King County Real Estate Project

What’s your house really worth?

Presentor: Zach Hyde

Overview

H_0: The square footage of a property will not have a significant increase of the price.

H_1: The square footage of a house will have a significant increase of the price.

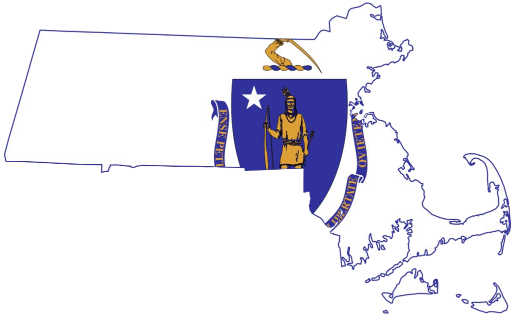
H_0: Having a waterfront feature will not increase the value of the property.

H_1: Having a waterfront feature will increase the value of the property.

H_0: The older the building is, the higher the value.

H_1: The newer the building is, the higher the value.

Data Utilization



Data utilized was provided via Flatiron School - ‘kc_housing.csv’

Facts about dataset:

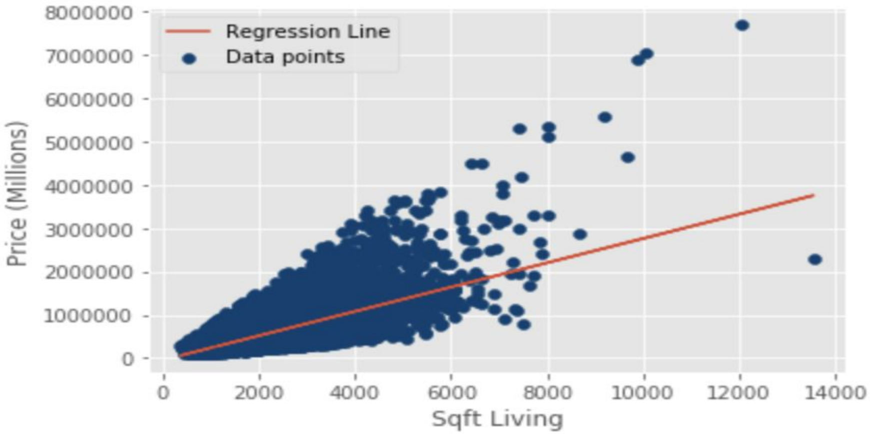
Volume - 21,597 listings

Value Range - $78,000 - $7.7M

Size Range - 370 sqft. - 13,540 sqft.

	bedrooms	grade	H2O_1_0	condition	sqft_living	yr_built	yr_renovated	price
count	21597.000000	21597.000000	21597.000000	21597.000000	21597.000000	21597.000000	21597.000000	2.159700e+04
mean	3.373200	7.657915	0.006760	3.409825	2080.321850	1970.999676	1972.945131	5.402966e+05
std	0.926299	1.173200	0.081944	0.650546	918.106125	29.375234	28.945393	3.673681e+05
min	1.000000	3.000000	0.000000	1.000000	370.000000	1900.000000	1900.000000	7.800000e+04
25%	3.000000	7.000000	0.000000	3.000000	1430.000000	1951.000000	1954.000000	3.220000e+05
50%	3.000000	7.000000	0.000000	3.000000	1910.000000	1975.000000	1977.000000	4.500000e+05
75%	4.000000	8.000000	0.000000	4.000000	2550.000000	1997.000000	1999.000000	6.450000e+05
max	33.000000	13.000000	1.000000	5.000000	13540.000000	2015.000000	2015.000000	7.700000e+06

EDA

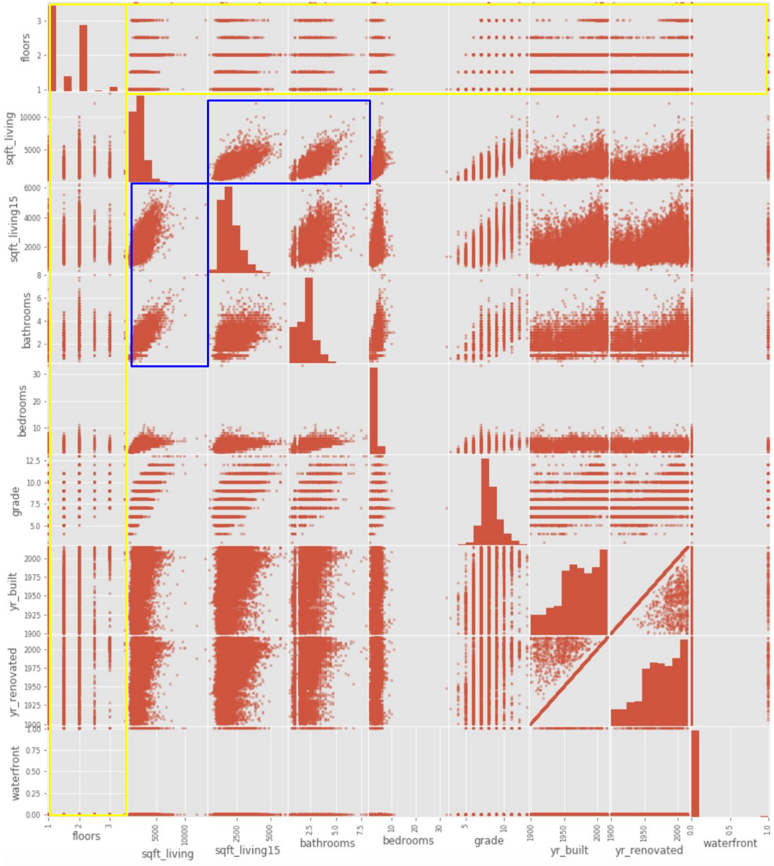



```
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   date           21597 non-null  object
 1   price          21597 non-null  float64
 2   bedrooms       21597 non-null  int64
 3   bathrooms      21597 non-null  float64
 4   sqft_living    21597 non-null  int64
 5   sqft_lot       21597 non-null  int64
 6   floors         21597 non-null  float64
 7   waterfront     19221 non-null  float64
 8   view           21534 non-null  float64
 9   condition      21597 non-null  int64
 10  grade          21597 non-null  int64
 11  sqft_above     21597 non-null  int64
 12  sqft_basement  21597 non-null  object
 13  yr_built       21597 non-null  int64
 14  yr_renovated   17755 non-null  float64
 15  zipcode        21597 non-null  int64
 16  lat            21597 non-null  float64
 17  long           21597 non-null  float64
 18  sqft_living15  21597 non-null  int64
 19  sqft_lot15     21597 non-null  int64
dtypes: float64(8), int64(10), object(2)
```

- Cleaned NaN values by either filling with 0 if categorical (waterfront) or replacing them with new value (yr_remodled)
- Simple linear regression to understand core belief that size of house will have positive correlation to price

Modeling

- floors doesn't add value to data
- sqft_living15 & bathrooms have multicollinearity to sqft_living



OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.650
Model:	OLS	**Adj. R-squared:**	0.650
Method:	Least Squares	**F-statistic:**	2866.
Date:	Sun, 01 Nov 2020	**Prob (F-statistic):**	0.00
Time:	20:54:28	**Log-Likelihood:**	-2.9605e+05
No. Observations:	21597	**AIC:**	5.921e+05
Df Residuals:	21582	**BIC:**	5.922e+05
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.681e+06	1.5e+05	44.506	0.000	6.39e+06	6.97e+06
bathrooms	5.362e+04	3477.037	15.421	0.000	4.68e+04	6.04e+04
bedrooms	-3.974e+04	2042.755	-19.455	0.000	-4.37e+04	-3.57e+04
grade	1.226e+05	2256.550	54.333	0.000	1.18e+05	1.27e+05
h2ofront_dummies	7.484e+05	1.83e+04	41.000	0.000	7.13e+05	7.84e+05
condition	1.838e+04	2510.038	7.324	0.000	1.35e+04	2.33e+04
sqft_living	164.4018	3.578	45.943	0.000	157.388	171.416
sqft_living15	36.2283	3.560	10.176	0.000	29.250	43.207
yr_built	-4184.6660	136.931	-30.560	0.000	-4453.061	-3916.271
yr_renovated	357.7732	141.349	2.531	0.011	80.718	634.829
floors_1_5	-6854.6759	5713.301	-1.200	0.230	-1.81e+04	4343.816
floors_2_0	-7346.6034	4150.909	-1.770	0.077	-1.55e+04	789.484
floors_2_5	1.238e+05	1.76e+04	7.053	0.000	8.94e+04	1.58e+05
floors_3_0	1.424e+05	9878.872	14.411	0.000	1.23e+05	1.62e+05
floors_3_5	2.507e+05	8.23e+04	3.048	0.002	8.95e+04	4.12e+05

Omnibus:	16138.034	**Durbin-Watson:**	1.978
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1091931.789
Skew:	3.002	**Prob(JB):**	0.00
Kurtosis:	37.313	**Cond. No.**	4.15e+05

Modeling continued

- All variables were significant (p-value < 0.05)
- More bedrooms = negative effect
- Older building = negative effect
- Waterfront gave highest value increase

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.640
Model:	OLS	**Adj. R-squared:**	0.640
Method:	Least Squares	**F-statistic:**	5480.
Date:	Sun, 01 Nov 2020	**Prob (F-statistic):**	0.00
Time:	20:54:29	**Log-Likelihood:**	-2.9636e+05
No. Observations:	21597	**AIC:**	5.927e+05
Df Residuals:	21589	**BIC:**	5.928e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.557e+06	1.25e+05	44.590	0.000	5.31e+06	5.8e+06
bedrooms	-3.52e+04	2021.289	-17.414	0.000	-3.92e+04	-3.12e+04
grade	1.363e+05	2129.487	64.006	0.000	1.32e+05	1.4e+05
h2ofront_dummies	7.581e+05	1.85e+04	40.947	0.000	7.22e+05	7.94e+05
condition	2.008e+04	2519.892	7.970	0.000	1.51e+04	2.5e+04
sqft_living	195.1264	2.952	66.110	0.000	189.341	200.912
yr_built	-3943.3168	136.357	-28.919	0.000	-4210.588	-3676.046
yr_renovated	684.8684	140.986	4.858	0.000	408.526	961.210

Omnibus:	15750.942	**Durbin-Watson:**	1.973
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	992747.271
Skew:	2.910	**Prob(JB):**	0.00
Kurtosis:	35.701	**Cond. No.**	2.93e+05

Conclusions

Variable	Coefficient
Waterfront	$ 758,100
Grade	$136,300
Condition	$20,080
Year Remodeled	$684
Sqft Living	$195
Year Built	-$3,943

With these coefficients, the null hypotheses can be rejected due to all p-value variables < 0.05. Having a waterfront will drastically increase property value. Having a larger home will, slightly, increase value and having an older home decreases the value.

RMSE: 448,139.02

Future Analysis

- Expand comparables to greater than 15 closest neighbors. Entire county or zip code?
- Only evaluate properties that do not have waterfront features due to extreme value and rarity.
- Categorize properties based on decade built for comparison within each decade
- Restrict model to accurately predict average sized houses (1,500-2,500 sqft. living)

Thank You!

References:

StackOverflow
Introduction to Linear Regression in Python | by Lorraine Li
Google’s 7 steps of Machine Learning in practice: a TensorFlow example for structured data
What is One-Hot Encoding and how to use Pandas get_dummies function

Contact information:

Email: zacharyhyde14@gmail.com
Github: zhyde23
Linkedin: Profile