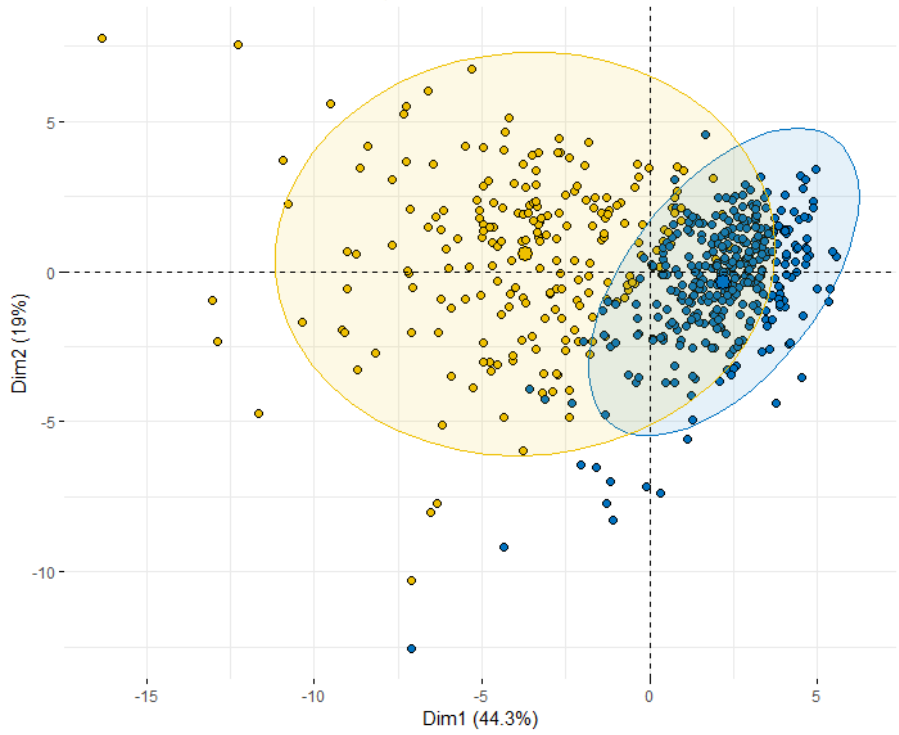
Seminário - Estatística Multivaridada Computacional

Code

Estatística Multivaridada Computacional: Teresina - Piauí



Gabriel Augusto Narciso Barreiros
Adriano José de Barros
Nelson Rios

https://github.com/Gabrielj96/Trabalho-R-UFMG-Multivariada (https://github.com/Gabrielj96/Trabalho-R-UFMG-Multivariada)

0. Bibliotecas e configurações

Show

1. Análise Descritiva

Tamanho populacional

Cidades maiores tendem a ser mais heterogêneas do que cidades menores. Isso significa que as análises PCA realizadas em cidades maiores são mais propensas a identificar padrões complexos e não lineares.

Seguindo essa lógica, talvez, Teresina padrão mais simples como menor variância, menores correlações e menor multicolinearidade facilitando a interpretação dos dados.

Show

```
[1] 224188
```

Show

	ID <dbl>	N <dbl>
3	3	8094
8	8	8279
12	12	8381
2	2	8529
1	1	8581
4	4	8606
9	9	8610
13	13	8803
14	14	9221
10	10	9407

1-10 of 16 rows Previous **1** 2 Next

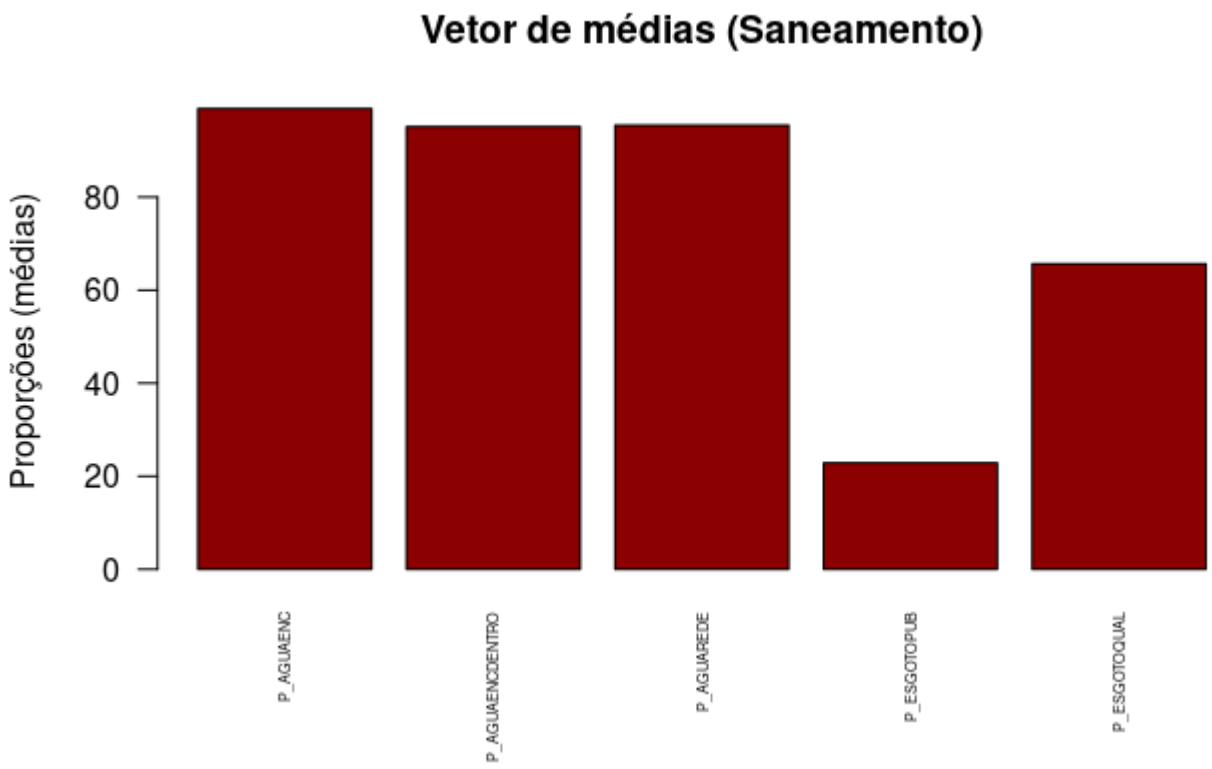
Vetor de médias

Saneamento

Analisando as variáveis de saneamento percebemos que a cidade tem um bom sistema de água encanada, atingindo um total de 99% de encaneamento de água e não ficando abaixo de 95%.

Já o sistema de esgoto com 88% fica um pouco abaixo, mas nada que podemos considerar ruim. Talvez uma enfâse na rede pública pudesse melhorar a situação.

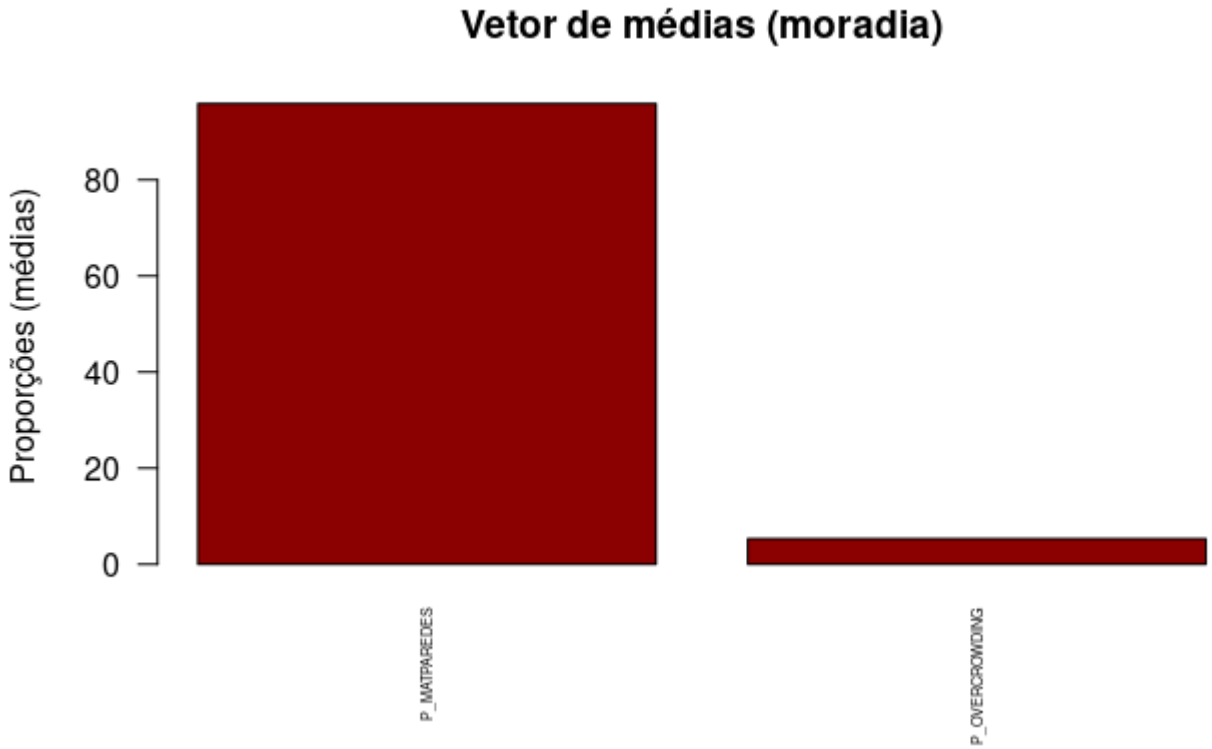
Show



Moradia

Analisando as variáveis de moradia é perceptível que a cidade não tem muitos problemas com isso. 95% da população tem algum tipo de moradia e apenas 5% com mais de 3 pessoas no lar.

Show

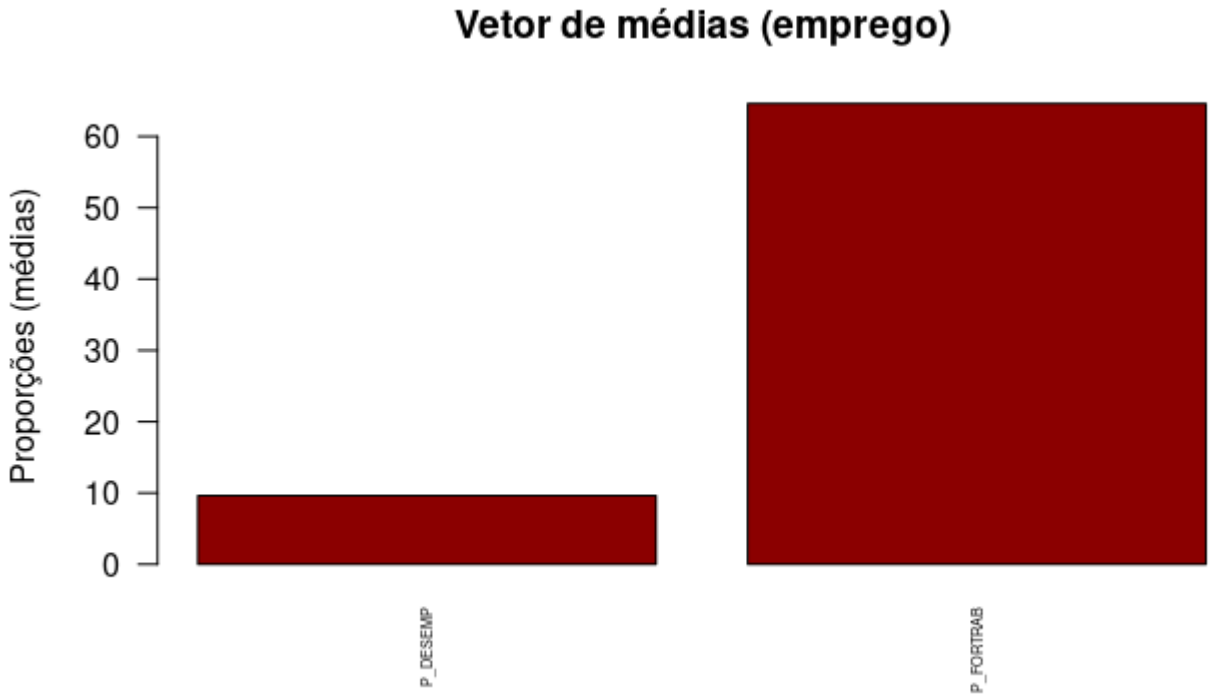


Emprego

Considerando que 7% de desemprego é considerado pleno emprego no Brasil, Teresina com 9% ainda tem um pouco a melhorar.

De acordo com a Organização Internacional do Trabalho (OIT), uma taxa de participação da força de trabalho de 60% ou mais é considerada elevada. Seguindo esse parâmetro Teresina está acima da média com 65%.

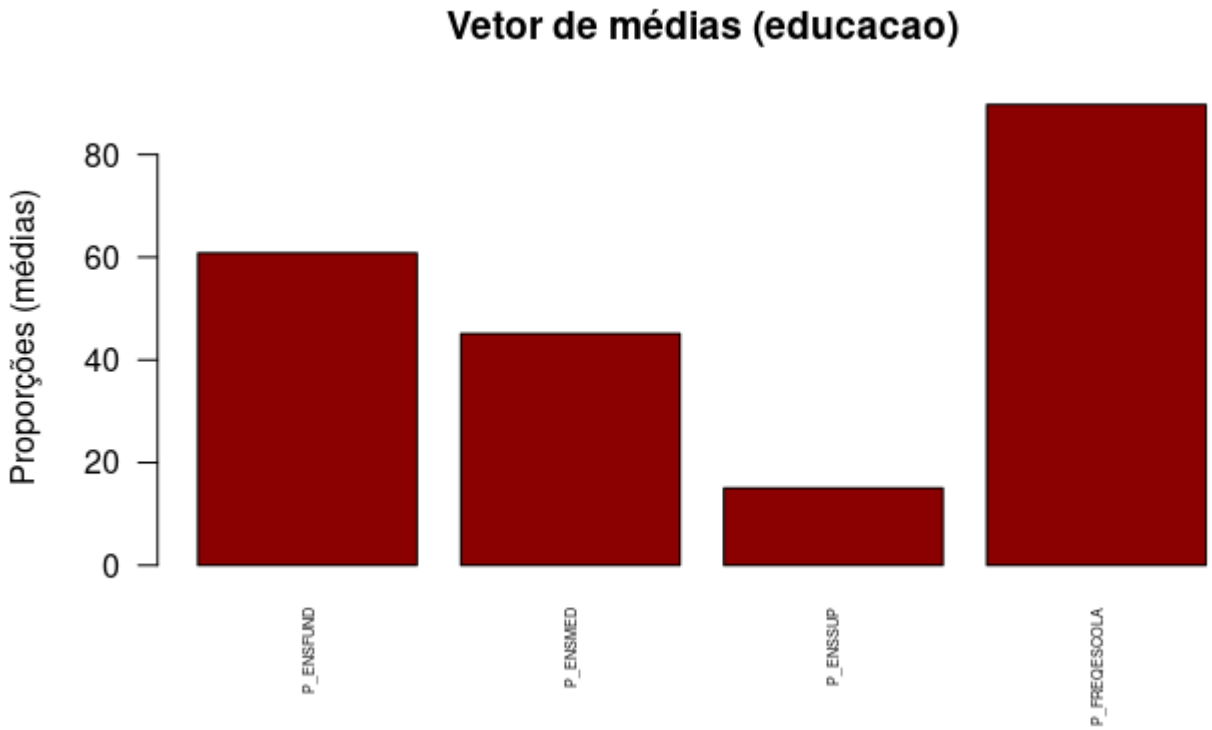
Show



Educação

A cidade tem boa parte da população frequentando a escola (89%), mas essa taxa cai consideravelmente ao aumentar o nível de escolaridade, chegando a 15% no ensino superior. O que da espaço para políticas públicas para a educação.

Show



Matriz de correlação

Educação vs Overcrowding = Correlação Negativa

Quanto menor a educação, maior o número de residências com superlotação.

Overcrowding vs Material Paredes = Correlação Negativa

Quanto maior o número de residências com superlotação, menor o número de domicílios com paredes feitas de materiais duráveis.

Educação vs Material Paredes = Correlação Positiva

Quanto maior a educação, maior o número de residências com paredes feitas de materiais duráveis.

Educação vs Esgoto = Correlação Positiva

Quanto maior a educação, maior o acesso a esgoto.

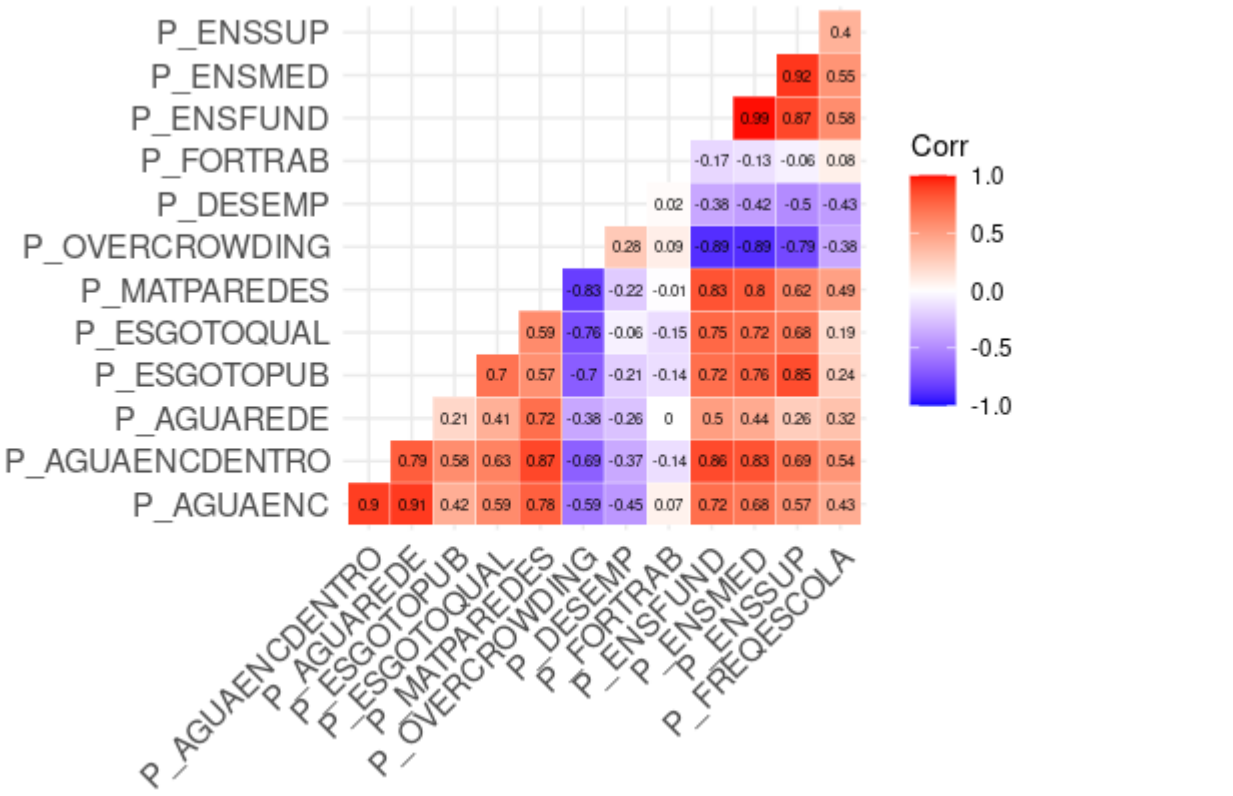
Material Paredes vs Residência Encanada = Correlação Positiva

Quanto mais residências com paredes de materiais duráveis, mais residências encanadas.

Educação vs Residência Encanada = Correlação Positiva

Quanto maior a educação, mais residências encanadas.

Show



2. Análise Qualitativa PCA

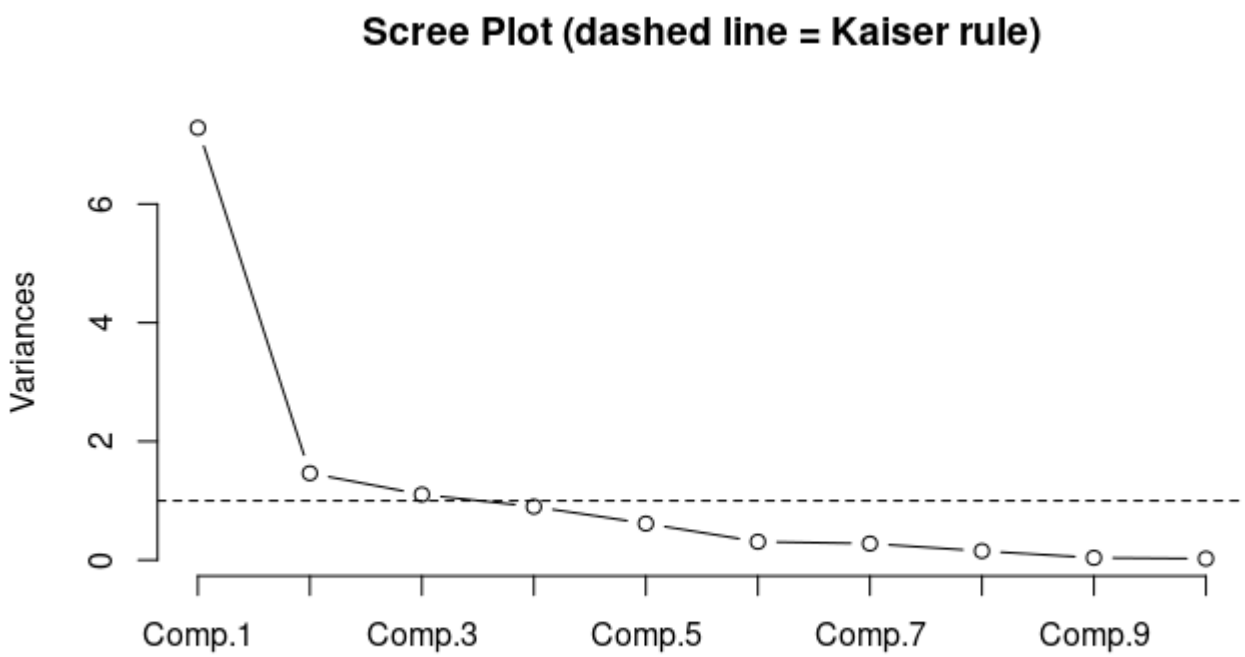
Show

```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5     Comp.
6     Comp.7     Comp.8      Comp.9     Comp.10     Comp.11      Comp.12       Comp.1
3
Standard deviation     2.6987995 1.2095882 1.05079758 0.94833494 0.78426232 0.5554105
3 0.52690299 0.38998252 0.194472377 0.153693768 0.127856020 0.0750477266 0.0260899008
7
Proportion of Variance 0.5976221 0.1200495 0.09059902 0.07379193 0.05046707 0.0253112
5 0.02277963 0.01247888 0.003103139 0.001938197 0.001341306 0.0004621261 0.0000558509
1
Cumulative Proportion  0.5976221 0.7176716 0.80827062 0.88206255 0.93252962 0.9578408
7 0.98062050 0.99309938 0.996202521 0.998140717 0.999482023 0.9999441491 1.0000000000
0
```

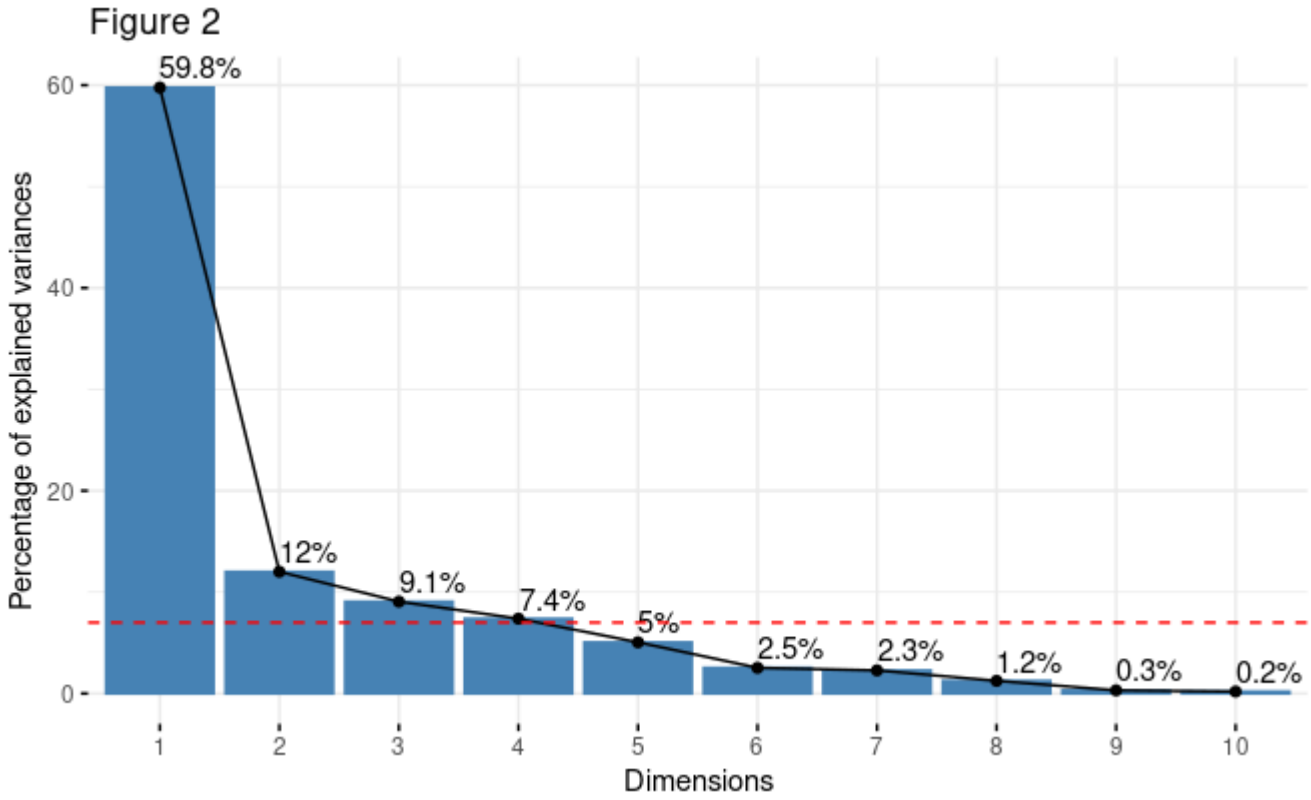
O primeiro componente explica 59.8% da variância total, enquanto o segundo 12% e o terceiro 9.1%.

E segundo a regra de Kaiser os primeiros 3 componentes, 81% da variação, podem precisamente representar os dados.

Show



Show



Proximidade

As variáveis com maior proximidade são os educação e acesso a sistema de esgoto, ou seja, o acesso a educação está relacionada com acesso à algum tipo de sistema de esgoto na residência.

Distância

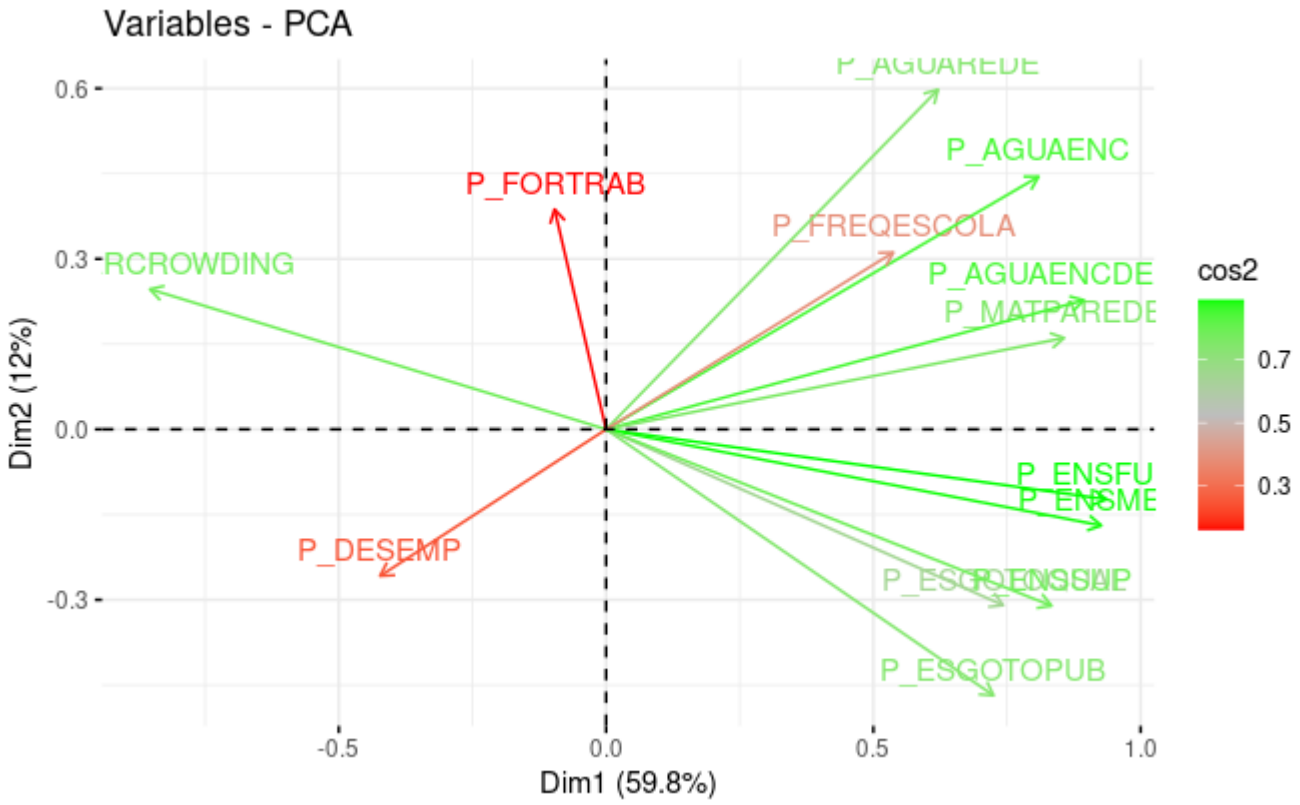
As variáveis com maior distância são superlotação e educação, ou seja, o acesso a educação está inversamente relacionada com a superlotação nas residências.

Também percebemos a distância entre a taxa de desemprego, encanamento e frequência escolar. O aumento da taxa de desemprego está relacionada negativamente com esses problemas sociais.

Distância da Origem e Coloração

Com base na distância da variável em relação ao centro do gráfico percebemos que as variáveis mais bem representadas por esses componentes principais são: educação e saneamento.

Show



3. Análise Qualitativa EFA

Show

Critério de Kaiser–Meyer–Olkin

Segundo o critério de Kaiser, podemos questionar se análise fatorial é a mais adequada para esses dados já que o teste retornou MSA = 0.5.

Um número baixo significa que as variáveis analisadas não apresentam correlações fortes o suficiente para serem agrupadas em fatores, 0.5 é um valor aceitável, porém não tão alto.

Show

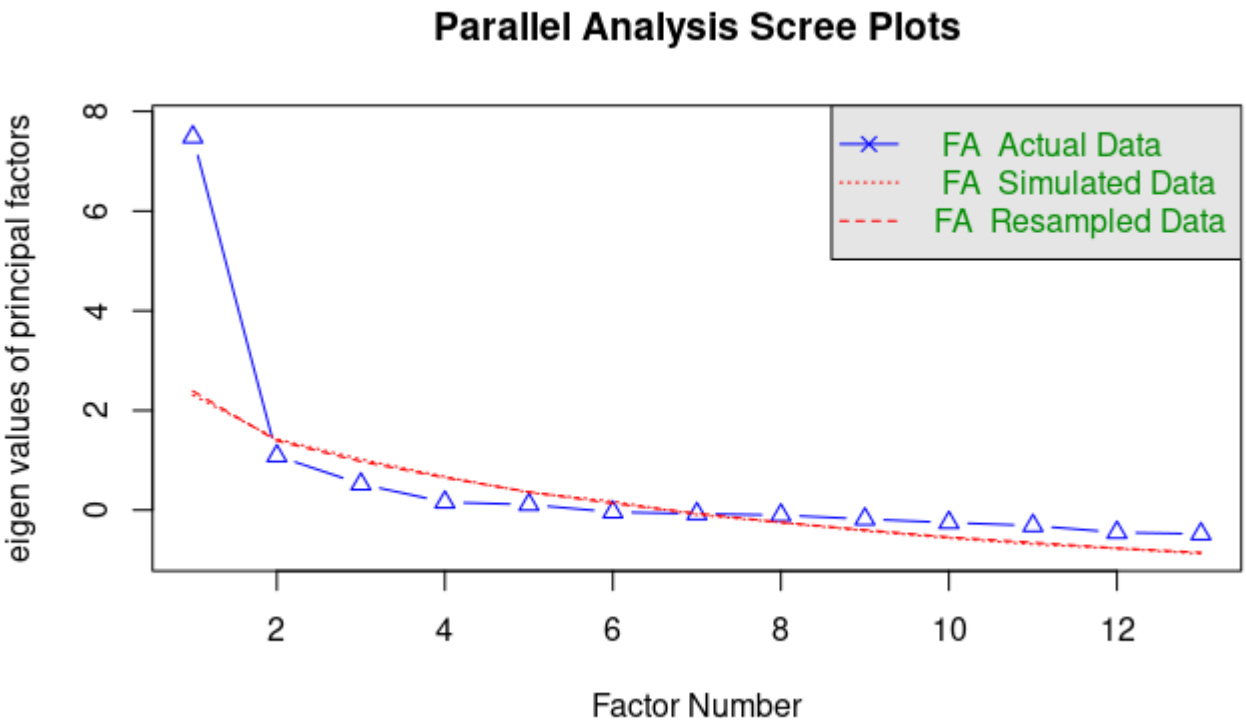
```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = df_teresina_efa)
Overall MSA =  0.55
MSA for each item =
      P_AGUAENC P_AGUAENCDENTRO      P_AGUAREDE     P_ESGOTOPUB    P_ESGOTOQUAL     P_
MATPAREDES  P_OVERCROWDING        P_DESEMP       P_ENSFUND        P_ENSMED        P_E
NSSUP       P_FORTRAB    P_FREQESCOLA
           0.48            0.63            0.48            0.70            0.64
0.81            0.73            0.22            0.59            0.60            0.69
0.03            0.32
```

Scree Plots

O gráfico mostra que a curva dos dados reais (linha azul) se aproxima da curva de scree após o quarto fator. Isso sugere que os primeiros quatro fatores explicam a maior parte da variação nos dados.

Show

```
Parallel analysis suggests that the number of factors =  1  and the number of compone
nts =  NA
```



Análise Fatorial

O scree plot sugere que 4 ou 5 fatores são capazes de explicar boa proporção da variância dos dados.

A regra de Kaiser indica apenas 3 fatores, e 3 fatores explicam apenas 72% da variação total.

Já 5 fatores explicam 83% da variação total, porém o p-valor fica um tanto alto. Neste caso podemos aceitar 72% de explicação da variação total e um p-valor menor com 3 fatores.

Show

```
Call:
factanal(x = df_teresina_efa, factors = 5, scores = "regression",     nstart = 100)

Uniquenesses:
      P_AGUAENC P_AGUAENCDENTRO      P_AGUAREDE     P_ESGOTOPUB    P_ESGOTOQUAL    P_
MATPAREDES  P_OVERCROWDING        P_DESEMP       P_ENSFUND        P_ENSMED        P_E
NSSUP       P_FORTRAB    P_FREQESCOLA
          0.005           0.075           0.040           0.188           0.005
0.005           0.093           0.395           0.005           0.005           0.005
0.829           0.500

Loadings:
               Factor1 Factor2 Factor3 Factor4 Factor5
P_AGUAENC        0.345   0.882   0.310
P_AGUAENCDENTRO  0.484   0.727   0.293   0.255   0.102
P_AGUAREDE               0.961   0.100          -0.106
P_ESGOTOPUB      0.885           0.136
P_ESGOTOQUAL     0.794   0.434  -0.128  -0.266   0.298
P_MATPAREDES     0.560   0.648           0.511
P_OVERCROWDING  -0.810  -0.336          -0.318  -0.193
P_DESEMP        -0.154  -0.181  -0.729           0.115
P_ENSFUND        0.724   0.430   0.292   0.298   0.335
P_ENSMED         0.767   0.353   0.346   0.294   0.278
P_ENSSUP         0.866   0.137   0.463
P_FORTRAB                                       -0.400
P_FREQESCOLA     0.163   0.291   0.353   0.423   0.291

               Factor1 Factor2 Factor3 Factor4 Factor5
SS loadings      4.661   3.406   1.306   0.878   0.601
Proportion Var   0.359   0.262   0.100   0.068   0.046
Cumulative Var   0.359   0.621   0.721   0.789   0.835

Test of the hypothesis that 5 factors are sufficient.
The chi square statistic is 26.56 on 23 degrees of freedom.
The p-value is 0.275
```

Show

```
objective
0.2379858
```

Show

```
objective
0.1858346
```

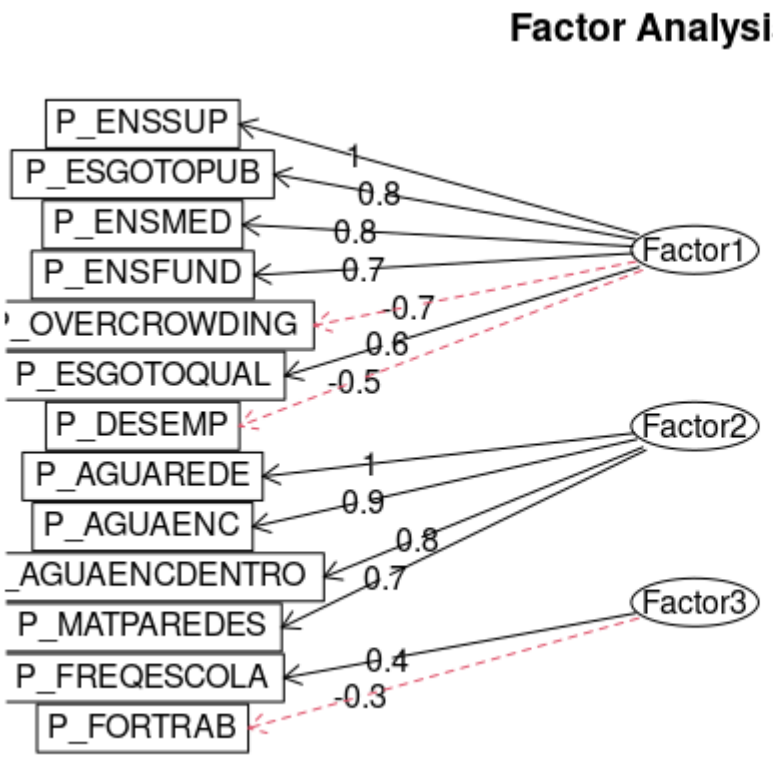
Gráfico de Análise Fatorial

Podemos interpretar os fatores da seguinte forma:

- **Fator 1:** Educação, saneamento, moradia e emprego
- **Fator 2:** Saneamento, moradia

- **Fator 3:** Educação e emprego

Show



4. Efeitos das Regras de Seleção

Regras de seleção PCA

Uma regra comum é selecionar o número de componentes que explicam uma certa porcentagem da variância total (75%).

Por exemplo, utilizamos essa regra e obtivemos PCA que explica 81% da variação e apenas 3 componentes. Considero um bom resultado com poucas variáveis considerando a quantidade na base de dados.

No entanto, esta regra pode levar à seleção de um número excessivo de componentes se as variáveis originais estiverem altamente correlacionadas.

Outra regra comum é selecionar o número de componentes com base na regra de Kaiser.

No caso dos nossos dados o número de componentes seria o mesmo nas duas regras, mas ela pode levar a escolher um número insuficiente de componentes.

Regras de seleção EFA

Uma regra comum é selecionar o número de componentes que explicam uma certa porcentagem da variância total (75%).

No caso dos dados de Teresina utilizamos a **Regra do scree plot** + **Teste de hipótese** isso nos ajuda a obter uma Análise Fatorial mais confiável, combinando % de variação e significância da análise.

Resultado: Variação Total Explicada = 72%, P-valor = 0.18.

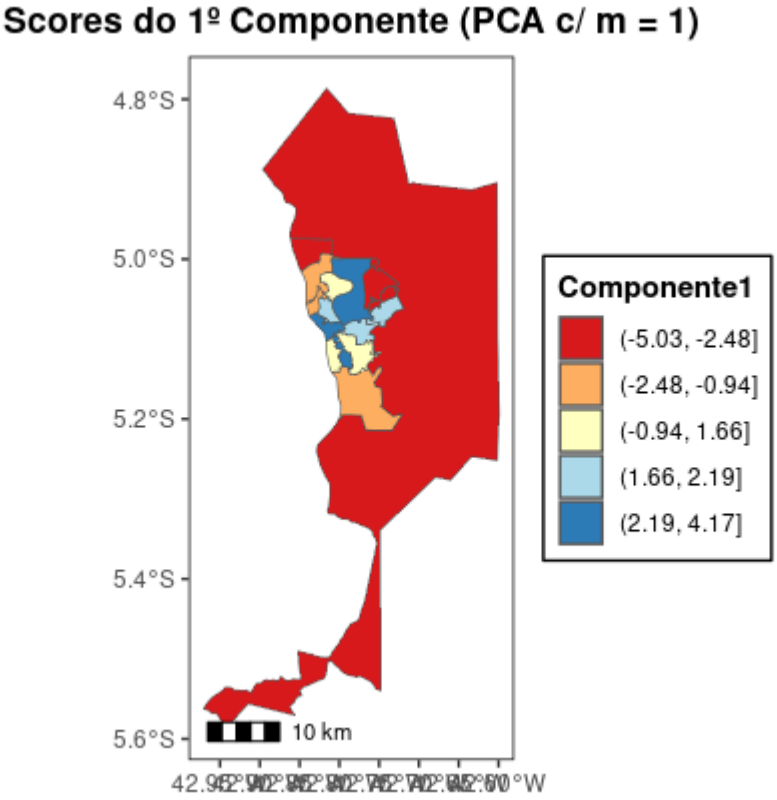
Poderíamos ter utilizado a **Regra da porcentagem da variância explicada** o que nos traria maior explição da variação total, mas com confiabiliade menor já que o P-valor se elevaria bastante.

5. Índice de Status Socioeconômico

Show

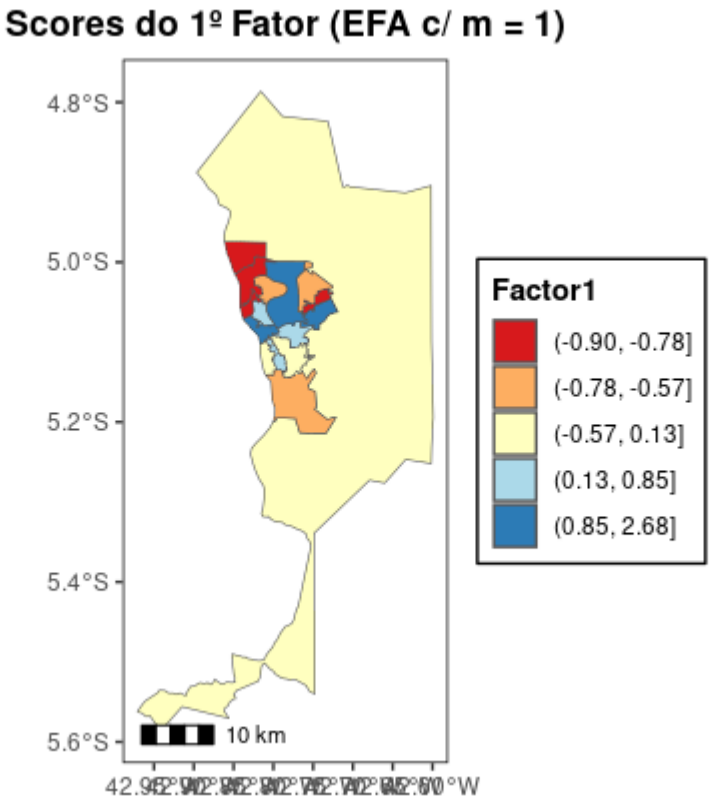
Índice Geral PCA

Show



Índice Geral EFA

Show



6. Conclusões

- A cidade, no geral, tem boas métricas de desenvolvimento, mas o setor da educação precisa ser melhorado.
- Melhorando a educação melhoraria boa parte dos problemas sociais da cidade já que está fortemente correlacionada com várias variável.
- Teresina poderia melhorar o sistema de esgoto de rede pública.
- A melhorando as oportunidades de emprego poderia melhorar a frequência escolar e consequentemente os outros problemas sociais.
- Tanto no mapa PCA quanto no EFA percebemos melhor qualidade de vida no centro da cidade enquanto nos extremos norte e sul uma situação mais precária.