Inside AIRBNB

Customer Sentiments, Networks, and Rental Forecast

Prepared by: Prashant Sanghal

Team Project Goal: Use Airbnb dataset (listings, reviews, neighborhoods) to predict rental prices in Seattle, New York and San Francisco area.

Teamwork: To do this we looked at topic modeling, sentiment analysis, network analysis, time series, regression analysis, and added search functionality to help users locate their next AirBnB rental with an estimated rental cost for the next one year.

Users can and also customize their preferences in the regression model to see predicted prices as well as interact with the dashboard to explore what's available in the neighborhood.

Workflow & Activities:

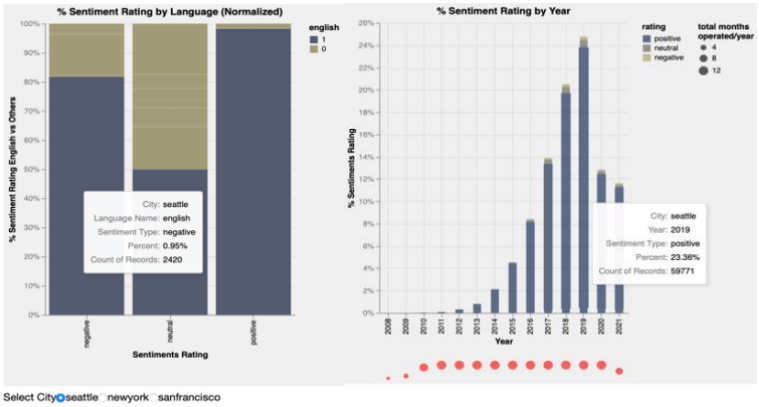
Data Preparation → **Feature Engineering** → **Modelling & Evaluation** → **Interactive Dashboard**

Text pre-processing *Exploratory data analysis* *Topic Modeling* *Random Forest Regressor* *Handling Missing values* ***Sentiment Analysis*** *Ridge/Lasso* *Outliers removal* ***Network Analysis*** *Grid Tuning & Evaluation* ***Time Series Forecasting***

Let's dive-in a few of these workflow activities.

1. Sentiment Analysis:

This notebook explores the shift in Air BnB customer sentiments by language and year when they left a review. As shown in the plot, we can see the percentage breakdown of sentiments ratings in English Vs other (50+ languages) from 2008 to current time frame. As an example, just in Seattle area, positive sentiments were pretty high at 96.5%, which is indicative of many happy customers and continued business growth that BnB has enjoyed over all these years. In fact, last year in 2020 March when the pandemic was announced, it was interesting to note that the decline shown in graph by years between 2019 to 2021, was due to lower booking counts while the rate of positive customer sentiments remained unchanged, almost hitting 2017 levels again.

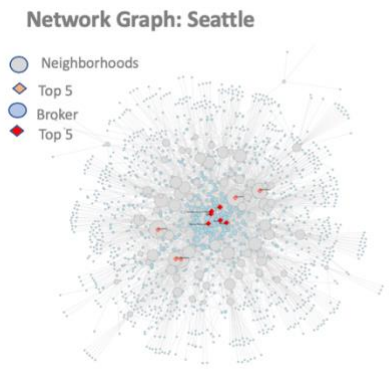


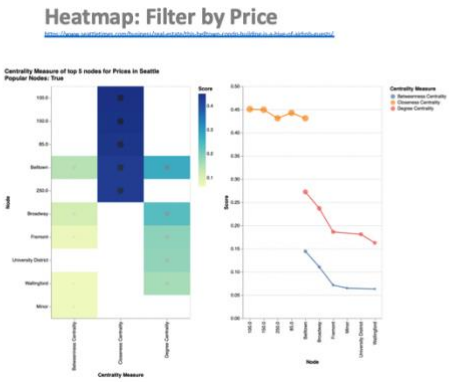
We found a similar pattern in other cities also, and so, were curious why is Air BnB so successful? We will try answer this question using network analysis, which takes us to our next notebook.

Failure in Sentiment Analysis: Since, positive sentiments account for more than 96%, we were wondering if there is perhaps a presence of sample bias in the dataset. It could be possible due to the way customer survey was designed to prompt customers to leave more positive reviews than negative, which was not explored as a part of this study.

2. Network Analysis:

Why Air BnB is so successful, was the key question we wanted to explore. In this notebook, we explored a network graph between neighborhoods and other features such as brokers, price, availability and customer sentiments in all three cities. To keep network graph readable, we chose to show only top 5 observations in the graph. As an example, this network graph belongs to Seattle area. The grey nodes are the neighborhoods and blue nodes are the brokers. A bigger neighborhood node represents high degree, which means a neighborhood has multiple brokers listings in the area. We have built a similar network graph for other features such as price, availability, customer sentiments and have explored network degree relationships using both popular and unpopular nodes, which helped us think critically "why a $2000/day rental in Seattle area would be considered a unique listing compared to all other listings"





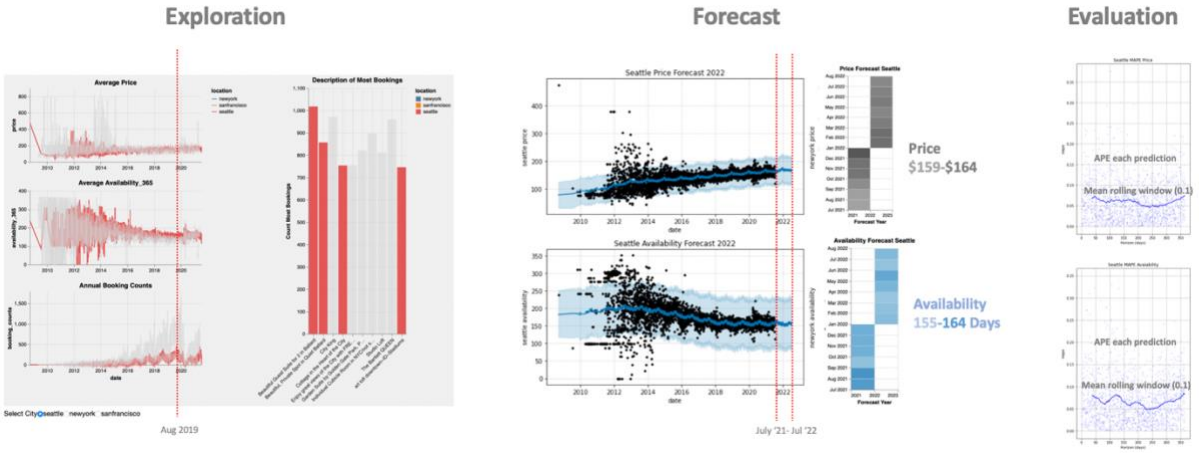
However, it can be hard to answer many such questions directly from the network graph itself, so, we built a heat map, where we can see variation in 3 centrality measures (degree, closeness and betweenness) by price/broker/availability or other features to get deeper insights. Example, a popular price range in Seattle is $85- $250, and popular neighborhoods are 'Belltown', 'Broadway', and popular brokers are Turnkey Vacation rentals, and vacasa washington.

Then, we wanted to investigate broker effectiveness, and found positive correlation between edge betweenness and closeness centrality as shown in the scatter plot above. What it means is that, that BnB might be using well connected community brokers or super hosts (as referred in the dataset) to build highly close-knit neighborhoods where customers like to stay. This pattern can be seen in all 3 cities and is likely the reason behind 96% positive sentiments as noted earlier during our sentiment analysis. In order to investigate this further, we plotted a network graph between neighborhoods and customer sentiments to see where we can probably empower community brokers to overcome negative customer sentiments and increase revenue potential for Air BnB business.

Failure in Network Analysis: However, one drawback related to relationship between individual & property managers with respect to customer sentiments was a 'mixed bag'. Example, brokers with fewer listing between 1 to 4 showed high customer sentiments as well as the property managers with multiple listings between (100 to 300). This trend, however was not found in property manager with higher listings (above 300) in the network graph. Due to long runtime, we did not reproduce the network graph for this analysis in the notebook, but it can be re-produced by calling the function "build_network" in the notebook and replacing "source_column" with 'host_listings_count", where we assumed that high listing counts would most likely be managed by the property managers. The other area we wanted to explore was the success stories behind individual brokers (1 to 4) who were able to generate very high customers sentiment (closer to 1). Maybe it would be possible to use these nodes as a reference for other broker nodes for information sharing and process enhancement.

Time Series Analysis

We used time series analysis to help estimate the likely cost and availability of an AirBnB rental from Aug 2021 to Aug 2022. To build this forecast, in the initial exploration stage, we noticed a high variation in the average prices and availability from 2011 to 2016, which was indicative of the presence of outliers in the time series dataset. The forecast plot containing black dots show the actual data points, blue line shows the model forecast with 95% confidence interval and black dots outside of the blue area shows outliers,



which were removed from the training data. We chose a more recent time period instead, beyond 2016-present, to train our time series model. The evaluation plot on the extreme right, shows the mean absolute percentage error which means if you plan to stay in Seattle right now (August 2021), you would be able to find a property that has an availability of 155 days and has an average cost of $159 +- 6% to 7% (Mean Absolute Percentage Error) which is calculated based on a 30-day rolling window for the next 365 days.

The bar plot shows the description of top 10 most booked rentals in Seattle, New York and San Francisco area, in case you are planning a visit.

Failure in Time Series Analysis: There could be a possibility of post pandemic cost overheads and sudden increase in rental demand possibly pushing the forecasted availability and prices higher than the estimated MAPE score as shown in the forecasted model above. For example, a 155 days availability could get all booked up fast, if all overseas travel restrictions are lifted in these cities in the near future. This will also have a significant increase in the rental prices for Air BnB properties, which our model currently does not account for, in the forecast.