



Exploring King County Housing Data



King County Housing Dataset

- KC housing dataset after data cleaning
 - The KC housing set has a list of homes in the city with id numbers and columns with various identifying factors for each. In this project, I seek to find a relationship between one or more of the factors and the price of the house.

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_above	sqft_basement	yr_built
0	7129300520	10/13/2014	221900.0	3	1.00	1180	5650	1.0	0.0	0.0	...	7	1180	0.0	1955
1	6414100192	12/9/2014	538000.0	3	2.25	2570	7242	2.0	0.0	0.0	...	7	2170	400.0	1951
2	5631500400	2/25/2015	180000.0	2	1.00	770	10000	1.0	0.0	0.0	...	6	770	0.0	1933
3	2487200875	12/9/2014	604000.0	4	3.00	1960	5000	1.0	0.0	0.0	...	7	1050	910.0	1965
4	1954400510	2/18/2015	510000.0	3	2.00	1680	8080	1.0	0.0	0.0	...	8	1680	0.0	1987

```
Index(['id', 'date', 'price', 'bedrooms', 'bathrooms', 'sqft_living',  
      'sqft_lot', 'floors', 'waterfront', 'view', 'condition', 'grade',  
      'sqft_above', 'sqft_basement', 'yr_built', 'yr_renovated', 'zipcode',  
      'lat', 'long', 'sqft_living15', 'sqft_lot15'],  
      dtype='object')
```

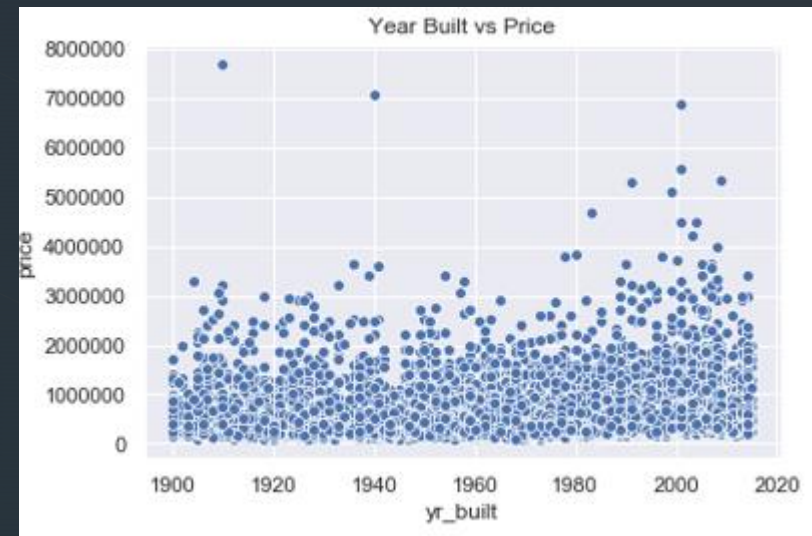
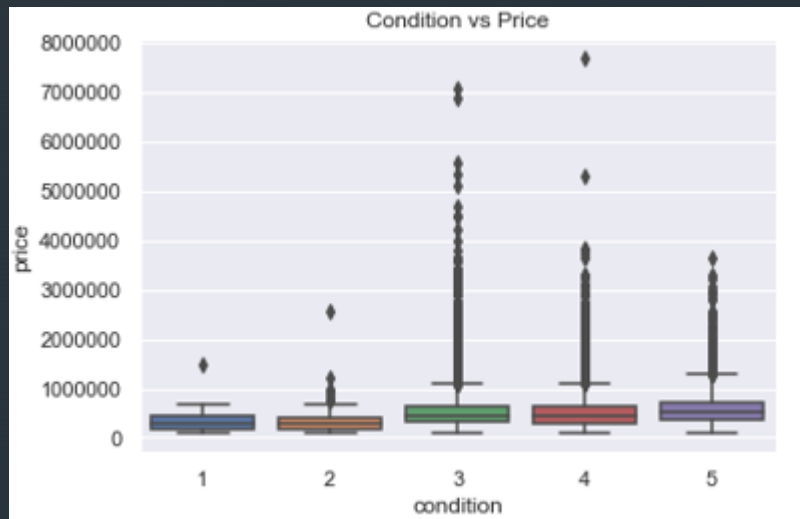
Data Visualization

- Are larger homes more expensive? Are homes on the water typically more expensive?
 - There appears to be a linear relationship between sqft_above and price and homes on the waterfront (value of 1) appear to generally have higher prices as well.



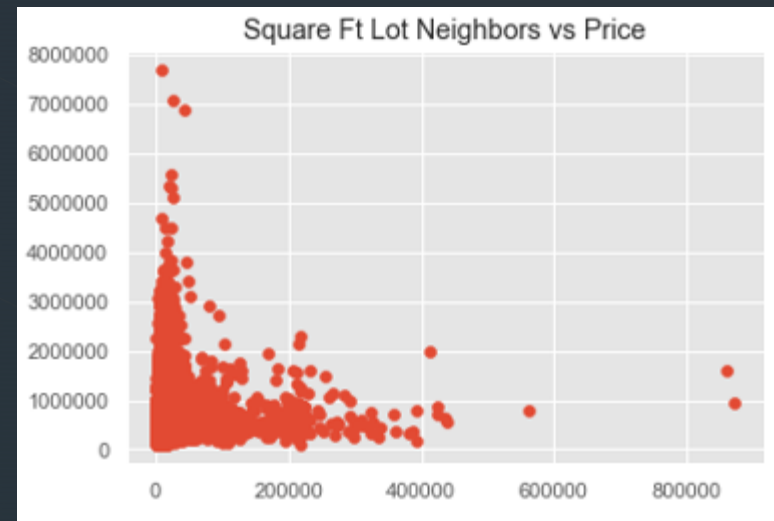
Data Visualization

- Exploring the relationship between condition and price, if any.
 - Note on condition: many more instances of 3 or 4 rated houses, but the average price does appear to be a bit higher for better (higher) conditions
- Price vs. year built- are newer homes more expensive, or is the opposite more true?
 - There does not appear to be a linear relationship between year built and price



Data Visualization

- Exploring the relationship between neighboring homes and price.
 - There seems to be a similar relationship between sqft_living vs price and sqft_living15 vs price, though perhaps not as strong, while sqft_lot15 (property size) does not have a strong linearity with price



Linear Regression

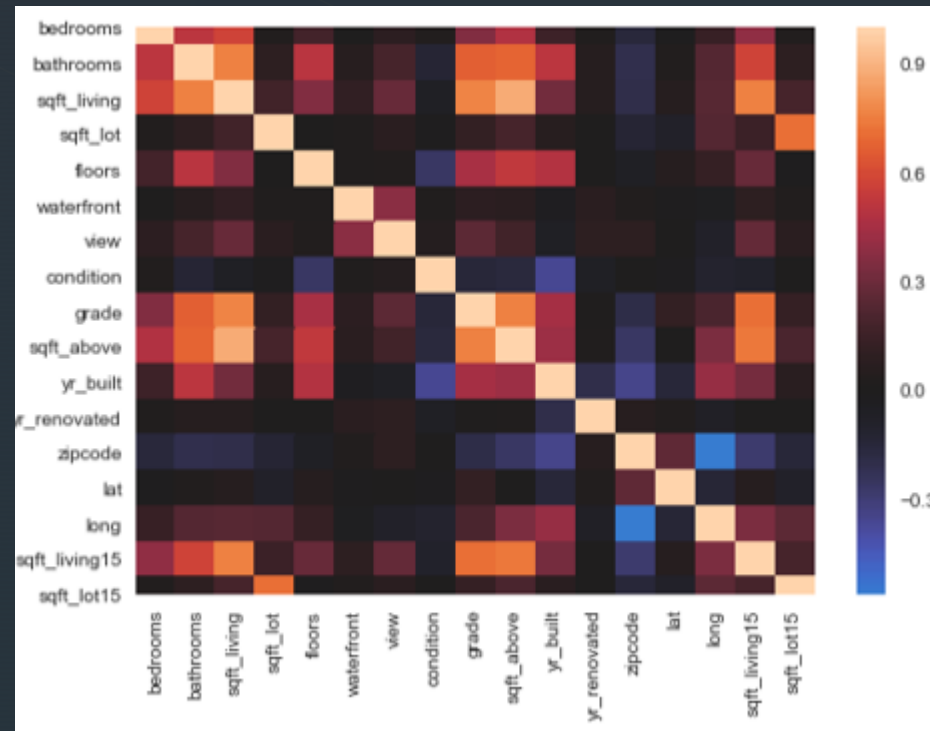
- Given the strong linear visual between sqft_living and price, these will be the columns used in the linear regression model to test the strength of this relationship.
 - R-squared of 0.493 does show a relationship exists between sqft_living and price but it is not especially strong

OLS Regression Results

Dep. Variable:	price	R-squared:	0.493			
Model:	OLS	Adj. R-squared:	0.493			
Method:	Least Squares	F-statistic:	2.097e+04			
Date:	Tue, 24 Mar 2020	Prob (F-statistic):	0.00			
Time:	14:30:35	Log-Likelihood:	-3.0006e+05			
No. Observations:	21597	AIC:	6.001e+05			
Df Residuals:	21595	BIC:	6.001e+05			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	-4.399e+04	4410.023	-9.975	0.000	-5.26e+04	-3.53e+04
sqft_living	280.8630	1.939	144.819	0.000	277.062	284.664
Omnibus:	14801.942	Durbin-Watson:	1.982			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	542662.604			
Skew:	2.820	Prob(JB):	0.00			
Kurtosis:	26.901	Cond. No.	5.63e+03			

Exploring Multicollinearity

- Looking at correlations between variables.
 - Some expected correlation between the different measures of square footage
 - Somewhat strong correlation between number of bathrooms and sqft_living, as well as grade and sqft_living



Linear Regression

- Run a regression predicting price based on sqft_living, waterfront and condition
 - R-squared value of 0.535 is higher than the single variable regression, indicating an improved model, but the value still does not suggest this is an especially strong model

OLS Regression Results

Dep. Variable:	price	R-squared:	0.535
Model:	OLS	Adj. R-squared:	0.535
Method:	Least Squares	F-statistic:	8279.
Date:	Sat, 29 Feb 2020	Prob (F-statistic):	0.00
Time:	16:14:02	Log-Likelihood:	-2.9912e+05
No. Observations:	21597	AIC:	5.983e+05
Df Residuals:	21593	BIC:	5.983e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	-1.786e+05	1.01e+04	-17.654	0.000	-1.98e+05	-1.59e+05
sqft_living	274.6070	1.871	146.785	0.000	270.940	278.274
waterfront	8.575e+05	2.09e+04	40.976	0.000	8.16e+05	8.99e+05
condition	4.16e+04	2626.112	15.841	0.000	3.65e+04	4.67e+04

Omnibus:	13445.792	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	444311.006
Skew:	2.471	Prob(JB):	0.00
Kurtosis:	24.664	Cond. No.	2.79e+04

Linear Regression

- Run a regression predicting price based on sqft_living, grade, bathrooms, waterfront, yr_renovated and view
 - R-squared value of 0.595 is higher than the single variable regression and first multi-variable regression, indicating an improved model, but the value still does not suggest this is an especially strong model

OLS Regression Results

Dep. Variable:	price	R-squared:	0.595			
Model:	OLS	Adj. R-squared:	0.595			
Method:	Least Squares	F-statistic:	5288.			
Date:	Tue, 24 Mar 2020	Prob (F-statistic):	0.00			
Time:	15:05:07	Log-Likelihood:	-2.9763e+05			
No. Observations:	21597	AIC:	5.953e+05			
Df Residuals:	21590	BIC:	5.953e+05			
Df Model:	6					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	-5.642e+05	1.25e+04	-45.053	0.000	-5.89e+05	-5.4e+05
sqft_living	179.2472	3.163	56.675	0.000	173.048	185.446
grade	1.009e+05	2152.495	46.895	0.000	9.67e+04	1.05e+05
bathrooms	-3.129e+04	3238.444	-9.662	0.000	-3.76e+04	-2.49e+04
waterfront	6.098e+05	2.1e+04	29.014	0.000	5.69e+05	6.51e+05
yr_renovated	70.8761	4.399	16.112	0.000	62.254	79.499
view	6.778e+04	2343.256	28.925	0.000	6.32e+04	7.24e+04
Omnibus:	15685.327	Durbin-Watson:	1.977			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	923152.843			
Skew:	2.914	Prob(JB):	0.00			
Kurtosis:	34.494	Cond. No.	3.01e+04			

Final Thoughts

- Takeaways: there does appear to be a linear relationship between sqft and price within the KC housing set, but the models created are not especially strong given R-squared values of 0.493, 0.535 and 0.595.
- The multiple variable regression slightly improved the r-squared value, but not by much. This could be due to slight correlations between the values or that there could be stronger factors in the dataset or elsewhere to predict price.
- Limitations and Next Steps: the r-squared values are not especially strong, suggesting these models aren't very strong. One next step could be to look into the locations of the houses. Do different zip codes have higher prices? Do these variables become stronger/weaker when looking into specific zip codes?