Disha Saha

Project 1: Milestone 4

Recently there has been a lot of buzz around the real estate market as prices have skyrocketed in recent months due to a lot of unique challenges in the market. There have been changes in residents moving from urban to suburban areas within cities, a shortage of inventory due to the Coronavirus pandemic and individuals not wanting to sell or move for the past year. Interest rates have dropped to an all time low and individuals have also been moving from high taxed states to low taxed states as remote working became the norm. With so many unique changes it may be difficult to quantify just how this has impacted real estate prices in recent months. My plan is to attempt to quantify this change and predict where prices might land once the short issues have been resolved. In this paper, I will look into these recent price changes and build a model using various metrics in order to see how these recent changes have impacted the real estate market.

To start with low inventory seems to have contributed to the increase in prices. According to Business Insider, there are a few reasons why the housing shortage may continue for some time. The first being millennials are now entering the home buying stage in their life, the second being housing starts were at their lowest levels the past few years and have just now started to pick up, and finally foreclosures which were expected to occur during the recent COVID recession won't appear to have a meaningful impact on supply (Sheffey, A). The below chart which has been used in my analysis shows housing starts in the US since 1960. As you can see after the collapse of the housing market in 2008 we had a steep decline in new homes being built in the US, which has since picked up but is still nowhere near the peaks that came before. According to the article from bankrate.com, we are short about 3 million homes while at the current pace we are only building 1.42 million homes per year, which would contribute to the real estate market being squeezed (Wichter, Z).

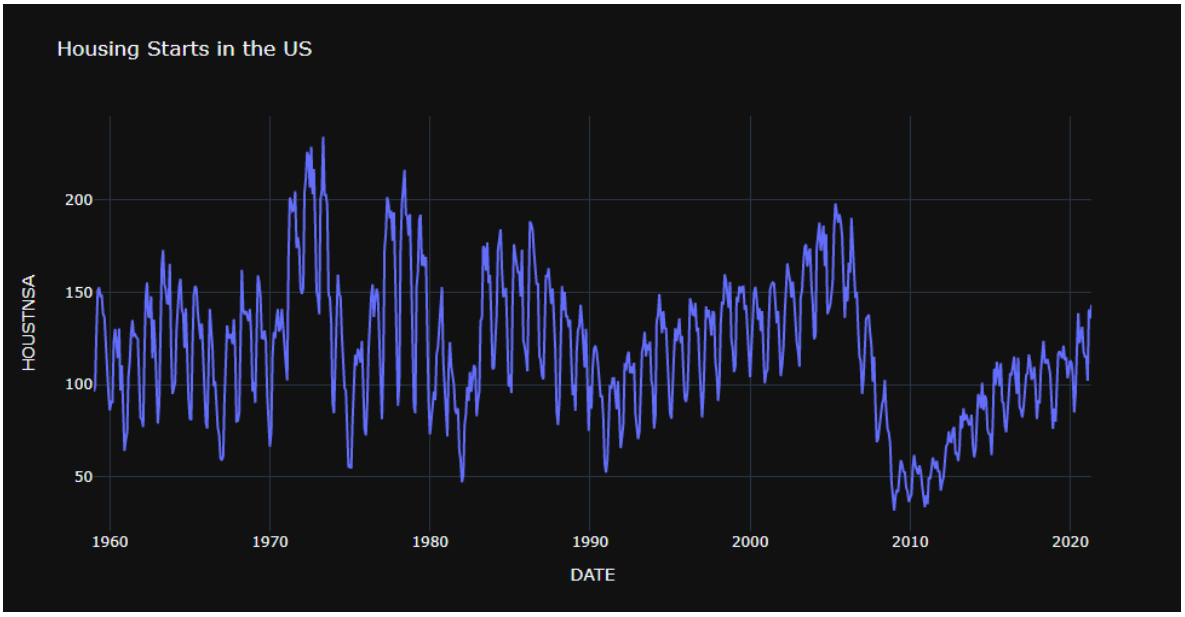


Fig 1. Housing Starts in the US

Another factor in low inventory has been the decrease in listings over the coronavirus pandemic. According to redfin and data published by the St. Louis Fed, "Home sales in April and May 2020 dropped to their lowest levels since the housing and financial crisis that began in 2007, with many homeowners hesitant to sell in the wake of the pandemic" (Gascon, C. S., & Haas, J.). However, in late spring demand began to pick back up but inventory and new listings remained low causing prices to surge. (st louis fed.) As you can see in the chart below I pulled data to compare the top 4 real estate markets; Florida, Texas, New York, and California. All 4 of these markets have had declines in their active listings over the past year and half, with Florida and Texas having the steepest declines. While demand continues to pick back up, the active listings in these markets have not yet recovered.

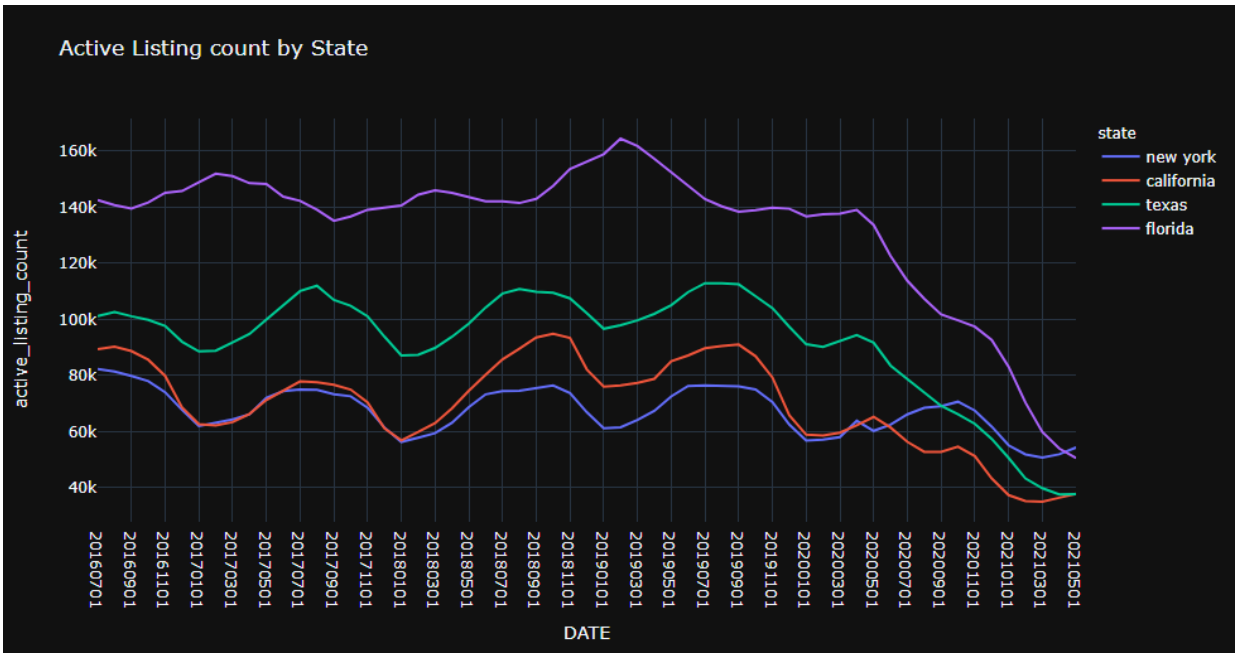
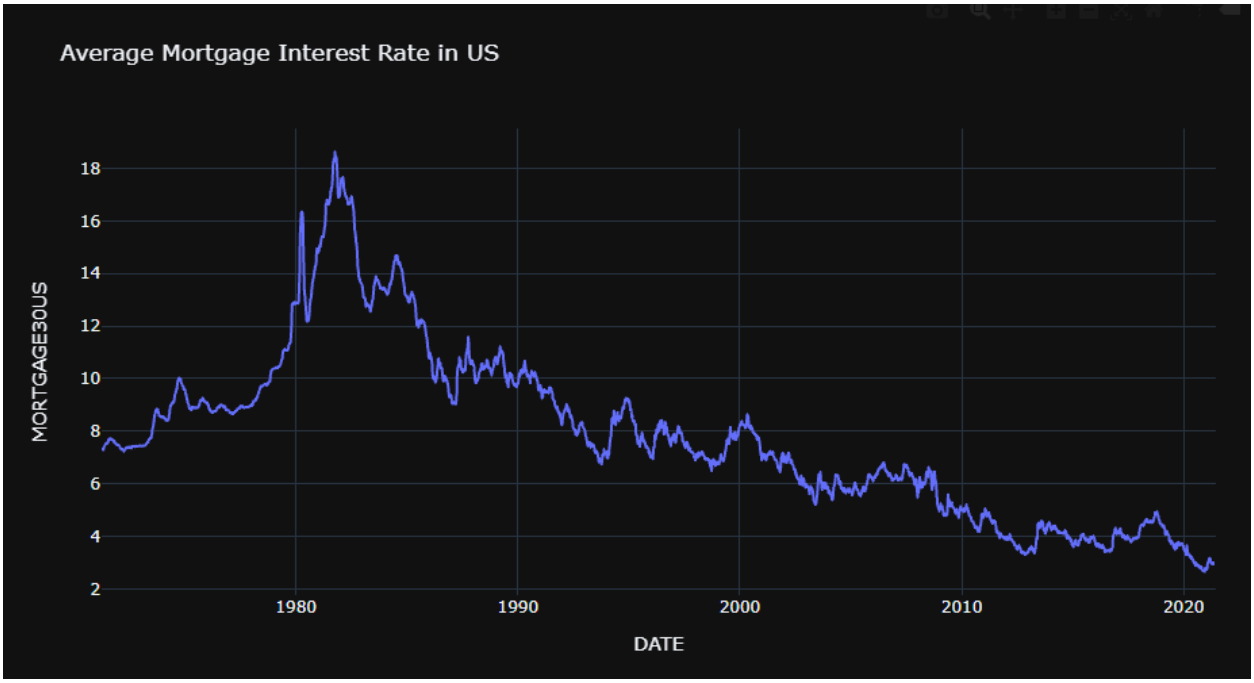


Fig 2. Active Listings in the US by State

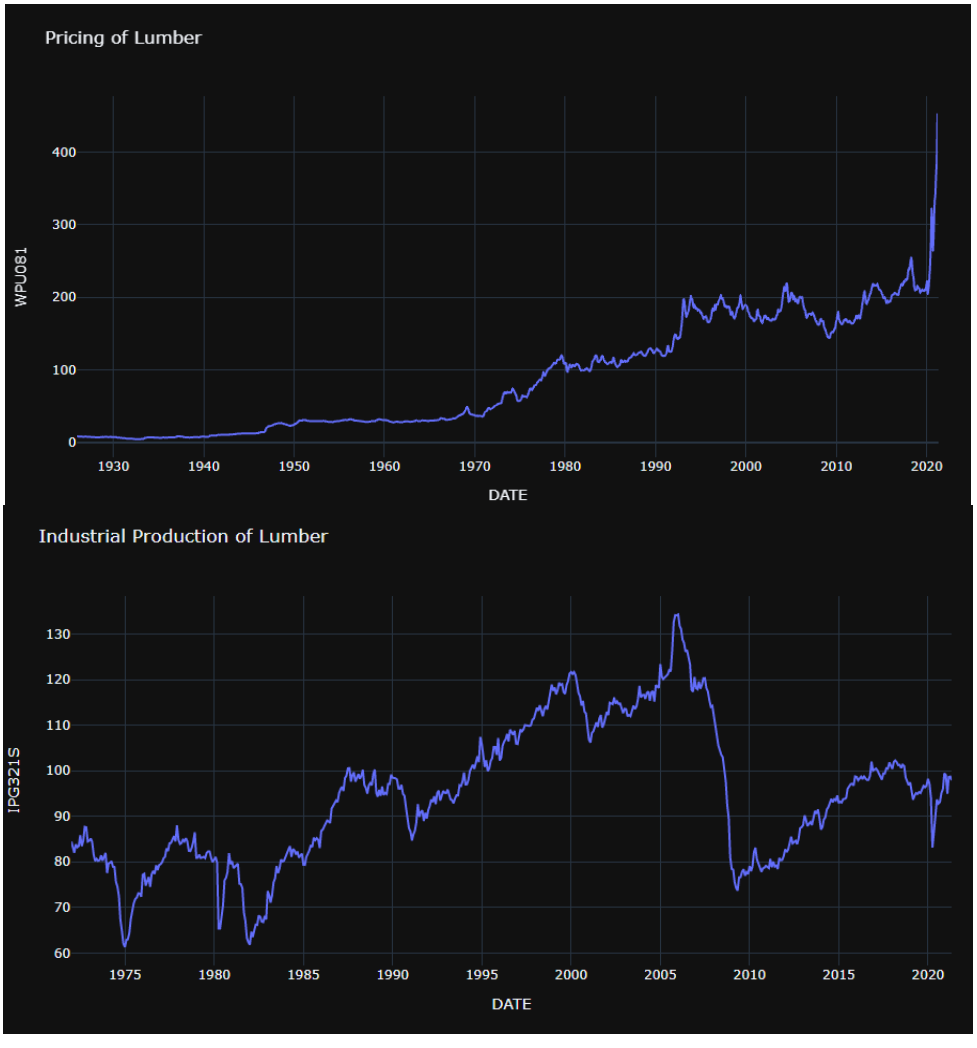
Another situation which seems to be fueling demand is the decrease in interest rates. Interest rates for mortgages during the end of the pandemic dropped to less than 3% from a high of 4.86% in November 2018. Due to interest rates dropping, more individuals can afford a more expensive home which drives demand up as some individuals look to trade up their home.

Fig 3. Mortgage Interest Rates in the US



The third and perhaps the most influential metric that was considered in the analysis was lumber supply and lumber pricing. According to the charts lumber supply took a steep decline in April 2020 as the nation shutdown for the COVID pandemic, it has since recovered but due to the dip in supply it has caused a backlog in the supply chain as demand increased. Due to this lumber prices have skyrocketed to new highs which would impact new homes being built. According to an article by yahoo finance, "at the same time that state-mandated lockdowns caused sawmills to halt production, bored, quarantining Americans were rushing to Home Depot and Lowe's to buy up materials for do-it-yourself projects. That caused lumber inventory to plummet. It only got worse from there: Recession-induced record-low interest rates helped to spur a housing boom. As of this spring, home construction was at its highest levels since 2006. Production has swung back at sawmills, but capacity is too limited to match the explosion on the demand side" (Lambert, L)

Fig 4. Lumber Prices and Production



Now that I have examined some of the underlying data of the real estate market I will begin the analysis of looking at how each of these factors has impacted real estate prices and examine our model to determine how things may look in the future. To start the main metrics I have looked at for my regression analysis are listed below along with their correlation with average listing price. As you can see lumber price has the largest correlation with correlation coefficient of 0.85 and total listing count having the second largest impact with a negative correlation on price of -0.75.

Fig 5. Correlation Matrix

average_listing_price	1.000000
WPU081	0.856473
pending_ratio	0.811223
price_increased_count	0.689253
HOUSTNSA	0.621110
pending_listing_count	0.347459
new_listing_count	-0.320936
IPG321S	-0.341273
median_square_feet	-0.430699
median_days_on_market	-0.566468
price_reduced_count	-0.672568
MORTGAGE30US	-0.688241
total_listing_count	-0.750958

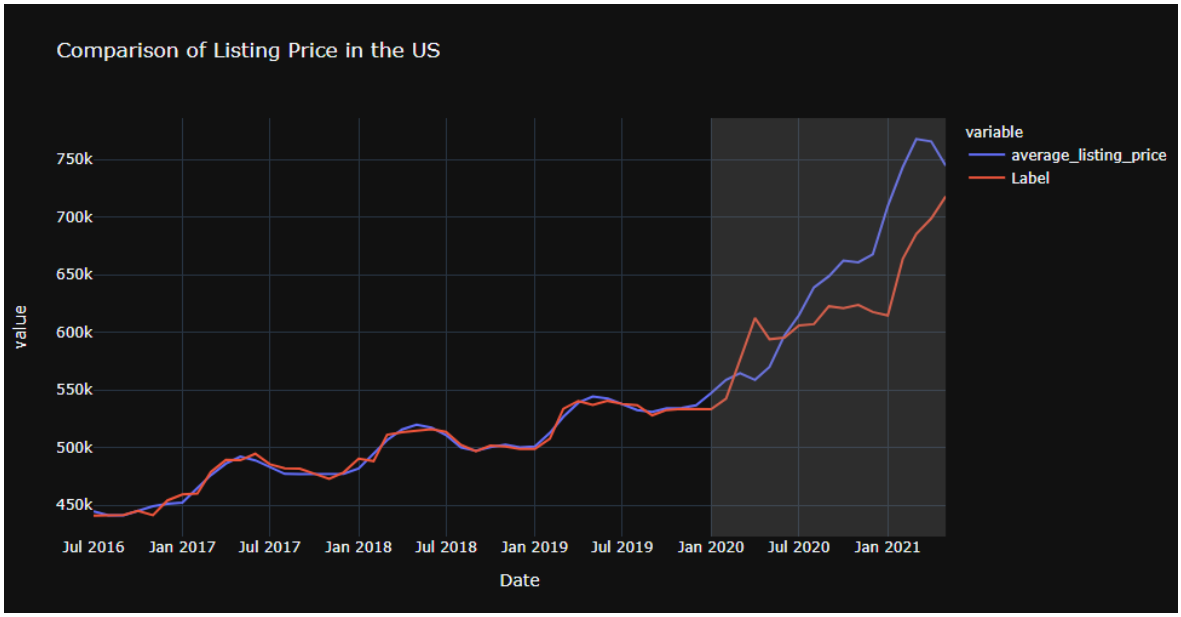
My next step in the analysis was to create train and test sets of the data. To train the model I looked at years 2020 and below and my test set included 2020 through May 2021. To create a regression model I used pycaret which was able to run multiple machine learning algorithms and come up with the best model to use in this scenario. Regressing on the average listing price, I found the Lasso Regression the best model to use.

Fig 6. Comparison of machine learning regressions using pycaret

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lasso	Lasso Regression	5331.4312	39539790.3000	5985.5941	-183.3589	0.0122	0.0109	0.1450
lr	Linear Regression	5547.7881	42899654.9000	6262.5692	-250.4146	0.0128	0.0113	0.2130
ridge	Ridge Regression	5480.2650	42295283.2000	6218.4830	-251.1697	0.0127	0.0112	0.0030
llar	Lasso Least Angle Regression	5385.6501	41748437.9945	6148.7971	-252.5129	0.0125	0.0110	0.0040
en	Elastic Net	5903.9367	55036149.6000	6950.7856	-293.3231	0.0143	0.0122	0.0040
gbr	Gradient Boosting Regressor	12081.8703	336408387.7623	13472.2280	-632.1976	0.0275	0.0246	0.0060
ada	AdaBoost Regressor	13772.6247	418066327.9448	15340.4838	-731.3711	0.0313	0.0282	0.0090
et	Extra Trees Regressor	15561.9524	409614975.3316	16485.4906	-759.7855	0.0335	0.0319	0.0180
rf	Random Forest Regressor	14991.2143	399231467.0079	16282.0436	-1032.0155	0.0331	0.0305	0.0230
br	Bayesian Ridge	13306.5526	300502329.2709	15820.1372	-1820.7274	0.0317	0.0266	0.0040
omp	Orthogonal Matching Pursuit	19840.3194	535469876.4762	20909.9244	-3767.8480	0.0422	0.0401	0.0030
dt	Decision Tree Regressor	15042.9189	451144571.2096	17305.8951	-3836.8792	0.0354	0.0308	0.0040
knn	K Neighbors Regressor	27563.7301	1081367016.4000	28912.9452	-4050.9856	0.0583	0.0562	0.0050
lightgbm	Light Gradient Boosting Machine	27511.5732	1094124605.2506	28703.7568	-4378.6559	0.0579	0.0557	0.0460
huber	Huber Regressor	45298.5798	3410986998.5995	49078.7864	-4394.5606	0.1006	0.0920	0.0090
par	Passive Aggressive Regressor	64484.0343	5359542527.5875	68906.9346	-13156.5432	0.1437	0.1314	0.0030
lar	Least Angle Regression	306372.9688	615638370469.6193	332801.5841	-14326.7240	0.3548	0.6057	0.0040

Using this model I then plotted the estimated price using the train and test sets. As you can see the model does a decent job following the curves up to the middle of 2020. As of May 1st we have a delta of an average listing price in the US of roughly 30k.

Fig 7. Comparison of the train and test models on listing price in the US



In order to mitigate the effect of the recent impacts on the real estate prices I modified the underlying data to mimic similar 2019 data such as interest rates and lumber pricing. The result of this analysis using our regression model would peak around 600k, which I believe once the short term issues resolve to indicate what our relative home price would be in the future. Based on this analysis, I would estimate that the coronavirus pandemic has contributed a 25% increase in home prices considering that the pandemic has sparked low interest rates, low inventory in homes, and a backlog in the lumber supply chain.

Questions and Answers

1) Where did the data come from?

   The data for mortgage interest rates, housing starts, and lumber pricing come from https://fred.stlouisfed.org/ which is the federal reserve bank of St. Louis. The other features and pricing come from http://www.realtor.com.

2) What other factors could be considered in the analysis?

   Other factors or features that could have been considered are more housing specific metrics such as square footage, bathrooms, and bedroom counts. The current analysis that has been done has been on average pricing in the US. The model has also been used on State specific information as well which was included in the jupyter notebook.

3) How long would it take for the impacts from the pandemic to dissipate?

   According to the lumber commodity price index, lumber prices have been falling since their high in May 2021 which I think would play a key role in home prices coming down (https://www.nasdaq.com/market-activity/commodities/lbs). Mortgage interest rates have also been slowly increasing but it may take a year or two for them to return to pre recession levels. As mentioned earlier bankrate believes that the shortage will continue over the next few years (https://www.bankrate.com/real-estate/why-are-house-prices-going-up/). Overall I think we may see home prices decrease in the short term due to lumber pricing, however long term they will continue to go up in value.

4) Have you modelled the other markets and what was the outcome?

   Yes, I have also modelled the top 4 markets that were discussed earlier, Florida, Texas, California, and New York. The result from running pycaret with the state level data produced an R^2 of 0.95 with extra tree regressor being the best regression model. The results range from 100k gap in Texas, 150k gap in Florida, 200k gap in New York, and a 1m gap in California.

5) Why is the data modelled only from 2016 - 2021?

   While I had previous year data from the FRED database, realtor.com only offered data from 2016 onwards.

6) What is lasso regression?

   Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination. (https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/)

7) What is the pending ratio?

   The ratio of the pending listing count to the active listing count within the specified geography during the specified month. (https://www.realtor.com/research/data/)

8) What is pycaret and how does it work?

   Pycaret is an open-source, low-code machine learning library in Python that automates machine learning workflows. It is an end-to-end machine learning and model management tool that speeds up the experiment cycle exponentially and makes you more productive. (https://pycaret.readthedocs.io/en/latest/). In short Pycaret integrates multiple machine learning libraries and allows individuals to run different algorithms very quickly and in fewer lines of code.

9) What does squeezing the real estate market mean?

   Squeezing the real estate market would happen with demand vastly outpacing supply causing a sharp increase in price.

10) What is causing the interest rates to decrease in the real estate market?

    After the coronavirus pandemic was declared and multiple cities in the United States had to lock down to prevent the spread of the virus, an economic recession had started. In order to combat the recession the federal reserve needed to lower interest rates to encourage borrowing and spending to keep the economy running.

References:

1) Lambert, L. (n.d.). The unprecedented lumber shortage, in 8 charts. Retrieved June 26, 2021, from https://www.yahoo.com/now/unprecedented-lumber-shortage-8-charts-223000018.html
2) Sheffey, A. (2021, May 19). 3 reasons why the housing shortage will last for years, Goldman Sachs says. Retrieved June 26, 2021, from https://www.businessinsider.com/housing-shortage-real-estate-inventory-foreclosures-builders-millennials-goldman-sachs-2021-5
3) Wichter, Z. (n.d.). Why Are House Prices Going Up So Much? Retrieved from https://www.bankrate.com/real-estate/why-are-house-prices-going-up/
4) Gascon, C. S., & Haas, J. (2021, January 04). The Impact of COVID-19 on the Residential Real Estate Market. Retrieved from https://www.stlouisfed.org/publications/regional-economist/fourth-quarter-2020/impact-covid-residential-real-estate-market