Housing Prices in King's County

Methodology: Step 1
Acquiring and Cleaning Data

- Before the data can be properly analyzed, we must make sure it doesn't have missing values.
  - We notice that there are nulls in waterfront and yr_renovated
  - It can be expected that renovating a house or living near a waterfront will increase the price of the house
    - Therefore the people to leave that information out likely have not renovated or live near the waterfront

```
df.isna().sum()
```

id	0
date	0
price	0
bedrooms	0
bathrooms	0
sqft_living	0
sqft_lot	0
floors	0
waterfront	2339
view	61
condition	0
grade	0
sqft_above	0
sqft_basement	0
yr_built	0
yr_renovated	3754
zipcode	0
lat	0
long	0
sqft_living15	0
sqft_lot15	0

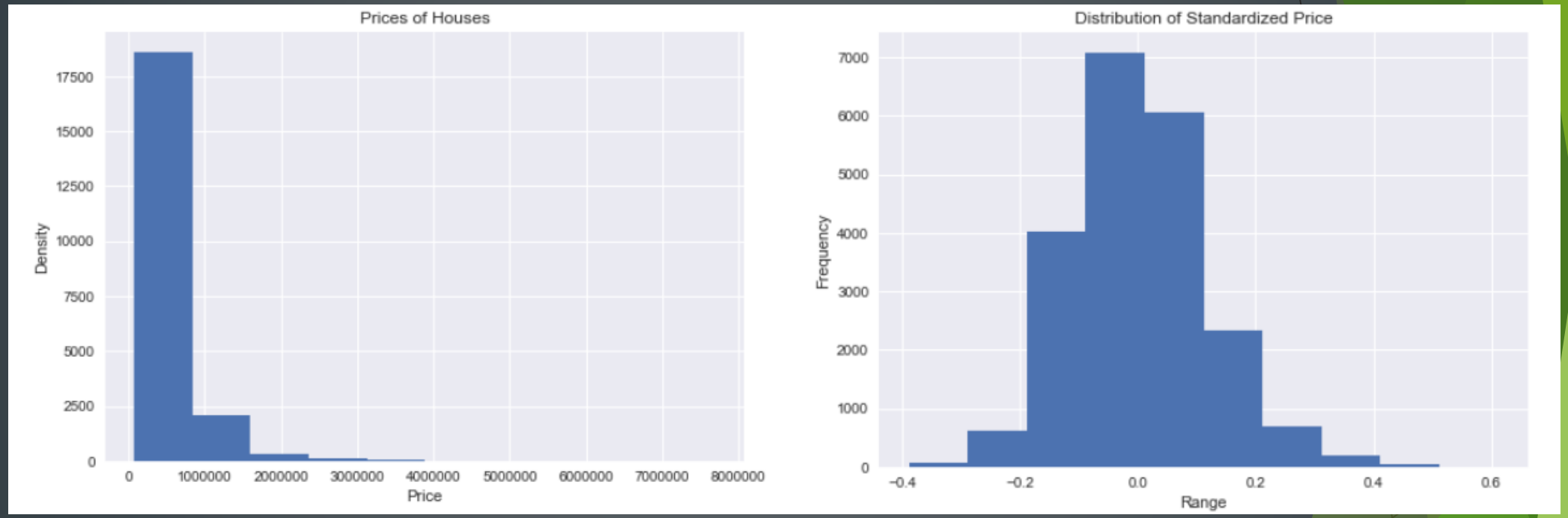
dtype: int64


```
df.isna().sum()
```

id	0
date	0
price	0
bedrooms	0
bathrooms	0
sqft_living	0
sqft_lot	0
floors	0
waterfront	0
view	0
condition	0
grade	0
sqft_above	0
sqft_basement	0
yr_built	0
yr_renovated	0
zipcode	0
lat	0
long	0
sqft_living15	0
sqft_lot15	0

dtype: int64

Methodology: Step 2
Data Exploration

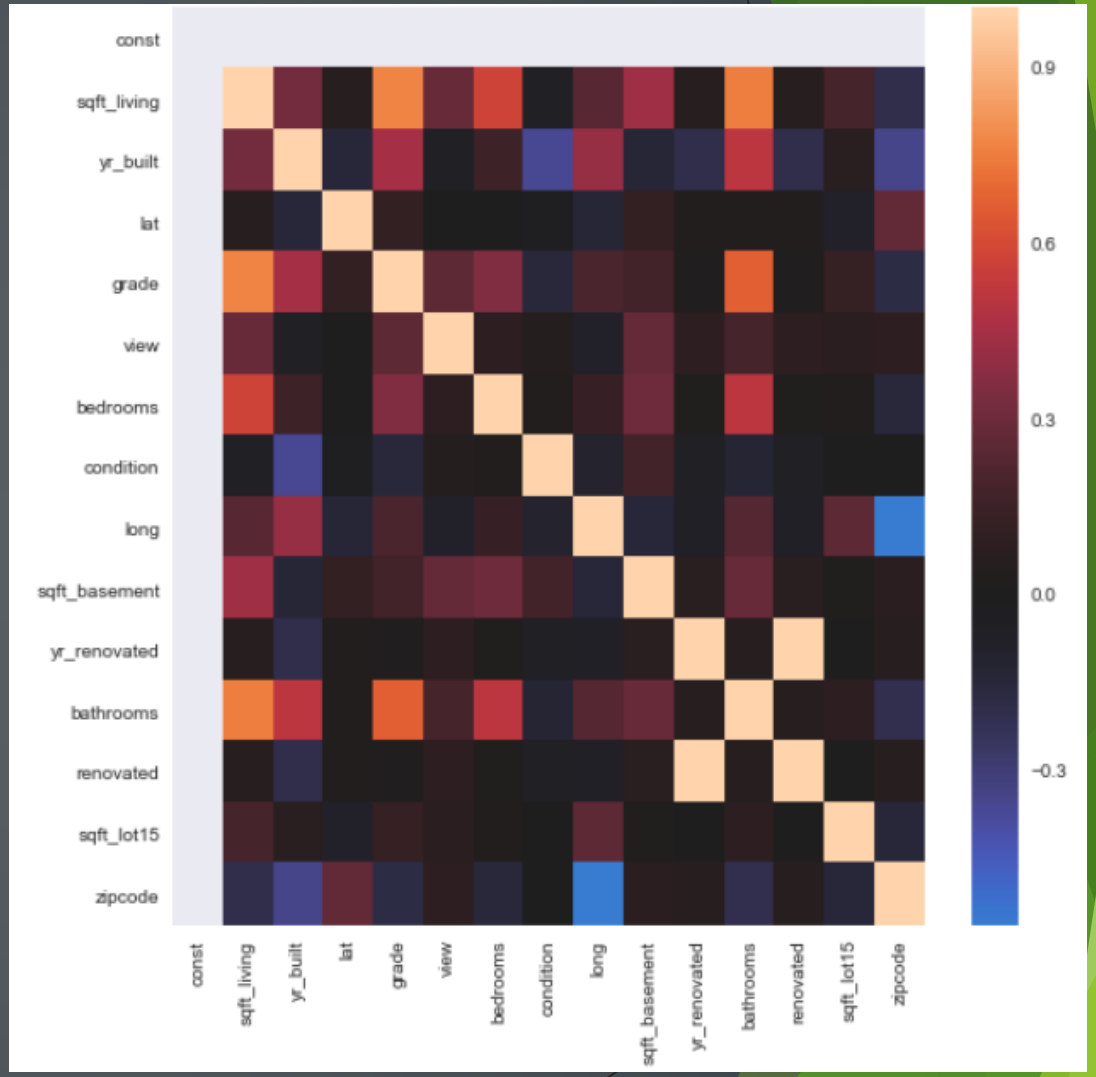


Here we can observe that the data is not normally distributed. Which is bad for modeling and regressions.

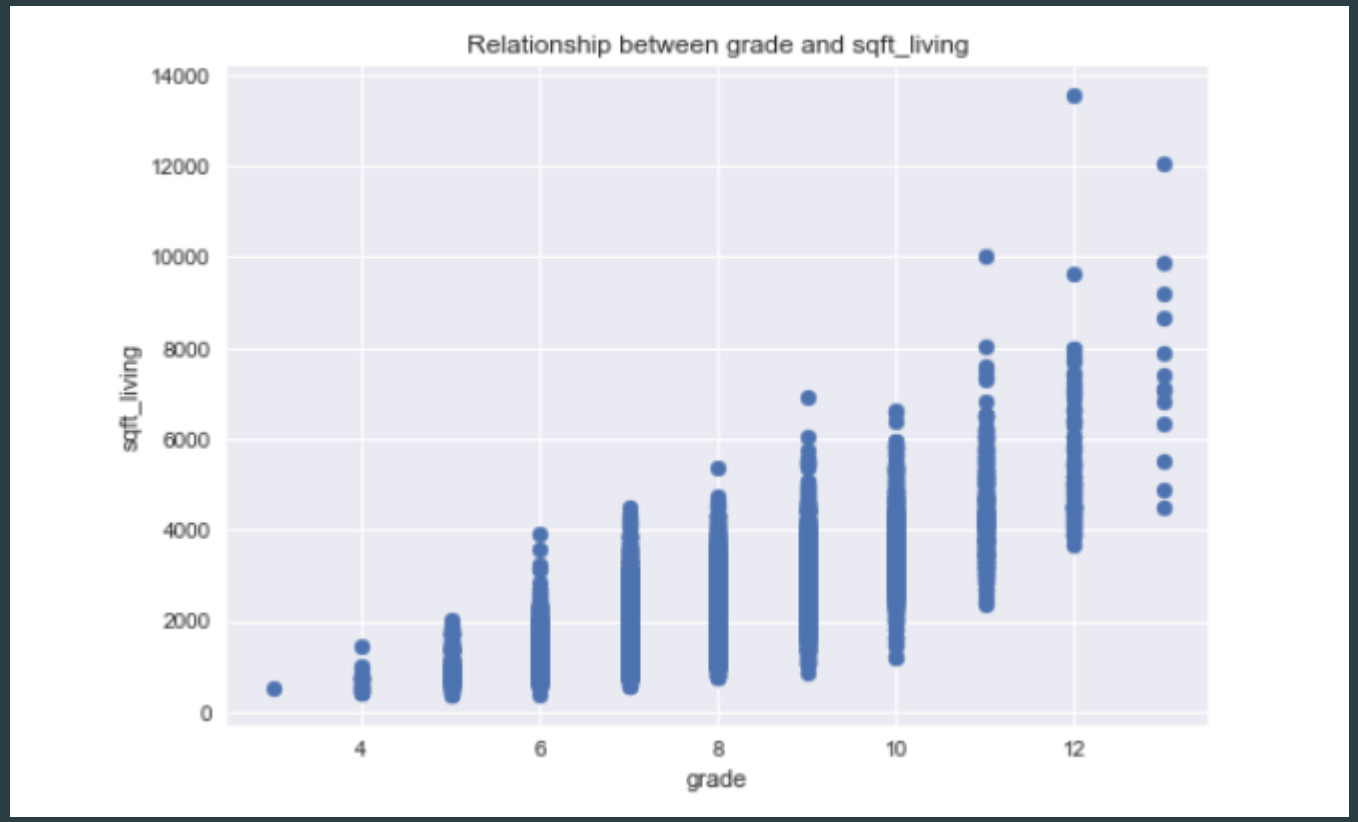
After Standardizing the data we get a much more preferable distribution.

Methodology 3: Selecting Variables

- We start off with all our variables and do stepwise selection.
  - This is a method that goes through all the variables and drops insignificant variables and tries to re-add them later on if they are significant.
- Next we deal with 2 independent variables correlating with each other.
  - This is a heatmap of correlation between all independent variables
  - One potential worry is grade and sqft_living



Methodology: Part 3
Selecting Variables Part 2



Our visualization shows us that as grade increases, sqft_living also goes up. This is bad because the impact of each individual effect gets hidden by the other one. The best option is generally to drop one of the variables and, generally drop the variable that drops the R2 by the least.

Running a Regression

- Once we have all of the variables we want to use, we can create a proper regression that can grant us insights.
  - Key thinks to pay attention to:
  - Rsquared
    - Tells you how much of the total variation is measured by the model
  - Coefficient
    - Tells you how each individual variable, on average, effects price.

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.634
Model:	OLS	**Adj. R-squared:**	0.634
Method:	Least Squares	**F-statistic:**	3044.
Date:	Mon, 29 Oct 2018	**Prob (F-statistic):**	0.00
Time:	11:35:34	**Log-Likelihood:**	-2.8942e+05
No. Observations:	21082	**AIC:**	5.789e+05
Df Residuals:	21069	**BIC:**	5.790e+05
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.54e+07	3.13e+06	8.125	0.000	1.93e+07	3.15e+07
waterfront	6.446e+05	2.03e+04	31.741	0.000	6.05e+05	6.84e+05
yr_built	-2251.5998	71.464	-31.507	0.000	-2391.675	-2111.524
lat	5.839e+05	1.19e+04	49.067	0.000	5.61e+05	6.07e+05
grade	1.973e+05	1678.713	117.558	0.000	1.94e+05	2.01e+05
view	6.81e+04	2332.350	29.200	0.000	6.35e+04	7.27e+04
bedrooms	2.477e+04	1862.101	13.300	0.000	2.11e+04	2.84e+04
condition	2.307e+04	2605.691	8.854	0.000	1.8e+04	2.82e+04
long	-6.136e+04	1.42e+04	-4.334	0.000	-8.91e+04	-3.36e+04
sqft_basement	76.8474	3.913	19.640	0.000	69.178	84.517
yr_renovated	40.9250	4.389	9.325	0.000	32.323	49.527
sqft_lot15	0.1773	0.059	3.009	0.003	0.062	0.293
zipcode	-585.4052	36.431	-16.069	0.000	-656.812	-513.999

2 Types of Regressions

- Unstandardized regressions are helpful for human interpretability.
  - Advantage:
    - You can look directly at a coefficient and tell what impact it will have on the dependent variable
  - Disadvantage:
    - You sacrifice accuracy by not standardizing data.
- Standardized Regressions will:
  - Give more accurate results, however the interpretations cannot be made by humans as effectively.

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.708
Model:	OLS	**Adj. R-squared:**	0.708
Method:	Least Squares	**F-statistic:**	4265.
Date:	Mon, 29 Oct 2018	**Prob (F-statistic):**	0.00
Time:	11:37:35	**Log-Likelihood:**	28742.
No. Observations:	21082	**AIC:**	-5.746e+04
Df Residuals:	21069	**BIC:**	-5.735e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0005	0.000	-1.070	0.285	-0.001	0.000
lat	0.2165	0.002	106.531	0.000	0.213	0.221
view	0.0882	0.003	34.259	0.000	0.083	0.093
sqft_living	1.0403	0.009	116.765	0.000	1.023	1.058
yr_built	-0.0374	0.002	-16.036	0.000	-0.042	-0.033
condition	0.0819	0.004	19.685	0.000	0.074	0.090
waterfront	0.0685	0.006	12.087	0.000	0.057	0.080
floors	0.0590	0.003	22.100	0.000	0.054	0.064
zipcode	-0.0399	0.002	-19.788	0.000	-0.044	-0.036
sqft_lot15	0.0376	0.014	2.609	0.009	0.009	0.066
yr_renovated	0.0250	0.002	10.089	0.000	0.020	0.030
long	-0.0631	0.005	-13.066	0.000	-0.073	-0.054
sqft_basement	-0.0641	0.006	-10.241	0.000	-0.076	-0.052

Recommendations:

- 1: Renovate your house!
  - Most of the variables in this dataset like the Zipcode, the Latitude, or the whether it is in front of water are immutable traits that you as a homeowner cannot change. However, renovating a house in the year 2000 has an estimated $8000 impact on the price of the home!
- 2: Things you may consider when renovating:
  - Floors, the size of your living room in square feet tend to increase the price of your home as well. It may be worth to make these changes as they will increase the resale value of your homes.
- 3: Keep your home in the best possible condition:
  - The only other variable left that you are able to have an impact on is the amount of care you put into your home. Condition also correlates positively with price and each additional grade they have on condition increases the price of the home on average by $23,070.