Airbnb Cost Analysis in Seattle and Boston

Moshiul Azam, Jinhao Zhang
CSC 440
April 27, 2020

Introduction

Airbnb is the world's leading online marketplace for arranging and offering rental accommodations, providing short-term rental housing and tourism-related services[1]. For travelers, it has become an increasingly popular alternative to regular hotel booking websites. It helps to directly connect those who have extra rooms or apartments with travelers who need short -term accommodation. Each year there are millions of housing lease transactions all over the world that happens on Airbnb.

Under the circumstance, Airbnb built a large database for storing those data since 2008. We found the data on Inside Airbnb site which is sourced from publicly available information from the Airbnb site[2]. The data has been analyzed, cleansed, and aggregated where we can do some further analysis considering the relationship between price and other stuff. In this paper, we will delve into these datasets, which include three .csv files each: calendar, listings, and reviews.

We'll mainly focus on the data from two major cities: Seattle and Boston, which respectively stand for the west coast and east coast of the United States. We'll manage to figure out the reasons that rental price changes what it leads to.

Previous work

Previous research[3][4] on predicting the price of Airbnb fell in two fields: econometrics and machine learning, which mainly focus on predicting itself with no practical significance. The main goal here is to analyze the reasons behind rental prices, find interesting patterns of the reviews, figure out what makes the price change, which could be used to assist Airbnb to develop a better strategy of attracting customers.

Dataset description

We used two datasets which are the following:

- Boston[5]
- Seattle[6]

These datasets are part of data inside Airbnb, we'll analyze them and do the comparison.

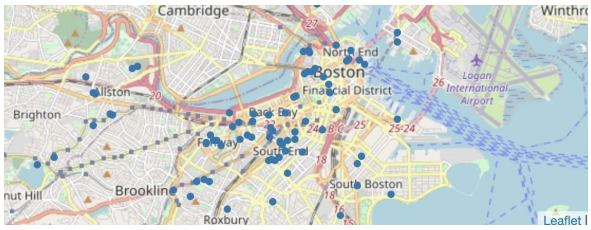
The following Airbnb activity is included in each dataset:

- Listings, including full descriptions and average review score
- Reviews, including unique id for each reviewer and detailed comments
- Calendar, including listing id and the price and availability for that day

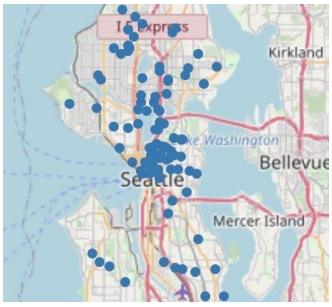
We'll mainly use Listings files for most of the analysis, including finding the rules behind the change of rental price, basic reviews analysis, and so on. For Calendar, we'll use regression in order to predict the review score based on the former comments. We also will discuss the availability of the properties in both cities throughout the year. For Reviews, we'll use the word cloud to find the common patterns in the comments.

Based on their locations, we use some samples to draw two general maps to demonstrate their distributions.

Distribution map of Boston:



Distribution map of Seattle:



Methodology and Experiment

First, We imported some libraries and we used all for different purposes like:

- Numpy[7] for numeric calculations on numeric columns
- Pandas for data input and structure building
- Wordcloud[8] for showing the description column with the most used words
- Matplotlib[9] and seaborn for the visualization of the data columns

After that We loaded both Boston and Seattle datasets.

Exploratory Data Analysis

We did some EDA exploratory data analysis which is the following:

- We checked all the columns of both datasets
- Data types of both. There are three data types which are object, float, and integer
- We checked the null/missing values. Boston has 3585 null values and Seattle has 3818.
- Information of both dataset
- Total length rows in the datasets. Boston 3585 and Seattle 3818
- Boston data shape is (3585 rows and 95 columns) and Seattle (3818 rows and 92 columns)
- Data description of both datasets with their numeric columns

Data Cleaning

We did data cleaning and pre-processing on both datasets which are the following:

- We checked the columns of Boston that contain 40% missing values.

```
{'access',
 'has_availability',
 'interaction',
 'jurisdiction_names',
 'license',
 'monthly_price',
 'neighbourhood_group_cleansed',
 'notes',
 'security_deposit',
 'square_feet',
 'weekly_price'}
```

- We checked the columns of Boston that contain 40% missing values.

```
{'license',
 'monthly_price',
 'notes',
 'security_deposit',
 'square_feet',
 'weekly_price'}
```

- After checking we remove these columns with the drop null function.
- We removed the dollar $ sign from PRICE column
- After removing the null values, there were also missing values in the columns of both datasets so fill the values with the forward fill method of pandas.
- When we check again, there are 0 zero null missing values.

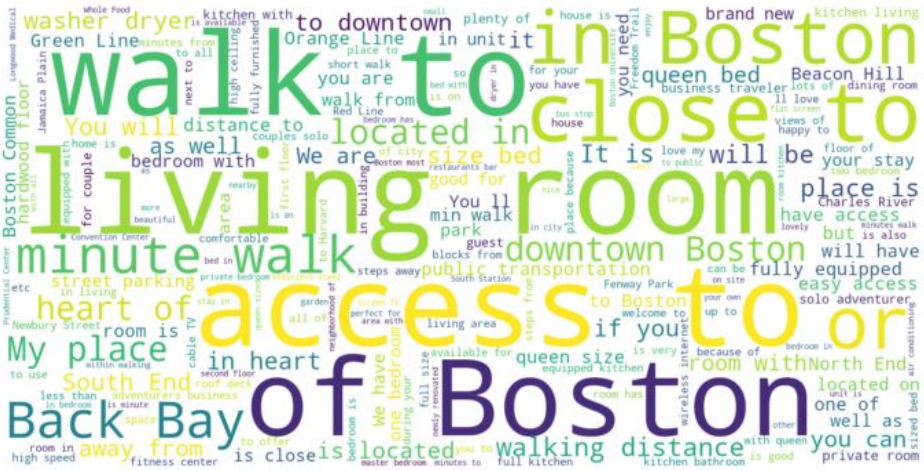
Patterns mining of Descriptions and Comments

Descriptions word cloud

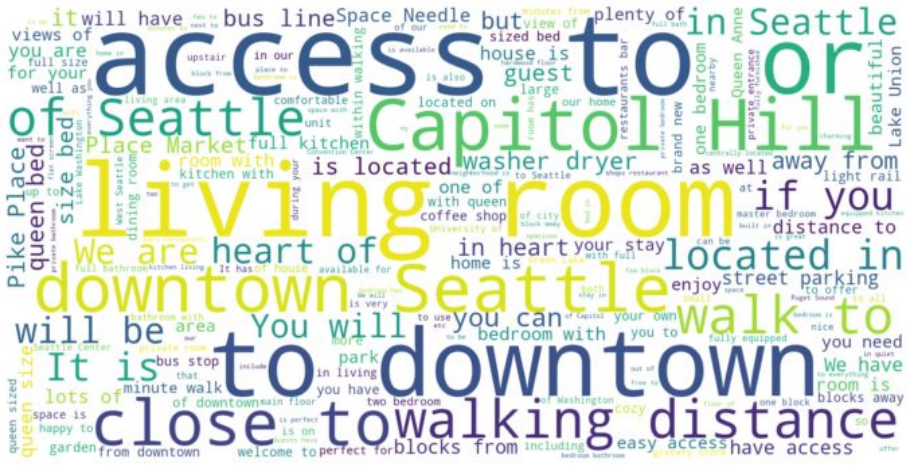
We used the word cloud library and created word clouds of description columns to look at which words are being used in both datasets.

In order to find the specific words that the owners used to describe their housing, block some words deliberately is necessary. Here we blocked the words such as 'apartment', 'the', 'and', 'there', 'this'.

Boston words cloud on the description:



Seattle words cloud on the description:



As we can see, homeowners in both cities tend to attract customers by explaining the location of the house and the conditions of the living room. They used ‘close to’, ‘walking distance’, ‘access to’ to describe how excellent their locations are. Typically, for Boston, it mentions ‘Back Bay’ a lot of time, which is a fashionable shopping destination. It also suggests the rental price there could be higher than others. for Seattle, descriptions tend to include ‘downtown’, ‘Capital Hill’ to attract customers, where there are many shops and restaurants.

Comments word cloud

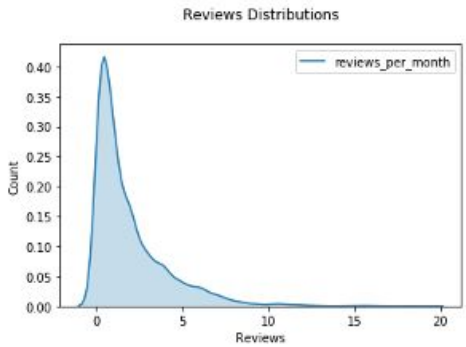
We also added comment word clouds to make it more interesting.
Boston words cloud on the comments:

Seattle words cloud on the comments:

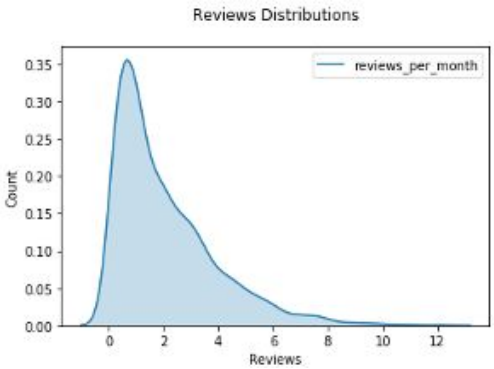
From the above pictures, we know that customers in both cities also mentioned ‘great location’, ‘walking distance’ as owners do. Also, they like to use some words to express praise such as ‘great host’, highly recommended’. Other words are literally identical to these words, which imply how scarce are the words used by customers to express complement.

Reviews distribution

Reviews distribution of Boston



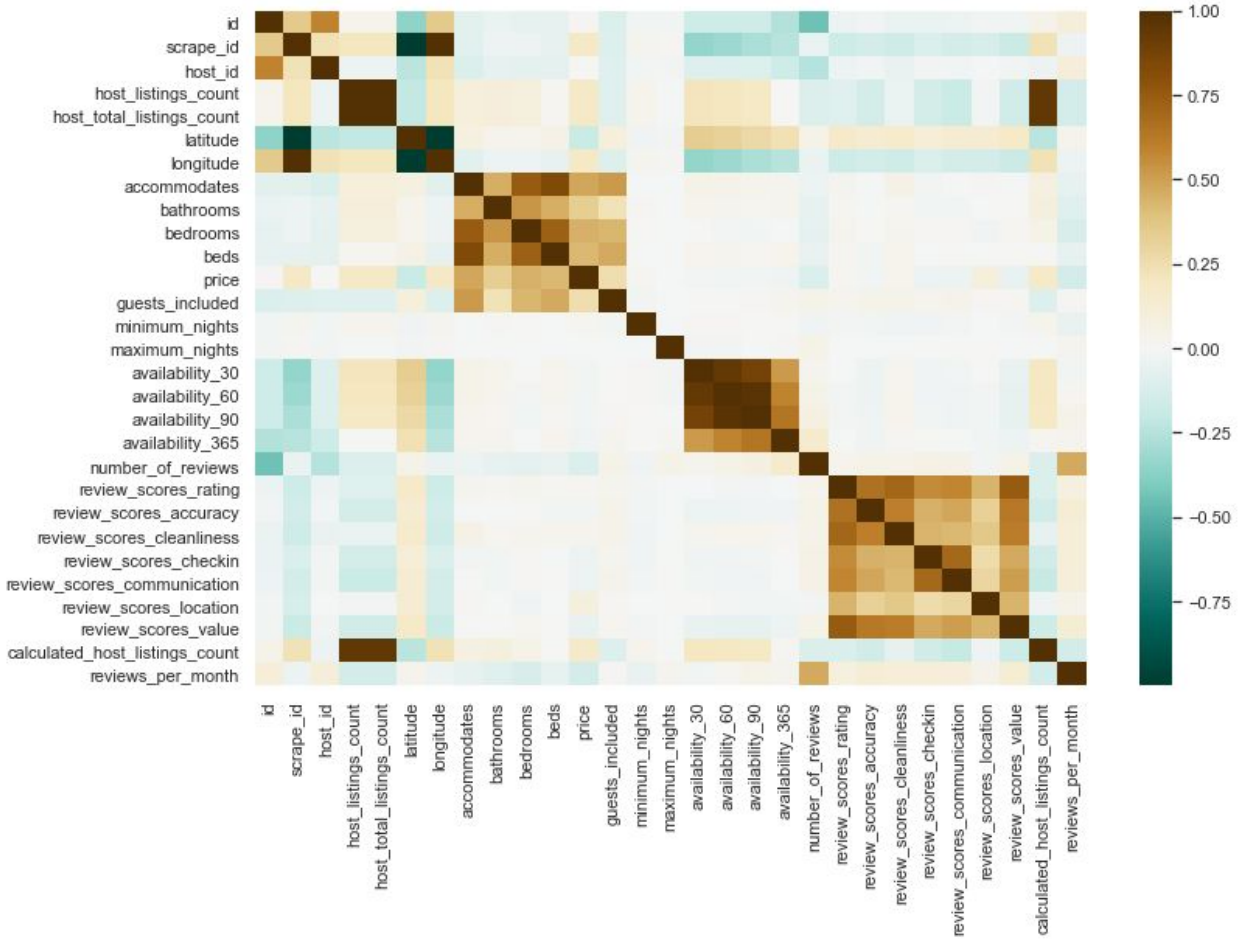
Reviews distribution of Seattle



As we can see that the reviews of Seattle are more than the Boston data. We also can conclude that generally customer turnover frequency is lower than one month based on the distribution around 1 review.

Relationship on price

To figure out what are the properties that has strong relation with the rental price, we build up a heatmap as follows:

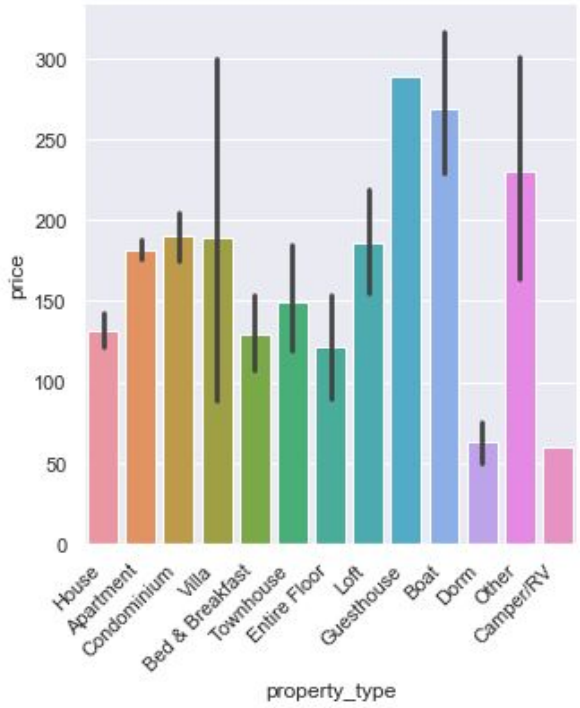


In this clustered heat map, we found that the price argument becomes more ‘brown’ when it meets accommodations, bathrooms, bedrooms, beds, and guests_inculded. It suggests that these arguments have some influence on price. We’ll talk about these arguments in the next few sections.

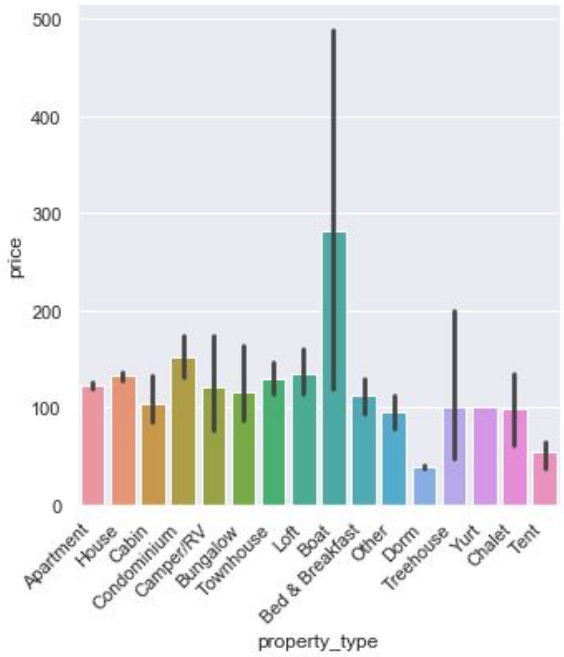
Property, bed and room type relationship on price

We checked the price relationship with the bed and property type of both datasets.

The property type with price of Boston:

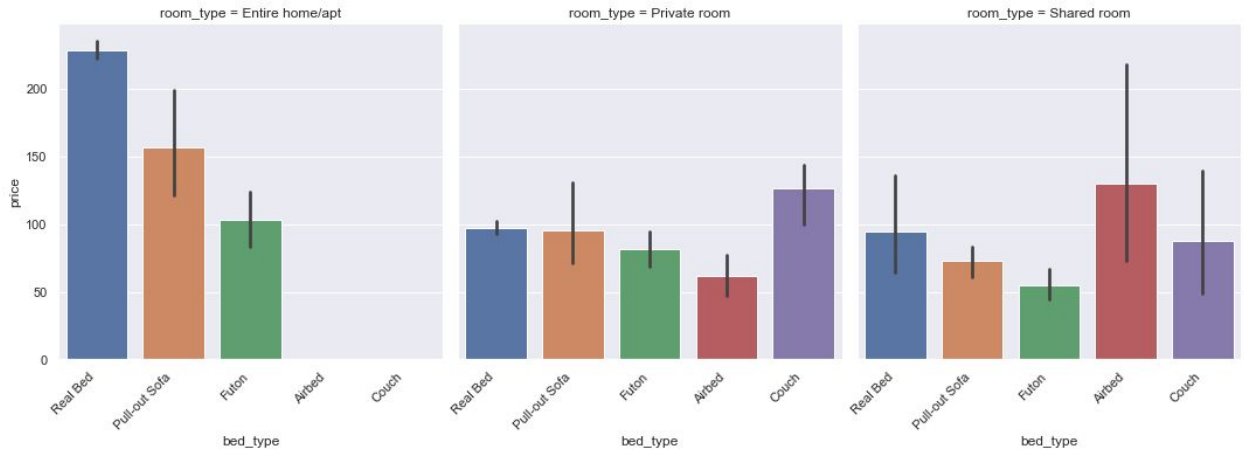


The property type with price of Seattle:

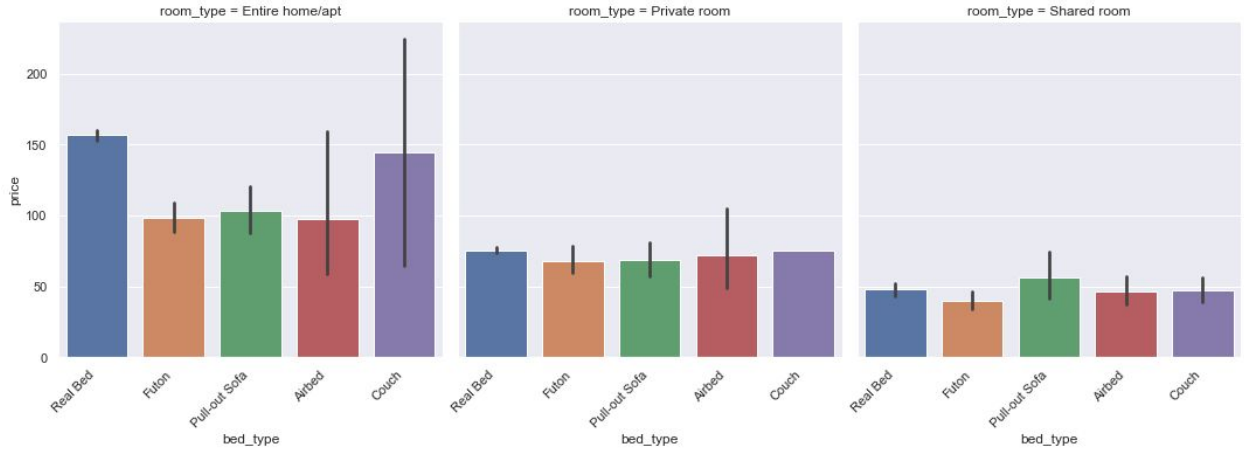


As we can see that property type with price in Boston is highest on GUESTHOUSE, and BOAT while by looking at Seattle, BOAT has the highest prices. Dorm, tent, RV have the lowest price in both cities because of their humble environment. The Prices of villas in Boston and boats in Seattle tend to change on a large scale, which means the quality of these types could cause a great impact on their price. But in both datasets, prices are in the range of 100$ in relation to property type.

Bed type with the price of Boston:



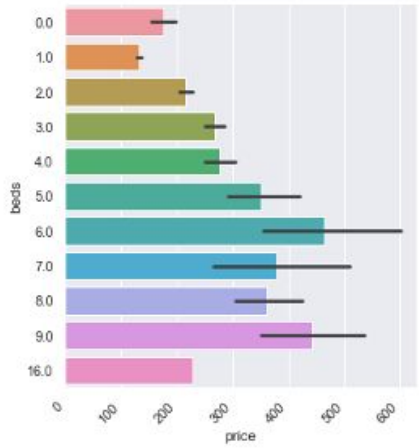
Bed type with the price of Seattle:



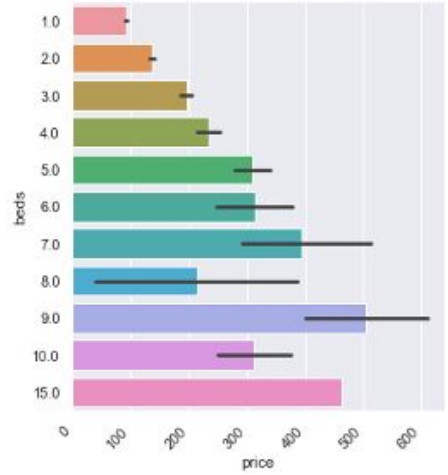
As we can see that the prices are little high with real beds in both datasets if the room type is entire home/apt. The room type also has influence for the rental price. Price of an entire home/apt is much higher than a private room or a shared room especially in Seattle. Couch for the entire home in Seattle looks like an outlier for deciding the price. Same as Airbed for the shard room in Boston.

Number of beds, bathroom and accommodates relationship on price

Number of beds with the price of Boston

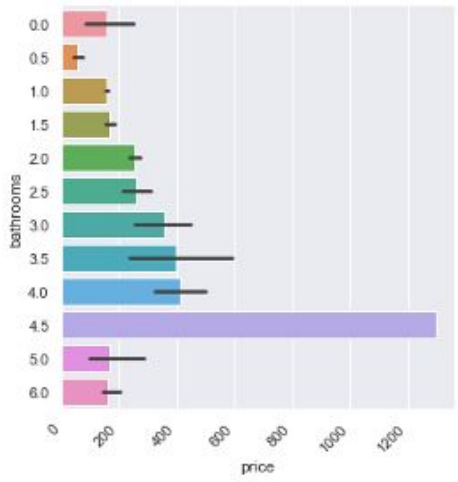


Number of beds with the price of Seattle

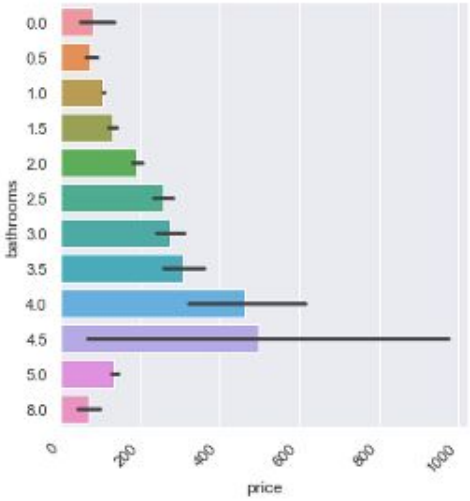


In Boston, from bed 0 to 5, the prices are in the range of 200$ to 300$ but from bed 6 to 9 are in the range of 400$ to 500$. While in Seattle, the prices are a little bit lower than Boston bed prices. In general, the price increases with the number of beds. The growth becomes unstable after the number of 6, which might be because of lacking samples for large numbers.

Number of bathrooms with the price of Boston

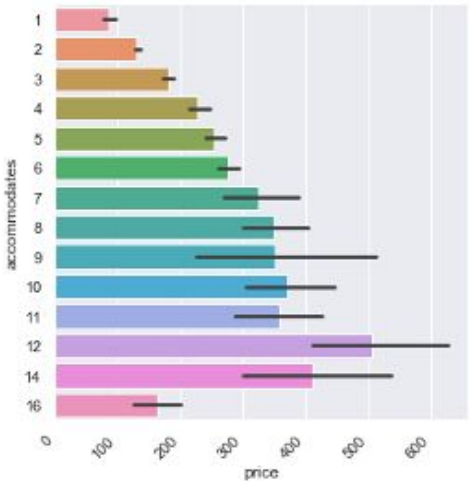


Number of bathrooms with the price of Seattle

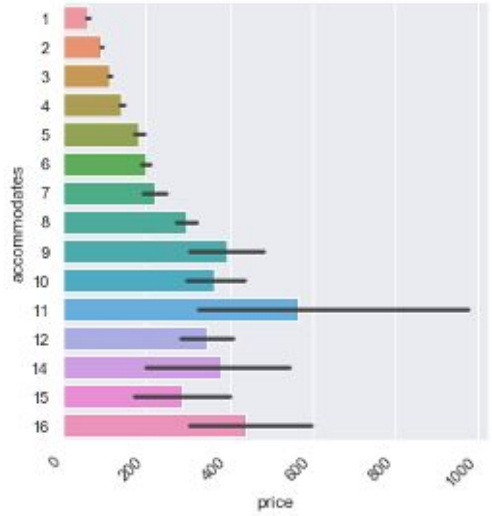


As we can see, the bathroom has a large influence on the price compared to beds. Price fluctuating after 4 probably not because of the number of the bathrooms but the size of the housing.

Accommodates with the price of Boston



Accommodates with the price of Seattle



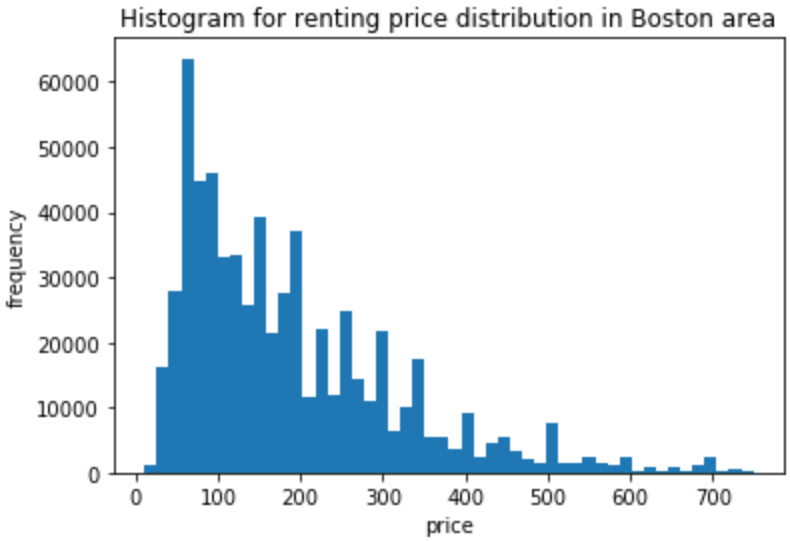
As we can see, the price increases with the number of accommodations. Seattle, however, has a lower price than Boston in each level. 'Accommodates' has the least influence on price among three features since it needs a higher number to match the certain price.

Analysis reasons behind price changing

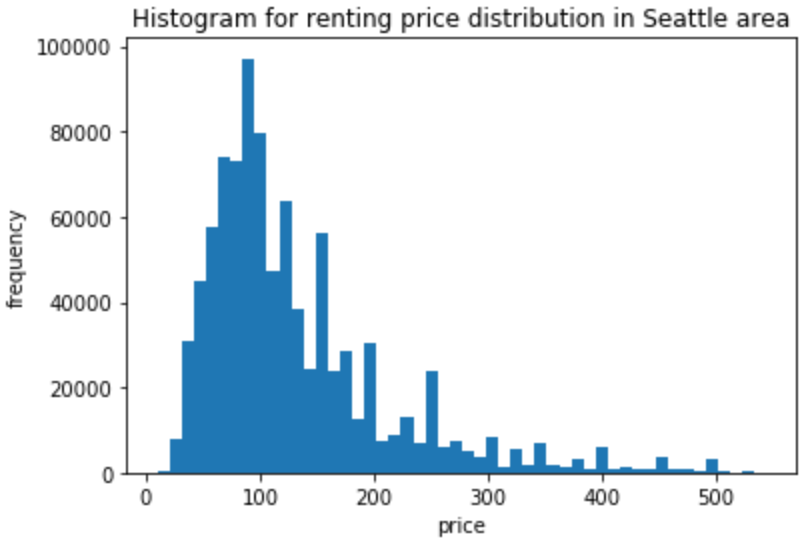
Although we have found arguments that would have an impact on the rental price, these arguments cannot explain the price changing over time. We draw some figures which suggest this changes in order to find out the reason behind it.

Price distribution

Rental price distribution in Boston:



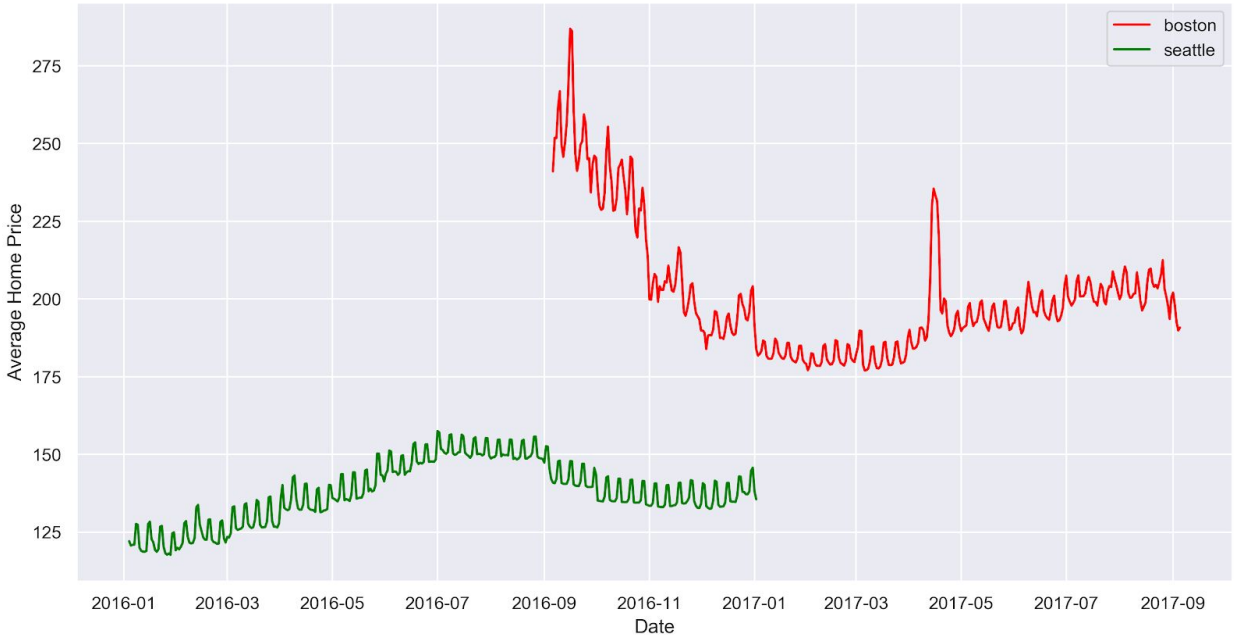
Rental price distribution in Seattle:



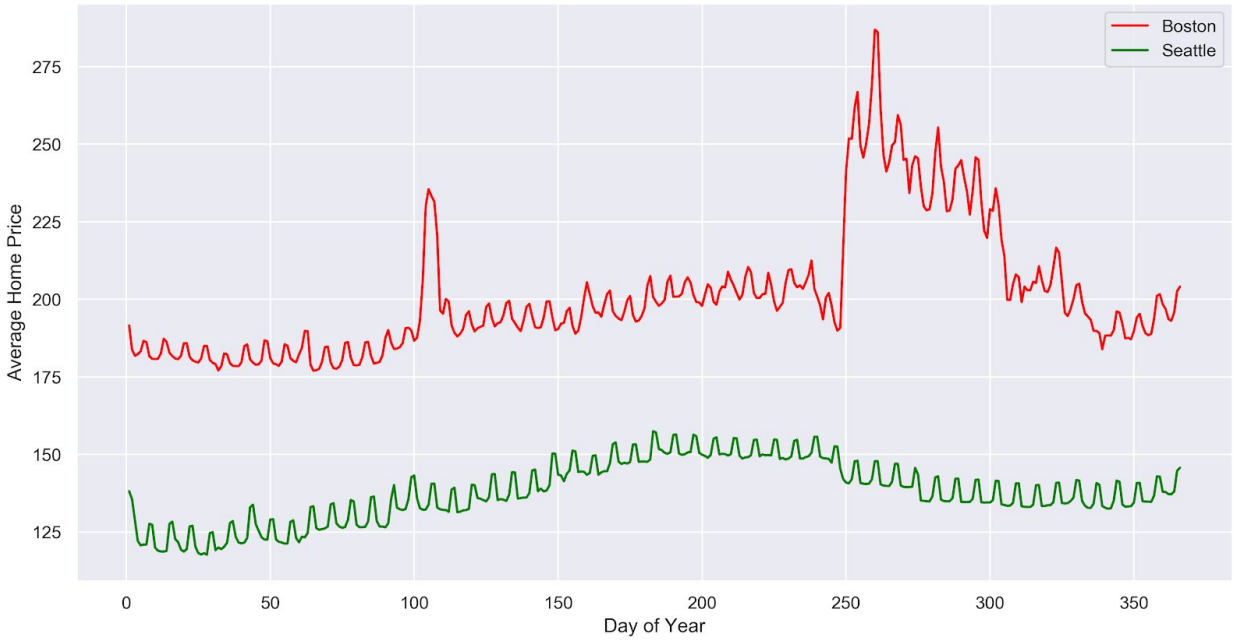
As we can see, the price distribution of two cities mainly falls into the range from 50$ to 200$. Based on what we implemented, the median rental price for Boston is $150 per night; while for Seattle it is $109 per night. We consider it is because Boston is a more popular tourist city.

Analysis reasons for change in prices during the year

Throughout the year, the rental cost in Boston is significantly higher than Seattle. We drawed a price-changing figure during its actual time period as follows:



Another figure suggests the price-changing relatively in a year:



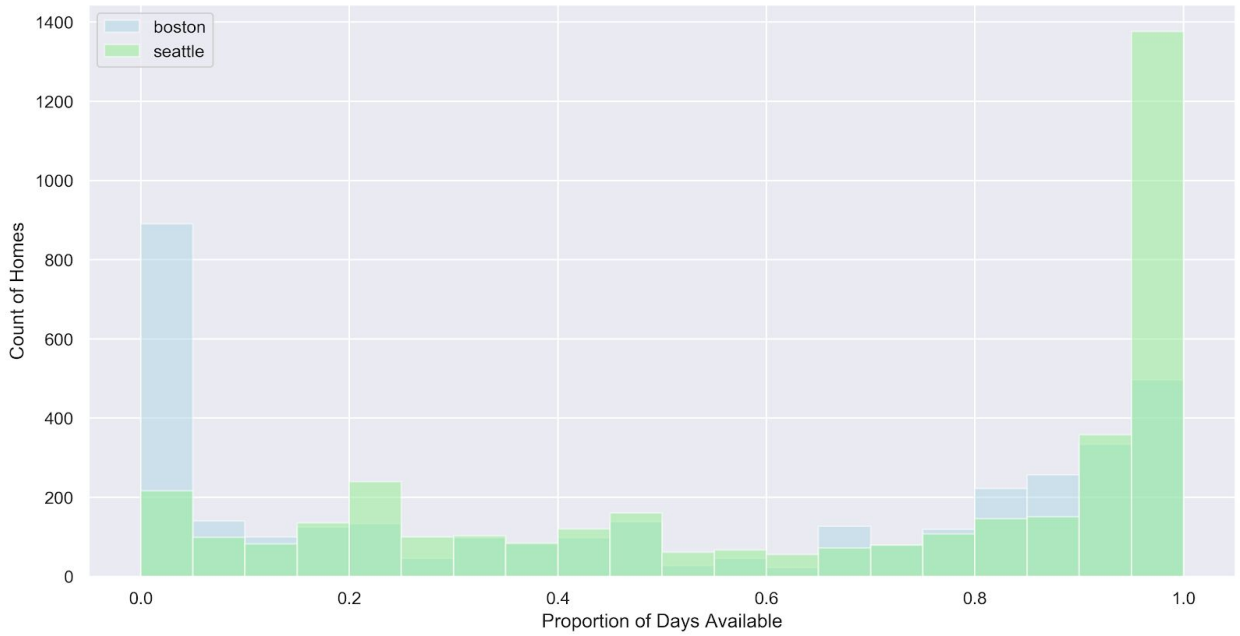
Although there is no significant peak in Seattle data, a dramatic peak occurs in the Boston dataset. What made this happen? We checked the events happening in Boston during that time

and found there was a big marathon in April. During that time, more than thirty-thousand runners attended. Because of the race, local rental prices appear to have risen around $34 per night in that area.

Except for the incident, rental price in Seattle tends to rise during the summer period. This could be related to an explanation that summer is a tourist season for Seattle. While during 2016-2017 the rental price in Boston has an explicit decrease. It might suggest tourism became worse in Boston during that time.

Analysis availability property during the year

We also checked the availability property of both cities. The proportions of days available are as following figure shows:



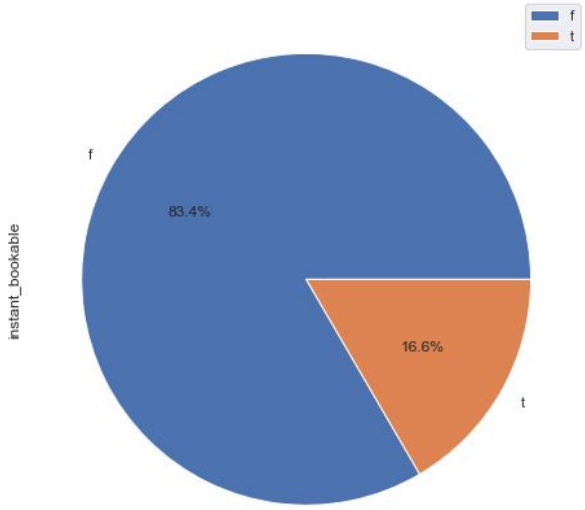
In Boston, there may be a 49% chance that any housing in the list is available at any time, while in Seattle, this number even rises to 67%. This can be explained by the higher supply-demand ratio in Boston than in Seattle. This may mean that Boston is a favorite destination for tourists, which leads to difficulties booking a house here than Seattle.

In terms of housing availability, Seattle and Boston are relatively evenly distributed. However, in Seattle, more than 95% of housing estates are available throughout the year. This may be because the prices of these houses are too high while their owners have no other means to attract tourists.

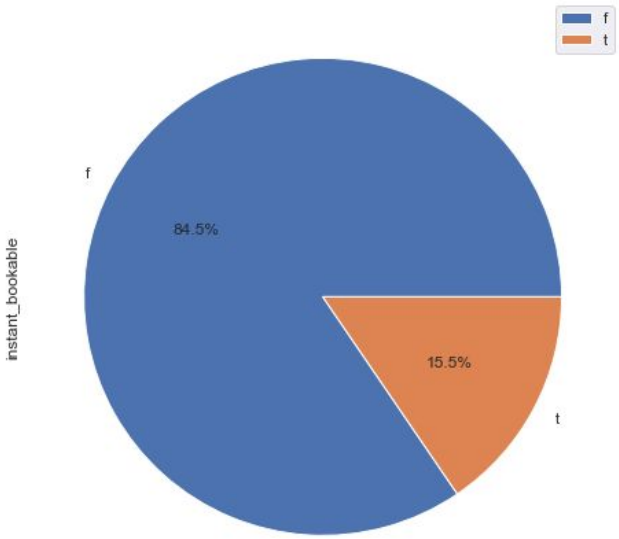
In Boston, 891 houses are very popular and their availability is less than 5% of the entire year. Due to various factors such as location, value and cleanliness, these houses will be the first choice.

Even though there is a high availability of housing, the demand for popular apartments is still at a high level. The picture below reflects the long-term unavailability of these apartments. They suggest that in Boston, popular apartments have only 16.6% chance of successful booking, while in Seattle that chance drops down to 15.5%.

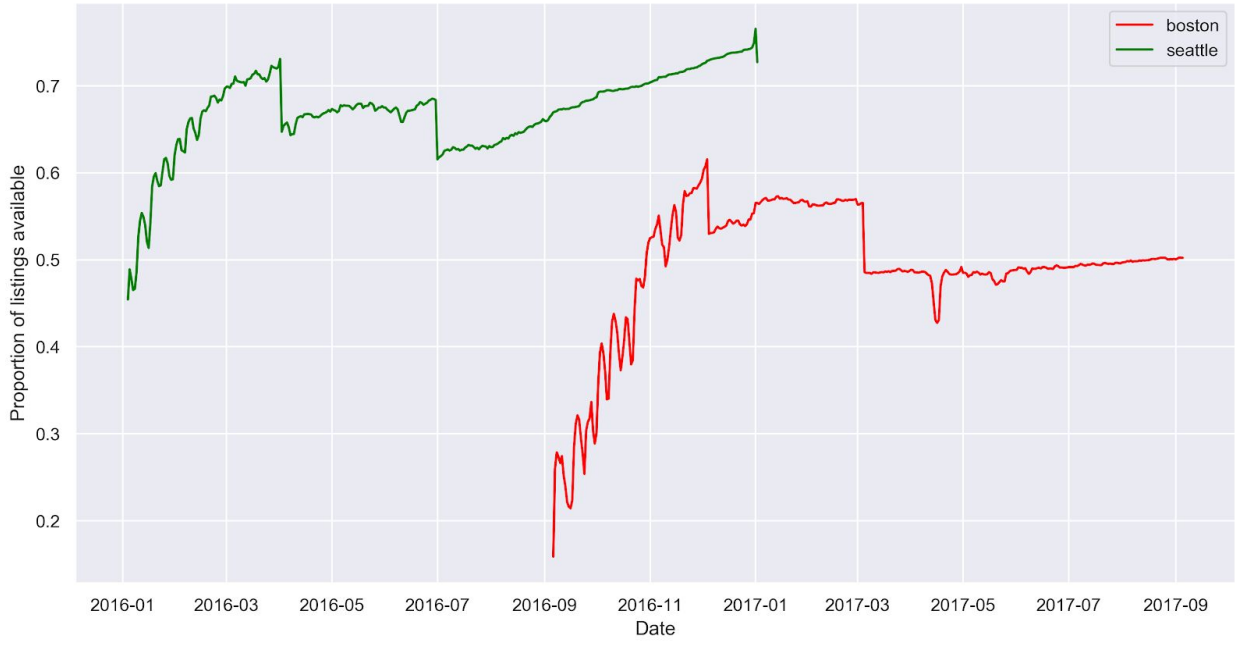
Instant bookable for Boston:



Instant bookable for Seattle:



Probably the most straightforward part of the proportion of listings available graph below is that in the first three months of the data there is a constant upward trend. The reason for this is that the data provided is a snapshot that was taken and that the closer dates to the snapshot date tend to have a higher booking rate.



Conclusion

After analysis on the both datasets Seattle and Boston, both cities have massive traffic. Guests and hosts both used Airbnb for a much better traveling experience. Boston, however, seems a more popular tourist city than Seattle according to the median rental price.

From what we saw on the word clouds considering the descriptions and comments, we found location and living conditions are the crucial factors of attracting guests. For Hosts, they usually attracted customers by describing the location advantage of their houses. This includes if it is near to the main street or area, if it has restaurants and shopping malls in the neighborhood, if it has a great public transportation system around the area. Hosts often tried their best to brag their housings' living room to attract customers. For guests, they always used some cliches such as 'cozy and convenient','great place to stay','good experience' on reviews. They also mentioned the locations are convenient for them to access airports, scenic spots and restaurants.

Based on the heatmap, we found out that the type of room, bed, property, and the number of beds, bathrooms, accommodations, are the main factors that decide the value of the housing rental. From the catplots, we concluded that these factors nearly have the same influence on the price in both Boston and Seattle datasets. For the property type, the price of the boat house and the guest house are relatively higher than others. Meanwhile, the price of RVs, tents, dorms tends to be low. It implies that people prefer comfortable and functional tenements when they consider a vacation there. For the room type, the entire house/apt costs the highest price since its large space. Moreover, if the bed type is the real bed, it costs even higher. However, the real bed doesn't have a significant impact when it comes to other room types. For the number of beds, bathrooms, and accomodations, they have an explicit effect to cause the ascending of price when they are small. The ascending curve will become flat or fluctuate when the number is large. From most important to the least, their order of impact on price should be: bathrooms, beds, accomodations.

As a matter of fact, what we discuss above is not truly the 'price', but the 'value' of the housing. The price of tenements changes daily. We want to figure out what's the rule behind it. First, we manage to build a price distribution chart. As we can see, the price per night in two cities mainly fell into the range from 50$ to 200$. Then, we build two graphs which stand for the price-changing in real date and in day of year. We found a dramatic peak for the Boston curve in April,2017. During that time, Boston held a large marathon which attracted thousands of contestants. It can explain why the peak appears. Also, curves suggest that summer could be a tourist season for Seattle, and tourism was getting worse for Boston in 2016-1017.

While analyzing the availability of these housing, we found out Seattle has a higher chance than Boston that any housing in the list is available at any time. It shows Boston is a favorite place for tourists which leads to the difficulties of booking housings. For Seattle, there were more than 95% of housing available throughout the year. These houses may have an unreasonably higher pricing and no attractive aspect to support that price. For Boston, there are 891 houses which are very popular and are available less than %5 of the time. These houses would be the ones that are highly preferred due to multiple factors such as location, value and cleanliness.

Our project has great potential to be extended and applied in rental price prediction, which is the main field of related work. To improve the services of Airbnb, it's also a potential idea to analyze how locality affects the price. We could measure the distance from house to the city center, the airports, the main shopping area and so on, and analyze them to see if they related to the price.

References

[1] Airbnb, wikipedia, https://en.wikipedia.org/wiki/Airbnb/.

[2] "Inside Airbnb, adding data to the debate", http://insideairbnb.com/get-the-data.html/.

[3] Laura Lewis, "Predicting Airbnb prices with machine learning and deep learning", 2019, https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-deep-learning-f46d44afb8a6/.

[4] Graciela Carrillo, "Predicting Airbnb prices with machine learning and location data", 2019, https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-location-data-5c1e033d0a5a/.

[5] AirBNB, "A sneak peek into the Airbnb activity in Boston, MA, USA," Kaggle, 2008.

[6] AirBNB, "A sneak peek into the Airbnb activity in Seattle, WA, USA," Kaggle, 2008.

[7] N. developers, "NumPy," https://numpy.org/.

[8] A. Mueller, "Wordcloud," https://amueller.github.io/word_cloud/.

[9] J. Hunter, "Matplotlib," https://matplotlib.org/.

ACKNOWLEDGMENTS

Thanks Professor Luo for his expert advice and help during the whole process of this project. Also, we really appreciate classmates who discuss with us and give us advice and ideas during the whole course. Specifically, we should appreciate kaggle, a website which gives us these wonderful datasets. From there, we came out with our original idea, and implemented eventually.

We found lots of excitement and interest in this collaborative work, which strengthens our collaborative skills as well as ability to treat data in another totally different perspective from before. With all the treasure gained from this final project, we will keep moving on the road to discover deep facts of the world.