databnb

Introduction

Airbnb is an online marketplace that allows individuals to rent housing for homestays or person travel. We are interested in understanding the relationship between Airbnb listings and the real estate market in the United States. In particular, we wanted to address the following overarching question: how does the real estate market impact the number of Airbnb listings in a particular area?

Many have called for government regulation of Airbnb, complaining that the spread of Airbnb has gentrified residential neighborhoods. Thus, it is worthwhile for the government as well as Airbnb marketing itself to understand how Airbnbs are related to real estate value.

Hypotheses

We hypothesize that when the real estate market is successful in a particular neighborhood, the Airbnb market will also be successful in that neighborhood.

To address this hypothesis, we had to define the meaning of Airbnb “success” in a neighborhood. In terms of the Airbnb market, we defined “success” by the number of Airbnbs that are rented in a particular zip code. In terms of the real estate market, we used the average price of homes per zip code to define a “successful” real estate market. That is, we interpret more expensive houses in a given zip code as an indicator of a successful neighborhood.

Data

We collected our data from Inside Airbnb and Zillow. Inside Airbnb has periodically web scraped Airbnb since 2015 and made this data publicly available. Our project looks at data from January 2015 to January 2020 from the following cities: Asheville, Austin, Boston, Chicago, Denver, Los Angeles, Nashville, New York, San Francisco and Seattle.

Zillow is the leading online real estate marketplace in the United States. Using their own real estate listing information, they have generated and published Zillow estimates, or “zestimates”, of average housing and rental prices per zip code in the United States over the last few years.

Challenges

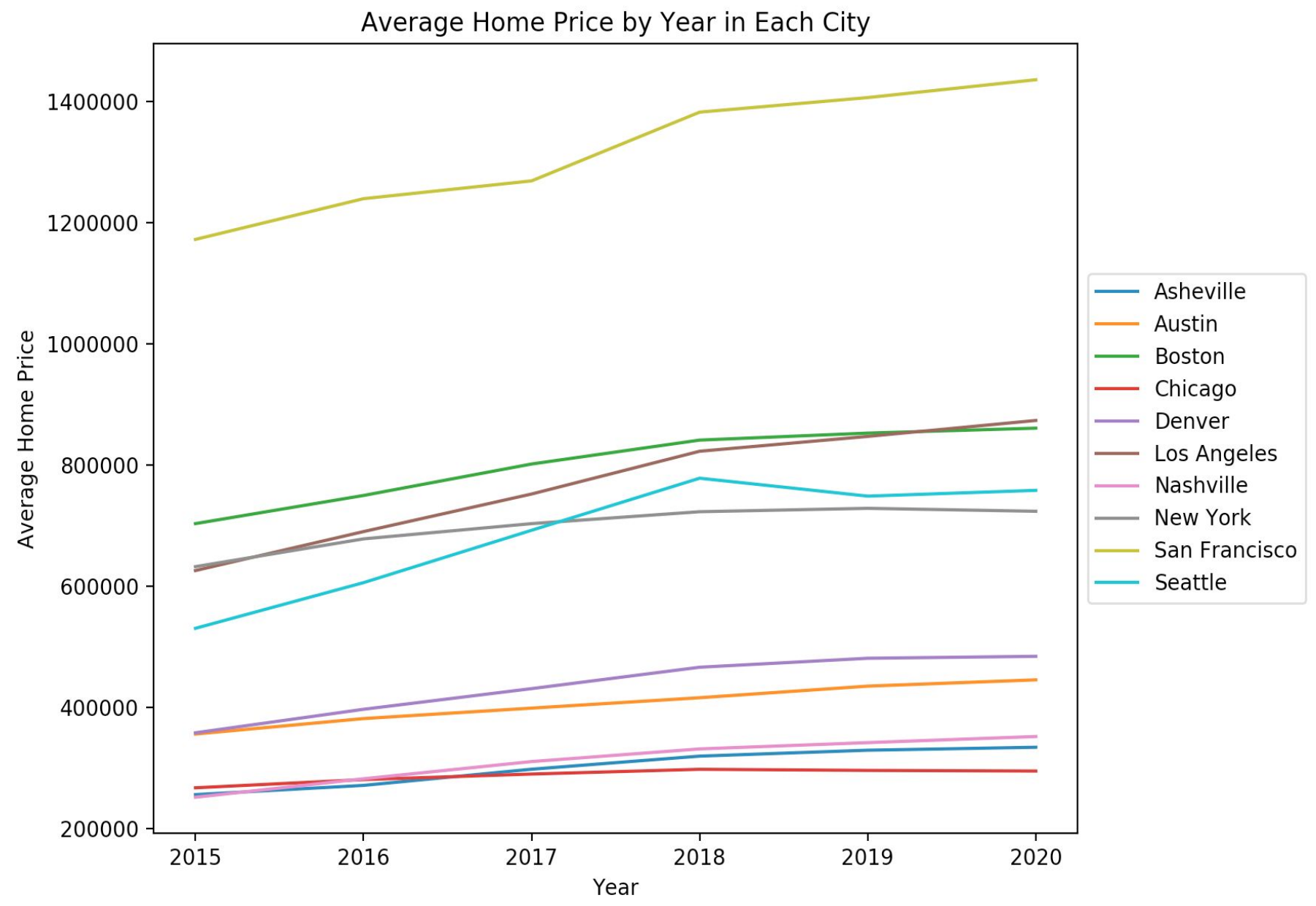
When constructing our data set, our New York and Los Angeles data files were too large to run on our computers, push to our Github or load into our data set using Google Drive links. Eventually, we solved this issue by loading in the data from Dropbox.

When cleaning our data, we had to further update the Airbnb data. This entailed removing Nan values on a case by case basis to ensure that we did not “doctor our data”, converting all city names to lowercase characters, and reformatting how suburbs are classified within a city.

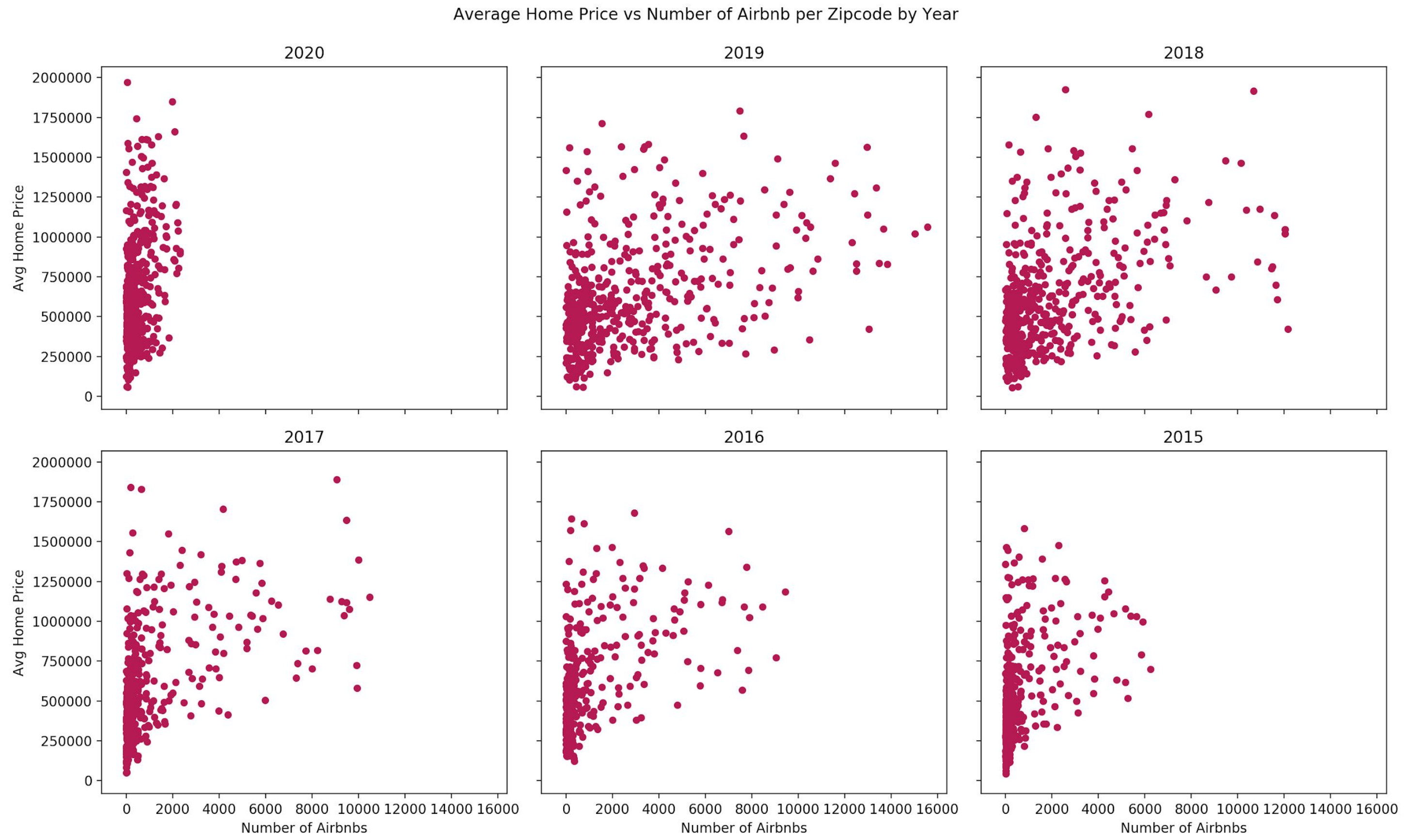
After creating our initial scatterplots, we ran a linear regression with the idea that it would give us stronger evidence for our hypothesis. However, this was not the case. We found that the real-estate market data we had just didn’t really explain the variation in the number of airbnbs.

Correlation Results

We graphed the average housing prices (aggregated by city) by year to gain an understanding of how housing prices may have changed over time. Some cities, like San Francisco and Los Angeles, show a trend of increasing prices of homes over the course of these years. On the other hand, some cities, like New York, have a very stable and flat curve, which indicates that home prices on average over the whole city are changing much less, if at all.



Next, we determined the overall distribution of the number of Airbnbs compared with average housing prices by zip code.



To further our understanding of this relationship, for each of these plots we calculated the Pearson’s correlation coefficient between the number of Airbnbs and the average housing price. If there is a strong correlation (meaning that, for a given city, a higher average price of a house is correlated with a higher number of Airbnbs rented), then it provides some evidence supporting our hypothesis.

Year	Correlation	P-Value
2015	0.4392	3.999e-22
2016	0.4979	1.64e-25
2017	0.5126	1.03e-31
2018	0.4375	7.28e-24
2019	0.4267	7.76e-23
2020	0.4046	2.26e-20

From our correlation analysis, there is a weak positive correlation between the number of Airbnbs and the average housing price. Since the p-values are close to zero, there is a very small probability that our correlation test results occurred due to chance, especially given a large sample of data points. This demonstrates that the correlation coefficients calculated from the sample size are reliable and suggests that we cannot reject the null hypothesis.

Multiple Linear Regression Results

We ran a linear regression to see how well the number of Airbnbs per zip code could be predicted using the average Zillow home price and average Zillow rental price as explanatory variables. We obtained the following regression:

Number of Airbnbs = 50.03 + 1,412.88 * *Average Zillow Home Price* - 241.06 * *Average Zillow Rental Price*

This regression yields an R-squared value of 0.129, which indicates that our housing and rental prices alone were unable to explain large amounts of the variation in the number of Airbnbs. Therefore, we decided to add the zip code data as indicator variables in our model. The coefficients attached to each of these zip code indicator variables are interpreted as the predicted difference in the number of Airbnbs from other zip codes with the same home and rental prices. A brief snippet of what this model looks like is attached below:

Number of Airbnbs = -185.54 + 1,703.91 * *Average Zillow Home Price* + 484.61 * *Average Zillow Rental Price* - 720.15 * *zipcode_02108* + ... -347.11 * *zipcode_98199*

When we ran the multiple regression with these added variables, we obtained an R-squared value of 0.927. This R-squared value indicates that the independent variables of our model account for much more variation in the independent variables than the previous model we made.

Implications

Given our analyses, we fail to reject the null hypothesis.

We are unable to confidently claim that the success in the real estate market has a significant impact on success in the Airbnb market. In our multiple linear regression, when zip code information was added, the model adjusted its predictions and captured more variation in the data. However, the lack of correlation found between the number of Airbnb rentals and housing prices, taken together with the analysis of our linear model, suggests that the success of Airbnb is not directly and solely connected to the success of the real estate market. Instead, it appears that there is a confounding factor in this relationship -- namely the desirability of a neighborhood -- which is not directly represented in our data but which can be noisily estimated with zip codes and housing prices. We believe that the number of Airbnbs rented is most impacted by the desirability of the location and perhaps concentration of housing, not necessarily just the housing prices.

Future Directions

There are many possible sources and forms of data that could help to identify and study neighborhood desirability more closely. Possible data sources are measuring the number of people moving into a neighborhood, evaluating the opinion surveys by residents (note: this has some ethical implications), and analyzing the traditional neighborhood “quality” metrics such as school district ranking and access to parks. We believe that future projects should analyze neighborhood desirability to determine its relationship with the real estate and Airbnb markets.

Furthermore, it may be worthwhile to reconsider the definition of “success” of Airbnb markets. For example, the definition of a successful Airbnb market could also take into account whether the listed Airbnb is an active listing (i.e. booked or description updated) during a given time period.