UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

1. Tema

Compreendendo fatores que podem influenciar o preço dos combustíveis dos estados brasileiros. Período de 2004 a 2021.

2. Equipe

Os reveladores das faces dos dados

● Caique Salvador Noboa, 1904949, caique2noboa, noboa@alunos.utfpr.edu.br, Eng. Comp, UTFPR;

● Maria Gabriela Rodrigues Policarpo, 2450151, mariagabrielapolicarpo, mariagabrielapolicarpo@alunos.utfpr.edu.br, BSI, UTFPR;

● Moisés Bryan Carneiro Roja, 2243814, moisesbryan, moisesbryan@alunos.utfpr.edu.br , Eng. Mecatrônica, UTFPR;

● Bruno Souza Zabot, 2238039, bzabot, zabot@alunos.utfpr.edu.br, Eng. Mecânica, UTFPR;

3. Introdução

O valor do preço dos combustíveis afeta muito a população brasileira, desde o preço de locomoção até o preço dos bens de consumo.

Buscamos entender melhor quais os fatores que influenciam o preço dos combustíveis, e encontramos uma base de dados com preços de 2004 a 2021, medido semanalmente para cada estado brasileiro.

Realizamos as seguintes perguntas de pesquisa:

- O etanol é mais barato nos estados que o produzem?
- A pandemia influenciou o preço dos combustíveis no Brasil?
- O lucro bruto dos postos de combustíveis está relacionado com a desigualdade social no estados?
- O que ocorreu com o poder de compra de combustíveis ao longo do tempo?

4. Processamento de dados

Fizemos o processamento de dados para as seguintes tabelas:

- **Tabela principal - Preço dos combustíveis dos estados brasileiros de 2004 a 2021**:
    - **Fontes dos dados:** A tabela foi obtida no *Kaggle*.
    - **Procedimentos de limpeza e processamento:** A tabela com os dados de preço médio de revenda e distribuição dos estados entre 2004 a 2021 foi formatada pelo arquivo "*Leitura_e_Procedimento de_Limpeza*". Nesse arquivo, primeiramente, foram renomeadas as colunas que possuíam acento e "*ç*", além de adequar o nome dos óleos (que havia sido cadastrado com dois tipos, com e sem acento) para ficar sem acento em todos os casos. Finalizando as formatações textuais, todos os estados, que estavam escritos por extenso, foram substituídos pelas respectivas siglas.

      Além disso, foi observado que seis colunas numéricas estavam com tipo de dado 'objeto', isso também foi ajustado para numérico. Além disso, foram checados os tipos de combustíveis analisados e excluídos da tabela todos os gasosos.

      Após isso a tabela foi exportada para CSV com o nome *'br_oil_prices_formatado.csv'.*
- **População por estado**
    - **Fontes dos dados:** A tabela foi obtida através de uma base de dados no Big Query, e foi utilizada a seguinte query para obter os dados:
      SELECT ano, sigla_uf, SUM(populacao) as populacao FROM `basedosdados.br_ibge_populacao.municipio` GROUP BY sigla_uf, ano;
      Então foi salvo o resultado em CSV, e não foi necessária limpeza nos dados.
- **PIB dos estados**
    - **Fontes dos dados:** A tabela foi obtida através deste link, porém as colunas eram: "*Estado, 2004, 2005.... 2019*", e o valor das células era o valor do PIB por estado e por ano. Então foi necessário utilizar o comando *melt* do Pandas para realizar a conversão para possuir as colunas: "*Estado, ano, pib*".
    - **Procedimentos de limpeza e processamento:** Foi convertido o nome do estado para a sigla, para seguir o padrão da tabela de População por estado.
- **Preço do barril do petróleo mundial**
    - **Fontes dos dados:** A tabela foi obtida no *Kaggle*, os dados abrangiam até o mês de janeiro de 2021, então de fevereiro a maio de 2021 os dados foram coletados de forma manual.
    - **Procedimentos de limpeza e processamento:** Foi feita a conversão de coluna para tipo data, filtro do período de datas necessário, média mensal da coluna preço do petróleo.

- **Preço do dólar**
    - **Fontes dos dados:** A tabela foi obtida através deste site que possui histórico da cotação do dólar.
    - **Procedimentos de limpeza e processamento:** Na parte de limpeza, colunas que tinham acento no nome foram modificadas, a numeração de valores estava em vírgula e foi trocada por ponto, foi feita conversão de colunas de tipo string para float. Como não havia uma coluna para média diária, foi criada a partir da média do valor mínimo e máximo do dólar

naquele dia, em seguida foi feita uma coluna com média mensal. As datas foram convertidas para o formato internacional.

- **Salário mínimo**
  - **Fontes dos dados:** A tabela foi preenchida manualmente, compreende 3 colunas: a data que o salário se tornou vigente, a data final e o salário mínimo em reais.
  - **Procedimentos de limpeza e processamento:** Por ser uma tabela pequena criada por nós mesmos, não foi necessário realizar a limpeza dos dados.
- **Estados produtores de etanol**
  - **Fontes dos dados:** A tabela foi obtida através deste site, compreende 3 colunas: a data, estado brasileiro e a produção em metros cúbicos.
  - **Procedimentos de limpeza e processamento:** Na parte de limpeza, colunas que tinham acento no nome foram modificadas, a numeração de valores estava em vírgula e foi trocada por ponto, foi feita conversão de colunas de tipo string para float. E também foi adicionado a coluna da sigla do estado, para padronizar com as outras tabelas obtidas.

5. Resultados

a) O etanol é mais barato nos estados que o produzem?

Descrição: Esta pergunta tinha como objetivo analisar se o combustível etanol tende a ser mais barato nos estados que produzem etanol.

Hipóteses: O etanol é mais barato nos estados que o produzem.

Dados: Tabela de produtores de etanol por estado, e preço do combustível por estado. Toda esta hipótese foi feita com o notebook "hipotese_etanol_mais_barato".

Modelo: Para responder essa pergunta verificamos primeiramente os dados visualmente para entender se havia alguma relação clara, e em seguida fizemos a correlação entre as colunas de produção e o preço médio do etanol.

Análises: Analisando a tabela, vemos:
Maiores produtores de etanol de 2004 à 2021 (em metros cúbicos):
SP 588653,144875
GO 191267,683087
MG 121116,133314
MS 110893,939042
MT 76172,782137

Menores preços de etanol (em reais):
SP: 1.911203
MT: 2.076379
PR: 2.082645
GO: 2.126071
MG: 2.223109

E vemos que 4 estados se repetem, SP, GO, MG e MT. Ainda na visualização de dados, criamos o mapa de estados brasileiros e sua produção de etanol:

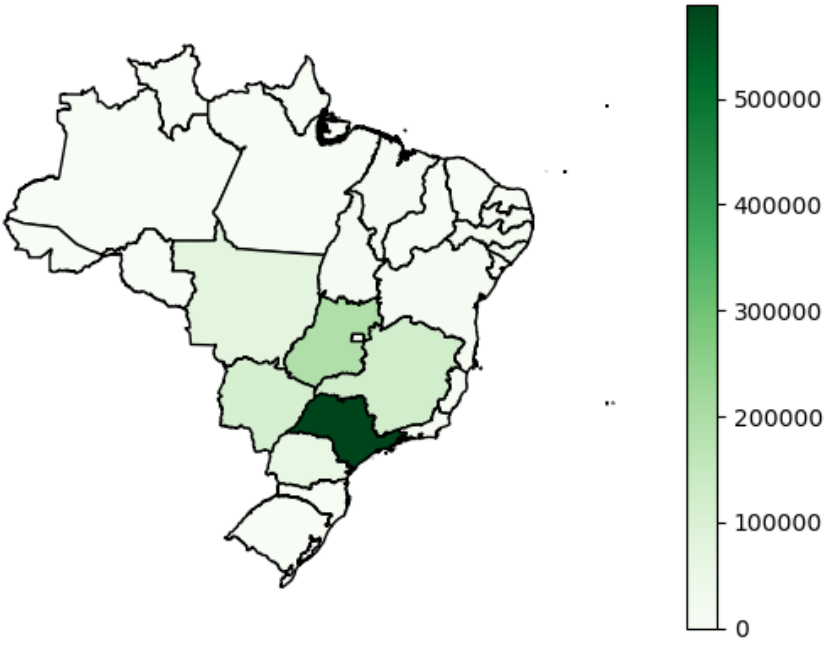


Figura 1: Mapa do Brasil, produtores de Etanol por estado

Agora, analisando o mapa de preços de etanol:

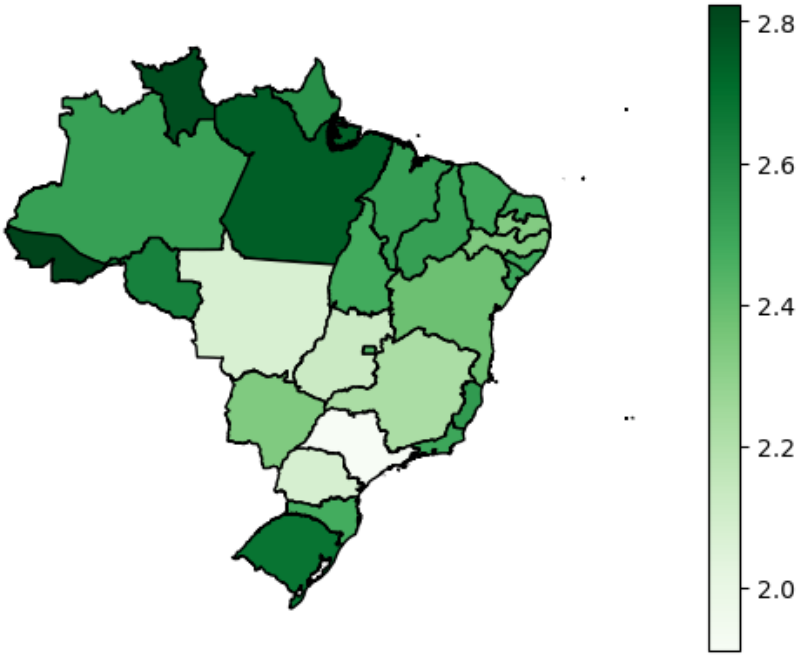


Figura 2: Mapa do Brasil, preço do Etanol por estado

Resultados: Foi visto semelhança visual entre os maiores produtores de etanol e seu preço mais baixo, então seguimos para o teste estatístico.

A correção foi de "-0.7468672742666072" com o p-value de "-0.00000764721568656832". Por conta do valor alto de correlação e o p-value muito baixo, podemos concluir que a hipótese é verdadeira.

Limitações:

- Não tínhamos dados de exportação e de importação de Etanol por estado, o que poderia influenciar no preço praticado no estado.
- Não tínhamos dados de consumo, para tentar normalizar os dados de produção.

Trabalhos futuros: Utilizar dados socioeconômicos de cada estado para ter mais segurança na hipótese. Além de buscar dados de consumo, e de importação e exportação.

b) A pandemia influenciou o preço da gasolina comum no Paraná?

Descrição: Esta pergunta tinha como objetivo analisar se a pandemia diminuiu ou aumentou o preço dos combustíveis.

Hipótese: O preço da gasolina comum no Paraná diminuiu no pico da pandemia.

Dados: Para criar o modelo de regressão antes foi necessário formatar os dados, isso foi feito com o notebook “Ajustes_DataFrame_para_Regressao”. Inicialmente, essa base continha o dado de vários tipos de combustíveis em todos os estados brasileiros. Para aumentar o nível de detalhe, foi filtrado somente os valores de gasolina comum no Paraná e o intervalo das datas 05/2004 até 04/2021, que estava organizado semanalmente com o valor da respectiva semana no domingo.

Além disso, também foi tratada a base de dados do dólar, tendo feito a média entre o valor mínimo e máximo para obter o valor do dia e foi filtrado as datas de acordo com o range do valor da gasolina. Por último, foram feitos ajustes no DataFrame com os dados do valor do petróleo filtrando as datas.

Feito isso, essas três últimas bases foram agrupadas semanalmente no domingo de acordo com a média da semana. Esse DataFrame foi exportado e utilizado para a regressão no arquivo “Regressao_BrunoZ_FINAL”.

Modelo:

Inicialmente, a hipótese foi formulada pensando que no pico da pandemia e do lockdown a demanda pela gasolina comum diminuiu expressivamente, com esse cenário, os postos se sentiram obrigados a diminuir o preço da gasolina para vender mais. No entanto, o pico do lockdown não causou uma redução no preço do petróleo e do dólar diretamente visto que não há relação direta entre o lockdown e a demanda de petróleo/dólar.

Pensando nisso, foi criado um modelo de Machine Learning que, treinado com os dados da gasolina, petróleo e dólar antes da pandemia, previa o preço da gasolina para o ano de 2020. Esse valor previsto foi comparado ao valor real para 2020 e, caso houvesse uma disparidade entre os dois modelos, a hipótese estaria satisfeita.

Resultados:

No nosso primeiro modelo de regressão foram utilizados os dados da gasolina, petróleo e dólar entre 2004 até o final de 2019. Baseado nisso, com o notebook *‘Regressao_BrunoZ_FINAL’*, foram obtidos os seguintes valores de correlação:

- $R^2_{PETRÓLEO\text{-}GASOLINA} = 0{,}042$

- $R^2_{DÓLAR\text{-}GASOLINA} = 0{,}678$

Como a relação entre o petróleo e a gasolina obtidos são pequenos foi feita a tentativa de adicionar um delay na correlação entre o petróleo a gasolina, que também não foi bem-sucedida.

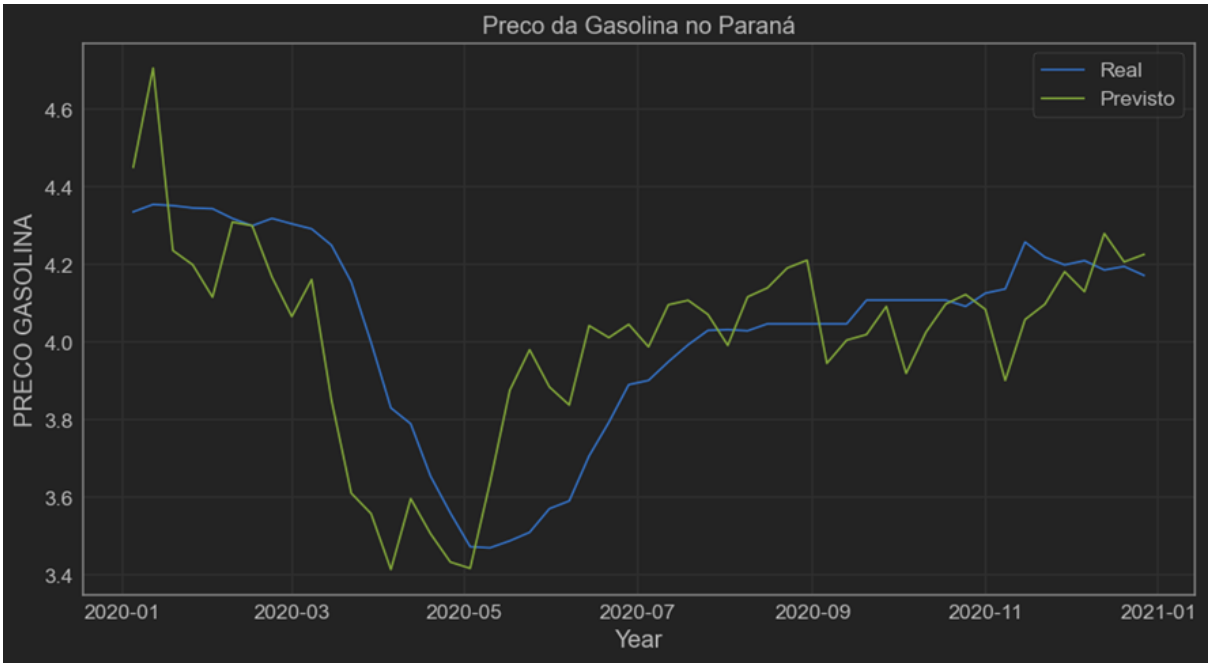
Então, baseado na seguinte notícia, datada de outubro de 2016, que cita que a Petrobras adotou uma nova política de preços baseado no mercado internacional, os dados para o modelo de regressão foram filtrados de 2017 até o final de 2019, com isso foram obtidos os seguintes valores:

- $R^2_{PETRÓLEO\text{-}GASOLINA} = 0{,}790$

- $R^2_{DÓLAR\text{-}GASOLINA} = 0{,}513$

- $R^2_{PETRÓLEO+DÓLAR\text{-}GASOLINA} = 0{,}881$

Com o modelo criado a partir do petróleo e dólar, com a gasolina como variável dependente, foi previsto o valor da gasolina para o ano de 2020 e comparado com o valor real da gasolina comum no Paraná.

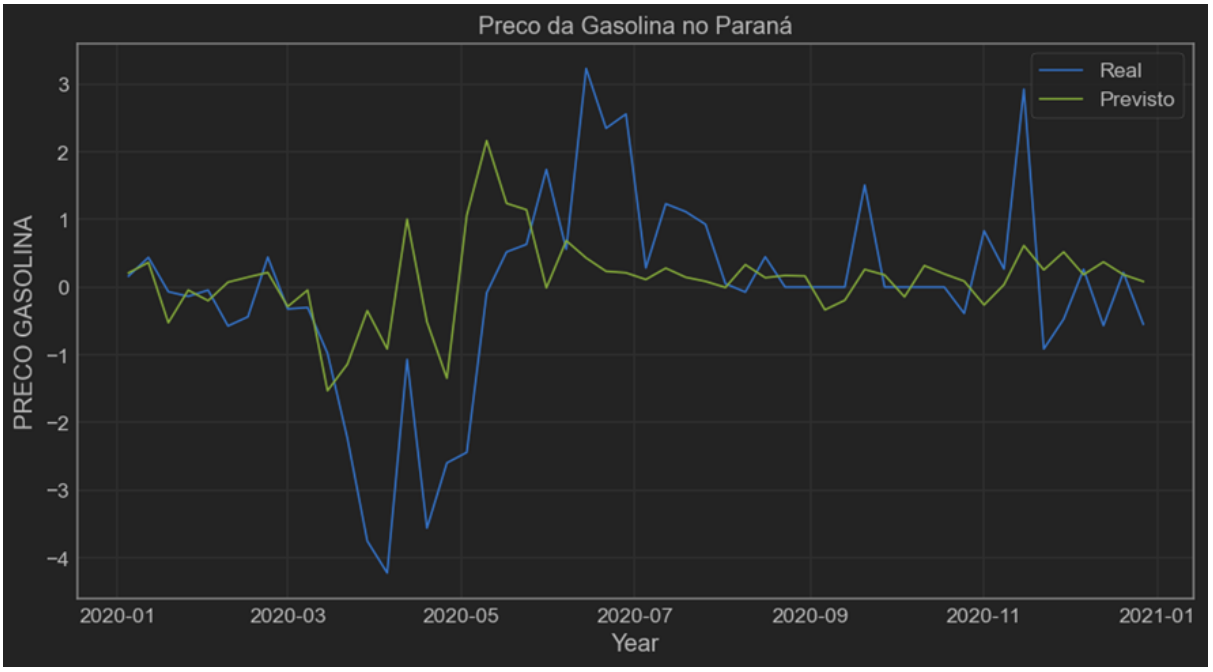


Na imagem acima percebe-se que o modelo previu bem o comportamento da gasolina comum no Paraná para o ano de 2020, no entanto, essa grande correlação entre os valores está mascarada pela inflação, portanto, foi necessário criar outro modelo, baseado na diferença percentual do combustível ao longo do tempo.

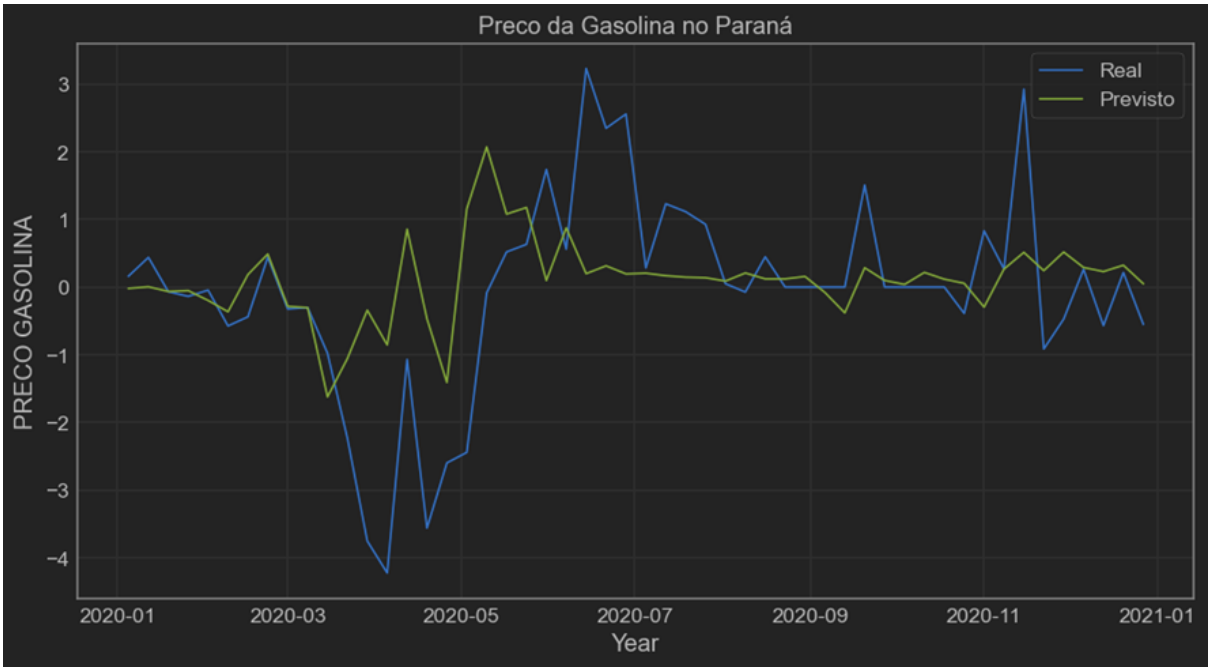
Criando a diferença percentual percebe-se que o coeficiente de relação entre as variáveis é praticamente nulo (0.000 para o dólar e 0.008 para a gasolina), baseado nisso,

foram adicionadas diferenças semanais nas variáveis visto que uma diferença percentual no dólar ou no petróleo pode demorar algum tempo para impactar no preço da gasolina.

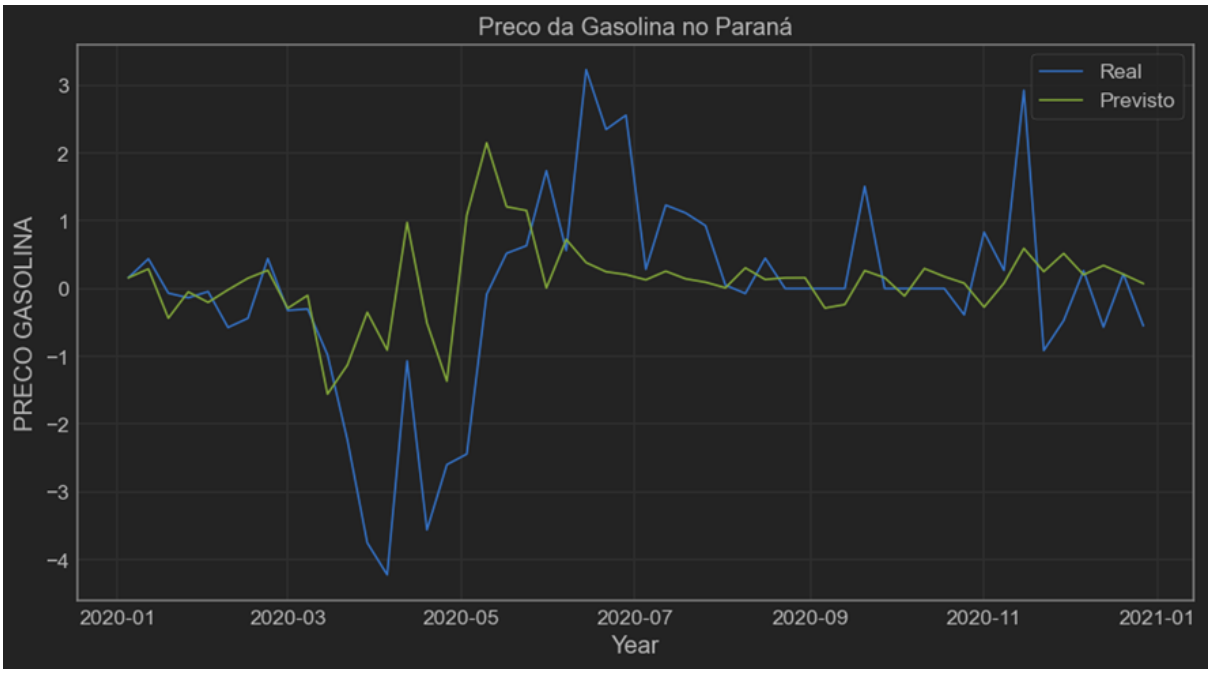
Com uma semana de diferença, o R^2 encontrado foi de 0.027 e os seguintes valores previstos foram obtidos, comparando com os valores reais:



Agora, fazendo a mesma coisa, no entanto adicionando duas semanas de delay entre as variáveis temos um R^2 de 0.024 e o seguinte gráfico:



Realizando a mesma coisa para três semanas de delay, obtemos um R^2 de 0.025 e o seguinte gráfico:



Portanto, percebe-se que nos três gráficos, entre os meses de março e julho de 2020, há uma grande diferença entre o previsto e o valor real da gasolina comum no Paraná. Baseado nisso, é possível concluir que com o lockdown (consequentemente a redução da demanda) os postos se sentiram obrigados a reduzir o preço do combustível, e isso não foi impactado pelo preço do dólar e do petróleo pois o modelo não previu essa redução. A partir de julho, o previsto e o real passam a assumir valores semelhantes.

Limitações/Trabalhos Futuros:

Uma clara limitação dessa parte do trabalho é mensurar o tempo de impacto das variáveis sobre o preço da gasolina comum no Paraná. É difícil concluir que as variáveis impactaram em 1, 2 ou 3 semanas o preço. Além disso, outra limitação é mensurar qual o impacto da inflação sobre as variáveis, como estamos utilizando duas variáveis internacionais (dólar e petróleo) para prever o preço da gasolina comum em um estado brasileiro é difícil relacionar a inflação no Brasil com essas outras duas variáveis impactadas pela inflação internacional.

Como trabalho futuro, talvez seria interessante analisar mais profundamente o tempo de impacto das variáveis sobre o preço da gasolina. Além disso, nos nossos dados iniciais do preço dos combustíveis há, além do preço de revenda, o preço de distribuição (preço que os postos de combustível compram a gasolina), com esses dados poderíamos verificar a hipótese pela variação da diferença entre essas duas variáveis.

c) O lucro bruto dos postos de combustíveis está relacionado com a desigualdade social no estados?

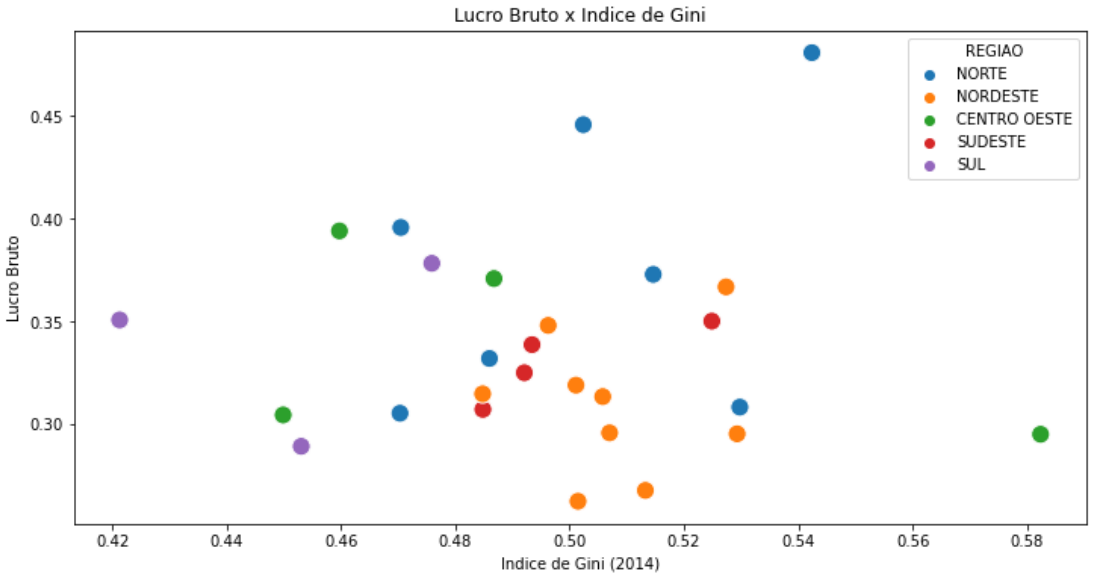
Hipótese: Quanto maior o lucro bruto dos postos de combustível, maior o índice de Gini para aquele estado.

Modelo: Para responder essa pergunta foi criado um scatterplot do lucro bruto (em um dos eixos) e o índice de Gini (no outro eixo), e caso houvesse uma correlação linear (uma linha inclinada) a hipótese estaria satisfeita.

Dados: Na amostragem do relacionamento entre o lucro bruto de 2014 dos postos de combustíveis e a desigualdade social nos estados, foram usados tanto dados do lucro bruto de postos de acordo com cada estado, já usado no último trabalho, quanto o índice de gini, obtido a partir do site do IPEA, para os dados de lucro bruto foi necessário retirar os valores nulos e agrupá-los por estado e ano, de acordo com a média do ano, já para o dataset do índice de gini foram utilizadas somente as colunas com as siglas dos estados e o índice de gini.

Resultados: Com base no Scatter Plot apresentado logo abaixo, observa-se que há uma pequena relação linear entre os valores, entretanto essa relação não é suficiente para ter alguma significância.

É um pouco aparente que, quanto maior o índice de Gini, maior o lucro bruto dos postos, entretanto, observe que em certos pontos essa relação não é tão clara; Além disso, o nordeste aparenta ter uma pequena concentração no gráfico, enquanto as outras regiões aparentam estarem mais dispersas.



Limitações:

- Dados somente de 2014.

d) O que ocorreu com o poder de compra de combustíveis ao longo do tempo?

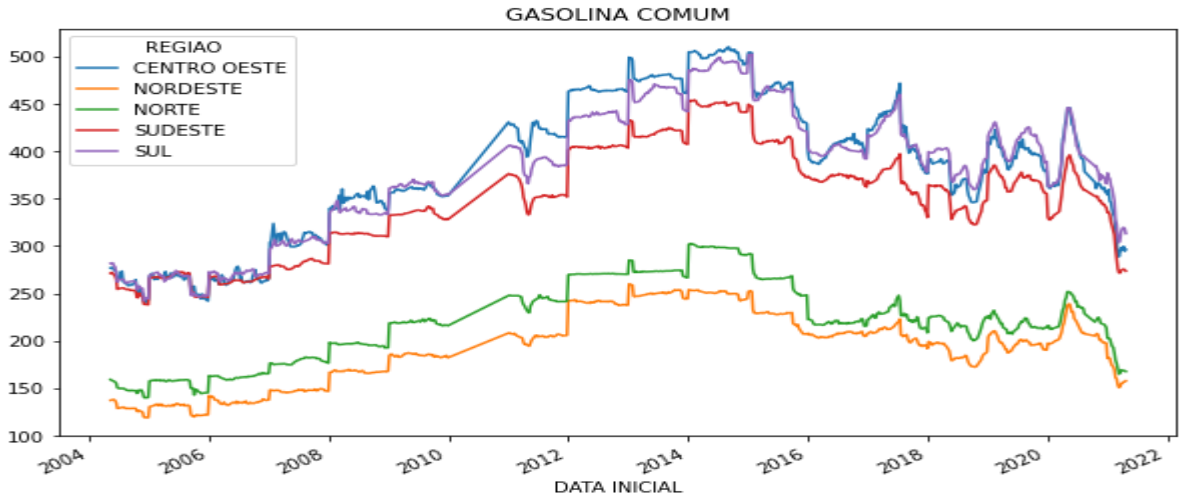
Descrição: Esta pergunta tinha como objetivo analisar se o poder de compra dos brasileiros relacionado a cada combustível aumentou ou diminuiu.

Hipótese: Antigamente o poder de compra dos brasileiros em relação aos combustíveis era bem menor.

Processamento de dados: Foi utilizada a tabela principal com o preço dos combustíveis e nela as colunas de data, preço médio de revenda, região e estado, além disso também foi utilizada a tabela de renda coletada manualmente através de dados do IBGE, em que foram utilizadas as colunas data, renda e estado. Então foi criado um

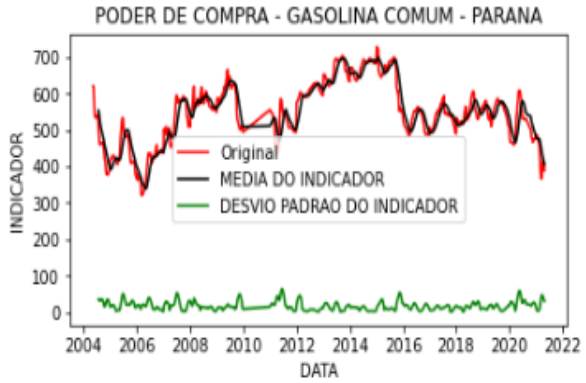
indicador para mensurar quanto cada estado podia comprar quantos litros de cada combustível de cada ano baseado na renda média familiar daquele ano e daquele estado. Por fim, foi feita uma média dos estados para representar a região. O processo de limpeza se encontra nos notebooks “renda_regiao” na pasta 04-Pesquisa e limpeza_tabela_br_oil_prices na pasta 02 - Análise Exploratória.

Resultados: Não rejeitamos a hipótese.



Limitações:

- Não foram coletados dados de renda de 2010.
- Os dados de renda utilizados são de renda mensal familiar por estado, sendo assim esses dados representam a média de renda do estado de uma forma geral e não as diferentes classes sociais. Uma pesquisa mais minuciosa consideraria diferentes faixas de valores - por exemplo, faixa 1: até 1 salário mínimo, faixa 2: de 1 a 3 salários mínimos e assim por diante - e quantos % da população está representada em cada faixa.
- Não é possível tirar conclusões precisas apenas com visualização. Levei em consideração o que foi comentado na apresentação e que o p-valor e teste estatístico poderia ter sido mostrado. Tirei o p-valor dos dados das cinco regiões, mas a granularidade ficou bem alta. Fiz um filtro de um combustível e de um estado e obtive p-value ~ 0.08, então não rejeitamos a hipótese para esse caso.




```
Test statistic:  -2.6564469124351873
p-value:  0.08188279018781491
Critical Values: {'1%': -3.438369485934381, '5%': -2.865079774912655, '10%': -2.5686548826226527}
```