Expedia Group RecTour Research Dataset

ADAM WOZNICA[*] and JAN KRASNODEBSKI[*], Expedia Group, Switzerland

This document provides details on the dataset that Expedia Group released to the RecTour community at the 15th ACM Conference on Recommender Systems. This dataset is based on real traveler lodging searches and bookings on Brand Expedia websites, which have been anonymized to protect identities of consumers and suppliers. The intention is to provide the recommendation system research community, and more specifically travel researchers, an open and rich dataset for their work. The motivation for this dataset was multiple requests originating from Expedia Group-sponsored competitions, where participants wanted to use the data that was provided for research purposes. This dataset was designed to meet that specific demand while preserving confidentiality.

Additional Key Words and Phrases: datasets

1 INTRODUCTION

Expedia Group is the world's travel platform that offers consumers a broad selection of travel products across brands such as Expedia, Hotels.com and Vrbo. 2019 bookings were over $107 billion while serving hundreds of millions of travelers [4].

To foster research in recommendation systems for travel, Expedia Group has provided a real world dataset that consists of lodging shopping and purchase data. This builds upon Expedia Group's previous efforts in the area of sharing data for recommendation system and tourism researchers via competitions [6, 7] and educational challenges [1, 3]. Participants were often interested in using the data from the contest for additional research of their own. However, datasets from contests are not directly fit for general research as they are designed for the smooth operation of a specific competition. This places various requirements on them not related to research uses such as doctorate theses or academic research. The authors consulted with leading researchers from the RecTour community [2] to create a dataset inspired by these competitions that was oriented towards research use. There was also a perceived desire within the wider RecSys community for datasets similar in concept to MovieLens [5] in other fields, in order to provide diversity and additional avenues for recommendation research.

The dataset is available under a Creative Commons license, subject to appropriate acknowledgement.

2 DATASET

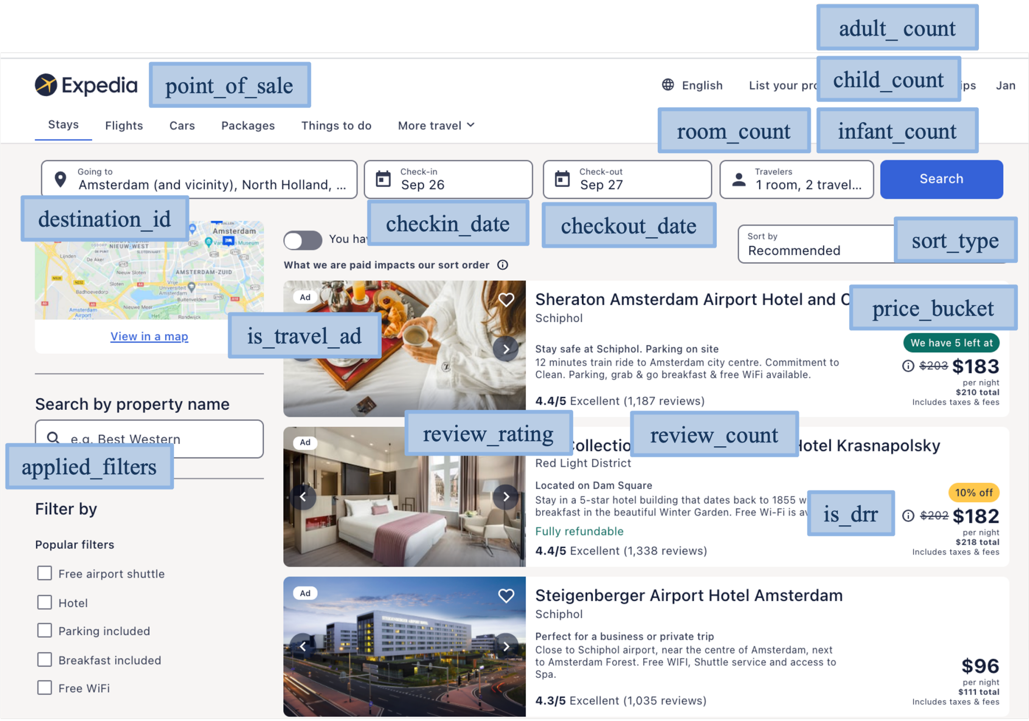
The Expedia Group dataset consists of global lodging shopping and purchase data from consumers in multiple countries across tens of thousands of destinations. The data are organized around a set of "search result impressions", i.e. the ordered list of properties that a consumer sees after a lodging search at one of the Brand Expedia websites. The user response is provided as a click on a property or/and a purchase of a property room. Only clicks and purchases that occurred after a search and before the next search within a 180 minute time limit are attributed to a search.

A property refers to one of over a million hotels, vacation rentals, apartments, B&Bs, hostels and other properties appearing on Brand Expedia's websites. Room types are not distinguished and the data can be assumed to apply to the least expensive room type.

The data span a period from 2021-06-01 to 2021-07-31 and contain searches for a random sample of consumers who made at least one click during the above time frame. Consumers who booked more than 4 distinct properties during

[*] Both authors contributed equally to this research.

Authors' address: Adam Woznica, awoznica@expediagroup.com; Jan Krasnodebski, jkrasnodebski@expediagroup.com, Expedia Group, Rue du 31 Décembre 40-42, Geneva, Switzerland, 1207.



Going to
Amsterdam (and vicinity), North Holland, ...

Check-in
Sep 26

Check-out
Sep 27

Travelers
1 room, 2 travel...

You ha

What we are paid impacts our sort order

Sort by
Recommended

Search by property name

e.g. Best Western

Filter by

Popular filters

□ Free airport shuttle

□ Hotel

□ Parking included

□ Breakfast included

□ Free WiFi

Sheraton Amsterdam Airport Hotel and C
Schiphol

Stay safe at Schiphol. Parking on site
12 minutes train ride to Amsterdam city centre. Commitment to Clean. Parking, grab & go breakfast & free WiFi available.

4.4/5 Excellent (1,187 reviews)

We have 5 left at
~~$203~~ $183
per night
$210 total
Includes taxes & fees

ollectio ... lotel Krasnapolsky
Red Light District

Located on Dam Square
Stay in a 5-star hotel building that dates back to 1855 w
breakfast in the beautiful Winter Garden. Free Wi-Fi is av

Fully refundable

4.4/5 Excellent (1,338 reviews)

10% off
~~$202~~ $182
per night
$218 total
Includes taxes & fees

Steigenberger Airport Hotel Amsterdam
Schiphol

Perfect for a business or private trip
Close to Schiphol airport, near the centre of Amsterdam, next to Amsterdam Forest. Free WIFI, Shuttle service and access to Spa.

4.3/5 Excellent (1,035 reviews)

$96
per night
$111 total
Includes taxes & fees

Fig. 1. Data labels as seen on Brand Expedia sites.

this period are excluded. The data span more than 800k unique users and approx. 2.5M searches and include desktop and mobile device traffic. The data include traveler inputs such as adding filters and selecting specific sort types, such as price ascending.

Figure 1 outlines the relationship between the search and property data in the dataset with the values impressed on the Brand Expedia site. Figure 2 outlines the click and purchase pathways on Brand Expedia's site.

2.1 Data Anonymization and Resampling

Several steps have been taken to anonymize the data and obfuscate the true data distribution to protect users and commercial sensitivities.

First, the *point_of_sale*, *geo_location_country* and *destination_id* columns were mapped to frequency based indexes. The *prop_id* column was indexed based on a random order. Next, distributions of the following categorical attributes were obfuscated by randomly changing proportions of users:

- *point_of_sale*
- *geo_location_country*
- *destination_id*
- *sort_type*
- *is_mobile*

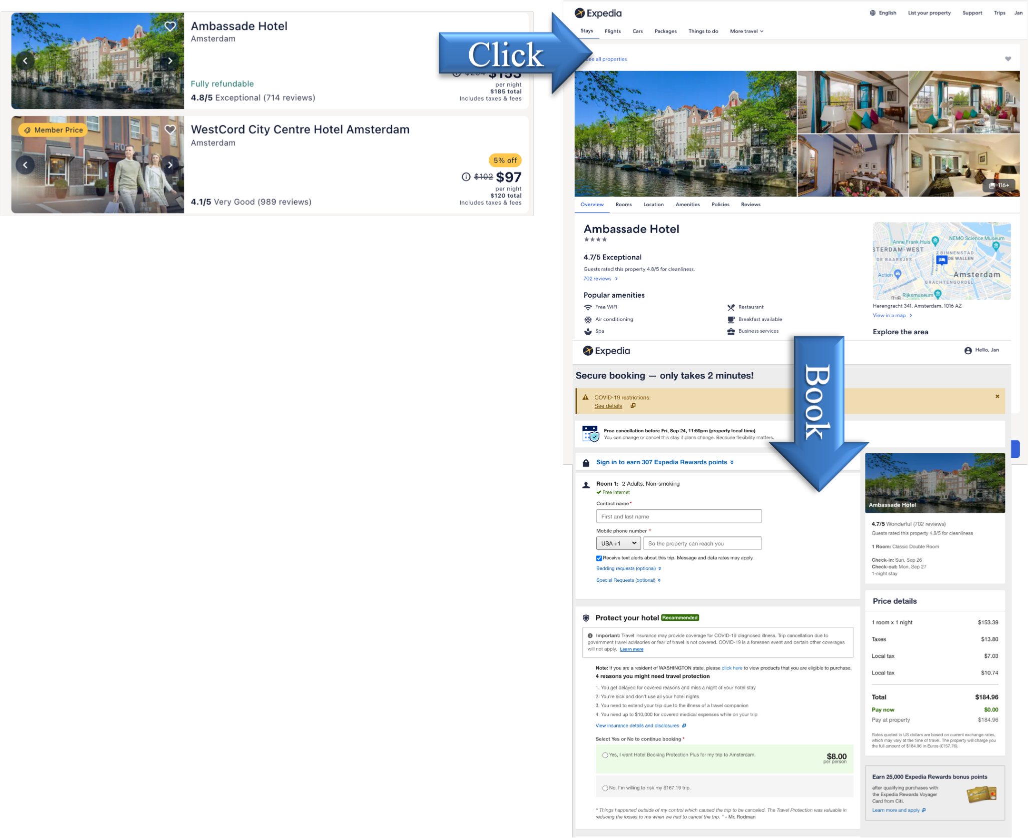


Fig. 2. Representation of click and purchase on Brand Expedia sites.

For example, the proportion of mobile searches (given by the *is_mobile* column) is similar but not identical to the ”true” proportion. Finally, we changed proportions of the *num_clicks* and *is_trans* ”label” attributes at the property (*prop_id*) level. In other words, the click through rate (CTR) and conversion rate (CVR) at the property level computed based on the above attributes do not exactly match the ”true” CTR and CVR values.

2.2 Attributes

In this section we provide a detailed list of attributes.

Table 1. Attribute description.

Attribute Name	DataType	Description	Comments
user_id	String	Unique user id (i.e. browser cookie)	
search_id	String	Unique search id	
search_timestamp	Timestamp	Date and time of the search	Rounded to minutes

point_of_sale	Integer	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.fr, ...)	Frequency based indexing. Obfuscated true distribution.
geo_location_country	Integer	The ID of the country the consumer is located	Frequency based indexing. Obfuscated true distribution.
is_mobile	Boolean	Whether the search was made from a mobile device	Obfuscated true distribution.
destination_id	Integer	ID of the destination where the hotel search was performed	Obfuscated true distribution.
checkin_date	Date	Stay start date	
checkout_date	Date	Stay stop date	
adult_count	Integer	The number of adults specified in the search	
child_count	Integer	The number of children specified in the search	
infant_count	Integer	The number of infants specified in the search	
room_count	Integer	Number of rooms specified in the search	
sort_type	String	Sort type	Obfuscated true distribution.
applied_filters	String	Pipe delimited list of applied filters. Each filter is identified by its name and value. Sample value: *STAR:4.0|LODGING:HOTEL*	Anonymized Property Name and Point of Interest filters.
impressions	List[Impr]	”|” delimited list of impressions. Each impression consist of the following ”,” delimited attributes: • rank • prop_id • is_travel_ad • review_rating • review_count • star_rating • is_free_cancellation • is_drr • price_bucket • num_clicks • is_trans	
Impr.rank	Integer	Hotel position on Expedia’s search results page.	
Impr.prop_id	Long	The ID of the property. It matches prop_id from Table 2.	Indexed based on a random order.
Impr.is_travel_ad	Boolean	If the impressed property is a travel ad (labelled "Ad", pay per click advertisement).	

Impr.review_rating	Float	The mean customer review score for the property on a scale out of 5, rounded to nearest integers. A 0 means there have been no reviews, null that the information is not available.	
Impr.review_count	Integer	The number of reviews for the property rounded to the nearest 25.	
Impr.star_rating	Float	The star rating of the hotel, from 1 to 5. A null indicates the property has no stars, the star rating is not known or cannot be publicized.	
Impr.is_free_cancellation	Boolean	If a booking can be cancelled without extra fees.	
Impr.is_drr	Boolean	If the property had a discount price reduction specifically displayed ("strikeout" price).	
Impr.price_bucket	Integer	Price bucket (1-5) based on percentile of the distribution of impressed prices; lower values of price_bucket correspond to lower prices. A null value means that the property was not available.	
Impr.num_clicks	Integer	Number of clicks within 180 minutes	Obfuscated true distribution.
Impr.is_trans	Boolean	If there was a transaction within 180 minutes	Obfuscated true distribution.

2.2.1 Property amenities. In addition to the main dataset from Table 1 we also released a property amenities dataset described in Table 2. This dataset spans approximately 1.5 million properties. Properties from the main table which cannot be matched with properties from the amenities table can be assumed to have missing amenities.

3 CONCLUSIONS

Expedia Group has provided a dataset based on real traveler behavior specifically for academic researchers and students. This dataset should address the demand that has been expressed in the past for it during competitions and events. This dataset can also be used by instructors for courses. Feedback is welcome on how we can improve this dataset in the future, and what other datasets may be useful for the RecTour and recommendation system research community.

4 ACKNOWLEDGEMENTS

We would like to acknowledge Julia Niedhardt for her initiative with the idea of creating an industry-based real world dataset for recommendation system and tourism researchers. And for her efforts to make it a reality at RecTour 2021. We also thank Dr. Wolfgang Wörndl for his contribution to this project.

Table 2. Property amenities table.

Attribute Name	DataType	Comments
prop_id	Long	It matches Impr.prop_id from Table 1.
AirConditioning	Boolean	
AirportTransfer	Boolean	
Bar	Boolean	
FreeAirportTransportation	Boolean	
FreeBreakfast	Boolean	
FreeParking	Boolean	
FreeWiFi	Boolean	
Gym	Boolean	
HighSpeedInternet	Boolean	
HotTub	Boolean	
LaundryFacility	Boolean	
Parking	Boolean	
PetsAllowed	Boolean	
PrivatePool	Boolean	
SpaServices	Boolean	
SwimmingPool	Boolean	
WasherDryer	Boolean	
WiFi	Boolean	

REFERENCES

[1] 2021. EXPEDIA GROUP X ENTER21 Data Science Competition Socially Responsible and Inclusive Tourism. https://enter-conference.org/compete/expedia-group-x-enter21/.

[2] 2021. RecTour: Workshop on Recommenders in Tourism. https://recsys.acm.org/recsys21/rectour/.

[3] American Statistical Association. 2017. ASA DataFest 2017. https://www.dropbox.com/s/eafdup47fpcqvam/UofT%20Stats%20data%20than%20v5%20-%20FINAL.mp4?dl=0.

[4] Expedia Group. 2020. Form 10-K. https://s27.q4cdn.com/708721433/files/doc_financials/2020/ar/Expedia-Group-Annual-Report.pdf.

[5] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (Dec. 2015), 19 pages. https://doi.org/10.1145/2827872

[6] Adam Woznica and Jan Krasnodebski. 2013. Personalize Expedia Hotel Searches - ICDM 2013 Learning to rank hotels to maximize purchases. https://www.kaggle.com/c/expedia-personalized-sort.

[7] Adam Woznica and Jan Krasnodebski. 2016. Expedia Hotel Recommendations. Which hotel type will an Expedia customer book? https://www.kaggle.com/c/expedia-hotel-recommendations.