Home Price Prediciton of King County, WA single-family homes using Linear Regression

Overview

The project's goal is to predict the sales price of a home in King County, WA based on the features of the home. One of the most common methods to predict continuous values is through linear regression. Linear regression explains the relationship of independent predictor variables to a dependent predicted variable using a linear equation. The linear equation represents the best fit among the data, where given a set of predictor variables the predicted value would be found. Ordinary least squares linear regression finds this linear equation by minimizing the sum of the squared difference between the actual observed dependent variable and the predicted value.

This project will use a dataset of home sales and design three linear regression models. The models will be compared and the best model will be used for a backend function for a dashboard to predict the sales range for a home.

Business Problem

Bon Jovi Real Estate Advisors is a residential real estate broker in King County, WA. Many of their clients come to them needing to sell their homes but are unsure of the price to list. The real estate broker wants us to design a model where they can take in the features of their client's home and determine which price to begin listing discussion with their client.

Stakeholders

- President and Managing Director of Bon Jovi Real Estate Advisors
- Bon Jovi real estate agents that will use the model for price predictions

Deliverables

- Presenatation to stakeholders
- Jupyter Notebook
- GitHub Repository

Repository Navigation

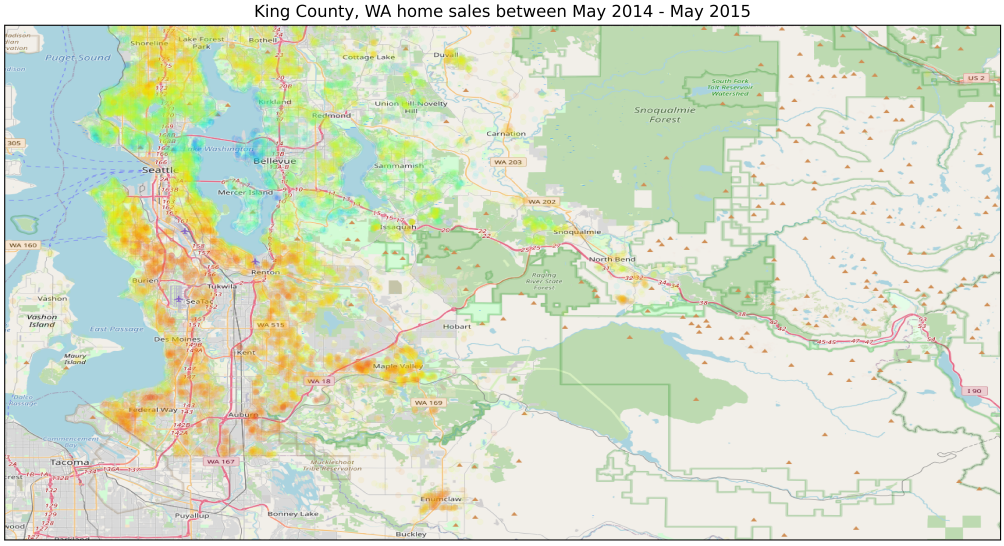
- /img - contains image files
- /data - house sale data from King County, WA
- /license - license information
- project_notebook.ipynb - Jupyter notebook

The Data

This project uses the King County House Sales dataset, which can be found as `kc_house_data.csv` in the data folder in this repository. The description of the column names can be found in `column_names.md` in the same folder.

King County Home Sales May 2014 - May 2015

King County, WA home sales between May 2014 - May 2015



Variable Names and Descriptions for King County Data Set

See the King County Assessor Website for further explanation of each condition code

Variable	Data Type	Description
`id`	catagorical	Unique identifier for a house
`date`	continuous	Date house was sold
`price`	continuous	Sale price (prediction target)
`bedrooms`	discrete	Number of bedrooms
`bathrooms`	discrete	Number of bathrooms
`sqft_living`	continuous	Square footage of living space in the home

Variable	Data Type	Description
sqft_lot	continuous	Square footage of the lot
floors -	discrete	Number of floors (levels) in house
waterfront	ordinal	Whether the house is on a waterfront
view	ordinal	Quality of view from house
condition	ordinal	How good the overall condition of the house is. Related to maintenance of house
grade	ordinal	Overall grade of the house. Related to the construction and design of the house
sqft_above	continuous	Square footage of house apart from basement
sqft_basement	continuous	Square footage of the basement
yr_built	catagorical	Year when house was built
yr_renovated	catagorical	Year when house was renovated
zipcode	catagorical	ZIP Code used by the United States Postal Service
lat	catagorical	Latitude coordinate
long	catagorical	Longitude coordinate
sqft_living15	continuous	The square footage of interior housing living space for the nearest 15 neighbors
sqft_lot15	continuous	The square footage of the land lots of the nearest 15 neighbors

Data Cleanup

The median house sold in King County, WA between 2014 to 2015 was for $450,000. The median house sold was 1910 square feet, 3 bedroom, 2.25 bathrooms, and 47 years old. The home sale price range was $78,000 and to $7,700,000.

Missing Data

The data in the `waterfront`, `view`, and `yr_renovated` had `NaN` values. There are placeholder values in `yr_renovated` and `sqft_basement`. `yr_renovated` has 0 for over 95% of values which means the property was not renovated. `sqft_basement` has `?` for about 2% of its values.

Strategy for clean-up

- `NaN`
  - `waterfront`
    - Binary categorical variable (`YES` or `NO`)
    - replace `NaN` with mode of `NO` as most likely these properties are not waterfront
  - `view`
    - Ordinal categorical variable
    - replace `NaN` with `NONE`
  - `yr_renovated`
    - Will be converted to a countable numerical variable
    - `0` is the most common value with over 95% of values.
    - Replace `NaN` with 0 value
- Placeholder
  - `yr_renovated` has `0` for missing or unknown values.
  - `sqft_basement` has `?` for missing or unknown values.

Encoding Variables

For regression analysis, catagorical data needs to be in numerical format so catagorical variables were encoded to to meet this requirement but still maintain their binary, ordinal, and count information.

variable	Data Type	Plan
`condition`	ordinal	Recode to dictionary. `{'Poor': 0, 'Fair': 1, 'Average': 2, 'Good': 3, 'Very Good': 4}`
`grade`	ordinal	Delete the descriptor, keep the number, and convert it to `int` datatype. Example: `7 Average` becomes `7`

variable	Data Type	Plan
`basement`	binary	If there is a basement (sq.ft > 0) the value will be set to `1`. No basement (sq.ft = 0) set to `0`. `?` makes up about 2% of values and the current value of `0` makes up almost 60%. Replace `?` with the mode of `0`.
`view`	oridinal	Recode to dictionary. `{'NONE': 0, 'FAIR': 1, 'AVERAGE': 2, 'GOOD': 3, 'EXCELLENT': 4}`
`waterfront`	binary	Recode to dictionary. `{NO': 0, 'YES': 1}`.
`home_age`	discrete	Create variable from `yr_built`. Subtract current year from `yr_built`. Drop `yr_built`
`yr_since_reno`	discrete	Create variable from `yr_renovated`. Subtract current year from `yr_renovated`. `0` is the most common value with over 95% of values. If never renovated then subtract from `yr_built`. Drop `yr_renovated`.
`zipcode`	catagorical	Used dummy variables

Outliers

An obvious outlier was found that I believe was a data entry error. A property is listed as having 33 bedrooms but only having 1,620 square feet in size. I edited this observation to 3 to match the median number of bedrooms in the dataset.

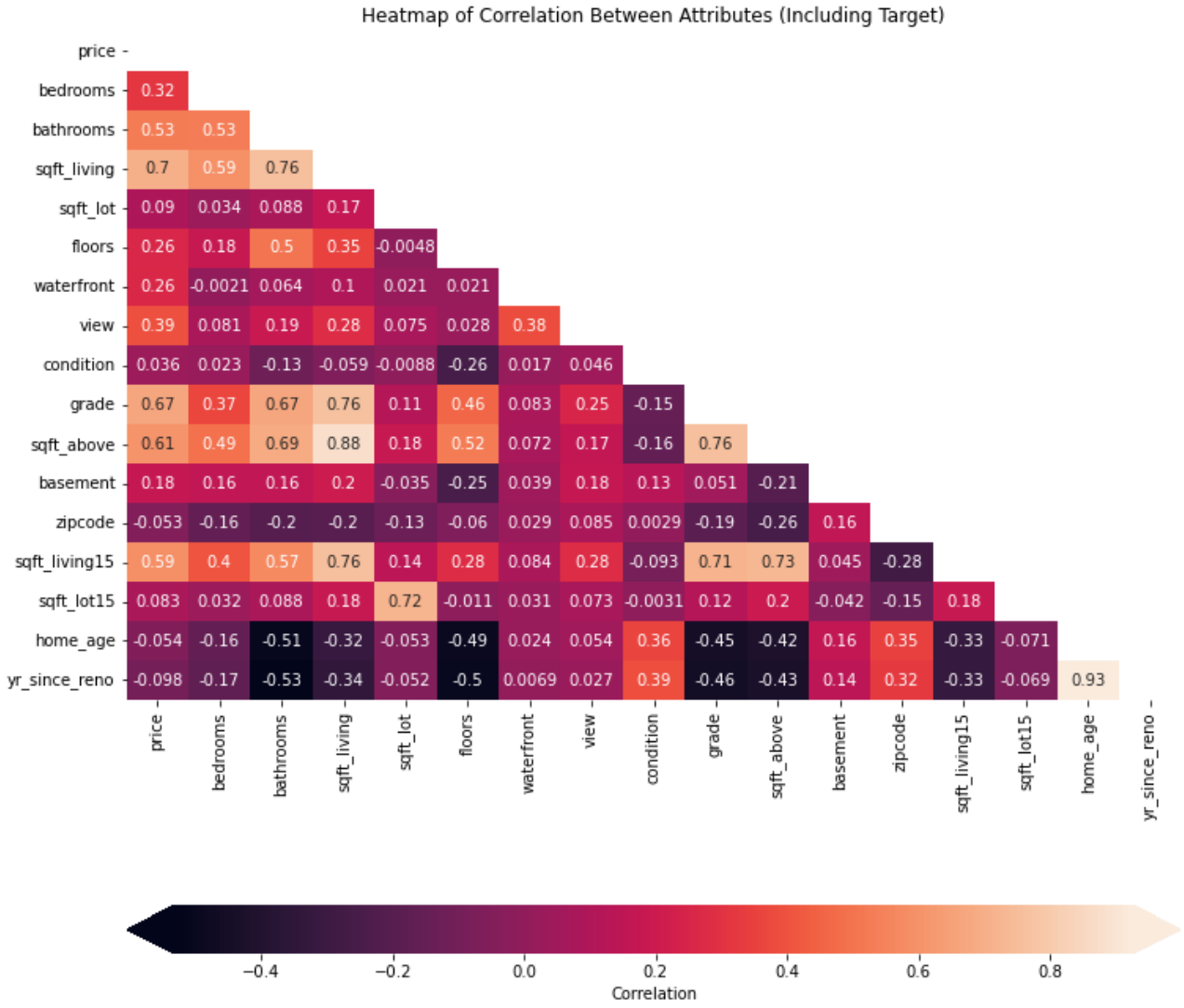
Data Limitations

- Data is only from 2014 to 2015. Models to predict future sales price would need to be updated with newer data.
- Some data might be missing, such as for-sale-by-owner or owner-financed sales.
- Ordinal data might be highly variable based on examinter's subjective experience.

Modeling

Model 1 (M1)

M1 is the baseline model and it was chosen from variables with a Pearson's correlation greater than 0.6 with sales price of a house.



Independent variables correlated with `price` with Pearson's correlation (r) greater than 0.6 are `sqft_living`, `sqft_above`, `grade`. As the heatmap shows, `sqft_living` and `sqft_above` are highly correlated with each other (r = 0.88) so this will likely create multicollinearity.

Model 2 (M2)

M2 uses an automated stepwise regression strategy. All the variables are fed into the model and fitted. The predictor variable with highest p-value is removed if the p-value is greater than 0.05. This is repeated until all the p-values of the predictor variables are less than 0.05.

M2 has many more features than M1 and utulizes the zipcode dummy variables.

Model 3

M3 builds on M1 by adding interaction effects to the main effects. It also removes `sqft_above` as a predictor variable because of multicollinearity. Interaction effects occur when the effect of one predictor variable depends on the value of another variable. For example, condition of a home may be dependent on the age of the home. There may be a dependency between the square feet of living space and the number of bathrooms.

Statistics used to compare models

R-squared

R-squared is a statistical measurement of how close the observed data is to the regression line. It tells us if the model has can explain the variance seen in the data. Adjusted R-squared is similar to R-squared but it takes into account the number of independent variables used in the model as R-squared has a tendency to increase each time a new variable is added to the model. R-squared range from 0 to 1 with the higher the value, the more the model explains the variance.

$$R^2 = 1 - \frac{\text{Residual Sum of Squares (RSS)}}{\text{Total Sum of Squares (TSS)}}$$

$$= 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Prediction interval

A prediction interval (PI) is the range where a single new observation is likely to fall given specific values of the indpendent variables. The prediciton interval can be use to assess if the predicitons are sufficiently in a narrow range to satisfy the client's requirement. Prediction intervals can be compared across models. Smaller intervals indicate tighter predictive range. Large prediction intervals tell us the model could have a wide range in its predictions and would not meet the client's needs.

$$PI = 1.96s,$$

where s is the sample standard deviation calculated by

$$s = \sqrt{\frac{1}{N-2} RSS}$$

Root Mean Squared Error (RMSE)

RMSE is a measure of the mean error rate of a regression model that penalizes larger errors. It is the square root of the average squared difference between the predicted dependent value and the actual values in the dataset. The smaller the RMSE value, the closer the fitted line from the linear equation is to the actual data. Like Mean Squared Error (MSE), this statistic squares the residual error before it is averaged, which gives a high weight to large errors, but because it the square root is taken, the statistic is in the same units as the dependent variable, sale price ($USD). The lower the score of RMSE the closer the model fits the data.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2},$$

where $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$ are the predicted values,
$(y_1, y_2, \ldots, y_n)$ are the observed values
and $(n)$is the number of observations

Regression Results

	R_squared	PI	RMSE
M1	0.554774	0.688652	0.351381
M2	0.883369	0.352465	0.180738
M3	0.629711	0.628029	0.323571

M1

The p-values indicate that each of the variables chosen has a statistical significant relationship with the dependent variable, but the r-squared value is low at 0.55. For a predictive model the R-squared needs to be higher. The function of M1 would be:

$$\hat{y} = 0.58X_{\substack{sqft \\ living}} + 0.22X_{grade} - 0.23X_{\substack{sqft \\ above}} + 8.6531$$

The coefficient for `sqft_living` is 0.58. This means for every 1% increase in square feet of living space in a house, there is a 0.58% increase in sale price. For every 1% increase in the square feet in the above ground area of a house there is a 0.22% reduction in sale price. Interestingly, the area of the space above the home is penalized. This may be due to multicolinearity between the `sqft_living` and `sqft_above` . For every one-unit increase in the grade of the home, there is a 24.6% increase in sales price.

M2

R-squared and adjusted R-squared is 0.89 for M2. This is tremendous improvement over M1. There are many more independent variables used in M2 as compared to M1, though. This model used all the independent variables expect for some of the Zipcodes after recoding this variable to dummy variables. Thought there are many variables, the p-values indicate that each of the variables chosen has a statistical significant relationship with the dependent variable. This R-squared score is good for a predictive model.

M2 has the lowest predictive intervals. The smaller the predictive interval the more confidence the true sale price is in that region. M1 and M2 had resonably similar prediction intervals and twice the value as M2.

M2 is almost half the RMSE score of M1 and M2 indicating it produces less error between the actual and predicted values.

The coefficient for `condition` for M2 is 0.0527. This means that for ever increase in one-unit of the condition value there is about a 5% increase in the home sale price. Some of the highest coefficients in in the Zipcode variables. For example, with all other independent variables held constant a home in 98039 would sell for 266% more than Zipcode 98003.

M3

R-squared and adjusted R-squared for M3 is 0.63. This is an improvement over M1 but is worse than M2. The interaction effects between some variables does seem to help increase R-squared.

RMSE is similar to M1 and less than M2.

Recommendations and Next Steps

Summary

- Our client wants to be able to predict sales price.
- Ordinary least squares linear regression was used to create three models.
- The three models were compared using R^2, Prediction Intervals (PI), and Root Mean Squared Error (RMSE).
- Model 2 (M2) is the best model as it has the best predictive capabilities, R-squared 0.88, low RMSE and PI.
- M2 could be used to prototype a client dashboard for real estate agents to predict sales price for new data.
- More data and variables should be collected to improve the model's predictive power.
- Communicate with client about internal real estate data that can be used to further train the model.

Actionable Recommendations

1. M2 could be used for a client dashboard prototype for Bon Jovi real estate agents to predict sales price.
2. M2 can be used to measure the cost-benefit analysis of making improvements to the home. For example, a one-unit increase in the condition of the home will increase the sale price by about 5%.

3. M2 can help Bon Jovi real estate agents locate customers and properties that have the highest sale price potential. For example, homes in Zipcode 98039 sold for over 200% more than homes in Zipcode 98003 so those customers in 98039 likely have a higher sales price.

Next Steps

This model could be improved to make better predictions by adding more data additional features, such as crime rate in the geographic location of the home, the zoned public school ranking, and time the house was on the market until it was sold. The GPS coordinates of the sold house could be used to collect the first two of these variables. The Multiple Listing Service may be a source for more recent data and on how long a house was on the market from day of listing to closing date.

Anther source of data could be in the internal data of our brokerage client. They possibly have data of properties they have sold or bid on, this would include the data of the asking and bidding price of the property.

Summary

- Model 2 is the best model to continue with as it has the best predictive capabilities.
- Model 2 could be used as a prototype for a dashboard the client can use to a range for a asking price.
- Gather more data and variables to improve the model's predictive power
- Communicate with client about internal data that can be used to train the model