1. *(10 points) Under Zellner's g-prior* $\Sigma_o = \sigma^2(X^TX)^{-1}$, *with* $\underset{\sim}{\beta_o} = \underset{\sim}{0}$, *state your prior for* σ^2 *and derive the marginal distribution* $p(\sigma 2|\underset{\sim}{y}, X)$. $\gamma = \frac{1}{\sigma^2} \sim \text{GAM}(\frac{\nu_o+n}{2}, (\nu_o\sigma_o^2 + SSR_g)/2)$

   $\gamma = \frac{1}{\sigma^2} \sim \text{GAM}(\frac{\nu_o}{2}, \nu_o\sigma_o^2/2)$

   $\text{p}(\gamma|\underset{\sim}{y}, \underset{\sim}{x}) \propto \text{p}(\gamma)p(\underset{\sim}{y}|\underset{\sim}{X}, \gamma)$

   Generally we think $\text{p}(\gamma|\underset{\sim}{y}, \underset{\sim}{x}, \beta) \propto \text{p}(\underset{\sim}{y}|\underset{\sim}{X}, \gamma, \beta)p(\gamma|\beta)p(\beta)$,

   but for the same reason as the previous homework, *which I still don't understand*, we can simplify and remove the β terms.

   We have: $\text{p}(\frac{1}{\sigma^2}|\underset{\sim}{y}, \underset{\sim}{x}) \propto \text{p}(1\frac{}{\sigma^2)p(\underset{\sim}{y}|\underset{\sim}{X}, \frac{1}{\sigma^2})}$

   $\propto (1\frac{}{\sigma^2)^{\nu_o/2-1}exp(-\frac{1}{\sigma^2}[\nu_o\sigma_o^2/2)\times(\frac{1}{\sigma^2})^{n/2}exp(-\frac{1}{2\sigma^2}(\underset{\sim}{Y}-\underset{\sim}{x}\beta)^T(\underset{\sim}{Y}-\underset{\sim}{x}\beta))}$

   $\propto 1\frac{}{\sigma^{2(\nu_o+n)/2}\times exp(-\frac{1}{2\sigma^2}[\nu_o\sigma_o^2+(\underset{\sim}{Y}-\underset{\sim}{X}\beta)^T(\underset{\sim}{Y}-\underset{\sim}{X}\beta)^T])}$

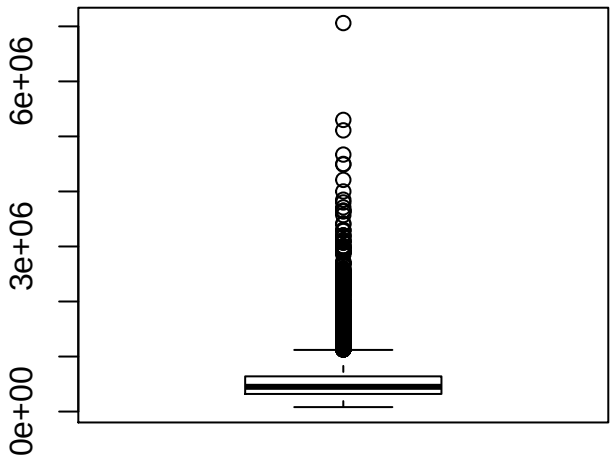
   $\propto 1\frac{}{\sigma^{2(\nu_o+n)/2}\times exp(-\frac{1}{2\sigma^2}[\nu_o^2{}_o+SSR])}$

2. *(10 points) Describe the process for choosing priors in Bayesian hypothesis testing with Bayes Factors.*

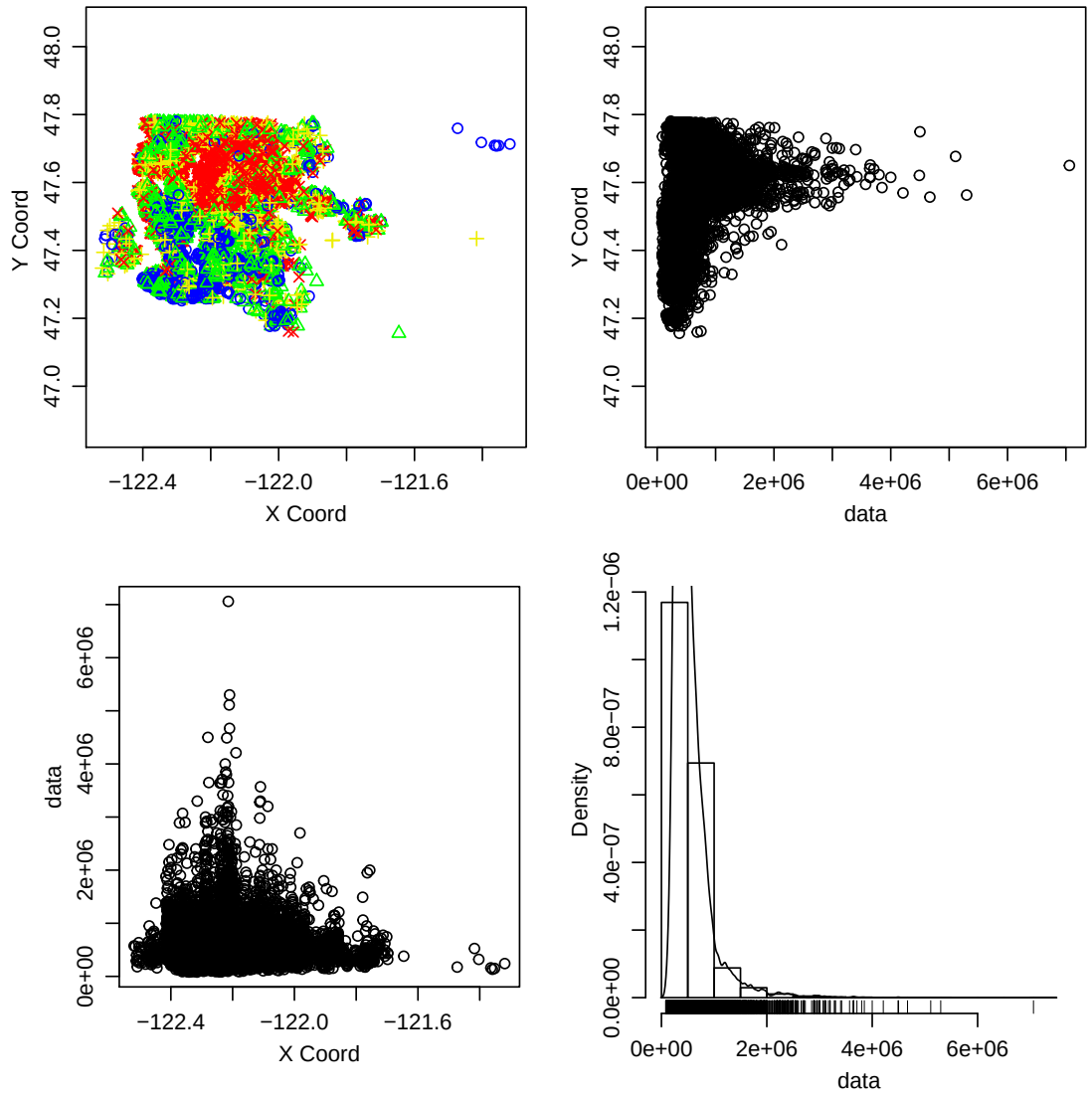
3. *Download the King County housing dataset from D2L. You have free rein to use this question to apply some of the ideas we have learned in class. The goal is to create a predictive model that best captures the housing dynamics in King County. Note, there is a dataset called predictHouse, that contains all of the covariates, but will require a predicted price. This will be part of the homework submission and all of the entries will be compared. Note you will only be asked to provide a point estimate, but I could as for a posterior predictive distribution for each home.*

   (a) *(10 points) What factors in the dataset to you anticipate being useful predictors of housing price?*

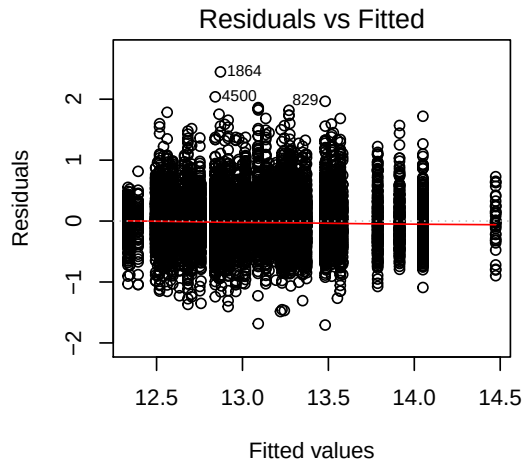
   Taken from Windermere Real Estate's website, below is a picture of King County Washington.

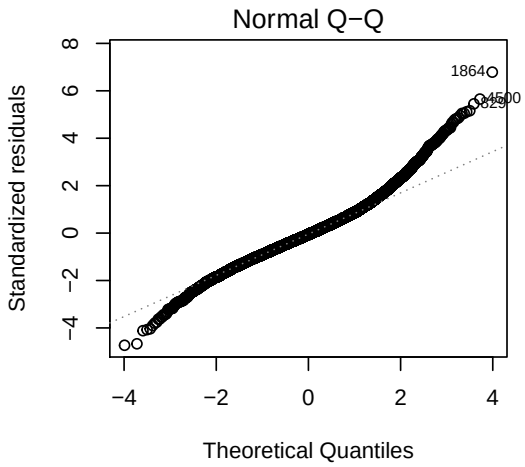


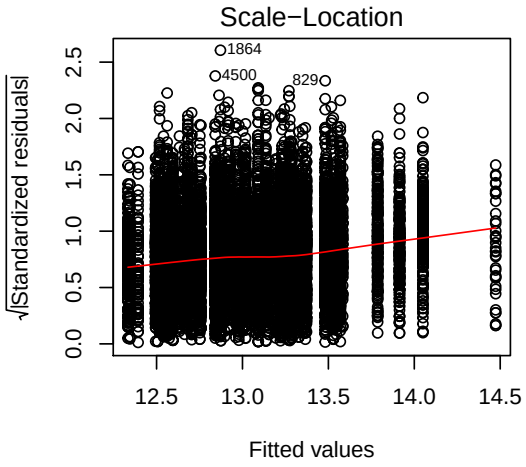

```
## as.geodata: 398 replicated data locations found.
##  Consider using jitterDupCoords() for jittering replicated locations.
## WARNING: there are data at coincident or very closed locations, some of the geoR's functions
##  Use function dup.coords() to locate duplicated coordinates.
##  Consider using jitterDupCoords() for jittering replicated locations
## [1] 379
```

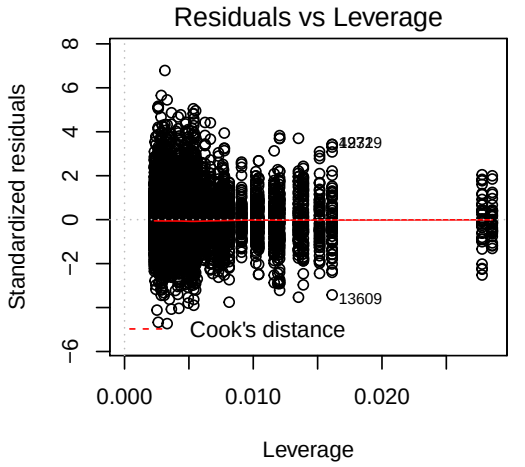


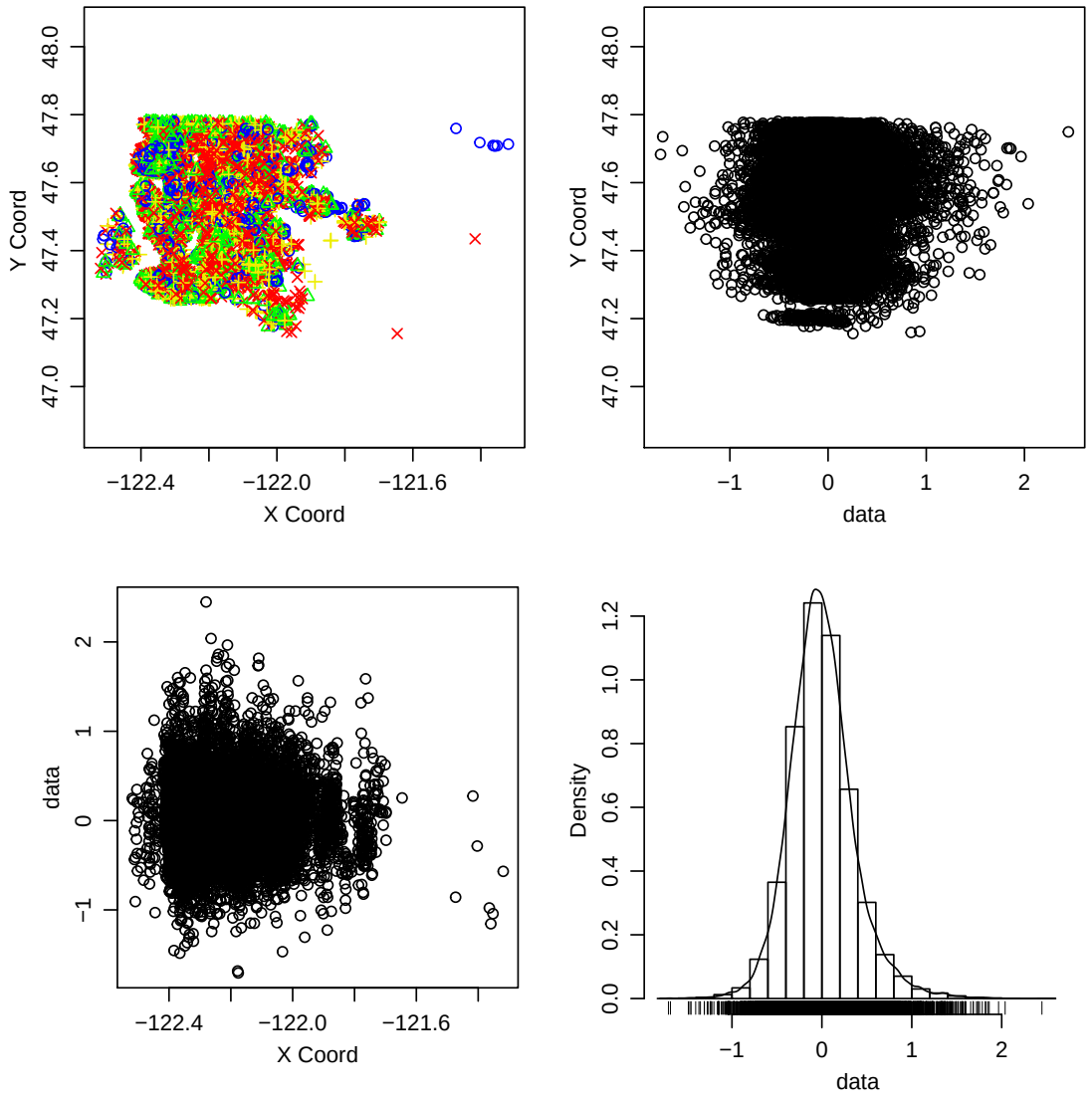
The geo plot visualizes the area considered. If location coordinates did not affect price, the north east and south west plots would show random scatter. Prices at certain location coordinates are higher than others. There are also a few houses located farther from others. Perhaps zipcode will serve as a good enough proxy for location coordinates. A SLR model was fit to model price as a function of zipcode. Due to the increasing variation and the large right skew in the residuals, price was logged. Doing so made the constant variance and normality assumptions more reasonable, but not perfect.


Residuals vs Fitted
Residuals
Fitted values
1864
4500
829


Normal Q−Q
Standardized residuals
Theoretical Quantiles
1864
4500


Scale−Location
√|Standardized residuals|
Fitted values
1864
4500
829

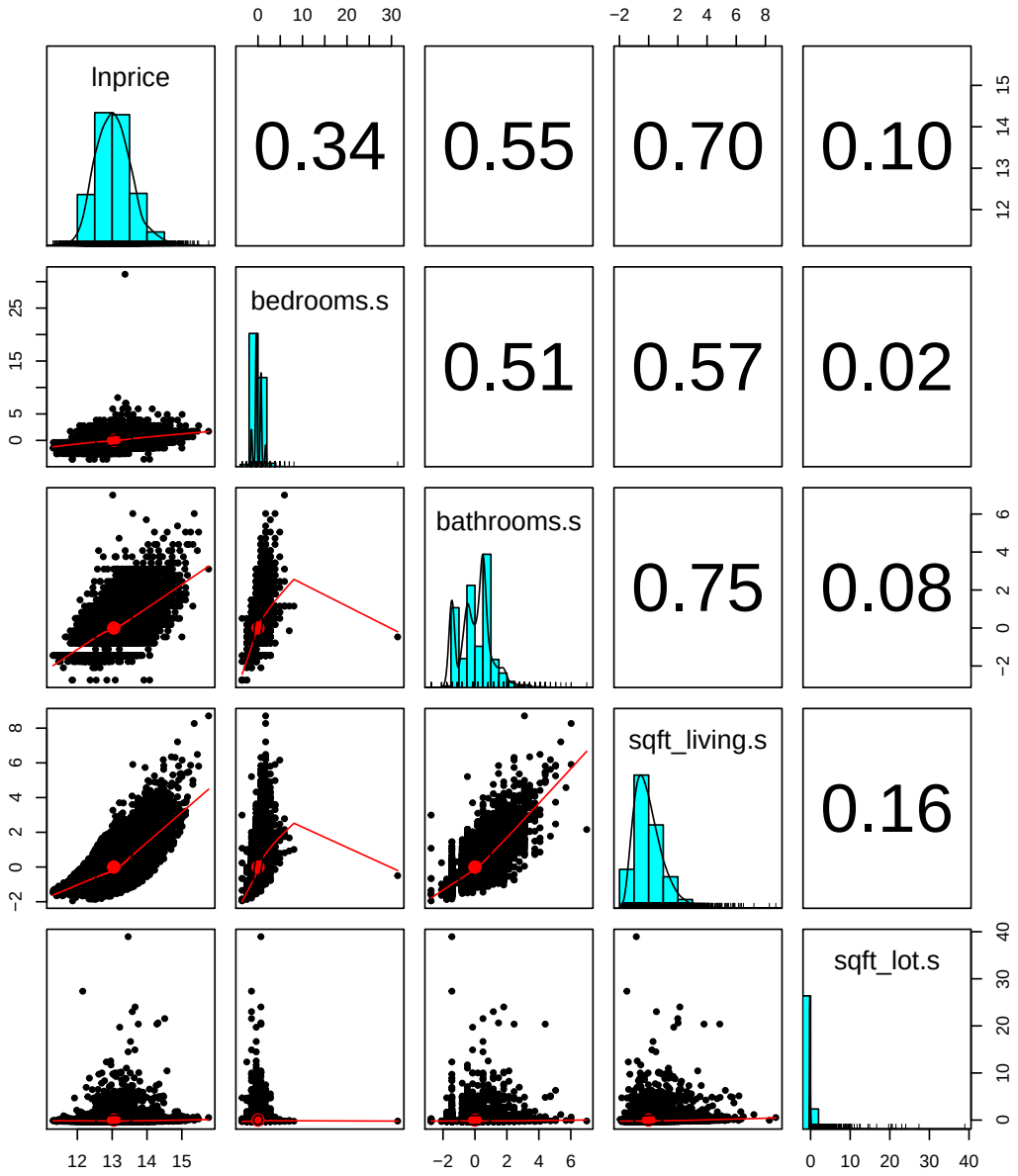

Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
13609

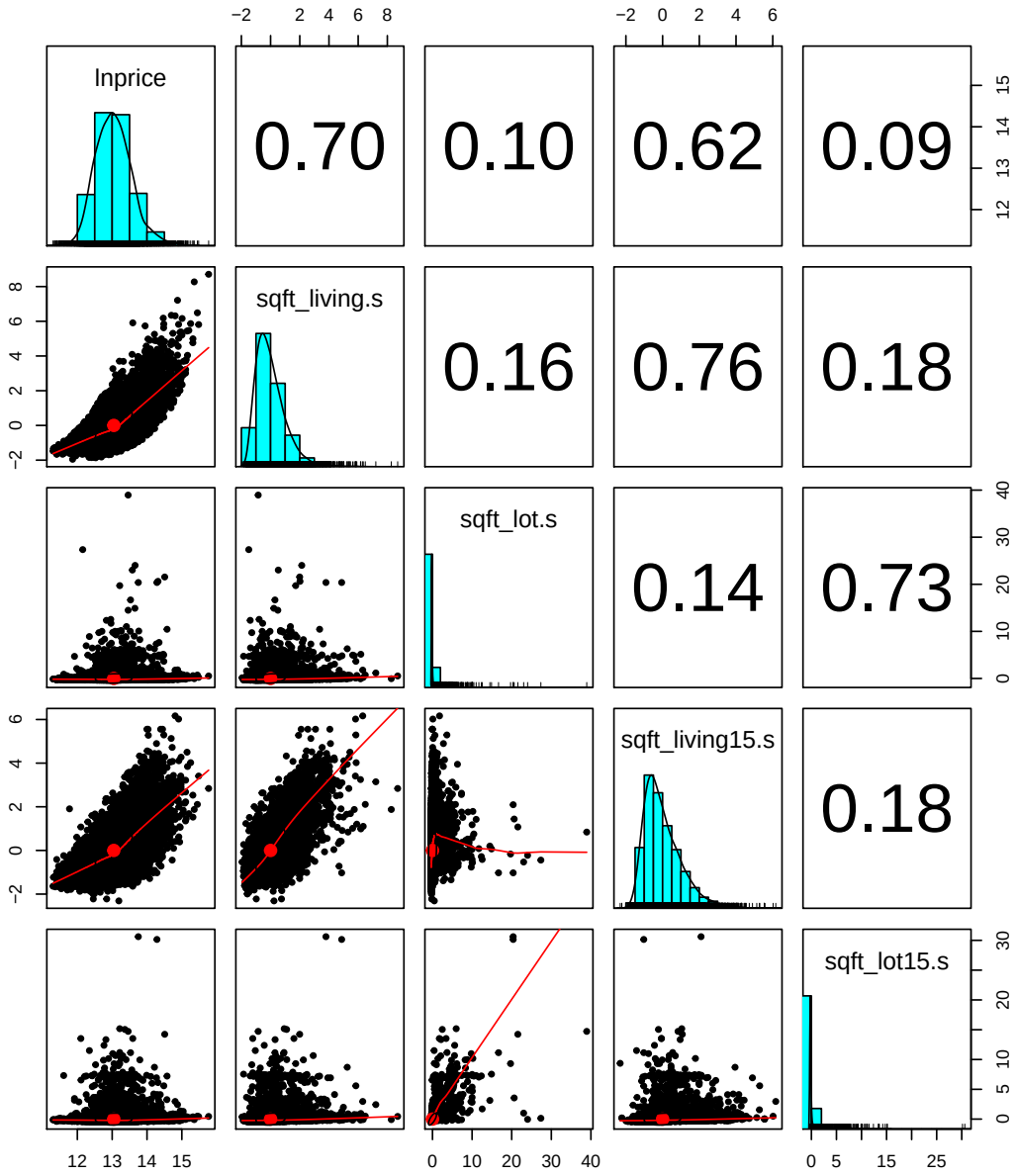


The off diagonal plots show random scatter, so relationships between price and location coordinates may be accounted for by including zipcode in the final model.

Note: zipcode is categorical!

With 23 possible covariates, I chose six continuous or reasonably approximately continuous variables to see how correlated they were.


lnprice
0.34
0.55
0.70
0.10
bedrooms.s
0.51
0.57
0.02
bathrooms.s
0.75
0.08
sqft_living.s
0.16
sqft_lot.s



One of the levels forth year renovated is "0", which is most likely a miss code. Zeros for the renovated variable imply not renovated. Only a small proportion (4.2897746%) of the houses were renovated.

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

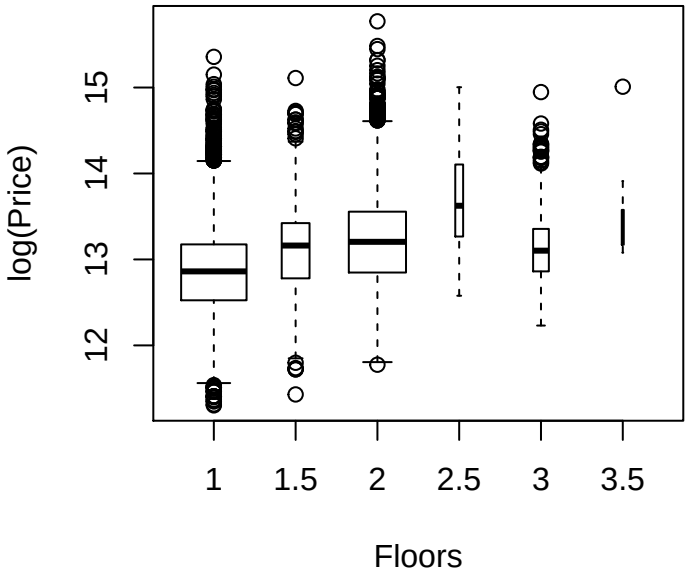
Based on the EDA, I think zipcode will be important to counteract location coordinates, though a model with variation relative to spatical coordinates would be most accurate. After accounting for zipcode, I do not think location coordinates will affect variation in prices.

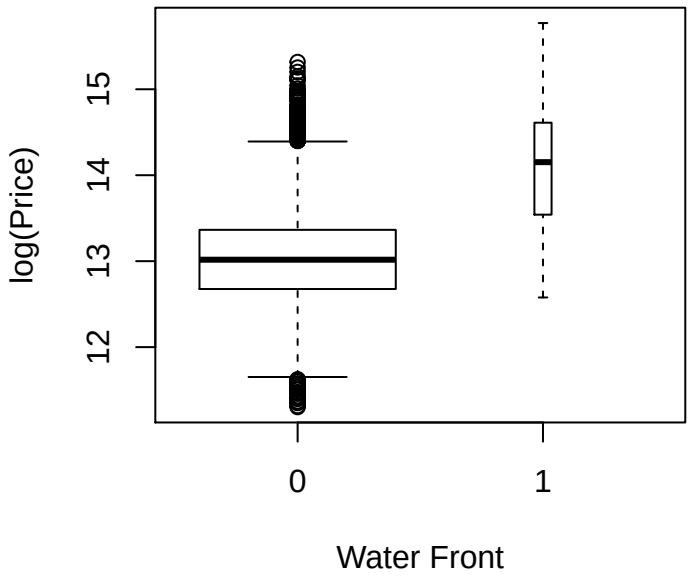
Many of the continuous variables are on much different scales and the effects of variables on

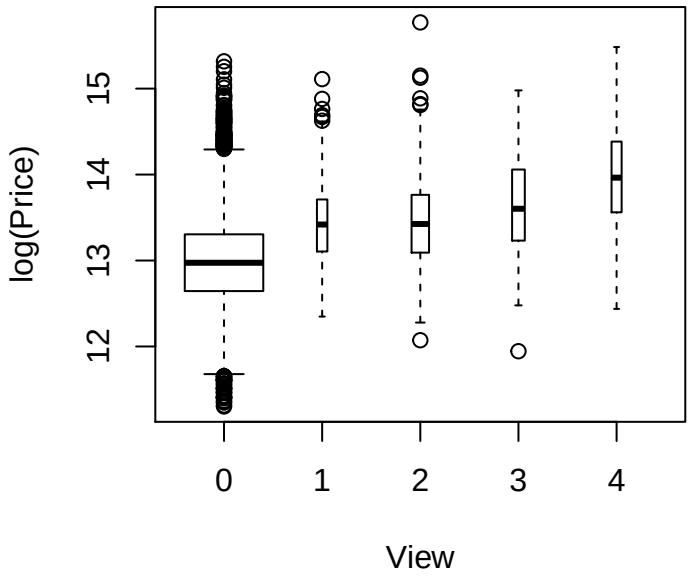
a small scale may be masked by effects of variables on a large scale, therefore, all continuous variables were scaled and centered. Bathrooms have fractional levels, and so bathrooms was also considered as continuous, scaled and centered.

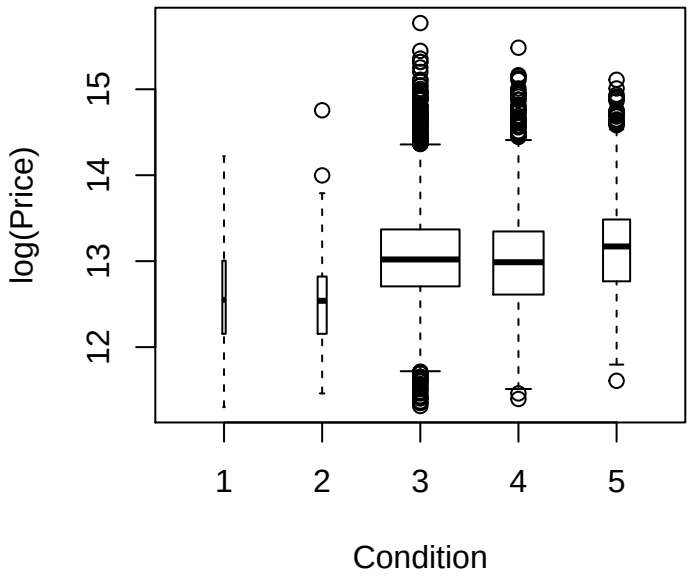
Continuous (or the like) Variables:

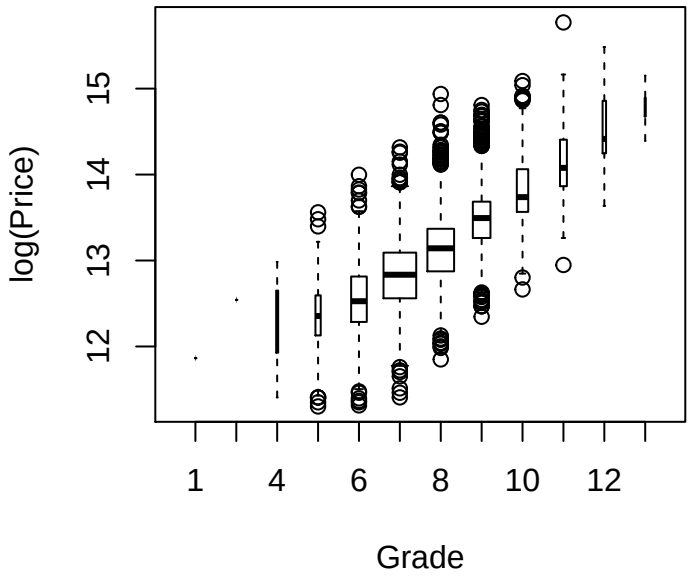
Based on the pairwise correlation plots, I believe either square footage or the average square footage of the 15 nearest houses will be important. The number of bathrooms and number of bedrooms also may be important with correlations to price of 0.52 and 0.3 respectively, as the two are highly correlated to each other, the effect of bathrooms may mask the effect of bedrooms.




log(Price)
12
13
14
15
0
1
Water Front

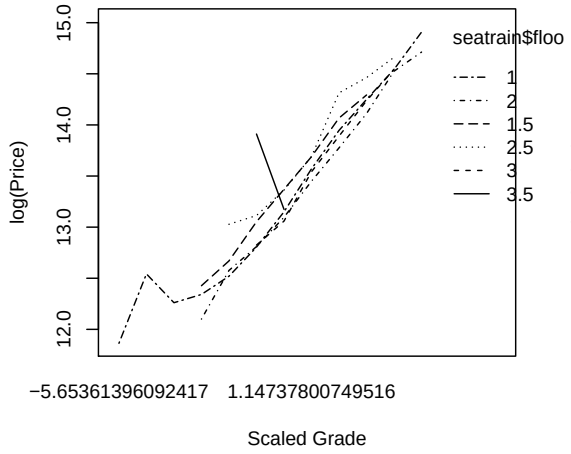

log(Price)
12
13
14
15
0
1
2
3
4
View

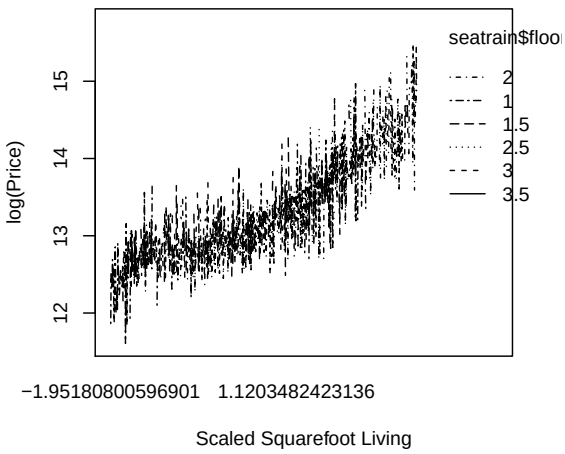


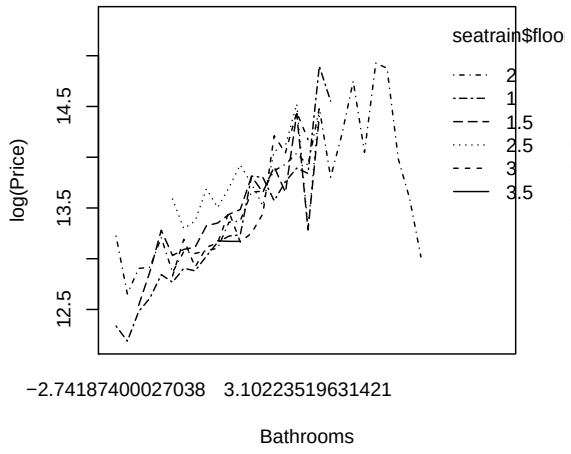


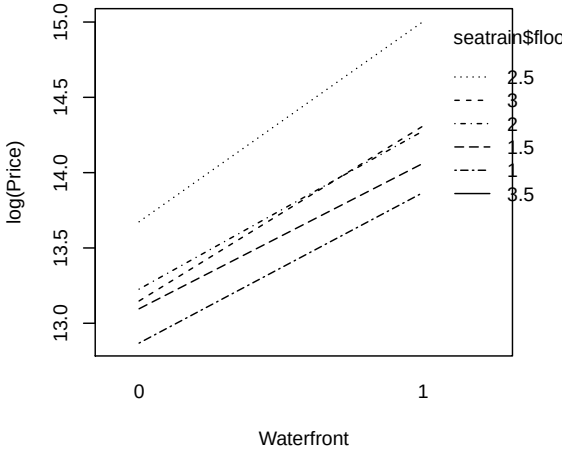
Waterfront, view, and grade all look like they may affect price, with view and grade on an exponential level.

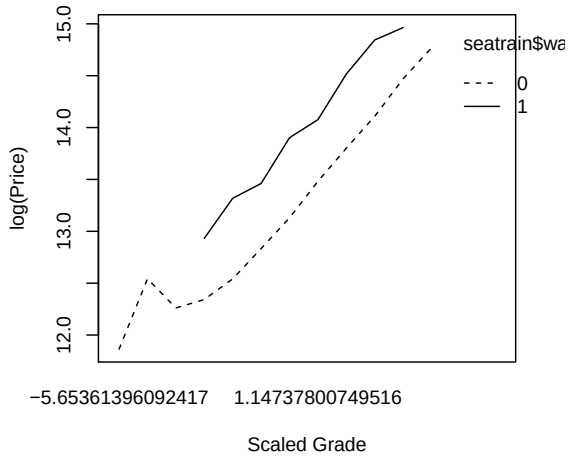
```
## Loading required package:  stringr
```

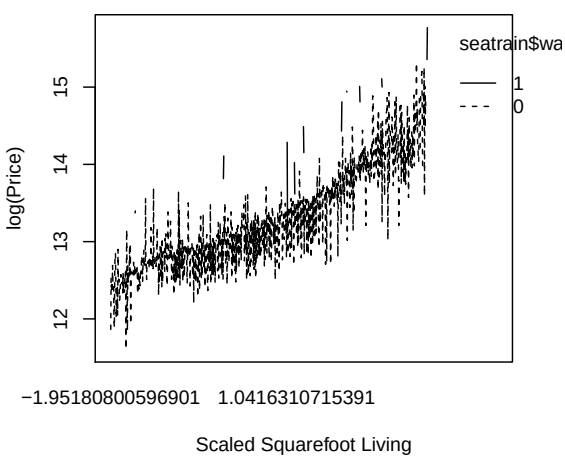


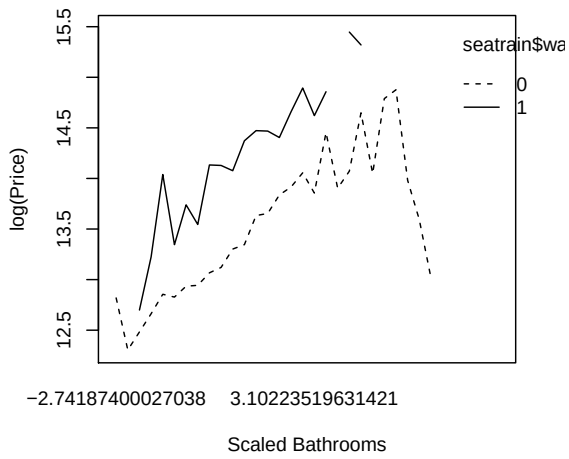


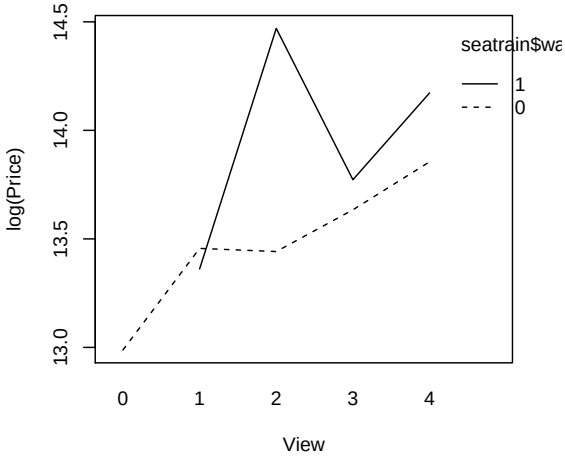


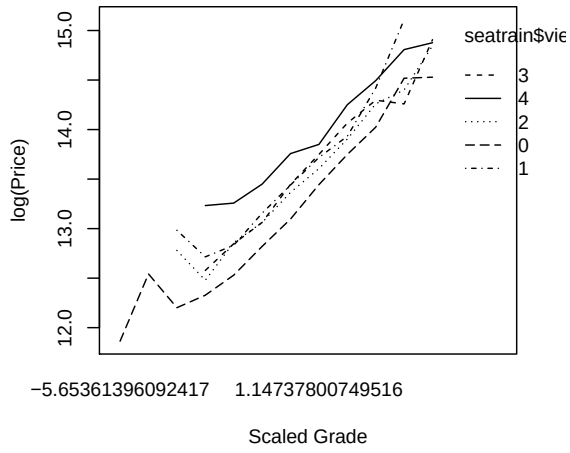


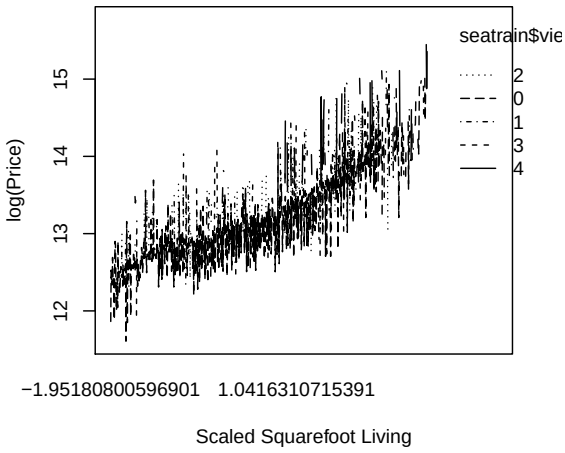

log(Price)
15.0
14.0
13.0
12.0
seatrain$wa
0
1
−5.65361396092417 1.14737800749516
Scaled Grade

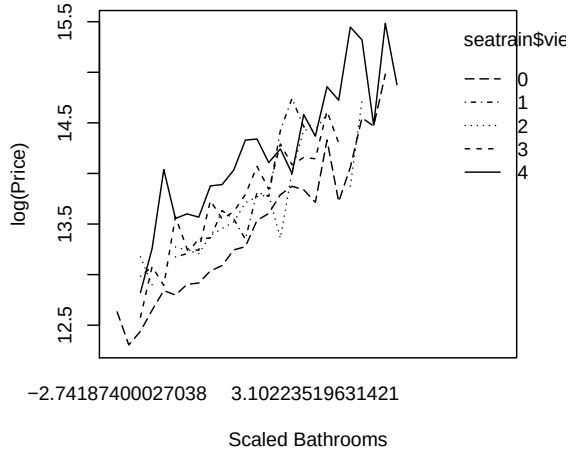

log(Price)
15
14
13
12
seatrain$wa
1
0
−1.95180800596901 1.0416310715391
Scaled Squarefoot Living


log(Price)
15.5
14.5
13.5
12.5
seatrain$wa
0
1
−2.74187400027038 3.10223519631421
Scaled Bathrooms


log(Price)
14.5
14.0
13.5
13.0
seatrain$wa
1
0
0 1 2 3 4
View







(b) *(10 points) What challenges did you face in understanding or processing the data?*

EDA took a long time, so it took a lot to understand the variables and get to the point where I could try out model selection in terms of Bayes.

R Code