PREDICTING HOUSE PRICES WITH THE KING COUNTY HOUSING DATASET

Tori Magin

BUSINESS PROBLEM

Mid-size local real estate agency needs data insights to advise homeowners and/or investors what type of renovations will increase house values.

THE DATA

To provide these insights and create an accurate model for predicting house prices, the King County Housing Data Set was analysed.

This dataset contains 20 features of over 20,000 houses sold in King County between May 2014 – May 2015.

1. id - unique identified for a house
2. Date - house was sold
3. Price - is prediction target
4. Bedrooms - Number of Bedrooms
5. Bathrooms - Number of bathrooms
6. sqft_livingsquare - footage of the home
7. sqft_lotsquare - footage of the lot
8. Floors - total floors (levels) in house
9. waterfront - House which has a view to a waterfront
10. view - Has been viewed
11. condition - How good the condition is (Overall)
12. grade - overall grade given to the housing unit, based on King County grading system
13. sqft_above - square footage of house apart from basement
14. sqft_basement - square footage of the basement
15. yr_built - Built Year
16. yr_renovated - Year when house was renovated
17. zipcode - zip
18. lat - Latitude coordinate
19. long - Longitude coordinate
20. sqft_living15 - The square footage of interior housing living space for the nearest 15 neighbours
21. sqft_lot15 - The square footage of the land lots of the nearest 15 neighbours

DATA PREPARATION

DROPPED VARIABLES

ID, date, latitude, longitude, and **zipcode** we're removed as they were not considered reasonable predictors

CLEANING

Null values replaced with 0.

GROUPING

Year Renovated and **Year Built** were bucketed into groups to make them easier to work with.

- Yr_Renovated grouped in 'Yes' or 'No as most entries had no renovation year.
- Year Built grouped into four periods.

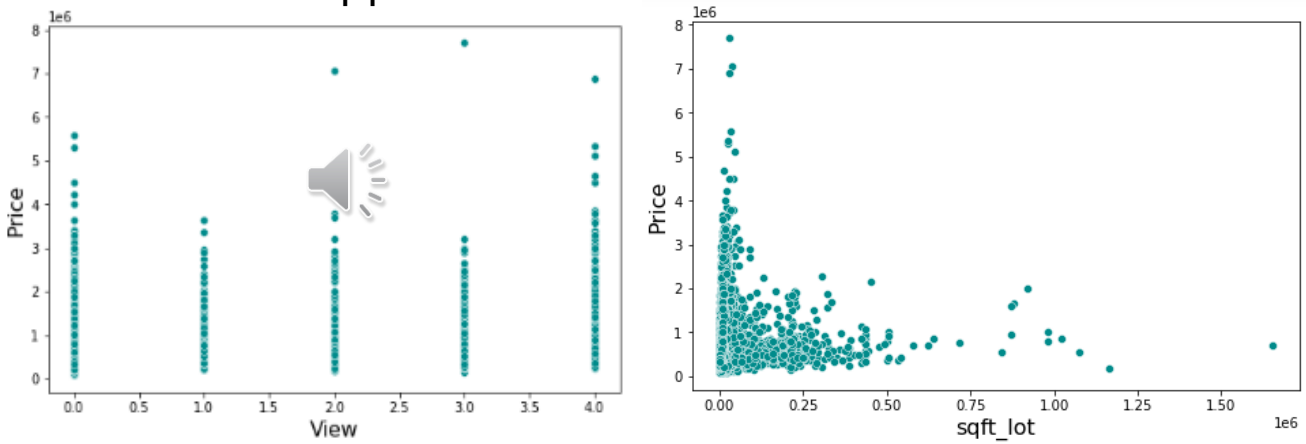
Conditions were consolidated from 11 to 6 groups and given qualitative names (such as Poor, Average, Excellent) to make them more meaningful.

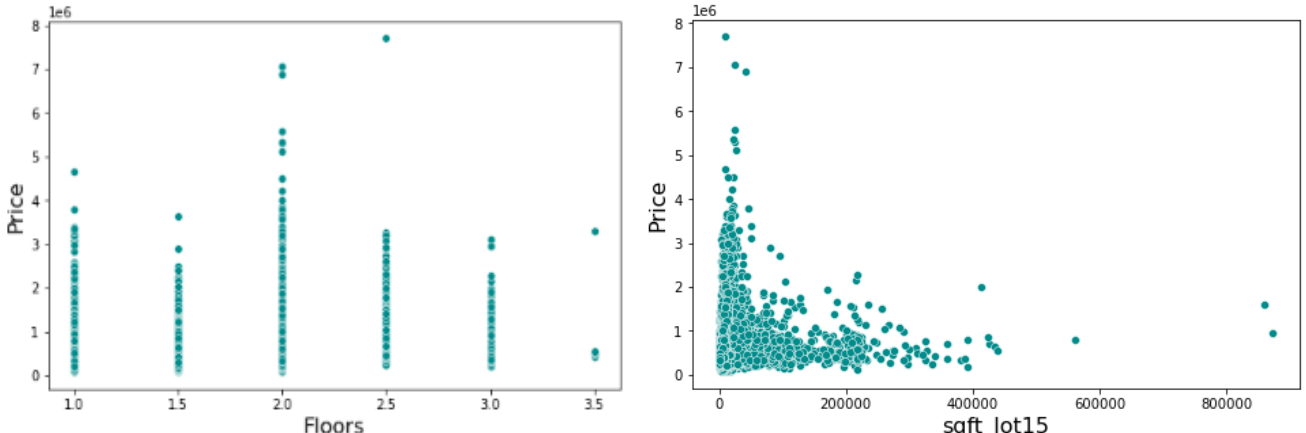
EXPLORING THE DATA

By examining the scatter plots between price and the potential predictors, we can see there is no clear linear relationship between price:

- View
- Floors
- Sqft Lot
- Sqft Lot15

These features were dropped from the dataset for the baseline model.





Multicollinearity and Correlation with Price

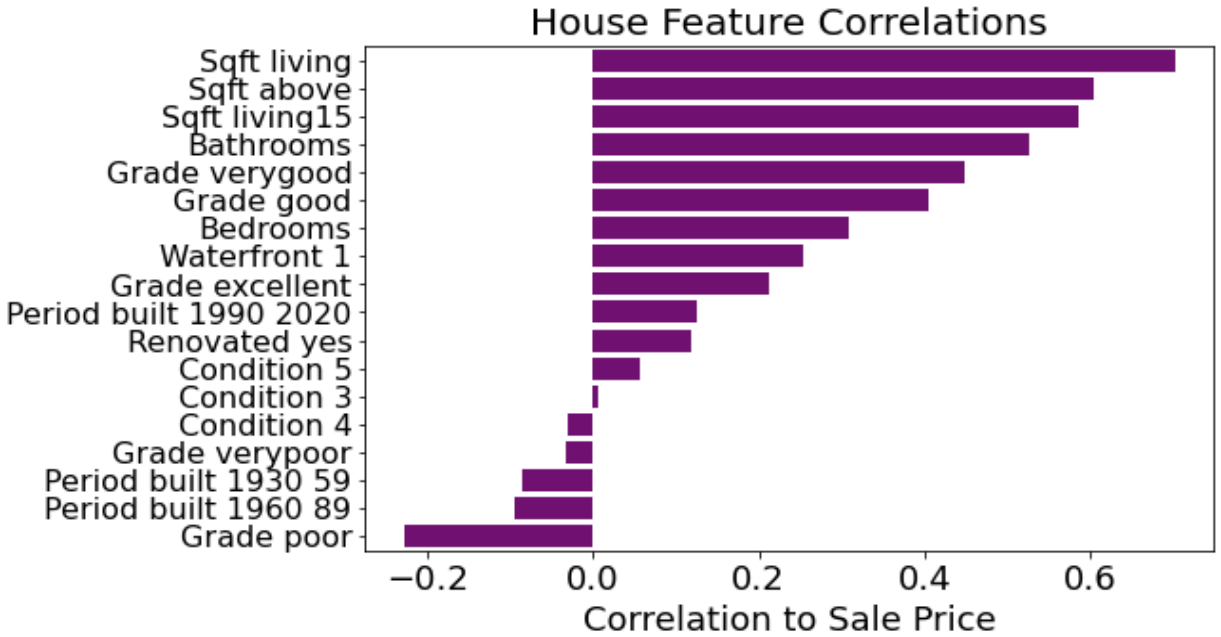
By identifying highly correlated house features, and these features correlation with price, it was decided to remove:

- Sqft_above
- Sqft_living
- Condition_3

To resolve possible multicollinearity issues.

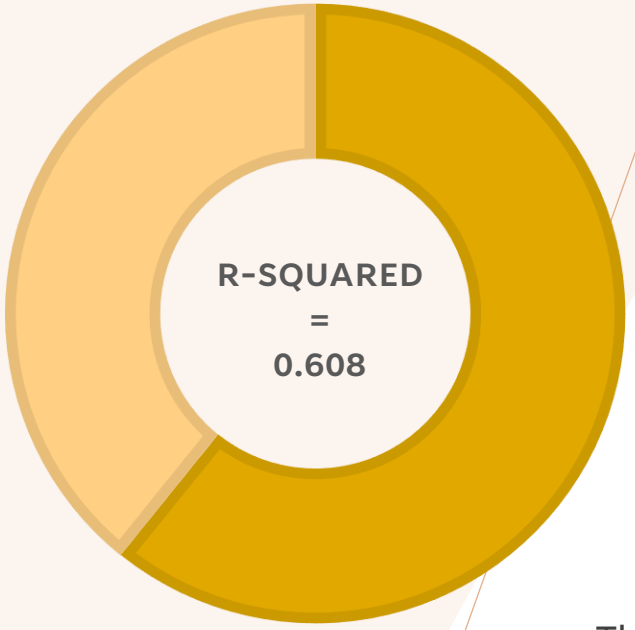
Highly Correlated Features

pairs	cc
(sqft_above, sqft_living)	0.876448
(condition_3, condition_4)	0.812294
(sqft_living, sqft_living15)	0.756402
(bathrooms, sqft_living)	0.755758



EXPLORING THE DATA

FINAL MODEL



log_price ~

bathrooms + condition_4 + condition_5 + grade_excellent + grade_good + grade_poor + grade_verygood + grade_verypoor + period_built_1930_59 + period_built_1960_89 + period_built_1990_2020 + renovated_yes + waterfront_1 + log_sqft_living

PREDICTIVE FEATURES

i. Square feet of living area
ii. No. of bathrooms
iii. Grade, based on King County grading system
iv. Period built
v. View of the waterfront (or not)
vi. Renovated (or not)

The model can account for about 61% of variability in the sale price.

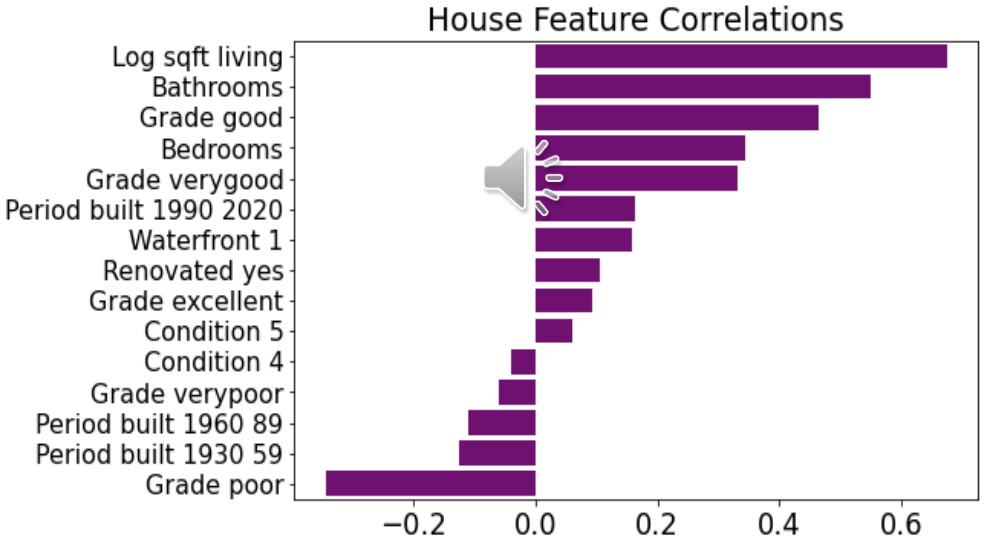
A p-value of less than 0.05 means that we can reject the hypothesis that there is no relationship between price and the predictor variables.

RECOMMENDATIONS

1. EXPAND THE LIVING AREA

2. ADD A BATHROOM

3. FOCUS ON HIGH CONSTRUCTION QUALITY

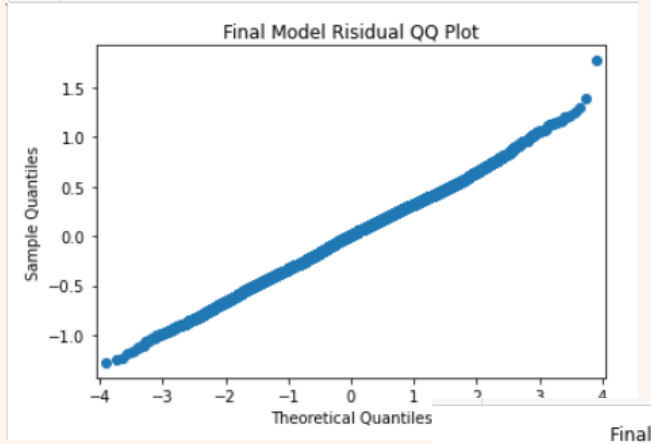


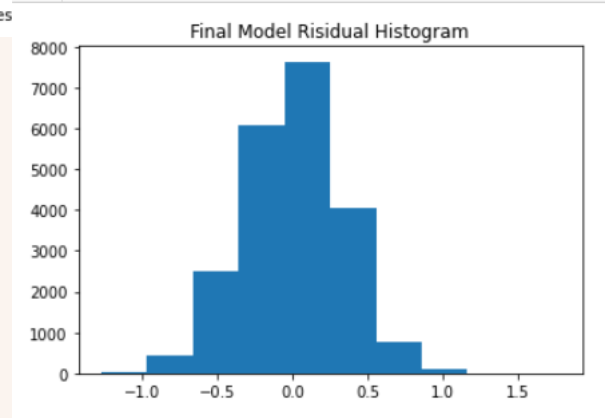
The square footage of a house, its grade, and number of bathrooms are among the strongest predictors of house prices.

FINAL ASSUMPTION CHECKS

Normality:

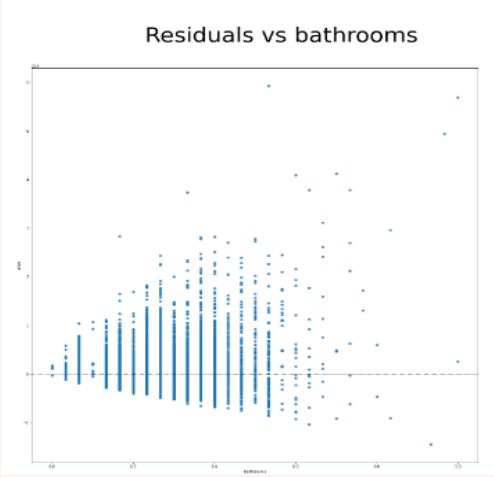
the residuals follow a normal distribution.

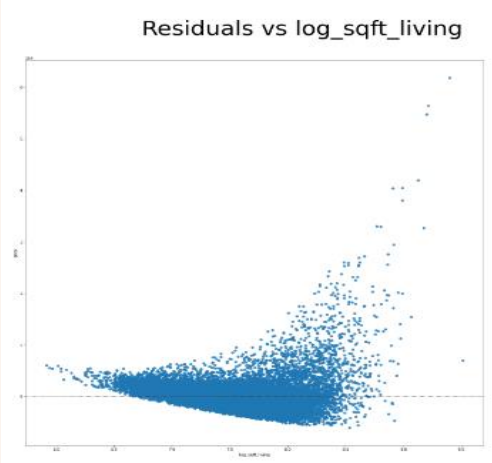




Homoscedasticity:

After log transforming and min-max scaling, the residuals for the predictors have **do not** have perfectly equal variance along the regression line.





SUGGESTIONS FOR FUTURE ANALYSIS

1. Resolve the Homoscedasticity problem to create a reliable model.

2. Analyse the King County house prices by suburb to provide insights to which areas are most valuable

3. Remove the outliers from the the dataset.