PROJECT 2:

Multilinear Regression Analysis of King County Housing Data

AGENDA

1. Introduction
2. The Dataset
3. The Approach
4. Data Modelling
5. Conclusions

INTRODUCTION

The Business Problem

This project analyses King County housing data, to provide information for potential buyers and sellers. Through multilinear regression analysis, we aim to determine what factors affect house pricing in the area.

This analysis will provide buyers and sellers with more information on the market. What factors are most valuable in increasing the value of your house.

THE DATASET

The King County Dataset provides information on the following variables:

- id - unique identified for a house
- Date - house was sold
- **Price** -prediction target
- bedroomsNumber - number of bedrooms per house
- bathroomsNumber - of bathrooms/bedrooms
- sqft_livingsquare - footage of the home
- sqft_lotsquare - footage of the lot
- floorsTotal - floors (levels) in house
- waterfront - House which has a view to a waterfront
- view - Has been viewed
- condition - How good the condition is (Overall)
- grade - overall grade given to the housing unit, based on King County grading system
- sqft_above - square footage of house apart from basement
- sqft_basement - square footage of the basement
- yr_built - Built Year
- yr_renovated - Year when house was renovated
- zipcode - zip
- lat - Latitude coordinate
- long - Longitude coordinate
- sqft_living15 - The square footage of interior housing living space for the nearest 15 neighbors
- sqft_lot15 - The square footage of the land lots of the nearest 15 neighbors

THE APPROACH

So how do we determine how much each factor affects house price?

The answer is **multilinear regression analysis:**

Multilinear regression analysis is a method used to understand how multiple factors influence an outcome. It helps us identify which factors matter most and how much they affect the result. This information is valuable for home owners and potential buyers to make informed decisions.

Essentially we take all the data, and provide a line of best fit. We will go through 3 iterations of our model, to improve it's reliability and robustness.

DATA MODELLING

Iteration 1

- After initial cleaning of the data, we get our first regression model.
- Our first iteration produces a good adjusted R squared value of **0.66.** Showing there is a fairly high correlation between all the data and the price of housing. Essentially, 66% of the housing price can be explained by our variables.
- This is a great start, but are lot of variables that have a high p value, which are not likely to be adding to the model.

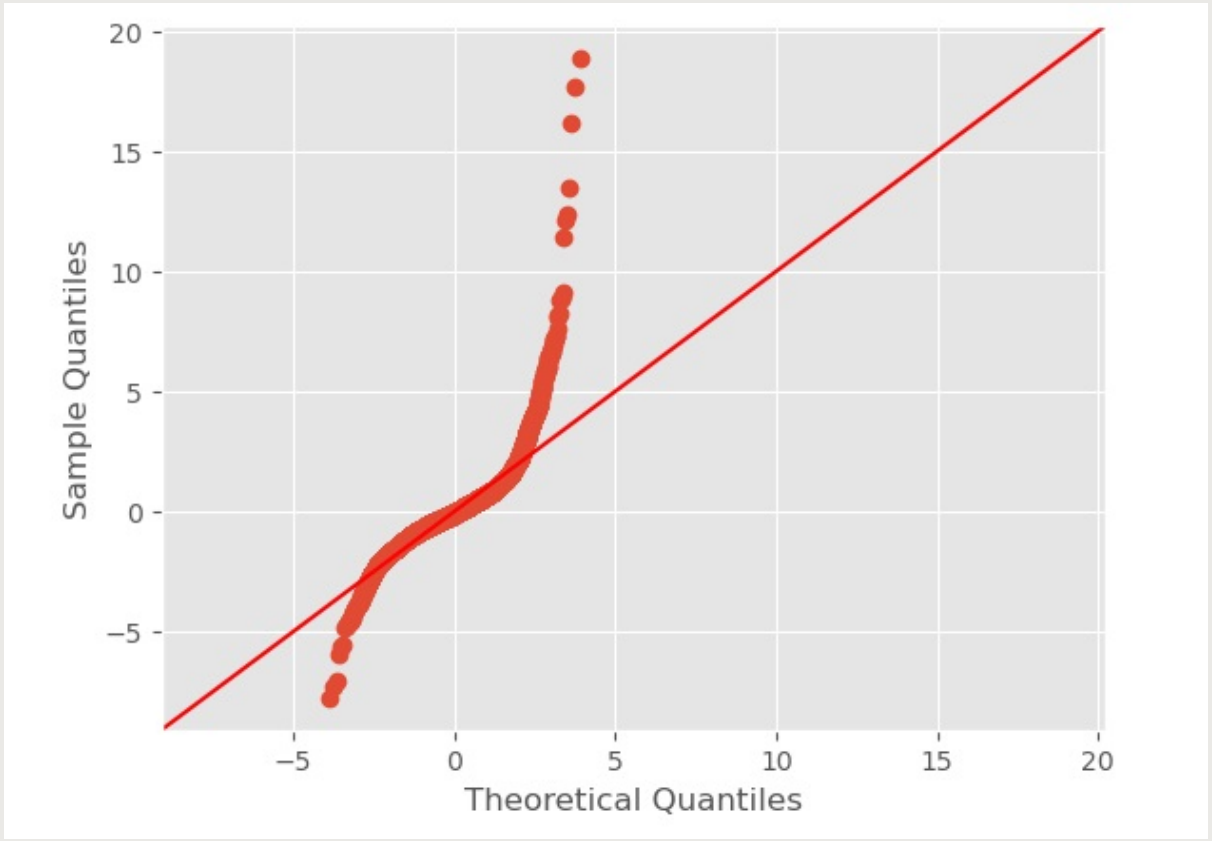
DATA MODELLING

Iteration 2

- In the second iteration the categorical variables are split into their dummies, this means we get closer look at how each category value affects the price.
- Variables with strong multicollinearity are removed, otherwise, we risk small changes to the model causing big fluctuations, making the model less reliable.
- After modelling, we QQ plot our model.

DATA MODELLING

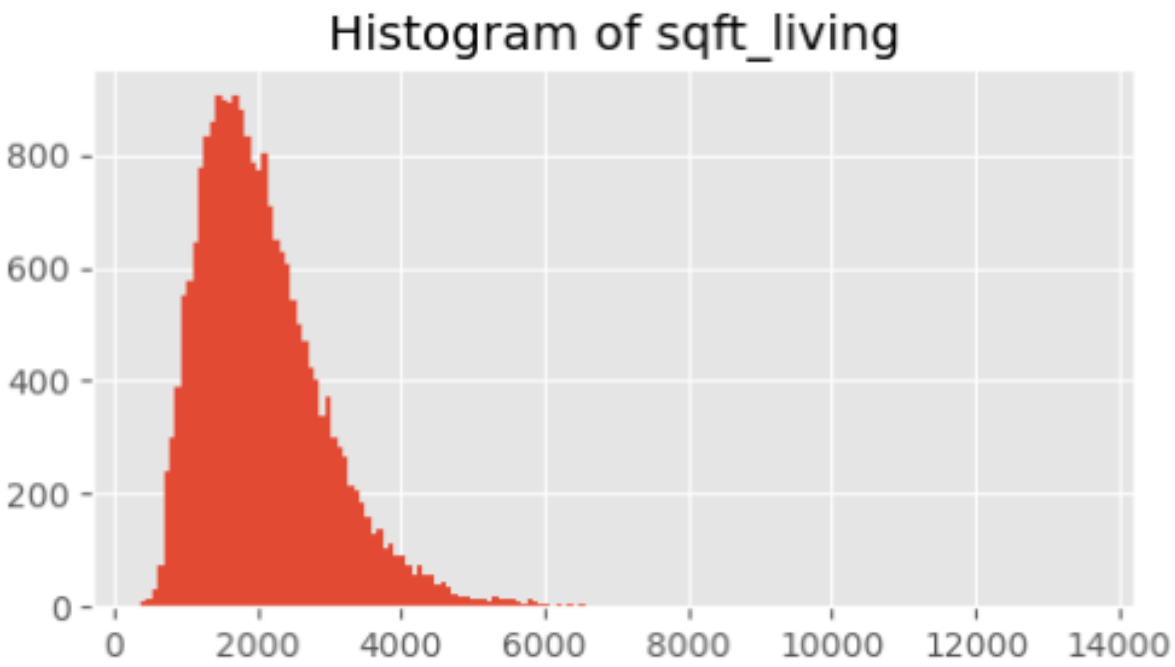
Iteration 2

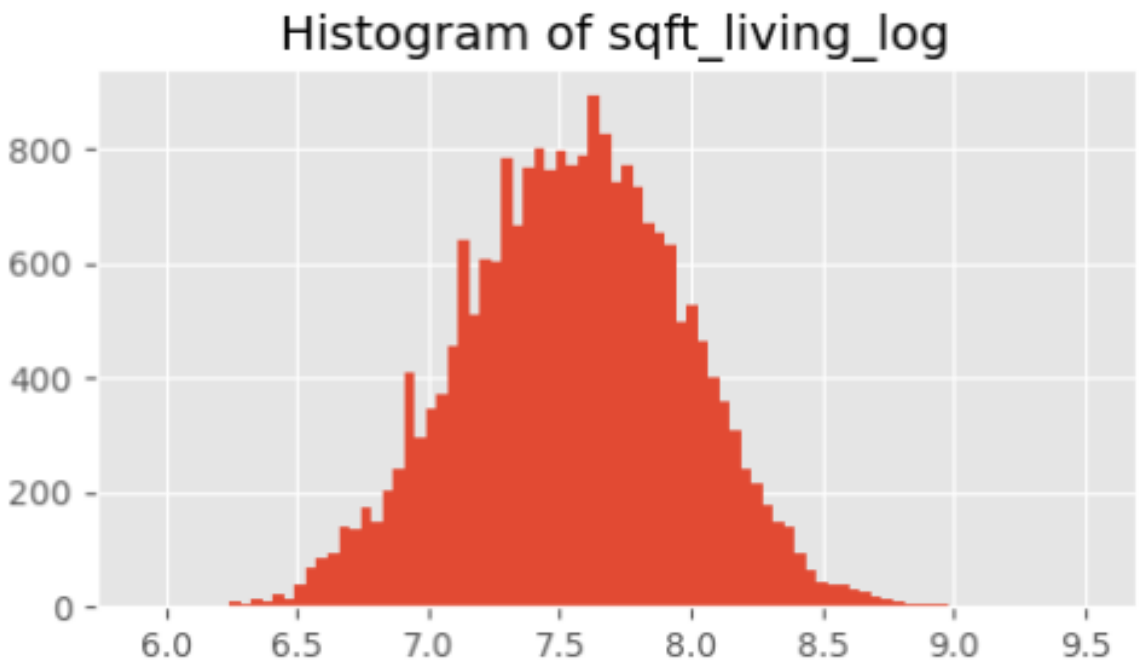


- Our QQ plot of iteration 2 is seen to our left.
- The QQ plot suggests our data is very heavy tailed. Compared to a normal distribution, there is much more data at the ends of the distribution, and less data in the centre. This can affect the validity of our model.

DATA MODELLING

Iteration 3

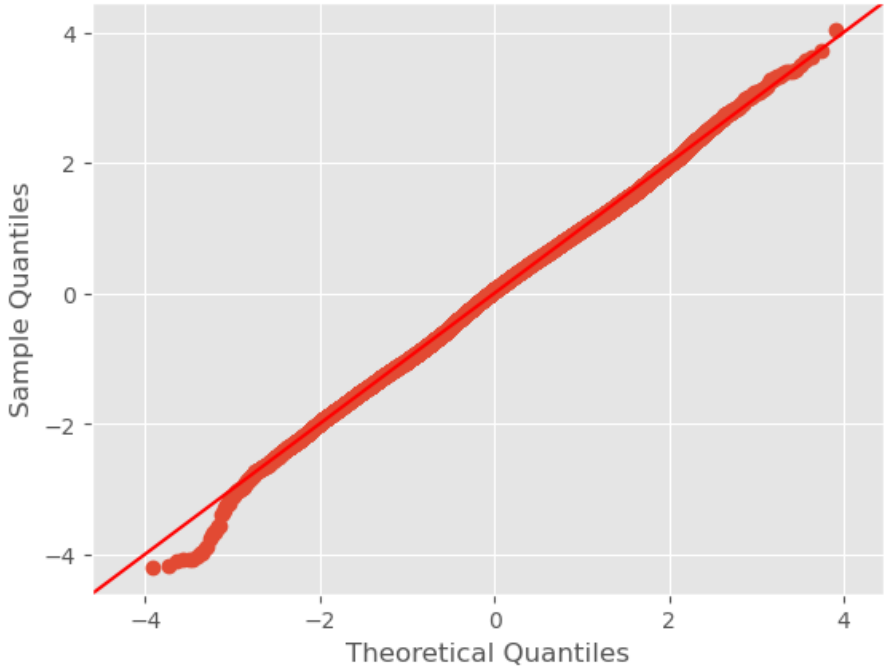




Our third iteration focuses on logarithmic transformations, and scaling. Our example above shows how one of our variables has become far more normalised. This makes the data more compatible with our model.

DATA MODELLING

Iteration 3



- As you can see from our QQ plot, the data has become more normalised.
- After cross validation, the model is neither underfitted or over fitted, making it robust, and suitable for modelling any new data.
- With p values of close to 0, we can confidently say that all or variables used in the final iteration are correlated with house price.
- Our final r squared value of 0.57 is smaller than our initial one, however, we have reduced the noise and made our model far more robust.


CONCLUSIONS

Sqft of the living area.

Is the variable with the largest affect on house price. This makes sense and seems like an obvious correlation. That is followed by **sqft of the lot**. Again, an obvious correlation, and sellers aren't really able to change or improve the size of the lot. If homeowners are considering an extension of their property, they should know this is one of the best ways to increase the value of their house.

House grade

House grade has a clear impact on house price, which represents construction quality. I would recommend sellers research the King County house grading system and find ways to increase the grade of their house before attempting to sell.

Properties with a view of the waterfront

are worth more. If potential buyers should are interested in buying near the coast, they should bear this in mind. When looking at areas to live, buyers will need to consider how valuable a waterfront view is for them, and are they willing to pay the extra for this.

Older houses

are more valuable than newer houses. Buyers should take this into consideration when investigating property.

Future of the model:

This model would be useful for king county residents looking to sell to get an initial estimate of what their house might be worth.