

Ajuste um modelo linear generalizado de Poisson para a base de dados “motorins.dat”. A variável resposta “Claims” representa o número de reclamos de uma seguradora. As variáveis preditoras são: Kilometers: Kilômetros percorridos em um ano.

Zone: Região Geográfica (Zona).

Bonus: Igual ao número de anos sem reclamos mais 1.

Make: Representa 8 tipos de modelos de automóveis mais o tipo 9 (outros)

Insured: Número de veículos segurados

Claims: Número de reclamos

Payment: Pagamento Total do

1: less than 1000

2: from 1000 to 15 000

3: 15 000 to 20 000

4: 20 000 to 25 000

5: more than 25 000

1: Stockholm, Göteborg, Malmö with surroundings

2: Other large cities with surroundings

3: Smaller cities with surroundings in southern Sweden

4: Rural areas in southern Sweden

5: Smaller cities with surroundings in northern Sweden

6: Rural areas in northern Sweden

7: Gotland

Roteiro da atividade

Roteiro

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)

3- Análise exploratória

a. Dados missing

b. Distribuição da variável

c. Scatter plot das variáveis numéricas com Claims

4- 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

5- Treinamento do modelo

6- Análise dos resíduos

7- Conclusão

1- Carregando bibliotecas

Pacotes

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

import statsmodels.api as sm
import statsmodels.formula.api as smf

from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import PoissonRegressor
from sklearn.preprocessing import PolynomialFeatures
from sklearn.pipeline import make_pipeline

from scipy.stats import mannwhitneyu, chi2_contingency, norm, kruskal

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

2- Informações básicas do dataset

a. Print das 5 primeiras linhas

```
In [ ]: df1.head()
```

	Kilometers	Zone	Bonus	Make	Insured	Claims	Payment
0	less than 1000	Stockholm, Göteborg, Malmö with surroundings	1	modelo 1	455.13	108	392491
1	less than 1000	Stockholm, Göteborg, Malmö with surroundings	1	modelo 2	69.17	19	46221
2	less than 1000	Stockholm, Göteborg, Malmö with surroundings	1	modelo 3	72.88	13	15694
3	less than 1000	Stockholm, Göteborg, Malmö with surroundings	1	modelo 4	1292.39	124	422201
4	less than 1000	Stockholm, Göteborg, Malmö with surroundings	1	modelo 5	191.01	40	119373

b. Tipo de dado em cada coluna

```
df1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2182 entries, 0 to 2181
Data columns (total 7 columns):
#   Column      Non-Null Count  Dtype  
---  -
0   Kilometers  2182 non-null   object  
1   Zone        2182 non-null   object  
2   Bonus       2182 non-null   int64   
3   Make        2182 non-null   object  
4   Insured     2182 non-null   float64  
5   Claims      2182 non-null   int64   
6   Payment     2182 non-null   int64   
dtypes: float64(1), int64(3), object(3)
memory usage: 119.5+ KB
```

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

```
df1.describe()
```

	Bonus	Insured	Claims	Payment
count	2182.000000	2182.000000	2182.000000	2.182000e+03
mean	4.015124	1092.195270	51.865720	2.570076e+05
std	2.000516	5661.156245	201.710694	1.017283e+06
min	1.000000	0.010000	0.000000	0.000000e+00
25%	2.000000	21.610000	1.000000	2.988750e+03
50%	4.000000	81.525000	5.000000	2.740350e+04
75%	6.000000	389.782500	21.000000	1.119538e+05
max	7.000000	127687.270000	3338.000000	1.824503e+07

3- Análise exploratória

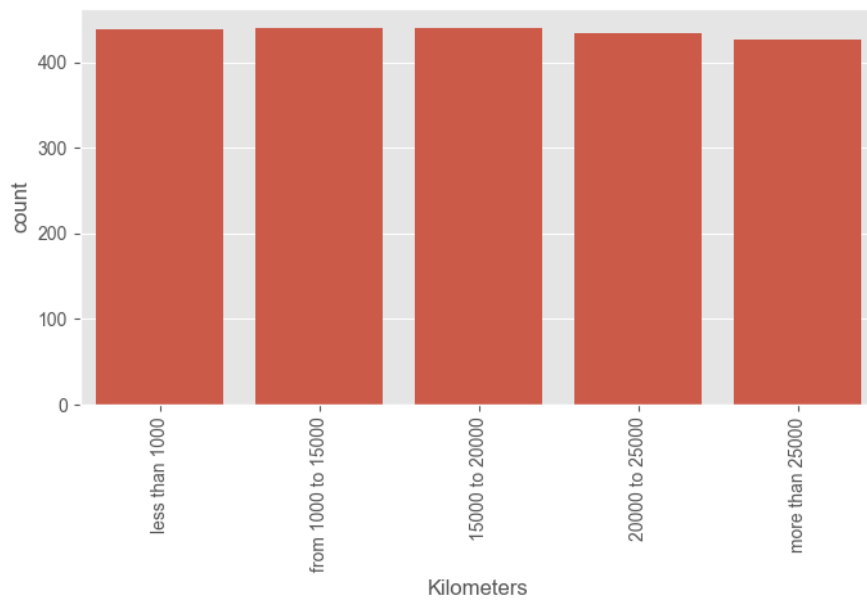
a. Dados missing

```
missing_values_table(df1)
```

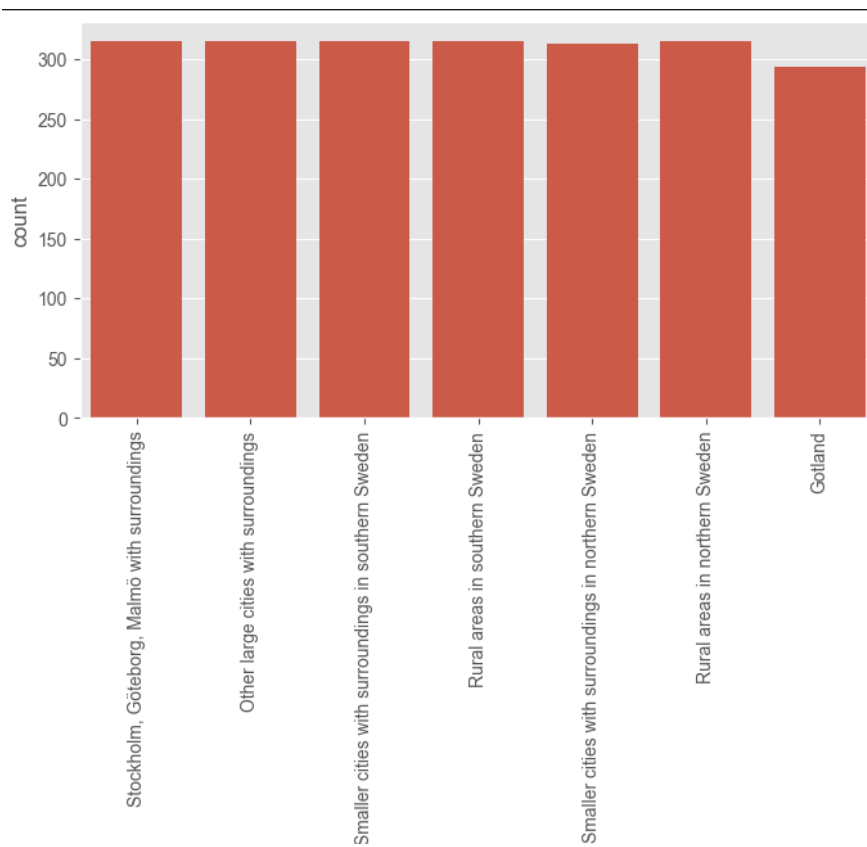
Your selected dataframe has 7 columns.
There are 0 columns that have missing values.

Missing Values	% of Total Values
----------------	-------------------

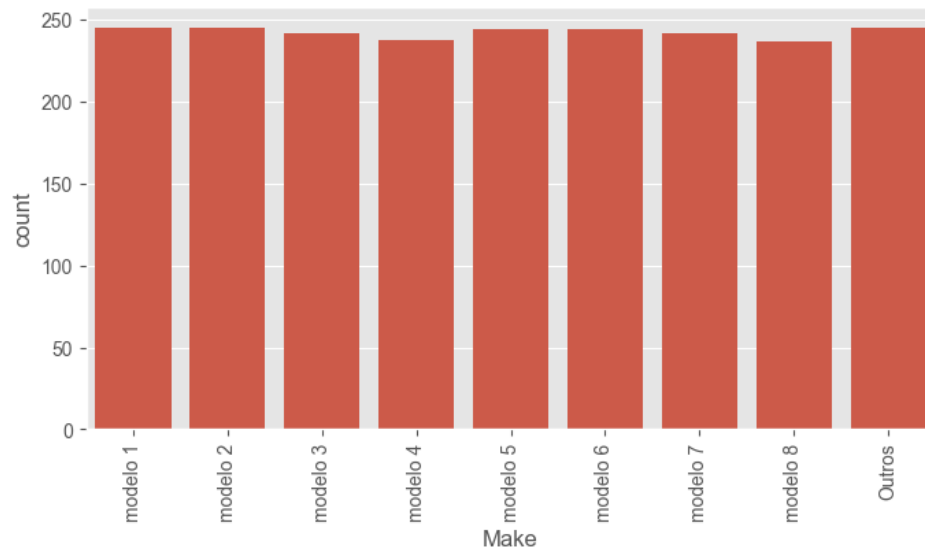
b. Countplots



A distribuição das categorias de quilômetros é bastante uniforme.

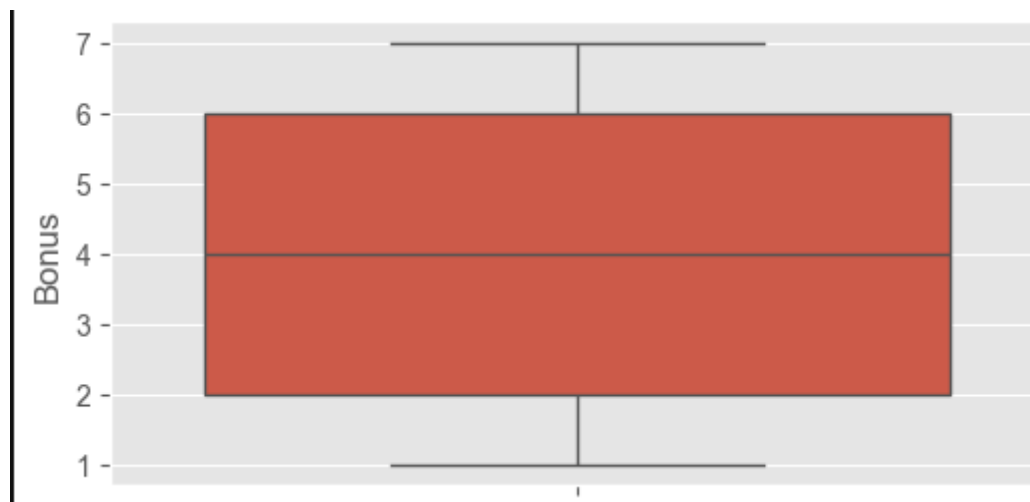


A distribuição das zonas é também relativamente uniforme. Gotland possui um entorno de 20 casos a menos

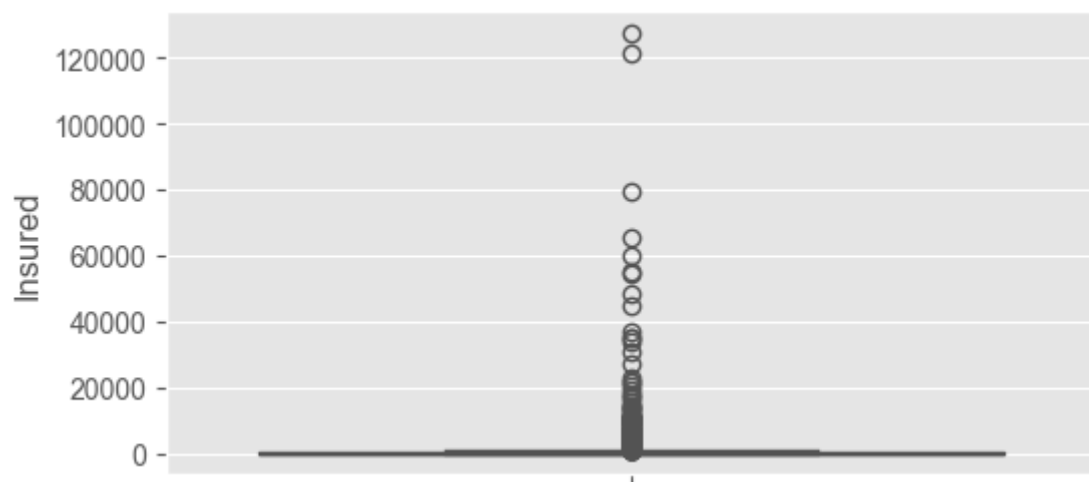


A distribuição dos modelos de automóveis também é bastante uniforme.

c. Boxplots

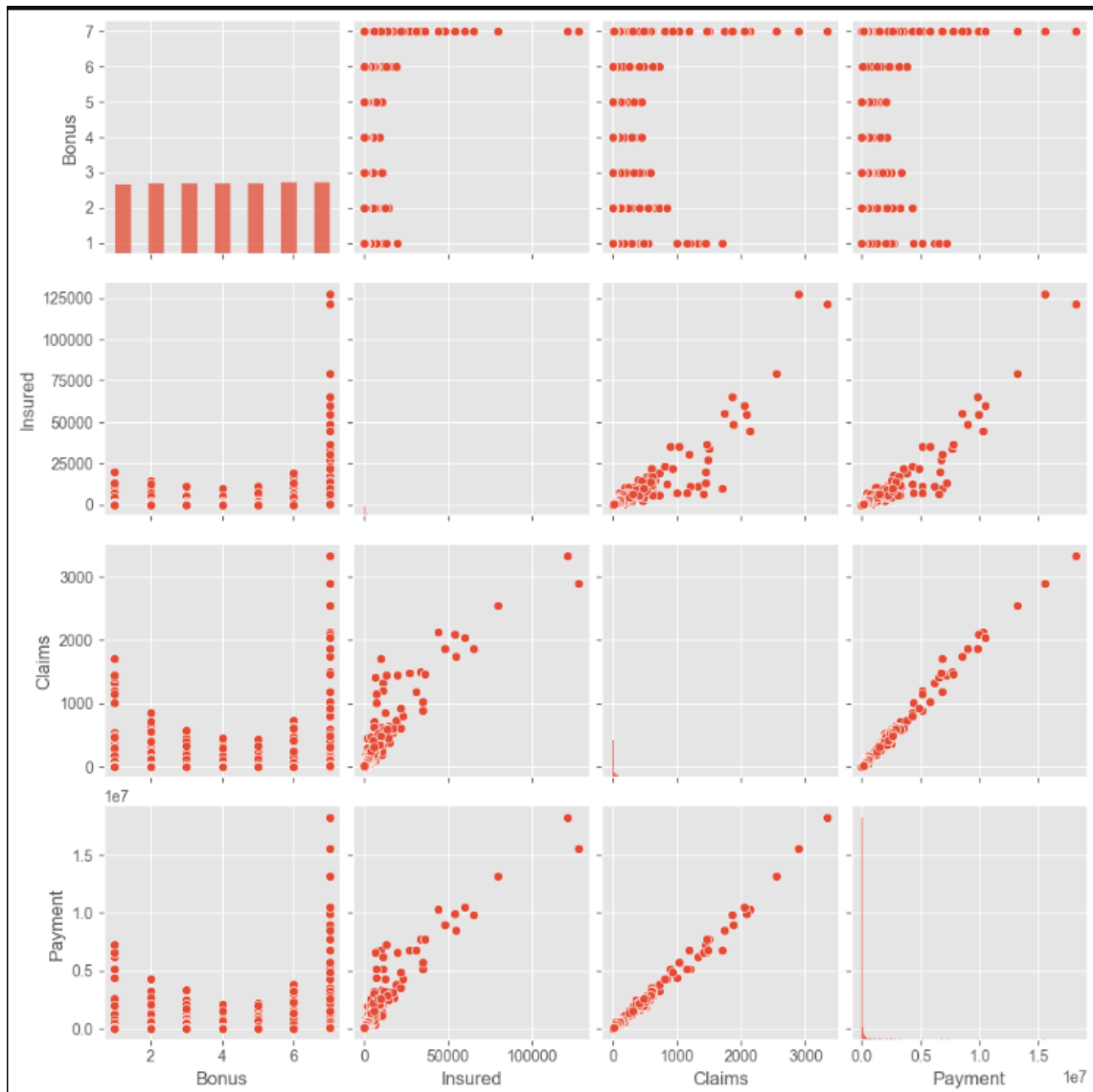


A mediana do bônus é de 4 anos, com um intervalo interquartil de 2 a 6 anos.



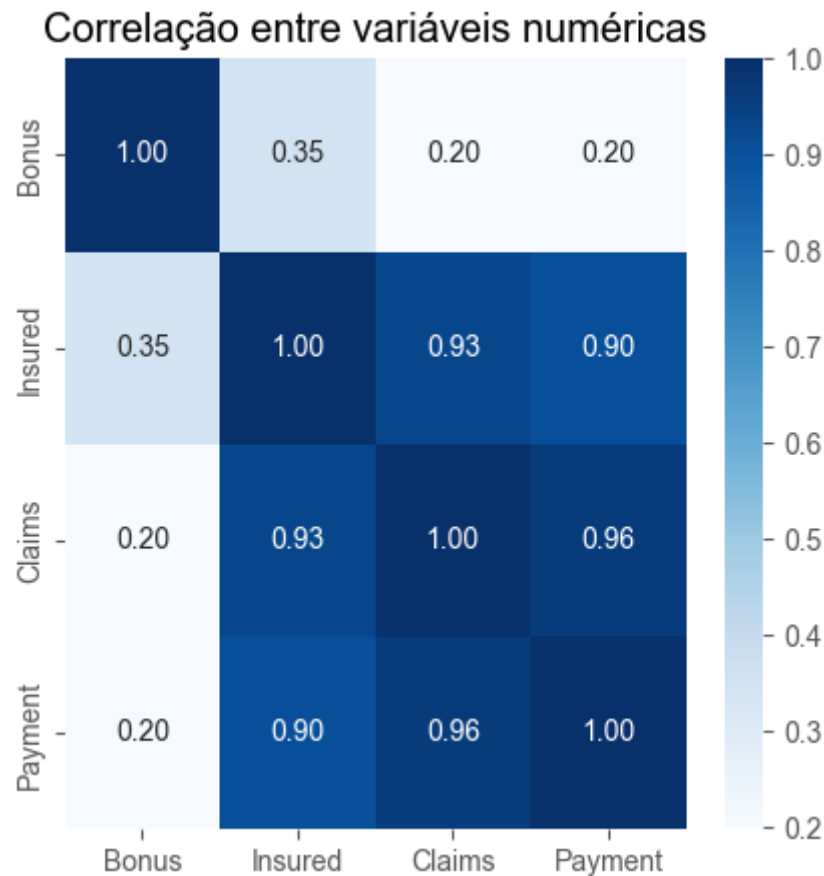
No caso de insured temo a presenta de alguns valores outliers apesar de grande parte dos valores estarem concentrados abaixo de 22

d. Scatter plot das variáveis numéricas com Claims



4- Modelagem

a. Correlação de Spearman



O destaque que merece ser feito aqui é na multicolinearidade entre a variável **Insured** e **Claims**, nossa variável a ser predita.

b. Teste de Shapiro-wilk

O teste de Shapiro-Wilk para a variável **Claims** resultou em:

- $W = 0.2469$
- $p\text{-valor} = 0.0$

Isso indica que a variável **Claims** não segue uma distribuição normal.

c. Teste de Kruskal-Wallis

O teste de Kruskal-Wallis foi realizado para verificar se há diferenças significativas entre as categorias das variáveis "Kilometers", "Zone" e "Make" em relação à variável resposta "Claims".

Teste de Kruskal-Wallis para Kilometers:

KruskalResult(statistic=195.72553167450306, pvalue=3.117263894895815e-41)

Teste de Kruskal-Wallis para Zone: KruskalResult(statistic=542.9041459870552, pvalue=4.7798986253134086e-114)

Teste de Kruskal-Wallis para Make: `KruskalResult(statistic=691.6308951948237, pvalue=4.5332978426564537e-144)`

Há uma diferença significativa entre as diferentes categorias de quilômetros percorridos anualmente em relação ao número de reclamações. O mesmo aconteceu com as demais variáveis Zone e Make

d. Teste Mann-Whitney U

Variáveis numéricas com resultados significativos no teste Mann-Whitney U:

Bonus: p-valor = 1.4647389332811162e-05

Insured: p-valor = 9.184641280479907e-170

O teste de Mann-Whitney U foi realizado para comparar duas variáveis numéricas em relação à variável resposta "Claims". Há uma diferença significativa entre o bônus (número de anos sem reclamações mais 1) e o número de reclamações. E o mesmo acontece com o número de veículos segurados e o número de reclamações

e. Teste de qui-quadrado

Variáveis categóricas com resultados significativos no teste de qui-quadrado:

Kilometers: p-valor = 0.0011081654037870712

Zone: p-valor = 2.1050504276507513e-24

Make: p-valor = 7.774440399334877e-14

O resultado do teste de qui-quadrado corroborou com o de Kruskal-Wallis

5- Treinamento do modelo

Foi implementado a normalização dos dados na variável Bonus. A variável Insured não foi utilizada como uma variável direta do modelo, foi implementado como um offset.


```
# Normalizar variáveis numéricas
scaler = StandardScaler()
df1[['Bonus']] = scaler.fit_transform(df1[['Bonus']])

# Ajustar o modelo de Poisson com offset
model = smf.poisson('Claims ~ Kilometers + Zone + Bonus + Make', data=df1).fit(offset=np.log(df1['Insured']))

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

Optimization terminated successfully.

Current function value: 15.980153

Iterations 12

Poisson Regression Results

```
=====
Dep. Variable:      Claims    No. Observations:      2182
Model:              Poisson  Df Residuals:          2162
Method:             MLE      Df Model:             19
Date:               Sun, 23 Jun 2024    Pseudo R-squ.:      0.8426
Time:               16:01:23    Log-Likelihood:     -34869.
converged:          True      LL-Null:            -2.2157e+05
Covariance Type:    nonrobust    LLR p-value:        0.000
=====
```

	coef	std err	z	P> z	[0.025	0.975]
Intercept	2.5427	0.041	62.586	0.000	2.463	2.622
Kilometers[T.20000 to 25000]	-0.9732	0.012	-78.763	0.000	-0.997	-0.949
Kilometers[T.from 1000 to 15000]	0.4998	0.008	60.937	0.000	0.484	0.516
Kilometers[T.less than 1000]	0.3289	0.008	38.760	0.000	0.312	0.346
Kilometers[T.more than 25000]	-1.1311	0.013	-86.330	0.000	-1.157	-1.105
Zone[T.Other large cities with surroundings]	3.5339	0.041	86.741	0.000	3.454	3.614
Zone[T.Rural areas in northern Sweden]	2.8036	0.041	67.791	0.000	2.723	2.885
Zone[T.Rural areas in southern Sweden]	3.9382	0.041	97.120	0.000	3.859	4.018
Zone[T.Smaller cities with surroundings in northern Sweden]	2.2609	0.042	53.579	0.000	2.178	2.344
Zone[T.Smaller cities with surroundings in southern Sweden]	3.4678	0.041	85.035	0.000	3.388	3.548
Zone[T.Stockholm, Göteborg, Malmö with surroundings]	3.6182	0.041	88.910	0.000	3.538	3.698
Make[T.modelo 1]	-1.9761	0.010	-199.649	0.000	-1.996	-1.957
Make[T.modelo 2]	-3.4185	0.019	-176.306	0.000	-3.457	-3.381
Make[T.modelo 3]	-3.8153	0.024	-162.191	0.000	-3.861	-3.769
Make[T.modelo 4]	-3.7027	0.022	-166.223	0.000	-3.746	-3.659
Make[T.modelo 5]	-3.2995	0.018	-180.236	0.000	-3.335	-3.264
Make[T.modelo 6]	-2.8891	0.015	-192.039	0.000	-2.919	-2.860
Make[T.modelo 7]	-3.6496	0.022	-168.226	0.000	-3.692	-3.607
Make[T.modelo 8]	-4.3305	0.030	-142.886	0.000	-4.390	-4.271
Bonus	0.4322	0.003	137.392	0.000	0.426	0.438

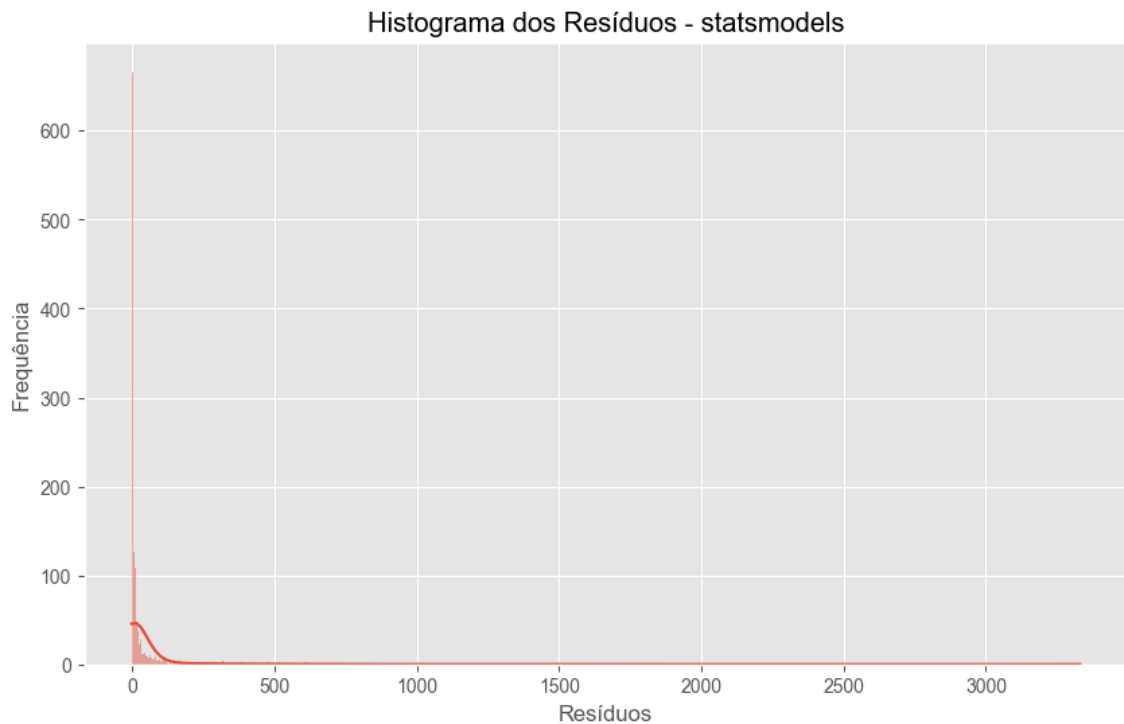
```
=====
```

Os coeficientes negativos para algumas categorias de "Kilometers" (Kilometers[T.20000 to 25000] e Kilometers[T.more than 25000]) "Make"(todos os modelos) indicando que influência de forma negativa para o resultado da nossa regressão.

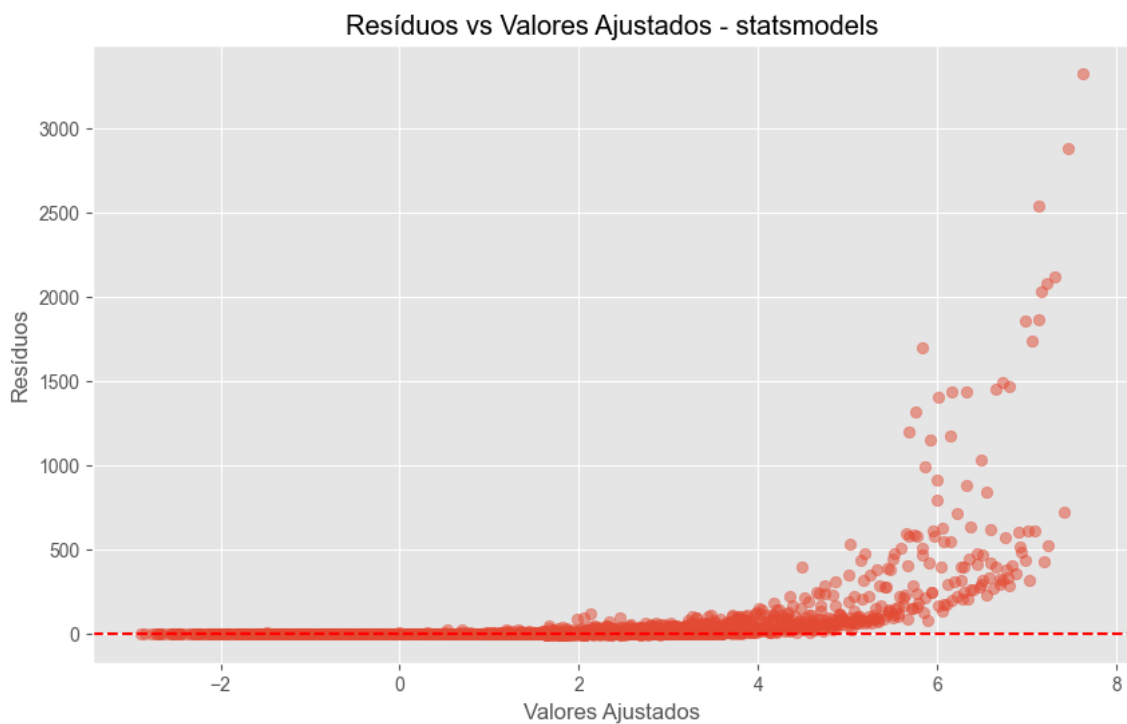
Dialogando com os testes executados anteriormente todos os coeficientes são estatisticamente significativos ($p < 0.001$), indicando que as variáveis preditoras escolhidas têm um efeito significativo no número de reclamações.

Por fim, o valor de R-quadrado (0.8426) indica que o modelo explica uma proporção substancial da variabilidade nos dados de reclamações.

6 - Análise dos resíduos

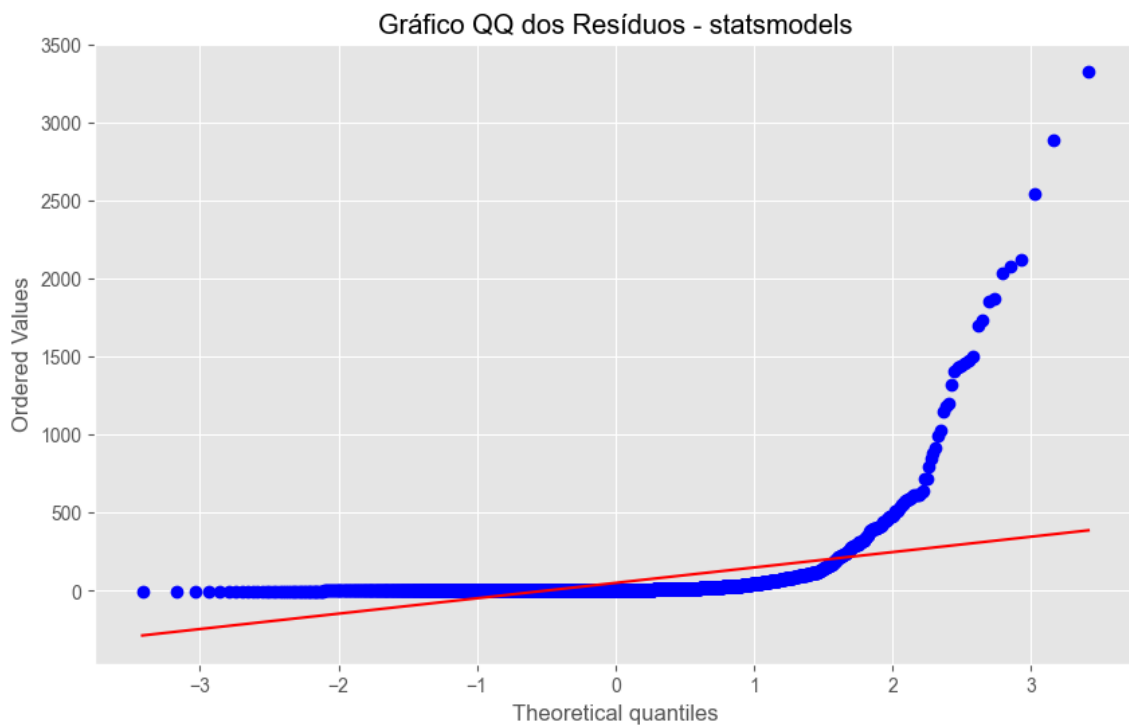


O histograma dos resíduos mostra uma distribuição assimétrica, com muitos resíduos próximos de zero e alguns valores bastante altos. Talvez o modelo tenha subestimado alguns valores.



O gráfico de dispersão dos resíduos versus os valores ajustados revela um padrão claro de heterocedasticidade (cone). Os resíduos aumentam conforme os valores

ajustados aumentam, o que indica que a variabilidade dos resíduos não é constante. Por fim, destacar a presença de alguns outliers para os valores mais altos



O gráfico QQ dos resíduos mostra que os resíduos não seguem uma distribuição normal. Os pontos divergem significativamente da linha de normalidade, especialmente nas caudas. Isso reforça a evidência de que os resíduos são assimétricos e possuem valores extremos (outliers).

7 - Conclusão

A análise descritiva inicial dos dados revelou importantes insights sobre as variáveis preditoras e a variável resposta. Primeiro a uniformidade na distribuição das variáveis categóricas ("Kilometers"), zonas geográficas ("Zone") e modelos de automóveis ("Make").

Já quando olhamos para as numéricas Bonus apresentou uma mediana de 4 anos, indicando que muitos segurados tiveram períodos consideráveis sem reclamações. A variável Insured mostrou uma alta variabilidade e presença de outliers.

Os testes estatísticos realizados (Kruskal-Wallis, Mann-Whitney U e qui-quadrado) indicaram que todas as variáveis preditoras (Kilometers, Zone, Make, Bonus, Insured) possuem associações significativas com a variável resposta "Claims".

O modelo de regressão de Poisson foi ajustado com as variáveis preditoras "Kilometers", "Zone", "Bonus" e "Make", e um offset logarítmico da variável "Insured". Todos os coeficientes das variáveis preditoras foram estatisticamente significativos ($p < 0.001$), indicando que essas variáveis têm um efeito significativo no número de reclamações.

O valor de 0.8426 do R-quadrado indica que o modelo explica uma proporção substancial

Os resíduos mostraram um padrão claro de heterocedasticidade, onde a variância dos resíduos aumenta com os valores ajustados.

O histograma e o gráfico QQ indicaram que os resíduos são assimétricos e possuem valores extremos, sugerindo que o modelo subestima significativamente o número de reclamações em alguns casos.