Bellevue University

Winter 2020

Let Data Decide Your Next Destination

December 15, 2020

Kevin Angotti

Executive Summary

Travel is a billion-dollar industry, and the hotel or lodging sector takes a large portion of the revenue. Airbnb has quickly become a staple in the travel sector, and with their bookable locations in just about every travel destination around the globe, you can easily see why they have become o successful. This project looks to finding ways a customer can locate the best host to book based on location and affordability. An overview of the methods and findings of this project are summarized below.

By looking at Airbnb data to determine if they are better options available than traditional hotel booking sites, I want to see if questions like, are there be just as many available places as you would find during a traditional hotel search? Does the price of the booking correlate with reviewing ratings? What determines if a host is successful? Is there one or a few variables that can predict if a host will be booked? Do hosts with instant booking available produce more booking? Answering some of these questions can determine if this project was successful or if a different approach is be needed.

Will the returned search for a particular property be of quality or money well-spent feeling? I think customer reviews significantly impact customer choices; however, can this type of data provide good predictions into choosing a suitable property or host to book with.

The majority of people looking to book a place to stay while on vacation or traveling tend to use travel sites or bargain sites to book their stay. Airbnb properties do not advertise that way; outside of using Google for a place to stay, Airbnb properties do not show up on sites like Travelocity or Expedia; you have to go directly to the Airbnb site.

Methods

From my project, I will utilize various predictive analytics methods to understand relationships within the data and identify factors that would allow a person to search through a site to choose a location based on specific criteria. Key factors examined during the project are price, location, host, number of properties, number of reviews, and if the customer could use instant book.

Findings

The project yielded several interesting insights worth noting. More information regarding the findings can be found down below. The main goal of creating a plugin application for travel sits would be possible with some further analysis or other locations and assistance from software engineers.

Abstract

While looking for the most affordable places to stay while on vacation, a weekend trip, or while on business can become a stressful task on its own. There are numerous travel sites out there that can provide you deals on hotels; Orbitz, Travelocity, and Priceline just a few that can offer deals on hotel stays. There is currently no sites or plugin application that can be used to help narrow down Airbnb locations by specific search criteria, such as price, or review ratings, to name a few.

This project aims to look at Airbnb data and determine if predictive analytics could help determine if a particular host would be a better choice than another based on search criteria a user could implement. Taking that information n and running a plugin application while searching the Airbnb site for a place to stay in was the primary goal.

Keywords: Predictive Analysis, Airbnb, K-Modes Algorithm, Decision Tree Algorithm

Introduction

Since August 2008, Airbnb has become a multi-Billion-dollar company, and people around the world now mee them as an alternative for booking their travel stays. While other lodging companies are still the first choice of places to stay, Airbnb is becoming more and more popular as the year's pass. The main reason is that renting a whole home or even a private room is better than a hotel.

Currently, there is no bargain site to help the customer look for the best deals, and you would have to search through the Airbnb site and skim reviews to determine if the location was right for you. However, what if there are ways to filter through Airbnb locations for recommendations based on their reviews, price, and even property upkeep? That might help the traveler choose an Airbnb over a hotel.

The project will attempt to find these filters through predictive modeling and see if they can be used to create an application plugin for the Airbnb site or even a bargain site like Travelocity.

Background Information

Methods

Data Understanding

When approaching the data for this project, I began by examining variables within the dataset. This took time due to the number of variables with complex levels to sift through. I used a combination of the R programming language, Python programming language, and Power BI to start this process. In R, I conducted summary statistics of the datasets to see initial trends and what each variable might have to tell me. I followed this up with a linear regression model to see what these trends might mean. Within Power BI, I used the combination of my summary statistics and linear model to conduct a series of visuals for the most significant variables. Doing this at the beginning of the project allowed me to see what the data looks like before running any model type. Initial visualizations provided evidence that price may be a possible target for the models. Other areas of interest were review rating, host response, reviews per month, availability, room type (whole home or private room), and guests included. I then used Python to create all of the data models that would be needed for analysis and predictions. The first model I created was a K-Modes clustering model; I then followed that with two linear regression models and, finally, a decision tree prediction model. These models provide insight into the questions, allowed for future analysis, and eventually implemented a plugin application.

Data Preparation

Extensive data preparation was performed on the dataset to transform key categorical predictors into a numeric form. This was done to prepare the data for modeling and further analysis. During the data cleaning process, certain variables were transformed into numeric values from categorical ones. This would allow for the summary statistics and linear models to run and provide insights into the data. The variables were transformed in Python on the data set by the use of the formula, (df['columnName'] = le.fit_transform(df.columnName.values). A new dataset created with the transformed variables was saved and then loaded into the Rstudio to run the summary stats and the initial liner model. The main variables that were transformed were room_type and neighborhood.

The next process I conducted was to create each of the highlighted variables (price, host response, reviews per month, availability, room type) visuals to get a better look at the data. As you will see in Appendix A, Figures 1 through 6 provide a good sense of what the data suggests a customer might choose for a place to stay based on the current reviews or location.

Modeling

The extensive data preparation allowed for linear regression models to be performed. R's initial linear model was for the price variable as a function of the instant book, availability, minimum nights, number of reviews, reviews per month, and host response. These variables returned some good insights for significance, and all but one returned a p-value under the 0.05 threshold. The overall P-value of the model yielded < 2.2e-16 or (0.00000000000000022) with an F-statistic of 27.68.

I then shifted to Python to run another linear model. I utilized a clustering algorithm that will better understand some of the categorical variables that could be important. I chose to deploy a K-Modes clustering model to better deal with the complex variables contained within the dataset.   This model allows for matching clusters based on the number of its matching categories between each data point. This clustering model works well with data with a high amount of categorical mixed in numeric data.

I then worked o a decision tree model that would take in the dataset with a target in mind and return the predicted outcome based on threshold criteria of each node in the tree. The mixture of data types is thought to provide the right prediction for properties a customer might choose by answering simple yes-no questions.

Results

From the preliminary results, things are looking promising. While further analysis is needed to draw definitive conclusions, the results of initial models are shown below.

Linear Model

```
lm(formula = price ~ instant_bookable + availability_365 + minimum_nights +
    number_of_reviews + reviews_per_month + host_response_rate,
    data = df2)

Residuals:
   Min     1Q Median     3Q    Max
-266.6 -108.8  -50.3   29.9 7820.7

Coefficients:
                      Estimate    Std. Error t value      Pr(>|t|)
(Intercept)        278.148607236   9.197847725  30.241       < 2e-16 ***
instant_bookable    -6.950344577   6.050437632  -1.149       0.25070
availability_365    -0.073083931   0.024048295  -3.039       0.00238 **
minimum_nights      -0.000002051   0.000002538  -0.808       0.41904
number_of_reviews   -0.170234684   0.052290551  -3.256       0.00114 **
reviews_per_month  -12.200102032   2.053149427  -5.942 0.00000000294 ***
host_response_rate  -0.272869678   0.098061170  -2.783       0.00541 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 253.7 on 7568 degrees of freedom
Multiple R-squared:  0.02148,           Adjusted R-squared:  0.0207
F-statistic: 27.68 on 6 and 7568 DF,  p-value: < 2.2e-16
```

Table 1. Linear Regression Model

After running the linear regression model in R, which need different summary statistics to determine significance, such as an F-Statistic, P-Value, and R-Squared, these measures only provided some of the information needed to determine if the data could predict on a model. I also created visuals in Power BI (Appendix A) from these statistics to get a peek at what a model might conclude before I started building my models. The first model was created from a form of K-Means clustering model called K-Modes. This type of model handles categorical data very well. A K-Modes model takes in the entire dataset and returns its prediction on each cluster's most common outcome. These results each time the model ran returned 5 clusters; 3 of the 5 were houses while the other two were apartments. I found it surprising that none of the clusters returned were private rooms, which is a large portion of the dataset. Most of the clusters returned a price range of $100 to $150, while one cluster returned a $250 price. A note to point out is that there was a long list of accommodations in three of the five clusters that could also have some significance level that I may need to look into further.

When building my next model(s), Python's first linear regression model took the price as the target variable and utilized a 60/40 test train split. The result, which is shown in figure 1, returned predictions for each of the neighborhood types. Unfortunately, the model was only able to provide an accuracy of .26 to 0.31 percent. This indicates that the model was not extremely accurate, and adjustments may need to be made to account for skew or other currently unknown factors. There can be many reasons for this, the linear relationships are not appropriate, or the model itself is not a good fit. Other regression models, like logistic regression, may have been a better choice.

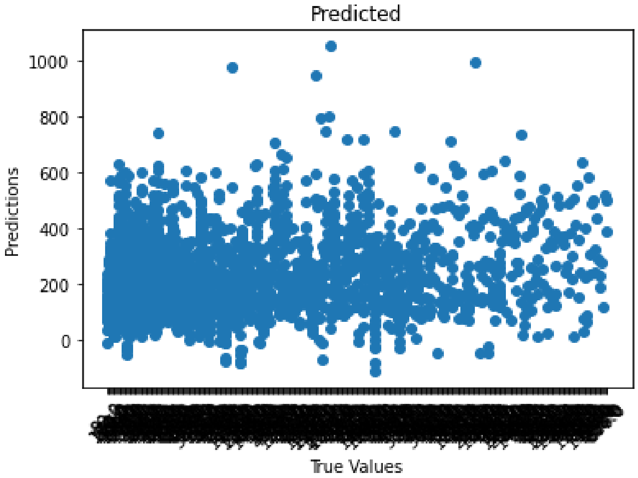


Figure 1: Model 1 predictions

I then created a second linear model, which also took the price variable as the target variable; however, it looked at the mean squared error (MSE) instead of accuracy. On the dataset as a whole, the predictions when plotted look similar (see figure 2) and return an MSE of 22110.00, which is exceptionally high and will need to be reevaluated for the final results.

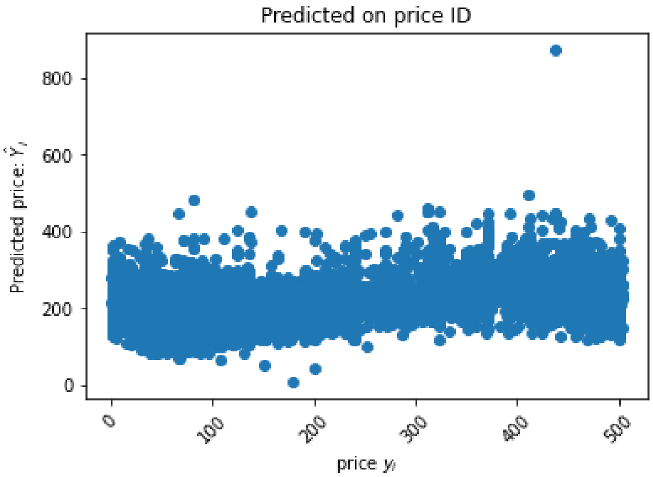


Figure 2: Model 2 Predictions

This model also ran against two variables in a separate run price against the number of minimum nights you could book. The training and testing data was split 60/40 with a random state of 5 and returned an MSE of 25349.17, also exceptionally high. The residuals are plotted in figure 3, which indicates that the error is contained to price above $200.

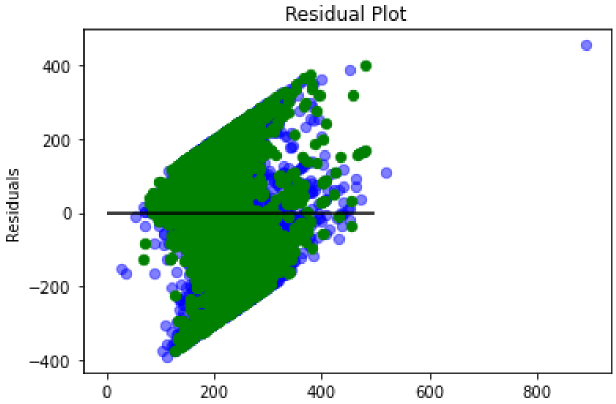


Figure 3: Residuals

The last model I created was a decision tree. This model typically provides a good prediction for a dataset of this type. The algorithm was set up to predict the neighborhood best suited for the search criteria of the customer. You can see in figure 4 (below), a host with more than 8.5 listings (top node) and the customer wanted to stay less than 17.5 nights (2nd node right), the search would narrow down(predict) 371 locations (terminal node bottom second from the right).

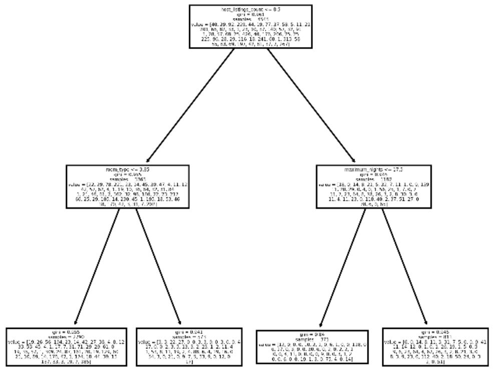


Figure 4: Decision Tree

Discussion

Upon running the summary statistics and the models, some of the insights I noticed around the key variables were the average price was around $215, with a $469 security deposit and $108 for cleaning fees and included 2 guests per stay. The average review rating was 95%, and the hosts had, on average, 20 properties/listings. On average, a host was reviewed 1.8 times a month, and the host had a 95% response rate.

Both of the linear models helped determine how price factored into how hosts were chosen based on the other variables. Getting the coefficient values of each of the key variables allows me to determine which of these variables should be included in the plugin tool. If the number of reviews or availability are positive coefficients, then as they increase, the price would also increase, making that particular property either a good or bad choice.

The K-Modes model might be one of the better models to help create the plugin tool as it looks at the entire dataset and returns the best prediction for a set of clusters. From those clusters, I am able to see what the best property type pricing and accommodations were. This can be helpful to the plugin as a way to determine the setting and search criteria when in use.

The last and maybe the most helpful part of my project was the use of Power BI for not only its visual ability but also some of its analytic tools. When creating the visuals for host locations and neighborhoods, the mapping of host properties became extremely helpful. Within Power BI, the relationship tools between visuals on the same page allow you to see different aspects of the data on one screen. My dashboard Appendix A. *Figure 9* shows how neighborhood, reviews, the value of a host's property, and room type all shown on a map can help determine which property to choose. The neighborhood is color-coordinated based on a value scale. In the figure, you can see that the Fisherman’s Wharf neighborhood in San Francisco is valued green, meaning a good value for its location, instant booking (key variable) was not a factor, and the average host had 11 listings in that neighborhood. The map showed which properties were above the green threshold, meaning more expensive for the area, allowing a customer to find a better deal for the area. If you know the area, Fisherman’s Wharf is one of the most desired areas to visit, and there are plenty of good places to stay if one was able to find the right deal.

The decision tree, though, overall, did not produce very good predictions 14 to 20 percent; specific neighborhoods did return good predictions. Four, in particular, returned 40% or higher, with one at 74%. In each of the four cases, the price was between $160 to $229 per night. Overall, the model determined the neighborhood based on the number of listing a host had and the minimum number of nights per stay. Taking the four neighborhoods and plugging them into the Power BI tool allows me to see the best locations in those areas. Plugin in the for areas Mission District, Downtown, Soma, and Outer Sunset all returned yellow to red in value. However, by digging down deeper into the areas, you can find useful insights into each host and where deals could be had.

Conclusion

Looking at Airbnb data and trying to find insights is a real possibility and the analysis I was able to perform is just a small picture of the whole aspect of the questions I looked to answer. Digging deeper into which neighborhood and or host was very successful and finding the best price and location can be possible. I will need a software engineer's assistance to help create the plugin tool for web browsers to make this fully achievable.

Questions

1. How much will this cost to implement?
2. Is there a rating system for the results?
3. What platforms will this be available on?
4. What safety features do you have in place if a host is bad?
5. is there a way to see the health precautions of property, i.e., COVID rules?
6. Will there be detailed information about the host/property?
7. Will there be copyright issues with the plugin application?
8. what type of funding do you have for creating this application?
9. Can you book through the application?
10. Will the application include points of interest for the areas of recommendation?

Acknowledgments

I want to thank Professor Catie Williams for her continued guidance and support throughout the project. I would also like to thank Data World and Airbnb for making the project data publicly available.

References

1. Lilly. (2017). BIGGEST REASONS WHY AIRBNB IS SO POPULAR. The Frugal Gene. Retrieved from https://www.thefrugalgene.com/airbnb-popular/

2. Folger. J. (2020). Airbnb: Advantages and Disadvantages. Investopedia. Retrieved from https://www.investopedia.com/articles/personal-finance/032814/pros-and-cons-using-airbnb.asp

3. Ryerson. (2016). Why tourists choose Airbnb over hotels. Ryerson University. Retrieved from https://www.ryerson.ca/news-events/news/2016/10/why-tourists-choose-airbnb-over-hotels/

4. Wilson, A. (2018). Guide to Airbnb vs Hotels for 2019. Skyscanner. Retrieved from https://www.skyscanner.com/tips-and-inspiration/hotels/airbnb-vs-hotels

5. Airbnb. (2020). How do I choose what type of place to stay?. Help Center. Retrieved from https://www.airbnb.com/help/article/5/how-do-i-choose-what-type-of-place-to-stay

6. Sobeh, M. (2019). The pros and cons of renting out the entire house. Mashvisor. Retrieved from https://www.mashvisor.com/blog/airbnb-income-property-vs-room/

7. Project Pro. (2020). How Data Science increased AirBnB's valuation to $25.5 bn? Retrieved from https://www.dezyre.com/article/how-data-science-increased-airbnbs-valuation-to-25-5-bn/199

8. Newman, R. (2015). How we scaled data science to all sides of Airbnb over 5 years of hypergrowth. Retrieved from https://venturebeat.com/2015/06/30/how-we-scaled-data-science-to-all-sides-of-airbnb-over-5-years-of-hypergrowth/

9. Pate, N. (2020). How Airbnb Uses Data Science to Improve Their Product and Marketing. Retrieved from https://neilpatel.com/blog/how-airbnb-uses-data-science/

10. Carrillo, G. (2019). Predicting Airbnb prices with machine learning and location data. Retrieved from https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-location-data-5c1e033d0a5a

Appendix A

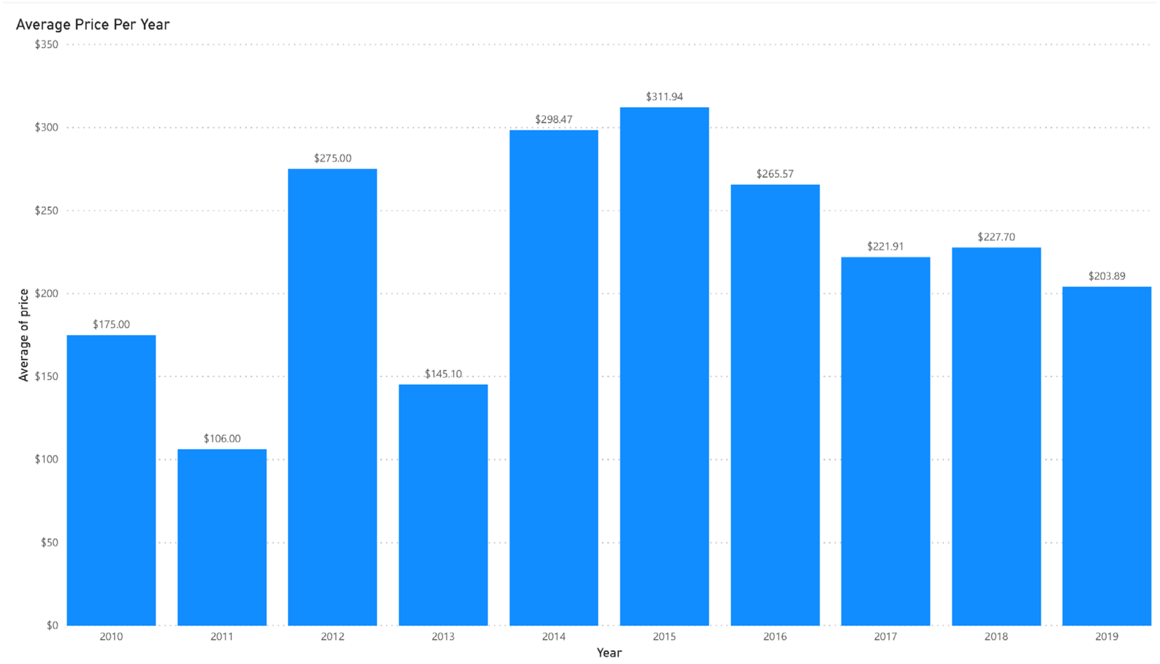


Figure 5: Price

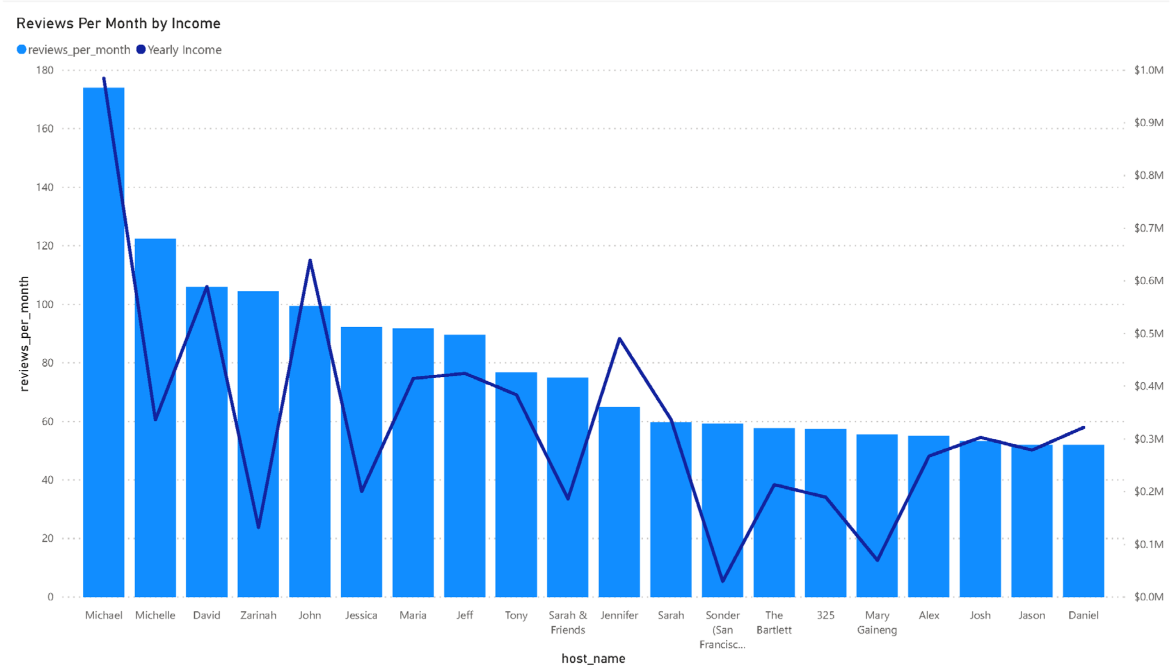


Figure 6: Reviews per Month

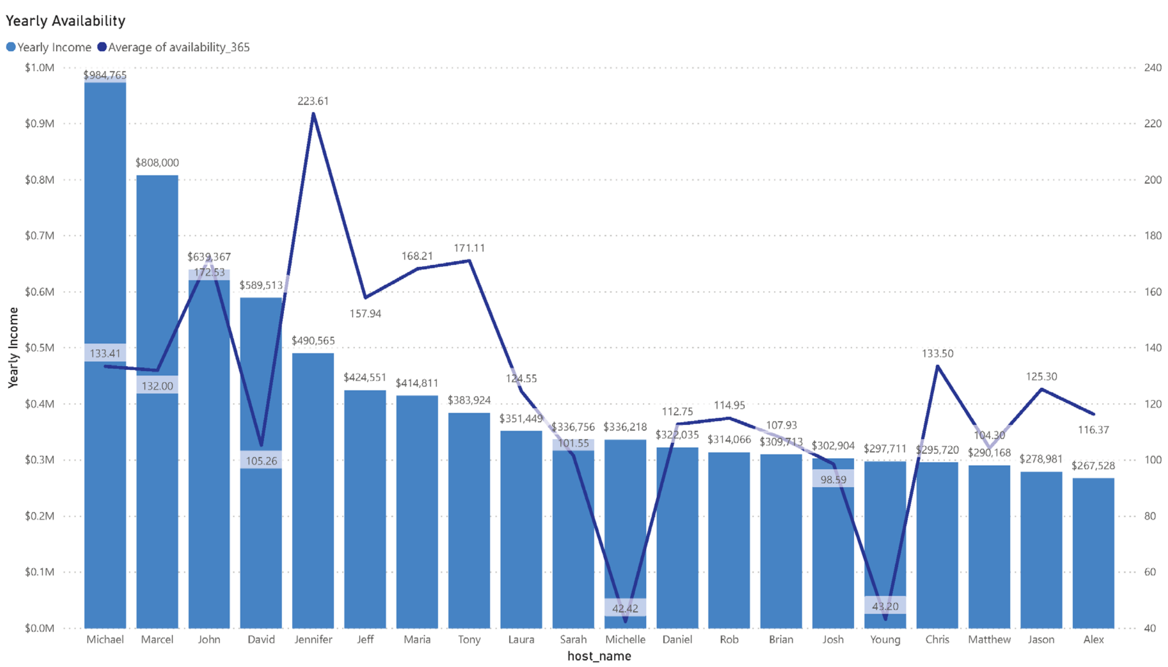


Figure 7: Availability

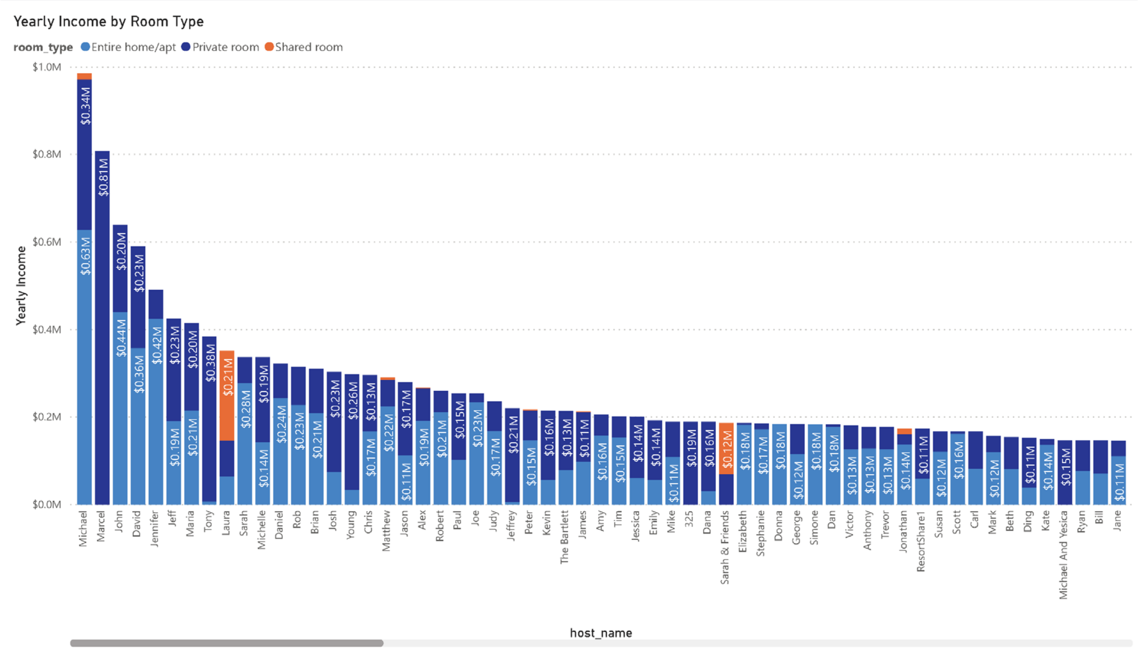
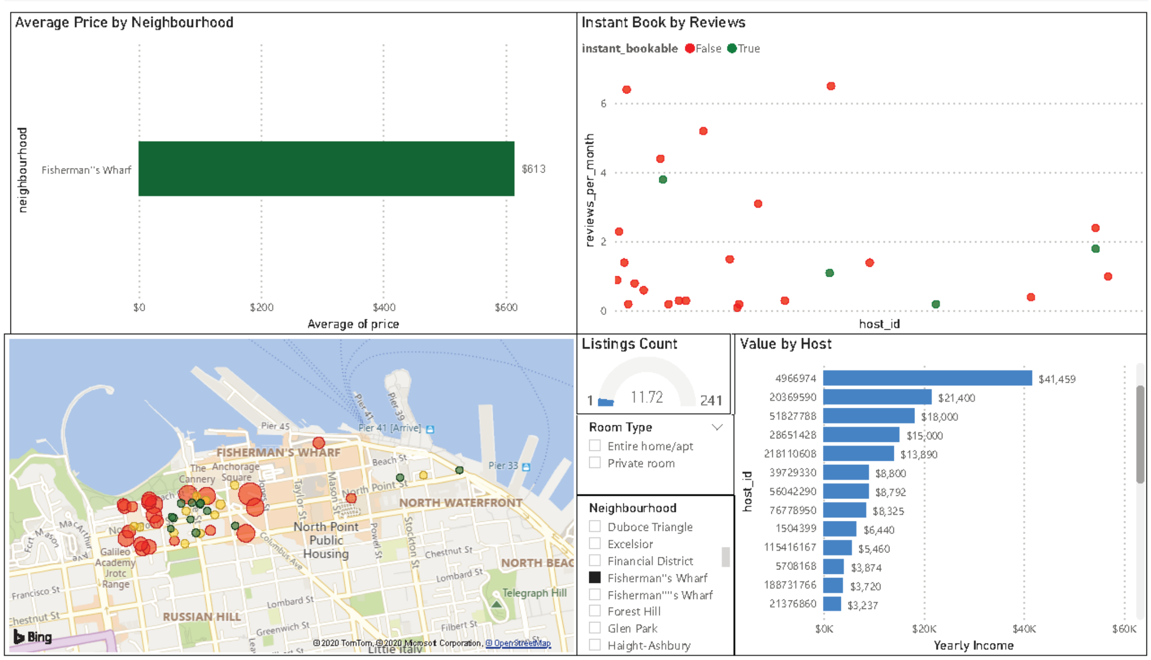


Figure 8: Room Type



Room Type

☐ Entire home/apt
☐ Private room

Neighbourhood

☐ Duboce Triangle
☐ Excelsior
☐ Financial District
■ Fisherman''s Wharf
☐ Fisherman''''s Wharf
☐ Forest Hill
☐ Glen Park
☐ Haight-Ashbury

Figure 9: Dashboard