# Categorização do Conjunto de Dados House Sales in King County

Resumo:

O seguinte documento se trata de uma ampla análise de dados sobre o conjunto de dados “House Sales in King County, USA” disponível publicamente no website Kaggle. Sendo assim, os dados analisados mensuram características de casas do condado King County localizado em Washington nos Estados Unidos da América.

Logo, verificar a estruturação e a integridade dos dados, gerar uma perspectiva visual geográfica, correlacionar as variáveis com o preço e estimar o valor financeiro de casas são os pontos chaves a serem abordados neste estudo. Além disso, tenha em mente que a pesquisa foi desenvolvida com a linguagem de programação R sem procurar atender às questões de nenhum público específico.

Introdução:

O conjunto de dados “Venda de Casas em King County” é um subconjunto das informações sobre transações imobiliárias ocorridas no condado de King, localizado no estado de Washington dos Estados Unidos da América. Sendo assim, coletadas e mantidas pelo Escritório de Avaliação de Propriedades do Condado de King, essas informações abrangem uma ampla gama de detalhes, como preços de venda, tamanho das propriedades, número de quartos e banheiros, bem como informações geográficas.

Os dados foram disponibilizados publicamente pelo próprio Escritório de Avaliação de Propriedades no portal online “King County GIS Center”. O escritório coleta e mantém essas informações para fins de avaliação de propriedades e cálculo de impostos. No entanto, eles também tornaram os dados disponíveis ao público em geral, reconhecendo o valor dessas informações para análises, estudos e pesquisas relacionadas ao mercado imobiliário.

Então, foi dessa forma que em 2016 o usuário HARLFOXEM do Kaggle publicou o conjunto de dados “Venda de Casas em King County” contendo 21.613 registros e 21 variáveis, contabilizando as vendas de maio de 2014 até maio de 2015. Abaixo segue a descrição das colunas:

| | |
|---|---|
| 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 | Condição 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 |

Objetivo:

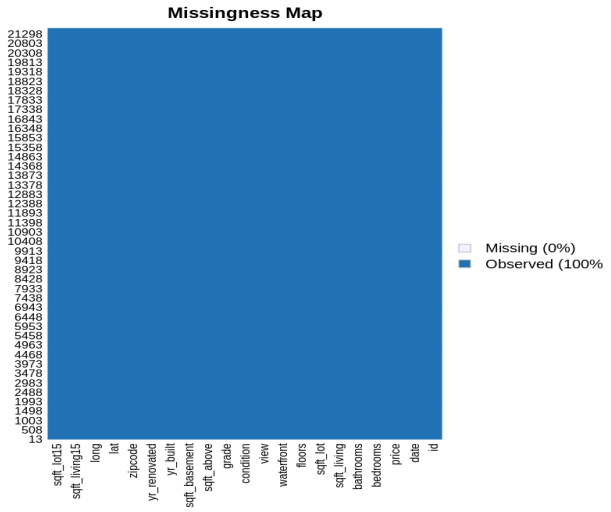
Esta é uma base de dados onde as variáveis abrem possibilidades de prospecções de análises com diferentes enfoques. Por exemplo, os dados fornecidos podem responder perguntas como: Para qual região estão as casas com a melhor qualidade de vida? Qual é a melhor temporada climática para vender casas perto do lago? Ou seja, diferentes questões orientam diferentes análises.

No entanto, nesse cenário, analisar a forma que as variáveis afetam o preço melhor atende a usabilidade deste conjunto de dados. Uma vez que esta é uma base de dados rica em detalhes com uma ampla abrangência das características das casas. Assim, correlacionar estes dados ao preço é o estudo que melhor atende uma ampla análise apropriada para este conjunto de dados.

Ou seja, o preço é a variável mais dependente das outras, necessitando uma ampla análise do banco de dados para entendê-la. Desse modo, as variáveis possuem diferentes graus de impacto e correlação com o valor da venda, sendo um modelo de previsibilidade fundamental para enquadrar essas as questões e entender como as diversas variáveis se relacionam.

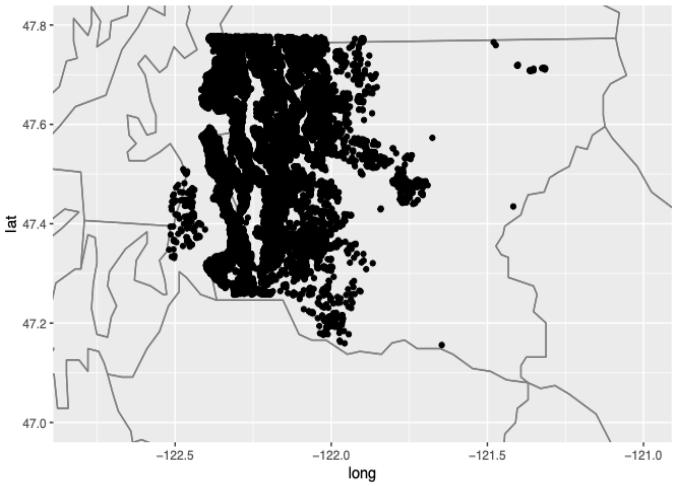
Integridade e Estruturação dos Dados:

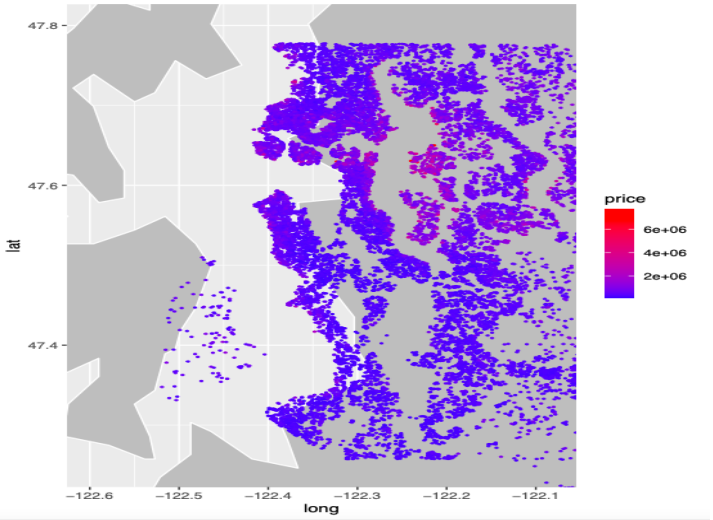
```
Rows: 21,613
Columns: 21
$ id            <dbl> 7129300520, 6414100192, 5631500400, 2487200875, 19544005…
$ date          <chr> "20141013T000000", "20141209T000000", "20150225T000000",…
$ price         <dbl> 221900, 538000, 180000, 604000, 510000, 1225000, 257500,…
$ bedrooms      <int> 3, 3, 2, 4, 3, 4, 3, 3, 3, 3, 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.…
$ sqft_living   <int> 1180, 2570, 770, 1960, 1680, 5420, 1715, 1060, 1780, 189…
$ 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…
$ 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, 3, 0, 0, 0, 0,…
$ condition     <int> 3, 3, 3, 5, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3, 3, 3, 4, 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, 189…
$ 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, 20…
$ yr_renovated  <int> 0, 1991, 0, 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…
$ long          <dbl> -122.257, -122.319, -122.233, -122.393, -122.045, -122.0…
$ sqft_living15 <int> 1340, 1690, 2720, 1360, 1800, 4760, 2238, 1650, 1780, 23…
$ sqft_lot15    <int> 5650, 7639, 8062, 5000, 7503, 101930, 6819, 9711, 8113, …
```




O primeiro passo da análise teve como objetivo identificar os tipos de dados e quantos valores perdidos a base de dados apresenta. Contabilizando a base de dados é composta de 21 colunas, 5 variáveis contínuas (price, sqft_liv, sqft_lot, sqft_above, sqft_basmt), 7 variáveis discretas (bedrooms, floors, lat, long, squft_liv15, squft_lot15, bathrooms), 4 variáveis nominais (id, yr_built, yr_renov, zipcode) e 5 variáveis variáveis ordinais (date, waterfront, view, condition, grade). Sendo que nenhum dado faltante está presente.

Conceitualização Visual:



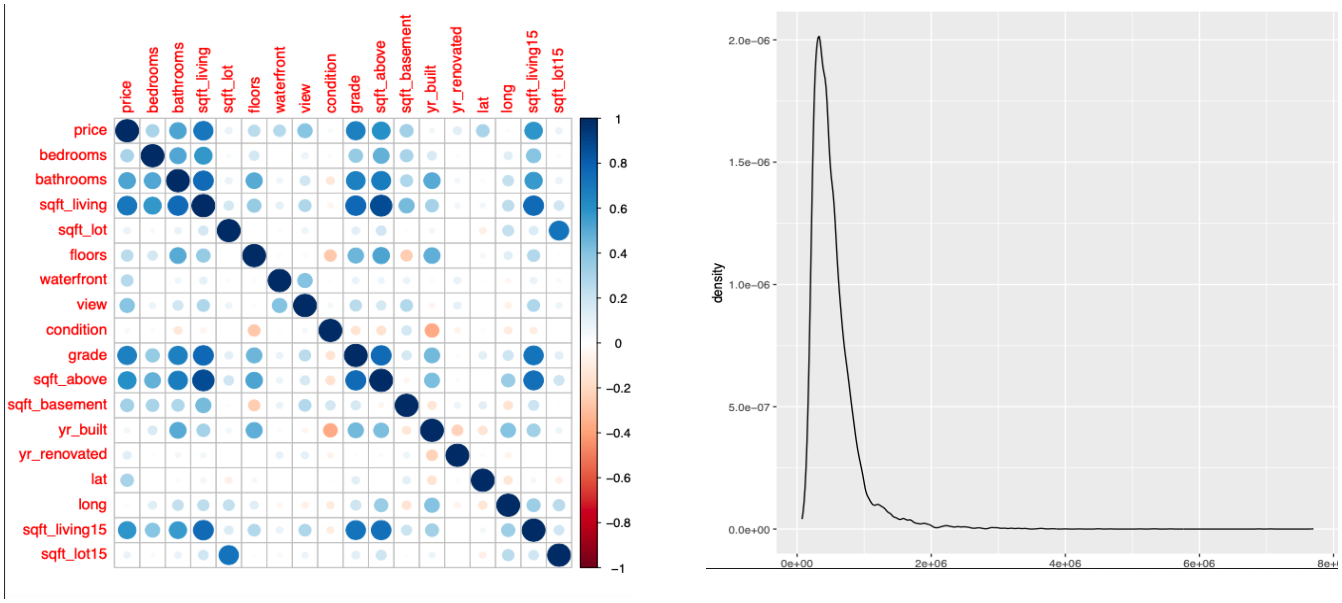




Embora os gráficos geográficos não estejam diretamente relacionados ao entendimento da precificação, estes são um critério chave para a compreensão da organização dos dados. Desse modo, são de grande utilidade para fundamentar hipóteses e agregar sentido a futuros resultados.

Nesse caso, o mapa demonstra que os dados estão organizados de certa forma lógica, apesar de que seja possível destacar uma irregularidade na dispersão das vendas com de altos e baixos valores.

Qualificação do Modelo de Previsão:

As análises até o momento mostraram que as 21 colunas representam uma grande diversidade de características nas vendas. Sendo que quase todas apresentam fundamental impacto na venda. Além de serem variáveis com complexas relações entre si, por exemplo, a relação de como os dados temporais e geográficos afetam as vendas. Desse modo, esses fatores favorecem as análises serem elaboradas sobre um modelo de regressão linear múltipla.




Medindo a correlação entre as variáveis do modelo e traçando a distribuição normal para a vendas realizadas, é possível perceber que os dados se encaixam para a implementação do modelo de regressão linear múltipla. Uma vez que, a relação entre as variáveis independentes e a log odds da variável dependente é linear. Além de que 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
```

O modelo desenvolvido fornece uma previsão estatisticamente significativa dos preços das casas, como mostrado pelo p-valor menor que 2.2e-16. E sendo este modelo caracterizado pela área útil (sqft_living) e pela categoria (grade) contribuindo positivamente, enquanto quartos (bedrooms) e banheiros (bathrooms) parecem contribuir negativamente (embora provavelmente devido a interações com outras variáveis).

No entanto, o erro padrão residual relativamente alto e o fato de que apenas cerca de 54% da variação de preço é explicada pelo modelo sugerem que pode haver outras variáveis importantes não incluídas no modelo, ou relacionamentos não lineares não capturados pelo modelo linear.

Conclusão:

Em suma, a base de dados “House Sales in King County” apresenta uma rica quantidade de informações que caracterizam as vendas de propriedade em King County no período de maio de 2014 a maio de 2015. Dessa forma, essa pesquisa dispensou correlações temporais por ter enfoque no preço das vendas, embora tenha utilizado os mapas geográficos para perceber a organização.

Ademais, pela independência das variáveis que influenciam o preço e pela quantidade de vendas seguir a distribuição de uma normal padrão foi possível contactar a possibilidade da modelação de uma regressão multilinear para analisar os dados. Assim, o modelo desenvolvido realiza previsões sobre o preço das vendas de forma significativa, embora possua um padrão residual relativamente alto.

Por fim, a pesquisa não buscou responder questões específicas, nem mesmo teve um público alvo a quem se direcionar. Apesar de que, caso a metodologia desse estudo tivesse sido direcionada a visualização dos dados ou a optimização das variáveis na regressão multilinear, poderia ter-se chegado a mapas geográficos e a modelos de predição com altos níveis de precisão.