



Universidade do Minho
Escola de Engenharia

Mestrado em Engenharia Informática

Perfil de Machine Learning: Fundamentos e Aplicações

Sistemas Baseados em Similaridade

Trabalho Prático de Grupo – 2ª Parte

4º Ano, 1º Semestre

Ano letivo 2020/2021

Diogo Alexandre Rodrigues Lopes

PG42823

Fábio Gonçalves

PG42827

Joel Costa Carvalho

PG42837

Conteúdo

| | |
|--|-----------|
| Introdução | 3 |
| Tarefa 1. Análise, Tratamento e Exploração de dados do <i>dataset</i> (2019 Iowa Liquor Sales)..... | 4 |
| a. Sobre o <i>dataset</i> selecionado | 4 |
| b. Carregar, no <i>Knime</i> , o <i>dataset</i> selecionado | 7 |
| c. Aplicar de nodos, de modo a fazer Tratamento de Dados | 10 |
| d. Aplicar de nodos, de modo a fazer Análise de Dados | 12 |
| 1. Metanode <i>Data Listing Information</i> | 14 |
| 2. Metanode <i>Data Sales Information</i> | 16 |
| 3. Metanode <i>Data TOP's</i> | 18 |
| e. Aplicar de nodos, de modo a fazer a Otimização do Modelo | 22 |
| 4. Solução de Otimização Ótima do Modelo..... | 22 |
| 5. Evidência da Implementação de outras Soluções de Otimização do Modelo | 25 |
| Tarefa 2. Conceção e Implementação de um Sistema de Recomendação | 27 |
| a. Aplicar de nodos, baseados em Regras de Associação no Modelo..... | 27 |
| 1. Regras de Associação Implementadas baseadas na <i>feature</i> ' <i>Product Name</i> ' | 27 |
| 2. Regras de Associação Implementadas baseadas na <i>feature</i> ' <i>Category Name</i> ' | 33 |
| b. Aplicar de nodos, baseados em <i>Clusters</i> no Modelo | 36 |
| 1. <i>Clusters</i> baseados na <i>feature</i> ' <i>Category Name</i> ' | 39 |
| 2. <i>Clusters</i> baseados na <i>feature</i> ' <i>Product Name</i> ' | 40 |
| 3. <i>Clusters</i> baseados na <i>feature</i> ' <i>Price</i> ' | 41 |
| 4. <i>Clusters</i> baseados na <i>feature</i> ' <i>Month (name)</i> ' | 42 |
| 5. <i>Clusters</i> baseados na <i>feature</i> ' <i>City</i> ' | 43 |
| c. Avaliação do Sistema de Recomendação | 44 |
| a. Regras de Associação | 44 |
| Conclusão | 47 |

Introdução

Para a elaboração desta 2ª Parte do Trabalho Prático de Grupo, que consistia na conceção e implementação de um Sistema de Recomendação, tendo o mesmo como principal objetivo indicar ao utilizador, da maneira mais precisa e robusta, os diversos produtos que se encontram de acordo com as suas preferências. Para a implementação deste sistema, escolhemos um *dataset* da plataforma *Iowa Data (Internal)* (<https://mydata.iowa.gov>), mais concretamente, o *dataset 2019 Iowa Liquor Sales*. Com o intuito de tornar de mais fácil perceção todo o trabalho elaborado, resolvemos dividir a implementação do mesmo em duas partes. Inicialmente será feita uma análise, tratamento e exploração de dados do *dataset* selecionado. Depois disto, numa segunda fase, será elaborado o sistema de recomendação.

A primeira parte do trabalho, resumiu-se à estruturação e “limpeza” do *dataset*, desde remoção de colunas, *casting* e eliminação de valores nulos. Após o *dataset* estar devidamente estruturado, algumas análises e cálculos foram realizados sobre o mesmo. Relativamente à última fase e a mais significativa do projeto, o Sistema de Recomendação, foi dividido em dois tipos, sendo eles baseados em *Clusters* e em *Regras Associativas*. De salientar que este relatório contém todas estas etapas descritas, assim como os aspetos importantes relacionados com todo este processo. Os principais objetivos com a implementação/resolução deste trabalho são:

1. Consultar, Analisar e Selecionar um *dataset* sobre os quais seja possível desenvolver um Sistema de Recomendação;
2. Utilizar a plataforma *KNIME* para desenvolver um, ou vários, *workflows* para Exploração, Análise e Tratamento dos Dados assim como para Extração de Informação dos mesmos;
3. Desenvolver um Sistema de Recomendação seguindo uma abordagem híbrida implementando paradigmas como top-N, filtragem colaborativa, baseada em conteúdo, baseada em conhecimento, entre outros;
4. Implementar Métodos para Controlo/Avaliação da Qualidade das Recomendações;
5. O sistema desenvolvido deverá ser capaz de receber e tratar novos *inputs* de um utilizador, devolvendo uma, ou mais, recomendações. Deverão também ser implementados métodos que permitam ao utilizador perceber o porquê da recomendação.

Tarefa 1. Análise, Tratamento e Exploração de dados do *dataset* (2019 Iowa Liquor Sales)

a. Sobre o *dataset* selecionado

O *dataset* selecionado, disponível em <https://t.ly/OnWe>, contém uma visualização filtrada de dados com informações relativas à compra e venda de bebidas alcoólicas dos vendedores licenciados da Classe E por produto e data de compra para o ano civil de 2019. Esta licença de bebidas (Classe E), para supermercados, lojas de bebidas, lojas de conveniência, etc., permite a estes estabelecimentos comerciais a venda de bebidas para consumo externo em recipientes originais fechados, de modo a que no *dataset*, é possível observar por loja o volume de vendas de um determinado artigo/produto.

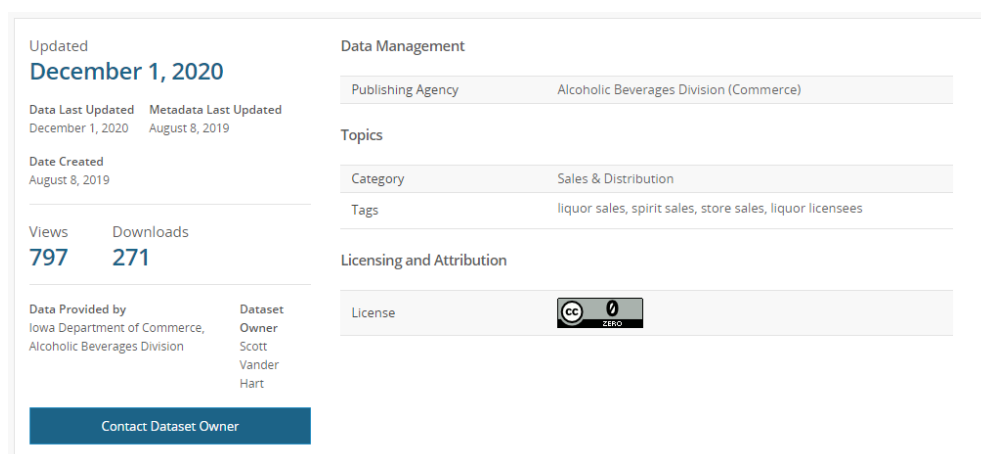


Figura nº1 – Informação sobre o *dataset*

Tal como é possível observar na imagem acima e como já foi referido, este conjunto de dados é fornecido pelo **Departamento de Comércio de Iowa** e a **Divisão de Bebidas Alcoólicas**. Os principais tópicos do mesmo são as vendas e distribuição de bebidas alcoólicas feitas por parte de estabelecimentos e vendedores devidamente licenciados. Salientamos também que o *dataset* foi criado em Agosto de 2019 e a última atualização feita foi em Dezembro de 2020.

Este *dataset* é composto por 24 colunas (*features*) com cerca de 2 380 000 registos, todos referentes ao ano de 2019.

| Nome da Coluna | Descrição | Tipo de Dados |
|----------------------------|---|---------------|
| Invoice/Item Number | Fatura Concatenada e Número da Linha associado ao pedido. Isto fornece um identificador único para os produtos incluídos no pedido | Plain Text |
| Date | Data do Pedido | Date & Time |
| Store Number | Número único atribuído à Loja que efetuou o pedido do Licor. | Plain Text |
| Store Name | Nome da Loja que encomendou o Licor | Plain Text |
| Address | Morada da Loja que encomendou o Licor | Plain Text |
| City | Cidade onde a Loja está localizada | Plain Text |
| Zip Code | Código Postal da loja que encomendou o Licor | Plain Text |
| Store Location | Local da Loja que efetuou o pedido do Licor. O endereço, cidade, estado e código postal são <i>geocodificados</i> para fornecer coordenadas geográficas | Point |
| County Number | Número do Município de Iowa para o Município onde a Loja que pediu a bebida está localizada | Plain Text |
| County | Município onde está localizada a Loja que fez o pedido | Plain Text |
| Category | Código da Categoria associado à bebida pedida | Plain Text |
| Category Name | Categoria da bebida pedida. | Plain Text |
| Vendor Number | Número do Fornecedor da empresa para a marca de bebida alcoólica solicitada | Plain Text |
| Vendor Name | Nome do Fornecedor da empresa para a marca de Licor pedido | Plain Text |
| Item Number | Número do Item para o produto de Licor individual pedido. | Plain Text |
| Item Description | Descrição do Licor individual pedido. | Plain Text |
| Pack | Número de Garrafas numa caixa da bebida pedida | Number |

| | | |
|------------------------------|--|--------|
| Bottle Volume (ml) | Volume de cada Garrafa de bebida solicitada em mililitros. | Number |
| State Bottle Cost | Valor que a Divisão de Bebidas Alcoólicas pagou por cada garrafa de Bebida Alcoólica pedida | Number |
| State Bottle Retail | Valor que a Loja pagou para cada Garrafa de Bebida Alcoólica pedida | Number |
| Bottles Sold | Número de Garrafas de Bebidas Alcoólicas encomendadas pela Loja | Number |
| Sale(Dollars) | Custo total do pedido de bebidas (número de garrafas multiplicado pelo custo de garrafas em cada estado) | Number |
| Volume Sold (Liters) | Volume total de Licor pedido em litros. (ou seja, (Volume da garrafa (ml) x garrafas vendidas) / 1.000) | Number |
| Volume Sold (Gallons) | Volume total de Licor pedido em gallons. (ou seja, (Volume da garrafa (ml) x garrafas vendidas) / 3785.411784) | Number |

Tabela nº1 – Colunas do *dataset*

| Invol... | Da... | Store... | Store... | Addr... | City | Zip C... | Store... | Coun... | County | Cate... | Cate... | Ven |
|-------------|------------|----------|--------------|---------------|--------------|----------|----------|---------|---------|---------|---------------|-----|
| INV-1668... | 01/02/2019 | 5286 | Sauce | 108, Colle... | Iowa City | 52240 | | 52 | JOHNSON | 1052100 | Imported ... | 420 |
| INV-1668... | 01/02/2019 | 5286 | Sauce | 108, Colle... | Iowa City | 52240 | | 52 | JOHNSON | 1022100 | Mixto Teq... | 395 |
| INV-1668... | 01/02/2019 | 5286 | Sauce | 108, Colle... | Iowa City | 52240 | | 52 | JOHNSON | 1012200 | Scotch W... | 055 |
| INV-1668... | 01/02/2019 | 2524 | Hy-Vee F... | 3500 Dod... | Dubuque | 52001 | | 31 | DUBUQUE | 1031100 | American... | 297 |
| INV-1669... | 01/02/2019 | 4449 | Kum & G... | 12041 Do... | Urbandale | 50322 | | 77 | POLK | 1031100 | American... | 434 |
| INV-1668... | 01/02/2019 | 2524 | Hy-Vee F... | 3500 Dod... | Dubuque | 52001 | | 31 | DUBUQUE | 1071100 | Cocktails ... | 065 |
| INV-1668... | 01/02/2019 | 2524 | Hy-Vee F... | 3500 Dod... | Dubuque | 52001 | | 31 | DUBUQUE | 1012100 | Canadian... | 260 |
| INV-1668... | 01/02/2019 | 2524 | Hy-Vee F... | 3500 Dod... | Dubuque | 52001 | | 31 | DUBUQUE | 1031100 | American... | 434 |
| INV-1669... | 01/02/2019 | 5151 | IDA Liquor | 500, Hwy ... | Ida Grove | 51445 | | 47 | IDA | 1012100 | Canadian... | 065 |
| INV-1669... | 01/02/2019 | 4116 | Lake View... | 223 Main ... | Lake View | 51450 | | 81 | SAC | 1062200 | White Rum | 035 |
| INV-1669... | 01/02/2019 | 5151 | IDA Liquor | 500, Hwy ... | Ida Grove | 51445 | | 47 | IDA | 1081600 | Whiskey ... | 421 |
| INV-1669... | 01/02/2019 | 5004 | Ida Grove... | 200 Susa... | Ida Grove | 51445 | | 47 | IDA | 1081200 | Cream Li... | 305 |
| INV-1669... | 01/02/2019 | 5151 | IDA Liquor | 500, Hwy ... | Ida Grove | 51445 | | 47 | IDA | 1031100 | American... | 434 |
| INV-1670... | 01/02/2019 | 3705 | Liquor Lo... | 507 1st A... | Rock Rapi... | 51246 | | 60 | LYON | 1031100 | American... | 260 |

< Previous Next >

Showing Rows 1 to 14 out of 2,380,345

Figura nº2 – Visão Geral do *dataset*

b. Carregar, no *Knime*, o *dataset* selecionado

Tendo em conta que o *dataset* tem cerca de 2 380 000 registos, e uma vez que para o processamento de dados no *KNIME* esta quantidade de registos tornava todo o processamento bastante demoroso, optámos por fazer uma a preparação no conjunto de dados. Com o objetivo de reduzir a quantidade de dados e aumentar a performance do modelo, decidimos extrair do *dataset* 2 000 registos mensais referentes aos meses de Janeiro, Fevereiro, Março, Abril, Maio e Junho do *dataset* original e assim construir um *dataset* novo para ser utilizado na construção do nosso modelo com 12 000 registos.

Para este processo de construção do *dataset*, criámos um componente (***Preparação Dataset***) onde se foi implementado o respetivo *workflow* para extração dos 2 000 registos mensais pretendidos. De acordo com a Figura nº 4 é possível visualizar que para fazer a extração dos registos pretendidos de cada mês, recorremos ao nodo ***Excel Reader (XLS)***. De modo a construir o novo conjunto de dados, aplicamos também o nodo ***Excel Writer (XLS)*** para extrair os registos mensais intencionados. Posteriormente fizemos uma junção manual, destes dados num único ficheiro do tipo ***.xlsx***.



Figura nº3 – Componente criado para a Preparação do *dataset*



Figura nº4 – Workflow implementado no Componente

À semelhança da configuração anterior, na Figura nº 5 em todos os nodos de leitura do *dataset*, foram lidos os 2 000 primeiros registos referentes a um determinado mês presentes no conjunto de dados original. De salientar que a exceção dos outros meses, quando efetuamos a extração dos dados referentes ao mês de Janeiro tivemos de ler 2001 registos (ver o destacado na imagem), uma vez que na primeira linha do *dataset* estão presentes os nomes das colunas que identificam os dados.

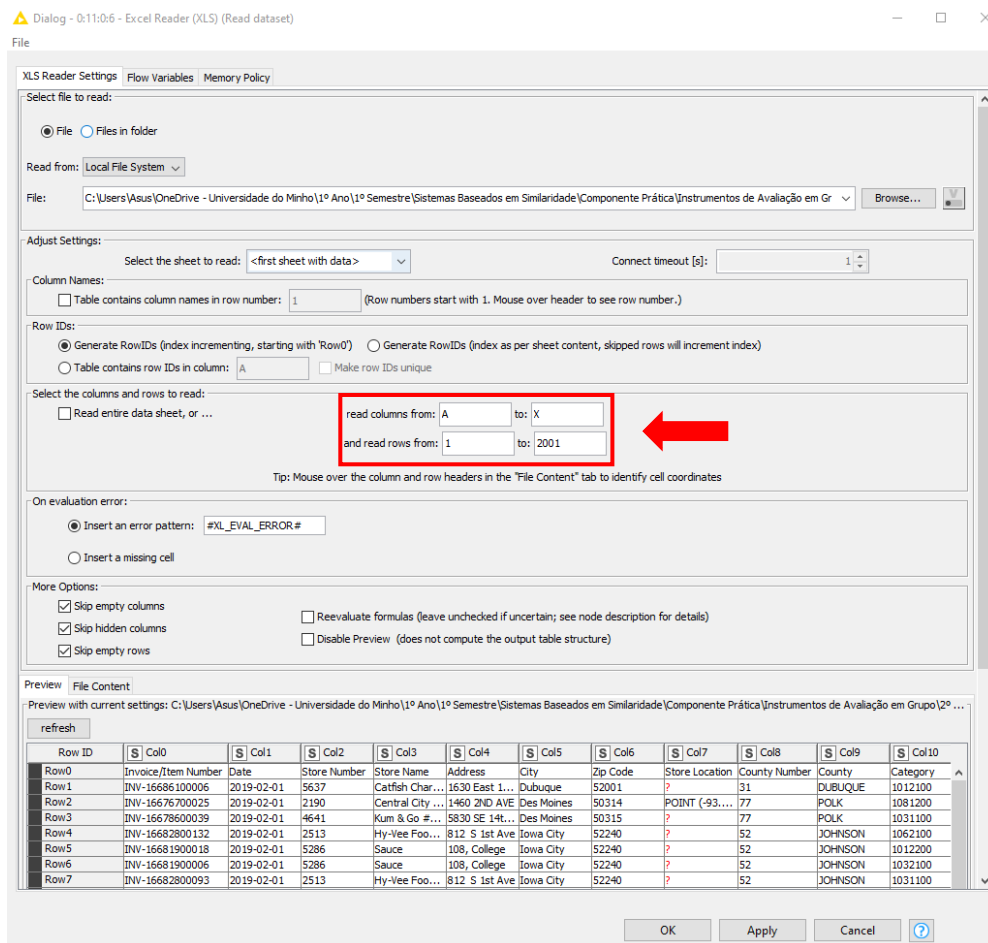


Figura nº5 – Configuração do nodo *Excel Reader (XLS)* (Mês de Janeiro)

| Invoice/Item Number | Date | Store Number | Store Name | Address | City | Zip Code | Store Location | County Number | County | Category |
|---------------------|------------|--------------|------------------|----------------|------------|----------|----------------|---------------|---------|----------|
| INV-16686100006 | 2019-02-01 | 5637 | Catfish Char... | 1630 East 1... | Dubuque | 52001 | ? | 31 | DUBUQUE | 1012100 |
| INV-16676700025 | 2019-02-01 | 2190 | Central City ... | 1460 2ND AVE | Des Moines | 50314 | POINT (-93... | 77 | POLK | 1081200 |
| INV-16678600039 | 2019-02-01 | 4641 | Kum & Go #... | 5830 SE 14t... | Des Moines | 50315 | ? | 77 | POLK | 1031100 |
| INV-16682800132 | 2019-02-01 | 2513 | Hy-Vee Foo... | 812 S 1st Ave | Iowa City | 52240 | ? | 52 | JOHNSON | 1062100 |
| INV-16681900018 | 2019-02-01 | 5286 | Sauce | 108, Colle | Iowa City | 52240 | ? | 52 | JOHNSON | 1012200 |
| INV-16681900006 | 2019-02-01 | 5286 | Sauce | 108, Colle | Iowa City | 52240 | ? | 52 | JOHNSON | 1032100 |
| INV-16682800093 | 2019-02-01 | 2513 | Hy-Vee Foo... | 812 S 1st Ave | Iowa City | 52240 | ? | 52 | JOHNSON | 1031100 |

Figura nº6 – Conjunto de dados extraído (Mês de Janeiro)

Posteriormente ao preparo do *dataset* que servirá para a implementação do nosso sistema, foi necessário carregar o mesmo no *KNIME*. Para isso, recorremos mais uma vez ao nodo *Excel Reader (XLS)* para ler os 12001 registos. Por fim, obtivemos o *output* pretendido que se encontra evidenciado na Figura nº8

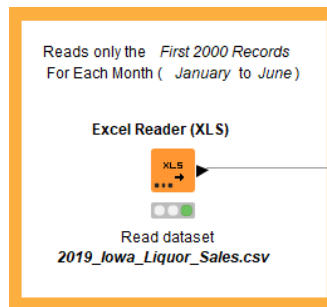


Figura nº7 – Leitura do dataset

| Table "C:\Users\AurOneDrive - Universidade do Minho\1º Ano\1º Semestre\Sistemas Baseados em Similaridade\Componente Prática\Instrumentos de Avaliação em Grupo\2º Trabalho Prático\KDM-WE Project\2019_Iowa_Liquor_Sales (2000 Registos Mensale).xlsx [Folha1]" - Rows: 11999 | | | | | | | | | | | | | | | | | | | | | | | | |
|---|-----------------|------------|--------------|--------------------|-------------------|--------------|------------|---------------|-------------|------------|------------|---------------------|-------------|--------------------|-------------|-----------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Soec - Columns: 24 | | | | | | | | | | | | Flow Variables | | | | | | | | | | | | |
| Row ID | S Invoice/Item | S Date | S Store N... | S Store Name | S Address | S City | S Zip Code | S Store L... | S County... | S County | S Category | S Category N... | S Vendor... | S Vendor Name | S Item N... | S Item D... | S Item D... | S Item D... | S Item D... | S Item D... | S Item D... | S Item D... | S Item D... | S Item D... |
| Row0 | INV-1668610006 | 2019-02-01 | 5637 | Carfish Charles | 1630 East 16th St | Dubuque | 52001 | ? | 31 | DUBUQUE | 1012100 | Canadian Whiskies | 260 | DIAGEO AMERICAS | 11297 | Crown Royal | | | | | | | | |
| Row1 | INV-1667670025 | 2019-02-01 | 2190 | Central City LI... | 1460 2ND AVE | Des Moines | 50314 | POINT (-93... | 77 | POLK | 1081200 | Cream Liqueurs | 305 | MHW LTD | 73052 | Rumchata | | | | | | | | |
| Row2 | INV-16676800039 | 2019-02-01 | 4641 | Kum & Go #57... | 5830 SE 14th St | Des Moines | 50315 | ? | 77 | POLK | 1031100 | American Vodkas | 301 | FIFTH GENERATI... | 38176 | Titos Handm... | | | | | | | | |
| Row3 | INV-16682800132 | 2019-02-01 | 2513 | Hy-Vee Food ... | 812 S 1st Ave | Iowa City | 52240 | ? | 52 | JOHNSON | 1062100 | Gold Rum | 434 | LUXCO INC | 45245 | Paramount ... | | | | | | | | |
| Row4 | INV-16681900018 | 2019-02-01 | 5286 | Sauce | 108, College | Iowa City | 52240 | ? | 52 | JOHNSON | 1012200 | Scotch Whiskies | 55 | SAZERAC NORTH... | 8824 | Lauder's | | | | | | | | |
| Row5 | INV-16681900006 | 2019-02-01 | 5286 | Sauce | 108, College | Iowa City | 52240 | ? | 52 | JOHNSON | 1032100 | Imported Vodkas | 260 | DIAGEO AMERICAS | 14457 | Henl One | | | | | | | | |
| Row6 | INV-16682800093 | 2019-02-01 | 2513 | Hy-Vee Food ... | 812 S 1st Ave | Iowa City | 52240 | ? | 52 | JOHNSON | 1031100 | American Vodkas | 301 | FIFTH GENERATI... | 38174 | Titos Handm... | | | | | | | | |
| Row7 | INV-16682800096 | 2019-02-01 | 2513 | Hy-Vee Food ... | 812 S 1st Ave | Iowa City | 52240 | ? | 52 | JOHNSON | 1031100 | American Vodkas | 301 | FIFTH GENERATI... | 38171 | Titos Handm... | | | | | | | | |
| Row8 | INV-16685400051 | 2019-02-01 | 2524 | Hy-Vee Food ... | 3500 Dodge St | Dubuque | 52001 | ? | 31 | DUBUQUE | 1062500 | Flavored Rum | 370 | PERNOD RICARD... | 42717 | Malibu Coco... | | | | | | | | |
| Row9 | INV-16685400024 | 2019-02-01 | 2524 | Hy-Vee Food ... | 3500 Dodge St | Dubuque | 52001 | ? | 31 | DUBUQUE | 1041100 | American Dry Gns | 370 | PERNOD RICARD... | 32237 | Seagrams E... | | | | | | | | |
| Row10 | INV-16685400061 | 2019-02-01 | 2524 | Hy-Vee Food ... | 3500 Dodge St | Dubuque | 52001 | ? | 31 | DUBUQUE | 1051100 | American Brandies | 125 | CEDAR RIDGE VI... | 53629 | Cedar Ridge... | | | | | | | | |
| Row11 | INV-16685400032 | 2019-02-01 | 2524 | Hy-Vee Food ... | 3500 Dodge St | Dubuque | 52001 | ? | 31 | DUBUQUE | 1032100 | Imported Vodkas | 65 | Jim Beam Brands | 34579 | Pinnacle | | | | | | | | |
| Row12 | INV-16685400047 | 2019-02-01 | 2524 | Hy-Vee Food ... | 3500 Dodge St | Dubuque | 52001 | ? | 31 | DUBUQUE | 1031200 | American Flavor... | 259 | Heaven Hill Brands | 41326 | Burnetts Citrus | | | | | | | | |
| Row13 | INV-16685400057 | 2019-02-01 | 2524 | Hy-Vee Food ... | 3500 Dodge St | Dubuque | 52001 | ? | 31 | DUBUQUE | 1062200 | White Rum | 434 | LUXCO INC | 46351 | Hawkeye Lig... | | | | | | | | |
| Row14 | INV-16684200023 | 2019-02-01 | 5359 | Fareway Stor... | 8400 NICC Drive | Peosta | 52068 | ? | 31 | DUBUQUE | 1062500 | Flavored Rum | 260 | DIAGEO AMERICAS | 72903 | Captain Mor... | | | | | | | | |
| Row15 | INV-16690300019 | 2019-02-01 | 4449 | Kum & Go #12... | 12041 Douglas | Urbandale | 50322 | ? | 77 | POLK | 1031100 | American Vodkas | 434 | LUXCO INC | 36308 | Hawkeye Vo... | | | | | | | | |
| Row16 | INV-16692900009 | 2019-02-01 | 4116 | Lake View Foods | 223 Main St | Lake View | 51450 | ? | 81 | SAC | 1031100 | American Vodkas | 434 | LUXCO INC | 36308 | Hawkeye Vo... | | | | | | | | |
| Row17 | INV-16693100026 | 2019-02-01 | 5151 | IDA Liquor | 500, Hwy 175 | Ida Grove | 51445 | ? | 47 | IDA | 1081600 | Whiskey Liqueur | 421 | SAZERAC COMPA... | 64866 | Fireball Cnn... | | | | | | | | |
| Row18 | INV-16693000015 | 2019-02-01 | 5004 | Ida Grove Foo... | 200 Susan Lawr... | Ida Grove | 51445 | ? | 47 | IDA | 1081200 | Cream Liqueurs | 305 | MHW LTD | 73055 | Rumchata | | | | | | | | |
| Row19 | INV-16692600009 | 2019-02-01 | 2200 | Sac Liquor Store | 619 E Main St | Sac City | 50583 | POINT (-94... | 81 | SAC | 1031100 | American Vodkas | 434 | LUXCO INC | 36304 | Hawkeye Vo... | | | | | | | | |
| Row20 | INV-16697700016 | 2019-02-01 | 3993 | New Star Liqu... | 1625 West 4th St | Waterloo | 50701 | ? | 7 | BLACK HAWK | 1011100 | Blended Whiskies | 297 | Laird & Company | 23823 | Five Star | | | | | | | | |
| Row21 | INV-16693100033 | 2019-02-01 | 5151 | IDA Liquor | 500, Hwy 175 | Ida Grove | 51445 | ? | 47 | IDA | 1022100 | Mixto Tequila | 395 | PROVINO | 89199 | Jose Cuervo... | | | | | | | | |
| Row22 | INV-16702300030 | 2019-02-01 | 3705 | Liquor Locker | 507 1st Ave #100 | Rock Rapids | 51246 | ? | 60 | LYON | 1031100 | American Vodkas | 260 | DIAGEO AMERICAS | 37997 | Smooff 80prf | | | | | | | | |
| Row23 | INV-16701400012 | 2019-02-01 | 4255 | Fareway Stor... | 512 8th SE | Orange City | 51041 | ? | 84 | SIOUX | 1031100 | American Vodkas | 421 | SAZERAC COMPA... | 36978 | Nikolai Vodka | | | | | | | | |
| Row24 | INV-16702300026 | 2019-02-01 | 3705 | Liquor Locker | 507 1st Ave #100 | Rock Rapids | 51246 | ? | 60 | LYON | 1011200 | Straight Bourbon... | 65 | Jim Beam Brands | 19667 | Jim Beam | | | | | | | | |
| Row25 | INV-16702300004 | 2019-02-01 | 3705 | Liquor Locker | 507 1st Ave #100 | Rock Rapids | 51246 | ? | 60 | LYON | 1012100 | Canadian Whiskies | 115 | CONSTELLATION... | 10350 | Black Velvet | | | | | | | | |
| Row26 | INV-16693000022 | 2019-02-01 | 5004 | Ida Grove Foo... | 200 Susan Lawr... | Ida Grove | 51445 | ? | 47 | IDA | 1011100 | Blended Whiskies | 492 | WESTERN SPIRIT... | 27474 | Bird Dog Bla... | | | | | | | | |
| Row27 | INV-16697300003 | 2019-02-01 | 5504 | Neighborhood... | 2102 Lafayette St | Waterloo | 50703 | POINT (-92... | 7 | BLACK HAWK | 1011100 | Blended Whiskies | 297 | Laird & Company | 23626 | Five Star PET | | | | | | | | |
| Row28 | INV-16684200007 | 2019-02-01 | 5359 | Fareway Stor... | 8400 NICC Drive | Peosta | 52068 | ? | 31 | DUBUQUE | 1062400 | Spiced Rum | 260 | Heaven Hill Brands | 43028 | Admiral Nels... | | | | | | | | |
| Row29 | INV-16685400044 | 2019-02-01 | 2524 | Hy-Vee Food ... | 3500 Dodge St | Dubuque | 52001 | ? | 31 | DUBUQUE | 1031100 | American Vodkas | 209 | MIDWAS MICRO D... | 38381 | Wisconsin Cl... | | | | | | | | |
| Row30 | INV-16687700006 | 2019-02-01 | 3666 | Target Store... | 3400 Edgewood... | Cedar Rapids | 52404 | POINT (-91... | 57 | LYNN | 1012100 | Canadian Whiskies | 115 | CONSTELLATION... | 11776 | Black Velvet | | | | | | | | |
| Row31 | INV-16676900001 | 2019-02-01 | 4617 | Lickety Liquor | 2501 HUBBELL ... | Des Moines | 50317 | POINT (-93... | 77 | POLK | 1041100 | American Dry Gns | 434 | LUXCO INC | 31658 | Paramount Gn | | | | | | | | |
| Row32 | INV-16678400017 | 2019-02-01 | 3696 | Wal-Mart 172... | 5101 SE 14th St | Des Moines | 50313 | ? | 77 | POLK | 1062400 | Spiced Rum | 260 | DIAGEO AMERICAS | 43334 | Captain Mor... | | | | | | | | |
| Row33 | INV-16693100019 | 2019-02-01 | 5151 | IDA Liquor | 500, Hwy 175 | Ida Grove | 51445 | ? | 47 | IDA | 1031100 | American Vodkas | 434 | LUXCO INC | 36306 | Hawkeye Vo... | | | | | | | | |
| Row34 | INV-16702300041 | 2019-02-01 | 3705 | Liquor Locker | 507 1st Ave #100 | Rock Rapids | 51246 | ? | 60 | LYON | 1062200 | White Rum | 35 | BACARDI USA INC | 43127 | Bacardi Sup... | | | | | | | | |
| Row35 | INV-16681900029 | 2019-02-01 | 5286 | Sauce | 108, College | Iowa City | 52240 | ? | 52 | JOHNSON | 1011100 | Blended Whiskies | 260 | DIAGEO AMERICAS | 25607 | Seagrams 7 ... | | | | | | | | |

Figura nº8 – Conjunto de Dados

c. Aplicar de nodos, de modo a fazer Tratamento de Dados

Após visualização do conjunto de dados, ficou perceptível que era necessário efetuar o tratamento de algumas *features* presentes no *dataset*. Assim implementamos no nosso *workflow* um metanodo (**Data Treatment**) onde efetuamos toda a parte de tratamento de dados.

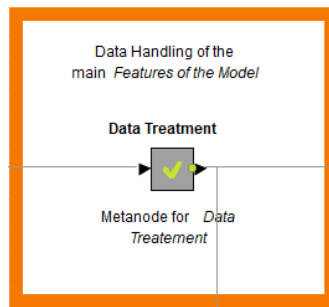


Figura nº9 – Metanodo criado para o Tratamento de Dados do *dataset*

Ao nível de tratamento de dados, e no que a *feature* 'Date' diz respeito, fizemos o *cast* da mesma para o formato *Date&Time* recorrendo ao nodo **String to Date&Time**. Posteriormente, e uma vez que achamos que seria relevante, extraímos da data o mês (número e nome), o dia do ano, do mês e da semana.

Dado o extenso número de *features* que o nosso conjunto de dados possuía, e com o objetivo de melhorar a performance do mesmo, resolvemos eliminar algumas *features*. Recorrendo ao nodo **Column Filter** e aplicando-o no *workflow*, as *features* 'Date', 'Store Name', 'Store Location', 'County Number' e 'Volume Sold (Gallons)' foram removidas dos conjuntos de dados. É importante salientar que o que esteve na génese da remoção destas *features* foi o facto de as mesmas replicarem informação já presente, por exemplo, mesmo removendo o nome da loja conseguimos identificar a mesma através do seu número identificador único. Depois de efetuado o processo anterior, e tendo verificado que existiam alguns registos com valores em falta, fizemos a remoção destes recorrendo ao nodo **Missing Value**.

Tendo em conta que para a otimização dos dados para a implementação do sistema de recomendação é necessário que os atributos sejam do tipo *number*, foi necessário fazer a conversão de algumas *features*. Deste modo, convertimos as *features* 'Store Number', 'Category', 'Vendor Number', 'Item Number' e 'Pack' para o tipo de dados *integer*. As *features* 'Bottle Volume (ml)', 'State Bottle Cost', 'State Bottle Retail', 'Bottles Sold', 'Sales (Dollares)' e 'Volume Sold (Liters)' foram convertidas para o formato do tipo *double*, devido ao facto de possuírem casas decimais.

Ainda no que ao tratamento de dados diz respeito, numa fase mais adiantada do trabalho, nomeadamente na parte do cálculo do *MAE* e *MSE*, apercebemo-nos que o *KNIME* entrava em conflito com o nome de algumas *features* devido ao facto de estas possuírem

'...Number'. Deste modo, recorrendo ao nodo **Column Rename**, solucionámos este problema renomeando as colunas em questão colocando um '_' antes da palavra 'Number' de acordo com o exemplo seguinte 'Store_Number'.

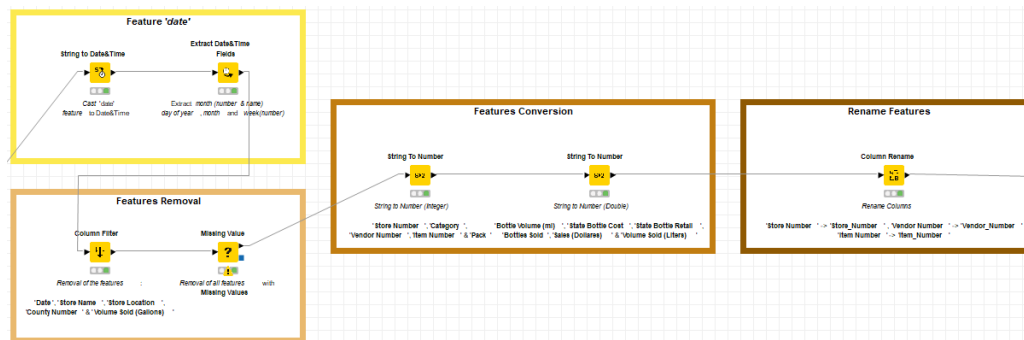


Figura nº10 – Workflow implementado para o Tratamento de Dados

d. Aplicar de nodos, de modo a fazer Análise de Dados

Depois de feito o tratamento de dados, foi necessário fazer uma análise ao *dataset*, com o objetivo de perceber algumas das particularidades do mesmo, de modo a facilitar na implementação do sistema de recomendação. Assim implementamos no nosso *workflow* um componente (**Data Analysis**) onde efetuamos toda a parte de tratamento de dados.

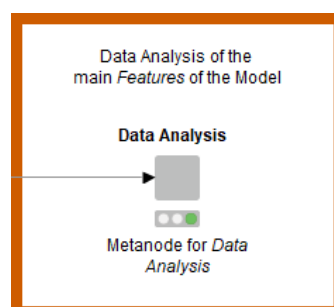


Figura nº11 – Componente criado para o Análise de Dados do *dataset*

Ao nível de análise de dados, numa primeira abordagem decidimos tentar perceber um pouco mais sobre a relação entre as diversas *features* recorrendo ao nodo **Rank Correlation**. De seguida, recorrendo a dois nodos **GroupBy**, onde fizemos 5 *Unique Count* com o objetivo de

saber o número de Licores, número de Categorias de Licores, número de Lojas, Cidades e Vendedores presentes no *dataset*. Com isto ficámos a perceber que existem 1161 Licores distribuídos por 46 Categorias diferentes. Sobre as Lojas existem no total 845, distribuídas por 270 Cidades e 86 Vendedores, o que permite concluir que o mesmo vendedor pode ter diversas lojas. Ao nível da relação entre *features* (**Figura nº13**), conseguimos perceber que existem *features*, algumas já esperadas, que se relacionam bem umas com as outras, como é o caso da *feature Bottles Sold* e a *feature Sale*. Por outro lado, existem *features* que não se relacionam tão bem, como é o caso das *features Pack* e *State Bottle Cost*.

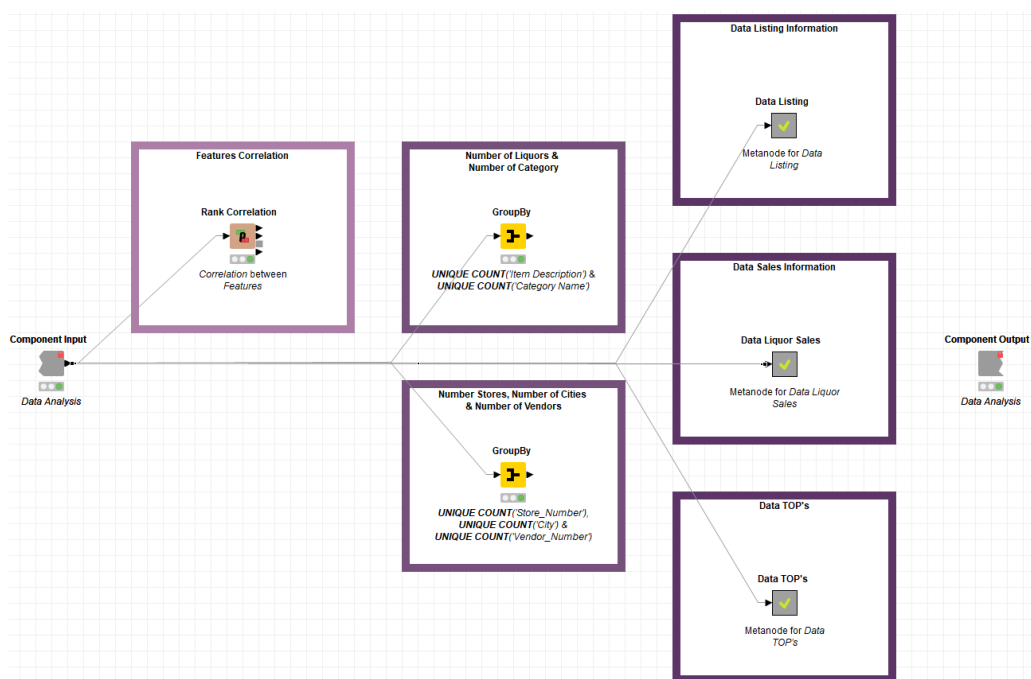


Figura nº12 – Workflow implementado para o Análise de Dados

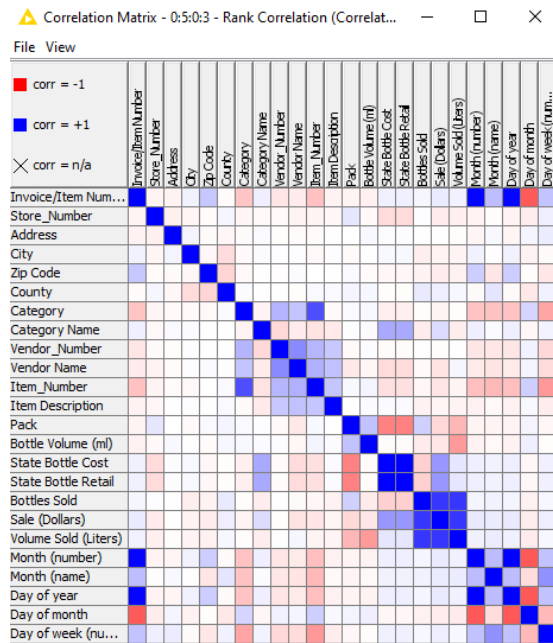


Figura nº13 – Relação entre *features* do *dataset*

De acordo com o apresentado na **Figura nº12**, é perceptível que para além dos já enunciados nodos, foram também criados mais 3 metanodos com o objetivo de fazer uma análise de dados mais detalhada, que permitirá tirar conclusões sobre como poderá ser implementado o sistema de recomendação. Cada um destes, e os seus respetivos resultados encontram-se devidamente explicados nos tópicos seguintes.

1. Metanode *Data Listing Information*

O objetivo da implementação deste metanode, é como o próprio nome indica fazer a listagem da informação contida no *dataset*. Com a implementação de nodos feita neste *workflow*, conseguimos perceber visualizar as seguintes listagens de dados:

- Listagem de Licores por Categoria
- Listagem de Licores por Vendedor
- Listagem de Licores por Cidade
- Listagem de Categorias por Vendedor
- Listagem de Categorias por Cidade

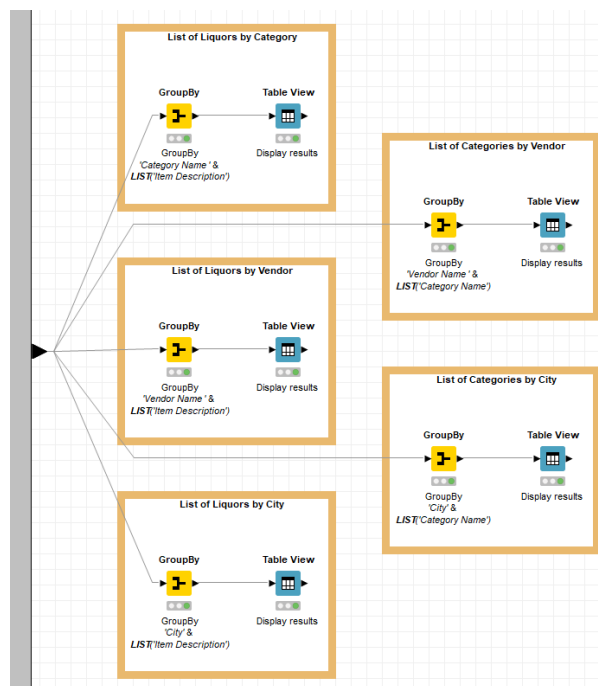


Figura nº14 – Workflow implementado no Metanode *Data Listing Information*

Group table - 0:5:0:119:60 - GroupBy (GroupBy)

File Edit Hilite Navigation View

Table "default" - Rows: 46 Spec - Columns: 2 Properties Flow Variables

| Row ID | [S] Category Name | [...] List(Item Description) |
|--------|-------------------------|---|
| Row0 | 100% Agave Tequila | [Patron Silver Mini,Casamigos Reposado,Margaritaville Silver Tequila,...] |
| Row1 | Aged Dark Rum | [Ron Zacapa 23 Year,Bacardi Black,Cross Keys Rum,...] |
| Row2 | American Brandies | [Cedar Ridge Apple Brandy,Korbel Brandy,Korbel Brandy,...] |
| Row3 | American Cordials ... | [Kinky Blue Mini,Kinky Mini,Paramount Amaretto,...] |
| Row4 | American Cordials ... | [SOOH Dr McGillicuddys Butterscotch] |
| Row5 | American Distilled S... | [Saints N Sinners Apple Pie,Evan Williams Honey,Evan Williams Honey...] |
| Row6 | American Dry Gins | [Seagrams Extra Dry Gin,Paramount Gin,Five O'Clock Gin,...] |
| Row7 | American Flavored ... | [Burnetts Citrus,Burnetts Mango Pineapple,Smirnoff Strawberry,...] |
| Row8 | American Schnapps | [Arrow Peach Schnapps,Dekuyper Hot Damn!,Paramount Peach Schn... |
| Row9 | American Sloe Gins | [Paramount Sloe Gin,Paramount Sloe Gin] |
| Row10 | American Vodkas | [Titos Handmade Vodka,Smirnoff 80prf Mini,Titos Handmade Vodka,...] |
| Row11 | Blended Whiskies | [Five Star,Bird Dog Blackberry,Five Star PET,...] |
| Row12 | Bottled in Bond Bo... | [Wild Turkey Rare Breed,Evan Williams Bottled in Bond,Wild Turkey R... |
| Row13 | Canadian Whiskies | [Crown Royal,Black Velvet Toasted Caramel,Black Velvet,...] |

Figura nº15 – Listagem de Licores por Categoria

Group table - 0:5:0:119:71 - GroupBy (GroupBy)

File Edit Hilite Navigation View

Table "default" - Rows: 270 Spec - Columns: 2 Properties Flow Variables

| Row ID | [S] City | [...] List(Category Name) |
|--------|----------|--|
| Row0 | Ackley | [Imported Flavored Vodka,Whiskey Liqueur,American Vodkas,...] |
| Row1 | Adair | [American Vodkas,American Vodkas] |
| Row2 | Adel | [Straight Bourbon Whiskies,Canadian Whiskies,Cocktails /RTD,...] |
| Row3 | Afton | [Spiced Rum,American Schnapps,American Vodkas,...] |
| Row4 | Albion | [Whiskey Liqueur,Whiskey Liqueur,Cream Liqueurs,...] |
| Row5 | Alden | [Coffee Liqueurs,American Vodkas,Canadian Whiskies,...] |
| Row6 | Algona | [Canadian Whiskies,American Flavored Vodka,American Distille... |
| Row7 | Allison | [Spiced Rum,Canadian Whiskies,Spiced Rum,...] |
| Row8 | Alta | [Straight Bourbon Whiskies,Canadian Whiskies,Canadian Whis... |
| Row9 | Altoona | [Triple Sec,Imported Flavored Vodka,American Flavored Vodka... |
| Row10 | Amana | [Spiced Rum,American Cordials & Liqueur,American Brandies,...] |
| Row11 | Ames | [American Schnapps,American Vodkas,American Vodkas,...] |
| Row12 | Anamosa | [Whiskey Liqueur,Imported Vodkas,Straight Bourbon Whiskies... |
| Row13 | Anita | [Mixto Tequila,American Schnapps,Mixto Tequila,...] |

Figura nº16 – Listagem de Licores por Cidade

Com esta análise de dados efetuada, e após observação dos resultados obtidos, destacamos a Listagem de Licores por Categoria (**Figura nº15**) e a Listagem de Licores por Cidade (**Figura nº16**), pois achamos interessante para a conceção do sistema de recomendação o utilizador introduzir uma categoria de licores, e mediante as regras implementadas, o sistema sugerir um *TOP* de licores de acordo com a categoria inserida. Por outro lado, também pensamos que seja interessante, o utilizador introduzir o nome da cidade, e o sistema recomendar por exemplo, os mais vendidos nessa determinada cidade.

2. Metanode *Data Sales Information*



Figura nº17 – Workflow implementado no Metanode *Data Sales Information*

Como o próprio nome indica, este *metanode* foi implementado com o objetivo de perceber um pouco mais sobre as vendas dos licores, pois alguns dos aspetos aqui observados poderão estar na génese das regras do sistema de recomendação. Com a implementação de nodos feita neste *workflow*, conseguimos visualizar as seguintes informações:

- Vendas de cada Licor
- Vendas de cada Vendedor
- Vendas de cada Cidade

- Vendas de cada Mês
- Volume Vendido em Litros de cada Licor
- Volume Vendido em Litros por cada Vendedor
- Volume Vendido em Litros por cada Cidade
- Volume Vendido em Litros por cada Mês
- Garrafas Vendidas de cada Licor
- Garrafas Vendidas por cada Vendedor
- Garrafas Vendidas por cada Cidade
- Garrafas Vendidas por cada Mês

JavaScript Table View

Sales for Each Liquor

Sales(Dollars)

Show: 10 entries Search:

| RowID | Item Description | Sum(Sale (Dollars)) |
|---------|-----------------------------------|---------------------|
| Row141 | Black Velvet | 104558.44000000013 |
| Row315 | Crown Royal | 100867.27999999996 |
| Row577 | Jack Daniels Old #7 Black Label | 74703.08000000001 |
| Row324 | Crown Royal Regal Apple | 70843.60999999997 |
| Row1096 | Titos Handmade Vodka | 49780.95 |
| Row590 | Jameson | 42705.49999999998 |
| Row693 | Kirkland Signature French Vodka | 37514.88 |
| Row236 | Captain Morgan Spiced Barrel | 33912 |
| Row690 | Kirkland Signature American Vodka | 31347 |
| Row603 | Jim Beam | 26933.450000000004 |

Showing 1 to 10 of 1,161 entries

Previous 1 2 3 4 5 ... 117 Next

Figura nº18 – Vendas de cada Licor

JavaScript Table View

Volume Sold for Each Liquor

Volume Sold (Liters)

Show: 10 entries Search:

| RowID | Item Description | Sum(Volume Sold (Liters)) |
|---------|-----------------------------------|---------------------------|
| Row141 | Black Velvet | 11126.530000000008 |
| Row690 | Kirkland Signature American Vodka | 4725 |
| Row315 | Crown Royal | 3446.779999999984 |
| Row540 | Hawkeye Vodka | 3234.8599999999999 |
| Row693 | Kirkland Signature French Vodka | 3024 |
| Row1096 | Titos Handmade Vodka | 2791.37 |
| Row577 | Jack Daniels Old #7 Black Label | 2597.4500000000003 |
| Row324 | Crown Royal Regal Apple | 2344.5199999999999 |
| Row964 | Seagrams 7 Crown | 2208.19 |
| Row236 | Captain Morgan Spiced Barrel | 2198 |

Showing 1 to 10 of 1,161 entries

Previous 1 2 3 4 5 ... 117 Next

Figura nº19 – Volume Vendido de cada Licor

JavaScript Table View

Bottles Sold for Each Liquor

Show 10 entries

Search:

| RowID | Item Description | Sum(Bottles Sold) |
|-------|--|-------------------|
| Row0 | 1800 Anejo | 6 |
| Row1 | 1800 Peach Margarita | 63 |
| Row2 | 1800 Reposado | 30 |
| Row3 | 1800 Silver | 35 |
| Row4 | 1800 Ultimate Mango Margarita | 6 |
| Row5 | 1800 Ultimate Margarita | 18 |
| Row6 | 2 Gingers | 26 |
| Row7 | 3-Oaks Distillery Straight Bourbon Whiskey | 6 |
| Row8 | 360 Bing Cherry | 12 |
| Row9 | 360 Double Chocolate | 3 |

Showing 1 to 10 of 1,161 entries

Previous 1 2 3 4 5 ... 117 Next

Figura nº20 – Garrafas Vendidas de cada Licor

3. Metanode *Data TOP's*



Figura nº21 – Workflow implementado no Metanode *Data TOP's*

Ao contrário dos outros 2 metanodes implementados, cujo seu principal objetivo era ajudar a estudar o *dataset* de modo a perceber alguns aspetos relevantes que possam estar na base do sistema de recomendação, este centra-se essencialmente na análise de dados, uma vez com a implementação destes *TOP's*, conseguimos visualizar e perceber alguns dados estatísticos com o volume de vendas e número de garrafas vendidas de cada vendedor, destacando-se pela positiva o **DIAGEO AMERICAS** ao nível de garrafas de licor vendidas, e pela negativa de acordo com o volume de vendas, o vendedor **BAD BEAR ENTERPRISES LLC**. Ao nível das vendas destaca-se pela positiva o vendedor **DIAGEO AMERICAS** com um total de vendas de 442 056,31\$, e pela negativa o vendedor **Jackson Hole Still Works** com um valor de vendas de 26,25\$.

Filtered - 0:5:0:120:80 - Row Filter (Filter 25)

File Edit Hilite Navigation View

Table "default" - Rows: 25 Spec - Columns: 2 Properties Flow Variables

| Row ID | S Vendor Name | D Sum(Bottles Sold) |
|--------|---------------------------|---------------------|
| Row17 | DIAGEO AMERICAS | 22,456 |
| Row9 | CONSTELLATION BRANDS INC | 13,422.182 |
| Row69 | SAZERAC COMPANY INC | 12,825 |
| Row40 | Jim Beam Brands | 11,488 |
| Row41 | LUXCO INC | 9,703 |
| Row70 | SAZERAC NORTH AMERICA | 6,544 |
| Row58 | PERNOD RICARD USA | 5,522 |
| Row42 | Laird & Company | 5,194 |
| Row6 | Brown Forman Corp. | 5,121 |
| Row51 | McCormick Distilling Co. | 4,168 |
| Row33 | Heaven Hill Brands | 3,815 |
| Row21 | E & J Gallo Winery | 3,742 |
| Row60 | PROXIMO | 3,035 |
| Row3 | BACARDI USA INC | 2,594 |
| Row23 | FIFTH GENERATION INC | 2,584 |
| Row64 | Phillips Beverage | 2,354 |
| Row46 | MISA Imports Inc | 1,903.728 |
| Row48 | MOET HENNESSY USA | 1,319 |
| Row74 | Sky Spirits Inc | 925 |
| Row43 | Levecke Corporation | 812.7 |
| Row50 | Mast-Jagermeister US, Inc | 695 |
| Row63 | Patron Spirits Company | 692 |
| Row36 | Infinium Spirits | 678 |
| Row85 | William Grant & Sons Inc | 666 |
| Row66 | REMY COINTREAU USA INC | 545 |

Figura nº22 – Vendedores com mais Garrafas Vendidas

Filtered - 0:5:0:120:92 - Row Filter (Filter 25)

File Edit Hilite Navigation View

Table "default" - Rows: 25 Spec - Columns: 2 Properties Flow Variables

| Row ID | S Vendor Name | D Sum(Bottles Sold) |
|--------|---------------------------|---------------------|
| Row4 | BAD BEAR ENTERPRISES LLC | 1 |
| Row34 | Hood River Distillers | 1 |
| Row38 | Jackson Hole Still Works | 1 |
| Row83 | W J Deutsch & Sons LTD | 1 |
| Row25 | Famous Brands | 2 |
| Row32 | Hawaii Sea Spirits LLC | 2 |
| Row57 | PALM BAY INTERNATIONAL | 2 |
| Row62 | Paterno Imports LTD | 2 |
| Row79 | USA Wine West, LLC | 2 |
| Row13 | Cats Eye Distillery | 4 |
| Row71 | SERRALLES USA | 4 |
| Row0 | 3-Oaks Distillery, LLC | 6 |
| Row73 | Shaw-Ross International | 6 |
| Row78 | Two Sons Imports LLC | 6 |
| Row80 | VBJ Beverages LLC | 6 |
| Row82 | Vision Wine & Spirit LLC | 6 |
| Row1 | AIKO IMPORTERS INC | 9 |
| Row28 | GoAmericaGo Beverages LLC | 9 |
| Row31 | HOTALING & CO | 9 |
| Row2 | AMERICAN VINTAGE BEVERAGE | 12 |
| Row5 | BRECKENRIDGE DISTILLERY | 12 |
| Row8 | CHATHAM IMPORTS INC | 12 |
| Row11 | Caribbean Distillers, LLC | 12 |
| Row15 | Colorado Gold Distillery | 12 |
| Row44 | Levecke Corporation JJB | 12 |

Figura nº23 – Vendedores com menos Garrafas Vendidas

Sobre os Municípios com mais vendas destacamos o **POLK** com um valor de 387 557,51\$ e o com menos vendas é o **DAVIS** com um valor de 91,32\$. Ao nível de garrafas vendidas, o município que vende mais é novamente o **POLK** com um total de 27 120 garrafas, e o que vende menos é o novamente o **DAVIS** com 6 garrafas vendidas.

Filtered - 0:5:0:120:93 - Row ...

File Edit Hilite Navigation View

| Properties | | Flow Variables |
|----------------------------|------------|-----------------------|
| Table "default" - Rows: 25 | | Spec - Columns: 2 |
| Row ID | S County | D Sum(Sale (Dollars)) |
| Row84 | POLK | 387,557.51 |
| Row85 | POTTAWATTA | 140,427.95 |
| Row32 | Dallas | 117,028.68 |
| Row63 | LINN | 115,409.3 |
| Row57 | JOHNSON | 89,941.14 |
| Row92 | SCOTT | 75,390.37 |
| Row88 | Polk | 71,480.59 |
| Row61 | KOSSUTH | 64,710.51 |
| Row31 | DUBUQUE | 48,876.66 |
| Row5 | BLACK HAWK | 38,773 |
| Row95 | STORY | 36,419.28 |
| Row19 | CLARKE | 27,940.76 |
| Row75 | MUSCATINE | 27,298.71 |
| Row62 | LEE | 26,843.28 |
| Row101 | WARREN | 23,183.35 |
| Row81 | PAGE | 23,044.74 |
| Row105 | WINNESHIEK | 20,519.55 |
| Row98 | UNION | 20,151.3 |
| Row45 | HARDIN | 18,606.18 |
| Row30 | DICKINSON | 17,852.36 |
| Row43 | HAMILTON | 17,729.16 |
| Row22 | CLINTON | 17,727.38 |
| Row67 | MADISON | 17,037.45 |
| Row71 | MILLS | 15,235.27 |
| Row108 | WRIGHT | 15,167.68 |

Figura nº24 – Municípios com mais Vendas

Filtered - 0:5:0:120:99 - Row ...

File Edit Hilite Navigation View

| Properties | | Flow Variables |
|----------------------------|------------|-----------------------|
| Table "default" - Rows: 25 | | Spec - Columns: 2 |
| Row ID | S County | D Sum(Sale (Dollars)) |
| Row26 | DAVIS | 91.32 |
| Row2 | AUDUBON | 294.4 |
| Row77 | Marshall | 442.8 |
| Row89 | Poweshiek | 494.69 |
| Row99 | VAN BUREN | 503.22 |
| Row50 | Henry | 511.98 |
| Row3 | Adair | 584.68 |
| Row106 | WOODBURY | 667.14 |
| Row35 | EMMET | 674.58 |
| Row96 | Scott | 690.3 |
| Row64 | LOUISA | 847.42 |
| Row53 | Iowa | 854.28 |
| Row16 | CERRO GORD | 855.13 |
| Row42 | GUTHRIE | 1,004.11 |
| Row109 | Wapello | 1,013.57 |
| Row78 | Mitchell | 1,217.2 |
| Row60 | KEOKUK | 1,248.01 |
| Row73 | MONONA | 1,253.91 |
| Row39 | Fayette | 1,287.35 |
| Row76 | Madison | 1,381.58 |
| Row55 | JASPER | 1,402.22 |
| Row34 | Des Moines | 1,435.38 |
| Row103 | WEBSTER | 1,454.54 |
| Row33 | Delaware | 1,482.4 |
| Row59 | Jackson | 1,577.6 |

Figura nº25 – Municípios com menos Vendas

Sobre as lojas com mais garrafas vendidas destacamos a com o identificador 2512 (**Hy-Vee Wine and Spirits / Iowa City**) com um total de 3 755 garrafas vendidas, e a com o indentificador 3565 (**Hartig Drug Store #10 / Iowa City**) com menos garrafas vendidas

Filtered - 0:5:0:...

File Edit Hilite Navigation View

| Properties | | Flow Variables |
|----------------------------|--|-------------------|
| Table "default" - Rows: 25 | | Spec - Columns: 2 |

| Row ID | I Store_... | D Sum(Bo... |
|--------|-------------|-------------|
| Row18 | 2512 | 3,755 |
| Row83 | 2633 | 3,246 |
| Row198 | 3814 | 2,716.428 |
| Row509 | 4829 | 2,313 |
| Row74 | 2620 | 1,829 |
| Row55 | 2585 | 1,775 |
| Row42 | 2561 | 1,578 |
| Row602 | 5144 | 1,470 |
| Row2 | 2190 | 1,416 |
| Row79 | 2626 | 1,367 |
| Row129 | 3477 | 1,350 |
| Row186 | 3773 | 1,332 |
| Row344 | 4312 | 1,318 |
| Row104 | 2666 | 1,004 |
| Row235 | 3952 | 991 |
| Row132 | 3494 | 936 |
| Row105 | 2670 | 919 |
| Row125 | 3443 | 875 |
| Row21 | 2521 | 857 |
| Row64 | 2602 | 844 |
| Row78 | 2625 | 756 |
| Row3 | 2191 | 703 |
| Row90 | 2648 | 691 |
| Row257 | 4073 | 689 |
| Row143 | 3618 | 684 |

Figura nº26 – Lojas com mais Garrafas Vendidas

Filtered - 0:5:0:...

File Edit Hilite Navigation View

| Properties | | Flow Variables |
|----------------------------|--|-------------------|
| Table "default" - Rows: 25 | | Spec - Columns: 2 |

| Row ID | I Store_... | D Sum(Bo... |
|--------|-------------|-------------|
| Row140 | 3565 | 2 |
| Row351 | 4336 | 2 |
| Row801 | 5644 | 2 |
| Row359 | 4371 | 3 |
| Row495 | 4792 | 3 |
| Row561 | 5028 | 3 |
| Row725 | 5447 | 3 |
| Row774 | 5584 | 3 |
| Row271 | 4132 | 4 |
| Row276 | 4146 | 4 |
| Row373 | 4411 | 4 |
| Row569 | 5056 | 4 |
| Row594 | 5112 | 4 |
| Row328 | 4267 | 5 |
| Row339 | 4299 | 5 |
| Row115 | 3013 | 6 |
| Row163 | 3690 | 6 |
| Row181 | 3742 | 6 |
| Row249 | 4023 | 6 |
| Row338 | 4296 | 6 |
| Row353 | 4359 | 6 |
| Row474 | 4718 | 6 |
| Row480 | 4736 | 6 |
| Row502 | 4804 | 6 |
| Row640 | 5255 | 6 |

Figura nº27 – Lojas com menos Garrafas Vendidas

e. Aplicar de nodos, de modo a fazer a Otimização do Modelo

No que a otimização de dados diz respeito, implementamos 4 abordagens diferentes, 2 com o **Método de Cotovelo (Elbow Method)**, onde aplicamos diversos nodos com o objetivo de calcular o *MAE* e o *MSE*. Nas outras abordagens recorremos ao **Clustering k-Means**, onde à semelhança do anterior, calculamos o *MAE* e *MSE*. O objetivo destas implementações, foi conseguir determinar qual a otimização ótima para o nosso modelo, ou seja, aquela que apresenta um menor valor de erro tendo todas por base configurações semelhantes, como por exemplo o número de *loops*. Nos tópicos seguintes apresentamos a solução de otimização ótima escolhida, e evidenciamos a implementação e teste de outras otimizações efetuadas.

4. Solução de Otimização Ótima do Modelo

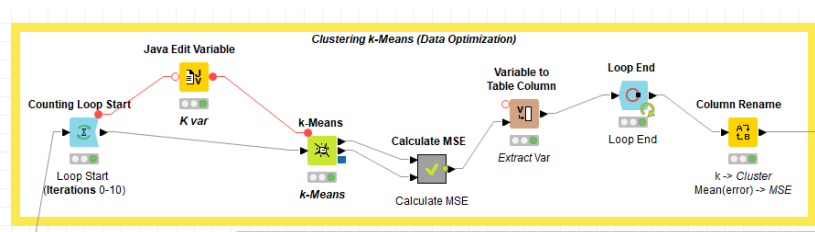


Figura nº28 – Nodos implementados para a Otimização de Dados

De acordo com a Figura apresentada, é possível verificar que aquela que consideramos como solução ótima de otimização, é recorrendo ao *Clustering k-Means* e efetuando o cálculo do *MSE*, uma vez que é aquela que apresenta um menor valor para o erro, como podemos verificar na Tabela nº 2.

Ao nível de configurações implementadas, foram efetuados 10 *loops*. Recorrendo ao nó **Java Edit Variable** definimos o nosso *k*, que posteriormente é utilizado como uma *flow variable* no nó **k-Means** para definir o número de *clusters*. Ainda na configuração deste nó filtramos as *features* que não são relevantes para o nosso Sistema de Recomendação baseado em *clusters*, tais como a 'Store_Number', 'Vendor_Number', 'Item_Number', 'State Bottle Cost', 'Bottles Sold', 'Sale (Dollars)',

‘Volume Sold (Liters)’, ‘Month (number)’, ‘Day of year’, ‘Day of month’ e ‘Day of week (number)’.

| Tipo de Otimização | Cálculo do Erro | Erro Obtido |
|---------------------------|-----------------|-------------|
| Elbow Method | <i>MAE</i> | 0.064 |
| Elbow Method | <i>MSE</i> | 0.001 |
| Clustering k-Means | <i>MAE</i> | 0.072 |
| Clustering k-Means | <i>MSE</i> | 0.001 |

Tabela nº2 – Erro Obtido na Otimização do Modelo

Output variables - 0:85 - Java Edit Variable (K var)

| Index | Owner ID | Name | Value |
|--------|----------|------------------|-------------------------------|
| 0:0:85 | | k | 10 |
| 0:0:86 | | maxIterations | 10 |
| 0:0:86 | | currentIteration | 9 |
| 0:0:86 | | Loop-Execute | |
| 0:0:86 | | Loop (0) | |
| 0 | | krime.workspace | C:\Users\Asus\krime-workspace |

Figura nº29 – Output obtido com o nodo Java Edit Variable

Clusters - 0:96 - k-Means (K-Means)

| Row ID | D Store... | D Category | D Vendor... | D Item_N... | D Pack | D Bottle V... | D State B... | D State B... | D Bottles ... | D Sale (D... | D Volume ... | D Month (...) | D Day of ... | D Day of ... | D Day of ... |
|-----------|------------|------------|-------------|-------------|--------|---------------|--------------|--------------|---------------|--------------|--------------|---------------|--------------|--------------|--------------|
| cluster_0 | 0.232 | 0.04 | 0.245 | 0.048 | 0.25 | 0.772 | 0.068 | 0.068 | 0.007 | 0.002 | 0.001 | 0 | 0 | 1 | 0.4 |
| cluster_1 | 0.245 | 0.039 | 0.24 | 0.044 | 0.241 | 0.042 | 0.074 | 0.074 | 0.013 | 0.004 | 0.002 | 0 | 0 | 1 | 0.4 |
| cluster_2 | 0.289 | 0.022 | 0.222 | 0.035 | 0.218 | 0.848 | 0.071 | 0.071 | 0.008 | 0.003 | 0.001 | 0.294 | 0.287 | 0 | 0.8 |
| cluster_3 | 0.296 | 0.02 | 0.219 | 0.031 | 0.213 | 0.111 | 0.075 | 0.075 | 0.011 | 0.004 | 0.003 | 0.304 | 0.297 | 0 | 0.8 |
| cluster_4 | 0.236 | 0.807 | 0.235 | 0.159 | 0.193 | 0.861 | 0.121 | 0.121 | 0.011 | 0.006 | 0.002 | 0.493 | 0.483 | 0 | 0.605 |
| cluster_5 | 0.304 | 0.022 | 0.219 | 0.034 | 0.985 | 0.209 | 0.012 | 0.012 | 0.031 | 0.002 | 0.001 | 0.623 | 0.618 | 0 | 0.643 |
| cluster_6 | 0.255 | 0.034 | 0.246 | 0.046 | 0.222 | 0.467 | 0.072 | 0.072 | 0.011 | 0.004 | 0.002 | 0.6 | 0.593 | 0 | 0 |
| cluster_7 | 0.23 | 0.032 | 0.239 | 0.048 | 0.245 | 0.784 | 0.068 | 0.068 | 0.009 | 0.003 | 0.001 | 0.8 | 0.793 | 0 | 0.4 |
| cluster_8 | 0.253 | 0.017 | 0.176 | 0.019 | 0.251 | 0.757 | 0.074 | 0.074 | 0.009 | 0.004 | 0.001 | 1 | 1 | 0 | 1 |
| cluster_9 | 0.24 | 0.027 | 0.202 | 0.035 | 0.156 | 0.006 | 0.093 | 0.093 | 0.014 | 0.007 | 0.005 | 0.897 | 0.893 | 0 | 0.69 |

Figura nº30 – Clusters obtidos (Nodo k-Means)

Para o cálculo do *MSE*, implementamos o *workflow* apresentado na imagem seguinte. O nodo **Joiner**, efetua a junção dos elementos presentes no conjunto de dados com os *clusters* anteriormente definidos. Com o nodo **Column Filter**, filtramos as *features* que consideramos irrelevantes para o cálculo do erro, tais como o ‘Invoice/Item Number’, ‘Store_Number’, ‘Address’, ‘City’, ‘Zip Code’, ‘County’, ‘Category Name’, ‘Vendor_Number’, ‘Vendor Name’, ‘Item_Number’, ‘Item Description’, ‘State Bottle

Cost, *Bottles Sold*, *Sale (Dolars)*, *Volume Sold (Liters)*, *Month (number)*, *Month (name)*, *Day of year*, *Day of month* e *Day of week (number)*. É importante salientar que no cálculo do erro efetuado em todas as otimizações testadas, removemos as mesmas *features* de modo a não existirem discrepâncias no cálculo da função matemática que nos permite obter o erro.

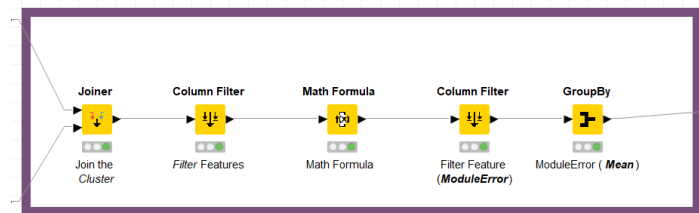


Figura nº31 – Nodos implementados para o Cálculo do MSE

Recorrendo ao nodo **Math Formula** aplicamos a respetiva fórmula para o cálculo do MSE, de acordo com o evidenciado na imagem seguinte. De seguida, fizemos uma filtragem de *features* de modo a ficarmos apenas com o **ModuleError**, que é o valor calculado e definido no nodo anterior, para cada um dos registos presentes no *dataset*. Por fim, recorrendo ao nodo **GroupBy** fizemos uma agregação manual da coluna **ModuleError**, onde o tipo de agregação foi a média, e obtivemos o valor de 0.009 já enunciado na Tabela nº2.

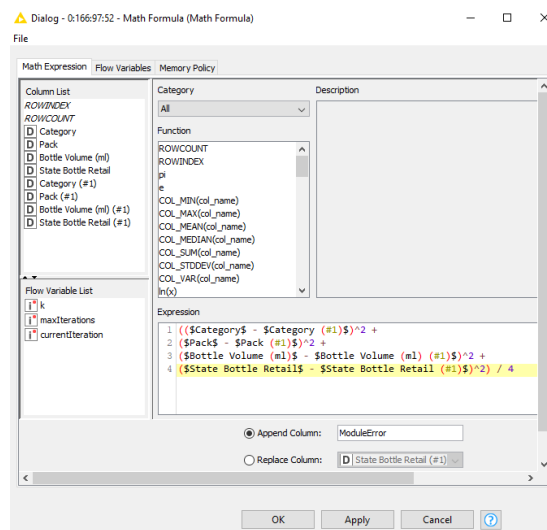


Figura nº32 – Configurações aplicadas no nodo **Math Formula**

Group table - 0:97:50 - GroupBy (Mod... — □ ×

File Edit Hilite Navigation View

Table "default" - Rows: 1 Spec - Column: 1 Properties Flow Variables

| Row ID | D Mean(ModuleError) |
|--------|---------------------|
| Row0 | 0.009 |

Figura nº33 – Média do Erro Obtido

5. Evidência da Implementação de outras Soluções de Otimização do Modelo

De acordo com o que já foi referido, implementamos e testamos várias formas de otimização do conjunto de dados com o objetivo de identificar a otimização ótima. Depois de termos escolhido o nosso método de otimização, armazenamos os restantes num componente (**Data Optimization (TESTS)**). Ao nível de configurações destas outras otimizações, as mesmas são semelhantes à ótima, uma vez que possuem o mesmo número de *loops*, e as suas únicas diferenças residem no método aplicado (*Elbow Method* ou *Clustering k-Means*) e nos casos do cálculo do *MAE*, uma vez que a fórmula matemática é diferente da do cálculo do *MSE*.

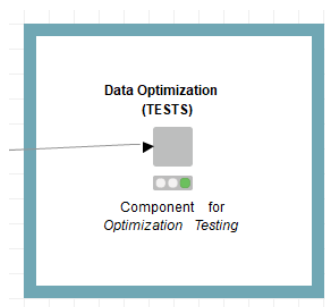


Figura nº34 – Componente criado para Teste de outras Otimizações

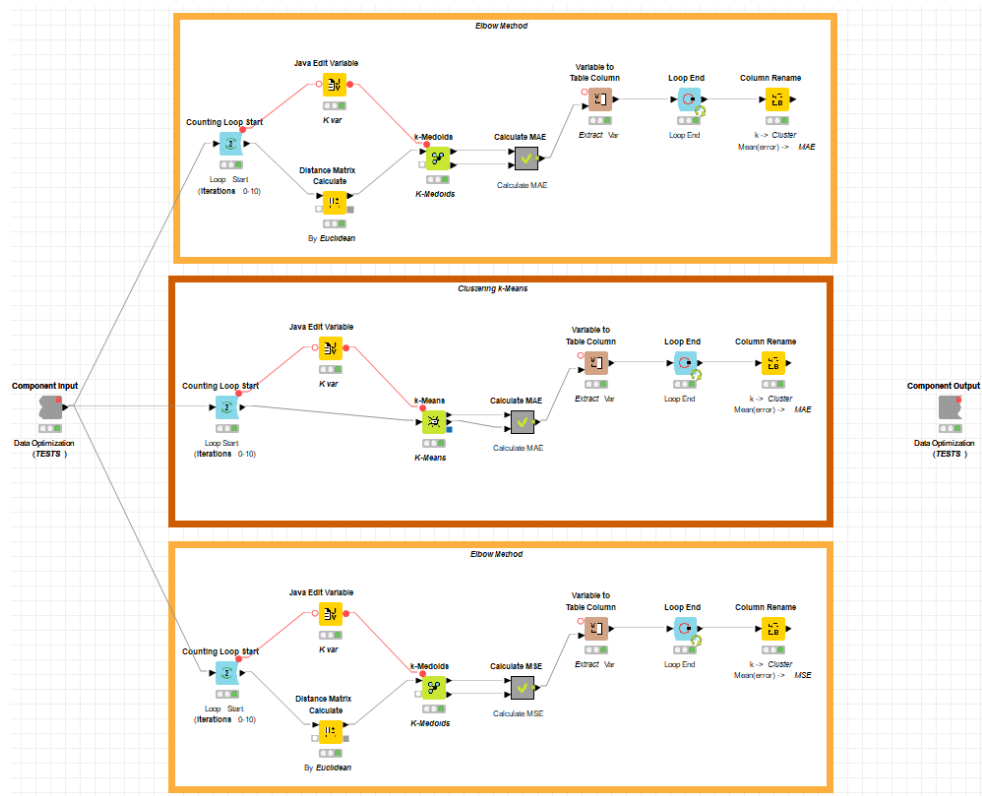


Figura nº35 – Workflow implementado para Teste de outras Otimizações

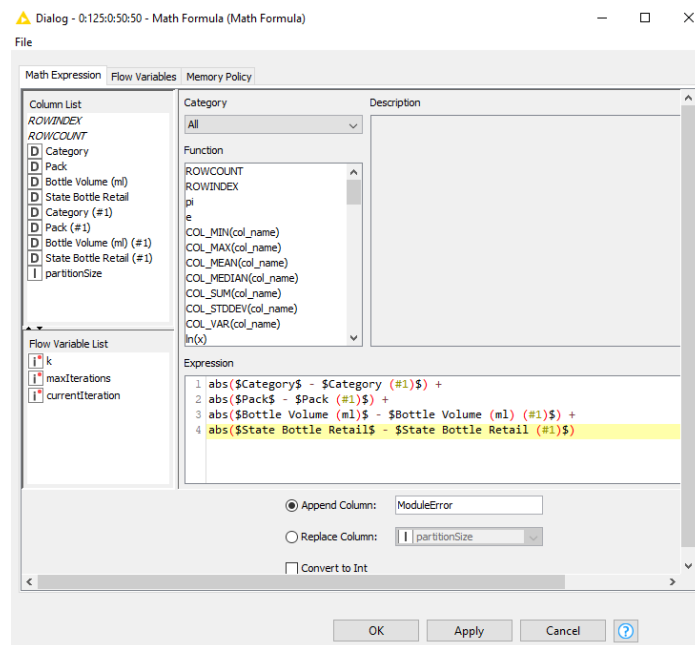


Figura nº36 – Configurações aplicadas no nodo Math Formula (Cálculo MAE)

Tarefa 2. Conceção e Implementação de um Sistema de Recomendação

a. Aplicar de nodos, baseados em Regras de Associação no Modelo

Para este processo de conceção e implementação de um Sistema de Recomendação baseado em Regras de Associação, criámos um metanode (**Recommendation System**) onde foram implementados dois *workflows*, um baseado na *feature* 'Product Name' e outro baseado na *feature* 'Category Name'. Cada um destes *workflows* implementados e as suas particularidades, encontram-se devidamente explicados nos tópicos seguintes.

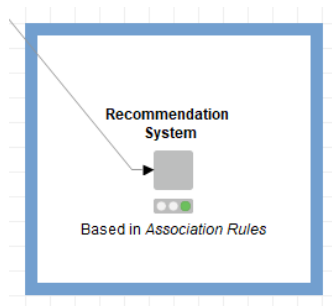


Figura nº37 – Componente criado para o Sistema de Recomendação Baseado em Regras de Associação

1. Regras de Associação Implementadas baseadas na *feature* 'Product Name'

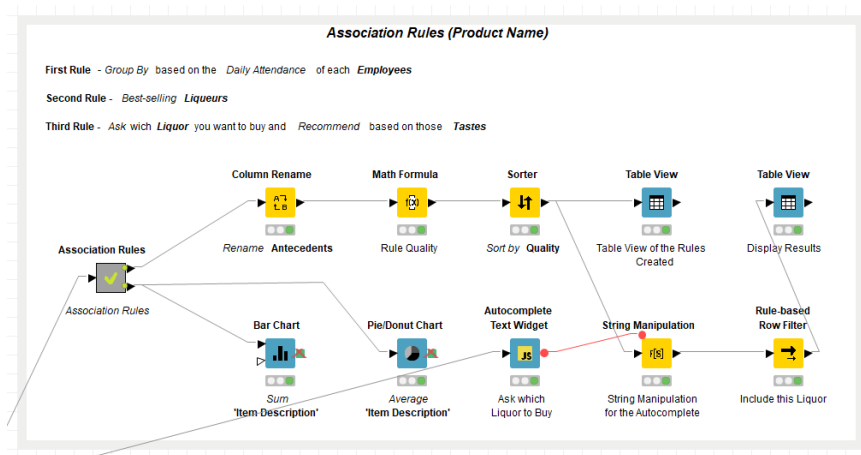


Figura nº38 – Workflow implementado para a Definição de Regras de Associação

De acordo com a imagem apresentada, podemos observar que implementamos no *workflow*, as referidas Regras de Associação. O objetivo da 1ª Regra ('*Group By* based on the *Daily Attendance* of each *Employees*') era conseguir agrupar dados, baseado no atendimento diário de cada funcionário. A 2ª Regra ('*Best-selling Liqueurs*') permite-nos saber quais os licores e bebidas alcoólicas mais vendidas. Já a nossa 3ª Regra ('*Ask wich *Liquor* you want to buy and *Recommend* based on those *Tastes**') foi desenvolvida com o propósito de saber qual o produto que o utilizador pretende adquirir, e com base nesse produto, fazer recomendações de acordo com os seus gostos.

Ao nível de implementações realizadas no *KNIME*, visando a implementação das regras de associação enunciadas, criámos um metanode (***Association Rules***), onde trabalhamos a parte a parte do vendedor de cada dia recorrendo ao nodo ***Group By***, agrupando as colunas '*Vendor_Number*', '*Month (number)*' e '*Day of month*', fazendo uma agregação manual de uma listagem dos itens da feature '*Item Description*', para que depois aplicando o ***Association Rule Learner (Borgelt)*** seja possível aprender as regras definidas.

Na parte da definição e tratamento da 2ª regra, começamos por fazer um count dos diversos licores, agrupando a coluna '*Item Description*', fazendo uma agregação manual com uma contagem dos itens da feature '*Item_Number*'. Com o objetivo de saber qual o mais vendido, recorrendo ao ***Sorter***, fizemos uma ordenação pela contagem do atributo '*Item_Number*'. Para finalizar, fizemos uma filtragem do Top 10.

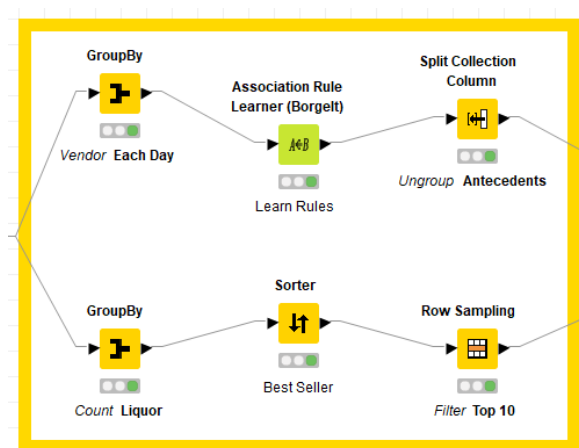


Figura nº39 – Nodos implementados no Metanode *Association Rules*

| Input data with newly appended columns - 0:127:43 - Split Collection Column (Ungroup Antecedents) | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|-----------------------|---|------------------|---|-----------|------|------------|------|-----------|-------|------------|---|-------------------------|---|----------|---|-----------|---|-------------------------|---|--------------------------|--|--|--|
| File Edit Hilite Navigation View | | | | | | | | | | | | | | | | | | | | | | | | | |
| Table "default" - Rows: 16769 Spec - Columns: 11 Properties Flow Variables | | | | | | | | | | | | | | | | | | | | | | | | | |
| Row ID | S | Consequent | S | Split Val... | 1 | ItemSe... | D | Relativ... | D | RuleCo... | D | Absolut... | D | RelativeBodySetSupport% | D | RuleLift | D | RuleLift% | D | AbsoluteHeadItemSupport | D | RelativeHeadItemSupport% | | | |
| Row0 | | Blue Ox Silver Rum | | Blue Ox Vodka | 2 | 0.673 | 50 | 4 | 1.35 | | 74.25 | 7,425 | 2 | 0.673 | | | | | | | | | | | |
| Row1 | | Four Roses Small... | | Four Roses ... | 2 | 0.673 | 66.7 | 3 | 1.01 | | 99 | 9,900 | 2 | 0.673 | | | | | | | | | | | |
| Row2 | | SOOH Bartender... | | Bartenders ... | 2 | 0.673 | 66.7 | 3 | 1.01 | | 99 | 9,900 | 2 | 0.673 | | | | | | | | | | | |
| Row3 | | Sweet Revenge | | Bartenders ... | 2 | 0.673 | 66.7 | 3 | 1.01 | | 99 | 9,900 | 2 | 0.673 | | | | | | | | | | | |
| Row4 | | Jewish Cream Liq... | | Cody Road ... | 2 | 0.673 | 66.7 | 3 | 1.01 | | 99 | 9,900 | 2 | 0.673 | | | | | | | | | | | |
| Row5 | | Stolchnaya Razberi | | Stolchnaya ... | 2 | 0.673 | 50 | 4 | 1.35 | | 74.25 | 7,425 | 2 | 0.673 | | | | | | | | | | | |
| Row6 | | Margaritaville Gol... | | Ryan's Crea... | 2 | 0.673 | 50 | 4 | 1.35 | | 74.25 | 7,425 | 2 | 0.673 | | | | | | | | | | | |
| Row7 | | Patron Reposado | | Patron Silver | 2 | 0.673 | 50 | 4 | 1.35 | | 74.25 | 7,425 | 2 | 0.673 | | | | | | | | | | | |
| Row8 | | Cody Road Rye | | Cody Road ... | 2 | 0.673 | 66.7 | 3 | 1.01 | | 49.5 | 4,950 | 4 | 1.347 | | | | | | | | | | | |
| Row9 | | Cody Road Bour... | | Cody Road ... | 2 | 0.673 | 50 | 4 | 1.35 | | 49.5 | 4,950 | 3 | 1.01 | | | | | | | | | | | |
| Row10 | | Patron Anejo | | Patron Silver | 2 | 0.673 | 50 | 4 | 1.35 | | 74.25 | 7,425 | 2 | 0.673 | | | | | | | | | | | |
| Row11 | | Patron Silver | | Patron Silver... | 3 | 1.01 | 100 | 3 | 1.01 | | 74.25 | 7,425 | 4 | 1.347 | | | | | | | | | | | |
| Row12 | | Patron Silver Mini | | Patron Silver | 3 | 1.01 | 75 | 4 | 1.35 | | 74.25 | 7,425 | 3 | 1.01 | | | | | | | | | | | |
| Row13 | | Margaritaville Sil... | | Ryan's Crea... | 2 | 0.673 | 50 | 4 | 1.35 | | 74.25 | 7,425 | 2 | 0.673 | | | | | | | | | | | |

Figura nº40 – Output Obtido após Definição da 1ª Regra (Split Collection Column)

The sampled table - 0:127:22 - Row Sampling...

File Edit Hilite Navigation View

Table "default" - Rows: 10 Spec - Columns: 2 Properties Flow Variables

| Row ID | S | Item Description | I | Count*(Item_Number) |
|---------|---|---------------------------------|---|---------------------|
| Row141 | | Black Velvet | | 512 |
| Row315 | | Crown Royal | | 345 |
| Row324 | | Crown Royal Regal Apple | | 282 |
| Row540 | | Hawkeye Vodka | | 244 |
| Row964 | | Seagrams 7 Crown | | 234 |
| Row577 | | Jack Daniels Old #7 Black Label | | 219 |
| Row603 | | Jim Beam | | 193 |
| Row472 | | Five Star | | 181 |
| Row1096 | | Titos Handmade Vodka | | 166 |
| Row470 | | Five O'Clock Vodka | | 150 |

Figura nº41 – Output Obtido após Definição da 2ª Regra (Row Sampling)

Após a saída dos dados proveniente do componente implementado, uma das saídas está ligada ao nodo **Column Rename**, com o objetivo renomear os antecedentes. De modo a calcularmos a qualidade da relação, recorreremos à fórmula ' $\$ItemSetSupport\$ * \$RuleConfidence\%$ ', de acordo com o evidenciado na imagem seguinte que diz respeito ao nodo **Math Formula**. De seguida é feita uma ordenação descendente de registos, mediante a feature 'Rule Quality', implementada anteriormente.

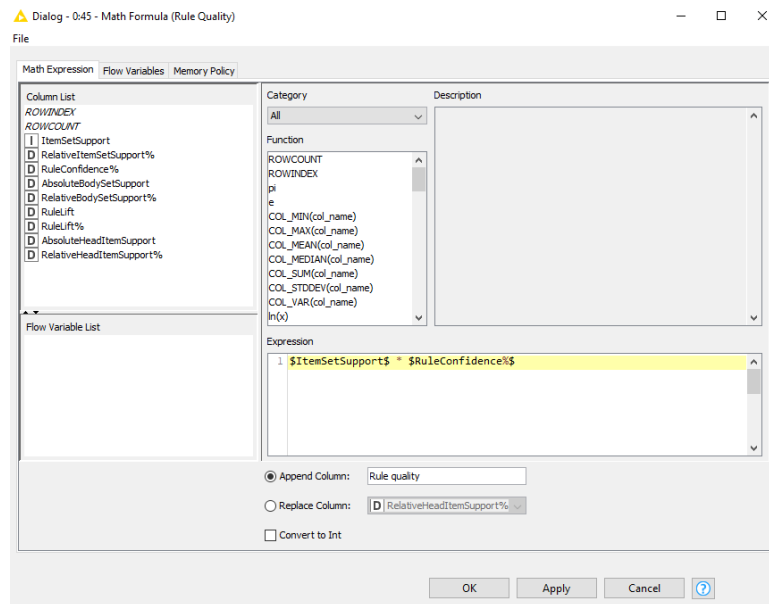


Figura nº42 – Configurações aplicadas no nodo *Math Formula*

O **Table View** apresentado e implementado depois do nodo **Sorter**, apenas serviu como como teste, de modo a conseguir visualizar se mediante as regras implementadas no nosso sistema, o mesmo recomendava outros licores mediante o licor comprado. De salientar que para a visualização de dados deste **Table View**, não são necessário quaisquer *Input* de dados por parte do utilizador, uma vez que são mostrados todos os dados de produtos relacionados.

JavaScript Table View

Recommendation

Similar to Purchased Liquor

Show: 10 entries Search:

| <input type="checkbox"/> | RowID | Recommended | Who buys | Rule quality |
|--------------------------|---------|------------------------|------------------------|--------------|
| <input type="checkbox"/> | Row300 | Five O'Clock Gin | Five Star | 600 |
| <input type="checkbox"/> | Row301 | Five Star | Five O'Clock Gin | 600 |
| <input type="checkbox"/> | Row302 | Five O'Clock Vodka PET | Five Star | 600 |
| <input type="checkbox"/> | Row303 | Five Star | Five O'Clock Vodka PET | 600 |
| <input type="checkbox"/> | Row316 | Five O'Clock Vodka PET | Five O'Clock Gin | 600 |
| <input type="checkbox"/> | Row317 | Five O'Clock Gin | Five O'Clock Vodka PET | 600 |
| <input type="checkbox"/> | Row3235 | Black Velvet Traveler | Black Velvet Reserve | 600 |
| <input type="checkbox"/> | Row3236 | Black Velvet Reserve | Black Velvet Traveler | 600 |
| <input type="checkbox"/> | Row3237 | Black Velvet | Black Velvet Reserve | 600 |
| <input type="checkbox"/> | Row3238 | Black Velvet Reserve | Black Velvet | 600 |

Showing 1 to 10 of 16,769 entries

Previous 1 2 3 4 5 ... 1677 Next

Reset Apply Close

Figura nº43 – *Table View* do *Output* das Regras de Associação

Tendo em conta que o principal objetivo do sistema de recomendação é mediante um *Input* do utilizador, que neste caso concreto será o nome do licor/bebida comprado (*'Product Name'*), serem feitas recomendações de outros licores mediante as regras de associação implementadas, foi necessário implementar no *workflow* o nodo **Autocomplete Text Widget** que irá receber os pretendidos *Input's*. Ao nível de configurações relevantes deste nodo, selecionamos a *feature* do *dataset* que nos dava o nome dos licores (*'Item Description'*) e definimos com valor *default* o licor *'Five Star'*.

Levando em consideração que uma determinada pessoa pode não saber o nome completo, ou até mesmo não saber escrever o nome de um determinado licor, implementamos o nodo **String Manipulation**, com o objetivo de fazer um *join* com o que o utilizador escreve, e com o que é guardado na variável do *Autocomplete* (*'join(""*, \$\$\${Autocomplete string input}\$\$, "*)'*) Por fim, com a implementação do nodo **Roled-based Row Filter**, implementamos a regra que nos permite fazer a associação entre o produto comprado e o produto a recomendar. Isto só é possível, mediante uma operação *LIKE* entre as variáveis *'who buys'*, que armazena o nome do licor comprado, e *'iWant'*, que armazena o nome dos licores a recomendar (*'\$Who buys\$ LIKE \$iWant\$ => TRUE'*).

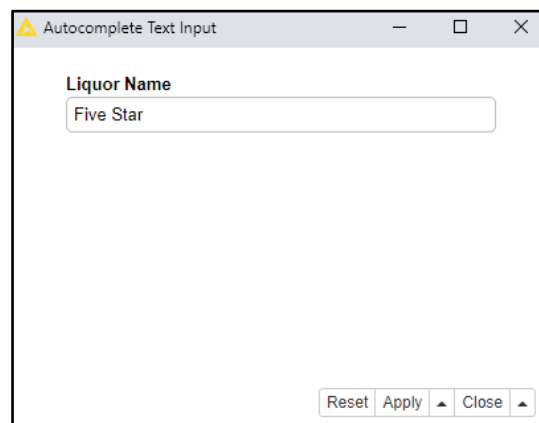


Figura nº44 – *Input* para o Nome do Licor

JavaScript Table View

Recommendation

Similar to Purchased Liquor

Show entries

Search:

| <input type="checkbox"/> | RowID | Recommended | Who buys |
|--------------------------|--------|------------------------|---------------|
| <input type="checkbox"/> | Row300 | Five O'Clock Gin | Five Star |
| <input type="checkbox"/> | Row302 | Five O'Clock Vodka PET | Five Star |
| <input type="checkbox"/> | Row227 | Five Star | Five Star PET |
| <input type="checkbox"/> | Row229 | Five O'Clock Gin | Five Star PET |
| <input type="checkbox"/> | Row231 | Five O'Clock Vodka PET | Five Star PET |
| <input type="checkbox"/> | Row220 | Five O'Clock Vodka | Five Star |
| <input type="checkbox"/> | Row228 | Five Star PET | Five Star |
| <input type="checkbox"/> | Row218 | Five O'Clock Vodka | Five Star PET |

Showing 1 to 8 of 8 entries

Previous 1 Next

Reset Apply Close

Figura nº45 – *Output* gerado pelo Sistema de Recomendação mediante o Nome do Licor

Ainda sobre os dois restantes nodos implementados neste *workflow* (**Bar Chart** e **Pie/Donut Chart**), os menos têm como objetivo a visualização de dados, segundo um *COUNT* de licores, uma ordenação pelo mais vendido e um filtro de *TOP 10*, implementados nos nodos que se encontram na parte inferior do *workflow* do metanode **Association Rules**.

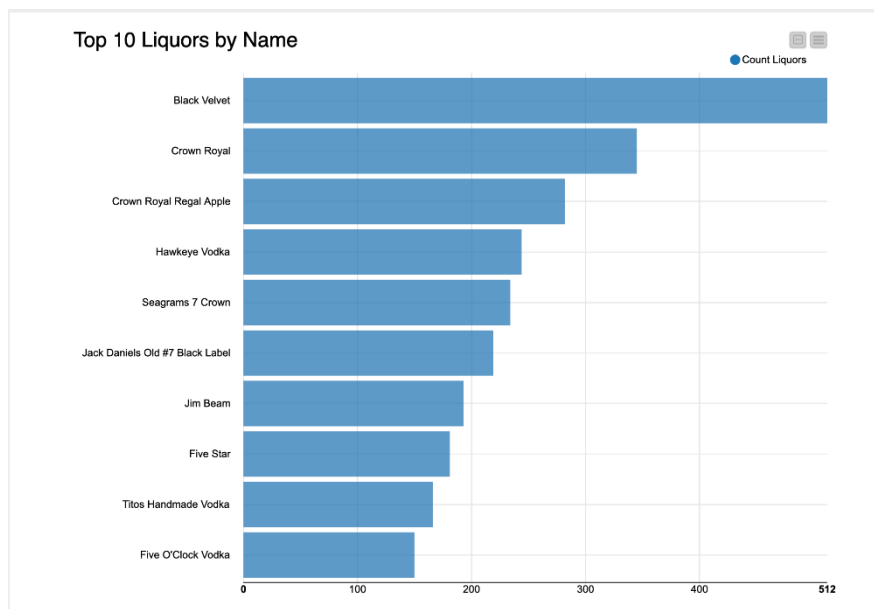


Figura nº46 – *Output* do Bar Chart (TOP 10 de Licores)

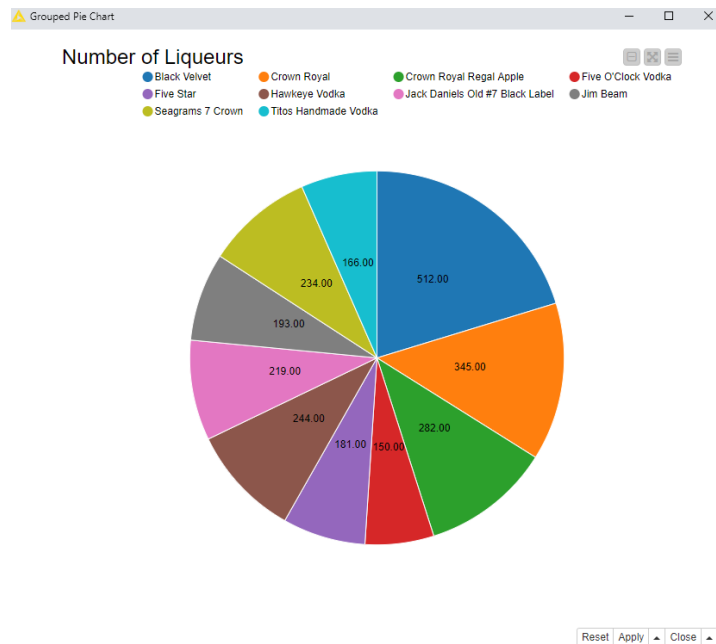


Figura nº47 – Output do Pie/Donut Chart (TOP 10 de Licores)

Foi adicionada nesta secção para análise do sistema de recomendação baseada no *feedback* obtido pelos utilizadores, isto é, após a visualização dos dados é perguntado ao utilizador se concorda com as recomendações, mediante isso seria possível obter estatísticas como:

1. Quais os produtos mais recomendados?
2. Nível de precisão da recomendação?
3. Número de clientes com recomendações.

Mediante estas questões e análises o sistema de recomendação poderia ser refinado gradualmente.

2. Regras de Associação Implementadas baseadas na *feature* 'Category Name'

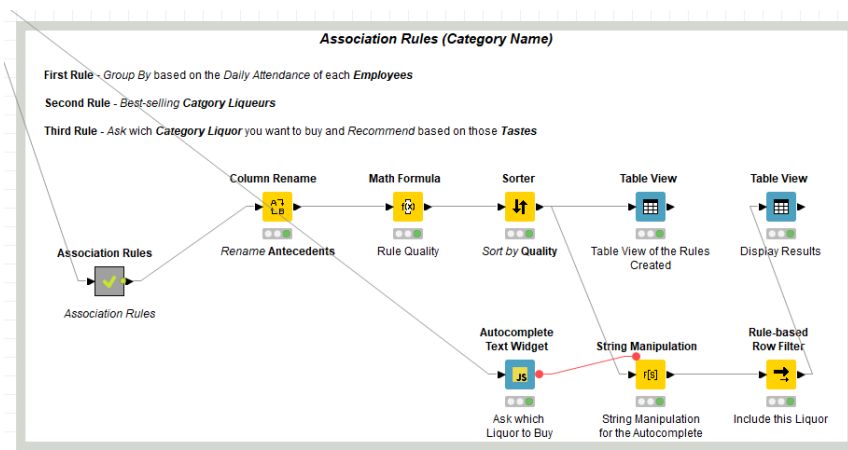


Figura nº48 – Workflow implementado para a Definição de Regras de Associação

À semelhança do implementado nas regras de associação segundo o nome do produto, foi feita uma abordagem semelhante, tendo as principais diferenças residido na implementação de nodos no metanode **Association Rules**, e na configuração do nodo **Autocomplete Text Widget**. O objetivo da implementação das regras de associação baseadas na *feature* 'Category Name' continuou a ser em primeiro lugar a agrupação dados, baseado no atendimento diário de cada funcionário, depois conseguir saber quais as categorias de licores e bebidas alcoólicas mais vendidas, de modo numa fase final saber qual o produto que o utilizador pretende adquirir, e com base nesse produto, fazer recomendações de acordo com os seus gostos.

Ao nível da configuração do metanode **Association Rules**, e com o objetivo de não tornar a informação demasiado repetitiva, não implementamos os nodos que nos permitiram anteriormente obter o TOP 10, ficando assim apenas com os nodos que nos permitem fazer e implementar o nosso sistema de recomendação baseado em associações.

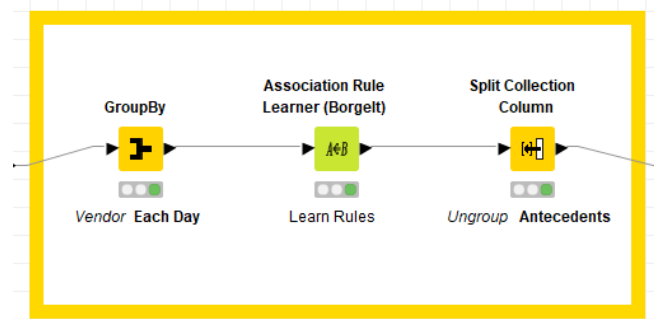


Figura nº49 – Nodos implementados no Metanode Association Rules

No nodo **Column Rename**, mais uma vez são renomeados os antecedentes, e com o objetivo de calcularmos a qualidade da relação, recorremos à fórmula anterior ' $\$ItemSetSupport\$ * \$RuleConfidence\%$ ', implementada no nodo **Math Formula**. Posteriormente é feita uma ordenação descendente de registos, mediante a *feature* '*Rule Quality*'. O **Table View** implementado, serviu mais uma vez como teste, de modo a conseguir visualizar se mediante as regras implementadas no nosso sistema, o mesmo recomendava outros licores mediante a categoria do licor comprado.

Sobre a configuração do nodo **Autocomplete Text Widget**, esta sim, com alguma diferença da implementada no sistema de recomendação anterior baseada no '*Product Name*', irá receber os pretendidos *Input's* relacionados com a categoria do produto. Para isso selecionamos a *feature* do *dataset* que nos dava o nome dos licores ('*Category Name*') e definimos com valor *default* a categoria de licores '*American Vodkas*'.

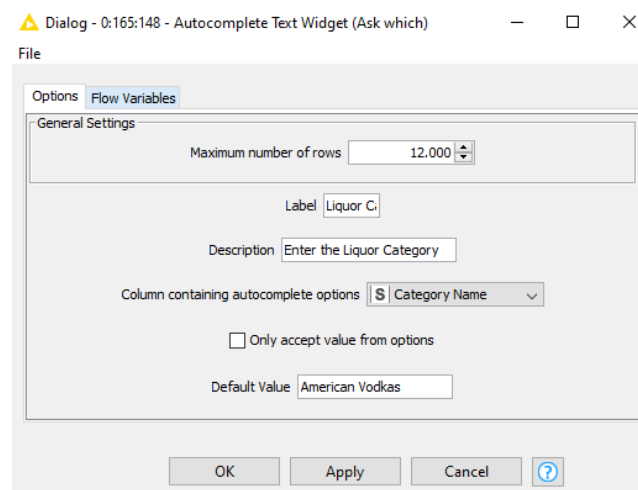


Figura nº50 – Configurações aplicadas no nodo **Autocomplete Text Widget**

O nodo **String Manipulation**, tem igualmente como objetivo fazer um *join* com a categoria de licores que o utilizador escreve, com o que é guardado na variável do *Autocomplete*. Já no nodo **Roled-based Row Filter**, implementamos a regra que nos permite fazer a associação entre a categoria do produto comprado e a categoria do produto a recomendar.

JavaScript Table View

Recommendation

Similar to Purchased Category Liquor

Show 10 entries Search:

| <input type="checkbox"/> | RowID | Recommended | Who buys |
|--------------------------|--------|-------------------|-----------------|
| <input type="checkbox"/> | Row331 | American Dry Gins | American Vodkas |

Showing 1 to 1 of 1 entries

Previous 1 Next

Reset Apply Close

Figura nº51 – *Output* gerado pelo Sistema de Recomendação mediante a Categoria do Licor

b. Aplicar de nodos, baseados em *Clusters* no Modelo

Para este processo de conceção e implementação de um Sistema de Recomendação baseado em *Clusters*, criámos um metanode (**Recommendation System**) onde foram implementados cinco sistemas baseados em clusters, para as *features* ‘Category Name’, ‘Product Name’, ‘Price’, ‘Month (name)’ e ‘City’. Cada um destes *workflows* implementados e as suas particularidades, encontram-se devidamente explicados nos tópicos seguintes.

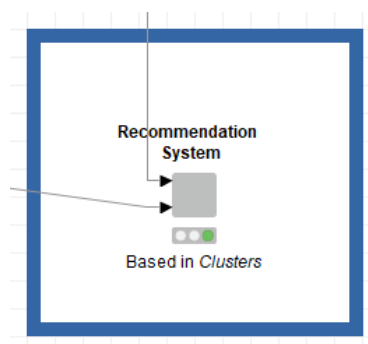


Figura nº52 – Componente criado para o Sistema de Recomendação Baseado em *Clusters*

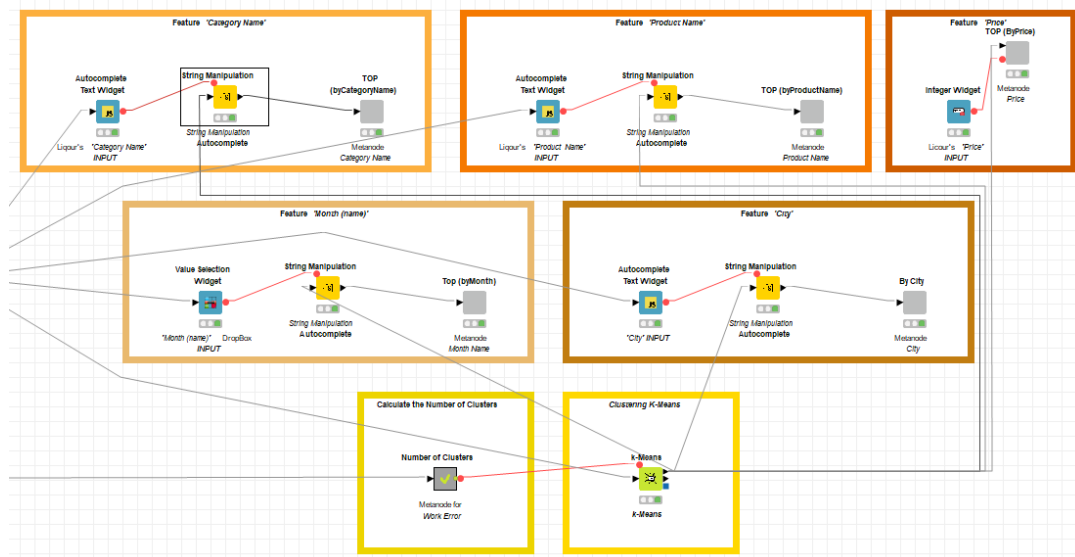


Figura nº53 – Workflow Implementado para o Sistema de Recomendação Baseado em Clusters

Sobre o *workflow* acima evidenciado, como já foi referido na **Tarefa 1 e** (Aplicar nodos, de modo a fazer a Otimização do Modelo), a solução ótima de otimização encontrada foi recorrendo ao *Clustering k-Means* com o cálculo do erro segundo a fórmula matemática do *MSE*. A forma de como foi feita a implementação deste tipo de otimização encontra-se devidamente explicada no tópico referenciado, pelo que de modo a não ter informação repetida, não é necessário voltar a explicar abordagem feita, e quais foram os motivos pelos quais optamos por este tipo de *clustering*.

Sobre o metanode **Number of Clusters**, que foi implementando com o objetivo de determinar o número ideal de clusters, foi calculada a distância entre clusters em primeiro lugar. Para isso recorreremos ao nodo **Lag Column**, onde definimos que a *Column to lag* seria a do *MSE*, com um *Lag* e *Lag interval* de 1. De seguida, foi calculada a diferença entre a **distância clusters** e o **valor obtido no erro (MSE)**, recorrendo à função implementada em *Java* presente na **Figura nº54**. Recorrendo ao **Sorter**, foi feita uma ordenação de forma descendente do valor da diferença obtido no nodo anterior. Mediante o output obtido no **Sorter**, foi perceptível que *Row0#1* é aquele que apresenta uma maior diferença entre a distância e o erro, pelo que foi o selecionado no **Row Filter**, onde o número ideal de *clusters* é 2. No nodo **Table Row to Variable**, foi armazenado numa variável o número ideal de *clusters* a passar para o nodo *k-Means*.

Através do output visualizado na **Scatter Plot**, foi possível confirmar que um número ideal de *clusters* era de facto 2, dado que a maior diferença entre o número de *clusters* é entre 1 e 2.

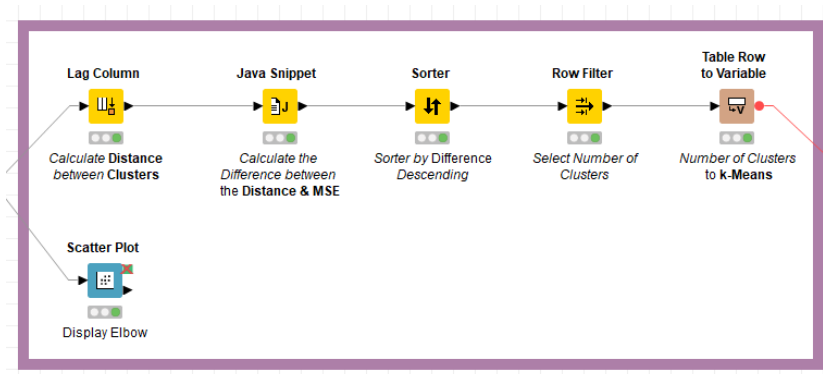


Figura nº54 – Nodos implementados no Metanode *Number of Clusters*

Output - 0:166:168:54 - Lag Column (Calcul...

File Edit Hilite Navigation View

Table "default" - Rows: 10 Spec - Columns: 4 Properties Flow Variables

| Row ID | D MSE | I Cluster | I Iteration | D MSE(-1) |
|--------|-------|-----------|-------------|-----------|
| Row0#0 | 0.048 | 1 | 0 | ? |
| Row0#1 | 0.015 | 2 | 1 | 0.048 |
| Row0#2 | 0.007 | 3 | 2 | 0.015 |
| Row0#3 | 0.006 | 4 | 3 | 0.007 |
| Row0#4 | 0.003 | 5 | 4 | 0.006 |
| Row0#5 | 0.003 | 6 | 5 | 0.003 |
| Row0#6 | 0.002 | 7 | 6 | 0.003 |
| Row0#7 | 0.002 | 8 | 7 | 0.002 |
| Row0#8 | 0.001 | 9 | 8 | 0.002 |
| Row0#9 | 0.001 | 10 | 9 | 0.001 |

Figura nº55 – *Output* obtido com o Cálculo da Distância em *Clusters*

```

1 // system imports
13 // Your custom imports:
14
15 // system variables
27 // Your custom variables:
28
29 // expression start
31 // Enter your code here:
32 try{
33     double diff = c_MeanMSE1 - c_MeanMSE;
34
35     if(diff < 0.0){
36         diff = 0.0-diff;
37         out_DiffMSE = diff;
38     }
39     else{
40         out_DiffMSE = diff;
41     }
42 } catch(Exception e){
43     out_DiffMSE = 0.0;
44 }
45 // expression end
46
49

```

Figura nº56 – Função em Java, para Cálculo da Diferença entre a Distância e o Erro

Appended table - 0:166:168:53 - Java Snippet (Calculate the)

File Edit Hilite Navigation View

Table "default" - Rows: 10 Spec - Columns: 5 Properties Flow Variables

| Row ID | D MSE | I Cluster | I Iteration | D MSE(-1) | D Diff MSE |
|--------|-------|-----------|-------------|-----------|------------|
| Row0#0 | 0.048 | 1 | 0 | ? | 0 |
| Row0#1 | 0.015 | 2 | 1 | 0.048 | 0.033 |
| Row0#2 | 0.007 | 3 | 2 | 0.015 | 0.008 |
| Row0#3 | 0.006 | 4 | 3 | 0.007 | 0.001 |
| Row0#4 | 0.003 | 5 | 4 | 0.006 | 0.003 |
| Row0#5 | 0.003 | 6 | 5 | 0.003 | 0 |
| Row0#6 | 0.002 | 7 | 6 | 0.003 | 0.001 |
| Row0#7 | 0.002 | 8 | 7 | 0.002 | 0 |
| Row0#8 | 0.001 | 9 | 8 | 0.002 | 0 |
| Row0#9 | 0.001 | 10 | 9 | 0.001 | 0 |

Figura nº57 – Output obtido com o Cálculo da Diferença entre a Distância e o Erro

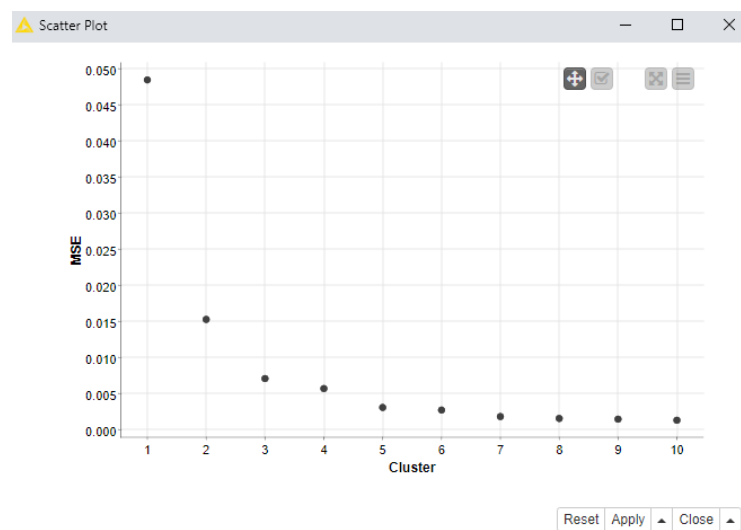


Figura nº58 – Output Scatter Plot

1. Clusters baseados na feature 'Category Name'

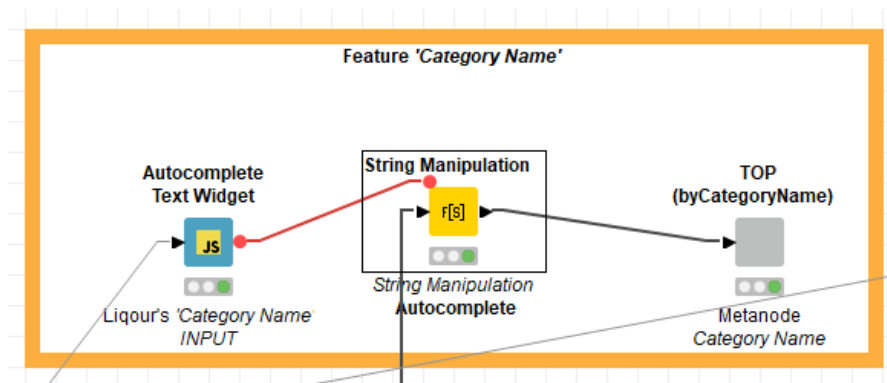


Figura nº59 – Nodos implementados para o Sistema de Recomendação baseado na Categoria do Licor

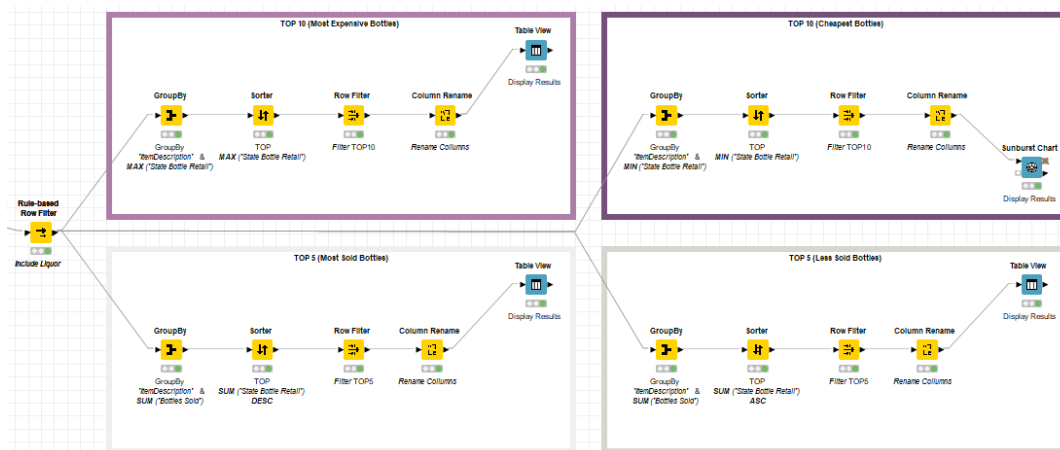


Figura nº60 – Workflow Implementado

De acordo com as **Figuras nº59 e 60**, podemos observar que recorrendo ao nodo **Autocomplete Text Widget**, recebemos o **INPUT** do utilizador sobre a Categoria de Licores a pesquisar. O valor *default* introduzido é o ‘Gold Rum’. No nodo **String Manipulation**, fazemos um *join* com a categoria de licores que o utilizador escreve, com o que é guardado na variável do **Autocomplete**. Já no nodo **Roled-based Row Filter**, implementamos a regra que nos permite fazer a associação entre a categoria do licor e o licor a recomendar. No caso concreto deste caso, fazemos recomendações mediante o TOP 10 de Garrafas mais e menos caras, e o TOP 5 de Garrafas mais e menos vendidas.

2. Clusters baseados na feature ‘Product Name’

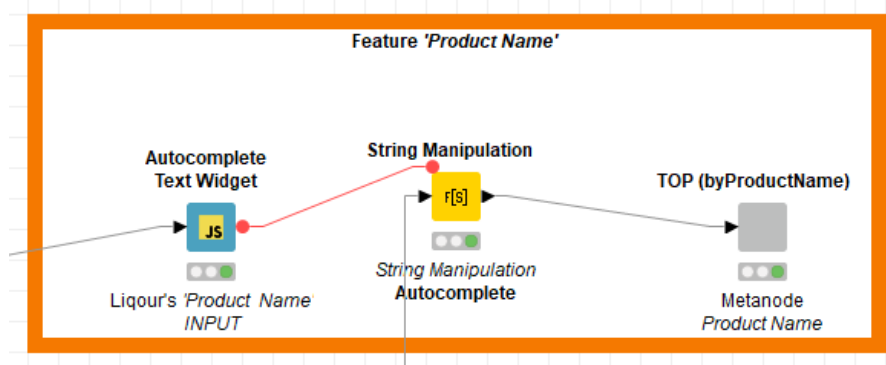


Figura nº61 – Nodos implementados para o Sistema de Recomendação baseado no Nome do Licor

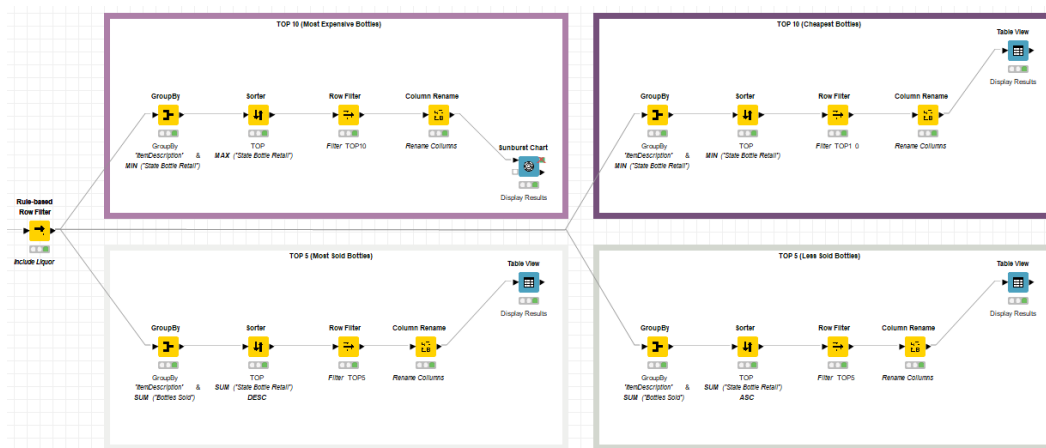


Figura nº62 – Workflow Implementado

De acordo com as **Figuras nº61 e 62**, e à semelhança das configurações efetuadas anteriormente, podemos observar que recorrendo ao nodo **Autocomplete Text Widget**, recebemos o **INPUT** do utilizador sobre o Nome de Licores a pesquisar. O valor *default* introduzido é o *'Five Star'*. O nodo **String Manipulation** e o **Roled-based Row Filter**, têm o mesmo objetivo dos implementados anteriormente. No caso concreto deste caso, fazemos recomendações mediante o TOP 10 de Garrafas mais e menos caras, e o TOP 5 de Garrafas mais e menos vendidas.

3. Clusters baseados na feature 'Price'

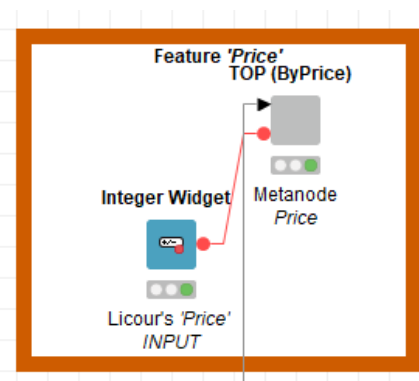


Figura nº63 – Nós implementados para o Sistema de Recomendação baseado no Preço

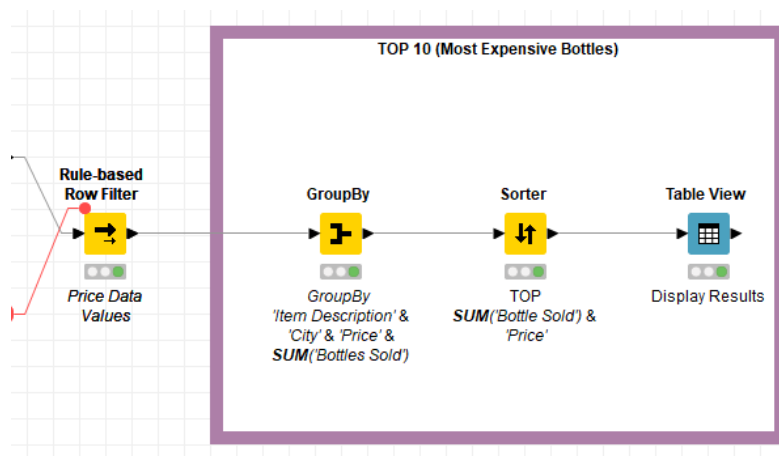


Figura nº64 – Workflow Implementado

De acordo com as **Figuras nº63 e 64**, e à semelhança das configurações efetuadas anteriormente, podemos observar que recorrendo ao nodo **Integer Widget**, onde recebemos o **INPUT** do utilizador sobre o Preço de Licores a pesquisar. O valor *default* introduzido é o '10 \$'. O **Roled-based Row Filter**, tem o mesmo objetivo dos implementados anteriormente. No caso concreto deste caso, fazemos recomendações mediante o TOP Preços de acordo com o valor introduzido.

4. Clusters baseados na feature 'Month (name)'

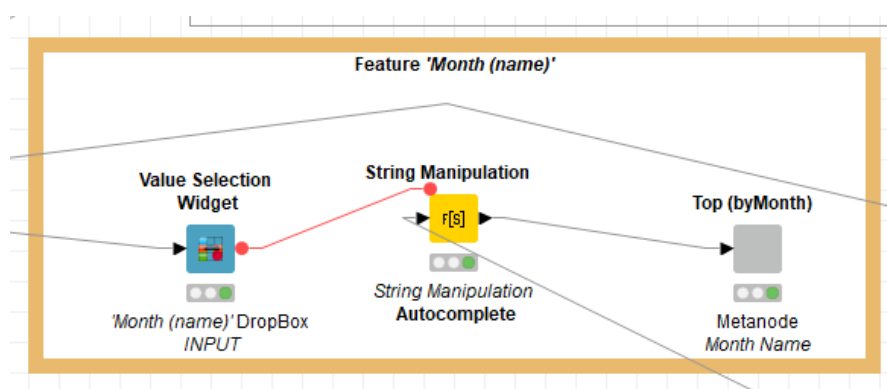


Figura nº65 – Nodos implementados para o Sistema de Recomendação baseado no Mês de Venda

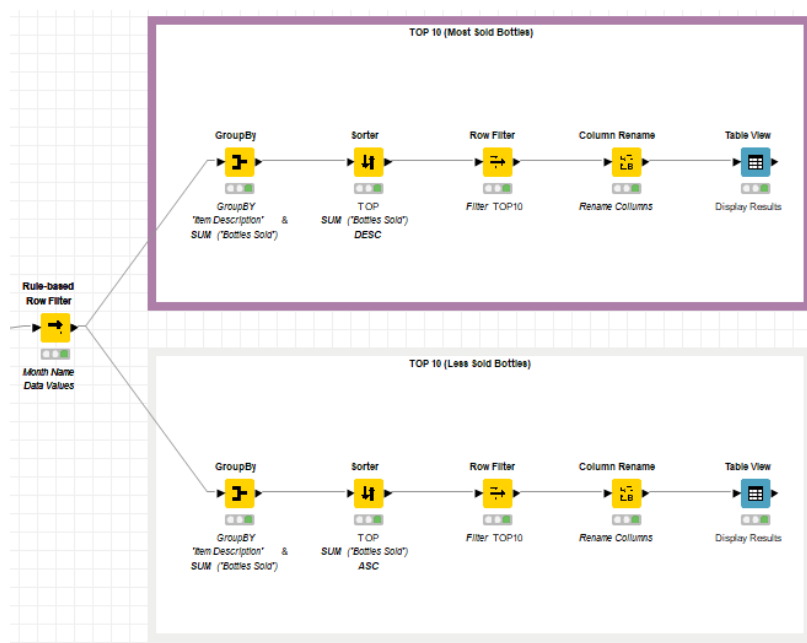


Figura nº66 – Workflow Implementado

De acordo com as Figuras nº65 e 66, recorrendo ao nodo **Value Selection Widget**, onde recebemos o *INPUT* do utilizador sobre o mês da venda de Licores a pesquisar. O valor *default* introduzido é o 'Janeiro'. O **Roled-based Row Filter**, tem o mesmo objetivo dos implementados anteriormente. No caso concreto deste caso, fazemos recomendações mediante o TOP 10 de Garrafas mais e menos vendidas.

5. Clusters baseados na feature 'City'

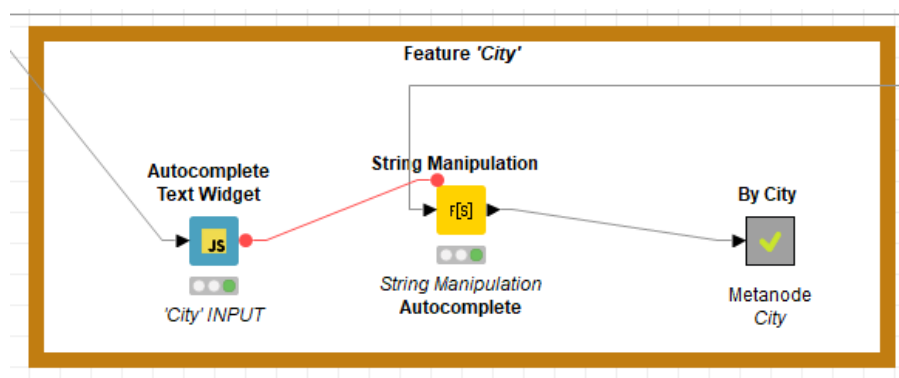


Figura nº67 – Nodos implementados para o Sistema de Recomendação baseado na Cidade de Venda

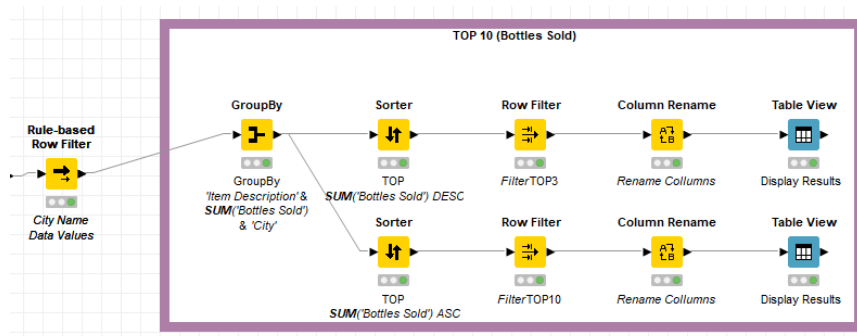


Figura nº68 – Workflow Implementado

De acordo com as **Figuras nº67 e 68**, podemos observar que recorrendo ao nodo **Autocomplete Text Widget**, recebemos o **INPUT** do utilizador sobre a Cidade de Venda de Licores a pesquisar. O valor *default* introduzido é o *'Dubuque'*. No nodo **String Manipulation**, fazemos um *join* com o que é guardado na variável do *Autocomplete*. Já no nodo **Rule-based Row Filter**, implementamos a regra que nos permite fazer a associação. No caso concreto deste caso, fazemos recomendações mediante o TOP 10 de Garrafas mais e menos vendidas.

c. Avaliação do Sistema de Recomendação

a. Regras de Associação

De seguida demonstramos a avaliação do Sistema de Recomendação baseado em Regras Associativas:

| Categoria Pesquisada | Categorias Recomendadas | Conclusão |
|-------------------------------|--|--|
| <i>Aged Dark Rum</i> | <i>Flavored Rum</i> <i>White Rum</i> <i>Spiced Rum</i> | Podemos observar que quem consome “ <i>Aged Dark Rum</i> ” o sistema recomenda mais 3 tipos de Rums. |
| <i>Whiskey Liqueur</i> | <i>Canadian Whiskies</i> <i>Straight Bourbon Whiskies</i> | Podemos observar que quem consome “ <i>Whiskey Liqueur</i> ” o sistema recomenda mais 2 tipos de whiskies. |

Tabela nº3 – Avaliação do Sistema de Recomendação (Categorias de Licores)

| Nome Pesquisado | Licores Recomendados | Conclusão |
|--------------------------------------|--|---|
| <i>Black Velvet Reserve</i> | <i>Black Velvet Traveler</i> <i>Black Velvet</i> <i>Black Velvet Toasted Caramel</i> <i>Black Velvet Mini</i> | Podemos observar que quem consome “ <i>Black Velvet Reverse</i> ” o sistema recomenda mais 4 tipos da mesma gama. |
| <i>Five O'Clock Vodka PET</i> | <i>Five Star</i> <i>Five O'Clock Gin</i> | Podemos observar que quem consome “ <i>Five O'Clock Vodka PET</i> ” o sistema recomenda mais 2 tipos da mesma gama. |

Tabela nº4 – Avaliação do Sistema de Recomendação (Nomes de Licores)

b. Baseado em *Clusters*

| Nome Pesquisado | Licores Recomendados | Conclusão |
|------------------------------|---|--|
| <i>Cream Liqueurs</i> | <i>Kirkland Signature Irish cream</i> <i>McCormick's Irish cream</i> <i>Baileys Original Irish cream</i> <i>Ryan's cream Liqueur</i> | Podemos observar que quem consome licores cremosos o sistema recomenda mais 4 tipos da mesma gama. |
| <i>White Rum</i> | <i>Paramount White Rum</i> <i>Barton Rum Light</i> <i>Paramount White Rum PET</i> <i>Bacardi Superior (RUM)</i> | Podemos observar que quem consome Rum branco o sistema recomenda mais 4 tipos de Rum. |

Tabela nº5 – Avaliação do Sistema de Recomendação (Nomes de Licores)

Claro está que futuramente deveria ser implementado um limite, de modo a que a regra de qualidade seja confiável. Para concluir, de modo geral, as recomendações são bastante coerentes e parecem assertivas.

Conclusão

A elaboração deste Projeto/Trabalho Prático, cujo objetivo principal recaía sobre a construção de um Sistema de Recomendação, aplicando e explorando todo aquilo que nos foi passado ao longo do semestre, com o propósito de ter um Sistema de Recomendação funcional e fiável.

Durante o desenvolvimento do sistema, foi crucial ir fazendo uma análise dos resultados que íamos obtendo, uma vez que no processo de desenvolvimento de modelos de *Machine Learning* é importante, não só saber analisar os dados e obter conclusões sobre estes, mas também, encontrar formas de moldar os dados de modo a obtermos os resultados esperados. Após a “limpeza” dos dados e do *dataset* devidamente tratado, iniciamos a implementação do Sistema de Recomendação. Este processo foi bastante demorado e custoso, tem sido a maior adversidade encontrada na realização deste trabalho, uma vez que residia a dúvida sobre qual ou quais as melhores recomendações baseadas nos *inputs* definidos pelos utilizadores. Porém com todo o empenho e dedicação tido pelos elementos do grupo, esta dificuldade foi superada com sucesso.

Qualquer projeto prático contém possíveis melhorias a adicionar e este não é exceção, a principal melhoria será a avaliação do k ótimo para a clusterização através do cálculo dos centroides, que infelizmente devida a escassez de tempo não conseguimos implementar. Para concluir, com a realização deste trabalho prático e com o fim deste relatório, consideramos ter atingido todos os objetivos propostos e descrito todo o desenvolvimento da forma mais intuitiva possível.