FUNDAÇÃO GETULIO VARGAS - FGV

TRABALHO - Previsão com Regressão Múltipla

Ana Paula Pudo
Fábio Monteiro
Lucas Sena Alves
Marcos Soares

MBA Business Analytics e Big Data | Estatística Preditiva

Prof. Abraham Laredo Sicsú

Brasília
2021

Sumário

ÍNDICE DE ILUSTRAÇÕES

GRÁFICOS

1. Introdução

O Condado de Wake está localizado na Carolina do Norte, nos EUA, e foi fundado no ano de 1791. Era o lar dos nativos americanos da tribo Tuscarora, que eram conhecidos à época como cultivadores de cânhamo e usavam a planta para tecer longas vestimentas, calafetar navios e produção de cordas. Após anos de luta, acabaram se unindo aos colonizadores ingleses e estabeleceram os primeiros povos do condado, no final do século XVIII.

A maior cidade do condado é Raleigh, que também é a capital da Carolina do Norte. É a segunda cidade mais populosa do estado, com 467 mil habitantes, de acordo com o censo realizado em 2020. Já o condado de Wake County possui uma população estimada de 823 mil moradores, um crescimento de 31,18% desde 2000.

Um dos motivos que explica o avanço da população é o crescimento econômico da área nos últimos anos, boa parte dele atribuído ao parque industrial localizado entre Raleight e Durham, conhecido como Research Triangle Park. Há uma forte atuação de empresas de tecnologia e biotecnologia, como IBM, GSK e Cisco, mas o condado também conta com companhias agroindustriais atuantes em milho, trigo, algodão e soja.

Com franca ascensão econômica e mais moradores no condado, existe uma expectativa de valorização dos imóveis residenciais na área. Para avaliar tal hipótese, este estudo vai analisar uma base de dados disponível no site do Condado de Wake, com variáveis que consideram o ano de construção dos imóveis, área útil e área total, andares, código postal, banheiros e lareiras, terreno e valor total do imóvel. Com base em 100 amostras retiradas desta base, será realizado um trabalho de previsão, utilizando técnicas e conceitos de Análise Preditiva. Os dados foram extraídos em fevereiro de 2008.

2. Objetivos

2.1 Objetivo Geral

- Analisar as variáveis disponíveis para realizar uma análise preditiva dos preços dos imóveis no Condado de Wake.

2.2 Objetivos Específicos

- Realizar as análises univariada e bivariada das informações disponíveis;
- Verificar a existência de outliers e missing values, adotando medidas para que eles não prejudiquem a análise

- Rodar modelo de regressão múltipla, verificando multicolinearidades, pontos influentes, outliers e detalhar as predições possíveis.

3. Metodologia

Este estudo foi produzido a partir de dados levantados pelo professor Roger Woordard, professor da Universidade North Carolina State, com o apoio do pesquisador Jason Leone, da mesma universidade. Os dados foram disponibilizados pela revista científica Journal of Statistics Education, criada na Carolina do Norte em 1992.

A base de dados deste estudo é composta por 10 variáveis: ano de construção, área útil (originalmente na base de dados como Square Feet), Andares, Área Total (em Acres, no original), Banheiros, Lareiras, Código Postal/CEP, Terreno e Valor Total.

As análises que serão descritas tiveram como variável alvo o valor total do imóvel. A partir do alvo foram realizados os cruzamentos, nas análises com os demais elementos disponibilizados na base, na identificação de elementos estatisticamente significantes na extração de informações.

Como se trata de uma base de dados de origem norte-americana, foram realizados ajustes na base para a utilização das unidades de medidas adotadas no Brasil, com a finalidade de padronização, facilitando assim as análises e simplificando os insights descritos neste estudo. Foram convertidas as unidades métricas de de pés quadrados e de acres na adoção da escala em metros quadrados.

Também foram realizados ajustes globais na base de dados utilizada e localmente das variáveis nas análises das variáveis descritas no tópico 4.

Todos os dados foram trabalhados no RStudio, onde também realizamos a plotagem de gráficos e tabelas. Para calcular todas as variáveis e realizar as análises necessárias, utilizamos códigos de programação da linguagem R.

4. Análises das Variáveis

4.1 Valor Total do Imóvel

A variável originalmente chamada de Total$ é a nossa variável resposta desta análise. O nome desta variável foi traduzido para valor_total. Trata-se de uma variável quantitativa contínua.

4.1.1 Análise Univariada

A maior parte dos imóveis da região estão na média de valor entre USD 100.000 a USD 200.000. Na análise inicial encontramos alguns outlier abaixo de USD 60.000 e acima de USD 300.000, os quais foram removidos para uma possível melhora das análises.

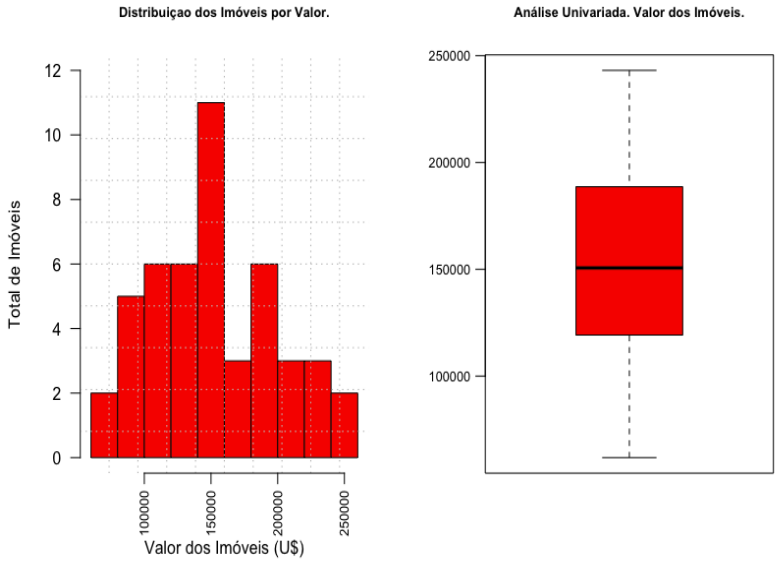

Gráfico 1: Análise Univariada - Valor Total dos Imóveis

4.2 Ano de Construção

A variável Year Built é quantitativa discreta. O nome desta variável foi traduzido para ano_construção.

4.2.1 Análise Univariada

A maior parte dos imóveis da região foi construída entre os anos 1980 e 1990. Os imóveis mais antigos foram erguidos na primeira metade do século passado, mas a grande maioria foi feita após a década de 1950.

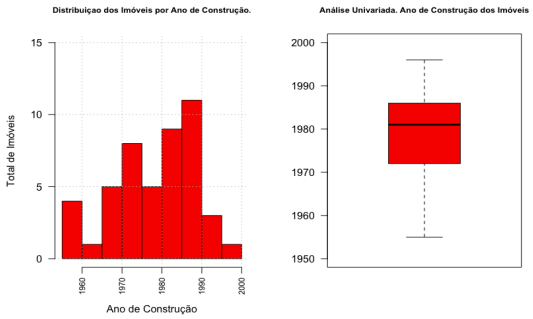

Gráfico 2: Análise Univariada: Ano de Construção

4.2.2 Análise Bivariada

A partir do gráfico 3, observa-se que os pontos estão muito dispersos, não mostrando uma relação linear entre a variável objeto e a variável alvo. É possível verificar imóveis construídos na década de 70 com valores muito próximos a imóveis construídos no final da década de 80.

Essa informação se confirma no resultado do teste de correlação de 37%, ou seja, o ano de construção do imóvel não é uma variável relevante para a definição do valor do imóvel.

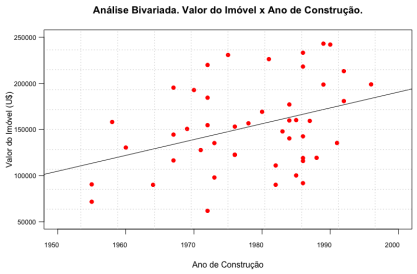

Gráfico 3: Análise Bivariada: Ano de Construção x Valor do Imóvel

4.3 Área Útil/Sqr.Feet

Esta é uma variável quantitativa discreta. Os dados originais estavam identificados como em Square Feets. Foi realizada a conversão para metros quadrados e traduzimos esta variável para área_útil, para melhor compreensão e adequação aos padrões brasileiros.

4.3.1 Análise Univariada

Entre os imóveis analisados, a maior parte tem entre 100 e 170 m² de área útil. Foram removidos os outiliers de imóveis com área útil acima de 250 m².

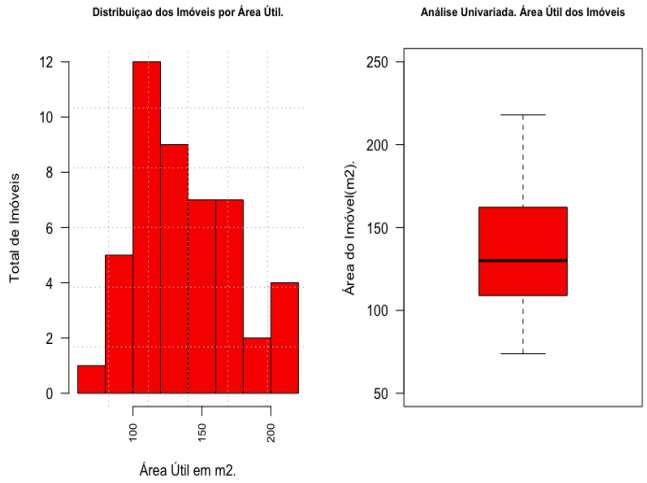

Gráfico 4: Análise Univariada: Área Útil

4.3.2 Análise Bivariada

A partir do gráfico 5, observa-se uma relação linear positiva entre a área útil e o valor total do imóvel. Ou seja, quanto maior a área útil do imóvel, maior seu valor em dólares. Essa informação se confirma no resultado do teste de correlação de 77%.

O gráfico 5, abaixo, traz mais detalhes:

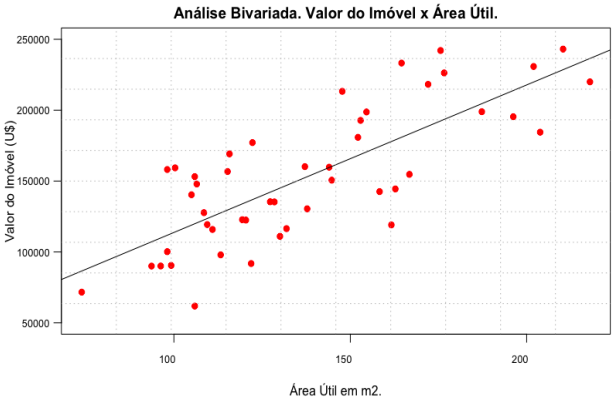

Gráfico 5: Análise Bivariada: Área Útil x Valor do Imóvel

Desta forma, é possível concluir que a área construída dos imóveis na região tem um peso importante para determinar o valor total.

4.4 Número de Andares

A variável originalmente chamada de Story é uma variável qualitativa nominal. Esta variável corresponde ao total de andares e possui 5 níveis de classificação, com valores oscilando entre 1, 1.5, 1.75, 2 e 2.5. A variável story foi traduzida para total_andares.

Comentado [FM1]: Não seria ordinal?

Comentado [UdMO2R1]: Não professor respondeu que é quali

4.4.1 Análise Univariada

Para melhor interpretação, os dados na figura abaixo apresentam os resultados em porcentagem. A maior concentração de imóveis na distribuição possui entre 1 e 2 andares. Podemos observar que:

- ~60% dos imóveis possuem 1 andar
- ~228% dos imóveis possuem 2 andares

Os valores fracionados da classificação dos elementos registrados (1.5; 1.75 e 2.5) representam menos de 20% do volume total do observado.

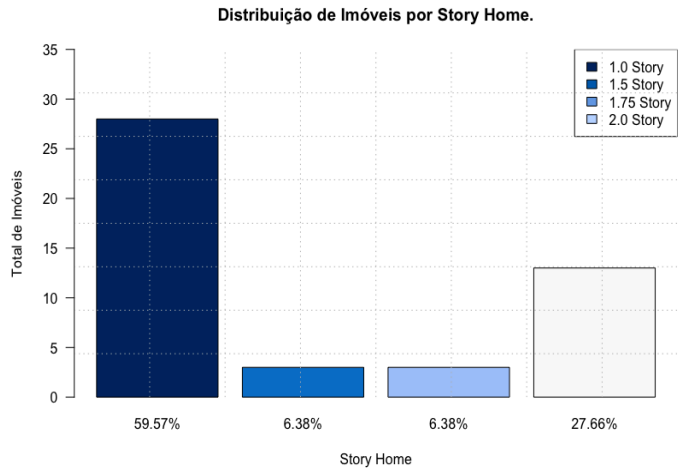

Gráfico 6: Análise Univariada: Número de Andares

4.4.2 Análise Bivariada

Na análise entre os andares dos imóveis e os valores totais, é notável que a mediana indica uma possibilidade de relação entre tais variáveis.

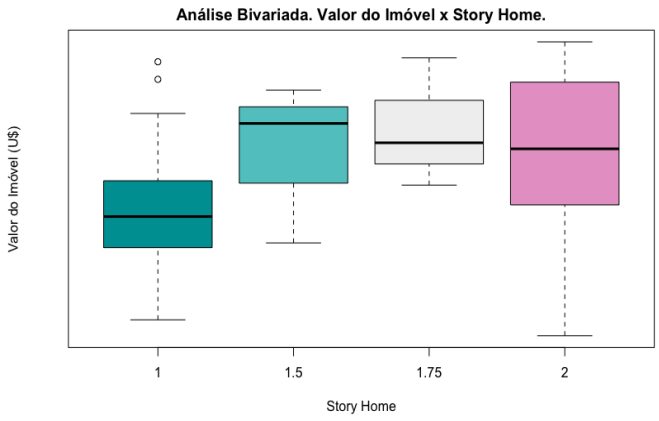

Gráfico 7: Análise Bivariada: Número de Andares (Story Home) x Valor do Imóvel

4.5 Área Total

Originalmente denominada de Acres na base de dados, esta é uma variável quantitativa discreta. Mantendo o critério adotado em variáveis anteriores, foi realizada a conversão para o sistema métrico brasileiro, em metros quadrados, e foi considerado que esta variável se refere à área total do terreno do imóvel.

4.5.1 Análise Univariada

A partir do gráfico 8, é possível verificar que a maior concentração de imóveis está em terrenos entre 700 m² a 1400 m² de tamanho registrado. No tratamento dos dados não foram excluídas 3 observações que possuem área de 0 m², porém outliers com área total acima de 2000 m² foram removidos.

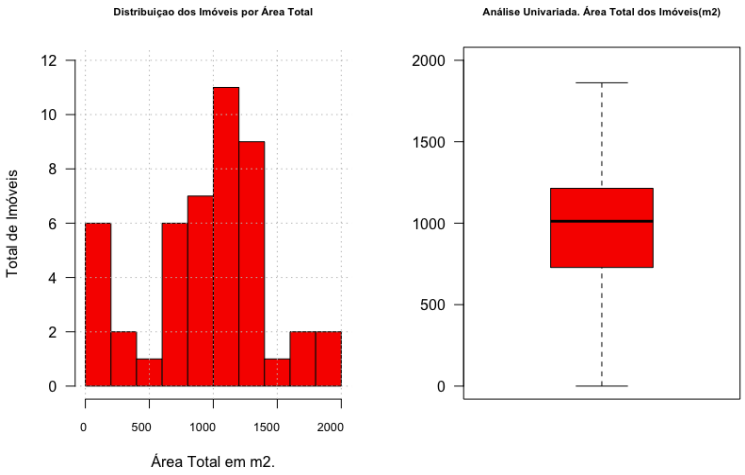

Gráfico 8: Análise Univariada Área Total

4.5.2 Análise Bivariada

Os dados aparecem bem dispersos no gráfico 9. Também é possível verificar os mesmos valores com área total muito distintas.

Essa informação se confirma no resultado do teste de correlação de 16%. Isso significa que a área total dos imóveis na região não tem um peso importante para determinar o valor total de cada imóvel analisado.

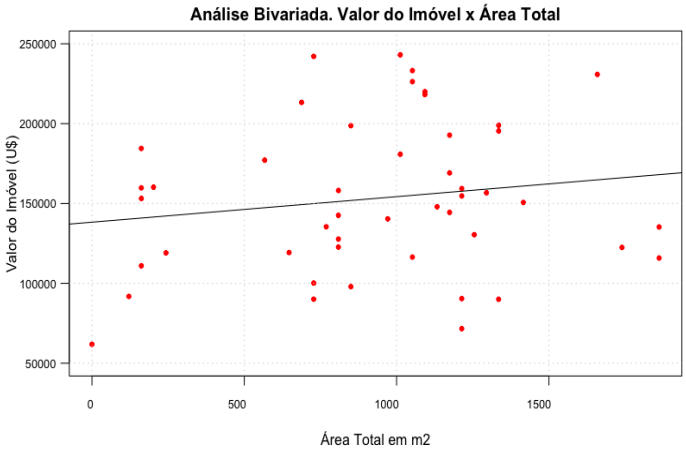

Gráfico 9: Análise Bivariada: Área Total x Valor do Imóvel

4.6 Total de Banheiros

Esta é uma variável quantitativa discreta. A base de dados trouxe valores de 1, 1.5, 2, 2.5 e 3 banheiros. O nome da variável foi traduzido para Número de Banheiros.

Comentado [FM3]: Qualitativa nominal ou quantitativa discreta?

4.6.1 Análise Univariada

Como é possível verificar a seguir, a maior parte dos imóveis possui 2 ou mais banheiros. Analisando os dados temos o seguinte:

- ~38% dos imóveis possuem 2.5 banheiros
- ~36% dos imóveis possuem 2 banheiros.

Foi excluída uma observação com missing value.

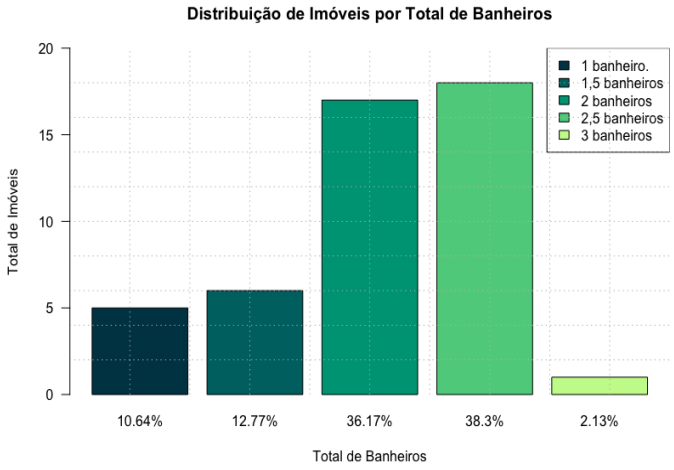

Gráfico 10: Análise Univariada - Número de Banheiros

4.6.2 Análise Bivariada

Nota-se uma inclinação positiva entre o total de banheiros dos imóveis analisados e seus respectivos valores. O resultado do teste de correlação apresentou um resultado de 63 %, o que não é necessariamente tão forte, mas o cruzamento desta variável com a alvo apresenta uma relação linear.

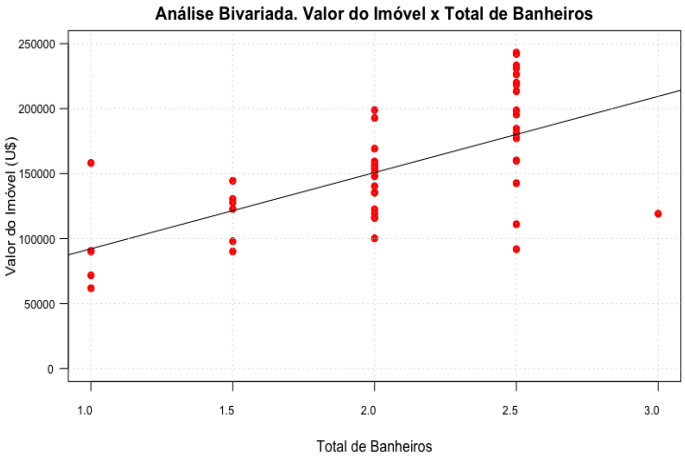

Gráfico 11: Análise Bivariada: Total de banheiros x Valor do Imóvel

4.7 Número de Lareiras

Item que não é muito comum no Brasil, as lareiras podem ser encontradas em várias residências dos EUA, dependendo da região do imóvel. Na base de dados original, esta variável chama-se Fireplaces e é uma qualitativa nominal. O nome da variável foi traduzido para total_lareiras.

Comentado [FM4]: Idem banheiros. Quantitativa nominal ou Qualitativa nominal?

4.7.1 Análise Univariada

Aproximadamente 69% dos imóveis possuem uma lareira. Apenas um imóvel foi considerado excluído como outlier, pois possuía duas lareiras.

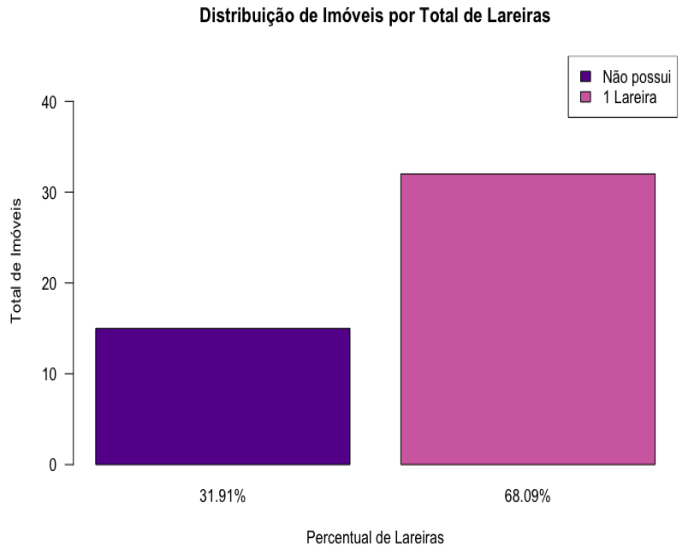

Gráfico 12: Análise Univariada - Número de Lareiras

4.7.2 Análise Bivariada

Na análise entre o número de lareiras e o valor do imóvel, é possível observar que imóveis com lareira nesta região são consideravelmente mais valorizados. Por consequência, este item tende a estar presente nos imóveis com valores mais altos.

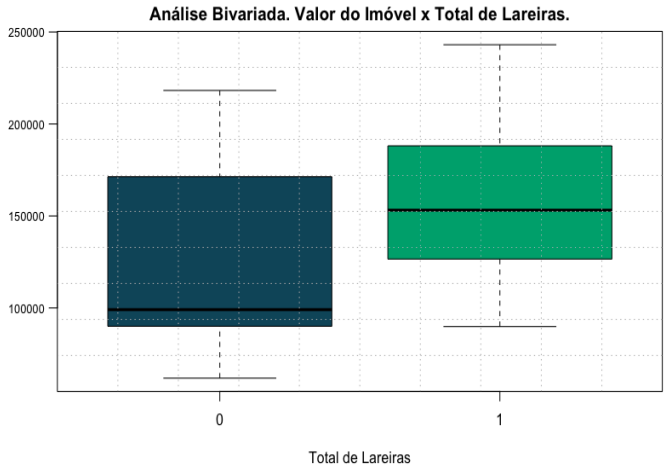

Gráfico 13: Análise Bivariada: Total de lareiras x Valor do Imóvel

4.8 CEP/Código Postal

Esta variável estava identificada na base de dados original com o nome Zip, que é o equivalente a código postal/CEP. Trata-se de uma variável qualitativa ordinal. Para as análises, foram utilizados apenas os 3º e 4º dígitos do código de identificação.

4.8.1 Análise Univariada

A maior concentração de imóveis verificada na base está entre os códigos 60 e 61.

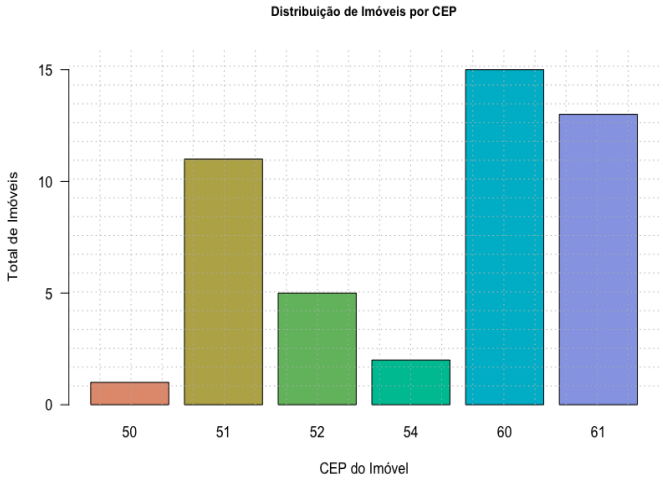

Gráfico 14: Análise Univariada: Imóveis por CEP

4.8.2 Análise Bivariada

Apesar da maior concentração entre os identificadores 60 e 61, a média de imóveis com valores mais altos está concentrada dígito 50, conforme gráfico 14 a seguir:

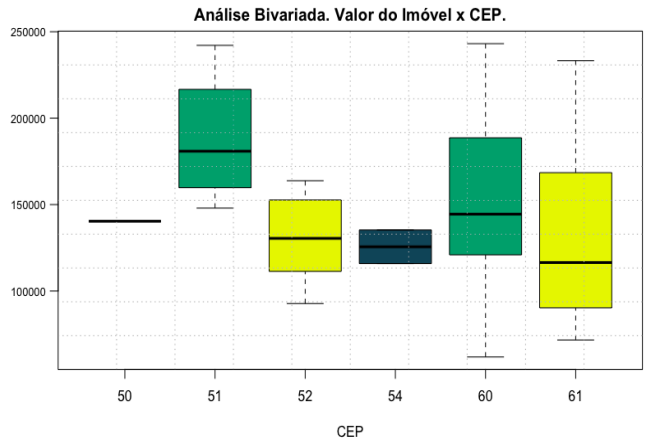

Gráfico 15: Análise Bivariada: Imóveis por CEP x Valor do Imóvel

4.9 Valor do Terreno

Esta é a variável do valor do terreno. Na base de dados original foi identificada como Land $, para especificar que trata-se de valores em dólares. Esta é uma variável quantitativa contínua.O nome da variável foi traduzida para valor_terreno e a moeda foi mantida.

Comentado [FM5]: Não seria quantitativa contínua? No slide do João ele cita preço nessa categoria.

4.9.1 Análise Univariada

A maior parte dos terrenos verificados valem cerca de 40 mil dólares, mas também foram verificados espaços com valor próximo aos 60 mil dólares. Outiliers com valores acima 100 mil dólares foram removidos.

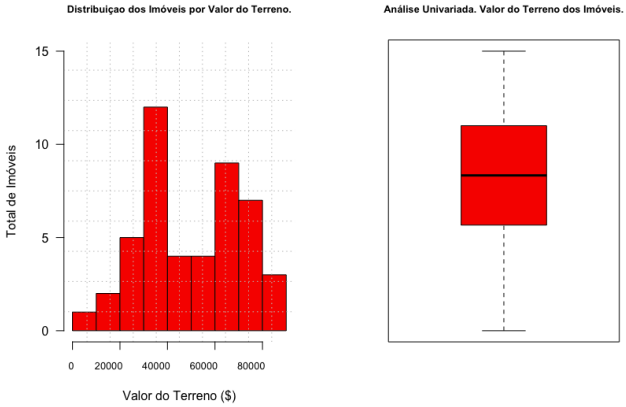

Gráfico 16: Análise Univariada - Valor dos Terrenos

4.9.2 Análise Bivariada

O valor dos terrenos impacta diretamente no valor total do imóvel, conforme é possível verificar no gráfico 16, de dispersão.

Essa informação se confirma no resultado do teste de correlação, que registrou 89%. Assim, o valor total dos terrenos na região tem um peso importante para determinar o valor total de cada imóvel analisado.

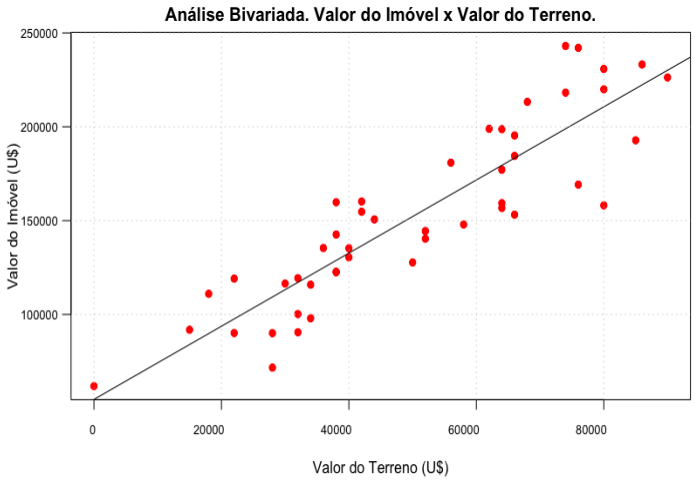

Gráfico 17: Análise Bivariada: Valor dos Terrenos x Valor do Imóvel

5. Modelo Preditivo

5.1 Regressão Múltipla

O percentual do p-valor adotado para análise dos modelos é de 15%.
A partir do modelo de equação de regressão da variável resposta sobre as demais variáveis (lm (formula = valor_total ~ ., data = data_wood), os resultados apresentaram variáveis como ano_construção, area_util, total_andares, total_banheiros e total de lareiras com p-valor abaixo de 15%, o que significa que são variáveis provavelmente selecionáveis para a formação do modelo que defina o valor total do imóvel. Nesta etapa o R2 foi de 0,976.

```
Coefficients:
                     Estimate     Std. Error t value
Pr(>|t|)
(Intercept)     -1741941.46868   374811.42350  -4.648
0.0000489700 ***
ano_construcao       881.90844      191.82620   4.597
0.0000567816 ***
area_util            498.54389       66.70463   7.474
0.0000000113 ***
total_andares      13842.58238     4685.98506   2.954
0.00566 **
area_total             1.57028        5.28256   0.297
0.76808
total_banheiros    -9342.94811     5929.65624  -1.576
0.12437
total_lareiras      7817.36040     3681.94239   2.123
0.04110 *
cep51               4829.75730     9133.46570   0.529
0.60038
cep52               5613.34750     9946.60413   0.564
0.57622
cep54               2297.83825    11726.94541   0.196
0.84582
cep60              -2015.50843     9492.47744  -0.212
0.83312
cep61               1595.51950     9276.88894   0.172
0.86447
valor_terreno          1.45706        0.08352  17.446 <
0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8549 on 34 degrees of freedom
Multiple R-squared:  0.976,   Adjusted R-squared:  0.9676
F-statistic: 115.4 on 12 and 34 DF,  p-value: < 0.00000000000000022
```

5.2 Seleção das Variáveis - Critério AKAIKE

A equação selecionada para o modelo foi determinada pela fórmula, lm(formula = valor_total ~ ano_construcao + area_util + total_andares + area_total + total_lareiras + valor_terreno, data = data_wood), gerando os seguintes resultado no R:

quote
```
Residuals:
   Min     1Q Median     3Q    Max
-16772  -4977   1322   4717  18208

Coefficients:
                   Estimate    Std. Error t value
Pr(>|t|)
(Intercept)    -1362773.9167   281074.2882  -4.848
0.0000191671393 ***
ano_construcao      685.2038      143.0531   4.790
0.0000230734782 ***
area_util           431.8322       46.9673   9.194
0.0000000000206 ***
total_andares     13161.9637     3956.0142   3.327
0.00189 **
area_total            5.2145        3.5223   1.480
0.14660
total_lareiras     7405.3703     2942.4859   2.517
0.01596 *
valor_terreno         1.4473        0.0703  20.586 <
0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8409 on 40 degrees of freedom
Multiple R-squared:  0.9727,  Adjusted R-squared:  0.9686
F-statistic: 237.8 on 6 and 40 DF,  p-value: < 0.00000000000000022
```
unquote

O valor estimado gerado pelo R significa que, para cada ano de construção, o valor do imóvel aumenta 685,20 dólares a cada m² em área útil, o valor do imóvel aumenta 431,83 dólares e assim sucessivamente para cada variável. Cabe ressaltar que cada variável possui o seu referencial, portanto, é importante tomar cuidado para não comparar proporções distintas.

O R2 apresentado é de 0,9727.

5.3 Pontos Influentes

A partir dos gráficos de diagnóstico 17 e 18 encontramos no cook's distance a observação 38 como ponto influente. Este ponto aparece com um erro nos totais das áreas, ou seja, a área total está menor que a área útil, o que pode ter sido a causa do aparecimento deste ponto.

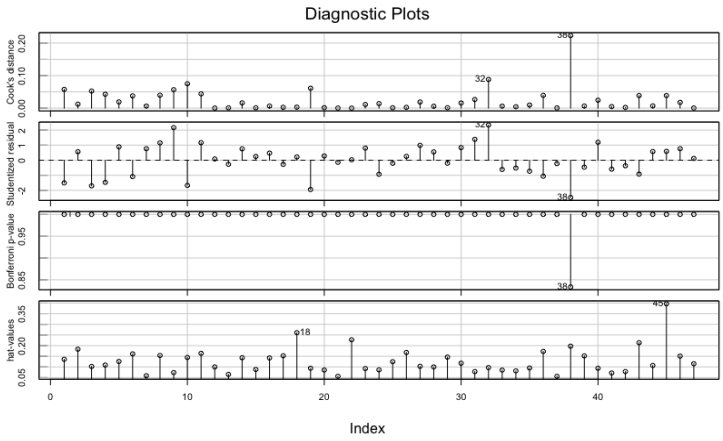

Gráfico 18: Gráfico de Análises de Pontos Influentes 1

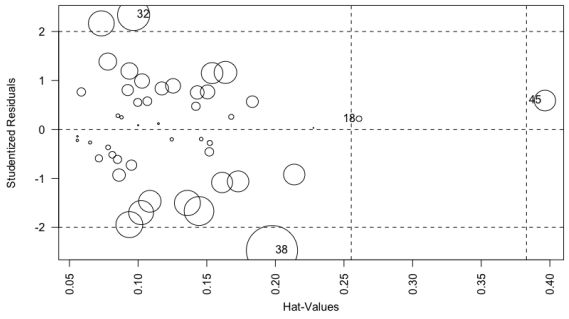

Gráfico 19: Gráfico de Análises de Pontos Influentes 2

5.4 Análise do poder preditivo do modelo - Erros percentuais

A partir das variáveis disponíveis para esta análise, foi possível chegar em uma estimativa de valor total do imóvel com margem de erro entre -11% e 5%. Ou seja, o preço encontrado pelo modelo está até 11% mais barato e até 5% mais caro que o valor alvo.

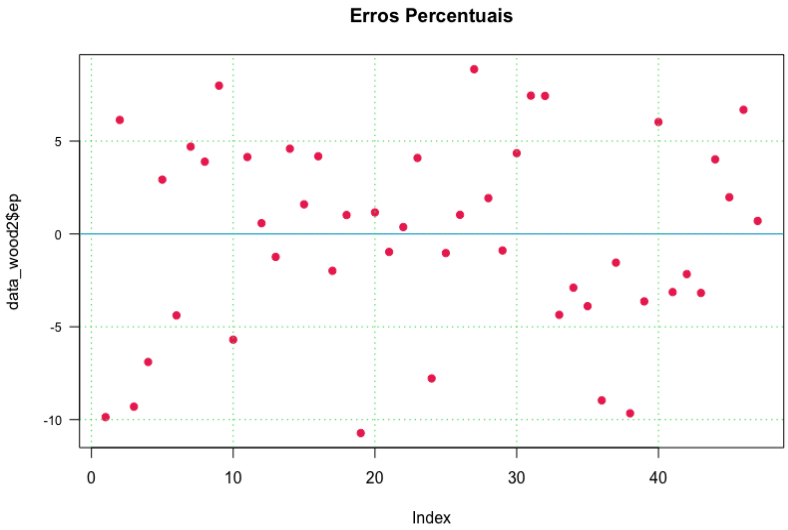

Gráfico 20: Gráfico de Erros Percentuais

5.5 Análise de Multicolinearidade

Está análise não apresentou evidências de multicolinearidade, **considerando que nenhuma** variável apresenta número >5.

ano construção	1.390696
área útil	1.722907
total andares	1.467462
total lareiras	1.242806
valor terreno	1.44306

6. Conclusões

Este estudo foi dividido em duas partes principais: análise exploratória de uma base de dados de imóveis no Condado de Wake, na Carolina do Norte (EUA), e a construção de um modelo preditivo baseado nos mesmos dados.

O objetivo principal foi criar um modelo preditivo capaz de estimar o valor total dos imóveis a partir das variáveis disponibilizadas, ainda que algumas delas precisaram de ajustes e tratamento.

Foi necessário fazer um tratamento cuidadoso na base de dados e também rodar varias vezes o modelo para tentar diminuir a margem de erro. Entretanto, foi possível chegar a uma margem de erro entre -10 e 5, o qual entendemos que ficou dentro dos limites aceitáveis conforme a base de dados fornecida.