



Universidade de Coimbra
Faculdade de Ciências e Tecnologia
Departamento de Engenharia Informática

Mestrado em Engenharia Informática
Segurança em Tecnologias da Informação
Projeto 3, 2020/2021, 2º Semestre

Docentes

Professor Pedro Furtado
(PNF@DEI.UC.PT)

David Silva de Paiva	davidpaiva@student.dei.uc.pt	2020178529
Ricardo David Da Silva Briceño	briceno@student.dei.pt	2020173503

Registo de Trabalhos

Lista de funcionalidades/objetivos implementados

Funcionalidades/Objetivos	Responsável	Esforço (horas)
EDA	Ricardo	12
Clustering	Ricardo	8
Regressão	Ricardo	5
Análise dos Resultados	David	16

Autoavaliação individual e global do projeto

O aluno, David Paiva, autoavalia o desempenho do grupo com uma nota de dezanove valores - numa escala de zero a vinte. Relativamente a uma autoavaliação individual, avalia o seu desempenho e o do colega, Ricardo Briceño, com 17 valores para ambos.

O aluno, Ricardo Briceño, autoavalia o desempenho do grupo com uma nota de dezanove valores - numa escala de zero a vinte. Relativamente a uma autoavaliação individual, avalia o seu desempenho e o do colega, David Paiva, com 17 valores para ambos.

Índice de Figuras

Figura 1 - Receita média gerada por cada país entre 1998 e 1999	3
Figura 2 - Receita média gerada por mês em Africa entre 1998 e 1999	3
Figura 3 - Receita média gerada por mês na América 1998 e 1999	4
Figura 4 - Receita média gerada por mês na Europa 1998 e 1999	5
Figura 5 - Receita média gerada por mês na Ásia entre 1998 e 1999	5
Figura 6 - Receita média gerada por mês no Médio Oriente entre 1998 e 1999	6
Figura 7 - Variação do número de encomendas realizadas entre 1998 e 1999 em cada país	7
Figura 8 - Variação do número de produtos comprados por país entre 1998 e 1999	7
Figura 9 - Variação do número de encomendas realizadas por mês em 1998 e 1999	8
Figura 10 - Variação do número de produtos comprados por mês em 1998 e 1999	9
Figura 11 - Número de vendas da melhor marca de cada mês nos anos de 1998 e 1999	10
Figura 12 - Número de vendas da melhor marca em cada estação nos anos de 1998 e 1999	11
Figura 13 - Número de vendas da marca menos vendida em cada mês nos anos de 1998 e 1999	12
Figura 14 - Número de vendas da marca menos vendida em cada época nos anos de 1998 e 1999	12
Figura 15 - Categorias mais vendidas em cada mês nos anos de 1998 e 1999	13
Figura 16 - Número de vendas da melhor categoria em cada época nos anos de 1998 e 1999	14
Figura 17 - Número de vendas da categoria menos vendida em cada mês nos anos de 1998 e 1999	15
Figura 18 - Número de vendas da categoria menos vendida em cada época nos anos de 1998 e 1999	16
Figura 19 - Cotovelo (Elbow) para agrupamento K-Means	17
Figura 20 - Critério de escolha para nomenclatura dos grupos	18
Figura 21. Clustering – valor total de venda em função da quantidade de produtos.....	19
Figura 22. Comparação dos modelos - Média e Desvio Padrão.....	20
Figura 23. Comparação dos algoritmos – BoxPlot.....	21
Figura 24. Regressão linear - valor da compra em função da quantidade de produtos..	21
Figura 25. Métricas para análise.....	22

Índice de Tabelas

Tabela 1 - Valores estatísticos das vendas da melhor marca de cada mês nos anos de 1998 e 1999	10
Tabela 2 - Valores estatísticos das vendas da melhor categoria de cada mês nos anos de 1998 e 1999	13
Tabela 3 - Valores estatísticos das vendas da melhor categoria de cada época nos anos de 1998 e 1999	14

Índice

Registo de Trabalhos	i
Lista de funcionalidades/objetivos implementados	i
Autoavaliação individual e global do projeto	i
Índice de Figuras	ii
Índice de Tabelas	ii
1. Enquadramento do Trabalho	1
2. Análise dos dados.....	2
2.1 Receitas geradas em cada país com a venda de produtos	2
2.2 Receitas geradas em cada região com a venda de produtos.....	3
2.3 Variação no número de encomendas e produtos comprados	6
2.4 Marcas mais e menos vendidas	9
2.5 Categoria mais e menos vendidas	12
3. Clustering	17
3.1 Clustering – valor total de venda em função do valor total da compra	18
4. Regressão	20
5. Considerações Finais.....	23

Esta página foi deixada propositadamente em branco.

1. Enquadramento do Trabalho

Nesta secção é apresentado o objetivo do terceiro projeto prático da unidade curricular de Sistemas e Gestão de Dados.

Durante as aulas práticas da unidade curricular houve a oportunidade de interação com diferentes técnicas de tratamento e análise de dados. Com isso em mente, neste projeto é pedido que se analise e compre os dados de dois data sets relativos com dados relativos a dois anos de compras do SSB/TPC-H.

De modo a concretizar os objetivos propostos no projeto, foi realizado um tratamento fino aos dados e de seguida foi feita uma exploratory data analysis (EDA) que é apresentada na secção 2. Para além disso, foi feito um cluster e duas regressões de modo a serem aplicadas as técnicas estudadas, no entanto na regressão não foi possível retirar grande informação. Os resultados obtidos em ambas são apresentados na secção 3 e 4, respetivamente.

2. Análise dos dados

Nesta secção são apresentados os resultados extraídos após uma análise dos dados dos dois data sets de vendas.

2.1 Receitas geradas em cada país com a venda de produtos

A análise do rendimento gerado pelas exportação e venda de bens e serviços é fundamental aquando de uma análise macroeconómica de um país/região. Com isso em mente, procurou-se encontrar diferenças no valor médio gerado por cada um dos países e regiões.

Do ponto de vista macroscópico e comparando o valor médio gerado pela venda de produtos nos anos de 1998 e 1999, é possível concluir – pela observação do gráfico apresentado na Figura 1 – quais os países que conseguiram vender mais artigos e, consequentemente, aumentar o valor de riqueza gerado de um ano para o outro. Com a observação do gráfico, concluímos que foram seis os países que aumentaram as receitas no ano de 1999, face ao ano de 1998. São eles, a Jordânia, Marrocos, Reino Unido, Iraque, Irão e a Indonésia. O Japão manteve em 1999 o mesmo valor de receita gerada, face ao ano de 1998. Todos os restantes países mencionados no gráfico, baixaram ligeiramente o valor gerado com as vendas no ano de 1999, comparativamente com o ano de 1998.

Por fim, é possível ver que a Etiópia foi o país que mais riqueza gerou com a venda de produtos nos anos dois anos, tendo gerado mais de 1.2 mil milhões de dólares. Isto pode parecer estranho uma vez que a Etiópia é um país em vias de desenvolvimento e considerado um dos países mais pobres do Mundo. No entanto, isto pode ser possível uma vez que a mão de obra nestes países é extremamente barata ou até gratuita, levando a que muitas empresas situem os seus centros de produção nestes países, aumentando assim a receita com a venda dos produtos produzidos. Para além disso, a receita média encontra-se dentro de um intervalo de confiança de 95 %.

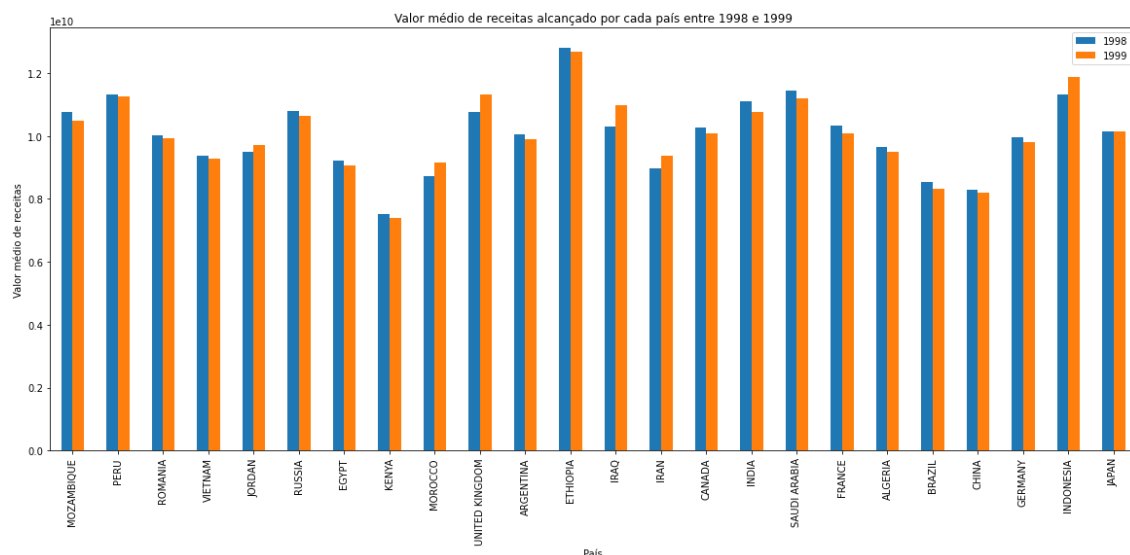


Figura 1 - Receita média gerada por cada país entre 1998 e 1999

2.2 Receitas geradas em cada região com a venda de produtos

Feita a comparação entre países durante os dois anos, é interessante observar a variação das receitas ao longo dos meses em cada uma das regiões.

Relativamente ao continente africano e observando o gráfico apresentado na Figura 2, é possível concluir que em ambos os anos as receitas não tiveram grande variação. No entanto, no mês de janeiro observa-se uma quebra nas vendas no ano de 1999 face ao período homólogo do ano anterior. O mesmo acontece no mês de novembro, se bem que com uma diferença menos acentuada. Já no mês de fevereiro o comportamento foi contrário e no ano de 1999 registou-se um ligeiro aumento na receita gerada pela venda de produtos, comparativamente ao mesmo período do ano anterior.

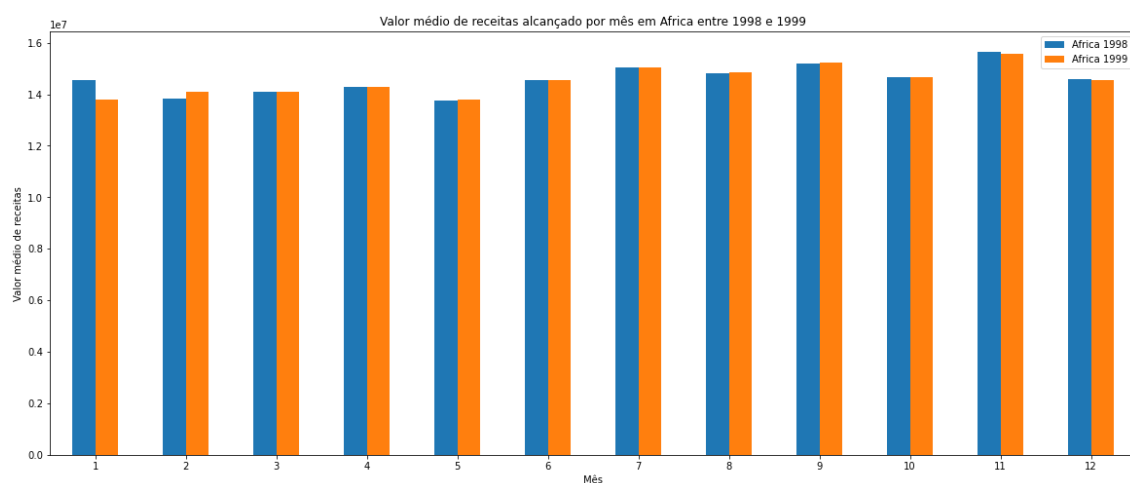


Figura 2 - Receita média gerada por mês em Africa entre 1998 e 1999

Na América, após observação do gráfico apresentado na Figura 3, conclui-se que, não há grande flutuação nas receitas em 1999 face ao ano de 1998. No entanto, importa realçar

Trabalho Prático nº 3

que nos meses de janeiro, fevereiro e julho houve um ligeiro aumento nas receitas no ano de 1999 comparativamente ao período homólogo do ano anterior. Em contraste, nos meses de março, abril, agosto, novembro e dezembro ocorreu uma ligeira quebra nas receitas do ano de 1999 comparativamente ao mesmo período do ano de 1998. Nos restantes meses as receitas mantiveram-se nos mesmos valores no ano de 1999. Por fim, os meses de março, agosto e novembro são os que apresentam um maior encaixe de receitas no ano de 1998 e, no ano de 1999 são os meses de janeiro, agosto e novembro.

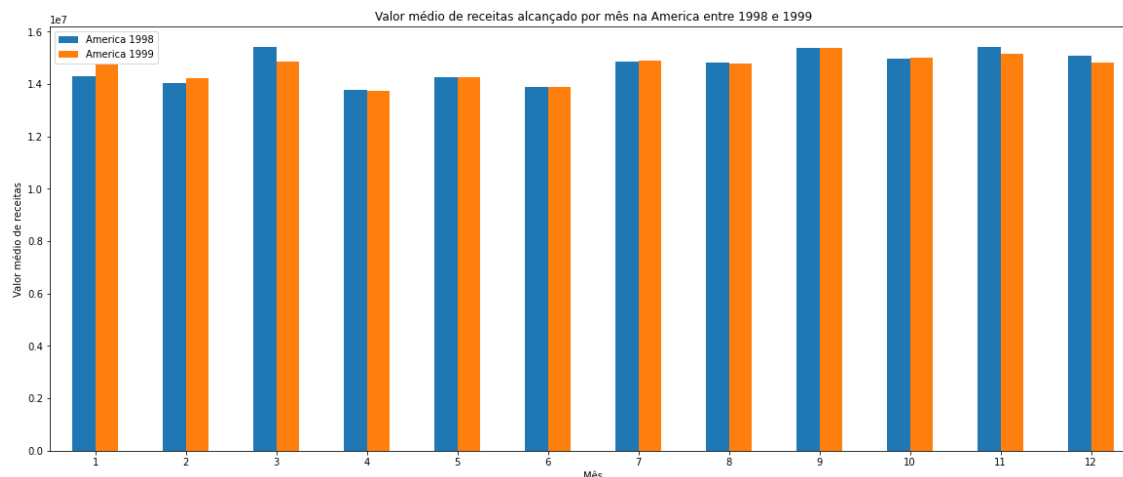


Figura 3 - Receita média gerada por mês na América 1998 e 1999

Analisando o gráfico apresentado na Figura 4 é possível concluir que as receitas nos anos de 1999 e 1998 não apresentam grande flutuação. No entanto, apenas foram três os meses em que se manteve o mesmo valor de receita no ano de 1999 comparativamente a 1998, sendo que foram os meses de julho, agosto e setembro. Também foram nestes meses que se registaram a maior faturação na Europa, rondando os 1.5 milhões de dólares. Por sua vez, nos meses de janeiro, março, abril, maio, junho e setembro registou-se uma quebra nas receitas, destacando-se o mês de janeiro que baixou dos mais de 1.4 milhões de dólares para os cerca de 1.3 milhões. Em contraste, nos meses de fevereiro, novembro e dezembro, houve um ligeiro aumento nas receitas regeradas na venda de produtos.

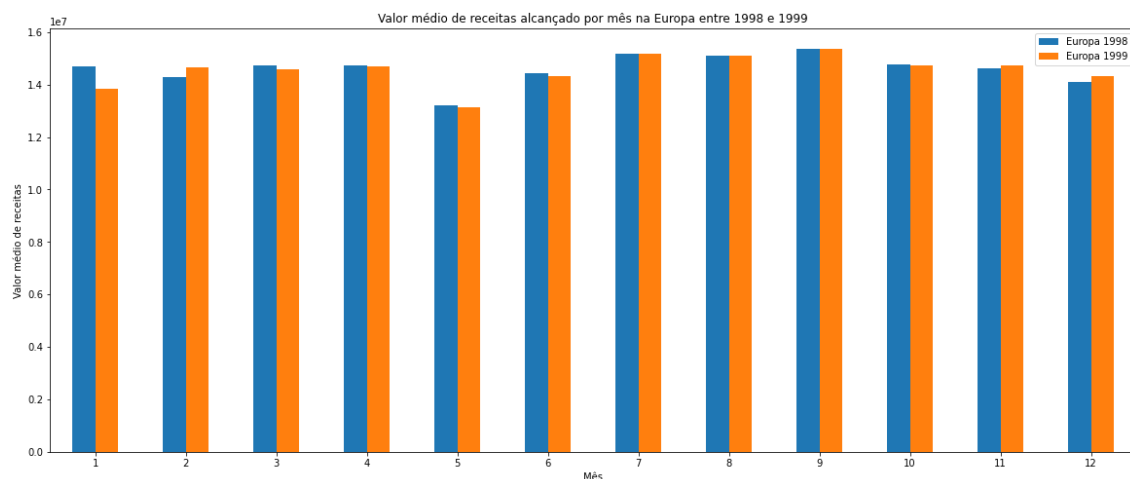


Figura 4 - Receita média gerada por mês na Europa 1998 e 1999

Relativamente à Ásia e analisando o gráfico apresentado na Figura 5 é possível concluir que há uma tendência de manter valores muito semelhantes no que toca às receitas geradas pela venda de produto nos dois anos. Nos meses de março, abril, julho, agosto, setembro e outubro, o valor médio das receitas manteve-se igual no ano de 1999 face ao período homólogo do ano anterior. Já nos meses de janeiro, maio, junho, novembro e dezembro ocorreu uma ligeira quebra nas receitas face ao mesmo período de 1998. Em contraste, apenas no mês de fevereiro é que houve um ligeiro aumento na faturação em 1999 comparativamente ao ano de 1998.

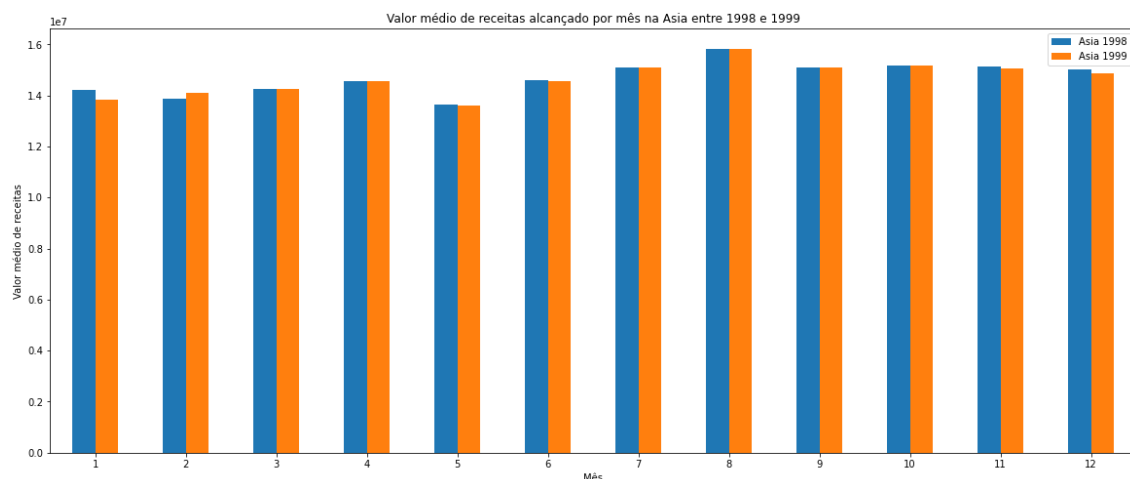


Figura 5 - Receita média gerada por mês na Ásia entre 1998 e 1999

Relativamente ao Médio Oriente, após observação do gráfico apresentado na Figura 6, conclui-se uma vez mais que não há grande diferença nas receitas obtidas no ano de 1999 face a 1998. No entanto, nos meses de janeiro, fevereiro, maio e novembro é possível observar um ligeiro aumento das receitas obtidas, sendo que, o mês de fevereiro, foi o que obteve uma maior subida nas receitas, chegando a ultrapassar os 1.4 milhões de

dólares comparativamente com o mesmo período do ano anterior. Num cenário inverso, os meses de março e dezembro apresentam uma quebra no valor médio das receitas arrecadadas, face ao período homólogo de 1998.

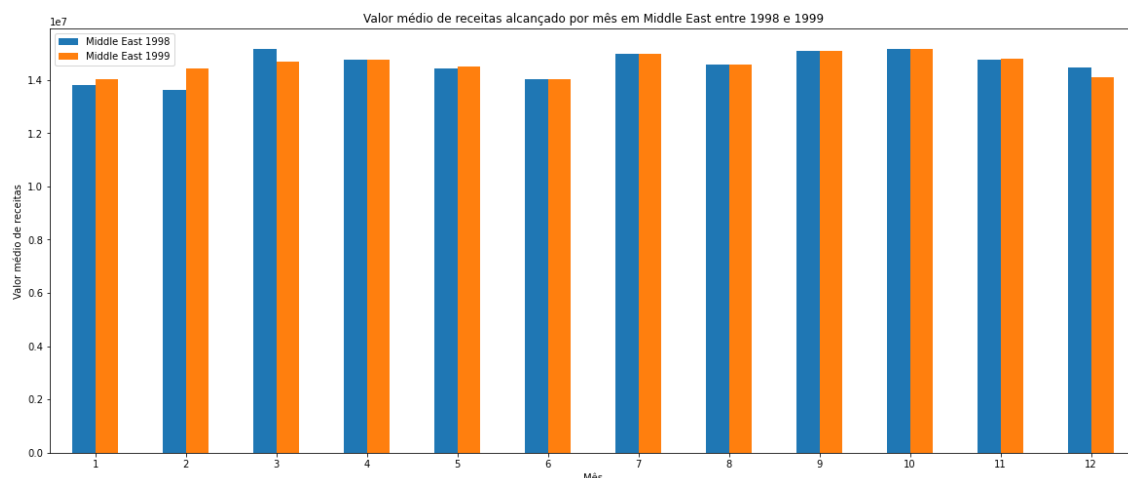


Figura 6 - Receita média gerada por mês no Médio Oriente entre 1998 e 1999

2.3 Variação no número de encomendas e produtos comprados

Uma vez feita a comparação do valor médio arrecadado com as receitas nos anos de 1998 e 1999, é importante comparar a variação de número de encomendas realizadas nos dois anos e, também, entre países de modo a perceber um pouco os padrões de consumo dos consumidores.

Relativamente à variação do número de encomendas realizadas entre países nos dois anos é possível retirar algumas conclusões após observação do gráfico apresentado na Figura 7. Destacam-se quatro países com um elevadíssimo número de encomendas realizadas em ambos os anos, sendo eles a Roménia, Argélia, Jordânia e o Brasil que apresentam um valor médio de 1300 encomendas anuais. No entanto, apesar de serem os países com o maior número de encomendas realizadas, no ano de 1999 a Roménia, Argélia e a Jordânia registaram uma pequena quebra nesse número, face ao ano anterior. Para além disso, é possível observar que o Japão é o país que regista um menor número de encomendas realizadas, seguido do Vietnã e do Reino Unido. Por fim, importa referir que todos os restantes países apresentam um número de encomendas realizadas igual em ambos os anos e a maioria deles entre as 400 e 600 encomendas realizadas.

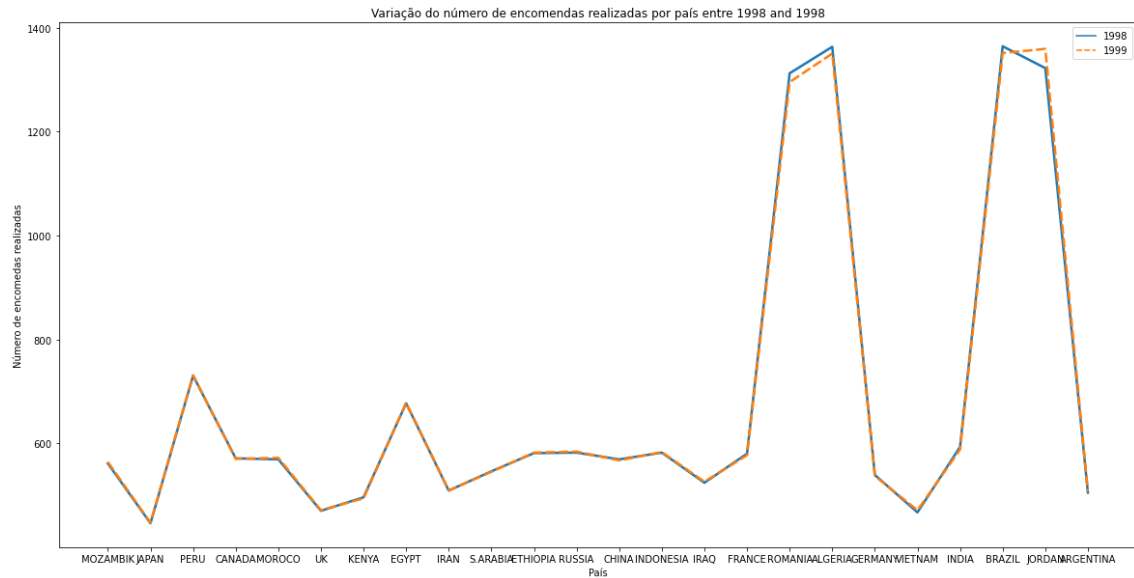


Figura 7 - Variação do número de encomendas realizadas entre 1998 e 1999 em cada país

Observando a variação do número de produtos comprados por país nos anos de 1998 e 1999 apresentada no gráfico da Figura 8, é possível corroborar as conclusões retiradas da análise anterior uma vez que, o número de encomendas realizadas está relacionado com o número de produtos comprados. Com isso em mente, é possível observar o mesmo comportamento e variação no número de produtos comprados entre países nos dois anos.

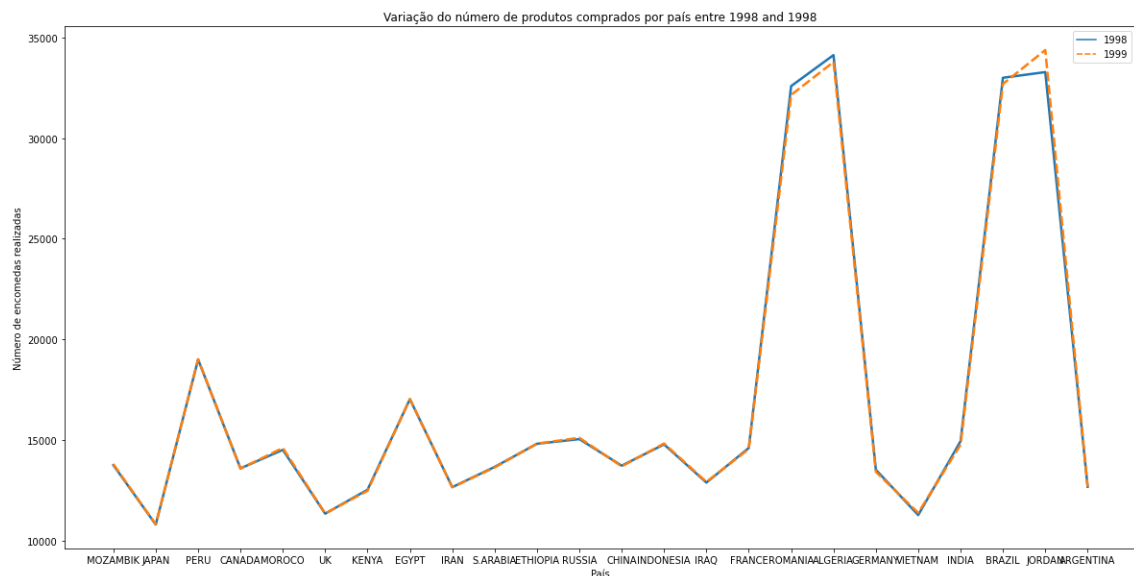


Figura 8 - Variação do número de produtos comprados por país entre 1998 e 1999

Analisando com detalhe o número de encomendas realizadas por mês nos dois anos – apresentado na Figura 9 – é possível concluir que há alguma flutuação do número de

Trabalho Prático nº 3

encomendas realizadas ao longo do ano – esta flutuação ocorre em ambos os anos, refletindo-se de forma diferente.

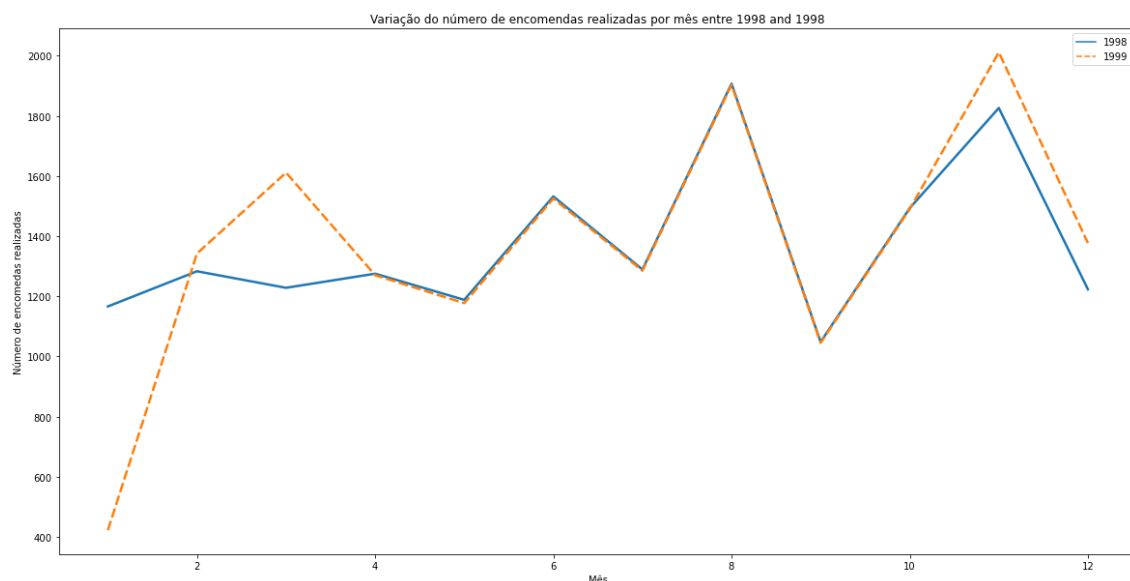


Figura 9 - Variação do número de encomendas realizadas por mês em 1998 e 1999

Olhando para o ano de 1998 é possível observar que o número de encomendas em janeiro situa-se próximo das 1200, oscilando em torno deste valor até ao mês de maio. De maio a agosto ocorrem alguns picos, sendo que no mês de agosto é registado o maior número de encomendas realizadas – o valor ultrapassa as 1800 encomendas. Em setembro o número de encomendas volta a descer para as cerca de 1000 encomendas e no mês seguinte volta a subir. No entanto, no fim do ano desde e dezembro encerra com um número total de vendas a rondas as 1300 encomendas.

Relativamente ao ano de 1999, observa-se que o ano não começa da melhor maneira, uma vez que em janeiro foram registadas poucas mais que 400 encomendas. Este valor melhora de forma contínua até ao mês de março, quando se atinge cerca de 1600 encomendas realizadas nesse mesmo mês. Comparativamente com o os três primeiros meses de 1998, observa-se que o início do ano piorou, mas que nos meses de fevereiro e março o número de encomendas realizadas foi superior ao mesmo período do ano anterior. De março ao fim de setembro, o número de encomendas realizadas foi exatamente igual em ambos os anos. No entanto, de outubro a dezembro registou-se um aumento no número de encomendas realizadas, comparativamente ao período homologado de 1998. Mas este aumento mantém o mesmo comportamento do ano anterior, ou seja, de outubro a novembro à um aumento do número de encomendas realizadas, mas de novembro para dezembro há uma quebra no número de comendas realizadas. Importa

realçar que o mês de novembro de 1999 é o mês que regista o maior valor de encomendas realizadas – 2011 encomendas, mais precisamente.

Analisando com detalhe o número de produtos comprados por mês nos dois anos – apresentado na Figura 10 – conclui-se que o comportamento da compra de produtos acompanha a análise que foi feita, anteriormente, relativa ao número de encomendas realizadas.

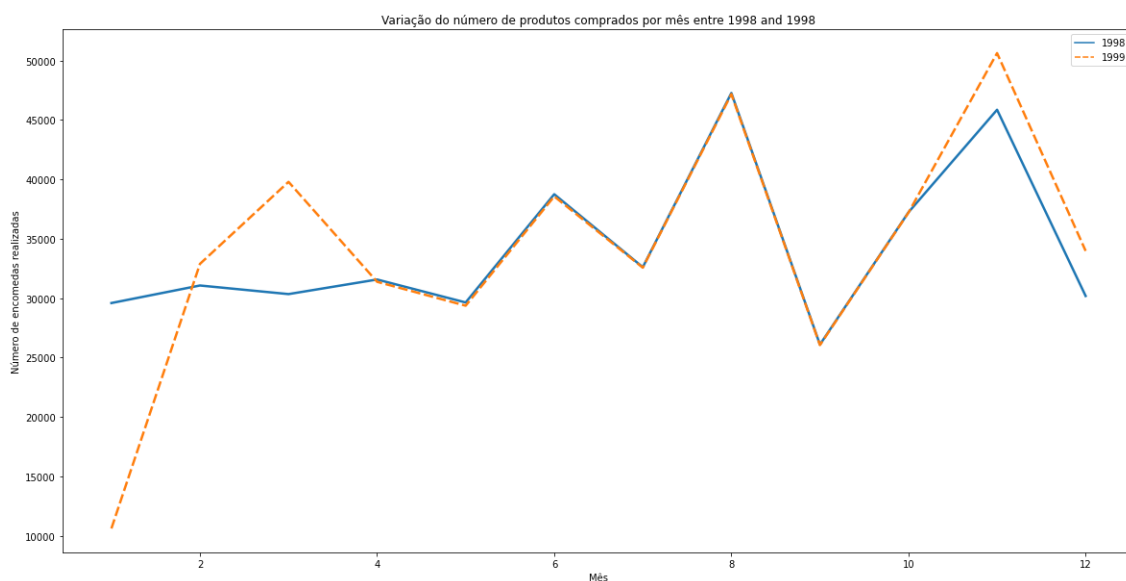


Figura 10 - Variação do número de produtos comprados por mês em 1998 e 1999

2.4 Marcas mais e menos vendidas

Feita uma análise mais geral relativa às receitas geradas e ao número de encomendas realizadas, assim como o número de produtos vendidos, é fundamental analisar quais as marcas que se destacam ou não ao nível das vendas.

Observando o gráfico apresentado na Figura 11 e focando apenas no ano de 1998, conclui-se que no mês de janeiro do ano 1998 a marca que mais vendeu foi a BRR#4340 – realizando mais de 800 vendas nesse mês. Olhando para o mês de fevereiro, conclui-se que a marca mais vendida foi BRR#2310 com mais de 800 vendas registadas. No mês de março a marca mais vendida foi a BRR#3340 somando perto de 800 vendas. Seguindo, no mês de abril a marca mais vendida foi a BRR#2330 somando um valor um pouco superior às 800 vendas. No mês de maio, a marca mais vendida foi a BRR#3410 com um valor superior às 800 vendas. Entrando no verão, no mês de junho a marca que registou um maior número de vendas foi a BRR#1310, tendo registado também um valor superior um pouco superior às 800 vendas. Já no mês de julho, a marca que mais vendeu foi a

Trabalho Prático nº 3

marca BRR#2230. No mês de agosto é quando se vê o maior aumento no número de vendas, conseguindo assim a marca BRR#1430 atingir o magnífico recorde de vendas que ultrapassa as 1400 num mês. Nos restantes meses do ano, setembro, outubro, novembro e dezembro as marcas que mais venderam foram as BRR#3310, BRR#3210, BRR#3210 e BRR#4310, respetivamente.

No ano de 1999 a marca mais vendida em todos os meses do ano, excepto o mês de dezembro, foi a marca BRR#4310 – registando um número médio de vendas a rondar as 800. Já no mês de dezembro a marca que recebeu o maior número de encomendas foi a BRR#1240, registando o valor mais elevado desse mesmo ano que foi de 814 vendas como é possível ver na Tabela 1.

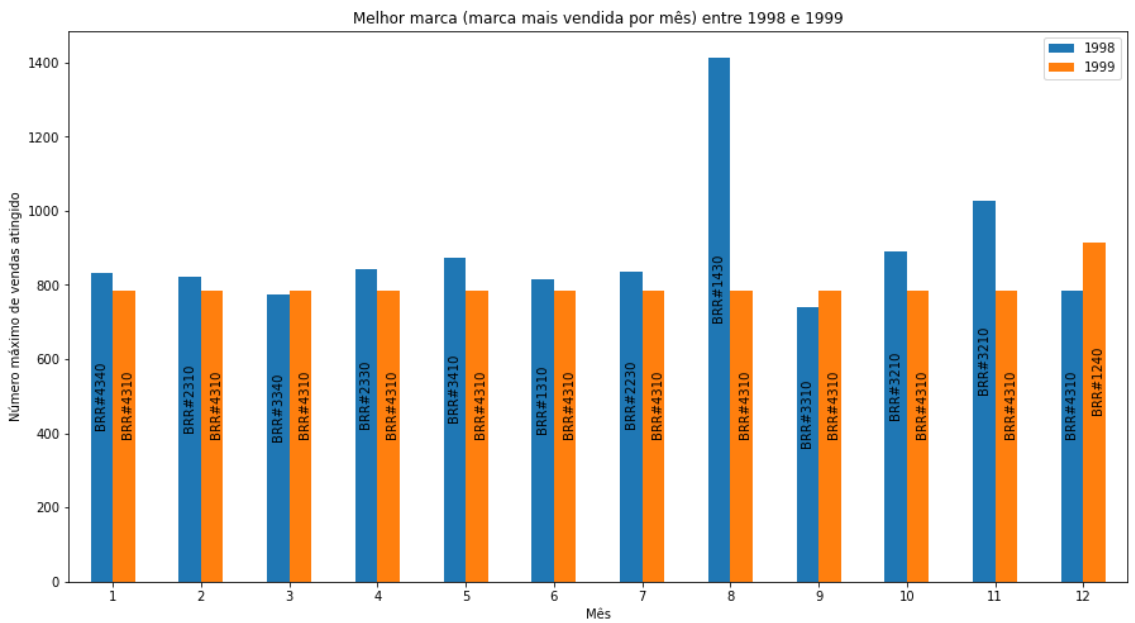


Figura 11 - Número de vendas da melhor marca de cada mês nos anos de 1998 e 1999

	1998	1999
Média	887.666	794.833
Desvio padrão	180.7420	37.527
Valor mínimo	741	784
Valor máximo	1414	914

Tabela 1 - Valores estatísticos das vendas da melhor marca de cada mês nos anos de 1998 e 1999

Olhando um pouco para o número de vendas em cada uma das estações nos anos de 1998 e 1999 - Figura 12 – conclui-se que no ano de 1998 a marca que mais vendeu numa única

estação foi a BRR#1430, registrando um valor superior às 3000 vendas no outono desse ano. Relativamente, ao ano de 1999 a marca que mais vendeu numa única estação foi a marca BRR#2230 que apenas na época do Natal registou um valor próximo das 3300 vendas, registrando assim o recorde de vendas por estação nos dois anos.

No inverno e na primavera a marca que mais vendeu no ano de 1998 foi a marca BRR#3210 com um valor médio de 1500 vendas. Já no período homologado do ano seguinte, a marca que mais vendeu foi a BRR#1140 registrando um valor próximo das 2000 encomendas. No verão, a marca mais vendida do ano 1988 foi a BRR#2330 registrando o valor mais baixo comparativamente com as outras estações – 842 vendas.

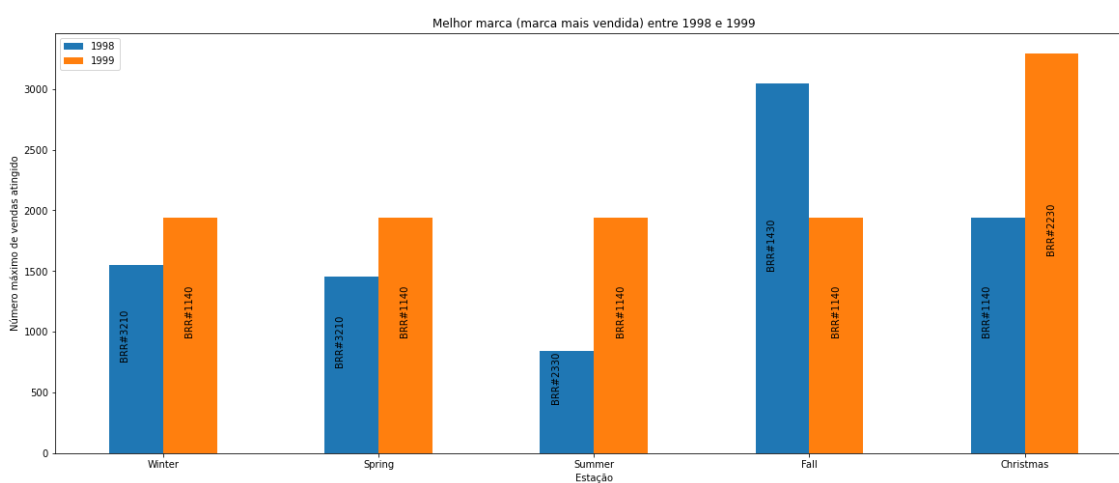


Figura 12 - Número de vendas da melhor marca em cada estação nos anos de 1998 e 1999

Feita a análise às marcas mais vendidas por mês e estação de cada um dos anos é necessário olhar também para as marcas que menos venderam nos mesmos períodos.

Analisando as marcas que menos venderam em cada mês do ano - Figura 13 – conclui-se que a marca que menos vendeu no ano de 1998 foi a BRR#4130 que registou apenas 130 vendas no mês de janeiro. Apesar disso, no mesmo ano, a marca BRR#2320 conseguiu registar um número de vendas máximo de 382 vendas no mês de agosto sendo a marca que obteve um melhor resultado na amostra de marcas com pior desempenho a nível de vendas.

Relativamente ao ano de 1999, a marca BRR#3130 foi a marca que menos vendeu o ano todo, apresentado um número médio de 215 vendas mensais

Trabalho Prático nº 3

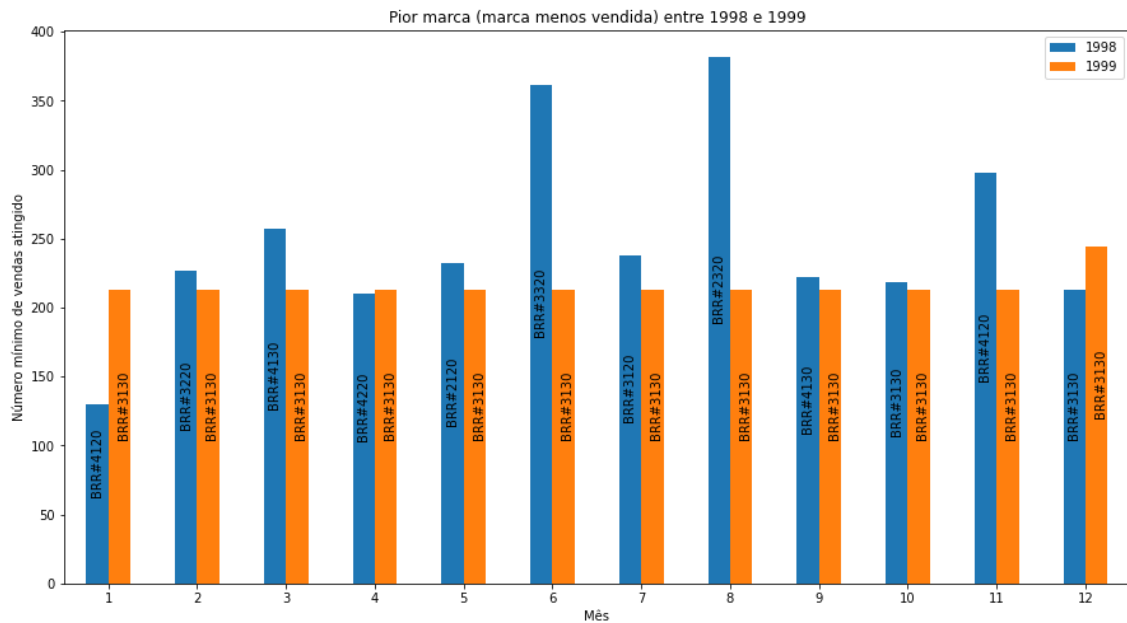


Figura 13 - Número de vendas da marca menos vendida em cada mês nos anos de 1998 e 1999

Com foco no número de vendas realizadas pelas marcas com menos volume de negócio em cada época do ano, é possível observar na Figura 14 que no ano de 1998 a marca que menos vendeu no inverno, primavera, verão, outono e Natal foi a BRR#3220, BRR#2110, BRR#4120, BRR#4110 e BRR#4120, respetivamente.

Já no ano de 1999, a marca que menos vendeu em todas as épocas foi a marca BRR#4120, apresentando um valor médio de 819 vendas em todas as épocas do ano.

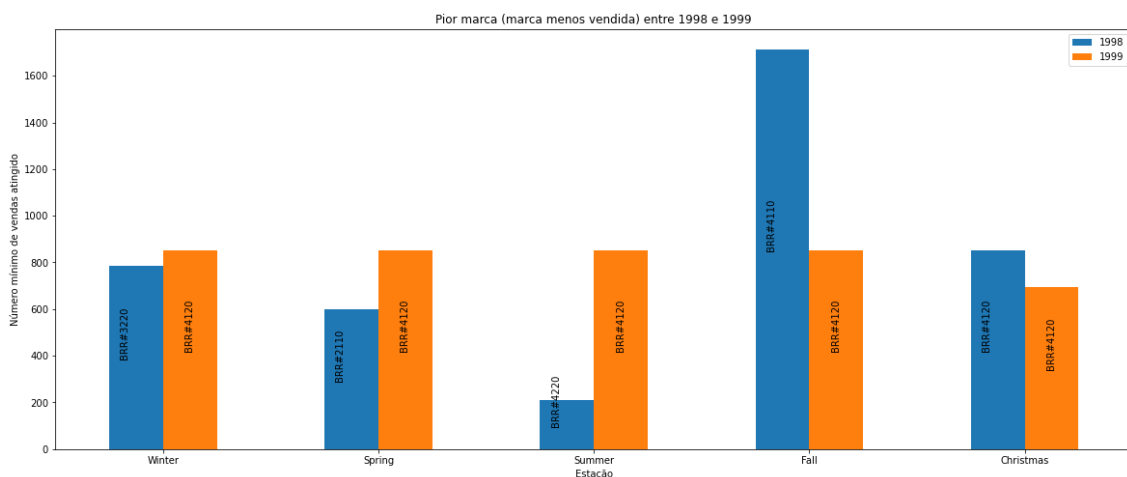


Figura 14 - Número de vendas da marca menos vendida em cada época nos anos de 1998 e 1999

2.5 Categoria mais e menos vendidas

Para além das marcas que mais e menos vendem é também interessante analisar quais as categorias de produtos mais procuradas pelos consumidores. Isto pode ajudar, por

exemplo, as marcas a introduzirem um novo produto no mercado, percebendo como é que o mercado se comporta e quais os hábitos e gostos dos consumidores.

A Figura 15 apresenta o número de vendas da categoria mais vendida em cada um dos meses dos anos 1998 e 1999.

Relativamente às categorias é possível observar que no ano de 1999 a categoria CATR#430 é a mais vendida de janeiro a novembro, vendendo em média 2419 vendas. Já no mês de dezembro do mesmo ano, é a CATR#120 a categoria mais vendida com as vendas a atingir o valor máximo desse ano, 2748 unidades vendas - Tabela 2.

Olhando agora para o ano de 1998, observa-se que três categorias se destacam em três meses diferentes, sendo que uma delas atinge o máximo de vendas nos dois anos, valor que corresponde a 3601 vendas - Tabela 2 – neste caso, trata-se a categoria CATR#330. De seguida, seguem-se as categorias CATR#420 e CATR#220, pela ordem correspondente.

Por fim, comparativamente com o ano de 1999, em 1998 de uma forma geral as categorias venderam em média mais. Uma diferença ligeiramente superior a 200 vendas.

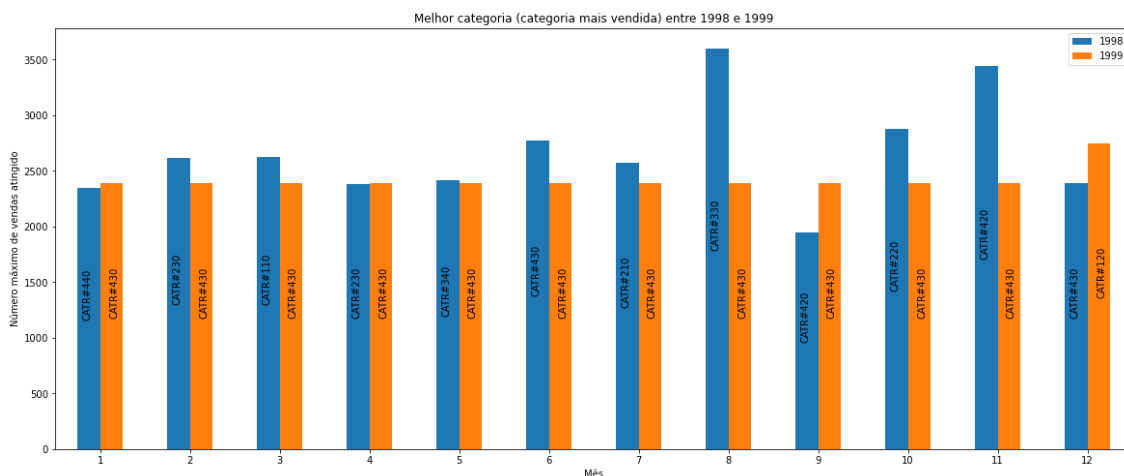


Figura 15 - Categorias mais vendidas em cada mês nos anos de 1998 e 1999

	1998	1999
Média	2666	2419
Desvio padrão	465.067	103.346
Valor mínimo	1950	2390
Valor máximo	3601	2748

Tabela 2 - Valores estatísticos das vendas da melhor categoria de cada mês nos anos de 1998 e 1999

Trabalho Prático nº 3

Após uma análise mais focada em cada mês do ano, é interessante analisar o desempenho das melhores marcas em cada uma das épocas do ano.

No ano de 1999 há uma constante assegurada pela categoria CATR#110 que segura o melhor número de vendas durante quatro épocas seguidas, nomeadamente no inverno, na primavera, no verão e no outono, perdendo na época do Natal para a categoria CATR#220. Comparativamente ao ano de 1998, em 199 há uma menor volatilidade no número de vendas das melhores categorias apresentado um desvio padrão de apenas 214 vendas, face ao desvio padrão de 2892 vendas do ano de 1998 – ver Tabela 3.

Por fim, comparando os dois anos e olhando para todas as estações a categoria que obteve melhor resultado foi a CATR#340 obtendo o máximo em vendas que chegou ao simpático valor de 10161 vendas.

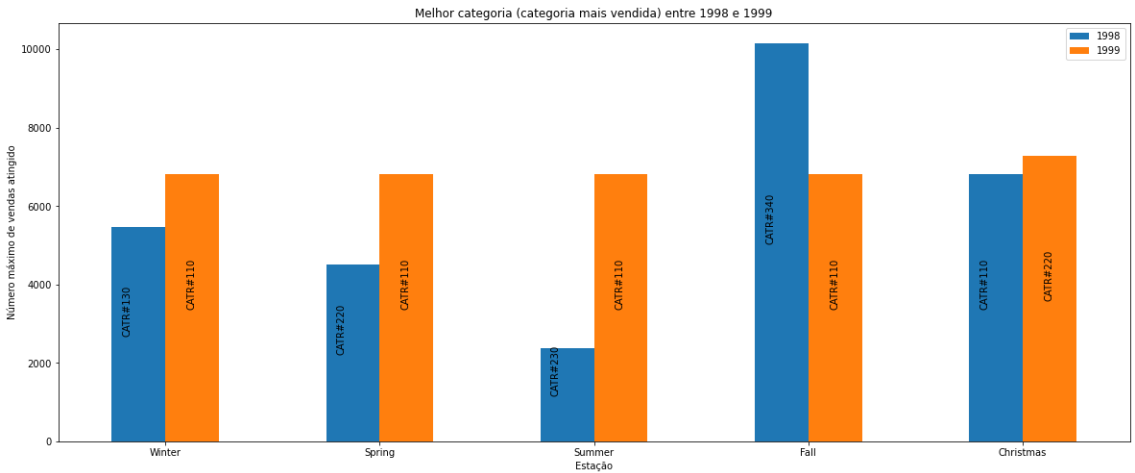


Figura 16 - Número de vendas da melhor categoria em cada época nos anos de 1998 e 1999

	1998	1999
Média	5866	6901
Desvio padrão	2891.656	214.663
Valor mínimo	2384	6805
Valor máximo	10161	7285

Tabela 3 - Valores estatísticos das vendas da melhor categoria de cada época nos anos de 1998 e 1999

Vistas as categorias com melhor desempenho a nível de vendas nos dois anos é necessário ver o reverso da medalha, ou seja, as categorias que menos venderam nos mesmos períodos de 1998 e 1999.

Na Figura 17 é possível observar as categorias que menos venderam nos dois anos em cada um dos meses do respectivo ano. No geral, a categoria que menos vendeu nos dois anos foi a CATR#410 – que nos mês de janeiro de 2018 vendeu apenas 1134 unidades.

Relativamente ao ano de 1999, mantém-se a não volatilidade do mercado, uma vez que a categoria CATR#340 foi a categoria menos vendida em todos os meses do ano, acabando o mês de dezembro com um máximo de 1693 vendas num mês.

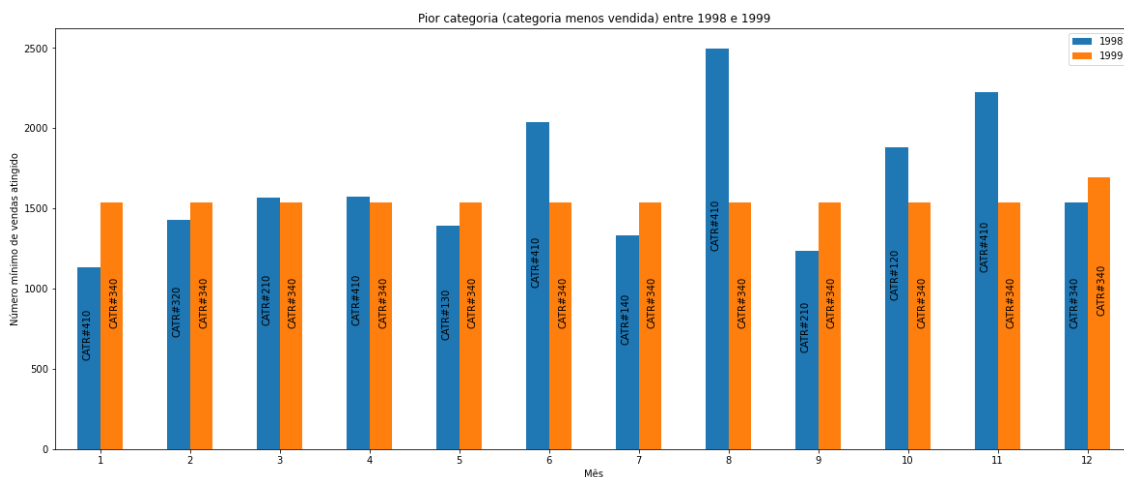


Figura 17 - Número de vendas da categoria menos vendida em cada mês nos anos de 1998 e 1999

Por fim, a Figura 17 apresenta as categorias menos vendidas em cada época do ano de 1998 e 1999. Analisando o gráfico apresentado, é possível concluir que no ano de 1999 a categoria menos vendida em todas as estações foi a CATR#410, com um valor médio de 4455 vendas em cada estação. Para além disso, nesse mesmo ano a época onde obteve o melhor resultado foi no inverno, tendo sido feitas 4477 vendas.

Em contraste, no ano de 1998 há uma volatilidade visível no mercado uma vez que, o número de vendas a categoria menos procurada, varia de época para época do ano. No entanto, o valor mais baixo foi registado no verão pela categoria CATR#410 que registou apenas 1570 vendas, tendo sido assim a pior época para esta categoria que aparece também no último lugar no inverno, e no Natal.

Trabalho Prático nº 3

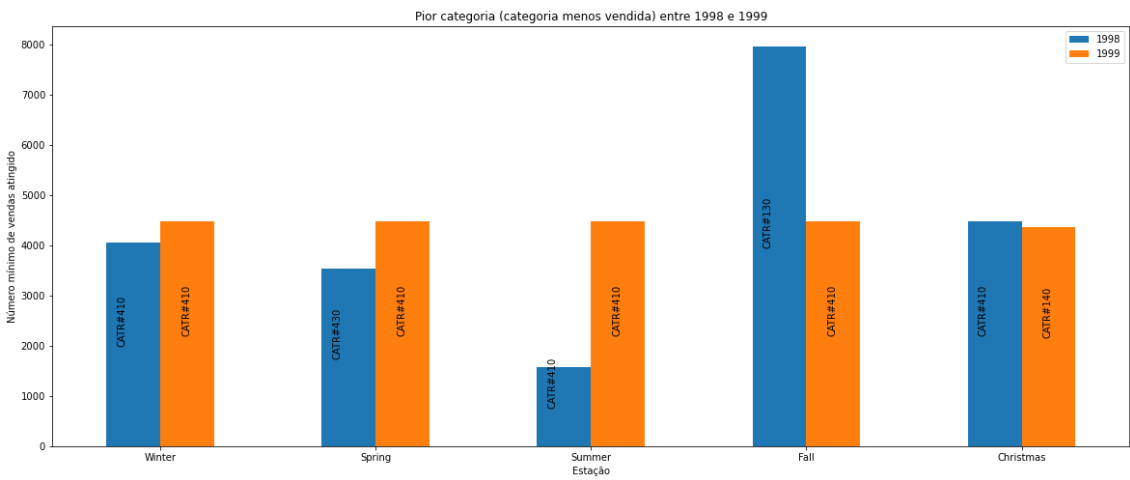


Figura 18 - Número de vendas da categoria menos vendida em cada época nos anos de 1998 e 1999

3. Clustering

Realizaram-se algumas análises de Clustering recorrendo a atributos que achamos que iriam ter mais impacto nos resultados, partindo dos resultados obtidos nas análises anteriores. Desta forma, escolhemos os meses (*month*), as estações do ano (*sellingseason*), o ano (*year*), a quantidade de produtos (*l_quantity*), o valor total da compra (*l_ordertotalprice*), o valor total de venda (*l_extendedprice*) e as categorias dos produtos (*p_category*).

Foram feitos muitos exercícios de Clustering, mas os resultados não estavam a ser de todo interessantes, pouca coisa se conseguia concluir. Com auxílio do docente da disciplina percebemos que estávamos a cometer um erro na preparação da Serie (*Dataframe*). Desta forma, começou-se por fazer agrupar os dados por 4 categorias, 3 delas temporais – *month*, *sellingseason*, *year* e *p_category*.

Depois, fizemos uma agregação dos dados (soma) de todos os restantes atributos de forma a obtermos classificações mais corretas.

Fizemos também um pequeno exercício comum em análises de Clustering para conhecermos um valor adequado para o número de cluster que devíamos escolher. Tudo indica que 3 clusters será suficiente para termos um agrupamento adequado. A escolha de 3 clusters deve-se ao facto de a média da soma dos quadrados do cluster começarem a tender para valores constantes (próximos de 0), o valor 3 está num ponto “intermédio”, isto é, é o ponto (cotovelo) no qual a média começa a alisar.

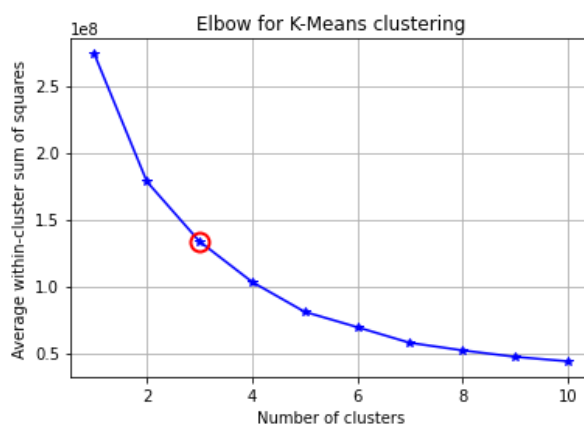


Figura 19 - Cotovelo (Elbow) para agrupamento K-Means

Após termos descoberto o número de cluster a utilizar, variaram-se alguns “random states” para obter melhores resultados, variou-se valores entre o 1000 e 9000, mas

acabamos por escolher um random state de 1426 por apresentar resultados mais direcionados para aquilo que pretendíamos.

Feita a computação para o agrupamento tivemos de escolher a nomenclatura da classificação feita pelo algoritmo, neste caso fizemos vários groupby's e determinamos algumas médias com atributos do nosso interesse para ver como o algoritmo se estava a comportar. Achamos que uma boa classificação poderia estar direcionada para a média de quantidade de produtos das encomendas – procura por parte dos clientes. Deste modo, as classificações foram:

0. Pouco procurado
1. Procurado
2. Muito Procurado

```
In [266]: df.groupby('clusteringGroups').l_quantity.sum()
Out[266]: clusteringGroups
0      99011.0
1     240500.0
2     480725.0
Name: l_quantity, dtype: float64
```

Figura 20 - Critério de escolha para nomenclatura dos grupos

3.1 Clustering – valor total de venda em função do valor total da compra

Nesta análise de Clustering tivemos resultados muito interessantes e com uma relação bastante significativa em termos de comportamento e distribuição da classificação. Podemos observar que há uma clara relação de proporcionalidade entre a quantidade de produtos (*l_quantity*) e o valor total da compra (*l_ordertotalprice*) o que faz sentido. Contudo, para a nossa surpresa, esperávamos obter uma distribuição diferente, na que os produtos muito procurados estivessem com os valores mais altos de compra.

Estes resultados fazem-nos refletir sobre o quotidiano – “procuramos produtos de qualidade que não sejam muito caros (acessíveis)”. Estes resultados representam que a maioria dos clientes procuraram muito mais produtos que não fossem tão caros (verde). Relativamente aos menos procurados (amarelo) notamos que são aqueles que têm um valor de compra mais baixo, tal como esperado por serem os menos procurados no mercado. Desta forma, podemos assumir que aqueles que estão no topo de valor de compra (vermelho) são produtos caros do quotidiano, mas que são estritamente

necessários para o dia a dia o que faz com que estejam num nível intermedio de procura e que por sua vez estejam no topo do valor de compra.

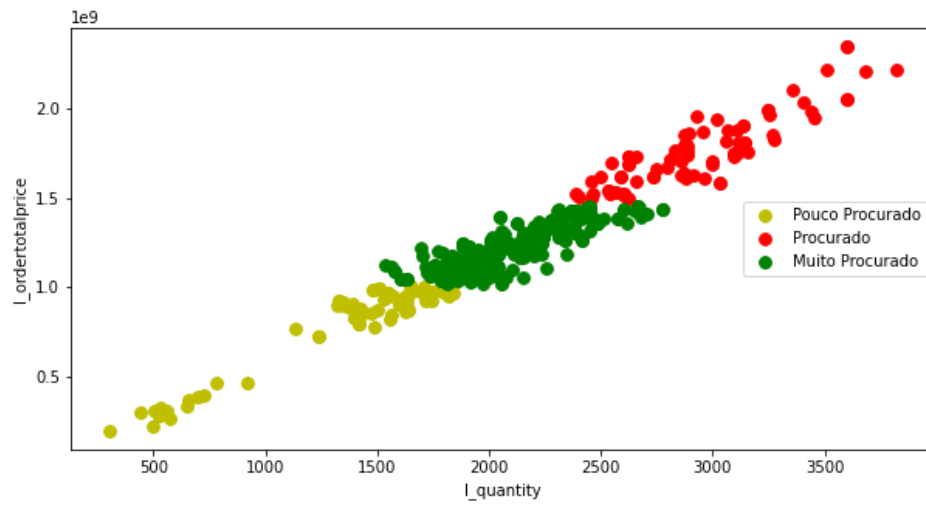


Figura 21. Clustering – valor total de venda em função da quantidade de produtos

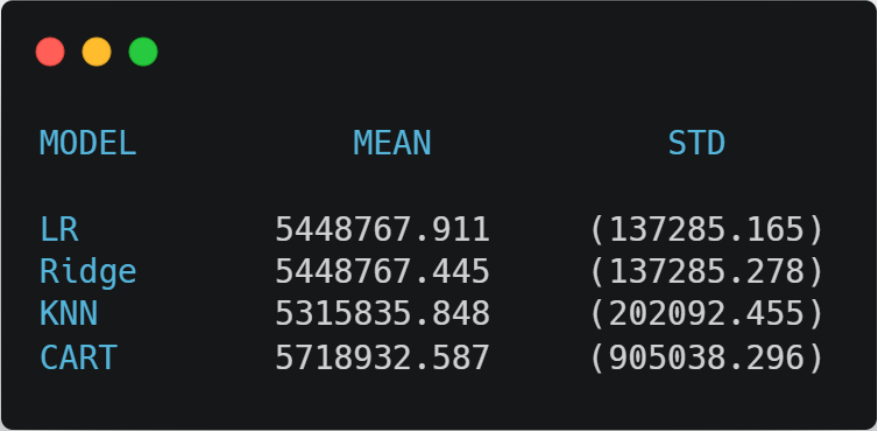
4. Regressão

Numa análise de regressão, decidimos que para prever as receitas, poderíamos basear-nos em alguns parâmetros (marcas, ano, estação do ano, países e regiões). A ideia era fazer uma avaliação entre vários algoritmos e perceber qual deles apresentava melhores resultados.

Contudo, os dados obtidos foram muito desagradantes, tanto os algoritmos lineares (*Linear Regression (LR)* e *Ridge*) como os não lineares (*K-Nearest Neighbors (KNN)* e *Classification and Regression Trees (CART)*) não apresentaram resultados bons quanto a métrica escolhida para avaliar o modelo – Perda da regressão do erro absoluto médio (*neg_mean_absolute_error*). Esperávamos resultados muito diferentes aos obtidos, na casa dos 10^4 pelo menos.

Uma possível justificação para a péssima previsão por parte dos algoritmos são os atributos escolhidos para a análise, talvez uma melhor preparação dos dados conseguia influenciar de forma benéfica a previsão. Outra possível justificação para os resultados é o facto de os dados utilizados serem fictícios e ser difícil prever situações que não têm uma relação direta.

Ignorando os resultados pouco satisfatórios e fazendo uma análise mais direccionada para os algoritmos, notamos que o melhor algoritmo para este estudo foi o KNN e o pior o CART. Por outro lado, o LR e o Ridge tiveram resultados muito semelhantes, não havendo uma distinção significativa entre ambos. Isto significa que o erro absoluto médio foi mais baixo para o KNN e o mais alto para o CART como se pode observar na Figura 22.



MODEL	MEAN	STD
LR	5448767.911	(137285.165)
Ridge	5448767.445	(137285.278)
KNN	5315835.848	(202092.455)
CART	5718932.587	(905038.296)

Figura 22. Comparação dos modelos - Média e Desvio Padrão

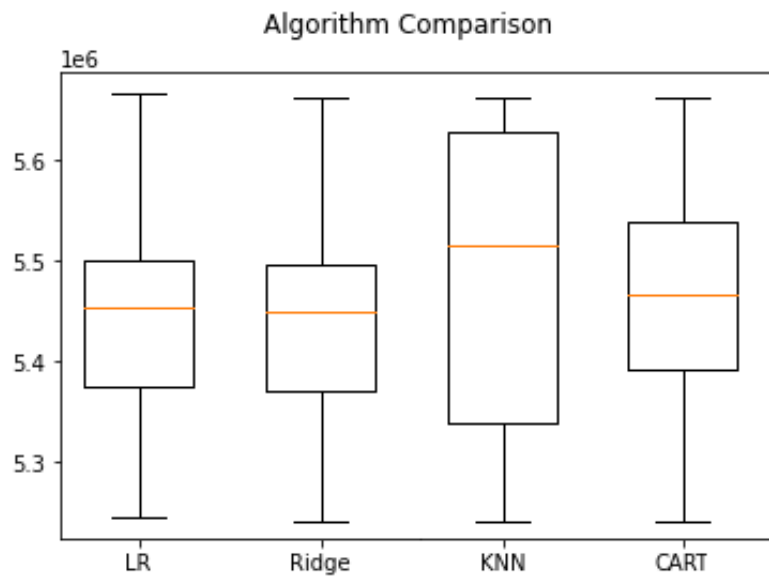


Figura 23. Comparação dos algoritmos – BoxPlot

Dada a nossa insatisfação com os resultados obtidos, tentamos fazer algo mais simples. A nossa tentativa foi prever as receitas baseando-nos em 2 parâmetros, os meses e a quantidade de produtos.

Com esta análise conseguimos obter resultados bastante satisfatórios apesar de serem pouco interessantes por serem óbvios. Na Figura 24 estão representados os valores reais agregados do dataset (pontos vermelhos) e a regressão obtida (linha azul) que consegue prever o valor da compra baseando-se na quantidade de produtos.

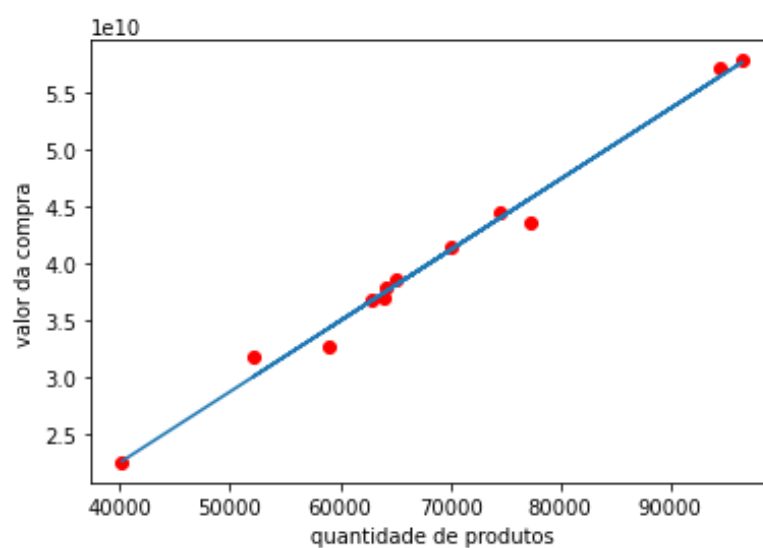


Figura 24. Regressão linear - valor da compra em função da quantidade de produtos

Trabalho Prático nº 3

Feita a representação gráfica da regressão, tentamos analisar os resultados de forma mais detalhada com algumas métricas – a média do erro dos quadrático, coeficiente de determinação (r^2), média do erro absoluto (MAE), média do erro do quadrático (MSE) e a média da raiz do erro quadrático (RMSE). Todas as métricas escolhidas são típicas em análises de regressão pelo que não poderíamos deixar de as analisar uma a uma.

Relativamente ao coeficiente de determinação, podemos observar que tem um valor quase perfeito (próximo de 1), o que não nos surpreende devido a relação óbvia entre os parâmetros escolhidos para análise.

Relativamente ao MAE parece ser um valor muito grande, contudo, estamos a considerar valores agregados e muito grandes. Desta forma é completamente aceitável este tipo de valores assim como para o MSE e o RMSE.

Baseando-nos no valor do coeficiente de determinação, não podemos afirmar que os nossos resultados são surpreendentes e inovadores, mas podemos afirmar que estamos a conseguir prever com rigor o valor da compra.

A terminal window with a dark background and three colored window control buttons (red, yellow, green) in the top left corner. It displays a table of regression metrics.

Critério	Valor
r^2	0.9896
MAE	700067616.3759
MSE	9.529187585006106e+17
RMSE	976175577.7014

Figura 25. Métricas para análise

5. Considerações Finais

Dado por terminado o projeto é necessário fazer um balanço dos objetivos alcançados, assim como as competências adquiridas.

Com as aulas práticas de introdução a análise EDA, Clustering e principalmente com o esclarecimento e ajuda do docente podemos afirmar que de uma forma geral, conseguimos fazer uma análise comparativa entre 2 datasets (2 anos) do SSB do TPC-H.

É necessário referir que, apesar dos exemplos terem sido bastante variados e úteis, faltou a experiência “na área”. Tivemos dificuldades em realizar o trabalho por falta de experiência em ciência de dados e com a linguagem de programação Python, o que provocou dispensar muito tempo a tentar que o código funcionasse e que os resultados aparecessem, assim como a posterior análise de cada um dos exercícios. A grande variedade de bibliotecas que existem facilitou, em parte, a superação das nossas dificuldades.

O trabalho prático ajudou-nos a compreender a complexidade que uma análise de dados brutos tem, assim como a preparação dos dados e representação dos resultados.

Consideramos, que as nossas soluções não estão da forma mais eficiente possível, pois como já foi referido a pouca experiência com a linguagem e área tornou-se um desafio no desenvolvimento do trabalho. Sentimos mais dificuldades em todo o processo de Clustering e de Regressão, incluindo a sua posterior análise do que a análise e representação dos resultados do EDA. Contudo, apesar das dificuldades sentidas, tentamos procurar sempre soluções ajustadas aquilo que nos é exigido e esperado no âmbito da disciplina de Sistemas de Gestão de Dados.