Lista 09

Matheus Cougias e Klysman Rezende

25/09/2020

Leitura de Dados

Realiza a leitura da base de dados motorins.dat, selecionando somente as variáveis desejadas para análise (excluindo a sétima variável de Payment). Dessa maneira, as variáveis consideradas como factor são transformadas e trocou-se seus valores para o significado real daquela variável para melhor compreender os resultados gerados. Para facilitar a análise futura, serão criadas duas variáveis, uma da quantidade de clientes que não acionaram o seguro (total de clientes - número de reclamos) e da proporção de reclamos por clientes totais.

```
dados <- read.csv("motorins.dat", sep="\t")
dados <- subset(dados, select = -c(7))

dados$Kilometres <- factor(dados$Kilometres)
levels(dados$Kilometres) <-  c("< 1000", "1000 ~ 15000", "15000 ~ 20000", "20000 ~ 25000", "> 25000")

dados$Zone <- factor(dados$Zone)
levels(dados$Zone) <- c("Stockholm...", "Large cities", "Pequenas cidades do sul", "Area rural do sul",
                                  "Area rural do norte", "Gotland")

dados$Make <- factor(dados$Make)
levels(dados$Make) <- c("Tipo 1", "Tipo 2", "Tipo 3", "Tipo 4", "Tipo 5", "Tipo 6", "Tipo 7", "Tipo 8",

dados$Bonus <- as.factor(dados$Bonus)

dados$Prop <- dados$Claims/dados$Insured
```

Análise Exploratória

A partir da análise dos gráficos, pode-se transformar a variável numérica Bonus em uma variável de factor, pois os valores apresentados são números inteiros no intervalo de 1 a 7 anos, sendo assim, foi feita a alteração na parte anterior do código para melhor visualização dos dados.

A partir do primeiro gráfico plotado (historgama das reclamos), é possível identificar que a o baixo número de reclamos é a principal faixa no banco de dados. Pelo fato do número de clientes ser diferente em cada uma das linhas do banco de dados, também é interessante a análise em relação à proporção de reclamos em relação ao número total de clientes para cada amostra coletada. Assim como no primeiro gráfico, o segundo gráfico (histograma das proporções) também demonstra que as faixas de baixo índice de reclamos são predominantes na base de dados. Ambos os gráficos apresentam um comportamento que pode ser levado como uma distribuição exponencial.

Causados pelo comportamento dos histograma aspresentados, os demais gráficos de análise do comportamento das variáveis preditoras e a variável de interesse possuem seus 3 primeiros quantis nas faixas mais

baixas de valor. Resta então realizar uma análise de quais grupos se encontram um pouco destoados dos demais.

Em relação à variável Kilometres, dois intervalos se destacaram com uma maior quantidade de pontos de alto valor em seu quarto quantil: os carros com rodagem entre 15000 e 20000km e os carros com rodagem acima dos 25000km. Uma provável explicação para o resultado gerado é o fato de veículos com uma maior rodagem são os que possuem maior desgaste de seus componentes, aumentando as probabilidades de ocorrência de um defeito (ou avaria) que levará ao acionamento do seguro por parte do assegurado.

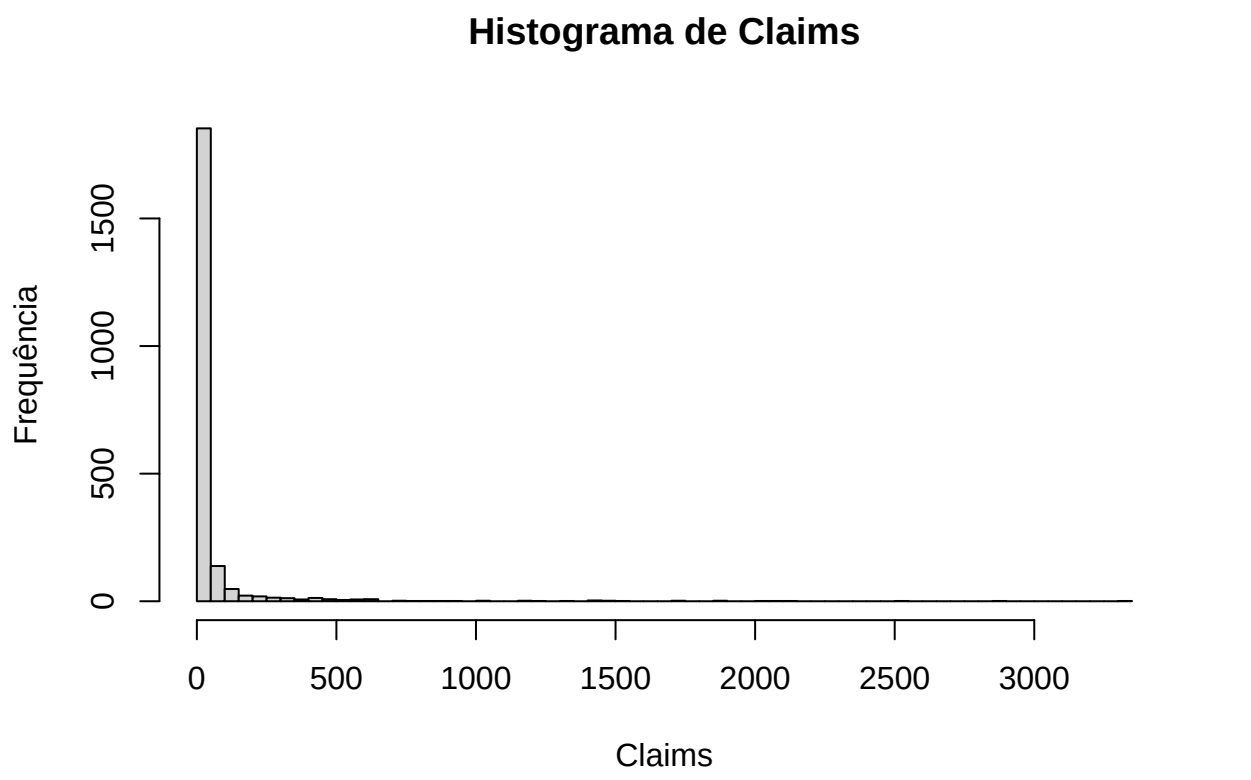
Quanto tomamos a comparação da proporção de acionamento com as Zonas do país, a Zona 7 (Gotland) se destacou com a maior quantidade de altos valores acima do terceito quantil. Por outro lado, a Zona 1 (Estolcomo) apresentou a maior média em comparação com as demais zonas, apesar de não estar muito acima delas.

Ao analisar a variável Bonus, tem-se que a maior variação de valores de reclamos ,juntamente com a maior média, é dada pelos clientes que estão na primeira categoria (Bonus 1). Esta faixa é composta pelos novos clientes da seguradora e aqueles que, no último ano, realizaram um acionamento da mesma. Essa característica consegue explicar o motivo pelo grupo apresentar uma maior média, pois os novos clientes (provavelmente inexperientes) encontram-se no grupo, juntamente com os que realizam reclamos recorrentes.

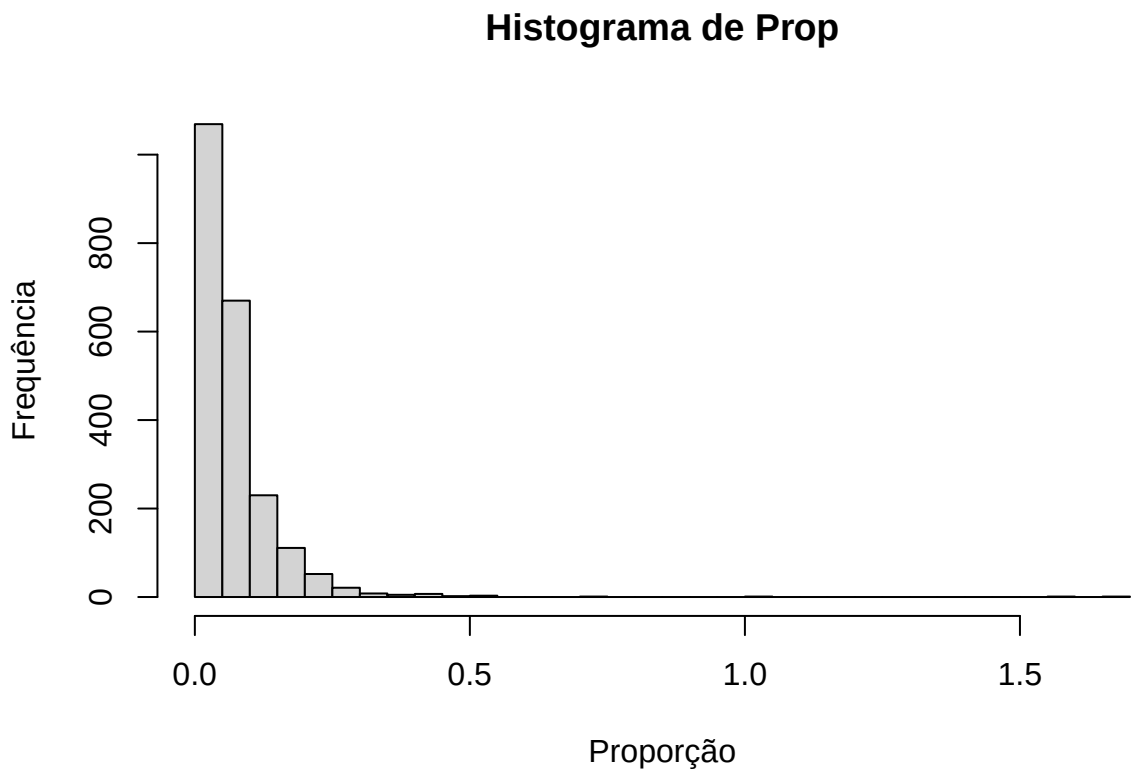
Sobre os valores apresentados pela variável Make, os tipos 5, 7 e 8 de veículos são os que apresentam as maiores médias de propabilidade de reclamos. O tipo 7 também se descata por possuir os maiores valores no quarto quantil do gráfico.

No caso do último gráfico, foi utilizada a comparação entre a variável resposta Claims e a variável preditora Insured. O gráfico nos mostra que, quanto maior a quantidade de clientes assegurados, maior tende a ser a quantidade de reclamos realizados.

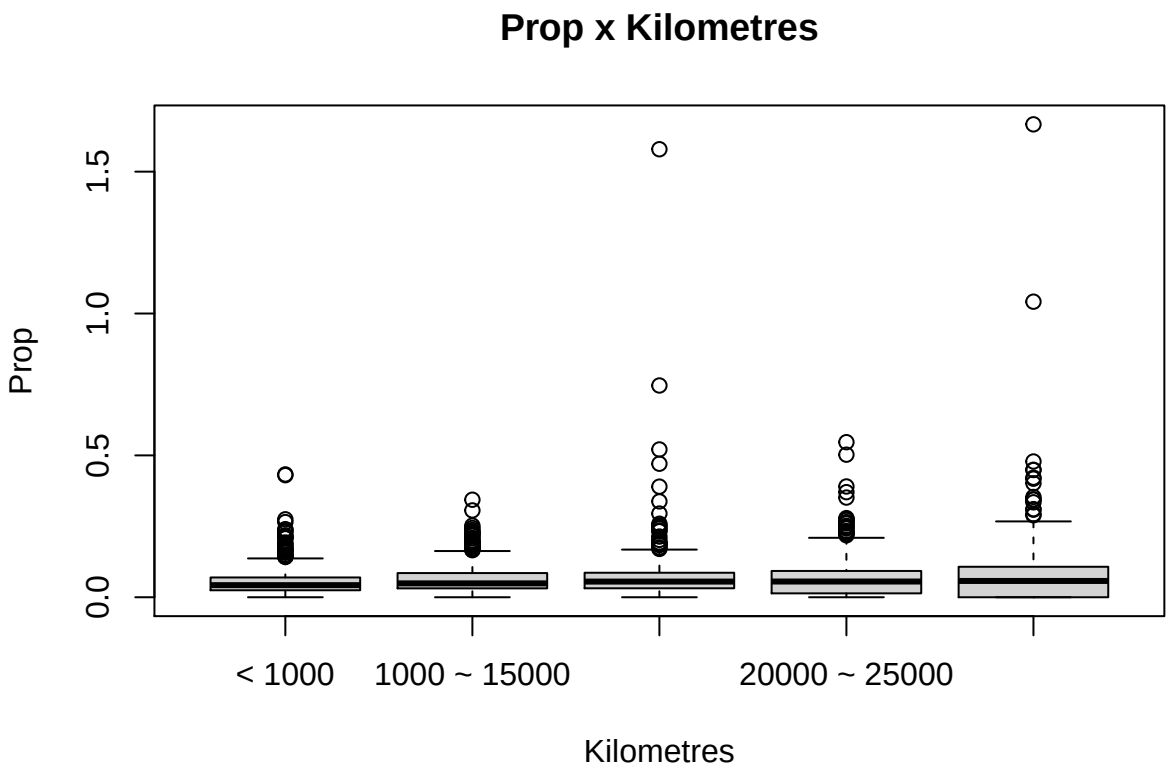
```
hist(dados$Claims, main = "Histograma de Claims", ylab = "Frequência", xlab = "Claims", breaks = 50)
```



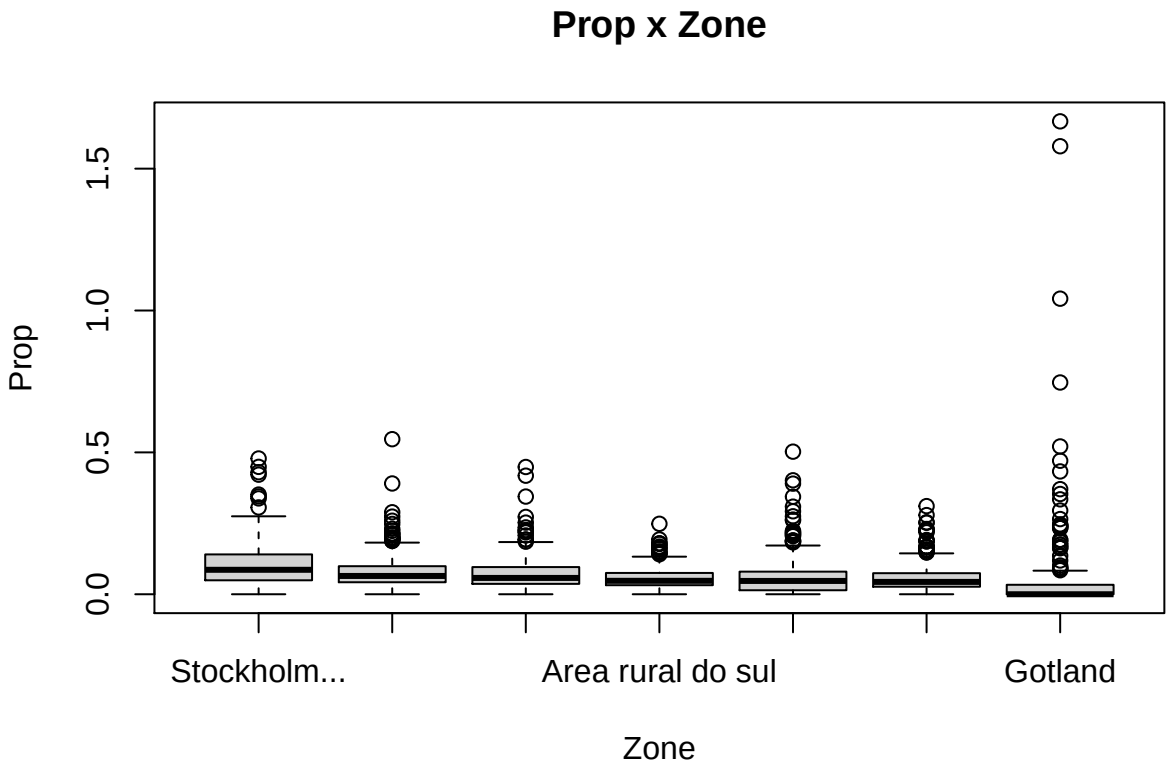
```r
hist(dados$Prop, main = "Histograma de Prop", ylab = "Frequência", xlab = "Proporção", breaks = 50)
```



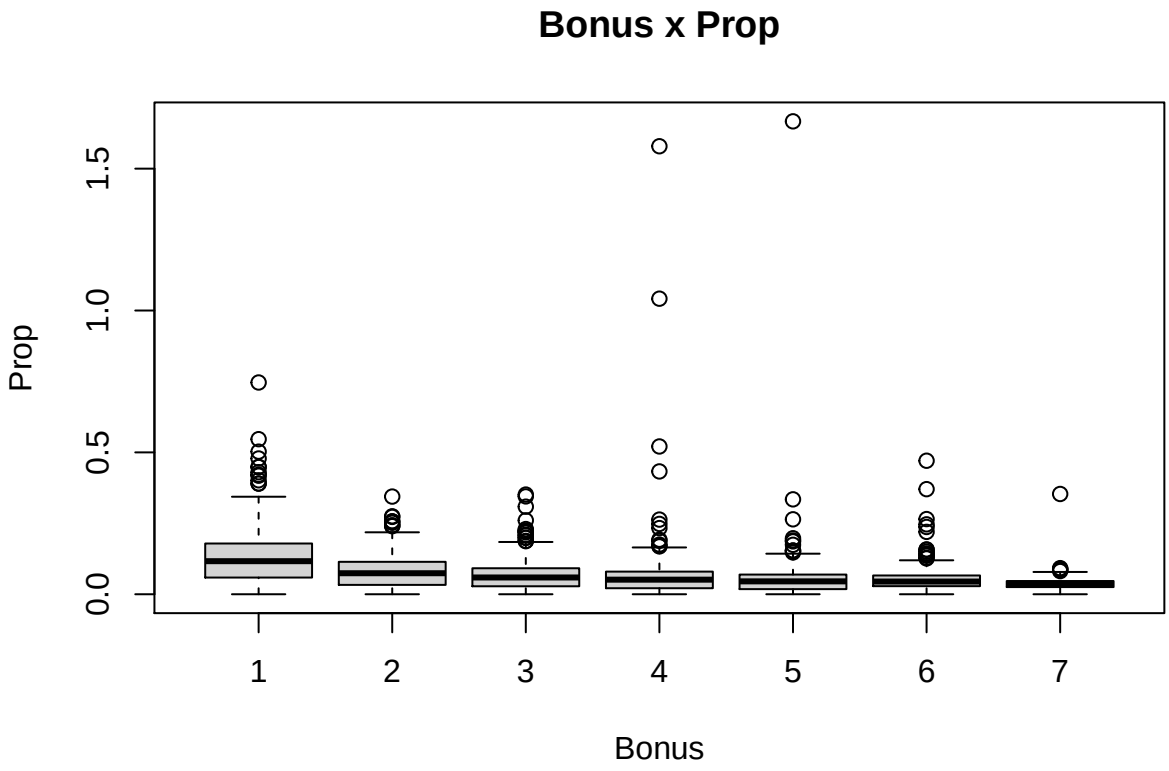

```r
plot(Prop ~ Kilometres, data=dados, main = "Prop x Kilometres")
```



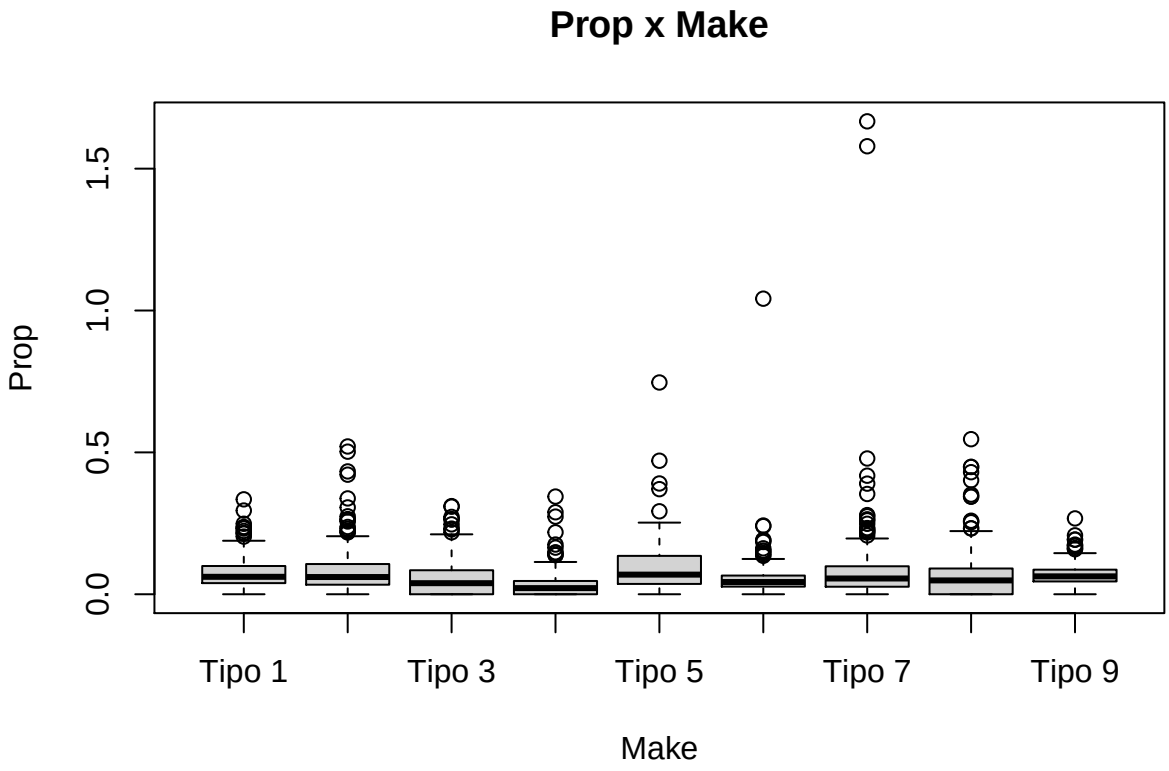

```
plot(Prop ~ Zone, data=dados, main = "Prop x Zone")
```



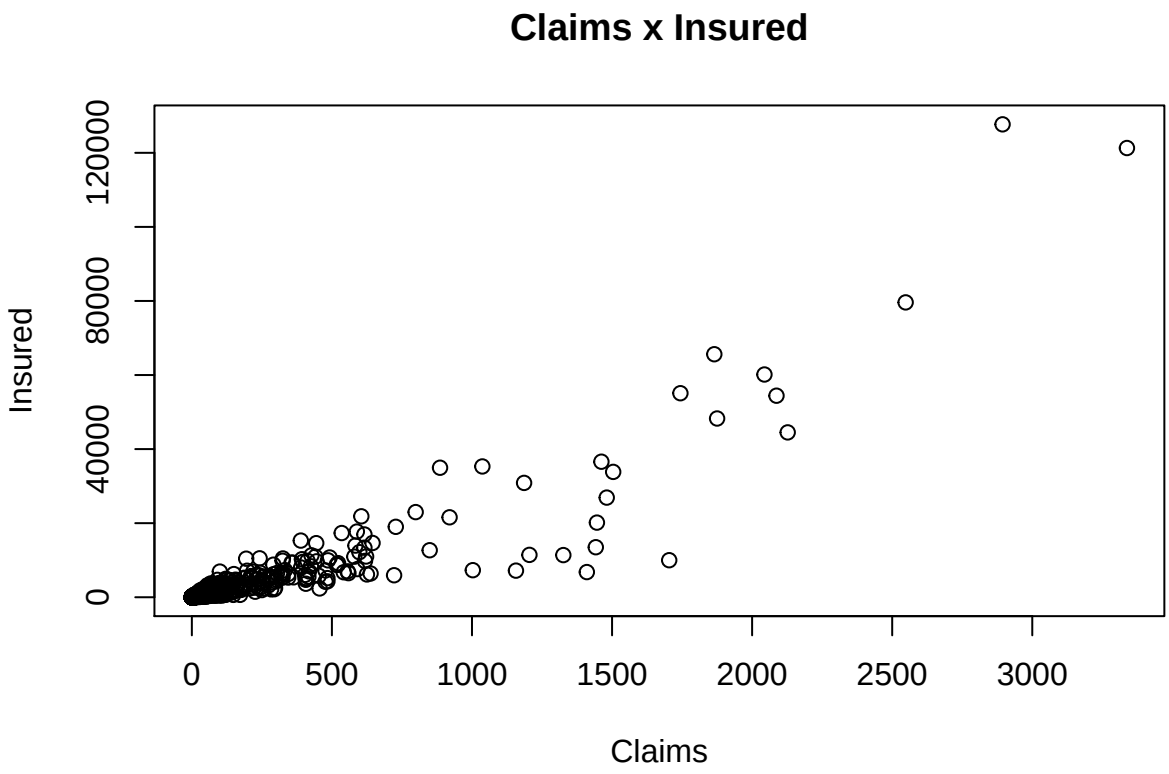

```
plot(Prop ~ Bonus, data=dados, main = "Bonus x Prop")
```




```
plot(Prop ~ Make, data=dados, main = "Prop x Make")
```




```
plot(Insured ~ Claims, data=dados, main = "Claims x Insured")
```



Regressão Linear Simples

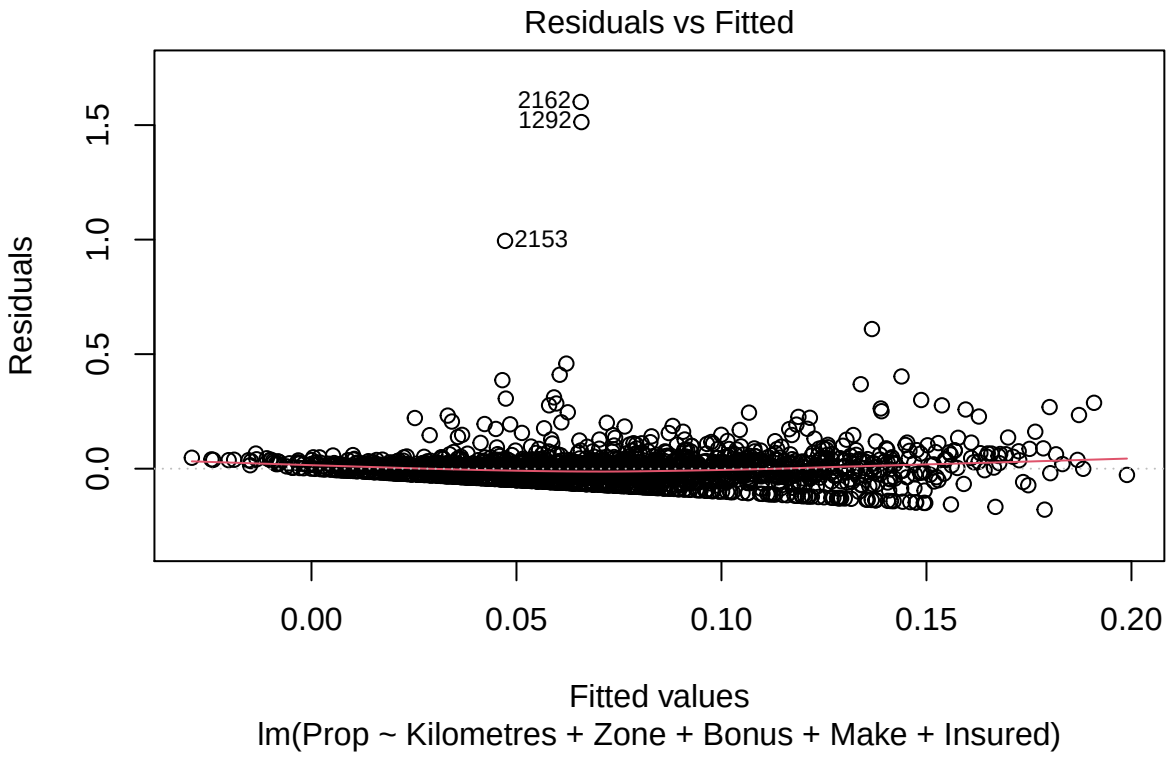
Ao analisarmos o modelo da Regressão Linear Simples, apesra do R2 apresentado pelo modelo ser de 87.32%, tanto o gráfico de Residuals vs Fitted quanto o de Normal Q-Q demonstram que o modelo não compreende bem os dados, existindo diversos pontos fora da curva desejada. Por questão de espaço, os gráficos foram retirados e deixados em forma de comentário. Se tomarmos como verdadeiro o resultado da regressão, temos a seguinte combinação de clientes para que o número de reclamos seja maior que os demais: -> Clientes que a variável Kilometres está entre 1000 ~ 15000 km; -> Clientes da Zona de Estolcomo; -> Clientes na primeira faixa de bônus; -> Clientes com veiculo do tipo 9; -> Clientes presentes nas áreas com maior quantidade de clientes assegurados.

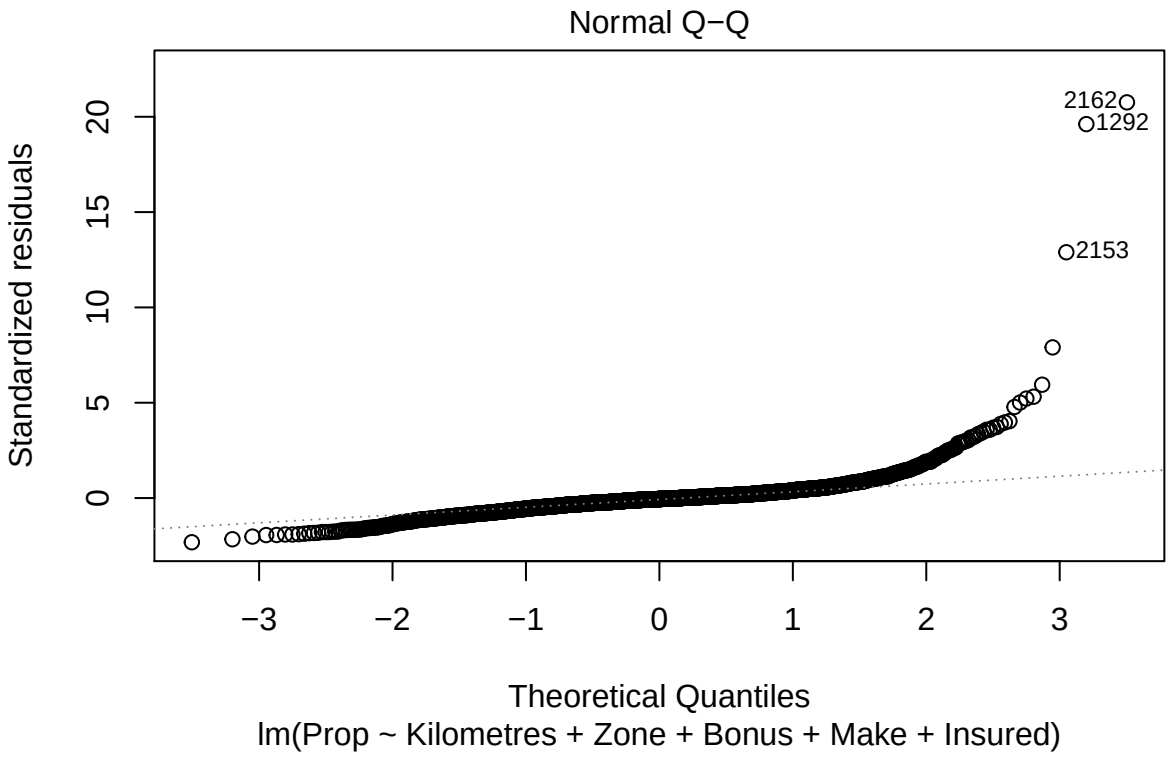
```
modeloRLN <- lm(Prop ~ Kilometres + Zone + Bonus + Make + Insured, data = dados)
summary(modeloRLN)
```

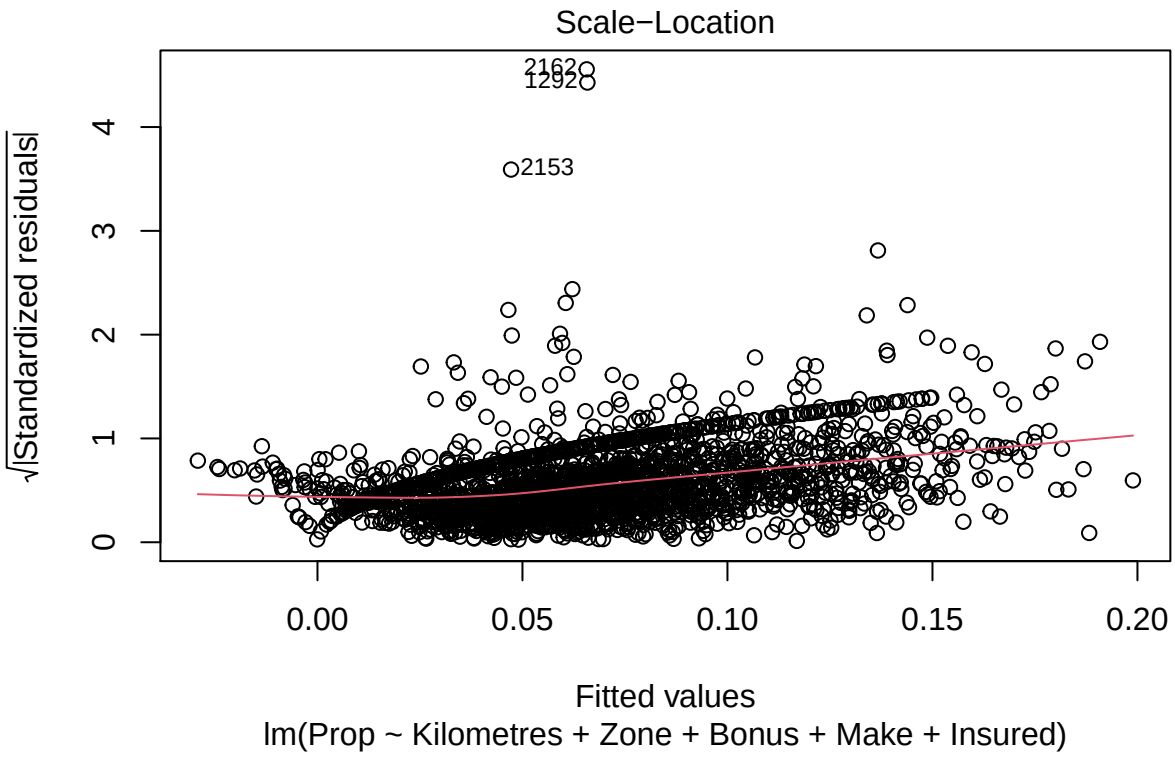
```
##
## Call:
## lm(formula = Prop ~ Kilometres + Zone + Bonus + Make + Insured,
##     data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17882 -0.02709 -0.00391  0.01524  1.60100
##
## Coefficients:
```

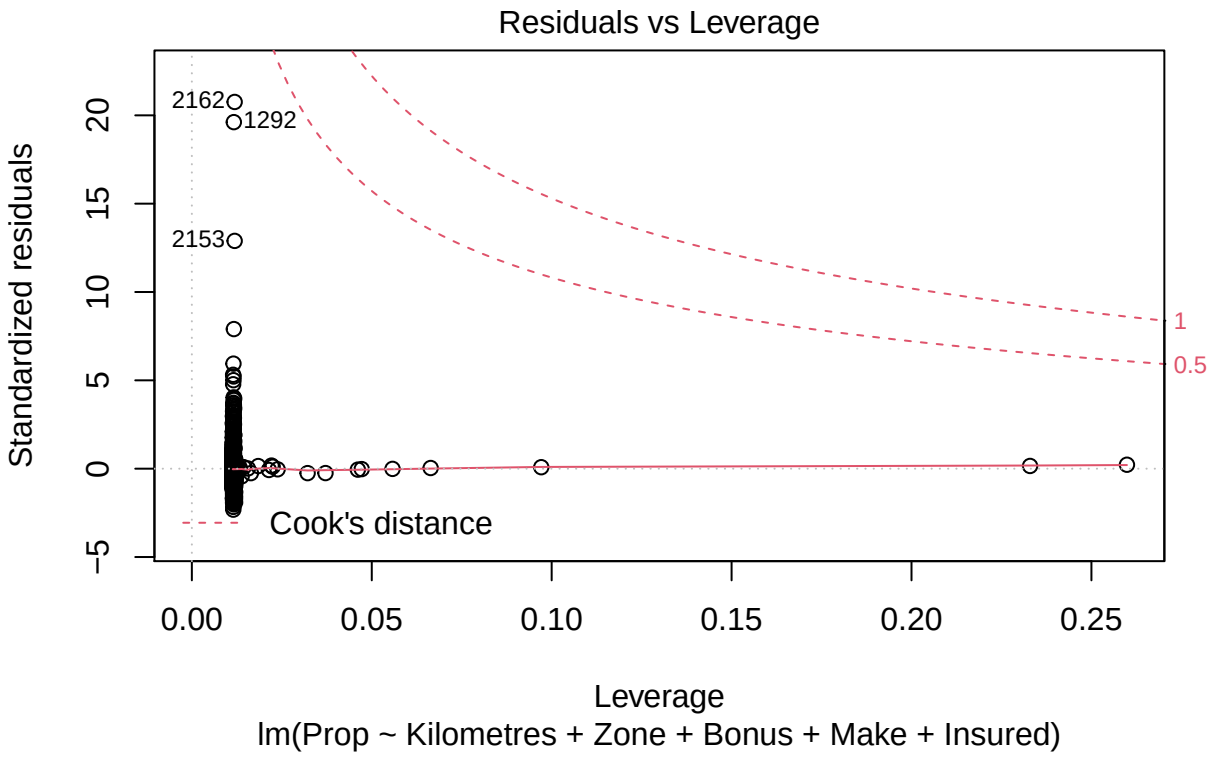
```
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1.569e-01  8.296e-03  18.914  < 2e-16 ***
## Kilometres1000 ~ 15000         8.975e-03  5.232e-03   1.716 0.086395 .
## Kilometres15000 ~ 20000        1.560e-02  5.238e-03   2.979 0.002923 **
## Kilometres20000 ~ 25000        1.421e-02  5.276e-03   2.694 0.007118 **
## Kilometres> 25000              2.627e-02  5.304e-03   4.953 7.89e-07 ***
## ZoneLarge cities              -2.406e-02  6.183e-03  -3.891 0.000103 ***
## ZonePequenas cidades do sul   -3.132e-02  6.184e-03  -5.065 4.43e-07 ***
## ZoneArea rural do sul         -4.596e-02  6.208e-03  -7.404 1.88e-13 ***
## ZonePequenas cidades do norte -4.117e-02  6.197e-03  -6.644 3.85e-11 ***
## ZoneArea rural do norte       -4.667e-02  6.183e-03  -7.547 6.53e-14 ***
## ZoneGotland                   -5.153e-02  6.308e-03  -8.169 5.22e-16 ***
## Bonus2                        -4.955e-02  6.238e-03  -7.942 3.17e-15 ***
## Bonus3                        -6.126e-02  6.248e-03  -9.805  < 2e-16 ***
## Bonus4                        -6.284e-02  6.248e-03 -10.057  < 2e-16 ***
## Bonus5                        -7.368e-02  6.234e-03 -11.819  < 2e-16 ***
## Bonus6                        -7.616e-02  6.225e-03 -12.233  < 2e-16 ***
## Bonus7                        -9.195e-02  6.381e-03 -14.410  < 2e-16 ***
## MakeTipo 2                     4.040e-03  7.016e-03   0.576 0.564801
## MakeTipo 3                    -1.821e-02  7.038e-03  -2.587 0.009746 **
## MakeTipo 4                    -4.259e-02  7.068e-03  -6.026 1.98e-09 ***
## MakeTipo 5                     1.574e-02  7.023e-03   2.241 0.025144 *
## MakeTipo 6                    -2.158e-02  7.020e-03  -3.074 0.002138 **
## MakeTipo 7                     7.709e-03  7.038e-03   1.095 0.273476
## MakeTipo 8                    -3.130e-03  7.079e-03  -0.442 0.658386
## MakeTipo 9                    -4.624e-03  7.301e-03  -0.633 0.526643
## Insured                       -5.144e-08  3.372e-07  -0.153 0.878772
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07759 on 2156 degrees of freedom
## Multiple R-squared:  0.1878, Adjusted R-squared:  0.1784
## F-statistic: 19.94 on 25 and 2156 DF,  p-value: < 2.2e-16
```

```r
plot(modeloRLN)
```


Residuals vs Fitted
2162
1292
2153
Residuals
0.0
0.5
1.0
1.5
0.00
0.05
0.10
0.15
0.20
Fitted values
lm(Prop ~ Kilometres + Zone + Bonus + Make + Insured)


Normal Q−Q
Standardized residuals
0
5
10
15
20
2162
1292
2153
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Prop ~ Kilometres + Zone + Bonus + Make + Insured)


Scale−Location
2162
1292
2153
√|Standardized residuals|
0
1
2
3
4
0.00
0.05
0.10
0.15
0.20
Fitted values
lm(Prop ~ Kilometres + Zone + Bonus + Make + Insured)


Residuals vs Leverage
Standardized residuals
Leverage
lm(Prop ~ Kilometres + Zone + Bonus + Make + Insured)
2162
1292
2153
Cook's distance
1
0.5
−5
0
5
10
15
20
0.00
0.05
0.10
0.15
0.20
0.25