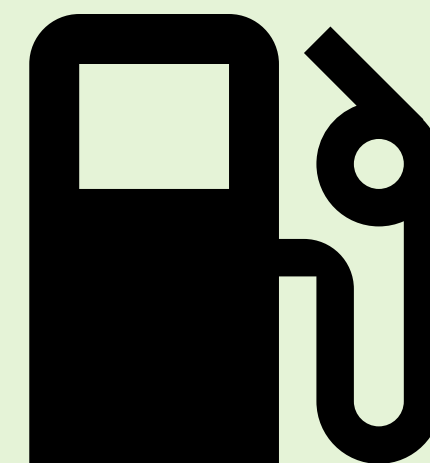




Resilia Educação

DADOS PARA ABASTECER O CARRO

4º Módulo



SQUAD:

PATRICK ELEOTERIO

CO-FACILITADOR



 **patrickeleoterio**

 **patrickeleoterio**

JAQUELINE DAMASCENO

GESTORA DE CONHECIMENTO



 **jaquelinesindie**

 **jaquelinesindie**

ISABELLE CAVALCANTE

GESTORA DE GENTE E ENGAJAMENTO



 **isabelle-cavalcante**

 **isa-sputnik**

THIAGO VASCONCELOS

COLABORADOR II



 **thiago-vasconcelos**

 **Avext**

LAURA ROMANO

COLABORADORA I



 **laura-romano**

 **lauramsromano**

RESILIA

Apresentação do Projeto:

O PROJETO:



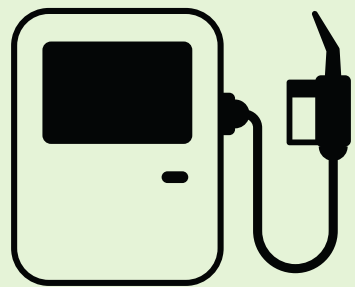
Fomos escadados pela Agência Nacional de Petróleo e Gás Natural e Biocombustíveis (ANP) para realizar uma análise exploratória relacionada à série histórica de preço de venda da gasolina e do etanol.

Foram utilizados os arquivos dos dois últimos meses do ano atual contendo a série histórica dos preços da gasolina e do etanol em todo o Brasil que estão disponíveis no portal dados.gov.

Expectativas e Desafios:

EXPECTATIVAS:

- Responder as questões levantadas no projeto utilizando os conteúdos aprendidos durante o módulo.
- Desenvolver o projeto em grupo de forma harmoniosa, organizada e pacífica.

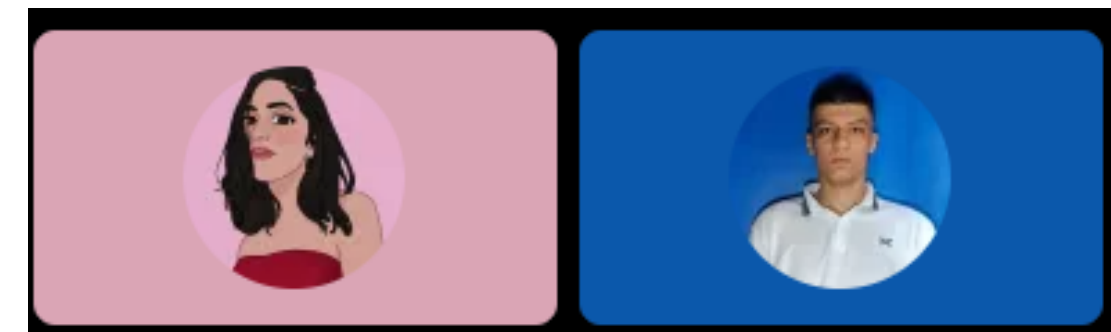


DESAFIOS:

- Elaboração do código (bibliotecas: funções e comandos).
- Compreensão estatística dos dados.
- Desgaste em relação ao tempo disponível para encontros.

- Encontro 1: Divisão de tarefas, criação do arquivo no Colab e tratativas iniciais.
- Encontro 2: Definições iniciais da apresentação (slide) e tratativa das perguntas do projeto.
- Encontro 3: Limpeza e tratamento dos dados.
- Encontro 4: Padronização do código e finalização das perguntas do projeto.
- Encontro 5: Definição das perguntas a serem elaboradas pelo grupo e elaboração inicial do slide.
- Encontro 6: Finalização do slide e apresentação.

REUNIÕES:



ORGANIZAÇÃO: TRELLO

Modulo 4 - Dados para abastecer o carro:

Backlog - Levantamento das atividades a serem feitas.

+ Adicionar um cartão

A fazer: Atividades que devem ser iniciadas.

Ultima semana: Dar uma editada no markdown.

Realizar o comentário nas próprias perguntas realizar a conclusão

Fazendo: Atividades já iniciadas.

Criação do gráfico Burndown:

Sugestão de estrutura para apresentação do projeto nos slides:

Criação do slide e apresentação.

Feito: Atividades feitas.

Criação do notebook compartilhado no Google Colab.

Divisão das tarefas do squad.

Responder as perguntas:

Criar repositório no Github.

Definir o tema dos Slides:

Limpeza da Dataframe

Amanhã: Padronização ou normalização dos dados

Reuniões:

Marcação das reuniões iniciais.

Reunião 1: Verificando informações iniciais e divisão de tarefas:

Reunião 2: Verificando as perguntas.

Reunião 3 - Limpeza dos dados.

Reunião 4: Padronizando o código.

Reunião 5: Perguntas do grupo.

Legenda e informações:

INFORMAÇÕES DO PROJETO:

Alta prioridade:

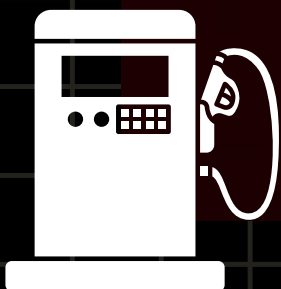
Média prioridade:

Concluído

Como a divisão de tarefas foi feita: Time de 5 pessoas. Atividades gerais (todos fazem individualmente): To do 7, divisão das tarefas, etc. | Atividades gerais (divisão de tarefas conforme demandas e experiência): diferentes partes da análise dos dados.

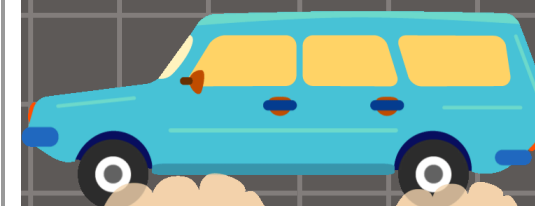
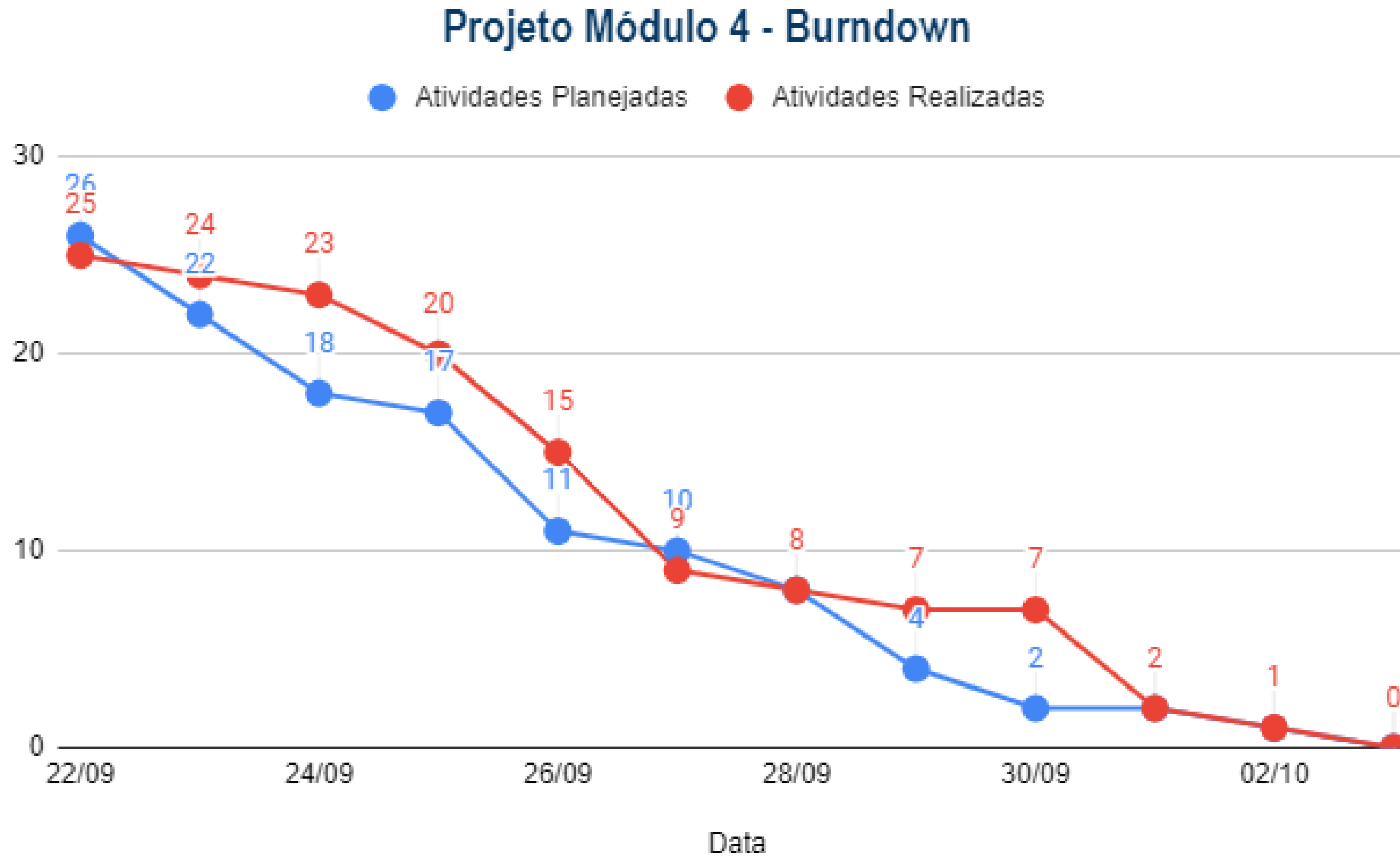
Quantidade máxima de atividades em cada coluna: Conforme demanda do projeto e dos participantes do squad.

criar com base em modelo...



RESILIA

GRÁFICO BURNDOWN:



FERRAMENTAS:

TECH:



SOFT:



O CÓDIGO:

▼ Dados para abastecer o carro!

Fomos escalados pela Agência Nacional de Petróleo e Gás Natural e Biocombustíveis (ANP) para realizar uma análise exploratória relacionada à série histórica de preço de venda da gasolina e do etanol. Serão utilizados os arquivos dos dois últimos meses do ano atual contendo a série histórica dos preços da gasolina e do etanol em todo o Brasil que estão disponíveis no portal dados.gov.

► 1. Coletando os dados:

Nesta etapa coletamos e realizando uma análise inicial para melhor compreensão dos dados de forma geral.

Os dados foram retirados do Portal Brasileiro de Dados Abertos do Governo Federal: dados.gov.br.

O link direto aos dados você encontra abaixo:

- [Dados Julho 2022](#)
- [Dados Agosto 2022](#)

Caso queira visualizar os dados de outros períodos, você pode encontrar no link abaixo:

- [Dados Gerais](#)

[] ↴ 12 células ocultas

► 2. Limpeza do dados:

Nesta etapa buscamos identificar e tratar possíveis informações nulas (NaN), duplicadas e outliers para uma melhor análise dos dados.

▶ ↴ 23 células ocultas



O CÓDIGO:

▶ 3. Analisando os dados e respondendo as perguntas:

Nesta etapa realizando uma análise descritiva dos dados através de algumas perguntas direcionadoras.

[] ↳ 60 células ocultas

▶ 4. Conclusão:

↳ 1 célula oculta

▶ 5. Referências:

↳ 1 célula oculta



TRATANDO OUTLIERS:



2.2 Tratando possíveis outliers:

Nesta etapa, por meio da aplicação de técnicas de estatística descritiva como análise de valores centrais **verificamos se há a presença de outliers no dataframe a ser analisado e como devem ser tratados.**

```
# Verificando algumas informações estatísticas da coluna 'Valor de Venda', única coluna numérica presente nos dados:  
df.describe()
```

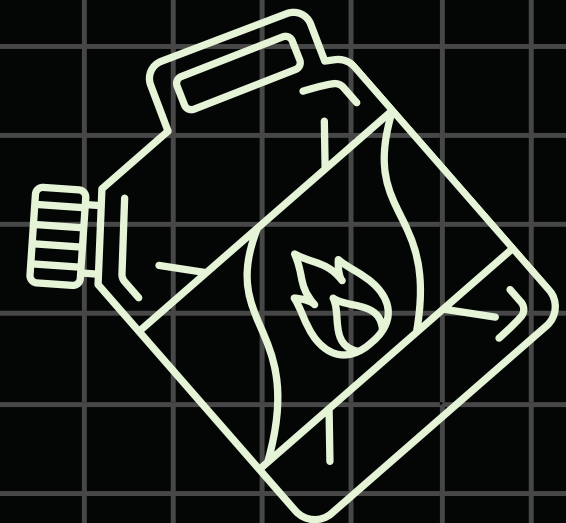
	Valor de Venda
count	127176.000000
mean	5.355503
std	0.860449
min	2.890000
25%	4.880000
50%	5.490000
75%	5.890000
max	9.270000

```
[53] # Verificando o valor de mediana da coluna 'Valor de Venda':  
df['Valor de Venda'].median()
```

5.49

Verificamos nas informações acima uma média (*mean*) de **5.35** um desvio padrão (*std*) de **0.86** e uma mediana (*median*) de **5.49**.

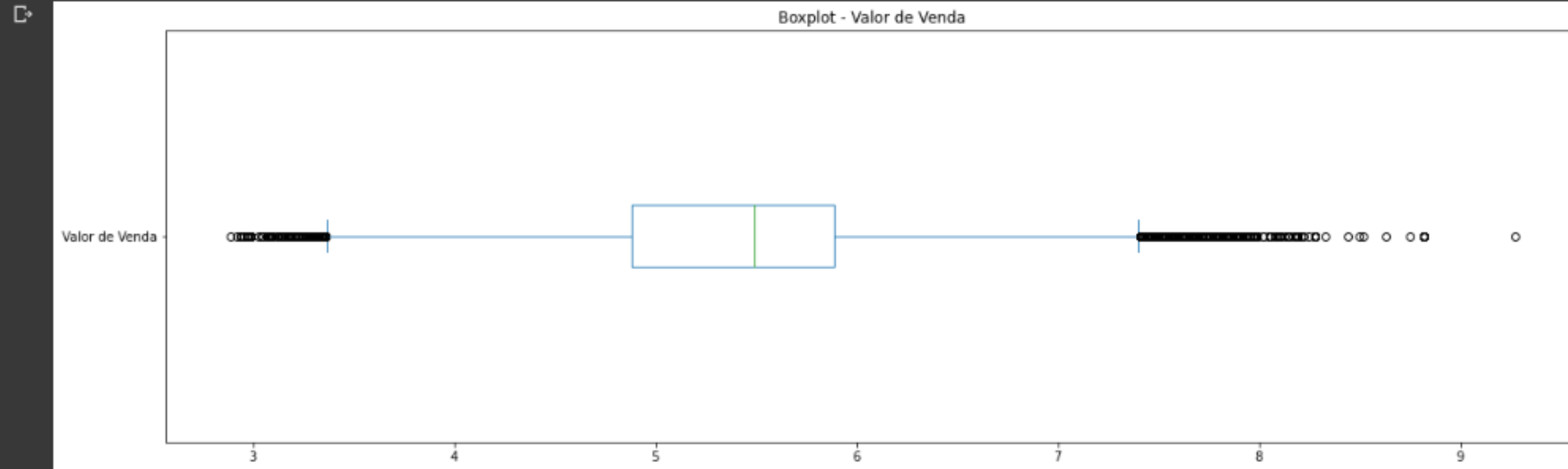
Contudo, podemos verificar que há um valor máximo (*max*) de **9.2** e um valor mínimo (*min*) de **2.89**, indicando possíveis outliers nos dados.



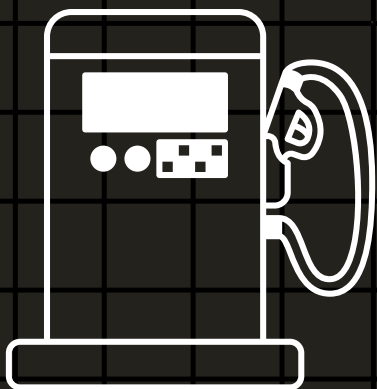
TRATANDO OUTLIERS:

Para verificar mais a fundo essa informação, plotamos o gráfico do tipo Boxplot, conforme abaixo:

```
#Gráfico Boxplot para identificar os outliers:  
df['Valor de Venda'].plot.box(vert=False, figsize = (20,6));  
plt.title('Boxplot - Valor de Venda');
```



Com o gráfico fica visível a presença de outliers, considerando valores abaixo de aproximadamente 3.4 e acima de aproximadamente 7.3



TRATANDO OUTLIERS:

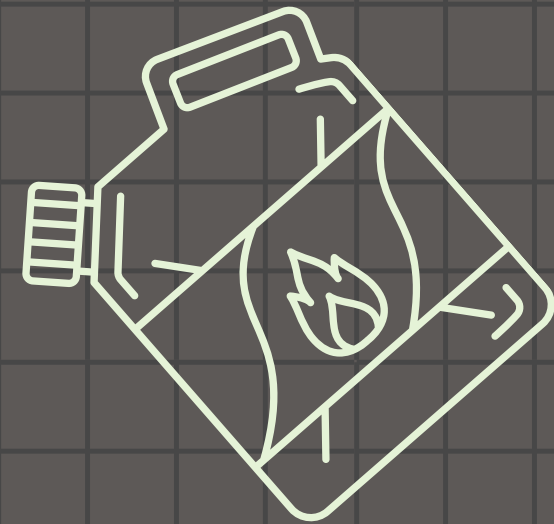
Temos agora que verificar se eles interferem diretamente nas grandezas estatística a ponto de haver a necessidade de retirá-los da amostra de dados a ser analisada.

```
# Verificando algumas informações estatísticas da coluna 'Valor de Venda',  
#Única coluna numérica presente nos dados para valores menores que 3.14 e maiores que 7.3  
df[(df['Valor de Venda'] >= 3.4) & (df['Valor de Venda'] <= 7.3)].describe()
```

	Valor de Venda
count	124756.000000
mean	5.349248
std	0.814622
min	3.400000
25%	4.880000
50%	5.490000
75%	5.890000
max	7.300000

```
[95] df[(df['Valor de Venda'] >= 3.4) & (df['Valor de Venda'] <= 7.3)].median()
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: FutureWarning: DataFrame.mean and DataFrame.median with numer  
    """Entry point for launching an IPython kernel.  
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: FutureWarning: Dropping of nuisance columns in DataFrame redu  
    """Entry point for launching an IPython kernel.  
Valor de Venda    5.49  
dtype: float64
```



TRATANDO OUTLIERS:

Verificando as informações estatística considerando esse recorte percebemos que **não há grandes mudanças em relação aos valores de tendência centrais como média, mediana e desvio padrão.**

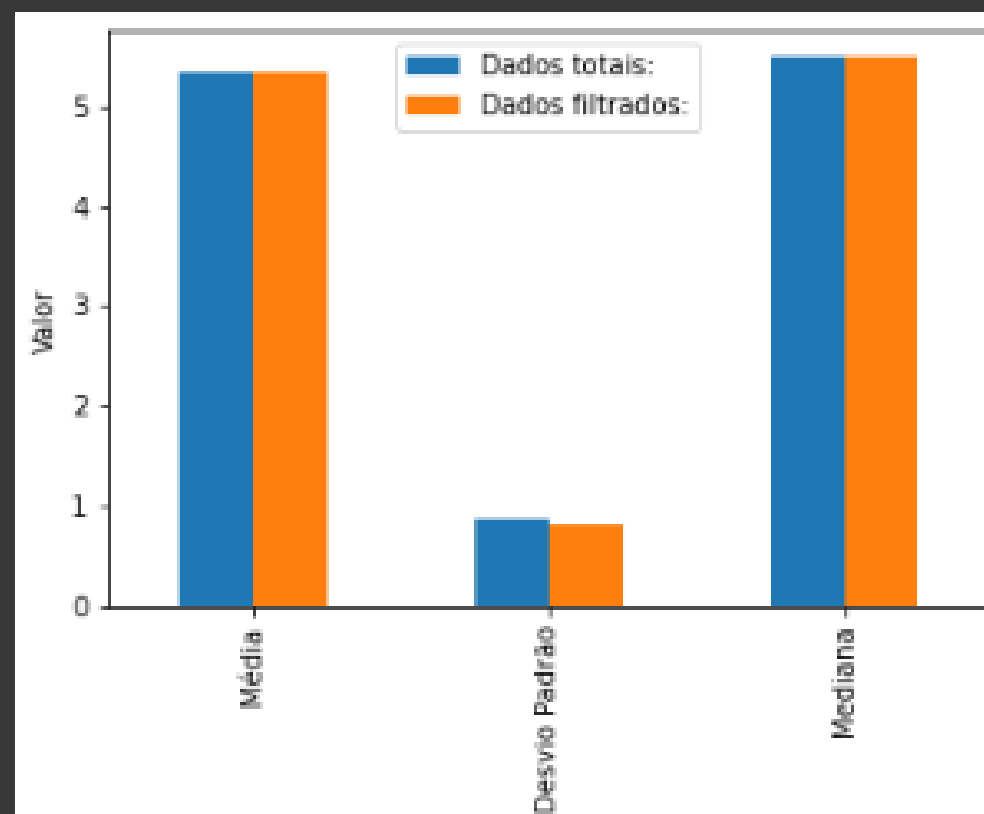
```
✓ [100] #Gráfico das médias, desvio padrão e mediana dos dados completos e filtrados:
```

```
# Criando dicionário para nomeação dos index:
leg_dados = {0: 'Dados totais:', 1: 'Dados filtrados:'}

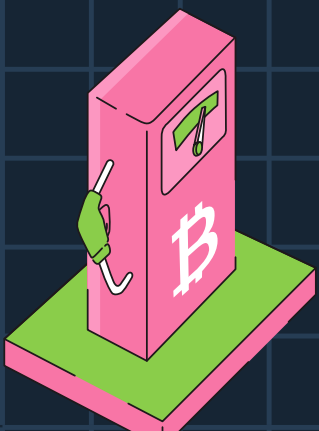
# Criando dataframe com os dados:
dados = pd.DataFrame({'Média': [5.35, 5.34], 'Desvio Padrão': [0.86, 0.81], 'Mediana': [5.49, 5.49]})

# Renomeando os index conforme dicionário criado:
dados.rename(index = leg_dados, inplace=True )

#Transpondo os dados e plotando:
dados.T.plot.bar()
plt.ylabel('Valor');
```



Dessa forma, como não há diferença considerável nos valores de tendência central com os dados total (contendo outliers) e os dados filtrados (sem os outliers), **decidimos seguir com nossas análises com a base de dados completa.**



AS PERGUNTAS:

1. Como se comportaram o preço dos combustíveis durante os dois meses citados? Os valores do etanol e da gasolina tiveram uma de queda ou diminuição?

```
[58] # Criando uma cópia do df para tratamento dos dados em paralelo:
df_copia = df.copy()
df_copia['Data da Coleta'] = pd.to_datetime(df_copia['Data da Coleta'], format="%d/%m/%Y")

# Filtrando as informações conforme coluna 'Produto' (Etanol, Gasolina e Gasolina Aditivada):
etanol = df_copia.query('Produto == "ETANOL"')
gasolina = df_copia.query('Produto == "GASOLINA"')
gasolina_aditiv = df_copia.query('Produto == "GASOLINA ADITIVADA"')

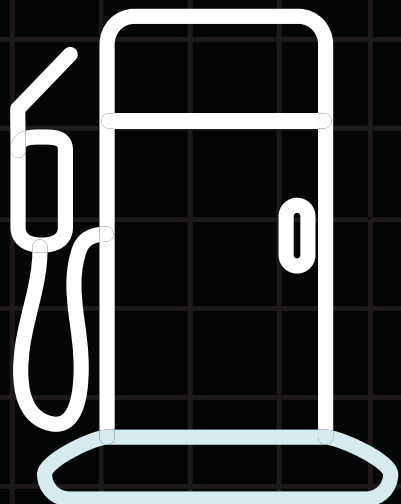
# Agrupando e organização as informações pela coluna 'Data de Coleta' e verificando a média do valor de venda conforme coluna 'Valor de Venda':
evolucao_precos_etanol = etanol.sort_values(['Data da Coleta']).groupby('Data da Coleta')['Valor de Venda'].mean()
evolucao_precos_gasolina = gasolina.sort_values(['Data da Coleta']).groupby('Data da Coleta')['Valor de Venda'].mean()
evolucao_precos_gasolina_aditiv = gasolina_aditiv.sort_values(['Data da Coleta']).groupby('Data da Coleta')['Valor de Venda'].mean()

# Criando gráfico com as informações encontradas:

# Plotando os gráficos:
fig, ax = plt.subplots(figsize=(25, 10))
fig.autofmt_xdate()

#Título do gráfico e dos eixos:
plt.title("Comportamento dos valores da gasolina e etanol nos meses de junho e julho de 2021")
plt.xlabel("Datas")
plt.ylabel("Valores")

# Plotando os gráficos:
plt.plot(evolucao_precos_etanol)
plt.plot(evolucao_precos_gasolina)
plt.plot(evolucao_precos_gasolina_aditiv)
```

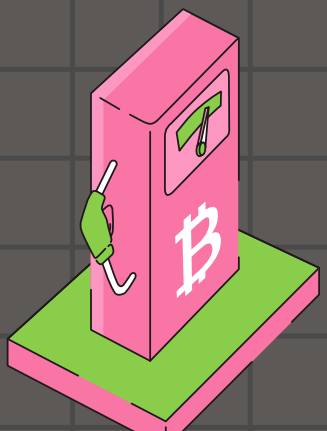


AS PERGUNTAS:

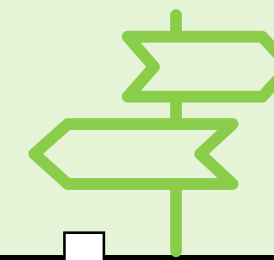
```
# Plotando os gráficos:
plt.plot(evolucao_precos_etanol)
plt.plot(evolucao_precos_gasolina)
plt.plot(evolucao_precos_gasolina_aditiv)

# Editando informações necessárias no gráfico:
plt.axhline(y = 4.52, color = 'r', linestyle = 'dashed')
plt.axhline(y = 5.77, color = 'r', linestyle = 'dashed')
plt.axhline(y = 5.63, color = 'r', linestyle = 'dashed')

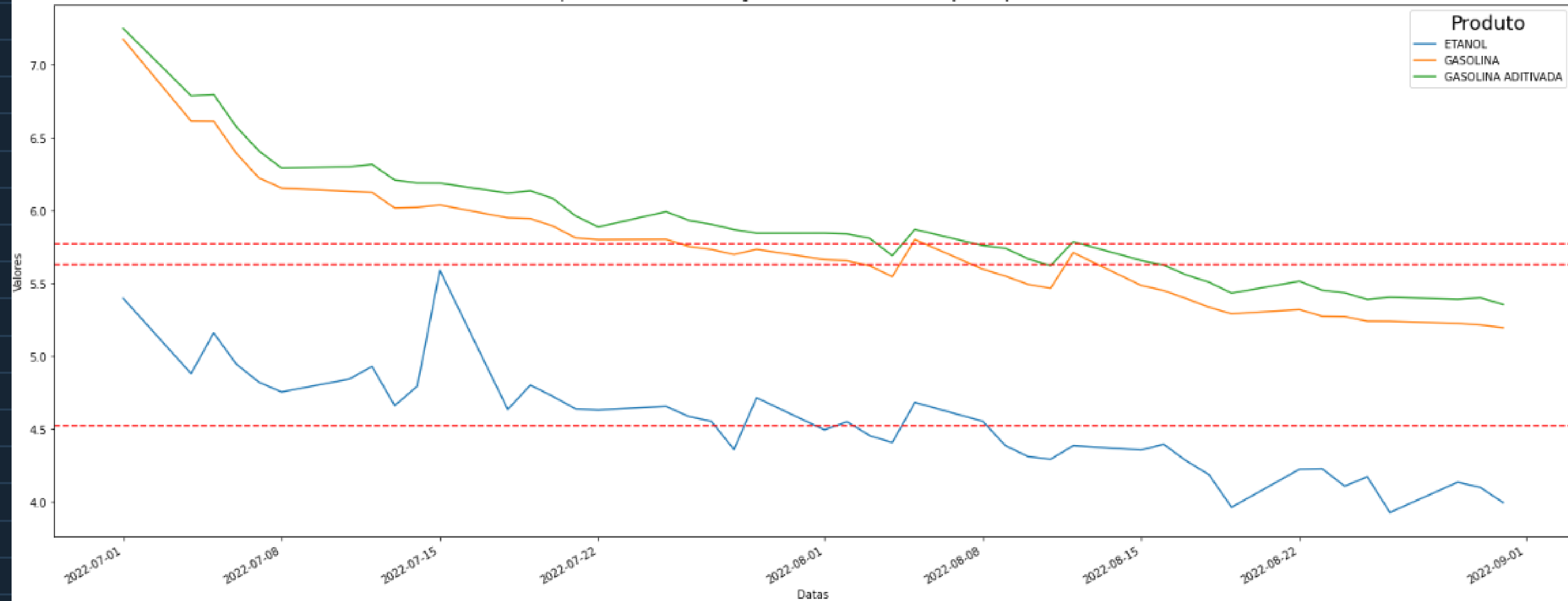
# Legenda do gráfico:
legend = plt.legend(['ETANOL', 'GASOLINA', 'GASOLINA ADITIVADA'], title = 'Produto', title_fontsize = 18)
```



AS PERGUNTAS:



Comportamento dos valores da gasolina e etanol nos meses de junho e julho de 2021



AS PERGUNTAS:

8. Qual a região que possui o menor valor médio do etanol?

```
[73] # Filtrando as informações conforme coluna 'Produto' (Etanol, Gasolina e Gasolina Aditivada):  
# agrupando pela coluna 'Regiao - Sigla' e verificando a média do valor de venda pela coluna 'Valor de Venda':  
  
dfEtanol = df[df['Produto'] == 'ETANOL'].groupby('Regiao - Sigla')[['Valor de Venda']].mean()  
dfEtanol.head()
```

Valor de Venda	
Regiao - Sigla	
CO	4.107305
N	5.340633
NE	5.246069
S	4.842372
SE	4.149964

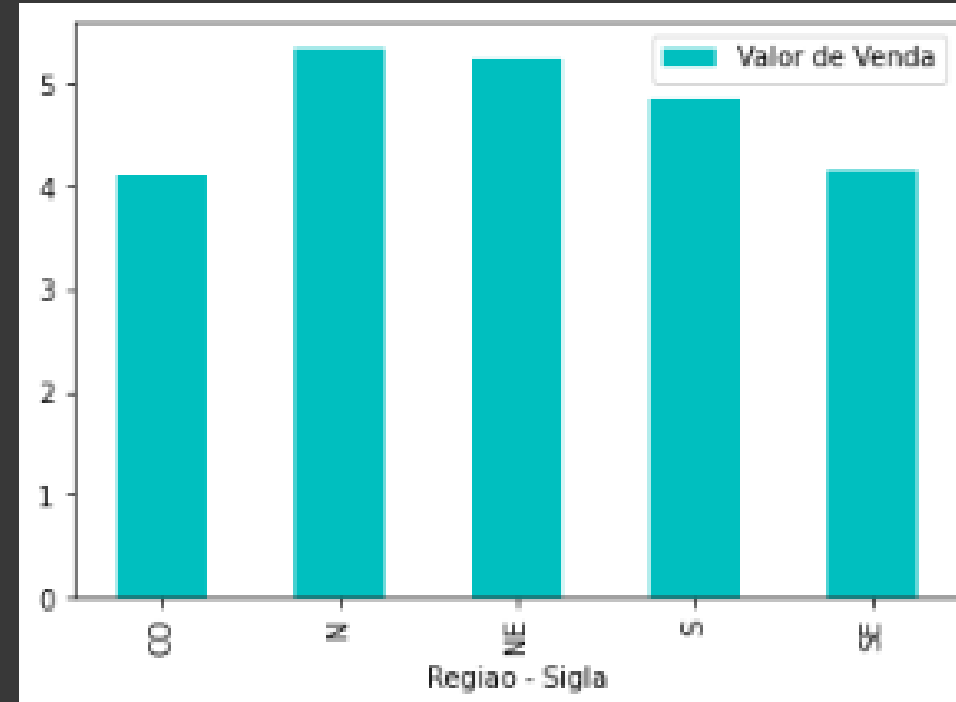


AS PERGUNTAS:

Também podemos visualizar esses dados na forma de gráfico:

```
[ ] dfEtanol.plot.bar( color="c")
```

<matplotlib.axes._subplots.AxesSubplot at 0x7f0391453110>




▶ # Mostrando a sigla da região e o maior valor para produto 'Etanol':

```
print(dfEtanol.idxmin())  
print(dfEtanol.min())
```


```
Valor de Venda    CO  
dtype: object  
Valor de Venda    4.107305  
dtype: float64
```

AS PERGUNTAS:

11. Qual é maior a média de preço de venda por Bandeira?

✓ 0s  #Agrupamos pela coluna "Bandeira" e calculamos a média dos valores de venda pela coluna "Valor de Venda":

```
df.groupby('Bandeira')[['Valor de Venda']].mean().sort_values(by=['Valor de Venda'], ascending=False).round(2).head(5)
```

Valor de Venda 	
Bandeira	
PETRONAC	6.19
EQUADOR	5.94
SP	5.93
REJAILE	5.91
FAN	5.89



AS PERGUNTAS:

12. Qual revenda vendeu mais caro em média, por região:

```
# Filtramos pelas regiões na coluna "'Regiao - Sigla'",  
# agrupamos pelas colunas 'Regiao - Sigla' e 'Revenda' e calculamos  
# e média dos valores de venda pela coluna "Valor de Venda" para cada :  
  
# Região SE - Sudeste  
  
df[df['Regiao - Sigla'].isin(["SE"])]\  
.groupby(['Regiao - Sigla', 'Revenda'])['Valor de Venda']\  
.mean().sort_values(by=['Valor de Venda'], ascending=False).round(2).head(5)
```

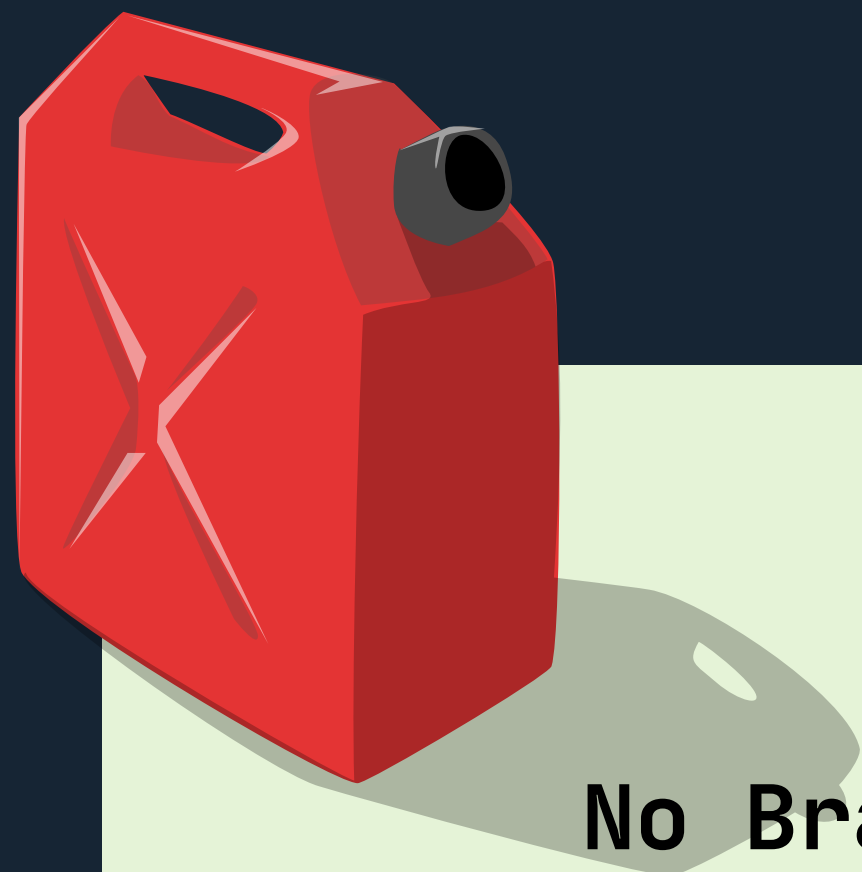
Regiao - Sigla		Valor de Venda
Regiao - Sigla		Revenda
SE	3POSTO RONCADOR LTDA.	7.24
	POSTO DE SERVICO NINO S LTDA	6.76
	POSTO SUPERSONICO LTDA	6.72
	POSTO CARIOQUINHA LTDA	6.69
	CENTRO AUTOMOTIVO ALPHA CENTER LTDA	6.69

Regiao - Sigla		Valor de Venda
Regiao - Sigla		Revenda
S	OP DERIVADOS DE PETROLEO LTDA	7.04
	BARROS, DIAS & CIA LTDA	6.96
	SLD COMERCIO DE COMBUSTIVEIS LTDA	6.90
	COMERCIAL INTERNACIONAL LTDA	6.77
	SANTA MARIA DISTRIBUIDORA DE DERIVADOS DE PETROLEO LTDA	6.68

Regiao - Sigla		Valor de Venda
Regiao - Sigla		Revenda
CO	POSTO PAULISTA PNEUS LTDA	6.27
	MARINHO & CIA LTDA	6.17
	WMR MINEIROS LTDA	6.16
	AUTO POSTO SUPER SOL LTDA	5.95
	ARAUJO & DUENHA LTDA	5.93

Regiao - Sigla		Valor de Venda
Regiao - Sigla		Revenda
N	M C D CARVALHO & CIA LTDA	7.50
	A. M. DE FARIAS - EPP	7.50
	BRITO E BARRA LTDA	7.29
	M T COMERCIO DE COMBUSTIVEIS LTDA	7.26
	M DOS S TELLO SOBRINHO	7.26

Regiao - Sigla		Valor de Venda
Regiao - Sigla		Revenda
NE	AUTOPOSTO CONFIANCA LTDA	7.43
	J.C COMERCIO VAREJISTA DE COMBUSTIVEIS LTDA	7.27
	COMERCIAL DE PETROLEO MOREIRA PEQUENO LTDA	7.14
	ALIANCA COMERCIO DE COMBUSTIVEIS LTDA	7.09
	COMERCIO DE COMBUSTIVEIS PETROSOJA II LTDA	7.09



CONCLUSÃO

No Brasil, o preço pago pelo consumidor final nos combustíveis derivados de petróleo é composto por diversos tipos de impostos, custos de transporte, taxas e etc. Como podemos verificar, a progressão dos valores é notável, afetando diretamente a vida de milhares de Brasileiros.



CONCLUSÃO

- A tratativa de outliers não foi necessária nos dados analisados.
- Ao longo do período analisado o preço médio dos aditivos diminuiu, com o Etanol tendo a maior variação.
- O valor médio do Etanol está diretamente ligado a região de revenda (correlação de 0.5) enquanto a Gasolina e a Gasolina Aditivada não possuem correlação com essa variável.
- Em geral, os aditivos são em média mais caros nas regiões Norte e Nordeste.

Lições Aprendidas e Resultados:

LIÇÕES APRENDIDAS:

- Análise estatística dos dados utilizando as ferramentas apresentadas durante o módulo.
- Gestão ágil do projeto (daily e gráfico Burndown)

RESULTADOS:

- Responder todas as perguntas solicitadas no projeto com as ferramentas apresentadas durante o módulo.
- Segurança e boa relação entre os integrantes do squad em todos os momentos do projeto.



AGRADECIMENTOS:

Facilitadores: Rafael Pílan e Esli Queiroz

Monitor: Guilherme Ribeiro

Ex-Resiliente: Mateus Sartorio

Engenheiro de Dados: Vinicius Damião

Sucesso do Estudante: Ana Guimarães



