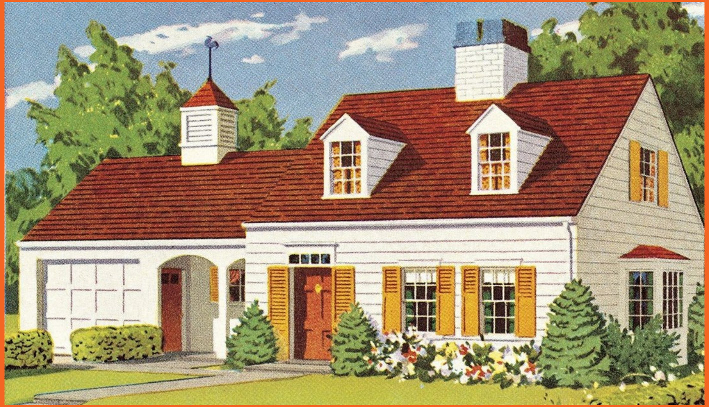
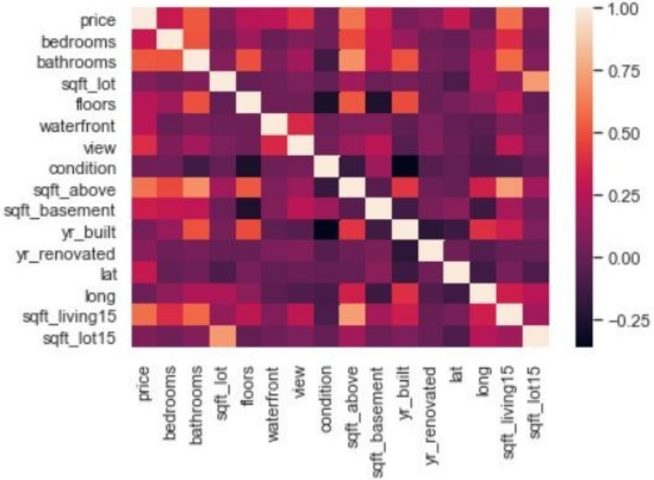
King’s County Housing Dataset



BY Sarah M

Cleaning the data

- Making sure each column was encoded as the correct data type
- Finding and eliminating (or replacing) Null Values
- Addressing multicollinearity (right figure)
- Using mean-normalization to standardize our data
- One-hot encoding our data



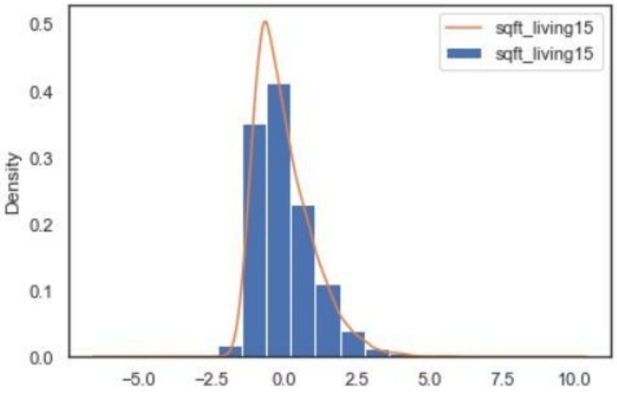
Exploratory Data Analysis

➔ **Posing meaningful questions -** A few questions that we would like answered:

- What is more important to a home buyer, the size of the lot or the size of the house (sqft_above)?
- Do people usually get a bargain on a house because it's old (looking at the variable yr_built)?
- Are there any negative relationships in our data?

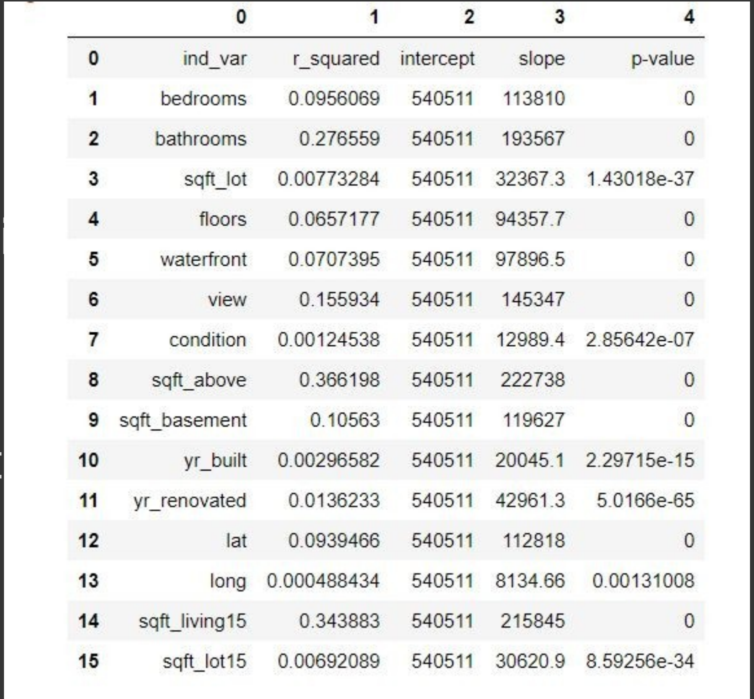
➔ Checking for normality with KDE plots and assumption of linearity with scatter plots

➔ The power of joint plots!



Modeling the Data

- Ordinary Least Squares
- Experimenting with log transformati
- Dealing with categorical data
- Which predictors make the final cut



	0	1	2	3	4
0	ind_var	r_squared	intercept	slope	p-value
1	bedrooms	0.0956069	540511	113810	0
2	bathrooms	0.276559	540511	193567	0
3	sqft_lot	0.00773284	540511	32367.3	1.43018e-37
4	floors	0.0657177	540511	94357.7	0
5	waterfront	0.0707395	540511	97896.5	0
6	view	0.155934	540511	145347	0
7	condition	0.00124538	540511	12989.4	2.85642e-07
8	sqft_above	0.366198	540511	222738	0
9	sqft_basement	0.10563	540511	119627	0
10	yr_built	0.00296582	540511	20045.1	2.29715e-15
11	yr_renovated	0.0136233	540511	42961.3	5.0166e-65
12	lat	0.0939466	540511	112818	0
13	long	0.000488434	540511	8134.66	0.00131008
14	sqft_living15	0.343883	540511	215845	0
15	sqft_lot15	0.00692089	540511	30620.9	8.59256e-34

Holdout Validation

- How well can we predict new data?

- Feature ranking on the data - (5,66,10)
  - Extracting the best features, muting the noisy ones

- Test-Train-Split
  - Training the model on 20% of the data, comparing the Mean of Squared Errors

- k-fold Cross Validation
  - Sample divided into k sub-samples
  - One retained for testing, the rest for training

Interpretations

- The final model included predictors bathrooms, sqft_above, lat, sqft_living15 and zipcode
- The had r-squared values of 30.3%, 36.1%, 20.2%, 38.5%, and 53.1%, respectively
- They were all statistically significant with p-values < .05
- Bathrooms had a co-efficient of 20%
- Sqft_above had a co-efficient of 30%
- Sqft_living15 had a co-efficient of 32%
  - This means for a unit increase in any one of these variables, there was an increase in the price that a house sold for by about 30 units.