
Housing Data

Zaria and Mahima

Who Are We?

Z&M Co. is a property agency in Seattle, catering to upper middle class and high income families (>\$130,000)

The aim of this presentation is to predict housing sale price for this target group in suburban Seattle



Initial Investigation

We began with the assumption that house price would be positively correlated with:

- Views
- Condition
- Size

However, we didn't find particularly meaningful relationships, as we were trying to use all of the data at once.

Subsetting the Data

We decided to subset our data focusing on houses most marketable to upper-middle class families ($HI > \$150,000$). This allowed us to make more accurate observations of the data, and allowed us to refine our analysis of the data.

We used a function to determine outliers in our data, and removed them from our new dataset, as well as removing any columns we didn't think added great value.



Marketable:

We defined the most-marketable properties for our demographic as:

1. Houses graded 8 and above
2. Houses with a condition of 4 and above
3. Houses with 2 or more bedrooms

Further Exploration

We set out to investigate three questions from our dataset:

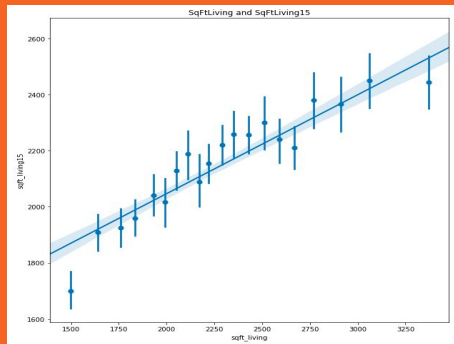
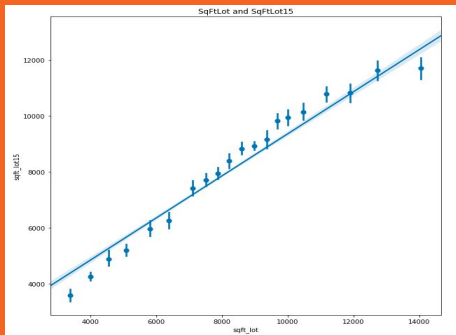
Are properties typically surrounded by similar properties?

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

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

OLS Regression Results

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



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

Question 1

Are properties typically surrounded by similar properties?

We decided to compare SqFtLiving and SqFtLiving15, and SqFtLot and SqFtLot15.

We noticed correlation on both graphs, but much higher with the size of the lot. This was reinforced when we compared the R^2 values of the two variables.

This could suggest people typically buy houses surrounded by houses on a similar plot size, although the living size of each property varies



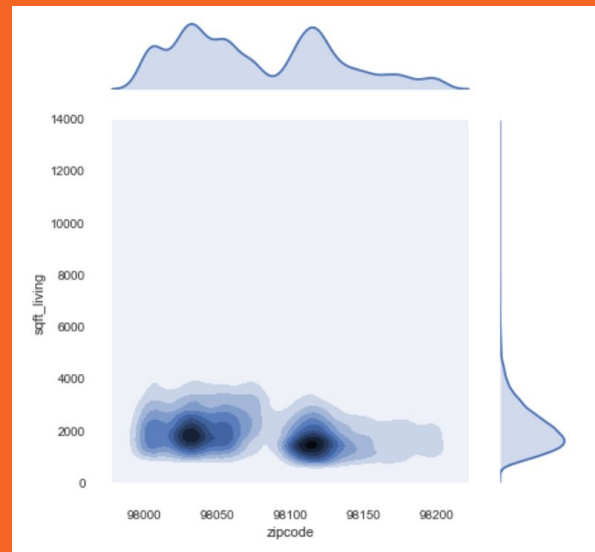
Question 2

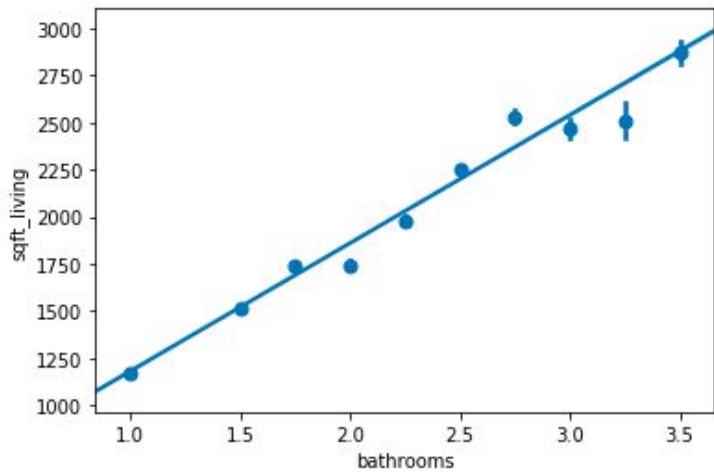
Which zip-codes would be most desirable to our client?

We noticed that some zip-codes had a higher frequency of large houses than others, which led us to question which zip-codes would we recommend for our clients.

We manipulated our data to group by postcode, and totalled how many desirable properties were in each area. We listed the top five values, which we would consider the most desirable area for our clients.¶

Our findings showed that the most desirable areas were 98006, 98052, 98040, 98023 and 98008, which are located in *XYZ*..



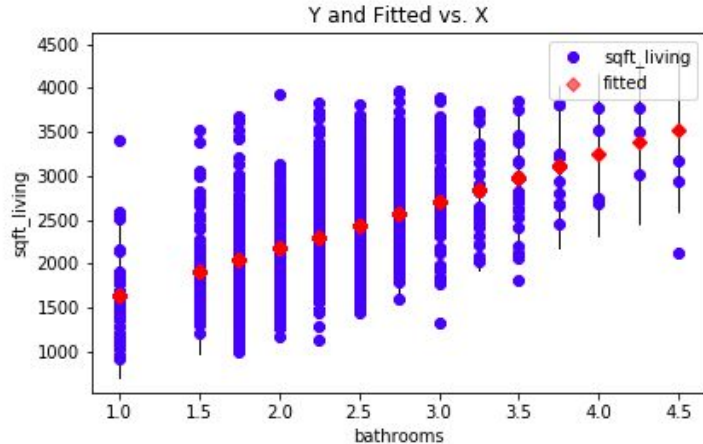


Question 3

Can you predict the 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 SqFtLiving value was known.

We intended for this model to be used once the ideal SqFt of the new property has been decided, so the clients can see how many bathrooms they can expect from the property. Should this number be inadequate, they could see how much the living space would need to increase before they have a satisfactory number of bathrooms.



Building Our Linear Model

We built a linear model to optimise the price of a property, based

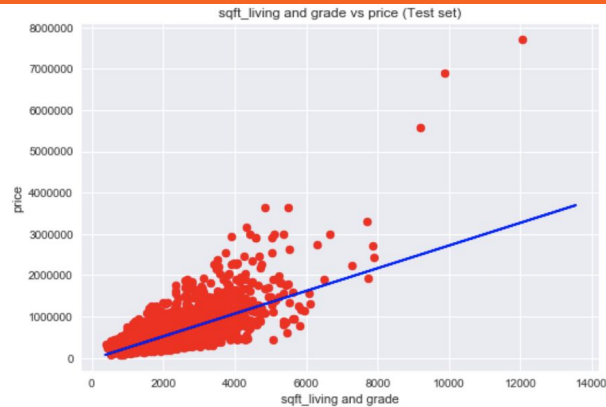
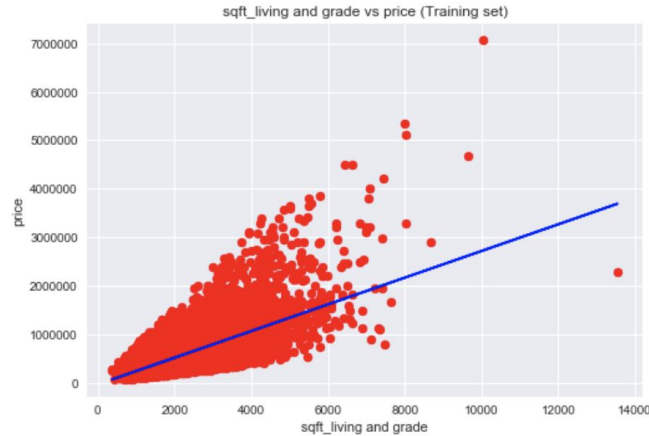


Question 3

Which variables have an impact on price?

We considered `sqft_living` and `grade` as the two variables which had highest correlation with price and were not correlated with each other.

We tested the linear regression between these variables to determine their relationship with price. Our findings revealed that in actual fact there is a significant variation between living size and grade of a house with the price. Hence, while these variables have an impact on the actual price, it is not significantly strong.





Thank You!