Data\_Exploration\_and\_Cleaning\_*RR*.R

rakeshravi

Mon Dec 3 18:05:51 2018

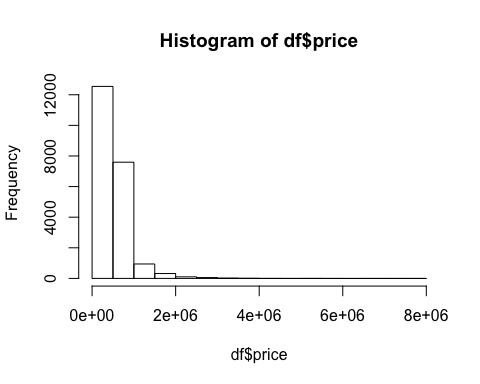
### Model Building  
  
# reading data into R  
setwd("/Users/rakeshravi/Documents/Linear Models - R/Project/")  
df <- read.csv("kc\_house\_data.csv", stringsAsFactors = FALSE)  
  
# Splitting into test and training samples  
set.seed(999)  
index<-sample(nrow(df),0.70\*nrow(df),replace=F)  
train<-df[index,]  
test<-df[-index,]  
  
#running a basic model;  
modview <- lm(price~.,data=train)  
summary(modview)

##   
## Call:  
## lm(formula = price ~ ., data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1246915 -97894 -9122 75913 4165633   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.603e+06 3.402e+06 1.353 0.176080   
## id -1.679e-06 5.610e-07 -2.993 0.002766 \*\*   
## date1/12/15 9.695e+04 1.967e+05 0.493 0.622038   
## date1/13/15 1.054e+05 1.967e+05 0.536 0.591967   
## date1/14/15 7.358e+04 1.958e+05 0.376 0.707114   
## date1/15/15 1.091e+05 1.960e+05 0.557 0.577817   
## date1/16/15 6.868e+04 1.960e+05 0.350 0.726040   
## date1/17/15 1.113e+05 2.736e+05 0.407 0.684288   
## date1/19/15 5.080e+04 2.068e+05 0.246 0.805977   
## date1/2/15 1.006e+05 1.962e+05 0.513 0.608083   
## date1/20/15 1.164e+05 1.962e+05 0.593 0.553032   
## date1/21/15 1.286e+05 1.958e+05 0.657 0.511349   
## date1/22/15 1.185e+05 1.969e+05 0.602 0.547316   
## date1/23/15 1.539e+05 1.966e+05 0.783 0.433692   
## date1/24/15 -8.757e+04 2.370e+05 -0.370 0.711719   
## date1/25/15 1.135e+05 2.369e+05 0.479 0.631928   
## date1/26/15 3.685e+04 1.966e+05 0.187 0.851306   
## date1/27/15 9.142e+04 1.963e+05 0.466 0.641439   
## date1/28/15 9.749e+04 1.958e+05 0.498 0.618593   
## date1/29/15 1.325e+05 1.966e+05 0.674 0.500239   
## date1/30/15 2.463e+05 1.976e+05 1.246 0.212686   
## date1/31/15 -3.071e+04 2.737e+05 -0.112 0.910677   
## date1/5/15 1.043e+05 1.959e+05 0.533 0.594364   
## date1/6/15 1.500e+05 1.968e+05 0.763 0.445772   
## date1/7/15 1.591e+05 1.958e+05 0.813 0.416309   
## date1/8/15 6.162e+04 1.963e+05 0.314 0.753600   
## date1/9/15 1.034e+05 1.973e+05 0.524 0.600189   
## date10/1/14 9.740e+04 1.951e+05 0.499 0.617677   
## date10/10/14 1.122e+05 1.952e+05 0.575 0.565322   
## date10/11/14 1.144e+06 2.371e+05 4.826 1.40e-06 \*\*\*  
## date10/12/14 6.973e+04 2.369e+05 0.294 0.768532   
## date10/13/14 6.987e+04 1.959e+05 0.357 0.721428   
## date10/14/14 9.826e+04 1.950e+05 0.504 0.614431   
## date10/15/14 1.187e+05 1.950e+05 0.608 0.542897   
## date10/16/14 1.239e+05 1.949e+05 0.636 0.524940   
## date10/17/14 1.327e+05 1.955e+05 0.679 0.497266   
## date10/18/14 2.673e+05 2.234e+05 1.197 0.231439   
## date10/19/14 8.650e+04 2.234e+05 0.387 0.698637   
## date10/2/14 1.042e+05 1.953e+05 0.533 0.593715   
## date10/20/14 1.751e+05 1.951e+05 0.898 0.369350   
## date10/21/14 8.493e+04 1.949e+05 0.436 0.663021   
## date10/22/14 9.957e+04 1.954e+05 0.509 0.610412   
## date10/23/14 5.956e+04 1.952e+05 0.305 0.760292   
## date10/24/14 7.967e+04 1.958e+05 0.407 0.684120   
## date10/25/14 1.589e+05 2.119e+05 0.750 0.453478   
## date10/26/14 2.519e+04 2.163e+05 0.116 0.907302   
## date10/27/14 7.733e+04 1.948e+05 0.397 0.691347   
## date10/28/14 1.205e+05 1.946e+05 0.619 0.535684   
## date10/29/14 1.142e+05 1.950e+05 0.586 0.558022   
## date10/3/14 1.136e+05 1.956e+05 0.581 0.561355   
## date10/30/14 1.550e+05 1.950e+05 0.795 0.426650   
## date10/31/14 9.349e+04 1.967e+05 0.475 0.634526   
## date10/4/14 6.030e+04 2.370e+05 0.254 0.799156   
## date10/6/14 5.883e+04 1.952e+05 0.301 0.763168   
## date10/7/14 9.164e+04 1.950e+05 0.470 0.638415   
## date10/8/14 8.605e+04 1.954e+05 0.440 0.659675   
## date10/9/14 9.538e+04 1.951e+05 0.489 0.624938   
## date11/1/14 2.280e+05 2.119e+05 1.076 0.282125   
## date11/10/14 1.249e+05 1.951e+05 0.640 0.522253   
## date11/11/14 8.422e+04 1.962e+05 0.429 0.667741   
## date11/12/14 9.252e+04 1.953e+05 0.474 0.635708   
## date11/13/14 8.619e+04 1.949e+05 0.442 0.658335   
## date11/14/14 1.141e+05 1.953e+05 0.584 0.559169   
## date11/15/14 2.069e+05 2.370e+05 0.873 0.382664   
## date11/16/14 -4.284e+04 2.370e+05 -0.181 0.856547   
## date11/17/14 5.916e+04 1.954e+05 0.303 0.762079   
## date11/18/14 8.186e+04 1.949e+05 0.420 0.674499   
## date11/19/14 1.061e+05 1.951e+05 0.544 0.586468   
## date11/2/14 8.420e+04 2.736e+05 0.308 0.758315   
## date11/20/14 1.147e+05 1.951e+05 0.588 0.556576   
## date11/21/14 6.024e+04 1.952e+05 0.309 0.757576   
## date11/22/14 1.357e+05 2.090e+05 0.650 0.516016   
## date11/23/14 1.153e+05 2.163e+05 0.533 0.593926   
## date11/24/14 1.205e+05 1.953e+05 0.617 0.537319   
## date11/25/14 1.215e+05 1.952e+05 0.622 0.533709   
## date11/26/14 1.046e+05 1.963e+05 0.533 0.594082   
## date11/28/14 2.042e+05 2.370e+05 0.862 0.388830   
## date11/29/14 8.503e+04 2.234e+05 0.381 0.703475   
## date11/3/14 1.128e+05 1.953e+05 0.577 0.563657   
## date11/4/14 7.715e+04 1.954e+05 0.395 0.692921   
## date11/5/14 8.930e+04 1.951e+05 0.458 0.647224   
## date11/6/14 1.017e+05 1.952e+05 0.521 0.602272   
## date11/7/14 8.007e+04 1.955e+05 0.410 0.682076   
## date11/8/14 1.493e+05 2.234e+05 0.668 0.504037   
## date11/9/14 1.779e+04 2.370e+05 0.075 0.940151   
## date12/1/14 1.089e+05 1.948e+05 0.559 0.575958   
## date12/10/14 9.158e+04 1.952e+05 0.469 0.639051   
## date12/11/14 1.071e+05 1.954e+05 0.548 0.583577   
## date12/12/14 7.656e+04 1.956e+05 0.391 0.695524   
## date12/13/14 2.395e+05 2.741e+05 0.874 0.382234   
## date12/14/14 4.138e+04 2.234e+05 0.185 0.853051   
## date12/15/14 1.011e+05 1.951e+05 0.518 0.604424   
## date12/16/14 5.537e+04 1.958e+05 0.283 0.777322   
## date12/17/14 9.615e+04 1.961e+05 0.490 0.623994   
## date12/18/14 7.300e+04 1.953e+05 0.374 0.708544   
## date12/19/14 1.804e+05 1.967e+05 0.917 0.358921   
## date12/2/14 1.085e+05 1.949e+05 0.556 0.577878   
## date12/20/14 1.975e+04 2.052e+05 0.096 0.923330   
## date12/21/14 1.300e+05 2.736e+05 0.475 0.634721   
## date12/22/14 6.593e+04 1.955e+05 0.337 0.735897   
## date12/23/14 8.766e+04 1.958e+05 0.448 0.654307   
## date12/24/14 1.138e+05 1.978e+05 0.575 0.565145   
## date12/26/14 5.764e+04 1.972e+05 0.292 0.770041   
## date12/27/14 3.400e+03 2.737e+05 0.012 0.990087   
## date12/29/14 5.873e+04 1.958e+05 0.300 0.764173   
## date12/3/14 9.620e+04 1.951e+05 0.493 0.621895   
## date12/30/14 9.983e+04 1.965e+05 0.508 0.611374   
## date12/31/14 1.262e+05 1.962e+05 0.643 0.520208   
## date12/4/14 8.246e+04 1.961e+05 0.420 0.674193   
## date12/5/14 1.450e+05 1.959e+05 0.740 0.459132   
## date12/6/14 6.931e+04 2.090e+05 0.332 0.740150   
## date12/7/14 7.930e+03 2.370e+05 0.033 0.973306   
## date12/8/14 1.167e+05 1.952e+05 0.598 0.549816   
## date12/9/14 1.060e+05 1.952e+05 0.543 0.587086   
## date2/1/15 1.458e+05 2.736e+05 0.533 0.594210   
## date2/10/15 1.105e+05 1.961e+05 0.563 0.573130   
## date2/11/15 1.033e+05 1.961e+05 0.527 0.598190   
## date2/12/15 2.085e+05 1.961e+05 1.064 0.287477   
## date2/13/15 1.332e+05 1.955e+05 0.681 0.495689   
## date2/14/15 -1.110e+05 2.737e+05 -0.406 0.684939   
## date2/15/15 1.457e+05 2.736e+05 0.533 0.594296   
## date2/16/15 2.794e+04 2.369e+05 0.118 0.906123   
## date2/17/15 8.381e+04 1.954e+05 0.429 0.667924   
## date2/18/15 9.158e+04 1.950e+05 0.470 0.638692   
## date2/19/15 1.218e+05 1.953e+05 0.624 0.532746   
## date2/2/15 8.926e+04 1.961e+05 0.455 0.649047   
## date2/20/15 1.280e+05 1.956e+05 0.655 0.512760   
## date2/21/15 2.915e+05 2.370e+05 1.230 0.218633   
## date2/22/15 9.535e+04 2.029e+05 0.470 0.638408   
## date2/23/15 1.262e+05 1.956e+05 0.646 0.518591   
## date2/24/15 1.008e+05 1.948e+05 0.517 0.604963   
## date2/25/15 1.277e+05 1.947e+05 0.656 0.511831   
## date2/26/15 1.179e+05 1.958e+05 0.602 0.547055   
## date2/27/15 1.918e+05 1.964e+05 0.977 0.328646   
## date2/28/15 1.677e+05 2.119e+05 0.791 0.428723   
## date2/3/15 7.641e+04 1.965e+05 0.389 0.697322   
## date2/4/15 1.318e+05 1.961e+05 0.672 0.501549   
## date2/5/15 1.003e+05 1.962e+05 0.511 0.609272   
## date2/6/15 7.458e+04 1.962e+05 0.380 0.703883   
## date2/7/15 1.458e+05 2.370e+05 0.615 0.538355   
## date2/9/15 1.116e+05 1.957e+05 0.570 0.568467   
## date3/1/15 1.858e+05 2.164e+05 0.859 0.390574   
## date3/10/15 1.308e+05 1.954e+05 0.670 0.503152   
## date3/11/15 1.145e+05 1.951e+05 0.587 0.557052   
## date3/12/15 1.053e+05 1.953e+05 0.539 0.589868   
## date3/13/15 1.498e+05 1.954e+05 0.767 0.443089   
## date3/14/15 1.271e+05 2.090e+05 0.608 0.543028   
## date3/15/15 2.180e+05 2.369e+05 0.920 0.357682   
## date3/16/15 1.668e+05 1.950e+05 0.855 0.392450   
## date3/17/15 1.499e+05 1.949e+05 0.769 0.441993   
## date3/18/15 8.077e+04 1.950e+05 0.414 0.678775   
## date3/19/15 1.148e+05 1.952e+05 0.588 0.556457   
## date3/2/15 2.103e+05 1.968e+05 1.069 0.285134   
## date3/20/15 1.021e+05 1.957e+05 0.522 0.601871   
## date3/21/15 1.284e+05 2.021e+05 0.635 0.525139   
## date3/22/15 2.398e+05 2.163e+05 1.109 0.267658   
## date3/23/15 1.296e+05 1.949e+05 0.665 0.506089   
## date3/24/15 1.353e+05 1.948e+05 0.695 0.487264   
## date3/25/15 1.262e+05 1.947e+05 0.648 0.516705   
## date3/26/15 1.503e+05 1.948e+05 0.772 0.440244   
## date3/27/15 1.534e+05 1.948e+05 0.788 0.430809   
## date3/28/15 1.902e+05 2.163e+05 0.879 0.379148   
## date3/29/15 7.028e+04 2.069e+05 0.340 0.734058   
## date3/3/15 2.093e+05 1.956e+05 1.070 0.284607   
## date3/30/15 1.490e+05 1.950e+05 0.764 0.445067   
## date3/31/15 1.738e+05 1.953e+05 0.890 0.373566   
## date3/4/15 1.600e+05 1.950e+05 0.821 0.411830   
## date3/5/15 1.129e+05 1.950e+05 0.579 0.562621   
## date3/6/15 1.672e+05 1.961e+05 0.853 0.393709   
## date3/7/15 3.680e+04 2.234e+05 0.165 0.869153   
## date3/9/15 1.314e+05 1.958e+05 0.671 0.502313   
## date4/1/15 1.615e+05 1.950e+05 0.828 0.407584   
## date4/10/15 1.764e+05 1.955e+05 0.902 0.367007   
## date4/11/15 1.513e+05 2.052e+05 0.737 0.460967   
## date4/12/15 2.002e+04 2.040e+05 0.098 0.921814   
## date4/13/15 1.456e+05 1.950e+05 0.746 0.455525   
## date4/14/15 1.287e+05 1.947e+05 0.661 0.508382   
## date4/15/15 1.210e+05 1.953e+05 0.620 0.535465   
## date4/16/15 1.280e+05 1.953e+05 0.655 0.512211   
## date4/17/15 1.121e+05 1.950e+05 0.575 0.565289   
## date4/18/15 1.610e+05 2.120e+05 0.759 0.447625   
## date4/19/15 8.473e+04 2.234e+05 0.379 0.704541   
## date4/2/15 1.446e+05 1.949e+05 0.742 0.458328   
## date4/20/15 1.066e+05 1.954e+05 0.545 0.585478   
## date4/21/15 1.293e+05 1.947e+05 0.664 0.506554   
## date4/22/15 1.194e+05 1.946e+05 0.613 0.539601   
## date4/23/15 1.506e+05 1.947e+05 0.774 0.439235   
## date4/24/15 1.482e+05 1.948e+05 0.761 0.446768   
## date4/25/15 2.447e+05 2.021e+05 1.211 0.225938   
## date4/26/15 8.073e+04 2.039e+05 0.396 0.692198   
## date4/27/15 1.006e+05 1.946e+05 0.517 0.605085   
## date4/28/15 1.350e+05 1.946e+05 0.694 0.487866   
## date4/29/15 1.634e+05 1.947e+05 0.839 0.401537   
## date4/3/15 1.548e+05 1.951e+05 0.794 0.427363   
## date4/30/15 1.239e+05 1.952e+05 0.635 0.525637   
## date4/4/15 2.215e+05 2.163e+05 1.024 0.305893   
## date4/5/15 1.959e+05 2.163e+05 0.906 0.365160   
## date4/6/15 1.223e+05 1.950e+05 0.627 0.530471   
## date4/7/15 1.890e+05 1.949e+05 0.970 0.332169   
## date4/8/15 1.826e+05 1.947e+05 0.938 0.348237   
## date4/9/15 1.275e+05 1.950e+05 0.654 0.513050   
## date5/1/15 1.747e+05 1.954e+05 0.894 0.371196   
## date5/10/14 1.253e+05 2.234e+05 0.561 0.574951   
## date5/10/15 9.270e+03 2.369e+05 0.039 0.968793   
## date5/11/14 5.837e+04 2.737e+05 0.213 0.831133   
## date5/11/15 2.109e+05 1.970e+05 1.070 0.284501   
## date5/12/14 1.290e+05 1.951e+05 0.661 0.508677   
## date5/12/15 1.968e+05 1.963e+05 1.003 0.316000   
## date5/13/14 1.211e+05 1.951e+05 0.621 0.534876   
## date5/13/15 9.056e+04 1.983e+05 0.457 0.647827   
## date5/14/14 1.005e+05 1.952e+05 0.515 0.606518   
## date5/14/15 3.058e+05 2.068e+05 1.479 0.139244   
## date5/15/14 8.508e+04 1.950e+05 0.436 0.662657   
## date5/15/15 1.314e+05 2.736e+05 0.480 0.630960   
## date5/16/14 9.328e+04 1.956e+05 0.477 0.633380   
## date5/18/14 1.081e+05 2.163e+05 0.500 0.617174   
## date5/19/14 8.964e+04 1.952e+05 0.459 0.646025   
## date5/2/14 8.071e+04 1.954e+05 0.413 0.679573   
## date5/2/15 8.854e+04 2.163e+05 0.409 0.682325   
## date5/20/14 9.734e+04 1.946e+05 0.500 0.616935   
## date5/21/14 1.124e+05 1.948e+05 0.577 0.563977   
## date5/22/14 9.675e+04 1.951e+05 0.496 0.619907   
## date5/23/14 9.399e+04 1.953e+05 0.481 0.630334   
## date5/24/14 1.576e+05 2.068e+05 0.762 0.446200   
## date5/24/15 1.962e+05 2.737e+05 0.717 0.473411   
## date5/25/14 -9.967e+02 2.119e+05 -0.005 0.996248   
## date5/26/14 6.103e+04 2.090e+05 0.292 0.770252   
## date5/27/14 1.012e+05 1.947e+05 0.520 0.603322   
## date5/27/15 4.069e+05 2.737e+05 1.487 0.137154   
## date5/28/14 1.046e+05 1.946e+05 0.537 0.591094   
## date5/29/14 1.112e+05 1.953e+05 0.570 0.568934   
## date5/3/14 3.296e+05 2.370e+05 1.391 0.164353   
## date5/3/15 1.696e+05 2.053e+05 0.826 0.408618   
## date5/30/14 1.383e+05 1.956e+05 0.707 0.479354   
## date5/31/14 7.780e+04 2.234e+05 0.348 0.727643   
## date5/4/14 4.451e+04 2.163e+05 0.206 0.836981   
## date5/4/15 1.542e+05 1.949e+05 0.791 0.428784   
## date5/5/14 9.637e+04 1.953e+05 0.494 0.621616   
## date5/5/15 1.591e+05 1.950e+05 0.816 0.414323   
## date5/6/14 8.234e+04 1.951e+05 0.422 0.672922   
## date5/6/15 1.243e+05 1.950e+05 0.637 0.523982   
## date5/7/14 9.837e+04 1.951e+05 0.504 0.614146   
## date5/7/15 1.409e+05 1.954e+05 0.721 0.470809   
## date5/8/14 1.242e+05 1.952e+05 0.636 0.524579   
## date5/8/15 1.817e+05 1.963e+05 0.926 0.354598   
## date5/9/14 7.825e+04 1.953e+05 0.401 0.688715   
## date5/9/15 2.427e+05 2.370e+05 1.024 0.305857   
## date6/1/14 1.738e+05 2.163e+05 0.803 0.421712   
## date6/10/14 1.227e+05 1.950e+05 0.629 0.529086   
## date6/11/14 8.599e+04 1.949e+05 0.441 0.659120   
## date6/12/14 1.422e+05 1.956e+05 0.727 0.467202   
## date6/13/14 1.228e+05 1.954e+05 0.629 0.529670   
## date6/14/14 9.072e+04 2.068e+05 0.439 0.660966   
## date6/15/14 1.685e+05 2.120e+05 0.795 0.426674   
## date6/16/14 1.080e+05 1.949e+05 0.554 0.579343   
## date6/17/14 1.245e+05 1.946e+05 0.639 0.522518   
## date6/18/14 1.403e+05 1.948e+05 0.720 0.471344   
## date6/19/14 1.123e+05 1.948e+05 0.576 0.564400   
## date6/2/14 1.077e+05 1.952e+05 0.552 0.581103   
## date6/20/14 1.163e+05 1.948e+05 0.597 0.550378   
## date6/21/14 1.659e+05 2.069e+05 0.802 0.422475   
## date6/22/14 7.897e+04 2.052e+05 0.385 0.700371   
## date6/23/14 8.720e+04 1.944e+05 0.448 0.653822   
## date6/24/14 9.998e+04 1.946e+05 0.514 0.607437   
## date6/25/14 1.224e+05 1.945e+05 0.629 0.529347   
## date6/26/14 6.969e+04 1.945e+05 0.358 0.720060   
## date6/27/14 6.347e+04 1.951e+05 0.325 0.744924   
## date6/28/14 1.520e+05 2.021e+05 0.752 0.452078   
## date6/29/14 3.405e+04 2.119e+05 0.161 0.872361   
## date6/3/14 1.176e+05 1.949e+05 0.604 0.546133   
## date6/30/14 1.292e+05 1.951e+05 0.662 0.507835   
## date6/4/14 1.488e+05 1.948e+05 0.764 0.444830   
## date6/5/14 1.059e+05 1.950e+05 0.543 0.587046   
## date6/6/14 8.709e+04 1.954e+05 0.446 0.655894   
## date6/7/14 -1.112e+05 2.370e+05 -0.469 0.638914   
## date6/8/14 1.267e+05 2.029e+05 0.624 0.532336   
## date6/9/14 1.061e+05 1.950e+05 0.544 0.586136   
## date7/1/14 1.341e+05 1.947e+05 0.689 0.490903   
## date7/10/14 1.251e+05 1.949e+05 0.642 0.521089   
## date7/11/14 1.133e+05 1.951e+05 0.581 0.561328   
## date7/12/14 1.036e+05 2.090e+05 0.496 0.619939   
## date7/13/14 9.671e+04 2.369e+05 0.408 0.683167   
## date7/14/14 1.198e+05 1.946e+05 0.616 0.538168   
## date7/15/14 1.014e+05 1.946e+05 0.521 0.602439   
## date7/16/14 9.949e+04 1.946e+05 0.511 0.609197   
## date7/17/14 8.700e+04 1.951e+05 0.446 0.655582   
## date7/18/14 1.018e+05 1.949e+05 0.522 0.601587   
## date7/19/14 1.647e+05 2.119e+05 0.777 0.436982   
## date7/2/14 1.680e+05 1.950e+05 0.861 0.389019   
## date7/20/14 3.408e+04 2.069e+05 0.165 0.869161   
## date7/21/14 1.066e+05 1.949e+05 0.547 0.584321   
## date7/22/14 4.117e+04 1.947e+05 0.211 0.832539   
## date7/23/14 8.114e+04 1.947e+05 0.417 0.676913   
## date7/24/14 8.544e+04 1.948e+05 0.439 0.660975   
## date7/25/14 1.480e+05 1.948e+05 0.759 0.447573   
## date7/26/14 1.161e+05 2.120e+05 0.548 0.583915   
## date7/27/14 5.830e+04 2.737e+05 0.213 0.831294   
## date7/28/14 7.110e+04 1.949e+05 0.365 0.715290   
## date7/29/14 8.803e+04 1.948e+05 0.452 0.651293   
## date7/3/14 1.351e+05 1.954e+05 0.691 0.489309   
## date7/30/14 1.231e+05 1.951e+05 0.631 0.528122   
## date7/31/14 1.290e+05 1.949e+05 0.662 0.508172   
## date7/5/14 3.582e+05 2.090e+05 1.714 0.086542 .   
## date7/6/14 1.754e+05 2.739e+05 0.641 0.521839   
## date7/7/14 9.060e+04 1.953e+05 0.464 0.642685   
## date7/8/14 1.050e+05 1.946e+05 0.540 0.589348   
## date7/9/14 9.533e+04 1.946e+05 0.490 0.624220   
## date8/1/14 1.375e+05 1.951e+05 0.705 0.480991   
## date8/10/14 1.617e+05 2.370e+05 0.682 0.495100   
## date8/11/14 8.255e+04 1.948e+05 0.424 0.671716   
## date8/12/14 1.109e+05 1.947e+05 0.570 0.568990   
## date8/13/14 1.092e+05 1.949e+05 0.560 0.575318   
## date8/14/14 7.148e+04 1.949e+05 0.367 0.713774   
## date8/15/14 1.856e+05 1.956e+05 0.949 0.342707   
## date8/16/14 5.198e+04 2.370e+05 0.219 0.826393   
## date8/17/14 -1.875e+04 2.163e+05 -0.087 0.930925   
## date8/18/14 8.859e+04 1.951e+05 0.454 0.649846   
## date8/19/14 1.159e+05 1.950e+05 0.594 0.552395   
## date8/2/14 1.425e+05 2.163e+05 0.659 0.509941   
## date8/20/14 8.165e+04 1.947e+05 0.419 0.674986   
## date8/21/14 9.003e+04 1.953e+05 0.461 0.644770   
## date8/22/14 7.884e+04 1.949e+05 0.404 0.685865   
## date8/23/14 7.643e+04 2.234e+05 0.342 0.732291   
## date8/25/14 9.906e+04 1.946e+05 0.509 0.610829   
## date8/26/14 8.172e+04 1.946e+05 0.420 0.674598   
## date8/27/14 7.523e+04 1.948e+05 0.386 0.699428   
## date8/28/14 1.270e+05 1.957e+05 0.649 0.516175   
## date8/29/14 9.501e+04 1.964e+05 0.484 0.628546   
## date8/3/14 2.099e+05 2.736e+05 0.767 0.442996   
## date8/30/14 3.388e+05 2.736e+05 1.238 0.215648   
## date8/31/14 8.831e+04 2.234e+05 0.395 0.692637   
## date8/4/14 1.098e+05 1.951e+05 0.563 0.573460   
## date8/5/14 1.096e+05 1.947e+05 0.563 0.573501   
## date8/6/14 5.585e+04 1.952e+05 0.286 0.774757   
## date8/7/14 1.483e+05 1.956e+05 0.758 0.448525   
## date8/8/14 1.309e+05 1.953e+05 0.670 0.502601   
## date8/9/14 2.663e+04 2.736e+05 0.097 0.922477   
## date9/1/14 1.302e+05 2.120e+05 0.614 0.538992   
## date9/10/14 8.058e+04 1.952e+05 0.413 0.679693   
## date9/11/14 1.073e+05 1.952e+05 0.550 0.582314   
## date9/12/14 5.259e+04 1.959e+05 0.268 0.788333   
## date9/13/14 9.660e+04 2.164e+05 0.447 0.655241   
## date9/14/14 1.484e+05 2.369e+05 0.626 0.531080   
## date9/15/14 1.070e+05 1.954e+05 0.547 0.584087   
## date9/16/14 1.158e+05 1.949e+05 0.594 0.552492   
## date9/17/14 7.439e+04 1.951e+05 0.381 0.703006   
## date9/18/14 1.140e+05 1.952e+05 0.584 0.559059   
## date9/19/14 1.612e+05 1.952e+05 0.826 0.408988   
## date9/2/14 8.359e+04 1.952e+05 0.428 0.668537   
## date9/20/14 2.085e+05 2.163e+05 0.964 0.335010   
## date9/21/14 1.117e+05 2.163e+05 0.516 0.605595   
## date9/22/14 1.110e+05 1.950e+05 0.569 0.569192   
## date9/23/14 1.103e+05 1.948e+05 0.566 0.571151   
## date9/24/14 1.041e+05 1.948e+05 0.534 0.593227   
## date9/25/14 9.966e+04 1.950e+05 0.511 0.609290   
## date9/26/14 5.990e+04 1.949e+05 0.307 0.758554   
## date9/27/14 3.452e+04 2.120e+05 0.163 0.870648   
## date9/28/14 7.759e+04 2.736e+05 0.284 0.776718   
## date9/29/14 1.210e+05 1.953e+05 0.620 0.535521   
## date9/3/14 1.052e+05 1.953e+05 0.539 0.590120   
## date9/30/14 1.432e+05 1.963e+05 0.729 0.465734   
## date9/4/14 1.016e+05 1.954e+05 0.520 0.603038   
## date9/5/14 7.752e+04 1.951e+05 0.397 0.691045   
## date9/6/14 7.343e+04 2.163e+05 0.339 0.734285   
## date9/7/14 6.835e+04 2.736e+05 0.250 0.802727   
## date9/8/14 8.974e+04 1.951e+05 0.460 0.645547   
## date9/9/14 8.079e+04 1.951e+05 0.414 0.678783   
## bedrooms -2.813e+04 2.195e+03 -12.816 < 2e-16 \*\*\*  
## bathrooms 3.907e+04 3.799e+03 10.284 < 2e-16 \*\*\*  
## sqft\_living 1.292e+02 5.128e+00 25.194 < 2e-16 \*\*\*  
## sqft\_lot 1.592e-01 5.753e-02 2.767 0.005670 \*\*   
## floors 1.194e+04 4.214e+03 2.833 0.004613 \*\*   
## waterfront 5.387e+05 2.036e+04 26.465 < 2e-16 \*\*\*  
## view 5.403e+04 2.523e+03 21.417 < 2e-16 \*\*\*  
## condition 3.014e+04 2.741e+03 10.998 < 2e-16 \*\*\*  
## grade 1.022e+05 2.523e+03 40.499 < 2e-16 \*\*\*  
## sqft\_above 2.700e+01 5.081e+00 5.314 1.09e-07 \*\*\*  
## sqft\_basement NA NA NA NA   
## yr\_built -2.555e+03 8.465e+01 -30.187 < 2e-16 \*\*\*  
## yr\_renovated 1.428e+01 4.286e+00 3.331 0.000869 \*\*\*  
## zipcode -5.415e+02 3.813e+01 -14.202 < 2e-16 \*\*\*  
## lat 5.972e+05 1.251e+04 47.731 < 2e-16 \*\*\*  
## long -1.987e+05 1.535e+04 -12.946 < 2e-16 \*\*\*  
## sqft\_living15 2.928e+01 4.010e+00 7.302 2.98e-13 \*\*\*  
## sqft\_lot15 -2.870e-01 9.218e-02 -3.113 0.001853 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 193400 on 14733 degrees of freedom  
## Multiple R-squared: 0.7115, Adjusted R-squared: 0.704   
## F-statistic: 94.89 on 383 and 14733 DF, p-value: < 2.2e-16

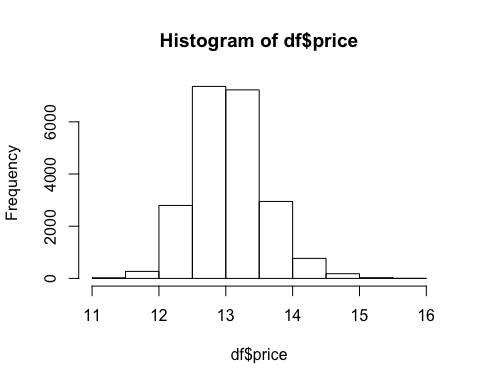
#model contains a number of levels of dates  
#split date into year and month  
#The number of levels in Date are too many as house prices generally vary from month to month, we can use just the month and  
# the year component from the date  
  
#Extracting year and month from the date  
df$year = as.numeric(format(as.Date(df$date, format="%m/%d/%Y"),"%Y"))  
df$month = as.numeric(format(as.Date(df$date, format="%m/%d/%Y"),"%m"))  
df$date = NULL  
  
#running a basic model once again  
modview <- lm(price~.,data=df)  
summary(modview)

##   
## Call:  
## lm(formula = price ~ ., data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1317169 -98523 -8703 77640 4334977   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.165e+06 2.928e+06 2.106 0.03525 \*   
## id -1.344e-06 4.814e-07 -2.792 0.00524 \*\*   
## bedrooms -3.576e+04 1.896e+03 -18.858 < 2e-16 \*\*\*  
## bathrooms 4.145e+04 3.259e+03 12.719 < 2e-16 \*\*\*  
## sqft\_living 1.499e+02 4.385e+00 34.190 < 2e-16 \*\*\*  
## sqft\_lot 1.182e-01 4.786e-02 2.469 0.01356 \*   
## floors 7.173e+03 3.593e+03 1.996 0.04590 \*   
## waterfront 5.833e+05 1.732e+04 33.686 < 2e-16 \*\*\*  
## view 5.260e+04 2.136e+03 24.622 < 2e-16 \*\*\*  
## condition 2.808e+04 2.351e+03 11.943 < 2e-16 \*\*\*  
## grade 9.742e+04 2.156e+03 45.191 < 2e-16 \*\*\*  
## sqft\_above 3.090e+01 4.350e+00 7.104 1.25e-12 \*\*\*  
## sqft\_basement NA NA NA NA   
## yr\_built -2.629e+03 7.253e+01 -36.243 < 2e-16 \*\*\*  
## yr\_renovated 2.072e+01 3.648e+00 5.679 1.37e-08 \*\*\*  
## zipcode -5.814e+02 3.291e+01 -17.669 < 2e-16 \*\*\*  
## lat 6.044e+05 1.071e+04 56.416 < 2e-16 \*\*\*  
## long -2.132e+05 1.314e+04 -16.218 < 2e-16 \*\*\*  
## sqft\_living15 2.124e+01 3.442e+00 6.172 6.87e-10 \*\*\*  
## sqft\_lot15 -4.030e-01 7.333e-02 -5.495 3.94e-08 \*\*\*  
## year 3.742e+04 4.703e+03 7.955 1.88e-15 \*\*\*  
## month 1.290e+03 7.047e+02 1.831 0.06717 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 200700 on 21576 degrees of freedom  
## Multiple R-squared: 0.7018, Adjusted R-squared: 0.7015   
## F-statistic: 2539 on 20 and 21576 DF, p-value: < 2.2e-16

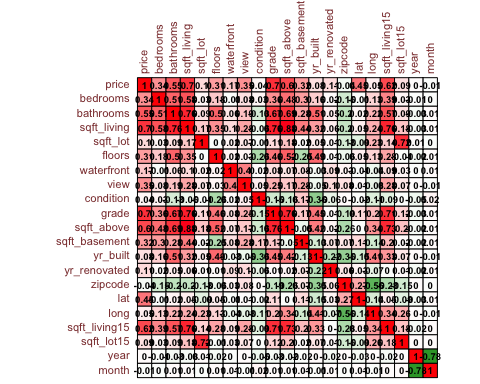
#the number of levels derived from month and year is manageble  
  
#check the distribution of prices (response variables)  
hist(df$price)



#the result is heavily skewed and manjority of the data points lie near the origin. Log transformation might help  
  
df$price =log(df$price)  
hist(df$price)  
#the distribution looks more smooth now  
  
#removing missing values from data  
# Check for NA and missing values  
# is.na return a vector with value T for missing values.  
numberOfNA = length(which(is.na(df) == T))  
if(numberOfNA > 0)  
{  
 cat('Number of missing values: ', numberOfNA)  
 cat('\nRemoving missing values...')  
 df = df[complete.cases(df), ]  
}  
  
# Removing columns id as we do not need it  
df$id = NULL  
  
#Correlation plots  
library(pacman)  
pacman:: p\_load(Metrics, car, corrplot, caTools, ggplot2, DAAG)



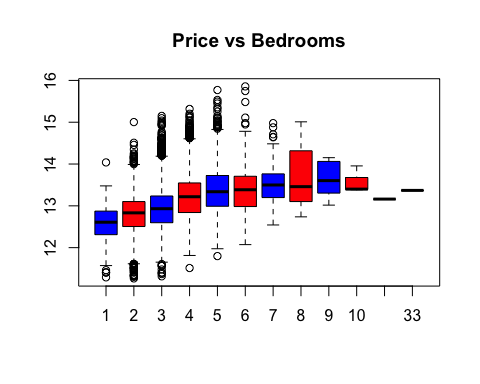
df$date = NULL  
corr = cor(df[, 1:21])  
corrplot(corr, method = "color", outline = T, cl.pos = 'n', rect.col = "black", tl.col = "indianred4", addCoef.col = "black", number.digits = 2, number.cex = 0.60, tl.cex = 0.7, cl.cex = 1, col = colorRampPalette(c("green4","white","red"))(100))



#There are lots of correlations between variables that have over 0.6 as their coefficients  
  
#correlations between response variable and the regressors  
# Correlation between price and sqft\_living is (0.7)  
# Correlation between price and grade is (0.67)  
# Correlation between price and sqft\_above is (0.61)  
# Correlation between price and sqft\_living15 is (0.59)  
# Correlation between price and bathrooms is (0.53)  
  
## date, sqft\_lot, yr\_built, long, sqft\_lot15 happen to have the lowest correlations with the response variable (less than 0.1)  
  
#feature-feature correlation (Potentially multicollinear)  
# Correlation between sqft\_living and sqft\_above is (0.88)  
# Correlation between sqft\_living and sqft\_living15 is (0.76)  
# Correlation between sqft\_living and grade is (0.76)  
# Correlation between sqft living and bathrooms is (0.75)  
  
# Removing columns  
df$date = NULL  
df$sqft\_lot = NULL  
df$sqft\_lot15 = NULL  
df$yr\_built = NULL  
head(df)

## price bedrooms bathrooms sqft\_living floors waterfront view condition  
## 1 12.30998 3 1.00 1180 1 0 0 3  
## 2 13.19561 3 2.25 2570 2 0 0 3  
## 3 12.10071 2 1.00 770 1 0 0 3  
## 4 13.31133 4 3.00 1960 1 0 0 5  
## 5 13.14217 3 2.00 1680 1 0 0 3  
## 6 14.02252 4 4.50 5420 1 0 0 3  
## grade sqft\_above sqft\_basement yr\_renovated zipcode lat long  
## 1 7 1180 0 0 98178 47.5112 -122.257  
## 2 7 2170 400 1991 98125 47.7210 -122.319  
## 3 6 770 0 0 98028 47.7379 -122.233  
## 4 7 1050 910 0 98136 47.5208 -122.393  
## 5 8 1680 0 0 98074 47.6168 -122.045  
## 6 11 3890 1530 0 98053 47.6561 -122.005  
## sqft\_living15 year month  
## 1 1340 14 10  
## 2 1690 14 12  
## 3 2720 15 2  
## 4 1360 14 12  
## 5 1800 15 2  
## 6 4760 14 5

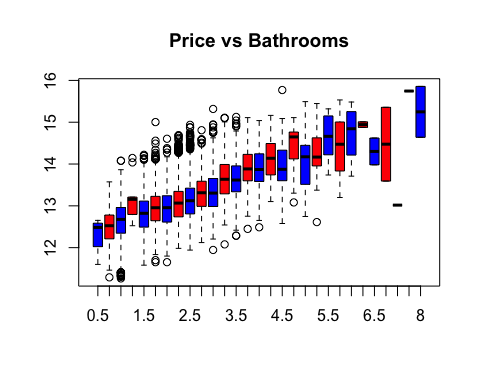
#Box plots  
# Boxplot between price and bedrooms  
boxplot(df[, 1] ~ df[, 2], main = 'Price vs Bedrooms', col=c("blue","red"))



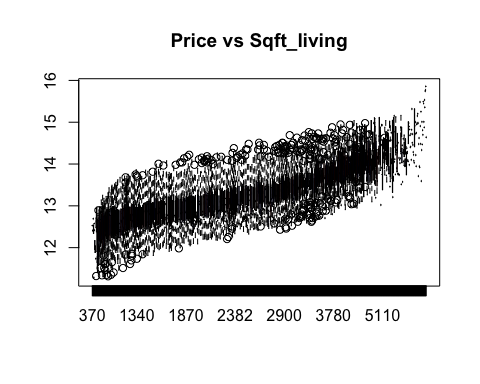
#bedrooms does not seem to have a linear relationship with price  
#for 11 and 33 number of bedrooms, there are hardly any houses.  
  
#Lets examine a subset of the houses that has more than 10 bedrooms  
print(subset(df, df$bedrooms > 10))

## price bedrooms bathrooms sqft\_living floors waterfront view  
## 8749 13.16158 11 3.00 3000 2 0 0  
## 15857 13.36922 33 1.75 1620 1 0 0  
## condition grade sqft\_above sqft\_basement yr\_renovated zipcode  
## 8749 3 7 2400 600 1999 98106  
## 15857 5 7 1040 580 0 98103  
## lat long sqft\_living15 year month  
## 8749 47.5560 -122.363 1420 14 8  
## 15857 47.6878 -122.331 1330 14 6

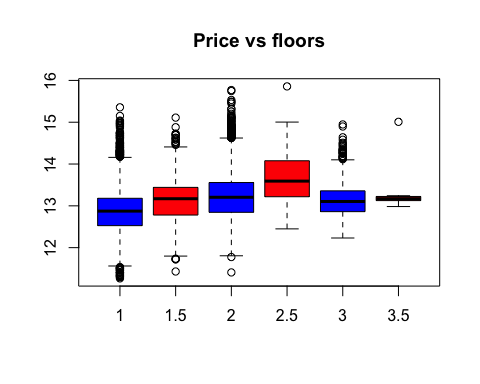
#there are only two houses and it looks like they do not have nearly enough bathrooms for this house to   
#really exist in real life. These looks obvious data entry error and will be removed from the dataset.  
df = df[df$bedrooms <= 10, ]  
  
#because there are only ten levels of bedrooms, it makes sense to convert into a factor without increasing the   
#dimensionality of the dataset.  
df$bedrooms = as.factor(df$bedrooms)  
  
# Boxplot between price and bathrooms  
boxplot(df[, 1] ~ df[, 3], main = 'Price vs Bathrooms', col=c("blue","red"))



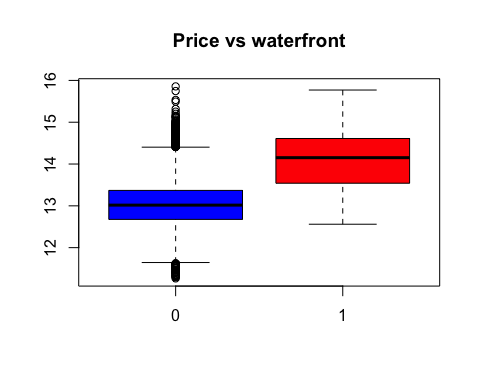
#there seems to be a linear relationship betweem Price and Bathrooms  
  
# Boxplot between price and sqft\_living  
boxplot(df[, 1] ~ df[, 4], main = 'Price vs Sqft\_living', col=c("blue","red"))



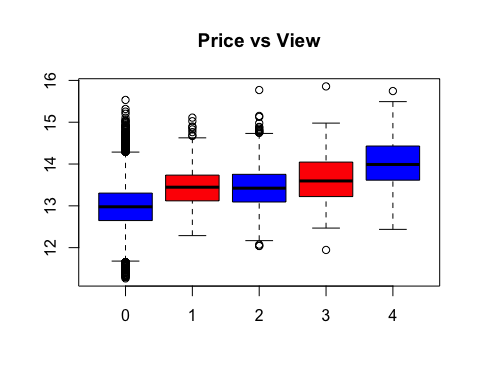
#there is a pseudo-linear relationship  
  
# Boxplot between price and floors  
boxplot(df[, 1] ~ df[, 5], main = 'Price vs floors', col=c("blue","red"))



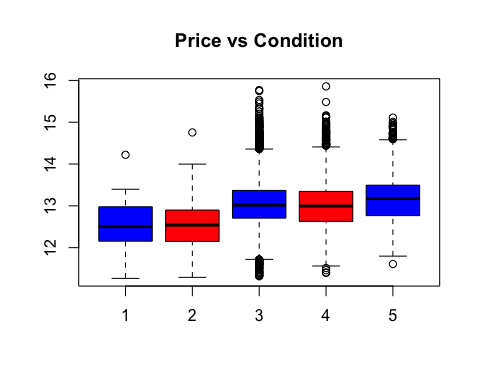
#Floors dont show a pattern and are clearly categorical so converting it to a factor  
  
df$floors = as.factor(df$floors)  
  
# Boxplot between price and waterfront  
boxplot(df[, 1] ~ df[, 6], main = 'Price vs waterfront', col=c("blue","red"))



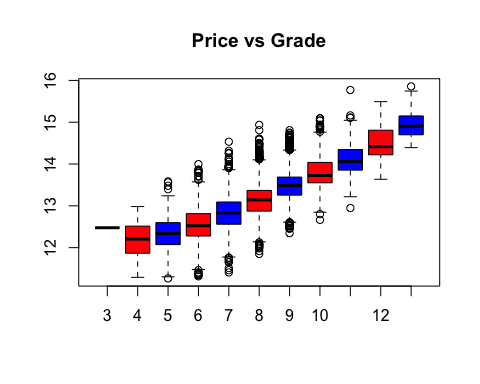
# Waterfront are clearly a categorical value.   
# Converting it into factor  
df$waterfront = as.factor(df$waterfront)  
  
# Boxplot between price and view  
boxplot(df[, 1] ~ df[, 7], main = 'Price vs View', col=c("blue","red"))



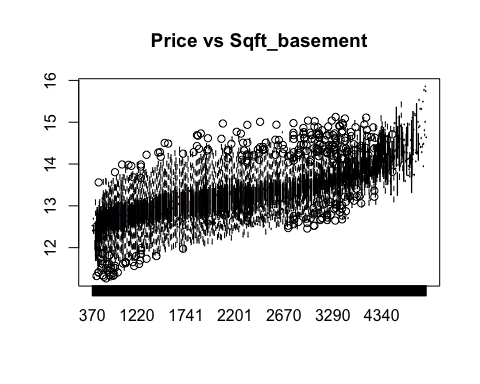
# View is a categorical value.   
# Converting it into factor  
df$view= as.factor(df$view)  
  
# Boxplot between price and condition  
boxplot(df[, 1] ~ df[, 8], main = 'Price vs Condition', col=c("blue","red"))



# Condition is clearly a categorical value.   
# Converting it into factor  
df$condition = as.factor(df$condition)  
  
# Boxplot between price and grade  
boxplot(df[, 1] ~ df[, 9], main = 'Price vs Grade', col=c("blue","red"))



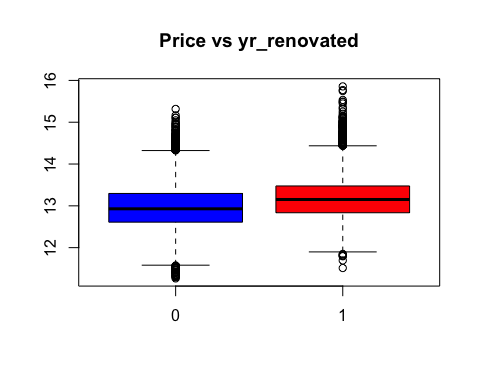
#there is a linear relationship  
  
# Boxplot between price and sqft\_basement  
boxplot(df[, 1] ~ df[, 10], main = 'Price vs Sqft\_basement', col=c("blue","red"))



#there is a linear relationship  
  
  
# Checking the number of houses that do not have a basement  
length(df$sqft\_basement[df$sqft\_basement == 0])

## [1] 13110

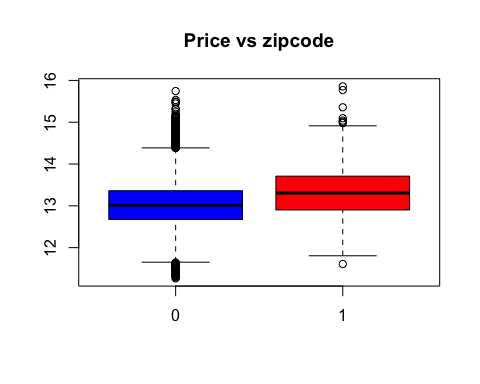
# 13110 houses do not have basement.   
# This can also be converted into a factor with either having a basement or not having a basement  
df$sqft\_basement[df$sqft\_basement != 0] = 1  
# Create factor  
df$sqft\_basement = as.factor(df$sqft\_basement)  
  
# Boxplot between price and yr\_renovated  
boxplot(df[, 1] ~ df[, 11], main = 'Price vs yr\_renovated', col=c("blue","red"))



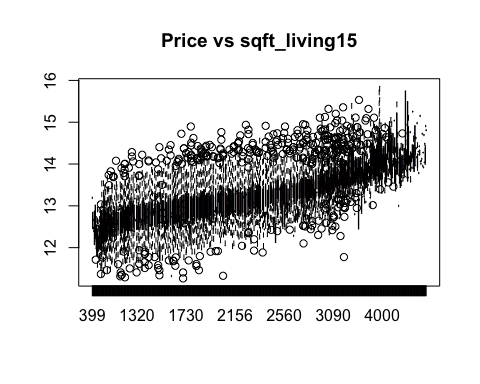
#Lets check how many rows do not have a yr\_renovated  
length(df$yr\_renovated[df$yr\_renovated == 0])

## [1] 20682

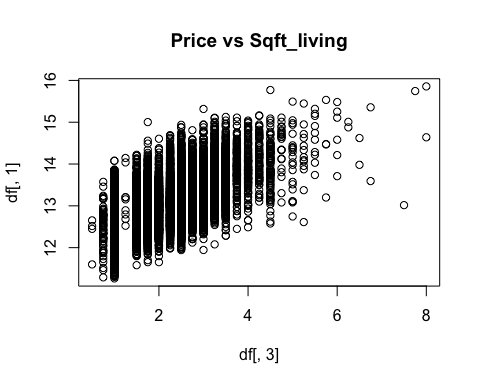
#Over 90% of the houses have not been renovated and we can just convert the years to one to convert this into   
#a factor of a smaller number of levels  
df$yr\_renovated[df$yr\_renovated != 0] = 1  
# Create factor  
df$yr\_renovated = as.factor(df$yr\_renovated)  
  
  
# Boxplot between price and zipcode  
boxplot(df[, 1] ~ df[, 12], main = 'Price vs zipcode', col=c("blue","red"))



# Condition is clearly a categorical value.   
# Converting it into factor  
df$zipcode = as.factor(df$zipcode)  
  
#boxplot between price and sqft\_kliving15  
boxplot(df[, 1] ~ df[, 16], main = 'Price vs sqft\_living15', col=c("blue","red"))



#there is indication of a possible linear relationship  
  
# Plot for Price  
plot(df[, 3], df[, 1], main = 'Price vs Sqft\_living')



# there is a house at the bottom right with a really high value of sqft but low price which seems a bit odd  
  
  
# running a basic model to set benchmark for future improvments on the linear models

# Splitting dataset into training set and test set  
set.seed(123) # Seed initializes the randomness  
sample = sample.split(df, SplitRatio = 0.7) # Returns a vector with T for 70% of data  
trainingSet = subset(df, sample == T)  
testSet = subset(df, sample == F)  
  
# Create model   
model = lm(formula = price ~ bedrooms + bathrooms + floors + waterfront + view + condition +  
 + sqft\_basement + yr\_renovated + zipcode + sqft\_living15 + sqft\_living + grade,  
 data = trainingSet)  
summary(model)

##   
## Call:  
## lm(formula = price ~ bedrooms + bathrooms + floors + waterfront +   
## view + condition + +sqft\_basement + yr\_renovated + zipcode +   
## sqft\_living15 + sqft\_living + grade, data = trainingSet)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.27801 -0.09919 0.00505 0.10511 1.08862   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.089e+01 4.863e-02 223.884 < 2e-16 \*\*\*  
## bedrooms2 4.188e-02 1.739e-02 2.408 0.01604 \*   
## bedrooms3 7.381e-02 1.719e-02 4.294 1.77e-05 \*\*\*  
## bedrooms4 8.120e-02 1.756e-02 4.625 3.78e-06 \*\*\*  
## bedrooms5 3.977e-02 1.856e-02 2.142 0.03219 \*   
## bedrooms6 1.836e-02 2.271e-02 0.808 0.41886   
## bedrooms7 -1.141e-01 4.401e-02 -2.592 0.00955 \*\*   
## bedrooms8 -3.604e-02 7.307e-02 -0.493 0.62184   
## bedrooms9 -1.950e-01 9.616e-02 -2.028 0.04260 \*   
## bedrooms10 -1.869e-01 1.096e-01 -1.706 0.08804 .   
## bathrooms 2.884e-02 3.639e-03 7.925 2.46e-15 \*\*\*  
## floors1.5 2.688e-02 6.100e-03 4.407 1.06e-05 \*\*\*  
## floors2 -1.489e-02 4.815e-03 -3.093 0.00199 \*\*   
## floors2.5 2.400e-02 1.937e-02 1.239 0.21521   
## floors3 -1.254e-01 1.107e-02 -11.326 < 2e-16 \*\*\*  
## floors3.5 -9.068e-02 1.323e-01 -0.685 0.49314   
## waterfront1 4.417e-01 2.323e-02 19.013 < 2e-16 \*\*\*  
## view1 1.212e-01 1.327e-02 9.135 < 2e-16 \*\*\*  
## view2 1.096e-01 7.872e-03 13.928 < 2e-16 \*\*\*  
## view3 1.597e-01 1.072e-02 14.900 < 2e-16 \*\*\*  
## view4 2.701e-01 1.721e-02 15.693 < 2e-16 \*\*\*  
## condition2 1.437e-01 4.641e-02 3.096 0.00197 \*\*   
## condition3 2.806e-01 4.313e-02 6.506 7.99e-11 \*\*\*  
## condition4 3.287e-01 4.316e-02 7.615 2.81e-14 \*\*\*  
## condition5 3.873e-01 4.342e-02 8.921 < 2e-16 \*\*\*  
## sqft\_basement1 -2.346e-02 4.060e-03 -5.779 7.67e-09 \*\*\*  
## yr\_renovated1 8.620e-02 8.005e-03 10.769 < 2e-16 \*\*\*  
## zipcode98002 -3.881e-02 2.008e-02 -1.933 0.05322 .   
## zipcode98003 -7.928e-03 1.775e-02 -0.447 0.65510   
## zipcode98004 1.088e+00 1.764e-02 61.686 < 2e-16 \*\*\*  
## zipcode98005 7.097e-01 2.118e-02 33.512 < 2e-16 \*\*\*  
## zipcode98006 5.972e-01 1.577e-02 37.859 < 2e-16 \*\*\*  
## zipcode98007 6.377e-01 2.309e-02 27.614 < 2e-16 \*\*\*  
## zipcode98008 6.298e-01 1.800e-02 34.997 < 2e-16 \*\*\*  
## zipcode98010 2.647e-01 2.551e-02 10.379 < 2e-16 \*\*\*  
## zipcode98011 4.377e-01 2.050e-02 21.350 < 2e-16 \*\*\*  
## zipcode98014 3.278e-01 2.287e-02 14.335 < 2e-16 \*\*\*  
## zipcode98019 3.315e-01 2.085e-02 15.899 < 2e-16 \*\*\*  
## zipcode98022 7.621e-02 1.942e-02 3.925 8.72e-05 \*\*\*  
## zipcode98023 -4.718e-02 1.561e-02 -3.023 0.00251 \*\*   
## zipcode98024 4.644e-01 2.599e-02 17.871 < 2e-16 \*\*\*  
## zipcode98027 5.008e-01 1.648e-02 30.384 < 2e-16 \*\*\*  
## zipcode98028 4.022e-01 1.796e-02 22.399 < 2e-16 \*\*\*  
## zipcode98029 5.735e-01 1.796e-02 31.929 < 2e-16 \*\*\*  
## zipcode98030 3.506e-02 1.854e-02 1.891 0.05865 .   
## zipcode98031 5.902e-02 1.798e-02 3.283 0.00103 \*\*   
## zipcode98032 -5.399e-02 2.482e-02 -2.176 0.02960 \*   
## zipcode98033 7.544e-01 1.605e-02 46.989 < 2e-16 \*\*\*  
## zipcode98034 5.291e-01 1.551e-02 34.103 < 2e-16 \*\*\*  
## zipcode98038 1.620e-01 1.509e-02 10.737 < 2e-16 \*\*\*  
## zipcode98039 1.222e+00 3.316e-02 36.859 < 2e-16 \*\*\*  
## zipcode98040 8.415e-01 1.836e-02 45.838 < 2e-16 \*\*\*  
## zipcode98042 4.844e-02 1.514e-02 3.200 0.00138 \*\*   
## zipcode98045 3.326e-01 1.917e-02 17.352 < 2e-16 \*\*\*  
## zipcode98052 6.223e-01 1.522e-02 40.897 < 2e-16 \*\*\*  
## zipcode98053 5.796e-01 1.658e-02 34.966 < 2e-16 \*\*\*  
## zipcode98055 1.220e-01 1.764e-02 6.917 4.81e-12 \*\*\*  
## zipcode98056 3.030e-01 1.666e-02 18.191 < 2e-16 \*\*\*  
## zipcode98058 1.509e-01 1.580e-02 9.549 < 2e-16 \*\*\*  
## zipcode98059 3.231e-01 1.590e-02 20.320 < 2e-16 \*\*\*  
## zipcode98065 3.816e-01 1.758e-02 21.712 < 2e-16 \*\*\*  
## zipcode98070 3.472e-01 2.436e-02 14.251 < 2e-16 \*\*\*  
## zipcode98072 4.771e-01 1.808e-02 26.386 < 2e-16 \*\*\*  
## zipcode98074 5.368e-01 1.622e-02 33.093 < 2e-16 \*\*\*  
## zipcode98075 5.305e-01 1.721e-02 30.825 < 2e-16 \*\*\*  
## zipcode98077 4.553e-01 2.020e-02 22.544 < 2e-16 \*\*\*  
## zipcode98092 2.923e-02 1.694e-02 1.726 0.08445 .   
## zipcode98102 9.245e-01 2.675e-02 34.555 < 2e-16 \*\*\*  
## zipcode98103 8.252e-01 1.548e-02 53.299 < 2e-16 \*\*\*  
## zipcode98105 9.168e-01 1.992e-02 46.026 < 2e-16 \*\*\*  
## zipcode98106 3.057e-01 1.751e-02 17.463 < 2e-16 \*\*\*  
## zipcode98107 8.239e-01 1.887e-02 43.659 < 2e-16 \*\*\*  
## zipcode98108 3.282e-01 2.038e-02 16.100 < 2e-16 \*\*\*  
## zipcode98109 9.547e-01 2.406e-02 39.675 < 2e-16 \*\*\*  
## zipcode98112 1.022e+00 1.884e-02 54.240 < 2e-16 \*\*\*  
## zipcode98115 7.905e-01 1.523e-02 51.913 < 2e-16 \*\*\*  
## zipcode98116 7.312e-01 1.748e-02 41.828 < 2e-16 \*\*\*  
## zipcode98117 7.904e-01 1.540e-02 51.331 < 2e-16 \*\*\*  
## zipcode98118 4.210e-01 1.582e-02 26.610 < 2e-16 \*\*\*  
## zipcode98119 9.319e-01 2.140e-02 43.556 < 2e-16 \*\*\*  
## zipcode98122 7.804e-01 1.849e-02 42.211 < 2e-16 \*\*\*  
## zipcode98125 5.418e-01 1.621e-02 33.434 < 2e-16 \*\*\*  
## zipcode98126 5.251e-01 1.719e-02 30.550 < 2e-16 \*\*\*  
## zipcode98133 4.399e-01 1.552e-02 28.347 < 2e-16 \*\*\*  
## zipcode98136 6.537e-01 1.844e-02 35.447 < 2e-16 \*\*\*  
## zipcode98144 6.274e-01 1.712e-02 36.643 < 2e-16 \*\*\*  
## zipcode98146 2.377e-01 1.829e-02 12.993 < 2e-16 \*\*\*  
## zipcode98148 1.125e-01 3.145e-02 3.576 0.00035 \*\*\*  
## zipcode98155 4.125e-01 1.606e-02 25.683 < 2e-16 \*\*\*  
## zipcode98166 2.823e-01 1.851e-02 15.252 < 2e-16 \*\*\*  
## zipcode98168 7.236e-02 1.814e-02 3.989 6.67e-05 \*\*\*  
## zipcode98177 5.614e-01 1.880e-02 29.862 < 2e-16 \*\*\*  
## zipcode98178 1.258e-01 1.827e-02 6.884 6.07e-12 \*\*\*  
## zipcode98188 9.589e-02 2.419e-02 3.964 7.40e-05 \*\*\*  
## zipcode98198 4.114e-02 1.812e-02 2.270 0.02322 \*   
## zipcode98199 8.151e-01 1.774e-02 45.957 < 2e-16 \*\*\*  
## sqft\_living15 8.655e-05 4.100e-06 21.109 < 2e-16 \*\*\*  
## sqft\_living 1.890e-04 3.957e-06 47.759 < 2e-16 \*\*\*  
## grade 9.229e-02 2.522e-03 36.590 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1865 on 14297 degrees of freedom  
## Multiple R-squared: 0.8767, Adjusted R-squared: 0.8759   
## F-statistic: 1038 on 98 and 14297 DF, p-value: < 2.2e-16

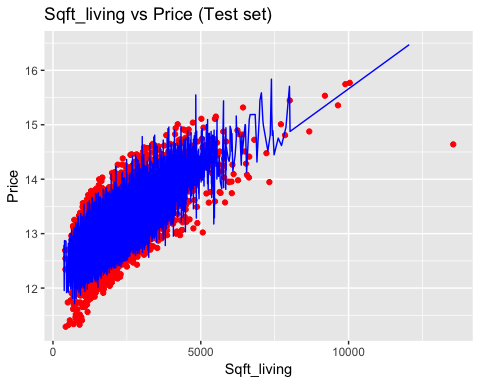
# Bedrooms are not statisticaaly significant, so we should omit it as removing it does not decreases the adjusted  
# r-squared value by a lot  
model = lm(formula = price ~ bathrooms + floors + waterfront + view + condition +  
 + sqft\_basement + yr\_renovated + zipcode + sqft\_living15 + sqft\_living + grade,  
 data = trainingSet)  
summary(model)

##   
## Call:  
## lm(formula = price ~ bathrooms + floors + waterfront + view +   
## condition + +sqft\_basement + yr\_renovated + zipcode + sqft\_living15 +   
## sqft\_living + grade, data = trainingSet)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.27417 -0.09910 0.00636 0.10574 1.06522   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.093e+01 4.627e-02 236.218 < 2e-16 \*\*\*  
## bathrooms 2.820e-02 3.566e-03 7.908 2.80e-15 \*\*\*  
## floors1.5 3.337e-02 6.046e-03 5.519 3.47e-08 \*\*\*  
## floors2 -1.096e-02 4.818e-03 -2.274 0.022961 \*   
## floors2.5 1.829e-02 1.934e-02 0.946 0.344261   
## floors3 -1.202e-01 1.110e-02 -10.830 < 2e-16 \*\*\*  
## floors3.5 -7.410e-02 1.329e-01 -0.557 0.577201   
## waterfront1 4.404e-01 2.332e-02 18.887 < 2e-16 \*\*\*  
## view1 1.197e-01 1.332e-02 8.988 < 2e-16 \*\*\*  
## view2 1.069e-01 7.898e-03 13.535 < 2e-16 \*\*\*  
## view3 1.584e-01 1.075e-02 14.735 < 2e-16 \*\*\*  
## view4 2.628e-01 1.727e-02 15.216 < 2e-16 \*\*\*  
## condition2 1.450e-01 4.662e-02 3.111 0.001871 \*\*   
## condition3 2.850e-01 4.332e-02 6.580 4.86e-11 \*\*\*  
## condition4 3.350e-01 4.335e-02 7.728 1.16e-14 \*\*\*  
## condition5 3.960e-01 4.360e-02 9.083 < 2e-16 \*\*\*  
## sqft\_basement1 -2.008e-02 4.062e-03 -4.944 7.75e-07 \*\*\*  
## yr\_renovated1 8.766e-02 8.033e-03 10.913 < 2e-16 \*\*\*  
## zipcode98002 -4.179e-02 2.016e-02 -2.073 0.038233 \*   
## zipcode98003 -1.024e-02 1.783e-02 -0.574 0.565716   
## zipcode98004 1.080e+00 1.770e-02 61.028 < 2e-16 \*\*\*  
## zipcode98005 7.061e-01 2.127e-02 33.199 < 2e-16 \*\*\*  
## zipcode98006 5.899e-01 1.583e-02 37.267 < 2e-16 \*\*\*  
## zipcode98007 6.258e-01 2.316e-02 27.021 < 2e-16 \*\*\*  
## zipcode98008 6.258e-01 1.806e-02 34.648 < 2e-16 \*\*\*  
## zipcode98010 2.625e-01 2.562e-02 10.246 < 2e-16 \*\*\*  
## zipcode98011 4.319e-01 2.059e-02 20.977 < 2e-16 \*\*\*  
## zipcode98014 3.208e-01 2.293e-02 13.993 < 2e-16 \*\*\*  
## zipcode98019 3.275e-01 2.093e-02 15.653 < 2e-16 \*\*\*  
## zipcode98022 7.303e-02 1.950e-02 3.745 0.000181 \*\*\*  
## zipcode98023 -4.784e-02 1.568e-02 -3.051 0.002282 \*\*   
## zipcode98024 4.611e-01 2.609e-02 17.672 < 2e-16 \*\*\*  
## zipcode98027 4.932e-01 1.653e-02 29.833 < 2e-16 \*\*\*  
## zipcode98028 4.005e-01 1.804e-02 22.203 < 2e-16 \*\*\*  
## zipcode98029 5.707e-01 1.804e-02 31.636 < 2e-16 \*\*\*  
## zipcode98030 3.581e-02 1.862e-02 1.923 0.054502 .   
## zipcode98031 5.926e-02 1.806e-02 3.281 0.001036 \*\*   
## zipcode98032 -5.752e-02 2.493e-02 -2.308 0.021020 \*   
## zipcode98033 7.511e-01 1.612e-02 46.580 < 2e-16 \*\*\*  
## zipcode98034 5.276e-01 1.558e-02 33.858 < 2e-16 \*\*\*  
## zipcode98038 1.626e-01 1.515e-02 10.728 < 2e-16 \*\*\*  
## zipcode98039 1.212e+00 3.328e-02 36.407 < 2e-16 \*\*\*  
## zipcode98040 8.303e-01 1.841e-02 45.087 < 2e-16 \*\*\*  
## zipcode98042 4.821e-02 1.520e-02 3.171 0.001523 \*\*   
## zipcode98045 3.317e-01 1.924e-02 17.235 < 2e-16 \*\*\*  
## zipcode98052 6.179e-01 1.528e-02 40.437 < 2e-16 \*\*\*  
## zipcode98053 5.733e-01 1.659e-02 34.550 < 2e-16 \*\*\*  
## zipcode98055 1.158e-01 1.771e-02 6.541 6.32e-11 \*\*\*  
## zipcode98056 3.000e-01 1.673e-02 17.933 < 2e-16 \*\*\*  
## zipcode98058 1.490e-01 1.587e-02 9.386 < 2e-16 \*\*\*  
## zipcode98059 3.219e-01 1.597e-02 20.152 < 2e-16 \*\*\*  
## zipcode98065 3.789e-01 1.764e-02 21.475 < 2e-16 \*\*\*  
## zipcode98070 3.373e-01 2.444e-02 13.801 < 2e-16 \*\*\*  
## zipcode98072 4.745e-01 1.816e-02 26.133 < 2e-16 \*\*\*  
## zipcode98074 5.331e-01 1.629e-02 32.723 < 2e-16 \*\*\*  
## zipcode98075 5.266e-01 1.728e-02 30.471 < 2e-16 \*\*\*  
## zipcode98077 4.538e-01 2.028e-02 22.373 < 2e-16 \*\*\*  
## zipcode98092 2.911e-02 1.702e-02 1.711 0.087155 .   
## zipcode98102 9.108e-01 2.683e-02 33.944 < 2e-16 \*\*\*  
## zipcode98103 8.113e-01 1.549e-02 52.386 < 2e-16 \*\*\*  
## zipcode98105 8.983e-01 1.989e-02 45.150 < 2e-16 \*\*\*  
## zipcode98106 2.933e-01 1.754e-02 16.716 < 2e-16 \*\*\*  
## zipcode98107 8.080e-01 1.888e-02 42.804 < 2e-16 \*\*\*  
## zipcode98108 3.160e-01 2.044e-02 15.456 < 2e-16 \*\*\*  
## zipcode98109 9.410e-01 2.413e-02 38.991 < 2e-16 \*\*\*  
## zipcode98112 1.008e+00 1.885e-02 53.479 < 2e-16 \*\*\*  
## zipcode98115 7.811e-01 1.526e-02 51.195 < 2e-16 \*\*\*  
## zipcode98116 7.222e-01 1.753e-02 41.197 < 2e-16 \*\*\*  
## zipcode98117 7.764e-01 1.540e-02 50.418 < 2e-16 \*\*\*  
## zipcode98118 4.073e-01 1.584e-02 25.709 < 2e-16 \*\*\*  
## zipcode98119 9.168e-01 2.143e-02 42.786 < 2e-16 \*\*\*  
## zipcode98122 7.648e-01 1.852e-02 41.304 < 2e-16 \*\*\*  
## zipcode98125 5.353e-01 1.626e-02 32.914 < 2e-16 \*\*\*  
## zipcode98126 5.119e-01 1.720e-02 29.761 < 2e-16 \*\*\*  
## zipcode98133 4.314e-01 1.557e-02 27.711 < 2e-16 \*\*\*  
## zipcode98136 6.413e-01 1.847e-02 34.712 < 2e-16 \*\*\*  
## zipcode98144 6.125e-01 1.713e-02 35.759 < 2e-16 \*\*\*  
## zipcode98146 2.301e-01 1.836e-02 12.534 < 2e-16 \*\*\*  
## zipcode98148 1.106e-01 3.159e-02 3.501 0.000465 \*\*\*  
## zipcode98155 4.068e-01 1.613e-02 25.227 < 2e-16 \*\*\*  
## zipcode98166 2.781e-01 1.859e-02 14.959 < 2e-16 \*\*\*  
## zipcode98168 6.408e-02 1.820e-02 3.521 0.000432 \*\*\*  
## zipcode98177 5.563e-01 1.888e-02 29.469 < 2e-16 \*\*\*  
## zipcode98178 1.219e-01 1.834e-02 6.648 3.09e-11 \*\*\*  
## zipcode98188 9.308e-02 2.430e-02 3.831 0.000128 \*\*\*  
## zipcode98198 3.622e-02 1.820e-02 1.990 0.046582 \*   
## zipcode98199 8.044e-01 1.777e-02 45.263 < 2e-16 \*\*\*  
## sqft\_living15 8.791e-05 4.103e-06 21.425 < 2e-16 \*\*\*  
## sqft\_living 1.850e-04 3.719e-06 49.740 < 2e-16 \*\*\*  
## grade 9.606e-02 2.484e-03 38.677 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1873 on 14306 degrees of freedom  
## Multiple R-squared: 0.8755, Adjusted R-squared: 0.8747   
## F-statistic: 1131 on 89 and 14306 DF, p-value: < 2.2e-16

#making predictions  
y\_pred = predict(model, newdata = trainingSet)  
  
  
# Visualizing the training set results  
library(ggplot2)  
ggplot() +  
 geom\_point(aes(x = trainingSet$sqft\_living, y = trainingSet$price),  
 colour = 'red') +  
 geom\_line(aes(x = trainingSet$sqft\_living, y = y\_pred),  
 colour = 'blue') +  
 ggtitle('Sqft\_living vs Price (Training set)') +  
 xlab('Sqft\_living') +  
 ylab('Price')



# Visualizing the test set results  
ggplot() +  
 geom\_point(aes(x = testSet$sqft\_living, y = testSet$price),  
 colour = 'red') +  
 geom\_line(aes(x = trainingSet$sqft\_living, y = y\_pred),  
 colour = 'blue') +  
 ggtitle('Sqft\_living vs Price (Test set)') +  
 xlab('Sqft\_living') +  
 ylab('Price')



# Checking accuracy on test set  
pricePrediction = predict(model, newdata = testSet)  
modelOutput <- cbind(testSet, pricePrediction)  
  
#Test with RMSE  
library(hydroGOF)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

##   
## Attaching package: 'hydroGOF'

## The following objects are masked from 'package:Metrics':  
##   
## mae, mse, rmse

rmse(modelOutput$price, modelOutput$pricePrediction)

## [1] 0.1863146

#Influence Diagnostics and Outlier Treatment  
  
#Perform Influence Diagnostics  
library(olsrr)

##   
## Attaching package: 'olsrr'

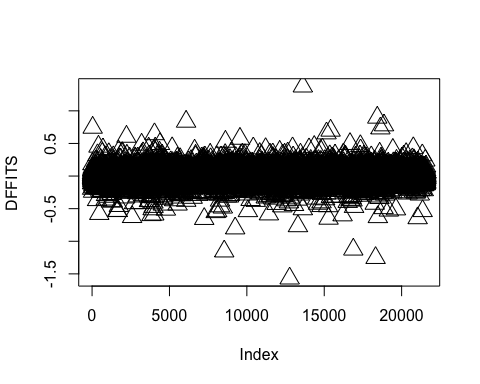
## The following object is masked from 'package:datasets':  
##   
## rivers

df$price = log(df$price)  
model = lm(formula = price ~.,  
 data = df)  
  
summary(model)

##   
## Call:  
## lm(formula = price ~ ., data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.104490 -0.007418 0.000582 0.007927 0.086093   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.701e+00 5.405e-01 -4.997 5.86e-07 \*\*\*  
## bedrooms2 3.240e-03 1.057e-03 3.065 0.002178 \*\*   
## bedrooms3 5.686e-03 1.046e-03 5.435 5.53e-08 \*\*\*  
## bedrooms4 6.448e-03 1.069e-03 6.031 1.66e-09 \*\*\*  
## bedrooms5 3.899e-03 1.132e-03 3.444 0.000574 \*\*\*  
## bedrooms6 1.582e-03 1.393e-03 1.136 0.256033   
## bedrooms7 -9.469e-03 2.576e-03 -3.675 0.000238 \*\*\*  
## bedrooms8 -2.431e-03 4.128e-03 -0.589 0.556021   
## bedrooms9 -1.009e-02 5.947e-03 -1.696 0.089810 .   
## bedrooms10 -9.503e-03 8.284e-03 -1.147 0.251325   
## bathrooms 2.585e-03 2.259e-04 11.443 < 2e-16 \*\*\*  
## sqft\_living 7.490e-06 4.445e-07 16.849 < 2e-16 \*\*\*  
## floors1.5 1.820e-03 3.859e-04 4.715 2.43e-06 \*\*\*  
## floors2 -1.983e-03 3.137e-04 -6.322 2.64e-10 \*\*\*  
## floors2.5 -2.189e-03 1.183e-03 -1.850 0.064325 .   
## floors3 -9.915e-03 6.897e-04 -14.377 < 2e-16 \*\*\*  
## floors3.5 -1.090e-02 5.414e-03 -2.014 0.044055 \*   
## waterfront1 3.166e-02 1.420e-03 22.301 < 2e-16 \*\*\*  
## view1 8.901e-03 8.057e-04 11.047 < 2e-16 \*\*\*  
## view2 8.196e-03 4.922e-04 16.651 < 2e-16 \*\*\*  
## view3 1.307e-02 6.741e-04 19.389 < 2e-16 \*\*\*  
## view4 1.938e-02 1.035e-03 18.723 < 2e-16 \*\*\*  
## condition2 1.000e-02 2.862e-03 3.496 0.000474 \*\*\*  
## condition3 2.006e-02 2.657e-03 7.552 4.46e-14 \*\*\*  
## condition4 2.428e-02 2.659e-03 9.133 < 2e-16 \*\*\*  
## condition5 2.924e-02 2.675e-03 10.930 < 2e-16 \*\*\*  
## grade 6.668e-03 1.576e-04 42.325 < 2e-16 \*\*\*  
## sqft\_above 8.155e-06 4.954e-07 16.463 < 2e-16 \*\*\*  
## sqft\_basement1 3.112e-03 3.804e-04 8.180 3.00e-16 \*\*\*  
## yr\_renovated1 6.863e-03 4.940e-04 13.893 < 2e-16 \*\*\*  
## zipcode98002 -1.794e-03 1.277e-03 -1.404 0.160266   
## zipcode98003 -5.186e-04 1.142e-03 -0.454 0.649906   
## zipcode98004 7.192e-02 2.076e-03 34.634 < 2e-16 \*\*\*  
## zipcode98005 4.726e-02 2.219e-03 21.299 < 2e-16 \*\*\*  
## zipcode98006 4.098e-02 1.814e-03 22.589 < 2e-16 \*\*\*  
## zipcode98007 4.251e-02 2.291e-03 18.559 < 2e-16 \*\*\*  
## zipcode98008 4.224e-02 2.175e-03 19.422 < 2e-16 \*\*\*  
## zipcode98010 2.651e-02 1.947e-03 13.614 < 2e-16 \*\*\*  
## zipcode98011 1.954e-02 2.830e-03 6.905 5.15e-12 \*\*\*  
## zipcode98014 2.467e-02 3.106e-03 7.943 2.07e-15 \*\*\*  
## zipcode98019 1.850e-02 3.067e-03 6.031 1.65e-09 \*\*\*  
## zipcode98022 1.706e-02 1.692e-03 10.081 < 2e-16 \*\*\*  
## zipcode98023 -5.542e-03 1.051e-03 -5.272 1.36e-07 \*\*\*  
## zipcode98024 3.608e-02 2.731e-03 13.211 < 2e-16 \*\*\*  
## zipcode98027 3.750e-02 1.863e-03 20.128 < 2e-16 \*\*\*  
## zipcode98028 1.599e-02 2.748e-03 5.817 6.07e-09 \*\*\*  
## zipcode98029 4.292e-02 2.126e-03 20.194 < 2e-16 \*\*\*  
## zipcode98030 3.500e-03 1.256e-03 2.788 0.005315 \*\*   
## zipcode98031 3.800e-03 1.308e-03 2.905 0.003676 \*\*   
## zipcode98032 -5.985e-03 1.518e-03 -3.942 8.11e-05 \*\*\*  
## zipcode98033 4.679e-02 2.358e-03 19.846 < 2e-16 \*\*\*  
## zipcode98034 2.706e-02 2.528e-03 10.706 < 2e-16 \*\*\*  
## zipcode98038 1.733e-02 1.411e-03 12.280 < 2e-16 \*\*\*  
## zipcode98039 7.775e-02 2.806e-03 27.711 < 2e-16 \*\*\*  
## zipcode98040 5.556e-02 1.837e-03 30.240 < 2e-16 \*\*\*  
## zipcode98042 6.209e-03 1.202e-03 5.163 2.45e-07 \*\*\*  
## zipcode98045 3.472e-02 2.606e-03 13.322 < 2e-16 \*\*\*  
## zipcode98052 3.848e-02 2.407e-03 15.988 < 2e-16 \*\*\*  
## zipcode98053 3.790e-02 2.580e-03 14.691 < 2e-16 \*\*\*  
## zipcode98055 6.392e-03 1.455e-03 4.393 1.12e-05 \*\*\*  
## zipcode98056 1.848e-02 1.583e-03 11.672 < 2e-16 \*\*\*  
## zipcode98058 1.064e-02 1.376e-03 7.733 1.09e-14 \*\*\*  
## zipcode98059 2.201e-02 1.553e-03 14.175 < 2e-16 \*\*\*  
## zipcode98065 3.238e-02 2.400e-03 13.492 < 2e-16 \*\*\*  
## zipcode98070 2.128e-02 1.807e-03 11.778 < 2e-16 \*\*\*  
## zipcode98072 2.493e-02 2.815e-03 8.856 < 2e-16 \*\*\*  
## zipcode98074 3.614e-02 2.278e-03 15.868 < 2e-16 \*\*\*  
## zipcode98075 3.727e-02 2.190e-03 17.017 < 2e-16 \*\*\*  
## zipcode98077 2.438e-02 2.929e-03 8.326 < 2e-16 \*\*\*  
## zipcode98092 5.727e-03 1.142e-03 5.016 5.33e-07 \*\*\*  
## zipcode98102 5.797e-02 2.423e-03 23.926 < 2e-16 \*\*\*  
## zipcode98103 4.751e-02 2.275e-03 20.882 < 2e-16 \*\*\*  
## zipcode98105 5.656e-02 2.333e-03 24.246 < 2e-16 \*\*\*  
## zipcode98106 1.338e-02 1.689e-03 7.920 2.50e-15 \*\*\*  
## zipcode98107 4.800e-02 2.348e-03 20.443 < 2e-16 \*\*\*  
## zipcode98108 1.687e-02 1.862e-03 9.062 < 2e-16 \*\*\*  
## zipcode98109 6.019e-02 2.411e-03 24.966 < 2e-16 \*\*\*  
## zipcode98112 6.458e-02 2.136e-03 30.233 < 2e-16 \*\*\*  
## zipcode98115 4.683e-02 2.312e-03 20.251 < 2e-16 \*\*\*  
## zipcode98116 4.386e-02 1.882e-03 23.302 < 2e-16 \*\*\*  
## zipcode98117 4.466e-02 2.342e-03 19.067 < 2e-16 \*\*\*  
## zipcode98118 2.519e-02 1.642e-03 15.338 < 2e-16 \*\*\*  
## zipcode98119 5.834e-02 2.278e-03 25.614 < 2e-16 \*\*\*  
## zipcode98122 4.885e-02 2.030e-03 24.067 < 2e-16 \*\*\*  
## zipcode98125 2.731e-02 2.501e-03 10.918 < 2e-16 \*\*\*  
## zipcode98126 2.955e-02 1.730e-03 17.083 < 2e-16 \*\*\*  
## zipcode98133 1.692e-02 2.583e-03 6.551 5.84e-11 \*\*\*  
## zipcode98136 3.964e-02 1.774e-03 22.346 < 2e-16 \*\*\*  
## zipcode98144 3.823e-02 1.893e-03 20.195 < 2e-16 \*\*\*  
## zipcode98146 1.049e-02 1.583e-03 6.628 3.49e-11 \*\*\*  
## zipcode98148 5.557e-03 2.155e-03 2.578 0.009936 \*\*   
## zipcode98155 1.477e-02 2.686e-03 5.497 3.90e-08 \*\*\*  
## zipcode98166 1.603e-02 1.450e-03 11.057 < 2e-16 \*\*\*  
## zipcode98168 -3.083e-03 1.531e-03 -2.014 0.044068 \*   
## zipcode98177 2.601e-02 2.696e-03 9.648 < 2e-16 \*\*\*  
## zipcode98178 3.893e-03 1.581e-03 2.463 0.013799 \*   
## zipcode98188 1.228e-03 1.624e-03 0.756 0.449667   
## zipcode98198 -3.367e-04 1.232e-03 -0.273 0.784634   
## zipcode98199 4.807e-02 2.225e-03 21.602 < 2e-16 \*\*\*  
## lat 3.908e-02 5.602e-03 6.976 3.13e-12 \*\*\*  
## long -2.599e-02 4.010e-03 -6.481 9.34e-11 \*\*\*  
## sqft\_living15 6.439e-06 2.565e-07 25.104 < 2e-16 \*\*\*  
## year 5.322e-03 3.333e-04 15.970 < 2e-16 \*\*\*  
## month 2.417e-04 4.992e-05 4.841 1.30e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01418 on 21491 degrees of freedom  
## Multiple R-squared: 0.8755, Adjusted R-squared: 0.8749   
## F-statistic: 1467 on 103 and 21491 DF, p-value: < 2.2e-16

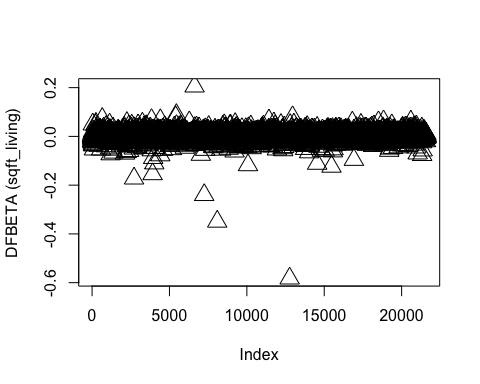
#influence(model)

# The values of DFFITS may be calculated using the function "dffits":   
plot(dffits(model), pch=2, bg='orange', cex=2, ylab="DFFITS")

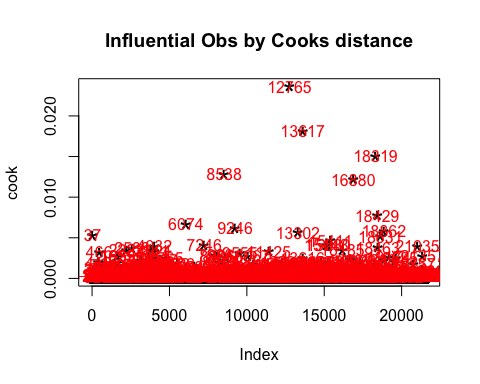


#the plot doesnt show too many outliers and influence points but it does warrant further investigqtion

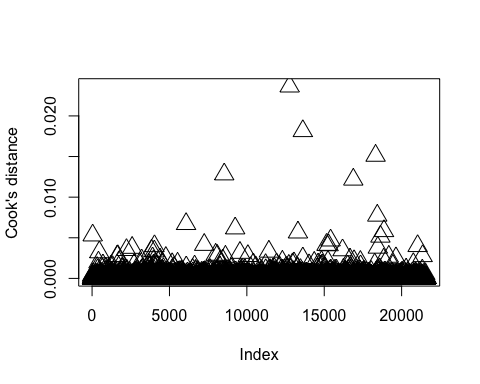
# The values of DFBETAS may be calculated using the function "dfbetas":   
plot(dfbetas(model)[,'sqft\_living'], pch=2, bg='orange', cex=2, ylab="DFBETA (sqft\_living)")



#Cook's Distance  
cook <- cooks.distance(model)  
  
# The associated cutoff value for identifying an   
# influential observation is   
cut.inf <- 1  
  
#plot to indicate outliers  
plot(cook, pch="\*", cex=2, main="Influential Obs by Cooks distance") # plot cook's distance  
abline(h = 4\*mean(cook, na.rm=T), col="red") # add cutoff line  
text(x=1:length(cook)+1, y=cook, labels=ifelse(cook>4\*mean(cook, na.rm=T),names(cook),""), col="red") # add labels



#another way of visualizing the same  
plot(cooks.distance(model), pch=2, bg='orange', cex=2, ylab="Cook's distance")



#displaying influential points  
df[which(cooks.distance(model) > 0.1),]

## [1] price bedrooms bathrooms sqft\_living floors   
## [6] waterfront view condition grade sqft\_above   
## [11] sqft\_basement yr\_renovated zipcode lat long   
## [16] sqft\_living15 year month   
## <0 rows> (or 0-length row.names)

influential <- as.numeric(names(cook)[(cook > 4\*mean(cook, na.rm=T))]) # influential row numbers  
nrow(df[influential, ])

## [1] 955

#removing outliers from the dataset as the number of outliers is lower than 5% of the dataset  
  
df.after.cd = df[-influential, ]  
  
  
model2 = lm(formula = price ~ bathrooms + floors + waterfront + view + condition +  
 + sqft\_basement + yr\_renovated + zipcode + sqft\_living15 + sqft\_living + grade,  
 data = df.after.cd)  
  
summary(model2)

##   
## Call:  
## lm(formula = price ~ bathrooms + floors + waterfront + view +   
## condition + +sqft\_basement + yr\_renovated + zipcode + sqft\_living15 +   
## sqft\_living + grade, data = df.after.cd)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.120436 -0.007464 0.000527 0.007982 0.071213   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.397e+00 3.188e-03 752.039 < 2e-16 \*\*\*  
## bathrooms 2.652e-03 2.165e-04 12.251 < 2e-16 \*\*\*  
## floors1.5 2.771e-03 3.680e-04 7.529 5.34e-14 \*\*\*  
## floors2 -5.240e-04 2.898e-04 -1.808 0.07056 .   
## floors2.5 -1.108e-03 1.168e-03 -0.949 0.34247   
## floors3 -9.377e-03 6.634e-04 -14.134 < 2e-16 \*\*\*  
## floors3.5 -1.158e-02 5.118e-03 -2.262 0.02371 \*   
## waterfront1 3.434e-02 1.514e-03 22.675 < 2e-16 \*\*\*  
## view1 8.181e-03 7.864e-04 10.403 < 2e-16 \*\*\*  
## view2 7.629e-03 4.786e-04 15.939 < 2e-16 \*\*\*  
## view3 1.188e-02 6.620e-04 17.940 < 2e-16 \*\*\*  
## view4 1.675e-02 1.041e-03 16.086 < 2e-16 \*\*\*  
## condition2 2.226e-02 3.248e-03 6.854 7.40e-12 \*\*\*  
## condition3 2.938e-02 3.030e-03 9.696 < 2e-16 \*\*\*  
## condition4 3.315e-02 3.031e-03 10.936 < 2e-16 \*\*\*  
## condition5 3.764e-02 3.044e-03 12.363 < 2e-16 \*\*\*  
## sqft\_basement1 -1.456e-03 2.436e-04 -5.977 2.30e-09 \*\*\*  
## yr\_renovated1 6.887e-03 4.863e-04 14.162 < 2e-16 \*\*\*  
## zipcode98002 -2.633e-03 1.219e-03 -2.160 0.03078 \*   
## zipcode98003 8.807e-04 1.101e-03 0.800 0.42390   
## zipcode98004 8.201e-02 1.079e-03 76.027 < 2e-16 \*\*\*  
## zipcode98005 5.684e-02 1.295e-03 43.906 < 2e-16 \*\*\*  
## zipcode98006 4.689e-02 9.709e-04 48.294 < 2e-16 \*\*\*  
## zipcode98007 5.030e-02 1.380e-03 36.455 < 2e-16 \*\*\*  
## zipcode98008 5.043e-02 1.094e-03 46.099 < 2e-16 \*\*\*  
## zipcode98010 1.994e-02 1.634e-03 12.204 < 2e-16 \*\*\*  
## zipcode98011 3.527e-02 1.219e-03 28.941 < 2e-16 \*\*\*  
## zipcode98014 2.891e-02 1.493e-03 19.367 < 2e-16 \*\*\*  
## zipcode98019 2.697e-02 1.238e-03 21.788 < 2e-16 \*\*\*  
## zipcode98022 6.114e-03 1.177e-03 5.194 2.08e-07 \*\*\*  
## zipcode98023 -2.649e-03 9.539e-04 -2.777 0.00550 \*\*   
## zipcode98024 3.767e-02 1.769e-03 21.301 < 2e-16 \*\*\*  
## zipcode98027 3.979e-02 1.000e-03 39.778 < 2e-16 \*\*\*  
## zipcode98028 3.304e-02 1.090e-03 30.314 < 2e-16 \*\*\*  
## zipcode98029 4.547e-02 1.057e-03 43.019 < 2e-16 \*\*\*  
## zipcode98030 4.712e-03 1.120e-03 4.208 2.59e-05 \*\*\*  
## zipcode98031 5.927e-03 1.100e-03 5.390 7.12e-08 \*\*\*  
## zipcode98032 -3.138e-03 1.444e-03 -2.173 0.02979 \*   
## zipcode98033 5.910e-02 9.838e-04 60.075 < 2e-16 \*\*\*  
## zipcode98034 4.142e-02 9.382e-04 44.145 < 2e-16 \*\*\*  
## zipcode98038 1.384e-02 9.224e-04 15.007 < 2e-16 \*\*\*  
## zipcode98039 8.913e-02 2.224e-03 40.084 < 2e-16 \*\*\*  
## zipcode98040 6.428e-02 1.124e-03 57.204 < 2e-16 \*\*\*  
## zipcode98042 4.621e-03 9.339e-04 4.948 7.55e-07 \*\*\*  
## zipcode98045 2.842e-02 1.190e-03 23.886 < 2e-16 \*\*\*  
## zipcode98052 4.881e-02 9.300e-04 52.479 < 2e-16 \*\*\*  
## zipcode98053 4.628e-02 1.006e-03 46.005 < 2e-16 \*\*\*  
## zipcode98055 1.076e-02 1.117e-03 9.636 < 2e-16 \*\*\*  
## zipcode98056 2.464e-02 1.003e-03 24.557 < 2e-16 \*\*\*  
## zipcode98058 1.268e-02 9.746e-04 13.015 < 2e-16 \*\*\*  
## zipcode98059 2.657e-02 9.702e-04 27.389 < 2e-16 \*\*\*  
## zipcode98065 3.108e-02 1.078e-03 28.824 < 2e-16 \*\*\*  
## zipcode98070 2.958e-02 1.558e-03 18.992 < 2e-16 \*\*\*  
## zipcode98072 3.895e-02 1.112e-03 35.020 < 2e-16 \*\*\*  
## zipcode98074 4.267e-02 9.863e-04 43.262 < 2e-16 \*\*\*  
## zipcode98075 4.215e-02 1.045e-03 40.331 < 2e-16 \*\*\*  
## zipcode98077 3.674e-02 1.237e-03 29.708 < 2e-16 \*\*\*  
## zipcode98092 2.932e-03 1.030e-03 2.846 0.00443 \*\*   
## zipcode98102 7.162e-02 1.592e-03 44.982 < 2e-16 \*\*\*  
## zipcode98103 6.308e-02 9.425e-04 66.925 < 2e-16 \*\*\*  
## zipcode98105 7.043e-02 1.174e-03 59.997 < 2e-16 \*\*\*  
## zipcode98106 2.371e-02 1.057e-03 22.423 < 2e-16 \*\*\*  
## zipcode98107 6.386e-02 1.129e-03 56.543 < 2e-16 \*\*\*  
## zipcode98108 2.645e-02 1.248e-03 21.195 < 2e-16 \*\*\*  
## zipcode98109 7.451e-02 1.531e-03 48.659 < 2e-16 \*\*\*  
## zipcode98112 7.740e-02 1.124e-03 68.838 < 2e-16 \*\*\*  
## zipcode98115 6.197e-02 9.316e-04 66.516 < 2e-16 \*\*\*  
## zipcode98116 5.732e-02 1.061e-03 54.028 < 2e-16 \*\*\*  
## zipcode98117 6.120e-02 9.399e-04 65.116 < 2e-16 \*\*\*  
## zipcode98118 3.413e-02 9.604e-04 35.543 < 2e-16 \*\*\*  
## zipcode98119 7.289e-02 1.278e-03 57.021 < 2e-16 \*\*\*  
## zipcode98122 6.034e-02 1.105e-03 54.623 < 2e-16 \*\*\*  
## zipcode98125 4.327e-02 9.959e-04 43.442 < 2e-16 \*\*\*  
## zipcode98126 4.085e-02 1.037e-03 39.386 < 2e-16 \*\*\*  
## zipcode98133 3.533e-02 9.583e-04 36.863 < 2e-16 \*\*\*  
## zipcode98136 5.191e-02 1.131e-03 45.904 < 2e-16 \*\*\*  
## zipcode98144 4.890e-02 1.051e-03 46.519 < 2e-16 \*\*\*  
## zipcode98146 1.990e-02 1.113e-03 17.881 < 2e-16 \*\*\*  
## zipcode98148 1.331e-02 2.042e-03 6.518 7.30e-11 \*\*\*  
## zipcode98155 3.318e-02 9.813e-04 33.811 < 2e-16 \*\*\*  
## zipcode98166 2.431e-02 1.150e-03 21.133 < 2e-16 \*\*\*  
## zipcode98168 5.281e-03 1.121e-03 4.710 2.50e-06 \*\*\*  
## zipcode98177 4.491e-02 1.131e-03 39.724 < 2e-16 \*\*\*  
## zipcode98178 1.050e-02 1.128e-03 9.305 < 2e-16 \*\*\*  
## zipcode98188 7.771e-03 1.426e-03 5.451 5.07e-08 \*\*\*  
## zipcode98198 4.928e-03 1.109e-03 4.443 8.92e-06 \*\*\*  
## zipcode98199 6.452e-02 1.070e-03 60.303 < 2e-16 \*\*\*  
## sqft\_living15 6.502e-06 2.488e-07 26.130 < 2e-16 \*\*\*  
## sqft\_living 1.365e-05 2.255e-07 60.541 < 2e-16 \*\*\*  
## grade 7.296e-03 1.496e-04 48.780 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01348 on 20550 degrees of freedom  
## Multiple R-squared: 0.8825, Adjusted R-squared: 0.882   
## F-statistic: 1734 on 89 and 20550 DF, p-value: < 2.2e-16

#R-squared jumps from 0.87 to 0.88