Analysis of Airbnb house pricing in NYC & LA

Final Project in Machine Learning for Cities

Yue Jin (yj1672)

Yushu Rao (yr872)

Chenjie Su (cs5998)

Kunru Lu (kl3743)

Github:https://github.com/carajumpshigh/Analysis_of_Airbnb_house_pricing_in_NYC_and_LA

Abstract

Airbnb is an online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences since 2008. NYC and LA are two of the most populous cities in the United States and also one of the most popular tourism and business places in the world. This research compared Airbnb listings in NYC & Los Angeles and analyzed the factors that affect pricing in the two cities. A series of machine learning models including linear regression, random forest, bagging, extra trees, and gradient boost have been used to forecast pricing. These models will help city officials with valuable information on the characters of the two cities and their neighborhoods for better city tourism development and management. The comparison between NYC and LA also illustrates innovative insight to help people get to know the differences between the two cities.

Introduction

Airbnb is one of the world’s largest marketplaces for unique, authentic places to stay and things to do, offering over 7 million accommodations, all powered by local hosts. It was founded in 2007 by two university graduates with the concept to save money. Currently AirBnB is accessible in 62 languages across 220+ countries and regions, with high concentration in Paris(59,881), London, United Kingdom (87,235), New York (50,378), LA(39,486).

This report seeks to find out the key factors including house conditions and demographics of the neighborhood that affect Airbnb pricings in NYC and LA and their neighborhoods to have a deep understanding of urban systems, inner interactions and city differences, to identify long-term trends and seasonal changes of Airbnb house pricing through time series analysis, offering insights of the short-term rent market of NYC & LA to guide related policy decision and for city planners to bring up effective and actionable policy decisions on how to drive local community tourism development.

Literature Review

Erudite studies have been researching evaluating the impact of Airbnb to the local community. Gurran, N., & Phibbs, P. (2017) take Sydney, Australia as an experient city, address the questions how effective existing planning controls on tourist and residential accommodations are and the extent to which increasing tourism demand puts pressure on the local housing market. Dudás, G., Vida, G., Kovalcsik, T., & Boros, L. (2017) use correlation, regression analysis to determine the socio-economic conditions of areas and reveal those factors that may affect the spatial distribution of Airbnb listings in the city. Coles, P. A., Egesdal, M., Ellen, I. G., Li, X., & Sundararajan, A. (2017). Identify the actual Airbnb usage patterns and how they vary across neighborhoods in New York City through empirical analysis. Sans, A. A., & Quaglieri, A. (2016). focuses on Airbnb in Barcelona, measures Airbnb's contribution to the revitalization of neighbourhoods and evaluates the urban impacts of Airbnb's business model. Davis, P. (2016) examines the ways in which uber and Airbnb have grown since their founding in terms of domestic and international market expansion and hiring for primary and support functions. Chatterjee, D., Dandona, B., Mitra, A., & Giri, M. (2019) seeks to understand Indian tourists' perceptions of Airbnb compared to other hospitality options, and the factors driving their purchase intentions.

All of these publications focus on the impact of Airbnb to the local community, but none of them research the comparison of Airbnb in different cities and less are involved with robust machine learning models. Our research focuses on NYC & LA, compare their differences and conduct sentiment analysis on the reviews to evaluate the reasonableness of Airbnb house pricing for better results.

Data

- **Airbnb Open Data**

  The Airbnb open data include listings, reviews and calendar. The Airbnb listing data including listing geological information, housing condition, pricing and customer review for each listing data, are scraped from inside airbnb open data[1]. There are 51097 rows of listings in NYC and 38851 in LA scraped by February 13th, 2020 (before COVID-19 widely recognized). The reviews data includes detailed contents, customer id and related listing information. The calendar data consists of the prices and maximum and minimum nights to stay for each listing throughout the year.

- **Census Data**

  For further research, we collect census data for a better conclusion of how some major social components will influence the pricing of Airbnb. For the income and education indicators, we collect income per capita and percentage of population over 25 years old with education over high school in each zip code of NYC in 2018 from U.S. Census ACS[2], and collect median household income and calculate percentage of population over 25 years old with education over high school in each council district of LA in 2018 from Los Angeles Office of the Controller[3]. For the crime indicator, we calculate the counts of arrests in 2019 in each zip code of NYC and LA based on the arrest data from New York Police Department (NYPD)[4] and Los Angeles Police Department (LAPD)[5].

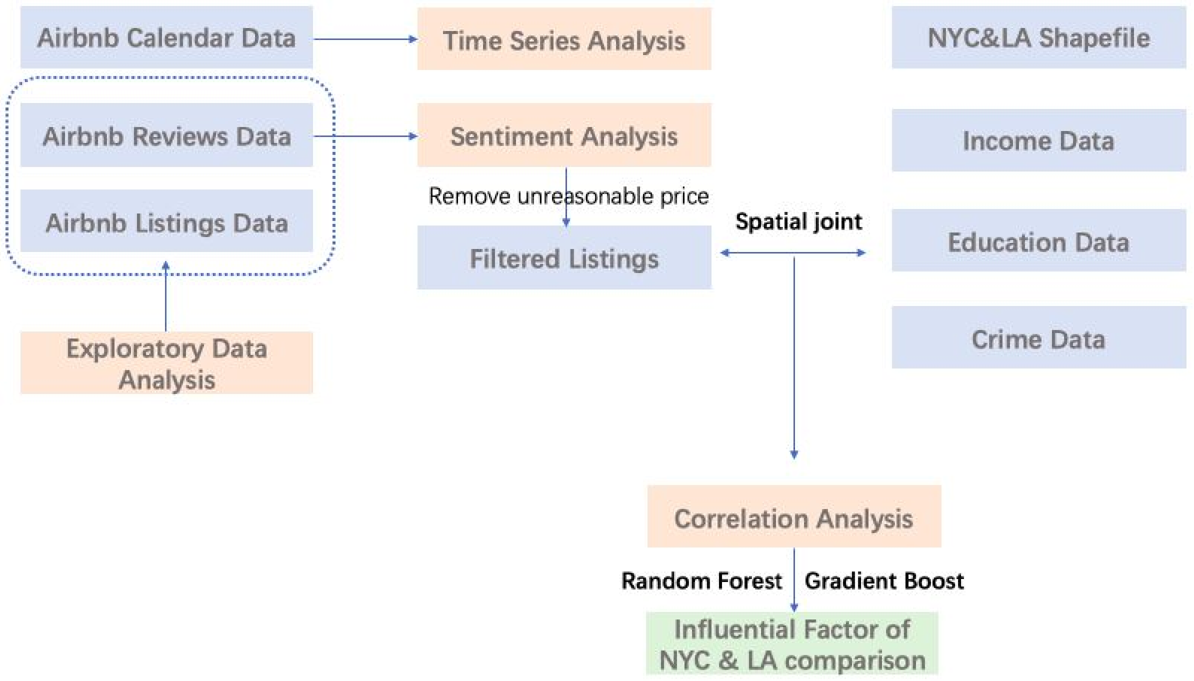
- **Spatial Data**

  To involve neighborhood indicators in the correlation analysis, boundaries of zip codes of New York City and Los Angeles from U.S. Census[6] and council districts of Los Angeles from the Bureau of Engineering[7] are collected.

- **Others**

  We also use US federal holidays data from 2011 to 2020[8] from the U.S. Office of Personnel Management (OPM) for time series analysis.

Methodology



The methodology of our project is as shown in the flow chart above. For airbnb calendar data, we conduct time series analysis to compare the long-term trend and seasonal changes of Airbnb prices in NYC and LA. For reviews and listings data, we firstly do exploratory data analysis to summarize the main characteristics of Airbnb prices in the two cities. Then, we conduct sentiment analysis on the reviews for the customers' general impressions on each listing house to determine whether the house is reasonably priced or not. After that, we filter out the houses with unreasonable prices based on the result of sentiment analysis, and spatial join the filtered listings with local income, education and crime indicators. Finally, we do correlation analysis using multiple regression models, including Linear Regression, Random Forest, Bagging, Extra Tree and Gradient Boost, for LA dataset and Random Forest classification for NYC dataset to get the influential factors of LA and NYC for further comparison.

Data Preparation & EDA

In the data cleaning stage, replacing missing values(NaN) with 0 and removing $ in price fields. Before building the model, exploratory data analysis is carried out to learn more about the data. From Listing Count by Borough (Appendix B, Figure 1), it is easy to see that in NYC, the most listings are located in Manhattan, and Brooklyn is slightly lower than Manhattan, while in LA, most listings are within the city of LA. The top 5 popular neighbourhoods in NYC and LA (Appendix B, Figure 2) account for 29.7% and 21.6% for each city respectively. To have a better understanding of listing pricing and space, the visualized heatmap (Appendix B, Figure 3,4) presents the price in each spot. It is clear to identify that airbnb housing lists converge in Manhattan, and there are several extremely high list pricings in Manhattan. The majority of the listings are in the city of LA, while the rest are located along the beach which is reasonable since the tourist may prefer sea view in their housing. Also, in the city of LA, there are multiple extremely higher pricings. The airbnb room type distribution (Appendix B, Figure 5) is similar in each city. The entire home/apartment and private room account for the majority of the total type. It can be clearly seen that listings with higher prices have extra amenities such as air conditioning, smoke detectors, essentials shampoo, Wi-Fi air, carbon monoxide. Like NYC, listings with higher prices have extra amenities such as air conditioning, carbon detectors, essentials shampoo, Wi-Fi air, carbon monoxide. And it is worthwhile to point out that the parking premise in LA amenity keyword while not in NYC amenity keyword(Appendix B, Figure 6). The real value of listing pricing distribution in NYC and LA (Appendix B, Figure 7) are right skewed. After log transformation, it nearly follows normal distribution.

Time Series Analysis

We implement Airbnb calendar data, which is the detailed calendar data for listings, to analyze the time series pricing. Firstly, we split the data into groups based on date and get the average prices by month in two cities (Appendix B, Figure 9, 10). We find that the two graphs have clearly different trends. For New York City, the highest average price appears in October 2020. But in Los Angeles the highest average price appears in August 2020. Since May, the average prices in New York City will be relatively stable. But in Los Angeles, the average prices begin to rise in March, peak in August, and then gradually decline.

Then, we add event dates like US federal holidays with the *Prophet* package in Python to fit our calendar data. We analyze the trend, holidays, yearly seasonality, and weekly seasonality of the time series (Appendix B, Figure 11). We apply holiday data since holidays can be an essential factor influence the tourism and Airbnb market. We could notice that the trend of LA Airbnb prices is smoother than in NYC. There is no significant yearly seasonality since the data we used only includes the next year's information. From weekly seasonality, we can see that the prices on Friday and Saturday in both two cities are significantly higher than other days of week. Those are the times when people would love to travel or spend time outside. As two most popular tourism cities, we'd love to see trends like this. We also think NYC can improve holidays tourism to attract more tourists.

Sentiment Analysis

Pricing by individual hosts can be subjective. For more precise analysis results later on to better understand the influential factors of the Airbnb pricing in NYC and LA, we conduct sentiment analysis to computationally identify and categorize the opinions expressed in the customer review for the listing products. More specifically, this step helps us determine whether the general customers' attitude towards the product is positive, negative, or neutral, which indicates whether the price for the product is reasonable and can effectively reflect its market-recognized value or not.

Firstly, after exploratory data analysis, we notice that the reviews we collected contain both null reviews and reviews in other languages. To solve this problem, we clean the null records and take a trade-off to filter out non-English reviews, since around 90% of the reviews are in English and processing multi-language can be quite time-consuming.

Secondly, we use the built-in analyzer in the *nltk* python library to generate the polarity scores. The results are in the appendix (Appendix B, Figure 12, 13, 14). As shown in the charts, the distributions of reviews' sentiment scores of New York City and Los Angeles are of similar pattern, with slightly left skewed distribution of neutral reviews, slightly right skewed distribution of positive reviews and lots of reviews are not negative at all.

Thirdly, we merge the scored reviews with original reviews data to get corresponding listing ID and calculate the mean sentiment of the listings with reviews to present the general opinion of the customers. Then, we identify houses with generally more negative than positive and significantly negative (score of negativity > 0.1, this threshold is selected by the distribution of negativity score) reviews and filter them out of the listings data as unreasonable pricing products to get the filtered listings dataset.

Correlation Analysis

Spatial Join

To take the factors of the neighborhoods into consideration, we conduct spatial join to map the listings with the local income, education and crime indicators.

For income and education indicators, we map NYC income and education data from the U.S. Census with Airbnb listings data by zip code. As for Los Angeles, we collect census data by council district from the Los Angeles Office of the Controller and spatial join it with listings data based on the latitude and longitude of the listing houses provided and the boundaries shapefile of LA council districts.

For the crime indicator, we firstly collect historical arrest data from NYPD and LAPD respectively and filter the arrest records in 2019. Then, we calculate the counts of arrests in 2019 in each zip code based on the locations of the arrests and the boundaries shapefiles of NYC and LA zip codes and spatial join them with listings data as the crime indicator.

Data preprocessing and sampling

Data preprocessing and sampling is essentially the understanding of data, which is integral to initialize analysis. However, when doing sampling and preprocessing of data, it is highly possible to have selection bias. This may affect generalizability of results and identifiability of model parameters. First of all, according to our estimation, we rule out non-essential features of original datasets and convert categorical values into dummy variables. Second, we use the sklearn package to split training, validation and test data.

When it comes to model selection, every algorithm has requirements on what type of data can be used as the input. For instance, regression based methods require numeric or binary variables while tree based methods can accept any type. In this practice, we will try both of these two methods.

Regression

For Los Angeles, Linear Regression, Random Forest, Bagging, Extra Trees and Gradient Boosting algorithms are used to predict prices.

- Comparison with alternative methods

- Linear Regression: R square is 0.444 after running through training and validating datasets.
- Random Forest: R square is 0.632 in training data and 0.609 in holdout data.
  Apparently, Random Forest Regressor performs better than Linear Regression.

- Bagging: R square is 0.901 in training data and 0.608 in holdout data.
- Extra Tree: R square is 0.999 in training data and 0.683 in holdout data.
  R square values in training data are much higher than ones in holdout data. Both of these two methods are overfitting with absurd median absolute error.

- Gradient Boosting: R square is 0.680 in training data and 0.677 in holdout data.
  Compared with above algorithms, Gradient Boosting is a robust model, and works quite well both with categorical and numerical values. The downside might be less interpretable and computationally expensive.

However, for New York City, we find that NYC airbnb data should have different strategies to do processing. The differences in housing prices really affect a lot of prediction results. Therefore, I divide prices into three basic ranges - low, medium and high, in order to do more robust prediction. Prices that are lower than $69 are in the low pool, $70-175 is in the medium pool and prices that are larger than 176 is in the high pool. Then, we use a Random Forest Classifier to predict price range, and tune the model to improve accuracy. The ultimate accuracy for Random Forest Classifier is 0.761. Next, we use Random Forest Classifier and Random Forest Regressor to get feature importance results, and see what can best determine the housing prices of NYC and LA.

Conclusion

We check the feature importance and find that it is essential to know and develop the neighborhoods since the demographic factors we selected are very critical in our feature engineering results. To be specific, both crime and education have an impact on NYC and LA, while income has the most influence on NYC which is different than LA. That's why we should consider more demographic or socio-economic data in our future research. And also, we can see for the Airbnb house condition itself, the most important factors have some differences in NYC and LA. For example, tourists who travel to LA more may care about bedrooms or bathrooms while tourists who travel to NYC tend to consider reviews and room type, which means that an entire home or an apartment really matters for tourists coming to NYC and the pricing for an Airbnb based on the special living condition in NYC.

Mitigation Strategy

In sentiment analysis, we filter out all the non-English reviews, which may cause some bias on involving foreign customers' opinions. For more inclusive analysis, we will use the *langdetect* python library and other techniques in future work to analyze reviews of all languages.

For our model building process, we only considered three neighborhood indicators, which may affect the accuracy of the model. In the future, we will involve more potential indicators to the analysis for more accurate fitting, like employment data and other socio-economic data.

When we were dealing with the census data, we spent a lot of time cleaning and processing our data. Since the census data of LA and NYC are collected from different sources, which causes some indicators are of different calculation methods. So we will use more comprehensive data sources with indicators of both cities, like the US Census.

References

1.Gurran, N., & Phibbs, P. (2017). When tourists move in: how should urban planners respond to Airbnb?. Journal of the American planning association, 83(1), 80-92. Retrieved from https://www.tandfonline.com/doi/full/10.1080/01944363.2016.1249011?scroll=top&needAccess=true#aHR0cHM6Ly93d3cudGFuZGZvbmxpbmUuY29tL2RvaS9wZGYvMTAuMTA4MC8wMTk0NDM2My4yMDE2LjEyNDkwMTE/bmVlZEFjY2Vzcz10cnVlQEBAMA==

2. Dudás, G., Vida, G., Kovalcsik, T., & Boros, L. (2017). A socio-economic analysis of Airbnb in New York City. Regional Statistics, 7(1), 135-151. https://www.ceeol.com/search/article-detail?id=578374

3. Coles, P. A., Egesdal, M., Ellen, I. G., Li, X., & Sundararajan, A. (2017). Airbnb usage across New York City neighborhoods: Geographic patterns and regulatory implications. Forthcoming, Cambridge Handbook on the Law of the Sharing Economy. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3048397

4. Sans, A. A., & Quaglieri, A. (2016). Unravelling airbnb: Urban perspectives from Barcelona. Reinventing the local in tourism: Producing, consuming and negotiating place, 73, 209.

5. Davis, P. (2016). How do sharing economy companies grow? A comparison of internal and external growth patterns of Airbnb and Uber. https://trace.tennessee.edu/cgi/viewcontent.cgi?article=2925&context=utk_chanhonoproj

6. Chatterjee, D., Dandona, B., Mitra, A., & Giri, M. (2019). Airbnb in India: comparison with hotels, and factors affecting purchase intentions. International Journal of Culture, Tourism and Hospitality Research.

APPENDIX

Appendix A: Data Sources

1. Airbnb Open Data (scraped by February 13th, 2020), Inside Airbnb
   http://insideairbnb.com/get-the-data.html
2. NYC Census Data (2018), U.S. Census ACS
   https://www.census.gov/programs-surveys/acs/
3. LA Economy Panel (2018), Los Angeles Office of the Controller
   https://controllerdata.lacity.org/Statistics/Economy-Panel-LA-Data/4ndk-mmc8/data
4. NYC Arrest Data (2006 - 2019), New York Police Department (NYPD)
   https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u
5. LA Arrest Data (2016 - Present), Los Angeles Police Department (LAPD)
   https://data.lacity.org/A-Safe-City/Arrest-data-2016-present/sniz-4n2f
6. Cartographic Boundary Files - Shapefile (2018), U.S. Census
   https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html
7. LA Council Districts (2020), Bureau of Engineering
   https://data.lacity.org/A-Well-Run-City/Council-Districts/5v3h-vptv
8. US federal holidays data (2011 - 2020), U.S. Office of Personnel Management (OPM)
   https://data.world/sudipta/us-federal-holidays-2011-2020

Appendix B: Figures

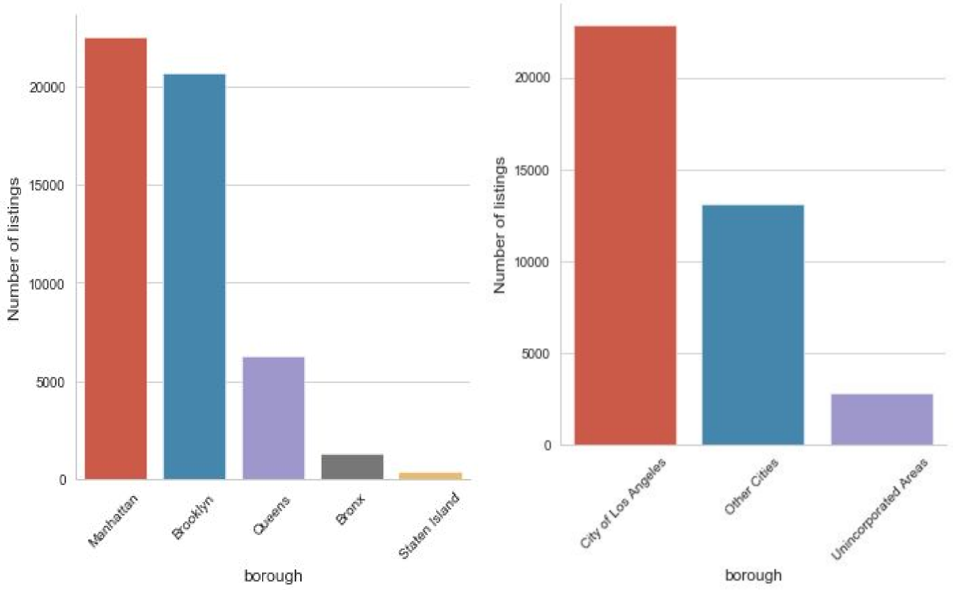


Fig.1 Listing Count by Borough in NYC and LA

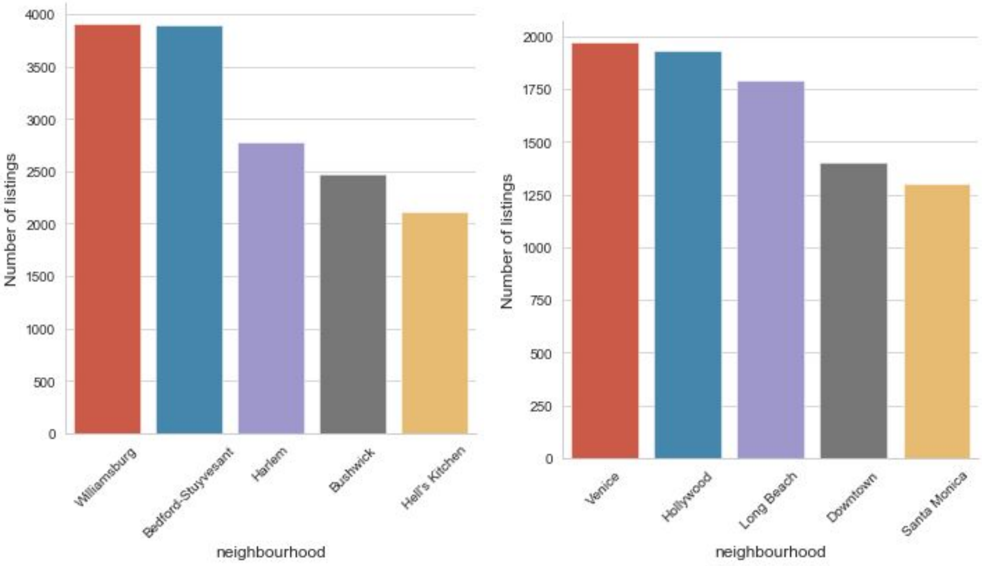


Fig.2 Top 5 Neighbourhood in NYC and LA

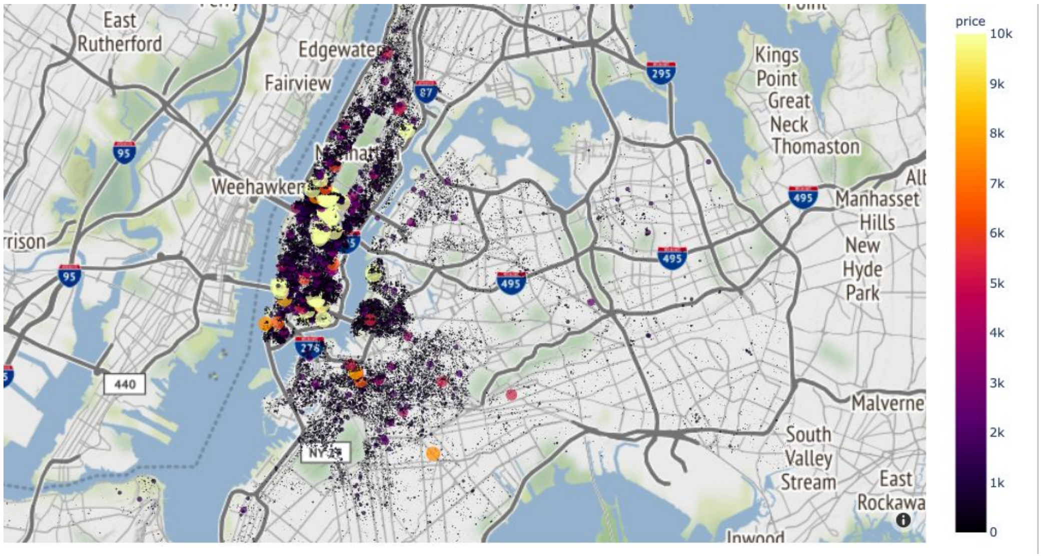


Fig.3 Price Heatmap in NYC

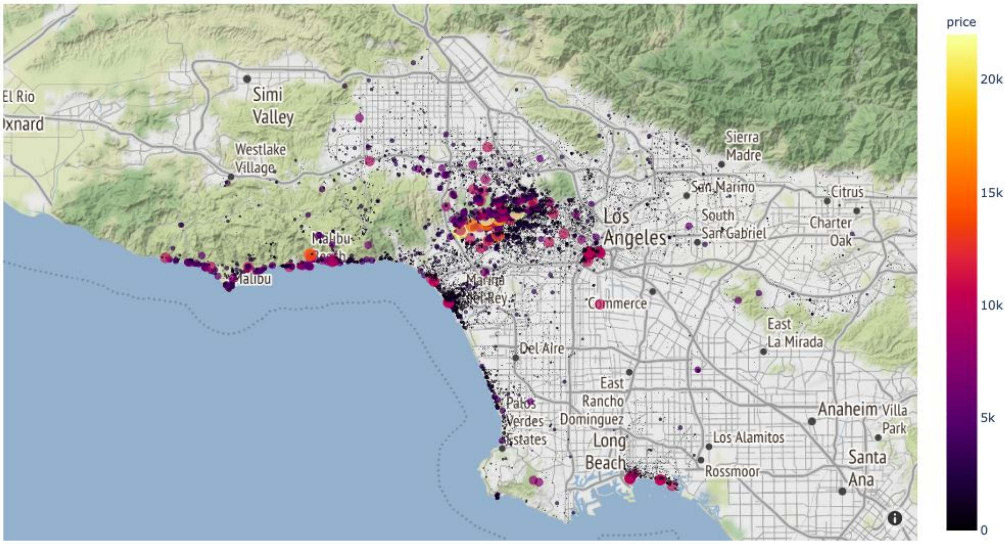


Fig.4 Price Heatmap in LA

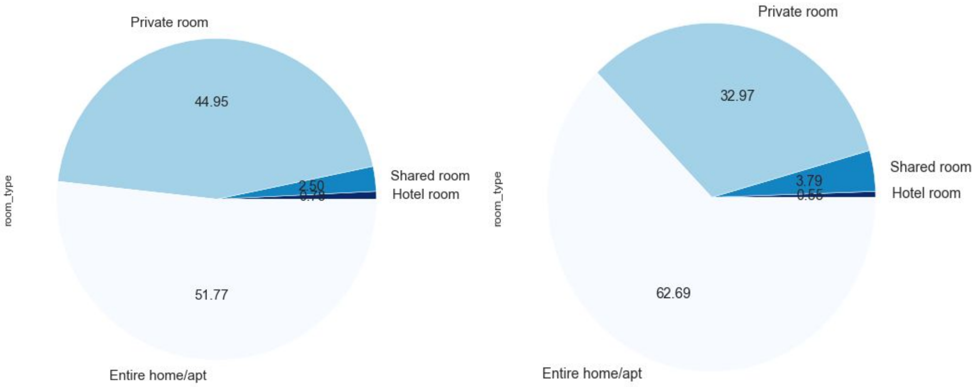


Fig.5 AirBnB Type in NYC and LA

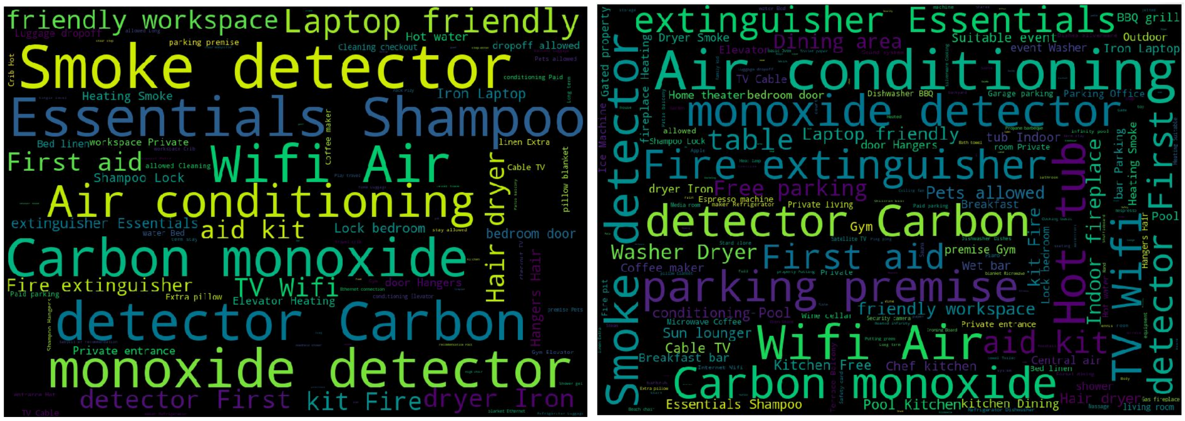


Fig.6 Amenity Keyword Analysis in NYC and LA

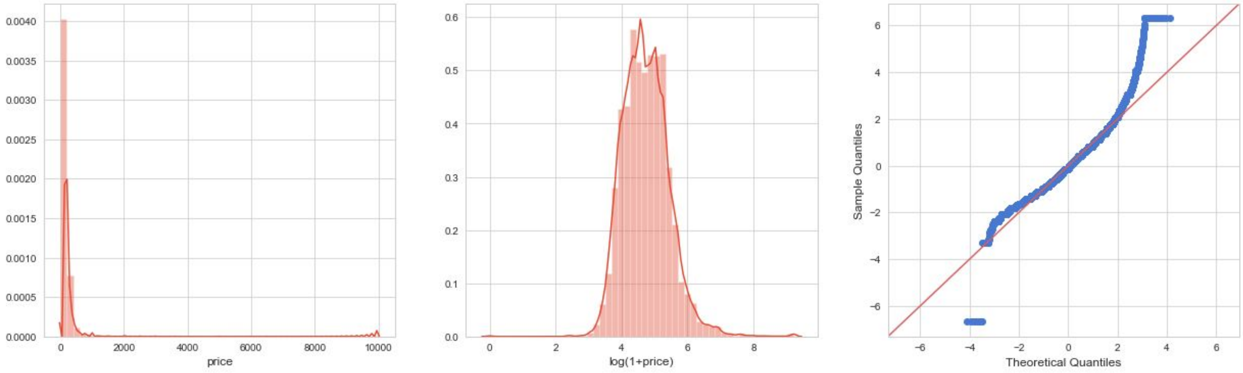


Fig. 7 List Pricing Distribution in NYC

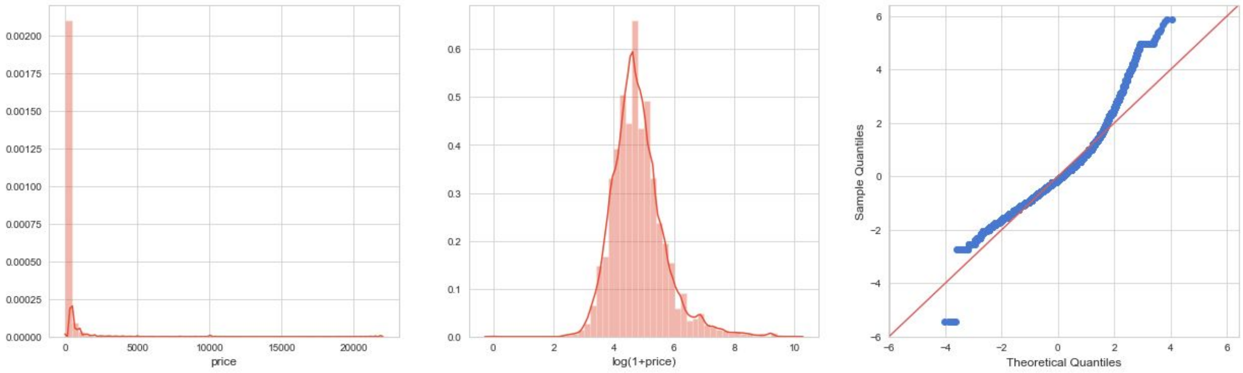


Fig. 8 List Pricing Distribution in LA

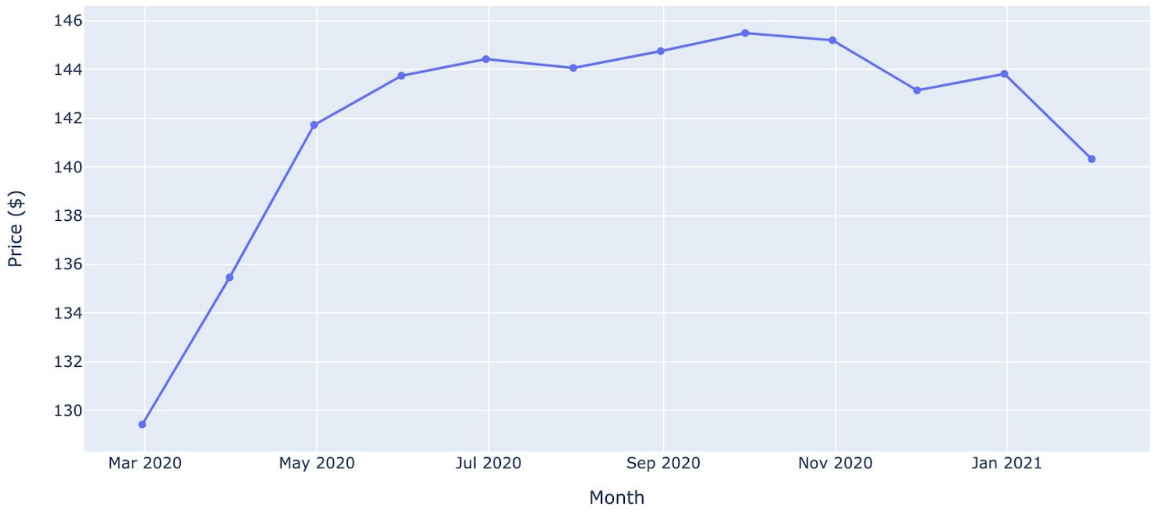


Fig. 9 Average Prices by Month in NYC

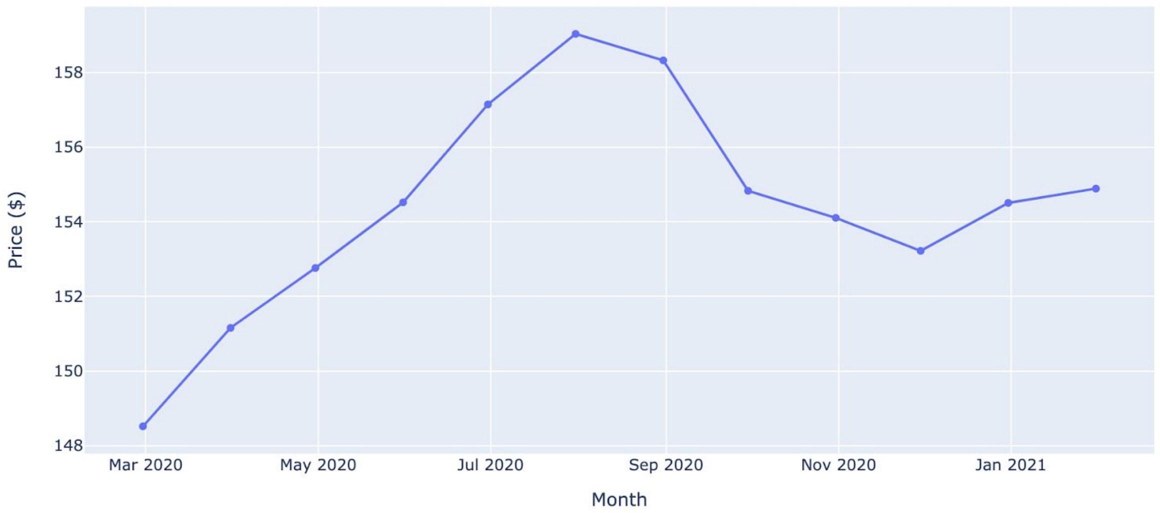


Fig. 10 Average Prices by Month in LA

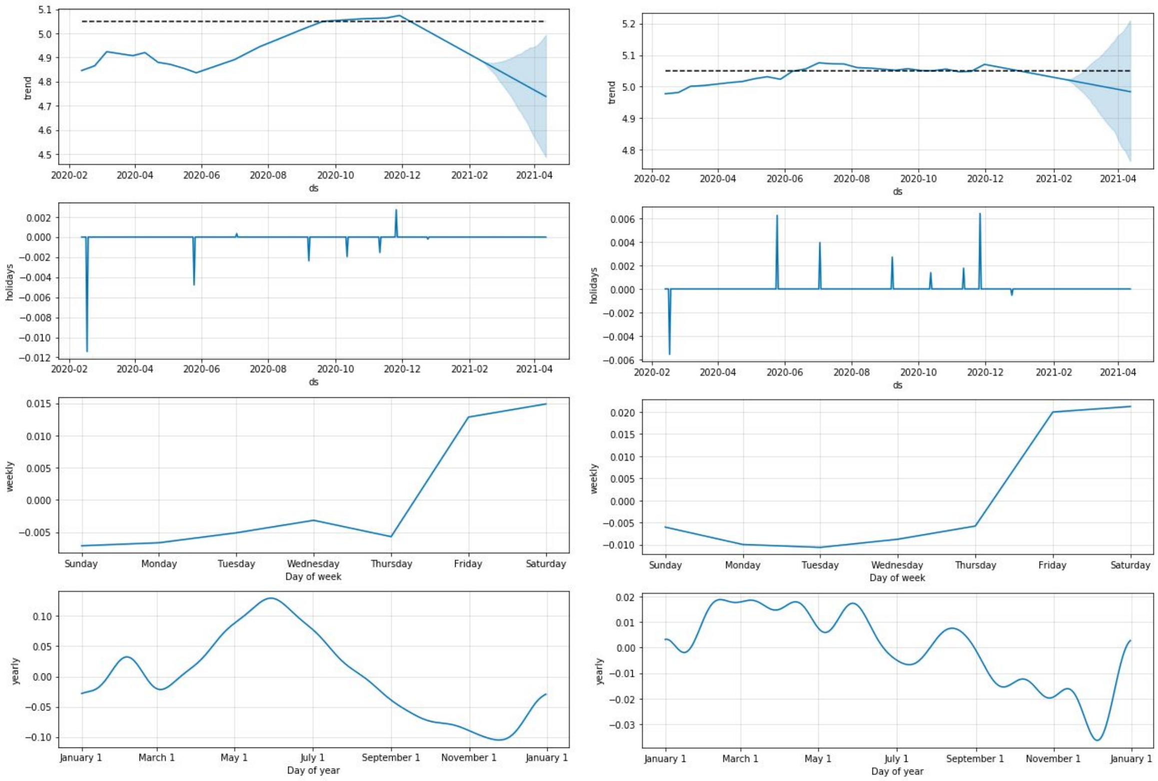


Fig. 11 Time Series in NYC and LA

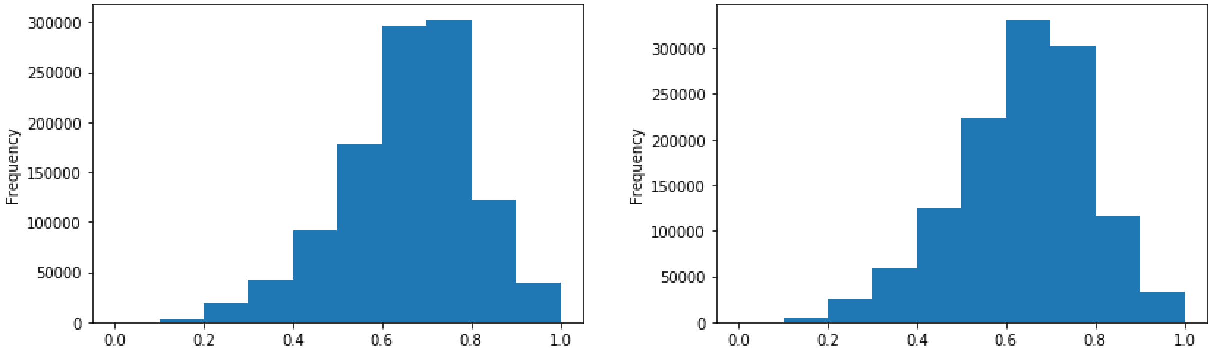


Fig. 12 Score on Neutrality (NYC & LA)

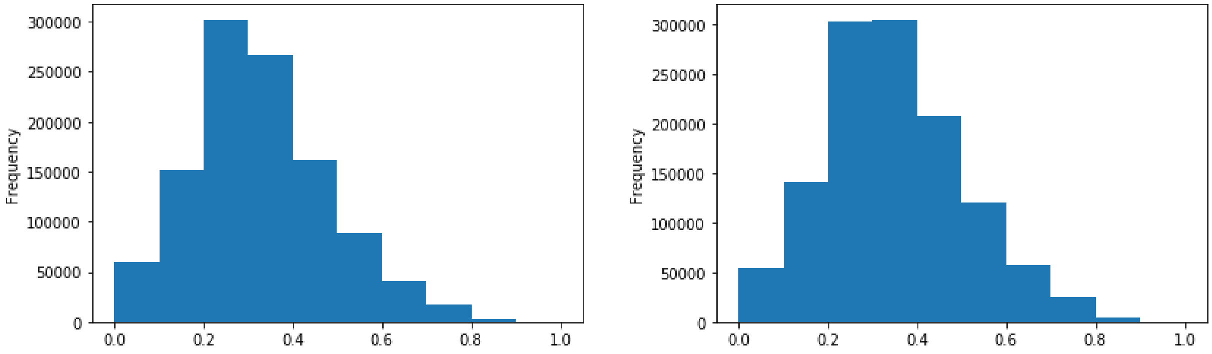


Fig. 13 Score on Positivity (NYC & LA)

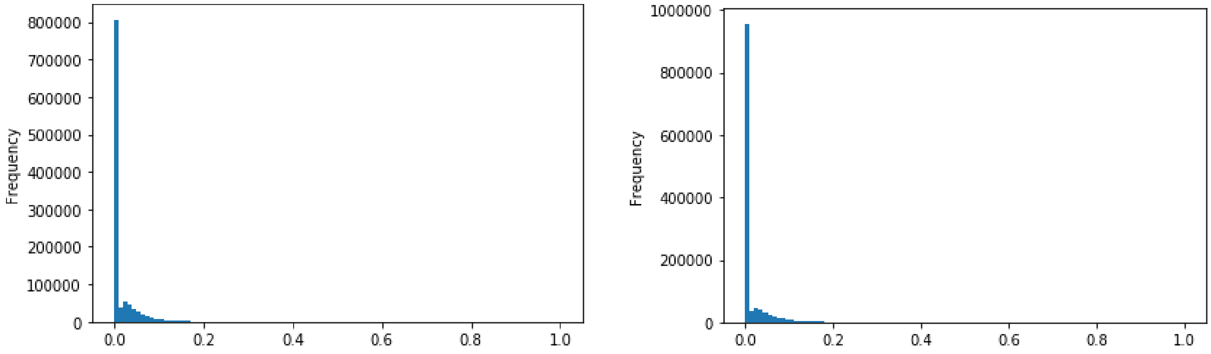


Fig. 14 Score on Negativity (NYC & LA)

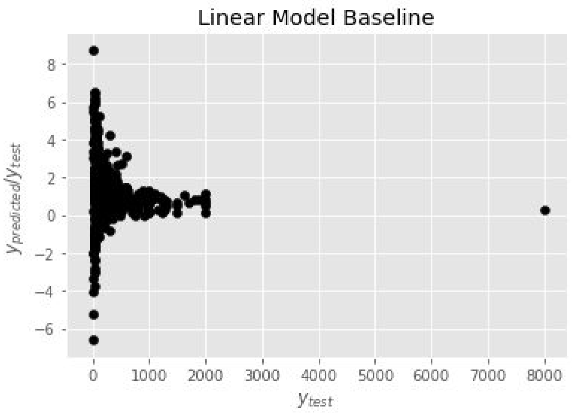
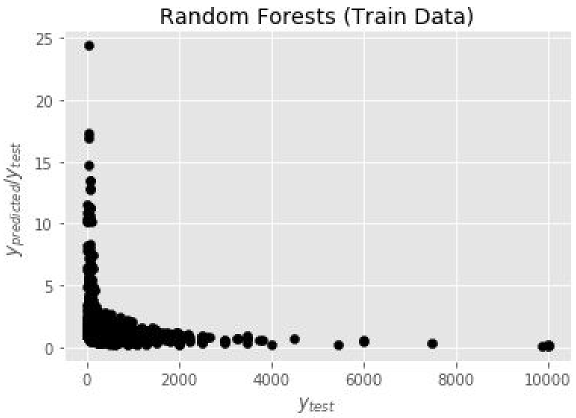


Fig. 15 Linear Model Results



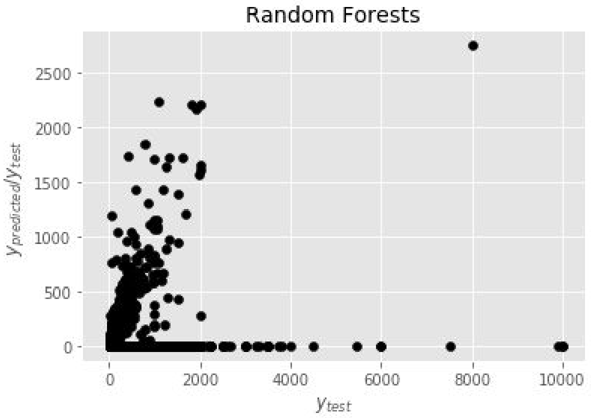
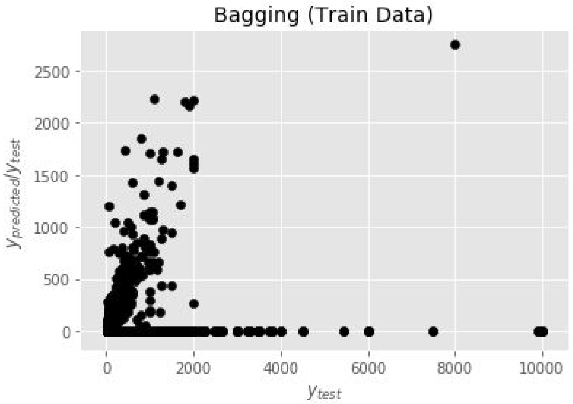


Fig. 16 Random Forest Results (Train data)

Fig. 17 Random Forest Results (Validation data)



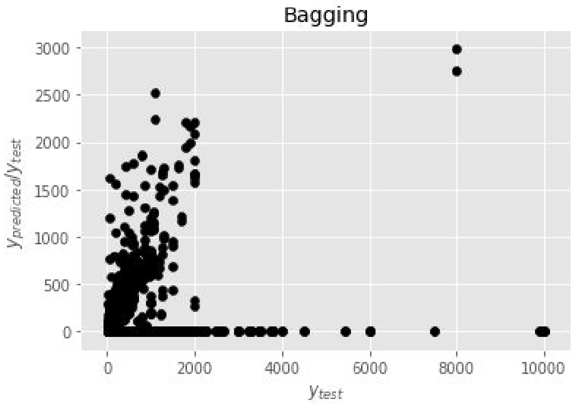


Fig. 18 Bagging Results (Train data)

Fig. 19 Bagging Results (Validation data)

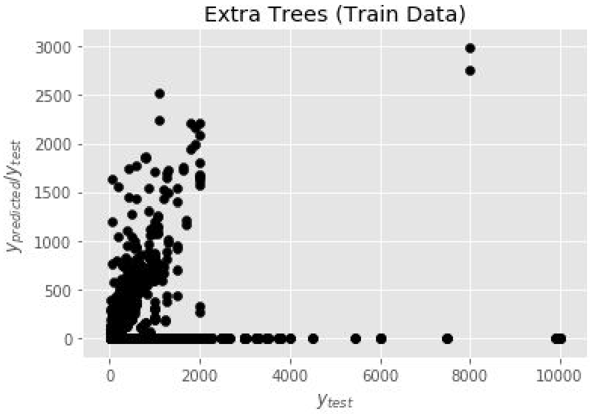


Fig. 20 Extra Trees Results (Train data)

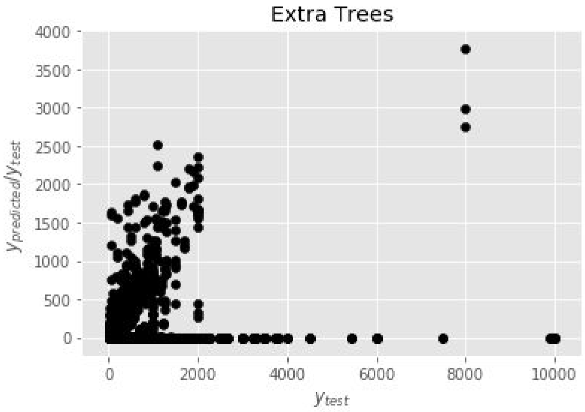


Fig. 21 Extra Trees Results (Validation data)

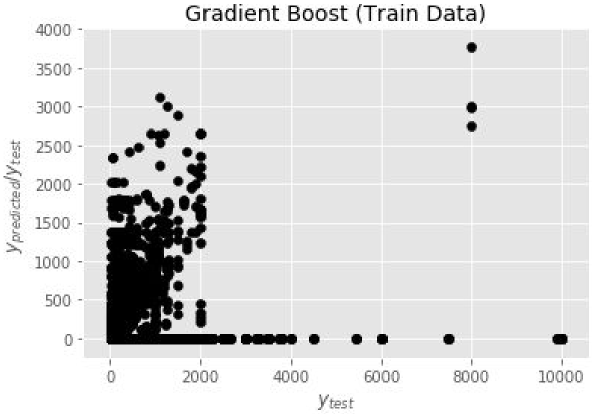


Fig. 22 Gradient Boosting Results (Train data)

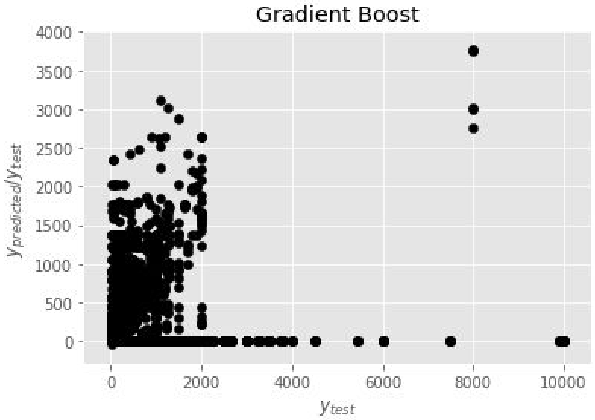
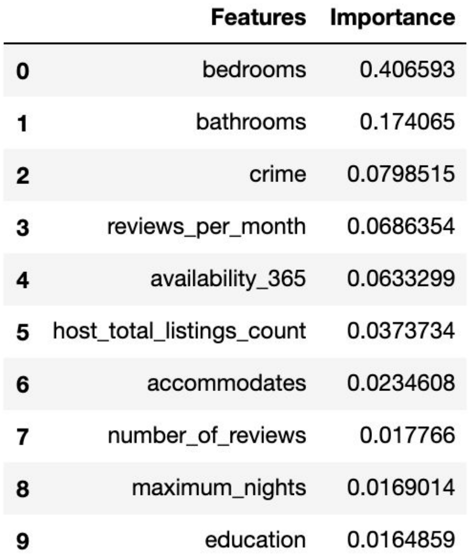


Fig. 23 Gradient Boosting Results (Validation data)

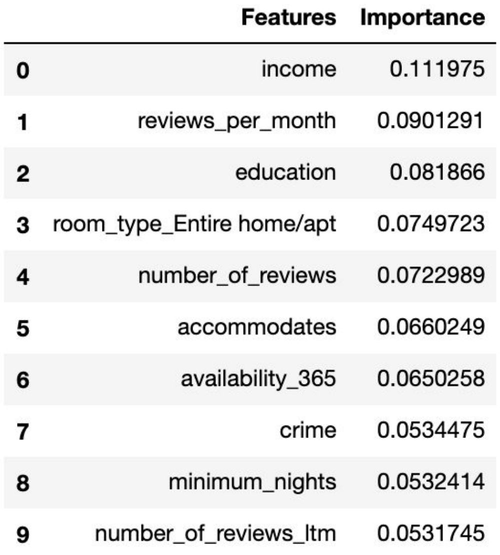
Appendix C: Tables

Modeling Evaluation Metrics (Comparison between test data and predicted data):



	Features	Importance
0	bedrooms	0.406593
1	bathrooms	0.174065
2	crime	0.0798515
3	reviews_per_month	0.0686354
4	availability_365	0.0633299
5	host_total_listings_count	0.0373734
6	accommodates	0.0234608
7	number_of_reviews	0.017766
8	maximum_nights	0.0169014
9	education	0.0164859

Table. 1 Feature Importance of Los Angeles



	Features	Importance
0	income	0.111975
1	reviews_per_month	0.0901291
2	education	0.081866
3	room_type_Entire home/apt	0.0749723
4	number_of_reviews	0.0722989
5	accommodates	0.0660249
6	availability_365	0.0650258
7	crime	0.0534475
8	minimum_nights	0.0532414
9	number_of_reviews_ltm	0.0531745

Table. 2 Feature Importance of New York City

Appendix D: Individual Contributions

Yue Jin: Data Processing, Modeling, Correlation Analysis, Paper Writing, Report

Yushu Rao: Literature Review, Exploratory Data Analysis, Paper Writing, Report

Chenjie Su: Time Series Analysis, Sentiment Analysis, Paper Writing, Report

Kunru Lu: Sentiment Analysis, Correlation Analysis, Paper Writing, Report