LITERACIA FAMILIAR NO PERÍODO PRÉ E PANDÊMICO DE COVID-19: CENÁRIO BRASILEIRO E ESTRATÉGIAS DE INTERVENÇÃO

Cliente: Natália Viana - Mestranda em Neuropsicologia do Desenvolvimento

Alunos

Bianca Caravelli de Sá

Luiz Avelar

Thomas Santiago

Orientador

Adrian Luna

QUESTIONÁRIO

EXAMINANDO A INFLUÊNCIA DAS RESTRIÇÕES DA COVID-19 NO AMBIENTE DE LITERACIA FAMILIAR

1) Rastreio sobre contexto familiar;

2) Educação Musical;

3) Práticas e atividades de aprimoramento de aprendizagem;

4) Histórico familiar de condição de saúde;

5) Práticas e atividades de literacia;

6) Histórico de alfabetização.

BANCO DE DADOS

PRÉ-TRATADO

- *2223 registros - 1 registro por criança*
- *167 variávies*
- *Variáveis dummy - Perguntas do tipo checkbox*
- *Predominância de variáveis categóricas*

DIFICULDADES E DESAFIOS

- *Banco de dados denso e complexo*
  - *Banco pré-tratado diferente do code book*
- *Como resumir os dados*
  - *Análise de cluster*
  - *Pré e pós ou período completo*
- *Muitas observações NAs*
- *Tratamento das variáveis*
  - *Variávies checkbox*
  - *Muitas categorias por variável*
- *Interpretação dos resultados*
- *Fatores pouco explicativos - MCA*

PLANO DE ATIVIDADES

- *Análise descritiva: Variáveis que podem impactar os hábitos de literacia durante a pandemia*
- *Blocos de Interesse*
- *Divisão em pré e pós-pandemia*
- *PCA - Análise de componentes principais*
- MCA - Análise de correspondência múltipla

BLOCO 1 - READING HISTORY BR

BLOCO 2 - HOME LITERACY RESOURCES BR

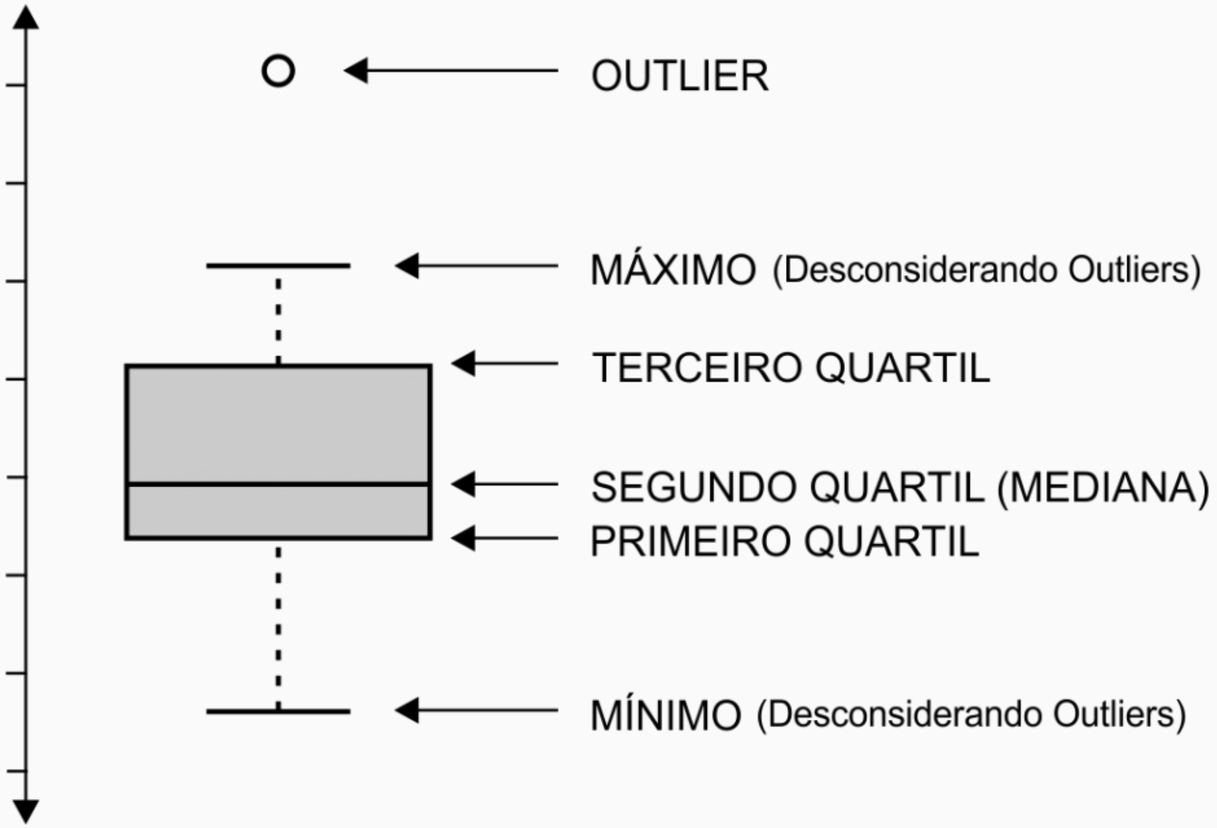
BLOCO 3 - ENRICHMENT ACTIVITIES BR

METODOLOGIA

TÉCNICAS EMPREGADAS NO DESENVOLVIMENTO DO TRABALHO

- Principal Components Analysis
- Multiple Correspondence Analysis
- Tabelas de Frequências
- Boxplots
- Correlação de Pearson

BOXPLOT



https://operdata.com.br/blog/como-interpretar-um-boxplot/

- ***Posição*** – *Observa-se a linha central do retângulo (a mediana ou segundo quartil).*
- ***Dispersão*** – *A dispersão dos dados pode ser representada pelo intervalo interquartílico que é a diferença entre o terceiro quartil e o primeiro quartil (tamanho da caixa)*
- ***Simetria*** – *Um conjunto de dados que tem uma distribuição simétrica, terá a linha da mediana no centro do retângulo.*
- ***Caudas*** – *As linhas que vão do retângulo até aos outliers podem fornecer o comprimento das caudas da distribuição.*
- ***Outliers*** – *Já os outliers indicam possíveis valores discrepantes. No boxplot, as observações são consideradas outliers quando estão abaixo ou acima do limite de detecção de outliers.*

CORRELAÇÃO DE PEARSON

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \cdot \operatorname{var}(Y)}}$$

- ***0.9*** *para mais ou para menos indica uma correlação muito forte.*
- ***0.7 a 0.9*** *positivo ou negativo indica uma correlação forte.*
- ***0.5 a 0.7*** *positivo ou negativo indica uma correlação moderada.*
- ***0.3 a 0.5*** *positivo ou negativo indica uma correlação fraca.*
- ***0 a 0.3*** *positivo ou negativo indica uma correlação desprezível.*

PRINCIPAL COMPONENTS ANALYSIS PT.1

- *É definida como uma transformação linear ortogonal;*
- *A maior variância por alguma projeção escalar dos dados venha a residir na primeira coordenada (chamada de primeiro componente principal), a segunda maior variância na segunda coordenada e assim por diante;*
- *Os componentes no PCA são obtidos através da diagonalização da matriz de correlação;*
- *Extrai-se os autovetores e autovalores associados.*
- *Autovalor é interpretado como a inércia da nuvem NI projetada (variância explicada" para o componente de classificação s.*

PRINCIPAL COMPONENTS ANALYSIS PT.2

*"**Step-by-step**"*

1. *Padronize o intervalo de variáveis iniciais contínuas;* $$z = \frac{value - mean}{standard\ deviation}$$

2. Calcular a matriz de covariância para identificar correlações $$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

3.Calcule os autovetores e autovalores da matriz de covariância para identificar os componentes principais;

4. Crie um vetor de recursos para decidir quais componentes principais manter;

5. Reformule os dados ao longo dos eixos dos componentes principais

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

Fonte: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

MULTIPLE CORRESPONDENCE ANALYSIS

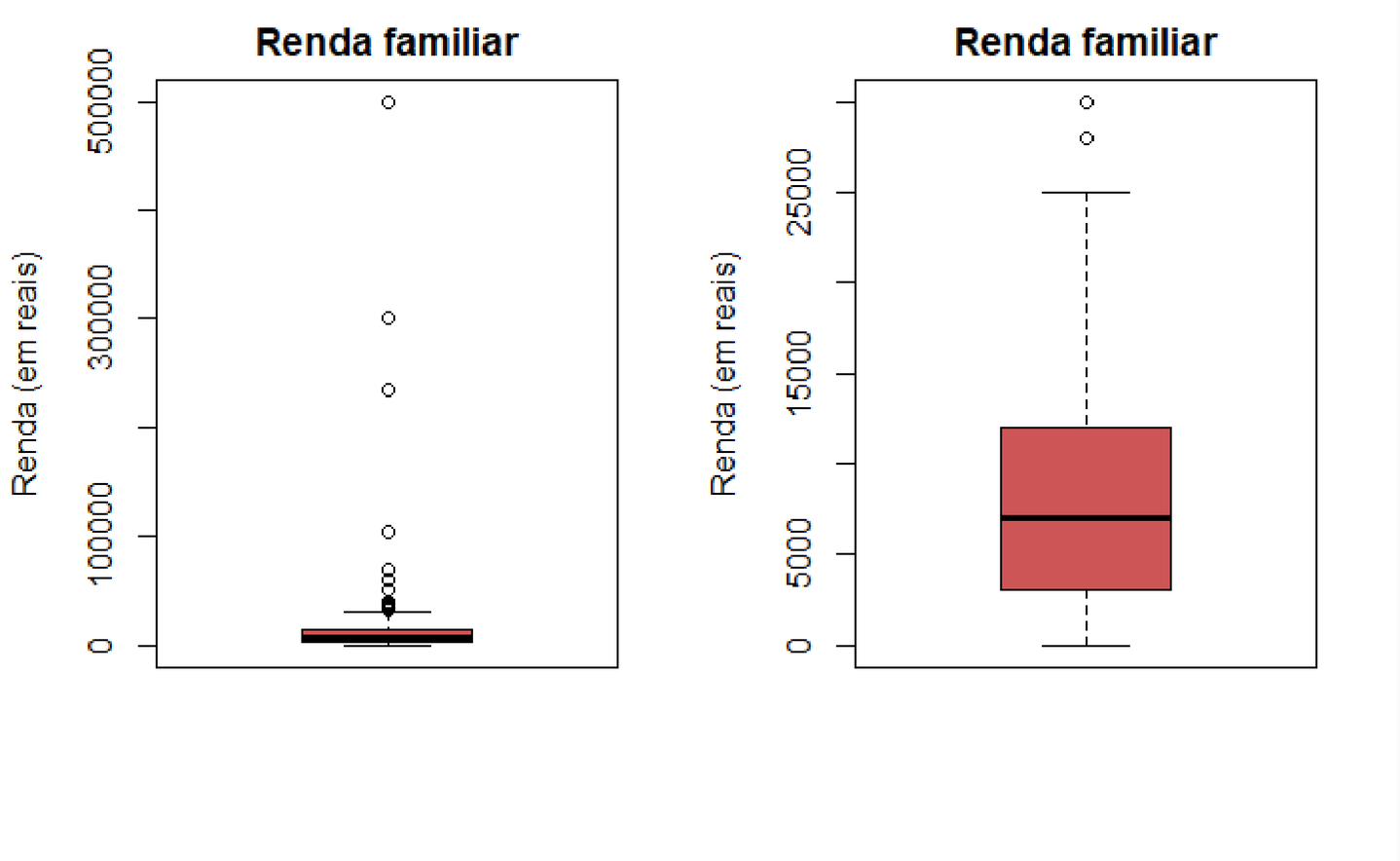
- *Técnica de análise de dados para dados categóricos nominais;*
- *Aplica-se o algoritmo de Corresponde Analysis a uma matriz de indicadores;*
- *Uma matriz de indicadores é uma matriz de indivíduos × variáveis, onde as linhas representam os indivíduos e as colunas são variáveis fictícias que representam categorias das variáveis;*
- *Permite a representação direta de indivíduos como pontos no espaço geométrico.*

ANÁLISE DESCRITIVA

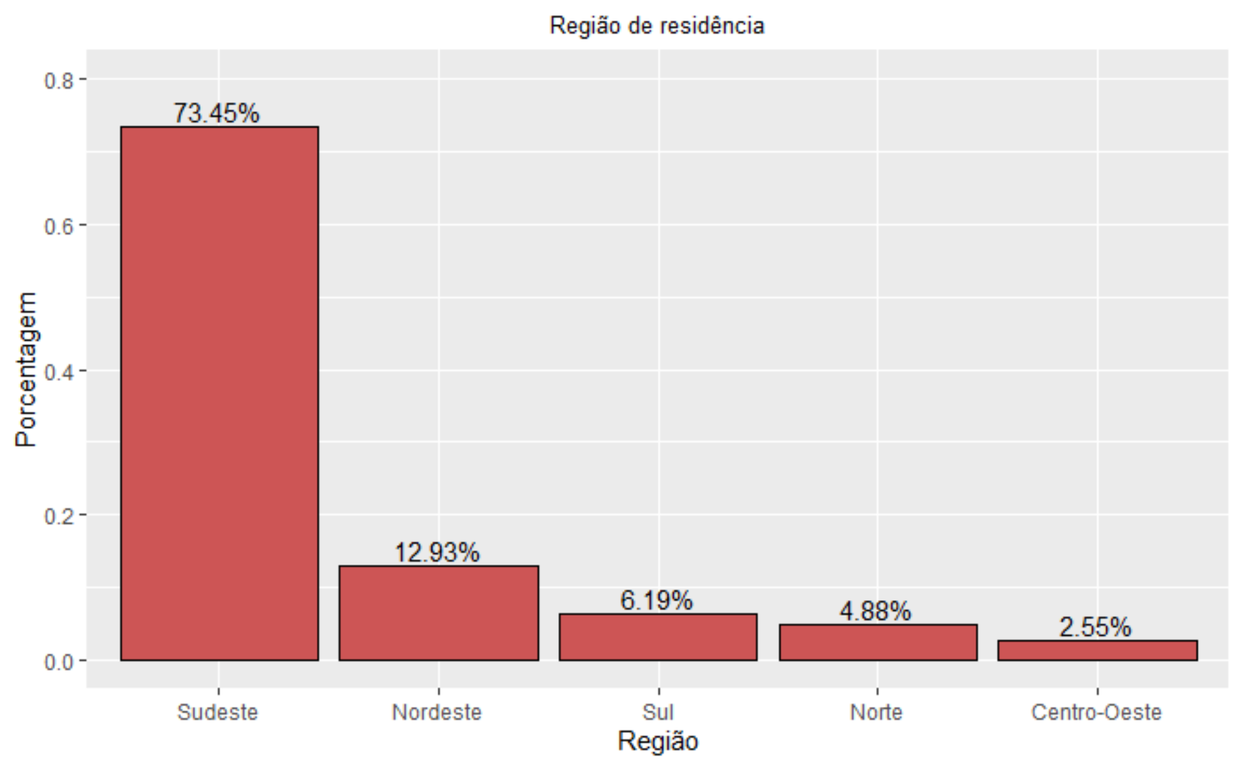
FOCO EM VARIÁVEIS SOCIOENOCÔMICAS

- *Gráficos de barra*
- *Boxplot*
- *Histograma*
- *Tabelas de frequência*

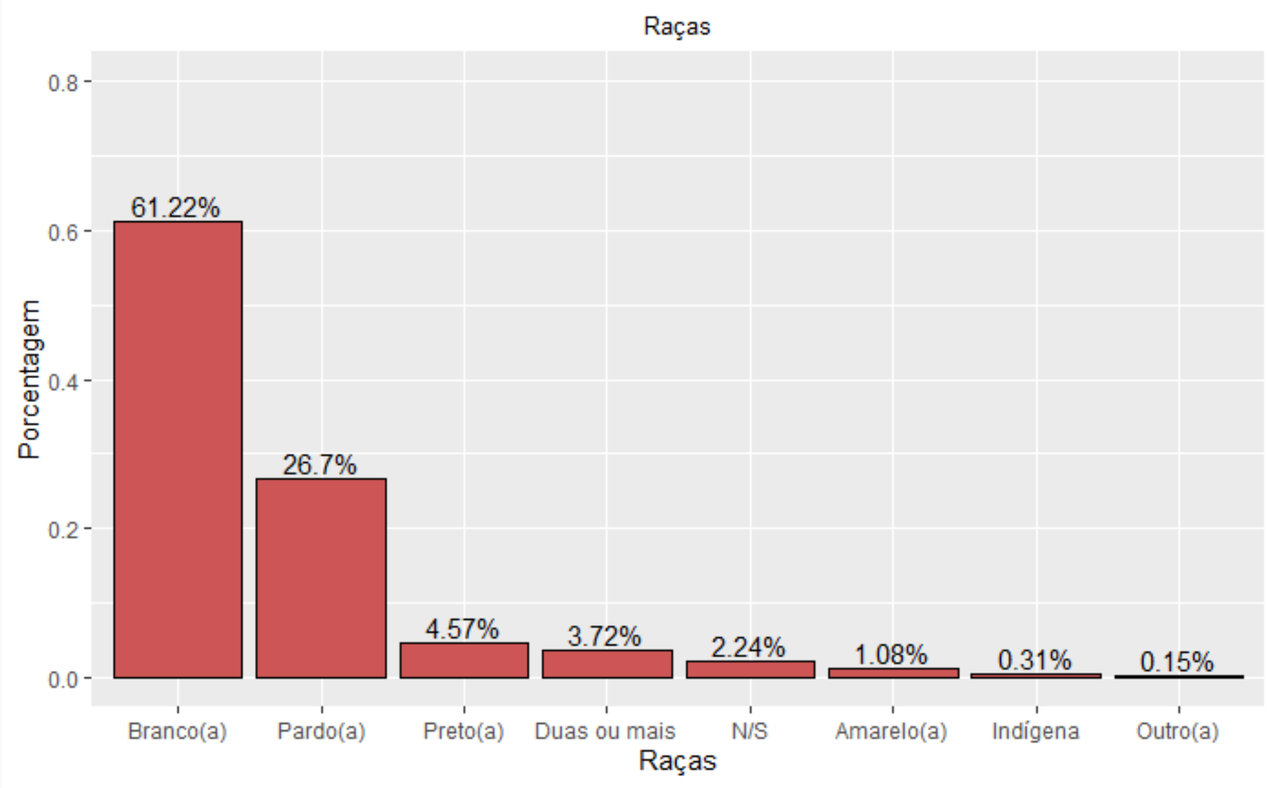
RENDA FAMILIAR



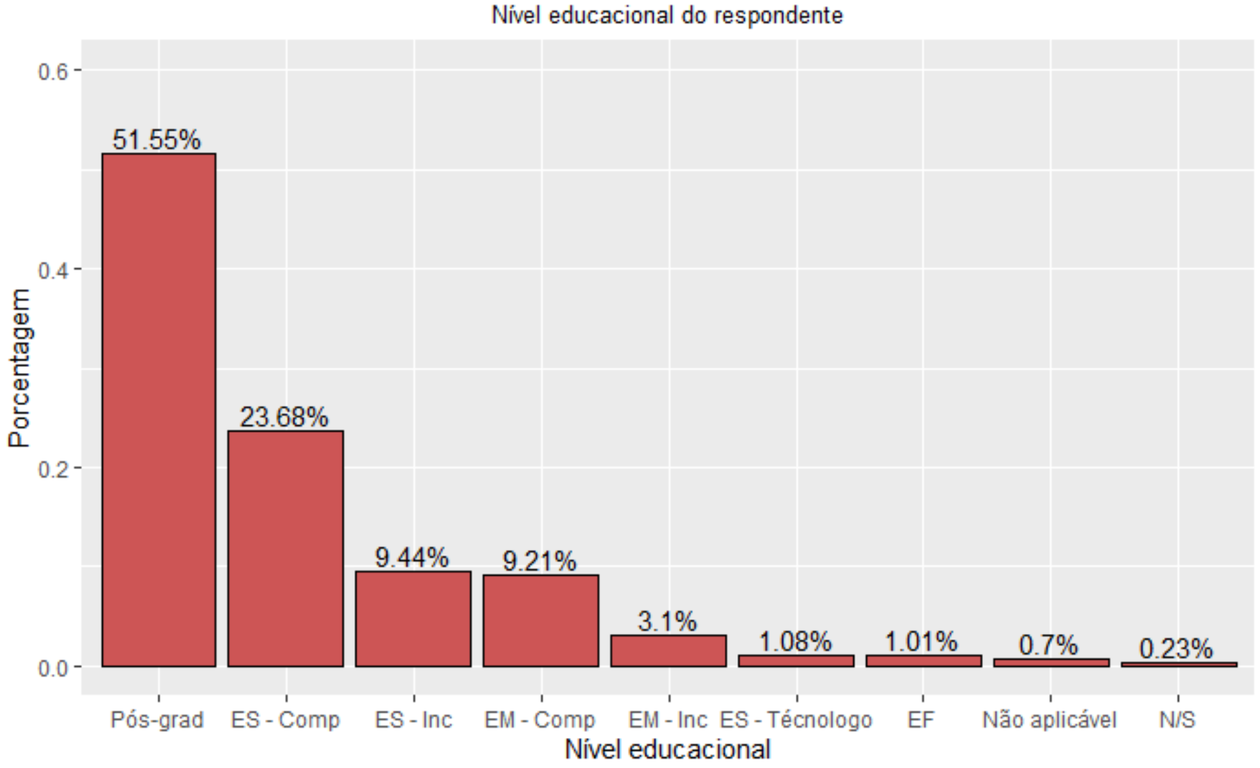
REGIÃO DE RESIDÊNCIA



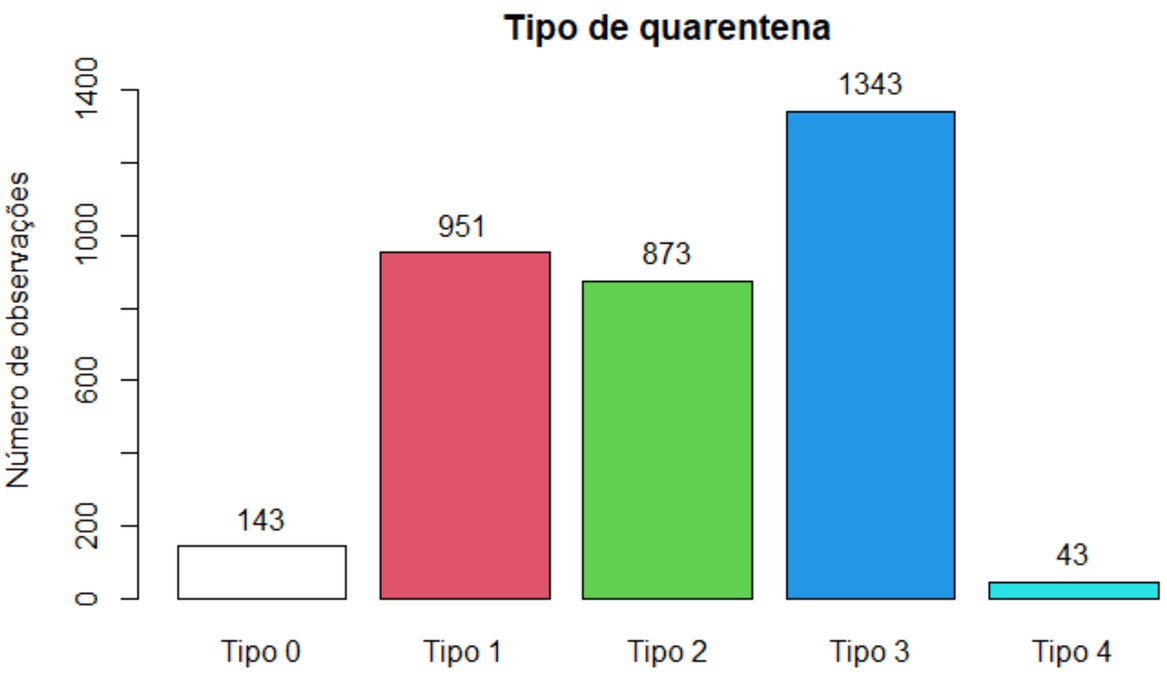
COR/RAÇA



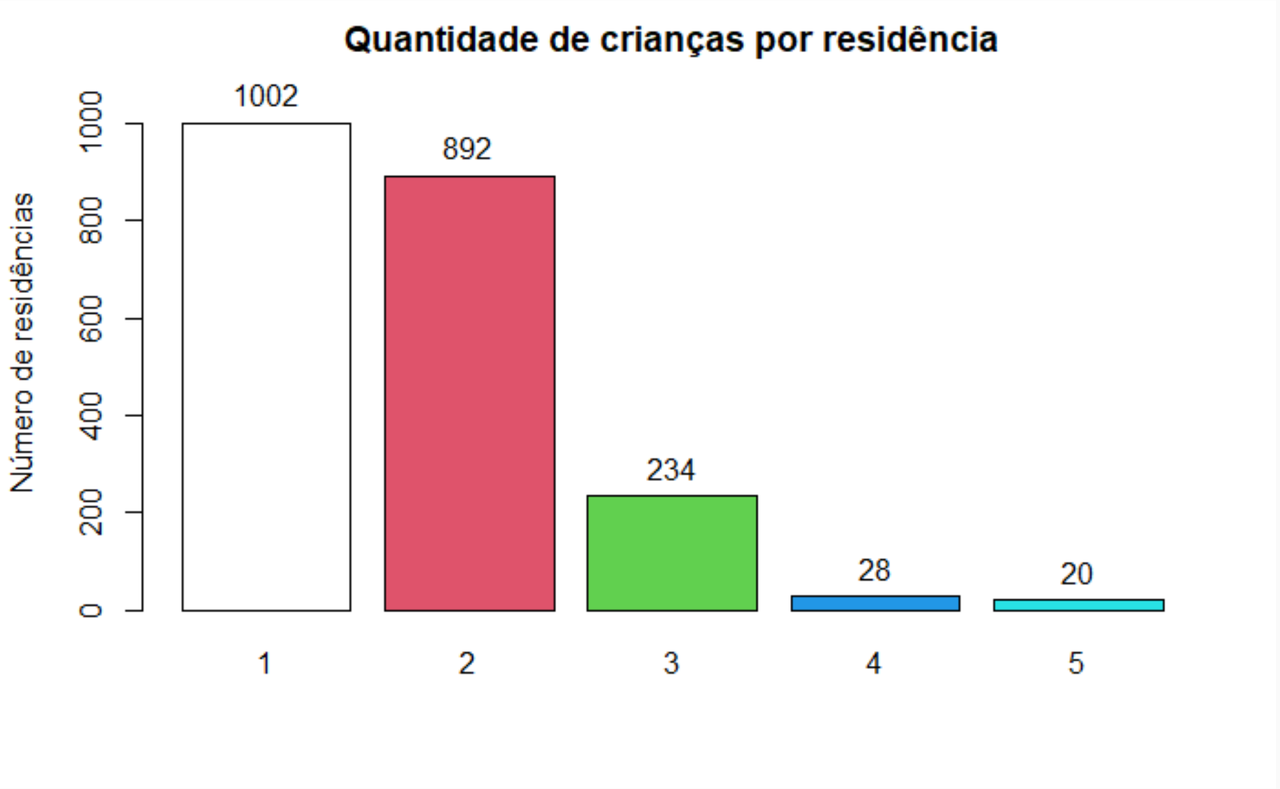
NÍVEL EDUCACIONAL DO RESPONDENTE



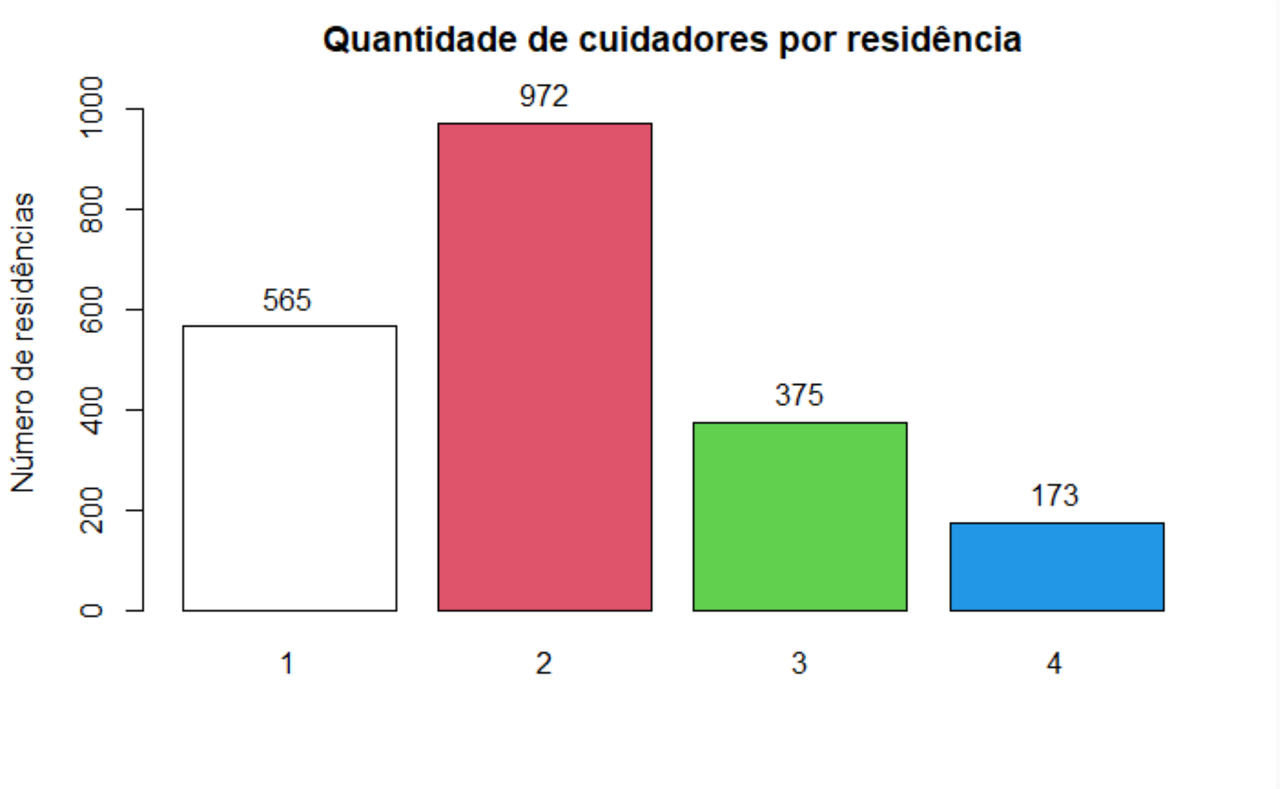
TIPO DE QUARENTENA



NÚMERO DE CRIANÇAS EM CASA



NÚMERO DE CUIDADORES EM CASA



DIFICULDADES LDA

Um problema particular dos modelos de tópicos Bayesianos é que para uma estrutura complexa do modelo, a inferência para a distribuição de tópicos latentes (pesos de mistura de tópicos) é muitas vezes intratável.
Vários métodos de aproximação têm sido propostos, como a abordagem variacional e o método de amostragem, embora a inferência ainda seja lenta.

Autores propõem trabalhar com métodos de linguagens naturais, utilizando *um modelo Bayesiano para supervisionar o treinamento de um modelo neural.*

ETAPAS LDA

1) Calcule os vetores médios d-dimensionais para as diferentes classes do conjunto de dados.

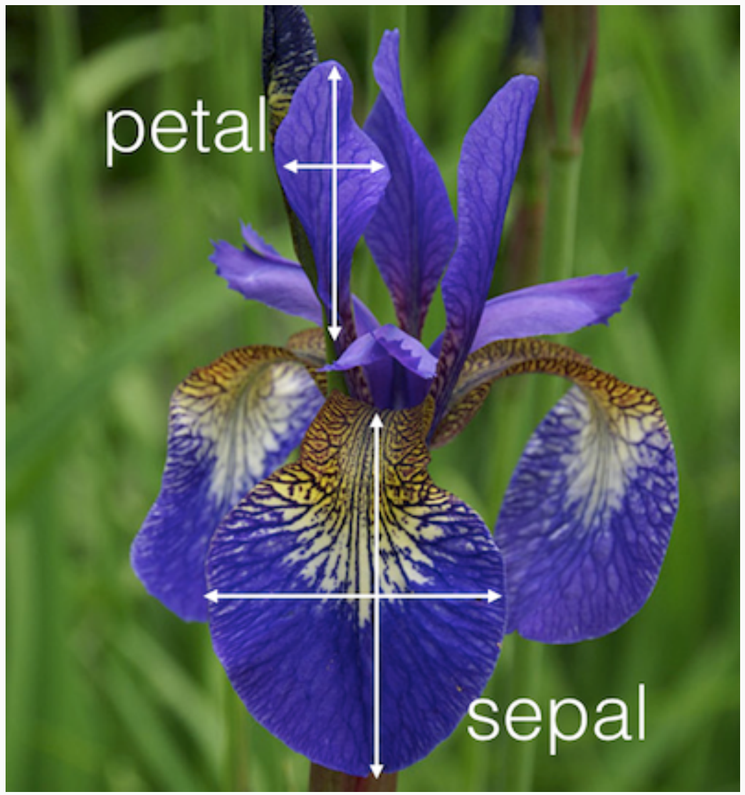
2) Calcule as matrizes de dispersão (matriz de dispersão entre classe e dentro da classe).

3) Calcule os autovetores
e os autovalores correspondentes $(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d)$
para as matrizes de dispersão $(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_d)$

4) Ordene os autovetores diminuindo os autovalores e escolha k autovetores com os maiores autovalores para formar uma d x k matriz dimensional W (onde cada coluna representa um autovetor).

5) Use esta matriz de autovetores d×k para transformar as amostras no novo subespaço.

EXEMPLO



3 CLASSES

1) Iris-setosa (n=50)

2) Iris-versicolor (n=50)

3) Iris-virginica (n=50)

4 FEATURES

1. *sepal length in cm*
2. *sepal width in cm*
3. *petal length in cm*
4. *petal width in cm*

I. CALCULANDO OS VETORES MÉDIOS D-DIMENSIONAIS

vetores médios, (i=1,2,3) das 3 classes de flores diferentes:

$$\boldsymbol{m}_i = \begin{bmatrix} \boldsymbol{\mu}_{\omega_i}(\text{sepal length}) \\ \boldsymbol{\mu}_{\omega_i}(\text{sepal width}) \\ \boldsymbol{\mu}_{\omega_i}(\text{petal length}) \\ \boldsymbol{\mu}_{\omega_i}(\text{petal width}) \end{bmatrix}$$

II. CALCULANDO DUAS MATRIZES DE DIMENSÃO 4X4: A MATRIZ DE DISPERSÃO DENTRO DA CLASSE E A MATRIZ DE DISPERSÃO ENTRE AS CLASSES.

MATRIZ DE DISPERSÃO DENTRO DA CLASSE

$$S_W = \sum_{i=1}^{c} S_i$$

$$S_i = \sum_{\boldsymbol{x} \in D_i}^{n} (\boldsymbol{x} - \boldsymbol{m}_i)\,(\boldsymbol{x} - \boldsymbol{m}_i)^T$$ *(matriz de dispersão para cada classe)*

$$\boldsymbol{m}_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in D_i}^{n} \boldsymbol{x}_k$$

MATRIZ DE DISPERSÃO ENTRE AS CLASSES

$$S_B = \sum_{i=1}^{c} N_i (\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^T$$

III. O PROBLEMA DE AUTOVALOR GENERALIZADO PARA A MATRIZ

A seguir, devemos resolver o problema de autovalor generalizado para a matriz $S_W^{-1} S_B$ *para obter os discriminantes lineares.*

Informações sobre a distorção de uma transformação linear:

Autovetores: *direção da distorção, formarão os novos eixos do novo subespaço de características*

Autovalore: *fator de escala para os autovetores que descrevem a magnitude da distorção, nos dirão quão "informativos" são os novos "eixos"*

IV. SELECIONANDO DISCRIMINANTES LINEARES PARA O NOVO SUBESPAÇO DE RECURSO

SELECIONAR OS AUTOVETORES

Os autovetores com os menores autovalores carregam menos informações sobre a distribuição dos dados, e esses são os que queremos eliminar.

A abordagem comum é classificar os autovetores do maior para o menor autovalor correspondente e escolher os principais k autovetores.

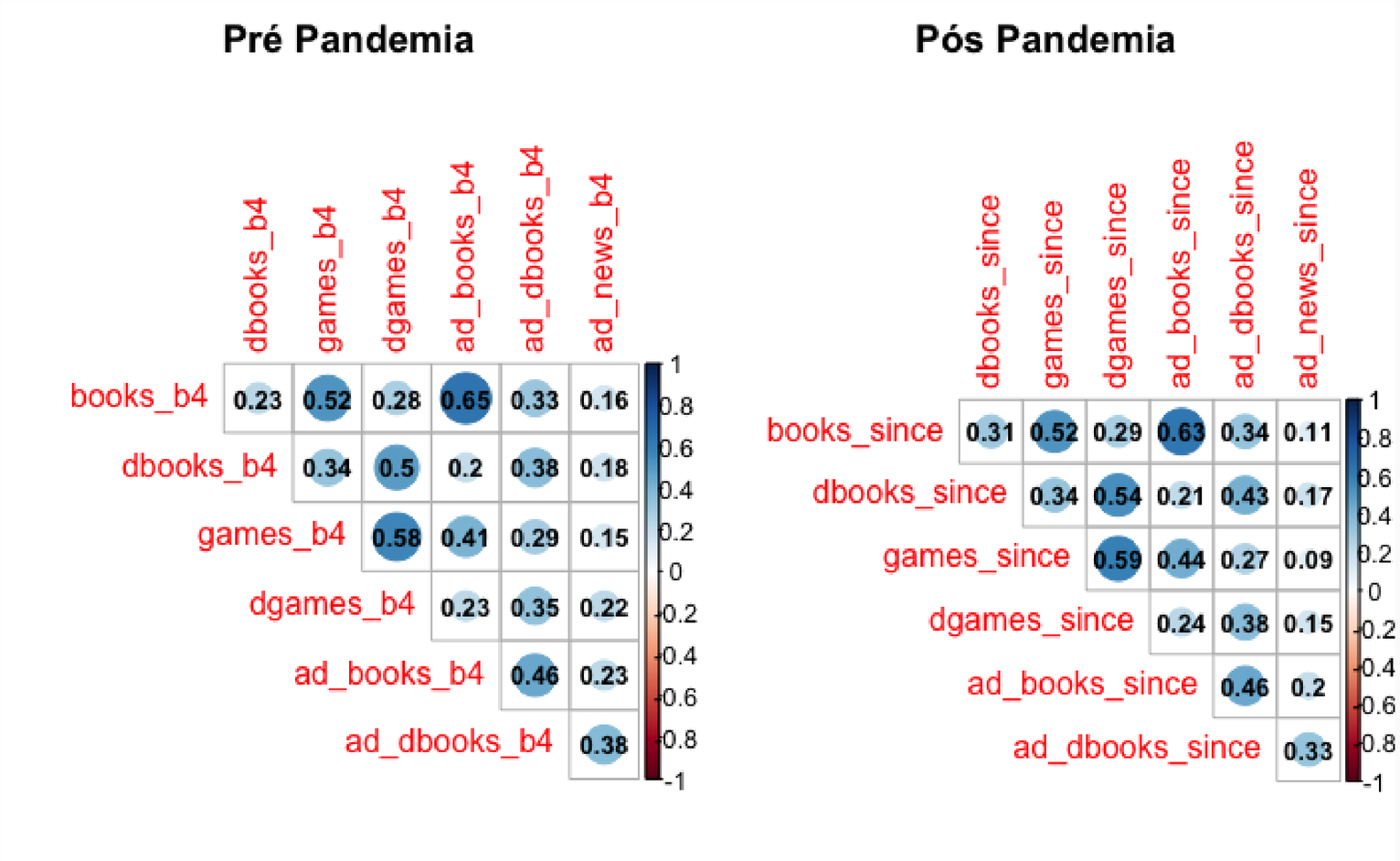
V. TRANSFORMANDO AS AMOSTRAS NO NOVO SUBESPAÇO

$$\boldsymbol{Y} = \boldsymbol{X} \times \boldsymbol{W}.$$

Latent Dirichlet allocation

CORRELAÇÕES PRÉ X PÓS PANDEMIA

BLOCO 2- HOME LITERACY RESOURCES BR

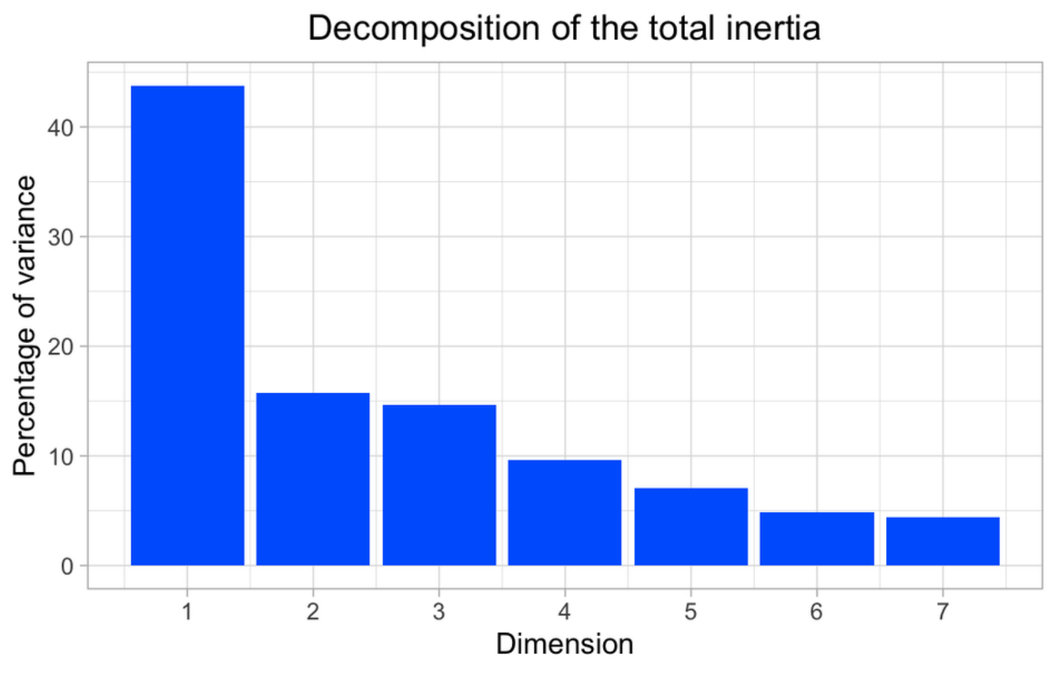


PCA BLOCO 2

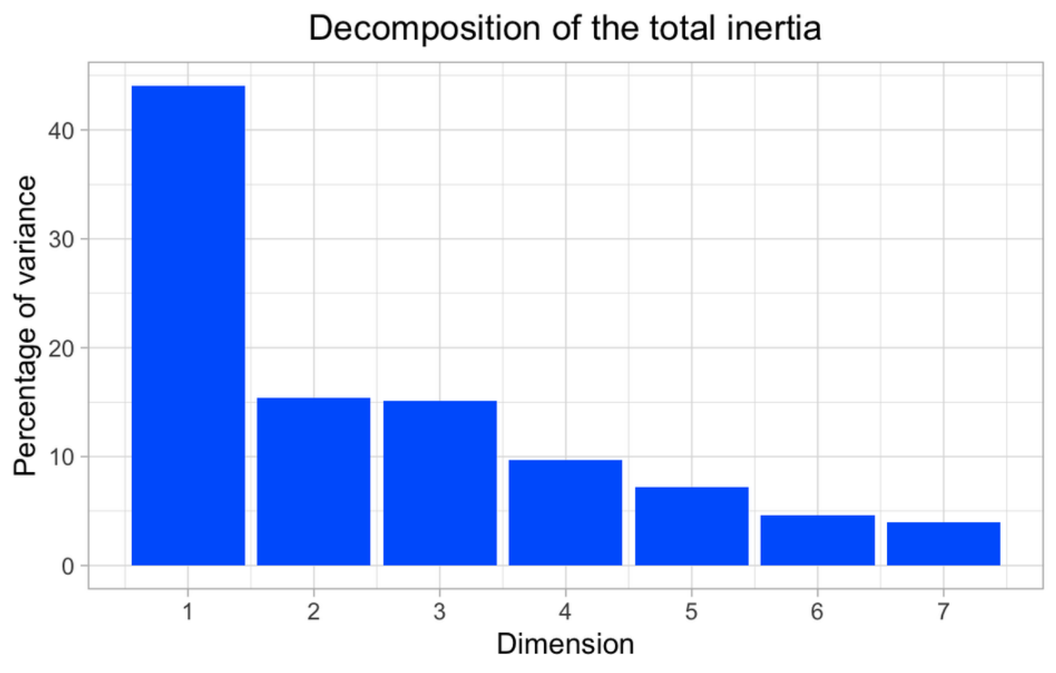
As **duas primeiras dimensões** para ambos os blocos **expressam cerca 60% da variabilidade** total da nuvem dos indivíduos (ou variáveis) é explicada pelo plano.

Pela % da variabilidade explicada pela **dimensão 1** sugere que apenas este eixo **já está carregando a informação real.**

PRÉ - PANDEMIA



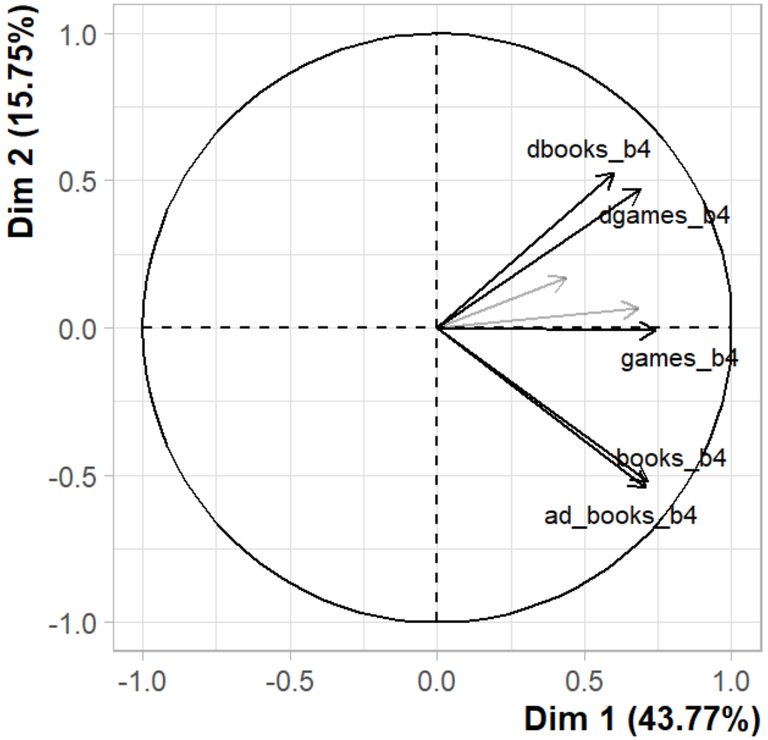
PÓS - PANDEMIA



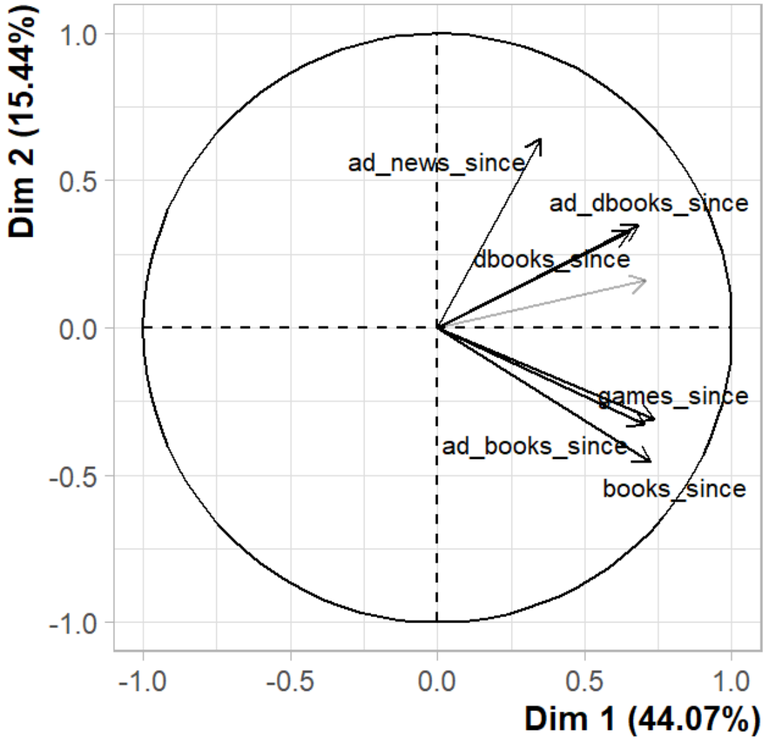
PCA - FACTOR MAP BLOCO 2

Tanto para o bloco **pré** quanto para o bloco **pós-pandemia** a primeira dimensão é caracterizada por indivíduos que opõem coordenadas positivas elevadas (**eixo superior direito**) com coordenadas negativas elevadas (**eixo inferior direito**)

PRÉ - PANDEMIA



PÓS - PANDEMIA



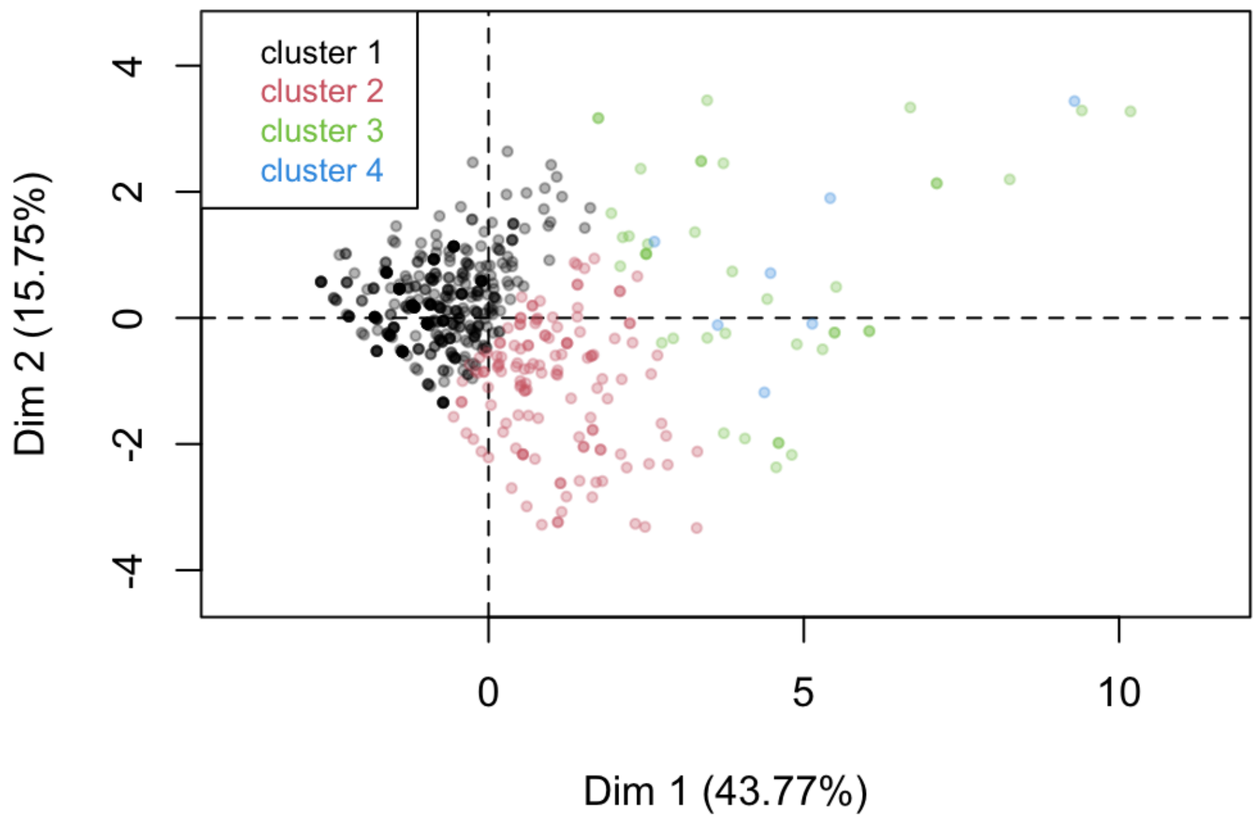
O **cluster 1** é composto por indivíduos que compartilham: valores **baixos** para as variáveis **ad_books_b4**, **books_b4**, **ad_dbooks_b4**, **games_b4**, **dgames_b4**, **dbooks_b4** e **ad_news_b4.**

O **cluster 2** é composto por indivíduos que compartilham: valores **altos** para as variáveis **ad_books_b4**, **books_b4**, **ad_dbooks_b4** e **games_b4**.

O **cluster 3** é composto por indivíduos que compartilham: valores **altos** para as variáveis **dbooks_b4**, **games_b4**, **dgames_b4**, **ad_dbooks_b4**, **books_b4**, **ad_books_b4** e **ad_news_b4**.

O **cluster 4** é composto por indivíduos que compartilham: valores **altos** para as variáveis **ad_news_b4**, **ad_dbooks_b4**, **ad_books_b4**, **dgames_b4**, **books_b4** e **games_b4.**

PRÉ-PANDEMIA



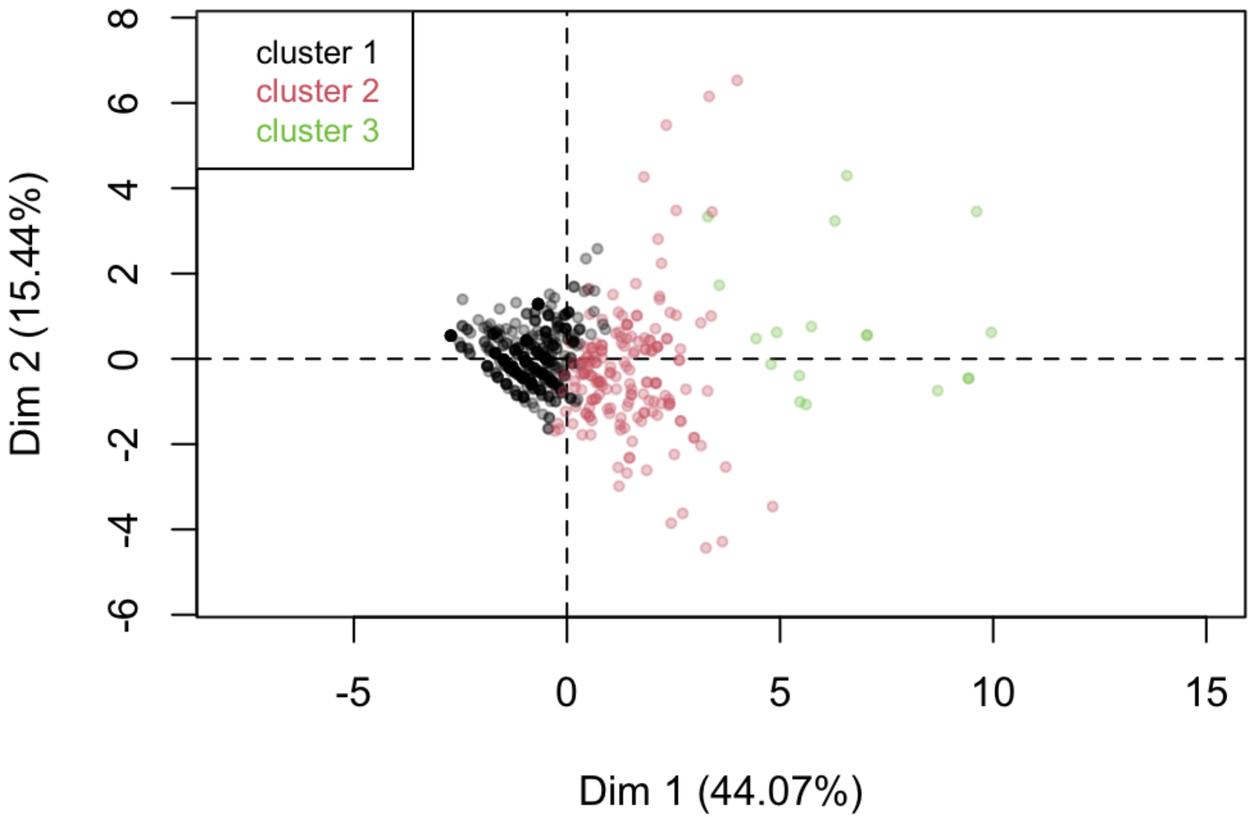
Classificação Hierárquica Ascendente dos indivíduos. A classificação feita em indivíduos revela 4 clusters.

O **cluster 1** é composto por indivíduos que compartilham: valores **baixos** para as variáveis **ad_books_since**, **books_since**, **ad_dbooks_since**, **games_since**, **dgames_since**, **dbooks_since** e **ad_news_since.**

O **cluster 2** é composto por indivíduos que compartilham: valores **altos** para as variáveis **ad_books_since**, **books_since**, **ad_dbooks_since**, **games_since**, **ad_news_since**, **dgames_since** e **dbooks_since.**

O **cluster 3** é composto por indivíduos que compartilham: valores **altos** para as variáveis **dbooks_since**, **dgames_since**, **ad_dbooks_since**, **games_since**, **books_since**, **ad_books_since** e **ad_news_since**.

PÓS-PANDEMIA



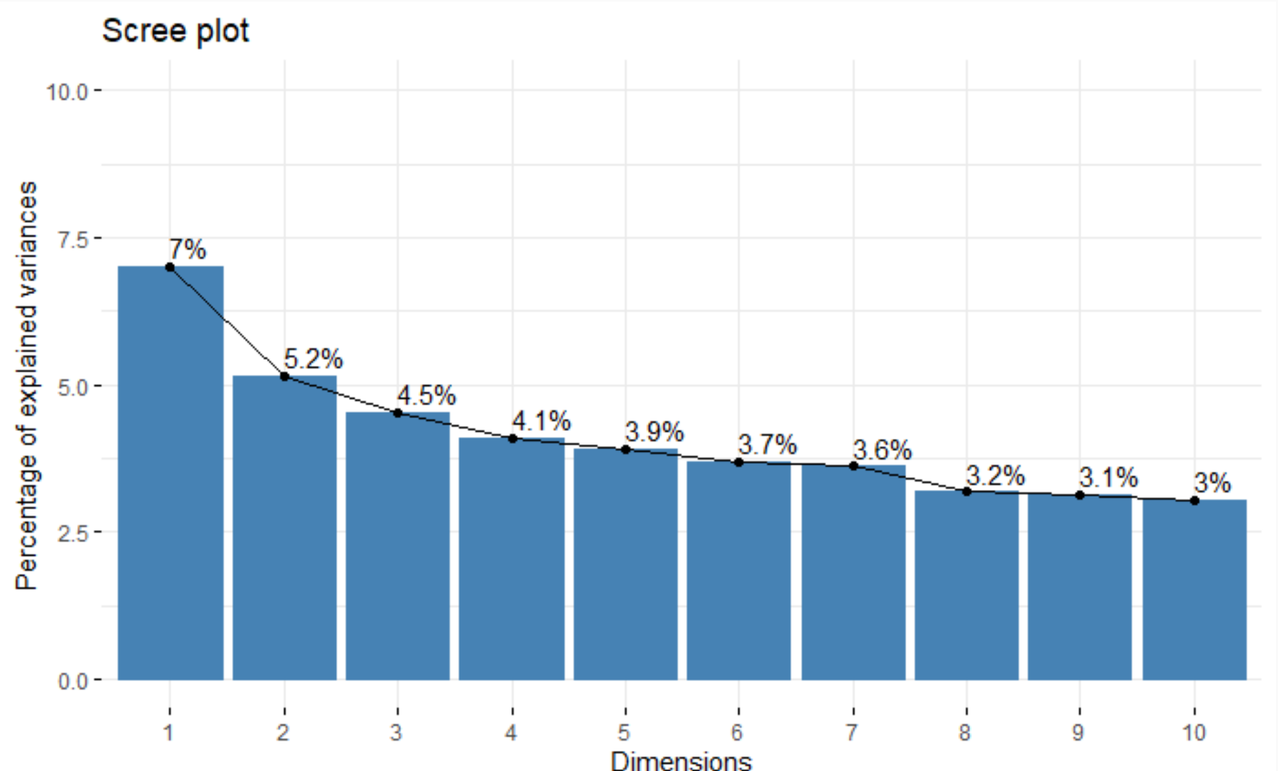
Classificação Hierárquica Ascendente dos indivíduos. A classificação feita em indivíduos revela 3 clusters.

MCA BLOCO 2

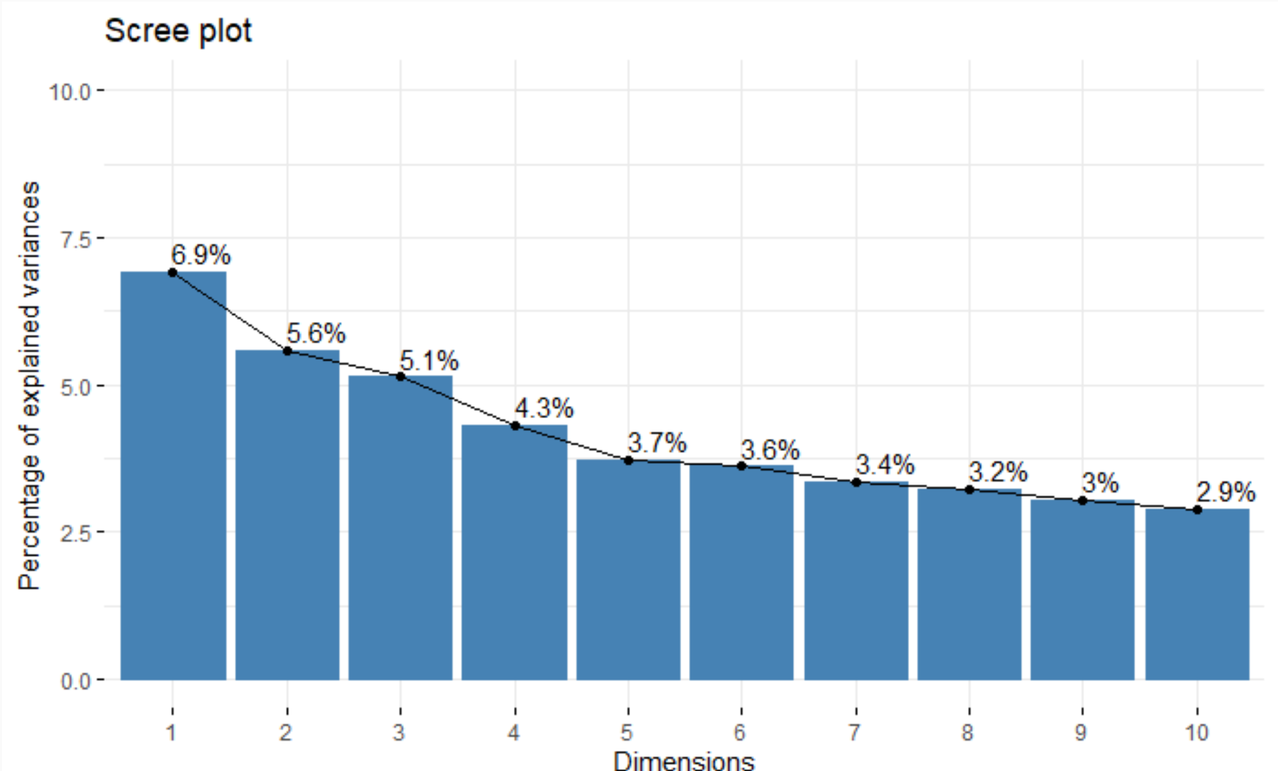
Entre ambos os blocos o maior valor percentual que uma dimensão expressa a variabilidade dos dados foi **7%**.

São necessárias muitas dimensões para explicar bem os dados.

PRÉ - PANDEMIA



PÓS - PANDEMIA

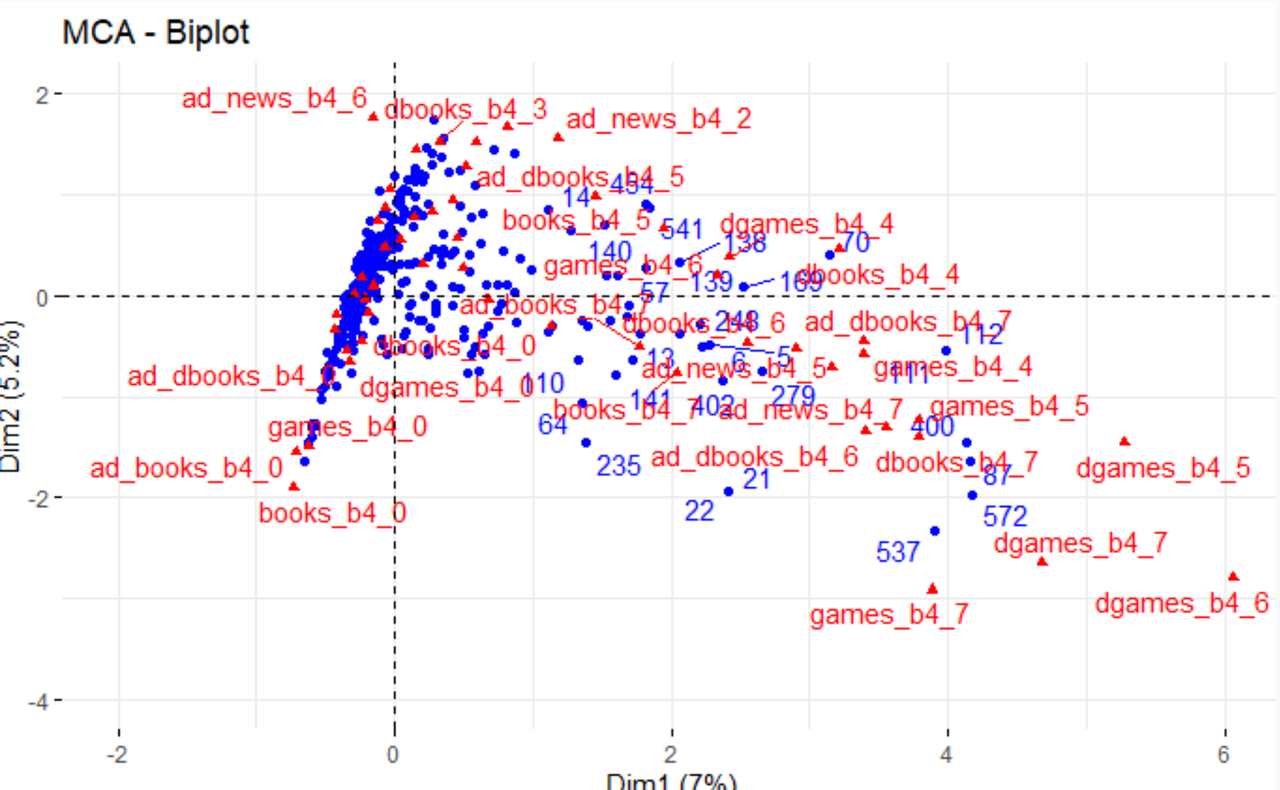


MCA
BLOCO 2

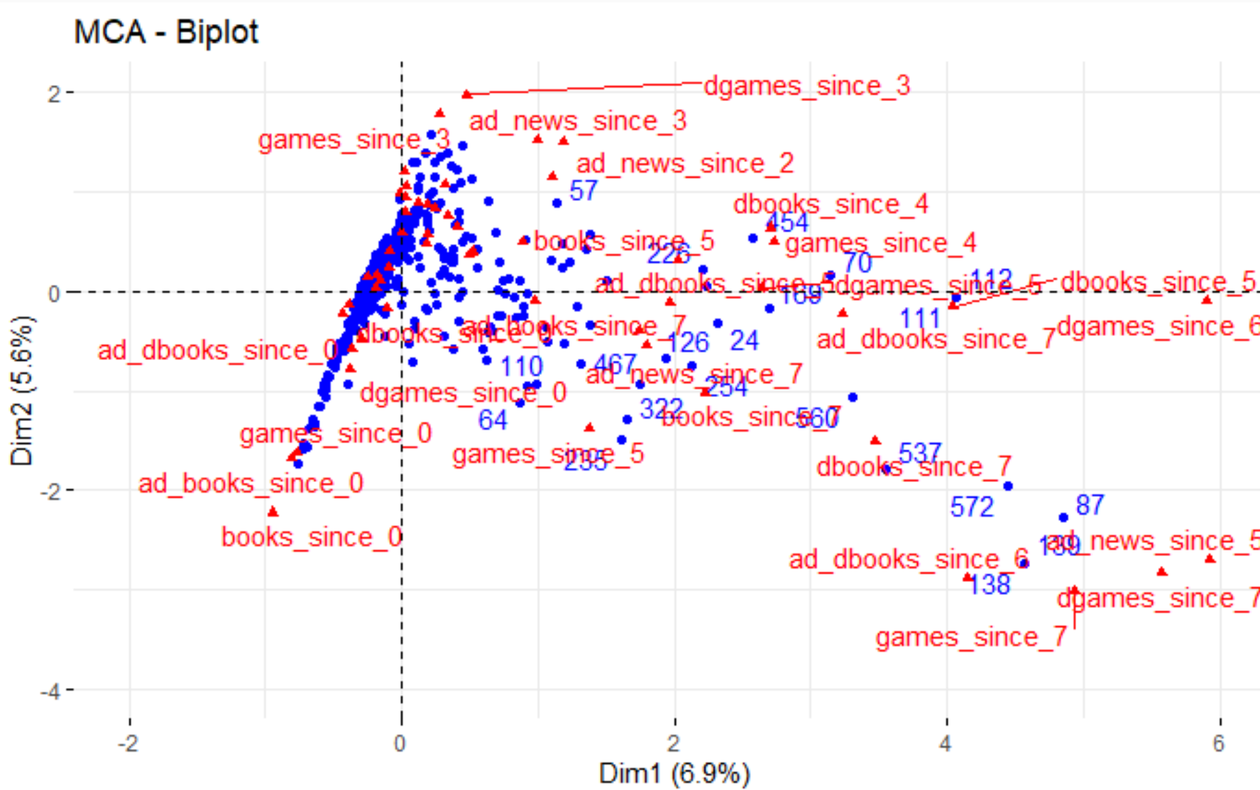
O biplot permite visualizar padrões dentro dos dados. Cada **indivíduo** (linha) é representado por um **ponto azul** e cada **variável (coluna)** é representada por um **ponto vermelho**.

A distancia entre cada observação representa a similaridade (ou dissimilaridade) entre elas, assim quanto mais próximos, mais similares são.

PRÉ - PANDEMIA



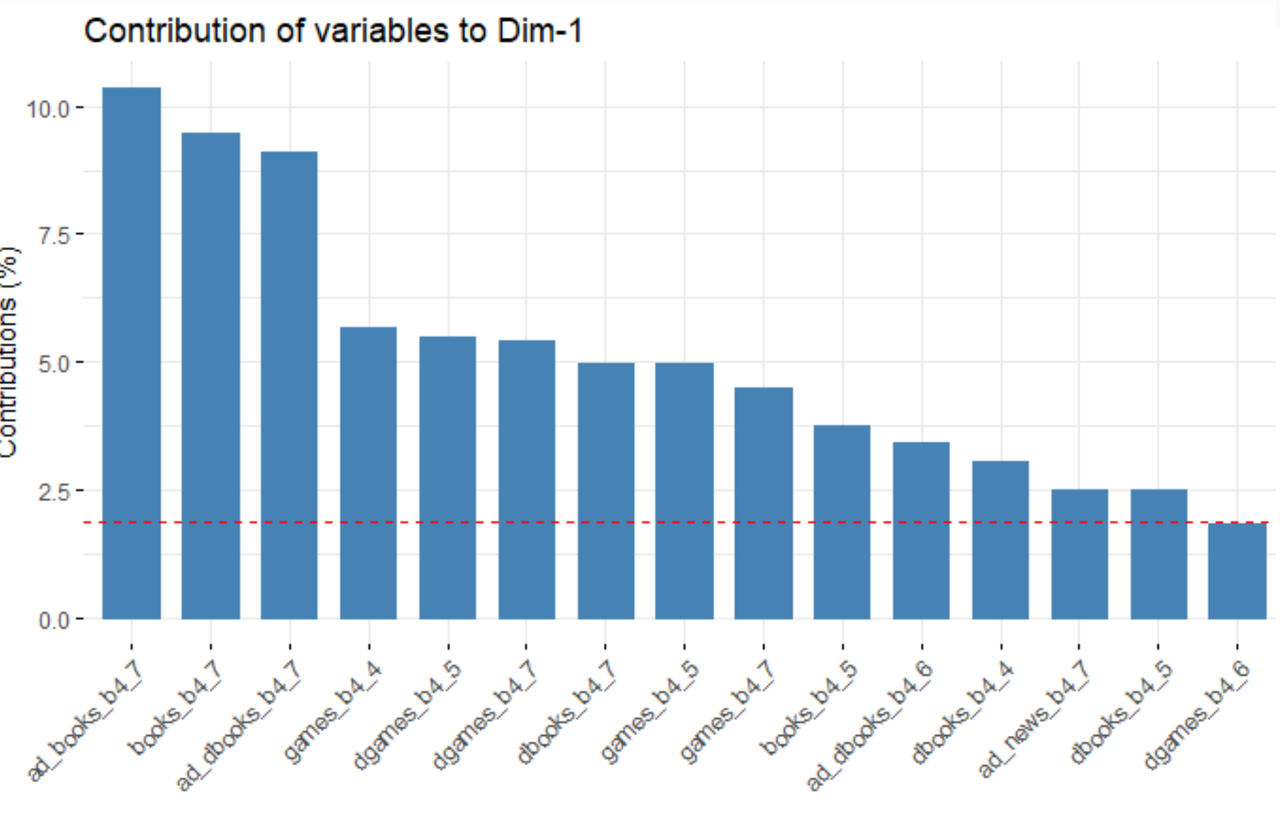
PÓS - PANDEMIA



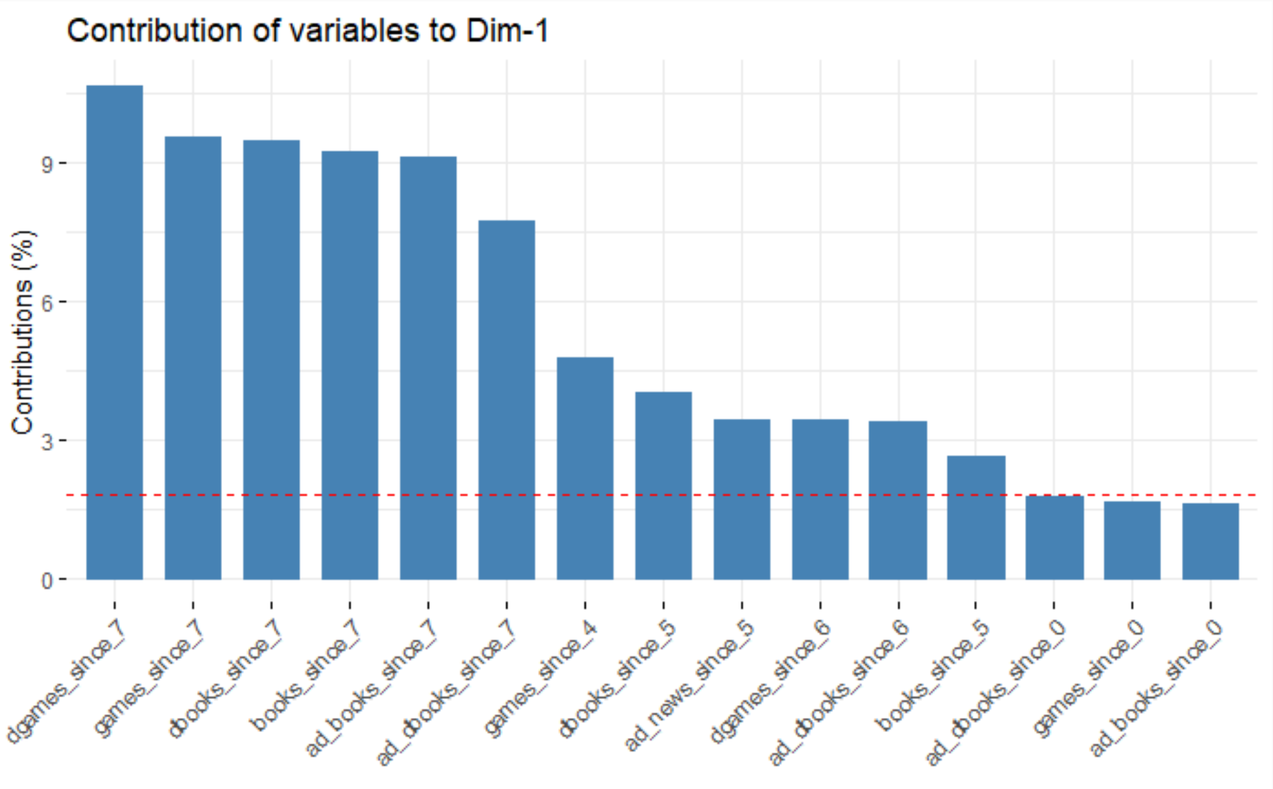
MCA
BLOCO 2

Os gráficos abaixo representam o quanto cada **categoria de cada variável** contribuiu para a construção da **dimensão 1**

PRÉ - PANDEMIA



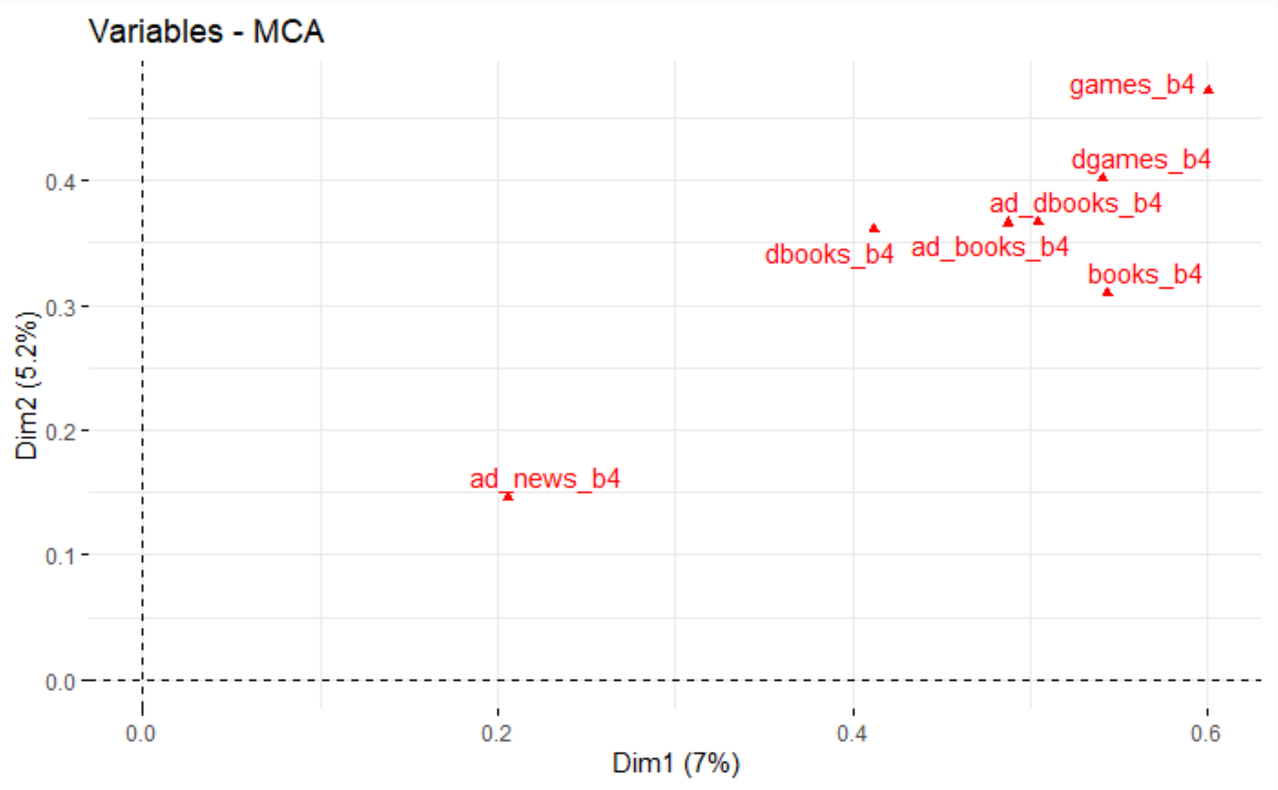
PÓS - PANDEMIA



MCA BLOCO 2

A correlação entre as variávies e as **duas principais dimensões** é, no geral, de **moderada a desprezíve**l, variando entre 0,6 a 0,2. Nenhuma variável possui correlação forte com as dimensões mais explicativas.

PRÉ - PANDEMIA



PÓS - PANDEMIA