

Universidade do Minho

Departamento de Informática

Mestrado em Engenharia Informática

Perfil de Machine Learning Fundamentos e Aplicações

Sistemas Baseados em Similaridade

1º Ano, 1º Semestre

Ano letivo 2020/2021

Enunciado Prático nº 4

09 de novembro de 2020

Joel Costa Carvalho

PG42837

Índice

[Tarefas - 4 -](#_Toc55849060)

[T1 – Carregar, no *Knime*, o *dataset* descarregado e explorar os dados; - 4 -](#_Toc55849061)

[T2 – Tratar os dados: - 6 -](#_Toc55849062)

[**a)** **Fazer *cast* do atributo “*quality*” para inteiro;** - 6 -](#_Toc55849063)

[**b)** **Normalizar todos os atributos numéricos utilizando a transformação linear Min-Max de forma a produzir um input normalizado entre 0 e 1;** - 6 -](#_Toc55849064)

[**c)** **Criar 4 *bins* de igual frequência para a *feature* “*citric acid*”, substituindo a *feature* original;** - 7 -](#_Toc55849065)

[**d)** **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*.** - 7 -](#_Toc55849066)

[T3 – Aplicar: - 7 -](#_Toc55849067)

[**a)** **Uma Análise de Componentes Principais (PCA) de forma a projetar os dados em apenas duas dimensões;** - 7 -](#_Toc55849068)

[**b)** **Utilizar um *scatter plot* para visualização dos resultados obtidos pelo PCA.** - 8 -](#_Toc55849069)

[T4 - Segmentar o *dataset*: - 8 -](#_Toc55849070)

[**a)** **Aplicando o método *k-means;*** - 8 -](#_Toc55849071)

[**b)** **Atribuir diferentes cores por qualidade do vinho e diferentes formas aos *clusters*;** - 9 -](#_Toc55849072)

[**c)** **Criar *scatter plots* e *scatter matrixes* que permitam ter uma noção gráfica, em duas dimensões, dos atributos e dos *clusters* criados;** - 9 -](#_Toc55849073)

[**d)** **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;** - 10 -](#_Toc55849074)

[**e)** **Guardar o resultado da atribuição num ficheiro csv** - 10 -](#_Toc55849075)

[T5 – Parametrizar o *workflow*, utilizando variáveis de fluxo para definir o número de *bins*, o número de *clusters* e os títulos dos gráficos criados; - 11 -](#_Toc55849076)

[T6 - Produzir o *workflow* de maneira a que seja possível visualizar, numa única página, todos os componentes visuais implementados; - 11 -](#_Toc55849077)

[T7 - Experimentar, avaliar e comparar outros métodos de segmentação. - 12 -](#_Toc55849078)

Índice de figuras

[Figura 1 - Carregar dataset e explorar dados - 4 -](#_Toc55930033)

[Figura 2 - Data Explorer - 4 -](#_Toc55930034)

[Figura 3 - Correlação entre as features - 5 -](#_Toc55930035)

[Figura 4 - Análise da quantidade de álcool sobre a qualidade - 5 -](#_Toc55930036)

[Figura 5 - Metanode T2 - 6 -](#_Toc55930037)

[Figura 6 - Cast atributo "quality" - 6 -](#_Toc55930038)

[Figura 7 - Normalizar atributos - 6 -](#_Toc55930039)

[Figura 8 - Criação de 4 bins para a feature "citric acid" - 7 -](#_Toc55930040)

[Figura 9 - Output das bins renomeadas - 7 -](file:////Users/joelcarvalho/Desktop/Mestrado/Cadeiras/SBS/TP4/TP4_JoelCarvalho_PG42837.docx#_Toc55930041)

[Figura 10 - Definição dos novos valores para cada bin - 7 -](#_Toc55930042)

[Figura 11 - PCA em duas dimensões - 7 -](#_Toc55930043)

[Figura 12 - Scatter Plot para visualizar os dados do PCA - 8 -](#_Toc55930044)

[Figura 13 - Método k-means - 8 -](#_Toc55930045)

[Figura 14 - Diferentes formas por clusters - 9 -](file:////Users/joelcarvalho/Desktop/Mestrado/Cadeiras/SBS/TP4/TP4_JoelCarvalho_PG42837.docx#_Toc55930046)

[Figura 15 - Cores por qualidade do vinho - 9 -](#_Toc55930047)

[Figura 16 - Scatter matrixes - 9 -](file:////Users/joelcarvalho/Desktop/Mestrado/Cadeiras/SBS/TP4/TP4_JoelCarvalho_PG42837.docx#_Toc55930048)

[Figura 17 - Scatter plots e matrixes - 9 -](#_Toc55930049)

[Figura 18 - Ler e tratar os dados, com base no modelo densenvolvido - 10 -](#_Toc55930050)

[Figura 19 - Guardar o resultado num ficheiro csv - 10 -](#_Toc55930051)

[Figura 20 - Variáveis de fluxo - 11 -](#_Toc55930052)

[Figura 21 - Atribuição da variável global "nr\_bins" ao nodo "Auto-Binner" - 11 -](#_Toc55930053)

[Figura 22 - Vista geral do workflow - 11 -](#_Toc55930054)

[Figura 23 - Vista geral do componente - 12 -](#_Toc55930055)

[Figura 24 - Output do componente - 12 -](#_Toc55930056)

[Figura 25 - Diferentes métodos de segmentação - 13 -](#_Toc55930057)

**Tarefas**

**T1 – Carregar, no *Knime*, o *dataset* descarregado e explorar os dados;**

Após o carregamento do *dataset*, os dados foram explorados com o auxílio do nodo *Data Explorer*, permitindo efetuar médias, medianas, modas, variâncias, desvios padrões e entre outros. No final, olhando para a Figura 4., concluímos, na generalidade, que quanto mais álcool mais qualidade tem.

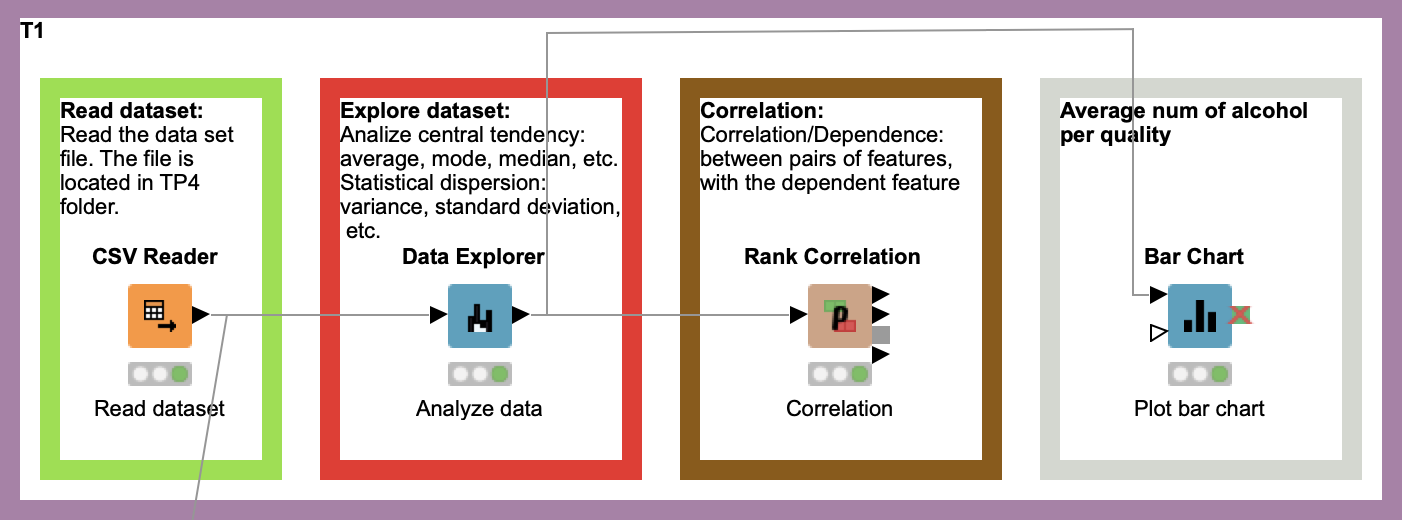


Figura 1 - Carregar dataset e explorar dados

Uma imagem com mesa

Descrição gerada automaticamente

Figura 2 - Data Explorer

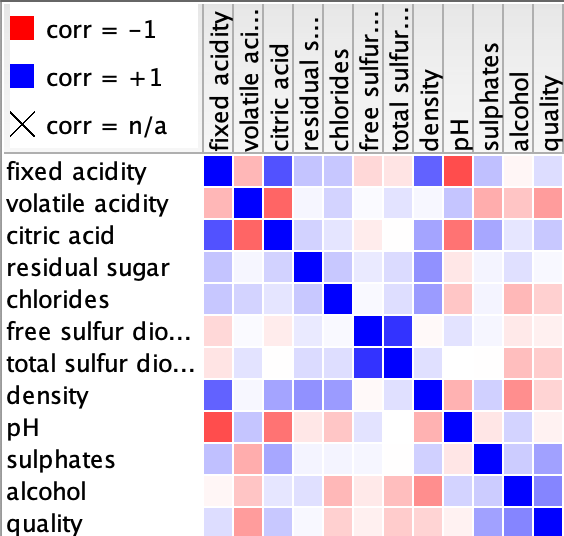


Figura 3 - Correlação entre as features

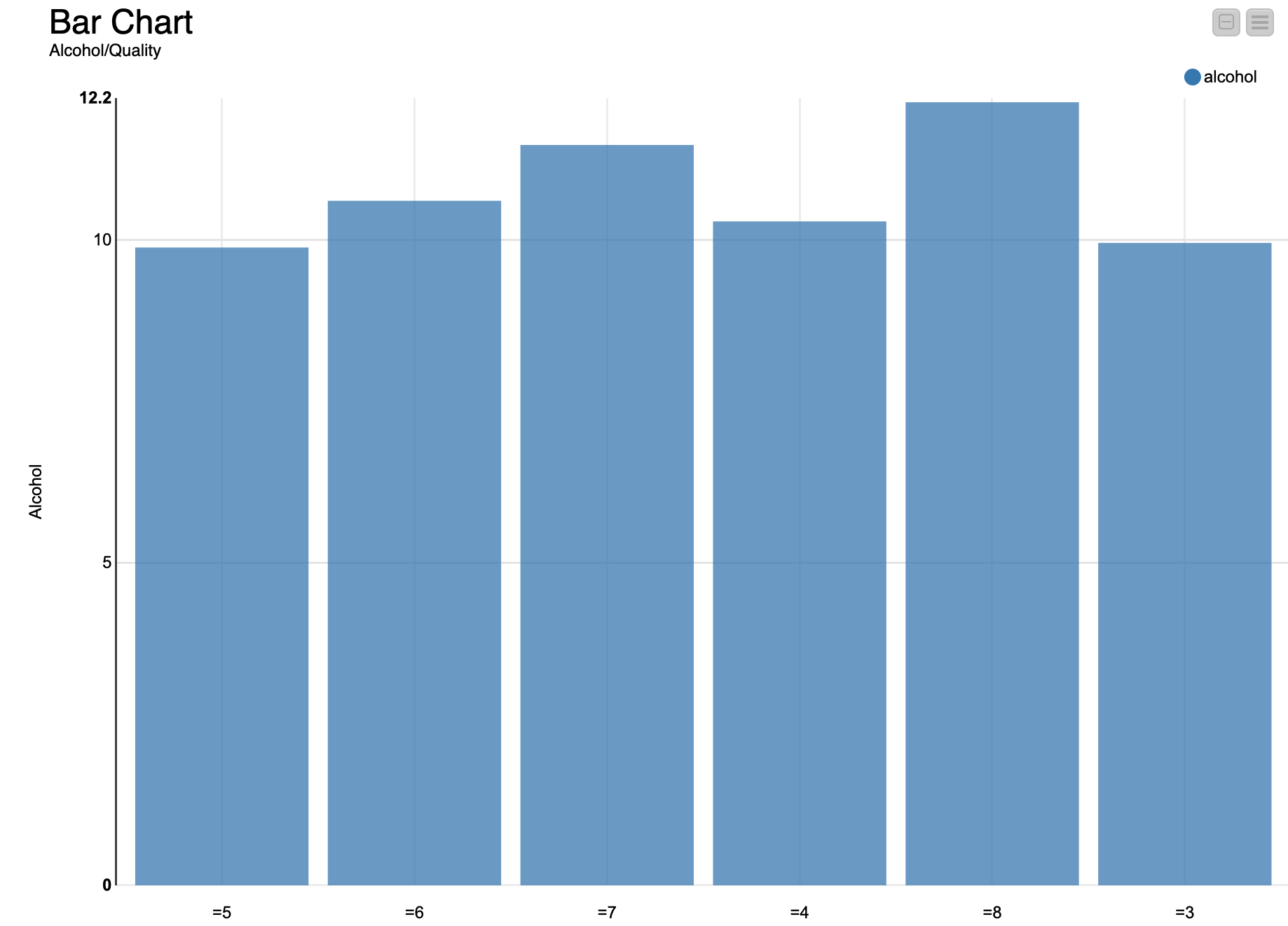


Figura 4 - Análise da quantidade de álcool sobre a qualidade

**T2 – Tratar os dados:**

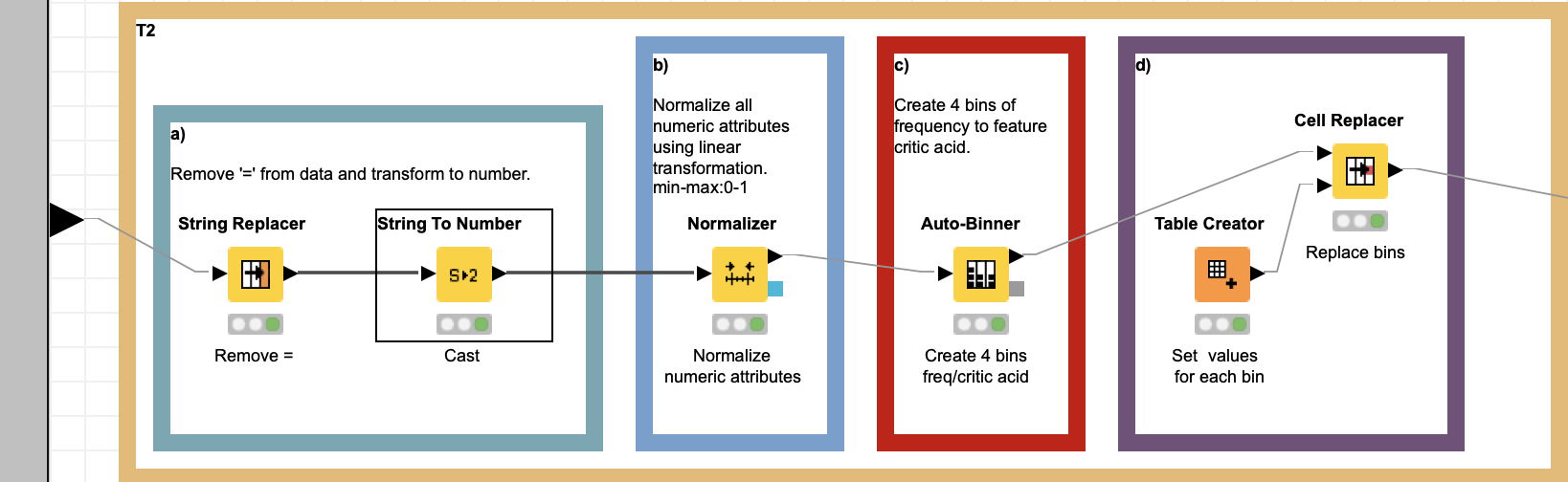
****

Figura 5 - Metanode T2

1. **Fazer *cast* do atributo “*quality*” para inteiro;**

**Uma imagem com texto

Descrição gerada automaticamente**

Figura 6 - Cast atributo "quality"

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

**Uma imagem com texto

Descrição gerada automaticamente**

Figura 7 - Normalizar atributos

1. **Criar 4 *bins* de igual frequência para a *feature* “*citric acid*”, substituindo a *feature* original;**

**Uma imagem com texto

Descrição gerada automaticamente**

Figura 8 - Criação de 4 bins para a feature "citric acid"

1. **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*.**

Uma imagem com mesa

Descrição gerada automaticamente

Uma imagem com mesa

Descrição gerada automaticamente

Figura 9 - Output das bins renomeadas

Figura 10 - Definição dos novos valores para cada bin

**T3 – Aplicar:**

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

**Uma imagem com texto

Descrição gerada automaticamente**

Figura 11 - PCA em duas dimensões

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

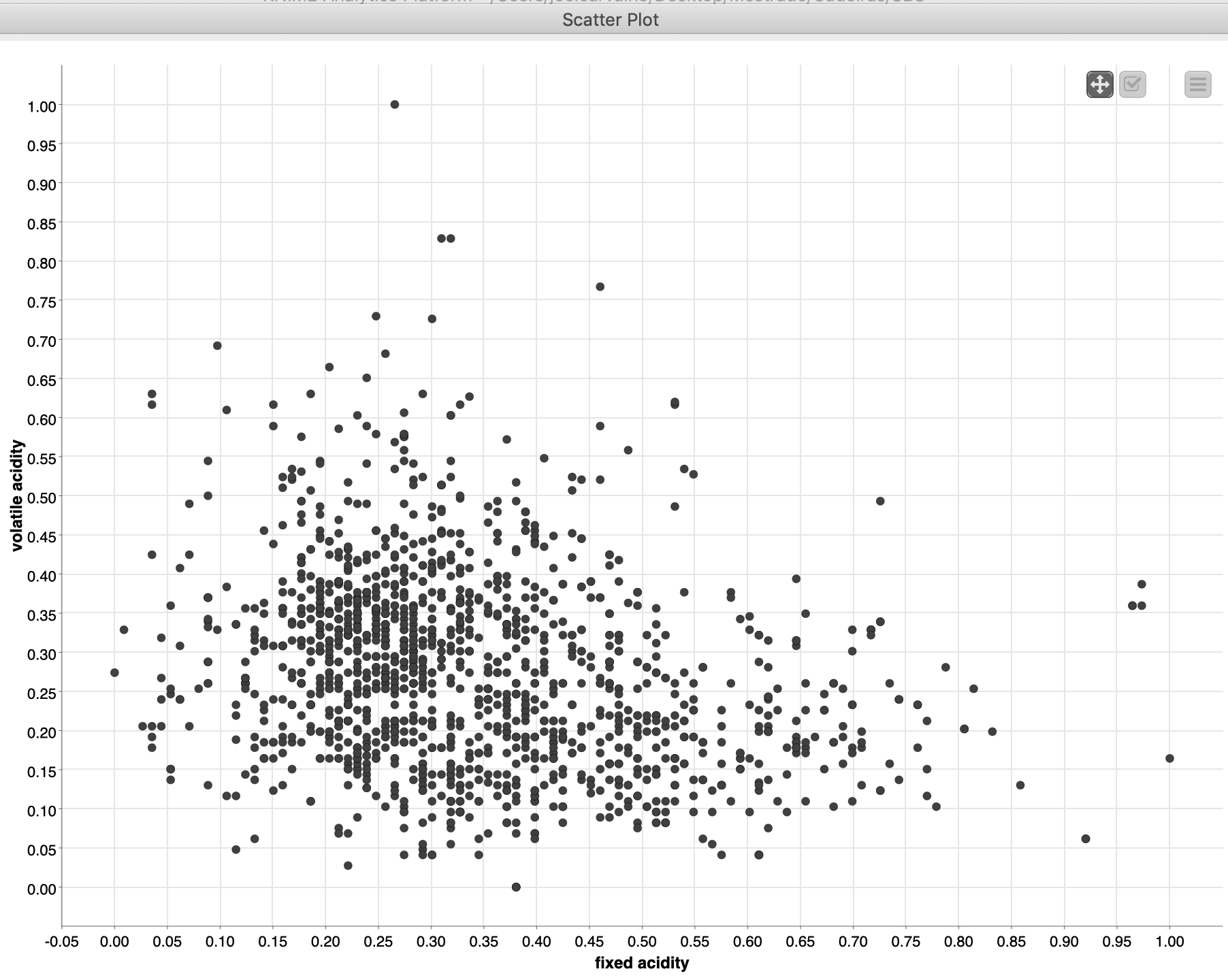


Figura 12 - Scatter Plot para visualizar os dados do PCA

**T4 - Segmentar o *dataset*:**

1. **Aplicando o método *k-means;***

Uma imagem com texto

Descrição gerada automaticamente

Figura 13 - Método k-means

1. **Atribuir diferentes cores por qualidade do vinho e diferentes formas aos *clusters*;**

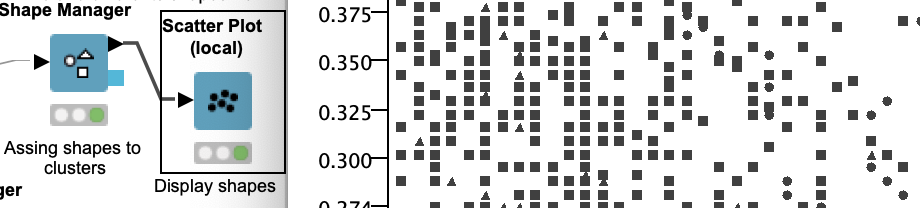
****

Figura 14 - Diferentes formas por clusters

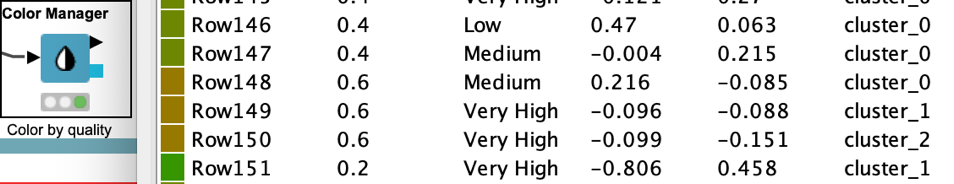


Figura 15 - Cores por qualidade do vinho

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

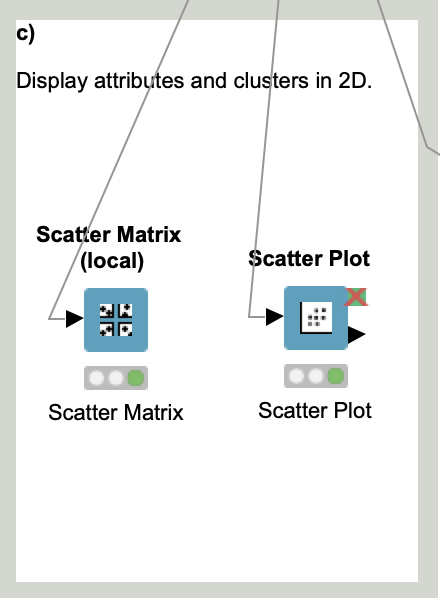
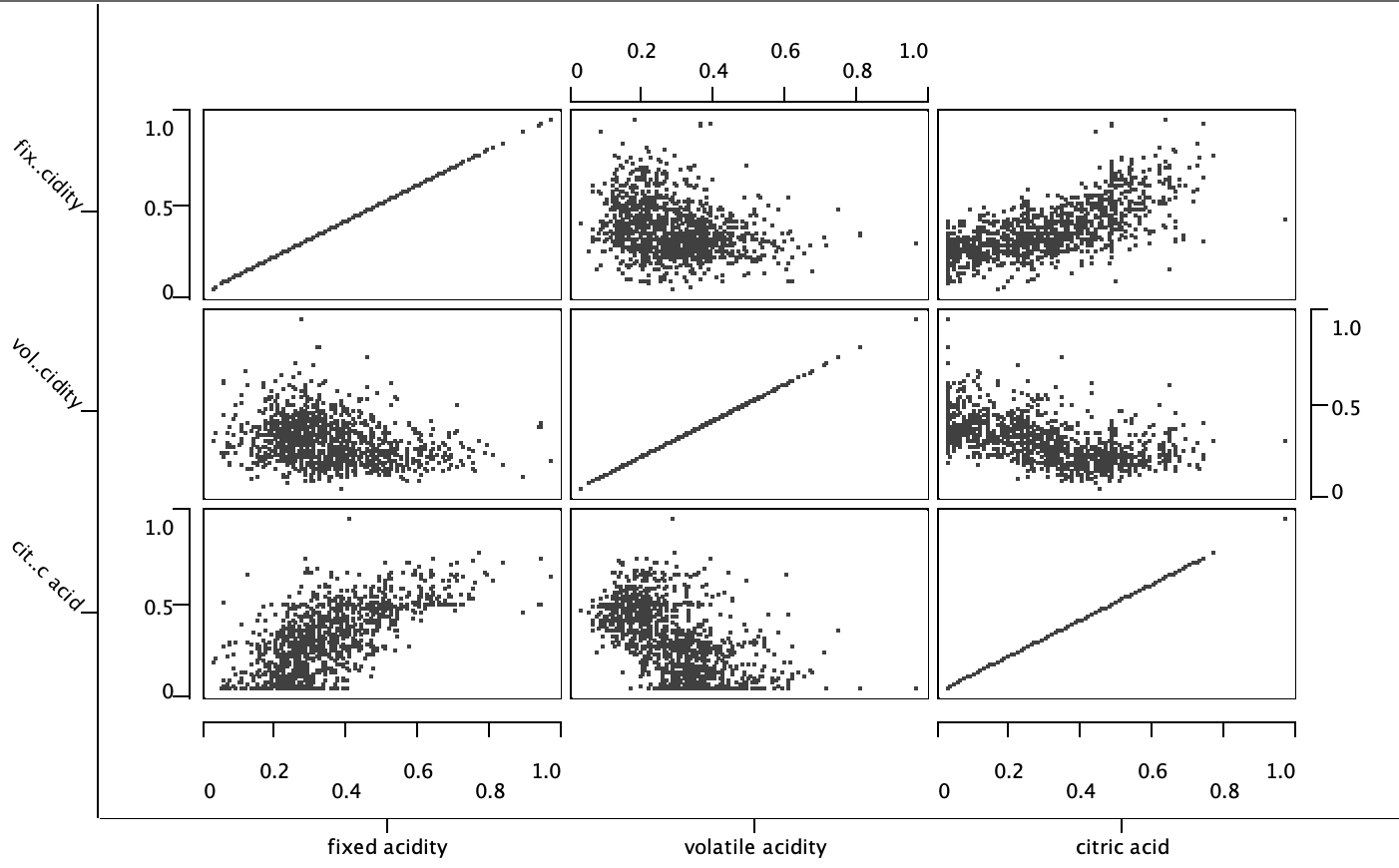


Figura 16 - Scatter matrixes

Figura 17 - Scatter plots e matrixes

1. **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;**

Uma imagem com mesa

Descrição gerada automaticamente

Figura 18 - Ler e tratar os dados, com base no modelo densenvolvido

1. **Guardar o resultado da atribuição num ficheiro csv**

Uma imagem com texto

Descrição gerada automaticamente

Figura 19 - Guardar o resultado num ficheiro csv

**T5 – Parametrizar o *workflow*, utilizando variáveis de fluxo para definir o número de *bins*, o número de *clusters* e os títulos dos gráficos criados;**

**Uma imagem com mesa

Descrição gerada automaticamente**

Figura 20 - Variáveis de fluxo

Depois da criação das variáveis de fluxo, foi necessário declará-las nos respetivos nodos, visível na Figura 21.



****

Figura - Atribuição da variável global "nr\_bins" ao nodo "Auto-Binner"

**T6 - Produzir o *workflow* de maneira a que seja possível visualizar, numa única página, todos os componentes visuais implementados;**

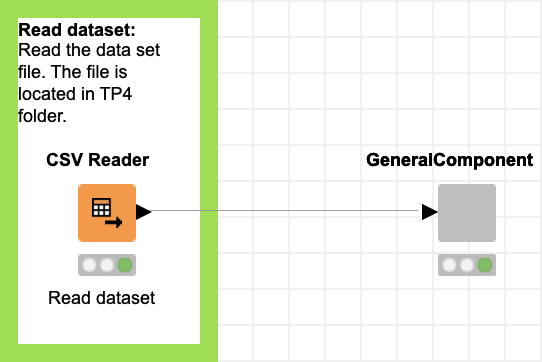


Figura - Vista geral do workflow



Figura - Vista geral do componente

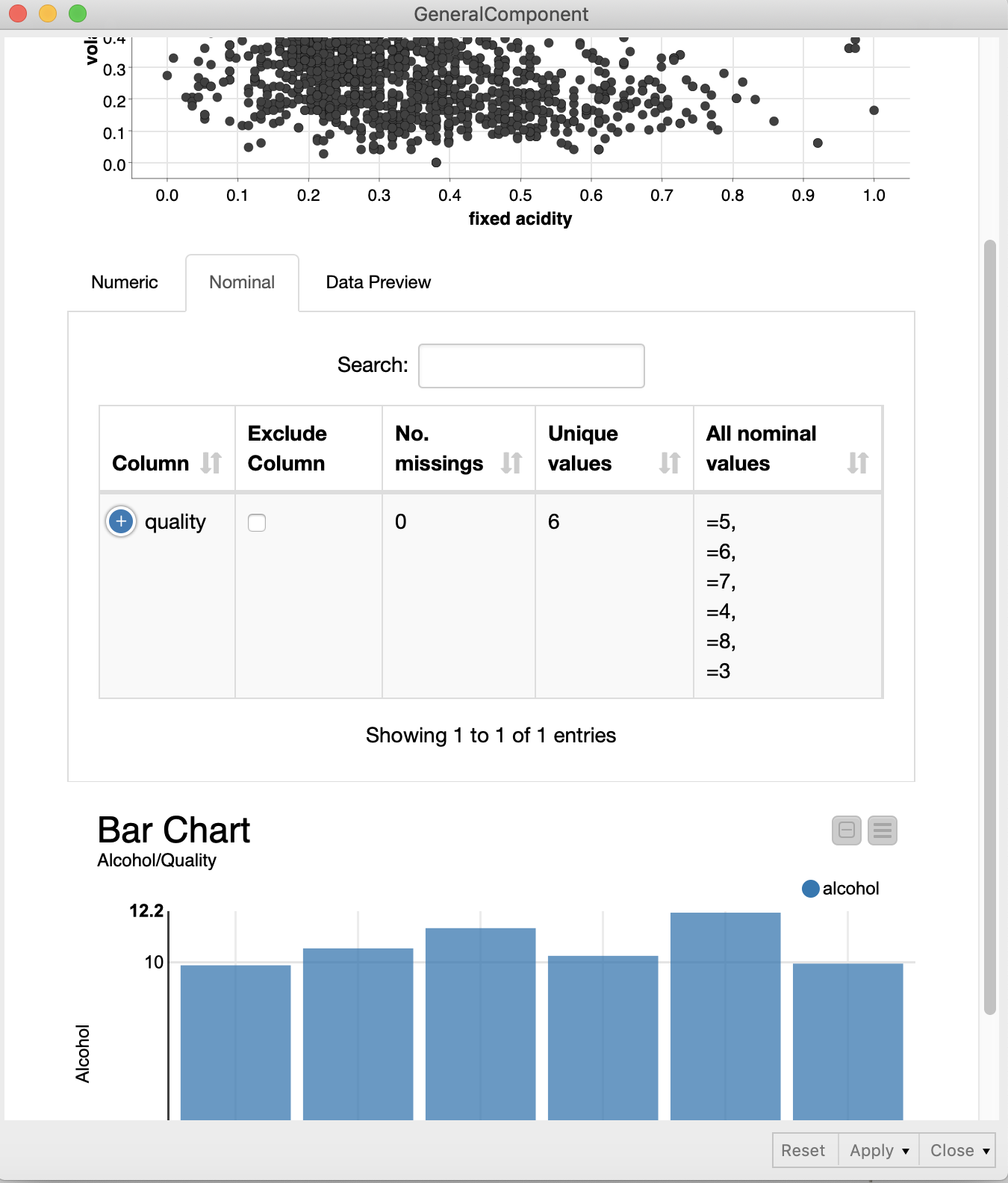


Figura - Output do componente

**T7 - Experimentar, avaliar e comparar outros métodos de segmentação.**

**Avaliação geral:**

* *Hierarchical Clustering* não lida tão bem como o *K-Means;*
* *Hierarchical Clustering* pode ser caro e lento;
* *Hierarchical Clustering* não tem número fixo de *clusters*;
* *DBSCAN* número de *clusters* não é necessário ser especificado;
* *DBSCAN* necessita de um raio e de um número de pontos mínimos;
* *DBSCAN* lida bem com dados ruidosos e *outliers*.

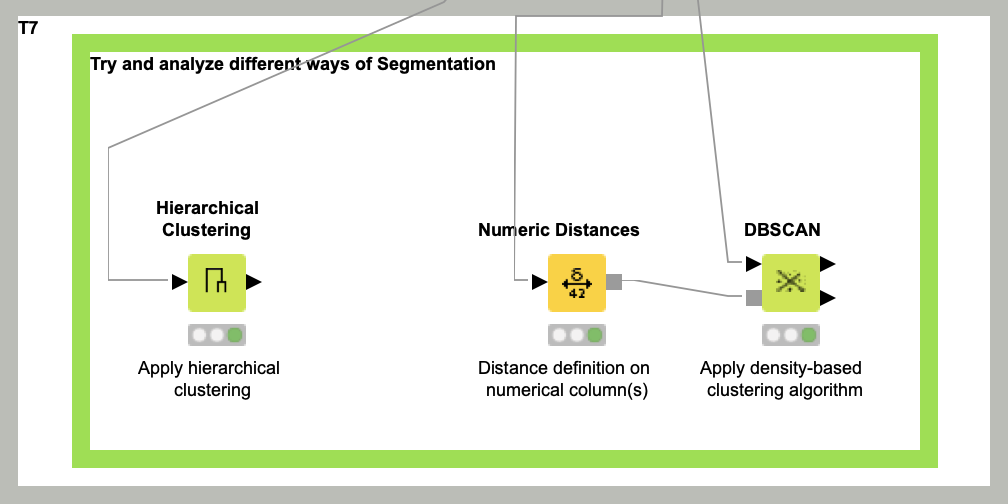


Figura - Diferentes métodos de segmentação