Midterm Report: Where and When to Buy a House?

Group member: Adrianne Fu, Grant Liu, Linyi Lyu 10/31/2021

1. Problem Statement

Our project is to provide useful information on the best county/regional investment in the U.S. to various stakeholders like individual buyers. The primary focuses are to analyze the important influencing factors of real estate prices and to predict the real estate price of a county in the next few months or years.

2. Data Overview

This part contains a detailed introduction of datasets since we collected data from various sources and combined them. The house price data we used is from Zillow Housing Data. It includes the region information (state, county) and Zillow Home Value Index (ZHVI) in each month. ZHVI reflects a smoothed, seasonally adjusted measure of the typical home value within the 35th to 65th percentile in a given region.

The unemployment data includes the unemployment rate in different regions in each month. The hospital data lists all registered hospitals in a given region together with the rating for each hospital. Similar to the hospital data, the school data contains information of all the schools in each region. We also collected the consumer price index (CPI) and interest rate data since they reflect the national economics level. The income data includes the average income and income inequality score. The education background data shows the percentage of residents with a bachelor's degree or higher and a high school degree or higher. The population data includes the estimated migration and total residents size.

All variables generated from above-mentioned datasets are categorized as follows:

- Variables constant in all months for a given county: hospital count, hospital rating, private schools count, public schools count.
- Variables constant in all counties for a given year: CPI, interest rate.
- Variables with different values in different months and different counties: ZHVI, unemployment.
- Variables with different values in different years and different counties: personal income, income inequality[1], percentage of population with Bachelor's Degree or higher, percentage of population with high school graduation or higher, net migration rate[2], resident population[3].

[1] Income inequality is the ratio of the mean income for the highest quintile (top 20 percent) of earners divided by that for the lowest quintile (bottom 20 percent).

[2] Net migration rate is data collected by the American Community Survey (ACS) using a series of monthly samples to produce estimates over a 60-month period.

[3] Resident population is estimates updated annually using current data on births, deaths, and migration to calculate population change since the most recent decennial census.

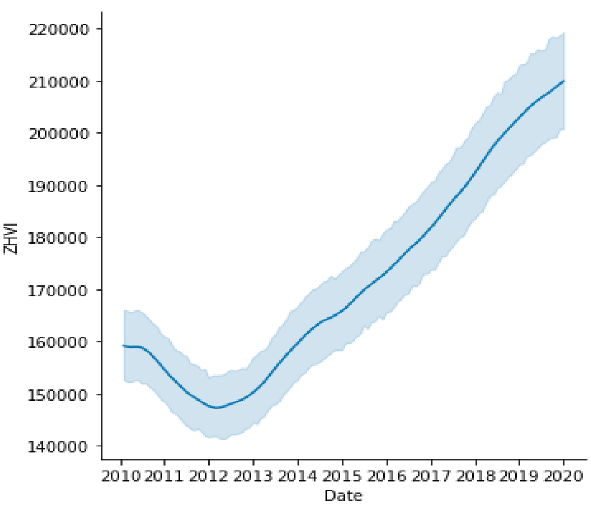
3. Data Processing

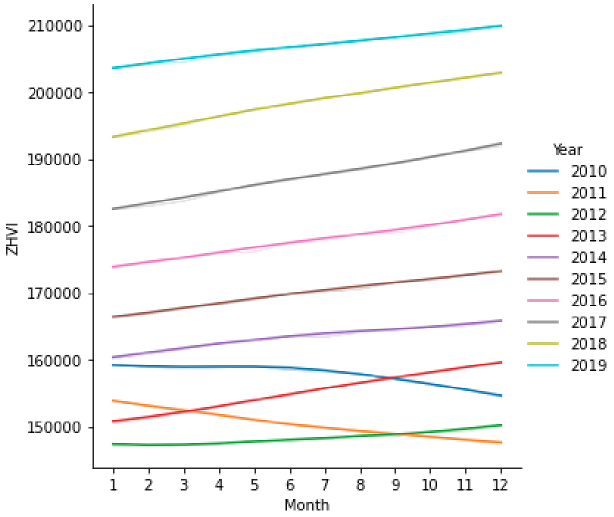
There are a total of more than 280 thousand observations in original house price data available on zillow, including regions that are not covered by American Community Survey (ACS). Besides, some variables such as personal income are missing federal census records of 2020 possibly due to the pandemic.

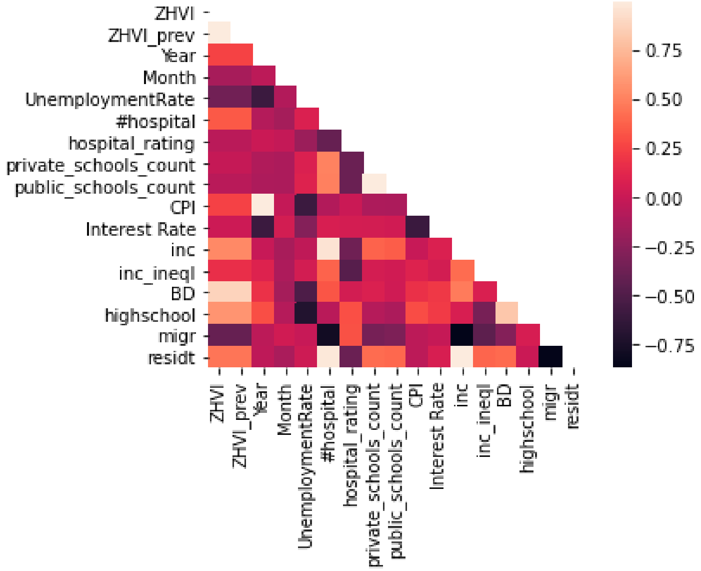
Thus, we only retained regions and times with valid ACS reports -- a total of 93835 rows of per county per month house price data from Jan. 2010 to Dec. 2019 of 781 distinct counties across the United States, along with according variables listed above.

4. Exploratory Data Analysis

The plot below shows the trend of consecutive monthly average house prices across all counties from 2010 to 2020. House price decreased from 2010 to approximately the first quarter of 2012, and then continued to rise until 2020. Except in the years where house prices decreased, the house price increments constantly during the year, and we did not observe obvious seasonal fluctuations.





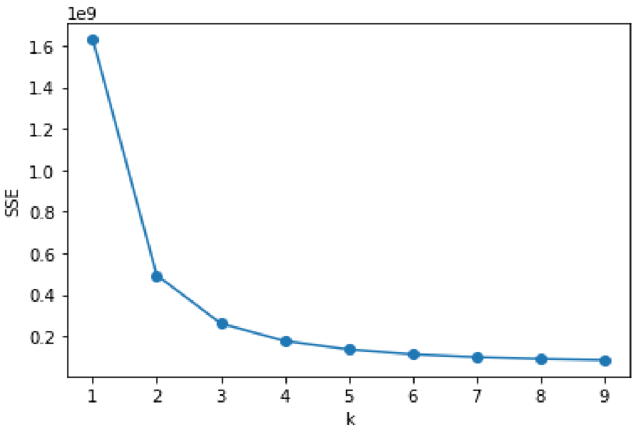


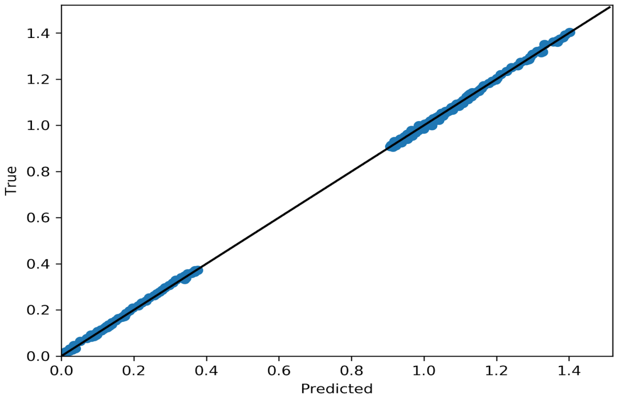
The heat map on the left shows correlations between each pair of numeric variables in our dataset. We can see that some features like ZHVI from the previous month, proportion of population with higher education degree, income, the number of hospitals, and resident population has a positive correlation with ZHVI, while unemployment rate and migration population are very negatively correlated.

5. Preliminary models

We standardized variables because of their very large distribution intervals, for example, the net migration rate spans from -87102 to 44753 (thousands of persons). For features with nominal values, for example, county and state, we converted those features into integer values using one-hot encoding to enable regression model fitting.

All observations could be split into three clusters according to the characteristics of each county. We will attempt to fit models on each cluster respectively which might be beneficial to avoid underfitting. The lower left diagram shows the result of KMeans classifying all the 781 counties based on 13 features including geographical information model.





We first attempted to fit a linear regression model to the dataset. We split the dataset to a training set (80%) and test set (20%) (We are still deciding whether to use random shuffling for train-test split or leave out the data from the last two years for testing). An offset is added to the feature data matrix. Above on the right is the plot of true value against our predictions. Since the dots almost perfectly lying along the line, plus the MSE for both training (9.11e-05) and testing dataset (7.69e-05) are very small, it is suspected that the house value from the previous month included in our feature matrix acts like a cheating column, therefore making weights of o less influential.

6. Validation

To avoid overfitting for the linear model. We will consider penalized regression such as ridge or lasso. In the next steps, we might fit regression tree models on the data. In order to optimize the parameters, we will use k-fold cross validation to tune the parameters. To avoid underfitting, we might include more possibly relevant features that we have not collected such as market demand, supply, and crime rate in each region etc. Another concern is that some features do not show a linear relationship with the house value. Thus we might need to scale (eg. log scale) some features or the response variable for the linear models.

To estimate the effectiveness of our model, we plan to leave out the data from the year 2018 to 2019 or just random 20% of the data for test use. Since the response variable is numeric, we will compute and assess the MSE of the prediction result in the test dataset.

7. Moving forward

Our next step is to modify linear regression models, examples include using lasso regularizer to generate more interpretable coefficients, and try out cross validation methods to enhance generalizability. Given that it’s not possible to collect some of the variables of 2020, we will try to predict house price of 2021 using predictions of 2020, however, the model's accuracy may be compromised as additional variables might be needed in 2020 to include the impact of the pandemic. To answer the main question: where and when to buy a house, we plan to use the house price prediction to calculate the growth rate in a certain future period of time in each region and figure out which place has the largest investment potential.