Descoberta de Padrões em Dados de Apartamentos para Alugar nos EUA usando Itemsets Frequentes e Descoberta de Subgrupos

Alexis Mariz[1], **Bernnardo Seraphim**[1], **Gabriel Castelo Branco**[1], **Henrique R. S. Ferreira**[1], **Luisa Toledo**[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)

{alexismariz, beseraphim, gabriel.castelo, henriqueferreira, luisatoledo}@dcc.ufmg.br

Abstract. *This paper presents the project developed for the course Aprendizado Descritivo (Discriptive Learning), taught by Professor Renato Vimieiro. The project focused on applying discriptive learning techniques to a dataset of rental apartments in the United States. To discover relevant patterns, two approaches were employed: frequent itemset mining, focusing on the amenities offered by the apartments, and subgroup discovery.*

Resumo. *O presente artigo detalha o projeto da disciplina de Aprendizado Descritivo, ministrada pelo Prof. Dr. Renato Vimieiro, que teve como escopo a aplicação de técnicas de aprendizado descritivo a uma base de dados de apartamentos para locação nos Estados Unidos. Para a descoberta de padrões de interesse, foram utilizadas duas técnicas abordadas em aula: a mineração de itemsets frequentes, focada nas comodidades oferecidas pelos apartamentos, e a identificação de subgrupos relevantes.*

1. Introdução

O mercado imobiliário é um setor de grande interesse na economia mundial. Notadamente tratado como termômetro do sistema financeiro, ele também reflete diversas características das múltiplas regiões de um país. Além disso, atrai muitos investidores que buscam solidez e segurança para investir.

No campo da computação, a indústria imobiliária apresenta-se como uma fonte rica de dados, que pode servir de arcabouço para diversas tarefas de mineração de dados e descoberta de padrões. Os padrões descobertos podem, inclusive, ser utilizados para direcionamento de investimentos de maneira mais assertiva.

Tendo em vista o contexto apresentado, este relatório técnico descreve a tarefa de descoberta e análise de padrões relevantes em um conjunto de dados de apartamentos para aluguel nos Estados Unidos. O conjunto de dados está disponível em duas grandes plataformas: *UC Irvine* [UCI Machine Learning Repository 2019] e *Kaggle* [adithyaawati 2025].

O *dataset* é composto por aproximadamente 100.000 entradas, que descrevem anúncios retirados de *sites* populares para divulgação de apartamentos. Esses dados vieram predominantemente da plataforma online *RentDigs.com*; ou seja, das 99.442 entradas, 90.912 foram obtidas nesse local. O *RentDigs.com* é um *site* de anúncios de aluguel onde proprietários podem postar suas propriedades para locação gratuitamente. 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 mostrou-se interessante para a descoberta de padrões.

Após uma descrição exploratória da análise de dados na Seção 2, as Seções 3 e 4 abordarão as tarefas propostas. Finalmente, a Seção 5 resume os resultados encontrados e conclui o relatório.

2. Análise Exploratória de Dados

2.1. Visualização dos dados

Tendo em vista que uma das questões mais relevantes da 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 essa tarefa, inicialmente foi criada uma cópia do conjunto de dados, onde se descartaram 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 essas mesmas colunas, calculando a média do preço nos grupos. A visualização dos dados foi feita na forma de um mapa de calor, o que permitiu uma compreensão visualmente agradável.

O primeiro tratamento feito foi a criação de um agrupamento de coordenadas de latitude e longitude em supergrupos, alterando 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.

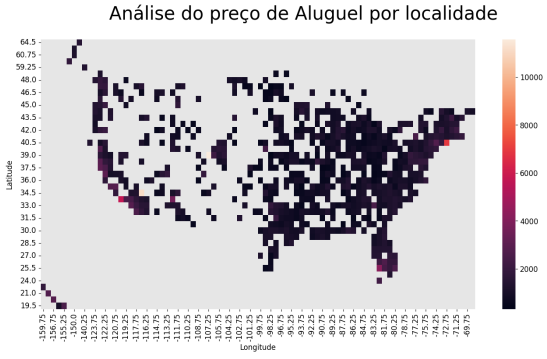

Figura 1. Mapa com agrupamento de Coordenadas e Outliers

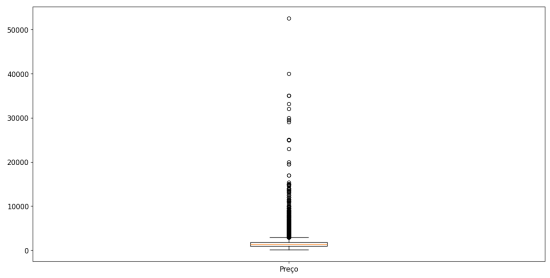

Figura 2. Boxplot com preços

Mesmo partindo de um mapa que não permite uma boa visualização devido à discrepância de preço em algumas regiões, essa visualização já nos possibilita perceber uma grande diferença 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.

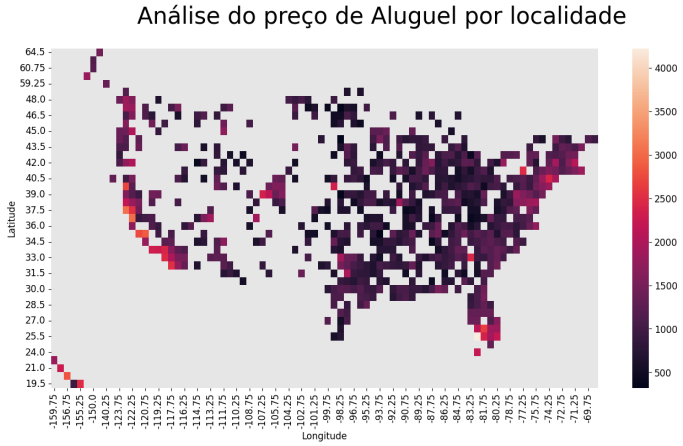

Figura 3. Heat Map Sem Outliers

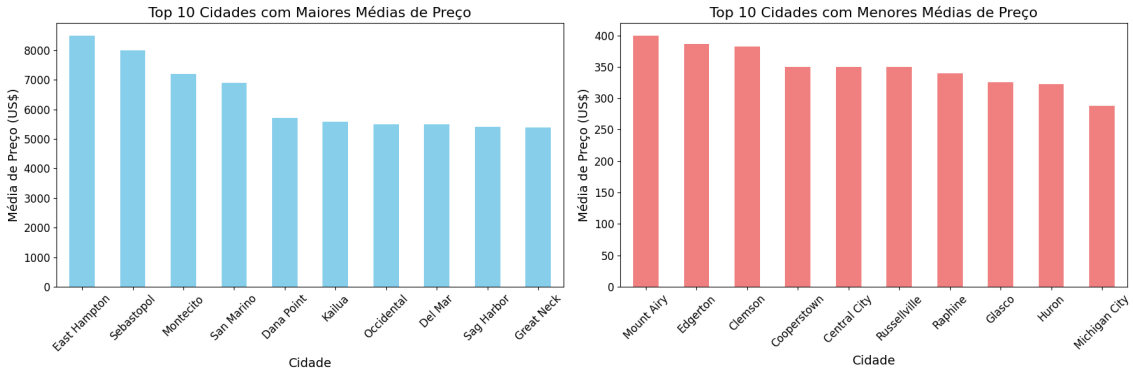

Figura 4. Gráficos de barras com 10 cidades mais caras e mais baratas

Os estados de economia mais marcante, ou com qualidade de vida mais alta, notadamente têm imóveis com preços mais elevados. Todavia, os valores exorbitantes dos *outliers* não permitem maiores deduções sobre essas visualizações. Conforme demonstrado na figura 2, a grande maioria dos valores se encontra abaixo de US$ 10.000,00, enquanto alguns se encontram acima dessa faixa. Portanto, um tratamento visando à remoção dos *outliers* foi aplicado. A remoção dos *outliers* se deu pela exclusão dos 5% mais altos e dos 5% mais baixos, por estado. Com isso, valores extremos foram desconsiderados.

Apesar da remoção dos *outliers*, verifica-se que os estados com maiores preços se mantêm de forma aproximada em suas posições, conforme o esperado.

Fica notável, portanto, uma concentração de apartamentos caros em regiões como Nova York (NY), Califórnia (CA) e Flórida (FL).

Um refinamento da exploração, após a remoção dos *outliers*, é descobrir as cidades mais caras e mais baratas do país, informação extremamente relevante para entender a distribuição de preços dos apartamentos de aluguel nos EUA.

Como pode ser visto na figura 4, as cidades mais caras de fato estão nos estados esperados (as 3 primeiras cidades estão em NY, CA e CA, respectivamente).

2.2. 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 influenciando uma disparada nos preços em um ponto específico do país. Finalmente, pudemos verificar, após a remoção dos *outliers*, uma consistência entre as cidades mais caras e os estados mais caros.

3. Mineração de Itemsets e Regras de Associação

Nesta seção, utilizamos a mineração de *itemsets* para identificar padrões significativos entre as comodidades oferecidas pelos apartamentos presentes no conjunto de dados. Também iremos gerar as regras de associação, considerando um *threshold* mínimo de 1.0 sobre o valor do *Lift*; essas duas técnicas permitem que diversas análises sejam feitas sobre os resultados. Para isso, vamos analisar três casos distintos: conjunto de dados completo, as 10 cidades mais caras e as 10 cidades mais baratas. É importante ressaltar que não usaremos as cidades mais caras e baratas do *dataset*, e sim de uma pesquisa feita pela *Zumper* [Zumper 2024], uma plataforma online que facilita a busca e o aluguel de imóveis.

3.1. Conjunto Completo

83.448 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* frequentes, utilizando o algoritmo Apriori [Agrawal and Srikant 1994], considerando uma frequência mínima de 10%. Também encontramos as regras de associação. Podemos visualizar os top 5 *itemsets* na Tabela 1 e as regras de associação na Tabela 2.

Tabela 1. Top 5 Itemsets com Maior Suporte no Conjunto de Dados Inteiro

Itemsets	Freq. Rel.
(pool, gym)	0.340476
(pool, parking)	0.251965
(gym, parking)	0.213414
(pool, patio/deck)	0.198099
(gym, washer dryer)	0.190646

Tabela 2. Top 5 Regras de Associação no Conjunto de Dados Inteiro

Antecedentes	Consequentes	Lift	Confiança
refrigerator	dishwasher	3.792461	0.754193
dishwasher	ac	2.830584	0.535161
gym, patio/deck	washer dryer	2.048958	0.637193
dishwasher	patio/deck	1.849401	0.587346
fireplace	washer dryer	1.838256	0.571668

3.2. As 10 Cidades Mais Caras

Continuando as análises, é interessante observar os casos das cidades mais caras e mais baratas dos EUA, pois queremos descobrir se existe diferença no que é regularmente ofertado nos anúncios. Assim, foram seguidos os mesmos passos da análise anterior, porém agora considerando apenas as 10 cidades com os preços de aluguel mais altos nos EUA, dentre as 100 cidades mais populosas do país. Esse conjunto de dados possui 5.499 transações das cidades Nova York, Jersey City, San Francisco, Boston, Miami, San Jose, Arlington, San Diego, Los Angeles, Chicago. Os resultados podem ser vistos nas Tabelas 3 e 5.

Tabela 3. Top 5 Itemsets com Maior Suporte Considerando as 10 Cidades Mais Caras

Itemsets	Freq. Rel.
(pool, gym)	0.260047
(gym, parking)	0.204765
(pool, parking)	0.189489
(gym, patio/deck)	0.169667
(patio/deck, parking)	0.162939

Tabela 4. Top 5 Itemsets com Maior Suporte Considerando as 10 Cidades Mais Baratas

Itemsets	Freq. Rel.
(pool, gym)	0.313031
(pool, clubhouse)	0.249292
(pool, parking)	0.239377
(gym, clubhouse)	0.228754
(gym, parking)	0.211048

Tabela 5. Top 5 Regras de Associação Considerando as 10 Cidades Mais Caras

Antecedentes	Consequentes	Lift	Confiança
refrigerator	dishwasher	2.983060	0.639576
dishwasher	ac	2.569804	0.600509
washer dryer	ac	1.765093	0.412465
pool, patio/deck	gym	1.731096	0.714286
pool	gym, patio/deck	1.681235	0.285250

3.3. As 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 1.412 transações das cidades Lincoln, El Paso, Memphis, Albuquerque, Tucson, Tulsa, Oklahoma City, Shreveport, Wichita, Akron. Os resultados podem ser vistos nas Tabelas 4 e 6.

3.4. Conclusão

Após todos esses dados fornecidos, podemos observar que, em geral, algumas comodidades são comuns para o país inteiro, como piscina, academia e estacionamento. Em contrapartida, também fica evidenciada uma divergência no que é mais comum nas cidades mais caras e nas mais baratas. Por exemplo, nas cidades caras as frequências relativas são mais baixas, o que pode indicar uma maior diversidade de comodidades. Ao mesmo tempo, podemos ver a presença da comodidade *patio/deck*; ou seja, nessas áreas é mais comum os imóveis possuírem essas áreas externas, algo que em geral é associado ao luxo

Tabela 6. Top 5 Regras de Associação Considerando as 10 Cidades Mais Baratas

Antecedentes	Consequentes	Lift	Confiança
pool, ac	gym, dishwasher	3.795277	0.596708
pool, gym, dishwasher	ac	3.074728	0.810056
ac, gym	pool, dishwasher	3.067726	0.617021
ac, refrigerator	dishwasher	3.045928	0.886598
refrigerator, patio/deck	dishwasher	3.034712	0.883333

e ao bem-estar.
Já nas cidades mais baratas, temos um suporte mais elevado dos *itemsets*, indicando que pode haver uma menor diversidade de amenidades. Isso ficou claro, pois amenidades como *luxury, doorman, alarm, golf* apareceram, respectivamente, 3, 1, 1 e 1 vezes. As cidades caras, em contraponto, excluindo a amenidade *golf*, não tiveram nenhuma amenidade que ocorreu menos de 13 vezes.

Além dos *itemsets*, podemos comparar as regras de associação entre os dois tipos de cidades. Ao fazer essa análise, fica evidenciado que em cidades mais baratas as associações envolvem mais de um item, o que pode sugerir que é comum existirem "pacotes de amenidades". Já nas cidades caras, as associações de um item geralmente geram outro item, ou seja, são mais diretas. É importante mencionar que, em geral, as métricas de *Lift* foram altas, mostrando que a chance dos consequentes acontecerem ficou muito maior caso o antecedente aconteça. E os valores de *Confiança* mostraram que as regras eram relativamente confiáveis, e não porque o consequente era muito comum, por mais que alguns casos sejam assim.

4. Descoberta de Subgrupos

A última parte deste trabalho foi realizar uma descoberta de subgrupos de interesse nos dados, buscando entender fatores locais que impactam nos preços. Para isso, utilizamos a biblioteca *pysubgroup* [Lemmerich and Becker 2018] e definimos um *target* sobre a variável preço do aluguel. Além disso, é importante ressaltar que, para essa análise, foram utilizadas as colunas: *amenities*, *bathrooms*, *bedrooms* e *state*. Para a realização desta tarefa, foi necessário adotar uma abordagem dividida por iterações, pois, ao longo do processo de descoberta de subgrupos, percebeu-se a influência de alguns itens que dominavam a caracterização dos subgrupos. Ao todo, foram feitas três iterações sobre o conjunto de dados.

4.1. Primeira Iteração

Em um momento inicial, foi realizada uma busca de subgrupos de interesse no conjunto de dados completo, tendo como variável alvo a maximização do preço de aluguel. Os subgrupos foram buscados nas colunas descritas na introdução dessa seção. O algoritmo escolhido para a tarefa foi o *Beam Search*, conforme implementado na biblioteca *pysubgroup*.

O uso dos dados em sua completude, todavia, mostrou-se problemático no âmbito da descoberta de subgrupos. Na Tabela 7, é possível ver como os subgrupos retornados são redundantes, dando grande importância para o estado da Califórnia, o que resultou na

desconsideração dos demais atributos. Estes resultados, portanto, não revelaram nenhum subgrupo de interesse.

4.2. Segunda Iteração

Buscando diminuir o viés regional encontrado na seção 4.1, na segunda iteração, a variável alvo foi alterada para o *desvio padrão*, que indica a quantos desvios padrões o apartamento está da média da cidade. A intenção por trás dessa mudança foi capturar características intrínsecas ao apartamento que fizessem seu preço ser diferente dos demais naquela localização.

Fica evidente, através dos resultados demonstrados na Tabela 8, que os subgrupos resultantes da normalização estabelecida trouxeram muito mais significado à tarefa de exploração. Todavia, o problema de subgrupos aparentemente redundantes, como nos casos de *bathrooms* ≥ 2.50, permaneceu. Além disso, a métrica de qualidade nativa da biblioteca *pysubgroup* não estava apresentando resultados satisfatórios.

4.3. Terceira Iteração

Mesmo após notáveis melhorias, ainda havia redundâncias explícitas nos subgrupos. Ao investigar de maneira mais profunda a implementação da métrica de qualidade utilizada pela biblioteca *pysubgroup*, percebeu-se uma diferença entre a métrica documentada, que dividia o tamanho do subgrupo pelo tamanho do conjunto de dados, e o que estava de fato implementado no código, que ignorava a normalização pelo tamanho total do conjunto.

Dessa forma, realizou-se a alteração da implementação da métrica de qualidade, fazendo com que ela refletisse a população relativa. Além disso, visando reduzir ainda mais a redundância entre os subgrupos descobertos, implementou-se uma penalização dos subgrupos semelhantes, através da *Similaridade de Jaccard*, que mede a similaridade entre dois conjuntos. Os subgrupos que interseccionassem subgrupos já escolhidos seriam, portanto, descartados.

Como consequência das alterações descritas, foi possível encontrar subgrupos menos redundantes. Ademais, os subgrupos encontrados possuíam descrições mais simples e, portanto, mais compreensíveis em um cenário real. Os resultados descobertos, explicitados na Tabela 9, refletem, inclusive, cenários notadamente conhecidos, onde se correlaciona um número maior de quartos e banheiros com um preço maior de casas e apartamentos.

Tabela 7. Subgrupos Iteração I

SG	Descrição	Qualidade	Tamanho
1	Playground==0 AND state==’CA’	95583.33	9790
2	state==’CA’	95245.22	10311
3	Basketball==0 AND Playground==0 AND state==’CA’	95180.46	9558
4	Basketball==0 AND state==’CA’	95056.33	9964
5	Playground==0 AND Tennis==0 AND state==’CA’	94033.86	9242

Tabela 8. Subgrupos Iteração II

SG	Descrição	Qualidade	Tamanho
1	$bathrooms \geq 2.50$ AND state=='CA'	54578.03	526
2	$amenity\_Elevator == 1$ AND $bedrooms \geq 4.0$	52235.98	32
3	$amenity\_Playground == 0$ AND $bathrooms \geq 2.50$	52113.61	3109
4	$bathrooms \geq 2.50$	51485.72	3370
5	$amenity\_Basketball == 0$ AND $bathrooms \geq 2.50$	51411.01	3276

Tabela 9. Subgrupos Iteração III

SG	Descrição	Qualidade	Tamanho	Desvio Padrão
1	bathrooms: [2.0:2.50[	0.2535	36263	0.5096
2	bedrooms: [3.0:4.0[	0.1467	10373	0.7080
3	amenity_Parking==1 AND bathrooms: [2.0:2.50[	0.1620	16423	0.5671
4	amenity_Parking==0 AND bathrooms: [2.0:2.50[	0.1507	19840	0.4620

5. Conclusão

Este projeto aplicou técnicas de mineração de dados para extrair informações de um conjunto de dados sobre aluguel de apartamentos nos Estados Unidos, com foco na identificação de padrões significativos de comodidades oferecidas pelos apartamentos.

A análise exploratória inicial nos permitiu encontrar as cidades em que os apartamentos de maior e menor preço se concentram e também nos mostrou como os *outliers* influenciam na visualização dos dados. Com isso, foi possível verificar que os estados com maiores preços são Nova York (NY), Califórnia (CA) e Flórida (FL).

Em seguida, o algoritmo Apriori foi aplicado para realizar a mineração de *itemsets* em 3 casos distintos: *dataset* completo, 10 cidades mais caras e 10 cidades mais baratas, e em todos houve destaque para um mesmo conjunto de comodidades frequentes: piscina, academia e estacionamento. Além disso, os apartamentos das cidades mais caras apresentam uma frequência maior das amenidades *patio/deck*, enquanto os dados sugerem que as cidades mais baratas apresentam menor diversidade de amenidades. Já as regras de associação obtidas com alta confiança e *lift* envolvendo mais de um item indicam a existência de "pacotes de amenidades" nas cidades mais baratas.

Por fim, a descoberta de subgrupos foi realizada sobre a variável de preço do aluguel utilizando o algoritmo Beam Search. Durante a tarefa, foram necessários alguns refinamentos, pois o estado da Califórnia dominou a caracterização dos subgrupos. Primeiro, a variável alvo foi alterada para o *desvio padrão*, que indica a quantos desvios padrões o apartamento está da média da cidade, visando extrair características que levam um apartamento a ter um preço diferente da sua cidade. Com isso, os subgrupos deixaram de ser dominados pela Califórnia, mas ainda se mantiveram redundantes, dominados pela quantidade de banheiros superior a 2.5. Na última iteração, a métrica de qualidade do algoritmo foi alterada para refletir a população relativa e penalizar subgrupos semelhantes, obtendo resultados menos redundantes.

6. Referências

Referências

adithyaawati (2025). Apartments for rent classified. Dataset disponível no Kaggle.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*, pages 487–499, Santiago, Chile. Morgan Kaufmann Publishers Inc.

Lemmerich, F. and Becker, M. (2018). pysubgroup: Easy-to-use subgroup discovery in python. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 658–662.

UCI Machine Learning Repository (2019). Apartment for rent classified. Dataset disponível via UCI Machine Learning Repository.

Zumper (2024). Annual rent report 2024.