Insights into Housing in King County, Washington

Tina, Mina, Mo

introduction

End customer in mind

When exploring the data for insights, we have two customers in mind

- **Ben who owns a real estate development company**
- **Amy who is a real estate broker**

Study Outline:

Using residential property sales data from 2014 May to 2015 May, we aim to help Ben and Amy to determine

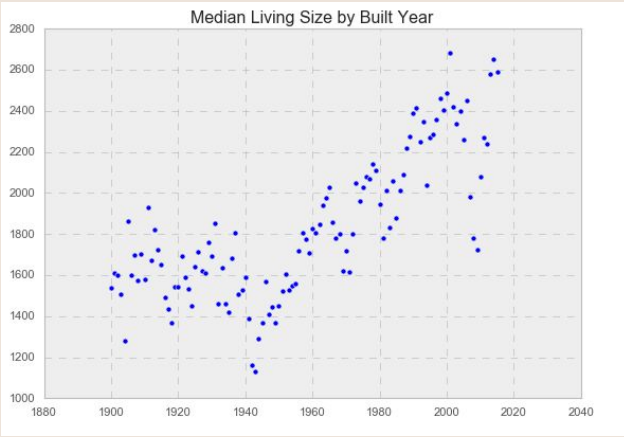
- What and where to build
- Where the affluent areas are

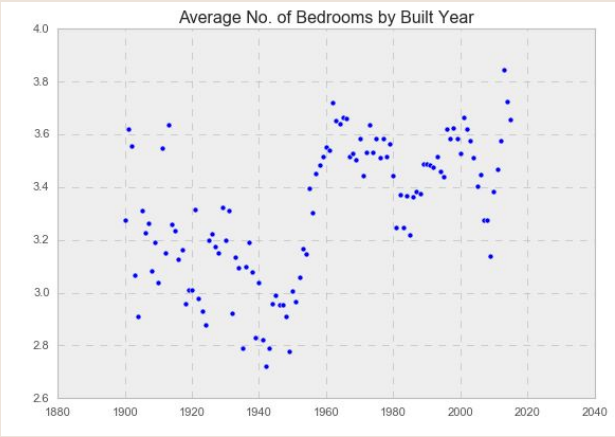
And ultimately try to predict the house price.

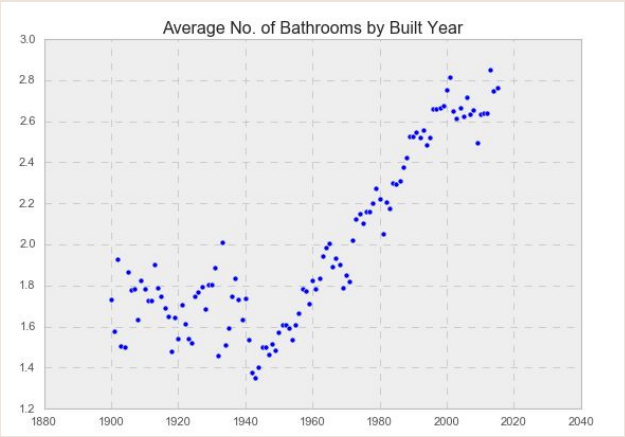
Are houses getting bigger over time?

Yes, although the living size of houses decreased during the great depression in the 30s, post-war they have been on an increasing trend until the recent financial crisis.

- Average number of bedrooms increased post-war as baby boomers arrived and families require more bedrooms
- Average number of bathrooms continue to rise, signalling more of a lifestyle change that consumer want more bedrooms

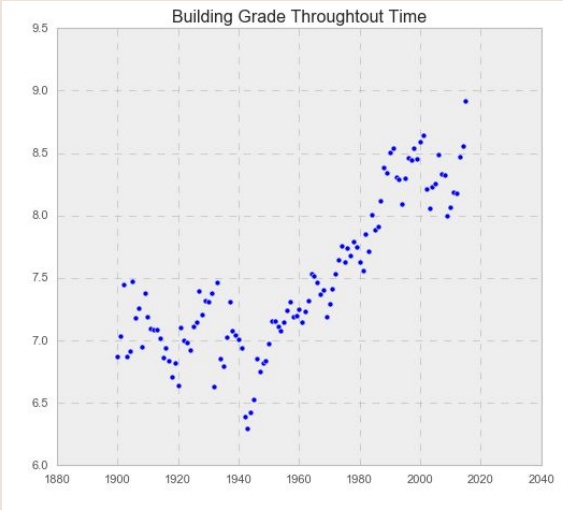






Are new houses more expensive than old houses?

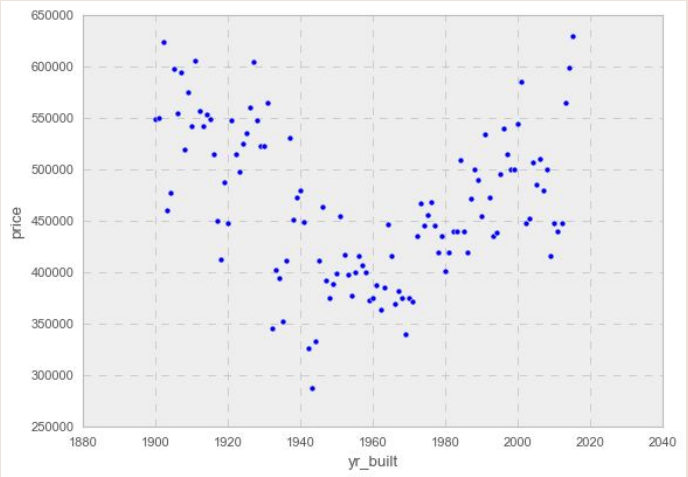
Not really, although there is a clear trend that newer houses are much better in terms of construction quality



Business Insight

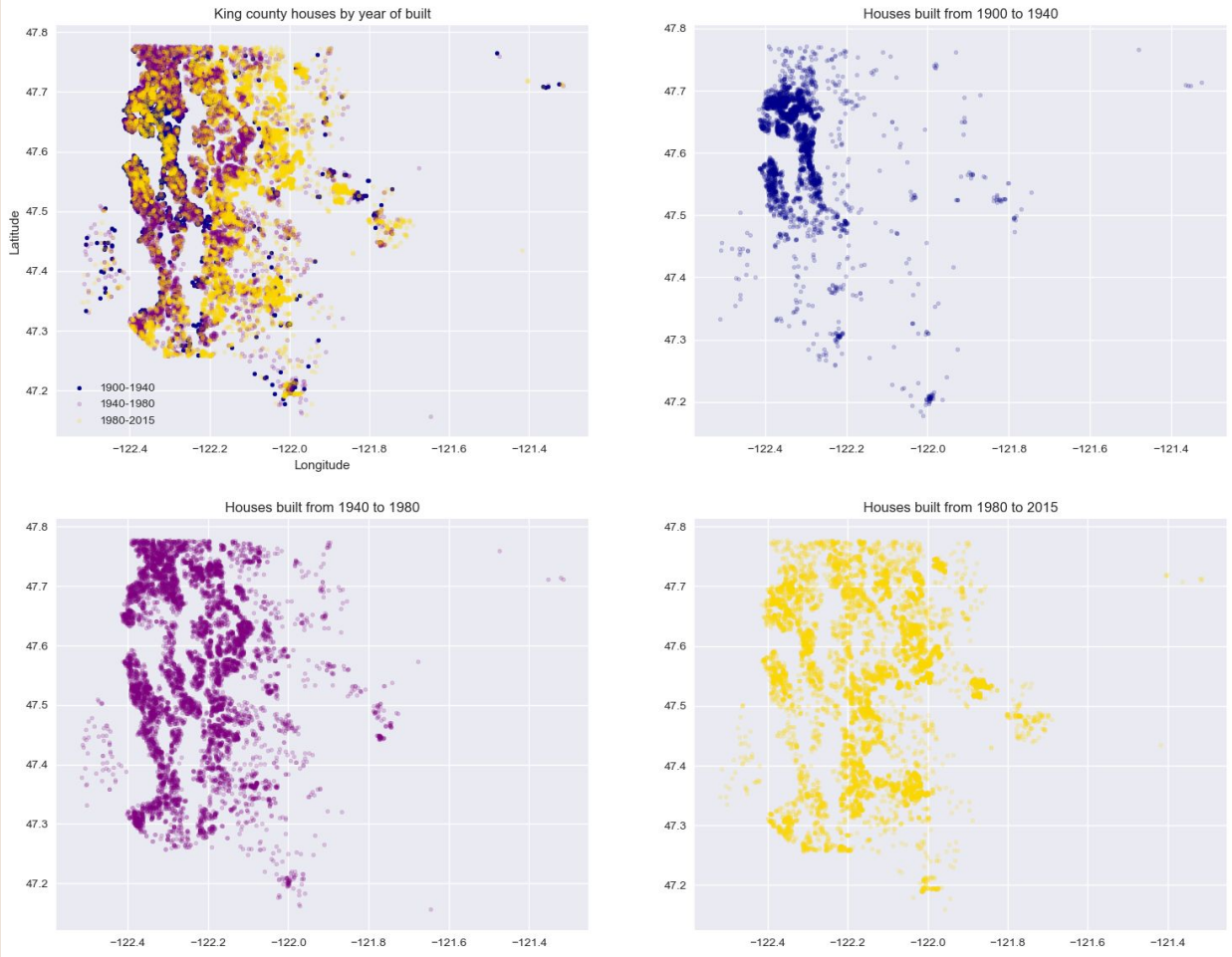
For house builder **Ben**, we'd advice that he

- include enough bathrooms
- could run marketing highlighting the benefit of better quality construction of new houses over period houses



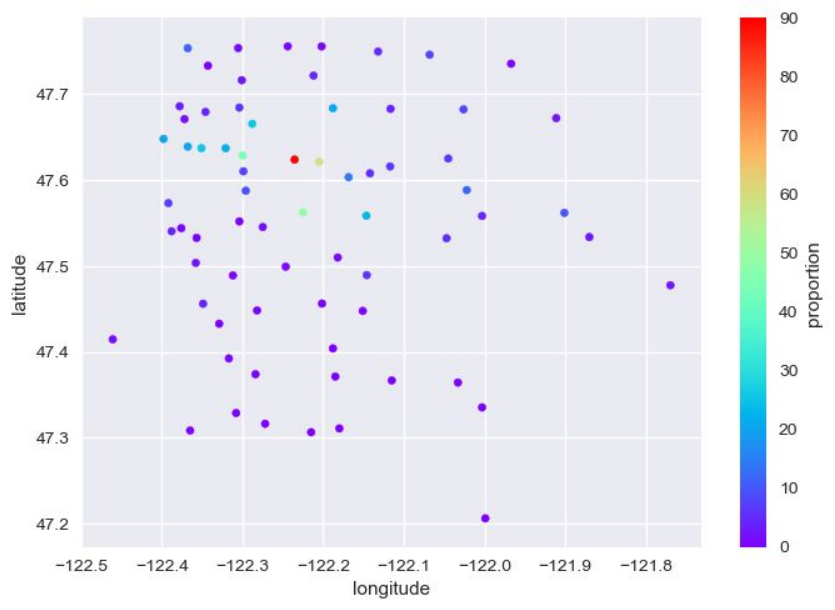
Where to find a nice historic property?

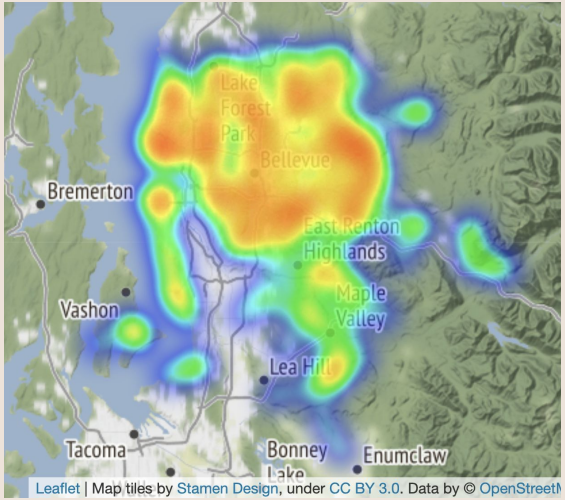
Most are in the city of Seattle



Where are the affluent areas?

In Medina, WA 98039, 90% of the houses are over $1million





Above: density of 25% most expensive houses

Business Insight

For real estate broker **Amy**, she could advice her clients that are looking for period houses to concentrate their efforts in the city or Seattle, and tell her wealthy clients to limit their search areas to only 4-5 zipcodes.

what drives house price

Our predictive model

Found that
- **Size** of living space in sq. ft.
- Being on the **waterfront**
- Being located **north** in the county

All drive up the house price.

Model R-squared: 65%

For example,

- A property of 1,000 sq. ft in size, located in the middle of the King County is predicted to be worth $280k
- If it were built to be 1,100 sq. ft in size, its valuation would increase by $22k
- If it were built on a waterfront, it could be worth up to $600k

Business Insight

House builder Ben could use this to estimate the sale price of a property he plans to build, in an aim to increase his **profitability.**

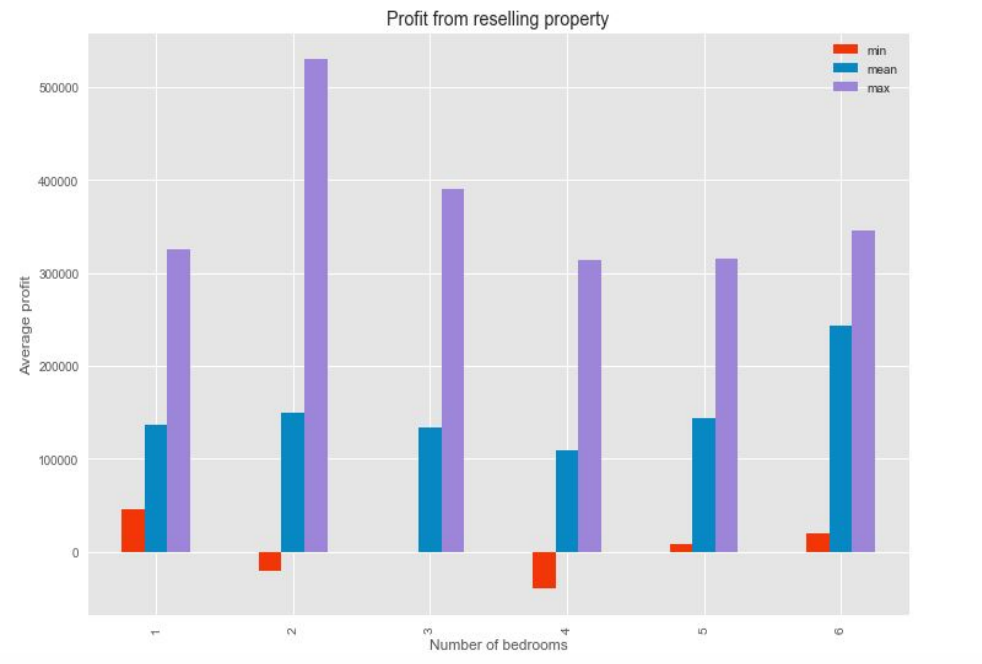
Q & A

Appendix

Model Details

How to make the most profit?

In the dataset, we noticed that there were multiple houses that had been bought and resold within the year. With this data, we wanted to explore the profit earned on a property by reselling



Out of the 21000+ house sales, we only had a total of 176 houses that were resold.

We can see that the average profit does not vary much. **However the maximum and minimums indicate the risk and reward involved in re-selling a property. The greatest profit earned was on a 2 bedroom property.**

Caution: this could be influenced by the location of the property. We did not have enough data to explore how the area affects the potential profit.

Data Cleaning

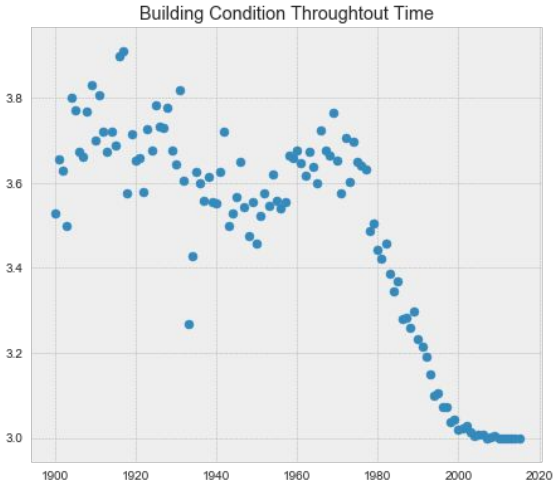
- **Dropped the data column ‘view’:** too many missing values which does not make intuitive sense
- **Dropped the data column ‘sqft_basement’:** feel this is a subset of sqft_living

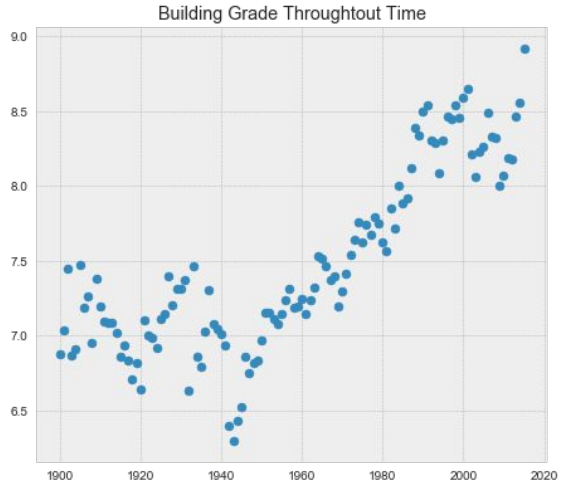
- Convert date from string object to datetime
- Created a new column ‘is_renovated’: 1 if it has a value in yr_renovation, 0 if its zero or null
- Nulls from waterfront were excluded from model

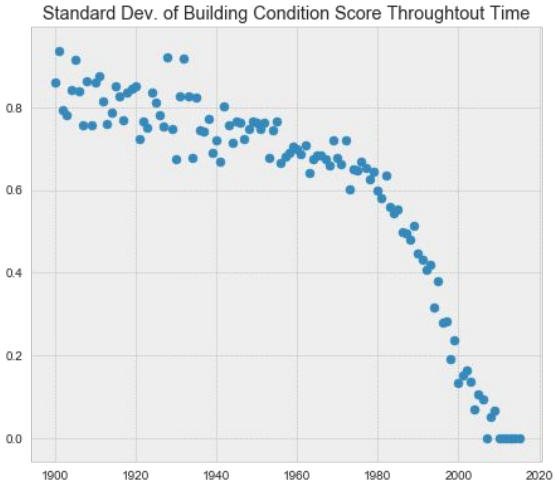
Data Transformation

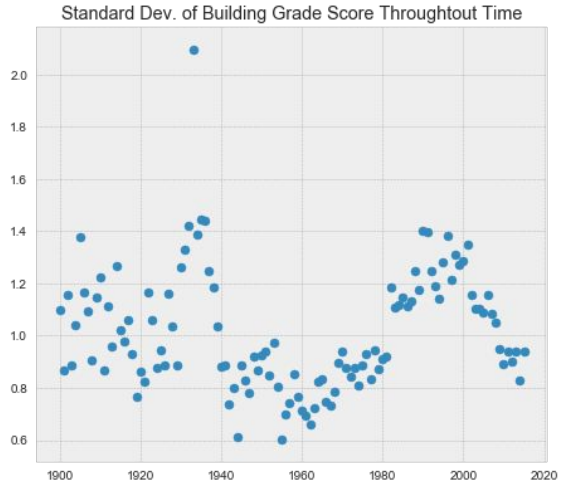
- Created a new column converting yr_built into 10 bins.

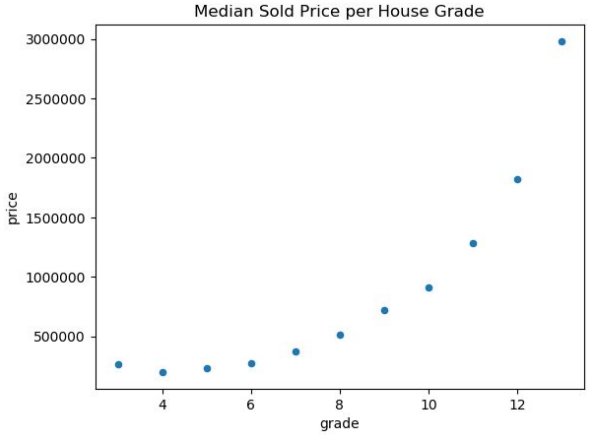
Study of building condition & grade over time

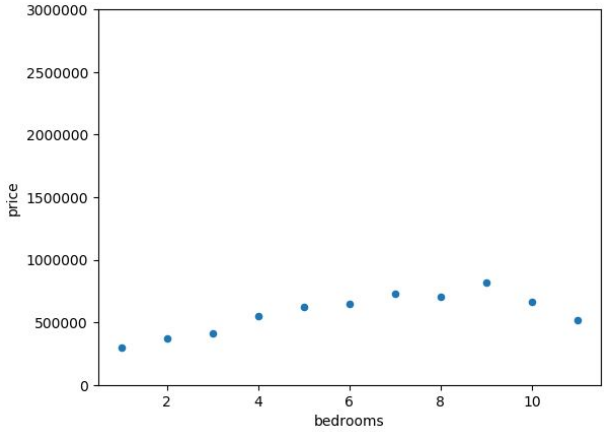


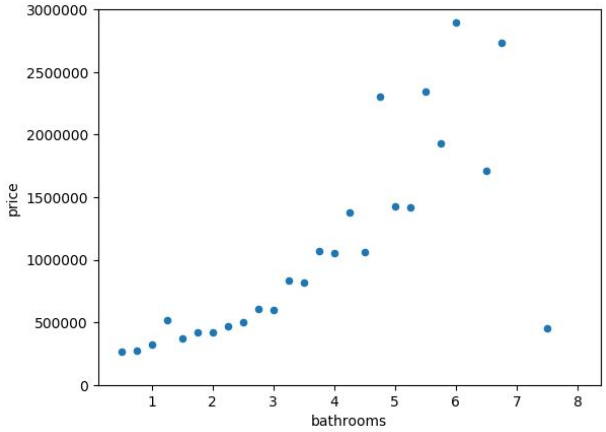


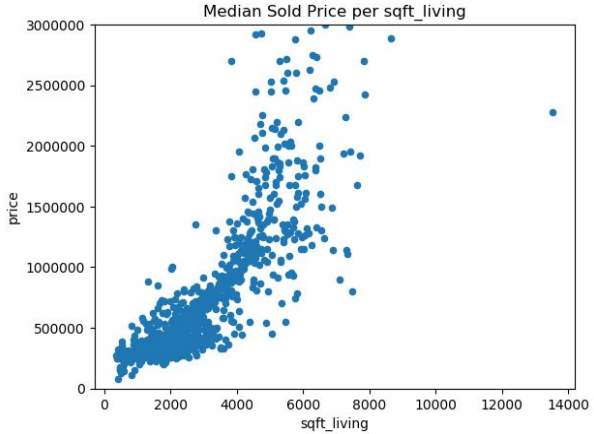


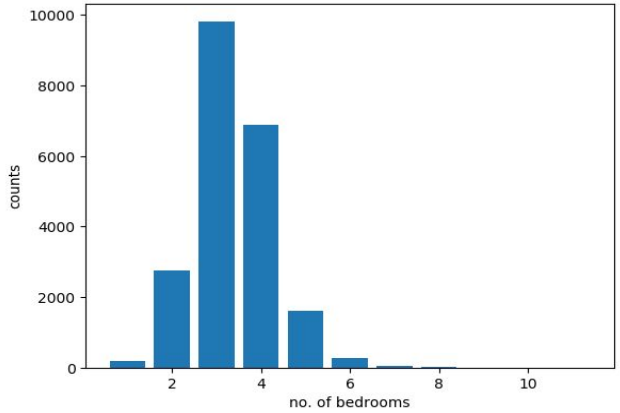


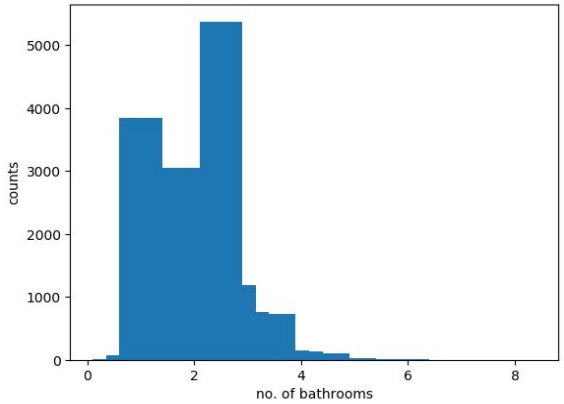








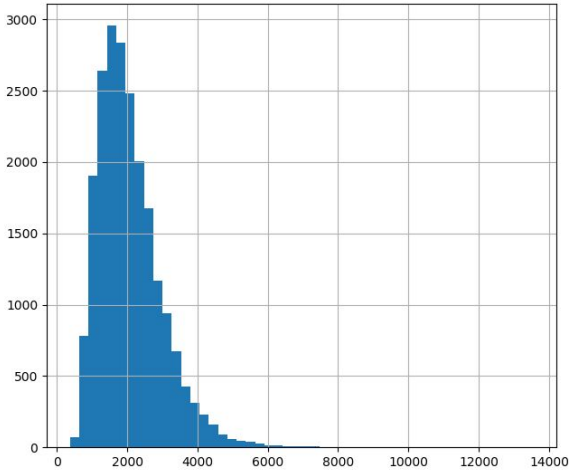


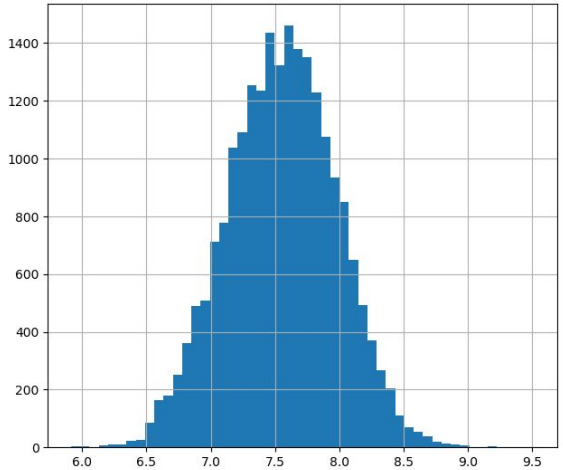


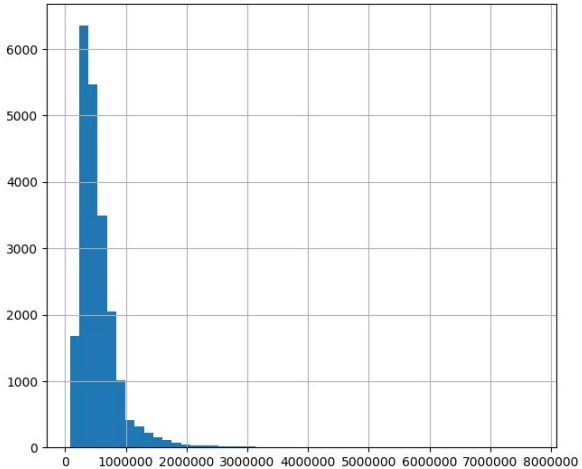
Initial look at potential predictor variables

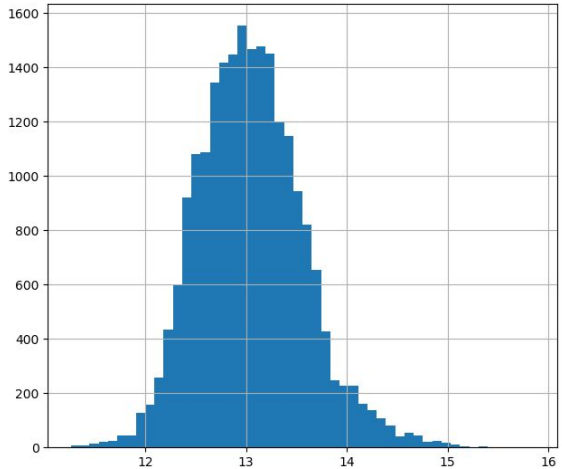
Log transformations *Of*

Sqft_living & Price



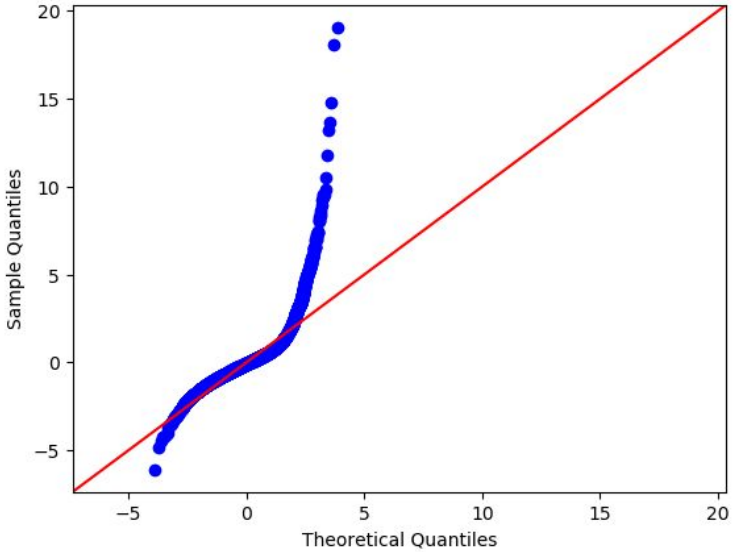




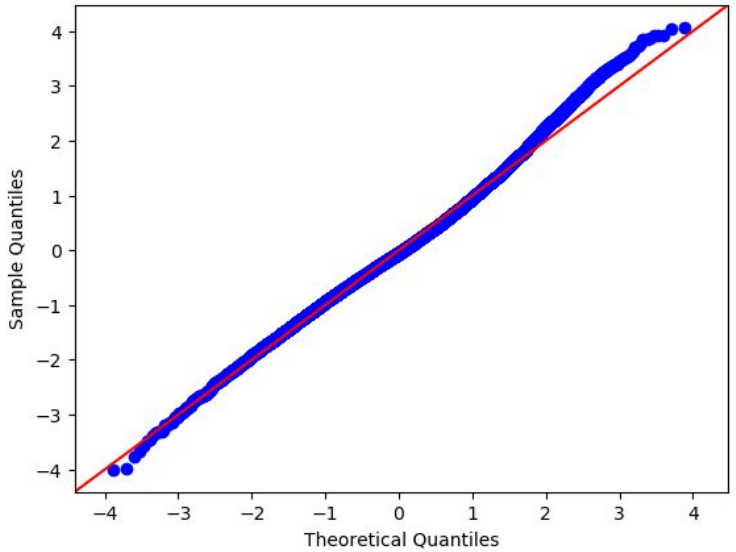


QQ plot to check normality assumption

Before log transformation



After log transformation



Predictive model *using* Linear Regression

ln(price) =
-70.1387
+ 0.8026 * ln(sqft_living)
+ 0.8021 * waterfront
+ 1.6216 * lat

OLS Regression Results

Dep. Variable:	price_log	**R-squared:**	0.653
Model:	OLS	**Adj. R-squared:**	0.653
Method:	Least Squares	**F-statistic:**	1.203e+04
Date:	Wed, 23 Oct 2019	**Prob (F-statistic):**	0.00
Time:	10:10:56	**Log-Likelihood:**	-4814.4
No. Observations:	19220	**AIC:**	9637.
Df Residuals:	19216	**BIC:**	9668.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-70.1387	0.769	-91.197	0.000	-71.646	-68.631
sqft_liv_log	0.8026	0.005	151.466	0.000	0.792	0.813
waterfront	0.8021	0.026	30.943	0.000	0.751	0.853
lat	1.6216	0.016	100.196	0.000	1.590	1.653

Omnibus:	372.358	**Durbin-Watson:**	2.012
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	464.164
Skew:	0.270	**Prob(JB):**	1.62e-101
Kurtosis:	3.536	**Cond. No.**	1.65e+04

Zipcodes with most above $1m houses

	med_prices	houses	houses > 1000000	proportion	longitude	latitude
98039	1.895e+06	50	45	90.00	-122.236	47.624
98004	1.150e+06	317	184	58.04	-122.205	47.621
98040	9.938e+05	282	136	48.23	-122.225	47.562
98112	9.150e+05	269	120	44.61	-122.300	47.629
98105	6.750e+05	229	59	25.76	-122.288	47.666
98109	7.360e+05	109	28	25.69	-122.351	47.637
98006	7.602e+05	498	117	23.49	-122.147	47.559
98102	7.100e+05	104	23	22.12	-122.321	47.637
98033	6.784e+05	432	90	20.83	-122.188	47.684
98119	7.450e+05	184	38	20.65	-122.368	47.639