

Análise de dados de uma estação de tratamento de água

Fernando Karchiloff Gouveia de Amorim - 10387644
Priscila Shibata Mendes - 8540501



Sobre o banco de dados

O conjunto de dados corresponde a uma estação genérica de tratamento de águas, sendo apresentadas diversas variáveis de dados coletados durante o processo de tratamento de água, desde a entrada até sua saída

O total de entradas do conjunto é de 527 dias, com 38 variáveis. Entretanto, algumas variáveis não foram medidas em alguns, o que é requisito para aplicar algumas análises, fazendo com que as observações passassem a girar em torno de 380. Tal medida foi necessária, pois senão aconteceriam inconsistências nas análises.



As variáveis levantadas pelo banco

- Data : Data da medição
- Q-E : Vazão de água que entra na estação
- ZN-E : Quantidade de Zinco na água que entra na estação
- PH-E : pH (potencial Hidrogeniônico) da água que entra na estação
- DBO-E : Demanda bioquímica de oxigênio da água que entra na estação
- DQO-E : Demanda química de oxigênio da água que entra na estação
- SS-E : Sólidos suspensos na água que entra na estação
- SSV-E : Sólidos suspensos voláteis na água que entra na estação
- SED-E : Quantidade de sedimentos na água que entra na estação
- COND-E : Condutividade da água que entra na estação
- PH-P : pH (potencial Hidrogeniônico) da água no primeiro “decantador”
- DBO-P : Demanda bioquímica de oxigênio da água no primeiro “decantador”



As variáveis levantadas pelo banco

- SS-P : Sólidos suspensos na água no primeiro “decantador”
- SSV-P : Sólidos suspensos voláteis na água no primeiro “decantador”
- SED-P : Quantidade de sedimentos na água no primeiro “decantador”
- COND-P: Condutividade da água no primeiro “decantador”
- PH-D : pH (potencial Hidrogeniônico) da água no segundo “decantador”
- DBO-D : Demanda bioquímica de oxigênio no segundo “decantador”
- DQO-D : Demanda química de oxigênio no segundo “decantador”
- SS-D : Sólidos suspensos na água no segundo “decantador”
- SSV-D : Sólidos suspensos voláteis na água no segundo “decantador”
- SED-D : Quantidade de sedimentos na água no segundo “decantador”
- COND-D : Condutividade da água no segundo “decantador”
- PH-S : pH (potencial Hidrogeniônico) da água na saída da estação



As variáveis levantadas pelo banco

- DBO-S : Demanda bioquímica de oxigênio da água na saída da estação
- DQO-S : Demanda química de oxigênio da água na saída da estação
- SS-S : Sólidos suspensos na água na saída da estação
- SSV-S : Sólidos suspensos voláteis na água na saída da estação
- SED-S : Quantidade de sedimentos na água na saída da estação
- COND-S : Condutividade da água na saída da estação
- RD-DBO-P : Performance da Demanda bioquímica de oxigênio da água no primeiro “decantador”
- RD-SS-P : Performance de Sólidos suspensos na água no primeiro “decantador”
- RD-SED-P : Performance da Quantidade de sedimentos na água no primeiro “decantador”
- RD-DBO-S : Performance da Demanda bioquímica de oxigênio da água no segundo “decantador”



As variáveis levantadas pelo banco

- RD-DQO-S : Performance da Demanda química de oxigênio da água no segundo “decantador”
- RD-DBO-G : Performance global da Demanda bioquímica de oxigênio da água na estação
- RD-DQO-G : Performance global da Demanda química de oxigênio da água na estação
- RD-SS-G : Performance global de Sólidos suspenso na água na estação
- RD-SED-G : Performance global de sedimentos da água na estação

Exemplo do nosso conjunto de dados

	Data	Q-E	ZN-E	PH-E	DBO-E	DQO-E	SS-E	SSV-E	SED-E	COND-E	PH-P	DBO-P	SS-P	SSV-P
1	D-1/3/90	44101	1.50	7.8	NA	407	166	66.3	4.5	2110	7.9	NA	228	70.2
2	D-2/3/90	39024	3.00	7.7	NA	443	214	69.2	6.5	2660	7.7	NA	244	75.4
3	D-4/3/90	32229	5.00	7.6	NA	528	186	69.9	3.4	1666	7.7	NA	220	72.7
4	D-5/3/90	35023	3.50	7.9	205	588	192	65.6	4.5	2430	7.8	236	268	73.1
5	D-6/3/90	36924	1.50	8.0	242	496	176	64.8	4.0	2110	7.9	NA	236	57.6
6	D-7/3/90	38572	3.00	7.8	202	372	186	68.8	4.5	1644	7.8	NA	248	66.1
7	D-8/3/90	41115	6.00	7.8	NA	552	262	64.1	5.0	1603	7.8	NA	320	67.5
8	D-9/3/90	36107	5.00	7.7	215	489	334	40.7	6.0	1613	7.6	NA	304	53.9
9	D-11/3/90	29156	2.50	7.7	206	451	194	69.1	4.5	1249	7.7	206	220	61.8
10	D-12/3/90	39246	2.00	7.8	172	506	200	69.0	5.0	1865	7.8	208	248	66.1

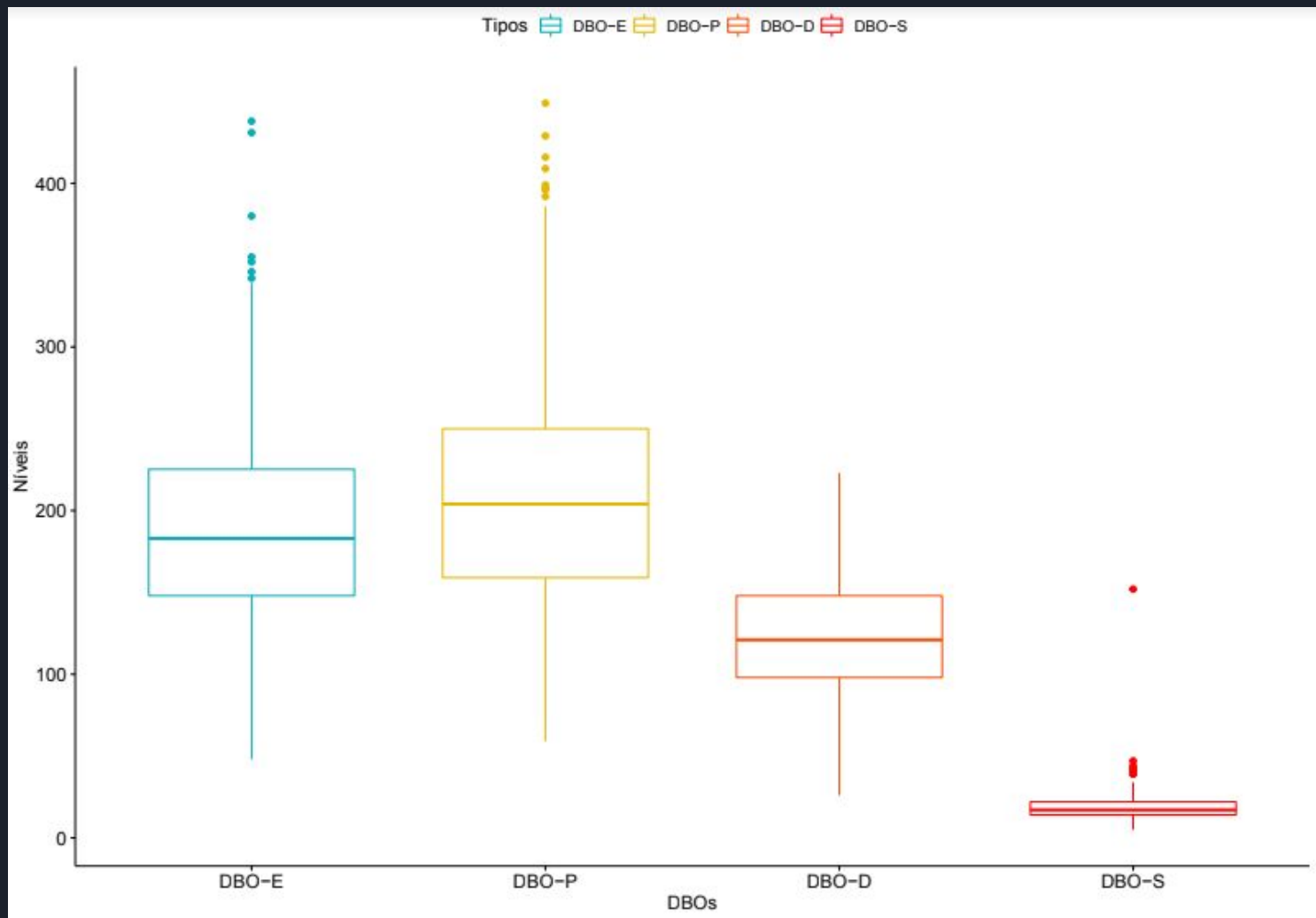
ANOVA

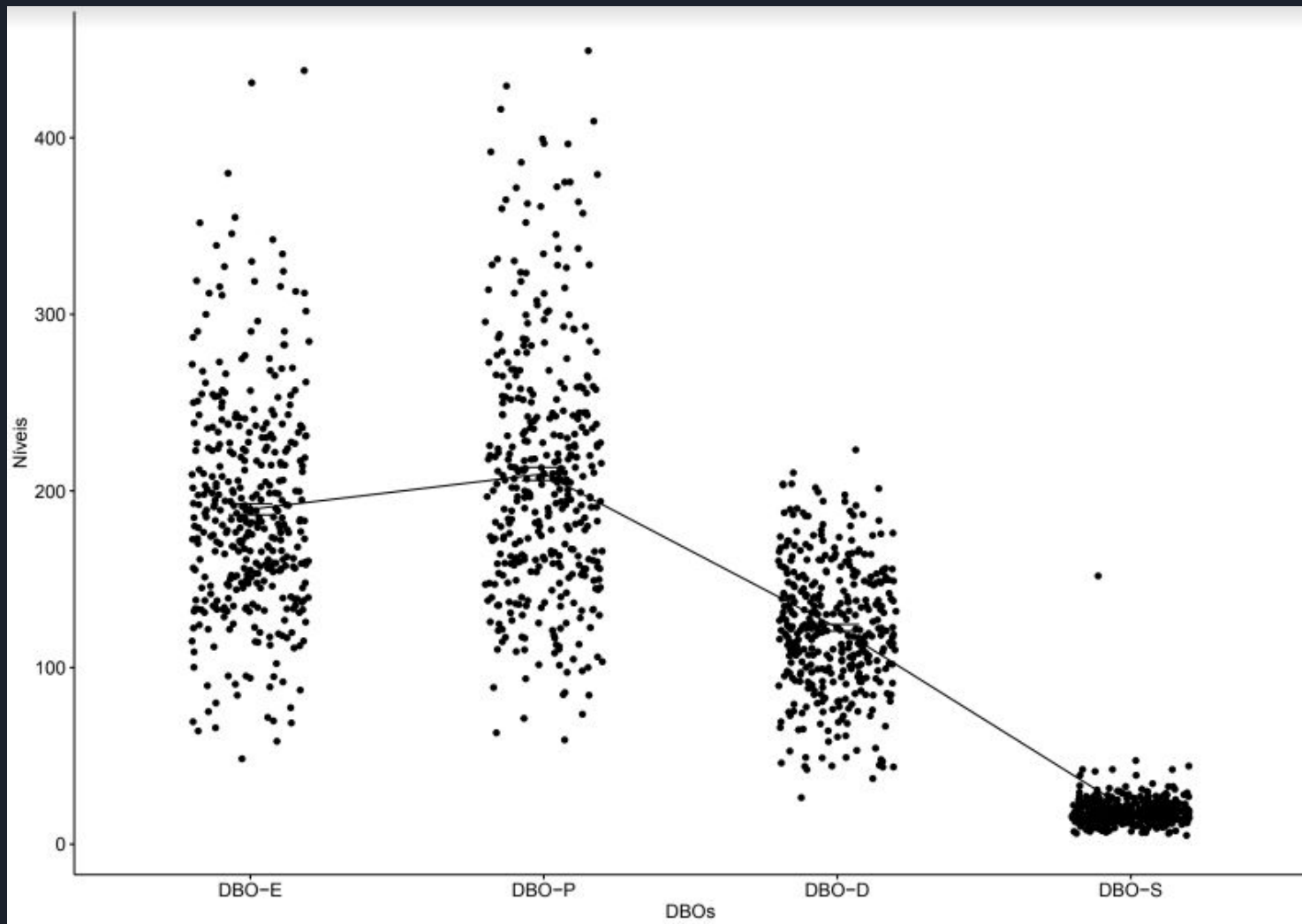





DBO

- Foi escolhida uma variável para ser analisada, no caso, o DBO
- Foram definidas as duas hipóteses:
 - H0- Todas as etapas do tratamento tem o mesmo nível de DBO;
 - H1- Existe diferença entre os 4 DBO's coletados desde a entrada, até a saída da estação de tratamento de água







	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Tipos	3	8447914	2815971	1087	<2e-16 ***
Residuals	1516	3928515	2591		

- p-valor pequeno consideravelmente pequeno, o que rejeita a hipótese nula

Análise de Regressão Simples



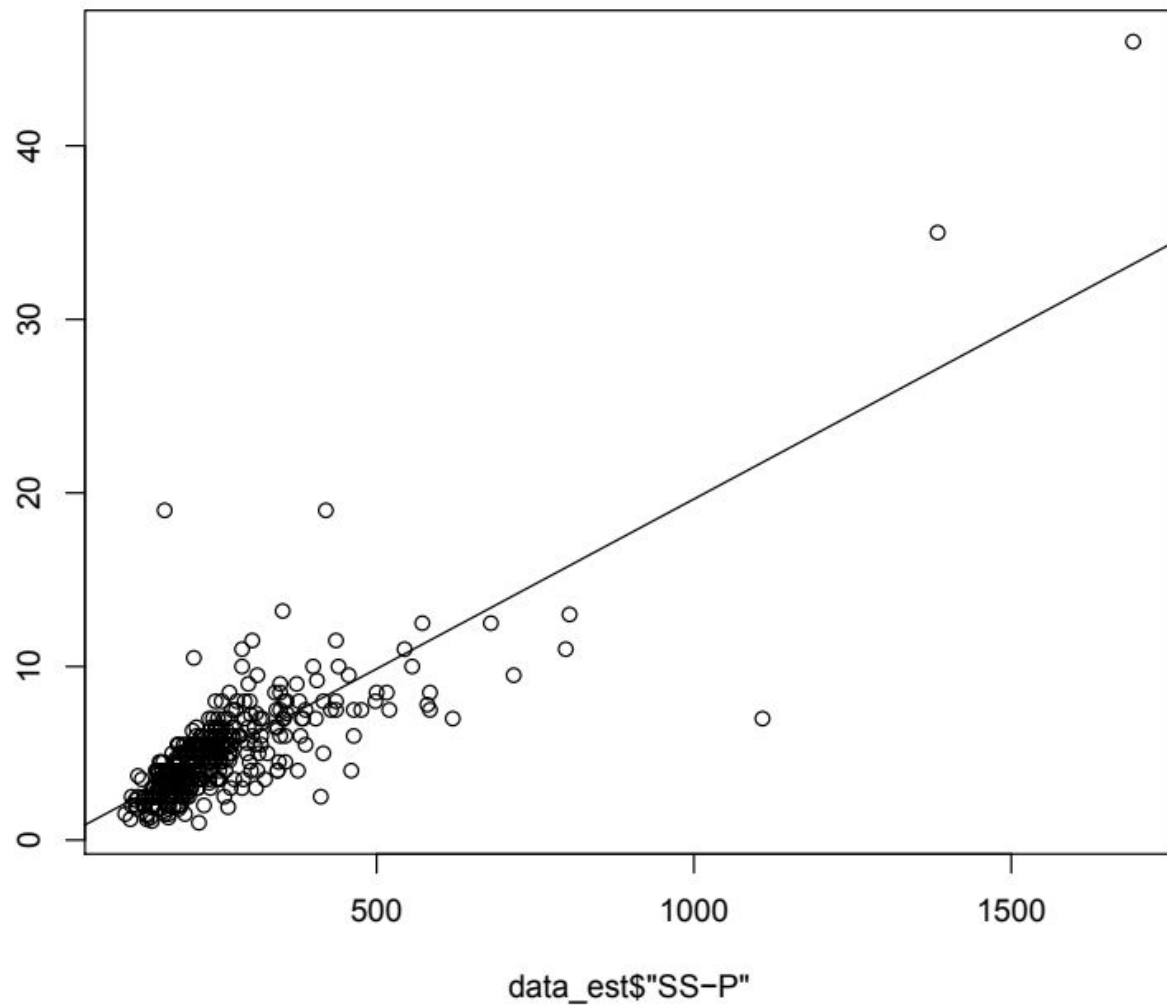


SS-P e SED-P

- Supondo que haja relação entre *Sólidos suspensos na água no primeiro “decantador”* e *Quantidade de sedimentos na água no primeiro “decantador”*.
- Foram definidas as duas hipóteses:
 - H0: não existe correlação entre sólidos suspensos e quantidade de sedimentos na água
 - H1: existe correlação entre sólidos suspensos e quantidade de sedimentos na água



data_est\$"SED-P"



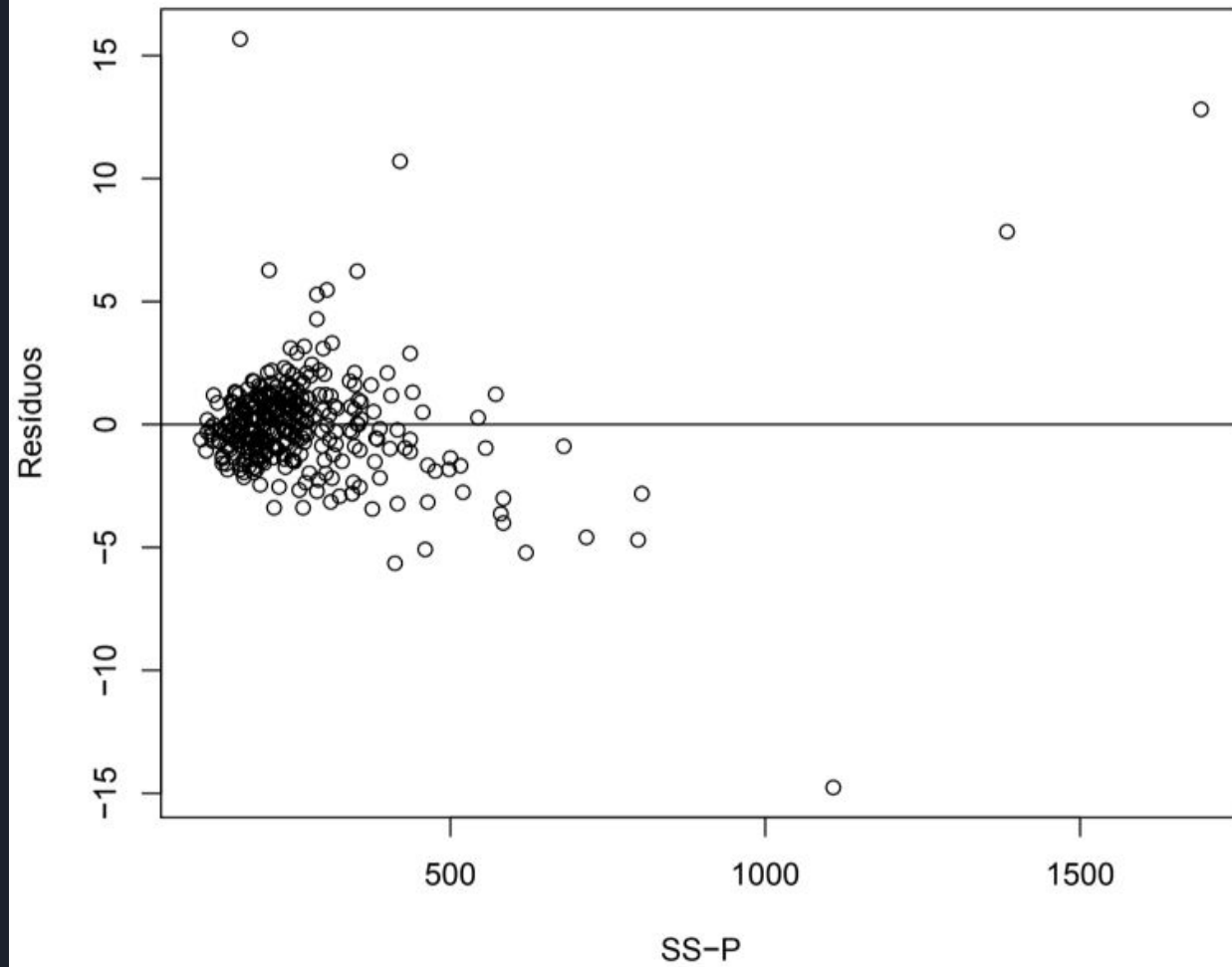


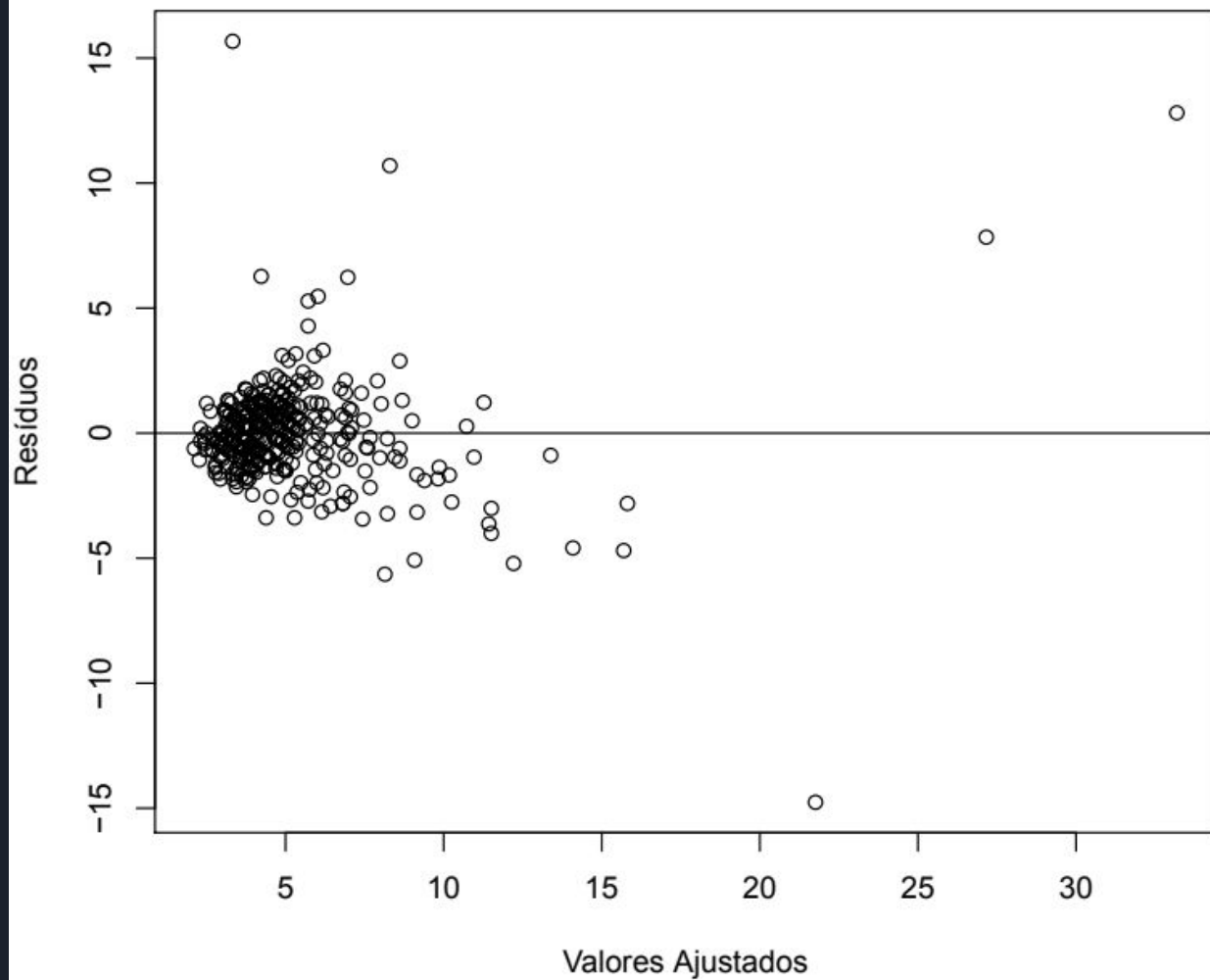
Pearson's product-moment correlation

data: data_est\$"SS-P" and data_est\$"SED-P"
t = 26.455, df = 378, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.7674074 0.8384337
sample estimates:
cor
0.8057995

F test to compare two variances

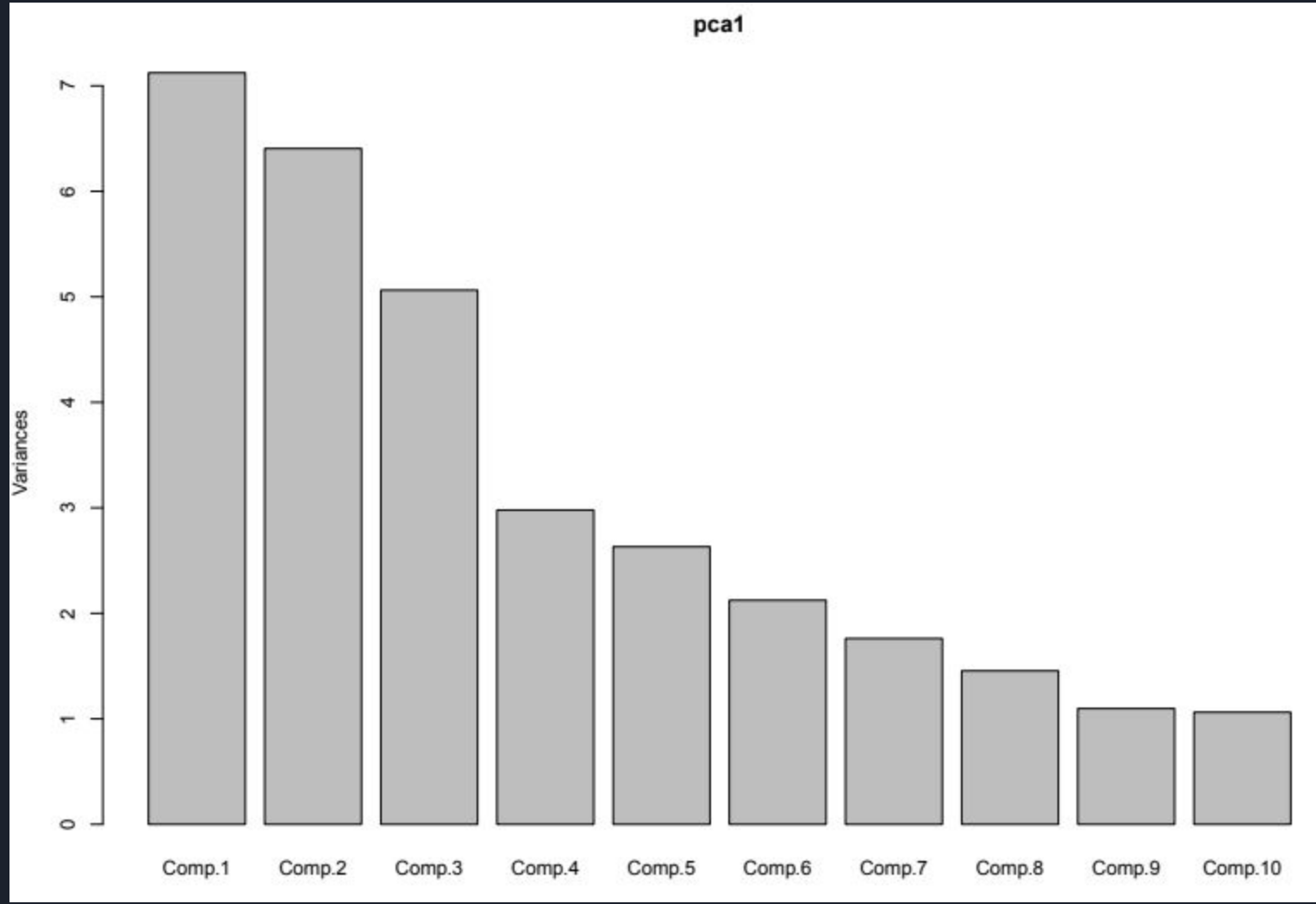
data: residuals(ajuste)[data_est\$"SS-P" > median(data_est\$"SS-P")] and
residuals(ajuste)[data_est\$"SS-P" < median(data_est\$"SS-P")]
F = 2.681, num df = 185, denom df = 188, p-value = 3.994e-11
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
2.010604 3.576153
sample estimates:
ratio of variances
2.681007



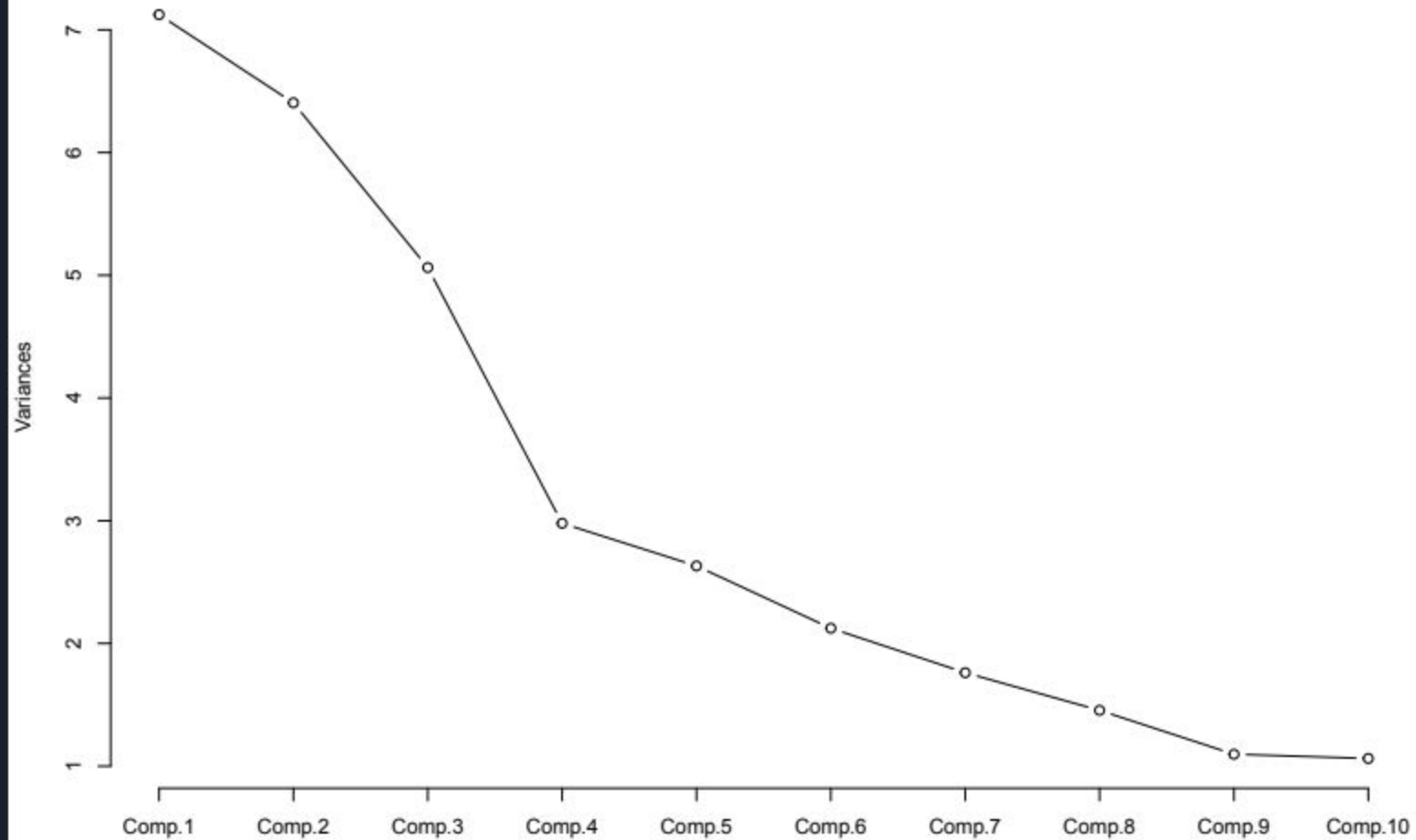


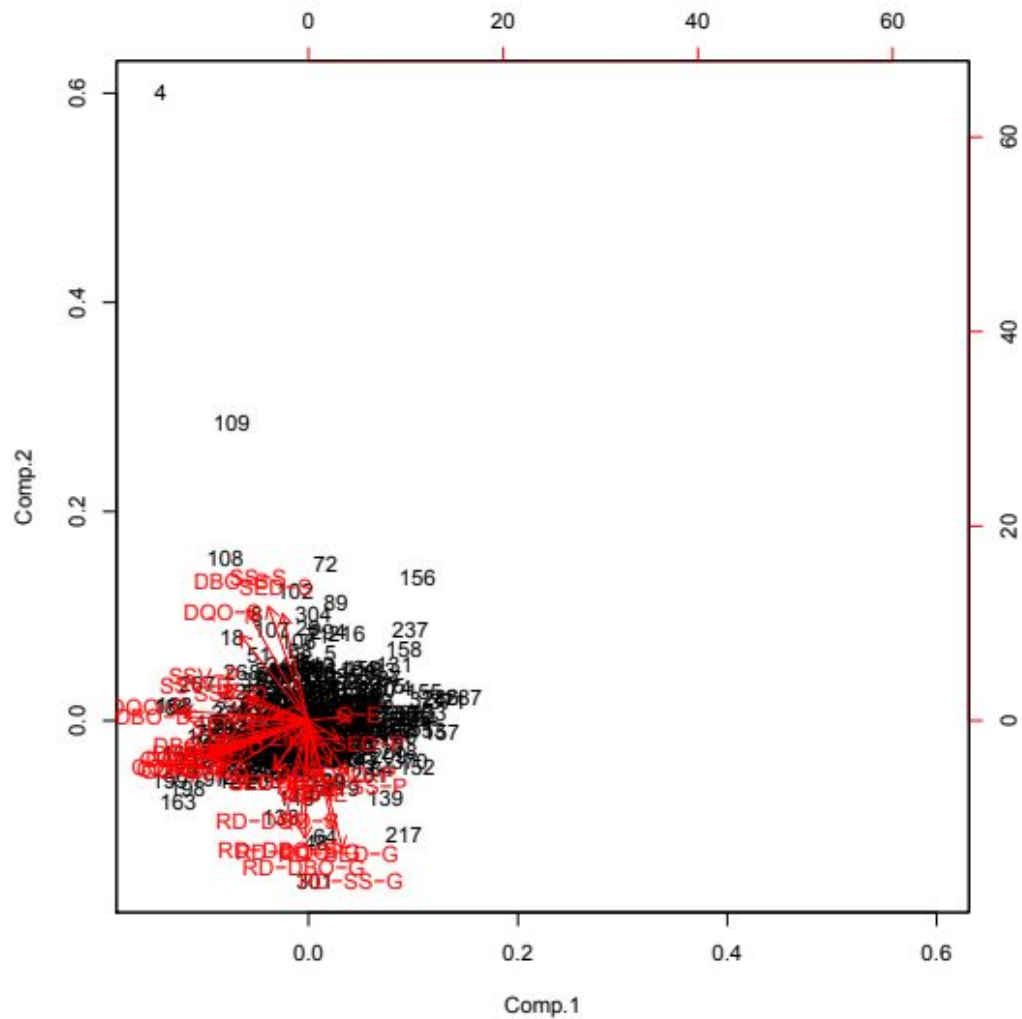
Análise de Componentes Principais

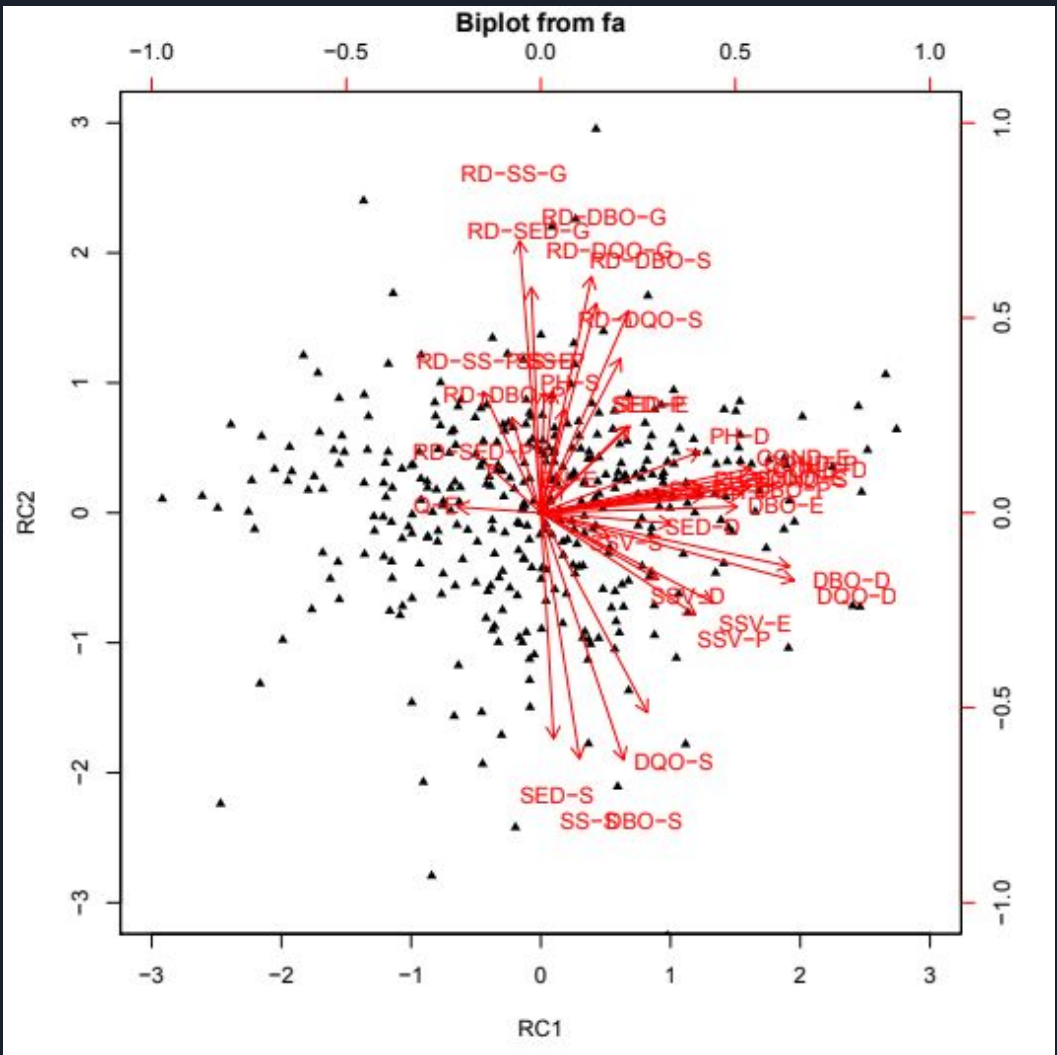




Scree Plot





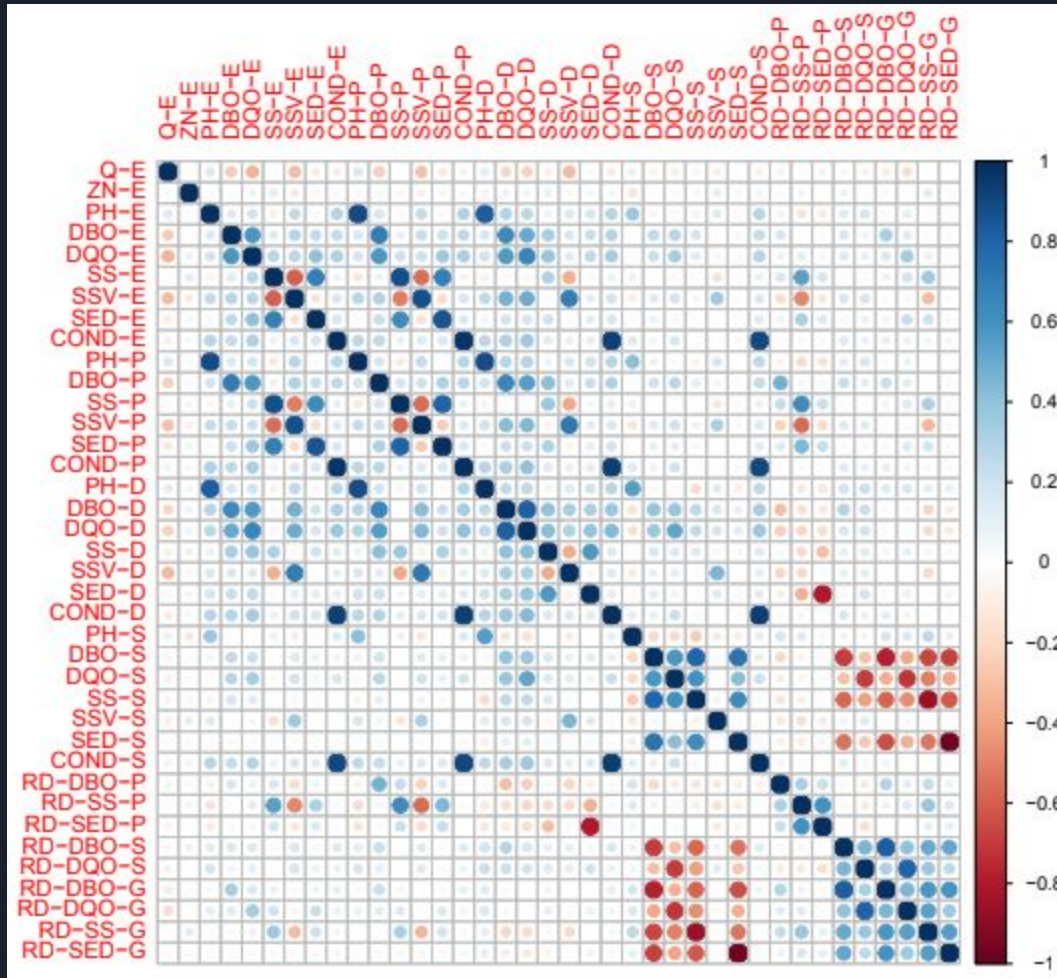


Análise Fatorial



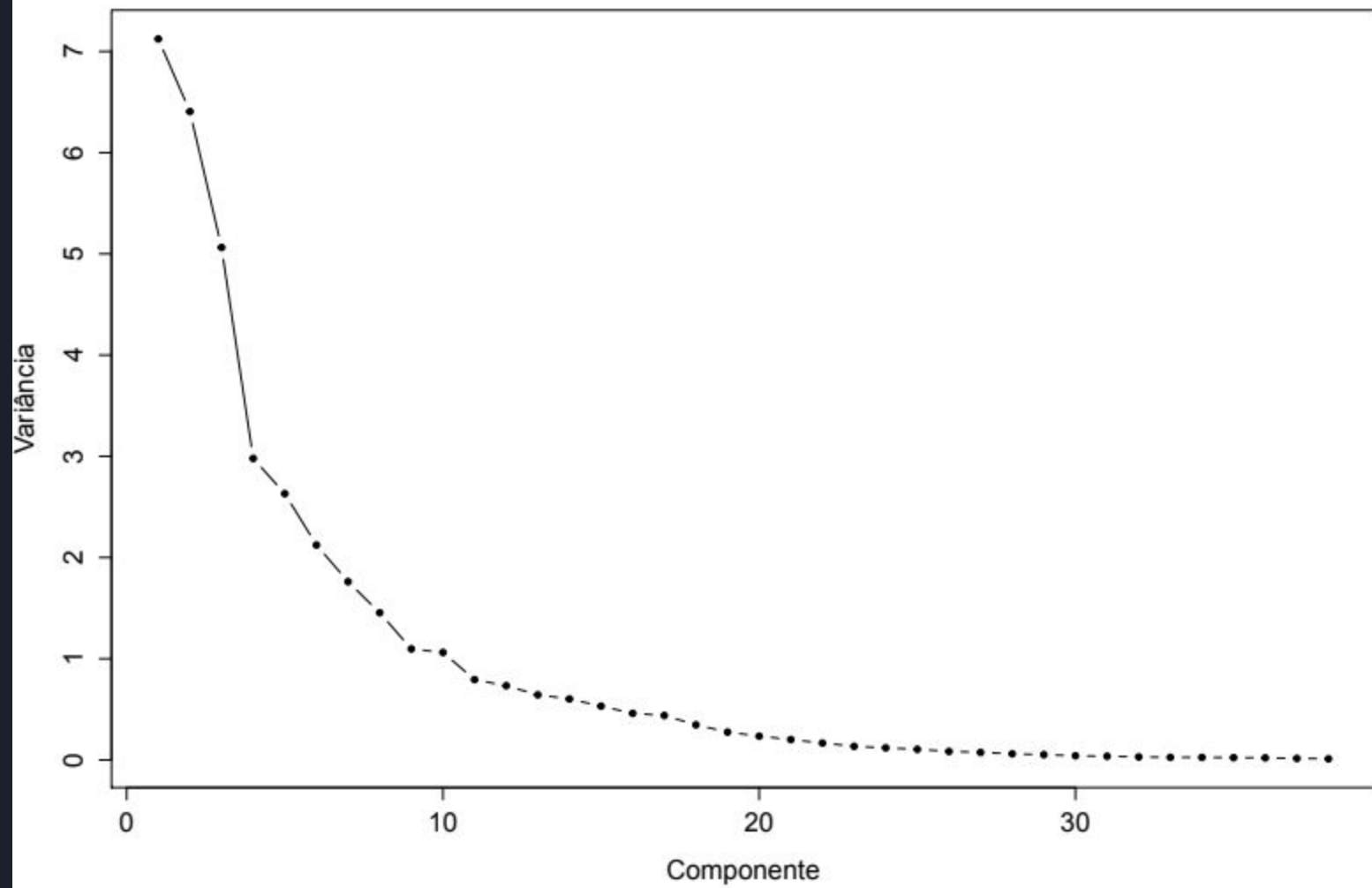
Análise Fatorial Exploratória

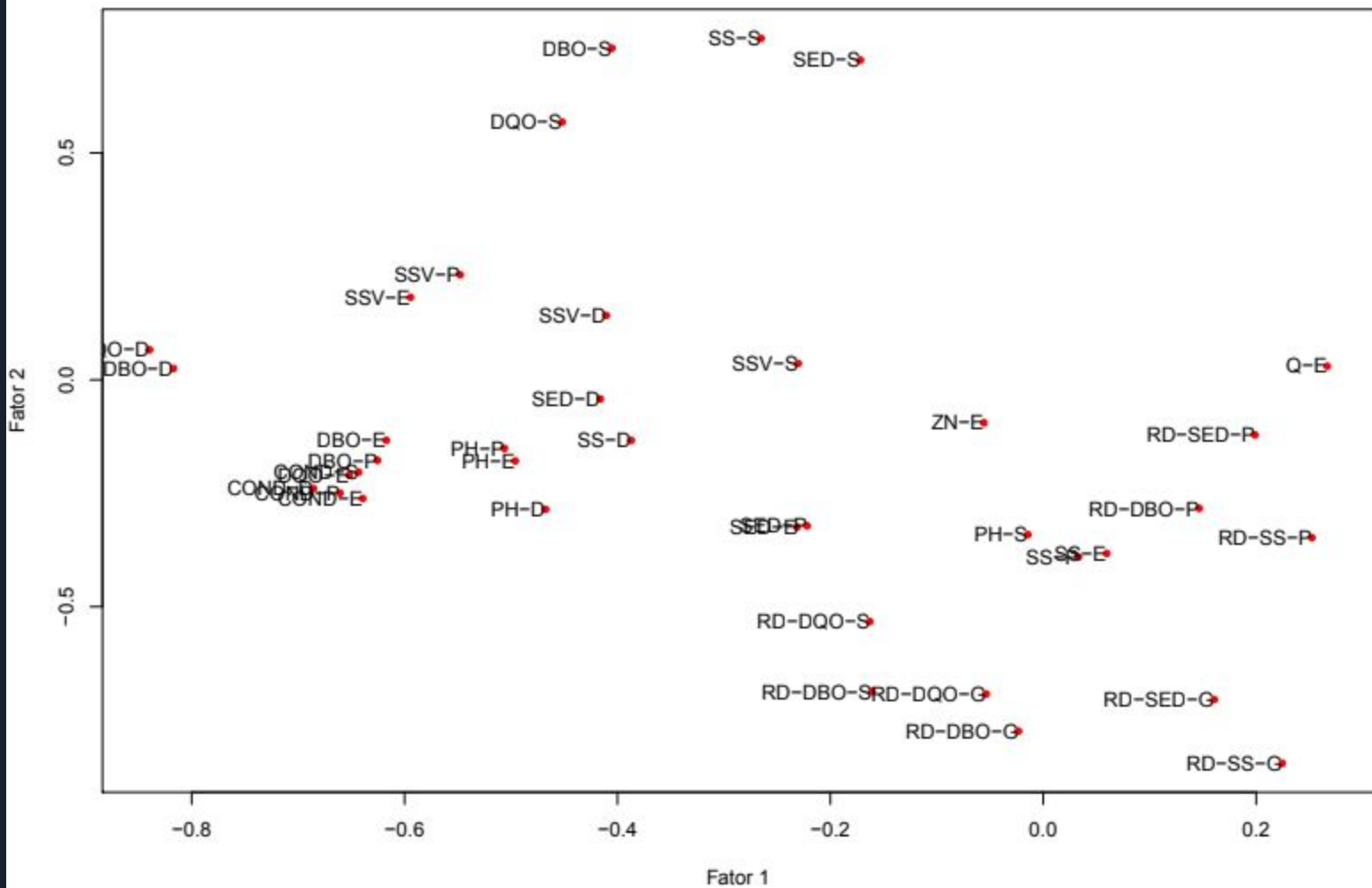




p-valor < $2,2e^{-16}$
(Teste de esfericidade
de Bartlett)

KMO: 0,7078904



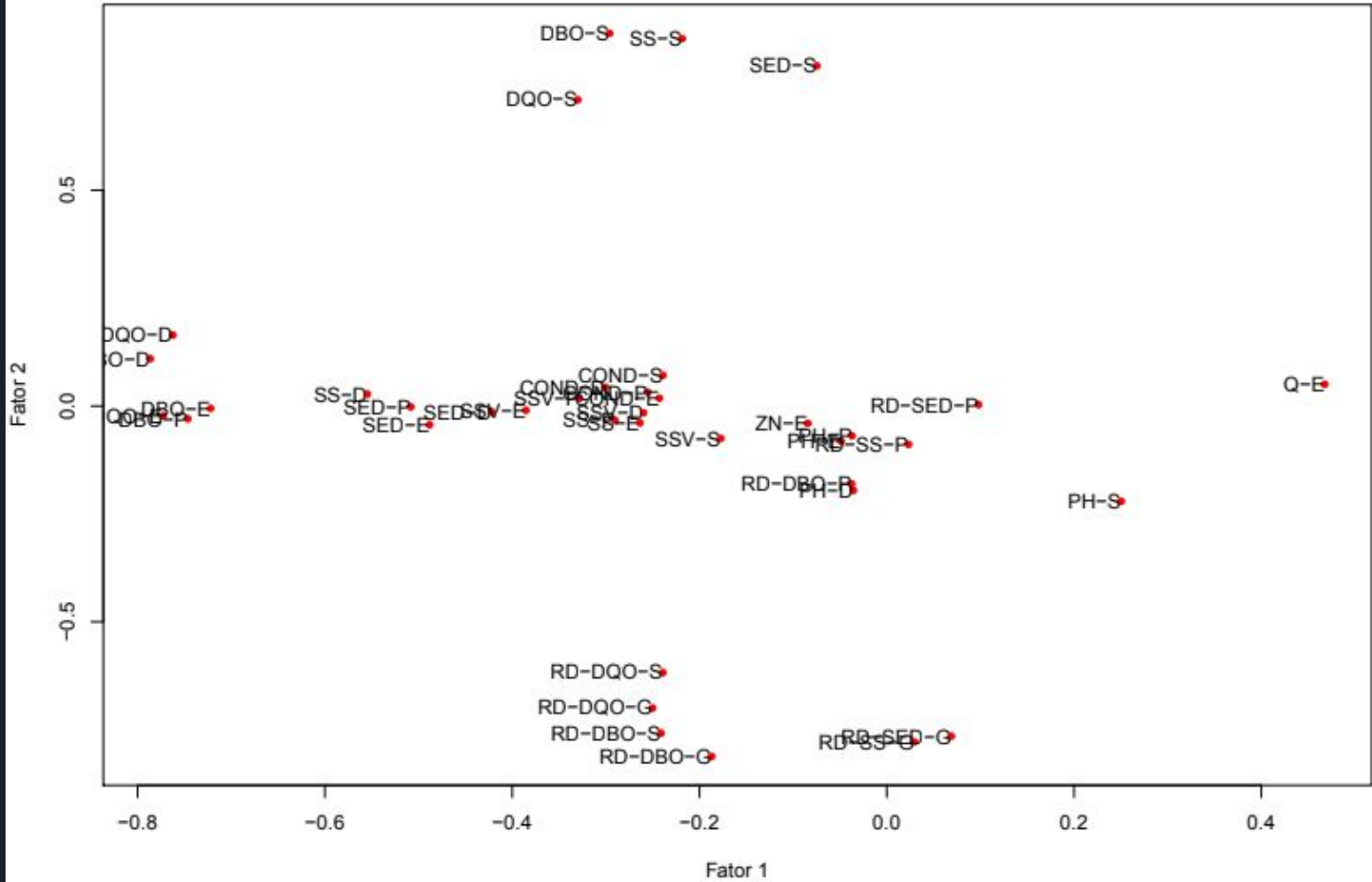




Após aplicar rotação
VARIMAX

Grupos:

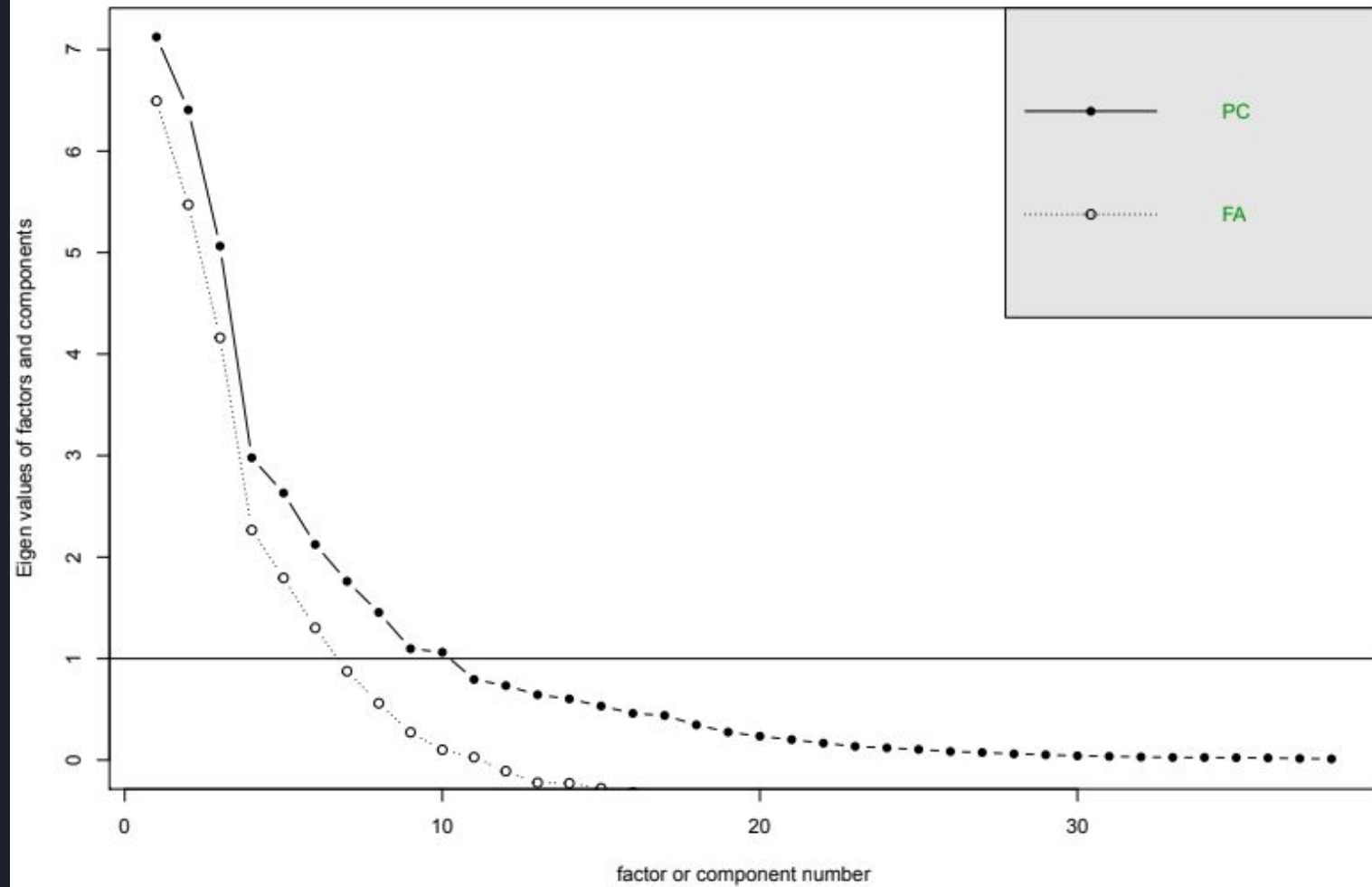
- Performace
- Saída
- Intermediária
- Demanda



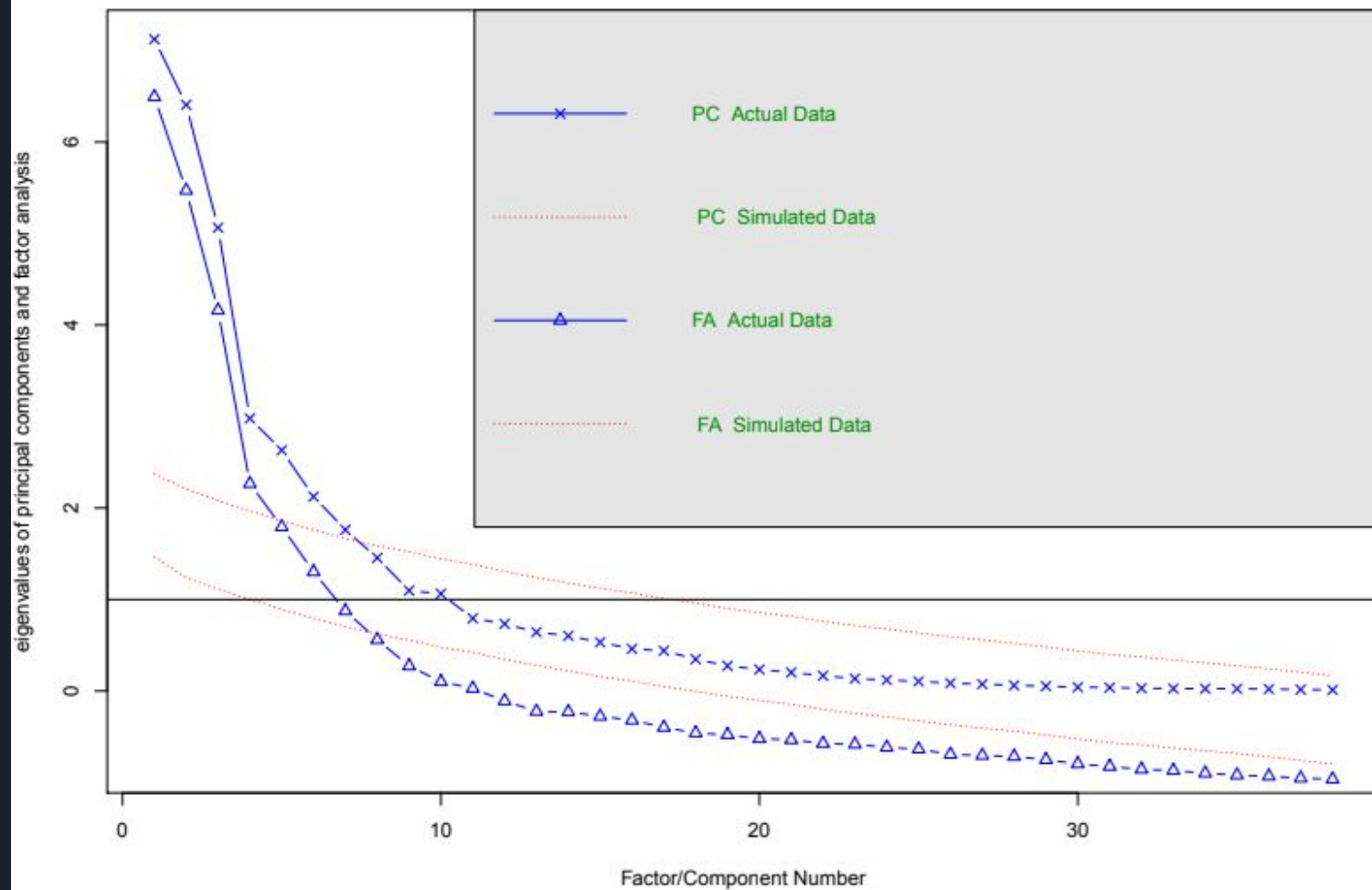
Análise Fatorial Confirmatória



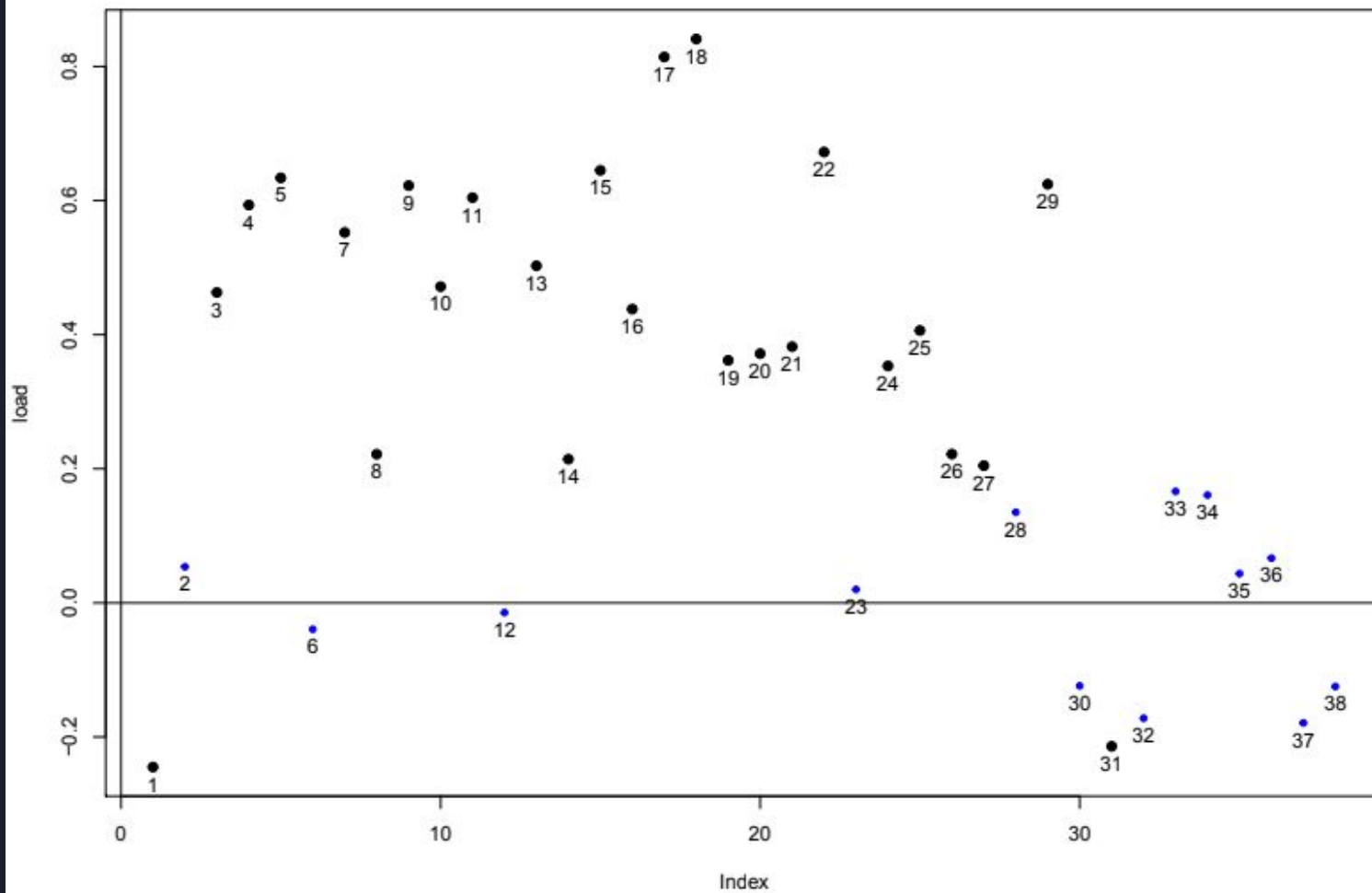
Scree plot

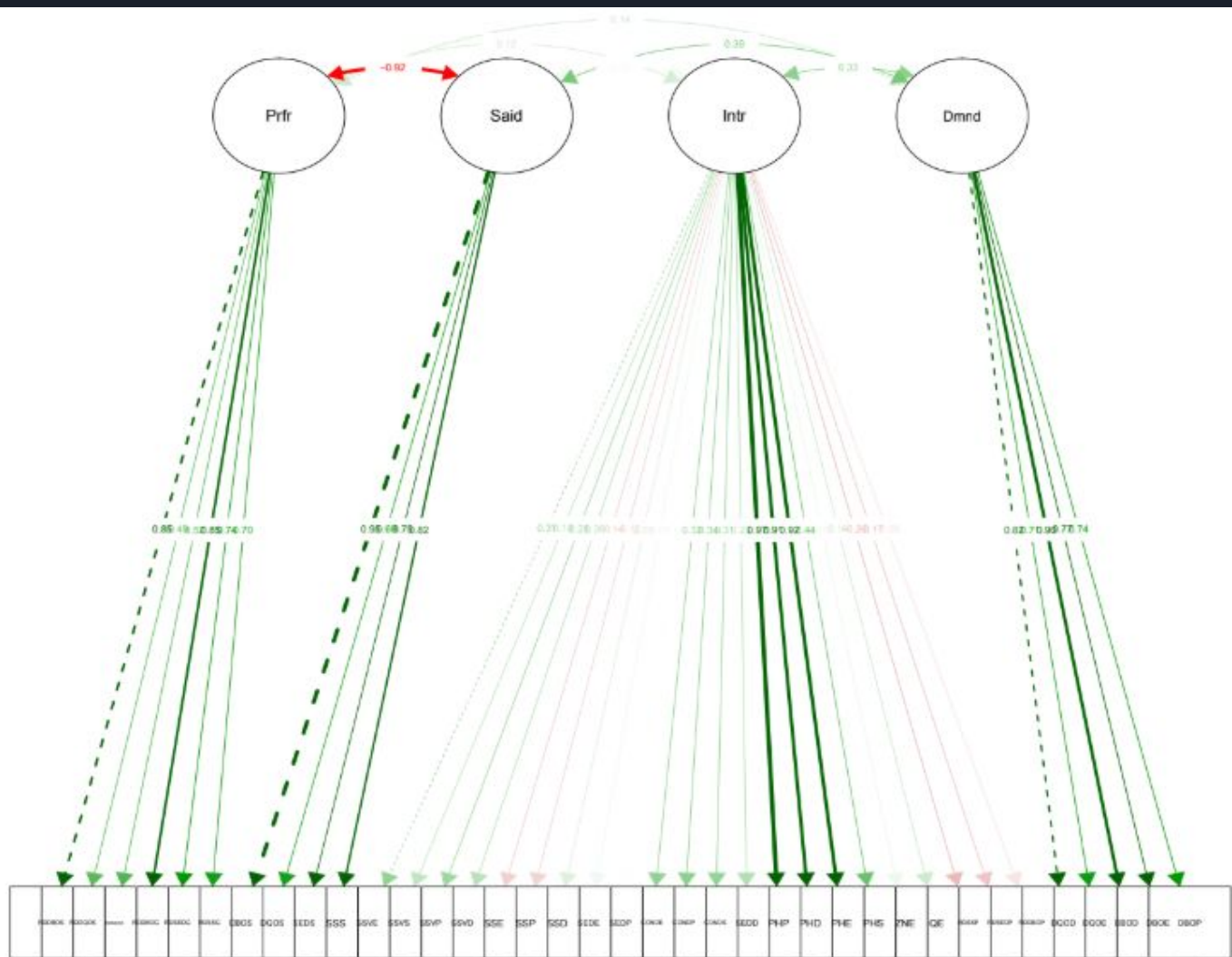


Parallel Analysis Scree Plots



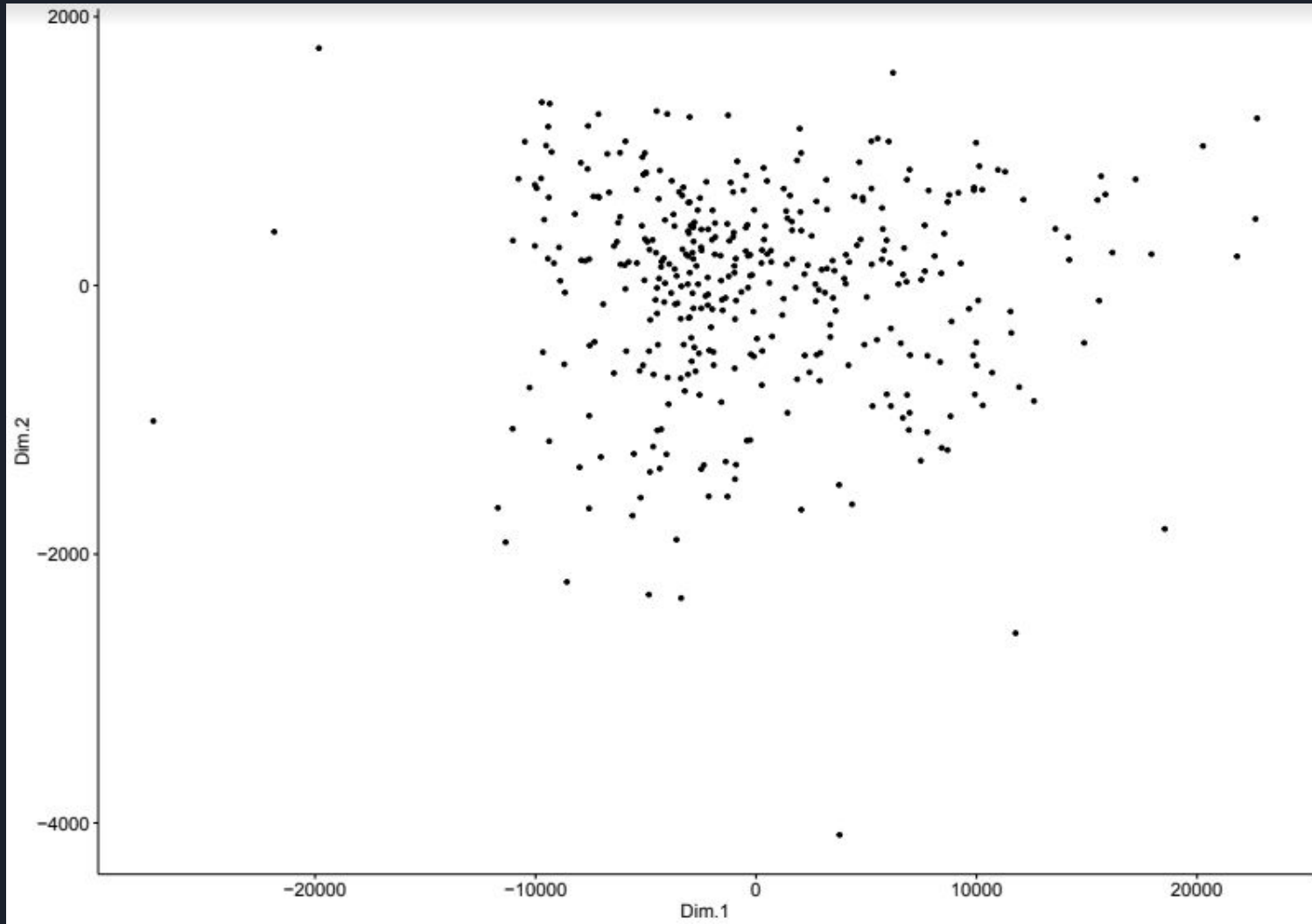
Factor Analysis

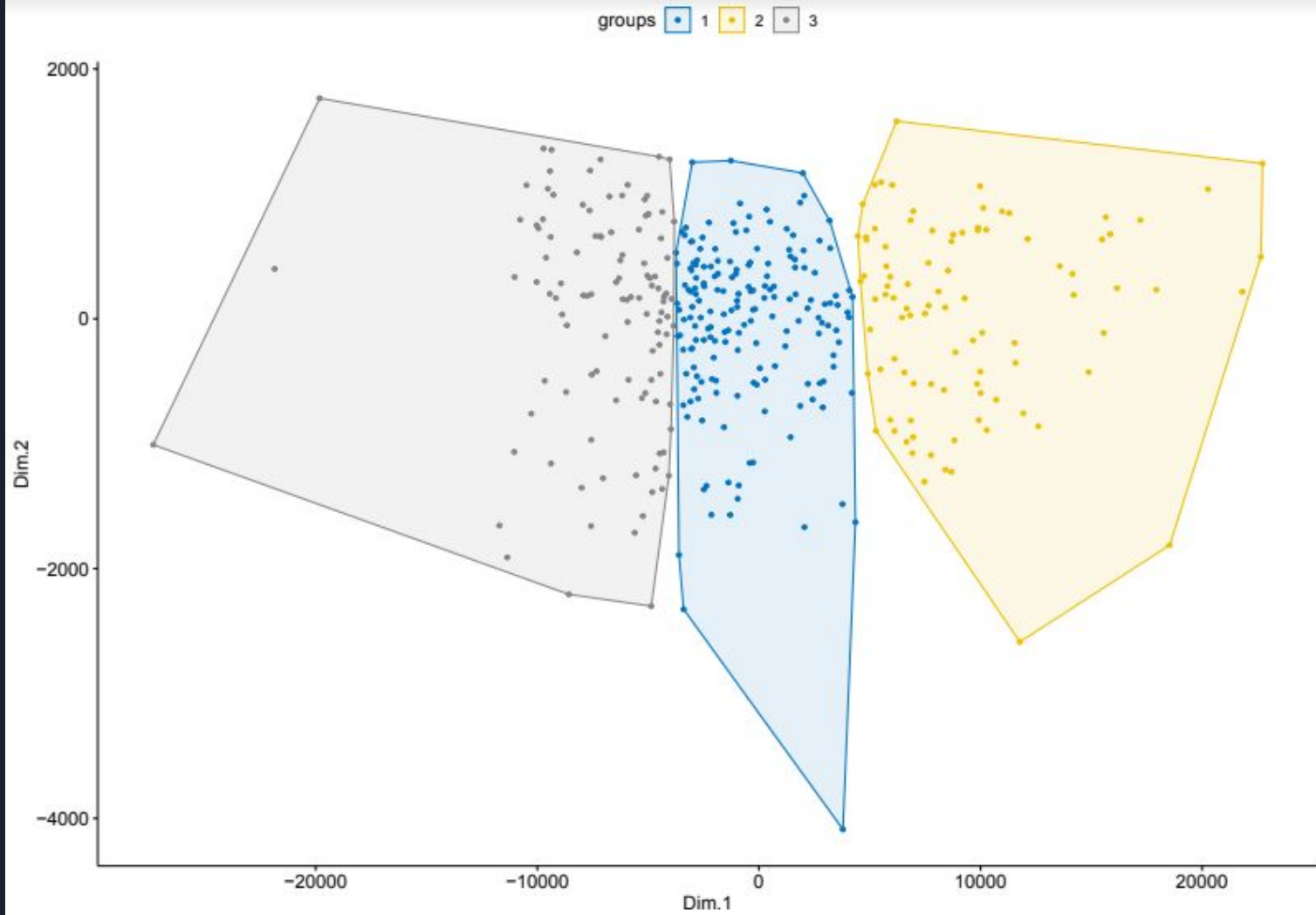




Escalonamento Multidimensional



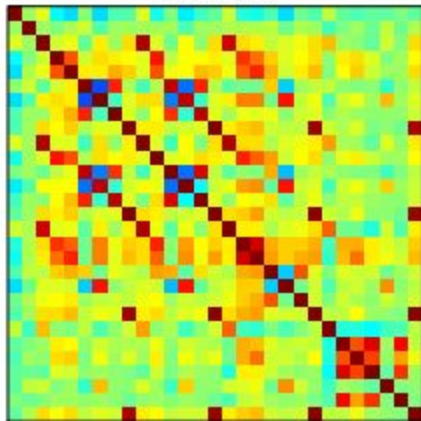




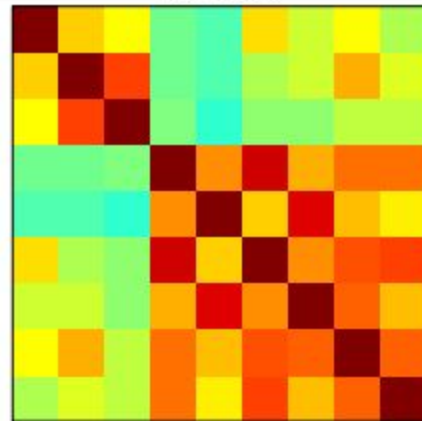
Correlação Canônica



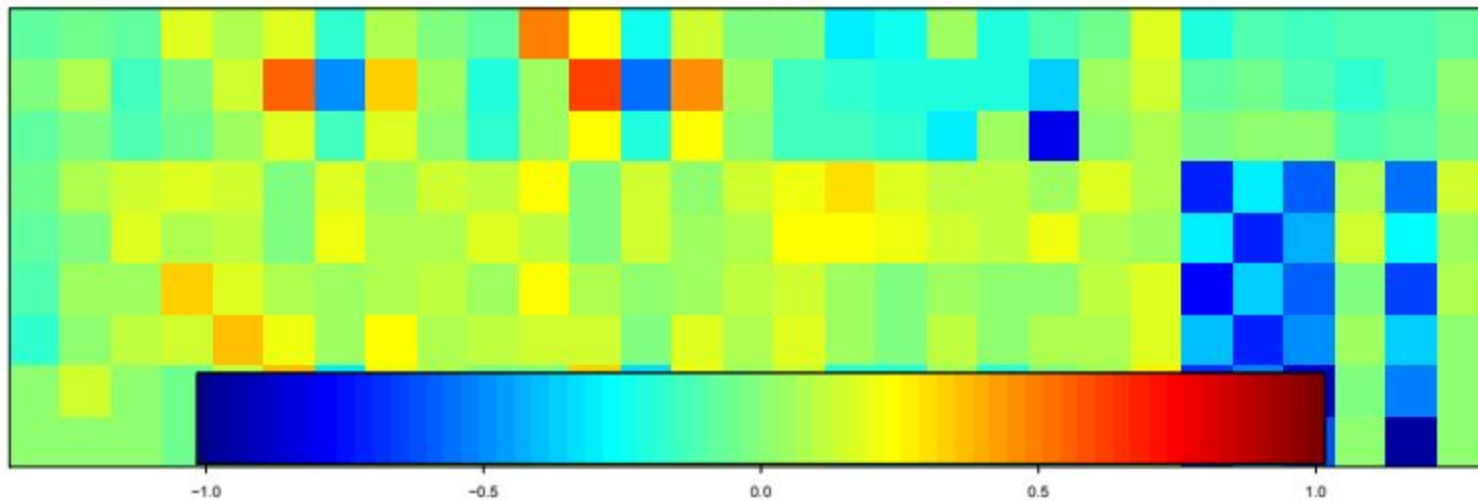
X correlation

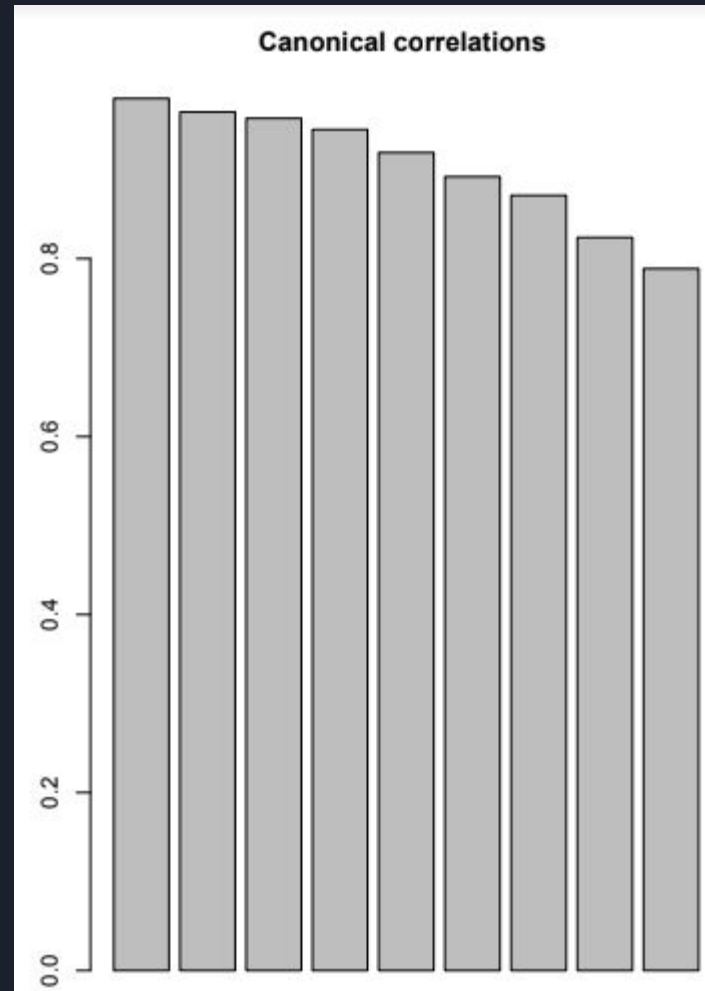


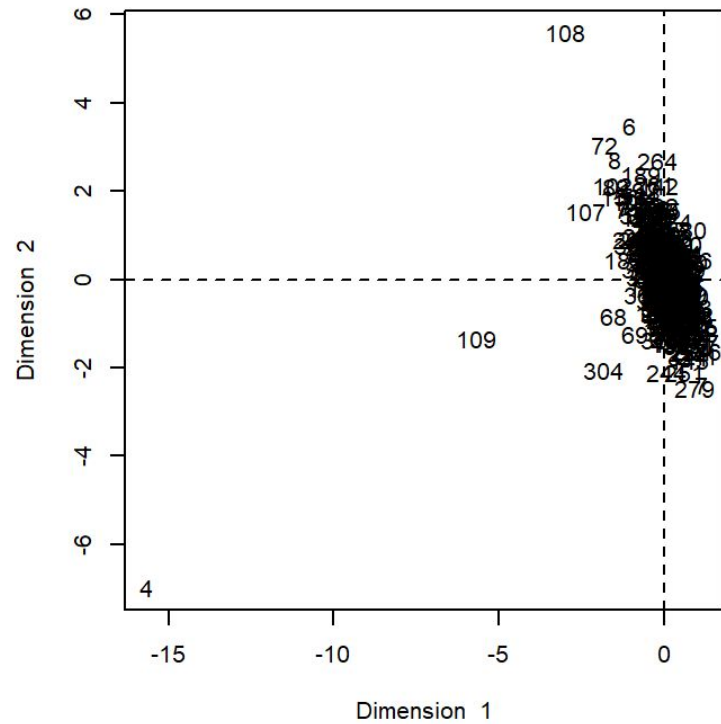
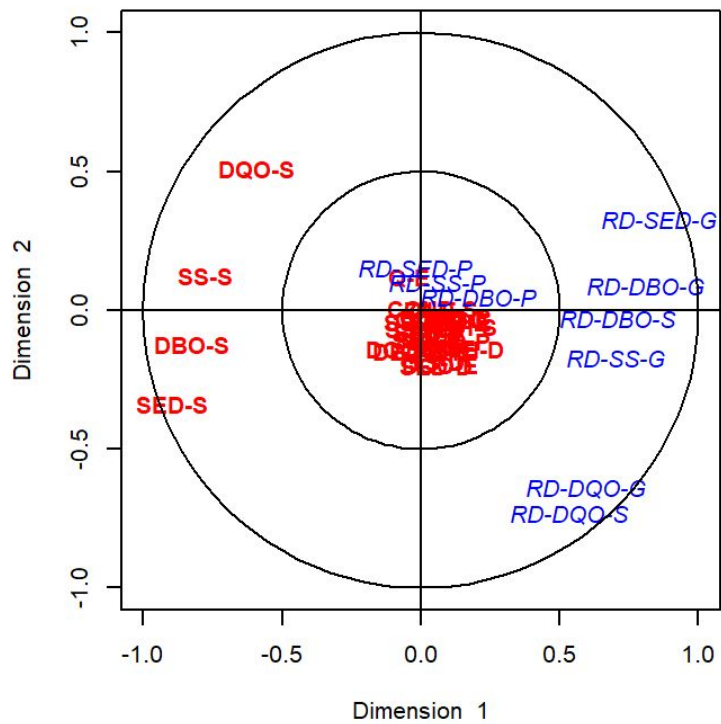
Y correlation

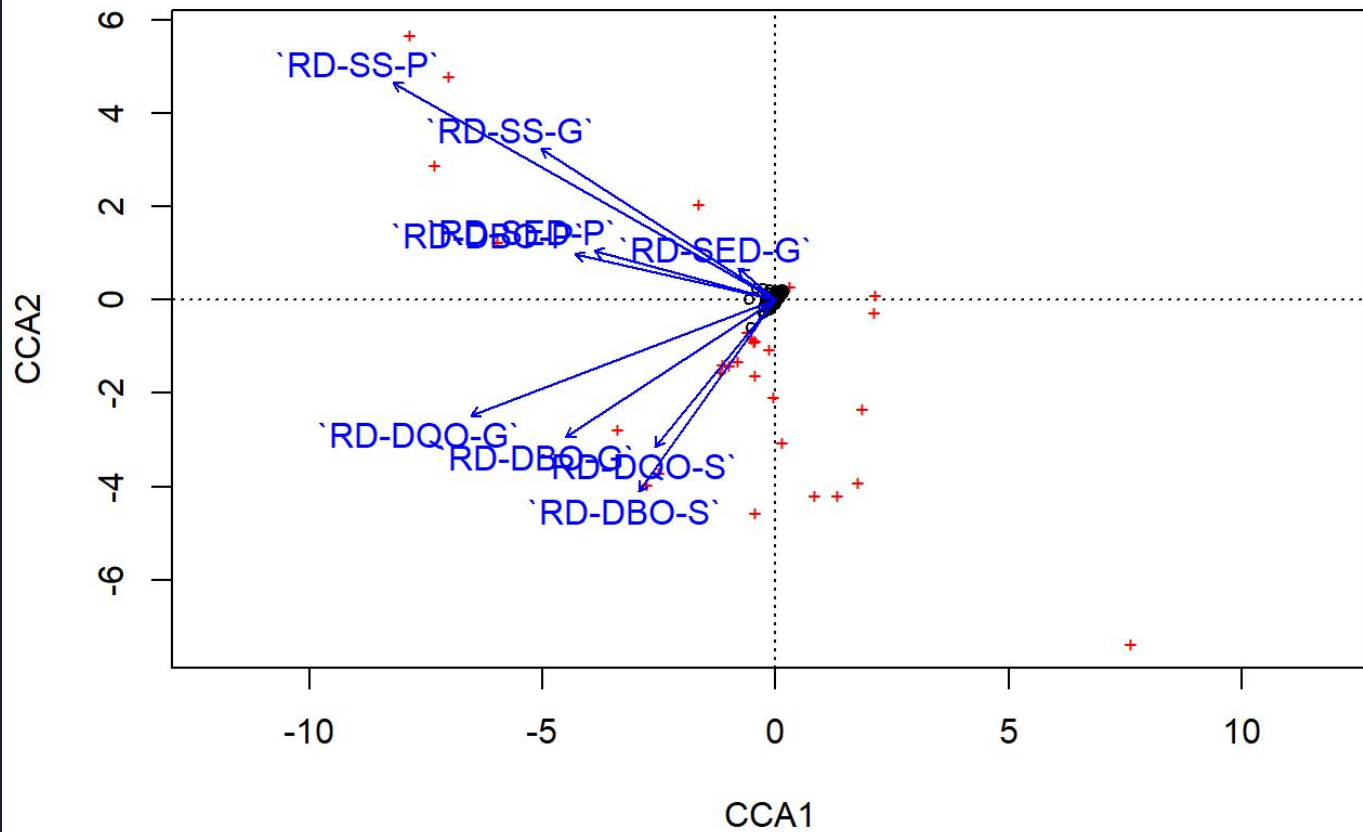


Cross-correlation










Co-clustering





An object of class Biclust

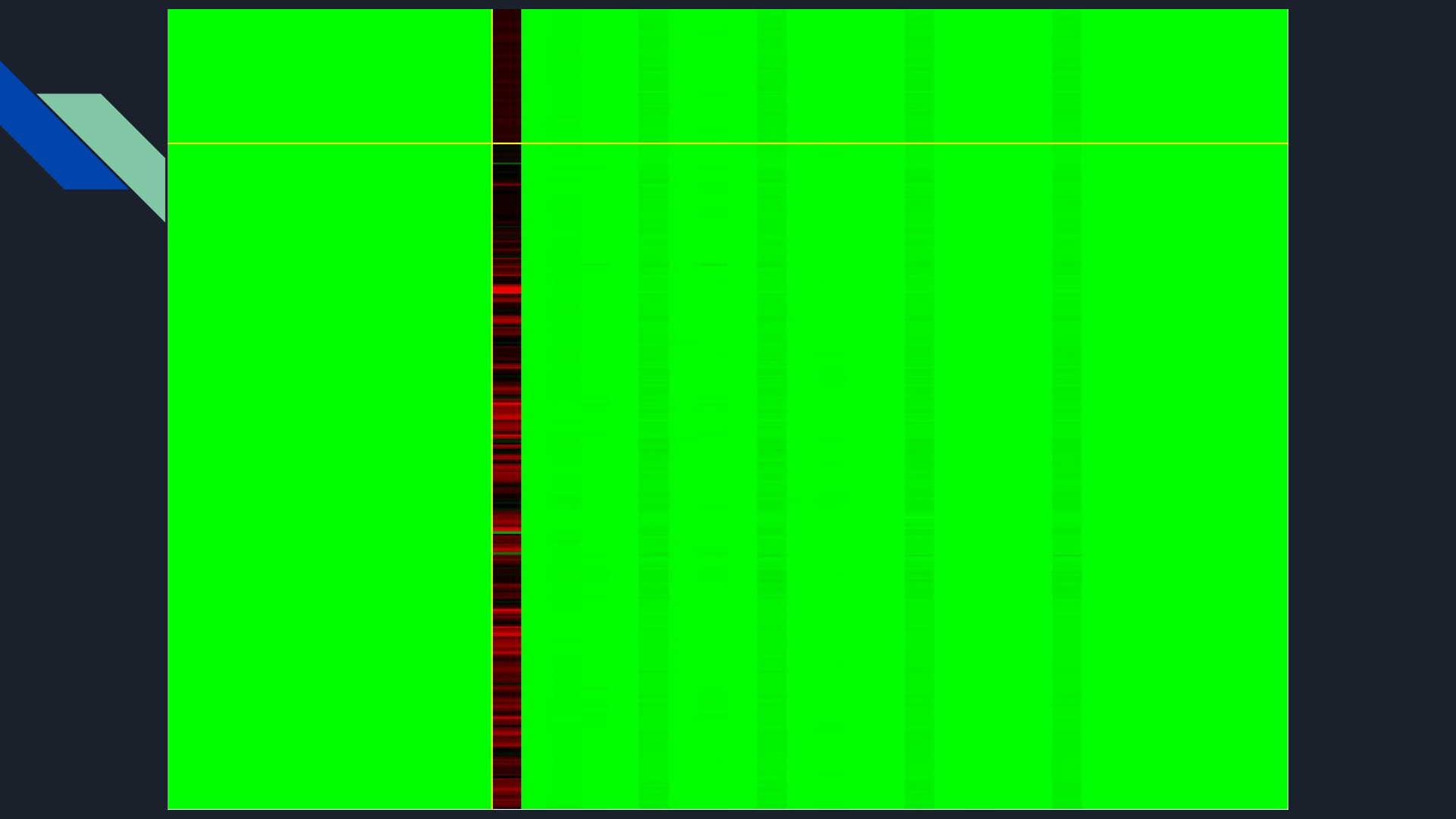
call:

```
biclust(x = x, method = method)
```

Number of Clusters found: 9

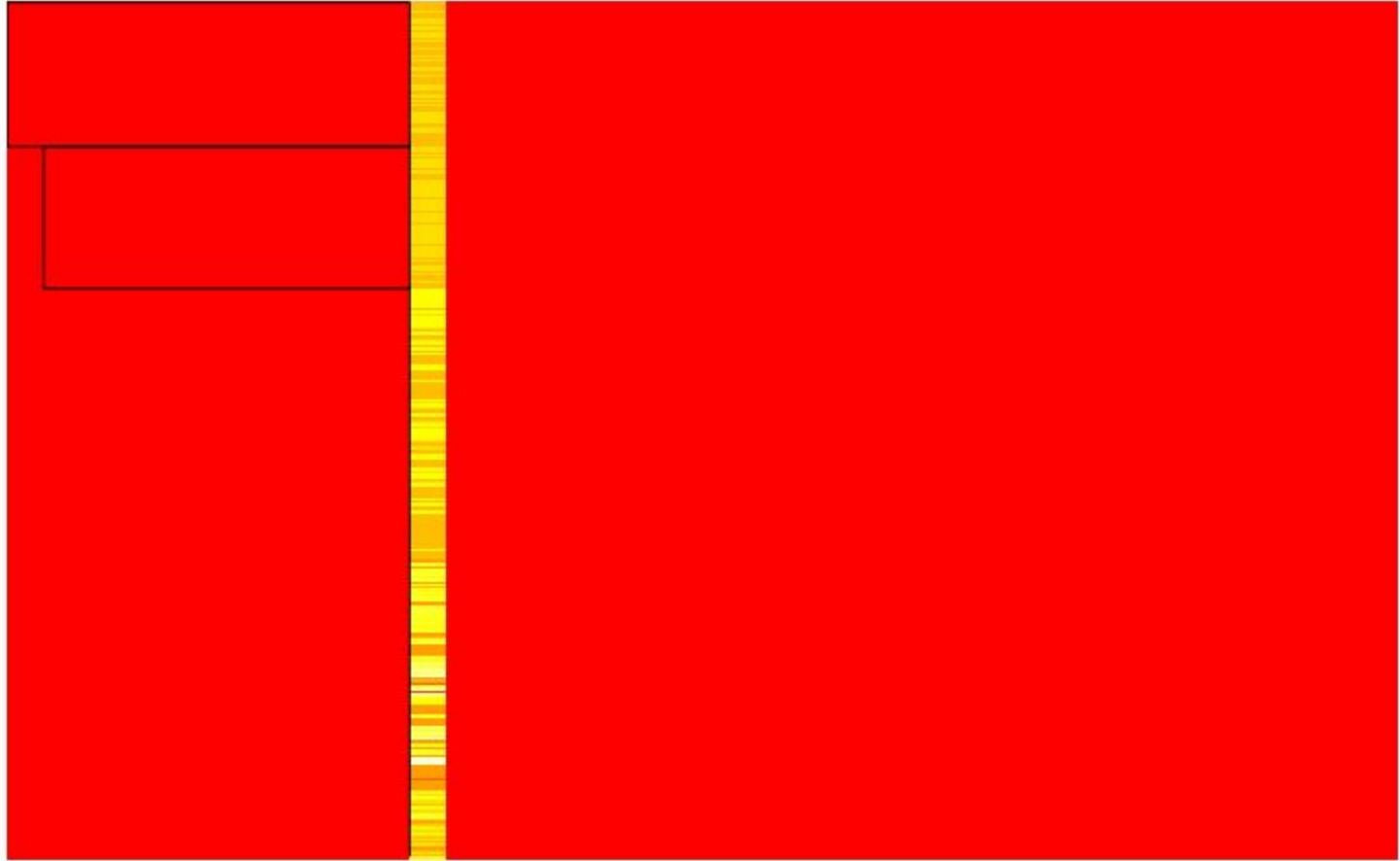
First 5 Cluster sizes:

	BC 1	BC 2	BC 3	BC 4	BC 5
Number of Rows:	64	63	49	71	42
Number of Columns:	11	10	10	10	10






1:length(bicRows)



1:length(bicCols)

Observações finais



- 
- Random Forest
 - Não foi possível aplicar a metodologia pois não possível criar um conjunto de dados que pudesse ser treinado devido à falta de medições das variáveis preditoras que deveriam ser usadas para análise. Impedindo analisar qualquer outra variável.
 - Regressão Logística e Análise Discriminante
 - Não aplicamos a metodologia destas análises, pois não conseguimos identificar “variável-resposta” para cada tipo de metodologia
 - Regressão Múltipla e Teste de Aderência
 - Não aplicamos as metodologias destas análises, porque não conseguimos identificar como utilizar sobre nosso banco de dados