Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Departamento de Estatística

Tratamento Multivariado de dados de Cartão de Crédito

Grupo 2: Antônio M. dos Santos Jr. - 744845
Douglas de Paula Nestlehner - 752728

Outubro, 2021

Sumário

Capítulo 1

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, será abordado no presente relatório: 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.

Para isso, dar-se-a o uso da base de dados: "Clientes de Cartão de Crédito"retirada da plataforma Kaggle (`kaggle.com/sakshigoyal7/credit-card-customers`).

1.1 Banco de dados

Essa base refere-se a informações de 10127 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.

Para a realização desse estudo, as covariáveis categóricas do banco de dados foram separadas. Logo, serão analisadas 10 covariáveis contínuas, quais estão especificadas abaixo. E para essas análises iremos utilizar como ferramenta/linguagem o software estatístico R.

Variáveis 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 (Q4 sobre Q1);
$\mathbf{X}_7$ **:** 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 (Q4 sobre Q1);

$\mathbf{X}_{10}$ **:** Taxa de utilização média do cartão.

1.2 Objetivo

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

Capítulo 2

Análise Descritiva dos Dados

2.1 Análise Univariada

Como citado na Introdução, a base de dados contém 10127 observações. Com isso, na Tabela 2.1 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 2.1: Base de dados

Para facilitar a análise, foram construídos gráficos e calculadas algumas medidas descritivas: mínimo; 1º quartil; mediana; média; 3º quartil; e máximo, de todas as covariaveis em estudo:

Tabela 2.2: Medidas descritivas das 10 variáveis 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

Através das medidas descritivas, já é possível fazer algumas observações:

- Nas variáveis x3, x5 e x7: o valor do 3º quartil é bem menor em relação ao valor máximo observado, indicativo de presença de outliers;
- Nas variáveis x4, x5 e x7 o valor mínimo observado é bem menor do que o valor máximo, indicando grande amplitude;
- Nas variáveis x3, x5, e x7: O valor médio é bem maior do que as demais, mas isso indica apenas que são variáveis que assumem valores maiores, não podendo compará-las pois são distintas e não padronizadas ainda.

No mais, primeiramente analisa-se estas variáveis de maneira individual:

2.1.1 X_1 : Idade do cliente (em anos)

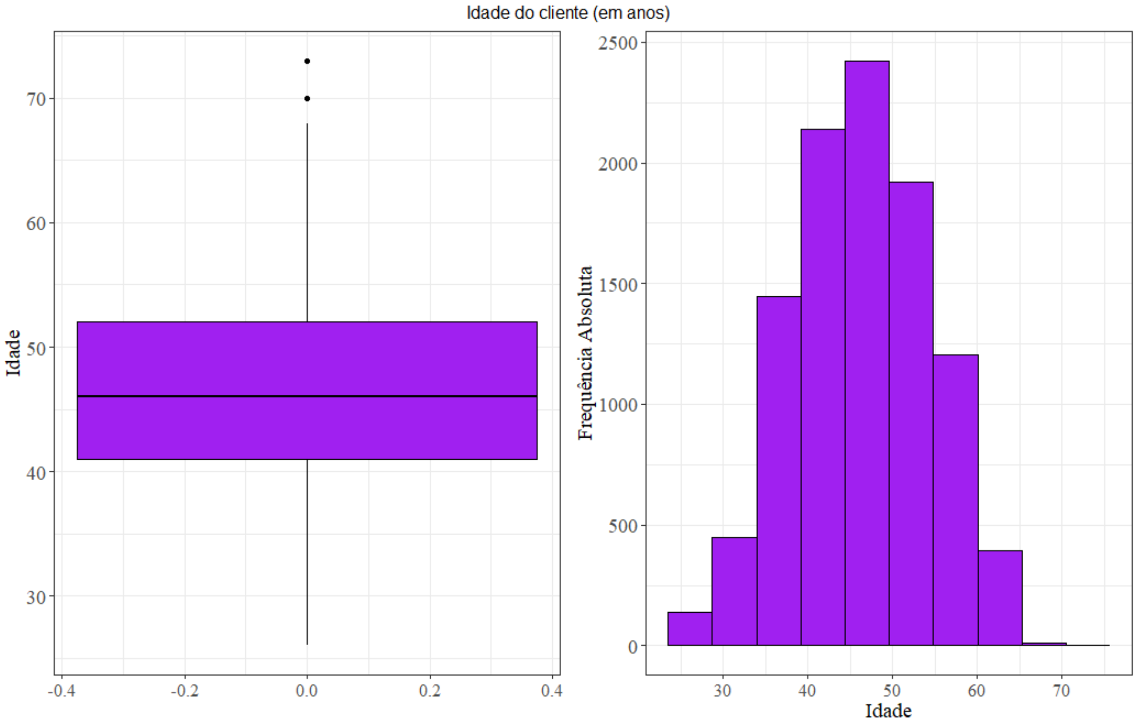

Figura 2.1: Boxplot e Histograma da variável X1

A idade do cliente é distribuída conforme os gráficos, com mediana de 46 anos e média de 46,33.

Nota-se uma simetria nos dados, dando indícios de que tais são normalmente distribuídos.

No mais, a menor idade dos clientes é 26 anos e têm-se duas idades discrepantes acima de 70, sendo 73 a idade máxima.

2.1.2 X_2 : Período de relacionamento com o Banco (em meses)

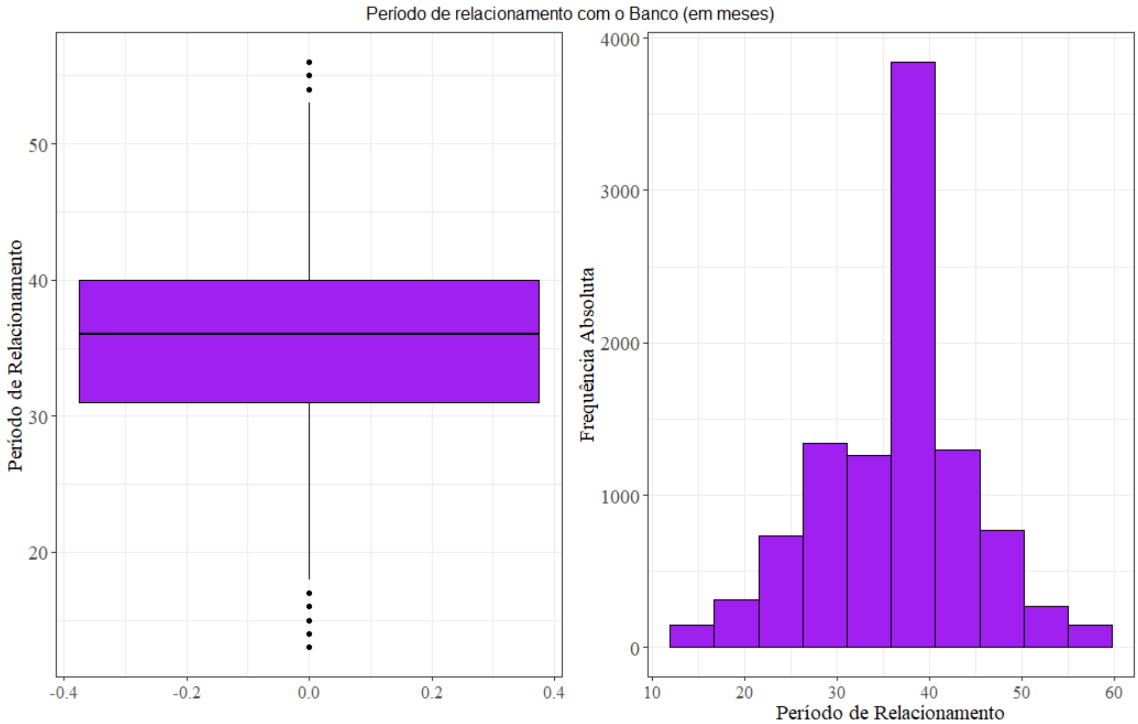

Figura 2.2: Boxplot e Histograma da variável X2

Relacionado em meses, o período de relacionamento dos clientes com o Banco é distribuído conforme os gráficos, com mediana de 36 meses e média de 35,93.

Percebe-se uma grande concentração de clientes com período de relacionamento próximo a 40 meses, mas no geral nota-se uma simetria nos dados, dando indícios de que tais são normalmente distribuídos.

São clientes que possuem de 13 (mínimo) a 56 (máximo) meses de relacionamento com o Banco.

2.1.3 X_3 : Limite de crédito no cartão de crédito (U$)

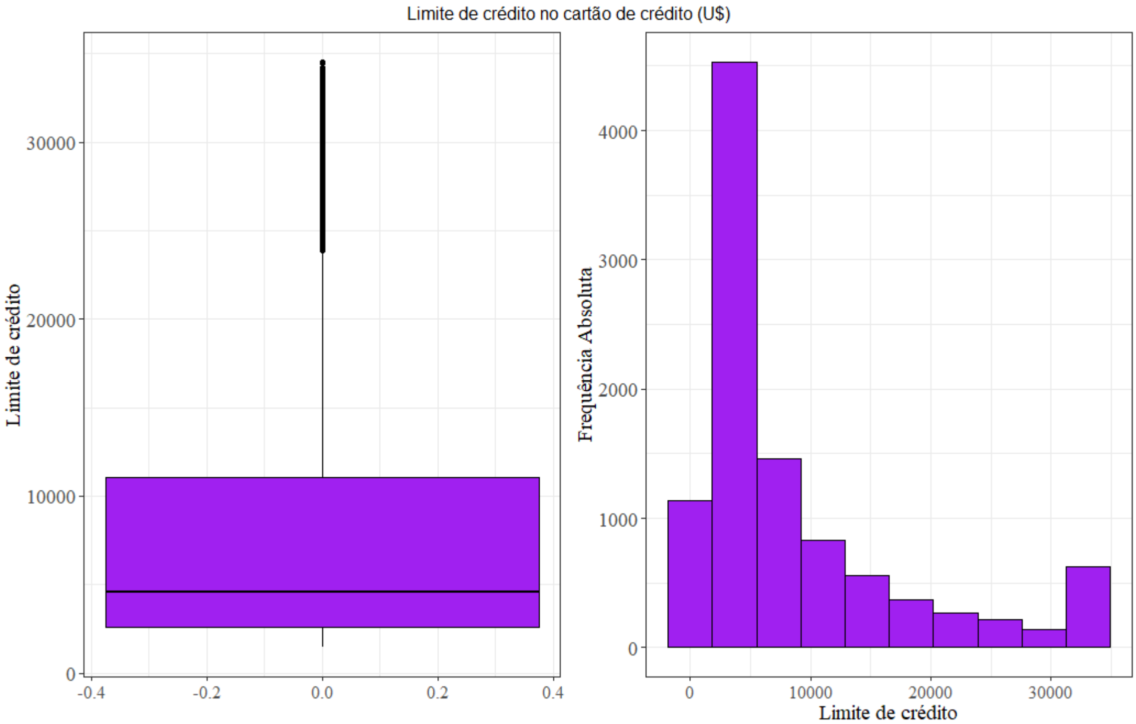

Figura 2.3: Boxplot e Histograma da variável X3

Quanto ao limite de crédito no cartão, temos um comportamento curioso.

Nota-se uma concentração dos dados próximo a 5000 dólares de limite, com uma assimetria à direita. Além disso, há uma quebra na suavidade dessa assimetria, pois existe um concentração forte e discrepante acima de 30000 dólares de limite, o que provavelmente faça com que a hipótese de normalidade seja rejeitada.

No mais, o menor limite liberado no cartão é de 1438 dólares, e o maior é de 34516 dólares

2.1.4 X_4 : Saldo rotativo total no cartão de crédito (U$)

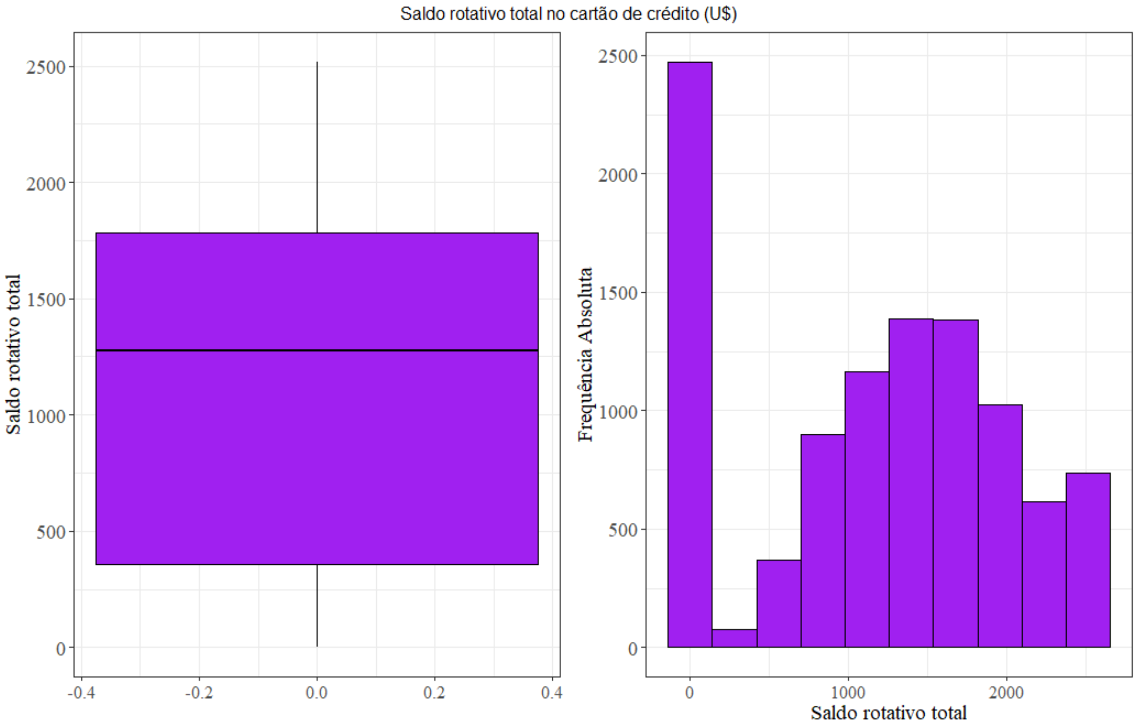

Figura 2.4: Boxplot e Histograma da variável X4

O saldo rotativo no cartão de crédito observado, varia de 0 a 2517 dólares, com mediana de 1276 e média de 1163.

Desconsiderando os clientes com saldo rotativo igual a zero, nota-se uma simetria nos dados em torno de 1500 dólares. Entretanto, a grande quantidade de usuários que não possuem saldo rotativo, pode vir a causar uma rejeição na normalidade desta variável.

2.1.5 X_5 : Linha de crédito aberta para compra (média dos últimos 12 meses)

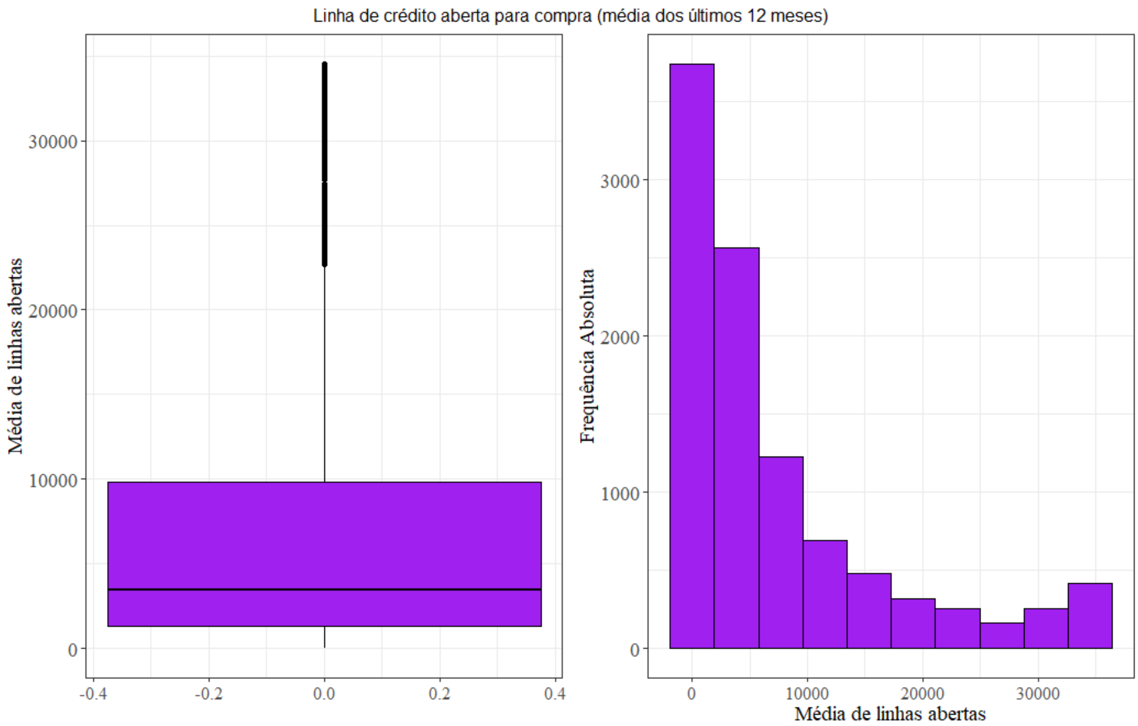

Figura 2.5: Boxplot e Histograma da variável X5

A média de linha de crédito aberta para compra durante um ano é distribuída conforme os gráficos a cima, com mediana de 3474 e média de 7469, quais se destoam bastante considerando a amplitude da variável que vai de 3 a 34516.

Essa diferença é dada por conta de seu comportamento assimétrico à direita e grande concentração próxima a zero. Sendo assim, é nítida a não normalidade dessa variável.

2.1.6 X_6 : Mudança no valor da transação (Q4 sobre Q1)

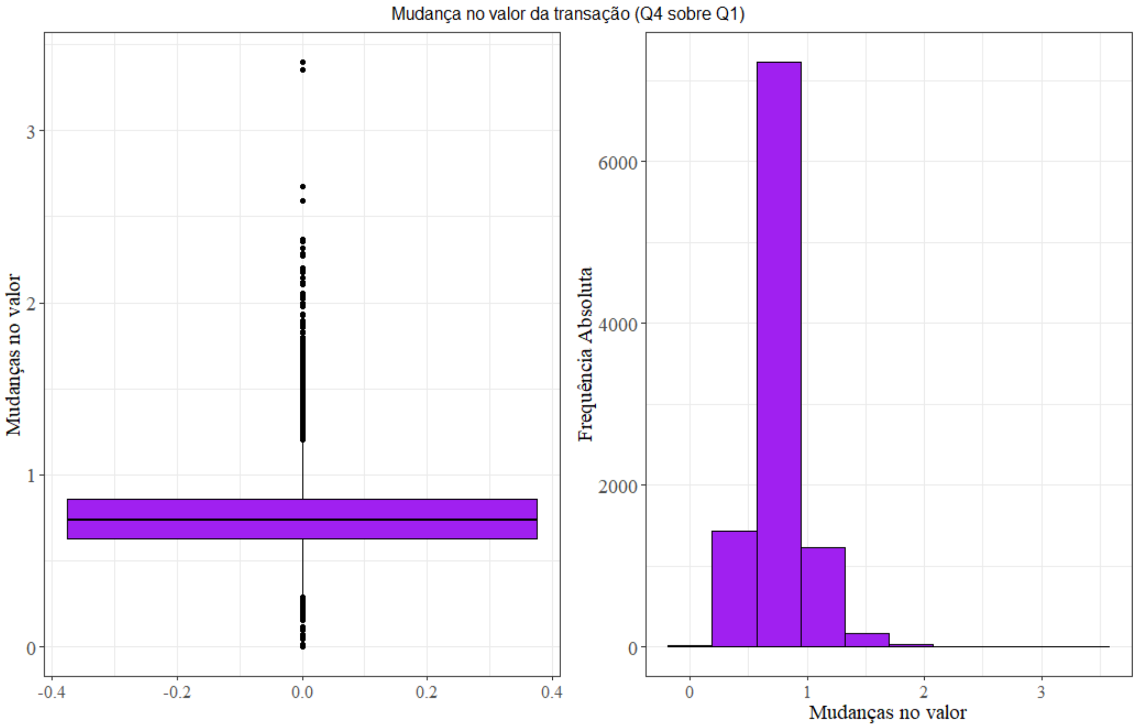

Figura 2.6: Boxplot e Histograma da variável X6

Tal variável apresenta-se com uma enorme concentração entre 0,5 e 1, o que acaba por resultar muitos valores discrepantes. Mas, sua média e mediana são próximas, respectivamente 0,76 e 0,736, e a distribuição dos valores, por mais que concentrados, aparentam uma considerável simetria tendendo a aceitar a hipótese de normalidade.

No mais, em se tratando de amplitude, essa variável assume valores de 0 a 3,397.

2.1.7 X_7 : Valor total da transação (últimos 12 meses)

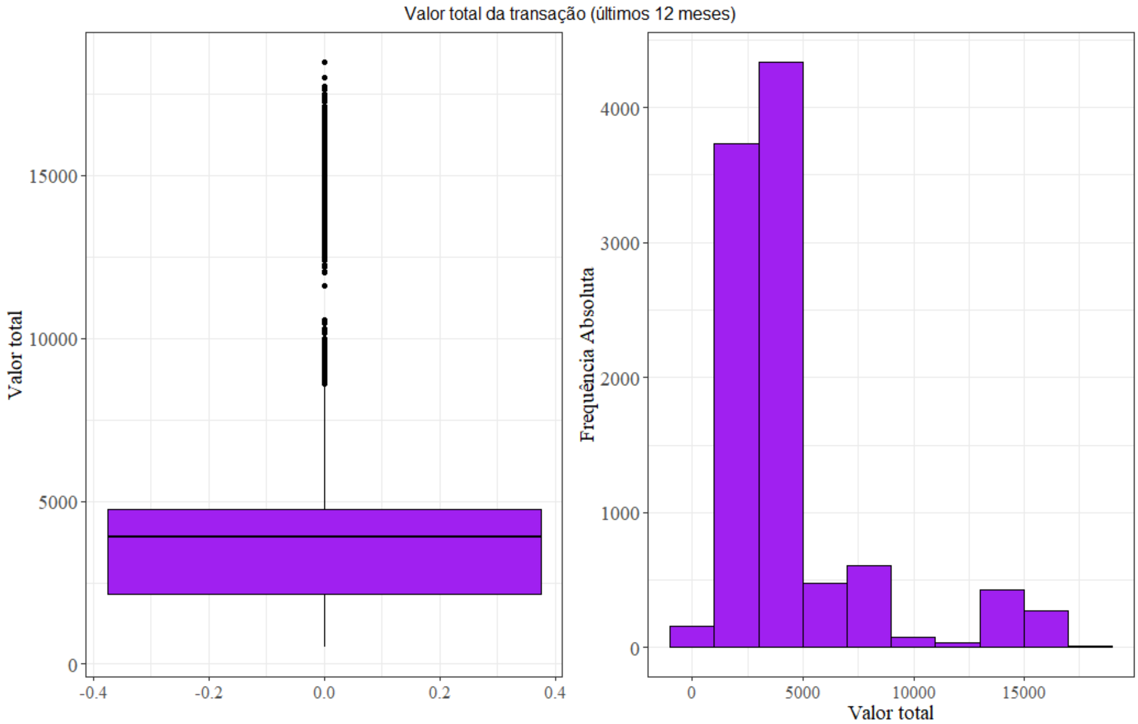

Figura 2.7: Boxplot e Histograma da variável X7

O valor total das transações dos clientes no último ano, é distribuído também de forma curiosa.

Tem-se uma enorme concentração entre 1000 e 5000 dólares, e uma certa quantidade distribuída de forma não homogênea até 18484 dólares (valor máximo observado).

Essa assimetria à esquerda tende a quebrar a hipótese de normalidade da variável, que será futuramente testada.

2.1.8 X_8 : Contagem total de transações (nos últimos 12 meses)

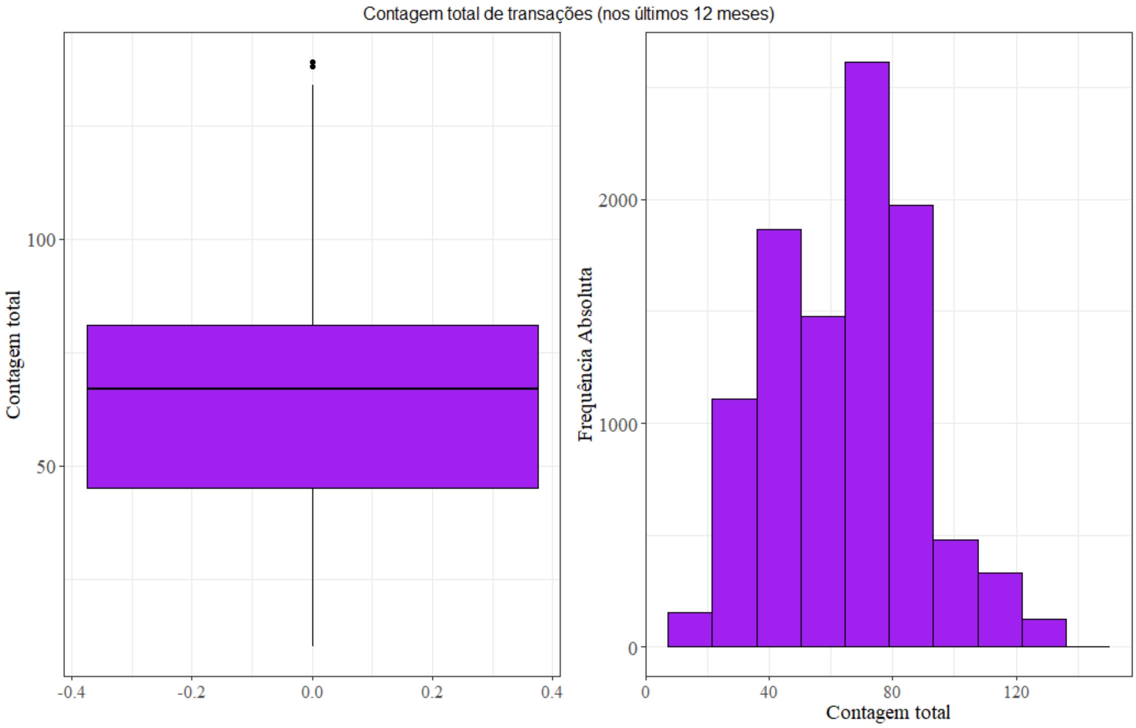

Figura 2.8: Boxplot e Histograma da variável X8

A contagem total de transações dos clientes no último ano é distribuída conforme os gráficos, com mediana de 67 e média de 64,86 transações.

Nota-se uma certa simetria nos dados, não tão forte, mas nos dá indícios de que tais são normalmente distribuídos.

No mais, o menor número de transações realizadas é igual a 10 e o usuário mais ativo nessa variável, fez 139 transações.

2.1.9 X_9 : Mudança na contagem de transações (Q4 sobre Q1)

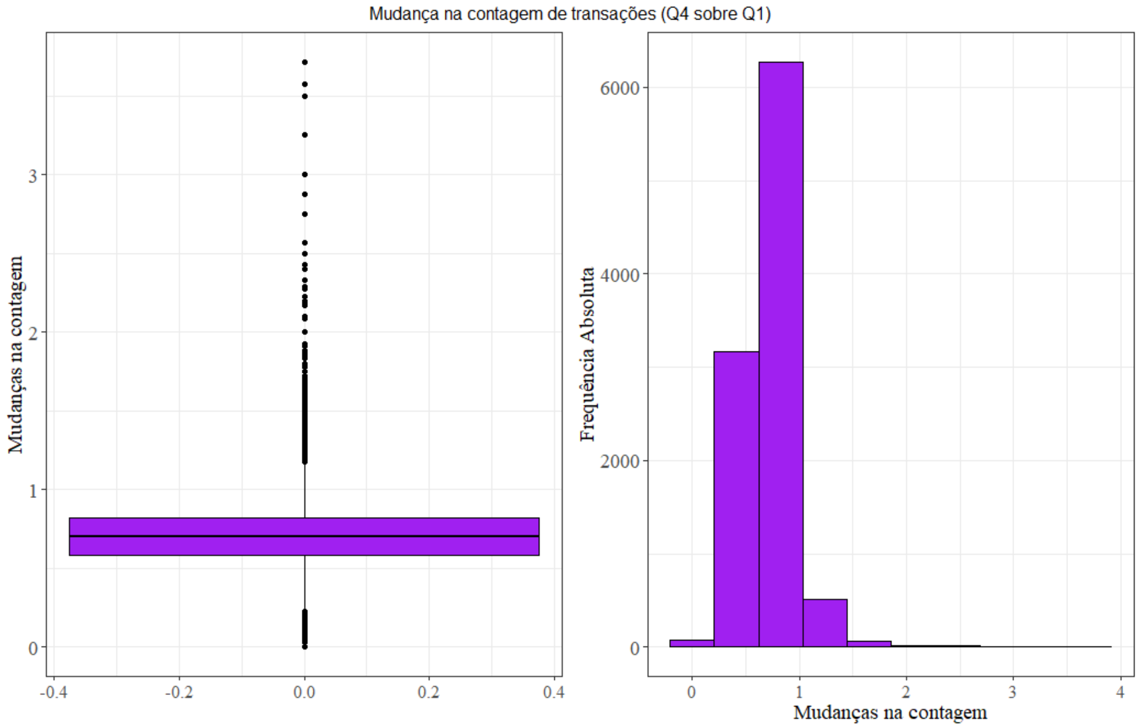

Figura 2.9: Boxplot e Histograma da variável X9

Parecido com X_6, essa variável apresenta-se com uma enorme concentração de observações entre 0,2 e 1, o que acaba por resultar muitos valores discrepantes à esquerda, e alguns à direita também. Mas, sua média e mediana são próximas, respectivamente 0,71 e 0,70, e a distribuição dos valores, por mais que concentrados, aparentam uma considerável simetria tendendo a aceitar a hipótese de normalidade que será testada posteriormente.

No mais, em se tratando de amplitude, essa variável assume valores de 0 a 3,714.

2.1.10 X_{10} : Taxa de utilização média do cartão.

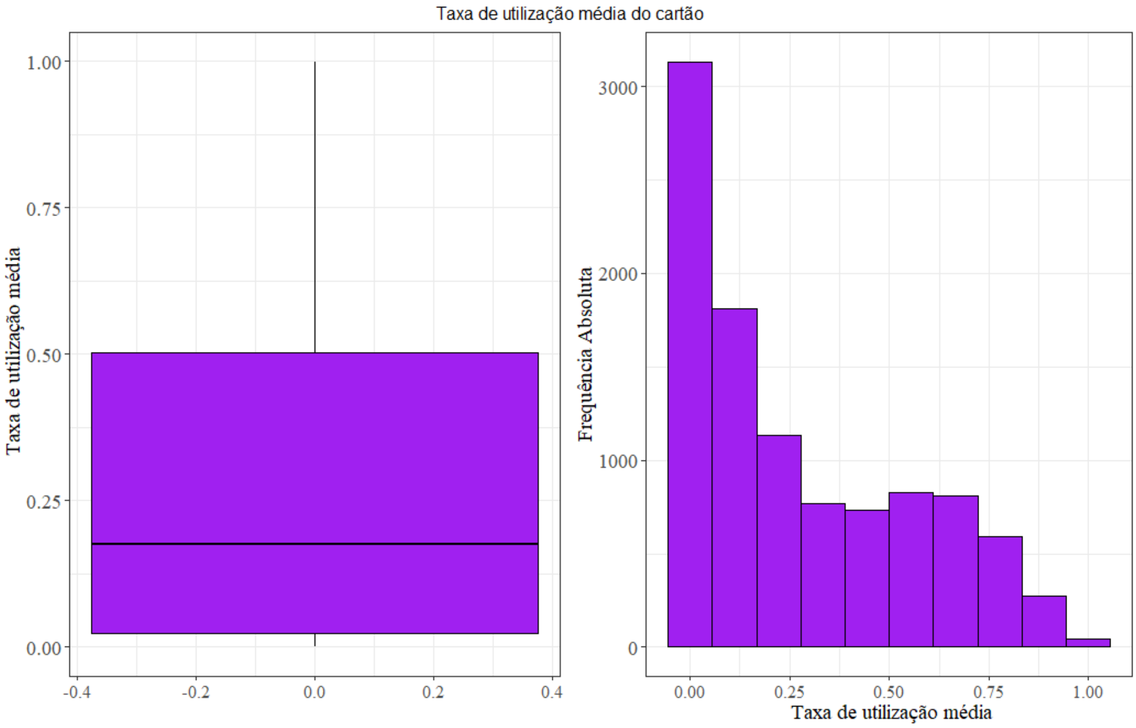

Figura 2.10: Boxplot e Histograma da variável X10

A taxa de utilização média do cartão durante o ano é distribuída conforme os gráficos, com mediana de 0,176 e média de 0,275, quais se destoam consideravelmente dado a amplitude da variável, que vai de 0 a 1.

Essa diferença é dada por conta de seu comportamento assimétrico à direita e grande concentração próxima a zero. Sendo assim, é nítida a não normalidade dessa variável.

2.2 Medidas Descritivas Apresentadas no Curso

2.2.1 Vetor de Médias

Para realizar o cálculo do vetor de médias amostrais foi utilizada a expressão:

$$\bar{x}_p = \sum_{i=1}^{n} \frac{x_{ip}}{n},$$

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

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

Esse vetor nos dá então, informações sobre a média de cada variável, por exemplo: Como já visto anteriormente, a média de idade ($\bar{x}_1$) dos clientes é de pouco mais de 46 anos, a média do período de relacionamento com o banco ($\bar{x}_2$) é de quase 36 meses, a média da taxa de utilização média do cartão ($\bar{x}_{10}$) pelos clientes é de 0,275.

Foi apresentado sua forma de cálculo, pois o mesmo é utilizado para obter a matriz de variâncias e covariâncias das variáveis em estudo, conforme segue.

2.2.2 Matriz de Variâncias e Covariâncias (Var-Cov)

Para calcular a matriz de variâncias e covariâncias amostrais, utilizamos a seguinte expressão,

$$S_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{n},$$

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

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

Tal matriz traz muitas informações, em sua diagonal principal, carrega a variância de cada variável (de X_1 a X_{10}), e fora da diagonal principal apresenta as covariâncias entre tais variáveis.

Consegue-se por exemplos então, visualizar que a variância de idade dos clientes (X_1) é de 64,26 anos, que para essa circunstância pode ser considerada baixa, com desvios de aproximadamente 8 anos em torno da média anteriormente apresentada. Outro exemplo é o que acontece com X_3 (Limite de crédito no cartão do cliente), que apresenta uma variância de 80597704,01 resultando desvios de aproximadamente 9 mil dólares em torno da média (que é de 8.632), ou seja, uma alta variabilidade no limite de crédito.

2.2.3 Variância Total

A variância total é uma forma de sintetização de todas variâncias apresentadas anteriormente, já que esta é a soma das variâncias de todas as variáveis analisadas no banco de dados. Ou seja, é simplesmente a soma das variâncias (listadas na diagonal da matriz S), sendo então representada por:

$$\text{traço}(S) = tr(S) = S_{11} + S_{22} + \cdot + S_{pp}.$$

Nesse contexto, no banco de dados analisado, temos que a variância total é de:

$$tr(S) = 177451791.$$

A soma dessas variabilidades, por si só, no momento, como as variáveis em estudo possuem escalas distintas, não nos dá uma interpretação além da já descrita separadamente. Entretanto, essa soma será útil para comparações futuras, onde pode-se haver interesses em diminuir essa variabilidade através de algum método multivariado.

2.2.4 Variância Generalizada

Outro método de sintetizar a variabilidade multivariada dos dados em um único valor numérico, é através da variância generalizada. Entretanto, essa, além de trazer informações apenas da variância (como a Total), também resume as covariâncias. Tal medida é definida como o determinante da matriz S. Sendo definida como:

$$\det(S) = \sum_{i=1}^{n} (-1)^{i+j} \cdot s_{ij} \cdot \det(S_{ij}),$$

onde $i, j = 1, 2, ..., 10$.

Assim, a partir da matriz Var-Cov gerada anteriormente, temos que:

$$\det(S) = 244049967398863$$

Da mesma forma que na Variância Total, por mais que seja um valor aparentemente grande, a Variância Generalizada, por si só, não nos acrescenta muitas informações além das já vistas, entretanto será útil para comparações futuras onde poderá haver interesses em diminuir o resultado desse determinante através de algum método multivariado.

2.2.5 Matriz de Desvios Padrão

Calculamos a matriz de desvio padrão utilizando as informações apresentadas em aula:

$$D^{1/2} = \begin{bmatrix} \sqrt{S_{11}} & 0 & ... & 0 \\ 0 & \sqrt{S_{22}} & & \\ ... & & ... & \\ 0 & ... & & \sqrt{S_{pp}} \end{bmatrix}$$

.

Implementando então para os dados que estão sendo utilizados, obtivemos os seguintes resultados:

$$D^{1/2} = \begin{bmatrix} 8.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & 7.98 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & 9088.32 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & 814.94 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & 9090.23.24 & 0 & 0 & 0 & 0 & 0 \\ & & & & & 0.21 & 0 & 0 & 0 & 0 \\ & & & & & & 3396.96 & 0 & 0 & 0 \\ & & & & & & & 23.47 & 0 & 0 \\ & & & & & & & & 0.23 & 0 \\ & & & & & & & & & 0.27 \end{bmatrix}$$

.

Tendo calculado a matriz de desvios padrão podemos encontrar outros resultados, que podem ser obtidos por meio das seguintes relações:

$$S = D^{1/2} R D^{1/2}$$

$$R = D^{-1/2} S D^{-1/2}$$

Ao realizar essas multiplicações de matrizes, conseguimos obter o mesmo resultado para a matriz de Var-Cov (S) já calculada anteriormente, e também para a matriz de correlação (R) que será calculada na próxima seção.

2.2.6 Matriz de Correlações

Para analisar o grau de relacionamento entre duas variáveis, pode ser calculada a correlação destas. A matriz de correlações nada mais é então que uma tabela que indica os coeficientes de conexão entre as variáveis, onde cada célula mostra a conexão entre duas variáveis respectivas da linha e coluna.

No caso, é utilizada a Correlação Linear de Pearson, dada por:

$$r_{jk} = \frac{S_{jk}}{\sqrt{S_{jj}}\sqrt{S_{kk}}} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}},$$

onde $j, k = 1, 2, ..., 10$.

Nesse contexto, no banco de dados analisado, a matriz de correlações é dada por:

	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$	$\mathbf{X}_5$	$\mathbf{X}_6$	$\mathbf{X}_7$	$\mathbf{X}_8$	$\mathbf{X}_9$	$\mathbf{X}_{10}$
$\mathbf{X}_1$	1.000	0.789	0.002	0.015	0.001	-0.062	-0.046	-0.067	-0.012	0.007
$\mathbf{X}_2$	0.789	1.000	0.008	0.009	0.007	-0.049	-0.039	-0.050	-0.014	-0.008
$\mathbf{X}_3$	0.002	0.008	1.000	0.042	0.996	0.013	0.172	0.076	-0.002	-0.483
$\mathbf{X}_4$	0.015	0.009	0.042	1.000	-0.047	0.058	0.064	0.056	0.090	0.624
$\mathbf{X}_5$	0.001	0.007	0.996	-0.047	1.000	0.008	0.166	0.071	-0.010	-0.539
$\mathbf{X}_6$	-0.062	-0.049	0.013	0.058	0.008	1.000	0.040	0.005	0.384	0.035
$\mathbf{X}_7$	-0.046	-0.039	0.172	0.064	0.166	0.040	1.000	0.807	0.086	-0.083
$\mathbf{X}_8$	-0.067	-0.050	0.076	0.056	0.071	0.005	0.807	1.000	0.112	0.003
$\mathbf{X}_9$	-0.012	-0.014	-0.002	0.090	-0.010	0.384	0.086	0.112	1.000	0.074
$\mathbf{X}_{10}$	0.007	-0.008	-0.483	0.624	-0.539	0.035	-0.083	0.003	0.074	1.000

Para visualização mais clara, a matriz foi construída também em sistema de cores, onde o termômetro ao lado direito indica as correlações, sendo tonalidades vermelhas as correlações negativas, e azuis as positivas. Lembrando que a matriz é simétrica, percebe-se que além da diagonal principal, são poucas as cores fortes (correlações altas).

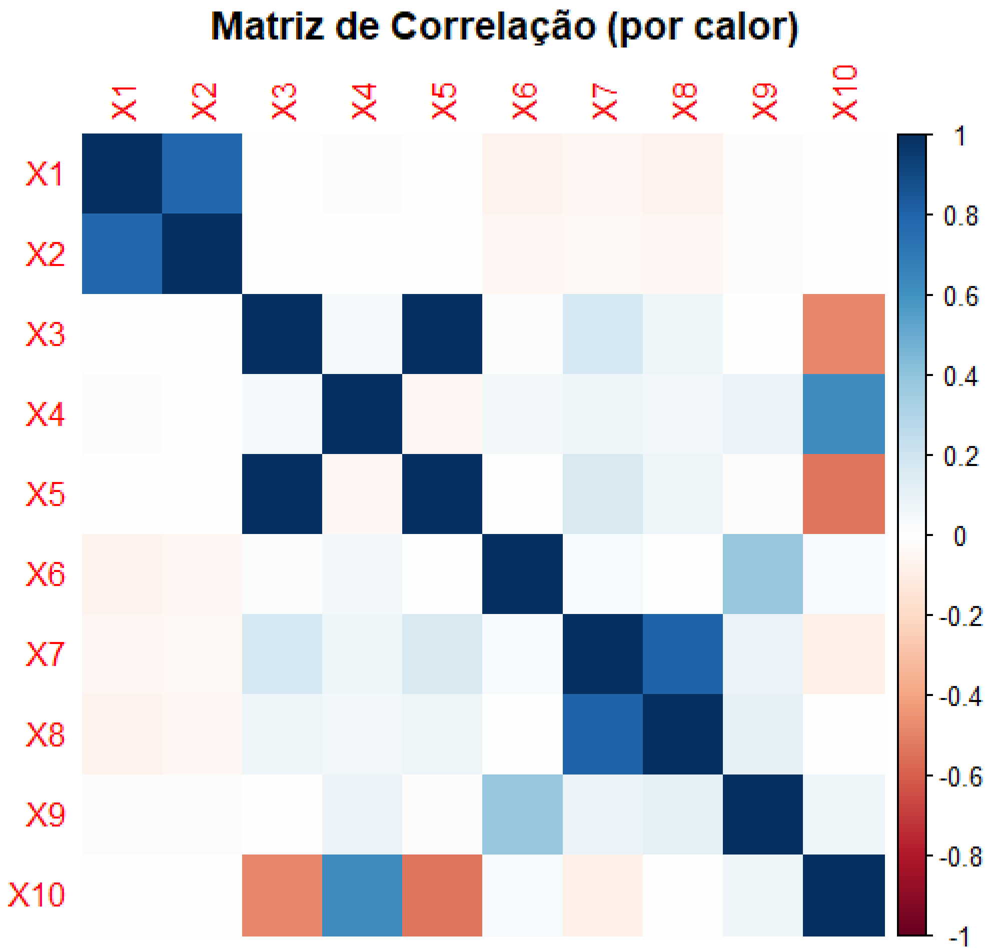

Figura 2.11: Matriz de Correlação - Método de calor

Algumas observações em que podemos fazer olhando na matriz de correlação ou no grafico de calor:

- Podemos observar na diagonal principal as correlações da cada covariavel com ela mesma, o que é igual a 1;
- Forte correlção entre as covariaveis x1 e x2: Ou seja, a covariavel "Idade do cliente"tem 78.9% de correlação com a covariável "Periodo de relacionamento com banco";
- Forte correlação entre as covaraiveis x3 e x5: Ou seja, a covaraivel "Limite de crédito no cartão"tem 99.6% de correlação com a covariavel "Linha de crédito aberta para compra"
- Forte correlação entre as covariavel x7 e x8: Ou seja, a covariavel "Valor total de transação"tem 80.7% de correlação entre a covariavel "Contagem total de transações".

Capítulo 3

Técnicas Multivariadas

3.1 Análise de Componentes Principais

A análise de componentes principais (ACP) consiste em estudar a estrutura de interdependência das variáveis 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.

Para realizar a ACP é necessário saber a variância total dos dados, que será a soma das variâncias das variáveis em questão, X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9 e X_{10} .

Depois disso é necessário encontrar os autovalores da matriz de variância e covariância das variáveis, com o intuito de observar qual a porcentagem que cada um explica da variância total. Porém, em alguns casos as variâncias são bastante distintas, umas muito altas e outras baixas, ou as unidades de medidas são diferentes, o que acaba distorcendo o uso adequado da ACP, pois os primeiros componentes são definidos pelas variáveis com maior variabilidade. Para resolver isso, é preciso realizar uma normalização nos dados, calcular a matriz de variância e covariância e, assim, encontrar os autovalores. Ou, de modo alternativo, utilizar a matriz de correlação dos dados originais para encontrar os autovalores.

Após encontrar os autovalores é necessário calcular os autovetores correspondentes, em que eles serão ortogonais e normalizados.

Não há uma regra específica que ajude a decidir quantos componentes serão necessários para realizar uma boa análise dos dados originais, mas existem alguns critérios, como: o critério da raiz latente, em que só são utilizados os componentes principais cujo autovalor é maior que 1; outro critério é proposto por Jollife (2002) que diz que é necessário utilizar um número de componentes que expliquem no mínimo 70% da variância total; entre outros.

É possível também realizar a ACP por meio de novos eixos que expliquem da melhor forma todas as variáveis e que contenham a maior variabilidade dos dados, esses que podem ser identificados por meio do gráfico de dispersão entre as variáveis estudadas.

Um dos novos eixos fará um ângulo de θ graus com o eixo de alguma das variáveis e a projeção de cada ponto do primeiro na variável dará as coordenadas dessas observações com respeito a ele.

A coordenada das observações com respeito ao novo eixo é uma combinação linear das coordenadas (antigas) do ponto com respeito aos eixos originais. Por exemplo:

$$X_1{}^* = X_1 cos(\theta) + X_2 sen(\theta),$$

sendo $X_1{}^*$ o novo eixo e X_1 e X_2 possível variáveis

Variando o ângulo entre X_1 e $X_1{}^*$ é possível ver a porcentagem da variância total explicada pela nova variável. Isso se torna necessário, pois o objetivo é encontrar o valor de θ que maximize a porcentagem, para que a nova variável (componente principal) explique melhor a variância total dos dados do que as variáveis originais.

Caso $X_1{}^*$ não explique toda a variabilidade dos dados, é necessário identificar um segundo eixo que corresponda a uma segunda nova variável e que explique o máximo da variância que não foi explicada por $X_1{}^*$. Se o ângulo entre X_1 e $X_1{}^*$ é θ, o ângulo entre X_2 e $X_2{}^*$ também será θ e a combinação linear para será:

$$X_2{}^* = -X_1 sen(\theta) + X_2 cos(\theta)$$

É interessante ressaltar que a correlação entre as duas novas variáveis é zero, isto é, $X_1{}^*$ e $X_2{}^*$ não são correlacionadas.

Primeiramente, como já apresentado, tem-se a Variância Total dos dados igual a:

$$tr(S) = 177451791.$$

Foi possível notar também, que as variabilidades são bastante distintas e que as unidades de medidas não são as mesmas, por isso calculou-se os autovalores da ACP a partir da matriz de correlação dos dados originais.

Tabela 3.1: 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

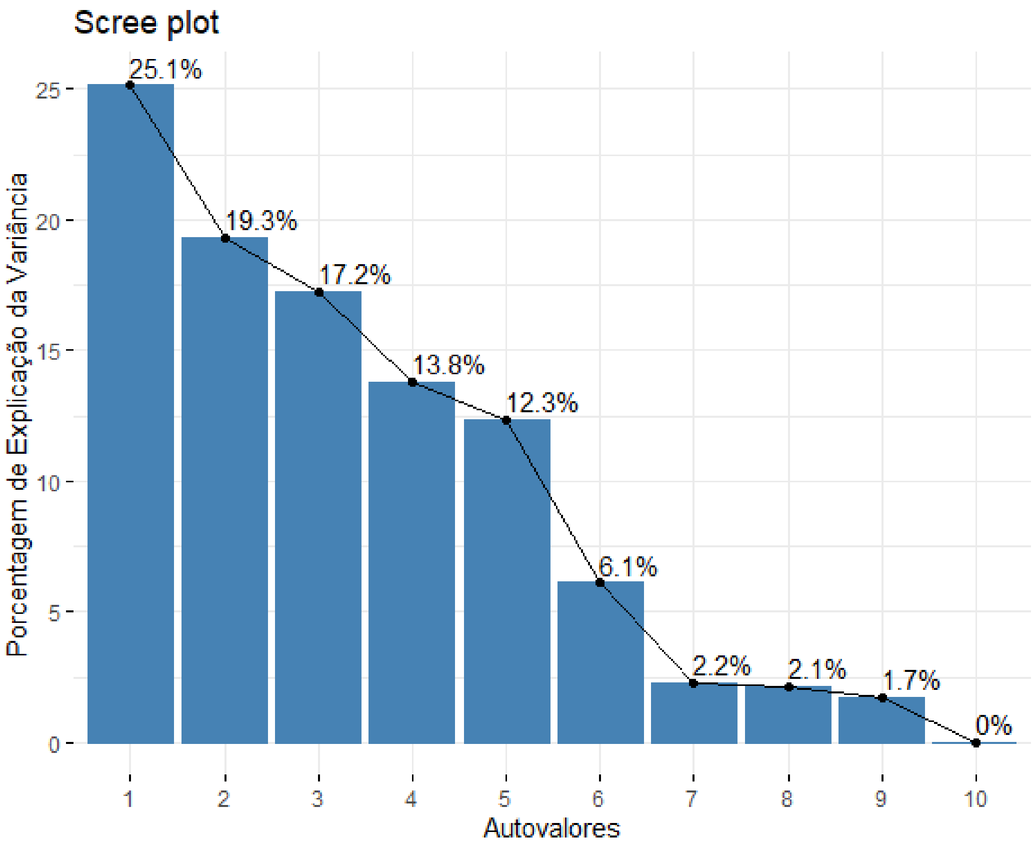

Figura 3.1: Gráfico de Cotovelo dos Autovalores da ACP

Percebe-se que o "cotovelo"aparece por volta do sexto/sétimo autovalor, o que indica que os autovalores menores que o quinto já são relativamente pequenos e que possuem valores parecidos, tendendo, a partir desse ponto, a ser paralelo ao eixo da abcissa. Nota-se também, na tabela, que a partir do sexto componente, o autovalor passa a ser menor do que 1 (limite recomendado por alguns teóricos). Assim, apenas os cinco primeiros componentes principais resumem de modo efetivo a variabilidade total dos dados com explicação de 87,8% desta.

Entretanto, mesmo com esse diagnóstico, é importante observar o comportamento dos Autovetores, para assim estipular a quantidade de componentes usuais que faça sentido.

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

Tabela 3.2: 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

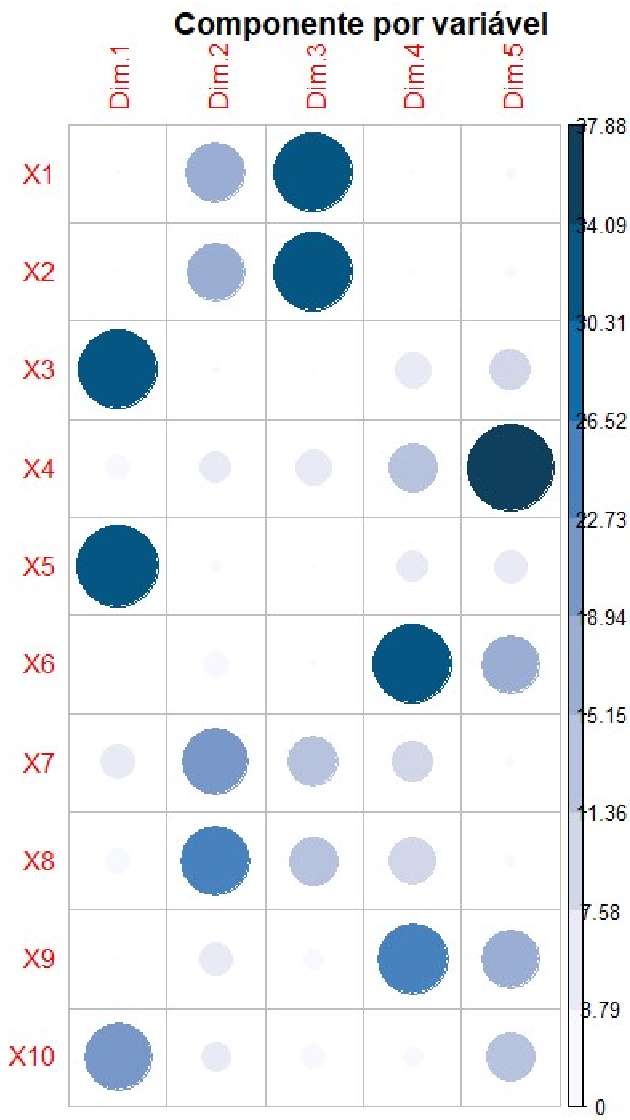

Figura 3.2: Contribuição da Variável em cada Componente

Confirmando o que já havia sido analisado nos autovetores, os autovalores, calculando a distribuição da variabilidade das variáveis perante os componentes, evidencia a representatividade de todas as variáveis.

É notório, por exemplo, a representatividade de X_1 e X_2 no Componente 3, de X_3 no Componente 1, de X_4 no Componente 5, e assim por diante.

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$);

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

Comparando as variáveis com os componentes, nesse indivíduo em específico, conseguimos perceber o comportamento.

No componente **Score do Cartão de Crédito** observa-se valores extremamente altos de X_3 e X_5 e um valor baixo de X_{10}, fazendo com que o valor do componente seja relativamente alto, mas não tão próximo ao máximo.

Já no componente **Maturidade do cliente**, tem-se um cliente de idade, e relação com o Banco, mediana/elevada, resultando também em um valor um pouco a cima da média no componente.

E assim segue a interpretação para as demais variáveis.

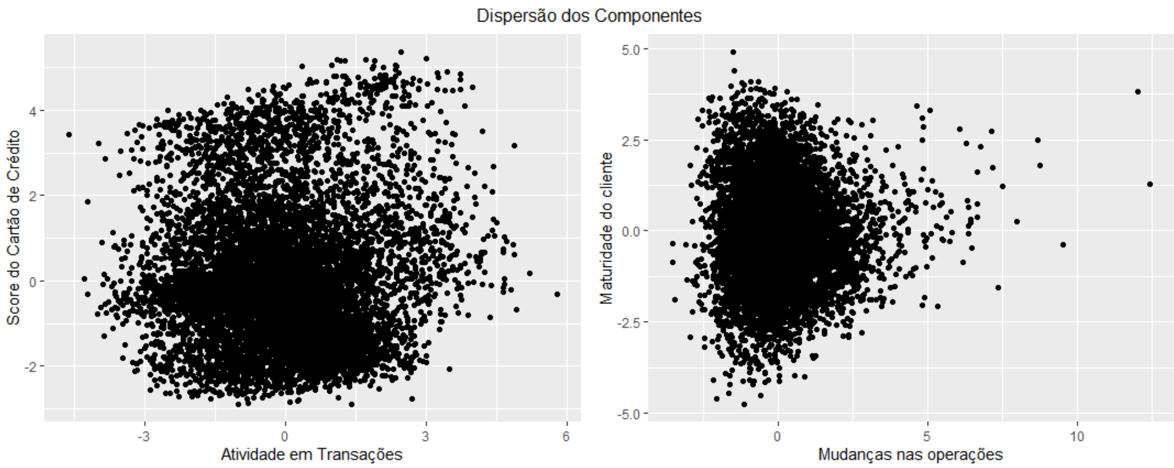

Figura 3.3: Gráfico de Dispersão dos componentes

Nos gráficos temos dois exemplos de dispersão. Entre Componente 1 e Componente 2, e o outro entre Componente 3 e Componente 4. As demais comparações seguem a mesma interpretação: Os Componentes Principais não são correlacionados. Isso acontece pois a ACP agrupa dentro de componentes as variáveis que são correlacionadas, fazendo com que haja correlação inter, mas não entre.

No mais, com relação aos componentes de forma individual, percebe-se que: o **Score do Cartão de Crédito** possui uma grande concentração abaixo de 2, ou seja, uma minoria de clientes porta um Score alto; Já a **Atividade em transações** dos clientes aparenta ser bem distribuída em torno de zero, ou seja, simétrica em torno da média; o componente **Mudança nas operações** apresenta uma assimetria à direita, onde poucos clientes realizam muitas mudanças; e por fim, a **Maturidade do Cliente** também aparenta ser simetricamente distribuída em torno de zero, mostrando uma maturidade geral mediana.

3.2 Distancia Estatística Generalizada

Apos calcular algumas medidas estatísticas (em vetores e matrizes), calcula-se também a distância estatística generalizada (distância centrada na origem e também a distância centrada no vetor de médias) dessas covariáveis. Para isso utiliza-se dois métodos diferentes que foram apresentados no curso, em um deles foram construídos os gráficos da elipse e da rotação desses dados, para poder definir alguns parâmetros.

Em ambos os métodos será utilizado as variáveis $\mathbf{X}_1$ e $\mathbf{X}_2$ a fim de entender o comportamento das observações quanto a **Maturidade do cliente** (componente criado anteriormente) e assim poder direcionar algumas atitudes do Banco.

3.2.1 1º Método

O primeiro método que utilizamos consiste em calcular a distancia de um ponto $P = (\tilde{x}_1, \tilde{x}_2)$ à origem $O = (0, 0)$, através da seguinte formula:

$$d(O, P) = \sqrt{a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2}, \tag{3.1}$$

em que,

$$a_{11} = \frac{\cos^2(\theta)}{S_{11}\cos^2(\theta) + S_{22}\sin^2(\theta) + 2\cos(\theta)\sin(\theta)S_{12}} + \frac{\sin^2(\theta)}{S_{22}\cos^2(\theta) + S_{11}\sin^2(\theta) - 2\cos(\theta)\sin(\theta)S_{12}}$$

$$a_{22} = \frac{\sin^2(\theta)}{S_{11}\cos^2(\theta) + S_{22}\sin^2(\theta) + 2\cos(\theta)\sin(\theta)S_{12}} + \frac{\cos^2(\theta)}{S_{22}\cos^2(\theta) + S_{11}\sin^2(\theta) - 2\cos(\theta)\sin(\theta)S_{12}}$$

$$a_{12} = \frac{\cos(\theta)\sin(\theta)}{S_{11}\cos^2(\theta) + S_{22}\sin^2(\theta) + 2\cos(\theta)\sin(\theta)S_{12}} - \frac{\sin(\theta)\cos(\theta)}{S_{22}\cos^2(\theta) + S_{11}\sin^2(\theta) - 2\cos(\theta)\sin(\theta)S_{12}}$$

Contudo, é necessário encontrar qual o θ mais adequado para ser utilizado ao longo dos cálculos. Nesse sentido, temos uma aproximação de qual será o θ apresentado na Figura 3.4.

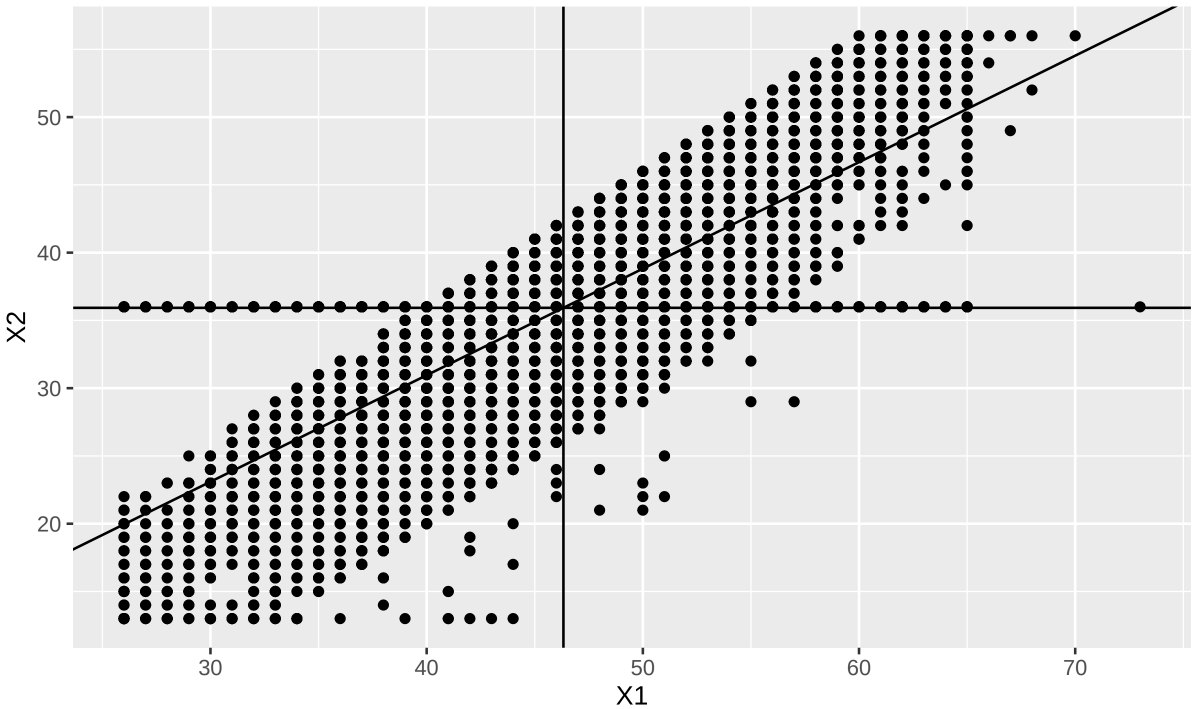

Figura 3.4: Gráfico de dispersão das covariáveis X_1 e X_2 com reta traçada

Nesse sentido, analisando a Figura 3.4, podemos escolher um θ que tenha 35º graus, conseguimos obter os valores:

$$a_{11} = 0.02597891;$$
$$a_{22} = 0.04366909;$$
$$a_{12} = -0.02430169;$$

Com isso utilizando a formula 3.1 calculamos a distancia generalizada d(0, P) do par de covariaveis X1 e X2, a qual esta representada na Figura 3.5

```
> options(max.print = 200)
> d
  [1] 5.807672 6.490751 5.910202 5.093833 4.471855 5.468324 6.778748 4.055102
  [9] 5.236466 5.697712 4.950879 8.154883 6.330924 4.482500 7.213892 5.563440
 [17] 5.697712 5.137987 8.220990 5.609426 6.202953 7.553500 5.046367 5.634327
 [25] 6.523368 4.702174 7.136303 8.281537 5.297735 6.202953 5.960126 7.066920
 [33] 5.341078 6.176223 7.356069 6.240726 6.240726 5.379000 7.314762 8.394780
 [41] 6.028417 6.561138 6.418268 5.834418 4.476173 5.498235 6.886165 7.707052
 [49] 5.950774 6.246799 6.327110 6.375089 8.202486 5.764907 6.330924 5.764907
 [57] 5.764907 7.386010 5.379648 5.186776 7.029332 5.629954 4.951466 5.155460
 [65] 6.646431 6.036222 7.643710 4.999184 5.557195 5.971518 6.363568 6.583668
 [73] 6.606180 5.484536 5.379648 6.069157 5.468324 4.487696 5.498235 5.799996
 [81] 5.713942 5.468324 6.835405 7.707052 6.025789 6.922814 6.025789 5.327435
 [89] 5.095701 6.940298 5.910202 5.910503 5.519508 6.466523 5.060001 7.920783
 [97] 5.436140 6.025789 7.386010 5.713942 6.380612 5.137987 5.989403 5.155460
[105] 4.911684 5.159695 6.069157 5.468324 8.249745 6.153429 6.540708 5.379000
[113] 5.379000 6.153429 5.629954 5.764907 7.038930 6.049081 5.764907 6.646431
[121] 5.835781 6.240726 6.954766 6.093969 6.583668 5.634327 7.673754 6.384933
[129] 6.704102 5.835781 5.379000 5.740020 7.026466 6.504920 5.713942 5.664671
[137] 6.069157 5.915384 7.033703 4.491028 5.853292 6.363568 6.082777 6.617716
[145] 7.817692 5.853292 4.961862 7.355247 5.634327 4.582382 4.622450 8.484757
[153] 5.619434 6.093969 6.954766 5.379000 5.853292 5.910503 5.379648 5.240071
[161] 6.721390 6.617716 5.574884 6.093969 6.132787 6.459729 5.634327 7.166957
[169] 5.361631 6.954766 5.519508 6.704102 6.069157 6.634726 5.341078 5.519508
[177] 6.466523 8.438349 6.423900 6.617716 5.359627 5.910202 7.932295 6.744303
[185] 5.634327 5.910503 5.598156 8.076629 5.835781 4.884839 7.117729 4.841617
[193] 7.117729 6.446268 6.561138 5.495515 5.341078 6.553764 8.145316 6.617716
 [ reached getOption("max.print") -- omitted 9927 entries ]
```

Figura 3.5: Distância estatística generalizada do par de covariáveis X_1 e X_2.

3.2.2 2º Método

Uma outra forma de obter a distancia generalizada de um ponto $(\tilde{x}_1, \tilde{x}_2)$ à origem $O = (0, 0)$ é utilizando a seguinte formula:

$$d(O, P) = \sqrt{x' S^{-1} x} \tag{3.2}$$

Em que:

- x': Matriz transposta do vetor das covariáveis;
- S^{-1}: Matriz inversa de S (Var-Cov);
- x: Matriz do vetor das covariáveis.

Esse método é conhecido como distância de Mahalanobis, e assim como o primeiro método apresentado anteriormente, é uma métrica de distancia multivariada que mede a distancia entre um ponto e uma distribuição, com o objetivo de detectar anomalias multivariadas; classificação em conjuntos de dados e entre outras possíveis aplicações.

Nesse contexto, utilizando a fórmula 3.2 chegamos aos resultados apresentados na Figura 3.6:

```
> options(max.print = 200)
> d2
  [1] 5.661820 6.213674 6.415772 5.016773 5.423816 5.496080 6.474909 4.009381
  [9] 4.825941 5.997715 5.254952 8.129688 7.173533 4.395272 7.139938 5.510562
 [17] 5.997715 5.127056 7.784177 5.623063 5.952709 7.733945 5.116703 5.865795
 [25] 6.736439 5.186127 7.359917 7.968871 5.489739 5.952709 6.842221 6.736382
 [33] 5.171841 6.652894 7.268211 7.016785 7.016785 5.274326 7.161436 8.275579
 [41] 5.729373 6.411527 6.282390 6.119894 4.755246 6.353370 6.988221 7.448425
 [49] 5.791253 6.488778 6.486423 6.754163 8.244121 6.133516 7.173533 6.133516
 [57] 6.133516 7.091092 5.489130 5.242881 6.903552 6.013092 4.888266 5.513594
 [65] 6.628367 6.375758 7.259940 6.240976 6.254443 6.242480 6.373623 6.865125
 [73] 6.736810 5.301738 5.489130 6.710729 5.496080 4.615328 6.353370 5.866668
 [81] 5.862692 5.496080 6.870101 7.448425 6.748695 6.605617 6.748695 5.369347
 [89] 5.536743 7.110176 6.415772 6.113038 5.614687 6.611481 5.810425 7.990742
 [97] 5.613699 6.748695 7.091092 5.862692 6.180243 5.127056 6.792563 5.513594
[105] 5.825298 5.366170 6.710729 5.496080 7.874450 6.862461 7.207523 5.274326
[113] 5.274326 6.862461 6.013092 6.133516 6.647172 6.237183 6.133516 6.628367
[121] 6.272948 7.016785 6.700253 5.920807 6.865125 5.865795 7.351985 6.612447
[129] 6.540785 6.272948 5.274326 5.467592 7.114277 6.500900 5.862692 5.532514
[137] 6.710729 5.692051 8.326894 4.921326 5.987680 6.373623 6.951901 7.656823
[145] 7.482520 5.987680 5.114453 6.998067 5.865795 5.381710 5.038418 8.498164
[153] 6.461913 5.920807 6.700253 5.274326 5.987680 6.113038 5.489130 5.363890
[161] 6.988359 7.656823 5.738024 5.920807 6.238541 6.389355 5.865795 7.240550
[169] 5.773781 6.700253 5.614687 6.540785 6.710729 6.344260 5.171841 5.614687
[177] 6.611481 8.385153 7.332550 7.656823 5.620104 6.415772 7.521983 6.646171
[185] 5.865795 6.113038 6.214418 7.653062 6.272948 5.120316 7.124206 5.539526
[193] 7.124206 6.742228 6.411527 5.893170 5.171841 6.616793 8.487265 7.656823
 [ reached getOption("max.print") -- omitted 9927 entries ]
```

Figura 3.6: Distância estatística do par de covariáveis X_1 e X_2 em torno da origem.

Como citado anteriormente, utilizamos a formula 3.2 para calcular a distancia, porem no R temos o comando mahalanobis() que faz esse calculo sem precisarmos criar um código especifico para cada problema, ou seja, se utilizarmos o comando: mahalanobis(X, 0, cov(X)) iremos obter o mesmo resultado.

3.2.3 Distancia em torno da média

Uma outra distância que podemos ter interesse em calcular, é a distância generalizada estatística para o par de covariáveis X_1 e X_2 em torno da média $(\bar{x}_1, \bar{x}_2)$. Sendo assim, utilizando o segundo método apresentado, a fórmula da distância passa a ser:

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

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).

assim, obtendo a distância que está representada pela Figura 3.7.

```
> options(max.print=200)
> d3
  [1] 0.85437423 1.26135923 0.93731275 1.00358842 2.17684596 0.48373537
  [7] 1.42813032 1.84901598 1.90459558 0.32842559 0.62304618 2.42992527
 [13] 1.95223488 1.53716745 1.52834771 0.65822311 0.32842559 0.80803963
 [19] 2.52540700 0.46135578 1.13234120 1.96115742 0.71203935 0.12563455
 [25] 0.95727847 1.01139627 1.58109342 2.51818645 0.29085542 1.13234120
 [31] 1.86195888 1.62104408 1.09265573 1.05415901 1.65859573 1.74924839
 [37] 1.74924839 0.88967523 1.63803619 2.62768035 1.25666681 1.06913800
 [43] 0.96792683 0.39408739 1.07621464 1.66985623 1.24326677 2.01239484
 [49] 0.88276972 0.71216378 0.71867774 1.03718240 2.51103192 0.53136773
 [55] 1.95223488 0.53136773 0.53136773 1.78701106 0.34319039 0.61731301
 [61] 1.38643326 0.50267957 1.08046774 0.51705302 1.01228499 0.66817321
 [67] 2.05507543 2.88772834 1.27315113 0.48751514 0.76072720 1.09259267
 [73] 0.98043648 1.05891529 0.34319039 1.34327756 0.48373537 1.16616691
 [79] 1.66985623 0.32522543 0.13877703 0.48373537 1.17546133 2.01239484
 [85] 1.50848352 1.52190924 1.50848352 0.53858639 0.69517878 1.33870241
 [91] 0.93731275 0.33360990 0.28080653 0.84936115 1.41287909 2.24897420
 [97] 0.16701185 1.50848352 1.78701106 0.13877703 1.06621360 0.80803963
[103] 1.68216764 0.51705302 1.87686952 0.41979214 1.34327756 0.48373537
[109] 2.51361637 1.54626252 1.71659295 0.88967523 0.88967523 1.54626252
[115] 0.50267957 0.53136773 1.70493127 0.45982420 0.53136773 1.01228499
[121] 0.73433571 1.74924839 1.44998687 0.92926535 1.09259267 0.12563455
[127] 2.02369206 0.83325707 1.17650092 0.73433571 0.88967523 1.23053848
[133] 1.37489182 0.88557951 0.13877703 0.84590360 1.34327756 1.05501879
[139] 3.37314840 1.11915099 0.25947849 0.76072720 1.88327984 2.56119664
[145] 2.14445261 0.25947849 0.66641543 1.83185976 0.12563455 1.71182219
[151] 0.99278327 2.77351333 1.65721148 0.92926535 1.44998687 0.88967523
[157] 0.25947849 0.33360990 0.34319039 0.43775750 1.21156766 2.56119664
[163] 0.07819008 0.92926535 0.51792187 0.90518629 0.12563455 1.50660792
[169] 0.54601360 1.44998687 0.28080653 1.17650092 1.34327756 1.34083158
[175] 1.09265573 0.28080653 0.84936115 2.69394425 2.15522183 2.56119664
[181] 0.24488998 0.93731275 2.28706329 1.14058336 0.12563455 0.33360990
[187] 1.07782977 2.40555756 0.73433571 0.68136371 1.43929440 1.35187858
[193] 1.43929440 0.97700868 1.06913800 0.50788431 1.09265573 0.91189302
[199] 2.71067668 2.56119664
 [ reached getOption("max.print") -- omitted 9927 entries ]
```

Figura 3.7: Distância estatística generalizada do par de covariáveis X_1 e X_2 em torno da média.

3.2.4 Analise dos resultados

Para que seja possível observarmos os resultados obtidos "lado a lado"e podermos realizar comparações, criamos uma tabela com as informações obtidas, o qual esta representada abaixo:

$X1$	$X2$	$Método1$	$Método2$	$Mahalanobis$	$Distancia\_Média$
45	39	5.8076	5.6618	5.66618	0.8543
49	44	6.4907	6.2136	6.2136	1.2613
51	36	5.9102	6.4157	6.4157	0.9373
40	34	5.0938	5.0167	5.0167	1.0035
...	...	...	...	...	...
41	25	4.598399	5.321797	5.321797	1.5260818
44	36	5.468324	5.496080	5.496080	0.4837354
30	36	5.242568	4.517804	4.517804	3.3255088
43	25	4.804119	5.655160	5.655160	1.7442264

Tabela 3.4: Distâncias

Ao compararmos os resultados das distâncias, vemos que a distância estatística generalizada das covariáveis X_1 e X_2 calculadas pelo método 1 e método 2 (igualmente ao mahalanobis) possuem resultados muito próximos. Contudo, quando comparamos os essas distâncias com a distância que possui a origem em torno da média, vemos que os valores são menores e diferentes dos outros.

Nesse sentido, após realizar os cálculos da distância estatística generalizada de duas formas diferentes, plotamos o gráfico de dispersão das covariáveis X_1 e X_2 com a elipse centrada na média e também plotamos o gráfico de dispersão com a rotação da elipse, que são representados, respectivamente, pelas Figuras 3.8 e 3.9.

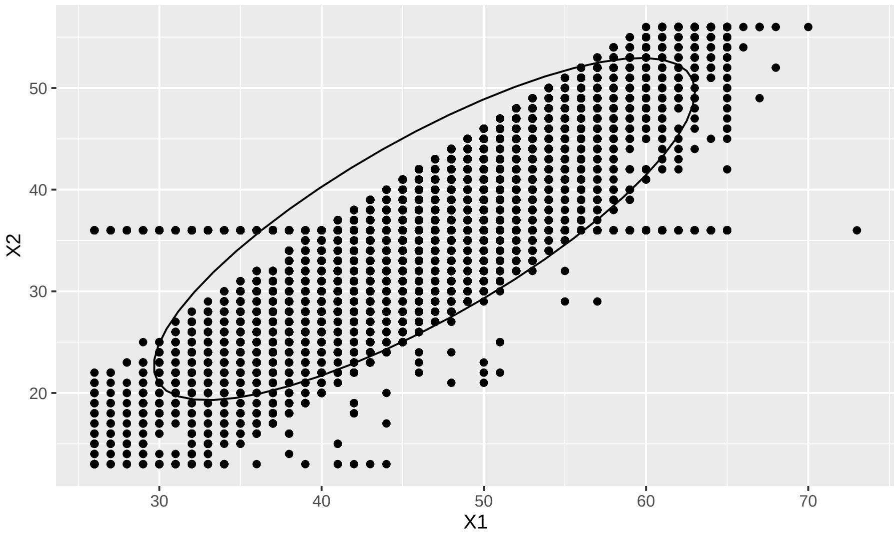

Figura 3.8: Gráfico de dispersão e elipse do par de covariáveis X_1 e X_2.

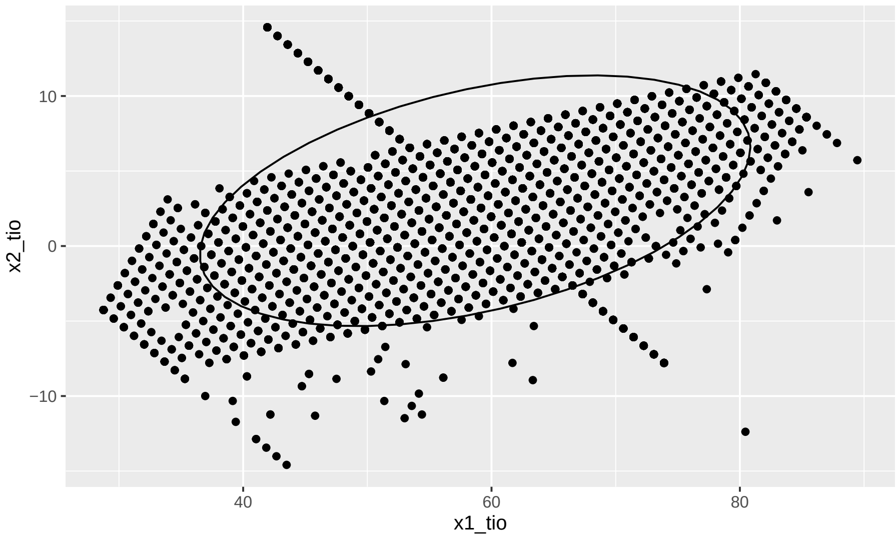

Figura 3.9: Gráfico da rotação do par de covariáveis X_1 e X_2.

3.2.5 Possíveis aplicações

Com as distancias obtidas podemos montar critérios para definir quais clientes se destacam dos demais em relação a idade e período de relacionamento com o banco. Ou seja, quais clientes tem idade avançada (em relação aos demais clientes) e possui baixo período de relacionamento com o banco; ou também, quais clientes possuem idade baixa (em relação aos demais clientes), e um alto período de relacionamento com o banco.

Utilizando os resultados da distancia estatística generalizada (em torno da média) representados na Figura 3.7, criamos o seguinte "rank"representado na Tabela 3.5, em que quanto mais elevado é o valor da distancia estatística generalizada, maior é a indiferença entre Idade x Período de relacionamento com o banco em relação a média dessas covariaveis (X1 e X2).

$X1$	$X2$	$Distancia\_Média$	ID
73	36	5.4030	252
44	13	4.3092	3758
43	13	4.1600	7571
26	36	4.1374	614
26	36	4.13746	991
26	36	4.13746	1090
26	36	4.13747	1125
...	...	...	...

Tabela 3.5

Tendo então uma classificação dos clientes que se diferem da média, podendo o Banco realizar ações (estudos, campanhas, promoções, entre outros) especificas para esse grupo de clientes.

Pode ser também realizadas outras aplicações com a distancia estatística generalizada, nesse caso utilizamos apenas as covariáveis $\mathbf{X}_1$ e $\mathbf{X}_2$, referidas à **Maturidade do Cliente**, mas também é possível aplicar a mesma análise nos outros Componentes.

3.3 Tratamento de Normalidade

Muitas técnicas de análise de dados e inferência, têm como pré requisito a normalidade multivariada dos dados.

Para contemplarmos a normalidade multivariada, é necessários que exista o comportamento de normalidade em todas as dimensões inferiores à total. Como possuímos 10 variáveis em estudo, para validar a normalidade multivariada, em tese, deveria ser verificada a normalidade univariada, bivariada, trivariada, etc, até chegar na décima dimensão. Entretanto, ao normalizar as primeiras dimensões já é possível que exista indícios de normalidade multivariada.

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

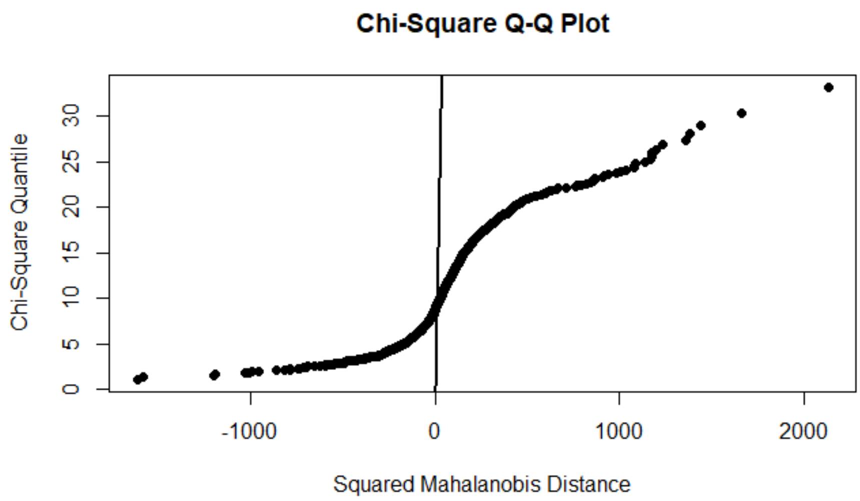

Figura 3.10: Gama Plot - Quantis QuiQuadrado para normalidade Multivariada

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 da QuiQuadrado.

Isso acontece pois, já na primeira dimensão (normalidade univariada), há variáveis que não se distribuem de forma normal.

Algumas variáveis, como é o caso de X_1, não apresentam esse problema:

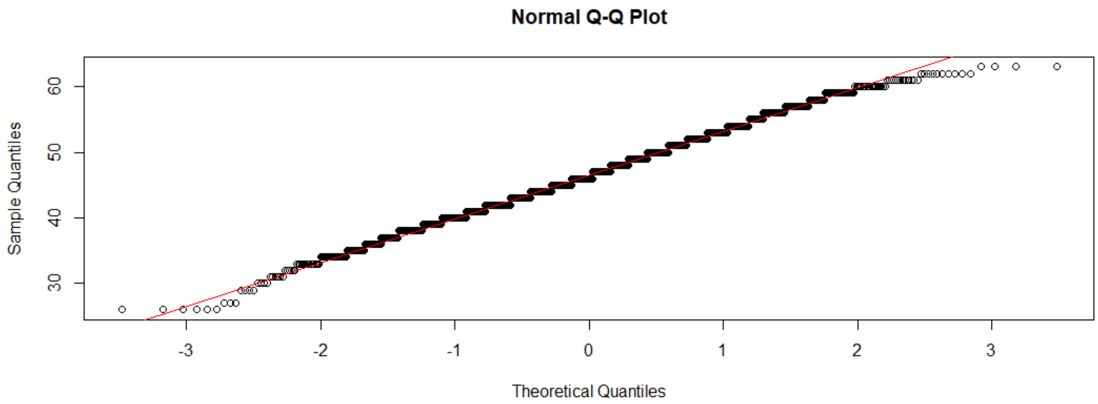

Figura 3.11: Gráfico Quantil-Quantil Normalidade Univariada X1

No caso desta variável, por exemplo, a normalidade é aceita, não rejeitando a hipótese nula dos Testes de Shapiro, KS e Anderson Darling. Percebe-se no qqplot, que há um distribuição linear quanto aos quantis teóricos da normal.

Entretanto, algumas variáveis, como é o caso de X_4 e X_{10} apresentam um distribuição, como vista na análise descritiva, relativamente complexa:

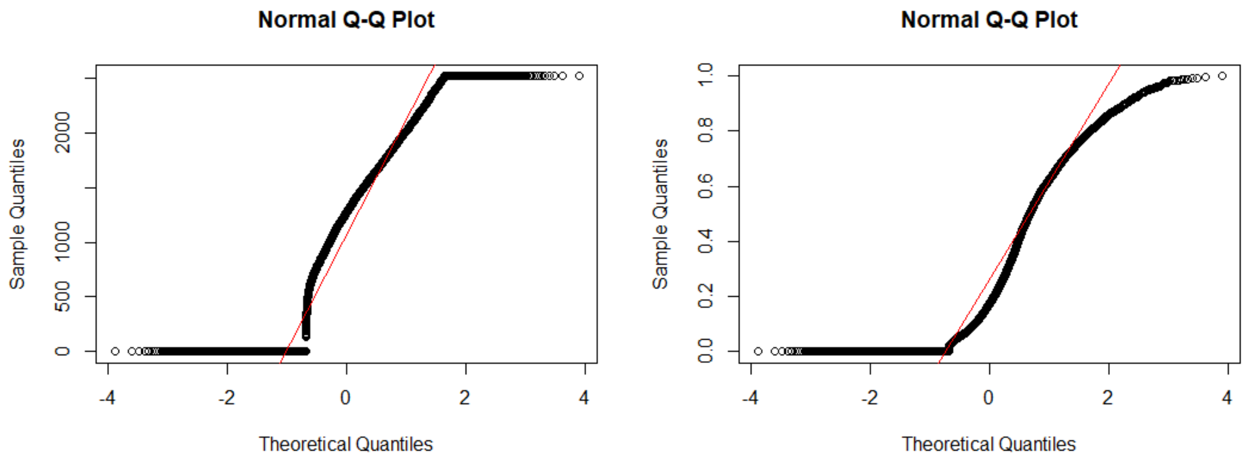

Figura 3.12: Gráfico Quantil-Quantil Normalidade Univariada X4 e X10 respectivamente

De fato, os testes de normalidade são rejeitados para ambas as variáveis. Através do gráfico de comparação com os quantis da normal já é possível perceber isso com nitidez, pois os pontos não se comportam de forma linear.

Sendo assim, torna-se necessária, para buscar a normalidade multivariada, uma transformação de acordo, que faça estas se comportarem como normais univariadas.

Entretanto, após tentativas de transformação (BoxCox, Exp, CenterScale, Logarítmica, Raiz, Arcsinh), **o Grupo não conseguiu** transformar tais variáveis a ponto de aceitar o teste de normalidade, nem mesmo de colaborar para a Normalidade Multivariada.

Estimamos que essa ocorrência possa ser dada por motivo de que estas variáveis são dicotômicas à atividade, ou seja, há uma grande concentração em um valor nulo (o zero), por não participação, e um outro grupo que se distribui entre os valores normalmente.

Sendo assim, para análises futuras, indica-se métodos que não atribuam normalidade como pré requisito. Ou, caso necessário, se possuir um objetivo específico a alguma variável, retirar essas observações de não atividade, e trabalhar com as demais. Claro, se isso não for enviesar a análise.

Capítulo 4

Conclusão

Dado as informações coletadas pela Instituição Financeira, muitas observações puderam ser feitas de forma univariada, analisando o comportamento dos clientes quanto as variáveis em estudo (tanto as de perfil, quanto as de atividade).

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.

Através desses indicadores, tornou-se possível olhar para as informações de forma mais resumida e direta. Considerando então um destes (o de Maturidade do Cliente), através da análise de distâncias, foram elencados clientes que mais se destacam com relação a Idade e Período de Relacionamento, assim possibilitando tomadas de decisões como a criação de ofertas direcionadas, personalização na comunicação e tratamento, dentre outros.

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. Indicando então análises que não possuam esse pré requisito, ou então uma personalização na base de dados visando objetivos específicos.

Apêndice A

Código

```
### ( SEMINARIO 1 ) ###

#Estat stica Multivariada 1 - Grupo 2
#An lise de Cart o de Cr dito
#

#leitura e organiza  o dos dados####
original = read.csv("BankChurners.csv", sep=",", header = TRUE,
                    stringsAsFactors = FALSE, encoding = "UTF-8")

dados = original[,-c(22,23)]
nomes=c("Identificador","Atividade","Idade","Sexo","Dependentes","N vel Educacional","Estado Civil",
        "Renda Anual", "Tipo do Cart o", "Per odo de Relacionamento", "N  de Produtos Mantidos",
        "Meses inativos U.A","N  de Contatos U.A","Limite de Cr dito", "Saldo Rotativo","M dia de Cr dito Aberto U.A",
        "Mudan a no Valor Transacional","Valor Total da Transa  o U.A","N  de Transa  es U.A","Mudan a no N  Transacional", "Taxa de Utiliza  o M dia")

library("data.table")
setnames(dados, nomes)
dados

#-- Salvando essa base
write.csv(dados, "dadoscartao.csv", row.names = F)

#verificando dados faltantes
any(is.na(dados))

#separando num rica de categ rica
dadosnum=dados[,c(3,10,14,15,16,17,18,19,20,21)]
dadoscat=dados[,-c(3,10,14,15,16,17,18,19,20,21)]
```

```
#Organizando base Num rica
nomes2=c("X1","X2","X3","X4","X5","X6","X7","X8","X9","X10")



library("data.table")
setnames(dadosnum, nomes2)

#Cabe alho
str(dadosnum)
head(dadosnum)

#AlgumasEstat sticas
options(scipen = 999)
summary(dadosnum)
var(dadosnum)


#analise descritiva
library(gridExtra)

tema = theme(axis.title.x = element_text(size = 14),axis.text.x =
    element_text(size = 12),axis.title.y = element_text(size = 14))

## X1
box_X1 = ggplot(dadosnum)+
  geom_boxplot(aes(x = X1), fill = "purple", col = "black")+
  theme_bw()+
  coord_flip()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = "Idade", y = "") + tema

his_X1 = ggplot(dadosnum)+
  geom_histogram(aes(x = X1), fill = "purple", col = "black", bins = 10)
    +
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = "Idade", y = "Frequ ncia Absoluta") + tema

grid.arrange(box_X1 , his_X1, ncol=2, top= "Idade do cliente (em anos)")


## X2
box_X2 = ggplot(dadosnum)+
  geom_boxplot(aes(x = X2), fill = "purple", col = "black")+
  theme_bw()+
  coord_flip()+
```

```
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = "Per odo de Relacionamento", y = "") + tema

his_X2 = ggplot(dadosnum)+
  geom_histogram(aes(x = X2), fill = "purple", col = "black", bins = 10)+
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = "Per odo de Relacionamento", y = "Frequ ncia Absoluta") + tema

grid.arrange(box_X2 , his_X2, ncol=2, top= "Per odo de relacionamento com o Banco (em meses)")



## X3
box_X3 = ggplot(dadosnum)+
  geom_boxplot(aes(x = X3), fill = "purple", col = "black")+
  theme_bw()+
  coord_flip()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = "Limite de cr dito", y = "") + tema

his_X3 = ggplot(dadosnum)+
  geom_histogram(aes(x = X3), fill = "purple", col = "black", bins = 10)+
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = "Limite de cr dito", y = "Frequ ncia Absoluta") + tema

grid.arrange(box_X3 , his_X3, ncol=2, top= "Limite de cr dito no cart o de cr dito (U$)")


## X4
box_X4 = ggplot(dadosnum)+
  geom_boxplot(aes(x = X4), fill = "purple", col = "black")+
  theme_bw()+
  coord_flip()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = "Saldo rotativo total", y = "") + tema

his_X4 = ggplot(dadosnum)+
```

```
  geom_histogram(aes(x = X4), fill = "purple", col = "black", bins = 10) +
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = "Saldo rotativo total", y = "Frequ ncia Absoluta") + tema

grid.arrange(box_X4 , his_X4, ncol=2, top= "Saldo rotativo total no cart o de cr dito (U$)")



## X5
box_X5 = ggplot(dadosnum)+
  geom_boxplot(aes(x = X5), fill = "purple", col = "black")+
  theme_bw()+
  coord_flip()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = "M dia de linhas abertas", y = "") + tema

his_X5 = ggplot(dadosnum)+
  geom_histogram(aes(x = X5), fill = "purple", col = "black", bins = 10) +
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = "M dia de linhas abertas", y = "Frequ ncia Absoluta") + tema

grid.arrange(box_X5 , his_X5, ncol=2, top= "Linha de cr dito aberta para compra (m dia dos  ltimos  12 meses)")


## X6
box_X6 = ggplot(dadosnum)+
  geom_boxplot(aes(x = X6), fill = "purple", col = "black")+
  theme_bw()+
  coord_flip()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = "Mudan as no valor", y = "") + tema

his_X6 = ggplot(dadosnum)+
  geom_histogram(aes(x = X6), fill = "purple", col = "black", bins = 10) +
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
```

```
  labs(x = "Mudan as no valor", y = "Frequ ncia Absoluta") + tema

grid.arrange(box_X6 , his_X6, ncol=2, top= "Mudan a no valor da transa  o (Q4 sobre Q1)")


## X7
box_X7 = ggplot(dadosnum)+
  geom_boxplot(aes(x = X7), fill = "purple", col = "black")+
  theme_bw()+
  coord_flip()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = "Valor total", y = "") + tema

his_X7 = ggplot(dadosnum)+
  geom_histogram(aes(x = X7), fill = "purple", col = "black", bins = 10) +
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = "Valor total", y = "Frequ ncia Absoluta") + tema

grid.arrange(box_X7, his_X7, ncol=2, top= "Valor total da transa  o ( ltimos  12 meses)")


## X8
box_X8 = ggplot(dadosnum)+
  geom_boxplot(aes(x = X8), fill = "purple", col = "black")+
  theme_bw()+
  coord_flip()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = "Contagem total", y = "") + tema

his_X8 = ggplot(dadosnum)+
  geom_histogram(aes(x = X8), fill = "purple", col = "black", bins = 10) +
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = "Contagem total", y = "Frequ ncia Absoluta") + tema

grid.arrange(box_X8 , his_X8, ncol=2, top= "Contagem total de transa  es (nos  ltimos  12 meses)")


## X9
```

```
box_X9 = ggplot(dadosnum)+
  geom_boxplot(aes(x = X9), fill = "purple", col = "black")+
  theme_bw()+
  coord_flip()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = "Mudan as na contagem", y = "") + tema

his_X9 = ggplot(dadosnum)+
  geom_histogram(aes(x = X9), fill = "purple", col = "black", bins = 10)
   +
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = "Mudan as na contagem", y = "Frequ ncia Absoluta") + tema

grid.arrange(box_X9 , his_X9, ncol=2, top= "Mudan a na contagem de
    transa  es (Q4 sobre Q1)")


## X10
box_X10 = ggplot(dadosnum)+
  geom_boxplot(aes(x = X10), fill = "purple", col = "black")+
  theme_bw()+
  coord_flip()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = " Taxa de utiliza  o m dia", y = "") + tema

his_X10 = ggplot(dadosnum)+
  geom_histogram(aes(x = X10), fill = "purple", col = "black", bins =
   10)+
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 18, family ="serif"))+
  labs(x = " Taxa de utiliza  o m dia", y = "Frequ ncia Absoluta") +
    tema

grid.arrange(box_X10 , his_X10, ncol=2, top= " Taxa de utiliza  o
    m dia do cart o")



#Normalidade

library(MVN)
library(AID)
library(dgof)
require("MVA")
```

```
require(GGally)
require(CCA)


mvn(dadosnum, mvnTest= c("royston"), covariance=TRUE, scale=FALSE, desc=
   TRUE,
    transform="none", R=1000, univariateTest = c("AD"), univariatePlot =
   "qq", multivariatePlot = "qq",
    multivariateOutlierMethod = "quan", bcType="optimal", showOutliers =
    TRUE, showNewData = FALSE)


library("nortest")

shapiro.test(dadosnum2$X1)
ad.test(dadosnum2$X1)
qqnorm(dadosnum2$X1)
qqline(dadosnum2$X1, col="red")
qqnorm(dadosnum$X2)
qqline(dadosnum$X2, col="red")
qqnorm(dadosnum$X3)
qqline(dadosnum$X3, col="red")
qqnorm(dadosnum$X4)
qqline(dadosnum$X4, col="red")
qqnorm(dadosnum$X5)
qqline(dadosnum$X5, col="red")
qqnorm(dadosnum$X6)
qqline(dadosnum$X6, col="red")
qqnorm(dadosnum$X7)
qqline(dadosnum$X7, col="red")
qqnorm(dadosnum$X8)
qqline(dadosnum$X8, col="red")
qqnorm(dadosnum$X9)
qqline(dadosnum$X9, col="red")
qqnorm(dadosnum$X10)
qqline(dadosnum$X10, col="red")
ks.test(dadosnum$X1,mean(dadosnum$X1),sd(dadosnum$X1))
ks.test(dadosnum$X2,mean(dadosnum$X2),sd(dadosnum$X2))
ks.test(dadosnum$X3,mean(dadosnum$X3),sd(dadosnum$X3))
ks.test(dadosnum$X4,mean(dadosnum$X4),sd(dadosnum$X4))
ks.test(dadosnum$X5,mean(dadosnum$X5),sd(dadosnum$X5))
ks.test(dadosnum$X10,mean(dadosnum$X10),sd(dadosnum$X10))

bestNormalize(dadosnum$X4)
bestNormalize(dadosnum$X10)

## Escolhendo 2 covariaveis
df_ATV2 = data.frame(dadosnum$X1, dadosnum$X2)
nomes3=c("X1","X2")
```

```r
setnames(df_ATV2, nomes3)
df_ATV2

#distancia estatistica
d3 = sqrt(mahalanobis(df_ATV2, 0, cov(df_ATV2)))
d3

generalizada$mahalanobis = d3

# Distancia do vetor de m dias

# Obtendo o vetor X1 - X1_mean
novoX1 = 0
for(i in 1:nrow(df_ATV2)){
  novoX1[i] = df_ATV2$X1[i] - mean(df_ATV2$X1)
}

novoX1

# Obtendo o vetor X2 - X2_mean
novoX2 = 0
for(i in 1:nrow(df_ATV2)){
  novoX2[i] = df_ATV2$X2[i] - mean(df_ATV2$X2)
}

novoX2

novoX = data.frame(novoX1, novoX2)

# calculando a distancia
C = data.matrix(novoX)
d3 = 0
for(i in 1:nrow(df_ATV2)){

  d3[i] = sqrt(t(C[i,])%*%solve(s)%*%C[i,])

}

options(max.print=100)
d3

d4 = sqrt(mahalanobis(df_ATV2, colMeans(df_ATV2), cov(df_ATV2)))
d4

generalizada$distancia_media = d3
generalizada$mahalanobis_media = d4
View(generalizada)

x = generalizada
```

```r
x$distancia = NULL
x$Distancia_2 = NULL
x$mahalanobis = NULL
x$mahalanobis_media = NULL
x$ID = c(1:10127)
library(dplyr)
x%>%
  arrange(desc(x$distancia_media))

View(x)


###Componentes principais
library(FactoMineR)
acp = PCA(dadosnum, scale.unit=TRUE, graph=TRUE)

#autovalores
acp$eig

#autovetores
acp$var$coord

#escores fatoriais
acp$ind$coord
summary(acp$ind$coord)


# Scree Plot
library(factoextra)
fviz_eig(acp, addlabels=TRUE, xlab = "Autovalores",ylab = "Porcentagem de Explica  o da Vari ncia")


library ("corrplot")
#Qualidade de representa  o da vari veis
corrplot(acp$var$coord, is.corr=FALSE,title = "
               Vari vel por componente")

# Contribui  o das Vari veis para os CP
col3 = hcl.colors(10, "PuBu", rev = TRUE)
corrplot(acp$var$contrib, is.corr= FALSE,title = "
               Componente por vari vel",col = col3)


#graficos acp
dadosacp= data.frame(acp$ind$coord)

d1d2 =ggplot(dadosacp, aes(x=Dim.2, y=Dim.1)) +
  xlab("Atividade em Transa  es")+ ylab("Score do Cart o de Cr dito"
```

```
    )+
  geom_point()

d3d4= ggplot(dadosacp, aes(x=Dim.4, y=Dim.3)) +
  xlab("Mudan as nas opera  es")+ ylab("Maturidade do cliente")+
  geom_point()

grid.arrange(d1d2 , d3d4, ncol=2, top= "Dispers o dos Componentes")
```