Análise da relação de conectividade com as notas do ENEM

Equipe: TL2F

Membros:

- Felipe Fontoura de Moraes
- Felippe Vernizze Sousa Nadolny

Relatório

Neste trabalho, buscou-se entender a relação entre a disponibilidade de conectividade de internet aos alunos participantes do Exame Nacional do Ensino Médio - Enem e seu impacto no desempenho na prova, antes e depois da pandemia de covid19. Considerou-se a hipótese de que os alunos com acesso à conectividade tiveram melhor desempenho, especialmente no período de pandemia.

Inicialmente o objetivo era mais amplo, abrangendo por exemplo como a relação variava entre municípios ou estados, e também considerando dados de fontes adicionais, como por exemplo os da Agência Nacional de Telecomunicações - Anatel. Também se esperava fazer uma análise mais geral da conectividade com a educação levando em conta também o Exame Nacional de Desempenho de Estudantes - Enade. Entretanto, devido a diversas dificuldades no decorrer do trabalho, foi necessário limitar o escopo.

Quanto ao impacto do estudo, considerou-se que, caso comprovada a correlação, isto poderia balizar as políticas públicas educacionais e de telecomunicações. Por exemplo, subsidiar infraestruturas de internet nos municípios com baixo índices educacionais, ou programas sociais para fornecer acesso à internet para alunos de baixa renda.

Na análise exploratória dos dados, identificamos que houve um aumento das faltas em 2020, que foi o ano do auge da pandemia de covid19.

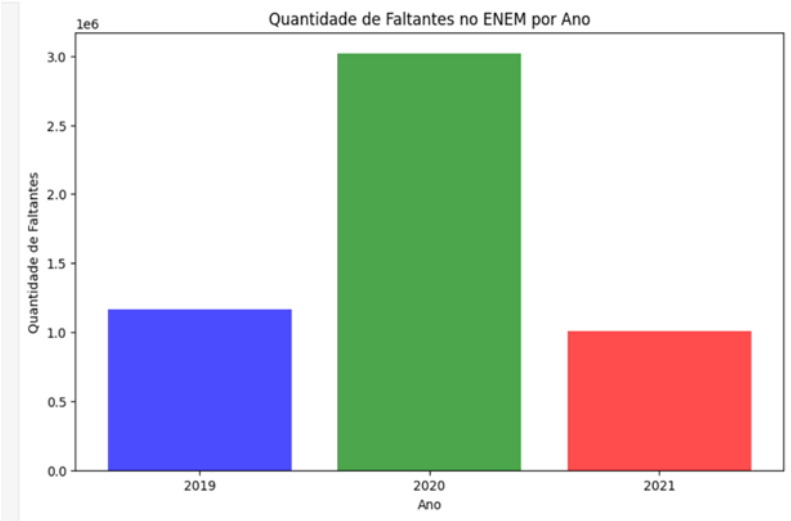

Figura 1 - Faltantes no Enem por ano

Também se identificou uma variação na distribuição da conectividade à internet nos estados, como se observa no gráfico abaixo, onde se observa que em todos os estados, mais da metade dos participantes possui acesso à internet.

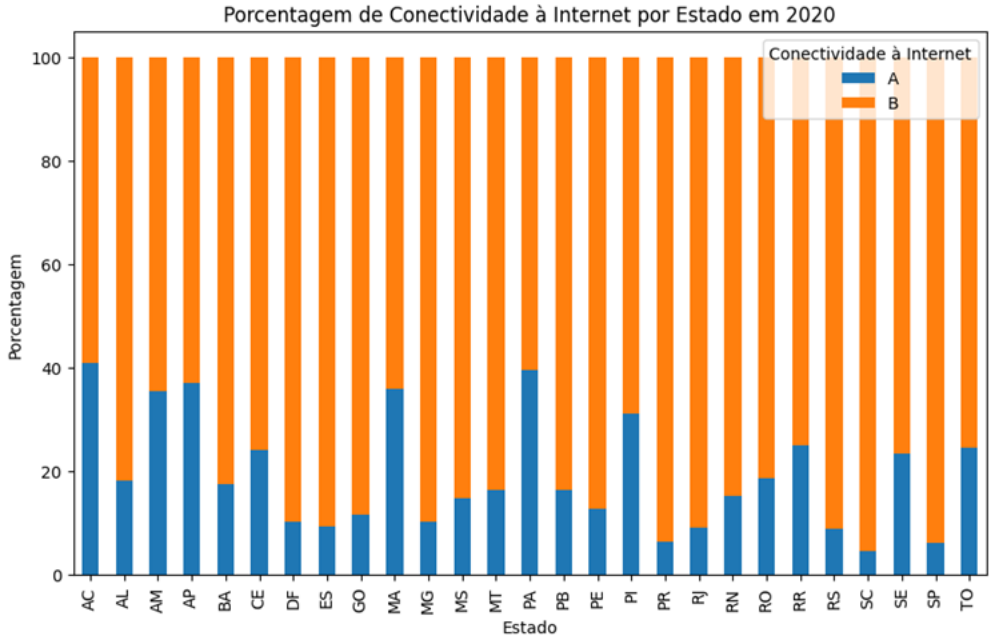

Figura 2 - Proporção de conectividade em cada estado

No decorrer da análise, observamos diferença nas médias dos candidatos com e sem acesso à internet.

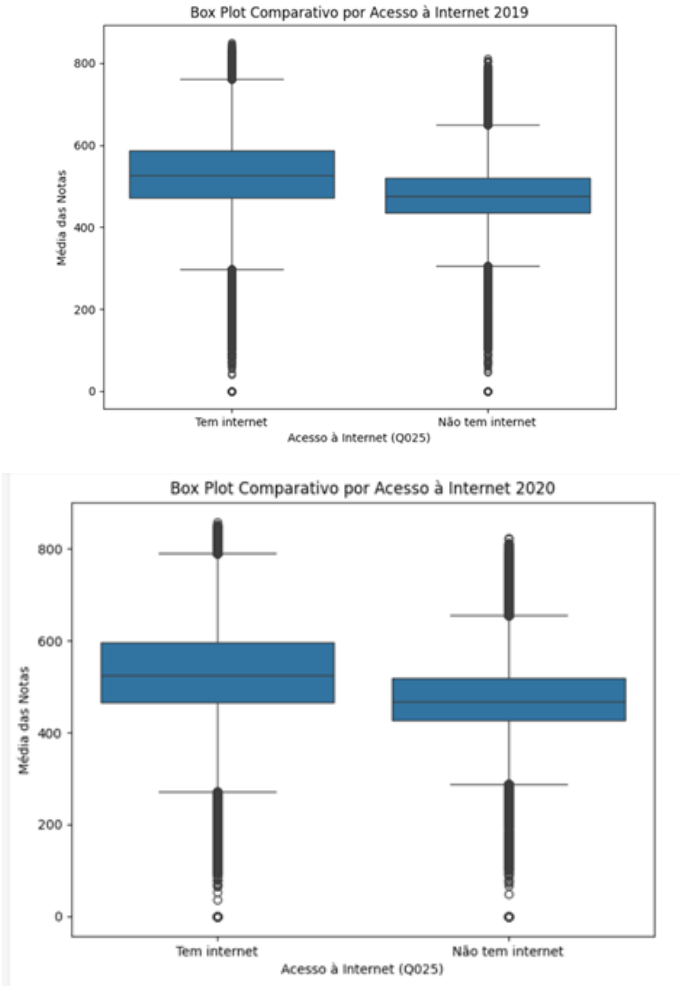

Figura 3 - Comparação das médias entre os participantes com e sem conectividade em 2019 e 2020

Foi utilizado um modelo de regressão linear para entender a contribuição de alguns fatores da pesquisa do Enem para a nota do candidato. Nossa análise também empregou um modelo de regressão logística para compreender os fatores associados à conectividade entre os participantes. Consideramos variáveis como sexo, raça, tipo de escola frequentada, renda e média das notas no Enem.

OLS Regression Results

Dep. Variable:	media_notas	**R-squared:**	0.214
Model:	OLS	**Adj. R-squared:**	0.214
Method:	Least Squares	**F-statistic:**	1.073e+05
Date:	Mon, 11 Dec 2023	**Prob (F-statistic):**	0.00
Time:	23:32:33	**Log-Likelihood:**	-2.2641e+07
No. Observations:	3931471	**AIC:**	4.528e+07
Df Residuals:	3931460	**BIC:**	4.528e+07
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	473.7150	0.269	1758.184	0.000	473.187	474.243
TP_ESCOLA[T.Privada]	40.8211	0.181	225.777	0.000	40.467	41.176
TP_ESCOLA[T.Publica]	-13.1690	0.088	-149.163	0.000	-13.342	-12.996
TP_SEXO[T.M]	5.6886	0.079	71.898	0.000	5.534	5.844
TP_COR_RACA[T.Branca]	13.4839	0.264	51.084	0.000	12.967	14.001
TP_COR_RACA[T.Indígena]	-30.6144	0.563	-54.360	0.000	-31.718	-29.511
TP_COR_RACA[T.Não declarado]	4.1294	0.375	11.019	0.000	3.395	4.864
TP_COR_RACA[T.Parda]	-6.2045	0.262	-23.670	0.000	-6.718	-5.691
TP_COR_RACA[T.Preta]	-9.7002	0.279	-34.808	0.000	-10.246	-9.154
conectividade	30.6717	0.098	312.550	0.000	30.479	30.864
Renda	0.0067	1.13e-05	592.694	0.000	0.007	0.007

Figura 4 - Resultados da Regressão Linear do Enem 2019

Logit Regression Results

Dep. Variable:	conectividade	**No. Observations:**	3931471
Model:	Logit	**Df Residuals:**	3931461
Method:	MLE	**Df Model:**	9
Date:	Mon, 11 Dec 2023	**Pseudo R-squ.:**	0.1679
Time:	23:31:45	**Log-Likelihood:**	-1.6972e+06
converged:	True	**LL-Null:**	-2.0397e+06
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.2267	0.009	-24.853	0.000	-0.245	-0.209
TP_SEXO[T.M]	0.1348	0.003	48.650	0.000	0.129	0.140
TP_COR_RACA[T.Branca]	0.3285	0.009	35.992	0.000	0.311	0.346
TP_COR_RACA[T.Indígena]	-0.6136	0.017	-36.760	0.000	-0.646	-0.581
TP_COR_RACA[T.Não declarado]	-0.1041	0.013	-8.116	0.000	-0.129	-0.079
TP_COR_RACA[T.Parda]	-0.2598	0.009	-29.285	0.000	-0.277	-0.242
TP_COR_RACA[T.Preta]	-0.1809	0.009	-19.374	0.000	-0.199	-0.163
TP_ESCOLA[T.Privada]	0.9478	0.013	70.900	0.000	0.922	0.974
TP_ESCOLA[T.Publica]	-0.0697	0.003	-24.085	0.000	-0.075	-0.064
Renda	0.0009	1.96e-06	470.112	0.000	0.001	0.001

Figura 5 - Resultados da Regressão Logística do Enem 2019

OLS Regression Results

Dep. Variable:	media_notas	**R-squared:**	0.208
Model:	OLS	**Adj. R-squared:**	0.208
Method:	Least Squares	**F-statistic:**	7.191e+04
Date:	Mon, 11 Dec 2023	**Prob (F-statistic):**	0.00
Time:	23:35:44	**Log-Likelihood:**	-1.5969e+07
No. Observations:	2732472	**AIC:**	3.194e+07
Df Residuals:	2732461	**BIC:**	3.194e+07
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	471.2748	0.362	1302.829	0.000	470.566	471.984
TP_ESCOLA[T.Privada]	35.3039	0.224	157.269	0.000	34.864	35.744
TP_ESCOLA[T.Publica]	-13.5059	0.118	-114.207	0.000	-13.738	-13.274
TP_SEXO[T.M]	6.0276	0.104	58.121	0.000	5.824	6.231
TP_COR_RACA[T.Branca]	17.4962	0.350	50.023	0.000	16.811	18.182
TP_COR_RACA[T.Indígena]	-34.8431	0.763	-45.640	0.000	-36.339	-33.347
TP_COR_RACA[T.Não declarado]	6.9843	0.497	14.051	0.000	6.010	7.959
TP_COR_RACA[T.Parda]	-8.3001	0.348	-23.868	0.000	-8.982	-7.619
TP_COR_RACA[T.Preta]	-14.2924	0.369	-38.703	0.000	-15.016	-13.569
conectividade	35.1856	0.141	249.109	0.000	34.909	35.462
Renda	0.0072	1.48e-05	484.285	0.000	0.007	0.007

Figura 6 - Resultados da Regressão Linear do Enem 2020

Logit Regression Results

Dep. Variable:	conectividade	**No. Observations:**	2732472
Model:	Logit	**Df Residuals:**	2732462
Method:	MLE	**Df Model:**	9
Date:	Mon, 11 Dec 2023	**Pseudo R-squ.:**	0.1725
Time:	23:39:20	**Log-Likelihood:**	-1.0040e+06
converged:	True	**LL-Null:**	-1.2133e+06
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.0640	0.012	5.252	0.000	0.040	0.088
TP_SEXO[T.M]	0.1001	0.004	27.059	0.000	0.093	0.107
TP_COR_RACA[T.Branca]	0.3425	0.012	27.749	0.000	0.318	0.367
TP_COR_RACA[T.Indígena]	-0.6111	0.022	-27.788	0.000	-0.654	-0.568
TP_COR_RACA[T.Não declarado]	-0.0566	0.017	-3.277	0.001	-0.090	-0.023
TP_COR_RACA[T.Parda]	-0.2496	0.012	-20.861	0.000	-0.273	-0.226
TP_COR_RACA[T.Preta]	-0.1969	0.013	-15.677	0.000	-0.222	-0.172
TP_ESCOLA[T.Privada]	1.1963	0.021	57.518	0.000	1.156	1.237
TP_ESCOLA[T.Publica]	0.0244	0.004	6.144	0.000	0.017	0.032
Renda	0.0011	2.69e-06	391.199	0.000	0.001	0.001

Figura 7 - Resultados da Regressão Logística do Enem 2020

Notamos que o sexo masculino demonstrou uma ligeira tendência a estar associado a uma maior probabilidade de conectividade em comparação com o sexo feminino. Além disso, identificamos tendências distintas entre diferentes grupos raciais. Por exemplo, indivíduos que se identificaram como pertencentes à raça branca tiveram uma tendência ligeiramente maior de conectividade em relação a outras categorias raciais.

Adicionalmente, a variável que indica o tipo de escola frequentada apresentou-se como um fator significativo. Participantes que frequentaram escolas privadas mostraram uma probabilidade maior de conectividade, em comparação com aqueles provenientes de escolas públicas. Essa descoberta sugere que o ambiente escolar pode influenciar a conectividade dos participantes.

Outros fatores como renda e desempenho médio nas provas do Enem também se mostraram relevantes. A renda teve uma associação positiva com a conectividade, indicando que participantes com renda mais alta tendem a apresentar maior conectividade. Similarmente, um melhor desempenho nas provas do Enem também foi associado a uma probabilidade mais alta de conectividade entre os participantes.

Os p-values associados à análise se mostraram significativos de forma geral. Os modelos apresentaram fator R^2 em torno de 0,20. Em trabalhos futuros, considera-se interessante utilizar uma gama maior de variáveis e fontes de dados para enriquecer a análise.

Além disso, notou-se que existiu um grande aumento na dependência da internet em 2020 comparado a 2019, onde o coeficiente da variável de estudo “Conectividade” na regressão linear foi de 30.67 para 35,18.

Foi realizada também a análise por clusterização hierárquica dos dados de 2020, agregando-se por estado. Utilizamos a proporção de candidatos com acesso a internet, média de faixa de renda e, além da nota geral do Enem, as notas de cada prova específica, para entender se haveria diferença por área do conhecimento.

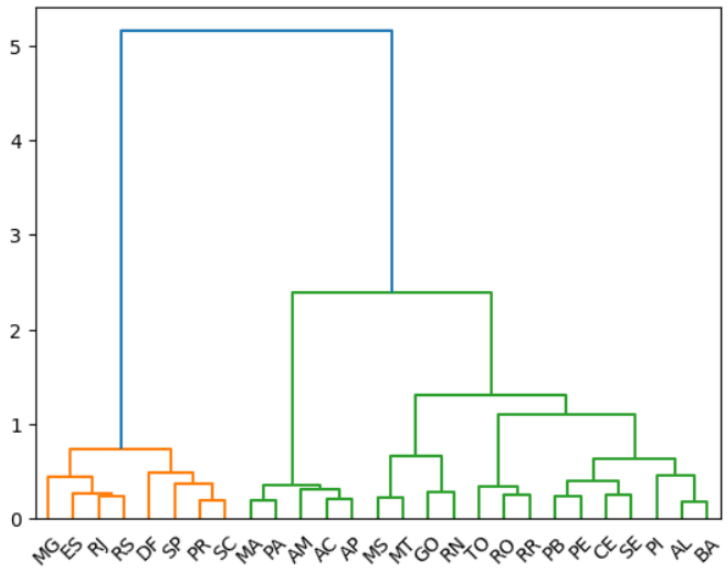

Figura 6 - Dendograma de clusterização por estado

Analisamos um agrupamento em 3 clusters, cujas características podemos observar nos gráficos a seguir.

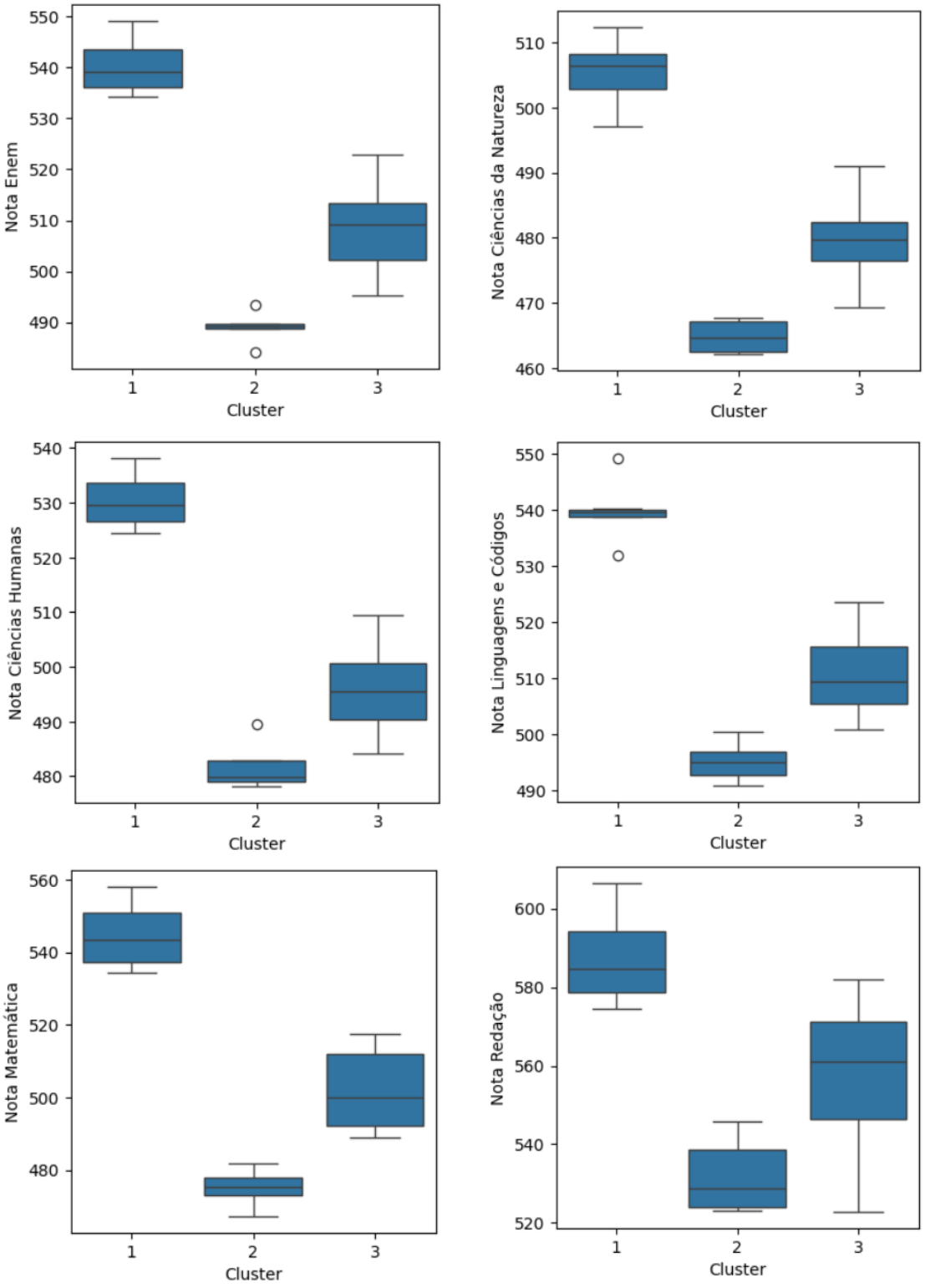

Figura 7 - Distribuições das notas de cada cluster no Enem

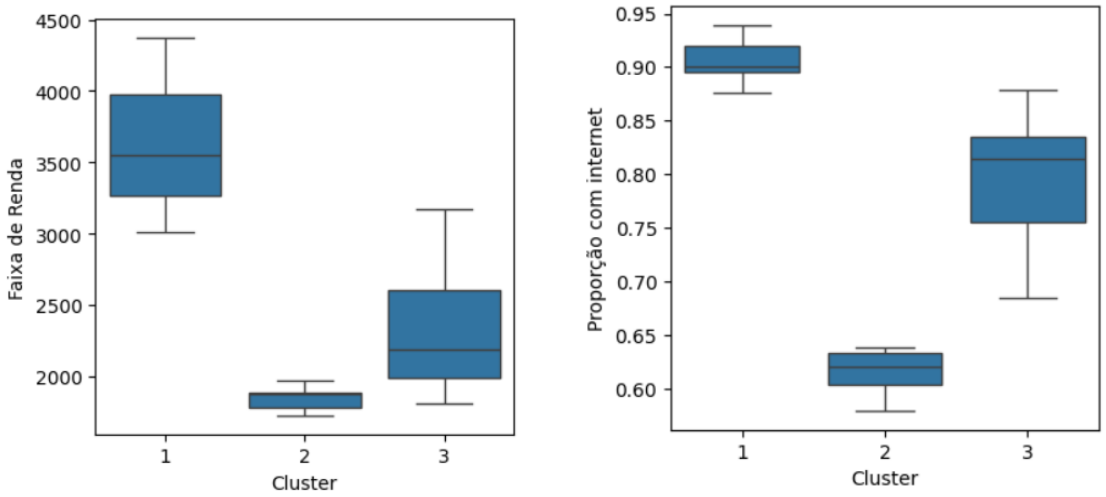

Figura 7 - Distribuições de faixa de renda e acesso à internet de cada cluster

Percebe-se que o cluster 1 (DF, ES, MG, PR, RJ, RS, SC, SP) agrupa estados com os candidatos de maior poder aquisitivo, ficando acima dos outros grupos, ainda que com certa variação. Também possui mais acesso a internet, com índices altos e distribuição concentrada. Com relação às notas na prova, percebe-se que fica acima dos outros grupos, com destaque para a prova de Linguagens e Códigos, em que apresenta comportamento de alta concentração de bom desempenho.

O cluster 2 (AC, AM, AP, MA, PA) se caracteriza pela grande concentração de baixas notas gerais no Enem, bem como muito menor proporção de acesso à internet e baixo poder aquisitivo. Já o terceiro cluster (AL, BA, CE, GO, MS, MT, PB, PE, PI, RN, RO, RR, SE, TO) abrange os casos intermediários, tanto em relação ao desempenho na prova como nos aspectos socioeconômicos de faixa de renda e conectividade com internet dos candidatos. É interessante observar que um aumento na renda familiar ocorre ao mesmo tempo de um aumento muito maior no acesso à internet deste cluster em relação ao cluster 2.