BOMA YANGU HOUSING AGENCY

House Price Prediction Regression Model

Antony Brian
abrian@bomayangu.com

Overview

This is a model that predicts house prices given features such as number of floors, proximity to water body and number of bedrooms to give the managers an accurate value to work on based on price.

Then analysis focuses on providing answers to : -

Business Impact
Prevent losses by over or undervaluing house prices

Solutions
Develop a statistical model to estimate house market price given certain features of the house

Next steps
Improve valuation on house price by improving the model

DATA SETS

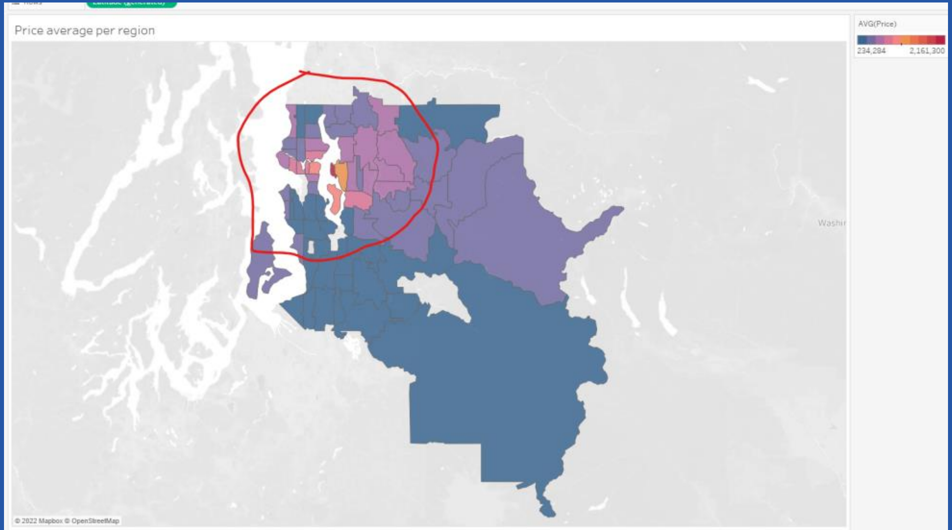
This model has been build and modeled by King County House Sales Dataset which contains 21597 rows and 21 columns of data about house prices and features like bedroom and bedroom numbers of a particular house. The dataset includes homes sold between September 9 2014 and January. This is an American Town hence the Imperial measurements are used in many features.

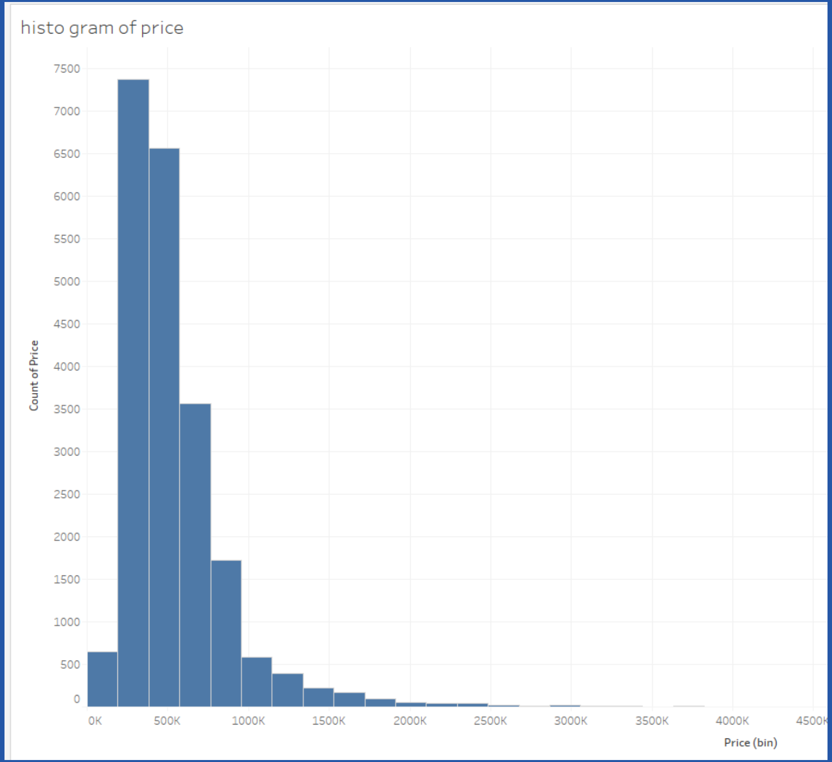
#		#	#	#	#	#	#	Abc	Abc	Abc
kc_house_data.csv	kc_house_data.csv	kc_house_data.csv	kc_house_data.csv	kc_house_data.csv	kc_house_data.csv	kc_house_data.csv	kc_house_data.csv	kc_house_data.csv	kc_house_data.csv	kc_house_dat
Id	Date	Price	Bedrooms	Bathrooms	Sqft Living	Sqft Lot	Floors	Waterfront	View	Condition
1321400060	6/27/2014	257,500.00	3	2.25000	1,715	6,819	2.00000	NO	NONE	Average
2008000270	1/15/2015	291,850.00	3	1.50000	1,060	9,711	1.00000	NO	*null*	Average
2414600126	4/15/2015	229,500.00	3	1.00000	1,780	7,470	1.00000	NO	NONE	Average
3793500160	3/12/2015	323,000.00	3	2.50000	1,890	6,560	2.00000	NO	NONE	Average
1736800520	4/3/2015	662,500.00	3	2.50000	3,560	9,796	1.00000	*null*	NONE	Average
9212900260	5/27/2014	468,000.00	2	1.00000	1,160	6,000	1.00000	NO	NONE	Good
114101516	5/28/2014	310,000.00	3	1.00000	1,430	19,901	1.50000	NO	NONE	Good
6054650070	10/7/2014	400,000.00	3	1.75000	1,370	9,680	1.00000	NO	NONE	Good
1175000570	3/12/2015	530,000.00	5	2.00000	1,810	4,850	1.50000	NO	NONE	Average

#	Id	kc_house_data.csv	id
	Date	kc_house_data.csv	date
#	Price	kc_house_data.csv	price
#	Bedrooms	kc_house_data.csv	bedrooms
#	Bathrooms	kc_house_data.csv	bathrooms
#	Sqft Living	kc_house_data.csv	sqft_living
#	Sqft Lot	kc_house_data.csv	sqft_lot
#	Floors	kc_house_data.csv	floors
Abc	Waterfront	kc_house_data.csv	waterfront
Abc	View	kc_house_data.csv	view
Abc	Condition	kc_house_data.csv	condition
Abc	Grade	kc_house_data.csv	grade
#	Sqft Above	kc_house_data.csv	sqft_above
#	Sqft Base...	kc_house_data.csv	sqft_basement
#	Yr Built	kc_house_data.csv	yr_built
#	Yr Renov...	kc_house_data.csv	yr_renovated
	Zipcode	kc_house_data.csv	zipcode
	Lat	kc_house_data.csv	lat
	Long	kc_house_data.csv	long
#	Sqft Livin...	kc_house_data.csv	sqft_living15
#	Sqft Lot15	kc_house_data.csv	sqft_lot15

DATA EXPLORATION

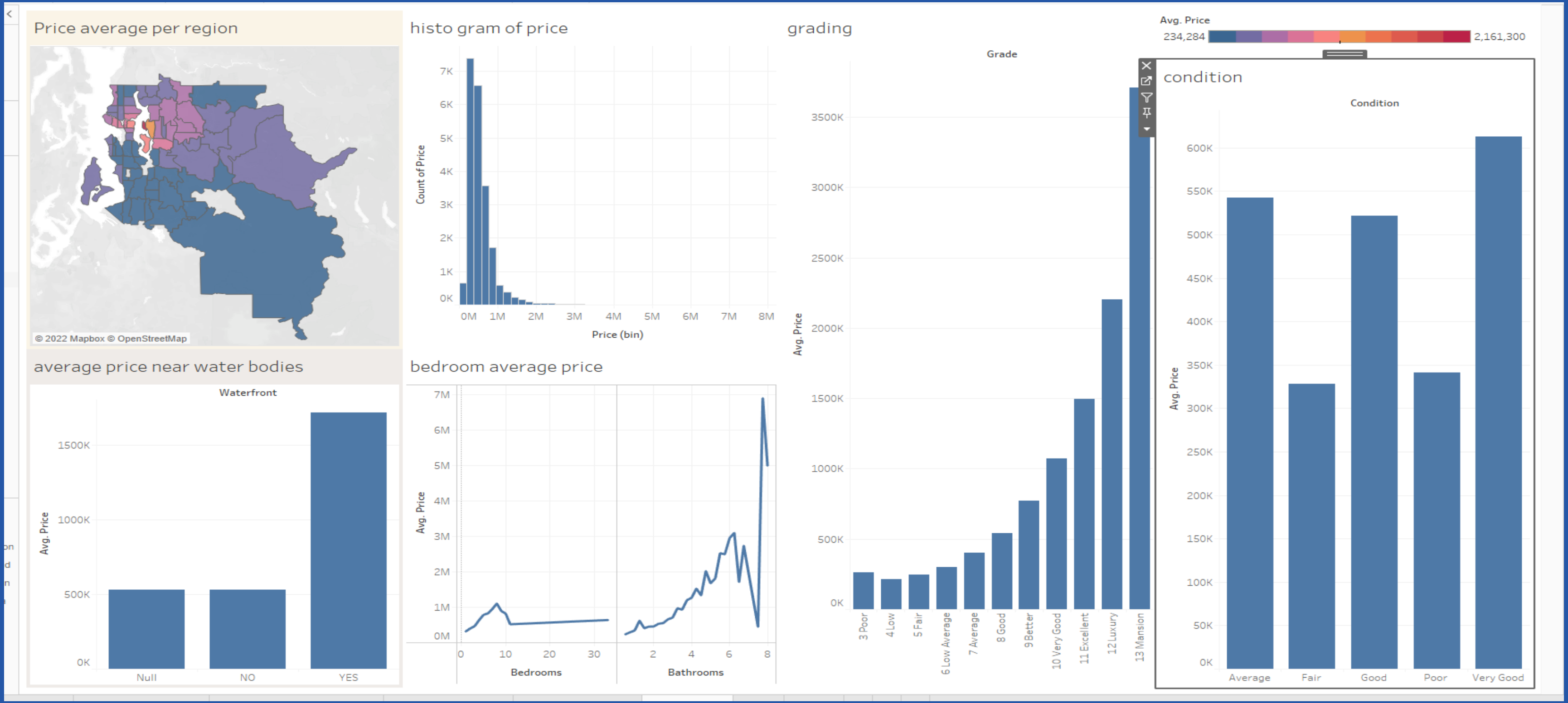
Analyzing the dataset shows some trends like house prices are higher near water bodies. This can be illustrated by the map below which chows high prices near the lakes.
Price distribution also shows that houses valued at about $299k to $500k tend to sell than other house





DATA EXPLORATION

The dashboard below shows features that influence the prices of a house. These features were used to develop the house price prediction model for this project Dashboard link



MODEL

Multiple linear regressor is the model used for this project since it allows us to work with more house features to come up with an accurate model to predict the house prices.

This process requires great data cleaning to avoid errors in the model since the slogan GIO (Garbage In Garbage Out) still applies to get a good performance model.

This include handling empty coloums and rows to transforming categorical variables to numbers to enable the model to analyze them.

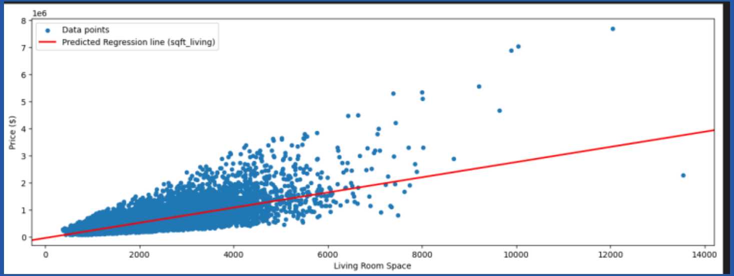
Modeling

The model used in this analysis is a multi regression Linear model which uses several features of a house to predict the house price.

The accuracy of the model is measured with by checking the R-Squared value above.

A value close to one indicates a good performing model which explains every variance in the house price value.

Example of how the model tries to fit one feature to predict price is shown below:



Dep. Variable:	price	R-squared:	0.725
Model:	OLS	Adj. R-squared:	0.725
Method:	Least Squares	F-statistic:	2030.
Date:	Fri, 30 Sep 2022	Prob (F-statistic):	0.00
Time:	14:26:54	Log-Likelihood:	-2.9345e+05
No. Observations:	21597	AIC:	5.870e+05
Df Residuals:	21568	BIC:	5.872e+05
Df Model:	28		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.261e+07	5.24e+05	-43.116	0.000	-2.36e+07	-2.16e+07
bedrooms	-1.845e+04	1856.336	-9.936	0.000	-2.21e+04	-1.48e+04
bathrooms	4.601e+04	3110.856	14.792	0.000	3.99e+04	5.21e+04
sqft_living	128.0980	3.297	38.855	0.000	121.636	134.560
sqft_lot	-0.1110	0.033	-3.398	0.001	-0.175	-0.047
floors	2.44e+04	3122.041	7.816	0.000	1.83e+04	3.05e+04
yr_built	-2245.4965	66.139	-33.951	0.000	-2375.135	-2115.858
yr_renovated	33.8124	3.842	8.801	0.000	26.282	41.343
lat	5.783e+05	9984.762	57.921	0.000	5.59e+05	5.98e+05
sqft_living15	24.5460	3.186	7.704	0.000	18.301	30.791
waterfront_YES	5.544e+05	1.96e+04	28.342	0.000	5.16e+05	5.93e+05

...

Model Summary and Result

Our model model is accurate by about 73%

Taking the difference between the predicated values from our model and the real price form the dataset we get an average difference of about $119k.
This means that on average this model may be off by about $119k in its predictions on home price. This difference is due to the outliers in our data such as mansions which lie outside the average price of many housing units that affect the model predictive ability.

Our model also suggests that for each increase in 1 bedroom space we see an associated decrease in House price by about $19k. This can be brought about by the fact that houses in towns tend to be small and expensive while those out of town are big and relatively cheap cheaper.

We also see a corresponding increase in price of about $45k for every 1 bathroom added to a house. This implies that bathroom number influence a positive price on a house

For each increase in 1 inch of squarefoot of living room space we also see an associated increase in House price by about $129
.

Model Summary

Our model model is accurate by about 73%

Taking the difference between the predicated values from our model and the real price form the dataset we get an average difference of about $119k.
This means that on average this model may be off by about $119k in its predictions on home price. This difference is due to the outliers in our data such as mansions which lie outside the average price of many housing units that affect the model predictive ability.

Our model also suggests that for each increase in 1 bedroom space we see an associated decrease in House price by about $19k. This can be brought about by the fact that houses in towns tend to be small and expensive while those out of town are big and relatively cheap cheaper.

We also see a corresponding increase in price of about $45k for every 1 bathroom added to a house. This implies that bathroom number influence a positive price on a house

For each increase in 1 inch of squarefoot of living room space we also see an associated increase in House price by about $129
.

Conclusion and recommendation

The average error of our model of about $119K can greatly affect the house price of small homes or undervalue big expensive house.

Though the model is accurate by about 73% an improvement can be done by using non-linear regression models to improve the models accuracy.

Solutions to improve on

The models overall performance can be improved by getting more data.
More accurate results can be gained by improving the model from just a liner model to deep learning models which will perform better on the data.

\+ Boma Yangu

Thank you