Relatório Final

Análise de Regressao Linear

Davi Wentrick Feijó - 20016806, Ana Beatriz de Carvalho Lopes - 200034812, Leonardo dos Reis Andrade -211029030

July 25, 2023

1 Introdução

Neste estudo, conduziremos uma análise detalhada de um conjunto de dados proveniente de uma cidade americana. O conjunto de dados é composto por informações de venda de 522 casas que foram vendidas no último ano. O propósito desse estudo é a construção de um modelo preditivo que busca prever o preço de venda de residências em função de características da casa e sua vizinhança.

A estrutura deste trabalho é organizada em seções distintas para facilitar a compreensão dos resultados obtidos. Na segunda seção, conduziremos uma análise descritiva abrangente das variáveis presentes no conjunto de dados. Nessa etapa, exploraremos as características individuais, identificando tendências e padrões que possam influenciar o valor de venda das casas. Serão realizadas estatísticas descritivas, como média, mediana, valor mínimo dos dados, valor máximo dos dados, quartis, e gráficos para visualização dos dados.

Na terceira seção, abordaremos a seleção criteriosa das variáveis que serão adotadas no modelo preditivo. Utilizaremos técnicas de Regressão Linear, para identificar quais características têm maior relevância. Essa etapa é crucial para garantir a eficiência e a precisão do modelo final.

A seção quatro é dedicada ao desenvolvimento do modelo preditivo propriamente dito, onde utilizaremos regressão para criar um modelo que seja capaz de generalizar bem os dados e fazer previsões precisas.

Por fim, na seção cinco, apresentaremos os resultados obtidos com o modelo preditivo.

Com essas etapas claramente definidas e uma abordagem metodológica sólida, esperamos que este estudo contribua significativamente para o entendimento do mercado imobiliário dessa cidade americana.

2 Objetivos

Para esse trabalho, usaremos Regressivo Linear Múltipla no banco de dados que apresenta informações observacionais sobre onze características diferentes de cada casa. Para podermos ter controle sobre o modelo, dividimos a train em 2 partes, a primeira, composta por 300 trains selecionadas aleatoriamente, para a construção do modelo, e a segunda com as 222 restantes visando a validação do modelo preditivo.

Como forma de mensurar a seleção de variáveis do modelo usaremos na seleção de variáveis os critérios R^2, Cp de Mallows e regressão “Stepwise”. As variáveis que constam no banco de dados são o preço de venda, tamanho da casa, número de quartos, número de banheiros, presença de ar-condicionado, tamanho da garagem, presença de piscina, idade de casa, obtida a partir do ano de construção, qualidade da construção, tamanho do terreno e proximidade da “Highway”(proximidade da rodovia). Essas variáveis representam elementos que podem influenciar o preço de venda das casas.

3 Metodologia

Coleta de dados: Os dados utilizados neste estudo foram obtidos a partir de um conjunto de informações sobre a venda de 522 casas em uma cidade americana no último ano. As variáveis incluídas no banco de dados são: preço de venda, tamanho da casa, número de quartos, número de banheiros, presença de ar-condicionado, tamanho da garagem, presença de piscina, idade da casa (obtida a partir do ano de construção), qualidade da construção, tamanho do terreno e proximidade da “Highway” (proximidade da rodovia).

Análise descritiva das variáveis: Foi realizada uma análise descritiva das variáveis presentes no banco de dados. Foram apresentados histogramas e gráficos de dispersão para as variáveis, permitindo uma melhor compreensão da distribuição e da relação entre as variáveis.

Seleção de variáveis: Para essa seleção, foram utilizados critérios como o coeficiente de determinação (R^2), o critério Cp de Mallows e o método de regressão “Stepwise”. Com base nessas análises, foram escolhidas as variáveis mais relevantes para a construção do modelo preditivo.

Modelo e validação: O modelo de regressão linear múltipla foi desenvolvido usando as variáveis selecionadas na etapa anterior. A train foi dividida em duas partes: uma continha 300 trains selecionadas aleatoriamente para construir o modelo e a outra continha as 222 trains restantes para validar o modelo.

Resultados: Os resultados obtidos com o modelo. Foram apresentados os coeficientes estimados para cada variável e a qualidade de ajuste do modelo, medido pelo coeficiente de determinação (R^2) e pelo F-statistic. Também foram analisados os resíduos do modelo para verificar a adequação das premissas da regressão linear.

Transformação dos dados: Durante a modelagem, foi realizada uma transformação nos dados, sendo o valor do preço de venda convertido em escala logarítmica, para melhorar o cumprimento das premissas do modelo e obter resultados mais precisos.

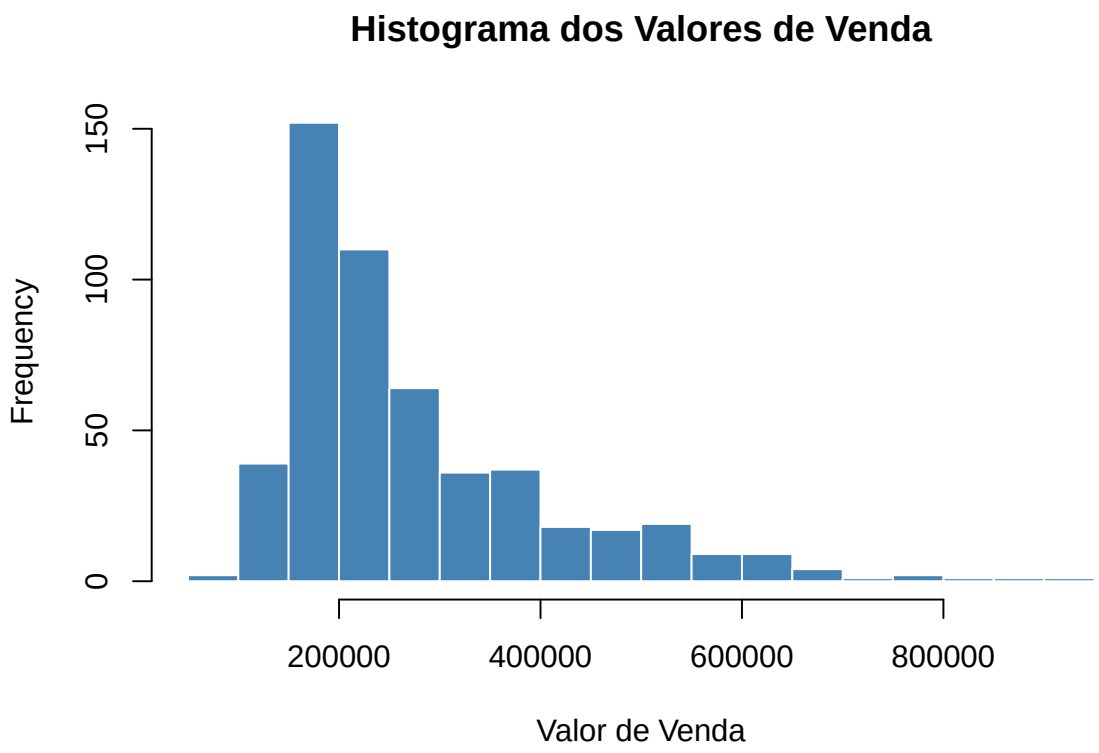
Redução do Modelo: Com base em análises adicionais, o modelo foi reduzido, considerando apenas as variáveis mais importantes e relevantes para prever o preço de venda das casas.

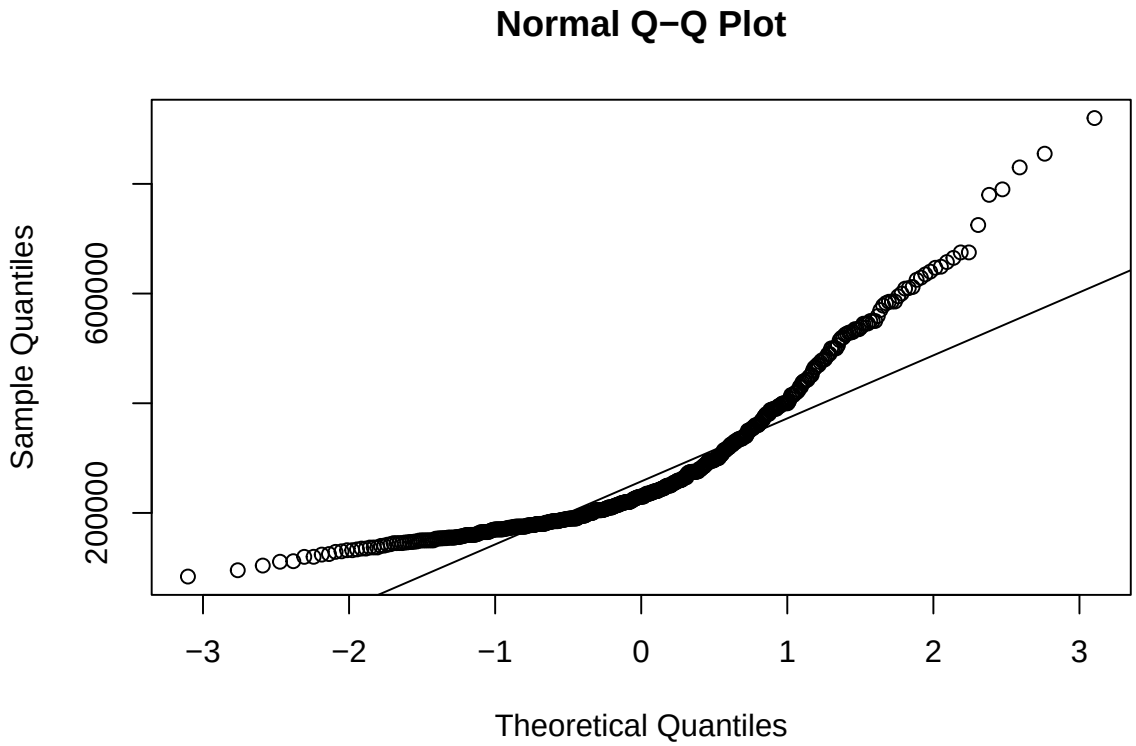
Análise e discussão: Os resultados do modelo foram analisados em relação aos objetivos propostos, avaliando a relevância das variáveis selecionadas e a eficácia do modelo preditivo.

4 Resultados

4.1 Analise descritiva das variáveis

4.1.1 Valor de Venda - Variavel Resposta - X1



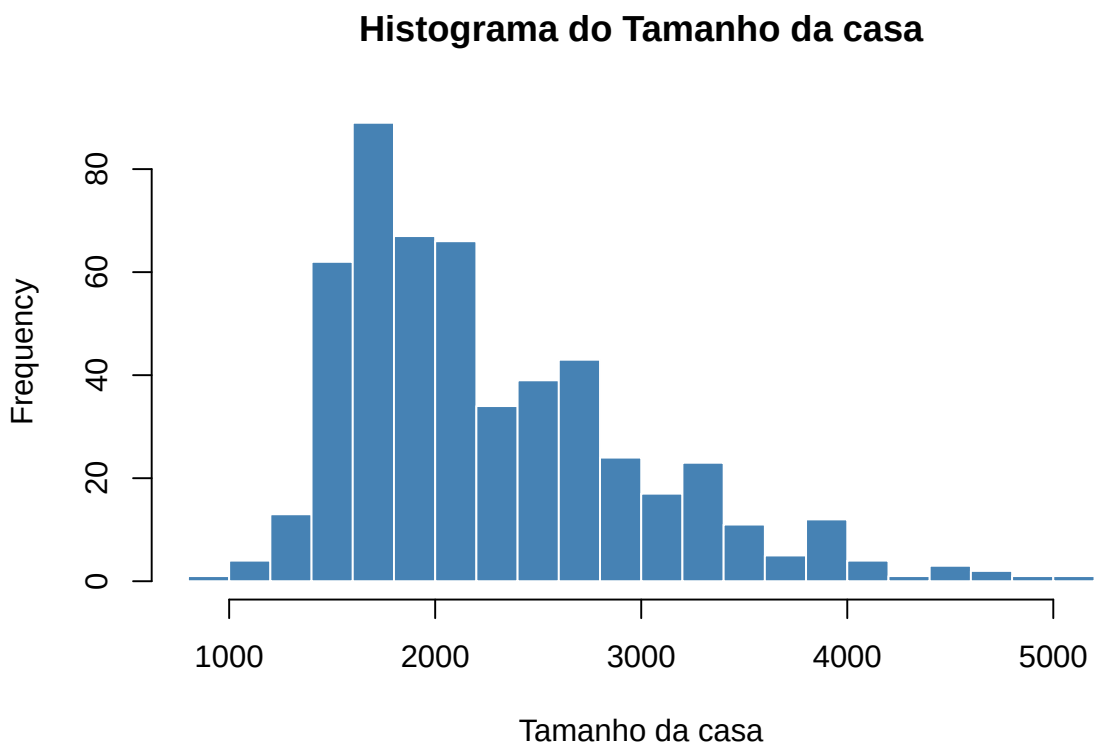


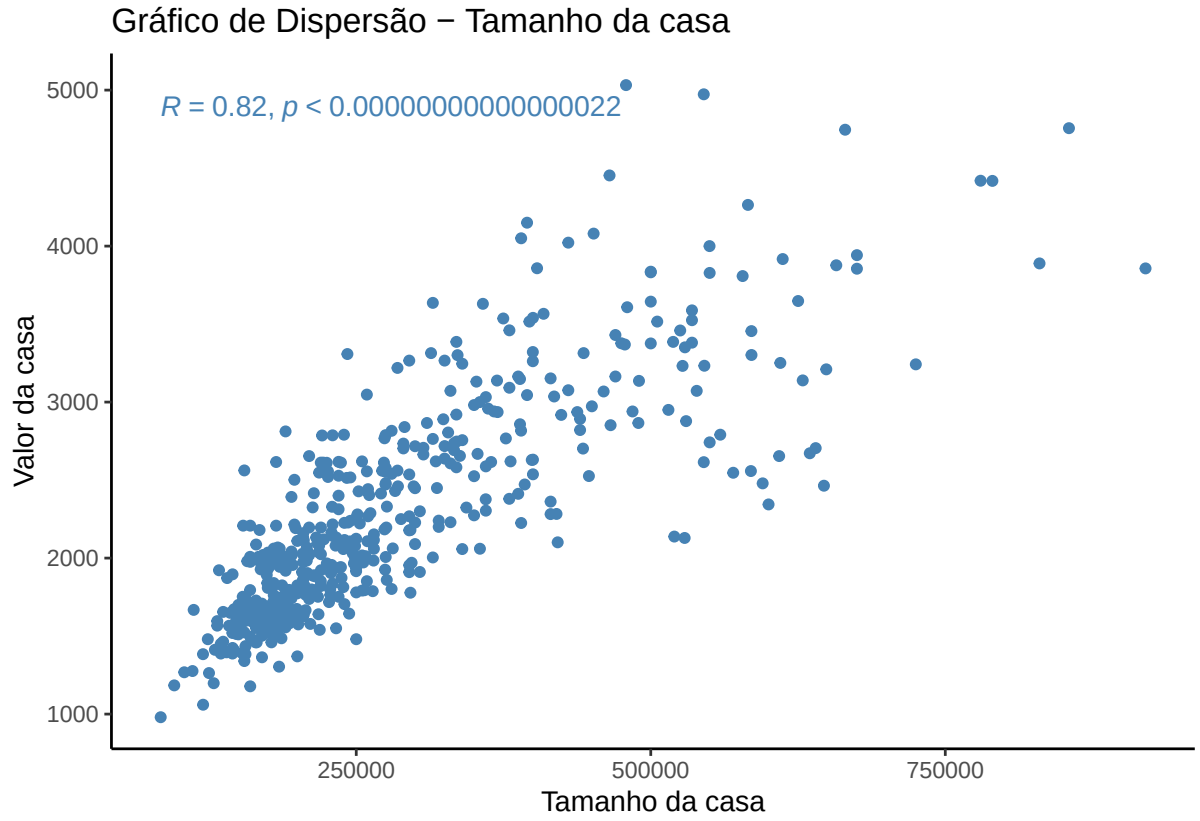
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   84000  180000  229900  277894  335000  920000
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dados$X1
## W = 0.84372, p-value < 0.00000000000000022
```

Das onze variáveis apresentadas no banco de dados, a variável “preço de venda” é nossa variável resposta, sendo ela uma variável quantitativa e continua. Esses da medida resumo nos mostra que a maior parte dos preços estão concentrados e R$ 180.000 até R$ 230.000, como indicado pela mediana., conforme podemos observar no gráfico 1. Ao realizar o teste de normalidade de Shapiro-Wilk na variável “preço de venda”, o resultado indica que ela não segue uma distribuição normal, pois o p-valor associado ao teste foi menor que 2.2e-16 (um valor extremamente pequeno). No entanto, a distribuição da variável parece possuir uma calda alongada, provavelmente devido à presença de um valor atípico (outlier) com o valor de R$ 920.000 Essa informação sugere que a distribuição dos preços de venda é assimétrica e é afetada pela presença de preços muito altos (outliers). Portanto, é importante levar em consideração essa característica ao analisar e modelar essa variável, para evitar que ela distorça as análises estatísticas e as conclusões feitas a partir dos dados.

4.1.2 Tamanho da Casa - X2





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     980    1701    2061    2261    2636    5032
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dados$X2
## W = 0.91518, p-value < 0.00000000000000022
```

A variável “tamanho da casa”, também é uma variável quantitativa e contínua, medida em metros quadrados. A maior parte dos tamanhos das casas concentra-se na faixa em torno de 2.000 pés quadrados, conforme evidenciado pela mediana e o intervalo entre o primeiro quartil e o terceiro quartil, e o histograma de tamanho das casas mostra a distribuição dos valores, sendo possível perceber uma concentração de dados em torno da faixa de 2.000 pés quadrados. Além disso, foi realizado o teste de normalidade de Shapiro-Wilk na variável “tamanho da casa”. O resultado indica que a distribuição não segue uma distribuição normal, com p-valor muito pequeno (p < 2,2e-16). Isso significa que a distribuição dos tamanhos das casas é assimétrica e não pode ser considerada normal. A partir do gráfico de dispersão entre o tamanho da casa e o valor da casa, podemos analisar a relação entre essas duas variáveis, e ver que temos uma correlação positiva moderada

4.1.3 Número de Quartos - X3

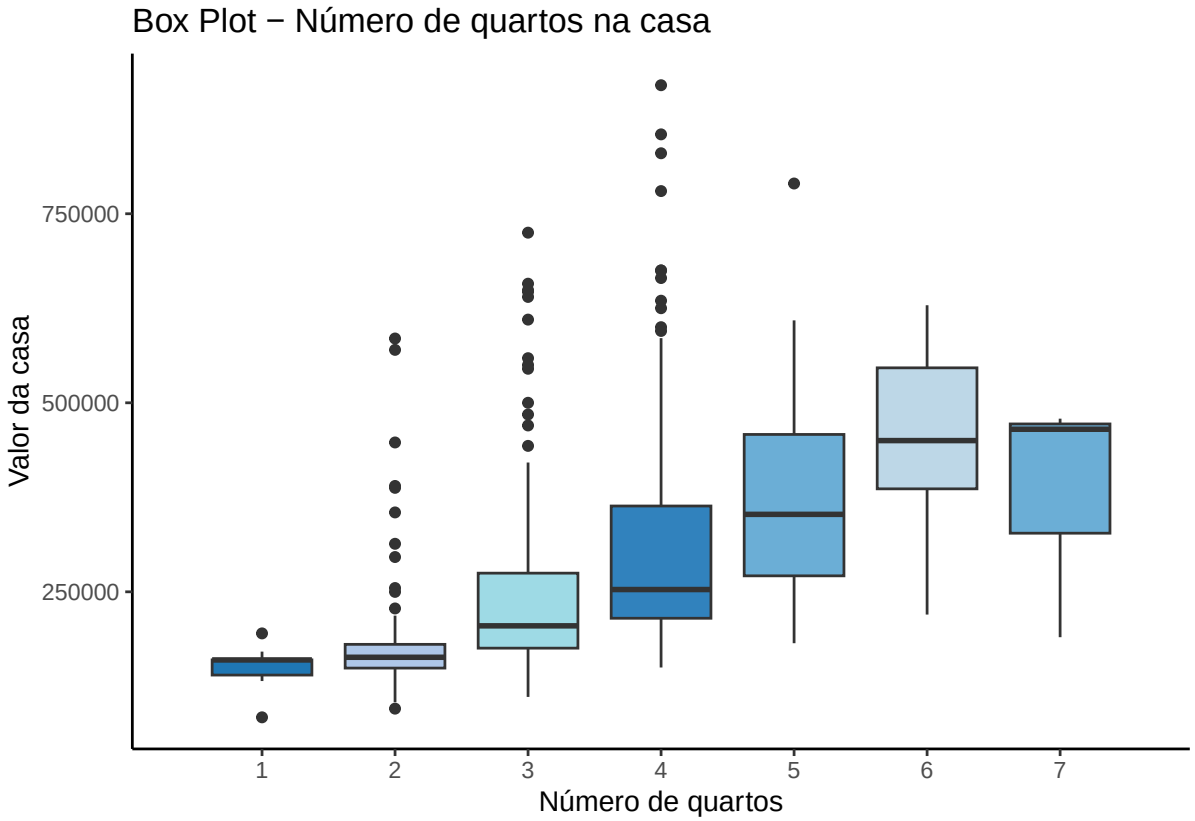


Table 1: Tabela de frequencia do Número de quartos

	Número de quartos	Frequencia	Porcentual
1	0.00	1	0.19
2	1.00	9	1.72
3	2.00	64	12.26
4	3.00	202	38.70
5	4.00	179	34.29
6	5.00	52	9.96
7	6.00	12	2.30
8	7.00	3	0.57

```
##
##  Shapiro-Wilk normality test
##
## data:  dados$X3
## W = 0.91065, p-value < 0.00000000000000022
```

Nossa terceira variável refere-se ao número de cômodos que as casas estudadas possuem e é uma variável quantitativa discreta. Acredita-se que casas com maior número de cômodos sejam mais valorizadas. A categoria com menor número de casas é a de 7 quartos, que possui apenas 3 casas, e a categoria com maior número de casas é a de 3 quartos, que possui 202 casas. Além disso, foi realizada uma análise do box plot gráfico para a variável “número de quartos” em

relação ao “valor da casa”. O box plot mostra a distribuição e dispersão dos dados em relação às diferentes categorias de quartis. É possível visualizar a mediana, os quartis e os possíveis valores atípicos. Por mais que o teste de normalidade de Shapiro-Wilk seja aplicado à variável “número de trimestres”, o resultado indica que a distribuição não segue uma distribuição normal, pois o p-valor associado ao teste é menor que 2,2e-16. Isso significa que a distribuição do número de trimestres não é uma distribuição normal. Em suma, uma análise revela uma distribuição dos quartos nas casas estudadas e destaca que a maioria das casas tem entre 2 e 4 quartos.

4.1.4 Número de Banheiros - X4

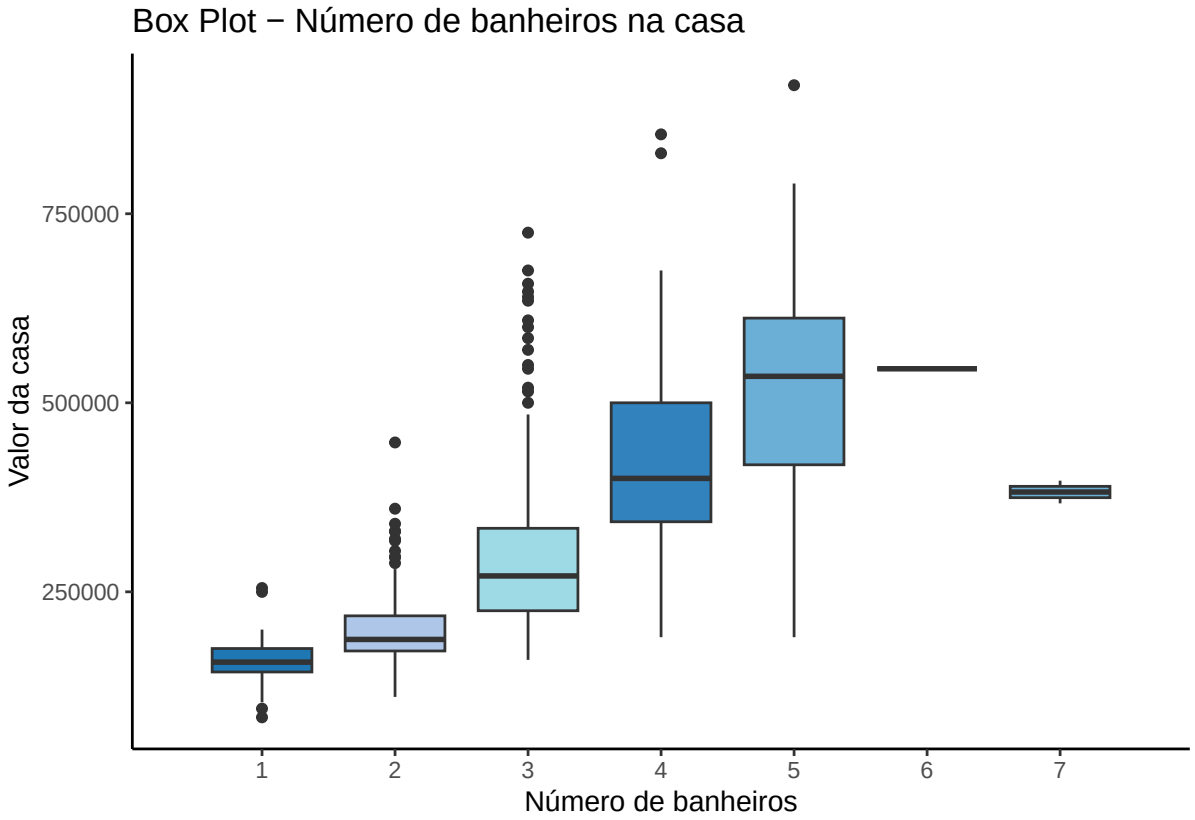


Table 2: Tabela de frequencia do Número de banheiros

	Número de banheiros	Frequencia	Porcentual
1	0.00	1	0.19
2	1.00	71	13.60
3	2.00	171	32.76
4	3.00	175	33.52
5	4.00	84	16.09
6	5.00	17	3.26
7	6.00	1	0.19
8	7.00	2	0.38

```
##
##  Shapiro-Wilk normality test
##
## data:  dados$X4
## W = 0.91053, p-value < 0.00000000000000022
```

A variável “número de banheiros” representa o número de banheiros que as casas estudadas possuem e é uma variável quantitativa discreta. Podemos observar pela tabela que a categoria com menor número de casas é a que possui 7 banheiros, e o maior número de casas está na categoria com 3 banheiros. Além disso, foi realizada uma análise do box plot gráfico, para a variável “número de banheiros” em relação ao “valor da casa”. O box plot mostra a distribuição e dispersão dos

dados em relação às diferentes categorias de banheiros. É possível visualizar a mediana, os quartis e os possíveis valores atípicos. Quanto ao teste de normalidade de Shapiro-Wilk aplicado à variável “número de banheiros”, o resultado indica que a distribuição não segue uma distribuição normal, pois o p-valor associado ao teste foi menor que 2,2e-16. Isso significa que a distribuição do número de banheiros não é uma distribuição normal. Em resumo, a análise descritiva da variável revela uma distribuição dos banheiros nos domicílios estudados e destaca que a maioria dos domicílios possui entre 2 e 4 banheiros.

4.1.5 Presença de Ar-Condicionado - X5

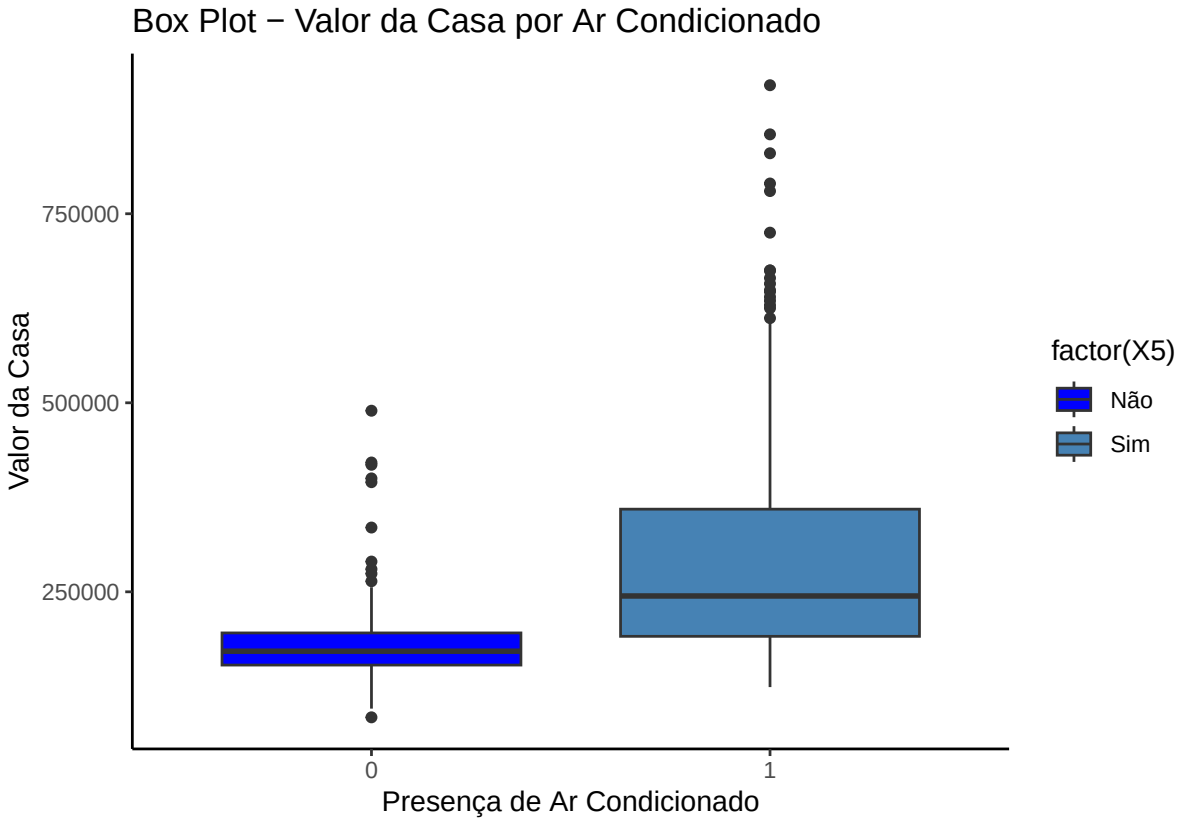


Table 3: Tabela de frequencia do Ar Condicionado

	Ar Condicionado	Frequencia	Porcentual
1	0	88	16.86
2	1	434	83.14

A variável “Ar-Condicionado” é uma variável qualitativa, representando a presença ou ausência de ar-condicionado nas residências estudadas. É codificado com 1 para indicar que tem ar-condicionado e 0 para indicar que não tem ar-condicionado. A maioria das residências (83,14%) possui ar-condicionado, enquanto uma proporção menor (16,86%) não possui ar-condicionado. Uma variável qualitativa não é normal porque não possui uma escala numérica contínua e, portanto, não pode ser representada por valores passíveis de cálculos matemáticos, como a média ou o desvio padrão.

4.1.6 Tamanho da Garagem - X6

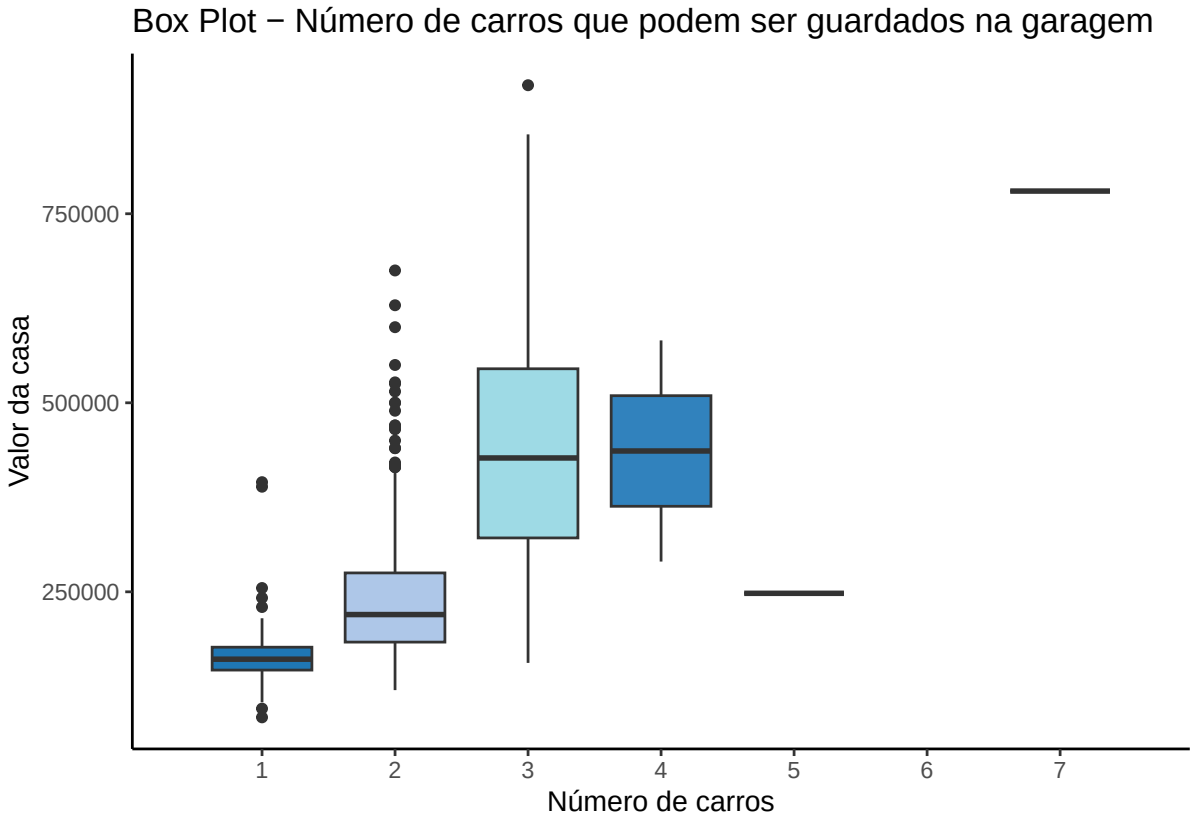


Table 4: Tabela de frequencia do Tamanho da Garagem

	Número de Garagens	Frequencia	Porcentual
1	0.00	7	1.34
2	1.00	52	9.96
3	2.00	353	67.62
4	3.00	106	20.31
5	4.00	2	0.38
6	5.00	1	0.19
7	7.00	1	0.19

```
##
##  Shapiro-Wilk normality test
##
## data:  dados$X6
## W = 0.73498, p-value < 0.00000000000000022
```

A variável “tamanho da garagem” representa a capacidade de carros que podem ser guardados na garagem da casa. É uma variável quantitativa discreta, pois é composta por números inteiros que representam a quantidade de carros. As casas com garagem têm valor 0 atribuído. Além disso, foi realizada uma análise do box plot gráfico para esta variável em relação ao “valor da casa”. O box plot mostra a distribuição e dispersão dos dados em relação às diferentes categorias de tamanhos de garagem. É possível verificar que a maioria das casas tem capacidade para garagem

para 2 carros, sete casas não têm garagem, enquanto apenas 3 casas têm capacidade para 4 ou mais carros. Quanto ao teste de normalidade de Shapiro-Wilk aplicado à variável “tamanho da garagem”, o resultado indica que a distribuição não segue uma distribuição normal, pois o p-valor associado ao teste é menor que 2,2e-16. Isso significa que a distribuição de tamanhos de garagem não é uma distribuição normal.

4.1.7 Piscina - X7

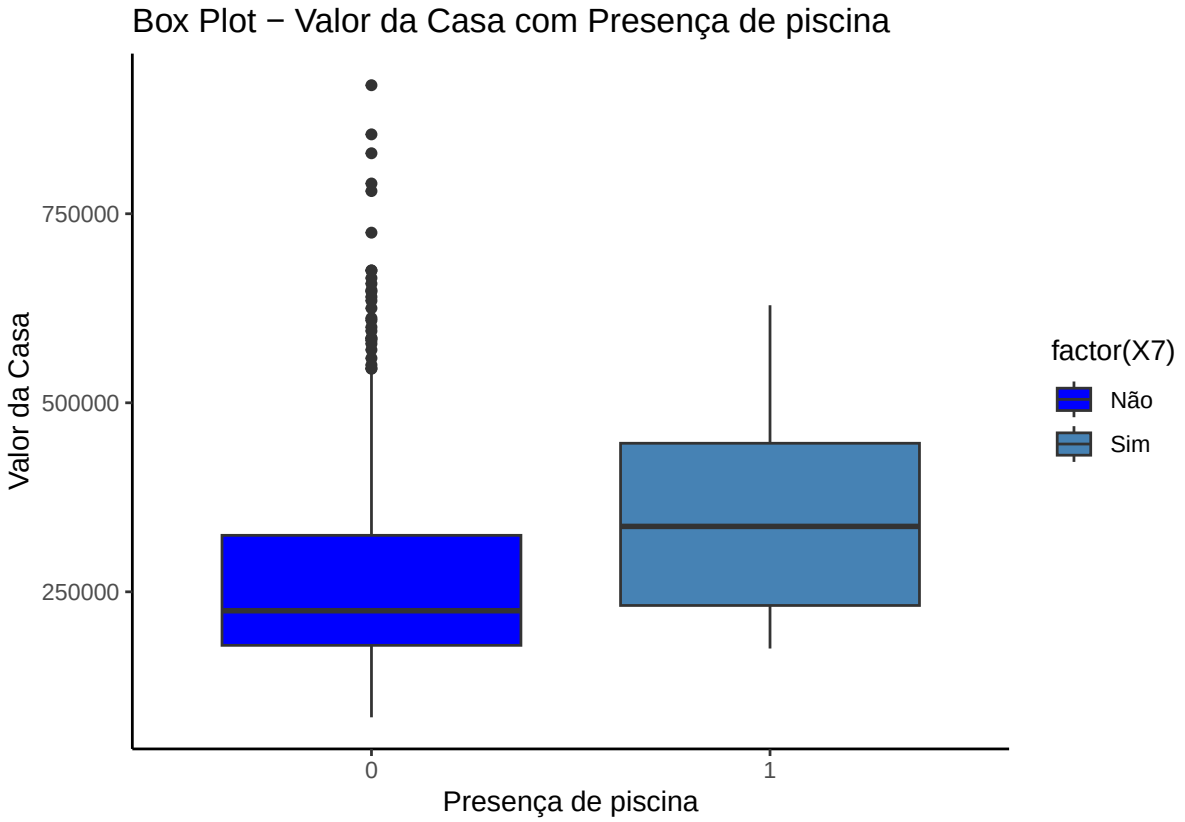
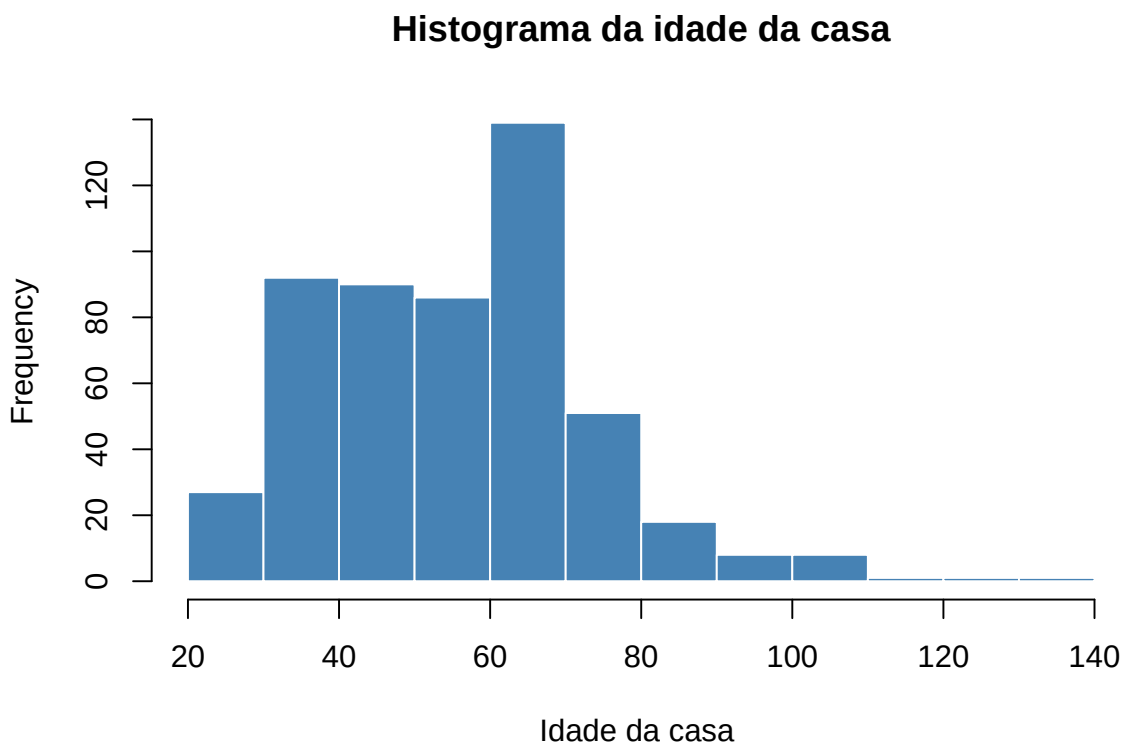


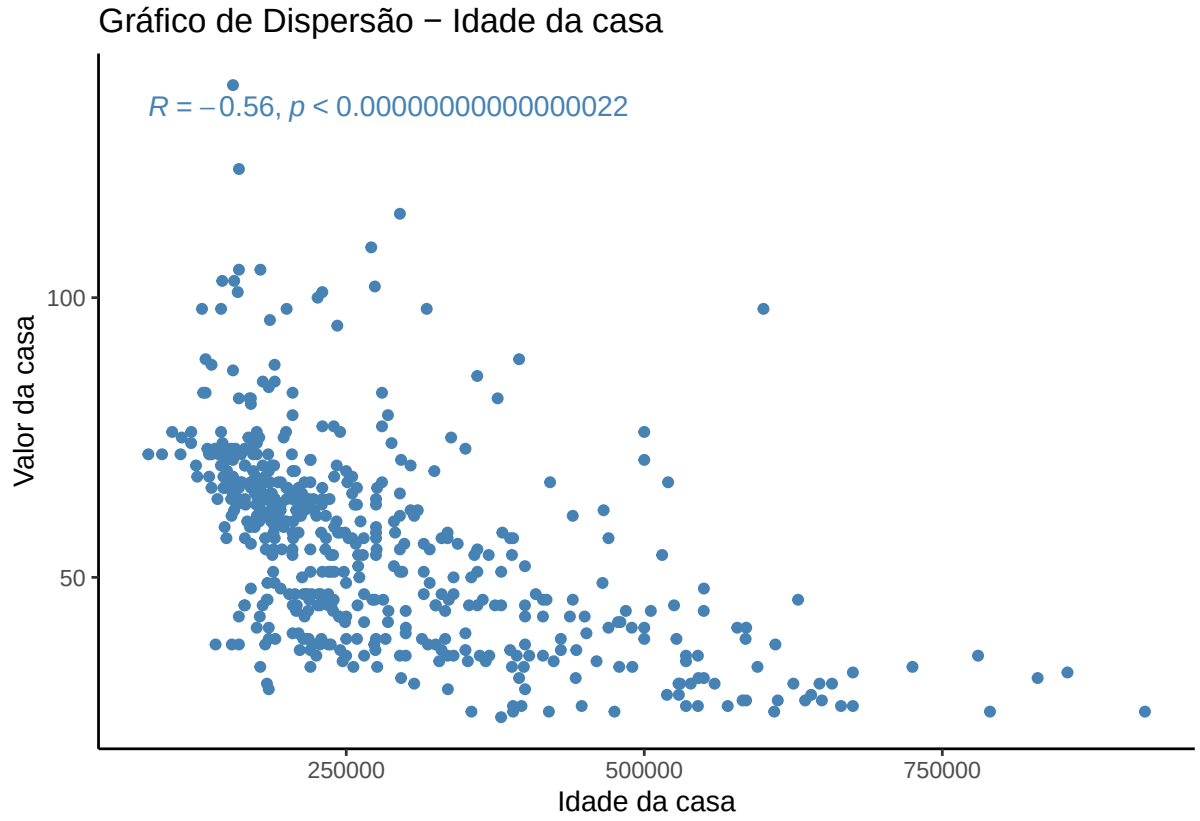
Table 5: Tabela de frequencia da Presença de piscina

	Presença de piscina	Frequencia	Porcentual
1	0	486	93.10
2	1	36	6.90

A variável “presença de piscina” é uma variável qualitativa, indicando se o domicílio possui piscina ou não. É codificado com 1 para indicar que a casa tem piscina e 0 para indicar que não tem piscina. Uma variável qualitativa não é normal porque não possui uma escala numérica contínua e, portanto, não pode ser representada por valores passíveis de cálculos matemáticos, como a média ou o desvio padrão. No box plot apresentado, é possível visualizar a distribuição dos valores das casas com base na presença ou ausência de piscina. E mostra-nos que a maioria dos agregados familiares (93,10%) não tem piscina, e apenas um pequeno número de agregados familiares (6,90%) tem piscina.

4.1.8 Idade do Imóvel - X8





```
##
##  Shapiro-Wilk normality test
##
## data:  dados$idade
## W = 0.96141, p-value = 0.0000000001833
```

A variável “idade da casa” refere-se ao ano em que foi construída, o ano de referência “2023” foi utilizado para calcular a idade de cada casa antes de iniciar a análise e só depois trabalhar com ela. Essa variável é quantitativa discreta, pois contém valores numéricos inteiros que representam a idade do domicílio. O histograma mostrado mostra a distribuição dos valores de idade do agregado familiar em diferentes categorias de idade. A frequência de domicílios de 40 a 60 anos é maior, observa-se também que o número de domicílios diminui com o aumento da idade, chegando a 120 domicílios com cerca de 100 anos. Além disso, gráficos de dispersão entre as variáveis “idade da casa” e “valor da casa” foram analisados. O enredo disperso mostra que você é uma pessoa fraca e negativa. Quanto ao teste de normalidade de Shapiro-Wilk aplicado à variável “idade do domicílio”, o resultado indica que a distribuição não segue uma distribuição normal. O p-valor associado ao teste foi inferior a 0,05 (p-valor = 1,833e-10), o que significa que a distribuição dos dados é significativamente diferente da normal.

4.1.9 Qualidade de Construção - X9

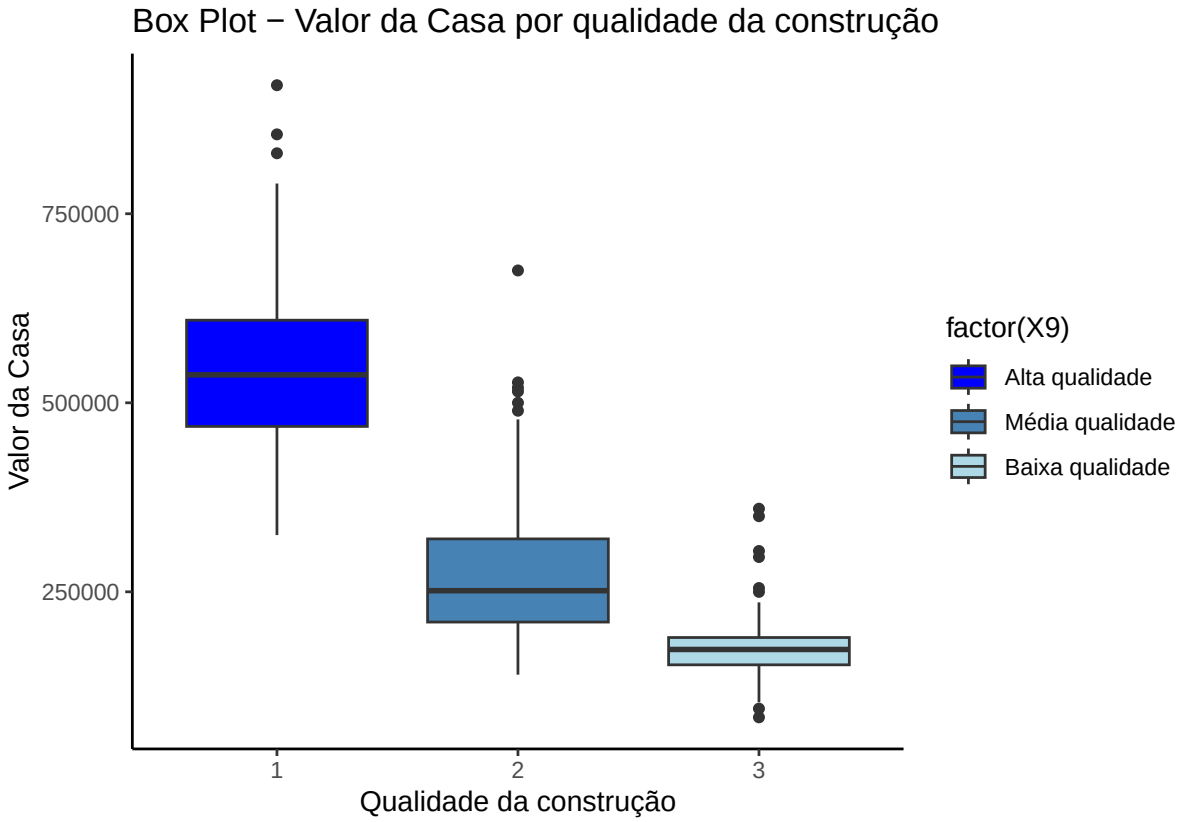
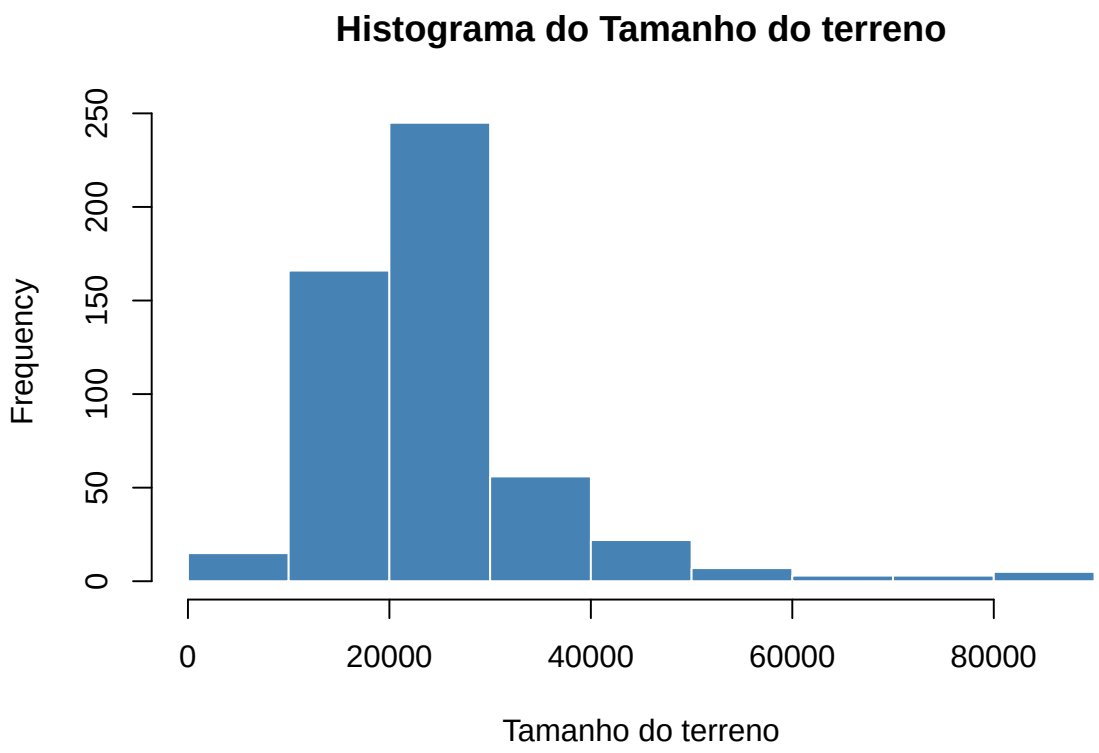


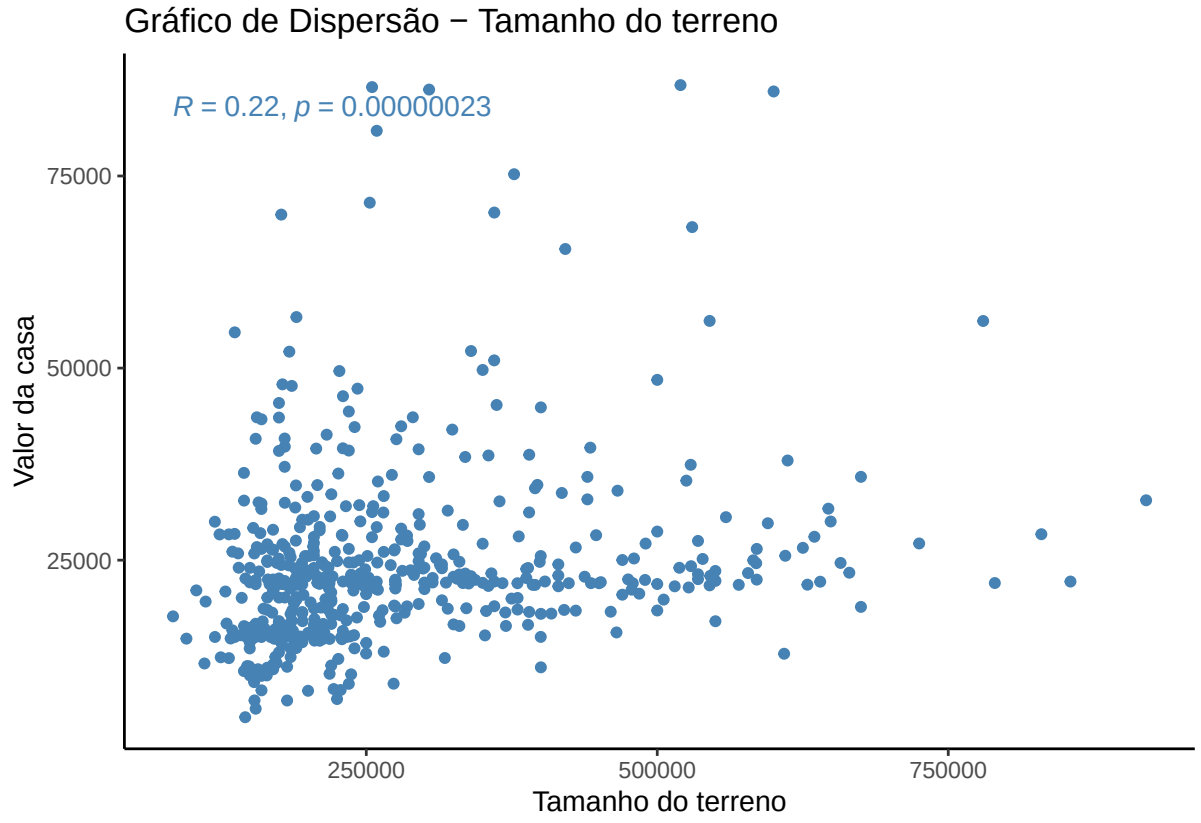
Table 6: Tabela de frequencia da qualidade da construção

	qualidade da construção	Frequencia	Porcentual
1	1	68	13.03
2	2	290	55.56
3	3	164	31.42

A Qualidade da construção foi selecionada sendo uma variável qualitativa, que tenta mensurar a qualidade do material empregado na construção, essa variável conta com bastante subjetividade por parte do avaliador, sendo pontuada com 1 para o alta qualidade, 2 para média qualidade e 3 para imóveis de construção avaliados como baixa qualidade. A maioria das residências (55,56%) foi classificada como “Qualidade de construção média” e uma proporção menor de residências (13,03%) recebeu uma classificação de “Qualidade de construção alta” O box plot apresentado mostra a distribuição dos valores das casas em relação à qualidade da construção. Esta visualização permite verificar que quanto mais casas mais carros estão associados a uma maior qualidade de construção.

4.1.10 Tamanho do Terreno - X10





```
##
##  Shapiro-Wilk normality test
##
## data:  dados$X10
## W = 0.79804, p-value < 0.00000000000000022
```

A variável “tamanho do terreno” representa o tamanho do terreno em metros quadrados. É uma variável quantitativa contínua, pois consiste em valores numéricos que podem variar ao longo de uma escala contínua. Com o histograma podemos ver que os terrenos em torno de 50.000 a 100.000 m² possuem uma frequência de pico. Além disso, foi realizada uma análise do gráfico de dispersão entre esta variável “e o” valor da casa “. O gráfico de dispersão mostra como varia o valor da casa em relação ao tamanho do terreno. É possível observar que existe uma forte correlação positiva. Na medida em que o teste de normalidade de Shapiro-Wilk é aplicado à variável”Tanmanho do Terreno” (X10), o resultado indica que a distribuição não segue uma distribuição normal. O p-valor associado ao teste é inferior a 0,05 (p-valor < 2,2e-16), o que significa que os dados têm uma distribuição significativamente diferente da normal.

4.1.11 Proximidade da “Highway” - X11

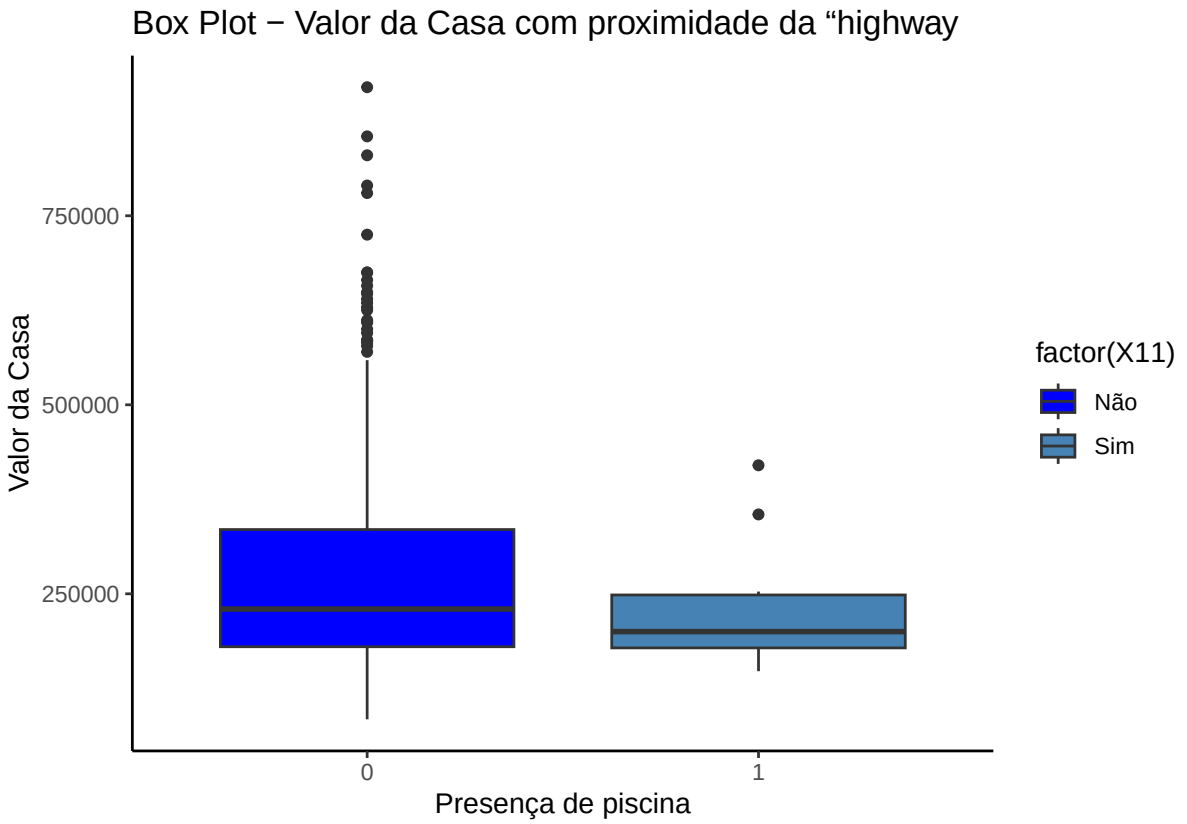


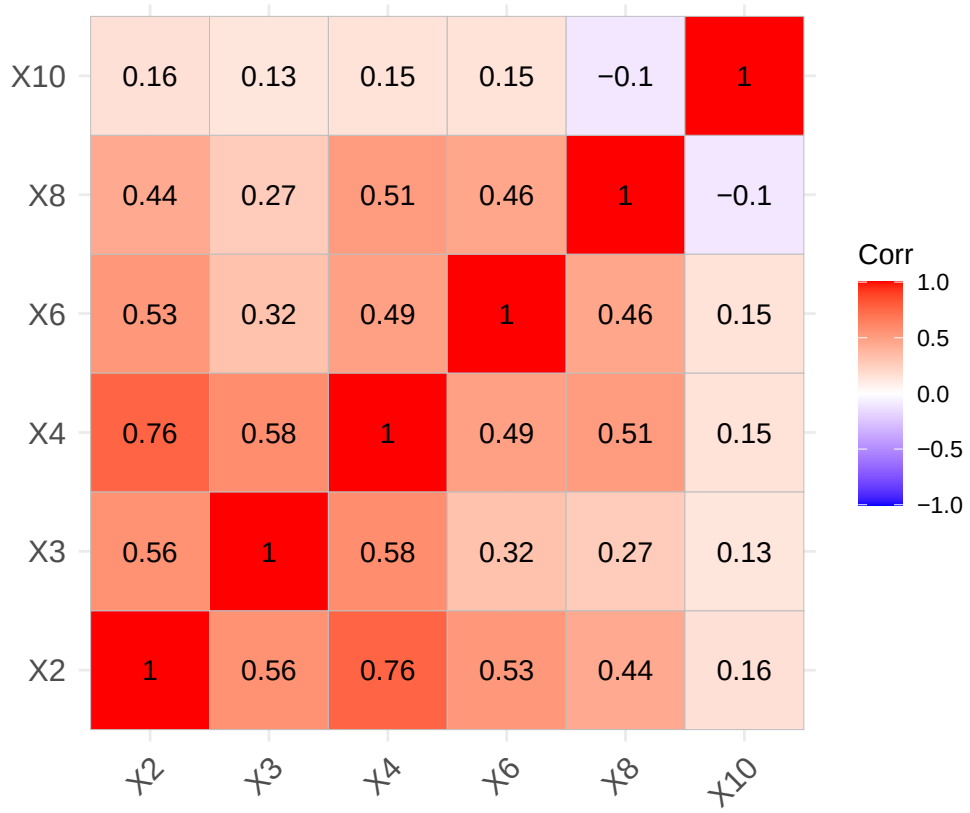
Table 7: Tabela de frequencia da proximidade da highway

	proximidade da highway	Frequencia	Porcentual
1	0	511	97.89
2	1	11	2.11

Por último, a variável proximidade da “Highway” avalia se a casa se encontra próxima ou distante da avenida.

4.2 Modelagem

4.2.1 Correlação entre as Variaveis



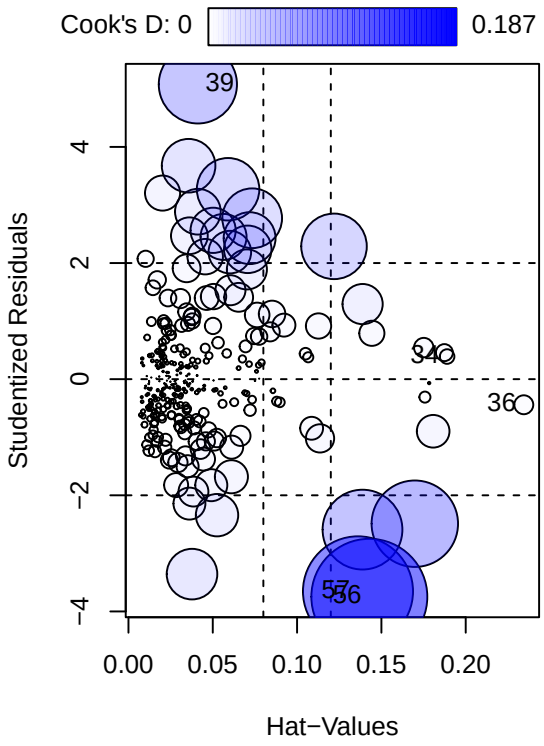
Podemos perceber algumas correlaçoes moderadas mas nenhuma muito forte entre as variaveis. Logo o problema de multicolinearidade nao está presente.

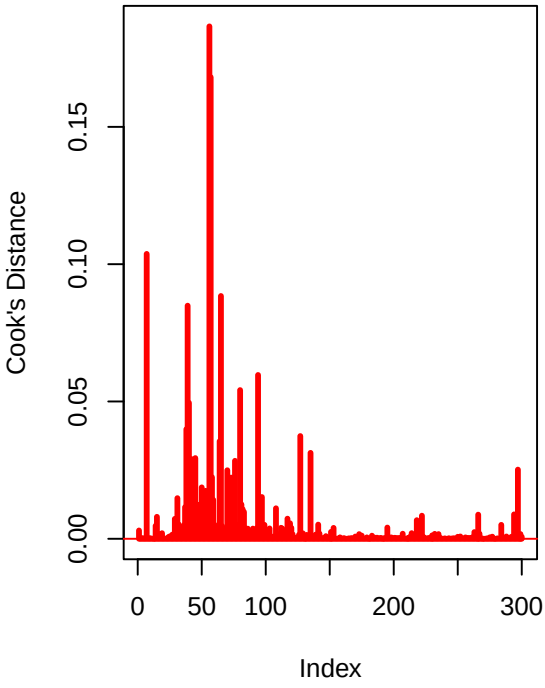
4.2.2 Modelo Completo

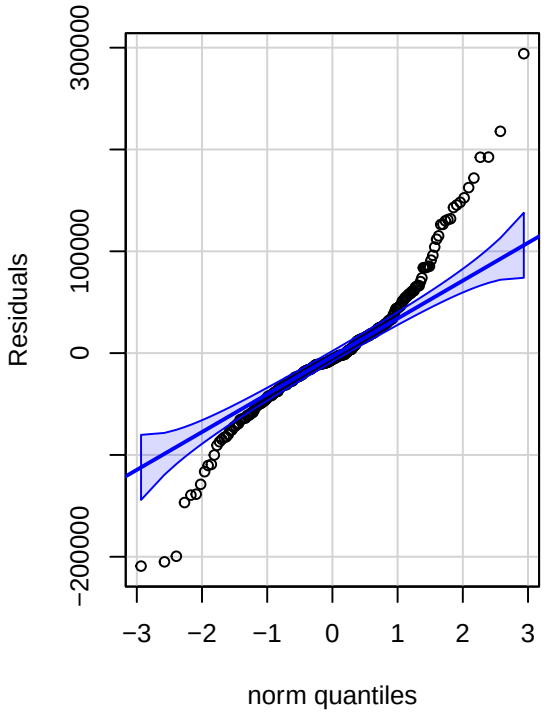
Aqui vamos rodar o modelo utilizando todas as variaveis disponiveis, a ideia é observar os resultados e ter uma base para saber quais variaveis manter e quais variaveis retirar do modelo.

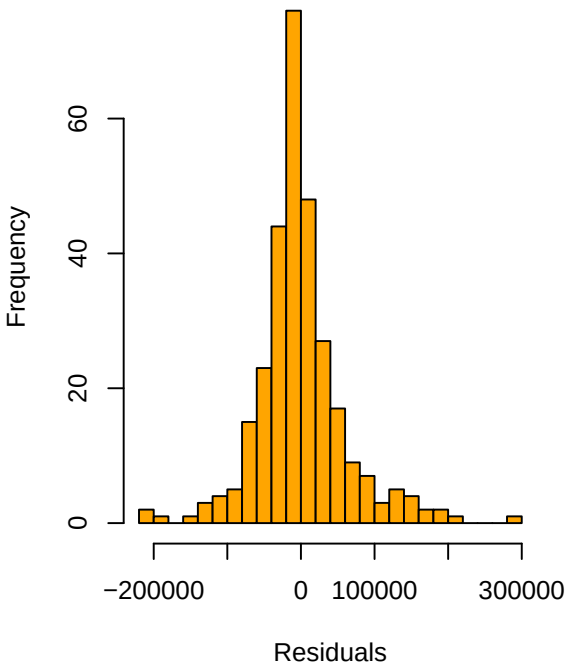
```
##
## Call:
## lm(formula = X1 ~ X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 +
##     X11, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -209192  -28227   -5787   21906  294076
##
## Coefficients:
##                   Estimate    Std. Error t value             Pr(>|t|)
## (Intercept) -2042843.7148   534812.1224  -3.820             0.000164 ***
## X2                101.2214        8.9750  11.278 < 0.0000000000000002 ***
## X3              -8356.8616     4493.4118  -1.860             0.063933 .
## X4               5767.8837     5627.8154   1.025             0.306276
## X51              6430.0087    11454.3102   0.561             0.574988
## X6               9692.4048     7841.7181   1.236             0.217464
## X71              8941.6902    15613.7601   0.573             0.567308
## X8               1108.3759      271.2807   4.086    0.000057018175886 ***
## X92           -139377.3210    14787.4109  -9.425 < 0.0000000000000002 ***
## X93           -149705.9698    19622.4273  -7.629    0.000000000000347 ***
## X10                1.0371        0.3389   3.060             0.002420 **
## X111          -39347.6356    25578.2952  -1.538             0.125068
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61640 on 288 degrees of freedom
## Multiple R-squared:  0.8022, Adjusted R-squared:  0.7947
## F-statistic: 106.2 on 11 and 288 DF,  p-value: < 0.00000000000000022
```

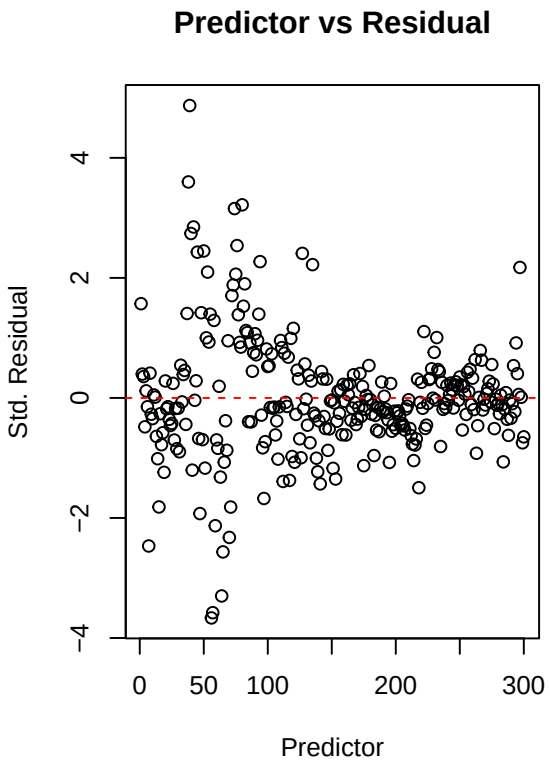
Vamos analisar os residuos desse modelo:

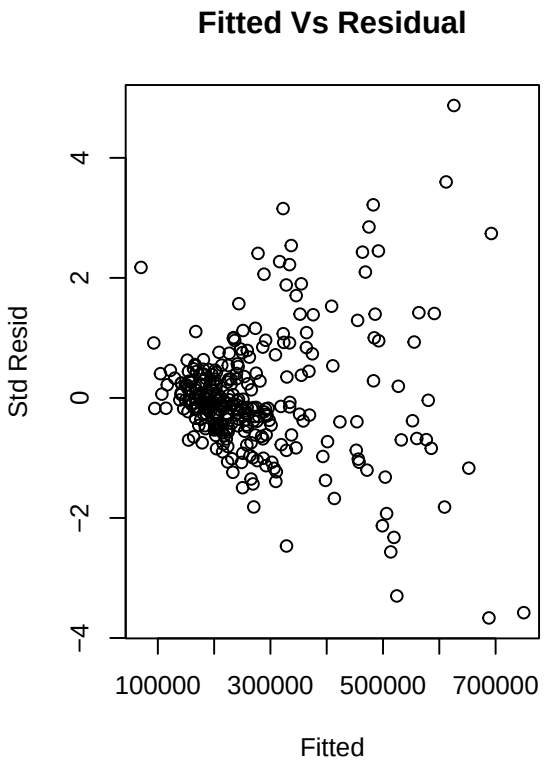


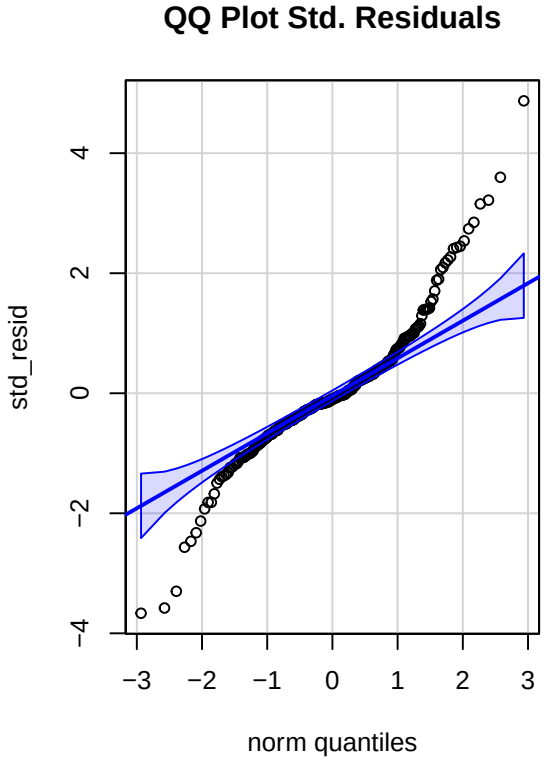


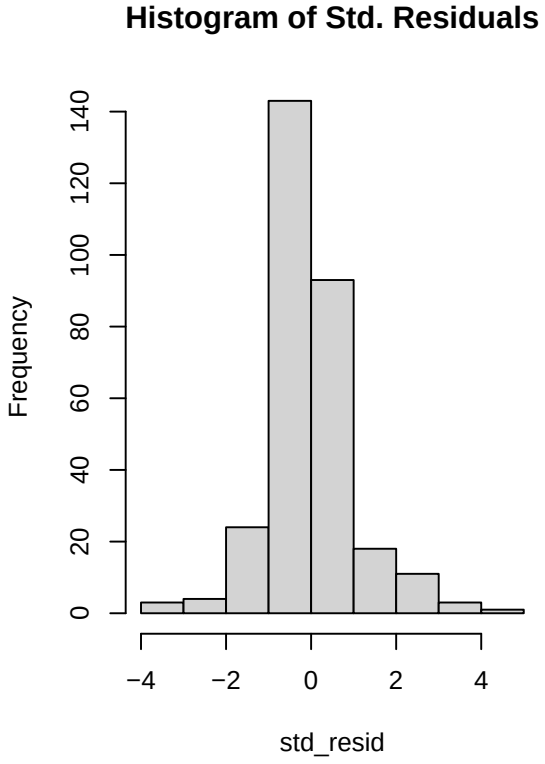





Predictor vs Residual
Std. Residual
Predictor
0
50
100
200
300
−4
−2
0
2
4


Fitted Vs Residual
Std Resid
Fitted
100000
300000
500000
700000
−4
−2
0
2
4


QQ Plot Std. Residuals
std_resid
norm quantiles
−3
−2
−1
0
1
2
3
−4
−2
0
2
4


Histogram of Std. Residuals
Frequency
std_resid
0
20
40
60
80
100
120
140
−4
−2
0
2
4

Vamos observar os VIFs e a Tolerancia para ver se temos multicolinearidade entre as variaveis

```
##    Variables Tolerance      VIF
## 1         X2 0.3458757 2.891212
## 2         X3 0.6312300 1.584209
## 3         X4 0.3766045 2.655306
## 4        X51 0.7306932 1.368564
## 5         X6 0.6247095 1.600744
## 6        X71 0.9719433 1.028867
## 7         X8 0.5316521 1.880929
## 8        X92 0.2347140 4.260505
## 9        X93 0.1495576 6.686389
## 10       X10 0.8865604 1.127955
## 11      X111 0.9877557 1.012396
```

Vale notar que VIFs muito altos pode indicar multicolinearidade, assim como uma tolerancia muito baixa (<0.1). Nos Resultados obtido podemos percerber que o VIF na variavel categorica 9 na categoria 3 está alto porem nao o suficiente para indicar multicolinearidade ja que sua tolerancia esta acima de 0.1

Vamos realizar o teste de Breusch-Pagan e Normalidade dos residuos

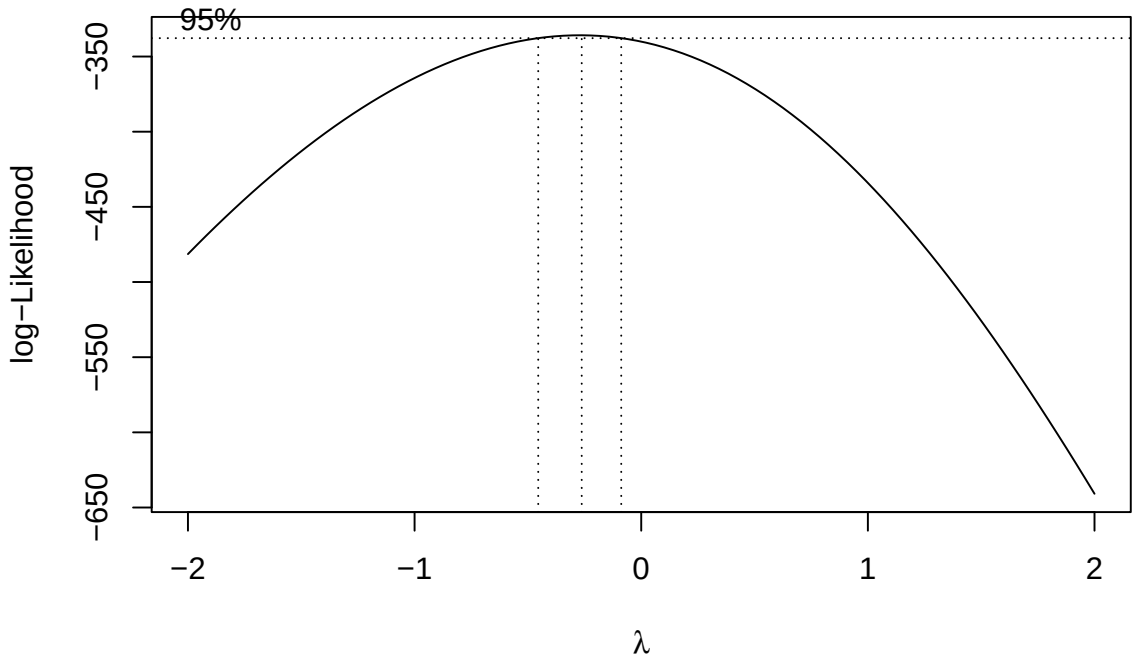
```
##
##  Shapiro-Wilk normality test
##
## data:  modelo_completo$residuals
## W = 0.92376, p-value = 0.00000000003001
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_completo
## BP = 102.93, df = 11, p-value < 0.00000000000000022
```

Pelos resultados obtidos podemos perceber que o modelo nao atende os pressupostos de normalidade nem de homocedasticidade dos residuos, podemos ver claramente uma tendencia do aumento da variancia em relacao aos valores ajustados. Logo faz-se necessario uma transformacao das variaveis com o objetivo de atender esses pressupostos.

4.2.3 Modelo Transformado

Nessa etapa vamos utilizar a transformação de Box-Cox para ajudar a melhorar a aderência dos dados aos pressupostos de normalidade e homocedasticidade, permitindo que os resultados da regressão sejam mais válidos e precisos. É importante lembrar que a interpretação dos resultados após a transformação deve ser feita considerando a escala dos dados transformados, o que pode ser uma dificuldade.



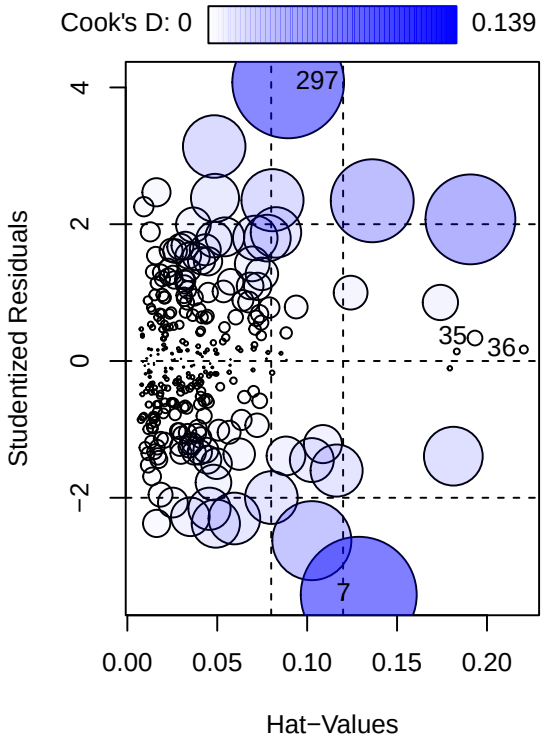

```
## O lambda otimo obtido pelo metodo de Box-Cox é: -0.2626263
```

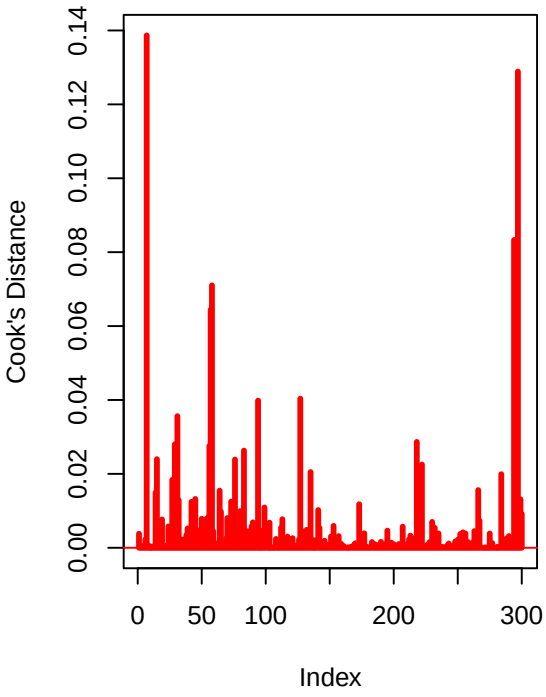
Uma vez obtido o λ da transformacao podemos aplicar no nosso banco de dados de teste e treino.

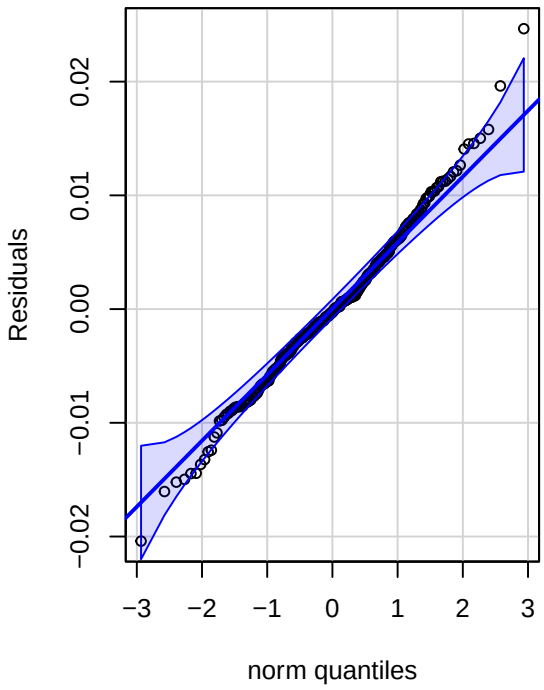
Em seguida podemos obter nosso modelo completo com base nos dados transformados

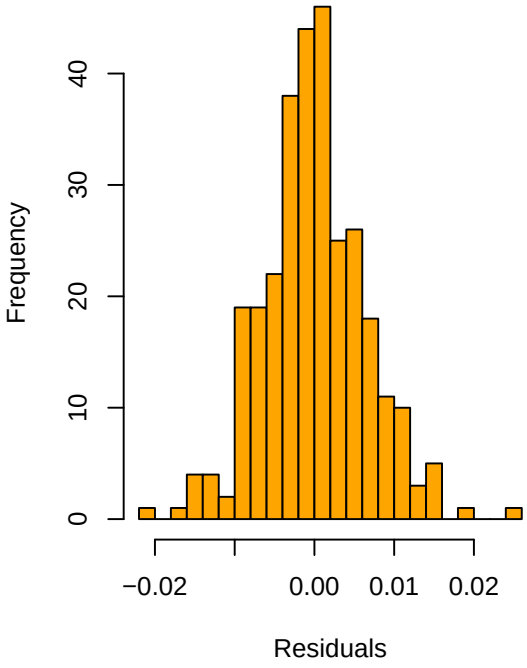
```
##
## Call:
## lm(formula = X1 ~ X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 +
##     X11, data = train_bc)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0204017 -0.0038820 -0.0002565  0.0039487  0.0246542
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept) -3.582481   1.339586  -2.674              0.00792 **
## X2           0.223436   0.016750  13.340 < 0.0000000000000002 ***
## X3           0.001497   0.001797   0.834              0.40525
## X4           0.002583   0.001688   1.530              0.12713
## X51          0.002231   0.001216   1.834              0.06762 .
## X6           0.002703   0.001849   1.462              0.14480
## X71          0.001838   0.001644   1.118              0.26449
## X8           1.920349   0.405434   4.737      0.0000034175782 ***
## X92         -0.010360   0.001483  -6.985      0.0000000000197 ***
## X93         -0.012751   0.002035  -6.265      0.0000000013575 ***
## X10          0.055201   0.013692   4.032      0.0000709902594 ***
## X111        -0.006330   0.002697  -2.347              0.01962 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.006511 on 288 degrees of freedom
## Multiple R-squared:  0.8345, Adjusted R-squared:  0.8282
## F-statistic: 132.1 on 11 and 288 DF,  p-value: < 0.00000000000000022
```

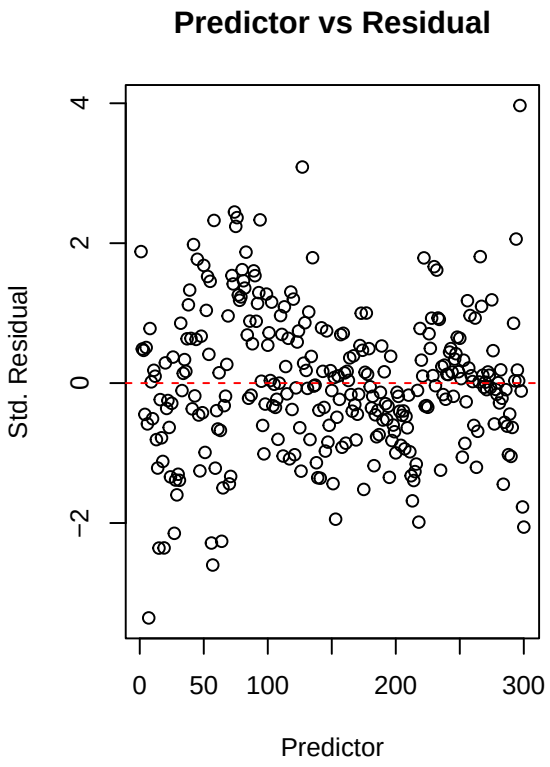
Vamos analisar os residuos desse modelo:

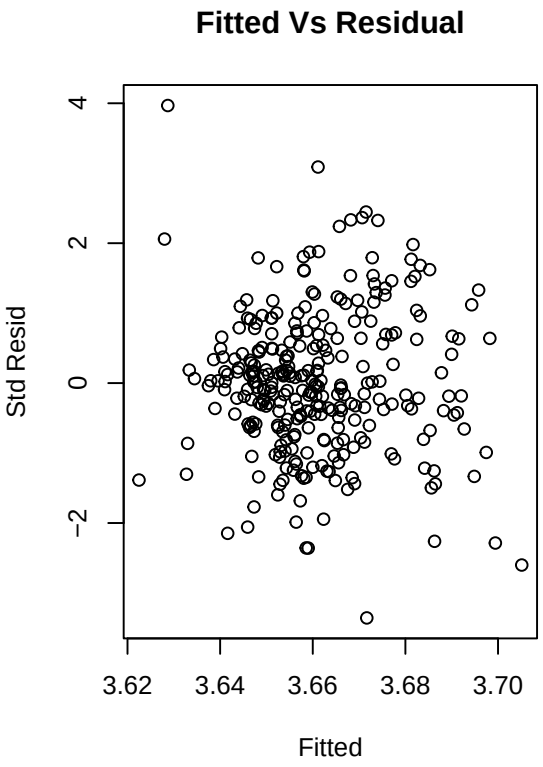


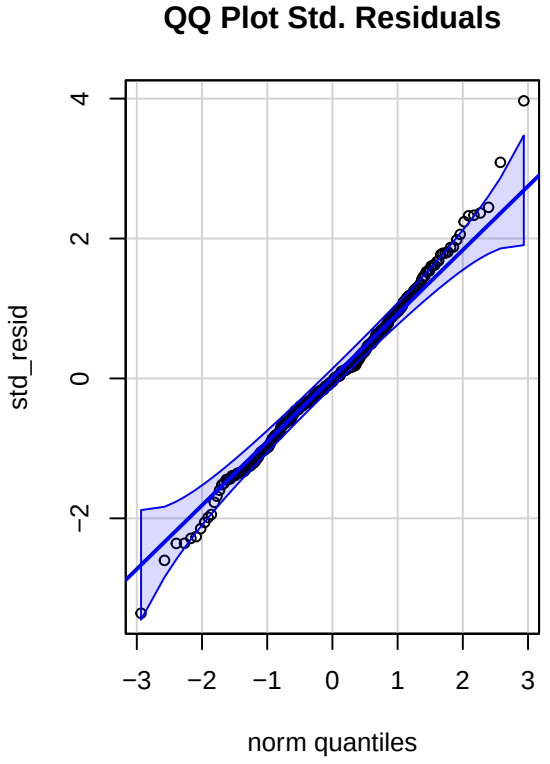


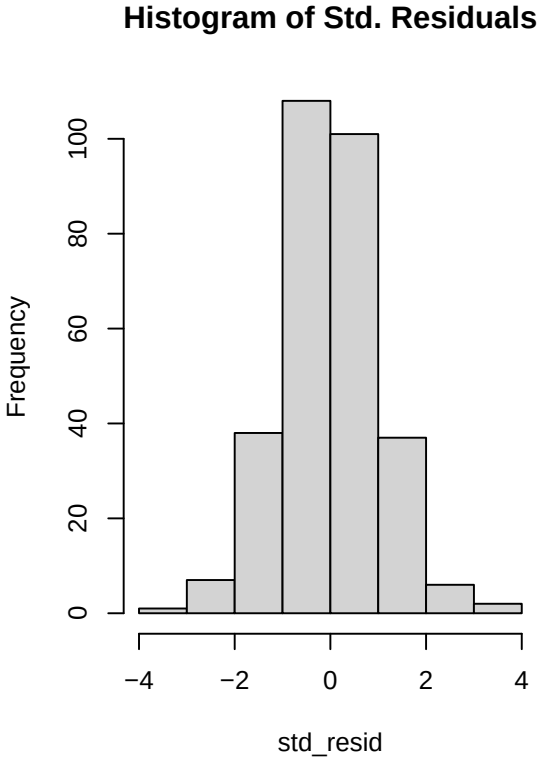





Predictor vs Residual
Std. Residual
Predictor
0
50
100
200
300
−2
0
2
4


Fitted Vs Residual
Std Resid
Fitted
3.62
3.64
3.66
3.68
3.70
−2
0
2
4


QQ Plot Std. Residuals
std_resid
norm quantiles
−3
−2
−1
0
1
2
3
−2
0
2
4


Histogram of Std. Residuals
Frequency
std_resid
−4
−2
0
2
4
0
20
40
60
80
100

Vamos realizar o teste de Breusch-Pagan e de Shapiro para normalidade

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo_completo$residuals
## W = 0.9916, p-value = 0.08614
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_completo
## BP = 24.668, df = 11, p-value = 0.01019
```

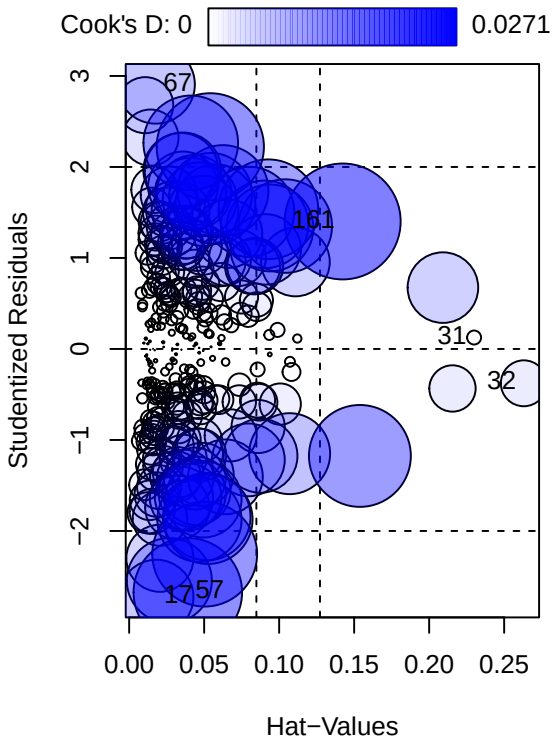
Os resultados mostram uma melhora significativa na adequação do modelo aso pressupostos da regressao! Contudo ainda podemos melhorar mais ainda removendo as observações extremas (influentes) que podemos notar pelo gráfico da distancia de cook.

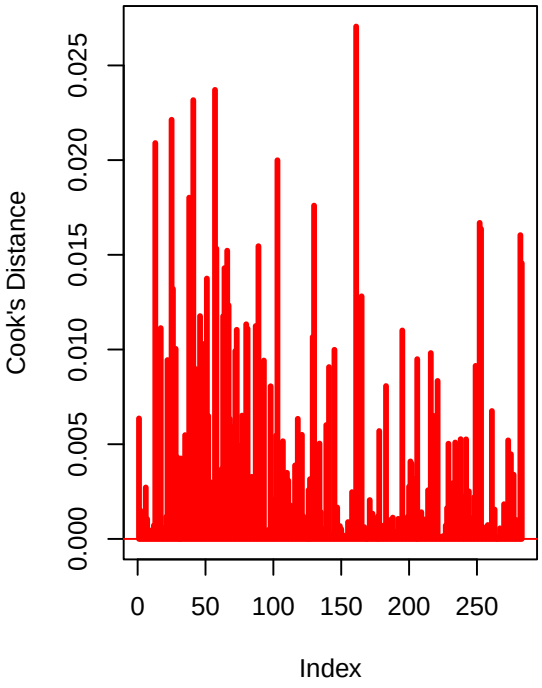
```
## Observações influentes 7 15 29 31 56 57 58 76 83 94 127 135 218 222 284 294 297
```

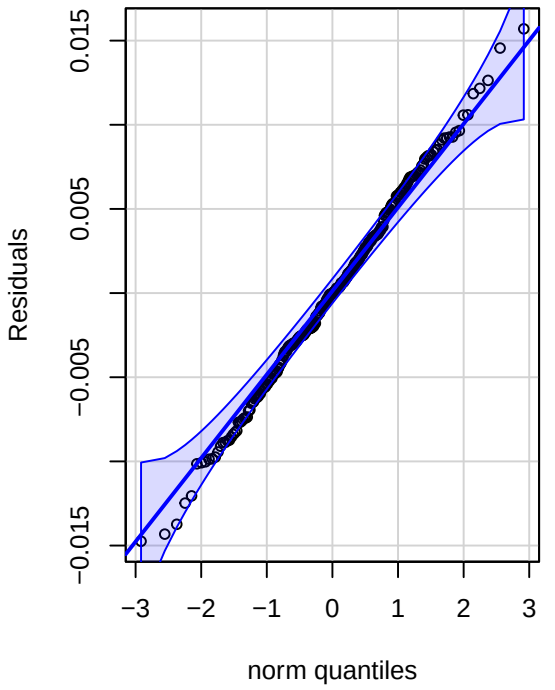
Podemos rodar nosso modelo novamente e ver os resultados.

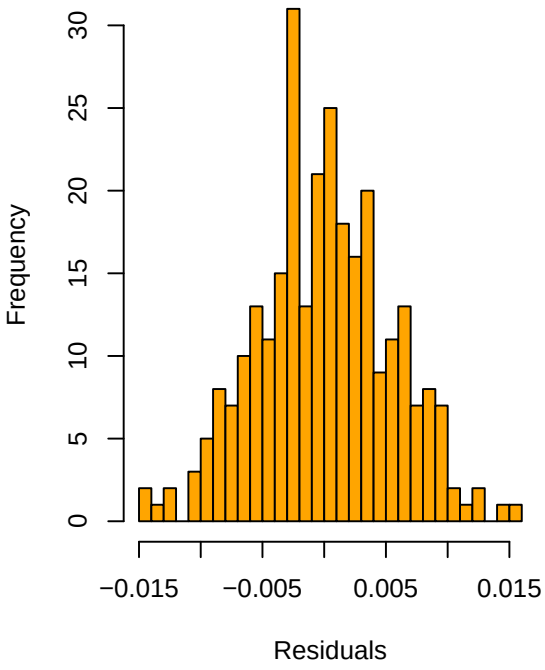
Agora obtemos nosso novo modelo sem observações extremas

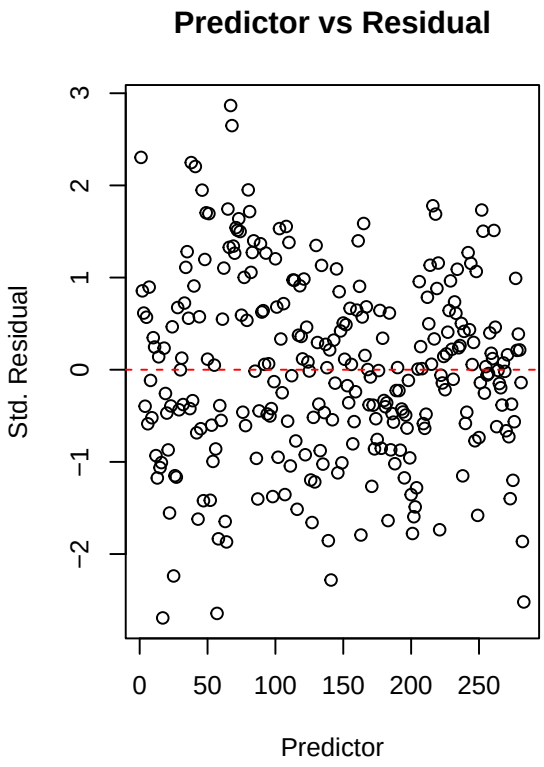
```
##
## Call:
## lm(formula = X1 ~ X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 +
##     X11, data = train_bc)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0147444 -0.0032233 -0.0000061  0.0034724  0.0156987
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept) -4.719308   1.286941  -3.667             0.000295 ***
## X2           0.235323   0.014906  15.788 < 0.0000000000000002 ***
## X3           0.001421   0.001847   0.770             0.442203
## X4           0.003406   0.001519   2.243             0.025714 *
## X51          0.002893   0.001089   2.656             0.008370 **
## X6           0.003568   0.001633   2.185             0.029728 *
## X71          0.003150   0.001587   1.985             0.048161 *
## X8           2.232600   0.388885   5.741     0.00000002518045 ***
## X92         -0.009719   0.001306  -7.439     0.00000000000134 ***
## X93         -0.010380   0.001799  -5.769     0.00000002171574 ***
## X10          0.074434   0.013045   5.706     0.00000003026835 ***
## X111        -0.004326   0.002507  -1.726             0.085548 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005525 on 271 degrees of freedom
## Multiple R-squared:  0.8774, Adjusted R-squared:  0.8725
## F-statistic: 176.4 on 11 and 271 DF,  p-value: < 0.00000000000000022
```

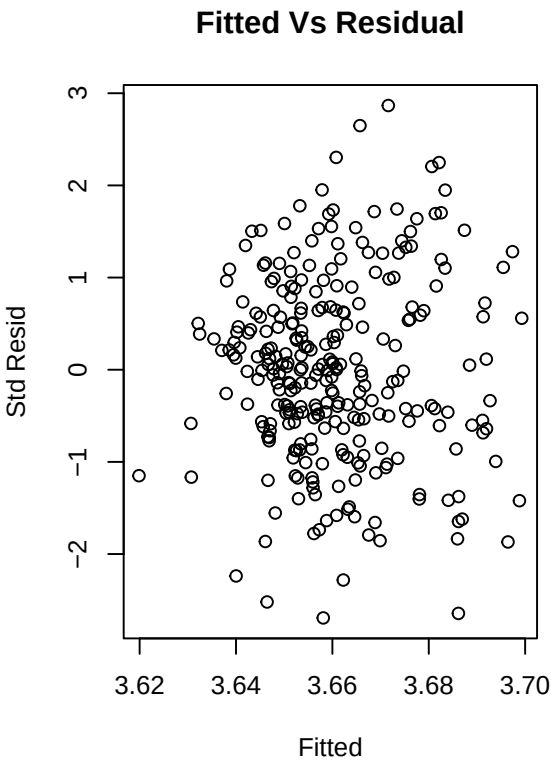

Cook's D: 0
0.0271
Studentized Residuals
Hat−Values
67
161
31
32
1757
0.00
0.05
0.10
0.15
0.20
0.25
3
2
1
0
−1
−2

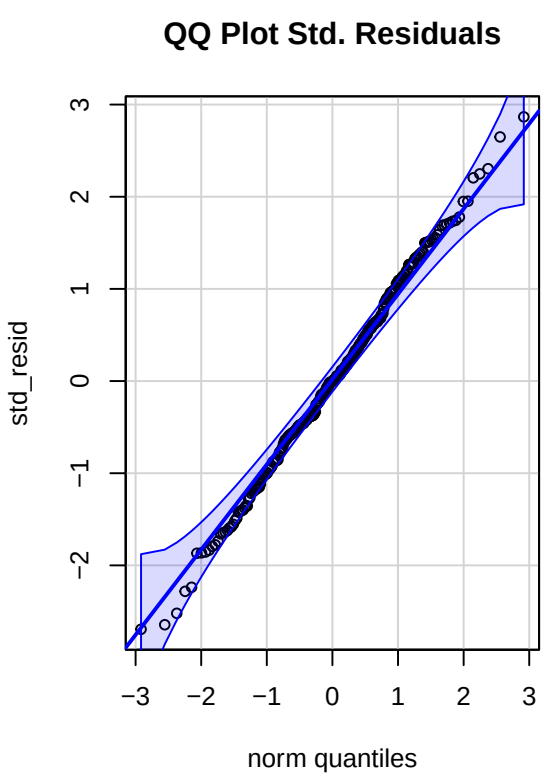

Cook's Distance
Index
0.000
0.005
0.010
0.015
0.020
0.025
0
50
100
150
200
250

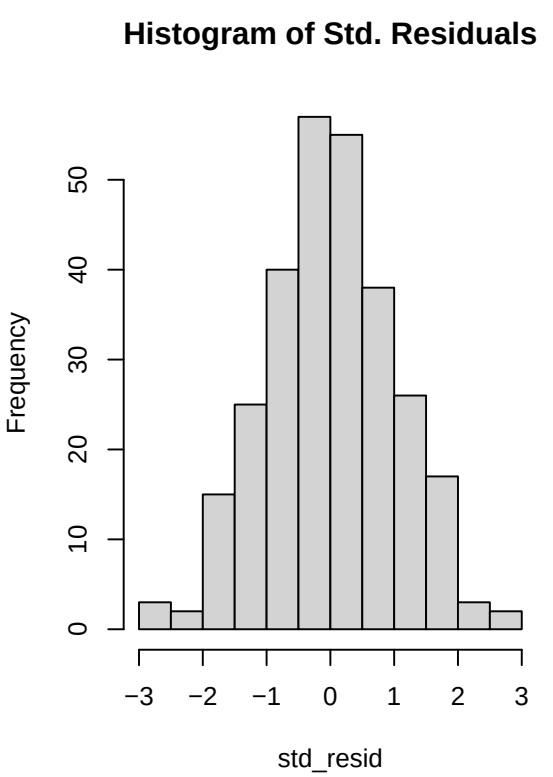

Residuals
norm quantiles
0.015
0.005
−0.005
−0.015
−3
−2
−1
0
1
2
3


Frequency
Residuals
0
5
10
15
20
25
30
−0.015
−0.005
0.005
0.015


Predictor vs Residual
Std. Residual
Predictor
0
50
100
150
200
250
−2
−1
0
1
2
3


Fitted Vs Residual
Std Resid
Fitted
3.62
3.64
3.66
3.68
3.70
−2
−1
0
1
2
3


QQ Plot Std. Residuals
std_resid
norm quantiles
−3
−2
−1
0
1
2
3


Histogram of Std. Residuals
Frequency
std_resid
0
10
20
30
40
50
−3
−2
−1
0
1
2
3

Teste de Breusch-Pagan e Shapiro

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_completo
## BP = 25.08, df = 11, p-value = 0.008875
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo_completo$residuals
## W = 0.99754, p-value = 0.9487
```

Podemos notar que melhorarmos muito a aderencia a normalidade, apesar de ter tido uma queda no p-valor de teste de Breusch-Pagan. Contudo nao muda pois nao chegamos perto de atingir os 5% necessarios para assumir homocedasticidade.

4.2.4 Modelo Reduzido

Para obter os modelo reduzidos, vamos rodar os algoritmos forward,backward e stepwise para selecao de parametros, assim como obter todas as combinacoes possiveis e ver os melhores subsets de variavies que mantenha a regressao simples e efetiva, ou seja onde o gnaho de se adicionar mais uma variavel nao é tao grande.

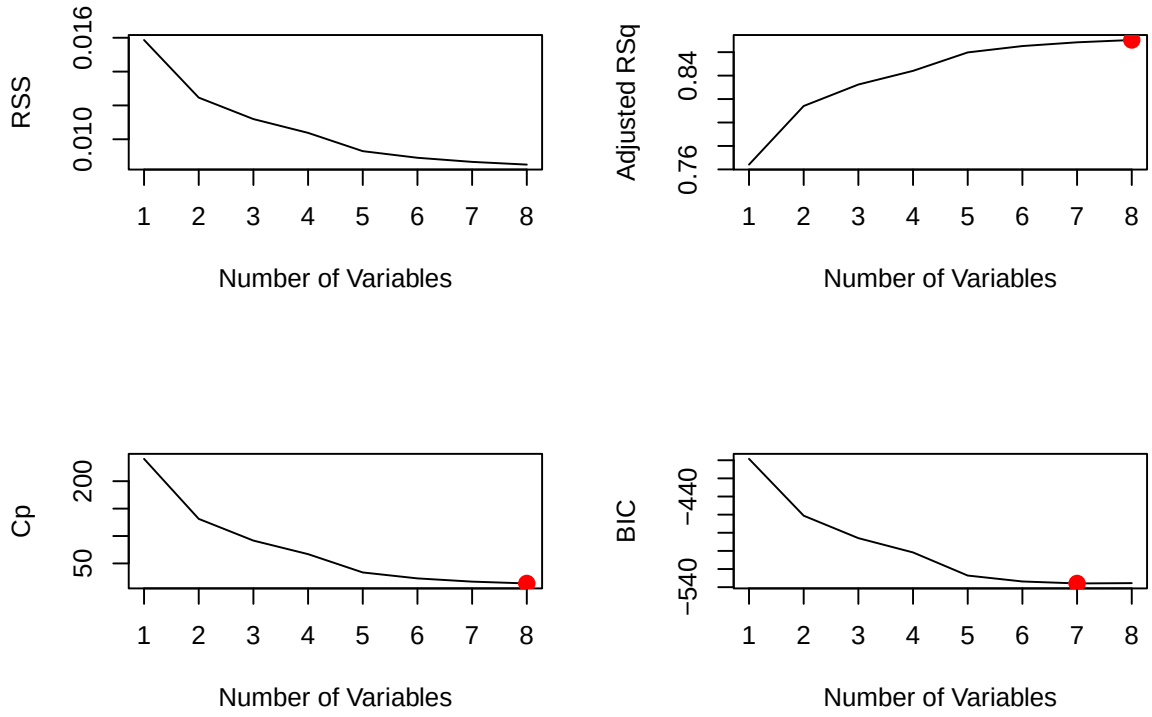
Table 8: Tabela do resultado da seleção de variavel e seu coeficiente

	X.Intercept.	X2	X3	X4	X51	X6	X71	X8	X92	X93	X10	X111
Forward	-4.72	0.235	0.001	0.003	0.003	0.004	0.003	2.233	-0.01	-0.01	0.074	-0.004
Backward	-4.69	0.238	NA	0.004	0.003	0.004	0.003	2.221	-0.01	-0.01	0.074	-0.004
Both	-4.69	0.238	NA	0.004	0.003	0.004	0.003	2.221	-0.01	-0.01	0.074	-0.004

Analisando todas as possiveis combinacoes das variaveis presentes no banco

```
##          X2  X3  X4  X51 X6  X71 X8  X92 X93 X10 X111
## 1  ( 1 ) "*" " " " " " " " " " " " " " " " " " " " "
## 2  ( 1 ) "*" " " " " " " " " " " " " "*" " " " " " " "
## 3  ( 1 ) "*" " " " " " " " " " " " " "*" " " " " "*" " "
## 4  ( 1 ) "*" " " " " " " " " " " " " "*" "*" "*" " " " "
## 5  ( 1 ) "*" " " " " " " " " " " " " "*" "*" "*" "*" " "
## 6  ( 1 ) "*" " " " " " "*" " " " " " "*" "*" "*" "*" " "
## 7  ( 1 ) "*" " " "*" "*" " " " " " "*" "*" "*" "*" " "
## 8  ( 1 ) "*" " " "*" "*" "*" " " " "*" "*" "*" "*" " "
```

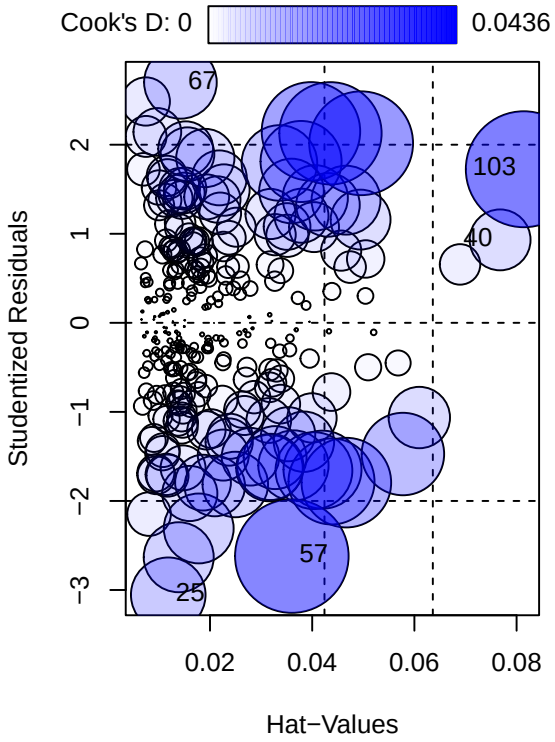
Essa tabela nos traz os melhores modelos com 1 até 8 variavies e as respectivas variaveis presentes nesse modelo (marcadas com um *)

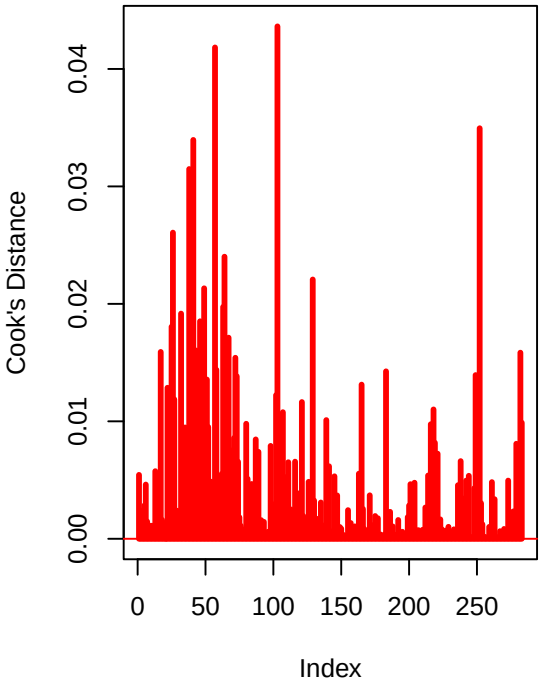


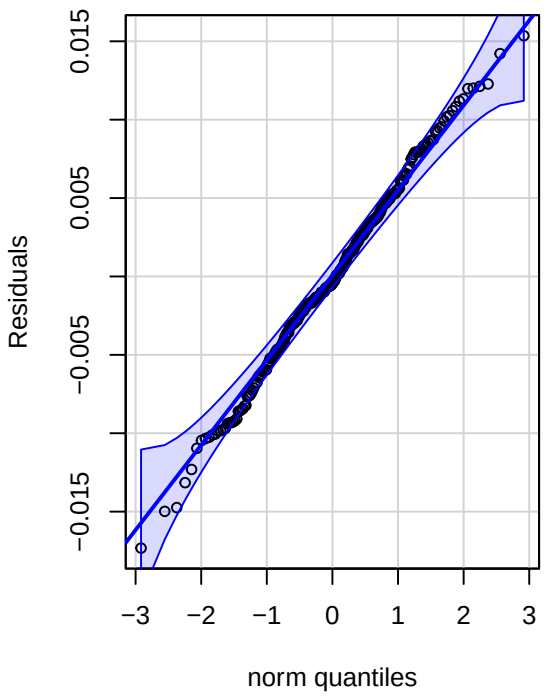
Apartir dos graficos podemos notar que o ganho de explicação do modelo com mais de 5 variavies se torna bem pequena. Indicando que o número ideal e minimo seja 5. Logo nosso modelo reduzido seria:

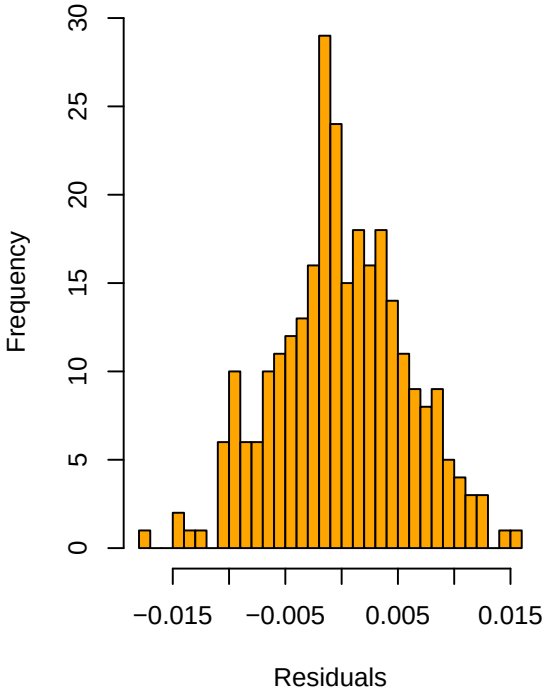
```
##
## Call:
## lm(formula = X1 ~ X2 + X8 + X9 + X10, data = train_bc)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.017327 -0.003573 -0.000445  0.003735  0.015357
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept) -6.216805   1.299233  -4.785     0.00000278467399 ***
## X2           0.261853   0.013353  19.610 < 0.0000000000000002 ***
## X8           2.663462   0.393131   6.775     0.00000000007455 ***
## X92         -0.009944   0.001338  -7.434     0.00000000000132 ***
## X93         -0.012400   0.001806  -6.865     0.00000000004348 ***
## X10          0.075257   0.013263   5.674     0.00000003507656 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005793 on 277 degrees of freedom
## Multiple R-squared:  0.8623, Adjusted R-squared:  0.8598
## F-statistic: 346.9 on 5 and 277 DF,  p-value: < 0.00000000000000022
```

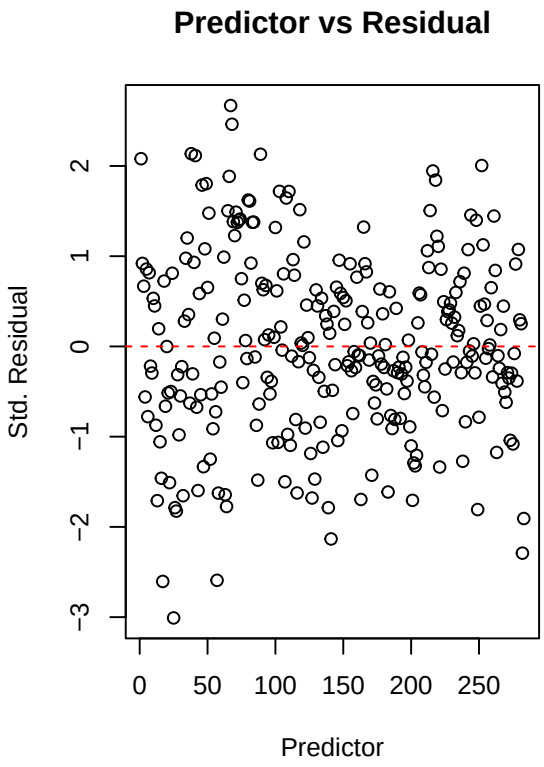
Vamos analisar os residuos desse modelo:

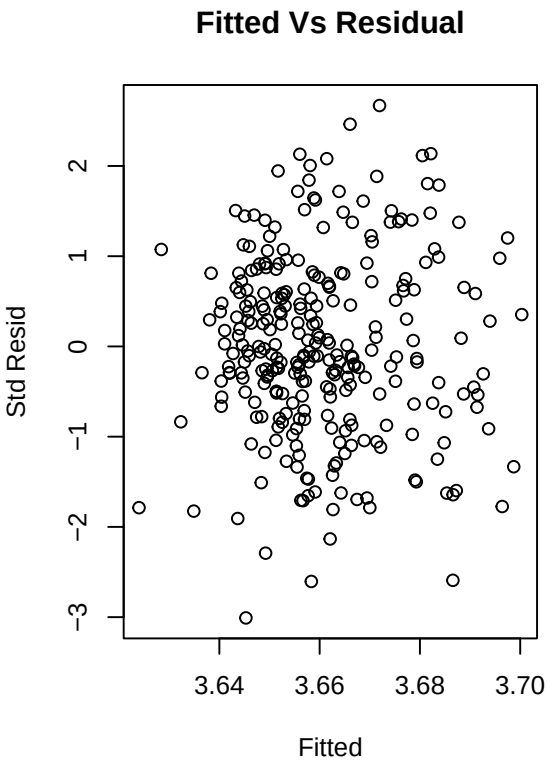


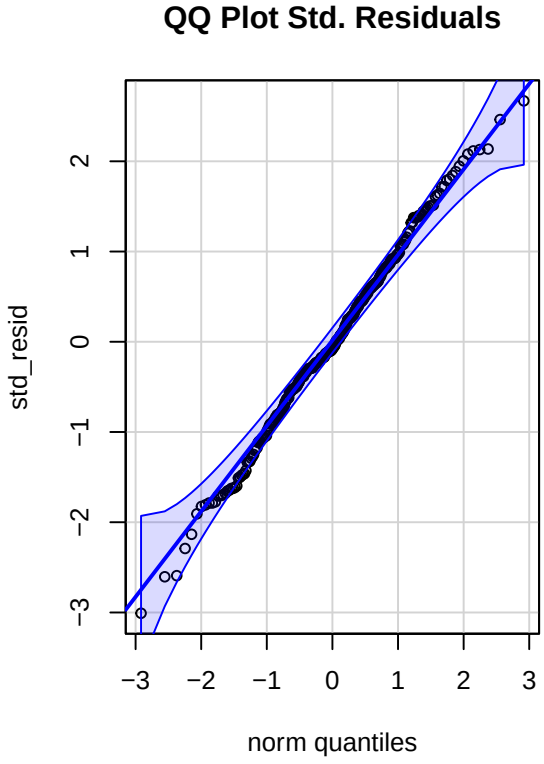


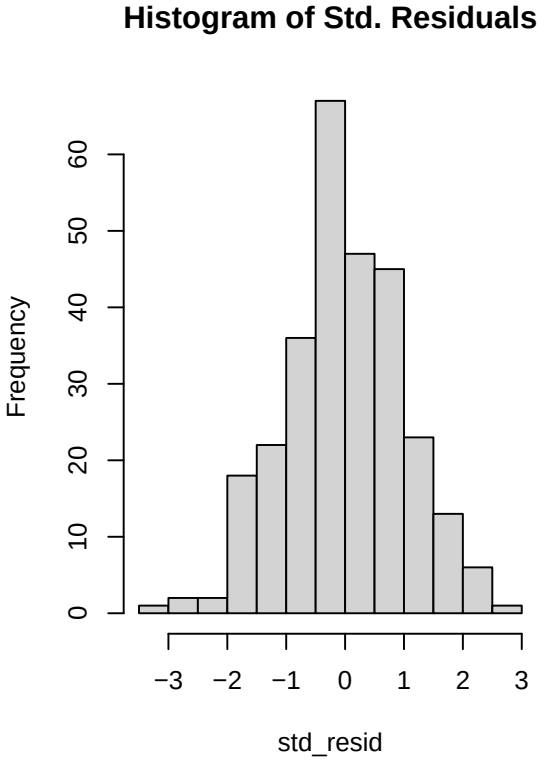





Predictor vs Residual
Std. Residual
Predictor
0
50
100
150
200
250
−3
−2
−1
0
1
2


Fitted Vs Residual
Std Resid
Fitted
3.64
3.66
3.68
3.70
−3
−2
−1
0
1
2


QQ Plot Std. Residuals
std_resid
norm quantiles
−3
−2
−1
0
1
2
3


Histogram of Std. Residuals
Frequency
std_resid
0
10
20
30
40
50
60
−3
−2
−1
0
1
2
3

Teste de Breusch-Pagan e Shapiro

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_completo
## BP = 25.08, df = 11, p-value = 0.008875
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo_completo$residuals
## W = 0.99754, p-value = 0.9487
```

4.2.5 Validação do Modelo

Vamos usar os 2 modelo obtidos e transformados por meio da transformação de Box-Cox para aplicar nos bancos de teste e verificar o Raiz do Erro Quadratico Médio e o Erro Médio Absoluto.

Modelo completo e transformado

```
##          R2        REQM         EMA
## 1 0.8548086 0.006496541 0.004992057
```

```
## Taxa de erro da predição do modelo completo 0.001773817
```

Modelo Reduzido e transformado

```
##          R2        REQM         EMA
## 1 0.8465789 0.006521815 0.005040489
```

```
## Taxa de erro da predição do modelo reduzido 0.001780718
```

Modelo completo original

```
##         R2    REQM     EMA
## 1 0.839498 56497.4 39650.4
```

```
## Taxa de erro da predição do modelo completo 86881.56
```

Podemos perceber que o modelo transformado apresenta um desempenho muito superior ao modelo com as variáveis originais. Isso se deve ao fato dos pressupostos não estarem sendo atendidos, e as transformações ajudam nisso. Os modelos transformados estão com as taxas de erros bem baixas, além de que vale notar que, com 5 variáveis, obtemos um desempenho muito parecido com o modelo completo. Em outras palavras, podemos simplificar e obter um resultado praticamente igual.

5 Conclusão

Podemos concluir que a regressão ajustada usando os dados originais não atende aos pressupostos de uma regressão, e isso causa diversos problemas na predição do modelo, que foi mostrado nos resultados de validação. Logo, foi necessário buscar uma transformação que aproxime os dados para a normalidade, melhorando assim a aderência dos dados aos pressupostos. Para isso, foi feita a transformação de Box-Cox, que busca, por meio da máxima verossimilhança, o valor de lambda ideal que mais aproxima o modelo da normalidade. Uma vez obtido esse lambda e transformados os dados, foi notada uma grande melhora no modelo, apesar de ainda não aceitar o pressuposto de variância constante, que foi o maior problema observado nos modelos. Apesar disso, conseguimos mostrar que o modelo apresenta um bom desempenho e que ele pode ser reduzido para apenas 5 parâmetros (4 variáveis, sendo uma delas categórica com 3 níveis), mantendo o desempenho muito próximo do modelo completo.