BedVal: A Visualization of Airbnb Pricing Factors

[1]

Team 17 Final Report

CSE 6242 Data Visualization and Analytics, Spring 2019

Team Members: Usman Ashraf, Azlan Shah Bin Abdul Jalil,
Jiun-Yu Lee, Jing Shi, He Zhang

[1] The copyright of this picture belongs to Airbnb.

Introduction and Motivation

The sharing economy is experiencing a drastic growth and reshaping the structure of the whole economy. Airbnb is unarguably a representative of the sharing economy. This online sharing platform has infused the features of E-commerce, online advertisements with the prototype of the hospitality industry, and created a new model for the business of accommodation services. In order to obtain a more comprehensive understanding of this model, we are trying to empirically examine some obscured facets of this house-sharing platform.

More precisely, we are trying to answer the following questions:

1. Does the ease of access to public transportation contributes to the listing price or the occupancy rate of a listed property?
2. Do the words with positive sentiments in listing titles have positive impacts on the listing price or occupancy rates?
3. How do the distances from listed properties to nearby famous attractions influence the listing prices and occupancy rates?

In the meantime, to better address those questions, we will try to present those effects, namely the effects from access to public transportation, the effects from listing titles with positive sentiments, and the effects from the distance to local attractions, with dynamic and interactive visualizations.

The ultimate goal of this project is not only to provide theoretical answers to the above questions, but also to help the hosts to determine their strategies of listing, as well as to help the guests to better understand the rationale behind the pricing mechanism of Airbnb. Thereby, at the end of this project, we will also have brief discussions on how to optimize the listing strategy as property owners, and what are the takeaways for future Airbnb users. Those discussions should add practical values to both parties involved in this business.

Literature Survey

There are many published studies which analyze different aspects of Airbnb:

- **Gunter (2018)**

  Gunter investigates factors that make an Airbnb host a superhost, a symbol of ultimate hospitality which encourages users to rent from these hosts.

- **Ke (2017)**

Ke studies the distribution of house types and found that 68.5% of listings are entire homes and identifies a bias towards high ratings and positive reviews.

- **Fang (2016), Mao (2019)**

  Fang uses polynomial regression to model Airbnb listings against tourism industry employment while Mao studied Airbnb effects on local economy. Both confirm a relationship between Airbnb and the local economy and found that its entry benefited the area through job generation.

- **Cheng (2016), Kaker (2018), Ma (2017)**

  Cheng claims that access to digital profiles of hosts results in discrimination around sex and ethnicity. Ma found that high quantity host self-description and specific topic inclusion enhance trustworthiness. Kaker found that hosts being of asian or hispanic descent negatively impact occupancy rates. This research identifies the relationship between host attributes and guest choice, but restricts analysis to host profiles.

- **Liu (2017)**

  Liu examines the effect of advertising appeal and sense of power on click-through and purchase intention and found that different advertising words attract different customer types which can lead to price discrimination. However, Liu doesn't detail how price discrimination occurs.

The above papers give us an awareness of various factors that can affect our research, but they don't explore the relationship between those factors and price/demand in detail. We'll explore more factors which can impact pricing/demand.

- **Cheng (2019), Gibbs (2018), Lawani (2018), Neumann (2017)**

  Cheng found that location, amenities, and host are the key attributes influencing Airbnb users' experiences. Gibbs and Lawani found that these characteristics significantly impact price and show a negative correlation between review and price. Although location's importance was stressed, impacts of access to public transportation and tourist attractions weren't separately analyzed nor do they investigate owners raising price as ratings improve. None of these papers looked at demand impact.

- **Gutiérrez (2017), Li (2016)**

  Gutiérrez analyzed spatial patterns of Barcelona Airbnb against hotels and sightseeing spots. Li proposed a multi-scale clustering algorithm to aggregate homes in similar price zones by distance to attractions and attraction popularity. This research gives insight to potential public transportation and attraction distance effects on price/demand and we can leverage Li's method of clustering.

- **Lee (2015)**

  Lee analyzed the correlation between room sales and social factors, though he didn't analyze the impact on prices. It's worth further exploration before we build our regression model.

- **Perez-Sanchez (2018), Wang (2016), Zhang (2017)**

  Wang identified host, site and property, amenities and services, rental rules, and online review ratings as price determinants in a sharing economy. Perez-Sanchez investigates the relationship between Airbnb accommodation attributes and listing price. Zhang uses a general linear and a geographically weighted regression model to calculate listing price key factors, but is limited to only two location variables, which is not suitable for big cities. We are doing similar analysis of location (Features 1 and 3) and are able to compare against these study methods, but their research doesn't look at the impact on demand as we plan to do.

- **Wen (2009)**

  Wen summarizes theories of factors affecting purchases of travel products. Like Wen, we will be looking at positive sentiment in titles, however we look at its relation with price listings as well as demand. We can utilize Wen's method of extracting wording in our analysis.

Proposed Method

As indicated in our proposal, our main goal is to test: 1, whether the sentiments in the listing title or description has significant impact on the demand or listing price of the property; 2, whether the ease of access to public transportation has significant impacts on the demand or listing price of the property; 3, whether the distance to attractions has significant impact on the demand or listing price of the property.

We plan to combine both the supervised learning and visualization for the purpose of analytics. More specifically, we do the following:

1. Use opinion lexicon to classify the words in the listing title and description into different sentimental groups;
2. Find if there is any relation between commonly used words and the market demand of the housing

3. Calculate the distance from the property to the attractions and clean up the data.
4. Use the transportation map API to pin down the nearest access to local transportation and calculate the distance.
5. Use the attraction map API to pin down the nearest attractions and calculate the distance and the number of attractions within walkable distance.
6. Build up a linear model based upon Wang (2016) and Zhang et al. (2017), and then we run Ordinary Least Square estimation on the model and perform t-tests on the coefficients.
7. Test the predictive power(specifically, predictive MSE) of the model with the variables of interests against the predictive power of the model without those variables by using random forest regression and other Machine Learning techniques.
8. Design an interactive map interface for our regression model that provides airbnb hosts the information about the suggested price and estimated demand for the property they want to host with.
9. Design a choropleth map visualization to help airbnb users finding out the density, average price and demand of the airbnb properties in San Francisco Bay Area.

Innovation of Methodology

- The current works run sentimental analysis only for the reviews and observe its impacts on price while our work also sheds some lights on the demand.
- The current works don't look into the effects from nearby attractions or the access to transportation while we scrape the data for transportation access and nearby attractions, and we take those factors into account.
- Our work expands the sentimental analysis to listing title and description, and we firstly visualize the effects of local attractions on the listing price/demand.
- The current works don't look into the prediction model for best listing price and anticipated number of bookings while we experiment with different predictive models and build up a prediction UI for the hosts.
- The current works mostly do analytics by using supervised learning while we combine the supervised learning method and visualization method to do analytics.

Experiments and Evaluation

Data and preprocessing

- **Airbnb Listing:** Our airbnb listing data is free from Inside Airbnb, which contains most of the attributes the past model are used and that we are going to analyze. The following two preprocessing are done to further extract necessary attributes from this data.

  - We use the opinion lexicon in the Natural Language Processing Toolkit (NLTK) module to do sentiment analysis and abstract the numbers of the words with positive sentiments and negative sentiments in both the listing titles and the listing descriptions.
  - Since there is no attribute for demand in the listing data, we used an approach proposed by Lee (2015) to estimate the monthly booking of a property based in the number of rating.
- **Transportation Data:** We collected BART, SFMTA, and Caltrain data from TransitWiki, which has a collection of public transportation data in the format of General Transit Feed Specification (GTFS). The following features with respect to each airbnb property are extracted for our analysis:
  - **Top1:** The distance to the nearest transportation station.
  - **Top5 Average:** The average of the distance of top 5 nearest transportation location.
  - **Within Radius Count:** The number of stations within a walkable distance radius (0.7 miles.)
  - **With Radius Average:** The average distance of the stations within the walkable distance.
  - **Total Average:** The average of the distance to all the transportation station.
- **Attraction Data:** For attraction data, we use HERE, a public attraction API to retrieve the nearest attractions for each Airbnb property. Using only the Free API, we were able to obtain at most 100 most popular attractions (based on HERE ranking) sorted by distance to each Airbnb property. The original raw data was very large as each attraction entry contains their name, category, distance to Airbnb property, url and opening hours. Since most Airbnb entry will have similar nearby attractions, we can minimize the data by only match the Airbnb entry with the attraction ID. Separating the files greatly reduce the total memory size. Then we calculate the same features as we did for transportation data.
- **San Francisco Map:** For the the map of San Francisco, the data was obtained from Census Bureau website. They provide various US .shp boundary files and for out project, we took the ZipCode Tabulated Area (ZCTA). The raw map data was in .shp format and the following steps were taken to get our final data structure.
  - The .shp file was converted to GeoJSON format using shp2json and filtered to only include the ZCTA in San Francisco.
  - The GeoJSON map was merged with the AirBnB listing data where the coordinates for each listing will correspond to a Point in the GeoJSON data.
  - The density value was added to each ZCTA entry as a property by doing another merge.

- The merged GeoJSON data was projected and rotated according to the Albers projection.
- Then the projected GeoJSON data was converted to TopoJSON format using geo2topo to minimize the file size. This reduce the file size by about 80% by converting the coordinate floating points into integers.
- Finally Map Shapper was used to verify our TopoJSON data. It will show two Features where one is the ZCTA boundary map, and the other is the Point of all the AirBnB listings.

In this project, we restricted ourselves to only analyze the data in San Francisco, since we believe that data from different cities will be influenced by other factors such as the housing price. Due to this restriction and the natural of Airbnb, we are not able to obtain a tremendous amount of data like we usually want to for a data science project. However, our experiment still shows a reasonable result and we believe that our application can be easily scaled up to different cities.

Experiments Conducted

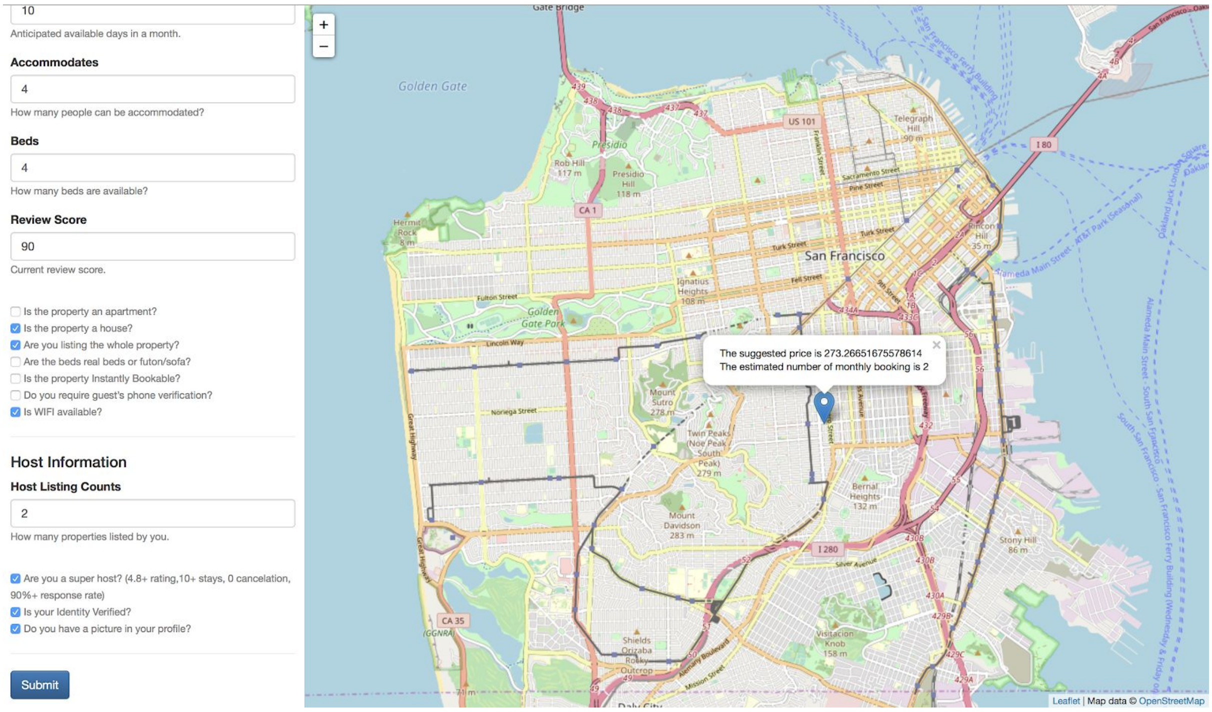
1. To analyze and evaluate effect of positive sentiments, we add positive word count variables we extracted to the baseline model built by Wang (2016) and Zhang et al. (2017) and run a OLS regression on it. The results are appended in Appendix 2.
2. We perform linear regression on the transportation we extracted in the previous section and the results are appended in Appendix.
3. We use the attraction features to perform another linear regression and perform t-test on the coefficients. The results are shown in the Appendix.
4. We visualized the most commonly used words in the ‘description’ using a word cloud.Then we run a OLS regression model based on select words to find how well-related are our selected words and the demand for a housing.
5. We conducted a data science experiment to compare the model built by other researchers in previous literature. More precisely, we divided the predictors into training and testing datasets, and then we used the training dataset to train the linear regression model, random forest regression model and the shallow neural network regression model. We measured the predictability of the model by out-of-sample Mean squared error(MSE). Then we compared the MSE values of the model built by previous researchers and the MSE values of the model built by us. Results are shown in Appendix 2.
6. We analysed how well some commonly used words are related to the demand of a listing. The results are attached in the appendix.
7. We also compared the out-of-sample MSE vales of the linear regression model, the random forest regression model and the shallow neural network model to determine which model is the best for predictions of price and demand.

8. We tuned the parameters of the random forest regression model by using the Grid Search Cross-validation technique. And then we build up the prediction interface based upon the tuned model.

Evaluation and Results

- **Supervised Learning & Perdiction model analysis**
    1. More positive words in the listing title or description don't significantly affect the listing price. However, more positive words in the listing title and more neutral words in the listing description does increase the demand(Estimated number of bookings per month).
    2. The more stations within the walkable distance there are, the higher the demand will be. However, transportation access doesn't seem to have huge impacts on the pricing decision made by the hosts.
    3. The more attractions within walkable distance there are, the higher the demand will be. However, attractions don't have significant impacts on listing price.
    4. From the word cloud we got, we realized that words with the highest frequency in commonly used words list are trivial words. They included words such as "San Francisco", "bed", "private" and "bathroom". None of these words told us much about the listing. Therefore, we then hand-picked 18 words from the list of 100 most common words, these words were descriptive words that tell us more about the housing. They included words as "furnished", "spacious" and "garden".
    5. Commonly used words being tested have a fair connection $(R^2 = 0.56)$ with the demand of a housing. This means most hosts know what words can make their properties attractive.
    6. We found that no matter which model we looked at, our model are better than the original model built by previous researchers. The MSE is reduced by 3%-12%.
    7. Based upon the experiment results in 4, we also identified that, judging with the MSE values, for the price prediction, linear regression will be the best model. For the demand prediction, random forest regression is the best.
- **Visualization and Interface**

Based on the model we have built up, we designed interfaces and visualizations for both airbnb hosts and users to get a better understanding about the Airbnb economic.



10

Anticipated available days in a month.

Accommodates

4

How many people can be accommodated?

Beds

4

How many beds are available?

Review Score

90

Current review score.

□ Is the property an apartment?
☑ Is the property a house?
☑ Are you listing the whole property?
□ Are the beds real beds or futon/sofa?
□ Is the property Instantly Bookable?
□ Do you require guest's phone verification?
☑ Is WIFI available?

Host Information

Host Listing Counts

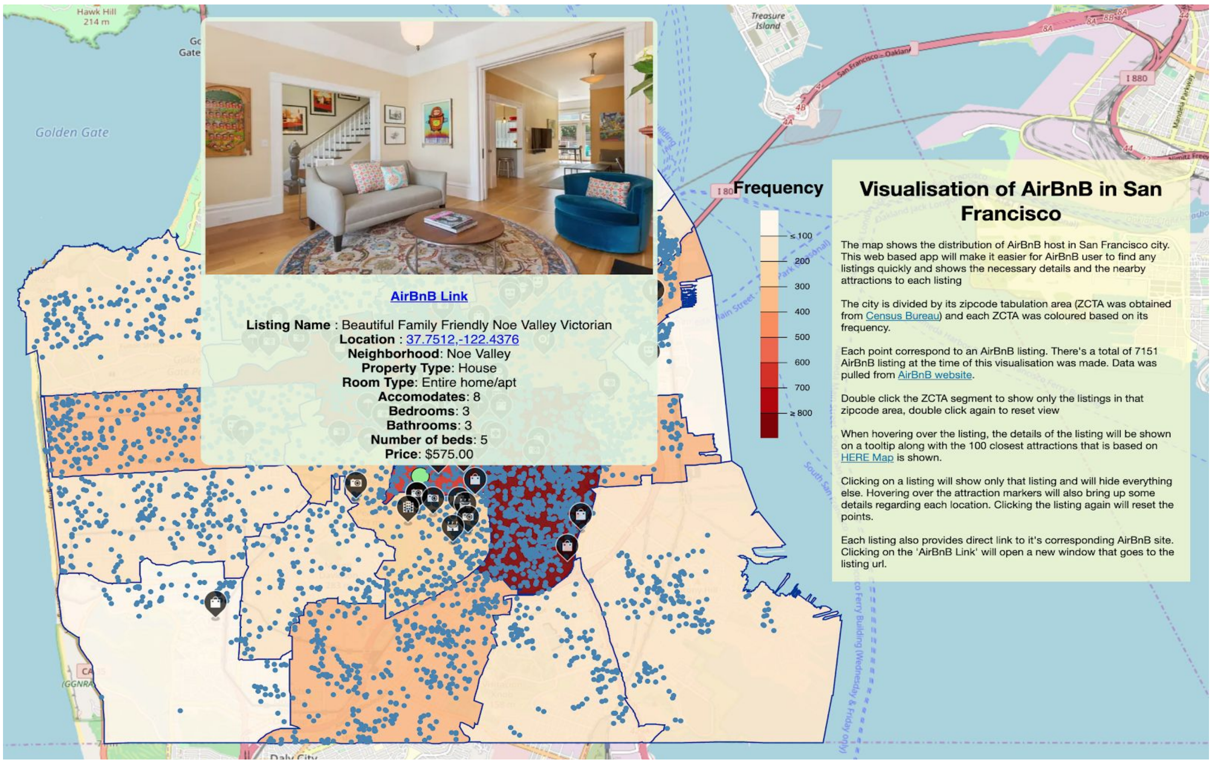
2

How many properties listed by you.

☑ Are you a super host? (4.8+ rating,10+ stays, 0 cancelation, 90%+ response rate)
☑ Is your Identity Verified?
☑ Do you have a picture in your profile?

The suggested price is 273.26651675578614
The estimated number of monthly booking is 2

For Airbnb hosts, we provide an interactive map application to locate the potential property they want to host on Airbnb and discover its suggested price and estimated demand using our regression model. Hosts can first pin down the location where they want to host an Airbnb on the map. We also ask hosts to fill out the form on the left to provide us more information about the property status to ensure the quality of regression. After hosts complete the form and submit, the property location and status will be sent back to our back end server. We then use our regression model to determine the suggested price for this property and the estimated monthly demand booking. Finally, the results will be sent back and presented as the popup on the pin shows.



AirBnB Link

Listing Name : Beautiful Family Friendly Noe Valley Victorian
Location : 37.7512,-122.4376
Neighborhood: Noe Valley
Property Type: House
Room Type: Entire home/apt
Accomodates: 8
Bedrooms: 3
Bathrooms: 3
Number of beds: 5
Price: $575.00

Visualisation of AirBnB in San Francisco

The map shows the distribution of AirBnB host in San Francisco city. This web based app will make it easier for AirBnB user to find any listings quickly and shows the necessary details and the nearby attractions to each listing

The city is divided by its zipcode tabulation area (ZCTA was obtained from Census Bureau) and each ZCTA was coloured based on its frequency.

Each point correspond to an AirBnB listing. There's a total of 7151 AirBnB listing at the time of this visualisation was made. Data was pulled from AirBnB website.

Double click the ZCTA segment to show only the listings in that zipcode area, double click again to reset view

When hovering over the listing, the details of the listing will be shown on a tooltip along with the 100 closest attractions that is based on HERE Map is shown.

Clicking on a listing will show only that listing and will hide everything else. Hovering over the attraction markers will also bring up some details regarding each location. Clicking the listing again will reset the points.

Each listing also provides direct link to it's corresponding AirBnB site. Clicking on the 'AirBnB Link' will open a new window that goes to the listing url.

For Airbnb user, we provide a quick and easy way to visualize all the available AirBnB listings in San Francisco. At a glance, user will immediately see the distribution of the listings all over the city. The city map is divided according to the ZipCode Tabulation Area (ZCTA) and each area is coloured according to its listing frequency. User can determine which general area have the highest listing densities.

Each point on the map corresponds to a valid Airbnb entry and user can quickly see the necessary information by hovering the cursor on it. Details such as price, number of bedrooms/bathrooms, number of beds and even a picture will accompany for every listed entry. If user is interested in the entry, they can click on the AirBnB Link in the provided tooltip to be taken directly to the Airbnb website for that specific entry. Clicking on the coordinates will open Google Map that will show the direction to that listing address.

Another feature for the user is the "nearby attractions". The HERE API has a built in rating that they sort the attractions based on popularity. By using this, we can obtain at most 100 most popular attractions that is nearby to the Airbnb property. By hovering the cursor on the property, the nearby attractions will be immediately visible on the map by a unique marker that was also obtained from HERE database. User can click on a property to hide all the other entries and focus on the nearby attractions. By hovering on the unique marker, user can get details on the attraction such as the distance, opening hours and also a hyperlink that will open Google Map to that attraction from the Airbnb property.

The map will also adjust accordingly when zooming in. The coordinates will be displayed bigger if zoomed in more. The legend, svg map color and description will be hidden when zoom in as well.

Conclusions

The primary conclusions that we have until this point include the following:

1. More positive words in the listing title and more neutral words in the description can significantly increase the demand of the property drastically.
2. Easy access to public transportation can improve the demand of the property tremendously.
3. With sentiment analysis and transportation information, the predictions for demand and price can be more precise.
4. For price prediction, the linear regression model gives the best precision, while for demand precision, the random forest regression model is the best.
5. Commonly used descriptive word such as beautiful, quiet, comfortable etc helps get more demand for housing.

Work Distribution

Member Name	Works Done
Azlan Shah Bin Abdul Jalil	Converted the .shp file and coordinates from the Airbnb database into a TopoJSON format. Extracted nearby attractions from HERE API for each Airbnb entry. Built the web-based user visualisation using Leaflet and D3.
Jing Shi	Built supervised learning models for the relationship between price/demand and distance to public transportation; computed statistics for the price/demand of Airbnb properties in each zip code tabulation area; prepared choropleth maps to show geographical distributions for those statistics.
Jiun-Yu Lee	Collected transportation data from multiple sources, cleaned and extracted transportation features for Airbnb listing data, built up a backend server with Restful API to host our prediction model, designed and built up the host frontend interface, integrated frontend and backend.
Usman Ashraf	Cleaned description data, found most commonly used words in description, built word cloud of commonly used words, analysed the impact commonly used words have on the demand, ran supervised learning models to see the relation between those words and demand, found number of housings in each zipcode
He Zhang	Cleaned the listing data, merged the transportation data and the attraction data with the listing data, conducted the sentiment analysis, conducted supervised learning experiments, built up the back-end prediction function for the host interface, and participated in the design of the host interface.

Special notes: **1, One of the member dropped the class and quit the group so we have to redistribute the work; 2, All team members have contributed similar amount of effort.**

Appendix (Experiment results):

```
                    OLS Regression Results
==============================================================================
Dep. Variable:              price   R-squared:                  0.178
Model:                        OLS   Adj. R-squared:             0.177
Method:             Least Squares   F-statistic:                171.9
Date:            Tue, 26 Mar 2019   Prob (F-statistic):          0.00
Time:                    21:44:40   Log-Likelihood:           -85578.
No. Observations:           12711   AIC:                    1.712e+05
Df Residuals:               12694   BIC:                    1.713e+05
Df Model:                      16
Covariance Type:        nonrobust
==================================================================================================
==
                              coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                           -185.0170     43.750     -4.229      0.000    -270.774     -99.260
host_is_superhost               -13.0363      3.841     -3.394      0.001     -20.566      -5.506
host_listings_count              0.2489      0.015     16.310      0.000       0.219       0.279
host_has_profile_pic            -30.8018     30.204     -1.020      0.308     -90.007      28.404
host_identity_verified           11.8035      3.767      3.134      0.002       4.420      19.187
apt                             -16.1063      4.316     -3.731      0.000     -24.567      -7.646
house                            26.0475      4.863      5.356      0.000      16.515      35.580
entire                           46.2175      4.579     10.094      0.000      37.242      55.193
accommodates                     32.7834      1.572     20.860      0.000      29.703      35.864
beds                              0.8318      2.296      0.362      0.717      -3.668       5.331
realbed                           6.5336     14.590      0.448      0.654     -22.066      35.133
WIFI                             -2.2435     15.266     -0.147      0.883     -32.167      27.680
review_scores_rating              2.4418      0.271      9.021      0.000       1.911       2.972
instant_bookable                -13.0199      3.748     -3.474      0.001     -20.366      -5.674
require_guest_phone_verification     17.1626      8.333      2.059      0.039       0.828      33.497
namepos                           2.1979      2.359      0.932      0.352      -2.426       6.822
despos                            0.5312      0.454      1.169      0.242      -0.359       1.422
==============================================================================
Omnibus:                    27042.839   Durbin-Watson:                1.818
```

```
Prob(Omnibus):              0.000   Jarque-Bera (JB):       215298487.103
Skew:                      18.432   Prob(JB):                    0.00
Kurtosis:                 639.516   Cond. No.                3.43e+03
==============================================================================
```

Table1. Sentimental analysis results on price

```
                OLS Regression Results
==============================================================================
Dep. Variable:       Monthly_demand   R-squared:                 0.185
Model:                          OLS   Adj. R-squared:            0.184
Method:               Least Squares   F-statistic:               179.5
Date:              Tue, 26 Mar 2019   Prob (F-statistic):         0.00
Time:                      21:44:40   Log-Likelihood:          -32261.
No. Observations:             12711   AIC:                   6.456e+04
Df Residuals:                 12694   BIC:                   6.468e+04
Df Model:                        16
Covariance Type:          nonrobust
====================================================================================
==
                          coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const                      -0.8142      0.663     -1.227      0.220      -2.114       0.486
host_is_superhost           1.9792      0.058     34.103      0.000       1.865       2.093
host_listings_count        -0.0031      0.000    -13.566      0.000      -0.004      -0.003
host_has_profile_pic        0.2468      0.455      0.542      0.588      -0.645       1.139
host_identity_verified     -0.7076      0.057    -12.454      0.000      -0.819      -0.596
apt                        -0.0804      0.065     -1.234      0.217      -0.208       0.047
house                      -0.2811      0.073     -3.834      0.000      -0.425      -0.137
entire                     -0.4211      0.069     -6.101      0.000      -0.556      -0.286
accommodates                0.0681      0.024      2.876      0.004       0.022       0.115
beds                       -0.1326      0.035     -3.832      0.000      -0.200      -0.065
realbed                     0.1280      0.220      0.582      0.561      -0.303       0.559
WIFI                        0.9652      0.231      4.183      0.000       0.513       1.418
review_scores_rating        0.0072      0.004      1.769      0.077      -0.001       0.015
instant_bookable            1.3017      0.056     23.042      0.000       1.191       1.412
require_guest_phone_verification  -0.8841  0.126  -7.042      0.000      -1.130      -0.638
namepos                     0.0645      0.035      1.840      0.066      -0.004       0.133
desneu                      0.0096      0.001     15.782      0.000       0.008       0.011
==============================================================================
Omnibus:                 3602.378   Durbin-Watson:                1.826
Prob(Omnibus):              0.000   Jarque-Bera (JB):         10292.022
Skew:                       1.493   Prob(JB):                      0.00
Kurtosis:                   6.242   Cond. No.                  5.24e+03
==============================================================================
```

Table 2. Sentimental analysis results on demand

Regression results of distance to public transportation to price

```
                OLS Regression Results
==============================================================================
Dep. Variable:                price   R-squared:                 0.001
Model:                          OLS   Adj. R-squared:            0.001
Method:               Least Squares   F-statistic:               2.328
Date:              Wed, 27 Mar 2019   Prob (F-statistic):       0.0539
Time:                      23:50:21   Log-Likelihood:          -51239.
No. Observations:              7151   AIC:                   1.025e+05
Df Residuals:                  7146   BIC:                   1.025e+05
Df Model:                         4
Covariance Type:          nonrobust
=================================================================================
                coef    std err          t      P>|t|      [0.025      0.975]
```

```
---------------------------------------------------------------------------------
const           186.8586   83.347    2.242    0.025    23.473   350.244
top1            48.6577   192.741    0.252    0.801   -329.172   426.487
top5_avg          -6.8737   205.864   -0.033    0.973   -410.429   396.681
total_avg        -17.3856    7.977   -2.179    0.029   -33.023    -1.748
within_radius_avg  192.4644  160.196    1.201    0.230   -121.567   506.495
==============================================================================
Omnibus:                14285.385   Durbin-Watson:             1.912
Prob(Omnibus):                0.000   Jarque-Bera (JB):       47679385.777
Skew:                    16.149   Prob(JB):                  0.00
Kurtosis:                401.719   Cond. No.                   289.
==============================================================================
```

Regression results of distance to public transportation to demand

```
OLS Regression Results
==============================================================================
Dep. Variable:       Monthly_demand   R-squared:                 0.011
Model:                     OLS   Adj. R-squared:             0.010
Method:              Least Squares   F-statistic:               19.31
Date:            Wed, 27 Mar 2019   Prob (F-statistic):        8.07e-16
Time:                 23:50:21   Log-Likelihood:            -18488.
No. Observations:            7151   AIC:                   3.699e+04
Df Residuals:             7146   BIC:                  3.702e+04
Df Model:                 4
Covariance Type:          nonrobust
=====================================================================================
              coef    std err        t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const            3.6186     0.855     4.234    0.000     1.943     5.294
top1            -5.7306     1.977    -2.899    0.004    -9.605    -1.856
top5_avg         10.5456     2.111     4.995    0.000     6.407    14.684
total_avg         0.3340     0.082     4.083    0.000     0.174     0.494
within_radius_avg   -6.5138     1.643    -3.965    0.000    -9.734    -3.293
==============================================================================
Omnibus:                 2324.053   Durbin-Watson:             1.812
Prob(Omnibus):                0.000   Jarque-Bera (JB):          6594.569
Skew:                     1.733   Prob(JB):                  0.00
Kurtosis:                  6.182   Cond. No.                  289.
==============================================================================
```

Regression results of effects from attractions on price

```
                    OLS Regression Results
==============================================================================
Dep. Variable:                price   R-squared:                 0.199
Model:                     OLS   Adj. R-squared:             0.197
Method:              Least Squares   F-statistic:               94.67
Date:            Fri, 19 Apr 2019   Prob (F-statistic):        1.77e-261
Time:                 11:31:41   Log-Likelihood:            -39773.
No. Observations:            5726   AIC:                   7.958e+04
Df Residuals:             5710   BIC:                  7.968e+04
Df Model:                 15
Covariance Type:          nonrobust
==========================================================================================
==
                          coef    std err        t      P>|t|      [0.025      0.975]
```

```
-----------------------------------------------------------------------------------------------
const                  -244.5911   79.691    -3.069     0.002   -400.815    -88.367
host_is_superhost            -5.2032     7.037    -0.739     0.460    -18.998      8.591
host_listings_count          -0.0397     0.028    -1.416     0.157     -0.095      0.015
host_has_profile_pic         -13.0887     48.332    -0.271     0.787    -107.838     81.660
host_identity_verified        21.0500     7.027     2.996     0.003      7.275     34.825
apt                    -34.8358     8.159    -4.269     0.000    -50.831    -18.840
house                   -19.4635     8.910    -2.185     0.029    -36.930     -1.997
entire                  38.4545     8.159     4.713     0.000     22.460     54.449
accommodates                 57.9966     3.362    17.248     0.000     51.405     64.588
beds                    -1.2307     5.157    -0.239     0.811    -11.340      8.879
realbed                  14.9862    30.829     0.486     0.627    -45.451     75.423
WIFI                    10.3824    31.198     0.333     0.739    -50.778     71.542
review_scores_rating           2.5370     0.503     5.046     0.000      1.551      3.523
instant_bookable              -2.3328     7.204    -0.324     0.746    -16.455     11.789
require_guest_phone_verification   -2.0561    13.493    -0.152     0.879    -28.507     24.395
within_radius_count_attr          0.2091     0.278     0.753     0.452     -0.335      0.754
==============================================================================
Omnibus:                 12363.905   Durbin-Watson:               1.925
Prob(Omnibus):                 0.000   Jarque-Bera (JB):         73077471.821
Skew:                    19.096   Prob(JB):                     0.00
Kurtosis:                555.122   Cond. No.                 3.35e+03
==============================================================================
```

Regression results of effects from attractions on demand

```
                       OLS Regression Results
==============================================================================
Dep. Variable:          Monthly_demand   R-squared:                   0.182
Model:                      OLS   Adj. R-squared:              0.180
Method:             Least Squares   F-statistic:                 84.63
Date:            Fri, 19 Apr 2019   Prob (F-statistic):          3.06e-235
Time:                  11:31:41   Log-Likelihood:              -14439.
No. Observations:             5726   AIC:                    2.891e+04
Df Residuals:               5710   BIC:                    2.902e+04
Df Model:                  15
Covariance Type:          nonrobust
==================================================================================================
==
                         coef    std err        t     P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                    1.4160     0.955     1.483     0.138     -0.456      3.288
host_is_superhost             1.7281     0.084    20.498     0.000      1.563      1.893
host_listings_count          -0.0034     0.000   -10.187     0.000     -0.004     -0.003
host_has_profile_pic          -0.2269     0.579    -0.392     0.695     -1.362      0.908
host_identity_verified        -0.5856     0.084    -6.956     0.000     -0.751     -0.421
apt                     -0.4811     0.098    -4.922     0.000     -0.673     -0.289
house                    -0.0995     0.107    -0.932     0.351     -0.309      0.110
entire                  -0.9167     0.098    -9.378     0.000     -1.108     -0.725
accommodates                  0.0821     0.040     2.038     0.042      0.003      0.161
beds                    -0.2931     0.062    -4.743     0.000     -0.414     -0.172
realbed                  -0.2500     0.369    -0.677     0.498     -0.974      0.474
WIFI                     1.4779     0.374     3.954     0.000      0.745      2.211
review_scores_rating           0.0128     0.006     2.130     0.033      0.001      0.025
instant_bookable               1.1391     0.086    13.198     0.000      0.970      1.308
require_guest_phone_verification   -1.0047     0.162    -6.215     0.000     -1.322     -0.688
within_radius_count_attr         -0.0119     0.003    -3.571     0.000     -0.018     -0.005
==============================================================================
Omnibus:                  1550.043   Durbin-Watson:               1.823
Prob(Omnibus):                 0.000   Jarque-Bera (JB):          4059.340
Skew:                     1.458   Prob(JB):                     0.00
Kurtosis:                  5.918   Cond. No.                 3.35e+03
==============================================================================
```

Using the commonly used descriptive words to model the demand of a housing in the next 90 days.

```
list_of_words_used = ['large', 'kitchen', 'restaurants', 'garden', 'great', 'quiet',\
                      'transportation', 'view', 'comfortable', 'furnished', 'spacious',\
                      'beautiful', 'easy', 'enjoy', 'sunny', 'downtown', 'perfect', 'best']
```

```
                      OLS Regression Results
==============================================================================
===
Dep. Variable:          demand_90     R-squared:                0.560
Model:                    OLS          Adj. R-squared:              0.559
Method:             Least Squares     F-statistic:              504.2
Date:             Sun, 21 Apr 2019     Prob (F-statistic):          0.00
Time:                   03:54:50      Log-Likelihood:            -36707.
No. Observations:            7151      AIC:                  7.345e+04
Df Residuals:              7133       BIC:                 7.357e+04
Df Model:                  18
Covariance Type:          nonrobust
==============================================================================
```

Comparisons of the predictability across models

	Price		Demand	
	New model MSE	Baseline model MSE	New model MSE	Baseline model MSE
Linear regression	52761.37	54425.80	8.64	9.20
Random forest regression	59651.82	69206.61	6.89	8.69
Shallow neural network	54096.71	54557.81	8.69	10.96

Citations

Chen, Yong. (2017). “Consumer evaluation of Airbnb listings. A hedonic pricing approach” https://www.emeraldinsight.com/doi/full/10.1108/IJCHM-10-2016-0606#_i29

Cheng, M., & Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*, *76*, 58-70.

Fang, Bin, Qiang Ye, and Rob Law. "Effect of sharing economy on tourism industry employment." *Annals of Tourism Research*(double blind peer review) 57.3 (2016): 264-267.

Gunter, U. (2018). “What makes an Airbnb host a superhost? Empirical evidence from San Francisco and the Bay Area.”

https://www.sciencedirect.com/science/article/pii/S026151771730242X?via%3Dihub

Gutiérrez, Javier, et al. "The eruption of Airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in Barcelona." *Tourism Management* 62 (2017): 278-291.

Kakar, Venoo, et al. "The visible host: Does race guide Airbnb rental rates in San Francisco?." *Journal of Housing Economics(Single blind peer review)* 40 (2018): 25-40.

Ke, Q. (2017). "Sharing Means Renting?: An Entire-marketplace Analysis of Airbnb" https://arxiv.org/pdf/1701.01645.pdf

Lee, Donghun, et al. "An analysis of social features associated with room sales of Airbnb." *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*. ACM, 2015.

Lawani, A. (2018). "Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston" https://www.sciencedirect.com/science/article/pii/S016604621730340X

Li, Yang, et al. "Reasonable price recommendation on Airbnb using Multi-Scale clustering." *2016 35th Chinese Control Conference (CCC)*. IEEE, 2016.

Liu, Stephanie Q., and Anna S. Mattila. "Airbnb: Online targeted advertising, sense of power, and consumer decisions." *International Journal of Hospitality Management* (double blind peer review) 60 (2017): 33-41.

Ma, Xiao, et al. "Self-disclosure and perceived trustworthiness of Airbnb host profiles." *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 2017.

Mao, Yifei, et al. (2019). "Real Effects of Peer-to-Peer Rental: Evidence from Airbnb" https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3111975

Neumann, Jürgen, et al. (2017). "Theory and Empirical Evidence for Optimal Pricing Conditional on Online Ratings" https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1064&context=ecis2017_rp

Perez-Sanchez, V. (2018). "The What, Where, and Why of Airbnb Price Determinants" https://www.mdpi.com/2071-1050/10/12/4596/htm

Wang, D. (2016). “Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com” https://www.sciencedirect.com/science/article/pii/S0278431916305618

Wen, I. (2009). Factors affecting the online travel buying decision: a review. *International Journal of Contemporary Hospitality Management*, *21*(6), 752-765.

Zhang, Z., Chen, R., Han, L., & Yang, L. (2017). Key factors affecting the price of Airbnb listings: A geographically weighted approach. *Sustainability*, *9*(9), 1635.