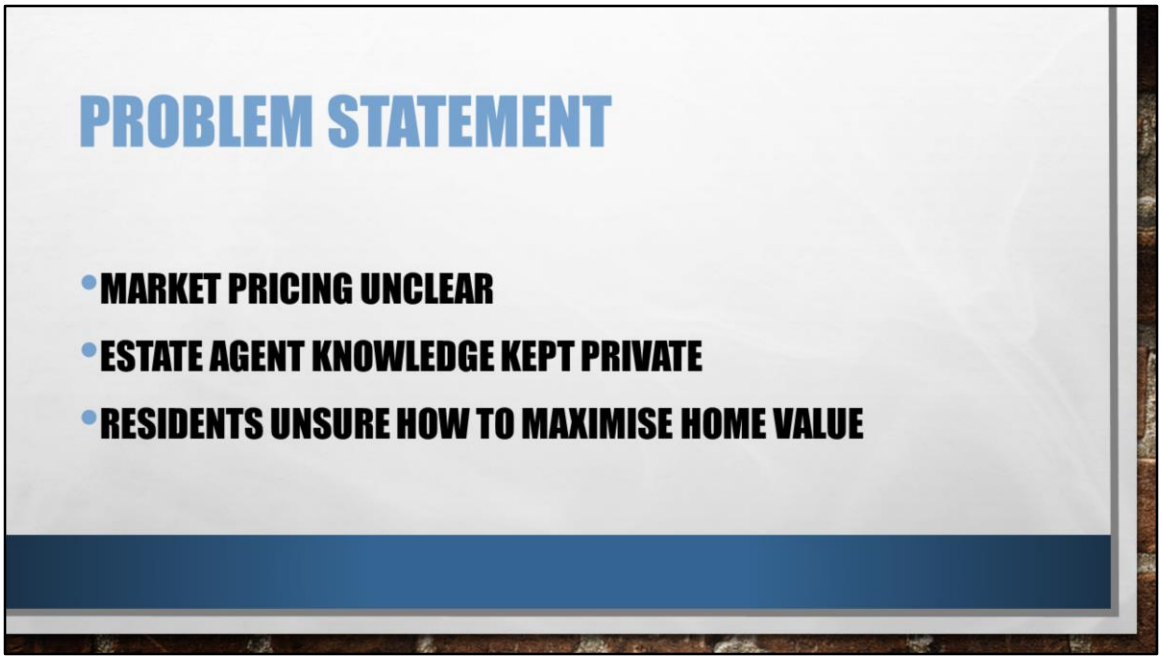
Hello, my name is Matthew Freeman and today I will be talking to you about how I have been using machine learning techniques to investigate house prices in King County and the best ways you can improve your house.



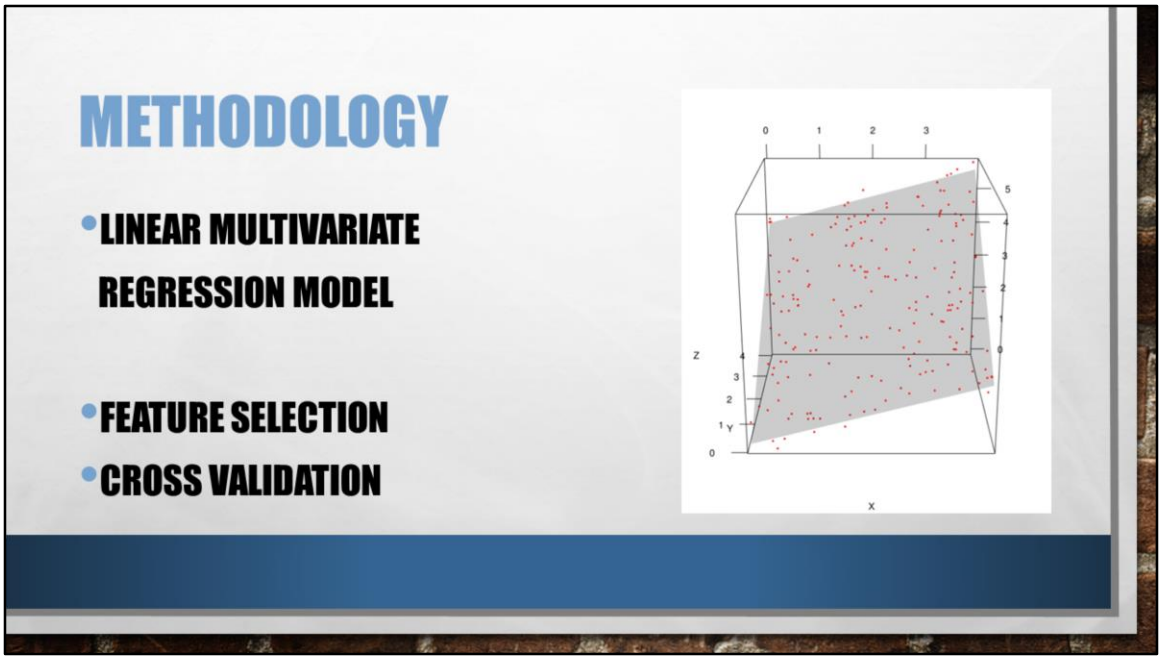
Residents in King County, like everywhere, do not know exactly how and why homes are priced the way they are.
You may want to sell your house, or are considering it, but don't know how the price has changed since you acquired it.
Estate agents may know, but their dark arts are a mystery to normal people. Seeking their help can be expensive and time consuming.
You may also know that you can improve your home's value with a calculated investment but don't know what would be the best use of money.

We want to solve these problems by:
1) Simplifying the pricing mechanism so that residents know what their house's value is and why.
2) Increasing public understanding of what prices houses across the county are selling for and what is behind their price.
3) Make sure residents can find out what the best ways are to maximise their homes value.

I believe we can reach these solutions and this is how I have been able to do that so far. Firstly, you must understand that a big database has been publicised. This database includes details on over 21,000 house sales between May 2014 and June 2015 within King County across all areas.

These details include things like price sold at, square footage of house, number of floors, whether is has a waterfront view, how many times it was viewed, which zip code it is in, when it was built and much more.

This much data is very useful and modern big data methods can be applied to this in interesting ways. This is exactly what I have done for you.



I have created a “linear multivariate regression model” which is technically a kind of machine learning method.

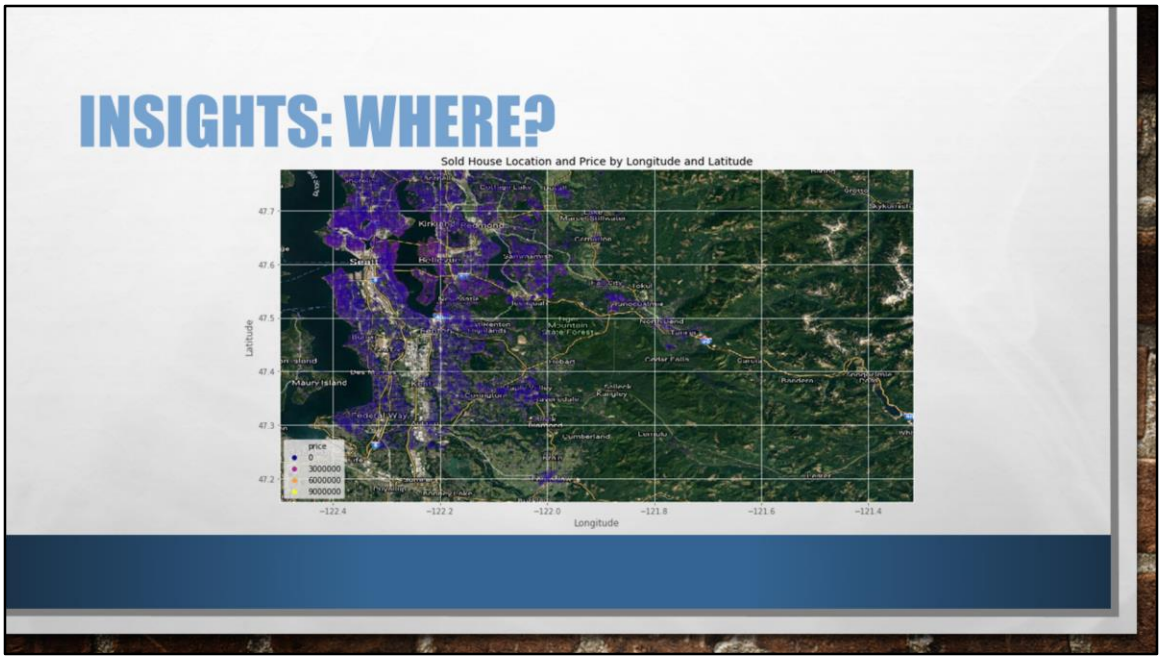
Linear multivariate regression model is a complicated word but what it means is that we have taken all the details of these houses, such as number of floors, and tried to draw a line of best fit against the sale price of the house. It is a bit like this image here where a statistical relationship between, for example, the number of floors on the x-axis, and price on the y-axis can be determined with some simple mathematical formula- except that we are doing this in many dimensions at the same time so instead of a “straight line” of best fit, it is a multidimensional shape.

We can then find out how all these relationships fit together and influence price, not just for a sample house from our data, but for any house in general.

I have also used what is called “feature selection” to adapt the linear regression to work best with a slightly modified choice of which house details to use. This helps to make sure that the details we are putting into the model do not have an undue bias on our result. For example, if we have both square footage of the house, and square footage of living space as details going into our model we may end up with a result

which counts square footage multiple times, and this needs to be managed.

I have also employed k-folds cross-validation in my model to ensure that my model has avoided over or under-fitting the data set- essentially I have made sure lines of best fit have not been drawn to show relationships which may only exist within the sample of data available to me.
I have also used other techniques of machine learning which can be explained later if necessary.



I shall now begin explaining some of the insights gained from creating this model. These are only SOME of the insights which are available to me and were gained alongside the main aim of the model, during the process of creating it.

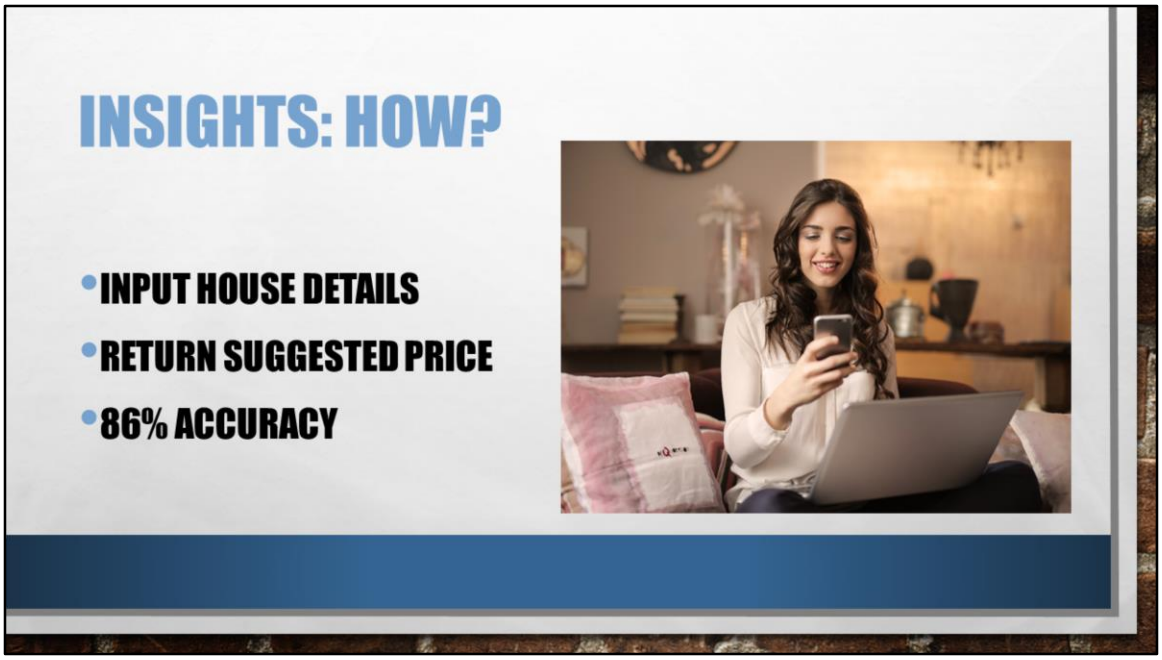
Here is a picture of the area around King County. We can see the points in blue are lower value houses and warmer colours are higher value houses. These points also represent every house sale included in the database, and thus also my model.

We can see from this that the highest priced houses are selling around Bellevue and waterside central north Seattle. I have also analysed details of house sale prices by zip code and found zip codes 98119 and 98109 to have houses selling for more than they would be worth based on their features alone.

I also found that houses with twice as many bathrooms were 15% more valuable, and houses which were renovated were 6% more valuable.

This is of course, on average and does not take into account how old the houses were before being renovated but it could identify ways of improving your house value.

I have also found that for every “government-assigned grade” you go up (this is a scale of 1-13 which is assigned to every house based on its condition and quality) the house price on average goes up by 18%.



You can gain more information on how much prices can be increased by changing features of your house by using my model or asking for further analysis from me.

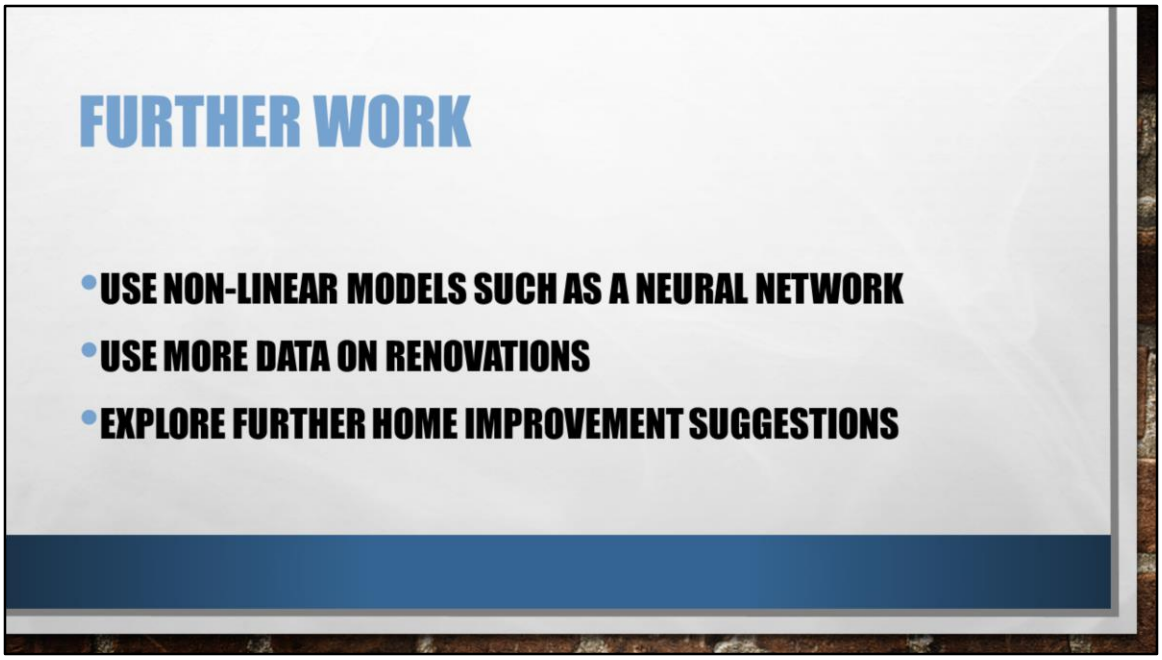
You can also employ my model for it's main purpose, which is of course to predict the price of your house should it enter the market.

I have tested my model and found it to be 86% accurate. What this means, is that if you enter details about your house into my model I can tell you what price your house can sell for, with only 14% discrepancy being unexplained in my model.

If you want to know more, please look up the R-squared measure, which is what this number represents. I have also checked my model against other statistical tests.

How should you use this information? Well for starters you should definitely use my model to help you value your house.

I can also find statistics such as those mentioned earlier which imply some increases in value you can expect by say, adding more bathrooms, renovating your house, or getting a higher grade confirmed to your house. You can use those statistics to help you decide on what investments are best to undertake by the return they may yield.

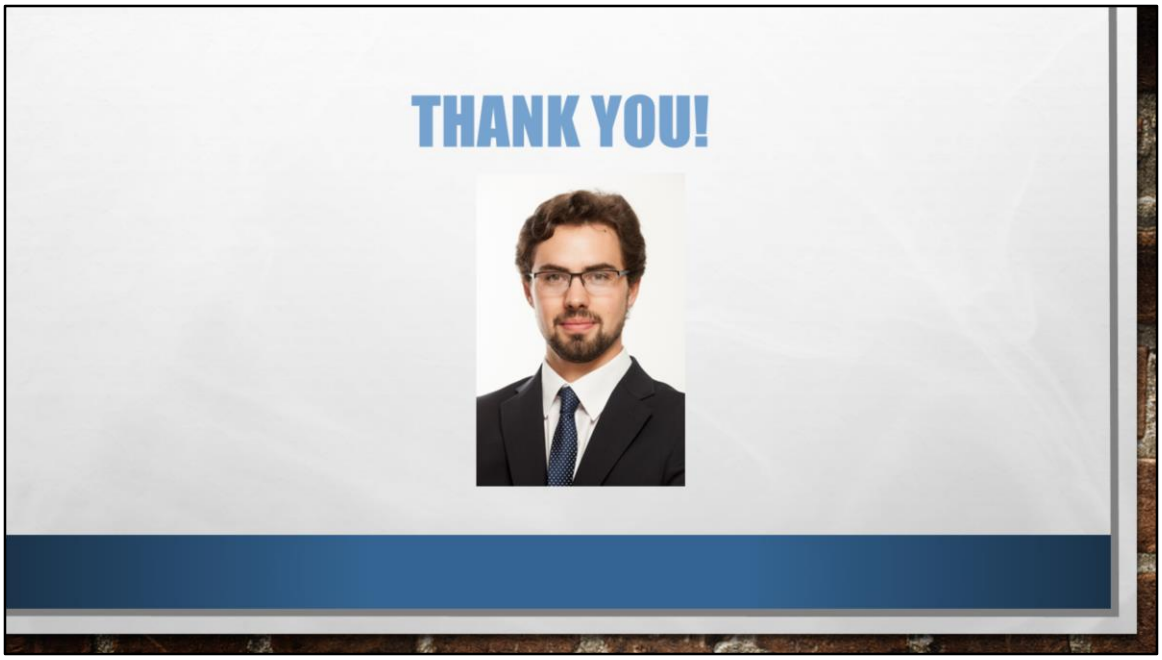


In the future I could also undertake more work on my models. I could adapt my model to be non-linear one, such as a neural network, a kind of advanced modern machine learning model similar to a human brain. This would allow more features to be used effectively, to give a more precise price estimate.

By “non-linear” I am referring to the line of best fit used in my modelling. This is currently always a straight line for each feature, which is a draw back because it uses average relationships across all houses. For example there is a big difference in price between houses one mile from the centre of Seattle and two miles out, and a small difference between houses 10 miles out and 11 miles out, but my model can only estimate the price using the average of the price-distance relationship.

I could also improve my model if I have more data on house renovations or other limited data sets, where not many people have renovated their house or have done so in small ways not represented in the data.

I could also find more suggestions on relationships between house features and house prices which lend themselves to quantifiable house improvement suggestions. There are more available from my model right now.



That is all I have to say today, thank you very much for listening! I am happy to any other questions now.