**Title of your project proposal:**

Predicting America’s Future Popular Neighborhoods

**Team Name:**

Data$cientists

**Github 1:**

austinwu0

**Github 2:**

elizahale

**Github 3:**

heiscody

**Github 4:**

azhong1

**Background and Motivation:**

One of our team members has an interest in urban planning and currently works at a startup that provides a platform for cities to promote redevelopment of economically underutilized properties. Since we’re all seniors about to enter the real world and rent or buy homes and apartments for the first time, we’re all interested in trying to pick the “right” neighborhood in live in, and make a good financial investment in the future.

**Project Objectives:**

We hope to be able to predict what neighborhoods in America’s largest cities are slated to be the most popular in the upcoming years, and what sort of factors can be used to predict popularity. “Popularity” is a subjective measure depending on the demographic being considered, so we will do some data exploration to determine a suitable and objective metric which can be related to popularity in a broad context; for now, we plan on using some metric that combines the distribution of housing prices and rental prices.

In writing our narrative at the conclusion of our project, we will pretend to be real-estate speculators trying to cash in on the next real estate boom. Our project, if successful in being able to predict which neighborhoods are on the verge of making it big, would have significant ramifications for real estate investment, as well as businesses and individuals who want to place themselves at the center of rapid growth and economic prosperity.

**What Data?**

We will use datasets that detail city-specific information over time. Ideally we will be able to get data down to the neighborhood level; if not, we will adjust our approach accordingly. We anticipate utilizing various U.S. Census APIs in order to retrieve information on demographics, crime rates, industry employment, university enrollment, migration, investment, economic indicators etc. in different cities/counties in different years (http://www.census.gov/data/developers/data-sets.html). The U.S. Census API also offers housing prices broken down by city/county, which we will try and predict in our model. We will look how change in our predictors over the years tracks housing prices and build a model to predict how these factors relate to housing prices. Zillow, Enigma.io, and other sources might also be able to offer other useful factors (internet usage, restaurant distribution, transportation, etc.) broken down by county/city.

**Must-Have Features**

The goal is to have our model predict which neighborhoods will be the most popular. Right now our plan is to define popular by housing prices, and we may also include the amount of time houses stay on the market, and the rate of construction.

Our project must be able to take in a set of features from historical data from many different neighborhoods, and return predictions for how popular those neighborhoods will become in the next few years. Our project must be able to validate predictions by comparing them to actual data.

We would like to build a model that predicts housing prices in different cities/counties. We would like to build a model that beats the RMSE of the baseline case in which we simply take the previous years housing price for that region.

**Optional Features**

One optional feature could be that rather than predicting just numbers for how popular a neighborhood will be, we could have a model that returns distributions. Meaning that we would have a distribution for each neighborhood representing the likelihood of that neighborhood being popular or not. This might be more difficult, as we would be implementing bayesian modeling techniques.

Another additional feature could be some sort of neighborhood clustering. We could cluster neighborhoods based on different house prices, and then investigate what other features these neighborhoods had in common.

Long-term popularity predictions. Predicting other things besides price that affect how desirable a neighborhood would be to live in. These could be things such as the amount of traffic in a neighborhood, the crime rates, the average age of a neighborhood, etc.

**Design Overview**

We would love to incorporate multiple models into an ensemble method; the exact methods will be determined at a later stage depending on the data we ultimately decide is relevant. At the very least, we will be using generalized linear regression (or a log transformed housing price variable, because we expect the distribution of housing prices across the country to be right skewed) and classification models to group similar cities together (for example we might expect San Francisco to behave similarly to San Jose). We will use several variable selection criterion (R^2, AIC, BIC, etc.) to make sure that the factors that we are considering are statistically significant. It will also be important to do some time series analysis so that we might be able to remove regular trends such as inflation and nationally increasing wages from housing prices.

For now, we do not believe that it will be necessary to use AWS. However, if we find that it is easy to wrangle our data, we may add many more cities to our dataset and use AWS to speed up our calculations.

**Verification**

We will train our model on a subset of our aggregated data against house prices (the subset being about 75% of the cities which we have collected data on), then cross-validate and test on another subset (the remaining 25% of the cities we have data on).

**Visualization and Presentation**

We will present our findings in a scroll-through website. Firstly, we will visually portray the motivations behind our project. For example, we might want to graphically depict the dramatic change in San Francisco housing prices over the past few decades as a reason for us wanting to dig deeper into factors that drove this trend. This will be followed by the questions and/or hypotheses that we set out to tackle in the project. Next, we will describe our data sets and the statistical methods that we used to analyze the data, after which we show visualizations of correlations that we found as well as conclusions that we drew from these results. Finally, we will display the predictions on our test sets and how these predictions performed when validated against actual data. We will draw most attention to the motivation, results and predictions because we want the website to focus on the storytelling aspect of this piece. In our effort to uncover interesting findings on American house price trends, we hope that any type of audience will be able to read and engage with our data analysis. We might end up documenting our statistical methods and other technicalities in more detail on a separate tab linked from the main page to keep the focus on the story as told by the data. We will use a website template for a professional aesthetic and if time permits, we might use D3 and Javascript to animate our graphs in order to enhance the storytelling aspect. The video will follow a similar format.

**Schedule & Timeline**

We have four weeks from now until the project due date.

Week 1: We will focus on data cleaning exploration. We have already identified Zillow and Enigma.io and the U.S Census API as potential data sources. We will get house price data from Zillow. We hope to get data that we will use as features to drive our predictions from Enigma.io. In this first week we will gather data, plot features against house prices to look for correlations, explore the variance and distributions of the data. We will use our data exploration to drive the features we select and the prediction strategies we will choose.

Week 2: This week we will solidify how we are going to train, test, and validate our prediction models. We will also decide on which features and prediction strategies to use, and begin implementation. The goal for this week will be to build a baseline prediction model using a few features. We will start simple, and then start adding more features, feature interactions, regularizers, and model complexity once we have a working baseline model.

Week 3: Improve the baseline model by adding features, feature interactions, regularizers, etc. We also might use some more complex prediction models such as random forests, neural nets, etc. We will compare the accuracy of the models against each other, and will try building an ensemble predictor to further improve.

Week 4: Extra steps. This week will be reserved for probable spill-over from week 3 (meaning that things often take longer than we think they will!). During this week we will also work on visualizations, and do some deep thinking about how our findings can be useful, and how best to communicate them.

**Team Member Contributions**

Eliza (elizahale) Hails from the dry deserts of Utah. She will focus on feature selection and data exploration. Eliza is especially excited to explore the relationship between house prices and features that we might not expect to be correlated/related to house prices such as the presence of bowling alleys, most popular household pet, amount of time spent brushing teeth (just kidding... but if they have that data she’d take it!), etc. Austin (austinwu0), from the sunny hills of Elk Grove, California, will be our resident statistics-expert and provide some domain knowledge on real estate and city planning. His cheerful disposition and can-do attitude will surely make him a great team-member. Anna (azhong1) brings serious programming skills to the team. She comes to us from the far-off land of Australia. Anna will be leading the data visualization portion of the project, as well as helping the rest of us learn to write decent code. Cody (heiscody) the Michigander will design our training, testing, and validation strategy. He will also be deciding what kinds of models we use for our predictions. Cody has a wealth of research experience; his analytic abilities and attention to detail will make him invaluable to the team. Though we may all take point on certain parts of the project, we are all excited to contribute to every stage of the project, and our roles will likely be very overlapping.