King County House Sales

Regression Model

Summary

- **Aim:**

To build regression model to predict house prices based on property sales dataset in King County, Seattle, WA to aid investment decisions in real estates.

- Business problem:

Real estate agency wants to give accurate appraisal and buying/selling advices to their clients

- Data:

King County house sales dataset from 2014- 2015

Our focus

1. What features add value to properties?
2. Does renovation add to property value?
3. Does neighbourhood add values to property?

Approach

- 5 different multilinear regression models were built and evaluated
- Each independent variables/ features relationship with prices were analysed
- 5 significant features affect property values are: sqft_living, sqft_living15, bathrooms, grade and sqft_above

	id	price
id	1.000000	-0.016772
price	-0.016772	1.000000
bedrooms	0.001150	0.308787
bathrooms	0.005162	0.525906
sqft_living	-0.012241	0.701917
sqft_lot	-0.131911	0.089876
floors	0.018608	0.256804
waterfront	-0.003599	0.264306
view	0.011772	0.393497
condition	-0.023803	0.036056
grade	0.008188	0.667951
sqft_above	-0.010799	0.605368
sqft_basement	-0.004359	0.321108
yr_built	0.021617	0.053953
yr_renovated	-0.010621	0.117543
zipcode	-0.008211	-0.053402
lat	-0.001798	0.306692
long	0.020672	0.022036
sqft_living15	-0.002701	0.585241
sqft_lot15	-0.138557	0.082845
day_sold	0.002143	-0.014684
month_sold	-0.011572	-0.009928
year_sold	0.009915	0.003727

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.679e+07	9.97e+06	-6.699	0.000	-8.63e+07	-4.72e+07
id	-1.349e-06	4.81e-07	-2.802	0.005	-2.29e-06	-4.05e-07
bedrooms	-3.611e+04	1895.492	-19.051	0.000	-3.98e+04	-3.24e+04
bathrooms	4.188e+04	3253.716	12.871	0.000	3.55e+04	4.83e+04
sqft_living	103.7713	18.027	5.756	0.000	68.437	139.105
sqft_lot	0.1176	0.048	2.457	0.014	0.024	0.211
floors	7650.4990	3591.439	2.130	0.033	611.013	1.47e+04
waterfront	6.188e+05	1.81e+04	34.211	0.000	5.83e+05	6.54e+05
view	5.317e+04	2117.373	25.109	0.000	4.9e+04	5.73e+04
condition	2.804e+04	2344.312	11.959	0.000	2.34e+04	3.26e+04
grade	9.727e+04	2155.487	45.126	0.000	9.3e+04	1.01e+05
sqft_above	77.0543	18.018	4.276	0.000	41.737	112.372
sqft_basement	46.9031	17.858	2.626	0.009	11.900	81.907
yr_built	-2641.1743	71.733	-36.820	0.000	-2781.775	-2500.573
yr_renovated	4.836e+04	7919.305	6.107	0.000	3.28e+04	6.39e+04
zipcode	-584.8477	32.902	-17.776	0.000	-649.338	-520.358
lat	6.026e+05	1.07e+04	56.257	0.000	5.82e+05	6.24e+05
long	-2.155e+05	1.31e+04	-16.403	0.000	-2.41e+05	-1.9e+05
sqft_living15	21.5336	3.441	6.258	0.000	14.789	28.279
sqft_lot15	-0.4006	0.073	-5.463	0.000	-0.544	-0.257
day_sold	-360.8515	159.165	-2.267	0.023	-672.827	-48.876
month_sold	1129.2739	708.781	1.593	0.111	-259.988	2518.536
year_sold	3.656e+04	4721.714	7.744	0.000	2.73e+04	4.58e+04

Omnibus:	18404.912	Durbin-Watson:	1.990
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1877255.867
Skew:	3.576	Prob(JB):	0.00
Kurtosis:	48.111	Cond. No.	3.95e+13

Final Model

- Produce predicted value with almost 0 residual errors
- Can only explained 36-37% of variance in property prices

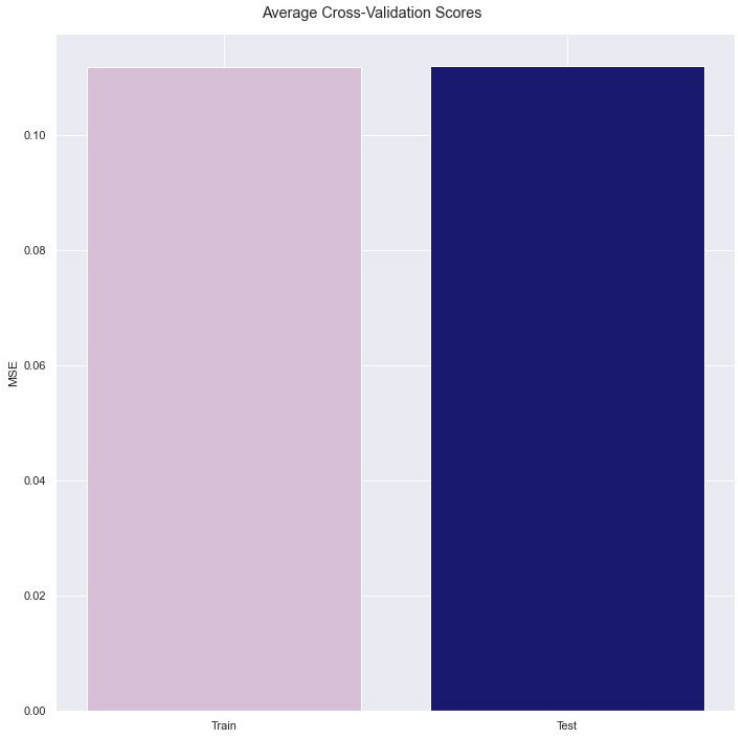
OLS Regression Results

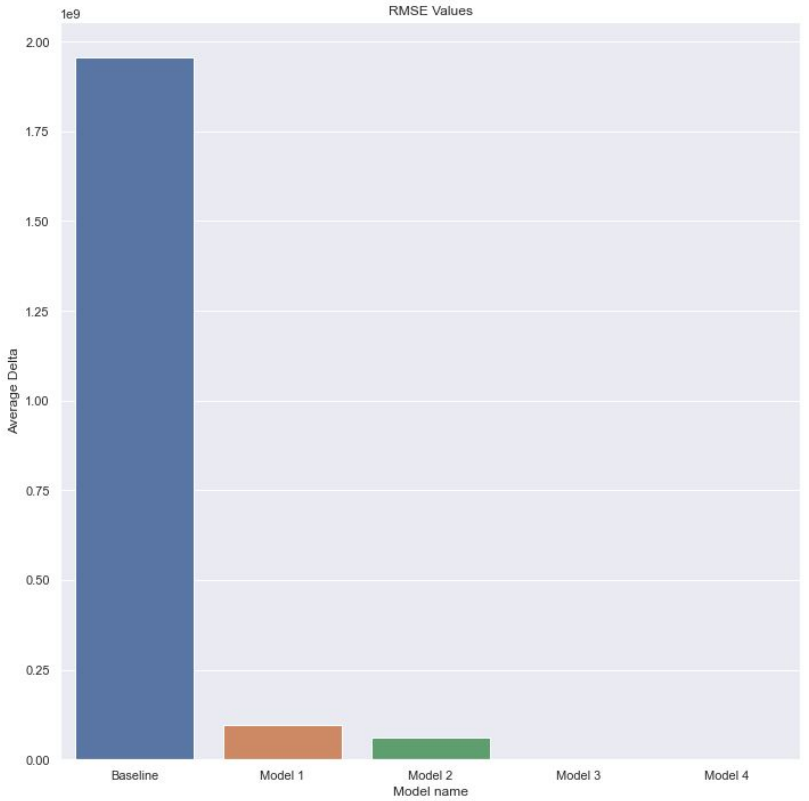
Dep. Variable:	price	**R-squared:**	0.369
Model:	OLS	**Adj. R-squared:**	0.369
Method:	Least Squares	**F-statistic:**	3689.
Date:	Thu, 09 Feb 2023	**Prob (F-statistic):**	0.00
Time:	19:26:17	**Log-Likelihood:**	-6128.9
No. Observations:	18945	**AIC:**	1.227e+04
Df Residuals:	18941	**BIC:**	1.230e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	12.9502	0.002	5329.884	0.000	12.945	12.955
sqft_living_sc	0.1088	0.004	29.947	0.000	0.102	0.116
sqft_living15_sc	0.0477	0.003	13.722	0.000	0.041	0.054
bathroom_sc	0.0672	0.002	41.346	0.000	0.064	0.070
grade_sc	0.0672	0.002	41.346	0.000	0.064	0.070

Omnibus:	210.682	**Durbin-Watson:**	1.974
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	136.179
Skew:	-0.055	**Prob(JB):**	2.69e-30
Kurtosis:	2.600	**Cond. No.**	1.75e+16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.89e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.


Average Cross-Validation Scores
MSE
0.00
0.02
0.04
0.06
0.08
0.10
Train
Test


RMSE Values
1e9
Average Delta
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Baseline
Model 1
Model 2
Model 3
Model 4
Model name

Conclusions

- Sqft_living - biggest feature
- Renovation/Extension can potentially add value to the property
- The size of the neighbouring houses have a positive impact on property prices

Limitations:

- Multilinear regression model perhaps not ideal - explore other models for this type of data
- Limitation of dataset only includes sale 2014-2015 - Further analysis into larger dataset with more data point from wider range of time

Thank You!

Email: yen.ho993@gmail.com
GitHub: @NBYH
LinkedIn: linkedin.com/in/yenho93/