UC Davis Graduate School of Management
BAX 422 Winter Quarter, 2023

Powering rental bookings for Airbnb's biggest competitor: VRBO

Collecting the data fuel for our machine learning engine

Submitted By:
Yuto Takeda, Tamalika Basu, Shulang (Simon) Ning

Table of Contents

Executive Summary

One of the key features of operating in online short-term rental businesses is ensuring that both strata of customers operating on the website, i.e. users looking to book a property and property owners wanting to rent their resources, have a pleasant experience. To augment the given use case in hand, we undertook the project to solve it in two halves. First, the project aims to scrape data from VRBO[1].com, one of Airbnb's largest competitors, to collect different attributes that determine the pricing for properties, across famous cities in the US such as San Francisco, New York, Seattle, and Texas, among others. Second, the collected data will be used to perform two major sub-tasks: first, developing an ensemble of machine learning algorithms to analyze pricing trends on VRBO.com, and further, using the models to predict the prices of different types of rental properties based on factors such as amenities, number of bedrooms, location of the property, and many more. The second sub-task is to use the algorithms to provide personalized recommendations for owners based on customer reviews, in order for them to improve their property listings for better profitability and increased booking chances. In this report, we will dive deep into identifying our source of data, gathering the required information for our analysis, and finally preparing the data for business readiness.

The business value that we aim to create using this project is to improve users' experience and thus boost engagement and popularity among customers. The insights generated from this project in the end will be beneficial for home rental business owners, enabling them to make informed decisions regarding prices and amenities to optimize their rental properties for the market. Additionally, the project's focus on popular tourist destinations in the US makes it particularly relevant for the hospitality industry, as it allows stakeholders to gain valuable insights into consumer behavior and preferences. Overall, this project has the potential to contribute to the growth and evolution of the short-term rental market, while also providing significant value to users and industry stakeholders.

Introduction of the Domain

The vacation rental industry is a rapidly growing industry with significant potential for innovation and disruption. The rise of short-term rental platforms like Airbnb and VRBO has revolutionized the way people travel and find accommodations, making it easier and more affordable for travelers to find comfortable and affordable accommodations while also providing property owners with new ways to monetize their properties. According to Allied² market research, the global vacation rental market size was evaluated at $91.2 billion in 2021 and is estimated to reach $315 billion by 2031. Further, Statistica² says that over 1,000 distinct properties were available only within San Francisco during April 2022, thereby projecting the user base to be over 62.99 million by 2027.

However, this industry also faces many challenges, such as fluctuating demand, varying pricing strategies, and the need to provide personalized experiences for customers. By combining the power of data (in BAX 422) and machine learning techniques (in BAX 452), we can help property owners optimize their rental properties for the market while enhancing the user experience for renters. The insights generated from this project can help industry stakeholders better understand consumer behavior and preferences, which can lead to more efficient and effective rental markets. Hence, the chosen domain provides an exciting and challenging context for data-driven innovation and value creation.

The Scraping Process

As part of the data design part of this project, firstly we shortlisted the top ten popular cities within the US that are among the largest market place for VRBO, namely: San Francisco, Las Vegas, Los Angeles, New York, Chicago, Boston, Miami, Orlando, Honolulu, and Washington DC. We scraped essential features, as mentioned below, from all of the properties listed in these cities using relevant Python libraries (BeautifulSoup, Selenium). Finally, we designed a database on MongoDB to store the data for driving further business insights using Machine Learning. The next part of our project to analyze prices and recommendations has been covered in the report for BAX 452.

- **Standardize the data:** In order to ensure our data is standardized for further analysis, we decided to use a *per-day rate* of different properties for convenient comparison (1st April ‘23 to 2nd April ‘23). Additionally, we based our search results on a standard *2 adults, 0 children per room* basis, and also included a filter to *exclude pets,* as the inclusion of pets is a highly relative situation based on the type of pet (big, small, noisy, non-noisy)
- **Defining the main loop:** The first essential loop in our code is a *for loop,* which first loops through a *list* containing all the cities required, followed by 10 pages under each city which comprises the property listings. Next, a *sub for loop* runs through a total of 50 properties on each of the 10 pages. Finally, in case the loops fail to access any of the mentioned information, an *exception handling* ensures the smooth running of the code
- **Saving web pages to local:** To prevent ourselves from either being blocked on the website for scraping heavy loads of data or ensure we are able to scrape details from a static web page, we downloaded all properties across the 10 cities in our local system. Using *selenium,* we find the appropriate class identifier to navigate to the web page of that property. Next, we account for one city and one page at a time and make the *selenium* driver click on properties one-by-one and download the *HTML* version of the property web page. In the end, we have a total of 2506 downloaded pages (out of which only 2503 seem to contain information, and 3 are blank pages potentially due to some website error)
- **Accessing property features:** Here, we finally extract the different features under each property that will be used in the analysis later. Using *beautiful soup*, we identify the necessary *HTML tag identifiers* that will give us the targetted information, as mentioned below:
    - Rank: Describes the rank of the property listing under each city, based on its popularity amongst customers
    - Name: Describes the title of the property for identification
    - VRBO_City: Describes the city under which the property is located

- VRBO_Text: Describes the details about the property as updated by the owner. This section gives an overview of the uniqueness and offerings of the listing to a customer
- VRBO_Type: Describes the type of property offered by VRBO namely, hotel, studio, guest house, apartment, house, building, hotel suite, condo, resort, cottage, townhome, yacht, villa, estate, recreational vehicle, bungalow, cabin, houseboat, hostel, corporate apartment, mobile home, and boat
- Number_of_Bedrooms: Describes the number of bedrooms available in the property
- Star Rating: Describes the rating of the property as rated by the visitors
- VRBO_Near: Describes the top 6 famous tourist attractions or important landmarks around the property (typically within 0-3 miles range)
- VRBO_Price: Describes the price of the property based on occupancy for 1 day, 2 adults, 0 children, and 0 pets
- VRBO_Number_Images: Describes the number of images posted by the owner of a property. More images would indicate more confidence from the owner, in showcasing the property to ensure transparency between the owner and the customer
- VRBO_Area_SQ: Describes the area covered by the property in square feet
- Number_Beds: Describes the number of beds available in the property
- Number_Sleepers: Describes the sleeping capacity available in the property
- Number_Bathrooms: Describes the number of bathrooms available on the property
- Number_Baths: Describes the number of baths available on the property
- Number_Reviews: Contains all the reviews as shared by visitors on the property
- Reviews_Text: Contains a concatenated format of all the reviews for a particular property, separated by '|||' to distinguish between two different reviews
- Number_Amenities: Describes the number of amenities supported for each property, namely, microwave, fridge, room heater, hairdryer, etc.

- Amenities_Text: Describes the type of amenities supported for each property, namely, microwave, fridge, room heater, hairdryer, etc.
- Number_Facilities: Describes the type of facilities available namely, shower, sofa, dining table, etc.
- Facilities_Text: Describes the number of facilities available namely, shower, sofa, dining table, etc.

Here is a snippet of the property, wherein all the different features extracted are highlighted for ease of understanding:

Fig 1: Capturing property name, booking date options, price, images, city, guests, and property features

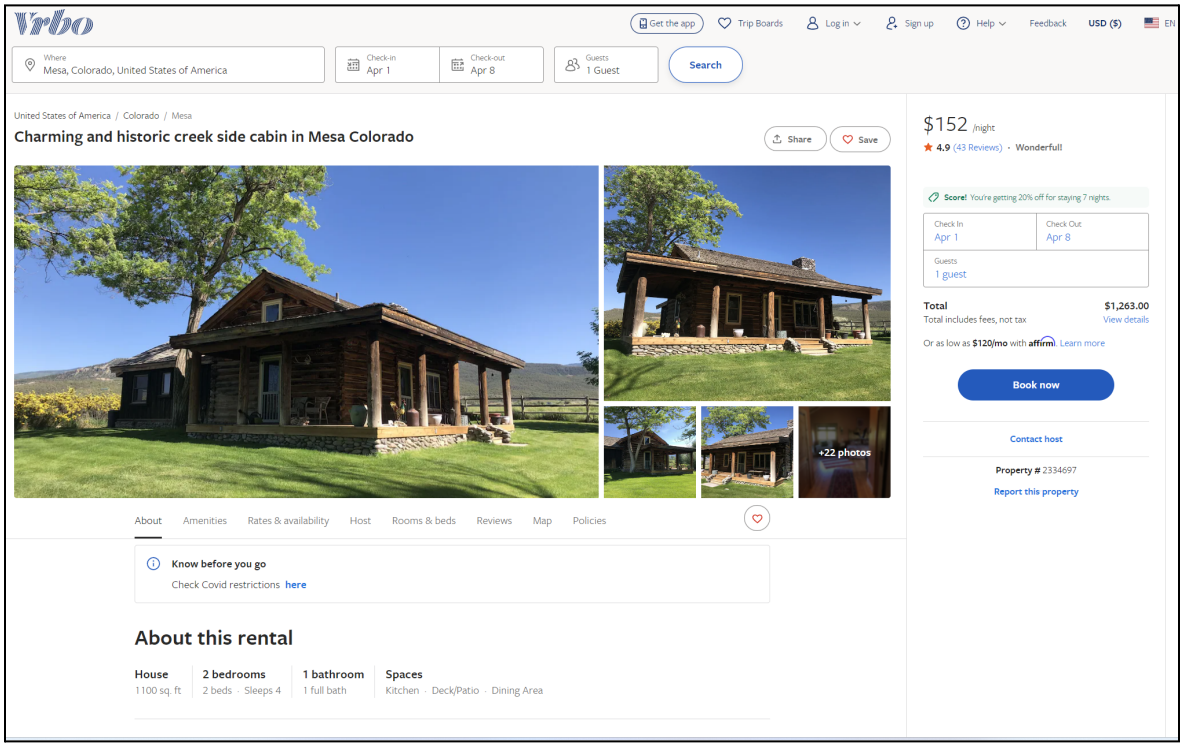


Fig 2: Capturing the property, amenities, host information, property rating, number of reviews

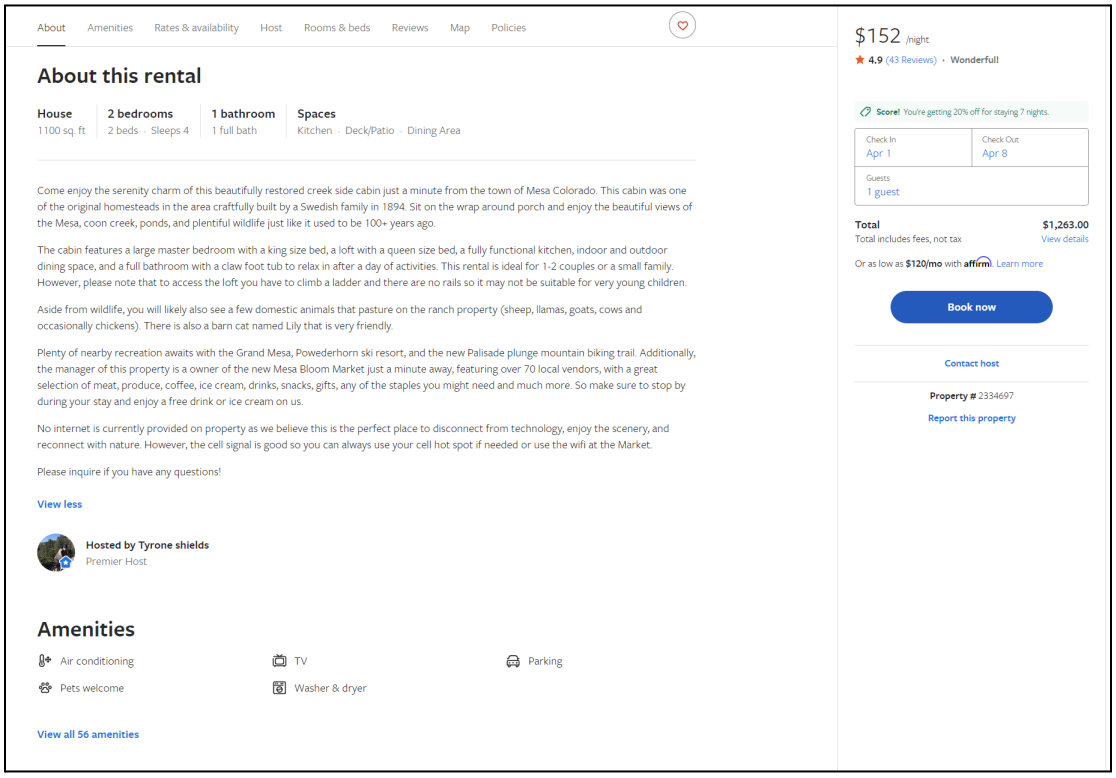


Fig 3: Capturing facilities, review texts for the property

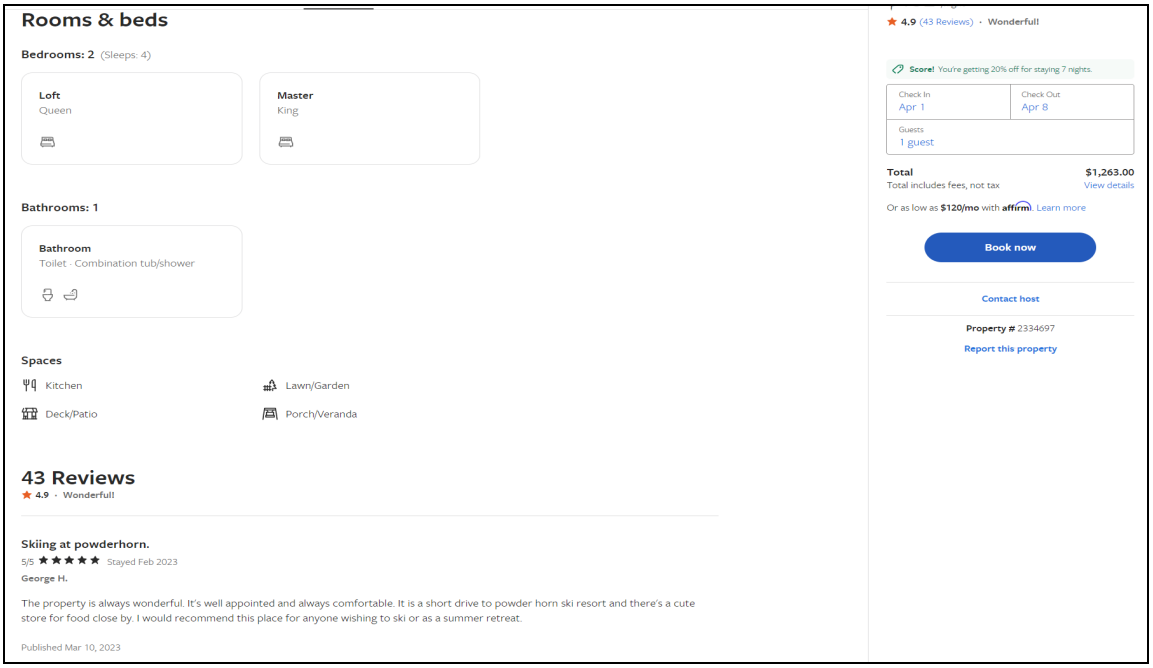
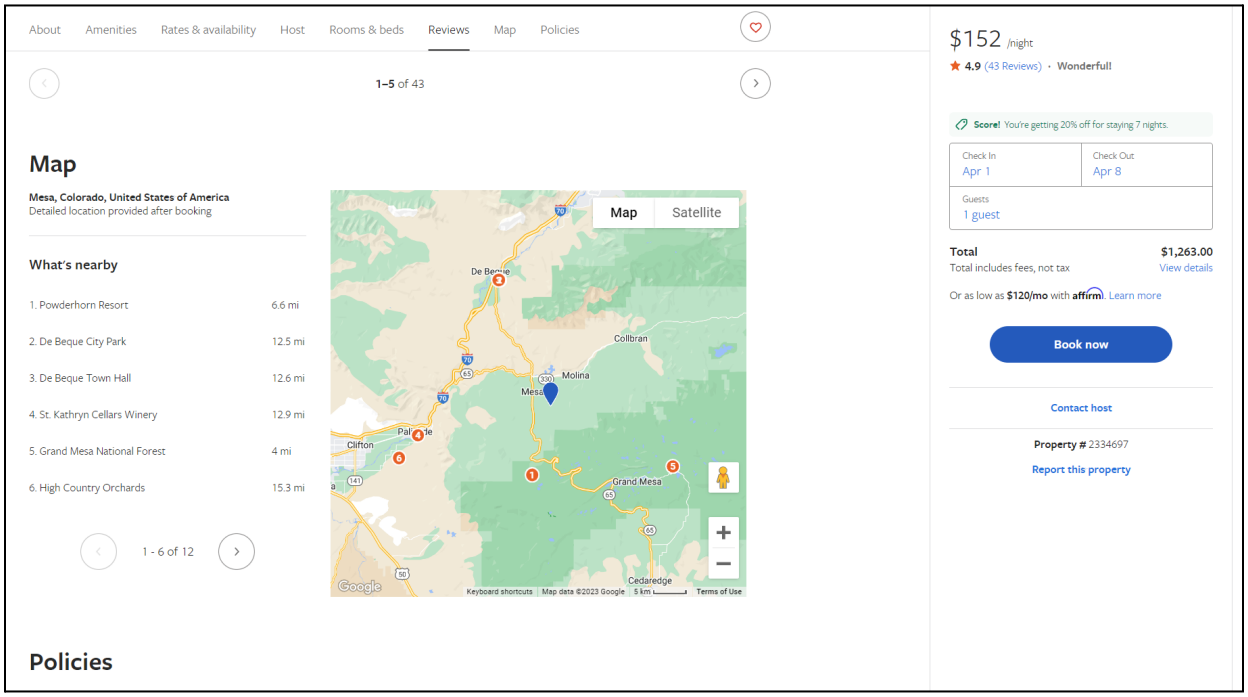


Fig 4: Capturing nearby place and high-level location of the property



- **Collecting the Data:** After extracting the aforementioned details using *beautiful soup,* we store the details in a *list of python dictionaries,* which can finally be pushed into the *MongoDB database*

Database Design

We chose MongoDB over MySQL because of two important reasons, first, the ability of MongoDB to scale horizontally and handle large amounts of data and if needed, distribute across multiple servers and ensure fast processing. Second, MongoDB's flexibility as a document-oriented database to store data in a schema-less format. This allows for easy modification of data structures, without having to change the underlying database and thus more adaptability to changing data requirements.

The different features for properties across 10 cities are firstly stored in the MongoDB database under a collection called *"vrbo"*. However, on inspecting the data from the websites, we observed that a lot of

these data were either not in a business consumable format, or, present in lists and nested list elements, or, included several noise characters in the texts present. To tackle this problem, we decided to perform some cleaning operations using *regex.* We also ensured that in the case of nested information in fields such as amenities, facilities, and reviews, we are able to extract such nested data to *CSV formats* hence we un-nested the nested information and parsed the different texts using a pipe symbol "|||" so that on exporting the files to *CSV formats,* the data is not distorted.

Therefore, the code first generates a collection called "*vrbo*" which contains the unformatted fields of different features. We perform the data formatting exercises in *python* for each field*,* in order to generate updated formats consumable for business analysis and store the final version of the business-ready data in a new collection called *"vrbo_formatted"*. The final database created contains 2503 rows and 21 fields and this collection has the following breakdown of entries per city:

San Francisco (290), Las Vegas (316), Los Angeles (326), New York (280), Chicago (133), Boston (236), Miami (183), Orlando (178), Honolulu (397), and Washington DC (167).

Conclusion and Next Steps

In this report, we gathered the data required for our analysis by performing routine web scraping exercises on VRBO.com and collected vital attributes that may potentially determine the pricing of short-term rental properties in popular tourist destinations in the US. This data will be used to build machine-learning models that can analyze pricing trends and provide personalized recommendations for owners. By collecting and analyzing this data, we can gain valuable insights into consumer behavior and preferences, ultimately helping to optimize the rental properties for the market. The use of web scraping technology in this project demonstrates how businesses can leverage data to gain a competitive advantage and enhance the user experience for both property owners and renters.

Appendix

1. Link to VRBO.com : https://www.vrbo.com/. Some facts about VRBO:

   VRBO stands for Vacation Rentals By Owner and was founded in 1995 by husband and wife team, David and Lynn Clouse, who wanted to create a platform for property owners to advertise their vacation homes. VRBO was acquired by HomeAway in 2006, and then by Expedia Group in 2015. It now operates as a part of Vrbo, a global vacation rental online marketplace.

   VRBO has over 2 million vacation rentals in more than 190 countries around the world, making it one of the largest vacation rental platforms. The site allows property owners to list and rent out their entire homes, apartments, cabins, and villas, as well as individual rooms in shared spaces across a range of unique and unusual vacation rentals, such as treehouses, houseboats, and even castles.

   VRBO's upcoming potential has stirred the establishment of the market leader Airbnb. Travelers and customers are increasingly using VRBO as a replacement, and therefore there seems to be a huge untapped market laying with VRBO to explore. Refer to the link here to see a comparison shown between the 2 sites: https://travelfreak.com/airbnb-vs-vrbo/

2. References to data stated in "Introduction":
   https://www.grandviewresearch.com/industry-analysis/vacation-rental-market

   https://www.igms.com/vacation-rental-sites/#

   statista.com/outlook/mmo/travel-tourism/vacation-rentals/united-state

   https://www.igms.com/2023-vacation-rental-industry-trends/#

   https://www.lodgify.com/blog/online-vacation-rental-rates/