# King County Real Estate Project

What's your house really worth?


By: 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:

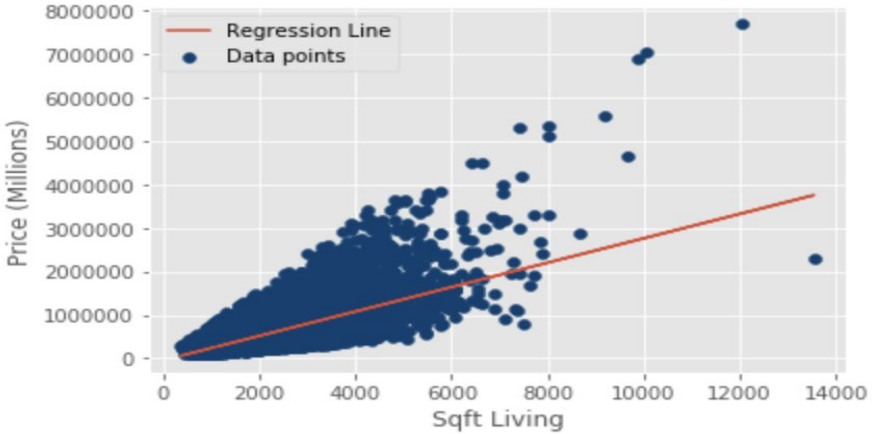
Volume - 21,597 listings

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

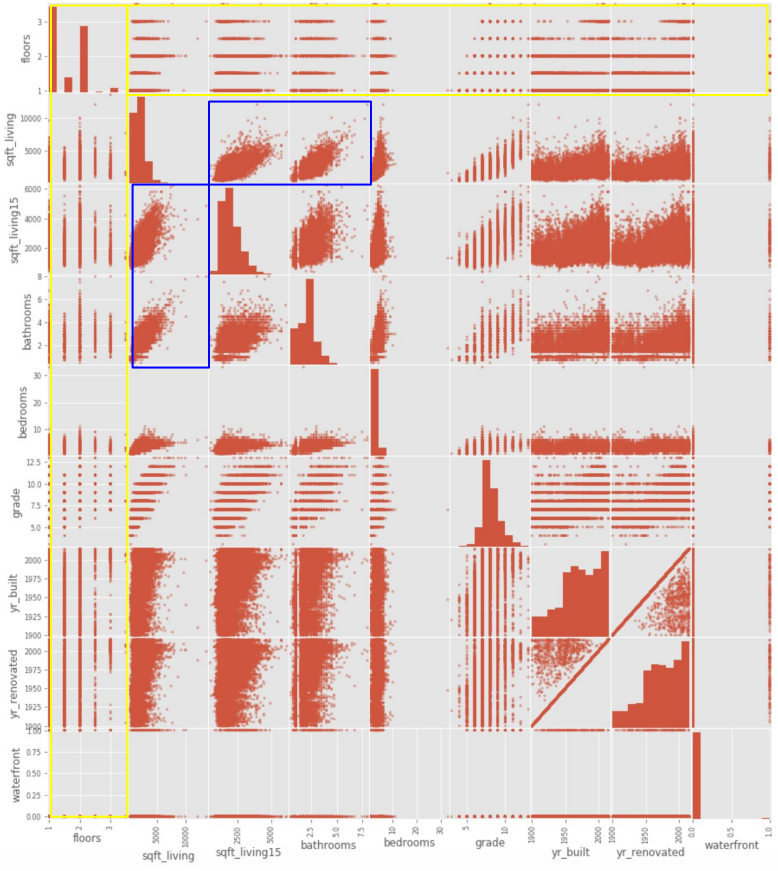



- Cleaned NaN values by either filling with 0 if categorical (waterfront) or replacing them with new value (yr_remodled)
- Simple linear regression to understand core belief that size of house will have positive correlation to price

```
 #   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)
```

# 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)
- Older building = negative effect
- Waterfront gave highest value increase
- Sqft_living = positive effect

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | price | **R-squared:** | 0.645 |
| **Model:** | OLS | **Adj. R-squared:** | 0.645 |
| **Method:** | Least Squares | **F-statistic:** | 3922. |
| **Date:** | Fri, 06 Nov 2020 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 09:40:19 | **Log-Likelihood:** | -2.9621e+05 |
| **No. Observations:** | 21597 | **AIC:** | 5.924e+05 |
| **Df Residuals:** | 21586 | **BIC:** | 5.925e+05 |
| **Df Model:** | 10 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 5.844e+06 | 1.26e+05 | 46.456 | 0.000 | 5.6e+06 | 6.09e+06 |
| **bedrooms** | -3.436e+04 | 2007.735 | -17.112 | 0.000 | -3.83e+04 | -3.04e+04 |
| **grade** | 1.321e+05 | 2127.986 | 62.059 | 0.000 | 1.28e+05 | 1.36e+05 |
| **h2ofront_dummies** | 7.535e+05 | 1.84e+04 | 40.978 | 0.000 | 7.17e+05 | 7.9e+05 |
| **condition** | 2.057e+04 | 2502.155 | 8.220 | 0.000 | 1.57e+04 | 2.55e+04 |
| **sqft_living** | 200.4878 | 2.958 | 67.779 | 0.000 | 194.690 | 206.286 |
| **yr_built** | -4000.3482 | 135.725 | -29.474 | 0.000 | -4266.380 | -3734.317 |
| **yr_renovated** | 602.1300 | 140.067 | 4.299 | 0.000 | 327.588 | 876.672 |
| **floors_2_5** | 1.283e+05 | 1.75e+04 | 7.348 | 0.000 | 9.41e+04 | 1.63e+05 |
| **floors_3_0** | 1.471e+05 | 9306.570 | 15.809 | 0.000 | 1.29e+05 | 1.65e+05 |
| **floors_3_5** | 2.601e+05 | 8.28e+04 | 3.142 | 0.002 | 9.78e+04 | 4.22e+05 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 15725.019 | **Durbin-Watson:** | 1.980 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 979586.921 |
| **Skew:** | 2.907 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 35.478 | **Cond. No.** | 2.98e+05 |

# Conclusions

| Variable | Coefficient(Effect) |
|---|---|
| Waterfront | $ 753,500 |
| Grade | $132,100 |
| Condition | $20,570 |
| Year Remodeled | $602 |
| Sqft Living | $200 |
| Year Built | -$4,000 |


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 increase value and each year your home ages, it decreases value.

RMSE: $211,235.91

# 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