Causual Relation between Airbnb and Housing Levels

Jung Suh, Oliver Xu, Hyun Bin Kim, Jason Yuan

Executive Summary

Airbnb's original slogan was "Belong Anywhere." Airbnb started off as a renegade, trying to revolutionize the accommodation business. The founders of Airbnb saw an opportunity for tourists not to just stay in hotels, but to veer away from the mainstream tourist destinations and blend in with the locals, experience the unknown and feel like one with the local community.

However, as Airbnb grew from its founders' humble room to a global company, their original motto seems to be in question. Airbnb seems to be blending in with the hotel industry and we attempt to seek out the truth and shed light on these matters through our analysis.

Through our analysis and geospatial visualizations, we were able to draw the conclusion that Airbnb is turning into the likes of the hotel industry. Instead of keeping to its innovative promise of tourists being able to mingle around with locals, it appears to be clustering up around hot spots, be it the city center for Austin, New Orleans, Nashville, and even coastal areas for Los Angeles. This phenomenon of clustering also does not seem to be benefiting the local economy, especially for the southern states.

The visualizations provide clear proof of correlation, but to solve the classical issue of "Correlation does not equal causation," we borrowed techniques from dynamic systems theory and machine learning. In the report, we show that there is a strong probability that there is a causal relationship between the amount of Airbnbs, the daily prices of Airbnbs, and the Zillow Housing Index (ZHVI). While our initial predictive models are weak, we have shown that it is a problem worth investigating.

Key Terms: Machine Learning, Data Science, Data Visualization, Geospatial Data,Granger Causality, Convergent Cross Mapping, Sugihara Causality, Regression, Multilayered Perceptrons

Contents

1. Introduction

Airbnb has been received much negative attention in news outlets. Many articles blame Airbnb for exacerbating the housing crisis, and even causing it. While there are many anecdotal stories of less than reputable Airbnb tactics implemented by owners, we wanted to investigate this issue with a quantitative mindset to answer the following question: "Has Airbnb contributed to the rise in home prices and rental inequality across the Southern United States, or is the company simply a media scapegoat?"
Here we adopt a three-stage pipeline to analyze the situation:

1. We first implement Exploratory Data Analysis to visualize potential issues
2. We then try to find causal relationships between Airbnb and the housing crisis through time-series analysis and machine learning.
3. Finally, we try to come up with potential predictive models to strengthen our point.

2. Data

Data is the most pivotal step in data analysis. In addition to the data provided, we used additional datasets that allowed us to observe Airbnb prices and quantities over time as opposed to a single snapshot in time. The follow is the data used in this study.

- **Airbnb Price:** The daily rental rate of an Airbnb listing
- **Airbnb Amount:** The number of Airbnb rentals in a geographical area

- **Airbnb Features:** Amenities provided by each Airbnb listing
- **Geographic Coordinates:** Latitude and longitude of each Airbnb listing
- **ZHVI:** Zillow Home Value Index is a smoothed, seasonally adjusted measure of the typical home value and market changes across a given region and housing type
- **GDP:** Gross Domestic Product of a state in a given quarter

3. Exploratory Data Analysis

3.1 Local Economy

Through some Exploratory Data Analysis (EDA), we realized the rating's values spanned from 1 to 5. However, in these cities, much of the ratings were NaN values (48% for Los Angeles, 38% for Austin, 49.8% for New Orleans, 49.5% for Nashville). We assumed that these NaN values were due to the venues' unpopularity and rare visits by people. Therefore, we filled these NaN values with the ratings of 0. Proceeding onto the plotting of these venues and Airbnb locations, we have decided to focus on the few southern states that we could find in the datasets. Los Angeles, Austin, New Orleans and Nashville seemed to be the top four most common cities in the southern states.

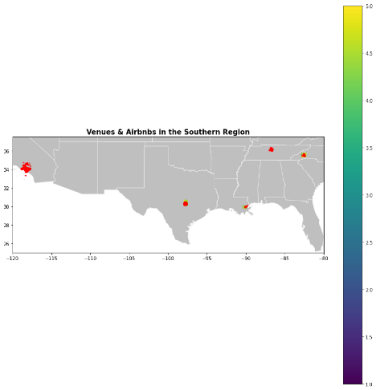

Fig. 1. Clusters of Airbnbs in the South US

Therefore, we have decided to analyze these four cities in depth. For each of these four cities, we also filtered the respective cities from the venues dataset to obtain the venues and Airbnb locations for the same city. For plotting of the points, we decided to use geopandas with the help of each respective cities' map in the form of a Shapefile (.shp) As we can see for the Airbnb locations, it is obvious that, currently, they have clustered up together. For Los Angeles, these locations are very concentrated in the city center and the coastal area such as Venice Beach and also Santa Monica Airport. These can be the effect of Airbnb chasing profitability and forgetting about its initial motto. They may have decided to allow listings congregate hot spots, instead of intervening to ensure a localized experience. For the other states, it is also visually obvious that Airbnbs are located near the city centers.

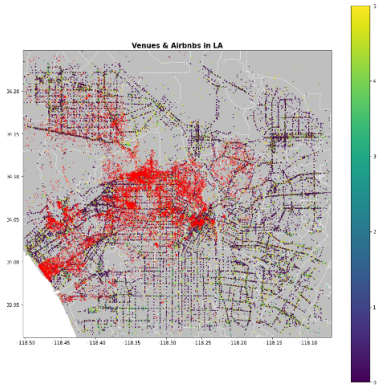

Fig. 2. Clusters of Airbnbs in Los Angeles

Looking at the venues plots, we also can detect a problem. For Los Angeles, many of the venues are congregated around Airbnb locations. However, those that are not, are showcasing mainly low ratings and even 0. This seems to be problematic as it may be an effect of the absence of Airbnb. Without Airbnb in its vicinity, it does not experience the same amount of traffic as what other venues in the vicinity of Airbnb might. Those venues in the vicinity of Airbnb are more likely to be chain businesses and even big global brands who can afford the high land prices (which we will be talking about in the later section of our report), which will contribute to the worsening of small businesses and the local economy. For the other southern cities, this seems to be a bigger problem. Many of the venues that we can see, are not even located in the cluster of Airbnb locations.

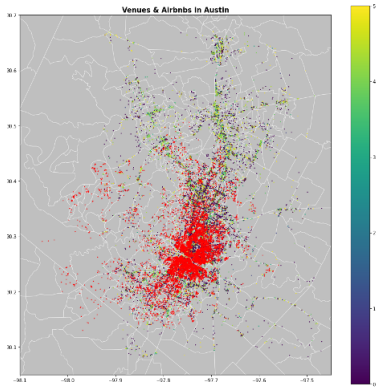

Fig. 3. Clusters of Airbnbs in Austin

For instance, Austin seems to have the worst scenario, with a lot of venues located far south of the Airbnb cluster. With venues being located away from the Airbnb cluster, Airbnb have very little positive impact on the local economy, through increased traffic of tourists and visitors.

Additionally, in the case of Nashville, the venues that are located in the cluster seem to be very dark blue in color, suggesting low to no ratings. This is in contrast to what many would believe, that Airbnb clusters would benefit the venues around them. Therefore, we arrived at the hypothesis that Airbnb is shifting away from its original motto and ignoring the local economy, in the chase for greater profitability.

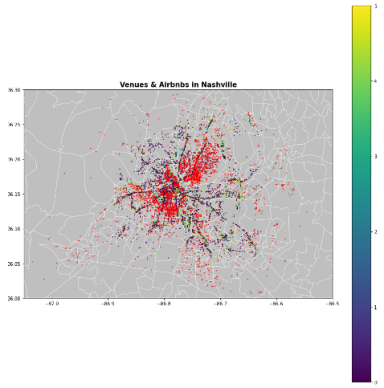

Fig. 4. Clusters of Airbnbs in Nashville

3.2 Real Estate

There have been numerous studies to show that the presence of Airbnb in a certain location is detrimental to the prices of local real estate. The problem is with supply and demand. As Airbnb's popularity rises, more and more 'professional rentals' appear. Professional rentals result in houses in that region being bought for the sole purpose of renting it out to Airbnb users. This caused greater demand and diminishing supplies for homes and this can result in rising house prices. We attempt to seek out if there is indeed any correlation between Airbnb and the home values.

We first wished to see if there was any visual trends that could support our hypothesis. Thus, we chose two regions in LA and Austin where the blue line represented a region where there was a cluster of Airbnb, while the orange line represented a region where there was barely any Airbnb. Data was prepared using the real estate dataset and transposing them to make sure we could plot a time series visualization.

As we can see, the general trend is an increasing one, as expected. However, the increase is of a greater rate for the region where Airbnb is clustered in.

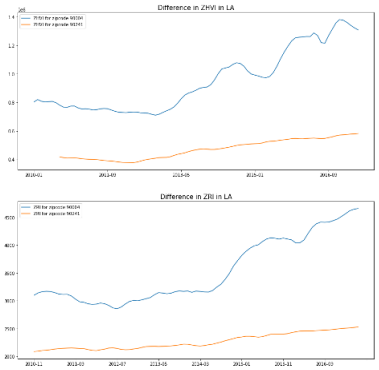

Fig. 5. Los Angeles Housing Index: Airbnb Clusters vs Non-Clusters

For instance, in LA, the blue line increased 120000/year while the orange line increased by 40000/year for the ZHVI. In Austin, the blue line increased 40000/year while the orange line increased by 10000/year for the ZHVI. Through

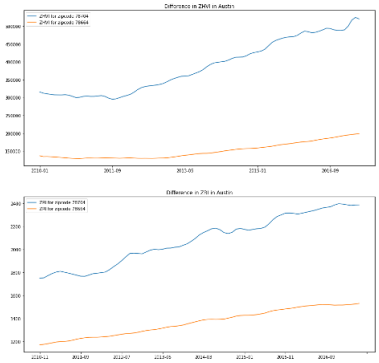

Fig. 6. Austin Housing Index: Airbnb Clusters vs Non-Clusters

this evidence, we believed that there was some relationship. Thus, we proceeded onto look at this in a deeper technical level.

3.3 Potential Limitations with initial Exploratory Data Analysis

1. Venues having low or 0 ratings near clusters does not necessarily mean its Airbnb's fault. It could have been the fault of these venues which might have unfavorable business conditions, such as unhygienic conditions or seldom sought for products or services.
2. Changes in home values are generally not influenced by a single factor but are caused by several different factors, such as economic conditions, geographic location, demographics in that region and so much more. Limiting the scope only to the number of Airbnbs in that region seems to be limited in terms of scope, but the dataset that we had on economy was limited in utility as it was not a time series dataset.

4. Causality Detection Models

4.1 Granger Causality

The age old saying in statistics "Correlation does not equate causation," is an important consideration here. While we saw visual patterns, this could be Airbnb owners following economic trends as opposed to Airbnb properties partially causing them. A powerful form to quantify causality is the Granger Causality. The Granger Causality test is used to determine whether one time series is useful in forecasting another. While correlation between two time series variables does not automatically equate causation, predictive power of one using the other can indicate causation. In other words, variable A causes B if we have a model that improves prediction of B if we include A.

Granger Causality, however, makes many hopeful assumptions. The major one is that the time series variables in question are completely separable. Especially in something as interactive and robust as real estate and the US economy, this is an assumption that is impossible to accept.

4.2 Cross Convergence Mapping

Mathematically, we can treat our system of real estate, US economy, and Airbnb as a partially deterministic dynamical system. While there are many factors that can help determine values in these time series, there are too many factors for it to be truly deterministic. Together, they can be seen as a dynamical system since they interact with each other and change over time.

In a system $Y = f(X,Y)$, cross mapping means given the points on the manifold of one variable M_y, we look for the corresponding points on M_x, i.e. points at the same time t. If X causes Y, then we say some information about X gets stored in Y. If this is the case, then we can predict the values of X given values from Y. We contrast this with Granger causality which says that X causes Y if we can predict Y better given X.

Cross Convergent Mapping allows us to find the causal relationships in these systems. Given two time series $X = \{X(1),X(2),...,X(L)\}$ and $Y = \{Y(1),Y(2),...,Y(L)\}$ where L is time series length, the algorithm is as follows.

1. Compute lagged-coordinate vectors $\underline{x}(t) =< X(t),X(t-\tau),X(t-2\tau),...,X(t-(E-1)\tau) >$ for $t \in [1+(E-1)\tau,L]$ where E is an "embedding" dimension. Search algorithms e.g. grid search can be used to identify what the best E value is. τ is the lag step.
2. Define the "shadow manifold" $M_x = \{\underline{x}(t)$ for each $t \in [1+(E-1)\tau,L]\}$
3. At t, locate $\underline{x}(t)$ in M_x and find the $E+1$ nearest neighbor vectors from selected vector $\underline{x}(t)$.
4. Denote the time indices of the $E+1$ nearest neighbors of $\underline{x}(t)$ by $t_1,...,t_{E+1}$. These indices will be used to identify the corresponding points in Y.
5. Define the model that predicts Y given M_x as $\hat{Y}|M_x = \Sigma w_i Y(t_i)$ where $i = 1...E+1$ where w_i is the weight multiplied by the corresponding $Y(t_i)$

If X and Y are dynamically "coupled" i.e. one influences the other, a clump of points in M_x should correspond to a clump of points in M_y. And as L increases, we get a denser cluster of $E+1$ points from M_x that correspond to a denser cluster of points in M_y. Given this, $\hat{Y}|M_x$ intuitively should converge to Y as L increases, e.g. we get more data/samples. Hence, we can test for convergence of the nearest neighbors to test for correspondence between states on M_x and M_y. We then plot the correlation coefficients between Y and $\hat{Y}|M_x$. If there is significant correlation, this means we have enough information from Y that is stored in X. We say that Y influences X. It sounds counter intuitive but another way to think about it is: if Y influences X then information from Y gets stored in X. We can thus use X to predict Y. Same goes for X influencing Y.

4.3 Results

When we plot the ZHVI values and the number of Airbnbs over time, we get the following.

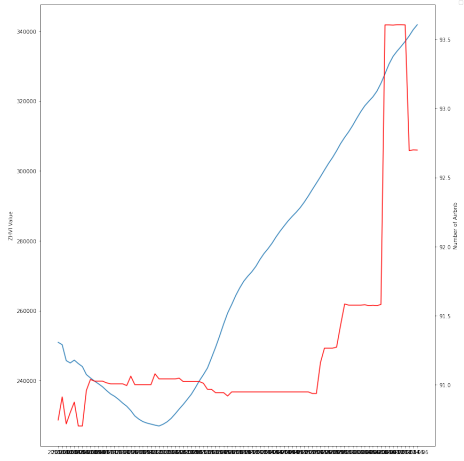

Fig. 7. ZHVI vs Number of Airbnbs

While we can see a semblance of the two variables moving together, it's impossible to determine how they are interacting. When we run it through our cross convergence mapping model, we get the following results,

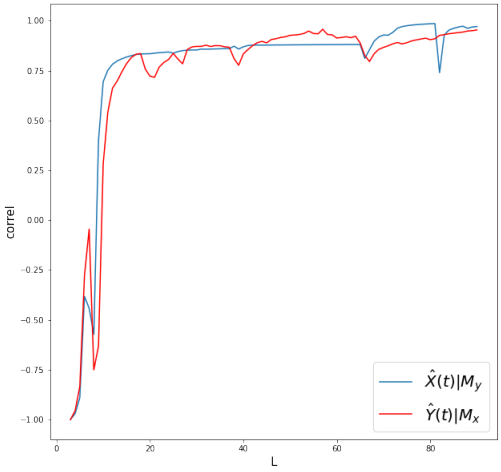

Fig. 8. Causality: ZHVI vs Number of Airbnbs

As our L value increases, we can see a clear convergence in causality. This indicates that there is a high probability that there is a causal relationship between ZHVI values and the number of Airbnbs in a specific region of the Southern US Region.

We also plot the ZHVI values and the price of Airbnbs over time, we get the following.

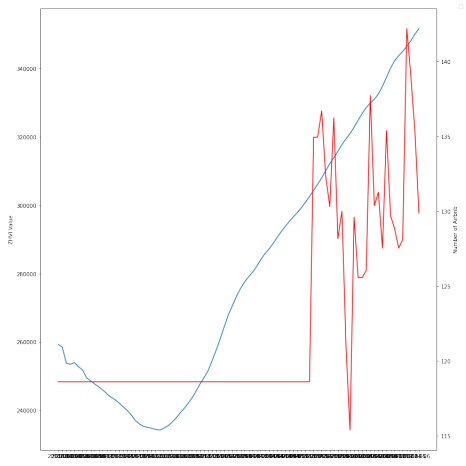

Fig. 9. ZHVI vs Price of Airbnbs

Again, we can see a semblance of the two variables moving together, it's impossible to determine how they are interacting. When we run it through our cross convergence mapping model, we get the following results,

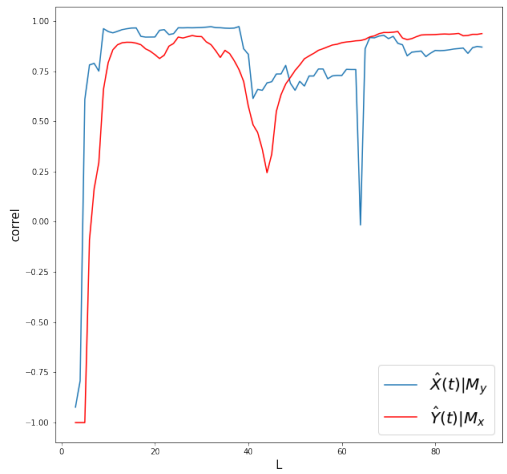

Fig. 10. Causality: ZHVI vs Price of Airbnbs

As our L value increases, we can see weaker convergence in causality. This indicates that there is a some probability that there is a causal relationship between ZHVI values and the price of Airbnbs in a specific region of the Southern US Region. However, it is not as strong as the relationship between the quantity of Airbnbs.

Finally, we plot the price of Airbnbs and the number of Airbnbs over time, and get the following.

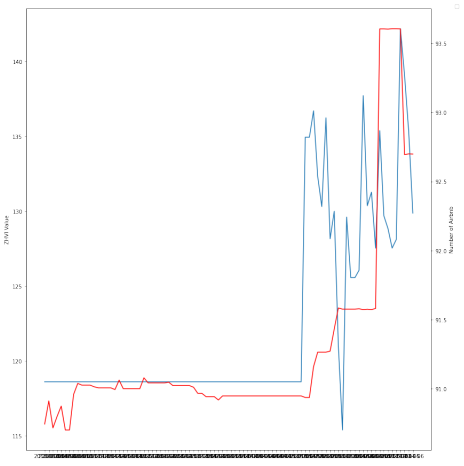

Fig. 11. Amount of Airbnbs vs Price of Airbnbs

The price behavior is much more varied, and the directions seem to move independently from each other.

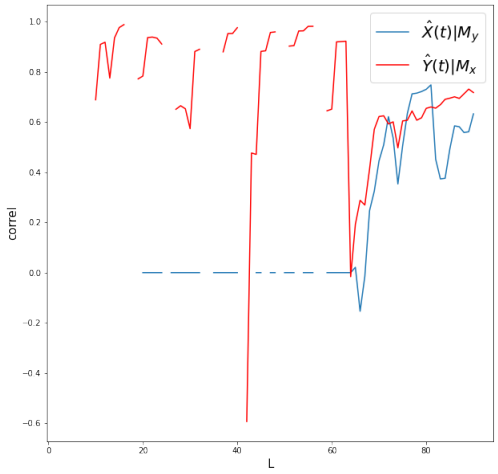

Fig. 12. Causality: Amount of Airbnbs vs Price of Airbnbs

We can see an odd relationship between price and the amount of Airbnbs. While not conclusive, it seems there is little to none causal relationship between the price and the amount of airbnbs on the market.

5. Early Prediction Models

5.1 Daily Price of Airbnb Rentals

Once we determined that which variables have causal relationships, we set out to build early regression models to predict future values. The first model we ran was to test the following hypothesis: "If the quantity of Airbnb rentals and the price do not have causal a relationship, does that imply price is not determined by factors such as room features, supply, demand, etc?"

Upon running basic regression techniques (linear regression, ridge regression, and ridge regression with a degree two polynomial transform), we get the results in Fig 13.

```
===Linear Regression ()===
Weights:
[172.23732461  79.62837671  -2.9154619  -20.19064992  10.04548173
  -7.65541021  21.65756366   0.32439547 -17.60299709 -23.64215924]
Intercept:
-60.702634898108016


Cross Validation Score
[0.22815604 0.37331713 0.29714146 0.35170695 0.31642972]


===Ridge Regression ()===
Weights:
[ 0.00000000e+00  2.38582471e+01  6.48574511e+00  4.86601777e+01
  2.45803795e+01 -1.82167783e+01  2.92800271e+01 -5.63638157e+01
 -1.07504274e+01  9.22889330e+01 -1.45511198e+02  1.48924485e+01
  6.45204937e+01 -7.66951046e+00 -2.63107450e+01  1.69623856e+01
  4.33584799e+00  1.05837169e+01 -1.48699107e+00 -4.63791718e+00
 -2.78350310e+01  1.50921585e+01 -2.58026983e+01 -1.29680136e+01
 -2.26564851e+00  7.09088279e+00  4.02415124e+00 -5.06071176e-01
  7.58060474e+00 -3.14109919e+01  1.44524783e+00  6.64788069e+00
  5.95728973e-01 -8.63297641e+00  3.92223667e+00  1.20153904e+00
 -1.34358071e+01  1.67876824e+01 -6.12902991e-02 -7.42575610e-01
  4.72443931e-01 -2.27012002e+00  1.06336882e-01  1.88622159e+00
  1.42382776e+01  1.63696557e+00  9.64249944e-01 -1.40698328e+00
 -2.78183439e-01  1.98869786e+00 -3.05489579e+00  3.90566887e-01
 -1.07828677e-01 -3.80284883e-01 -1.63382662e+00  5.13805237e+00
  3.72742202e+00  3.08728150e-01 -2.31180712e-01 -5.44854670e+00
  1.21486496e-01 -6.13843295e-01 -9.15408165e-01 -3.27695536e+00
 -3.24638910e+00  1.86867043e+02]
Intercept:
12.137402862742448


Cross Validation Score
[0.22815564 0.37331822 0.29714216 0.35170697 0.31642931]


===Ridge Regression After Degree 2 Polynomial Transform ()===
Weights:
[ 0.00000000e+00  2.38582471e+01  6.48574511e+00  4.86601777e+01
  2.45803795e+01 -1.82167783e+01  2.92800271e+01 -5.63638157e+01
 -1.07504274e+01  9.22889330e+01 -1.45511198e+02  1.48924485e+01
  6.45204937e+01 -7.66951046e+00 -2.63107450e+01  1.69623856e+01
  4.33584799e+00  1.05837169e+01 -1.48699107e+00 -4.63791718e+00
 -2.78350310e+01  1.50921585e+01 -2.58026983e+01 -1.29680136e+01
 -2.26564851e+00  7.09088279e+00  4.02415124e+00 -5.06071176e-01
  7.58060474e+00 -3.14109919e+01  1.44524783e+00  6.64788069e+00
  5.95728973e-01 -8.63297641e+00  3.92223667e+00  1.20153904e+00
 -1.34358071e+01  1.67876824e+01 -6.12902991e-02 -7.42575610e-01
  4.72443931e-01 -2.27012002e+00  1.06336882e-01  1.88622159e+00
  1.42382776e+01  1.63696557e+00  9.64249944e-01 -1.40698328e+00
 -2.78183439e-01  1.98869786e+00 -3.05489579e+00  3.90566887e-01
 -1.07828677e-01 -3.80284883e-01 -1.63382662e+00  5.13805237e+00
  3.72742202e+00  3.08728150e-01 -2.31180712e-01 -5.44854670e+00
  1.21486496e-01 -6.13843295e-01 -9.15408165e-01 -3.27695536e+00
 -3.24638910e+00  1.86867043e+02]
Intercept:
12.137402862742448


Cross Validation Score
[0.32316311 0.41793419 0.42654519 0.46382757 0.49425404]
```

Fig. 13. Prediction Results for Airbnb Rental Prices

We can clearly see that without the geographic location feature from our model, features of an individual Airbnb rental is a poor predictor of the price.

5.2 Housing Index from Airbnb Levels

We saw in the previous section that there is a strong causal relationship between Zillow's housing index and the number of Airbnb properties in an area. We also saw there was a weaker, but still significant causal relationship between Zillow's housing index and the price of Airbnb properties in an area. We wanted to see if Airbnb data can predict future Housing Index levels. When we ran it through the same three basic regression techniques (linear regression, ridge regression, and ridge regression with a degree two polynomial transform), we get the following results in Fig 14.

```
===Linear Regression ()===
Weights:
[12422.95303683  3158.72519083]
Intercept:
-1250132.366897277


Cross Validation Score
[-1.61163509  0.72076436  0.79751999  0.76561112  0.65704083]


===Ridge Regression ()===
Weights:
[     0.           2734.11017277 -16121.50115027   -878.77861841
   1312.17602964   -394.60943093]
Intercept:
581626.2270865735


Cross Validation Score
[-1.61529469  0.7222711   0.79610593  0.76595246  0.65707104]


===Ridge Regression After Degree 2 Polynomial Transform ()===
Weights:
[     0.           2734.11017277 -16121.50115027   -878.77861841
   1312.17602964   -394.60943093]
Intercept:
581626.2270865735


Cross Validation Score
[-6.17849445  0.6853684   0.1911773   0.75245677  0.67619853]
```

Fig. 14. Prediction Results for ZHVI

The cross validation scores for the models are mixed. This could be an indicator that there are other features to consider for the model. While we did see a clear causal relationship, the cross convergence mapping model does not tell us the exact nature of the relationship. In addition, we ran our dataset throough Elastic Net Regression and Multi-Layered Perceptron Regression models and came to the same conclusion: "While there is a causal relationship between Airbnb data and housing levels, we need to consider more features to get an accurate prediction model."

6. Conclusion

In this report, we explored the potential impacts Airbnb has on local economies and the housing market. While we came to some promising conclusions, there are potential flaws in our approach that we can fix.

6.1 Data

As with any data analysis study, Data Quality is everything. While we had access to robust data, our time series data quality was not high. There were many missing values in the time series. Furthermore, the time line between data did not match perfectly. For example, we used monthly data for some, but had to consolidate from quarterly data for others. Finally, Airbnb was founded in 2008. This means not enough time elapsed for us to make statistically significant conclusions.

6.2 Flaws with the Cross Convergent Mapping Model

While the cross convergent mapping model is powerful, it has limitation. The main limitation is that it only determines causation, not the actual relationship between the variables. Furthermore, it provides information containment but not the time in which it takes for one variable to affect the other. There are more powerful versions of causality testing, such as Reservoir Computing, which not only determines causation, but also the time lag between information containment.

Since we can assume effects between Airbnb levels and housing levels takes time, the next step of this study would be determine the time lag between the change of the two factors.

6.3 Predictive Models

Aside from potential issue with data, the major flaw of our predictive models was the lack of considered features. Our goal was to see the degree of prediction of Airbnb levels, but not a conclusive predictive model. While we saw some predictive power, it is clear we need more features in order to create a more powerful model.

Reference

[1] Yu Huang, Z., Christian L.E. Franzke (2020). Detecting Causality from Time Series in a Machine Learning Framework. Chaos, 30.

[2] C.W. Granger (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37, 424-438.

[3] G.Sugihara, R., H.Ye, C., E.Deyle, M., S.Munch (2012). Detecting causality in complex ecosystems. Science.

[4] H.Ye, E., L.J.Gilanrranz, G.Sugihara (2015). Distinguishing time-delayed causal interactions using convergent cross mapping. Sci.Rep 5, 14750.

[5] Z.Lu, J., B.Hunt, R., E.Ott (2017). Reservoir Observers: Model-free inference of unmeasured variables in chaotic systems. Chaos, 27(4)(041102).