Assesing Regression Assignment

Dean D'souza

2016-06-19

## Regression on Data Set 1: Diamonds

## carat cut color clarity depth table price x y z  
## 1 0.23 Ideal E SI2 61.5 55 326 3.95 3.98 2.43  
## 2 0.21 Premium E SI1 59.8 61 326 3.89 3.84 2.31  
## 3 0.23 Good E VS1 56.9 65 327 4.05 4.07 2.31  
## 4 0.29 Premium I VS2 62.4 58 334 4.20 4.23 2.63  
## 5 0.31 Good J SI2 63.3 58 335 4.34 4.35 2.75  
## 6 0.24 Very Good J VVS2 62.8 57 336 3.94 3.96 2.48

#### The Analysis

As the dataset consists of more than two numeric rows, we go forward with determining the correlation among variables. Also, in order to build the correlation matrix, we take a subset of the data, removing the columns that are not numeric.

datadia1<-datadia[,c(-2,-3,-4)]  
diacor<-cor(datadia1)  
diacor

## carat depth table price x y  
## carat 1.00000000 0.02822431 0.1816175 0.9215913 0.97509423 0.95172220  
## depth 0.02822431 1.00000000 -0.2957785 -0.0106474 -0.02528925 -0.02934067  
## table 0.18161755 -0.29577852 1.0000000 0.1271339 0.19534428 0.18376015  
## price 0.92159130 -0.01064740 0.1271339 1.0000000 0.88443516 0.86542090  
## x 0.97509423 -0.02528925 0.1953443 0.8844352 1.00000000 0.97470148  
## y 0.95172220 -0.02934067 0.1837601 0.8654209 0.97470148 1.00000000  
## z 0.95338738 0.09492388 0.1509287 0.8612494 0.97077180 0.95200572  
## z  
## carat 0.95338738  
## depth 0.09492388  
## table 0.15092869  
## price 0.86124944  
## x 0.97077180  
## y 0.95200572  
## z 1.00000000

We can see that this data set contains a number of variables that are correlated with a few exceptions. We continue with using 'findLinearCombos()' on the subsetted data.

findLinearCombos(datadia1)

## $linearCombos  
## list()  
##   
## $remove  
## NULL

We can see that none of the columns reuire to be removed and hence continue with partioning the data by using 'createDataPartition()' into training and test sets.We consider price to be our response variable.(we also set the seed to '564738')

set.seed(564738)  
dia\_sv<-createDataPartition(datadia1$price, p=0.75, list = FALSE)  
dia\_train<-datadia1[dia\_sv,]  
dia\_test<-datadia1[-dia\_sv,]

We now build the linear model as follows:

dia\_lm<-lm(dia\_train$price ~ . , data = dia\_train)  
summary(dia\_lm)

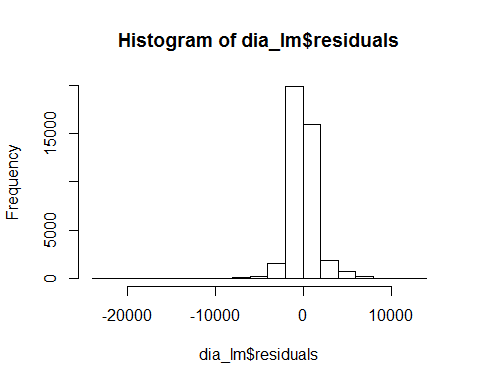
##   
## Call:  
## lm(formula = dia\_train$price ~ ., data = dia\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23945.9 -605.8 -53.8 342.6 12765.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 21176.067 508.333 41.658 <2e-16 \*\*\*  
## carat 10738.730 74.149 144.827 <2e-16 \*\*\*  
## depth -205.941 6.168 -33.388 <2e-16 \*\*\*  
## table -102.991 3.545 -29.053 <2e-16 \*\*\*  
## x -1336.412 46.563 -28.701 <2e-16 \*\*\*  
## y 61.389 25.764 2.383 0.0172 \*   
## z 37.023 44.772 0.827 0.4083   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1493 on 40450 degrees of freedom  
## Multiple R-squared: 0.8593, Adjusted R-squared: 0.8593   
## F-statistic: 4.117e+04 on 6 and 40450 DF, p-value: < 2.2e-16

Here we can see that the value of "z" does not seem to have a significant impact on the value of price. We also take a look at the residuals.

summary(dia\_lm$residuals)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -23950.00 -605.80 -53.79 0.00 342.60 12770.00

hist(dia\_lm$residuals)



We now rebuild the model, taking all the variables having a lesser impact on price out.

dia\_lm2<-lm(dia\_train$price ~ . -z, data = dia\_train)  
summary(dia\_lm2)

##   
## Call:  
## lm(formula = dia\_train$price ~ . - z, data = dia\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23940.3 -605.8 -53.8 342.3 12765.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 21046.250 483.484 43.530 <2e-16 \*\*\*  
## carat 10739.016 74.148 144.833 <2e-16 \*\*\*  
## depth -203.779 5.586 -36.479 <2e-16 \*\*\*  
## table -103.032 3.545 -29.068 <2e-16 \*\*\*  
## x -1316.624 39.944 -32.962 <2e-16 \*\*\*  
## y 64.181 25.542 2.513 0.012 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1493 on 40451 degrees of freedom  
## Multiple R-squared: 0.8593, Adjusted R-squared: 0.8593   
## F-statistic: 4.941e+04 on 5 and 40451 DF, p-value: < 2.2e-16

As all the variables have low p-values and have an effect on the estimate we consider this model to be that of the best fit.

anova(dia\_lm2,dia\_lm, test="Chisq")

## Analysis of Variance Table  
##   
## Model 1: dia\_train$price ~ (carat + depth + table + x + y + z) - z  
## Model 2: dia\_train$price ~ carat + depth + table + x + y + z  
## Res.Df RSS Df Sum of Sq Pr(>Chi)  
## 1 40451 9.0133e+10   
## 2 40450 9.0131e+10 1 1523703 0.4083

We now try to predict the values for our test data set.

dia\_p<-predict(dia\_lm2,dia\_test)  
summary(datadia1$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 326 950 2401 3933 5324 18820

summary(dia\_train$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 326 950 2401 3926 5324 18820

summary(dia\_test$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 326.0 949.5 2401.0 3953.0 5326.0 18820.0

summary(dia\_p)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1662.0 937.9 2965.0 3959.0 5681.0 37170.0

The above summaries shows that for the most part our initial, tranining and test data sets have the same median, with means that are close. The predicted data shows the mean value is only slightly above those of the other summaries.The median on the other hand, seems to be off by a large margin and hence we should conclude that there are other factors (like the categorical factors omitted before) that should be considered.

set.seed(564738)  
dia\_sv1<-createDataPartition(datadia$price, p=0.75, list = FALSE)  
dia\_train1<-datadia[dia\_sv1,]  
dia\_test1<-datadia[-dia\_sv1,]  
dia\_lm4<-lm(dia\_train1$price ~ ., data = dia\_train1)  
summary(dia\_lm4)

##   
## Call:  
## lm(formula = dia\_train1$price ~ ., data = dia\_train1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21430.2 -590.6 -185.8 372.7 10681.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6128.805 454.531 13.484 < 2e-16 \*\*\*  
## carat 11303.827 57.230 197.516 < 2e-16 \*\*\*  
## cut.L 596.045 26.254 22.703 < 2e-16 \*\*\*  
## cut.Q -314.067 20.969 -14.977 < 2e-16 \*\*\*  
## cut.C 146.197 17.909 8.163 3.36e-16 \*\*\*  
## cut^4 -17.915 14.275 -1.255 0.20946   
## color.L -1939.555 20.026 -96.854 < 2e-16 \*\*\*  
## color.Q -658.050 18.221 -36.114 < 2e-16 \*\*\*  
## color.C -163.851 16.999 -9.639 < 2e-16 \*\*\*  
## color^4 42.163 15.612 2.701 0.00692 \*\*   
## color^5 -90.561 14.748 -6.141 8.29e-10 \*\*\*  
## color^6 -32.315 13.429 -2.406 0.01612 \*   
## clarity.L 4088.428 34.881 117.211 < 2e-16 \*\*\*  
## clarity.Q -1910.942 32.550 -58.708 < 2e-16 \*\*\*  
## clarity.C 1004.223 27.862 36.043 < 2e-16 \*\*\*  
## clarity^4 -360.728 22.228 -16.228 < 2e-16 \*\*\*  
## clarity^5 237.696 18.174 13.079 < 2e-16 \*\*\*  
## clarity^6 7.177 15.877 0.452 0.65123   
## clarity^7 98.146 13.976 7.023 2.21e-12 \*\*\*  
## depth -65.823 5.134 -12.821 < 2e-16 \*\*\*  
## table -28.664 3.360 -8.531 < 2e-16 \*\*\*  
## x -1054.816 35.704 -29.543 < 2e-16 \*\*\*  
## y 14.431 19.555 0.738 0.46052   
## z -29.072 33.930 -0.857 0.39155   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1130 on 40433 degrees of freedom  
## Multiple R-squared: 0.9194, Adjusted R-squared: 0.9194   
## F-statistic: 2.005e+04 on 23 and 40433 DF, p-value: < 2.2e-16

dia\_lm5<-lm(dia\_train1$price ~ . -clarity^6, data = dia\_train1)  
summary(dia\_lm5)

##   
## Call:  
## lm(formula = dia\_train1$price ~ . - clarity^6, data = dia\_train1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23535.1 -583.5 -107.0 381.3 12447.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11961.343 554.936 21.554 < 2e-16 \*\*\*  
## carat 11379.885 69.727 163.206 < 2e-16 \*\*\*  
## cut.L 1037.863 31.940 32.494 < 2e-16 \*\*\*  
## cut.Q -509.706 25.502 -19.987 < 2e-16 \*\*\*  
## cut.C 329.798 21.906 15.055 < 2e-16 \*\*\*  
## cut^4 42.104 17.511 2.404 0.0162 \*   
## color.L -1636.354 24.392 -67.086 < 2e-16 \*\*\*  
## color.Q -762.954 22.267 -34.263 < 2e-16 \*\*\*  
## color.C -105.855 20.873 -5.071 3.97e-07 \*\*\*  
## color^4 101.669 19.166 5.305 1.13e-07 \*\*\*  
## color^5 -133.397 18.122 -7.361 1.86e-13 \*\*\*  
## color^6 -136.429 16.464 -8.287 < 2e-16 \*\*\*  
## depth -118.875 6.280 -18.929 < 2e-16 \*\*\*  
## table -41.318 4.128 -10.009 < 2e-16 \*\*\*  
## x -1439.764 43.538 -33.069 < 2e-16 \*\*\*  
## y 40.873 24.035 1.701 0.0890 .   
## z 35.048 41.702 0.840 0.4007   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1389 on 40440 degrees of freedom  
## Multiple R-squared: 0.8782, Adjusted R-squared: 0.8781   
## F-statistic: 1.822e+04 on 16 and 40440 DF, p-value: < 2.2e-16

dia\_lm6<-lm(dia\_train1$price ~ . -clarity^6-z, data = dia\_train1)  
summary(dia\_lm6)

##   
## Call:  
## lm(formula = dia\_train1$price ~ . - clarity^6 - z, data = dia\_train1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23530.0 -583.4 -106.8 381.4 12447.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11839.036 535.512 22.108 < 2e-16 \*\*\*  
## carat 11380.138 69.726 163.211 < 2e-16 \*\*\*  
## cut.L 1037.743 31.939 32.491 < 2e-16 \*\*\*  
## cut.Q -509.778 25.502 -19.990 < 2e-16 \*\*\*  
## cut.C 329.980 21.905 15.064 < 2e-16 \*\*\*  
## cut^4 42.537 17.503 2.430 0.0151 \*   
## color.L -1636.343 24.392 -67.086 < 2e-16 \*\*\*  
## color.Q -762.828 22.267 -34.259 < 2e-16 \*\*\*  
## color.C -105.692 20.872 -5.064 4.13e-07 \*\*\*  
## color^4 101.600 19.165 5.301 1.16e-07 \*\*\*  
## color^5 -133.366 18.122 -7.359 1.88e-13 \*\*\*  
## color^6 -136.447 16.464 -8.288 < 2e-16 \*\*\*  
## depth -116.846 5.797 -20.155 < 2e-16 \*\*\*  
## table -41.348 4.128 -10.017 < 2e-16 \*\*\*  
## x -1420.953 37.344 -38.050 < 2e-16 \*\*\*  
## y 43.461 23.837 1.823 0.0683 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1389 on 40441 degrees of freedom  
## Multiple R-squared: 0.8782, Adjusted R-squared: 0.8781   
## F-statistic: 1.944e+04 on 15 and 40441 DF, p-value: < 2.2e-16

dia\_lm7<-lm(dia\_train1$price ~ . -clarity^6-z-y, data = dia\_train1)  
summary(dia\_lm7)

##   
## Call:  
## lm(formula = dia\_train1$price ~ . - clarity^6 - z - y, data = dia\_train1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23540.0 -583.4 -107.0 381.2 12447.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11873.183 535.200 22.185 < 2e-16 \*\*\*  
## carat 11382.664 69.715 163.275 < 2e-16 \*\*\*  
## cut.L 1038.434 31.938 32.514 < 2e-16 \*\*\*  
## cut.Q -511.291 25.489 -20.059 < 2e-16 \*\*\*  
## cut.C 331.969 21.878 15.173 < 2e-16 \*\*\*  
## cut^4 43.418 17.497 2.481 0.0131 \*   
## color.L -1636.379 24.392 -67.086 < 2e-16 \*\*\*  
## color.Q -762.909 22.267 -34.261 < 2e-16 \*\*\*  
## color.C -105.646 20.873 -5.061 4.18e-07 \*\*\*  
## color^4 101.554 19.166 5.299 1.17e-07 \*\*\*  
## color^5 -133.214 18.123 -7.351 2.01e-13 \*\*\*  
## color^6 -136.395 16.464 -8.284 < 2e-16 \*\*\*  
## depth -117.184 5.795 -20.223 < 2e-16 \*\*\*  
## table -41.498 4.127 -10.055 < 2e-16 \*\*\*  
## x -1378.671 29.272 -47.098 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1389 on 40442 degrees of freedom  
## Multiple R-squared: 0.8782, Adjusted R-squared: 0.8781   
## F-statistic: 2.082e+04 on 14 and 40442 DF, p-value: < 2.2e-16

The final model made here should provide the best fit considering all variables included.

dia\_p1<-predict(dia\_lm7,dia\_test1)  
summary(datadia$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 326 950 2401 3933 5324 18820

summary(dia\_train$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 326 950 2401 3926 5324 18820

summary(dia\_test$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 326.0 949.5 2401.0 3953.0 5326.0 18820.0

summary(dia\_p1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -2826 1011 2870 3955 5829 36460

From the above summary we can see that we have made a better estimate, with less variation in median and mean across the initial, training, test and predicted values.

## Regression on Data Set 2 : Sacramento

## city zip beds baths sqft type price latitude longitude  
## 1 SACRAMENTO z95838 2 1 836 Residential 59222 38.63191 -121.4349  
## 2 SACRAMENTO z95823 3 1 1167 Residential 68212 38.47890 -121.4310  
## 3 SACRAMENTO z95815 2 1 796 Residential 68880 38.61830 -121.4438  
## 4 SACRAMENTO z95815 2 1 852 Residential 69307 38.61684 -121.4391  
## 5 SACRAMENTO z95824 2 1 797 Residential 81900 38.51947 -121.4358  
## 6 SACRAMENTO z95841 3 1 1122 Condo 89921 38.66260 -121.3278

#### The Analysis

As the data set consists of more than two numeric rows, we go forward with determining the correlation among variables.Also, in order to build the correlation matrix, we take a subset of the data, removing the columns that are not numeric.

datasac1<-datasac[,c(-1,-2,-6)]  
saccor<-cor(datasac1)  
saccor

## beds baths sqft price latitude  
## beds 1.00000000 0.65677220 0.7159079 0.4634122 -0.05362760  
## baths 0.65677220 1.00000000 0.7607516 0.5756415 0.04723159  
## sqft 0.71590791 0.76075156 1.0000000 0.7673609 0.04224950  
## price 0.46341217 0.57564155 0.7673609 1.0000000 0.15480838  
## latitude -0.05362760 0.04723159 0.0422495 0.1548084 1.00000000  
## longitude 0.08368503 0.17991708 0.2364463 0.3630227 0.38234665  
## longitude  
## beds 0.08368503  
## baths 0.17991708  
## sqft 0.23644634  
## price 0.36302273  
## latitude 0.38234665  
## longitude 1.00000000

Here we can see that the sqft variable has a greater correlation with the value of price. We go ahead with using the 'findLinearCombos()' function to see if there are any variables that can be removed.

findLinearCombos(datasac1)

## $linearCombos  
## list()  
##   
## $remove  
## NULL

Here we see that none of the variables are eligible for removal so we continue with partioning the data into test and training sets.We consider price to be our response variable.(we also set the seed to '564738')

set.seed(564738)  
sac\_sv<-createDataPartition(datasac1$price, p=0.75, list = FALSE)  
sac\_train<-datasac1[sac\_sv,]  
sac\_test<-datasac1[-sac\_sv,]

We now build the linear model:

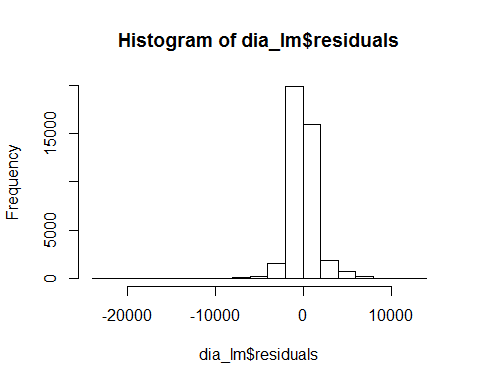
sac\_lm<-lm(price ~ . , data = sac\_train)  
summary(sac\_lm)

##   
## Call:  
## lm(formula = price ~ ., data = sac\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -231015 -50423 -11043 39879 571456   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.541e+07 3.350e+06 4.599 5.05e-06 \*\*\*  
## beds -1.966e+04 4.897e+03 -4.015 6.60e-05 \*\*\*  
## baths 4.669e+03 6.367e+03 0.733 0.464   
## sqft 1.419e+02 6.976e+00 20.348 < 2e-16 \*\*\*  
## latitude 2.601e+04 2.439e+04 1.067 0.287   
## longitude 1.347e+05 2.363e+04 5.700 1.77e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 77380 on 694 degrees of freedom  
## Multiple R-squared: 0.638, Adjusted R-squared: 0.6354   
## F-statistic: 244.6 on 5 and 694 DF, p-value: < 2.2e-16

summary(dia\_lm$residuals)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -23950.00 -605.80 -53.79 0.00 342.60 12770.00

hist(dia\_lm$residuals)



From the above summary we can see that the varaibles baths and latitude have lesser effect on determining the value of price, so we rebuild the linear model taking out the variables one at the time starting with baths.

sac\_lm2<-lm(price ~ . -baths, data = sac\_train)  
summary(sac\_lm2)

##   
## Call:  
## lm(formula = price ~ . - baths, data = sac\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -233387 -50776 -10851 40168 579847   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.538e+07 3.349e+06 4.594 5.17e-06 \*\*\*  
## beds -1.880e+04 4.753e+03 -3.955 8.43e-05 \*\*\*  
## sqft 1.447e+02 5.863e+00 24.681 < 2e-16 \*\*\*  
## latitude 2.690e+04 2.435e+04 1.105 0.27   
## longitude 1.348e+05 2.363e+04 5.705 1.72e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 77360 on 695 degrees of freedom  
## Multiple R-squared: 0.6377, Adjusted R-squared: 0.6356   
## F-statistic: 305.8 on 4 and 695 DF, p-value: < 2.2e-16

sac\_lm3<-lm(price ~ . -baths-latitude, data = sac\_train)  
summary(sac\_lm3)

##   
## Call:  
## lm(formula = price ~ . - baths - latitude, data = sac\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -233895 -51322 -10050 39534 583845   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.766e+07 2.642e+06 6.684 4.76e-11 \*\*\*  
## beds -1.933e+04 4.730e+03 -4.087 4.89e-05 \*\*\*  
## sqft 1.448e+02 5.863e+00 24.707 < 2e-16 \*\*\*  
## longitude 1.450e+05 2.176e+04 6.662 5.50e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 77370 on 696 degrees of freedom  
## Multiple R-squared: 0.6371, Adjusted R-squared: 0.6355   
## F-statistic: 407.2 on 3 and 696 DF, p-value: < 2.2e-16

Now that all the variables have low p-values we consider the 3rd model built to be accurate, we continue to check with anova.

anova(sac\_lm3,sac\_lm,test="Chisq")

## Analysis of Variance Table  
##   
## Model 1: price ~ (beds + baths + sqft + latitude + longitude) - baths -   
## latitude  
## Model 2: price ~ beds + baths + sqft + latitude + longitude  
## Res.Df RSS Df Sum of Sq Pr(>Chi)  
## 1 696 4.1662e+12   
## 2 694 4.1557e+12 2 1.0524e+10 0.4153

We can conclude that the final model is of best fit after reducing the degrees of freedom and the Residual sum of squares. We now test our model by predicting the values of price for the test data set

sac\_p<-predict(sac\_lm3,sac\_test)  
summary(datasac1$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 30000 156000 220000 246700 305000 884800

summary(sac\_train$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 30000 156000 220000 245000 305000 884800

summary(sac\_test$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 30000 157300 221600 251600 305200 879000

summary(sac\_p)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 119300 175900 218400 242600 279900 578500

From the above set of summaries we can see that the value of the mean of the predicted values is less that the mean of the test data. The difference is significant (9000) and hence we should conclude that other factors (such as the categorical factors which were removed earlier) may play more of an important role than thought before.

set.seed(564738)  
sac\_sv1<-createDataPartition(datasac$price, p=0.75, list = FALSE)  
sac\_train1<-datasac[sac\_sv1,]  
sac\_test1<-datasac[-sac\_sv1,]  
sac\_lm4<-lm(price ~ . , data = sac\_train1)  
summary(sac\_lm4)

##   
## Call:  
## lm(formula = price ~ ., data = sac\_train1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -236845 -40195 -6518 28898 270328   
##   
## Coefficients: (31 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.996e+07 1.926e+07 1.555 0.120379   
## cityAUBURN 7.785e+04 6.834e+04 1.139 0.255038   
## cityCAMERON\_PARK -3.653e+04 6.266e+04 -0.583 0.560130   
## cityCARMICHAEL 8.104e+04 2.715e+04 2.985 0.002949 \*\*   
## cityCITRUS\_HEIGHTS 1.467e+03 2.173e+04 0.067 0.946215   
## cityCOOL 2.732e+04 9.890e+04 0.276 0.782460   
## cityDIAMOND\_SPRINGS -9.190e+04 1.084e+05 -0.848 0.396723   
## cityEL\_DORADO -5.410e+04 9.105e+04 -0.594 0.552608   
## cityEL\_DORADO\_HILLS 5.759e+04 5.269e+04 1.093 0.274835   
## cityELK\_GROVE 1.626e+04 5.779e+04 0.281 0.778566   
## cityELVERTA -2.795e+04 3.687e+04 -0.758 0.448755   
## cityFAIR\_OAKS 6.514e+04 3.392e+04 1.921 0.055247 .   
## cityFOLSOM 8.338e+04 4.156e+04 2.006 0.045235 \*   
## cityFORESTHILL -7.958e+04 1.257e+05 -0.633 0.526877   
## cityGALT -2.998e+04 8.120e+04 -0.369 0.712124   
## cityGOLD\_RIVER 2.033e+04 6.153e+04 0.330 0.741234   
## cityGREENWOOD -4.155e+04 1.092e+05 -0.380 0.703840   
## cityLINCOLN 7.992e+03 3.843e+04 0.208 0.835318   
## cityLOOMIS 3.295e+05 6.193e+04 5.321 1.44e-07 \*\*\*  
## cityMATHER -6.129e+04 7.321e+04 -0.837 0.402806   
## cityNORTH\_HIGHLANDS -2.687e+04 2.461e+04 -1.092 0.275376   
## cityORANGEVALE 2.924e+04 3.290e+04 0.889 0.374570   
## cityPLACERVILLE -3.767e+04 9.653e+04 -0.390 0.696472   
## cityPOLLOCK\_PINES -1.654e+05 1.382e+05 -1.196 0.231962   
## cityRANCHO\_CORDOVA -5.612e+04 4.422e+04 -1.269 0.204846   
## cityRANCHO\_MURIETA -3.820e+04 7.257e+04 -0.526 0.598768   
## cityRIO\_LINDA 1.782e+04 2.933e+04 0.608 0.543595   
## cityROCKLIN 7.285e+04 3.614e+04 2.016 0.044253 \*   
## cityROSEVILLE 1.042e+05 2.823e+04 3.692 0.000241 \*\*\*  
## citySACRAMENTO 1.058e+05 4.257e+04 2.486 0.013185 \*   
## cityWALNUT\_GROVE 1.635e+05 1.123e+05 1.456 0.145755   
## cityWEST\_SACRAMENTO 2.202e+04 5.679e+04 0.388 0.698373   
## cityWILTON 1.600e+05 6.861e+04 2.331 0.020049 \*   
## zipz95608 NA NA NA NA   
## zipz95610 -6.548e+03 3.042e+04 -0.215 0.829643   
## zipz95614 NA NA NA NA   
## zipz95619 NA NA NA NA   
## zipz95621 NA NA NA NA   
## zipz95623 NA NA NA NA   
## zipz95624 -1.568e+04 2.053e+04 -0.764 0.445277   
## zipz95626 NA NA NA NA   
## zipz95628 NA NA NA NA   
## zipz95630 NA NA NA NA   
## zipz95631 NA NA NA NA   
## zipz95632 NA NA NA NA   
## zipz95635 NA NA NA NA   
## zipz95648 NA NA NA NA   
## zipz95650 NA NA NA NA   
## zipz95655 NA NA NA NA   
## zipz95660 NA NA NA NA   
## zipz95661 -5.443e+03 3.897e+04 -0.140 0.888969   
## zipz95662 NA NA NA NA   
## zipz95667 NA NA NA NA   
## zipz95670 4.901e+04 3.344e+04 1.466 0.143256   
## zipz95673 NA NA NA NA   
## zipz95677 -2.582e+04 4.014e+04 -0.643 0.520221   
## zipz95678 -5.926e+04 2.691e+04 -2.202 0.027993 \*   
## zipz95682 NA NA NA NA   
## zipz95683 NA NA NA NA   
## zipz95690 NA NA NA NA   
## zipz95691 NA NA NA NA   
## zipz95693 NA NA NA NA   
## zipz95726 NA NA NA NA   
## zipz95742 NA NA NA NA   
## zipz95747 NA NA NA NA   
## zipz95757 1.527e+04 1.807e+04 0.845 0.398532   
## zipz95758 NA NA NA NA   
## zipz95762 NA NA NA NA   
## zipz95765 NA NA NA NA   
## zipz95811 1.781e+05 7.411e+04 2.403 0.016538 \*   
## zipz95814 3.096e+04 5.336e+04 0.580 0.561957   
## zipz95815 -1.229e+05 3.886e+04 -3.164 0.001633 \*\*   
## zipz95816 1.458e+05 7.697e+04 1.894 0.058750 .   
## zipz95817 -1.096e+05 4.862e+04 -2.253 0.024593 \*   
## zipz95818 1.468e+04 4.627e+04 0.317 0.751118   
## zipz95819 1.943e+05 5.726e+04 3.394 0.000732 \*\*\*  
## zipz95820 -1.019e+05 3.937e+04 -2.588 0.009864 \*\*   
## zipz95821 -7.505e+04 4.370e+04 -1.718 0.086360 .   
## zipz95822 -8.890e+04 4.404e+04 -2.019 0.043949 \*   
## zipz95823 -1.269e+05 4.231e+04 -3.000 0.002805 \*\*   
## zipz95824 -1.367e+05 4.200e+04 -3.255 0.001194 \*\*   
## zipz95825 -7.689e+04 3.826e+04 -2.010 0.044899 \*   
## zipz95826 -1.032e+05 3.736e+04 -2.762 0.005917 \*\*   
## zipz95827 -1.387e+05 4.026e+04 -3.445 0.000609 \*\*\*  
## zipz95828 -1.230e+05 3.997e+04 -3.078 0.002174 \*\*   
## zipz95829 -1.187e+05 4.710e+04 -2.520 0.011983 \*   
## zipz95831 3.548e+03 5.053e+04 0.070 0.944054   
## zipz95832 -1.354e+05 4.902e+04 -2.763 0.005898 \*\*   
## zipz95833 -7.273e+04 4.187e+04 -1.737 0.082833 .   
## zipz95834 -7.219e+04 4.275e+04 -1.688 0.091822 .   
## zipz95835 -5.020e+04 4.288e+04 -1.171 0.242092   
## zipz95838 -1.173e+05 3.735e+04 -3.142 0.001760 \*\*   
## zipz95841 -7.370e+04 4.574e+04 -1.611 0.107604   
## zipz95842 -1.297e+05 4.001e+04 -3.241 0.001254 \*\*   
## zipz95843 NA NA NA NA   
## zipz95864 NA NA NA NA   
## beds -1.193e+04 4.771e+03 -2.501 0.012623 \*   
## baths 1.127e+04 6.146e+03 1.834 0.067192 .   
## sqft 1.093e+02 7.313e+00 14.942 < 2e-16 \*\*\*  
## typeMulti\_Family 6.101e+03 2.581e+04 0.236 0.813248   
## typeResidential 4.661e+04 1.296e+04 3.595 0.000349 \*\*\*  
## latitude -7.088e+04 1.780e+05 -0.398 0.690627   
## longitude 2.242e+05 1.564e+05 1.433 0.152300   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 64630 on 628 degrees of freedom  
## Multiple R-squared: 0.7715, Adjusted R-squared: 0.7457   
## F-statistic: 29.86 on 71 and 628 DF, p-value: < 2.2e-16

sac\_lm5<-lm(price ~ . -latitude, data = sac\_train1)  
summary(sac\_lm5)

##   
## Call:  
## lm(formula = price ~ . - latitude, data = sac\_train1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -236870 -39840 -6613 28906 270551   
##   
## Coefficients: (31 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.828e+07 1.878e+07 1.506 0.132665   
## cityAUBURN 6.139e+04 5.438e+04 1.129 0.259347   
## cityCAMERON\_PARK -3.767e+04 6.255e+04 -0.602 0.547284   
## cityCARMICHAEL 8.594e+04 2.419e+04 3.552 0.000411 \*\*\*  
## cityCITRUS\_HEIGHTS 2.012e+03 2.167e+04 0.093 0.926054   
## cityCOOL 1.040e+04 8.924e+04 0.116 0.907303   
## cityDIAMOND\_SPRINGS -9.484e+04 1.080e+05 -0.878 0.380352   
## cityEL\_DORADO -5.569e+04 9.091e+04 -0.613 0.540350   
## cityEL\_DORADO\_HILLS 5.779e+04 5.265e+04 1.098 0.272787   
## cityELK\_GROVE 3.753e+04 2.198e+04 1.708 0.088148 .   
## cityELVERTA -2.739e+04 3.682e+04 -0.744 0.457268   
## cityFAIR\_OAKS 6.808e+04 3.308e+04 2.058 0.039996 \*   
## cityFOLSOM 8.458e+04 4.142e+04 2.042 0.041556 \*   
## cityFORESTHILL -1.060e+05 1.067e+05 -0.994 0.320840   
## cityGALT 7.284e+02 2.542e+04 0.029 0.977152   
## cityGOLD\_RIVER 2.889e+04 5.761e+04 0.501 0.616215   
## cityGREENWOOD -6.182e+04 9.658e+04 -0.640 0.522327   
## cityLINCOLN -3.565e+03 2.517e+04 -0.142 0.887434   
## cityLOOMIS 3.203e+05 5.744e+04 5.577 3.63e-08 \*\*\*  
## cityMATHER -5.016e+04 6.762e+04 -0.742 0.458484   
## cityNORTH\_HIGHLANDS -2.443e+04 2.382e+04 -1.025 0.305529   
## cityORANGEVALE 3.015e+04 3.280e+04 0.919 0.358429   
## cityPLACERVILLE -4.517e+04 9.461e+04 -0.477 0.633230   
## cityPOLLOCK\_PINES -1.762e+05 1.354e+05 -1.301 0.193648   
## cityRANCHO\_CORDOVA -4.483e+04 3.390e+04 -1.322 0.186504   
## cityRANCHO\_MURIETA -2.542e+04 6.504e+04 -0.391 0.696039   
## cityRIO\_LINDA 2.032e+04 2.863e+04 0.710 0.478093   
## cityROCKLIN 6.477e+04 2.988e+04 2.168 0.030547 \*   
## cityROSEVILLE 9.923e+04 2.525e+04 3.930 9.44e-05 \*\*\*  
## citySACRAMENTO 1.148e+05 3.613e+04 3.177 0.001559 \*\*   
## cityWALNUT\_GROVE 1.981e+05 7.107e+04 2.787 0.005482 \*\*   
## cityWEST\_SACRAMENTO 3.366e+04 4.866e+04 0.692 0.489344   
## cityWILTON 1.812e+05 4.319e+04 4.195 3.12e-05 \*\*\*  
## zipz95608 NA NA NA NA   
## zipz95610 -7.101e+03 3.037e+04 -0.234 0.815178   
## zipz95614 NA NA NA NA   
## zipz95619 NA NA NA NA   
## zipz95621 NA NA NA NA   
## zipz95623 NA NA NA NA   
## zipz95624 -1.655e+04 2.040e+04 -0.812 0.417349   
## zipz95626 NA NA NA NA   
## zipz95628 NA NA NA NA   
## zipz95630 NA NA NA NA   
## zipz95631 NA NA NA NA   
## zipz95632 NA NA NA NA   
## zipz95635 NA NA NA NA   
## zipz95648 NA NA NA NA   
## zipz95650 NA NA NA NA   
## zipz95655 NA NA NA NA   
## zipz95660 NA NA NA NA   
## zipz95661 -3.172e+03 3.853e+04 -0.082 0.934413   
## zipz95662 NA NA NA NA   
## zipz95667 NA NA NA NA   
## zipz95670 4.565e+04 3.233e+04 1.412 0.158529   
## zipz95673 NA NA NA NA   
## zipz95677 -2.403e+04 3.986e+04 -0.603 0.546766   
## zipz95678 -5.844e+04 2.681e+04 -2.180 0.029632 \*   
## zipz95682 NA NA NA NA   
## zipz95683 NA NA NA NA   
## zipz95690 NA NA NA NA   
## zipz95691 NA NA NA NA   
## zipz95693 NA NA NA NA   
## zipz95726 NA NA NA NA   
## zipz95742 NA NA NA NA   
## zipz95747 NA NA NA NA   
## zipz95757 1.767e+04 1.703e+04 1.038 0.299865   
## zipz95758 NA NA NA NA   
## zipz95762 NA NA NA NA   
## zipz95765 NA NA NA NA   
## zipz95811 1.796e+05 7.396e+04 2.428 0.015451 \*   
## zipz95814 3.262e+04 5.316e+04 0.614 0.539708   
## zipz95815 -1.242e+05 3.870e+04 -3.209 0.001402 \*\*   
## zipz95816 1.505e+05 7.601e+04 1.979 0.048196 \*   
## zipz95817 -1.059e+05 4.770e+04 -2.219 0.026814 \*   
## zipz95818 1.820e+04 4.539e+04 0.401 0.688593   
## zipz95819 1.965e+05 5.697e+04 3.449 0.000601 \*\*\*  
## zipz95820 -9.721e+04 3.754e+04 -2.590 0.009827 \*\*   
## zipz95821 -7.702e+04 4.339e+04 -1.775 0.076364 .   
## zipz95822 -8.144e+04 3.983e+04 -2.045 0.041302 \*   
## zipz95823 -1.176e+05 3.526e+04 -3.336 0.000899 \*\*\*  
## zipz95824 -1.310e+05 3.941e+04 -3.323 0.000941 \*\*\*  
## zipz95825 -7.641e+04 3.822e+04 -1.999 0.045988 \*   
## zipz95826 -1.005e+05 3.673e+04 -2.737 0.006384 \*\*   
## zipz95827 -1.373e+05 4.008e+04 -3.425 0.000654 \*\*\*  
## zipz95828 -1.150e+05 3.441e+04 -3.341 0.000883 \*\*\*  
## zipz95829 -1.096e+05 4.114e+04 -2.664 0.007924 \*\*   
## zipz95831 1.202e+04 4.581e+04 0.262 0.793089   
## zipz95832 -1.263e+05 4.323e+04 -2.920 0.003620 \*\*   
## zipz95833 -7.347e+04 4.180e+04 -1.758 0.079254 .   
## zipz95834 -7.441e+04 4.236e+04 -1.757 0.079464 .   
## zipz95835 -5.472e+04 4.132e+04 -1.324 0.185906   
## zipz95838 -1.207e+05 3.634e+04 -3.322 0.000946 \*\*\*  
## zipz95841 -7.862e+04 4.401e+04 -1.786 0.074533 .   
## zipz95842 -1.363e+05 3.630e+04 -3.756 0.000188 \*\*\*  
## zipz95843 NA NA NA NA   
## zipz95864 NA NA NA NA   
## beds -1.197e+04 4.767e+03 -2.511 0.012298 \*   
## baths 1.138e+04 6.136e+03 1.854 0.064164 .   
## sqft 1.093e+02 7.308e+00 14.950 < 2e-16 \*\*\*  
## typeMulti\_Family 5.529e+03 2.576e+04 0.215 0.830099   
## typeResidential 4.652e+04 1.295e+04 3.591 0.000355 \*\*\*  
## longitude 2.330e+05 1.548e+05 1.505 0.132772   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 64590 on 629 degrees of freedom  
## Multiple R-squared: 0.7714, Adjusted R-squared: 0.746   
## F-statistic: 30.33 on 70 and 629 DF, p-value: < 2.2e-16

sac\_lm6<-lm(price ~ . -latitude-longitude, data = sac\_train1)  
summary(sac\_lm6)

##   
## Call:  
## lm(formula = price ~ . - latitude - longitude, data = sac\_train1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -244770 -41014 -6725 29089 267704   
##   
## Coefficients: (31 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.804e+03 2.062e+04 0.378 0.705237   
## cityAUBURN 1.261e+05 3.334e+04 3.782 0.000171 \*\*\*  
## cityCAMERON\_PARK 4.688e+04 2.756e+04 1.701 0.089481 .   
## cityCARMICHAEL 9.357e+04 2.368e+04 3.951 8.66e-05 \*\*\*  
## cityCITRUS\_HEIGHTS 1.395e+04 2.019e+04 0.691 0.489959   
## cityCOOL 1.002e+05 6.644e+04 1.508 0.132052   
## cityDIAMOND\_SPRINGS 3.345e+04 6.645e+04 0.503 0.614847   
## cityEL\_DORADO 6.023e+04 4.835e+04 1.246 0.213282   
## cityEL\_DORADO\_HILLS 1.291e+05 2.295e+04 5.626 2.78e-08 \*\*\*  
## cityELK\_GROVE 2.077e+04 1.897e+04 1.095 0.273846   
## cityELVERTA -4.021e+04 3.586e+04 -1.121 0.262555   
## cityFAIR\_OAKS 9.251e+04 2.886e+04 3.206 0.001414 \*\*   
## cityFOLSOM 1.357e+05 2.373e+04 5.719 1.66e-08 \*\*\*  
## cityFORESTHILL 1.970e+04 6.647e+04 0.296 0.767006   
## cityGALT 1.507e+04 2.359e+04 0.639 0.523276   
## cityGOLD\_RIVER 6.797e+04 5.148e+04 1.320 0.187151   
## cityGREENWOOD 4.275e+04 6.716e+04 0.637 0.524663   
## cityLINCOLN 1.255e+04 2.281e+04 0.550 0.582375   
## cityLOOMIS 3.665e+05 4.863e+04 7.535 1.70e-13 \*\*\*  
## cityMATHER -3.181e+04 6.658e+04 -0.478 0.632965   
## cityNORTH\_HIGHLANDS -2.781e+04 2.374e+04 -1.172 0.241770   
## cityORANGEVALE 6.114e+04 2.556e+04 2.392 0.017068 \*   
## cityPLACERVILLE 9.046e+04 2.887e+04 3.133 0.001809 \*\*   
## cityPOLLOCK\_PINES 1.213e+03 6.673e+04 0.018 0.985501   
## cityRANCHO\_CORDOVA -1.413e+04 2.710e+04 -0.521 0.602310   
## cityRANCHO\_MURIETA 3.950e+04 4.873e+04 0.811 0.417808   
## cityRIO\_LINDA -9.374e+02 2.493e+04 -0.038 0.970018   
## cityROCKLIN 8.551e+04 2.653e+04 3.223 0.001335 \*\*   
## cityROSEVILLE 1.029e+05 2.516e+04 4.092 4.84e-05 \*\*\*  
## citySACRAMENTO 1.109e+05 3.607e+04 3.073 0.002207 \*\*   
## cityWALNUT\_GROVE 1.615e+05 6.687e+04 2.416 0.015983 \*   
## cityWEST\_SACRAMENTO -6.818e+03 4.059e+04 -0.168 0.866665   
## cityWILTON 2.145e+05 3.712e+04 5.779 1.18e-08 \*\*\*  
## zipz95608 NA NA NA NA   
## zipz95610 3.945e+03 2.950e+04 0.134 0.893634   
## zipz95614 NA NA NA NA   
## zipz95619 NA NA NA NA   
## zipz95621 NA NA NA NA   
## zipz95623 NA NA NA NA   
## zipz95624 -2.317e+03 1.809e+04 -0.128 0.898109   
## zipz95626 NA NA NA NA   
## zipz95628 NA NA NA NA   
## zipz95630 NA NA NA NA   
## zipz95631 NA NA NA NA   
## zipz95632 NA NA NA NA   
## zipz95635 NA NA NA NA   
## zipz95648 NA NA NA NA   
## zipz95650 NA NA NA NA   
## zipz95655 NA NA NA NA   
## zipz95660 NA NA NA NA   
## zipz95661 2.057e+04 3.519e+04 0.585 0.559076   
## zipz95662 NA NA NA NA   
## zipz95667 NA NA NA NA   
## zipz95670 3.027e+04 3.071e+04 0.986 0.324601   
## zipz95673 NA NA NA NA   
## zipz95677 -1.046e+04 3.886e+04 -0.269 0.787913   
## zipz95678 -4.582e+04 2.549e+04 -1.798 0.072709 .   
## zipz95682 NA NA NA NA   
## zipz95683 NA NA NA NA   
## zipz95690 NA NA NA NA   
## zipz95691 NA NA NA NA   
## zipz95693 NA NA NA NA   
## zipz95726 NA NA NA NA   
## zipz95742 NA NA NA NA   
## zipz95747 NA NA NA NA   
## zipz95757 1.613e+04 1.701e+04 0.948 0.343362   
## zipz95758 NA NA NA NA   
## zipz95762 NA NA NA NA   
## zipz95765 NA NA NA NA   
## zipz95811 1.560e+05 7.236e+04 2.156 0.031431 \*   
## zipz95814 6.626e+03 5.033e+04 0.132 0.895306   
## zipz95815 -1.406e+05 3.716e+04 -3.785 0.000169 \*\*\*  
## zipz95816 1.212e+05 7.356e+04 1.648 0.099827 .   
## zipz95817 -1.262e+05 4.579e+04 -2.756 0.006018 \*\*   
## zipz95818 -8.474e+03 4.183e+04 -0.203 0.839526   
## zipz95819 1.817e+05 5.617e+04 3.234 0.001284 \*\*   
## zipz95820 -1.144e+05 3.580e+04 -3.194 0.001473 \*\*   
## zipz95821 -7.962e+04 4.340e+04 -1.835 0.067026 .   
## zipz95822 -1.076e+05 3.589e+04 -2.997 0.002833 \*\*   
## zipz95823 -1.322e+05 3.393e+04 -3.897 0.000108 \*\*\*  
## zipz95824 -1.453e+05 3.828e+04 -3.795 0.000162 \*\*\*  
## zipz95825 -8.353e+04 3.796e+04 -2.200 0.028151 \*   
## zipz95826 -9.676e+04 3.668e+04 -2.638 0.008542 \*\*   
## zipz95827 -1.246e+05 3.922e+04 -3.177 0.001560 \*\*   
## zipz95828 -1.202e+05 3.426e+04 -3.509 0.000482 \*\*\*  
## zipz95829 -1.034e+05 4.098e+04 -2.525 0.011830 \*   
## zipz95831 -2.396e+04 3.911e+04 -0.613 0.540393   
## zipz95832 -1.518e+05 3.980e+04 -3.814 0.000150 \*\*\*  
## zipz95833 -1.038e+05 3.666e+04 -2.832 0.004776 \*\*   
## zipz95834 -1.075e+05 3.626e+04 -2.964 0.003149 \*\*   
## zipz95835 -8.846e+04 3.475e+04 -2.546 0.011139 \*   
## zipz95838 -1.373e+05 3.467e+04 -3.960 8.36e-05 \*\*\*  
## zipz95841 -7.053e+04 4.373e+04 -1.613 0.107239   
## zipz95842 -1.307e+05 3.614e+04 -3.616 0.000323 \*\*\*  
## zipz95843 NA NA NA NA   
## zipz95864 NA NA NA NA   
## beds -1.234e+04 4.765e+03 -2.590 0.009827 \*\*   
## baths 1.099e+04 6.137e+03 1.790 0.073879 .   
## sqft 1.097e+02 7.308e+00 15.017 < 2e-16 \*\*\*  
## typeMulti\_Family 7.036e+03 2.576e+04 0.273 0.784864   
## typeResidential 4.715e+04 1.296e+04 3.638 0.000297 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 64650 on 630 degrees of freedom  
## Multiple R-squared: 0.7706, Adjusted R-squared: 0.7455   
## F-statistic: 30.67 on 69 and 630 DF, p-value: < 2.2e-16

sac\_lm7<-lm(price ~ . -latitude-longitude-baths, data = sac\_train1)  
summary(sac\_lm7)

##   
## Call:  
## lm(formula = price ~ . - latitude - longitude - baths, data = sac\_train1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -254828 -40535 -5873 28965 261818   
##   
## Coefficients: (31 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.684e+04 2.003e+04 0.841 0.400919   
## cityAUBURN 1.263e+05 3.340e+04 3.782 0.000170 \*\*\*  
## cityCAMERON\_PARK 4.863e+04 2.759e+04 1.763 0.078446 .   
## cityCARMICHAEL 8.906e+04 2.359e+04 3.776 0.000175 \*\*\*  
## cityCITRUS\_HEIGHTS 1.166e+04 2.018e+04 0.577 0.563860   
## cityCOOL 9.938e+04 6.655e+04 1.493 0.135894   
## cityDIAMOND\_SPRINGS 3.362e+04 6.657e+04 0.505 0.613674   
## cityEL\_DORADO 5.543e+04 4.836e+04 1.146 0.252119   
## cityEL\_DORADO\_HILLS 1.289e+05 2.299e+04 5.606 3.10e-08 \*\*\*  
## cityELK\_GROVE 1.948e+04 1.898e+04 1.026 0.305267   
## cityELVERTA -4.003e+04 3.592e+04 -1.115 0.265478   
## cityFAIR\_OAKS 9.138e+04 2.890e+04 3.162 0.001641 \*\*   
## cityFOLSOM 1.329e+05 2.372e+04 5.604 3.13e-08 \*\*\*  
## cityFORESTHILL 2.030e+04 6.658e+04 0.305 0.760571   
## cityGALT 1.365e+04 2.362e+04 0.578 0.563566   
## cityGOLD\_RIVER 5.954e+04 5.135e+04 1.159 0.246696   
## cityGREENWOOD 3.322e+04 6.706e+04 0.495 0.620564   
## cityLINCOLN 1.328e+04 2.284e+04 0.581 0.561221   
## cityLOOMIS 3.751e+05 4.848e+04 7.737 4.06e-14 \*\*\*  
## cityMATHER -3.662e+04 6.664e+04 -0.549 0.582874   
## cityNORTH\_HIGHLANDS -3.111e+04 2.371e+04 -1.312 0.189927   
## cityORANGEVALE 5.825e+04 2.556e+04 2.279 0.022979 \*   
## cityPLACERVILLE 8.976e+04 2.892e+04 3.104 0.001995 \*\*   
## cityPOLLOCK\_PINES -9.730e+03 6.657e+04 -0.146 0.883842   
## cityRANCHO\_CORDOVA -1.607e+04 2.713e+04 -0.592 0.553895   
## cityRANCHO\_MURIETA 3.329e+04 4.869e+04 0.684 0.494380   
## cityRIO\_LINDA -3.450e+03 2.493e+04 -0.138 0.889992   
## cityROCKLIN 8.213e+04 2.651e+04 3.098 0.002037 \*\*   
## cityROSEVILLE 1.051e+05 2.517e+04 4.175 3.40e-05 \*\*\*  
## citySACRAMENTO 1.030e+05 3.586e+04 2.872 0.004210 \*\*   
## cityWALNUT\_GROVE 1.480e+05 6.656e+04 2.224 0.026481 \*   
## cityWEST\_SACRAMENTO -1.267e+04 4.053e+04 -0.313 0.754679   
## cityWILTON 2.044e+05 3.675e+04 5.562 3.93e-08 \*\*\*  
## zipz95608 NA NA NA NA   
## zipz95610 6.782e+03 2.950e+04 0.230 0.818269   
## zipz95614 NA NA NA NA   
## zipz95619 NA NA NA NA   
## zipz95621 NA NA NA NA   
## zipz95623 NA NA NA NA   
## zipz95624 -2.217e+03 1.812e+04 -0.122 0.902660   
## zipz95626 NA NA NA NA   
## zipz95628 NA NA NA NA   
## zipz95630 NA NA NA NA   
## zipz95631 NA NA NA NA   
## zipz95632 NA NA NA NA   
## zipz95635 NA NA NA NA   
## zipz95648 NA NA NA NA   
## zipz95650 NA NA NA NA   
## zipz95655 NA NA NA NA   
## zipz95660 NA NA NA NA   
## zipz95661 1.252e+04 3.496e+04 0.358 0.720422   
## zipz95662 NA NA NA NA   
## zipz95667 NA NA NA NA   
## zipz95670 3.336e+04 3.071e+04 1.086 0.277835   
## zipz95673 NA NA NA NA   
## zipz95677 -8.674e+03 3.892e+04 -0.223 0.823710   
## zipz95678 -4.977e+04 2.544e+04 -1.956 0.050865 .   
## zipz95682 NA NA NA NA   
## zipz95683 NA NA NA NA   
## zipz95690 NA NA NA NA   
## zipz95691 NA NA NA NA   
## zipz95693 NA NA NA NA   
## zipz95726 NA NA NA NA   
## zipz95742 NA NA NA NA   
## zipz95747 NA NA NA NA   
## zipz95757 1.604e+04 1.704e+04 0.941 0.347035   
## zipz95758 NA NA NA NA   
## zipz95762 NA NA NA NA   
## zipz95765 NA NA NA NA   
## zipz95811 1.548e+05 7.248e+04 2.136 0.033055 \*   
## zipz95814 2.309e+04 4.957e+04 0.466 0.641453   
## zipz95815 -1.387e+05 3.721e+04 -3.728 0.000210 \*\*\*  
## zipz95816 1.321e+05 7.344e+04 1.799 0.072509 .   
## zipz95817 -1.240e+05 4.585e+04 -2.704 0.007044 \*\*   
## zipz95818 -6.901e+03 4.190e+04 -0.165 0.869218   
## zipz95819 1.833e+05 5.626e+04 3.257 0.001185 \*\*   
## zipz95820 -1.126e+05 3.585e+04 -3.142 0.001759 \*\*   
## zipz95821 -7.991e+04 4.347e+04 -1.838 0.066494 .   
## zipz95822 -1.034e+05 3.587e+04 -2.881 0.004100 \*\*   
## zipz95823 -1.255e+05 3.378e+04 -3.715 0.000221 \*\*\*  
## zipz95824 -1.437e+05 3.834e+04 -3.748 0.000195 \*\*\*  
## zipz95825 -8.159e+04 3.801e+04 -2.146 0.032219 \*   
## zipz95826 -9.218e+04 3.665e+04 -2.515 0.012154 \*   
## zipz95827 -1.173e+05 3.908e+04 -3.002 0.002790 \*\*   
## zipz95828 -1.140e+05 3.415e+04 -3.339 0.000890 \*\*\*  
## zipz95829 -9.529e+04 4.079e+04 -2.336 0.019805 \*   
## zipz95831 -1.685e+04 3.898e+04 -0.432 0.665624   
## zipz95832 -1.456e+05 3.972e+04 -3.666 0.000267 \*\*\*  
## zipz95833 -9.701e+04 3.652e+04 -2.656 0.008103 \*\*   
## zipz95834 -9.916e+04 3.602e+04 -2.753 0.006075 \*\*   
## zipz95835 -8.217e+04 3.463e+04 -2.373 0.017954 \*   
## zipz95838 -1.327e+05 3.463e+04 -3.832 0.000140 \*\*\*  
## zipz95841 -6.983e+04 4.380e+04 -1.594 0.111406   
## zipz95842 -1.247e+05 3.605e+04 -3.461 0.000575 \*\*\*  
## zipz95843 NA NA NA NA   
## zipz95864 NA NA NA NA   
## beds -1.090e+04 4.705e+03 -2.316 0.020856 \*   
## sqft 1.160e+02 6.424e+00 18.061 < 2e-16 \*\*\*  
## typeMulti\_Family 1.508e+04 2.541e+04 0.594 0.553049   
## typeResidential 4.744e+04 1.298e+04 3.654 0.000279 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 64760 on 631 degrees of freedom  
## Multiple R-squared: 0.7694, Adjusted R-squared: 0.7446   
## F-statistic: 30.97 on 68 and 631 DF, p-value: < 2.2e-16

We have removed as many unnecessary variables step by step as possible.

#sac\_p1<-predict(sac\_lm7,sac\_test1)  
#summary(datasac$price)  
#summary(sac\_train1$price)  
#summary(sac\_test1$price)  
#summary(sac\_p1)

While the new model may be considered to be accurate, it introduces new levels for categorical variables (namely city) and hence we can use the model which ommits the categorical variables till the training data set can account for all levels.