Citadel Datathon - Team 7 Submission

Bradley Dice, Shannon Moran, Vyas Ramasubramani, Pengji Zhou

Topic question

How do Airbnb densities compare with the rent in their area, socioeconomic diversity in a ZIP code, and income levels in that ZIP code?

Non-technical executive summary

- Key findings: Airbnbs are popular in "city-center" regions, which command high rent. Socioeconomically diverse regions are often nearby regions with a higher density of Airbnbs, suggesting that outside of tourist-heavy regions, areas with high socioeconomic diversity may be areas that attract the type of folks that would be interested in hosting an Airbnb.
- Further work: There are a number of interesting directions we could pursue.
    - How do locations of Airbnb density compare with hotels? Is Airbnb competing with hotels on location?
    - Gentrification - could we use booking data and the emergence of new Airbnb locations, coupled with growth in Zillow Rent Index (ZRI) to predict either Airbnb rental locations or rental values over time?

Technical executive summary

- Question: What are the dominant factors contributing to the variable popularity of Airbnb in different regions?
- Approach: We used principal component analysis and geospatial analysis to understand correlations. Our approach suggests strongly that property value is the dominant factor in determining Airbnb popularity. The most popular Airbnb region (the 25% most popular zip codes for Airbnb) are clearly distinguished from the remainder of zip codes by these analyses. For the less popular regions, an interesting direction to look into would be the venues available and how they contributed to the relative popularity of these zip codes.
- Further work: We were unable to match location venues to the Airbnb locations in a computationally efficient manner. We would like to extend our analysis in this way to further identify the contributions that distinguish Airbnb popularity in less popular regions. Additionally, there are a number of interesting business questions we could explore further, as detailed in the "Further work" section in this report.

Topic and Background

We were provided with a set of current listing data from Airbnb for Los Angeles (~30k listings), as a well as several smaller cities in the U.S. south (Asheville: ~800, Austin: ~9.5k, Nashville: ~3.2k, and New Orleans: ~5.3k). We planned to analyze the Los Angeles listings to then see if we could predict the distribution of Airbnbs in smaller cities.

Given a set of Airbnb locations in a metropolitan area, we ask the question,
How do Airbnb densities compare with the rent in their area, socioeconomic diversity in a ZIP code, and income levels in that ZIP code?

Airbnb was originally founded to serve as an extra source of income for its founders to be able to afford their rent in an expensive city. Today, many hosts on their platform seek to solve that same problem by subletting their apartments out while not using them. However, Airbnb has also begun to compete with the traditional hotel market. We hypothesized that one marker of this will be that higher densities of Airbnbs exist where there is more for a visiting tourist to do.

Why do we care about this question? We can think about this from several perspectives.

Corporate strategy

As Airbnb looks to expand from a way of connecting hosts and interested visitors and move to taking on hotels, a natural question is: how much saturation can you expect in a given market? By combining data on tourism with key tourism sites (here, we use "nightlife" as a proxy), a hypothetical Airbnb saturation can be calculated.

City expansion teams

Within a city, regional Airbnb managers may be interested in understanding where within a city more locations are needed, relative to where renters want to stay.

From the perspective of supply and demand, we can imagine using a zip code or address "tourist location score" to optimize a pricing strategy to ensure full capacity, as the hotel industry does to maintain keep its rooms at capacity.

Airbnb host relations

While many Airbnb hosts genuinely do rent out their own space, Airbnb's growth has been accompanied by the rise of the hosts that manage multiple locations. One point of potential interest, then, is choosing location to offer a new Airbnb location. Assuming that Airbnbs have already clustered in response to demand, understanding correlations behind where current Airbnbs are located may allow a prospective Airbnb host where to open a new location.

Exploration

Our initial explorations proceeded along multiple lines.

Time series data

Initially, we hoped to use information about the availability of the Airbnbs over time as an indicator of when specific areas became popular. In doing so, we made two key assumptions that proved to be mistaken. Firstly, we assumed that the availability data also provided information about when a particular listing was actually being rented out, not simply when it was *available* to be rented out. Secondly, we assumed that listing information was actually available for the full 10 year 2007-2017 time span as indicated in the schema. Since neither of these assumptions proved true, we were unable to effectively test our hypotheses since the calendar data did not overlap the data for which demographic data were provided (2011-2015) and the lack of actual Airbnb utilization precluded a clear analysis of how much Airbnbs were utilized in a particular area.

As a result, we pivoted to focusing on a study of how Airbnb popularity correlated with the presence of nearby popular venues, as well as what determined the success of Airbnbs outside peak zones. The first step in our analysis was cleaning the data. For the most part this consisted of converting string lists of columns into proper lists and the converting those into columns, but we also had to do some formatting of dates and zip codes. It was in the process of completing this cleaning that we noticed that the calendar data was not as extensive as expected. Having completed this task, we proceeded along two fronts: principal component analysis and geospatial analysis.

Key findings

Principal Component Analysis

To identify the characteristics of a region that make it a hotbed for Airbnb listings, we distinguished regions by ZIP code and studied Airbnb performance and its correlation with other economic or demographic factors in each of these ZIP code regions. We first identified five dimensions that can be obtained by the provided data sets to construct a multi-dimensional data set for each ZIP code region, these includes:

- Airbnb average price
- Gini coefficient
- Median Household Income
- Airbnbs per 10,000 Households
- Zillow Rent Index

With these multi-dimensional data constructed around each ZIP code region, we then performed Principal Component Analysis (PCA) to identify the dominant factors that makes becoming an Airbnb Host favorable in that region. PCA is a linear dimensionality reduction technique that helps identify which dimensions are most relevant in separating the data, where separation is defined according to the variance in the data set along this dimension.

The results of our PCA indicated that the most important indicators of regions where properties were likely to be listed on Airbnb were related to property value. In particular, the first principal component, which explained 30% of the variance in the data, was dominated by the Zillow Rent Index and the average price of the Airbnb. Both of these dimensions are strong indicators of property value, which we would expect to correlate with both location and higher quality housing. The second principal component is primarily dominated by the actual density of Airbnbs. Although the composition of the first principal component is unsurprising in and of itself, it is notable that the classes we identify in our PCA plot are not necessarily well separated along our principal components; rather, the different classes are best identified by the amount of spread along these dimensions. For example, the lack of spread in the red dots in **Figure 1** suggests that areas with low property values are almost always areas that Airbnb can ignore. However, the fact that some yellow dots cluster close to points of other colors on the plot indicates that there are some areas that are popular Airbnb locations despite not being outstanding in these respects. Therefore, our PCA data can be used to identify areas that will be successful. In particular, areas with high property values will almost invariably be good places to rent out Airbnbs with a high expected revenue from renting, but areas with low property may or may not prove successful.

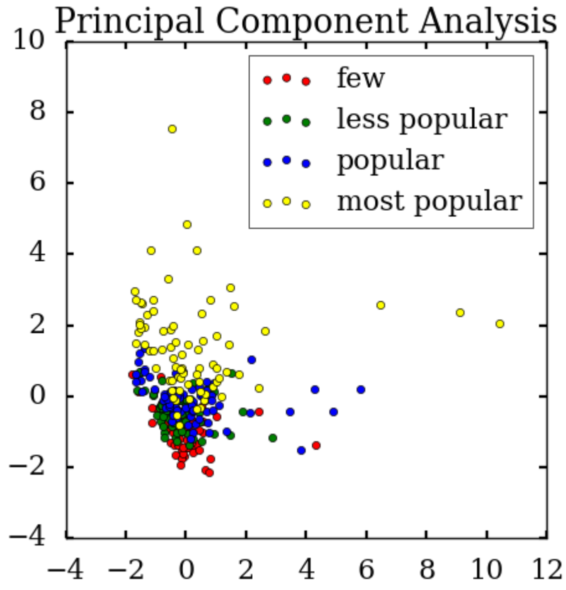


Figure 1, above, show the spread of Airbnb Listings in Los Angeles along the first 2 principal components.

Geographic Distribution of Airbnbs in Los Angeles

We seek to understand how the distribution of Airbnbs correlates with socioeconomic status and socioeconomic diversity in Los Angeles. We hypothesize that Airbnbs will be more common in areas with high socioeconomic diversity.

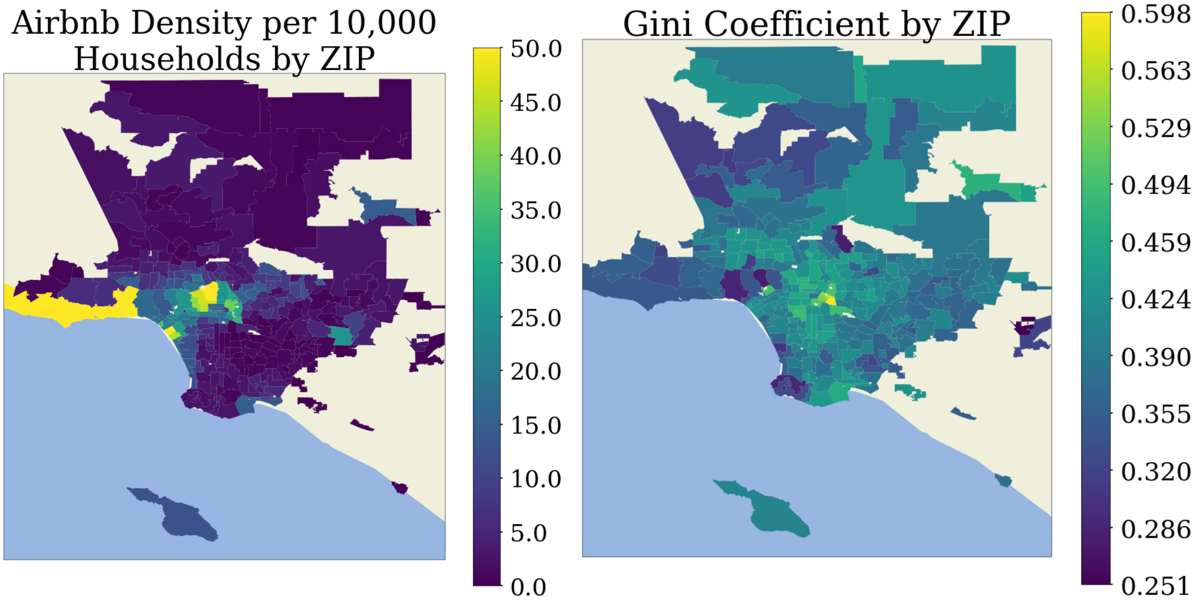
Gini coefficient

The Gini coefficient is a measure of statistical dispersion intended to represent the income or wealth distribution of a nation's residents. This is commonly used as a measure of income inequality. Specifically, this is described by:

$$G = \frac{1}{2\mu} \sum_{i=1}^{n} \sum_{j=1}^{n} f(y_i) f(y_j) |y_i - y_j|$$

$$\mu = \sum_{i=1}^{n} y_i f(y_i)$$

Equation 1, above, shows how the Gini coefficient is calculated. Here, y_i represents the mean income level of each bracket, and $f(y_i)$ represents the proportion belonging to that bracket. The top bracket had no upper bound ($200,000+), so a mean income level of $225,000 was used.



Figures 2 and 3, above left and right, show the geographic distributions of Airbnb Density and Gini Coefficient, respectively.[1]

[1] ZIP shapefiles came from the US Census Bureau's "ZIP Code Tabulation Areas." https://www.census.gov/geo/maps-data/data/cbf/cbf_zcta.html

We see that neighborhoods with the *highest* Airbnb density **(Figure 2)** and those with the *highest* Gini coefficient **(Figure 3)** do not overlap strongly. Neighborhoods near East Los Angeles are the most unequal while the downtown neighborhoods see the most Airbnbs. However, there is a clear pattern where downtown ZIP codes show relatively high Gini coefficients, Airbnb density, and mean income, suggesting that Airbnbs are largely hosted by upscale renters or lower-income renters in a high-income neighborhood. Furthermore, by comparing to mean income by ZIP **(Figure 4)**, we see that the areas with the *highest* income near Beverly Hills are not offering as many Airbnbs as their neighbors to the east (who have beachfront property) and west (who are located in the downtown area).

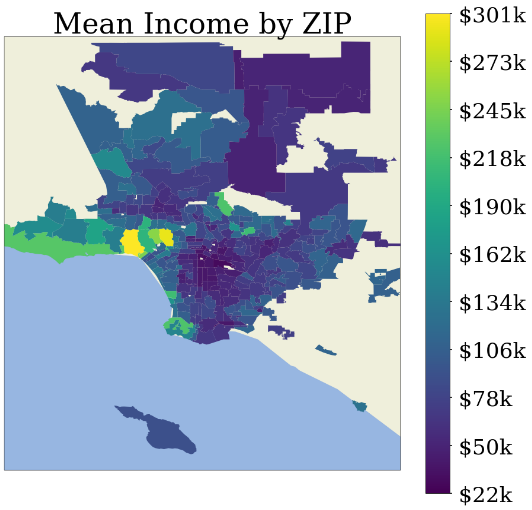


Figure 4, above, shows the mean income of ZIP codes in Los Angeles.

Next Steps and Future Directions

Apart from the economical and demographical factors that dominates the airbnb popularity in among different regions, we would also like to investigate the impact of the presence of nearby recreational activities. Especially interesting would be those venues that are typically frequented by tourists. We propose that a couple core facilities that are important for tourist including transportation venues (bus stations, airports, subway stations), nightlife venues(bars, night clubs, and casinos) and food venues (restaurants and cafes) may also be the other contributing factor that provide extra incentive for becoming a host on Airbnb. These extra dimensions might help us understand the popularity of Airbnb locations that are outside the city center. Moreover, these dimensions might distinguish regions outside the city center in our PCA plot.

We can also imagine exploring the following questions of interest: How do locations of Airbnb density compare with hotels? Is Airbnb competing with hotels on location? Additionally, could we use booking data and the emergence of new Airbnb locations, coupled with growth in ZRI to predict either Airbnb rental locations or rental values over time?