## Introduction

Have you ever used Airbnb to rent a place to stay? Airbnb provides accommodations for visitors.

If you look for a place on Airbnb website, do you prefer to make an appointment for the first place you see? I can guess you do not as most of us use to.

Well, which criteria do you take into consideration? For instance, do you check for the Internet services? Maybe you are one who reads reviews and make the selection.

Can your choice be a model that is generated by your brain? When you see a data to make decision, our brains start to process this data and try to find an optimal solution.

This is like a piece of cake is in front of you and have chance to eat it or not. This decision depends on which idea has more weight in your brains model as enjoying the moment, keeping a healthy life, maintaining to look good to boys/girls around etc. You will probably prefer not to eat if the latter has ten times more weight than others in the model that your brain build.

It is sure, decisions are not such easy, formula behind our brain is much more complex. However, we will simplify Airbnb room selection problem to understand the general tendency. Main purpose is to apply basic data preprocessing steps and interpret the data.

Three points will be marked in this text. These are:

1. Which property types are preferred by customers? Is there a demand-supply gap?
2. Which services is more effective to attract customers?
3. Can we guess the reservation rates according to the amenities or review scores?

## Demand by Property Types

First of all, property types of hosts in Seattle is observed to see properties’ distribution and understand the guests’ tendency. At the below chart, property type rate means proportion of the property type that is stated in the first column among all properties in Seattle. Then, busy day rate is the proportion of appointed days for this property type among all reservations.

For example, houses cover about 45 percent of all properties, and about 42 percent of all reservation to Seattle is made for a house type property.

This graph clearly tells us, Airbnb hosts in Seattle consist mostly of houses and apartment. Furthermore, we can see that houses can not attract much customers as hosts supply while apartments, townhouses and condominiums, next three most chosen properties, take the big piece.

Can we conclude that investing in Airbnb with an apartment is more reasonable than houses? Or… Might there be other factors that affect the customers’ choices like price, location?

Second question seems more logical, just one graph is insufficient to conclude a problem. Anyway, this plot gives some idea about property preferences in Seattle and makes door open to new questions.

Machine generated alternative text:
d iff 
3 
5 
property type 
House 
Apartment 
Townhouse 
Condominium 
Bed & Breaktast 
Cabin 
Bungalow 
Camper/RV 
Treehouse 
Chalet 
Yurt 
busy_day_rate 
0.426832 
0.470218 
0_034675 
0_030479 
0_010260 
0_007937 
0_002349 
0_005628 
0_003133 
0_002880 
0_001405 
0_002072 
0_000919 
0_000139 
0_000662 
0100410 
btw_demandsupply 
-0,027189 
0_003761 
006638 
-0_000220 
-0_001756 
-0_003415 
0_000126 
-0_000273 
-0_000525 
-0_000691 
0_000762 
0_000133 
-0_000385 
0_000138 
0,000148 
property_type_rate 
0.454021 
0 _ 447472 
0_030914 
0_023841 
0_010479 
0_009693 
0_005764 
0_005502 
0_003406 
0_003406 
0_002096 
0_001310 
0_000786 
0_000524 
0_000524 
0100262 

## Never Without Internet!

I guess most people check the services while renting a place. Gripping question is which services are wished at most. Fortunately, Airbnb dataset has a column that tells the amenities that host serve.

Since all the services are included only in a column, it is hard to understand general trend in the data. In detail, even if a host serve ten different amenities, customers should probably take care only a few of them like the Internet.

In order to see effects of each service, the data is modified to show services existence in separate columns. In short, it is easier to interpret columns that show each amenity, like heating, is available for a host rather than a list of services.

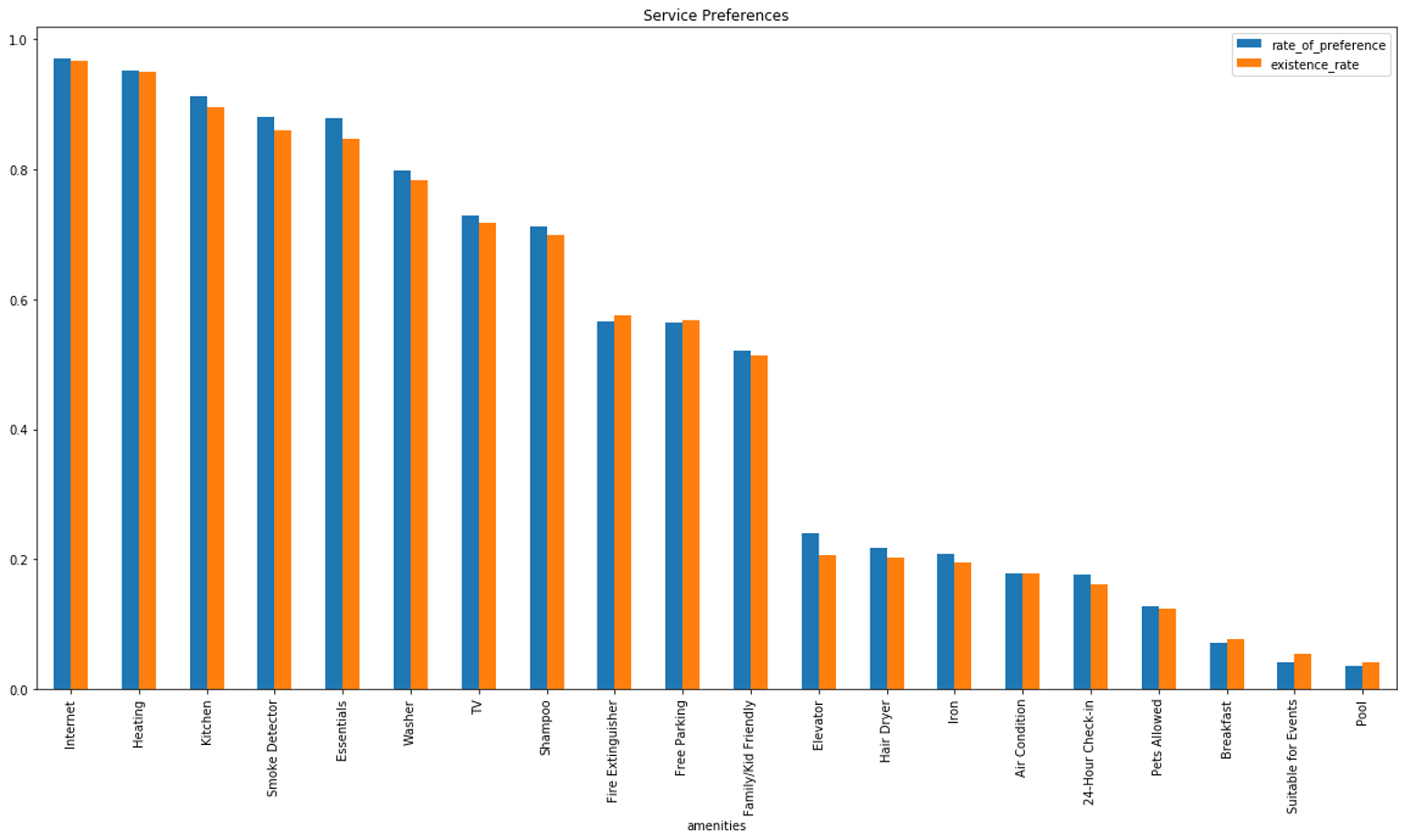
We can see that some of services are preferred by both hosts and guests. After shampoo, there is dramatic difference for existence of a service and its preference.

Moreover, we can say from this graph host are generally aware of which amenities they should serve to attract customers since there is no huge difference between preference and existence rate for all columns.

In the last question, this data has been used to see effect of amenities on appointment rates. However, only services before the dramatic difference is taken into consideration.

Well, what is bad to see influences of other amenities? Can’t iron service also attract customers?

Try to guess this, last part will try to answer it.



## Time to Bet: Forecasts

Now this is most exciting part! Let’s see whether reviews or amenities play the crucial role in customers’ decisions.

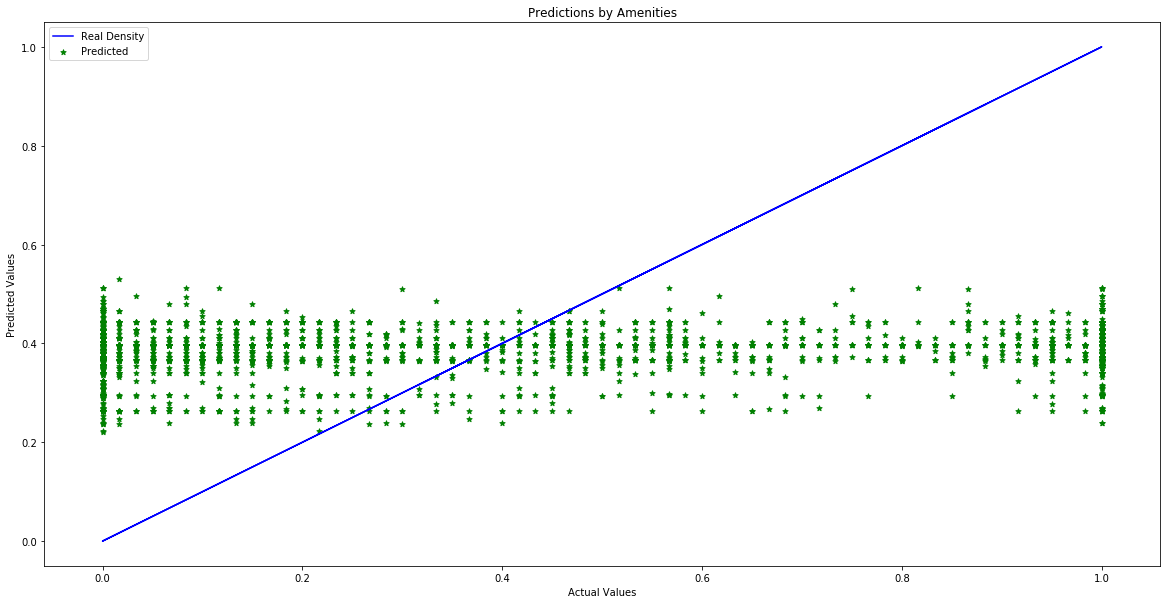
Our purpose is to forecast ‘two-month density’ for hosts. ‘two-month density’ is fullness of host for next two month and value between 0 and 1. Airbnb data actually has one-year data but most of the appointment seen to be arranged for next two months. The conjecture is two-month density data is very similar, three-month density or a-year density, so I think it is enough to guess just one value for now.

A linear regression model is fit to our data for all cases. This model is open to discussion but the main purpose of this project is to apply data preprocessing steps and to see general process as a first data science project. Models will be crucial to interpret results, so also the logic and math behind it.

At the first scenario, it is assumed only amenities influence the two-month density rates. I care only most requested services. Here is the place to answer question asked at the end of previous part.

I try to avoid overfitting and enable the model not to think an irrelevant factor as a determiner. However, the results are not as pleasant as I guessed.

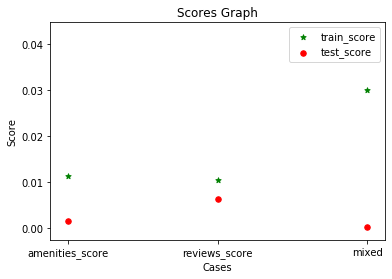
Below chart, there are actual fulness rates for hosts at the horizontal line and predicted values by first case. Green dots show prediction, and the blue line shows what they actually should. As you can see trends in predictions and reality is irrelevant.



At the second model, I tried to build a model that only depends on review points of host. However, the outcome is still not heartwarming.

For the last case, model is trained with both review and amenities. Moreover, property is also added them. Since the graph does not change so much, even goes worse, I prefer not to put them here.

Below chart shows r -square, a method to measure the performance of a model, scores can be seen for each case. Unfortunately, all the results are so close to zero! ☹



## Conclusion

Does it look like we could not get any outcome from there?

These trials have at least shown us our simple linear regression models are not sufficient to link this input with desired outcomes. Maybe assumption was totally wrong, input is absolutely unrelated with behaviors of customer and these trials provide such an experience to eliminate wrong ways.

Data is like an ocean, it seems scary to step in but when dive into it, the seabed is full of beauties and very intriguing. More you dive, more you will desire to explore around.