Analysis of Airbnb's Rental Scene

Annie Chen, Mariana Curiel, Monique Duong, Usmon Muslimbekov
Department of Information Systems, California State University
Los Angeles
Jongwook Woo
12/13/2019

Abstract: In this project, we will analyze Los Angeles and San Francisco data from *InsideAirbnb*, which is an independent, noncommercial open-source data tool, to understand the overall rental marketplace/landscape by applying different measurements to produce static and interactive visualizations.

With this analysis, you can answer fundamental questions about Airbnb within the Los Angeles and San Francisco region. Questions include:

- Which city is the most popular in Los Angeles/San Francisco?
- Which city is the most expensive in Los Angeles/San Francisco?
- Which city is the most affordable in Los Angeles/San Francisco?
- How has Airbnb grown in terms of popularity?
- Is there seasonality based on the reviews? In other words, is there a specific season that people rent Airbnbs in Los Angeles?

1. Introduction

1.1 Airbnb background

Airbnb is an online business where owners can list the spaces they want to rent on the Airbnb website and connect with users who are willing to rent them. Airbnb's roots started in San Francisco in 2007, when roommates Brian Chesky and Joe Gebbia couldn't afford rent and had the idea of renting out spaces with three air mattresses to visitors in their apartment [1].

After finding their idea to be lucrative, they contacted Nathan Blecharczyck who helped them formally start and found their business in August of 2008 [1]. Since then, Airbnb has grown to have more than 7 million unique listings worldwide and become available in over 191 countries. It's also estimated to have around 150 million users worldwide along with a growth rate of 153% in the past ten years [2].

1.2 Why we chose Airbnb

We decided to choose the topic of Airbnb due to the scale that it has grown[1] over the years, becoming a worldwide open platform that lets frequent travelers easily rent homes online versus traditional hotel renting [2]. As this online marketplace for arranging to lodge continues to grow, we thought that it would be ideal to analyze the data of Los Angeles and San Francisco, two of the most popular cities that users rent from. In doing so, we can get an understanding of the rental landscape and be able to closely look at key figures representing which city is more popular and more expensive to book rentals in.

1.3 Why our work is important

We think that our work is important because it provides an in-depth overview of how Airbnb is being used and how it is affecting the neighborhoods of cities across the nation. A key importance of our project is that it enables us to better analyze the data that is around us at all times. Companies are constantly collecting data from our phones, apps, and daily habits and using that information to sell products to us. Applying what we learned from our work to different situations allow us to understand the data and get an insight into how our data is being used.

2. Related Work

The related data analyses we found were about Boston and Seattle Airbnb. The biggest difference between this analysis and ours was their use of the Anaconda distribution version of Python in Boston Airbnb and Python 3.x, Python libraries, and iPython notebook for Seattle Airbnb. The 3 objectives in the Boston data analysis were: determining whether the price of a listing can be predicted based on a property's features, finding the best location to invest in Airbnb in Boston for maximum return rates and if reviews can be predicted according to different characteristics of property [3]. The objectives of the Seattle analysis were to search for; any trends in the rental of spaces according to each season, look at what type of spaces are popular to rent on Airbnb, and how the features in each type of space can influence pricing [4].

[1] Bookings in the US have grown 45% each year according to Airbnb Statistics on IPropertyManagement.

Different visualization methods[2] were used in both analyses as well. Additionally, different measures[3] were also used in both analyses[4].

3. Background/existing work

3.1 Background of our work

After looking through many data sets, we decided on choosing Airbnb as our data source. The background of our work was to utilize the data provided by *InsideAirbnb* which was analyzed, cleansed and aggregated for the public. In order to make the data easier to view, Airbnb categorizes its data by county. This data is then divided into neighborhoods which further simplifies the process of analyzing the data. The two key files that we worked with include listings which contain detailed listings data and reviews, which show the reviews for each of the listings. Even though these files aren't very large, they contain a wide range of useful information to be able to analyze key attributes.

When analyzing the listings, some of the key attributes we used include listing_id, neighborhoods, latitude, longitude, reviews_per_month, and price. As for the reviews file, the key attributes were listing_id, date, reviewer_id, and reviewer_name. Both Los Angeles and San Francisco used the same attributes which made the analysis easier to understand and allowed us to discover useful information about Airbnb.

4. Our work

4.1 Measurements we used to gather insight

We analyzed the rental landscape within both Los Angeles and San Francisco using the measure, Price, by the dimension, Neighborhood. From these measurements, we can gather insight on which cities are more expensive and which cities are more affordable.

Similarly, we analyzed popularity within both Los Angeles and San Francisco using the measure, reviews_per_month, by the dimension, Neighborhood. From these measurements, we can gather insight on which neighborhoods within both Los Angeles/San Francisco are most popular to rent an Airbnb according to their reviews per month.

Lastly, we analyzed seasonality within Los Angeles using the measure, reviews_per_month, by the time dimension, date. From this analysis, we want to see if there is growth in Airbnb usage by looking at the years of reviews from 2009 to 2019. In addition, we want to see if there is a specific season where people tend to rent an Airbnb in Los Angeles, so we filtered the reviewer date by Quarter, Year.

4.2 Specifications

In our specifications, we used AWS to analyze our data. We used 1 node and the node name is ip-10-7-251-11.us-west-2.compute.internal. The release label is emr-5.27.0. The Hadoop Distribution is Amazon 2.8.5. And our CPU speed is 2.50 GHz. After analysis, we generated 2 worksheets that has a total of 124 MB. The applications we used for analysis were Apache Pig, Microsoft Excel 3D Maps, and Tableau.

4.3 Flow Chart of Data Analysis

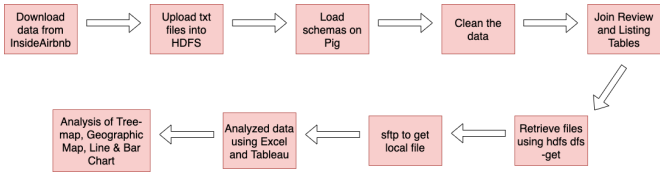

Figure 1. Flow chart of our Data Analysis

Figure 1 illustrates the entire process of our project. It begins by downloading the listings and reviews files from the Los Angeles and San Francisco region, manually from our Github, which is available in the References. Then, we import the necessary Airbnb CSV files into our local by performing the wget command.

Before we start editing the data, it needs to be put in the correct directories for organizational purposes. We have a directory for each region, one for Los Angeles and one for San Francisco. After that, we created 4 schemas: la_listings, la_reviews, sf_listings, and sf_reviews and loaded them on Grunt Shell.

Within our analysis, we cleansed the data by eliminating any unnecessary fields such as host_id, host_name, and comments by using the FOREACH GENERATE operator. We used the GROUP BY operator to group price by neighborhood. From there, we used the FILTER BY operator to filter listings that contain a location and reviews. From this, we gained insight into the top 15 most expensive cities by using the ORDER BY operator on the field, Price.

Aside from finding out which cities are the most expensive or affordable in both Los Angeles and San Francisco, we also want to produce 2 worksheets that join la_listings with la_reviews and sf_listings with sf_reviews. After joining, we used FOREACH GENERATE operator to eliminate the extra listing_id column. From there, we stored the joined relations in HDFS and then retrieved the files using the get command. After retrieving the local files using sftp,

[2] The tools used to visualize Seattle's findings were line charts, bar charts, and heatmaps, tools used for Boston were pie charts, bar charts, and heat maps.

[3] Measures in Boston Data Analysis were average price per neighborhood and ratings per neighborhood.

[4] Measures in Seattle Data Analysis include instant bookable and cancelation policy.

we loaded the CSV files into Excel and Tableau to generate our static and interactive visualizations.

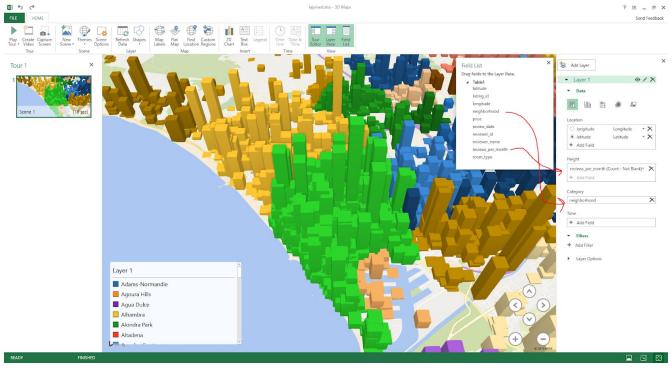

Figure 2.1 Popularity: Reviews by Listings in Los Angeles

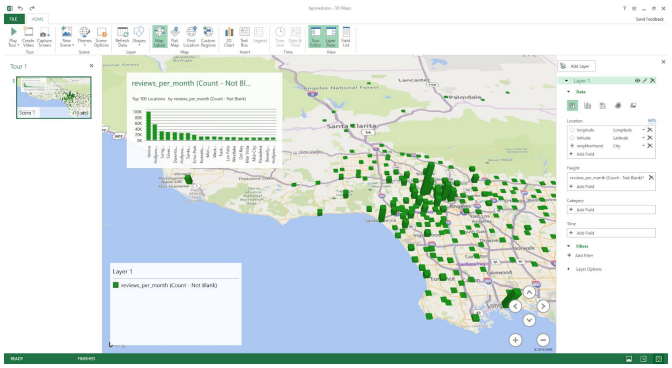

Figure 2.2 Total amount of reviews per general neighborhood in Los Angeles

In Figure 2.1, we analyzed Los Angeles data to get a visualization of the most and least popular Airbnb listings using the dimension, Listings, by the measure, reviews_per_month. It is color coded by the dimension, Neighborhood, and the 3D Maps shows a "bar" per listing. Figure 2.2 analyzes the total amount of reviews per general neighborhood in the Los Angeles county.

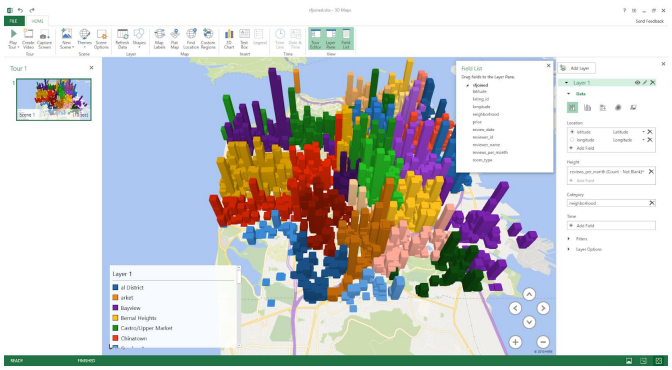

Figure 3. Popularity: Reviews by Listings in San Francisco

Similarly, in Figure 3, we performed the same measurements for San Francisco. The bars height indicate the amount of reviews each specific listing has. The amount of reviews that a listing has determines how popular the Airbnb is because more reviews means that more people booked it. The bars are color coded for the user to easily distinguish the differences between the neighborhoods in the county.

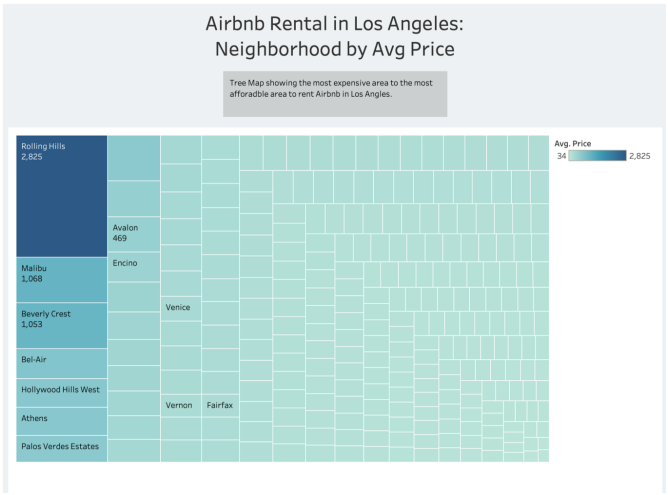

Figure 4. Price: Neighborhood by Avg Price in Los Angeles

In Figure 4, we used Tableau to create interactive visualizations, analyzing the rental landscape using the dimension, Neighborhood, by the measure, Average Price. With Los Angeles, we used a Treemap to illustrate the most expensive area to the most affordable area to rent Airbnb. We used color and size to represent the average of price. So the darker the box, the more expensive it is. Similarly, the bigger the box, the more expensive it is.

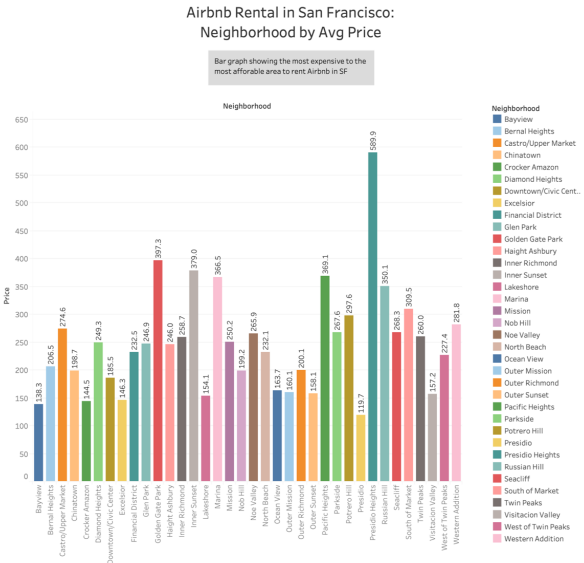

Figure 5. Price: Neighborhood by Avg Price in San Francisco

In Figure 5, we also used Tableau to create a Bar Chart to illustrate the most expensive area to the most affordable to rent in San Francisco. We analyzed the rental landscape using the dimension, Neighborhood, by the measure, Average Price, in San Francisco. In this bar chart, the colors show details about various cities within San Francisco. The marks on top of the bars are the average price.

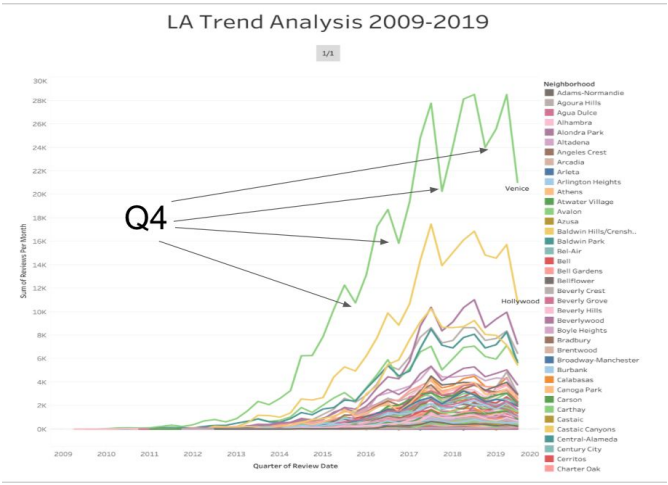

Figure 6. Seasonality: Sum of Reviews by Date in Los Angeles

In Figure 6, we created a line chart that illustrates how popular Airbnb has become from 2009 to 2019. In our visualization, it is using the measure, the SUM of the number of reviews by the time dimension, review date. Through this chart, we are able to answer the question of whether there is growth in Airbnb usage in Los Angeles.

4.4 Data Source URL

In our Github, it contains the presentation slides, flow chart of our data analysis, datasets, pig script, and the tutorial [6].

5. Conclusion

We analyzed the data from *InsideAirbnb* using Apache Pig and was able to perform further analysis using Excel and Tableau. From this, we were able to draw conclusions on the questions that were asked in our abstract.

It was concluded that popular beaches in Los Angeles County, such as Venice Beach, were booked the most. An insight from the San Francisco analysis was that the neighborhood, Mission, has the most reviews about their sites. With Mission being in downtown San Francisco, we concluded that the downtown areas of both counties were popular with bookings. The 3D graphs accurately confirmed our predictions about bookings being more popular in congested areas such as the coast and Downtown.

We created interactive visualizations using Tableau to answer the questions of which areas are most expensive in Los Angeles and San Francisco. In Los Angeles county, we concluded that Rollings Hills, Malibu, and Beverly Crest were the most expensive cities to rent Airbnbs. In San Francisco, the most expensive city to rent Airbnbs is Presidio Heights.

Our analysis also gave us an understanding of the Los Angeles rental scene and how popular Airbnb has become from 2009 to 2019. As you can see in Figure 6, we conclude that there has been tremendous growth within Airbnb usage between the years of 2009 to 2019 in Los Angeles. We used the sum of the number of reviews by review date and noticed that there were consistent drops within Quarter 4, which is between the months of October to December from 2016 to 2019. This provides insight into how people rent Airbnb less in Los Angeles from October to December.

From this project, we learned that the data engineering phase was difficult as it took many trials and errors. We also realized that there were many tools available to use for analyzing the finalized data such as Excel, Tableau and R Data.

One key takeaway is that Airbnb has significantly grown and it's apparent by our reviews trend analysis that they continue to dominate the rental market and disrupt the hotel industry around the world.

References

[1] Aydin, Rebecca. *How 3 Guys Turned Renting Air Mattresses in Their Apartment Into a $31 Billion Company, Airbnb*. 20 Sept. 2019, https://www.businessinsider.com/how-airbnb-was-founded-a-visual-history-2016-2.

[2] Bustamante, Jaleesa. *Airbnb Statistics: User & Market Growth Data* [November 2019]. 7 Nov. 2019, https://ipropertymanagement.com/airbnb-statistics.

[3] Tang, Lucas Bo. "Project: Seattle Airbnb." *GitHub*, 16 Feb. 2019, https://github.com/LucasBoTang/Project_Seattle_Airbnb/blob/master/seattle_airbnb.ipynb.

[4] Gudapati, Susmitha. "Boston Airbnb Data Analysis." *GitHub*,25 Feb. 2019, https://github.com/susmithagudapati/Boston-Airbnb-Data-Analysis.

[5] Cox, Murray. "Inside Airbnb. Adding Data to the Debate." *Inside Airbnb*, 2019, http://insideairbnb.com/.

[6] https://github.com/anniechen61/CIS4560/