Daniel Amorim

PORTEFÓLIO · INICIATIVA PRÓPRIA · ANÁLISE DE DADOS

Segmentação RFM de clientes em Python

Como usei os dados de vendas de uma loja online para definir uma estratégia de marketing diferente para cada mercado internacional.

Categoria

Análise de dados e marketing

Ferramentas

Python, Pandas, Plotly, Seaborn, Jupyter, Gemini, PowerPoint

Origem

Iniciativa própria

Autor

Daniel Amorim

Contacto: daniel@damorimdigital.com

Ver mais casos: damorimdigital.com/projetos

ÍNDICE

Em resumo....3
01 O ponto de partida....3
02 Como fiz, passo a passo....4
03 Os dados....5
04 Os mercados....5
05 A segmentação RFM....6
06 Sazonalidade e produtos....8
07 A estratégia por mercado....9
08 O que esta análise não mostra....10
09 O que ficou....10

EM RESUMO

Pergunta	Uma loja online dependia demasiado do Reino Unido. Que mercados internacionais merecem mais investimento, e com que tipo de ação?
Método	Limpei os dados de vendas em Python, comparei os países por receita e ticket médio e segmentei 4 049 clientes por recência, frequência e valor (RFM) nos quatro mercados foco.
Resultado	O Reino Unido gera cerca de 8 vezes a receita dos quatro países seguintes juntos e concentra o maior número de clientes em risco. Os Países Baixos têm o ticket médio mais alto. A recomendação é retenção no Reino Unido, posicionamento premium nos Países Baixos e na Irlanda e aquisição na Alemanha.
Limite	A Irlanda tem 4 clientes e os Países Baixos 22, por isso a recomendação para esses mercados é uma pista a confirmar. Os dados cobrem um único ano.

8×
a receita do Reino Unido sobre a dos quatro países seguintes juntos

4 049
clientes segmentados por recência, frequência e valor

4 e 22
clientes na Irlanda e nos Países Baixos: uma amostra pequena

01

O ponto de partida

O dataset público Online Retail II é de uma loja online sediada no Reino Unido, cuja faturação vem sobretudo desse mercado. Depender de um só mercado é um risco, e faturação alta não quer dizer clientes mais valiosos. O ficheiro que usei tem 525 461 linhas de vendas, de 1 de dezembro de 2009 a 9 de dezembro de 2010.

Quis responder a uma pergunta de marketing: que mercados internacionais merecem mais investimento, e com que tipo de ação. Para isso precisava de duas leituras: o valor de cada país e a qualidade da base de clientes de cada um.

Termos usados neste documento

- **RFM:** segmentação de clientes por três critérios: recência, frequência e valor monetário
- **Recência:** os dias desde a última compra; quanto menor, melhor
- **Frequência:** o número de faturas do cliente; quanto maior, melhor
- **Valor monetário:** a receita total gerada pelo cliente; quanto maior, melhor
- **Ticket médio:** a receita a dividir pelo número de faturas, ou seja, o valor de uma compra típica
- **Score RFM:** a soma das pontuações de recência, frequência e valor, de 3 (pior) a 15 (melhor)

02

Como fiz, passo a passo

Segui sete passos, da limpeza dos dados à apresentação. As escolhas de método e as recomendações finais foram minhas.

1. **Preparei os dados num notebook Jupyter com Pandas.** Cerca de 20 % das linhas não tinham cliente identificado e havia quantidades negativas, que correspondem a devoluções. Fiquei só com vendas reais de clientes identificados, porque a segmentação depende de saber quem comprou.
2. **Comparei os países por receita e ticket médio.** O Reino Unido gerava cerca de 8 vezes a receita dos quatro países seguintes juntos. O ticket médio mostrou outra coisa: onde cada compra vale mais.
3. **Desenhei uma matriz de mercados com Plotly.** Cruzei receita e ticket médio num gráfico em quadrantes, que separa os mercados de volume (Reino Unido), os de valor (Países Baixos e Irlanda) e os de potencial por explorar (Alemanha e França).
4. **Segmentei os clientes por RFM em 4 mercados.** Avaliei cada cliente pela recência, pela frequência e pelo valor das compras, com uma pontuação de 1 a 5 em cada critério. Com a soma, agrupei-os em 5 segmentos, dos melhores clientes aos adormecidos, que é a base para decidir a quem falar e com que mensagem.
5. **Validei o código e a leitura dos resultados com o Gemini.** Usei-o para rever e corrigir linhas de código e para discutir a interpretação dos segmentos. As conclusões e as recomendações finais foram minhas.
6. **Cruzei a sazonalidade com os produtos.** A receita concentra-se em outubro e novembro, puxada pelo Reino Unido. Procurei os produtos que lideram ao mesmo tempo em receita e em frequência de venda, os candidatos naturais a campanhas nessa época.
7. **Apresentei a estratégia em PowerPoint e publiquei o notebook.** Os 10 slides traduzem a análise em ações por mercado. O notebook está no GitHub e no Binder, onde abre no navegador sem instalar nada.

03

Os dados

A tabela 1 mostra quantas linhas sobram em cada passo da limpeza. Para a ler, comece pelo total do ficheiro e siga as linhas para baixo: cada uma resulta de aplicar um filtro à anterior.

Passo	Linhas
Linhas no ficheiro	525 461
Linhas sem cliente identificado (retiradas)	107 927 (20,5 %)
Vendas de clientes identificados, só com receita positiva	407 664
Só produtos comerciais, sem portes nem testes	406 491

Tabela 1. Da base completa à base usada nas análises.

Os valores monetários estão em libras esterlinas, a moeda do dataset.

04

Os mercados

Comecei pela receita de cada país. A figura 1 mostra uma concentração forte no mercado de origem: o Reino Unido gerou 7 360 539 £, contra 920 992 £ dos quatro países seguintes juntos, o que dá cerca de 8 vezes (7,99). Qualquer instabilidade económica ou mudança na concorrência nesse mercado pesaria muito na empresa.

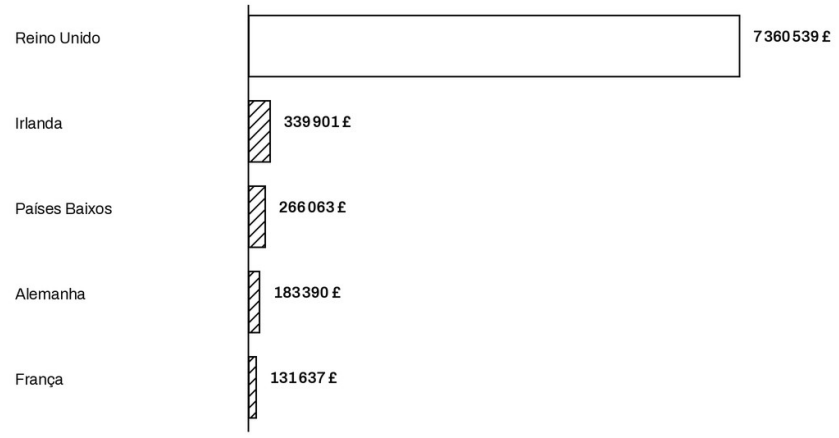


Figura 1. Receita dos cinco países com mais vendas (sem portes nem testes).

A receita, porém, não diz onde cada compra vale mais. Calculei o ticket médio de cada país, a receita a dividir pelo número de faturas. A tabela 2 junta as duas leituras e a figura 2 mostra o ticket médio. Os Países Baixos têm o ticket mais alto, 1 991 £, e a Irlanda o segundo, 1 127 £. O Reino Unido, que domina a receita, tem o mais baixo, 421 £.

País	Receita (£)	Faturas	Ticket médio (£)	Leitura
Reino Unido	7 414 756	17 612	421	Volume
Irlanda	356 085	316	1 127	Valor
Países Baixos	268 786	135	1 991	Valor
Alemanha	202 395	347	583	Potencial
França	146 215	236	620	Potencial

Tabela 2. Receita, faturas e ticket médio. Inclui todas as vendas de clientes identificados, por isso difere da figura 1, que exclui portes e testes.

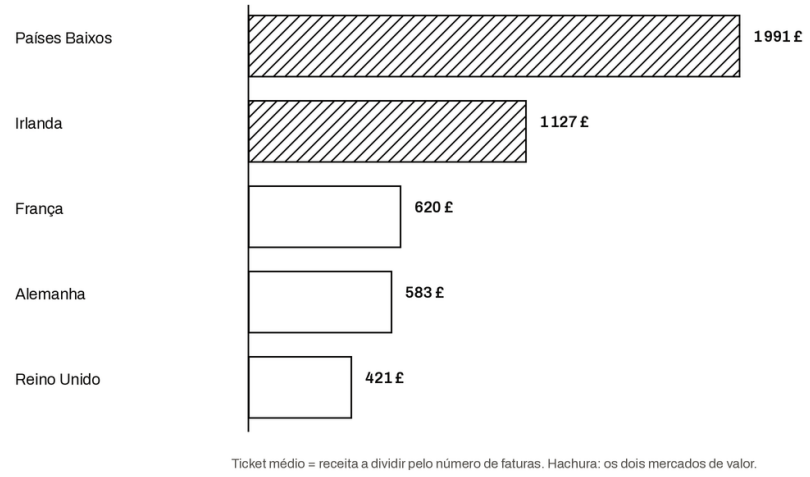


Figura 2. Ticket médio por país.

Cruzar as duas medidas dá os quadrantes da matriz de mercados. O Reino Unido é o mercado de volume. A Irlanda e os Países Baixos são mercados de valor: pouco volume de pedidos, mas compras grandes. A Alemanha e a França combinam receita baixa com ticket abaixo da média da matriz, e por isso ficam como mercados de potencial por explorar.

Há países com ticket ainda mais alto, como a Noruega (2 177 £, em 11 faturas) ou Israel (1 600 £, em apenas 2). Ficaram fora da leitura porque o volume é demasiado pequeno para sustentar uma decisão.

05

A segmentação RFM

A segmentação RFM classifica os clientes pelos hábitos de compra: quem comprou há pouco, quem compra muitas vezes e quem gasta mais. Aplicou-se aos quatro mercados foco: Reino Unido (volume), Irlanda e Países Baixos (valor) e Alemanha (potencial). Cada cliente foi associado ao país onde mais comprou, e a data de referência foi o dia seguinte à última venda do ficheiro.

A tabela 3 resume as estatísticas dos 4 049 clientes. Leia a diferença entre a média e a mediana: na frequência, a maioria comprou 2 vezes ou menos, e no valor, a média (2 013 £) é cerca de três vezes a mediana (679 £). Poucos clientes de valor muito alto puxam a média para cima.

Métrica	Média	Mediana	3.º quartil	Máximo
Recência (dias)	91	52	136	374
Frequência (faturas)	4,5	2	5	188
Valor monetário (£)	2 013	679	1 674	349 164

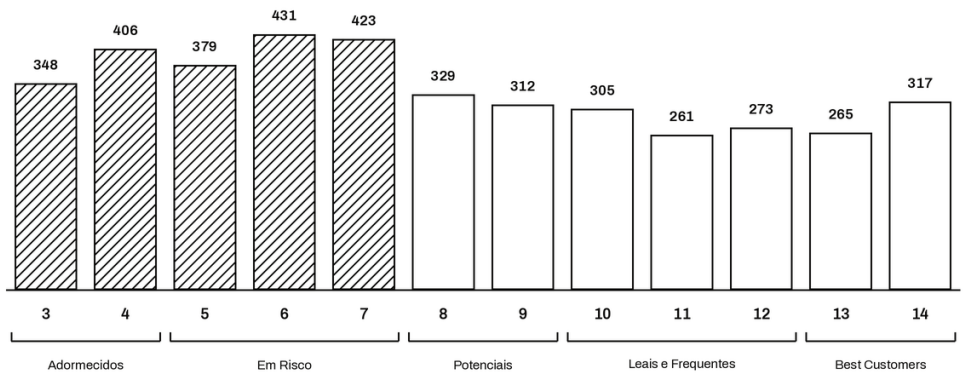
Tabela 3. Estatísticas das métricas RFM, 4 049 clientes.

Cada métrica foi dividida em cinco grupos iguais (quintis), com pontuação de 1 a 5. Na recência a escala é invertida, para que 5 seja a compra mais recente. A soma dos três scores dá o score RFM, e a figura 3 mostra quantos clientes há em cada valor.

```
# Score de Recência (invertido: 5 = compra mais recente)
rfm_interest['R_score'] = pd.qcut(
    rfm_interest['Recency'], q=5, duplicates='drop', labels=False) + 1
rfm_interest['R_score'] = 6 - rfm_interest['R_score']

# Score RFM final: soma simples dos três scores
rfm_interest['RFM_Score'] = (
    rfm_interest['R_score'] + rfm_interest['F_score'] + rfm_interest['M_score'])
```

Excerto 1. Atribuição dos scores por quintis e soma final (excerto simplificado).



Score RFM (soma de R, F e M) e número de clientes. Hachura: Adormecidos e Em Risco.

Figura 3. Clientes por score RFM e segmento.

Os cortes dos segmentos são fixos: 13 ou mais, Best Customers; de 10 a 12, Leais e Frequentes; 8 e 9, Potenciais; de 5 a 7, Em Risco; 3 e 4, Adormecidos. Chamei clientes de alta prioridade aos três primeiros segmentos. A tabela 4 mostra os segmentos em cada país e a figura 4 a percentagem de alta prioridade.

País	Best	Leais e Freq.	Potenciais	Em Risco	Adormec.	Total
Irlanda	3	0	0	0	1	4
Países Baixos	1	3	5	9	4	22
Alemanha	15	14	9	24	5	67
Reino Unido	563	822	627	1 200	744	3 956

Tabela 4. Número de clientes por segmento e país.

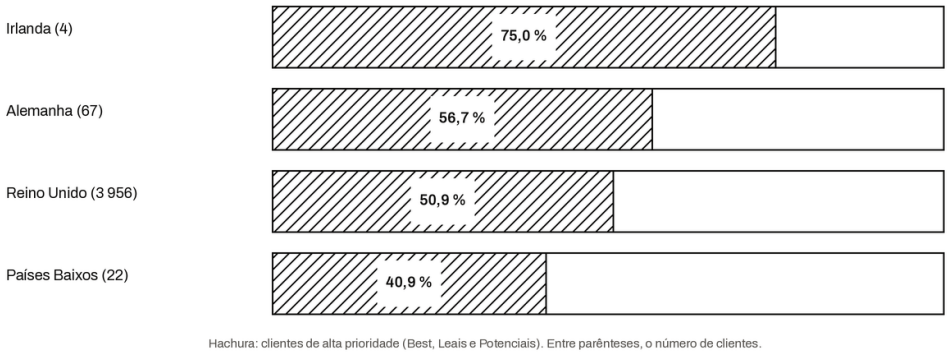


Figura 4. Percentagem de clientes de alta prioridade, por país.

No Reino Unido, metade dos clientes (1 944 em 3 956) está em risco ou adormecida, e é onde o número absoluto é maior. Na Irlanda e nos Países Baixos a percentagem é calculada sobre 4 e 22 clientes, e por isso não sustenta uma conclusão firme.

06

Sazonalidade e produtos

A figura 5 mostra a receita mensal dos quatro países foco. Há uma subida clara a partir de setembro, com o máximo em novembro (1 073 954 £), e dezembro de 2010 aparece baixo apenas porque o ficheiro só vai até ao dia 9. O Reino Unido e a Alemanha têm o pico em novembro, os Países Baixos em setembro e a Irlanda em janeiro: a concentração em outubro e novembro é puxada pelo Reino Unido.

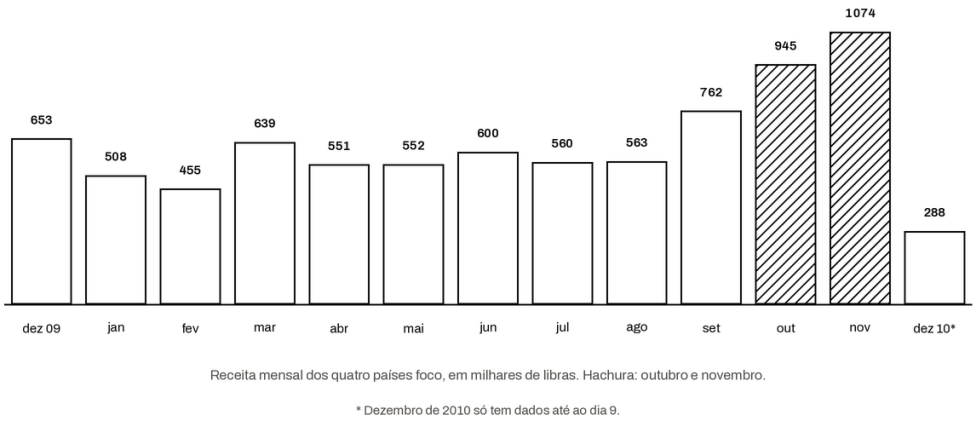


Figura 5. Receita mensal dos quatro países foco, de dezembro de 2009 a dezembro de 2010.

Nos produtos, cruzei a receita com a frequência de venda, o número de linhas de fatura em que o produto aparece. A tabela 5 mostra os que se destacam. O WHITE HANGING HEART T-LIGHT HOLDER lidera as duas listas e o REGENCY CAKESTAND 3 TIER é o segundo em ambas. O JUMBO BAG RED RETROSPOT está entre os que mais faturam, mas vende em menos linhas do que os produtos de ticket baixo.

Produto	Receita (£)	Posição (receita)	Linhas de venda	Posição (frequência)
WHITE HANGING HEART T-LIGHT HOLDER	151 624	1.º	3 153	1.º
REGENCY CAKESTAND 3 TIER	143 893	2.º	1 705	2.º
JUMBO BAG RED RETROSPOT	51 759	4.º	1 084	9.º
STRAWBERRY CERAMIC TRINKET BOX	29 780	13.º	1 407	3.º
HOME BUILDING BLOCK WORD	28 207	16.º	1 219	5.º

Tabela 5. Produtos que aparecem nas listas dos 20 mais vendidos em receita e em frequência.

07

A estratégia por mercado

A tabela 6 junta o que os dados dizem de cada mercado com a ação que propus. As recomendações são propostas, não resultados medidos.

Mercado	Posição	Objetivo	O que os dados dizem
Reino Unido	Volume	Retenção e reativação	1 200 clientes em risco e 744 adormecidos, em 3 956
Países Baixos e Irlanda	Valor	Posicionamento premium e mais frequência	Ticket de 1 991 £ e 1 127 £, mas só 22 e 4 clientes
Alemanha	Potencial	Aquisição de clientes	202 395 £ de receita e 67 clientes; ticket de 583 £

Tabela 6. Posição, objetivo e evidência por mercado.

Produtos e conteúdo

- Os produtos prioritários são o WHITE HANGING HEART T-LIGHT HOLDER e o JUMBO BAG RED RETROSPOT, que lideram receita ou frequência, e merecem prioridade de stock e de promoção.
- Países Baixos e Irlanda: comunicação de valor e exclusividade, com destaque para os produtos de ticket mais alto.
- Reino Unido: foco na recorrência, com produtos de ticket mais baixo e conteúdos rápidos para reativação.

Tráfego pago

- Países Baixos e Irlanda: campanhas para adquirir clientes de valor, com públicos semelhantes aos melhores clientes.
- Reino Unido: reativação dos segmentos inativos, com campanhas de baixo custo por conversão e ofertas para estimular a recompra.

Calendário

- Setembro e outubro: antecipar o investimento em anúncios e a organização do stock, para estar presente antes do pico de novembro.
- Início do ano: fidelizar os clientes de alta prioridade, quando o volume e a concorrência são menores.

08

O que esta análise não mostra

Os números respondem à pergunta de partida, mas têm limites que convém dizer:

- A Irlanda tem 4 clientes e os Países Baixos 22. As percentagens desses mercados mudam muito com um só cliente, por isso a recomendação de posicionamento premium é uma pista a confirmar com mais dados.
- Os dados cobrem um único ano, de dezembro de 2009 a dezembro de 2010, e dezembro de 2010 está incompleto. Com um ano só, não se separa a sazonalidade da tendência.
- O ticket médio é receita por fatura, não margem: o dataset não tem custos.
- O dataset não distingue clientes empresariais de particulares, o que pode inflacionar o ticket médio de alguns mercados.
- Os segmentos dependem de cortes fixos no score RFM. Cortes diferentes dariam percentagens diferentes.

09

O que ficou

A recomendação foi um marketing diferente por mercado: retenção e reativação no Reino Unido, que concentra o maior número de clientes em risco; posicionamento premium nos Países Baixos e na Irlanda, onde o ticket médio é mais alto; e aquisição na Alemanha. A Irlanda e os Países Baixos têm poucos clientes no dataset, por isso essa recomendação é uma pista a confirmar com mais dados.

O notebook completo está no Binder, que demora 1 a 3 minutos a abrir.

O QUE APRENDI

Aprendi que uma recomendação por mercado tem de dizer quantos clientes a sustentam: uma percentagem calculada sobre 4 clientes é só uma pista.