Booking.com Multi-Destination Trips Dataset

Dmitri Goldenberg
dima.goldenberg@booking.com
Booking.com, Tel Aviv, Israel

Pavel Levin
pavel.levin@booking.com
Booking.com, Amsterdam, Netherlands

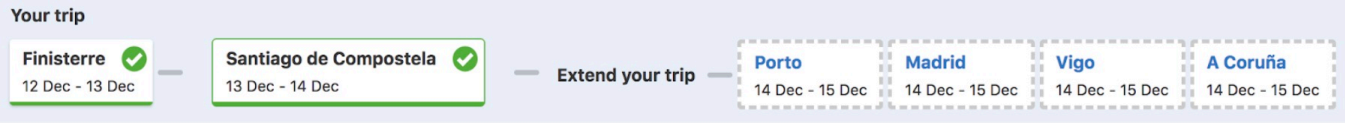

Figure 1: An example of trip extension recommendations on Booking.com website

ABSTRACT

We introduce a novel dataset of real multi-destination trips booked through Booking.com's online travel platform. The dataset consists of 1.5 million reservations representing 359,000 unique journeys made across 39,000 destinations. As such, the data is particularly well suited to model sequential recommendation and retrieval problems in a high cardinality target space. To preserve user privacy and protect business-sensitive statistics, the data is fully anonymized, sampled and limited to five user origin markets. Even so, the dataset is representative of the general travel purchase behavior and therefore presents a uniquely valuable resource for Machine Learning and information retrieval researchers. This work provides an overview of the dataset. It reports several benchmark results for relevant recommendation problems obtained as part of the recently held Booking.com data challenge during the WSDM WebTour workshop.

CCS CONCEPTS

• **Information systems** → **Personalization**; **Recommender systems**.

KEYWORDS

Personalization, Travel, Recommender Systems, Dataset

ACM Reference Format:
Dmitri Goldenberg and Pavel Levin. 2021. Booking.com Multi-Destination Trips Dataset. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3404835.3463240

1 INTRODUCTION

Booking.com is one of the leading online travel platforms, enabling millions of customers to book accommodation worldwide. Many such customers go to multiple destinations within every single trip, thus generating rich sequential purchase data. One specific application for such a dataset in the travel domain is building a trip extension recommendation system: given a partially booked trip, can we recommend the next destination to book?

The dataset presented in this paper contains about 1.5 million anonymized accommodation bookings, which represent 359,000 unique trips. The data is conveniently split into training and test sets for easier reproducibility of any results. The dataset was introduced publicly at Booking.com's challenge [**?**] as part of the WebTour 2021 ACM WSDM workshop on web tourism [**?**] at the 14th ACM international 2021 WSDM Conference. It was published on the Booking.com challenge website[1] and will remain accessible via the Booking GitHub repository [2]. Leading challenge solution papers are included in the workshop proceedings, including detailed method documentation and codebase.

Unlike previously published sequential datasets for recommendation tasks [**? ?**], this dataset contains only positive examples of completed trip reservations. Working with such data resembles the modeling done on real online recommendation tasks and adds an additional challenge to the classical modeling efforts [**?**]. Similar travel datasets were shared previously, such as Expedia's Kaggle

[1] https://www.bookingchallenge.com
[2] https://github.com/bookingcom/ml-dataset-mdt

Table 1: Dataset description

Column	Description
user_id	User ID
checkin	Reservation check-in date
checkout	Reservation check-out date
affiliate_id	An anonymized ID of affiliate channel where the booker came from (e.g. direct, 3rd party referrals, paid search engine, etc.)
device_class	desktop/mobile
booker_country	Country from which the reservation was made (anonymized)
hotel_country	Country of the hotel (anonymized)
city_id	unique ID of the hotel's city (anonymized)
utrip_id	unique ID (a group of multi-destinations bookings within the same trip)

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463240

competition [?] and Trivago's Recsys 2019 challenge [?]. However, while the datasets mentioned above focus on session-based user modeling, our resource focuses on reservation-only data, allowing us to focus on the trip entity and shedding light on customers' actual travel patterns.

2 DATASET

2.1 Multi-Destination Trips

Travelers often go to multiple destinations within a single trip to experience various cities and landmarks in one vacation. Unlike classical e-commerce platforms, a sequence of reservations on an online travel platform often represents a sequence of physical actions by the customers during the trip. The rich sequential bookings data encompass a travel pattern of real experiences.

Figure 2 demonstrates an example multi-destination trip in Italy. The trip starts from Milan, through Venice, Florence, Rome, Naples, Reggio, Calabria, and finishes in Palermo. Such sequential order is reasonable due to the geographical location of each of the destinations and has a high likelihood compared to alternative reshuffled sequences of the same cities. The itinerary plan is also highly dependant on the dates, the origin country, in addition to the personal preferences of each traveler.

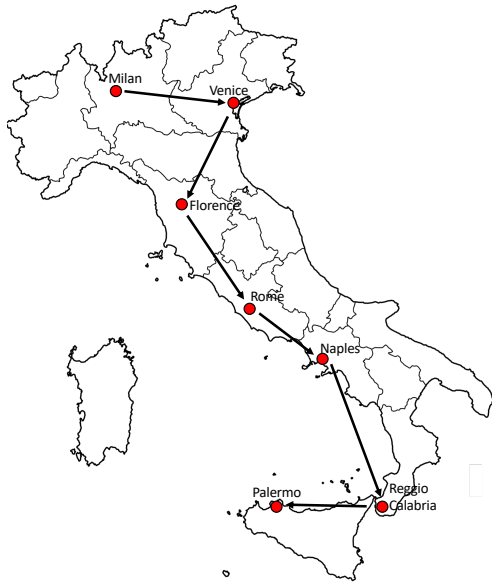

Figure 2: Example Multi-Destination Trip in Italy

2.2 Data Schema

The dataset consists of over one and a half million anonymized hotel reservations based on real data, as described in Table 1. The dataset contains information about reservation dates, destination, customers' countries of origin and the platform they used.

Each reservation is a part of a customer's trip (identified by utrip_id), including consecutive reservations. There are zero or more days between check-out and check-in dates of two consecutive reservations.

2.3 Exploratory Analysis

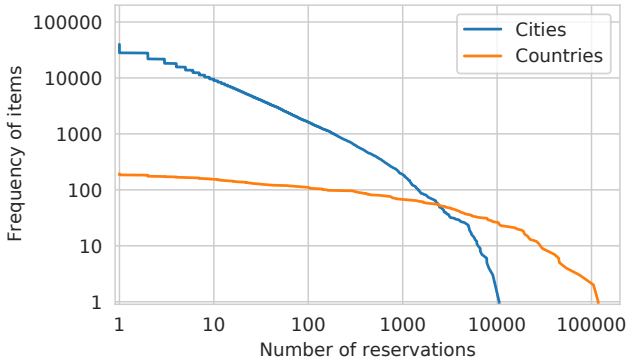

Figure 3: Cities and Countries Reservations Frequency

The dataset is split between a training set (about 75% of all reservations) and a test set. The training set consists of 1,166,835 reservations of 200,153 unique customers over 217,686 trips. This data represents a sample of anonymized Booking.com reservations stayed between 2016 and 2017. Additional key statistics of the data (such as values ranges, metrics, and the number of unique values) are described in table 2.

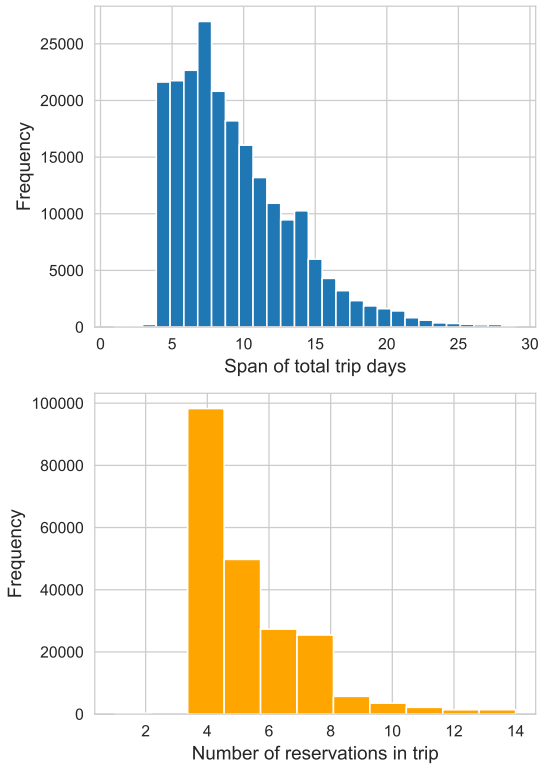

Figure 4: Trip length and days span distributions

Table 2: Training dataset columns and properties key statistics

Column / Property	Unique Values	Mean Value	Mode Value	Min Value	Max Value
user_id	200153	-	2209265	-	-
checkin	425	2016-08-01	2016-08-08	2015-12-31	2017-02-27
checkout	425	2016-08-03	2016-08-10	2016-01-01	2017-02-28
city_id	39901	-	47499	-	-
device_class	3	-	desktop	-	-
affiliate_id	3254	-	9924	-	-
booker_country	5	-	Gondal	-	-
hotel_country	195	-	Cobra Island	-	-
utrip_id	217686	-	3635431_3	-	-
Reservation days span	30	1.74	1	1	30
Trip span (days)	81	9.32	7	1	304
Trip span (# reservations)	41	5.36	4	1	48

The attached test dataset contains 378,667 reservations and is distributed similarly. The city_id of the final reservation of each trip is concealed and requires a prediction.

Figure 3 depicts the distribution of the number of reservations across the different cities and countries in the training set (on a log-log scale). The X-axis represents the total number of reservations in the city or country, and the Y-axis presents the number of cities or countries at each data-point. There is only a single city in the dataset with more than 10,000 reservations and a few countries above 100,000 reservations.

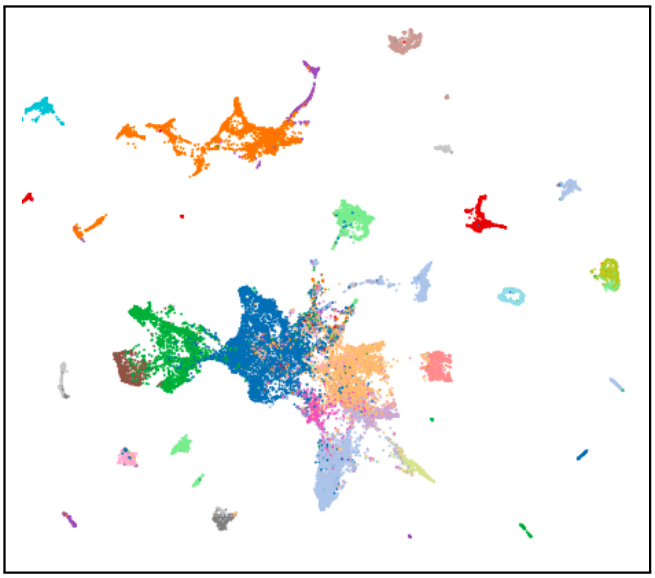

Figure 5: 2-D visualization of city embeddings learned by Cleora. Each color denotes a different hotel country out of top 20 countries. Taken from [?].

The distribution of trips' length and days span is presented in figure 4. On the upper plot, we observe that the most popular trip span is 7 days, with a quick decay. The distribution of the number of total reservations in a trip is shown on the lower plot - the trips in the dataset contain at least 4 reservations.

While the cities and the countries in the dataset are anonymized, many travel patterns can be learned because of the sequential nature of the data. For instance, Figure 5 depicts a two-dimensional UMAP [?] projection of city embeddings learned by Cleora [?] system, generated by one of the top competitors in the Booking.com data challenge. Each color denotes a different hotel country out of the top 20 countries in the dataset. The embeddings exhibit the awareness of geographic closeness of cities representing travel relationships between the cities in the dataset.

2.4 Data Files

The resource repository consists of the following files:

- **train_set.csv** - Training dataset of 1,166,835 reservations.
- **test_set.csv** - Test dataset of 378,667 reservations (with a concealed last destination) as used in Booking.com challenge.
- **ground_truth.csv** - The true last city values for 70,662 trips in the test dataset.
- **submission.csv** - an example submission for the test data based on top four popular cities benchmark.
- **evaluation_demo.ipynb** - a Jupyter notebook with sample code for data loading, formatting submission and evaluation.
- **Readme.md** - a documentation file with dataset description and usage terms and conditions.

3 DATASET CREATION

The dataset is a sample of Booking.com accommodation reservations with check-out dates between January 2017 and February 2018. In order to preserve user privacy, no personally identifiable information was included in the data.

Similarly, to protect business-sensitive statistics, the dataset is limited to only five user countries; all country names were anonymized by randomly mapping each of them to one of the fictional county names retrieved from Wikipedia[3] using an automated script. Both, Booking.com internal city ids, as well as affiliate ids, were randomly permuted and assigned their ordinal number in the permutation in order to create the city and affiliate ids shared in this dataset. Check-in, check-out and device classes are shared in their original form without any processing.

Each data point represents a stayed (non-canceled) reservation. All reservations are grouped into unique trips identified through *utrip_id* feature. Most trips (96%) consist of entirely consecutive bookings, meaning that the next booking's check-in date is the previous booking's check-out date. However, there are also a small number of trips containing a "gap". These gaps can be interpreted as the user staying with a friend or booking with a competitor. The dataset contains trips with at least four reservations, to allow a proper sequential recommendation task.

3.1 Evaluation dataset

The evaluation test dataset was created as a standard hold-out validation set for recommender algorithms, developed based on the train set. The data was divided into the train and test set with a 75%/25% random split, based on the *utrip_id* column. Such a split ensures that all the reservations of the same trip belong to the same dataset. However, the same user might appear in both train and test sets. In fact, about 15% of the trips in the test set belong to users who appear in the train set as well.

4 RESEARCH DIRECTIONS

4.1 Sequential Recommendations

The use of sequential models has increased rapidly in the field of recommendation systems [**? ? ? ?**]. Recent studies show how to use contextual data in a Recurrent Neural Network (RNN) based recommender system [**? ?**] and combine the recurrent part within a complex architecture. A related tutorial on *Personalizaton in Practice* [**?**] reviewed advances in sequence-based recommendations using real-world use cases, including the trip extension recommendations at Booking.com.

4.2 Next Destination Recommendation

The main recommendation task on the given dataset is to predict the final city (city_id) of each trip (utrip_id). The quality of the predictions is evaluated based on the top four recommended cities for each trip by using *Top-4 Accuracy metric* (4 represents the four suggestion slots at Booking.com website as shown in figure 1). When the actual city is one of the top four suggestions (regardless of the order), the prediction is considered correct.

4.3 Additional Research Directions

While the next destination in the trip is a key Machine Learning challenge in the context of trip recommendations, there are multiple additional questions to be solved.

[3]https://en.wikipedia.org/wiki/List_of_fictional_countries

This section provides a brief description of potential research directions on the provided dataset, which were explored within Booking.com.

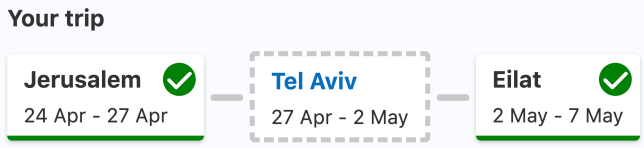

Figure 6: Gap filling recommendation

4.3.1 Gaps filling. Customers often perform reservations in an order that differs from the chronological order of the stays. This happens due to multiple reasons, such as starting from *anchor* bookings at arrival and department destinations, which happens prior to planning the rest of the trip.

Such behavior opens up a new recommendation challenge in the form of "gaps-filling", where the missing part of the trip is surrounded by two existing bookings, as exampled in figure 6.

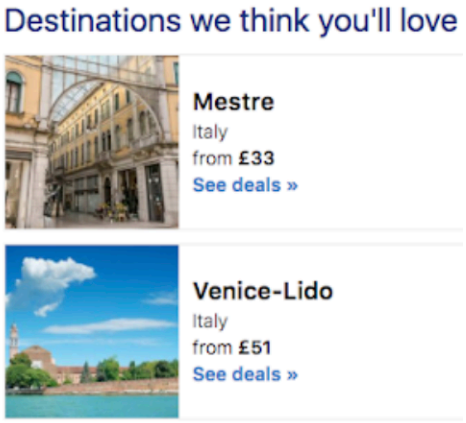

Figure 7: Alternative Destinations Recommendations

4.3.2 End of trip prediction. Showing trip extension recommendations might be relevant for some customers. However, it might be misleading and confusing to others. Therefore a personalization model might be beneficial, one that decides whether to keep recommending the trip extension or not. Recognizing the last step of the trip can be helpful for improving customer experience and gain potential business insights.

4.3.3 Length of stay recommendation. While planning a trip, the number of days to stay at each destination is an important (and hard) decision. Recommending the number of days to stay at each destination based on the existing information of the trip can provide a personalized experience to users and ease the booking process.

Table 3: Top performing Booking.com challenge teams

	Team	Accuracy@4	Method
1	**NVIDIA RAPIDS** [**?**]	**0.5939**	Ensemble of MLP, GRU and XLNet Transformers
2	**Synerise AI** [**?**]	**0.5780**	Cleora graph embedding and Sketch-Based EMDE model
3	**TEAM DASOU** [**?**]	**0.5741**	LSTM and GRU based recurrent neural architecture
4	**MBaigorria** [**?**]	0.5566	Many-to-many RNN Encoder-Decoder model
5	**APREC** [**?**]	0.5557	Attention-based Transformers and Lambdarank Loss
6	**hakubishin3, u++, yu-y4** [**?**]	0.5399	Weighted Average of LSTM models
7	**Anonymous**	0.5332	-
8	**Alexander Makeev**	0.5310	-
9	**YiNet**	0.5112	-
10	**Marlesson - MARS-Gym** [**?**]	0.4958	Neural Attentive Recommendation Machine

4.3.4 Trip groupings. The grouping of reservations into trips is a non-trivial task which relies on multiple assumptions and heuristics. It often relies on partial or unlabelled data, and therefore requires an unsupervised solution approach. Besides naive grouping techniques based on date gaps between the reservations, incorporating additional features may improve the accuracy of the grouping.

4.3.5 City similarities. Travel patterns data between countries and cities encompass hidden information about customer preferences and city properties. Using word or graph embedding techniques may offer similarity metrics between different cities. In turn, such metrics can be used for alternatives recommendations and incorporated into various machine learning models. Figure 7 demonstrates a usage of such similarity-based recommendation for alternative destinations for Venice.

5 BOOKING.COM CHALLENGE

The presented dataset was used at Booking.com's challenge [**?**], whose goal was to predict the final city of each trip on the test dataset. The quality of the challenge submissions was evaluated using Accuracy@4 metric as described in section 4.2. The contest allowed a single submission for each of the competing teams, with no opportunity to tune the modeling method on the ground-truth of the test dataset.

5.1 Naïve Benchmark

In order to estimate the potential predictability of the provided task, we evaluated two naïve benchmark results on the test dataset:

(1) Popular City Suggestions
(2) City-to-City pairs

5.1.1 Popular City Suggestions. Popular city benchmark evaluated the hit rate of the four most common cities in the training dataset. The test set evaluation resulted in an Accuracy@4 of 0.052, which means that 1 out of 20 prediction-sets will include the correct city.

5.1.2 City-to-City pairs. Another naïve approach that utilizes the sequential nature of the data is recommending the most popular city giving the previous city of the trip. Namely, given the current last city of the trip $City_l$ the next top four recommended cities $City_{n1}...City_{n4}$ are such that the transition $City_l \rightarrow City_{ni}$ is the most frequent in the training dataset. Evaluating this method on the test set achieved an Accuracy@4 of 0.4363.

5.2 Challenge Results

After two months of the contest, 40 research teams performed a final submission to the challenge. The top performing teams are listed in table 3. The median *Accuracy@4* score achieved 0.4562 on the hold-out test dataset, which is slightly better than the City-to-City benchmark described in the previous section.

The best-performing team achieved a 0.5939 score, providing a 30% improvement compared to the median result. Their solution put to use a blend of three different neural network architectures, using XLNet Transformers [**?**], Gated Recurrent Units [**?**], and feed-forward networks with a shared Session-based Matrix Factorization head (SMF). The authors' paper [**?**] describes a thorough experimentation method, alongside with insightful ablation study and publicly available code-repository. The authors discuss the superior performance of deep-learning based algorithms in the challenge, compared to classical recommender approaches relying on their past experience [**?**].

The runner-ups based their solution on a novel graph-embedding technique - Cleora [**?**], following an Efficient Manifold Density Estimator (EMDE) for the recommendation task [**?**], achieving 0.5780 Accuracy@4 score. Other leading solutions used various deep-learning and recommender techniques, including Transformers, LSTM networks [**? ?**], LambdaRank [**? ?**], and further state-of-the-art recommender methods. The top 10 performing teams introduced short papers and code repositories with a detailed description of their solution methodology.

6 SUMMARY

We introduce a novel travel dataset which unveils the complex dynamics of real multi-destination trips. The dataset is particularly suitable to model sequence-aware recommendation problems. Because the dataset consists of sequences of real customer purchases in a high-cardinality product space, we believe it will be useful to researchers interested in studying real-world recommendation systems. A recent contest conducted on the dataset provides benchmark results for the recommendation problem and opens up an opportunity to explore various state-of-the-art Machine Learning techniques. Besides the suggested next-destination problem, the dataset is highly useful for additional recommendation tasks and has the potential to become a standard benchmark to multiple research directions. The structured form of the data, together with the simple underlying booking process, allows multiple Machine Learning practitioners to explore the data in their research.

ACKNOWLEDGMENTS

The authors would like to thank Sarai Mizrachi, Maayan Kafry, Kostia Kofman and Guy Nadav for their help with the Booking.com challenge, to Steven Baguley for editing support and to challenge participants for submitting innovative applications of the dataset.