MGT 6203

Instructor: Jonathan Fan

Predicting Time on the Market for New Home Listings

Group Progress Report

Team #21

GitHub Repo:
https://github.gatech.edu/MGT-6203-Fall-2023-Canvas/Team-21

Introduction and Problem Statement

Housing finance in the United States is a twelve trillion-dollar industry that touches every corner of the American economy, from a deep relationship with the federal funds rate and employment to historically low-risk government-backed securities for purchase by investors. Often, interacting with buying and selling a home is the largest financial transaction in an individual's life which makes getting the right home at the right time a critical decision.

There are several factors that go into which home to pick to make an offer on, but a key component remains the time that a home has been on the market. For a buyer, a home that has been on the market for a considerable time demonstrates the opportunity to offer a below-asking price due to the presumption of low demand. For a seller, the longer their home is on the market, it triggers more anxiety or re-evaluation of the asking-price, listing details, or even triggering delisting the home from the market.

With the housing market as competitive as it is, it is becoming more and more difficult for buyers to know when to make their offer, and for sellers the same difficulty exists in timing changes in their expectations. By understanding how long a home is expected to be on the market within their metro area, buyers and sellers can act with more confidence in their decision making to remove listings from the market, identify opportunities to offer under-asking, or reevaluate the listing price.

Median time on the market provides a good barometer of the overall pace of transactions. This gives buyers and sellers a benchmark to compare their anticipated transaction against. While not a perfect measure of how long a home will take to sell, the median time on the market approach balances imperfect data availability with an easy to understand metric that can be easily compared against.

From a business perspective, this is a metric we are anticipating being useful in a number of ways. Given the size of the housing finance industry, a business insight that can drive even a component of the industry to a more efficient action is worth a great deal. Realtors often operate on qualitative knowledge of a market, and buyers and sellers typically have limited to no experience at all with buying and selling a home independently.

Creating metrics to get at a quantitative approach for evaluating time on the market for a specific listing is useful to many stakeholders. More knowledge about the time a specific home has been listed relative to the market could drive a buyer to identify a value investment or trigger a seller to lower asking-price to accelerate a sale. These small decisions often mean thousands of dollars of change in asking price. Being able to clearly define when these actions are probable or necessary to drive the desired impact using existing data, can lead to advantages in an often-asymmetric market.

Macroeconomically, time on the market should serve as an effective and measurable proxy for housing demand. Our anticipation is that with more housing demand, median time on the market decreases. A cooler housing market means less demand and our prediction would be more time on the market.

In short, predicting time on the market for new home listings should continue to drive more insight into a critical portion of home buyers and sellers.

Literature Review

There has been many studies in the areas of time on the housing market over the years, with the purpose and methods of these studies having taken various routes.

In 2002, Knight studied the revision of price of a listing during the marketing period to investigate its relationship with the listing price, selling price, and selling time. He used a maximum-likelihood probit model, and found evidence consistent with the theory of pricing behavior under demand uncertainty: He concluded that vacant homes and homes with high initial price, are more likely to change price during the marketing time, but homes that have any unusual features, are less likely to experience a price change.This change in price is very costly to the seller in both time and money. He stated that "homes with large percentage changes in list price take longer to sell and ultimately sell at lower prices" (Knight, 2002). This study identifies vacancies as one of the major reasons for a prolonged time on the market for a listing, and most importantly shows the business impact of price changes. This underlines the importance of making price change decisions in a timely manner to "stop the bleeding."

In 2015, Benefield and Hardin explored the ambiguity in the definition of time on the market by using Multiple Listing Service (MLS) from a medium-sized United States city. They modeled and evaluated five measures of time-on-market. Then they investigated the importance of relisted properties in defining and assessing time-on-market. They concluded that "both time-on-market definition and the handling of relisted properties substantially influence model outcomes and the statistical significance of dependent variables" (Benefield and Hardin, 2015). They also suggested using the term time-on-market needs to be reevaluated and defined for relisted properties. While this is a valid point, the definition of Days-On-Market (DOM) is still commonly defined as the number of days between a property's initial listing and the date it is either sold or taken off the market. It should be noted that many still use the technique of relisting to reset the count for the property. The concept of relisting of a home presents another important option that can be taken with a greater understanding of time-on-market.

In 2019, Castelli et al. attempted to predict days on the market to optimize real estate sales. Their study mainly focused on building an accurate predictive model for the number of days a published listing will be online, with the objective of identifying fake listings. They tested four different modeling approaches for this task: Lasso regression, Ridge regression, Elastic Net regression, and Artificial Neural Networks. The results, obtained on a vast dataset made available by the Bulgarian company Homeheed, showed the effectiveness of Lasso regression. The limitation of this study was focusing on one city only, in Bulgaria. In comparison of our differing outcomes, it will be interesting to see if our final modeling outcome will take the same path.

In 2020, McGreal et al. worked to study "to estimate value effects in relation to the time on the market for residential properties" within the Belfast (U.K.) metropolitan area. They hypothesized that sale price is "a function of marketing time and that properties with a shorter marketing time (TOM) are more likely to sell above list price" (McGreal et al, 2020) . The results suggested that the relationship between sales price and marketing period is mixed and "although sale price is influenced by TOM, the effects are uneven" with properties selling at or above list price likely to have shorter marketing periods compared to those selling below list price. This study embarks through a useful journey and its conclusion has been a common perception lately specifically in the post-COVID housing market.

McGreal mainly focuses on testing the machine learning algorithm's ability to predict the Time-On-Market by combining the real state data with U.S. census data, changes in interest rates and Consumer Confidence Index (CCI). Consequently, to offer a broader perspective, the aim of this study is to better understand different variables impacting the Time-On-Market in the past 6 years. The time period selected intentionally covers the unusual changes to the housing market from pre-pandemic, to pandemic, and post-pandemic era, each with very different landscapes; Bringing their insights to a dataset covering a broader economic time period could help generate insights that cut across economic trends.

Data

To form a well-rounded picture of each metro area to predict median days on the market for a metro area, we plan to use a number of different sources of data. Core to our analysis will be several different components of Realtor.com's data on housing market activity that includes the response variable as well as several predictors. This dataset is aggregated primarily based on regional Multiple Listing Service (MLS). MLS data is the best source of data at the individual region level for home sales, but given each MLS covers an inconsistent area of a region, Realtor.com's aggregation provides a significant amount of value for our purposes.

With the housing specific data will be a variety of components of overall macroeconomic data, with national- and state-level data both available. Census bureau data will bring different angles on demographics that may drive market differentiation, while other macroeconomic factors will help to explain supply and demand for homes. One of the first priorities of our analysis has been stitching these disparate datasets together.

Market Level Data - Zillow

Our first focus of market level analysis was actually on data that Zillow.com publicly made available, rather than the Realtor.com data we ended up pursuing. Zillow-sourced data made up a large part of our initial dataset as well as our initial response variable, mainly oriented around metropolitan statistical areas (MSAs). There were several reasons for this switch.

More of an issue, and core to why we made the switch overall, we soon found out that Zillow MSAs map with CBSA Codes (Core Based Statistical Areas). Most of our other macroeconomic and other external data is based around counties, identified uniquely by Federal Information Processing Standard (FIPS) county and state codes. An individual CBSA code can make up anywhere between 1 to more than 30 counties (FIPS).

This presented a larger question about how to match up other data sets with Zillow's MSAs: whether to aggregate county data to roll up to Zillow's MSAs, or to assume MSA data reflects down to the county level as well. Both represented transformations that could impact the accuracy of our final model appreciably, and limit explanatory power. Rolling up county level information to Zillow MSA could skew our perception of macroeconomic data in favor of lower volume counties, while rolling down from an MSA to a county would certainly limit the variability in our response variable on a county-by-county basis and would not be entirely accurate either.

Another issue with the Zillow data is its completeness (or lack thereof). Depending on the metro area, the fill rate of fields changes appreciably. We had a few options for remedying this, which we pursued to a basic extent. Some of the columns, like "Median Price Cut Amount" were missing data for large regions including San Francisco despite being ranked as the twelfth largest region. Others simply didn't have the volume to be used at all, appreciably exceeding a rule-of-thumb of 5% threshold for imputation. The remainder could be used but observations would need to either be thrown out or imputed, using either the mean of the rest of the data, or by creating a simple regression model based on the other explanatory factors to create more variation. Given these difficulties, we found that the assumptions that we would have to make to use the Zillow data would be strong enough to limit the usefulness of our overall model.

Market Level Data- Realtor.com

Given these issues, we moved on to the Realtor.com dataset. As discussed briefly earlier, the Realtor.com dataset is an aggregated view of MLS data from each different region, which suffers from neither of the issues that the Zillow data did. Fill rate is relatively consistent and high, while the data is ready at the county level, associated with county FIPS code. Additionally, this data is made available by ZIP code, giving us an option with even more location-based granularity.

There are a number of variables of interest that we'll be focusing on within this dataset. The first is the response variable- "Median Days on Market." Other selected data points are below, defined by Relator.com within their data dictionary:

- Avg Listing Price- The average listing price within the specified geography during the specified month
- Median List Price Per Sqft- The median listing price per square foot within the specified geography during the specified month
- Median Listing Sqft- The median listing square feet within the specified geography during the specified month.
- New Listing Count- The count of new listings added to the market within the specified geography. The new listing count represents a typical week's worth of new listings in a given month. The new listing count can be multiplied by the number of weeks in a month to produce a monthly new listing count.
- Price Decrease Count- The count of listings which have had their price reduced within the specified geography. The price decrease count represents a typical week's worth of listings which have had their price reduced in a given month. The price decrease count can be multiplied by the number of weeks in a month to produce a monthly price decrease count.
- Price Increase Count- The count of listings which have had their price increased within the specified geography. The price increase count represents a typical week's worth of listings which have had their price increased in a given month. The price increase count can be multiplied by the number of weeks in a month to produce a monthly price increase count.
- Total Listing Count- The total of both active listings and pending listings within the specified geography during the specified month. This is a snapshot measure of how many total listings can be expected on any given day of the specified month.
- LDP Unique Viewers Per Property (vs US)- The count of viewers a typical property receives in the specified geography divided by the count of views a typical property receives in the US overall during the same month.

Given the shared source and relationship to the overall housing market, we anticipate a significant amount of covariance for these variables.

We are still faced with decisions on data quality. The first is around the Realtor.com data and county population. Counties of very low populations are prone to have missing values for many of our variables. This is not unexpected, as there is likely not enough real estate activity in certain areas to have a meaningful sample size to derive the above metrics. Due to this, a population cutoff will be put in place around the data.

The second issue is one of completeness. Despite the superior fill rate with the Realtor.com data as compared to Zillow, we still face missing values in an appreciable number of observations. Initially, we are tackling this with a simple mean-based imputation approach. Depending on further analysis here, we may be comfortable with that or have to take a more nuanced approach.

Macroeconomic Data

The next area of analytical focus was the array of macroeconomic data we had sourced. There are multiple dimensions to our focus here. The first is forming an accurate picture of the overall macroeconomic climate at any one data point. As our dataset progresses through time, we'll see variations in both the Zillow data and the macroeconomic climate that will tune our model.

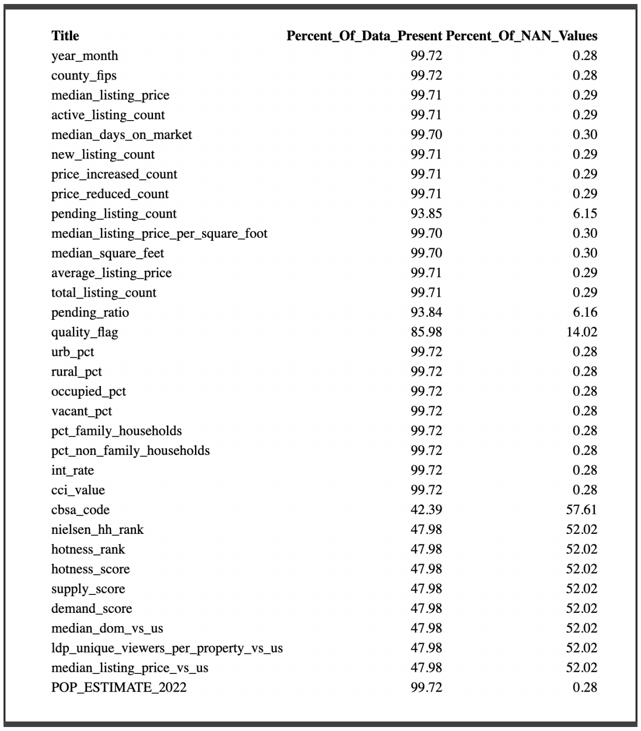
We expect that the most critical driver of housing economic activity will be changes in the federal funds rate. The period we are analyzing, 2017-2023, has had great shifts in interest rates which have impacted the volume of mortgages tremendously (Rothstein, 2023). We've gone from federal funds rates at effectively 0% in February 2022 to 5% and climbing in September 2023, which has a strong direct relationship with mortgage rates. Our expectation is that there will be a strong relationship between the federal funds rate and "Median Time on the Market", as a "hot" mortgage market with low rates means more offers coming in for a property, often over the asking price and in cash (Rothstein, 2023). This provides easy decisions for sellers and requires quick action for buyers to lock in advantageous mortgage terms. So, anticipating strong positive covariance between interest rates and "Median Time on the Market". From a usability perspective, given federal funds rate's impact on the entire US housing markets, this will be a column that applies to each MSA, making mapping straightforward. An interesting point of further exploration would be to see the differentiation between MSAs in their volume reaction to interest rates.

Another market-specific "Median Time on the Market" driver we'll explore is census bureau data. We have available the 2020 Census Demographic and Housing Characteristic Survey, as well as the American Community Survey data. Different markets will have very different household makeups, and the demographic makeup of households may drive greater demand and therefore time on the market. A rough guess on the relationship here would be that younger areas have higher demand for housing to support growing families, while areas that have higher prevalence of multigenerational households may see some of that demand lessened. This data is available at a county and state level by FIPS code.

Other areas of macroeconomic interest are Real Personal Income and the Consumer Confidence Index. Real personal income has a potential to be of importance given the impact of inflation on disposable income and housing affordability, and therefore demand for housing. Consumer Confidence Index on the other hand may impact consumer's propensity to spend on housing. If consumers feel an economic downturn is coming or feel squeezed by the rising costs of goods, they may be less likely to make a large financial commitment.

Exploratory Data Analysis

Table 1 shows the summary statistics of the numerical variables for our compiled dataset. We have sorted out our data to counties only with populations greater than 5,000 as of 2022. Below, we have a table to note the presence of data in our compiled dataset. We examined the frequency of having NaN values as opposed to actual values. We see that most of our variables are present in greater than 90% of our data points. We also see a smaller subset of variables that are only present in ~ 45% of our data points as well. There is possibly some bias in the data points in this group, but we will further explore.



Title	Percent_Of_Data_Present	Percent_Of_NAN_Values
year_month	99.72	0.28
county_fips	99.72	0.28
median_listing_price	99.71	0.29
active_listing_count	99.71	0.29
median_days_on_market	99.70	0.30
new_listing_count	99.71	0.29
price_increased_count	99.71	0.29
price_reduced_count	99.71	0.29
pending_listing_count	93.85	6.15
median_listing_price_per_square_foot	99.70	0.30
median_square_feet	99.70	0.30
average_listing_price	99.71	0.29
total_listing_count	99.71	0.29
pending_ratio	93.84	6.16
quality_flag	85.98	14.02
urb_pct	99.72	0.28
rural_pct	99.72	0.28
occupied_pct	99.72	0.28
vacant_pct	99.72	0.28
pct_family_households	99.72	0.28
pct_non_family_households	99.72	0.28
int_rate	99.72	0.28
cci_value	99.72	0.28
cbsa_code	42.39	57.61
nielsen_hh_rank	47.98	52.02
hotness_rank	47.98	52.02
hotness_score	47.98	52.02
supply_score	47.98	52.02
demand_score	47.98	52.02
median_dom_vs_us	47.98	52.02
ldp_unique_viewers_per_property_vs_us	47.98	52.02
median_listing_price_vs_us	47.98	52.02
POP_ESTIMATE_2022	99.72	0.28

Table1: Summary statistics of the numerical variables (county population >5000 in year 2022)

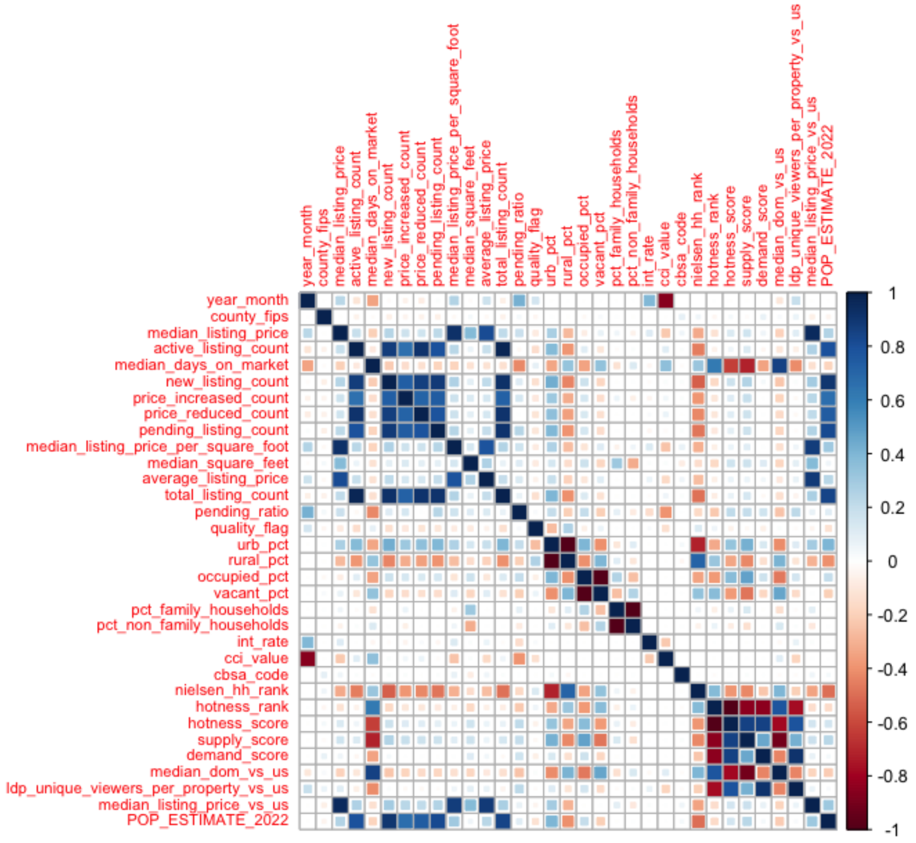


Figure1: Correlation matrix of main consolidated real estate dataset.

Figure 1 shows a simple correlation plot of our variables in our dataset. We see that there are some variables that are highly correlated, such as active_listing_count & new_listing_count, which make sense since they imply a popular housing market. We also see these values correlated with population, which makes for the assumption that the areas with larger populations are where most of our new active listings are coming from. This corresponds well with our high positive correlation with urban percentage & new active listing count, alluding that new listings are occuring in counties that are urban as opposed to rural. Most importantly, our dependent variable median days on market is strongly negatively correlated with pending ratio, urban percentage, occupied percentage, hotness score, supply score, & unique viewers per property We then see a strong positive correlation with hotness rank & median days-on-market. Table 2 shows the top five highly correlated variables. All pairs of highly correlated variables are expected and logically explainable. One interesting observation here is the 92% correlation between the "active listing count" and "price reduced count" which suggests that when there are more active listings for a particular house, there is a higher possibility that the seller is reducing the price in order to attract potential buyers, rather than increasing the price, as the correlation between the "active_listing_count" and "price_increased_count" is notably lower, and at approximately 65%.

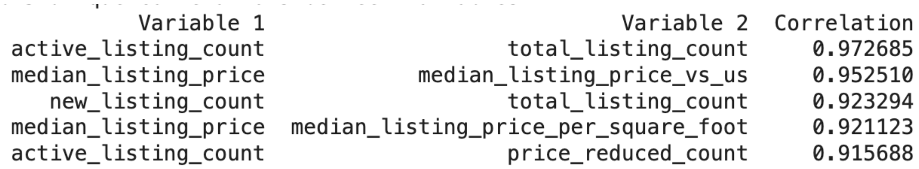

```
           Variable 1                            Variable 2  Correlation
  active_listing_count                    total_listing_count     0.972685
  median_listing_price               median_listing_price_vs_us     0.952510
     new_listing_count                    total_listing_count     0.923294
  median_listing_price  median_listing_price_per_square_foot     0.921123
  active_listing_count                    price_reduced_count     0.915688
```

Table 2: Top five Correlation Coefficient of main consolidated real estate dataset.

In Figure 2, we have the Consumer Confidence Index in the United States from 2016 to 2023. We can see that there is a significant decrease in CCI in 2020, at the time of the COVID-19 pandemic & its lasting economic effects. Since our data covers a span of both economic prosperity & downfall, we

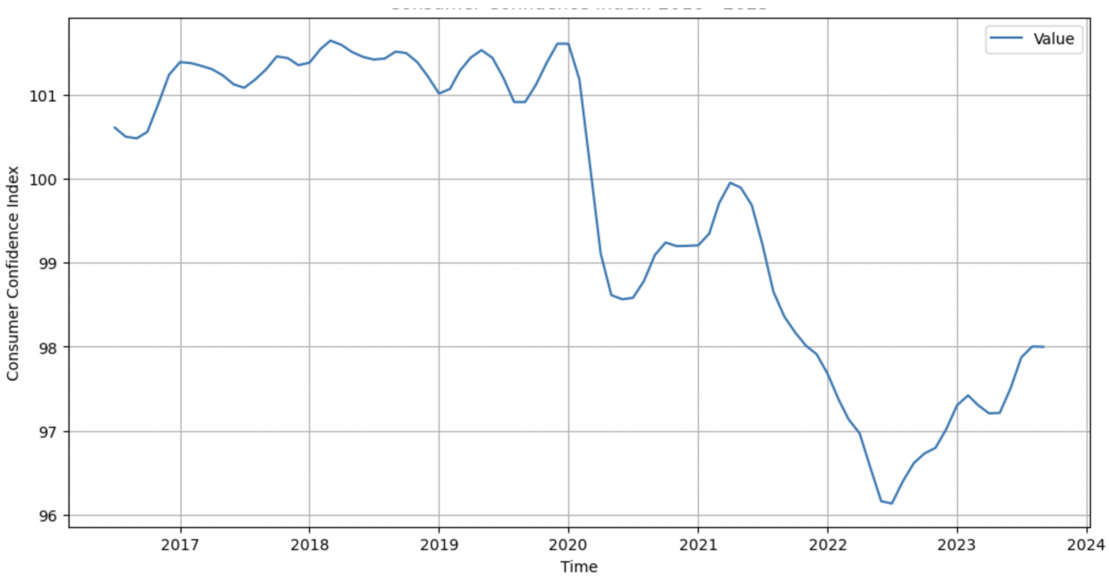


Figure2: Consumer Confidence Index from 2016 – 2023.

expect that time has a large impact on specific variables. Since our data is based monthly, time is also standardized when assessing other variables, since those data are taken at the same time. We also may assess different modeling based on the specific dates of pre-COVID & post-COVID to look at these different macroeconomic conditions.

Modeling

So far we have just created univariate regression models; looped through all the desired independent variables. In the upcoming weeks we will experiment with multivariate models, polynomial models and logistic transformations. Moreover, in terms of hyper-parameters, we will explore our options with lasso and ridge regression techniques in the Sklearn module to optimize the regression model and reduce overfitting.

Lastly, we will also try using regression trees and a vanilla ANN model. For the latter we will normalize each independent variable to increase model accuracy. To gauge each model's effectiveness, we will select the model with the lowest root mean squared error in the test set (we will create a stratified train, validate and test subsets).

Works Cited

Rothstein, Robin. “Mortgage Rate Forecast for 2023.” Forbes, October 31, 2023. https://www.forbes.com/advisor/mortgages/mortgage-interest-rates-forecast/.

Benefield, J.D., Hardin, W.G. Does Time-on-Market Measurement Matter?. J Real Estate Finan Econ 50, 52–73 (2015). https://doi.org/10.1007/s11146-013-9450-z

Knight, J.R. (2002), Listing Price, Time on Market, and Ultimate Selling Price: Causes and Effects of Listing Price Changes. Real Estate Economics, 30: 213-237. https://doi.org/10.1111/1540-6229.00038

Mauro Castelli, Maria Dobreva, Roberto Henriques, Leonardo Vanneschi, "Predicting Days on Market to Optimize Real Estate Sales Strategy", Complexity, vol. 2020, Article ID 4603190, 22 pages, 2020. https://doi.org/10.1155/2020/4603190

Stanley McGreal, Alastair Adair, Louise Brown & James Webb (2009) Pricing and Time on the Market for Residential Properties in a Major U.K. City, Journal of Real Estate Research, 31:2, 209-234, DOI: 10.1080/10835547.2009.12091248 https://doi.org/10.1080/10835547.2009.12091248

Realtor dictionary. https://www.realtor.com/research/data/