**Introduction**

The Florida Housing Project will be using information collected on housing in Florida in order to predict housing costs. Given these inputs, we should be able to tell whether a house is a good deal or not based on the predicted price. I will be the only one working on this project.

There are many different styles of house available, and while I would like to focus on any single-family domiciles (while ignoring dorms, student housing, etc.) it may eventually fall to us having to focus on a single style of home. I’m hoping that we can keep all types in, however.

**Background**

Ultimately, the final sale price of a home is based on what the market will bear. That being said, there are a number of tangible, measurable indicators that can have the effect on the price of a home. While some intangibles do exist, generally similar homes within the same region will sell for a similar price range. Set a price too high, and the house will sit on the market for months.

We do have some value indicators available to us in our data set. Mortgage payments and appraised value for property taxes stand out.

**Problem Statement**

Through analysis of publicly available housing records in the state of Florida, the goal is to narrow down fields that have the most effect on market price. Once the 230+ fields of data have been pared down to a more manageable set, the hope is to build a linear model that can predict the price of a home and, given that value, indicate whether a home is priced as expected, priced too low, or priced too high.

**Scope**

While data is available for all 50 states, the District of Columbia, and Puerto Rico, the variance between the different regions due to cost of living, availability of housing, etc. can result in wildly different values. In order to minimize this regional variation, we will be focusing only on data from the state of Florida. Data is available over both a five-year span and a single year span, and we will be focusing on a single year, the most recent data available.

Although information is provided that can be used to compare costs between multiple years, we will only be focusing on a single year’s worth of data.

**Preliminary Requirement**

**Technical Approach**

Our data will be read in the CSV format using Python. From there, we will transform and clean up the data. This outputted data will then be read into R which is what the rest of the analysis will be conducted in. We will split the data into a training set and a test set, using our model to predict the value for the test set.

As we currently have nearly 100,000 observations in our data set, we shouldn’t need fear not having enough data for analysis. However, certain programs are better suited to larger databases than others, so if we begin to run into problems, we may need to adjust our approach (or cut down our data).

**Data Sources**

The primary requirement for data for this project will be the housing data. We will be using data provided by the US Census. The ACS Public Use Microdata Sample files are a sample of the actual responses to the American Community Survey and include most population and housing characteristics. This information is freely provided in CSV format by the Census Bureau, and we will be using the 2018 housing data for the state of Florida.

**Analysis**

Once our data has been gathered and entered, the goal is to create a method that tests different variables and combinations of variables to determine the ones that have the greatest effect on the final value of homes. Ideally, we would train this model on our training set and test it on our test set in order to gauge performance.

**Requirement Development**

The data has already been collected but will need to be sliced. Since rent payments don’t technically map to the purchase value of the house (the landlord needs to make a profit and cover the cost of repairs, admin costs, etc.) we will focus on homes that are owned, not rented. We will need to identify these homes to focus our dataset a little bit more. We may do some preliminary culling of columns here – for instance, we don’t necessarily need the occupation of the occupants, or their annual income, as that may provide clues that would make our model less efficient.

Fortunately, the dataset provided has had a lot of this work done for us. The information is provided annually and is derived from the American Community Survey, which collects data each year from January to December. The Census Bureau does a great job of cleaning up the data and providing a summary document, so most of that work has been done for us.

**Model Deployment**

Several models will be built, with each model emphasizing different variables. Each home will have a different predicted value, and we’ll then be able to classify a home as a good buy, fair buy, or poor buy.

**Testing and Evaluation**

Each one will have their accuracy evaluated. Since it is unlikely that our model will predict the price exactly, we will be more focused on identifying the homes at the extremes – the most undervalued for their features, the most overvalued, etc. For phase 2 we may pull data from actual listings to see how our model would do.

**Expected Results**

Purchasing a home, and the price the home was purchased at, can vary wildly. My one concern is that certain intangibles may influence the price of a home outside the physical indicators – number of bathrooms, number of bedrooms, etc. Ideally, I’d love the model to be able to give a range of prices that are in the ballpark and be a decent indicator of whether a house is underpriced, overpriced, or a fair price given its physical characteristics. There may be some worth to a model that can quickly classify a purchase price as realistic or not, which would give a better idea of whether a house is worth pursuing or not.

**Management Approach**

As we are early in the project life, our direction may change as time goes on but here are what we are currently looking at. I have identified the dataset that I will be working with and over the next several weeks will be importing, refining, and getting to know our dataset. As we currently have 240 variables in our dataset, I will need to narrow that down for as much as I’d like to include everything the processor time investment could scale exponentially as our model gets more complex.

Currently, the site (American Fact Finder) on the Census Bureau’s web site will be shutting down on March 31st, 2020, but I have already saved my data and located it on the Census Bureau’s FTP repository. If need be, I can access that data later.

In order to evaluate our valuation analysis, we’ll need to determine thresholds at which a home is considered over, under, and just right valued. Since we’re dealing with large purchase prices, it may make more sense to set a percentage value, rather than a dollar amount. A $4 Million home that is undervalued by $13K is not nearly as enticing as a $110K home undervalued by $13K. As such, given the wide variety of properties we will have to set these thresholds which may change as we evaluate.

As we refine our models, we will choose the best performing ones to move on to the next round of testing while discarding the less efficient ones. Ideally, after several iterations we will have a more efficient model that is more accurate.

I’ll be maintaining the code in a DropBox which will allow me to work from any one of the devices I currently utilize. That way, I’ll always have access to the model even if my desktop computer is currently processing it.

**Project Risk**

There are many factors that can cause a house’s value to plummet, which wouldn’t necessarily be indicated in our data. If our dataset contains a number of houses that were foreclosures, which isn’t indicated in the data, that will indicate a lower valued house without being represented in the data.

With the current world situation, it’s likely that we could see some movement in house prices. Given that we are using 2018 data, a seismic shift like the one we are currently experiencing could very well prove difficult for our model to predict.

Care must be taken in the calculation of the value of the house. Mortgage payments consider the creditworthiness of the applicant, and so simply multiplying the mortgage payment by 360 doesn’t give us the whole picture. We will have to take care to value things correctly.

With luck, however, we’ll be able to come up with a model that will help price and evaluate homes in Florida quickly and efficiently.