Exploring Boston Housing Data

Boston Housing Dataset

- Boston housing dataset after data cleaning
  - The Boston 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

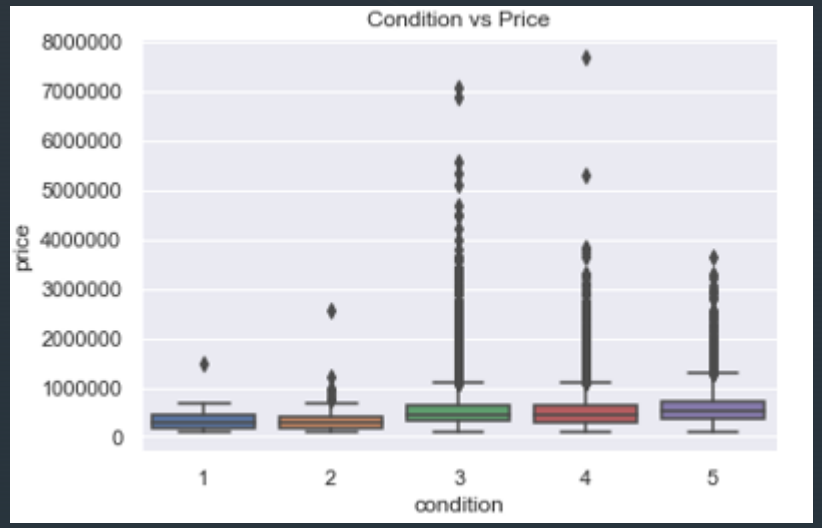
```
Index(['id', 'date', 'price', 'bedrooms', 'bathrooms', 'sqft_living',
       'sqft_lot', 'floors', 'waterfront', 'view', 'condition', 'grade',
       'sqft_above', 'sqft_basement', 'yr_built', 'yr_renovated', 'zipcode',
       'lat', 'long', 'sqft_living15', 'sqft_lot15'],
      dtype='object')
```

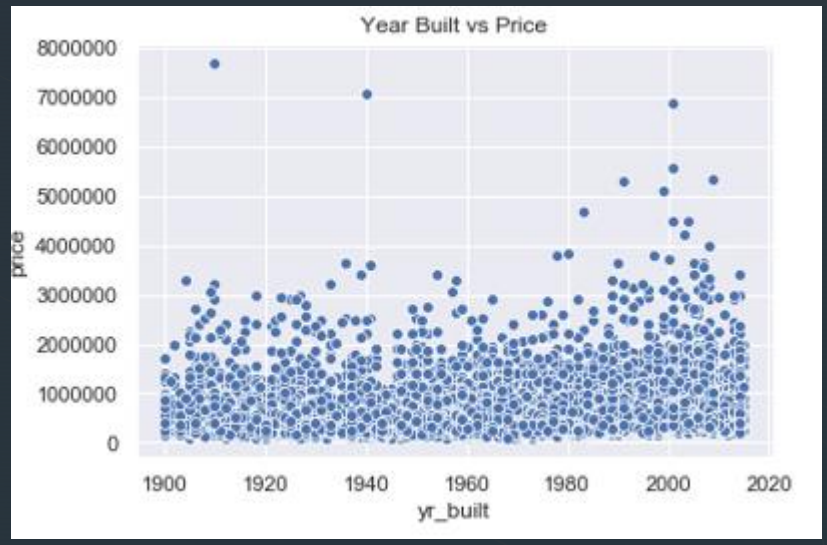
Data Visualization

- 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

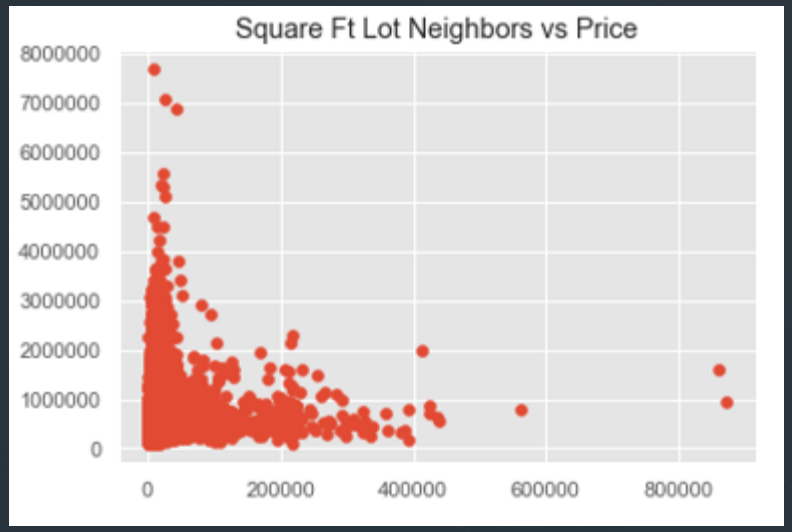
- Exploring the relationship between condition and price, if any.
  - Note on condition: many more instances of 3 or 4 rated houses, but the average price does appear to be a big higher for better (higher) conditions
- Price vs. year built- are newer homes more expensive, or is the opposite more true?
  - 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

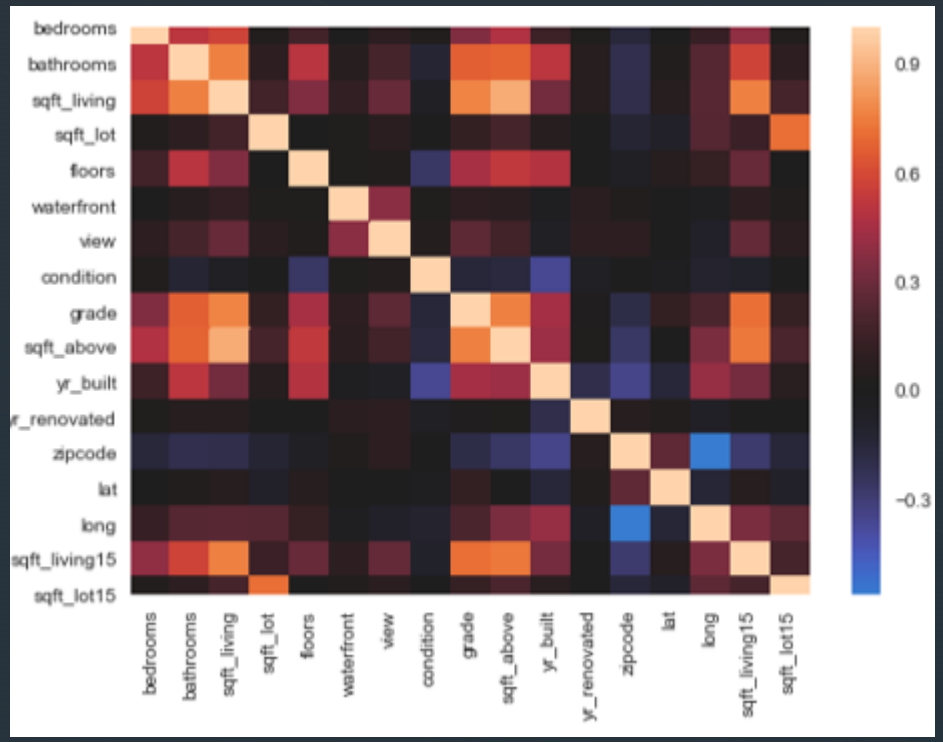
Dep. Variable:	sqft_living	**R-squared:**	0.493
Model:	OLS	**Adj. R-squared:**	0.493
Method:	Least Squares	**F-statistic:**	2.097e+04
Date:	Sat, 29 Feb 2020	**Prob (F-statistic):**	0.00
Time:	16:14:01	**Log-Likelihood:**	-1.7066e+05
No. Observations:	21597	**AIC:**	3.413e+05
Df Residuals:	21595	**BIC:**	3.413e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1132.5374	7.914	143.104	0.000	1117.025	1148.050
price	0.0018	1.21e-05	144.819	0.000	0.002	0.002

Omnibus:	2829.055	**Durbin-Watson:**	1.981
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	8794.373
Skew:	0.685	**Prob(JB):**	0.00
Kurtosis:	5.810	**Cond. No.**	1.16e+06

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

Final Thoughts

- Takeaways: there does appear to be a linear relationship between sqft and price within the Boston housing set, but the models created are not especially strong given R-squared values of 0.493 and 0.535.
- 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?