ANÁLISE - REGRESSÃO LINEAR MULTIPLA

Nosso objetivo será responder a seguinte questão: É possível prever a quantidade de pessoas transportadas no transporte público (Ônibus) de São Paulo, levando em consideração o clima do período ?

Ao validar modelos de regressão linear existem 4 premissas básicas que devem ser exploradas para verificar o bom funcionamento do modelo:

1-Linearidade

2-Indepêndencia dos resíduos

3-Homocedasticidade

4-Normalidade dos resíduos

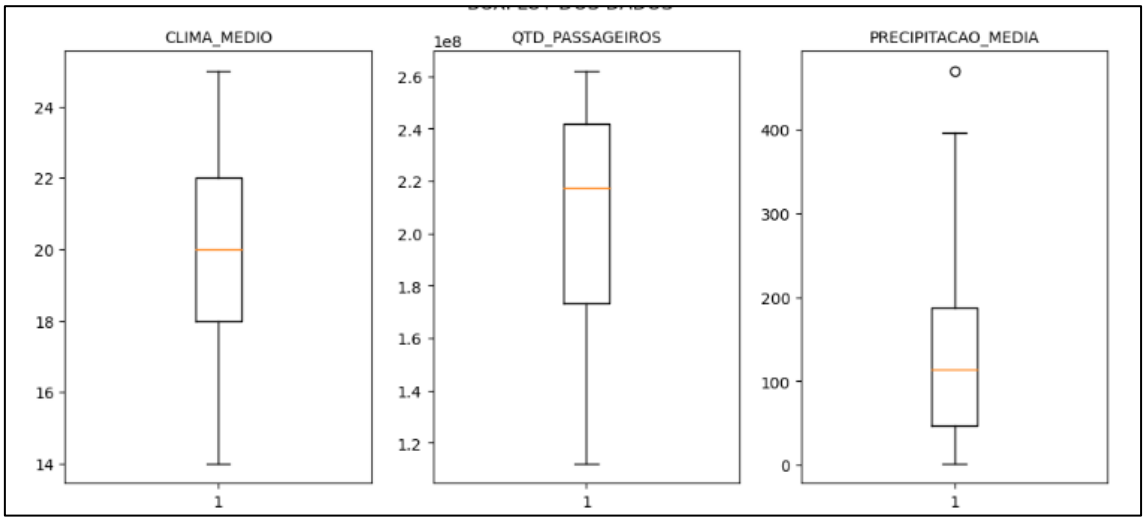
Sendo assim, vamos tentar prever a demanda de passageiros no transporte público de São Paulo utilizando variáveis como: Clima médio mensal, Precipitação de chuva média mensal e Período de férias.

Ao todo são 96 linhas no dataframe, no período de Jan2015 até Dez2023. Os dados sobre o clima foram coletados no site oficial da prefeitura de São Paulo (https://capital.sp.gov.br/) e a coluna do período de férias foi atribuído binário 1 para os meses: Dezembro, Janeiro e Julho de cada ano.

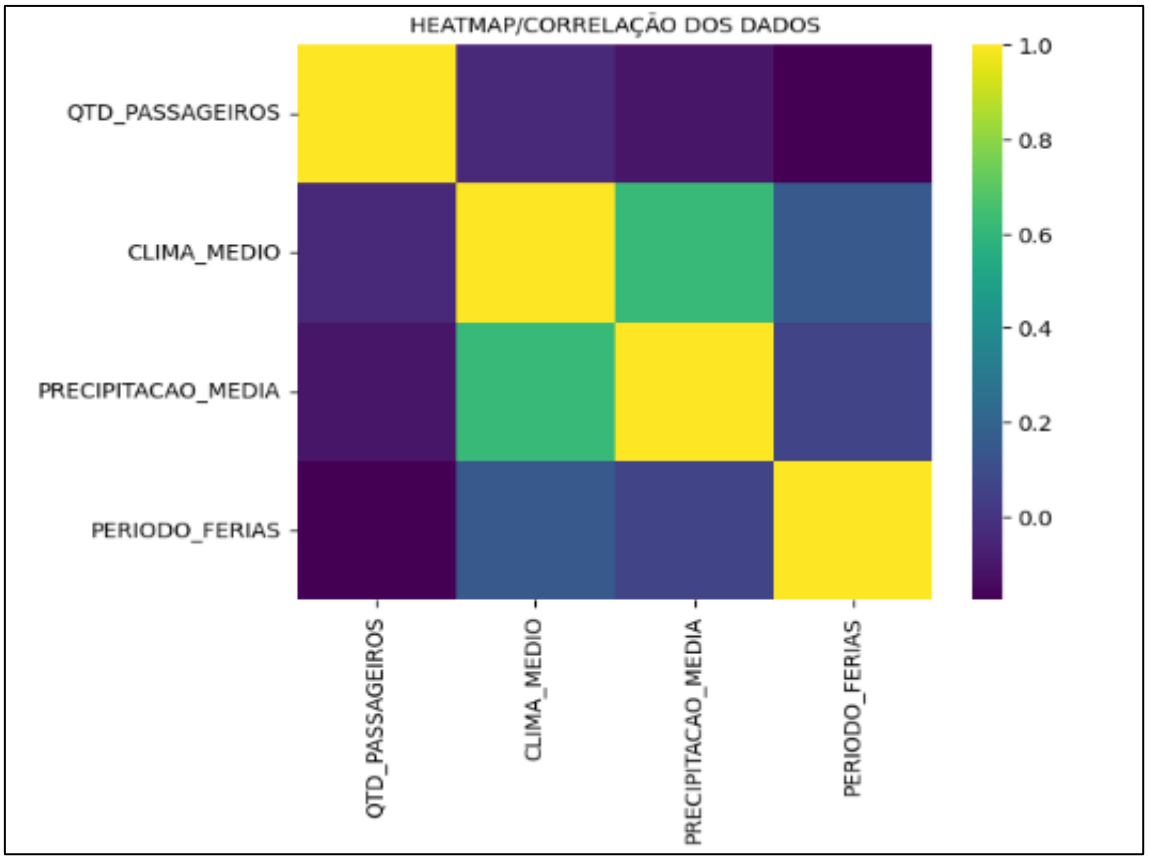
Abaixo podemos verificar as principais estatísticas básicas do conjunto de dados:

```
#######  -INFOS-  ##############
<class 'pandas.core.frame.DataFrame'>
Int64Index: 96 entries, 0 to 95
Data columns (total 5 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   MÊS                 96 non-null     object
 1   QTD_PASSAGEIROS     96 non-null     int32
 2   CLIMA_MEDIO         96 non-null     int32
 3   PRECIPITACAO_MEDIA  96 non-null     int32
 4   PERIODO_FERIAS      96 non-null     float64
dtypes: float64(1), int32(3), object(1)
memory usage: 3.4+ KB
None
#######  -ESTATISTICAS-  ##############
       QTD_PASSAGEIROS  CLIMA_MEDIO  PRECIPITACAO_MEDIA  PERIODO_FERIAS
count            96.00        96.00               96.00           96.00
mean      207484841.36        19.81              134.03            0.25
std        40385900.50         2.53              105.65            0.44
min       112113967.00        14.00                2.00            0.00
25%       173100664.75        18.00               46.75            0.00
50%       217407780.50        20.00              113.50            0.00
75%       242020789.75        22.00              188.00            0.25
max       262069342.00        25.00              470.00            1.00
#######  -BASE-  ##############
       MÊS  QTD_PASSAGEIROS  CLIMA_MEDIO  PRECIPITACAO_MEDIA  PERIODO_FERIAS
0   dez/23         169344196           23                 110            1.00
1   nov/23         169267045           22                 205            0.00
2   out/23         179184034           20                 242            0.00
3   set/23         176996200           21                  83            0.00
4   ago/23         191273935           18                  38            0.00
..     ...               ...          ...                 ...             ...
91  mai/15         247975278           18                  58            0.00
92  abr/15         240925285           20                  50            0.00
93  mar/15         258355557           21                 204            0.00
94  fev/15         214836388           23                 283            0.00
95  jan/15         217543645           24                 262            1.00

[96 rows x 5 columns]
```

Vamos verificar como está a distribuição dos dados de cada variável preditora, por meio do BOXPLOT:



Vamos verificar a correlação dos dados por meio do mapa de calor:

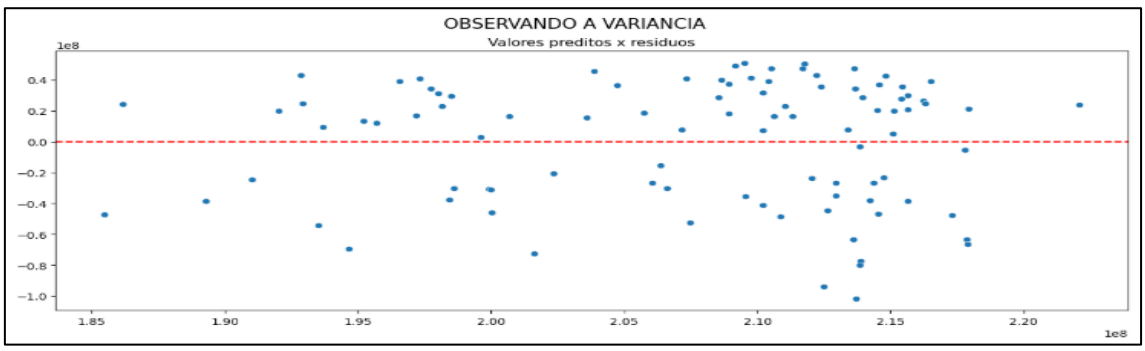


Após o modelo treinado, utilizando a biblioteca STATSMODEL, pode-se observar os seus resultados:

```
#######  -CRIANDO MODELO COM STATSMODEL-  ##############
#######  -RESULTADOS-  ##############
                            OLS Regression Results
==============================================================================
Dep. Variable:        QTD_PASSAGEIROS   R-squared:                       0.043
Model:                            OLS   Adj. R-squared:                  0.012
Method:                 Least Squares   F-statistic:                     1.389
Date:                Sat, 11 Jan 2025   Prob (F-statistic):              0.251
Time:                        21:02:23   Log-Likelihood:                -1814.9
No. Observations:                  96   AIC:                             3638.
Df Residuals:                      92   BIC:                             3648.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept              1.97e+08   3.75e+07      5.247      0.000    1.22e+08    2.72e+08
CLIMA_MEDIO          1.098e+06   2.08e+06      0.527      0.600   -3.04e+06    5.24e+06
PRECIPITACAO_MEDIA  -5.336e+04   4.95e+04     -1.078      0.284   -1.52e+05    4.49e+04
PERIODO_FERIAS      -1.646e+07   9.57e+06     -1.719      0.089   -3.55e+07    2.55e+06
==============================================================================
Omnibus:                       10.654   Durbin-Watson:                   0.222
Prob(Omnibus):                  0.005   Jarque-Bera (JB):                8.764
Skew:                          -0.641   Prob(JB):                       0.0125
Kurtosis:                       2.262   Cond. No.                     1.57e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.57e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

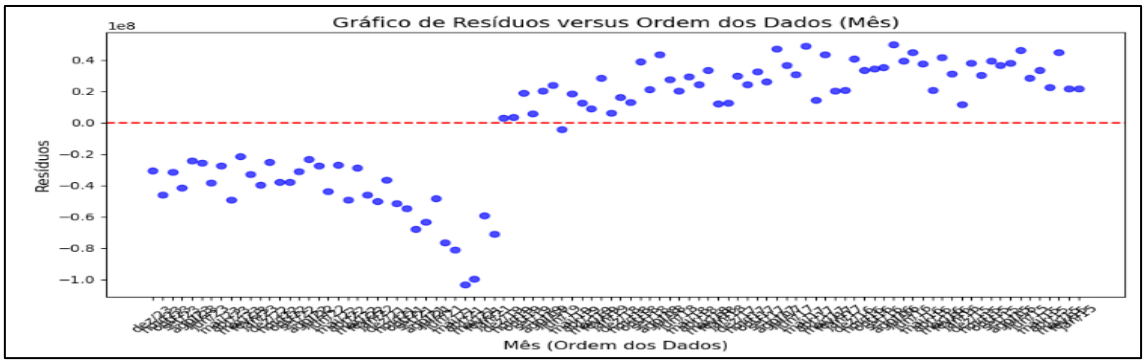
Iniciando a análise dos pressupostos pela Linearidade e Heterocedasticidade:

```
## VERIFICANDO PRESSUPOSTOS 1 E 3

'''
1-LINEARIDADE
A VARIAVEL RESPOSTA É LINEAR COM A VARIAVEL PREDITORA? NUM GRAFICO DE RETA AMBAS FICAM
PRÓXIMAS VISUALMENTE?

R: Este pressuposto não está sendo atendido, porque além de não possuir uma reta visualmente clara,
os residuos estão distribuidos de forma organizada em alguns pontos, indicando algum padrão nos residuos.


3-HOMOCEDASTICIDADE / IGUALDADE DAS VARIANCIAS
A VARIANCIA É IGUAL AO LONGO DA BASE DE DADOS ?  PORQUE DEVE HAVER DISPERSÃO
NOS DADOS, SEM PADRÕES OU CLUSTERS.

R: Este pressuposto não está sendo atendido, pois os residuos estão distribuidos de forma organizada
em alguns pontos, indicando algum padrão.
Isso se chama heterocedasticidade, o oposto da homocedasticidade, que seria nosso objetivo.
'''
```



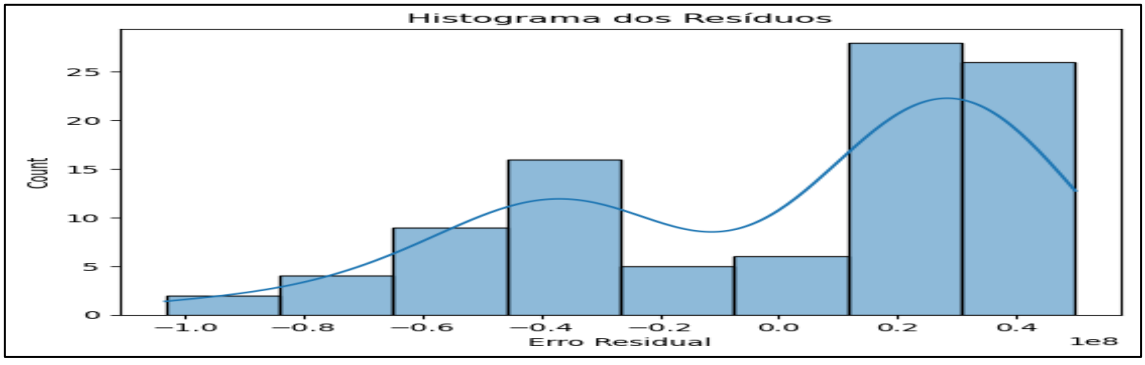
Analisando o pressuposto de Independência:

```
'''
2-INDEPENDENCIA DOS RESIDUOS, OS ERROS/RESIDUOS SÃO INDEPENDENTES ENTRE SI ? PORQUE SE TIVER ALGUM FATO
QUE MUDE O COMPORTAMENTO DOS DADOS, OS ERROS ENTÃO PODEM ESTAR INFLUENCIADOS.

        Durbin watson responde a essa questão, pois durbin Watson varia de 0 a 4
        se durbin Watson próximo de 2 não há correlação (AQUI É O OBJETIVO)

        Se Durbin Watson <2 correlação positiva
        Neste caso, os resíduos sucessivos tendem a ter valores semelhantes.
        Exemplo: se um resíduo é positivo, o próximo também tem alta probabilidade de ser positivo.
        indica que o modelo pode não estar capturando adequadamente a relação entre variáveis ou
        que há omissão de alguma variável importante.

        Se Durbin Watson >2 correlação negativa
        Neste caso, os resíduos sucessivos têm uma tendência a oscilar.
        Exemplo: se um resíduo é positivo, o próximo tende a ser negativo.
        Problemas potenciais: pode indicar instabilidade ou má especificação do modelo,
        o que exige revisão do modelo.

R: O modelo está com correlação positiva, pois tem resultado 0,222 em Durbin-Watson e visualmente ao
verificarmos o gráfico de RESIDUOS VERSUS A ORDEM (MÊS) temos muita alternância nos resíduos, muitos altos e baixos.
'''
```



Analisando o pressuposto de Normalidade:

```
'''
4- NORMALIDADE DOS RESIDUOS
O ERRO É NORMALMENTE DISTRUIDO? SE EU PLOTAR UM GRÁFICO ELE FICA PARECIDO COM UM SINO ?

        Para responder essa questão devemos analisar 3 situações (GRÁFICO, TESTE OMNIBUS e TESTE JARQUE-BERA)

        NO GRÁFICO: Visualmente a maioria dos resíduos está perto de 0 (pequenos desvios entre o
        valor observado e o previsto), tendo poucos resíduos muito positivos ou negativos

        NO TESTE OMNIBUS: p>0.05 não há evidência suficiente para rejeitar a normalidade, ou seja esta ok. Mas se
        p≤0.05 então rejeitamos a normalidade dos resíduos (pode haver problemas).

        NO TESTE DE JARQUE-BERA: p>0.05 não há evidência suficiente para rejeitar a normalidade, ou seja esta ok. Mas se
        p≤0.05 então rejeitamos a normalidade dos resíduos (pode haver problemas).


R: Este pressuposto não está sendo atendido em nenhuma das analises, OMNIBUS = 10.654, JARQUE-BERA = 8.764 e
visualmente o gráfico não se parece com um sino.
'''
```



Após realizado as análises dos pressupostos e verificação dos resultados do modelo, observamos que mesmo algumas variáveis tendo uma boa correlação entre si (CLIMA MÉDIO, PRECIPITAÇÃO e PERÍODO FÉRIAS) elas não são suficientes para prever/explicar o comportamento da quantidade de passageiros transportados no transporte público (Ônibus) de São Paulo.

Isso pode ser explicado, um pouco, devido ao tipo de modelo previamente utilizado para resolução deste objetivo. Modelos de regressão linear não são adequados para predições em variáveis temporais, existem outros tipos de modelos que trabalham melhor com series temporais como ARIMA/SARIMA.

Uma vez estressado as possibilidades com a regressão linear, o próximo passo é adequar o nosso problema objetivo, aos modelos que possuem esse comportamento mais assertivo no tocante as series temporais.

A principal mensagem aqui é que não existe uma receita de bolo perfeita, quando se trata de modelos de machine learning e que o teste de diferentes modelos, para um determinado objetivo é sempre o melhor caminho.

Os códigos em Python podem ser acessados no meu perfil do GITHUB, o link está disponível no texto da publicação, feita no LinkedIn:

www.linkedin.com/in/guilherme-lima-747355169

Bons estudos e obrigado por ter chegado até aqui !!