Trabalho final de Aprendizado Não Supervisionado.

1. Introdução

Neste trabalho estou utilizando um dataset retirado do UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/Air+quality.

O problema proposto é a análise dos dados de sensores químicos da qualidade do ar, a fim de verificar o comportamento dos poluentes nesta área classificada como poluída dentro de uma cidade Italiana. São fornecidos pelo dataset medições de gases nocivos tanto aos seres vivos, quanto ao meio ambiente. O objetivo inicial é avaliar como esses gases nocivos se comportam, dependendo da época do ano, temperatura, dia da semana e até pela reação com outros gases.

Os atributos disponíveis ao estudo são:

- Data em que foi realizada a mensuração

- Hora em que foi realizada a mensuração

- Concentração média horária real de CO em mg / m³ (analisador de referência), o monóxido de carbono (CO) é um gás inodoro e muito perigoso pois realiza ligações estáveis com a hemoglobina, sendo tóxico para o corpo, é formado por combustão.

- PT08.S1 (óxido de estanho) média de resposta horária do sensor (nominalmente CO direcionado)

- Concentração de hidrocarbonetos não metânicos em média horária real em microg / m³ (analisador de referência). São compostos muito reativos que interferem no ciclo global do carbono, produzidos naturalmente pelas florestas, principalmente por pinheiro, mas também esses gases são produzidos na combustão incompleta dos combustíveis, como a gasolina, o gás natural e o GLP.

- Concentração de benzeno média horária real em microg / m³ (analisador de referência). Produzidos quando materiais ricos em carbono passam por combustão incompleta, podendo ser originário de incêndios e vulcões.

- PT08.S2 (titânia) com média de resposta horária do sensor (nominalmente direcionada ao NMHC)

- Concentração de NOx média horária verdadeira em ppb (analisador de referência). NOx ou óxidos de nitrogênio são um ramo de gases tóxicos altamente reativos. Os gases de óxido de nitrogênio estimulam a formação de smog e chuva ácida, além de desempenhar um papel na influência do ozônio troposférico.

- PT08.S3 (óxido de tungstênio) resposta média horária do sensor (nominalmente direcionado a NOx)

- Concentração de NO2 média horária em microg / m^3 (analisador de referência). Processos de combustão tendem a emitir baixas concentrações de NO2 em relação aos valores de NO, mas ao entrar em contato com o oxigênio do ar, as moléculas de NO logo se convertem em NO2 e, por este motivo, as taxas de emissão são sempre calculadas considerando ambos os compostos como sendo apenas NO2. Atinge os revestimentos celulares das vias respiratórias, indo desde o nariz até os alvéolos pulmonares. Em casos de intoxicação grave, pode ainda causar hemorragias, insuficiência respiratória e até a morte.

- PT08.S4 (óxido de tungstênio) resposta média horária do sensor (nominalmente direcionado para NO2)

- PT08.S5 (óxido de índio) resposta média horária do sensor (nominalmente O3 direcionado)

- Temperatura em °C

- Umidade Relativa (%)

- Umidade Absoluta

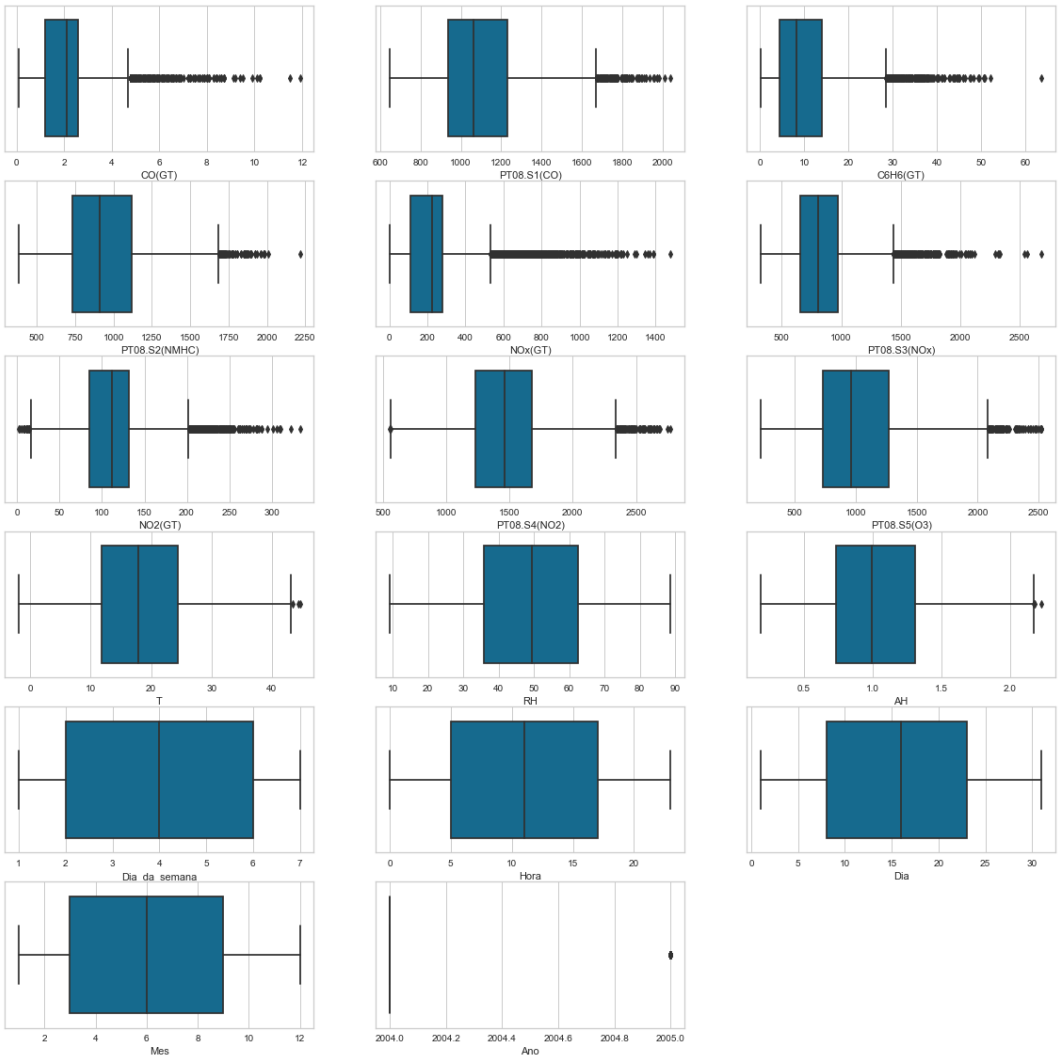
- Dia da semana: O dia é dado como um inteiro, variando de 1 (domingo) a 7 (sábado), por padrão. Foi utilizada a funcionalidade do Excel para nos trazer o dia da semana, pois a credita-se que pode ser uma informação importante.

2. Tratamento dos dados

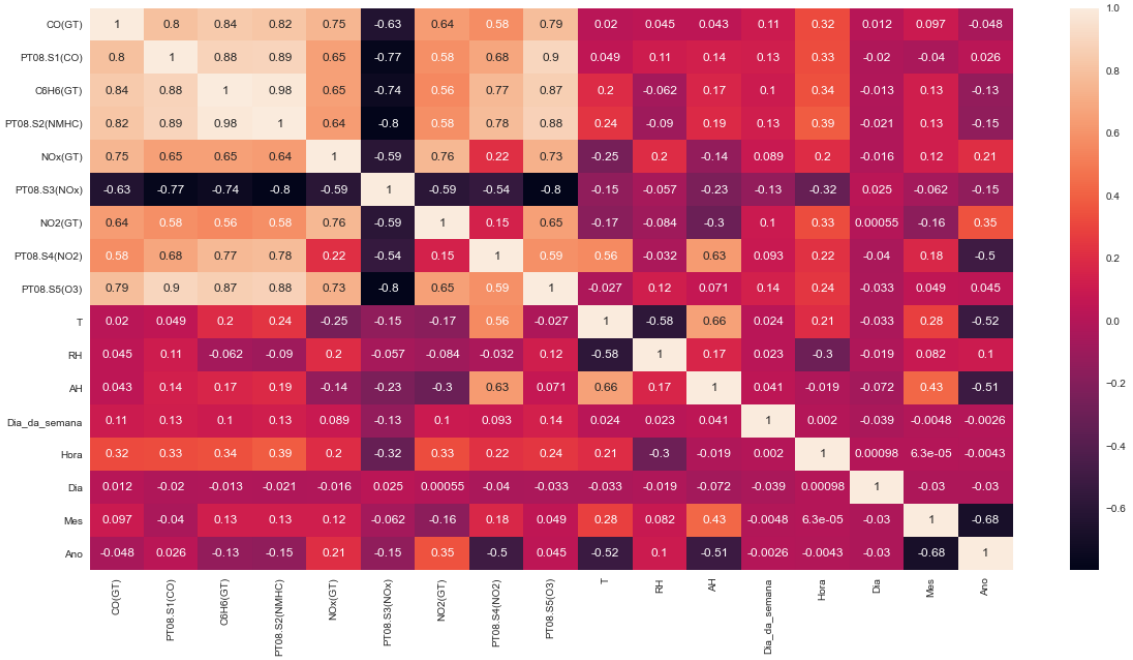
Inicialmente ao verificar os dados, não existem valores nulos no dataset, mas isto ocorre porque os erros de leitura são caracterizados como '-200'. Quando verificado, apenas 9.77% das 9357 leituras de NMHC foram recolhidas de forma correta, por este motivo a informação de NMHC foi desconsiderada no estudo. Outro ponto considerado na limpeza dos dados foi, se pelo menos 80% dos poluentes derem erro de leitura, os registros daquele momento serão desconsiderados. No total foram descartadas 340 linhas do dataset. Demais dados que não foram corretamente lidos, foram reparados com respectiva a média do atributo, a fim de impactar o mínimo possível a análise.

Devido a formatação original da coluna hora, ela precisou ser tratada para que pudesse ser utilizada. A coluna date foi separada em dia, mês e ano, já que o intuito é aprendizado supervisionado e desta maneira a data pode trazer mais informações. Por exemplo, se há maior concentração de poluentes em um mês específico, ou se no início dos meses ocorrem maiores poluições de carros.

Foram encontrados valores discrepantes, mas provavelmente eles são importantes à análise, por este motivo não foram desconsiderados.



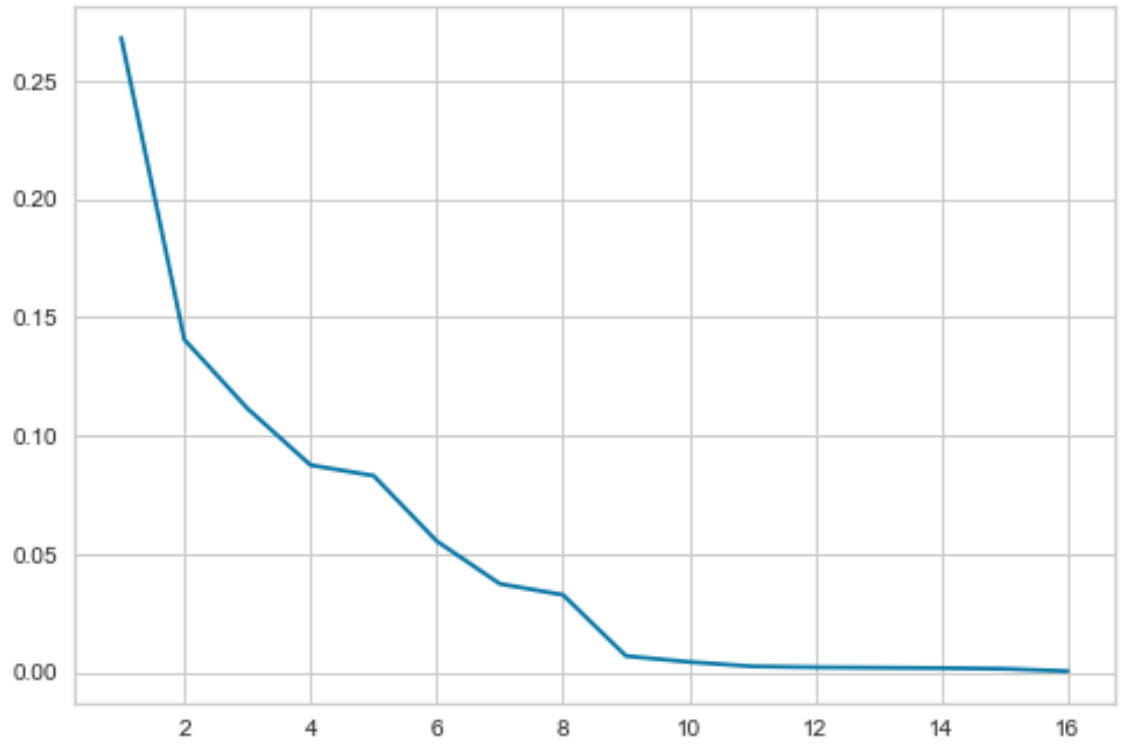
Ao analisar a correlação, foi retirada da análise o PT08.S2(NMHC), pois ele é fortemente correlacionado com várias outras variáveis.



Os dados estão em escalas muito distintas, desta forma a aplicação de métodos por distância serão negativamente impactadas, impossibilitando o desempenho ideal e o correto funcionamento. Fazendo a normalização, vamos colocar tudo na escala 0-1, tornando possível a comparação e análise de distância entre as variáveis.

Para melhorar a precisão do classificador, e com menor custo, é adotado o PCA para a redução de dimensionalidade. Com 6 componentes 0.89 da variância já é explicada. O primeiro cotovelo ocorre na 4º componente, onde apenas 0.73 da variância é explicada, já o segundo cotovelo ocorre com 7 componentes, onde 0.94 da variância é explicada.

Gráfico da variância explicada (PCA):

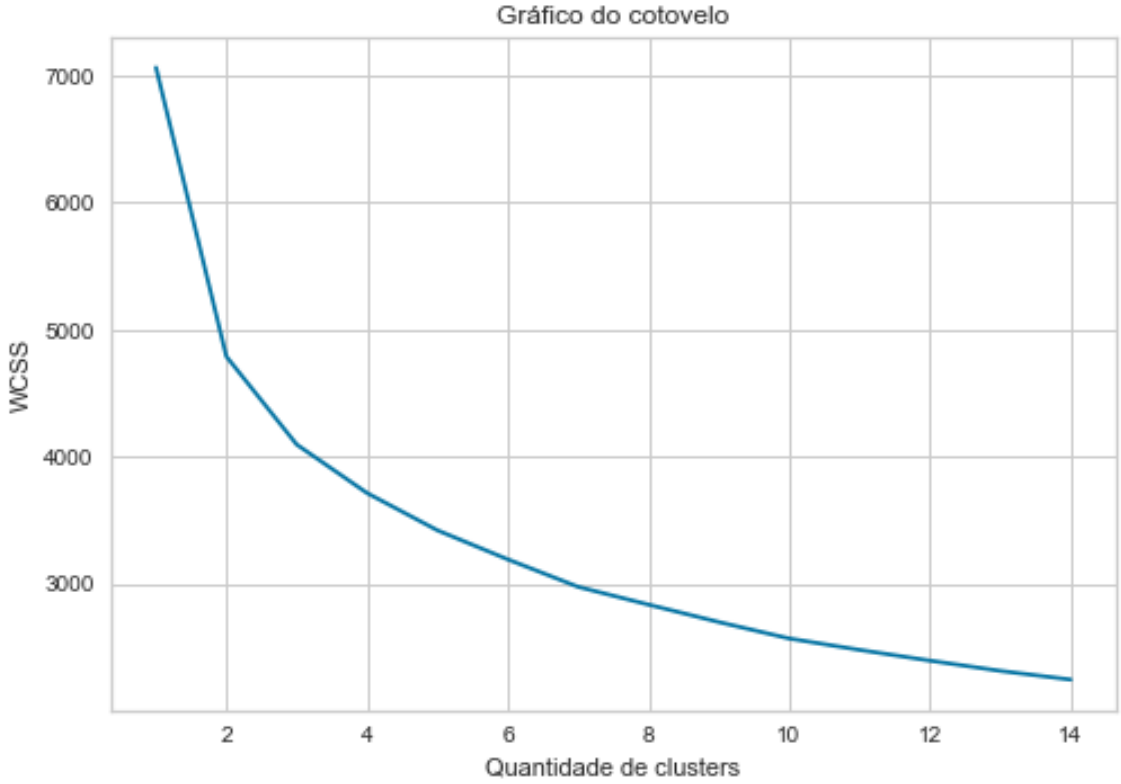


3. Modelagem

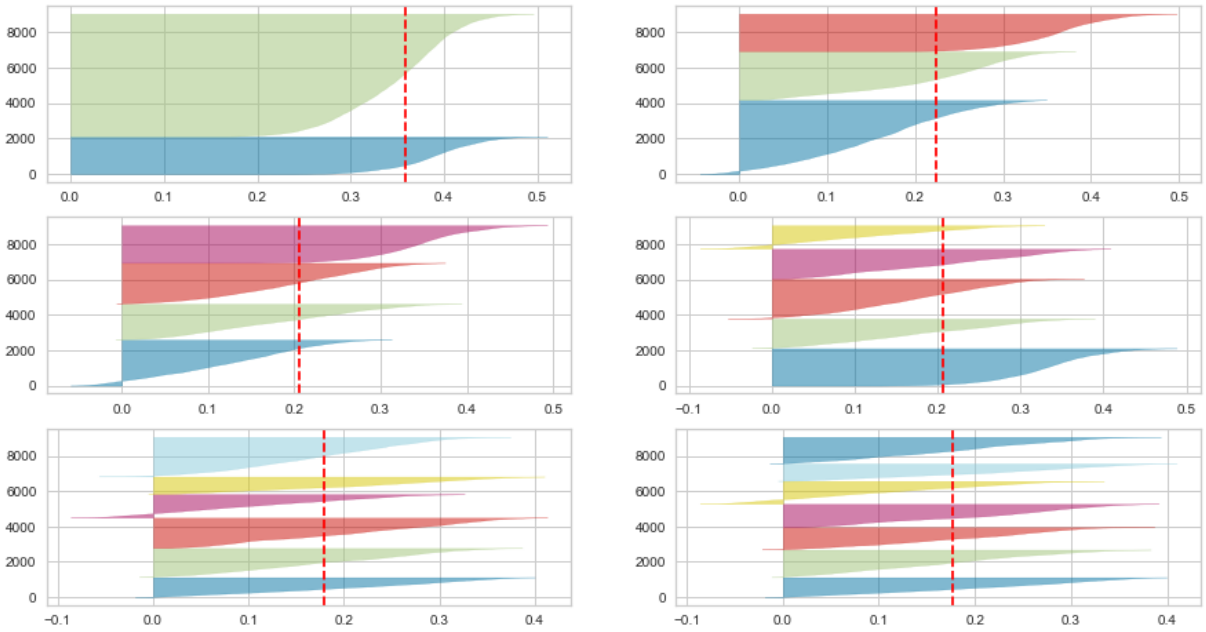
a. K-médias

É um algoritmo de particionamento, onde seus centróides de ajustam a fim de minimizar o erro do agrupamento que é definido pela soma da distância euclidiana entre cada indivíduo e o centróide do grupo.

O gráfico do cotovelo retorna o WCSS pela quantidade de Clusters. Onde o WCSS é a soma da distância quadrada entre cada ponto e o centróide em um cluster.



Pelo gráfico, o cotovelo ocorre com 2 clusters, mesmo assim vou utilizar o método silhuete para validar a melhor quantidade de clusters.

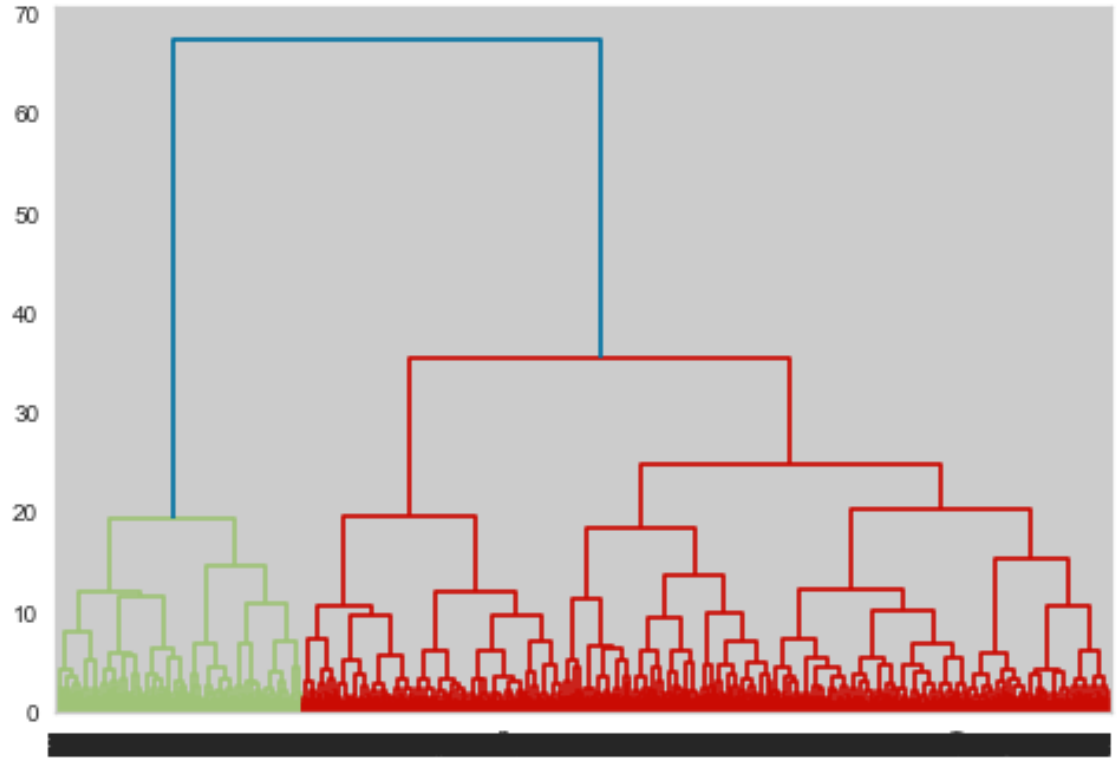


Pela análise da silhueta, realmente 2 clusters é o mais indicado para o problema.

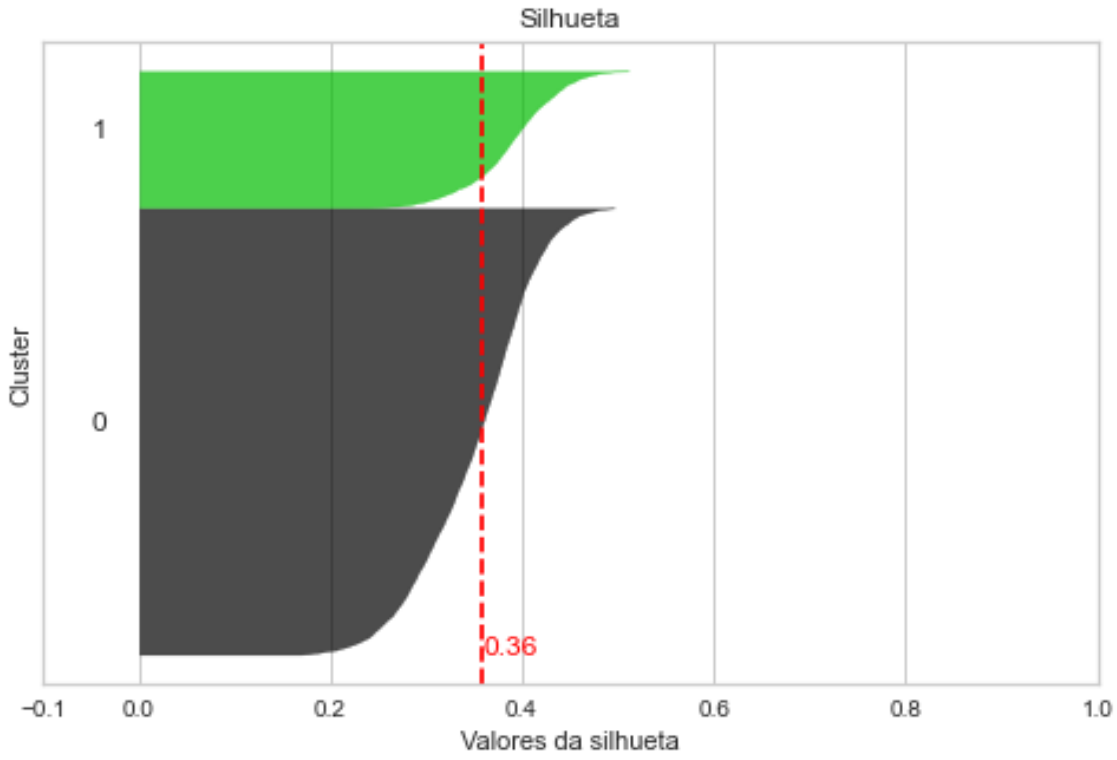
b. Hierárquico

Aqui foi escolhida a ligação de Ward. Neste método aglomerativo utilizado, a distância entre os grupos se baseia na distância intra e entre os clusters. Esse método tende a combinar clusters com poucas observações e aproximadamente o mesmo tamanho, além de possuir uma alta sensibilidade a ruídos.

Dendograma com ligação de Ward:



O dendrograma indica dois clusters, para validar e viabilizar a comparação com o método K-médias, aqui também é realizada a análise da silhueta.

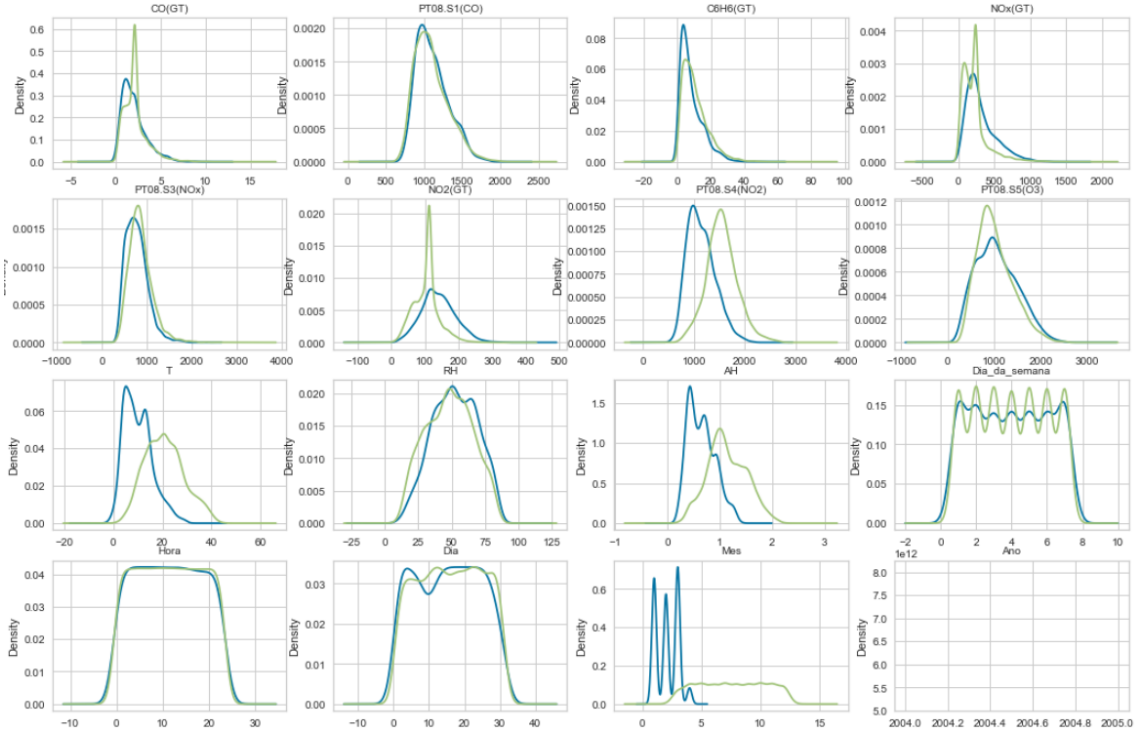


4. Avaliação

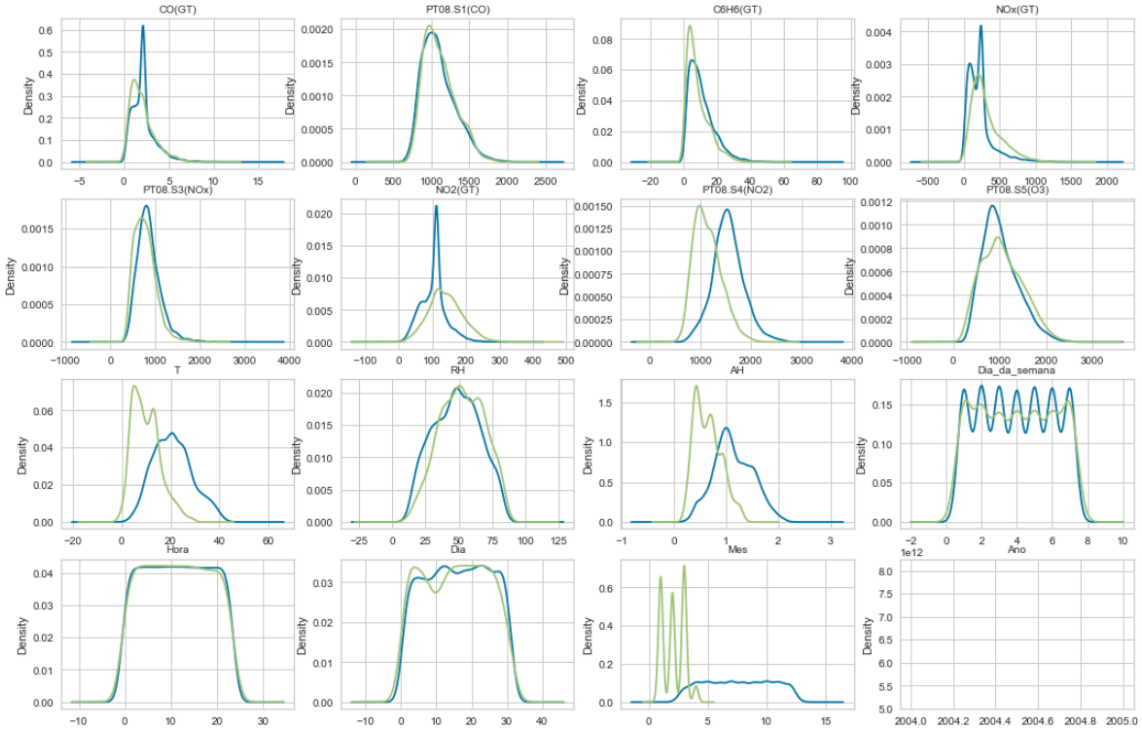
Tanto o método hierárquico quanto o k-médias indicam a criação de dois clusters, e os valores da silhueta são iguais nos dois métodos, 0.36.

Para uma melhor análise, os dados dos dois clusters são comparados na escala real das variáveis.

Clusters de K-médias:



Clusters de método hierárquico:



Não há diferença entre os algoritmos, ambos classificaram as observações nos mesmos clusters.

5. Conclusão

Não houve diferença entre as metodologias K-médias e Hierárquico no problema proposto. Ambas tiveram o mesmo desempenho e realizaram a mesma separação dos dados.

Foram encontrados dois clusters, o primeiro apresenta temperaturas em geral menores que 20º, humidade absoluta do ar em geral menor do que 1, fica concentrado entre os meses de janeiro a maio do ano de 2005. Já o segundo cluster parece ter uma concentração maior de NO2(GT), que fica por volta de 100, com maior concentração média de PTO8.S4.

Tabela do primeiro cluster:

	CO(GT)	PT08,S1(CO)	C6H6(GT)	NOx(GT)	PT08,S3(NOx)	NO2(GT)	PT08,S4(NO2)	PT08,S5(O3)	T	RH	AH	Dia_da_semana	Hora	Dia	Mes	Ano
count	2.109	2.109	2.109	2.109	2.109	2.109	2.109	2.109	2.109	2.109	2.109	2.109	2.109	2.109	2.109	2.109
mean	2,0	1.110,0	8,3	312,6	767,5	139,7	1.144,4	1.055,6	10,0	52,4	0,7	4	11	15	2	2.005
std	1,3	207,0	6,6	204,9	236,5	49,7	269,7	440,8	5,7	16,5	0,3	2	7	9	1	0
min	0,1	715,0	0,1	17,0	330,0	17,0	551,0	221,0	-1,9	9,9	0,2	1	0	1	1	2.005
25%	1,0	949,0	3,4	166,0	589,0	108,0	941,0	706,0	5,4	39,9	0,4	2	5	7	1	2.005
50%	1,7	1.076,0	6,3	250,0	740,0	137,0	1.109,0	1.011,0	9,1	52,2	0,6	4	11	16	2	2.005
75%	2,7	1.238,0	11,7	408,0	904,0	171,0	1.310,0	1.370,0	13,5	65,4	0,8	6	17	23	3	2.005
max	8,7	1.846,0	43,0	1.230,0	1.881,0	333,0	2.147,0	2.494,0	30,0	86,6	1,4	7	23	31	4	2.005

Tabela do segundo cluster:

	CO(GT)	PT08,S1(CO)	C6H6(GT)	NOx(GT)	PT08,S3(NOx)	NO2(GT)	PT08,S4(NO2)	PT08,S5(O3)	T	RH	AH	Dia_da_semana	Hora	Dia	Mes	Ano
count	6.908	6.908	6.908	6.908	6.908	6.908	6.908	6.908	6.908	6.908	6.908	6.908	6.908	6.908	6.908	6.908
mean	2,2	1.096,7	10,6	220,4	856,2	103,7	1.551,5	1.012,9	20,9	48,3	1,1	4	11	16	8	2.004
std	1,3	219,6	7,6	175,5	258,7	37,1	308,3	383,3	8,0	17,4	0,4	2	7	9	3	0
min	0,1	647,0	0,2	2,0	322,0	2,0	682,0	261,0	1,2	9,2	0,2	1	0	1	3	2.004
25%	1,3	932,0	5,0	100,0	682,0	79,0	1.356,0	737,0	14,8	34,6	0,9	2	6	9	5	2.004
50%	2,1	1.061,0	8,8	207,0	824,0	112,1	1.538,0	951,5	20,4	48,6	1,1	4	11	16	8	2.004
75%	2,5	1.228,0	14,4	242,0	988,0	118,0	1.733,0	1.245,0	26,0	61,4	1,4	6	17	24	10	2.004
max	11,9	2.040,0	63,7	1.479,0	2.683,0	288,0	2.775,0	2.523,0	44,6	88,7	2,2	7	23	31	12	2.004