FEBRUARY 2023

MULTILINEAR REGRESSION MODEL

King County House Sales Analysis.

Summary.

I’m tasked to assist a real estate agency to guide homeowners s on how renovations could potentially enhance the assessed value of their properties and by what degree.

Findings:

- A model is created through 5 iterations and it explains 86.6% of the variables in the house prices. It can help individuals to decide on where to put their investment budget case by case - depending on multiple variables such as the age of the house, the current lot space etc.
- “Square footage of house apart from basement", "the square footage of interior housing living space for the nearest 15 neighbors" and "the square footage of the lot" as the most important features.
- The homeowners should always look to maximise their liveable square footage and shouldn’t compromise on it by increasing the number of bedrooms or floors. The property value comes down the space size.

Outline

BUSINESS PROBLEM

DATA

METHOD

ITERATIONS

RESULTS

CONCLUSIONS

LIMITATIONS

NEXT STEPS

Business Problem.

In January 2023, the median listing home price in King County, WA was $800K, trending up 4.2% year-over-year.

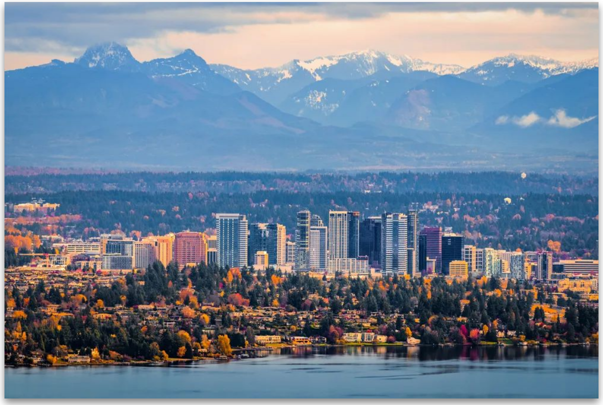
It presents a good business opportunity for the homeowners who wants to sell their homes and maximise their profits from their investments. However, to maximise the profit, they need to commit to renovations.

The question is how much they should spend and which features they should invest in.

Data.

The King County House Sales dataset is used for this study. It comes from **2014-2015 year's sales data** and consists of **21 features**, including an unique id, and **21,597 entries**. With the amount of entries, the modelling process should satisfy the minimum requirements of the regression models.

Method.



I took an **iterative** and **train-test split approach** to the regression model. Used both **OLS model** from **statsmodels** and **LinearRegression** from **sklearn**.

I used **plots to visualise the relationships** between the variables and **behaviours of the residuals**.

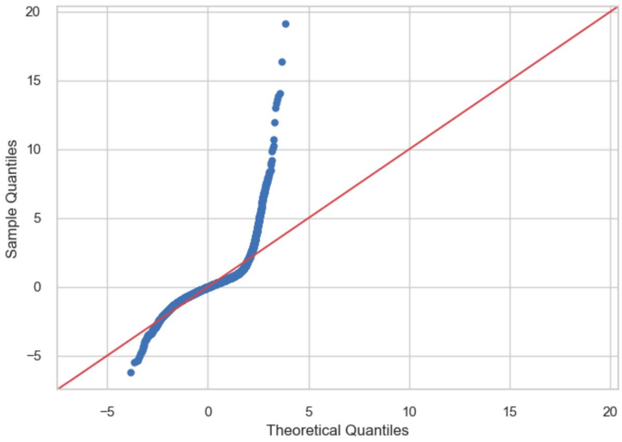
To prepare the data for the analysis, I also utilised the following methods:

- I removed columns and rows that were not part of the study ('id','lat','long', and 'view')
- Either filled the null values or drop the rows that consist of them.
- Transformed some of the features into more insightful ones. For example, the built year is transformed into the age of the house.
- Addressed the multicollinearity of features and removed the highly correlated ones from the dataset.
- Created dummy variables to include the categorical values to the study.

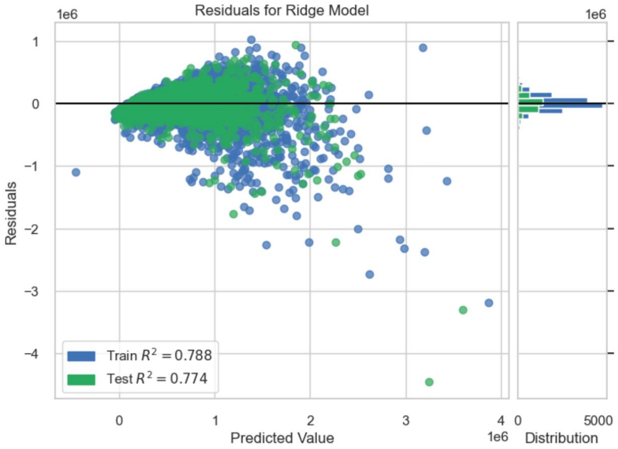
Iterations

Model 1.

The first model didn’t use any data transformations. I removed the predictors that have high multicollinearity and created dummy variables for the categorical predictors. The model explained 78.8% of the variables but couldn’t satisfy the primary assumptions for the linear (multilinear) regression.

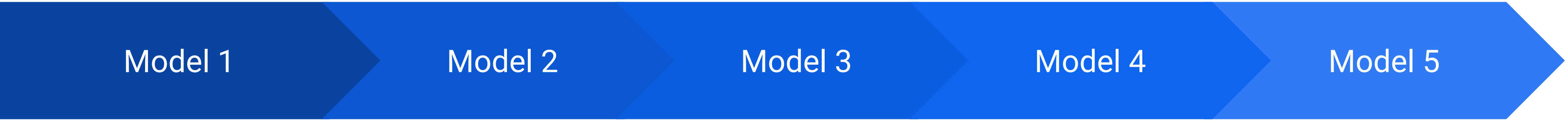


Not linear



Heteroscedastic

Iterations.



Model 1

Model 2

Model 3

Model 4

Model 5

Addressed multicollinearity and categorical variables.

The adj. R2 is 78.8%

Addressed the linear relationship between variables (log transformations)

The adj. R2 is 86.4%

Residuals are more linear and homoscedastic.

Normalised the data

Value between -1 and 1

The adj. R2 is 86.4%

Removed the insignificant features.

Decreased the training data size.

The adj. R2 is 86.3%

Better fit and more reliable model!

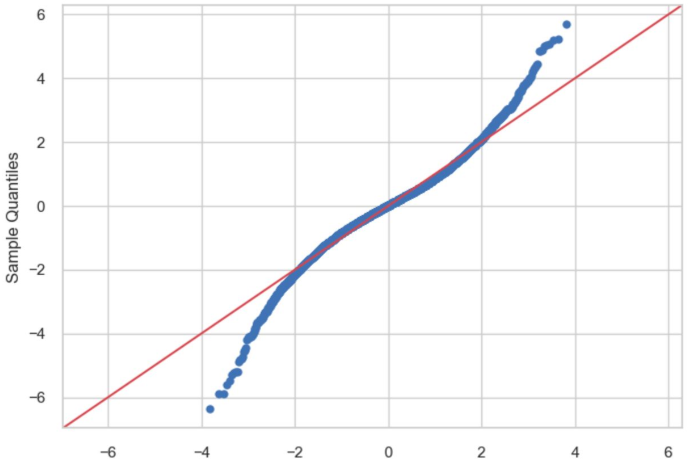
Eliminated the outliers

The adj. R2 is 86.6%

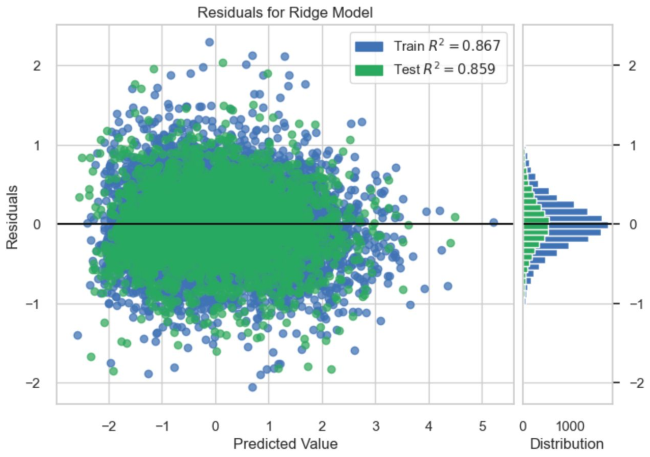
Final model.

Model 5.

The last model have more linearity between the variables, it satisfies primary assumption, and the model is a better fit which means it’s more reliable. It explains 86.6% of the variables.

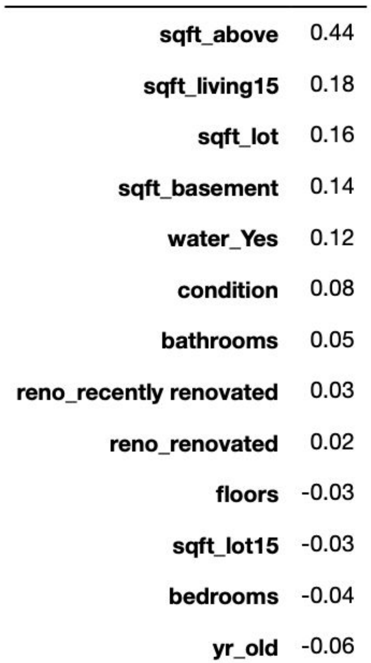


More linear



Homoscedastic

Results.
Space. Space. Space.



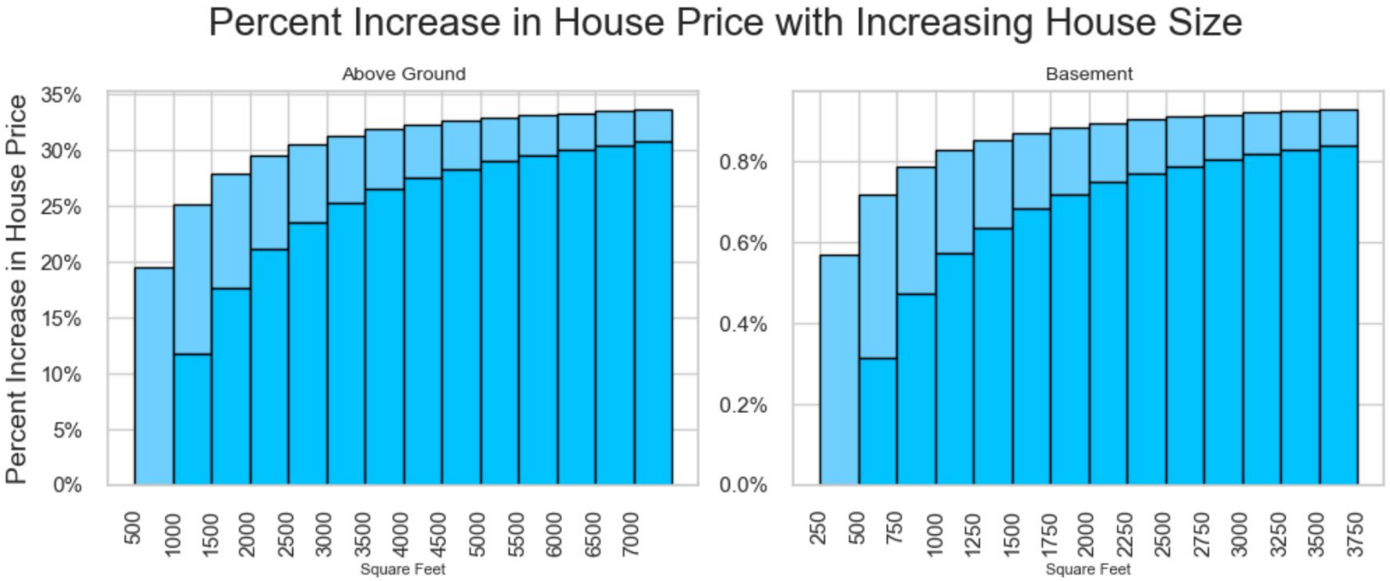
sqft_above	0.44
sqft_living15	0.18
sqft_lot	0.16
sqft_basement	0.14
water_Yes	0.12
condition	0.08
bathrooms	0.05
reno_recently renovated	0.03
reno_renovated	0.02
floors	-0.03
sqft_lot15	-0.03
bedrooms	-0.04
yr_old	-0.06

Looking at the coefficients - aka features’ weight on the sale price - and the most important features are related to the square feet space of the houses. In fact, the number of floors and bedroom have negative relationship with the price, meaning that when their numbers increases the sale value drops down.

I recommend the homeowners in the King County should invest in maximising their livable space in the houses - even it means knocking the walls down and decreasing the number of bedrooms.

Increasing House Size.

The below graphs show the percent increase of the prices with every incremental square feet update. As you can see, above ground is affecting more than the basement - and the price increase keeps getting smaller with each incremental upgrade.



Conclusions.

The model should be use case-by-case situations while guiding the homeowners. It can give different valuations by changing the existing features and help the real estate agency to find the most profitable renovation budget and identify the key feature improvements.

Rule of thumb:

- “Square footage of house apart from basement", "the square footage of interior housing living space for the nearest 15 neighbors" and "the square footage of the lot" as the most important features.
- The homeowners should always look to maximise their liveable square footage and shouldn’t compromise on it by increasing the number of bedrooms or floors. The property value comes down the space size.
- Having renovations in the past 5 years, good quality finish, having a waterfront view and high number of bathrooms all increase the overall sale price.

Limitations & Next Steps.

Moving forward, I highly recommend considering the followings:

- **The square footage of interior housing living space** for the nearest 15 neighbors is the second most important feature in the model. By itself, it doesn't give much insight on the renovations. If we transform this data into a categorical variable that states **whether a house has a larger living space than its neighbours** will be more meaningful.
- **The number of floors and the number of bedrooms** have **negative correlation** with house prices. It's an unexpected result from a business perspective. Their relationship with the square footage of living space and the sizes of the rooms and floors could give more actionable insights. It's possible that having larger rooms and floors could be increasing the value of the property more than the number of them.
- Finally, **Jarque-Bera score** of the study is **still at large** and needs addressing. **Transforming the features** by considering the above points could help - or the model could be more aggressive with **removing the outliers.**

Thank you!

Hazal Aydin

h.aydinhazal@gmail.com

https://www.linkedin.com/in/hazalaydin/

https://github.com/hazal-aydin