Housing-Price-Prediction.R

qianna

2021-03-09

```
# Housing Price Prediction
# Author : Na Qian

# Purpose for the analysis : Predict the selling prices of houses in the region:
# You are in market to buy 4 bedrooms, 2 baths and 2 storied houses with approx lot
# size of 5500 SFT in specific area You would like to gather historical sales data and
# analyze for bidding the right price for the house.

#Import the dataset
df<-read.csv('Housing.csv')
head(df)
```

```
##   X price lotsize bedrooms bathrms stories driveway recroom fullbase gashw airco gara
gepl prefarea
## 1 1 42000    5850        3       1       2      yes      no      yes    no    no
1       no
## 2 2 38500    4000        2       1       1      yes      no       no    no    no
0       no
## 3 3 49500    3060        3       1       1      yes      no       no    no    no
0       no
## 4 4 60500    6650        3       1       2      yes     yes       no    no    no
0       no
## 5 5 61000    6360        2       1       1      yes      no       no    no    no
0       no
## 6 6 66000    4160        3       1       1      yes     yes      yes    no   yes
0       no
```

```
any(is.na(df))
```

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

```
summary(df)
```

```
##        X             price           lotsize         bedrooms        bathrms
stories
##  Min.   :  1.0   Min.   : 25000   Min.   : 1650   Min.   :1.000   Min.   :1.000   Mi
n.   :1.000
##  1st Qu.:137.2   1st Qu.: 49125   1st Qu.: 3600   1st Qu.:2.000   1st Qu.:1.000   1st
Qu.:1.000
##  Median :273.5   Median : 62000   Median : 4600   Median :3.000   Median :1.000   Med
ian :2.000
##  Mean   :273.5   Mean   : 68122   Mean   : 5150   Mean   :2.965   Mean   :1.286   Mea
n   :1.808
##  3rd Qu.:409.8   3rd Qu.: 82000   3rd Qu.: 6360   3rd Qu.:3.000   3rd Qu.:2.000   3rd
Qu.:2.000
##  Max.   :546.0   Max.   :190000   Max.   :16200   Max.   :6.000   Max.   :4.000   Ma
x.   :4.000
##  driveway  recroom   fullbase  gashw     airco        garagepl      prefarea
##  no : 77   no :449   no :355   no :521   no :373   Min.   :0.0000   no :418
##  yes:469   yes: 97   yes:191   yes: 25   yes:173   1st Qu.:0.0000   yes:128
##                                                   Median :0.0000
##                                                   Mean   :0.6923
##                                                   3rd Qu.:1.0000
##                                                   Max.   :3.0000
```

```
str(df)
```

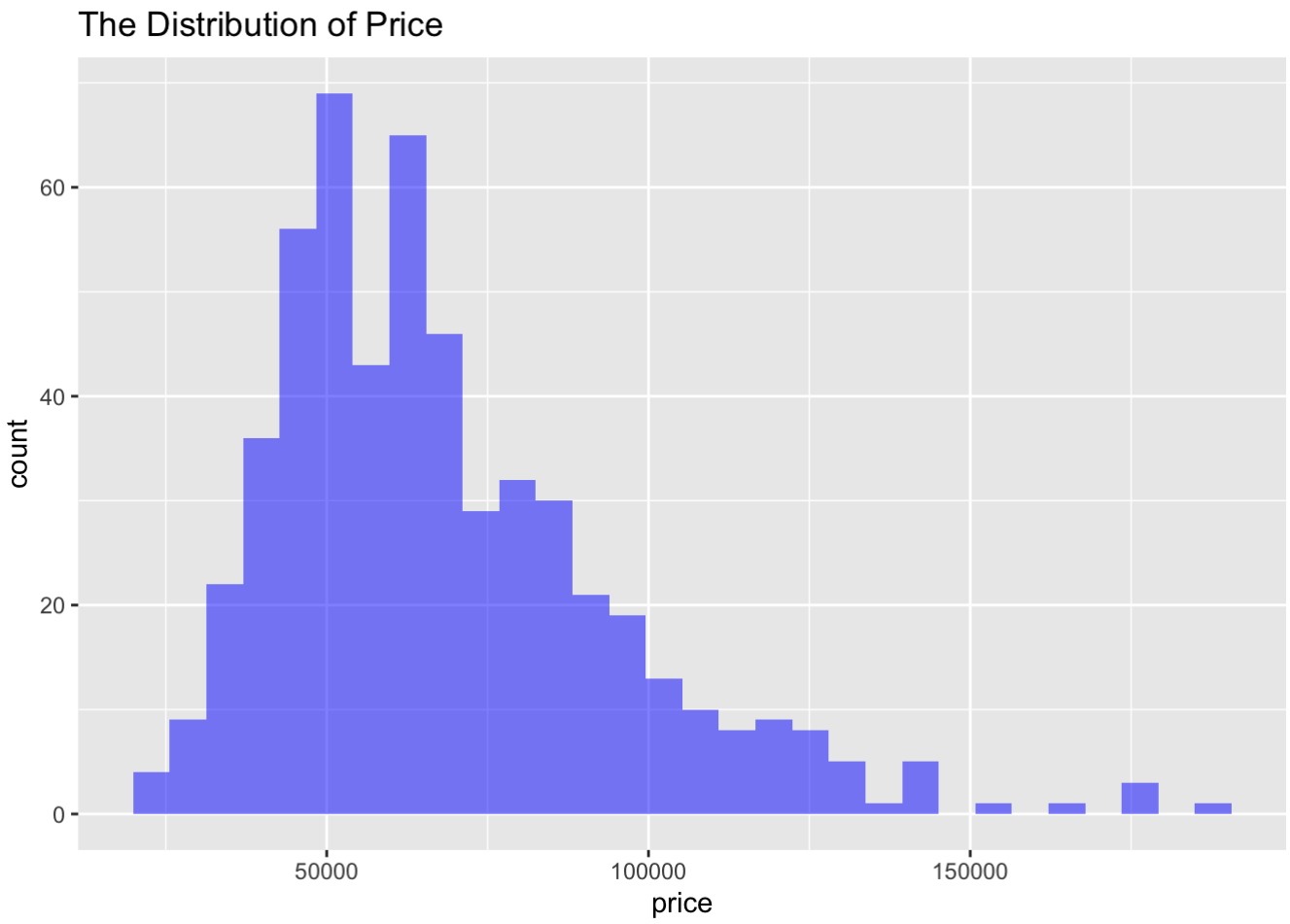
```
## 'data.frame':    546 obs. of  13 variables:
##  $ X       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ price   : num  42000 38500 49500 60500 61000 66000 66000 69000 83800 88500 ...
##  $ lotsize : int  5850 4000 3060 6650 6360 4160 3880 4160 4800 5500 ...
##  $ bedrooms: int  3 2 3 3 2 3 3 3 3 3 ...
##  $ bathrms : int  1 1 1 1 1 1 2 1 1 2 ...
##  $ stories : int  2 1 1 2 1 1 2 3 1 4 ...
##  $ driveway: Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ recroom : Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 2 2 ...
##  $ fullbase: Factor w/ 2 levels "no","yes": 2 1 1 1 1 2 2 1 2 1 ...
##  $ gashw   : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ airco   : Factor w/ 2 levels "no","yes": 1 1 1 1 1 2 1 1 1 2 ...
##  $ garagepl: int  1 0 0 0 0 0 2 0 0 1 ...
##  $ prefarea: Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```
#convert categorical data into dummy data
df$driveway <- ifelse(df$driveway == 'yes', 1, 0)
df$recroom <- ifelse(df$recroom == 'yes', 1, 0)
df$fullbase <- ifelse(df$fullbase == 'yes', 1, 0)
df$gashw <- ifelse(df$gashw == 'yes', 1, 0)
df$airco <- ifelse(df$airco == 'yes', 1, 0)
df$prefarea <- ifelse(df$prefarea == 'yes', 1, 0)
str(df)
```

```
## 'data.frame':    546 obs. of  13 variables:
##  $ X       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ price   : num  42000 38500 49500 60500 61000 66000 66000 69000 83800 88500 ...
##  $ lotsize : int  5850 4000 3060 6650 6360 4160 3880 4160 4800 5500 ...
##  $ bedrooms: int  3 2 3 3 2 3 3 3 3 3 ...
##  $ bathrms : int  1 1 1 1 1 1 2 1 1 2 ...
##  $ stories : int  2 1 1 2 1 1 2 3 1 4 ...
##  $ driveway: num  1 1 1 1 1 1 1 1 1 1 ...
##  $ recroom : num  0 0 0 1 0 1 0 0 1 1 ...
##  $ fullbase: num  1 0 0 0 0 1 1 0 1 0 ...
##  $ gashw   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ airco   : num  0 0 0 0 0 1 0 0 0 1 ...
##  $ garagepl: int  1 0 0 0 0 0 2 0 0 1 ...
##  $ prefarea: num  0 0 0 0 0 0 0 0 0 0 ...
```

```
library("ggplot2")

# the distribution of house price
pl<-ggplot(df,aes(x=price)) + geom_histogram(fill='blue',alpha=0.5)
print(pl+ggtitle('The Distribution of Price'))
```

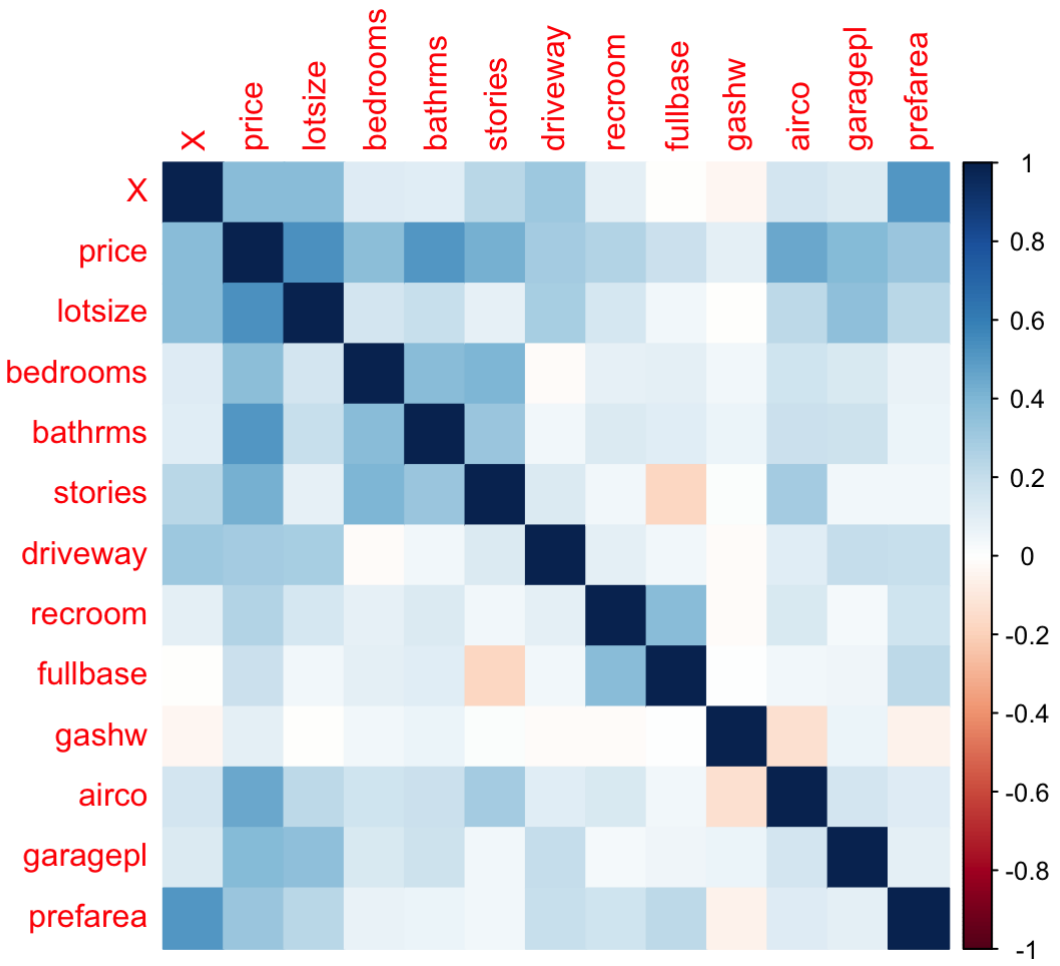
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



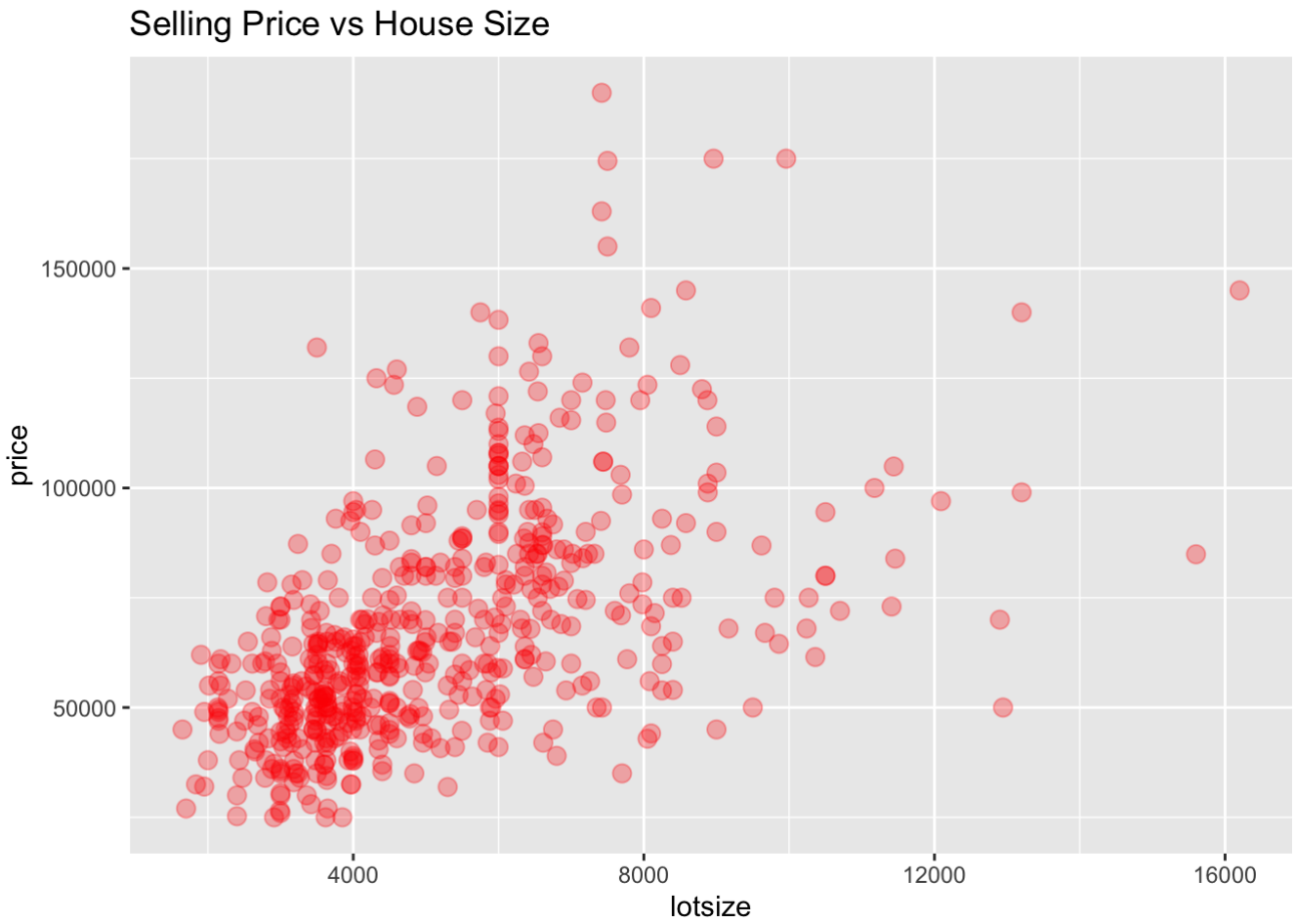
```
install.packages('corrplot')
```

```
## Error in install.packages : Updating loaded packages
```

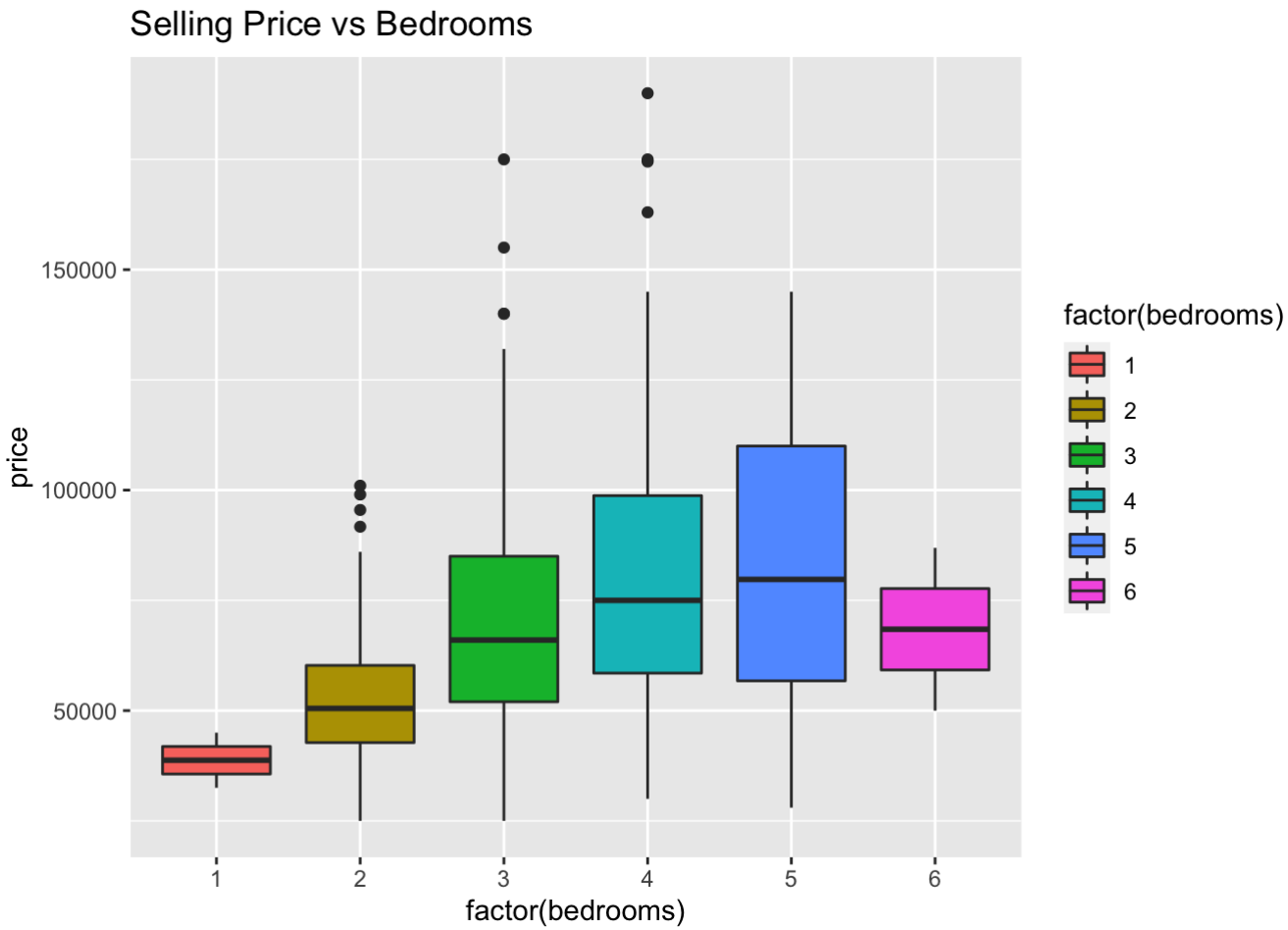
```
library(corrplot)
corr.data<-cor(df)
corrplot(corr.data, method='color')
```



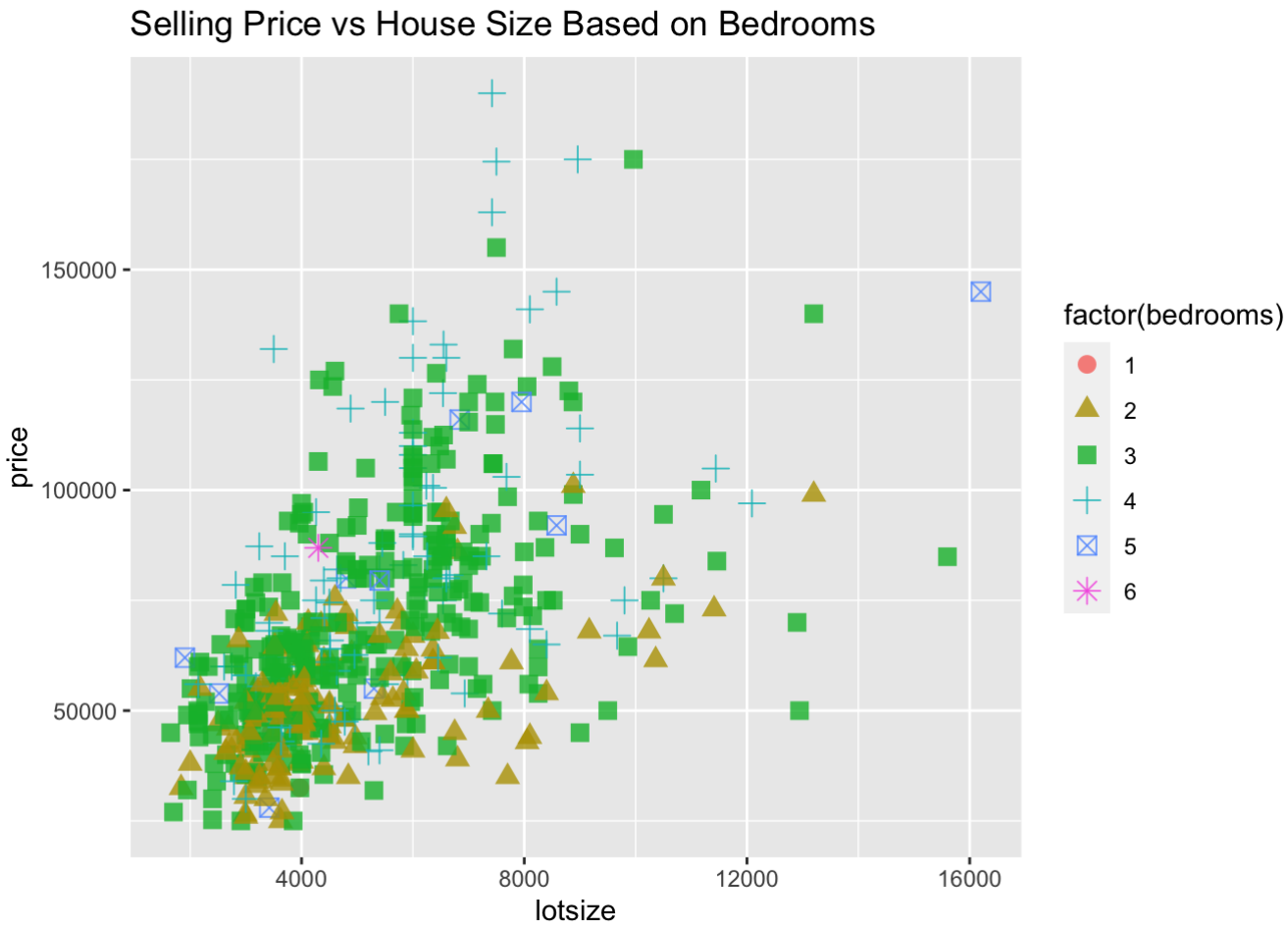

```
# price vs size
pl<-ggplot(df,aes(x=lotsize,y=price)) + geom_point(alpha=0.3, color='red',size=3)
print(pl+ggtitle('Selling Price vs House Size'))
```



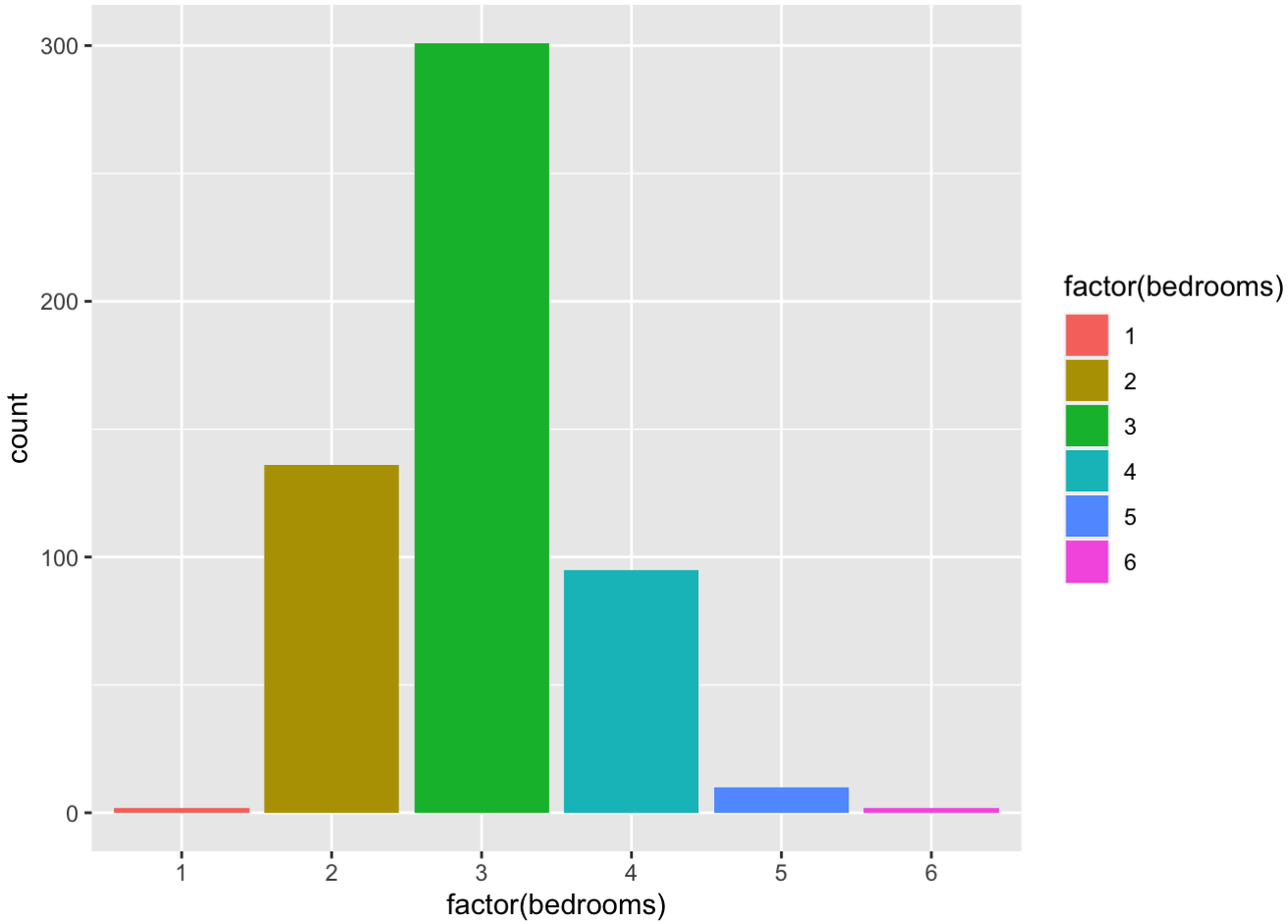

```
#price vs bedrooms
pl<-ggplot(df, aes(x=factor(bedrooms), y=price)) + geom_boxplot(aes(fill=factor(bedroom
s)))
print(pl+ggtitle('Selling Price vs Bedrooms'))
```



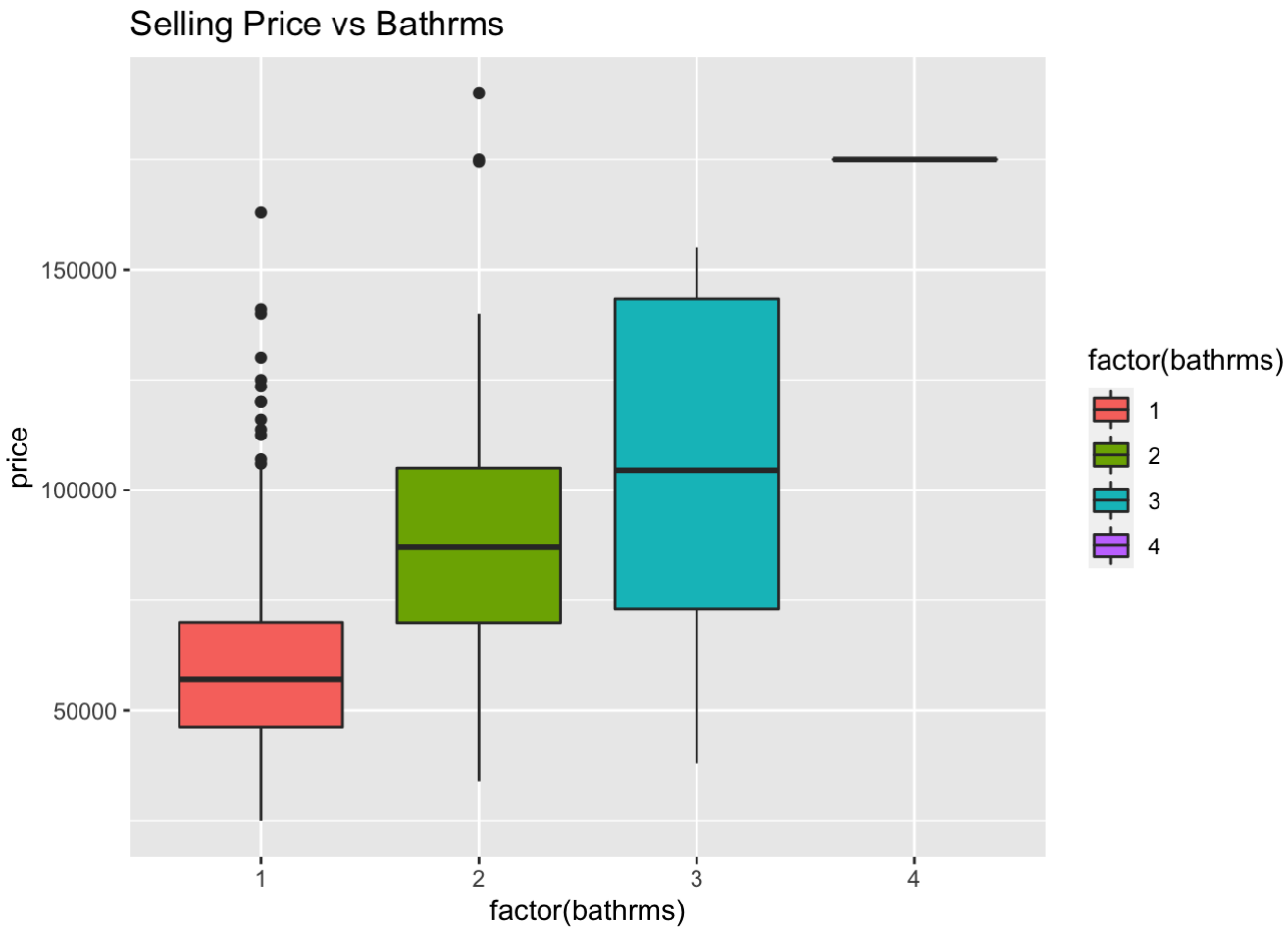

```
#price vs size but consider the bedrooms
pl<-ggplot(df,aes(x=lotsize,y=price)) + geom_point(alpha=0.8, size=3,
                    aes(shape=factor(bedrooms),color=factor(bedrooms)))
print(pl+ggtitle('Selling Price vs House Size Based on Bedrooms'))
```



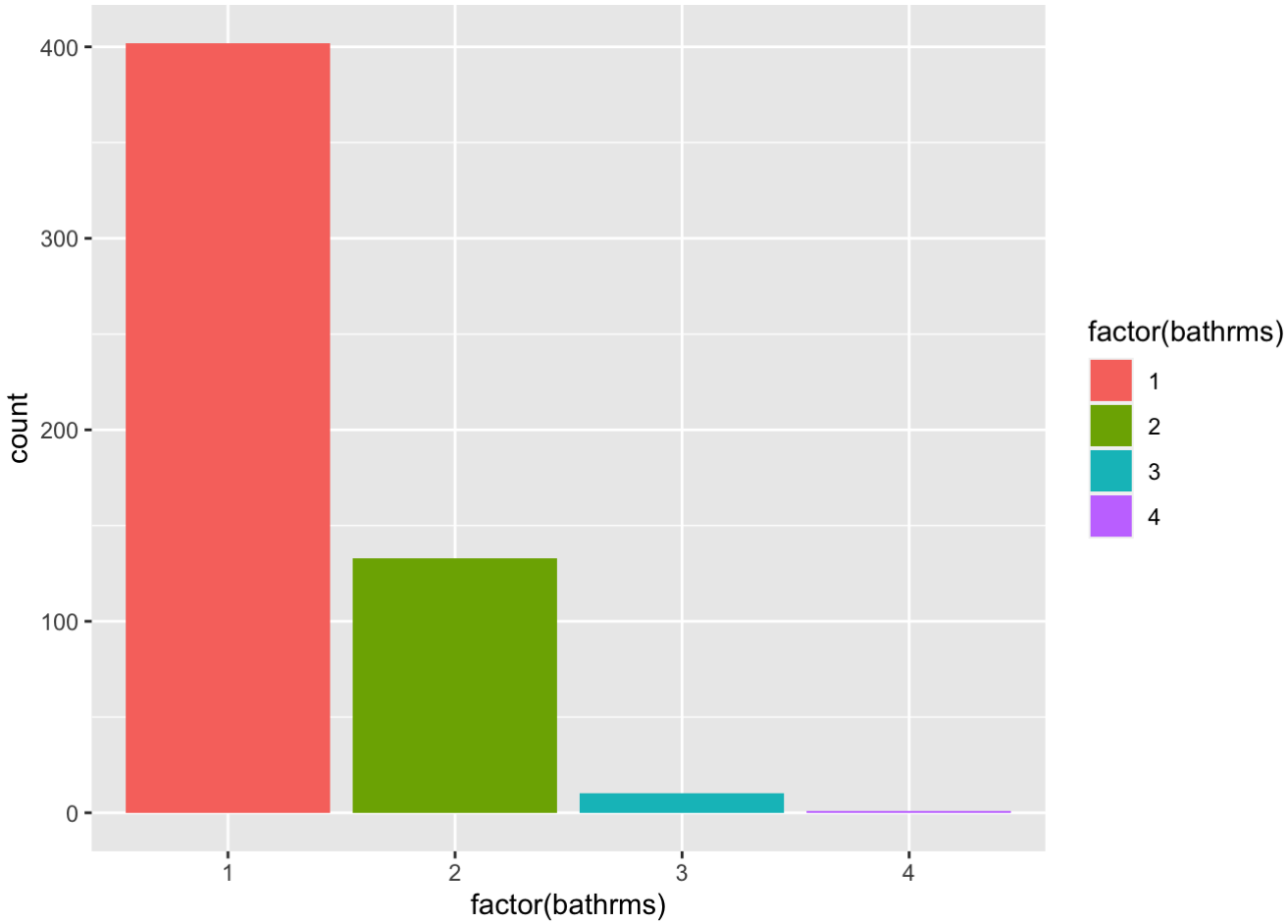

```
#bedroom count
pl<-ggplot(df,aes(x=factor(bedrooms)))+geom_bar(aes(fill=factor(bedrooms)))
print(pl)
```



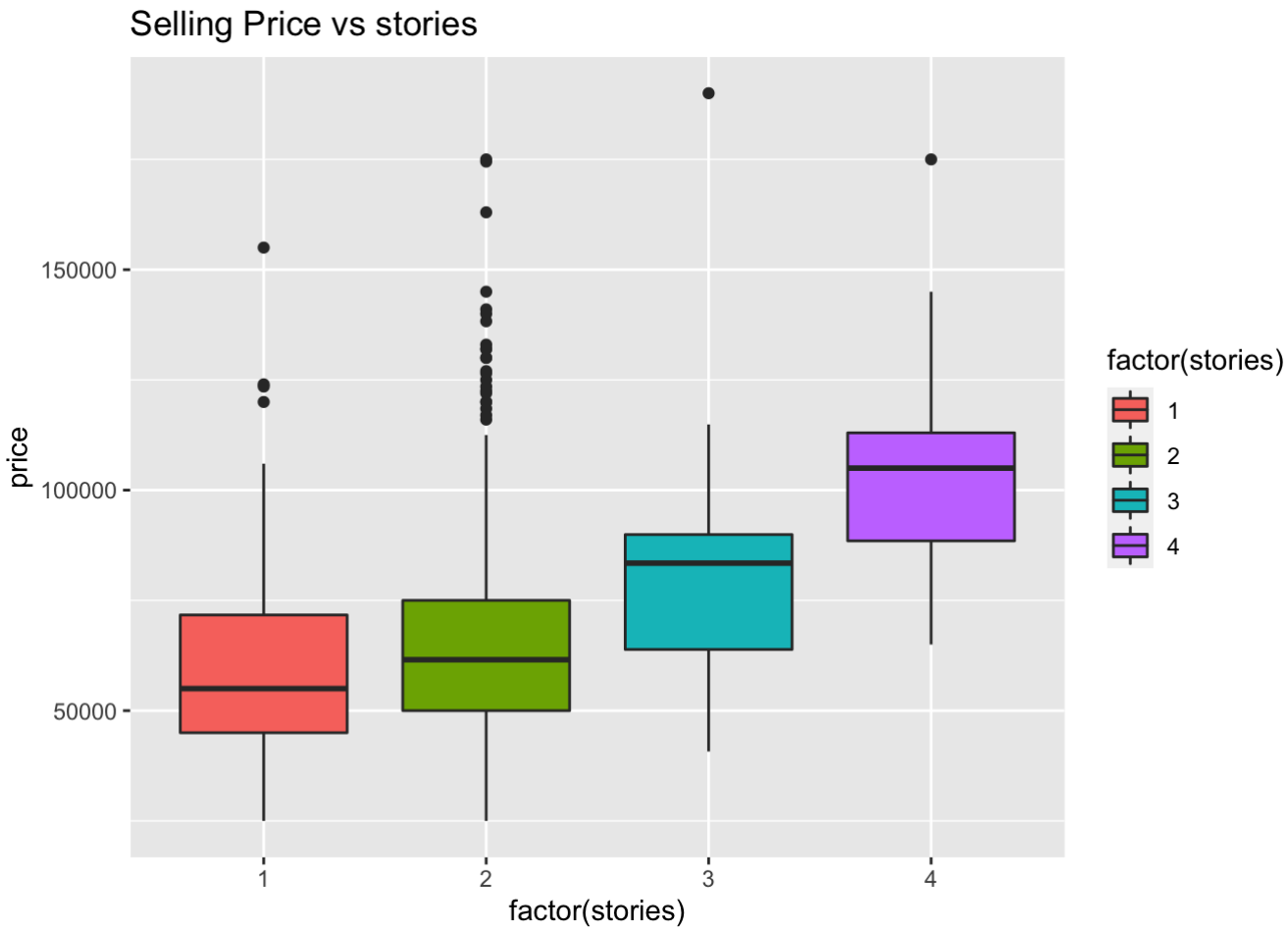

```
#price vs bathrooms

pl<-ggplot(df, aes(x=factor(bathrms), y=price)) + geom_boxplot(aes(fill=factor(bathrm
s)))
print(pl+ggtitle('Selling Price vs Bathrms'))
```




```
#bathroom count
pl<-ggplot(df,aes(x=factor(bathrms)))+geom_bar(aes(fill=factor(bathrms)))
print(pl)
```



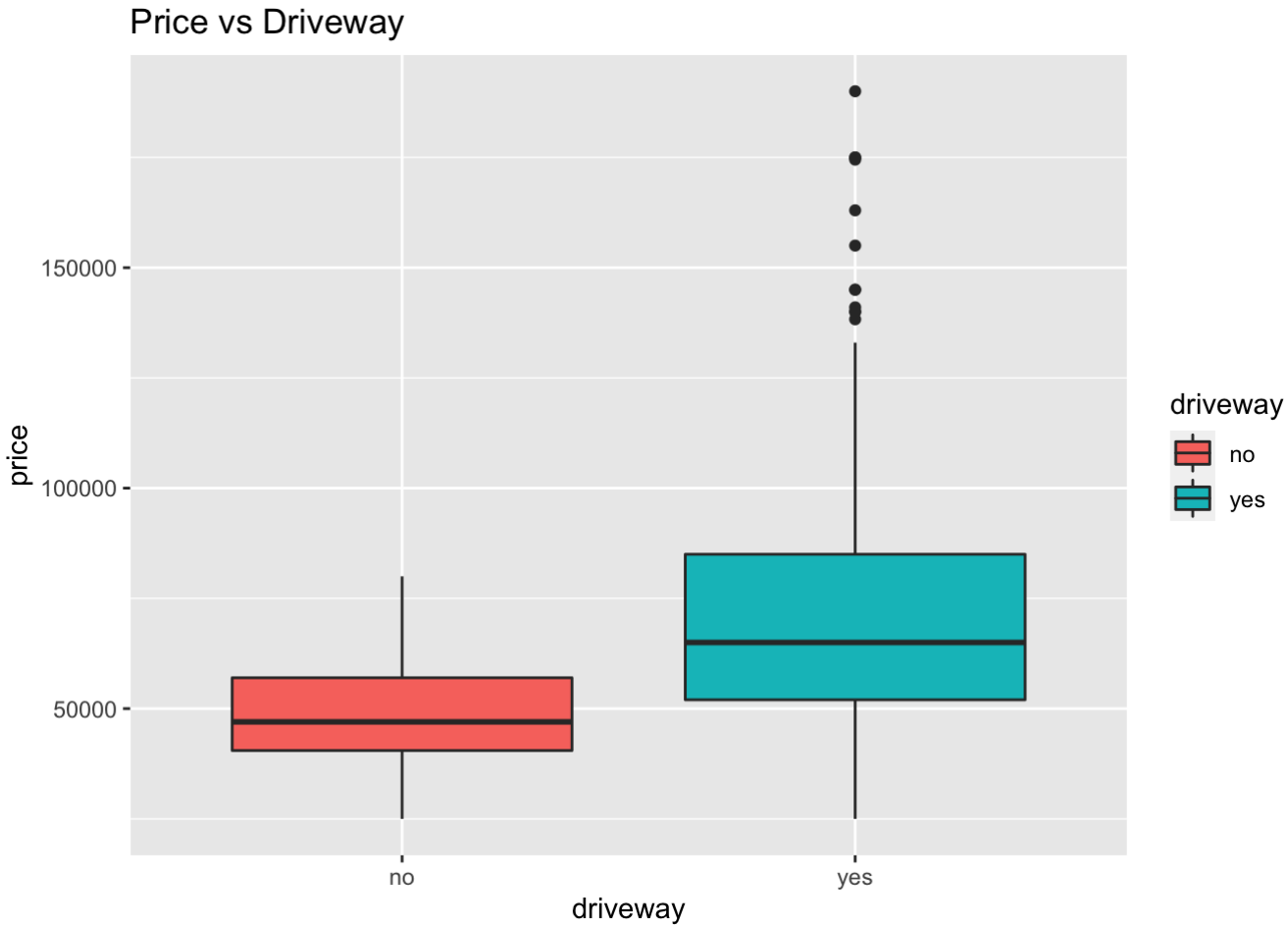

```
#price vs stories
pl<-ggplot(df, aes(x=factor(stories), y=price)) + geom_boxplot(aes(fill=factor(storie
s)))
print(pl+ggtitle('Selling Price vs stories'))
```



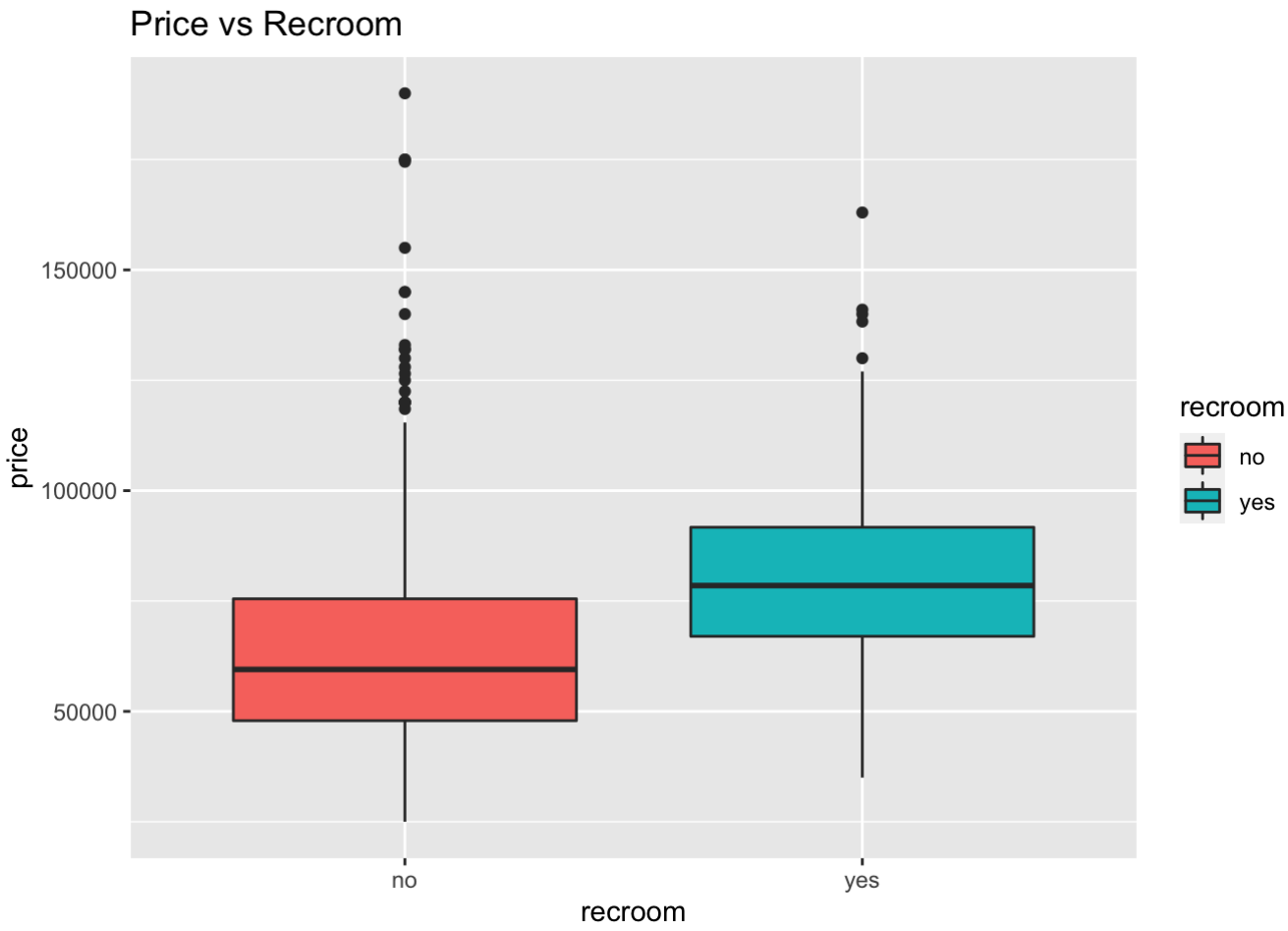

```
#categorical data
df2<-read.csv('Housing.csv')
head(df2)
```

```
##   X price lotsize bedrooms bathrms stories driveway recroom fullbase gashw airco gara
gepl prefarea
## 1 1 42000    5850        3       1       2      yes      no      yes    no    no
1       no
## 2 2 38500    4000        2       1       1      yes      no       no    no    no
0       no
## 3 3 49500    3060        3       1       1      yes      no       no    no    no
0       no
## 4 4 60500    6650        3       1       2      yes     yes       no    no    no
0       no
## 5 5 61000    6360        2       1       1      yes      no       no    no    no
0       no
## 6 6 66000    4160        3       1       1      yes     yes      yes    no   yes
0       no
```

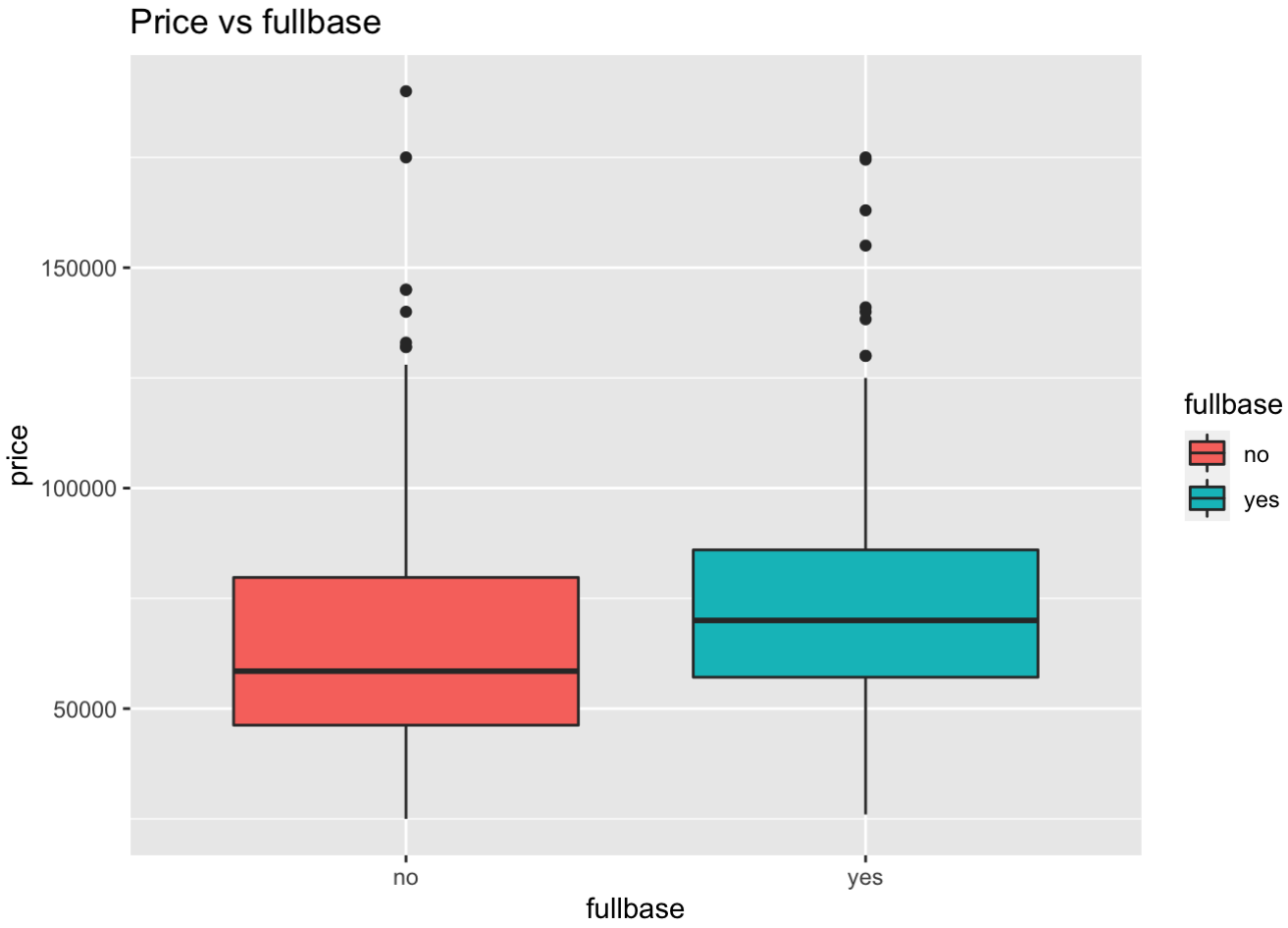
```
# driveway
pl<-ggplot(df2,aes(x=driveway,y=price))+geom_boxplot(aes(fill=driveway))
print(pl+ggtitle('Price vs Driveway'))
```



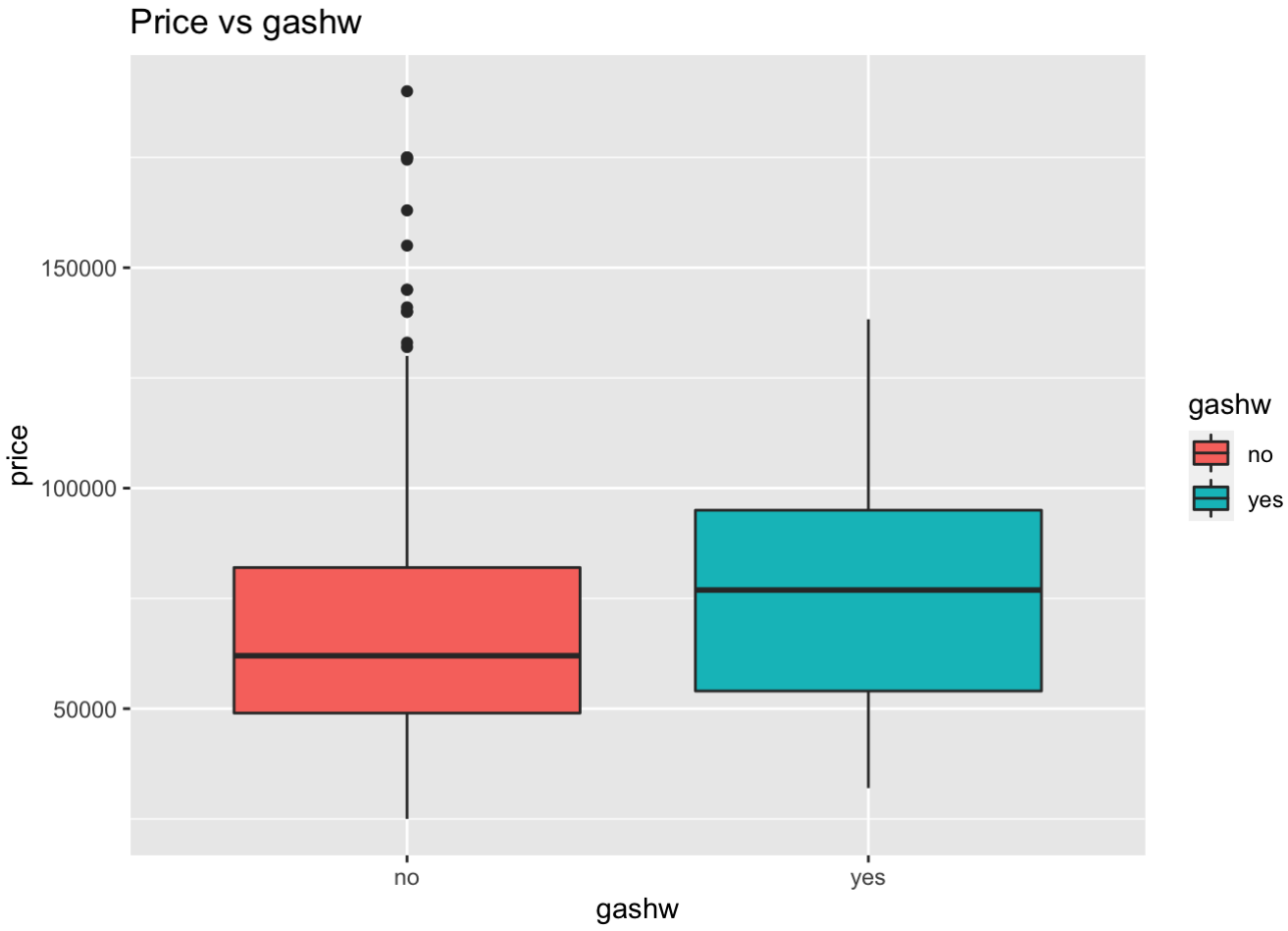

```
#recroom
pl<-ggplot(df2,aes(x=recroom,y=price))+geom_boxplot(aes(fill=recroom))
print(pl+ggtitle('Price vs Recroom'))
```



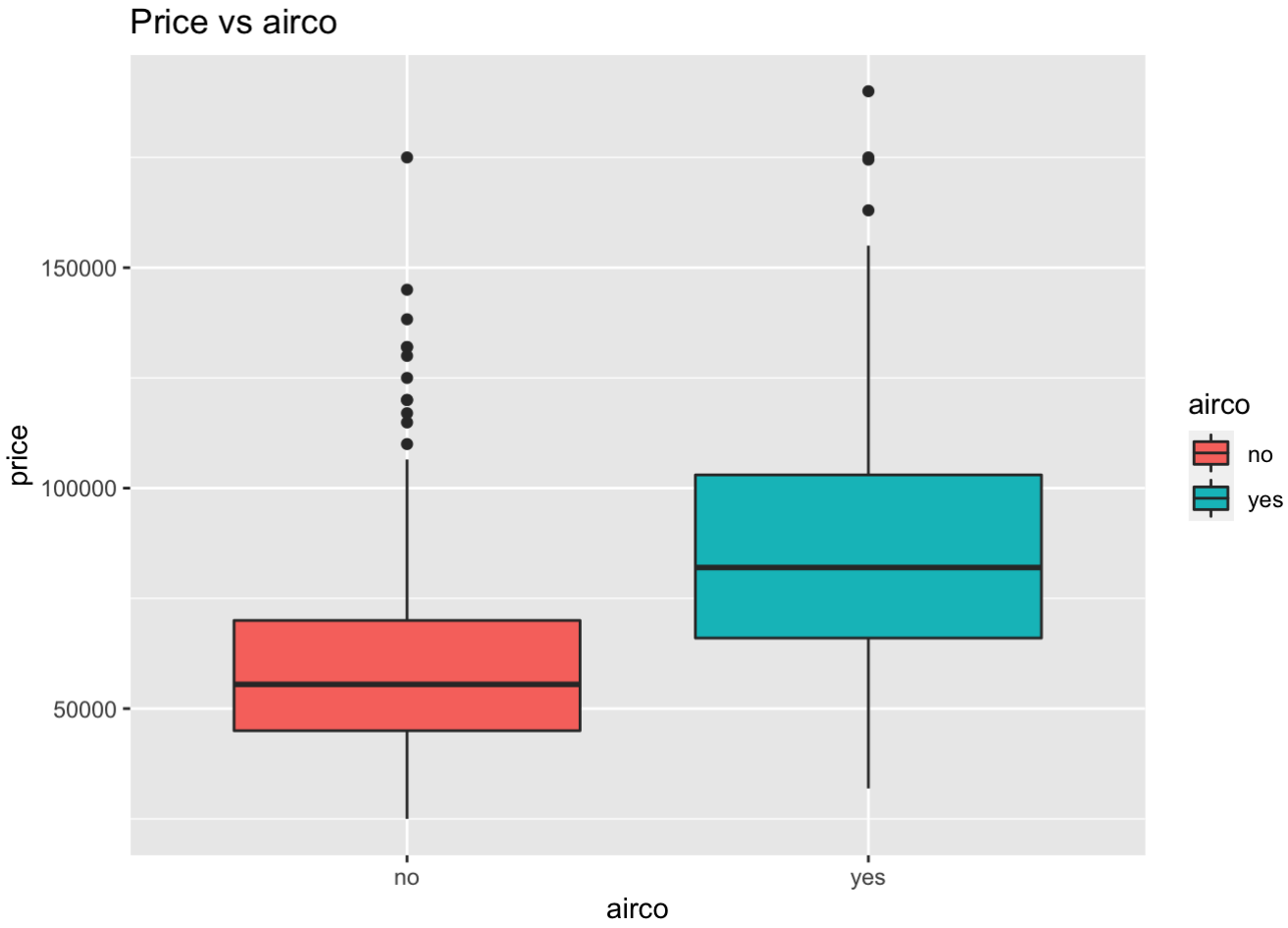

```
#fullbase
pl<-ggplot(df2,aes(x=fullbase,y=price))+geom_boxplot(aes(fill=fullbase))
print(pl+ggtitle('Price vs fullbase'))
```



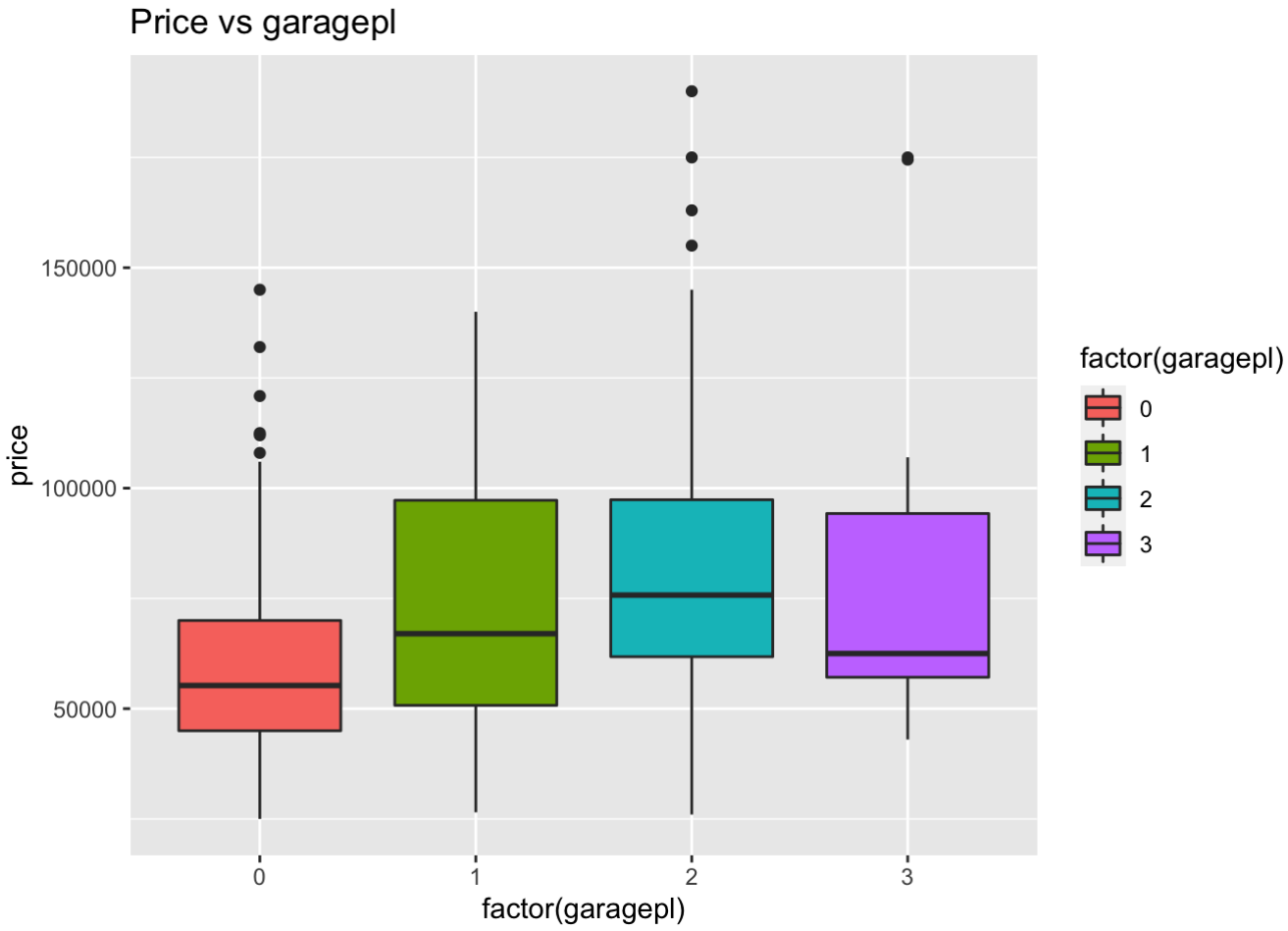

```
#gashw
pl<-ggplot(df2,aes(x=gashw,y=price))+geom_boxplot(aes(fill=gashw))
print(pl+ggtitle('Price vs gashw'))
```



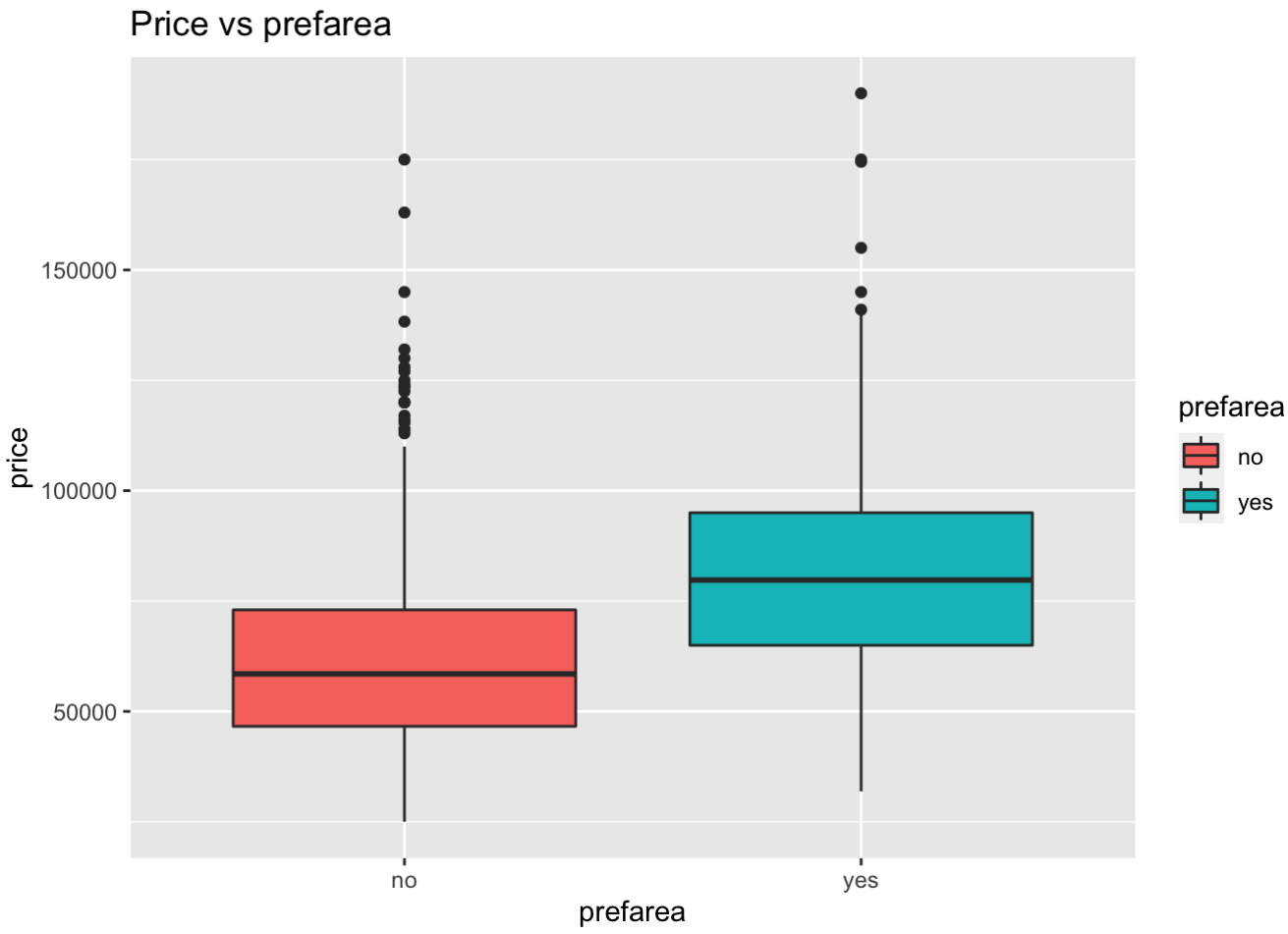

```
#airco
pl<-ggplot(df2,aes(x=airco,y=price))+geom_boxplot(aes(fill=airco))
print(pl+ggtitle('Price vs airco'))
```




```
#garagepl convert to categorical

pl<-ggplot(df2,aes(x=factor(garagepl),y=price))
pl2<-pl + geom_boxplot(aes(fill=factor(garagepl)))
print(pl2+ggtitle('Price vs garagepl'))
```




```
#prefarea
pl<-ggplot(df2,aes(x=prefarea,y=price))+geom_boxplot(aes(fill=prefarea))
print(pl+ggtitle('Price vs prefarea'))
```



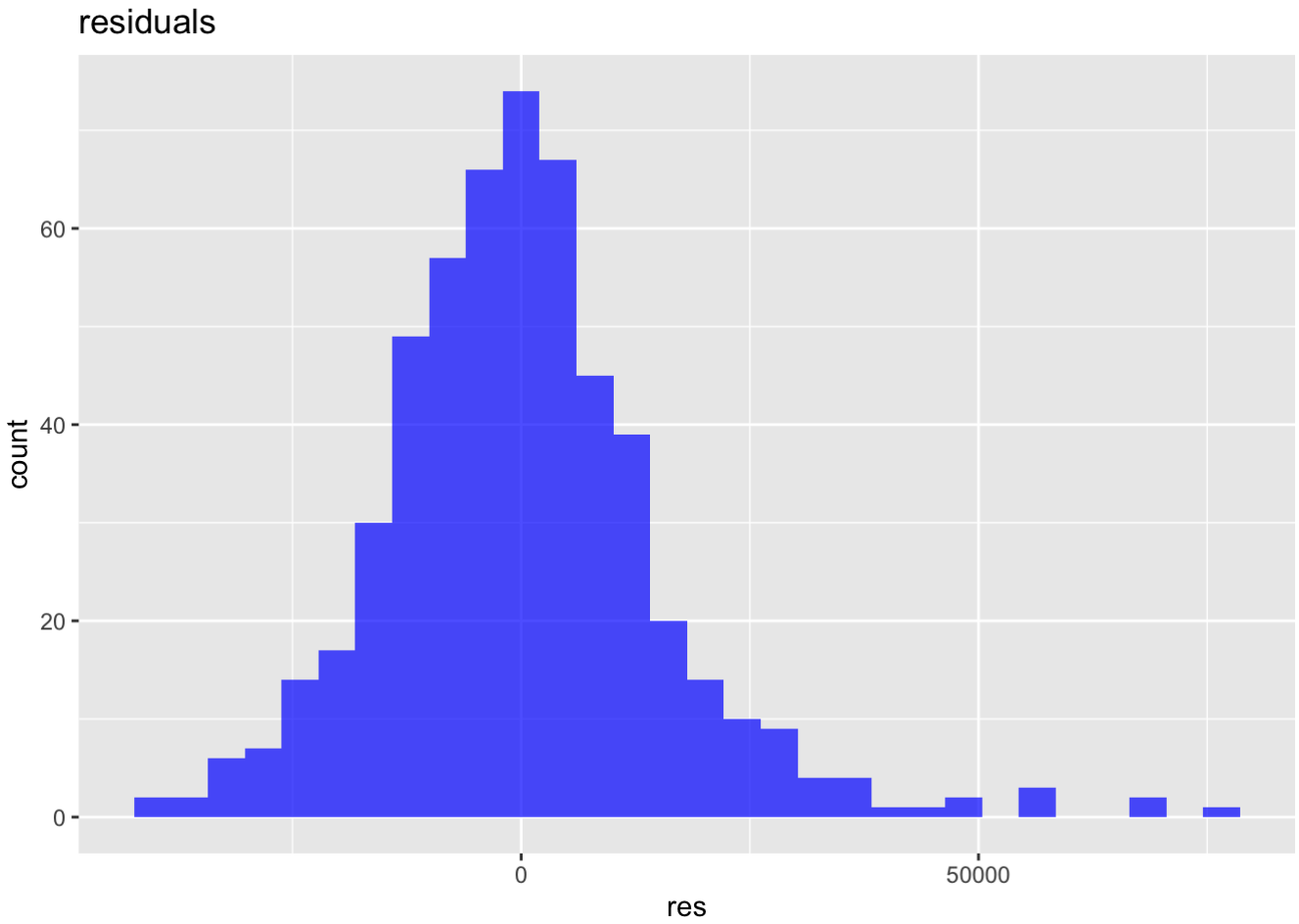

```
# split the data
library(caTools)
sample<-sample.split(df$price,SplitRatio = 0.7)
train<-subset(df,sample=TRUE)
test<-subset(df,sample=FALSE)

# fit the model
model<-lm(price ~. , data = train)
print(summary(model))
```

```
##
## Call:
## lm(formula = price ~ ., data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41272  -9312   -885   7346  75628
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4447.4245  3422.1157  -1.300 0.194296
## X               6.9709     5.4090   1.289 0.198036
## lotsize         3.4313     0.3613   9.498  < 2e-16 ***
## bedrooms     1840.8782  1046.3755   1.759 0.079102 .
## bathrms     14353.6882  1489.0661   9.639  < 2e-16 ***
## stories      6348.4461   938.7632   6.763 3.57e-11 ***
## driveway     6224.2509  2075.3841   2.999 0.002834 **
## recroom      4484.9187  1898.8932   2.362 0.018542 *
## fullbase     5671.0707  1596.0875   3.553 0.000414 ***
## gashw       12845.7083  3215.6268   3.995 7.39e-05 ***
## airco       12634.5979  1554.0602   8.130 3.02e-15 ***
## garagepl     4278.5911   840.4329   5.091 4.95e-07 ***
## prefarea     8192.0506  1901.8780   4.307 1.97e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15410 on 533 degrees of freedom
## Multiple R-squared:  0.6741, Adjusted R-squared:  0.6668
## F-statistic: 91.89 on 12 and 533 DF,  p-value: < 2.2e-16
```

```
#plot residuals
res<-residuals(model)
res<-as.data.frame(res)
head(res)
```

```
##           res
## 1 -22379.6694
## 2  -1399.7840
## 3  10977.7712
## 4  -1180.8642
## 5  12981.4794
## 6    891.8618
```

```
pl<-ggplot(res,aes(res))+geom_histogram(fill='blue',bins=30,alpha=0.7)
print(pl+ggtitle('residuals'))
```




```
#use test data make predictions
price.pred<-predict(model,test)
result<-cbind(price.pred,test$price)
colnames(result)<-c('predicted','True')
results<-as.data.frame(result)
head(results)
```

```
##   predicted  True
## 1  64379.67 42000
## 2  39899.78 38500
## 3  38522.23 49500
## 4  61680.86 60500
## 5  48018.52 61000
## 6  65108.14 66000
```

```
#mse
mse<-mean((results$True-results$predicted)^2)
print(mse)
```

```
## [1] 231923928
```

```
#rmse
print(mse^0.5)
```

```
## [1] 15229.05
```

```
# R^2
sse<-sum((results$predicted-results$True)^2)
sst<-sum((mean(df$price)-results$True)^2)
R2<-1-sse/sst
print(R2)
```

```
## [1] 0.6741391
```

```
# make prediction for the given requrment
colnames(df)
```

```
##  [1] "X"        "price"    "lotsize"  "bedrooms" "bathrms"  "stories"  "driveway" "re
croom"
##  [9] "fullbase" "gashw"    "airco"    "garagepl" "prefarea"
```

```
X<-c(1,2)
lotsize<-c(5500,5500)
bedrooms<-c(4,4)
bathrms<-c(2,2)
stories<-c(2,2)
driveway<-c(1,0)
recroom<-c(1,0)
fullbase<-c(1,0)
gashw<-c(1,0)
airco<-c(1,0)
garagepl<-c(2,0)
prefarea<-c(1,0)
task_data<-data.frame(X,lotsize,bedrooms,bathrms,stories,driveway,recroom,fullbase,gash
w,
           airco,garagepl,prefarea)
task_data
```

```
##   X lotsize bedrooms bathrms stories driveway recroom fullbase gashw airco garagepl p
refarea
## 1 1    5500        4       2       2        1       1        1     1     1        2
1
## 2 2    5500        4       2       2        0       0        0     0     0        0
0
```

```
is.data.frame(task_data)
```

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

```
#make prediction
task.pred<-predict(model,task_data)
result<-as.data.frame(task.pred)
rownames(result)<-c('highest','lowest')
result
```

```
##         task.pred
## highest 121809.15
## lowest   63206.35
```

```
# give categorical data requrment
X<-c(1)
lotsize<-c(5500)
bedrooms<-c(4)
bathrms<-c(2)
stories<-c(2)
driveway<-c(1)
recroom<-c(0)
fullbase<-c(0)
gashw<-c(1)
airco<-c(1)
garagepl<-c(1)
prefarea<-c(1)

requre_data<-data.frame(X,lotsize,bedrooms,bathrms,stories,driveway,recroom,fullbase,gas
hw,
                        airco,garagepl,prefarea)
requre_data
```

```
##   X lotsize bedrooms bathrms stories driveway recroom fullbase gashw airco garagepl p
refarea
## 1 1    5500        4       2       2        1       0        0     1     1        1
1
```

```
is.data.frame(requre_data)
```

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

```
#make prediction
requre.pred<-predict(model, requre_data)
requre.pred
```

```
##        1
## 107374.6
```

```
# 4bed 2 bath 2 stories
library("dplyr")
df3<-select(df, price,bedrooms,bathrms, stories,lotsize)
head(df3)
```

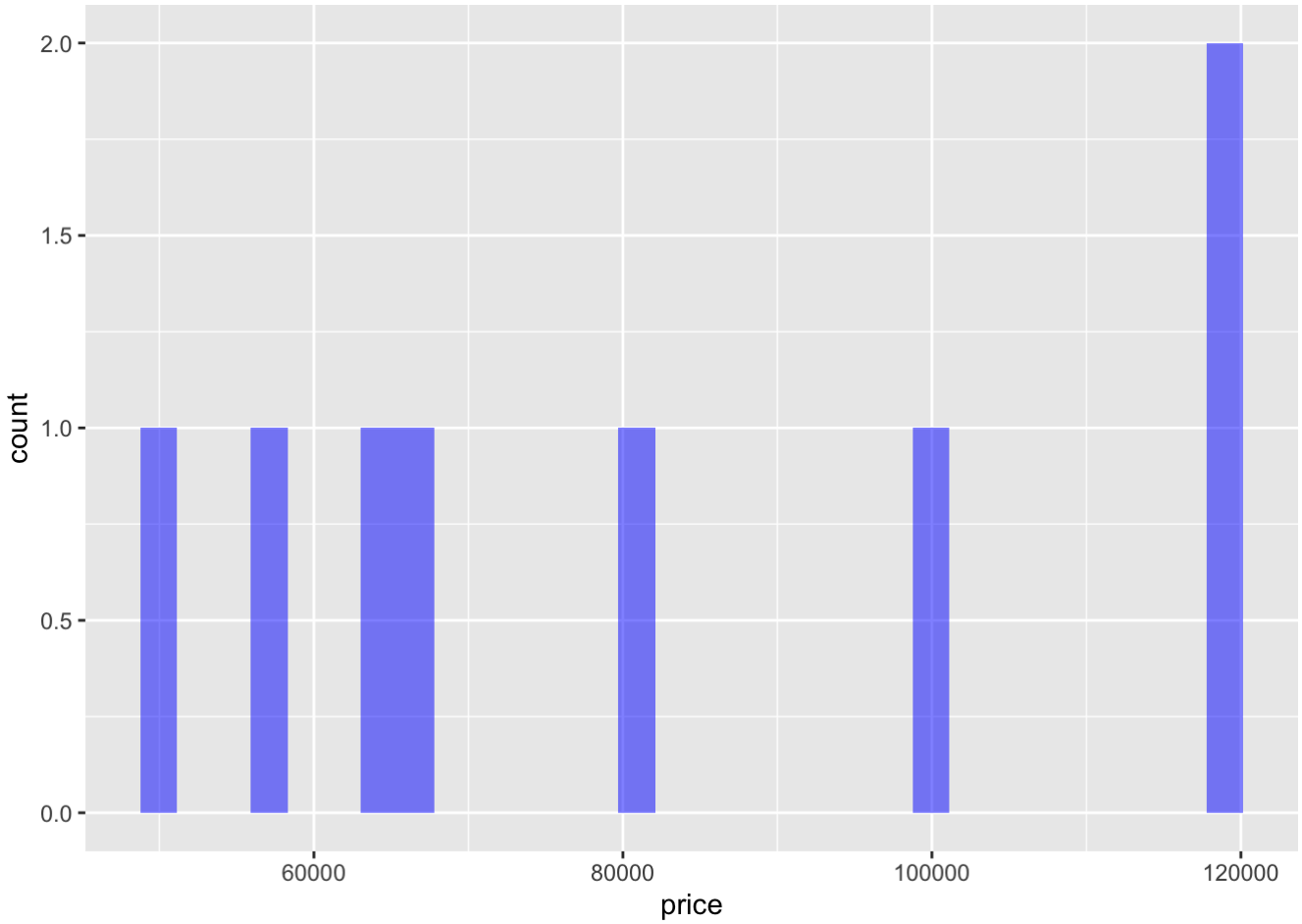
```
##   price bedrooms bathrms stories lotsize
## 1 42000        3       1       2    5850
## 2 38500        2       1       1    4000
## 3 49500        3       1       1    3060
## 4 60500        3       1       2    6650
## 5 61000        2       1       1    6360
## 6 66000        3       1       1    4160
```

```
his.data<-df3 %>% filter(bedrooms==4) %>% filter(bathrms==2) %>%
         filter(stories==2) %>% filter(lotsize<=6500 & lotsize>=4500) %>% arrange(desc(p
rice))
his.data
```

```
##    price bedrooms bathrms stories lotsize
## 1 120000        4       2       2    5500
## 2 118500        4       2       2    4880
## 3 101000        4       2       2    6240
## 4  82000        4       2       2    5400
## 5  65900        4       2       2    4510
## 6  64900        4       2       2    4990
## 7  58000        4       2       2    5900
## 8  51000        4       2       2    4500
```

```
ggplot(his.data,aes(x=price)) +geom_histogram(fill='blue',alpha=0.5)
```

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


2.0
1.5
1.0
0.5
0.0
count
60000
80000
100000
120000
price