Linear Data Modelling Project: ‘Predict how various property features affect property sale price’

Author: Evan James, Melbourne, Aus - 2024

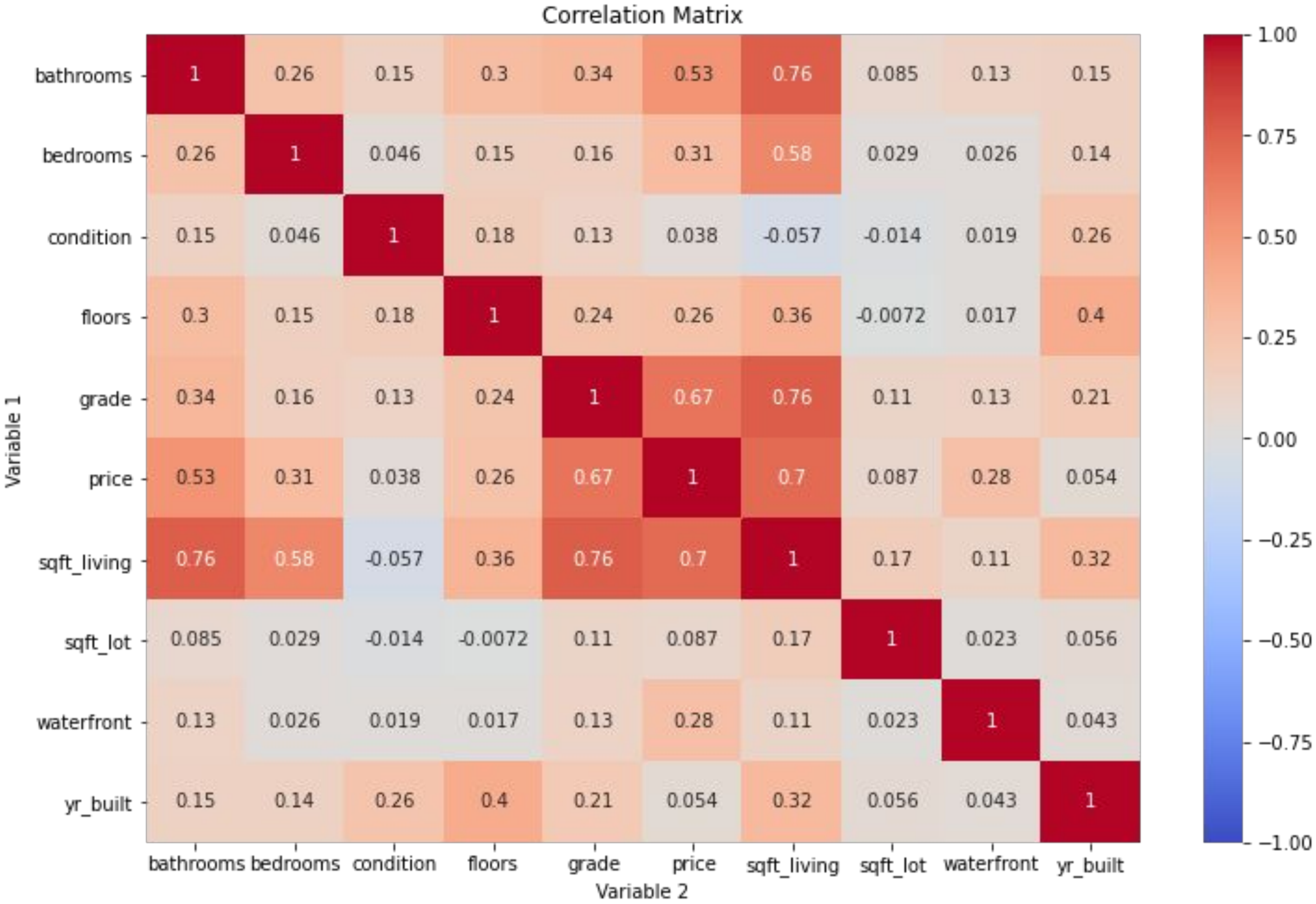
The purpose of this project is to test the null hypothesis that there is not a linear relationship between any features of properties, and their sold price.

The project is based on King County, USA sold prices in 2014-2015.

After checking the data and removing relevant outliers and Nan values there are approximately 18,000 rows. This is a substantial enough data sample for developing a robust price model.

Brief summary of data cleaning/preparation

- Select columns and load into dataframe.
- After checking all columns for Nan values and non-numeric values, it was discovered that the Waterfront Column had approximately 2000 Nan values. As most of the corresponding price values are found in both waterfront and non-waterfront properties it is not possible to determine from the available data which category (ie/ waterfront or non-waterfront) to assign these properties, so they were removed.
- Outliers removed from columns chosen to be utilised. Bathrooms, bedrooms, and sqft_living, in particular, needed outliers removed.
- Columns chosen to be excluded based on the correlation matrix heatmap and based on their relevance. Sqft_lot, condition, and yr built were excluded as correlations to price are very low.
- All columns checked and, where necessary changed to appropriate and consistent numeric values, so they would be usable in further analysis.
- Categorical variables verified, dummy and one hot encoding implemented to prevent multicollinearity issues.
- The two remaining continuous variables, sqft living and price, were checked for and found have no multicollinearity concerns.



Correlation Matrix

	bathrooms	bedrooms	condition	floors	grade	price	sqft_living	sqft_lot	waterfront	yr_built
bathrooms	1	0.26	0.15	0.3	0.34	0.53	0.76	0.085	0.13	0.15
bedrooms	0.26	1	0.046	0.15	0.16	0.31	0.58	0.029	0.026	0.14
condition	0.15	0.046	1	0.18	0.13	0.038	-0.057	-0.014	0.019	0.26
floors	0.3	0.15	0.18	1	0.24	0.26	0.36	-0.0072	0.017	0.4
grade	0.34	0.16	0.13	0.24	1	0.67	0.76	0.11	0.13	0.21
price	0.53	0.31	0.038	0.26	0.67	1	0.7	0.087	0.28	0.054
sqft_living	0.76	0.58	-0.057	0.36	0.76	0.7	1	0.17	0.11	0.32
sqft_lot	0.085	0.029	-0.014	-0.0072	0.11	0.087	0.17	1	0.023	0.056
waterfront	0.13	0.026	0.019	0.017	0.13	0.28	0.11	0.023	1	0.043
yr_built	0.15	0.14	0.26	0.4	0.21	0.054	0.32	0.056	0.043	1

Baseline model

The first linear model included all cleaned and prepared variables as a baseline. While the adj. R score was above 0.65 there were too many high P values and many inexplicable negative coefficients to draw any consistent conclusions. There were extremely high levels of multicollinearity as indicated by the condition no. and very high coefficients across all predictor variables.

Bathrooms had a large range of values which is likely to have made the results unstable and possibly contributing to an overfit model.

Sqft_living and Price were found to be skewed right.

Further modelling

Because of the unreliable and confounding results of the baseline model, backward elimination was not practical. So the stepwise selection method was adopted.

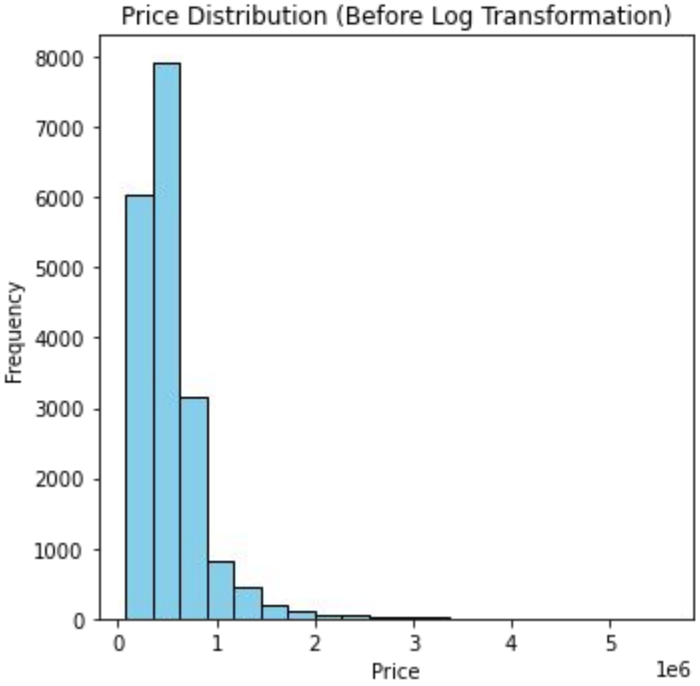
Simple linear models to analyse the individual predictors relationship to the dependent ‘Price’ variable were built. This provided deeper insights into the shape of the data.

Sqft living and price were both log transformed and visually checked to determine whether this would be enough to meet the distribution assumptions of the model.

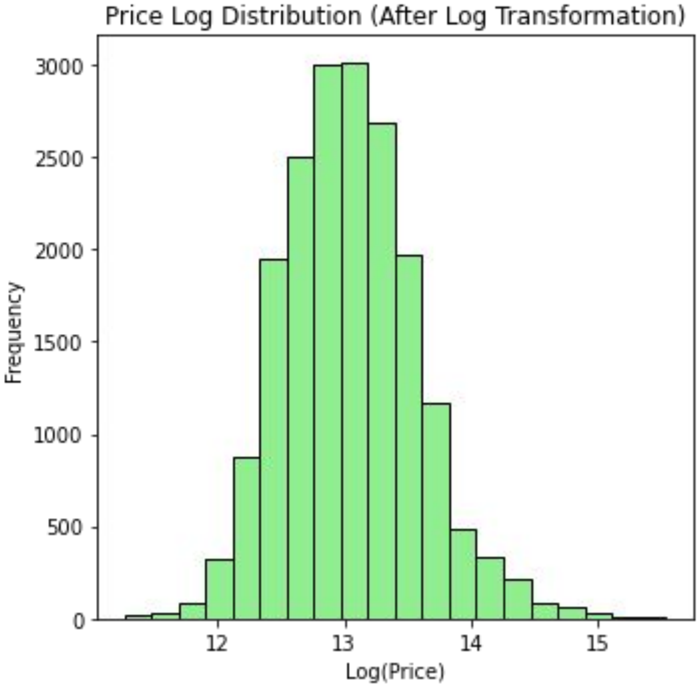
Then from the small models, other relevant variables were introduced into the model and analysed for P values, adj. R and coefficients etc.

Bathrooms were binned into logical whole number groups to avoid overfitting. Distributions were checked.

Log transformation before and after example.



Price Distribution (Before Log Transformation)



Price Log Distribution (After Log Transformation)

Final model

All coefficients are explainable in the final model. All P values are significant except for Floors_3.

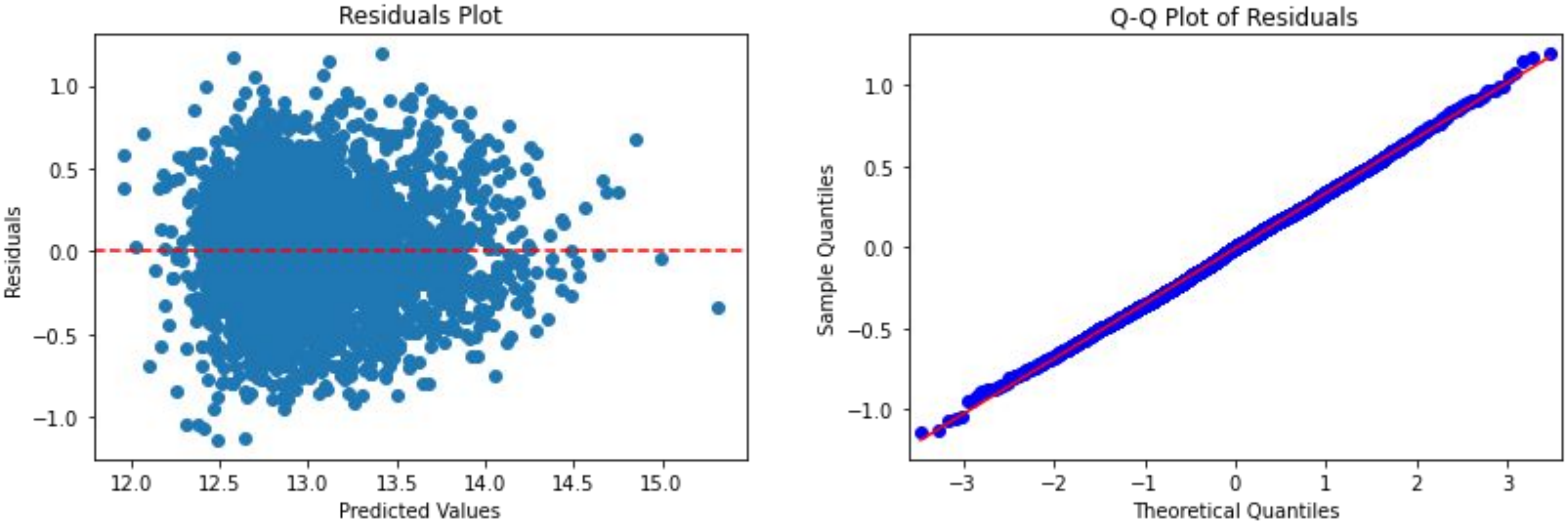
Multicollinearity is low

Adj. R of 0.577 means the model explains 57% of the increase in price. Unfortunately this is the highest level that could be achieved without producing confounding results. This means there are likely other factors not represented in this data that are affecting the price.

Final model results

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               price_log   R-squared:                       0.578
Model:                             OLS   Adj. R-squared:                  0.577
Method:                  Least Squares   F-statistic:                     1357.
Date:                 Sun, 25 Feb 2024   Prob (F-statistic):               0.00
Time:                         07:38:13   Log-Likelihood:                -6486.5
No. Observations:                18860   AIC:                         1.301e+04
Df Residuals:                    18840   BIC:                         1.317e+04
Df Model:                           19
Covariance Type:             nonrobust
```

Training and testing results show a reasonably robust model MSE: 0.1162 Training 80:20 Test



Results summary

After excluding irrelevant and confounding variables, and accounting for multicollinearity, a robust positive linear relationship has been modelled with an acceptable Mean squared error of 0.116 after splitting the data and testing. and almost uniformly 0.0 P values.

The Null hypothesis -that there is not a linear relationship between property features and price - can be rejected as certain property features were found to have significant linear relationships to price.

Property features interaction investigation.

Increasing sqft living without

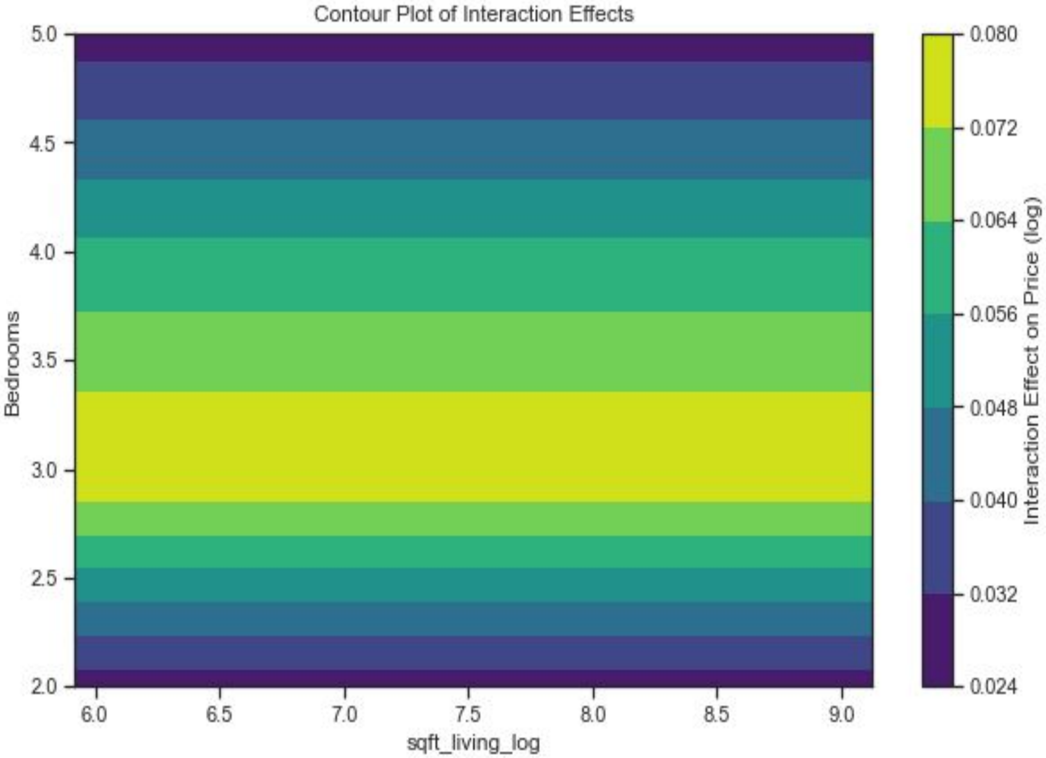
Increasing bedroom numbers

Will likely not maximise price.

Other interactions were investigated

But were not significant enough

To make recommendations.



Renovation Recommendations

Increasing bedroom numbers without increasing sqft_living is predicted to not increase the sale price. Therefore it is not recommended to convert living areas or larger bedrooms into smaller bedrooms as this is less desirable.

Increasing the sqft living alone can significantly increase the price. However, if bedroom numbers are increased in tandem then further price increase can be expected - Although with diminishing returns above 3 bedrooms.

The property feature with potentially the highest effect on price (that can be changed) is raising the Grade (ie. The building standard). By renovating an existing property in poor condition and, not changing any other features, bring the property to a high standard will likely have the largest impact on price.

Increasing bathroom numbers alone has a low effect on price in comparison to the other features discussed. As bathrooms are the most expensive rooms to add (p/sq.m) this cannot be recommended based on this data.

Limitations and further investigations

The modelling is robust but can only explain approx 57% of the Price.

Additional relevant domain knowledge and data may be useful for further investigation of other important factors contribution to price. For example, quality of school, income and other demographics, distance from transport connections and shops, crime rates, unemployment rates, scope and quality of local amenities, etc.