**Projeto LabEst: Consultoria em Análise Estatística para Alunos de Pós-Graduação**

**Relatório Preliminar da Análise Estatística:**

Literacia Familiar no período pré e pandêmico de Covid-19: cenário brasileiro e estratégias de intervenção

**Cliente**:

Natália Viana

Mestranda em Psicologia: Cognição e Comportamento - Faculdade de Filosofia e Ciências Humanas da Universidade Federal de Minas Gerais

**Consultores**

Bianca Caravelli de Sá

Luiz Felipe Avelar

Thomás de Santiago Silva

Alunos da Disciplina “Laboratório de Estatística I”

**Orientação: Prof. Adrian Luna**

**15/07/2022**

## 1) Introdução e metodologia

O objetivo dessa tese de doutorado é investigar o impacto da pandemia de COVID-19 nas práticas e recursos de literacia familiar no contexto brasileiro, ou seja, investigar as práticas de leitura e letramento feitas por pais e responsáveis com crianças pequenas feitas em casa durante o período de quarentena e o impacto para o desenvolvimento da alfabetização dessas crianças em um momento em que as escolas encontravam-se fechadas.

Para a obtenção dos dados, foi aplicado o questionário “Examinando a influência das restrições da COVID-19 no ambiente de literacia familiar” que consiste em perguntas de interesse que descrevem o perfil demográfico e socioeconômico dessas famílias, assim como os hábitos e atividades de literacia desenvolvidas no ambiente familiar.

A partir desses dados obtidos no questionário, as análises do dados encontram-se divididas em duas partes, a primeira consiste em uma análise descritiva, através de gráficos de barra, histogramas, boxplots e tabelas de frequência, das váriaveis de interesse principal que podem impactar os hábitos de literacia durante a pandemia. Cada variável foi analisada de maneira individual, sendo elas:

* Região de residência
* Raça
* Nível educacional do respondente
* Nível Socioeconômico da família - renda familiar
* Situação profissional do respondente antes e depois da covid
* Quantidade de cuidadores
* Quantidade de crianças
* Tipo de quarentena
* Nível de estresse com o isolamento

      Já a segunda parte consiste em análises fatorial exploratórias a partir de blocos de perguntas a fim de reduzir a quantidade de variávies e obter indíces que explicam a variabilidade dos dados. As variáveis de interesse para essa análise foram divididas nos seguintes blocos:

* Bloco 1 - Bloco Reading History Br (Histórico de alfabetização)
* Bloco 2 - Home Literacy Resources Br (Recursos)
* Bloco 3 - Enrichment Activities Br (Atividades de enriquecimento)

## 2) Análise Exploratória

### Região de residência

Pode-se observar que a grande maioria, 73,45%, das pessoas que respoderam ao questionário moram na região Sudeste do Brasil, enquanto que apenas 2,55% são residentes da região Centro-Oeste. Esta variável possui 931 observações NAs.

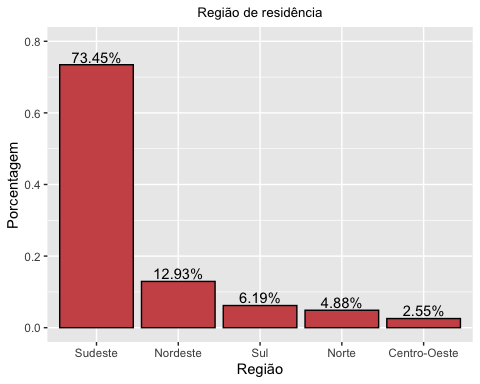


Figura 1 - Região de residência

### Raça

Pode-se observar que 87,92% das crianças são brancas ou pardas, sendo a raça branca a mais frequente. Para esta variável houve 931 obervações NAs.

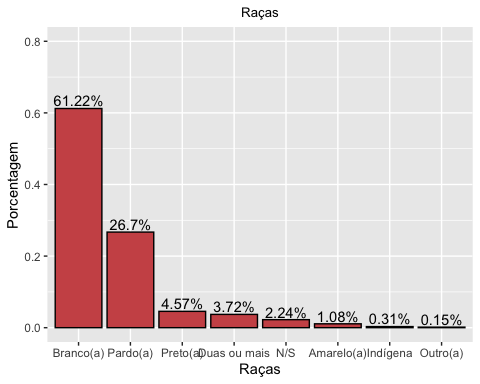


Figura 2 - Raça

### Nível educacional do respondente

Pode-se observar que a maioria dos respondentes do questionário possuem um nível educacional alto, sendo que 51,55% possuem pós-graduação, enquanto que apenas 1,01% dos respondentes possuiem Ensino Fundamental completo ou incompleto.

Vale ressaltar que 931 pessoas não responderam a essa pergunta.

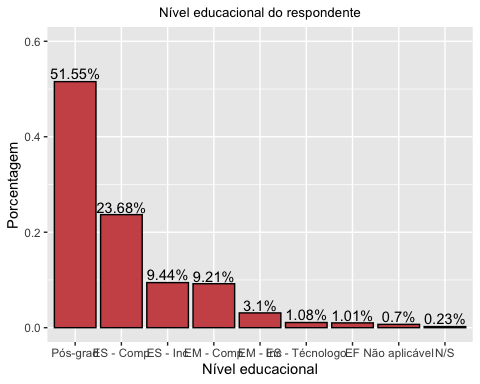


Figura 3 - Nível educacional do respondente

### Nível Socioeconômico da família - Renda familiar

Em relação a renda familiar, pode-se observar que possui uma distribuição com consideráveis valores outliers, os gráficos abaixo possuem duas versões, sendo a primeira os valores originais e a segunda removendo os outliers. A renda mediana familiar é de R$ 7000,00 e 978 pessoas não responderam a essa questão.

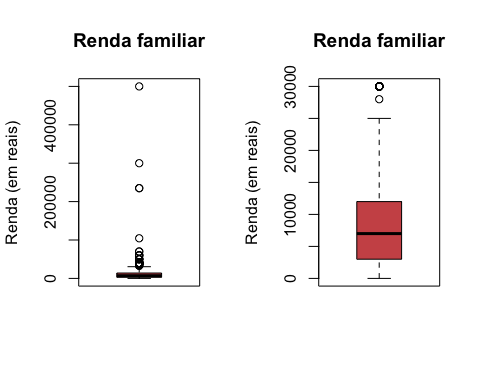


Figura 4 - Renda familiar

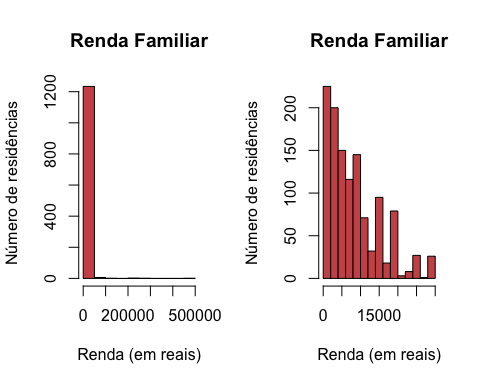


Figura 5 - Renda familiar

### Situação profissional do respondente antes e depois da covid

Para as duas questões sobre situação profissional, há 12 opções de resposta, sendo elas:

0 - Eu trabalhava em tempo integral fora de casa (mais do que 40 horas por semana)

1 - Eu trabalhava em tempo integral em casa (mais do que 40 horas por semana)

2 - Eu trabalhava em tempo integral, parcialmente em casa e parcialmente fora de casa (mais do que 40 horas por semana)

3 - Eu trabalhava meio período fora de casa (menos que 40 horas por semana)

4 - Eu trabalhava meio período em casa (menos que 40 horas por semana)

5 - Estudante

6 - Eu estava em licença de trabalho temporária (por exemplo suspenso do trabalho, licença maternidade)

7 - Era dono(a) de casa ou cuidava das crianças em tempo integral

8 - Desempregado

9 - Aposentado

10 - Beneficiário do INSS

11 - Prefiro não responder

Pode-se observar que a grande maioria dos respondentes possuia um emprego integral fora de casa anteriormente ao COVID-19, assim que a pandemia se inicia há uma mudança na distribuição dos tipos de emprego e o trabalho em casa passa a ser a categoria mais frequente.

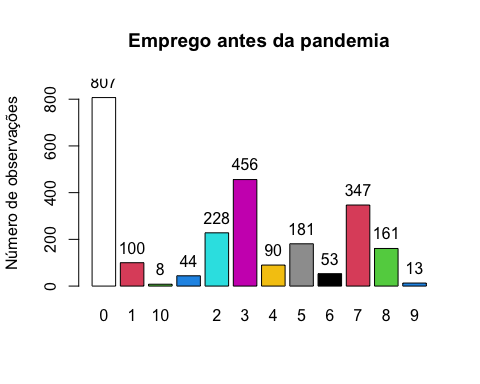


Figura 6 - Emprego pré pandemia

## job\_before :   
## Frequency Percent Cum. percent  
## 0 807 32.44 32.44  
## 1 100 4.02 36.45  
## 10 8 0.32 36.78  
## 11 44 1.77 38.55  
## 2 228 9.16 47.71  
## 3 456 18.33 66.04  
## 4 90 3.62 69.65  
## 5 181 7.27 76.93  
## 6 53 2.13 79.06  
## 7 347 13.95 93.01  
## 8 161 6.47 99.48  
## 9 13 0.52 100.00  
## Total 2488 100.00 100.00

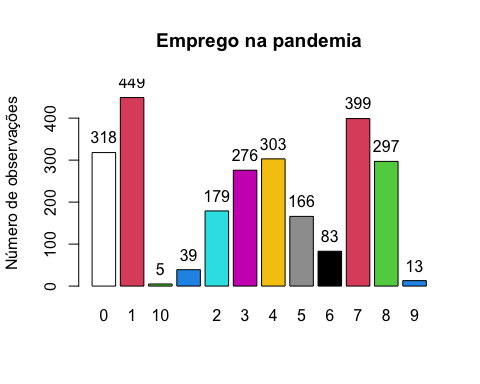


Figura 7 - Emprego na pandemia

## job\_after :   
## Frequency Percent Cum. percent  
## 0 318 12.58 12.58  
## 1 449 17.77 30.35  
## 10 5 0.20 30.55  
## 11 39 1.54 32.09  
## 2 179 7.08 39.18  
## 3 276 10.92 50.10  
## 4 303 11.99 62.09  
## 5 166 6.57 68.66  
## 6 83 3.28 71.94  
## 7 399 15.79 87.73  
## 8 297 11.75 99.49  
## 9 13 0.51 100.00  
## Total 2527 100.00 100.00

### Quantidade de cuidadores

Pode-se observar que das 2223 respostas ao questionário, 2085 respostas foram de residências que possuem cuidadores.

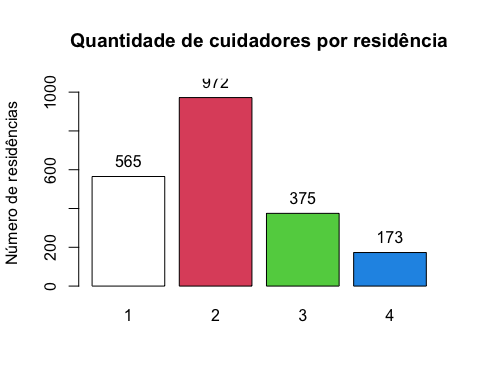


Figura 8 - Quantidade de cuidadores

## Quantidade\_de\_cuidadores :   
## Frequency Percent Cum. percent  
## 1 565 27.10 27.10  
## 2 972 46.62 73.72  
## 3 375 17.99 91.70  
## 4 173 8.30 100.00  
## Total 2085 100.00 100.00

### Quantidade de crianças

Pode-se observar que a maioria das residências possuem duas ou menos crianças, sendo que apenas 48 delas possuem quatro ou mais crianças. Para esta variável há 47 observações NAs.

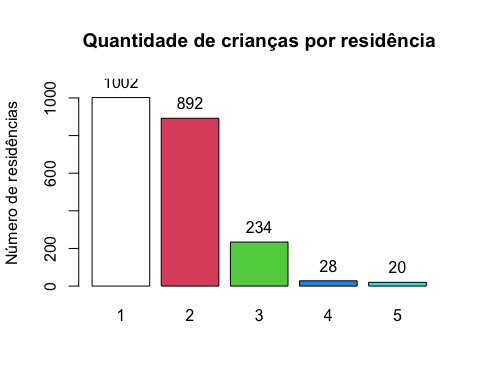


Figura 9 - Quantidade de crianças

## Quantidade\_de\_criancas :   
## Frequency Percent Cum. percent  
## 1 1002 46.05 46.05  
## 2 892 40.99 87.04  
## 3 234 10.75 97.79  
## 4 28 1.29 99.08  
## 5 20 0.92 100.00  
## Total 2176 100.00 100.00

### Tipo de quarentena

Para o tipo de quarentena e restrições recomendadas à família, temos as seguintes opções:

0 - Não há restrições agora

1 - Quarentena voluntária

2 - Pedido ou aviso depermanência em casa pelo governo local (caminhadas ou passeios com distanciament osocial permitidos)

3 - Pedido de permanência em casa pelo governo local (só saia de casa para fins essenciais)

4 - Outro

Pode-se observar que o tipo 3 foi o mais frequente, enquanto que o tipo 0 foi o menos frequente.

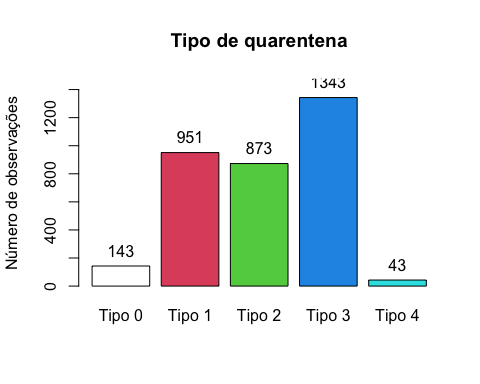


Figura 10 - Restrições familiares na pandemia

## restrictions :   
## Frequency Percent Cum. percent  
## Tipo 0 143 4.26 4.26  
## Tipo 1 951 28.36 32.63  
## Tipo 2 873 26.04 58.66  
## Tipo 3 1343 40.05 98.72  
## Tipo 4 43 1.28 100.00  
## Total 3353 100.00 100.00

### Nível de estresse com o isolamento

Para o nível de estresse com o isolamento, tem-se as seguintes categorias:

0 - Não tem - Não tem sido estressante

1 - Pouco - Um pouco estressante

2 - Estressante - Estressante

3 - Muito - Muito estressante

4 - Extremamente - Extremamente estressante

Pode-se notar que apenas 175 das respostas consideram que o isolamento não tem sido estressante, enquanto 712 o consideram pouco estressante.

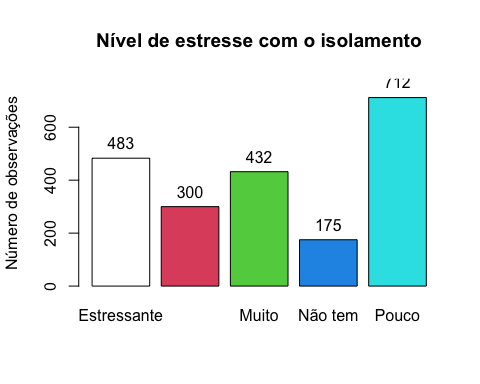


Figura 11 - Nível de estresse com o isolamento

## stress :   
## Frequency Percent Cum. percent  
## Estressante 483 22.98 22.98  
## Extremamente 300 14.27 37.25  
## Muito 432 20.55 57.80  
## Não tem 175 8.33 66.13  
## Pouco 712 33.87 100.00  
## Total 2102 100.00 100.00

## 3) Análise Fatorial - PCA e MCA

Ao trabalhar com um grande número de variáveis, a matriz de correlação agrupa uma grande quantidade de coeficientes (190 coeficientes para K = 20 variáveis). É, portanto, essencial ter uma ferramenta capaz de resumir as principais relações entre as variáveis de forma visual.

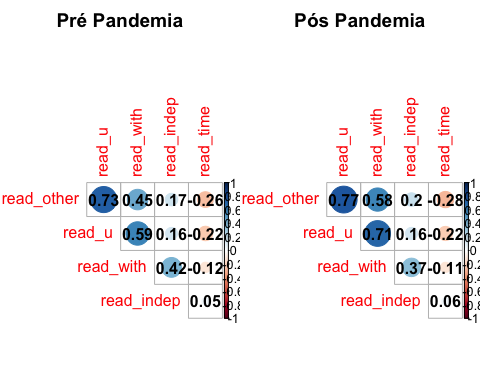
O objetivo do **PCA** é tirar conclusões das relações lineares entre variáveis detectando as dimensões principais de variabilidade. Como será analisado, estas conclusões serão complementadas pela definição das variáveis sintéticas oferecidas pelo PCA. Portanto, será mais fácil descrever os dados usando algumas variáveis sintéticas em vez de todas as variáveis originais.

Assim como na análise de componentes principais (PCA), o objetivo é resumir as relações entre as variáveis. Essas relações são estudadas em pares ou todas juntas. Neste último caso, procuramos variáveis sintéticas que resumem as informações contidas em um número de variáveis. A informação transportada por uma variável pode ser estudada em termosde suas categorias. Na **MCA**, focamos principalmente no estudo das categorias, como categorias representam tanto variáveis quanto um grupo de indivíduos (todos os indivíduos que selecionam esta categoria).

### Bloco 1 - Histórico de alfabetização

|  |  |
| --- | --- |
| Pré Covid | Com Covid |
| rh\_read\_b4\_yng\_br | rh\_read\_since\_yng\_br |
| rh\_read\_b4\_old\_br | rh\_read\_since\_old\_br |
| rh\_read\_with\_b4\_old\_br | rh\_read\_w\_since\_old\_br |
| rh\_readin\_b4\_old\_br | rh\_readin\_since\_old\_br |
| rh\_read\_time\_b4\_br | rh\_read\_time\_since\_br |

Correlação entre as variáveis do bloco:



### Multiple Correspondence Analysis

O MCA geralmente é usado para analisar um conjunto de dados de uma pesquisa. O objetivo é identificar:

1 - Um grupo de indivíduos com perfil semelhante em suas respostas às perguntas

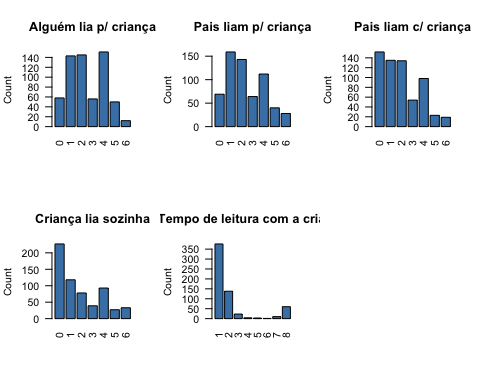
2 - As associações entre categorias de variáveis

Calcular e visualizar a análise de correspondência múltipla no software R usando *FactoMineR* (para a análise) e *factoextra* (para visualização de dados).

Revelar as variáveis mais importantes que mais contribuem para explicar as variações no conjunto de dados explicando como prever os resultados para indivíduos e variáveis suplementares. Filtrar os resultados do MCA para manter apenas as variáveis que mais contribuíram.

#### **Pré Pandemia**

## read\_other read\_u read\_with read\_indep read\_time   
## 0: 58 0: 69 0:152 0:227 1 :376   
## 1:143 1:159 1:135 1:118 2 :138   
## 2:145 2:143 2:134 2: 78 8 : 60   
## 3: 56 3: 64 3: 54 3: 39 3 : 23   
## 4:151 4:112 4: 98 4: 93 7 : 10   
## 5: 50 5: 40 5: 23 5: 27 4 : 4   
## 6: 12 6: 28 6: 19 6: 33 (Other): 4



Construindo o MCA

As variáveis podem ser representadas calculando as razões de correlação entre as coordenadas dos indivíduos em um componente e cada uma das categorias.

No gráfico biplot abaixo cada alternativa para cada resposta do questionário é plotada em vermelho, enquanto que cada indivíduo (pessoa que respondeu o questionário) é plotado em vermelho. Estas obervações são plotadas em relação as duas dimensões mais explicativas, no caso as dimensões 1 e 2 e indica o quão bem cada dimensão explica aquele observação. Pode-se observar que há uma quantidade grande de indivíduos no canto inferior esquerdo do gráfico que não são bem explicados por pelas duas dimensões.

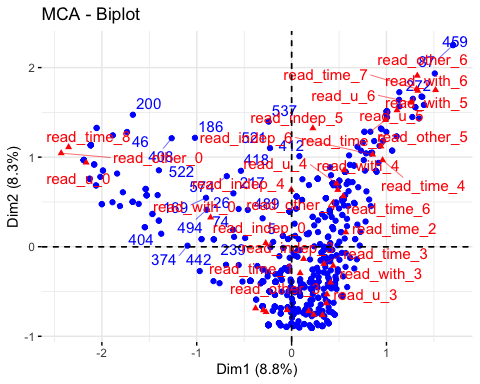


Figura n - Representação do quanto as duas principais dimensões explicam cada dividuo e variavel

Visualização e interpretação

Estudar a inércia dos componentes principais permite-nos, por um lado, ver se as variáveis estão estruturadas (presença de correlações entre variáveis) e, além disso, determinar o número de componentes a serem interpretados.

Os resultados em abaixo, produzidos em um gráfico de barras, correspondem ao autovalor (ou seja, a inércia ou a variação explicada) associada a cada um dos componentes; a porcentagem de inércia associada a cada componente e a soma cumulativa dessas porcentagens.

Porcentagens de inércia explicadas por cada dimensão MCA

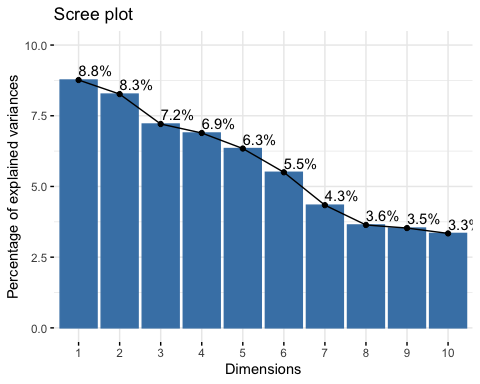
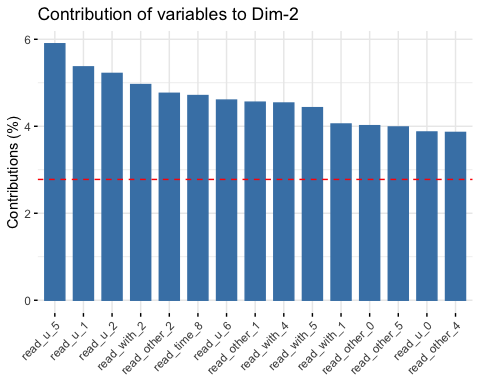
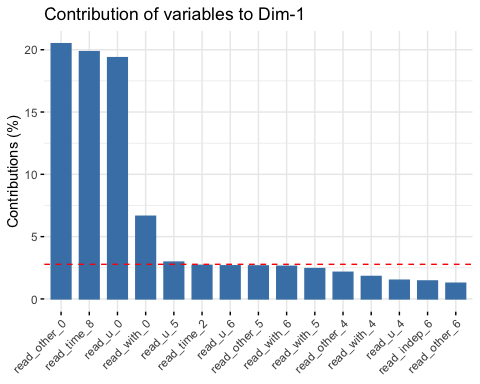
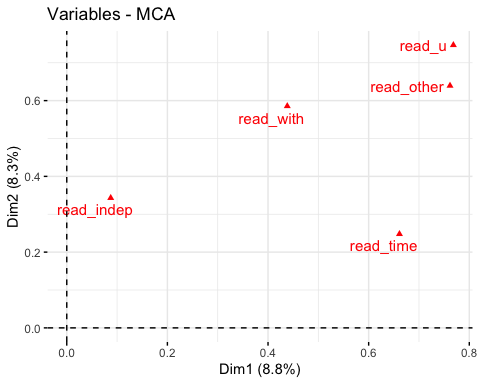


Figura n - Representação gráfica da porcentagem de variância explicada por cada dimensão

Dentre das 15 principais categorias do gráfico anterior, o plot de quanto elas contribuiem para as duas maiores dimensões

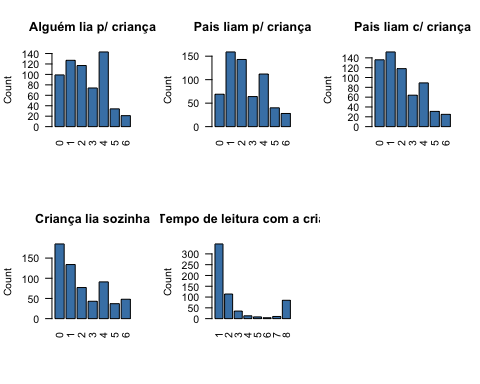


Correlação entre variáveis e dimensões principais



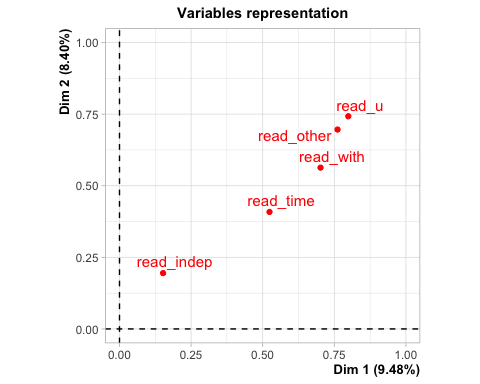
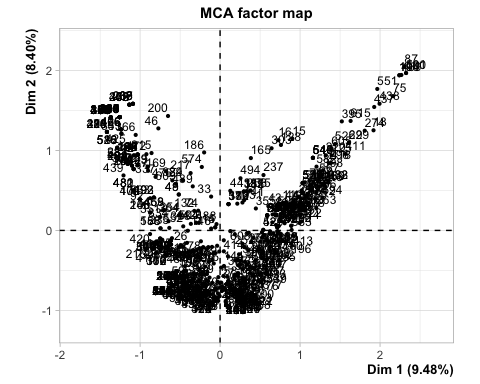
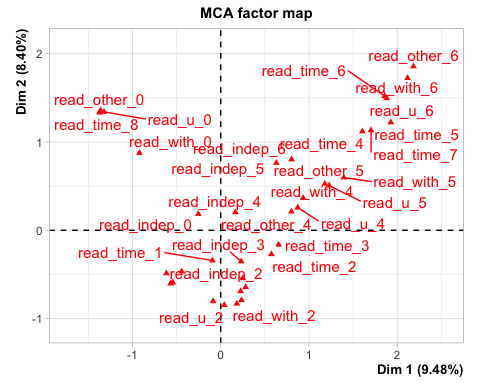
#### **Pós Pandemia**

## read\_other read\_u read\_with read\_indep read\_time   
## 0: 99 0:104 0:136 0:185 1 :346   
## 1:127 1:157 1:152 1:134 2 :114   
## 2:117 2:113 2:118 2: 77 8 : 85   
## 3: 74 3: 62 3: 64 3: 43 3 : 35   
## 4:143 4:109 4: 89 4: 91 4 : 13   
## 5: 34 5: 35 5: 31 5: 37 7 : 10   
## 6: 21 6: 35 6: 25 6: 48 (Other): 12



Construindo o MCA

As variáveis podem ser representadas calculando as razões de correlação entre as coordenadas dos indivíduos em um componente e cada uma das categorias.



No gráfico biplot abaixo cada alternativa para cada resposta do questionário é plotada em vermelho, enquanto que cada indivíduo (pessoa que respondeu o questionário) é plotado em vermelho. Estas obervações são plotadas em relação as duas dimensões mais explicativas, no caso as dimensões 1 e 2 e indica o quão bem cada dimensão explica aquele observação. Pode-se observar que há uma quantidade grande de indivíduos no canto inferior esquerdo do gráfico que não são bem explicados por pelas duas dimensões.

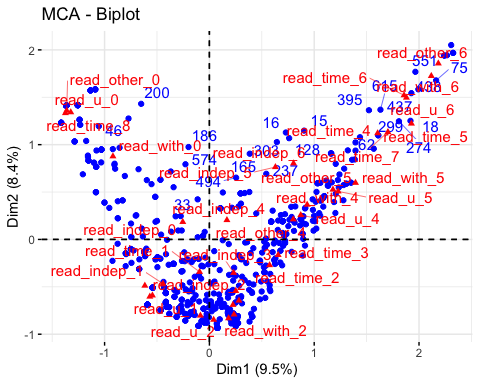


Figura n - Representação do quanto as duas principais dimensões explicam cada dividuo e variavel

Visualização e interpretação

Estudar a inércia dos componentes principais permite-nos, por um lado, ver se as variáveis estão estruturadas (presença de correlações entre variáveis) e, além disso, determinar o número de componentes a serem interpretados.

Os resultados em abaixo, produzidos em um gráfico de barras, correspondem ao autovalor (ou seja, a inércia ou a variação explicada) associada a cada um dos componentes; a porcentagem de inércia associada a cada componente e a soma cumulativa dessas porcentagens.

Porcentagens de inércia explicadas por cada dimensão MCA

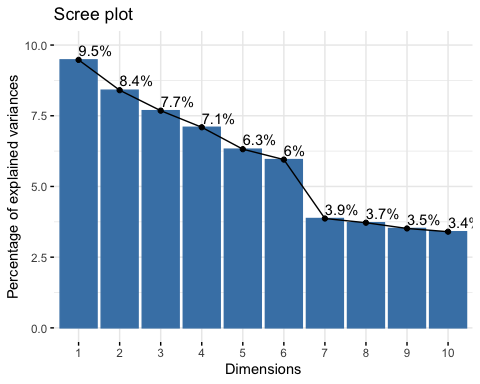
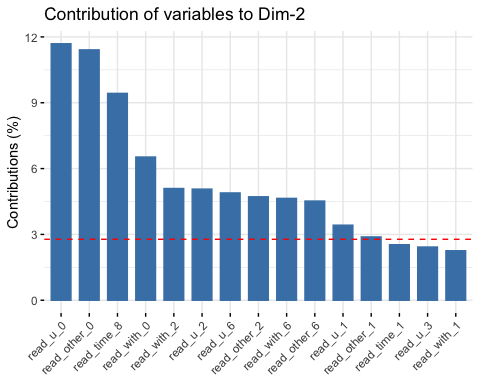
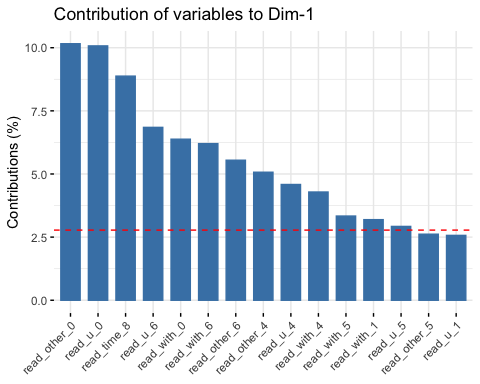
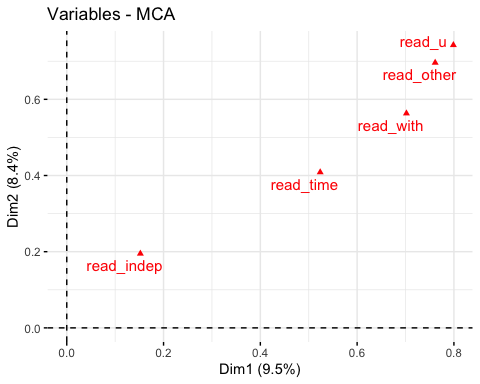


Figura n - Representação gráfica da porcentagem de variância explicada por cada dimensão

Dentre das 15 principais categorias do gráfico anterior, o plot de quanto elas contribuiem para as duas maiores dimensões

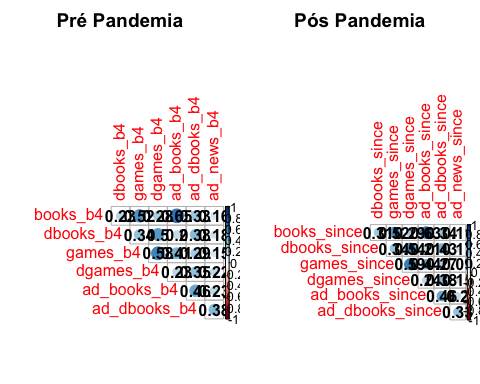


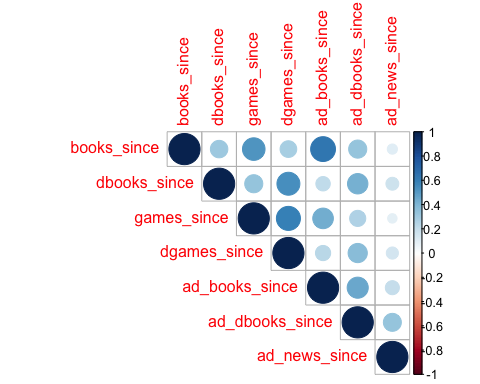
Correlação entre variáveis e dimensões principais



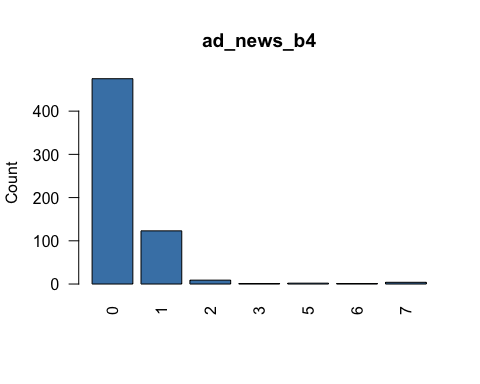
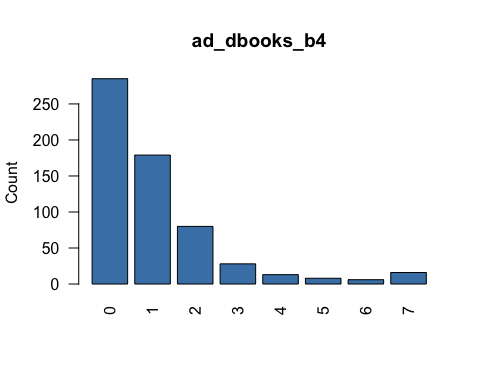
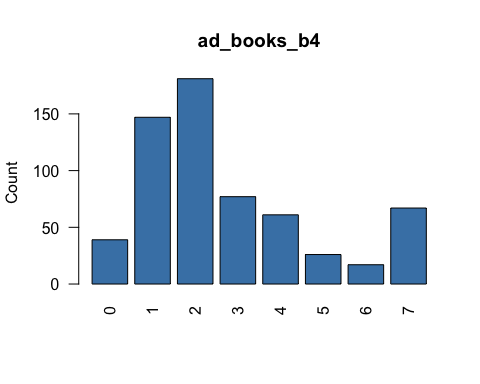
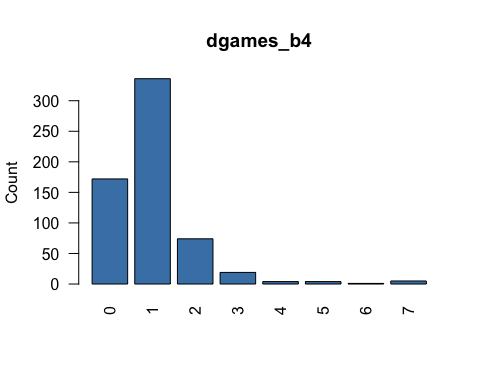
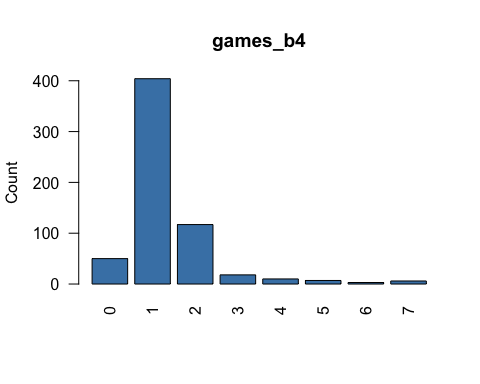
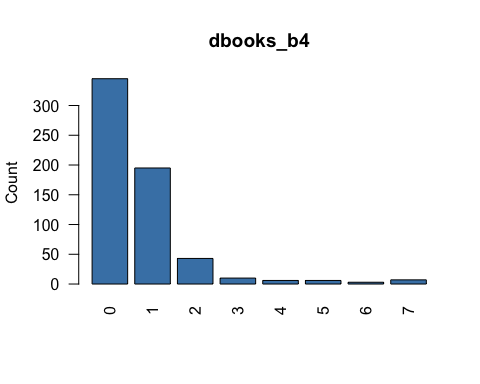
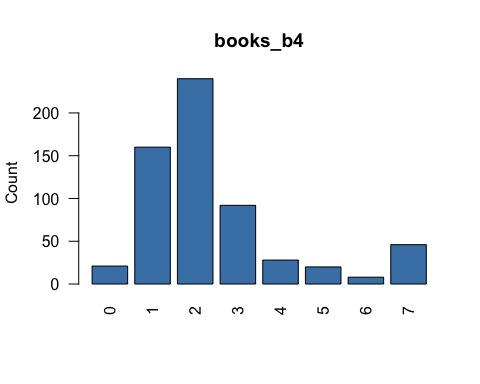
### Bloco 2 - Recursos de alfabetização inicial

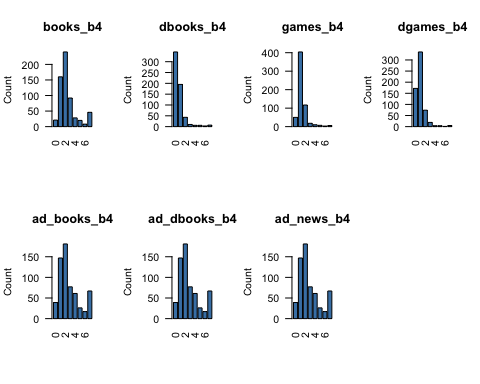
Mapa de calor para a correlação entre as variáveis do bloco 2





#### Pré Pandemia

Pelas frequências das categorias, a maioria dos recursos se concentram nas quantidades das categoria até 1. Apenas livros para crianças e adultos apresentam quantidades maiores. 

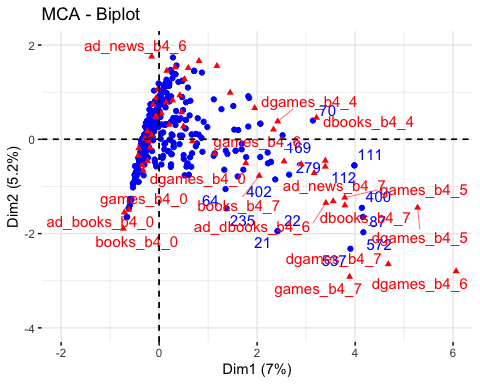


Há um grupo indivíduos (na extremidade positiva do segundo principal componente) muito próximos, pois compartilham muitas categorias em comuns. Assim como na extremidade negativa.

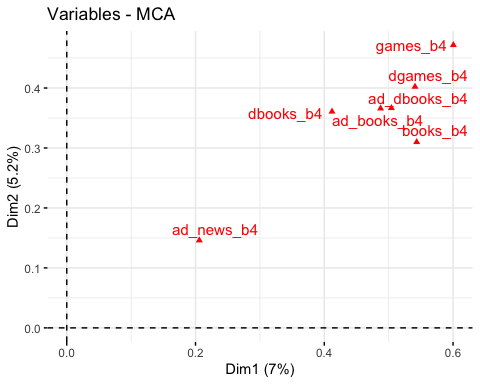
Já a primeira principal componente parece explicar um grupo seleto de pessoas.

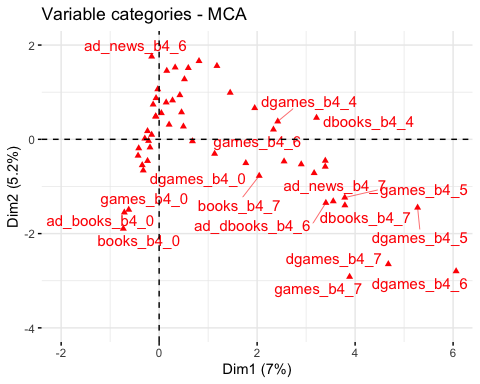
Analisando pelas categorias, o primeiro grupo citado apresenta categorias de quantidade entre 2 e 6 na sua extremidade, no segundo as categorias as extremidades se concetram em zero. E o terceiro categorias em 7. Além disso parece haver um grupo naa1aaa2ão bem explicado em ambas.

Construindo MCA

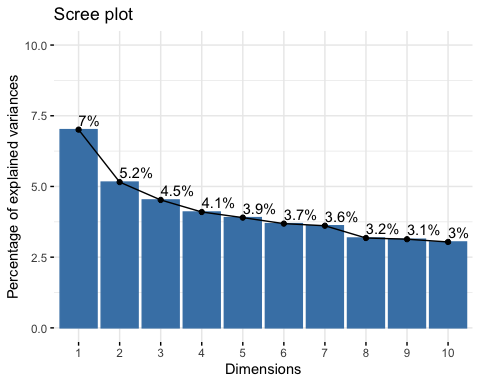


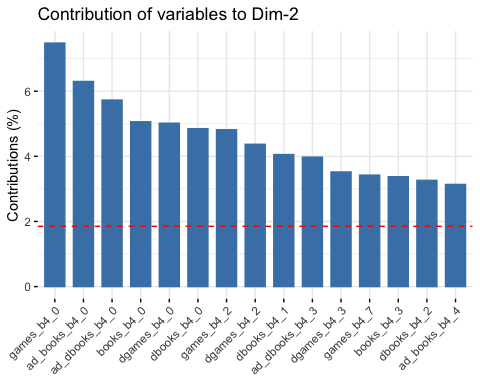
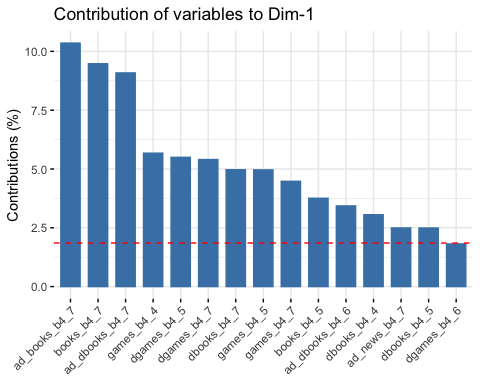
A variável **hlr\_ad\_news\_b4** é a menos correlacionada com ambas dimensões. As demais variáveis sao próximas nos dos primeiros componentes.

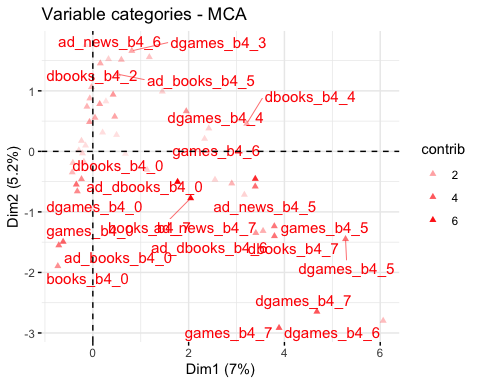




As 5 primeiras dimensões juntas explicam cerca de 25% da variância dos dados. E as 2 primeiras 12%. Porcentagens de inércia explicadas por cada dimensão MCA

 15 principais categorias de variáveis que contribuem para as dimensões:



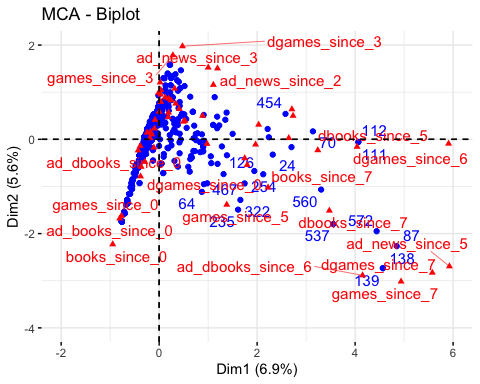


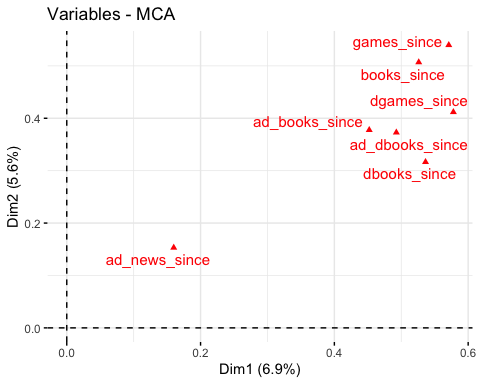
#### Pós Pandemia

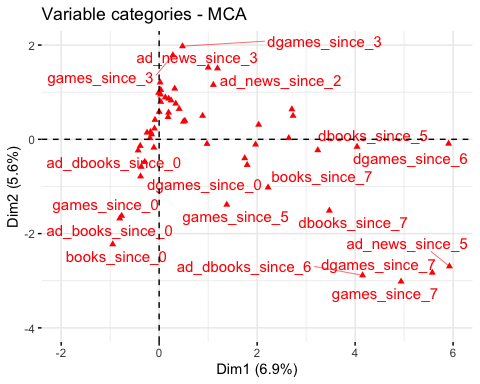
Há um grupo indivíduos (na extremidade positiva do segundo principal componente) muito próximos, pois compartilham muitas categorias em comuns. Assim como na extremidade negativa.

Já a primeira principal componente parece explicar um grupo seleto de pessoas.

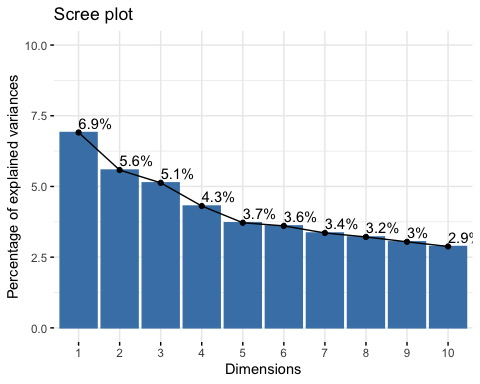
Analisando pelas categorias, o primeiro grupo citado apossenta categorias de quantidade entre 2 e 3 na sua extremidade superior, no segundo as categorias as extremidades se concetram em zero. E o terceiro categorias em 7. Além disso parece haver um grupo naa1aaa2ão bem explicado em ambas.

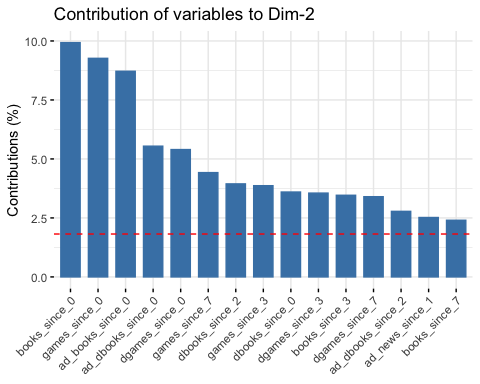
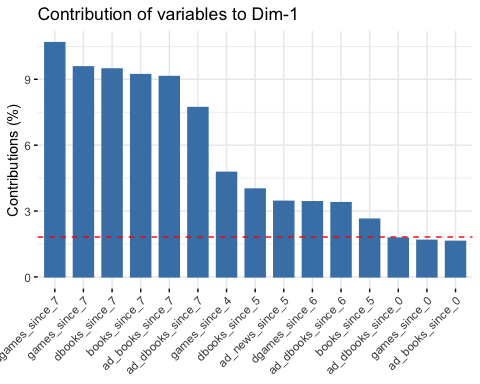
 A variável **hlr\_ad\_news\_since** é a menos correlacionada com ambas dimensões. As demais variáveis sao próximas nos dos primeiros componentes.

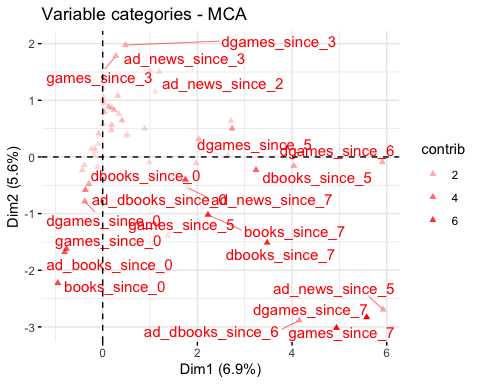


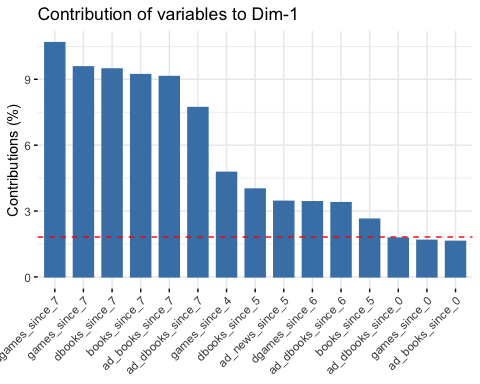
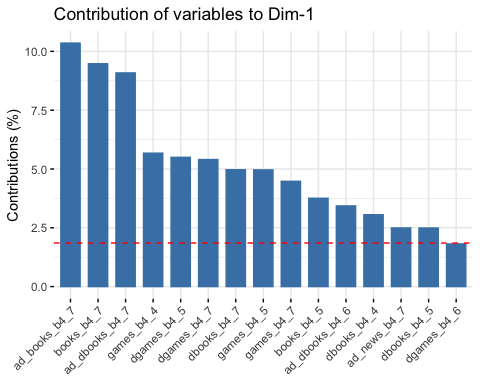


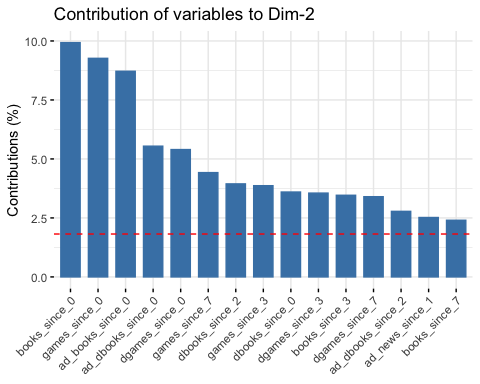
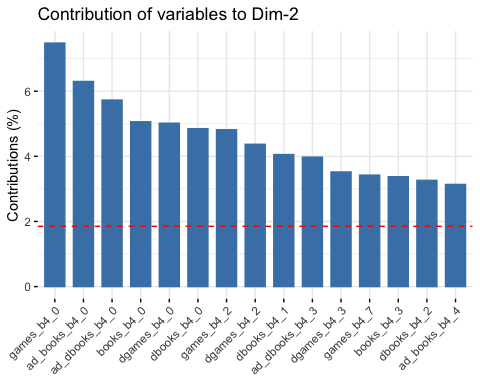
As 5 primeiras dimensões juntas explicam 25.7% da variância dos dados. E as 2 primeiras 12.5%. Porcentagens de inércia explicadas por cada dimensão MCA

 15 principais categorias de variáveis que contribuem para as dimensões:









### Bloco 3 - Atividades de enriquecimento

|  |  |
| --- | --- |
| Pré Covid | Com Covid |

ea\_read\_b4\_ALL\_br | ea\_read\_since\_yng\_br |  
ea\_indr\_b4\_ALL\_br | ea\_indr\_since\_yng\_br |  
ea\_letter\_b4\_yng\_br | ea\_letter\_since\_yng\_br |  
ea\_tv\_b4\_yng\_br | ea\_tv\_since\_yng\_br |  
ea\_boardg\_b4\_yng\_br | ea\_boardg\_since\_yng\_br |  
ea\_vidg\_b4\_yng\_br | ea\_vidg\_since\_yng\_br |  
ea\_edapp\_b4\_yng\_br | ea\_edapp\_since\_yng\_br |

Mapa de calor para a correlação entre as variáveis do bloco 3

Para o bloco pré-pandemia, pode-se observar que as variávies **ea\_read\_b4\_AAL\_br** e **ea\_edapp\_b4\_yng\_br** possuem a menor correlação, ou seja, crianças que se envolvem em alguma atividade de leitura e crianças que utilizam algum aplicativo educacional em tablet são pouco correlacionadas, enquanto que as variáveis **ea\_vidg\_b4\_yng\_br** e **ea\_read\_b4\_AAL\_br** têm a maior correlação, ou seja, crianças que se envolvem em alguma atividade de leitura são fortemente correlacionadas com crianças que assitem vídeos ou jogam jogos educacionais no computador.

Já para o bloco pós-pandemia, pode-se observar que a menor correlação continua sendo entre as variáveis **ea\_read\_b4\_AAL\_br** e **ea\_edapp\_b4\_yng\_br**, enquanto que a maior correlação passar a ser entre as variáveis **ea\_tv\_since\_yng\_br** e **ea\_vidg\_since\_yng\_br**, ou seja, entre crianças que assistem a vídeos/programas educacionais com computador e na TV.

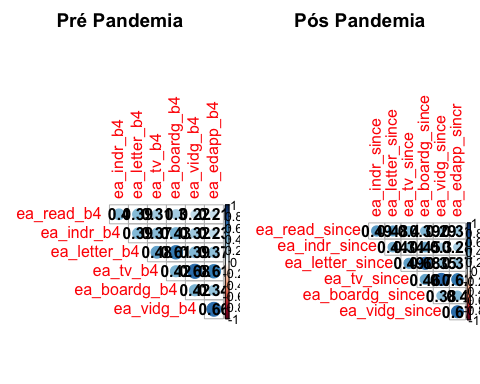


Figura n - Correlação entre variáveis para os blocos pré e pós pandemia

## Multiple Correspondence Analysis

#### **Pré Pandemia**

Podemos observar nos gráficos abaixo a frequência absoluta respostas para cada categoria de cada pergunta.

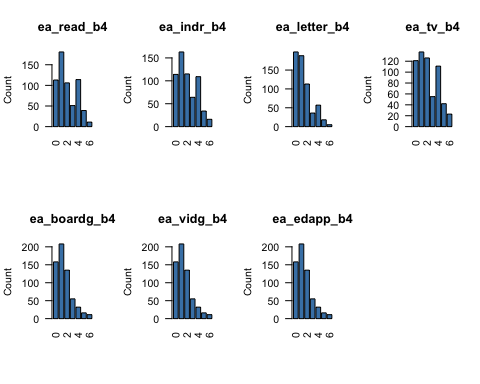
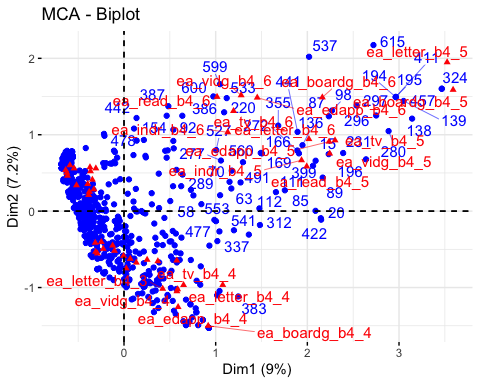
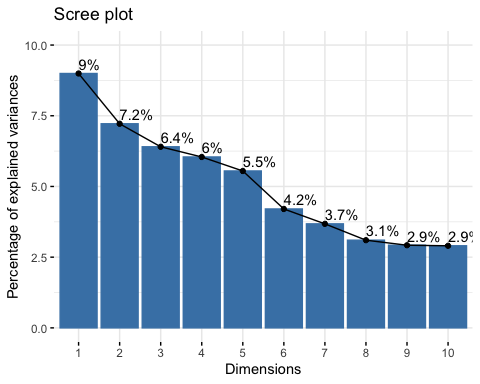


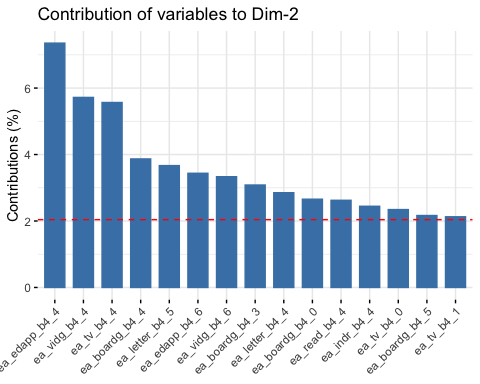
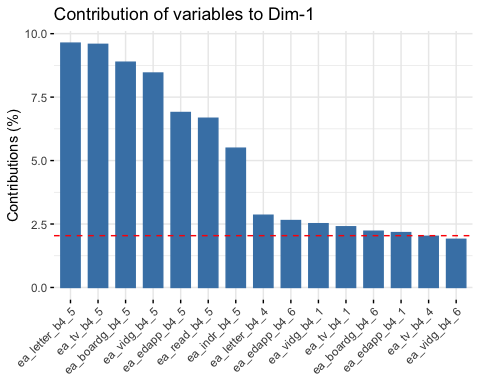
Figura n - Quantidade de respostas por alternativa

Construindo o MCA

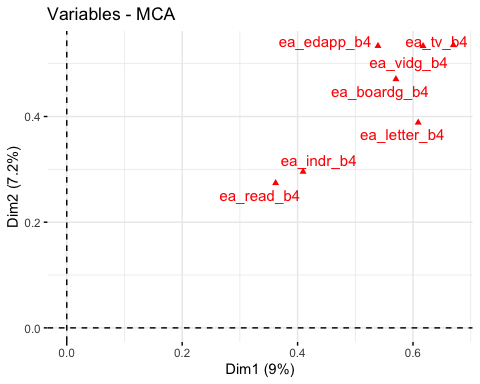
No gráfico biplot abaixo cada alternativa para cada resposta do questionário é plotada em vermelho, enquanto que cada indivíduo (pessoa que respondeu o questionário) é plotado em vermelho. Estas obervações são plotadas em relação as duas dimensões mais explicativas, no caso as dimensões 1 e 2 e indica o quão bem cada dimensão explica aquele observação. Pode-se observar que há uma quantidade grande de indivíduos no canto inferior esquerdo do gráfico que não são bem explicados por pelas duas dimensões. 

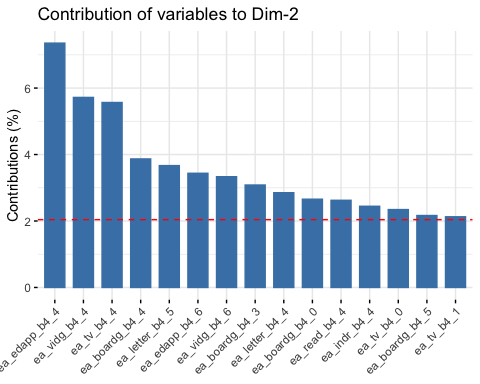
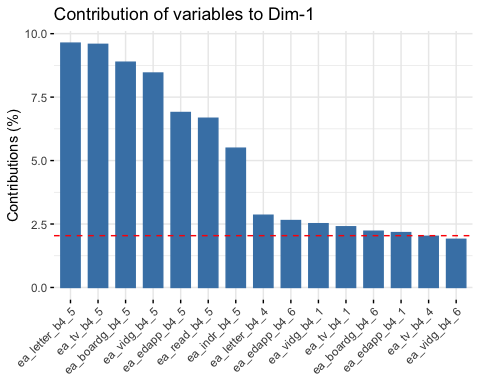
Visualização e interpretação

Representação gráfica da porcentagem de variância explicada por cada dimensão.  Quanto cada uma das alternativas do questionários contribuem para as dimensções 1 e 2

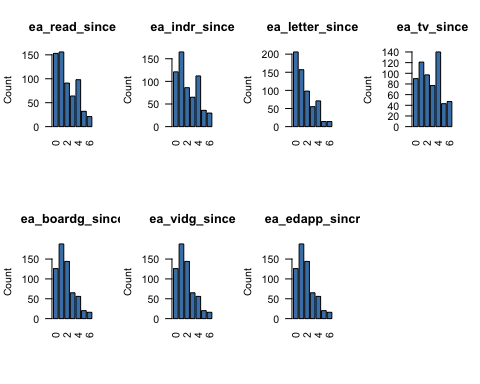


Correlação entre variáveis e dimensões principais

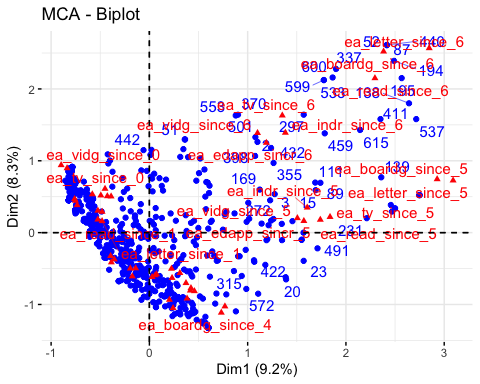
Pode-se observar que para as duas principais dimensões, a maioria das variáveis têm um correlação alta, destoando as variáveis **ea\_read\_b4\_AAL\_br** e **ea\_indr\_b4\_AAL\_br**. 

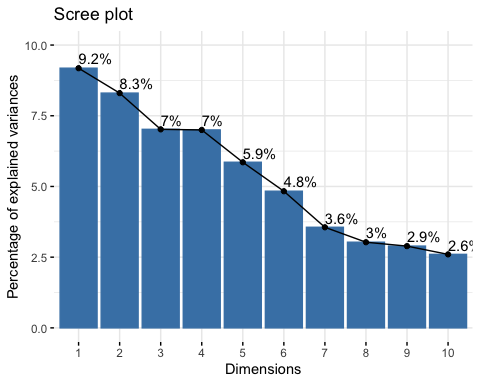
15 principais categorias de variáveis que contribuem para cada uma das principais dimensões: 

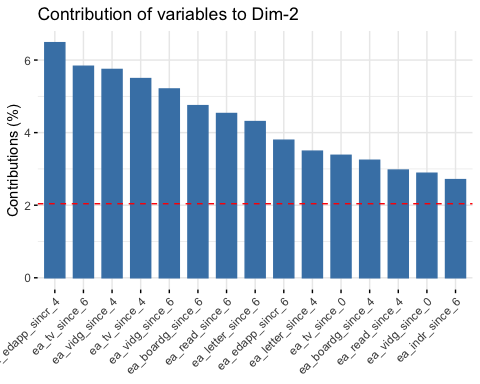
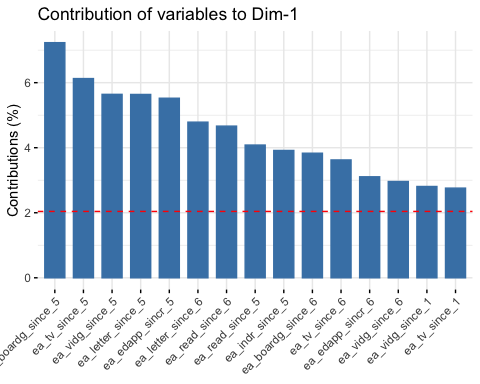
#### **PÓS-PANDEMIA**

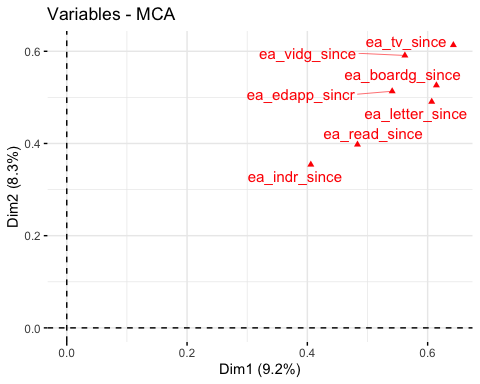
Podemos observar nos gráficos abaixo a frequência absoluta respostas para cada categoria de cada pergunta. 

Construindo o MCA

Assim com no biplot pré-pandemia, podemos observar que há uma grande quantidade de indivíduos que não são explicados pelas duas principais dimensões. 

Porcentagens de inércia explicadas por cada dimensão MCA 

 > Correlação entre variáveis e dimensões principais

Pode-se observar que para as duas principais dimensões, a maioria das variáveis têm um correlação alta, destoando as variáveis **ea\_indr\_since** e **ea\_read\_since**. 

15 principais categorias de variáveis que contribuem para cada uma das principais dimensões: 