An analysis of King County Housing Data

Zaria and Mahima

About Us

Z&M Co. is a residential property agency based in Seattle

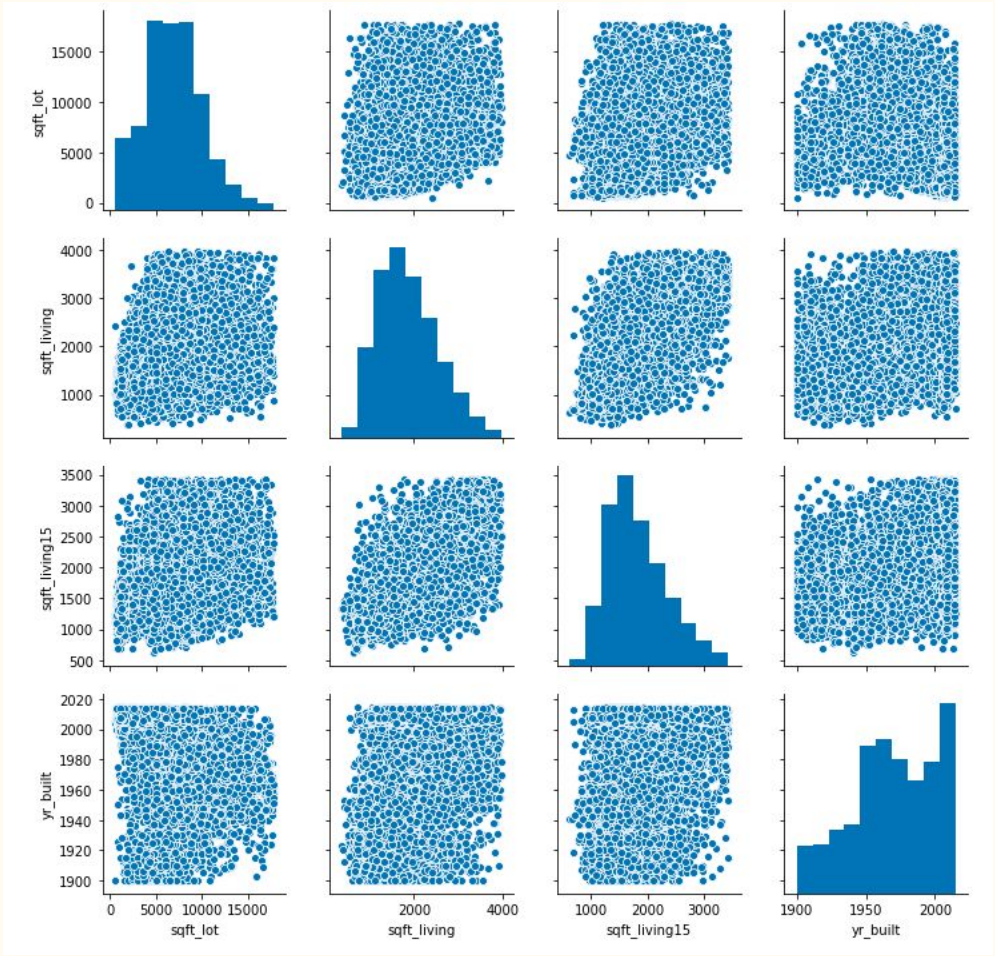
We buy and sell properties in the Greater Seattle area and also offer advisory services

The objective of this presentation is to predict which factors play the strongest role in influencing the housing sale price for our customers

Methodology

Exploratory Data Analysis

- We began by plotting all variables against each other and did not find any significant insights
- So instead of looking at the entire dataset, we decided to focus on houses that appeal to families earning over $130,000
- We also decided to drop ‘waterfront’ as it had significant null values and removed the NaN values from ‘view’


sqft_lot
sqft_living
sqft_living15
yr_built
0
5000
10000
15000
1000
2000
3000
4000
500
1900
1920
1940
1960
1980
2020
1950

Methodology

Identifying the dataset

We decided to subset our data on the basis of houses that would be most marketable to our target segment. We defined most-marketable as:

- Houses of grade 8 and above
- Houses with condition 4 and above
- Houses with >= 2 bedrooms

This gave us 792 properties (3% of the dataset)

Initial Investigation

We set out to investigate three questions from our dataset:

Are properties typically surrounded by similar properties?

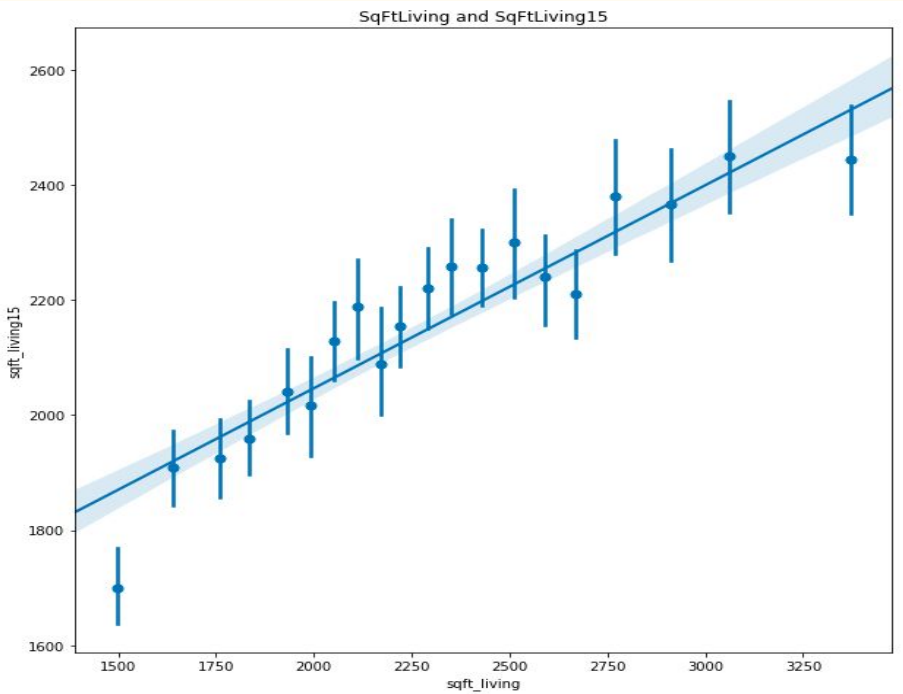
Which zip-codes would be most desirable to our client?

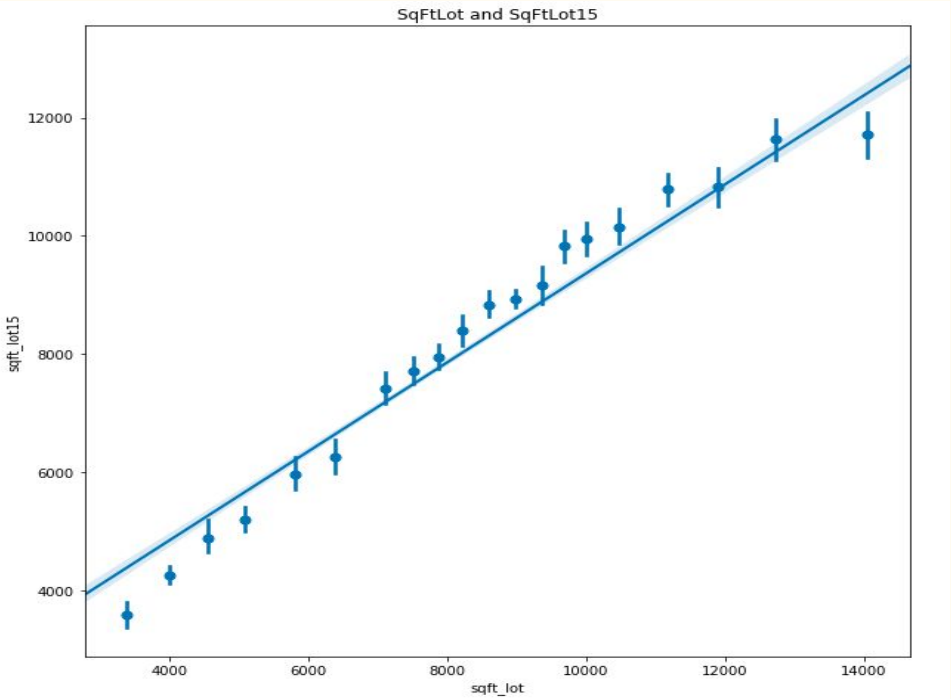
Can you predict the size of living space from the number of bathrooms?

Q.1. Are properties typically surrounded by similar properties?

We decided to compare the size of living space of the property with that of the neighbouring 15 properties

And we did the same for size of lot size of the property and that of the neighbouring 15 properties





OLS Regression Results

Dep. Variable:	sqft_living15	**R-squared:**	0.204
Model:	OLS	**Adj. R-squared:**	0.204
Method:	Least Squares	**F-statistic:**	424.8

OLS Regression Results

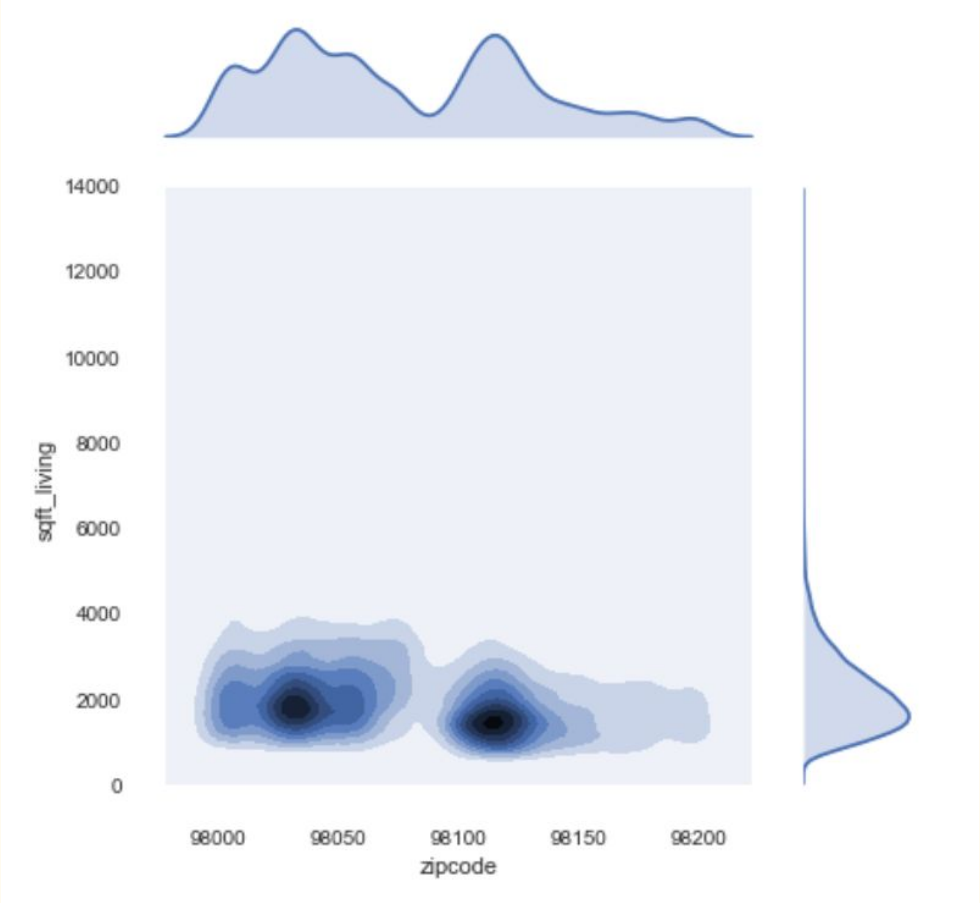
Dep. Variable:	sqft_lot15	**R-squared:**	0.746
Model:	OLS	**Adj. R-squared:**	0.745
Method:	Least Squares	**F-statistic:**	4849.

Q.2. Which zip-codes would be most desirable?

We manipulated our data to group by zip code, longitude and latitude, and totalled how many desirable properties were in each area.

Thereafter, we listed the top three values, which we would consider the most desirable area for our clients to advise our clients on specific areas they may want to move to.

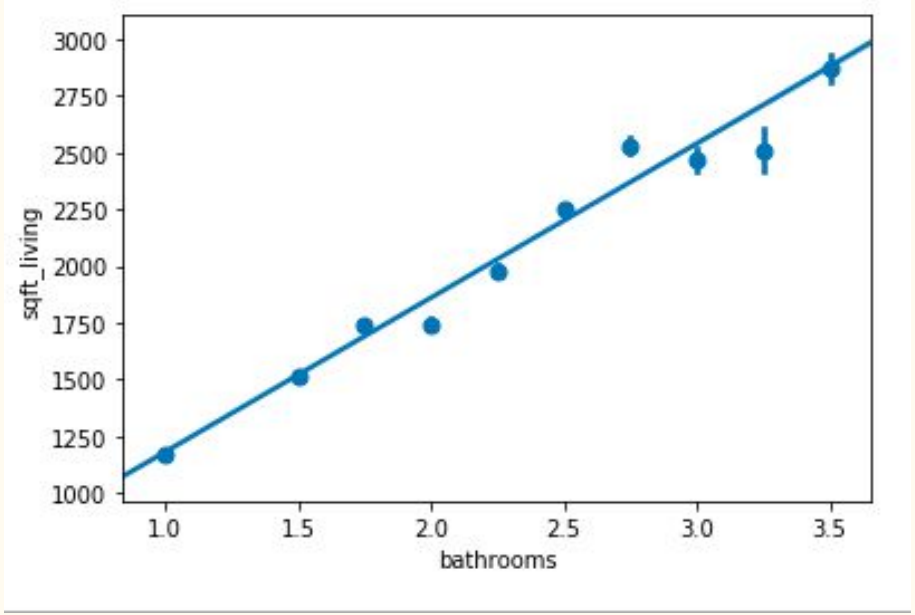
For our top zip-codes, the areas were Bellevue, Redmond and Mercer Island.

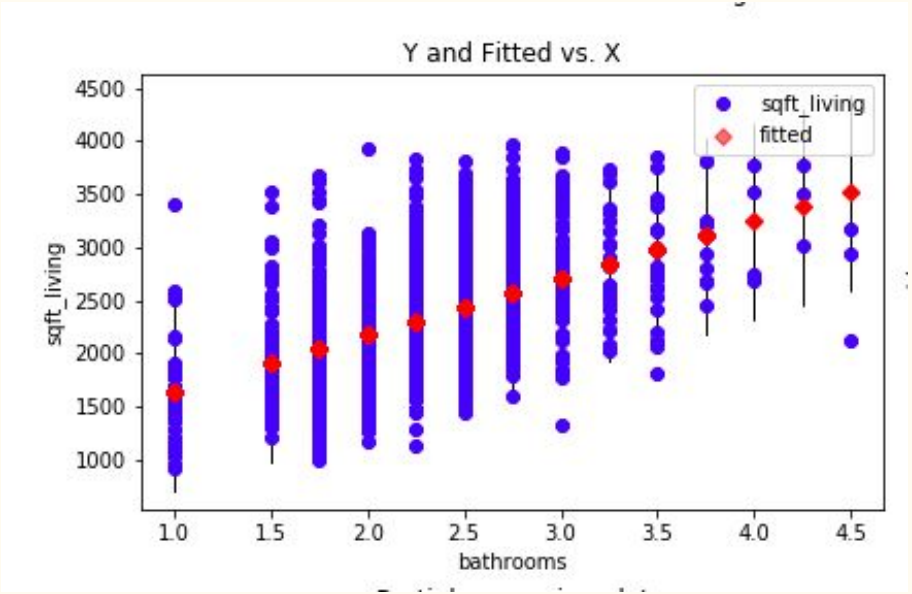

14000
12000
10000
8000
6000
4000
2000
0
sqft_living
98000
98050
98100
98150
98200
zipcode

Q.3. Can we predict living space from the number of bathrooms?

We built a simple model to estimate how many bathrooms could be expected in a property, when the living size was known.

We intended for this model to be used once the ideal number of bathrooms in the new property has been decided, so the clients can see can have an estimate of how many SqFt they can expect from the property. Should this number be inadequate, they could see how much more or less living space they would have to compromise to reach a satisfactory number of bathrooms.


sqft_living
bathrooms
3000
2750
2500
2250
2000
1750
1500
1250
1000
1.0
1.5
2.0
2.5
3.0
3.5


Y and Fitted vs. X
sqft_living
fitted
sqft_living
bathrooms
4500
4000
3500
3000
2500
2000
1500
1000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5

Building a Price Estimator Model

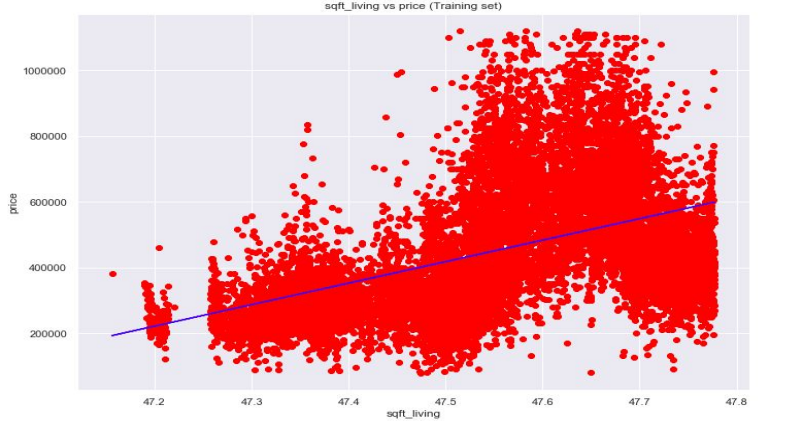
Methodology

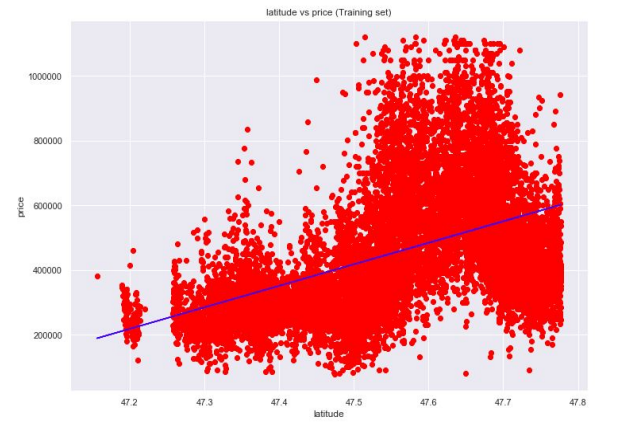
As our initial objective was to find which factors have the strongest influence on price, we decided to explore ‘living size’ and ‘latitude’ for the entire dataset

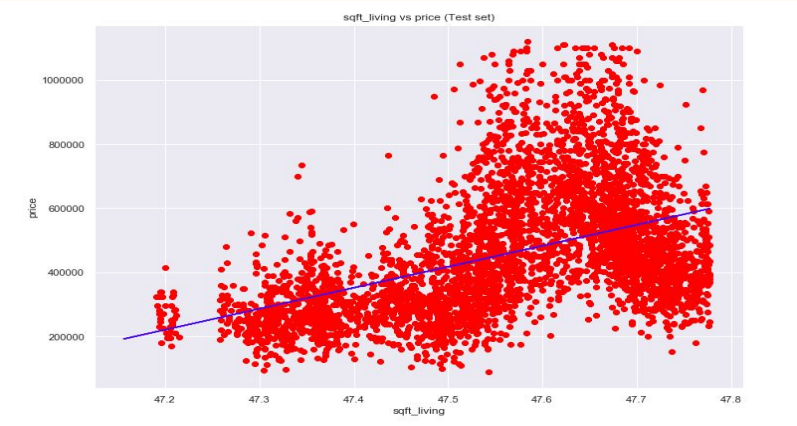
These were found to have the strongest correlation with price (sqft_living = 0.4) and (latitude = 0.5) without running the risk of multicollinearity (sqft_living and latitude = 0.06)

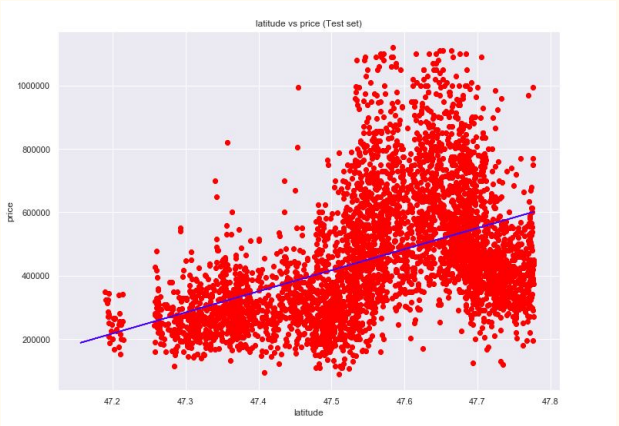
We then split the data into training (80%) and testing(20%)

Our Model


sqft_living vs price (Training set)
price
sqft_living
1000000
800000
600000
400000
200000
47.2
47.3
47.4
47.5
47.6
47.7
47.8


latitude vs price (Training set)
price
latitude
1000000
800000
600000
400000
200000
47.2
47.3
47.4
47.5
47.6
47.7
47.8


sqft_living vs price (Test set)
price
sqft_living
1000000
800000
600000
400000
200000
47.2
47.3
47.4
47.5
47.6
47.7
47.8


latitude vs price (Test set)
price
latitude
1000000
800000
600000
400000
200000
47.2
47.3
47.4
47.5
47.6
47.7
47.8

Conclusions and recommendations

Our findings revealed that both variables have somewhat of an impact on actual price i.e. they could be used to increase the sale price of a house or making a property more marketable to our target audience

1. Adding to the square footage through an additional room / bathroom / floor would increase the value of the property

2. Houses with latitudes 122.1277, 122.1182 and 122.2331 could command more price hence this will affect buyers looking to sell their properties in the medium to long run

   Additionally, changing the orientation of a property (south / west facing) would make it more energy-efficient leading to less amount spent on heating / air-conditioning makes it more valuable

Thanks for listening!

Appendix

R2 values

Coefficient

Intercept

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.318
Model:	OLS	**Adj. R-squared:**	0.318
Method:	Least Squares	**F-statistic:**	8078.
Date:	Wed, 23 Oct 2019	**Prob (F-statistic):**	0.00
Time:	14:21:56	**Log-Likelihood:**	-2.3199e+05
No. Observations:	17288	**AIC:**	4.640e+05
Df Residuals:	17286	**BIC:**	4.640e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.485e+05	3664.061	40.539	0.000	1.41e+05	1.56e+05
sqft_living	166.3940	1.851	89.876	0.000	162.765	170.023

Omnibus:	856.091	**Durbin-Watson:**	1.984
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	986.940
Skew:	0.579	**Prob(JB):**	4.88e-215
Kurtosis:	3.177	**Cond. No.**	5.86e+03

SqFt Living

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.215
Model:	OLS	**Adj. R-squared:**	0.215
Method:	Least Squares	**F-statistic:**	4741.
Date:	Wed, 23 Oct 2019	**Prob (F-statistic):**	0.00
Time:	14:21:56	**Log-Likelihood:**	-2.3321e+05
No. Observations:	17288	**AIC:**	4.664e+05
Df Residuals:	17286	**BIC:**	4.664e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.103e+07	4.57e+05	-67.853	0.000	-3.19e+07	-3.01e+07
lat	6.62e+05	9615.042	68.856	0.000	6.43e+05	6.81e+05

Omnibus:	1834.605	**Durbin-Watson:**	1.978
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2522.332
Skew:	0.856	**Prob(JB):**	0.00
Kurtosis:	3.755	**Cond. No.**	1.64e+04

Latitude