Predicting the prices of real estate

Omar Bassem Steitieh

23rd of August

I. Introduction

1.1 Background

Long before the last Los Angeles real estate bubble when real, inflation-adjusted house prices nearly tripled from 1997 to 2006 but then fell nearly in half by 2011... and long before the S&L bubble when real Los Angeles house prices fell over 40% from the peak in 1989 to 1997...There was the Los Angeles real estate bubble of the 1880s when real land prices increased 10-fold from 1882 to 1888 and then fell by one-third in one year, the next year, 1889. Even though the population of Los Angeles was skyrocketing at the time, real estate prices got way ahead of themselves and then busted hard. From land speculation in New York state after the Revolutionary War, to a huge bubble in Chicago in the 1830s, to our 21st century bubble, the United States has always been a nation of real estate speculators, according to Edward Glaeser of Harvard University in his brief and paper called, "A Nation of Gamblers: Real Estate Speculation and American History," from 2013. In addition to the bubbles mentioned above, the paper also covers bubbles in cotton farm land in Alabama in the early 1800s, and wheat farm land in Iowa and New York City housing in the early 1900s.

1.2 Problem

It is hard sometimes to look for real estate within your budget and its is very time consuming to go around visiting different real estates. This project aims to make this easier and being able to look for desired houses easily

1.3 Interest

The goal of the new technology is to make life easier for all people, therefore I believe this is a new way to explore houses for sale next to your area.

II. Data acquisition and cleaning

2.1 Data Sources

Houses around Washington data was provided by another course however the data was so unpredictable due to having many factors that affect the price of a certain real estate.

2.2 Data Cleaning

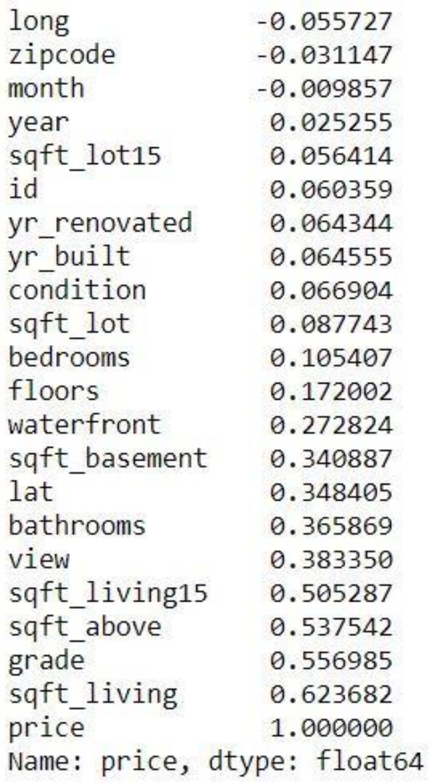
A couple problems came up. First, few real estates had large number of bedrooms which causes the data to be uncertain therefore number of bedrooms greater than 3 was not included in the study. Secondly, Data cannot be sorted due to the different categories each rea estate holds.

III. Exploratory Data

3.1 Calculation of Target variable

The main feature people tend to focus on while looking for real estate is the Price (our target variable). What affects the price the most? It is how bug is it and its grade.

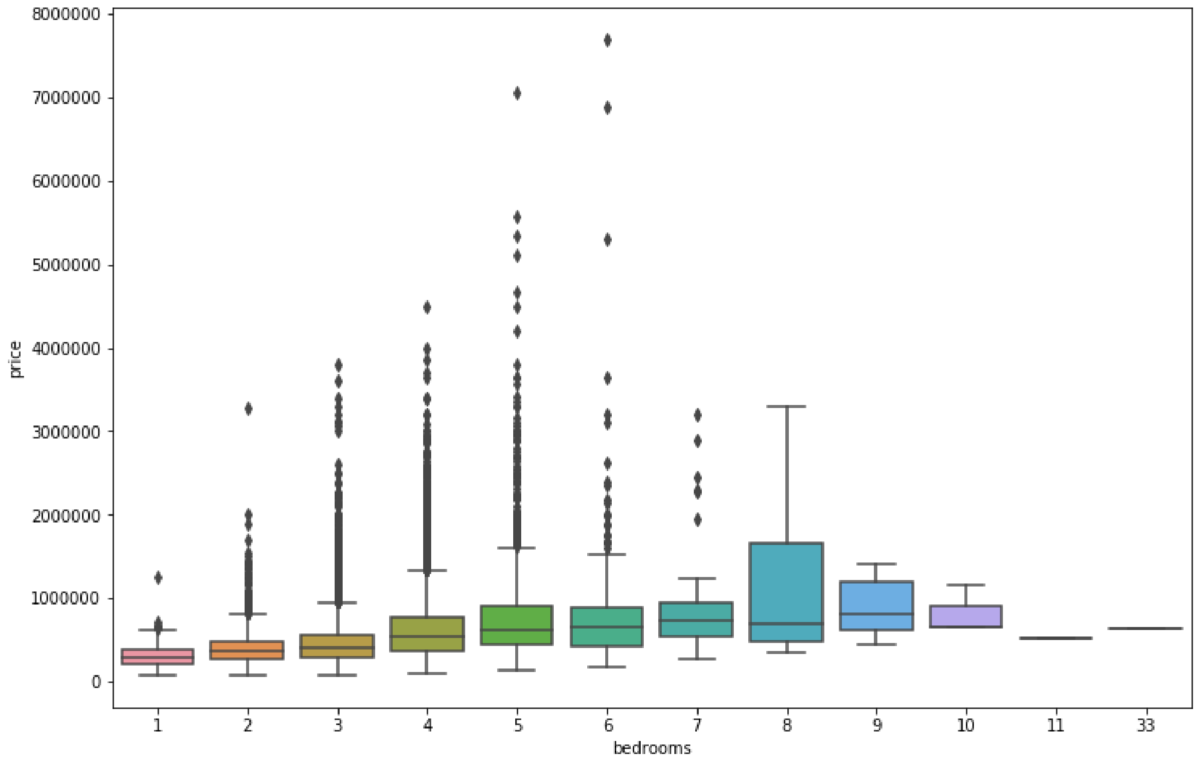
The following figure values will say sort underscore values and here I can see things that are either highly positively correlated or highly negatively correlated.



```
long            -0.055727
zipcode         -0.031147
month           -0.009857
year             0.025255
sqft_lot15       0.056414
id               0.060359
yr_renovated     0.064344
yr_built         0.064555
condition        0.066904
sqft_lot         0.087743
bedrooms         0.105407
floors           0.172002
waterfront       0.272824
sqft_basement    0.340887
lat              0.348405
bathrooms        0.365869
view             0.383350
sqft_living15    0.505287
sqft_above       0.537542
grade            0.556985
sqft_living      0.623682
price            1.000000
Name: price, dtype: float64
```

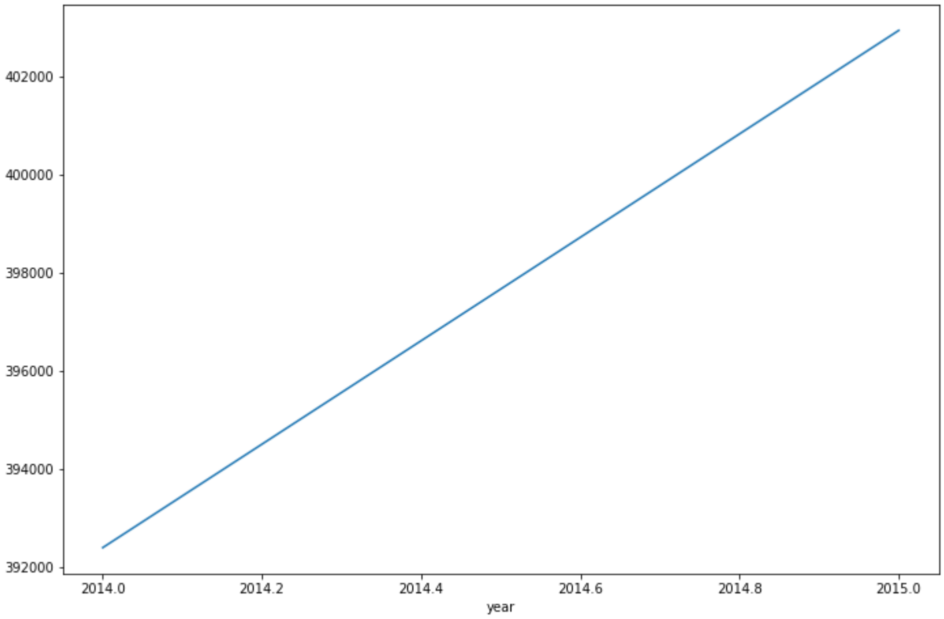
Figure 1: Correlation

3.2 Relationship between Price and number of bedrooms



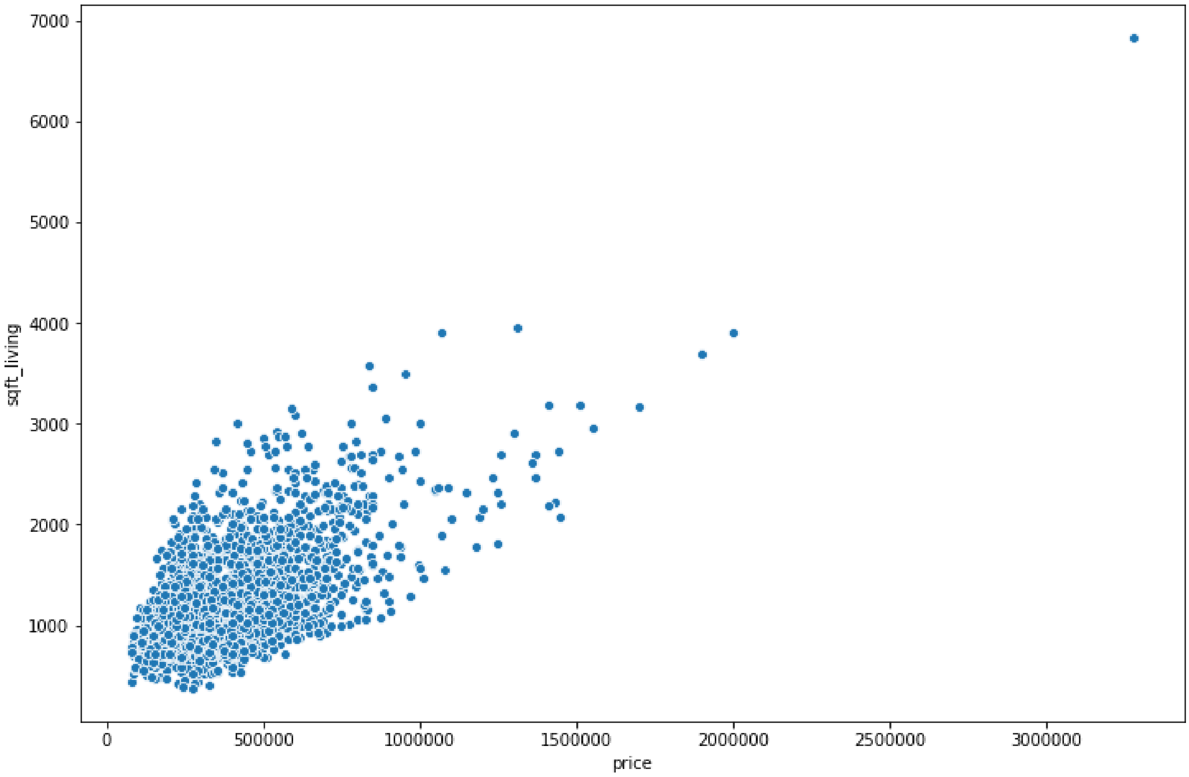
As you can see there are different varieties of bedrooms but at least we can have an idea of the correlation of the number of bedrooms against Price.

3.3 Relationship between Price and Year built



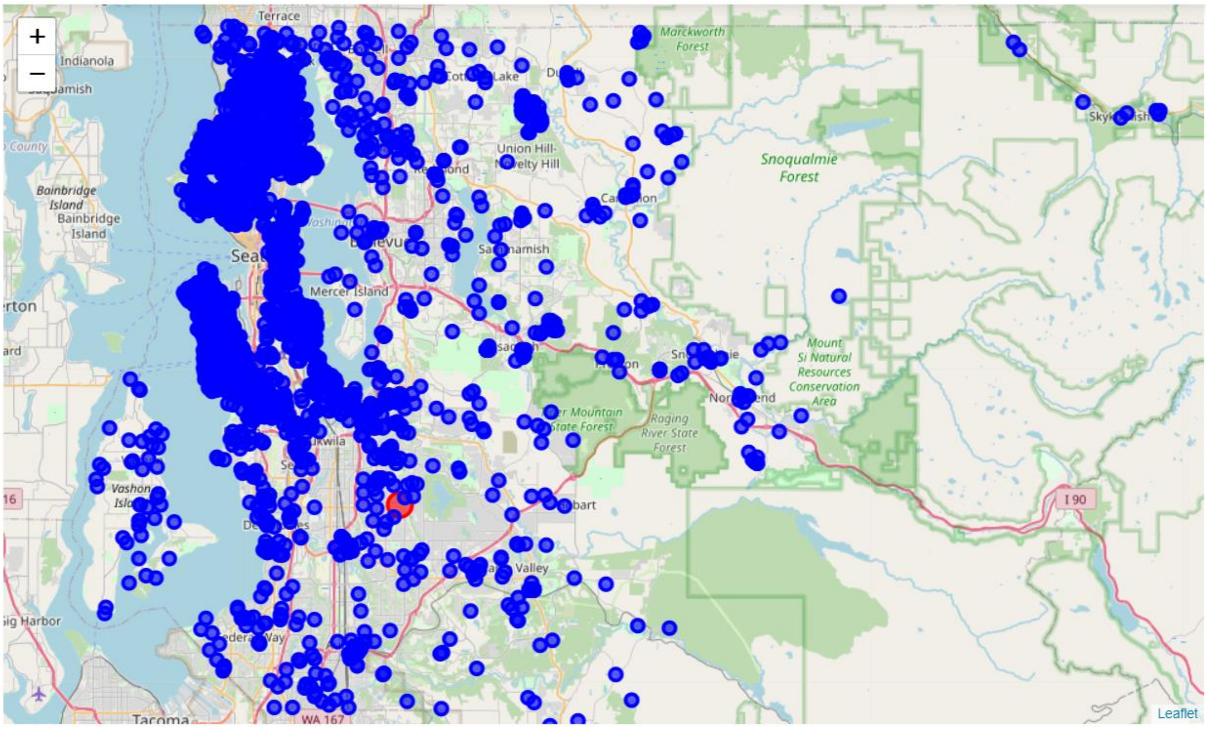
As we can see it is a linear relationship as newer as it gets the higher the price disregarding other categories.

3.4 Relationship between Price and Square feet of the real estate



This variable has the maximum correlation with the price according to figure 1.

3.5 Foursquare API

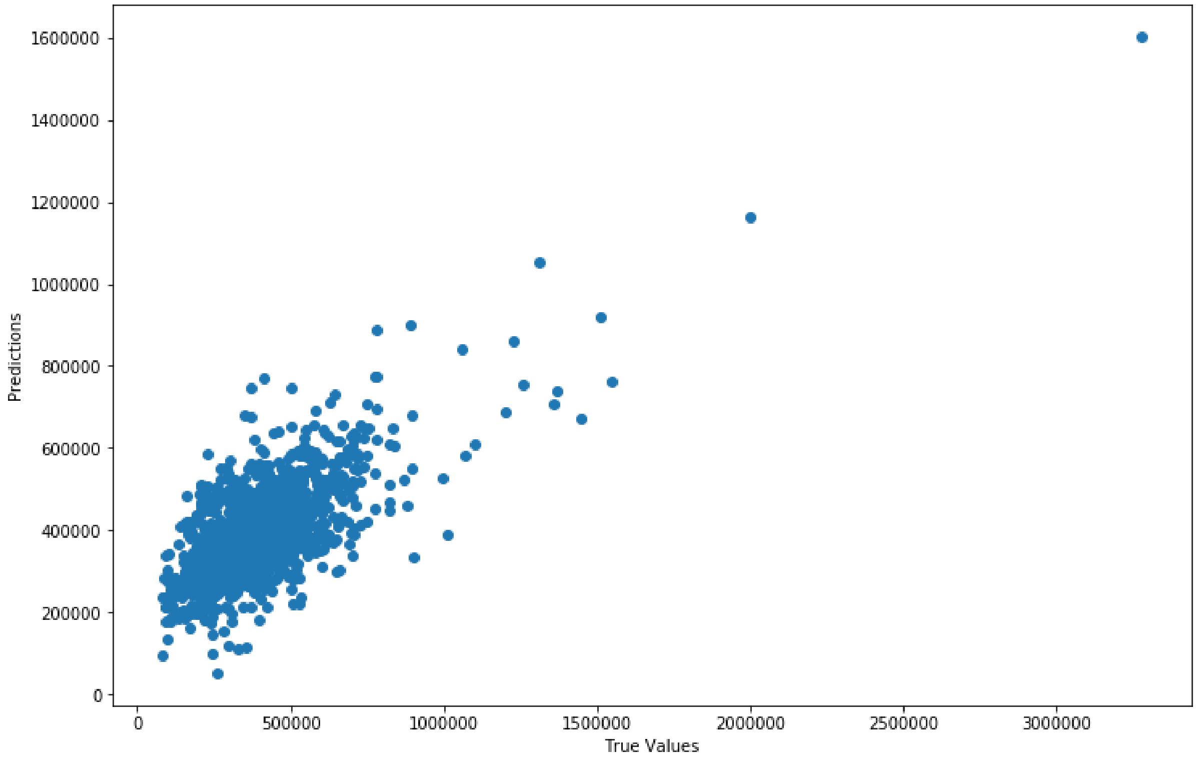


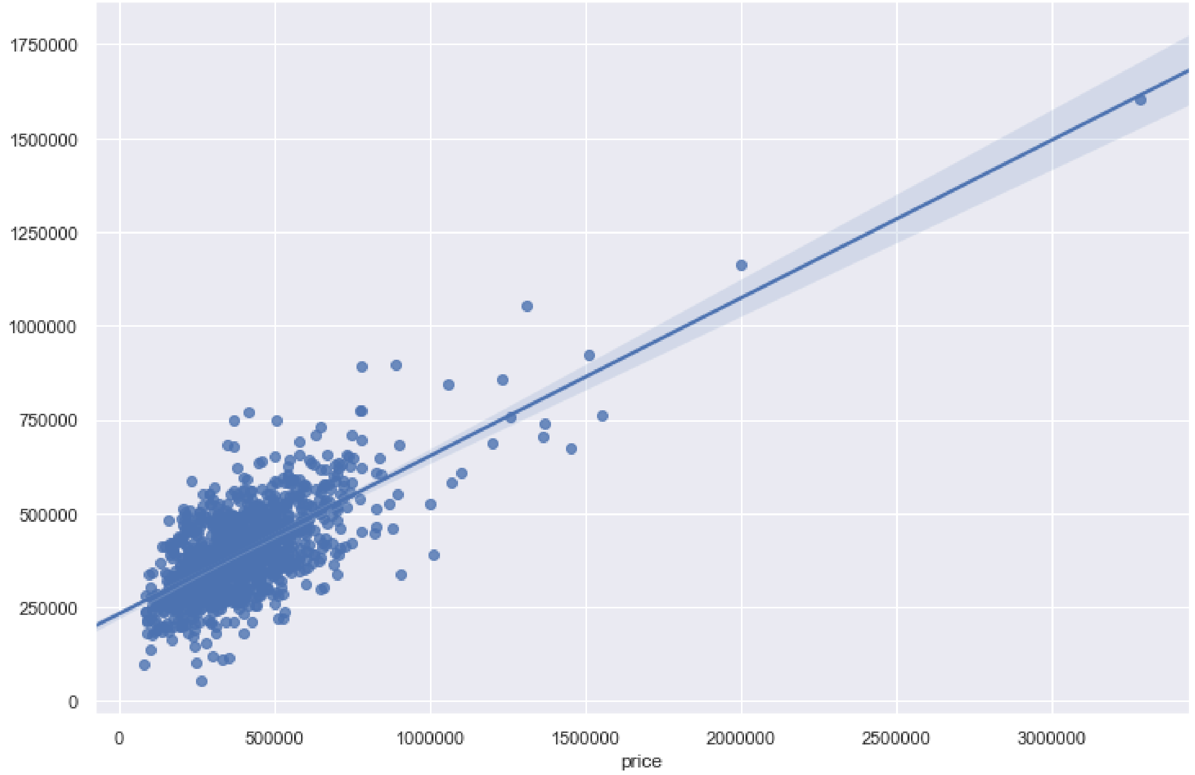
Foursquare API was used, to be able to determine all listed real estates surrounding my location (in red).

IV. Predictive Modeling

In predicting a continuous integer; we will be using Regression using train/test split approach

4.1 Regression model





After applying the linear regression. We have achieved an accuracy of 47.8%. due to the various categories available.

V. Discussion and conclusion

In this study we have achieved a low accuracy due to the various information that can be fed to the model. Also, the model in this study mainly focused on individual features. However for future references we might have to categorize the such huge data to be able to study each category alone and that can be done by classification.