Tratamento Multivariado de dados de Cartão de Crédito

Grupo 2
Antônio Marcos
Douglas Nestlehner

Outubro, 2021

Introdução

- A vertente de Crédito é um dos carros chefes dentro das instituições financeiras. É através da concessão de crédito aos clientes e cobrança de taxas, que se dá boa parte da receita dessas empresas. É fato que muitos dados são gerados nesse trâmite, pertencentes a inúmeras variáveis, seja de atividade, perfil, dentre outras.
- Sendo assim, linkando esse cenário ao conteúdo apresentado na disciplina de Estatística Multivariada, iremos aplicar técnicas multivariadas de tratamento, visualização e preparação para futuras análises, de um banco de dados sobre trâmites em cartões de crédito de uma instituição financeira.

Base de Dados

- Base de dados: "Clientes de Cartão de Crédito"retirada da plataforma Kaggle (link disponibilizado no relatório)
- Essa base refere-se a informações de clientes de um determinado banco (não informado), em que os dados foram coletados com o intuito de analisar o motivo dos clientes estarem desistindo do cartão de credito da empresa.
- Com isso foram coletadas 10127 observações com 21 covariaveis distintas.

Base de Dados

Covariáveis analisadas no estudo:

$\mathbf{X}_1$: Idade do cliente (em anos);
$\mathbf{X}_2$: Período de relacionamento com banco (em meses);
$\mathbf{X}_3$: Limite de crédito no cartão de crédito (U$);
$\mathbf{X}_4$: Saldo rotativo total no cartão de crédito (U$);
$\mathbf{X}_5$: Linha de crédito aberta para compra (média dos últimos 12 meses);
$\mathbf{X}_6$: Mudança no valor da transação (U$);
$\mathbf{X}_7$: Valor total da transação (últimos 12 meses);
$\mathbf{X}_8$: Contagem total de transações (nos últimos 12 meses);
$\mathbf{X}_9$: Mudança na contagem de transações (unidades);
$\mathbf{X}_{10}$: Taxa de utilização média do cartão (%).

Objetivo

- Utilizar de técnicas multivariadas para tratar, visualizar e analisar o comportamento das covariáveis presentes no banco de dados. Assim possibilitando e facilitando interpretações para que possam ser introduzidas sugestões à estudos futuros mais direcionados.

Análise Descritiva dos Dados

Análise Descritiva dos Dados

Na Tabela 1 estão representadas algumas observações das 10 variáveis anteriormente listadas.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
45	39	12691	777	11914	1.335	1144	42	1.625	0.061
49	44	8256	864	7392	1.541	1291	33	3.741	0.105
51	36	3418	0	341812	2.592	1887	20	2.333	0.000
40	34	3313	2517	796	1.405	1171	20	2.333	0.760
...	...	...	...	...	...	...	...		
41	25	42771	2186	2091	0.804	8764	69	0.683	0.511
44	36	5409	0	5409	0.819	10291	60	0.818	0.000
30	36	5281	0	5281	0.535	8395	62	0.722	0.000
43	25	10388	1961	8427	0.703	10294	61	0.649	0.189

Tabela: Base de dados

Análise Descritiva dos Dados

Para conhecermos melhor os nossos dados, construímos gráficos e calculamos algumas medidas descritivas: mínimo; 1º quartil; mediana; média; 3º quartil; e máximo, de todas as covariáveis em estudo:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Mínimo	26.00	13.00	1438	0	3	0	510	10.00	0	0
1º Quartil	41.00	31.00	2555	359	1324	0.6310	2156	45.00	0.5820	0.0230
Mediana	46.00	36.00	4549	1276	3474	0.7360	3899	67.00	0.7020	0.1760
Média	46.33	35.93	8632	1163	7469	0.7599	4404	64.86	0.7122	0.2749
3º Quartil	52.00	40.00	11068	1784	9859	0.8590	4741	81.00	0.8180	0.5030
Máximo	73.00	56.00	34516	2517	34516	3.3970	18484	139.00	3.7140	0.9990

X1

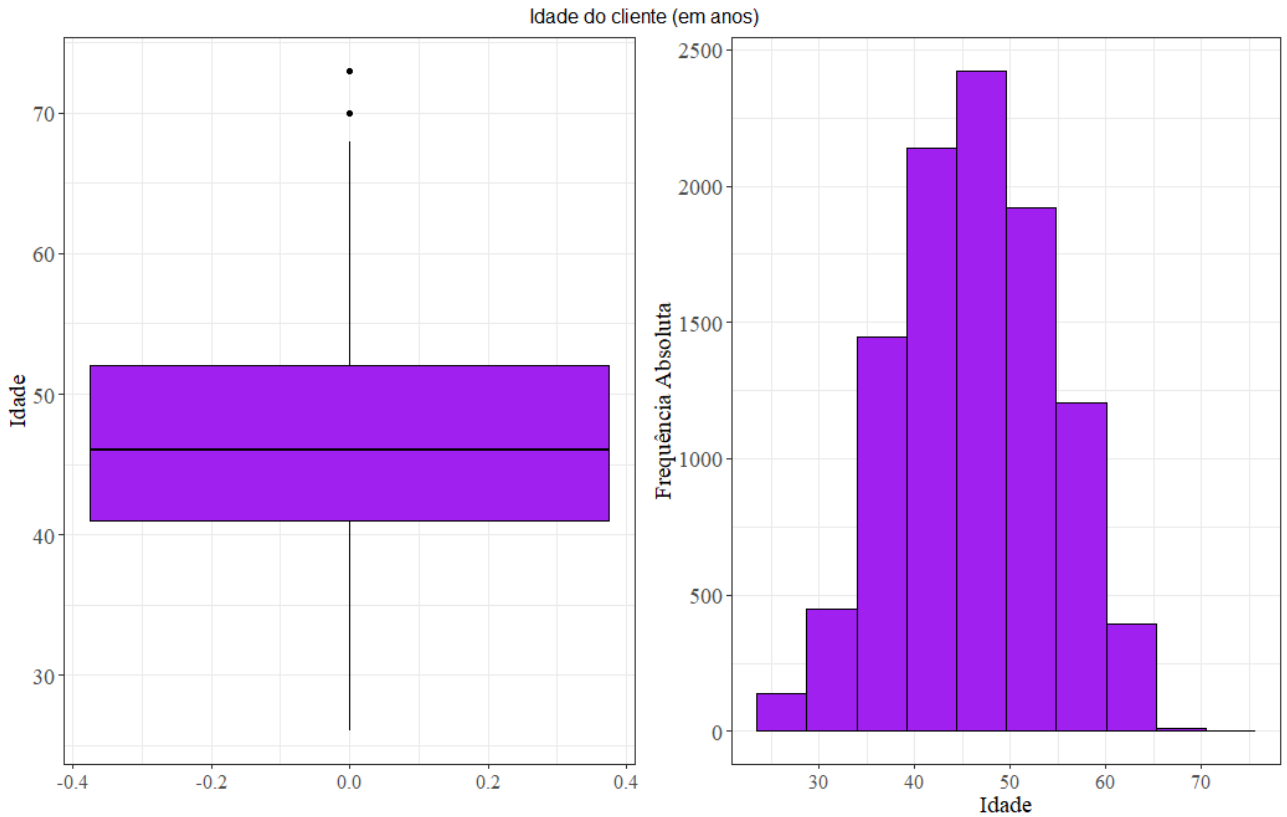

Figura: Boxplot e Histograma da Idade do cliente

X2

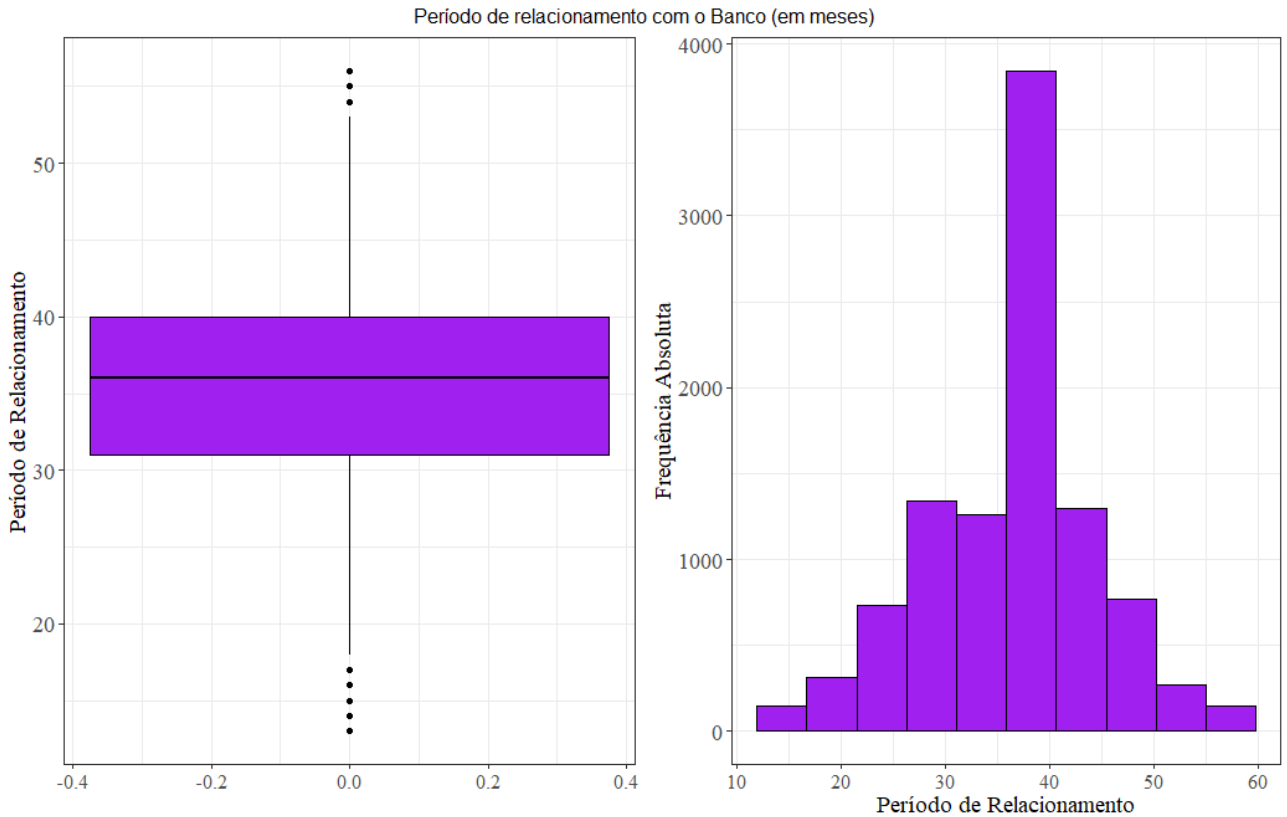

Figura: Boxplot e Histograma do Relacionamento com o Banco (meses)

X3

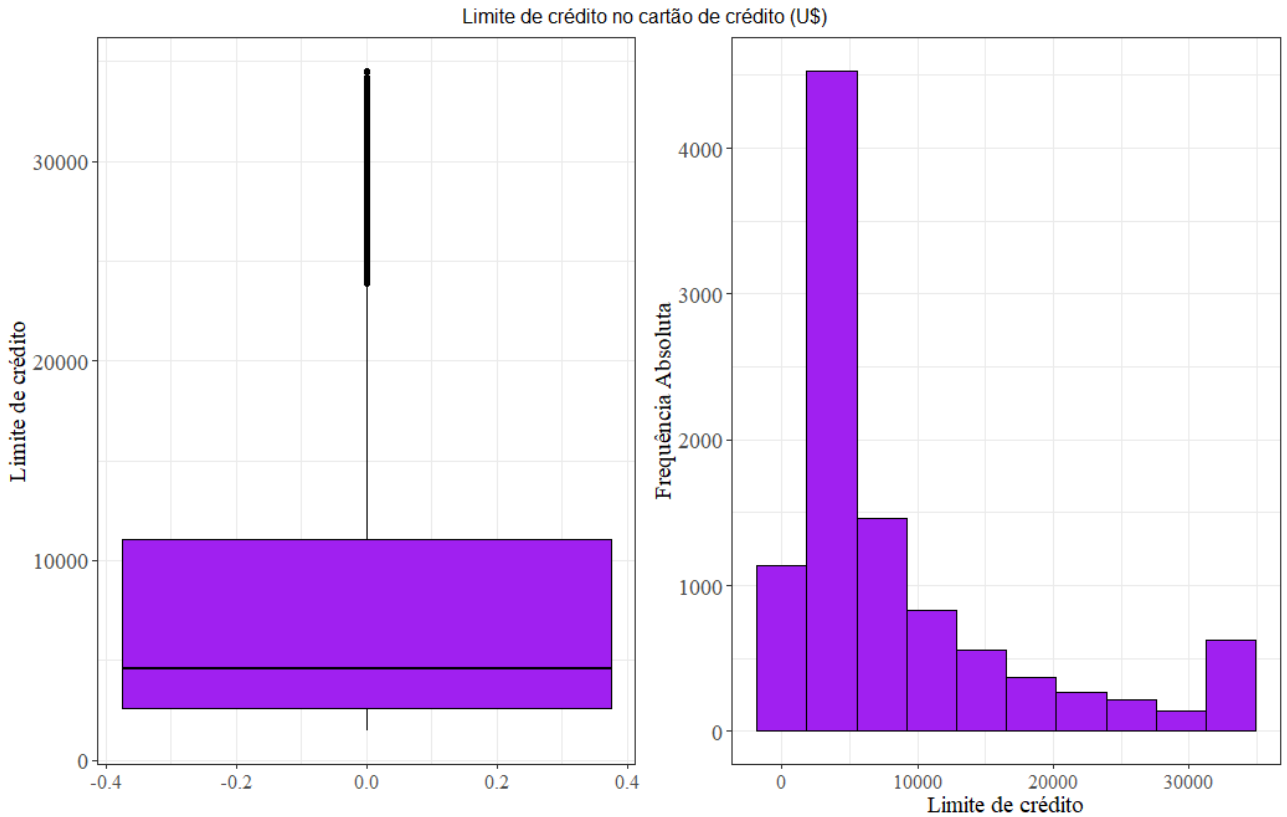

Figura: Boxplot e Histograma do Limite de Crédito

X4

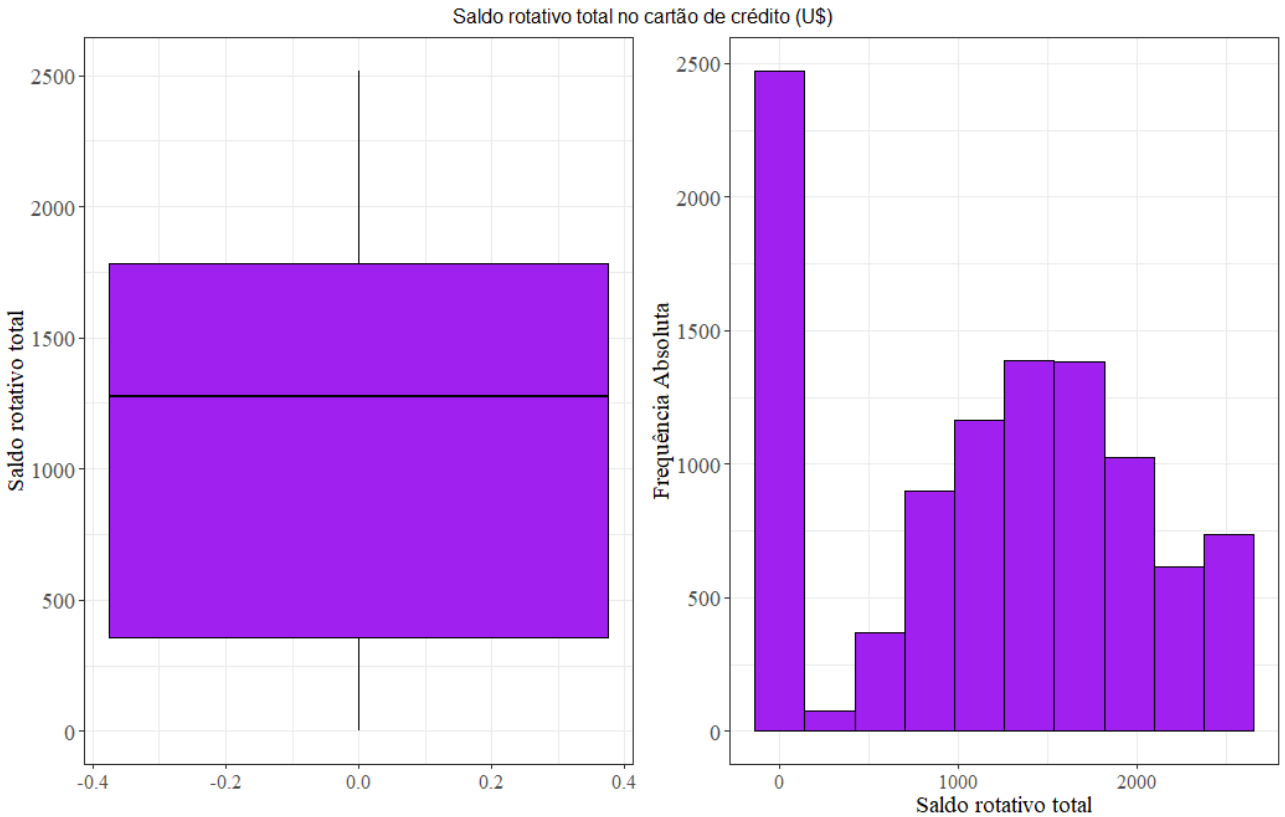

Figura: Boxplot e Histograma do Saldo rotativo total

X5

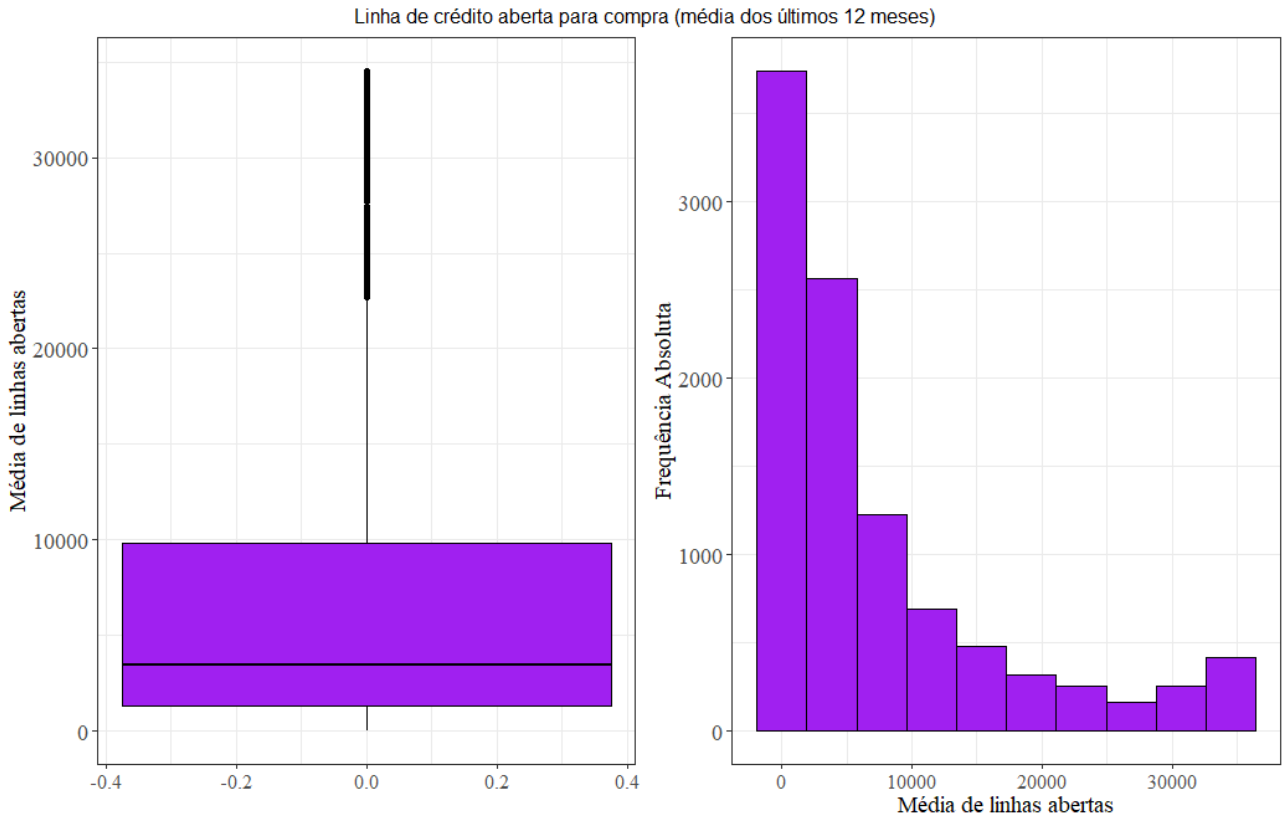

Figura: Boxplot e Histograma de Linhas de crédito aberta (último ano)

X6

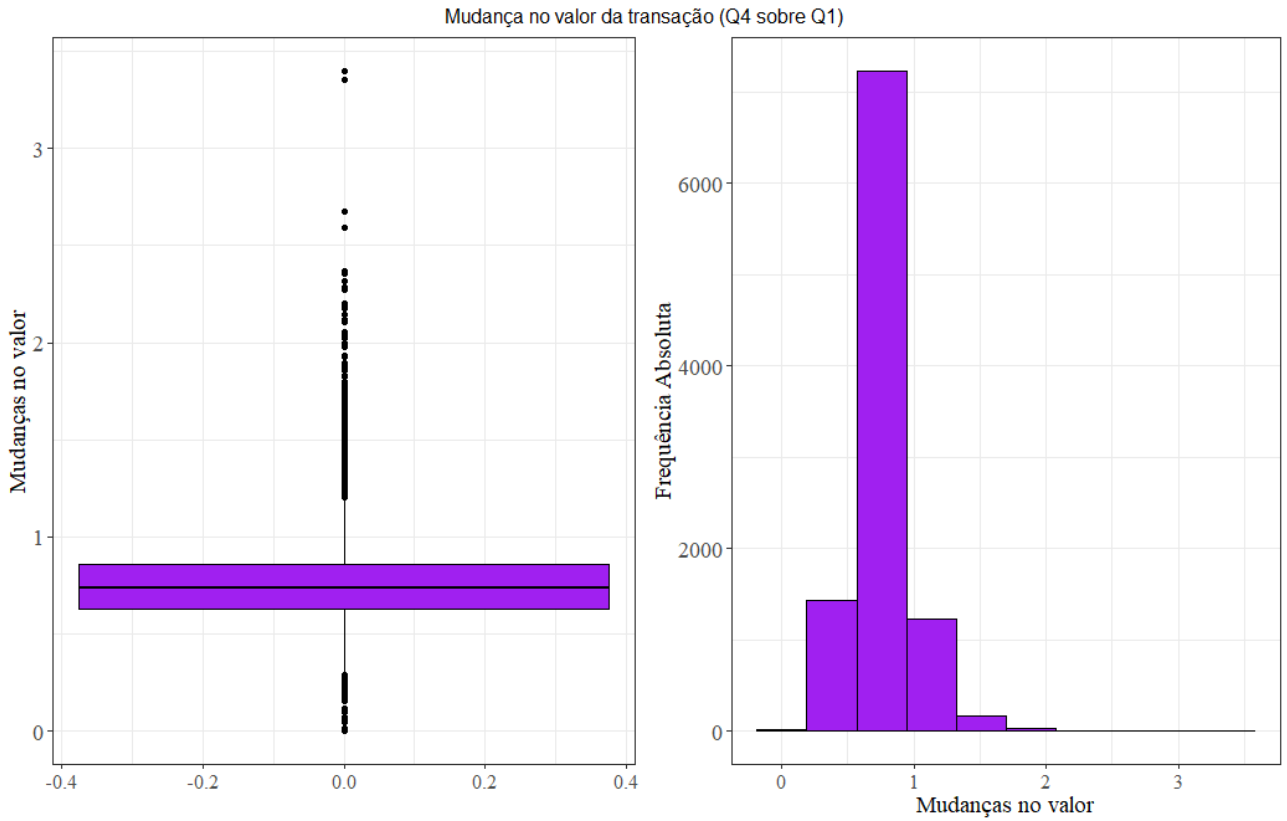

Figura: Boxplot e Histograma da Mudança no valor da Transação

X7

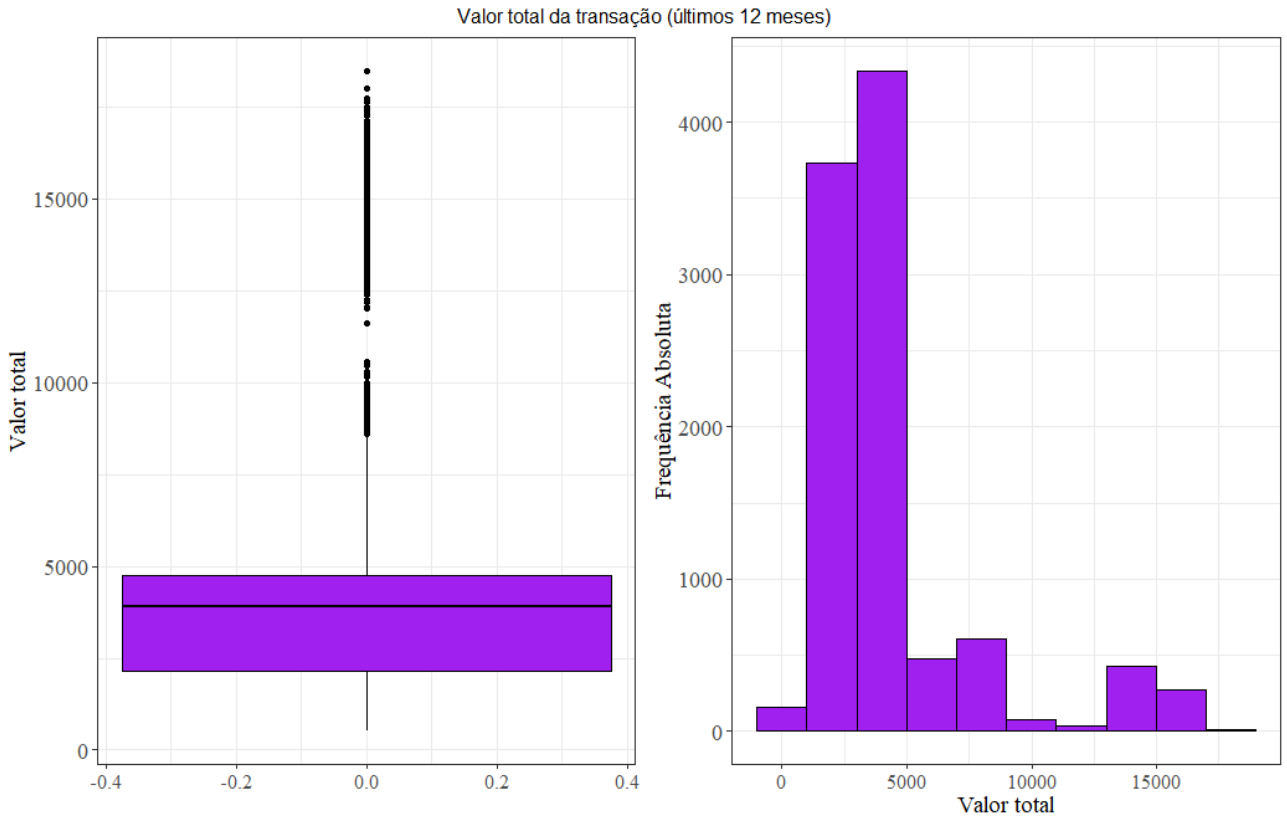

Figura: Boxplot e Histograma do Valor total da Transação

X8

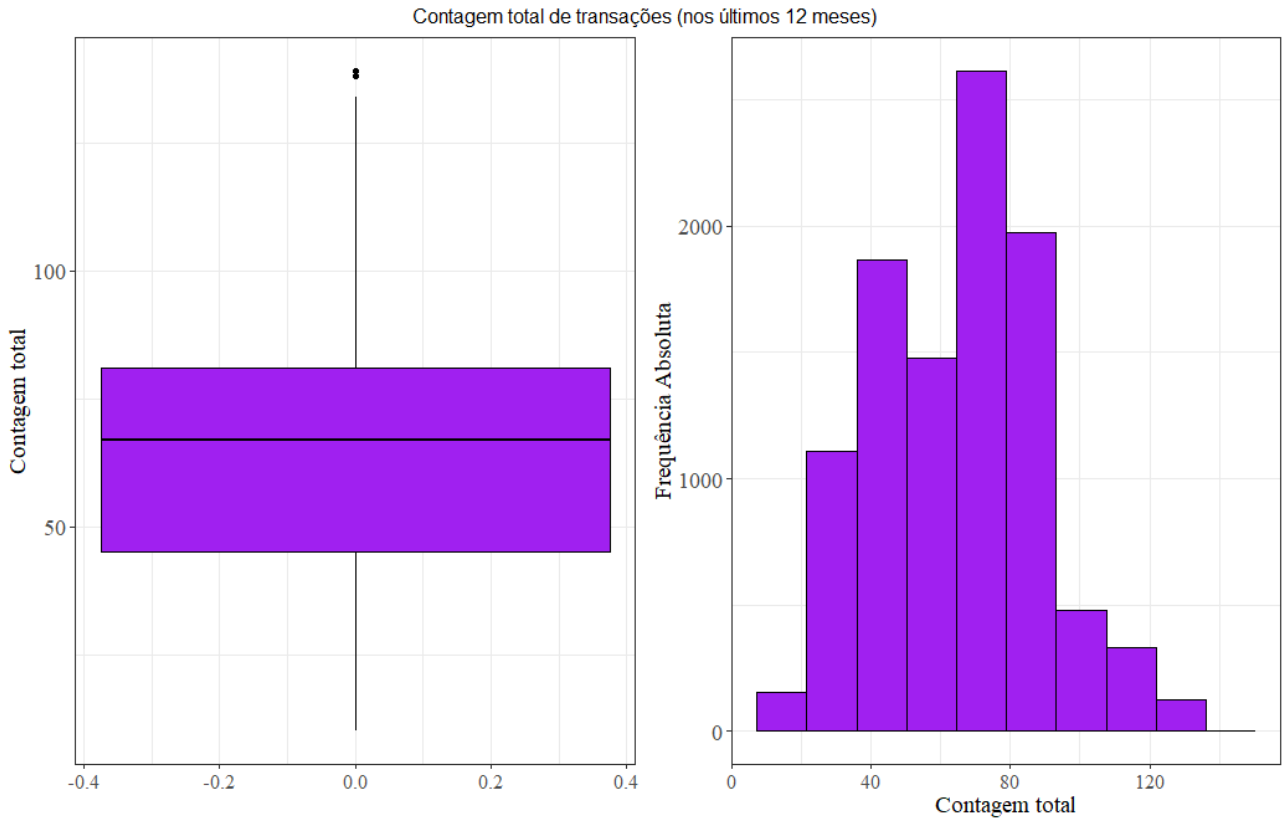

Figura: Boxplot e Histograma do Total de Transações (último ano)

X9

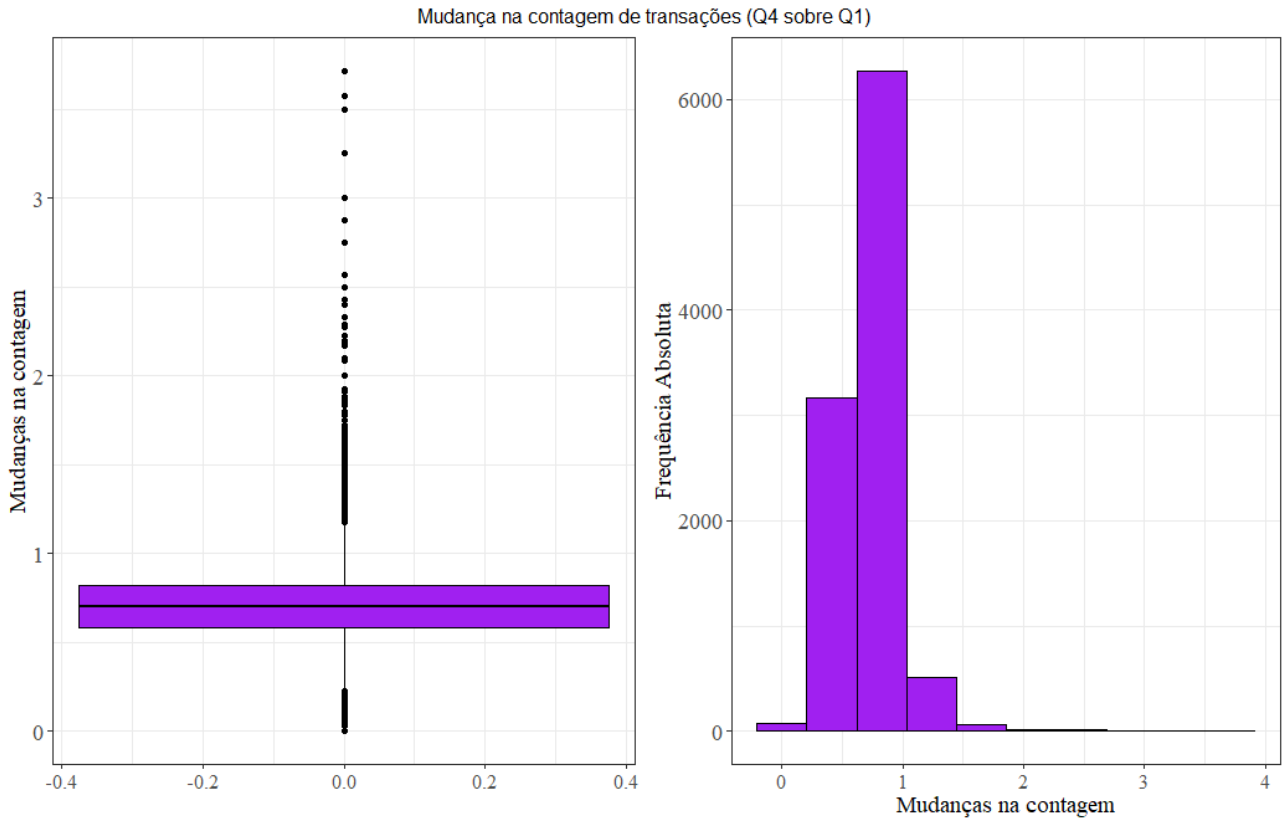

Figura: Boxplot e Histograma das Mudanças na contagem de Transações

X10

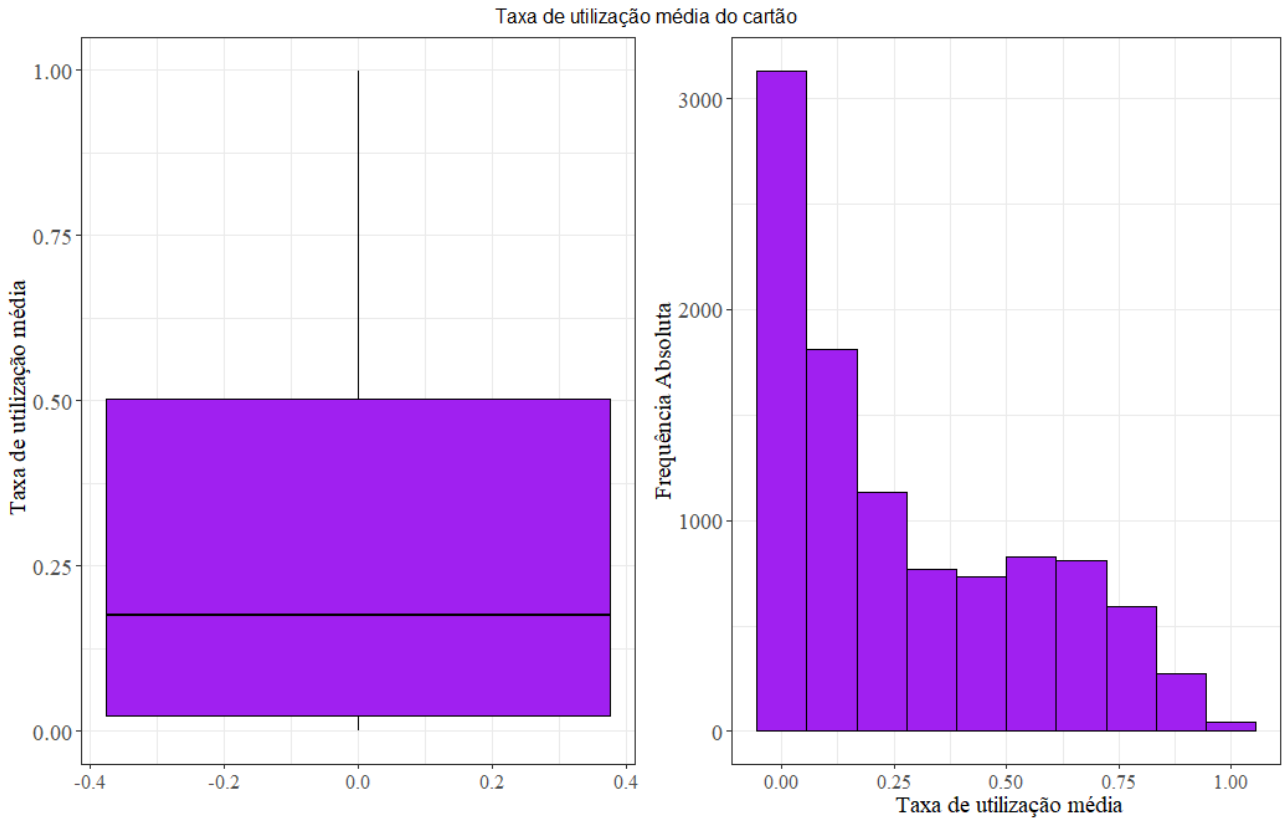

Figura: Boxplot e Histograma da Taxa de utilização média do Cartão

Medidas Descritivas

Vetor de Médias

Obteve-se o vetor de médias das 10 variáveis do banco de dados em estudo, sendo:

$$\bar{\mathbf{x}} = \begin{pmatrix} 46.326 \\ 35.928 \\ 8631.953 \\ 1162.814 \\ 7469.140 \\ 0.760 \\ 4404.086 \\ 64.859 \\ 0.712 \\ 0.275 \end{pmatrix}.$$

Matriz de Variâncias e Covariâncias (Var-Cov)

Obtendo a matriz das 10 covariáveis analisadas. Essa matriz é simétrica e serve para medir o grau de relacionamento linear entre duas variáveis.

$$\begin{bmatrix}
64.26 & 50.51 & 180.41 & 96.56 & 83.85 & -0.11 & -1264.81 & -12.62 & -0.02 & 0.02 \\
 & 63.78 & 544.86 & 56.12 & 488.74 & -0.09 & -1046.89 & -9.34 & -0.03 & -0.02 \\
 & & 82597704.01 & 314721.77 & 82282982.24 & 25.52 & 5301773.45 & 16196.42 & -4.37 & -1210.05 \\
 & & & 664138.77 & -349417.00 & 10.39 & 178199.62 & 1072.32 & 17.43 & 140.19 \\
 & & & & 82632399.24 & 15.13 & 5123573.83 & 15124.09 & -21.80 & -1350.24 \\
 & & & & & 0.05 & 29.54 & 0.03 & 0.02 & 0.00 \\
 & & & & & & 11539347.59 & 64358.62 & 69.21 & -77.76 \\
 & & & & & & & 550.91 & 0.63 & 0.02 \\
 & & & & & & & & 0.06 & 0.00 \\
 & & & & & & & & & 0.08
\end{bmatrix}$$

Figura: Matriz Var-Cov

Variância Total e Variância Generalizada

Afim de sintetizar a variabilidade multivariada dos dados em um único valor numérico. Calculamos:

- Variância Total:

$$tr(S) = 177451791.$$

- Variância Generalizada:

$$\det(S) = 244049967398863$$

Matriz de Desvio Padrão

Obtemos também a Matriz de Desvio Padrão:

$$\begin{bmatrix} 8.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & 7.98 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & 9088.32 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & 814.94 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & 9090.23.24 & 0 & 0 & 0 & 0 & 0 \\ & & & & & 0.21 & 0 & 0 & 0 & 0 \\ & & & & & & 3396.96 & 0 & 0 & 0 \\ & & & & & & & 23.47 & 0 & 0 \\ & & & & & & & & 0.23 & 0 \\ & & & & & & & & & 0.27 \end{bmatrix}$$

Figura: Matriz de desvios padrão

Matriz de Correlação

Para analisar o grau de relacionamento entre duas covariaveis, calculamos a matriz de correlação:

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1.000	0.789	0.002	0.015	0.001	-0.062	-0.046	-0.067	-0.012	0.007
X_2	0.789	1.000	0.008	0.009	0.007	-0.049	-0.039	-0.050	-0.014	-0.008
X_3	0.002	0.008	1.000	0.042	0.996	0.013	0.172	0.076	-0.002	-0.483
X_4	0.015	0.009	0.042	1.000	-0.047	0.058	0.064	0.056	0.090	0.624
X_5	0.001	0.007	0.996	-0.047	1.000	0.008	0.166	0.071	-0.010	-0.539
X_6	-0.062	-0.049	0.013	0.058	0.008	1.000	0.040	0.005	0.384	0.035
X_7	-0.046	-0.039	0.172	0.064	0.166	0.040	1.000	0.807	0.086	-0.083
X_8	-0.067	-0.050	0.076	0.056	0.071	0.005	0.807	1.000	0.112	0.003
X_9	-0.012	-0.014	-0.002	0.090	-0.010	0.384	0.086	0.112	1.000	0.074
X_{10}	0.007	-0.008	-0.483	0.624	-0.539	0.035	-0.083	0.003	0.074	1.000

Figura: Matriz de Correlação

Também representamos a matriz de correlação pelo gráfico de calor:

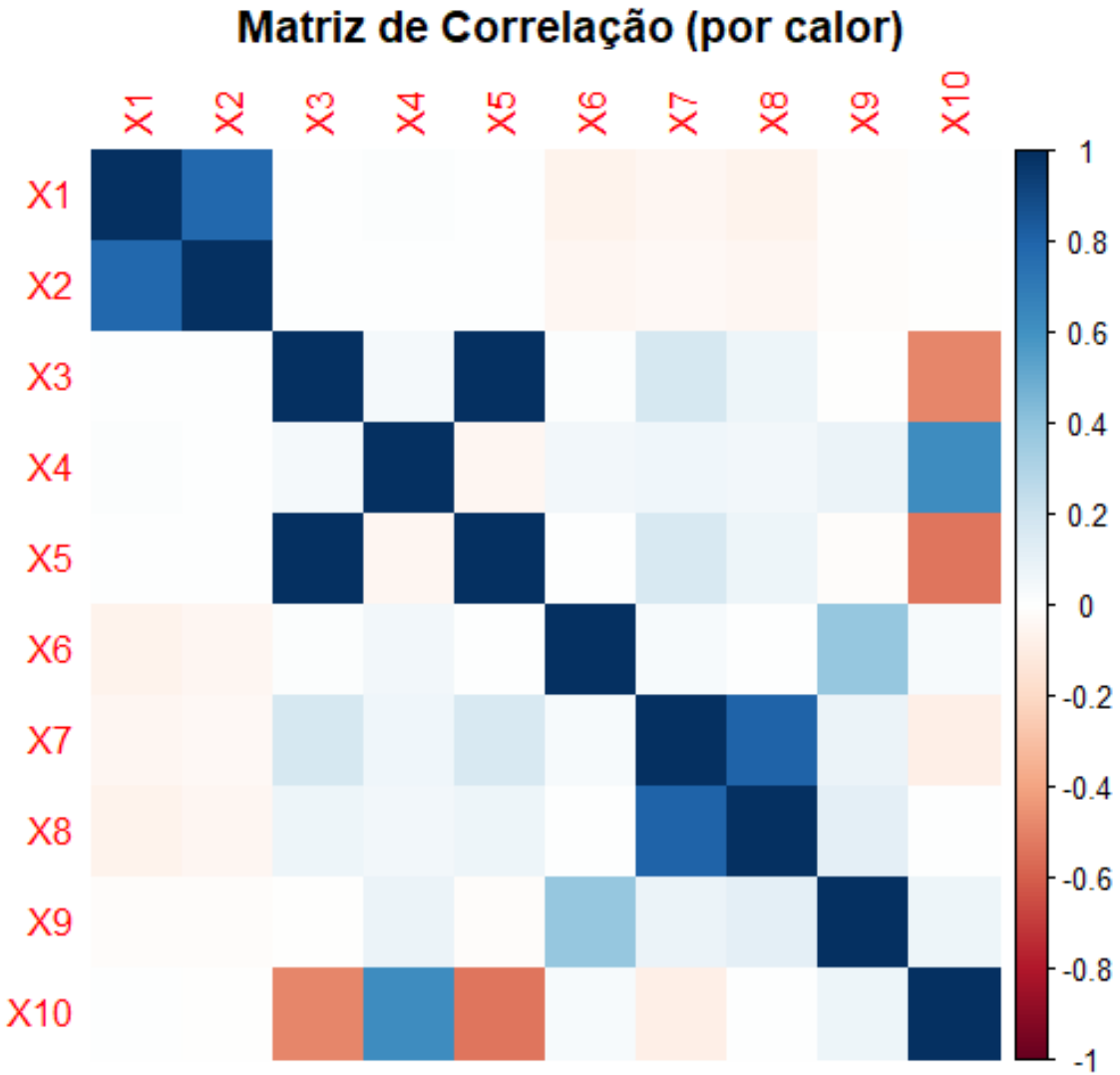

Figura: Matriz de Correlação - Método de calor

Analise de Componentes Principais

Analise de Componentes Principais

- A análise de componentes principais (ACP) consiste em estudar a estrutura de interdependência das covariáveis observadas em um conjunto de dados, com o objetivo de obter combinações lineares dessas para a construção de componentes principais (novas variáveis) que expliquem a maior variabilidade total dos dados.
- Assim, é possível reduzir a dimensão dos dados, o que pode facilitar na análise e interpretação dessas interdependências.

Analise de Componentes Principais

- Como já visto anteriormente: $\text{tr}(S) = 177451791$.
- Foi possível notar também, que as variabilidades são bastante distintas e que as unidades de medidas não são as mesmas, por isso, para obtermos uma padronização, calculou-se os autovalores da ACP a partir da matriz de correlação dos dados originais.

Tabela: Autovalores da ACP

	Autovalores	Porcentagem da Variância	Proporção Acumulada
Componente 1	2.51	25.14	25.14
Componente 2	1.93	19.31	44.45
Componente 3	1.72	17.23	61.69
Componente 4	1.38	13.77	75.46
Componente 5	1.23	12.34	87.8
Componente 6	0.60	6.09	93.9
Componente 7	0.22	2.23	96.14
Componente 8	0.21	2.11	98.26

Analise de Componentes Principais

Sendo assim, a partir dos autovalores, foram calculados seus respectivos autovetores:

Tabela: Autovetores da ACP

	Autovetor 1	Autovetor 2	Autovetor 3	Autovetor 4	Autovelor 5
X_1	-0.043757293	-0.5873186	0.73498329	-0.0002692969	-0.08512224
X_2	-0.030657285	-0.5800584	0.73931945	-0.0037678922	-0.09902996
X_3	0.892358539	-0.0755410	0.03273077	0.2945753164	0.31953389
X_4	-0.288905371	0.3141020	0.33678294	0.4026371092	0.68374467
X_5	0.918071781	-0.1036846	0.00253103	0.2584167232	0.25816853
X_6	-0.007823903	0.2620676	0.05950260	0.6481576430	-0.46223257
X_7	0.380673723	0.6463384	0.46196665	-0.3461654607	-0.07679890
X_8	0.279913101	0.6794745	0.45805952	-0.3867379536	-0.09735273
X_9	-0.020324134	0.3205309	0.19675242	0.5865052208	-0.45313654
X_{10}	-0.751822842	0.2897093	0.23811128	0.1687827493	0.38269612

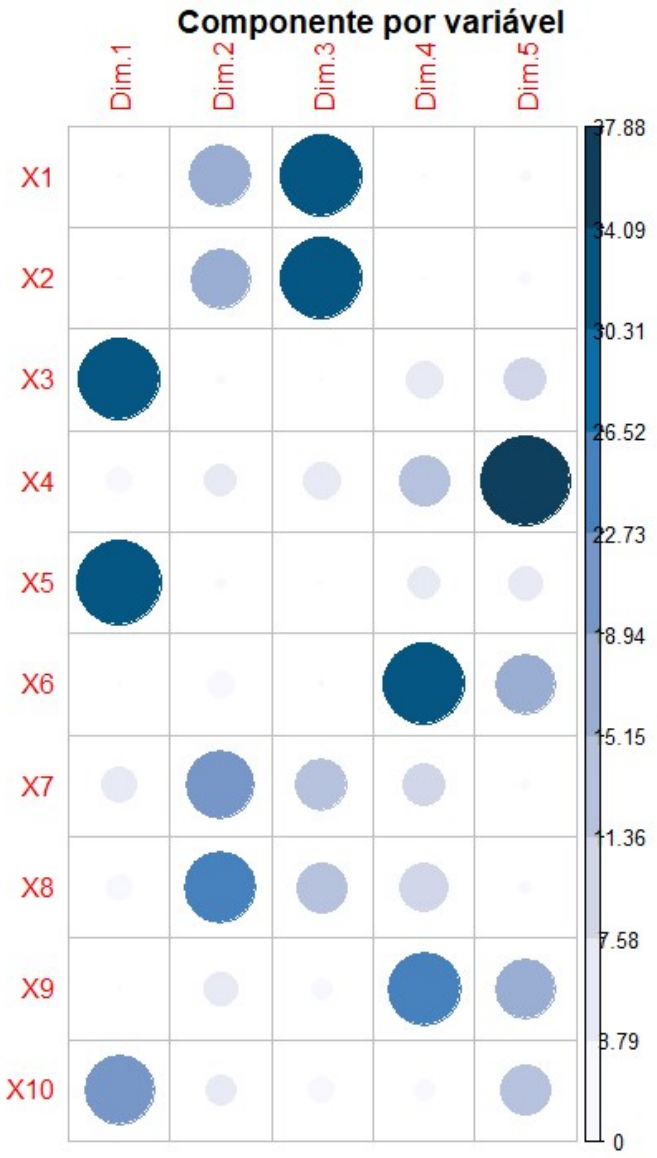

Figura: Contribuição da Variável em cada Componente

- Com isso, podem ser criados tipos de "indicadores", nomeando os componentes de acordo com as variáveis que mais impactam neste:

Componente 1: *Score do Cartão de Crédito*
$\mathbf{X}_3$ **:** Limite de crédito no cartão de crédito (U$); $\mathbf{X}_5$ **:** Linha de crédito aberta para compra (média dos últimos 12 meses); $\mathbf{X}_{10}$ **:** Taxa de utilização média do cartão.

Componente 2: *Atividade em Transações*
$\mathbf{X}_7$ **:** Valor total da transação (últimos 12 meses); $\mathbf{X}_8$ **:** Contagem total de transações (nos últimos 12 meses);

Componente 3: *Maturidade do cliente*
$\mathbf{X}_1$ **:** Idade do cliente (em anos); $\mathbf{X}_2$ **:** Período de relacionamento com banco (em meses);

Componente 4: *Mudanças nas operações*
$\mathbf{X}_6$ **:** Mudança no valor da transação (Q4 sobre Q1); $\mathbf{X}_9$ **:** Mudança na contagem de transações (Q4 sobre Q1);

Componente 5: *Saldo Rotativo no Cartão*
$\mathbf{X}_4$ **:** Saldo rotativo total no cartão de crédito (U$);

Analise de Componentes Principais

Tabela 3.3: Exemplo comparativo de uma observação

Originais			**ACP**		
Variáveis	Indivíduo A	Amplitude	**Componentes**	Indivíduo A	Amplitude
X_3	34516	1438 a 34516	**Score do**	2.75	-2.89 a 5.36
X_5	32252	3 a 34516	**Cartão de Crédito**		
X_{10}	0.06	0 a 1			
X_7	1330	510 a 18484	**Avidade**	-1.05	-4.6 a 5.8
X_8	31	10 a 139	**em Transações**		
X_1	51	26 a 73	**Maturidade**	0.75	-4.76 a 4.89
X_2	46	13 a 56	**do cliente**		
X_6	1.97	0 a 3.39	**Mudanças**	5.48	-3.5 a 12.4
X_9	0.7	0 a 3.71	**nas operações**		
X_4	2264	0 a 2517	**Saldo Rotativo no Cartão**	-0.26	-10.2 a 4.2

Analise de Componentes Principais

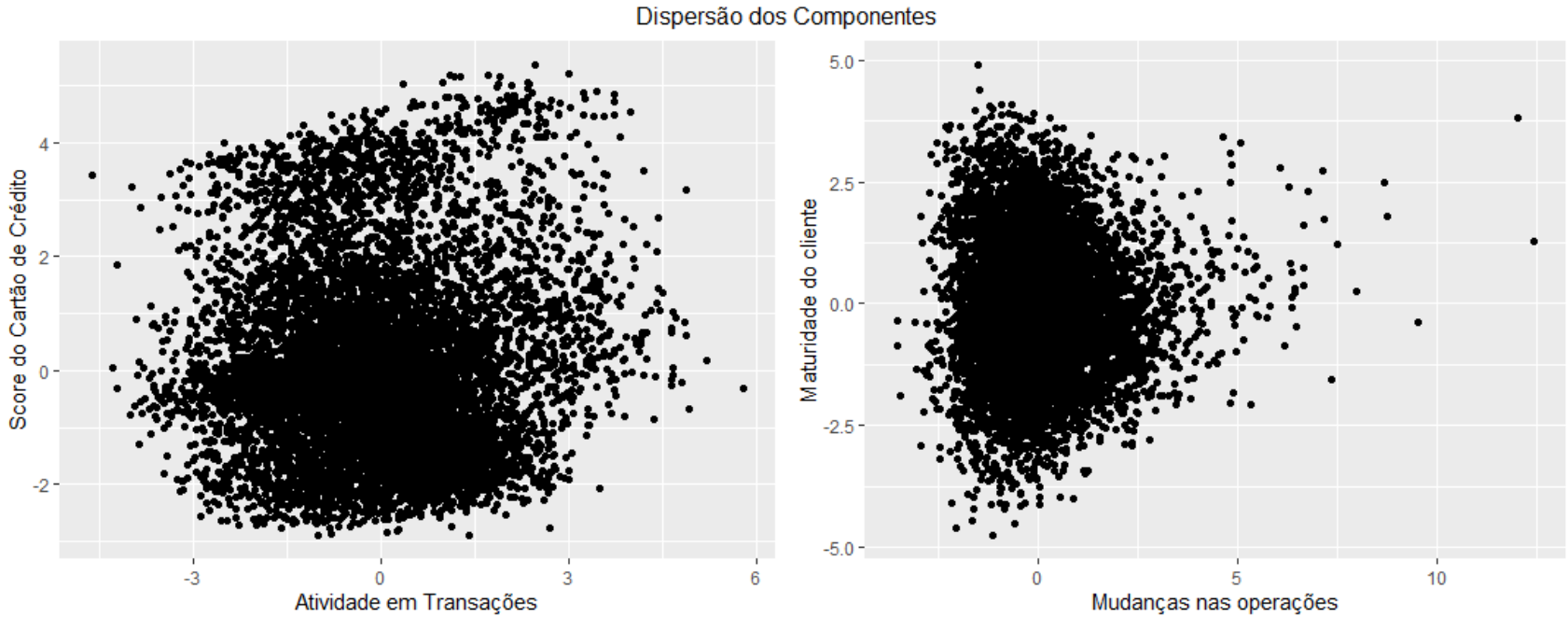

Figura: Gráfico de Dispersão dos componentes

Distância Estatística

Distancia Estatística Generalizada

- Com o intuito de detectar anomalias multivariadas (outliers), criar classificações, e outras possíveis aplicações, calculamos a distancia estatística generalizada, usando os dois métodos diferentes apresentados em aula.
- Em ambos os métodos procuramos entender sobre a **Maturidade do cliente** (componente criado anteriormente) que envolve $\mathbf{X}_1$ e $\mathbf{X}_2$, a fim de entender o comportamento das observações e assim poder direcionar algumas atitudes do Banco.

Distancia Estatística Generalizada

Como citado anteriormente, para o cálculo das distancias utilizamos os dois métodos apresentados em aula (motivo: apenas para comparações). Para essa apresentação, trouxemos a distância estatística generalizada em torno da média, calculada pelo método Mahalanobis:

$$d(Q,P) = \sqrt{(x-\bar{x})' S^{-1} (x-\bar{x})}$$

em que:

- x: Matriz do vetor de covariáveis;
- $\bar{x}$: Vetor de médias da matriz de covariáveis;
- S^{-1}: Matriz inversa de S (Var-Cov).

Resultado

Obtendo então,

```
> d4
  [1] 0.8543742 1.2613592 0.9373127 1.0035884 2.1768460 0.4837354 1.4281303 1.8490160 1.9045956 0.3284256 0.6230462 2.4299253
 [13] 1.9522349 1.5371674 1.5283477 0.6582231 0.3284256 0.8080396 2.5254070 0.4613558 1.1323412 1.9611574 0.7120394 0.1256346
 [25] 0.9572785 1.0113963 1.5810934 2.5181864 0.2908554 1.1323412 1.8619589 1.6210441 1.0926557 1.0541590 1.6585957 1.7492484
 [37] 1.7492484 0.8896752 1.6380362 2.6276803 1.2566668 1.0691380 0.9679268 0.3940874 1.0762146 1.6698562 1.2432668 2.0123948
 [49] 0.8827697 0.7121638 0.7186777 1.0371824 2.5110319 0.5313677 1.9522349 0.5313677 0.5313677 1.7870111 0.3431904 0.6173130
 [61] 1.3864333 0.5026796 1.0804677 0.5170530 1.0122850 0.6681732 2.0550754 2.8877283 1.2731511 0.4875151 0.7607272 1.0925927
 [73] 0.9804365 1.0589153 0.3431904 1.3432776 0.4837354 1.1661669 1.6698562 0.3252254 0.1387770 0.4837354 1.1754613 2.0123948
 [85] 1.5084835 1.5219092 1.5084835 0.5385864 0.6951788 1.3387024 0.9373127 0.3336099 0.2808065 0.8493612 1.4128791 2.2489742
 [97] 0.1670119 1.5084835 1.7870111 0.1387770
 [ reached getOption("max.print") -- omitted 10027 entries ]
```

Figura: Distância estatística generalizada

Uso da distância

Com as distancias obtidas podemos entender quais clientes destacam-se dos demais com relação à Maturidade (idade e período de relacionamento com o banco).

$X1$	$X2$	$Distancia\_Média$	ID
73	36	5.4030	252
44	13	4.3092	3758
43	13	4.1600	7571
26	36	4.1374	614
26	36	4.1374	991
26	36	4.1374	1090
...	...	...	...

Tabela

Tendo então uma classificação dos clientes que se diferem da média, podendo o Banco aprimorar estudos, campanhas, promoções, todo um tratamento especifico para esse perfil de cliente.

Normalidade

Normalidade

Utilizando do Pacote MVN no R, é realizado, a princípio, o teste de Royston, para a Normalidade Multivariada.

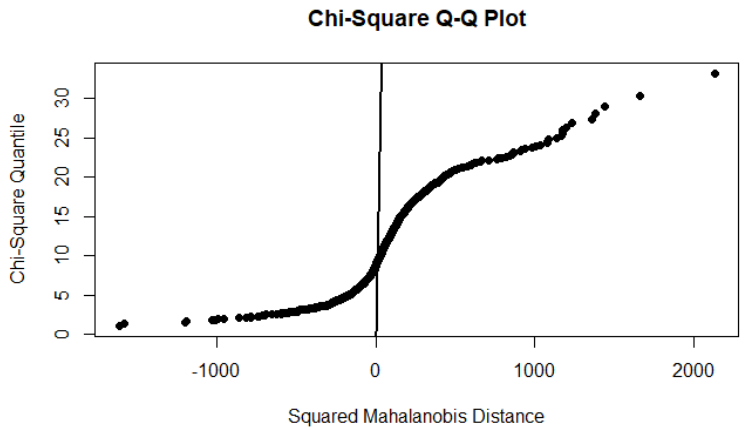

Figura: Gama Plot - Quantis QuiQuadrado para normalidade Multivariada

Como já alertado na própria análise descritiva, a normalidade multivariada não é aceita. No Gama Plot é possível notar que não há uma linearidade em comparação com os quantis teóricos da QuiQuadrado.

Normalidade - Univariada

Afim de entender a não normalidade multivariada, verificamos a normalidade univariada, através dos testes de LillieFors e Anderson-Darling:

Covariavel	LillieFors	Anderson-Darling
X1	0.1241	0.0821
X2	0.0000	0.0002
X3	0.0001	0.0000
X4	0.0001	0.0000
X5	0.0004	0.0001
X6	0.0000	0.0000
X7	0.0014	0,0020
X8	0.0031	0.0007
X9	0.0000	0.0000
X10	0.0000	0.0000

Normalidade

- Verificamos a não normalidade multivariada;
- Verificamos a não normalidade em quase todas as covariáveis;
- Sendo assim, para análises futuras, como a base de dados possui 10 mil observações, indica-se o uso do TLC; ou realizar transformações para termos então dados multivariados que seguem distribuição normal multivariada.

Conclusão

Conclusões

- Para possibilitar uma análise multivariada e facilitar algumas interpretações, criou-se 5 Indicadores, através da técnica de Componentes Principais: Score do Cartão de Crédito, Atividade em Transações, Maturidade do cliente, Mudanças nas operações e o Saldo Rotativo no cartão.
- Utilizamos a Distancia estatística generalizada para realizar uma classificação de clientes em relação ao componente "Maturidade do Cliente"e assim possibilitar atitudes direcionadas;
- Por fim, na tentativa de preparar a base de dados para futuras análises que venham a ser apresentadas no decorrer da disciplina, o grupo concluiu a não normalidade multivariada dos dados, mas indica o uso de outros formatos e técnicas para contornar a situação.

Obrigado !

Perguntas?