Averi Tanlimco (averi.tanlimco@sjsu.edu)

Sania Bandekar (saniavighneshvar.bandekar@sjsu.edu)

Professor Huynh-Westfall

CS-133 Section 02

7 November 2024

CS 133 Project Proposal

Dataset Selection:

Dataset Description:

The dataset we are using is the SF Bay Area House Prices dataset from the original project template. The dataset contains information about Bay Area houses for sale in June 2019. There are 7145 records and 15 variables. The variables listed in the dataset are: 'Address', 'City', 'State', 'Zip', 'Price', 'Beds', 'Baths', 'Home size', 'Lot size', 'Latitude', 'Longitude', 'SF time', 'PA time', 'School score', and 'Commute time'. The 'State' variable is California for all the entries, since all the houses are located in the Bay Area. The 'Price' variable is specifically the listing price of the house at the time of data gathering, not the price it was actually sold for. 'Beds' and 'Baths' are numerical values that represent the number of bedrooms and bathrooms respectively. 'Home size' and 'Lot size' are the measurements of the home and lot area in square feet, respectively. 'SF time', 'PA time', and 'Commute time' represent how long it takes for you to travel from the house to San Francisco, Palo Alto, and the general Bay Area during 8:00 AM traffic, respectively. 'School score' is a measurement of the quality of the schools near the house. The Machine Learning task is to predict the listing price of Bay Area houses.

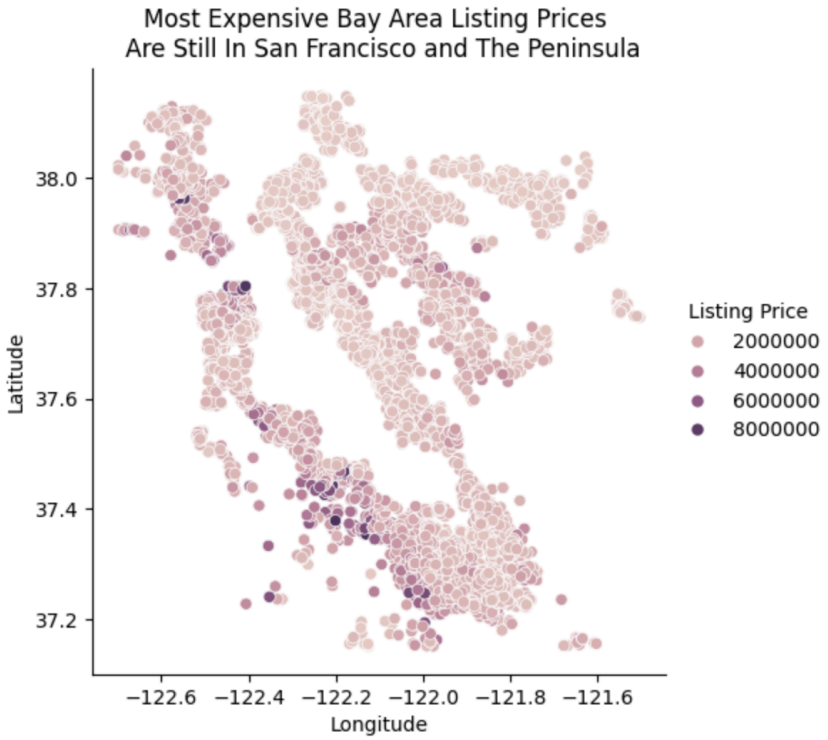
Source:

https://raw.githubusercontent.com/csbfx/cs133/main/sf_bayarea_house_prices.csv

Project Questions:

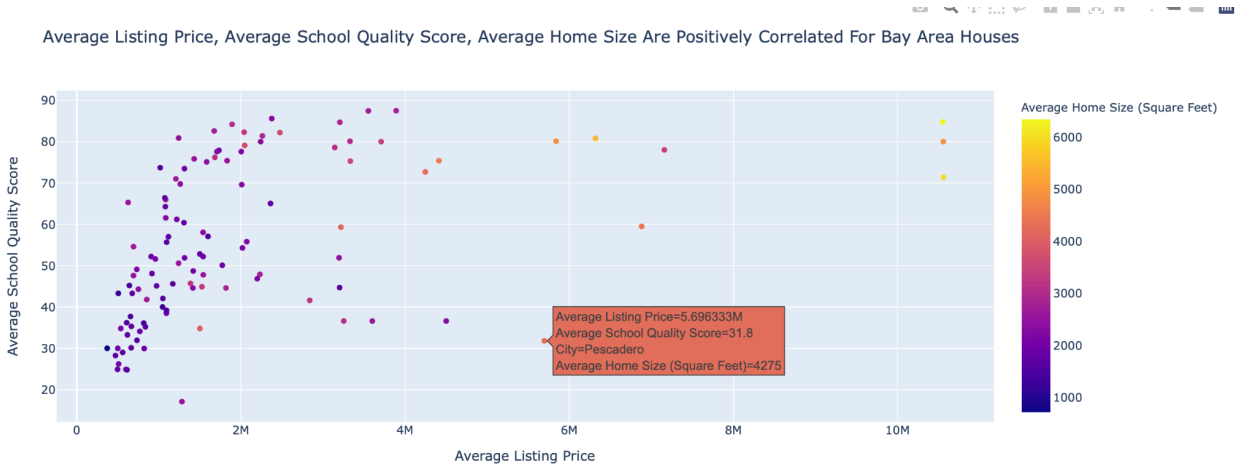
1. How is a house's coordinates related to its price?

The dataset contains the 'Latitude' and 'Longitude' coordinates for each house. Using the 'Longitude' as the x-axis and the 'Latitude' as the y-axis in a relational plot, the graph will be reminiscent of the actual Bay Area, since each dot would be in its actual coordinate location. The color of each dot would be based on the 'Price' of the house. This way, we can see which locations have more expensive houses and which locations have less expensive houses.



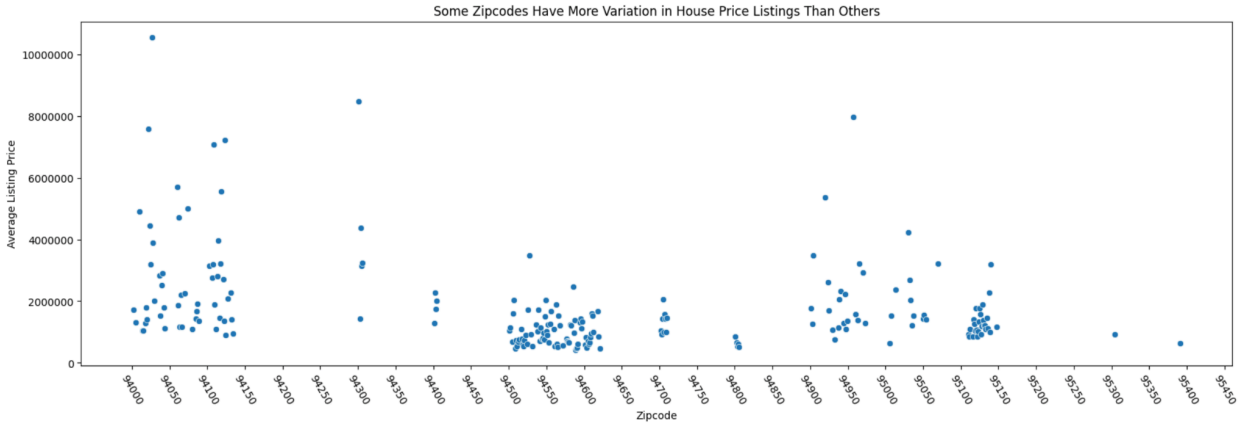
2. <u>What is the average house price, average school quality, and average house size in different cities?</u>

We would use the plotly library to create the interactive plot. We can group by ‘City’ and find the average House ‘Price’, ‘School score’, and ‘Home size’. Then, we can put these values into one dataframe together. We can then graph the dataframe as a scatterplot using plotly, with ‘Price’ as the x-axis, ‘School score’ as the y-axis, and ‘Home size’ as the color of the dots. When we hover over each dot, we would be able to see the name of the ‘City’, the average ‘Price’, average ‘School score’, and average ‘Home size’ as a text pop-up.



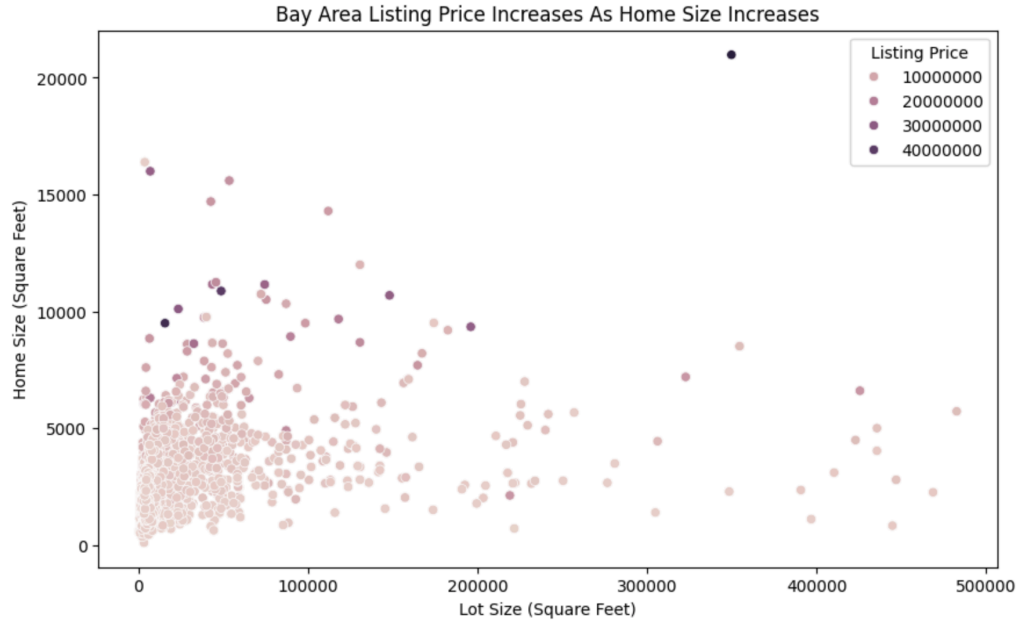
3. What is the relationship between zip codes and price?

We can group by ‘Zip’ Code and find the average House ‘Price’ for each ‘Zip’ Code. We can then set the ‘Zip’ Codes to be the x-axis and the average House ‘Price’ as the y-axis. This way, we would be able to see which ‘Zip’ Codes have the higher average House ‘Prices’ and which have the lower average ‘Prices’.



4. How do lot size and home size relate to the house price? Do bigger lot sizes and home sizes correspond to more expensive houses?

We can set the ‘Lot Size’ as the x-axis, the ‘Home Size’ as the y-axis, and the ‘Price’ as the color of the dots. This lets us see going from left to right if the ‘Price’ increases depending on ‘Lot Size’. This also lets us see going from bottom to top if the ‘Price’ increases depending on ‘Home Size’. We can also see if the ‘Home Size’ increases as the ‘Lot Size’ increases.



5. Does having more rooms (bedrooms, bathrooms) increase house price?

We can set the Number of 'Bedrooms' as the x-axis, the Number of 'Bathrooms' as the y-axis, and the 'Price' as the color of the dots. This way, we can see going from left to right if the 'Price' increases based on the Number of 'Bedrooms' and going from bottom to top if the 'Price' increases based on the Number of 'Bathrooms'. We can also see if the Number of 'Bedrooms' increases as the Number of 'Bathrooms' increases.

Proposed Methodology:

Training Dataset:

We will use the sklearn (scikit-learn) library to perform the machine learning task. The sklearn library has a method called train_test_split(), which splits the original dataset into four lists, two containing training data (for the influencing variable X and target variable Y) and two containing testing data (for the influencing variable X and target variable Y), based on the percentage we pass in that determines how to split the dataset.

From the graph visualizations, we can see which variables likely influence the house' listing 'Price', which is our target variable, so we can choose which variables will be our influencing

variables based on that. We will train the test dataset by calling the train_test_split() method and using 80% of the original dataset as the training data and 20% as the testing data.

Then, we will use the training lists produced by the train_test_split() method in different Machine Learning models, such as Linear Regression, Random Forests, and Decision Trees. After using each model, we will compare the Mean Squared Error from each model to determine which model was the most accurate. Since the Mean Squared Error is susceptible to outliers, we may remove the outliers from the dataset prior to splitting the dataset into training and testing lists. The model that is most accurate will be refined and used with the testing data set to predict the 'Price' values.

We may also graph the regression line and test data to illustrate how accurate the best model is at predicting the test 'Prices'.

ML:

The three Machine Learning models we will use are Linear Regression, Random Forests, and Decision Trees. These are all Supervised Machine Learning models because we already have an idea of what variables may impact the house 'Price' (due to the visualizations) and we already know what variable we want to predict ('Price').

Testing Performance:

We will use sklearn's cross_val_score() to compare the different ML models and decide which model is the best in order to refine it further. We will validate the ML models by using N fold cross validation to make sure that we are getting the same results regardless of which part of the dataset we are training on. We will also look at the Mean Squared Error and see which value from the models is the lowest in order to determine how close to the actual values we are getting.

Goals (Expected Outcomes):

Our goal is to determine which factors from Bay Area Houses are most effective in determining the price of the house. This will be done by creating visualizations that display the correlations between different variables and the 'Price' variable, as well as by creating Machine Learning models that try to predict the 'Price' variable.

We anticipate that larger house sizes, better school quality, and shorter commute times will lead to a higher house 'Price'. It may also be possible that two traits that may be positively correlated with a higher house 'Price' may be negatively correlated to each other.

Potential challenges we may face are null values and outliers in the data. These will have to be removed prior to creating the respective visualizations / machine learning models that use the columns they are located in.