# Math 158 Final Project

## Nick George and Spencer Louie

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 - 10.
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.
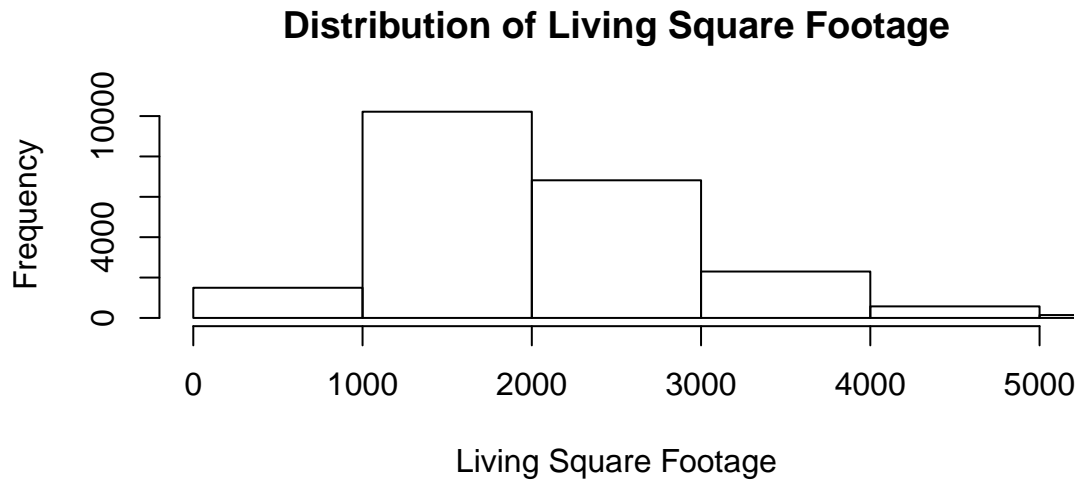
## Distribution of Price
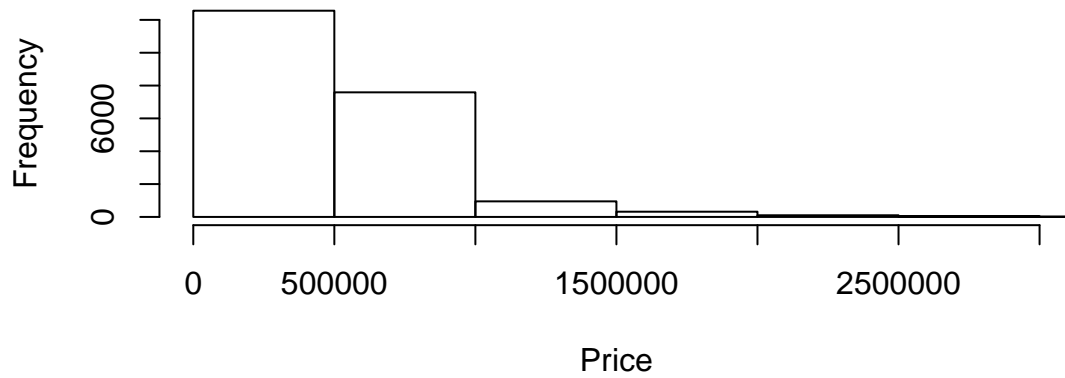


Price seems to be heavily skewed. By far the majority of the houses are sold for less than a million dollars, but still a few higher cases pull the mean up.

## Distribution of Bedrooms



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.

## Distribution of Living Square Footage



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.
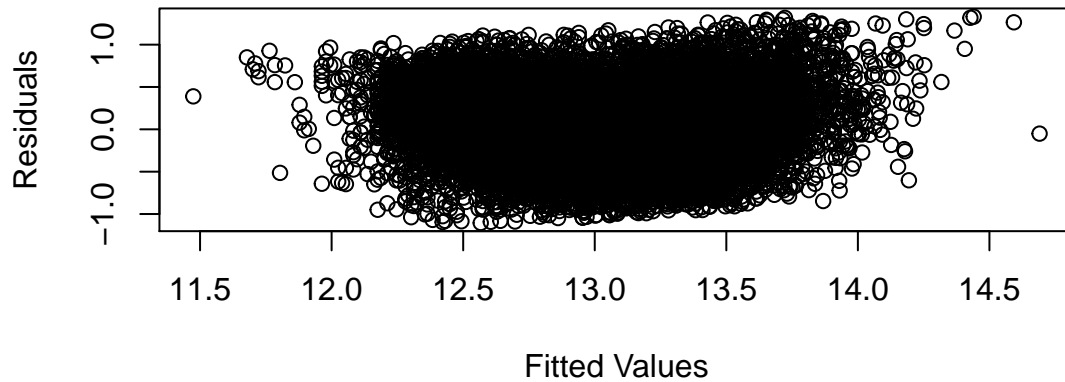
## Distribution of Floors
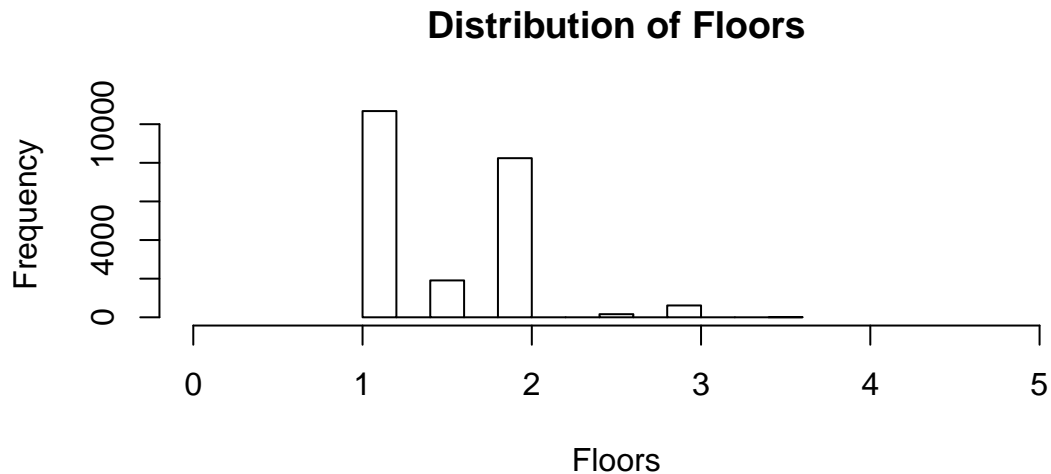


─────New Stuff─────

We're interested in testing whether or not theire is a positive relation between housing price and the square footage of living area. So $H_0 : \beta_1 \leq 0$ and $H_a : \beta_1 > 0$.

```r
kc.lm <- lm(log(price)~log(sqft_living), data=kc_data)
kc.resid <- resid(kc.lm)
kc.hat <- fitted(kc.lm)
plot(kc.hat, kc.resid, ylab="Residuals", xlab="Fitted Values", main="Housing Prices")
```

## Housing Prices



```
tidy(kc.lm)
```

```
##               term estimate   std.error statistic p.value
## 1      (Intercept) 6.729916 0.047061982  143.0011       0
## 2 log(sqft_living) 0.836771 0.006223257  134.4587       0
```

By taking log-log transformations of our data we were able to get a residual plot that seems to satisfy some of our necessary assumptions. The residuals appear to be fairly equally distributed positively and negatively and the variance seems to be constant across the fitted values.

The p-value is roughly 0 and the t-stastic is quite large at $134/2 = 67$ so we are able to reject the null hypothesis and say that there is some positive relation between housing price and living square footage. This implies that a doubling in square footage would be associated with a $2^{.836} = 1.786$ multiplicate change in the median of price.

```
plot(log(kc_data$price),log(kc_data$sqft_living),ylab="Price",xlab="Living Sqft")
```