METCS521 Final Project Submission

Predicting Housing Prices

Authors: Ayoub Amghar, Maroun Boussif, and Supawadee Phakdee

Introduction:

In this project housing prices were chosen to be analyzed. The analysis consisted of implementing a linear regression, then comparing different regression models, followed by forecasting those models, visualizing the results, and finally comparing the results to Zillow. Housing prices were chosen to be predicted because many claim to know how the market works and which way it will go. By using statistics, it can be shown which variables truly matter in the market, and can predict how housing prices will trend in the future.

Medel and Data Analysis:

The following scatter plot (figure 1) reflects the relationship between the neighborhood of the house and the sale price variable. From the plot, NoRidge neighborhood has the most expensive houses of any neighborhood in our database, which is over $700,000. The neighborhood with the lowest home prices is 34,900, which is Idotrr. The mean of housing price from 2006 to 2010 is $186,354.32. Because home prices are very much location dependent, there are many neighborhoods with near minimum values. NoRidge neighborhood has the largest range of values. Home prices having a large range of values typically mean a diversity of home types available. Blueste neighborhood had the least amount of observations. This is mostly likely due to a lack of homes on the market for that particular neighborhood.

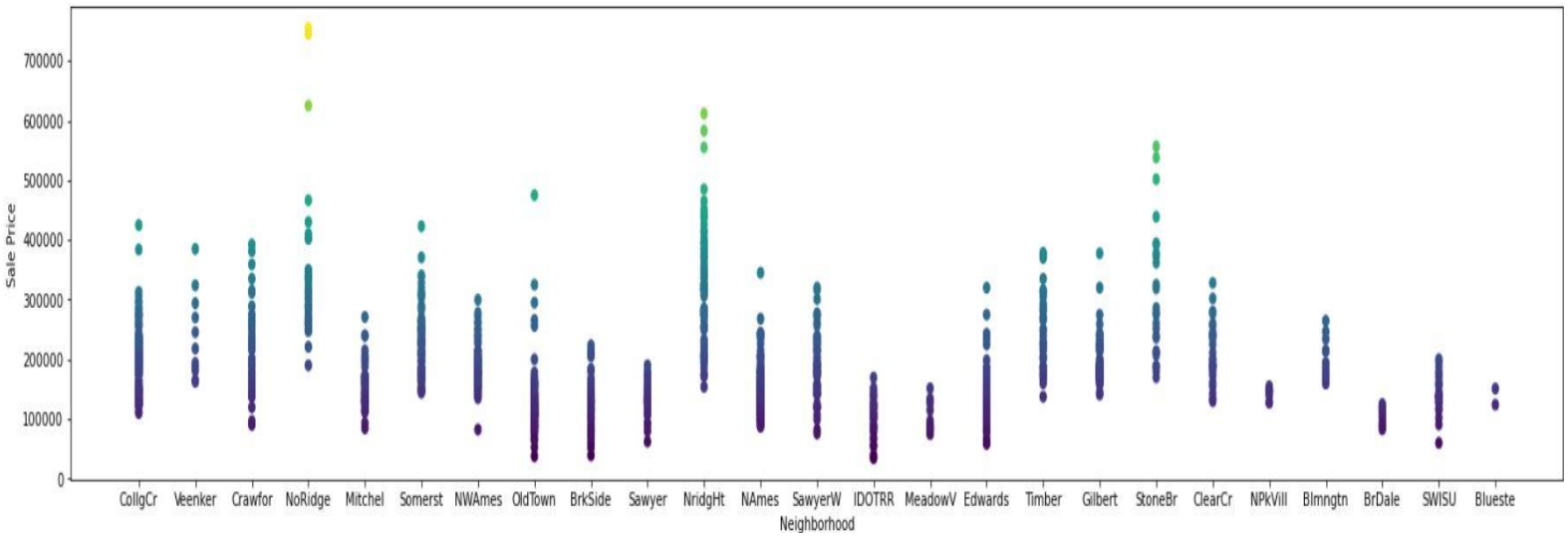

Figure 1: Scatter plot of the neighborhood and housing sale price model

The Histogram represents the sale prices density for houses, as shown in the graph that most of the houses density prices range between 100,000 and 400,000, as the bins describe that in the graph, because the kernel curves follow the flow of the data given in our data.
The peak of the density where we have most houses at the range price between 100,000 and 200, 000, as the data distribution were showing in the graph. After the 200,000 we can see that the density started falling down because we have less houses to sell at the range price between 200,000 and 400,000 and the data getting lesse until 600,000.
The blue bins represent the number of houses in the specific price range according to the y axis which present the density.

The green line in the Histogram represents the kernel density estimation,The kernel Density estimation is mathematical by plotting out data to show the changes of the density, the curve calculated by measuring the distance of all the points of the bins along the distribution. The kernel density calculates the density of the bins points to give us a clear idea about the peak and the lower data concentration of housing prices. The Kernel used to give a clear visualization about the highest peak point of the bins, and also give you a whole idea about the quantity of houses in a specific price range.

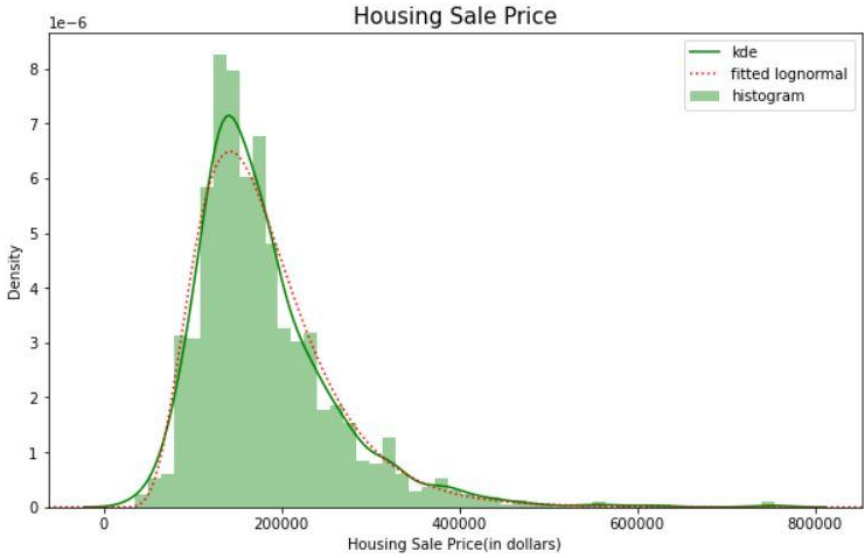

Figure 2: Histogram and density plot of the housing sale price

Extension 1: Linear Regression

The following linear regression shows the relationship between the sale price of a home and the size of the property. There are a significantly higher number of smaller sized lots or properties, and also a larger number of “low cost” houses. The linear regression shows a strong relationship between the variables for every year, 2006 through 2010. It is obvious that a larger property will have a higher cost. But as the properties get larger, so does the range of prices for these properties. The majority of the observations are below $400,000, but once properties start to go over 12,000 sq.ft the variation increases dramatically.

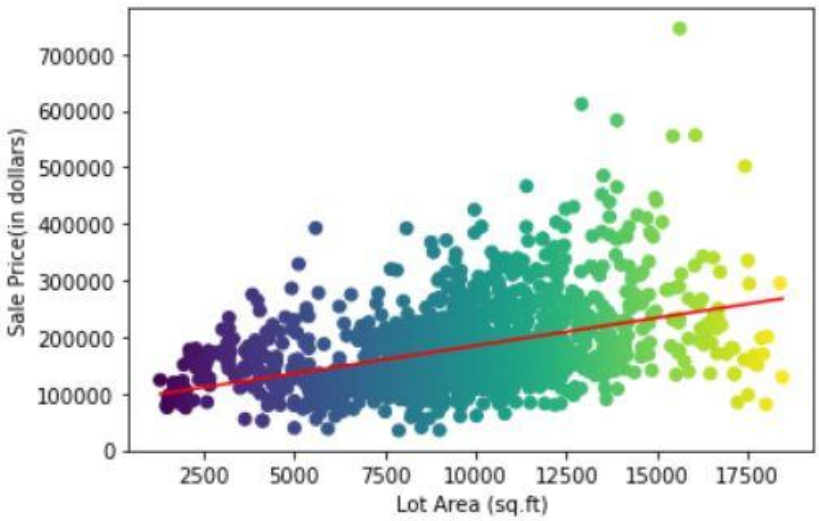

Figure 3: Linear regression model

Comparing the price of the home and the size of the area showed variation from year to year. The earlier the year, the more related the area is to home price. As time has gone on, the price of the homes have had less to do with the area. For example, the value of houses in the city has gone up and the price is high compared to the past.

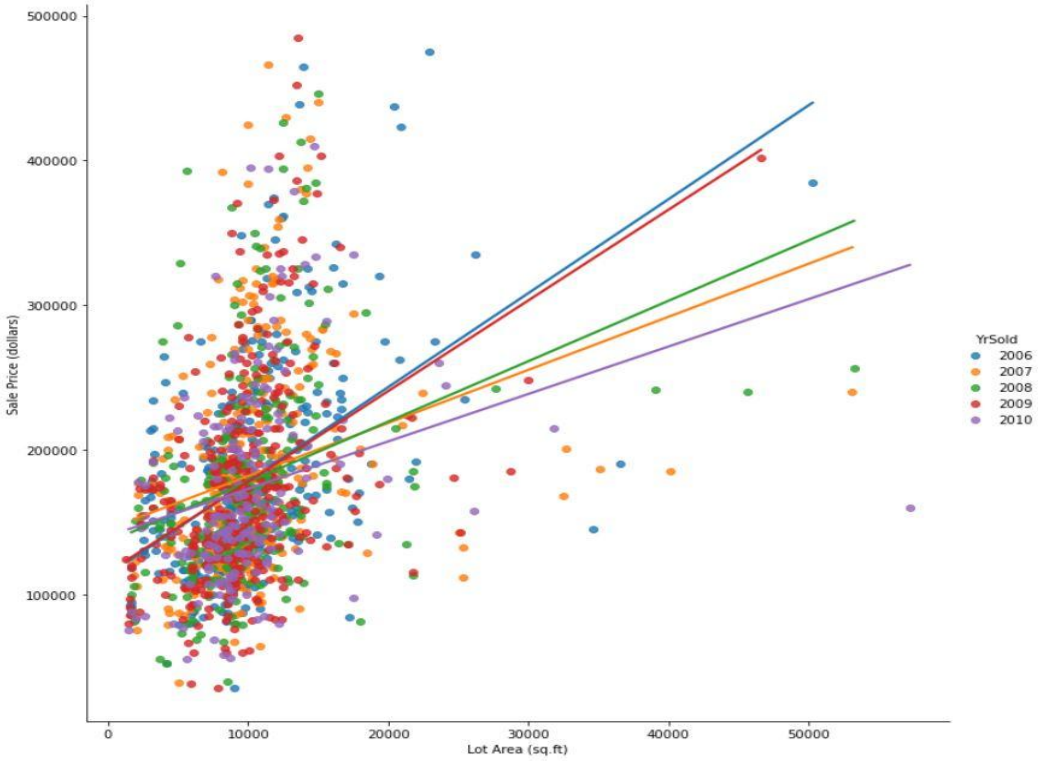

Figure 4: Sale price and Lot area of each year Linear regression model

Extension 2: Correlation Matrix

In figure 5, we only collect the variables that are important to the housing sale price. The two factors that are most correlated to the sale price are overall quality and ground level living area. Overall quality is a rating between 1-10 and ground level living area is the amount of floor space on the ground floor. Other factors that are correlated to sale price include the amount of cars that can fit in a garage, the garage area, and the first floor area.

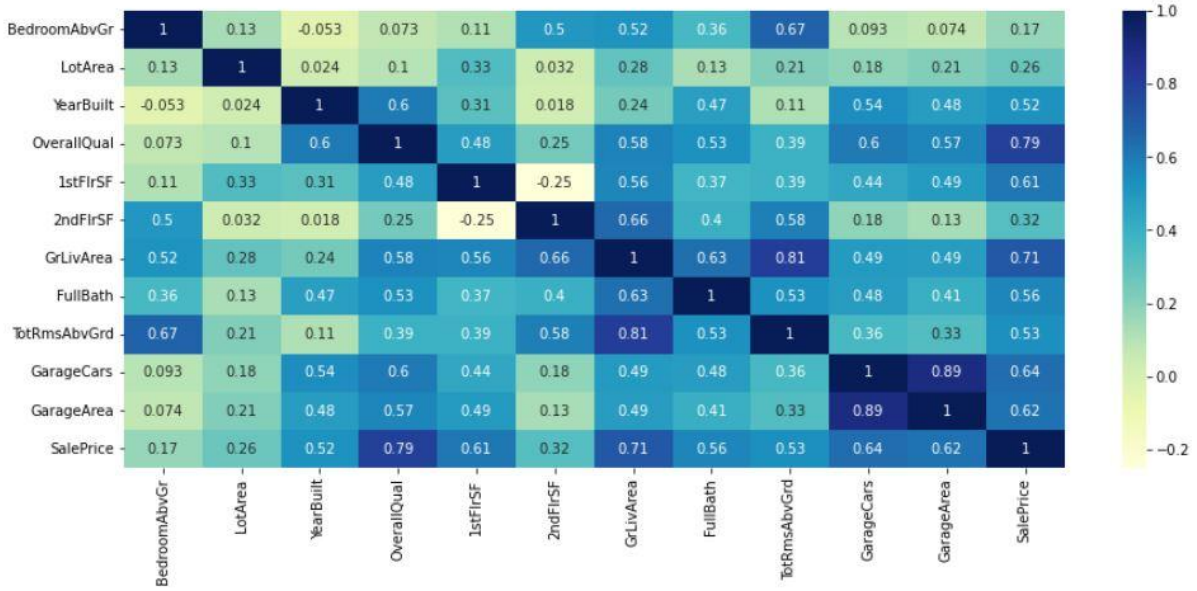

Figure 5: Correlation

Extension 3: Visually and descriptively compare results to Zillow.

The figure 6 draws a scatter plot comparing our housing DB of the sale price under 1,000,000 for the the lot area under 20,000, in order to compare them the zillow DB for the sale price less than 1,000,000 for the all the lot area less than 20,000, the orange plots presents our housing Db and the blue plots presents the zillow DB. clearly we can see that the x axis presents the surface Sqft and the y axis presents the sale price. There is clear correlation between the two data sets. that the more the plot above the ground level, the higher the price of the lot is. From the graph we can determine that there are more areas for our demo housing DB at the range sqft was specified than the zillow DB. Also we see that the prices for demo housing DB are cheaper than the zillow even if they get bigger in sqft, which was presented with blue color. On the other hand, the plots for the zillow DB were smaller in sq ft size but more expensive which reflected the location of that lot(city, near transportation…). Also most of the areas of the zillow db were between the range of 0 and 2,500 sqft, at the opposite in our data we see that lot areas were distributed almost all on the x axis almost for all the lot sqft . Also the main take out from the graph is how the prices have changed over time, that way can give us space to predict the price changes over years. The years difference between the two data sets is 10 years.

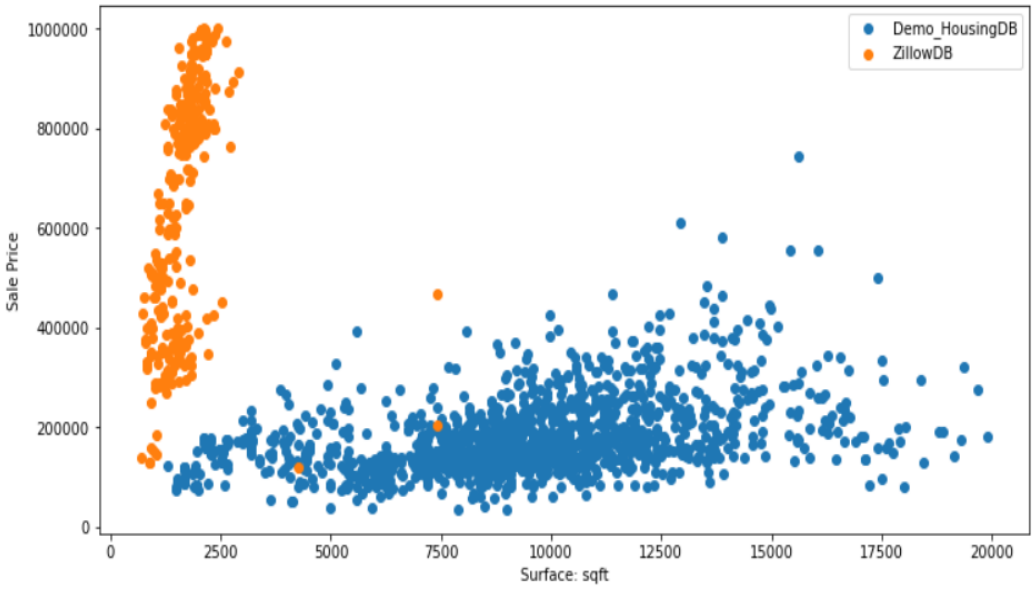

Figure 6: Comparison of housing sale prices between Zillow and demo housing data

Conclusion:

This project helped us to bring together all the skills we have learned throughout the course. We have worked on the House Price Predictor option. The dataset we have used contains housing data from 2006 to 2010, and the problem we tried to solve was how the pricing has changed after 2010. Therefore, the Zillow database was a helpful source to make this comparison and housing price prediction. We have observed that the prices are increasing over years at a high rate, even if the house area is small. We first programmatically downloaded the provided dataset and loaded it to dataframe. We have used Pandas, sklearn, and other tools to visualize our dataset to understand its content. The graphs produced from filtered data helped us analyze the data. We have included pushed through GIT a script to download dataset, the data files(Demo and Zillow), Data Analysis (Script to generate analysis), Machine learning Jypter file that includes all the 3 extensions (Linear regression, Correlation, and Visually & descriptively comparing Zillow and our data). And we concluded with a PDF write-up that explains the purpose of this project and all the methods we used to analyze and compare the data.