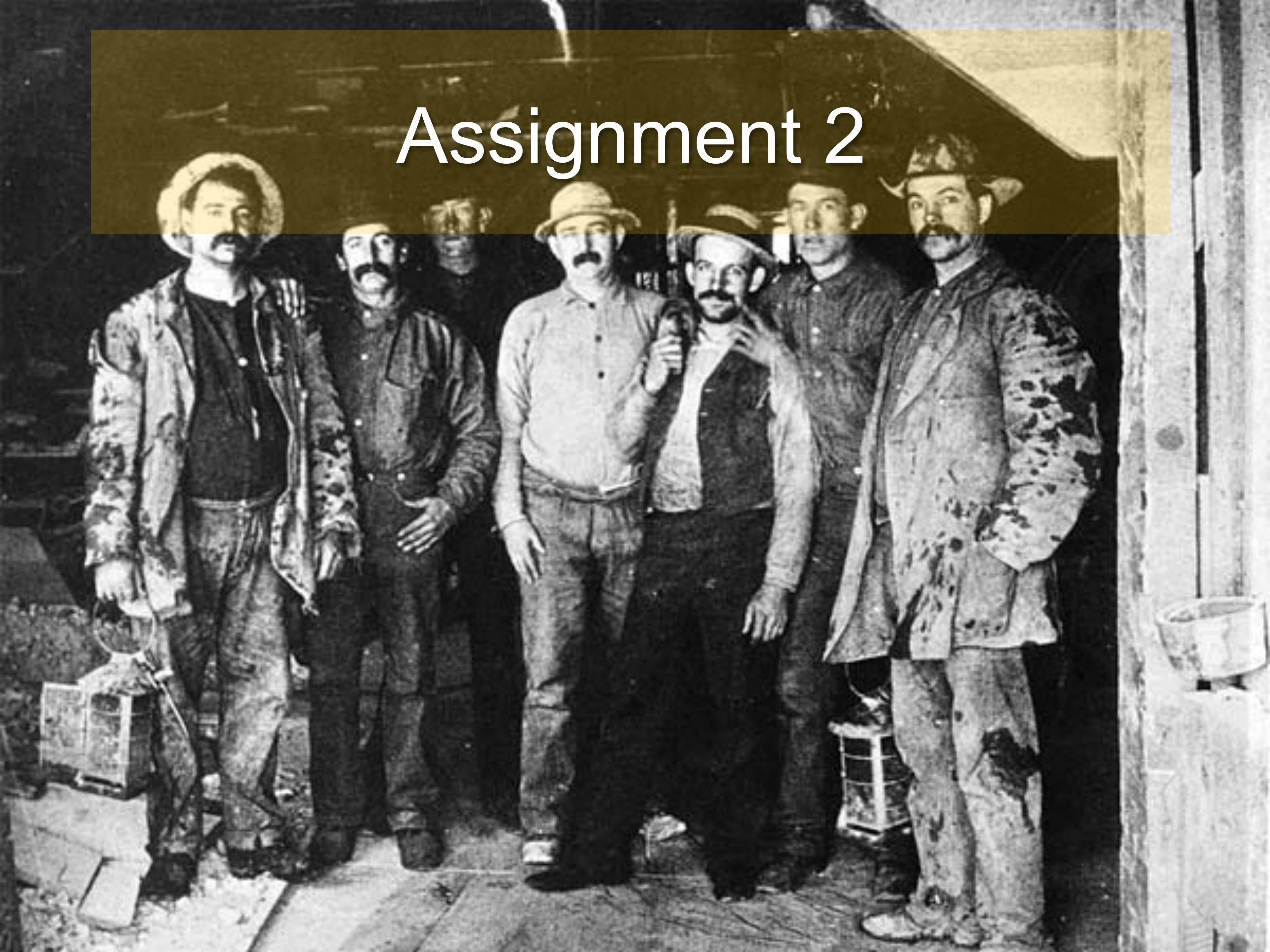
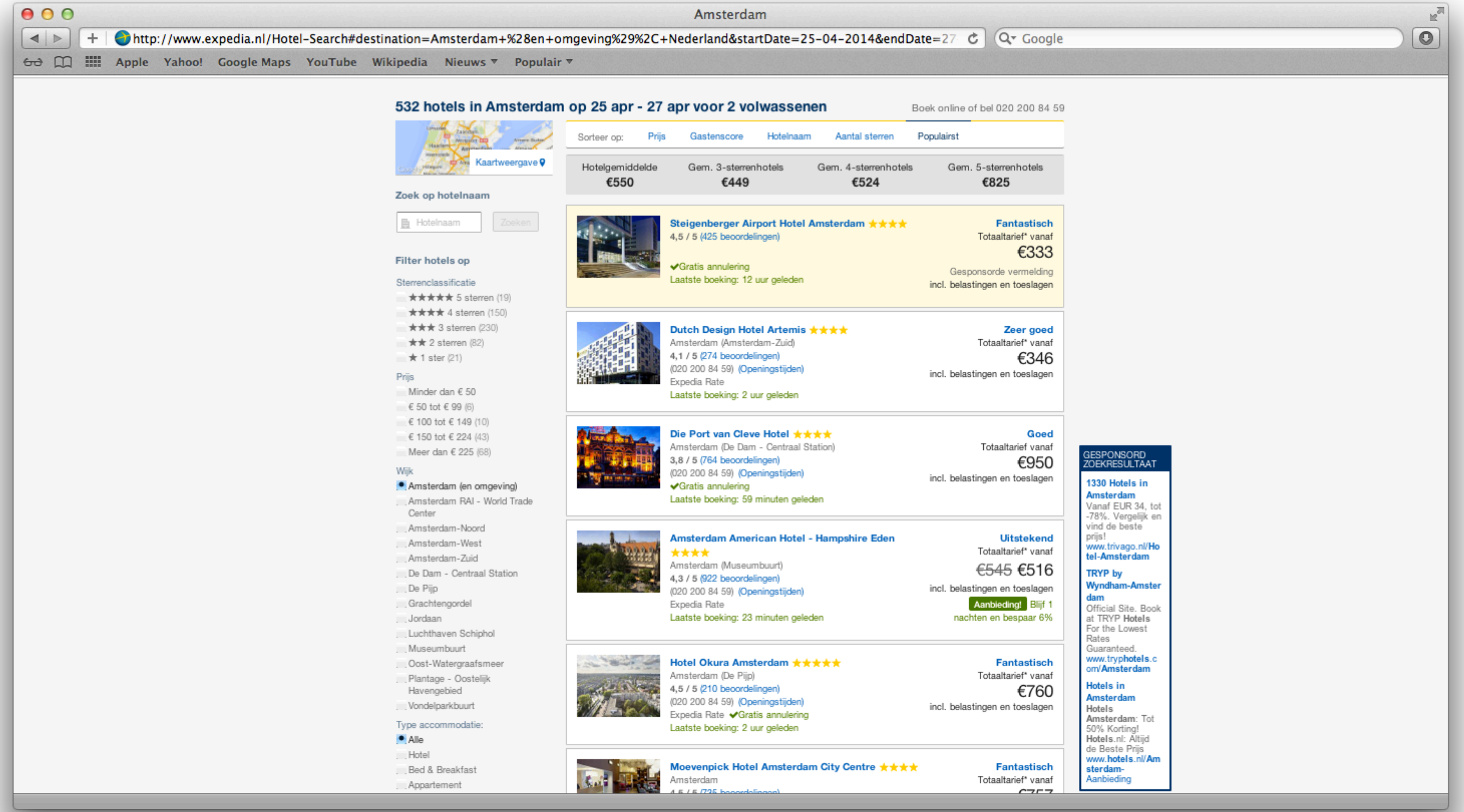
Assignment 2



Assignment 2

- Real life Data Mining assignment
- Your task: rank hotels on Expedia based on likelihood of booking
- Taken from Kaggle (DM competition website), dataset provided by Expedia
- Use ONLY the dataset provided via BB, not the Kaggle one (!)



Amsterdam

http://www.expedia.nl/Hotel-Search#destination=Amsterdam+%28en+omgeving%29%2C+Nederland&startDate=25-04-2014&endDate=27

532 hotels in Amsterdam op 25 apr - 27 apr voor 2 volwassenen

Boek online of bel 020 200 84 59

Sorteer op: Prijs | Gastenscore | Hotelnaam | Aantal sterren | Populairst

Hotelgemiddelde	Gem. 3-sterrenhotels	Gem. 4-sterrenhotels	Gem. 5-sterrenhotels
€550	€449	€524	€825

Zoek op hotelnaam

Hotelnaam | Zoeken

Filter hotels op

Sterrenclassificatie

- □ ★★★★★ 5 sterren (19)
- □ ★★★★ 4 sterren (150)
- □ ★★★ 3 sterren (230)
- □ ★★ 2 sterren (82)
- □ ★ 1 ster (21)

Prijs

- □ Minder dan € 50
- □ € 50 tot € 99 (6)
- □ € 100 tot € 149 (10)
- □ € 150 tot € 224 (43)
- □ Meer dan € 225 (68)

Wijk

- ● Amsterdam (en omgeving)
- ○ Amsterdam RAI - World Trade Center
- ○ Amsterdam-Noord
- ○ Amsterdam-West
- ○ Amsterdam-Zuid
- ○ De Dam - Centraal Station
- ○ De Pijp
- ○ Grachtengordel
- ○ Jordaan
- ○ Luchthaven Schiphol
- ○ Museumbuurt
- ○ Oost-Watergraafsmeer
- ○ Plantage - Oostelijk Havengebied
- ○ Vondelparkbuurt

Type accommodatie:

- ● Alle
- ○ Hotel
- ○ Bed & Breakfast
- ○ Appartement

Steigenberger Airport Hotel Amsterdam ★★★★
4,5 / 5 (425 beoordelingen)
✓Gratis annulering
Laatste boeking: 12 uur geleden
Fantastisch
Totaaltarief* vanaf
€333
Gesponsorde vermelding
incl. belastingen en toeslagen

Dutch Design Hotel Artemis ★★★★
Amsterdam (Amsterdam-Zuid)
4,1 / 5 (274 beoordelingen)
(020 200 84 59) (Openingstijden)
Expedia Rate
Laatste boeking: 2 uur geleden
Zeer goed
Totaaltarief* vanaf
€346
incl. belastingen en toeslagen

Die Port van Cleve Hotel ★★★★
Amsterdam (De Dam - Centraal Station)
3,8 / 5 (764 beoordelingen)
(020 200 84 59) (Openingstijden)
✓Gratis annulering
Laatste boeking: 59 minuten geleden
Goed
Totaaltarief* vanaf
€950
incl. belastingen en toeslagen

Amsterdam American Hotel - Hampshire Eden ★★★★
Amsterdam (Museumbuurt)
4,3 / 5 (922 beoordelingen)
(020 200 84 59) (Openingstijden)
Expedia Rate
Laatste boeking: 23 minuten geleden
Uitstekend
Totaaltarief* vanaf
~~€545~~ €516
incl. belastingen en toeslagen
Aanbieding! Blijf 1 nachten en bespaar 6%

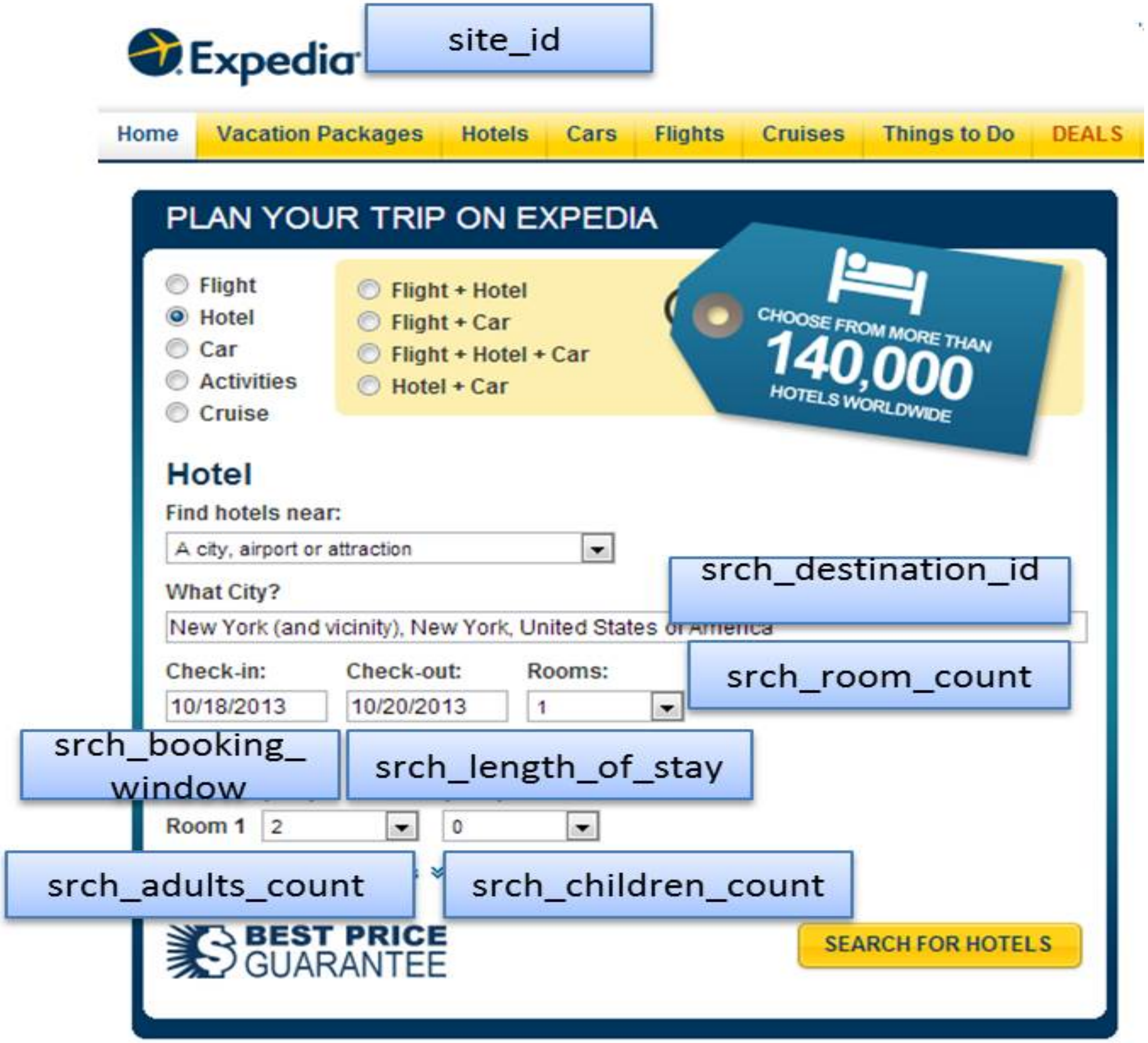
Hotel Okura Amsterdam ★★★★★
Amsterdam (De Pijp)
4,5 / 5 (210 beoordelingen)
(020 200 84 59) (Openingstijden)
Expedia Rate ✓Gratis annulering
Laatste boeking: 2 uur geleden
Fantastisch
Totaaltarief* vanaf
€760
incl. belastingen en toeslagen

Moevenpick Hotel Amsterdam City Centre ★★★★
Amsterdam
4,5 / 5 (735 beoordelingen)
Fantastisch
Totaaltarief* vanaf
€757

Assignment 2

- How could we do this using Data Mining techniques? Do you have ideas?

Assignment 2: The data - search fields



site_id

srch_destination_id

srch_room_count

srch_booking_window

srch_length_of_stay

srch_adults_count

srch_children_count

PLAN YOUR TRIP ON EXPEDIA

○ Flight
● Hotel
○ Car
○ Activities
○ Cruise

○ Flight + Hotel
○ Flight + Car
○ Flight + Hotel + Car
○ Hotel + Car

Hotel

Find hotels near: A city, airport or attraction

What City? New York (and vicinity), New York, United States of America

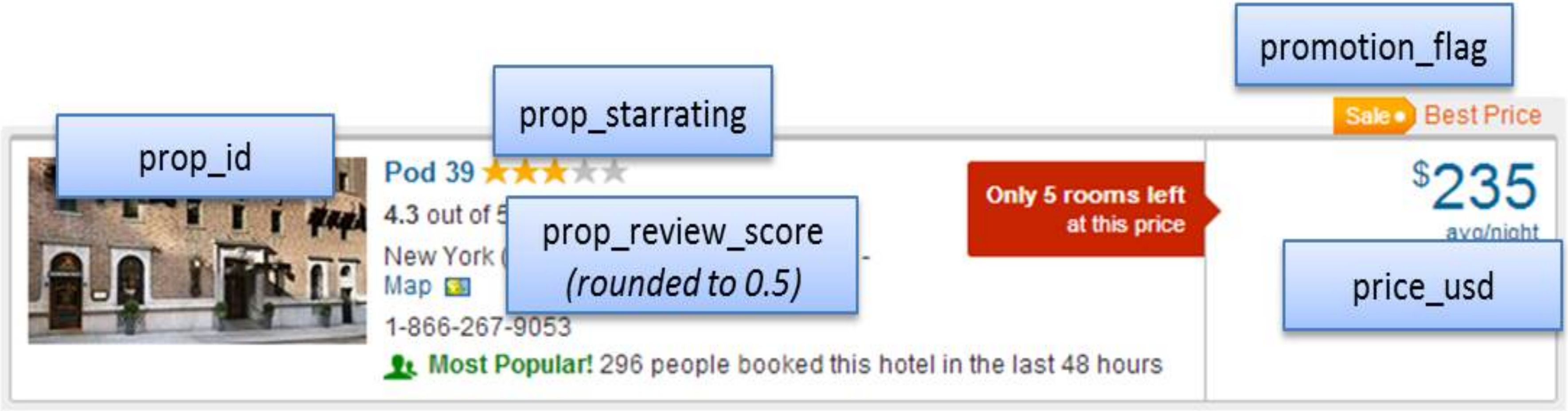
Check-in: 10/18/2013

Check-out: 10/20/2013

Rooms: 1

Room 1 2 0

Assignment 2: The data - resulting properties (more per search)



promotion_flag

prop_starrating

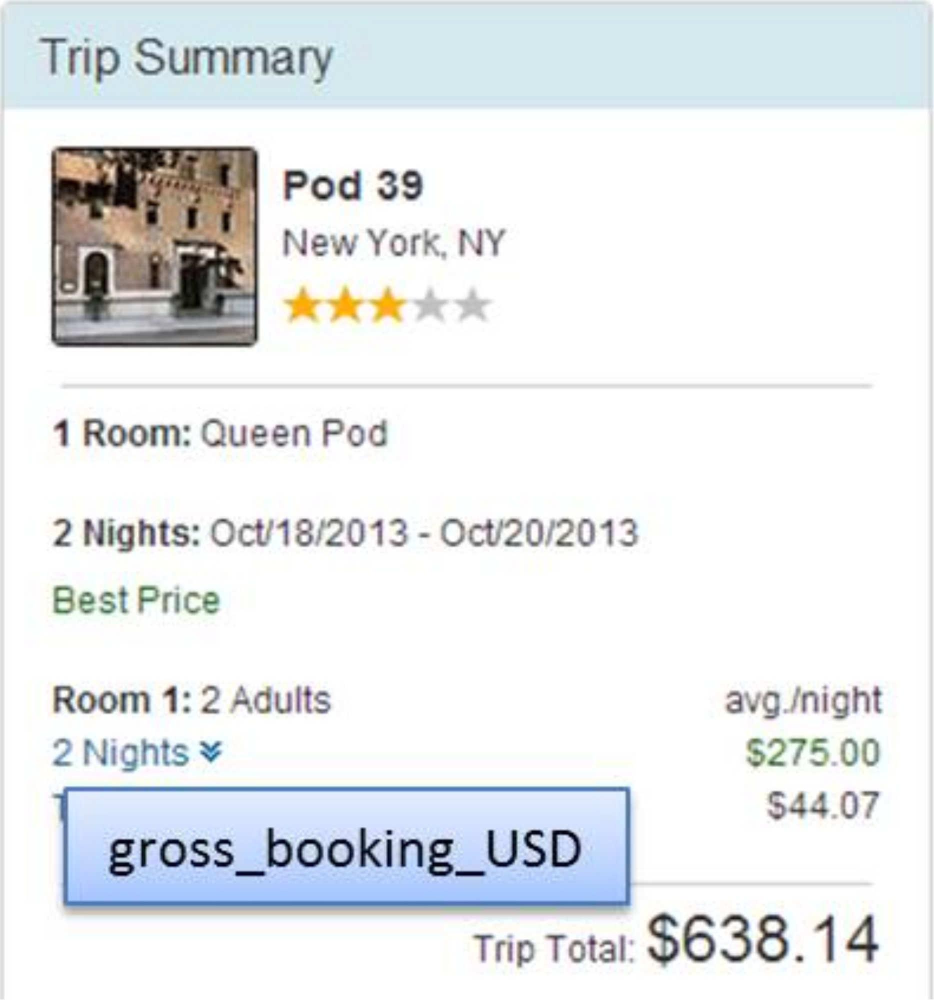
prop_id

prop_review_score *(rounded to 0.5)*

price_usd

Sale
Best Price
Pod 39
4.3 out of 5
New York
Map
1-866-267-9053
Only 5 rooms left at this price
$235
avg/night
Most Popular! 296 people booked this hotel in the last 48 hours

Assignment 2: The data (booking)



Trip Summary

Pod 39

New York, NY

★★★☆☆

1 Room: Queen Pod

2 Nights: Oct/18/2013 - Oct/20/2013

Best Price

Room 1: 2 Adults avg./night

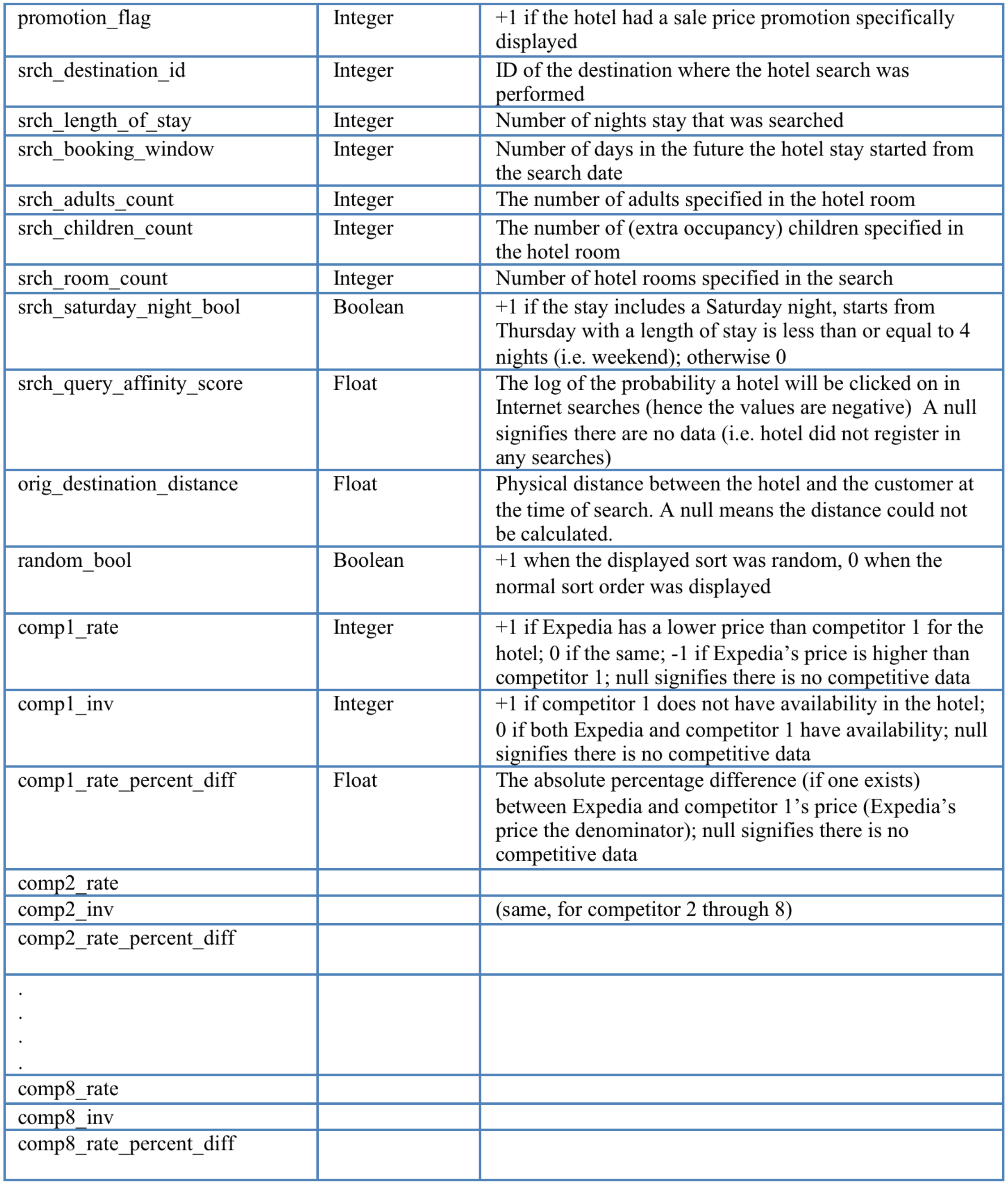
2 Nights $275.00

T $44.07

Trip Total: $638.14

gross_booking_USD

Field	Data Type	Description
srch_id	Integer	The ID of the search
date_time	Date/time	Date and time of the search
site_id	Integer	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ..)
visitor_location_country_id	Integer	The ID of the country the customer is located
visitor_hist_starrating	Float	The mean star rating of hotels the customer has previously purchased; null signifies there is no purchase history on the customer
visitor_hist_adr_usd	Float	The mean price per night (in US$) of the hotels the customer has previously purchased; null signifies there is no purchase history on the customer
prop_country_id	Integer	The ID of the country the hotel is located in
prop_id	Integer	The ID of the hotel
prop_starrating	Integer	The star rating of the hotel, from 1 to 5, in increments of 1. A 0 indicates the property has no stars, the star rating is not known or cannot be publicized.
prop_review_score	Float	The mean customer review score for the hotel on a scale out of 5, rounded to 0.5 increments. A 0 means there have been no reviews, null that the information is not available.
prop_brand_bool	Integer	+1 if the hotel is part of a major hotel chain; 0 if it is an independent hotel
prop_location_score1	Float	A (first) score outlining the desirability of a hotel's location
prop_location_score2	Float	A (second) score outlining the desirability of the hotel's location
prop_log_historical_price	Float	The logarithm of the mean price of the hotel over the last trading period. A 0 will occur if the hotel was not sold in that period.
price_usd	Float	Displayed price of the hotel for the given search. Note that different countries have different conventions regarding displaying taxes and fees and the value may be per night or for the whole stay



promotion_flag	Integer	+1 if the hotel had a sale price promotion specifically displayed
srch_destination_id	Integer	ID of the destination where the hotel search was performed
srch_length_of_stay	Integer	Number of nights stay that was searched
srch_booking_window	Integer	Number of days in the future the hotel stay started from the search date
srch_adults_count	Integer	The number of adults specified in the hotel room
srch_children_count	Integer	The number of (extra occupancy) children specified in the hotel room
srch_room_count	Integer	Number of hotel rooms specified in the search
srch_saturday_night_bool	Boolean	+1 if the stay includes a Saturday night, starts from Thursday with a length of stay is less than or equal to 4 nights (i.e. weekend); otherwise 0
srch_query_affinity_score	Float	The log of the probability a hotel will be clicked on in Internet searches (hence the values are negative) A null signifies there are no data (i.e. hotel did not register in any searches)
orig_destination_distance	Float	Physical distance between the hotel and the customer at the time of search. A null means the distance could not be calculated.
random_bool	Boolean	+1 when the displayed sort was random, 0 when the normal sort order was displayed
comp1_rate	Integer	+1 if Expedia has a lower price than competitor 1 for the hotel; 0 if the same; -1 if Expedia's price is higher than competitor 1; null signifies there is no competitive data
comp1_inv	Integer	+1 if competitor 1 does not have availability in the hotel; 0 if both Expedia and competitor 1 have availability; null signifies there is no competitive data
comp1_rate_percent_diff	Float	The absolute percentage difference (if one exists) between Expedia and competitor 1's price (Expedia's price the denominator); null signifies there is no competitive data
comp2_rate		
comp2_inv		(same, for competitor 2 through 8)
comp2_rate_percent_diff		
. . . .		
comp8_rate		
comp8_inv		
comp8_rate_percent_diff		

For training data only…

position	Integer	Hotel position on Expedia's search results page. This is only provided for the training data, but not the test data.
click_bool	Boolean	1 if the user clicked on the property, 0 if not.
booking_bool	Boolean	1 if the user booked the property, 0 if not.
gross_booking_usd	Float	Total value of the transaction. This can differ from the price_usd due to taxes, fees, conventions on multiple day bookings and purchase of a room type other than the one shown in the search

Assignment 2

- You should provide:
  - A ranking of hotels based on likelihood of booking
  - For each search you will get a number of hotels, and you should rank them using your algorithm
- Some initial questions:
  - Could you just use the data as it is, or should you combine multiple records?
  - What kind of algorithm could be suitable for this task?

Assignment 2

```
Perfect'    [      1]    [27.0662]
   '047'         [0.5186]    [      10]
   'Kaggle'      [0.5127]    [ 9.7903]
   '100'         [0.5105]    [ 9.7121]
   '040'         [0.5101]    [ 9.6973]
   '044'         [0.5065]    [ 9.5684]
   '015'         [0.5047]    [ 9.5043]
   '077'         [0.5000]    [ 9.3405]
   '042'         [0.4998]    [ 9.3310]
   '033'         [0.4995]    [ 9.3223]
   '030'         [0.4987]    [ 9.2928]
   '099'         [0.4948]    [ 9.1537]
   '080'         [0.4940]    [ 9.1265]
   '090'         [0.4921]    [ 9.0602]
   '060'         [0.4907]    [ 9.0081]
   '039'         [0.4902]    [ 8.9922]
   '009'         [0.4886]    [ 8.9337]
   '006'         [0.4871]    [ 8.8812]
   '011'         [0.4863]    [ 8.8519]
   '024'         [0.4835]    [ 8.7540]
   '005'         [0.4826]    [ 8.7214]
   '003'         [0.4785]    [ 8.5759]
   '036'         [0.4759]    [ 8.4849]
   '022'         [0.4758]    [ 8.4797]
   '001'         [0.4749]    [ 8.4489]
   '025'         [0.4748]    [ 8.4465]
```

```
'055'         [0.4737]    [ 8.4060]
'018'         [0.4671]    [ 8.1726]
'032'         [0.4655]    [ 8.1168]
'046'         [0.4647]    [ 8.0869]
'017'         [0.4626]    [ 8.0150]
'013'         [0.4576]    [ 7.8358]
'035'         [0.4572]    [ 7.8213]
'070'         [0.4482]    [ 7.5018]
'016'         [0.4375]    [ 7.1250]
'043'         [0.4341]    [ 7.0036]
'007'         [0.4273]    [ 6.7602]
'048'         [0.4197]    [ 6.4915]
'020'         [0.4194]    [ 6.4831]
'096'         [0.4184]    [ 6.4467]
'026'         [0.4115]    [ 6.2011]
'038'         [0.4082]    [ 6.0848]
'012'         [0.4018]    [ 5.8582]
'072'         [0.3600]    [ 4.3765]
'091'         [0.3500]    [ 4.0214]
'008'         [0.3499]    [ 4.0192]
'028'         [0.3494]    [ 4.0005]
'Random'      [0.3494]    [      4]
'027'         [0.3494]    [ 3.9988]
'051'         [0.3493]    [ 3.9964]
'031'         [0.3491]    [ 3.9878]
'021'         [0.3471]    [ 3.9201]
'023'         [0.3292]    [ 3.2854]
'037'         [0.3269]    [ 3.2039]
'010'         [0.3235]    [ 3.0826]
```

Assignment 2

- What is expected of you?
  - Prediction file with your answer (score counts 20%)
  - Final report (grading based on selected techniques, quality of evaluation, rationale, writing style and creativity) (score counts 60%)
    - Max 10 pages LNCS
  - Process report (score counts 20%)
    - Who did what and why, how did the cooperation between group members go
  - Presentation (top 3 and random 3)