Questão

Ajuste um modelo linear generalizado de Poisson para a base de dados “danishlc.dat”. Esta base é composta pelo número de casos de câncer de pulmão em quatro cidades na Dinamarca entre os anos de 1968 e 1971. As informações disponíveis são: Número de casos (y), faixas etárias, população por faixas etárias e cidade. Procure caracterizar as faixas etárias e as cidades que apresentam as maiores taxas da doença.

Faça uma análise dos resultados em um relatório sucinto. O relatório deve incluir: análise descritiva, ajuste do modelo, análise dos “resíduos” e principais conclusões.

Questão 2

1- **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)**
2- **Análise exploratória**
    a. **Dados missing**
    b. **Distribuição da variável**
    c. **Scatter plot das variáveis numéricas com Claims**
3- **Modelagem**
    a. **Correlação de Spearman**
    b. **Teste de Shapiro-wilk**
    c. **Teste de Kruskal-Wallis**
    d. **Teste Mann-Whitney U**
    e. **Teste de qui-quadrado**
4- **Treinamento do modelo**
5- **Análise dos resíduos**
6- **Conclusão**

1- Informações básicas do dataset

a. Print das 5 primeiras linhas

```
danishlc
```

	Cases	Pop	Age	City
0	11	3059	40-54	Fredericia
1	11	800	55-59	Fredericia
2	11	710	60-64	Fredericia
3	10	581	65-69	Fredericia
4	11	509	70-74	Fredericia
5	10	605	>74	Fredericia

b. Tipo de dado em cada coluna

```
danishlc.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24 entries, 0 to 23
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Cases   24 non-null     int64
 1   Pop     24 non-null     int64
 2   Age     24 non-null     category
 3   City    24 non-null     category
dtypes: category(2), int64(2)
memory usage: 984.0 bytes
```

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

```
danishlc.describe()
```

	Cases	Pop
count	24.000000	24.000000
mean	9.333333	1100.333333
std	3.157691	842.232730
min	2.000000	509.000000
25%	7.000000	628.000000
50%	10.000000	791.000000
75%	11.000000	954.750000
max	15.000000	3142.000000

2- Análise exploratória

a. Dados missing

Dados missing

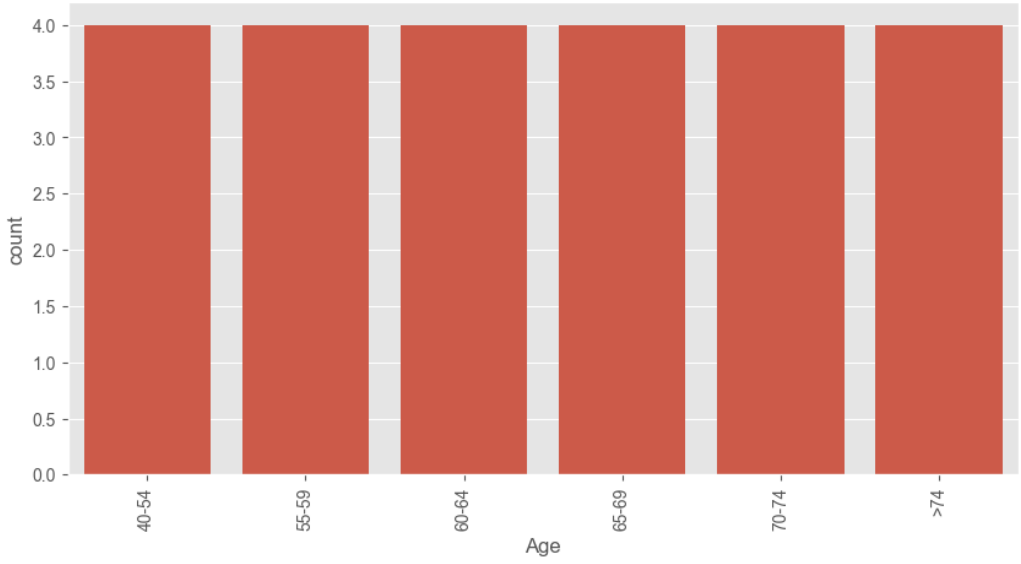
```
missing_values_table(danishlc)
```

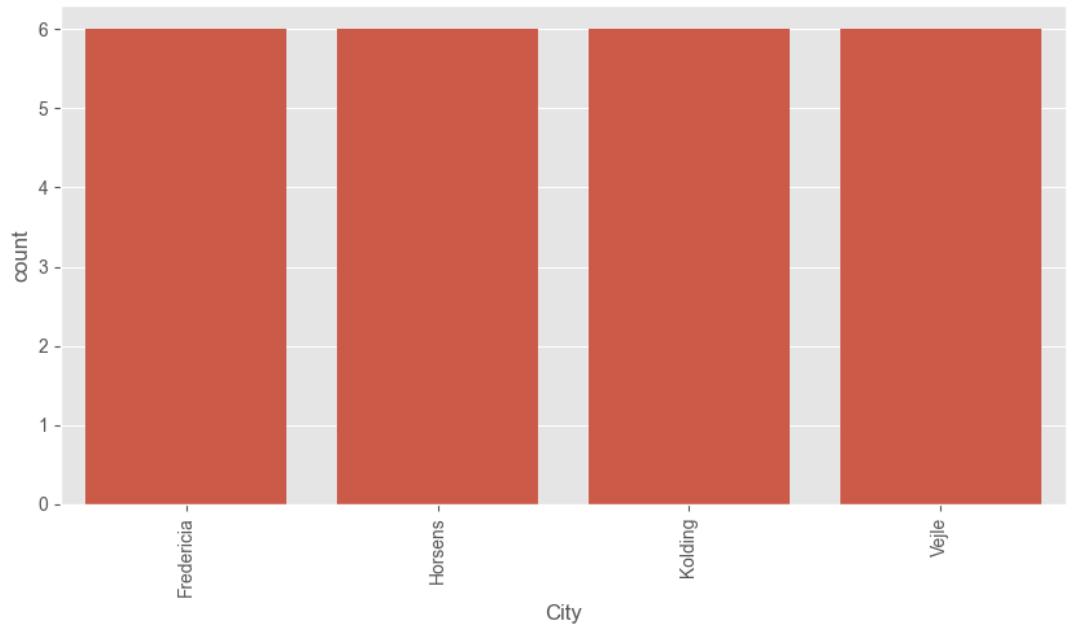
```
Your selected dataframe has 4 columns.
There are 0 columns that have missing values.
```

	Missing Values	% of Total Values

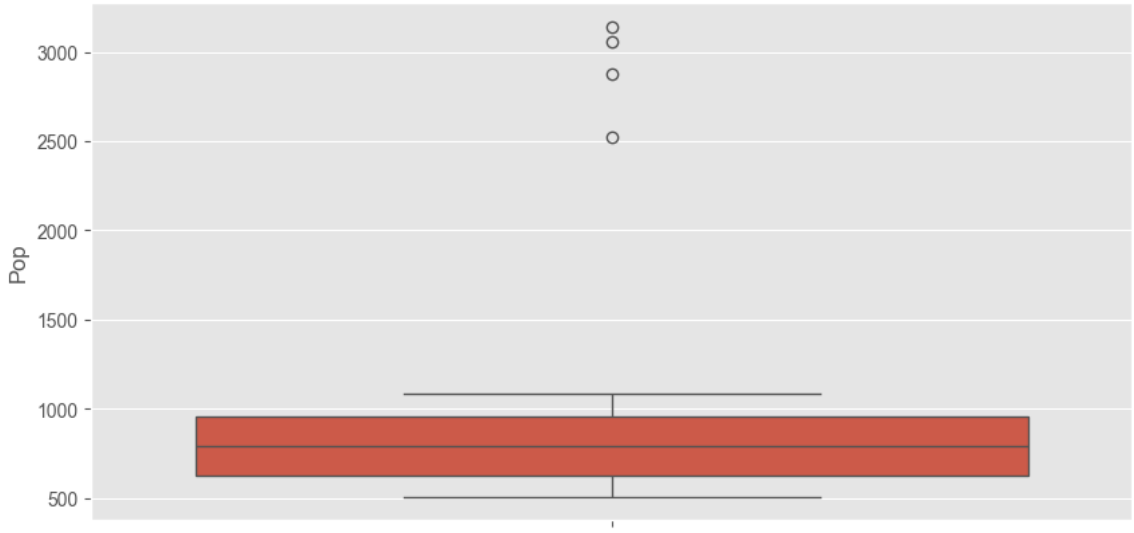
b. Gráficos

Countplots

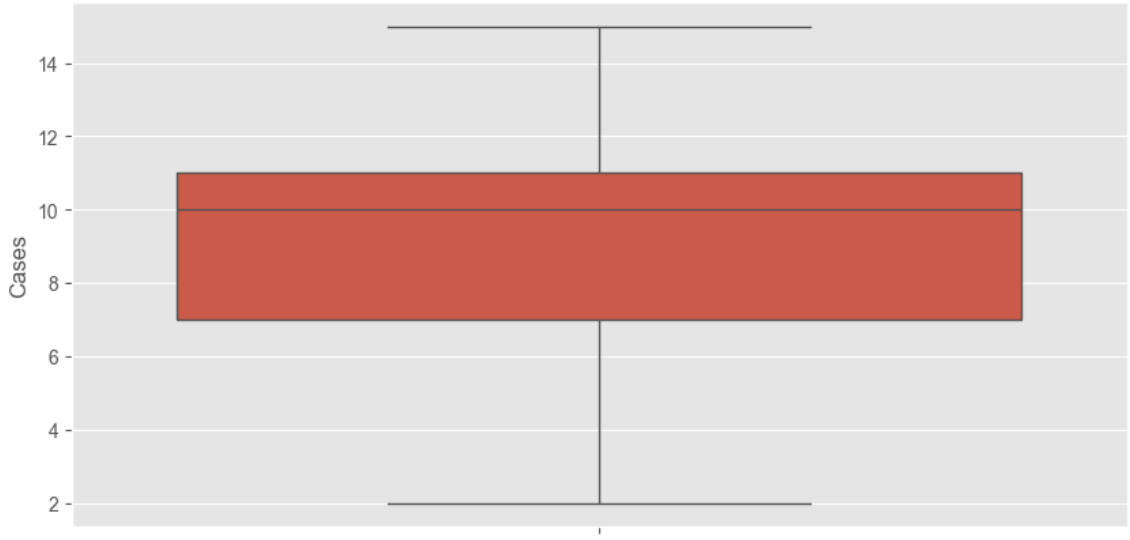




Boxplot



A variável Pop apresenta algumas discrepâncias, com outliers acima de 3000.

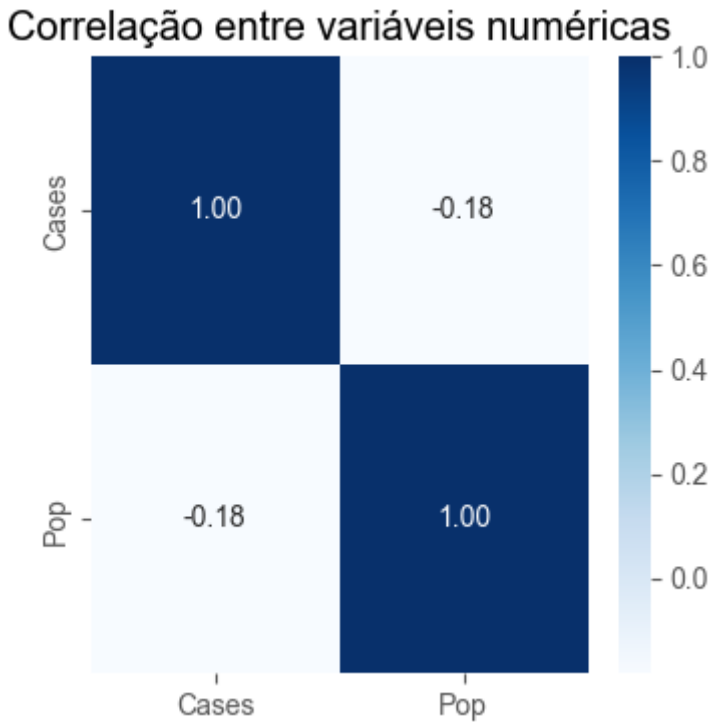


O número de casos de câncer de pulmão varia entre 2 e 15.

c. Scatter plot das variáveis numéricas com Claims

3- Modelagem

a. Correlação de Pearson



Correlação fraca

b. Teste de Shapiro-wilk

4- Treinamento do modelo

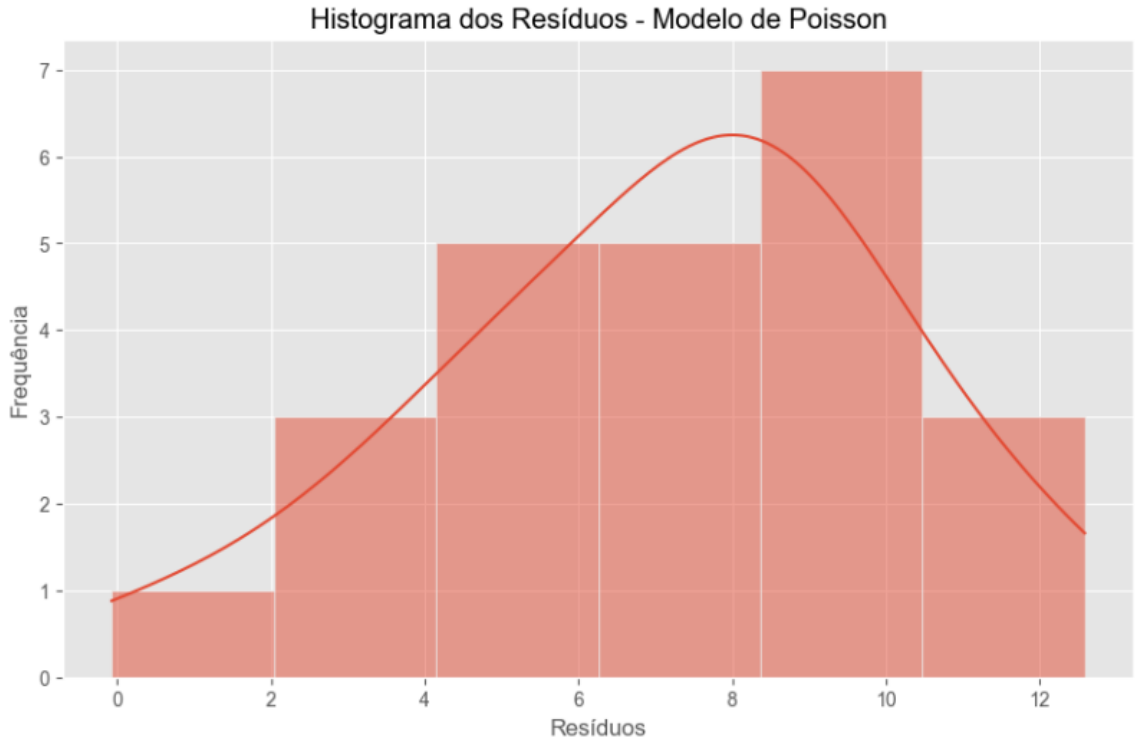
```python
# Normalizar variáveis numéricas
#scaler = StandardScaler()
#danishlc[['Pop']] = scaler.fit_transform(danishlc[['Pop']])

# Ajustar o modelo de Poisson
modelo_poisson = smf.poisson('Cases ~ Age + City', data=danishlc).fit(offset=np.log(danishlc['Pop']))

# Resumo do modelo
print(modelo_poisson.summary())
```

```
Optimization terminated successfully.
         Current function value: 2.438772
         Iterations 5
                          Poisson Regression Results
==============================================================================
Dep. Variable:                  Cases   No. Observations:                   24
Model:                        Poisson   Df Residuals:                       15
Method:                           MLE   Df Model:                            8
Date:                Sun, 23 Jun 2024   Pseudo R-squ.:                0.05666
Time:                        18:19:32   Log-Likelihood:                -58.531
converged:                       True   LL-Null:                       -62.046
Covariance Type:            nonrobust   LLR p-value:                    0.5333
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      2.2437      0.204     11.019      0.000       1.845       2.643
Age[T.55-59]  -0.0308      0.248     -0.124      0.901      -0.517       0.455
Age[T.60-64]   0.2647      0.231      1.144      0.253      -0.189       0.718
Age[T.65-69]   0.3102      0.229      1.353      0.176      -0.139       0.759
Age[T.70-74]   0.1924      0.235      0.818      0.413      -0.269       0.653
Age[T.>74]    -0.0625      0.250     -0.250      0.803      -0.553       0.428
City[T.Horsens] -0.0984    0.181     -0.543      0.587      -0.454       0.257
City[T.Kolding] -0.2271    0.188     -1.210      0.226      -0.595       0.141
City[T.Vejle]  -0.2271      0.188     -1.210      0.226      -0.595       0.141
```

O valor de Pseudo R-quadrado (0.05666) é baixo

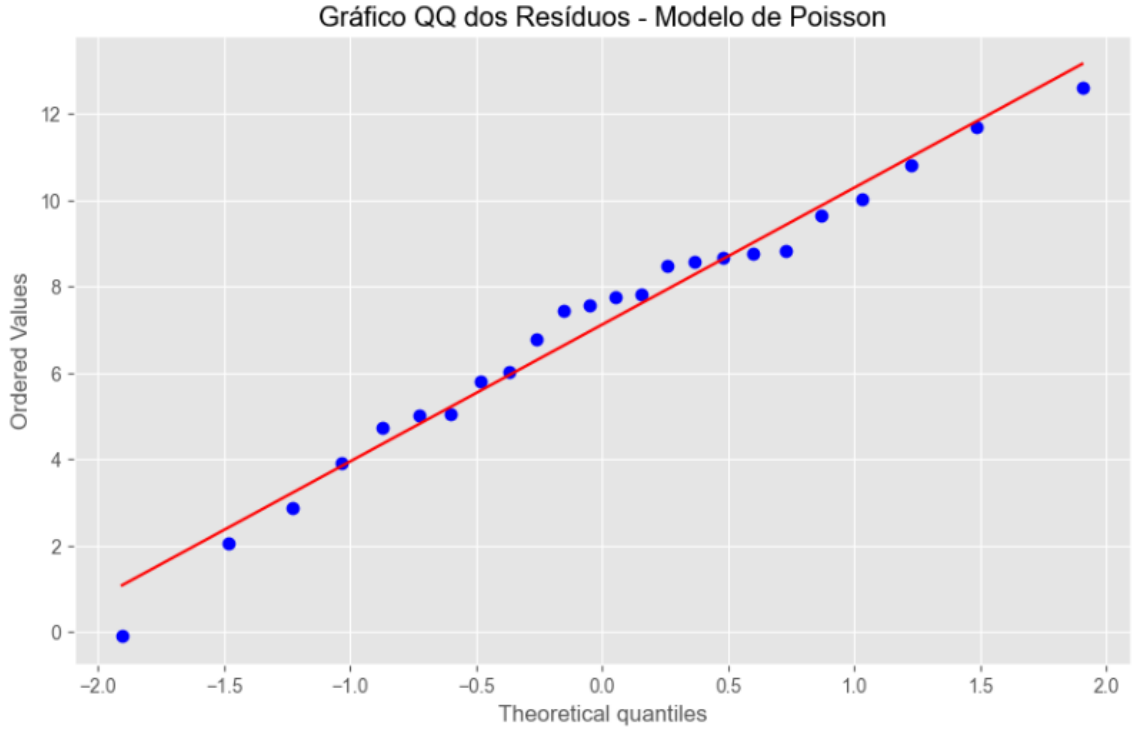
Nenhum dos coeficientes das variáveis categóricas são estatisticamente significativo,

5- Análise dos resíduos



O histograma dos resíduos mostra uma distribuição levemente assimétrica com uma tendência a valores mais altos, mas sem outliers extremos evidentes.

O gráfico de dispersão dos resíduos versus os valores ajustados mostra uma leve tendência a homocedaasticidade. Os resíduos estão razoavelmente distribuídos ao redor da linha zero, embora haja uma concentração de resíduos positivos.



O gráfico QQ dos resíduos mostra que os resíduos estão, em grande parte, alinhados com a linha de normalidade, exceto para os valores mais extremos. Isso sugere que os resíduos seguem aproximadamente uma distribuição normal.

6- Conclusão

Os resíduos estão razoavelmente distribuídos ao redor da linha zero, sem um padrão claro de heterocedasticidade. O gráfico QQ mostrou que os resíduos seguem aproximadamente uma distribuição normal, com algumas discrepâncias nas caudas.

O modelo de Poisson ajustado não identificou coeficientes estatisticamente significativos, sugerindo que as variáveis preditoras escolhidas não têm um impacto significativo.