MSSP 607
Final Project
Project Name: Boston & Seattle Airbnb Data Exploration & Comparative Research
Team Members: Xuechuan Song
Ziyu Zhao

Part 1: Introduction

I. Problem Statement

What are some of the distinct features of Boston and Seattle Airbnb markets? What are some similarities and differences of the two Airbnb markets from the perspective of property distribution, targeted customer types and price?

II. Sub questions

A. City property statistics
   1. The Airbnb in which city is more concentrated and popular?
   2. Is the room type distribution different in these two cities?
   3. What are Airbnb price distribution in two cities and characteristics among different neighborhoods?
   4. If Airbnb rooms are clustered with a small group of hosts?

B. Preference analysis
   1. What do the customers' reviews tell us about their preferences?
   2. What does the availability tell us about the Airbnb markets in Boston and Seattle, and about the customers' preferences of stays in the two cities?
   3. Given above, what conclusions can we draw about customers' preferences and overall conditions in Boston and Seattle Airbnb market?

C. Price Analysis
   1. What factors influence the price most? What is the range of the price and the average price for properties of about same features?
   2. Whether a property is overpriced in similar listings?

III. Audience

The audience of this project report is Airbnb the company and other companies in competition; the goal is to provide some general understanding of Airbnb business and some detailed insights of their markets in Boston and Seattle in order to facilitate their future decision-making.

IV. Structure

Three major parts will be included in our analyses: city property statistics, tenant preference analysis and price influencing factors analysis.

In the city property statistics part, Questions in this section will be answered with statistics, short descriptions, tables and charts to describe the general distribution of properties in the two cities.

In the customer preference analysis part, answers will be answered with analysis of customers' preferences with verbal descriptions, tables and charts to see the similarities and differences between Boston and Seattle Airbnb homestay markets.

In the price influencing factors part, we will analyze price influencing factors with statistical calculations and a decision tree to present our ideas of what price is appropriate for certain required conditions.

V. Motivation

Generally, the questions we mentioned above are important as the answers to which will give one insights into short-term Airbnb homestay rental and service industry in Boston and Seattle. And for us, we might know some "secrets" about Airbnb and learn how to read behind the prices and save money for our future travels by picking stays with best cost performance.

Airbnb business is a suitable object to be analyzed with data quantitatively. Firstly, Airbnb is an online marketplace that generates authentic firsthand rea-time data that might reveal patterns of people's and cities' economic behaviors. And, secondly, when customers' booking stays on Airbnb, it is exactly data that they see and they use to make purchases. It is only reasonable to take advantage of the data we have to guide future choices and activities interacting with data.

During our preparation, we referred to following studies and analyses:

- Airbnb Properties Analysis in Beijing (in Chinese)
- Airbnb User Profile (in Chinese)
- Airbnb Data Analysis Question List (in Chinese)
- Seattle Airbnb Case Analysis
- Random Browsing Airbnb Data Exploration on Kaggle: Boston Seattle NYC

Part 2: Dataset

This project is designed to explore two available datasets describing Airbnb listing activity of homestays in Boston and Seattle. Both datasets have three files storing data in the same format.

1. Each file and its contents are as following:

a) Listings: full descriptions and average review score

- Variables:

id	host_verifications	guests_included
listing_url	host_has_profile_pic	extra_people
scrape_id	host_identity_verified	minimum_nights
last_scraped	street	maximum_nights
name	neighbourhood	calendar_updated
summary	neighbourhood_cleansed	has_availability
space	neighbourhood_group_cleansed	availability_30
description	city	availability_60
experiences_offered	state	availability_90
neighborhood_overview	zipcode	availability_365
notes	market	calendar_last_scraped
transit	smart_location	number_of_reviews
thumbnail_url	country_code	first_review

medium_url	country	last_review
picture_url	latitude	review_scores_rating
xl_picture_url	longitude	review_scores_accuracy
host_id	is_location_exact	review_scores_cleanliness
host_url	property_type	review_scores_checkin
host_name	room_type	review_scores_communication
host_since	accommodates	review_scores_location
host_location	bathrooms	review_scores_value
host_about	bedrooms	requires_license
host_response_time	beds	license
host_response_rate	bed_type	jurisdiction_names
host_acceptance_rate	amenities	instant_bookable
host_is_superhost	square_feet	cancellation_policy
host_thumbnail_url	price	require_guest_profile_picture
host_picture_url	weekly_price	require_guest_phone_verification
host_neighbourhood	monthly_price	calculated_host_listings_count
host_listings_count	security_deposit	reviews_per_month
host_total_listings_count	cleaning_fee	

Size:

Seattle	Boston
3818 obs. of 92 variables	3585 obs. of 92 variables

Reviews: unique id for each reviewer and detailed comments

- Variables: listing_id, id, date, reviewer_id, reviewer_name, comments
- Size:

Seattle	Boston
84849 obs. of 6 variables	68275 obs. of 6 variables

b) Calendar: listing id and the price and availability for that day

- Variables: listing_id, date, available, price
- Size:

Seattle	Boston
1393570 obs. of 4 variables	1308890 obs. of 4 variables

2. Preprocessing

Before data analysis and exploration, we did data cleaning and preprocessing, including: 1) Removing null values; 2) Removing duplicate values; 3) Data Normalization. For example, here exists “$” in price column from the listings, and we only want to know the price value, so we need to remove the dollar sign from every cell. 4) Data extraction. For example, date value, we hope to know the value of year, month and day, but the format is yyyy/mm/dd, so we do data extraction.

3. Copyright and privacy concern

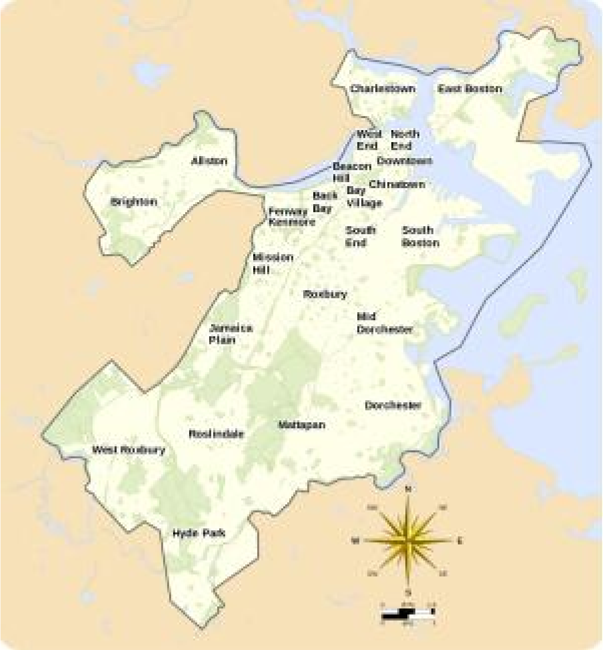
These datasets are from Kaggle, public domain and no copyright, and is one part of “Airbnb Inside”. The data behind the Inside Airbnb site is sourced from publicly available information from the Airbnb site. The data has been analyzed, cleansed and aggregated where appropriate to facilitate public discussion. Here is a privacy concern example, the open dataset provides id, name, basic info, geo info of every host, making it possible to analyze one certain host together with her/his economic status and active area.

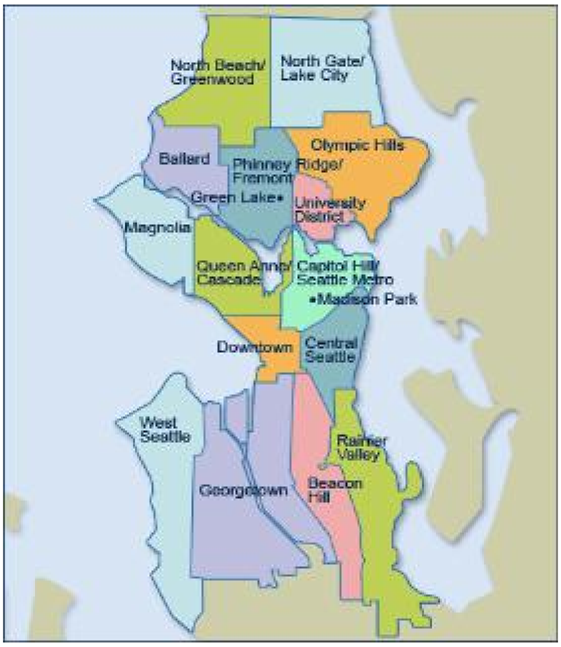
Part 3: Data Analysis

3.1 City property statistics

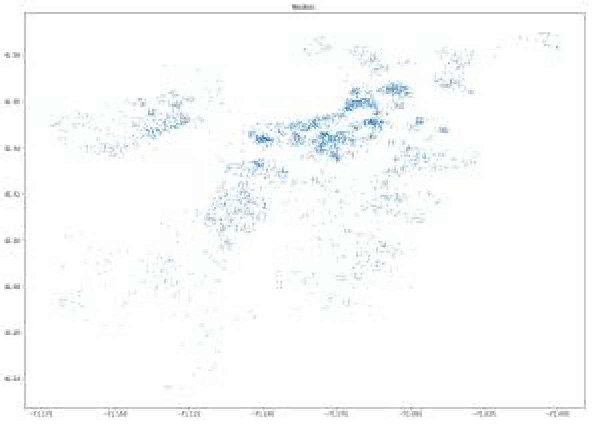
I. Geographical distribution——The Airbnb in which city is more concentrated and popular?

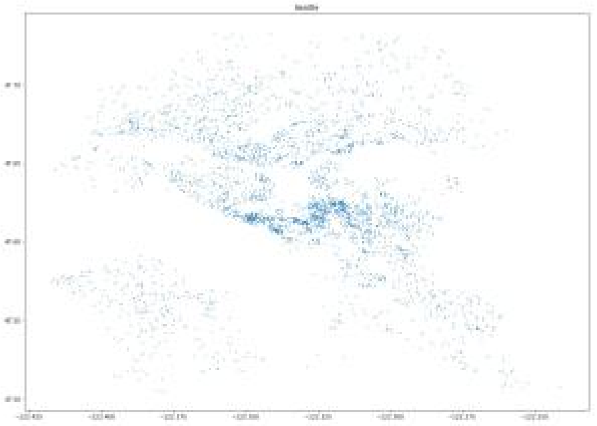
In this part, we will analyze the geographical distribution of Airbnb in both Boston and Seattle. Boston's diverse neighborhoods serve as a political and cultural organizing mechanism. The City of Boston's Office of Neighborhood Services has designated 23 Neighborhoods in the city, which are made up of approximately 84 sub-districts, squares, and neighborhoods within each official neighborhood. Seattle, Washington contains many districts and neighborhoods, even called “a city of neighborhoods”. In the Airbnb of Seattle dataset, here are totally 87 neighborhoods. Here are neighborhood maps.



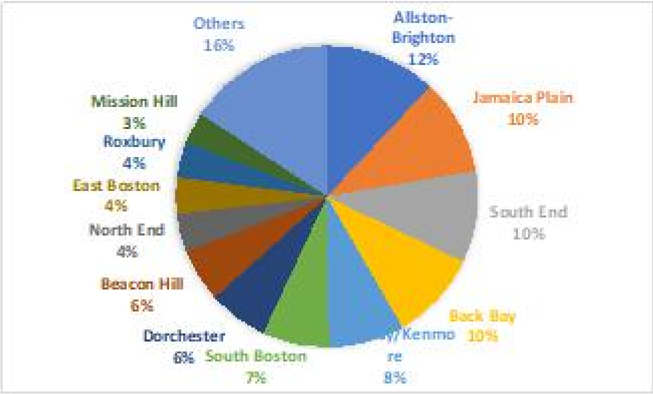


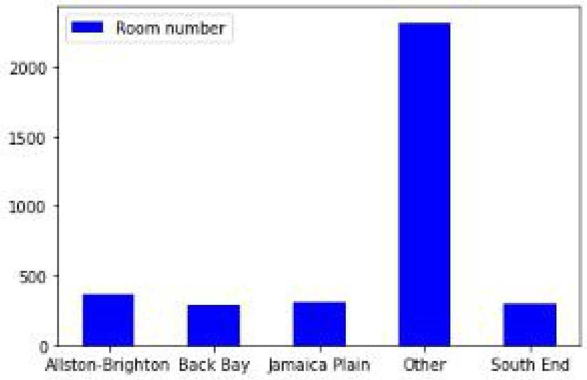
And geographical scatter pattern of Airbnb location in these two cities is as the follow, where we can see the obviously centralized distribution in certain neighborhoods.



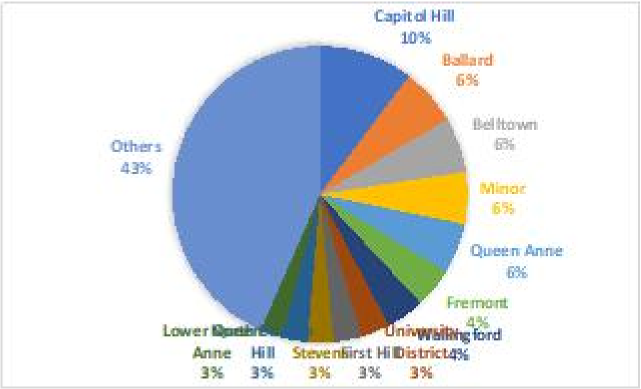


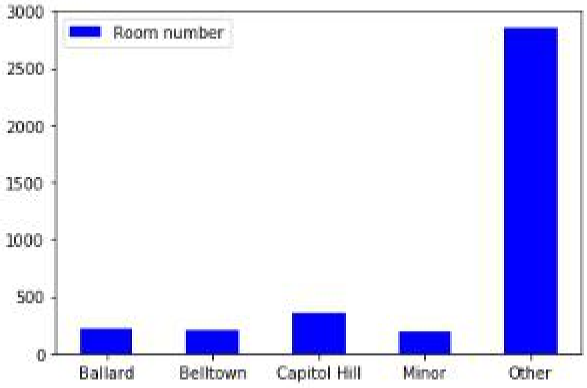
Using Pie and bar chart, we can have the quantitative comparison on Airbnb amount in each neighborhoods of these two cities. In Boston, the top four neighborhoods are Allston-Brighton, Jamaica Plain, South End, and Back Bay, occupying 41.36% Airbnb in Boston, and the Allston-Brighton neighborhood have the most, which is 364 Airbnb houses.





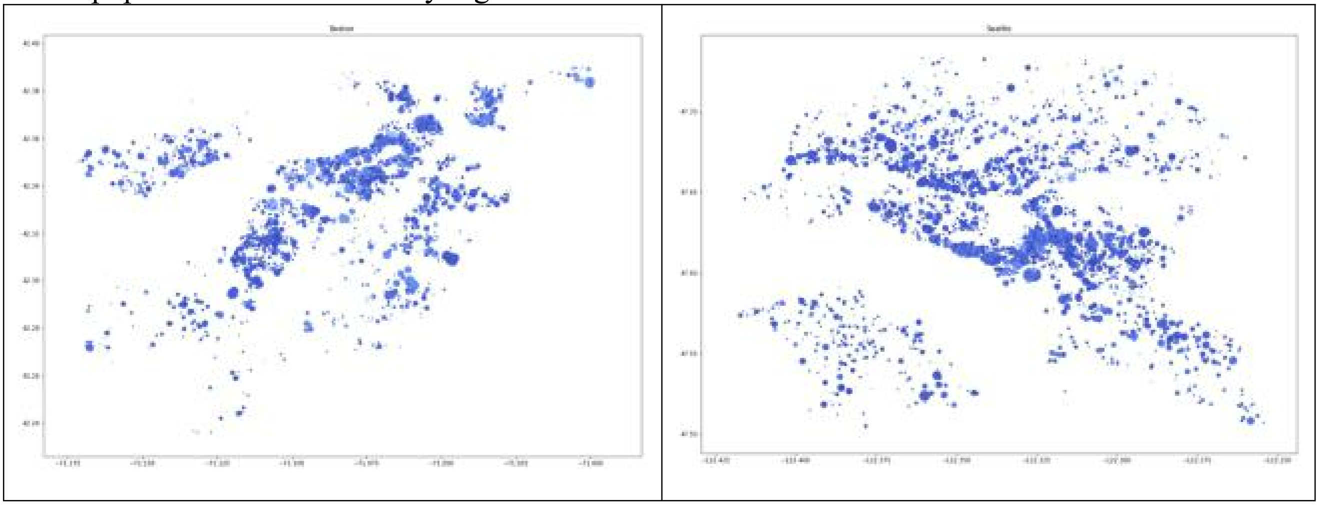
In Seattle, the top four neighborhoods are Capitol Hill, Ballard, Belltown, and Minor, occupying 28.22% Airbnb in Boston, and the Capitol Hill neighborhood have the most, which is 351 Airbnb houses.





Relatively, Airbnb in Boston is more concentrated, though not considering the fact that Seattle has more neighborhoods.

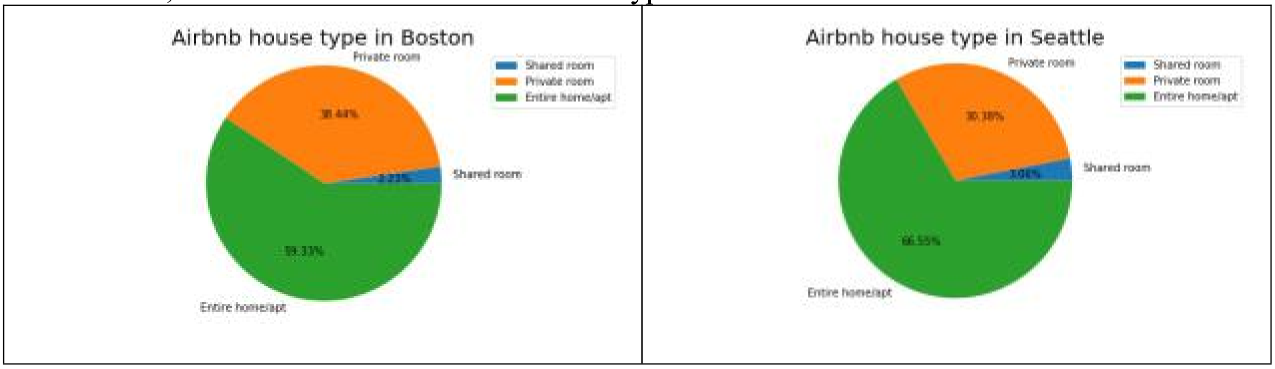
Given information about number of reviews and review scores, we can find the popular Airbnb location. The bigger and darker, the more popular the neighborhood is with high scores. Using the same criteria, it seems that Airbnb houses in Seattle are more popular and have relatively higher review scores than in Boston.



In a conclusion, the Airbnb in Boston is more concentrated, but in Seattle, Airbnb has higher star rate and more popular.

II. Room Type distribution——Are the room type distributions different in these two cities?

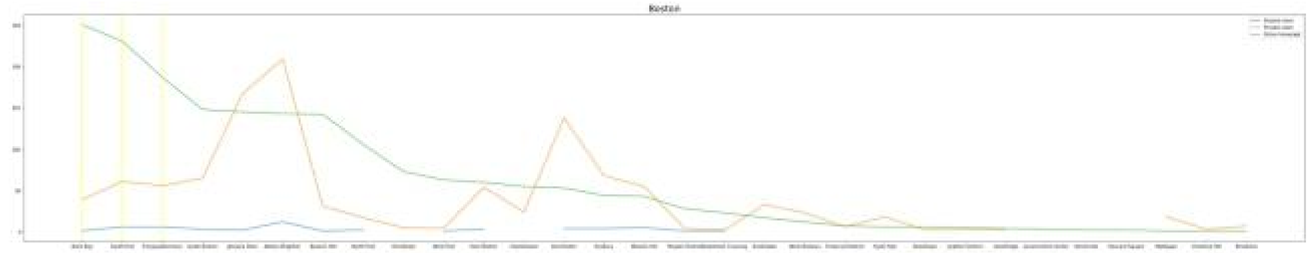
At first, we want to learn the overall room type distribution.



In both cities, entire home/apt ranks top with more than 50%, while private room ranks the second with more than 30%. Boston have relatively more entire home or apartment at proportion.

For Boston, sort descending Airbnb amount by entire home/apartment, the top 5 neighborhoods are Back Bay(251), South End(231), Fenway/Kenmore(187), South Boston(148), Jamaica Plain(145).

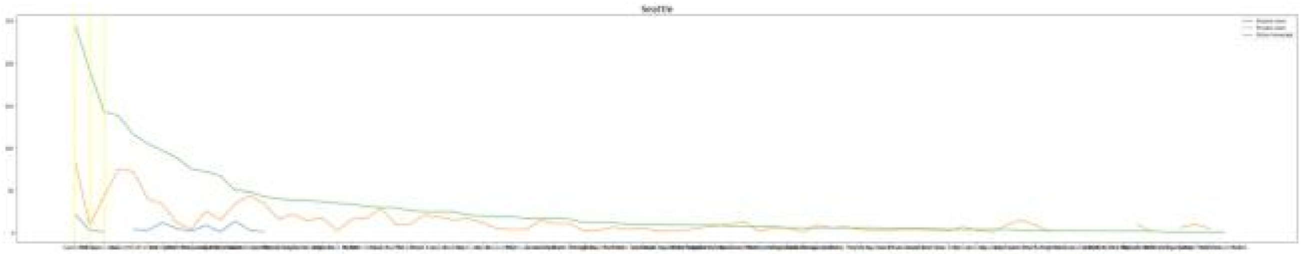
neighborhood	Entire home/apt	Private room	Shared room
Back Bay	251.0	39.0	1.0
South End	231.0	61.0	6.0
Fenway/Kenmore	187.0	56.0	6.0
South Boston	148.0	65.0	3.0
Jamaica Plain	145.0	167.0	2.0



We can see, except Jamaica Plain, Allston-Brighton, Dorchester, Roxbury .etc., most neighborhoods have more entire home and apartment than other two types.

For Seattle, sort descending Airbnb amount by entire home/apartment, the top 5 neighborhoods are Capitol Hill(245), Belltown(192), Queen Anne(143), Ballard(138), Minor(116).

neighborhood	Entire home/apt	Private room	Shared room
Capitol Hill	245.0	84.0	22.0
Belltown	192.0	9.0	3.0
Queen Anne	143.0	43.0	1.0
Ballard	138.0	75.0	NaN
Minor	116.0	72.0	4.0

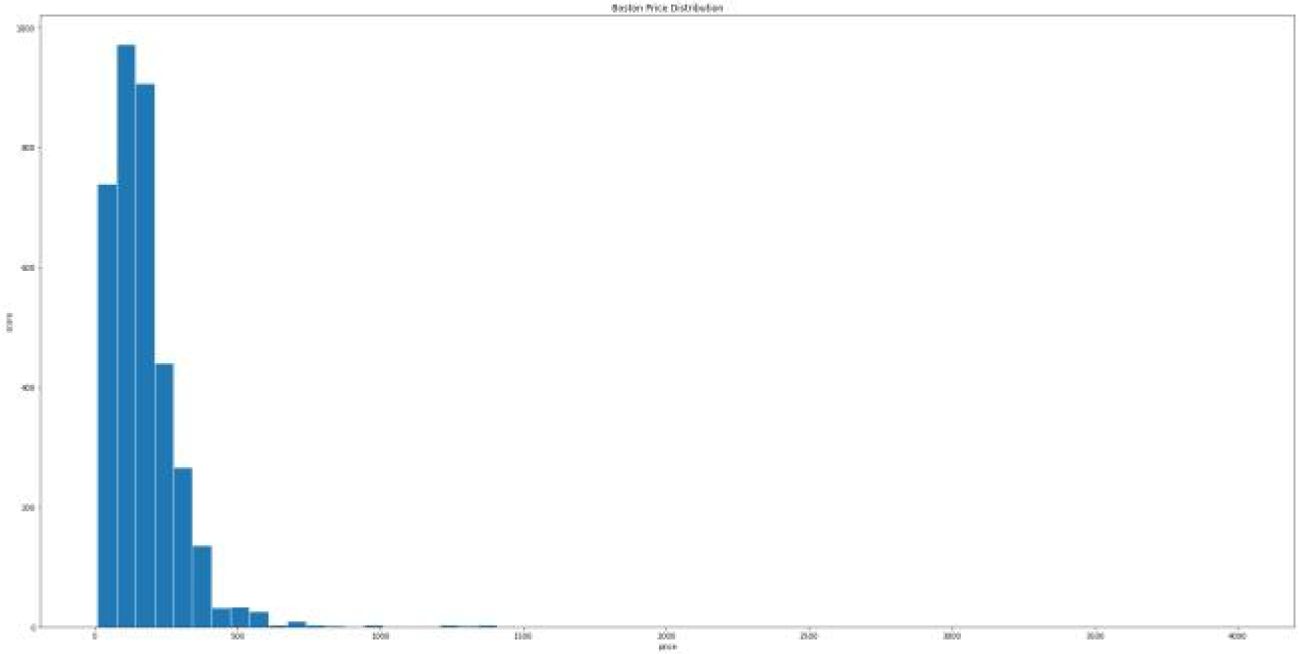


We can see the room type of entire home or apartment takes the lead, comparing with other two room types.

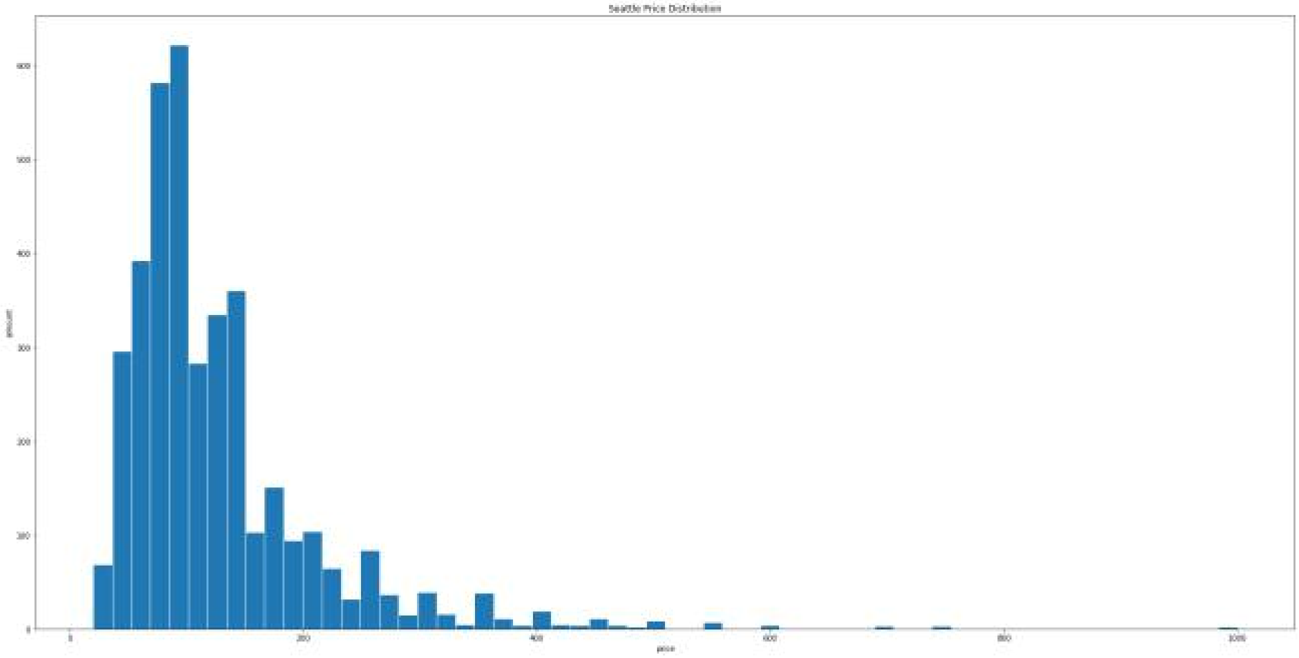
In a conclusion, the overall room type distribution is similar in Boston and Seattle, that most are entire home or apartment. In Seattle, the entire type ranks top in almost all the neighborhoods, while in Boston there are some neighborhoods having more private room type.

III. Price distribution——What are airbnb price distribution in two cities and characteristics among different neighborhoods.

At first, we want to learn the overall Airbnb price distribution.



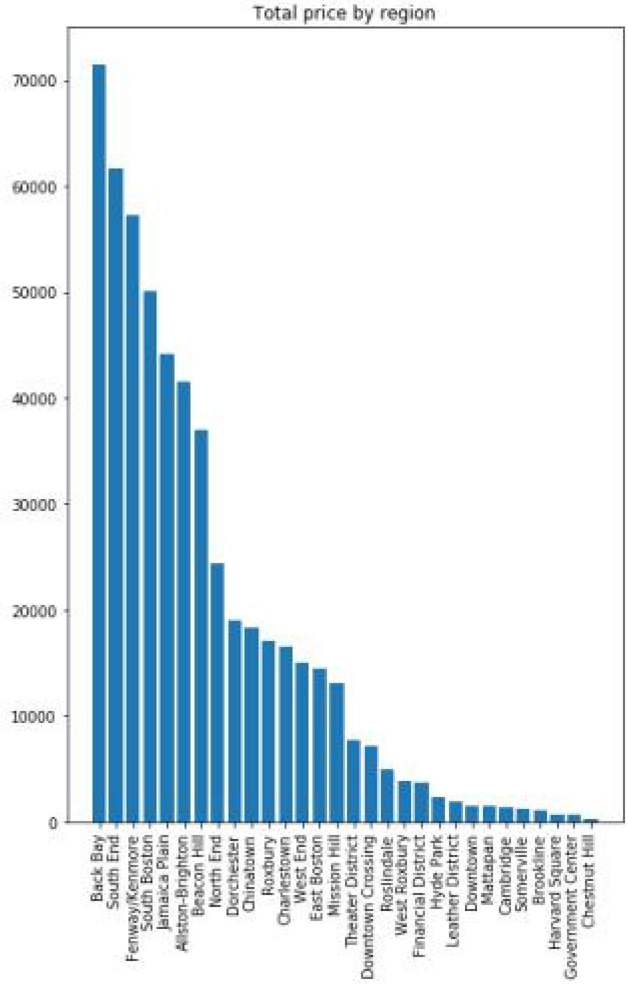
In Boston, the price represents the right skewed distribution. The price of most Airbnb rooms ranges from \$120 to \$200.

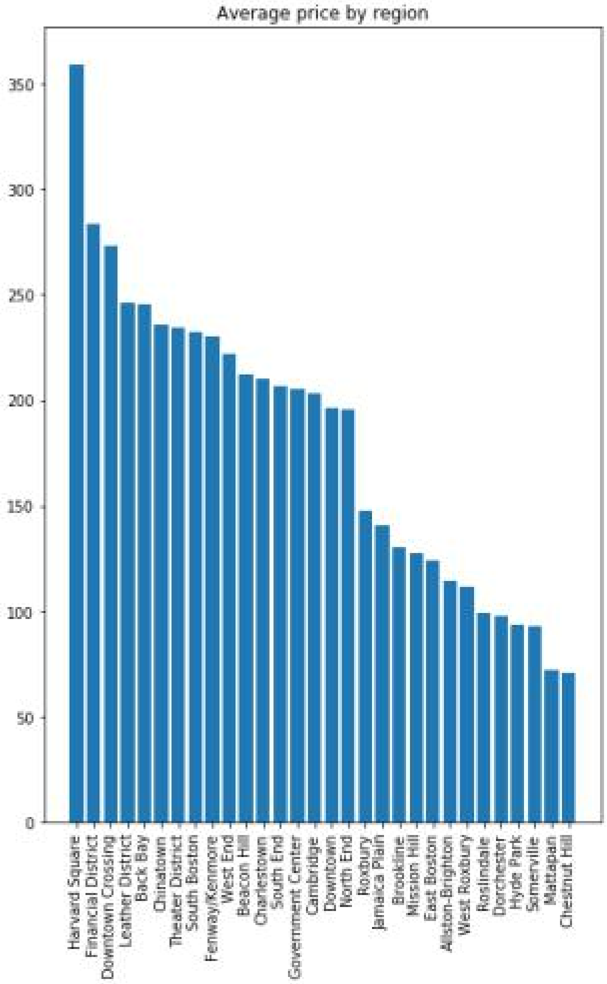


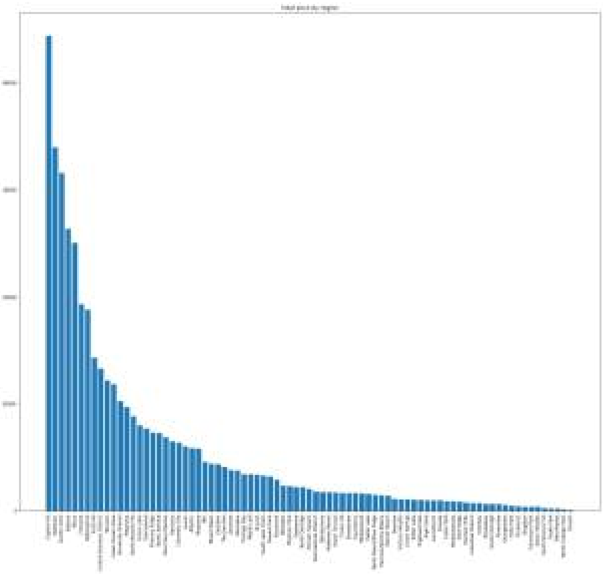
In Seattle, the price also represents the right skewed distribution. The price of most Airbnb rooms ranges from \$70 to \$150. Relatively lower than the Airbnb price in Boston.

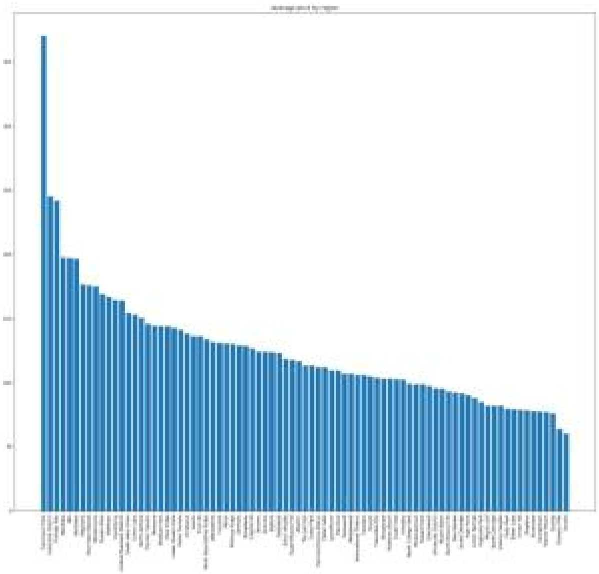
For each neighborhood, we want to see the total price and the average price, hoping to find the relationship between quantitative results and their geolocation.

In Boston, the top five neighborhoods with highest total price are Back Bay(71482), South End(61603), Fenway/Kenmore(57296), South Boston(50138), Jamaica Plain(44176). The top five with the highest average price are Harvard Square(359), Financial District(283.69), Downtown Crossing(273.5), Leather District(245.87), Back Bay(245.46).







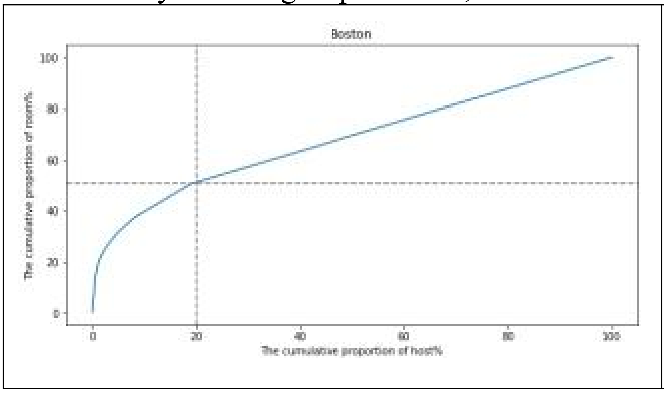


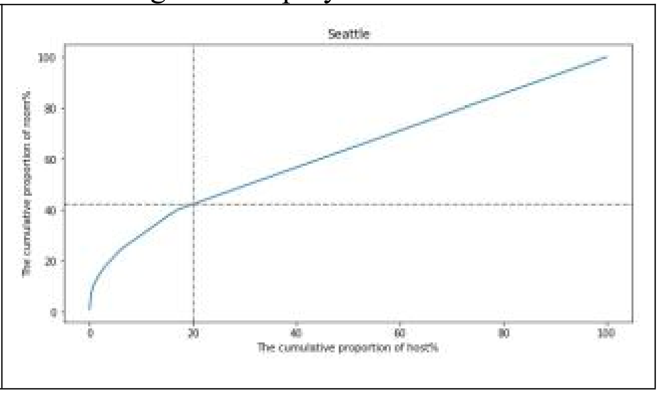
In Seattle, the top five neighborhoods with highest total price are Capitol Hill(44378), Belltown (33992), Queen Anne(31559), Ballard(26330), Minor(25017). The top five with the highest average price are Fairmount Park(370), Industrial District(245), Portage Bay(241.43), Westlake (197), Alki (196.65).

In conclusion, Airbnb price is relatively lower in Seattle than in Boston. Back Bay in Boston is the only neighborhood listed in both top five highest total price and the average price.

IV. Property distribution——If Airbnb rooms are clustered with a small group of hosts?

In this part, we are interested in if Airbnb houses in Boston and Seattle are in fact clustered by a small group of hosts, which is the same meaning of monopoly.





In Boston, we can see about 50% Airbnb rooms are accumulated by 20% hosts, while in Seattle, only about 40% Airbnb rooms are accumulated by 20% hosts, which shows the relatively high monopoly phenomenon in Boston Airbnb market.

In a conclusion, it is true that Airbnb rooms are clustered with a small group of hosts in both cities.

3.2 Preference Analysis

I. What do the customers' reviews tell us about their preferences?

A. Issues of concerns

By conducting word count analysis for comments is both Boston and Seattle review datasets, we identified following issues of great concerns that were frequently mentioned in reviewers comments (each category in order):

Issues of Concerns	
Value	stay, place, apartment, room, bathroom, kitchen, space, parking, breakfast, amenities
Location	location, neighborhood, restaurants, distance, station, subway, transportation, shops, towels, shower, beds, grocery
Host	host
Communication	communication, pictures

For Airbnb hosts reference, special attention should be paid to above customer concerns as they are potential price, business and rating affecting factors.

B. Sentiment analysis

Also through word count analysis, we identified following frequently used sentiment verbal indicators that reflect the customers' degree of satisfaction toward their stays:

Sentiment Indicators	
Positive adjectives	great, nice, clean, good, comfortable, easy, perfect, wonderful, helpful, beautiful, super, lovely, quite, friendly, convenient, excellent, welcoming, amazing, accommodating, responsive, safe, cozy

Positive actions	recommend, home, will, enjoyed, loved, thank, thanks, appreciated
Negative adjectives	bad, noisy, poor

(Note: To avoid mistakenly overlooking negative action indicators composed by positive actions and “not” or “no” prefixes, we checked the content of phrases starting with “not” and “no” and found that the number of cases as mentioned was very small. Thus, those special cases were ignored in our calculation.)

After identified sentiment indicators, we came up with a simple method to score comments by scoring each occurrence of positive adjectives 1 point, positive actions 2 points and deduct 5 points for each occurrence of negative adjectives. And we considered the properties whose average sentiment scores were greater than 10 as highly commended properties and whose average sentiment scores were less than 0 as poorly commended properties. The results were as the following:

Examples of Highly/Poorly Commended Properties	
Type	**Listing ID**
Boston Highly (15)	6796364, 7092874, 12603280, 6914622, 12125635, 14220964, 5584915, 5212321, 14125957, 11022736, 8201104, 14603013, 14813006, 7367795, 3593290
Boston Poorly (7)	6277566, 13751871, 6887926, 1867754, 10036037, 14532696, 4409653
Seattle Highly (55)	6915487, 8988178, 3593582, 3817141, 6022715, 1027860, 5968862, 7676574, 9522082, 8717068, 7278583, 9367465, 7459637, 8765219, 6348159, 3888986, 5648564, 8483744, 9694921, 6808970, 3728802, 670262, 9957555, 8088447, 609421, 7574864, 4825073, 7179783, 7619727, 3403858, 3490239, 4557204, 4238106, 6558980, 639130, 7420339, 5930473, 9426058, 9352778, 7203765, 8103365, 9216874, 5128160, 7828509, 6317449, 7513198, 7151107, 2158992, 4678866, 8047522, 6714817, 7151924, 1150519, 8147215, 319768
Seattle poorly (3)	1589461, 8474294, 8505421

(Note: the detailed information was stored in dataframes respectively: boston_high_review_score_df, boston_low_review_score_df, seattle_high_review_score_df and seattle_low_review_score_df.)

As the sizes of both review datasets are:

Seattle	Boston
84849 obs. of 6 variables	68275 obs. of 6 variables

From the result, it is safe to conclude that Seattle Airbnb market has more passionate and tolerant customers than Boston and that Seattle Airbnb market generally leaves its customers better impression than Boston market.

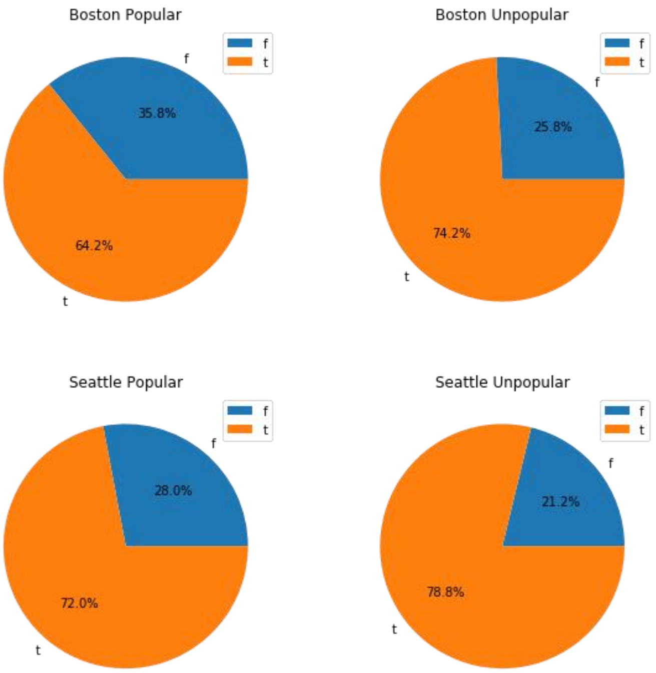
II. What does the availability tell us about the Airbnb markets in Boston and Seattle, and about the customers’ preferences of stays in the two cities?

In this part, we consider the number of days available for the next year’s booking as an indicator of a property’s popularity. The logic is that the less available for future booking a property, the more popular it is for advance booking. For example, the 0 in column available_365 means that the property is in annual rental or booked for a whole year, thus popular. The goal of this section is to identify some differences between popular and less popular (more available) properties in order to distinguish popularity affecting factors from each other. We consider the properties whose annual availability rates for the next year is less than 15% to be popular, more than 80% to be unpopular.

To emphasize, by “popular” in the following analyses, we mean popular for long-term bookings or annual rentals in advance.

A. Host Condition Analysis

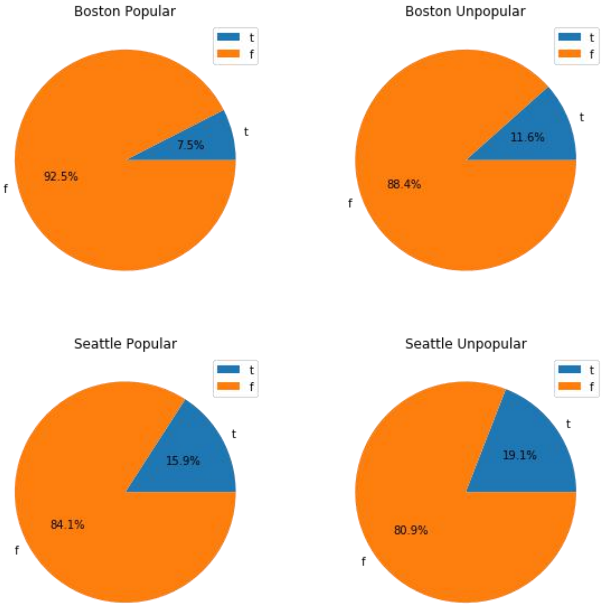
1. host identity verification status



These four pie charts show that less popular (more available next year) properties in both Boston and Seattle has greater identity verification rates, i.e. more hosts of less popular properties had their identity verified on Airbnb. One explanation is that the less popular properties hosts aim more for short-term rental than annual rental, and get their identity verified so as to improve New Year's business performance. The implication is that host's getting identity verified is considered an effective way to boost business and attract customers.

Also, the host identity verified rate is greater in both Seattle popular and unpopular groups, indicating that Seattle Airbnb short-term stay market is slightly more regulated than Boston's.

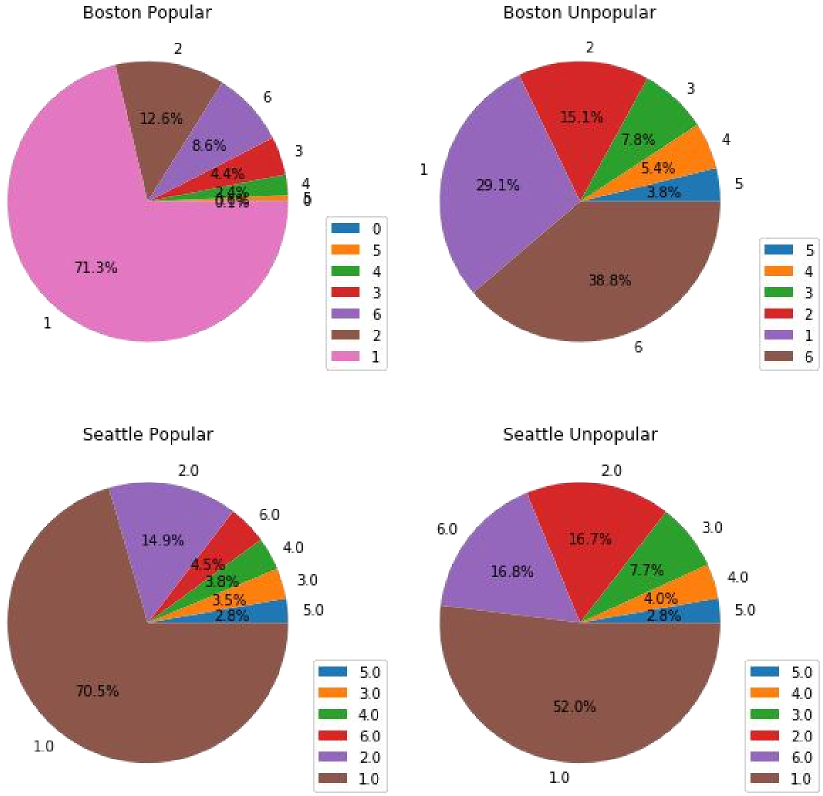
2. superhost rate



More available properties in both Boston and Seattle have more super hosts among them than popular ones. One possible explanation is that super hosts are slightly less inclined to accept annual rentals.

And again, the super host rate in Seattle is greater than Boston's, meaning that Seattle Airbnb market has more experienced hosts than Boston. This could imply that Seattle's tourism is more developed and Seattle's Airbnb market is more mature and competitive than Boston's.

3. host total listing



(Note: to improve the readability of these pie charts, all hosts with more than 5 listed properties (considered chain property hosts) are labeled with 6, regardless of their actual numbers of listing.)

From the charts, we read that:

a. for most of the groups, hosts own only one properties take the greatest percentage

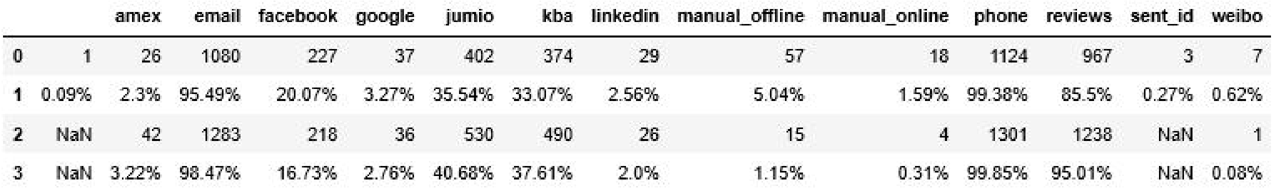
b. Boston more available properties has more hosts (38.8%) that own more than 5 listed properties, i.e. Boston Airbnb market is more dominated by big hosts

c. More big hosts run properties with more availability days, indicating that large hosts tend to prefer flexible short-term rental and their properties are less popular for pre booking

4. host verification accessibility on social media

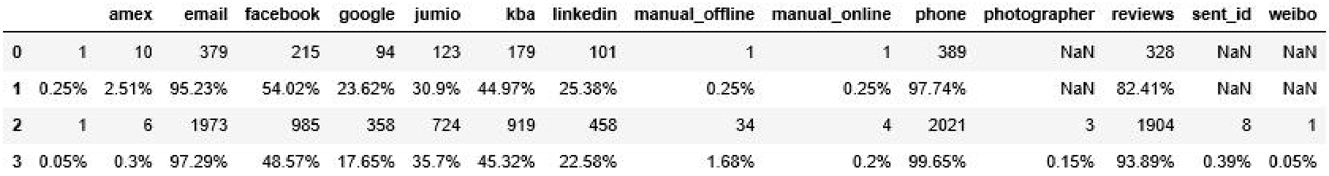
The results of frequency (popular row 1; less popular) the hosts choose a certain social media to verify their identity and the rate (row 2 and 4) a social media is chosen are as following:

a. Boston



		amex	email	facebook	google	jumio	kba	linkedin	manual_offline	manual_online	phone	reviews	sent_id	weibo
0	1	26	1080	227	37	402	374	29	57	18	1124	967	3	7
1	0.09%	2.3%	95.49%	20.07%	3.27%	35.54%	33.07%	2.56%	5.04%	1.59%	99.38%	85.5%	0.27%	0.62%
2	NaN	42	1283	218	36	530	490	26	15	4	1301	1238	NaN	1
3	NaN	3.22%	98.47%	16.73%	2.76%	40.68%	37.61%	2.0%	1.15%	0.31%	99.85%	95.01%	NaN	0.08%

b. Seattle



		amex	email	facebook	google	jumio	kba	linkedin	manual_offline	manual_online	phone	photographer	reviews	sent_id	weibo
0	1	10	379	215	94	123	179	101	1	1	389	NaN	328	NaN	NaN
1	0.25%	2.51%	95.23%	54.02%	23.62%	30.9%	44.97%	25.38%	0.25%	0.25%	97.74%	NaN	82.41%	NaN	NaN
2	1	6	1973	985	358	724	919	458	34	4	2021	3	1904	8	1
3	0.05%	0.3%	97.29%	48.57%	17.65%	35.7%	45.32%	22.58%	1.68%	0.2%	99.65%	0.15%	93.89%	0.39%	0.05%

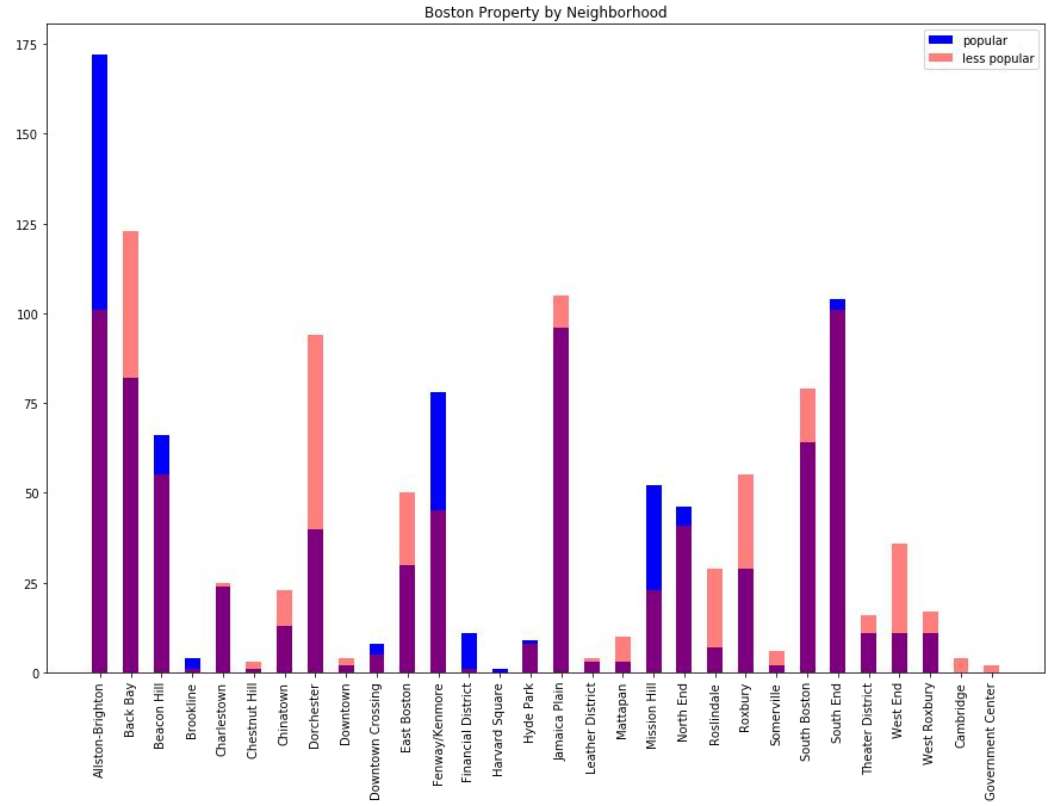
The results show that the most frequently chosen means of contacts are phone, email, reviews, facebook and kba in both cities. And for main stream contact methods, hosts of less popular for advanced booking properties have a higher rate of reachability. This particular kind of “unpopular” properties’ business depend more on instant anytime free booking than pre long-term booking, thus their hosts are more reachable and have higher responsive rates.

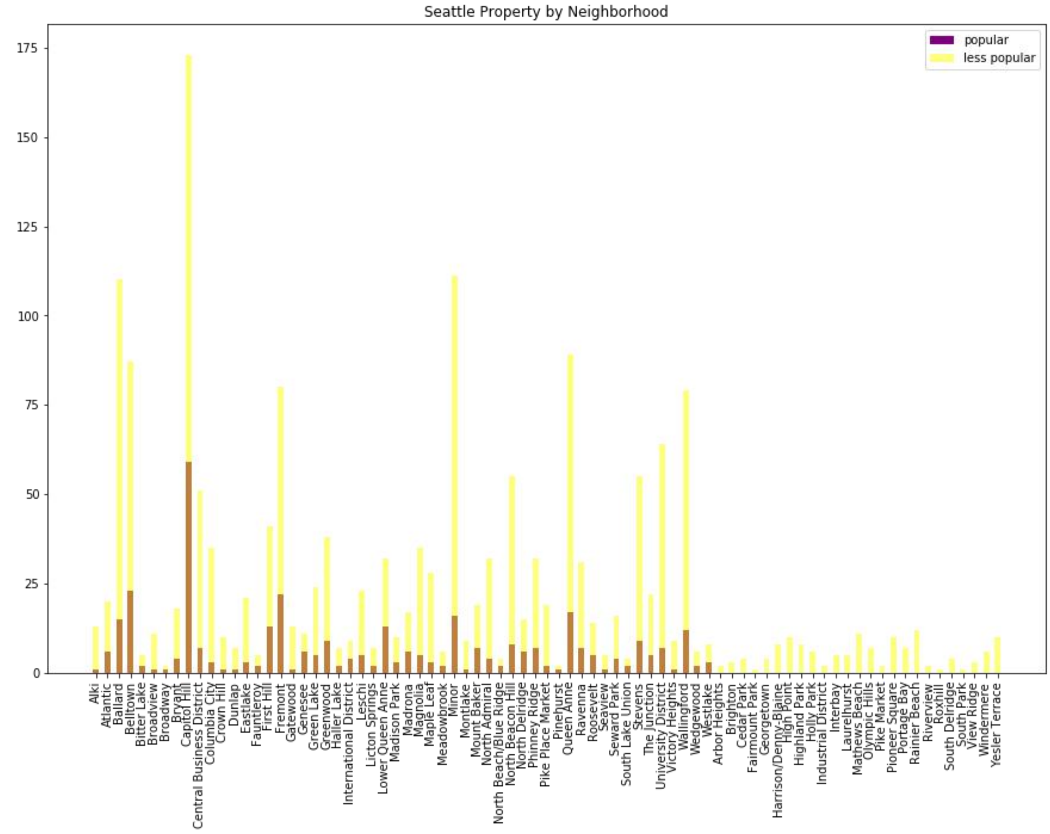
B. Property Condition Analysis

1. neighborhood

The first bar plot shows Boston properties’ distribution by neighborhood. We read that the neighborhoods such as Allston-Brighton, Back Bay, Jamaica Plain and South End are more popular for Airbnb business and touristy in Boston. Allston-Brighton, Fenway/Kenmore and Mission Hill are more popular for advance long-term bookings while most of the rest neighborhoods are more popular for short-term tourism bookings.

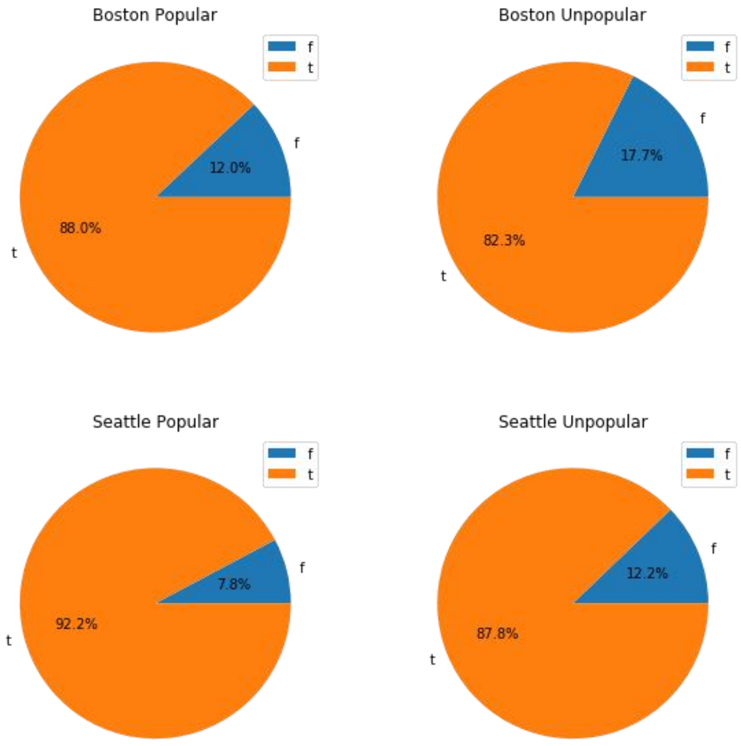
The second bar plot shows Seattle properties’ distribution by neighborhood. The most distinct feature is that the amount of properties more popular for short-term tourism bookings than long-term advance annual rental. We read that the neighborhoods such as Capitol Hill, Ballard, Minor, Bitter Lake, Fremont, Queen Anne and Wallingford are more popular for Airbnb business and touristy in Seattle.





We made an assumption that the reason that caused Seattle having comparatively more properties for short-term rental and Boston having more annual or long-term rental is relative to the composition of local Airbnb customers. While Seattle might attract more tourists and short-term visitors, Boston might have more college students, academic visitors and international students as long-term tenants. This assumption needs further evidential support on the composition of customers.

2. whether location description is exact



Unlike in host statuses analysis section, the popular and less popular for advance long-term booking properties groups in Boston and Seattle show an opposite pattern. Properties that are popular for long-term or annual rental have higher location accuracy rates. One possible implication is that there are more short-term properties entering the market and they are less experienced or double checked the accuracy of their location

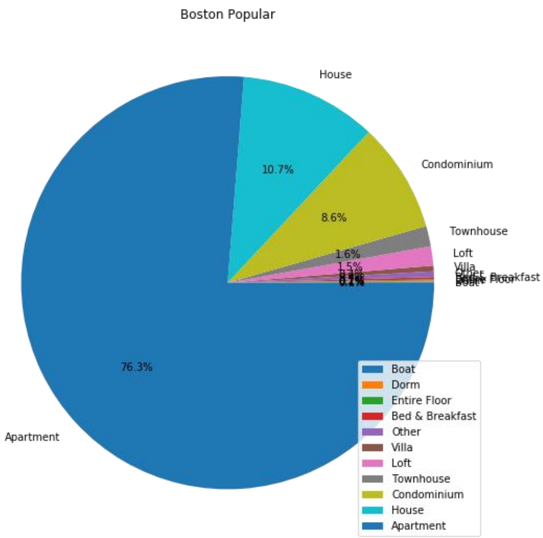
Again, Seattle Airbnb market shows more professionalism in this sense.

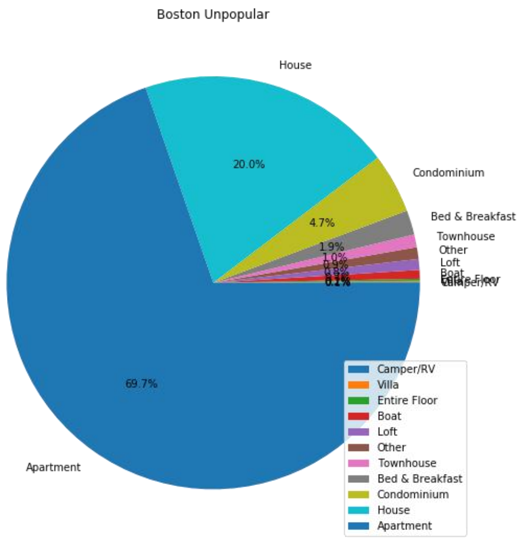
3. property type

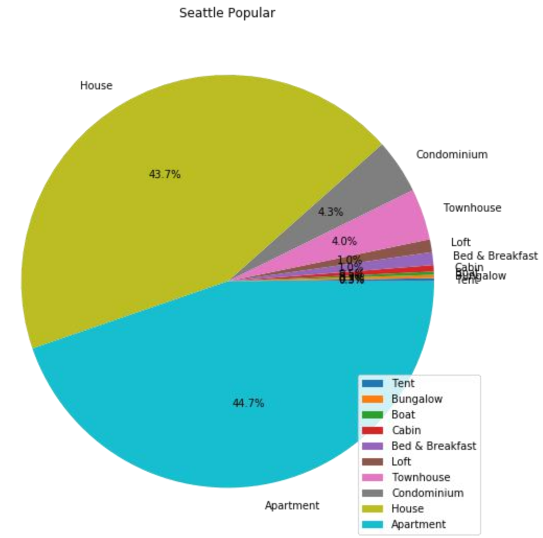
Property types of apartment and house are dominating the market.

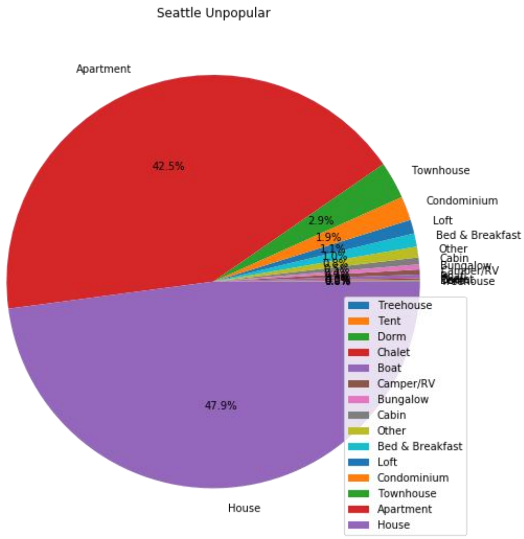
	Boston		Seattle	
	Popular	Unpopular	Popular	Unpopular
Apartment Rate	76.3%	69.7%	44.7%	42.5%
House Rate	10.7%	20.0%	43.7%	47.9%

Property type of apartment takes about 70% in both popular for long-term bookings and unpopular groups in Boston, while long-term bookings seem to prefer apartments slightly more. The apartment and house rate in Seattle's two groups are about the same level. The most distinct result of cross-regional comparison is that more houses are listed in Seattle than in Boston. Based on previous untested assumptions, we made a further assumption that, firstly, more residential properties in Boston are apartments, thus more for rent; secondly, tenants of long-term rental prefer apartments while tourists prefer houses due to budget concerns.





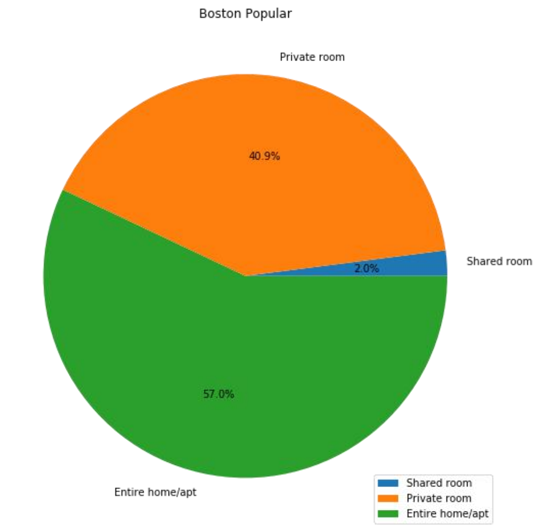


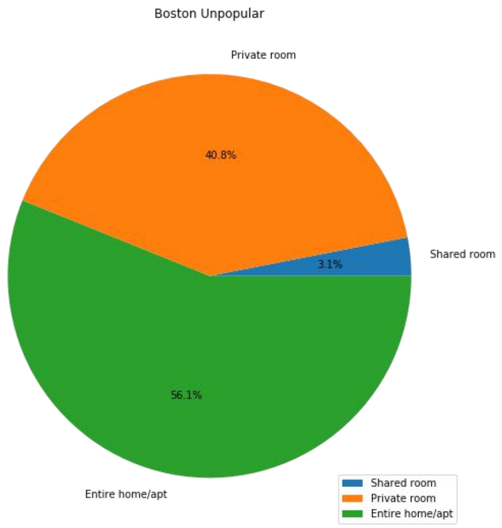


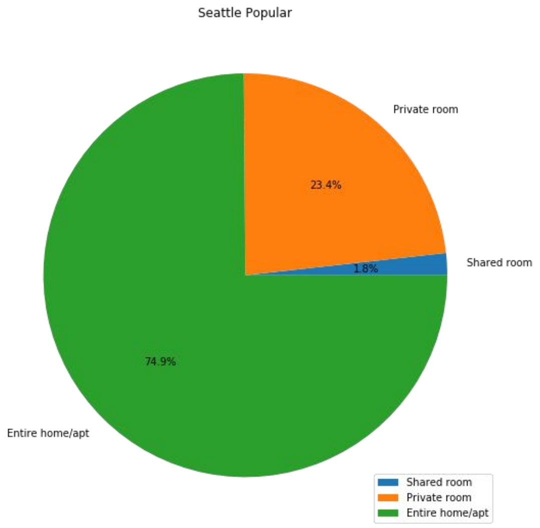
4. room type

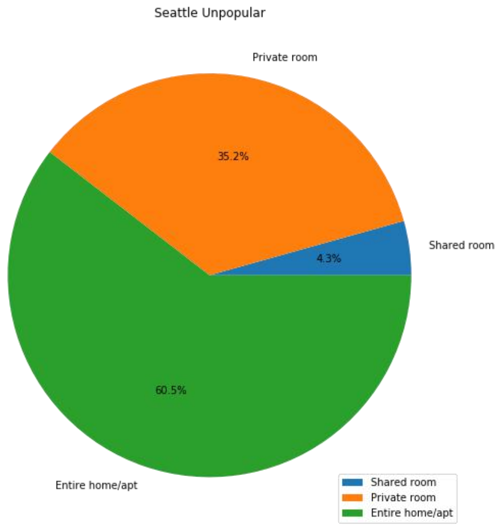
	Boston		Seattle	
	Popular	**Unpopular**	**Popular**	**Unpopular**
Entire Home/Apt	57.0%	56.1%	74.9%	60.5%
Private Room	40.9%	40.8%	23.4%	35.2%
Shared Room	2.0%	3.1%	1.8%	4.3%

Entire Home/Apt rental is the most popular for all the groups. While Boston has more private single rooms (shared apartments) listed, Seattle Airbnb rental is more concentrated on entire home. The result seems consistent with our customer profiling assumption, that comparative economical choice of private single rooms are more popular in Boston for customers such as students with limited budget, and Seattle with better developed tourism treats more short-term tourists who prefer better quality stay in entire homes and apartments.





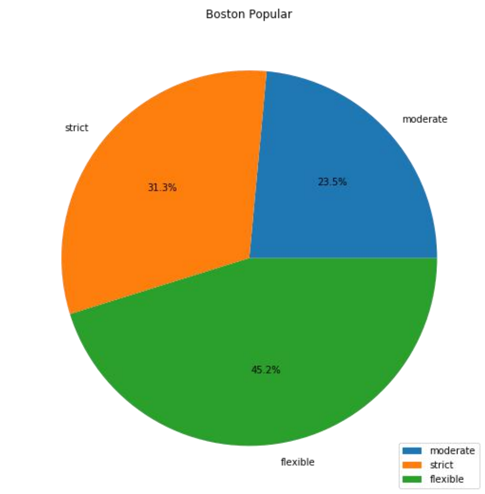


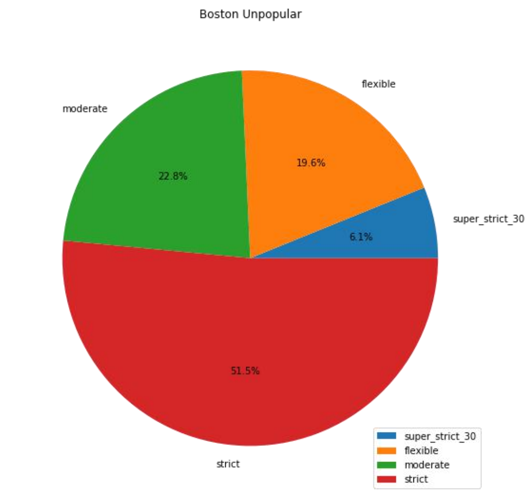


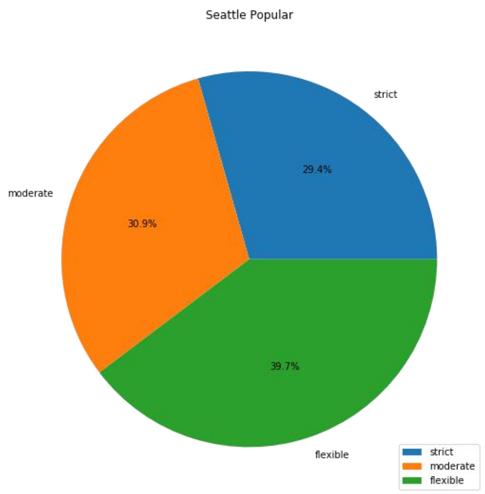
5. cancellation policy

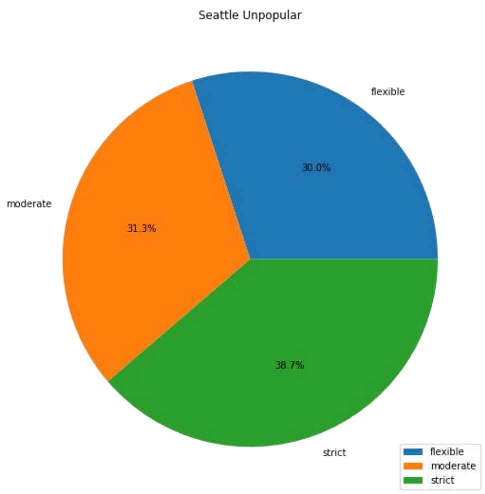
	Boston		**Seattle**	
	Popular	**Unpopular**	**Popular**	**Unpopular**
Flexible	45.2%	29.6%	39.7%	30.0%
Moderate	23.5%	22.8%	30.9%	31.3%
Strict	32.3%	51.5%	29.4%	38.7%
Super Strict		6.1%		

The distributions of cancellation policy in Boston popular, Seattle popular, and Seattle unpopular are quite even among flexible, moderate and strict. An exception is that for Boston unpopular for long-term stay group, the cancellation policy is distinctly stricter, with a percentage of almost 60%. The possible reason for setting strict cancellation policy is to protect the hosts' business from customers' mind changing. To certain extent, we think that this reflects the comparatively not so well-doing of Boston's short-term tourism.



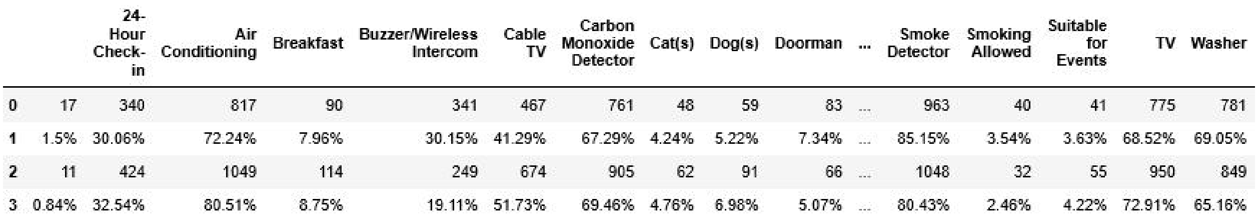






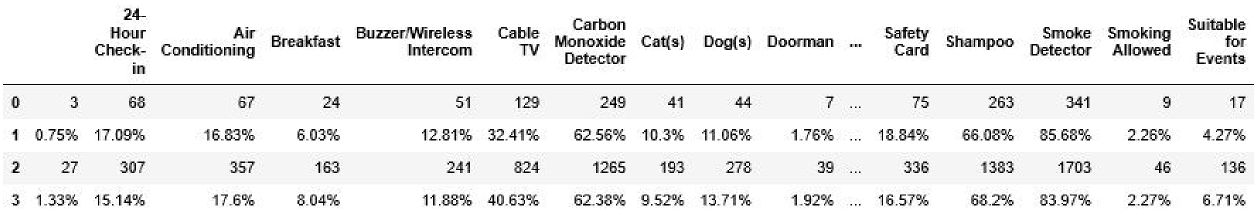
6. amenities

a. Boston



		24-Hour Check-in	Air Conditioning	Breakfast	Buzzer/Wireless Intercom	Cable TV	Carbon Monoxide Detector	Cat(s)	Dog(s)	Doorman	...	Smoke Detector	Smoking Allowed	Suitable for Events	TV	Washer
0	17	340	817	90	341	467	761	48	59	83	...	963	40	41	775	781
1	1.5%	30.06%	72.24%	7.96%	30.15%	41.29%	67.29%	4.24%	5.22%	7.34%	...	85.15%	3.54%	3.63%	68.52%	69.05%
2	11	424	1049	114	249	674	905	62	91	66	...	1048	32	55	950	849
3	0.84%	32.54%	80.51%	8.75%	19.11%	51.73%	69.46%	4.76%	6.98%	5.07%	...	80.43%	2.46%	4.22%	72.91%	65.16%

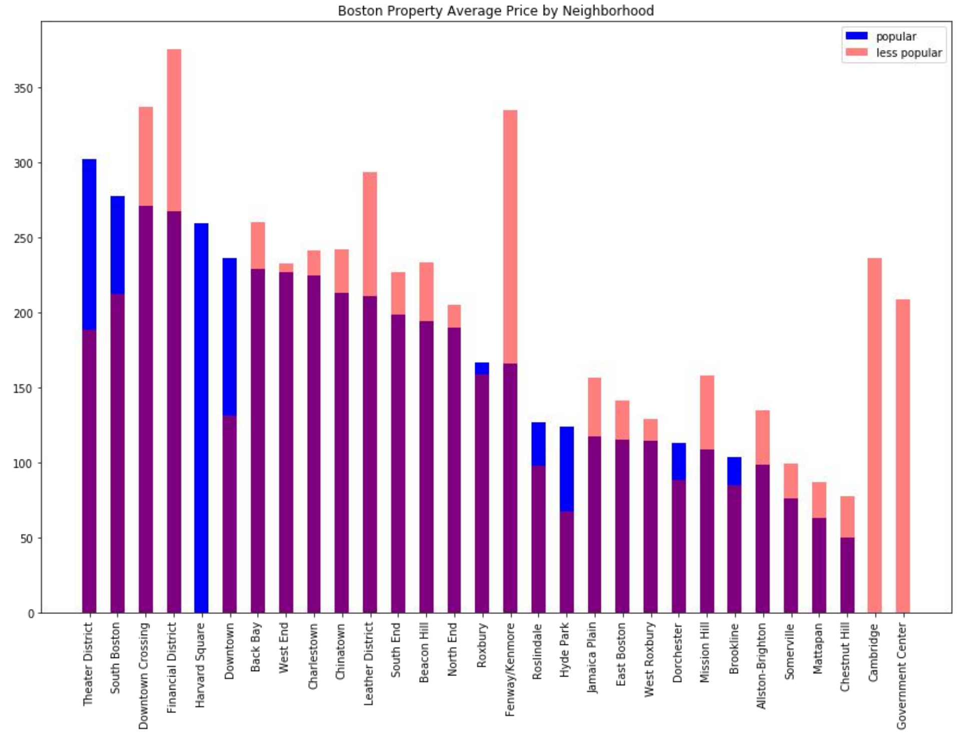
b. Seattle

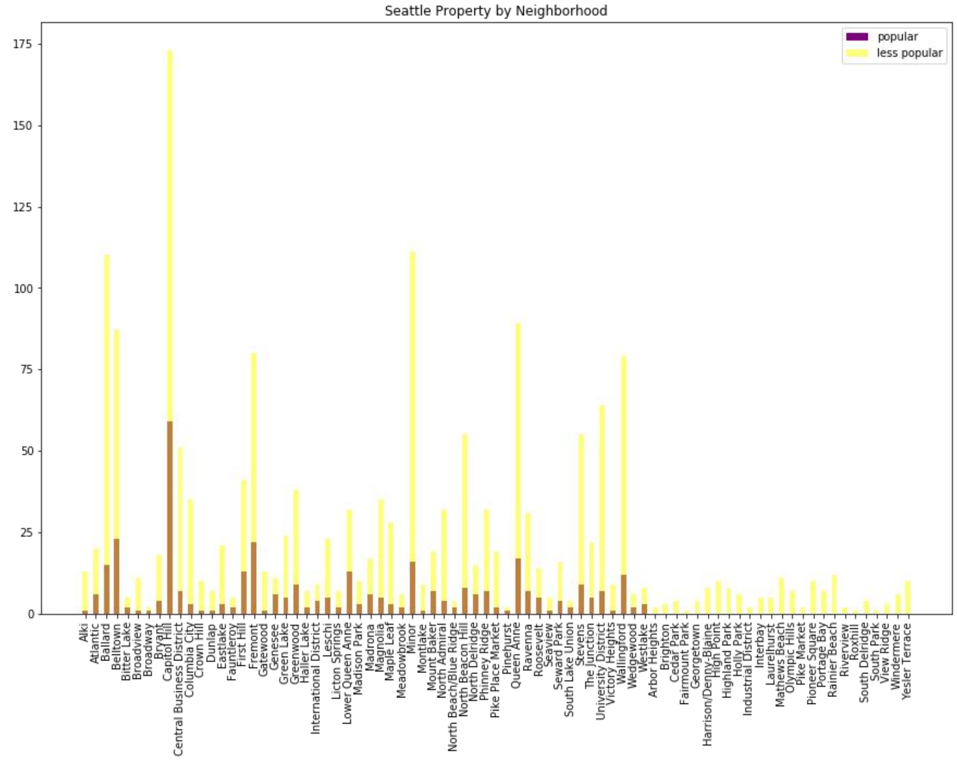


		24-Hour Check-in	Air Conditioning	Breakfast	Buzzer/Wireless Intercom	Cable TV	Carbon Monoxide Detector	Cat(s)	Dog(s)	Doorman	...	Safety Card	Shampoo	Smoke Detector	Smoking Allowed	Suitable for Events
0	3	68	67	24	51	129	249	41	44	7	...	75	263	341	9	17
1	0.75%	17.09%	16.83%	6.03%	12.81%	32.41%	62.56%	10.3%	11.06%	1.76%	...	18.84%	66.08%	85.68%	2.26%	4.27%
2	27	307	357	163	241	824	1265	193	278	39	...	336	1383	1703	46	136
3	1.33%	15.14%	17.6%	8.04%	11.88%	40.63%	62.38%	9.52%	13.71%	1.92%	...	16.57%	68.2%	83.97%	2.27%	6.71%

From results above, we concluded that some most common amenities features that Boston and Seattle share are wireless Internet, TV (w/o cable), smoke detector, carbon monoxide detector, kitchen, washer, air conditioning, and shampoo, some of which, such as kitchen and washer, are advantages of Airbnb compare to normal hotels.

7. price by neighborhood





From the two bar plots above, we read that:

a. properties' average prices in Boston are averagely much higher than in Seattle; this is possibly because that the prices depicted in the plots above are not divided by number of people the property accommodates and the number of nights (the duration of the stay), which also reflects that Boston Airbnb has more customers of longer stay

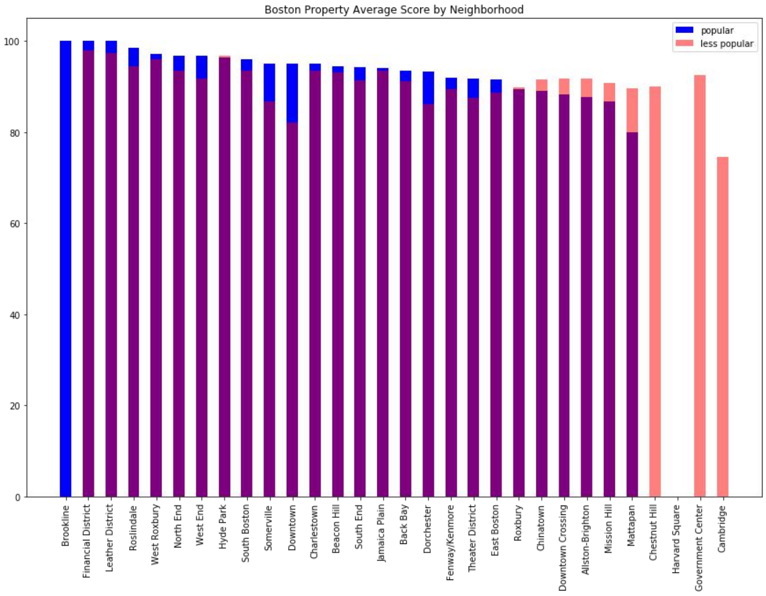
b. the prices of properties more popular for short-noticed bookings are averagely higher than long-term pre bookings' in both Boston and Seattle, while the differences are smaller in Boston comparing to Seattle

c. cross compared with the bar plot of number of properties by neighborhoods, the prices for stays in popular neighborhoods are higher than less popular and competitive markets

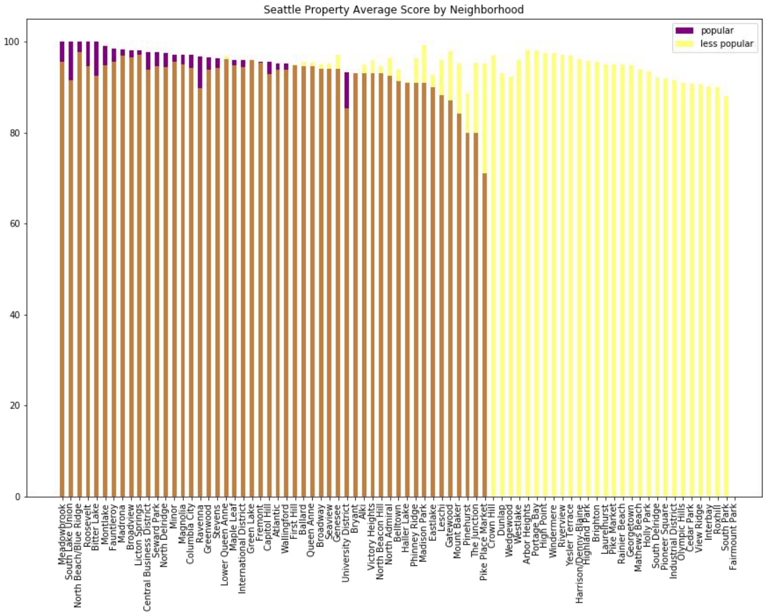
The prices for long-term and short-term stays in Boston and Seattle fit our previous analysis that long-term rental is more popular in Boston and shortly-advance bookings are the major type of bookings in Seattle.

8. review scores rating

a. average by neighborhood

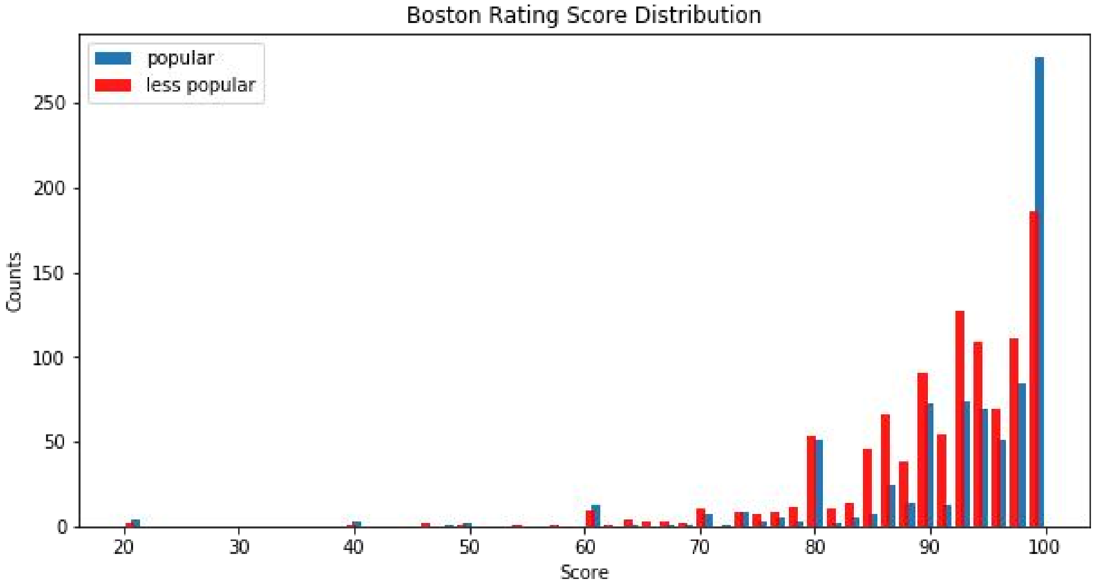


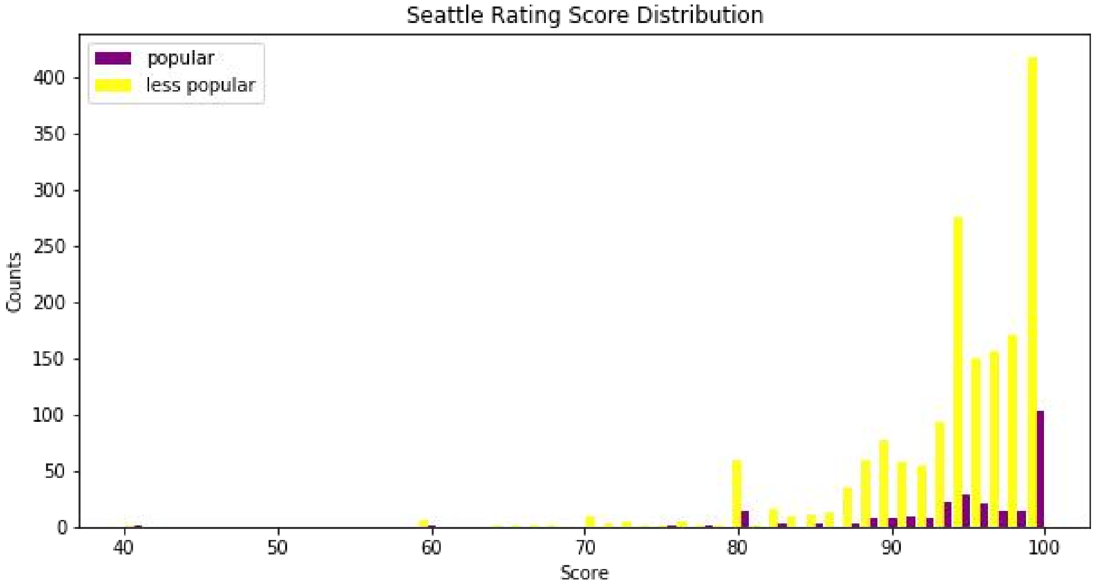
From the bar plot of average score rating in Boston, we concluded that there is no significant difference between the rating score of properties that are popular for long-term and short-term bookings respectively.



The bar plot of average score rating in Seattle reflects distinctly that hosts in some neighborhood focus only and entirely on short-term booking business.

b. score histogram





From both histograms above, we read that more of the properties that are less popular for long-term advance bookings receive high score rating from reviewers.

Pitifully, during analyzing, we realized that the histograms of Boston and Seattle Rating Score Distribution reflect a significant short coming of our research design. When deciding the availability rate to group popular and less popular for pre-booking properties, we did not control the size of both groups to be about the same in Seattle. That is to say, while the total amounts of properties in popular and unpopular groups in Boston were about the same, the selected sample in Seattle lean toward the unpopular group, which may compromise the precision of our results (only for a couple of topics).

III. Given above, what conclusions can we draw about customers' preferences and overall conditions in Boston and Seattle Airbnb market?

Concluding from previous analysis, we believe that Boston and Seattle Airbnb markets has very different targeted types of customers and both markets have developed unique features to suit different needs of Airbnb homestays. The major differences are as in the following table.

Feature	**Boston**	**Seattle**
Booking types	Long-term advance bookings and annual rental	Shortly-noticed (not so advance) bookings
Targeted customers (needs more backup evidence)	Longer-stay residents, such as college students and academic visitors, who prefer single rooms in apartments due to budget concerns	Shorter-stay customers, such as tourists and visitors, who prefer apartments and houses about equally and prefer entire room due to concerns of the quality of the stay
Market maturity	Less mature	More mature
Market regulation	Less regulated	More regulated, has more experienced and professional hosts
Tourism (untested)	Economy weigh less on tourism	More developed, more touristy
Gross price	More expensive, has a wider spread	Cheaper, more centered

3.3 Price Analysis

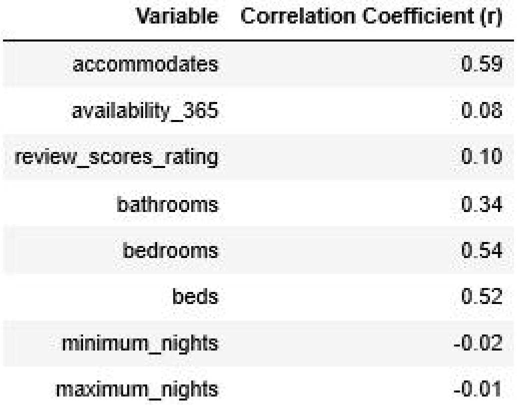
I. What factors influence the price most?

In this section, to determine what factors influences the prices of stays, we divided potential price influencing factors into two groups, quantitative and categorical. For quantitative influencing factors, we calculated the correlation coefficient to evaluate the association between the independent variables and the responsive variable, price. For categorical factors, we divided the price per person (because the results of analyses on quantitative factors shows that the price is correlated to number of people a property accommodates) into different price levels and calculated the chi-square statistics to see whether the two variables are independent from each other's influence.

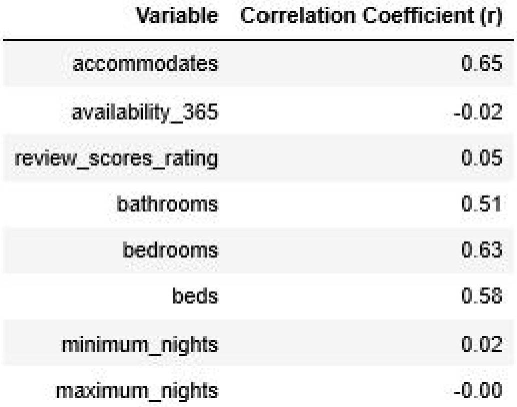
	Potential price affecting factors
Quantitative	accommodates, availability 365, review scores rating, bathrooms, bedrooms, beds, minimum nights, maximum nights
Categorical	neighborhood, property type, room type, cancellation policy

A. Quantitative variables

Results: Boston (left) & Seattle (right)



Variable	Correlation Coefficient (r)
accommodates	0.59
availability_365	0.08
review_scores_rating	0.10
bathrooms	0.34
bedrooms	0.54
beds	0.52
minimum_nights	-0.02
maximum_nights	-0.01



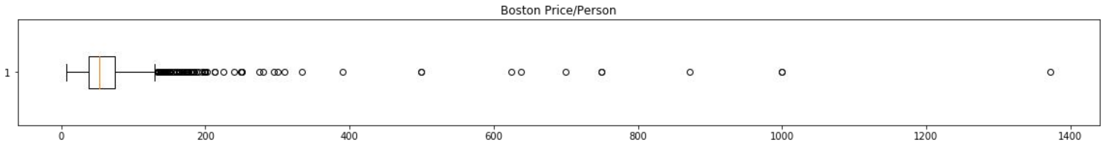
Variable	Correlation Coefficient (r)
accommodates	0.65
availability_365	-0.02
review_scores_rating	0.05
bathrooms	0.51
bedrooms	0.63
beds	0.58
minimum_nights	0.02
maximum_nights	-0.00

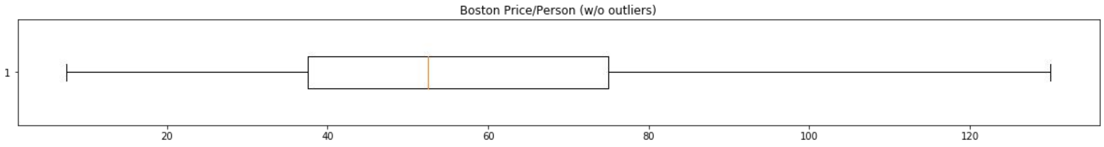
Among the selected quantitative variables, we read that the r statistics of the number of people accommodates, bathrooms, bedrooms and beds are closer to 1, indicating a moderately strong correlation between the variable and the price. And the number of people a property accommodates is the most direct and strong explanatory factor

B. Categorical variables

As the price is correlated to number of people accommodates, to identify real price influencing factors, we calculated price per person stay in this part and divided the price/person into subgroups for categorical analysis.

1. Boston price levels





After using the box plot to study the distribution of prices, we divided Boston prices into 3 groups: low (x<=35), moderate (35<x<=80), and high (80<x). (Boston prices are wider spread; if subcategorized into more than 3 groups, there will be a lot of empty values, compromising the accuracy of the results.)

a. Chi-square

Price level counts	
Low	646
Moderate	1774
High	619

Observed: [619, 646, 1774]

Uniform would be: [1013.0, 1013.0, 1013.0]

Significant difference? p=0.0000, chi^2=857.8934

The distribution of price is significantly different among price levels. More prices are centered in the range of moderate prices.

b. By neighborhood

Significant difference? p=0.0000, chi^2=769.1826

Prices are significantly different between different neighborhoods, meaning that location is a significant influencing factor of price.

c. By property type

Price level counts		
	Apartment	House
Low	407	178
Moderate	1304	256
High	490	40

Significant difference? p=0.0000, chi^2=104.2456

Prices are significantly different between apartments and houses. Apartments comparatively have more properties of high price. Therefore, property type is another significant influencing factor of price.

d. By room type

Price level counts			
	Entire home/apt	Private room	Shared room
Low	283	343	20
Moderate	1113	636	25
High	464	145	10

Significant difference? p=0.0000, chi^2=136.0938

Room type affects prices significantly as well.

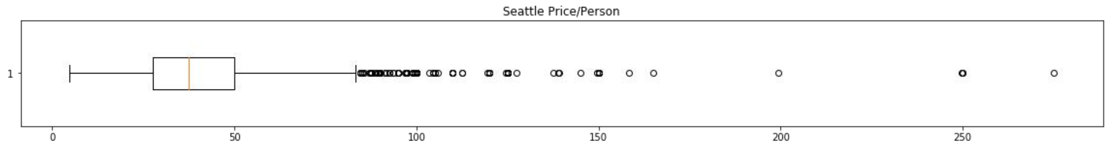
e. By cancellation policy

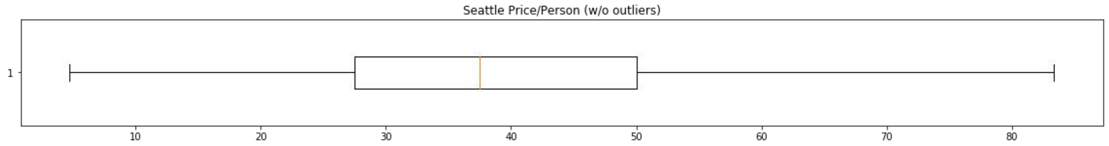
Price level counts				
	flexible	moderate	strict	Super strict
Low	220.0	165.0	261.0	0.0
Moderate	450.0	436.0	873.0	15.0
High	148.0	154.0	255.0	62.0

Significant difference? p=0.0000, chi^2=200.6714

Cancellation policy is another significant price affecting factor.

2. Seattle price levels





We divided more centered Seattle prices into 5 groups: economical (x<=25), lower moderate (25<x<=40), upper moderate (40<x<=80), expensive (80<x<=100) and luxury (100<x). Despite different actual numbers and more groups of classification, the conclusions drawn from the chi-square tests are same as conclusions drawn for Boston dataset.

II. Conclusion

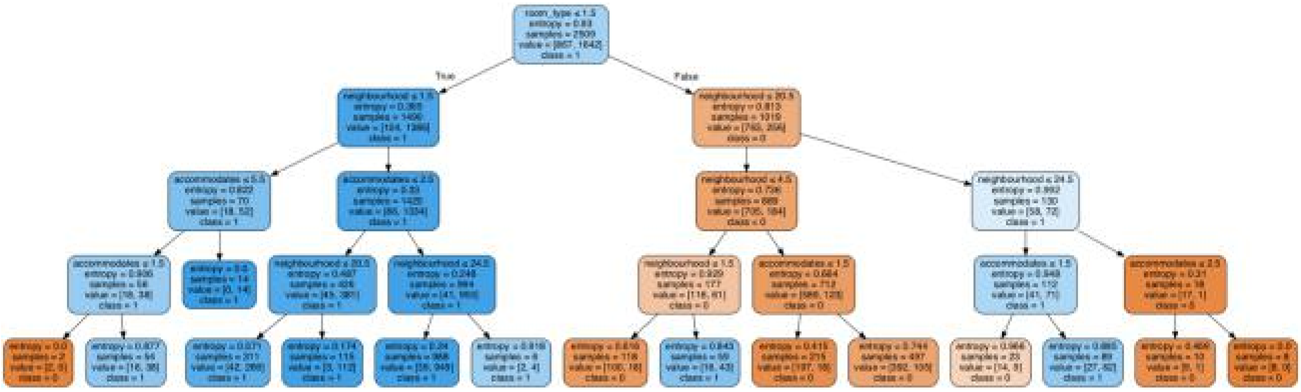
Categorical variables of neighborhood, property type, room type, cancellation policy are all associated with price levels at 95% significance level.

III. Overpriced decision tree——Whether a property is overpriced in similar listings?

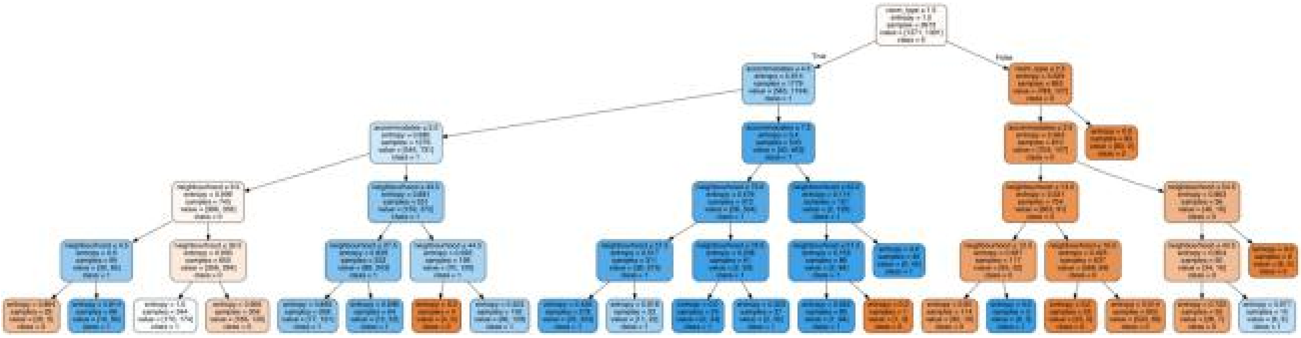
Here, we want to provide the customers with reference information on the prices for the sake of budget plan. We did data preprocessing for this part: If the Airbnb room price is larger than 100 then we labeled with 1, if not, we labeled with 0. Besides, considering decision tree can only process numeric number, we preprocess the room type and neighborhoods into number.

With the help of the decision tree, according to information about house type, accommodates, neighborhoods, we can know whether the house price is suitable.

For Boston, the accuracy of the unpruned tree is 88.38%, but it is unexplainable and not easy to understand. After pruning, we find when max depth is four, we get the biggest accuracy, which is 87.83%.



For Seattle, the accuracy of the unpruned tree is 78.18%, but it is unexplainable and not easy to understand as well. After pruning, we find when max depth is five, we



get the biggest accuracy, which is 78.62%.

Using existing information about accommodates, neighborhoods, house type and the decision tree above, we can find whether the price should be higher or lower than

$100.

Part 4: Conclusion

I. From comparison analysis above, we have these findings:

A. The Airbnb in Boston is relatively more concentrated, but in Seattle, Airbnb has higher star rate and more popular.

B. The overall room type distribution is similar in Boston and Seattle, and most are entire home or apartment. In Seattle, the entire type ranks top in almost every neighborhoods, while in Boston, here are some neighborhoods having more private room type.

C. Airbnb price is relatively lower in Seattle than in Boston. Back Bay in Boston is the only neighborhood listed in both top five highest total prices and the average price.

D. Airbnb rooms are clustered with a small group of hosts in both cities, and it shows the relatively high monopoly phenomenon in Boston Airbnb market.

E. Neighborhood, property type, room type, and cancellation policy are significant price influencing factors.

F. Feature comparisons

Feature	**Boston**	**Seattle**
Booking types	Long-term advance bookings and annual rental	Shortly-noticed (not so advance) bookings
Targeted customers (needs more backup evidence)	Longer-stay residents, such as college students and academic visitors, who prefer single rooms in apartments due to budget concerns	Shorter-stay customers, such as tourists and visitors, who prefer apartments and houses about equally and prefer entire room due to concerns of the quality of the stay
Market maturity	Less mature	More mature
Market regulation	Less regulated	More regulated, has more experienced and professional hosts
Tourism (untested)	Economy weigh less on tourism	More developed, more touristy
Gross price	More expensive, has a wider spread	Cheaper, more centered

Considering the relatively high price, monopoly together with less popular and clustering in Boston, we recommend product operators in Boston focus more on the price and user experience, customers using the decision tree to help avoid overpriced situation. Boston market should be further adjusted to long-term pre-booking customers and Seattle market should better fit for booming tourism and increasing needs as replacements of traditional hotels and hostels.

II. **Discussion**

The results about geographical distribution and property distribution deserve to special attention. Limiting the Airbnb price if a host provides too much houses maybe help to improve customer experience. A more mature decision tree can be used to help customers or host decide the suitable house price.

There are some limitations in this study.

The first one is about the neighborhood. According to our datasets, Seattle has 87 neighborhoods while Boston has 25 neighborhoods, both are overwhelming. To make results more readable, we simply exclude smaller neighborhoods, maybe leading to missing some important findings. Manually group neighborhoods into “regions” (all the neighborhoods grouped together geographically), or “neighborhood types” (wealthy, working-class, commercial, etc.) can work better.

The second is about the decision tree. We hope to have a decision tree to help find suitable price range with targeting conditions, such as neighborhoods, accommodates and so on. However, due to limiting time and knowledge about decision tree, this project can only judge whether the Airbnb price is higher than $100 or not.

There are also some problems with sample selection in the customer preference analysis part which may compromise the accuracy of the results (as described earlier).

If we have more time and budget to continue this study, we hope to do A/B test to find what attributes customer satisfaction, which can work as guide for Airbnb hosts. Besides, we will try different kinds of coupons with Airbnb pricing, to find which sale strategy can help improve customer amount.