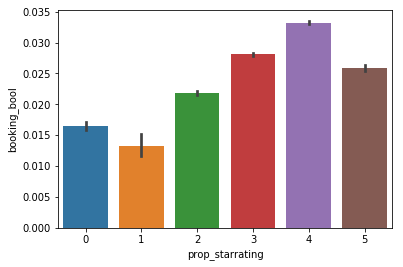
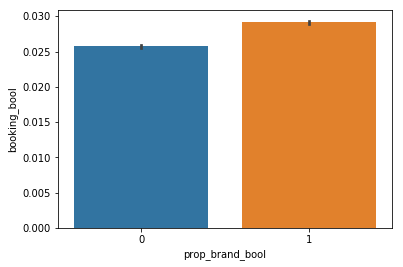
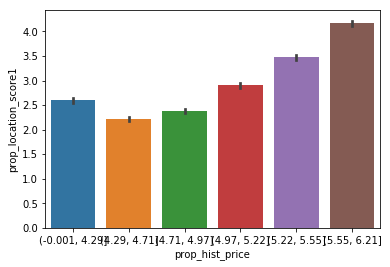
* Visitor\_location\_country id: 210 unique values, nothing missing.
* Site\_id: 34 unique values, nothing missing.
* Visitor\_hist\_starrating: 250000 values, lot missing (5% available). Floats between 0 and 5. Categorized in 10 bins (0-0.5 = 1, 0.5-1 = 2, etc.), missing values categorized as 0.
* visitor\_hist\_adr\_usd: 252988 values, lot missing (5,1% available). Represents mean price spend per night in previously visited hotels. Categorized in 4 quartiles (1, 2, 3, 4). Missing values categorized as 0.
* Prop\_country\_id: 172 unique values, nothing missing.
* Prop\_id: 129113 unique values, nothing missing.
* Prop\_starrating: 6 unique values (0, 1, 2, 3, 4, 5), nothing missing.



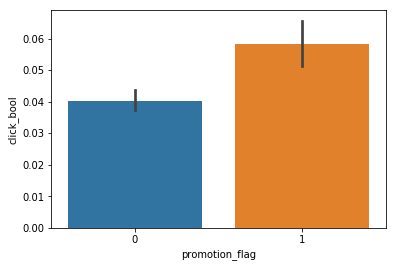
* Prop\_review\_score: 10 unique values (0, 1, 1.5, 2, etc., only 0.5 is missing), approx. 7000 missing values (>99% available).
* Prop\_brand\_bool: represents whether a hotel is part of a chain (1 if yes, 0 if no). No missing values.



* prop\_location\_score1: first location score of hotel, on a scale of 0 to 7. No missing values, 337 unique values. Scores are rounded to integers.
* Prop\_location\_score2: second location score of hotel, on a scale of 0 to 1. Approx. 1 million missing values, 78% available, 9342 unique values. Missing values are replaced with the mean location score. All scores are rounded to 1 decimal.
* prop\_log\_historical\_price: 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. 392 unique values, nothing missing. Categorized in 6 quantiles.



* Position: representing hotels position on the search page. No missing values, 40 unique values.
* Price\_usd: price displayed for this hotel in this search. No missing values, approx. 76000 unique values between 0 and 19.726.330 (??!). The value may be for the whole stay or a single night. Data is categorized in 6 quantiles.
* Promotion\_flag: 1 if a promotion flag was displayed, 0 if not. No missing values.



* Srch\_destination\_id: the id of the spot the user searched for a hotel. No missing values, approx. 18000 unique values (integers).
* Srch\_length\_of\_stay: number of nights that was searched. No missing values, 36 unique values. Since most of the search were for 1-4 nights, searches for 5 or more were grouped together (as ‘5’).
* Srch\_booking\_window: number of nights between the time of the search and the start of the stay. Integers between 0 and 492 (429 unique values) (half is less than 17). No missing values. The numbers were categorized by rounding to the number of weeks the booking was in advance, and when the booking took place 12 or more weeks beforehand they were grouped together (as ‘12’).
* Srch\_adults\_count: number of adults the hotel was searched for. No missing values, ranges between 1 and 9 (although at least 50% for 2). Since few users searched for 3 adults or more, these were grouped together (as ‘3’).
* Srch\_children\_count: number of children the hotel was searched for. No missing values, ranges between 0 and 9, average is 0.35 (mostly 0 children). 3 or more children searches are grouped together, as they are very few (as ‘3’).
* Srch\_room\_count: number of rooms searched for. No missing values, ranges between 1 and 8, average is 1.11. Since so few people looked for 2 or more rooms, these were grouped together (as ‘2’).
* Srch\_Saturday\_night\_bool: 1 if search included a Saturday night, 0 if not. No missing values. About 50-50.
* Srch\_query\_affinity\_score: 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). 317406 values available, which is approx. 6%, the rest did not register in searches. The values which were available were grouped in 4 quantile groups, labelled 1, 2, 3 and 4. The missing values were labelled as 0.
* Orig\_destination\_distance: distance between the customer at the time of the search and the hotel. 3,35 million values available, which is approx. 67% of the dataset. Some of the missing data are replaced by the mean distance between other data with the same visitor\_location\_country\_id and prop\_country\_id. This led to 84% of the data. The data was grouped into 6 quantiles (labelled 1, 2, 3 etc.), the missing data labelled as 0.
* Random\_bool: 1 if the order of search results was random, 0 if not. Mean is 0.29. No missing values.
* Comp\*\_rate (\* is 1-8, the different competitors): 1 if Expedia has a lower price than the competitor for the hotel, 0 if the same, -1 if Expedia has a higher price. All competitors are combined, first in all\_comp\_rates: here, the minimum value is taken of all competitors, so if the hotel has one -1 the score is -1 (since the customer would go there, this is the objective market position). Hotels were all values are missing are labelled ‘2’. Secondly, a category is created where all values are added (Nulls count as 0).
* Comp\*\_inv: 1 if Expedia has a room available and the competitors doesn’t, 0 if Expedia and competitor both have rooms available. The values of all rooms are added in category count\_comp\_inv, representing the number of competitors that do not have rooms available. Missing values are replaced with 0.
* Comp\*\_rate\_percent\_diff: difference in price, how much expedia is cheaper than competitors. Average of all competitors is taken, and divided in 4 quantile groups (labelled 1, 2, 3, 4). When all values are missing it is labelled 0.
* Gross\_bookings\_usd: 138390 values available, 2,7%. The difference between this price and the first shown price is calculated, and set to 0 if value is missing. Categorized in quantiles 1-4.
* Date\_time: three new categories created: srch\_month, srch\_year and srch\_daytime (0 = night, 1 = morning, 2 = afternoon, 3 = evening).
* New category: booking\_month. Based on which month was searched for, determined by the date of the search and the time between search and stay.