Real Estate Development

CAPSTONE PROJECT

MARCOS GERALDO

Table of contents

1. Introduction: Business Problem

2. Data

3. Methodology

4. Results

5. Discussion

6. Conclusion

Introduction: Business Problem

This project will provide insights about capital gain on real estate investments.

It wil be targeted to landlords who are evaluating the impact of home improvements projects in the selling price of their properties.

It will also provide a model to estimate the listing price that fits the market valuation of a particular house.

It will use current data published for the city of interest, and use it to stablish the relative weights of the key elements that drive the price of a house.

It will use Foursquare Data to evaluate the distance to relevant venues, and evaluate the weight of those elements in the listing price of a property.

Data

According to the problem definition, the relevant data to understand price valuation, are the following:

selling price

listing price (as a proxy for selling price, that might not be public)

number and distance of venues

To avoid market variations the data will come from current market conditions. The candidates are real state web sites that publish and share freely properties and listing prices:

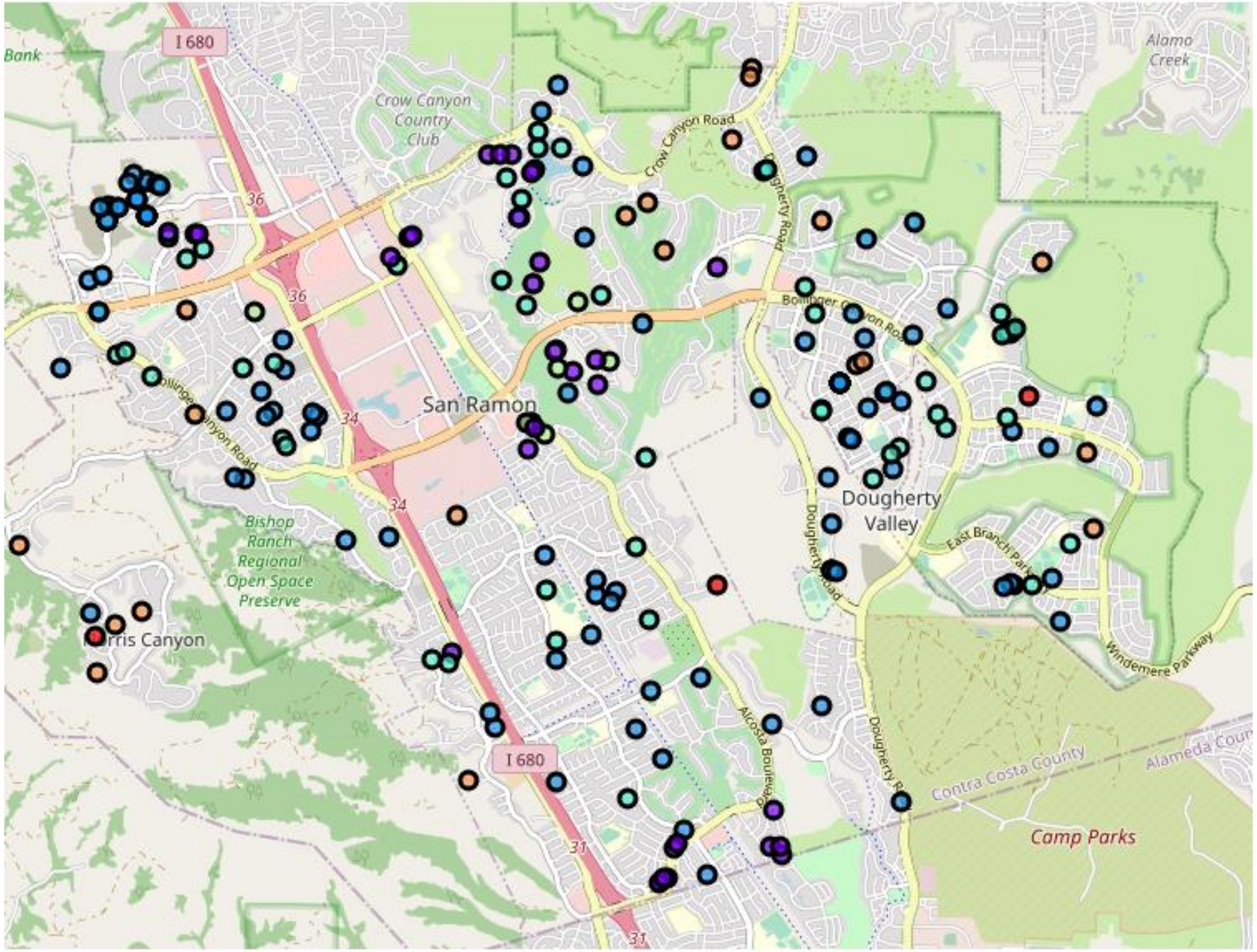
Realtor

FourSquare

Realtor API

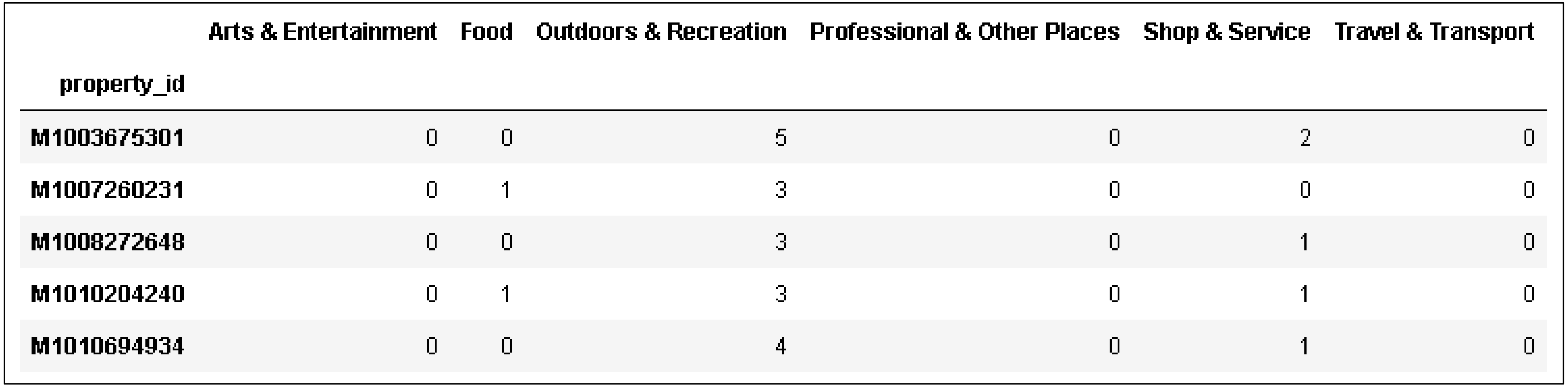
Realtor data provides

- Year of construction
- Constructed suface
- Bedrooms
- Bathrooms
- Garages
- Stories
- School ratings



FourSquare API

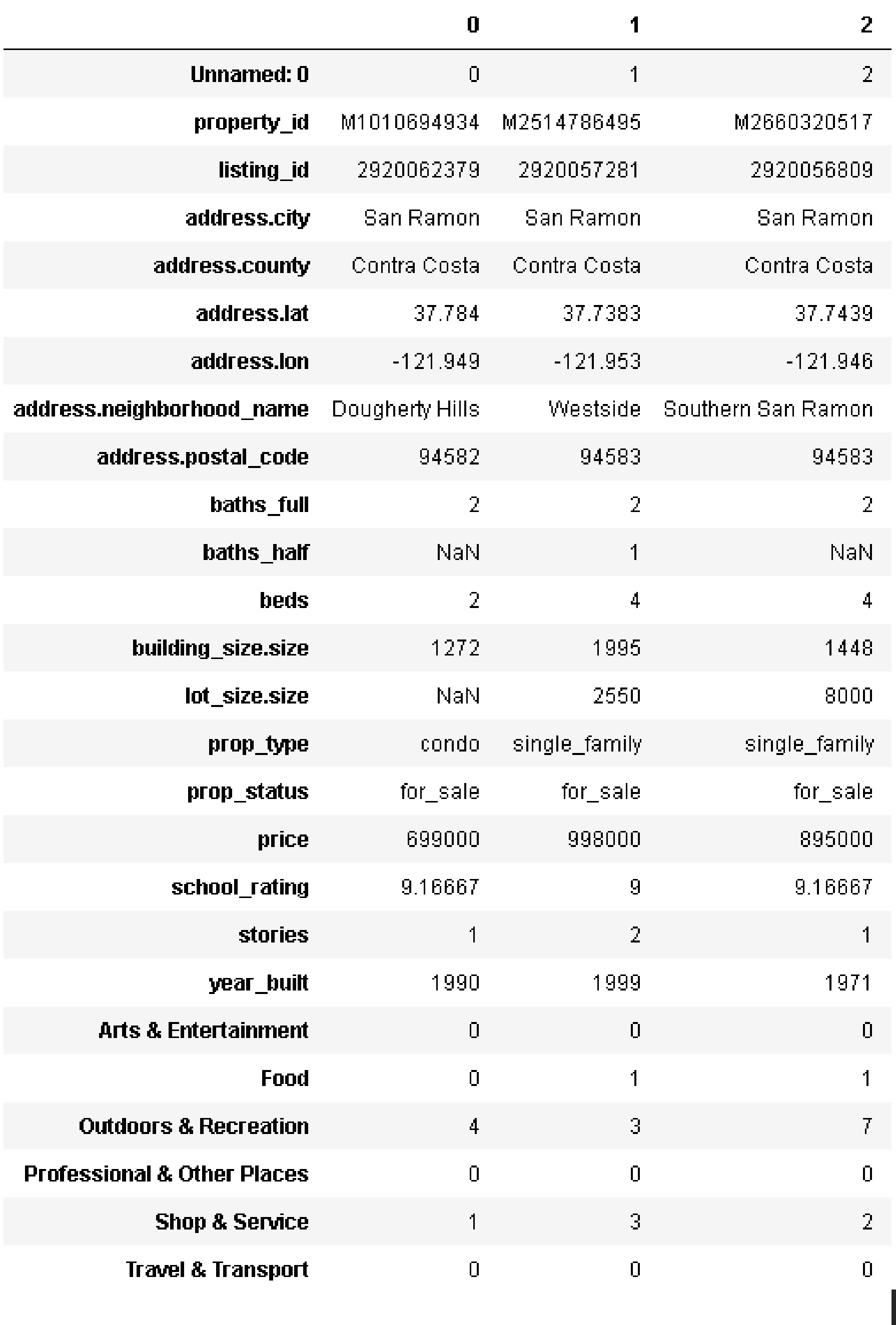
Using Foursquare data, each property got a set of 6 features describing the number of venues for each category in a radius of 1 Km



	Arts & Entertainment	Food	Outdoors & Recreation	Professional & Other Places	Shop & Service	Travel & Transport
property_id						
M1003675301	0	0	5	0	2	0
M1007260231	0	1	3	0	0	0
M1008272648	0	0	3	0	1	0
M1010204240	0	1	3	0	1	0
M1010694934	0	0	4	0	1	0

Final data set

The final dataset joins the information from both sources.



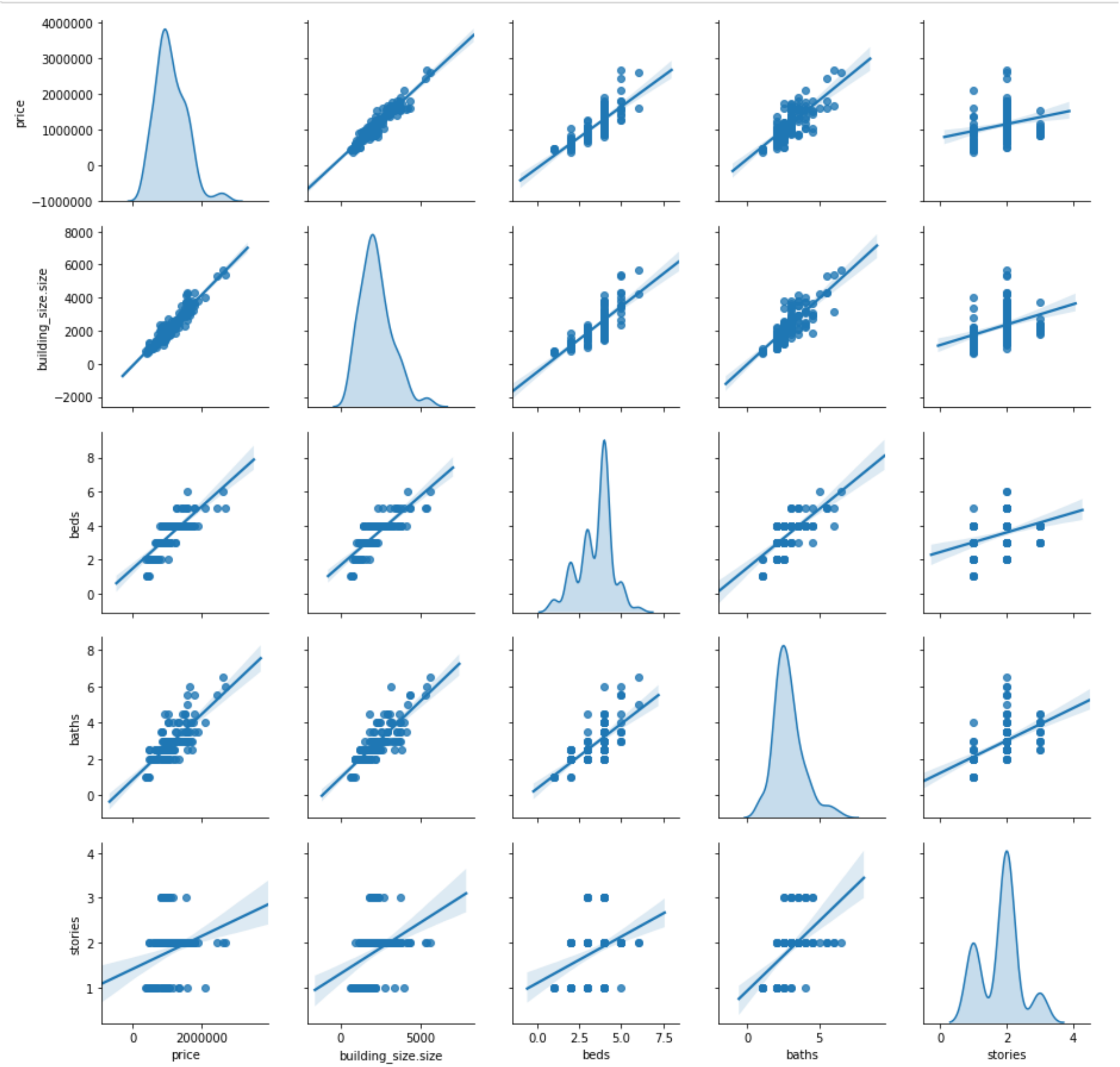
	0	1	2
Unnamed: 0	0	1	2
property_id	M1010694934	M2514786495	M2660320517
listing_id	2920062379	2920057281	2920056809
address.city	San Ramon	San Ramon	San Ramon
address.county	Contra Costa	Contra Costa	Contra Costa
address.lat	37.784	37.7383	37.7439
address.lon	-121.949	-121.953	-121.946
address.neighborhood_name	Dougherty Hills	Westside	Southern San Ramon
address.postal_code	94582	94583	94583
baths_full	2	2	2
baths_half	NaN	1	NaN
beds	2	4	4
building_size.size	1272	1995	1448
lot_size.size	NaN	2550	8000
prop_type	condo	single_family	single_family
prop_status	for_sale	for_sale	for_sale
price	699000	998000	895000
school_rating	9.16667	9	9.16667
stories	1	2	1
year_built	1990	1999	1971
Arts & Entertainment	0	0	0
Food	0	1	1
Outdoors & Recreation	4	3	7
Professional & Other Places	0	0	0
Shop & Service	1	3	2
Travel & Transport	0	0	0

Data Cleansing

- **Lot Size** : Too many NA values to be used. This was the case for the lot size, that presented >30% of missing data.
- **Neighborhood_name:** I will assign the neighborhood of the closest neighborhood.
- **Year Built:** In theory the year of construction of the house is relevant to determine the qualiy of the construction, and might be related to price. That field had 8% of empty values. The most likely
- **School Rating** The missing value in School rating was due to a failure in the API call for the details. I just repeated the call and fixed it manually.
- **County** Not used. Already contained in zipcode.

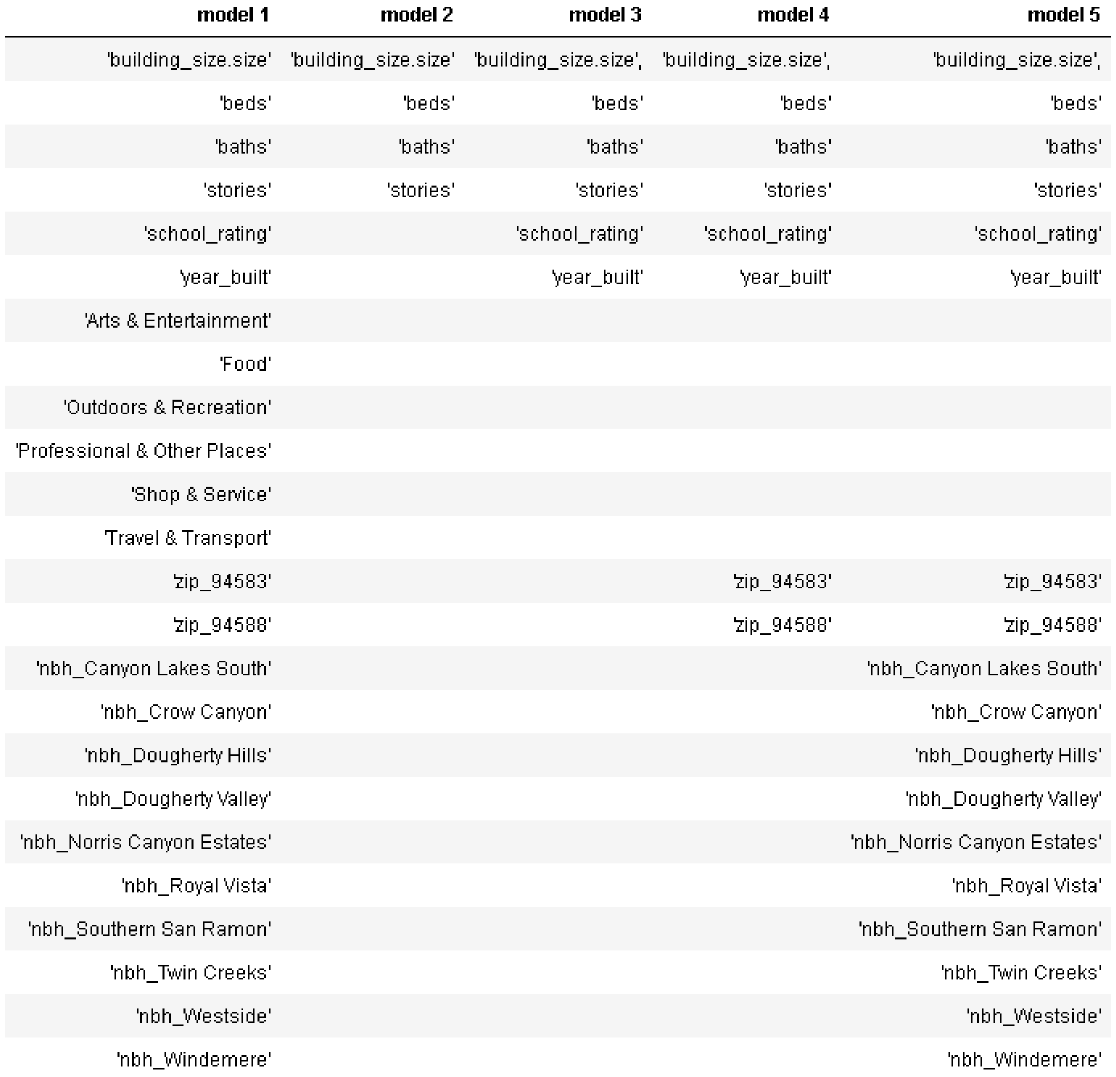
Data Analysis

Several variables are correlated among them. I wan to avoid collinearity in to get meaningful coefficients, and at the same time, I want to have an accurate model.



Feature Selection

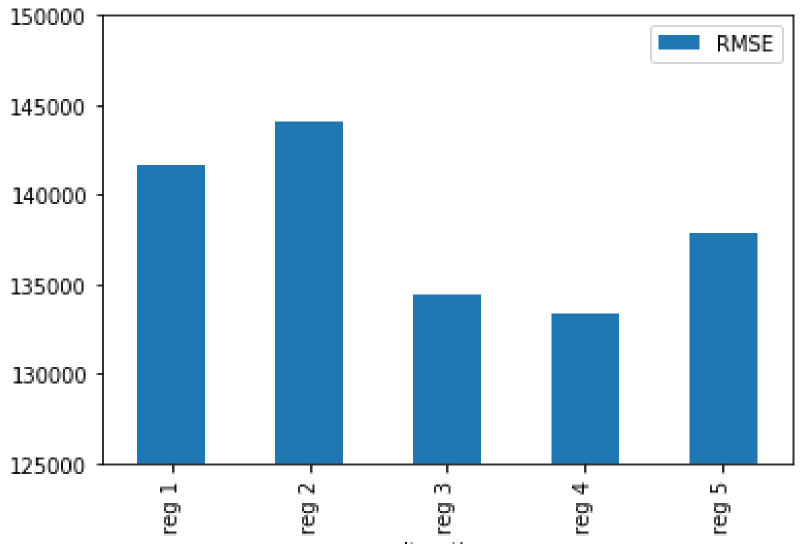
To determine the best set of features I will rum several combinations of features, and compare them based on RMSE and R2.

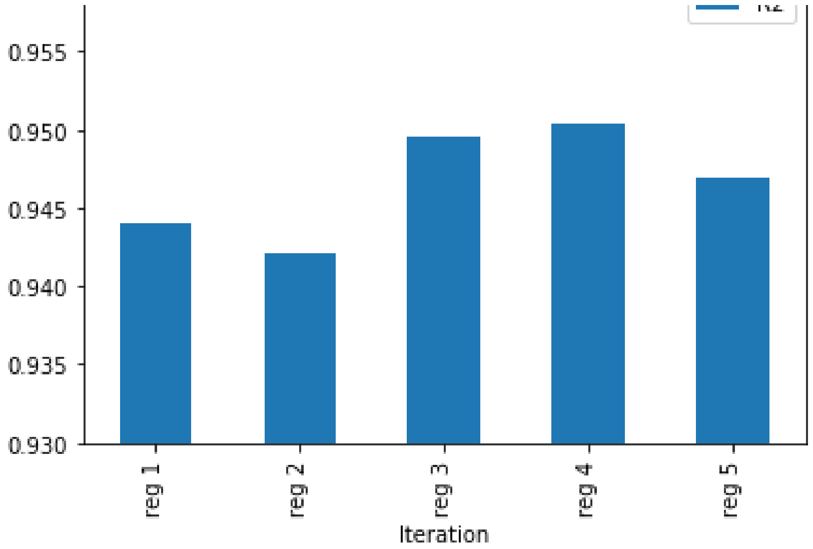


model 1	model 2	model 3	model 4	model 5
'building_size.size'	'building_size.size'	'building_size.size',	'building_size.size',	'building_size.size',
'beds'	'beds'	'beds'	'beds'	'beds'
'baths'	'baths'	'baths'	'baths'	'baths'
'stories'	'stories'	'stories'	'stories'	'stories'
'school_rating'		'school_rating'	'school_rating'	'school_rating'
'year_built'		'year_built'	'year_built'	'year_built'
'Arts & Entertainment'				
'Food'				
'Outdoors & Recreation'				
'Professional & Other Places'				
'Shop & Service'				
'Travel & Transport'				
'zip_94583'			'zip_94583'	'zip_94583'
'zip_94588'			'zip_94588'	'zip_94588'
'nbh_Canyon Lakes South'				'nbh_Canyon Lakes South'
'nbh_Crow Canyon'				'nbh_Crow Canyon'
'nbh_Dougherty Hills'				'nbh_Dougherty Hills'
'nbh_Dougherty Valley'				'nbh_Dougherty Valley'
'nbh_Norris Canyon Estates'				'nbh_Norris Canyon Estates'
'nbh_Royal Vista'				'nbh_Royal Vista'
'nbh_Southern San Ramon'				'nbh_Southern San Ramon'
'nbh_Twin Creeks'				'nbh_Twin Creeks'
'nbh_Westside'				'nbh_Westside'
'nbh_Windemere'				'nbh_Windemere'

The winning model's coeficients

- Built Surface
- Number of Bedrooms
- Number of Baths
- School ratings
- Year Built
- Postal code





Results

Extension	Additional Area	Cost	Est Gain	ROI
Bathroom (addition)	40	$22,000	$31,897	145%
Bathroom (repurpose)	0	$17,000	$17,371	102%
Half Bath (addition)	25	$20,000	$17,764	89%
Second floor, two rooms	700	$110,000	$297,889	271%
New Room	400	$60,000	$206,340	344%

Conclusions

- The listing price of houses depend strongly on the charateristics of the house itself, the postal_code, the rating of schools, but surprisingly not as strongly on the services surrounding them, the concentration of shops, or the closeness to recreation locations.

- Other variables that showed up as weak was the neighborhood, or al least less informative that the zip_code. That doesn no't mean necesarily that it is not important but simply correlated to a other more relevant variable.

- Another finding was that not all real state investemnt produces postive returs. Here the examples of the half bathroom is clearly returning less that the investment needed, and the bathroom built reusing space in the house barely makes the cut. Analyzing the coeficients we can see that many other combinations give negative retunrs, for example adding a second floor is not worth it unless you add enough surface and rooms to the building.

- Clearly, investment in Real Estate depends strongly on the cost of labor, the additioal built surface added, and the rooms added.