Gabarito Lista 04

Ciência de Dados - 2025

Alex Monito Nhancololo

Critérios usados na correção

Independentemente do software empregado, aplicam-se os seguintes critérios de avaliação:

- Justificar sempre as escolhas metodológicas, não bastando apresentar apenas o código e/ou sua saída.
- Manter o documento limpo, sem `#` desnecessários, warnings, erros ou descrições triviais.
- Incluir tabelas e gráficos bem formatados, com títulos, legendas e rótulos claros para interpretação autônoma.
- Organizar todo o código em apêndice/anexo, mantendo coerência, consistência e redação técnica objetiva.

Nota: Comentários sobre funções básicas (como set.seed() ou outras funções triviais) não devem ser incluídos.

Sugestões

- Use `pacman` para instalar e importar pacotes numa só vez, evitando `install.packages()`, `library()` ou `require` para cada pacote.
- Use `knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE, comment = "")`, para mostrar o código e a saída, ocultar mensagens automáticas, ocultar mensagens de aviso (warnings) na saída e remover o símbolo padrão (como ##) antes da saída do código, respectivamente.
- Use pacotes como `textreg` ou `stargazer`, `finalfit`, `flextable`, `equatiomatic`, entre outros, para extração das estimativas do modelo e suas métricas em diferentes formatos (LaTeX, HTML, Word, colocar tabela única para vários modelos etc.), geração de tabelas formatadas e extração automática das equações/expressões matemáticas do modelo ajustado com ou sem as estimativas, respectivamente.

Pacotes usados:

```
if (!require("pacman")) install.packages("pacman")
pacman::p_load(
  knitr, ggstatsplot,MASS,
  datasets, dplyr, psych, ggplot2, devtools, GPArotation,
  ISLR, ggbiplot, ggrepel, factoextra, corrplot, fastICA,
  ica, MASS, gtsummary, GGally, flextable, tidyr, kableExtra,
  patchwork, finalfit
)
```

EXERCÍCIO 01: Componentes principais

A Análise de Componentes Principais (PCA) é uma técnica exploratória de redução de dimensionalidade que transforma variáveis correlacionadas em um conjunto menor de variáveis não correlacionadas entre si, chamadas componentes principais. Esses componentes são combinações lineares das variáveis originais, ordenados da maior para a menor capacidade de explicar a variância nos dados. É importante lembrar que, se as variáveis originais estiverem em escalas diferentes, elas devem ser padronizadas (normalizadas) antes de aplicar o PCA. Caso contrário, variáveis com maior variância (ou amplitude) dominarão os primeiros componentes.

A Tabela 1 apresenta parcialmente os dados *Iris*. Embora todas as variáveis do conjunto `iris` compartilhem a mesma unidade (cm), as diferenças expressivas nas variâncias (Tabela 2) exigem padronização (função `scale()`) para a maioria das análises multivariadas (PCA, FA, ICA). A não padronização distorce os resultados ao super-representar variáveis com maior dispersão, como demonstrado nas discrepâncias entre os resultados de PCA com e sem padronização (Figura @pca-final).

Tabela 1: Banco de dados com duas observações de cada Espécie

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica

Tabela 2: Resumo estatístico do conjunto de dados

Variável	**setosa** N = 50[1]	**versicolor** N = 50[1]	**virginica** N = 50[1]
Sepal.Length	5,01 (0,35)	5,94 (0,52)	6,59 (0,64)
Sepal.Width	3,43 (0,38)	2,77 (0,31)	2,97 (0,32)
Petal.Length	1,46 (0,17)	4,26 (0,47)	5,55 (0,55)
Petal.Width	0,25 (0,11)	1,33 (0,20)	2,03 (0,27)

[1]Média (Desvio Padrão)

Na Figura 1, observa-se que o comprimento e a largura da pétala exibem correlação positiva forte (0,96), seguido do comprimento da sépala com ambas as medidas da pétala (0,87 e 0,82). Em contraste, a largura da sépala apresenta correlação negativa com o comprimento da pétala (–0,43) e sua largura (–0,37), bem como comprimento da sépala (–0,12). Os histogramas mostram distribuição simétrica para comprimento da sépala (4,5–8,0 cm) e ligeiramente assimétrica à esquerda para largura da sépala (2,0–4,5 cm), enquanto comprimento e largura da pétala exibem multimodalidade (picos em 1,5 cm e 4,5 cm para comprimento; 0,3 cm e 1,5 cm para largura).

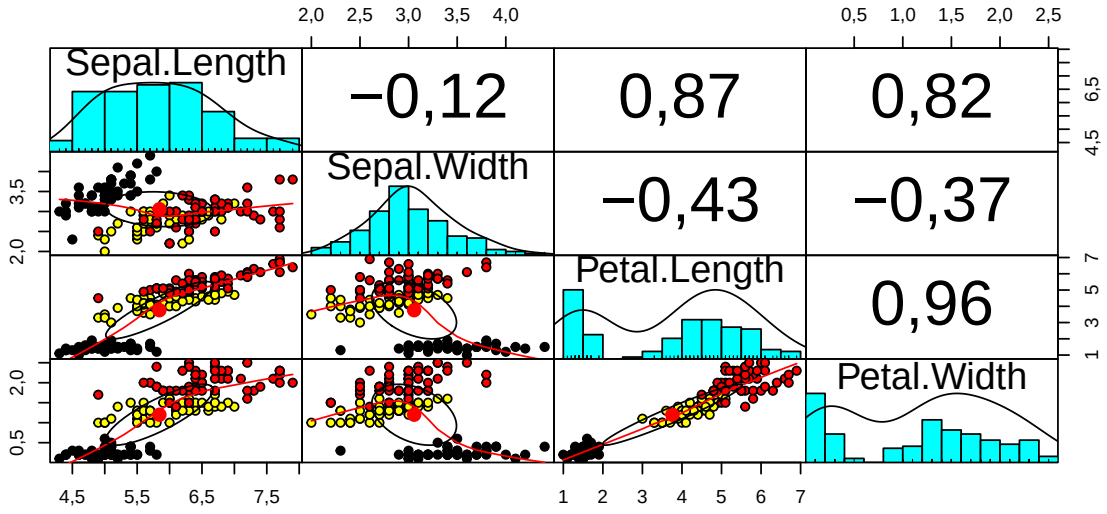

Figura 1: Matriz de dispersão das variáveis de sépala e pétala do iris — setosa (preto), versicolor (amarelo) e virginica (vermelho)

Os componentes principais CP1 e CP2 (Tabela 3) explicam juntos 96% da variância morfológica total entre as três espécies florais (73% e 23%, respectivamente). O CP1 é dominado por contribuições positivas das variáveis de comprimento e largura das pétalas (Petal.Length:

0,58; Petal.Width: 0,56) e de comprimento da sépala (Sepal.Length: 0,52), com contribuição negativa da largura da sépala (Sepal.Width: –0,27), representando um eixo de tamanho floral geral. O CP2 é determinado pela largura da sépala (Sepal.Width: –0,92) e, em menor grau, pelo comprimento da sépala (Sepal.Length: –0,38), refletindo variações na morfologia da sépala. Os demais componentes (CP3 e CP4) explicam apenas 4% e 1% da variância, respectivamente, com padrões residuais: o CP3 opõe comprimento da sépala (0,72) à largura da pétala (–0,63), enquanto o CP4 opõe comprimento da pétala (–0,80) à largura da pétala (0,52), mas sua relevância prática é mínima devido à baixa variância explicada.

Tabela 3: Componentes principais para o conjunto de dados

	Componentes principais (CP)			
	CP1	**CP2**	**CP3**	**CP4**
Variável				
Sepal.Length	0,52	-0,38	0,72	0,26
Sepal.Width	-0,27	-0,92	-0,24	-0,12
Petal.Length	0,58	-0,02	-0,14	-0,80
Petal.Width	0,56	-0,07	-0,63	0,52
Sd	1,71	0,96	0,38	0,14
% Variância	73,00	23,00	4,00	1,00
% Acumulada	73,00	96,00	99,00	100,00

Uma alternativa para a escolha dos componentes, além do uso da variância (%) acumulada, é o uso do screeplot. O screeplot mostra a variância explicada por cada componente principal (CP) em ordem decrescente, permitindo identificar quantos CPs são relevantes: o primeiro CP (à esquerda) tem a barra mais alta (maior variância), seguido de quedas abruptas; o ponto em que as barras se estabilizam (o "cotovelo") indica o limite de CPs ótimos. Nos dados *Iris* (Figura 2), o cotovelo está entre CP2 e CP3, validando que os dois primeiros componentes capturam a maior variabilidade dos dados (95,9% da variância), enquanto os demais (CP3 e CP4) podem ser descartados por adicionarem pouca informação.

A Figura @pca-final mostra que a padronização é essencial para equilibrar a contribuição das variáveis na variabilidade total e nos componentes. No PCA não padronizado (A), a dominância do comprimento da sépala (por possuir escala maior) distorce os resultados, capturando 92,46% da variância no PC1, mas falhando em discriminar versicolor e virginica. Já no PCA padronizado (B), todas as variáveis contribuem proporcionalmente: o PC1 (72,96%) reflete o tamanho floral global (pétalas + sépala) e o PC2 (22,85%) destaca a largura da sépala, permitindo separação nítida das três espécies, setosa (PC1 baixo/PC2 alto), virginica (PC1 alto/PC2 baixo) e versicolor (intermediário).

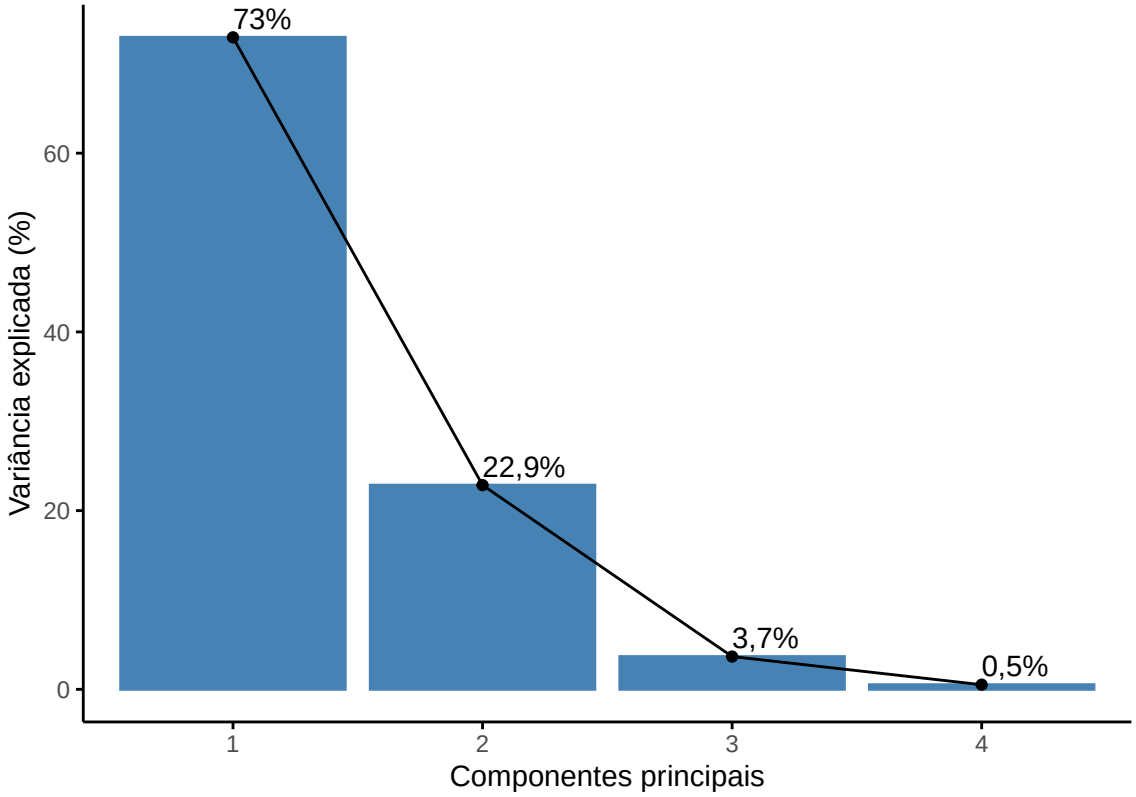

Figura 2: Screeplot.

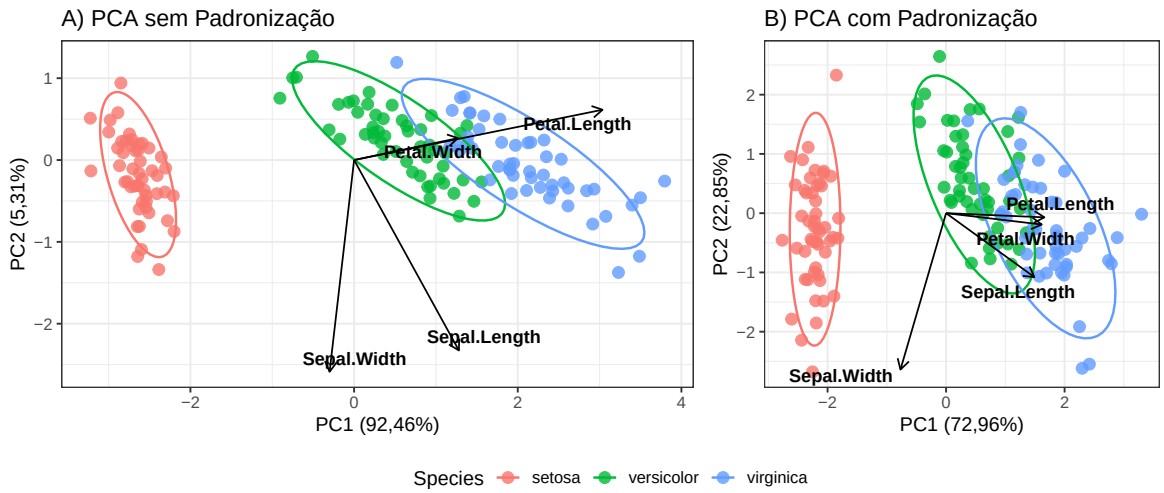

Figura 3: Análise PCA para dados Iris com biplots

Por fim, vamos criar o gráfico de dispersão baseado no PCA e verificar se o problema da multicolinearidade foi resolvido. Na Figura 4, os coeficientes de correlação são zero, indicando que não há problemas de multicolinearidade. Ademais, o PC1 concentra as maiores magnitudes (valores próximos de $\pm 0,5$), indicando forte contribuição integrada de todas as variáveis originais (com destaque para os comprimentos da pétala e da sépala), enquanto o PC2 mostra oposição entre direções (valores positivos e negativos), refletindo o contraste morfológico entre a largura da sépala (negativo) e as demais variáveis; já os PC3 e PC4 exibem magnitudes residuais ($< |0{,}2|$).

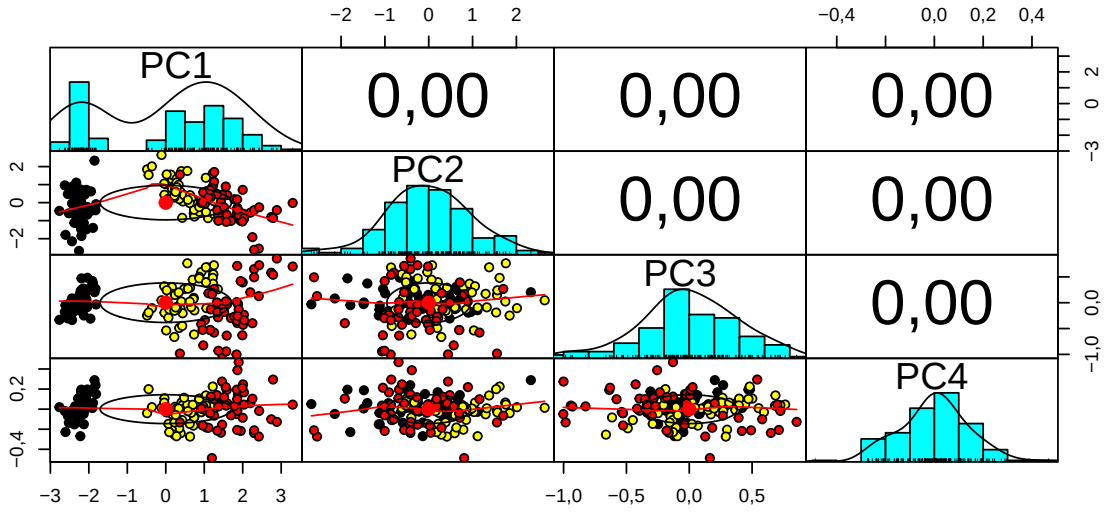

Figura 4: Matriz de dispersão baseado no PCA — setosa (preto), versicolor (amarelo) e virginica (vermelho)

EXERCÍCIO 02: Análise fatorial

Muitas pessoas costumam confundir análise fatorial com análise de componentes principais. A seguir, apresenta-se uma descrição de AF e sua diferença em relação ao PCA.

Leitura opcional: A Análise Fatorial (AF) é usada quando suspeitamos que várias variáveis que medimos (como perguntas de um questionário ou características observadas) são, na verdade, influenciadas por um número menor de "fatores escondidos" (também chamados de variáveis latentes ou construtos teóricos não medidos diretamente). Ela ajuda a descobrir essas causas subjacentes que explicam por que as variáveis originais estão correlacionadas entre si (por exemplo, notas altas em Português, Matemática e Ciências podem ser influenciadas por um fator latente chamado *Inteligência*).

Para aplicar a AF, verificamos primeiro se os dados são correlacionados (matriz de correlação), aplicamos o *Teste de Esfericidade de Bartlett* (que deve ser significativo, $p < 0{,}05$) e o *índice KMO* (ideal $> 0{,}7$). Decidimos quantos fatores extrair combinando três critérios: autovalores maiores que 1 (critério de Kaiser), o ponto de "cotovelo" no gráfico screeplot e a variância total explicada (a escolha é subjetiva). Extraímos os fatores (com métodos como máxima verossimilhança ou mínimos quadrados, etc.), aplicamos rotação (por exemplo, *Varimax*, que rotaciona os eixos para tornar os fatores não correlacionados, ou *Promax*, que rotaciona os eixos permitindo que os fatores sejam correlacionados) para facilitar a interpretação, e, em seguida, analisamos as *cargas fatoriais* (valores altos, próximos de $|1|$, indicam forte ligação da variável

ao fator) para *nomear cada fator* (por exemplo, um fator com cargas altas em "Comprimento da Pétala" e "Largura da Pétala" pode ser chamado de "Tamanho da Pétala"). Validamos o modelo verificando se os resíduos (diferenças entre correlações observadas e estimadas) são pequenos ($< |0{,}05|$) e, no caso de máxima verossimilhança, se o teste de Qui-quadrado apresenta $p > 0{,}05$, indicando bom ajuste.

Já a Análise de Componentes Principais (PCA) tem objetivo diferente: não busca causas escondidas, mas resume os dados originais criando novas variáveis (componentes principais), que são combinações lineares das variáveis originais e capturam o máximo possível da variância total. Enquanto a AF concentra-se em explicar a covariância (correlações) entre as variáveis por meio de fatores latentes, separando a variância única (ruído), a PCA foca em explicar a variância total (variância comum também designada comunalidade + variância única), sem separar o "ruído" ou assumir um modelo causal.

Assim, use PCA quando quiser *reduzir a dimensão dos dados mantendo o máximo de informação original* (para simplificar análises ou visualizações); use AF quando quiser investigar ou confirmar a existência de conceitos ou construtos teóricos subjacentes (variáveis latentes não observadas) que expliquem *por que* suas variáveis se relacionam.

A matriz de correlação anti-imagem (Q) (Figura 5) mostra as correlações parciais entre pares de variáveis, após remover a influência das demais variáveis do conjunto. Os valores apresentados são as correlações parciais ajustadas, indicando relações únicas não compartilhadas com outras variáveis; magnitudes elevadas ($> |0{,}5|$) sugerem que pares de variáveis têm associações não explicadas por fatores comuns.

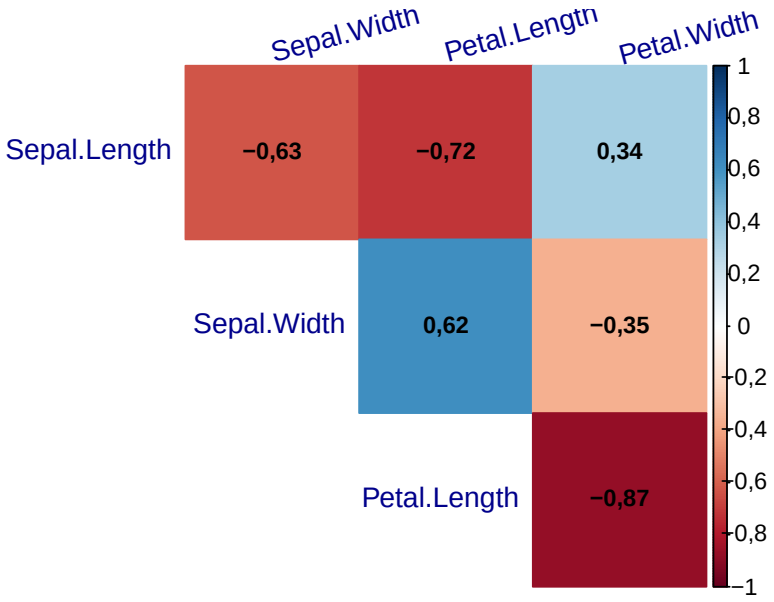

Figura 5: Matriz de correlação anti-imagem (Q).

Tabela 4: Medidas de Adequação Amostral (KMO)

[!h]

	Medidas de Adequação Amostral (KMO)
Variável	**MSAi**
Medidas Individuais (MSAi)	
Sepal.Length	0,584
Sepal.Width	0,270
Petal.Length	0,531
Petal.Width	0,634
MSA Global	
MSA	0,540

MSAi: Avalia se cada variável individual se relaciona adequadamente com as demais
MSA: Avalia se todo o conjunto de dados é adequado para Análise Fatorial.

Conforme descrito por Fávero e Belfiore (2024, p. 407), o índice KMO (0 a 1) classifica-se como: 0,9 = *Muito boa*, 0,8–0,9 = *Boa*, 0,7–0,8 = *Média*, 0,6–0,7 = *Razoável*, 0,5–0,6 = *Má*, $< 0{,}5$ = *Inaceitável*.

Nos resultados (Tabela 4), o MSA global (0,54) é *Má*, e os MSAs individuais revelam que `Sepal.Width` (0,27) é *Inaceitável*, `Petal.Length` (0,53) e `Sepal.Length` (0,58) são *Más*, enquanto `Petal.Width` (0,63) é *Razoável*, indicando falta de fatores comuns e inadequação dos dados para análise fatorial.

Os autores ressaltam que o Teste de Esfericidade de Bartlett (`cortest.bartlett`), que avalia se a matriz de correlações difere significativamente de uma matriz identidade, *deve ser sempre preferido ao KMO para decisão sobre adequação global da AF*, pois constitui um teste estatístico formal com nível de significância definido, enquanto o KMO é uma medida descritiva sem distribuição probabilística ou hipóteses testáveis que permitam inferência decisória (Fávero & Belfiore, 2024, p. 407).

A Tabela 5 apresenta os resultados do Teste de Esfericidade de Bartlett, que testa a hipótese nula de que a matriz de correlações populacional é uma matriz identidade (todas as correlações entre variáveis são nulas), contra a alternativa de que não o é. Observa-se que o p-valor é 0,0005 ($< 5\%$), o que permite rejeitar a hipótese nula ao nível de significância de 5%, concluindo-se que existem correlações estatisticamente diferentes de zero entre pelo menos alguns pares de variáveis, atendendo assim ao pré-requisito estatístico mínimo para realização de Análise Fatorial (AF). Portanto, a AF é apropriada, e podem ser extraídos fatores a partir das variáveis originais.

Nota: Dada a divergência entre o resultado do Teste de Esfericidade de Bartlett e o KMO, recomenda-se como critério adicional: (a) avaliar a exclusão da variável `Sepal.Width` devido ao seu baixo MSAi; ou (b) aplicar o *alpha de Cronbach* para verificar a fidedignidade da extração de fatores a partir das variáveis originais.

Tabela 5: Bartlett de homogeneidade de variâncias

Bartlett.s.K.squared	df	p.value
15.2	2	0.000499

Dada a natureza aberta da questão e as duas aplicações possíveis da Análise Fatorial confirmatória (AFC), que testa uma estrutura fatorial pré-definida teoricamente, e exploratória (AFE), que identifica padrões latentes através das covariâncias entre variáveis observadas, optou-se pela abordagem exploratória (AFE). Esta escolha visa identificar fatores distintos e interpretáveis sem pressupostos teóricos prévios, mantendo sua independência (ortogonalidade) para simplificar a análise. Para tanto, implementou-se o método de Máxima Verossimilhança (MLE) (vide Morettin & Singer, 2024), com rotação *Varimax* (vide o texto **Leitura opcional**). Devido à restrição matemática que limita a extração multifatorial com apenas quatro variáveis em alguns pacotes, utilizou-se o pacote `psych`. Sua função `fa()` supera essa limitação, oferecendo flexibilidade para: implementar métodos de extração (MLE, minres); processar diversos inputs (dados brutos, matrizes de correlação/covariância); calcular escores fatoriais (regressão, Bartlett, Thurstone); aplicar rotações (Varimax ortogonal ou Promax oblíqua); e gerenciar valores ausentes com controle de convergência (vide `help(psych)`).

Na Tabela 6, observa-se que apenas o primeiro fator apresenta autovalor maior que 1. Assim, pelo critério de Kaiser, o ideal seria considerar um fator, embora o segundo esteja próximo de 1.

Tabela 6: Retenção de fatores pelo critério de Kaiser

Fator	Autovalor	% Variância	% Acumulada
1	2.918	73.0	73.0
2	0.914	22.9	95.8
3	0.147	3.7	99.5
4	0.021	0.5	100.0

A Figura 6 mostra que, pelo critério de cotovelo, reteríamos dois fatores.

```
Parallel analysis suggests that the number of factors =  1  and the number of components =  1
```

A Tabela 7 mostra que, embora o modelo de um fator explique 72,7% da variância total, com comunalidades aceitáveis para as pétalas (0,95 em `Petal.Length`), apresenta desempenho insatisfatório para `Sepal.Width` (comunalidade = 0,27). Já o modelo de dois fatores com rotação

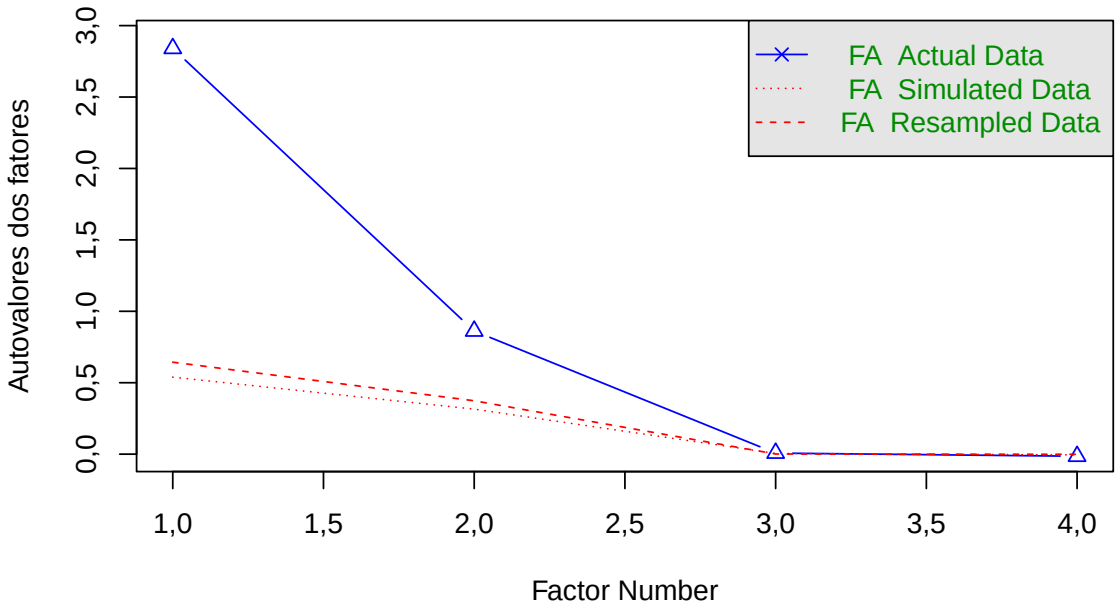

Figura 6: Screenplot da análise fatorial.

Varimax demonstra superioridade estatística ao explicar 95,7% da variância total, com ganhos expressivos em todas as comunalidades, especialmente em `Sepal.Width` (de 0,27 para 0,63). O Fator 1 representa dimensões de pétalas (cargas $> 0{,}95$ em `Petal.Length` e `Petal.Width`), e o Fator 2 caracteriza as sépalas (cargas $> 0{,}80$ em `Sepal.Length` e `Sepal.Width`), com ortogonalidade preservada (correlação entre fatores $= -0{,}003$). Portanto, recomenda-se trabalhar com dois fatores devido ao ganho de 23% na variância explicada.

Tabela 7: Cargas Fatoriais e Comunalidades

Variável	C_1Fator	Comun_1Fat	Carga_F1_2Fat	Carga_F2_2Fat	Comun_2Fat
Sepal.Length	0.872	0.760	0.997	0.006	0.995
Petal.Length	0.998	0.995	0.871	0.486	0.995
Petal.Width	0.965	0.931	0.818	0.514	0.932
Sepal.Width	-0.422	0.178	-0.115	-0.665	0.455

A Figura 7 mostra dois eixos ortogonais (Fator 1 vs. Fator 2, com rotação Varimax) e observa-se que `Petal.Length` e `Sepal.Length` estão fortemente associados ao Fator M1 (cargas positivas altas), formando um cluster que representa a dimensão de comprimento floral. Já `Petal.Width` está isoladamente vinculada ao Fator ML2, configurando um fator de *largura da pétala*. `Sepal.Width` aparece em posição oposta a M1 (carga negativa), indicando que sépalas mais largas correlacionam-se inversamente com estruturas alongadas.

A Figura 8 mostra que o Fator M1 está fortemente correlacionado com `Petal.Length` (carga 0,9) e moderadamente com `Sepal.Length`, indicando uma dimensão relacionada ao *comprimento estrutural da flor*. Já o Fator ML2 associa-se à variável `Petal.Width`, sugerindo um fator de *largura da pétala*. Nota-se ainda que `Sepal.Width` tem correlação negativa com o M1 (carga –0,7), o que implica que sépalas mais largas tendem a ocorrer em flores com pétalas e sépalas menos alongadas.

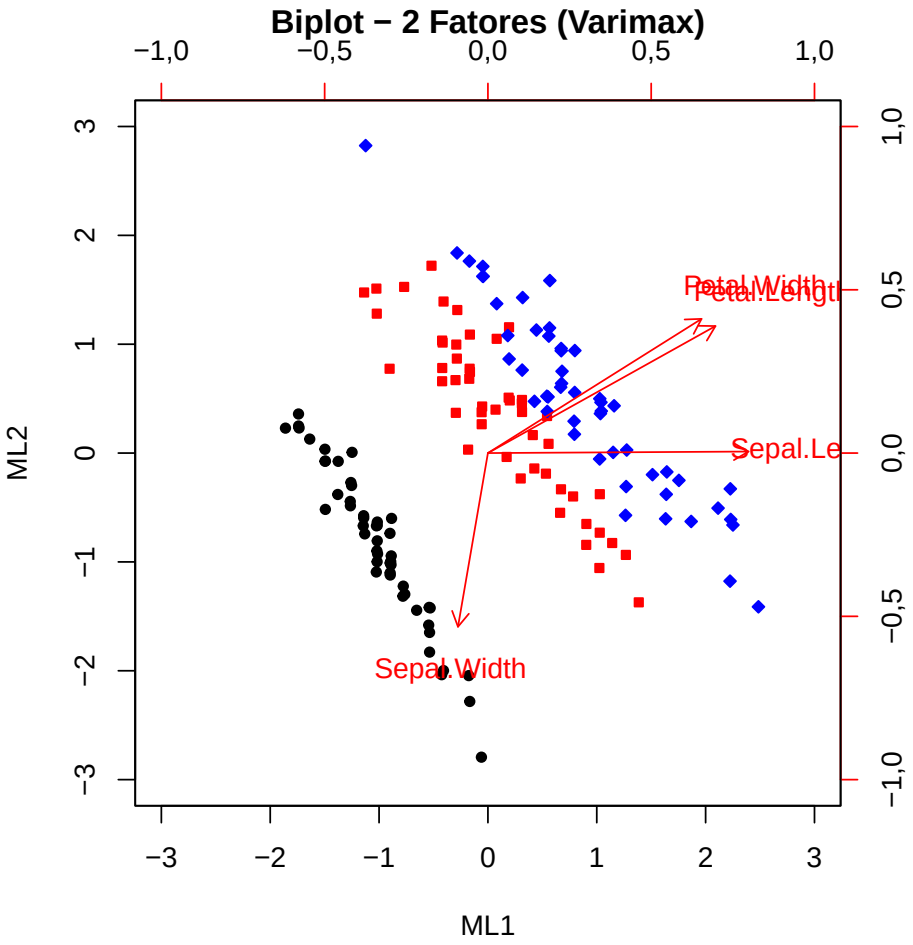

Figura 7: Biplot da análise fatorial-Setosa (preto), Versicolor (vermelho) e Virginica (azul)

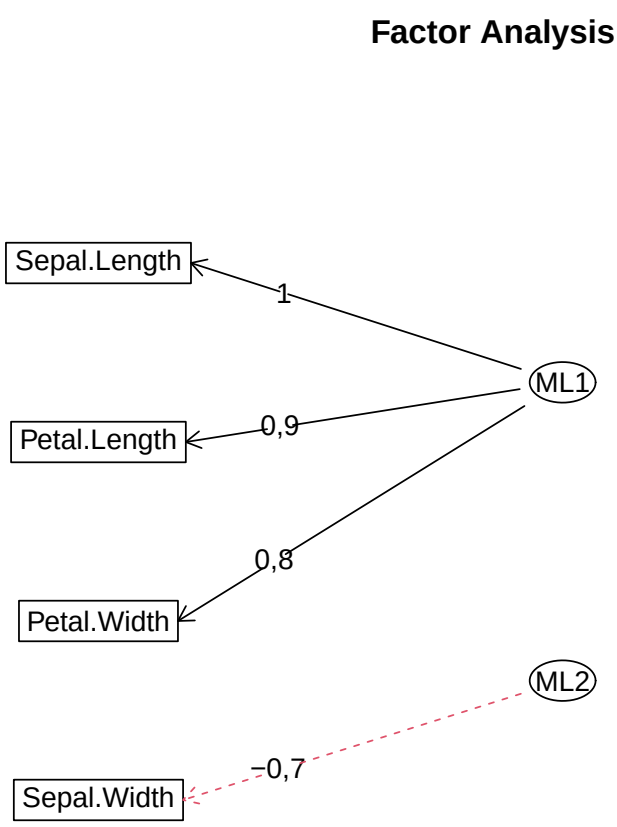

Figura 8: Matrizes de cargas fatoriais

EXERCÍCIO 03: Análise de componentes independentes (ICA)

A Análise de Componentes Independentes (ICA) é uma técnica que separa fontes originais não observadas a partir de sinais misturados, baseando-se na premissa de que essas fontes são estatisticamente independentes e não-gaussianas (exceto no máximo uma). A ICA busca componentes independentes, sendo ideal para cenários de *separação cega de fontes*, como isolar vozes em gravações com múltiplos falantes, limpar artefatos em sinais de EEG/ECG ou extrair padrões não-gaussianos em dados multivariados. Seu funcionamento envolve pré-processar os dados (centralização e *whitening* via PCA), estimar a matriz de separação W e recuperar as fontes originais S por meio de $S = WX$, onde X são os dados observados. As componentes resultantes (S) representam as fontes independentes estimadas, enquanto a matriz de mistura A (inversa de W) revela como cada fonte contribui para os sinais observados.

Considerando as variáveis originais do conjunto *Iris* X_1 (comprimento da sépala), X_2 (largura da sépala), X_3 (comprimento da pétala) e X_4 (largura da pétala) , a matriz de pesos W é dada por:

$$W = \begin{bmatrix} 0{,}031 & -0{,}580 & 0{,}553 & 0{,}597 \\ -0{,}147 & 0{,}797 & 0{,}306 & 0{,}499 \\ -0{,}960 & -0{,}156 & -0{,}212 & 0{,}094 \\ 0{,}235 & -0{,}059 & -0{,}746 & 0{,}620 \end{bmatrix},$$

e as equações para os componentes independentes S_1, S_2, S_3 e S_4, em termos das variáveis originais $X = [X_1, X_2, X_3, X_4]^T$, são:

$$\begin{aligned} S_1 &= 0{,}031X_1 - 0{,}580X_2 + 0{,}553X_3 + 0{,}597X_4 \\ S_2 &= -0{,}147X_1 + 0{,}797X_2 + 0{,}306X_3 + 0{,}499X_4 \\ S_3 &= -0{,}960X_1 - 0{,}156X_2 - 0{,}212X_3 + 0{,}094X_4 \\ S_4 &= 0{,}235X_1 - 0{,}059X_2 - 0{,}746X_3 + 0{,}620X_4 \end{aligned},$$

, substituindo os nomes das variáveis, ficam assim:

$$\begin{aligned} S_1 &= 0{,}031 \cdot \text{Sepal.Length} - 0{,}580 \cdot \text{Sepal.Width} + 0{,}553 \cdot \text{Petal.Length} + 0{,}597 \cdot \text{Petal.Width} \\ S_2 &= -0{,}147 \cdot \text{Sepal.Length} + 0{,}797 \cdot \text{Sepal.Width} + 0{,}306 \cdot \text{Petal.Length} + 0{,}499 \cdot \text{Petal.Width} \\ S_3 &= -0{,}960 \cdot \text{Sepal.Length} - 0{,}156 \cdot \text{Sepal.Width} - 0{,}212 \cdot \text{Petal.Length} + 0{,}094 \cdot \text{Petal.Width} \\ S_4 &= 0{,}235 \cdot \text{Sepal.Length} - 0{,}059 \cdot \text{Sepal.Width} - 0{,}746 \cdot \text{Petal.Length} + 0{,}620 \cdot \text{Petal.Width} \end{aligned}$$

Os componentes independentes resultantes da ICA sobre o conjunto *Iris* mostram perfis morfológicos bem distintos: S_1 realça flores com pétalas longas e largas (cargas 0,553 em `Petal.Length` e 0,597 em `Petal.Width`), associadas a sépalas relativamente estreitas (–0,580 em `Sepal.Width`), característica que se aproxima de *virginica*;S_2, dominado pela largura da

sépala (0,797), soma-se às pétalas de tamanho intermediário, típico de *versicolor*; S_3 capta indivíduos compactos, com sépalas e pétalas curtas (forte carga negativa em `Sepal.Length` e `Petal.Length`), identificando-se claramente com *setosa*; já o S_4 contrapõe pétalas largas (0,620 em `Petal.Width`) a pétalas relativamente curtas (–0,746 em `Petal.Length`), diferenciando contrastes morfológicos entre espécies como *setosa* e *virginica*. (**Sugestão:** consulte um biólogo para validar essas associações e identificar outras nuances).

A matriz de pesos A é dada por:

$$A = \begin{bmatrix} 0{,}957 & -0{,}118 & -0{,}251 & 0{,}015 \\ 0{,}155 & 0{,}975 & 0{,}130 & -0{,}043 \\ 0{,}211 & -0{,}135 & 0{,}889 & 0{,}376 \\ -0{,}094 & 0{,}102 & -0{,}352 & 0{,}922 \end{bmatrix}$$

A matriz de mistura A revela que o componente independente 1 (IC1) é quase inteiramente explicado pelo comprimento da sépala (`Sepal.Length`, carga 0,957), enquanto o IC2 capta sobretudo a largura da sépala (`Sepal.Width`, carga 0,975). O IC3 reflete uma combinação das dimensões da pétala, com contribuição principal de `Petal.Length` (0,889) e secundária de `Petal.Width` (0,376); já o IC4 é dominado pela largura da pétala (`Petal.Width`, carga 0,922). Esta estrutura revela que os ICs decompõem a morfologia floral em dois blocos independentes: dimensões da sépala (IC1–IC2) e configurações da pétala (IC3–IC4). (**Sugestão:** consulte um biólogo para mais esclarecimentos). A Figura 9 mostra a representação gráfica dos quatro componentes.

A escolha do número de componentes independentes (ICs) na ICA não é trivial, devido à ausência de métricas diretas como a variância explicada na PCA. A Norma de Frobenius é uma abordagem prática para quantificar o erro de reconstrução entre os dados originais padronizados ($\mathbf{X}_{\text{pad}}$) e a versão reconstruída ($\hat{\mathbf{X}}$) com K ICs; busca-se um ponto de "cotovelo" no gráfico da norma versus o número de ICs que indique saturação do ganho descritivo. Assim, com base na Figura 10, escolhemos dois componentes.

Note que vários gráficos como gráficos dos dados, CP's e CI's podem ser efitos (veja `help(fastICA)`).

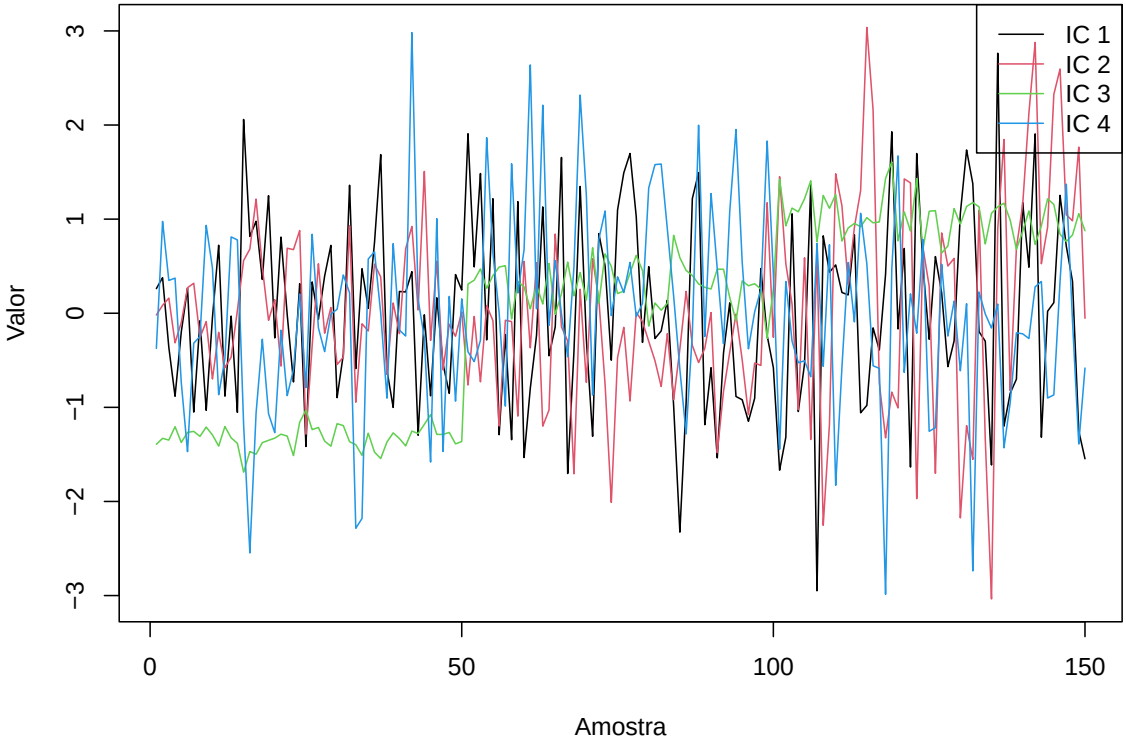

Figura 9: Componentes Independentes.

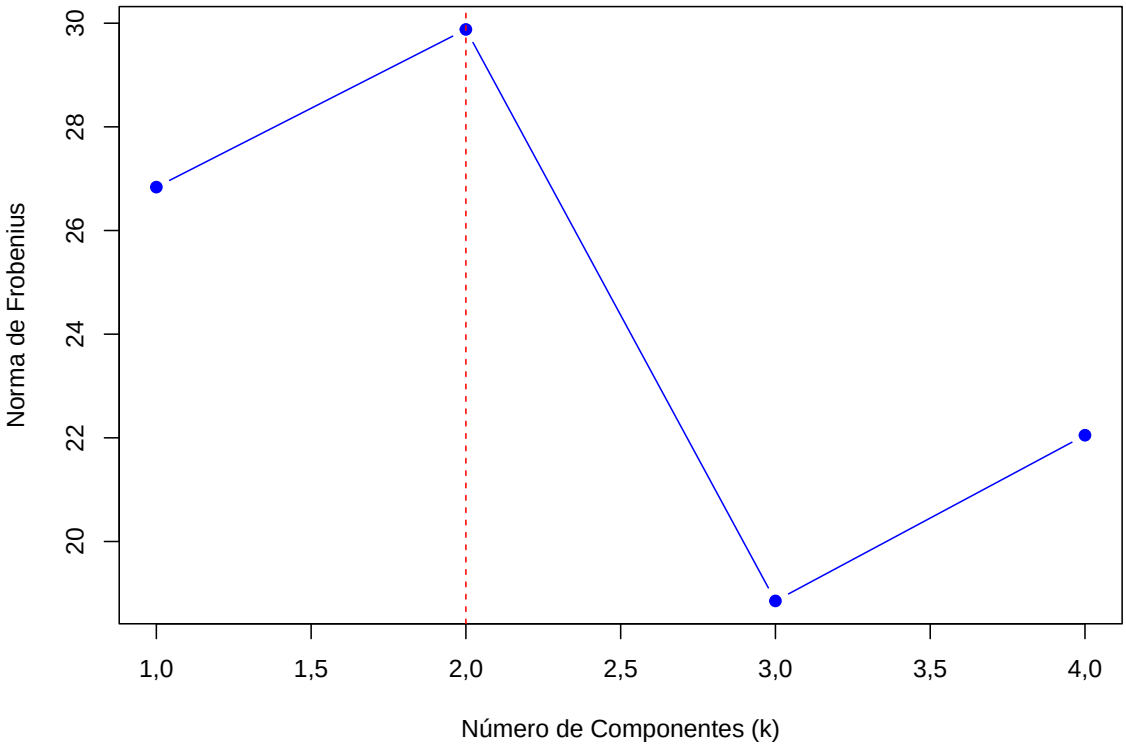

Figura 10: Seleção de Componentes ICA - Iris

EXERCÍCIO 04: Análise de componentes principais e fatorial para dados Boston

O conjunto de dados Boston Housing contém 506 observações e 14 variáveis que descrevem características socioeconômicas e urbanas de subúrbios de Boston, sendo:`crim` (taxa de criminalidade *per capita*), `zn` (proporção de terrenos residenciais zoneados para lotes $> 25.000\ ps^2$), `indus` (proporção de acres não comerciais por cidade), `chas` (dummy do rio Charles: 1 se limita o rio, 0 caso contrário),`nox` (concentração de óxidos de nitrogênio em ppm), `rm` (número médio de cômodos por residência),
`age` (proporção de unidades ocupadas pelo proprietário construídas antes de 1940), `dis` (distância média ponderada aos centros empregatícios), `rad` (índice de acesso a rodovias radiais), `tax` (taxa de imposto predial por U$10.000),`ptratio` (proporção aluno-professor),`black` (escala de proporção de residentes negros: $1000(B_k - 0.63)^2$), `lstat` (percentual de população de baixo status socioeconômico), e `medv` (valor mediano de casas ocupadas pelos proprietários em U$1.000), sendo `medv` a variável-alvo típica para modelagem de preços imobiliários.

Tabela 8: Banco de dados Boston

crim	zn	indus	chas	nox	rm	age	dis	rad
0.00632	18	2.31	0	0.538	6.58	65.2	4.09	1
0.02731	0	7.07	0	0.469	6.42	78.9	4.97	2
0.02729	0	7.07	0	0.469	7.18	61.1	4.97	2
0.03237	0	2.18	0	0.458	7.00	45.8	6.06	3
0.06905	0	2.18	0	0.458	7.15	54.2	6.06	3
0.02985	0	2.18	0	0.458	6.43	58.7	6.06	3

Realizaremos a análise de correlação para identificar as variáveis que têm uma relação significativa, facilitando a seleção das características mais relevantes (Figura 11). Cada gráfico de dispersão na Figura 11 permite visualizar a relação entre duas variáveis, enquanto os histogramas na diagonal representam a distribuição de cada variável individualmente.

Os coeficientes de correlação são apresentados acima de cada gráfico de dispersão, indicando a força e a direção da relação linear entre as variáveis. Note que a correlação Spearman's (Figura 12) não foi diferente da do Pearson.

Observe que a variável dummy (chas), referente a fazer limite com o rio Charles, apresentou baixa correlação. Por ser dummy, vamos considerar ela como variaűel resposta e regredir-a com base nas demais variávies (Figura 13). No entanto, observa-se que não existe diferença significativa entre as variáveis respostas (Intervalos de confiança sobrepostos).

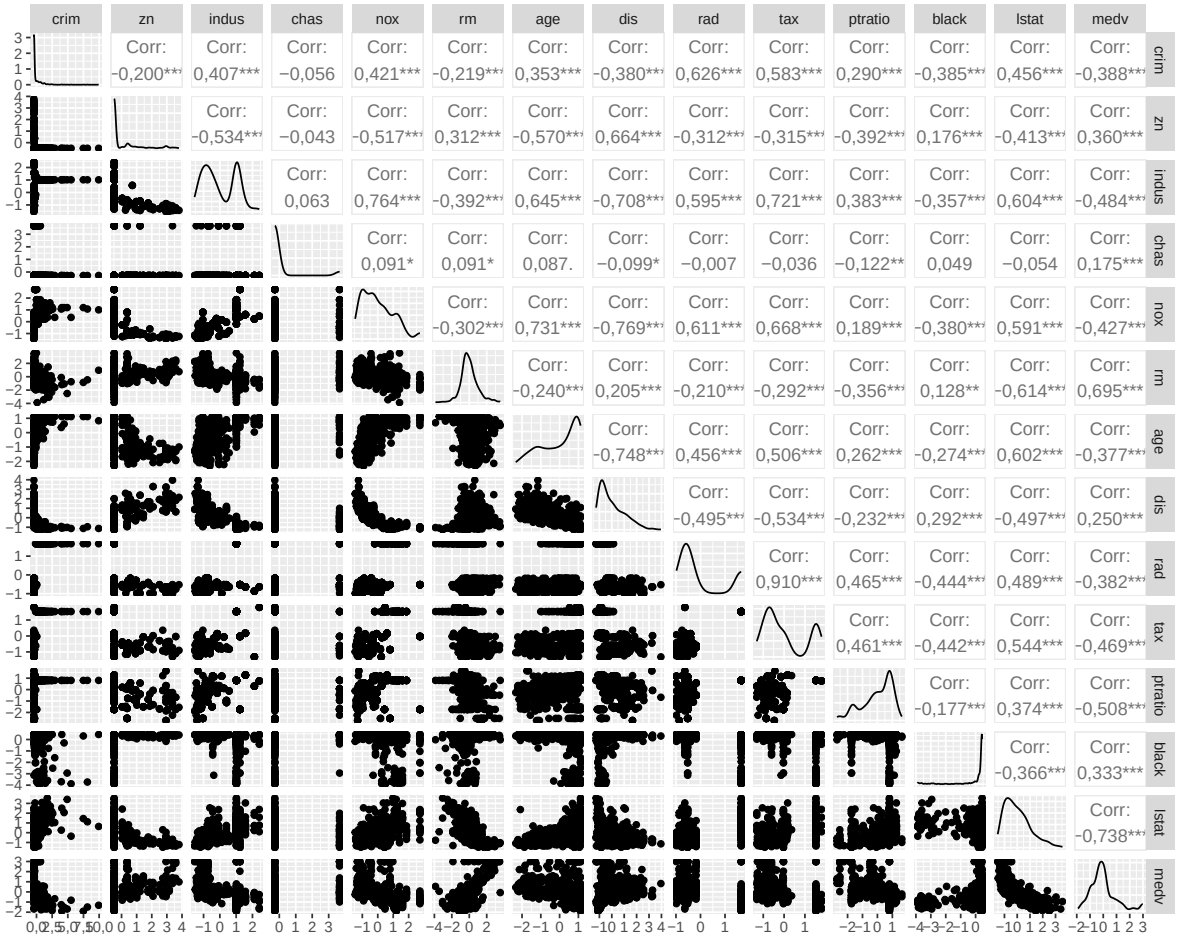

Figura 11: Matriz de gráficos dos dados Boston

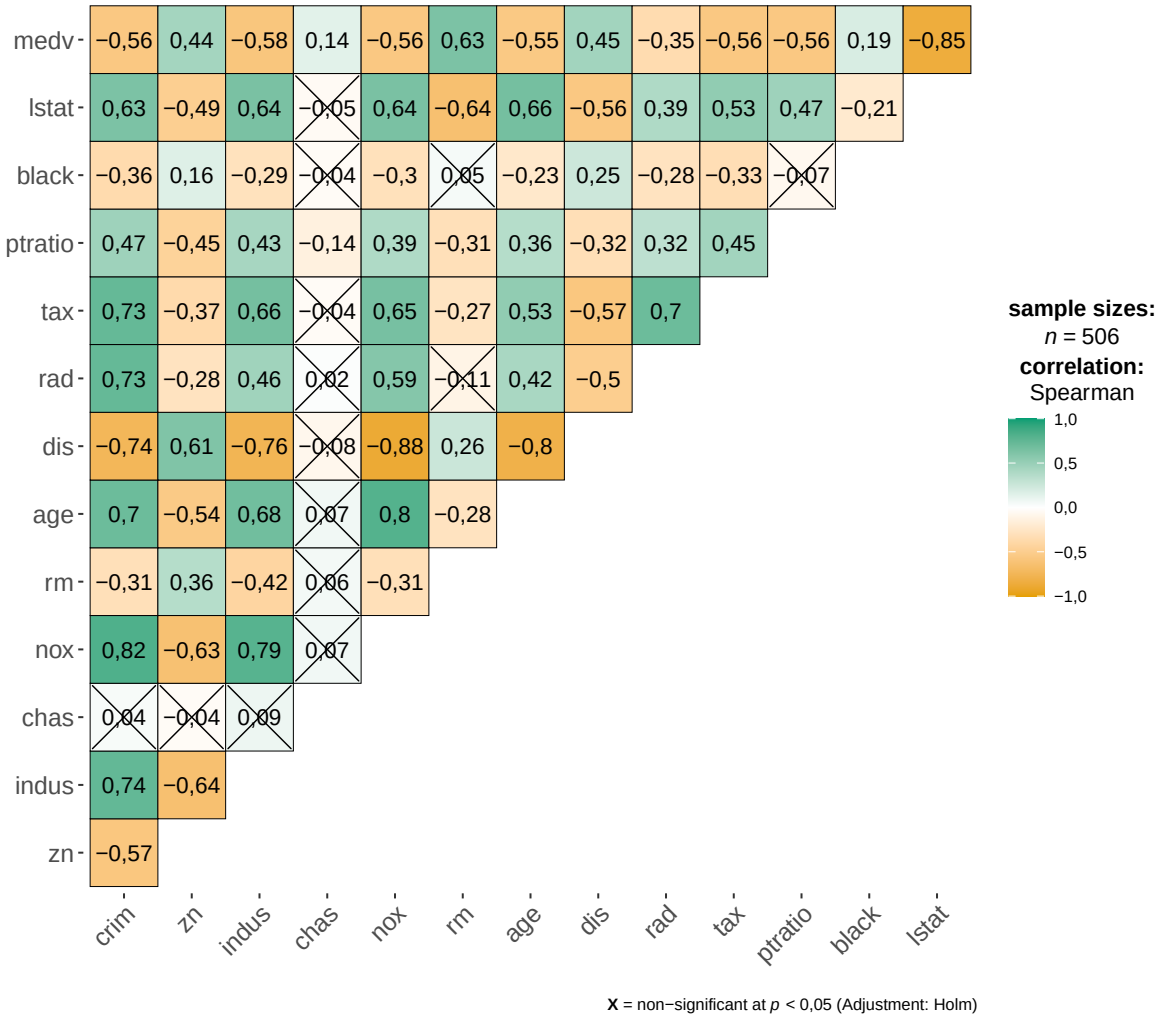

Figura 12: Matriz de correlação de Spearman para dados Boston

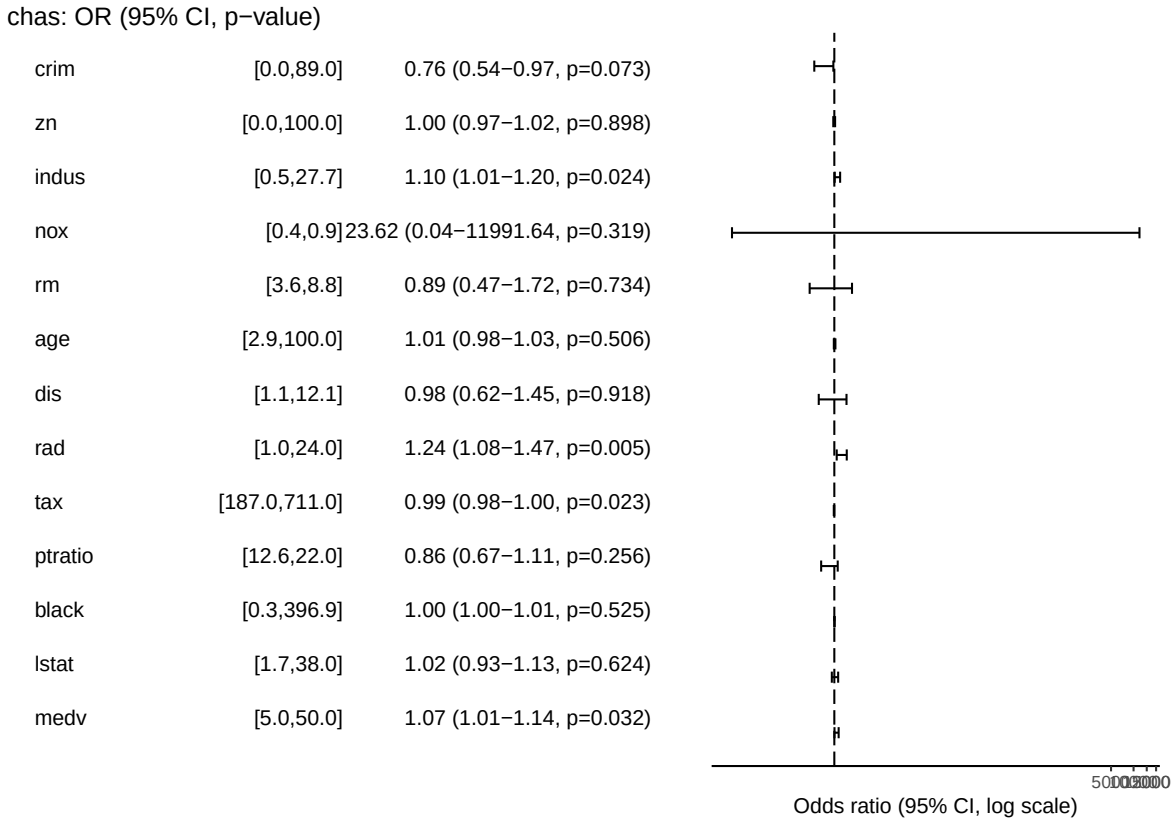

Figura 13: Tabela e gráfico da razão de chances

Portanto, vamos selecionar as variáveis baseando-se na correlação, conforme apresnetado na Figura 14.

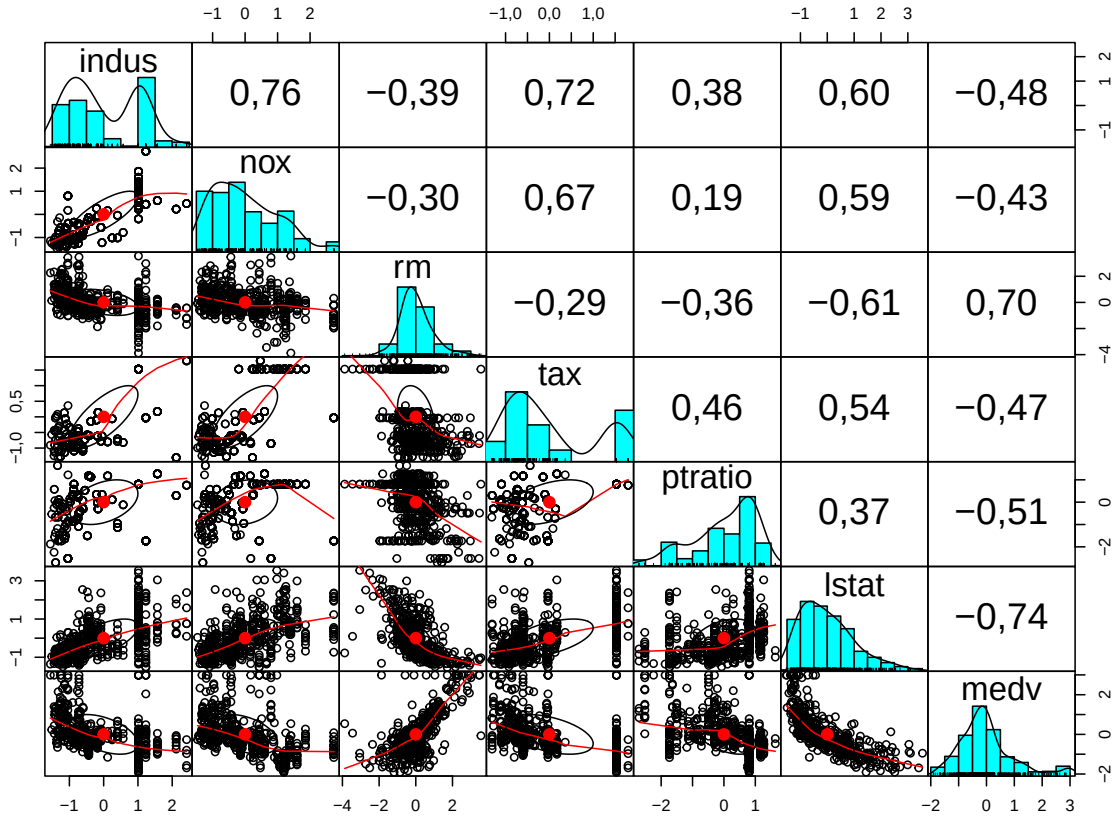

Figura 14: Matriz de dispersão das variáveis dos dados Boston

Tabela 9: Componentes principais para o conjunto de dados Boston

	Componentes principais (CP)						
	CP1	**CP2**	**CP3**	**CP4**	**CP5**	**CP6**	**CP7**
Variável							
indus	0,41	0,36	-0,01	0,38	0,33	0,19	-0,64
nox	0,38	0,48	0,23	0,02	0,30	-0,43	0,54
rm	-0,34	0,51	-0,33	-0,65	0,20	-0,08	-0,23
tax	0,39	0,36	-0,29	-0,05	-0,79	0,08	0,05
ptratio	0,29	-0,29	-0,82	0,05	0,32	0,02	0,24
lstat	0,42	-0,13	0,28	-0,54	0,15	0,62	0,13
medv	-0,40	0,39	-0,09	0,36	0,07	0,61	0,40
Sd	2,02	1,06	0,89	0,58	0,52	0,47	0,44
% Variância	58,00	16,00	11,00	5,00	4,00	3,00	3,00
% Acumulada	58,00	74,00	85,00	90,00	94,00	97,00	100,00

[!h]

Nota:
indus: Proporção de acres não comerciais por cidade;
nox: Concentração de NOx (partículas por milhão);
rm: Número médio de cômodos por residência;
tax: Taxa de imposto predial por U$10.000;
ptratio: Proporção aluno-professor;
lstat: % da população com baixo status socioeconômico;
medv: Valor mediano das casas (em milhares de dólares).

Análise de Componentes Principais (PCA)

A descrição é a mesma fornecida no exercício 1, desde a padronização dos dados, pacotes, etc. Sem querer ser tão rigoroso, com base na Tabela 9, dois componentes a 4 seriam suficientes. Note que não existe um ponto de corte padrão recomendável, sendo a escolha subjetiva.

Uma alternativa para a escolha dos componentes além do uso da variância (%) acomulada é o uso do `Screenplot`.

Assim, pela regra do cotovelo (Figura 15), o corte ideal ocorre após o terceiro componente principal. Note que o segundo componente (PC2) ainda aporta 11,2% de variância elevando o acumulado a 27,3%, enquanto o terceiro (PC3) adiciona apenas mais 4,8% (totalizando 32,1%). A partir daí, observa-se um decréscimo brusco no ganho marginal: todos os PCs subsequentes contribuem com menos de 4% cada, o que justificaria reter apenas os três primeiros componentes. Por simplicidade vamos considerar apenas dois componentes.

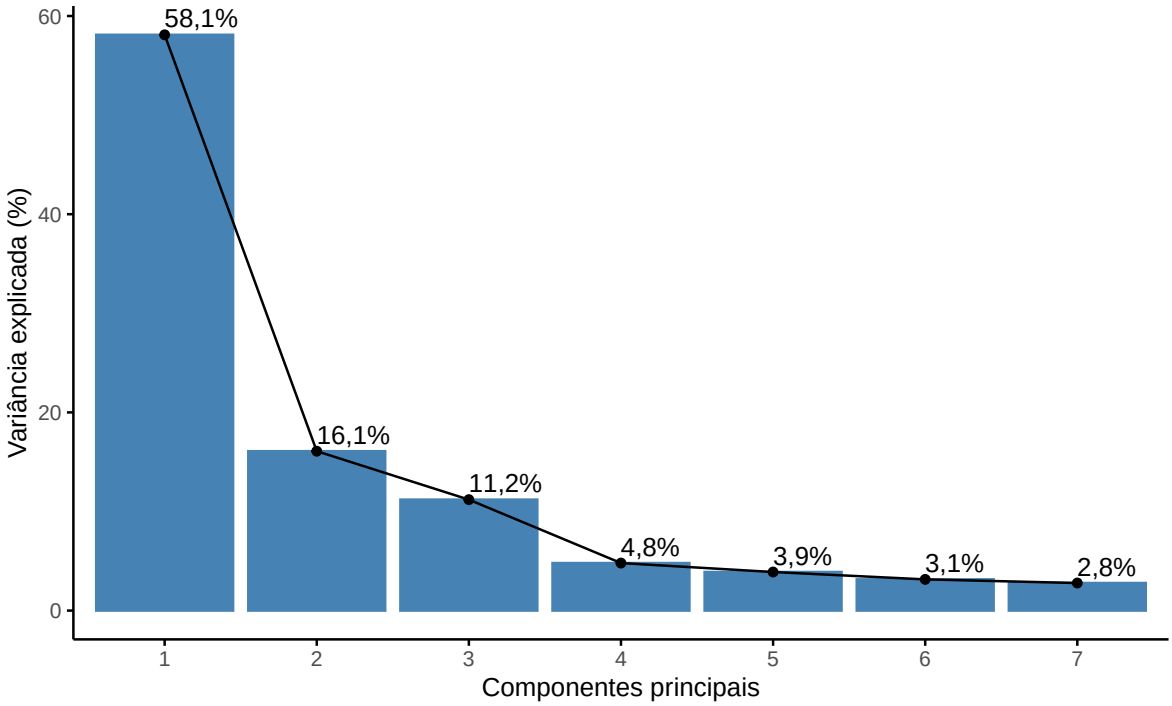

Figura 15: Screeplot para dados Boston.

Na Figura 16, observa-se que o primeiro componente principal (PC1, responsável por 58% da variância) diferencia áreas industriais (indus), com níveis elevados de óxidos de nitrogênio (nox), altos impostos (tax) e maior percentual de população de baixa renda (lstat), cujas casas são, em média, menores (rm) e menos valiosas (medv), das áreas essencialmente residenciais, onde a poluição e a pobreza são menores, as construções são maiores e os imóveis mais caros; já o segundo componente principal (PC2, responsável por 16% da variância), disposto verticalmente, distingue, no topo, locais cujas residências são maiores e mais valorizadas, mas ainda submetidas a níveis de poluição mais elevados, e, na parte inferior, bairros marcados por turmas escolares mais numerosas (ptratio elevado) e melhor qualidade do ar, sem poluição excessiva, sendo que cada ponto sobreposto no gráfico identifica um bairro específico posicionado conforme essas características, de modo que um ponto no quadrante inferior direito indicaria um bairro com casas caras e escolas menos sobrecarregadas, enquanto um ponto no quadrante superior esquerdo corresponderia a uma área industrial, poluída e de residências pequenas.

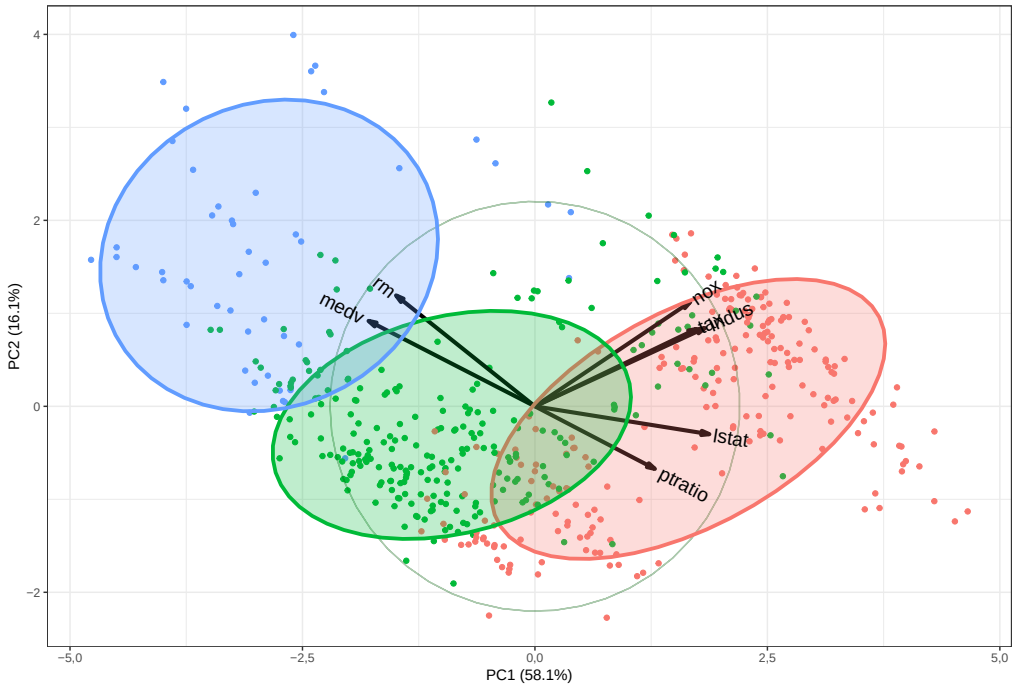

Figura 16: Biplot PCA para dados Boston

Tabela 10: Medidas de Adequação Amostral (KMO) para dados Boston

[!h]

	Medidas de Adequação Amostral (KMO)
Variável	**MSAi**
Medidas Individuais (MSAi)	
indus	0,827
nox	0,778
rm	0,824
tax	0,851
ptratio	0,753
lstat	0,872
medv	0,802
MSA Global	
MSA	0,820

MSAi: Avalia se cada variável individual se relaciona adequadamente com as demais
MSA: Avalia se todo o conjunto de dados é adequado para Análise Fatorial.

Análise Fatorial

Para esta parte, seguimos o mesmo descrito no exercício 2 acima.

Na Tabela 10, observa-se MSAi individuais, indus (0,827), rm (0,824), tax (0,851), lstat (0,872) e medv (0,802) com classificação Bom, enquanto nox (0,778) e ptratio (0,753) são Médios, indicando que o conjunto de variáveis é adequado para análise fatorial, com forte relação entre a maioria das variáveis (especialmente lstat, tax e indus).

Observa-se na Tabela @bl-bartletF4 que o valor-p<0.05, portanto, rejeita-se H_0 e portanto o AF é recomendável. Nota: Para mais detalhes vide Exercício 2.

Tabela 11: Teste de Bartlett

Bartlett.s.K.squared	df	p.value
2,139	21	0

Na Tabela 12, observa-se que apenas dois fatores apresentam autovalor maior que 1. Assim, pelo critério de Kaiser o ideal seria consideral dois fatores.

Tabela 12: Retenção de fatores pelo critério de Kaiser para dados Boston

Fator	Autovalor	% Variância	% Acumulada
1	4.067	58.1	58.1
2	1.126	16.1	74.2
3	0.784	11.2	85.4
4	0.336	4.8	90.2
5	0.272	3.9	94.1
6	0.220	3.1	97.2
7	0.195	2.8	100.0

A Figura 17 mostra que pelo critério de cotovelo, reteríamos dois fatores.

```
Parallel analysis suggests that the number of factors =  1  and the number of components =  1
```

Na Tabela 13 o primeiro fator está fortemente ligado à industrialização e custos urbanos, representado por maior presença de negócios não comerciais (`indus` = 0.85), poluição do ar (`nox` = 0.82) e altos impostos (`tax` = 0.74), enquanto o segundo fator reflete condições habitacionais e sociais, mostrando que áreas com casas de menor valor (`medv` = -0.90) e menos cômodos (`rm` = -0.73) tendem a ter mais população de baixa renda (`lstat` = 0.68). Estes dois fatores explicam bem as diferenças entre bairros, especialmente para valor das casas (87.7% explicado) e presença industrial (80.6%), mas têm menor "poder" para explicar a proporção aluno-professor (apenas 28.3% explicado), vide Figura 18 para uma visualização gráfica.

Assim, considerando um modelo de análise fatorial com dois fatores comuns F_1 e F_2 para as variáveis observadas $\{\mathsf{medv}, \mathsf{indus}, \mathsf{nox}, \mathsf{lstat}, \mathsf{tax}, \mathsf{rm}, \mathsf{ptratio}\}$, as respectivas equações ficam dadas por (note que o mesmo poderia ter sido feito no exercício 2):**

$$
\begin{cases}
\mathsf{medv} &= -0.264\,F_1 - 0.899\,F_2 + \varepsilon_{\mathsf{medv}}, \\
\mathsf{indus} &= 0.848\,F_1 + 0.296\,F_2 + \varepsilon_{\mathsf{indus}}, \\
\mathsf{nox} &= 0.820\,F_1 + 0.230\,F_2 + \varepsilon_{\mathsf{nox}}, \\
\mathsf{lstat} &= 0.489\,F_1 + 0.678\,F_2 + \varepsilon_{\mathsf{lstat}}, \\
\mathsf{tax} &= 0.741\,F_1 + 0.297\,F_2 + \varepsilon_{\mathsf{tax}}, \\
\mathsf{rm} &= -0.177\,F_1 - 0.728\,F_2 + \varepsilon_{\mathsf{rm}}, \\
\mathsf{ptratio} &= 0.245\,F_1 + 0.473\,F_2 + \varepsilon_{\mathsf{ptratio}}.
\end{cases}
$$

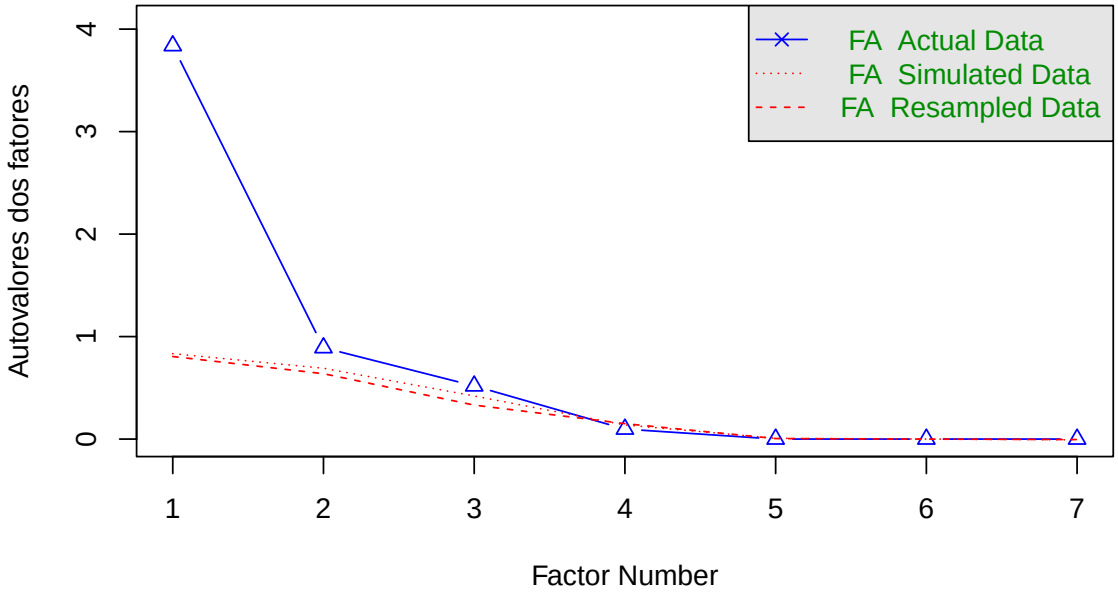

Figura 17: Screenplot da análise fatorial para dados Boston.

Tabela 13: Cargas Fatoriais e Comunalidades para dados Boston

Variável	Fator1	Fator2	Comunalidade
medv	-0.264	-0.899	0.877
indus	0.848	0.296	0.806
nox	0.820	0.230	0.726
lstat	0.489	0.678	0.699
tax	0.741	0.297	0.638
rm	-0.177	-0.728	0.561
ptratio	0.245	0.473	0.283

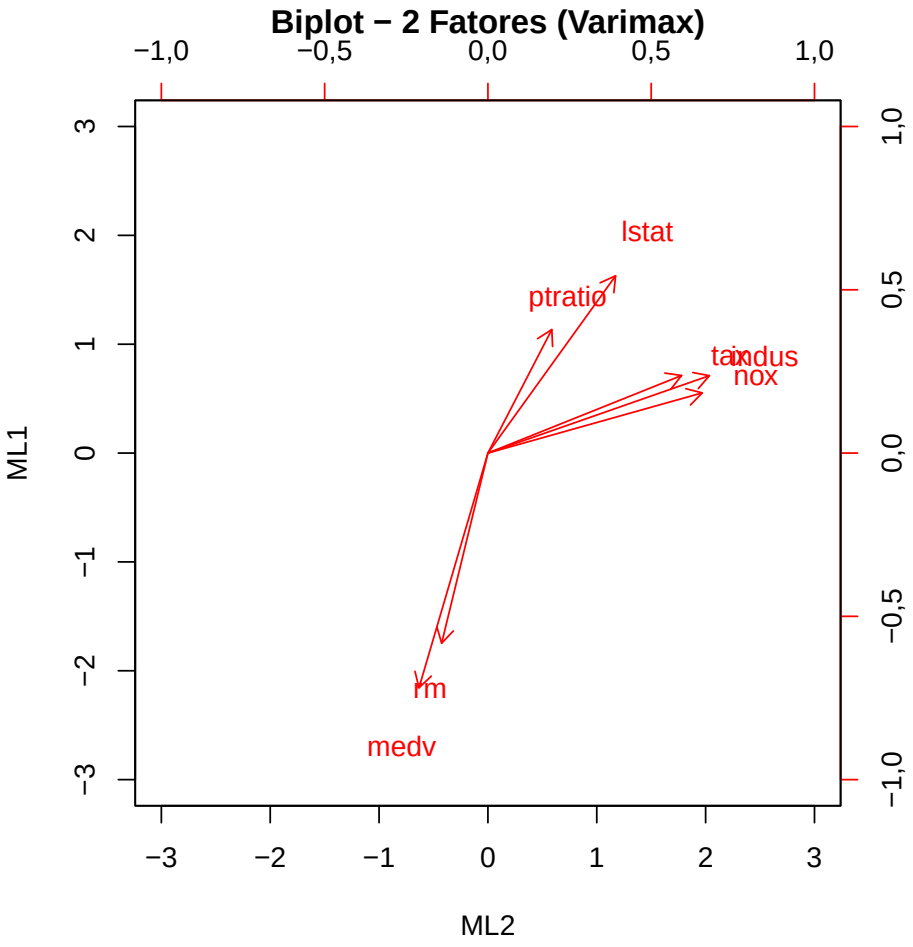

Figura 18: Biplot da análise fatorial dados Boston

Referência

Morettin, Pedro Alberto, and Júlio da Motta Singer. *Estatística e ciência de dados.* (2025).

Taherdoost, H. A. M. E. D., S. H. A. M. S. U. L. Sahibuddin, and N. E. D. A. Jalaliyoon. **Exploratory factor analysis; concepts and theory.** Advances in applied and pure mathematics 27 (2014): 375-382.

Bartlett, M. S., (1951), **The Effect of Standardization on a chi square Approximation in Factor Analysis**, Biometrika, 38, 337-344.

Fávero, Luiz Paulo, and Patrícia Belfiore. **Manual de Análise de Dados: Estatística e Machine Learning com Excel, SPSS, Stata, R e Python**. Elsevier Brasil, 2024.

Hyvärinen, A., & Oja, E. (2000). **Independent component analysis: algorithms and applications.** Neural networks, 13(4-5), 411-430.

Jolliffe, I. T., & Cadima, J. (2016). **Principal component analysis: a review and recent developments.** Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

Códigos

Exercício 1

Tabela de dados brutos

```
data(iris)

iris%>%
  group_by(Species)%>%
  slice_head(n=2)%>%
  ungroup()%>%
  flextable()
```

Estatísticas descritivas

```
iris %>%
  tbl_summary(
    by = Species,
    statistic = all_continuous() ~
      "{mean} ({sd})",
    digits = all_continuous() ~ 2
  )%>%
  modify_header(label ~ "**Variável**")%>%
  modify_footnote_header("Média (Desvio Padrão)", columns = all_stat_cols())
```

Correlações

```
pairs.panels(iris[,-5],
gap = 0,
bg = c("black", "yellow", "red")[iris$Species],
pch=21) #Quinta coluna é qualitativa (nome das espécies)
```

PCA

```r
iris_standar <- scale(iris[, 1:4])
pca <- prcomp(iris_standar, center = TRUE, scale. = TRUE)

componentes <- pca$rotation[, 1:4]
estatisticas <- data.frame(
  Sd = pca$sdev[1:4],
  Prop_Var = pca$sdev^2 / sum(pca$sdev^2),
  Cum_Prop = cumsum(pca$sdev^2 / sum(pca$sdev^2))
)

componentes <- round(componentes, 2)
estatisticas$Sd <- round(estatisticas$Sd, 2)
estatisticas$Prop_Var <- round(estatisticas$Prop_Var, 2)
estatisticas$Cum_Prop <- round(estatisticas$Cum_Prop, 2)

tabela_final <- rbind(
  componentes,
  "Sd" = c(estatisticas$Sd, rep(NA, 0)),
  "% Variância" = c(estatisticas$Prop_Var * 100, rep(NA, 0)),
  "% Acumulada" = c(estatisticas$Cum_Prop * 100, rep(NA, 0))
)

colnames(tabela_final) <- c("CP1", "CP2", "CP3", "CP4")
rownames(tabela_final) <- c(
  "Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width",
  "Sd", "% Variância", "% Acumulada"
)


kable(
  tabela_final,
  format.args = list(big.mark = ".", decimal = ","),
  booktabs    = TRUE,
  align       = "ccccc"
) %>%
  formartaKable() %>%
  pack_rows("Variável", 1, 4, bold = TRUE) %>%
  add_header_above(
    header = c(" " = 1, "Componentes principais (CP)" = 4),
    bold   = TRUE
  )
```

Scree plot

```
fviz_eig(pca,ggtheme = theme_classic(),addlabels = TRUE,
         xlab = "Componentes principais",ylab = "Variância explicada (%)", main = "") #Scree
```

PCA final

```
#Biplots
create_biplot <- function(pca, scores, title, var_exp, scale_arrows = 1) {
  loadings <- as.data.frame(pca$rotation[, 1:2] * scale_arrows)
  colnames(loadings) <- c("x", "y")
  loadings$variavel <- rownames(loadings)
  ratio <- max(abs(scores[, 1:2])) / max(abs(loadings[, 1:2])) * 0.8
  ggplot(scores, aes(PC1, PC2, color = Species)) +
    geom_point(size = 3, alpha = 0.8) +
    stat_ellipse(level = 0.95, linewidth = 0.7) +
    geom_segment(
      data = loadings,
      aes(x = 0, y = 0, xend = x * ratio, yend = y * ratio),
      arrow = arrow(length = unit(0.25, "cm")),
      color = "black",
      inherit.aes = FALSE
    ) +
    geom_text_repel(
      data = loadings,
      aes(x = x * ratio, y = y * ratio, label = variavel),
      color = "black",
      size = 4,
      fontface = "bold",
      inherit.aes = FALSE
    ) +
    labs(title = title,
         x = paste0("PC1 (", var_exp[1], "%)"),
         y = paste0("PC2 (", var_exp[2], "%)")) +
    theme_bw(base_size = 12) +
    theme(legend.position = "bottom") +
    coord_equal()
}

# PCA sem padronização
```

```r
pca_orig <- prcomp(iris[, -5], scale. = FALSE)
scores_orig <- as.data.frame(pca_orig$x)
scores_orig$Species <- iris$Species
var_exp_orig <- round(100 * pca_orig$sdev^2 / sum(pca_orig$sdev^2), 2)

# PCA com padronização
pca_scale <- prcomp(iris[, -5], scale. = TRUE)
scores_scale <- as.data.frame(pca_scale$x)
scores_scale$Species <- iris$Species
var_exp_scale <- round(100 * pca_scale$sdev^2 / sum(pca_scale$sdev^2), 2)


g1 <- create_biplot(pca_orig, scores_orig,
                    "A) PCA sem Padronização",
                    var_exp_orig,
                    scale_arrows = 1)

g2 <- create_biplot(pca_scale, scores_scale,
                    "B) PCA com Padronização",
                    var_exp_scale,
                    scale_arrows = 1)


g1+g2 + plot_layout(guides = "collect") & theme(legend.position = "bottom")
```

Correlações final

```r
pairs.panels(pca$x,
gap=0,
bg = c("black", "yellow", "red")[iris$Species],
pch=21)
```

Exercício 2

Tabela KMO

```r
data(iris)
iris_scaled <- scale(iris[, 1:4])
kmo <- KMO(iris_scaled)
```

```r
corrplot(kmo$Image,
         method = "color",
         type = "upper",
         order = "original",
         tl.col = "darkblue",
         tl.srt = 15,
         addCoef.col = "black",
         number.cex = 0.8,
         diag = FALSE)
```

Tabela MSA

```r
kmo_df <- data.frame(
  Variável = names(kmo$MSAi),
  MSAi = unname(kmo$MSAi)
)

msa_global <- data.frame(
  Variável = "MSA",
  MSAi = kmo$MSA
)

kmo_df_final <- rbind(kmo_df, msa_global)

kable(
  kmo_df_final,
  format      = "latex",
  format.args = list(big.mark = ".", decimal = ","),
  booktabs    = TRUE,
  align       = "lc"
) %>%
  formartaKable() %>%
  pack_rows("Medidas Individuais (MSAi)", 1, nrow(kmo_df_final) - 1) %>%
  pack_rows("MSA Global", nrow(kmo_df_final), nrow(kmo_df_final)) %>%
  add_header_above(
    header = c(" " = 1, "Medidas de Adequação Amostral (KMO)" = 1),
    bold   = TRUE
  ) %>%
  footnote(
    general = c(
```

```
      "MSAi: Avalia se cada variável individual se relaciona adequadamente com as demais",
      "MSA: Avalia se todo o conjunto de dados é adequado para Análise Fatorial."
    ),
    general_title = ""
  )
```

Teste de Bartlett

```
species <- iris$Species
dados <- as.vector(as.matrix(iris_scaled))
grupos <- factor(rep(species, each = ncol(iris_scaled)))
bartlett_result_psych <- bartlett.test(dados, grupos)

resultados_bartlett <- data.frame(
  "Bartlett's K-squared"=bartlett_result_psych$statistic,
  "df"=bartlett_result_psych$parameter,
  "p-value" = c(bartlett_result_psych$p.value
))

resultados_bartlett%>%
  flextable()
```

Autovalores

```
cor_matrix <- cor(iris_scaled)
eigen_values <- eigen(cor_matrix)$values

eigen_df <- data.frame(
  Fator = 1:length(eigen_values),
  Autovalor = round(eigen_values, 3),
  `% Variância` = round(eigen_values / sum(eigen_values) * 100, 1),
  `% Acumulada` = round(cumsum(eigen_values) / sum(eigen_values) * 100, 1)
)

colnames(eigen_df) <- c("Fator", "Autovalor", "% Variância", "% Acumulada")
eigen_df%>%
  flextable()
```

Scree plot AF

```
set.seed(123)
parallel <- fa.parallel(iris_scaled, nfactors=2, fa = "fa",
                        main="", ylabel="Autovalores dos fatores")
```

Cargas fatoriais

```
set.seed(123)
fa1_mle <- fa(
  r = iris_scaled,
  nfactors = 1,
  rotate = "none",
  fm = "mle",
  scores = "regression"
)

fa_varimax_mle <- fa(
  r = iris_scaled,
  nfactors = 2,
  rotate = "varimax",
  fm = "mle",
  scores = "regression"
)

results_fa1 <- data.frame(
  Variável = rownames(fa1_mle$loadings),
  Carga_Fator1 = unname(round(fa1_mle$loadings[,1], 3)),
  Comunalidade = unname(round(fa1_mle$communalities, 3))
)

vaccounted_fa1 <- data.frame(
  Componente = rownames(fa1_mle$Vaccounted),
  Valor = unname(round(fa1_mle$Vaccounted[,1], 3))
)

loadings_df <- data.frame(
  Variável = rownames(fa_varimax_mle$loadings),
  Fator1 = unname(round(fa_varimax_mle$loadings[,1], 3)),
```

```
  Fator2 = unname(round(fa_varimax_mle$loadings[,2], 3)),
  Comunalidade = unname(round(apply(fa_varimax_mle$loadings^2, 1, sum), 3))
)

vaccounted_fa2 <- data.frame(
  Componente = rownames(fa_varimax_mle$Vaccounted),
  Fator1 = round(fa_varimax_mle$Vaccounted[,"ML1"], 3),
  Fator2 = round(fa_varimax_mle$Vaccounted[,"ML2"], 3)
)

cor_fatores <- round(cor(fa_varimax_mle$scores), 4)


comunalidades <- data.frame(
  Variável = names(sort(fa1_mle$communality, decreasing = TRUE)),
  Mod_1Fat = round(sort(fa1_mle$communality, decreasing = TRUE), 3),
  Mod_2Fat = round(sort(fa_varimax_mle$communality, decreasing = TRUE), 3)
)

cargas_comunalidades <- data.frame(
  Variável = results_fa1$Variável,
  C_1Fator = results_fa1$Carga_Fator1,
  Comun_1Fat = results_fa1$Comunalidade,
  Carga_F1_2Fat = loadings_df$Fator1,
  Carga_F2_2Fat = loadings_df$Fator2,
  Comun_2Fat = loadings_df$Comunalidade
)

cargas_comunalidades <- cargas_comunalidades[order(cargas_comunalidades$Comun_2Fat, decreasi

cargas_comunalidades_formatted <- cargas_comunalidades %>%
  mutate(across(where(is.numeric), round, 3))

vaccounted_fa1 <- data.frame(
  Componente = rownames(fa1_mle$Vaccounted),
  Valor = unname(round(fa1_mle$Vaccounted[,1], 3))
)

vaccounted_fa2 <- data.frame(
  Componente = rownames(fa_varimax_mle$Vaccounted),
  Fat1 = round(fa_varimax_mle$Vaccounted[,"ML1"], 3),
  Fat2 = round(fa_varimax_mle$Vaccounted[,"ML2"], 3)
```

```
)

variancia_explicada <- data.frame(
  Modelo = c("1 Fat", "2 Fat", "2 Fat"),
  Componente = c("Prop Var", "Prop Var F1", "Prop Var F2"),
  Valor = c(
    vaccounted_fa1$Valor[vaccounted_fa1$Componente == "Proportion Var"],
    vaccounted_fa2$Fat1[vaccounted_fa2$Componente == "Proportion Var"],
    vaccounted_fa2$Fat2[vaccounted_fa2$Componente == "Proportion Var"]
  ),
  Cumulativa = c(
    vaccounted_fa1$Valor[vaccounted_fa1$Componente == "Cumulative Var"],
    vaccounted_fa2$Fat1[vaccounted_fa2$Componente == "Cumulative Var"] +
    vaccounted_fa2$Fat2[vaccounted_fa2$Componente == "Cumulative Var"]
  )
)

comunalidades_comparacao <- comunalidades %>%
  mutate(Diferença = Mod_2Fat - Mod_1Fat)

correlacao_fatores <- data.frame(
  Fat1 = c(1, cor_fatores[1,2]),
  Fat2 = c(cor_fatores[2,1], 1)
)
rownames(correlacao_fatores) <- c("F1", "F2")

cargas_comunalidades_formatted%>%
  flextable()
```

Biplot AF

```
biplot.psych(
  fa_varimax_mle,
  main = "Biplot - 2 Fatores (Varimax)",
  col = c("black", "red", "blue"),
  pch = c(21, 22, 23),
  group = as.numeric(iris$Species)
)
```

Diagrama AF

```
fa.diagram(fa_varimax_mle)
```

Exercício 3

Componentes Independentes

```
set.seed(123)
data(iris)
X <- iris[, 1:4]

X_centered <- scale(X, center = TRUE, scale = FALSE)

whitening <- function(X) {
  cov_matrix <- cov(X)
  eigen_decomp <- eigen(cov_matrix)
  whitening_matrix <- eigen_decomp$vectors %*%
    diag(1 / sqrt(eigen_decomp$values)) %*%
    t(eigen_decomp$vectors)
  X_whitened <- X %*% whitening_matrix
  return(X_whitened)
}


X_whitened <- whitening(X_centered)

n_components <- 4
ica_result <- fastICA(X_whitened, n.comp = n_components)

S <- ica_result$S

A <- round(ica_result$A, 3)

W <- round(ica_result$W, 3)

matplot(S, type = "l", lty = 1,
        col = 1:ncol(S),
        main = "",
```

```
        xlab = "Amostra",
        ylab = "Valor")
legend("topright",
       legend = paste("IC", 1:ncol(S)),
       col = 1:ncol(S),
       lty = 1)
```

Seleção de Componentes ICA

```
set.seed(123)
data(iris)
X <- iris[, 1:4]

X_centered <- scale(X, center = TRUE, scale = FALSE)
cov_matrix <- cov(X_centered)
eigen_decomp <- eigen(cov_matrix)
whitening_matrix <- eigen_decomp$vectors %*% diag(1/sqrt(eigen_decomp$values)) %*% t(eigen_de
X_whitened <- X_centered %*% whitening_matrix

frobenius_norm <- function(k) {
  S_k <- S[, 1:k, drop = FALSE]
  A_k <- A[, 1:k, drop = FALSE]
  X_whitened_hat <- S_k %*% t(A_k)
  X_centered_hat <- X_whitened_hat %*% solve(whitening_matrix)
  sqrt(sum((X_centered - X_centered_hat)^2))
}

norms <- sapply(1:4, frobenius_norm)

plot(1:4, norms, type = "b", pch = 19, col = "blue",
     xlab = "Número de Componentes (k)", ylab = "Norma de Frobenius",
     main = "")
abline(v = 2, lty = 2, col = "red")
```

Exercício 4

Seleção de variáveis

```
cor_matrix <- cor(Boston1)
ggpairs(scale(Boston1))
```

```
cor_matrix <- cor(Boston1)
cor_medv <- cor_matrix[,"medv"]
important_vars <- names(which(abs(cor_medv) > 0.4))
boston_selected <- Boston[, important_vars]
boston_selected <- scale(boston_selected)
ggcorrmat(Boston, type="nonparametric")
```

```
explanatory = c("crim", "zn","indus","nox","rm","age","dis",
                "rad","tax","ptratio", "black","lstat","medv")
dependent = "chas"
Boston %>%
  or_plot(dependent, explanatory)
```

```
pairs.panels(boston_selected,
gap = 0,
bg = c("black", "yellow", "red")[Boston$medv],
pch=21)
```

Análise de componentes principais

```
boston_selected_standar <- scale(boston_selected)
pcaBos <- prcomp(boston_selected_standar, center = TRUE, scale. = TRUE)

componentesB <- pcaBos$rotation[, 1:7]
estatisticasB <- data.frame(
  Sd = pcaBos$sdev[1:7],
  Prop_Var = pcaBos$sdev^2 / sum(pcaBos$sdev^2),
  Cum_Prop = cumsum(pcaBos$sdev^2 / sum(pcaBos$sdev^2))
)

componentesB <- round(componentesB, 2)
```

```r
estatisticasB$Sd <- round(estatisticasB$Sd, 2)
estatisticasB$Prop_Var <- round(estatisticasB$Prop_Var, 2)
estatisticasB$Cum_Prop <- round(estatisticasB$Cum_Prop, 2)

tabela_finalB <- rbind(
  componentesB,
  "Sd" = c(estatisticasB$Sd, rep(NA, 0)),
  "% Variância" = c(estatisticasB$Prop_Var * 100, rep(NA, 0)),
  "% Acumulada" = c(estatisticasB$Cum_Prop * 100, rep(NA, 0))
)

colnames(tabela_finalB) <- c("CP1", "CP2", "CP3", "CP4","CP5", "CP6", "CP7")

kable(
  tabela_finalB,
  format      = "latex",
  format.args = list(big.mark = ".", decimal = ","),
  booktabs    = TRUE,
  align       = paste0("l", rep("c", 7), collapse = "")
) %>%
  formartaKable() %>%
  add_header_above(
    header = c(" " = 1, "Componentes principais (CP)" = 7),
    bold   = TRUE
  ) %>%
  pack_rows("Variável", 1, 7) %>%
  pack_rows("", 8, 10) %>%
  row_spec(8:10, bold = FALSE) %>%
  footnote(
    general = c(
      "indus: Proporção de acres não comerciais por cidade;",
      "nox: Concentração de NOx (partículas por milhão);",
      "rm: Número médio de cômodos por residência;",
      "tax: Taxa de imposto predial por U$10.000;",
      "ptratio: Proporção aluno-professor;",
      "lstat: % da população com baixo status socioeconômico;",
      "medv: Valor mediano das casas (em milhares de dólares)."
    ),
    general_title = "Nota:"
  )

fviz_eig(pcaBos,ggtheme = theme_classic(),addlabels = TRUE,
```

```
         xlab = "Componentes principais",ylab = "Variância explicada (%)", main = "") #Scree
```

```
Boston$medv_cat <- cut(Boston$medv, breaks = 3)

ggbiplot(pcaBos, obs.scale = 1, var.scale = 1,
groups = Boston$medv_cat, ellipse = TRUE,
 circle = TRUE, varname.size = 5) +
scale_color_discrete(name = "") +
theme_bw()+
theme(legend.position = "none")
```

Análise fatorial

```
kmoBos <- KMO(boston_selected_standar)

kmo_Bos <- data.frame(
  Variável = names(kmoBos$MSAi),
  MSAi = unname(kmoBos$MSAi)
)

msa_globalBos <- data.frame(
  Variável = "MSA",
  MSAi = kmoBos$MSA
)

kmo_df_finalBos <- rbind(kmo_Bos, msa_globalBos)

n_total <- nrow(kmo_df_finalBos)

n_total <- nrow(kmo_df_finalBos)

kable(
  kmo_df_finalBos,
  format      = "latex",
  format.args = list(big.mark = ".", decimal = ","),
  booktabs    = TRUE,
  align       = "lc"
) %>%
  formartaKable() %>%
  add_header_above(
```

```
    header = c(" " = 1, "Medidas de Adequação Amostral (KMO)" = 1),
    bold   = TRUE
  ) %>%
  pack_rows("Medidas Individuais (MSAi)", 1, n_total - 1) %>%
  pack_rows("MSA Global", n_total, n_total) %>%
  footnote(
    general = c(
      "MSAi: Avalia se cada variável individual se relaciona adequadamente com as demais",
      "MSA: Avalia se todo o conjunto de dados é adequado para Análise Fatorial."
    ),
    general_title = ""
  )

bartlett_result_psychBo <- cortest.bartlett(cor(boston_selected_standar), n = nrow(boston_sel


resultados_bartlettBo <- data.frame(
  "Bartlett's K-squared"=bartlett_result_psychBo$chisq,
  "df"=bartlett_result_psychBo$df,
  "p-value" = c(bartlett_result_psychBo$p.value
))

resultados_bartlettBo%>%
  flextable()

cor_matrixBos <- cor(boston_selected_standar)
eigen_valuesBos <- eigen(cor_matrixBos)$values

eigen_dfBos <- data.frame(
  Fator = 1:length(eigen_valuesBos),
  Autovalor = round(eigen_valuesBos, 3),
  `% Variância` = round(eigen_valuesBos / sum(eigen_valuesBos) * 100, 1),
  `% Acumulada` = round(cumsum(eigen_valuesBos) / sum(eigen_valuesBos) * 100, 1)
)

colnames(eigen_dfBos) <- c("Fator", "Autovalor", "% Variância", "% Acumulada")
eigen_dfBos%>%
  flextable()
set.seed(123)
parallel <- fa.parallel(boston_selected_standar, nfactors=4,
                        fa = "fa", main="", ylabel="Autovalores dos fatores")
```

```r
set.seed(123)
fa_varimax_mle <- fa(
  r = boston_selected_standar,
  nfactors = 2,
  rotate = "varimax",
  fm = "mle",
  scores = "regression"
)

loadings_df <- data.frame(
  Variável = rownames(fa_varimax_mle$loadings),
  Fator1 = unname(round(fa_varimax_mle$loadings[,1], 3)),
  Fator2 = unname(round(fa_varimax_mle$loadings[,2], 3)),
  Comunalidade = unname(round(rowSums(fa_varimax_mle$loadings^2), 3))
)

loadings_df <- loadings_df[order(loadings_df$Comunalidade, decreasing = TRUE), ]

loadings_df %>%
  flextable()

biplot.psych(
  fa_varimax_mle,
  main = "Biplot - 2 Fatores (Varimax)",
  col = c("black", "red", "blue"),
  pch = c(21, 22, 23),
  group = as.numeric(Boston$medv)
)
```