The King of Home Renovation

By: Grant Edwards

Summary

King Renovators is looking for a way to help find homes that are below market value that they can buy, remodel, and sell for a profit, as well as what home features can help increase the properties values. King county Washington, the home to Seattle, is one of the hottest housing markets in the country with a lot of potential for profits to be made in buying homes that are below market, remodeling and adding features that improve the value of the property and selling for a profit. However, King county has a large number of properties and it would be impossible for someone to, quickly and thoroughly find homes that have potential to be profitable by hand. This is where we can use historical data to help predict the homes values based on the features and find homes that have potential to be profitable.

Why Build a Model?

Benefits:

- Identify what features in a home have the largest impact on the price of a home.
- Quickly identify potential properties that are below their predicted price and could be flipped or remodeled for a profit from the known variables.

Business Problem

Goal: Using historical data to identify opportunity to renovate and flip homes

- Can we accurately predict the price of a home based on the known variables?
- What features affect a home's value the most?
- Can we use our model to identify homes that are under their predicted market value?

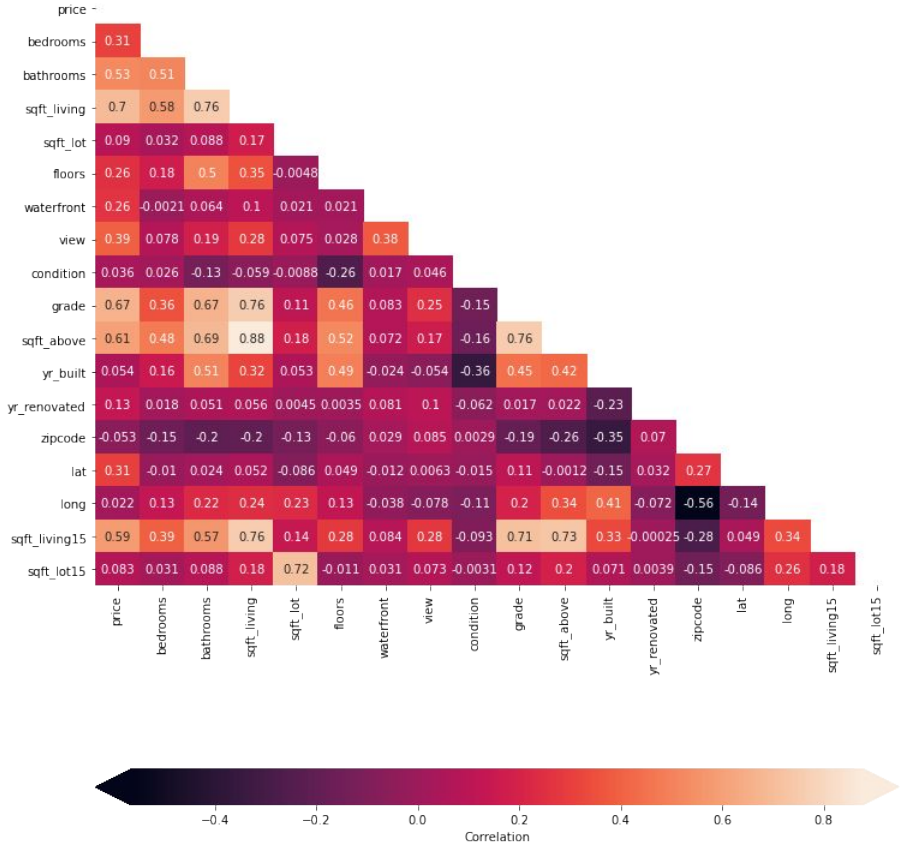
Data

- The data is historical home sales in King County, WA
- Contains information on various features of a home as well as the price.

Methods

- Used Python and Pandas in a Jupyter Notebook.
- Convert variables to numeric where needed so statistical analysis can be performed.
- Clean Data - remove outliers, clean missing values.
- Drop insignificant independent variables.
- Normalize data to get the same variance between variables.
- Reduce multicollinearity, heteroscedasticity, and improve normality.

Initial Look at Correlation of Data



- Focusing on price column we see the correlation to other features of the home.
- Square foot of living space (sqft_living) has highest correlation, with grade having a similar value.

Higher correlated values are lighter, or darker if there is a negative correlation. The closer a value is to 1 or -1, the more significant the variable while being closer to 0 indicates little to no correlation between the dependant and independent variable.

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.700
Model:	OLS	**Adj. R-squared:**	0.699
Method:	Least Squares	**F-statistic:**	3140.
Date:	Mon, 06 Feb 2023	**Prob (F-statistic):**	0.00
Time:	21:00:25	**Log-Likelihood:**	-2.9441e+05
No. Observations:	21597	**AIC:**	5.888e+05
Df Residuals:	21580	**BIC:**	5.890e+05
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.678e+06	2.93e+06	2.622	0.009	1.94e+06	1.34e+07
bedrooms	-3.624e+04	1901.570	-19.058	0.000	-4e+04	-3.25e+04
bathrooms	4.386e+04	3242.297	13.527	0.000	3.75e+04	5.02e+04
sqft_living	149.6716	4.399	34.022	0.000	141.049	158.295
sqft_lot	0.1258	0.048	2.623	0.009	0.032	0.220
floors	7934.2693	3600.589	2.204	0.028	876.848	1.5e+04
waterfront	6.23e+05	1.81e+04	34.349	0.000	5.87e+05	6.59e+05
view	5.363e+04	2123.300	25.256	0.000	4.95e+04	5.78e+04
condition	2.462e+04	2320.160	10.610	0.000	2.01e+04	2.92e+04
grade	9.73e+04	2161.253	45.022	0.000	9.31e+04	1.02e+05
sqft_above	31.1056	4.363	7.129	0.000	22.554	39.657
yr_built	-2758.4653	68.984	-39.987	0.000	-2893.678	-2623.252
zipcode	-588.4926	33.015	-17.825	0.000	-653.204	-523.781
lat	5.985e+05	1.07e+04	55.729	0.000	5.77e+05	6.2e+05
long	-2.153e+05	1.32e+04	-16.363	0.000	-2.41e+05	-1.9e+05
sqft_living15	20.7858	3.450	6.025	0.000	14.023	27.548
sqft_lot15	-0.3855	0.073	-5.252	0.000	-0.529	-0.242

Omnibus:	18377.350	**Durbin-Watson:**	1.988
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1847895.306
Skew:	3.572	**Prob(JB):**	0.00
Kurtosis:	47.749	**Cond. No.**	2.15e+08

First Model

We found that 70% of our home value is based on our known home features.

All features have a P-value less than 0.05, thus all features are currently significant.

Data appears to have multicollinearity and skewed

Distribution of Square Footage

Increasing variance on sale price as square footage increases.

This is not great for performing linear regression as is.

Shows that there is heteroskedasticity in our data

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.682
Model:	OLS	**Adj. R-squared:**	0.682
Method:	Least Squares	**F-statistic:**	3941.
Date:	Mon, 06 Feb 2023	**Prob (F-statistic):**	0.00
Time:	21:01:11	**Log-Likelihood:**	-2.6360e+05
No. Observations:	20206	**AIC:**	5.272e+05
Df Residuals:	20194	**BIC:**	5.273e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.647e+06	1.62e+06	1.631	0.103	-5.35e+05	5.83e+06
bedrooms	-1.013e+04	1113.212	-9.096	0.000	-1.23e+04	-7943.422
bathrooms	2.153e+04	1886.565	11.410	0.000	1.78e+04	2.52e+04
sqft_living	89.1670	1.905	46.812	0.000	85.434	92.901
floors	3.06e+04	1873.463	16.336	0.000	2.69e+04	3.43e+04
waterfront	1.435e+05	1.66e+04	8.663	0.000	1.11e+05	1.76e+05
view	3.251e+04	1366.774	23.786	0.000	2.98e+04	3.52e+04
condition	2.228e+04	1334.802	16.690	0.000	1.97e+04	2.49e+04
grade	7.936e+04	1219.064	65.097	0.000	7.7e+04	8.17e+04
yr_built	-1853.9186	39.753	-46.636	0.000	-1931.838	-1775.999
zipcode	-256.5777	16.724	-15.342	0.000	-289.358	-223.797
lat	5.411e+05	5982.036	90.457	0.000	5.29e+05	5.53e+05

Omnibus:	1319.480	**Durbin-Watson:**	1.973
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2183.617
Skew:	0.516	**Prob(JB):**	0.00
Kurtosis:	4.237	**Cond. No.**	2.02e+08

Refining the Model

Removed some variables so that we can reduce multicollinearity.

Removed outliers from our data to reduce skew and make our model more accurate and normal.

Standardize the Data

- Take individual values and subtract the mean of that variable, divide by standard deviation.
- Take the log of the price (the dependant variable)

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.699
Model:	OLS	**Adj. R-squared:**	0.699
Method:	Least Squares	**F-statistic:**	4263.
Date:	Mon, 06 Feb 2023	**Prob (F-statistic):**	0.00
Time:	21:01:50	**Log-Likelihood:**	189.61
No. Observations:	20206	**AIC:**	-355.2
Df Residuals:	20194	**BIC:**	-260.3
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	12.7874	0.010	1289.289	0.000	12.768	12.807
waterfront	0.3087	0.035	8.709	0.000	0.239	0.378
view	0.0678	0.003	23.193	0.000	0.062	0.074
condition	0.0497	0.003	17.327	0.000	0.044	0.055
bedrooms_log	-0.0307	0.002	-13.548	0.000	-0.035	-0.026
bathrooms_log	0.0339	0.003	11.665	0.000	0.028	0.040
sqft_living_log	0.1561	0.003	48.112	0.000	0.150	0.162
floors_log	0.0390	0.002	18.424	0.000	0.035	0.043
grade_log	0.1618	0.003	61.545	0.000	0.157	0.167
yr_built_log	-0.1018	0.002	-41.093	0.000	-0.107	-0.097
zipcode_log	-0.0273	0.002	-14.335	0.000	-0.031	-0.024
lat_log	0.1926	0.002	106.188	0.000	0.189	0.196

Omnibus:	365.419	**Durbin-Watson:**	1.983
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	749.643
Skew:	-0.044	**Prob(JB):**	1.65e-163
Kurtosis:	3.940	**Cond. No.**	75.8

Normalizing the Model

With our data normalized, we have reduced skew and multicollinearity.

Our model still supports that 70% of the homes price can be attributed to the variables being used. However with the work we have done we can be far more confident that our model is more accurate in predicting the value of a home.

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.699
Model:	OLS	**Adj. R-squared:**	0.699
Method:	Least Squares	**F-statistic:**	4263.
Date:	Mon, 06 Feb 2023	**Prob (F-statistic):**	0.00
Time:	21:01:50	**Log-Likelihood:**	189.61
No. Observations:	20206	**AIC:**	-355.2
Df Residuals:	20194	**BIC:**	-260.3
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	12.7874	0.010	1289.289	0.000	12.768	12.807
waterfront	0.3087	0.035	8.709	0.000	0.239	0.378
view	0.0678	0.003	23.193	0.000	0.062	0.074
condition	0.0497	0.003	17.327	0.000	0.044	0.055
bedrooms_log	-0.0307	0.002	-13.548	0.000	-0.035	-0.026
bathrooms_log	0.0339	0.003	11.665	0.000	0.028	0.040
sqft_living_log	0.1561	0.003	48.112	0.000	0.150	0.162
floors_log	0.0390	0.002	18.424	0.000	0.035	0.043
grade_log	0.1618	0.003	61.545	0.000	0.157	0.167
yr_built_log	-0.1018	0.002	-41.093	0.000	-0.107	-0.097
zipcode_log	-0.0273	0.002	-14.335	0.000	-0.031	-0.024
lat_log	0.1926	0.002	106.188	0.000	0.189	0.196

Omnibus:	365.419	**Durbin-Watson:**	1.983
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	749.643
Skew:	-0.044	**Prob(JB):**	1.65e-163
Kurtosis:	3.940	**Cond. No.**	75.8

The Most Influential Variables

Looking at the t-values for our independent variables, we can identify the highest influencers on the homes values. We can see that the latitude has a high impact on the sale price of a home, as well as the grade of the house and the square footage of the living space.

Finding potential Homes

	price	bedrooms	bathrooms	sqft_living	floors	waterfront	view	condition	grade	yr_built	zipcode	lat	resids
12539	90000.0	2	1.00	790	1.0	0	0	3	7.0	1973	98034	47.7351	-1.339540
2587	134000.0	2	1.50	980	2.0	0	0	3	7.0	1922	98014	47.7076	-1.292986
326	274975.0	3	2.50	3030	2.0	0	0	3	9.0	1987	98077	47.7721	-1.152139
12332	160000.0	2	1.00	1140	1.0	0	0	3	8.0	1980	98028	47.7637	-1.087647
1220	130000.0	3	1.00	1110	1.0	0	0	4	7.0	1960	98033	47.6830	-1.087423
18318	130000.0	3	1.00	1200	2.0	0	0	1	7.0	1908	98116	47.5883	-1.059802
7090	285000.0	4	3.50	2770	2.0	0	0	3	8.0	1940	98133	47.7412	-1.031400
16828	170000.0	1	0.75	850	1.0	0	2	3	6.0	1903	98019	47.7654	-1.029823
14255	130000.0	2	1.00	840	1.0	0	0	3	7.0	1951	98133	47.7319	-1.017974
9767	289275.0	3	2.00	2860	1.0	0	0	3	9.0	1985	98019	47.7718	-1.017711
16879	125000.0	3	1.00	1230	1.5	0	0	1	6.0	1916	98117	47.6941	-1.006182
12711	130000.0	3	1.00	1100	1.0	0	0	4	7.0	1913	98108	47.5231	-0.994533
4764	154000.0	2	1.00	1040	1.0	0	3	3	6.0	1949	98014	47.6981	-0.973902
16927	160000.0	3	1.00	1140	1.5	0	0	4	6.0	1910	98014	47.7093	-0.973152
7985	90000.0	1	1.00	780	1.0	0	0	3	5.0	1905	98108	47.5424	-0.947266
18973	140000.0	3	1.50	1200	2.0	0	0	3	8.0	1966	98055	47.4659	-0.940718
19173	345600.0	5	3.50	2800	2.5	0	0	3	9.0	1903	98122	47.6059	-0.938861
8267	82000.0	3	1.00	860	1.0	0	0	3	6.0	1954	98146	47.4987	-0.924726
5522	119500.0	3	1.00	1170	1.0	0	0	2	6.0	1980	98019	47.7346	-0.915658
18833	380000.0	3	2.50	1980	2.0	1	4	3	10.0	1984	98166	47.4551	-0.900046

Looking at the homes with the highest negative residual values, we can find some potential homes that would have potential for remodel or resale for a profit.

Just looking at the most influential variables, we can find homes at a higher latitude, that are a higher grade and have higher square footage of living space, to find several homes that have high potential for a profit.

Conclusions

Through our linear regression model we found that we can get a pretty decent idea on the price of the home.

We were also able to find that the latitude (location, location, location), the grade, and the square footage of the homes were the largest predictors in the price of the home.

We were also able to identify some potential homes that the price was well below the predicted values. This could be used to shortlist homes that could potentially be good investment opportunities for the business.

Thank You!

Email: grantedwards11@gmail.com
GitHub: @gzedwards
LinkedIn: www.linkedin.com/in/grant-edwards-25206914a