Predictive Listings of Airbnb Accommodations

Alexander Sosnovsky (as10507), Raymond Shi (rs5955)

December 11, 2019

Abstract

Since 2008, Airbnb has been providing alternatives to renting a room at a hotel, instead allowing users to rent a room provided by a person. The amount of listings here in NYC has exploded since Airbnb started, with almost 50,000 rooms, houses, or shared areas available to book on average. That is a very large amount of rooms to choose from, and while there are filters to help narrow the search down, there are many listings in similar locations with the same filter information. We propose an algorithm that, given some basic information about the user, ranks a given set of listings (with similar statistics) based on how likely the user is to choose that listing. This would result in a higher likeliness that the user would find a listing they are interested in, and therefore rent. Our methodology for testing the program includes having a survey in which participants are shown some mock Airbnb listings with their information, and told to choose one. Then, we compare the responses to the predicted responses generated by the algorithm.

Keywords

Airbnb, socioeconomics, accommodations, New York City

1. Introduction

Airbnb is an online marketplace for people to list, search, and book lodging. With over 150 million users across the globe and over 650,000 hosts registered, Airbnb provides about 500,000 stays every night. Compared to a single hotel room, an entire home on Airbnb costs between 6% to 17% less. The numerous affordable listings available at each city allows Airbnb to create an authentic experience tailored to each

individual. Consequently, Airbnb is projected to overtake hotels in revenue and users; the number of bookings increases by 45% each year. As the number of listings on the site increases, users will have more options to select. While this seems advantageous to the users, this will also increase the time spent searching for the ideal accommodation.

Currently, users spend an average of 11 minutes and 31 seconds on the Airbnb app. On Airbnb, there are over 4 million listings worldwide - 1.9 million of which are immediately bookable. With the Airbnb marketplace becoming increasingly popular, we expect the average time spent on the app to increase even more. Our goal is to reduce the inconvenience of searching for the best accommodations possible. The filters Airbnb provides allow users to select their preferences. However, the filters are not personalized to the user; the filters do not ask for personal information such as the user's residence, age, gender, etc. which may be critical in deciding what accommodations the user would actually choose.

We hypothesize that users' living conditions, primarily their hometown and current residence, are strong indicators of their socioeconomic status, and subsequently, their accommodation selections. Our goal is to take into consideration the factors that Airbnb generally overlooks when presenting their list of accommodations. First, the user will answer a few personal questions (our program will ask for the user's zip codes). Then, our program will select 10 accommodations we think the user will most likely select based on the answers, and present them in descending order of importance. Essentially, we plan to develop a simplified "Airbnb search engine" that requires only a few answers as input. Afterwards, we will analyze the user's selection and its proximity

to our “optimal” selection (the first option). We believe that users who live or have lived in wealthy neighborhoods and are accustomed to the urban lifestyle will select Airbnb accommodations in highly rated neighborhoods, which are near subways and popular hotspots. That is to say, we expect there to be a correlation between residence and selection of accommodation.

2. Literature

Our motivation for this project is partially inspired by Google’s pagerank algorithm, which provides users who want to search for something a list to choose from. There is significant discourse related to growth and effect of Airbnb. Prior research has labeled Airbnb as a technological phenomenon that has revolutionized the idea of lodging.

2.1. Airbnb Rental Listings Dataset Mining

Sarang Gupta notes the meteoric growth of Airbnb since its inception in 2008. He observes that the demand for Airbnb rentals has increased exponentially over the years; the number of unique listings receiving reviews has increased exponentially since 2008. Higher rated neighborhoods are better connected to subways and closer to hotspots, such as Times Square and Wall Street. There is a correlation between location score and the appeal of the neighborhood. Gupta observes that highly rated locations tend to be more expensive, due to supply and demand. Furthermore, the demand for Airbnb rentals is the lowest in January and the greatest in October, where it slowly falls until December. Similar to the demand of listings, average prices across listings increases across the year and spikes in December. Fridays and Saturdays tend to be more expensive compared to the other days of the week. Gupta also analyzes the Airbnb word cloud - keywords such

as neighborhood, location, area, subway, and walk are frequently mentioned. To summarize the word clouds: helpful "hosts" and "communication" correlate to a comfortable stay. Finally, Gupta notices that the "Superhost" title, awarded by Airbnb, requires a host to maintain a review rate above 50%, a response rate above 90%, and other criteria.

2.2. Sharing Means Renting?: An Entire-marketplace Analysis of Airbnb

Qing Ke discusses how, while Airbnb and similar online peer-to-peer marketplaces have proliferated over the recent years, there is also an increase in numbers of debates, regulatory challenges, and battles regarding the service. Supporters of Airbnb affirm that the marketplace allows established household owners to become small business owners and alleviate rental-related stress. On the contrary, critics argue that Airbnb enables illegal, short-term rentals. Ke observes that, across over 191 countries, 2million listings, and 60million total guests, there is an 18.6% guest-to-host review rate. He analyzes the geolocation, room type, star ratings, and reviews of the listings that appear on the marketplace. Interestingly, there is a "rich-get-richer" explanation of review growth. As the host's age increases, the number of reviews increases correspondingly. 66.3% of guests have left a single review, while only .63% of guests have left at least 10 reviews. Trust is also a crucial part of providing services; hosts establish trust by showing faces in their profile photos. Furthermore, characteristics between Airbnb multi-listers and other hosts differ greatly. Multi-listers tend to use the description section of the listing to advertise their homes. While there is a significant number of hosts on the marketplace, there is no clean distinction between professional

and non-professional hosts. Ultimately, monetary compensation and sociability are the two primary reasons for providing hospitality, and as a result, the main determinants of the Airbnb review system.

2.3. The High Cost of Short-Term Rentals in New York City

Wachsmuth et al. analyze Airbnb activity within New York City and surrounding areas, questioning how Airbnb is affecting the housing market and evolving social perceptions. They note that two-thirds of Airbnb listings in NY are likely illegal — 66% of revenue and 45% of all NY Airbnb reservations (the year before) were illegal. Airbnb has also removed about 13,500 units of housing from NYC's rental market. Consequently, rent has increased by $380, since the housing supply, for non Airbnb users, has decreased. For some Manhattan neighborhoods, rent has drastically increased by more than $700. Moreover, white neighborhoods are more profitable on Airbnb compared to non-white neighborhoods. Wachsmuth et al. refer to InsideAirbnb.com's "Airbnb as a Racial Gentrification Tool," where predominantly Black NYC neighborhoods were analyzed. The study found that "Across all 72 predominantly Black New York City neighborhoods, Airbnb hosts are 5 times more likely to be white. In those neighborhoods, the Airbnb host population is 74% white, while the white resident population is only 14%." Even in Black neighborhoods, white hosts earned 530% more than black hosts. As a result, groups and housing advocates are becoming more concerned about the negative effects of Airbnb on the city.

3. Data

We use three different datasets. We use the Kaggle dataset of about 50,000 different Airbnb listings in NYC, which contains information publicly found on the

Airbnb website. Of all 16 columns, our focus is on neighborhood_group, neighborhood, room_type, and price. Namely, we want to analyze how a user's living conditions factor into accommodation choice. We compile the selected columns from the original Airbnb dataset into "rooms.csv". Since the number of listings is quite large, we decided to only use about 5,000 of the total number of listings. We also use mean and median of households per zip code. The dataset contains 32,000 records on US Household Income Statistics & Geo Locations. The dataset is compiled based on the 2015 US Census Report, which we store in "incomes.csv". Finally, we use the median household income dataset per borough in NYC. The data is compiled by Renthop, a web platform that provides users with an apartment-search web facility, which we store in "nycincome.csv". The three datasets are used for our program, which uses the income datasets to predict the user's income and then selects the best listings from the Airbnb dataset. Lastly, we use the data we collect from the people we survey so that we can test our hypothesis. The data includes the zip codes of the user's current and past residence. If the user does not reside within the United States or does not have a valid zip code, then we disregard the information.

4. Experiments

We conduct our experiment using two main sources of data: the public Airbnb NYC dataset and a questionnaire that we ask potential or current Airbnb users to complete. The dataset presents social factors, such as the neighborhood, number of reviews, and the room type, alongside the price of the accommodation. Then, we observe the relationship, if any, between listing selection and affordability. From the Airbnb

dataset, we use only the id, name, borough, neighborhood, room type, price, and number of reviews per listing. On the other hand, the data we collect from the questionnaires is processed by a python program, which calculates the 10 best listings for the user. The user then selects a listing. Finally, the information is collated and stored in a csv file that we analyze further. We also use two additional datasets to inform us about household income across the nation and across different boroughs in NYC. This data is then used in our calculations to predict what the user's income may be and subsequently, what Airbnb listing the user decides to select. We conduct a one-tailed t-test to see if there is a correlation between current living condition, specifically where the user lives, and what Airbnb listing the user selects.

4.1. Algorithm

- The interface of this project is a PHP webpage that takes in various inputs about the user, which sends the input to a backend Python script. The script then returns a selection of rooms for the user, and their final choice is recorded.
- The backend script takes in, as command line arguments, the zip codes of the user's hometown and current place of residence. We reference this information against the US income dataset, and add 35% of their hometown's median income to 65% of their current residence's median income. This is then converted to an index score between 1 and 10. If one of the zip codes is invalid, the index only uses the other zip code. If both zip codes are invalid, their index is set to 1.

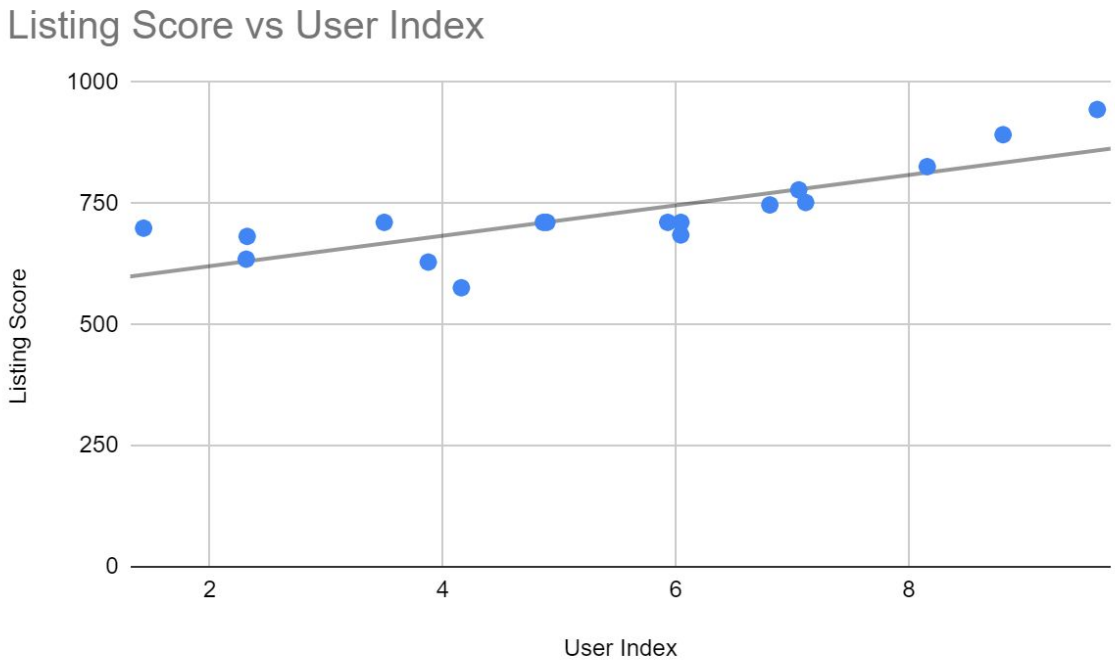
- We then read the dataset containing all the available rooms; a score out of 1000 is generated for each room based on the room's qualities and user's estimated income. This score is weighted as follows:
    - (65%) How close the room's area income is to the user's estimated income. A low score means the room's area income is lower than the user's, a medium score is if they match, and a high score means the room's area income is higher than the user's.
    - (35%) Deprioritizing rooms that are predicted to be too expensive.
        - The maximum a user is willing to spend on a room (nightly) is calculated with a formula based on their income. This is their income index squared, then multiplied by 500 and divided by 1 plus the income index. If a room is more expensive than that, it is deprioritized based on how much more expensive it is.
- We return the top 10 best-scoring listings, of which the user chooses one. Their responses (and the top 10 room scores) are saved for statistical analysis.

5. Results

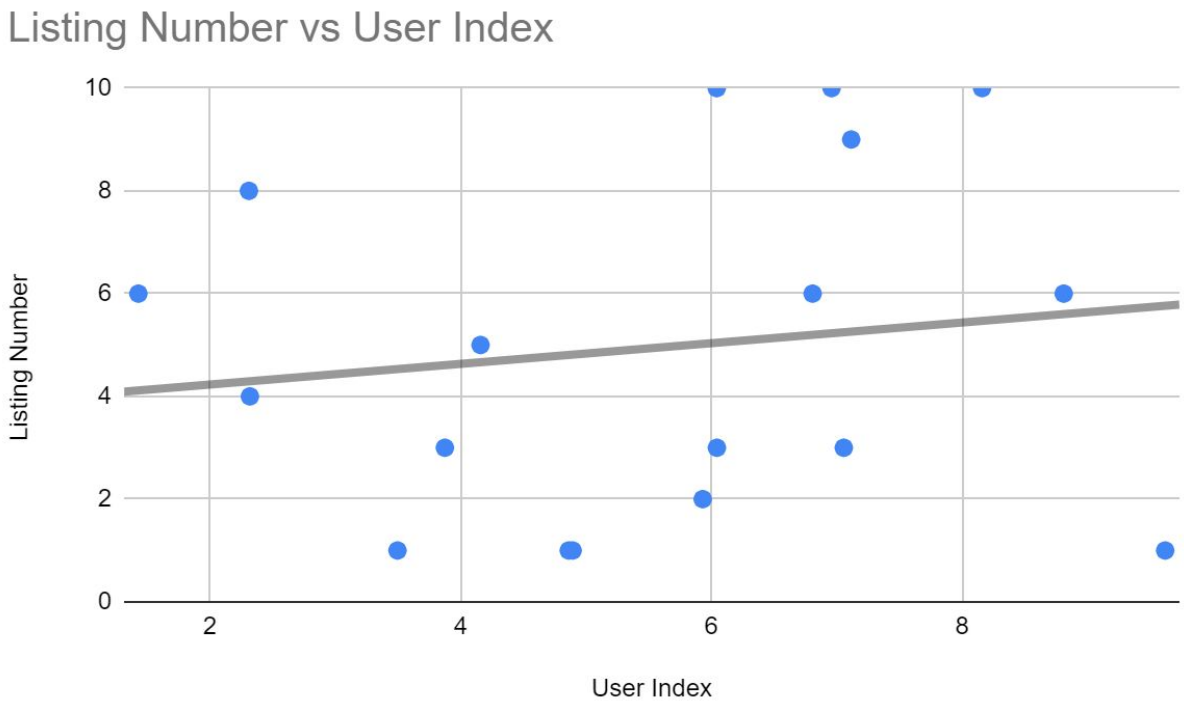
User #	User Index	Listing Selection Number	Selection ID	Listing Score
1	6.807206	6	7750	747
2	2.319895	4	7003697	682
3	3.49668	1	9782	711
4	4.158948	5	781486	576
5	6.043833	3	15341	711
6	8.809883	6	7750	892
7	8.158723	10	19319	826
8	7.114923	9	498859	752
9	5.930627	2	15341	711
10	1.430467	6	470498	699
11	4.893225	1	5203	711
12	6.04254	10	2618288	685
13	9.618925	1	5121	944
14	7.055863	3	444430	778
15	3.874598	3	5121	629
16	4.862617	1	5293	711
17	2.311775	8	470498	635
18	6.95859	10	498859	741

From this chart, we focus on the user index, which is affected by the user's zip codes, and how it affects the listing selection number. The user index, which reflects the user's estimated wealth, ranges from 1 to 10 inclusive (1 being least wealthy and 10 being most wealthy) and depends on the zip codes of the user's hometown and current residence. We notice that the user index and the listing selection numbers distributed uniformly. In addition, the listing scores are all at least 500, suggesting that users tend to pick listings located in well-off neighborhoods.

6. Data Analysis



From the Listing Score vs User Index graph, we notice a slight positive correlation between user index and listing score.



From the data we collected, the least-squares regression equation: y = 3.828 + .201*(user_index).

Regression Statistics	
Multiple R	0.139408
R Square	0.019435
Adjusted R Square	-0.04185
Standard Error	3.40287
Observations	18

Our correlation coefficient is .1394, which means that there is a weak relationship between the user index and listing selection number. In addition, only 1.94% of our values fit the regression analysis model.

6.1 Hypothesis Testing

Hypothesis

H_o: $\mu > 1$

Users' hometown and current residence do not affect their Airbnb selections. If users using our algorithm, which is dependent on the user's hometown and current residence, do not select the first option on average, then $\mu > 1$.

\ H_A: $\mu = 1$

User's hometown and current residence do affect their Airbnb selections. If users select the first option on average, then $\mu = 1$.

μ is the mean listing number Airbnb users select. The listings are sorted by a score our algorithm generates based on the zip codes that the user gives as input.

Level of Significance

\ $\alpha = .05$

One-Tailed T-Test

$df = 17$

$t^* = 1.740$

T-Statistic

$t = 5.01972$

Conclusion

Since t is less extreme than t*, we do not reject the null hypothesis. There is no statistical evidence that the population mean's Airbnb selections are based on their hometown and current residence.

7. Conclusion and Future Work

Based on the results of our experiment, we conclude that there is no correlation between users' living conditions and their Airbnb listing selection. We acknowledge that there were biases in our experiment. For example, we collected data from a small sample size, so there was not a large enough data to account for the general population of Airbnb users. Additionally, we surveyed Americans, which excludes Airbnb users from other parts of the world. There is also selection bias with our experiment due to how we conducted the survey and asked certain group of people, namely people we knew, to respond, which most likely skewed the results. Additionally, there could have been response bias with what information the users were willing to share. Specifically, the personal information we ask for in the survey — the zip codes — might not be information the user is willing to report honestly. If we make the Airbnb listings to be

even more specific, then we might ask more personal questions, which could make this bias more apparent.

To improve on this project, we could analyze a broader set of Airbnb users; we can include Airbnb users who reside outside the United States. We could increase our sample size and randomly select the people we ask. We could also increase the scalability of our program by asking different personal questions. To pivot, we could use a different metric to predict what the user might select. For example, we could use a personality test as an indicator of what Airbnb listing the user selects. Additionally, rather than observing income per region, we could observe the physical environment of where the user lives, such as the quality of the infrastructure and perhaps the climate and weather as well.

8. References

1. https://boosters.fsu.edu/sites/boosters.fsu.edu/files/documents/Events/About%20Airbnb.pdf
2. https://ipropertymanagement.com/airbnb-statistics
3. https://towardsdatascience.com/airbnb-rental-listings-dataset-mining-f972ed08ddec
4. https://arxiv.org/pdf/1701.01645.pdf
5. http://www.sharebetter.org/wp-content/uploads/2018/01/High-Cost-Short-Term-Rentals.pdf
6. http://snap.stanford.edu/class/cs224w-2014/projects2014/cs224w-44-final.pdf
7. https://www.researchgate.net/publication/320443736_A_socio-economic_analysis_of_Airbnb_in_New_York_City
8. http://insideairbnb.com/
9. https://www.census.gov/acs/www/data/data-tables-and-tools/narrative-profiles/2017/report.php?geotype=nation&usVal=us
10. https://medium.com/@kasiarachuta/basic-statistics-in-pandas-dataframe-594208074f85
11. https://medium.com/python-pandemonium/data-visualization-in-python-bar-graph-in-matplotlib-f1738602e9c4
12. https://project.wnyc.org/median-income-nation/
13. https://www.geeksforgeeks.org/python-pandas-dataframe-describe-method/
14. https://www.ablebits.com/office-addins-blog/2018/08/01/linear-regression-analysis-excel/
15. https://stattrek.com/multiple-regression/excel.aspx
16. https://www.youtube.com/watch?v=lNoxKsuJ6Xc
17. https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations
18. https://ny.curbed.com/platform/amp/2017/8/4/16099252/new-york-neighborhood-affordability
19. https://www.ablebits.com/office-addins-blog/2018/08/01/linear-regression-analysis-excel/