Instant Happiness or Regret?

The Impact of iBuyers in Real Estate Market Liquidity

Larasati Wulandari (1004751262)

Renato Mateus Magela Zimmermann (1004767978)

University of Toronto Department of Economics

April 14, 2021

Abstract

Real estate is one of the least liquid assets, known for taking considerable time to sell. This has led to the emergence of iBuyers and their algorithms who price the value of homes, offer home sellers an all-cash offer and resell it in the market. This raises the question of how the presence of iBuyers have impacted the US housing market, specifically its liquidity and price level. To answer this research question, this paper collected county-level data on the US socioeconomic and housing characteristics. Data were drawn from the Bureau of Economic Analysis, the US Census Bureau and US Department of Education amongst other origins. The county-level median listing price and time-on-market are analyzed using a combination of simultaneous equation systems and the difference-in-difference approach. This paper found that iBuyers did not have any significant impact on the price levels and time-on-market.

1 Introduction and Motivation

Land is and has been one of the strongest storages of value available in modern society. As is evident from the 2008 Financial Crisis, land, and more specifically real estate, has the scale to significantly impact the national economy. More recently in the 2020 COVID-19, real estate has become a safe haven to investors and individuals seeking stability amidst a volatile market. Still, one issue that has limited the real estate market for years is its lack of liquidity, which affects especially individuals whose purchase marks a significant portion of their wealth.

With the objective of tackling this issue, and backed by massive amounts of capital and data, Instant Buyers (commonly referred to as iBuyers) have grown to become prevalent investors in residential real estate. As their name suggests, IBuyers differ from traditional residential real estate investors as they reduce the time from offer to purchase to often less than a week. IBuyers leverage price data and statistical models to evaluate properties in bulk and without the need of a physical evaluator. Their hope is to buy underpriced properties quickly at a discount and hopefully resell it at competitive prices within a short amount of time.

Since the success of Opendoor, the first major iBuyer, in 2014, several other firms have joined the market for quick buying and selling of residential real estate. While most firms operate at a local level, others, such as real estate listing giant Zillow, have set up operations throughout the United States and brought with them billions of dollars in venture capital. While these companies appeal to individuals for their quick, all-cash offers, they often try advertising themselves as generators of higher market liquidity and fairer prices.

At their best, iBuyers back their claims with annecdotal evidence rather than numbers, and at the worst, IBuers could have a detrimental effect on markets. Though companies tout their qualities as home buyers, the true effect on a region's real estate market depends on their qualities as sellers. Given enough market share in a given area, iBuyers can create a local monopoly if they own a significant portion of available homes. If this is indeed an iBuyer's ultimate strategy, we should be seeing higher prices and a possible increase in the time needed for a property to be sold.

In this study, we aim to test this claim that iBuyers improve market liquidity and affect market prices in severl US counties. We will explore whether the iBuyer's claim is in fact true, and whether their presence in a county can in fact affect property prices and reduce the time needed to sell a house in their acting region.

Section 2 of the paper review past literature on the effects of specific variables on market prices and median time on market of real estate properties. Section 3 explores the datasets used to estimate the chosen models models, considering their origin, characteristics, limitations and summary statistics. Section 4 explores the empirical framework behind the utilized models and data processing decisions; this section considers the core equations used for the analysis and the assumptions behind the model choices. Section 5 states the results from estimating the models and answers the question posed in this introduction. Section 6 interprets said results and considers the reason behind the arrived-at figures.

2 Literature Review

Real estate is one of the least liquid asset, requiring a considerable amount of time listed in the market before it is sold. This has led to the emergence of real estate brokerage services that aim to reduce market frictions by connecting buyers and sellers, and providing legal expertise. The literature is unclear on the effect of brokerage firms on housing market liquidity, as measured by time on market (TOM). Sirmans, Turnbull and Benjamin (1991) [1] found that larger firms were able to sell houses faster than their smaller counterparts. Yet, Yang and Yavas (1995) [2] on two separate studies found that the size does not have any significant effect on TOM.

However, the literature suggests that TOM is endogenous, as it is affected by listing price and vice versa. Miller (1978) [3] treated TOM as a regression in explaining price. Miller finds that if time on the market is longer, the observed equilibrium price would be lower compared to the initial listing price. In contrast, Belkin et al. (1976) [4] considered TOM as a regressand. The researchers find the listing price is high, this would increase TOM. Wheaton (1990) [5] as well as Krainer and Leroy (2002) [6] use a search-theoretic model to argue that price and TOM depend simultaneously on the likelihood of a sale made. Essentially, there is a positive correlation between prices and TOM. This paper accounts for this endogeneity and uses a simultaneous system to estimate the impact of iBuyer on the price and TOM of the housing market.

In modeling the housing market, existing literature commonly treats houses as composite goods. This suggests that property price is jointly determined by consumer's evaluation of the value of the property's observable attributes and producer's offering price. Papers that models housing demand and supply in the micro-level focus on individual house characteristics, such as number of rooms, bathrooms, size, etc. However, another critical feature of houses as an economic good is the considerable cost associated with choosing another dwelling unit. Beyond just the costs of actual moving, there are search costs from transferring property possession and broker service fee. It is thus necessary to also account for quality of public services or accessibility of employment specific to the county of the house's location.

3 Data

The data used in the analysis was gathered from various origins, either directly through official sources or using web scraping techniques. The Subsection 3.1 discusses the general structure of this paper's final dataset. Section 3.2 breaks down the data sources and important features, as well as any significant flaws or limitations they might have. Finally, Subsection 3.3 outlines key statistics of the selected data.

3.1 Structure

The full dataset has a panel format, and contains 121 descriptive variables related to the economics, demographics and real estate market of 798 different US counties for the years of 2016 and 2019. Of these counties, 112 are categorized as part of the treatment group, meaning the counties have one or more active iBuyers in 2019. Of the aforementioned 121 variables, 23 are selected as our independent variables; a detailed discussion of this selection is provided in Section 1 in the Appendix.

The dataset is constrained to the years 2016 and 2019 due to limitations in the available data and unnecessary volatility in the year 2020. The "Number of iBuyers" dataset is composed of data scraped from the web in February 2021. The data is constrained as the website does not provide past historical data. For the same reason, time series observations are constrained to pre- and post-treatment.

The year 2019 was selected as our post-treatment sample date to avoid any noise from the COVID-19 pandemic. Several economic characteristics included in this paper's analysis were significantly impacted by the pandemic, which could introduce unnecessary unobserved variation in the model. This is further discussed in Section I.

The year 2016 was selected as our pre-treatment date to conform to limitations in the housing market data mentioned in IV. Note that the optimal pre-treatment date is 2013, which is one year before *Opendoor*, the first modern iBuyer, was founded. The three years in between 2013 and 2016 did not see much nationwide growth in the industry, with only one other major iBuyer - *Offerpad* - being founded in 2015. The year 2016 is still well-set as the pre-treatment year of the data, as it is one year before two large players, *RedfinNow* and *Orchard* were founded and two years before the online real estate database, *Zillow*, began operating in the iBuying market.

3.2 Composition

The full dataset is the result of merging seven datasets. This section briefly outlines the origin of said datasets, the variables selected from each of them, and their key limitations. For a discussion on the data processing and methods used to collect and merge each dataset, refer to Subsection 8.1 in the Appendix.

I Number of iBuyers

The Number of iBuyers dataset (referred to as n_ibuyers in the accompanying code) is composed of data scraped in February 15, 2021 from the website *ibuyer.com*[7]. The web scraping code accompanies this paper. The complete dataset contains the number of iBuyers in a given county at the time of collection (which we label 2019 for consistency with the remaining data).

The data collected outlines the number of ”local” and ”top” iBuyers in each US city surveyed by the website. Top iBuyers are a subset of iBuyer companies with significant capital backing who have active investments throughout the United States. This subset is composed of *Opendoor*, *RedfinNow*, *Orchard*, *Offerpad* and *Zillow Offers*, which were selected by the website authors in accordance to the previously mentioned standards. We do not include cities in which *Opendoor* or *Offerpad* were active, as these two companies were founded before 2016. Smaller iBuyers must register with the website to be included as a “Local iBuyer”, which in turn constrains our treatment group to counties with at least one IBuyer registered in the website.

We follow with the assumption that events in 2020 did not significantly affect the number of counties with active iBuyers compared to 2019. This is a fair assumption, as by 2020, the iBuying market was already active in most major counties. There is no reason to believe iBuyers significantly expanded their market share during the COVID-19 pandemic. We consider this issue further in Section 4, where we compare the effects of including the number of active iBuyers in our model, compared to a dummy variable stating iBuyers where present in a county.

One additional consideration is that the data available on the website is in the city level, while our analysis is on the county level. This further increases the need to treat the presence of iBuyers as a dummy variable, rather than an ordinal discrete data point. This need stems from how any omitted city would not translate into an omitted county, but rather introduce a downward bias in the number of iBuyers in the county.

II Education

This dataset of achievement scores in English and Mathematics for public schools is drawn from the US Department of Education (ED) Facts Data [8]. The raw data has cross-sectional format, so the final data is the result of merging data from 2016 and 2019.

Each state is obligated to report the results of assessments designed by said state, based on what they deem an appropriate test of understanding. Because of this, content on test and achievement standards are not comparable between states. There exists fixed factors in the education of each state that has to be accounted for in the empirical framework.

Students are tested annually from third to eight grade and at least once in high school. These data are aggregated by students and various subgroups. For both Mathematics and English, we retain the percentage of students in a county who score above a ”proficient” level. This percentage is often a range to account for measurement uncertainties in a given county. The data present in our final dataset is the mean of said range.

The reported data from each state is reviewed by OSEP and OESE for timeliness, completeness and accuracy. States with missing or inaccurate data are required to resubmit their data files and are reviewed before publication. The coordinated review is this dataset’s main strength.

III Characteristics

The combination of four county-level US Census datasets containing **Social**, **Demographic**, **Economic** and **Housing** characteristics make up 13 of the 23 selected descriptive variables. These were collected from the American Community Survey (ACS) taken by the U.S. Census Bureau annually [9]. Each of these datasets has panel format and are subset to only include the years 2016 and 2019. Estimates of populations, housing units and characteristics in both 2016 and 2019 use the same boundaries of areas as defined by the Census 2010 data.

These numbers are official county-level estimates of the population as inferred by the US Census Bureau, and based on the sample information of ACS. These estimates have a 90 percent margin of error, meaning that there is a 90% chance that the true value is contained by the estimate, give or take the margin of error. However, one of the ACS's limitation is in its sample size. The ACS continuously collects data monthly over 5 years which results in a total sample size of 12.5% of the population. This means analysis of specific subset of the population, such as the rural population, becomes more limited because of a small sample size and high margin of error.

IV Housing

This county-level dataset of real estate listing information is drawn from realtor.com. The realtor.com data library is in turn drawn from aggregating and analyzing data from "hundreds" database of MLS-listed for-sale homes in the industry. Realtor.com admits that some data points from smaller geographies or markets with limited or partial listing and sales coverage would be too volatile. However, because this data is originally monthly, this paper chose to aggregate the numeric column such that the final data shows the values for the month of December. Not only does this reduce the possible issue of volatility, the decision of taking the last month instead of an average is based on the objective of keeping values as close to the date the number of iBuyers was scraped as possible. Additionally, data points are continuously improved upon as realtor.com are able to build on their data breadth and accuracy.

The raw data has panel format and is subset to include only the years 2016 and 2019. This dataset has data ranging as far back as 2016 and is the reason for us selecting 2016 to be our pre-treatment year. This is an especially important dataset as it contains our regressand, the median days on market variable. It is worth mentioning that no variable in this dataset was cumulative, which does not warrant for using summation.

V GDP and Wages

The county-level GDP and personal income datasets are drawn from the Regional Economic Accounts program at the Bureau of Economic Analysis [10] [11]. In here, real GDP is in millions of chained 2012 dollars and calculations are done on "unrounded data". These measures of GDP and personal income are released both quarterly and annually and are used by the government to compare and monitor local economies consistently. The large impact that the government has as a policy maker makes it even more essential that the reported data is as accurate as possible. This is the main advantage of this dataset and this paper uses it to account for spending on constructions and employment characteristics that may encourage or discourage individuals to reside in that specific county. The raw data has panel format and is subset to include only the years 2016 and 2019.

VI Population Information

The county-level dataset of aggregate population statistics is drawn from the Demographic Analysis done by the United States Census Bureau annually [12]. The population estimate at a particular year is estimated by taking the last decennial census and then accounting for birth, death and net migration. This is used in federal funding allocations for community development and business planning. The raw data has panel format and is subset to include only the years 2016 and 2019.

VII Others

Other datasets are used for the purpose of merging the datasets discussed above. A more detailed treatment of this process is described in Section 8.1 of the Appendix. These datasets include:

- A table of information on all US cities used to merge counties and cities and area FIPS code. [**QCEW**]
- A map of US Counties and their respective LEAID codes. [8]

3.3 Summary Statistics

I General Statistics

The statistics referred to here are available in Table 3 in the Appendix. The first thing to notice in our selected dataset is the high range of standard deviations present in the selected variables. Regressors, such as the median listing price, median household income, population size estimate and construction spending have standard deviation values above five figures while variables such as homeowner vacancy rate and median rooms, which have standard deviations as low as 0.53. The same observation applies to the means of the data. This is a natural result of the various units of the various variables, and warrants normalizing the data. This normalization is further dicussed in Section 4.

We further highlight the mean of our regressands. As seen in Table 3, the mean median listing price amongst counties is 274354.04. This is to be expected in the US housing market, and follows what is expected to be seen in home sales. The high mean is accompanied by a significant variance of 162236.86 which illustrates the fairly broad range of listing prices in the market. Still, as we can see in Figure 1 this data is highly skewed to the right, following the shape of an exponential distribution.

The mean median days on market in turn has a fairly different statistical appearance compared to the median listing price. The mean median days on market is 78.55 (about two months and a half), and has standard deviation of 23.45. This is to be expected from an average marken in the United States. As seen in Figure 2, the data follows a relatively normal distribution, though it also presents a slight rightward skew.

II Treatment/Control Comparative Statistics

The statistics referred to here are aviailable in Table 4. This table compares the mean of each selected variable for control, treatment and both values. Both of the regressands, median listing price and median days on market, are fairly different. The treatment group has an average median listing price 81539.85 dollars (31%) higher than the control group. The treatment group also has an average median days on market 11.59 days below (−14.5%) that of the control group. These two characteristics are apparent from Figures 3 and 4 in the Appendix.

Other values that significantly differ when comparing the control and treatment groups are the population estimate (267214.26 vs 897316.43), rate of international migration (2.26 vs. 4.22), average household size (14475.26 vs. 48343.23) and construction spending (510128.97 vs. 1937689.42), where comarissons are for control group versus treatment group averages. These differences are consistent with a more urban profile for the treatment group compared to the control group. This makes economic sense, as larger, more urban markets are more likely to yield higher returns to real estate investors such as iBuyers. It also corresponds to areas with higher digital literacy and available data.

4 Empirical Framework

This section considers the empirical framework used to quantify the causal relationship between iBuyers and the median listing prices and TOM of the US housing market.

4.1 Preprocessing

As mentioned in Section 3.3, the disparity between the standard deviations of the covariates warrants for their normalization. For that reason, all numerical variables except for the number of top iBuyers, number of local iBuyers, median housing prices and TOM are all standardized.

4.2 Economic Model

In addressing the simultaneity between the price and TOM of the US housing market, this paper uses a system of simultaneous equations. It takes the linear assumption as given and from the system of equations, this paper derives the linear reduced form of each equation, running a linear panel data regression on median listing price and TOM against three sets of covariates:

- Average housing and public service characteristics at county-level
- Demand and supply shifters such as socioeconomic factors of the population
- Presence and number of iBuyers in the county

$$TOM_{it} = \beta_0 + \beta_1 house_{it} + \beta_2 dem_{it} + \beta_3 supply_{it} \tag{1}$$

$$+ \beta_4 Z_{it} + \beta_5 T_{it} + \beta_6 Z_{it} \cdot Tit + \beta_7 iBuyers \tag{2}$$

$$Price_{it} = \alpha_0 + \alpha_1 house_{it} + \alpha_2 dem_{it} + \alpha_3 supply_{it} \tag{3}$$

$$+ \alpha_4 Z_{it} + \alpha_5 T_{it} + \alpha_6 Z_{it} \cdot Tit + \alpha_7 iBuyers \tag{4}$$

where:

- $house_{it}$ is the set of variables that represent county-level housing and public goods characteristics at county i at time t.
- dem_{it} is set of demand shifter variables at county i at time t.
- $supply_{it}$ is set of supply shifter variables at county i at time t.
- $Z_{it} = 1$ for treatment group and 0 for control group
- $T_{it} = 1$ if time is 2019 and 0 if time is 2016

It is important to note that the observable attributes above do not fully represent the "full complexity" of the utilities gained from residing in that county. However, Kain and Quigley (1975) [13] and King (1976) [14] have shown that consumers evaluate goods according to a "reduced set of composite attributes". Thus, this is the best linear approximation of the model given the constraints.

In studying the dynamic causality between the presence of iBuyers and its corresponding effect towards the US housing market, this paper carries out a three-year difference-in-difference (DD) analysis. It treats the entering of iBuyers into specific counties as a "natural experiment", an exogenous shock to the US housing market. This divides the US counties into two groups: a treatment group which contains all the counties with at least one iBuyer company, and a control group which contains all counties without iBuyers. By using the DD method, this paper rids of time-invariant unobserved heterogeneity. It eliminates the fixed factors specific to each county that may impact the treatment and control group. DD allows for the comparison between changes in housing price level and TOM before and after iBuyer firms enter the market. For a given dependent variable like $price_{it}$ or TOM_{it}, the population DD treatment effect is the difference in the dependent variable for treated and control units before and after the intervention.

$$DID = \{E(Y_{it=1}|D_{it=1} = 1, Z_i = 1, X_i) - E(Y_{it=1}|D_{it=1} = 1, Z_i = 0, X_i)\} \tag{5}$$

$$-\{E(Y_{it=0}|D_{it=0} = 1, Z_i = 1, X_i) - E(Y_{it=1}|D_{it=0} = 0, Z_i = 0, X_i)\} \tag{6}$$

In studying the dynamic causality between the presence of iBuyers and its corresponding effect towards the US housing market, this paper carries out a three-year difference-in-difference (DD) analysis. It treats the entering of iBuyers into specific counties as a "natural experiment", an exogenous shock to the US housing market. This divides the US counties into two groups: a treatment group which contains all the counties with at least one iBuyer company, and a control group which contains all counties without iBuyers. By using the DD method, this paper rids of unobserved heterogeneity. It eliminates the fixed factors specific to each county that may impact the treatment and control group. DD enables us to compare the changes in housing price level and TOM before and after iBuyer firms enter the market.

4.3 Assumption

The above model relies on several assumptions.

I The US housing market is in equilibrium and is responsive to changes

This model assumes that county-level changes such as quality of public service, demographics or the number of iBuyers would affect the TOM and price level without any significant time lags. Essentially, changes in the market would affect the demand and supply of the housing market almost instantaneously.

II Parallel trend assumption – that both control and treatment group showed common trends in the housing market before the iBuyers entered the market

To test the validity of this assumption, we plot the housing price trend over time to check whether the two groups of counties showed similar trends. Unfortunately, there exists no county-level data on TOM and median price level before 2016 and so this paper makes the assumption that housing price index is a proxy in representing overall housing trend in the US. As shown in Figure 7, the two groups were almost identical up to 2012 but start showing diverging trends. There are two possible implications to this: first, the housing price index is not the best proxy to represent trends in median price level and liquidity. Second, the parallel trend assumption does not hold. It is necessary to note that adding data from before 2012 would not be practical, as the first iBuyer was founded in 2014. Additionally, as mentioned in Section IV, the housing dataset used for our independent variable does not go as far back as necessary to consider periods where this gap was smaller. Given this apparent breach of the parallel trend assumption, it would be necessary to utilize a synthetic control group to assure trends between the average treatment and control data are similar leading up to 2016. This would involve finding weights to compose a new synthetic control group which would in turn be used in the DD regression. Due to time and technical constraints, this will not be included in this paper, though this is an open suggestion for future work.

III The nature of its censored observation is insignificant in affecting the estimates

This paper recognizes that the model would be using censored data. A house that has not been successfully sold will not have a time on market although it will have a median listing price. This means the observed median TOM is taken from a sample of sold houses. An extension beyond this paper could choose to incorporate a selection equation in addition to the system of simultaneous equation.

IV The intervention or the entering of the iBuyer into specific county is random

This assumption is not reasonable; there exists the issue of sample selection. This paper recognizes that iBuyers choose to enter the market of specific counties based on their market potential. For example, iBuyers would want to operate in a market where they can buy homes at a cheaper price than their selling price and could do so relatively quick as to maintain their business cash flows. This means iBuyers tend to enter housing markets characterized by high liquidity and a large price discrepancy.

5 Results

This paper estimates the DD model specification in three different ways: without the covariates, with covariates including the number of iBuyers and with covariates but excluding the number of iBuyers. The results are shown in Table 6.

In the first model, TOM and log(price) are only regressed against the time dummy variable, treatment dummy variable and their interaction term. Results show that holding all else constant, an iBuyer in the market reduces the median listing price by 1.4% and TOM by 3.7 days. However, these coefficient estimates are not statistically significant. This paper argues that the model without covariates suffers from an omitted variable bias. Thus, this paper supplements the analysis by incorporating covariates into the other two models.

Looking at the estimates of the second and third models, the presence of an iBuyer increases the median listing price, holding all else constant, by 0.8 and 1%, respectively. Given that housing prices range from $100,000to900,000$, this small percentage might be significant in dollar terms. However, both models show that this effect on price is not statistically significant. Similarly, the presence of an iBuyer only reduces TOM by 3.6 to 4.3 days. This is arguably insignificant when compared against the average TOM of 78.5 days.

This paper proceeds to interpret the estimated coefficients by evaluating their sign, magnitude and statistical significance.

6 Interpretation

6.1 Dependent Variable Analysis

The first model is a single DD treatment-effect estimation that only includes five variables: the dependent variable (either TOM or log(price)), the treatment variable (Z_it), the time variable (T_it), and its interaction term. This model requires the assumption that only time-invariant unobserved heterogeneity is distorting the effect of iBuyers on TOM and house prices. The results indicate that the presence of iBuyers in the county reduces the median listing price by 1.4%. Additionally, the results suggest that, on average, the presence of iBuyers in the county reduced the TOM of house by 3.7 days.

This result aligns with expected reactions of the estate market market to the presence of iBuyers. The efforts of iBuyers in estimating the value of homes with their algorithmic pricing should reduce uncertainty and friction in the housing market. This would increase the number of houses transacted, which would in turn improve the market's liquidity and transparency. The increase in volume and scale of homes listed would decrease the market price. At the same time, this increase would allow iBuyers to drop their fees. This, the decrease in price and TOM in counties with iBuyers is economically significant.

However, the scale of this effect is small – given that the average time on market is 78.5 days (as seen in Table 3), a 3.7 day decrease in TOM is arguably insignificant. The percentage reduction in price is also small. However, as the average listing price in all counties is $274,000, this decrease is translated to be around $2,740 in absolute terms.

Still, the changes in TOM and housing price are not statistically significant in the context of this first model. As seen in Table 6, the coefficient for housing price and TOM has a p-value of 0.841 and 0.212 respectively. This shows us that comparatively, the treatment effect of listed prices is less statistically significant than its effect on TOM.

However, as mentioned in Section 4, there exists the issue of sample selection combined with the diverging trend in the housing price index of control groups. Essentially, iBuyers select market with higher growth trends in the housing prices. This unaccounted selection of cities introduce an upward biased in price differences. In other words, because iBuyers are more likely to select cities with the highest potential of growth in listing price, resulting estimates underestimate the effects of iBuyers on prices. Little can be said about the time on market due unavailable data on the historical trend of median time on market prior to 2016. However, under the assumption that the housing price index is a good proxy for trends in TOM, estimates for the effect of iBuyers toward TOM is also underestimated.

6.2 Independent Variable Analysis

This subsection accounts for changes in covariates that affect the changes in TOM and housing prices. As shown in Table 6, most coefficient of the covariates are statistically significant at the 10 % confidence level.

Most coefficients exhibit signs which are consistent with theory. For example, an increase in homeowner's vacancy rate decreases the median listing price. When the market is characterized by vacant homes, prospective buyers have more bargaining power as this almost signifies an oversupply. This is further reinforced by the positive correlation between homeowner's vacancy rate towards TOM. An increase in homeowner's vacancy rate by one standard deviation increases the TOM by 3.84 days. Another example is average household size. An increase in the number of people in a household by 1 standard deviation increase the price level by 34.4%. This aligns with theory, as large families or households require more spacious (and thus expensive) houses.

However, some estimates have the opposite or even no relation to what we expect. For example, this model incorporated the median household income and population number as demand shifters that may affect housing prices or liquidity. However, the model estimates these variables to be 0. This means holding all else constant, an increase in median household income and population number affect neither housing price nor TOM. After further analysis, this paper found that there is little significant growth in population from 2016 to 2019. This is not the case with median household income, which increased from and average of $\tilde{\$}$58,000 to $\tilde{\$}$65,000 per capita. There are two scenarios that explain this result: first, median household income do not affect housing price level at all, which is highly unlikely. Second, household and per-capita incomes might still affect price level, but just on a very small economic level. It is important to note that this estimate is supported by a high t-statistics, and thus high statistical significance.

Another specification showing the opposite sign to what we expect is the median rooms. From the DD regression results, an increase in one standard deviation in median rooms reduces price by 36.4%, holding all else constant. Higher prices in response to more rooms per listed house may make sense when considering individual home purchases. However, as this dataset is is the county-level, the average rooms per property might just incorporate unobserved effects of rural-urban separation. Cities are known for their high housing prices, yet the average rooms tend to be smaller than that of their rural counterparts. Additionally, when a person chooses to reside in a specific county, they are limited by their housing options in that location, which usually share similar characteristics. Ultimately, adding more urban/rural indicators could prevent this unexpected result from happening in future work.

It is important to point out that the number of both top or local iBuyers in the county had little statistical significance towards both affecting housing prices or TOM. With regards to housing prices, holding all else constant, an increase in the number of iBuyers in the county on average increases the median listing prices by 6.2%. This is contradictory to the argument that iBuyers reduce friction in the market which in turn, improves liquidity and increase the volume of transacted house. However, there are two points to be made. First, this assumes that changes in housing liquidity and volume are extremely responsive to the entrance of iBuyers into the market. However, this assumption does not seem to conform to theory. The assumption disregards the fact that iBuyers aim to make a 5-7% margin on each house they buy. If there exists many top iBuyers within a market, these large firms could have market power to significantly alter median listing prices. A similar argument can be made for the effect of the number of iBuyers on TOM. However, it is important to note that the number of iBuyers as covariates for TOM and price are not statistically significant, as they have high p-values. Still, future work could be done to test this hypothesis by re-estimating this model using new data on TOM and prices and see whether the entrance of iBuyers into a market would significantly impact the market's housing dynamics.

This raises the question of how the model would be affected when excluding the number of iBuyers, which is the third DD model specification of this paper. As observed in Table 6, there is almost no change in the coefficient estimates on all covariates. This reinforces the initial hypothesis stated in from the first DD model that iBuyers are not observed to have a significant impact on both housing prices and TOM.

7 Conclusion

This paper aims to quantify the causal impact of iBuyers towards the US housing market price level and liquidity. By using a system of simultaneous equations and deriving their linear reduced form, this paper accounts for the endogeneity between TOM and price. These two dependent variables are then regressed against three sets of covariates that covers socioeconomic and housing characteristics, as well as demand and supply shifters. Additionally, this paper also incorporates the difference-in-difference approach to eliminate time-invariant unobserved heterogeneity that exists in the county level.

All three difference-in-difference specifications result show that iBuyers do not have any significant impact towards the median listing price or TOM of the housing market. However, there exists limitations within this paper that originates from the lack of available data. One such limitation is the breach of the parallel trend assumption, originating from the lack of observations on key years. Another limitation is the sample selection bias that comes from iBuyers making decisions on which market they should enter. One possible solution is by adding a selection equation in addition to the simultaneous equation of housing prices and TOM.

Future studied on this topic should consider gathering longer-term trends in iBuyer dynamics and perhaps focus on conducting studies on the individual property level. Incorporating older data, optimally beginning in 2013, would also improve results. Additionally, novel techniques such as using synthetic controls groups could mitigate the limitations of the current DD model.

Ultimately, it is unclear whether this new trend in the real estate market is beneficial to the affected markets or just backed by shallow promises. Regardless, this trend will continue, and if this phenomenon is not better understood, the instant gratification could turn into regret in no time.

References

[1] C.F Sirmans, Geoffrey Turnbull, and John Benjamin. “The Markets for Housing and Real Estate Broker Services”. In: *Journal of Housing Economics* 1.3 (1991), pp. 207–217.

[2] Abdullah Yavas and Shiawee Yang. “The Strategic Role of Listing Price in Marketing Real Estate: Theory and Evidence”. In: *Real Estate Economics* 23.3 (1995), pp. 347–368.

[3] Norman G. Miller. “Time on the Market and Selling Price”. In: *Real Estate Economics* 6.2 (1978), pp. 164–174.

[4] Jacob Belkin, Donald J. Hempel, and Dennis McLeavey. “An Empirical Study of Time on Market using Multidimensional Segmentation of Housing Markets”. In: *Real Estate Economics* 4.2 (1976), pp. 57–75.

[5] William C. Wheaton. “Vacancy, Search and Prices in a Housing Market Matching Model”. In: *Journal of Political Economy* 98.6 (1990), pp. 1270–1292.

[6] John Krainer and Stephen Leroy. “Equilibrium Valuation of Illiquid Assets”. In: *Economic Theory* 19.2 (2002), pp. 223–242.

[7] *iBuyers Market.* URL: `\url{https://ibuyer.com/ibuyer-markets.html}` (visited on 04/13/2021).

[8] *EDFacts Data Files.* URL: `\url{https://www2.ed.gov/about/inits/ed/edfacts/data-files/index.html}` (visited on 04/13/2021).

[9] *American Community Survey.* URL: `\url{https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/}` (visited on 04/13/2021).

[10] *Regional Economic Accounts: Gross Domestic Product (GDP).* URL: `\url{https://apps.bea.gov/regional/downloadzip.cfm}` (visited on 04/13/2021).

[11] *Regional Economic Accounts: Personal Income (State and Local).* URL: `\url{https://apps.bea.gov/regional/downloadzip.cfm}` (visited on 04/13/2021).

[12] *Demographic Analysis (DA).* URL: `\url{https://www.census.gov/programs-surveys/decennial-census/about/coverage-measurement/da.html}` (visited on 04/13/2021).

[13] John F. Kain and John M. Quigley. “Housing Markets and Racial Discrimination: A Microeconomic Analysis”. In: *National Bureau of Economic Research* (1975), pp. 1–393.

[14] A.T. King. “The Demand for Housing: Integrating the Roles of Journey-to-Work, Neighborhood Quality, and Price”. In: *Household Production and Consumption* (1976), pp. 451–488.

8 Appendix

8.1 Data Processing and Methods

All data processing was performed using the Python programming language, and more specifically the Pandas library, the notebooks used are all available with this paper. Most processing can be separated into categories:

- **Cleaning and Standardizing:** This involves standardizing column names to enable propper merging and cleaning data points to ensure type consistency. Similar procedures where applied to all merged datasets to ensure all received the same treatment. Additionally, cleaning county names took a significant ammount of time, as naming standards diverged between datasets. This was partially autmated by subsituting regular expressions, but also involved partial cleaning by hand. Data aggregation was necessary for type subsets (cities to counties and months to years) in order to maintain the structural consistency of the data. Aggregation methods where performed on a per-variable basis.

- **Merging:** Once the data is cleaned, all different datasets where merged to form a single dataset. Given most of the work was done in the cleaning stage, this only required a small custom merging operation to left-merge all datasets onto the n_ibuyers dataset on county, state and year. This is saved under the name "merged.csv."

- **Manipulation:** The first manipulation performed on the data is its standardizing. This is saved under "merged_std.csv." Other manipulations mostly involve data aggregation to display summary statistics.

- **Visualization:** Some plots are created using the matplotlib backend of the Pandas library.

8.2 Charts

Table 1: Variable Descriptions

regressor_name	**dataset**
median_rooms	median number of rooms per listed property
educ_nohs	percentage of population with no high school diploma
white	percentage of white population
mean_cash_public_assistance_income_dollars	mean per-capita income from public assistance programs (dollars)
per_capita_income_dollars	mean per-capita income from all sources (dollars)
median_household_income_dollars	mean household income from all sources (dollars)
part of educ_further	percentage of population with degrees beyond high school
mean_retirement_income_dollars	mean per-capita retirement income
median_age_years	median age (years)
mean_travel_time_to_work_minutes	mean travel time from home to work
homeowner_vacancy_rate	homeowner vacancy
move_post2010	percentage of population who moved to current property after 2010
average_household_size	average number of people per household
english_prof_pct	percent of students proficient in English
math_prof_pct	percent of students proficient in English
construction	total expenditure in construction
median_listing_price	median property listing price
median_square_feet	median property size (square feet)
median_days_on_market	median days property is listed on market
nlocal_ibs	number of local iBuyers
ntop_ibs	number of top iBuyers
rnaturalinc	natural rate of population growth
rnetmig	rate of net migration
popestimate	population size estimate
annual_avg_emplvl	annual average employment level

Table 2: Dataset of Origin of Selected Regressors

regressor_name	dataset
median_rooms	characteristics
educ_nohs	characteristics
white	characteristics
mean_cash_public_assistance_income_dollars	characteristics
per_capita_income_dollars	characteristics
median_household_income_dollars	characteristics
part of educ_further	characteristics
mean_retirement_income_dollars	characteristics
median_age_years	characteristics
mean_travel_time_to_work_minutes	characteristics
homeowner_vacancy_rate	characteristics
move_post2010	characteristics
average_household_size	characteristics
english_prof_pct	educ
math_prof_pct	educ
construction	gdp
median_listing_price	h_vals
median_square_feet	h_vals
median_days_on_market	h_vals
nlocal_ibs	n_ibuyers
ntop_ibs	n_ibuyers
rnaturalinc	pop_info
rnetmig	pop_info
popestimate	pop_info
annual_avg_emplvl	wages

Table 3: General Summary Statistics

	mean	std
treatment	0.141414	3.485583e-01
median_listing_price	274354.042614	1.622369e+05
median_days_on_market	78.553030	2.345368e+01
median_rooms	5.693813	5.329772e-01
homeowner_vacancy_rate	1.639657	1.087029e+00
mean_travel_time_to_work_minutes	25.214646	5.247091e+00
math_prof_pct	46.819991	1.281202e+01
english_prof_pct	51.249663	1.155455e+01
median_square_feet	1827.165404	5.912998e+02
median_age_years	38.617677	4.654999e+00
move_post2010	43.302652	7.698905e+00
annual_avg_emplvl	154349.486427	2.943917e+05
median_household_income_dollars	61671.342487	1.647529e+04
per_capita_income_dollars	31669.158775	7.945113e+03
popestimate	356319.613952	7.038671e+05
rnaturalinc	2.533129	3.524344e+00
rnetmig	4.029436	1.067996e+01
educ_nohs	23.133649	1.455964e+01
educ_further	78.632797	4.086123e+01
average_household_size	2.672629	2.430024e-01
white	42.600758	4.056286e+01
mean_cash_public_assistance_income_dollars	2772.248037	1.138819e+03
construction	712006.203283	1.410527e+06
mean_retirement_income_dollars	26170.630997	5.861282e+03

Table 4: Treatment/Control Summary Statistics

treatment	0	1	All
median_listing_price	262823.15	344363.01	274354.04
median_days_on_market	80.19	68.60	78.55
median_rooms	5.73	5.45	5.69
homeowner_vacancy_rate	1.64	1.61	1.64
mean_travel_time_to_work_minutes	25.22	25.20	25.21
math_prof_pct	47.03	45.53	46.82
english_prof_pct	51.57	49.29	51.25
median_square_feet	1812.60	1915.57	1827.17
median_age_years	38.92	36.75	38.62
move_post2010	42.60	47.55	43.30
annual_avg_emplvl	110745.13	419090.23	154349.49
median_household_income_dollars	61437.53	63090.91	61671.34
per_capita_income_dollars	31415.57	33208.79	31669.16
popestimate	267214.26	897316.43	356319.61
rnaturalinc	2.26	4.22	2.53
rnetmig	4.05	3.88	4.03
educ_nohs	23.39	21.58	23.13
educ_further	78.07	82.06	78.63
average_household_size	2.66	2.76	2.67
white	43.19	39.04	42.60
mean_cash_public_assistance_income_dollars	2720.58	3049.60	2772.25
construction	510128.97	1937689.42	712006.20
mean_retirement_income_dollars	25906.35	27775.20	26170.63

Table 5: 2016/2019 Summary Statistics

treatment	0	1	All
median_listing_price	262823.15	344363.01	274354.04
median_days_on_market	80.19	68.60	78.55
median_rooms	5.73	5.45	5.69
homeowner_vacancy_rate	1.64	1.61	1.64
mean_travel_time_to_work_minutes	25.22	25.20	25.21
math_prof_pct	47.03	45.53	46.82
english_prof_pct	51.57	49.29	51.25
median_square_feet	1812.60	1915.57	1827.17
median_age_years	38.92	36.75	38.62
move_post2010	42.60	47.55	43.30
annual_avg_emplvl	110745.13	419090.23	154349.49
median_household_income_dollars	61437.53	63090.91	61671.34
per_capita_income_dollars	31415.57	33208.79	31669.16
popestimate	267214.26	897316.43	356319.61
rnaturalinc	2.26	4.22	2.53
rnetmig	4.05	3.88	4.03
educ_nohs	23.39	21.58	23.13
educ_further	78.07	82.06	78.63
average_household_size	2.66	2.76	2.67
white	43.19	39.04	42.60
mean_cash_public_assistance_income_dollars	2720.58	3049.60	2772.25
construction	510128.97	1937689.42	712006.20
mean_retirement_income_dollars	25906.35	27775.20	26170.63

8.3 Figures

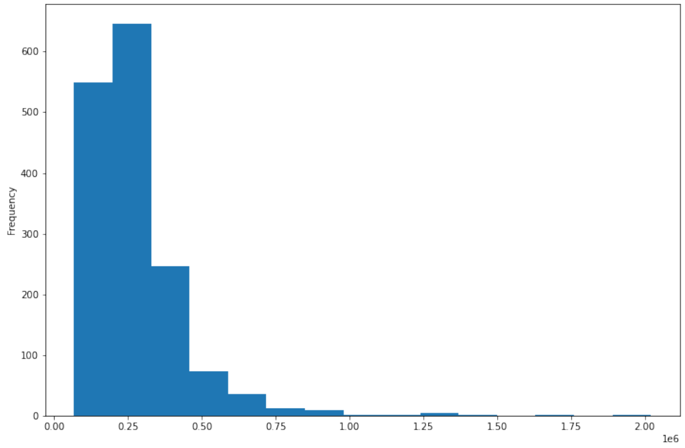

Figure 1: Histogram for Median Listing Price
(Aggregate of 2016 and 2019)

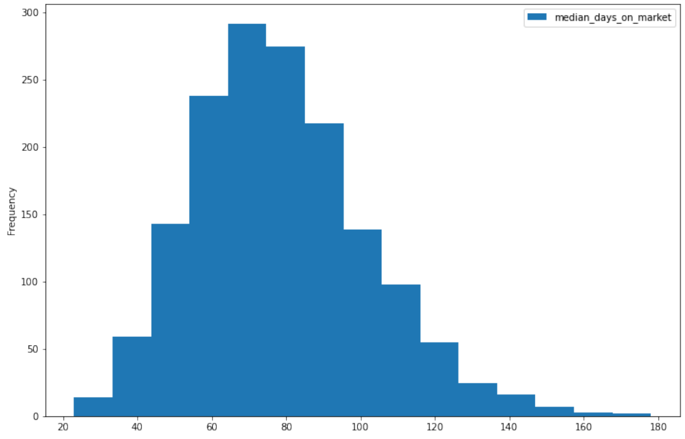

Figure 2: Histogram for Median Days on Market
(Aggregate of 2016 and 2019)

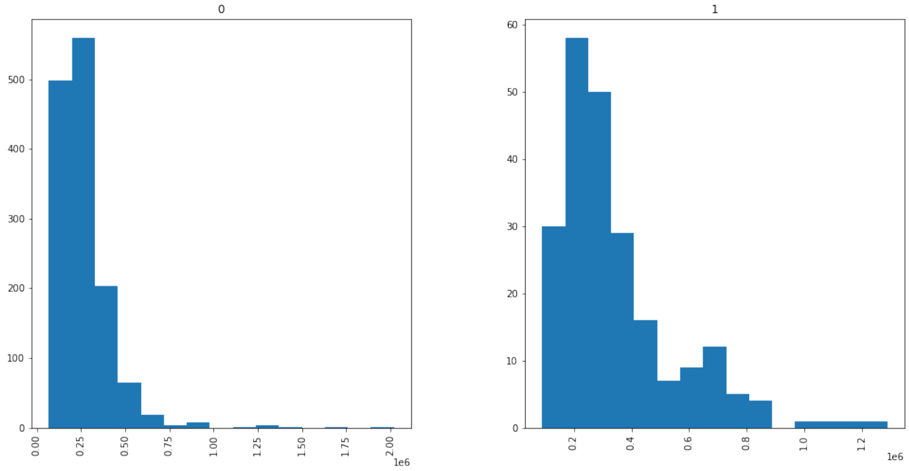

Figure 3: Control/Treatment Comparisson for Median Listing Price

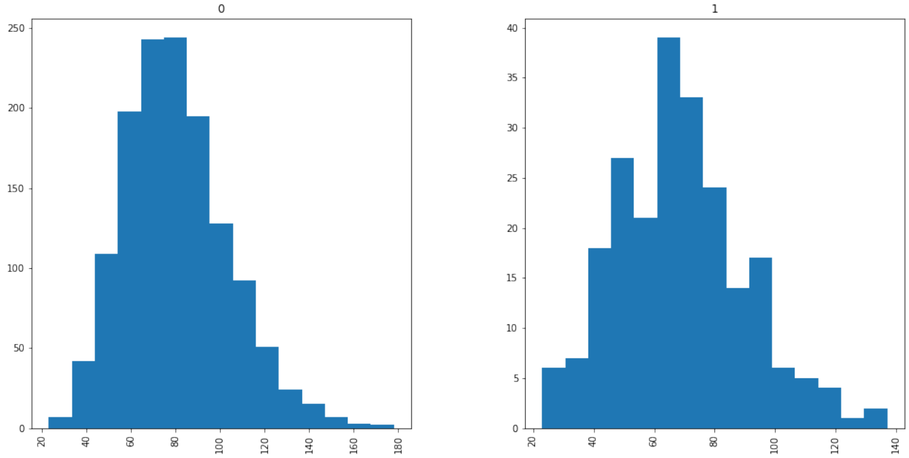

Figure 4: Control/Treatment Comparisson for Median Days on Market

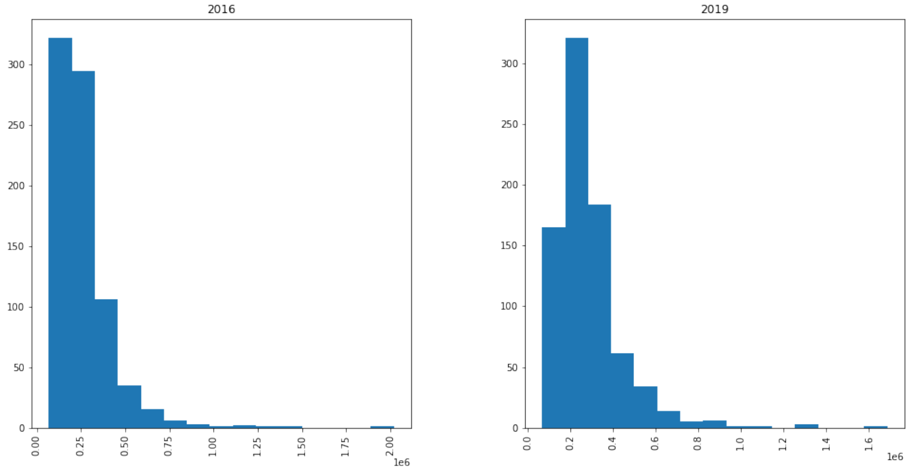

Figure 5: 2016/2019 Comparisson for Median Listing Price

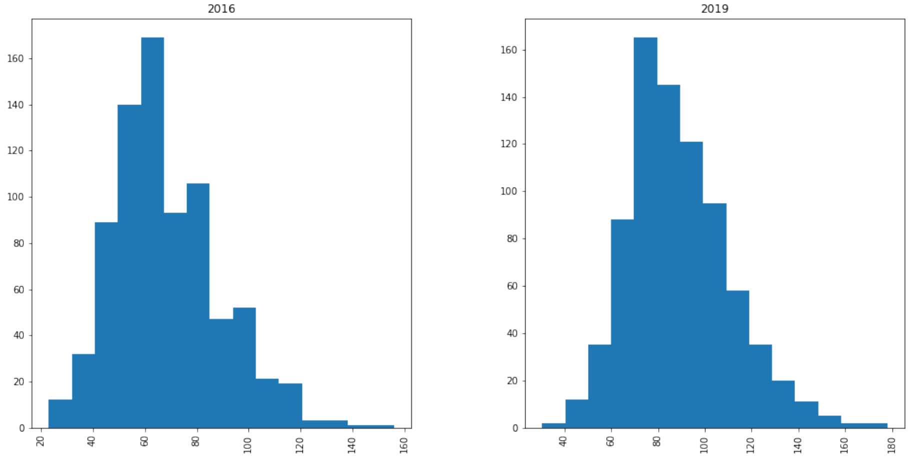

Figure 6: 2016/2019 Comparisson for Median Days on Market

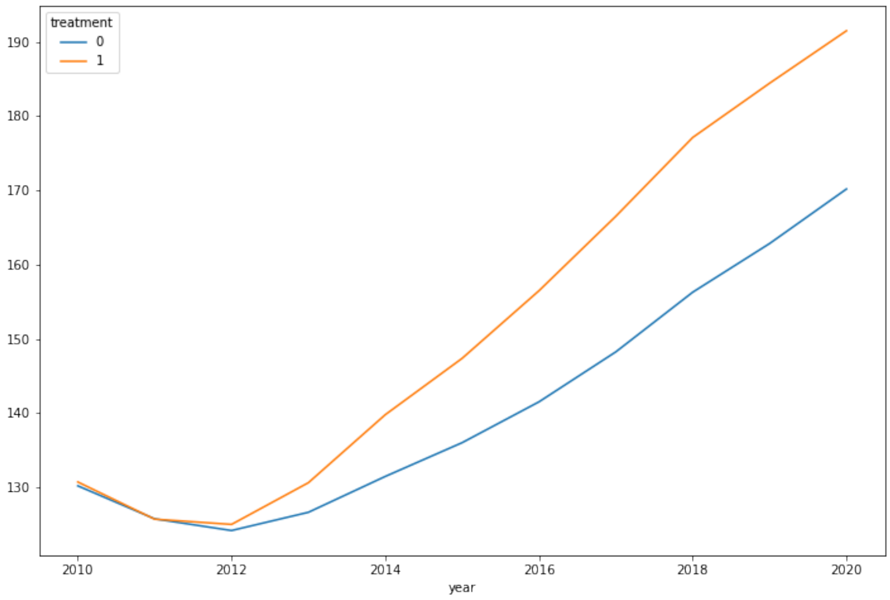

Figure 7: House Price Index Trends for Control and Treatment Data

8.4 Difference-In-Difference Results

Bellow is the results of six difference-in-difference regressions. The columns are as follows:

1. Regress listing price on no covariates
2. Regress time on market on no covariates
3. Regress listing price on all covariates and number of iBuyers
4. Regress time on market on all covariates and number of iBuyers
5. Regress listing price on all covariates and presence of iBuyers (binary)
6. Regress time on market on all covariates and presence of iBuyers (binary)

Table 6: Difference-in-Difference Regression Results

	(1)	(2)	(3)	(4)	(5)	(6)
	__000001	__000001	__000001	__000001	__000001	__000001
time	0.0992***	21.06***	0.0507	25.24**	0.0517	25.21**
	(0.03)	(1.12)	(0.13)	(8.67)	(0.13)	(8.66)
treatment	0.250***	-9.672***	0.0256	0.306	0.0257	0.325
	(0.05)	(2.11)	(0.03)	(1.82)	(0.03)	(1.82)
_diff	-0.0136	-3.726	0.00969	-3.658	0.00842	-4.337
	(0.07)	(2.99)	(0.05)	(3.10)	(0.04)	(2.52)
median_rooms			-0.364***	-5.296***	-0.364***	-5.254***
			(0.02)	(1.23)	(0.02)	(1.23)
homeowner_vacancy_rate			-0.0245***	3.843***	-0.0246***	3.838***
			(0.01)	(0.45)	(0.01)	(0.45)
mean_travel_time_to_work_minutes			0.00478**	-0.301*	0.00483**	-0.298*
			(0.00)	(0.12)	(0.00)	(0.12)
math_prof_pct			-0.00541***	0.105	-0.00545***	0.103
			(0.00)	(0.06)	(0.00)	(0.06)
english_prof_pct			0.00613***	-0.0158	0.00616***	-0.0147
			(0.00)	(0.07)	(0.00)	(0.07)
median_square_feet			0.000131***	-0.000470	0.000131***	-0.000470

		(0.00)	(0.00)	(0.00)	(0.00)
median_age_years		0.0179***	0.0802	0.0179***	0.0758
		(0.00)	(0.24)	(0.00)	(0.24)
move_post2010		-0.00178	-1.182***	-0.00177	-1.184***
		(0.00)	(0.12)	(0.00)	(0.12)
annual_avg_emplvl		-0.000000372***	-0.0000194**	-0.000000377***	-0.0000195**
		(0.00)	(0.00)	(0.00)	(0.00)
median_household_income_dollars		0.00000877***	-0.000404***	0.00000871***	-0.000410***
		(0.00)	(0.00)	(0.00)	(0.00)
per_capita_income_dollars		0.0000124***	-0.0000133	0.0000124***	-0.00000532
		(0.00)	(0.00)	(0.00)	(0.00)
popestimate		0.000000117***	0.00000327	0.000000117***	0.00000329
		(0.00)	(0.00)	(0.00)	(0.00)
rnaturalinc		0.0151**	-0.0984	0.0152**	-0.0936
		(0.00)	(0.32)	(0.00)	(0.32)
rnetmig		0.00610***	-0.0439	0.00610***	-0.0431
		(0.00)	(0.06)	(0.00)	(0.06)
educ_nohs		-0.00601***	0.145	-0.00597***	0.145
		(0.00)	(0.11)	(0.00)	(0.11)
educ_further		0.00375**	0.160	0.00380**	0.161
		(0.00)	(0.09)	(0.00)	(0.09)
average_household_size_of_owner_		0.344***	-4.176	0.343***	-4.176
		(0.05)	(3.47)	(0.05)	(3.47)
white		0.000435	0.105*	0.000482	0.107*
		(0.00)	(0.05)	(0.00)	(0.05)
mean_cash_public_assistance_inco		0.0000248***	-0.00149***	0.0000247***	-0.00149***
		(0.00)	(0.00)	(0.00)	(0.00)
construction		3.33e-08*	0.000000593	3.37e-08*	0.000000580
		(0.00)	(0.00)	(0.00)	(0.00)
mean_retirement_income_dollars		0.0000106***	0.000513***	0.0000106***	0.000515***

			(0.00)	(0.00)	(0.00)	(0.00)
ntop_ibs			0.0692	1.912		
			(0.07)	(5.01)		
nlocal_ibs			-0.00116	-0.189		
			(0.01)	(0.44)		
_cons	12.31***	69.60***	12.36***	65.41***	12.36***	65.42***
	(0.02)	(0.80)	(0.06)	(4.38)	(0.06)	(4.38)
N	1584	1584	1397	1397	1397	1397

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$