CITADEL DATA OPEN
SUMMER INVITATIONAL DATATHON 2021

HOUSING ECONOMY:

THE REPERCUSSIONS OF AIRBNB BUSINESS IN THE SOUTHERN U.S.

PREPARED BY TEAM 25

Shijia(Silvia) Huang, Keying(Coco) Chen,
Steven Zhang, Jiawen(Erica) Li

TABLE OF CONTENTS

03

EXECUTIVE SUMMARY

Through the analysis of the given datasets, we discovered the following key insights.

Firstly, the distribution of Airbnb listings varied with geographic data. Through visualizations, we successfully identified the Airbnb activity density for each zip code area.

We computed the Airbnb listing price index (**ALPI**) for each zip code region and compared it with the local Zillow rental index (**ZRI**). Next, we verified that there is a casual relationship between ALPI and ZRI (for a majority of the regions in North Carolina and Tennessee) using time series analysis.

The second insight relates to how Airbnb activities influence housing rental prices, and further impacts the **Gini Index** of a selected set of zip code areas.

The Gini index is a single number that demonstrates a degree of inequality in the distribution of income/wealth. Through causal inference analysis, we are 99% confident that Airbnb activities reduced rental prices and alleviated local income inequality by providing a way to fund low-income households.

04

BACKGROUND & RESEARCH QUESTION

In just over a decade, Airbnb has transformed hospitality around the world. Its platform now counts 500 million guests that stay in 81,000 cities, and in December 2019, it announced it had collected and dispersed $1 billion in tax revenue [1]. Airbnb, whose mission is "to democratize travel by allowing anyone to belong anywhere" argues that it provides an economic equalizer, helping low-income host families to boost incomes and manage otherwise unaffordable housing costs. Yet a growing army of critics allege that the proxy hotel service more often does the opposite, accelerating affordable housing crises and gentrification patterns that force out residents.

TOPIC QUESTION:
Does an influx of temporary Airbnb rentals have a positive or negative effect on the housing markets of the Southern U.S. regions?

In this report, we will be analyzing two factors in particular to evaluate the impacts of Airbnb activities:

1. Are Airbnb activities in Southern U.S. regions causing the growth of residential rental prices?
2. Does the impact of Airbnb on residential rental prices further aggregate the level of local income inequality?

By answering those questions, we may provide useful insights for local housing policymakers to adjust restrictions and tax policies towards Airbnb, preventing it from disrupting local rental markets.

05

DATASETS

The team utilized 4 of the given datasets. Used variables in each dataset are listed below.

Listings: Descriptive information on Airbnb listings in the U.S. South.

Source: Airbnb

id - ID of the listed property
latitude - Latitude of property
longitude - Longitude of property
price - One-night rental price of the property
state - State the property is located in
zipcode - Zip Code of the property's location

Calendar: Basic information on the Airbnb listing calendar, 2007 – 2017.

Source: Airbnb

listing_id - ID of the listed property
date - Date for which data is available.
available - Whether the property is available at a certain date
price - The price of a one-night stay in U.S. dollars

Demographics: Demographic data organized by zip code, 2011 – 2015.

Source: U.S. Census.

zip code - Zip Code of the area
households - Total households (including singles)
Income brackets (9 total) - % of households in the given income bracket
mean_household_income - Mean household income

Real Estate: Monthly data that represents real estate prices, organized alphabetically by city and state.

Source: Zillow

type - Real estate value type – 'ZHVI' or 'ZRI'
zipcode - Zip Code of the region
ZRI - Zillow Rent Index (ZRI) from Nov. 2010 – Jun. 2017

06

DATA WRANGLING

After selecting the datasets, we cleaned them, and then used and combined them to form new datasets.

CLEANING LISTINGS DATA

The original listing dataset had data from 'NC', 'TX', 'OR', 'TN', 'LA', 'CA', and 'DC'. We discarded the data that's not in the Southern states, thus the cleaned listing data has the listing information of states 'NC', 'TX', 'TN', 'LA'. We also dropped rows with NaN values in 'price', 'bathrooms', 'bedrooms', 'beds', 'name', and 'zipcode' columns so that we only use listings with complete data for our future data analysis.

CLEANING CALENDAR DATA

The original calendar dataset only has the price information filled out when the property is available. We dropped the rows where the property is not available as it did not offer any pricing data.

CLEANING DEMOGRAPHICS DATA

The original demographics dataset has the rows where income brackets columns filled with '-'. We dropped the rows where the income brackets columns have NAN or filled with '-', and rows that 100 filled in any of the income brackets columns.

CLEANING REAL ESTATE DATA

The original real_estate dataset has the ZHVI from Apr. 1996 to Jun. 2017 as well as the ZRI from Nov. 2010 to Jun. 2017. For our interest, only the ZRI data from the zip code areas that appeared in the selected data from the listing.csv (the four southern states) was selected.

MERGING LISTINGS AND CALENDAR DATA

We joined the two data tables based on the property id and generated a CSV file containing all the Airbnb listing prices over the given time range of 2016-04-20 to 2018-06-01.

07

COMPUTE AIRBNB LISTING PRICE INDEX (ALPI) FOR EACH ZIP CODE REGION TO COMPARE WITH THE ZRI

Using the merged listings and calendar data, we separated the data based on the zip code and the month. For each of these separated pricing data lists, we initially took the median for each list to represent that month's and zipcode's Airbnb listing price. However, this median did not reflect changes in the data over time well, so we chose to use a calculation method similar to the newly updated ZRI calculation [2], and took the mean of the 20th and 80th percentile of the data, as opposed to the median. Although all Airbnb listing prices fell between 2016-04-20 and 2018-06-01, the date range of filled-in and missing values varied for different states and regions.

COMPUTING GINI INDEX USING DEMOGRAPHICS DATA

We used the income brackets data region from the demographics dataset to compute the Gini Index in each zip code. Then we plotted the choropleth of the Gini Index with Airbnb listings to explore the patterns. Darker colors represent a higher Gini Index in a region.

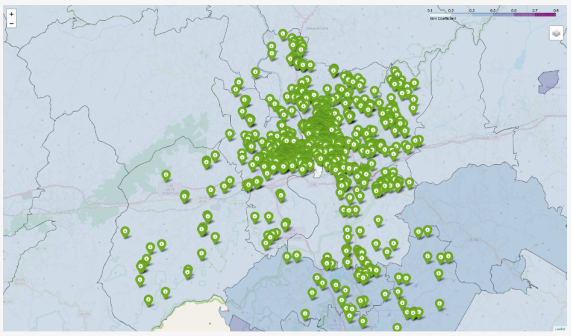

Figure 1. Gini Index and Listings Distribution (NC)

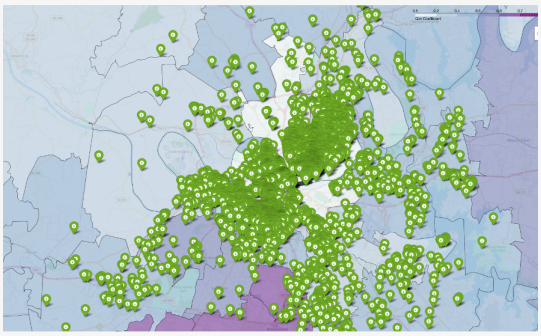

Figure 2. Gini Index and Listings Distribution (TX)

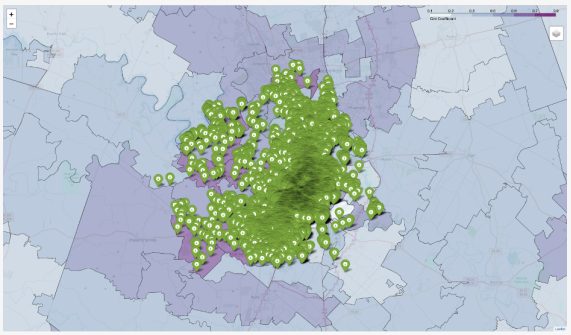

Figure 3. Gini Index and Listings Distribution (TN)

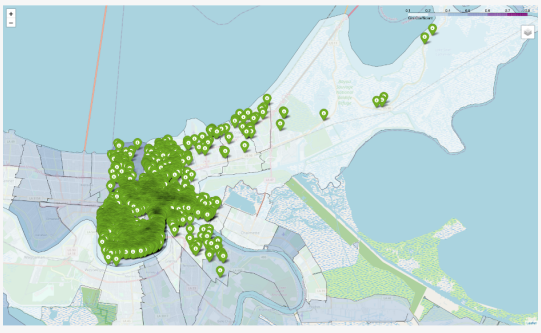

Figure 4. Gini Index and Listings Distribution (LA)

From the plotted choropleths, we noticed that the region where more Airbnb is more active tends to have a lighter Gini index color and thus lower levels of income inequality. Therefore, we decided to do a further analysis of the causality of Airbnb activity on income inequality.

08

EXPLORATORY DATA ANALYSIS

To explore the listings.csv file, we started off by plotting a histogram of all the Airbnb listing prices from the 4 Southern states (TX, NC, TN, LA) shown in Figure 5. The two most frequent daily listing prices are around $50 and $175. The majority of listing prices fall between $50 and $200.

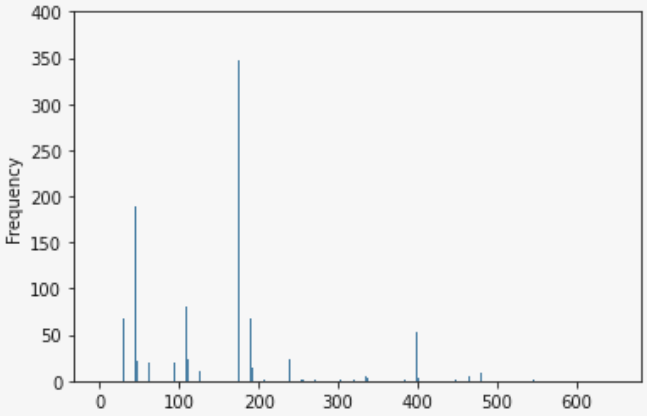

Figure 5. Listing prices distribution (exc.outliers)

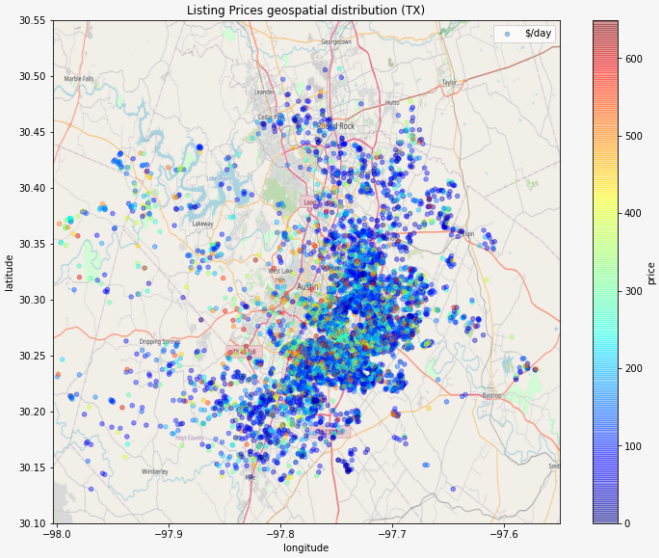

Figure 7. Listing prices geospatial distribution (TX)

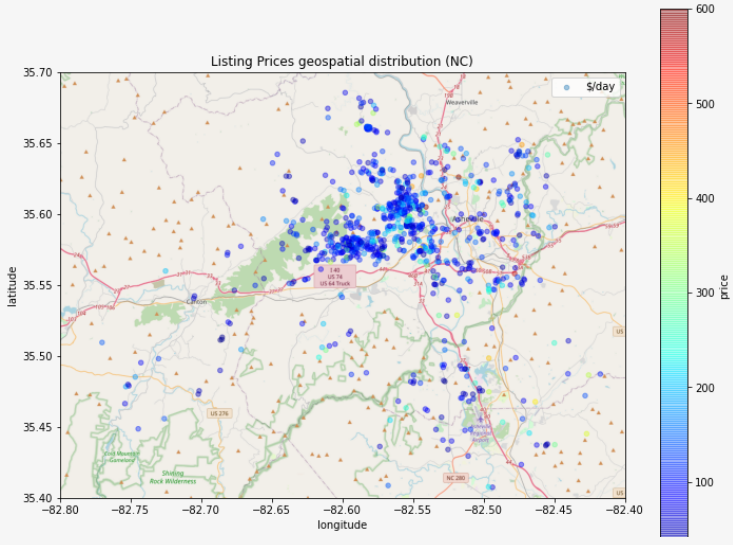

Figure 6. Listing prices geospatial distribution (NC)

In Figure 7, the Airbnb business is more active in the cities.

in Figure 6, the Airbnb listings are centered around the downtown areas and along the highways.

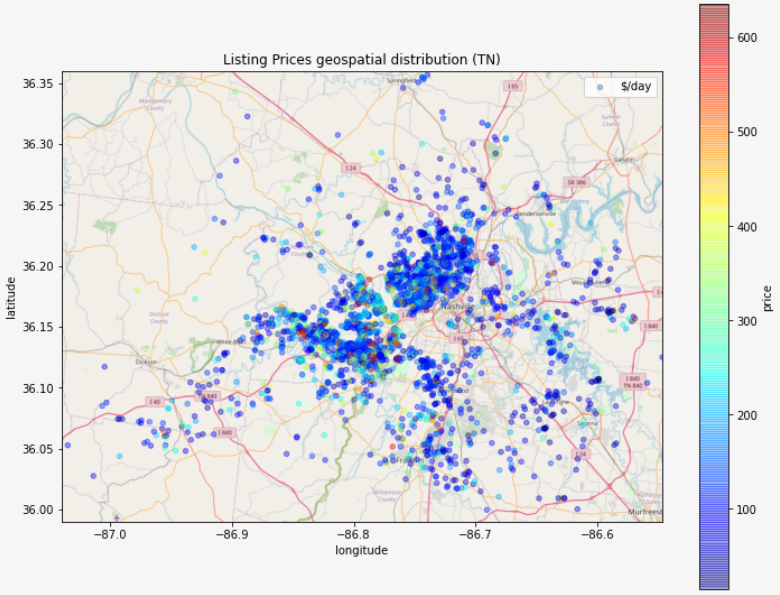

Figure 8. Listing prices geospatial distribution (TN)

In Figure 8, Airbnb listings are fanning out from the downtown areas and along the highways.

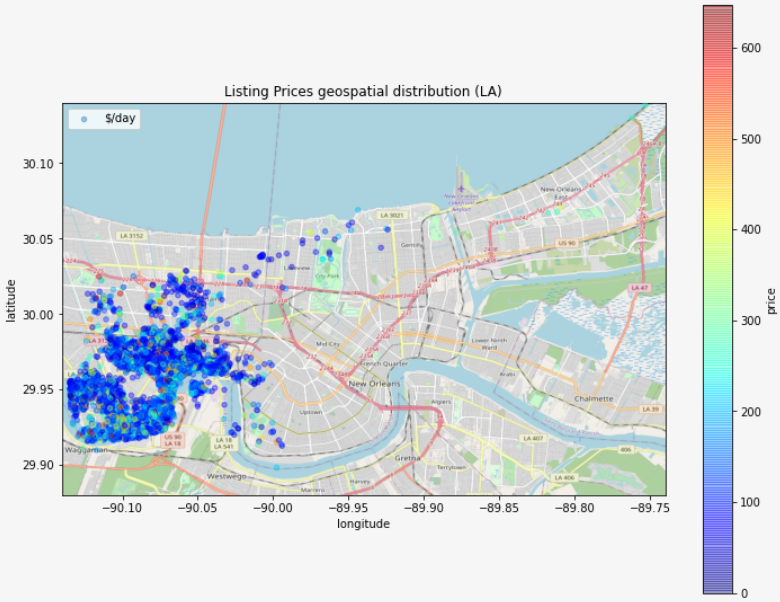

Figure 9. Listing prices geospatial distribution (LA)

In Figure 9, Airbnb listings are only active in some zip code areas in the state.

We generated a map of listing prices based on their given longitude and latitude and separated this into four maps based on the four Southern states, shown in Figure 6-9.

By observing the geospatial distributions of Airbnb listing prices, we noticed that Airbnb activities are more active in several zip code areas than other areas. In addition, listing prices are distributed quite uniformly across all states.

10

METHDOLOGY

We used the Granger Causality Test and Structural Equation Modeling to analyze the causal relationships of the ALPI, ZRI, and the Gini Index.

VERIFICATION OF THE CAUSAL RELATIONSHIP BETWEEN ALPI AND ZRI USING TIME SERIES ANALYSIS

Granger Causality Test is a statistical hypothesis test for determining whether one time series is useful in forecasting another. A time series X is said to Granger-cause Y if it can be shown through hypothesis tests on lagged values of X with lagged values of Y that those X values can provide statistically significant information about future values of Y. As mentioned in 3.4, we have ZRI and ALPI data across an overlapped time range, therefore, we decided to use the Granger Causality Test as a verification of the causal relationship between the ZRI and ALPI.

We selected the ZRI and ALPI data from the overlapped time period for each zip code area as the time series we would perform the test on. Then, we used the Granger Causality Tests function from the statsmodels package, which performs four tests (params_ftest, ssr_ftest, ssr_chi2test, lrtest) for Granger non-causality of two time series. The null hypothesis for the tests is that: the time series ALPI does not Granger cause the time series ZRI. The test is based on whether the past values of ALPI have a statistically significant effect on the current value of ZRI, taking the past value of ZRI into account as regressors [3]. We reject the null hypothesis that ALPI does not Granger cause ZRI if the p-values are below the desired size of the test. (In our case, we picked our confidence interval to be 95%, therefore the threshold for our p-value is 0.05). Multiple lagged values from ALPI were tested, the maximum lag value we tested depends on the length of the overlapped period of the two indices, which can be expressed as:

$$MaximumLagValue = ((OverlappedPeriod'sLength - 1)/3) - 1$$

11

STRUCTURAL EQUATION MODELLING

We chose to use the **Structural Equation Modelling (SEM)** technique to measure the causal inference between Airbnb rental activities, rental price level, and the second-order impact on income inequality. We also explored whether the potential increase in rental price caused by the Airbnb business could aggregate the level of local income inequality.

Scientists often study causal inference between observed variables to understand the causal effect of one event on another. However, not all events can be presented by observational data, and sometimes we are more interested in studying an underlying phenomenon. Structural Equation Modelling provides the advantage of analyzing the structural relationship between measured variables and **latent** (unobserved) constructs in a combination of multivariate regression and factor analysis. It can be used to estimate the multiple and interrelated dependence in a single analysis on the unobserved structure.

In our case, we are trying to explore the second-order effect of Airbnb activity on income inequality. Both of the variables cannot be represented by observed data, and there are many confounding variables and variables of constructs. The SEM allows us to select variables in our dataset as the observed constructs of the latent variables, and then explore their interrelated relationships and also measurement errors.

THE SEM MODEL

The SEM model is fitted using data where the listings data were collected, and we organized and standardized variables by zip code. The fitted model is below:

```
income_inequality =~ gini_coef
airbnb_activity =~ num_listings + avg_price
rental_price =~ income_inequality
rental_price ~ mean_income + airbnb_activity
airbnb_activity ~ num_households + avg_ZRI + mean_income
income_inequality ~ airbnb_activity
```

The latent variables are **airbnb_activity**, **rental_price**, and **income inequality**, which represent abstract phenomena that cannot be directly observed.

The latent variable **airbnb_activity** is measured by two observed variables: average listing price and the number of listings during the timeframe when the listings data were collected, which can be measurements of the level of Airbnb rental activity in that region. There is also a multivariate regression model between Airbnb activity level, mean households income, average ZRI from 2012-2017, and num of households in the zip code region.

The latent variable **income_ineuqality** is measured by the Gini index of the region. The higher the Gini index means higher the level of income disparity and then higher income inequality.

RESULTS

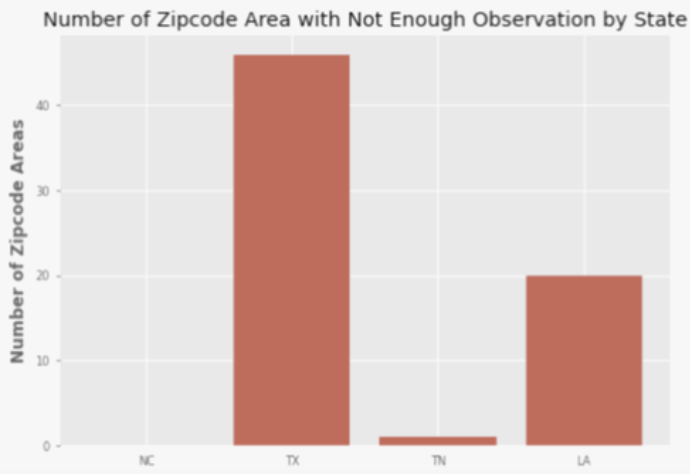

Figure 10. Number of Area with Not Enough Observations by State

There are 107 zip code areas with both ZRI and ALPI data available. Then, we performed the Granger Causality hypothesis test on those areas. The test requires time series with at least five months of observations, 67 of the 107 zip codes do not meet the requirement, these areas with not enough observations are mainly located in TX and LA state (shown in Figure 10). Additionally, one zip code area has a flat ALPI value across the time period, thus we excluded it from the test as well.

The remaining areas with enough observation to do the test are located in TX and NC. 30 of the 39 zip code areas have a p-value < 0.05 for at least one of the four Granger-Cause-Hypothesis-Test (shown in Figure 11). The detailed p-value results can be found in Appendix A.

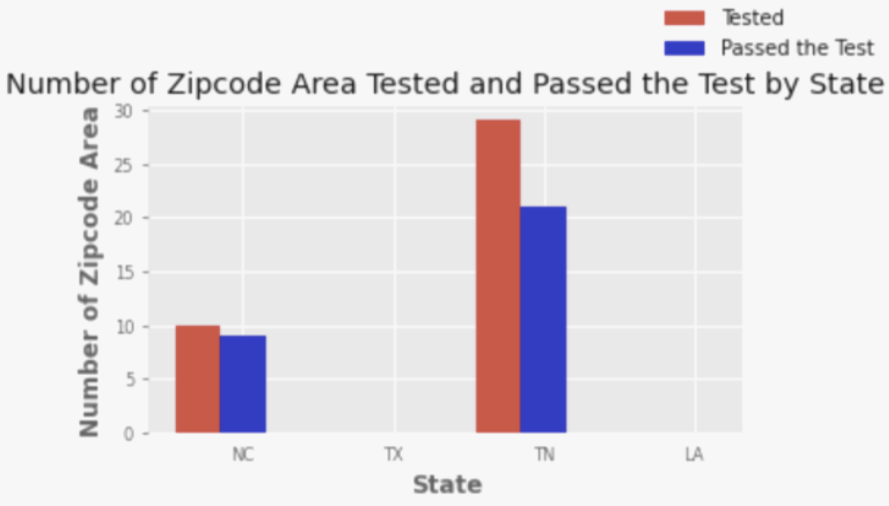

Figure 11. Number of Area Tested and Passed the Test by State

We concluded that for those zip code areas in North Carolina and Tennessee, at a 95% Confidence Interval, we reject the hypothesis that there is no Granger Causality between ALPI and ZRI.

AIRBNB ACTIVITY, RENTAL PRICE, AND INCOME INEQUALITY

Using Structural Equation Modeling, we fitted the dataset into the aforementioned model structure to get the estimate of parameters with their p-value (Figure 12). See Appendix B for detailed SEM model inspection results.

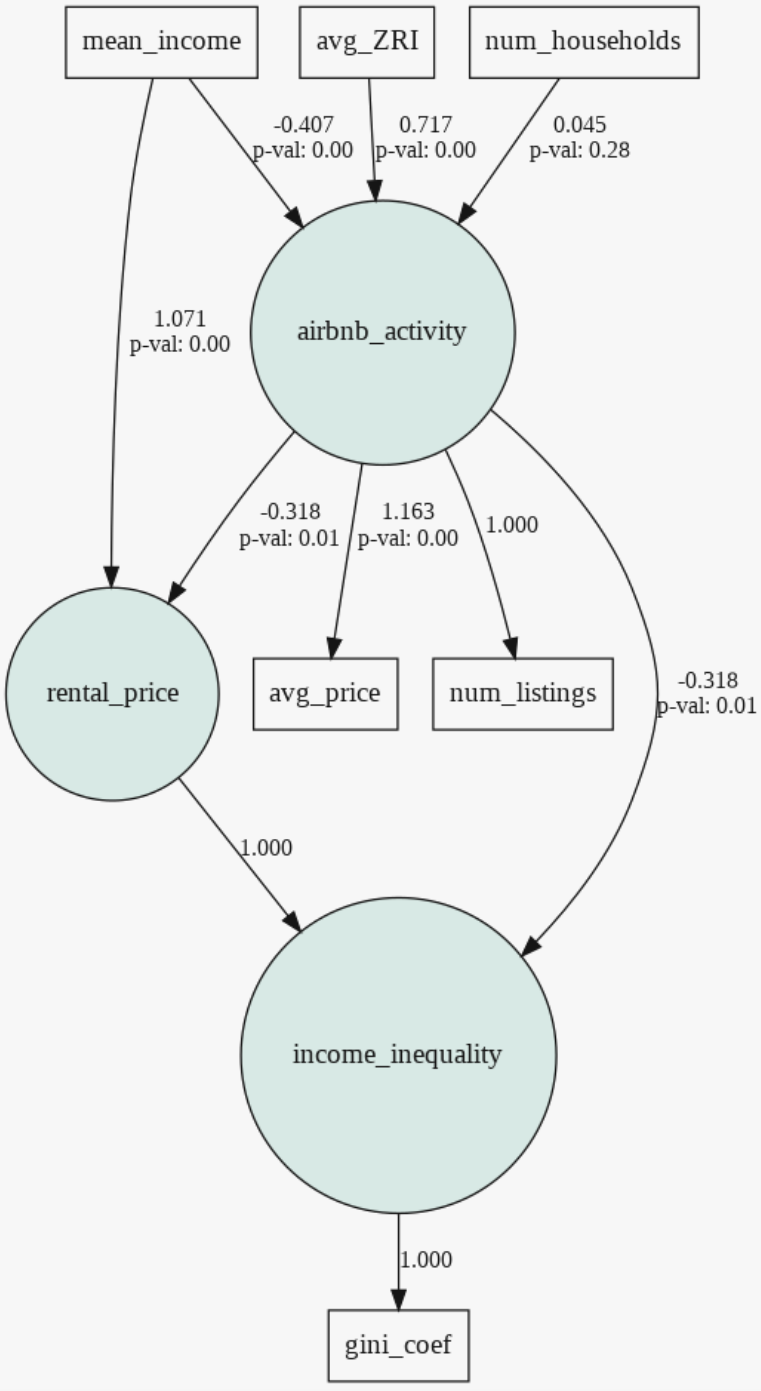

Figure 12. Visualization of the SEM Model

We noticed that the parameter for the causal effect of airbnb activity to rental price and economic inequality is -0.318 with a significantly small p-value.

Thus we can conclude that there is a negative causal effect of airbnb activity on rental price and income inequality with 99% confidence.

In addition, we also noticed that the measurements of the latent variables are mostly accurate with 99% confidence, except that the number of households may not be a good structural construct of airbnb rental activity.

15

CONCLUSION

In order to evaluate the impact of Airbnb activities on the U.S. South region's economy, we decided to answer two main questions: are Airbnb activities in Southern U.S. regions causing increased growth in residential rental prices, and does this potential impact further affect the local level of economic inequality?

To answer the first question, we calculated the Airbnb Listing Price Index (ALPI) for each zip code over the given data's time range. Next, we used time series analysis to compare it with the Zillow Rental Index (ZRI) over the same zip code areas and time range, and found that there is a causal relationship between them for North Carolina and Tennessee.

To answer the second question, we calculated the Gini Index, a number that represents the level of income inequality. Next, for each zip code region, we performed causal inference analysis using these Gini Index and the given data and found statistically significant evidence that Airbnb prices have a negative causal effect on the income inequality in U.S. southern regions.

Through our statistical analysis of the datasets, we found that Airbnb listings reduced rental prices and also helped provide a source of income to lower-income households, reducing income inequality,

16

LIMITATIONS

One limitation was that the listings.csv dataset given only contained Airbnb listing prices from five states. In the future, we could look for data from all of the Southern states, and could also compare this to the other states.

Moreover, we did not have enough overlapped ZRI and Airbnb Listing Price Index data for Texas and Louisiana to do the time series analysis, we could look for more ZRI data for these two states to perform the analysis in the future.

Another limitation was that the Airbnb listings data only contains data from the time range 2016-04-20 to 2018-06-01, and also has a lot of missing data, as listing prices were not shown when the Airbnb is occupied. In the future, we could address this problem by searching for more Airbnb in other time ranges online, and potentially using linear or other models to interpolate and extrapolate missing data.

17

REFERENCES & APPENDIX

[1]
"Is Airbnb Helping – or Worsening – Inequality in Cities?," U.S. News & World Report. [Online]. Available: https://www.usnews.com/news/cities/articles/2019-05-02/airbnbs-controversial-impact-on-cities. [Accessed: 18-Jul-2021].

[2]
"Zillow Rent Index Methodology (Most Current) - Zillow Research", Zillow Research, 2021. [Online]. Available: https://www.zillow.com/research/zillow-rent-index-methodology-2393/. [Accessed: 18- Jul- 2021].

[3]
J. Perktold, S. Seabold and J. Taylor, "statsmodels.tsa.stattools.grangercausalitytests — statsmodels", Statsmodels.org, 2021. [Online]. Available: https://www.statsmodels.org/stable/generated/statsmodels.tsa.stattools.grangercausalitytests.html. [Accessed: 18- Jul- 2021].

Appendix A.
Summary of P-Values for the Granger Causality Hypothesis Test

Zipcode	Lag	P_value_ssr	P_value_ssr	P_value_lrte	P_value_par
28704	1	0.0218495	0.0013963	0.0065661	0.0218495
28704	2	0.0916647	0.0012331	0.0125133	0.0916647
28704	3	0.3944634	0.0029097	0.0327575	0.3944634
28715	1	0.779025	0.7384788	0.739061	0.779025
28715	2	0.0332505	9.13E-06	0.0019497	0.0332505
28715	3	0.0362579	2.26E-25	1.22E-05	0.0362579
28732	1	0.2924284	0.1966286	0.2114982	0.2924284
28732	2	0.8062465	0.6640785	0.6737897	0.8062465
28732	3	0.6292477	0.0856117	0.1666574	0.6292477
28759	1	0.8367649	0.8029021	0.8031746	0.8367649
28759	2	0.1869359	0.0084047	0.034945	0.1869359
28787	1	0.0485738	5.26E-06	0.0027497	0.0485738
28801	1	0.0414216	0.0060533	0.0156016	0.0414216
28801	2	0.096378	0.0015095	0.0137181	0.096378
28801	3	0.1645628	1.12E-07	0.0017501	0.1645628
28803	1	0.0232166	0.0016184	0.0071291	0.0232166
28803	2	0.2188178	0.0265919	0.061681	0.2188178
28803	3	0.3574214	0.001225	0.0234167	0.3574214
28804	1	0.7763255	0.7353047	0.7359083	0.7763255
28804	2	0.3487986	0.0989284	0.1450063	0.3487986
28804	3	0.2614084	5.15E-05	0.00816	0.2614084
28805	1	0.2530515	0.1585198	0.1749783	0.2530515
28805	2	0.6994356	0.4985611	0.5192432	0.6994356
28805	3	0.3081268	0.0002966	0.0141762	0.3081268
28806	1	0.5921147	0.5212668	0.5247786	0.5921147
28806	2	0.1369576	0.0056842	0.0261261	0.1369576
28806	3	0.4287122	0.005777	0.0435619	0.4287122
37013	1	0.2572987	0.1253021	0.1484846	0.2572987
37013	2	0.823148	0.5745562	0.5951245	0.823148
37015	1	0.9288856	0.909258	0.9092906	0.9288856
37015	2	0.1985753	0.0004299	0.0134215	0.1985753
37027	1	0.1809751	0.0638284	0.0880527	0.1809751
37027	2	0.1634831	8.45E-05	0.0079909	0.1634831
37029	1	0.0152597	3.90E-05	0.002033	0.0152597
37029	2	0.3789719	0.0263007	0.0752117	0.3789719
37072	1	0.7478422	0.6801103	0.6815234	0.7478422
37072	2	0.1789068	0.0001844	0.0101626	0.1789068
37076	1	0.0790595	0.0096615	0.0252627	0.0790595
37076	2	0.5970154	0.1936682	0.2527133	0.5970154
37080	1	0.2143719	0.0890532	0.1133184	0.2143719
37080	2	0.6249062	0.2293581	0.2854343	0.6249062
37115	1	0.9907753	0.9882237	0.9882238	0.9907753
37115	2	0.4143317	0.0408781	0.0954103	0.4143317
37138	1	0.4971985	0.3762557	0.3862539	0.4971985
37138	2	0.4833581	0.0825357	0.1438967	0.4833581
37189	1	0.4488121	0.3211178	0.3337783	0.4488121
37189	2	0.7903898	0.5070544	0.5340439	0.7903898
37201	1	0.3337082	0.1979957	0.2174661	0.3337082
37201	2	0.517335	0.110035	0.1724745	0.517335
37203	1	0.2210913	0.0944639	0.1186321	0.2210913
37203	2	0.6997149	0.3412585	0.3858837	0.6997149
37204	1	0.1904594	0.0706932	0.095027	0.1904594
37204	2	0.1489191	3.58E-05	0.0062307	0.1489191
37205	1	0.5507812	0.4390121	0.4463982	0.5507812
37205	2	0.2503284	0.0023099	0.0248901	0.2503284
37206	1	0.321706	0.1860015	0.2061443	0.321706
37206	2	0.3874474	0.0294119	0.0797812	0.3874474
37207	1	0.6869379	0.604289	0.6069461	0.6869379
37207	2	0.2884513	0.0057276	0.036324	0.2884513
37208	1	0.5160835	0.3981892	0.4072223	0.5160835
37208	2	0.0603955	2.83E-10	0.0005615	0.0603955
37209	1	0.1203219	0.0266611	0.0477107	0.1203219
37209	2	0.2079648	0.000614	0.0151812	0.2079648
37210	1	0.6266264	0.5302266	0.5346327	0.6266264
37210	2	0.0361832	7.20E-15	0.0001432	0.0361832
37211	1	0.9697477	0.9613827	0.9613853	0.9697477
37211	2	0.3113071	0.00902	0.0445147	0.3113071
37212	1	0.0025581	1.26E-09	0.000129	0.0025581
37212	2	0.7819444	0.4902851	0.5189621	0.7819444
37214	1	0.6765575	0.5914617	0.5943817	0.6765575
37214	2	0.9473446	0.8633988	0.8656746	0.9473446
37215	1	0.093733	0.0148063	0.0327084	0.093733
37215	2	0.0160878	3.01E-26	1.65E-05	0.0160878
37216	1	0.0411736	0.001509	0.0093388	0.0411736
37216	2	0.2809154	0.0048655	0.0338481	0.2809154
37217	1	0.2031143	0.0802297	0.1045842	0.2031143
37217	2	0.1485398	3.50E-05	0.0061884	0.1485398
37218	1	0.1779553	0.0616959	0.085867	0.1779553
37218	2	0.3605255	0.0203124	0.0658409	0.3605255
37219	1	0.5496975	0.4377276	0.4451625	0.5496975
37219	2	0.5013896	0.0965333	0.15866	0.5013896
37220	1	0.0449784	0.0019844	0.010691	0.0449784
37220	2	0.0272839	3.71E-18	6.75E-05	0.0272839
37221	1	0.6946749	0.6138693	0.6163404	0.6946749
37221	2	0.1239467	5.61E-06	0.0038191	0.1239467

Appendix B.
The SEM Model Inspection Results

	lval	op	rval	Estimate	Std. Err	z-value	p-value
0	income_inequality	~	rental_price	1.000000	-	-	-
1	rental_price	~	mean_income	1.071313	0.0620887	17.2546	0
2	rental_price	~	airbnb_activity	-0.318240	0.114874	-2.77035	0.00559966
3	airbnb_activity	~	num_households	0.045155	0.0419669	1.07597	0.281938
4	airbnb_activity	~	avg_ZRI	0.717151	0.14491	4.94895	7.46141e-07
5	airbnb_activity	~	mean_income	-0.407110	0.106077	-3.83788	0.000124104
6	income_inequality	~	airbnb_activity	-0.318240	0.114874	-2.77035	0.00559966
7	gini_coef	~	income_inequality	1.000000	-	-	-
8	num_listings	~	airbnb_activity	1.000000	-	-	-
9	avg_price	~	airbnb_activity	1.162864	0.296108	3.92716	8.59553e-05
10	income_inequality	~~	income_inequality	0.006354	0.000602156	10.5524	0
11	airbnb_activity	~~	airbnb_activity	0.000000	0.00101762	0	1
12	rental_price	~~	rental_price	0.006354	0.000602156	10.5524	0
13	num_listings	~~	num_listings	0.015765	0.00238345	6.61419	3.73586e-11
14	gini_coef	~~	gini_coef	0.000156	0.000602156	0.259306	0.795399
15	avg_price	~~	avg_price	0.018266	0.00285128	6.40615	1.49242e-10