Resolução da Prova Prática

Candidata: Natália Augusto Keles

Este documento apresenta as repostas e justificativas solicitadas da Prova Prática para a vaga de cientista de dados da Cinnecta. O desafio consiste em fazer a análise exploratória dos dados “Retail Online.xlsx” disponibilizados pela empresa. O código fonte para responder as questões abaixo foi desenvolvido Python.

Questões do roteiro

- **Quantas linhas possui o conjunto de dados?**
  O dataframe analisado possui 541909 linhas e 8 colunas.

- **Quais são os nomes e os tipos de cada atributo do conjunto de dados?**

Atributos	Tipo
InvoiceNo	Object
StockCode	Object
Description	Object
Quantity	Int64
InvoiceDate	Datetime64[ns]
UnitPrice	Float64
CustomerID	Float64
Country	Object

Tabela 1 Atributos e Tipo das variáveis do dataframe.

- **Quantos produtos distintos existem?**
  Existem 4070 produtos distintos

- **Quantas linhas possuem a descrição com produto ausente ou nula? Após verificar, remova essas linhas do conjunto de dados**.
  Existem 1454 linhas com descrição ausente. Verificar no código fonte a maneira que foi utilizada para excluir esses atributos.

- **Crie um novo atributo de preço total que seja o resultado da multiplicação da quantidade pelo preço unitário.**
  Verificar no código fonte a maneira que foi feita a criação dessa nova variável.

- **Quantos objetos do conjunto de dados possuem o valor do atributo ‘*UnitPrice’* menor que zero? Apague esses objetos do conjunto de dados**.
  Existem dois produtos em que o preço unitário é menor que zero. Verifique no código fonte a maneira que foi utilizada para excluir esses valores.

- **Na sua opinião, quais são os melhores clientes dessa empresa? Justifique sua resposta.**

  O melhor cliente dessa empresa é o cliente com CustomerID = 14646, pois ele é o cliente com maior gasto no período e ao mesmo tempo é o cliente que compra mais quantidade de produtos.

  Considerando uma lista dos 10 melhores clientes por código podemos citar os clientes 14646, 18102, 17450, 16446, 14911, 12415, 14156, 17551, 16029, 12346, pois eles que contribuíram mais com o faturamento da empresa durante esse período.

- **Qual o produto mais caro vendido?**
  Considerando que a base está correta e todos os produtos válidos, o produto mais caro é o AmazonFee custando 13541,33.

- **Quais os países (Country) que fazem parte deste conjunto de dados?**
  Países que fazem parte desse conjunto de dados (nomes em inglês)

    - *United Kingdom;*
    - *France;*
    - *Australia;*
    - *Netherlands;*
    - *Germany;*
    - *Norway;*
    - *EIRE;*
    - *Switzerland';*
    - *Spain;*
    - *Poland;*
    - *Portugal;*
    - *Italy;*
    - *Belgium;*
    - *Lithuania;*
    - *Japan;*
    - *Iceland;*
    - *Channel Islands;*
    - *Denmark;*
    - *Cyprus;*
    - *Sweden;*
    - *Finland;*
    - *Austria;*
    - *Bahrain;*
    - *Israel;*

- *Greece*
- *Hong Kong;*
- *Singapore;*
- *Lebanon;*
- *United Arab Emirates;*
- *Saudi Arabia;*
- *Czech Republic;*
- *Canada;*
- *Unspecified;*
- *Brazil;*
- *USA;*
- *European Community;*
- *Malta;*
- *RSA.*

- **Qual a compra que possui mais itens?**
  O número de fatura com mais itens ('Quantity') comprados é com a InvoiceNo= 581483 com 80995 itens.

- **Qual a compra (InvoiceNo) que tem o maior valor de compra?**
  O número de fatura com maior valor de compra ('TotalPrice') comprados é com a InvoiceNo=581483 com valor total de 168469.60.

- **Faça um gráfico de barras mostrando a quantidade vendida total dos 10 p rimeiros produtos do conjunto**

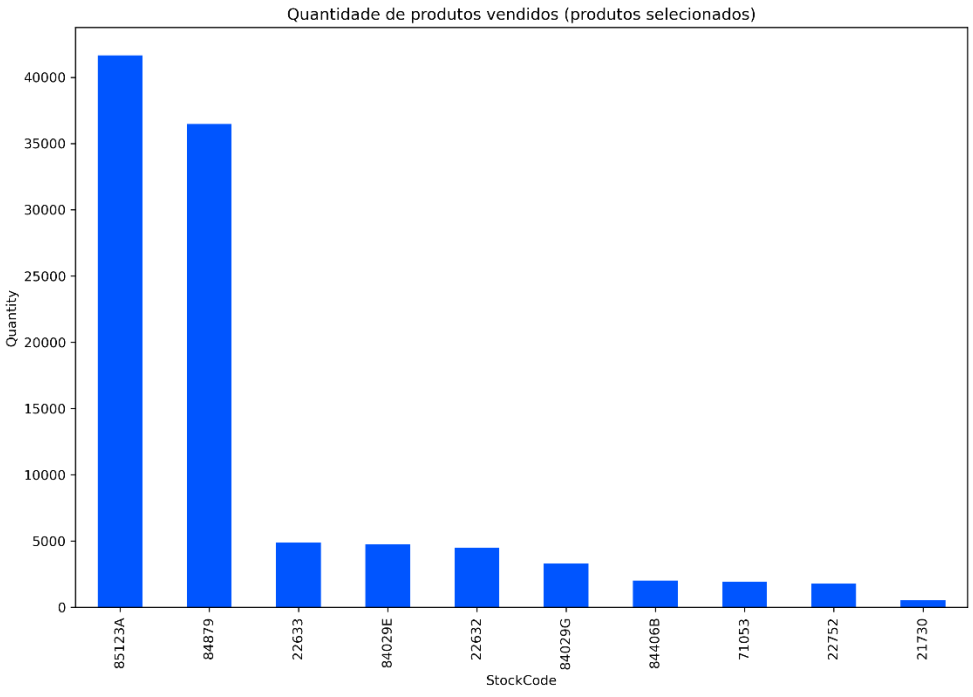

Figura 1 Quantidade de produtos vendidos dos 10 primeiros produtos do conjunto

- **Faça um Boxplot dos preços unitários de todos os produtos. Considere somente os 100 produtos mais vendidos**

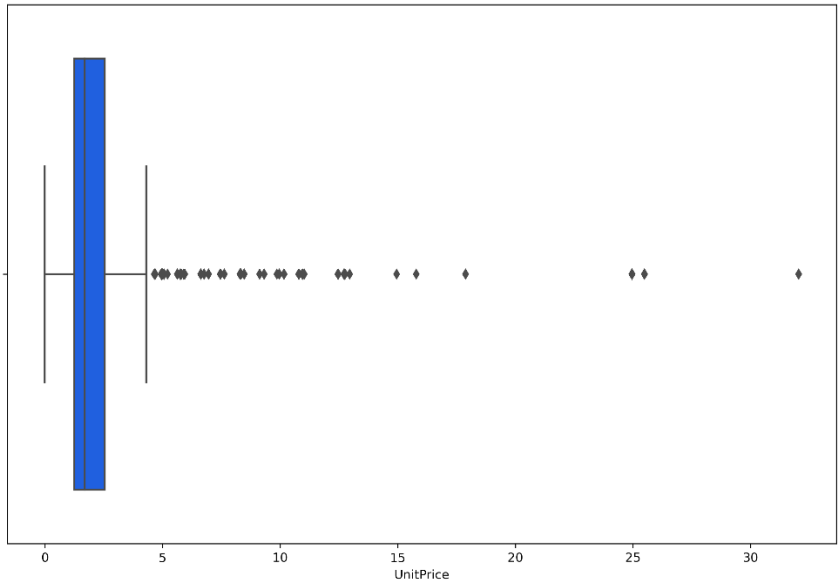

Figura 2 Boxplot dos preços unitários dos 100 produtos mais vendidos

- **Faça um gráfico de linhas mostrando o faturamento (total de vendas) por dia.**

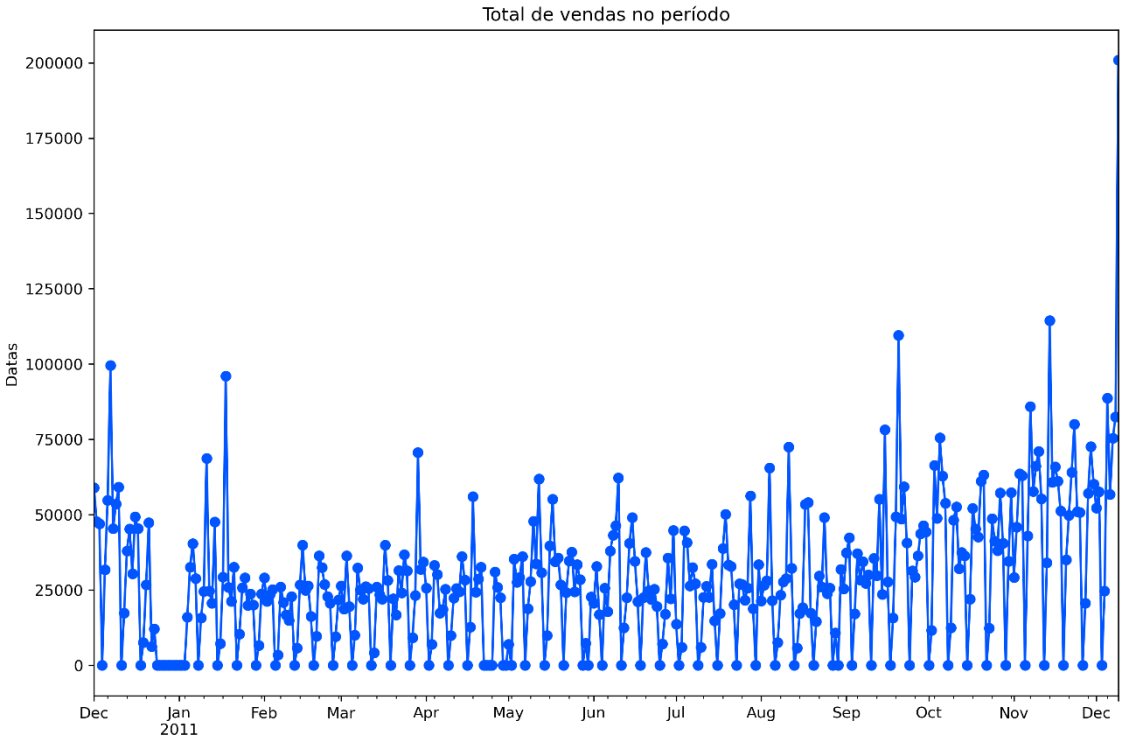

Figura 3 Total de faturamento por dia

Foram inseridos dias que não existiam no dataframe para que seja possível verificar de maneira mais clara os dias que não possuem vendas.

- **Faça o histograma dos preços unitários dos produtos.**
  No boxplot de preços unitários podemos verificar que existem muitos outliers.

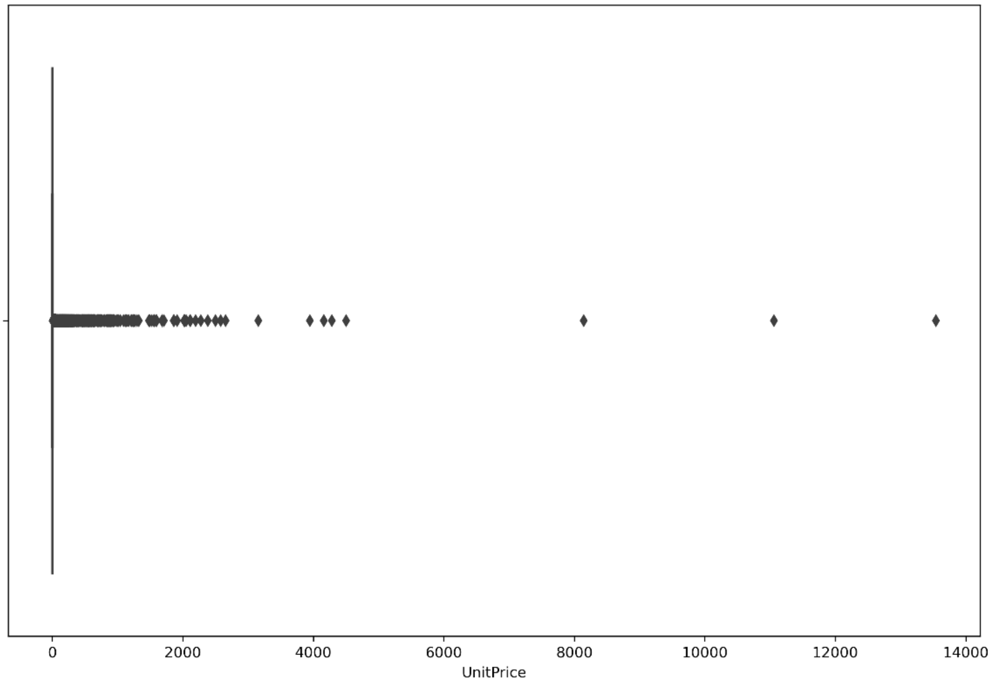

Figura 4 Boxplot mostrando a quantidade de outliers existentes nesse atributo.

Por essa razão a visualização com o histograma fica comprometida conforme mostrado na figura abaixo

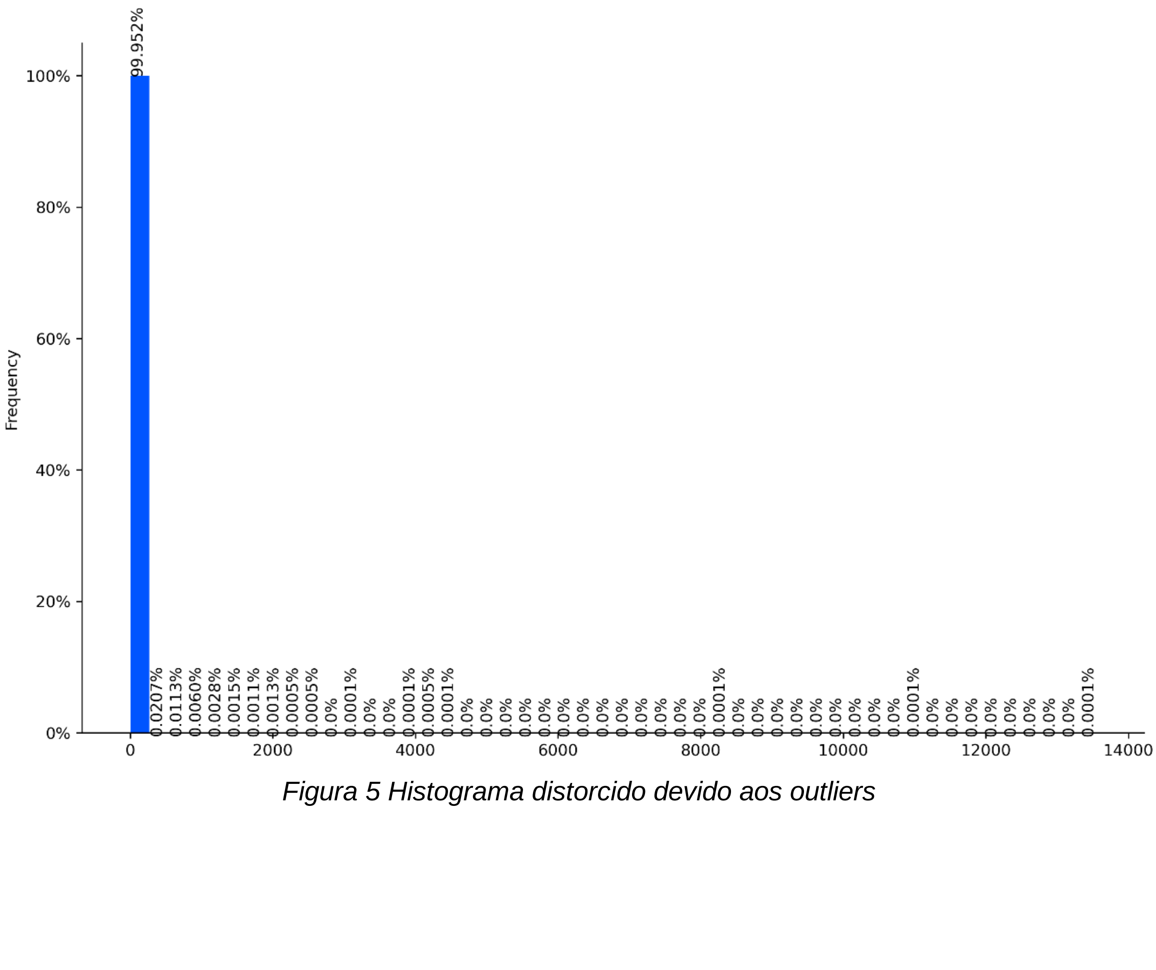

Figura 5 Histograma distorcido devido aos outliers

Por conter valores de média e mediana muito baixos e 75% dos dados serem menores do que 4.13, foi escolhido fazer o histograma com os produtos de preços unitários menores que 10. Resultando no histograma abaixo

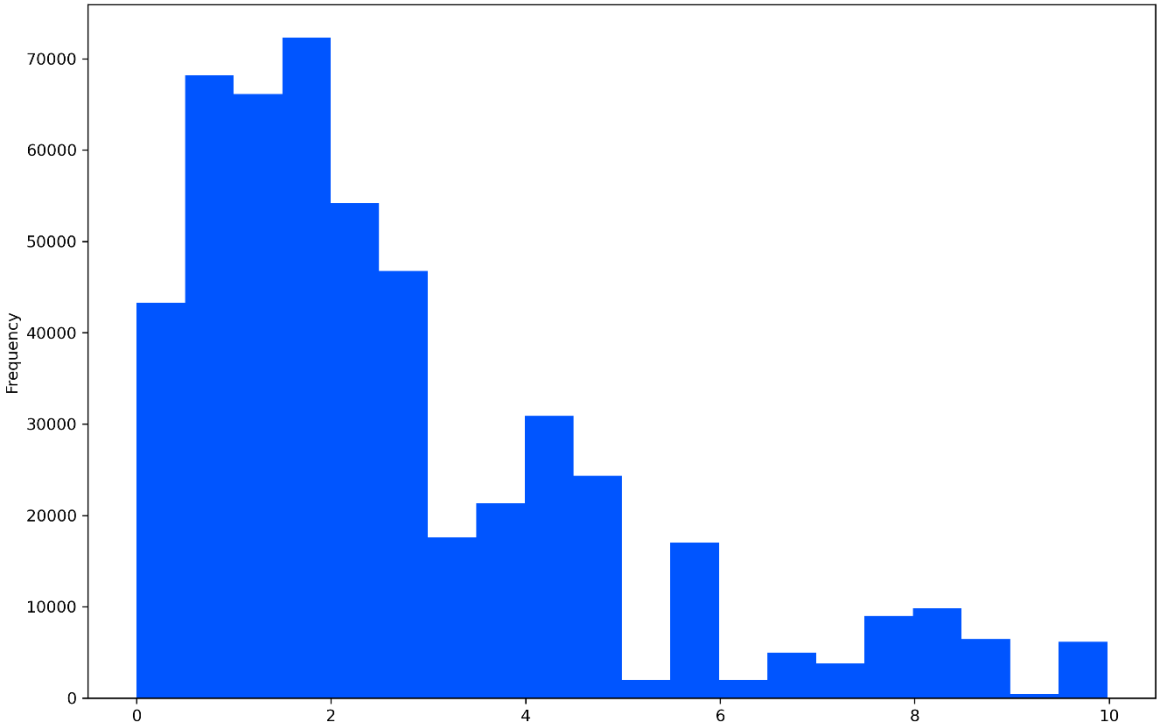

Figura 6 Histograma considerando os preços unitários menores que 10.

- **Faça dois gráficos de barras da quantidade de compras e do total de faturamento por país.**

O Reino Unido domina na categoria de quantidade de compras e na quantidade total do faturamento. Por essa razão foi escolhido plotar gráficos incluindo o Reino Unido, e também gráficos excluindo o Reino Unido para termos uma melhor visualização do comportamento dos outros países.

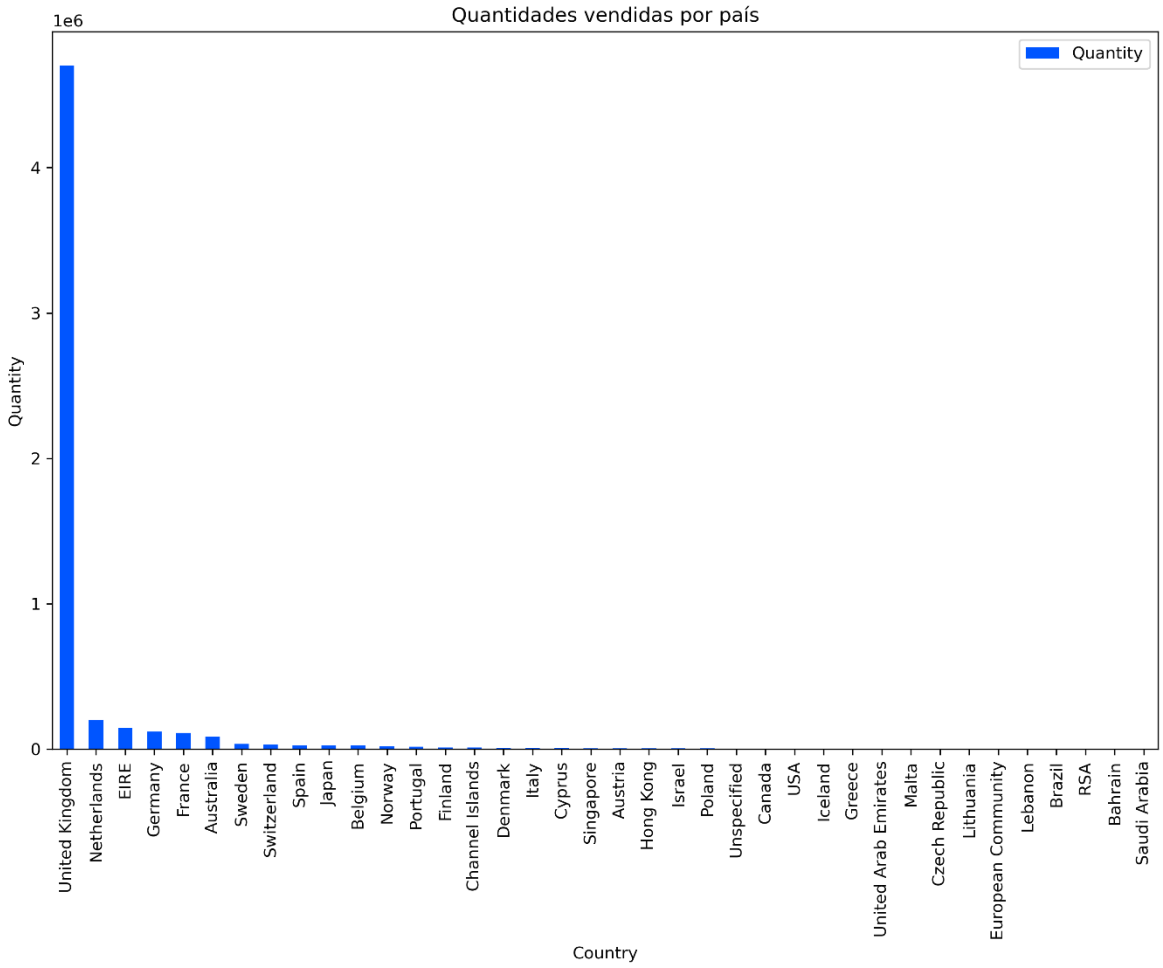

Figura 7 Quantidade vendidas por país (incluindo o Reino Unido)

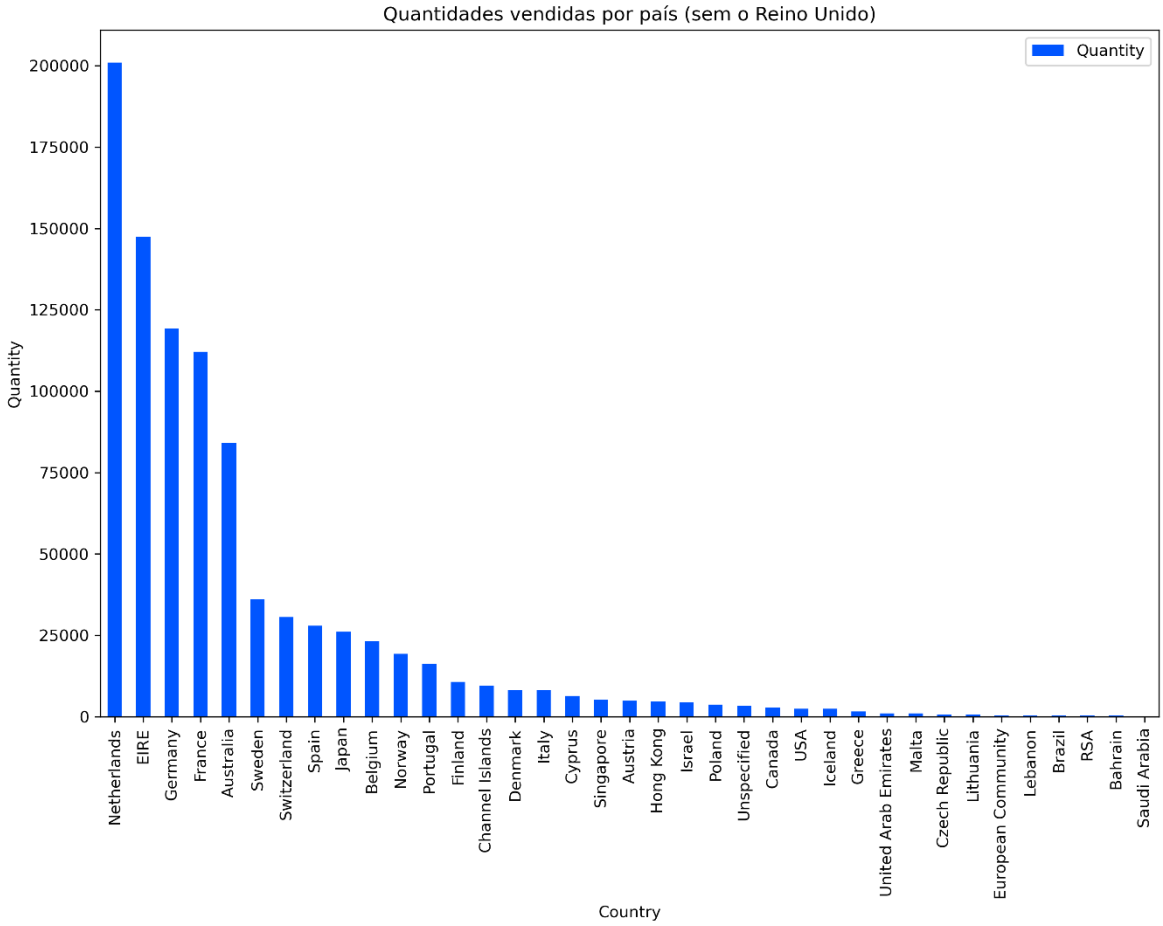

Figura 8 Quantidade vendidas por país (excluindo o Reino Unido).

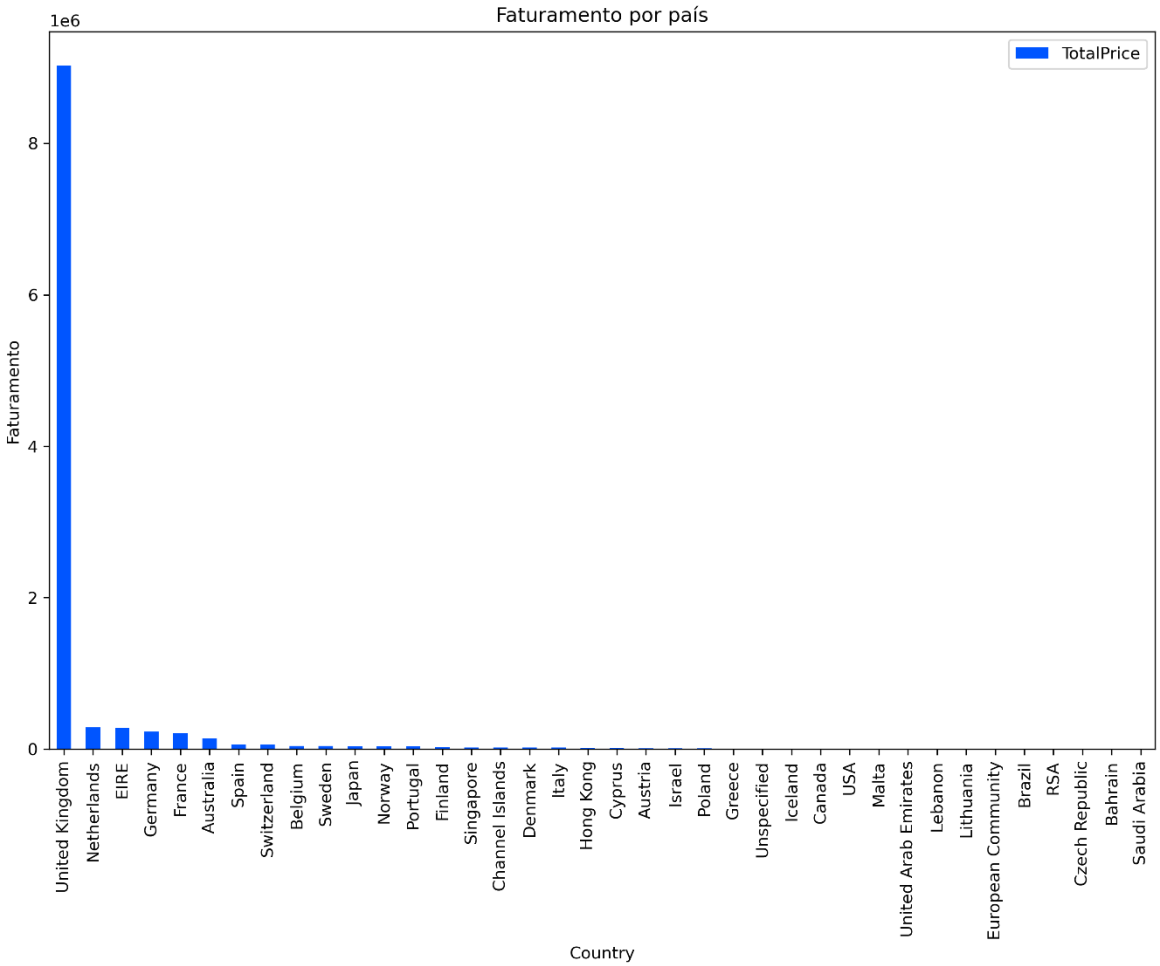

Figura 9 Faturamento por país (incluindo o Reino Unido)

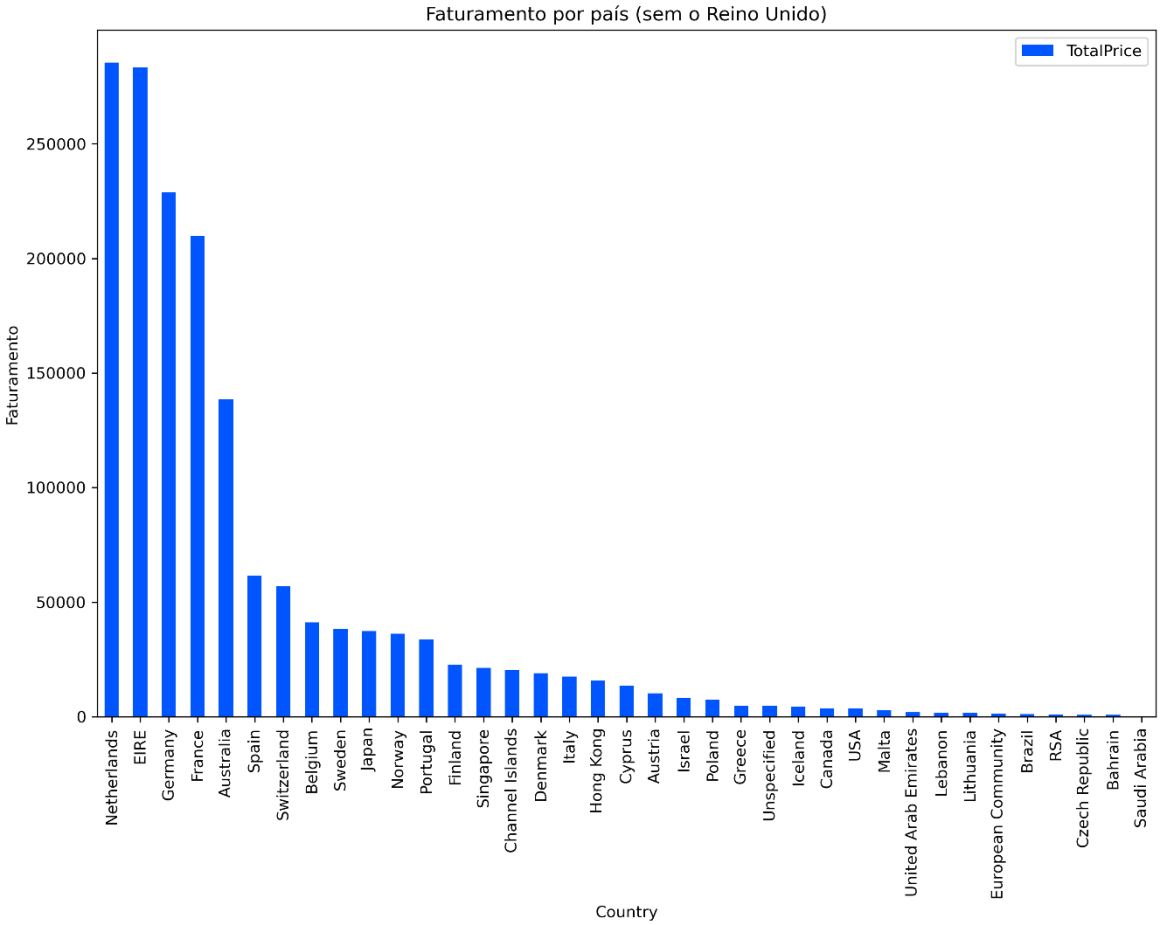

Figura 10 Faturamento por país (excluindo o Reino Unido)