King County House Sales

By: Eduardo Osorio

Overview

The purpose of this report is to provide recommendations on which houses in kings county would have the highest return on investment. We're taking a look at a range of different metrics to pinpoint which features are statistically significant for our model.

Questions

1. Are waterfront properties more valuable?
2. How does the square footage affect the price?
3. How does the number of floors affect the price?

What Data was used?

The data we're looking at today is the King County House Sales Datasheet. The data frame contained:

- 21,597 Houses .
- The houses were built between 1900 - 2015.
- Using 70 different Zip Codes in the King county area.
- Grades ranging from 3 - 13 (overall grade given to the housing unit, based on King County grading system. Assuming 13 is the highest.)
- Conditions ranging from 1 - 5 (How good the condition of the house is. Assuming 5 is the highest.)

EDA

- Removed Null values in the '*waterfront*' column
- Replaced the Null values in the '*yr_renovated*' column so that I could use it to engineer another column that counted how many years since the last renovation.
- Dropped the '*view*' column since it showed no purpose.
- Dropped '*bathrooms*, *sqft_lot15*, *and sqft_above*' since they showed collinearity to other columns.
- Created a month_sold column to show us the month the house was last sold.

Final model

The features used in this model were: price, sqft_living, sqft_lot, sqft_living15, yr_built, bedrooms, floors, waterfront, condition, grade, yrs_renovated, sale_month.

- Sqft_living15 is the square footage of interior housing living space for the nearest 15 neighbors.
- Yrs_renovated is a feature engineered for this model that counts how many years since the last renovation.

I used a stepwise function with backwards elimination to test every variable and delete the ones with p-values over .05.

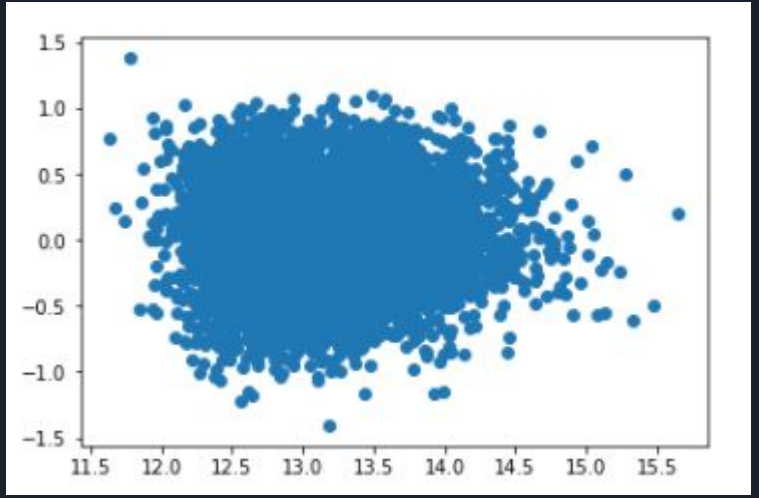
Finally I used Scikit-Learn's Recursive Feature Elimination to select the most relevant features for my model.

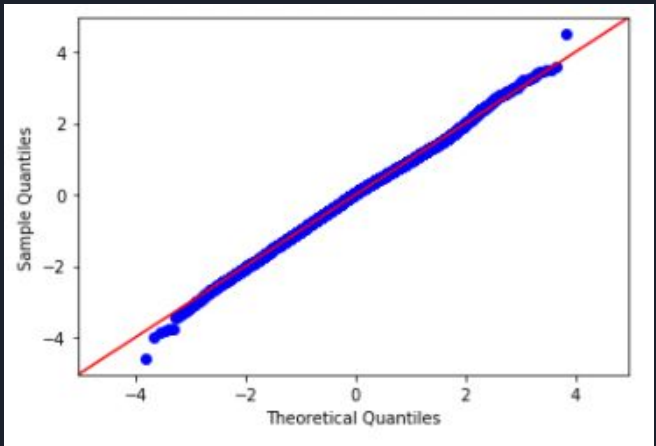
Dep. Variable:	price_log	R-squared:	0.660
Model:	OLS	Adj. R-squared:	0.659
Method:	Least Squares	F-statistic:	992.6
Date:	Sun, 01 Nov 2020	Prob (F-statistic):	0.00
Time:	10:49:54	Log-Likelihood:	-3627.1
No. Observations:	15372	AIC:	7316.
Df Residuals:	15341	BIC:	7553.
Df Model:	30		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	88.1215	1.804	48.852	0.000	84.586	91.657
sqft_living_log	0.4385	0.011	39.120	0.000	0.416	0.460
yr_built_log	-10.5552	0.238	-44.374	0.000	-11.021	-10.089
sqft_living15_log	0.2660	0.012	21.331	0.000	0.242	0.290
sqft_lot_log	-0.0572	0.003	-17.309	0.000	-0.064	-0.051
grade_6	-0.2239	0.010	-22.698	0.000	-0.243	-0.205
grade_8	0.2082	0.007	29.938	0.000	0.195	0.222
grade_5	-0.3854	0.025	-15.488	0.000	-0.434	-0.337
grade_9	0.4429	0.010	42.665	0.000	0.423	0.463
waterfront_1.0	0.5381	0.029	18.513	0.000	0.481	0.595

Final model

- The R squared for this model is .66 which is lower than ideal. The reason for this is because some features were taken out to avoid overfitting.
- The train error is 1.35 and the test error is 1.34
- The first plot seems to show that the data is evenly scattered
- The QQ plot shows us that the data sets come from common distributions.





summary/conclusions

The final conclusions for this model are:

- For every unit of sq ft living area the price increases by $1.56
- If a house is a waterfront property you can expect the price to go up by 80 percent.
- 2 bedroom homes make the price go up by 12% while 4 bedroom homes make the price go down by 7%
- For total floors, houses with both 1.5 and 2.5 floors, increase the value by 3%. Houses with 3 floors increase the value by 25%

Next steps

- Due to the number of zip codes in this county adding them to the model would lead to overfitting. With more time, I would like to test every zip code to see which increase the price of houses the most.

Thank you!

I would like to thank Yish for all the help she's provided and everyone for listening to my presentation.

Sources Used:

- https://towardsdatascience.com/linear-regression-in-6-lines-of-python-5e1d0cd05b8d
- https://machinelearningmastery.com/rfe-feature-selection-in-python/
- https://scikit-learn.org/

Contact info:

- Eduardo.osorio231@gmail.com