King County's House Sale Predictions

Milena Afeworki
June 6, 2021

Outlines

___

Overview

This is a presentation on the analysis based on the King County's houses sales price, to identify the significant factors affecting the value worth of Single family homes so the Real Estate agency XYZ could have a finest structure of how much a house entering the market would cost according to these specific factors.

Such statistically analysed information could increase the agency's ability to provide valuable knowledge and information at each step for the clients while also coming up with an unbiased valuation of their home and help set a listing/buying price.

Project objective:

Building a model for understanding the features of a property's relationship to the market sales price.

Data

- King County's House Sales data of over 20,000 homes for the years 2014 - 2015
- Number of bedrooms, bathrooms and floors
- Size of living space
- Size of house
- Condition and Grade based on County's standards
- Location by zip code and coordinate system.

Methods:
The OSMEN process

Step 1

- **Obtain**
- **Scrub**
- **Explore**

Step 2

- **Model**
1. Baseline model
2. Log transform
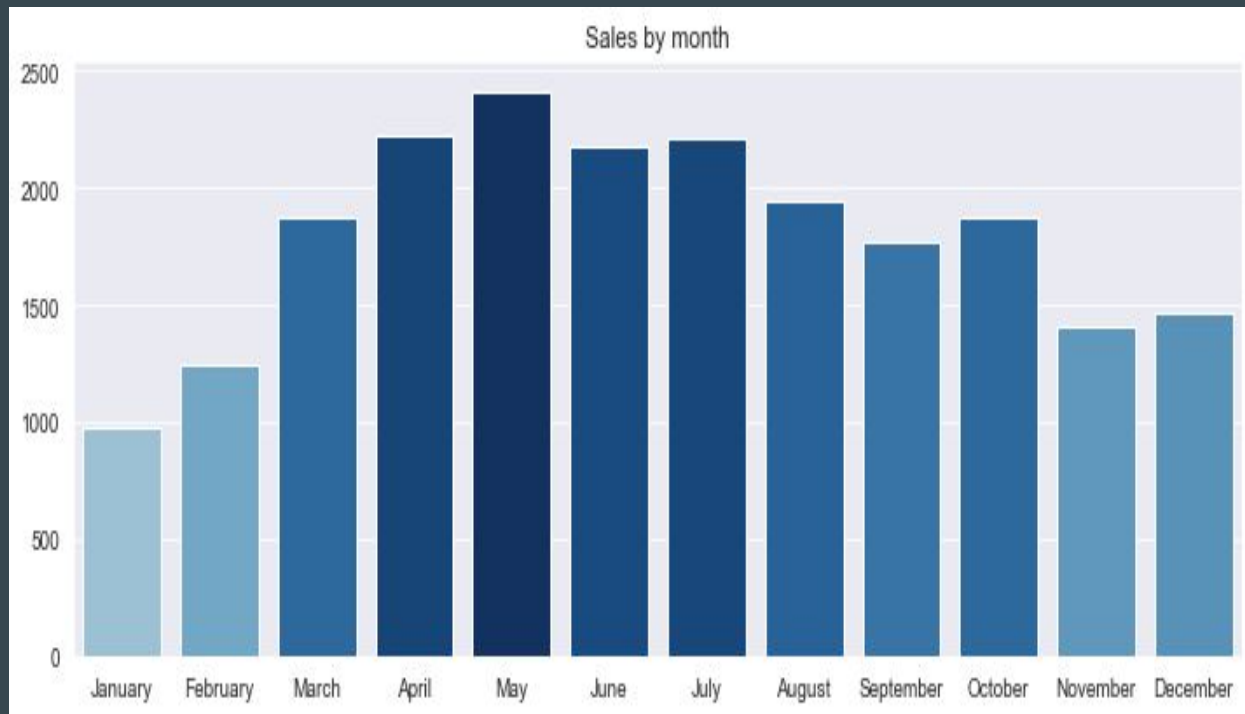3. One-Hot-Encoding
4. Binning Categories
5. Handle correlations

Step 3

- **Interpret**
1. Coefficients
2. MSE
3. R2

Results

Sales Months

- Spring and early Summer are the best possible seasons with higher housing transactions. Families would want to settle before the beginning of a school year.

Sales by month

Results

Sales by location

- Location has a significant role on the price of a property.
- House prices increase as we go higher north of the County.
- Properties around water fronts are more pricey.
- Medina, Mercer and Bellevue come up to be the County's preeminent cities in Housing Sales price.

Locations and Prices
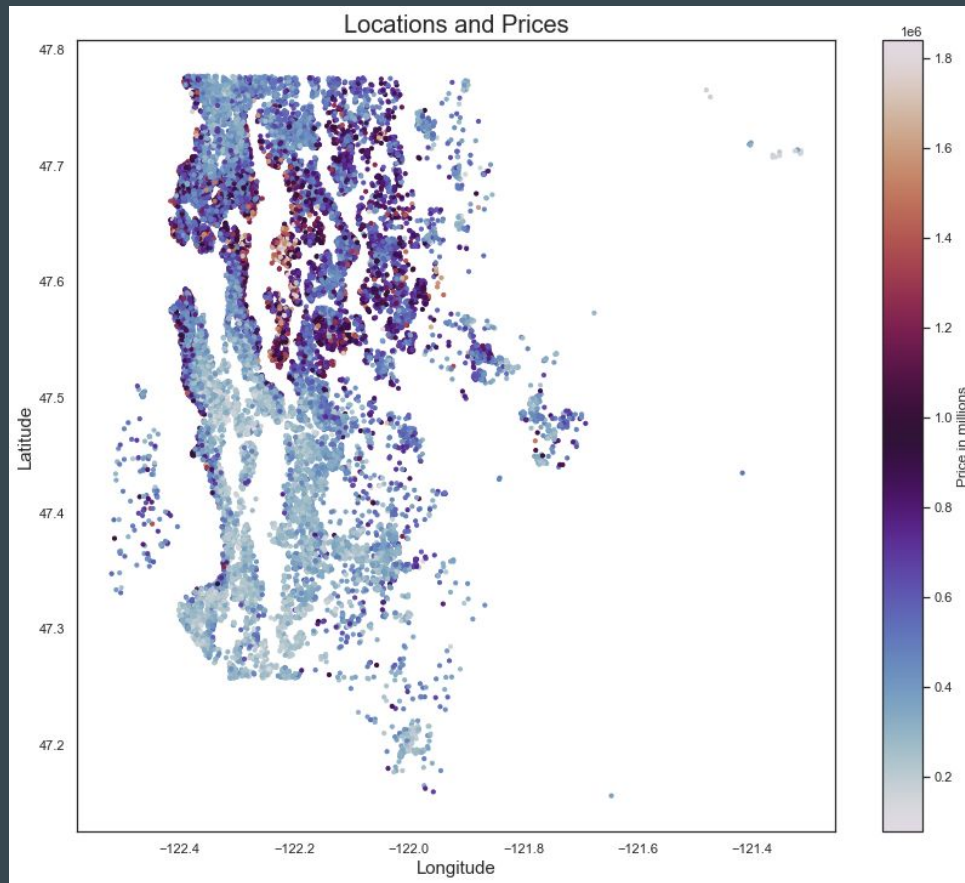
Sales per sqft of Living Area.

- We have a total variance of 78.2% of our features attributed to the price, explained by our model.
- For every 1% increase in sqft_living our price increase by 0.61%.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.782
Model:	OLS	Adj. R-squared:	0.781
Method:	Least Squares	F-statistic:	1532.
Date:	Tue, 08 Jun 2021	Prob (F-statistic):	0.00
Time:	16:31:18	Log-Likelihood:	762.86
No. Observations:	15879	AIC:	-1450.
Df Residuals:	15841	BIC:	-1158.
Df Model:	37		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-151.9193	4.886	-31.093	0.000	-161.496	-142.342
lat	1.2435	0.024	50.928	0.000	1.196	1.291
long	-0.8222	0.036	-22.582	0.000	-0.894	-0.751
sqft_basement	-6.173e-05	5.78e-06	-10.673	0.000	-7.31e-05	-5.04e-05
sqft_living_log	0.6134	0.010	63.736	0.000	0.595	0.632
sqft_lot_log	0.0508	0.003	17.660	0.000	0.045	0.056
Gra_Good	0.2235	0.006	36.111	0.000	0.211	0.236
Gra_Very Bad	-0.1244	0.007	-17.820	0.000	-0.138	-0.111
Con_Bad	-0.1180	0.021	-5.507	0.000	-0.160	-0.076
Con_Good	0.0775	0.004	18.022	0.000	0.069	0.086
Bed_Large	-0.0319	0.005	-6.933	0.000	-0.041	-0.023
Bed_Small	0.0676	0.006	10.538	0.000	0.055	0.080
Bat_High	0.0696	0.007	10.377	0.000	0.056	0.083
Bat_Med	0.0193	0.006	3.461	0.001	0.008	0.030
Bui_MED	0.0188	0.006	3.390	0.001	0.008	0.030
wat_1.0	0.4956	0.027	18.137	0.000	0.442	0.549
Rec_True	0.0401	0.006	6.654	0.000	0.028	0.052
C_Bellevue	0.4630	0.010	47.382	0.000	0.444	0.482
C_Black Diamond	0.3459	0.029	11.982	0.000	0.289	0.402
C_Bothell	-0.0928	0.022	-4.310	0.000	-0.135	-0.051
C_Carnation	0.1148	0.028	4.121	0.000	0.060	0.169
C_Enumclaw	0.3780	0.020	18.624	0.000	0.338	0.418
C_Fall City	0.4040	0.033	12.363	0.000	0.340	0.468

Results

Model performance

- As compared to our other models the final model has an outcome average error of $91,519.
- And the lowest average distance between our predicted values from the model and actual values in the data set (RMSE = 141,343)

Training Scores:
R2: 0.7827344660247418
Root Mean Squared Error: 135615.30440822302
Mean Absolute Error: 89542.21519979693

Testing Scores:
R2: 0.777342376471941
Root Mean Squared Error: 141343.56244896515
Mean Absolute Error: 91519.00345539676
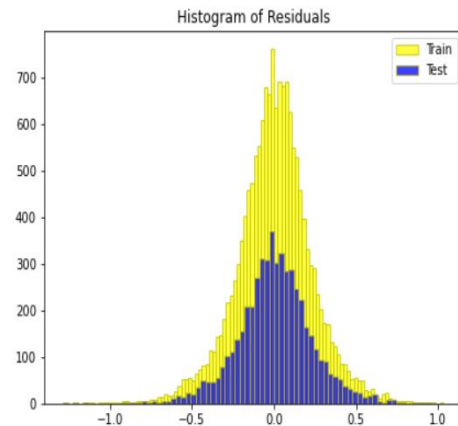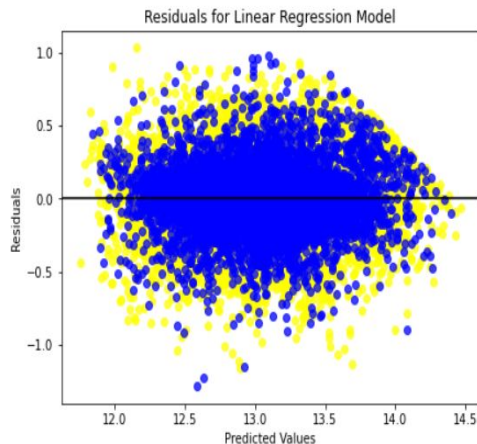
Recommendations

Square ft for Living Area	• Adding square footage to a property can add significant value to a house.
Location and Budget	• Direct buyers with higher budget and priority of living area to invest on properties towards the Northern region of Seattle.
Condition of House	• Engage only in transaction of homes with a 'Good' Condition.

Further Study

Number of rooms per sqft of living.	● Determine the value of different types of expansions (Bedrooms, Bathrooms) and investigate how that affects the sale price of a house.
Locations on the Rise for preferred settlement	● Identify areas of Seattle where housing prices are increasing and possibly predict which neighborhoods will be ideal for settlement in the future.
Updated data trends	● Improve the model with more data over the years after 2015 and observe if there is any change in the trend of the major factors.

Thank you!

Email : milena22peter@gmail.com
GitHub: @Milenaafeworki