Descoberta de Padrões em Apartamentos para Alugar Através de Mineração de Dados

Gabriel Castelo [1], Gabriel Teixeira Carvalho [2], Henrique Rotsen [3], Maria Luiza Leão Silva [4].

Departamento de Ciência da Computação, Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
[1] gcastelo.rocha@gmail.com
[2] gabrieltc@ufmg.com [3] henriqueferreira@dcc.ufmg.br [4] marialls@dcc.ufmg.br

I. Introdução

Este relatório técnico descreve a descoberta de vários padrões relevantes em um conjunto de dados de **apartamentos para alugar nos Estados Unidos**. O conjunto de dados está disponível em 2 grandes plataformas de dados, **UC Irvine** e **Kaggle**.

Esses dados vieram predominantemente da plataforma online RentDigs.com, ou seja, das 99.442 entradas, 90.912 foram obtidas a partir dessa plataforma. O RentDigs.com é um site de anúncios de aluguel onde proprietários podem postar suas propriedades para locação gratuitamente. Ele permite que os usuários criem anúncios detalhados com várias imagens e descrições para atrair potenciais inquilinos. Além disso, os anúncios publicados no RentDigs são distribuídos em outras plataformas como Oodle.com, Trovit.com, Claz.org, Mitula.com e RentJungle.com, aumentando a visibilidade dos imóveis em diferentes sites.

No que diz respeito aos outros 8.500 dados, eles foram retirados de outras 24 plataformas online: RentLingo, ListedBuy, RENTCafé, GoSection8, Listanza, RealRentals, RENTOCULAR, tenantcloud, Real Estate Agent, rentbits, Home Rentals, Nest Seekers, RentFeeder, vFlyer, Claz, Real Estate Shows, Seattle Rentals, BostonApartments, SpreadMyAd, Apartable, Z57, FreeAdsTime, AgentWebsite, HousesForRent. Vale dizer que este é um conjunto de dados bastante completo e, para fins didáticos, mostrou-se interessante para a descoberta de padrões.

Após uma descrição exploratória da análise de dados na Seção II, a Seção III investiga os dados por meio de um estudo aprofundado de agrupamento dos apartamentos. A Seção IV apresenta uma tarefa preditiva sobre os dados, descrevendo uma regressão realizada para prever o preço de um apartamento. A Seção V mostra padrões encontrados nas comodidades oferecidas nos anúncios, que foram encontrados por uma abordagem de mineração de *itemset*. Finalmente, a Seção VI resume os resultados encontrados e conclui o relatório.

A. Tratamento dos dados

Antes de mais nada vamos passar pelos tratamentos realizados. Descartamos as colunas: category, title, body, currency, fee, address, has_photo, price_display e price_type. As definições delas podem ser encontradas no dataset disponibilizado no final deste arquivo. Houve esse descarte, pois elas não agregavam para as tarefas que foram realizadas. Vamos falar um pouco, apenas, sobre os atributos que foram usados nas tarefas:

1) *id*: Id numérico único para cada anúncio
2) *amenities*: A coluna *amenities* consiste em uma série de comodidades que estão presente nas moradias. Não necessariamente existem comodidades em todas elas, e algumas moradias não tiveram essa coluna bem preenchida, portanto, há objetos no conjunto que apresentam discrepância em termos de valor (*price*) e *amenities* existentes. Para permitir bom uso na Seção V, foi realizado um *one hot encoding*, criando-se novas colunas, um referente a cada *amenity*, totalizando 27 colunas.
3) *bathrooms*: Coluna numérica, contendo número de banheiros de um objeto.
4) *bedrooms*: Coluna numérica, contendo número de quartos de um objeto.
5) *pets_allowed*: Coluna categórica. Denota a permissão ou não de pets dentro daquela residência. Assume quatro valores: "Cats", "Dogs", "Cats,Dogs", None. A coluna foi tratada através de um *one hot encoding*, e separada em duas colunas diferentes: *gatos*_permitidos e *cachorros_permitidos*
6) *price*: Coluna numérica. Preço da residência.
7) *square_feet*: Coluna numérica. Tamanho da residência.
8) *city_name*: Coluna categórica. Nome da cidade onde a residência está localizada.
9) *state*: Coluna categórica. Sigla do estado onde se localiza a residência. Possuí 51 valores distintos, pois existem alguns anuncios do Havaí.
10) *latitude* e *longitude*: Ambas colunas numéricas. Indicam as coordenadas de latitude e longitude referentes àquela residência.

Outro tratamento realizados em todas as seções foi a remoção dos outliers. Essa remoção ocorreu da seguinte maneira, para cada estado foram removidos os outliers, pois dessa maneira garantimos que um estado que seja mais caro ou mais barato, não seja prejudicado.

Vale mencionar que, apenas para fins de comparação, após a remoção desses outliers, a média de preço de aluguel do dataset foi de US$ 1339.54, um valor muito próximo do

divulgado pela mídia. O site **Apartments.com**, que é uma das maiores plataformas de anúncios de imóveis, indica um valor médio de US$ 1536. Isso sugere que o dataset não apresenta viés significativo.

II. Análise Exploratória de Dados

A. Visualização dos dados

Tendo em vista que uma das questões mais relevantes de mineração do preço de apartamentos para aluguel é entender a distribuição geográfica dos preços, para, posteriormente, buscar agrupamentos, um excelente ponto de partida para a exploração de dados é visualizá-los em um mapa.

Para tal tarefa, inicialmente, foi criada uma cópia do conjunto de dados, onde descartaram-se todas as colunas, exceto as colunas *latitude*, *longitude* e *price*. Em seguida, foi realizado um arredondamento das colunas *latitude* e *longitude* para uma casa decimal, e um agrupamento por estas mesmas colunas, tirando a média do preço nos grupos. A visualização dos dados foi feita na forma de um mapa de calor, que permitiria entender de forma visualmente agradável. O primeiro tratamento feito foi a criação de um agrupamento de coordenadas de latitude e longitude em supergrupos, mudando o padrão de variação de 0,1 em 0,1 para 0,75 em 0,75. A cor de cada quadrado representa o preço médio do mesmo, conforme visto na figura 1.

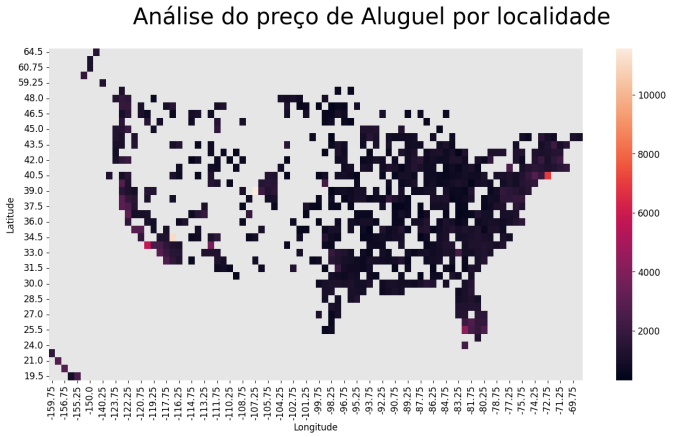

Figura 1. Mapa com agrupamento de Coordenadas e Outliers

Mesmo partindo de um mapa que não permite uma visualização boa por conta da discrepância de preço em algumas regiões, essa visualização já nos permite ver uma diferença grande entre os preços de cidades mais importantes (tais quais Nova York e Los Angeles) em relação às demais. Essa discrepância fica ainda mais evidente quando verificamos o preço médio por estado:

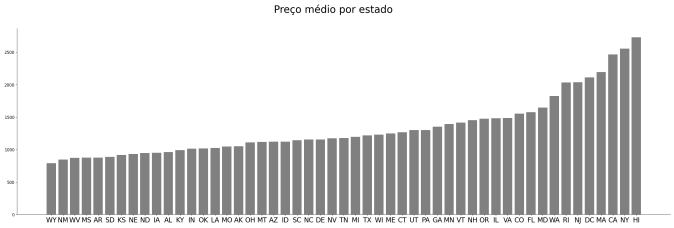

Figura 2. Gráfico de preço médio por Estado

Os estados de economia mais marcante, ou com qualidade de vida mais alta notadamente tem um preço dos imóveis mais caros. Todavia, os valores exorbitantes dos outliers não permitem maiores deduções sobre essas visualizações. Conforme demonstrado na figura 3, a grande maioria dos valores se encontra abaixo de US$10000,00, enquanto alguns se encontram acima dessa faixa. portanto um tratamento visando a remoção dos outliers foi aplicado. A remoção dos outliers se deu pela remoção dos 5% mais altos e dos 5% mais baixos, por estado. Com isso, valores extremos foram desconsiderados.

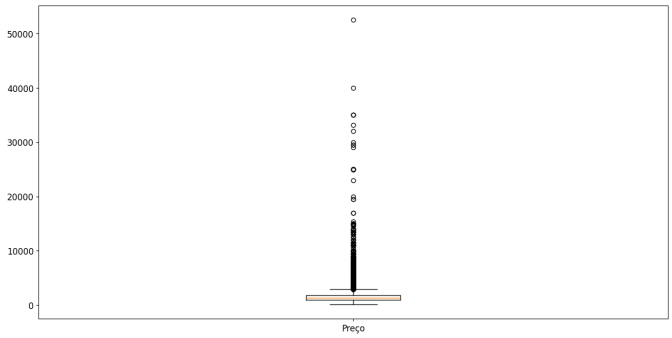

Figura 3. Boxplot com preços

Tendo em vista isso, é possível averiguar uma mudança nos preços médios:

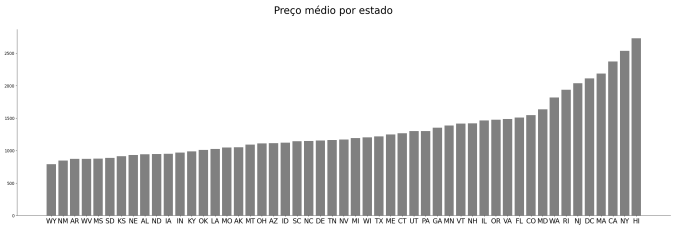

Figura 4. Preço médio nos estados após remoção de outliers

Apesar da remoção dos outliers, verifica-se que os estados com maiores preços se mantem de maneira aproximada em suas posições, conforme o esperado. Feito isso, é possível visualizar um mapa de calor dos Estados Unidos de maneira mais clara, como na figura 5:

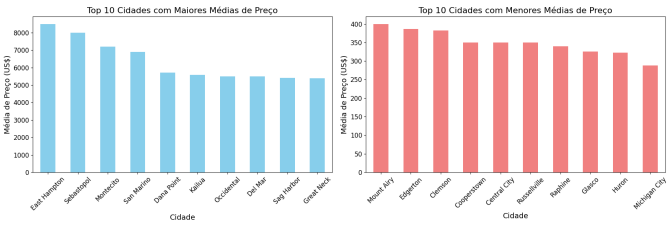

Figura 6. Gráficos de barras com 10 cidades mais caras e mais baratas

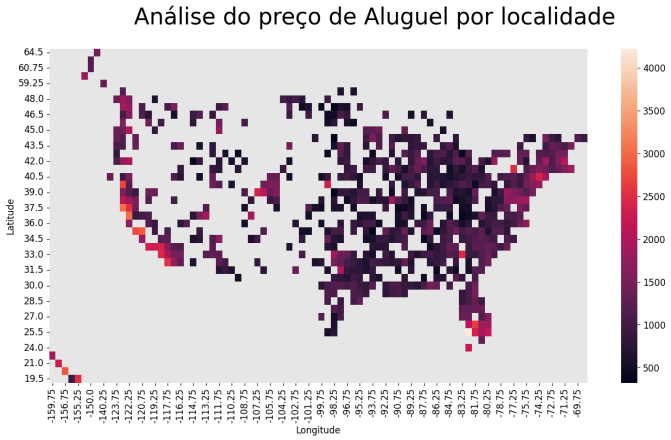

Figura 5. Heat Map Sem outliers

Fica notável, portanto, uma concentração de apartamentos caros em regiões como *Nova York*, *Califórnia* e *Flórida*.

Um refinamento da exploração, removidos os outliers, é descobrir, então, as cidades mais caras e mais baratas do país, informação extremamente relevante para entender a distribuição de preços dos apartamentos de alguel nos EUA.

Como pode ser visto na figura 6, as cidades mais caras de fato estão nos estados esperados (as 3 primeiras cidades estão em NY, CA e CA, respectivamente).

B. *Conclusão:*

A exploração de dados inicial nos permitiu mapear uma média de preços de apartamentos nos Estados Unidos e entender regiões onde os apartamentos de maior preço se concentram. Além disso, a Análise exploratória dos dados permitiu uma noção clara de como os outliers estavam influênciando em uma disparada nos preços em um ponto especifico do país. Finalmente, pudemos verificar, após remoção dos outliers, uma consistência entre as cidades mais caras e os estados mais caros.

III. Agrupamento

Nessa tarefa buscamos achar um agrupamento para os dados do dataset, usando as seguintes colunas: Banheiros, Quartos, Preço, Metragem, Gatos Permitidos e Cachorros Permitidos. O algoritmo escolhido foi o DBSCAN, já que ele se comporta muito bem em identificar clusters de tamanhos arbitrários e se comportou bem nos nossos testes.

Para realizar a tarefa os dados foram normalizados utilizando a configuração Z-Score e já havíamos removido os outliers como dito na introdução. Além disso, sempre que era necessário usar uma métrica de distância, a escolhida era a euclideana.

Seria interessante trabalhar em um agrupamento que fosse possível de ser visualizado, para isso seria necessário usar alguma técnica de redução de dimensionalidade, nesse caso, escolhemos o PCA. Para isto usamos 2 componentes principais, escolha esta que preservaria 72,1% de toda a informação e iria permitir uma visualização dos dados em um plano 2D.

A primeira componente principal é formada pelos seguintes pesos: Banheiros (0.539), Quartos (0.500), Preço (0.260), Metragem (0.550), Gatos Permitidos (0.200), Cachorros Permitidos (0.220). Isso mostra que a maior variância dos dados é captada pelas seguintes colunas: Banheiros, Quartos e Metragem.

Já a segunda componente é formada por: Banheiros (0.139), Quartos (0.222), Preço (0.044), Metragem (0.159), Gatos Permitidos (-0.677), Cachorros Permitidos (-0.667). Nessa componente percebemos que existe um grande peso **negativo** associado as colunas de *pets*, e já as outras colunas um peso mais baixo, porém **positivo**. O que sugere que os clusters podem ter uma separação evidente pelo eixo associado ao à segunda componente principal.

A. *DBSCAN*

No caso dele temos que fornecer um *épsilon* que é o raio de busca e um número de objetos mínimo que deve estar nesse raio. Nesse caso usamos *épsilon* = 0.5, número mínimo de objetos = 5, e a métrica de distância sendo a euclidiana. A escolha desses números aconteceu baseada em tentativa e erro, essa combinação foi a que mais chegou perto do resultado obtido nos outros algoritmos, um Coeficiente da Silhueta de 0.356. Além de que ao aumentar/abaixar os hiper parâmetros o Coeficiente da Silhueta apenas piorava o resultado. Veja, na Figura 7, que utilizando 2 componentes principais achamos três clusters e que o DBSCAN os agrupou bem.

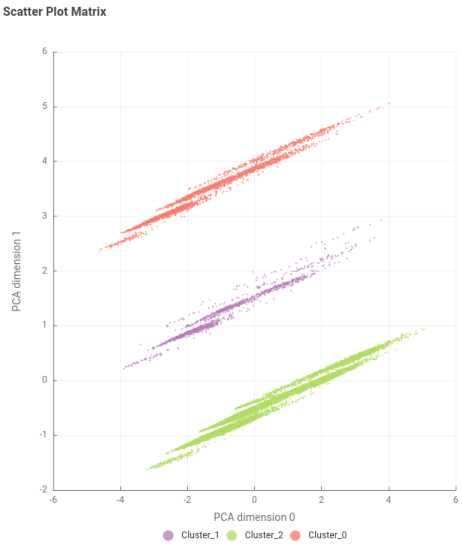

Figura 7. Resultado do DBSCAN

B. Análise dos Resultados

A partir dos pesos do PCA e do resultado do DBSCAN vimos que ele temos 3 clusters com uma forma quase linear, isso indica que os apartamentos de cada grupo podem ser bastante diferentes. Logo, o que divide os clusters é a permissão ou não de animais. Entretanto como temos 2 colunas de animais, deveríamos ter 4 clusters: não permite animais, permite apenas gatos, permite apenas cachorros, permite ambos os animais. Com isso ao fazer uma análise detalhada vimos que quase todo apartamento que permite cachorro, também permite gato, para ser mais preciso, apenas 61 apartamentos permitem cachorros e não gatos.

Isso pode ser pode refletir preferências comuns no mercado imobiliário e regulamentações condominiais. Além disso, características dos apartamentos e decisões dos proprietários geralmente levam à aceitação de ambos os animais, resultando na fusão dos clusters esperados. O pequeno número de apartamentos que permite apenas cachorros indica que as políticas de aceitação de animais são mais homogêneas do que inicialmente previsto.

Já analisando o Eixo X na 7, vemos que os dados são bem completos, contendo grandes variações de configurações diferentes de apartamentos. Provavelmente os apartamentos mais da esquerda são apartamentos menores e menos completos. Isso não quer dizer que eles não vão ser caros, tudo vai depender do preço do "metro quadrado"dessas regiões, ao fazer uma análise percebemos que os apartamentos que possuem um preço de aluguel por metro quadrado mais caro (entre 7 e 11 price/sqft), estão à esquerda do gráfico e sendo predominantemente da Califórnia (72%) com uma mensalidade média de US$ 2108.38 e tamanho médio de 273 sqft (25.3 m²). Já para os outros apartamentos eles possuem uma distribuição homogênea nos clusters, indicando que os dados são bem variados.

Uma observação é que se o DBSCAN identificou bem os grupos, por que o score dele foi tão baixo? Bom para responder isso devemos lembrar que o Coeficiente da Silhueta para um ponto é calculado como a diferença entre a distância média do ponto aos outros pontos em seu próprio cluster e a menor distância média do ponto ao cluster mais próximo, dividida pelo maior valor entre essas duas distâncias. Ao analisar o gráfico de dispersão dos dados, podemos ver que os pontos da extremidade vão estar mais mais próximo de outro cluster do que de alguns pontos do próprio cluster. Isso fica fácil de ser visualizado se colorirmos o gráfico usando o resultado do Coeficiente da Silhueta. Veja o que acontece na Figura 8.

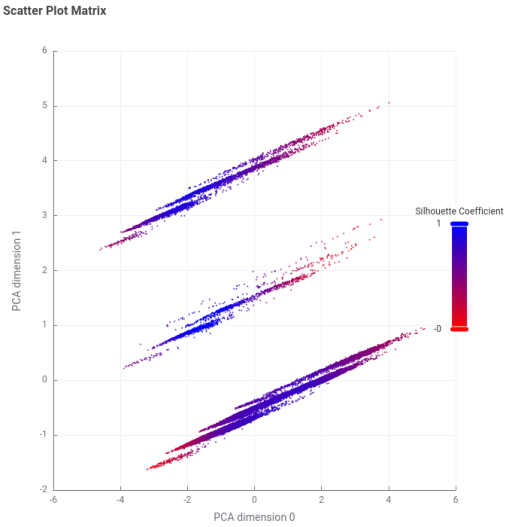

Figura 8. Scores para cada ponto

IV. Regressão

No campo da modelagem preditiva, a análise de regressão serve como uma técnica fundamental para entender as relações entre variáveis. Neste estudo, nosso objetivo é prever com precisão os preços de aluguel de apartamentos utilizando modelos de regressão linear e não linear. A motivação para o uso da regressão é descobrir e quantificar a influência de várias features, como metragem quadrada e comodidades, nos preços das propriedades. Ao aplicar tanto a regressão linear quanto técnicas avançadas como a Random Forest Regression, buscamos capturar relações tanto simples quanto complexas dentro dos dados. Essa abordagem dupla não só nos permite avaliar a eficácia de diferentes modelos, mas também fornece insights sobre a estrutura subjacente dos dados, levando a previsões mais precisas e confiáveis.

A. Metodologia

Primeiramente, consideramos somente as colunas numéricas e booleanas do data-set. Ademais, consideramos somente as entradas que continham no mínimo alguma comodidade e que não tinham nenhum valor faltando para as colunas que usaríamos no processo da regressão. Isso fez com que acabacem com um total de 84.202 instancias. Então, realizamos mais dois passos que serão explicados nas Subseções IV-A1 e IV-A2 a seguir.

1) Mediana do preço do pé quadrado por cidade: Segundo a Seção II-B, a cidade onde o apartamento está localizado é importante para prever o preço. Com isso em mente, adicionamos uma nova coluna ao data-set chamada 'sqf_price' e representa a mediana do preço do pé quadrado para a cidade onde esse apartamento está localizado. No total, são 2671 cidades diferentes no data-set e, para obter essa nova coluna, calculamos o preço do pé quadrado de cada entrada e depois agrupamos por cidade e obtemos a mediana desses valores. Essa nova coluna foi utilizada tanto no modelo de regressão linear quanto no modelo de regressão Random Forest.

2) Análise da distribuição: Para garantir que nosso modelo de regressão produza resultados confiáveis, é essencial abordar a distribuição da variável alvo, price. Em uma regressão linear, pressupõe-se que os resíduos do modelo seguem uma distribuição normal, uma condição fundamental para a validade das inferências estatísticas, como a significância dos coeficientes e a precisão das previsões.

Portanto, conduzimos uma análise da distribuição da coluna price. Conforme ilustrado na Figura 9, a distribuição original

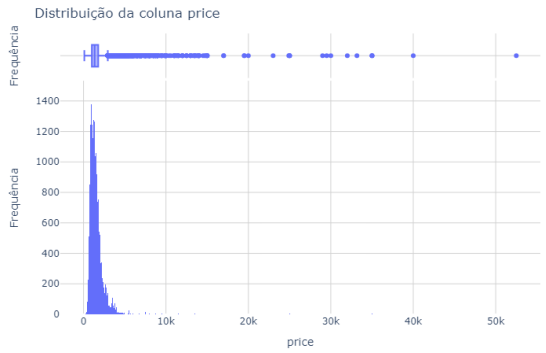

Figura 9. Histograma e boxplot do preço

de price é altamente assimétrica, com uma cauda longa à direita e a presença de outliers com valores extremamente elevados. Essa assimetria indica que muitos valores se concentram em uma faixa inferior, enquanto alguns preços são excepcionalmente altos, o que pode distorcer a interpretação e o desempenho do modelo de regressão linear.

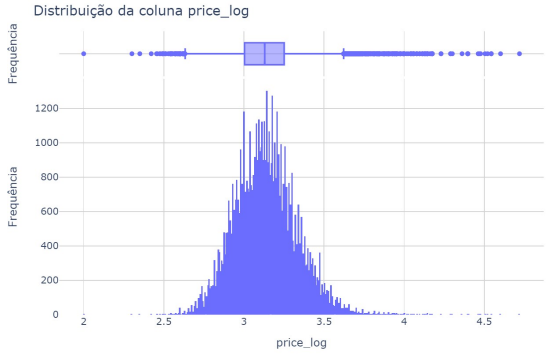

Figura 10. Histograma e boxplot do log do preço

Para mitigar esses problemas, realizamos uma transformação logarítmica na coluna price, como mostrado na Figura 10, usando um log de base 10. A transformação logarítmica resulta em uma distribuição mais simétrica e próxima da normalidade, alinhando-se melhor com as suposições da regressão linear. Ao normalizar a distribuição da variável alvo, melhoramos a robustez e a precisão do modelo, permitindo que ele capture de forma mais eficaz as relações subjacentes nos dados e produza previsões mais confiáveis.

B. Regressão Linear

Primeiro, implementamos uma regressão linear para prever os preços de aluguel dos apartamentos. Considerando que a regressão linear assume que as amostras são independentes e identicamente distribuídas, realizamos uma normalização das colunas numéricas e uma foward feature selection para identificar as variáveis mais relevantes para nosso modelo. Esses processos são detalhados a seguir.

1) Normalização: Para permitir uma interpretação dos coeficientes usados na regressão pra cada feature, normalizamos os as colunas 'bathrooms', 'bedrooms', 'square_feet', 'sqf_price', 'longitude' e 'latitude' utilizando o método Min-Max. Assim, teremos somente valores entre 0 e 1. Como as outras colunas são booleanas e só assumem os valores 0 ou 1, a normalização só foi aplicada nas colunas numéricas citadas.

2) Seleção de features: Para selecionar quais features adicionar no nosso modelo, utilizamos a função SequentialFeatureSelector do biblioteca sklearn. Esse método de seleção irá fazer uma foward selection das features seguindo uma estrategia gulosa. Ou seja, começamos com nenhuma feature selecionada e encontramos a feature que maximiza o R^2 score da validação cruzada com 5 folds de uma regressão linear que tenta prever o log10 do preço usando somente aquela feature. Uma vez que essa primeira feature é selecionada, esse processo é repetido adicionando uma nova feature ao conjunto de features já selecionadas. Como limite da quantidade de features para o modelo, optamos por um mínimo de 1 feature e um máximo de 10 features.

Como resultado, tivemos que o R^2 score foi maximizado ao ser utilizar 10 features, sendo elas 'bathrooms', 'square_feet', 'longitude', 'gym', 'parking', 'refrigerator', 'elevator', 'playground', 'ac', 'sqf_price'.

3) Resultados: Finalmente, executamos uma validação cruzada, usando um K-fold com k igual a 5, da regressão linear com as 10 features encontradas anteriormente. Obtivemos um R^2 médio de 0.6969, o que indica que nosso modelo explica 69.69% da variação nos preços dos apartamentos. Isso sugere que o modelo captura uma boa quantidade da variabilidade presente nos dados, mas ainda pode ser melhorado.

Para ajudar na interpretação dos resultados, passamos o valor previsto pela regressão, que corresponde ao log do preço, para o exponencial do log, ou seja, de volta para o preço(dólar). Na Figura 11, temos que para algumas entradas o valor previsto pelo modelo é muito distante do real valor. Essa diferença acontece em ocorre tanto para mais, pontos a cima da linha vermelham, como para menos, pontos a baixo da linha vermelha.

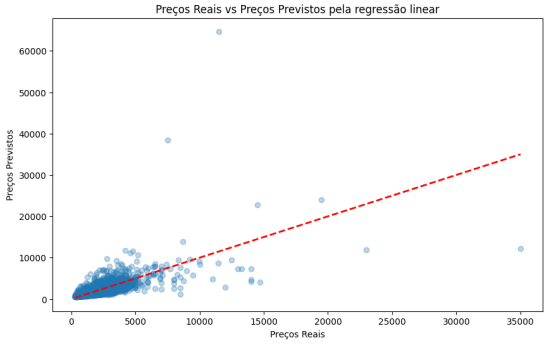

Figura 11. Dispersão dos preços reais e dos preços previstos pela regressão linear

No que diz respeito aos coeficientes das features no modelo, apresentados na Tabela I, os coeficientes estimados na regressão linear indicam que 'square_feet' e 'sqf_price' são as duas features mais influentes, com os maiores coeficientes, sugerindo que o tamanho do apartamento e a mediana do preço

Tabela I
COEFICIENTES DAS FEATURES NO MODELO DE REGRESSÃO LINEAR

Feature	Coeficiente
square_feet	2.3199
sqf_price	1.5062
bathrooms	0.3415
longitude	0.0567
elevator	0.0450
gym	0.0300
parking	0.0191
ac	-0.0203
playground	-0.0305
refrigerator	-0.0351

do pé quadrado naquela cidade contribuem para aumentar o log do preço previsto. As features 'longitude' e 'bathrooms' também têm uma influência positiva, embora menor.

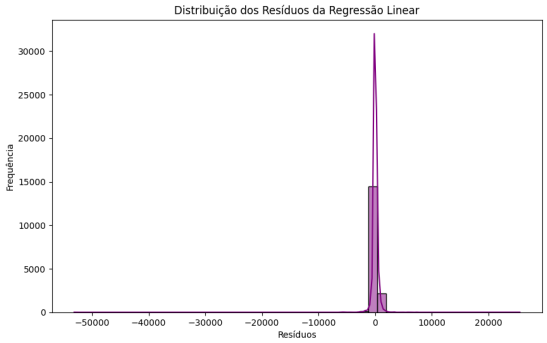

Figura 12. Distribuição dos resíduos da regressão linear.

A distribuição dos resíduos, mostrada na Figura 12, indica de o modelo produz resíduos negativos altos. Isso indica que ele produz algumas previsões com valores bem mais altos do que o valor real, chegando a aumentar mais de 50 mil no valor de aluguel para um apartamento. Ademais, temos também previsões menores que o real valor, chegando a prever valores até 20 mil dólares mais baratos que o real valor do aluguel. Tudo isso, sugere que ainda pode haver uma ligeira assimetria ou presença de outliers que precisam ser tratados para potencialmente melhorar o modelo.

C. Regressão Random Forest

Após analisar os resultados da regressão linear, decidimos explorar um modelo não linear robusto. Optamos pelo Random Forest Regression, um modelo de regressão não linear, devido à sua capacidade de lidar com interações complexas entre variáveis e por não ser afetado pela presença de outliers.

Executamos o modelo de Random Forest Regressor utilizando todas as features disponíveis no conjunto de dados e 'price_log' como target, configurado com os valores default da biblioteca sklearn, max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100. Obtivemos um modelo com R^2 de 0.8738, o que significa que o modelo explica aproximadamente 87.38% da variação nos preços dos apartamentos e é uma indicação de que o modelo tem um bom ajuste e é eficaz na captura das relações subjacentes nos dados.

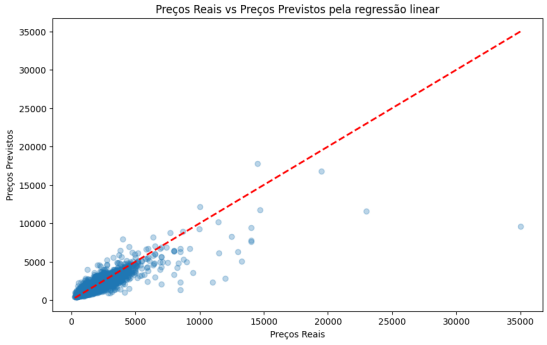

Figura 13. Dispersão dos preços reais e dos preços previstos pelo Random Forest Regressor

Tabela II
10 FEATURES COM MAIOR IMPORTÂNCIA NO MODELO DE RANDOM FOREST

Feature	Importância
sqf_price	0.5496
square_feet	0.2307
longitude	0.0592
latitude	0.0486
bathrooms	0.0262
bedrooms	0.0106
gym	0.0054
parking	0.00460
pool	0.0045
playground	0.0042

No que diz respeito a importância de cada feature no nosso modelo, temos na Tabela II a 10 features com maior importância, sendo a importância calculada como a media e desvio padrão da acumulação do decrescimento da impureza em cada árvore. Temos que, assim como na regressão linear, 'sqf_price' e 'square_feet' se destacam entre as features. Isso possivelmente indica que o tamanho e a tendencia do valor do pé quadrado também influenciam diretamente o preço previsto para o apartamento. 'longitude', 'bathrooms', 'playground' e 'gym' são outras features que aparecem em similarmente nos dois modelos. Contudo, 'latitude', 'bedrooms' e 'pool' são features que não foram utilizadas no modelo linear mas que aparecem entre as features mais importantes do modelo não linear.

O histograma dos resíduos, Figura 14, não tem mais os altos valores negativos que observamos no modelo linear, o que nos indica que esse modelo não prevê presos bem maiores que o valores reais. Contudo, ainda podemos perceber a presença de predições bem menores que o real valor, como também ocorreu no modelo linear. Também podemos observar isso na Figura 13.

D. Conclusão

Nossa análise de regressão demonstrou a eficácia variada de modelos lineares versus não lineares na previsão de preços de imóveis. O modelo linear apresentou uma boa performance com um R^2 perto de 70%. Já o modelo não linear forneceu um poder preditivo maior que o linear, com um R^2 que se

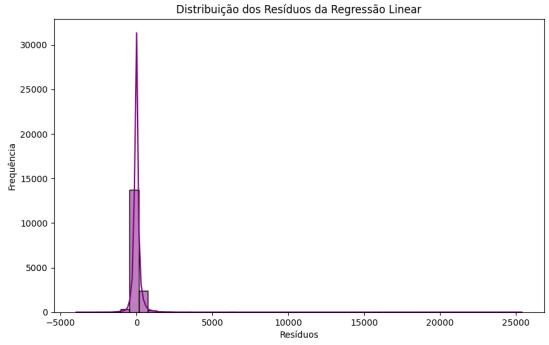

Figura 14. Distribuição dos resíduos do Random Forest Regressor

aproxima de 88%. De modo geral, fomos capazes de obter um bom modelo para prever o prever o preço de aluguel para os apartamentos.

V. Mineração de Itemsets

Nesta seção, utilizamos a mineração de itemsets [1] para identificar padrões significativos entre as comodidades oferecidas pelos apartamentos presentes no conjunto de dados. A mineração de itemsets é uma técnica fundamental na análise de dados que permite descobrir associações frequentes e relevantes entre conjuntos de itens.

A. Metodologia

84484 apartamentos têm pelo menos uma comodidade. Para esses, consideramos a presença de cada uma das 27 comodidades (ou seja, 27 atributos Booleanos) e listamos os itemsets fechados compostos de pelo menos duas comodidades, suportados por pelo menos 8449 apartamentos (10%) e com os cinco maiores lifts.

Durante a análise, observamos que os itemsets com mais de 2 itens apresentaram lifts extremamente baixos ou não existiram em nosso conjunto de dados. Isso indica que as associações entre as comodidades tendem a ser mais fracas ou menos frequentes quando consideramos a presença simultânea de três ou mais itens.

B. Itemsets com Maior Lift no Conjunto de Dados Inteiro

Primeiramente, analisando a distribuição de comodidades no conjunto de dados inteiro na figura 15, percebemos que as comodidades mais comuns são *parking*, *pool* e *gym*.

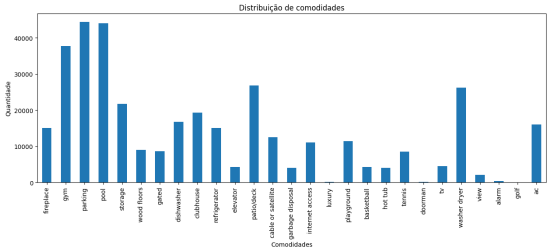

Figura 15. Distribuição das comodidades.

Em seguida, fizemos a análise dos itemsets com maior *lift* no conjunto de dados inteiro, com o objetivo de identificar as associações mais significativas nos apartamentos de cidades dos EUA como um todo.

Tabela III
Top 5 Itemsets com Maior Lift no Conjunto de Dados Inteiro

Itemsets	Freq. Rel.	Lift
(dishwasher, refrigerator)	0.134724	3.786168
(dishwasher, ac)	0.107275	2.837779
(patio/deck, dishwasher)	0.117028	1.850890
(fireplace, washer dryer)	0.101948	1.848020
(washer dryer, ac)	0.103357	1.757369

A Tabela III apresenta os 5 itemsets com o maior *lift* no conjunto de dados inteiro. Observamos que o itemset (dishwasher, refrigerator) tem o maior *lift*, 3.786168, indicando uma forte associação entre a presença de uma lava-louças e uma geladeira nos imóveis. Além disso, os itemsets (dishwasher, ac), (patio/deck, dishwasher), (fireplace, washer dryer) e (washer dryer, ac) também apresentam *lift* significativo. Os altos valores de *lift* nesses itemsets sugerem que a presença de um item é acompanhada de uma alta probabilidade de encontrar o outro, o que pode ser útil para entender as preferências dos inquilinos e as práticas comuns no mercado imobiliário.

Esses resultados mostram que estas comodidades são frequentemente oferecidas juntas nos apartamentos dos EUA. Nesses itemsets, é possível observar uma quantidade alta de comodidades relacionadas a eletrodomésticos, como lava-louças, geladeira, lavadora e secadora e ar-condicionado. Outros itens, como *patio/deck* e *fireplace*, também aparecem em associação com esses eletrodomésticos.

Esses resultados sugerem que, embora esses itens não sejam os mais comuns no mercado imobiliário dos EUA, eles são frequentemente oferecidos em conjunto nos apartamentos. No caso dos eletrodomésticos, isso pode indicar que eles são instalados simultaneamente, sendo considerados itens essenciais ou já beneficiados pela infraestrutura necessária para sua instalação. No caso de *patio/deck* e *fireplace*, comodidades mais comuns em apartamentos maiores ou mais caros, a associação com os eletrodomésticos pode ser explicada pela presença de espaços mais amplos nesses imóveis, ou pelo fato de que esses eletrodomésticos são vistos como essenciais.

C. Itemsets com Maior Lift nas 10 Cidades Mais Caras

Para entender como as associações entre comodidades se dão em áreas de alto custo de vida[2], analisamos os itemsets mais significativos nas 10 cidades com os preços de aluguel mais altos nos EUA, dentre as 100 cidades mais populosas do país. Esse conjunto de dados possui 6114 transações das cidades New York, Jersey City, San Francisco, Boston, Miami, San Jose, Arlington, San Diego, Los Angeles, Chicago.

Primeiramente, analisamos a distribuição de comodidades nas 10 cidades mais caras, conforme apresentado na figura 16.

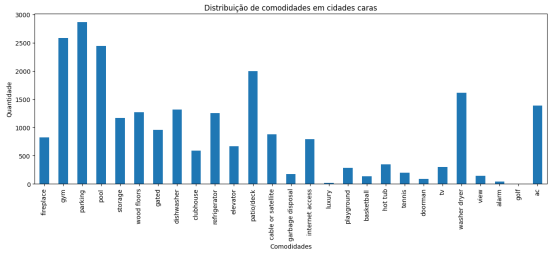

Figura 16. Distribuição das comodidades nas 10 cidades mais caras.

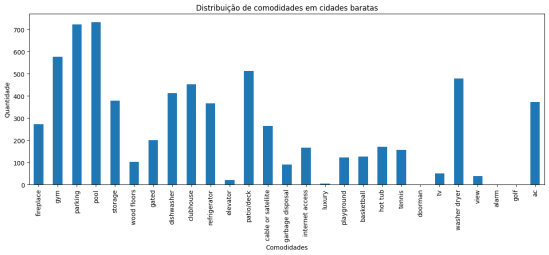

Figura 17. Distribuição das comodidades nas 10 cidades mais baratas.

Nessas cidades, novamente observamos que as comodidades mais comuns são *parking*, *gym* e *pool*.

Tabela IV
TOP 5 ITEMSETS COM MAIOR LIFT CONSIDERANDO AS 10 CIDADES MAIS CARAS

Itemsets	Freq. Rel.	Lift
(dishwasher, refrigerator)	0.135754	3.075035
(dishwasher, ac)	0.125940	2.582917
(washer dryer, ac)	0.107295	1.796447
(patio/deck, ac)	0.120379	1.626140
(patio/deck, dishwasher)	0.112038	1.592812

A Tabela IV apresenta os 5 itemsets com o maior *lift* no conjunto de dados das 10 cidades mais caras. Observamos que o itemset (dishwasher, refrigerator) aparece novamente como o itemset com o maior *lift*, 3.075035. Além disso, os itemsets (dishwasher, ac), (washer dryer, ac), (patio/deck, ac) e (patio/deck, dishwasher) também apresentam *lift* significativo.

Esses resultados sugerem as mesmas interpretações feitas para o conjunto de dados inteiro e mostram que apartamentos nessas cidades são frequentemente equipados com essas comodidades em conjunto, mesmo não sendo as comodidades que mais aparecem nos apartamentos. Além disso, os resultados podem indicar que essas comodidades são esperadas pelos inquilinos em áreas de alto custo de vida.

D. Itemsets com Maior Lift nas 10 Cidades Mais Baratas

Por outro lado, para as 10 cidades mais baratas dentre as 100 cidades mais populosas do país, a análise revela como as associações de comodidades diferem em regiões de menor custo. Nesse conjunto de dados foram consideradas 1421 transações das cidades Lincoln, El Paso, Memphis, Albuquerque, Tucson, Tulsa, Oklahoma City, Shreveport, Wichita, Akron.

Analisando a distribuição de comodidades nas 10 cidades mais baratas, conforme apresentado na figura 17, observamos novamente que as comodidades mais comuns são *parking*, *gym* e *pool*.

A Tabela V apresenta os 5 itemsets de maior *lift* considerando as 10 cidades mais baratas. Mais uma vez, as comodidades mais comuns não apareceram nos itemsets com maior *lift*. Observamos que o itemset (dishwasher, refrigerator) mais uma vez aparece como o itemset com o maior *lift*, 2.847173. Além desse, os itemsets (dishwasher, ac), (cable or satellite, ac), (ac, refrigerator) e (cable or satellite, dishwasher) também contam com um alto *lift*.

Tabela V
TOP 5 ITEMSETS COM MAIOR LIFT CONSIDERANDO AS 10 CIDADES MAIS BARATAS

Itemsets	Freq. Rel.	Lift
(dishwasher, refrigerator)	0.202674	2.700042
(dishwasher, ac)	0.171006	2.247540
(cable or satellite, ac)	0.102745	2.112516
(ac, refrigerator)	0.136524	2.019235
(cable or satellite, dishwasher)	0.102041	1.889766

Esses resultados sugerem as mesmas observações feitas para os outros conjuntos de dados, e mostra que as comodidades relacionadas a eletrodomésticos são frequentemente oferecidas juntas nos apartamentos das cidades mais baratas dos EUA. Isso indica que essas comodidades aparecem juntas mesmo em regiões de menor custo, o que pode sugerir que essas comodidades são consideradas básicas ou essenciais pelos inquilinos, independentemente do contexto econômico. Além disso, todas as comodidades desses itemsets possuem um valor relativamente baixo quando comparado com outras comodidades, o que pode indicar um padrão específico desse contexto econômico.

E. Conclusão

A análise dos *lift* dos itemsets revela padrões interessantes sobre a presença de comodidades em diferentes contextos econômicos. Em todos os conjuntos de dados, as comodidades mais frequentes não apareceram nos itemsets frequentes, o que mostra que mesmo sendo comuns, essas comunidades não são oferecidas em conjunto com outras. Além disso, nos três conjuntos de dados, os itemsets com maior *lift* incluem comodidades relacionadas a eletrodomésticos, como lava-louças, geladeira, lavadora e secadora e ar-condicionado. Esses resultados sugerem que essas comodidades são frequentemente oferecidas juntas nos apartamentos dos EUA, independentemente do contexto econômico e são consideradas essenciais pelos inquilinos. Além disso, outras comodidades, como *patio/deck* e *fireplace* também aparecem em associação com essas comodidades, o que pode indicar que essas comodidades são oferecidas em apartamentos maiores ou mais caros.

Essas descobertas fornecem uma visão valiosa sobre como as comodidades são agrupadas em diferentes contextos de mercado e podem informar decisões sobre marketing e desenvolvimento de empreendimentos no setor imobiliário.

VI. Conclusão

A exploração inicial dos dados foi crucial para identificar outliers e definir estratégias para lidar com eles, além de traçar possíveis caminhos para as etapas subsequentes do projeto. Durante a tarefa de agrupamento, enfrentamos desafios ao lidar com a coluna de Amenities, que se mostrou problemática, conforme discutido na Seção III. Ao optar por excluir essa coluna, conseguimos simplificar a análise, especialmente após a aplicação do PCA, que se tornou mais fácil de interpretar com menos variáveis. Esta abordagem resultou em um processo de agrupamento mais claro e eficiente, culminando em resultados satisfatórios.

Na análise de regressão, destacamos a importância de selecionar modelos adequados para prever preços de imóveis. A regressão linear, embora útil para uma visão inicial das relações entre variáveis, não capturou toda a complexidade dos dados. O Random Forest Regression mostrou-se mais eficaz, capturando padrões complexos e explicando uma parte significativa da variância. A normalização e a seleção criteriosa de features foram essenciais para otimizar o desempenho do modelo. Além disso, a partir da mineração de itemsets, identificamos padrões significativos entre as comodidades oferecidas nos apartamentos, revelando associações frequentes e relevantes entre conjuntos de itens. Com a análise dos *lifts* dos itemsets, foi possível analisar padrões de comodidades em diferentes contextos econômicos, oferecendo insights valiosos para o setor imobiliário. No futuro, desafios poderão incluir a integração de novas fontes de dados para enriquecer as análises e o desenvolvimento de modelos ainda mais sofisticados para capturar a dinâmica do mercado em constante evolução.

Referências

[1] L. Cerf, "Itemset mining," 2024, accessed: 2024-08-06. [Online]. Available: https://homepages.dcc.ufmg.br/~lcerf/slides/mda13.pdf

[2] Zumper, "Zumper national rent report," 2024, accessed: 2024-08-06. [Online]. Available: https://www.zumper.com/blog/rental-price-data/