Boston Airbnb Open Data Analysis Report

Tian Gao, Xinyue Li

Github: https://github.com/feiyue33/cs506-final-project

Background

Airbnb has successfully disrupted the traditional hospitality industry as more and more travelers decide to use Airbnb as their primary accommodation provider. Since its beginning in 2008, Airbnb has seen an enormous growth, with the number of rentals listed on its website growing exponentially each year. However, we have no idea about how Airbnb is really used in Boston. Therefore, we want to explore the public data set to help users get more information behind Airbnb.

Data Description

Our data set is an open source data set provided by http://insideairbnb.com/get-the-data.html. We downloaded the latest version and the listing file of 2018.

- listings_2019.csv: detailed listings data for Boston until September 22, 2019.
- listings_2018.csv: detailed listings data for Boston until December 13, 2019.
- calendar_2019.csv: detailed calendar data for listings in Boston
- reviews_2019.csv: detailed review data for listings in Boston
- neighborhoods.csv: neighborhood list for geo filter
- neighborhoods.geojson: GeoJSON file of neighborhoods of the city

Data Preprocessing

listings_2019.csv

This file contains 106 columns and many of them, like scrape information or listing url, are not useful for us. Therefore, we dropped those unnecessary columns and kept 27 columns in total which might be useful for future analysis.

The next thing we do is to check if duplicated data exists to make sure that every record is unique. Then, we checked the missing value in each column. According to the result, we replaced the missing value in "*reviews per month*"with zero, and replaced the missing value in descriptive columns like "*name*", "*neibourhood_overview*", "*description*" with the word "Unknown".

We also checked the type of each column. We noticed that the data type of "*price*" attribute is "*object*", and we converted it to numeric type for the next step. The head of this data set after preprocessing is as follows.

	id	name	summary	space	neighborhood_overview	transit	host_id	neighbourhood_cleansed	latitude	longitude	room_type
0	3781	HARBORSIDE-Walk to subway	Fully separate apartment in a two apartment bu...	This is a totally separate apartment located o...	Mostly quiet (no loud music, no crowed sidewa...	Local subway stop (Maverick Station on the Bl...	4804	East Boston	42.36524	-71.02936	Entire home/apt
1	6976	Mexican Folk Art Showcase in Boston Neighborhood	Come stay with me in Boston's Roslindale neigh...	This is a well-maintained, two-family house bu...	The LOCATION: Roslindale is a safe and diverse...	PUBLIC TRANSPORTATION: From the house, quick p...	16701	Roslindale	42.29244	-71.13577	Private room
2	8789	Curved Glass Studio/1bd facing Park	Bright, 1 bed with curved glass windows facing...	Fully Furnished studio with enclosed bedroom. ...	Beacon Hill is a historic neighborhood filled ...	The MBTA site is a great reference for public ...	26988	Downtown	42.35919	-71.06265	Entire home/apt
3	9273	Stay at "HARBORVIEW" Walk to subway	NaN	Available $200.00 per night/seven night minim...	NaN	NaN	4804	East Boston	42.36461	-71.02902	Entire home/apt
4	10730	Bright 1bed facing Golden Dome	Bright, spacious unit, new galley kitchen, new...	Bright one bed facing the golden dome of the S...	Beacon Hill is located downtown and is conveni...	The Red Park Street Train stop and the Green B...	26988	Downtown	42.35840	-71.06185	Entire home/apt

listings_2018.csv

The preprocessing step for *listings_2018.csv* is the same as *listings_2019.csv*.

calendar_2019.csv

This file includes 2,084,515 records and none of them have empty columns. However, "*price*" column in this file is not numeric. Therefore, we need to convert the data type to float number for future analysis. The head of this file after preprocessing is attached below.

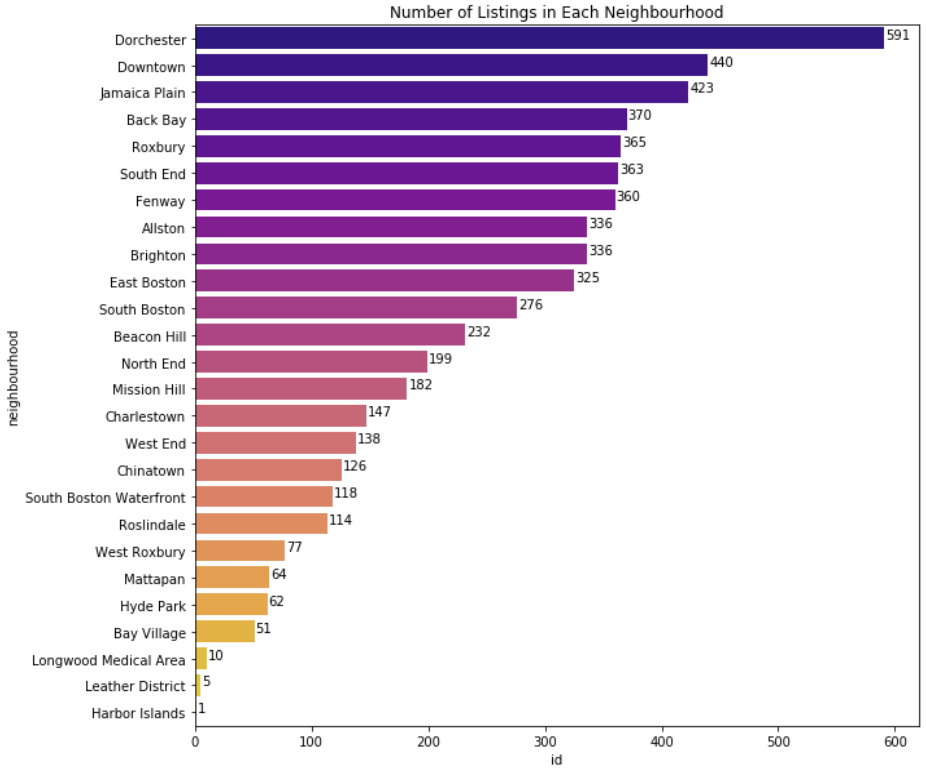
	listing_id	date	available	price	adjusted_price	minimum_nights	maximum_nights
0	957224	2019-09-22	f	275.0	$275.00	3	1125
1	957224	2019-09-23	f	275.0	$275.00	3	1125
2	957224	2019-09-24	f	275.0	$275.00	3	1125
3	957224	2019-09-25	f	275.0	$275.00	3	1125
4	957224	2019-09-26	f	275.0	$275.00	3	1125

Exploratory Data Analysis

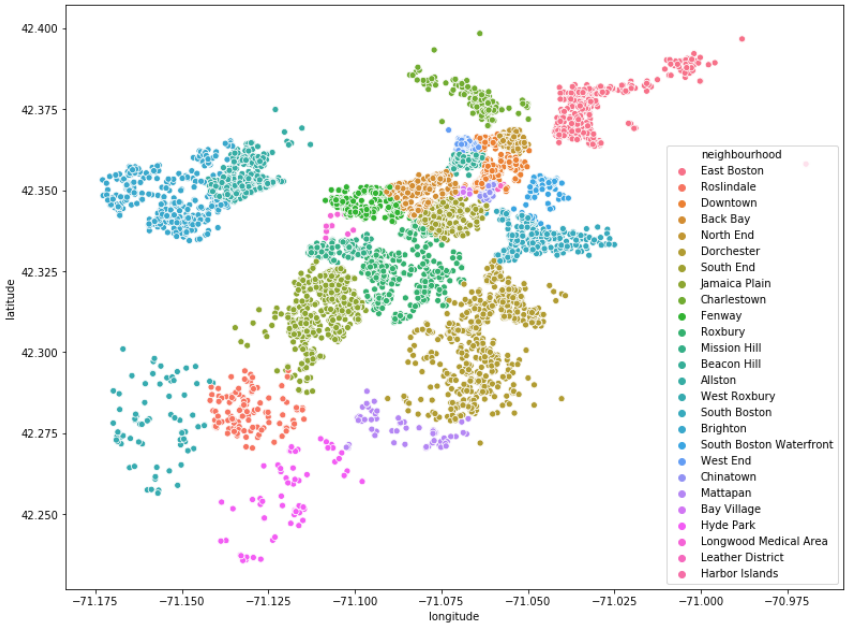
In this section, we explored the data set and used the result to answer some fundamental questions.

Q1: How many listings are in each neighborhood in Boston?

There are 5,711 Airbnb listings in Boston until September 2019. We calculated the sum of listings in each neighborhood and visualized the result. We could find that in 2019, Dorchester has the most listings in Boston. Some popular areas, like Downtown, Bacy Bay, and Fenway has many listings as well.

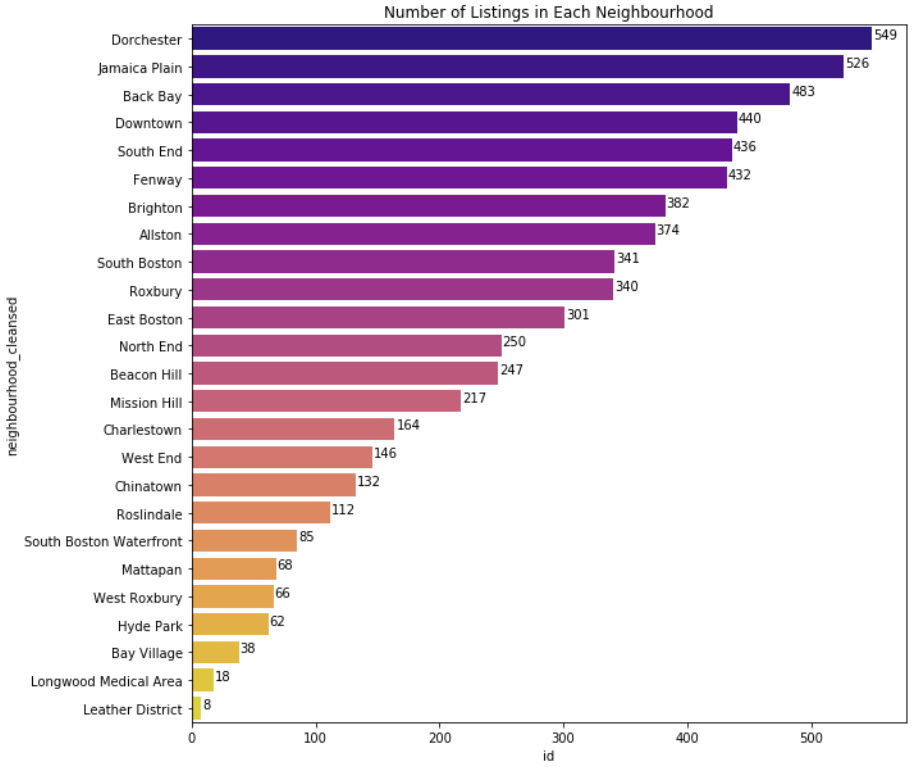


We also generated a scatter plot using coordinates and used different colors to represent each area.



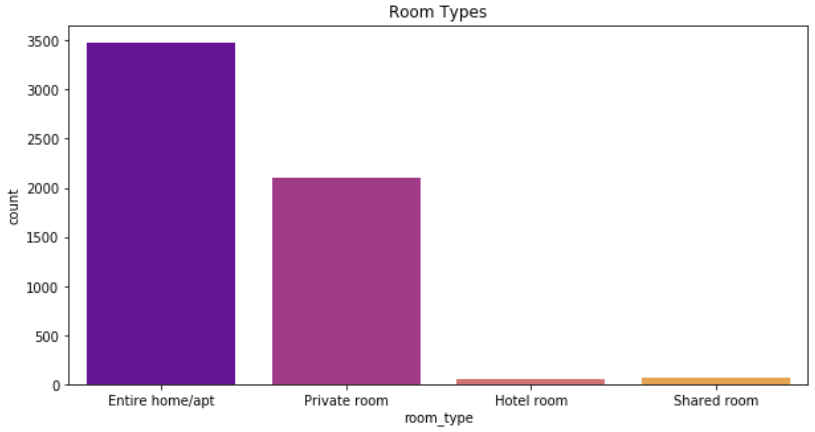
Q2: How does the number of listings change from 2018 to 2019?

There are 6,217 listings by the end of 2018. Therefore, compared with 2018, the number of listings in Boston actually decreased, especially in Jamaica Plain, South End and Fenway area.

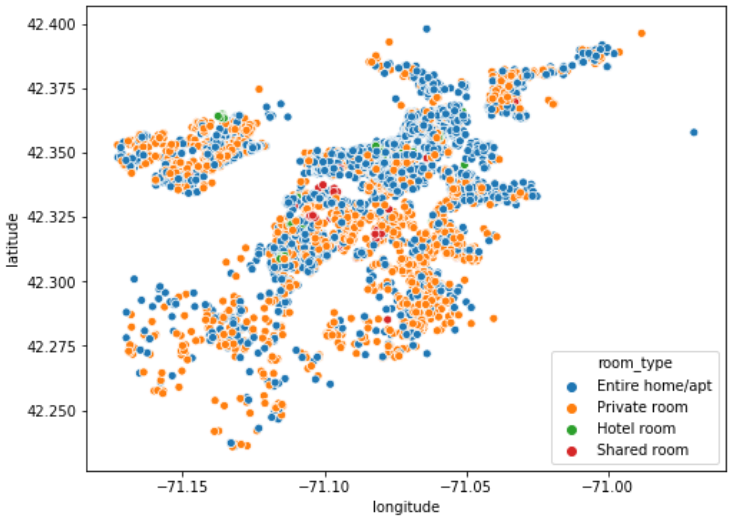


Q3: Which type of room are popular in Boston?

There are four types of room for Airbnb: entire home/apt, private room, hotel room and shared room. As for room type, we could find that more than 95% of the rooms are private room or entire room or apartment. Also, about 60% of the rooms are entire room or apartment.

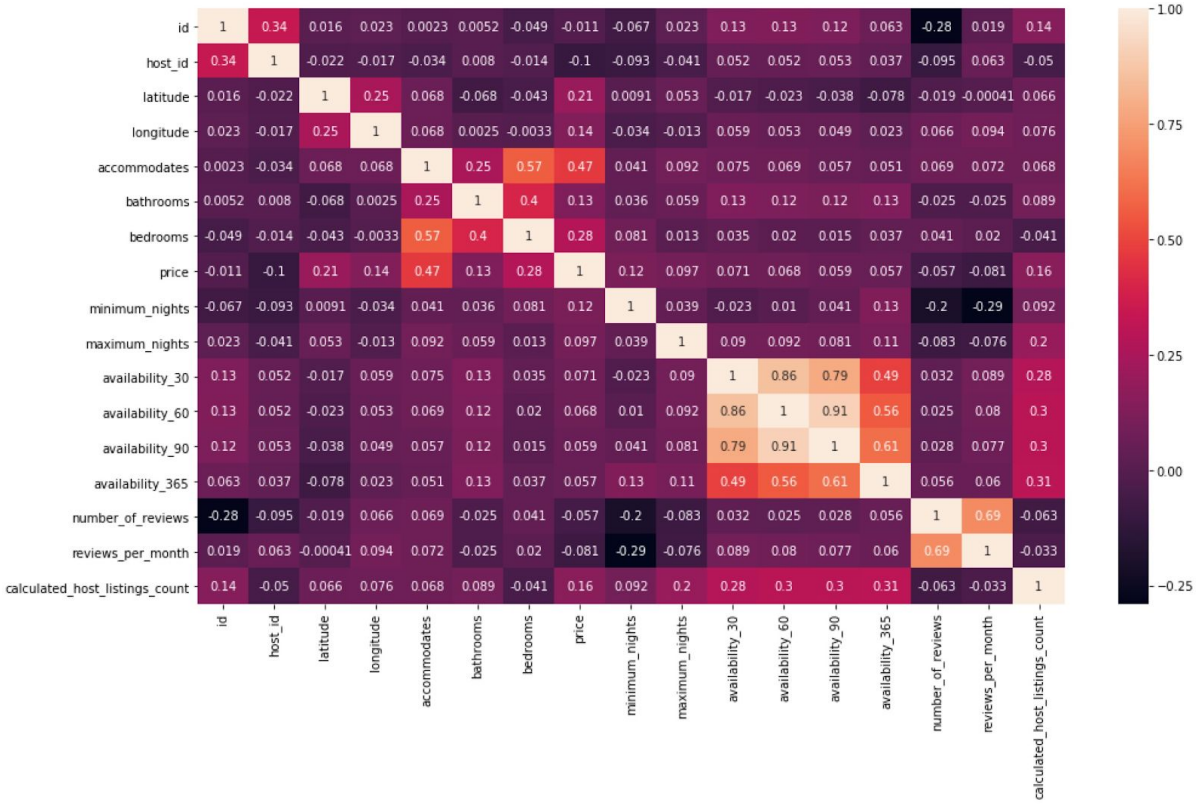


We also used the coordinates to see the distribution of room types. Combined the figure with Boston map, we could find that a large proportion of entire room or apartment listings are located in the Back Bay area.



Q4: Which attributes affect the price of Airbnb most?

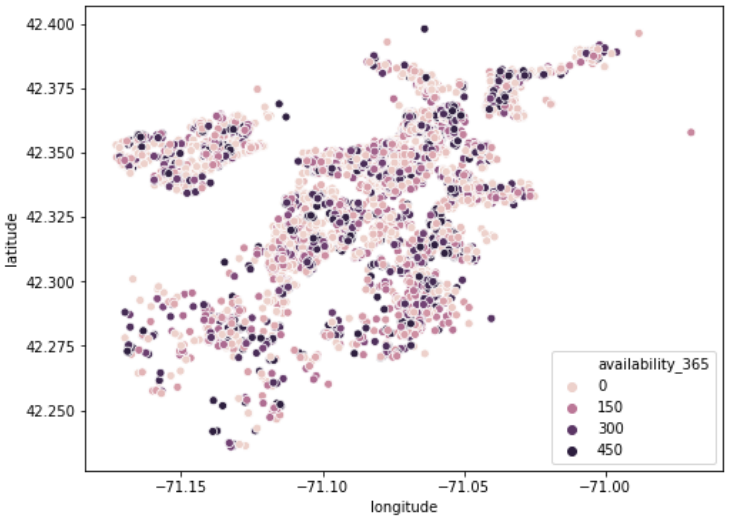
To answer this question, we calculated the correlation between different variables. From the result, we could see that accomodates, bedrooms and latitude(location) are the top three attributes affect the price most.



Q5: Which areas are popular in Boston?

We analyzed the total number of available days during a year to answer this question and plotted the map using coordinates. We found that Fenway and Back Bay are the most popular

areas in Boston because only a fews days in one year are available for rooms in these two areas.

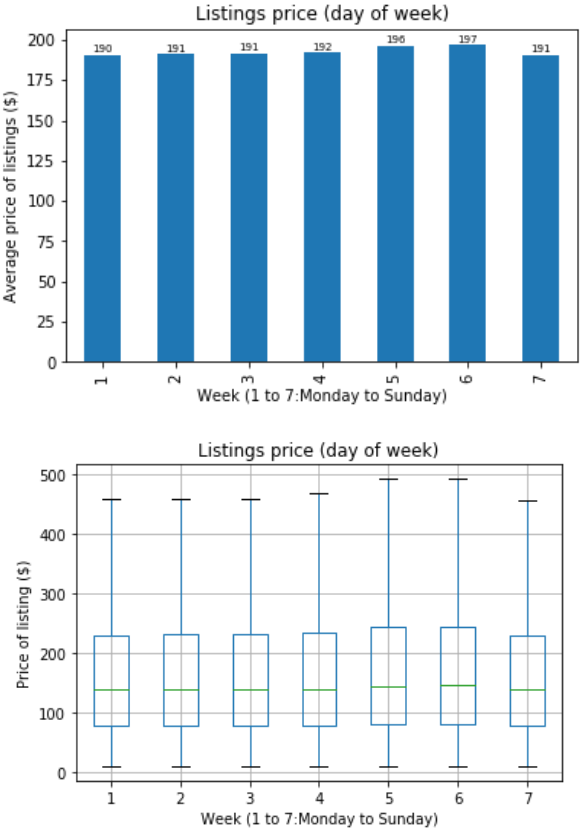


Q6: What is the trend of Airbnb price in Boston? What time is an economical choice for tourists visiting Boston?

We analyzed the calendar information to answer this question. The first step is to extract new features from "*date*" attribute. We built two new features "*month*" and "*days of week*" to explore the trend of price in two different angles.

First, we calculated the average price of each month. The result shows that the price of summer, from June to August, is the highest, and the price during winter is much lower. It is a reasonable result since the winter in Boston is so cold that lower tourists choose to visit Boston at that time.

Then, we tried to find the trend of price in a week. We could see that the price of Friday and Saturday is a little bit higher than the other days.

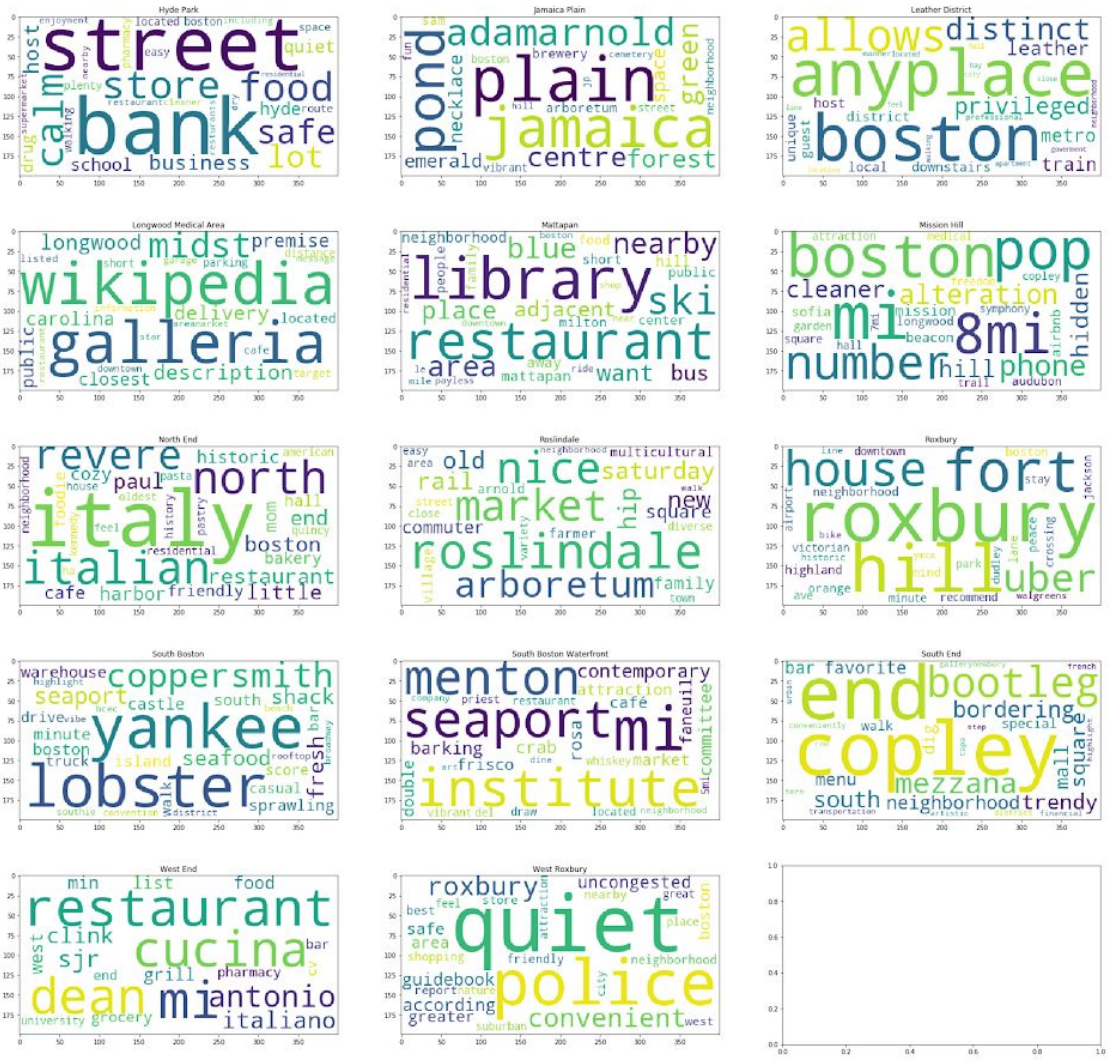


Neighborhood Vibe Analysis

We used "*neighborhood_overview*" provided by host in order to describe the vibe of each neighborhood. This column has already cleaned in preprocessing step and doesn't include missing value. Then, we built a model using techniques learnt from class, including TF-IDF vectorizer and classifier, to find phrases that characterize each neighborhood. We selected top 5 words to find if the model works. We could see that the result is not perfect, but it did extracted some useful information. For example, the vibe of Back Bay area is fashionable and a good choice for shopping, which is correspond to what we know about Back Bay.

```
Allston: greater, greater boston, harvard square, harvard, allston
Back Bay: historic bay, fashionable, shopping, newbury, bay
Bay Village: college, village, marathon, boston marathon, bay village
Beacon Hill: antique, state, hill, beacon hill, beacon
Brighton: college, brighton center, college boston, boston college, brighton
Charlestown: hill monument, monument, bunker, bunker hill, charlestown
Chinatown: rated craft, avery bar, avery, culture, chinatown
Dorchester: savin hill, diverse neighborhood, savin, ashmont, dorchester
Downtown: crossing highlight, downtown crossing, heart downtown, crossing, walk property
East Boston: end old, property le, airport, east, east boston
Fenway: sweet, kenmore, sweet cheek, cheek, fenway
Harbor Islands: community lived, beachfront, beachfront community, point shirley, oriented residential
Hyde Park: street school, school street, quiet family, hyde, hyde park
Jamaica Plain: jp, pond, plain, jamaica plain, jamaica
Leather District: feel privileged, located leather, privileged, leather, leather district
Longwood Medical Area: delivery ),, target star, distance target, restaurant delivery, ), cafe
Mattapan: lower, mattapan square, blue hill, quiet neighborhood, mattapan
Mission Hill: phone number, number hidden, mi phone, mission hill, mission
North End: north end, north, italian, little italy, italy
Roslindale: farmer market, arboretum, village, roslindale village, roslindale
Roxbury: lane, orange line, roxbury, fort, fort hill
South Boston: southie, seaport, beach, broadway, south boston
South Boston Waterfront: contemporary, committee, contemporary art, institute contemporary, seaport
South End: end highlight, copley, end, south, south end
West End: (., (. mi, west end, mi ),, university mi
West Roxbury: guidebook nearby, uncongested, uncongested guidebook, west, west roxbury
```

We used word cloud to better show our result. In word cloud, we selected top 30 words for each neighborhood.



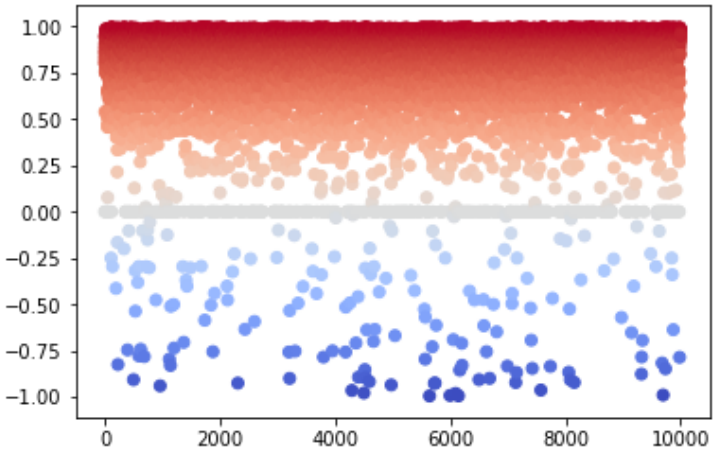
Sentiment Analysis

We tried sentiment analysis on comments to explore user's attitude. First, we joined "*listing_id*", "*latitude*", "*longitude*" from *listings_2019* and "*id*", "*comments*" from *reviews_2019* on "*listing_id*" ("*id*") to generate a new dataset for sentiment analysis. There are 218,319 records in our new data set.

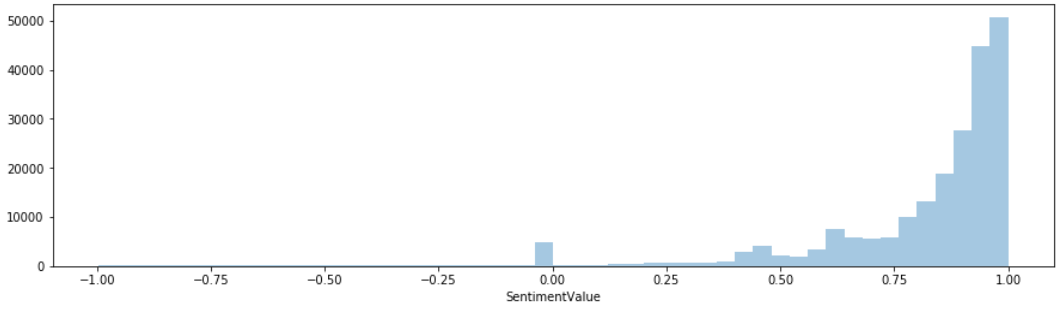
After that, we cleaned the new data set. There are 1,394 comments which contain "automated posting" are auto generated by Airbnb, so we dropped these rows because they do not include any useful information. Also, we dropped rows with missing value in "*comments*".

Before sentiment analysis, we need to make sure that the language of all comments is english. We tried to use ASCII to check characters, but then we realized that some languages have the same character as english. Therefore, we used the package *guess_language* to better check the language. If the language of the comment is not english, we used *Google Cloud Translation API* to translate it into english and then stored it back into data set.

We performed sentiment analysis after finishing those steps, and the output for each comment is a score ranging from -1 to 1. The more the score close to -1, the more negative the comment is; the more the score close to 1, the more positive the comment is. We randomly sampled 10,000 data and drew a scatter plot. We could see that a large proportion of comments are positive.

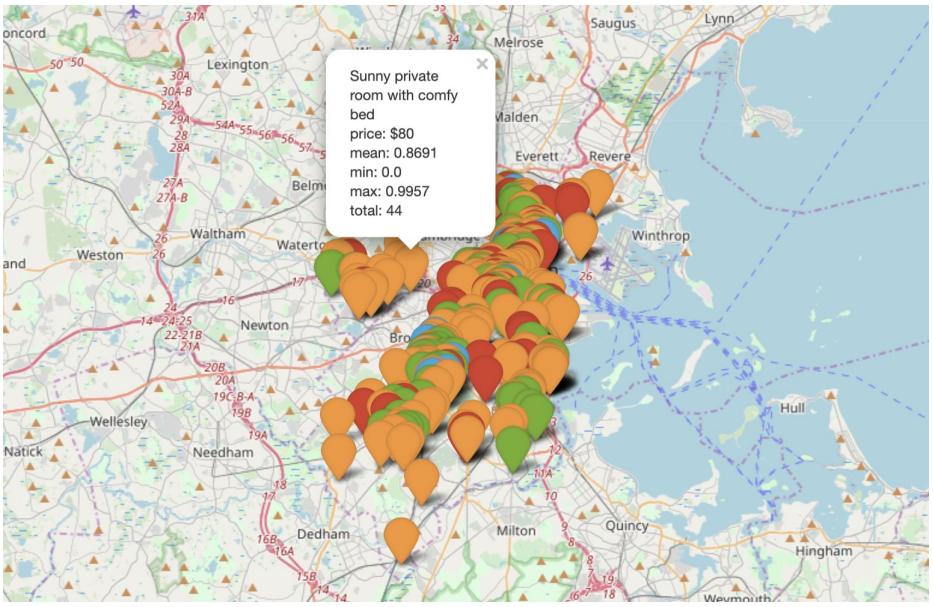


Also, we drew a histogram and calculated quartiles of the sentiment score. The quartiles are ***Q1=0.765000***, ***Q2=0.902000***, ***Q3=0.956500***, which also shows that most people are satisfied with their experience of Airbnb.



Data Visualization

Except for the plots we included in previous sections, we also created an interactive map using our results to help tourists filter listings. We only randomly sampled 300 listings because the html file will be very large if we include all records. We used the quartiles to stratify those listings, and marked it with different colors. Also, when clicking each marker, people could find the name, price and sentiment score and check the exact location on the map. We also generated a map with neighborhood filter. People could choose the neighborhood and check listings information.



Conclusion

In this project, we have done following works:

1) Preprocessed the data sets and generated a new data set for sentiment analysis.
2) Analyzed the current situation of Boston Airbnb using statistics method.
3) Analyzed the vibe of different neighborhood in Boston by machine learning.
4) Performed sentiment analysis for user's reviews and generated the interactive web application.

We could draw some conclusions from our results including:

1) The number of Airbnb listings in Boston from 2018 to 2019 actually decreased, especially in Jamaica Plain.
2) Back Bay and Downtown are the most popular areas in Boston. Also, it is very easy to find an entire apartment listing in these areas.
3) Visiting Boston in spring(March, April) or fall(September, October) is a comparable economical choice for tourists. Also, tourists could avoid Friday and Saturday if possible.
4) Back Bay is a good place for tourists to go shopping. Also, tourists could enjoy Italian food in North End area. More information is included in the word cloud.
5) Most users are satisfied with the quality of Boston Airbnb listings.

Future Work

1. Clean the content of neighborhood overview and remove more useless words, like the neighborhood name, to provide a better description of vibe.
2. Build an interactive web application including all the open data we have.
3. Build a model that can predict the price of Airbnb, which tourists could use this information to see if the listing they found is overpriced.

Reference

[1] https://cloud.google.com/translate/docs/apis
[2] https://github.com/cjhutto/vaderSentiment
[3] https://pypi.org/project/guess_language-spirit/
[4] https://python-visualization.github.io/folium/