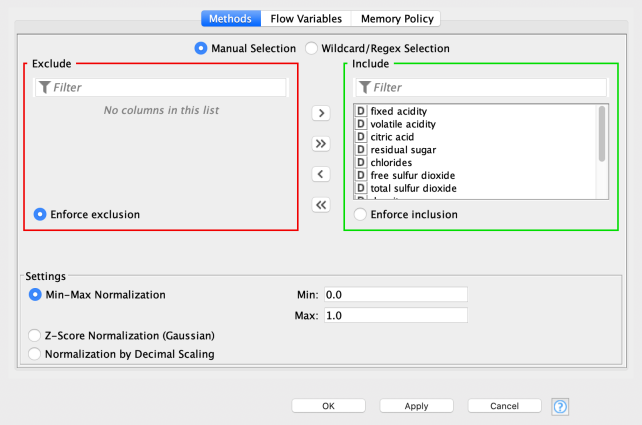

Figure 4: Normalização dos atributos numéricos

2.1.3 Criar 4 bins de igual frequência para a feature "citric acid", substituindo a feature original;

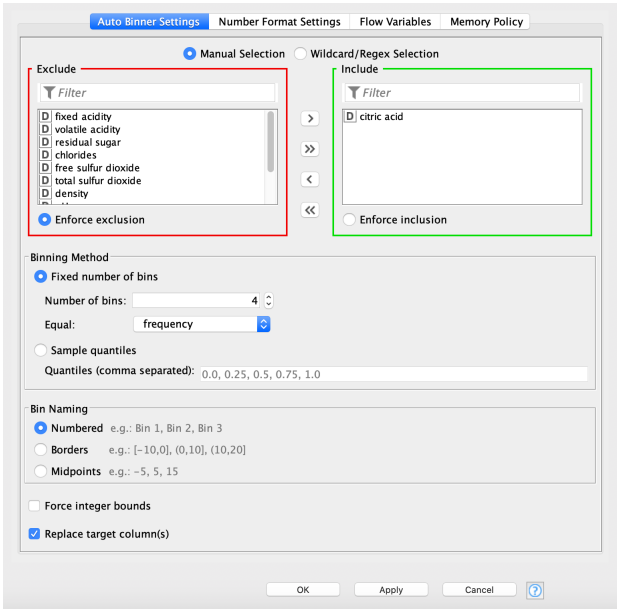

Figure 5: Normalização dos atributos numéricos

2.1.4 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.

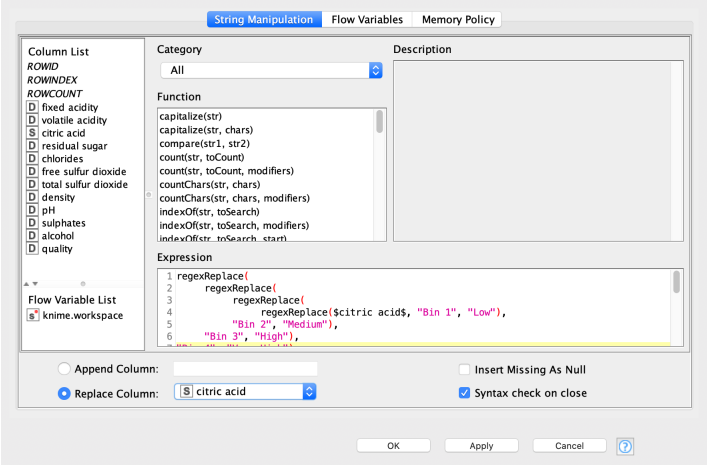

Figure 6: Renomeação de cada bin

3 Tarefa 3

3.1 Aplicar:

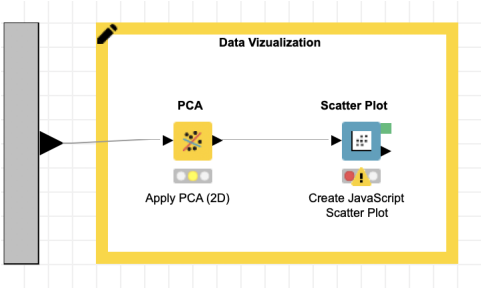

Figure 7: Metanodo Data Visualization

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

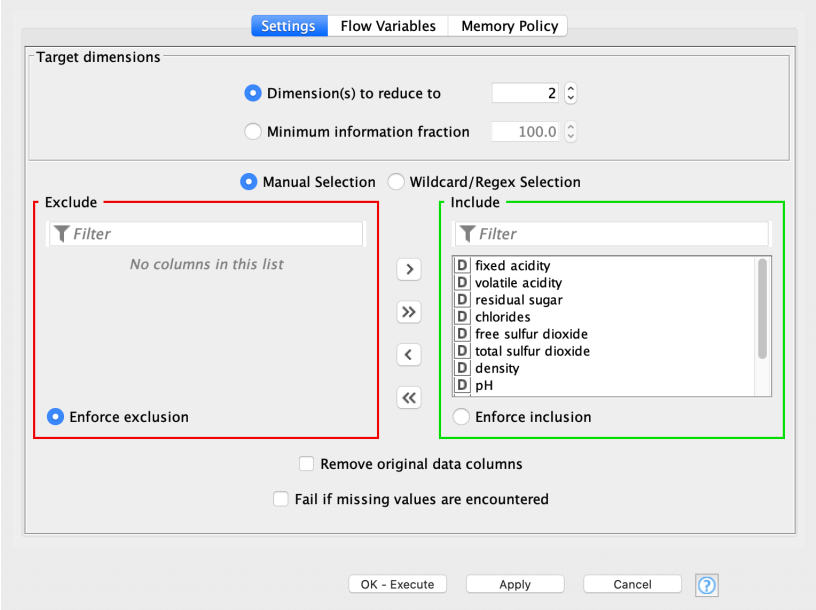

Figure 8: Aplicação de uma PCA

3.1.2 Utilizar um scatter plot para visualização dos resultados obtidos pelo PCA.

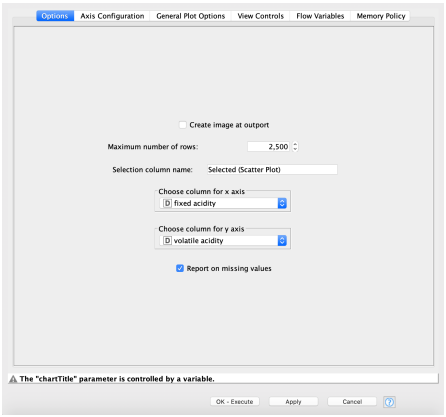

Figure 9: Criação do scatter plot

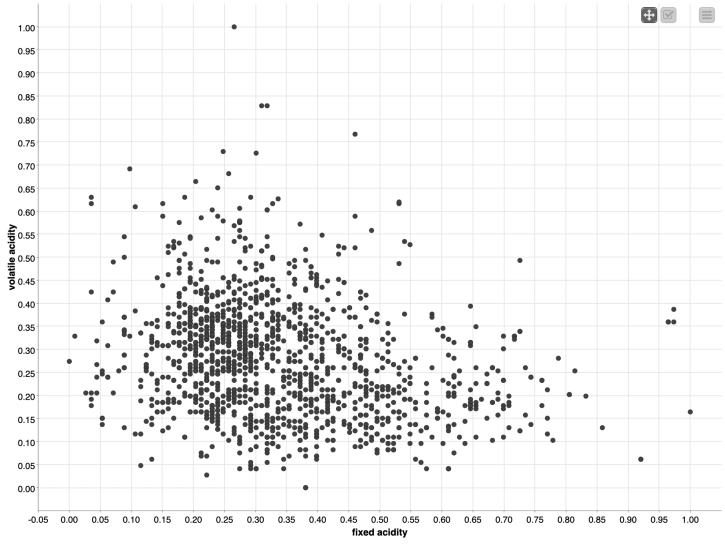

Figure 10: Scatter Plot da PCA

4 Tarefa 4

4.1 Segmentar o dataset:

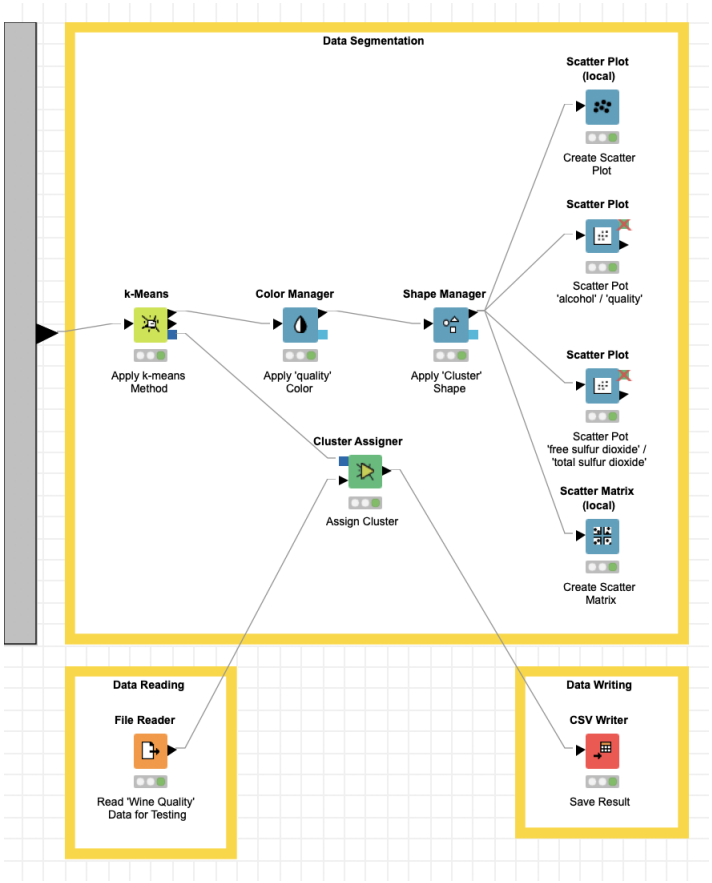

Figure 11: Metanode Data Segmentation

4.1.1 Aplicando o método k-means;

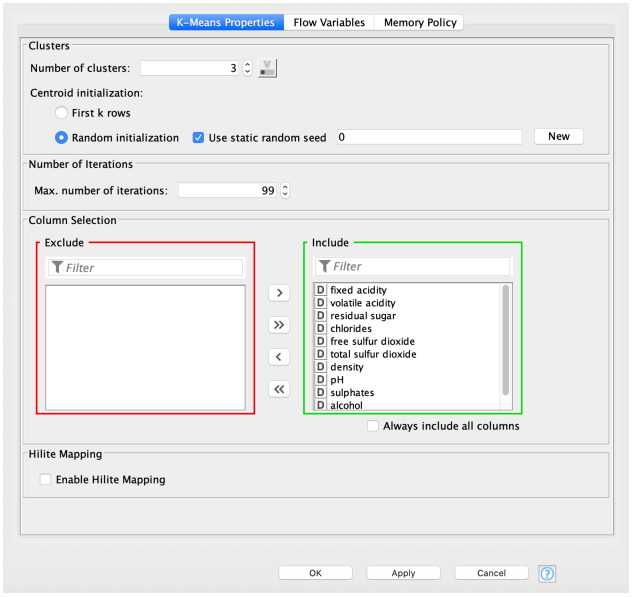

Figure 12: Método k-means

4.1.2 Atribuir diferentes cores por qualidade do vinho e diferentes formas aos clusters;

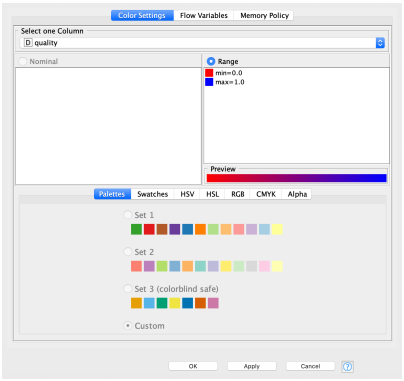

Figure 13: Atribuição de cores por qualidade do vinho

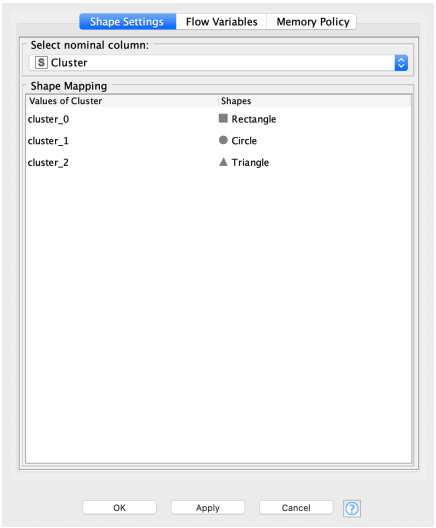

Figure 14: Atribuição de formas aos clusters

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

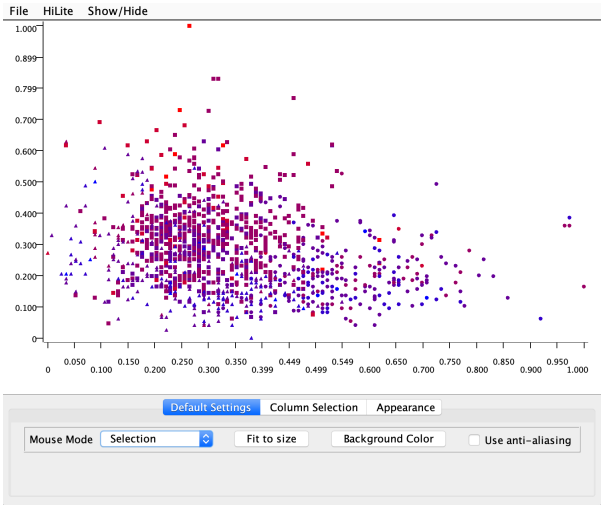

Figure 15: Scatter plot

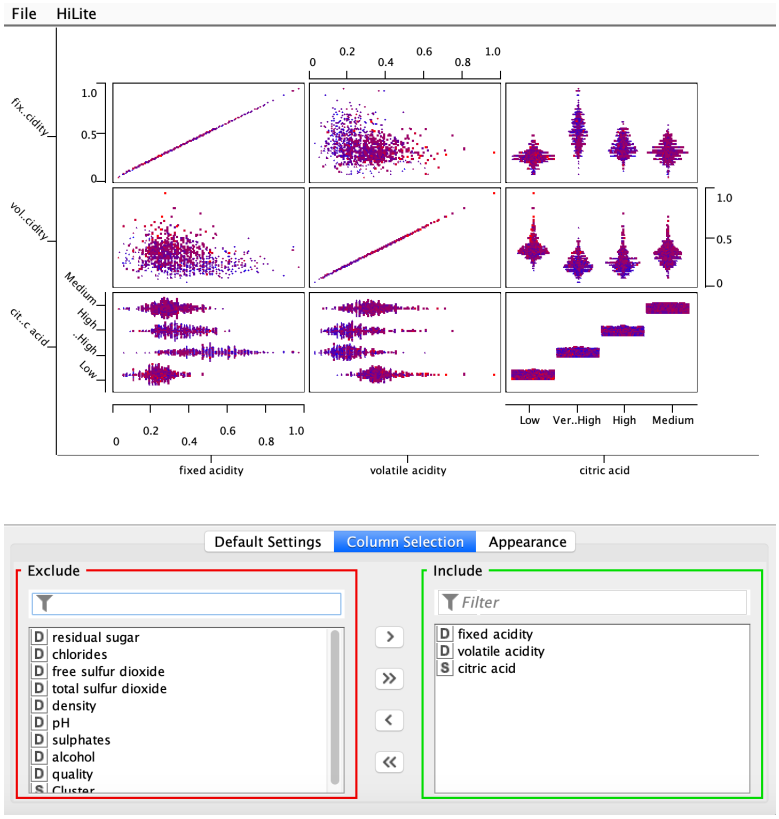

Figure 16: Scatter Matrix

4.1.4 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;

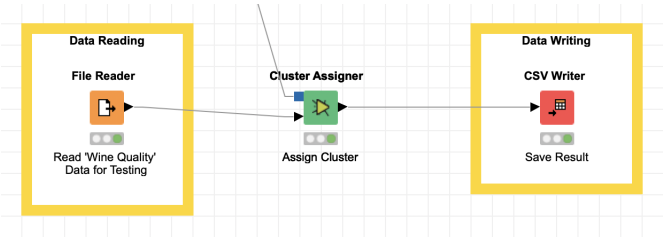

Figure 17: Leitura dos dados de teste e aplicação do modelo desenvolvido

4.1.5 Guardar o resultado da atribuição num ficheiro csv.

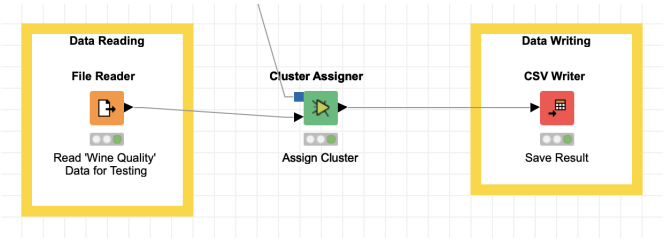

Figure 18: Guardar o resultado

5 Tarefa 5

5.1 Parametrizar o workflow, utilizando variáveis de fluxo para definir o número de bins, o número de clusters e os títulos dos gráficos criados;

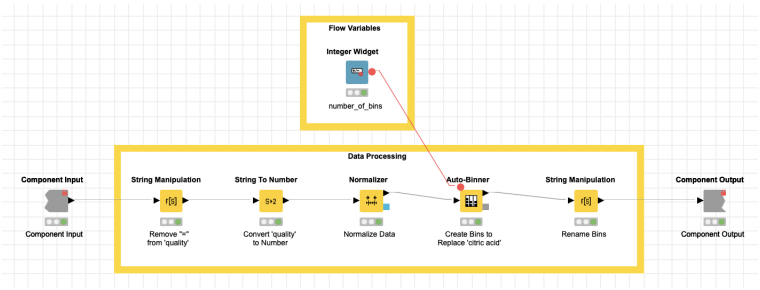

Figure 19: Variável para o número de bins

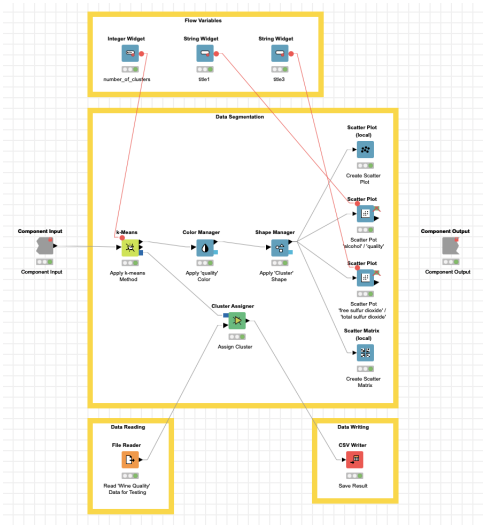

Figure 20: Variáveis para o número de clusters e os títulos dos gráficos

6 Tarefa 6

6.1 Produzir o workflow de maneira a que seja possível visualizar, numa única página, todos os componentes visuais implementados;

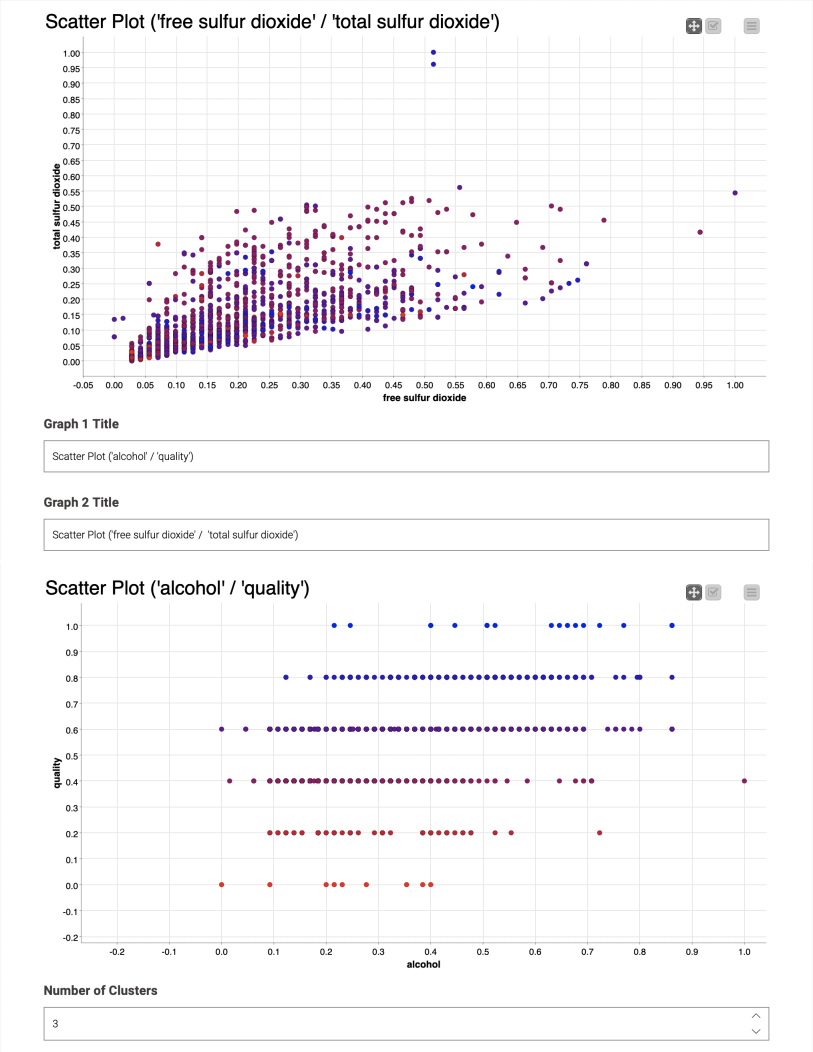

Figure 21: Página com os componentes visuais

7 Tarefa 7

7.1 Experimentar, avaliar e comparar outros métodos de segmentação.

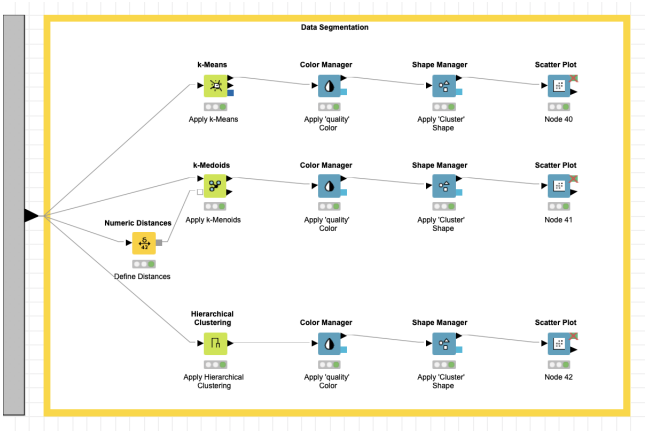

Figure 22: Metanodos de Segmentação

7.1.1 Particionamento

Método k-Means: Apesar de ser relativamente eficiente e terminar com ótimos locais, apenas é aplicável quando é possível calcular a média, necessita da identificação do número de segmentos à priori, é incapaz de lidar com ruído e é inadequado para a determinação de segmentos côncavos.

Método k-Menoids: Apesar de apresentar maior robustez relativamente à presença de dados ruidosos comparativamente ao método k-Means, a qualidade dos resultados diminui quanto maior a dimensão dos conjuntos de dados.

7.1.2 Hierarquização

Comparativamente ao método k-Means, apresenta melhor resultados e não necessita da especificação do número de segmentos. Ao contrário do particionamento, traduz alguma organização dos segmentos em vez de um simples conjunto de segmentos. Contudo, apresenta dificuldades com o aumento de atributos ou objetos.

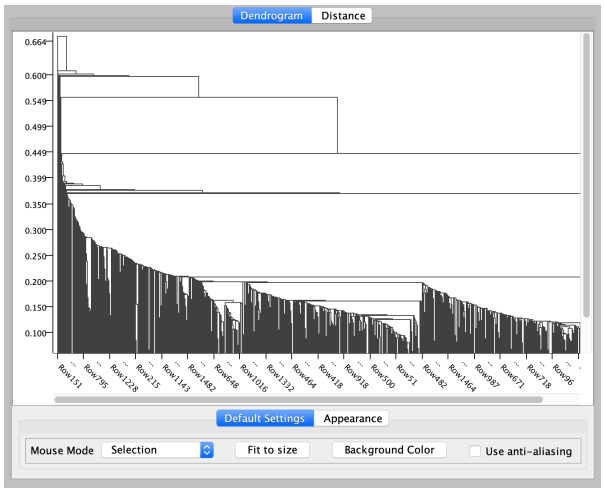

Figure 23: Dendograma

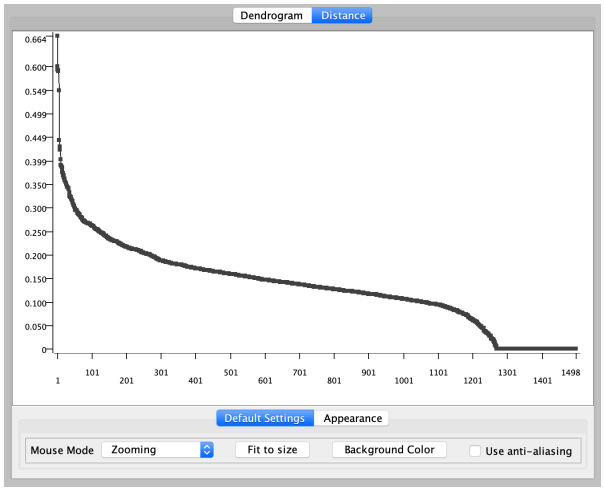

Figure 24: Visualização da distância