Alunos: Igor Silva, André Leal

Professor: Prof. Dr. João Paulo

Tarefa de Teoria do Aprendizado Estatístico

12 de agosto de 2025

BASE DE DADOS

A escolha do conjunto de dados (encontrado aqui) foi pautada na diversidade dos dados e na aplicabilidade dos dados, uma aplicação extraída da realidade e não dados hipotéticos e reais. Nesse conjunto, encontramos data no formato AAAAMM; valor e quantidade no formato inteiro; e uma série de outros dados no formato texto que requererão formas diversas para trabalhar esses dados

1. Qual foi o método utilizado para a coleta dos dados?

   R = Está estastísticas geradas pelo próprio banco central.

2. A base já foi empregada em trabalhos acadêmicos ou pesquisas científicas?
   R = O artigo *Instant Payments and Banks: the Impact of Pix on Bank Branches in Brazil* de Leal A. e Haase é um exemplo de pesquisa científica que utiliza a base de dados contendo informações do banco para analisar o impacto que surgimento do Pix teve para a forma de se fazer pagamentos.
3. A base possui alguma certificação oficial ou validação reconhecida?
   R = Open Data Commons Open Database License (ODbL).
4. Em sua opinião, a base selecionada é confiável? Justifique sua resposta
   R = Sim. Esse banco é elaborado e disponibilizado pelo banco central com estatísticas real obtidas pelas movimentações financeiras por meio do método de pagamento Pix, além disso, ele também já foi usado em outros estudos consagrados, fortalecendo a confiabilidade no banco de dados.

VARIÁVEIS

O banco de dados do Banco do Brasil forneceu um dicionário com valores de “inteiro”, “texto” e “decimal”, sendo estes convertidos respectivamente para “numeric”, “character” e “numeric”.

Há a possibilidade de conversão para *factor* de algumas variáveis, o que poderia ser útil para o armazenamento do *dataset* na memória RAM.

Abaixo segue os dicionários dos dados com os tipos de dados equivalentes em

Tabela 1 - Variáveis e sua descrição[1]

Nome	Tipo	Descrição	Unidade
anomes	Data (AAAAMM)	Data-base - ano/mês	adimensional
pagpfpj	qualitativa nominal	PF= Pessoa Física PJ= Pessoa Jurídica	adimensional
recpfpj	qualitativa nominal	PF= Pessoa Física PJ= Pessoa Jurídica	adimensional
pagregiao	qualitativa nominal	Região do domicílio do usuário pagador.	adimensional
recregiao	qualitativa nominal	Região do domicílio do usuário recebedor.	adimensional
pagidade	quantitativas contínuas	Idade em anos do usuário pagador	adimensional
recidade	quantitativas contínuas	Idade em anos do usuário recebedor	adimensional
formainiciacao	qualitativa nominal	Forma de iniciação das transações: iniciador com todas as informações do recebedor (INIC), QR Code estático (QRES), QR Code dinâmico (QRDN), inserção manual (MANU) e chave Pix (DICT).	adimensional
natureza	qualitativa nominais	P2P - Pessoa para Pessoa, B2B - Empresa para Empresa, P2B - Pessoa para Empresa, B2P - Empresa para Pessoa, P2G - Pessoa para Governo, B2G - Empresa para Governo	adimensional
finalidade	qualitativa nominal	Finalidade da transação Pix: transferência, saque ou troco.	adimensional
valor	quantitativa discreta	Volume financeiro de transações Pix liquidadas mensalmente	R$
quantidade	quantitativa discreta	Quantidade de transações Pix liquidadas mensalmente	adimensional

TRANSFORMAÇÕES

Observação: os códigos aqui contidos tiveram suas transformações feitas de duas formas diferentes: utilizando-se a biblioteca *Tidyverse* (e *janitor*) no R-Studio e, também, utilizando-se as funções bases de R, no Jupyter Notebook pelo Google Colab. Por questão de brevidade as transformações aqui citadas se referem ao código utilizando R base.

[1] Dicionário do Banco Central encontrável em: https://dadosabertos.bcb.gov.br/dataset/pix

As transformações feitas foram as seguintes:

1. Conversão de todos os dados para minúsculo, incluindo nome de colunas *(lapply + tolower);*
2. Valores como "nao informado" ou "nao disponivel" (exceto para a coluna *formainiciacao*) foram substituídos por NA *(ifelse);*
3. Retirou-se os underlines "_" dos títulos das colunas *(gsub);*
4. As colunas *pagidade* e *recidade* eram *strings* contendo faixa-etária e foram convertidas para uma categoria que representa essas faixas *(as.factor + ifelse);*
5. Modificou-se a coluna *anomes* para *factor;*
6. Conversão das colunas *valor* e *quantidade* para valores inteiros (*as.integer*) e reais (*as.numeric*);
7. As demais colunas não citadas foram convertidas para *factor*;
8. Modificou-se os *levels* da coluna *"pagregiao'* para siglas;
9. Criação da variável "pixmedio" correspondendo a divisão da "valor" por "quantidade" e correspondendo ao valor médio de um único pix para aquele PF/PJ naquele mês.

ANÁLISE

O código geratriz da análise se encontra no Colab. Nesta análise foram feitas Tabelas Cruzadas para duas variáveis qualitativas, histogramas para uma variável quantitativa, e gráficos de colunas para dados qualitativos e quantitativos. Para a análise efetiva do *dataset* e suas observações foram empregadas uma gama de técnicas estatísticas descritivas e indutivas.

Estatísticas Descritivas

Nessa seção, buscou-se elaborar estatísticas descritivas com a amostra de uma variável quantitativa, para o nosso caso, escolheu-se a variável quantidade.

Figura 1 - estatísticas descritivas gerada pelo *Summary*

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	36	800	292752	14132	207552841

Fonte: Elaboração própria gerada com base nos dados do BCB (2025).

Conforme a Figura 1, uma acerca quantitativa foi passada como parâmetro à função s*ummary* para calcular automaticamente algumas estatísticas descritivas como: média, mediana, máximo, mínimo, e por fim, primeiro ao terceiro quartil.

Tabela 2 - Estatística descritiva gerada pelas funções

```
-------- Por funções --------
|Mínimo:       |           1|
|1º Quartil:   |       36.00|
|Mediana:      |   292752.43|
|Média:        |   292752.43|
|3º Quartil:   |    14131.75|
|Máximo:       |207552841.00|
-------- Por funções --------
```

Fonte: Elaboração própria gerada com base nos dados do BCB (2025).

Também se calculou esses mesmos valores estatísticos por meio de funções separadamente pelo R, conforme a Tabela 2, elas aparecem respectivamente como média, mediana, desvio padrão, variância e moda, sem essa a única em que não há especificamente uma função. Por meio dessa tabela, percebe-se que os valores são aparecidos.

Tabela 3 - Tabela de frequência gerada pelo *summary*

valor	n_i	f_i	p_i
<chr>	<int>	<chr>	<chr>
(1,5.19e+07]	519918	0.999	99.943
(5.19e+07,1.04e+08]	244	0	0.047
(1.04e+08,1.56e+08]	38	0	0.007
(1.56e+08,2.08e+08]	12	0	0.002
Total	520212	1	100

Fonte: Elaboração própria gerada com base nos dados do BCB (2025).

Partindo dessa estatística descritiva, pode-se gerar uma tabela de frequência com a amostra da variável quantidade, conforme a Tabela 3.

Contingência entre variáveis

Selecionou-se duas variáveis quantitativas em que se desejava saber se há associação. Para a geração dessa tabela, relacionou-se as amostras das variáveis região e tipo de pagador a fim de saber se há associação, a Tabela 4 abaixo, é o resultado da primeira tabela:

Tabela 4 – Quantidade absoluta de PF e PJ por Região

	nao disponivel	pf	pj	Total_coluna
CO	0	75001	17561	92562
NE	0	78785	17747	96532
N	0	75719	16076	91795
SE	0	79671	19091	98762
S	0	74298	17478	91776
Total linha	0	383474	87953	471427

Fonte: Elaboração própria com base nos dados do BCB (2025).

A Tabela 4 são os resultados absolutos da associação da amostra das variáveis onde se obteve a frequência que pessoa jurídica ou física aparece em relação às regiões do país.

Tabela 5 – Quantidade relativa de PF e PJ por Região

	nao disponivel	pf	pj	Total_coluna
CO	0	15.91	3.73	19.64
NE	0	16.71	3.76	20.47
N	0	16.06	3.41	19.47
SE	0	16.90	4.05	20.95
S	0	15.76	3.71	19.47
Total linha	0	81.34	18.66	100.00

Fonte: Elaboração própria com base nos dados do BCB (2025).

Já a Tabela 5 é uma tabela das frequências relativas (em porcentagem) geradas a partir da tabela anterior, ela contém os decimais que geram os percentuais.

Tabela 5 – Gráfico de contingência com dados observados

	CO	NE	N	SE	S	Total coluna
nao disponivel	0	0	0	0	0	0
pf	75001	78785	75719	79671	74298	383474
pj	17561	17747	16076	19091	17478	87953
Total linha	92562	96532	91795	98762	91776	471427

Fonte: Elaboração própria com base nos dados do BCB (2025).

Nesse caso, a Tabela 5 foi elaborado como a transposta da anterior, dessa vez está se relacionando a região com o pagamento ter sido efetuado por uma pessoa jurídica ou física. Dessa forma, obteve-se uma tabela com o *layout* inverso da anterior.

Tabela 6 – Gráfico de contingência com dados relativos

	CO	NE	N	SE	S	Total coluna
nao disponivel	0.00	0.00	0.00	0.00	0.00	0.00
pf	15.91	16.71	16.06	16.90	15.76	81.34
pj	3.73	3.76	3.41	4.05	3.71	18.66
Total linha	19.64	20.47	19.47	20.95	19.47	100.00

Fonte: Elaboração própria com base nos dados do BCB (2025).

A partir da tabela anterior, gerou-se a tabela com as frequências relativas que podem gerar os percentuais, conforme a Tabela 6, que correspondem a associação entre as amostras das variáveis região e tipo do pagador.

Figura 2 – Chi-quadrado de Person

```
        Pearson's Chi-squared test

data:  df_na$pagregiao and df_na$pagidade
X-squared = 2.0759, df = 20, p-value = 1
```

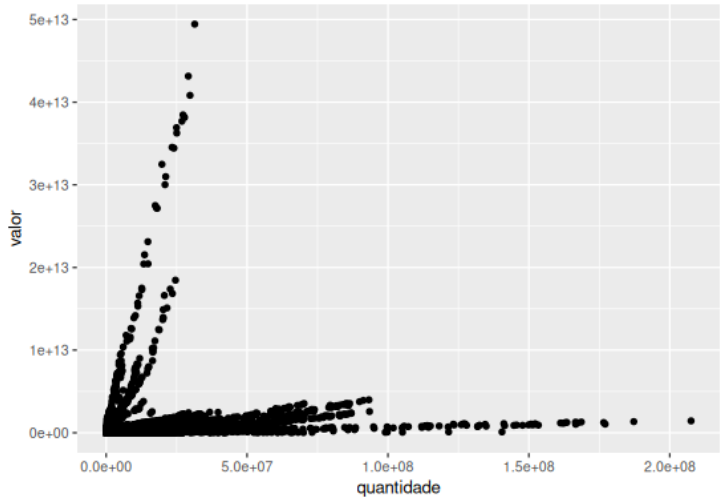
Fonte: Elaboração própria com base nos dados do BCB (2025).

Embora as tabelas demonstrassem que há indícios de associação entre as amostras, era necessário fazer também um teste chi-quadrado de Pearson que permite ter mais certeza acerca da associação entre as amostras, conforme a figura 2. A interpretação é que a curva tende ao infinito positivo, então, quanto maior o valor do chi-quadrado, maior a associação, então, pode-se considerar de que há indícios que haja associação.

Correlação entre Variáveis

Selecionou-se duas variáveis quantitativas a fim de investigar se há correlação entre essas variáveis. Para gerar as figuras e tabelas dessa seção, escolheu-se as variáveis valor e quantidade.

Figura 3 – Gráfico de dispersão entre ‘valor’ e ‘quantidade’



Fonte: Elaboração própria com base nos dados do BCB (2025).

Conforme a Figura 3, percebe-se uma relação entre os valores de quantidade e valores que segue padrão distinto de "retas”, com duas principais direções. Em geral, neste *dataset*, quanto mais transações para uma dada entrada, menor os valores de cada transação individualmente, e quanto maior o total das transações num dado mês, menor a quantidade de transações naquele mês.

Figura 4 – Chi-quadrado de Person

```
0.249538531099509

    Pearson's product-moment correlation

data:  valor and quantidade
t = 190.47, df = 546308, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2470503 0.2520235
sample estimates:
      cor
0.2495385
```

Fonte: Elaboração própria com base nos dados do BCB (2025).

Conforme a Figura 4, a correlação entre as amostras das variáveis é alta, p-valor desse teste verifica também a existência de correlação entre essas variáveis.

Determinação entre Variáveis

Selecionou-se duas amostras de variáveis, uma quantitativa e uma qualitativa, a fim de calcular associação de determinação entre essas variáveis. Para esse estudo se selecionou a idade

do pagador e a quantidade paga para ver se há associação entre essas amostras.

Figura 5 – Determinação pela a *Summary*

```
$`[19]`
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
       1       37      522   100478     7689 39911542

$`[20,29]`
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
        1       118      2122    531440     35249 177062118

$`[30,39]`
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
        1       110      2109    555360     34254 207552841

$`[40,49]`
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
        1        83      1530    402590     22922 162865781

$`[50,59]`
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
       1       47      814   196724    10860 79316934

$`[60)`
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
       1       29      452    98393     5387 31981290
```

Fonte: Elaboração própria com base nos dados do BCB (2025).

A FIgura 5 é um exemplo do cálculo de determinação entre duas variáveis utilizando o *summar*. Nesse caso, todas as estatísticas descritivas são calculadas para todos os valores que a amostra qualitativa pode assumir. Como a amostra idade do pagador é uma qualitativa ordinal, que tem uma ordem, ela se apresenta na forma de intervalos do menor para o menor.

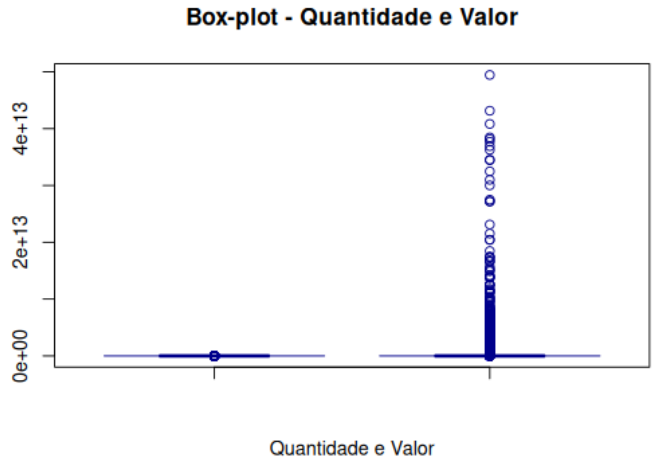
Figura 6 – Determinação com idade e quantidade

```
[19]:  66940 [20,29]:  76285 [30,39]:  76743 [40,49]:  75810 [50,59]:  73183 [60):  70502
```

Fonte: Elaboração própria com base nos dados do BCB (2025).

Já na Figura 6 ocorre algo semelhante, no entanto, em vez do parâmetro *summary*, aplicou-se o parâmetro *length*, por isso se obteve apenas a incidência de cada rótulo em relação a quantidade.

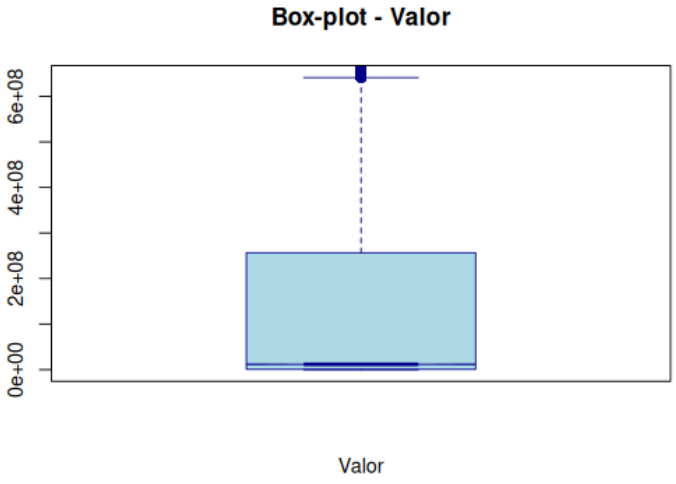
Figura 7 – *Boxplot* de determinação (quantidade e valor)



Fonte: Elaboração própria com base nos dados do BCB (2025).

O gráfico acima (Figura 6) demonstra um problema inerente a esse *dataset*: devido a distribuição dos dados, com *outliers* excessivos, tanto em quantidade quanto em escala, análises da disribuição dos dados se torna praticamente inviável sem antes definier limites. Note que mesmo a ordenada estando sendo compartilhada, o problema se repete a ambas variáveis.

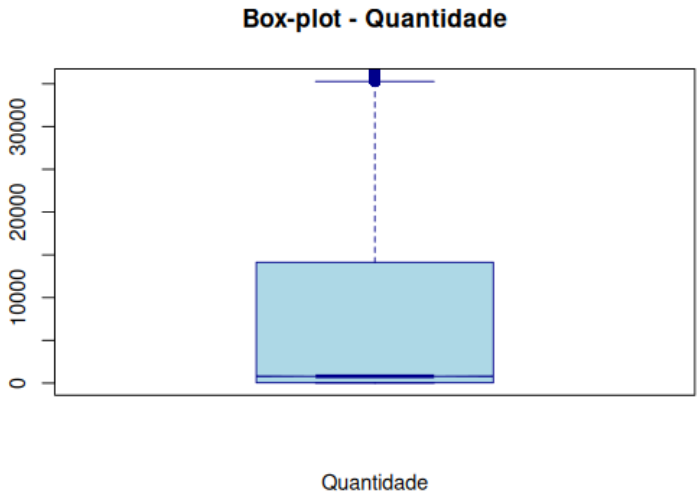
Figura 7 – *Boxplot* de determinação (valor até 2.5X o 3º Quartil)



Fonte: Elaboração própria com base nos dados do BCB (2025).

Acima (Figura 7) podemos verificar o *Boxplot* do "valor" limitado nas ordenadas para que varie de 0 a 2,5 x 3º Quartil. Aqui podemos observar o segundo problema com esse *dataset*, a distribuição tem carácter exponencial, o que leva à "distância" inter-quartil a ser exagerada à cada patamar (vide Figura 1 e Tabela 2).

Figura 8 – *Boxplot* de determinação (quantidade até 2.5X o 3º Quartil)



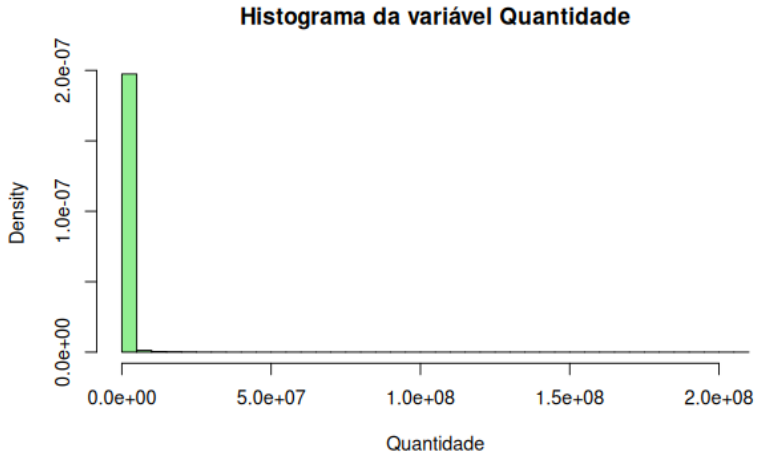
Fonte: Elaboração própria com base nos dados do BCB (2025).

O mesmo ocorre com a Figura 8 para a variável quantidade.

Histogramas

O histograma (por densidade) gerou para as duas variáveis quantitativas valores muitos distorcidos:

Figura 9 – Histograma da variável quantidade

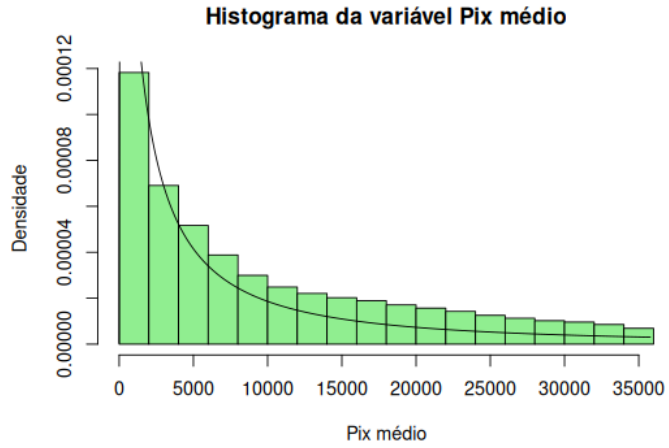


Fonte: Elaboração própria com base nos dados do BCB (2025).

A fim de uma análise mais aprofundada, decidiu-se olhar também uma subseção das variáveis (com quantidade de 0 até o 3º quartil daquela variável), conforme a Figura 9, de onde, por inspeção, decidiu-se utilizar a distribuição exponencial "lnorm" para ajustar uma curva

correspondente ao formato aproximado da distribuição total (note que o valor densidade é relativa ao ponto de interesse):

Figura 10 – Histograma do pix médio (reduzido e ajustado)

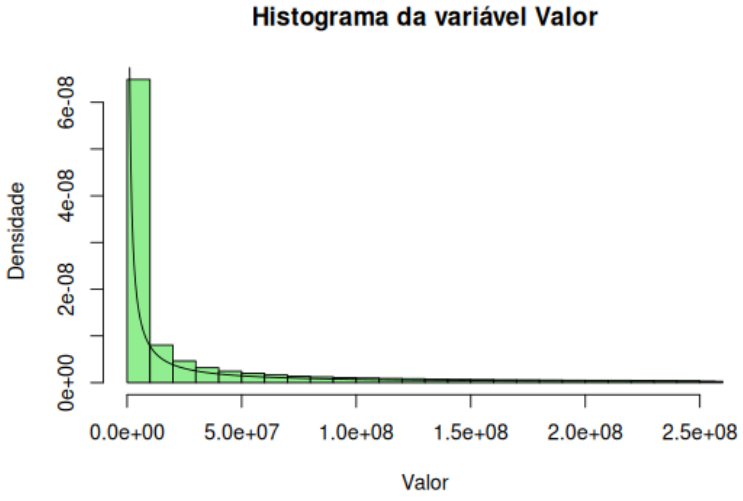


Fonte: Elaboração própria com base nos dados do BCB (2025).

Vale denotar que esse padrão exponencial se demonstra aproximadamente constante para qualquer intervalo na qual se pegar os dados, conforme a figura 10.

Para as demais variáveis quantitativas o histograma até o 3º quartil se encontra representado abaixo:

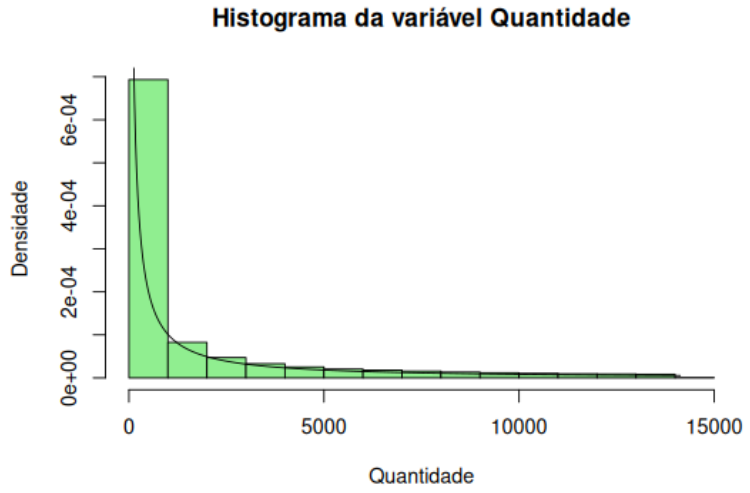
Figura 11 – Histograma do valor (reduzido e ajustado)



Fonte: Elaboração própria com base nos dados do BCB (2025).

O valor é, dentre as quantitativas, a que tem uma curva de densidade mais acentuada, conforme a figura 11, justificado pelo fato de que 25% dos seus dados (até o primeiro quartil) correspondem a um valor muito baixo (0 a 452385.50), mas o espaço entre o segundo e primeiro quartil é aproximadamente 25 vezes maior (e, portanto, mais disperso), por referência essa diferença é "apenas" 21 vezes no caso da variável "quantidade".

Figura 12 – Histograma da quantidade (reduzido e ajustado)

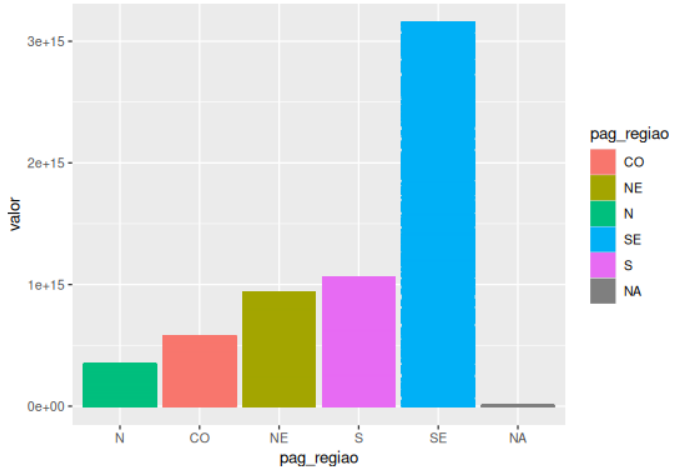


Fonte: Elaboração própria com base nos dados do BCB (2025).

A curva "lnorm" foi considerada a mais adequada, conforme a figura 12, dada a distribuição aparente dos dados, para as três variáveis quantitativas aqui expostas. Vale denotar que a variável pix média foi gerada com o único intuito de se comparar o seu gráfico ao das demais variáveis quantitativas.

Para a análise dos dados qualitativa por quantitativa, optou-se pelo uso de um gráfico de colunas:

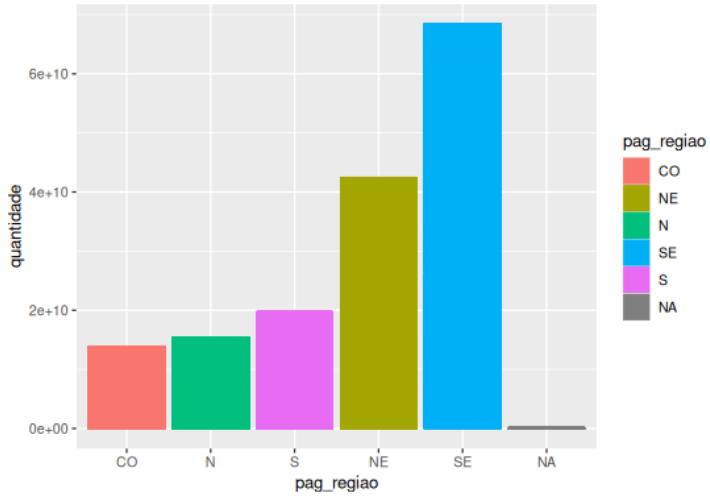
Figura 13 – Gráfico de colunas de valores agregados do PIX por região



Fonte: Elaboração própria com base nos dados do BCB (2025).

O gráfico, figura 13, mostra que os valores totais de transações PIX feitas no por região contidas no *dataset* em muito superam àquelas feitas nas demais regiões. Uma coisa notável nesse gráfico é que os valores mensais totais do Nordeste (NE) são o terceiro maior, atrás da região Sul (S), contudo, em termos de quantidade (Figura 6), o Nordeste é onde a segunda maior quantidade de PIXs realizados no período do *dataset*, o que sugere que está região tenha mais PIX de "menor valor" quando comparada com a região sudeste:

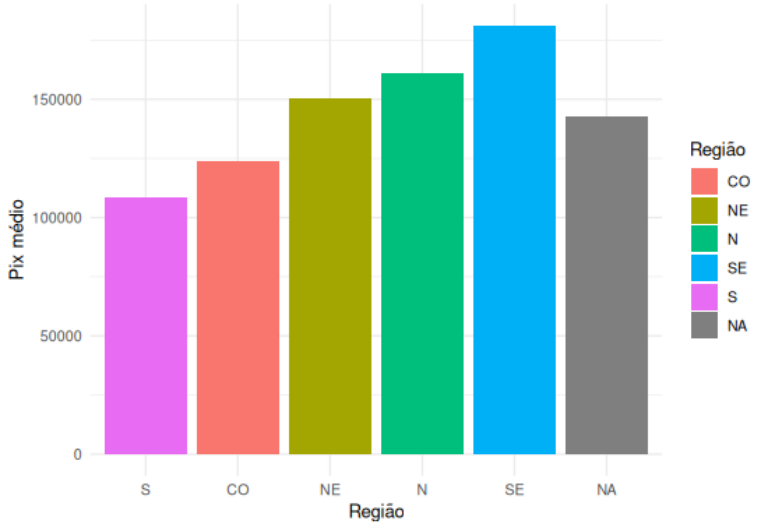
Figura 14 – Gráfico de colunas de quantidade agregada do PIX por região



Fonte: Elaboração própria com base nos dados do BCB (2025).

A teoria de que a região Nordeste tem PIX de menor valor no período considerado pode ser corroborada ao se olha o gráfico de colunas da figura 14, que considera a variável quantitativa como sendo o "Pix Médio":

Figura 15 – Gráfico de colunas do Pix Médio agregado por região

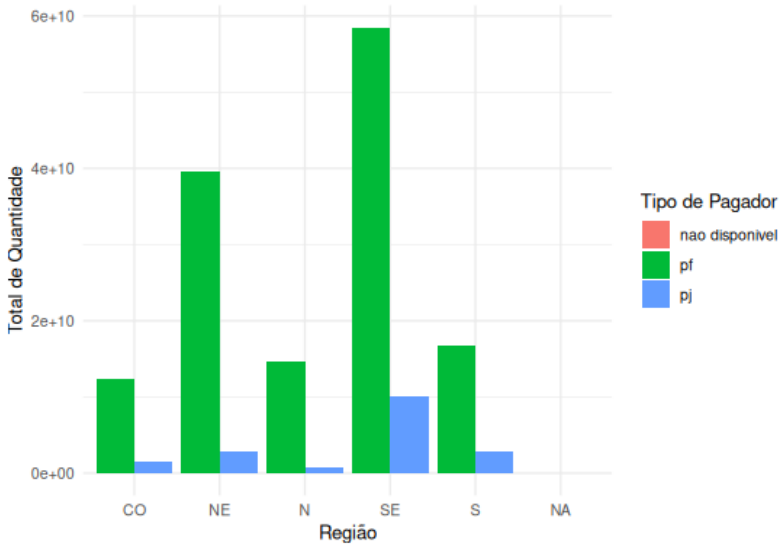


Fonte: Elaboração própria com base nos dados do BCB (2025).

O “Pix Médio” demonstra que não é o caso, já que o Pix médio mensal é maior no Nordeste, conforme a figura 15. O que possivelmente ocorreu é que a região Sul teve grandes transações de PIX no período considerado, passando a região Nordeste, que apesar de ter mais PIXs realizados e com valores médios maiores, não pode ultrapassar o valor total angariado pelo Sul neste mesmo período.

Em vista dessas transações grandes feitas no Sul, espera-se que haja maior quantidade de PIX por PJ lá do que no Nordeste, justificando o alto capital das empresas, quando comparado com o de indivíduos:

Figura 16 – Tipo de pagador por região



Fonte: Elaboração própria com base nos dados do BCB (2025).

A figura 16 mostra haver, relativo ao total de quantidade de transações, um número maior de transações de PJs do que de PFs no Sul do que comparado ao Nordeste. O Nordeste tem maior quantidade de transações de pessoas físicas do que o Sul, contudo, eles têm quantidades parecidas de transações de pessoas jurídicas. O número maior de transações de PFs era esperado, já que a população nordestina é maior do que a sulista, contudo, o número parecido de PJs implica em maior atividade comercial nesta região (quando comparada à região Nordeste) no período considerado.

Conclusões mais precisas acerca das distribuições dos dados se demonstram complexas, e demandam maior quantidade de tempo para serem plenamente sintetizadas.

CÓDIGO

- O *dataset* utilizado pode ser encontrado aqui (necessário utilizar e-mail institucional): estatisticas-de-transações-pix.csv.zip.
- O código fonte pode ser encontrado aqui: https://colab.research.google.com/drive/1SOQNpJ47FwGUL3sLn-pXYx3nq6kE1KLS?usp=sharing.
- O Github do projeto pode ser encontrado aqui: https://github.com/igorzeck/AnaliseRESTDoPix

REFERÊNCIAS BIBLIOGRÁFICAS

LEAL, Alan Marques Miranda; DE OLIVEIRA HAASE, Mariana Aparecida. Instant Payments and Banks: the Impact of Pix on Bank Branches in Brazil. Disponível em: https://www.researchgate.net/publication/393782221_Instant_Payments_and_Banks_the_Impact_of_Pix_on_Bank_Branches_in_Brazil. Disponível em: 27 de ago. de 2025.