House Sales in King County, USA

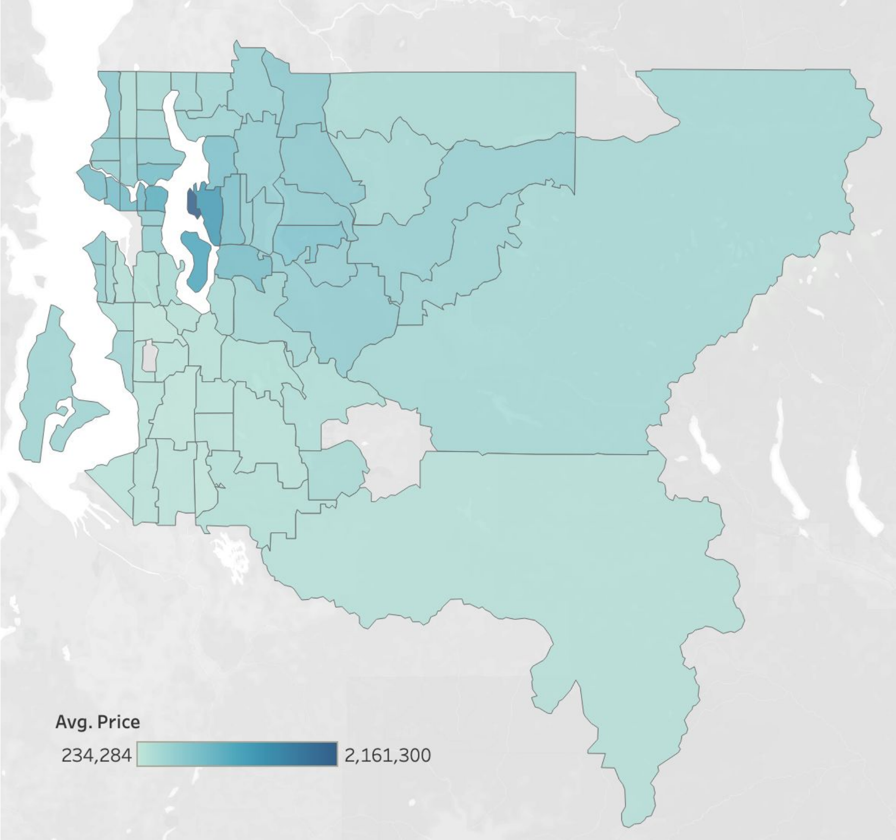
Deep dive on variables that affecting the housing price and housing trend in Washington State

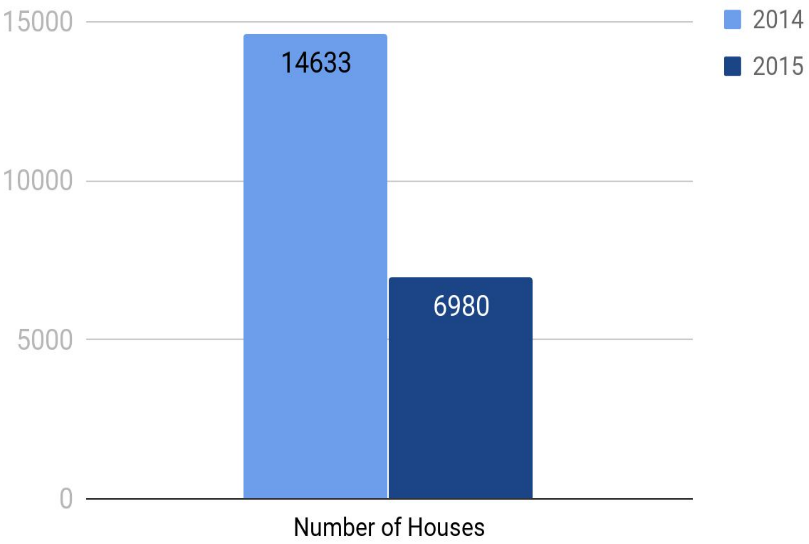
Aaron Bossard, Harry Heublum, Mer Arnel Manahan & Hansen Xu

Housing Price in King County

- Overview of the sample population
- Data Exploration/Cleanup
- Data source
- Identify variables
- Regression and visualization
- Prediction and trend

Overview





21,613 houses sold during May 2014 to May 2015

Average price range from **$ 234,284 to $ 2,161,300**

Data Cleanup

```
house_df = pd.read_csv("kc_house_data.csv")
```

```
house_df.head()
```

	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_above	sqft_basement	yr_built	yr_renovated	z
	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	0	...	7	1180	0	1955	0	
	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	0	...	7	2170	400	1951	1991	
	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	0	...	6	770	0	1933	0	
	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	0	...	7	1050	910	1965	0	
	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	0	...	8	1680	0	1987	0	

Month and Year should be parsed out for further exploratory analysis

Data Cleanup

Required Dependency is ‘datetime’

```
house_df['date'] = pd.to_datetime(house_df['date'], format = '%Y-%m-%d')
```

Results in the following ‘date’ column

```
house_df.head()
```

date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	
2014-10-13	221900.0	3	1.00	1180	5650	1.0	0	0	...	7	1180	0	1955	0	98178	47.
2014-12-09	538000.0	3	2.25	2570	7242	2.0	0	0	...	7	2170	400	1951	1991	98125	47.
2015-02-25	180000.0	2	1.00	770	10000	1.0	0	0	...	6	770	0	1933	0	98028	47.
2014-12-09	604000.0	4	3.00	1960	5000	1.0	0	0	...	7	1050	910	1965	0	98136	47.
2015-02-18	510000.0	3	2.00	1680	8080	1.0	0	0	...	8	1680	0	1987	0	98074	47.

Data Cleanup

Add columns to aid in further analysis

```
house_df['month'] = house_df['date'].dt.month
house_df['year'] = house_df['date'].dt.year
house_df['age'] = house_df['year'] - house_df['yr_built']
```

```
house_df.head()
```

oms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15	month	year	age
3	1.00	1180	5650	1.0	0	0	...	1955	0	98178	47.5112	-122.257	1340	5650	10	2014	59
3	2.25	2570	7242	2.0	0	0	...	1951	1991	98125	47.7210	-122.319	1690	7639	12	2014	63
2	1.00	770	10000	1.0	0	0	...	1933	0	98028	47.7379	-122.233	2720	8062	2	2015	82
4	3.00	1960	5000	1.0	0	0	...	1965	0	98136	47.5208	-122.393	1360	5000	12	2014	49
3	2.00	1680	8080	1.0	0	0	...	1987	0	98074	47.6168	-122.045	1800	7503	2	2015	28

Data Cleanup

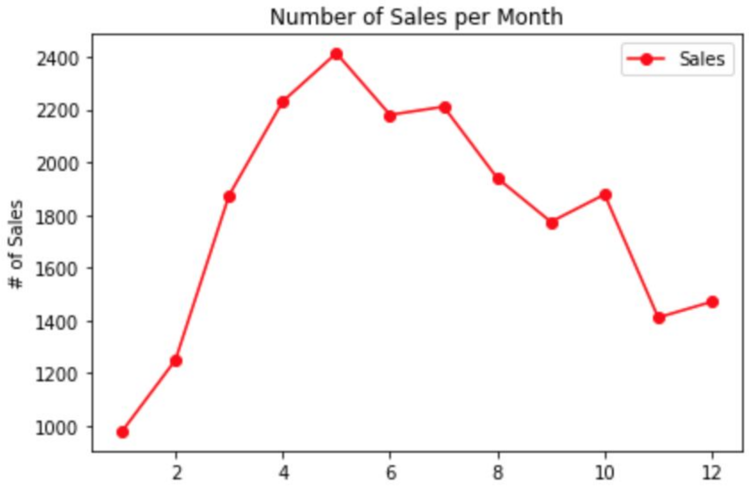
```
predicted_df['price'] = predicted_df['price'].map('${:,.2f}'.format)
```

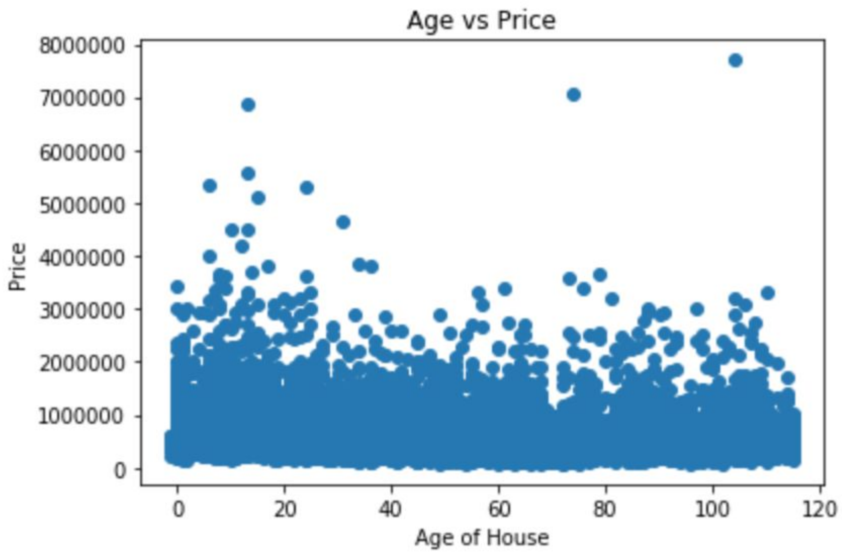
```
predicted_df['predicted'] = predicted_df['predicted'].map('${:,.2f}'.format)
```

```
predicted_df.head()
```

ing	sqft_lot	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15	month	year	age	price	predicted
60	9711	1060	0	1963	0	98198	47.4095	-122.315	1650	9711	1	2015	52	$291,850.00	$274,170.60
80	7470	1050	730	1960	0	98146	47.5123	-122.337	1780	8113	4	2015	55	$229,500.00	$390,993.50
30	19901	1430	0	1927	0	98028	47.7558	-122.229	1780	12697	5	2014	88	$310,000.00	$462,398.30
90	14040	1890	0	1994	0	98019	47.7277	-121.962	1890	14018	7	2014	21	$395,000.00	$347,725.70
00	9850	1200	0	1921	0	98002	47.3089	-122.210	1060	5095	12	2014	94	$189,000.00	$456,679.00

Data Exploration

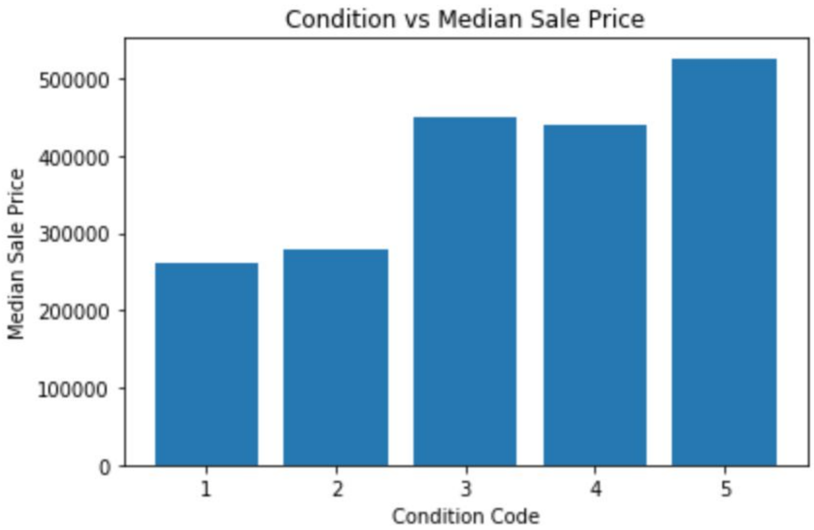


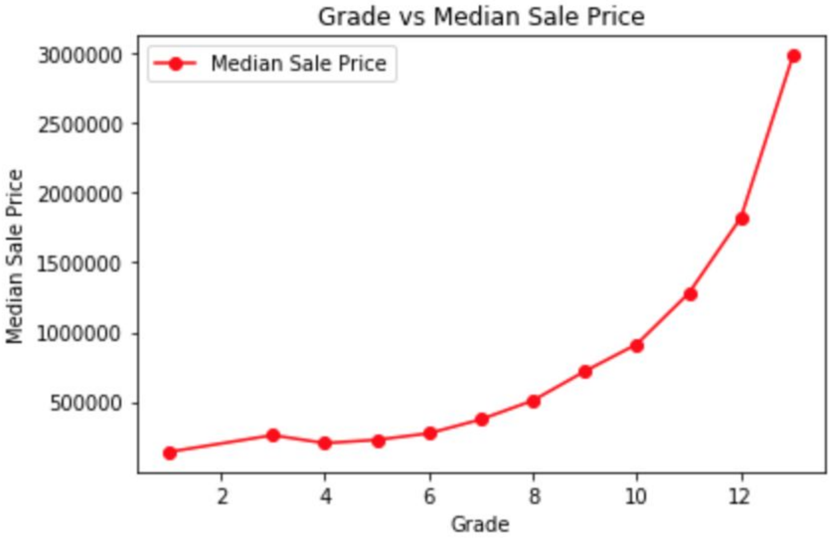


Most sales happen in the end of spring and beginning of summer

Slight distinction showing higher prices for younger homes

Data Exploration





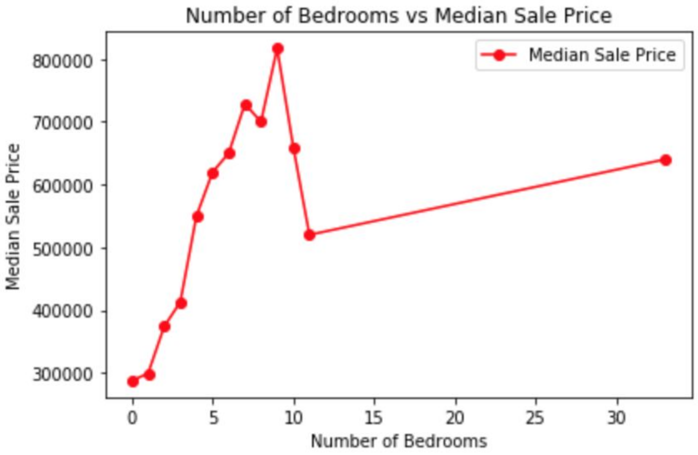
As condition code and grade increase, so does the median sale price

Data Exploration

found outlier

```
plt.plot(bdroom_x, bdroom_price_y, color="red", label='Median Sale Price',marker='o')
plt.legend()

plt.ylabel("Median Sale Price")
plt.xlabel("Number of Bedrooms")
plt.legend(loc="best")
plt.title("Number of Bedrooms vs Median Sale Price")
plt.tight_layout()
```



Data Exploration

Find row with this outlier and check

House has 33 bedrooms, but only 1.75 baths and 1 floor. Also, what is 1.75 bathroom???

```
house_df.loc[house_df['bedrooms'] == 33]
```

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	yr_built	yr_renovated	zipcode	lat	long
15870	2402100895	2014-06-25	640000.0	33	1.75	1620	6000	1.0	0	0	...	1947	0	98103	47.6878	-122.331

1 rows × 24 columns

What is a full bath?

A full bathroom is made up of four parts: a sink, a shower, a bathtub, and a toilet. Anything less than that, and you can't officially consider it a full bath.

www.realtor.com

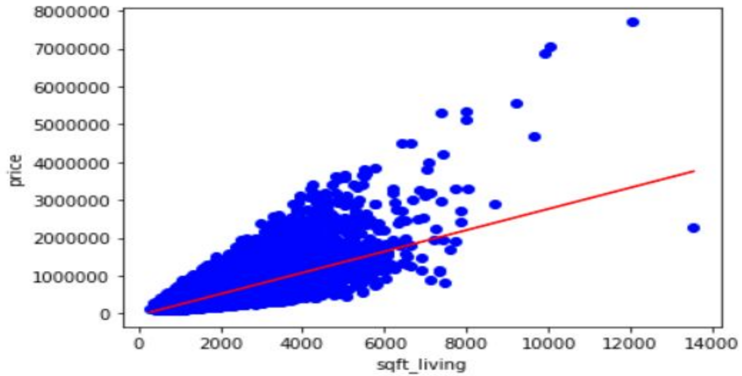
12

Total Variables

- Number of Bedrooms
- Number of Bathrooms
- Square feet of Living
- Number of Floors
- Waterfront
- View
- Grade
- Square feet of Above
- Square feet of Basement
- Year built
- 2015 Square feet of Living
- 2015 Square feet of Lot

Regression comparing price and sqft_living

```
### BEGIN SOLUTION
plt.scatter(X, y, c='blue')
plt.plot([x_min[0], x_max[0]], [y_min[0], y_max[0]], c='red')
plt.xlabel('sqft_living')
plt.ylabel('price')
plt.tight_layout()
### END SOLUTION
```

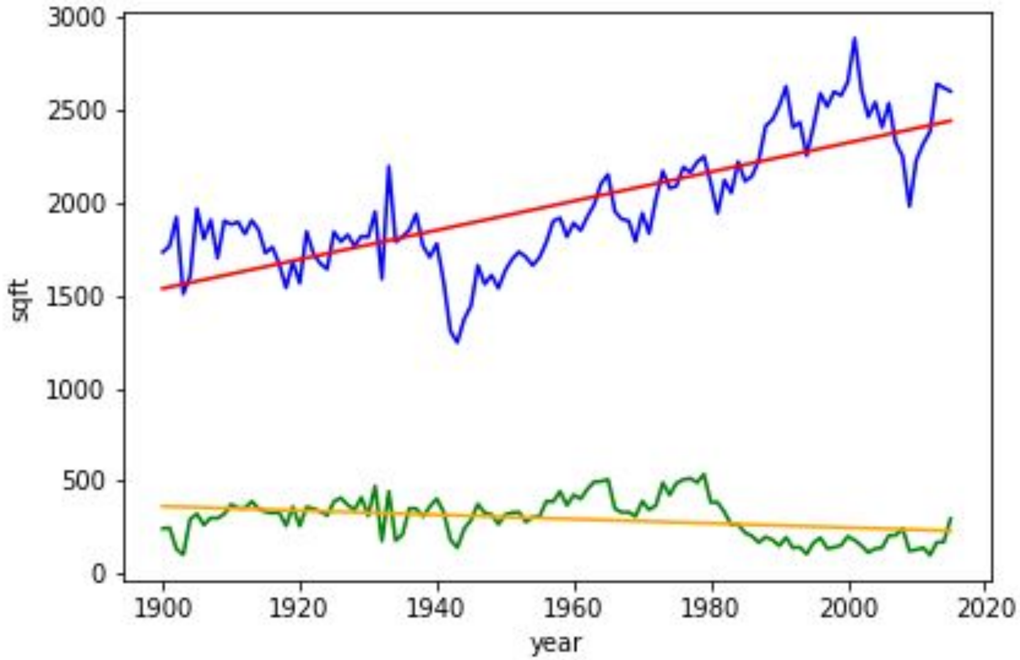



```
#This trend line shows that generally as sqft_living increases, price of the house increases.
#Other variables can affect the price of the house as well though.
#This is a simple linear regression that we did based on the equation y = b0+b1X
#Our graph would have the equation "price = b0+b1(sqft_living)".
#Price is the dependent variable, and sqft_living is an independent variable.
#If we wanted to get even more complex we could have added more independent variables to do multiple regre
ssion.
```

Multiple Regression - this model explains 65.3% of the variance in the dependent variable

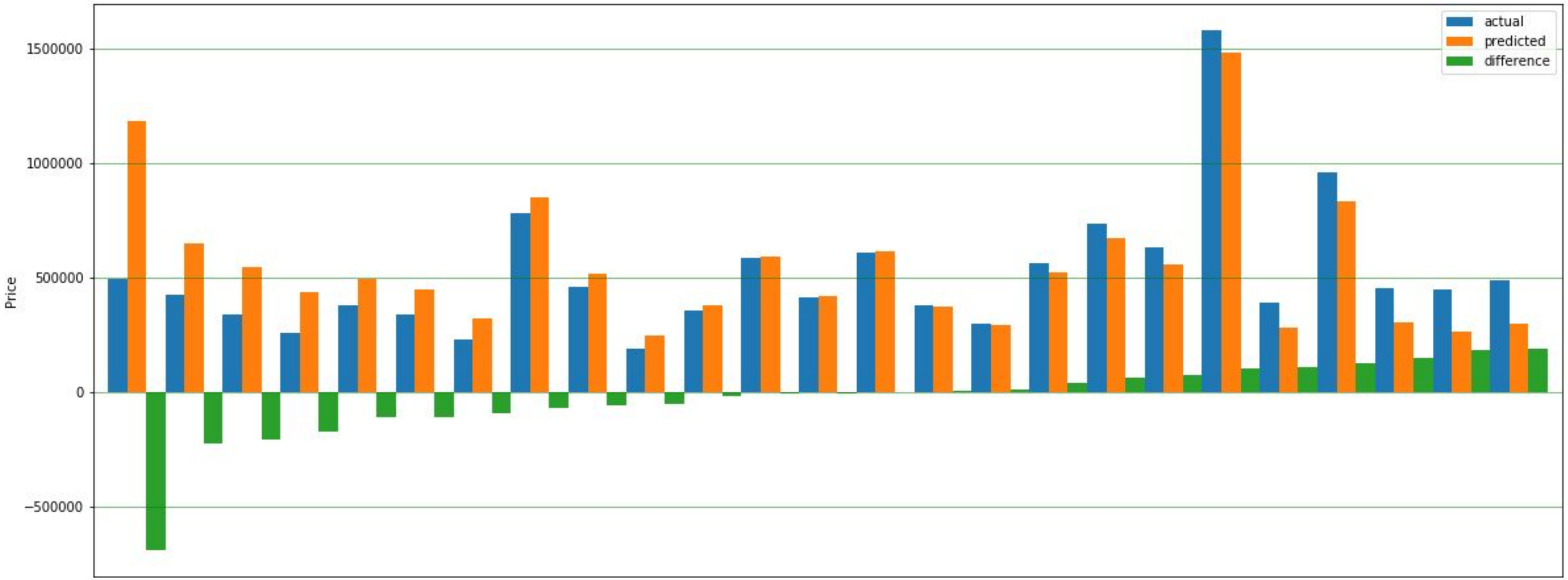
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.653
Model:                            OLS   Adj. R-squared:                  0.653
Method:                 Least Squares   F-statistic:                     3692.
Date:                Sat, 28 Sep 2019   Prob (F-statistic):               0.00
Time:                        12:21:11   Log-Likelihood:            -2.9619e+05
No. Observations:               21613   AIC:                         5.924e+05
Df Residuals:                   21601   BIC:                         5.925e+05
Df Model:                          11
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          6.643e+06   1.24e+05     53.710      0.000     6.4e+06    6.89e+06
bedrooms       -3.87e+04   2026.434    -19.099      0.000   -4.27e+04   -3.47e+04
bathrooms      4.817e+04   3464.934     13.903      0.000    4.14e+04     5.5e+04
sqft_living     109.7526      2.436     45.047      0.000     104.977     114.528
floors         2.462e+04   3769.055      6.533      0.000    1.72e+04     3.2e+04
waterfront     5.823e+05   1.86e+04     31.241      0.000    5.46e+05    6.19e+05
view           4.349e+04   2275.709     19.109      0.000     3.9e+04    4.79e+04
grade            1.2e+05   2253.019     53.275      0.000    1.16e+05    1.24e+05
sqft_above       50.7542      2.351     21.587      0.000      46.146      55.363
sqft_basement    58.9984      2.782     21.210      0.000      53.546      64.451
yr_built      -3765.7403     65.017    -57.919      0.000   -3893.179   -3638.302
sqft_living15    24.1480      3.599      6.710      0.000      17.094      31.202
sqft_lot15       -0.5419      0.056     -9.684      0.000      -0.652      -0.432
==============================================================================
Omnibus:                    16207.762   Durbin-Watson:                   1.980
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1161197.705
Skew:                           3.000   Prob(JB):                         0.00
Kurtosis:                      38.404   Cond. No.                     1.39e+17
==============================================================================
```

Sq Ft of Living vs Sq Ft of Basement



21,613 houses built between **1900 and 2015**

Price Prediction by Multi Linear Regression




```
Mean Absolute Error: 136243.38814812695
Mean Squared Error: 42340574552.13962
Root Mean Squared Error: 205768.25448095636
R Square: 0.6444505688835578
```

Improvements / Next Steps

Cloud Services:

- Use cloud storage for data
- Perform ETL using ZEPL w/ PySpark
- Host Transformed data on cloud database
- Create API

Build web application to allow user to enter either square footage, number bedrooms or other features in dataset to predict housing price.