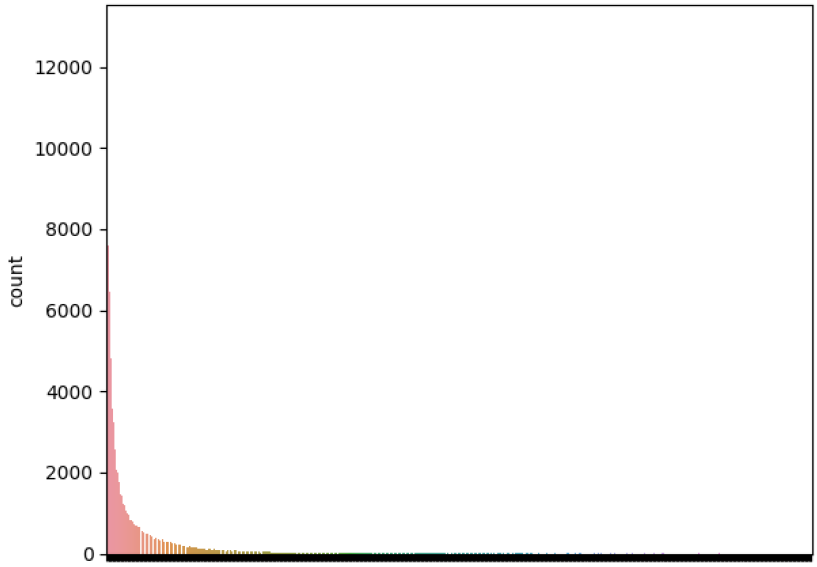
What to find? Sort hotel on the likeliness that it will be booked

Unique search: 199795 – All the unique ids
Booked ids: 138390 – Ids that actually have booked
Unique booked ids: 138390 – Meaning there are no ids with double booking
Unique hotels: 129113 – Number of hotels that are available in the train set
Number of unique listings: 7448 - Number of listings not in both train and test

Hypothesis: If a hotel is showed more often, it will be booked more often.

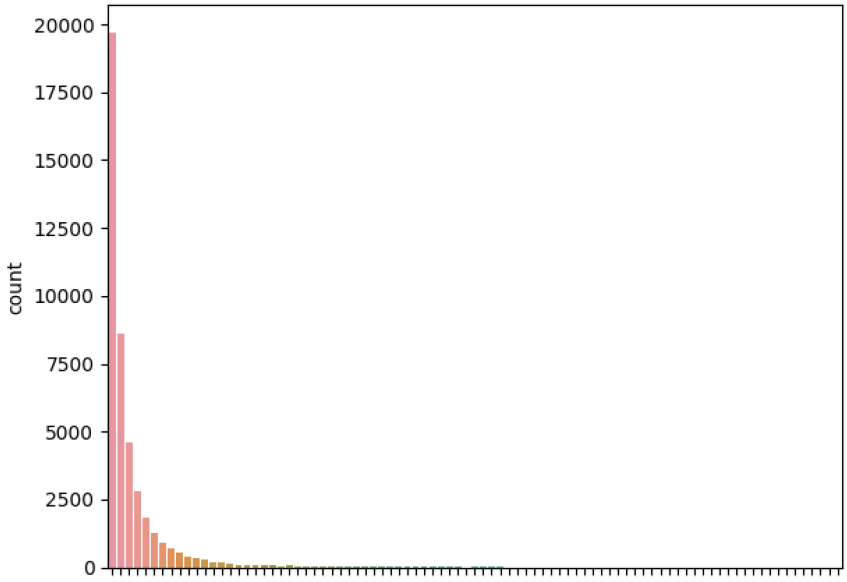
The amount of times a hotel is showed to a user



You can see that there are various hotels which are showed many times more often in comparison with other hotels. Difference between hotels that are only showed once (minimum) and hotels that are showed 2357 times (maximum).

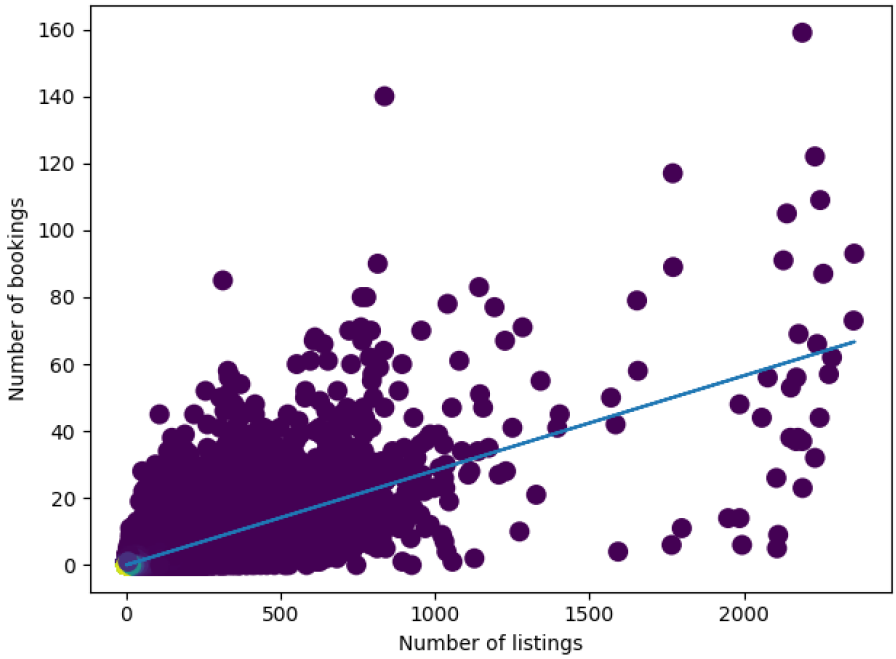
The question is, do these two figures correlate?

The number of times a hotel is booked



As you can see, not many (just ~25000) hotels are actually booked out of the 129113 total hotels. Each line on the x-axis represents the amount of times booked x=1 is 1 time booked up to x=159, which means that there is a hotel that is booked 159 times.

Correlation number of listing vs bookings



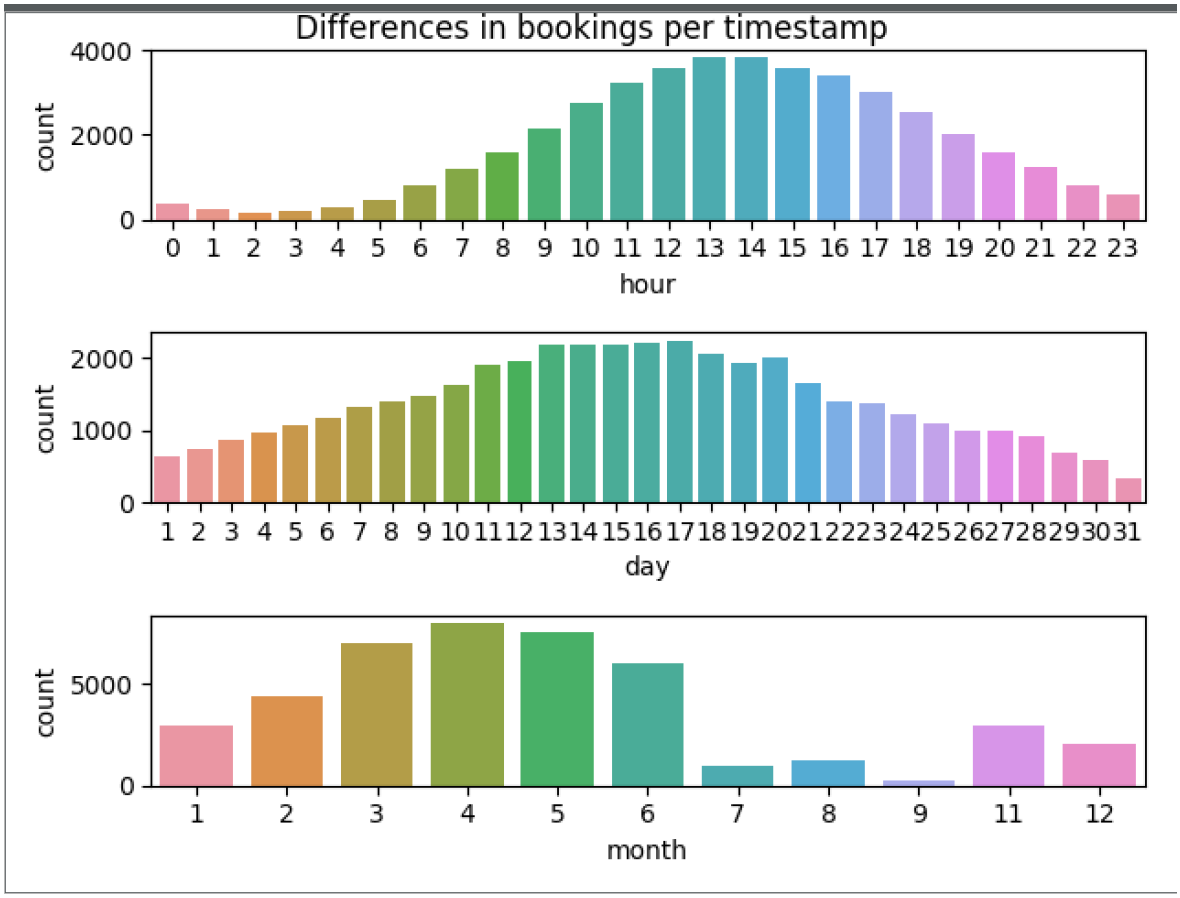
As you can see, there is a correlation between how many times a hotel is listed and actually booked. This isn't 1-on-1 but you probably want to correct for this correlation in your analysis.

Hypothesis; Each hotel has a higher probability being booked on a specific time pattern.

By plotting the time when a hotel is booked, we should see that there is a non-uniform distribution between the time series.

Node: If a hotel is booked more than once, we take the median of that time series.

We split the booking time series into 3; hour, day and month.



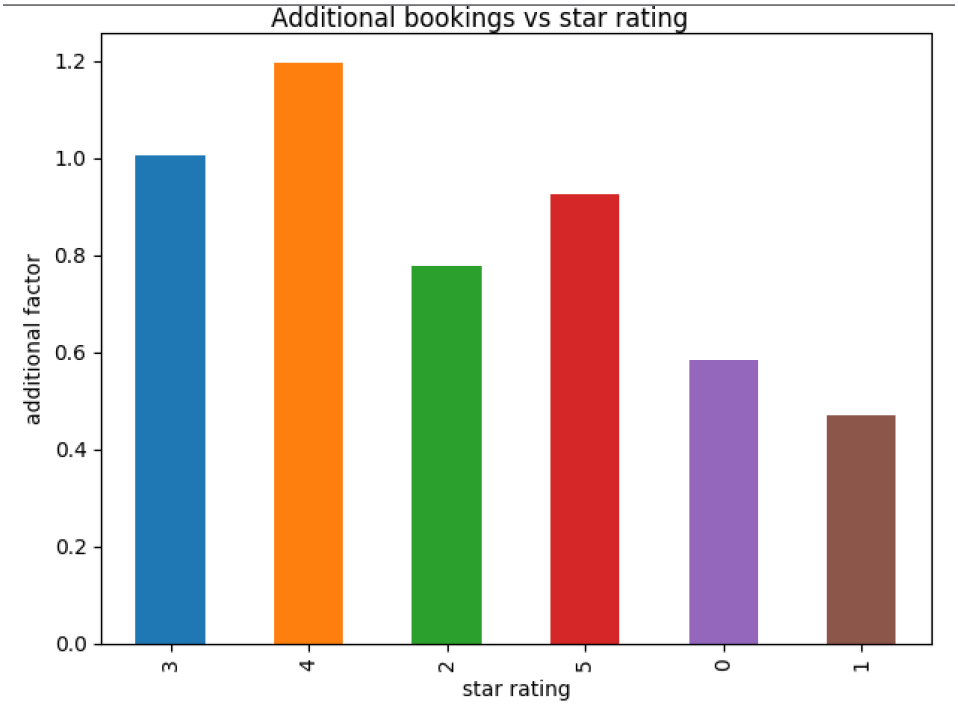
As you can see, the data is non-uniform distributed and there is probably information within the timestamp between different listings.

One downside of these features mentioned above is that all the listings in a search share the same properties, which means that you cannot distinguish different properties from these classes. So now we are going to focus first on the classes which are unique for each property id.
Classes that are unique for each property:
- **id (just an unique id)**
- **starrating**
- **review_score**
- **promotion_flag**
- **price_usd**

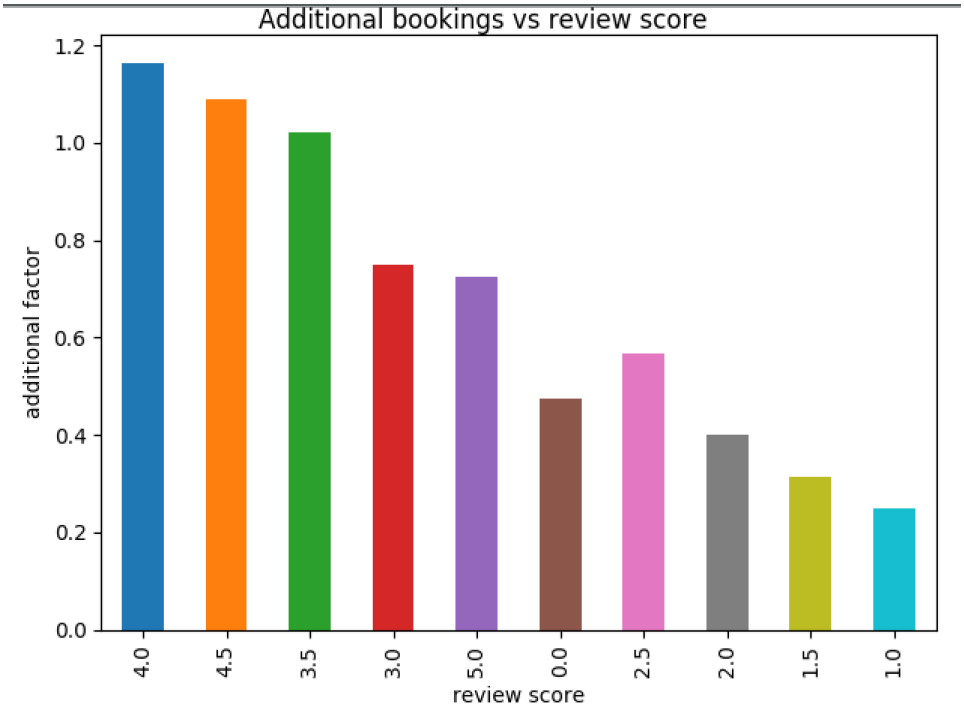
So lets focus on these first.

Star rating
Hypothesis: The star rating depends on the booking choice.

How can we visualize this?
- Count the bookings per start rating, divide them by their total and calculate the difference.



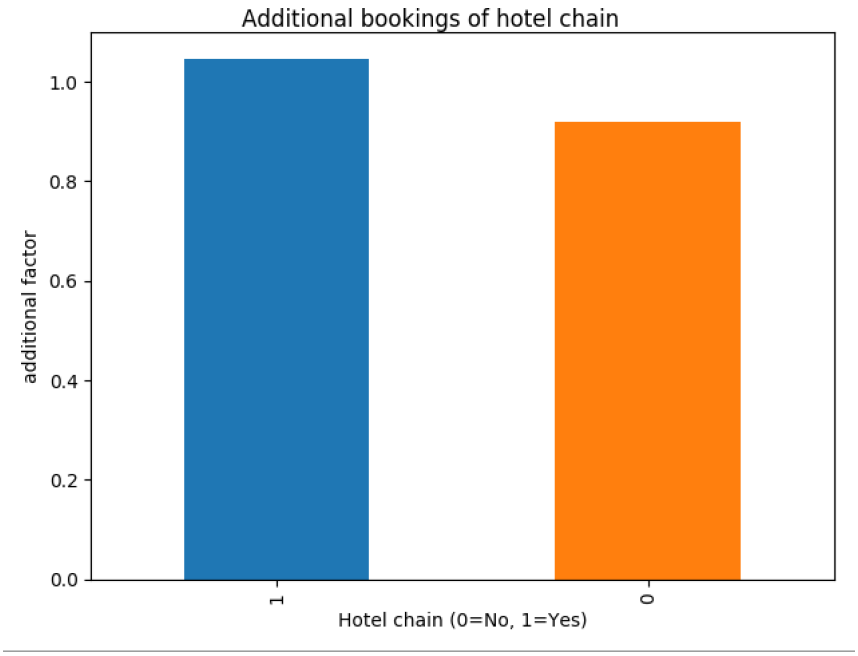
As you can see, the most additional booked hotels are the one with 4 star rating, which clearly says something about the probability of the choice of interest. (0 = no star rating)



Just like the star rating, you can expect the review score to say something about the probability of something gets booked. The 4 and 4,5 stars are ”overbooked” in comparison with the others.

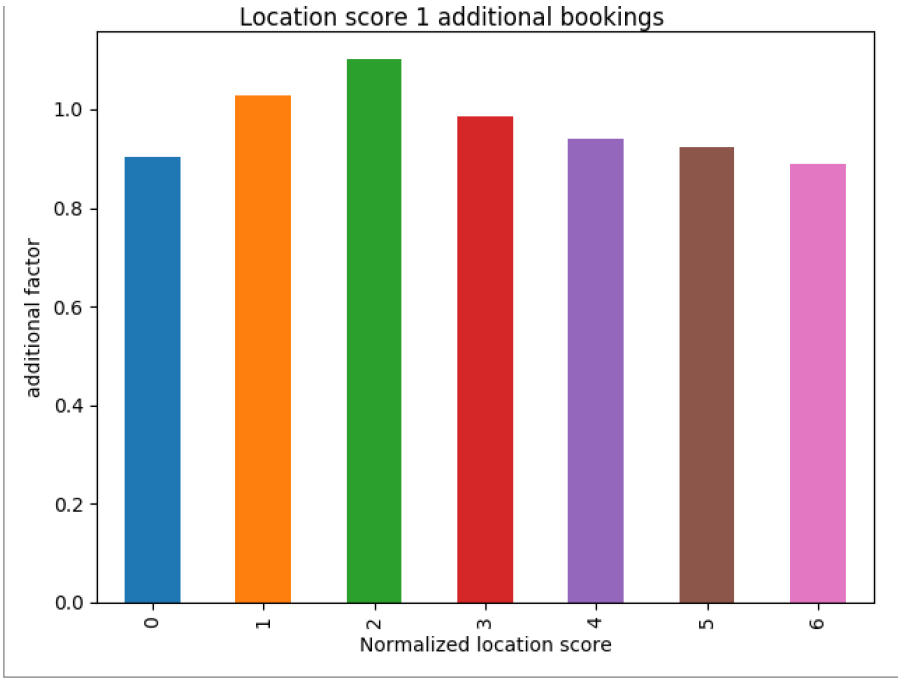
Just like the star rating per property, the review is an indicator as well for a hotel being booked. (0 = no review score)

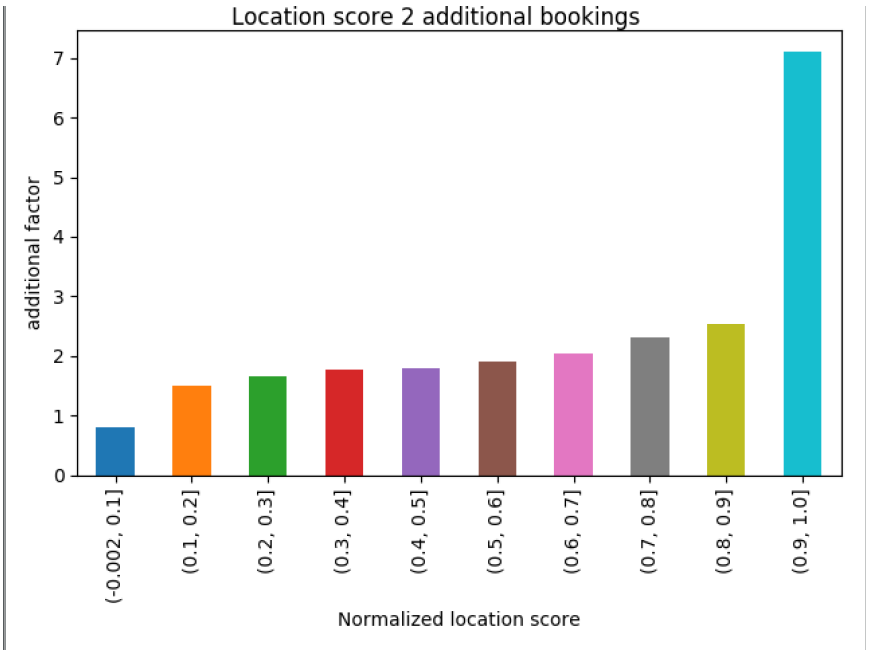
Another class to look at is if a hotel is part of a big hotel chain en compare if customers are more likely to book if this is the case.



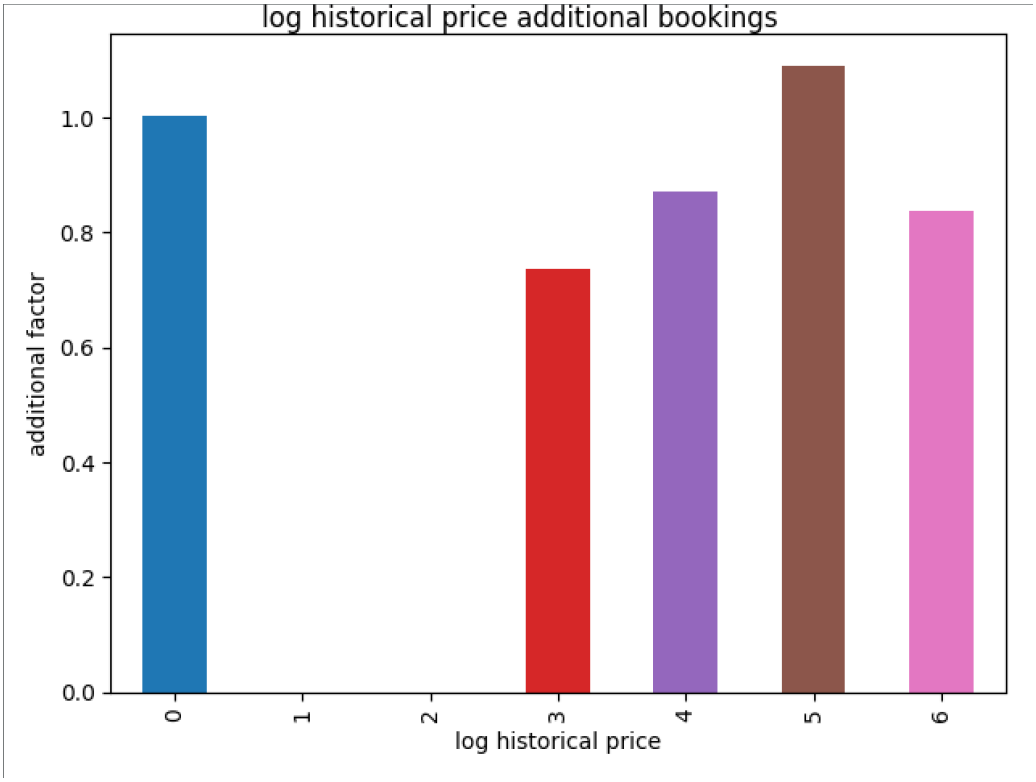
As you can see, customers are more likely to book a hotel if it is part of a bigger hotelchain.

Another feature is the Location score. Both of these features have a 0 which means there is none. Looking at the graphs below, I have to add that the 0 is overrepresented even though these location scores tell much about if a hotel is getting booked. Location score 2 is a factor 7 more likely to get booked in comparison with the others (which is likely zero).

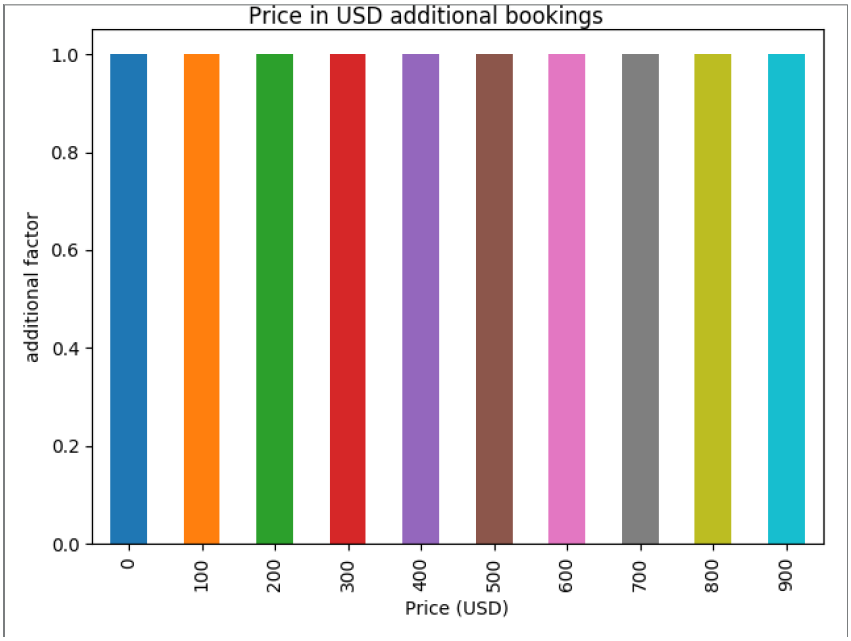




The price is obviously important for someone to look at before booking but we also go the log historical price, so this could be a factor of interest as well. (0 means no historical price)

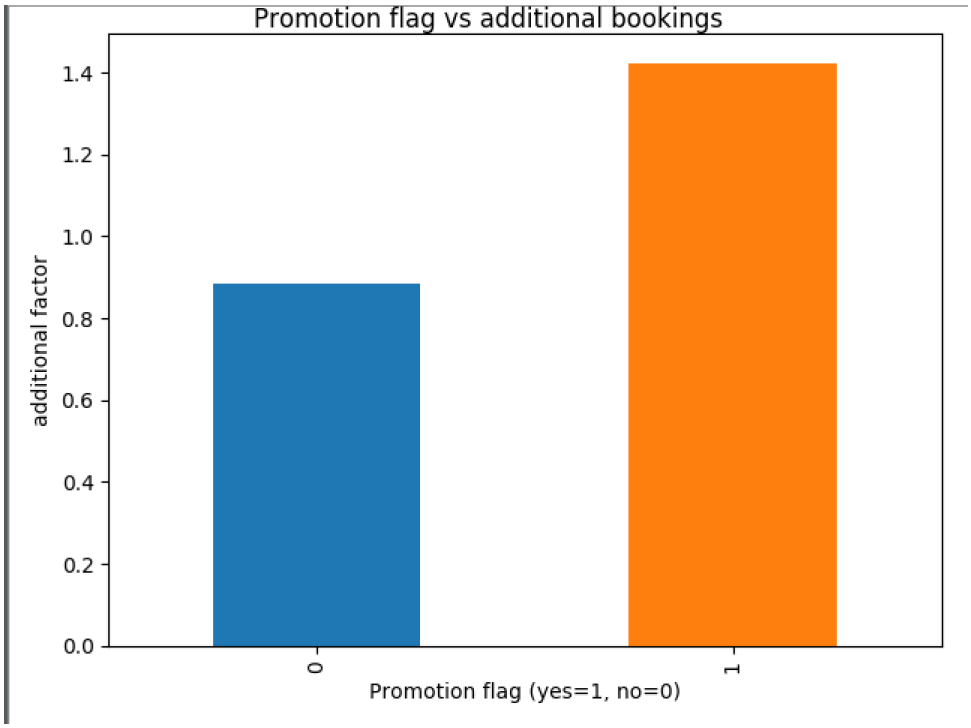


Besides the historical price, you got the actual price required for booking, lets plot this.



As you can see, there is almost no more interest in a specific price range in comparison with the others. So this is probably not a feature you want to focus on maybe.

Even though it is a lousy marketing trick, some hotels have a “promotion flag”, so this will probably increase the likelihood that someone buys that specific hotel.



Indeed, people are indeed more likely to click on such hotels.

These features listed above were all the hotel specific features.