Let Data Decide Your Next Destination

# **Kevin Angotti Winter 2020 https://github.com/kangotti/DSC-680**

# Which Domain?

The data I will be using is coming from Data World. The site provides a fair amount of data sets for the travel industry. Many of the data sets here are all organized and relatively easy to navigate for understanding.

1. <https://www.thefrugalgene.com/airbnb-popular/>

Help make sense of why this company has become a go-to for lodging searches.

2. <https://www.investopedia.com/articles/personal-finance/032814/pros-and-cons-using-airbnb.asp>

More information and helps weigh the pros and cons of using Airbnb over traditional sites.

3. <https://www.ryerson.ca/news-events/news/2016/10/why-tourists-choose-airbnb-over-hotels/>

What's driving customers to Airbnb? What are they doing right?

4. <https://www.skyscanner.com/tips-and-inspiration/hotels/airbnb-vs-hotels>

Additional compare and contrast to hotels.

5. <https://www.airbnb.com/help/article/5/how-do-i-choose-what-type-of-place-to-stay>

Are travels renting a whole property or just rooms, how to determine which is better?

6. <https://www.mashvisor.com/blog/airbnb-income-property-vs-room/>

Additional help on the stay choice, room, or whole place.

7. <https://www.dezyre.com/article/how-data-science-increased-airbnbs-valuation-to-25-5-bn/199>

This will help explain the data science side of things and why it is a useful tool for the future.

8. <https://venturebeat.com/2015/06/30/how-we-scaled-data-science-to-all-sides-of-airbnb-over-5-years-of-hypergrowth/>

Additional information on the data science side of things.

9. <https://neilpatel.com/blog/how-airbnb-uses-data-science/>

What has the company done already, and how can we expand. What have they not thought of yet?

10. <https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-location-data-5c1e033d0a5a>

Airbnb currently uses ML for pricing; using that train of thought, we can use ML for choosing locations or hosts to book.

# Which Data?

I will be examining the city of San Francisco's Airbnb listing data.

<https://data.world/ajsanne/sf-airbnb-listings>

This dataset provides 107 variables with over 7500 rows of data that can be used to make predictions or draw up some conclusions about listings in a particular part of the city or area. Within the dataset, key variables such as host, location, and review rating stand out in the data and can be used in different ways for making predictions.

# Research Questions? Benefits? Why analyze these data?

Which hosts are the busiest in the San Francisco area?

Does the price of the booking have a correlation to review rating?

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?

The main draw for looking into Airbnb data is to determine if there are better options than your traditional hotel bookings. Will there be just as many available places as you would find during a traditional hotel search? Will the returned search for a particular property be of quality or money well-spent feeling. Customer reviews will have a large impact on the analysis; 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 play 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.

Creating a project on this type of data and providing prediction for value and positive host properties, I feel it can then be used to create other tools to help travelers look outside the standard booking methods.

# What Method?

I want to start with regression analysis, dive into either cluster analysis or time series analysis, or even use all three. By predicting which properties are affordable and appealing to stay in, more people can choose this type of lodging over hotels.

Here, the primary thought process is to create a type of prediction tool that can help future customers with more options when searching for places to stay. Starting with the models, if successful, the next steps portion of the project could be implementing the models into a usable website or calculator for a customer to use when booking in the future.

# Potential Issues?

The main issue I see that could arise is only having 7500 rows of data. If I use time series, that may not be enough data to make a good prediction. Other areas that could be an issue are targeting the review rating and host variables for the bulk of the analysis. Many articles describe those are the main focal points of Airbnb data; however, I am not sure if that will translate to a data analysis project.

Potential issues that can cause the project to delay or potentially fail rest in the modeling phase. If the models do not work or run correctly, the whole project can become delayed. With the project's timeline already cut by 1/3 of the standard twelve-week timeline, the potential for delay increases as any pitfalls arise.

# Concluding Remarks

The hospitality industry is a large and lucrative area. Most people that are traveling tend to stay at a hotel of some sort. Airbnb has become a hot topic these past few years, and with the company going public and their IPO dropping sometime in December, this can become a good data analysis project.

The potential for obtaining predictions on more than just traditional hotel chains will give future customers a useful tool to aid in their lodging search before traveling. The use of regression and time series can provide predictions on how nice a booking property is and how often a said property has been booked over time.