Datascraping e EDA de Concursos Públicos no Brasil

Agenda

1. **Introdução**
2. **Objetivos**
3. **Coleta de Dados**
4. **Limpeza dos Dados**
5. **Feature Engineering e EDA**
6. **Conclusões**

1. Introdução

Nesse post vamos fazer uma coleta de dados de concursos públicos no Brasil e uma análise completa deles. Os dados foram extraídos no dia 24/01/2023 do site do G1. Como o Brasil é um país que oferta muitas vagas de concursos públicos todos os anos esse tópico é interessante de estudar mais a fundo, por exemplo, como as vagas são distribuídas pelo país, ou qual região possui mais oportunidades de trabalho e mais. Dos dados que foram coletados, temos a instituição/órgão da vaga, número de vagas, salário máximo, escolaridade, local de trabalho (cidade) e estado. Utilizando técnicas de webscraping e análise de dados em Python, foram estabelecidas diversas conclusões e relações entre os dados coletados.

2. Objetivos

Os principais objetivos dessa analise são:

- Coletar os dados por meio de webscraping
- Verificar qual região possui a maior quantidade de vagas
- Relação entre escolaridade e salário máximo
- Qual escolaridade requerida é a mais presente nos concursos públicos
- Se há diferença no salário máximo nas diferentes regiões do Brasil

3. Coleta de Dados

Observando o código do site, os dados não são estáticos, ou seja, são carregados por um script de Javascript, portanto é necessário usar outra biblioteca (não pode ser a BeautifulSoup) para extrai-los. Foi utilizada a biblioteca Selenium para a coleta dos dados.

Figura 1 – Tabela dos dados coletado

	Instituto_Orgao	Vagas	Salario_Maximo	Escolaridade	Cidade	Estado
0	Prefeitura de Boca do Acre (AM)	326	R$ 3.400,00	fundamental, médio e superior	Boca do Acre	Amazonas
1	Câmara de Tunápolis	1	R$ 1.789,92	superior	Tunápolis	Santa Catarina
2	Prefeitura de Ouroeste (SP)	31	R$ 2.992,36	fundamental, médio e superior	Ouroeste	São Paulo
3	Prefeitura de Olinda (PE)	230	R$ 3.867,47	superior	Olinda	Pernambuco
4	Prefeitura de Olinda (PE)	230	R$ 3.867,47	superior	Olinda	Pernambuco
...	...	...	...	...	...	...
166	Prefeitura de Bertioga (SP)	100	R$ 2.513,07	médio	Bertioga	São Paulo
167	Prefeitura de Palmas (TO)	50	R$ 3.440,77	médio	Palmas	Tocantins
168	Fundação Estadual de Proteção Ambiental Henriq...	56	R$ 6.646,96	médio, técnico e superior	várias cidades	Rio Grande do Sul
169	Prefeitura de Capivari (SP)	57	R$ 15.519,11	fundamental, médio e superior	Capivari	São Paulo
170	Instituto Federal do Sul de Minas	4	R$ 9.616,18	superior	Pouso Alegre, Passos e Poços de Caldas	Minas Gerais

Com isso, temos na figura 1 os dados numa ‘dataframe’. Podemos ver o número de vagas, salário máximo e etc.

4. Limpeza dos dados

Para podermos trabalhar com os dados, primeiramente precisamos deixá-los prontos para isso. Na figura 1, vemos que ‘Salario_Maximo’ está formatado com ‘R$’, ‘.’ e ‘,’. Além disso, o datatype dessa coluna está como object e não como int ou float. Vamos transforma-lo pra float para essa análise. Outra coluna que temos que arruma é a de ‘Escolaridade’, pois há palavras com letras maiúsculas e outras com minúsculas. Também vamos transformar os dados de ‘Escolaridade’, ‘Cidade’ e ‘Estado’ para string (str), a fim de futuras manipulações. Por fim, foi retirado da dataframe valores nulos (N/A).

Figura 2 – Tabela limpa

	Instituto_Orgao	Vagas	Salario_Maximo	Escolaridade	Cidade	Estado
0	Prefeitura de Boca do Acre (AM)	326.0	3400.00	fundamental, médio e superior	Boca do Acre	Amazonas
1	Câmara de Tunápolis	1.0	1789.92	superior	Tunápolis	Santa Catarina
2	Prefeitura de Ouroeste (SP)	31.0	2992.36	fundamental, médio e superior	Ouroeste	São Paulo
3	Prefeitura de Olinda (PE)	230.0	3867.47	superior	Olinda	Pernambuco
4	Prefeitura de Olinda (PE)	230.0	3867.47	superior	Olinda	Pernambuco
...	...	...	...	...	...	...
166	Prefeitura de Bertioga (SP)	100.0	2513.07	médio	Bertioga	São Paulo
167	Prefeitura de Palmas (TO)	50.0	3440.77	médio	Palmas	Tocantins
168	Fundação Estadual de Proteção Ambiental Henriq...	56.0	6646.96	médio, técnico e superior	várias cidades	Rio Grande do Sul
169	Prefeitura de Capivari (SP)	57.0	15519.11	fundamental, médio e superior	Capivari	São Paulo
170	Instituto Federal do Sul de Minas	4.0	9616.18	superior	Pouso Alegre, Passos e Poços de Caldas	Minas Gerais

Podemos ver o resultado na figura 2.

5. Feature Engineering e EDA

Com os dados obtidos na figura 2, vamos substituir todas as escolaridades disponíveis em 'Escolaridade' por dummies. Cada linha será classifica pela maior escolaridade requerida para a vaga, já que futuramente vamos relacionar essas informações com o 'Salario_Maximo'. Foram classificadas da seguinte maneira:

- 1: ensino fundamental
- 2: médio
- 3: técnico
- 4: superior
- 5: doutorado
- 6: magistério

Além disso, vamos adicionar uma nova coluna 'Regiao' para determinar em qual região do Brasil aquelas vagas se encontram. São elas:

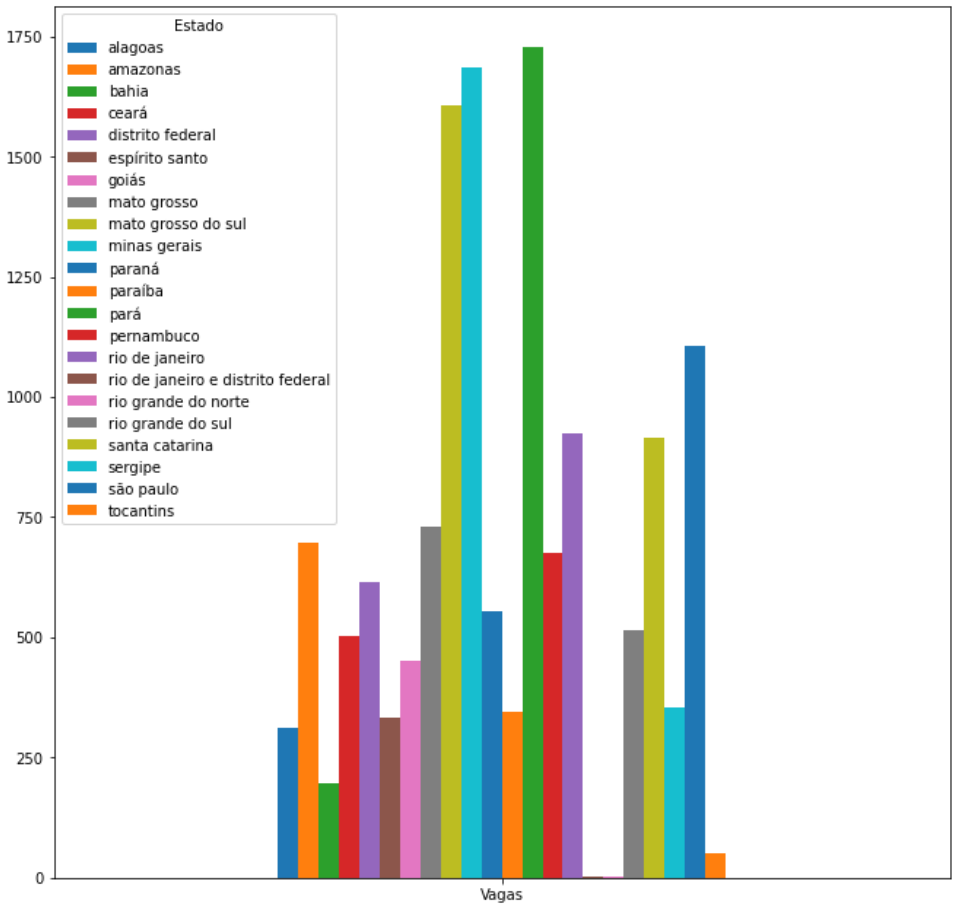
- Nordeste
- Norte
- Centro-Oeste
- Sudeste
- Sul

Figura 3 – Tabela EDA

	Instituto_Orgao	Vagas	Salario_Maximo	Escolaridade	Cidade	Estado	Regiao
0	Prefeitura de Boca do Acre (AM)	326.0	3400.00	4	boca do acre	amazonas	NORTE
1	Câmara de Tunápolis	1.0	1789.92	4	tunápolis	santa catarina	SUL
2	Prefeitura de Ouroeste (SP)	31.0	2992.36	4	ouroeste	são paulo	SUDESTE
3	Prefeitura de Olinda (PE)	230.0	3867.47	4	olinda	pernambuco	NORDESTE
4	Prefeitura de Olinda (PE)	230.0	3867.47	4	olinda	pernambuco	NORDESTE
...	...	...	...	...	...	...	...
162	Prefeitura de Bertioga (SP)	100.0	2513.07	2	bertioga	são paulo	SUDESTE
163	Prefeitura de Palmas (TO)	50.0	3440.77	2	palmas	tocantins	NORTE
164	Fundação Estadual de Proteção Ambiental Henriq...	56.0	6646.96	4	várias cidades	rio grande do sul	SUL
165	Prefeitura de Capivari (SP)	57.0	15519.11	4	capivari	são paulo	SUDESTE
166	Instituto Federal do Sul de Minas	4.0	9616.18	4	pouso alegre, passos e poços de caldas	minas gerais	SUDESTE

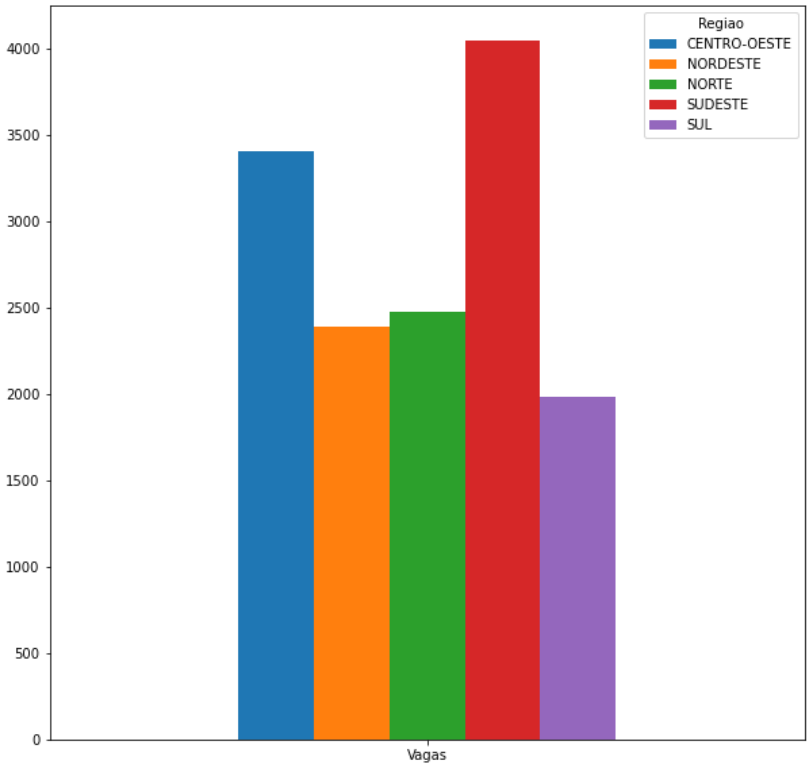
Podemos ver o resultado na figura 3.

Figura 4 – Número de vagas em cada estado



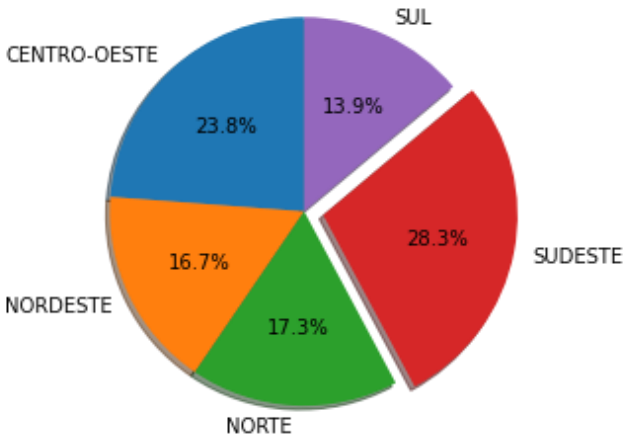
Na figura 4, temos que o estado com mais vagas de concursos públicos é a Bahia seguido por Minas Gerais. Também é possível observar que os outros estados do Sudeste (São Paulo e RJ) possuem um número grande de vagas.

Figura 5 – Regiões com mais vagas



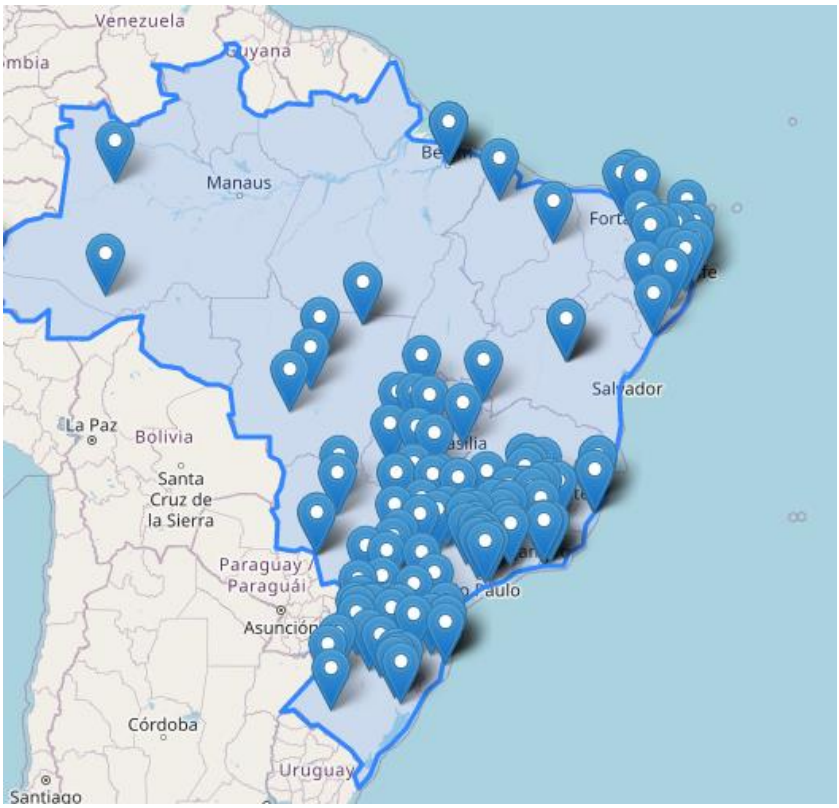
Confirmando o que foi visto na figura 4, temos que na figura 5 a região brasileira com mais vagas é o Sudeste e em segundo lugar o Centro-Oeste. A região com menos vagas é o Sul.

Figura 6 – % de vagas em cada região



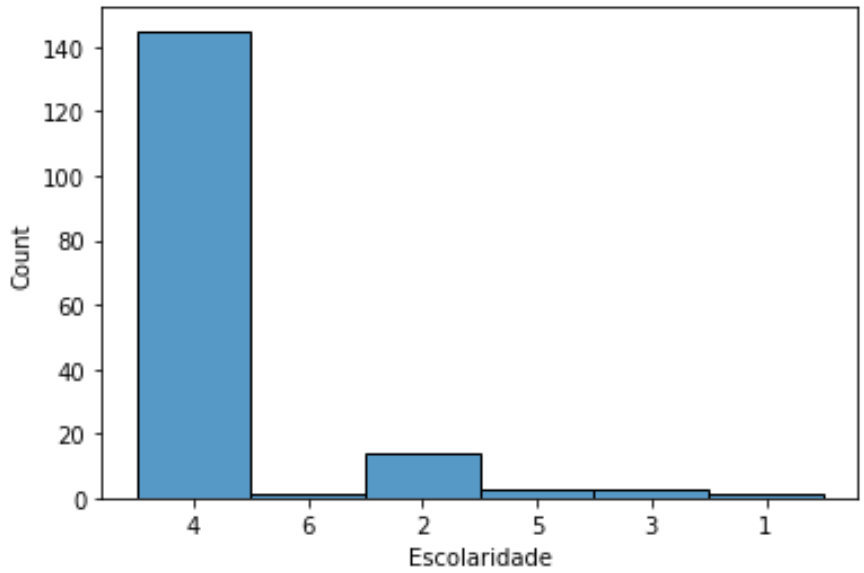
Na figura 6, temos que o Sudeste está em primeiro lugar com 28,3 % das vagas dentre todas as regiões. Logo em seguida temos Centro-Oeste com 23,8 %. E a região com menos vagas, como visto na figura 5, é o Sul com apenas 13,9 % das vagas.

Figura 7 – Cidades com concursos públicos



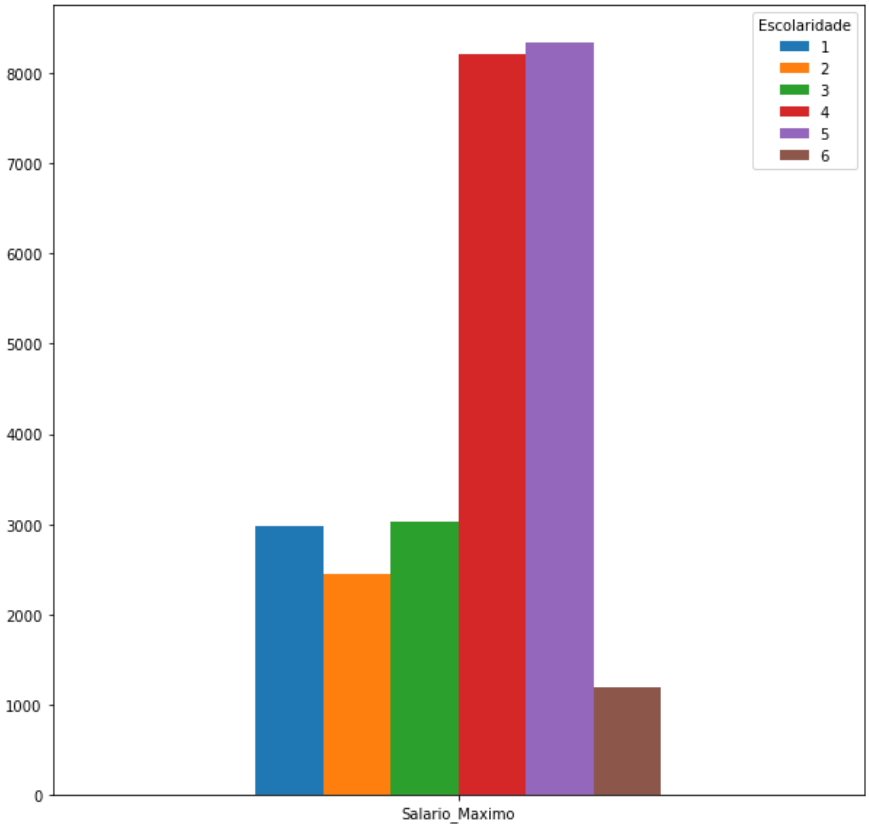
Como é visto na figura 7, a maioria das cidades com vagas para concursos públicos disponíveis ficam, realmente, no Sudeste e Centro-Oeste. No código temos algumas cidades (poucas que não interferem na análise) fora do Brasil, o motivo disso é que a biblioteca que foi utilizada para determinar a latitude e longitude de cada cidade usa o nome da mesma para pegar suas coordenadas e algumas cidades brasileiras possuem o mesmo nome de algumas cidades em outros países.

Figura 8 – Histograma de ‘Escolaridade’



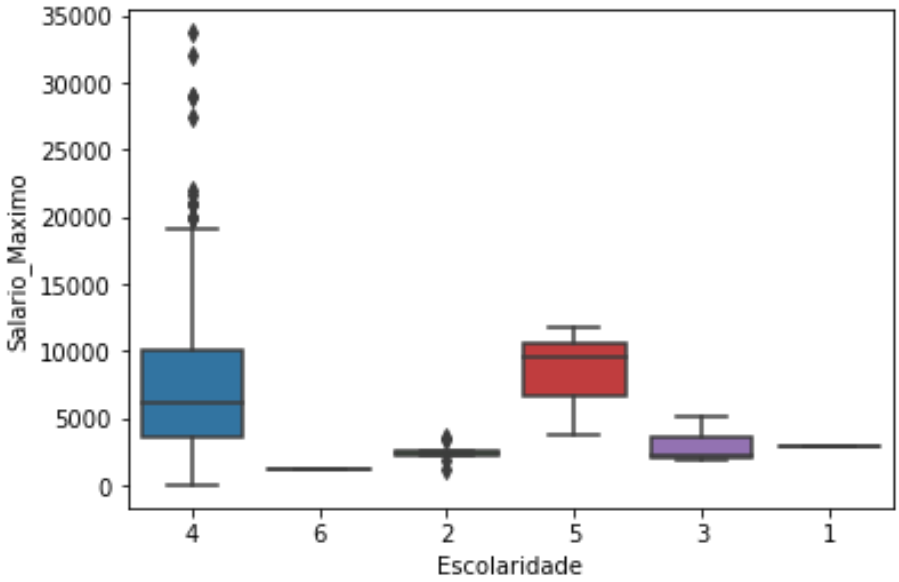
Para verificarmos qual a maior escolaridade requerida para as vagas, temos um histograma, figura 8, para essa análise. Com base na figura 8, vemos que a necessidade de escolaridade ‘superior’ está presente na maioria das vagas. Já o ensino médio está em segundo lugar, mas com um número bem menor que o de ensino superior.

Figura 9 – Média Salário Máximo e Escolaridade



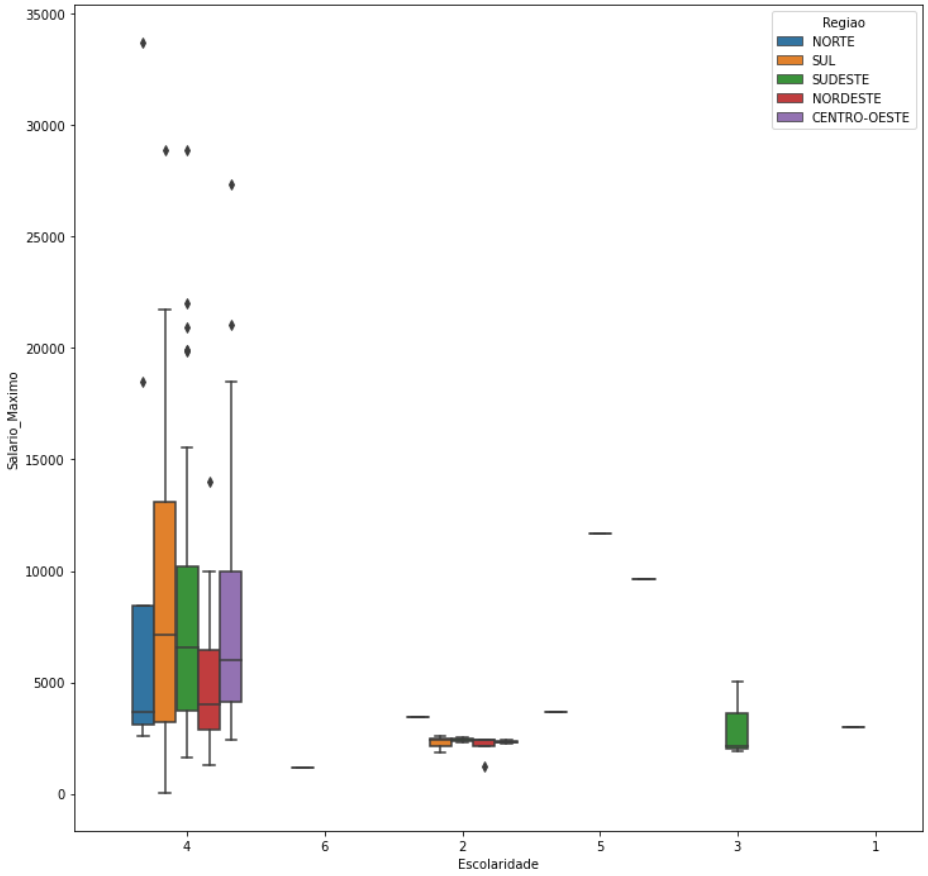
Na figura 9, vemos que as vagas para escolaridades ‘superior’ e ‘doutorado’, 4 e 5, respectivamente, são as vagas que possuem as maiores médias de salário máximo dentre todas.

Figura 10 – Boxplot de Salario_Maximo e Escolaridade



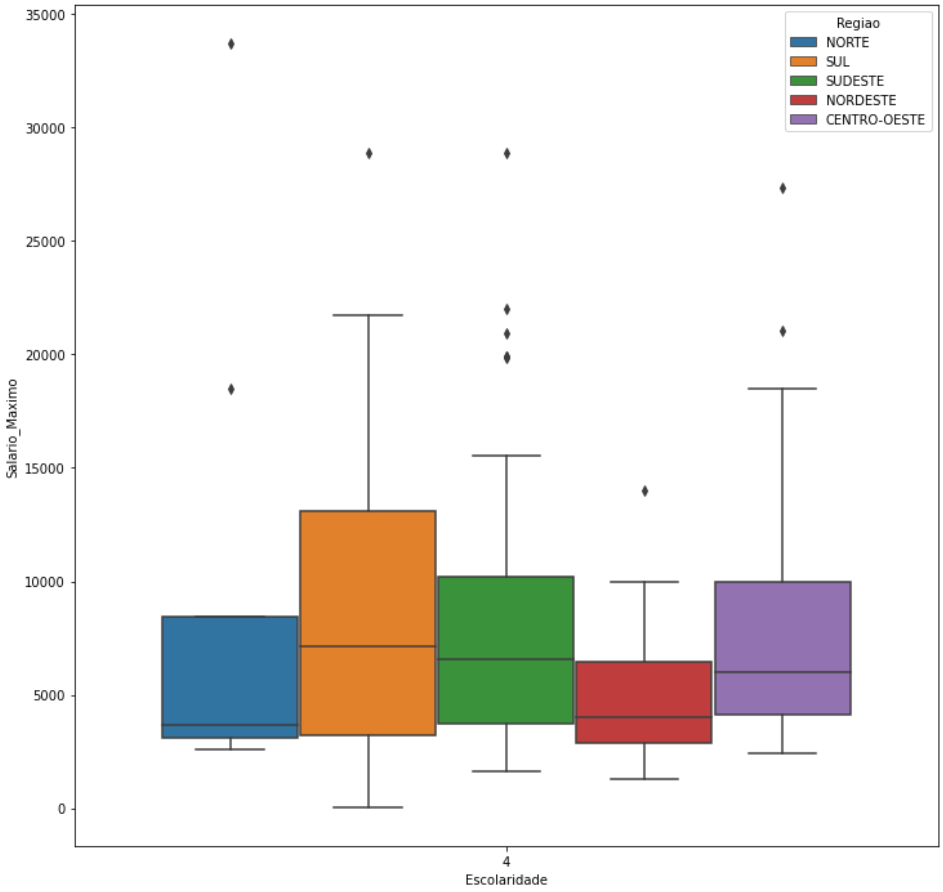
No boxplot da figura 10, vemos que quem aplica para as vagas de doutorado pode ter um salário maior que 25 % das pessoas que aplicam para vagas de ensino superior (1º quartil), além disso, aproximadamente 75 % das vagas para doutorado pagam quase 10 mil reais, ou seja, mais que 75 % das vagas para ensino superior. Por outro lado, as vagas de ensino superior podem pagar mais que as de doutorado, o ultimo quartil varia de 10 mil a quase 20 mil reais, quase o dobro que o máximo pago por vagas de doutorado. Já para as outras escolaridades, as diferenças salariais são irrelevantes.

Figura 11 – Boxplot de Salario_Maximo e Escolaridade com hue=Regiao



Na figura 11, vemos as diferenças entre os salários máximos entre as escolaridades, como na figura 10, porém nesse boxplot, temos essa comparações entre diferentes regiões. A parte desse boxplot que mais se destaca é na escolaridade ‘superior’, portanto vamos dar um zoom nessa parte.

Figura 12 – Boxplot de Salario_Maximo e Escolaridade='superior' com hue='Regiao'



Como dito acima, vamos colocar um foco no ensino superior e podemos ver essas diferenças na figura 12. Aqui podemos destacar que a região norte é a que menos paga dentre todas as regiões. A região Sul é que mais possui variações de salário máximo, podendo chegar a pouco mais de 20 mil reais. Além disso, temos que pouco mais de 50 % do salários máximos nas regiões Sul e Sudeste podem ser maiores que 75 % dos salários máximos das regiões Norte e Nordeste. Por fim, o maior do salário máximo na região Nordeste é aproximadamente igual ao maior valor salário máximo do 3º quartil das regiões Sudeste, Sul e Centro-Oeste.

6. Conclusões

Com o uso da biblioteca Selenium foi possível coletar os dados e foi possível analisar os dados. Vimos que a região Sudeste é a que mais possui vagas. A relação entre escolaridades e salario máximo é a de quanto maior a escolaridade maior o salário máximo e que a escolaridade mais presente nos concursos é a de ensino superior e, por fim, que há uma diferença nos salários máximos nas regiões do Brasil.