Where and When to Buy a House?

Group member: Adrianne Fu, Grant Liu, Linyi Lyu Date:10/3/2021

Background and Objectives:

Both individuals and institutions hope to invest in the real estate market with the best interest rate and minimal risk, and hope that it will be the best investment for the long term. Housing prices have maintained an increase of around 5% for many years, however, during the epidemic, there have been double-digit price increases, and the current housing supply is at its lowest level since the 1970s. These phenomena arise increasing concerns about the ability to buy houses while decision makers are being inundated with extremely messy data.

Thus, the purpose of this project is to provide useful information on the best county/regional investment in the U.S. to various stakeholders like individual buyers, national real estate developers, and financial institutions.

In this project, we will focus on two challenges: a) analyze the important influencing factors of real estate prices to answer the question “Where and When to Buy a House”; b) predict the real estate price of a county in the next few months or years. To achieve these two goals, we did some research and categorized the factors that might have relevance to the topic, and plan to take into account the following data sets:

Datasets:

a) Market Aspect: house monthly prices (response variable), market supply and demand.
The house price dataset contains monthly house prices (from Jan 2000 to Aug 2021) for 2822 counties in the US. Prices reflect the typical value for homes in the 35th to 65th percentile range.
b) Social Aspect: education, crime rate, schools, hospitals.
c) Finance Aspect: income level and distribution, business pattern. (United States Census)
d) Economic Aspect: price of commodities(CPI, GNP, etc.), employment data, GDP, international balance of payment.

Approaches:

We will focus on large cities only due to data availability. Because the housing price data to be predicted is continuous, we consider using a linear regression model to predict. Some of the features currently considered may have overlaps, such as income and payroll, so the most effective features need to be selected based on specific data. The problem that may also be encountered is that the data of year 2020 may have large deviations due to the impact of the pandemic, so pandemic-related features may need to be considered to adjust the model.