Module 1 Project:

Kings County Housing Dataset

By Joe and Greg

What we set out to understand and why

We wanted to maximise profits by seeing which variables will have the biggest effect on price.

1. Does the location affect the price?
2. Does the effect of each variable vary with location? IE Do neighbourhoods prioritize different property features?
3. What is the best combination of these features to maximise property value?

Our client will be buying existing properties or empty lots and altering them based on our findings.
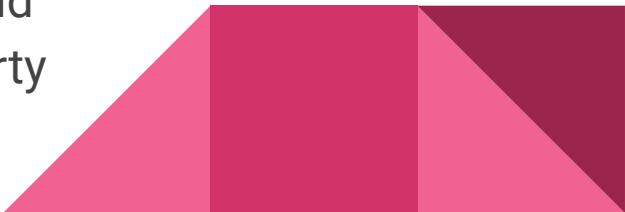
Cleaning the data

Some interesting features couldn't be examined as the data was incomplete:

- Waterfront view: 147 properties listed, many more in reality
- View grade: many listed as 0 so median value was used
- Grade: this is comprised of our other variables (high correlation)

What we fixed:

- Changed column names for clarity and accurate representation.
- Basement null values -> sqft "not basement" - sqft living
- 33 bedroom house -> based on size we could deduce this was a typo -> 3 bedroom property
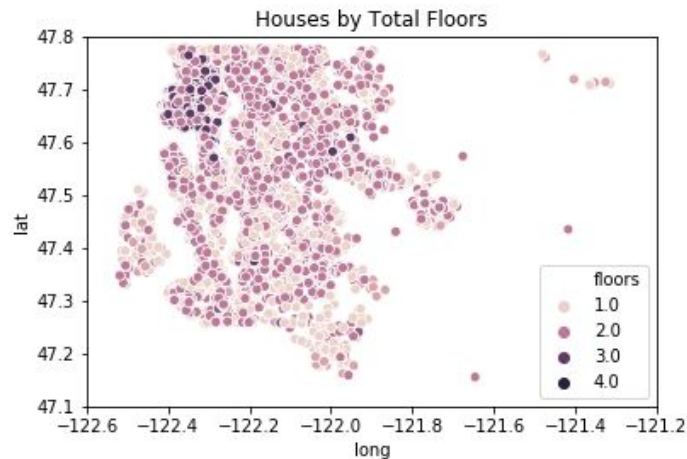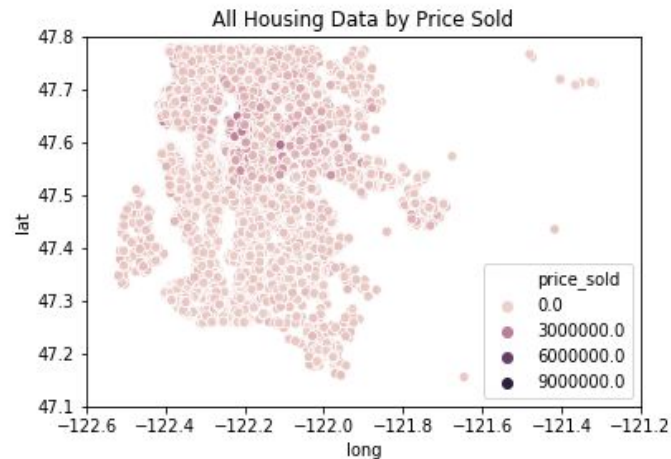
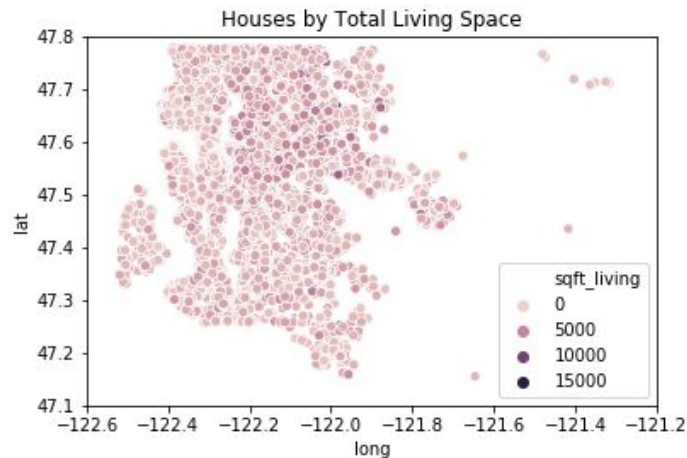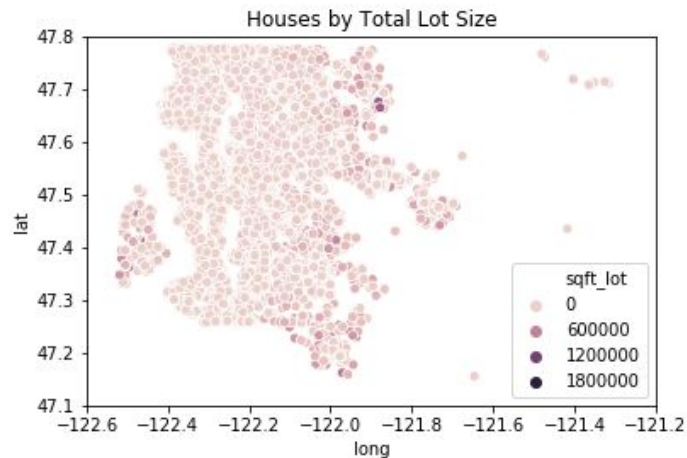# Does location affect the house price?
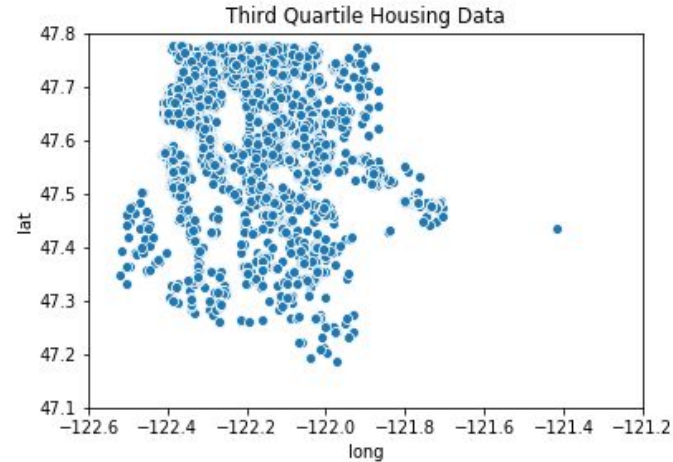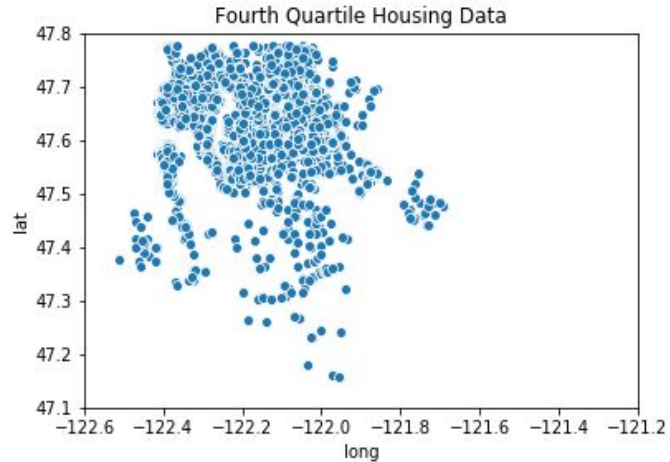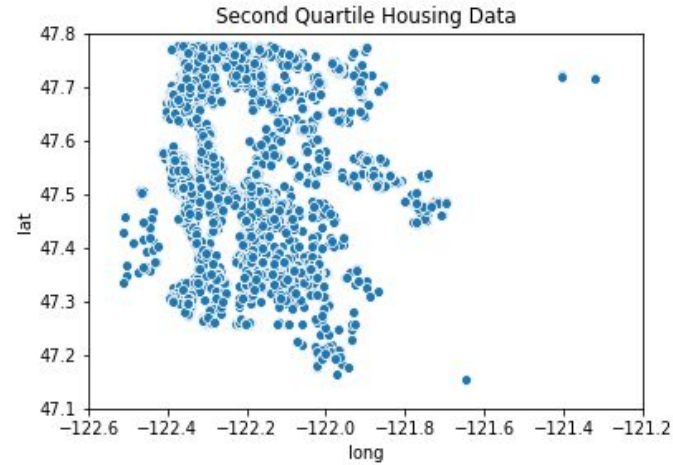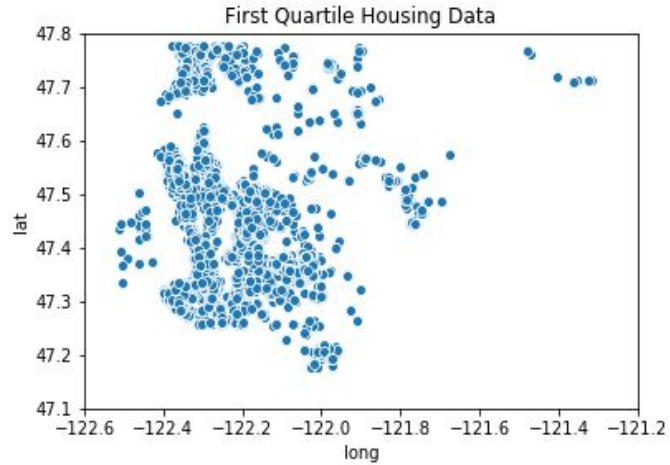
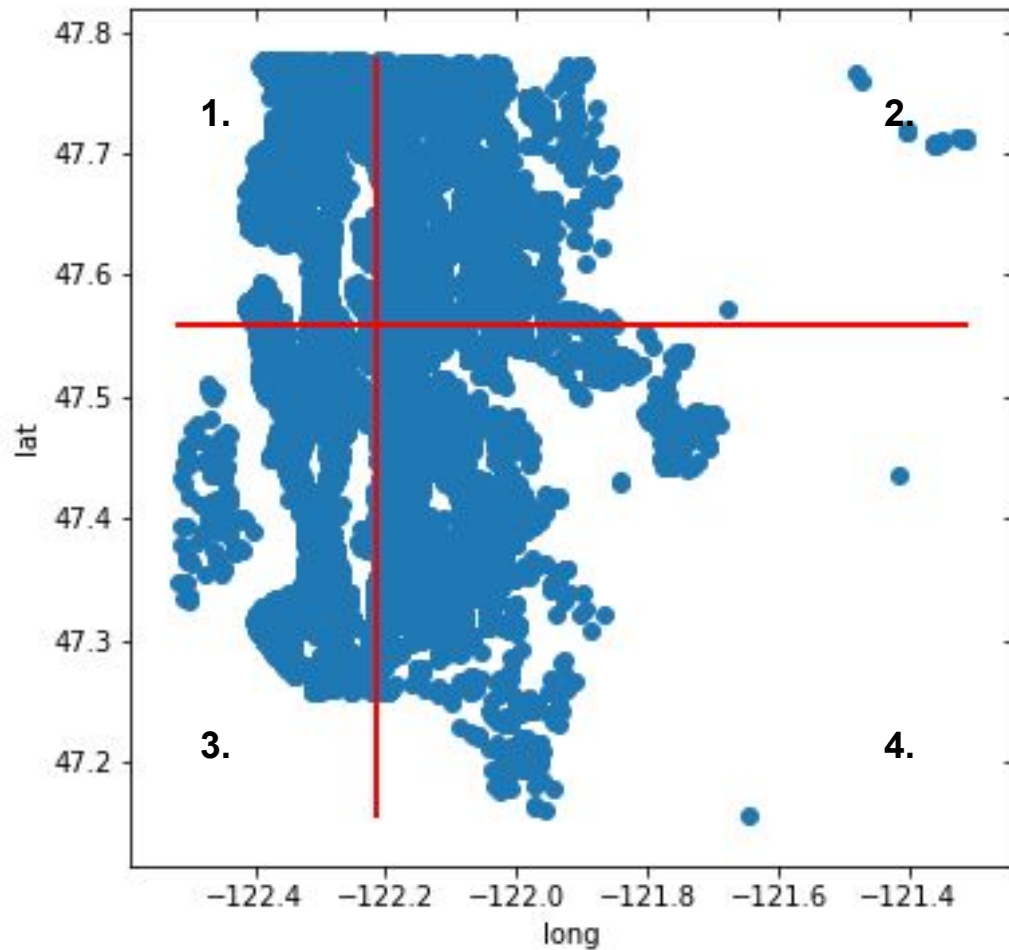Q1 < $322,000

Q1 < Q2 < $450,000

Q2 < Q3 < $645,000

$645,000 < Q4

How else can we categorise the neighbourhoods?
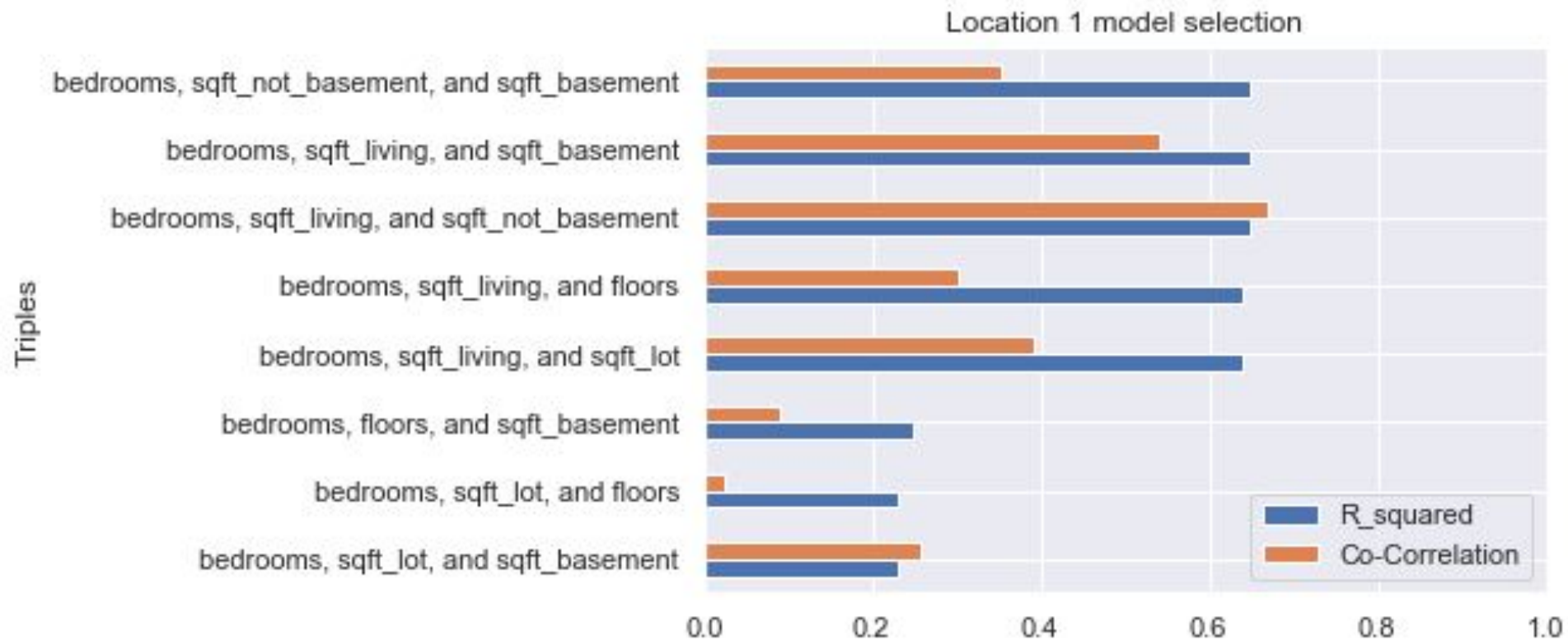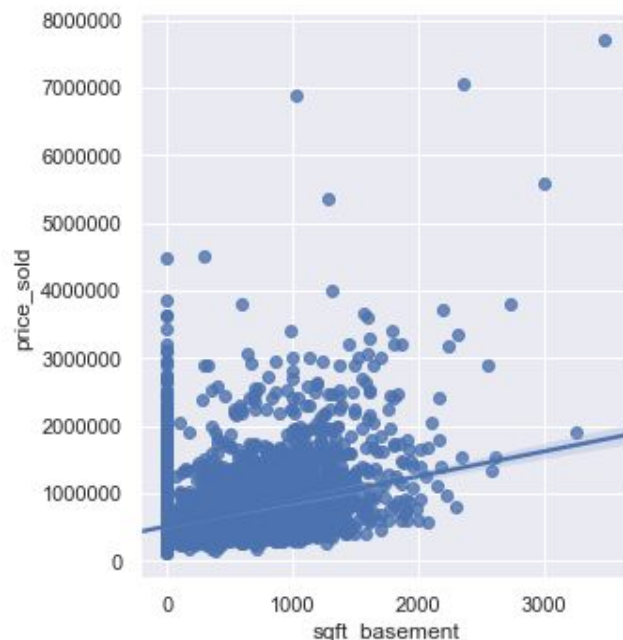

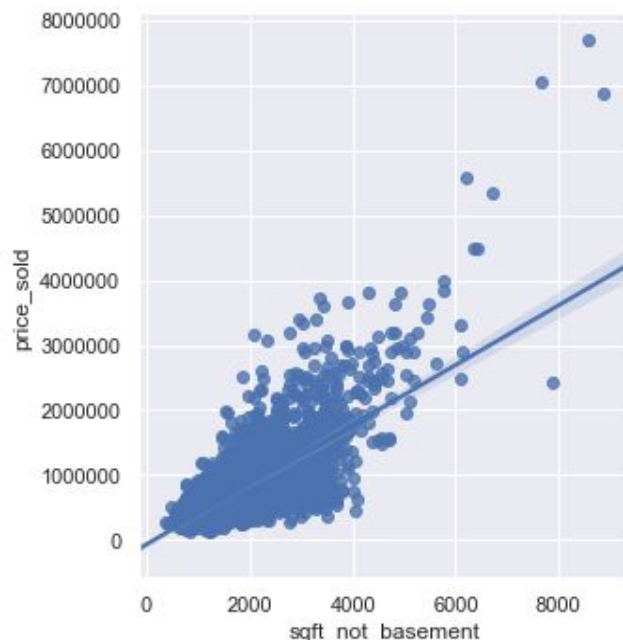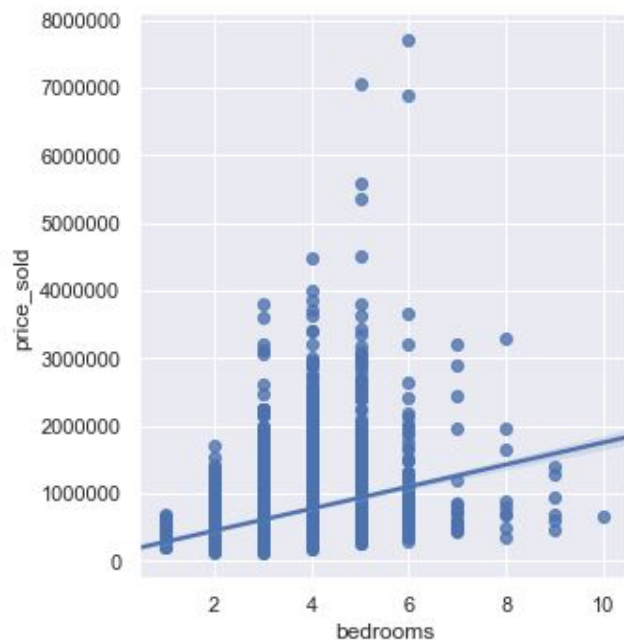Can we group by number of floors?

Roughly split into regions with same number of properties in each quadrant.

Will this allow us to more accurately predict the effect of the variables by looking in a more granular way.

What are the best three variables to choose?

Location 1 model selection

Seaborn linear regression for individual variables

Linear regression for our 'best' variable trio

R2 = 0.65

Not good!

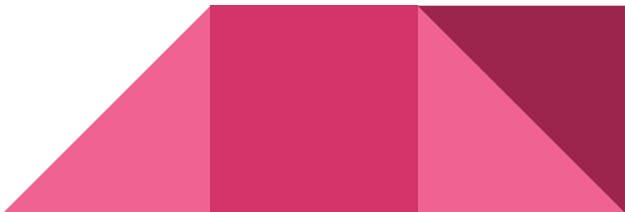const   30900

bedrooms   -85000

Even accounting for

sqft_not_basement   492

co-correlation (simply)

sqft_basement   332

What can we conclude?

- Clear split in property prices with location.

- Bigger lot size or more floors does NOT result in greater property value.

- Key factor is living space -> Focus on this predictor.

- The accuracy of each predictor varies with location.

- Model still gave strange results even after checking for co-correlation

- Need to be more careful in initially selecting variables

Thanks for listening!

Any questions?