UNIVERSIDADE FEDERAL DO CEARÁ - UFC
DEPARTAMENTO DE COMPUTAÇÃO

DEPARTAMENTO DE COMPUTAÇÃO

Modelagem de Preços de Combustíveis: Regressão e Clusterização Aplicadas

Daniel Oliveira
Manolidis Efstratios
Melissa Felipe

Tópicos

- Qual o foco do trabalho?
- Quais modelos iremos utilizar?
- Fonte de dados
- Metodologia
  - Mineração
  - Pré processamento
  - Estimação de densidade / Clusterização
  - Avaliação de outliers
  - Regressão
  - InferÊncia bayesiana
  - Avaliação do Modelo
- Conclusão

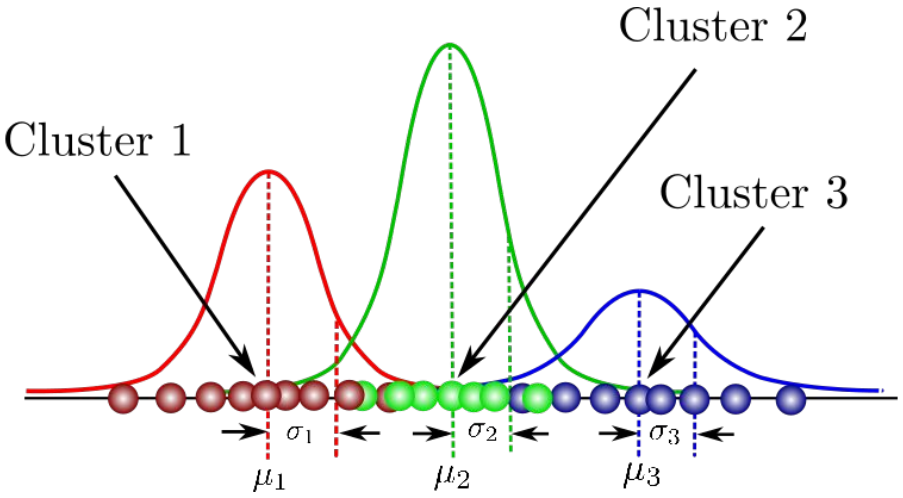
Qual o foco do trabalho ?

Realizamos uma Estimação de densidade/clusterização anual no dataset de preços de gasolina, considerando apenas a cidade de Fortaleza/CE e analisando a presença de outliers. Além disso, aplicamos um modelo de regressão para prever os preços futuros. Por fim, avaliamos o desempenho de cada modelo, garantindo uma análise abrangente dos resultados.

Quais modelos iremos utilizar?

Gaussian Mixture Model - GMM :

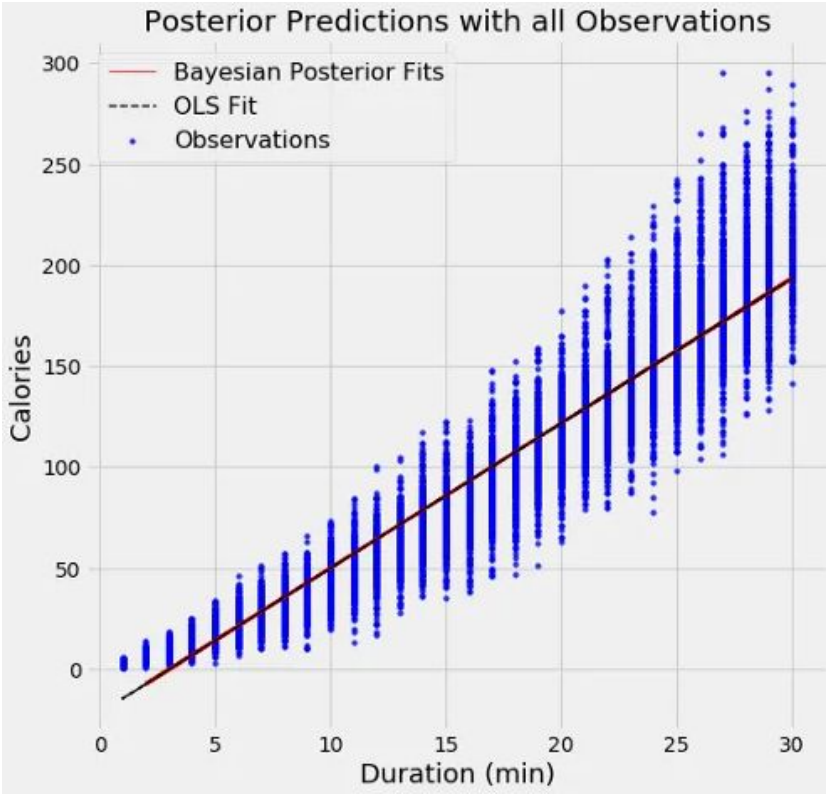
- Modela uma distribuição de probabilidades como uma combinação de várias distribuições gaussianas;
- A clusterização realizada pelo GMM é baseada na densidade, onde uma região de objetos com alta densidade é cercada por áreas de baixa densidade.



Quais modelos iremos utilizar?

Regressão Linear Bayesiana:

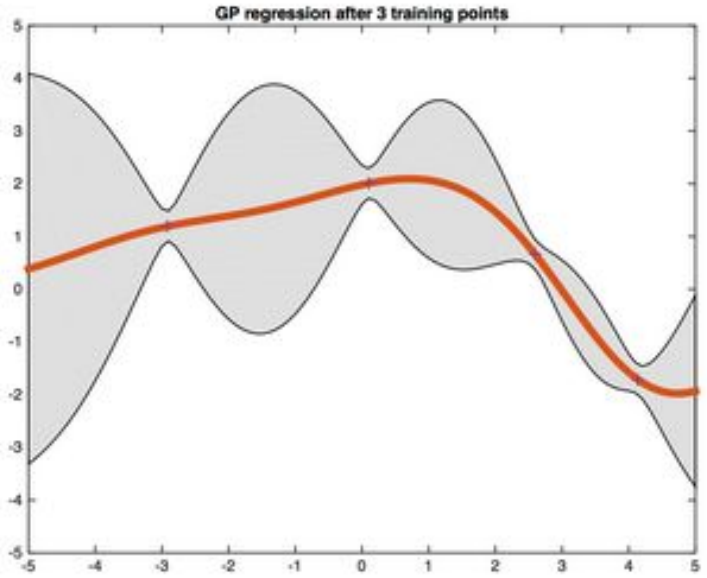
- Utiliza distribuições de probabilidade em vez de estimativas pontuais;
- A resposta ***y*** não é estimada como um único valor, mas sim como uma variável proveniente de uma distribuição de probabilidade.



Quais modelos iremos utilizar?

Gaussian Processes:

- Um processo gaussiano é uma distribuição de probabilidade sobre funções possíveis que se ajustam a um conjunto de pontos;
- Uma vantagem é a simplicidade matemática e a facilidade de manipulação, especialmente por sua fundamentação na extensão da distribuição normal multivariada.



Fonte de dados

gov.br

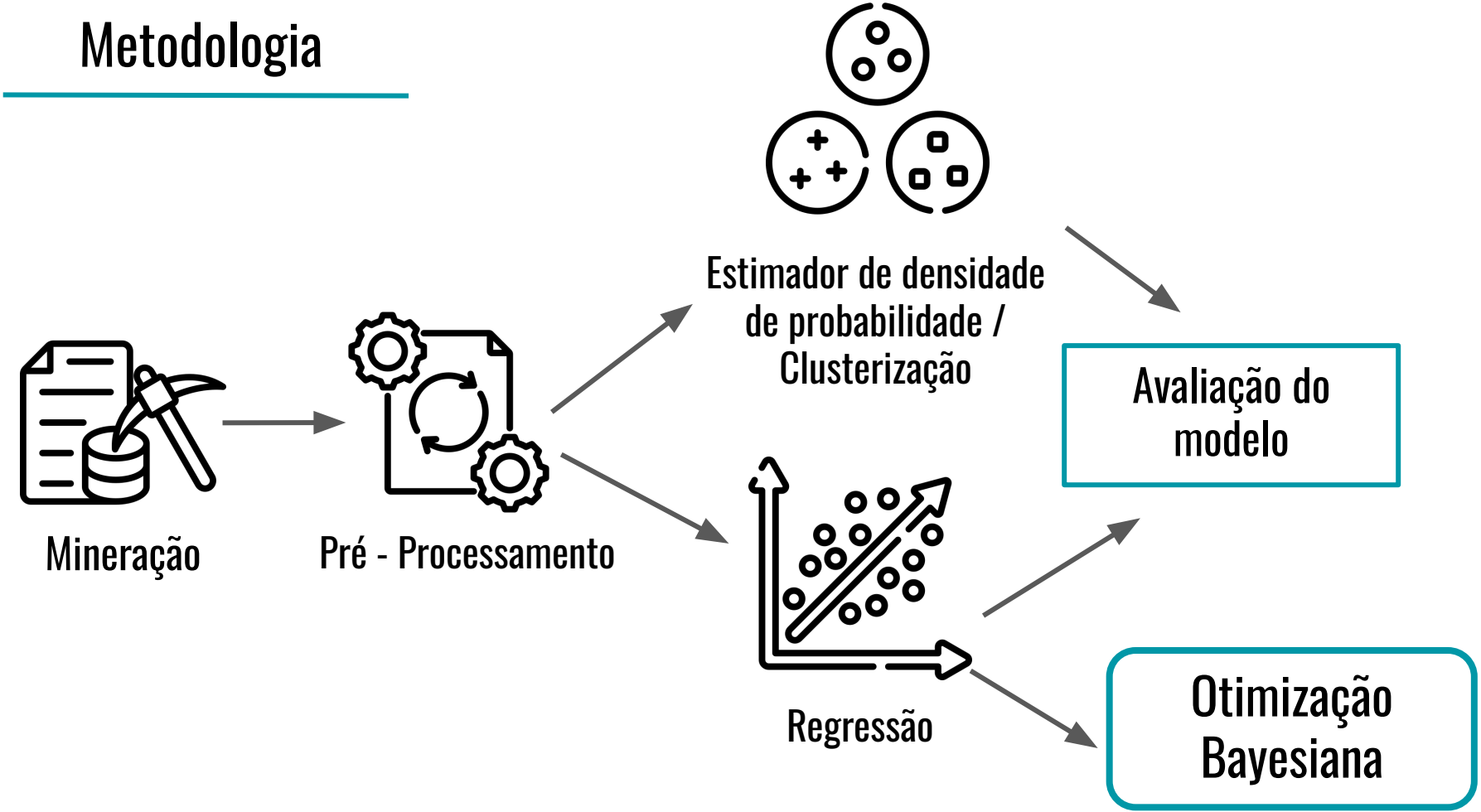
Estado - Sigla	Municipio	Revenda	CNPJ da Revenda	Nome da Rua	Numero Rua	Complemento	Bairro	Cep	Produto	Valor de Venda	Unidade de Medida	Bandeira	data	end
CE	FORTALEZA	DIAS COMERCIO DE DERIVADOS DE PETROLEO LTDA	08.509.458/0001-82	RODOVIA DOUTOR MENDEL STEINBRUCH	6750	NaN	ARACAPE	60765-005	GASOLINA	3.89	R$ / litro	RAIZEN	2016-01-05	ROD DOU MEI STEINBR 6 ARA
CE	FORTALEZA	IGUATEMI DERIVADOS DE PETROLEO LTDA	07.304.199/0003-52	AVENIDA OSORIO DE PAIVA	6800	NaN	SIQUEIRA	60720-001	GASOLINA	3.89	R$ / litro	IPIRANGA	2016-01-06	AVE OSORI PAIVA, 6 SIQUE F
CE	FORTALEZA	SOBRAL & PALACIO PETROLEO LTDA	07.240.641/0002-43	AVENIDA 13 DE MAIO	233	0	FATIMA	60040-530	GASOLINA	3.89	R$ / litro	RAIZEN	2016-01-05	AVENI DE MAIO - FAT FORTA
CE	FORTALEZA	,SOUSA & RIBAS COMÉRCIO DE COMBUSTÍVEIS LTDA.	07.796.363/0001-24	AVENIDA HERACLITO GRAÇA	20	NaN	CENTRO	60140-060	GASOLINA	3.82	R$ / litro	SP	2016-01-05	AVE HERAC GRAÇA CENT FORTA

Fonte de dados

yahoo! finance

Date	Open	High	Low	Close
2015-02-19 00:00:00+00:00	2.8345	2.8638	2.819200	2.8340
2015-02-20 00:00:00+00:00	2.8629	2.8808	2.849500	2.8633
2015-02-23 00:00:00+00:00	2.8641	2.8976	2.859700	2.8641
2015-02-24 00:00:00+00:00	2.8755	2.8822	2.853400	2.8747
2015-02-25 00:00:00+00:00	2.8547	2.8794	2.828100	2.8570
...	...	...	...	...
2025-02-13 00:00:00+00:00	5.7644	5.7982	5.753031	5.7644
2025-02-14 00:00:00+00:00	5.7669	5.7984	5.707000	5.7669
2025-02-17 00:00:00+00:00	5.7012	5.7290	5.693500	5.7012
2025-02-18 00:00:00+00:00	5.7121	5.7258	5.674600	5.7121
2025-02-19 00:00:00+00:00	5.6926	5.7324	5.676700	5.7277

Metodologia



Mineração

- Realizamos a uma mineração de dados mapeando cada cep com suas respectivas latitudes e longitudes;
- Desenvolvemos um script em Python utilizando a biblioteca Selenium, que automatiza a busca de coordenadas no Google Maps;
- O script python acessa a plataforma, pesquisa os endereços e extrai as coordenadas diretamente da URL, salvando os resultados em um novo arquivo CSV.

Pré-Processamento

- Incorporamos o preço do dólar dos últimos 10 anos e fizemos um merge com o nosso dataset;
- Restringimos as coordenadas geográficas ao Estado do Ceará;
- Filtramos o dataset para recuperar somente os postos(cnpj) que estão presentes em todos os anos;
- Realizamos um agrupamento por ano e por posto;
- Criamos novas variáveis.

Variáveis - Existentes

cnpj	Identificação da empresa.
data	Data do registro
latitude	Coordenada geográfica de latitude
longitude	Coordenada geográfica de longitude
bandeira	Identificação da bandeira do posto
dolar	Cotação do dólar correspondente ao período do registro.
preco	Valor associado ao produto ou serviço analisado
ano	Ano associado a data correspondente

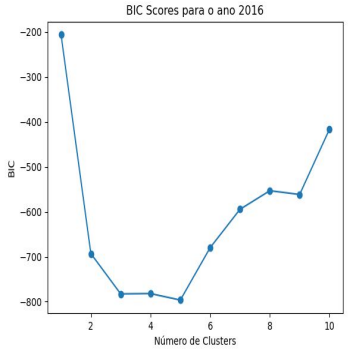
Variáveis - Criadas - Clusterização

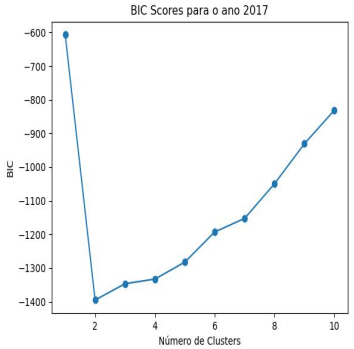
preco_ajustado	Preço ajustado de acordo com a cotação do dólar. (preço/dólar)
media_preco	Média do preço da gasolina
minimo_preco	Mínimo do preço da gasolina
maximo_preco	Máximo do preço da gasolina
desvio_padrao_preco	Desvio padrão da gasolina
media_dolar	Média do valor do dólar
media_preco_ajustado	Média do preço ajustado.
minimo_preco_ajustado	Mínimo do preço ajustado
maximo_preco_ajustado	Máximo do preço ajustado
amplitude_preco	Amplitude do preço da gasolina

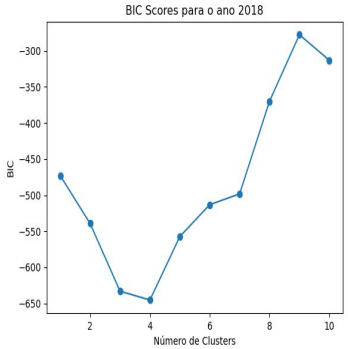
Estimação de Densidade / Clusterização

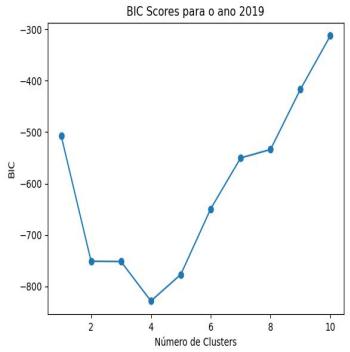
- Modelo de GMM para cada ano;
- Critério de escolha: minimização do BIC;
- Retreino por ano com o menor BIC;
- Testamos diferentes números de clusters [1, 10];
- Calculamos a log-verossimilhança para detectar os outliers, considerando os 5% menores valores como outliers.

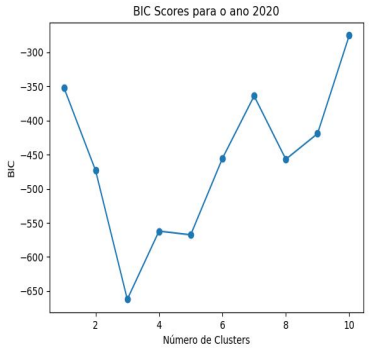
Bic score por ano

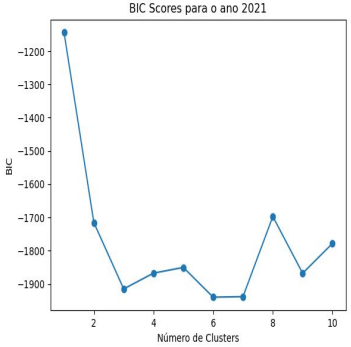


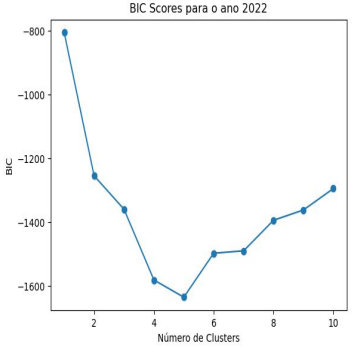


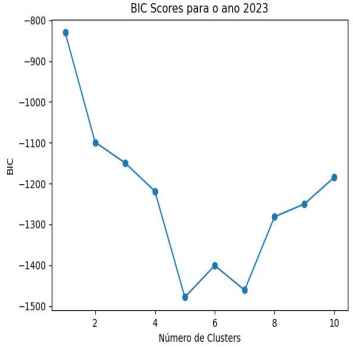


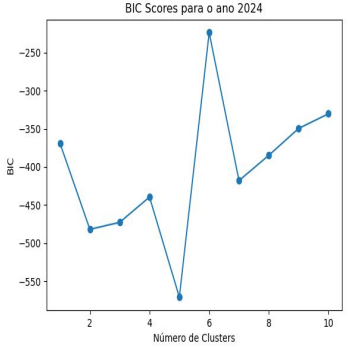






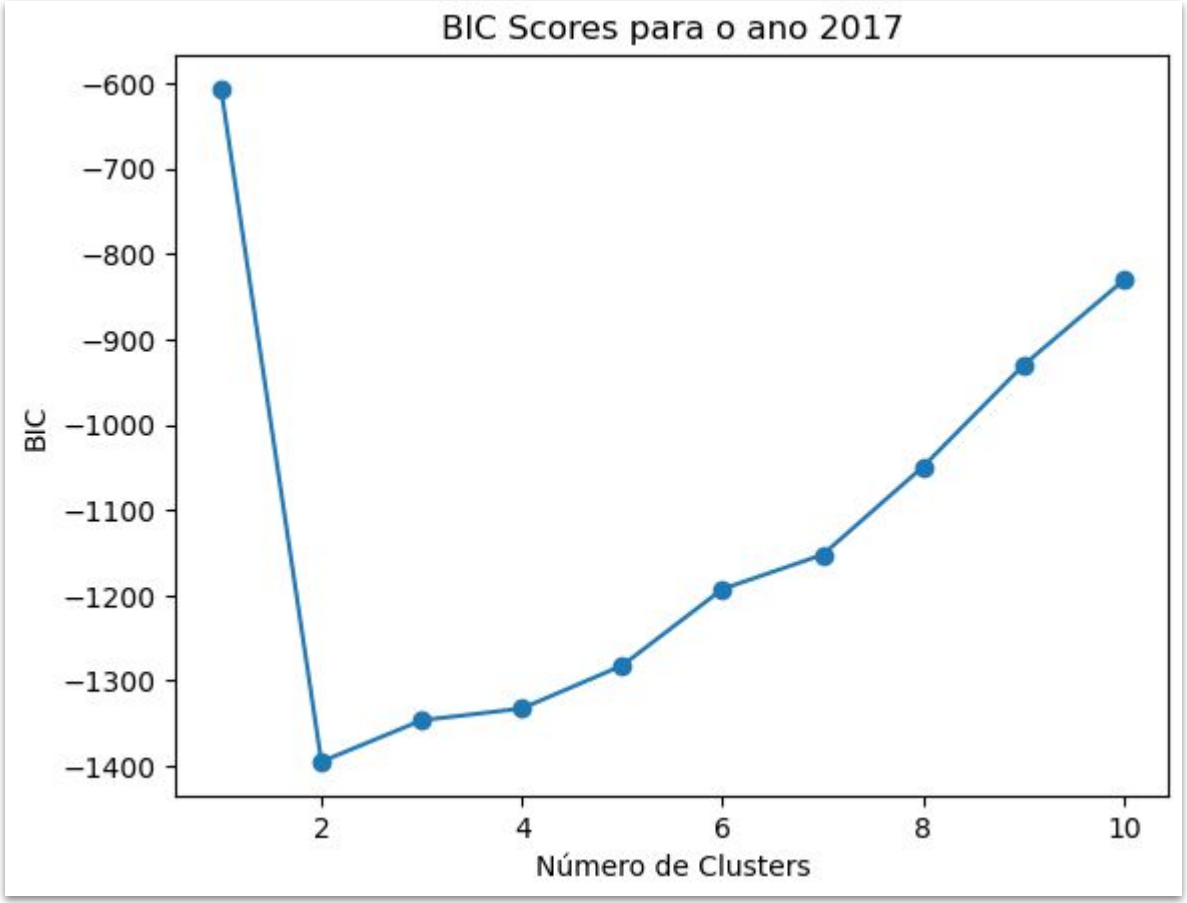






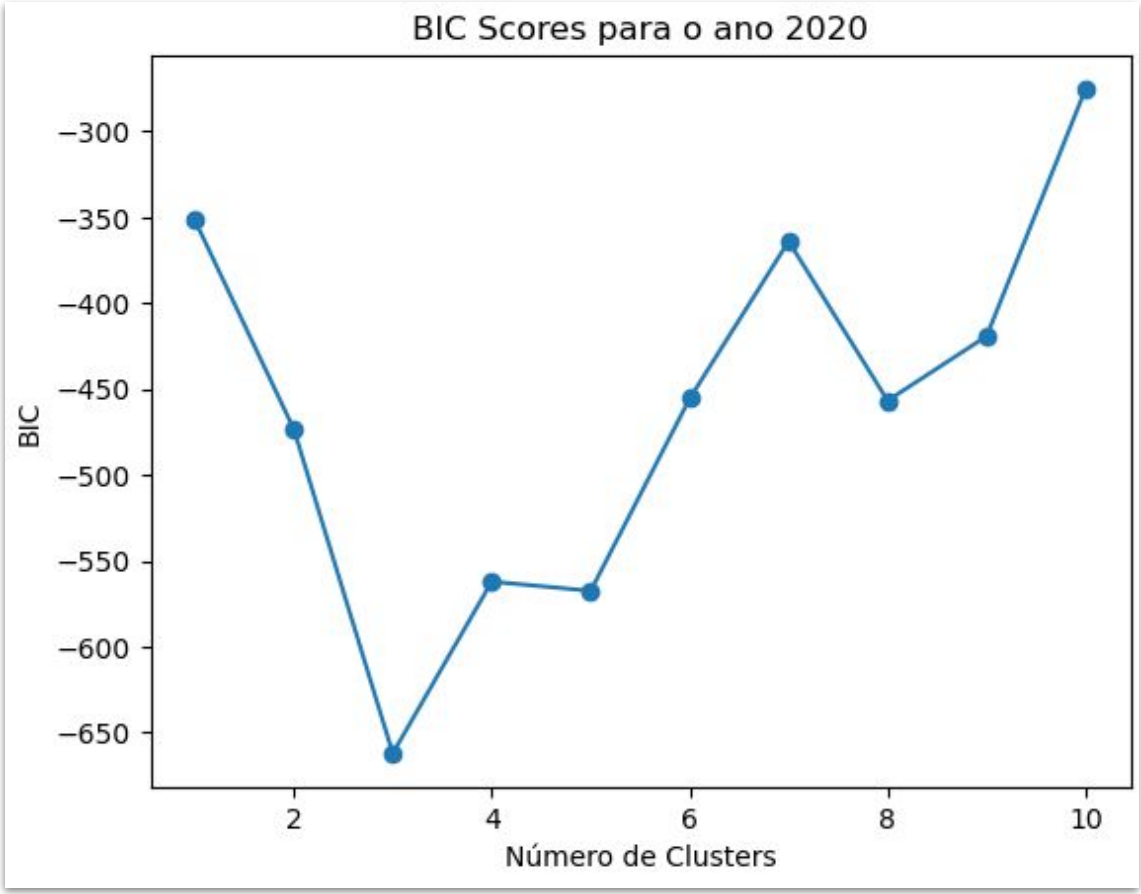
Bic score por ano

Definiu o número ideal de cluster = 2, para o ano de 2017



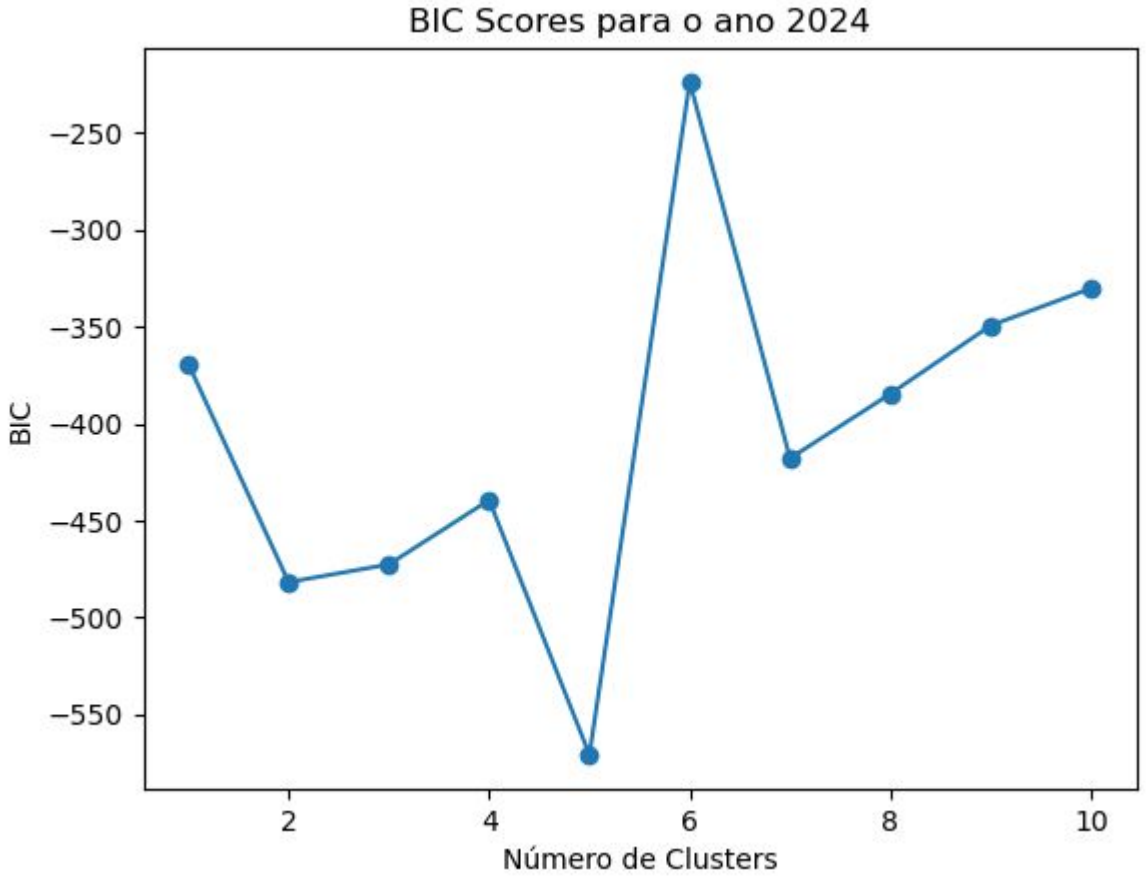
Bic score por ano

Definiu o número ideal de cluster = 3, para o ano de 2020



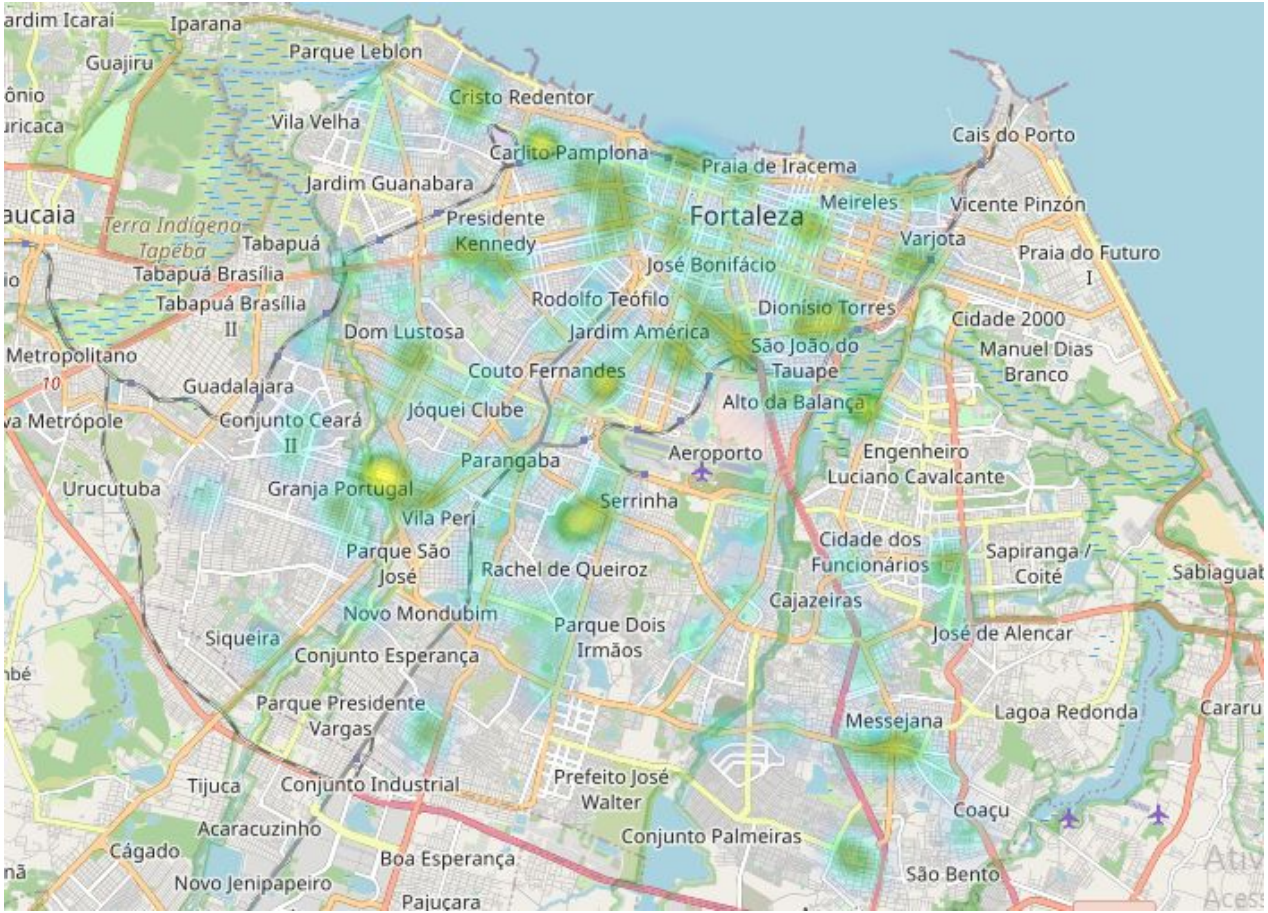
Bic score por ano

Definiu o número ideal de cluster = 5, para o ano de 2024



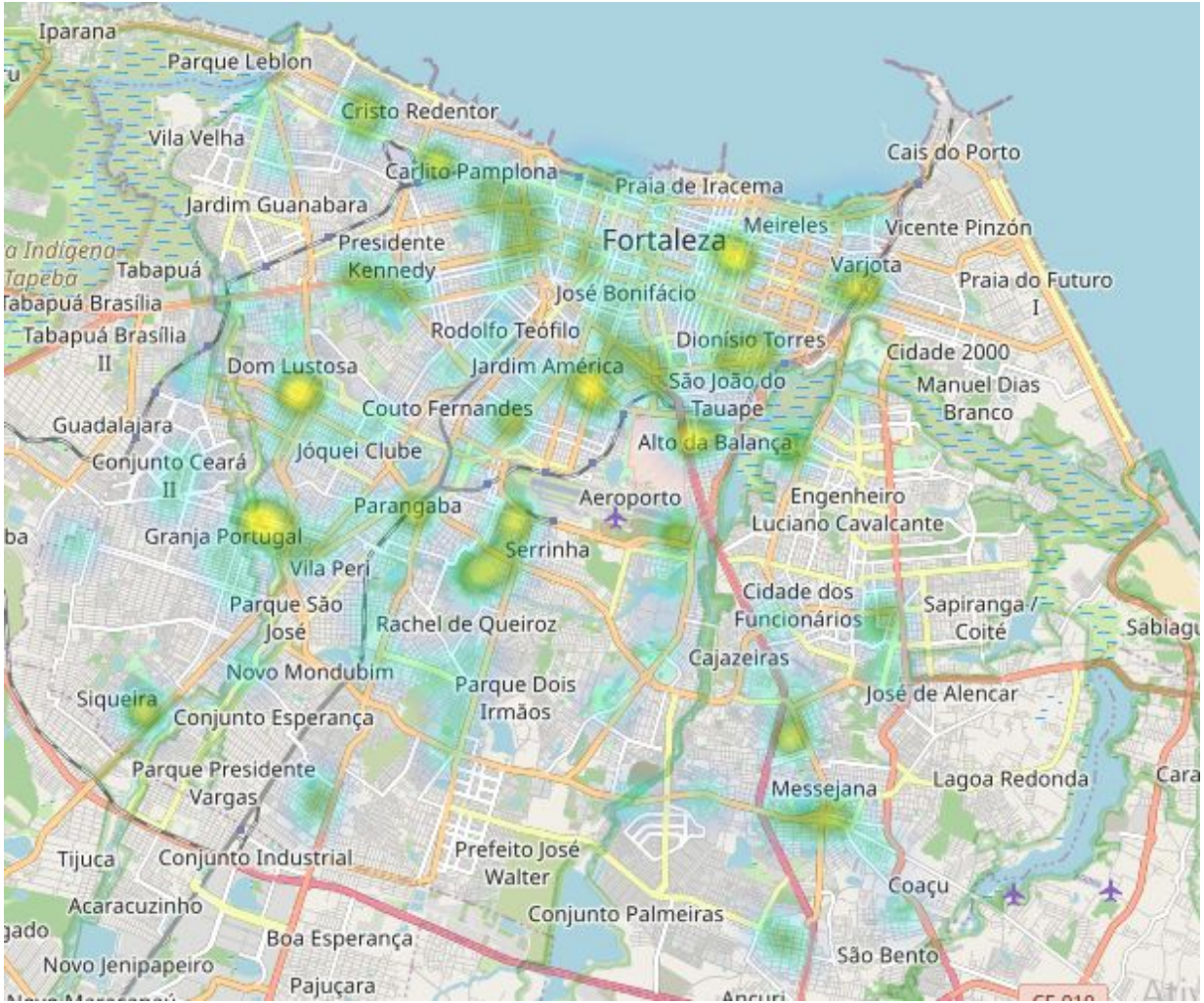
Mapa de Calor

Distribuição do preço médio da gasolina no ano de 2017



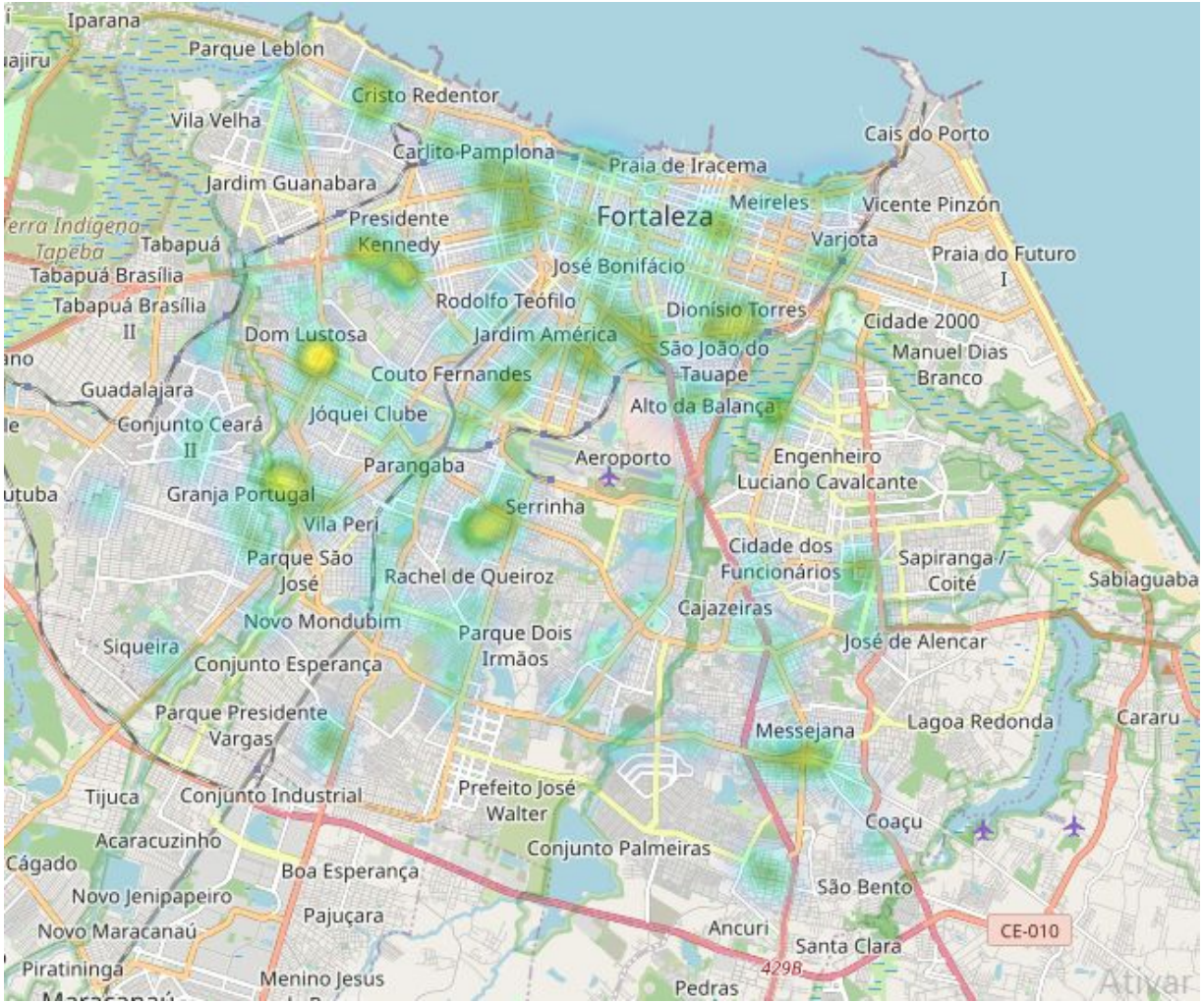
Mapa de Calor

Distribuição do preço médio da gasolina no ano de 2020

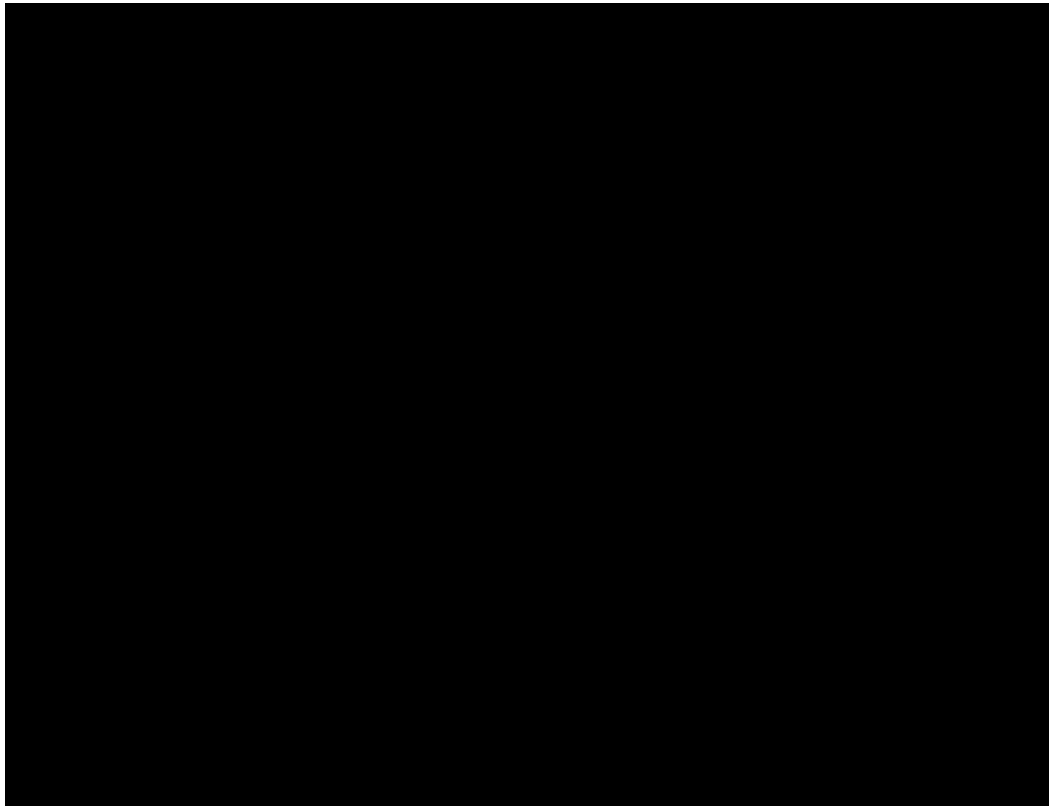


Mapa de Calor

Distribuição do preço médio da gasolina no ano de 2024



HeatMap - Preço da gasolina

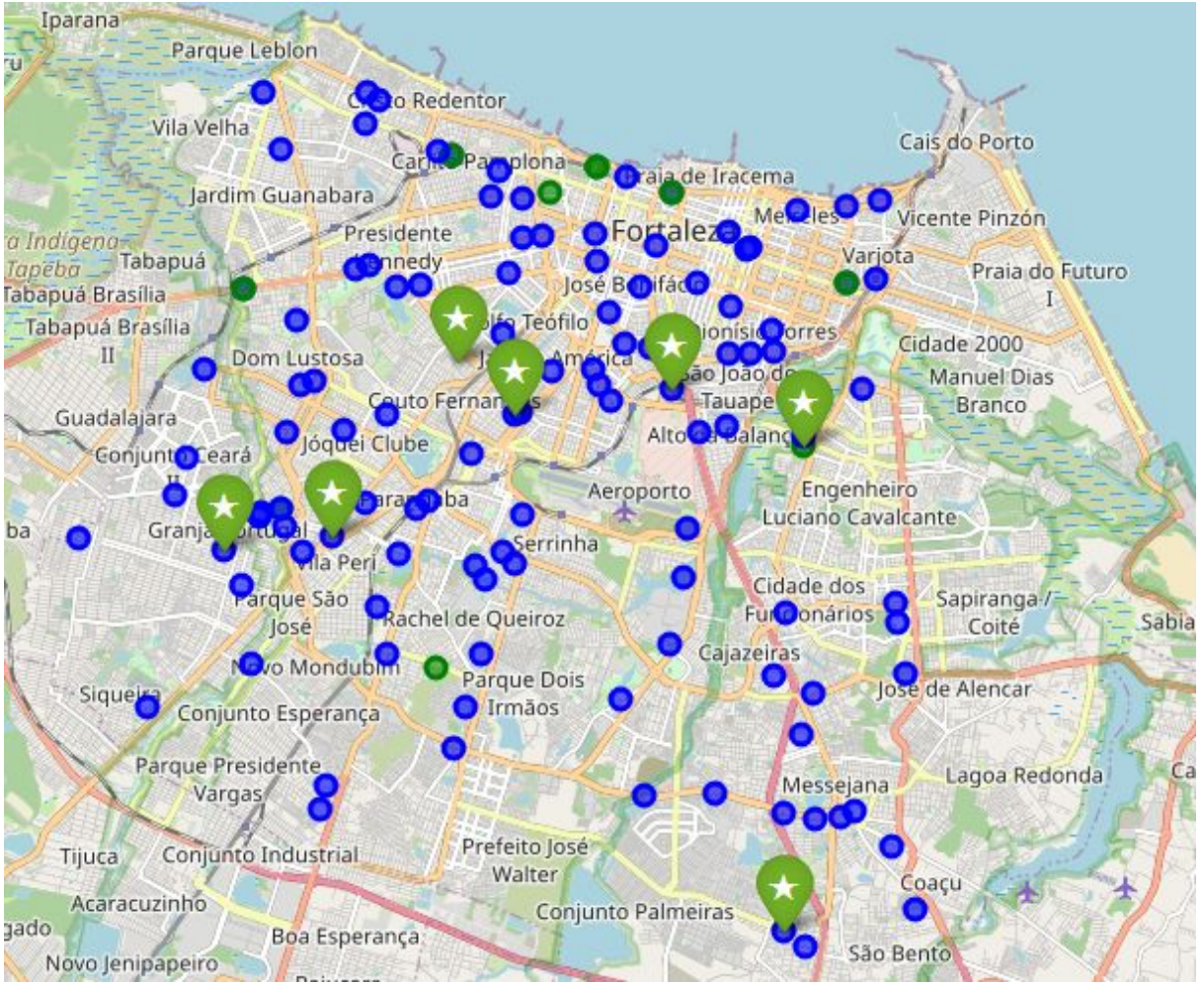


Cluster

Clusterização ao longo de todos os anos (2016-2024)

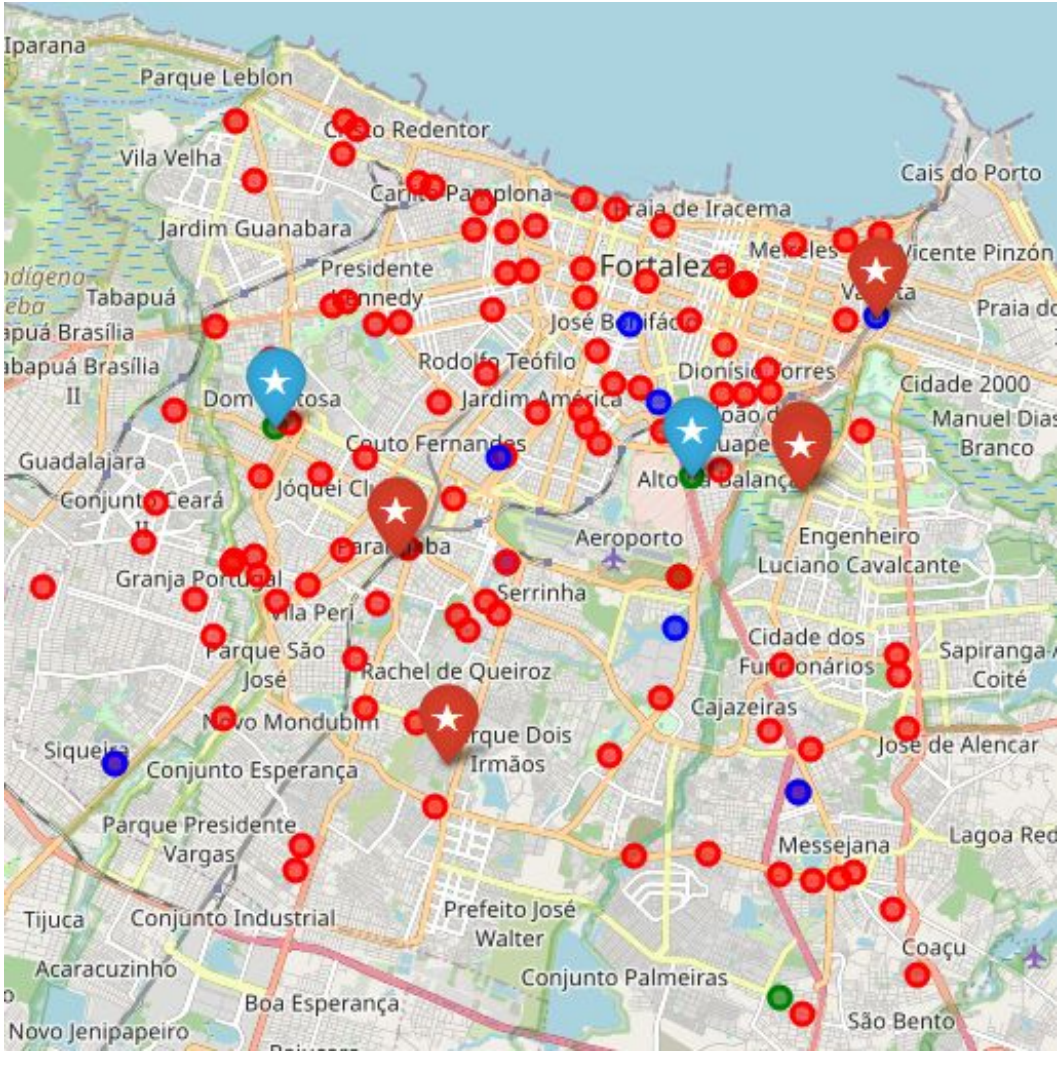
Cluster

Clusterização para o ano de 2017



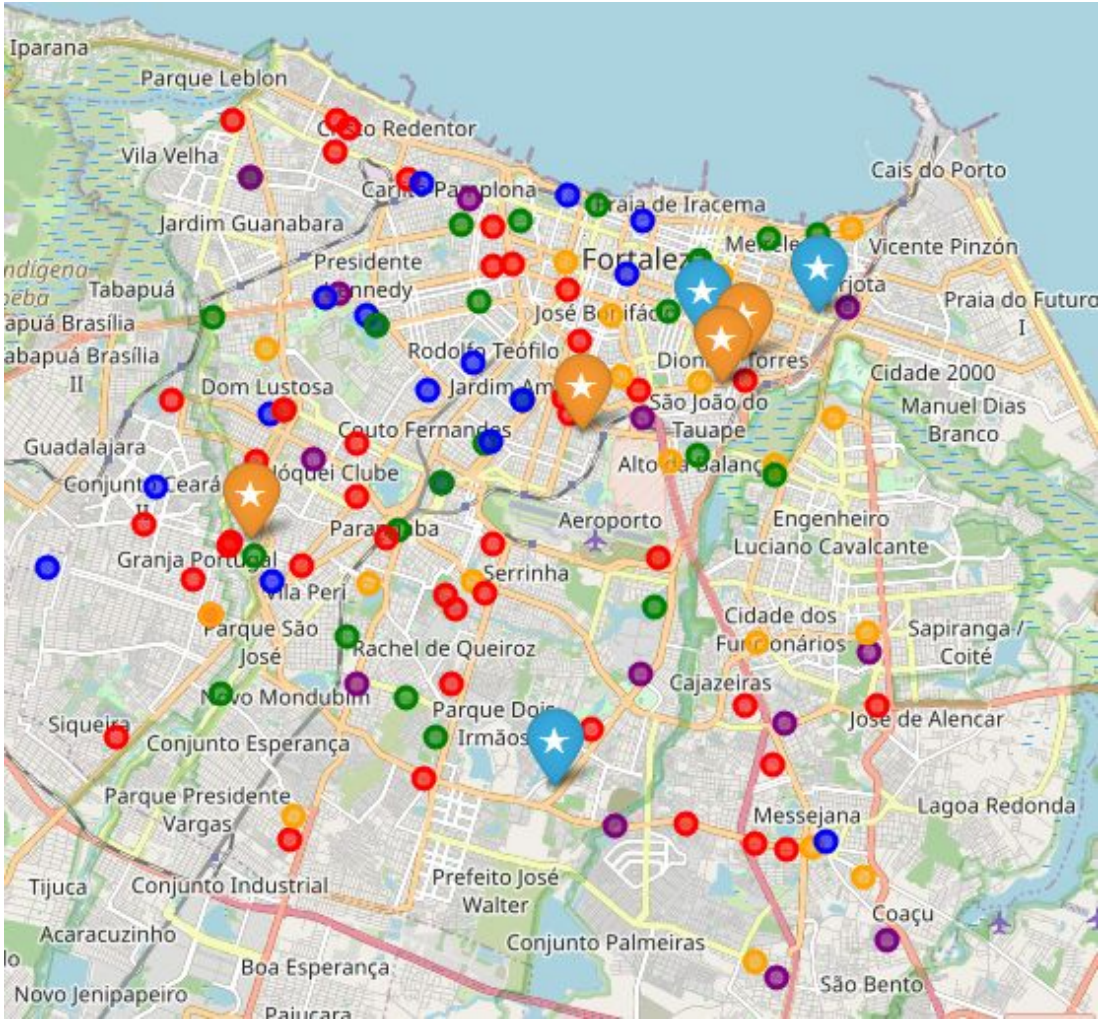
Cluster

Clusterização para o ano de 2020



Cluster

Clusterização para o ano de 2024



Regressão - Regressão Linear Bayesiana

- Dividimos os dados em treino (01/01/2021 - 30/09/2024) e teste (01/10/2024 - 27/12/2024);
- Criamos novas variáveis;
- Normalizamos os dados;
- Treinamos o modelo utilizando a biblioteca BayesianRidge, depois testamos o modelo;
- Definimos um intervalo de confiança de 95%.

Variáveis - Criadas - Regressão

dias_desde_inicio	Permite capturar padrões temporais que influenciam o preço ao longo do tempo e que está calculando quantos dias se passaram desde a data mínima presente no conjunto de dados
mes	Auxiliam na análise de sazonalidade mensal
dia_da_semana	Auxiliam na análise de sazonalidade semanal
preco_medio_posto_7d, preco_medio_posto_14d, preco_medio_posto_30d, preco_medio_posto_60d e delta_preco_posto_7d	Representa a variação do preço médio em um posto de gasolina nos últimos 7 dias em comparação com os últimos 30 dias e 60 dias, que ajudam a suavizar flutuações diárias, permitindo identificar tendências de variação nos preços ao longo do tempo
cnpj_media_preco	Captura o comportamento agregado dos preços por posto

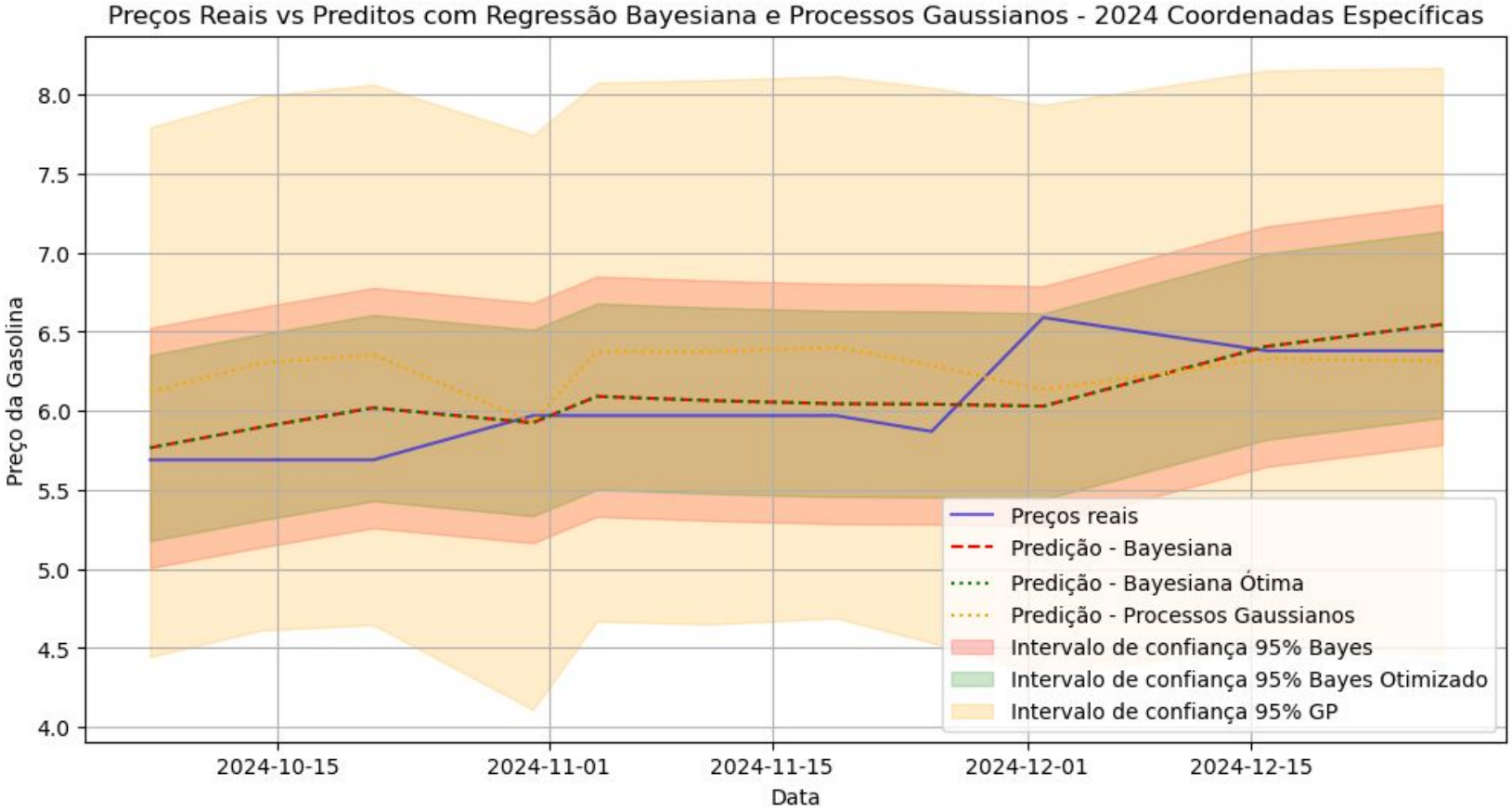
Regressão - Processos Gaussianos

- Dividimos os dados em treino (01/01/2021 - 30/09/2024) e teste (01/10/2024 - 27/12/2024);
- Criamos novas variáveis;
- Normalizamos os dados;
- Utilizamos uma combinação de 3 Kernels:
  - RBFKernel: Captura variações suaves
  - LinearKernel: Captura tendências lineares
  - PeriodicKernel: Captura sazonalidade (ciclo entre 30 e 365 dias)
- Treinamos e testamos o modelo utilizando a biblioteca GPModel.

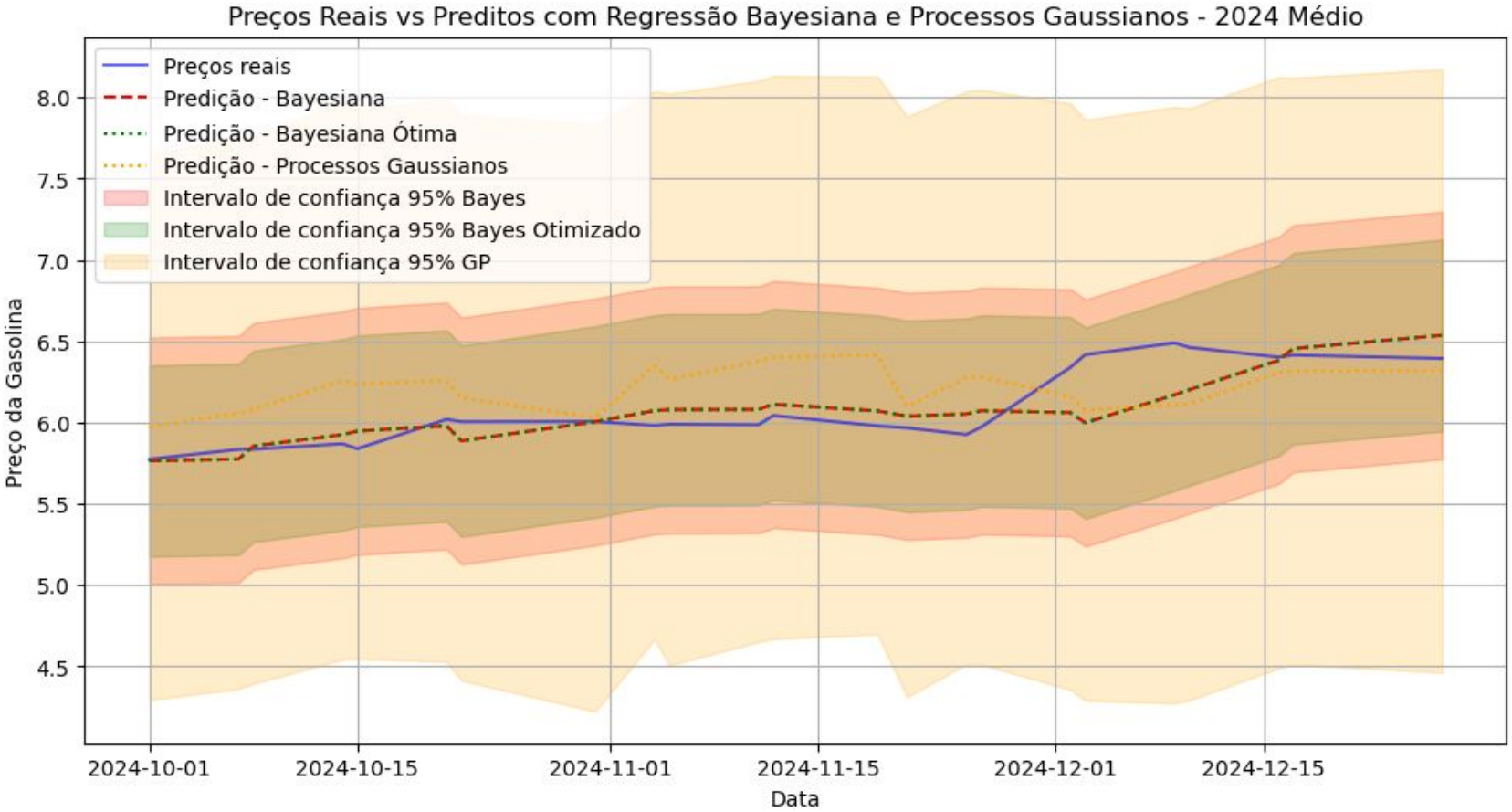
Inferência Bayesiana

- Definimos a função objetivo baseada na Densidade preditiva de log negativo (NLPD);
- Definimos um limite máximo de busca dos hiperparâmetros [10^-6, 10^-1];
- Utilizamos a biblioteca BayesianOptimization;
- Treinamos os modelos com os melhores hiperparâmetros.

Regressão - Caso específico



Regressão - Média



Avaliação dos modelos

	MSE	RMSE	NLPD
Regressão Linear Bayesiana	0.0450	0.2122	0.1206
Regressão Linear Bayesiana com Otimização	0.0450	0.2122	-0.0348
Processos Gaussianos	0.1054	0.3248	

Conclusão

- Conseguimos realizar um estimação de densidade/Clusterização utilizando GMM;
- A Regressão Linear Bayesiana demonstrou um resultado melhor em relação ao modelo Processos Gaussianos;
- Como trabalho futuro, pretende-se aprofundar a investigação sobre os outliers, buscando compreender suas causas e impactos na modelagem dos preços.

Obrigado (a) !