# PrelimProjectLookGroupC

*Sam Schneider, Raymond Saitoti, Brian Burrows*

*4/9/2018*

**Read in the data**

```
require(readr)
require(mosaic)
groupcdata <- read_csv("https://awagaman.people.amherst.edu/stat230/projectsS18/groupCdataS18.csv")
```

**Summary command on the data set**

## Put data in right format.

```
summary(groupcdata)
```

```
##       id                 date                          price
##  Length:21613       Min.   :2014-05-02 00:00:00   Min.   :  75000
##  Class :character   1st Qu.:2014-07-22 00:00:00   1st Qu.: 321950
##  Mode  :character   Median :2014-10-16 00:00:00   Median : 450000
##                     Mean   :2014-10-29 04:38:01   Mean   : 540088
##                     3rd Qu.:2015-02-17 00:00:00   3rd Qu.: 645000
##                     Max.   :2015-05-27 00:00:00   Max.   :7700000
##     bedrooms        bathrooms      sqft_living       sqft_lot
##  Min.   : 0.000   Min.   :0.000   Min.   :  290   Min.   :    520
##  1st Qu.: 3.000   1st Qu.:1.750   1st Qu.: 1427   1st Qu.:   5040
##  Median : 3.000   Median :2.250   Median : 1910   Median :   7618
##  Mean   : 3.371   Mean   :2.115   Mean   : 2080   Mean   :  15107
##  3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.: 2550   3rd Qu.:  10688
##  Max.   :33.000   Max.   :8.000   Max.   :13540   Max.   :1651359
##      floors        waterfront           view          condition
##  Min.   :1.000   Min.   :0.000000   Min.   :0.0000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000
##  Median :1.500   Median :0.000000   Median :0.0000   Median :3.000
##  Mean   :1.494   Mean   :0.007542   Mean   :0.2343   Mean   :3.409
##  3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000
##  Max.   :3.500   Max.   :1.000000   Max.   :4.0000   Max.   :5.000
##      grade          sqft_above   sqft_basement       yr_built
##  Min.   : 1.000   Min.   : 290   Min.   :   0.0   Min.   :1900
##  1st Qu.: 7.000   1st Qu.:1190   1st Qu.:   0.0   1st Qu.:1951
##  Median : 7.000   Median :1560   Median :   0.0   Median :1975
##  Mean   : 7.657   Mean   :1788   Mean   : 291.5   Mean   :1971
##  3rd Qu.: 8.000   3rd Qu.:2210   3rd Qu.: 560.0   3rd Qu.:1997
##  Max.   :13.000   Max.   :9410   Max.   :4820.0   Max.   :2015
##   yr_renovated       zipcode           lat             long
##  Min.   :   0.0   Min.   :98001   Min.   :47.16   Min.   :-122.5
##  1st Qu.:   0.0   1st Qu.:98033   1st Qu.:47.47   1st Qu.:-122.3
##  Median :   0.0   Median :98065   Median :47.57   Median :-122.2
##  Mean   :  84.4   Mean   :98078   Mean   :47.56   Mean   :-122.2
##  3rd Qu.:   0.0   3rd Qu.:98118   3rd Qu.:47.68   3rd Qu.:-122.1
##  Max.   :2015.0   Max.   :98199   Max.   :47.78   Max.   :-121.3
```

```
##  sqft_living15    sqft_lot15
##  Min.   : 399   Min.   :   651
##  1st Qu.:1490   1st Qu.:  5100
##  Median :1840   Median :  7620
##  Mean   :1987   Mean   : 12768
##  3rd Qu.:2360   3rd Qu.: 10083
##  Max.   :6210   Max.   :871200
```

```
#Change Year, Month and Date to right format
sales<-groupcdata %>%
  mutate(condition=as.factor(condition),id=1:n(),waterfront=as.factor(waterfront),zipcode=as.factor(zip
```

### Data Codebook

Our variables are:

bedrooms - The number of bedrooms in the house (numeric: 0-33)
bathrooms - The number of bathrooms in the house (numeric: 0-8)
sqft_lot - The total square footage of the property (numeric: 520-1651359)
floors - The number of floors in the house (numeric: 1-3.5)
waterfront - Indicator if the house has a waterfront view (categorical: 0-1, where 0 is no waterfront view)
condition - The overall condition of the house based upon the King County grading system (categorical: 0-5)
yr_built - The year in which the house was built (numeric: 5/2/14 - 5/27/15) zipcode - The ZIP code in which the house is located (categorical) price - The price a house was sold for (numeric: 75000-7700000)

### Analysis Plan

As stated in the proposal, we aim to create a model that can predict home values in King County, using 2014 and 2015 sales data. In addition, we will examine the following additional questions:

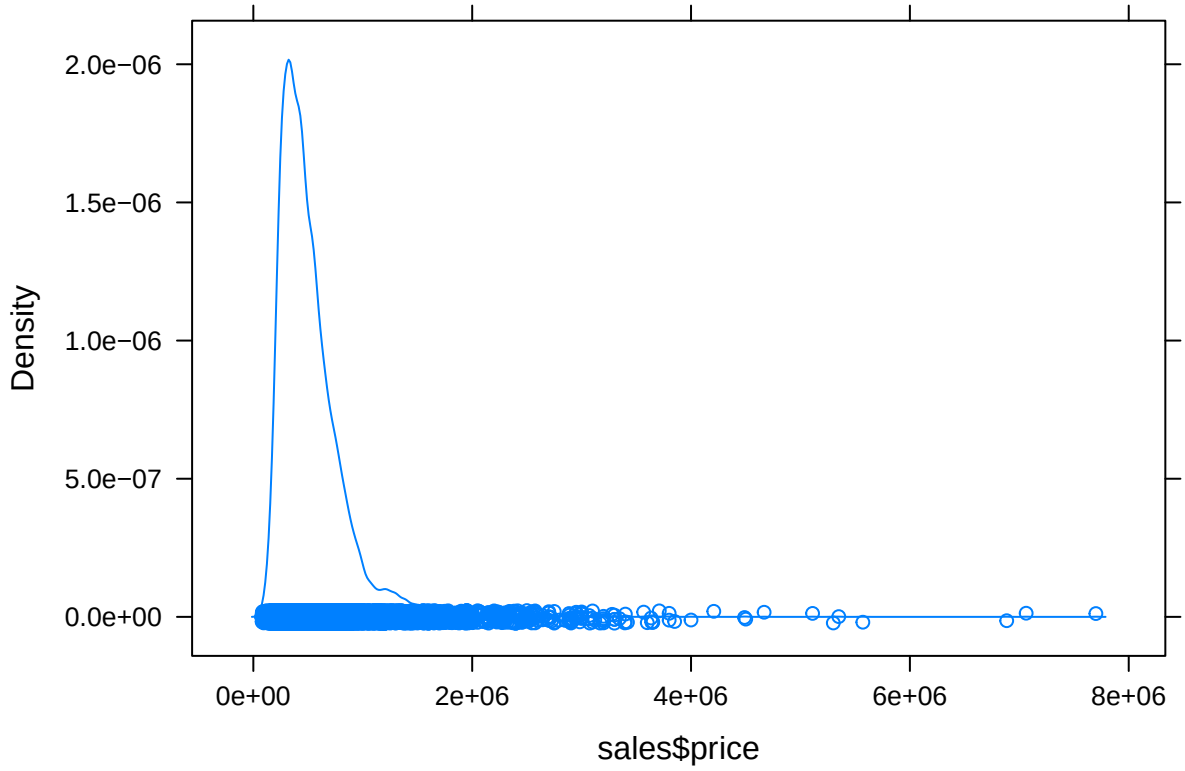
1. How much does the location of the home determine the value of that home? (Response Variable = Price , Explanatory Variable(s)= ZIP Code)
2. How much does having a waterfront view affect the impact of home location in determining home value?(Response Variable = Price , Explanatory Variable(s)= ZIP Code, Waterfront View, interaction term - ZIP Code : Waterfront View)
3. Is the square footage of a house a better predictor of home value than the number of bedrooms?(Response Variable = Price, Explanatory Variable(s)= Square Footage, Bedrooms)

### Prelim Univariate Analysis

```
fav_stats(sales$price)
```

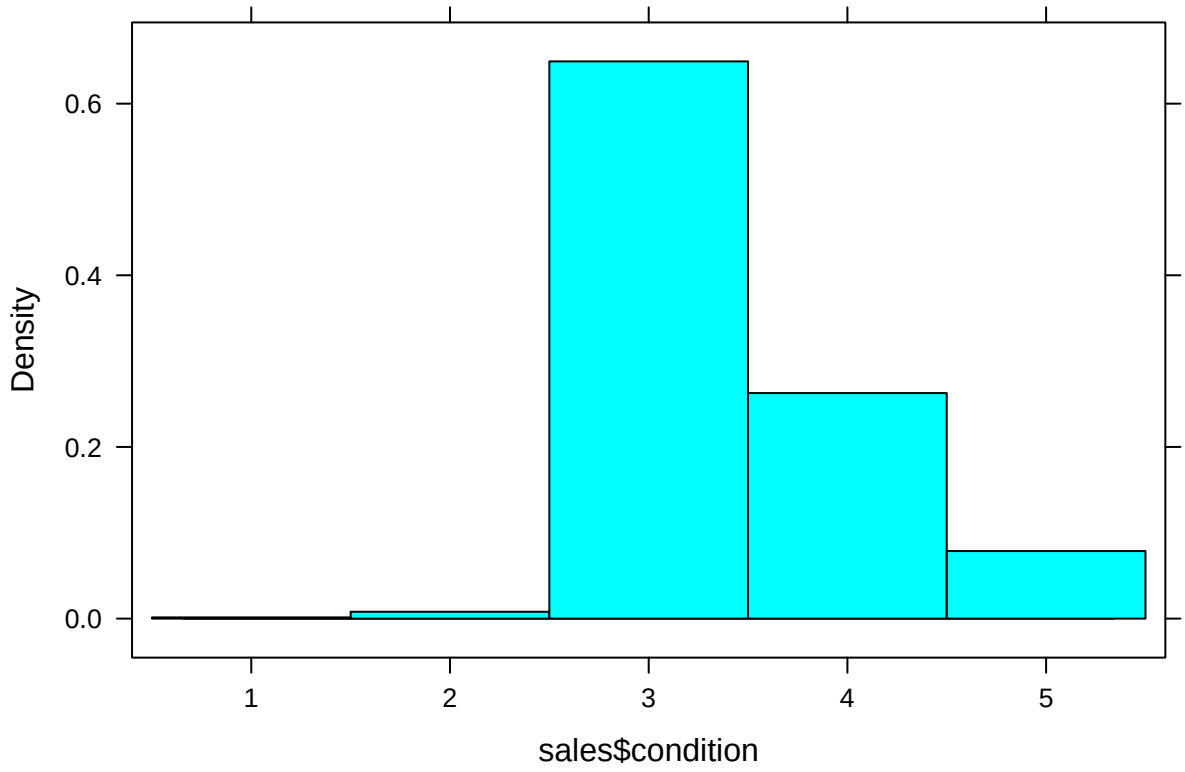
```
##    min     Q1 median     Q3     max     mean       sd     n missing
##  75000 321950 450000 645000 7700000 540088.1 367127.2 21613       0
```

```
densityplot(sales$price)
```

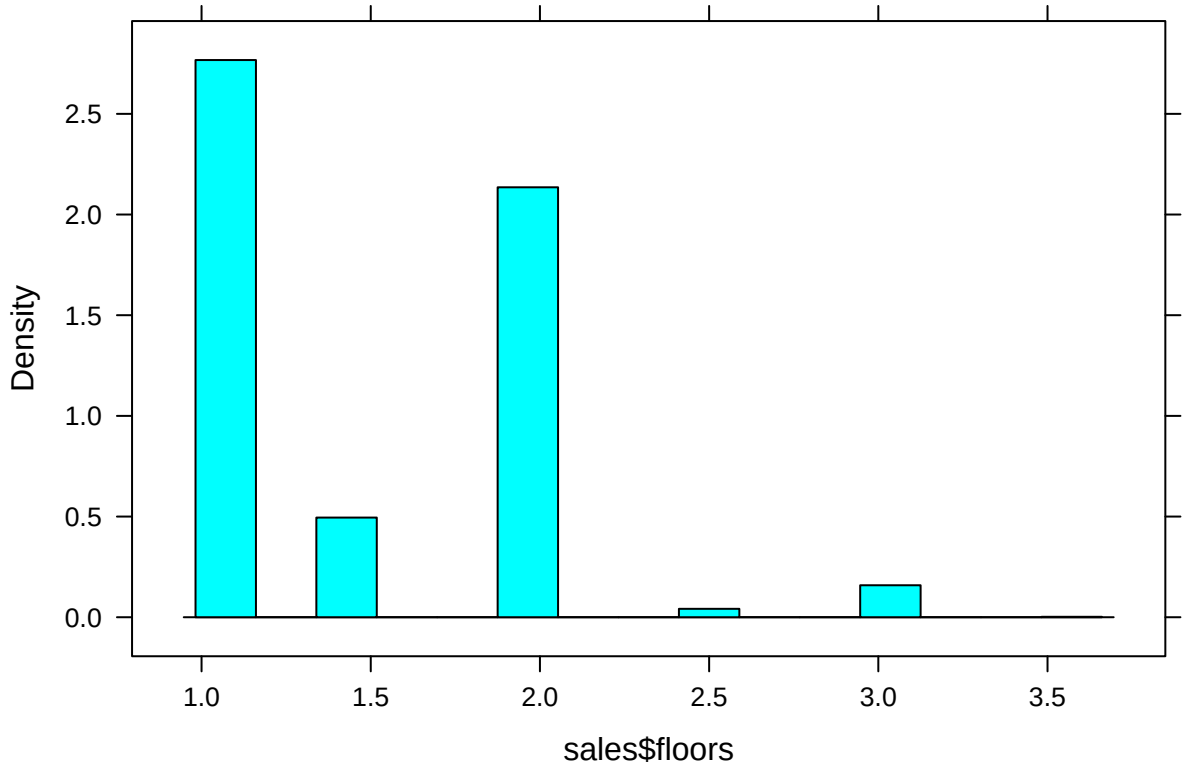



The average price of a home in this market is $540,000 although this number is likely subject to outliers as shown by the density plot. Accordingly, the median is a lower value of $450,000.

```
histogram(sales$condition)
```



```
histogram(sales$floors)
```

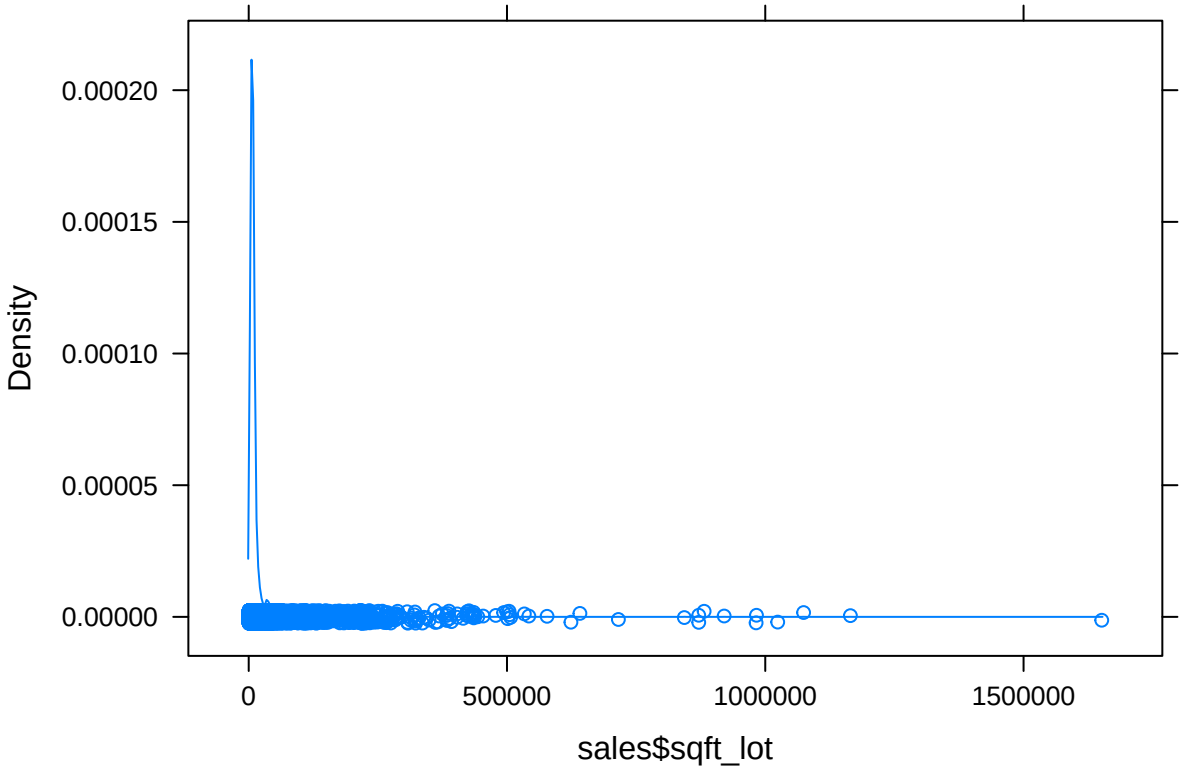



The majority of houses in this market only have one floor and it is very rare to have more than 2 floors. Furthermore, the majority of the homes are in above average condition as most homes are graded at a 3 or above.

```
favstats(sales$sqft_lot)
```

```
##  min   Q1 median    Q3     max     mean       sd     n missing
##  520 5040   7618 10688 1651359 15106.97 41420.51 21613       0
```

```
densityplot(sales$sqft_lot)
```




The values for sqft in this market is extremely distorted by outliers. The mean value is over 15,000 square feet while the median is roughly half of the mean. This will be an important factor to consider as we continue our analysis.

```
tally(sales$waterfront)
```

```
## X
##     0     1
## 21450   163
```

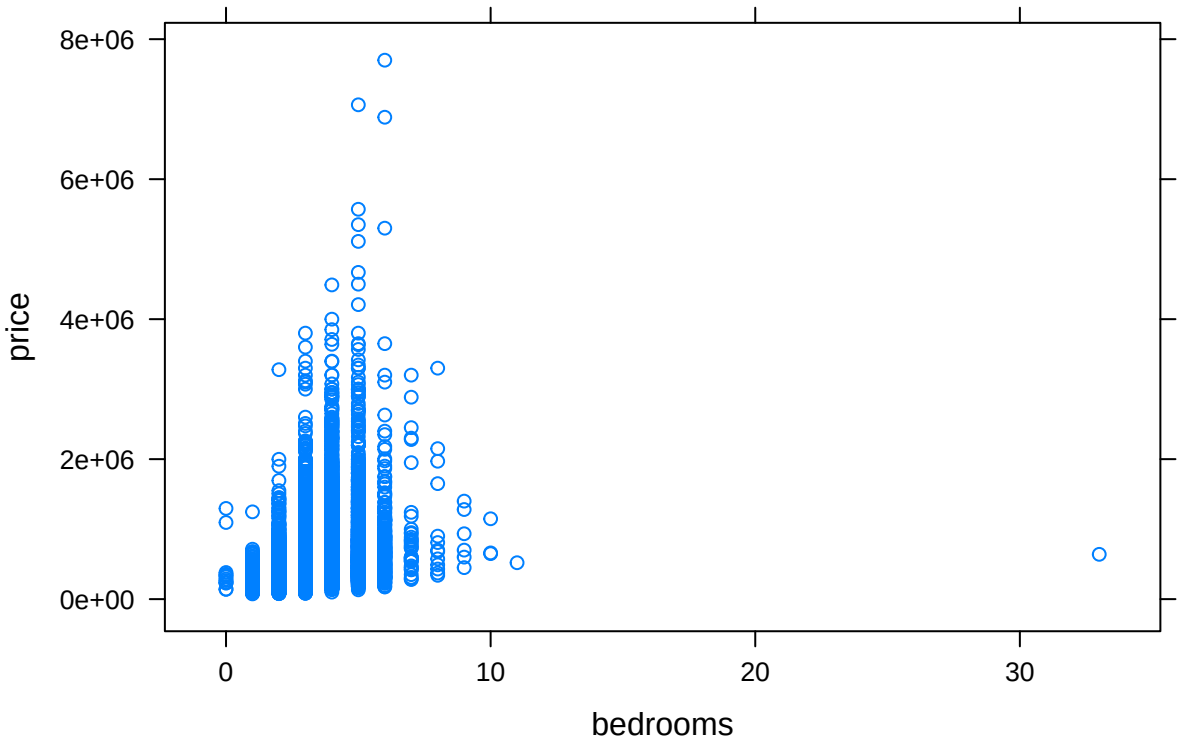
It is very rare to have a waterfront home in this market. This could make it very valuable in terms of the price of the home.

**Prelim Bivariate Analysis**

We begin with simple analyses of the relationships price has with each of our predictors to see if transformations are necessary.

First is price and bedrooms.

```
xyplot(price~bedrooms, main="House Price vs. Number of Bedrooms", data=sales)
```



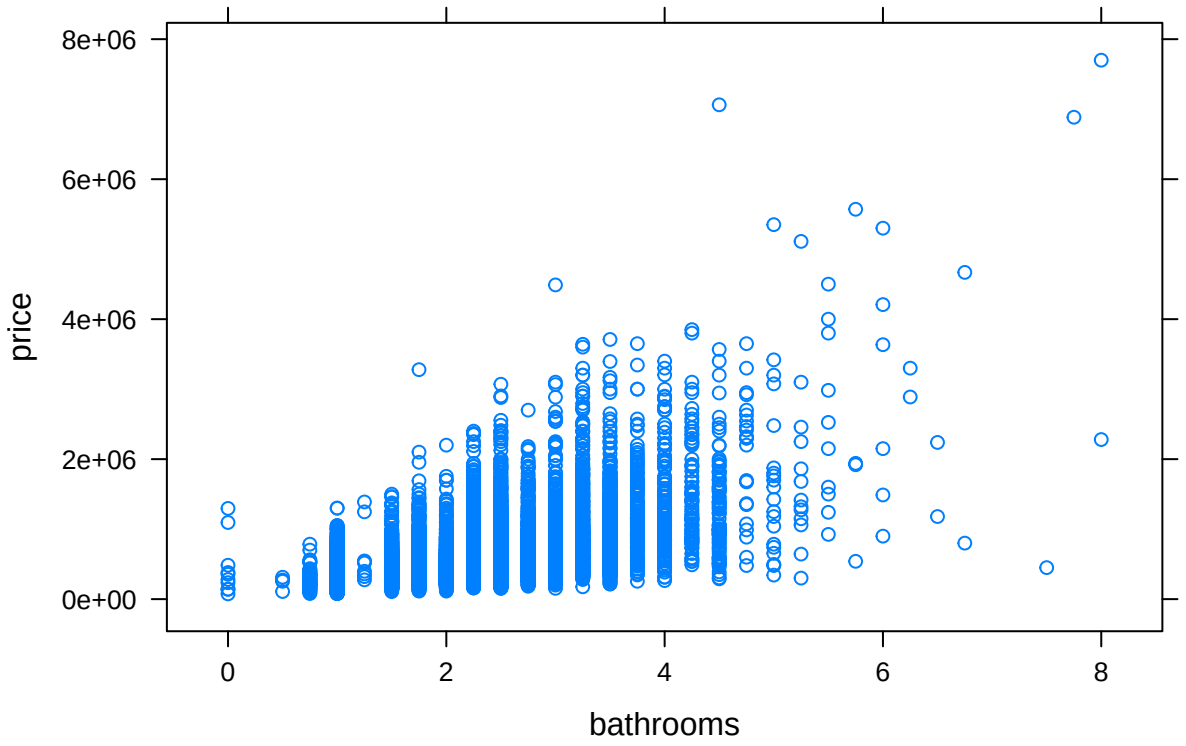

```
with(sales, cor(price,bedrooms))
```

```
## [1] 0.3083496
```

From the graph, we see that baseline home values seem to increase as the the number of bedrooms increases. However, home values seem to increase up to six bedrooms and decrease afterwards. This might most likely be the result of a small sample size for houses with more bedrooms than that. It might be worthwhile to consider bedrooms and zipcode together in future analysis. There is also one extreme outlier with more than thirty rooms. We might want to consider looking more closely at that point to determine if we should remove it. Also, as expected, home price and number of bedrooms seems to be positively correlated, although weakly so.

Second is price and bathrooms.

```
xyplot(price~bathrooms, main="House Price vs. Number of Bathrooms", data=sales)
```



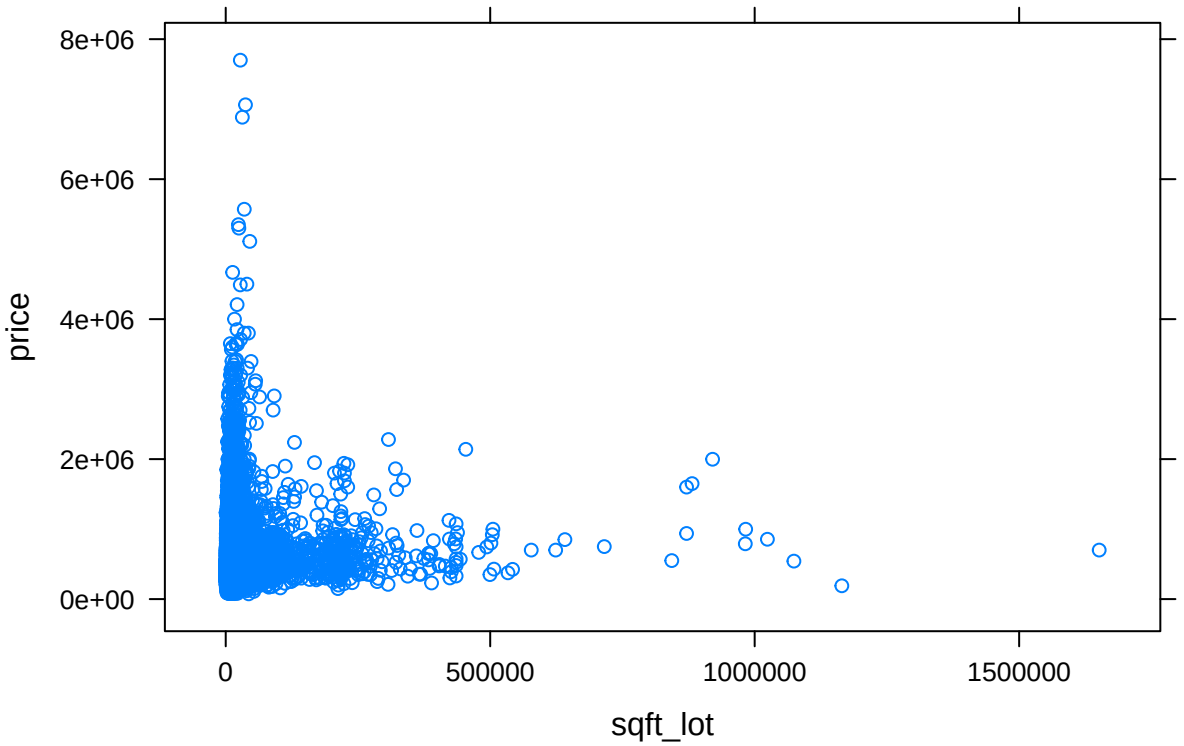

```
with(sales, cor(price,bathrooms))
```

```
## [1] 0.5251375
```

The graph indicates a moderate positive relationship between price and number of bathrooms. However, it is worthwhile noting that there are far fewer observations for houses with five or more bathrooms. This might limit the scope to which we may predictions out of fear of extrapolation beyond the data.

Third is price and square footage.

```
xyplot(price~sqft_lot, main="House Price vs. Square Footage", data=sales)
```



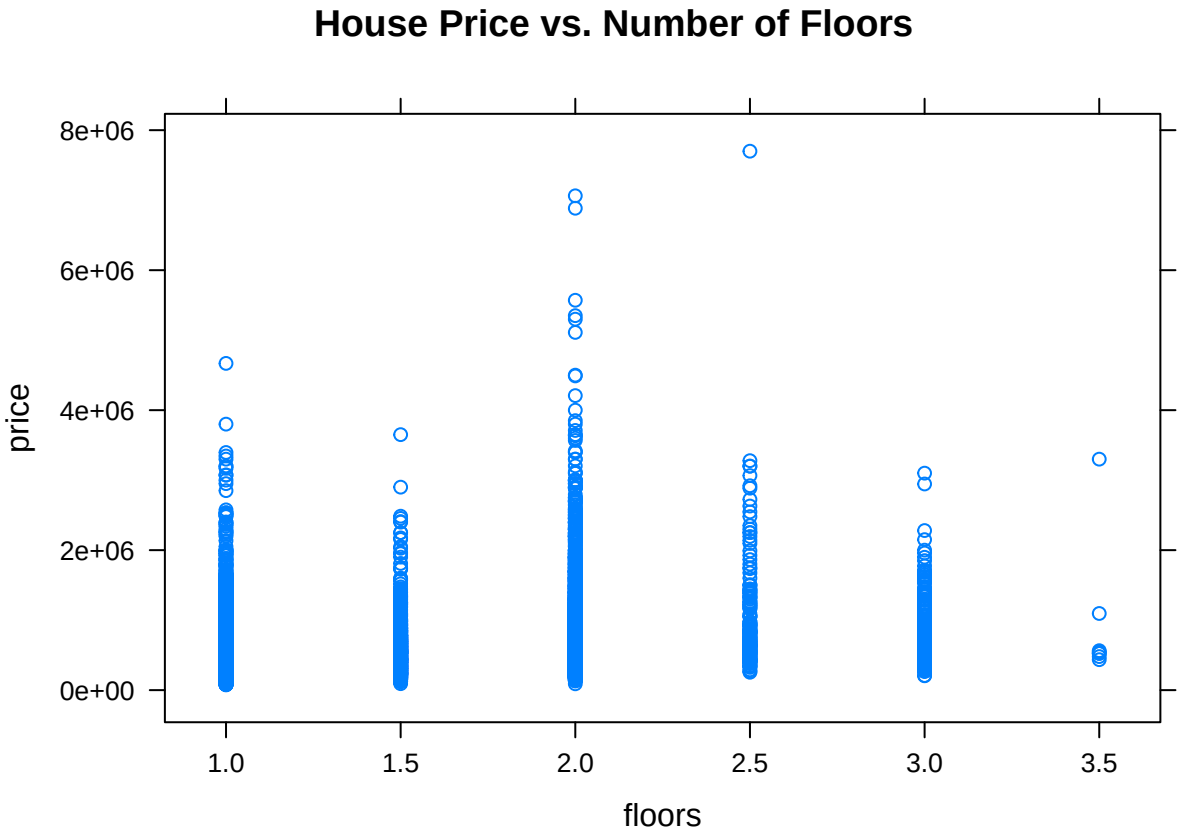

```
with(sales, cor(price,sqft_lot))
```

```
## [1] 0.08966086
```

The graph of house price and square footage indicates that transformations will most likely be required. Furthermore, very few observations exist for homes of more than 500,000 square feet. In addition, a correlation test indicates that price and square footage is very weakly positively correlated, which is expectated but not readily shown by the graph.

Fourth is price and floors.

```
xyplot(price~floors, main="House Price vs. Number of Floors", data=sales)
```



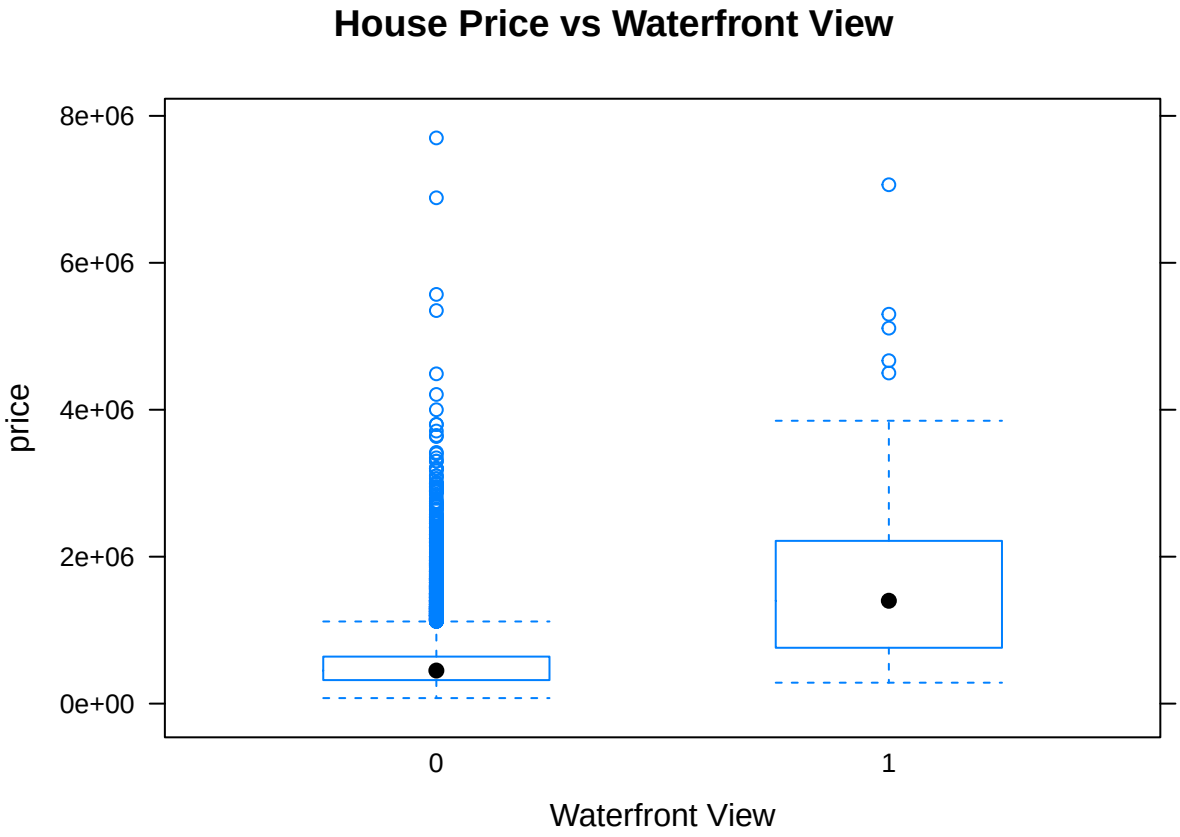

```
with(sales, cor(price,floors))
```

```
## [1] 0.2567939
```

While the scatterplot was not particularly informative, a correlation test shows that price and number of floors are weakly positively correlated. However, the scatterplot does elucidate a number of high outliers that should be examined more closely, especially the one for 2.5 floors.

Fifth is price and waterfront.

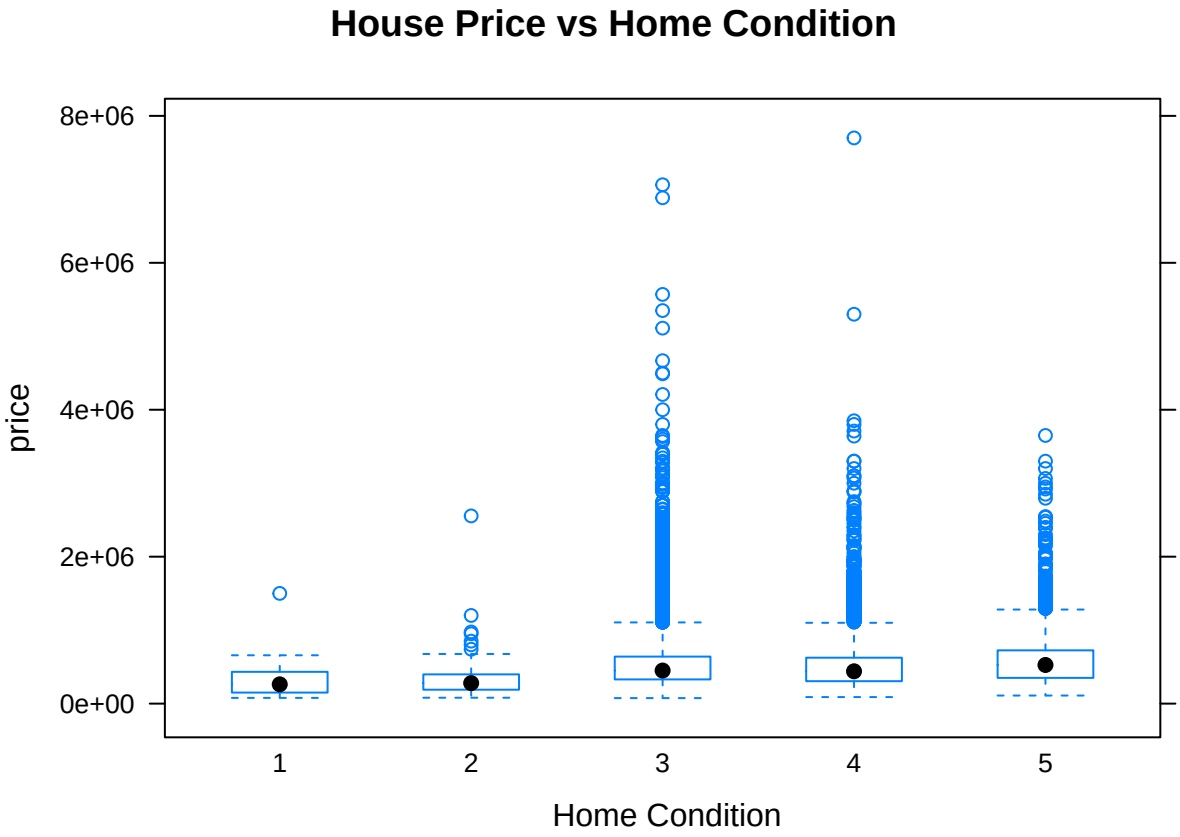
```
bwplot(price~waterfront, main="House Price vs Waterfront View", xlab="Waterfront View", data=sales)
```




Side by side boxplots of house price by waterfront view supports earlier univariate analysis as waterfront properties seem to have higher median price. Whether this difference is significant is something to be evaluated in a regression.

Sixth is price and condition.

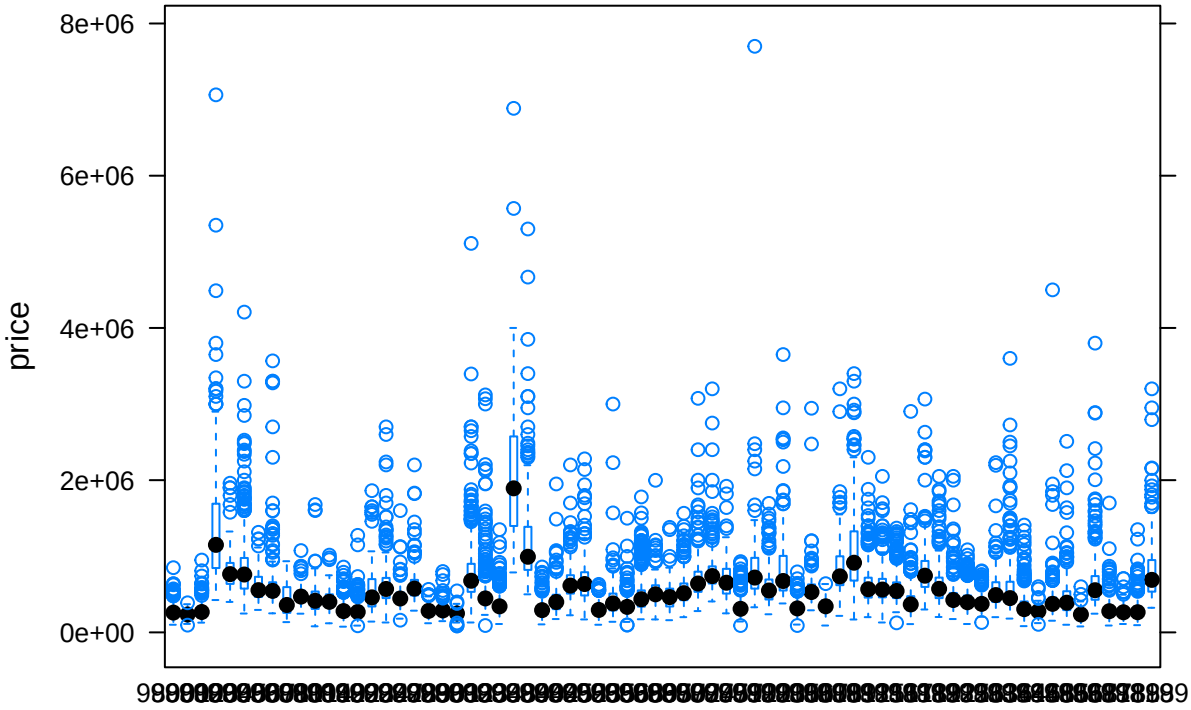
```
bwplot(price~condition, main="House Price vs Home Condition", xlab=" Home Condition", data=sales)
```




The side-by-side boxplots of price by house condition shows fairly similar medians for each home condition group. Interestingly, there are a number of outliers for homes with conditions rated three or higher. This might be explained by location or age data in additional analysis. Also, an ANOVA will be run to see if the differences in house price between condition groups is significant.

Last is price and zipcode.

```
bwplot(price~zipcode, main="House Price vs Zipcode", data=sales)
```




Boxplots of price by zip-code seem to indicate differences in house price by zip-code. An ANOVA will be run to see if any of the differences are significant. Also, for the sake of legibility, we might consider renaming the zip-codes, which should help to make the horizontal axis more readible.

**Randomization-Based Procedure Thoughts** We suspect that our price vs square footage plot might need a log transformation. However, we would like to see whether the results we see after fixing the conditions would be compatible with those obtained using a randomization test for the slope. The variable we would be shuffling in this context would be the square footage

**Questions and notes/concerns** 1. We are aware that the number of floors, bathrooms, and bedrooms in this context could also be used as categorical variables in an anova setting. Would it be more advisable to doa linear regression or ANOVA/ both?

2. What is the best way to handle overplotting in this scenario?
3. We might have too many zip codes than we anticipated.