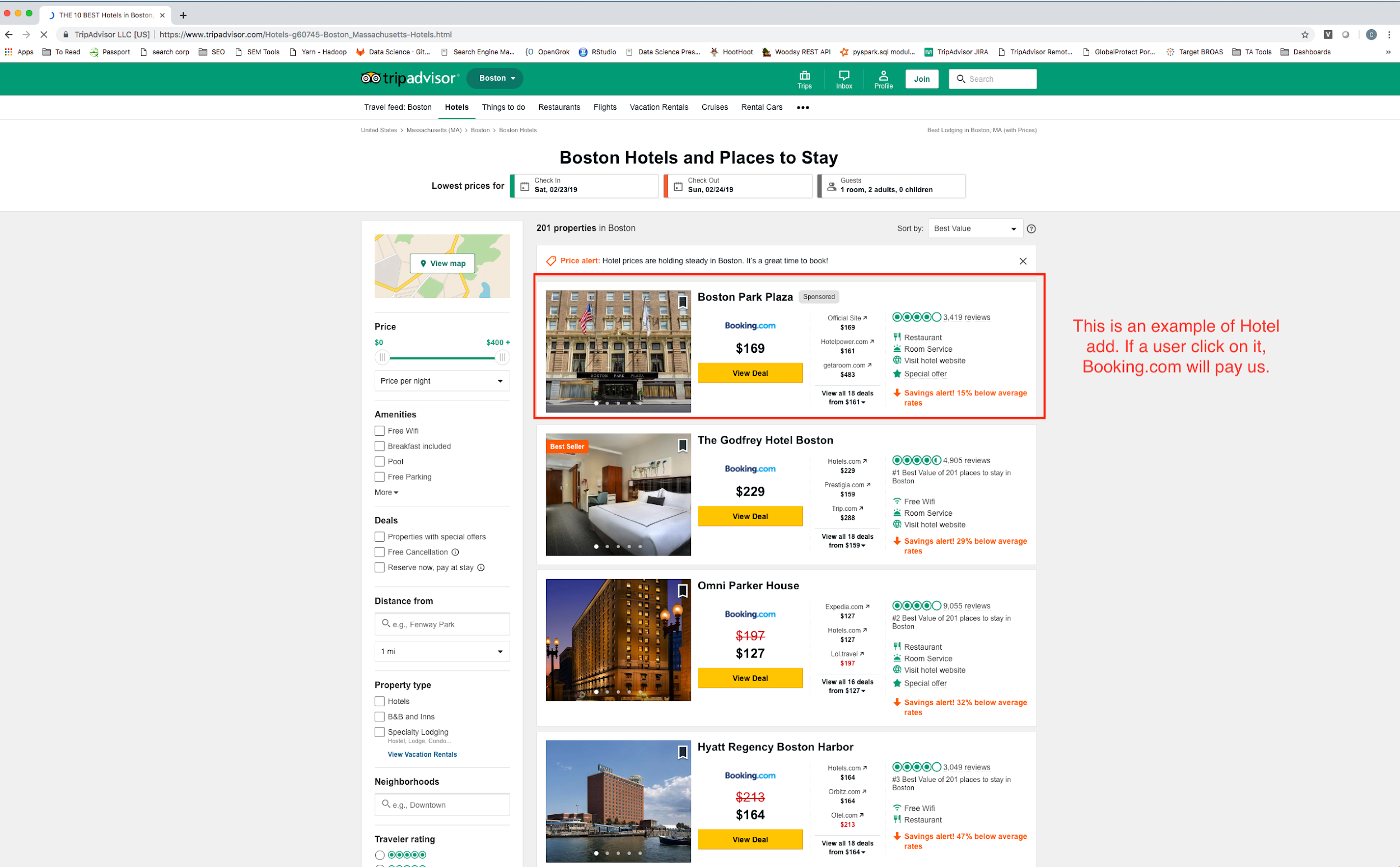
**Tripadvisor hotel booking challenge**

**A little bit of context...**

Tripadvisor is one of the leading hotel review and travel booking websites. We generate revenue via travel product related ads (Hotels, Restaurant, Attractions, Cruises, ...) which our partners (such as Kayak, Expedia, Marriott, ...) pay for everytime one of their ads gets clicked. See example below.



So that both our users and partners are satisfied, it is important that we only display ads that are likely to generate a booking. Indeed, users would stop visiting our website if the content was not relevant to them and partners would not be inclined to pay for ads that do not generate them revenue!

**Data**

For the scope of this Datathon, we’re providing you with two datasets featuring user-hotel interactions on [Tripadvisor.com](http://tripadvisor.com) in January 2019 from hotels in greater New York city area.

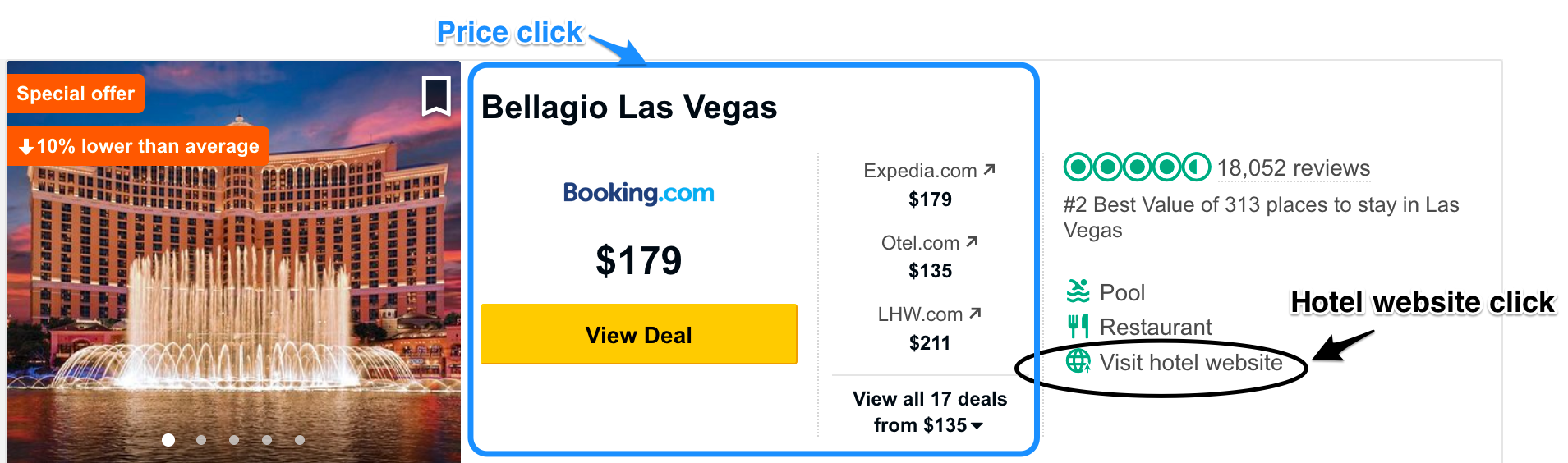
**activity\_data.csv**

This file is the user-hotel interaction data in January 2019. The dataset is unique on date, user\_id, hotel\_id, user\_action.

The user\_action column is a multi-tiered label describing how a user interacted with a hotel on a specific day.

1. ‘booking’: booked the hotel
2. ‘price\_click’: clicked on a price tag of the hotel.
3. ‘hotel\_website\_click’: click on the hotel website link.
4. ‘view’: viewed the hotel

Here is an example of a price\_click and a hotel\_website\_click



Other columns include:

* date: date of the action
* user\_id: unique user identifier
* user\_country: the country the user is residing in while searching on TripAdvisor
* device: the device (tablet, desktop, mobile) which the user is connected with
* hotel\_id: unique hotel identifier

**hotel\_data.csv**

This file has details for each unique hotel. Columns include:

* hotel\_id: unique hotel identifier
* hotel\_name: name of the hotel
* city\_name: name of the city
* star\_rating: star rating of the hotel, ranging from 1 to 5 in 0.5 increments
* bubble\_score: Aggregate Tripadvisor user review score, ranging from 10 to 50 in 5 increments
* review\_count: Tripadvisor user review count
* hotel\_type: hotel types. eg. hotel, hostel, etc
* brand\_name: hotel brand eg. Courtyard, Fairfield Inn, Marriott
* parent\_brand\_name: hotel parent company. eg. Courtyard, Fairfield Inn, Marriott are owned by Marriott International, Inc.

In the attached EDA.ipynb, you will find some basic exploratory analysis to help you get a first dive into the data. You are encouraged to use your own analysis to explore features further.

**Guidelines:**

In this challenge, you are responsible for defining your own problem using the dataset and creating training and validation datasets to evaluate your model. Ideally, you should come up with a simple baseline and improve upon that.

While you are welcome to let your imagination run free and come up with your own subject of investigation, here are some potential topics of exploration:

1. Which hotel is a particular user likely to book?
2. Which type of users are the most likely to book on our website?
3. Cluster hotels based on user activity.