Math 158 Final Project

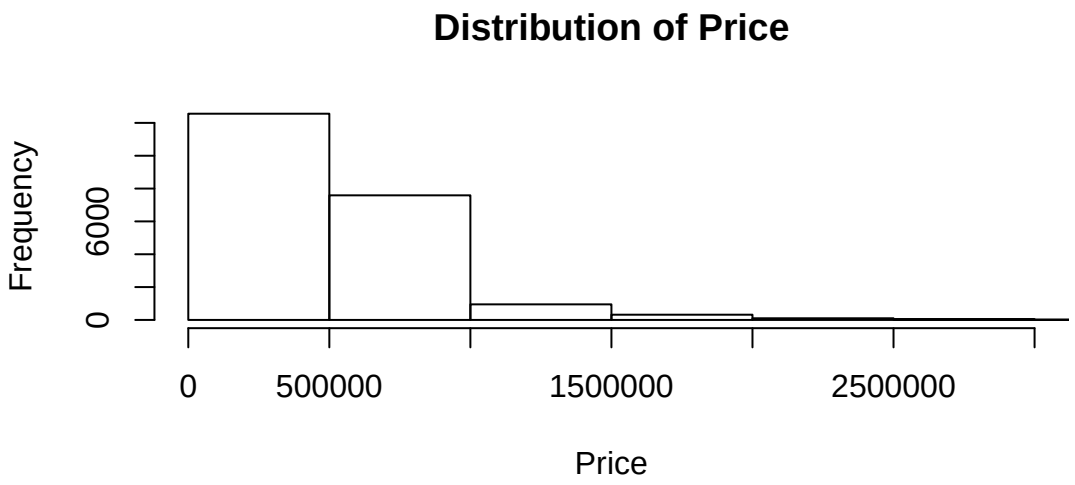
Nick George and Spencer Louie

The goal of our project is to figure out what factors affect housing prices. This data is from home sales between May 2014 and May 2015 in King County Washington. It includes 21613 observations, each one an individual home sale. The relevant variables are displayed and explained below:

Price - the sale price of the home.
Bedrooms - the number of bedrooms.
Bathrooms - the number of bathrooms per bedroom.
Sqft_Living - square footage of the house.
Sqft_lot - square footage of the lot.
Floors - the number of floors in the house. Waterfront - 0 or 1 depending on if the house has a waterfront view or not.
Condition - condition of the house 1 to 5.
Grade - grade given to house based on King County grading system 1 - 13.
Yr-built - year the house was built.
Yr-renovated - year the house was renovated if applicable. 0 if not renovated. - Could be treated as a dummy variable for renovations in the future.
Zipcode - the zipcode of the house sale. Could be converted into categorical areas for analysis.

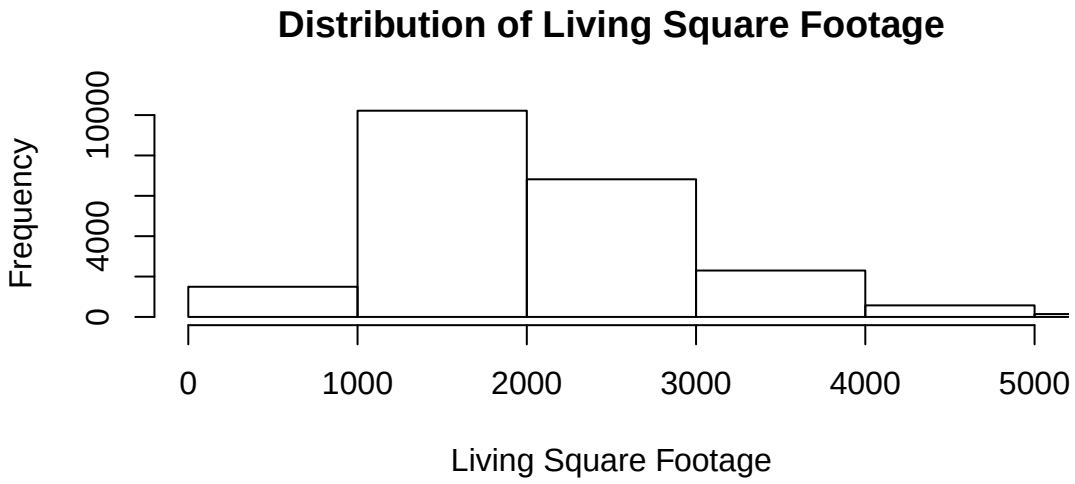
	Mean	Std. Dev.	25th Pct	Median	75th Pct
Bedrooms	3.37	0.93	3.00	3.00	4.00
Condition	3.41	0.65	3.00	3.00	4.00
Grade	7.66	1.18	7.00	7.00	8.00
Liv. Sq. Ft.	2079.90	918.44	1427.00	1910.00	2550.00
Lot Sq. Ft.	15106.97	41420.51	5040.00	7618.00	10688.00
Waterfront	0.01	0.09	0.00	0.00	0.00
Year Built	1971.01	29.37	1951.00	1975.00	1997.00
Bathrooms	2.11	0.77	1.75	2.25	2.50
Floors	1.49	0.54	1.00	1.50	2.00
Price	540088.14	367127.20	32190.00	450000.00	645000.00

This table includes all the relevant variables for which the statistics are useful. Looking at the median of year renovated for example is misleading because it is 0 for homes that have not been renovated. Similarly zipcode's statistics would be misleading. On average a house has 3.37 bedrooms, is built in 1971 and sells for $540,088.14. Furthermore we can see that some of the data is heavily skewed, like price and lot square footage, where the mean is far greater than the median. They also have fairly large standard deviations. Floors on the other hand seems much more symmetrically distributed.

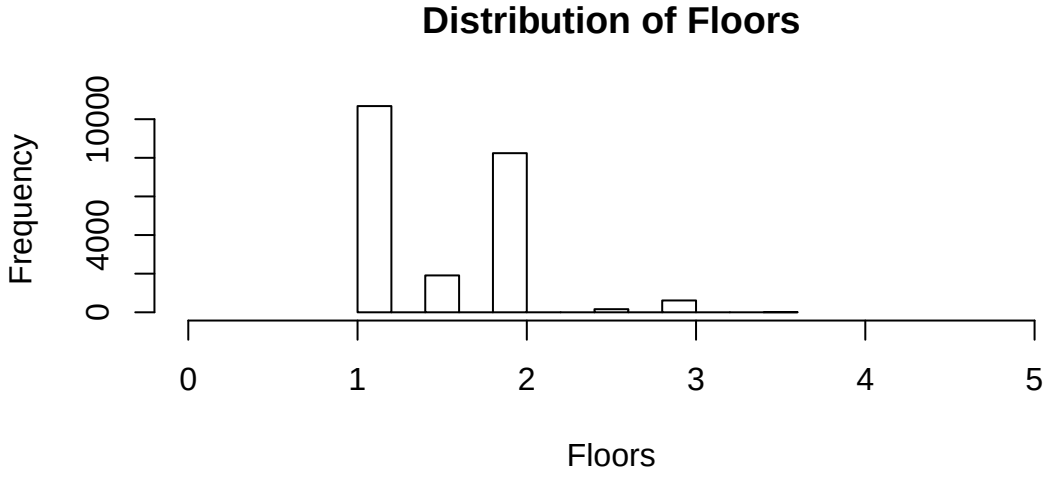


Price seems to be heavily skewed. By far the majority of the houses are sold for less than a million dollars, but still a few higher cases pull the mean up.

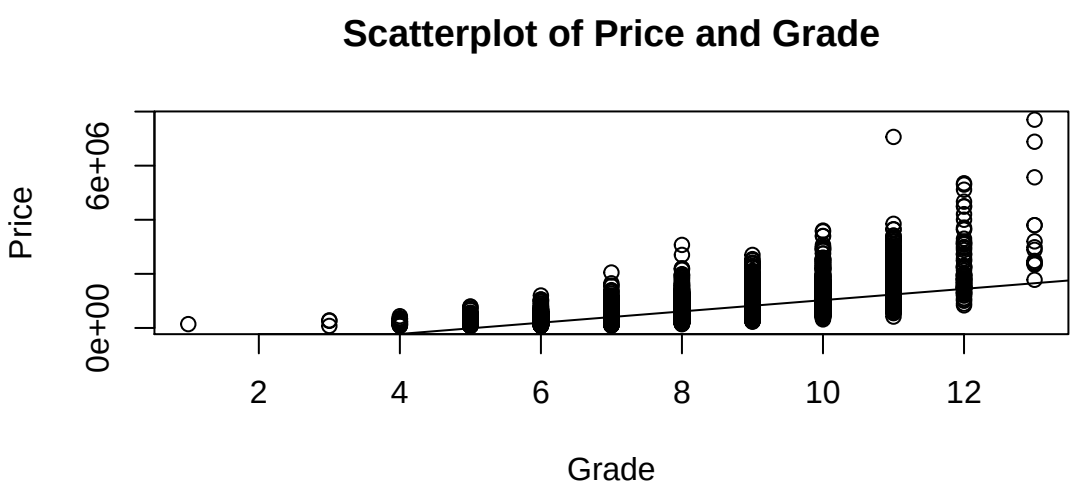
Bedrooms appears to be fairly symetrically distributed. A large portion is between 2 and 4, but then the weightings at the tails of 0 to 2 and 4 to 6 are pretty similar.



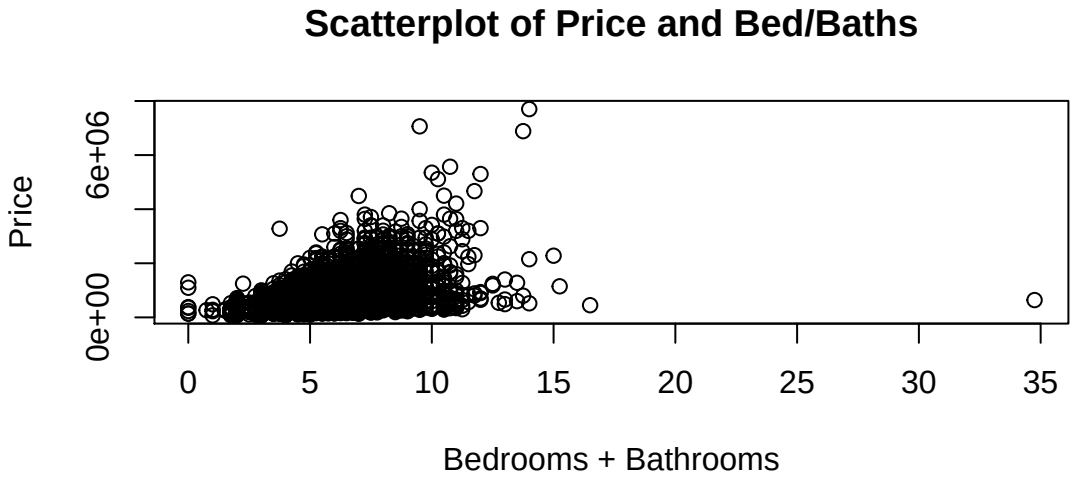
The distribution of living square footage, that is square footage of house itself, is somewhat similar to price, however not as drastic. While it's not as symetrically distributed as bedrooms, the majority of observations are between 1000 and 3000, with smaller tales.



The floor histogram is again similar to price. We see the majority of responses on the left, with a few observations at the higher end. However we do not see the same high end outliers like we do in price. The data is closer together.



Looking at the relationship between price and grade, we see a clear relationship between the grade of the house (based on King County grading system) and the price. As the grade increases, so does the price. The line on the graph is the line of best fit showing this relationship.



A scatterplot of bedrooms + bathrooms and price shows another expected relationship. As the number of bed/baths increases, the price seems to increase as well.

There were two aspects about the data that surprised us. The first was the relatively high average grade of the houses. At 3.4 it is well above what would be the pure middle of the 1 to 5 scale at 2.5. We somewhat expected there to me more lower-end quality houses in the dataset, although it's possible that renovated houses were originally at this lower level. Similarly it is possible that lower-end quality houses just don't sell so they don't end up in the dataset. Furthermore the standard deviation and thus variance of lot square footage was a lot higher than we expected. This number could be greatly affected by a few outliers. Some observations are

Our github repository is https://github.com/spencer-louie/MATH158Proj.