ISSN 2558-6052

COLETA E ANÁLISE DE DADOS DO AIRBNB

Luis Henrique Ferreira Costa[1]; Jônatas Freire[2]; Michele A. Brandão[3]

1 Luis Henrique Ferreira Costa, Bolsista (IFMG), Técnico Integrado em Informática, IFMG Campus Ribeirão das Neves, Ribeirão das Neves - MG; luishenrique30102005@yahoo.com
2 Jônatas Freire, Bacharelado em Sistemas de Informação, IFMG Campus Sabará, Sabará – MG
3 Michele A. Brandão, Pesquisadora do IFMG, IFMG Campus Ribeirão das Neves, Ribeirão das Neves – MG

RESUMO

O Airbnb é uma plataforma online com mais de 6,6 milhões de anúncios e 1,4 bilhões de hóspedes originados de diferentes localidades. Diante de um número tão acentuado de usuários, essa plataforma gera um grande volume de dados que podem ser utilizados em diferentes aplicações, desde a análise de impactos econômicos até a recomendação de imóveis. Dentre as diversas funcionalidades e recursos presentes no Airbnb, essa plataforma fornece aos proprietários a oportunidade de listar suas propriedades e oferece aos viajantes uma grande seleção de acomodações exclusivas e acessíveis. O impacto transformador do Airbnb na indústria de viagens, na economia compartilhada e na economia global é evidente em sua ampla popularidade e nos vários estudos acadêmicos que examinam sua influência. Além do impacto econômico, o Airbnb gera dados de imóveis, usuários (proprietários e viajantes), cidades, entre outros, que podem ser utilizados para análises de outros impactos gerados pelo uso dessa plataforma, por exemplo, no turismo. Por isso, este trabalho apresenta um conjunto de dados com imóveis brasileiros do Airbnb e respectivas avaliações, que foi coletado por meio do uso de uma biblioteca em Python. Vale destacar que o conjunto de dados construído possui imóveis de 10 cidades consideradas empreendedoras e 10 cidades identificadas como turísticas. As cidades São Paulo e Florianópolis estão em ambos os ranqueamentos, por isso, o conjunto de dados proposto possui anúncios de imóveis de 18 cidades diferentes. Aqui, esse conjunto de dados é descrito e caracterizado com o intuito de facilitar seu uso em outros estudos. Além disso, são apontadas possíveis aplicações e limitações do conjunto de dados construído. A análise inicial do conjunto de dados revela que ele pode ser ainda mais explorado em diferentes estudos. Como trabalhos futuros, planeja-se coletar informações sobre os usuários no Airbnb e também em diferentes períodos ao longo do ano.

INTRODUÇÃO:

O Airbnb[1] é uma plataforma online, estabelecida em 2008, possui mais de 6,6 milhões de anúncios ativos em mais de 220 países e regiões. Em dezembro de 2022, essa plataforma recebeu mais 1,4 bilhões de hóspedes[2]. Por esse volume de usuários, o Airbnb tem sido alvo de muitos estudos. Por exemplo, [Jain et al. 2021] utilizam diferentes dados dessa plataforma para quantificar e acompanhar mudanças em determinadas regiões.

Dentre as diversas funcionalidades e recursos presentes no Airbnb, essa plataforma fornece aos proprietários a oportunidade de listar suas propriedades e oferece aos viajantes uma grande seleção de acomodações exclusivas e acessíveis. O impacto transformador do Airbnb na indústria de viagens, na economia compartilhada e na economia global é evidente em sua ampla popularidade e nos vários estudos acadêmicos que examinam sua influência [Ding et al. 2023]. Além do impacto econômico, o Airbnb gera dados de imóveis, usuários (proprietários e viajantes), cidades, entre outros, que podem ser utilizados para análises de outros impactos gerados pelo uso dessa plataforma, por exemplo, no turismo [Jordan et al. 2023] e na qualidade de vida da vizinhança de imóveis anunciados no Airbnb [Mody et al. 2021].

Nesse contexto, este trabalho apresenta um conjunto de dados com imóveis brasileiros do Airbnb e respectivas avaliações. Vale destacar que esse conjunto de dados possui imóveis de 10 cidades consideradas empreendedoras e 10 cidades identificadas como turísticas. As cidades São Paulo e Florianópolis estão em ambos os ranqueamentos, por isso, o conjunto de dados proposto possui anúncios de imóveis de 18 cidades diferentes.

METODOLOGIA:

[1] Airbnb: https://www.airbnb.com.br/
[2] Sobre o Airbnb: https://news.airbnb.com/br/about-us/

ISSN 2558-6052

Esta seção apresenta as principais etapas para a construção do conjunto de dados do Airbnb, conforme mostra a Figura 1. Essas etapas consistem na seleção das cidades a terem os anúncios coletados, seleção das datas para coleta dos anúncios, coleta de dados do Airbnb e organização e armazenamento desses dados.

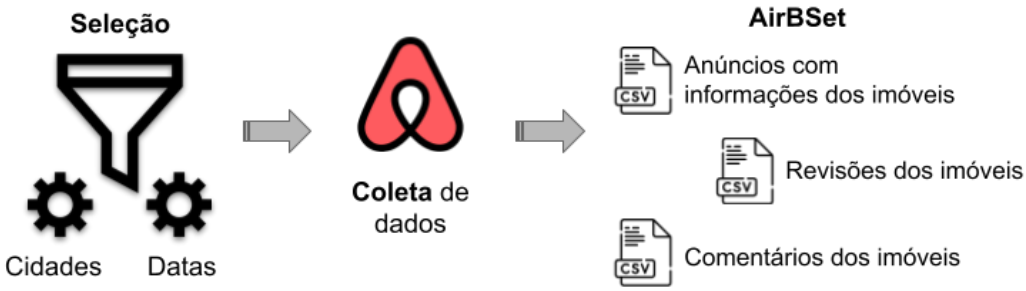

Figura 1. Principais etapas da metodologia para realização deste trabalho.

Seleção das cidades. No Airbnb, há anúncios de imóveis localizados em mais de 100 mil cidades ao redor do mundo[3], o que dificulta uma coleta total desses anúncios. Por isso, neste trabalho, foram selecionados anúncios de imóveis pertencentes a um top 10 cidades brasileiras consideradas empreendedoras e um top 10 cidades brasileiras apontadas como turísticas. A Tabela 1 apresenta o ranqueamento dessas dez cidades brasileiras empreendedoras de acordo com estudo conduzido pelo Enap (Escola Nacional de Administração Pública)[4] em 2021 e turísticas segundo o IBGE (Instituto Brasileiro de Geografia e Estatística)[5] em 2022.

Tabela 1. Top 10 cidades empreendedoras em 2021 e tur.sticas em 2022.

Top 10 cidades		
Posição	**Empreendedoras**	**Turísticas**
1º	São Paulo - São Paulo	Rio de Janeiro - Rio de Janeiro
2º	Florianópolis - Santa Catarina	São Paulo - São Paulo
3º	Curitiba - Paraná	Gramado - Rio Grande do Sul
4º	Vitória - Espírito Santo	Ubatuba - São Paulo
5º	Belo Horizonte - Minas Gerais	Porto Seguro - Bahia
6º	Porto Alegre - Rio Grande do Sul	Florianópolis - Santa Catarina
7º	São José dos Campos - São Paulo	Fortaleza - Ceará
8º	Osasco - São Paulo	Natal - Rio Grande do Norte
9º	Joinville - Santa Catarina	Porto de Galinhas - Pernambuco
10º	Cuiabá - Mato Grosso	Campos do Jordão - São Paulo

Seleção das datas de coleta. O Airbnb é uma plataforma que possibilita apenas a busca por imóveis selecionando datas posteriores ao dia corrente. Portanto, a coleta foi realizada selecionando feriados do ano de 2023 após o dia 01 de junho de 2023. Assim, os dados dos imóveis foram coletados para dois feriados, de Corpus Christi (de 08 de junho de 2023 até 11 de junho de 2023) e Natal (de 24 de dezembro de 2023 até 25 de dezembro de 2023).

Coleta de dados do Airbnb. Após a seleção das cidades e períodos a serem coletados, utilizou-se o pacote do Python chamado Selenium[6], esse mesmo pacote também foi utilizado para coletar dados em [Silva et al. 2021]. Essa biblioteca automatiza interações em um navegador web e, assim, simula um usuário acessando à plataforma Airbnb e buscando por um imóvel na cidade e período passados como parâmetro. Em seguida,

[3] Sobre o Airbnb: https://news.airbnb.com/br/about-us/
[4] Muda o ranking de melhores cidades para empreender no Brasil: bit.ly/43A3WMG
[5] IBGE confirma atividade turística como importante indutora da economia brasileira: bit.ly/3Cju7Lq
[6] Selenium: https://pypi.org/project/selenium/

seminário de iniciação científica
ISSN 2558-6052

os dados retornados sobre os imóveis e respectivas avaliações são armazenados em arquivos no formato CSV (Comma Separated Values).

Organização e armazenamento dos dados. Cada arquivo CSV foi guardado em pastas separadas, uma pasta por cidade, ou seja, um total de 18 pastas, e duas pastas para os feriados, uma para cada feriado, dentro da pasta da cidade. Ao todo, o conjunto de dados possui 108 arquivos CSV. A Tabela 2 apresenta os atributos presentes em cada arquivo coletado sobre os anúncios, revisões e comentários. É importante destacar que as revisões são um agregado de notas nos anúncios, ou seja, são notas calculadas para um imóvel com base nas revisões dos usuários para tal imóvel.

Tabela 2. Descrição das informações coletadas sobre os imoveis e respectivos atributos.

Sobre o Imóvel	Atributos
Anúncio	id, url, informação, avaliação, local, preço e comodidades
Revisão	id, limpeza, exatidão do anúncio, comunicação, localização, check-in e custo benefício
Comentário	id, id_quarto, nome, comentário, data e id_usuário

RESULTADOS E DISCUSSÕES:

Esta seção apresenta uma breve caracterização dos dados presentes no conjunto de dados. Em particular, a Tabela 3 descreve a quantidade de anúncios, revisões e comentários coletados para cada cidade, agrupados por cidades consideradas turísticas e empreendedoras, apenas empreendedora e apenas turística. As cidades estão ordenadas de acordo com o ranqueamento apresentado na Tabela 1. Como coletamos informações para dois feriados, cuidamos para que anúncios presentes nos dois períodos coletados não fossem contabilizados duas vezes. É importante destacar que a quantidade de anúncios e revisões são iguais pelo fato das revisões serem gerais para cada anúncio.

Tabela 3. Estatísticas sobre os anúncios, revisões e comentários presentes no conjunto de dados.

Tipo	Cidade	# de Anúncios	# de Revisões	# de Comentários
Turística e Empreendedora	São Paulo	344	344	1639
	Florianópolis	347	347	1623
	Total	691	691	3.262
	Média	346	346	1.631
Empreendedora	Curitiba	349	349	1.672
	Vitória	314	314	1.441
	Belo Horizonte	363	363	1.731
	Porto Alegre	277	277	1.376
	São José dos Campos	115	115	517
	Osasco	276	277	1.315
	Joinville	258	258	1.291
	Cuiabá	305	305	1.537
	Total	2.257	2.257	10.880
	Média	282	282	1.360
Turística	Gramado	283	283	1.289
	Ubatuba	348	348	1.479
	Porto Seguro	137	137	580
	Rio de Janeiro	365	365	1.622
	Fortaleza	341	342	1.623
	Natal	167	167	772
	Porto de Galinhas	274	274	1.206
	Campos do Jordão	251	251	1.043
	Total	1.825	1.825	9.614
	Média	228	228	1.201
Total Geral	—	4.773	4.773	23.756
Média Geral	—	265	265	1.319

ISSN 2558-6052

De maneira geral, na Tabela 3, é possível observar que as cidades consideradas empreendedoras e turísticas (São Paulo e Florianópolis) possuem mais anúncios, revisões e comentários em relação às cidades apenas empreendedoras ou apenas turísticas. Também é possível notar que as cidades empreendedoras juntas possuem cerca de 432 anúncios a mais que as cidades turísticas. Isso pode ser justificado pelo grande porte dessas cidades que acabam sendo turísticas também. Além disso, das cidades empreendedoras, São José dos Campos é a que possui menor quantidade de anúncios e revisões (115) e comentários (517), e Belo Horizonte é a que possui maior quantidade de anúncios e revisões (363) e comentários (1.731). Já para as cidades turísticas, Porto seguro é a que possui menor quantidade de anúncios e revisões (137) e comentários (580), e Rio de Janeiro é a que possui maior quantidade de anúncios e revisões (365), mas possui um comentário (1.622) a menos que Fortaleza (1.623).

CONCLUSÕES:

A API (Application Programming Interface ou Interface de Programação de Aplicação, em português) do Airbnb[7] fornece limitação de coleta de apenas 300 anúncios por conta. Por isso, foi necessário desenvolver um coletor, conforme descrito na metodologia. Entretanto, esse coletor não possibilita a coleta de dados de anúncios em datas retroativas ao dia da coleta. Assim, a coleta de datas retroativas é um desafio que gera uma limitação no conjunto de dados de não conter esse tipo de informação. Por exemplo, não foi possível obter dados do período da pandemia de COVID-19 e, assim, não é possível uma comparação entre os períodos de locação no Airbnb antes, durante e após esse acontecimento. Um outro desafio também é obter informações sobre os usuários do Airbnb, tanto de anunciantes quanto de revisores. Após diversas análises nos dados, foi possível coletar o campo identificador do usuário, mas o conjunto de dados descrito neste trabalho ainda não tem informações mais detalhadas sobre esses usuários.

Finalmente, este trabalho apresentou um conjunto de dados com imóveis brasileiros e respectivas avaliações de 18 cidades brasileiras, divididas em turísticas e empreendedoras. Esse conjunto de dados foi descrito e caracterizado com o intuito de facilitar seu uso em outros estudos. Também descrevemos possíveis aplicações, bem como os desafios e possíveis limitações. Em trabalhos futuros, planeja-se a coleta de mais períodos, incluindo feriados e dias úteis. Isso possibilitará o uso dos dados para um estudo mais amplo, principalmente, comparando cidades empreendedoras e turísticas. Também planeja-se coletar informações sobre os usuários do Airbnb.

REFERÊNCIAS BIBLIOGRÁFICAS:

Ding, K., Niu, Y., and Choo,W. C. (2023). The evolution of airbnb research: A systematic literature review using structural topic modeling. Heliyon, page e17090.

Jain, S., Proserpio, D., Quattrone, G., and Quercia, D. (2021). Nowcasting gentrification using airbnb data. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1):1–21.

Jordan, E. J., Vieira, J. C., Santos, C. M., and Huang, T.-Y. (2023). Do residents differentiate between the impacts of tourism, cruise tourism, and airbnb tourism? Journal of Sustainable Tourism, 31(2):265–283.

Mody, M., Suess, C., and Dogru, T. (2021). Does airbnb impact non-hosting residents' quality of life? comparing media discourse with empirical evidence. Tourism Management Perspectives, 39:100853.

Silva, M. O., Scofield, C., and Moro, M. M. (2021). Pportal: Public domain portuguese language literature dataset. In Anais do III Dataset Showcase Workshop - SBBD, pages 77–88. SBC.

[7] API do Airbnb: https://www.airbnb.com/partner