Base de dados "ModeloVigente_DadosCP62_2023.xlsx" contém valores médios, considerando o período 2018 a 2020, para as seguintes variáveis:

• Custos Operacionais (PMSO/Despesas com Pessoal, Materiais, Serviços e Outros – em R$ 1.000,00)
• Rede de alta tensão (Km)
• Rede subterrânea (Km)
• Rede de distribuição aérea (Km)
• Mercado ponderado (MWh – Mercado Atendido em termos de potência)
• Consumidores totais (unid)
• Consumidor Hora Interrompido - CHI (h – Tempo médio de consumidores sem energia)
• Perdas Não Técnicas - PNT (MWh – Energia perdida considerando fraudes, gatos, etc.)

Essas variáveis estão disponíveis para 52 empresas Brasileiras distribuidoras de energia elétrica. A Agência Nacional de Energia Elétrica (ANEEL) está utilizando essas informações para calcular a eficiência operacional das empresas de distribuição para o ano de 2023 e para os próximos anos.

Neste desafio, o objetivo é avaliar se os Custos Operacionais (PMSO) podem ser estimados a partir das demais variáveis. O objetivo é estimar o melhor modelo de regressão linear interpretável, considerando relevância estatística das variáveis e capacidade preditiva do modelo.

Roteiro da atividade

1- Carregando bibliotecas
2- Informações básicas do dataset
    a. Print das 5 primeiras linhas
    b. Tipo de dado em cada coluna
    c. Descrição das colunas numéricas (count, média, mediana, desvio padrão)
    d. Quantidade de linhas e colunas
    e. Quantidade de categóricas
3- Análise exploratória
    a. Dados missing
    b. Distribuição da variável PMSO
    c. Scatter plot das variáveis numéricas com PMSO
    d. Correlação de Pearson
    e. Teste de Shapiro-wilk
4- Treinamento do modelo
    a. Modelo com todas as variáveis
    b. Modelo sem as variáveis de multicolinearidade
    c. Modelo sem as variáveis não significativas e sem as variáveis de multicolinearidade
5- Melhor modelo
    a. Análise de resíduos
    b. Análise preditiva do modelo

6- Conclusão

1- Carregando bibliotecas

```
[1]: import warnings
     import matplotlib.pyplot as plt
     import missingno as msno
     import seaborn as sns
     import pandas as pd
     import numpy as np

     import statsmodels.api as sm
     from statsmodels.formula.api import ols
     from statsmodels.stats.outliers_influence import variance_inflation_factor
     from scipy import stats
```

2- Informações básicas do dataset

a. Print das 5 primeiras linhas

```
dfs.head()
```

	DMU	Codigo	PMSO	rede_alta	rede_subterranea	rede_aerea	mercadoP	cons	CHI	PNT
0	RGE SUL (FUSAO 3)	D01f	872875.834	4316.129	73.295	152713.282	8369799.117	2900561.667	14417450.720	394123.495
1	AMAZONAS	D02	810650.018	250.560	36.711	35967.862	2758708.609	1032967.000	6862142.925	1957040.451
2	ENEL RJ	D03	780334.203	3665.902	3283.996	54738.023	5618601.494	2688391.333	16483600.360	753479.604
3	BANDEIRANTE	D04	473808.857	1034.272	220.411	28192.829	5300379.297	1934474.000	3941057.287	269641.164
4	BOA VISTA ENERGIA	D05	162246.335	712.923	0.050	16307.496	642846.856	172172.000	1024922.675	135664.913

b. Tipo de dado em cada coluna

```
dfs.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 52 entries, 0 to 51
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   DMU               52 non-null     object
 1   Codigo            52 non-null     object
 2   PMSO              52 non-null     float64
 3   rede_alta         52 non-null     float64
 4   rede_subterranea  52 non-null     float64
 5   rede_aerea        52 non-null     float64
 6   mercadoP          52 non-null     float64
 7   cons              52 non-null     float64
 8   CHI               52 non-null     float64
 9   PNT               52 non-null     float64
dtypes: float64(8), object(2)
memory usage: 4.2+ KB
```

c. Descrição das colunas numéricas (count, média, mediana, desvio padrão)

```
dfs.describe()
```

	PMSO	rede_alta	rede_subterranea	rede_aerea	mercadoP	cons	CHI	PNT
count	52.000	52.000	52.000	52.000	52.000	52.000	52.000	52.000
mean	484353.942	2416.639	403.115	70862.938	3919206.856	1625136.628	6754251.617	244591.338
std	536654.056	3262.809	843.470	95913.700	4800544.842	1961824.386	8664597.534	412203.726
min	2320.103	0.000	0.000	71.673	10913.805	3773.667	0.000	0.000
25%	33776.577	58.466	0.000	3134.130	208507.833	101832.500	17527.267	3151.038
50%	338802.409	1249.778	22.391	33396.888	2316725.345	1037062.500	3367755.322	53871.998
75%	726569.718	3731.913	321.032	110571.630	5379934.846	2690253.250	11799042.047	291209.606
max	2448650.839	17436.323	3283.996	520266.354	21780368.080	8560418.667	40601260.630	1957040.451

d. Quantidade de linhas e colunas

```
print(f'Quantidade de Linhas:{ dfs.shape[0]}')
print(f'Quantidade de Colunas:{ dfs.shape[1]}')
```

```
Quantidade de Linhas:52
Quantidade de Colunas:10
```

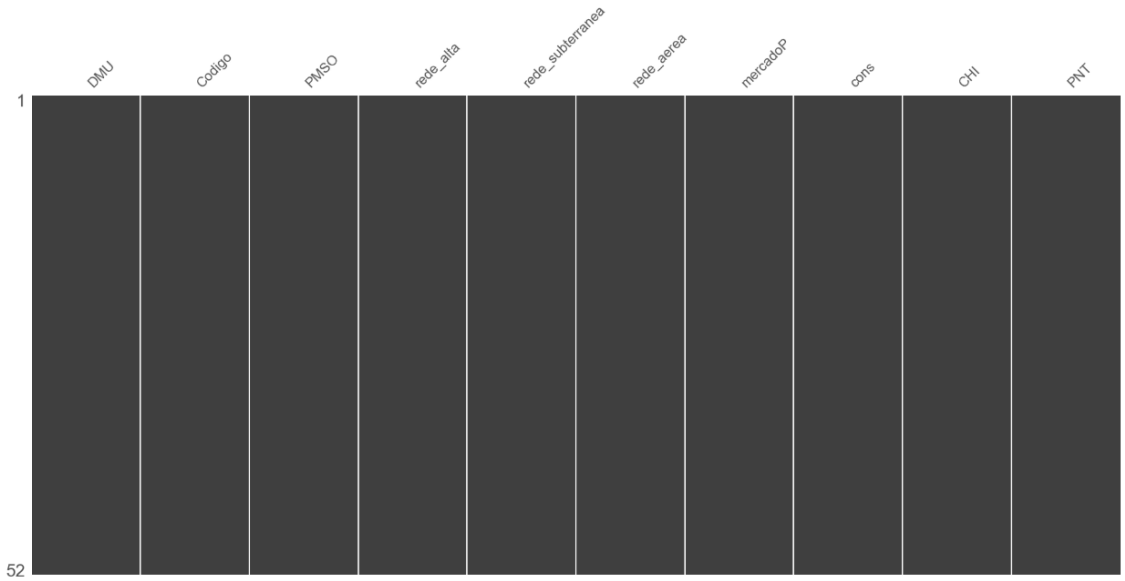
e. Quantidade de categóricas

Não possui categórica duplicada.

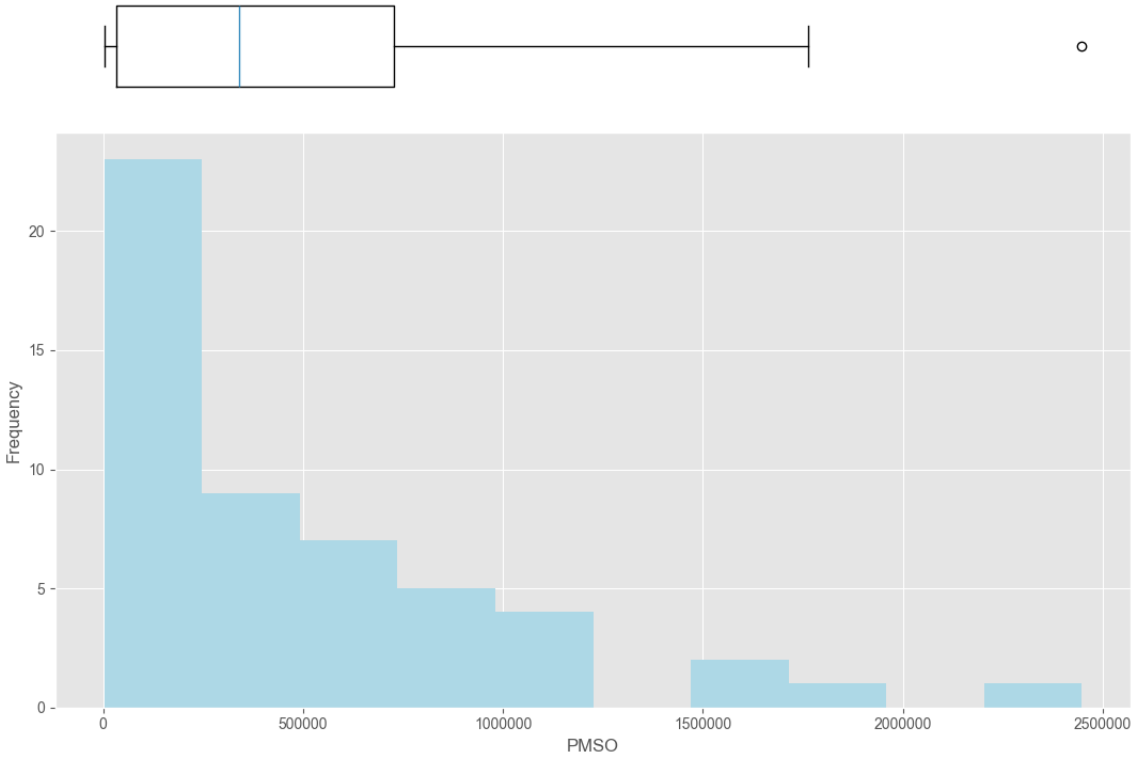
3- Análise exploratória

a. Dados missing

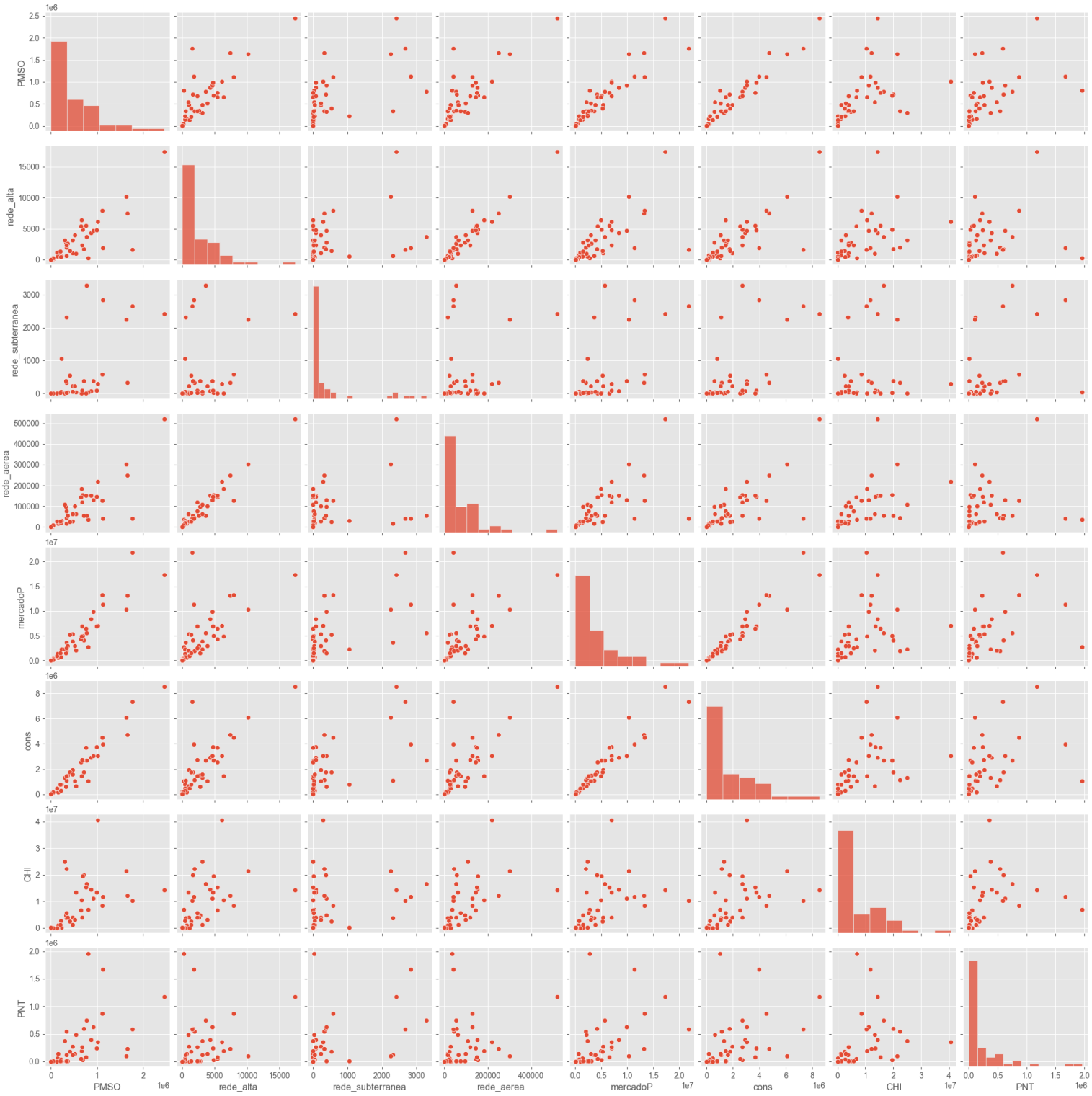
Não temos nenhum dado missing no dataset. No gráfico abaixo podemos visualizar



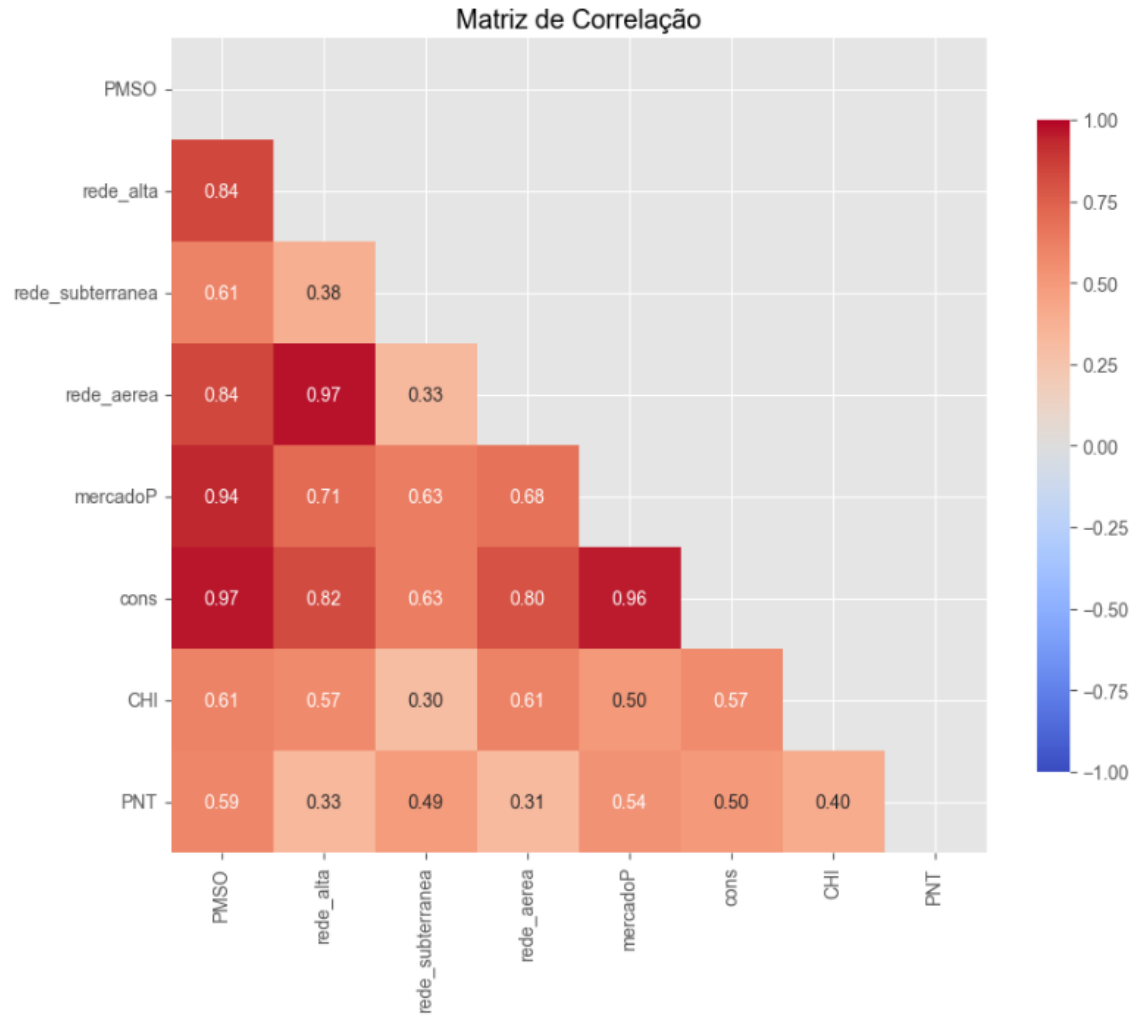
b. Distribuição da variável PMSO


Frequency
20
15
10
5
0
0
500000
1000000
1500000
2000000
2500000
PMSO

c. Scatter plot das variáveis numéricas com PMSO



d. Correlação de Pearson



Sinais de multicolinearidade da variável PMSO com cons (0,97) e mercadoP(0,94). E as variáveis rede aerea e rede alta possuem uma alta correlação o que merece um sinal de alerta

e. Teste de Shapiro-wilk

Shapiro-Wilk normality test: W=0.8286572694778442,

p-value=2.9977438771311427e-06

O teste de Shapiro-wilk permite afirmar que o a distribuição da PMSO não é normal

4- Treinamento do modelo

a. Modelo com todas as variáveis

```
VIF do Modelo 1:
            feature    VIF
0             const  1.898
1        rede_alta 21.951
2  rede_subterranea  2.146
3       rede_aerea 22.073
4         mercadoP 19.918
5             cons 33.588
6              CHI  1.799
7              PNT  1.613
                            OLS Regression Results
==============================================================================
Dep. Variable:                   PMSO   R-squared:                       0.980
Model:                            OLS   Adj. R-squared:                  0.977
Method:                 Least Squares   F-statistic:                     304.9
Date:                Sun, 16 Jun 2024   Prob (F-statistic):          4.11e-35
Time:                        17:43:22   Log-Likelihood:                -657.87
No. Observations:                  52   AIC:                             1332.
Df Residuals:                      44   BIC:                             1347.
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       2.171e+04   1.57e+04      1.384      0.173   -9903.886    5.33e+04
rede_alta     -39.3693     16.511     -2.384      0.021     -72.645      -6.093
rede_subterranea    21.4655     19.968      1.075      0.288     -18.778      61.709
rede_aerea      2.8962      0.563      5.142      0.000       1.761       4.031
mercadoP        0.0366      0.011      3.426      0.001       0.015       0.058
cons            0.0930      0.034      2.739      0.009       0.025       0.161
CHI             0.0003      0.002      0.196      0.846      -0.003       0.004
PNT             0.1914      0.035      5.403      0.000       0.120       0.263
==============================================================================
Omnibus:                        5.168   Durbin-Watson:                   2.270
Prob(Omnibus):                  0.075   Jarque-Bera (JB):                4.490
Skew:                           0.450   Prob(JB):                        0.106
Kurtosis:                       4.123   Cond. No.                     1.66e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.66e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Interpretação:

rede_alta: p=0.021 significativo.

rede_subterranea: p=0.288 não significativo

rede_aerea: p<0.001 altamente significativo

mercadoP: p=0.001 significativo.

cons: p=0.009 significativo .

CHI: p=0.846 não significativo.

PNT: p<0.001 altamente significativo.

Exceto a váriavel rede_alta, as demais impactam aumentando,ou seja, ou aumento de rede_aerea aumenta o PMSO

R-squared (R^2): 0.980 indica que 98%

Adj. R-squared (R^2 ajustado): 0.977

Kurtosis: 4.123 indica que tem uma calda mais pretuberante.Isso já era esperado.

Durbin-Watson: 2.270, indica que não há autocorrelação dos resíduos.

Outro destaque merece ser feito é na presença de multicolinearidade no modelo com todas as variáveis. Quando olhamos para o resultado VIF, vemos as varaiveis:

rede_alta = 21.951

rede_aerea = 22.073

mercadoP = 19.918

cons = 33.588

os valores que indicam multicolinearidade. Algo esperado tendo em vista o resultado da correlação de pearson

b. Modelo sem as variáveis de multicolinearidade

```
VIF do Modelo 2:
            feature   VIF
0             const 1.820
1       rede_aerea 1.643
2  rede_subterranea 1.375
3               CHI 1.714
4               PNT 1.448

Resumo do Modelo 2:
                            OLS Regression Results
==============================================================================
Dep. Variable:                   PMSO   R-squared:                       0.874
Model:                            OLS   Adj. R-squared:                  0.863
Method:                 Least Squares   F-statistic:                     81.65
Date:                Sun, 16 Jun 2024   Prob (F-statistic):           1.50e-20
Time:                        17:44:14   Log-Likelihood:                -705.42
No. Observations:                  52   AIC:                             1421.
Df Residuals:                      47   BIC:                             1431.
Df Model:                           4
Covariance Type:            nonrobust
================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const          6.518e+04   3.71e+04      1.757      0.085   -9439.164     1.4e+05
rede_aerea        3.6722      0.371      9.897      0.000       2.926       4.419
rede_subterranea 165.7922     38.603      4.295      0.000      88.132     243.452
CHI               0.0020      0.004      0.478      0.635      -0.006       0.010
PNT               0.3212      0.081      3.963      0.000       0.158       0.484
==============================================================================
Omnibus:                       45.508   Durbin-Watson:                   1.963
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              212.260
Skew:                           2.239   Prob(JB):                     8.10e-47
Kurtosis:                      11.826   Cond. No.                     1.47e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.47e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Interpretação

R-squared (R^2): 0.874

87.4% da variabilidade do PMSO é explicada pelo modelo

rede_aerea $p<0.001$ altamente significativo

rede_subterranea: $p=0.001$ altamente significativo

CHI: $p=0.635$ não é significativo.

PNT: $p<0.001$ altamente significativo .

Omnibus: 45.508 com $p<0.001$. Indica que os resíduos não seguem uma distribuição normal, sugerido por uma alta assimetria e curtose.

Jarque-Bera (JB): 212.260 com $p<0.001$. Confirma a não normalidade dos resíduos.

Durbin-Watson: 1.963 está próximo de 2, isso indica pouca ou nenhuma autocorrelação dos resíduos.

rede_subterranea (1.375), CHI (1.714), PNT (1.448). Todos os valores de VIF estão abaixo de 10, indicando baixa ou nenhuma multicolinearidade.

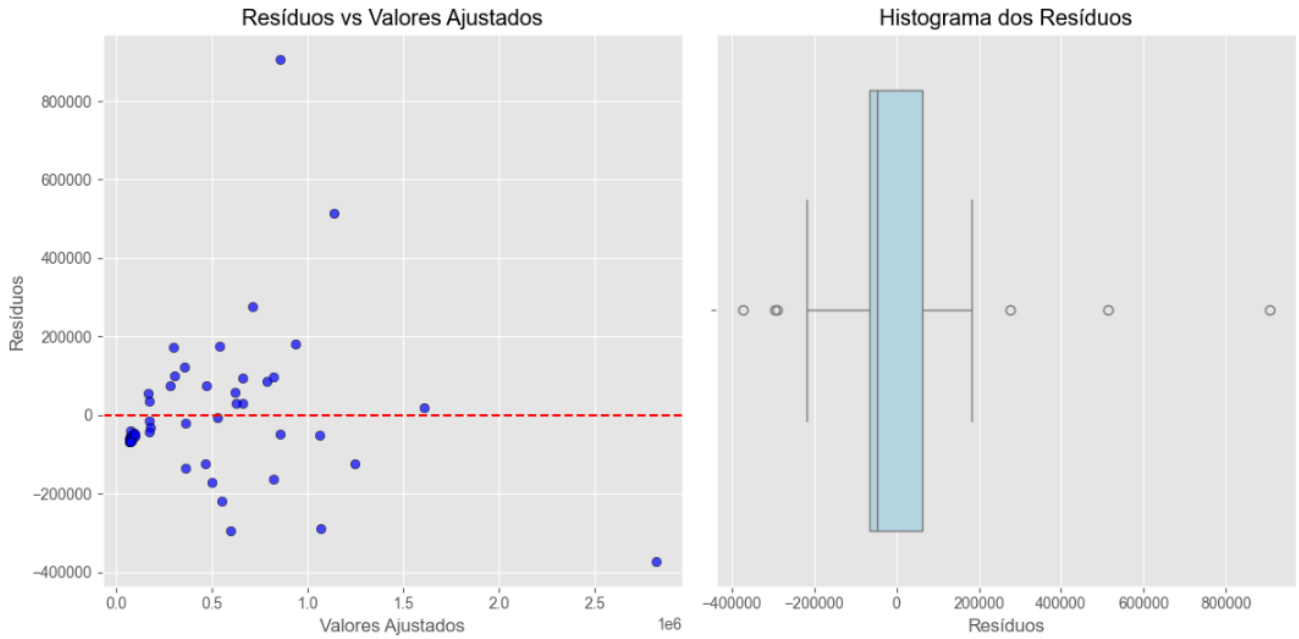
c. Modelo sem as variáveis não significativas e sem as variáveis de multicolinearidade

```
VIF do Modelo 3:
           feature   VIF
0            const 1.406
1  rede_subterranea 1.314
2              PNT 1.314

Resumo do Modelo 3:
                            OLS Regression Results
==============================================================================
Dep. Variable:                   PMSO   R-squared:                       0.483
Model:                            OLS   Adj. R-squared:                  0.462
Method:                 Least Squares   F-statistic:                     22.86
Date:                Sun, 16 Jun 2024   Prob (F-statistic):           9.71e-08
Time:                        17:45:58   Log-Likelihood:                -742.19
No. Observations:                  52   AIC:                             1490.
Df Residuals:                      49   BIC:                             1496.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.537e+05   6.48e+04      3.918      0.000    1.24e+05    3.84e+05
rede_subterranea   264.6187     74.928      3.532      0.001     114.046     415.191
PNT                0.5067      0.153      3.305      0.002       0.199       0.815
==============================================================================
Omnibus:                        9.018   Durbin-Watson:                   1.567
Prob(Omnibus):                  0.011   Jarque-Bera (JB):                8.300
Skew:                           0.861   Prob(JB):                       0.0158
Kurtosis:                       3.931   Cond. No.                     5.64e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.64e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

O R-squared do modelo não caiu indicando que a remoção da variável CHI não impactou. Portanto vamos selecionar o modelo 3 como o mais adequado. As demais estatística se mantiveram alinhadas com o resultado do modelo 2

5- Melhor modelo

a. Análise de resíduos



Shapiro-Wilk normality test: W=0.7887091040611267,

p-value=3.2705284525036404e-07

A variação dos resíduos aumenta com os valores ajustados, indicando heterocedasticidade. Isso sugere que a variância dos erros não é constante ao longo dos valores preditos indo contra a suposição de homocedasticidade do modelo

O boxplot dos resíduos deve idealmente se aproximar de uma distribuição normal (sino). No entanto, a distribuição é assimétrica e não segue uma distribuição normal. Confirmamos isso com o teste de Shapiro-Wilk e Jarque-Bera(JB).

b. Análise preditiva do modelo

```
# Selecionar as variáveis independentes e a variável dependente
X = dfs[['rede_aerea','rede_subterranea', 'PNT']]
y = dfs['PMSO']

# Adicionar uma constante ao modelo (intercepto)
X = sm.add_constant(X)
# Previsões com LOOCV
yhat = np.empty(len(dfs))

for cont in range(len(dfs)):
    X_train = X.drop(cont)
    y_train = y.drop(cont)
    modelo = sm.OLS(y_train, X_train).fit()
    X_test = X.iloc[cont].values.reshape(1, -1)
    yhat[cont] = modelo.predict(X_test)[0]  # Extrair o valor escalar

# Calcular SQT e SQe
SQT = np.sum((y - np.mean(y)) ** 2)
SQe = np.sum((y - yhat) ** 2)

# Calcular o R^2 preditivo
R2_pred = 1 - SQe / SQT
print(f"R^2 preditivo: {R2_pred}")
```

```
R^2 preditivo: 0.8028547064815165
```

R^2 Preditivo:0.8028

R^2 preditivo de 80,28% obtido via LOOCV que é ligeiramente menor que o R^2 do modelo completo, mas ainda indica um bom poder preditivo. É um modelo com uma boa capacidade de generalização

```
[41]: # Ajustar o modelo de regressão linear com todos os dados
      modelo_final = sm.OLS(y, X).fit()

      # Fazer a previsão para a primeira observação com intervalo de confiança de 95% e previsão
      newdata = X.iloc[0].values.reshape(1, -1)
      prediction = modelo_final.get_prediction(newdata)
      prediction_summary = prediction.summary_frame(alpha=0.05)  # 95% de intervalo

      print(prediction_summary)
```

```
        mean   mean_se  mean_ci_lower  mean_ci_upper  obs_ci_lower  \
0 787471.168 42806.009     701403.918     873538.417    382764.213

   obs_ci_upper
0   1192178.122
```

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
Previsão	787471.168	42806.009	701403.918	873538.417	382764.213	1192178.122

6- Conclusão

No modelo escolhido como melhor não foi identificado multicolinearidade significativa entre as variáveis independentes o que é positivo para o modelo e podemos confirmar isso analisando o resultado do VIF. R^2 ajustado de 0.866, indicando um bom ajuste aos dados. O teste de Shapiro-Wilk e o JB indica que que a distribuição dos resíduos não é normal, portanto os resíduos são heterocedásticos. E isso também confirmado quando olhamos para o scatter plot com o resíduos e para o boxplot.

O resultado do R^2 preditivo de 0.8029 obtido via LOOCV confirma a capacidade preditiva do modelo.

Por fim quando olhamos para a previsão para a primeira observação mostra intervalos de confiança e de previsão amplos, o que sugere variabilidade nos dados.

Vale destacar que abaixa amostra impacta na qualidade do modelo, uma amostra maior captaria de forma mais eficiente o comportamento do dados.

E por fim vale mencionar que o processo de regularização (Lasso ou Ridge) ou então inserir uma etapa de normalização dos dados antes do treinamento talvez melhore a capacidade do modelo representar a realidade.