Connor Lockman
22 April 2019
Si 618 – Final Project

Exploring Housing Prices: Analyzing the Changes in Real-Estate Cost

Motivation

This project serves as a jumping off point. It is the foundation for understanding not only what is happening in the American real-estate market today, but also what will happen tomorrow. Housing prices can tell us a tremendous amount about the world around us and in this project, I aim to lay a foundation for a baseline understanding of America's Housing Market from 1996 to 2017. To accomplish this task, I set out to answer four(ish) questions.

1. Which city was the most, and which was the least, expensive place to live in United States in the year 2017?
    a. States' housing prices vary greatly, for each state explore which city was the most and least affordable?
2. People care deeply about volatility in housing prices, determine which cities in the US experience the highest overall increase in price over the last year, five years, and ten year (starting in 2017).
3. Ann Arbor is a place of interest for me. Have prices for the same size homes changed at similar rates over the course of time? If not are some sizes more expensive or more affordable than other sizes?
4. Predict which cities are likely to experience an increase in prices in the coming years, where should you invest your money?

Data Source

The data I worked with comes from Zillow and was found on Kaggle at this link https://www.kaggle.com/zillow/zecon. The data is available in csv format when downloaded. The zip file contains multiple csv files (9) that each focus on different levels of information. For instance, one csv focuses on city level data while another focuses on state level. These were the two datasets that I used primarily. With that being said, each csv has a different number of rows, but the columns are relatively close to the same number. The state csv features 13,213 rows while the city one features 81 columns and 3,762,566 rows. The data covers the years 1996 – 2017 within the United States. The actual columns within the data consist of either date or housing price which is broken into different units of measure. I worked primarily with ZHVI_MiddleTier housing price as it is Zillow's smoothed measure for a median house in the region. I attempt to answer questions that are useful for the most people and thus I chose this unit vs. the bottom and top tier measures which were also available. When I wanted a more inclusive measure, I leveraged the price for square footage. Finally, I used specific pricing for different house sizes with the question regarding Ann Arbor prices.

Methods

Question 1: Most and Least Expensive Places to Live

A) Manipulation for this question regarded parsing the data frame down to the year 2017. I also used .loc (which was my technique for fixing on 2017) to also fix the data on June. I did this to make sure that there were not twelve different instances for each city.
B) I dropped all data that was null. It seemed to be the most reasonable approach to this question as it does not interfere with the other data.

C) The workflow follows a similar approach throughout the project. I have data frames at the top which serve as the basis of how I answer each question. In this question, I iterate over this generic data frame to harness the data needed to answer the question. This process included a sorting step on ZHVI_MiddleTier in both ascending and descending order after I controlled for instances only in June of 2017.
D) A major issue was that I was getting results for twelve months for each city. To address this, I decided to use only data for June as it is typically one of the most expensive times to purchase a home. I figured this was a good approach in order to yield results on the high end rather than low and because this is a preliminary exploration, I'm interested more in trends than minute accuracy.

Question 1a: Find the Most and Least Expensive City in Each State

A) The manipulation for this question followed many of the steps laid out in question one. I still used June of 2017 as my way to control the number of instances that existed within the data. I created a separate data frame so to not interfere with question one's results.
B) I once again dropped NA, because it would not provide any suitable information in helping answer this question.
C) I leveraged the power of groupby to sort the database by the max and min values. This resulted in producing both the highest and lowest price of ZHVIPerSqft_AllHomes. I switched to using this measure because it now includes all Homes, not just the median. I figured it would be helpful to use different metrics to see if results are similar.
D) Interestingly, Atherton California is still the most expensive in this Data which actually surprised me a bit. Other than that, the process went fairly smoothly in generating the list for this information.

Question 2: Change in Housing Prices

A) I followed similar steps with the first two questions above. I had to however create multiple data frames controlled for 2017, 2016, 2012, and 2007.
B) I dropped NA in this example and used the ZHVIPerSqft_AllHomes as my measure because it seemed to have less instances of null values throughout the years in question.
C) I broke this question into multiple steps revolving around the use of years. For each timespan in question, I performed the same control for the two years in question and fixed the data on June. I used pd.merge to push the two year's datasets into a single data frame using RegionName as the index. I then created a Dif column in the dataset through subtracting the more recent year (2017) and the previous year in question (2016,2012, and 2007) and inputting it in this new column. I then could sort that column using previously explained methods to show the largest differences in increase and decrease in price.
D) Something strange was happening with this data while I was attempting to produce these results. Records were duplicating and I realized that using City + State as the index was not a best practice for merging the two. I was getting duplicate results from this process. To curve this, I ended up using region name as it is the Zillow unique identifier for a city and that cleared up my problems.

Question 3: Change in Pricing of Ann Arbor Houses

A) This question involved a slightly different set of tasks compared with the ones above. For this question I had to bring the different sizes of houses into my data frame. The first thing I did was build this data frame with the 1 bedroom – 5+ bedroom measures.
B) I then dropped NA because it felt that it would not further answering this question. I did not fix the month to June because I wanted to take a more wholistic approach to answering the difference across years.
C) I controlled using loc onto the years in question and took the median for each size of bedroom (1, 2, 3, 4, 5 or more). I did this iteratively over a set of years spanning back from 2017 to 2002.
D) I thought this question would prove the most difficult, but I just used a little creative problem solving. While iterating over the years in question I was made a table in excel and plotted it to see how the changes were impacting the general trends and ultimately decided to use this table to help communicate the story.

Question 4: Predicting Where to Invest Your Money

A) In order to answer this question, I decided to leverage the process undertaken in question two. I used the dif column as a marker to indicate whether you'd want to invest in a property. I had to iterate over the years in question (2015 to 2017) to build a difference column that showed the price change.
B) I still dropped NA in this example in order to make the data clean and because I needed to as a step in the Naïve Bayes process.
C) I created another column in the data frame that was a binary label based on the dif column. If the price difference was equal to or greater than $100 and equal to or less than $500, I labeled it "yes" as a marker that one should invest. Everything else was labeled "no" as a marker to not invest. I then created test and train datasets and ran the train data in a Naïve Bayes model based on the housing price from the year prior to the increase which was 2015. I then ran the test data and measured the accuracy. Finally, I switched to use the price for housing in 2017 to try and predict where would be a good place to invest for the coming years.
D) This was a fairly arbitrary label as far as good or bad to invest in a property is concerned. I don't have much knowledge of the real-estate market. While this measure predicts where is likely for prices to increase, it may not represent the optimal place to invest your money. This could be improved on in the future.

Analysis and Results

Question 1: Most and Least Expensive Places to Live

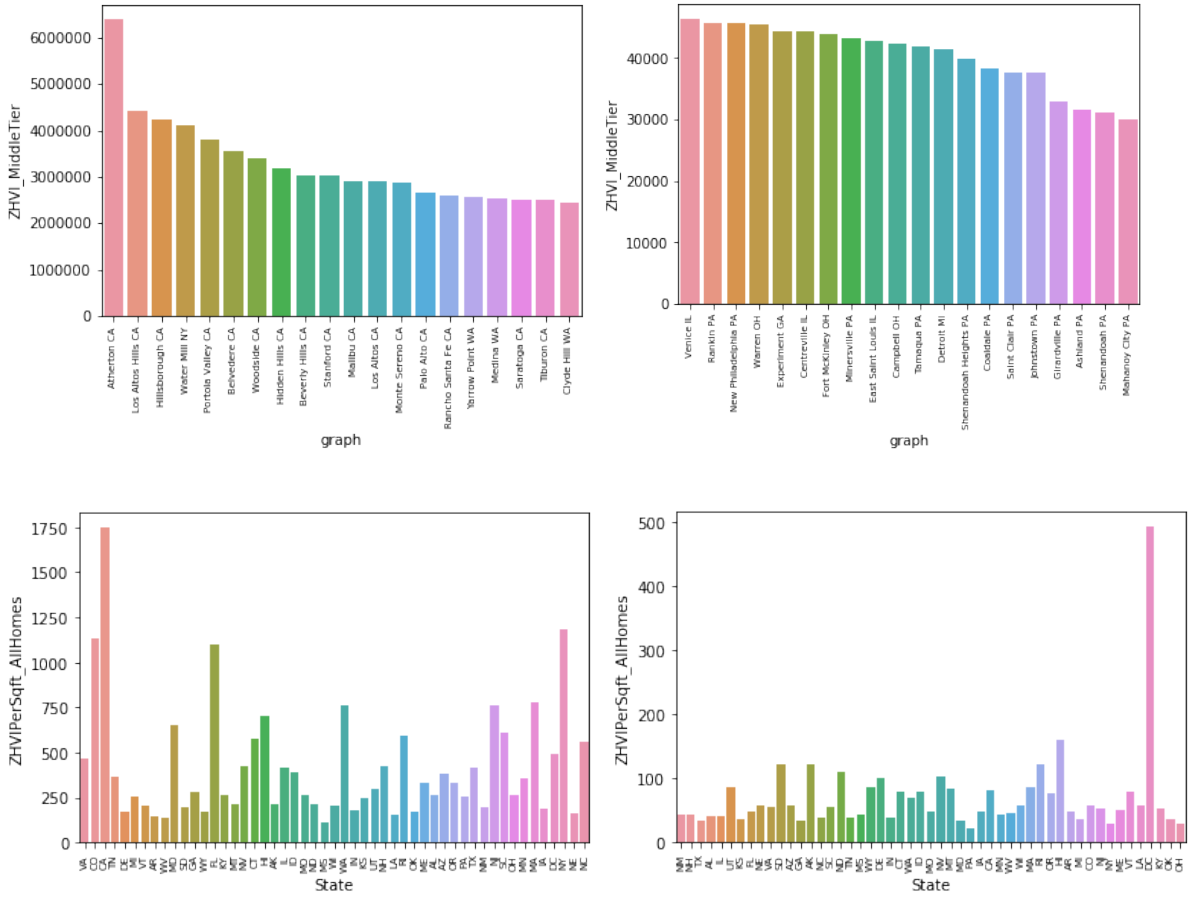
A) This question shed light on the pricing context of different states. When I sorted ZHVI_MiddleTier by the largest values (representing the most expensive median homes) one state rose to the top. California dominates the list with Atherton, a city near San Francisco, having a price of $6,399,100. The San Francisco area continued to dominate the top of the list and the only states outside of California to make an appearance were New York and Washington. When we plot this, we can see that Atherton is even an anomaly among the top of the list.

   On the Inverse, when we follow the same steps to plot the cheapest cities, we see a little more diversity in the list. The cheapest city is Mahanoy City in Pennsylvania with a price of $30,000. Pennsylvania has by far the most cities on this graph with twelve.

Other notable cities on the list include Detroit, Michigan with a price of $41,400. What these graphs truly make clear is the amount of housing price inequality that exists within the United States.

I also plotted the most expensive and least expensive city in each state to give context as to what the real estate market looks like across the country.

B) I chose to plot these in bar plot format because I felt that it was easier to convey the information opposed to my original proposal of a scatterplot.

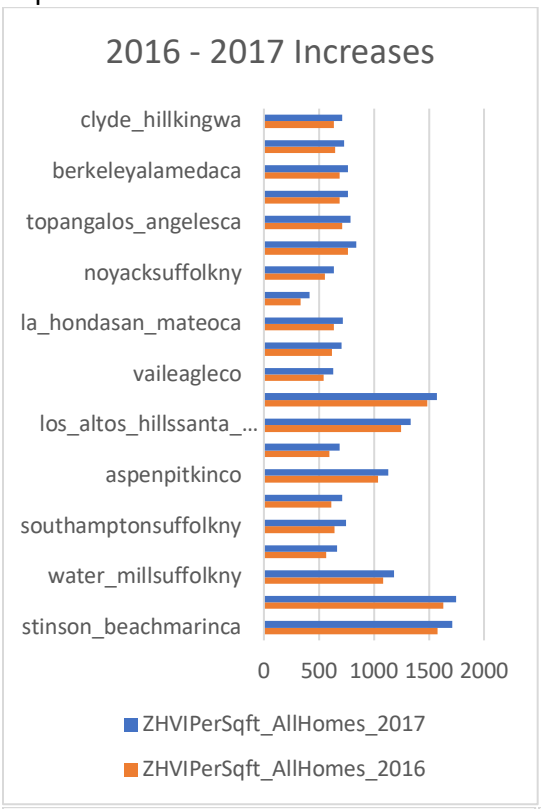


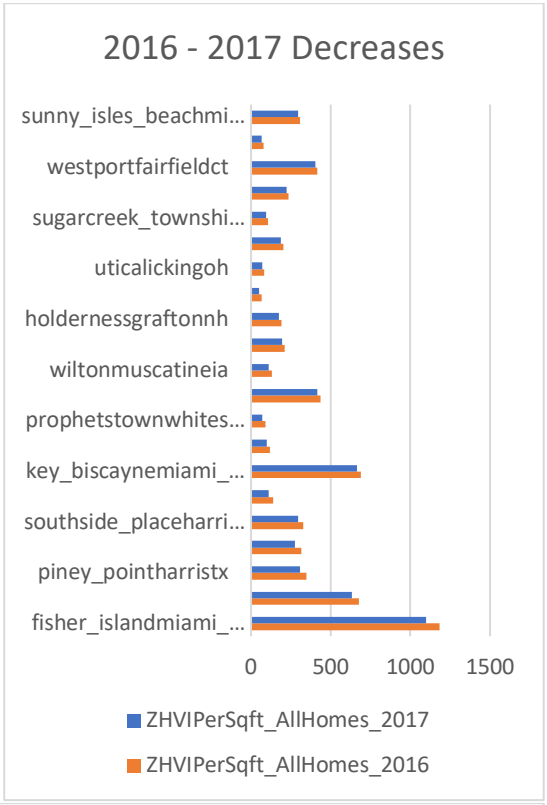
Question 2: Change in Housing Prices

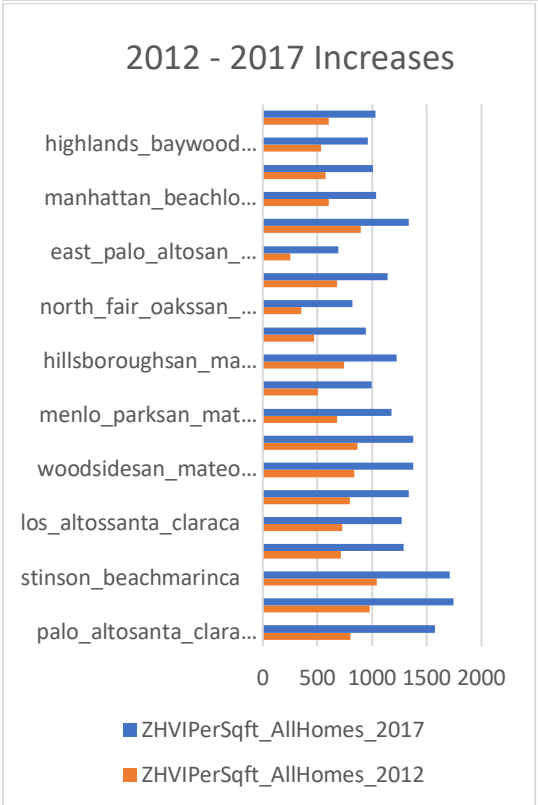
A) These results are broken into price differences for 2016 – 2017, 2012 – 2017, and 2007 - 2017. The city with the greatest increase in price for 2016 – 2017 was Stinson Beach, California while the largest decrease occurred in Fisher Island, Florida. In the time span between 2012 – 2017, Palo Alto, California rose to the top of the biggest increases in price while Roosevelt, Rhode Island fell to the place of largest decrease. Finally, between 2007 and 2017 Palo Alto, California retained the position of largest increase and Kirkwood, California took over the place of greatest decrease. These seem to make logical sense as Palo Alto has experienced significant growth over the last two decades with the emergence of the tech industry and Silicon Valley. Stinson Beach is just outside of San Francisco which shows us that the region is now, as a whole experiencing an

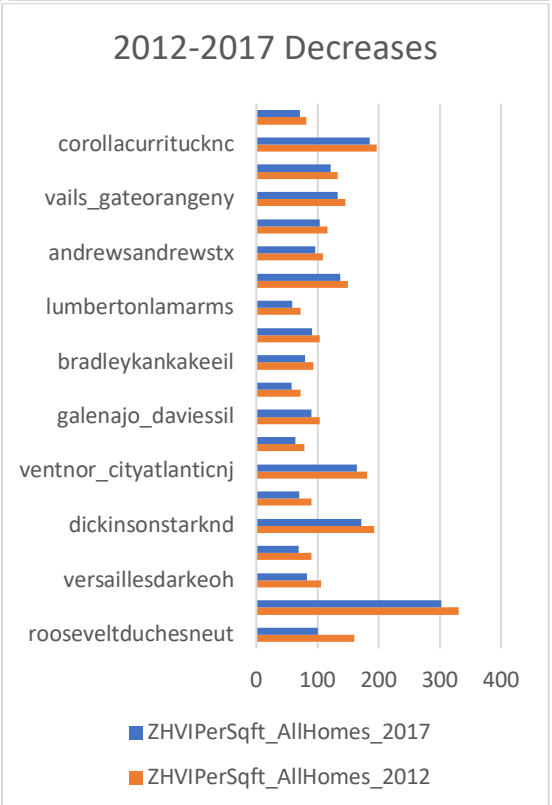
increase in prices. It's tougher to determine some of the reasoning behind the largest decreases and further investigation is needed to substantiate the reasons behind the fall in prices.

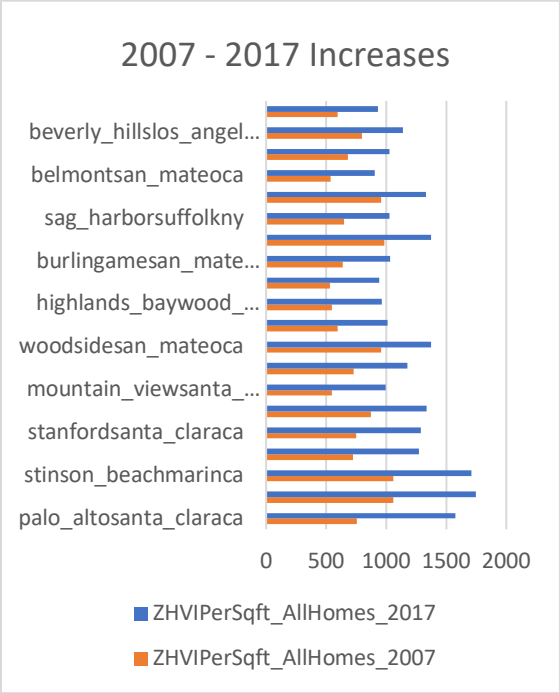
B)

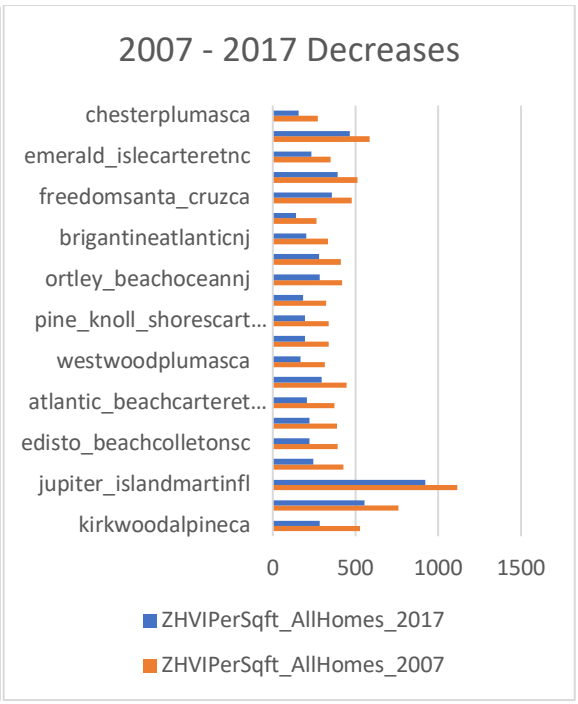








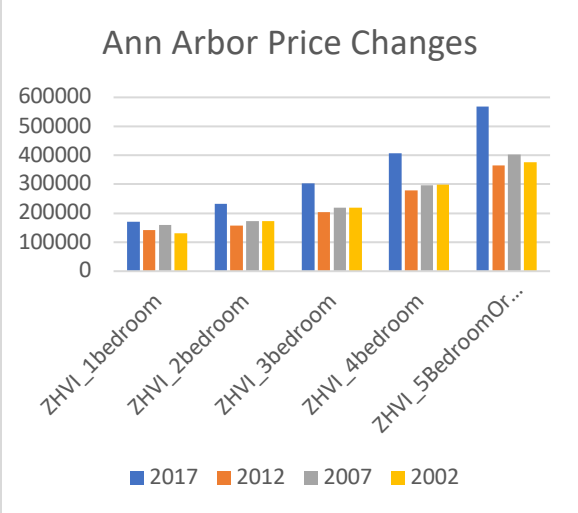


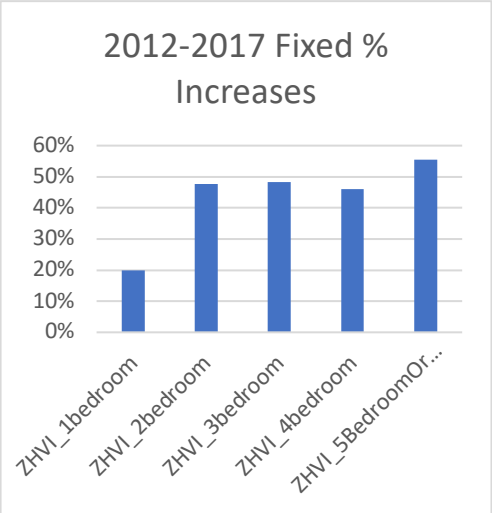


Question 3: Change in Pricing of Ann Arbor Houses

A) As outlined above, this table represents the change in median houses in each subset over the course of 15 years. The size of homes is grouped together, and color denotes the price of the house in a given year. We can see that 2017 proved to be the most expensive year for housing prices over the course of this data but the change in a one-bedroom house has not spiked the way a five or more-bedroom house has. It shows us that if you can buy a bigger house, it may make sense as the price is likely to continue to grow. We can also see that prices fell from 2007 to 2012 a sign of the recession that had occurred over that time period as well. I then normalized these prices because naturally an increase in a five-bedroom house is going to be larger than that of a one bedroom. We see that a one bedroom has in fact increased the least amount since 2012 and if you are trying to save money, might make the most financial sense to purchase.

B)





Question 4: Predicting Where to Invest Your Money

A) As outlined in the section above, I built a Naïve Bayes classifier for measuring where one should invest their money. I ran the train data on Square feet of the homes for 2015 and created a fairly arbitrary label that indicated if the increase was between $100 and $500 that the increase indicated that it would be a good place to invest. When I ran the model on the test data, I received an accuracy score of 95%. Satisfied with this performance, I ran another test leveraging the square feet of homes in 2017. If it predicted a label as yes, I considered it a good place to invest and pursued plotting it.

B) This map contains the cities that I plotted representing good places to purchase property as an investment in the year 2017. We can see California emerge as a good place, especially the Silicon Valley Region as neighboring cities experience increases in housing prices due to growth.