Housepricepredication

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
```

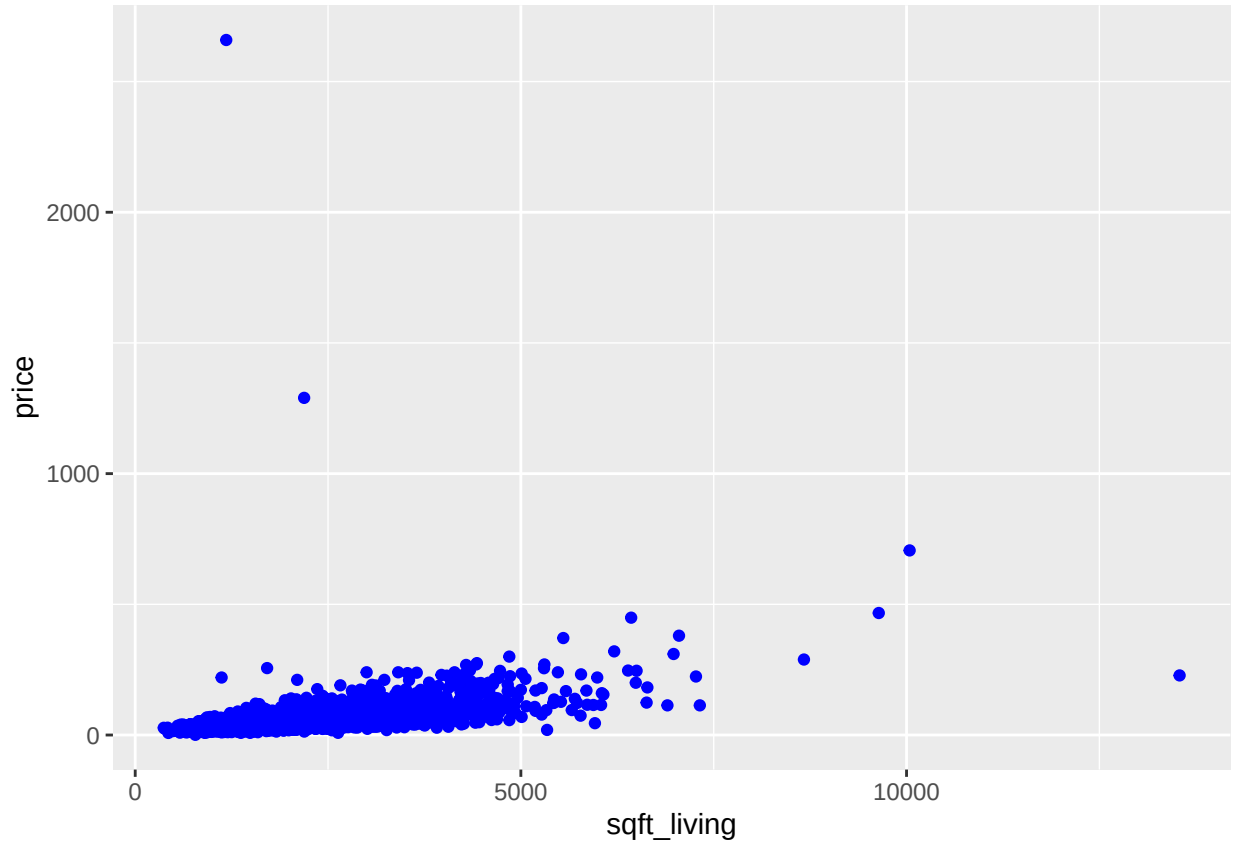
```
#normalizing price values
data$price<-data$price/10000
```

Visualization

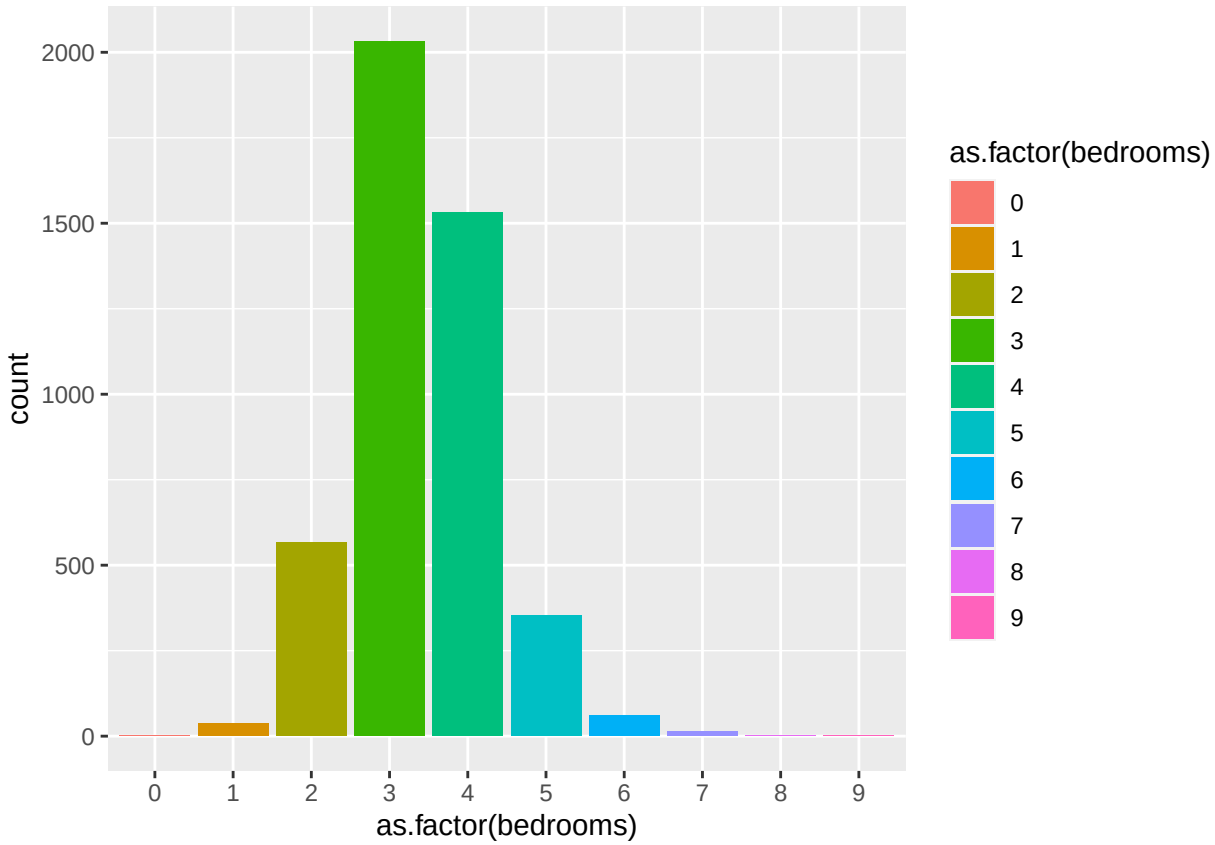
```
library(ggplot2)
```

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

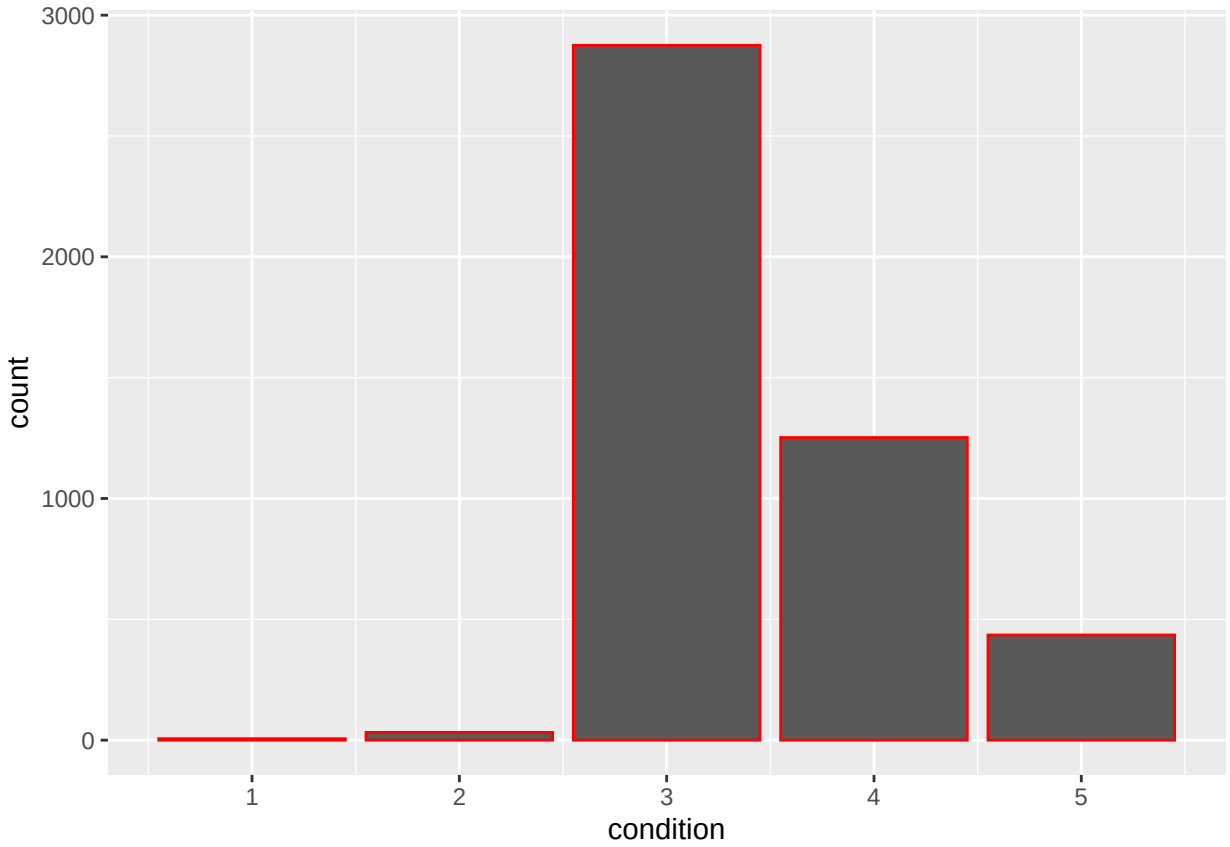
```
## Warning: Removed 49 rows containing missing values (geom_point).
```




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




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



Data splitting into test and train dataset

```
ratio<-floor((nrow(data)/5)*4)
data<-data[sample(nrow(data)),]
train_data<-data[1:ratio,]
test_data<-data[(ratio+1):nrow(data),]
```

```
#The dataset is now split into 20% test data and 80% train data
print(dim(train_data))
```

```
## [1] 3680   18
```

```
dim(test_data)
```

```
## [1] 920  18
```