Análise Exploratória de Dados – Cadeia de Perfumarias

2020 / 2021

DESCO – Descoberta do Conhecimento

Vasco Rodrigues 1171419, Hugo Correia 1170569, Tiago Andrade 1170677

Índice

Índice de figuras

1. Introdução

O presente trabalho foi realizado no âmbito da disciplina de Descoberta do Conhecimento, integrando o Mestrado de Sistemas de Informação e Conhecimento do Instituto Superior de Engenharia do Porto.

O projeto deve ser acompanhado de um relatório com a descrição o mais detalhada possível do processo que seguiram para obter as vossas soluções e deve incluir: a exploração gráfica dos dados e respetivas interpretações que considerarem mais relevantes, limpeza e pré-processamento efetuada aos dados, explicação dos objetivos de negócio pretendidos com os modelos criados, interpretação/avaliação dos modelos e compromissos assumidos no seu desenvolvimento.

O projeto deve seguir a metodologia CRISP-DM e como tal devem desenvolver scripts R adequados para cada uma das suas fases:

- Preparação e exploração de dados;
- Pré-processamento dos dados;
- Criação de modelos utilizando alguns dos algoritmos de mineração estudados;
- Avaliação dos modelos criados.

2. Objetivos

Os dados são de uma rede de perfumarias com mais de 30 anos de existência. Esta empresa comercializa produtos de perfumaria, cosmética, cuidados do corpo e desde cedo apostou na fidelização de clientes através de um cartão cliente.

O objetivo do projeto é identificar segmentos de clientes de modo a dirigir campanhas adequadas a cada um; desenvolver modelos que permitam caracterizar cada um dos segmentos e assim conhecer melhor o perfil dos clientes; saber quais os produtos mais adequados a cada segmento de clientes de modo a estimular o interesse dos clientes pelos produtos e consequentemente aumentar as vendas da empresa.

O score RFM (Recência, Frequência e Monetário) permite segmentar os clientes de forma objetiva. No entanto, este score ao considerar todos os produtos comprados, inclui também aqueles raramente comprados e não tem em consideração por exemplo, a variação do preço dos produtos, a frequência com que o cliente compra produtos mais caros ou mais baratos. Portanto, o score RFM fornece apenas índices de avaliação de soma total, e pode não ser preciso a quantificar a lealdade e a contribuição do cliente para os resultados da empresa.

Associações entre os produtos comprados pelos clientes podem ser usadas para prever padrões de consumo dos clientes no futuro, pelo que questões igualmente importantes para a empresa são:

- Quais os produtos que são comprados conjuntamente?
- Qual o perfil dos potenciais compradores de um determinado produto ou conjunto de produtos?
- Qual é o intervalo de consumo, a frequência e o valor gasto por um grupo de clientes num conjunto de produtos específico?

3. Preparação e pré-processamento dos dados

Após a leitura dos dados dos clientes e das compras, foi feita uma análise inicial para verificar existência de dados em falta nos datasets, ao qual não foi verificado tal facto. Posteriormente, foi feita uma análise da dimensão dos dados para verificar se foram importados corretamente e a verificação dos tipos de dados e informação adicional sobre os dados como média, mediana.

Finalmente, depois destas verificações iniciais, foi procedido à preparação e pré-processamento dos dados.

Para os dados dos clientes, foram feitas as seguintes transformações:

- Data de registo removendo a parte do tempo visto que seria sempre 0
- Verificação se existe clientes com filhos, mas o número de filhos era 0 e sua correção
- Colocação da tag "O"(other) nos clientes que não indicaram estado civil
- Verificação e tratamento/normalização das cidades do dataset dos clientes

Para os dados das compras, foram feitas as seguintes transformações:

- Verificação de valores vazios para os atributos: Suplier, Productype e brand e substituição por “No suplier”, “Productype” e “brand”, respetivamente
- Remoção de *whitespaces* e formatação de valores da coluna “Tax”
- Verificação e correção do valor Total de uma compra através da multiplicação da quantidade de produtos (Qnt) com o preço de venda do produto (Tableunitprice).
- Verificação e correção do valor ano da coluna ”Year” através do ano que consta da tabela “Date”
- Verificação e correção das linhas cuja quantidade do produto é inferior a 1

Após estes tratamentos, foram contabilizados o número de compras de cada cliente e a remoção de compras cujo cliente apenas tenha feito menos de 12 compras, considerando que o dataset é referente a 4 anos e comprado uma média de 3 perfumes por ano. Após este pré-processamento dos dados das compras, foi eliminado todos os clientes do dataset de clientes sem compras no dataset de compras. Também foi removida a variação de preços dos produtos. Estas alterações influenciam como o RFM funciona visto que este considera “todos os produtos comprados, inclui também aqueles raramente comprados e não tem em

consideração por exemplo, a variação do preço dos produtos, a frequência com que o cliente compra produtos mais caros ou mais baratos…".

Analisando os dados que restaram, podemos verificar que as colunas PACKAGE e PACKAGEFACT não variavam os seus valores, logo não teriam importância no âmbito do trabalho, o mesmo se passou com DESCOUNTVAL1 e o DESCOUNTPERCENT1 que apenas variavam o seu valor em 14 entradas de 285802 entradas ao qual consideramos pouco significativo.

No fim foi feito o merge dos dados para termos no mesmo dataset toda a informação não compartimentada.

4. Exploração dos dados

Neste capítulo é feita a exploração gráfica dos dados, com diferentes gráficos elaborados nesse âmbito para o projeto.

4.1 Boxplot do número de filhos

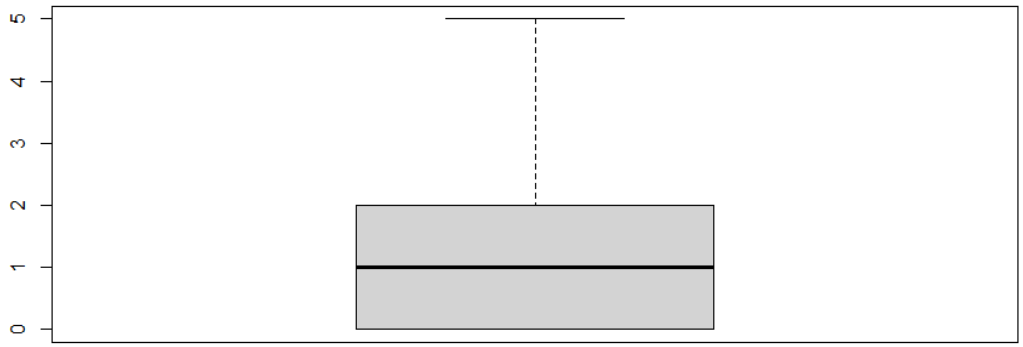

Figura 1 - Boxplot referente ao número de filhos dos clientes

Com este gráfico, Figura 1, podemos perceber que a mediana do número de filhos é 1 e que o número máximo é 5.

4.2 Percentagem de número de filhos

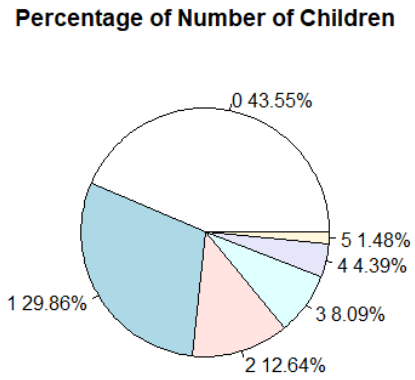

Figura 2 – Gráfico circular referente à percentagem do número de filhos dos clientes

Como podemos ver na Figura 2, a maior parte dos clientes não tem filhos, seguido de pessoas com apenas 1 filho. Isto pode significar que a compra de produtos é feita para si.

4.3 Número de crianças por zona

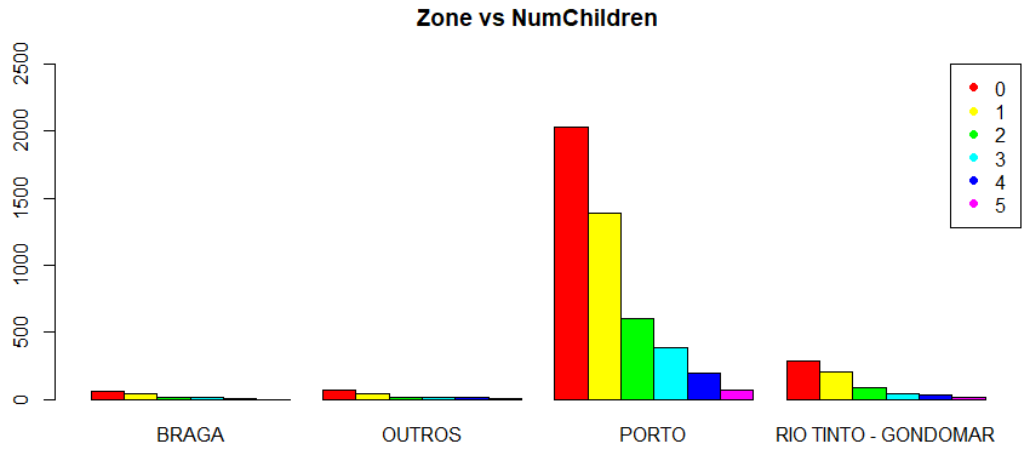

Figura 3 – Gráfico de número de crianças por zona

Com este gráfico, Figura 3, podemos perceber que a zona do Porto tem mais clientes e que maioritariamente, este não tem filhos, o mesmo acontece nas outras zonas onde 0 filhos é sempre a maioria.

4.4 Idade por género

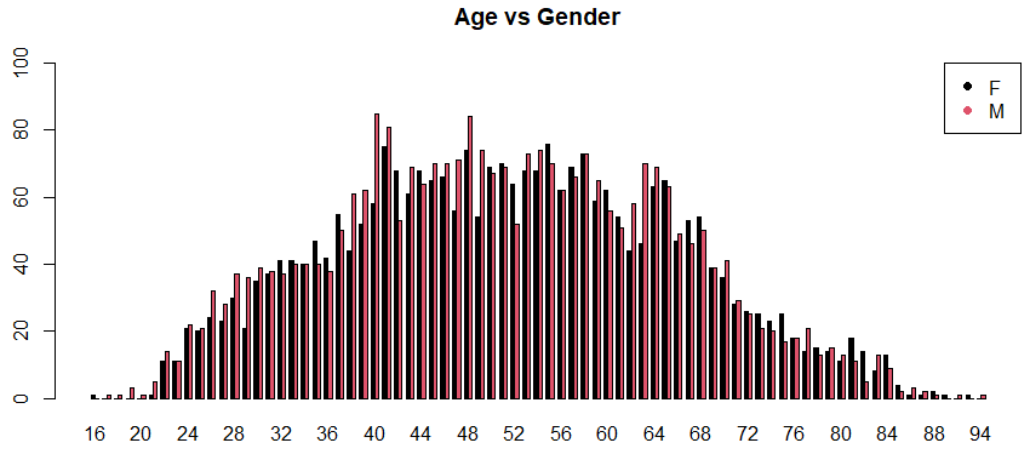

Figura 4 – Gráfico de relação entre idade e género dos clientes

Como este gráfico, Figura 4, podemos perceber melhor a demografia dos clientes, onde estes são de meia-idade e estão distribuídos quase igualmente em termos de género.

4.5 Estado civil por género

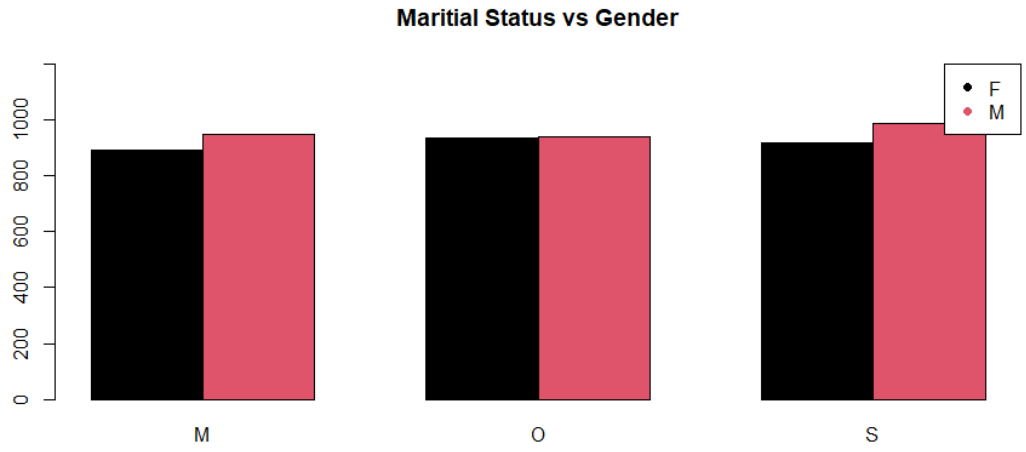

Figura 5 – Gráfico de relação entre estado civil e género

4.6 Se tem crianças por género

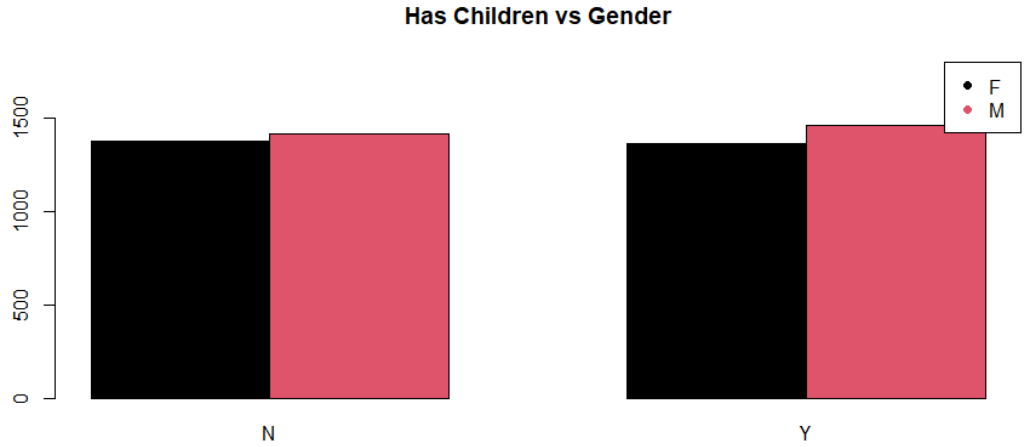

Figura 6 – Gráfico de relação entre ter crianças e género

4.7 Se tem crianças por estado civil

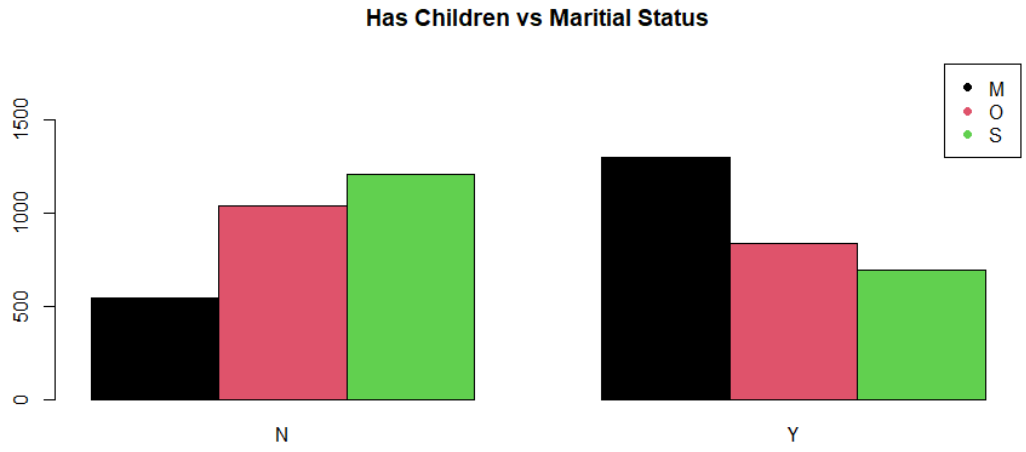

Figura 7 – Gráfico de relação entre ter crianças e estado civil

Continuando a estudar a demografia dos clientes, nas figuras Figura 5, Figura 6 e Figura 7, podemos perceber que estão distribuídos de forma quase igual em termos de ter crianças por género e estado civil por género, notando-se uma diferença significativa se tiver crianças com base no estado civil, onde é comprovado que normalmente quando uma pessoa está solteira não tem filhos que quando está casada, tem filhos.

4.8 Número de clientes por zona

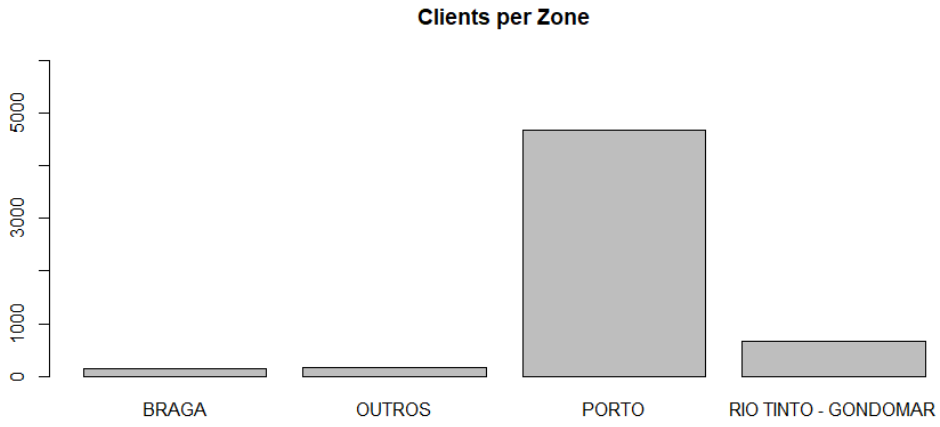

Figura 8 – Gráfico de relação entre número de clientes por zona

Como podemos ver na Figura 8, a maioria dos clientes desta perfumaria é da zona do Porto, seguida da zona do Rio Tinto – Gondomar.

4.9 Média de preço por marca

Análise do preço que, em média, cada marca associa aos seus produtos, podendo observar-se que existem marcas que, por norma, possuem produtos mais caros.

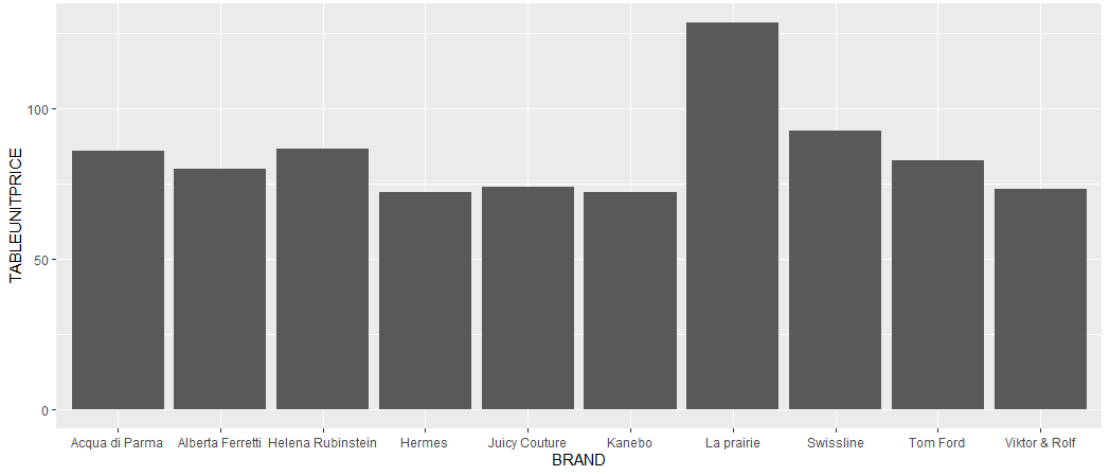

Figura 9 – Gráfico de Preço vs Marca

4.10 Número de compras efetuadas

Análise de compras efetuadas ao longo dos anos. Foram efetuados dois tipos de gráficos para esta análise:

- Um boxplot, que representa a mediana do número de compras, e os vários quartis associados, sendo que grande parte do número de compras encontra-se entre os 4000 e os 10000. É também realçado que valores abaixo de 4000 são pouco comuns e consideradas discrepâncias na norma.
- Um gráfico de barras que representa os valores de vendas de cada ano. É realçado o grande volume de compras efetuadas nos anos 2012 e 2013. Enquanto mais baixo, o ano de 2014 continua com valores minimamente aceitáveis, tendo em conta o diagrama anterior. No entanto, no ano de 2011, as compras estão muito abaixo do esperado.

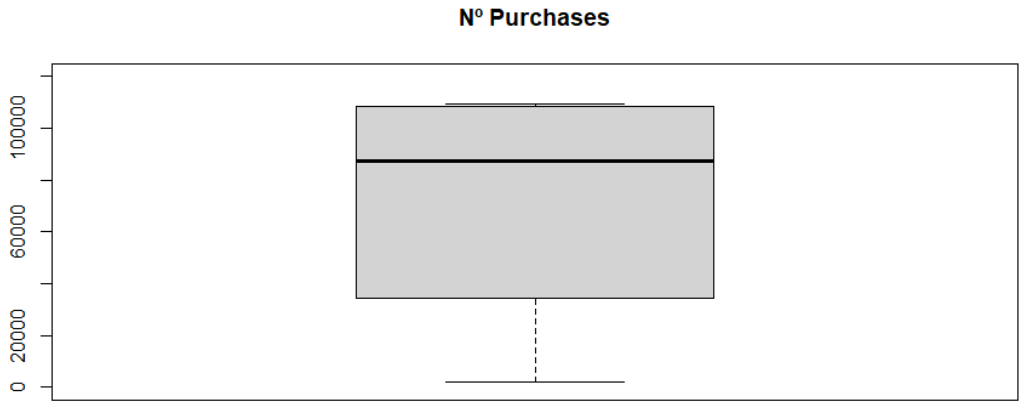

Figura 10 – Boxplot referente ao número de compras efetuadas nos vários anos

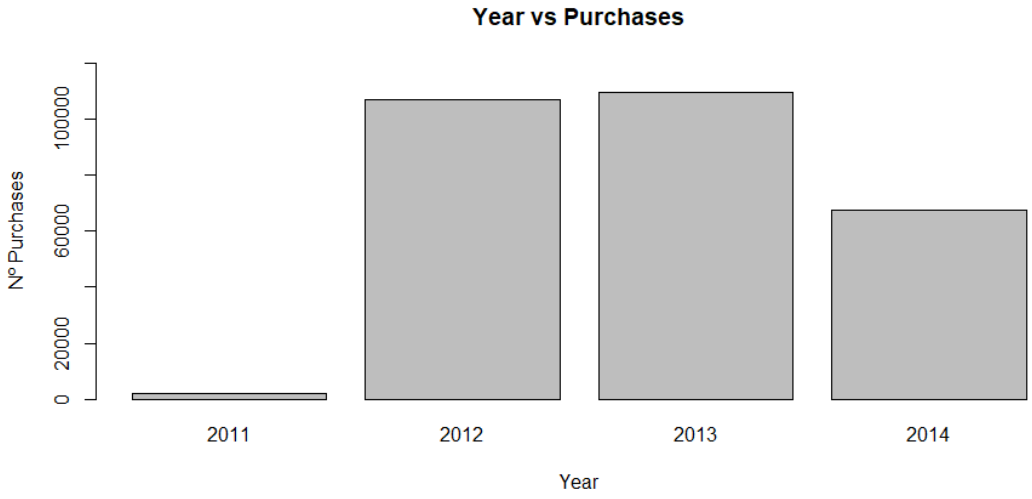

Figura 11 – Gráfico de Ano vs Compras

4.11 Melhores 10 clientes

Análise exploratória dos clientes com compras de maior valor, através dum gráfico horizontal de barras. Esta análise demonstra os clientes que, de forma geral, mais contribuem para os ganhos da perfumaria.

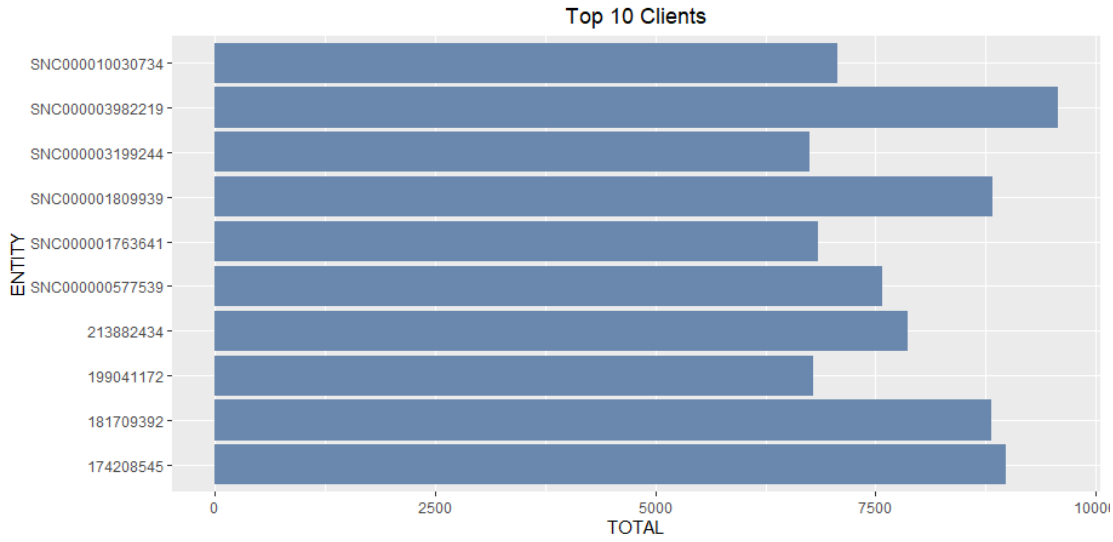

Figura 12 – Gráfico horizontal dos 10 melhores clientes

4.12 Compras efetuadas por marca

Análise do número de compras efetuadas por marca, em cada ano. Desta forma, torna-se possível destacar as marcas mais procuradas, assim como o seu crescimento/declínio ao longo dos anos. Foram efetuados dois gráficos distintos: Um *lineplot* com as 5 marcas mais procuradas, assim como um gráfico semelhante, com as 5 marcas divididas, de forma a facilitar a leitura dos dados.

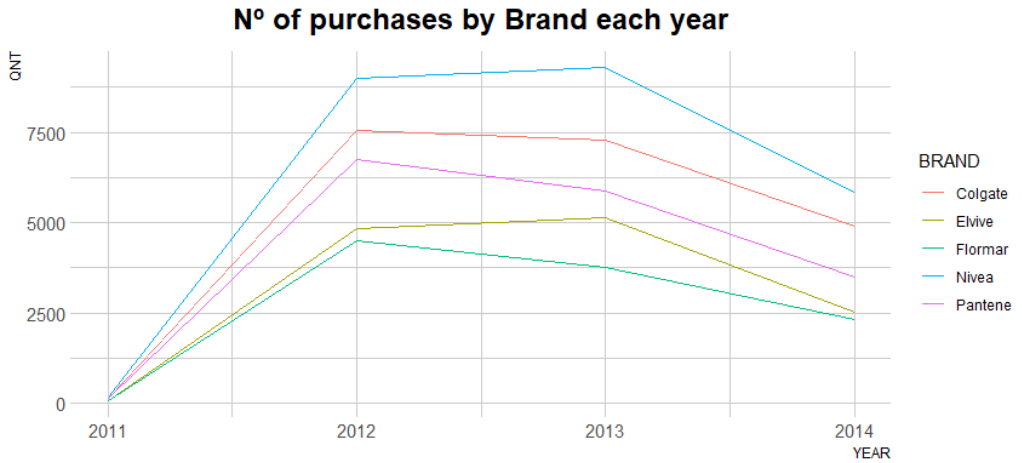

Figura 13 – Gráfico nº1 do número de compras subdividido por marca e por ano

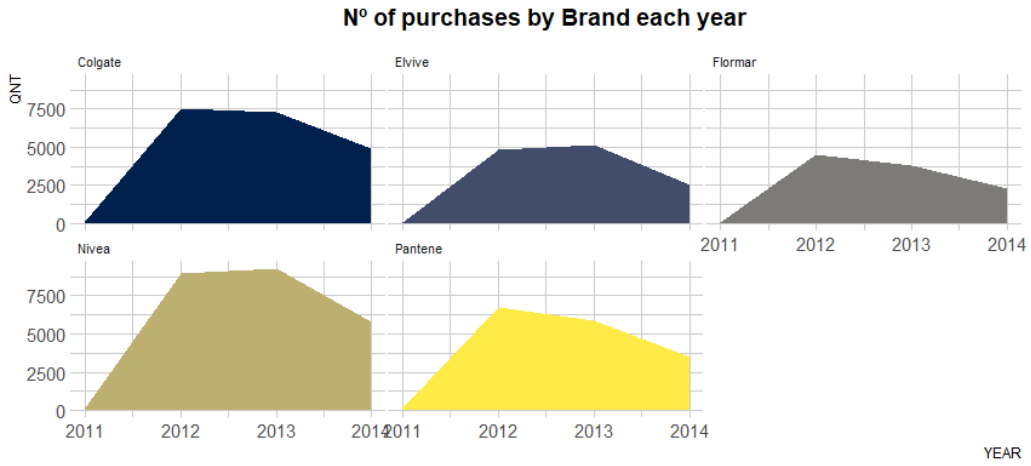

Figura 14 – Gráfico nº2 do número de compras subdividido por marca e por ano

4.13 Compras por tipo de produto

Nesta análise, são realçados, através de um gráfico de barras, a quantidade de produtos vendidos de cada tipo de produto diferente. O tipo de produto "Sem tipo" possui, com uma larga margem, o maior número de produtos vendidos, o que não possui um grande valor informativo. No entanto, esta análise continua a destacar alguns tipos de produtos que são mais predominantes nesta perfumaria, tais como: Champô, Desodorizantes e Pastas de dentes.

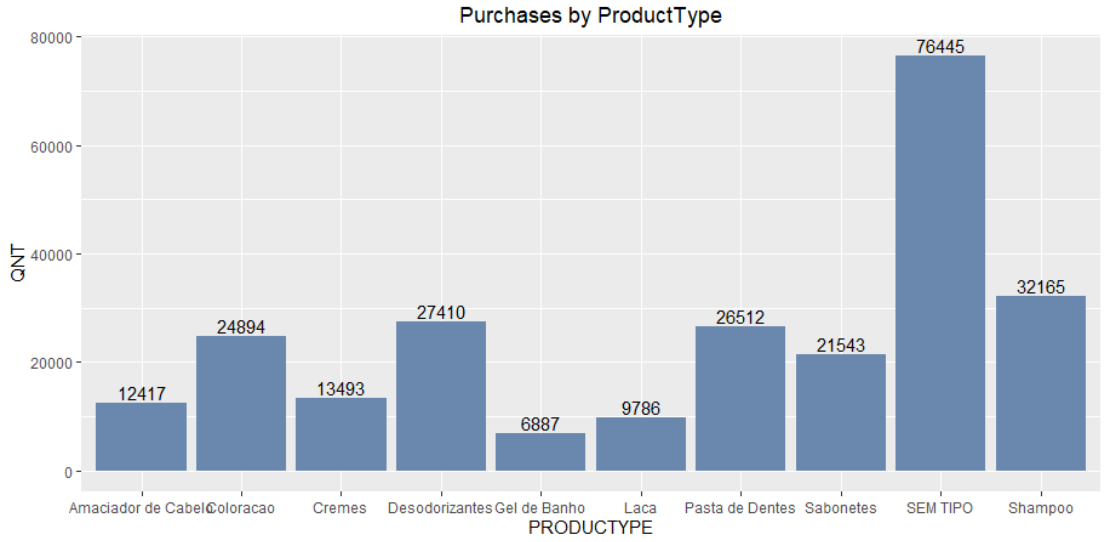

Figura 15 – Gráfico de compras de cada tipo de produto

4.14 Crescimento de vendas ao longo do tempo

Por último, foi efetuada uma análise da evolução das vendas da rede de perfumarias ao longo dos anos. Para tal, foi efetuado um gráfico interativo que demonstra, de forma acessível, o crescimento da empresa, enquanto disponibilizando de opções de *zoom* para avaliar os valores concretos em cada ponto do gráfico (Figura 17-18). Através deste gráfico observa-se que no início dos anos de 2013 e 2014 foram os picos de vendas desta empresa, realçando que a altura mais lucrativa será no início de cada ano. Observa-se também que a meio do ano de 2013 existe um aumento considerável de compras. Esta análise é útil para a empresa conseguir dar resposta a um aumento súbito de procura dos seus produtos.

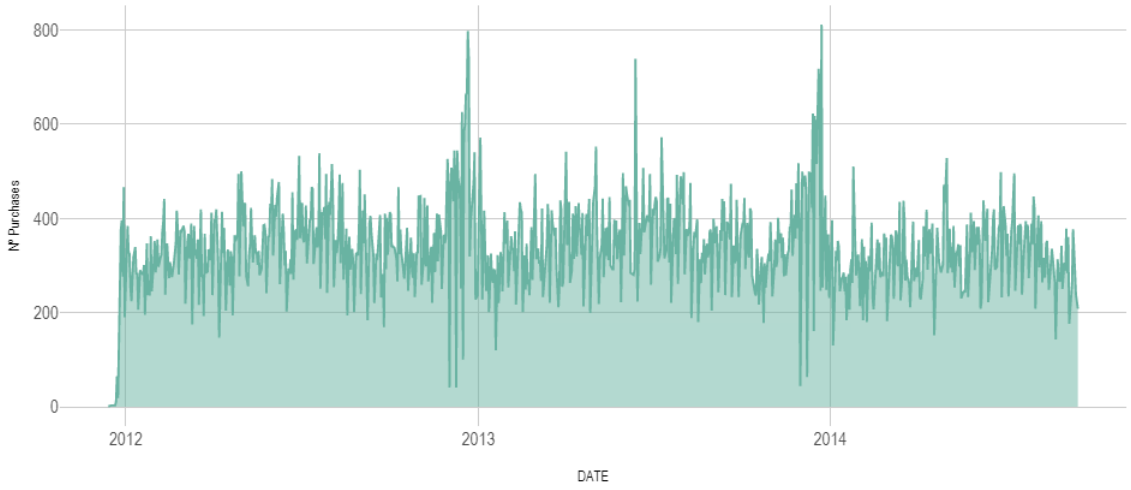

Figura 16 - Gráfico interativo do crescimento de vendas ao longo dos anos

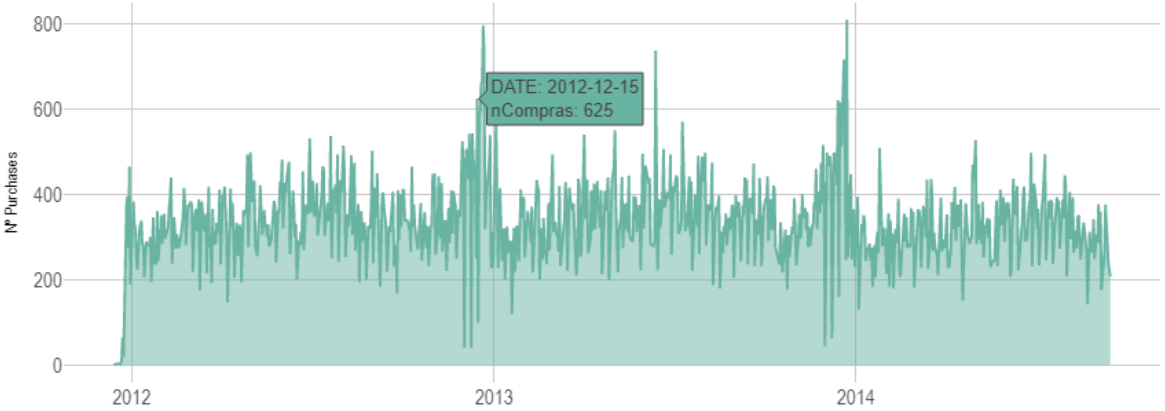

Figura 17 – Exemplo de visualização dos valores da figura 16

5. RFM

Visto que as datas das compras têm como intervalo "2011-12-14" e "2014-09-15", consideramos a data de análise do rfm de “2014-12-31”.

Considerando esta data de análise e os dados convergidos num dataset, realizamos o cálculo do rfm deste dataset. Também, para o desenvolvimento do modelo somente necessitamos dos seguintes dados ("CLIENT","DATE","TOTAL"). Com estes dados são assim efetuados os cálculos necessários para a obtenção das novas colunas “Recency”, “Frequency” e “Monetary”.

Como podemos ver na Figura 18, o cliente com o id “100132081” fez a sua última compra à 119 dias, totalizou 13 compras no valor total de 96.7. Assim, com estes valores, obteve o score rfm de 5-1-1, 5 para recency, 1 para frequency e 1 para monetary. Isto significa que o cliente realizou uma compra muito recentemente, mas o seu volume e valor de compras é bastante baixo.

	customer_id	date_most_recent	recency_days	transaction_count	amount	recency_score	frequency_score	monetary_score	rfm_score
1	100132081	2014-09-03	119	13	96.7000	5	1	1	511
2	100150624	2014-07-17	167	41	178.5400	3	3	2	332
3	100340814	2014-06-03	211	16	283.3900	2	1	3	213
4	100360289	2013-06-22	557	24	70.2100	1	2	1	121
5	100361110	2014-09-10	112	39	898.7900	5	3	5	535
6	100369529	2014-08-27	126	28	52.4000	4	3	1	431
7	100370616	2014-09-06	116	103	240.8600	5	5	2	552
8	100400710	2014-06-16	198	77	231.6300	2	5	2	252
9	100407161	2014-06-02	212	20	128.0100	2	2	1	221
10	100418708	2014-09-06	116	92	692.9500	5	5	4	554
11	100452663	2014-09-12	110	12	35.2800	5	1	1	511
12	100484123	2014-09-09	113	74	538.8500	5	5	4	554
13	100484484	2014-03-28	278	12	98.2300	2	1	1	211
14	100544916	2013-12-20	376	53	342.4500	1	4	3	143
15	100608582	2014-07-22	162	62	1700.7800	3	4	5	345
16	100621112	2014-09-03	119	32	157.7400	5	3	2	532
17	100621180	2014-08-22	131	139	729.6000	4	5	4	454
18	100785107	2014-08-20	133	83	741.5300	4	5	4	454
19	1008687252	2014-06-05	209	52	249.9500	2	4	2	242
20	101361050	2014-03-31	275	15	397.3500	2	1	3	213
21	101487622	2014-07-05	179	21	282.2900	2	2	3	223

Figura 18 – rfm data and score

Efetuamos uma análise global aos resultados do rfm, obtendo os seguintes gráficos.

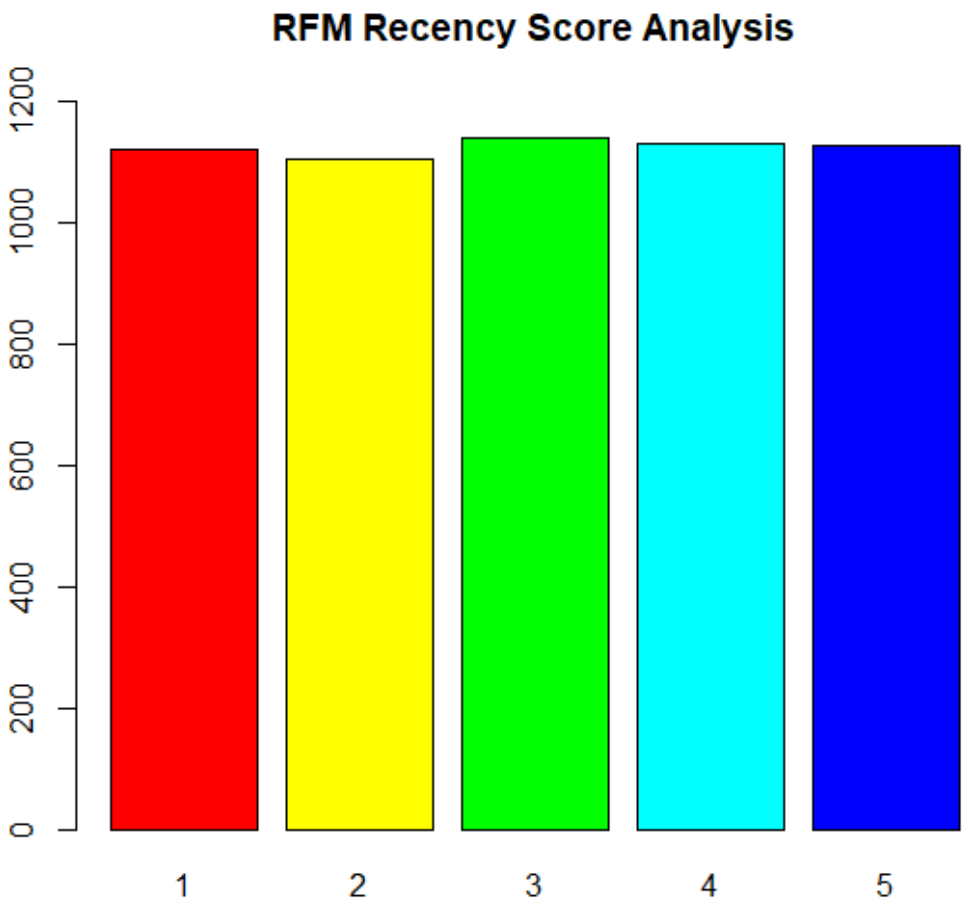

Figura 19 – Overall recency scores

Como podemos ver na Figura 19, em termos de recency, os clientes estão equilibrados, com maior número de cliente com recency score de 3.

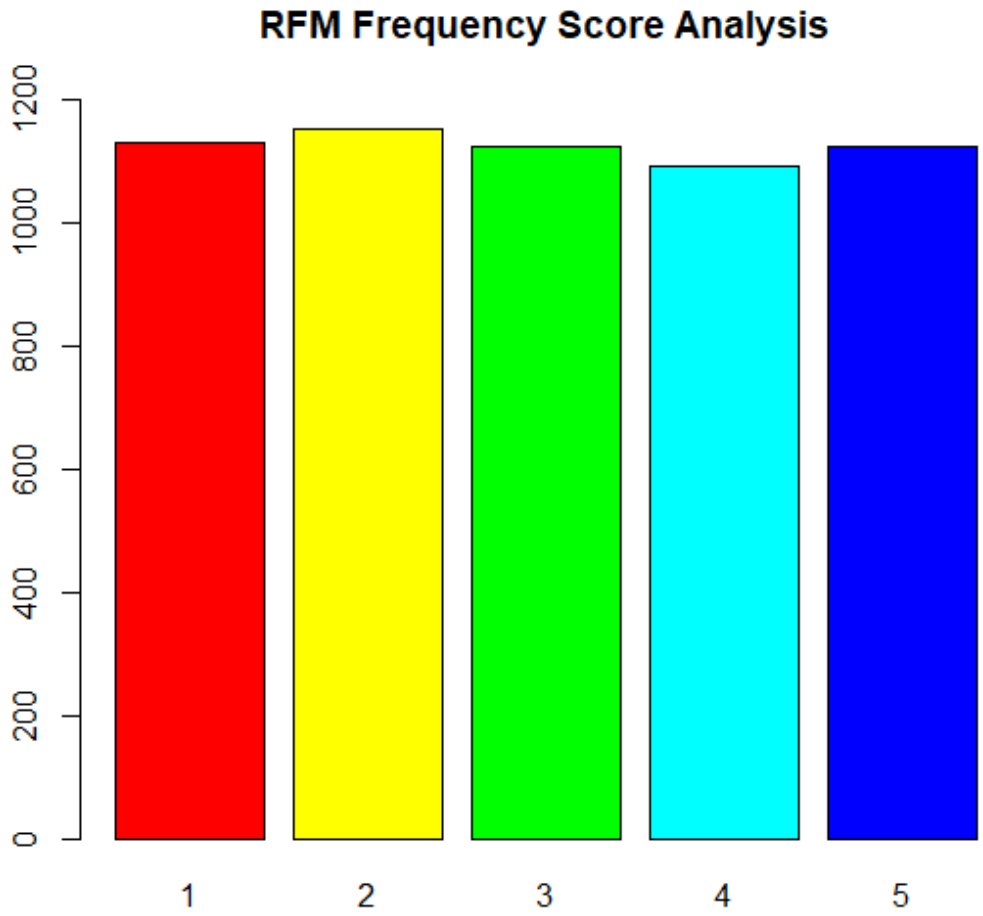

Figura 20 – Overall frequency scores

Como podemos ver na Figura 20, em termos de frequency, os clientes estão equilibrados, com maior número de cliente com frequency score de 2.

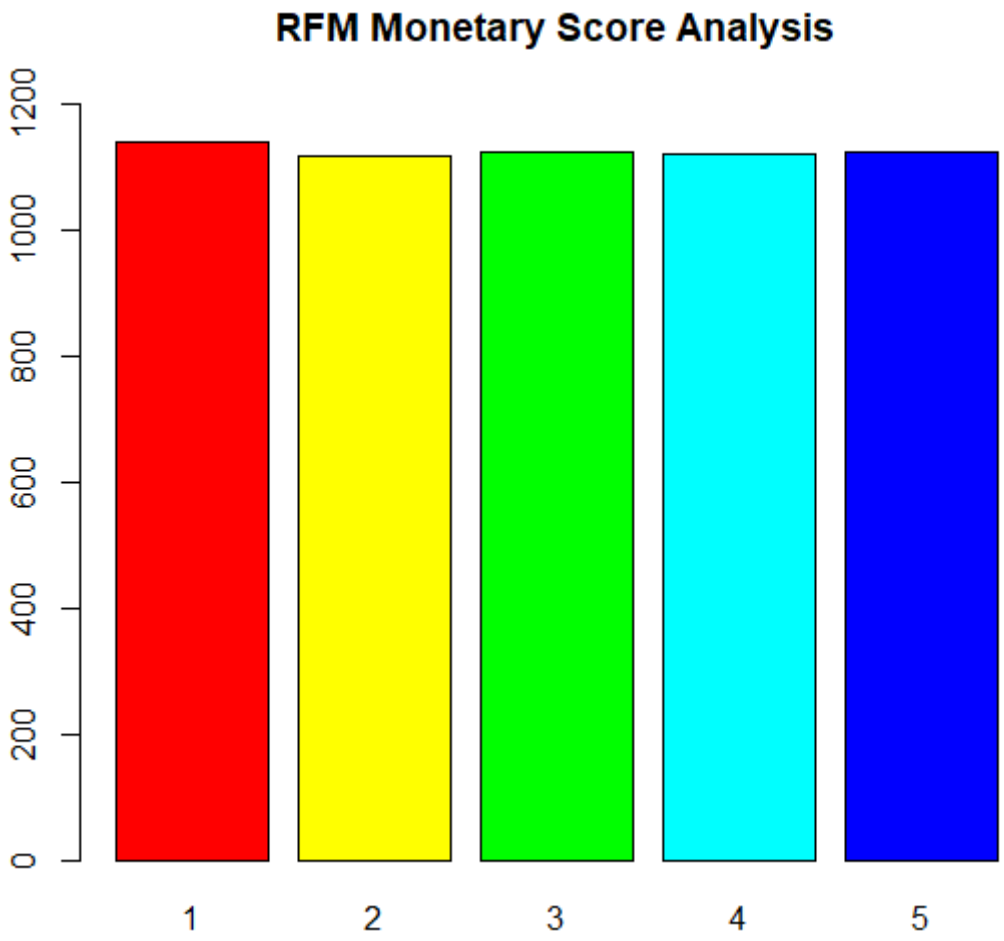

Figura 21 – Overall monetary score

Como podemos ver na Figura 21, em termos de monetary, os clientes estão equilibrados, com maior número de cliente com monetary score de 1.

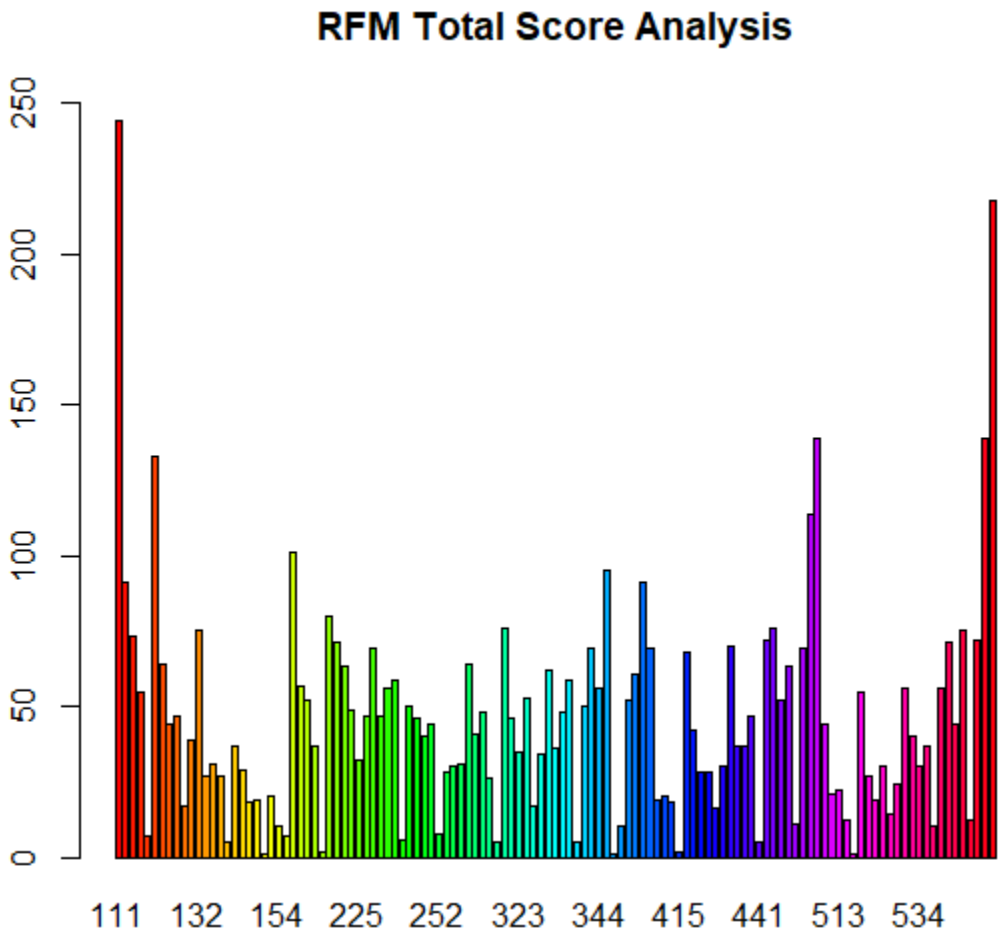

Figura 22 – Overall total score

Como podemos ver na Figura 22, existe uma grande quantidade de clientes nos extremos do score, 111 e 555, ou seja, clientes pouco frequentes, que já não fazem compras à muito tempo e não gastaram muito dinheiro nessas compras e no outro polo, clientes frequentes, que fazem compras frequentemente e que gastaram muito dinheiro nessas compras.

6. Clustering

Os métodos de agrupamento ou *Clustering* são usados para construção de grupos de objetos com base nas semelhanças e diferenças entre os mesmos, de tal maneira que os grupos obtidos sejam os mais homogéneos e bem separados possíveis.

6.1 Número de clusters a considerar

Para calcular o número correto de clusters a criar recorremos a duas técnicas normalmente utilizadas:

- Coeficiente de Silhouette

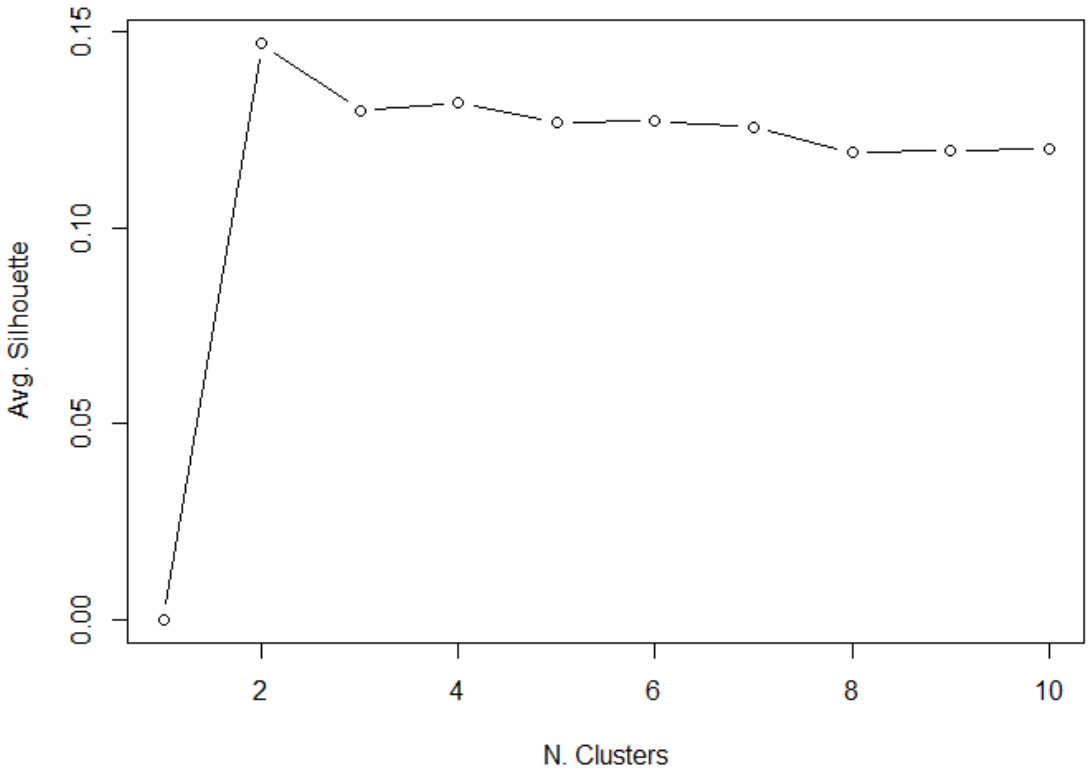

Figura 23 – Coeficiente de Silhouette

- Índice Calinski-Harabasz

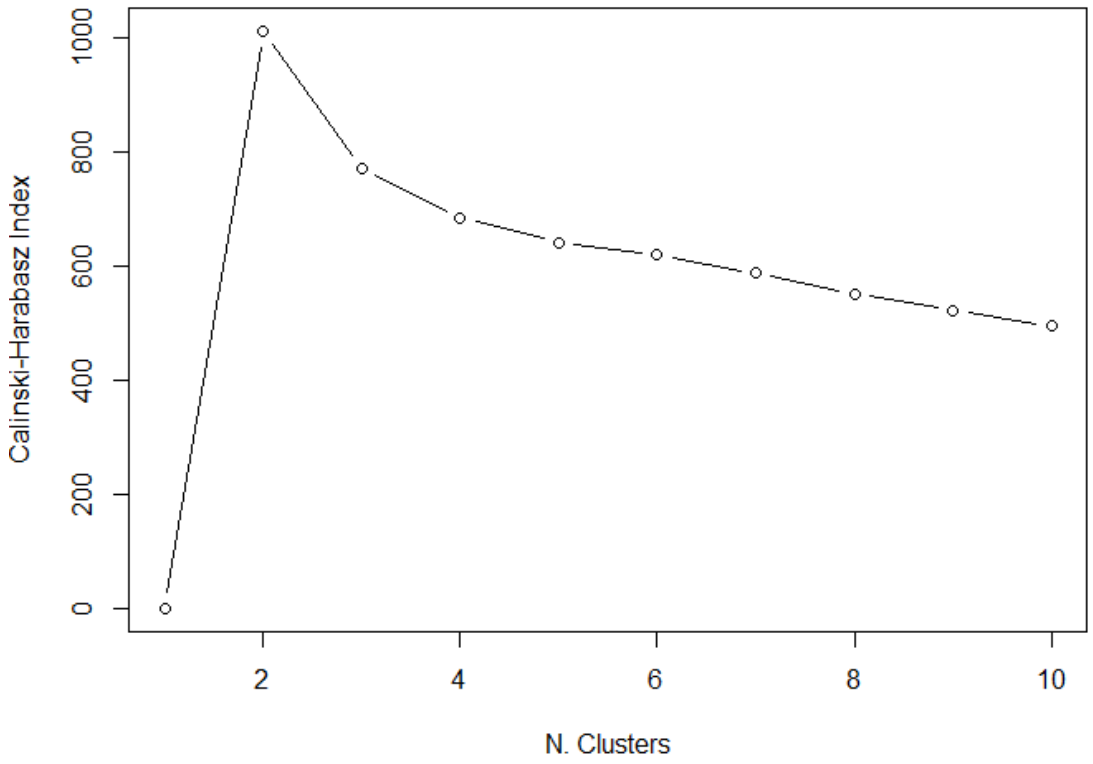

Figura 24 – Calinski-Harabasz

O número correto de clusters corresponde ao valor de K para o qual exista um "joelho" distinto, ao qual consideramos 3 clusters de acordo com o índice de Calinski-Harabasz.

6.2 K-means

Aplicando o algoritmo de k-means com K = 3 aos dados de rfm dos clientes, obtivemos os seguintes clusters:

- Cluster 1

R Score	F Score	M score
1.96	1.96	2.18

Este cluster representa clientes que não vão frequentemente às Perfumarias e foi há muito tempo que foram a uma. Quando vão, fazem compras com valor monetário relativamente baixo.

- Cluster 2

R Score	F Score	M score
3.10	2.35	2.36

Este cluster é parecido com o anterior só que clientes que a última vez que foram a uma Perfumaria é mais recente que o cluster anterior.

- Cluster 3

R Score	F Score	M score
3.82	4.19	4.01

Este cluster representa clientes que vão frequentemente às Perfumarias e foi há pouco tempo que foram a uma. Quando vão, fazem compras com valor monetário relativamente elevado.

Os seguintes gráficos apresentam a divisão dos dados de uma forma bastante visível nos 3 clusters definidos.

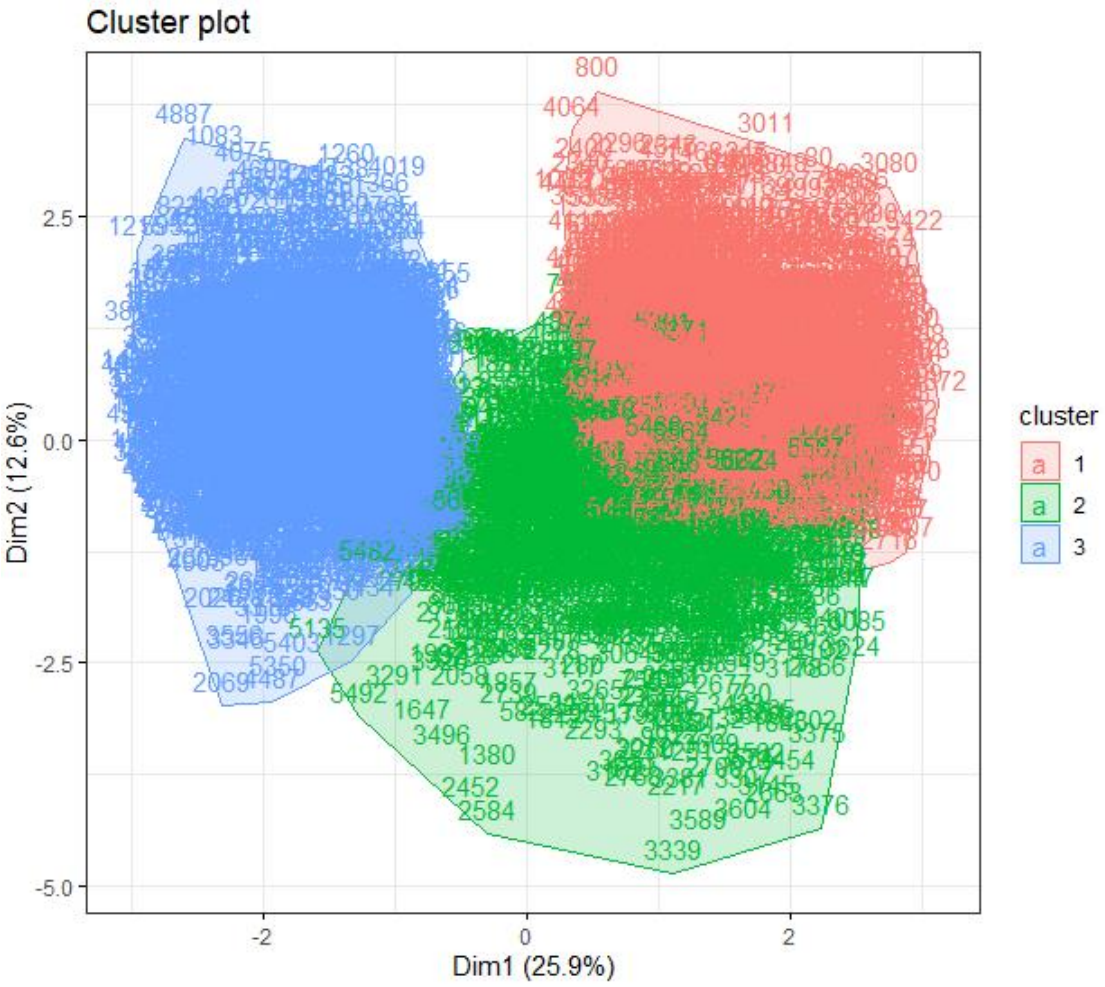

Figura 25 – Kmeans Clusters

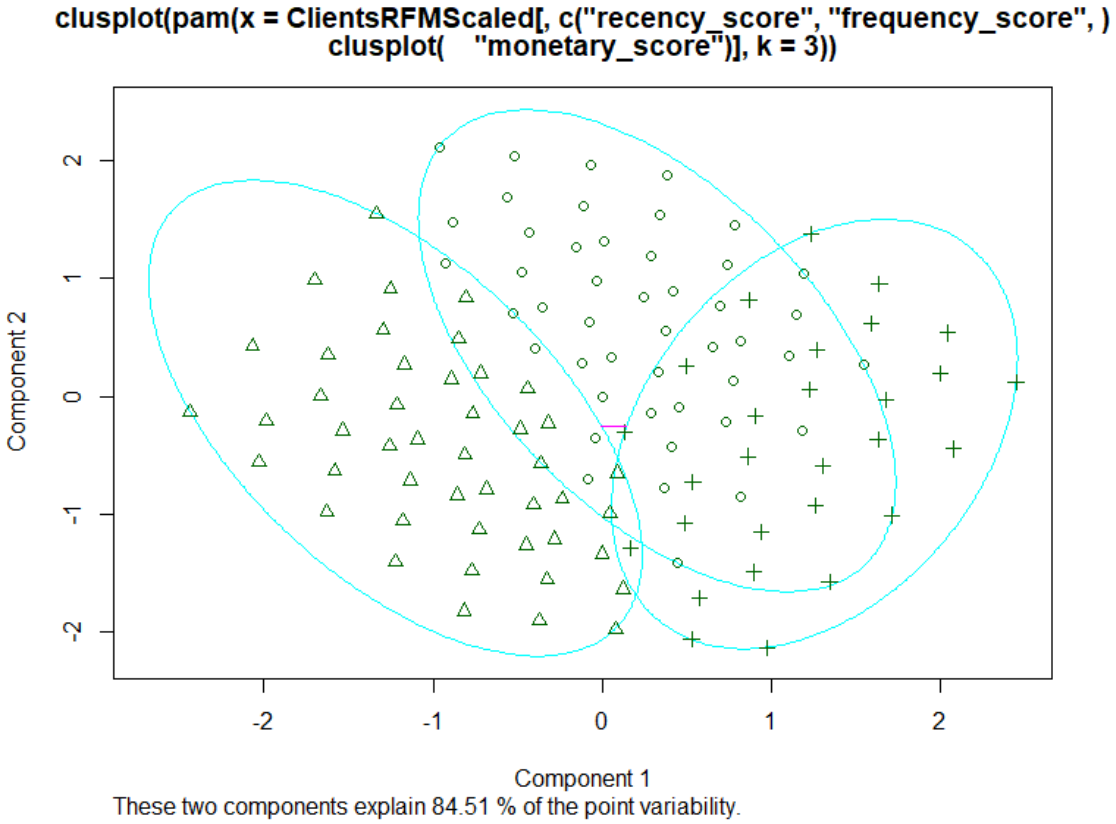

Figura 26 – Kmeans Clusters

7. Regras de associação

Neste capítulo, são demonstradas as regras de associação descobertas, indicando quais os produtos que são geralmente comprados em conjunto.

7.1 Apriori

Utilizando o algoritmo Apriori foi possível determinar vários grupos de produtos que são comprados conjuntamente. Na utilização deste algoritmo foi utilizado um suporte de 0.01 e um grau de confiança de 50% visto que, após vários testes, estes constituíam os melhores resultados.

Figura 27 – Visualização das regras de associação criadas

De acordo com a figura anterior, dividiram-se os produtos em grupos de produtos que são comprados em conjunto:

- Produtos 10058 e 10059 – *Group 1*
- Produtos 8434 e 8484 – *Group 2*
- Produtos 2693 e 2703 – *Group 3*
- Produtos 10414, 3463 e3462 – *Group 4*
- Produtos 10413, 3463 e 3462 – *Group 5*
- Produtos 3463 e 3462 – *Group 6*
- Restantes produtos – *Others*

Das regras mencionadas anteriormente, foram optadas por as regras que possuíam maior suporte/interesse para a criação de grupos de produtos. Foram também preferidas as regras que indicavam uma correlação bidirecional entre os produtos.

No caso de ser desejado um aumento de grupos de produtos, é possível introduzir os novos produtos desejados de forma simples, no script R realizado.

8. Modelos

8.1 Previsão do tipo de cliente obtido das regras de associação

Com base na pergunta "Qual o perfil dos potenciais compradores de um determinado produto ou conjunto de produtos" decidimos utilizar o seguinte modelo de previsão:

- Árvore de decisão (Modelo C5.0)

Como o dataset é de enormes dimensões, tentamos utilizar as Support Vector Machines mas sem sucesso visto que não conseguia alocar memória no sistema (17.7GB). Assim obtemos a seguinte *accuracy* com este modelo.

```
> results
  model accuracy
1  C5.0     0.82
```

Figura 28 – Resultado do C5.0 para Tipo de Cliente

Só é possível de calcular a *accuracy* visto que com o atributo objetivo contém 7 tipos, seria muito complicado calcular as restantes medidas com *precision, recall, kappa, f1*, etc.

8.2 Previsão do cluster do cliente

Com base na pergunta "Qual é o intervalo de consumo, a frequência e o valor gasto por um grupo de clientes num conjunto de produtos específico" decidimos utilizar os seguintes modelos:

- Árvore de decisão (Modelo C5.0)
- Support Vector Machines com as funçoes de kernel: **tanhdot, vanilladot e rbfdot**
- Naive Bayes

Utilizando o dataset obtido do ficheiro CLIENTS.txt e com algumas melhorias como remoção colunas desnecessárias (City, hasChildren, Client, PostalCode) para a previsão do cluster de um cliente, obtivemos os seguintes resultados:

```
> results[order(results$accuracy, decreasing = TRUE),]
                model accuracy
5         Naive Bayes    0.605
4     SVMmodel rbfdot    0.600
1                C5.0    0.580
3 SVMmodel vanilladot    0.580
2    SVMmodel tanhdot    0.376
```

Figura 29 - Resultados cluster do Cliente

Foi decidido calcular apenas a *accuracy* visto que como atributo objetivo só pode tomar 3 valores, não é necessário calcular as restantes medidas como *precision, recall, kappa, f1*, etc.

9. Avaliação dos Modelos

Visto que só foi possível de utilizar 1 modelo na previsão do tipo de cliente obtido das regras de associação, a avaliação que podemos realizar deste modelo é que existia uma grande quantidade de clientes com o tipo “Others” logo o modelo pode não ser o melhor devido aos dados de treino utilizados.

O *support vector machine* com a função de *kernel* rbfdot apresenta os melhores resultados, em conjunto com o algoritmo *Naive Bayes*, de acordo com a *accuracy*.

O *support vector machine* com a função de *kernel* tanhdot não é viável, pelo que apresenta uma *accuracy* inferior a 50%.

O modelo C5.0 e o *support vector machine* com a função de *kernel* vanilladot apresentam valores de *accuracy* semelhantes.

O *clustering* considera o RFM, e este tem em consideração o dataset e o ficheiro PURCHASES.txt. Tendo isto em consideração e os dados que foram fornecidos de treino aos modelos deste tipo de previsão, a *accuracy* destes podem ter sido influenciadas pela falta de dados utilizados no cálculo do RFM. Mesmo assim, o SVM com kernel rbfdot fornece os melhores resultados logo seria este modelo a utilizar para esta tarefa.

10. Conclusão

Com a realização deste trabalho, o grupo foi capaz de aprender e consolidar conhecimentos importantes a nível de tratamento e análise exploratória de dados. Foram também estudados os sistemas de recomendações, obtendo conhecimento acerca do seu funcionamento. De seguida, o grupo dedicou-se à elaboração de modelos de *machine learning* para *clustering*, regras de associação e classificação.

Utilizando os conceitos acima referidos, tornou-se possível responder a algumas das questões relativas ao negócio de uma rede de perfumarias, que poderá ser utilizada a favor da mesma.

Por último, foram ainda consolidados os conhecimentos sobre a avaliação e comparação de modelos, tendo estes sido utilizados para descobrir quais melhor preveem o conjunto de produtos que um cliente irá provavelmente comprar, e em que perfil se enquadra um dado cliente.