

Categorização do Conjunto de Dados "Venda de Casas em King County"

UNICEUB - Luiz Felipe Nast Roballo

introdução

- Escritório de Avaliação de Propriedades - arrecadação de impostos e contabilidade
- Portal King County GIS Center

The screenshot shows the King County GIS Center website. At the top is a navigation bar with the King County logo, a search bar, and links for Home, How do I..., Services, About King County, and Departments. A large map of King County is the background for the main content area. On the left, a sidebar lists services: Maps & Apps, Data Hub, Spatial Services, Blog: GIS & You, Property Research, Contact Us, Training, About us, and For KC Employees. The main content area features a heading "King County GIS Center" and a sub-headline "We put geographic information to work for King County and beyond". A green box contains a COVID-19 update: "King County GIS Center and COVID-19" and a note about staff telecommuting. Below this is a paragraph about the GIS Center's mission and a "Events" section with a "Full Calendar" link.

The screenshot shows a Kaggle dataset page titled "House Sales in King County, USA". The page includes a thumbnail of a house and coins, a download button ("Download (798 kB)", "2005", "New Notebook"), and a "Data Card" tab. The main content area has a heading "House Sales in King County, USA" and a subtitle "Predict house price using regression". It features sections for "About Dataset" (describing the dataset as containing house sale prices from May 2014 to May 2015), "Data Card" (with 1208 code snippets and 28 discussions), and "Code" (with 1208 snippets). On the right, there are "Usability" (7.06), "License" (CC0: Public Domain), "Expected update frequency" (Not specified), and "Tags" (Finance).

ID	IDENTIFICAÇÃO
DATE	DATA DA VENDA
PRICE	PREÇO DE VENDA
BEDROOMS	NÚMERO DE QUARTOS
BATHROOMS	NÚMERO DE BANHEIROS
SQFT_LIV	TAMANHO DO ESPAÇO INTERNO DA HABITAÇÃO EM METROS QUADRADOS
SQFT_LOT	TAMANHO DO LOTE EM METROS QUADRADOS
FLOORS	NÚMERO DE ANDARES
WATERFRONT	'1' SE A PROPRIEDADE TIVER BEIRA-MAR, '0' SE NÃO.
VIEW	UM ÍNDICE DE 0 A 4 DE QUÃO BOA ERA A VISTA DO IMÓVEL
CONDITION	CONDICÃO DA CASA, CLASSIFICADA DE 1 A 5
GRADE	CLASSIFICAÇÃO PELA QUALIDADE DA CONSTRUÇÃO
SQFT_ABOVE	PÉS QUADRADOS ACIMA DO SOLO
SQFT_BASMT	PÉS QUADRADOS ABAIXO DO SOLO
YR_BUILT	ANO EM QUE A PROPRIEDADE FOI CONSTRUÍDA
YR_RENOV	ANO EM QUE A PROPRIEDADE FOI RENOVADA
ZIPCODE	OS 5 PRIMEIROS DÍGITOS DO CÓDIGO POSTAL
LAT	LATITUDE
LONG	LONGITUDE
SQFT_LIV15	TAMANHO MÉDIO DO ESPAÇO INTERNO DA HABITAÇÃO PARA AS 15 CASAS MAIS PRÓXIMAS, EM PÉS QUADRADOS
SQFT_LOT15	TAMANHO MÉDIO DOS TERRENOS PARA AS 15 CASAS MAIS PRÓXIMAS, EM METROS QUADRADOS

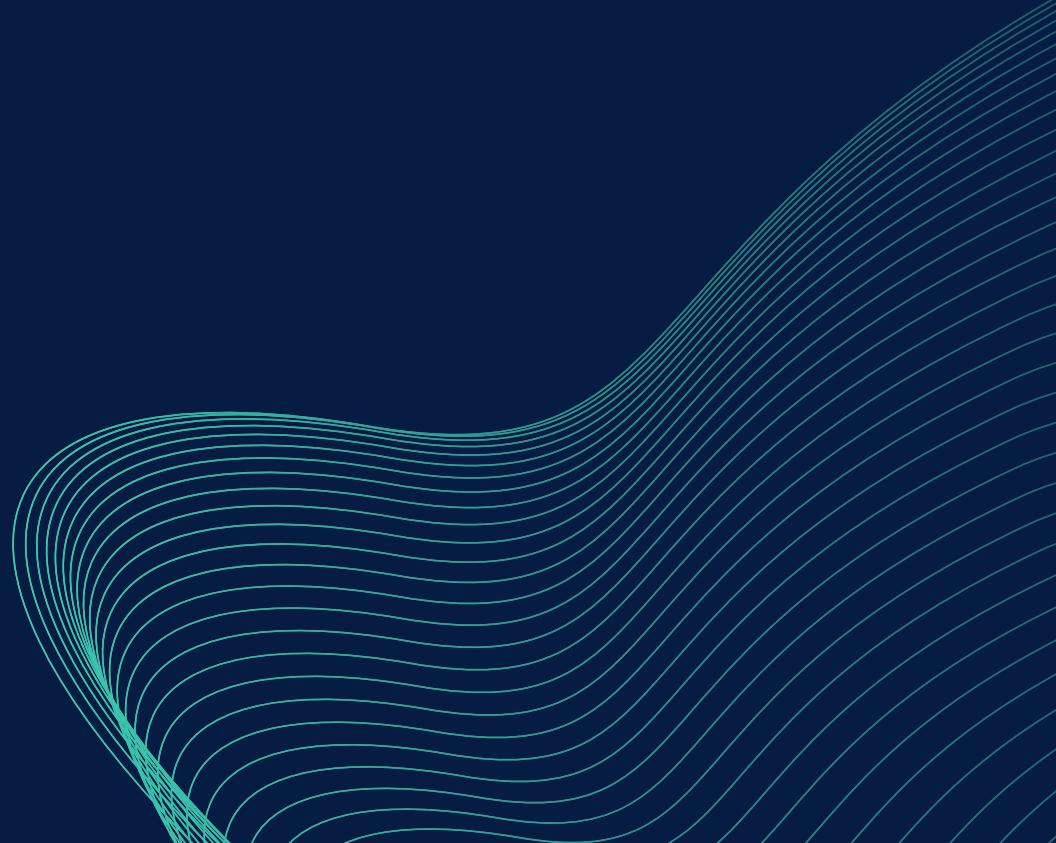
21 colunas X 21.613 registros

- 5 variáveis contínuas (price, sqft_liv, sqft_lot, sqft_above, sqft_basmt)
- 7 variáveis discretas (bedrooms, floors, lat, long, sqft_liv15, sqft_lot15, bathrooms)
- 5 variáveis ordinais (date, waterfront, view, condition, grade)
- 4 variáveis nominais (id, yr_built, yr_renov, zipcode)

Objetivo

ID	IDENTIFICAÇÃO
DATE	DATA DA VENDA
PRICE	PREÇO DE VENDA
BEDROOMS	NÚMERO DE QUARTOS
BATHROOMS	NÚMERO DE BANHEIROS
SQFT_LIV	TAMANHO DO ESPAÇO INTERNO DA HABITAÇÃO EM METROS QUADRADOS
SQFT_LOT	TAMANHO DO LOTE EM METROS QUADRADOS
FLOORS	NÚMERO DE ANDARES
WATERFRONT	'1' SE A PROPRIEDADE TIVER BEIRA-MAR, '0' SE NÃO.
VIEW	UM ÍNDICE DE 0 A 4 DE QUÃO BOA ERA A VISTA DO IMÓVEL
CONDITION	CONDICÃO DA CASA, CLASSIFICADA DE 1 A 5
GRADE	CLASSIFICAÇÃO PELA QUALIDADE DA CONSTRUÇÃO
SQFT_ABOVE	PÉS QUADRADOS ACIMA DO SOLO
SQFT_BASMT	PÉS QUADRADOS ABAIXO DO SOLO
YR_BUILT	ANO EM QUE A PROPRIEDADE FOI CONSTRUÍDA
YR_RENOV	ANO EM QUE A PROPRIEDADE FOI RENOVADA
ZIPCODE	OS 5 PRIMEIROS DÍGITOS DO CÓDIGO POSTAL
LAT	LATITUDE
LONG	LONGITUDE
SQUFT_LIV15	TAMANHO MÉDIO DO ESPAÇO INTERNO DA HABITAÇÃO PARA AS 15 CASAS MAIS PRÓXIMAS, EM PÉS QUADRADOS
SQUFT_LOT15	TAMANHO MÉDIO DOS TERRENOS PARA AS 15 CASAS MAIS PRÓXIMAS, EM METROS QUADRADOS

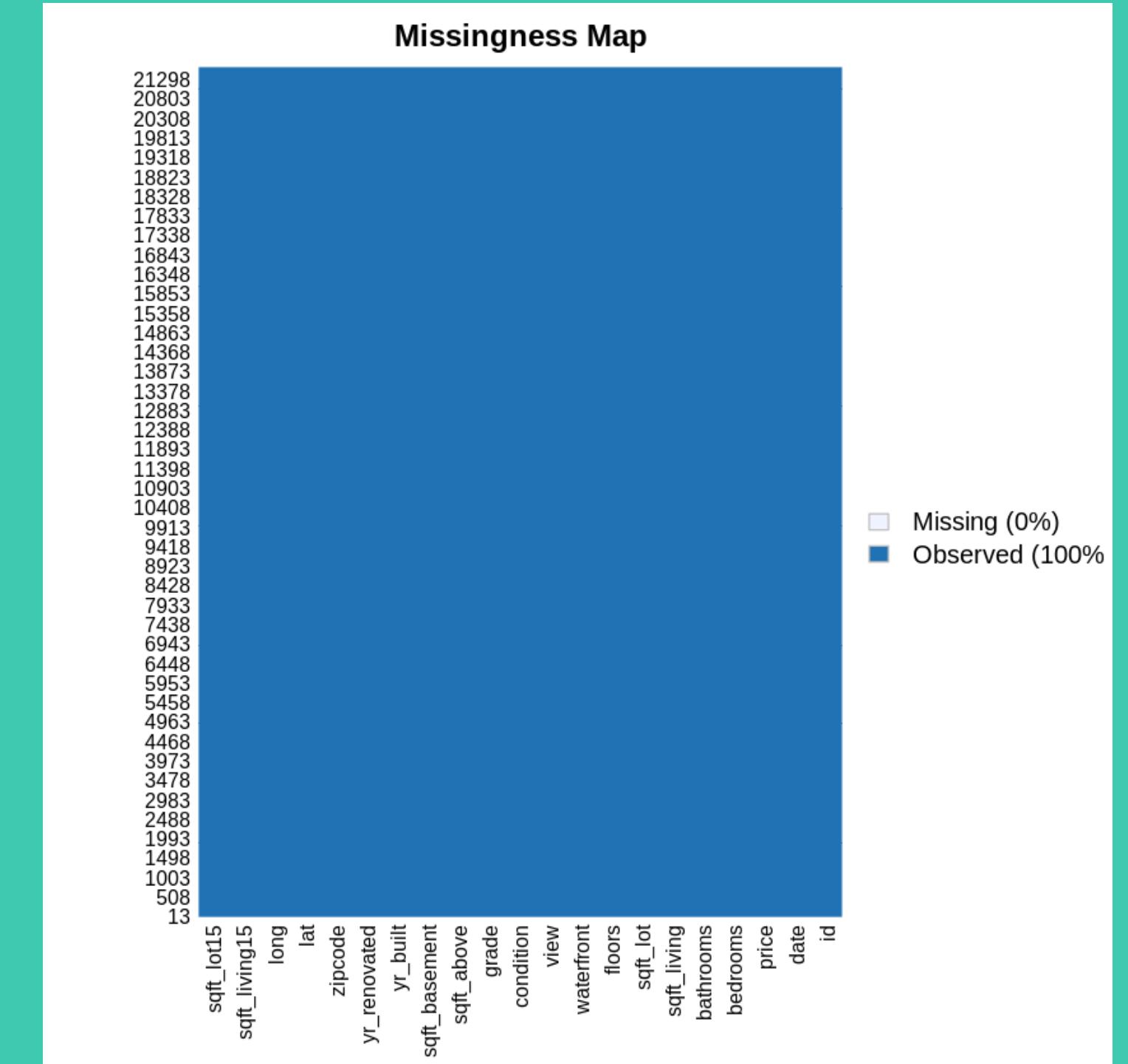
- Características X Valor da propriedade
- Variáveis independentes X Variáveis dependentes
- Modelo de previsão - boa opção de análise



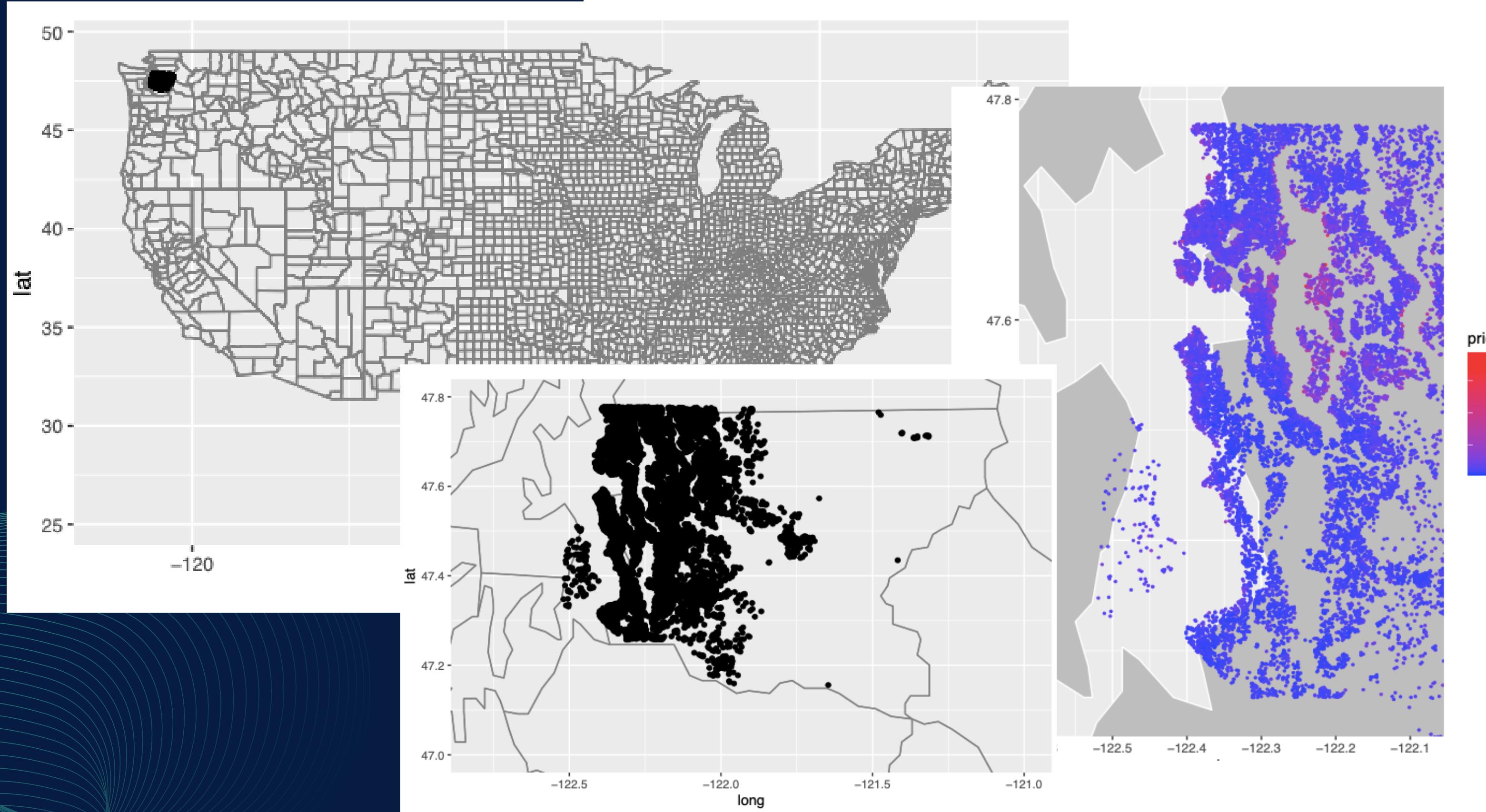
Integridade e Estruturação dos Dados

Rows: 21,613
Columns: 21

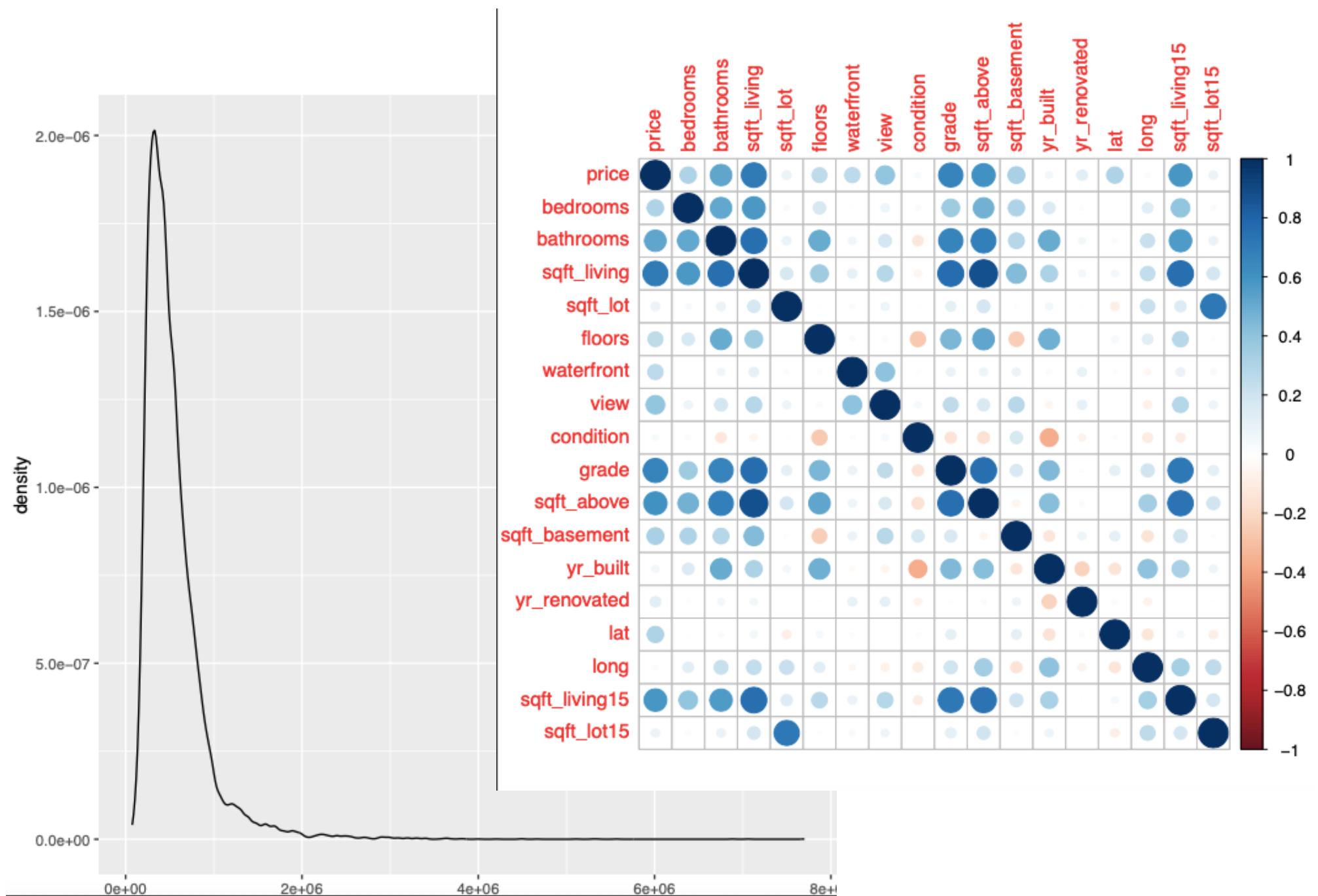
```
$ id <dbl> 7129300520, 6414100192, 5631500400, 2487200875, 19544005...
$ date <chr> "20141013T000000", "20141209T000000", "20150225T000000", ...
$ price <dbl> 221900, 538000, 180000, 604000, 510000, 1225000, 257500, ...
$ bedrooms <int> 3, 3, 2, 4, 3, 4, 3, 3, 3, 3, 2, 3, 3, 5, 4, 3, 4, 2, ...
$ bathrooms <dbl> 1.00, 2.25, 1.00, 3.00, 2.00, 4.50, 2.25, 1.50, 1.00, 2.00, ...
$ sqft_living <int> 1180, 2570, 770, 1960, 1680, 5420, 1715, 1060, 1780, 1890, ...
$ sqft_lot <int> 5650, 7242, 10000, 5000, 8080, 101930, 6819, 9711, 7470, ...
$ floors <dbl> 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 2.0, 1.0, 1.0, ...
$ waterfront <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ view <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ condition <int> 3, 3, 3, 5, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3, 3, 3, 4, 4, ...
$ grade <int> 7, 7, 6, 7, 8, 11, 7, 7, 7, 7, 8, 7, 7, 7, 7, 9, 7, 7, 7, ...
$ sqft_above <int> 1180, 2170, 770, 1050, 1680, 3890, 1715, 1060, 1050, 1890, ...
$ sqft_basement <int> 0, 400, 0, 910, 0, 1530, 0, 0, 730, 0, 1700, 300, 0, 0, ...
$ yr_built <int> 1955, 1951, 1933, 1965, 1987, 2001, 1995, 1963, 1960, 2000, ...
$ yr_renovated <int> 0, 1991, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ zipcode <int> 98178, 98125, 98028, 98136, 98074, 98053, 98003, 98198, ...
$ lat <dbl> 47.5112, 47.7210, 47.7379, 47.5208, 47.6168, 47.6561, 47.6000, ...
$ long <dbl> -122.257, -122.319, -122.233, -122.393, -122.045, -122.000, ...
$ sqft_living15 <int> 1340, 1690, 2720, 1360, 1800, 4760, 2238, 1650, 1780, 2300, ...
$ sqft_lot15 <int> 5650, 7639, 8062, 5000, 7503, 101930, 6819, 9711, 8113, ...
```



Conceitualização Visual



Qualificação do Modelo



- A relação entre as variáveis independentes e a log odds da variável dependente é linear
- Não existe alta correlação entre as variáveis independentes

Modelo de Regressão Multilinear

```
Call:  
lm(formula = price ~ sqft_living + grade + bedrooms + bathrooms,  
    data = df)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-1032063 -135075 -22930  97756  4647181  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -4.784e+05  1.489e+04 -32.121 < 2e-16 ***  
sqft_living   2.267e+02  3.577e+00  63.362 < 2e-16 ***  
grade         9.614e+04  2.315e+03  41.520 < 2e-16 ***  
bedrooms     -3.931e+04  2.291e+03 -17.158 < 2e-16 ***  
bathrooms    -2.674e+04  3.480e+03 -7.683 1.62e-14 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 248100 on 21608 degrees of freedom  
Multiple R-squared:  0.5434,    Adjusted R-squared:  0.5433  
F-statistic: 6428 on 4 and 21608 DF,  p-value: < 2.2e-16
```

- p-valor < 2.2e-16
- modelo caracterizado pela área útil (sqft_living) e pela categoria (grade) contribuindo positivamente, enquanto quartos (bedrooms) e banheiros (bathrooms) parecem contribuir negativamente
- 54% da variação de preço é explicada pelo modelo (pode ser melhorado)

Obrigado Pela Sua Atenção!

