Relatório Final

Análise de Regressao Linear

Davi Wentrick Feijó

July 22, 2023

1 Introdução

Nesse estudo, analisaremos um conjunto de dados de uma cidade americana, composta por informações de venda de 522 que foram vendidas no último ano. O propósito desse estudo é a construção de um modelo preditivo que busca prever a média da venda das casas anunciadas, dado algumas características da casa.

O trabalho possui como estrutura a segunda seção com analise descritiva das variáveis, a terceira seção conta com a seleção de variáveis a serem adotadas no modelo. O modelo junto com sua validação se encontra na seção quatro e, por fim, os resultados são apresentados na seção cinco.

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 amostra em 2 partes, a primeira, composta por 300 amostras selecionadas aleatoriamente, para a construção do modelo, e a segunda com as 222 restantes visando a validação do modelo preditivo.

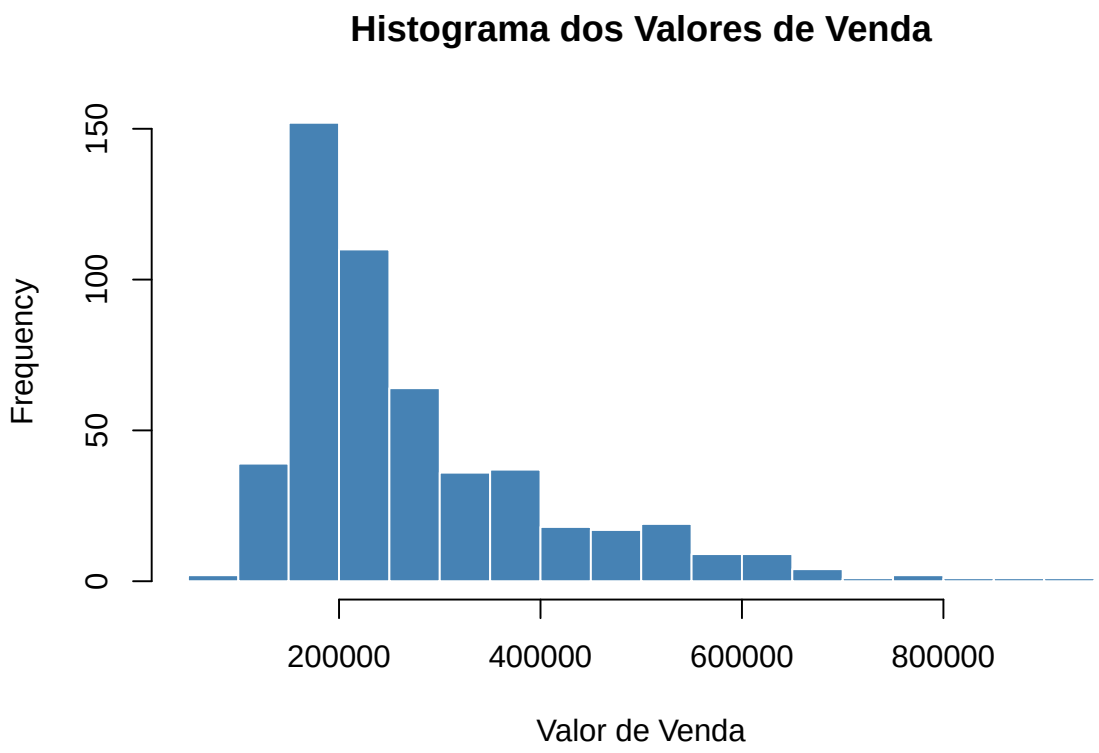
Como forma de mensurar a seleção de variáveis do modelo usaremos na seleção de variáveis os critérios R^2, Cp de Mallows e regressão “Stepwise”. As variáveis que constam no banco de dados são o preço de venda, tamanho da casa, número de quartos, número de banheiros, presença de ar-condicionado, tamanho 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.

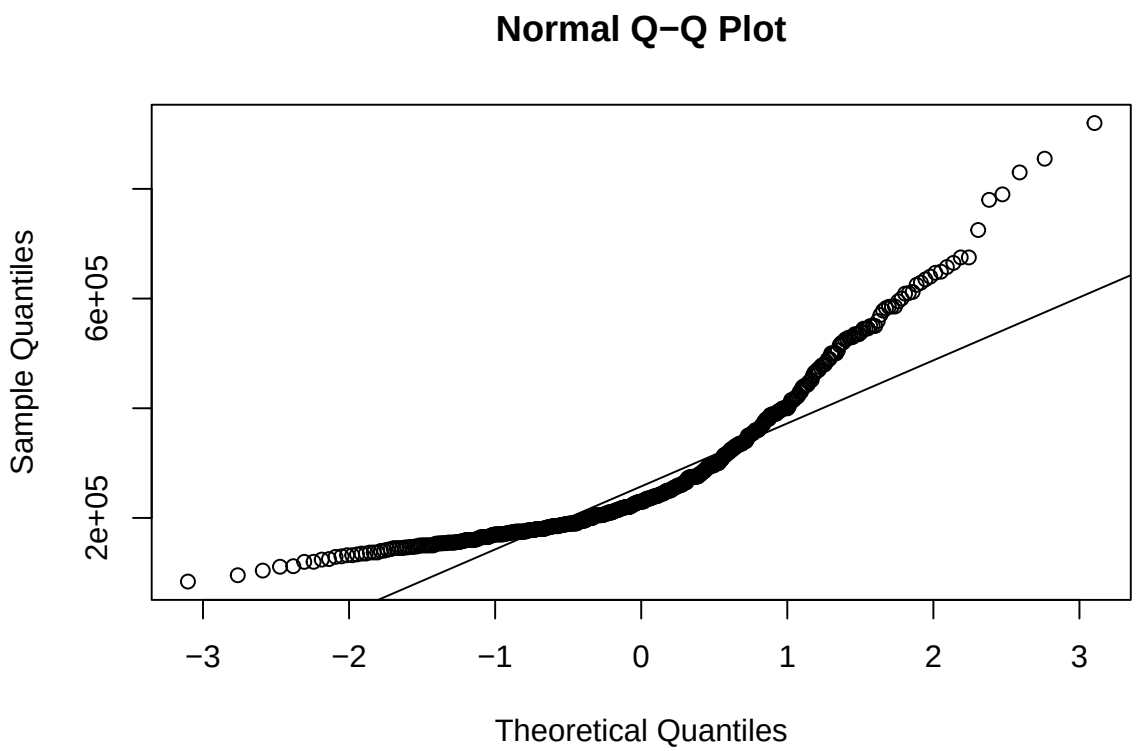
3 Metodologia

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
```

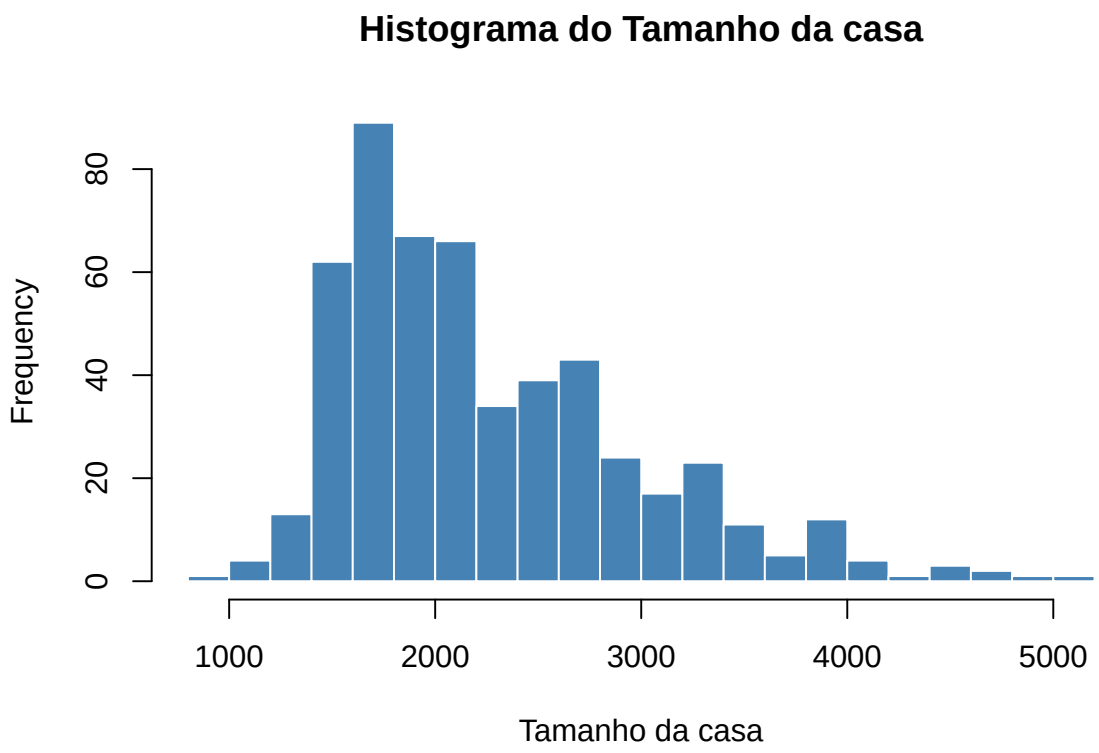
```
##
##  Shapiro-Wilk normality test
##
## data:  dados$X1
## W = 0.84372, p-value < 2.2e-16
```

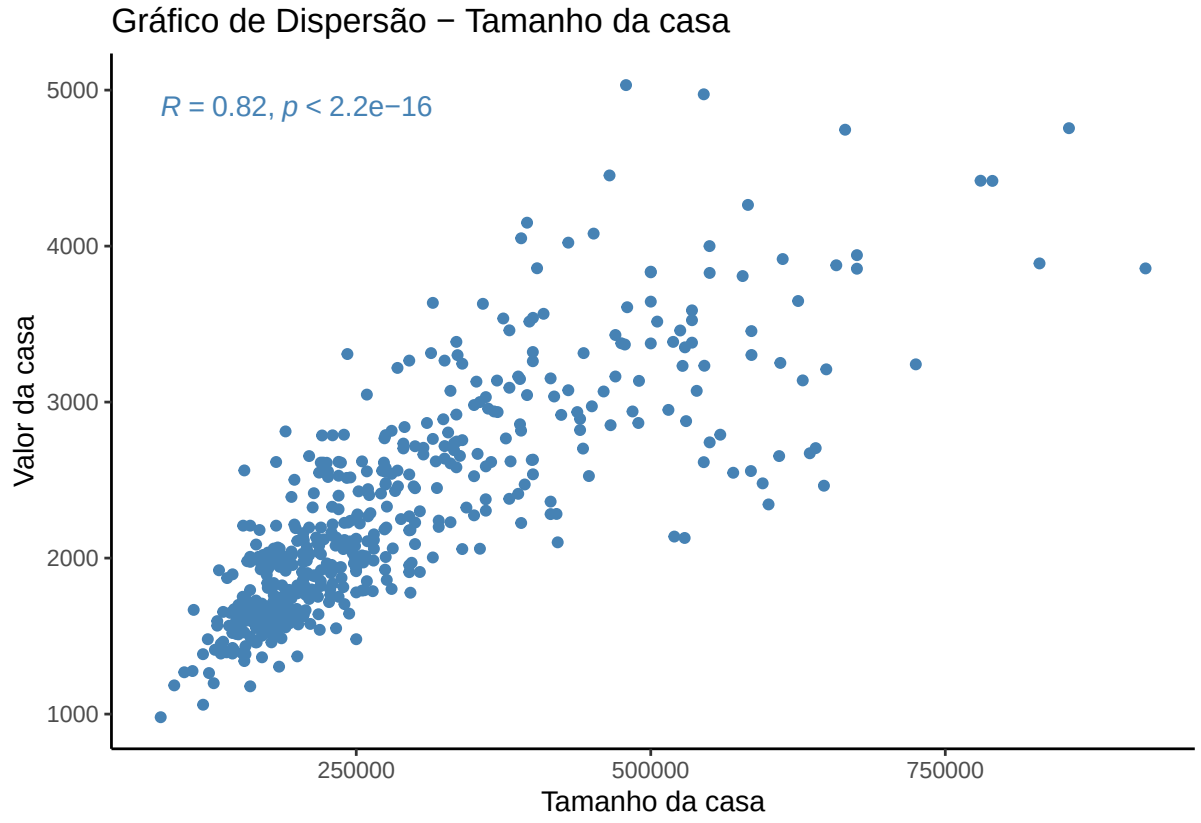
Das onze variáveis apresentadas no banco de dados, a variável preço de venda é nossa variável resposta, sendo ela uma variável quantitativa e continua, que apresenta como medidas as seguintes:

O que nos mostra que a maior parte dos preços estão concentrados na faixa 180.000 até 230.000, conforme podemos observar no gráfico 1

Nossa variável embora não pareça seguir uma distribuição normal, pelo método de Shapiro-Wilk vemos que ela segue a normal e possui uma calda alongada dado por possuir um valor outlier (920.000).

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 < 2.2e-16
```

A variável tamanho da casa, também é uma variável quantitativa e continua, medida em Pés-quadrado e que apresenta como medidas:

Observa-se que o maior valor é 5.032 pés-quadrado o que difere da média, a maior parte dos valores estão concentrados na faixa próxima de 2.000 pés, juntamente temos o histograma de dispersão em relação a variável resposta, conforme seguem os gráficos abaixo:

4.1.3 Número de Quartos - X3

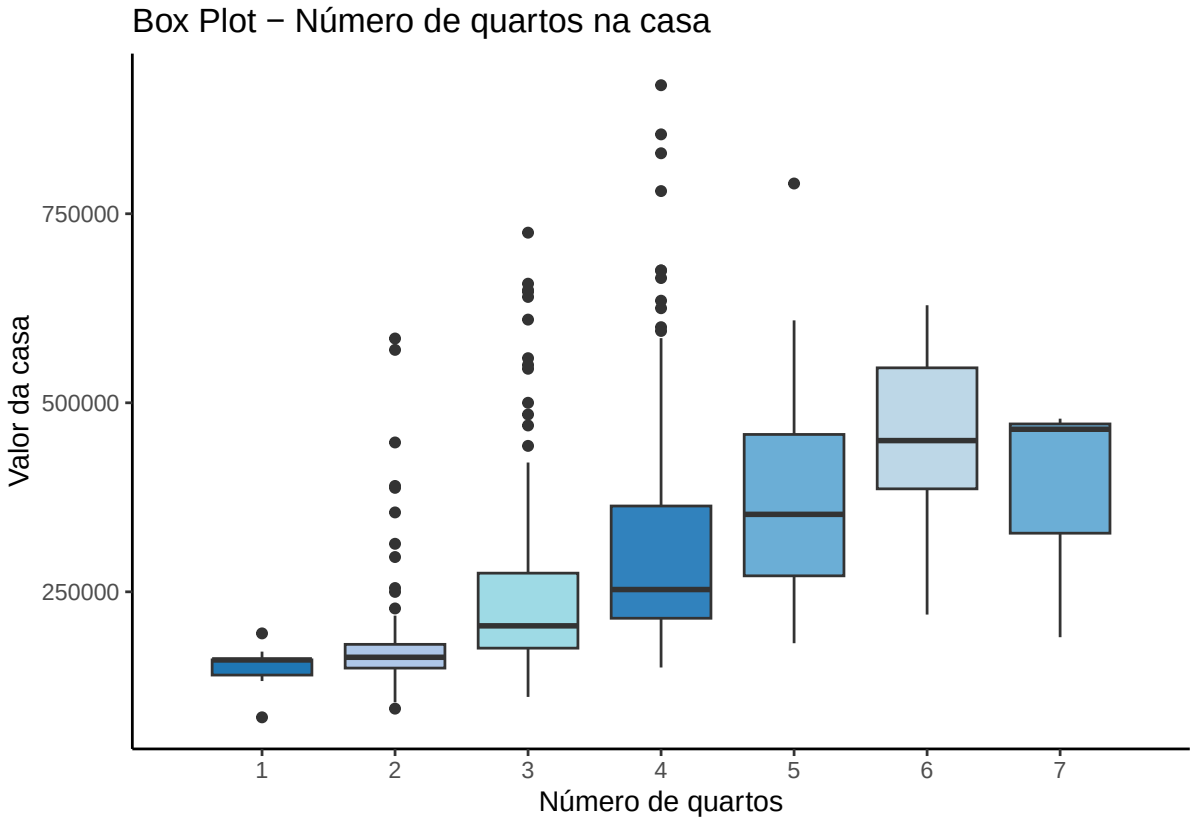


Table 1: Tabela de frequencia do Número de quartos

	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 < 2.2e-16
```

Nossa terceira variável refere-se ao número de quartos que as casas estudadas possuíam, acredita-se que casas com maior número de quartos são mais valorizadas, essa variável é uma variável quantitativa discreta com casas possuindo entre 0 e 7 quartos conforme tabela abaixo:

Número de quartos Quantidade 0 1 1 9 2 64 3 202 4 175 5 52 6 12 7 3

Ao se observar o gráfico x vemos que a variável segue uma distribuição normal, que se comprova ao aplicar o teste de normalidade.

4.1.4 Número de Banheiros - X4

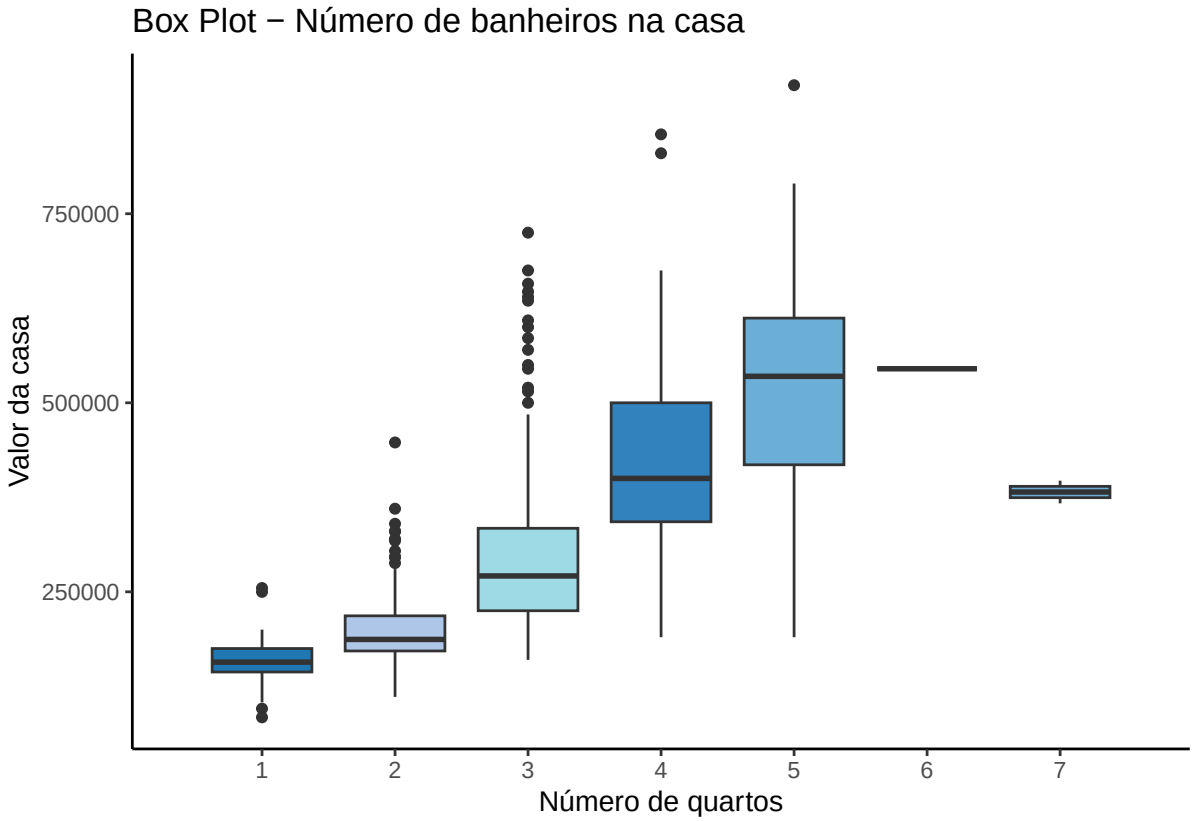


Table 2: Tabela de frequencia do Número de banheiros

	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 < 2.2e-16
```

A quarta variável, número de banheiros segue os mesmos princípios da quarta variável e tem como dados: Número de banheiros Quantidade 0 1 1 71 2 170 3 174 4 83 5 17 6 11 7 2

4.1.5 Presença de Ar-Condicionado - X5

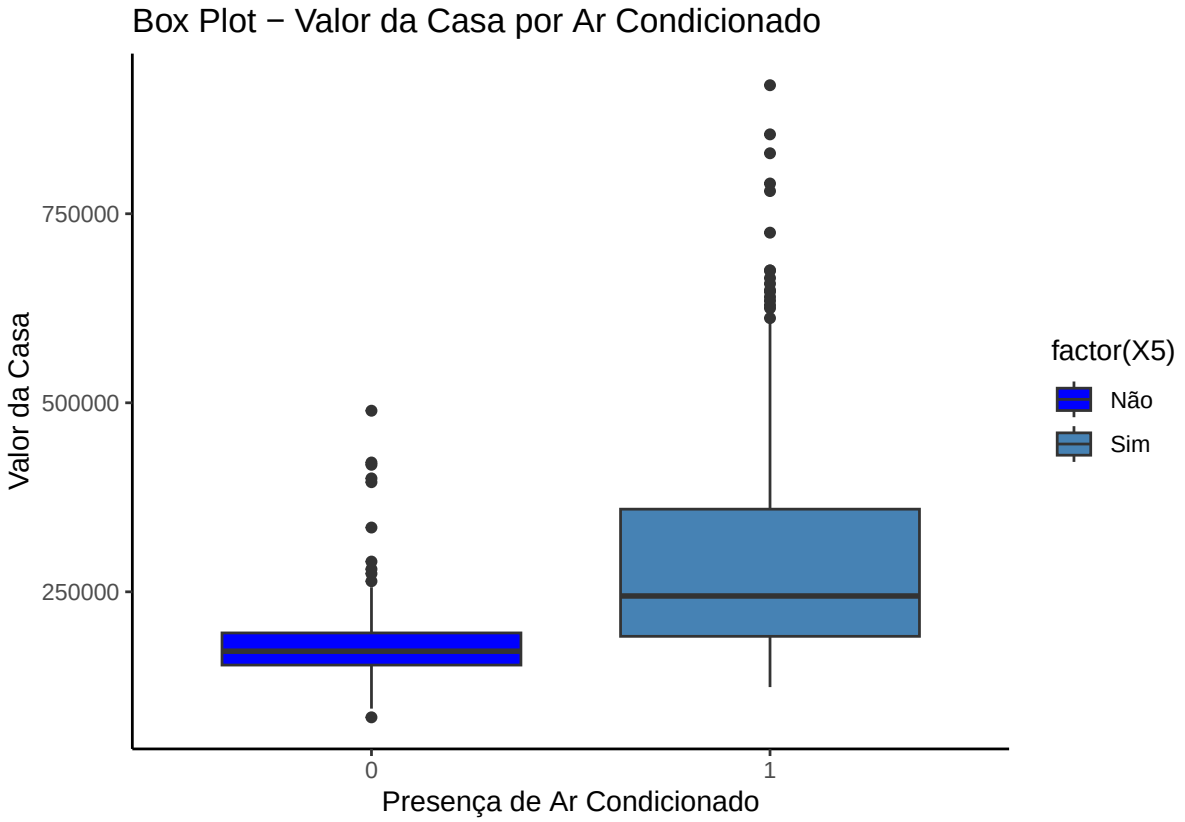


Table 3: Tabela de frequencia do Ar Condicionado

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

A presença de ar-condicionado foi selecionada como a quinta variável, sendo essa uma variável qualitativa, com 1 para o caso de haver e 0 para quando não há ar condicionado na residência.

4.1.6 Tamanho da Garagem - X6

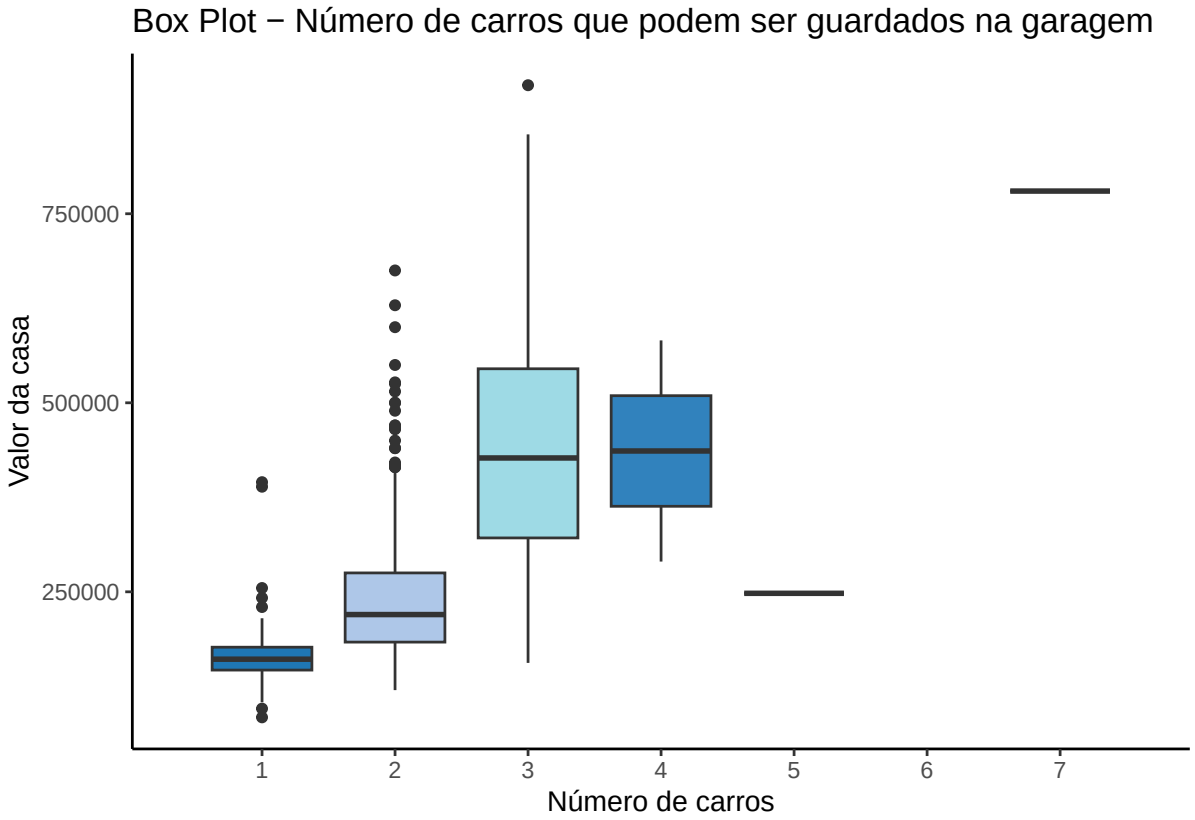


Table 4: Tabela de frequencia do Tamanho da Garagem

	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 < 2.2e-16
```

O tamanho da garagem, sexta variável, mede em número de carros o tamanho da garagem da casa, para casas que não possuem garagem o valor 0 foi imputado, a tabela abaixo é referente aos dados da variável:

Número de carros Quantidade 0 7 1 52 2 351 3 106 4 2 5 1 6 0 7 1

4.1.7 Piscina - X7

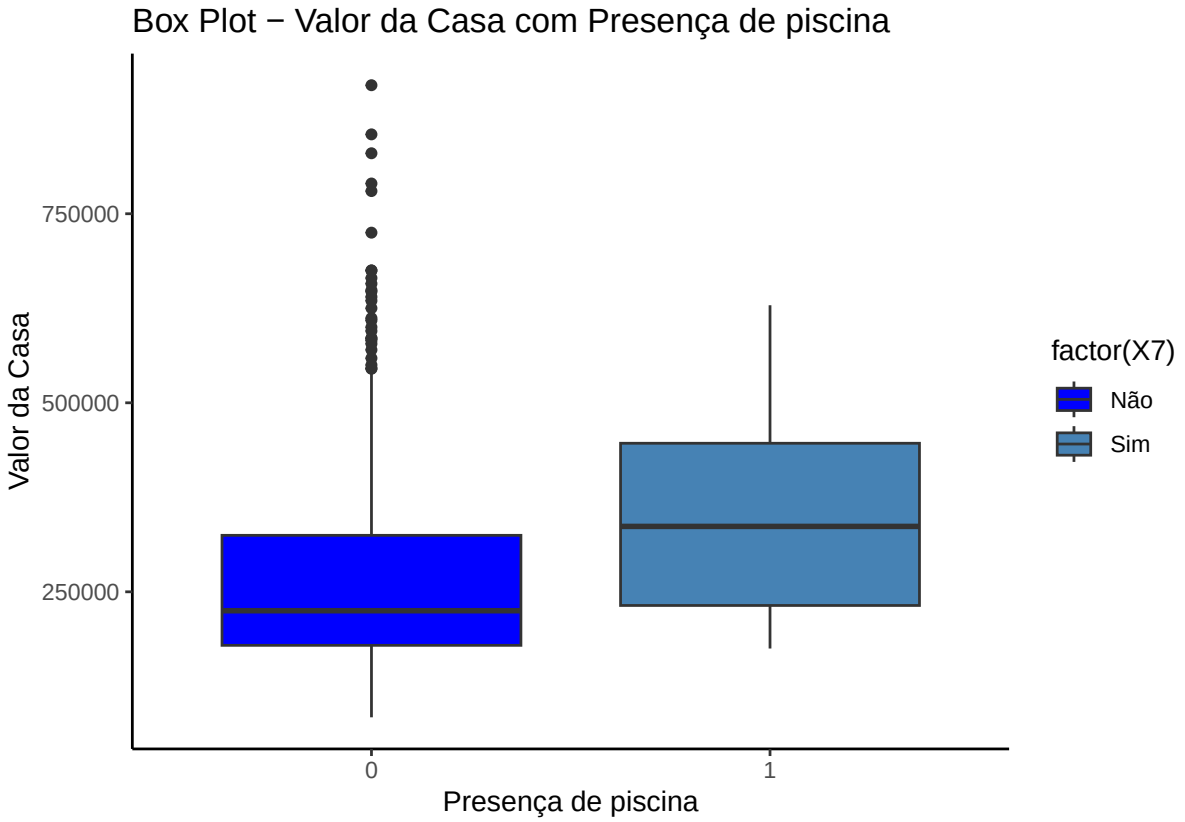
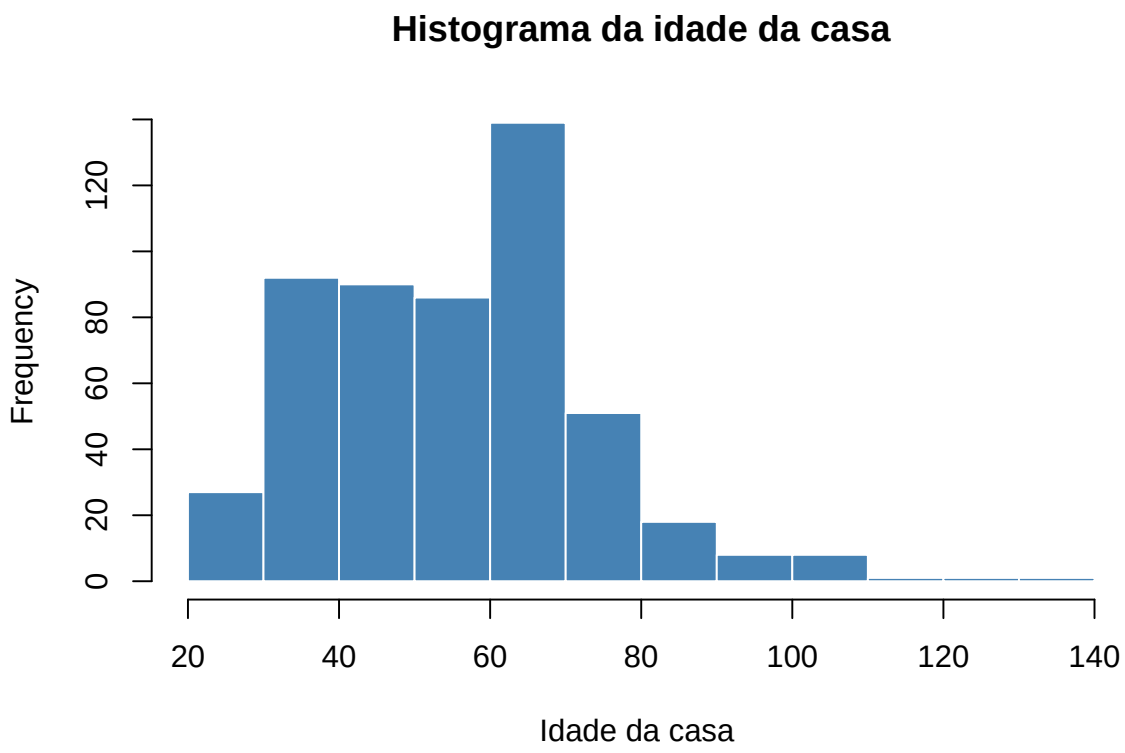


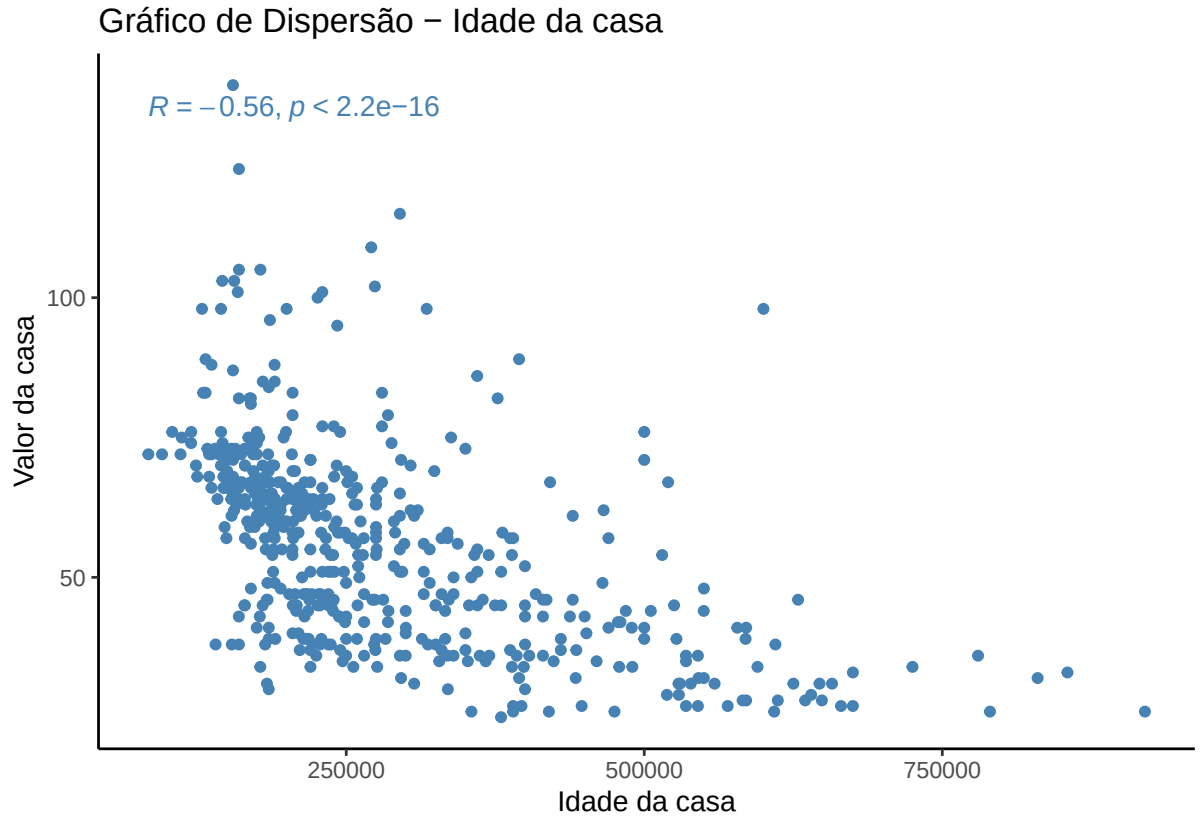
Table 5: Tabela de frequencia da Presença de piscina

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

A sétima variável avalia se a casa possui piscina, sendo essa outra variável qualitativa, com 1 para o caso de haver e 0 para quando não há piscina na residência.

4.1.8 Idade do Imóvel - X8





```
##
##  Shapiro-Wilk normality test
##
## data:  dados$idade
## W = 0.96141, p-value = 1.833e-10
```

A variável idade do imóvel foi retirada com base no dado apresentado de ano de construção, foi usado como base o ano de 2023 para realizarmos a medição dessa variável, nota-se que essa é a variável explicativa com uma grande distribuição de valores, conforme observado no gráfico abaixo:

4.1.9 Qualidade de Construção - X9

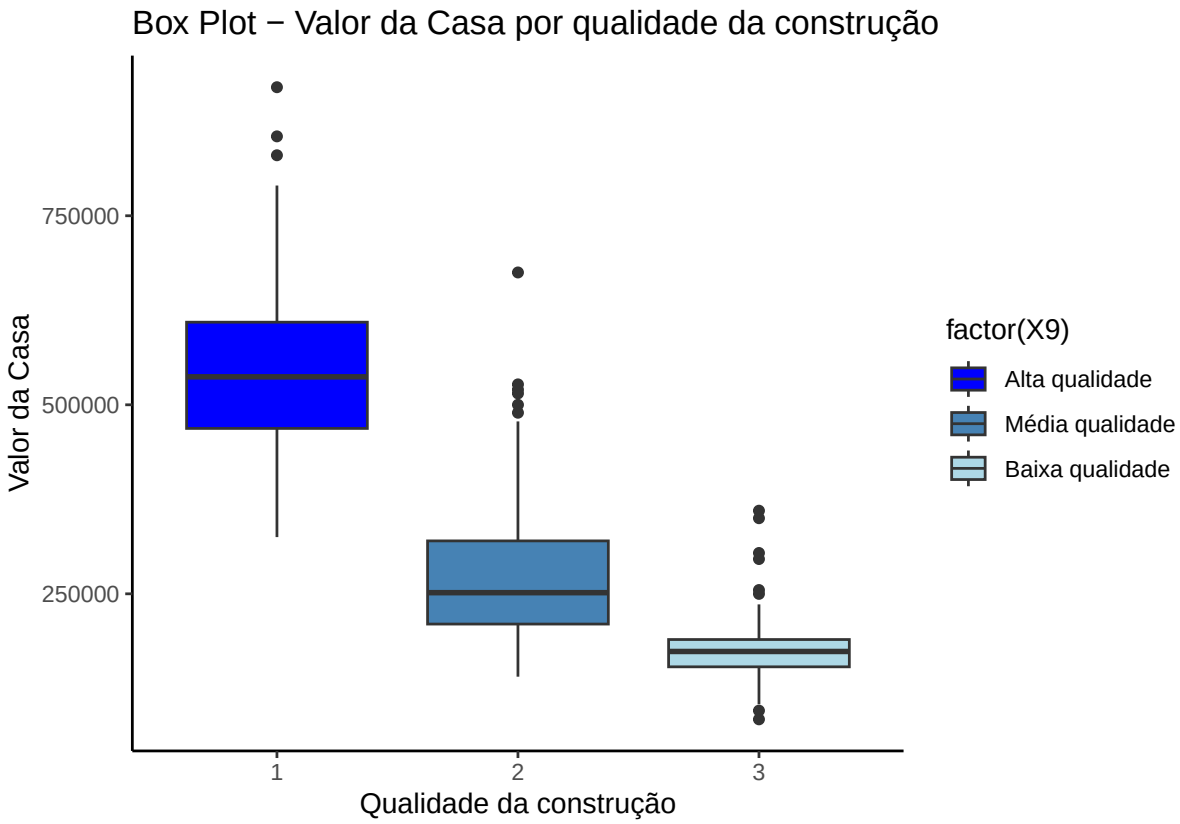
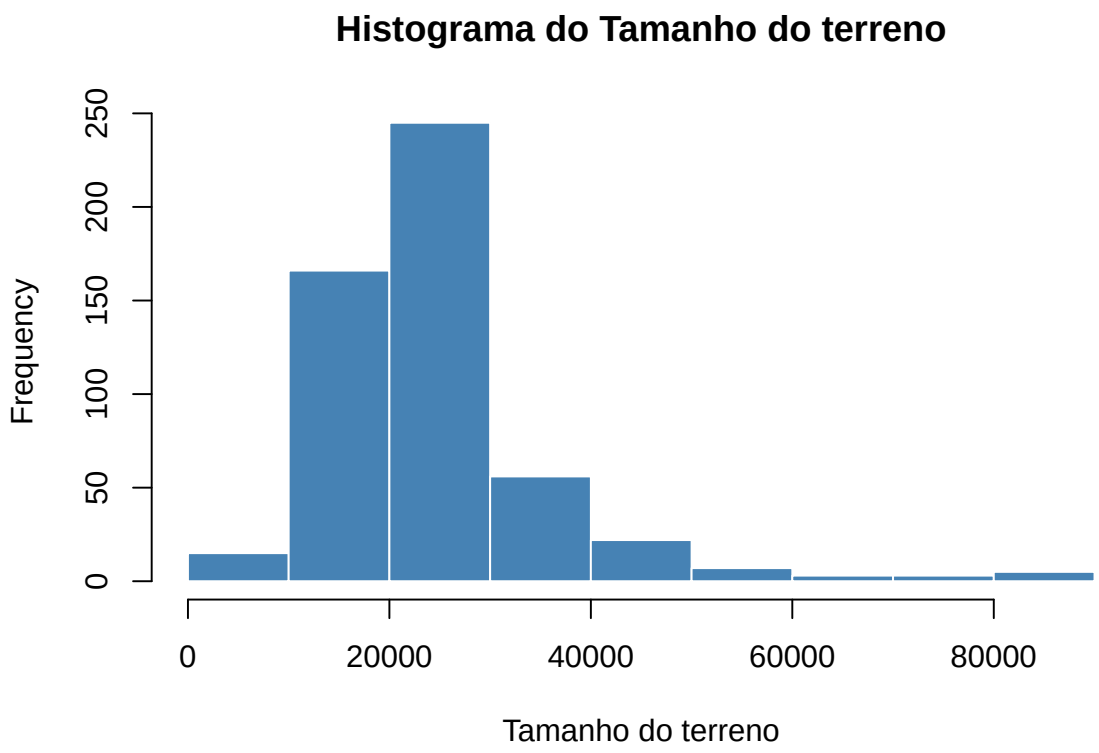


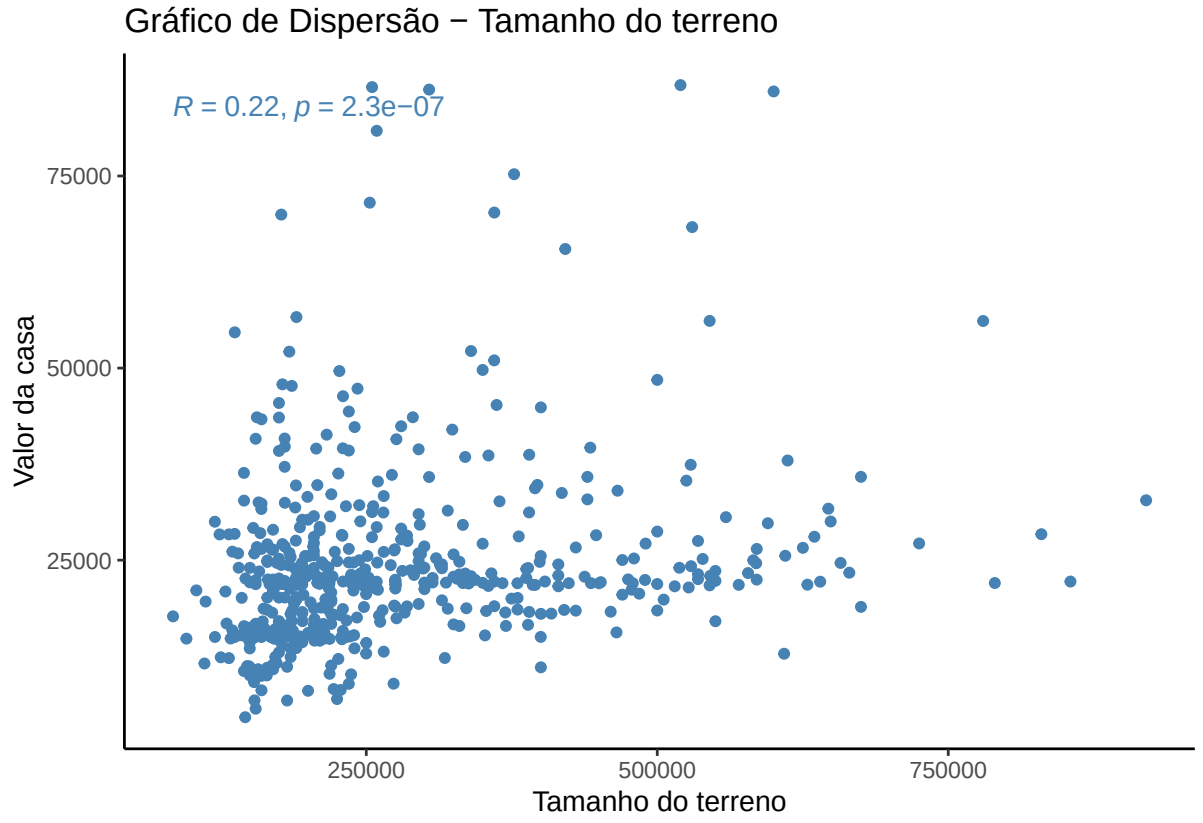
Table 6: Tabela de frequencia da qualidade da construção

	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.

4.1.10 Tamanho do Terreno - X10





```
##
##  Shapiro-Wilk normality test
##
## data:  dados$X10
## W = 0.79804, p-value < 2.2e-16
```

4.1.11 Proximidade da “Highway” - X11

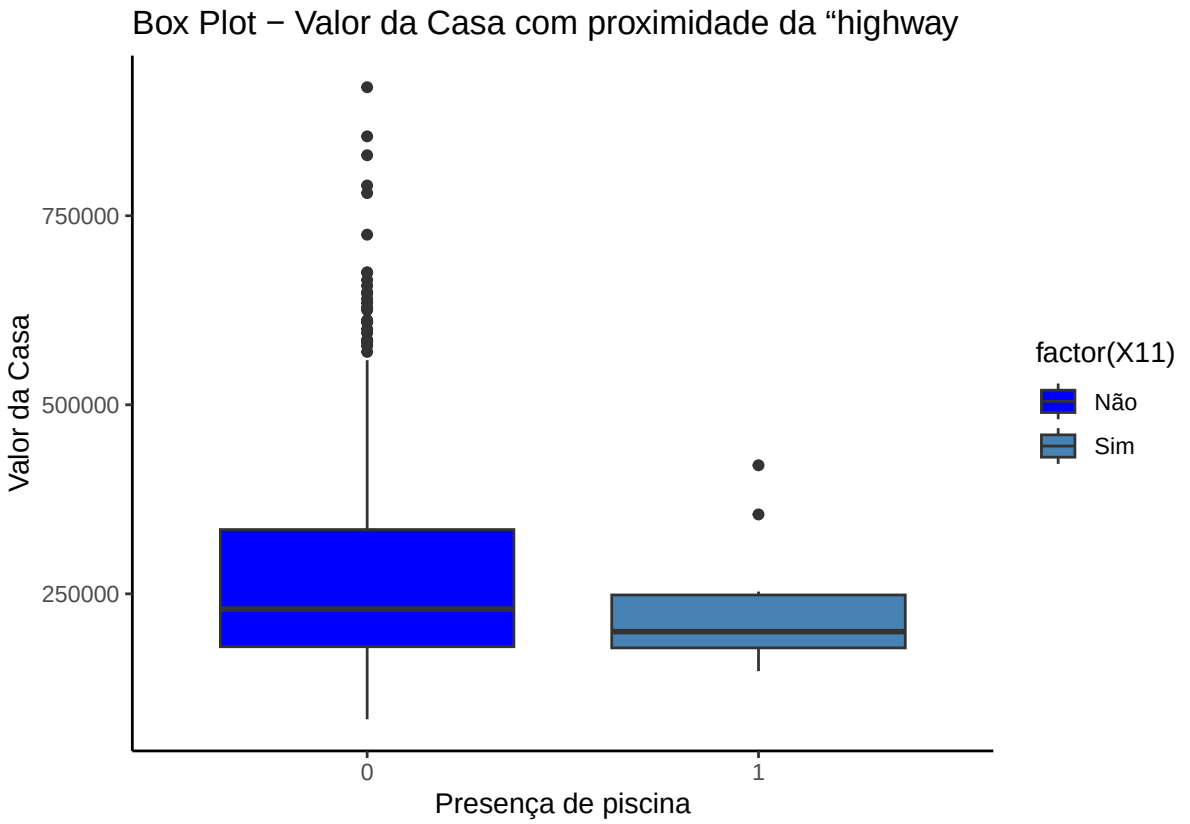


Table 7: Tabela de frequencia da proximidade da highway

	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