CS298 Data-Driven Decision-Making Final Project Paper

Factors Influencing Real Estate Prices

Vivian Xia and Andrii Liudvichuk

Abstract

Real estate is in a constant state of flux, so it is important to understand what factors have the biggest impact have on making the house prices. The topic of changing real estate prices was researched using credible sources. By researching, the factors that were found in articles and studies on the topic of real estate prices were the population estimate, population density, the average income, and the technological improvement of the region. And through an exploration of the data, statistical analysis, and data visualization, it was revealed that the average income had no significance in predicting real estate prices in the model. The factors, population density, population estimate, and technological improvement, on the other hand, did have an impact on the real estate prices.

Introduction

The purpose of this project is to address the factors that impact real estate price. As the pandemic progresses, the population continues to decline with over one million deaths up to this date. People are also moving out of cities where there is the highest population densities because of the pandemic. With the decline in both population estimate and population density in certain areas, the question remains, what is the predicted effect it will have on housing prices? With that, James R. Follan Ph.D. has produced "A Study of Real Estate Markets in Declining Cities," has found that "substantial price declines can be expected to occur at the MSA, [metropolitan statistical areas], level in places that suffer substantial and persistent declines in population and employment" (2010, p.33). The finding references the decline in population density, population estimate, and employment in cities. This conclusion has identified three factors that should be analyzed in this project in understanding the real estate prices including

population density and population estimate as well as employment, which would be associated with income.

Logically, the more people earn, the more people are willing to spend. Tejvan Pettinger supports that notion of a strong correlation between the income and the house prices, as the "demand for housing is often noted to be income elastic (luxury good); rising incomes leading to a bigger % of income being spent on houses" (Pettinger, 2019). In other words, people having higher than average income are willing to spend more on real estate. Real estate prices are usually driven by the earnings in the region. For example, New York, one of the most expensive home price locations, has the largest average salary among all cities in the United States.

Another factor that impacts the price of real estate is technological improvement. One instance of a state where the real estate prices boomed because of the technological companies is Seattle. It has not been too long ago since Amazon, one of the largest companies in the world, settled in Seattle. And according to the Census data "rent around the Seattle metro area had increased 17 percent from 2011 to 2015, triggering a supply and demand challenge" (U.S. Census Bureau, 2020). This instance showed that technological improvements in the region can impact the real estate prices in that same area. The states with a higher technological index have much higher house prices. Technological hubs have higher-priced prices for homes close to it due to a large number of highly-paid jobs there. People are willing to move from, for example, Idaho to California, as the tech index is much higher there. This project will not only explore the factors population density, population estimate, average income, but also technology of the state and their impact on real estate prices.

Data

The real estate prices data was found from Zillow. Zillow publishes their housing data and consistently updates it with a transparent methodology and calculation process. The real estate prices data contains the home values of the 35-65% of the United States by counties. Because the data provides a narrow selection of 35-65 percentile of home prices, the prices are do not include the lower 35%, which are the house transactions made for a nominal price to save on taxes and other fees when making the buy/sell operation as well as not include the top 35% which are the highest prices which are located generally in California, Massachusetts and New York. The data also included each counties' corresponding state, which was later used when totaling the total prices for each state. The data values represent all homes per square foot including for single family rooms, condo/co-ops and its value that is smoothed and seasonally adjusted from 2010.

The technology scores data comes from a credible source, the Milken Institute which is a nonprofit independent organization that works to improve lives around the world through conducting and producing economic research. The report containing this data also offers more in-depth analysis to explain the scores for each state. The data contains a composite technology score for each state for 2014. The composite technology score was found by scoring five sub-indexes (Human Capital Investment, Risk Capital and Entrepreneurial Infrastructure, Research and Development Inputs, Technology Concentration and Dynamism, Technology and Science Workforce) out of 100.

The population estimate data was found through the United States Census Bureau, a credible source as it is a United States government organization. The population

estimate data contains county-level population totals of 2010. It contained the population estimate for each county in the U.S. as well as identified its corresponding state, which allowed us to later group by the state to find the total population estimate for each state.

The population density data was found through the United States government organization, the United States Census Bureau. The population density data contains state-level values for 2010. The population density values are representative of people per square mile.

The income per capita data was found through an official website run by the United States government, The Bureau of Economic Analysis: U.S. Department of Commerce. The data contains state-level data on incomes per capita for 2010.

Data Analysis

	Average Income	Technology Scores	Population Density	Population Estimate	Real Estate Prices
Mean	29,801.5	52.9963	194.962	12,325,628.16	6,563,802.46
Median	38,374	52.81	98.75	8,872,739	5,430,279.5
Mode	N/A	N/A	153.9	N/A	N/A
Standard Deviation	6,376.574	14.603	261.091	13,696,436.64	4,692,938.519
Count	50	50	50	50	50

Figure 1. Descriptive Statistics of Each Variable.

The statistics told us how the data is distributed. Referencing Figure 1, the technology scores, average income, and real estate prices have a small difference between its mean and median. This small difference means that the data for those variables are distributed almost symmetrically since the data is distributed quite equally around the mean. Unlike those, population estimate and density are positively skewed distributions, because the mean is much higher than the median, which denotes that there are a few states with a much higher population density and population estimate. Each variable has a count of 50, which corresponds to the data collected pertaining to each state.

Regression Analysis

Summary	
Adjusted R Square	0.473
Significance F	1.00E-06
	P-value
Average Income	0.429
Population Density	0.182
Population Estimate	7.15E-07
Tech Scores	0.085

Figure 2. Ordinary Least Squares Regression Analysis with the following independent variables: Average Income, Population Density, Population estimate, Technology Scores.

The output of the regression analysis for the Ordinary Least Squares regression, shown in Figure 2, the relationships between the dependent variable and the independent variables can be observed. There is a positive relationship between real estate prices and the variables, population estimate and technology scores. There is a negative relationship

between real estate prices and the variables, population density and average income.

The regression analysis reveals which independent variables are significant and how successful the model is. The adjusted R squared was 47.3%, so 47.3% of the real estate prices variability is explained by this model. The Significance of F is very low, denoting that this model is useful in terms of predicting the real estate prices.

Looking at the p-values, population estimates and technology scores are statistically significant in predicting real estate prices due to their low p-values. The average income, on the other hand, is not that useful for the regression, as its p-value is quite high. Because the p-value of average income is high, average income is dropped from the model and the regression analysis of the new model is shown below in Figure 3.

Summary	
Adjusted R Square	0.478
Significance F	3.07E-07
	P-value
Population Density	0.053
Population Estimate	4.02E-07
Tech Scores	0.117

Figure 3. Ordinary Least Squares Regression Analysis with the following independent variables: Population Density, Population estimate, Technology Scores.

The new model's regression analysis, shown in Figure 3, shows a larger Adjusted R-squared at 47.8% compared to the 47.3% of the last model. This model's Adjusted R-squared represents that 47.8% of the real estate prices variability is explained by this model. The resulting Significance F is also smaller compared to the last model, so this updated model and its use of these three independent variables are more useful in terms of predicting the real estate

prices than the last model.

And with these three independent variables, it would be interesting to see their relationships with one another, so a correlation is created between the three variables and shown in Figure 4.

	Population Density	*Population Estimate*	*Tech Scores*
Population Density	1		
Population Estimate	0.172000838	1	
Tech Scores	0.471494638	0.348058198	1

Figure 4. Correlation between the following three independent variables: Population Density, Population Estimate, and Technology Scores.

The correlation shows that the population density and population estimate have a very weak correlation with one another. This observation does corroborate with why population estimate has a very low p-value of 4.02E-07, but population density has a larger p-value of 0.05. The population density and technology scores have a weak relationship but it is stronger than that of the relationship between population estimate and technology scores. Because none of the correlation values are above 0.8, it can be assumed that there is no multicollinearity -- no linear relationship -- between the variables in the model.

Data Visualizations

The data visualizations reaffirmed the relationships between each independent variable and the dependent variable that could be seen from the regression analysis.

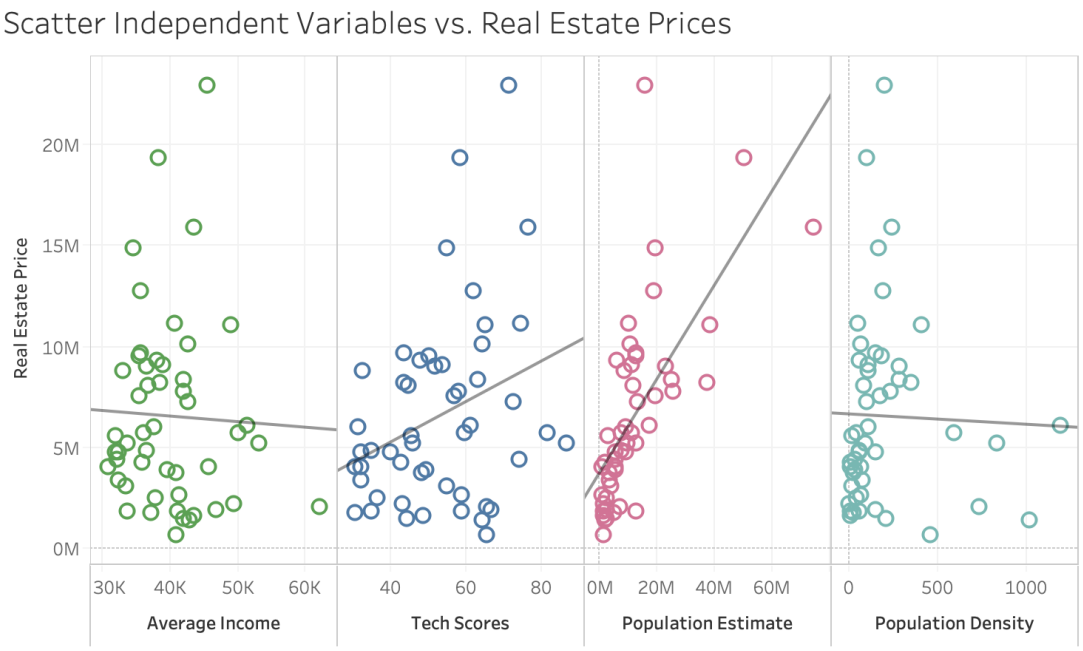

Figure 4. Average Income, Tech Scores, Population Estimate, and Population Density vs. Real Estate Price.

If you refer to Figure 4, the population estimate's linear regression was the steepest compared to the others, representing the stronger relationship between population estimate and real estate prices than with real estate price and the other independent variables. The population estimate and real estate price scatter plot also shows an evident pattern as the data are tightly packed along the trend line. The visualization of it corroborates what was seen in the regression analysis as the population estimate had the smallest p-value. Technology scores have a less steep linear regression than population estimate but steeper than average income and population density, which again supports the regression analysis. The scatter plot for average income and real estate prices is widely distributed with no real pattern, which is also why the trend line falls flatter than the other trend lines. This visualization supports that the average income is not statistically significant in the model.

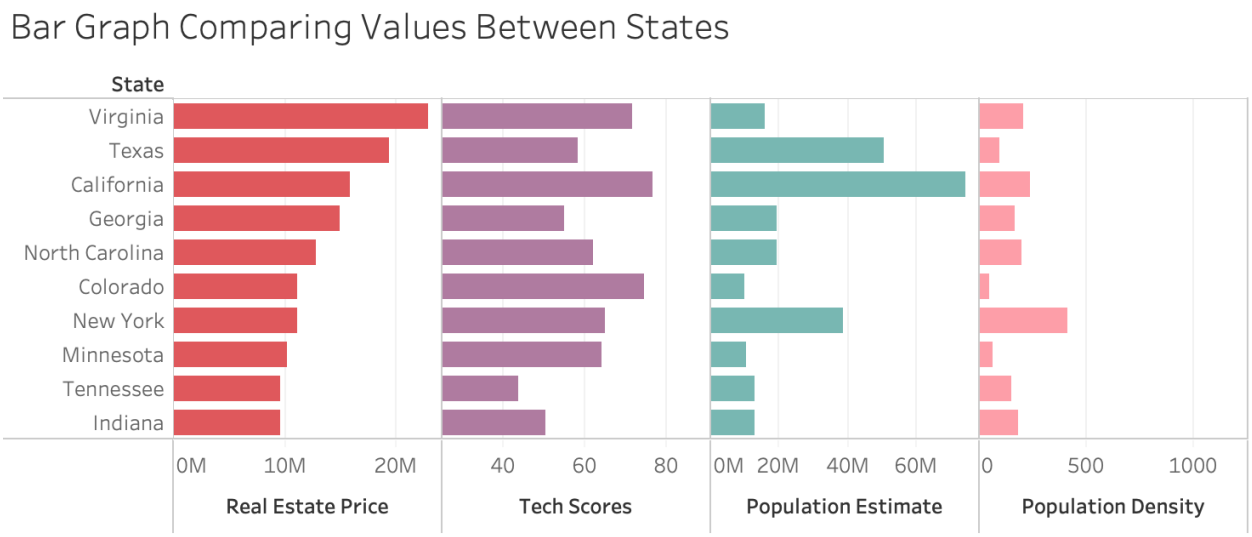

Figure 5. Real Estate Price, Tech Scores, Population Estimate, and Population Density for States with the highest real estate prices.

It can be easily seen from Figure 5, the bar graph, that Virginia, Texas, and California has the highest real estate prices compared to the rest of the states. It is interesting to see that despite California having a larger technology score, population estimate, and population density, it has lower real estate prices. The real estate price data only including the 35-65 percentile of housing prices may also play a part in why Virginia and Texas have higher real estate prices than states such as California with high values in each category of technology and population. It could also mean that there are other factors other than these three that have an effect on the resulting real estate price. This visualization allows for the conclusion and assumption that there are other factors that impact real estate prices than just technology score, population estimate, and population density.

Conclusion

Through the research conducted, it can be seen that there are factors that have

an impact on the real estate prices in the United States. With the data gathered and analysis conducted using it, the factors that have a significant impact on formation of the prices are the population estimate and population density, and the technological improvement of the region. Keeping in mind that the real estate prices are of the 35-65% home prices in the United States, these factors are statistically significant in predicting the mid-tier home prices in the country.

As a result of this research, it can be concluded that technological improvement and the population of the region are useful factors to look at when deciding where to buy real estate. After running the regression analysis, it appeared to be that average income is not a significant variable to use. Also, both the average income and the population density data have a negative correlation with the real estate prices, which contradicts the research from other studies that were cited earlier in the introduction. The data visualizations of the bar graph comparing values between states also shows that there are other factors other than population and technological improvement that impact real estate prices. So, when investing into real estate, the best things to research is probably the birth and immigration rates (which would influence the population), and the technological improvement of the region, which can be evaluated through the amount of technological institutes and the IPOs launched.

Bibliography

Bureau of Economic Analysis. (2010). *Regional Economic Accounts GDP and Personal Income* [Dataset]. Bureau of Economic Analysis. https://apps.bea.gov/iTable/drilldown.cfm?reqid=70&stepnum=40&Major_Area=4&State=00000&Area=XX&TableId=20&Statistic=3&Year=2018,2017,2016,2015,2014,2013,2012,2011,2010&YearBegin=-1&Year_End=-1&Unit_Of_Measure=Levels&Rank=0&Drill=1.

Dowell, E. (2019). *The Impact of the Tech Boom on Housing*. The United States Census Bureau. Retrieved 14 December 2020, from https://www.census.gov/library/stories/2019/04/impact-of-tech-boom-on-housing.html.

Follain, J. (2010). *A Study of Real Estate Markets in Declining Cities* [Ebook]. Research Institute for Housing America. Retrieved 14 December 2020, from https://www.mba.org/assets/Documents/Research/RIHA/75154_10296_Research_RIHA_ShrinkingCities_Report.pdf.

Keough, K., Klowden, K., & Barrett, J. (2014). *2014 State Tech and Science Index* [Dataset]. Milken Institute. https://milkeninstitute.org/sites/default/files/reports-pdf/StateTechScienceReport4Web.pdf.

Pettinger, T. (2019). *Factors that affect the housing market*. Economics Help. Retrieved 14 December 2020, from https://www.economicshelp.org/blog/377/housing/factors-that-affect-the-housing-market/.

US Census Bureau. (2010a, June 22). *County Population Totals: 2010-2019* [Dataset]. United

States Census Bureau. https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html#par_textimage_242301767.

US Census Bureau. (2010b, August 26). *2010 Census: Population Density Data (Text Version)* [Dataset]. US Census Bureau. https://www.census.gov/data/tables/2010/dec/density-data-text.html.

Zillow. (1996–2020, August 20). *Housing Data* [Dataset]. Zillow. https://www.zillow.com/research/data/.