THE UNIVERSITY OF CHICAGO

How Does Superhost Accreditation Benefit Airbnb Hosts in the City of Chicago?

By

Yongfei Lu

July 2021

A paper submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts in Computational Social Science

Faculty Advisor: Yanyan Sheng

Preceptor: Sanja Miklin

Abstract

In order to obtain a superhost accreditation, an Airbnb host has to meet a series of strict requirements, including a 90% response rate, 80% 5-star reviews, honor confirmed reservations, etc. However, is it worth the effort to get such an accreditation? How does the superhost status benefit the hosts? Based on signaling theory that effective information disclosure in markets can help provide positive encouragement and maintain benign market conditions, we assume that superhost status serves as a positive market signal, and thus can benefit the badge owner in some way. Choosing Chicago as our target city, where Airbnb hosts have earned tens of millions of dollars over the 5 biggest weekends in 2019, we take advantage of content analysis techniques, decision tree, random forest models, ordinary least squares (OLS) linear regression and further multilevel modeling to answer the question. Our findings confirm that superhosts do have significantly higher income in comparison with normal hosts. Nevertheless, such accreditation does not tend to bring up their rental price. One critical channel through which superhosts earn more consists in their capacity to maintain a relatively higher occupancy rate, which is achieved mainly through providing better accommodation services. Furthermore, even though Chicago has a notorious reputation for high crime rate and conspicuous residential segregation, neither the safety nor race issues reduce host revenues.

Keywords: Airbnb, superhost, price, revenue, segregation, crime

Introduction

Over the past decade, the 'sharing economy' has been unrolling swiftly via platforms like Airbnb. In 2009, the company launched the Superhost program to automatically recognize the top-performing hosts who satisfy the minimum performance standards[1] specified by the Superhost Program in the most recent 12 months from the review. The superhost badge appears on their listing and profile to help renters identify them. As the Airbnb business has experienced a long, steady boom, the program has provided a great natural experiment through which researchers can examine the external validity of signaling theory (Spence, 1973). Due to Akerlof's (1970) research on adverse selection, the opposite side of signaling theory, it has been widely acknowledged that in markets with severe information asymmetries, 'bad' goods can dominate the markets, dragging down the overall price. Considering the robust growth of Airbnb business, it is interesting to understand whether superhost accreditation serves as an important channel for low-cost information disclosure that builds up the market; and how such an accreditation benefits those top-performing hosts in the online rental market.

Signaling theory, as first proposed by Michael Spence (1973), stated that educational credentials send information about the ability of job seekers, thus effectively helping employers distinguish highly talented workers from low ability ones. In addition to labor market, signaling theory has been applied extensively to many other

[1] To be qualified as superhosts, hosts must maintain a 4.8 or higher average overall rating based on reviews from their Airbnb guests, complete at least 10 stays, has 0 cancellations of reservations with fewer than 100 reservations and respond to 90% of new messages within 24 hours in the past year.

fields. By analyzing the process of Initial Public Offering, Leland and Pyle (1977) demonstrate how promising firms should always send clear positive signals to the stock market, which are expected to be difficult to be imitated by other inferior companies. However, in the context of Internet shopping, a study shows that information used for price comparison weakens the pull of brands by reducing shopping at concentrated branded retailers by approximately one tenth (Waldfogel & Chen, 2006). Lewis (2011) investigates the operation of eBay Motors, and extends our understanding by finding that price is significantly influenced by online disclosures and that disclosure costs impact both the degree of disclosure and prices.

Signaling theory has also been applied to the analysis of Airbnb business. Research shows that Airbnb hosts with longer self-description on a mix of topics tend to be perceived as more trustworthy, an assertion of signaling theory (Ma et. al, 2017). Another research on signal attributes of Airbnb listings finds that superhost status, as well as price, extra fees, location competitiveness and house rules are all effective signals in Airbnb, especially for listings without review comments (Yao et. al, 2019). Moreover, in the hedonic price regression model, reputational aspects of the rooms, including rating scores and duration of membership, are found to be associated with economic reward (Teubner et al., 2017). While current research treats superhost accreditation as one aspect of the signal information, the author regards the status as the critical signal that separates top-performing hosts from ordinary ones in other listing attributes, including review numbers, regional advantages, duration of membership, etc. Therefore, evaluating the effectiveness of Superhost program becomes the highlight of

the research.

In terms of how superhost status benefits Airbnb hosts, research results vary across locations and with different scale of analysis. A report based on 2.2 million host data shows that on the whole, superhosts charge less on a nightly basis but maintain a higher occupancy rate (Shatford, 2018). Another study, based on 33 cities listed on Airbnb.com, shows contradictory results and demonstrates that for hosts with superhost status, more listings and verified identities usually set prices at a higher level since these are regarded as a type of quality signals (Wang et al., 2017). Current research as mentioned above tends to focus on large-scale analysis, and mostly use traditional parametric approaches, but such results derived from extensive regions ignore the evident features of a specific location and cannot avoid the limitation of parametric approaches[2]. The City of Chicago, the third largest city in the United States, where crime rate is notoriously high, residential segregation is evident but Airbnb has gained great success, is worth further exploration. The researcher, therefore, would like to further examine the external validity of signaling theory in such a context. More specifically, this study focuses on answering the following four questions: (1) Is superhost accreditation an effective market signal that lives up to consumers' expectation and gives superhosts more chance to increase their income? (2) Does superhost accreditation gives superhosts more market power to set a higher room price? (3) Is occupancy rate a critical channel through which superhosts can benefit more from the market? and (4)

[2] Parametric models are based on the assumption of some finite set of parameters, thus the model complexity is bounded in terms of capturing the complex non-linear dynamics (usually it's not easy for us to specify a very accurate model form). However, non-parametric models have more flexible assumptions on data distribution and allow infinite dimensions of parameters to capture more accurate information about a growing amount of data.

Does the conclusion still hold in the context of high crime rate and conspicuous residential segregation? To answer these research questions, the researcher applied geospatial data analysis to look for potential geospatial patterns and utilized analytical methods such as linear regression, logistic regression, random forest models, multilevel modeling.

Data and Methods

Data

Our main source of data comes directly from insideairbnb.com, an independent, non-commercial and widely used website which provides monthly scraped data from Airbnb. We select the Chicago listing data of August, 2019 to do our research, when huge amounts of income were generated due to a high concentration of various festivals. Moreover, this period of time is well before the outbreak of COVID-19 pandemic, and thus we can have a precise evaluation of Airbnb's regular commercial performance in its peak month. We also merge the dataset with the crime rate data (from 2018 to 2019) released by Chicago Police Department. After cleaning the raw data, we obtain 8646 observations[3] in our dataset. Moreover, we create a new variable occupancy rate based on the already-existent variable reviews per month. The variable is calculated based on the following equation[4] (Equation 1), which is supported by the research on the effects of short-term rentals on San Francisco housing market, conducted by Budget and Legislative Analyst's Office in 2015. Here, we take 0.5 for review rate, and 3 for average length of stay. The two estimated values are provided on the Airbnb website.

In this part, we conduct an exploratory data analysis to develop a big picture of

$$occupancy\ rate = \frac{reviews\ per\ month}{review\ rate} \times \frac{average\ length\ of\ stay}{30}$$

Equation 1. Calculating the 'occupancy rate' variable

[3] An observation is a listing here.

[4] Here we estimate the number of guests per month per listing by dividing the number of reviews per month (tool variable for the number of guests writing comments per month) by review rate. Multiplying this number by average length of stay per guest in that room, we get the number of days when the room is booked. Divide the number by 30 we get the occupancy rate.

the data we're going to use. Here, we mainly focus on variables, including host's monthly revenue (revenue), listing price (price), superhost status (host_is_superhost), monthly occupancy rate of the listing (occupancy_rate), yearly average number of crimes per thousand residents in the community where the listing is located from 2018 to 2019 (avg_yearly_cirmes) and others indicating accommodation service quality. We first explore the difference between superhosts and regular ones in these aspects, then the correlation between all the variables and finally the geospatial correlation among them. From Table 1, we can see that price and revenue have some extremely large values, so we'll exclude these outliers before further analysis.

To explore whether superhosts have more advantages in nightly room price, revenue and the critical performance indicator, occupancy rate, we plot the following 3 box plots (Figure 1, Figure 2, and Figure 3). These figures indicate that while superhosts tend to maintain a noticeable higher occupancy rate and earn more than their regular host counterparts, they do not seem to set an evidently higher price.

Table 1[5]

Descriptive Statistics of Key Variables

[5] For descriptive statistics of all variables, see Table D in the Appendix.

index	price	revenue	occupancy_rate	avg yearly crimes
count	8646.000000	8646.000000	8646.000000	8646.000000
mean	176.640180	1780.815184	38.909438	81.642112
std	397.146114	4288.507866	34.184191	48.361371
min	0.000000	0.000000	0.000000	17.000000
25%	69.000000	204.000000	7.600000	46.700000
50%	109.000000	918.000000	30.400000	59.400000
75%	189.000000	2241.000000	66.000000	105.900000
max	10000.000000	187200.000000	100.000000	263.300000

Figure 1

Boxplot of Occupancy Rate By Superhost Status

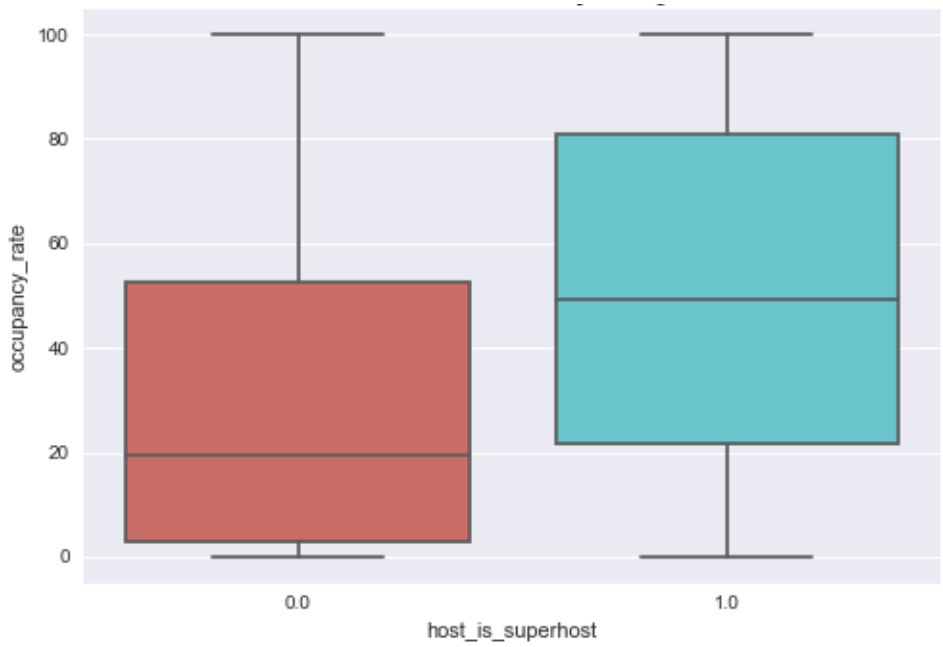


Figure 2

Boxplot of Revenue By Superhost Status

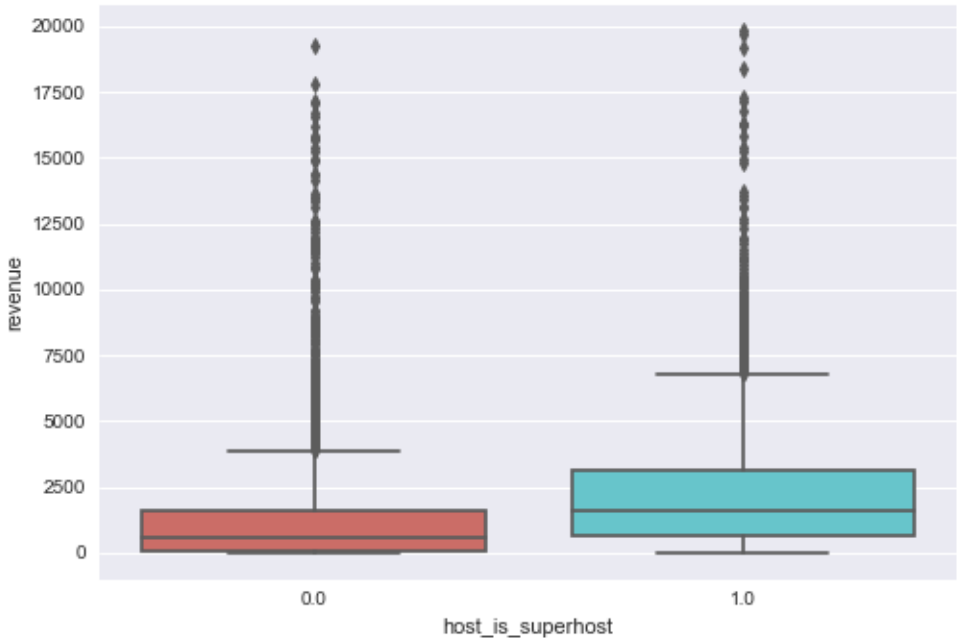


Figure 3

Boxplot of Price By Superhost Status

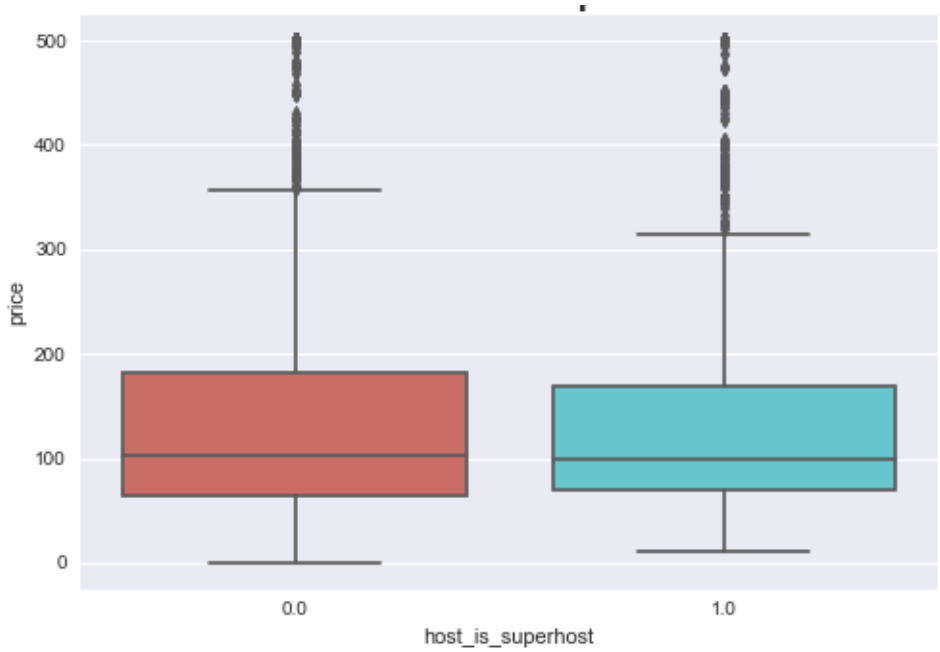
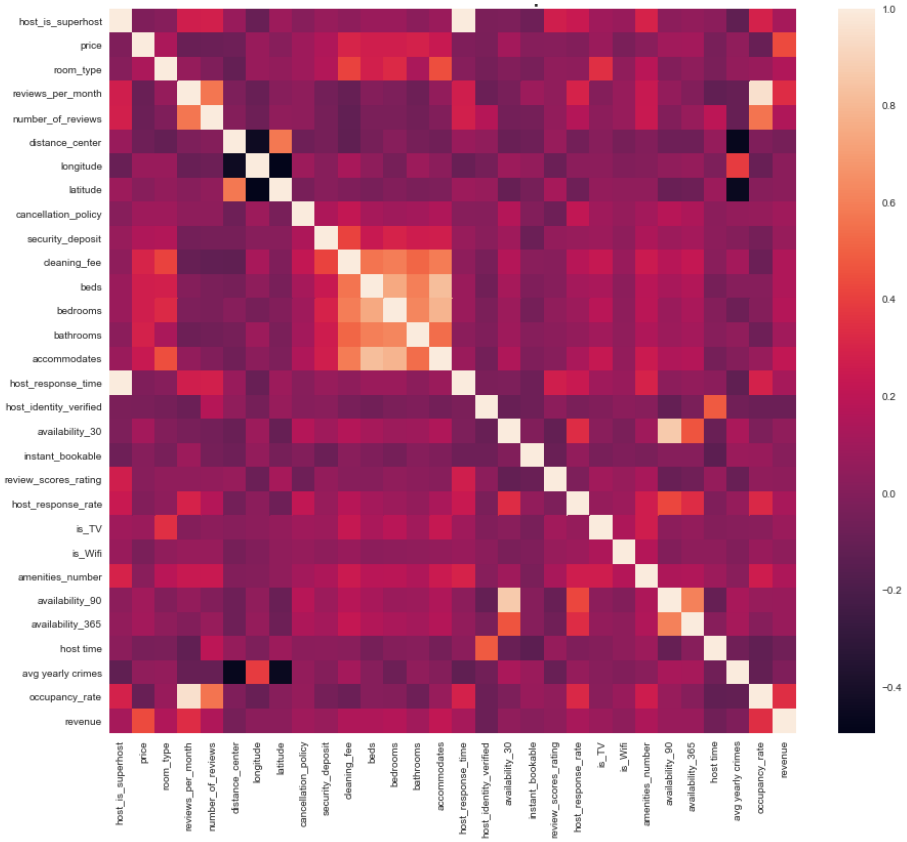


Figure 4

Correlation Heat Map for Main Variables

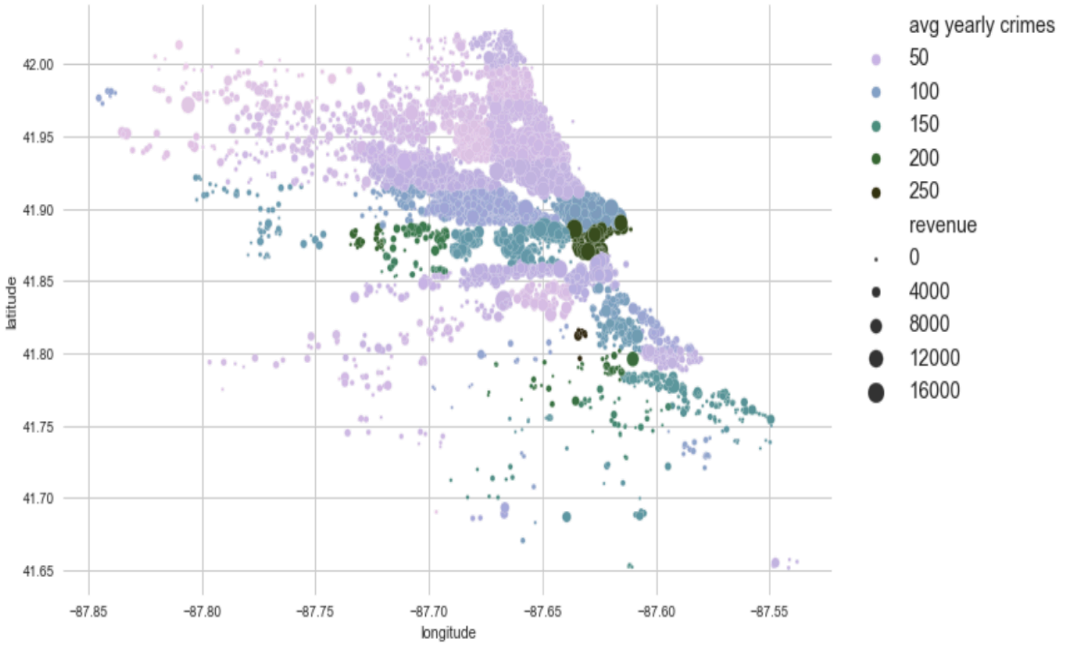


From Figure 4, it's obvious that there is a positive correlation between revenue and occupancy rate, while price doesn't seem to be correlated with the superhost status. Price appears to be influenced by cleaning fees and other variables demonstrating how comfortable the room is for living, including number of bedrooms, bathrooms and other amenities. Besides, even though superhost status does not seem to be correlated with rent rate, it is positively correlated with various review rating scores and revenue, indicating the potential difference between superhosts and normal ones. On the other hand, the average_yearly_crimes is indicated to be negatively correlated with center distance, which means that the farther the apartment is from the city center, the higher crime rate there is around the neighborhood. Figure 5 sheds light on the possible effect

of crime rate on host revenues. Somewhat counterintuitively, we haven't seen any negative effects of average yearly crimes per thousand residents on host revenues, room price or occupancy rate. Overall, we've had a big picture of potential relationships among the key variables, then we'll take advantage of more advanced geospatial data analysis to reinforce our understanding of the data.

Figure 5

Accommodations in Chicago by Crime Rate and Revenue



Methods

To answer the first research question on whether the superhost badge can live up to comsumers' expectation, we analyze the review text data for all the listings to learn about what they care about most with respect to a listing, and then check whether these aspects can serve as important predictors to determine a superhost. First, we divide the review dataset into 2 parts, one with review scores over 90 and the other with scores lower than 90. Then we remove the meaningless words from the review text using the

pyspark kernel in EMR[6] jupter notebook, which uses the Amazon Web Service's (AWS) clusters to parallelly speed up data wrangling. To count the top 100 frequent words respectively from the 2 wrangled review datasets, we use the python packages, MRJob and MRStep, to parallelly finish the work. Finally, we plot the 2 word clouds against frequency for both lowly and highly scored review dataset. Without the help of these large-scale computing techniques, we would not be able to finish the analysis in a reasonable amount of time. After knowing which aspects of a listing consumers care about most, we run the logistic regression and random forest models in EMR juper notebook using the pyspark kernel to check whether these attributes are important predictors to identify a superhost. In this way, we can see whether the listings of superhosts can really satisfy consumers' demands.

For the second research question, we try to answer whether superhost accreditation makes superhosts incline to setting a higher listing price through the linear regression model with price as the predicted variable and superhost status as one of the key predictors. Then to check whether occupancy rate is the critical channel through which superhosts can benefit more from the market (research question three), we first run the random forest model with revenue as the predicted variable and others including occupancy rate as the predictors to see the relative importance weights of each predictor and then compare the performance of regular and super hosts in these aspects through the Welch's t-test statistics. In this way, we can check whether occupancy rate is an important predictor of hosts' revenue and then if so, whether superhosts and regular

[6] Amazon EMR is a managed cluster platform on AWS where big data frameworks, such as Apache Hadoop and Apache Spark, are simplified to process and analyze vast amounts of data.

ones have a significant difference in occupancy rate.

All the machine learning models are built and ran using pyspark.ml packages in the EMR jupter notebook of AWS, and follow basically the same procedure. First, we vectorize all the variables, then specify the specific model along with the predictors and predicted variables, put all these into a pipeline, split the dataset for training and testing purposes, and finally train and test the machine learning models.

Finally, for the fourth research question on whether the conclusions above still hold in the context of high crime rate and conspicuous residential segregation, we first conduct a geospatial data analysis to identify potential geospatial patterns of superhost density, crime rate, black population density, rental price and occupancy rate, and then use multilevel modeling to estimate community effects and yield a more precise evaluation. We use R to build the models, where the explained variables and explaining ones are basically the same as those linear regression models except that we conduct centered and uncentered versions[7] of models and that we add the community variable to indicate to which neighborhood each listing belongs, and thus estimate the community effects including segregation and crime rate.

[7] We center the predictors because we expect the within-community slope and the between-community slope are expected to deviate, and our main interest is in the within-community slope; we also conduct the uncentered one so as to better estimate the effect of a level 2 predictor (community). We check whether the conclusions still hold in both versions of multi-level modelling.

Geospatial Data Analysis

The Context of Crime Rate

As another dimension of the research, here we conduct a geospatial data analysis to explore the potential relationship between crime rate, room price and superhost density (sphost_ds) in each community of the city. From the box map of sphost_ds (Figure A.1), we can clearly see the distribution pattern of sphost_ds: communities with a high superhost density are concentrated around the downtown area (northeast of the city), while those in the southern of the city have very low superhost density. We can see a similar pattern in the box map for the average room price (Figure A.2). However, in terms of crime rate (Figure A.3), the high crime rate appears in both the downtown area as well as the southern of the city. The conditional map (Figure 6) reveals a more obvious relationship between the three variables: the communities with higher superhost density tend to set higher room price, whereas crime rate does not seem to have an effect on rent. The pattern is further confirmed by the parallel coordinate plot (Figure A.4). From the scatter plot matrix (Figure A.5), crime rate is confirmed not having a significant effect on the average room price in the community.

Then we draw a scatter plot for sphost_ds and price (Figure 7) to further explore the relationship in detail, and we can see that in general, one unit increase in superhost density can lead to 31.500 dollars of increase in the average room price in that community, and that the positive effect is significant with a p-value less than .001. Besides, we also find that when the superhost density in the community exceeds about 3, a different effect becomes dominant: one unit increase in superhost density leads to

about 14.22 dollars of decrease in the community's room price. However, the negative effect is not significant with a p-value of 0.561. The result indicates that while superhosts themselves do not necessarily set higher rent rate, high superhost density gives hosts in the corresponding communities more market power to set higher room price, through which they can probably earn more (the hypothesis still needs to be checked). The result of the simple exploration is in line with the signaling theory first proposed by Michael Spence. The superhost status probably serves as a positive signal of the service quality in the market and thus boosts the Airbnb business in the City of Chicago.

Figure 6

Conditional Map for price

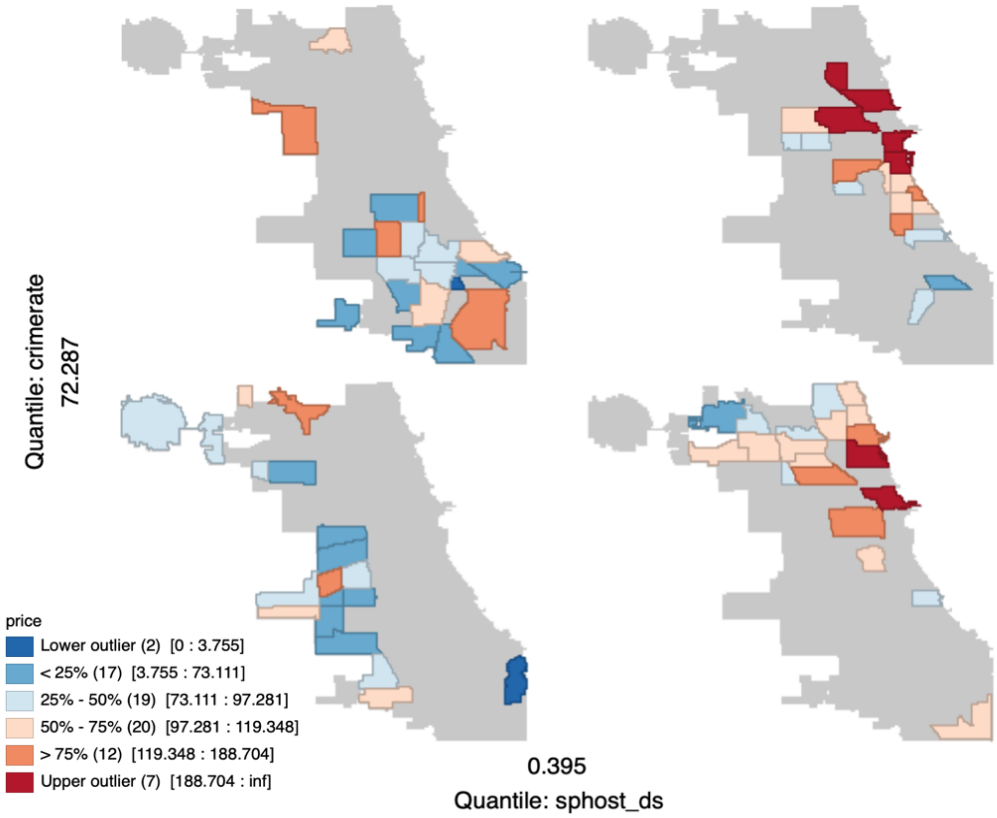
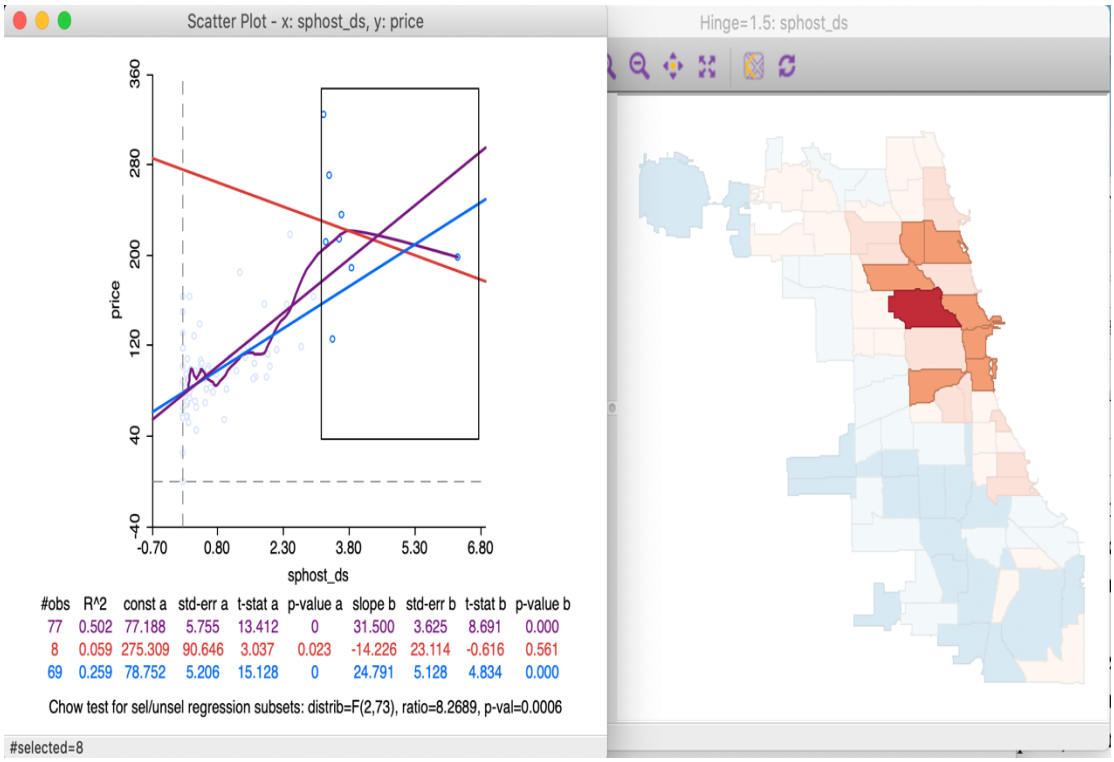


Figure 7

Scatter Plot with Map Brushing



The Context of Residential Segregation

Residential Segregation is still a conspicuous phenomenon in Chicago, which can be observed in Figure A.6. Communities with low black population ratio (black_pct) tend to appear in the northeast and middle west, while those with high ratio appear to be segregated in the south. Regarding the distribution of occupancy rate (Figure A.7), communities in relatively north tend to have higher occupancy rate than those in the south. However, looking in more details, high occupancy rates also appear in some southern communities whereas low occupancy rates appear in some north communities. To have a big picture of the general pattern for the four variables, we put the 4 box maps together and select the observations with extremely high superhost density (Figure A.8),

we can see that communities with a high superhost density and room price but low black population ratio gather around the northeast of the city, while occupancy rate seems not to have a clear association with the superhost density distribution. Then we plot the conditional maps (Figure 8) to further explore the relationship among the 3 variables, price, superhost density and black population ratio. We observe that price is positively associated with superhost density, while black population ratio does not have clear-cut associations with room price. However, a weak potential pattern is still observed that in communities where black_pct is high, extremely high room price is less likely to appear in communities even if the superhsot density is high.

In the bubble chart for sphost_ds and price (Figure A.9) with dark orange as large black_pct and light blue as small black pct, we can see that among communities with small black population proportion, room price increases with superhost density and that communities with large black_pct have an upper bound of room price and superhost density. This indicates that high black_pct can exert a negative influence on room price and superhost density, which matches our inference from Figure A.8. This can be conversely verified by the parallel coordinate plot (Figure B.1), communities with high superhost density and price only corresponds to low black population proportion.

Then we explore the bivariate relationship between sphost_ds and black_pct in more depth (Figure 9). Generally, black_pct is negatively associated with sphost_ds, with a very significant negative slope of -8.818. By selecting the lower black_pct bandth (less than 30 percent), we see that there is a significantly positive association between sphost_ds and black_pct, meaning that when black population ratio is low,

communities with higher superhost density tend to have slightly higher black population ratio and that when black_pct is relatively high, communities with higher superhost density tend to have lower black population ratio. The p-value of the Chow test statistic is <.001, meaning the difference between the selected and unselected part is significant. The observation consolidates our previous inference in Figure 8.

Figure 8

Conditional Map on black_pct and sphost_ds

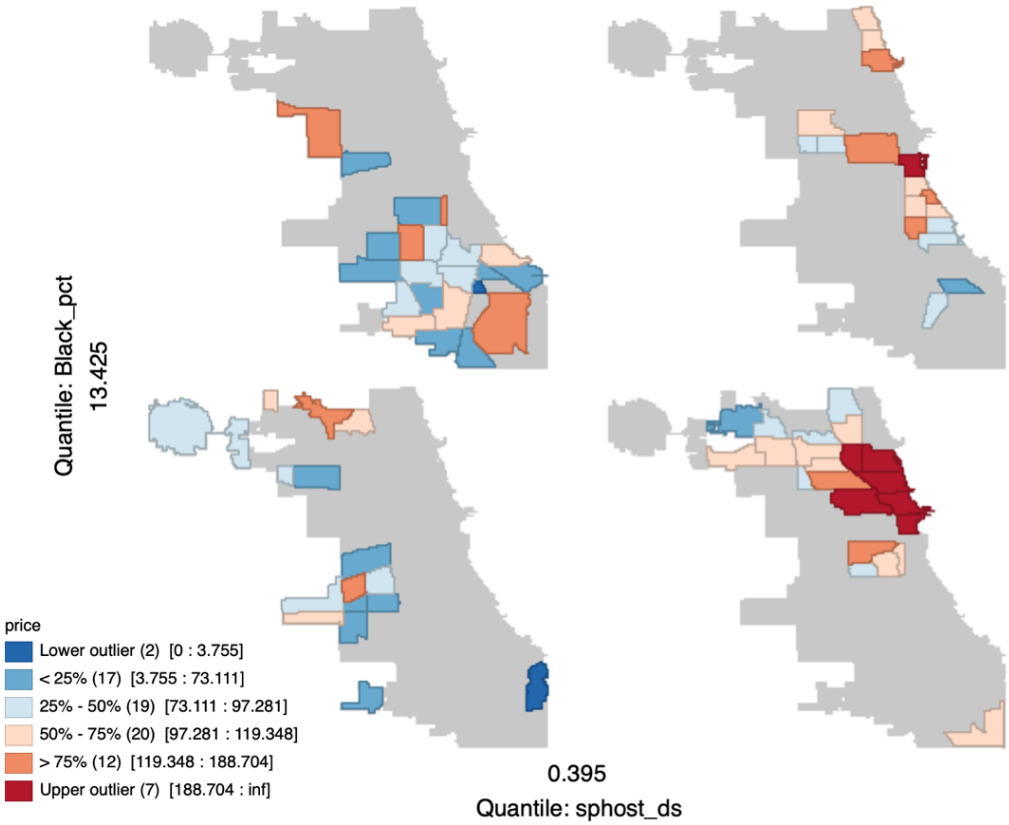
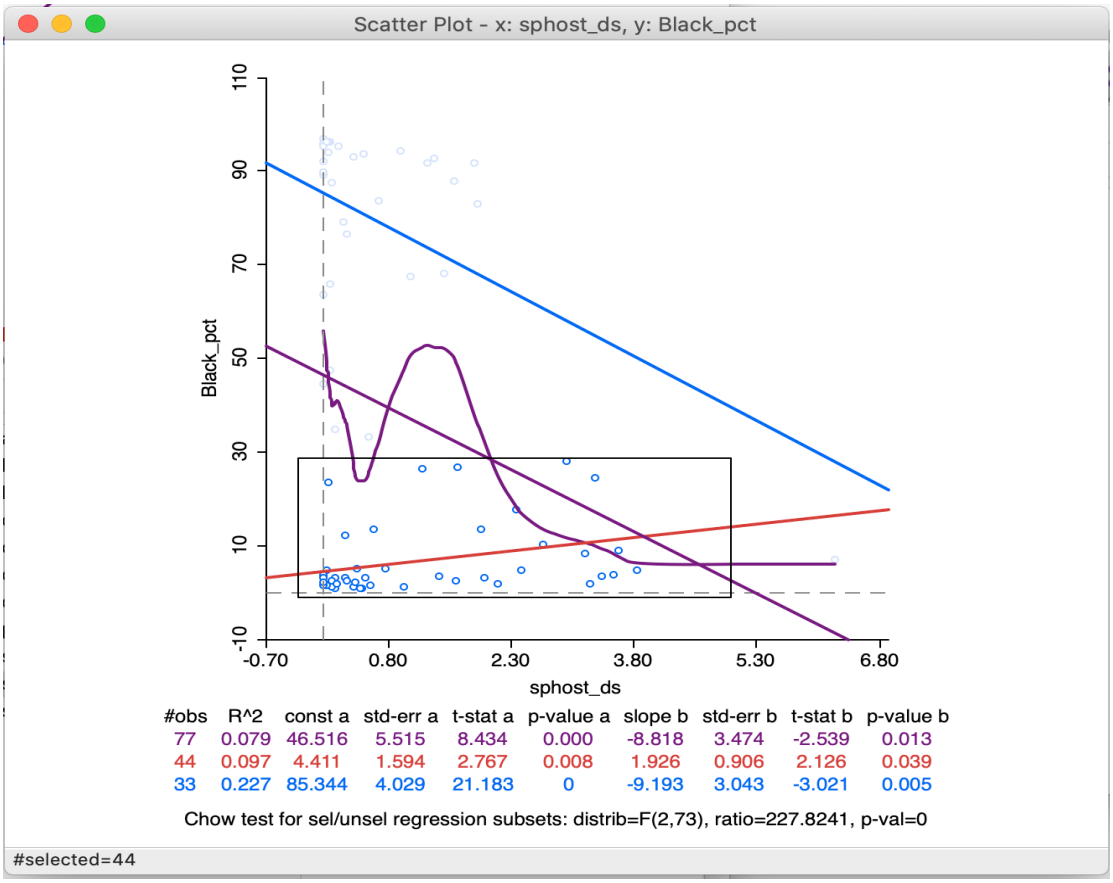


Figure 9

Scatter Plot for black_pct



Spatial Autocorrelation Analysis

In this section, we explore the spatial autocorrelation for the community superhost density, black_pct and price to see whether there are any local clusters or spatial outliers.

Here, we prepare two geographies. One is the original Chicago community polygon layer, and the other, the centroid point layer (Figure B.2) derived from the community map. Using alpha_id as the id variable, we respectively compute the queen contiguity weights, great-circle-distance-band weights (using arc distance as the distance metric due to simple latitude-longitude projection) and KNN weights (setting 6 as the number of neighbors) for the 2 most interesting variables, sphost_ds and black_pct. To explore

the local spatial autocorrelation in the maps, we employ the Univariate and Bivariate Local Moran, Univariate and Multivariate Local Geary as well as Multivariate Quantile LISA analysis.

The cluster map (Figure B.6) augments the significant locations with an indication of the type of spatial association, based on the location of the value and its spatial lag in the Moran scatter plot. Here, we analyze the superhost density cluster map derived from the queen contiguity weights since the patterns are consistent and stable across the 6 maps. By the default setting, 999 permutations are performed and the p-value is 0.05. In this example, three out of four categories are represented, with red for the high-high clusters (15 in our example), dark blue for the low-low clusters (19 locations), light blue for the low-high spatial outliers (1 location), and light red for the high-low spatial outliers (0 location). The pattern is that communities with a high superhost density tend to have neighboring communities with also high superhost density more than spatial randomness. On the other hand, low superhost density communities also tend to be surrounded by low density communities more so than would be randomly, but only a very few communities with low superhost density are surrounded by high density neighbors. In addition, high-high spatial clusters tend to appear in the northeastern of the city, while low-low clusters tend to gather around the southern and southwestern of the city. The low-high cluster is in the Humboldt Park community. By further exploration, we find some other interesting patterns to explain such phenomenon.

The cluster map for black population proportion per community (Figure B.7) also

has three out of four categories represented. The high-high clusters appear in the northern and middle part of the city, while the low-low clusters appear in the southern part of the city. This means that black communities tend to be surrounded by black ones more so than would be randomly, vice versa. Such a pattern indicates a conspicuous race segregation in the city: black people mainly gather around the south. Not coincidentally, the black_pct high-high clusters take place where superhost_ds low-low clusters appear. This further indicates that black people are less likely to be superhosts. Therefore, if superhost density somehow represents a greater chance to earn more via the Airbnb rental platform, black people seem to have less economic gains from the new economy. However, a causal relationship still remains to be confirmed, and such an inference needs more rigorous verification.

Figure 10 plots the Bivariate Local Moran cluster map for sphost_ds and black_pct, and confirms our observation in Figure B.6 and B.7. There is a strong low-high cluster in the southern. However, the interpretation of the bivariate Local Moran cluster map warrants some caution because this could either be attributed to black population proportion in surrounding locations, impacting the central location's superhost density or central location's black population ratio and affecting its own superhost density. A bit counterintuitively, there are also strong low-low and high-low clusters in the north, which means that both low and high superhost density communities in the north side can be surrounded by low black density communities more so than would be randomly.

The slope of the linear fit from the Moran scatter plot (Figure B.8) is -0.213, implying a negative spatial autocorrelation. Also, the significance map (Figure B.9)

with commarea_q as the spatial weight (queen contiguity weight) demonstrates that the 10 locations are securely significant with p = 0.001.

Next, we further explore the patterns by drawing the Local Geary cluster map (Figures 11 and 12 using queen contiguity and KNN weights respectively) for each individual variable. The result confirms a match among the three individual variables in both geographic and attribute space. Here we set permutations = 99999 and p = 0.05, and find that strong high-high clusters for sphost_ds and price correspond to the low-low clusters for black_pct in the northeast of the city and that a strong low-low cluster for sphost_ds and price corresponds to a negative cluster for black_pct in the southwest. Interestingly, we also find a common high-high cluster for the 3 variables in the middle west of the city. Moreover, for sphost_ds, there is a cluster being classified as other positive in addition to low-low and high-high.

The big picture is summarized in the co-location map in Figure C.1, which is constructed by first turning the significant locations in each significance map into an indicator variable for the selection, and then using these three variables in a co-location map. We can easily find the overlap among the significant locations distributed in the 3 Local Geary Univariate cluster map. Here 5 locations are identified as significant for all 3 variables. Moreover, 10 locations are discerned as commonly insignificant and 62 locations are only significant for less than 3 of the variables.

Then the Multivariate Local Geary functionality is employed. In our example, a p-value of 0.001 also corresponds roughly to the False Discovery Rate (FDR). From the significance and cluster maps (Figure C.2), we illustrate the effect of tightening the p-

value from 0.05 to 0.001. The number of significant location is reduced from 59 to 24. The significant locations cover all the 4 positive clusters identified in the previous co-location map.

However, the results of the Multivariate Local Geary significance or cluster map should be treated with caution because the multivariate statistic involves various tradeoffs. The significance map and a 3D data cube (Figure 13) are drawn to illustrate the complexity: observations in question that are very close in multi-attribute space still have some of their neighbors much further removed. Also, the PCP (Figure 14) can further illustrate the tradeoffs: In this instance, while the distances between the paths are small, there is much less of a close match. Then we select 3 highly significant neighboring locations in the right-hand panel, and we find that the neighbors track each other better and meets our expectation in the multivariate attribute space.

At last, we conduct the Multivariate Quantile LISA analysis by setting sphost_ds, price and black_pct as the 3 variables, as shown in Figure 15. By the default setting (999 permutations and p = 0.05), we see that 5 out of the 77 locations are identified as the cores of actual clusters. Then we set the significance level as the False Discovery Rate (0.0166667) and do 99999 permutations. The result (Figure 16) remains the same as before. Finally, we construct the co-location map (Figure 17) using the QT_sp, QT_pr and QT_blk (the fifth quantiles for sphost_ds, price, and black_pct) derived from the Multivariate Quantile LISA process. However, it turns out that none of the locations in the significance map can be recognized as the spatial clusters of co-locations. This is in line with our economic intuition, because while a high superhost density is associated

with a high room price, the high values of these 2 variables are usually associated with a low black population ratio in the communities. Therefore, we plot the co-location maps (Figure 18) for the fifth quantile of sphost_ds and price along with the 2nd and 3rd quantile of black_pct. And it turns out all the 5 significant locations in the significance map have corresponding significant locations in the co-location map, indicating a special spatial association among the 3 variables. The highlighted 6 communities gather around the northeast of the city. High superhost density, high room price and low or medium black population ratio tend to happen simultaneously in these locations. However, our research cannot identify any causality, which demands other rigorous social science methods. It's also worth mentioning that the quantile LISA method can cause loss of information due to concentration on extremes of the variable distributions.

In conclusion, our analysis indicates a strong spatial association among the superhost density, room price and black population ratio. More specifically, a community with a high superhost density is more likely to be surrounded by communities with high average room price and low or medium black population ratio. This implies conspicuous residential segregation in the city of Chicago, where black people have inferior economic opportunities in the emerging 'sharing economy'. On the other hand, we cannot see any spatial associations between the occupancy rate and the other 3 variables. Besides, regardless of the trivial discrepancy between Multivariate Local Geary and Multivariate Quantile LISA, they yield almost the same pattern, demonstrating the robustness of our analysis. Nevertheless, our analysis cannot

give a causality inference.

Figure 10

Bivariate Local Moran Cluster Map

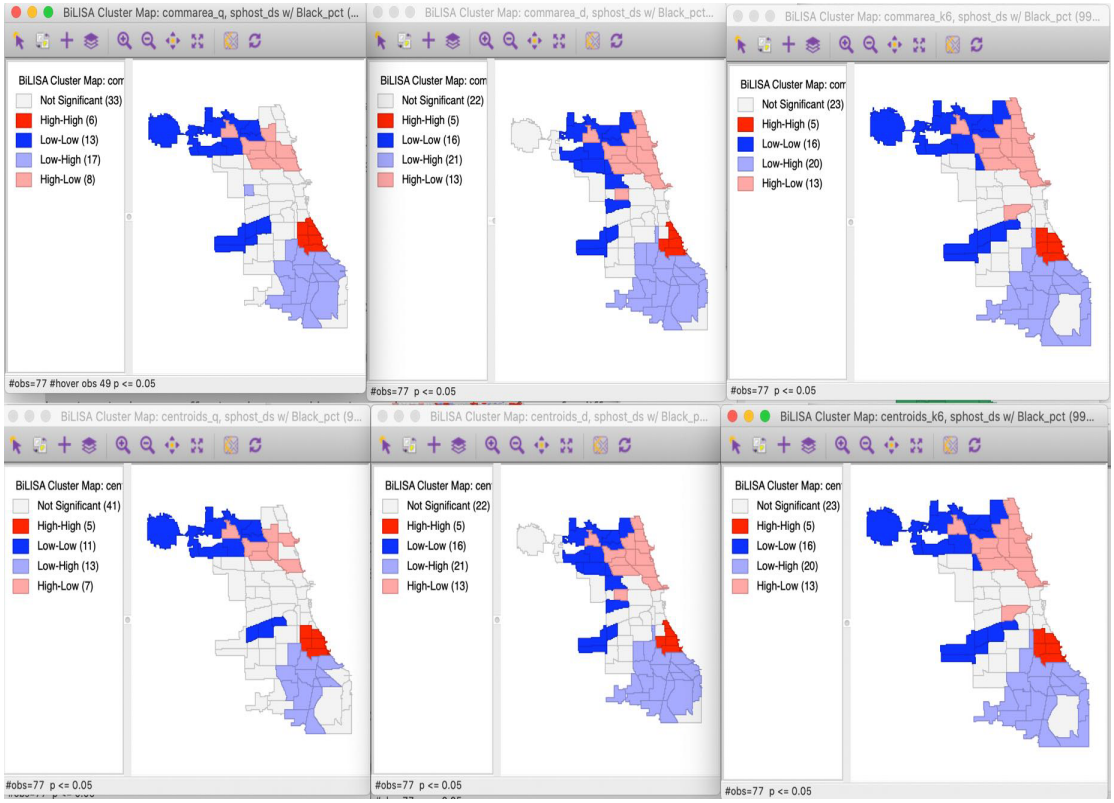


Figure 11

Univariate Local Geary Cluster Map (KNN)

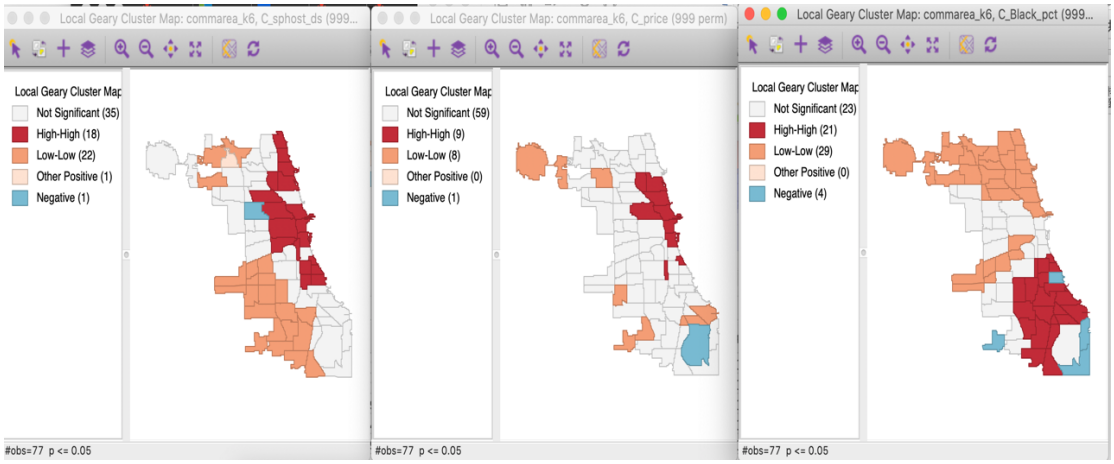


Figure 12

Univariate Local Geary Cluster Map (Queen)

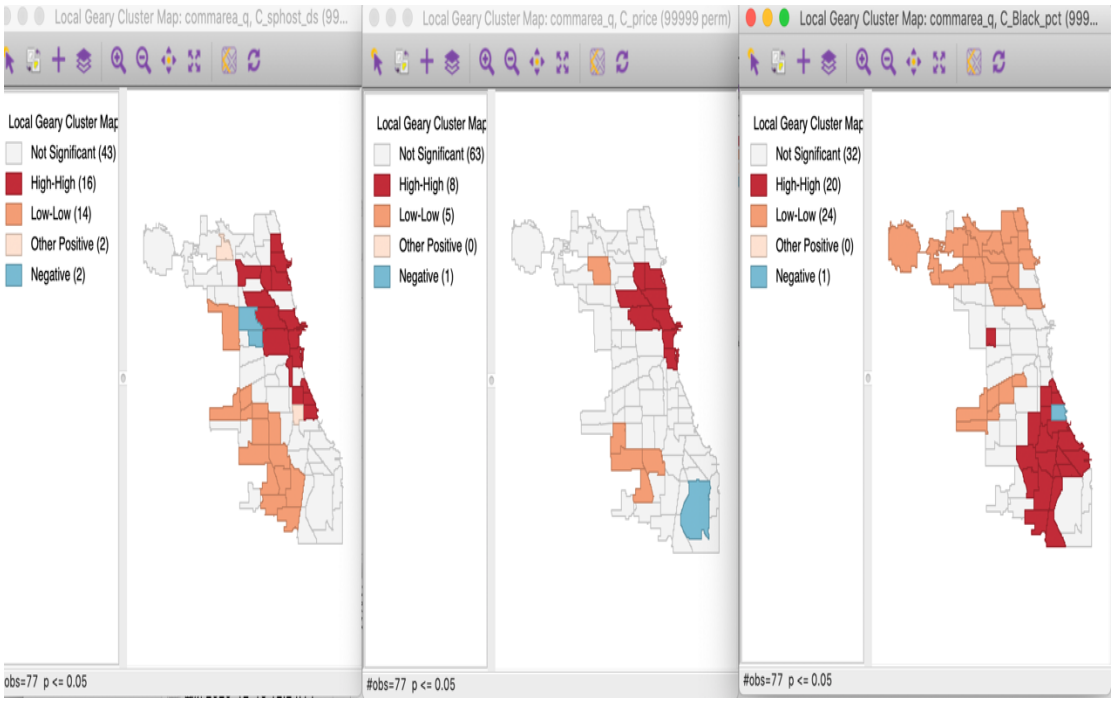


Figure 13

3D Plot

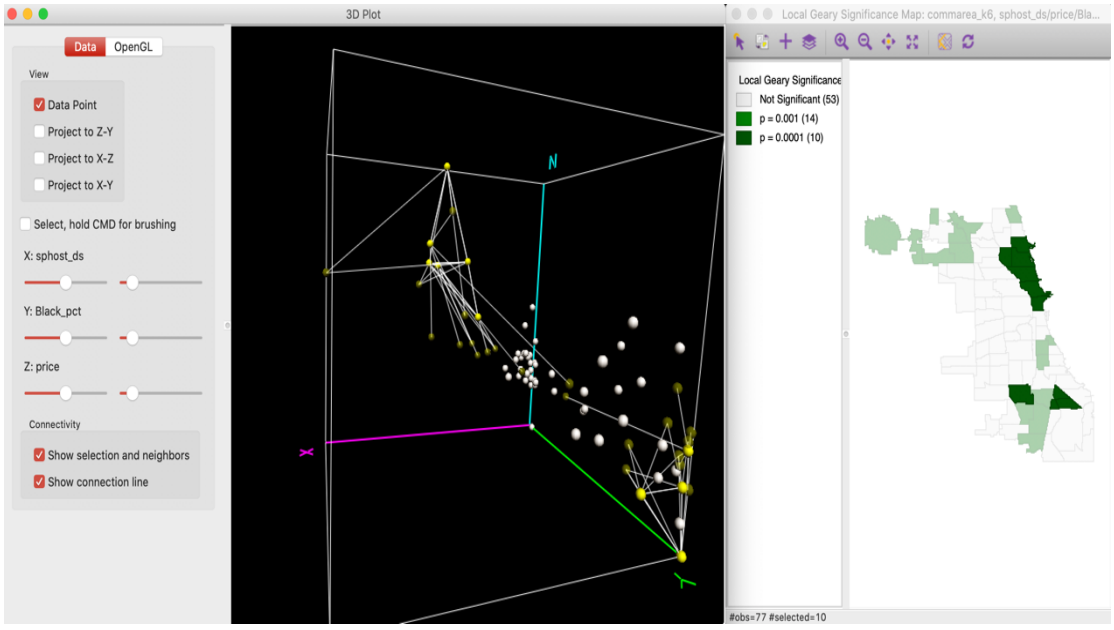


Figure 14

Parallel Coordinate Plots

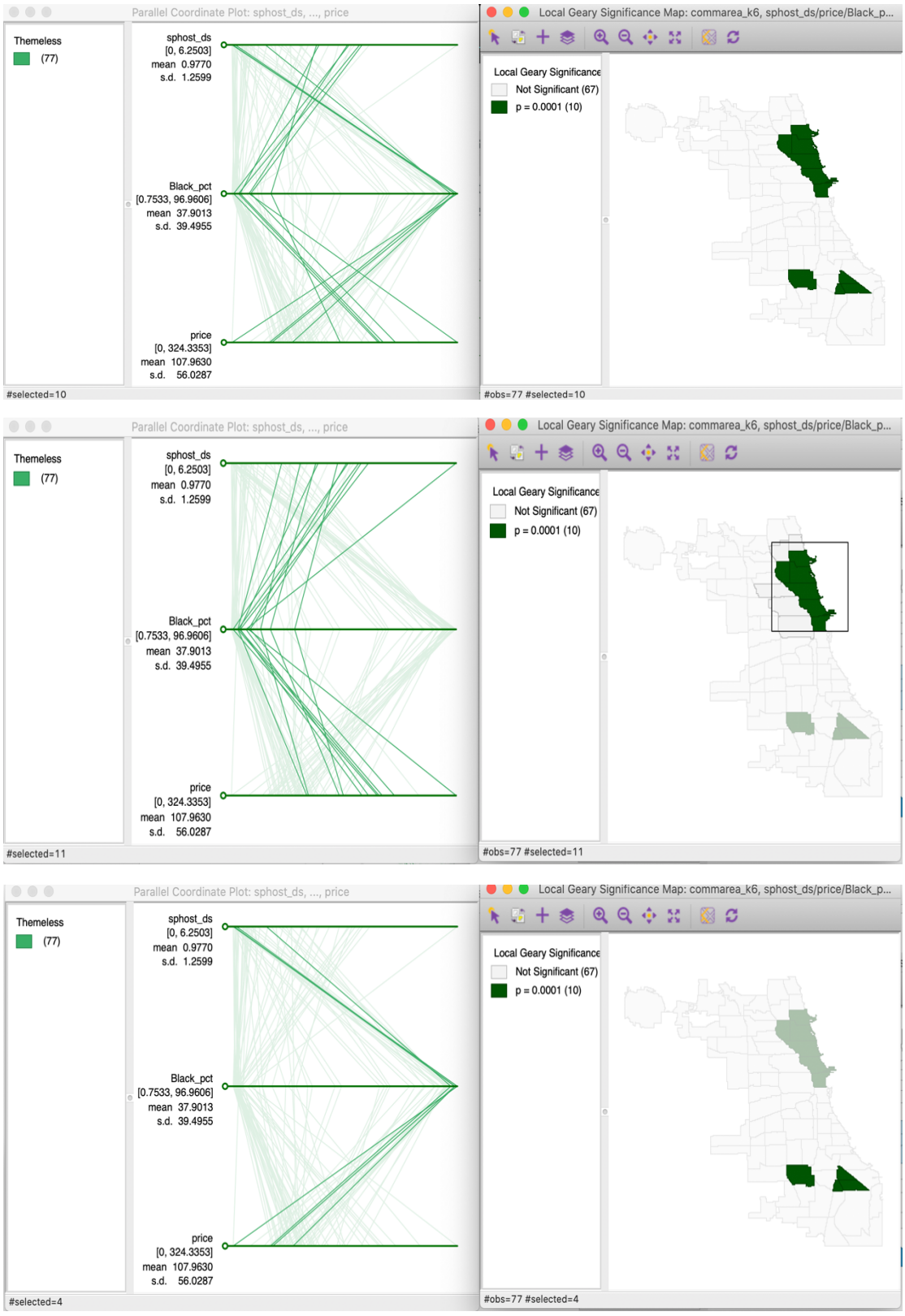


Figure 15

Mutivariate Quantile LISA

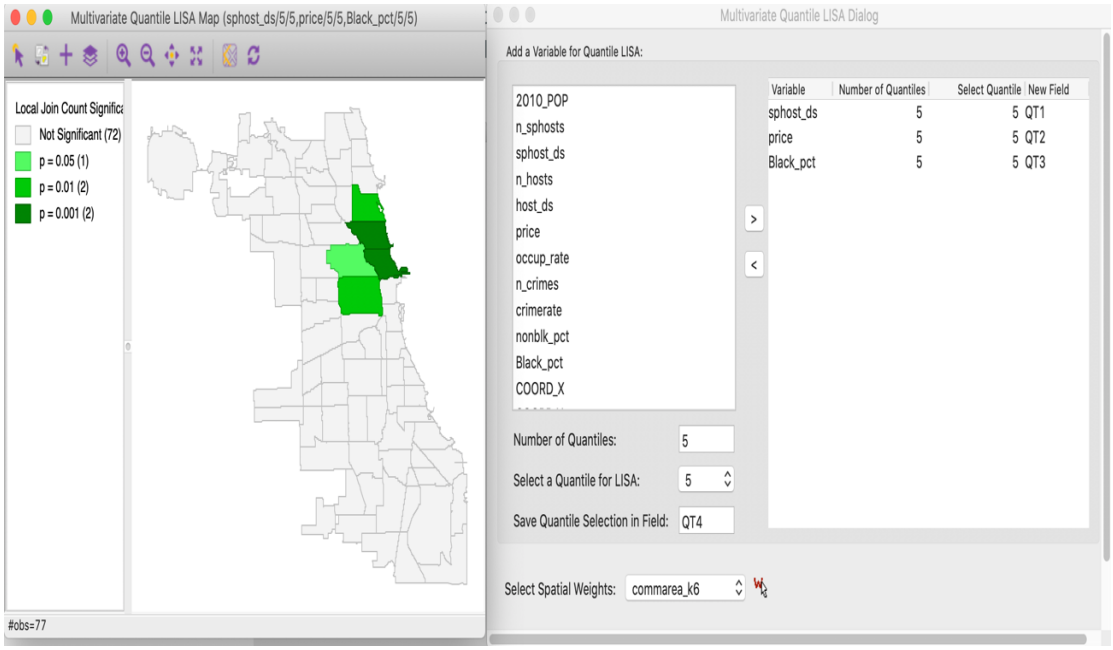


Figure 16

Mutivariate Quantile LISA (with FDR)

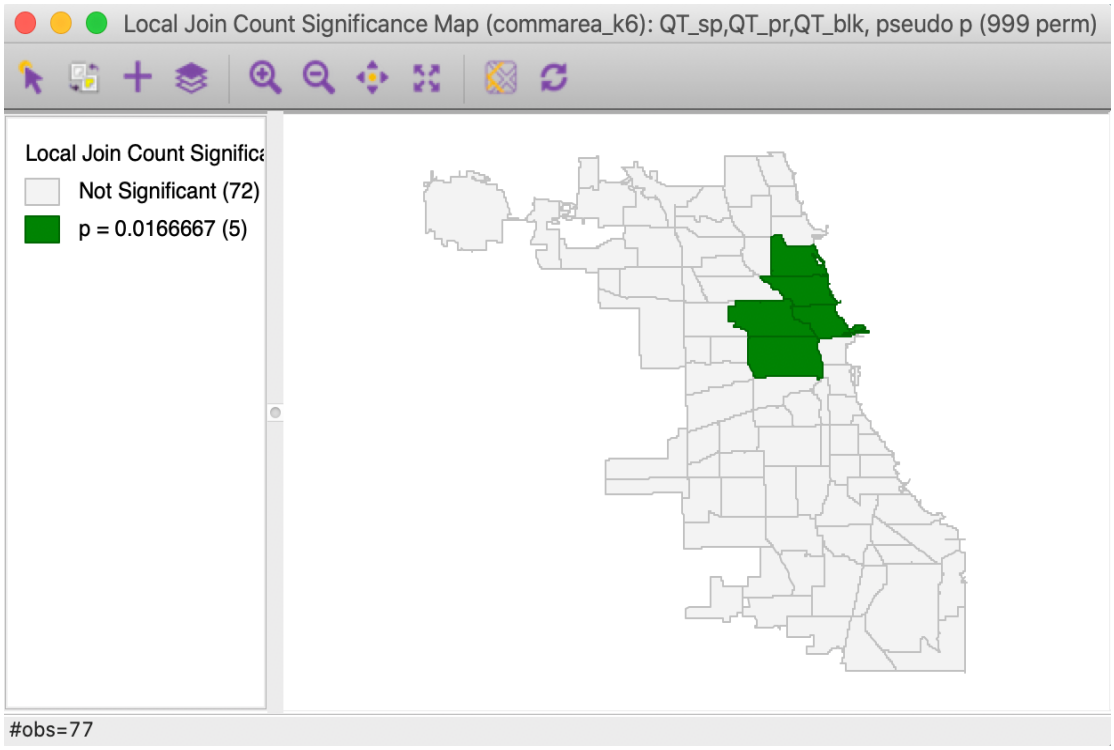


Figure 17

Co-location Map (All 5th Quantile)

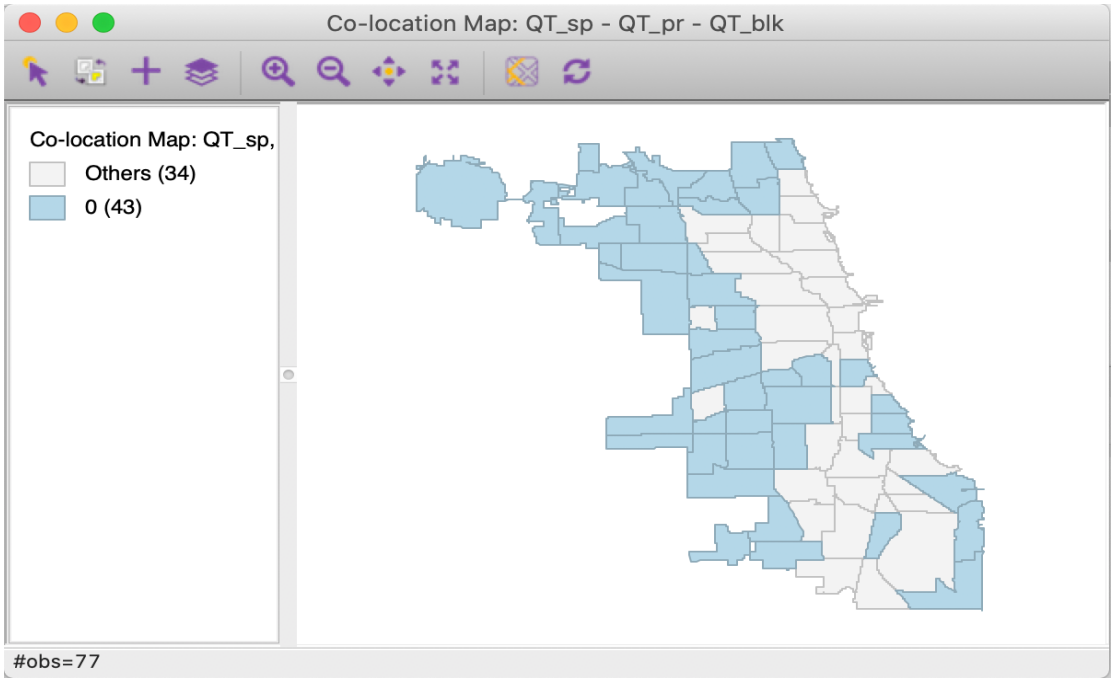
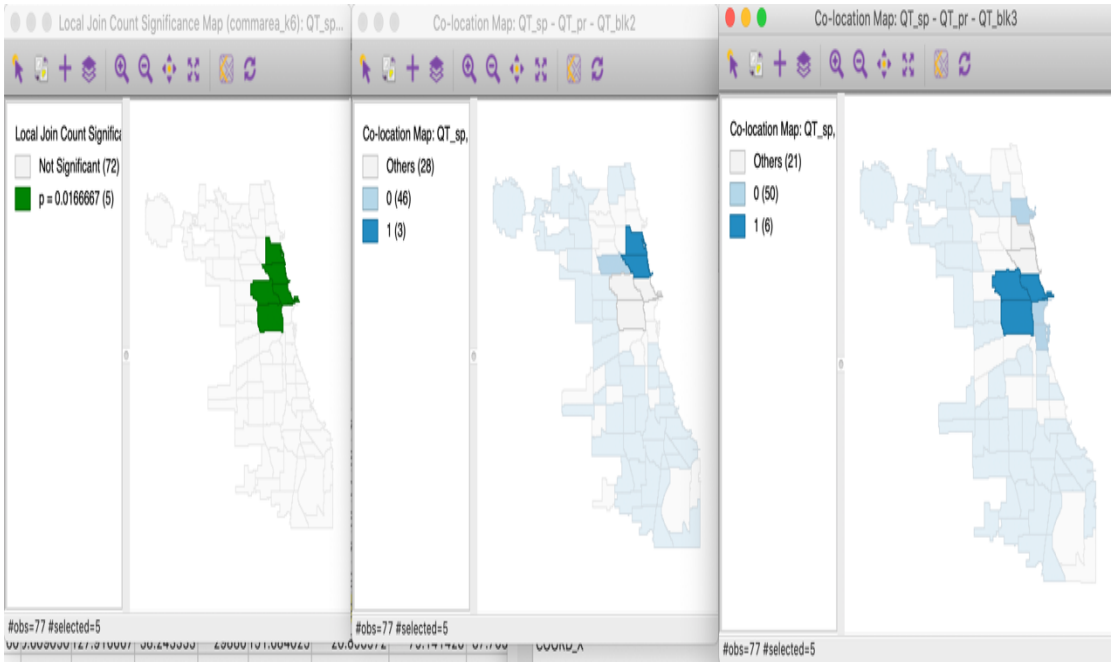


Figure 18

Co-location Map (2nd and 3rd Quantile of black_pct)



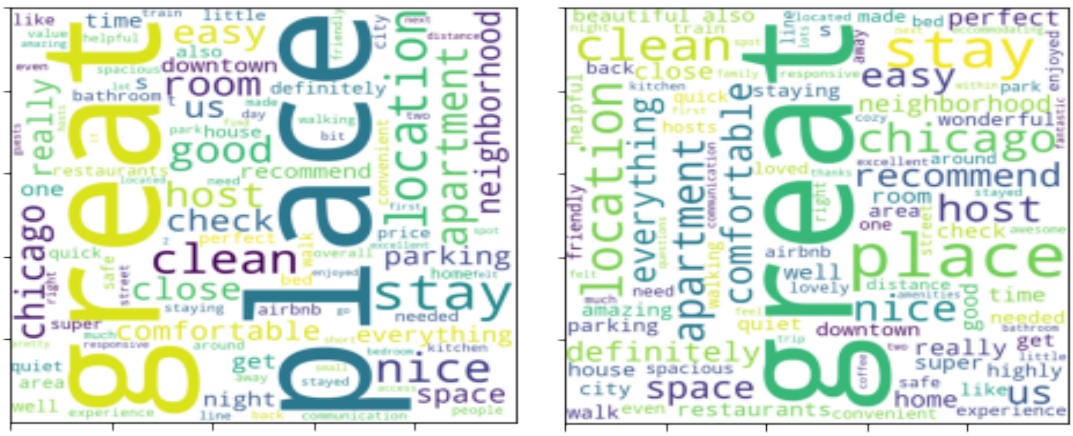
Analysis and Result

Comment Analysis

To see what consumers care about most in a Airbnb room, we export the review data as txt files and use mrjob[8] to find the most frequent 100 words in the reviews with review score ratings over 90 (out of 100) and those with review score ratings less than 90. In this way, we can see what consumers hate and like most in a listing. From the 2 word clouds against frequency, while we cannot see significant difference between the 2 cohorts, we see they share commonalities that the 2 groups of consumers attach great importance to geographic locations, and cleaning conditions of a room. They prefer apartments with kitchen, beautiful scenes and quiet environment. On the other hand, they cannot bear with rooms which are dirty, have difficult access to parking, and don't have easy-to-communicate hosts. While MRJob enables us to quickly analyze the comment words, such results can only give us a rough picture of consumers' thought. We'll then use other pyspark machine learning models to answer our questions.

Figure 19

Word Cloud for Reviews with Low & High Rating Scores



[8] Mrjob is python package that lets you write MapReduce jobs in Python 2.7/3.4+ and run them on several platforms.

Superhost Classification

In order to identify the difference between superhosts and normal ones, here we use superhost status as the predicted label, and price, room type index, reviews per month, cleaning fee, bathrooms, host response time, availability_30, review scores rating, revenue, amenities number, occupancy rate, host identity verified index, instant bookable index, cancellation policy index as predicting features to build logistic and forest random classification models.

Logistic Regression Classification Model

We use a 5-fold cross-validated grid search to identify the optimal regression parameters and elastic net parameters for the logistic model. Our optimal model has a training and testing area under the receiver operating characteristics (ROC) curve (AUC) of around 0.803. The testing accuracy is about 0.736. This indicates a good[9] predictive performance of our model. Therefore, we can roughly conclude that price, occupancy rate, revenue, and other features indicating the room's basic amenities combine to serve as good predictors of superhost status. However, we have no idea which of them are the most and least valid predictors. We answer this question by applying the random forest model.

[9] The accuracy of 0.736 is regarded as good by the author because the model can correctly identify superhosts for 73.6% of the listings in the testing dataset.

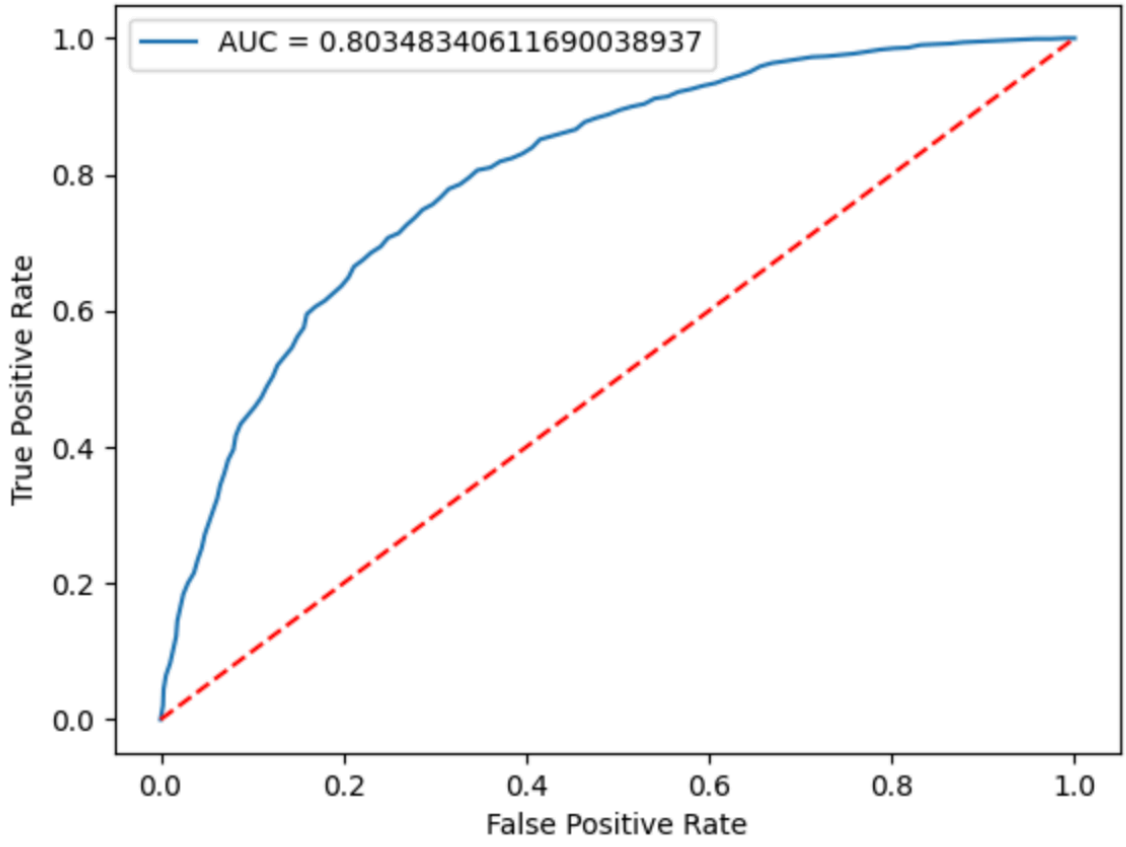
Table 2

Evaluation Result

Training AUC	0.809
Testing AUC	0.803
False positive rate by label (Training)	
label 0	0.380
label 1	0.180
True positive rate by label (Training)	
label 0	0.821
label 1	0.620
Training Accuracy	0.743
Testing Accuracy	0.736

Figure 20

ROC Curve



Random Forest Classification Model

It is not enough to develop a rough knowledge of features that can distinguish superhosts from normal ones; we also need to know which attributes play the more

important roles. Therefore, we use the random forest model, which, in essence, is an ensemble of decision trees, with each tree trained on a random subset of data. By exploiting this model, we can not only find the best decision tree model, but also identify the importance of each feature in our model. Here, we find that review scores rating, occupancy rate, and the number of amenities in the room are the 3 most important features to identify a superhost. Surprisingly, price is an unimportant feature to identify superhosts with an importance weight of 0.0269. Therefore, we assume that price is not the primary channel through which superhosts can benefit from their status, and that superhosts benefit from the top-performing accreditation mainly by maintaining a higher occupancy rate.

Table 3

Result of Random Forest Model for Price Prediction

Feature	Weight
bathrooms	0.0008
room_type_idx	0.0036
instant_bookable_idx	0.0036
availability_30	0.0109
cleaning_fee	0.0212
revenue	0.0223
price	0.0269
cancellation_policy_idx	0.0456
host_response_time	0.0711
reviews_per_month	0.0807
amenities_number	0.1220
occupancy_rate	0.1275
review_scores_rating	0.4639

Tree-based Revenue Prediction Model

The previous section has shown that superhosts don't set higher prices than normal ones. We now further our exploration to check other channels through which superhosts

could gain more benefits. This time we use the non-parametric supervised learning method, random forest and decision regression tree, to determine whether superhosts gain more, along with the possible mechanisms. We first use the random forest model to find the best decision tree model, along with the importance of each feature in our model. Here, we find that occupancy rate, price, room type, number of reviews and availability in the next 30 days are the most important features to predict the host revenue. Part of the decision tree is visualized in Figure 20. Therefore, if we can demonstrate that superhosts are significantly different from normal hosts in these features, then we can conclude that superhost accreditation does influence host revenues. Since the number of normal hosts is much smaller than that of superhsots in our dataset, we compute the Welch's t-test statistics, which is more reliable in the scenario of two samples with unequal sample sizes or variances, to determine the differences are significant. Among these predictor variables, we find those (including occupancy rate, and number of reviews) which are significantly different between superhosts and normal ones at the 5% significant level, as illustrated by Table 5. Therefore, we conclude that superhosts can maintain higher occupancy rates, thus earning more.

Table 4

Result of Random Forest Model for Revenue Prediction

Feature	Weight
super_host	0.0002
host_identity_verified_idx	0.0002
instant_bookable_idx	0.0002
cancellation_policy_idx	0.0010
review_scores_rating	0.0017
bathrooms	0.0045
bedrooms	0.0058
beds	0.0082
amenities_number	0.0119
cleaning_fee	0.0187
accommodates	0.0264
host_response_time	0.0326
availability_30	0.0632
number_of_reviews	0.0784
room_type_idx	0.1205
price	0.1742
occupancy	0.4523

Figure 20

Decision Tree Model for Revenue Prediction

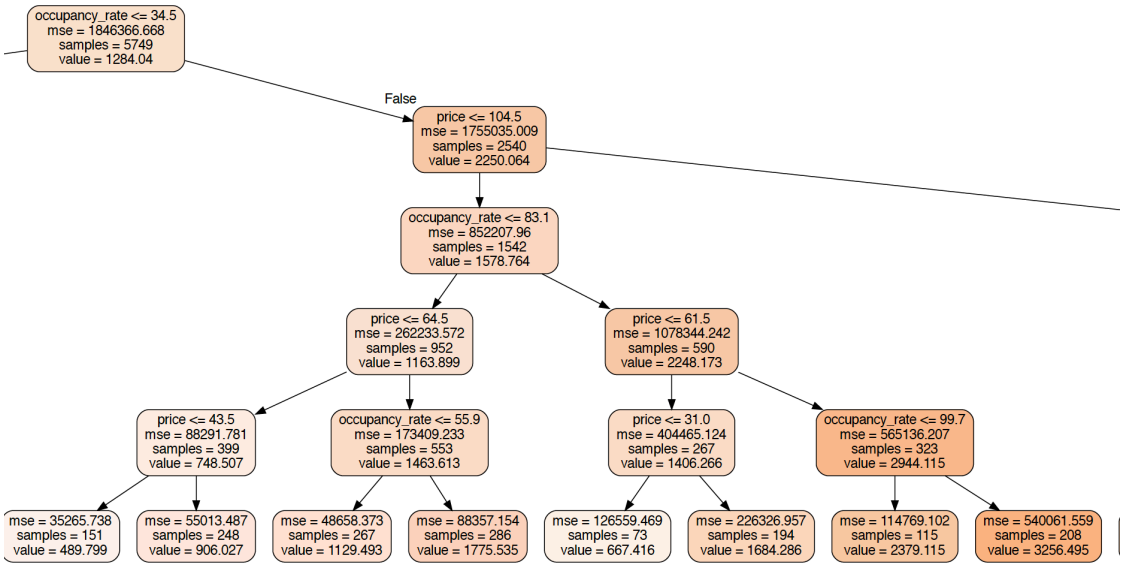


Table 5

Welch's T-test for Variables' Difference By Superhost Status

Variable	T Value	P Value
host_is_superhost	Inf	0.000000e+0
host_response_time	Inf	0.000000e+0
review_scores_rating	30.783945	6.041265e-196
amenities_number	28.617988	1.814308e-170
occupancy_rate	28.246235	1.922976e-166
number_of_reviews	24.733898	5.242382e-128
bedrooms	7.370896	1.879644e-13
accommodates	7.310573	2.982563e-13
beds	6.943957	4.212726e-12
cleaning_fee	3.762768	1.692601e-03
host_time	2.544214	1.097263e-02
bathrooms	2.446237	1.445767e-02

Linear Regression

Next, we use a linear regression model to do the robustness check. The models are specified as follows, where S is the superhost status[10], X includes all price-influencing variables, and X includes all revenue-influencing factors.

$$Price = \theta + \gamma S + \beta X + \mu$$

$$Revenue = \alpha + \lambda S + \beta X + \mu$$

Price

From the linear regression result, we can see that the effect of superhost status on price setting is not significant in Model 1 and 2. While the effect in Model 3 and 4 is significant at the 1% significant level, it is negative, which means that superhosts tend to set lower prices ($7.55 and $9.34 in Model 3 and 4 lower than that of normal hosts)

[10] S takes the value of 1 if the listing host is superhost, and 0 if not.

controlling for other variables. Besides, the number of bedrooms, and bathrooms exert significantly positive effects on price.

Revenue

From the comparison in the previous section, we know that superhosts maintain significantly higher occupancy rates. Now the regression model indicates that one percent increase in occupancy rate leads to an average increase of $36.43 in host revenue and that superhosts tend to earn on average $244.49 more than normal hosts, holding other factors constant. The two effects are both significant at the 1% significant level. We therefore conclude that superhosts can benefit from their accreditation by maintaining a relatively higher occupancy rate. Moreover, setting stricter cancellation policies is an effective way to increase income. It's also worth mentioning that renting out an entire apartment brings more revenues than renting private or publicly sharing rooms.

Crime and Residential Segregation

Even though Chicago has a notorious reputation for high crime rate, our results show that one percent increase in average yearly crimes per thousand residents leads to a $0.45 increase in price (Model 4), and $1.60 increase in host revenues (Model 7), holding the other variables constant. The crime effects on revenue is not significant; although significant in predicting price, the trivial economic effects verify our initial observation in the previous section that safety issues have not become a big concern for travelers enough to impair Chicago's Airbnb business.

Similarly, while the geospatial data analysis demonstrates that Chicago has a

conspicuous residential segregation, our regression results show that the black population ratio in each community has a significantly negative but trivial effect on listing price (Model 4: -$0.56) and host revenue (Model 7: -$3.18) above and beyond the other variables. Therefore, we conclude that residential segregation does not have a sufficiently negative effect on Airbnb business in the city of Chicago.

Table 6[11]

OLS Regression Results for Price Prediction

	Model 1	Model 2	Model 3	Model 4
const	140.48***	41.31*	-63.11***	-48.19***
	(4.03)	(22.18)	(18.36)	(18.27)
host_is_superhost	3.65	-2.07	-7.55***	-9.34***
	(2.98)	(3.14)	(2.59)	(2.54)
crm_rate	0.87***	0.77***		0.45***
	(0.13)	(0.13)		(0.10)
blk_pct	-0.43***	-0.46***		-0.56***
	(0.06)	(0.06)		(0.05)
sphost_ds	11.42***	12.36***		6.64***
	(0.93)	(0.92)		(0.75)
occupancy_rate	-0.70***	-0.75***	-0.70***	-0.72***
	(0.04)	(0.05)	(0.04)	(0.04)
host time		-0.01***	-0.01***	-0.01***
		(0.00)	(0.00)	(0.00)
host_response_rate		53.98***	8.09**	14.14***
		(4.48)	(3.82)	(3.76)
number_of_reviews		-0.14***	-0.03	-0.03
		(0.03)	(0.03)	(0.02)
review_scores_rating		0.85***	1.04***	0.75***
		(0.23)	(0.19)	(0.18)
cancellation_policy			8.64***	8.47***
			(1.35)	(1.33)
cleaning_fee			0.46***	0.39***
			(0.03)	(0.03)
security_deposit			0.02***	0.02***
			(0.00)	(0.00)
beds			-8.62***	-7.45***
			(1.13)	(1.11)
bedrooms			2.76	5.06***
			(1.78)	(1.75)
bathrooms			32.31***	33.72***
			(2.05)	(2.01)
accommodates			18.08***	16.85***
			(0.86)	(0.85)
R-squared	0.07	0.10	0.39	0.41
R-squared Adj.	0.07	0.10	0.39	0.41
R-squared	0.07	0.10	0.39	0.41
No. observations	8548	8548	8548	8548

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

[11] The numbers of observations for the price and revenue regression models are slightly different because we drop observations with extreme high values of price and revenue respectively for the 2 OLS regression models.

Table 7

OLS Regression Results for Revenue Prediction

	Model 5	Model 6	Model 7
const	-516.62***	-1390.12***	-2673.12***
	(44.58)	(228.38)	(216.69)
host_is_superhost	283.86***	285.16***	244.39***
	(30.39)	(32.17)	(28.77)
occupancy_rate	35.71***	37.84***	36.43***
	(0.44)	(0.56)	(0.50)
price	0.99***	0.97***	0.49***
	(0.04)	(0.04)	(0.04)
crm_rate	4.76***	4.36***	1.60
	(1.31)	(1.30)	(1.16)
blk_pct	-3.71***	-3.43***	-3.18***
	(0.65)	(0.65)	(0.58)
sphost_ds	90.67***	97.57***	57.37***
	(9.50)	(9.47)	(8.51)
availability_30	12.05***	10.80***	5.50***
	(1.41)	(1.51)	(1.36)
host time		-0.06***	-0.07***
		(0.02)	(0.02)
number_of_reviews		-2.77***	-2.00***
		(0.32)	(0.28)
review_scores_rating		9.62***	7.87***
		(2.33)	(2.08)
host_response_rate		81.41*	-120.77***
		(49.02)	(44.72)
cancellation_policy			95.79***
			(15.11)
security_deposit			0.11***
			(0.04)
cleaning_fee			1.57***
			(0.32)
beds			-101.68***
			(11.73)
bedrooms			87.70***
			(19.68)
bathrooms			93.12***
			(22.56)
accommodates			158.95***
			(9.07)
room_type			346.03***
			(29.04)
R-squared	0.50	0.51	0.61
R-squared Adj.	0.50	0.50	0.61
R-squared	0.50	0.51	0.61
No. observations	8559	8559	8559

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

Multilevel Modeling

The linear regression model ignores data nesting in our dataset. The listings are further nested within the neighborhood or community level of the city, so the unobserved community characteristics can lead to inter-correlation between outcomes for listings from the same community and hence violate the independence assumption with OLS regression. Therefore, we use multilevel modeling to estimate such community effects and develop a more accurate evaluation of the effects of the predictors we are interested in. The predicted variable and predictors are the same as the regression models above, and we only add the community variable to the models so that the community effects can be taken into consideration. Our level 1 (listings) is the lowest unit of observations, that are nested in the level-2 communities.

Our price multilevel regression model indicates that none of the superhost density, black population ratio and crime rate[12] in the community exerts a significant influence on the listing price, which confirms our previous conclusions that superhosts do not tend to set higher price and that residential segregation and crime rate do not impair the hosts' power to set prices. In addition, the revenue multilevel regression model also indicates that black population ratio and crime rate in the community do not have a significant effect on hosts' revenue. However, one percent increase in the occupancy rate can lead to an increase of $38.57 for the host per month and the effect is significant, which consolidates our previous observation that superhosts gain more mainly through

[12] In the OLS regression model, the effects of black population (-0.56) and crime rate (0.45) on price are significant but very trivial in magnitude. In this aspect, the result is consistent with multilevel modelling. However, given the evidence of spatial autocorrelation and the result of multilevel modelling, multilevel modelling should be the more accurate method to choose, which does not ignore the inter-correlations among listings in a community.

maintaining a higher occupancy rate.

Table 8[13]

Multilevel Modeling Result for Price Prediction (Model 1)

Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']						
Formula: price ~ host_is_superhost + occupancy_rate.C + sphost_ds.C + crm_rate.C + blk_pct.C + (sphost_ds.C + crm_rate.C + occupancy_rate.C + blk_pct.C | community) Data: listing						
AIC	BIC	logLik	deviance	df.resid		
127824.9	127980.3	-63890.4	127780.9	8624		
Scaled residuals:						
Min	1Q	Median	3Q	Max		
-0.8774	-0.2406	-0.1102	0.0431	25.2058		
Random effects:						
Groups	Name	Variance	Std.Dev.	Corr		
community	(Intercept)	7.975e+04	282.4026			
	sphost_ds.C	1.229e+04	110.8713	1.00		
	crm_rate.C	1.547e-01	0.3933	0.77	0.77	
	occupancy_rate.C	3.335e-01	0.5775	-0.76	-0.76	-0.19
	blk_pct.C	1.492e-02	0.1221	-0.97	-0.97	-0.69 0.82
Residual		1.335e+05	365.4081			
Number of obs: 8646, groups: community, 75						
Fixed effects:						
	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	137.81442	52.11125	372.70237	2.645	0.00852	**
host_is_superhost	16.26058	9.09048	8606.07338	.789	0.07369	.
occupancy_rate.C	-0.88144	0.16928	17.27996	-5.207	6.76e-05	***
sphost_ds.C	20.26638	21.36670	465.05975	0.949	0.34337	
crm_rate.C	0.27691	1.35073	143.16663	0.205	0.83786	
blk_pct.C	-0.02931	0.33293	311.14556	-0.088	0.92991	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						
Correlation of Fixed Effects:						
	(Intr)hst_s_occ_.C	sph_.C	crm_.C			
hst_s_sprhs	-0.101					
occpncy_r.C	-0.399	-0.197				
sphost_ds.C	0.963	-0.036	-0.391			
crm_rate.C	0.030	0.003	-0.002	-0.030		
blk_pct.C	-0.045	0.019	0.066	0.093	-0.427	

[13] For multi-level modelling, the '.C ' after a variable name means the corresponding variable has been centered before the model is run.

Table 9

Multilevel Modeling Result for Price Prediction (Model 2)

Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']						
Formula: price ~ host_is_superhost + host.time.C + host_response_rate.C + number_of_reviews.C + review_scores_rating.C + cancellation_policy.C + cleaning_fee.C + security_deposit.C + beds.C + bedrooms.C + bathrooms.C + accommodates.C + occupancy_rate.C + sphost_ds.C + crm_rate.C + blk_pct.C + (sphost_ds.C + crm_rate.C + occupancy_rate.C + blk_pct.C | community) Data: listing						
AIC	BIC	logLik	deviance	df.resid		
127162.3	127395.5	-63548.2	127096.3	8613		
Scaled residuals:						
Min	1Q	Median	3Q	Max		
-2.4999	-0.2224	-0.0640	0.0843	27.0024		
Random effects:						
Groups	Name	Variance	Std.Dev.	Corr		
community	(Intercept)	5.554e+04	235.666			
	sphost_ds.C	7.572e+04	275.1747	0.01		
	crm_rate.C	1.507e+05	388.1927	-0.02	0.09	
	occupancy_rate.C	7.804e-01	0.8834	0.46	-0.35	-0.85
	blk_pct.C	2.377e+03	48.7527	-0.08	0.12	-0.97 0.73
Residual		1.335e+05	365.4081			
Number of obs: 8646, groups: community, 75						
Fixed effects:						
	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	1.235e+02	2.160e+02	1.622e+01	0.572	0.57543	
host_is_superhost	-1.069e+00	9.197e+00	8.606e+03	-0.116	0.90748	
host.time.C	-2.930e-02	5.477e-03	8.600e+03	-5.350	9.00e-08	***
host_response_rate.C	-3.644e+01	1.351e+01	8.540e+03	-2.697	0.00701	**
number_of_reviews.C	2.392e-02	9.000e-02	8.449e+03	0.266	0.79037	
review_scores_rating.C	4.519e-01	6.668e-01	8.557e+03	0.678	0.49801	
cancellation_policy.C	1.973e+01	4.772e+00	8.641e+03	4.133	3.61e-05	***
cleaning_fee.C	7.713e-01	1.002e-01	8.401e+03	7.700	1.51e-14	***
security_deposit.C	1.667e-02	1.217e-02	8.632e+03	1.370	0.17066	
beds.C	2.589e+01	3.703e+00	8.588e+03	6.991	2.94e-12	***
bedrooms.C	4.353e+01	6.459e+00	8.608e+03	6.739	1.70e-11	***
bathrooms.C	6.164e+01	7.122e+00	8.569e+03	8.655	< 2e-16	***
accommodates.C	-1.558e+01	2.704e+00	8.641e+03	-5.761	8.67e-09	***
occupancy_rate.C	-5.246e-01	2.077e-01	2.144e+01	-2.526	0.01948	*
sphost_ds.C	5.448e+01	1.049e+02	1.147e+01	0.519	0.61349	
crm_rate.C	-7.603e+00	5.614e+01	2.222e+02	-0.135	0.89240	
blk_pct.C	2.363e+00	9.046e+00	9.520e+01	0.261	0.79444	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Table 10

Multilevel Modeling Result for Price Prediction (Model 3)

Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']						
Formula: price ~ host_is_superhost + host.time + host_response_rate + number_of_reviews + review_scores_rating + cancellation_policy + cleaning_fee + security_deposit + beds + bedrooms + bathrooms + accommodates + occupancy_rate + sphost_ds + crm_rate + blk_pct + (sphost_ds + crm_rate + occupancy_rate + blk_pct | community) Data: listing						
AIC	BIC	logLik	deviance	df.resid		
126887.0	127120.1	-63410.5	126821.0	8613		
Scaled residuals:						
Min	1Q	Median	3Q	Max		
-2.4913	-0.2224	-0.0636	0.0835	26.9889		
Random effects:						
Groups	Name	Variance	Std.Dev.	Corr		
community	(Intercept)	1.394e+05	373.3164			
	sphost_ds	1.001e+05	316.4233	-0.59		
	crm_rate	3.483e+01	5.9016	-0.97	0.38	
	occupancy_rate	5.594e-01	0.7479	-0.45	-0.46	0.65
	blk_pct	6.835e+00	2.6144	-1.00	0.53	0.99 0.51
Residual		1.335e+05	365.3861			
Number of obs: 8646, groups: community, 75						
Fixed effects:						
	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	-38.57035	92.76966	897.27685	-0.416	0.67768	
host_is_superhost	-0.87256	9.17806	8626.62847	-0.095	0.92426	
host.time	-0.02919	0.00546	8568.08705	-5.346	9.21e-08	***
host_response_rate	-36.82851	13.51426	8593.85289	-2.725	0.00644	**
number_of_reviews	0.02404	0.08998	8506.27138	0.267	0.78930	
review_scores_rating	0.44577	0.66200	8599.04204	0.673	0.50074	
cancellation_policy	19.71707	4.76061	8633.11009	4.142	3.48e-05	***
cleaning_fee	0.76989	0.10021	8561.90230	7.683	1.73e-14	***
security_deposit	0.01697	0.01217	8640.26378	1.395	0.16310	
beds	25.93323	3.70225	8575.83130	7.005	2.66e-12	***
bedrooms	43.38651	6.45561	8623.60819	6.721	1.92e-11	***
bathrooms	61.51567	7.10848	8578.64619	8.654	< 2e-16	***
accommodates	-15.52205	2.70104	8640.14858	-5.747	9.41e-09	***
occupancy_rate	-0.47916	0.20799	32.15776	-2.304	0.02785	*
sphost_ds	-4.35842	54.08290	273.83963	-0.081	0.93583	
crm_rate	-0.54999	1.90103	16.16548	-0.289	0.77602	
blk_pct	-0.50616	0.62287	79.87327	-0.813	0.41885	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Table 11

Multilevel Modeling Result for Revenue Prediction (Model 1)

Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']						
Formula: revenue ~ occupancy_rate.C + sphost_ds.C + crm_rate.C + blk_pct.C + (sphost_ds.C + crm_rate.C + occupancy_rate.C + blk_pct.C | community) Data: listing						
AIC	BIC	logLik	deviance	df.resid		
167848.6	167997.0	-83903.3	167806.6	8625		
Scaled residuals:						
Min	1Q	Median	3Q	Max		
-1.535	-0.182	-0.060	0.033	46.430		
Random effects:						
Groups	Name	Variance	Std.Dev.	Corr		
community	(Intercept)	2.059e+05	453.7462			
	sphost_ds.C	1.230e+04	110.9022	1.00		
	crm_rate.C	1.982e+01	4.4521	-1.00	-1.00	
	occupancy_rate.C	1.577e+02	12.5585	1.00	1.00	-1.00
	blk_pct.C	2.962e-01	0.5442	-0.95	-0.94	0.94 -0.96
Residual		1.562e+07	3952.4180			
Number of obs: 8646, groups: community, 75						
Fixed effects:						
	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	1482.192	95.373	34.602	15.541	<2e-16	***
occupancy_rate.C	36.245	2.472	64.999	14.663	<2e-16	***
sphost_ds.C	118.083	50.420	32.397	2.342	0.0255	*
crm_rate.C	7.211	4.745	49.837	1.520	0.1349	
blk_pct.C	-1.476	2.233	276.721	-0.661	0.5090	
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1						
Correlation of Fixed Effects:						
	(Intr)occ_.C	sph_.C	crm_.C			
occpncy_r.C	0.762					
sphost_ds.C	0.569	0.479				
crm_rate.C	-0.222	-0.166	-0.392			
blk_pct.C	0.033	0.014	0.423	-0.271		

Table 12

Multilevel Modeling Result for Revenue Prediction (Model 2)

Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']						
Formula: revenue ~ host_is_superhost + price.C + availability_30.C + room_type.C + host.time.C + host_response_rate.C + number_of_reviews.C + review_scores_rating.C + cancellation_policy.C + cleaning_fee.C + security_deposit.C + beds.C + bedrooms.C + bathrooms.C + accommodates.C + occupancy_rate.C + sphost_ds.C + crm_rate.C + blk_pct.C + (sphost_ds.C + crm_rate.C + occupancy_rate.C + blk_pct.C | community) Data: listing						
AIC	BIC	logLik	deviance	df.resid		
165177.4	165431.7	-82552.7	165105.4	8610		
Scaled residuals:						
Min	1Q	Median	3Q	Max		
-14.723	-0.176	0.001	0.141	39.513		
Random effects:						
Groups	Name	Variance	Std.Dev.	Corr		
community	(Intercept)	5.510e+04	234.7266			
	sphost_ds.C	2.614e+00	1.6166	1.00		
	crm_rate.C	7.472e-01	0.8644	1.00	1.00	
	occupancy_rate.C	1.869e+02	13.6704	1.00	0.99	1.00
	blk_pct.C	3.523e+00	1.8770	-1.00	-0.99	-1.00 -1.00
Residual		1.142e+07	3379.5630			
Number of obs: 8646, groups: community, 75						
Fixed effects:						
	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	1566.16756	62.69989	55.67188	24.979	< 2e-16	***
host_is_superhost	99.21064	84.31157	8638.30150	1.177	0.23934	
price.C	5.06307	0.09931	8581.55693	50.983	<2e-16	***
availability_30.C	-3.92334	4.00795	8558.60387	-0.979	0.32766	
room_type.C	81.46910	85.36964	8609.39500	0.954	0.33995	
host.time.C	0.06127	0.05002	8610.64868	1.225	0.22070	
host_response_rate.C	8.39407	131.10461	8583.48998	0.064	0.94895	
number_of_reviews.C	-3.80612	0.83380	8454.49501	-4.565	5.07e-06	***
review_scores_rating.C	9.54697	6.09905	8620.79269	1.565	0.11754	
cancellation_policy.C	101.09621	44.04621	8627.07834	2.295	0.02174	*
cleaning_fee.C	-0.08413	0.93179	8588.66782	-0.090	0.92806	
security_deposit.C	-0.03750	0.11229	8618.35876	-0.334	0.73840	
beds.C	-252.09086	34.17265	7986.61687	-7.377	1.78e-13	***
bedrooms.C	-45.08140	57.71090	8136.84389	-0.781	0.43473	
bathrooms.C	-233.30187	66.11381	8469.39829	-3.529	0.00042	***
accommodates.C	286.67321	26.46189	8631.19234	10.833	< 2e-16	***
occupancy_rate.C	38.56848	2.64402	93.20976	14.587	< 2e-16	***
sphost_ds.C	-53.47731	28.35538	38.85343	-1.886	0.06679	.
crm_rate.C	-0.56918	3.66886	50.98701	-0.155	0.87733	
blk_pct.C	-1.16879	1.81805	1042.22956	-0.643	0.52044	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Table 13

Multilevel Modeling Result for Revenue Prediction (Model 3)

Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']						
Formula: revenue ~ host_is_superhost + price + availability_30 + room_type + host_time + host_response_rate + number_of_reviews + review_scores_rating + cancellation_policy + cleaning_fee + security_deposit + beds + bedrooms + bathrooms + accommodates + occupancy_rate + sphost_ds + crm_rate + blk_pct + (sphost_ds + crm_rate + occupancy_rate + blk_pct | community) Data: listing						
AIC	BIC	logLik	deviance	df.resid		
165177.4	165431.8	-82552.7	165105.4	8610		
Scaled residuals:						
Min	1Q	Median	3Q	Max		
-14.723	-0.176	0.001	0.142	39.518		
Random effects:						
Groups	Name	Variance	Std.Dev.	Corr		
community	(Intercept)	1.005e+05	317.0153			
	sphost_ds	2.035e+02	14.2641	-1.00		
	crm_rate	7.575e-01	0.8704	-1.00	1.00	
	occupancy_rate	1.886e+02	13.7342	-1.00	1.00	1.00
	blk_pct	2.548e+00	1.5964	1.00	-1.00	-1.00 -1.00
Residual		1.142e+07	3379.5853			
Number of obs: 8646, groups: community, 75						
Fixed effects:						
	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	-2.168e+03	6.397e+02	7.452e+03	-3.389	0.000706	***
host_is_superhost	1.001e+02	8.431e+01	8.639e+03	1.187	0.235351	
price	5.062e+00	9.932e-02	8.581e+03	50.968	<2e-16	***
availability_30	-3.954e+00	4.008e+00	8.561e+03	-0.987	0.323839	
room_type	8.096e+01	8.536e+01	8.619e+03	0.948	0.342946	
host.time	6.153e-02	5.002e-02	8.613e+03	1.230	0.218682	
host_response_rate	7.593e+00	1.311e+02	8.587e+03	0.058	0.953813	
number_of_reviews	-3.800e+00	8.338e-01	8.459e+03	-4.557	5.26e-06	***
review_scores_rating	9.543e+00	6.099e+00	8.620e+03	1.565	0.117682	
cancellation_policy	1.011e+02	4.405e+01	8.627e+03	2.295	0.021772	*
cleaning_fee	-8.506e-02	9.317e-01	8.593e+03	-0.091	0.927265	
security_deposit	-3.730e-02	1.123e-01	8.620e+03	-0.332	0.739766	
beds	-2.520e+02	3.417e+01	8.025e+03	-7.375	1.81e-13	***
bedrooms	-4.452e+01	5.773e+01	8.043e+03	-0.771	0.440660	
bathrooms	-2.340e+02	6.613e+01	8.463e+03	-3.538	0.000405	***
accommodates	2.867e+02	2.646e+01	8.631e+03	10.834	<2e-16	***
occupancy_rate	3.861e+01	2.653e+00	9.234e+01	14.553	<2e-16	***
sphost_ds	-5.645e+01	2.919e+01	2.872e+01	-1.934	0.063009	.
crm_rate	-5.646e-01	3.690e+00	4.787e+01	-0.153	0.879043	
blk_pct	-1.095e+00	1.816e+00	1.136e+03	-0.603	0.546604	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Conclusion

Based on signaling theory, this research intends to explore the mechanisms through which Airbnb hosts benefit from the superhost accreditation. While results of similar research vary across locations, we focus on the case of the City of Chicago, where Airbnb business flourishes in the context of relatively higher crime rates and conspicuous residential segregation. At the very beginning, we conduct geospatial analysis to look for important patterns for variables of interest, including price, revenue, superhost density, crime rate and black population ratio. Then we analyze the review text data to roughly develop knowledge of what consumers hate or like most in a listing. Further, by applying the logistic regression and random forest models, we find that price is not an important feature to identify superhosts. However, depending on the tree-based revenue prediction models, we reach the conclusion that superhosts can maintain higher occupancy rates, thus earning more. In comparison with traditional parametric approaches, these machine learning methods can more precisely capture our target relationship. Finally, we use parametric OLS regression and multilevel modeling as robustness checks. The former has excellent performance in explaining the effects and the latter, removing the unobserved community effects from our analysis to yield a more accurate conclusion. Our research confirms that superhosts do not tend to ask for a higher price, but instead decrease their price a little bit and maintain a significantly higher occupancy rate, thereby increasing their revenues. Furthermore, the notoriously high crime rate and conspicuous residential segregation in the city do not impair the Airbnb business, negatively impacting neither price setting nor revenue gained.

While the research yields consistent results, it does have some limitations that need further disposal. First, as we calculate the occupancy rate based on certain already-existent variables, we may introduce minor errors to our estimation. Besides, we only focus on the period in a year when huge amounts of revenues are generated, therefore may omit some other important patterns. In the future, to attain a more reliable result, we need to develop a more precise model to estimate the occupancy rate, and use more extensive dataset that includes data of other periods of time to conduct our investigation. In addition, given the evidence of spatial autocorrelation and multilevel modelling, geospatial regression models can be built to improve this research.

References

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. Quarterly Journal of Economics, 84(3), 488–500.

Chattopadhyay, M. & Mitra, S. K. (2019). Do airbnb host listing attributes influence room pricing homogenously? *International Journal of Hospitality Management*, 81(8), 54-64.

Cheng, M. & Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*, 76(A), 58-70.

Leland, H. E. & Pyle D. H. (1977). Informational Asymmetries, Financial Structure, and Financial Intermediation. *The Journal of Finance*, 32(2), 371-387.

Lewis, G. (2011). Asymmetric Information, Adverse Selection and Online Disclosure: The Case of eBay Motors. *American Economic Review*, 101(4), 1535–1546.

Ma, X., Hancock, J. T., Kenneth Lim Mingjie, & Naaman, M. (2017). Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles. *CSCW '17: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2397-2409.

Spence, M. (1973). Job Market Signaling. *Quarterly Journal of Economics*, 87(3), 355–374.

Shatford, S. (2018). What is Airbnb's Superhost Status Really Worth? From https://www.airdna.co/blog/airbnb_superhost_status.

Teubner, T., Hawlitschek, F. & Dann, D. (2017). Price Determinants on Airbnb: How Reputation Pays Off in the Sharing Economy. *Journal of Self-Governance and Management Economics*, 5(4), 53-80.

Waldfogel, J. and Chen, L. (2006). Does information undermine brand? Information intermediary use and preference for branded web retailers. *The Journal of Industrial Economics*, 54(4), 425-449.

Wang, D. & Nicolau, J. L. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com. *International Journal of Hospitality Management*, 62, 120-131.

Yao, B., Qiu, R.T.R., Fan, D.X.F., Liu, A. & Buhalis, D. (2019). Standing out from the crowd – an exploration of signal attributes of Airbnb listings. *International Journal of Contemporary Hospitality Management*, 31(12), 4520-4542.

Appendix A

Geospatial Plots Part I

Figure A.1. Box Map for sphost_ds

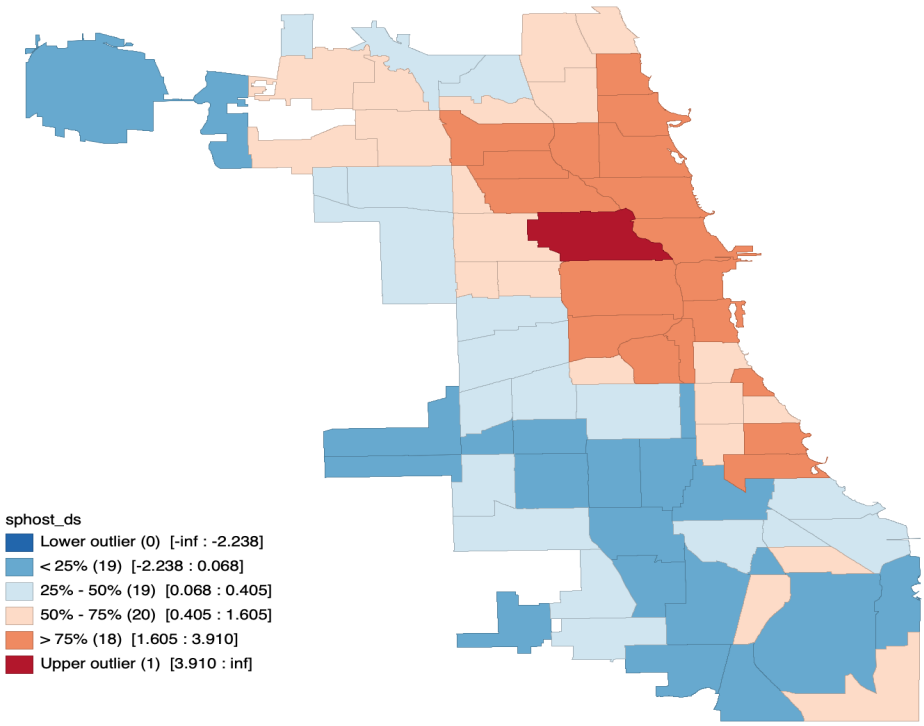


Figure A.2. Box Map for price

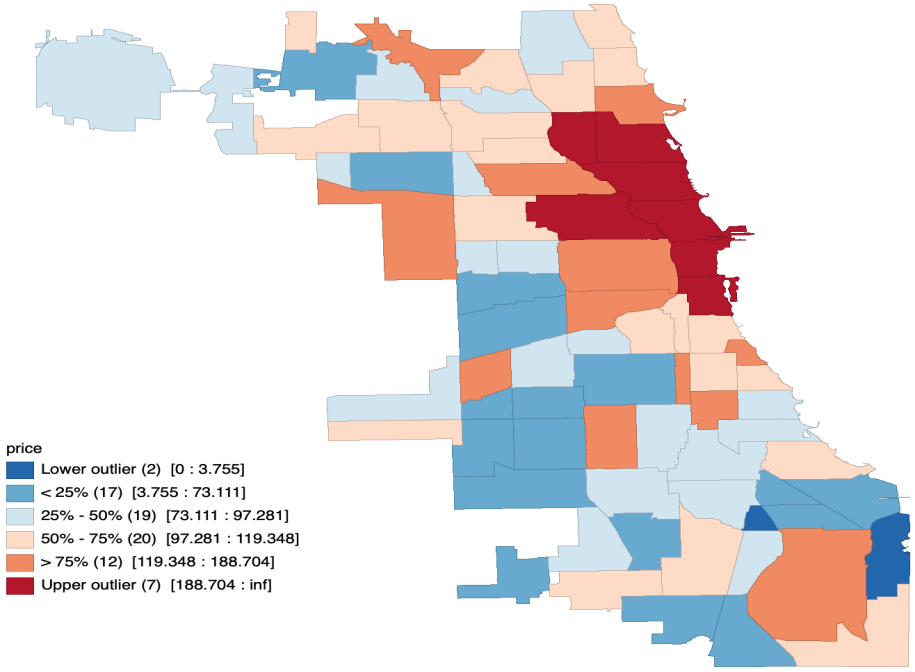


Figure A.3. Box Map for crimerate

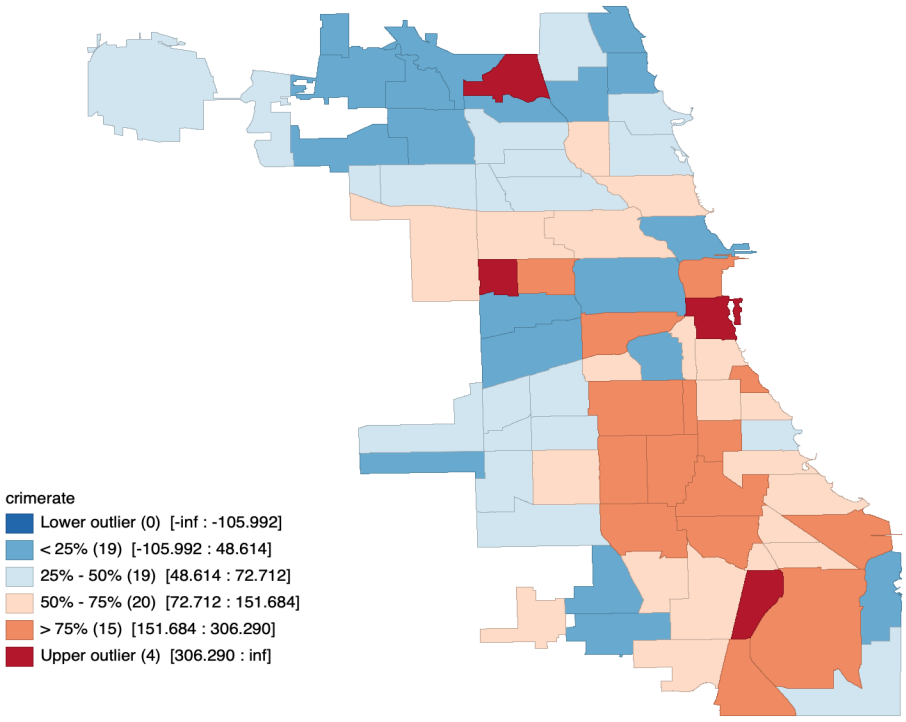


Figure A.4. Parallel Coordinate Plot

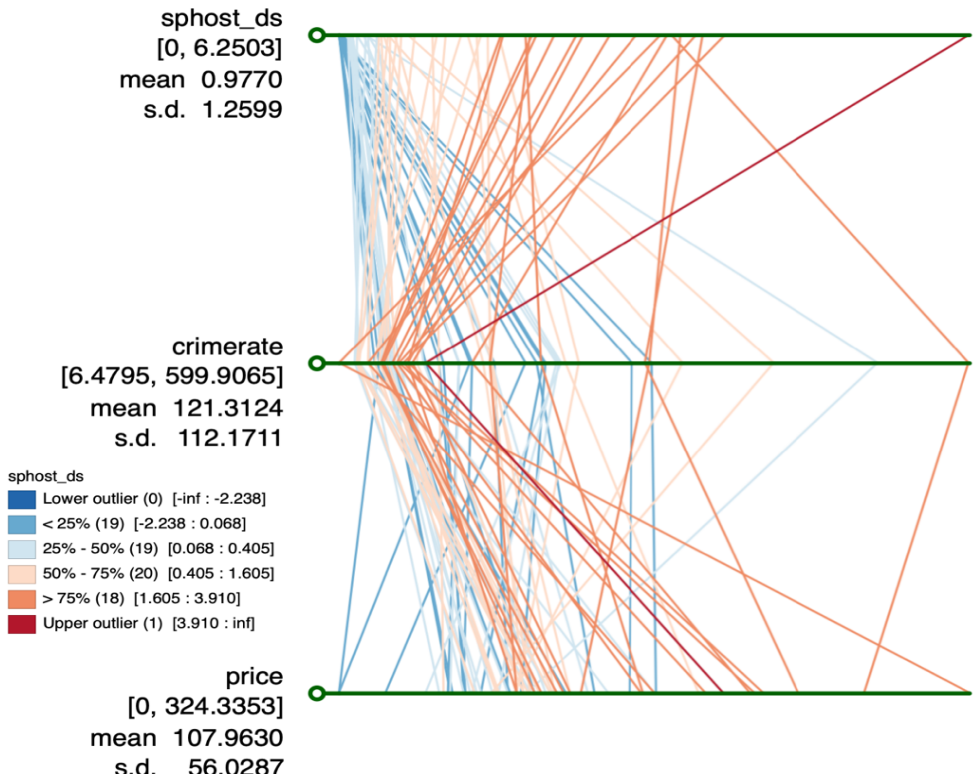


Figure A.5. Scatter Plot Matrix

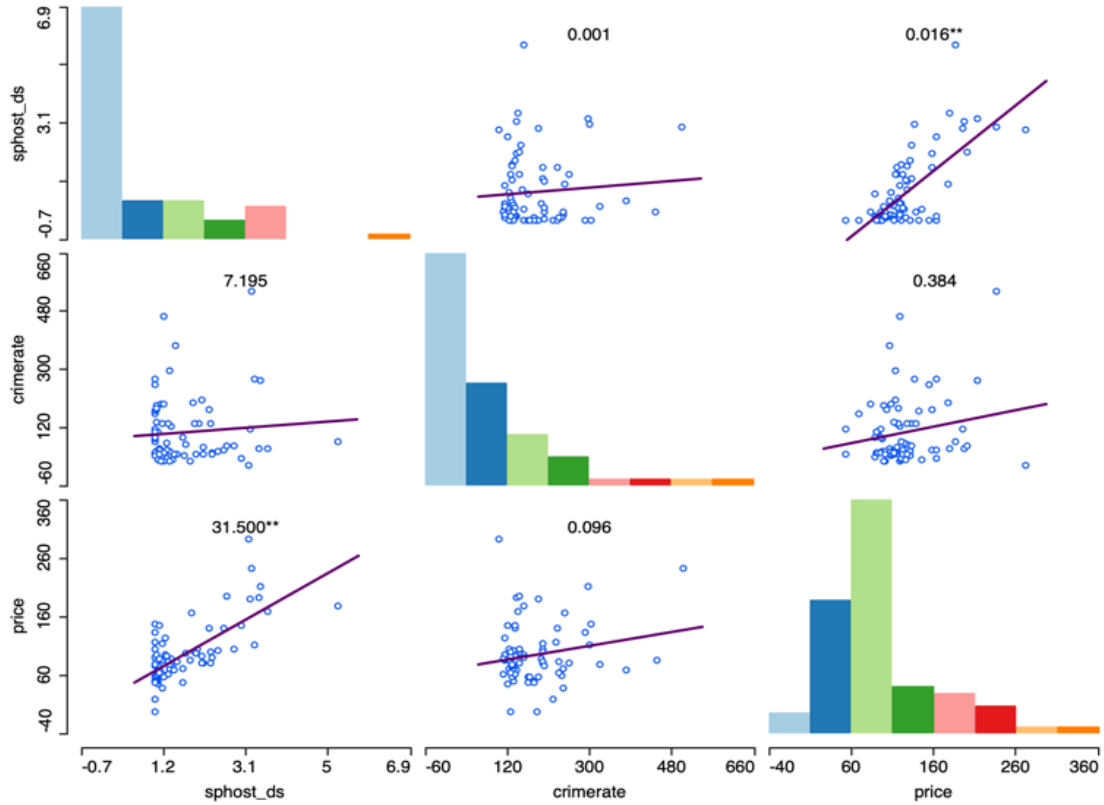


Figure A.6. Box Map for Airbnb Community Black Population Ratio

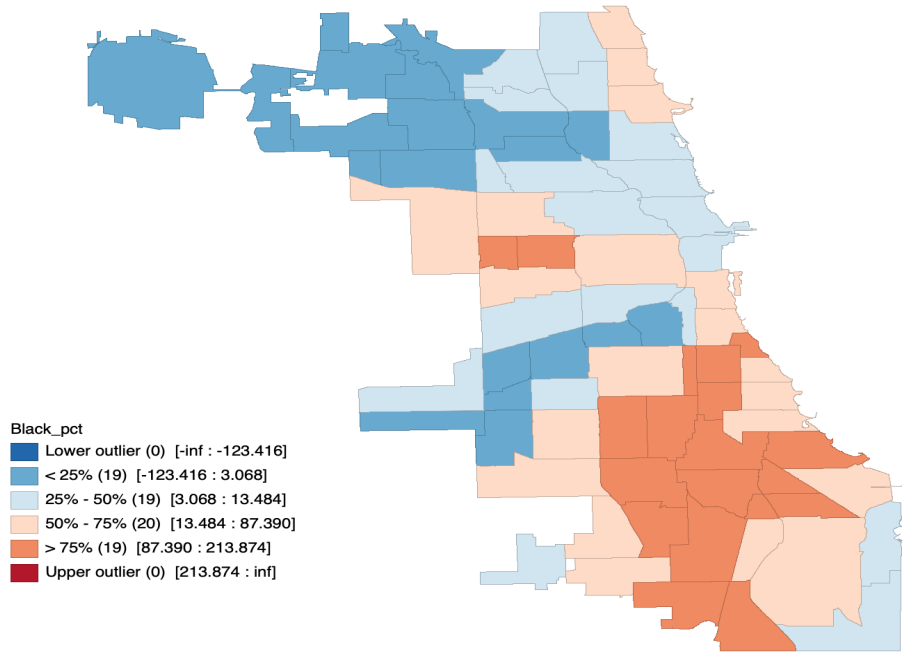


Figure A.7. Box Map for Airbnb Community Occupancy Rate

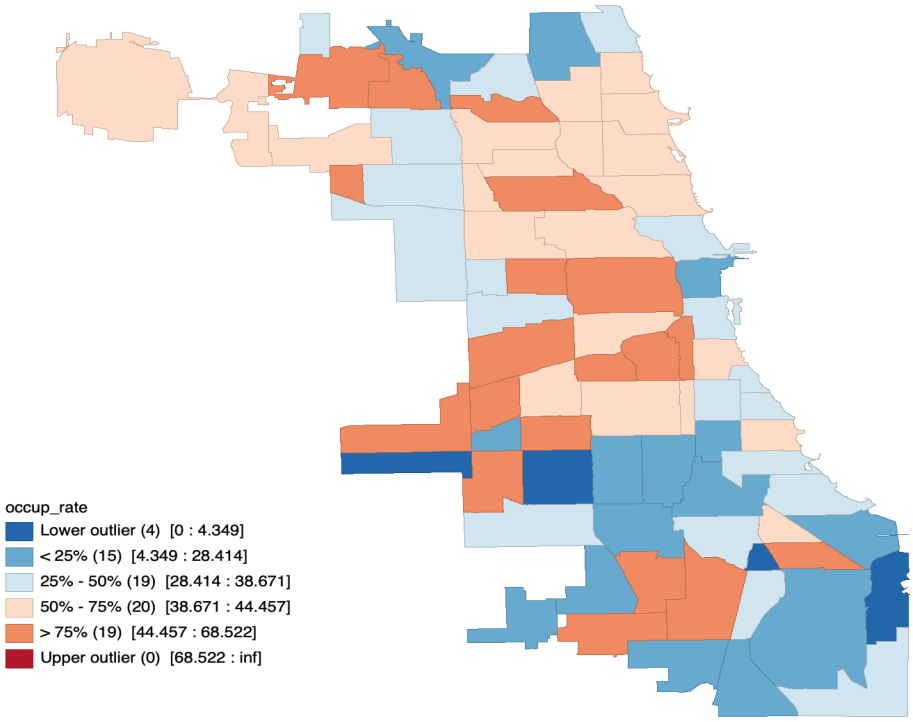


Figure A.8. All Box Maps for Airbnb with Box Plot for sphost_ds

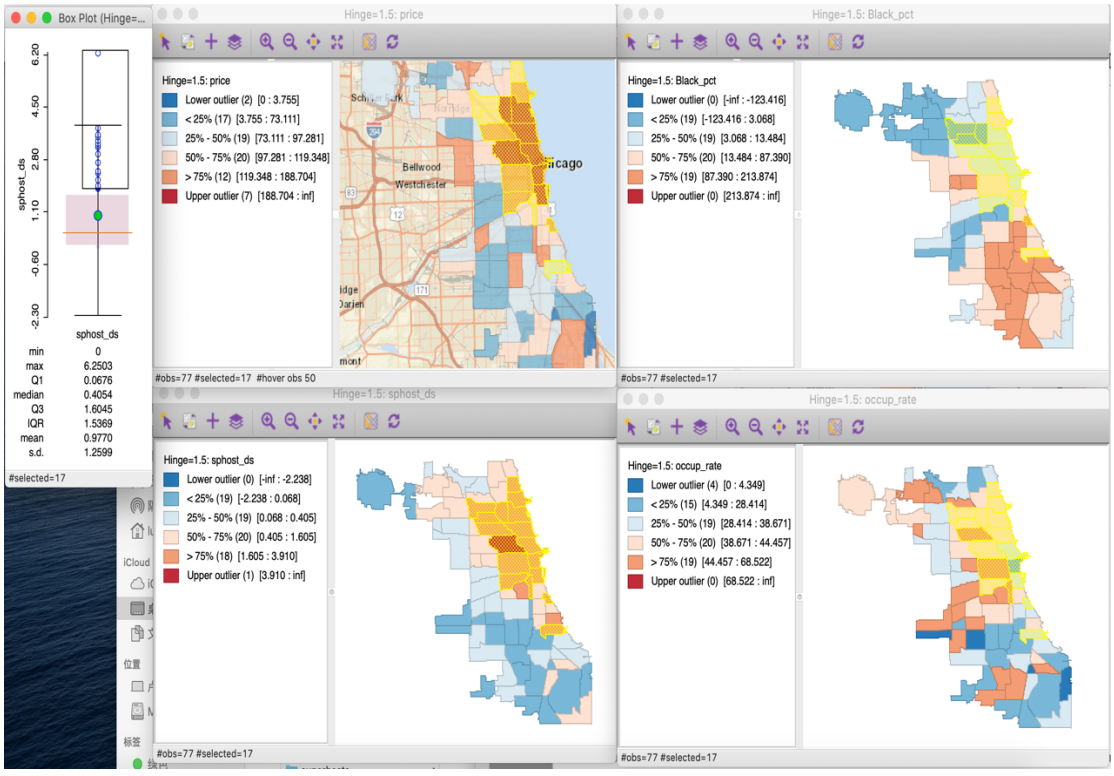
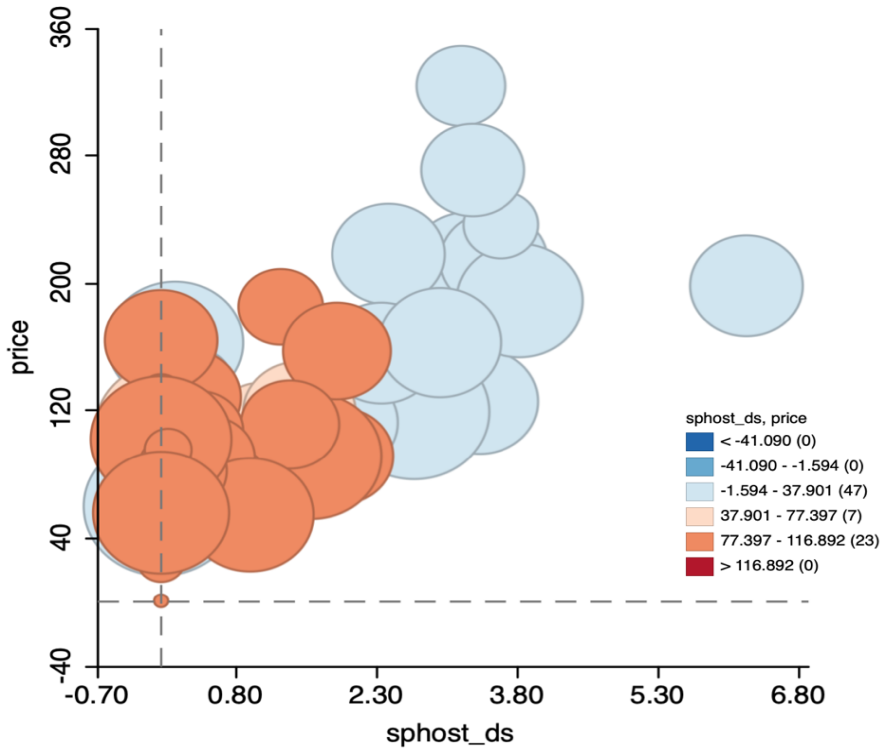


Figure A.9. Bubble Chart for price and sphost_ds



Appendix B

Geospatial Plots Part II

Figure B.1. Parallel Coordinate Plot for All 4 Variables

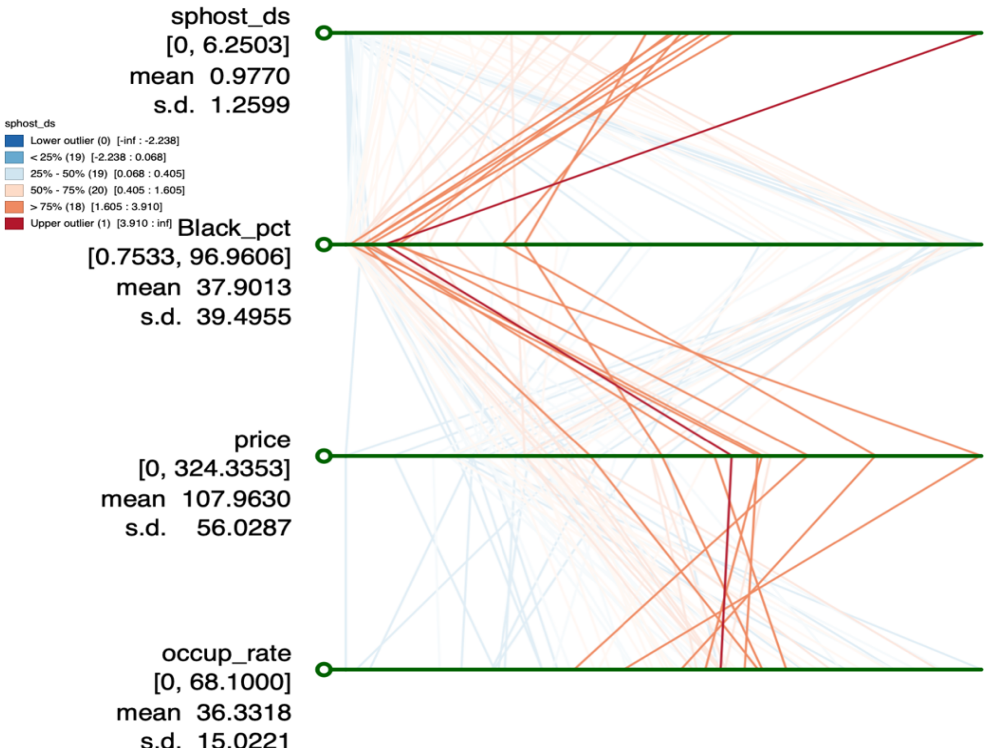


Figure B.2 Centroid Layer for Community Polygon

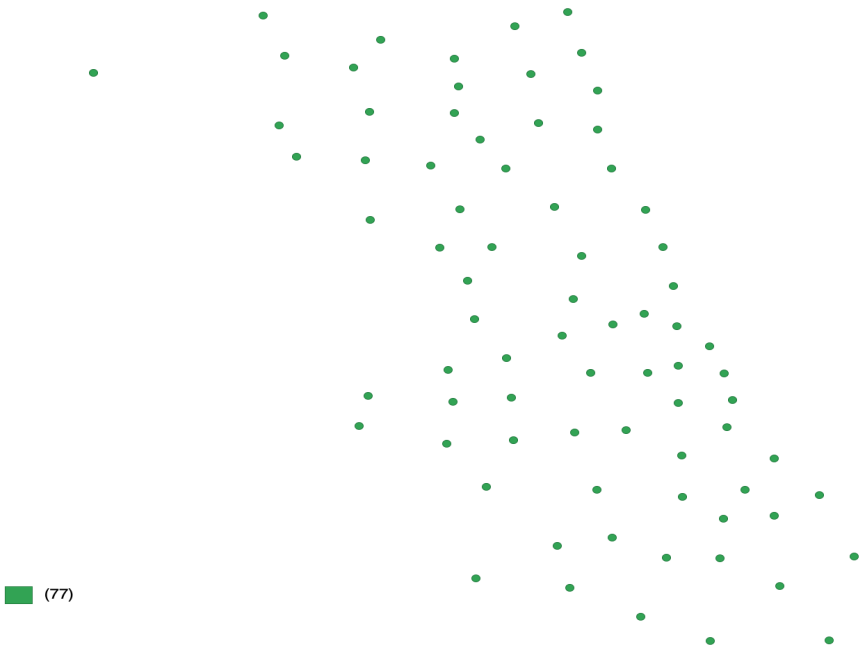


Table B.3. Spatial Weights (queen contiguity weights, great-circle-distance-band weights and KNN weights) for Polygon Layer

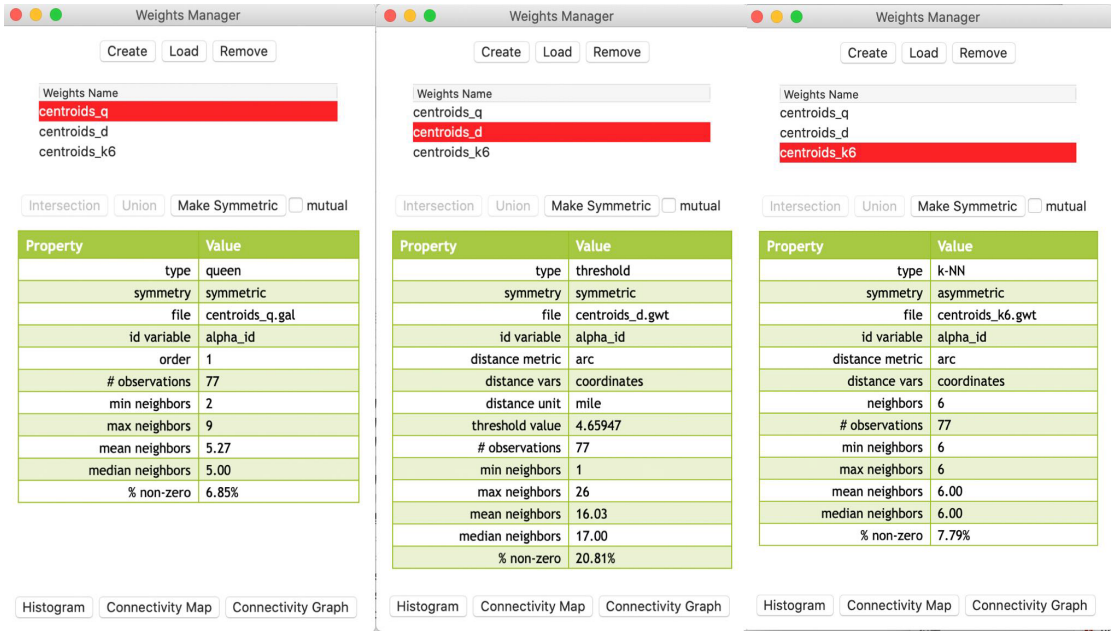


Table B.4. Spatial Weights for Centroid Layer

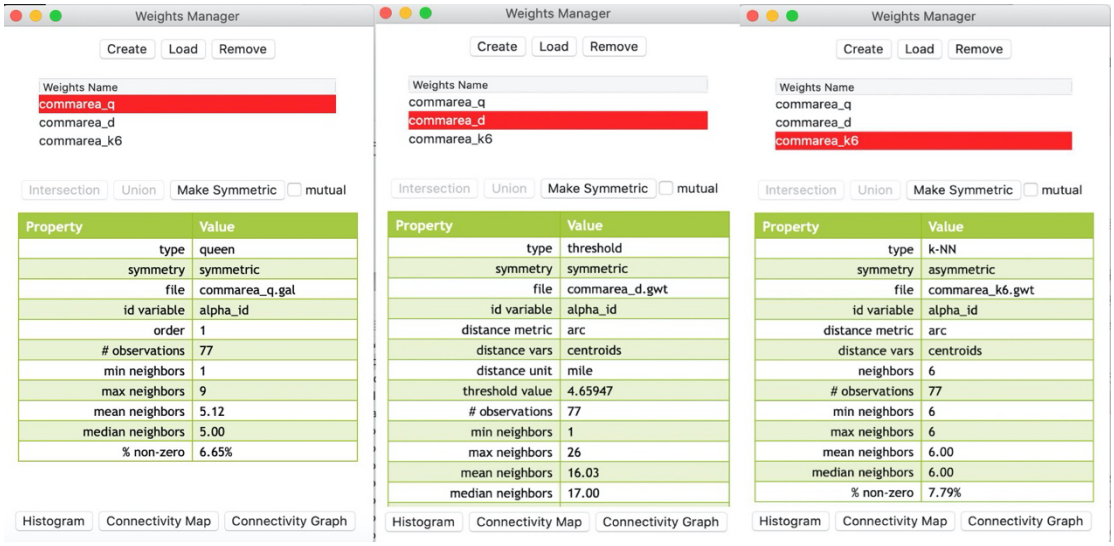


Figure B.5. Connectivity Histograms

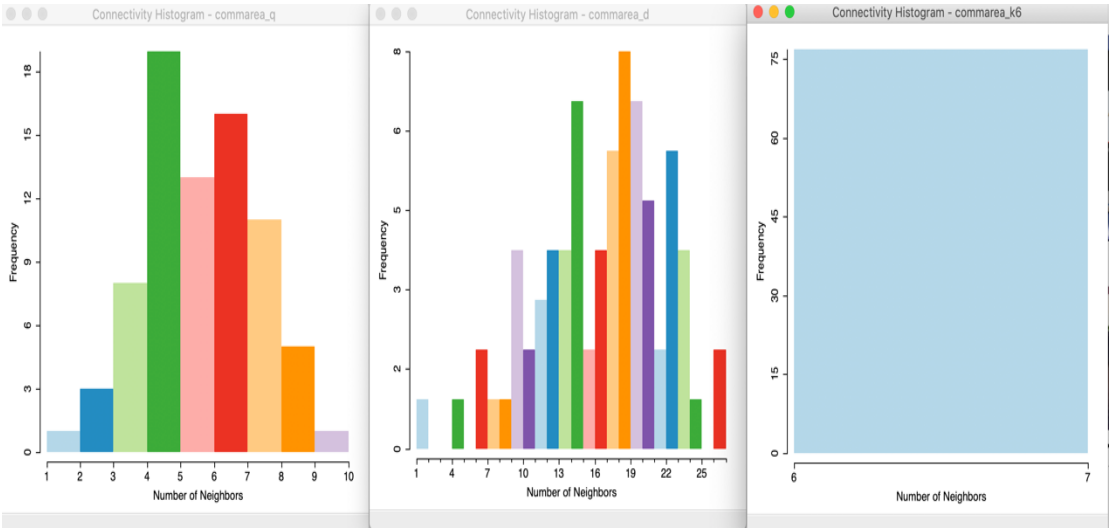


Figure B.6. Cluster Map of Local Moran's I for sphost_ds

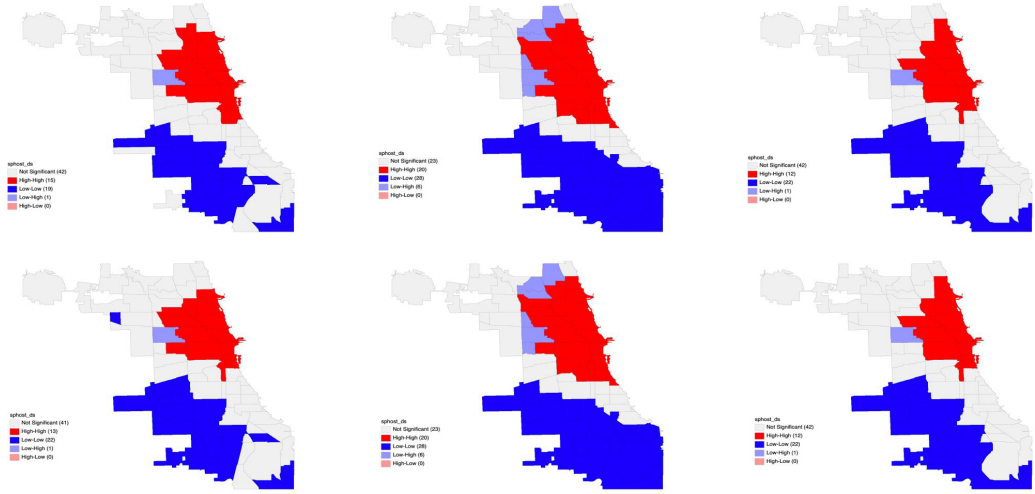


Figure B.7. Cluster Map of Local Moran's I for black_pct

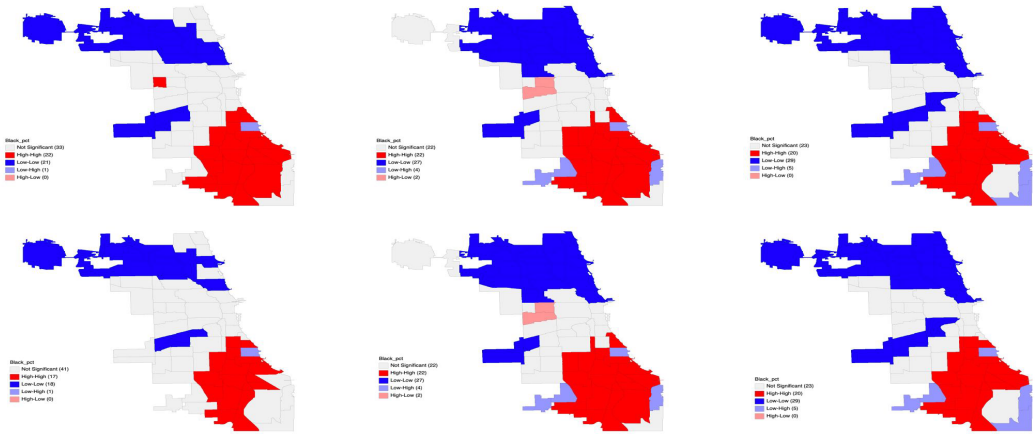


Figure B.8. Scatter Plot for Bivariate Moran's I

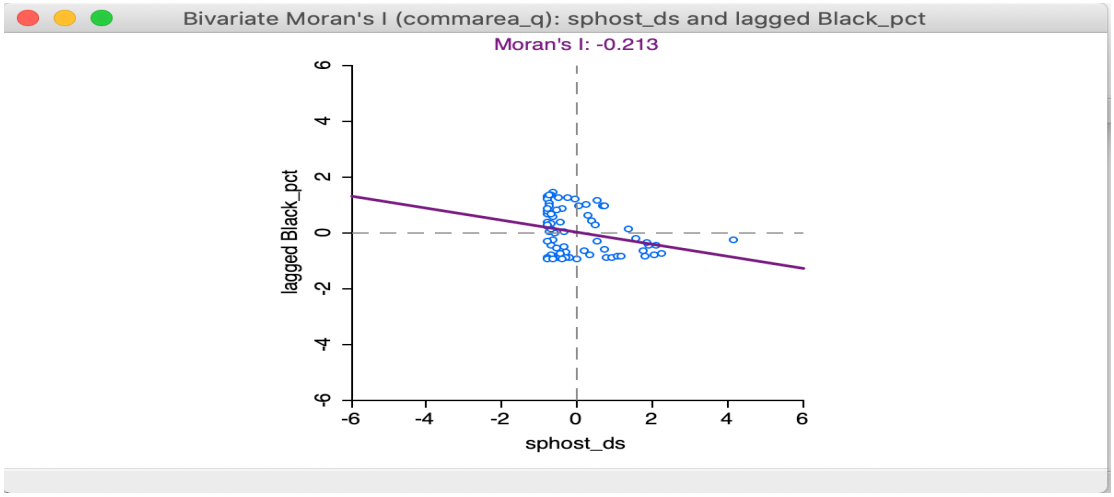
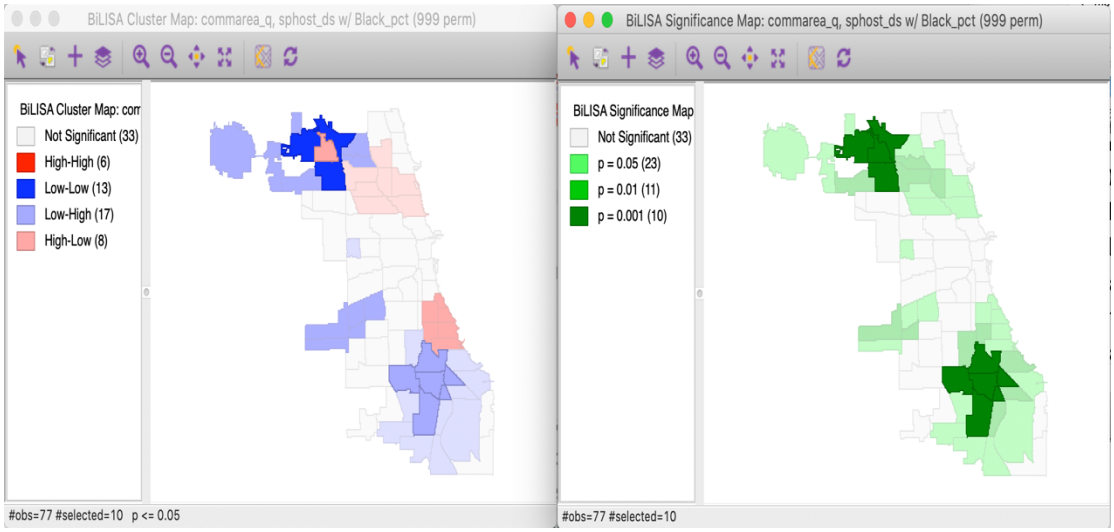


Figure B.9 Bivariate Local Moran Cluster Map and Significance Map



Appendix C

Geospatial Plots Part III

Figure C.1. Co-location Map (KNN)

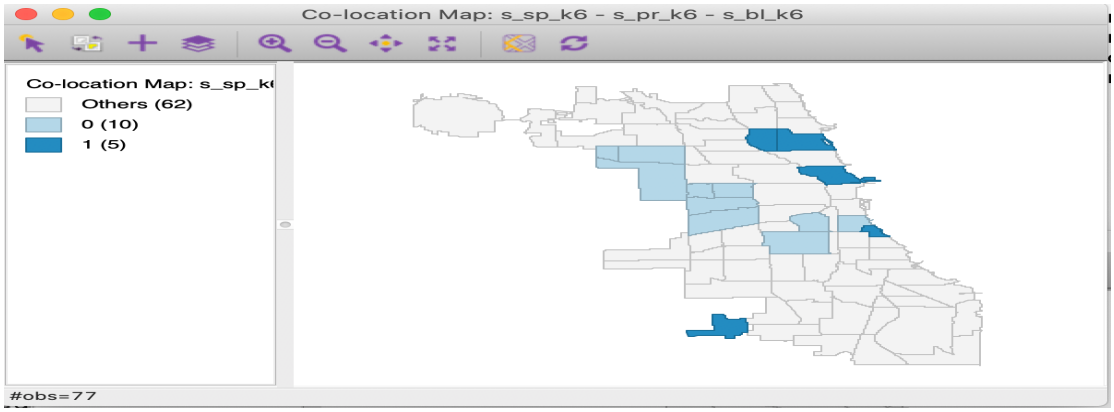
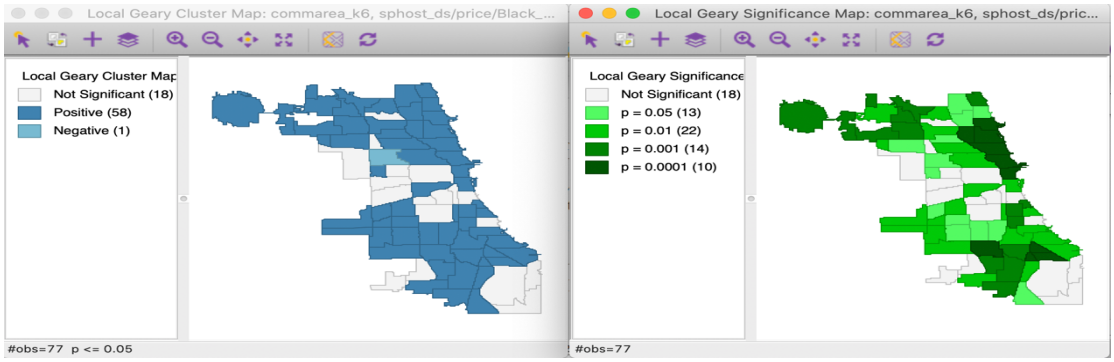
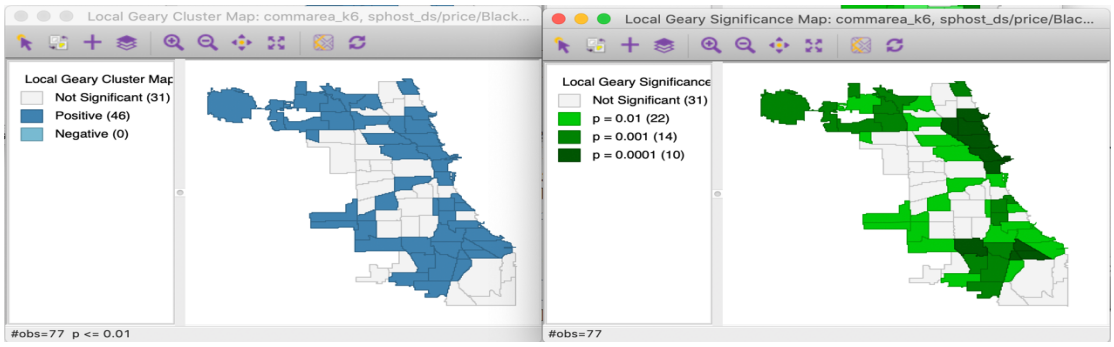
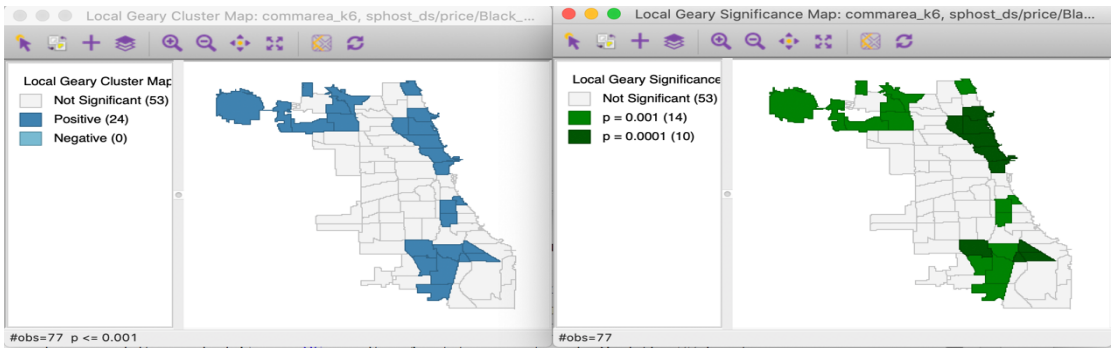


Figure C.2. Multivariate Local Geary (KNN)







Appendix D

Tables

Table D. ***Descriptive Statistics of All Variables***

	count	mean	std	min	25%	50%	75%	max
host_is_superhost	8646	0.391	0.4881	0.0	0.0	0.0	1.0	1.0
price	8646	176.640	397.146	0.0	69.0	109.0	189.0	10000
room_type	8646	2.661	0.522	1.0	2.0	3.0	3.0	3.0
reviews_per_month	8646	2.112	2.100	0.0	0.38	1.52	3.3	15.0
number_of_reviews	8646	39.114	57.724	0.0	3.0	16.0	51.0	588.0
distance_center	8646	8.149	4.358	0.341	4.932	7.362	10.907	23.816
longitude	8646	-87.664	0.041	-87.845	-87.686	-87.661	-87.635	-87.538
latitude	8646	41.902	0.056	41.652	41.876	41.904	41.941	42.022
cancellation_policy	8646	2.226	0.877	1.0	2.0	2.0	3.0	6.0
security_deposit	8646	163.584	360.652	0.0	0.0	0.0	250.0	5000.0
cleaning_fee	8646	59.091	58.968	0.0	19.0	50.0	90.0	1500.0
beds	8646	2.255	2.065	0.0	1.0	2.0	3.0	50.0
bedrooms	8646	1.615	1.170	0.0	1.0	1.0	2.0	24.0
bathrooms	8646	1.364	0.765	0.0	1.0	1.0	1.5	21.0
accommodates	8646	4.343	3.005	1.0	2.0	4.0	6.0	38.0
host_response_time	8646	0.392	0.488	0.0	0.0	0.0	1.0	1.0
host_identity_verified	8646	0.440	0.496	0.0	0.0	0.0	1.0	1.0
availability_30	8646	12.523	10.125	0.0	2.0	12.0	21.0	30.0
instant_bookable	8646	0.491	0.4999	0.0	0.0	0.0	1.0	1.0
review_scores_rating	8646	95.160	6.333	20.0	94.0	96.0	99.0	100.0

host_response_rate	8646	0.856	0.330	0.0	0.98	1.0	1.0	1.0
is_TV	8646	0.850	0.357	0.0	1.0	1.0	1.0	1.0
is_Wifi	8646	0.977	0.150	0.0	1.0	1.0	1.0	1.0
amenities_number	8646	29.775	11.125	1.0	21.0	30.0	36.0	88.0
availability_90	8646	46.695	30.936	0.0	16.0	53.0	73.0	90.0
availability_365	8646	166.542	131.197	0.0	43.0	149.0	314.0	365.0
host time	8646	1382.666	791.904	-16.0	752.0	1358.0	1980.75	4002.0
avg yearly crimes	8646	81.642	48.361	17.0	46.7	59.4	105.9	263.3
occupancy_rate	8646	38.909	34.184	0.0	7.600	30.400	66.0	100.0
revenue	8646	1780.815	4288.508	0.0	204.0	918.0	2241.0	187200.0

Table D. Variables of Interest

Variable Name	Data Type	Description	Source
alpha_id	integer	Community id	Created by the author
community	string	Community name	Chicago Data Portal: Boundaries - Community Areas (current)
sphost_ds	real	Superhost density in each community: the number of superhosts per 1000 people in the community	Inside Airbnb: detailed listing data of August 2019
black_pct	real	Black population percentage: the ratio of black popuplation over the total population in the community	CSV table: Chicago Community Area (CCA) CDS data
occup_rate	real	Occupancy rate: mean average occupancy rate of all the Airbnb hosts in the community	Inside Airbnb: detailed listing data of August 2019
price	real	average room price per night in the community	Inside Airbnb: detailed listing data of August 2019
crimerate	real	Crime rate: the number of crime incidents per 1000 people in the community from 2018 to 2019	Chicago Data Portal: Crimes – 2001 to Present