FlatIron Module 1 Final Project – KMC House Prices

Evan Okin
July 25, 2019

Agenda

1) Dataset And Business Objective

2) Visualizing The Data

3) Cleaning The Data

4) Processing For Multiple Regression

5) Interpreting Model Results

6) Recommendations

7) Further Information

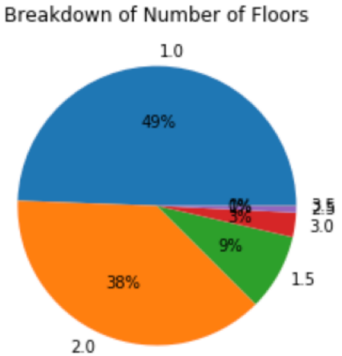
1. The Dataset and our Business Objective

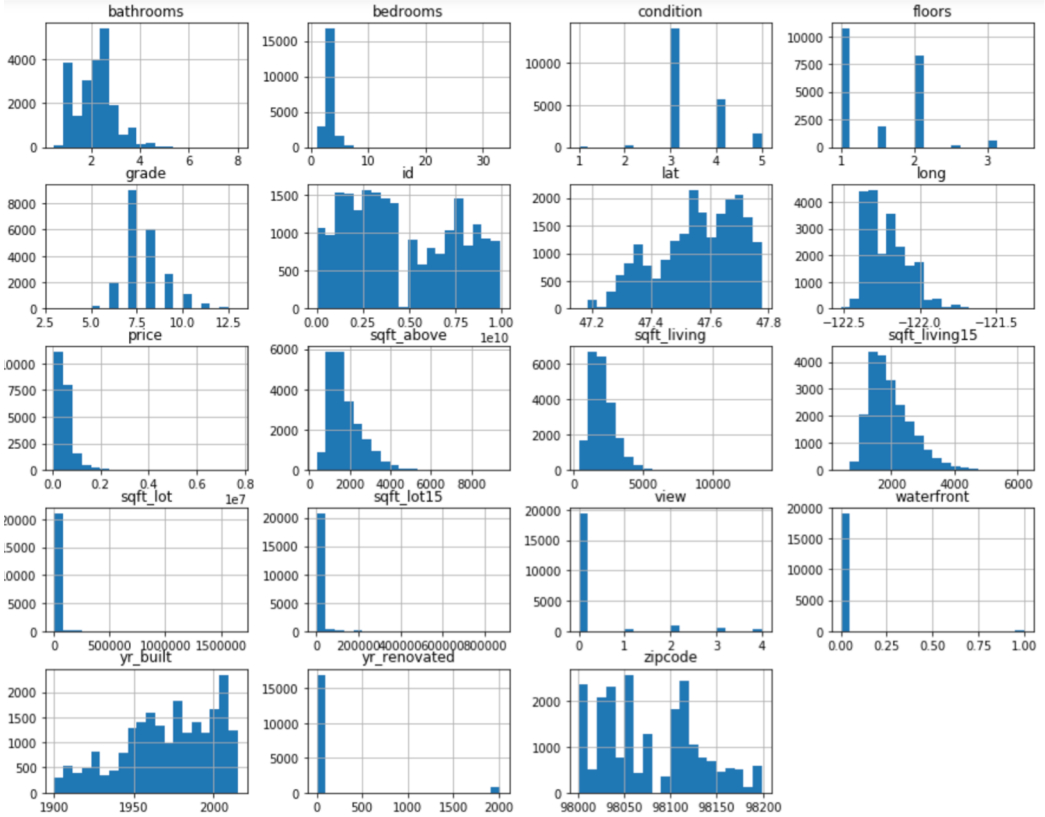
- The KC House Dataset includes over 21,000 homes with variables such as square footage, number of bedrooms, and whether the home has a waterfront view.
- The objective was to come up with a model to predict home prices based on the variables in the dataset.

Unique ID	Date sold	House price	Number of bedrooms	Number of bathrooms	Square footage (home)	Square footage (lot)
Number of floors	Waterfront view	Viewed or not	Overall condition	Overall grade by KC	Square footage (upper)	Square footage (basement)
Year built	Year renovated	Zip code	Latitude coordinates	Longitude coordinates	Square footage of living space (neighbors)	Square footage of lot (neighbors)

2. Visualizing The Data

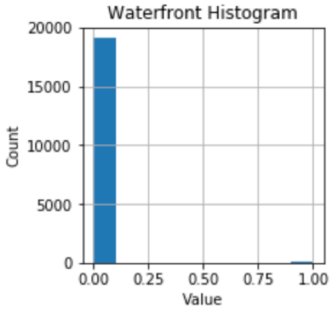
- The data includes both categorical variables (such as the number of bedrooms) and continuous variables (such as the square footage).

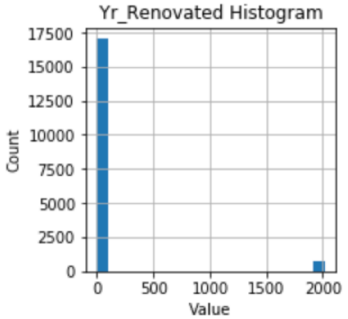




3. Cleaning The Data

- Several variables had to be analyzed due to missing values.
    - 12% of Waterfront View were missing
        - Filled with zeros – overwhelming majority
    - <1% of View were missing
        - Filled with zeros – overwhelming majority
    - 22% of Year Renovated were missing
        - Filled with zeros – majority not renovated





- Filled in ”?’s” for basement square footage (2% of values) by taking the difference between the total square footage and upstairs square footage.
- Rounded bathrooms to nearest 0.5.
- Created a “recently renovated” indicator and “post-war” building indicator.

4. Processing for Multiple Regression

- Some variables are categorical in nature (they only take on discrete values) and needed to be adjusted in the model.
- Some variables exhibit high correlation (multicollinearity) and needed to be dropped.
- Non-categorical variables needed to be scaled so that movements in results are standardized.
- Changed variable naming to comply with python regression.

Categorical Variables To Scale:	Number of bedrooms	Number of bathrooms	Number of floors	Waterfront view	Viewed or not	Overall condition	Overall grade by KC
Drop Due To Correlation:	Square footage of living space (neighbors)	Square footage of lot (neighbors)					
Scale:	Square footage (home	Square footage (lot)	Square footage (basement				

5. Interpreting Model Results

- Good amount of variability explained by the model (R-squared = 0.59).

OLS Regression Results

Dep. Variable:	price	R-squared:	0.586
Model:	OLS	Adj. R-squared:	0.585
Method:	Least Squares	F-statistic:	2179.
Date:	Mon, 22 Jul 2019	Prob (F-statistic):	0.00
Time:	13:08:46	Log-Likelihood:	-2.9788e+05
No. Observations:	21597	AIC:	5.958e+05
Df Residuals:	21582	BIC:	5.959e+05
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.742e+05	8795.562	65.286	0.000	5.57e+05	5.91e+05
sqftabove	2.33e+05	2104.673	110.725	0.000	2.29e+05	2.37e+05
sqftbasement	238.4921	3.956	60.285	0.000	230.738	246.246
postwarindicator	-1.665e+05	5192.975	-32.065	0.000	-1.77e+05	-1.56e+05
renovatedindicator	1.45e+05	1.24e+04	11.736	0.000	1.21e+05	1.69e+05
floors1p5	-3.103e+04	6592.977	-4.707	0.000	-4.4e+04	-1.81e+04
floors2p0	-3.355e+04	4506.039	-7.446	0.000	-4.24e+04	-2.47e+04
floors2p5	1.414e+05	1.93e+04	7.346	0.000	1.04e+05	1.79e+05
floors3p0	1.23e+05	9994.186	12.310	0.000	1.03e+05	1.43e+05
floors3p5	2.637e+05	8.95e+04	2.948	0.003	8.83e+04	4.39e+05
water0p0	2.607e+04	8220.313	3.171	0.002	9955.403	4.22e+04
water1p0	5.482e+05	1.6e+04	34.179	0.000	5.17e+05	5.8e+05
view1p0	1.649e+05	1.32e+04	12.471	0.000	1.39e+05	1.91e+05
view2p0	1.086e+05	7976.209	13.610	0.000	9.29e+04	1.24e+05
view3p0	1.83e+05	1.09e+04	16.798	0.000	1.62e+05	2.04e+05
view4p0	3.878e+05	1.66e+04	23.417	0.000	3.55e+05	4.2e+05

5. Interpreting Model Results

- The following was the final model to predict house prices:
- Variable interpretation:
  - Recently Renovated: +$145K if renovated since 2000
  - Number of Floors: +$123K if home has three floors
  - Square Footage: +$233K if home has more square footage (1 s.d.) than other homes

```
House Price = 574,226.9 +
              233,039.0 * sqftabove +
              1,149,532.0 * sqftbasement +
              -166,513.6 * postwarindicator +
              145,024.6 * renovatedindicator +
              -31,029.99 * floors1.5 +
              -33,551.63 * floors2.0 +
              141,425.3 * floors2.5 +
              123,033.1 * floors3.0 +
              263,706.1 * floors3.5 +
              26,067.82 * waterfront0.0 +
              548,159.1 * waterfront1.0 +
              164,943.2 * view1.0 +
              108,552.5 * view2.0 +
              183,034.1 * view3.0 +
              387,793.5 * view4.0
```

6. Recommendations

- In order to maximize profit when selling your home, implement the following:
    - Renovate the home if it was not renovated recently (2000 or later).
    - Renovate the home to have at least 3 floors, and a basement area if possible.
    - If possible, a waterfront view goes a long way for home value.
    - Home price goes up a lot when it is viewed by many people (this might be a function of increased demand), so show the home to as many prospective buyers as possible.

7. Further Information

- Python code can be found on my Github page, in the following repository: https://github.com/eokin324/FlatIron-Module-1-Final-Project
- This presentation can be found on my Github page, in the following repository: https://github.com/eokin324/FlatIron-Module-1-Final-Project
- Contact information:
    - Personal (including FlatIron) email address: eokin324@gmail.com
    - MBA email address: eo919@stern.nyu.edu

Questions?