Capstone Project
The Battle of Neighborhoods

Using location data to predict house sale prices

Elin Ansin

August 2020

1. Introduction and Business Problem

Location, location, location! You have heard it before. It's a common mantra in real estate. Many factors can influence the value of a home, but the surrounding area is one of the most influential. The value of the home rises and falls with the value of the properties around it and the commercial or recreational activities that develop nearby. It would therefore be beneficial to use location data when predicting house prices.

Traditionally historical sales data is used to predict prices in individual neighborhoods. Can we combine historical sales data with location data to make even better predictions? An accurate prediction on the house price is important to a lot of different stakeholders as prospective homeowners, real estate agencies, developers, investors, mortgage lenders and insurers etc.

2. Data

2.1 House Sales in King County, USA

This dataset from Kaggle contains house sale prices for King County, which includes Seattle. It includes 21613 homes sold between May 2014 and May 2015. The dataset includes attributes and features for the sold objects such as square footage, number of bedrooms, building year etc. It also contains the latitude and longitude coordinates for the location of the house.

Example of the data:

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...
0	7129300520	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	0	...
1	6414100192	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	0	...
2	5631500400	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	0	...
3	2487200875	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	0	...
4	1954400510	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	0	...

2.2 Foursquare Location Data

The initial idea was that since we know the location of each house we could use the Foursquare API to find out which services and entertainment or recreation opportunities like movie theatres, parks, and golf courses can be found in the immediate area of each house and include that information when

building our prediction model. Unfortunately, due to limited calls on the Foursquare API I could not collect that information for 21613 houses separately so that information was instead collected for each of the 70 zip codes in the data and then added to each house with that zip code. The data collected was the top 100 venues within 500 m from the centre of the zip code calculated by taking the average latitude and longitude for all houses in the zip code.

Example of data collected for one zip code:

	name	categories	location.lat	location.lng
0	Seattle Urban Academy	High School	47.537993	-122.284066
1	Amazing Grace Christian School	School	47.510888	-122.259854
2	Summit Sierra High School	High School	47.597844	-122.318325
3	Coinstar	Bank	47.519900	-122.268300
4	KeyMe	Locksmith	47.520774	-122.268436

From this I created a variable for each of the 203 different venue categories found in the collected data where the number of venues in each category was summarized for each zip code.

Example data for one zip code:

zipcode	ATM	American Restaurant	Antique Shop	Arcade	Art Museum	Arts & Crafts Store	Asian Restaurant	Auto Garage	Automotive Shop	...
98004	1	2	0	1	1	0	0	0	0	...

3. Dimension Reduction

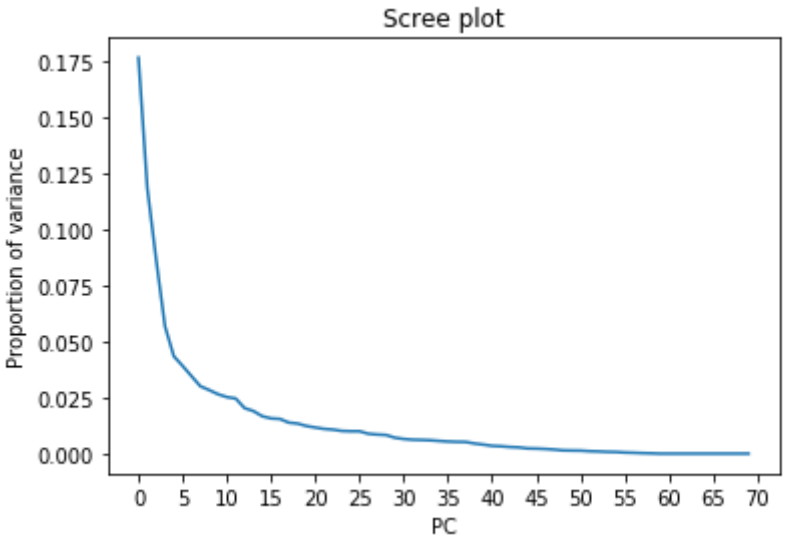
The location dataset contains 203 variables indicating what venue categories are presented in each neighborhood. This is a large number of variables where many of them are highly correlated to each other since they often appear together. This would be a problem when fitting the model since a linear model require the independent variables to be independent of each other. I therefore performed a Principal Component Analysis (PCA) to reduce the number of variables in to a set of new uncorrelated variables (principal components). Where the principal components are a linear combination of the original variables. A few principal components account for most of the variance in the dataset.

3.1 Applying Principal Component Analysis

PCA is affected by scale so before applying the principal component analysis I normalize the data to unit scale (mean=0 and variance=1).

3.1.1 Deciding on number of principal components to use

I used the elbow-method to find how many principal components to use. This was done by calculating the proportion of variance explained by each feature and making a scree plot and find the “elbow”, the point where we have a large drop of extra variance explained by adding an extra principal component.



In this case we see that drop after 4 principal components so that is what I will use. These principal components account for 50 percent of the variance in the dataset.

3.1.2 Interpretation of the Principal Components

The principal components are linear combinations of the original variables which in this case are the occurrence of 203 different venue categories. The larger the absolute value of the coefficient the more important the corresponding variable is when calculating the principal component.

Principal Component 1

The variables with the largest effect on this principal component are:

Original variable	***Coefficient***
Clothing Store	0,157987497
Bank	0,147104038
Hotel	0,146727962
Cosmetics Shop	0,146308489
Lingerie Store	0,145308154
Spa	0,144867286
Mobile Phone Shop	0,142071063
Jewelry Store	0,138485711
Electronics Store	0,137040001
Seafood Restaurant	0,136800701
Movie Theater	0,136008091
Pizza Place	0,134677804
Shopping Mall	0,131013364
American Restaurant	0,130853579
Café	0,128211411
Gourmet Shop	0,125103165
Sandwich Place	0,122072291
Toy / Game Store	0,121845659
Furniture / Home Store	0,119164974

Mexican Restaurant	0,117840915
Arcade	0,111348905
Art Museum	0,111348905
Brazilian Restaurant	0,111348905
Cheese Shop	0,111348905
Cocktail Bar	0,111348905
Cycle Studio	0,111348905
Dumpling Restaurant	0,111348905
Hotel Bar	0,111348905
Hotpot Restaurant	0,111348905
Mattress Store	0,111348905
Men's Store	0,111348905
New American Restaurant	0,111348905
Stationery Store	0,111348905
Steakhouse	0,111348905
Taiwanese Restaurant	0,111348905
Yoga Studio	0,111348905
Sporting Goods Shop	0,109875914
Women's Store	0,107466939
Ramen Restaurant	0,105822553
Burger Joint	0,103439207
Gift Shop	0,100322066

My interpretation is that a high value on principal component 1 indicates an area with a lot of shops, restaurants, services and entertainments available nearby.

Principal Component 2

The variables with the largest positive and negative effect on this principal component are:

Original variable	*Coefficient*
BBQ Joint	0,151976
Bookstore	0,151976
Burrito Place	0,151976
Department Store	0,151976
Laser Tag	0,151976
Lighting Store	0,151976
Miscellaneous Shop	0,151976
Snack Place	0,151976
Video Game Store	0,151976
Weight Loss Center	0,151976
Big Box Store	0,151976
Shoe Store	0,147627
Burger Joint	0,143469
Korean Restaurant	0,143351

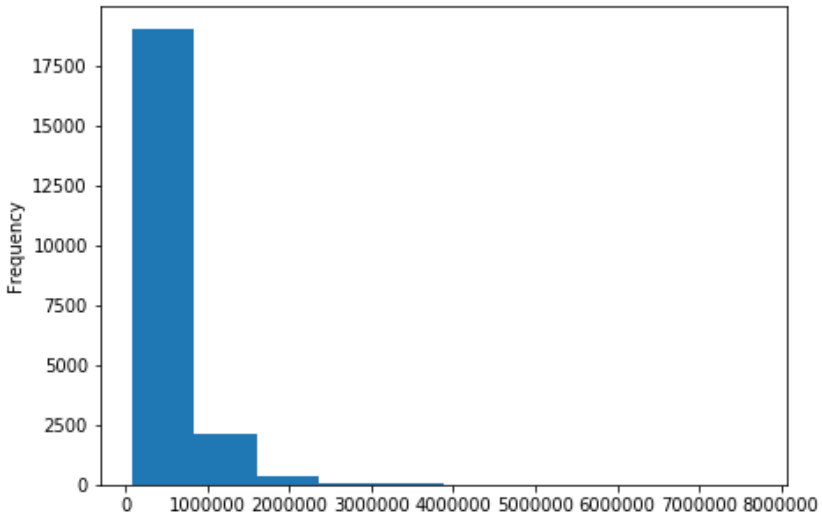
Sporting Goods Shop	0,139455
Japanese Restaurant	0,137589
Shipping Store	0,130878
Ice Cream Shop	0,129654
Optical Shop	0,12942
Wings Joint	0,114572
Supplement Shop	0,113826
Discount Store	0,112253
Diner	0,112175
Hawaiian Restaurant	0,11129
Gas Station	0,107207
Pet Store	0,100676
Mexican Restaurant	0,095146
Shopping Mall	-0,09093
Poke Place	-0,09224
Supermarket	-0,09224
Irish Pub	-0,09232
Juice Bar	-0,09549
Gastropub	-0,09769
Bowling Alley	-0,0978
Gourmet Shop	-0,09853
Bubble Tea Shop	-0,1009
Toy / Game Store	-0,10977
Arcade	-0,13512
Art Museum	-0,13512
Brazilian Restaurant	-0,13512
Cheese Shop	-0,13512
Cocktail Bar	-0,13512
Cycle Studio	-0,13512
Dumpling Restaurant	-0,13512
Hotel Bar	-0,13512
Hotpot Restaurant	-0,13512
Mattress Store	-0,13512
Men's Store	-0,13512
New American Restaurant	-0,13512
Stationery Store	-0,13512
Steakhouse	-0,13512
Taiwanese Restaurant	-0,13512
Yoga Studio	-0,13512

The interpretation of this principal component is not as straight forward. For high values on principal component 2 there are some food places and stores in the area. There is not a lot of entertainments and there is no shopping mall, yoga studio or fancy restaurants in the area.

4. Exploratory Data Analysis

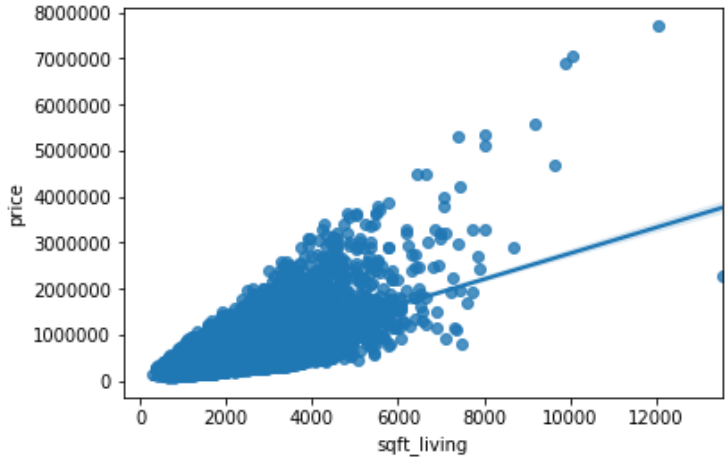
4.1 Target variable

The target variable is the house sale price. The distribution of the house sales prices in the data set are shown in the plot bellow.



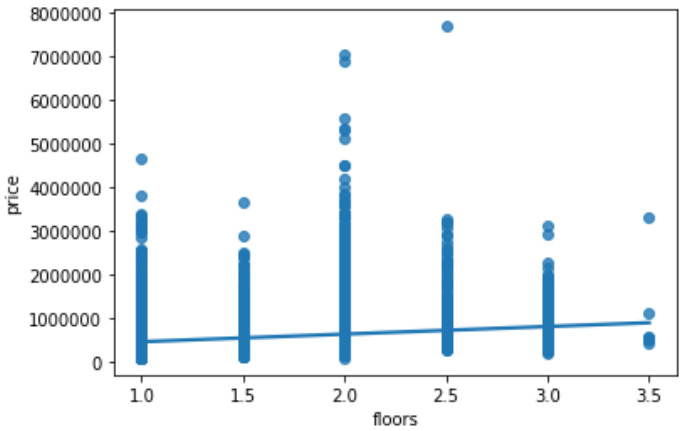
4.2 Relationship between square footage living area and house sale price

It is well known that a house with large living area renders a higher sales price than a smaller one. From the regression plot we can see that this is true for the houses in the dataset. The funnel shaped pattern of the data suggests that the data suffers from heteroskedasticity, i.e. the error terms have non-constant variance. Since we have a large sample size of over 20 000 observations the variance of the least square estimator may still be sufficiently small to obtain precise estimates without correcting for heteroskedasticity.

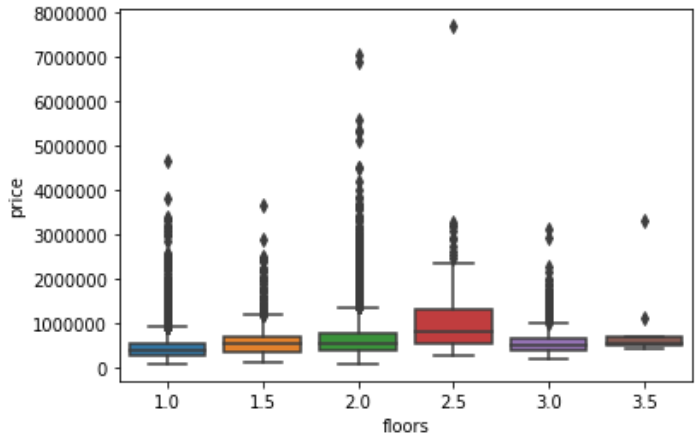


4.3 Relationship between number of floors and house sale price

The regression plot shows that we can't assume a linear relationship between the number of floors and house sales price. I solve that by creating dummies of the number of floors.

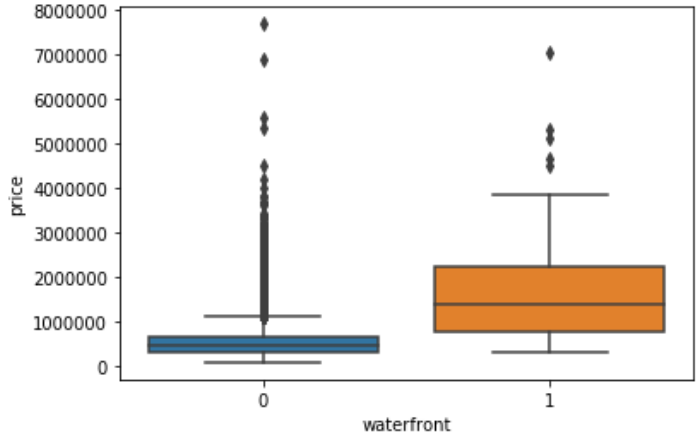


From the boxplot we see that there is for sure not a linear relationship between number of floors and house sales price. The houses with 2.5 floors seem to have higher house sales prices than the other houses.



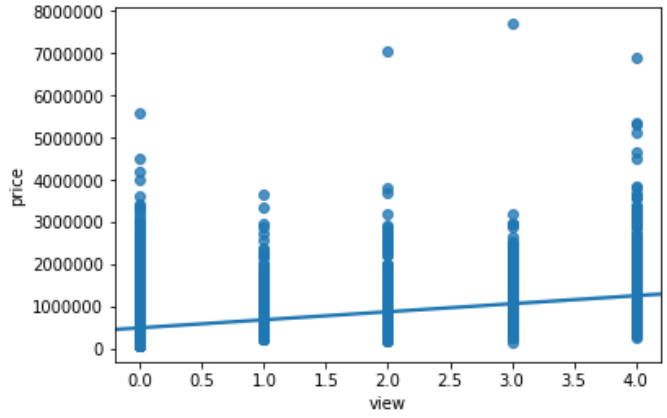
4.4 Relationship between waterfront and house sale price

From the boxplot it is obvious that houses with waterfront have a significantly higher selling price.

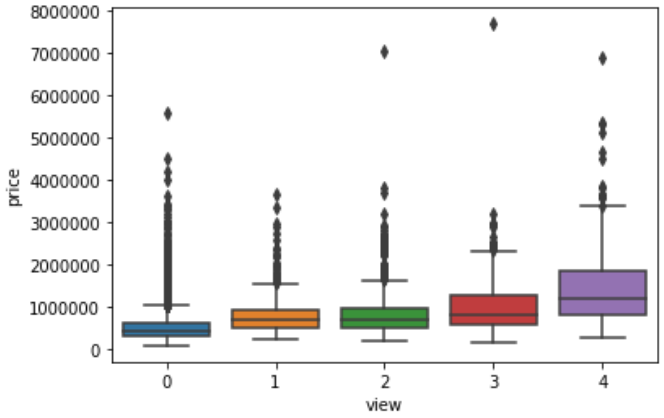


4.5 Relationship between view and house sale price

The variable view is a rating of the view from the house on a scale from 0 to 4, where 4 is the best view. The regression plot shows that we can't assume a linear relationship between the rating of the view and the house sale price. I solve that by creating dummies of the view variable.

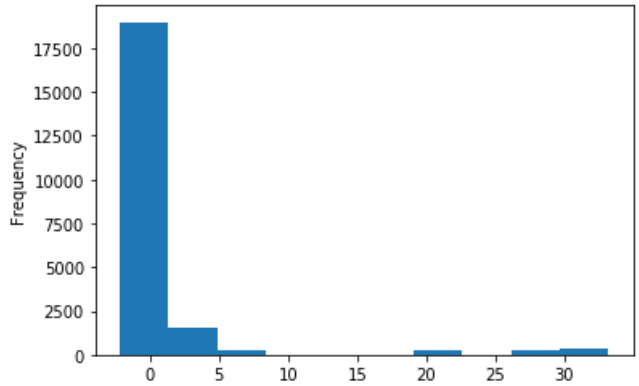


The boxplot of the dummy variables shows that a higher rating of the view renders a higher house sales price, especially houses the highest rating of the view have a much higher house sales price.

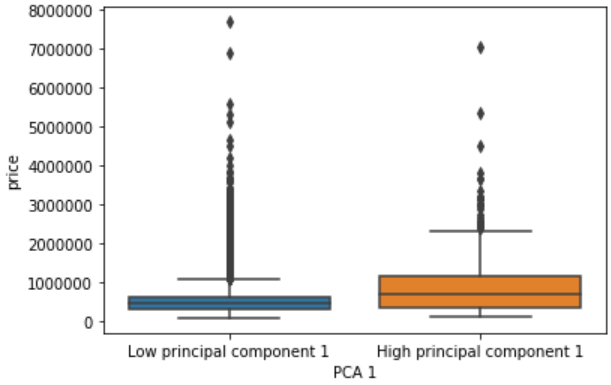


4.5.1 Relationship between principal component 1 and house sales price

Using a regression plot I concluded that principal component 1 and house sale prices do not have a linear relationship. Investigating the variable closer with a histogram showed me that it would be reasonable to categorise PC 1 into high and low values.

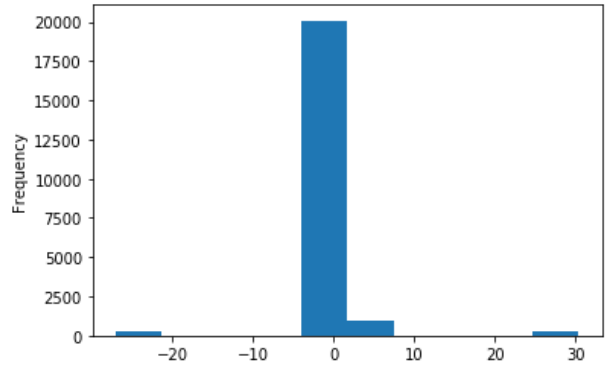


The boxplot of the new categorical variable showed that houses with high values of PC 1 i.e. houses in an area with a lot of shops, restaurants, services and entertainments available nearby have higher house sale prices.

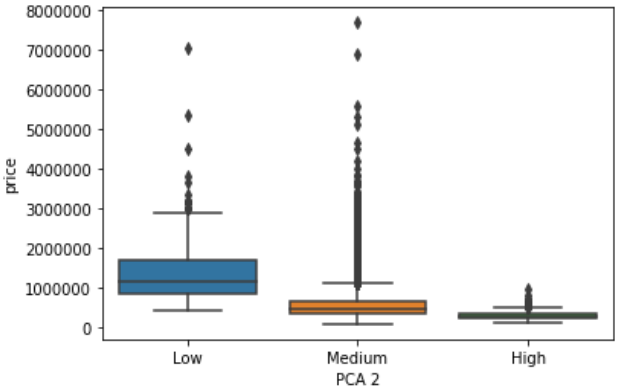


4.5.2 Relationship between Principal Component 2 and house sales price

As for PC 1, PC 2 do not have a linear relationship with house sale price. According to the histogram it is reasonable to categorise PC 2 into high, low and medium values.

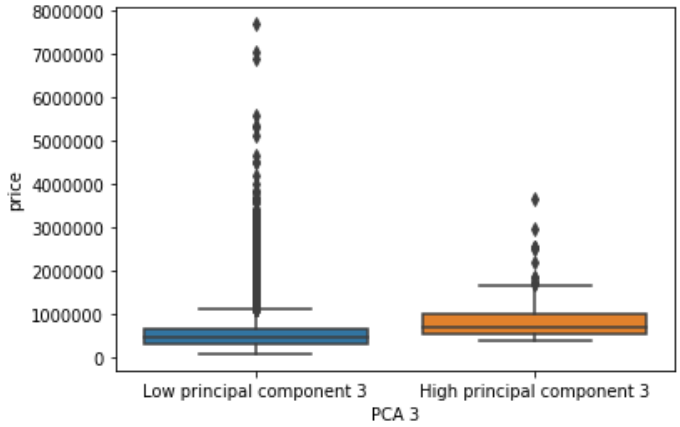


The boxplot of the new categorical variable showed that houses with high values of PC 2 have lower house sales prices and the houses with low values on PC 2 have much higher house sale prices.



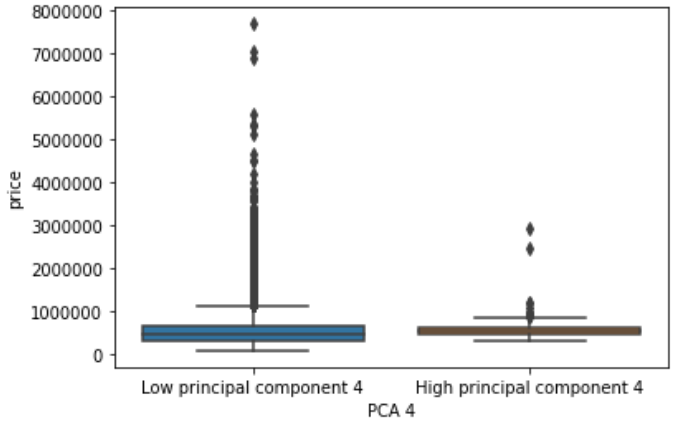
4.5.3 Relationship between Principal Component 3 and house sales price

The difference in house sale prices between houses with high and low values on PC 3 are small.



4.5.4 Relationship between Principal Component 4 and house sales price

There is no significant difference in house sale prices between houses with high and low values on PC 4.



5. Predictive Modelling

A regression model can be used to predict house sale price. I randomly selected 75% of the observations from the data set as a training data set and the remaining 15% observations to a test data set to validate the model on. I applied a linear regression model to the training data set using R-square as evaluation metric. For comparison I also built a linear regression model without principal components.

5.1 Resulting Model

I found the following model to be the best model:

Variable	*Coefficient*
Intercept	-489 492
Square foot living area	247
Waterfront	503 721
View 1	157 784
View 2	119 850
View 3	207 197
View 4	402 652
1.5 Floors	66 597
2 Floors	-5 342
2.5 Floors	246 415
3 Floors	119 168
3.5 Floors	301 002
High Principal Component 1	290 764
Low Principal Component 2	790 288
Medium Principal Component 2	473 721

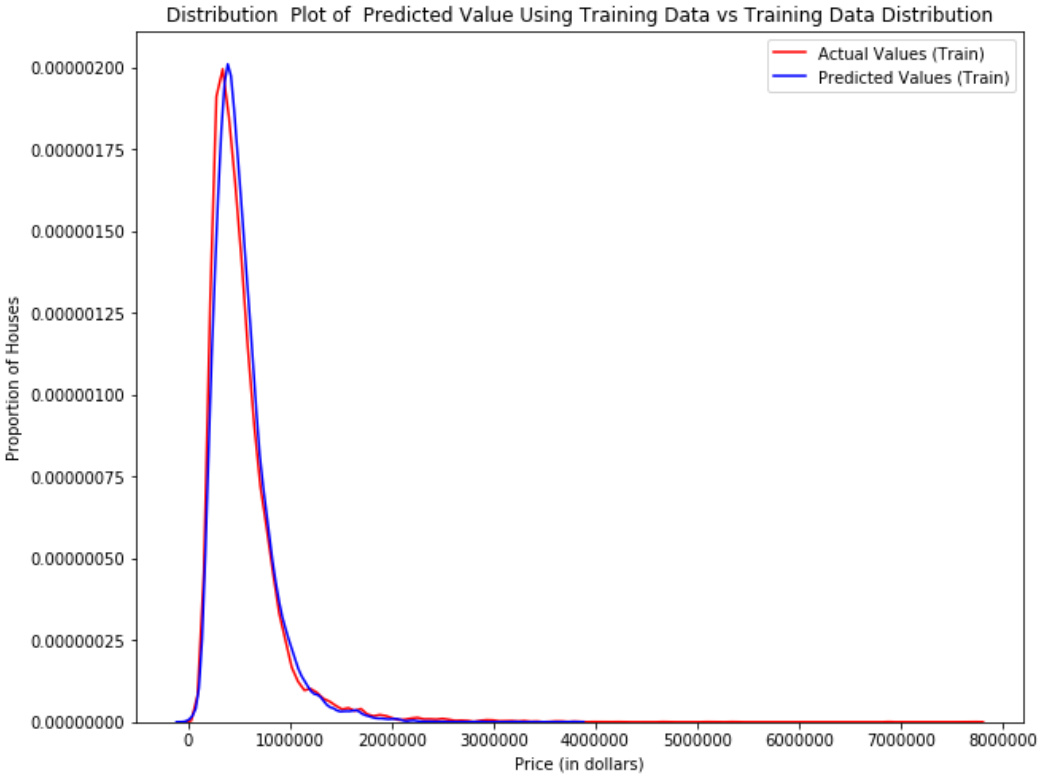
Principal component 3 and 4 did not add any information to the model and where not included.

5.2 Model Performance

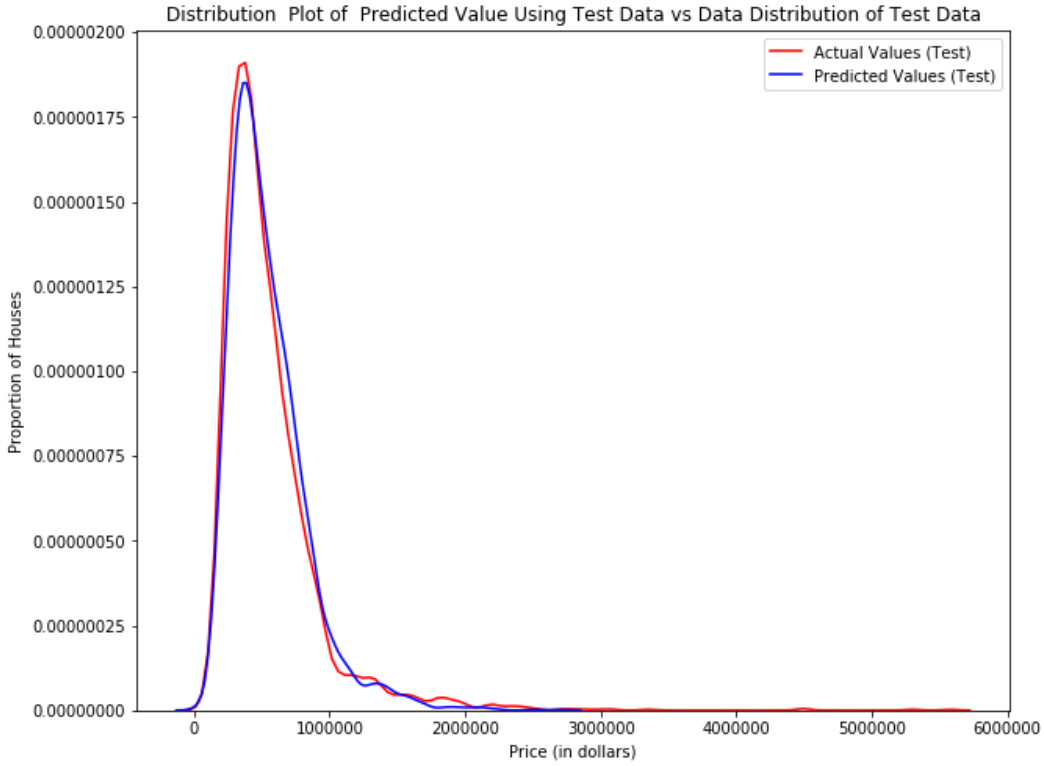
By adding principal components to the model, I gained a 0.05 increase in the R-square value.

Model	*R-square*
Model without Principal Components	0.56
Model with Principal Components	0.61

We can see that the distribution of predicted house prices follows the distribution of actual house prices quite well, indicating that we have a good model. But we can see from the plot that houses with selling prices over 4 million dollars have their selling price underestimated, so for these houses this is not a good model.



Even for the test data we have similar results, indicating that the model does not suffers of over fitting.



6. Discussion

My findings suggest that there is a relation between the house sales price and the range of services found nearby that can be used when modelling house sales prices. I was able to achieve a 0.05 improvement of the R-square of the model by including the location data. These results should however be interpreted with caution. The sales data was from 2014 and 2015 while the location data was collected in 2020, this is of course not ideal. The areas could have changed since the houses was sold, venues could have closed down and new venues could have been developed. This should however only have a noticeable effect if the total composition of venues would have changed drastically for example by lot of new development in an area.

I also made a simplification getting location data for each postal code instead of each house separately, giving same information for all houses in the postal code, this could have had an effect on the results, even though a postal code area is a quite limited area so the composition of venues should not differ that much for different locations in the area. It would be of interest to redo the study making separate calls for each house and use house sales data and location data from same time period.

6.1 Suggestions for further research

In this study I only used data from one area, a natural next step would bet to find out if the results can be generalized to be used for other cities or areas by replicating the study for another or preferable multiple other areas. Can we see that the same structure of venues renders similar effect in different locations?

Using principal components is of course not the only way to handle the location data, is there some specific venues in the surrounding area effecting the sales price? It could also be interesting to see if the rating of the venues nearby effects the house sales price. It is reasonable to believe that people prefer to live close to popular restaurants etc.

7. Conclusions

In this study, I explored the possibility of using location data to improve house sale price predictions. I discovered that I with using principal components was able to identify areas with a specific composition of venues that had an impact on the house sale price and thus improve the house sales price prediction from R-square 0.56 to 0.61. This improvement is quite small but there is still a lot left to investigate in the use of location data for this purpose and this study gives us an indication that it could be possible to use location data to improve the house sales prediction, but further studies are needed.