Housepriceprediction

Ramana

```
#Loading the dataset
data=read.csv("data.csv")
```

```
#Summary of dataset
summary(data)
```

```
##      date               price           bedrooms       bathrooms
##  Length:4600        Min.   :       0   Min.   :0.000   Min.   :0.000
##  Class :character   1st Qu.:  322875   1st Qu.:3.000   1st Qu.:1.750
##  Mode  :character   Median :  460943   Median :3.000   Median :2.250
##                     Mean   :  551963   Mean   :3.401   Mean   :2.161
##                     3rd Qu.:  654962   3rd Qu.:4.000   3rd Qu.:2.500
##                     Max.   :26590000   Max.   :9.000   Max.   :8.000
##   sqft_living       sqft_lot           floors        waterfront
##  Min.   :  370   Min.   :    638   Min.   :1.000   Min.   :0.000000
##  1st Qu.: 1460   1st Qu.:   5001   1st Qu.:1.000   1st Qu.:0.000000
##  Median : 1980   Median :   7683   Median :1.500   Median :0.000000
##  Mean   : 2139   Mean   :  14852   Mean   :1.512   Mean   :0.007174
##  3rd Qu.: 2620   3rd Qu.:  11001   3rd Qu.:2.000   3rd Qu.:0.000000
##  Max.   :13540   Max.   :1074218   Max.   :3.500   Max.   :1.000000
##       view          condition       sqft_above   sqft_basement
##  Min.   :0.0000   Min.   :1.000   Min.   : 370   Min.   :   0.0
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:1190   1st Qu.:   0.0
##  Median :0.0000   Median :3.000   Median :1590   Median :   0.0
##  Mean   :0.2407   Mean   :3.452   Mean   :1827   Mean   : 312.1
##  3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.:2300   3rd Qu.: 610.0
##  Max.   :4.0000   Max.   :5.000   Max.   :9410   Max.   :4820.0
##     yr_built     yr_renovated       street              city
##  Min.   :1900   Min.   :   0.0   Length:4600        Length:4600
##  1st Qu.:1951   1st Qu.:   0.0   Class :character   Class :character
##  Median :1976   Median :   0.0   Mode  :character   Mode  :character
##  Mean   :1971   Mean   : 808.6
##  3rd Qu.:1997   3rd Qu.:1999.0
##  Max.   :2014   Max.   :2014.0
##    statezip           country
##  Length:4600        Length:4600
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```
head(data)
```

```
##                  date   price bedrooms bathrooms sqft_living sqft_lot floors
## 1 2014-05-02 00:00:00  313000        3      1.50        1340     7912    1.5
## 2 2014-05-02 00:00:00 2384000        5      2.50        3650     9050    2.0
## 3 2014-05-02 00:00:00  342000        3      2.00        1930    11947    1.0
## 4 2014-05-02 00:00:00  420000        3      2.25        2000     8030    1.0
## 5 2014-05-02 00:00:00  550000        4      2.50        1940    10500    1.0
## 6 2014-05-02 00:00:00  490000        2      1.00         880     6380    1.0
##   waterfront view condition sqft_above sqft_basement yr_built yr_renovated
## 1          0    0         3       1340             0     1955         2005
## 2          0    4         5       3370           280     1921            0
## 3          0    0         4       1930             0     1966            0
## 4          0    0         4       1000          1000     1963            0
## 5          0    0         4       1140           800     1976         1992
## 6          0    0         3        880             0     1938         1994
##                     street      city statezip country
## 1     18810 Densmore Ave N Shoreline WA 98133     USA
## 2          709 W Blaine St   Seattle WA 98119     USA
## 3 26206-26214 143rd Ave SE      Kent WA 98042     USA
## 4          857 170th Pl NE  Bellevue WA 98008     USA
## 5        9105 170th Ave NE   Redmond WA 98052     USA
## 6           522 NE 88th St   Seattle WA 98115     USA
```

Data Cleaning

```
sum(is.na(data))
```

```
## [1] 0
```

```
#There are no missing values in the dataset taken
```

```
colSums(data==0)
```

```
##          date         price      bedrooms     bathrooms   sqft_living
##             0            49             2             2             0
##      sqft_lot        floors    waterfront          view     condition
##             0             0          4567          4140             0
##    sqft_above sqft_basement      yr_built  yr_renovated        street
##             0          2745             0          2735             0
##          city      statezip       country
##             0             0             0
```

```
#There are 49 rows with price values as 0 . We need to remove these rows
```

```
sum(is.na(data))
```

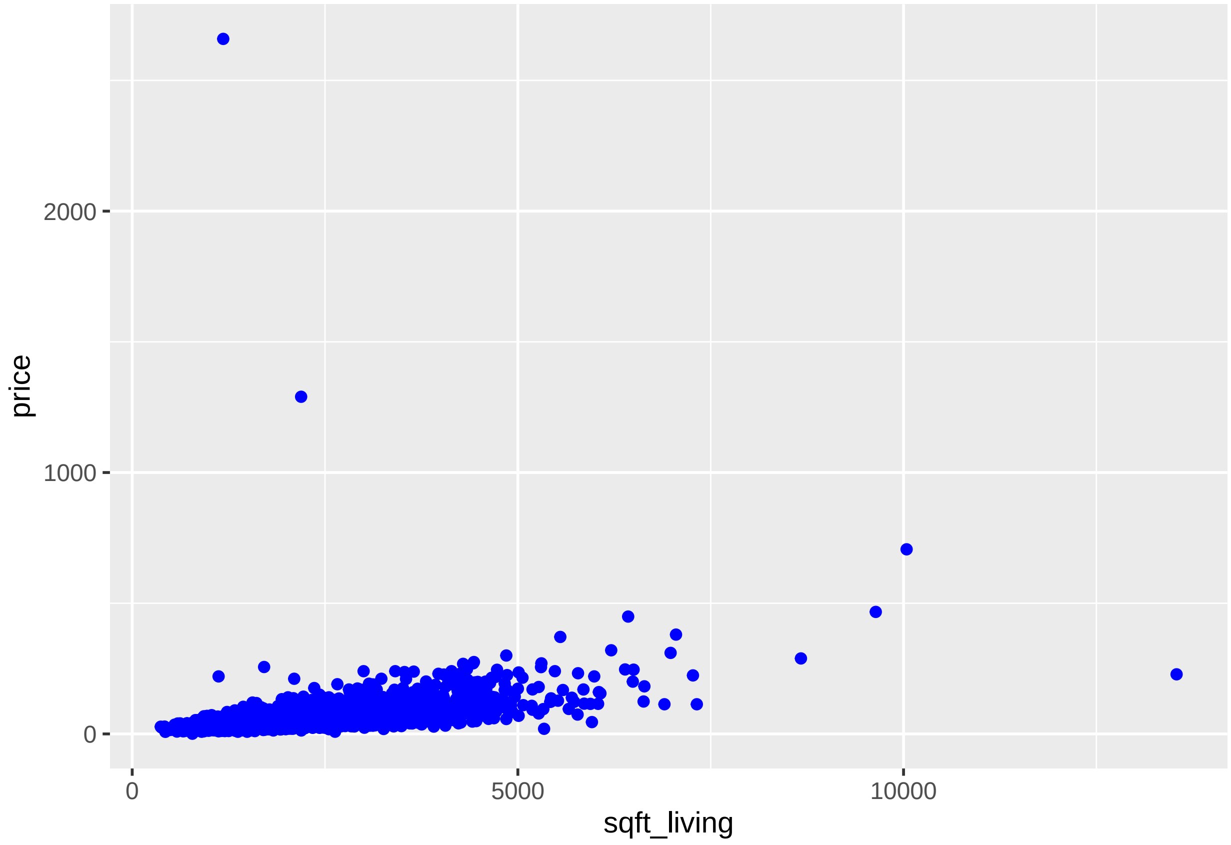
```
## [1] 49
```

```r
#normalizing price values
data$price<-data$price/10000
```

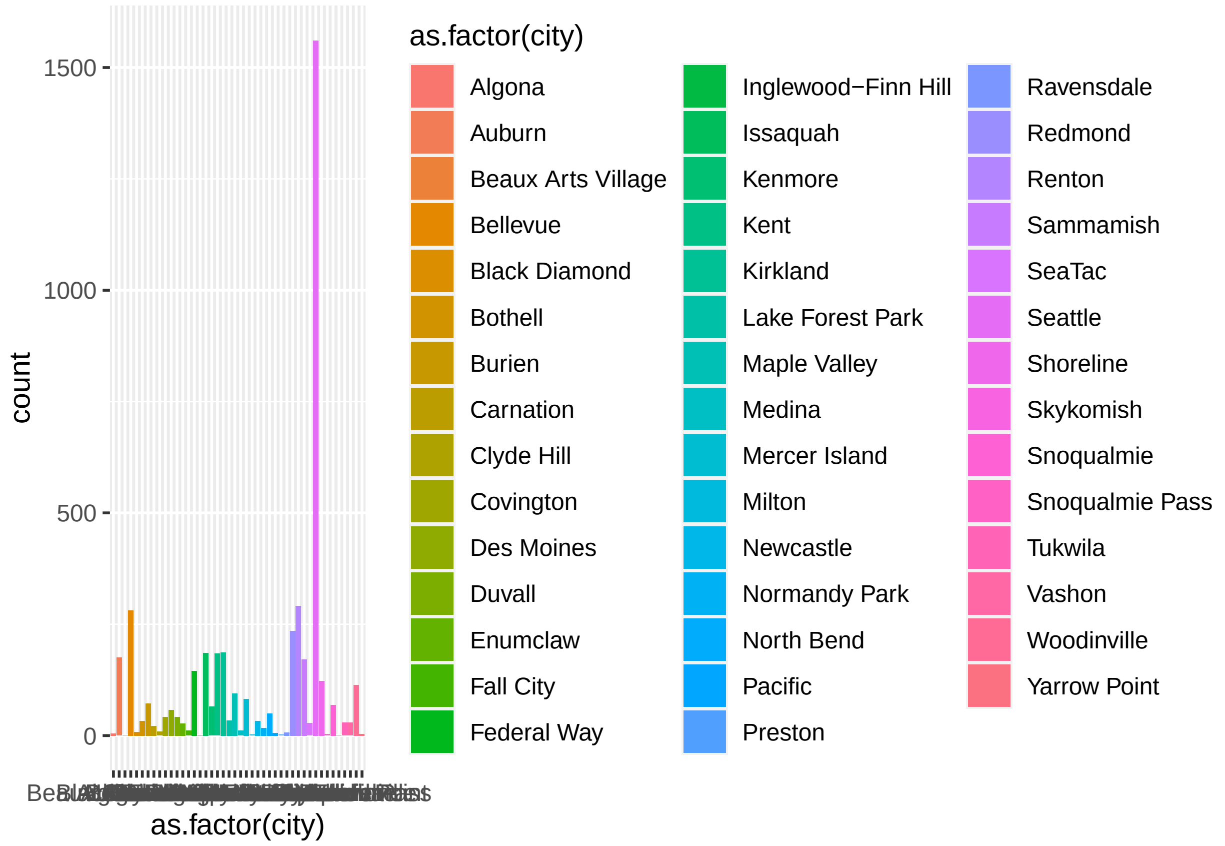
Visualization

```r
library(ggplot2)
```

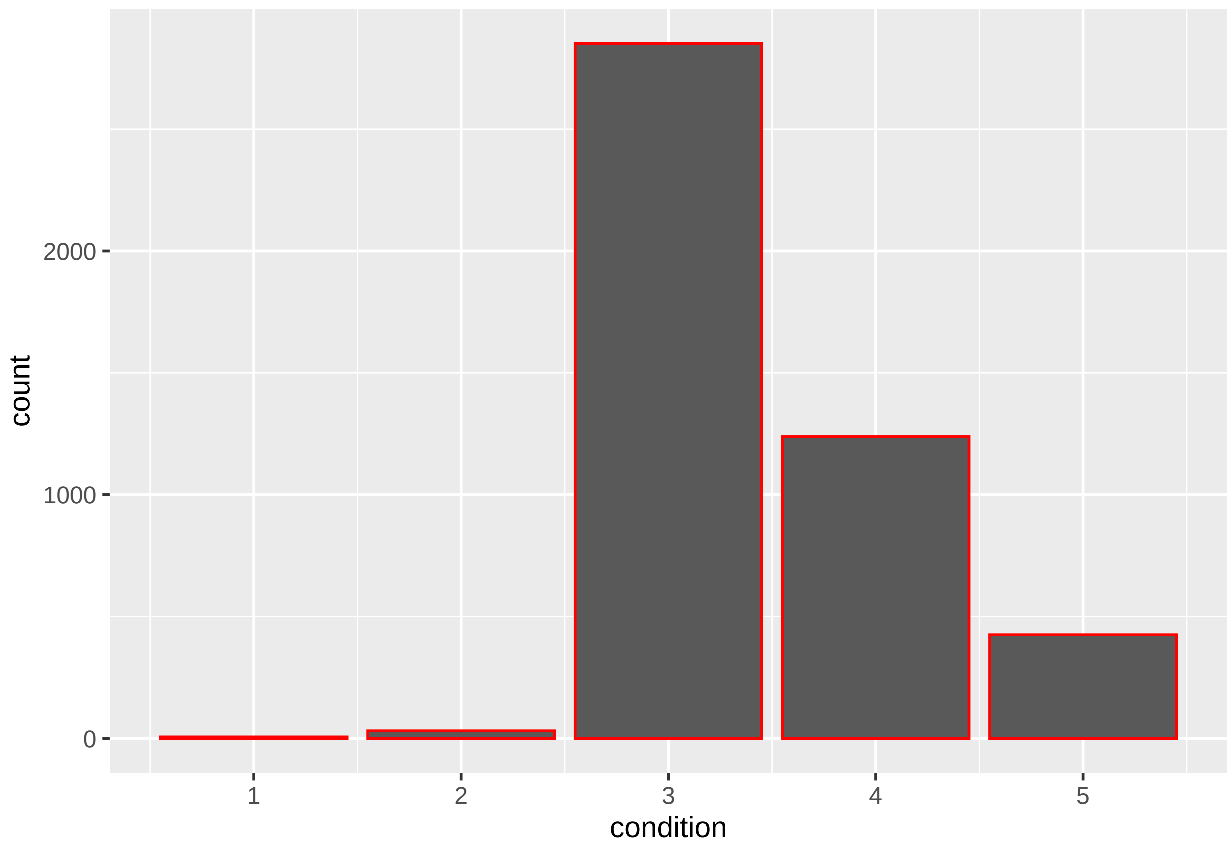
```r
ggplot(data,aes(sqft_living,y=price))+
  geom_point(color="blue")
```



```r
ggplot(data,aes(x=as.factor(city),fill=as.factor(city)))+
  geom_bar()
```

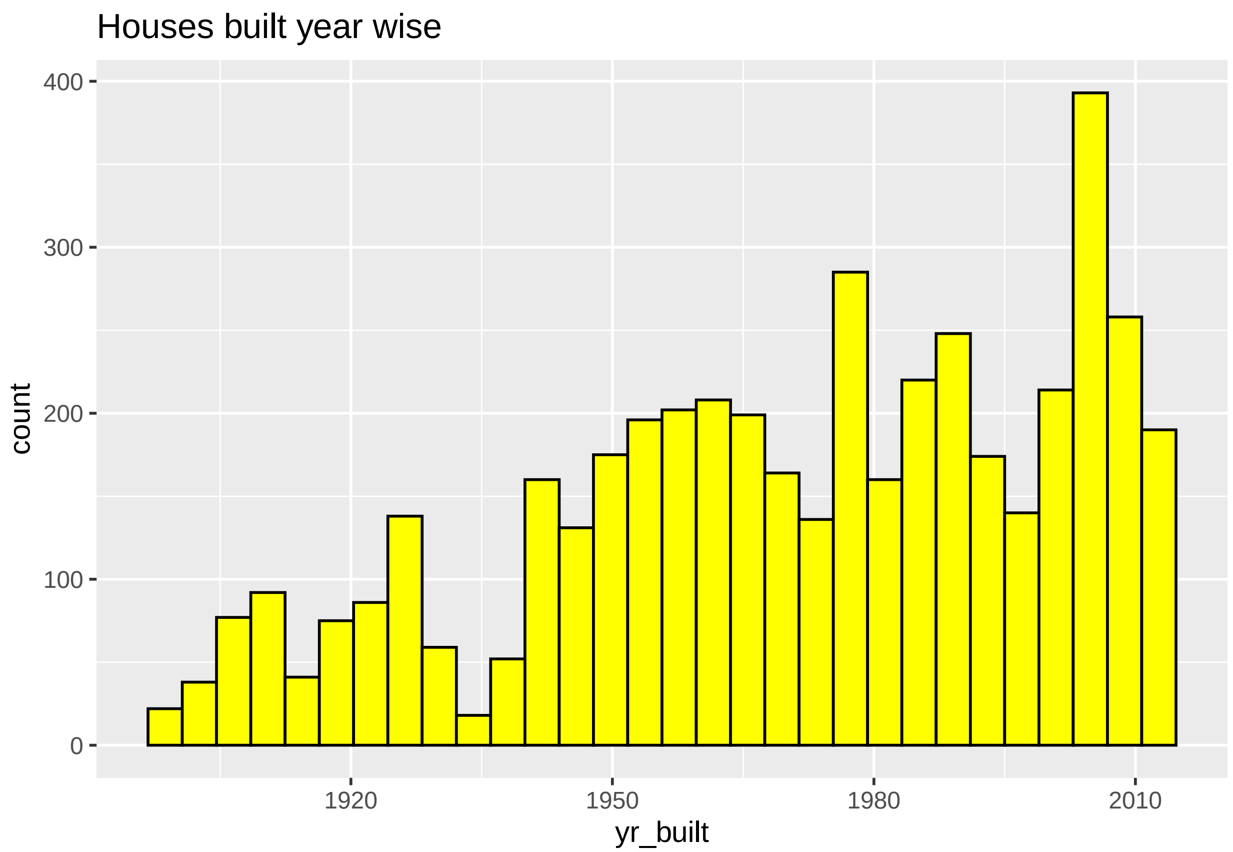


```
ggplot(data,aes(condition))+
  geom_bar(color="red")
```



```
ggplot(data, aes(x=yr_built)) +
  geom_histogram(color="black",fill="yellow")+labs(title = "Houses built year wise")
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



```
cols <- sapply(data, is.numeric)
numeric_data <- data[,cols]

corr_matrix <- cor(numeric_data)
corr_matrix
```

```
##                    price    bedrooms  bathrooms sqft_living      sqft_lot
## price         1.00000000  0.21022759  0.3411259  0.44549371  0.0513473301
## bedrooms      0.21022759  1.00000000  0.5476115  0.59605329  0.0711375009
## bathrooms     0.34112592  0.54761150  1.0000000  0.75721332  0.1093313311
## sqft_living   0.44549371  0.59605329  0.7572133  1.00000000  0.2132675559
## sqft_lot      0.05134733  0.07113750  0.1093313  0.21326756  1.0000000000
## floors        0.15275831  0.17621907  0.4895482  0.34351326  0.0042453119
## waterfront    0.15008259 -0.00552121  0.0633104  0.10775814  0.0174078256
## view          0.24258747  0.11508022  0.2055361  0.30934312  0.0725267888
## condition     0.03889172  0.02301785 -0.1207655 -0.06252868  0.0009289773
## sqft_above    0.38066094  0.48567166  0.6872080  0.87565653  0.2191932396
## sqft_basement 0.21778233  0.33510295  0.2958325  0.44967103  0.0358943837
## yr_built      0.02175681  0.14149772  0.4642395  0.28473287  0.0491634718
## yr_renovated -0.02903374 -0.06221932 -0.2181595 -0.12158915 -0.0210677332
##                    floors   waterfront       view     condition  sqft_above
## price         0.152758308  0.150082587  0.24258747  0.0388917213  0.38066094
## bedrooms      0.176219070 -0.005521210  0.11508022  0.0230178464  0.48567166
## bathrooms     0.489548206  0.063310399  0.20553611 -0.1207654863  0.68720805
## sqft_living   0.343513264  0.107758137  0.30934312 -0.0625286789  0.87565653
## sqft_lot       0.004245312  0.017407826  0.07252679  0.0009289773  0.21919324
## floors         1.000000000  0.015804402  0.03198006 -0.2737856691  0.52221450
## waterfront     0.015804402  1.000000000  0.34757197  0.0061115110  0.07250229
## view           0.031980062  0.347571972  1.00000000  0.0625603230  0.17462889
## condition     -0.273785669  0.006111511  0.06256032  1.0000000000 -0.17654863
## sqft_above     0.522214500  0.072502289  0.17462889 -0.1765486285  1.00000000
## sqft_basement -0.255042074  0.088880236  0.31711737  0.1971442683 -0.03759685
## yr_built       0.466690558 -0.032017059 -0.06634405 -0.3988864426  0.40643626
## yr_renovated  -0.235968880  0.015821146  0.02584557 -0.1844831086 -0.16128119
##               sqft_basement    yr_built yr_renovated
## price            0.21778233  0.02175681  -0.02903374
## bedrooms         0.33510295  0.14149772  -0.06221932
## bathrooms        0.29583249  0.46423948  -0.21815955
## sqft_living      0.44967103  0.28473287  -0.12158915
## sqft_lot         0.03589438  0.04916347  -0.02106773
## floors          -0.25504207  0.46669056  -0.23596888
## waterfront       0.08888024 -0.03201706   0.01582115
## view             0.31711737 -0.06634405   0.02584557
## condition        0.19714427 -0.39888644  -0.18448311
## sqft_above      -0.03759685  0.40643626  -0.16128119
## sqft_basement    1.00000000 -0.16253754   0.04669836
## yr_built        -0.16253754  1.00000000  -0.32293836
## yr_renovated     0.04669836 -0.32293836   1.00000000
```

There is strong positive co relation between sqft_living and price.

Scaling the values

```
data$price=data$price*10000
cols_scale <- c("price","sqft_living", "sqft_lot", "sqft_above", "sqft_basement")
scaled_data <- as.data.frame(scale(data[cols_scale]))

data$price=scaled_data$price
data$sqft_living=scaled_data$sqft_living
data$sqft_above=scaled_data$sqft_above
data$sqft_basement=scaled_data$sqft_basement
data$sqft_lot=scaled_data$sqft_lot
data$yr_built=2023-data$yr_built
head(data)
```

```
##                  date       price bedrooms bathrooms sqft_living    sqft_lot
## 1 2014-05-02 00:00:00 -0.43428432        3      1.50  -0.8288848 -0.19250544
## 2 2014-05-02 00:00:00  3.23815814        5      2.50   1.5875603 -0.16086275
## 3 2014-05-02 00:00:00 -0.38285948        3      2.00  -0.2116976 -0.08031015
## 4 2014-05-02 00:00:00 -0.24454441        3      2.25  -0.1384720 -0.18922439
## 5 2014-05-02 00:00:00 -0.01401929        4      2.50  -0.2012368 -0.12054475
## 6 2014-05-02 00:00:00 -0.12041550        2      1.00  -1.3100817 -0.23510350
##   floors waterfront view condition sqft_above sqft_basement yr_built
## 1    1.5          0    0         3 -0.5643631   -0.67133944       68
## 2    2.0          0    4         5  1.8114261   -0.06526261      102
## 3    1.0          0    0         4  0.1261372   -0.67133944       57
## 4    1.0          0    0         4 -0.9622786    1.49322069       60
## 5    1.0          0    0         4 -0.7984310    1.06030866       47
## 6    1.0          0    0         3 -1.1027193   -0.67133944       85
##   yr_renovated                  street      city statezip country
## 1         2005     18810 Densmore Ave N Shoreline WA 98133     USA
## 2            0          709 W Blaine St   Seattle WA 98119     USA
## 3            0 26206-26214 143rd Ave SE      Kent WA 98042     USA
## 4            0          857 170th Pl NE  Bellevue WA 98008     USA
## 5         1992        9105 170th Ave NE   Redmond WA 98052     USA
## 6         1994           522 NE 88th St   Seattle WA 98115     USA
```

Data splitting into test and train dataset

```
train <- sample(nrow(data), floor(0.7*nrow(data)), replace = FALSE)
test <- setdiff(1:nrow(data), train)
train_data<-data[train, ]
test_data<-data[test, ]
```

```
#The dataset is now split into 30% test data and 70% train data
print(dim(train_data))
```

```
## [1] 3185   18
```

```
dim(test_data)
```

```
## [1] 1366   18
```

1. Linear Regression Model

```
lin_reg<-lm(price~ bedrooms+bathrooms+sqft_living+sqft_lot+floors+ waterfront+view+ condition +sqft_abo
summary(lin_reg)
```

```
##
## Call:
## lm(formula = price ~ bedrooms + bathrooms + sqft_living + sqft_lot +
##     floors + waterfront + view + condition + sqft_above + sqft_basement +
##     yr_built + yr_renovated, data = train_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.672 -0.238 -0.039  0.153 46.659
##
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.792e-01  1.577e-01  -3.039 0.002395 **
## bedrooms      -9.216e-02  2.369e-02  -3.890 0.000102 ***
## bathrooms      1.224e-01  3.928e-02   3.116 0.001847 **
## sqft_living    3.595e-01  4.688e-02   7.668 2.30e-14 ***
## sqft_lot      -4.983e-02  1.664e-02  -2.995 0.002764 **
## floors         4.562e-02  4.206e-02   1.085 0.278153
## waterfront     6.570e-01  2.020e-01   3.252 0.001160 **
## view           1.113e-01  2.508e-02   4.438 9.39e-06 ***
## condition      6.046e-02  2.976e-02   2.032 0.042269 *
## sqft_above     6.507e-02  4.097e-02   1.588 0.112376
## sqft_basement         NA         NA      NA       NA
## yr_built       3.973e-03  7.730e-04   5.141 2.90e-07 ***
## yr_renovated   1.026e-05  1.943e-05   0.528 0.597443
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.928 on 3173 degrees of freedom
## Multiple R-squared:  0.2091, Adjusted R-squared:  0.2064
## F-statistic: 76.27 on 11 and 3173 DF,  p-value: < 2.2e-16
```

```
lin_reg_pred<-predict(lin_reg,newdata=test_data)
lm_rmse <- sqrt(mean((lin_reg_pred - test_data$price)^2))
print(paste("Linear Regression RMSE:", lm_rmse))
```

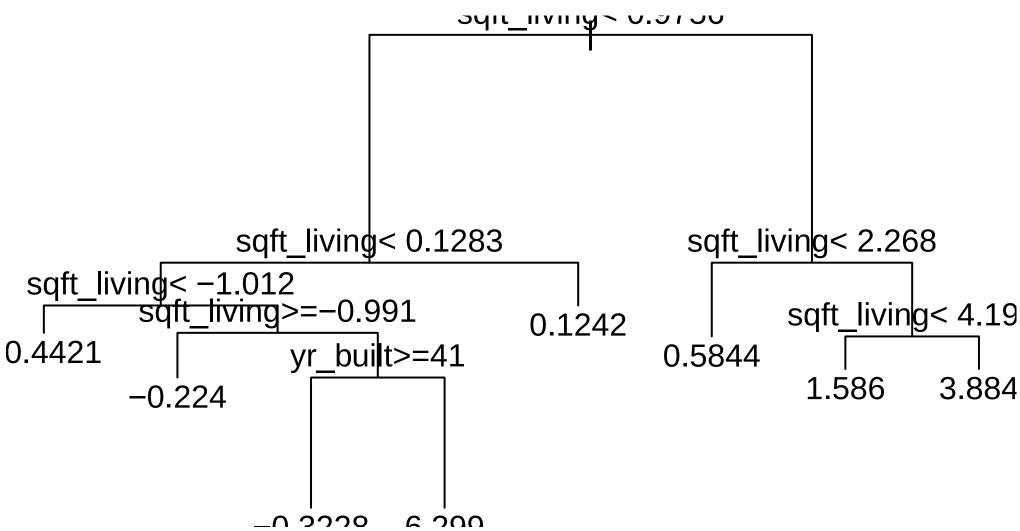
```
## [1] "Linear Regression RMSE: 0.737985376364015"
```

2. Building a Decision Tree Model

```
library(rpart)
tree_model <- rpart(price ~bedrooms+bathrooms+sqft_living+sqft_lot+floors+ waterfront+view+ condition +
printcp(tree_model)
```

```
##
## Regression tree:
## rpart(formula = price ~ bedrooms + bathrooms + sqft_living +
##     sqft_lot + floors + waterfront + view + condition + sqft_above +
##     sqft_basement + yr_built + yr_renovated, data = train_data,
##     method = "anova")
##
## Variables actually used in tree construction:
## [1] sqft_living yr_built
##
## Root node error: 3454.8/3185 = 1.0847
##
## n= 3185
##
##         CP nsplit rel error  xerror    xstd
## 1 0.111926      0   1.00000 1.00079 0.61787
## 2 0.036240      1   0.88807 0.89368 0.62178
## 3 0.028059      2   0.85183 0.86810 0.62159
## 4 0.015985      6   0.73960 0.90512 0.62204
## 5 0.010000      7   0.72361 0.90263 0.62539
```

```
plot(tree_model)
text(tree_model)
```



```r
head(test_data)
```

```
##                    date      price bedrooms bathrooms sqft_living   sqft_lot
## 2  2014-05-02 00:00:00  3.2381581        5      2.50   1.5875603 -0.1608628
## 7  2014-05-02 00:00:00 -0.3952724        2      2.00  -0.8184240 -0.3413206
## 15 2014-05-02 00:00:00  1.1386063        5      2.75   0.8134610 -0.1489064
## 17 2014-05-02 00:00:00 -0.2463177        3      1.50  -0.5882864 -0.2262057
## 18 2014-05-02 00:00:00 -0.3376411        4      3.00   1.0226770 -0.2114410
## 22 2014-05-02 00:00:00 -0.2179454        4      1.00  -0.7138160 -0.1678141
##    floors waterfront view condition sqft_above sqft_basement yr_built
## 2     2.0          0    4         5  1.8114261   -0.06526261      102
## 7     1.0          0    0         3 -0.5526597   -0.67133944       47
## 15    1.5          0    0         3  1.2730699   -0.67133944       84
## 17    1.0          0    0         4 -0.2951850   -0.67133944       67
## 18    2.0          0    0         3  1.5071379   -0.67133944       26
## 22    1.0          0    0         4 -0.4356258   -0.67133944       69
##    yr_renovated             street     city statezip country
## 2             0   709 W Blaine St  Seattle WA 98119     USA
## 7             0 2616 174th Ave NE  Redmond WA 98052     USA
## 15         1969  3534 46th Ave NE  Seattle WA 98105     USA
## 17            0   15424 SE 9th St Bellevue WA 98007     USA
## 18            0 11224 SE 306th Pl   Auburn WA 98092     USA
## 22         1979 3922 154th Ave SE Bellevue WA 98006     USA
```

```
predicted_price <- predict(tree_model, test_data[,-c(1)],method="anova",type="vector")
rmse <- sqrt(mean((predicted_price - test_data$price) ^ 2))
print(paste("Decision Tree RMSE:", rmse))
```

```
## [1] "Decision Tree RMSE: 0.83957861487738"
```