BootCamp Ciência de Dados

Aluno: Jorge Leandro Piva

Desenvolvimento da Solução do Desafio Final

No primeiro passo carregamos os arquivos, tanto do consumo quanto dos estados, após isso fizemos o merge deles em um único dataframe.

Na sequencia nós identificamos valores duplicados e valores nulos, no caso dos duplicados nós dropamos imediatamente, sobre os nulos, conduzimos uma análise para saber se poderíamos tentar fazer uma inferência visto que a quantidade de valores era muito alta, após entender que eles eram de dois tipos de consumo específicos que seriam cativos e Total, nós dropamos os valores nulos e seguimos com as demais análises.

1. Visão de consumo ano a ano
2. Total de Consumidores por Ano
3. Consumo por região por Ano
4. Consumidores por região por ano
5. Consumo por consumidor por Estado
6. Distribuição de Consumo por Tipo
7. Evolução do Consumo Residencial
8. Top 10 Estados de Consumo de 2023
9. Top 10 Estados em número de Consumidores de 2023
10. Consumo por Mês (Análise de Sazonalidade)
11. Consumo por Tipo ao Longo do Tempo
12. Scatter Plot Consumidores vs Consumo
13. Consumo Total por Região em 2023
14. Crescimento percentual do consumo por região (2004-2023)
15. Consumo médio por consumidor por região ao longo dos anos
16. Consumo médio por consumidor por mês em 2023
17. Consumo médio por consumidor por estado em 2023 (heatmap)
18. Consumo médio por consumidor por estado por Ano (heatmap)

Por fim utilizamos o K-means para agrupar os estados em clusters, tivemos um resultado interessante:

Cluster 0 [ES, MT, DF, AM, RN, MS, PB, AL, SE, PI, RO, TO, AC, AP, RR] São estados com população e consumo menores

Cluster 1 SP o estado é tão grande tão fora da curva em consumo e população que ele teve um custer só para ele não haviam vizinhos próximos nos quesitos população e consumo

Cluster 2 [BA, RS, PR] São estados com bom consumo e boa População.

Cluster 3 [RJ, MG] Se SP é fora da curva em relação ao restante do Brasil RJ e MG são fora da curva quando removemos SP, eles são muito grandes em termos de população e industrialização então tem um cluster só para eles dois.

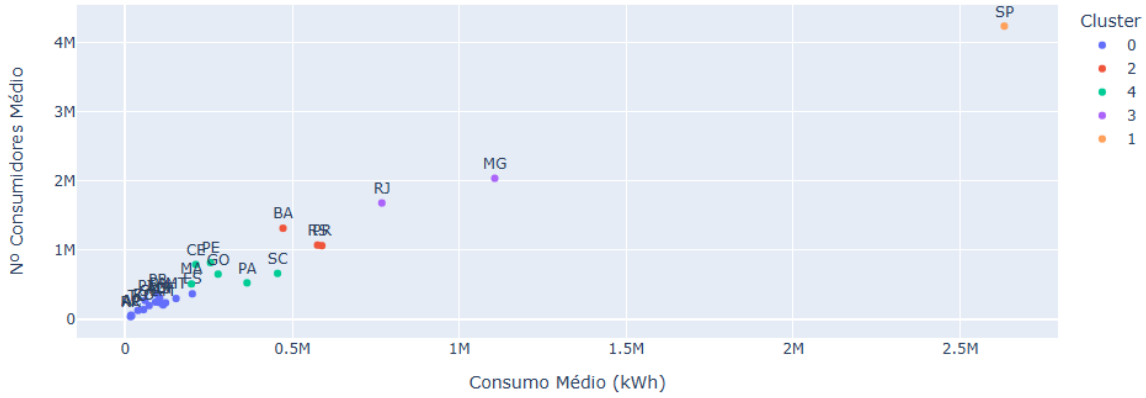
Cluster 4 [SC, PA, GO, PE, CE, MA] São estados com dados de consumo mais medianos, nem tão fortes nem tão fracos.

Conclusão

Foram aplicados diversos conhecimentos adquiridos ao longo do curso como a manipulação dos dados, diversas visualizações diferentes, tratamentos de valores ausentes, valores duplicados, e diversos tipos de visualização.

A inferência de dados para tratamento de valores nulos seria uma tratativa interessante mas não foi possível de ser aplicada porque não havia nenhum registro com população para os dois tipos de consumo citados então a decisão foi pela remoção, porém, esta decisão foi uma das mais difíceis dentro do modelo.

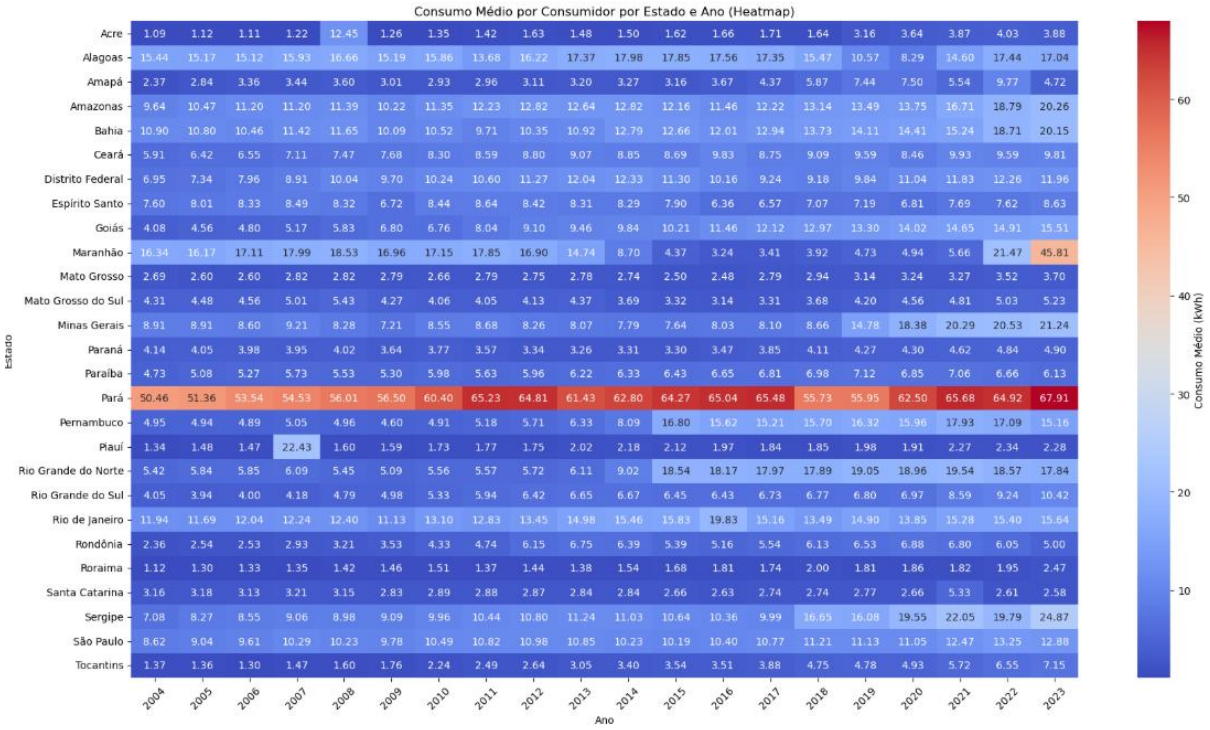
Ao longo do desafio pudemos encontrar diversas informações valiosas, como por exemplo ao clusterizar os estados nós identificamos que com 5 clusters, 1 fica para são Paulo, 1 para MG e RJ e 1 para BA, RS, PR ou seja foram 3 clusters para jogar 6 estados e outros 2 clusters para os demais estados que seriam 21.



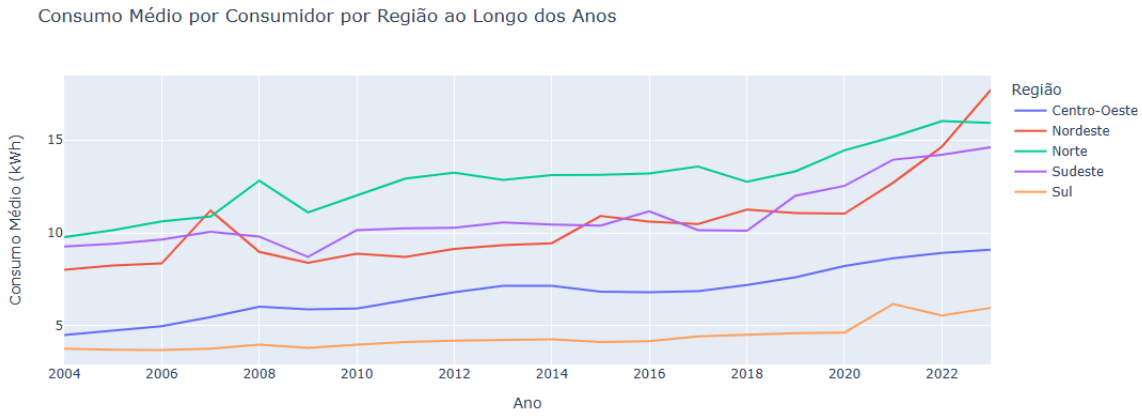
Essa imagem ficou bem característica disso, temos estados que despontam dos 3

clusters mencionados e os outros são estados médios e estados pequenos em termos de população e consumo.

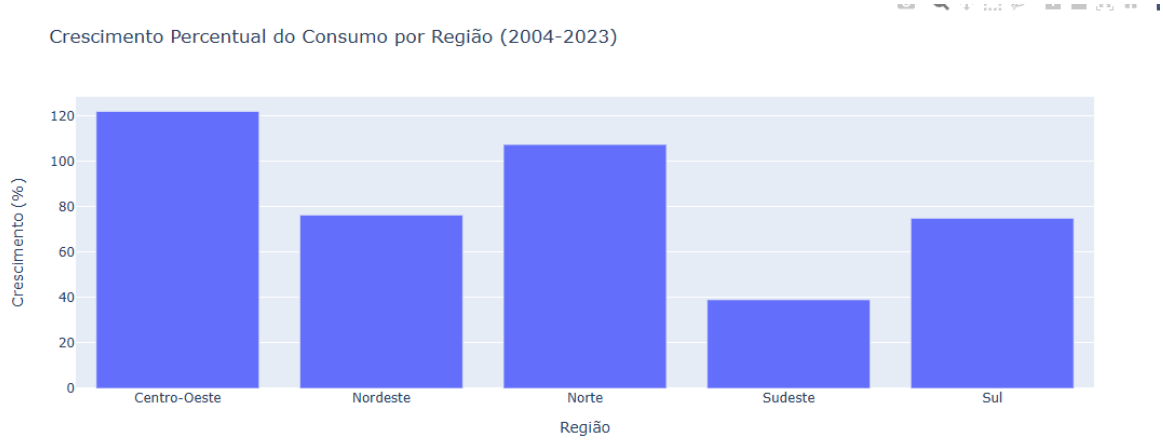
Uma informação que era totalmente desconhecida antes desta condução é o fato do Pará e do Maranhão terem o maior consumo médio de todo o país, estados do norte e nordeste em geral com esse indicador muito alto.



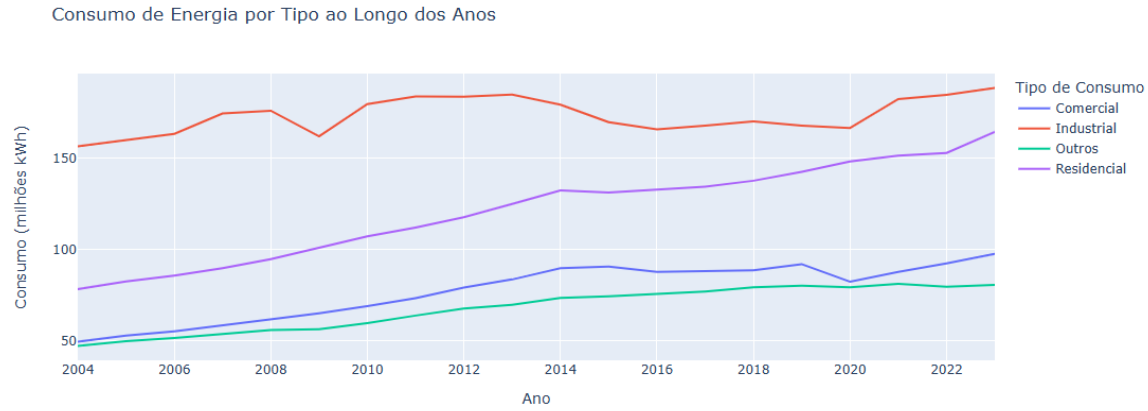
É interessante também ver como o nordeste de forma geral cresceu seu consumo médio ficando em primeiro lugar nesse quesito.



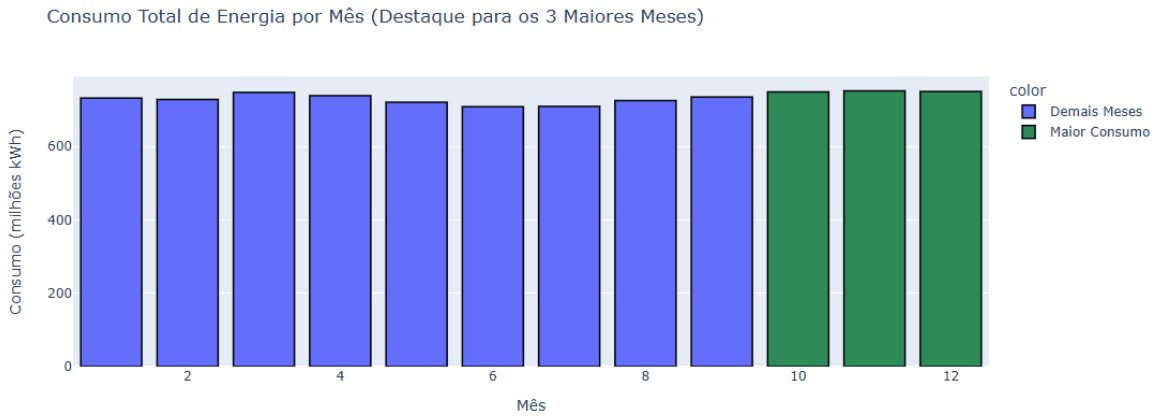
Sobre crescimento percentual a região centro-oeste e norte tiveram o maior crescimento, esta também foi um descoberta interessante.



Já em questão de crescimento o consumo residencial está acelerando muito nos últimos anos, isso talvez se dê por uma informatização muito grande, muitas pessoas com computadores hoje em dia, mais tvs, videogames, e também trabalhando de home-office, quando mais acessíveis os eletrônicos se tornam maior o consumo de energia.



Também identificamos que os meses que mais tem consumo de energia são de outubro a dezembro uma possível consequência do ar condicionado já que o Brasil é um país muito quente, mas até aqui apenas uma suposição.



Foram diversos insights e muitos outros podem ser identificados em nosso arquivo do jupyter notebook, no entanto para melhoria incial devido ao tempo e também porque tivemos que fazer essa entrega em uma semana de feriado onde muitos alunos estavam com viagens marcadas deixamos para um segundo momento o cruzamento da base com outras bases como PIB, IDH, População, entre outras mas já tivemos diversas informações uteis com as bases que foram utilizadas, outro ponto interessante foi a Clusterização com o K-means, apesar de não abordada neste módulo ela também pode ser um técnica de análise exploratória e foi aprendida no bootcamp de machine learning para enriquecer esta entrega.