Milwaukee 2021 Property Sales Data

Code

Hide

```
housing = read.csv("armslengthsales_2021_valid.csv")
summary(housing)
```

```
   PropertyID       PropType             taxkey            Address          CondoProject     
 Min.   :861425   Length:6508        Min.   :3.017e+07   Length:6508        Length:6508       
 1st Qu.:868280   Class :character   1st Qu.:2.300e+09   Class :character   Class :character  
 Median :872021   Mode  :character   Median :3.220e+09   Mode  :character   Mode  :character  
 Mean   :871938                      Mean   :3.471e+09                                        
 3rd Qu.:875743                      3rd Qu.:4.720e+09                                        
 Max.   :881476                      Max.   :7.160e+09                                        
   District             nbhd              Style              Extwall             Stories         
 Length:6508        Length:6508        Length:6508        Length:6508        Length:6508       
 Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  



  Year_Built           Rooms           FinishedSqft           Units             Bdrms          
 Length:6508        Length:6508        Length:6508        Min.   :  0.000   Length:6508       
 Class :character   Class :character   Class :character   1st Qu.:  1.000   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Median :  1.000   Mode  :character  
                                                          Mean   :  1.864                     
                                                          3rd Qu.:  2.000                     
                                                          Max.   :781.000                     
     Fbath           Hbath           Lotsize           Sale_date          Sale_price       
 Min.   :0.000   Min.   :0.0000   Length:6508        Length:6508        Length:6508       
 1st Qu.:1.000   1st Qu.:0.0000   Class :character   Class :character   Class :character  
 Median :1.000   Median :0.0000   Mode  :character   Mode  :character   Mode  :character  
 Mean   :1.466   Mean   :0.2961                                                           
 3rd Qu.:2.000   3rd Qu.:1.0000                                                           
 Max.   :6.000   Max.   :3.0000                                                           
```

Hide

```
#removed propertyID, taxkey, and address since it is not relevant to the analysis
housing$PropertyID = NULL
housing$taxkey = NULL
housing$Address = NULL
```

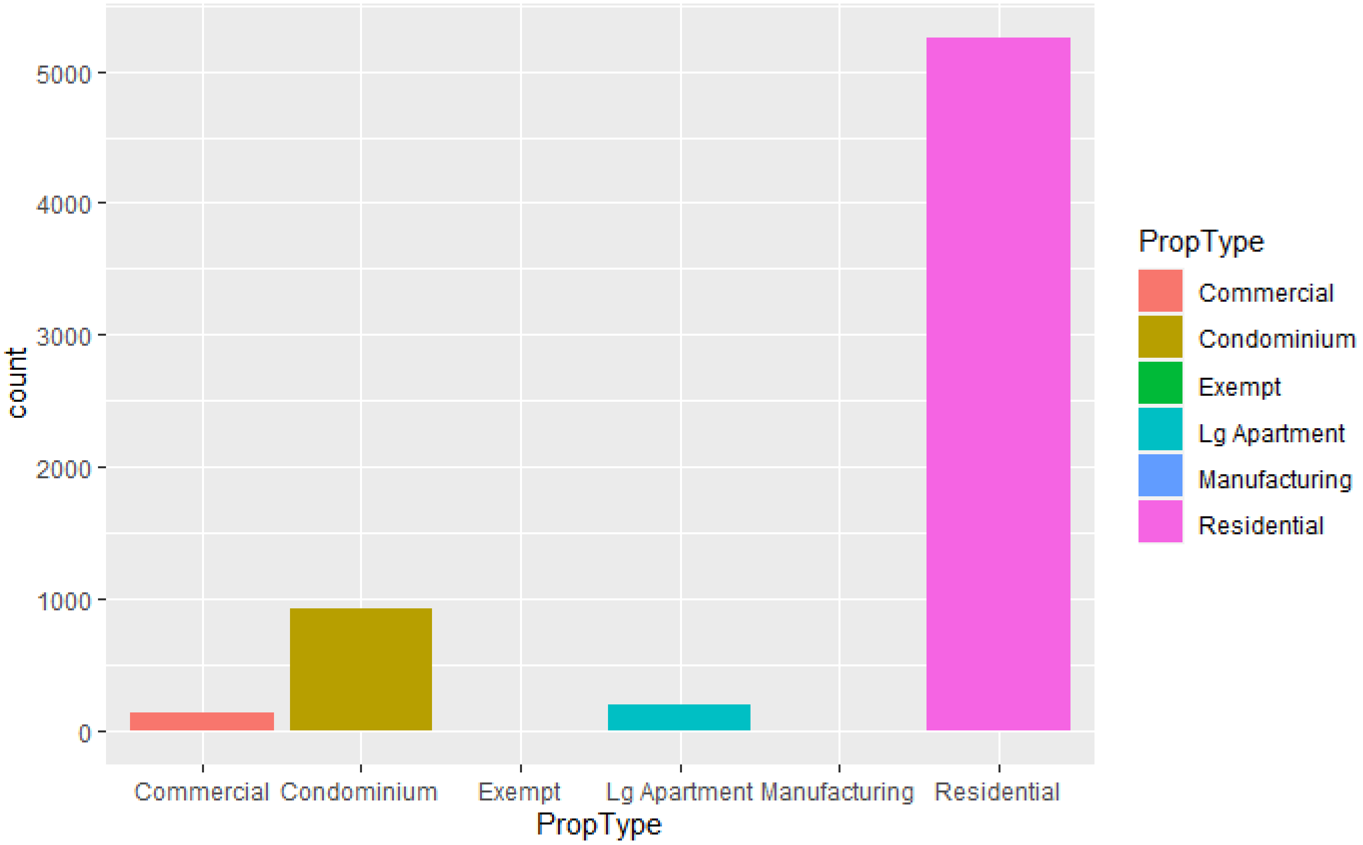
Hide

```
table(housing$PropType)
```

```
   Commercial   Condominium        Exempt  Lg Apartment Manufacturing   Residential 
          133           928             2           191             2          5252 
```

Hide

```
library(ggplot2)
ggplot(housing, aes(x=PropType, fill=PropType)) + geom_bar()
```



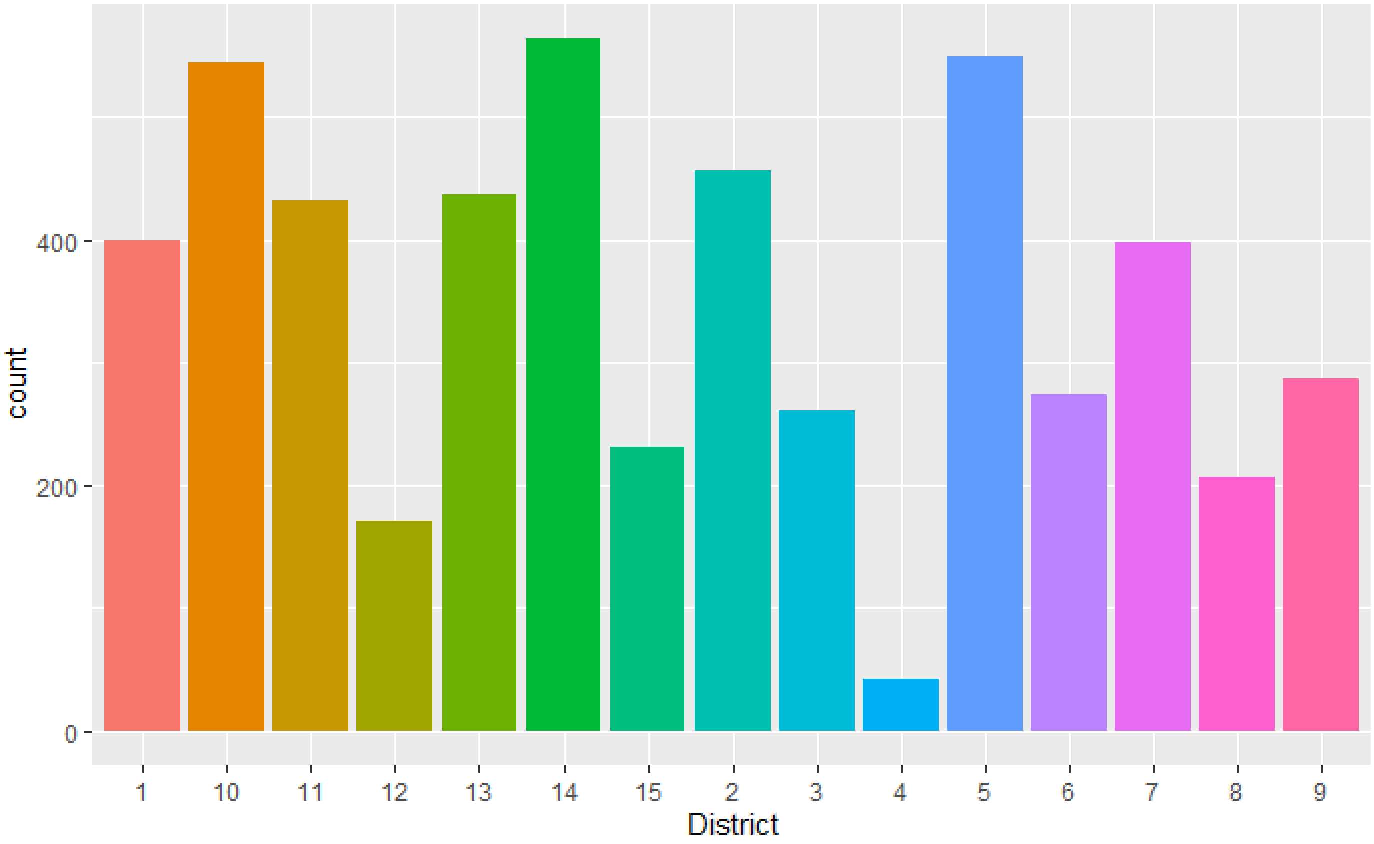
We can see that the majority of our data comes from the information of residential homes. For this analysis, I will remove all other property types other than residential homes. We will also remove the CondoProject column since it is no longer relevant. We are now left with 5252 observations of 16 variables.

Hide

```
library(dplyr)
housing2 = filter(housing, PropType == "Residential")
housing2 = subset(housing2, select = -CondoProject)
```

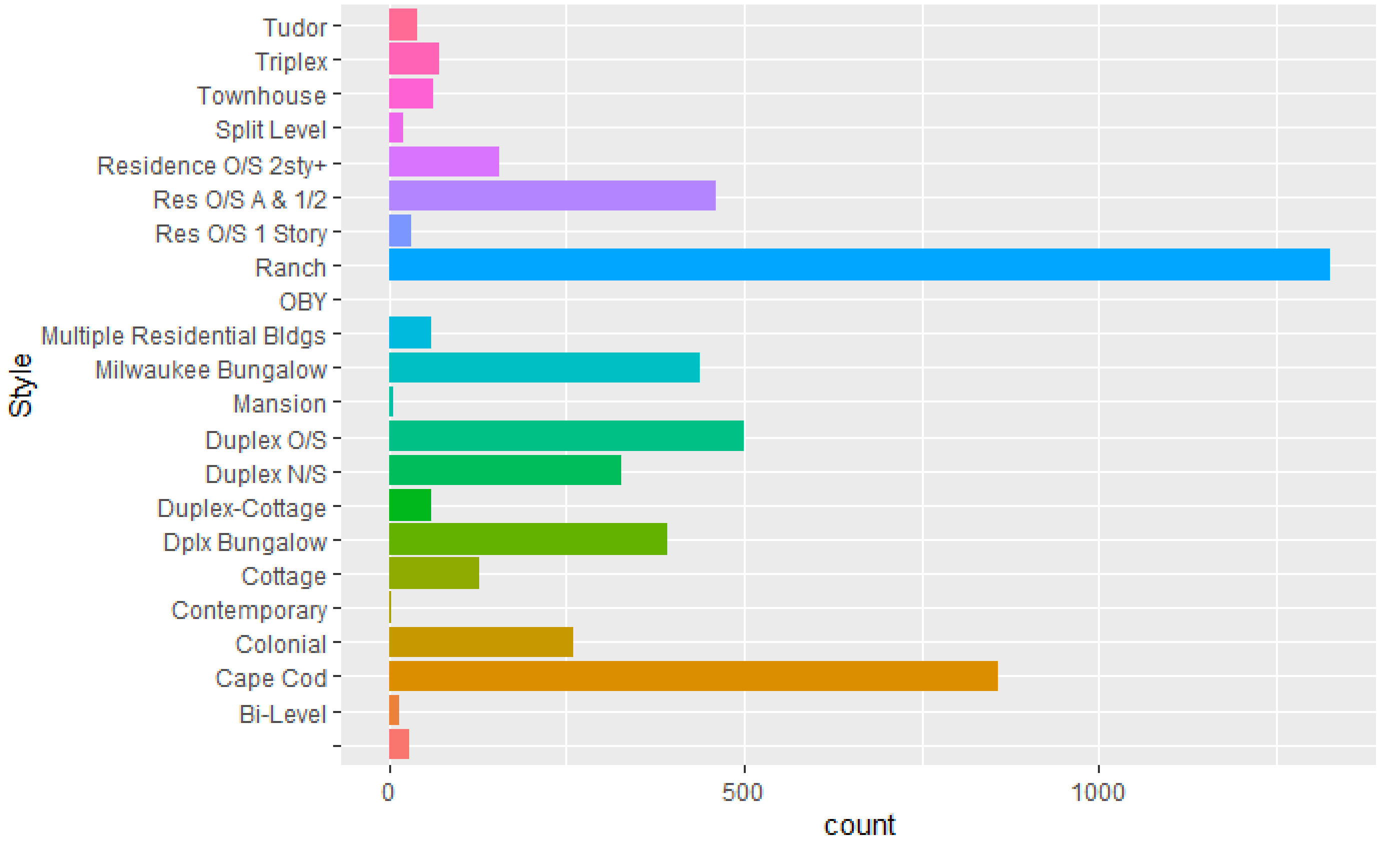
Hide

```
ggplot(housing2, aes(District, fill=District)) +geom_bar()+theme(legend.position="none")
```



Hide

```
ggplot(housing2, aes(Style, fill=Style)) + geom_bar(stat="count") + coord_flip()+theme(legend.po
sition = "none")
```



Converting some character columns into int columns and coerced NULLS to NAs

Hide

```
table(housing2$Rooms)
```

```

   0   10   11   12   13   14   15   16   17   18   19   20    3    4    5    6    7    8    9 N
ULL 
  27  562  148  200   33   33   12    8    2    2    3    1   10  405 1395 1059  557  462  293 
40 
```

Hide

```
table(housing2$Stories)
```

```

        0    1  1.5    2  2.5    3 
   8   22 2983  853 1377    4    5 
```

Hide

```
table(housing2$Bdrms)
```

```

   0    1    2    3    4    5    6    7    8    9 NULL 
   4   31  683 2358 1374  336  376   24   21    6   39 
```

Hide

```
housing2$Rooms = as.integer(housing2$Rooms)
```

```
Warning: NAs introduced by coercion
```

Hide

```
housing2$Stories = as.integer(housing2$Stories)
housing2$Bdrms = as.integer(housing2$Bdrms)
```

```
Warning: NAs introduced by coercion
```

Hide

```
any(is.na(housing2))
```

```
[1] TRUE
```

The FinishedSqft, Lotsize, and Sale_price should be numerical, so we must remove the commas and convert them. NULLs will become NAs.

Hide

```
housing2$FinishedSqft = gsub(",", "", housing2$FinishedSqft)
housing2$FinishedSqft = as.integer(housing2$FinishedSqft)
```

```
Warning: NAs introduced by coercion
```

Hide

```
housing2$Lotsize = gsub(",", "", housing2$Lotsize)
housing2$Lotsize = as.integer(housing2$Lotsize)
```

Hide

```
housing2$Sale_price = gsub(",", "", housing2$Sale_price)
```

Hide

```
housing2$Sale_price = substring(housing2$Sale_price, 2)
```

Hide

```
housing2$Sale_price = as.integer(housing2$Sale_price)
```

Converting Sale_date to date format and then to month

Hide

```
housing2$Sale_date = format(as.Date(housing2$Sale_date, format = "%m/%d/%Y"), format="%m")
```

Hide

```
table(housing2$Sale_date)
```

```

 01  02  03  04  05  06  07  08  09  10  11  12 
359 337 509 523 588 630 625 591 481 244 212 153 
```

Hide

```
housing2 = housing2 %>% rename(Sale_month = Sale_date)
```

Removing PropType since they are all now Residential.

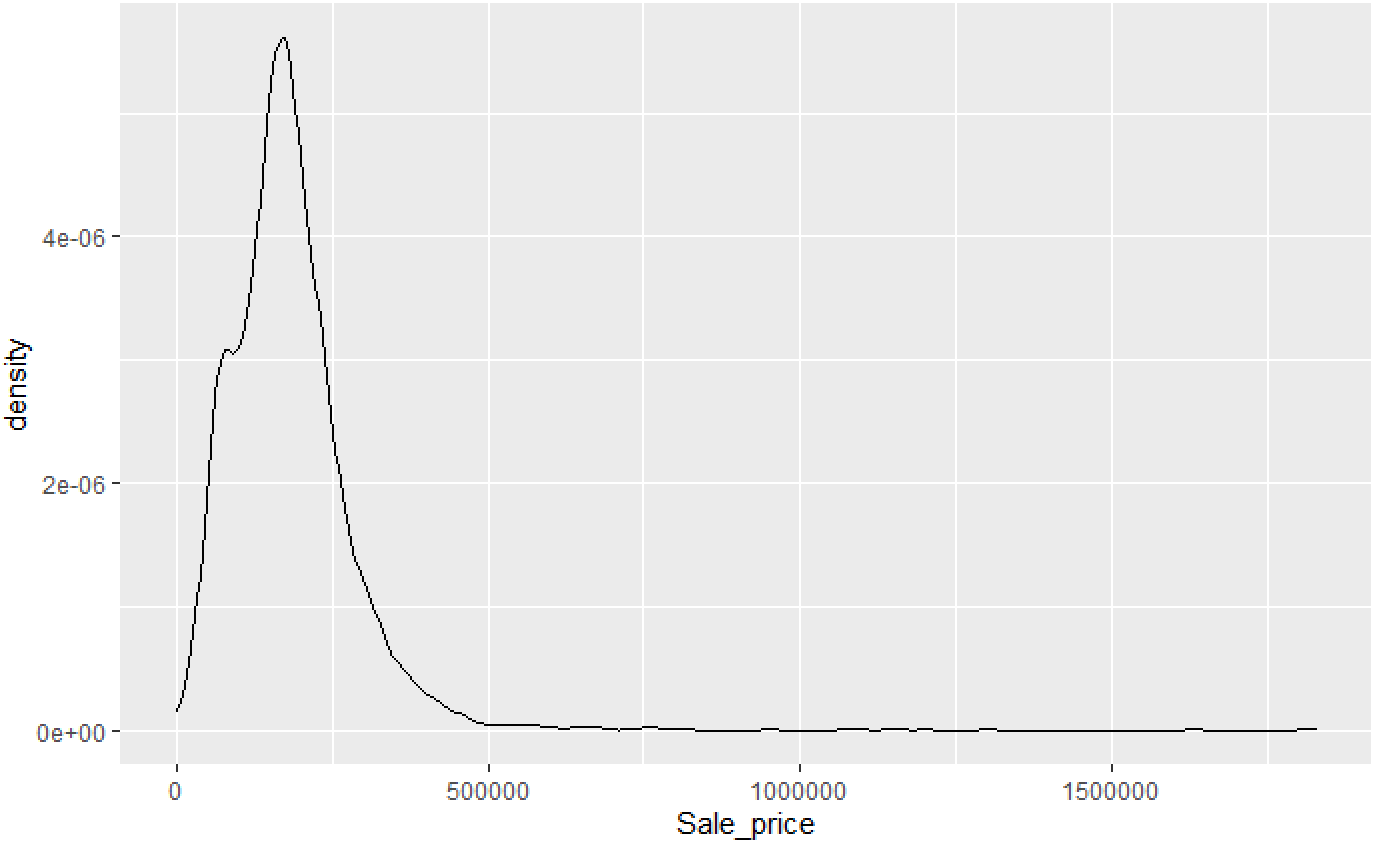
Hide

```
housing2$PropType = NULL
```

Density Curve of Sale_price

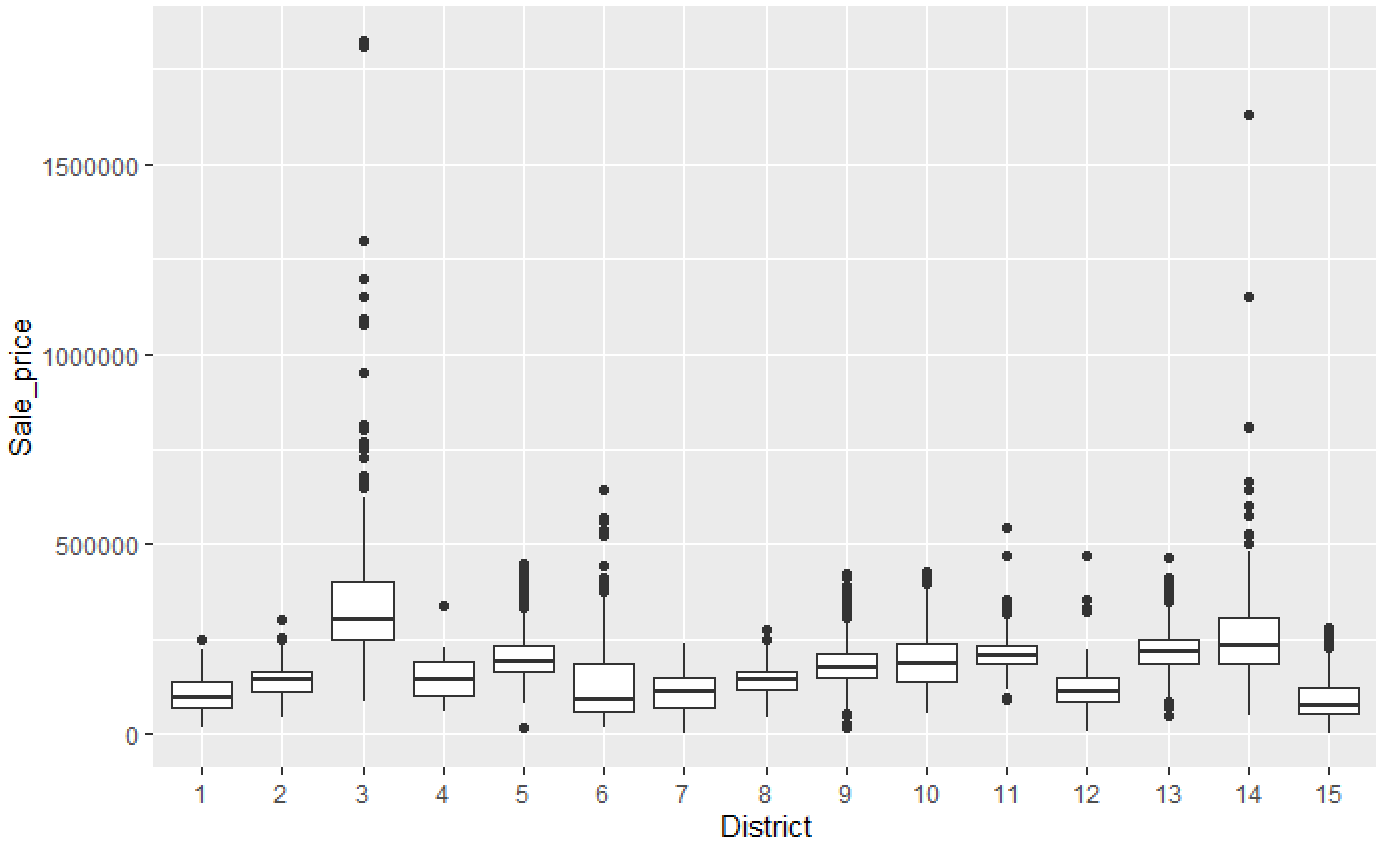
Hide

```
ggplot(housing2, aes(Sale_price))+geom_density()
```



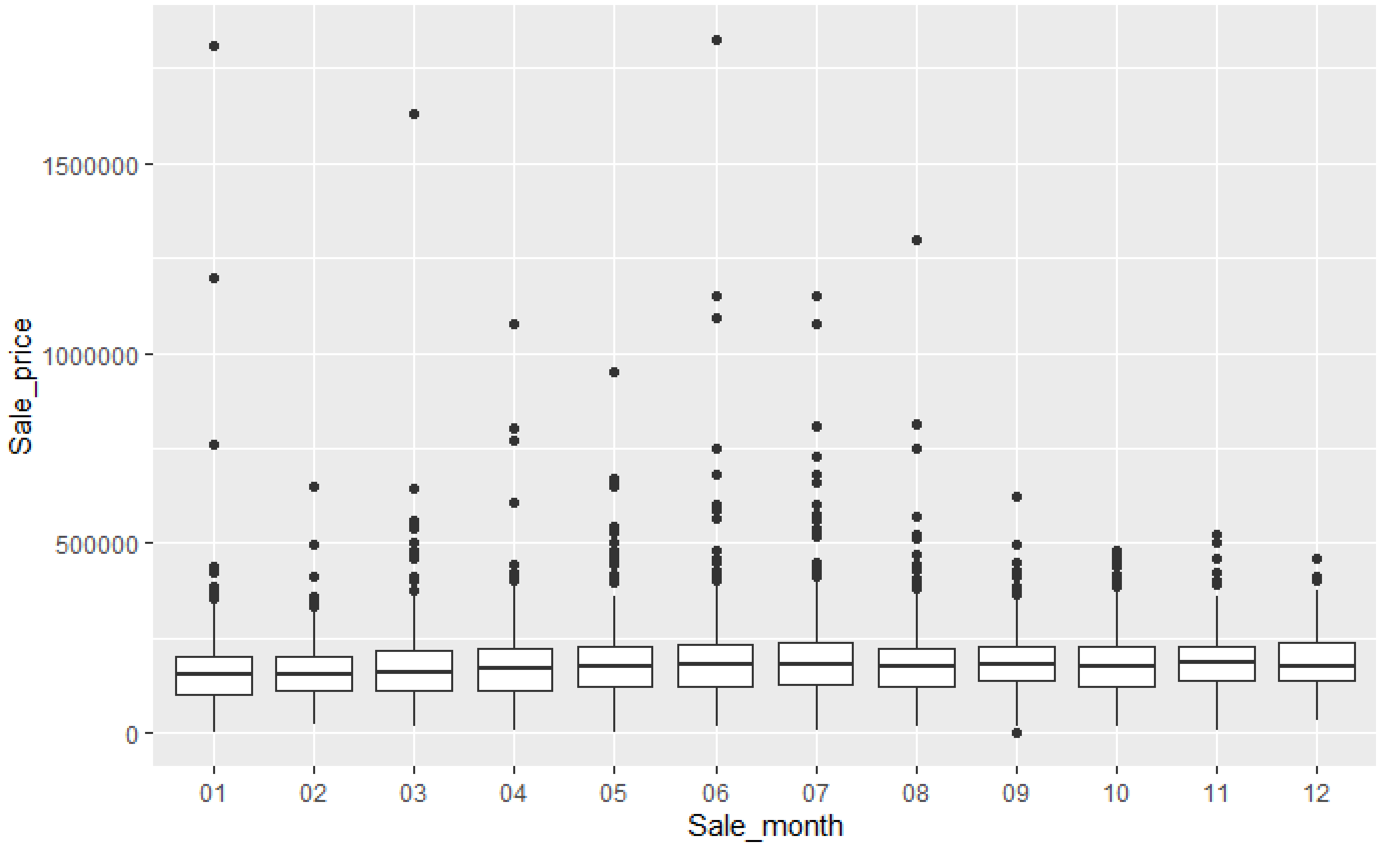
Hide

```
ggplot(housing2, aes(x=District, y=Sale_price))+geom_boxplot()+scale_x_discrete(limits=c("1",
"2", "3", "4", "5", "6","7", "8", "9", "10", "11", "12", "13", "14", "15"))
```



Hide

```
ggplot(housing2, aes(x=Sale_month, y=Sale_price))+geom_boxplot()
```



Hide

```
housing2$District = as.factor(housing2$District)
housing2$nbhd = as.integer(housing2$nbhd)
```

```
Warning: NAs introduced by coercion
```

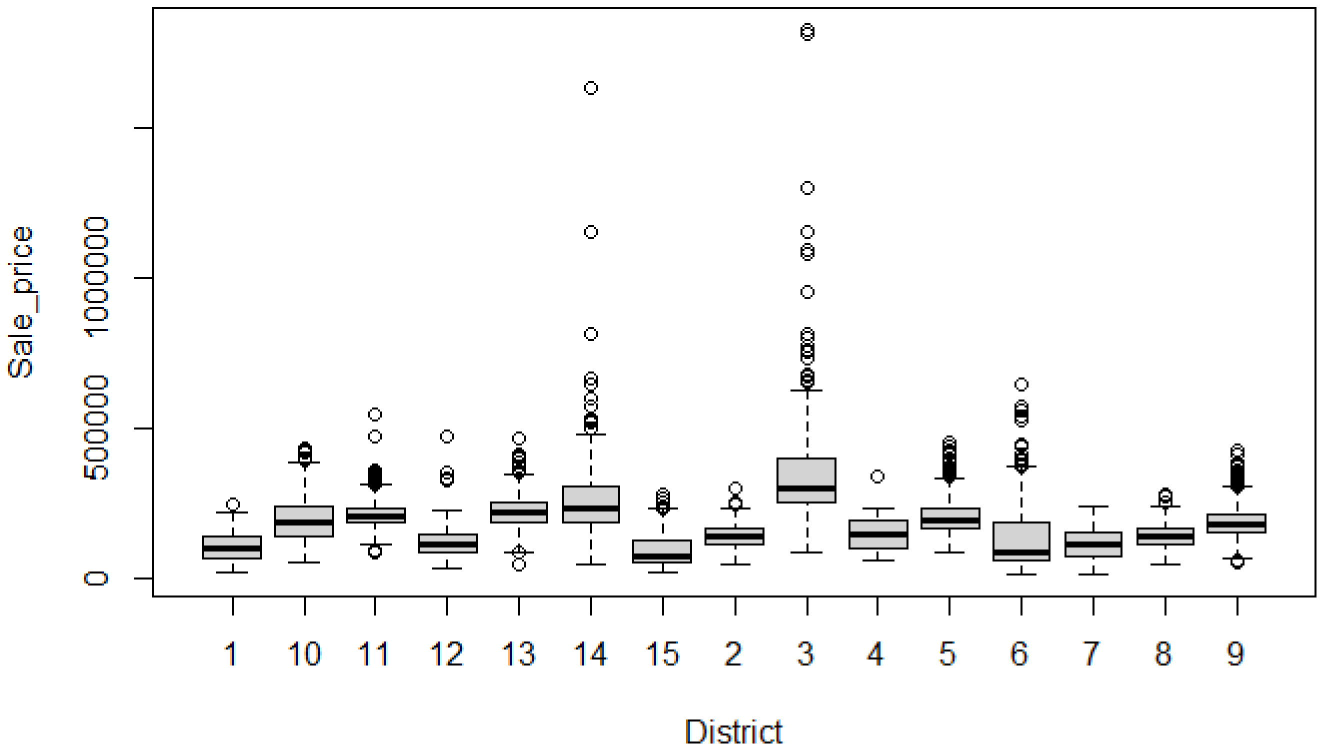
Hide

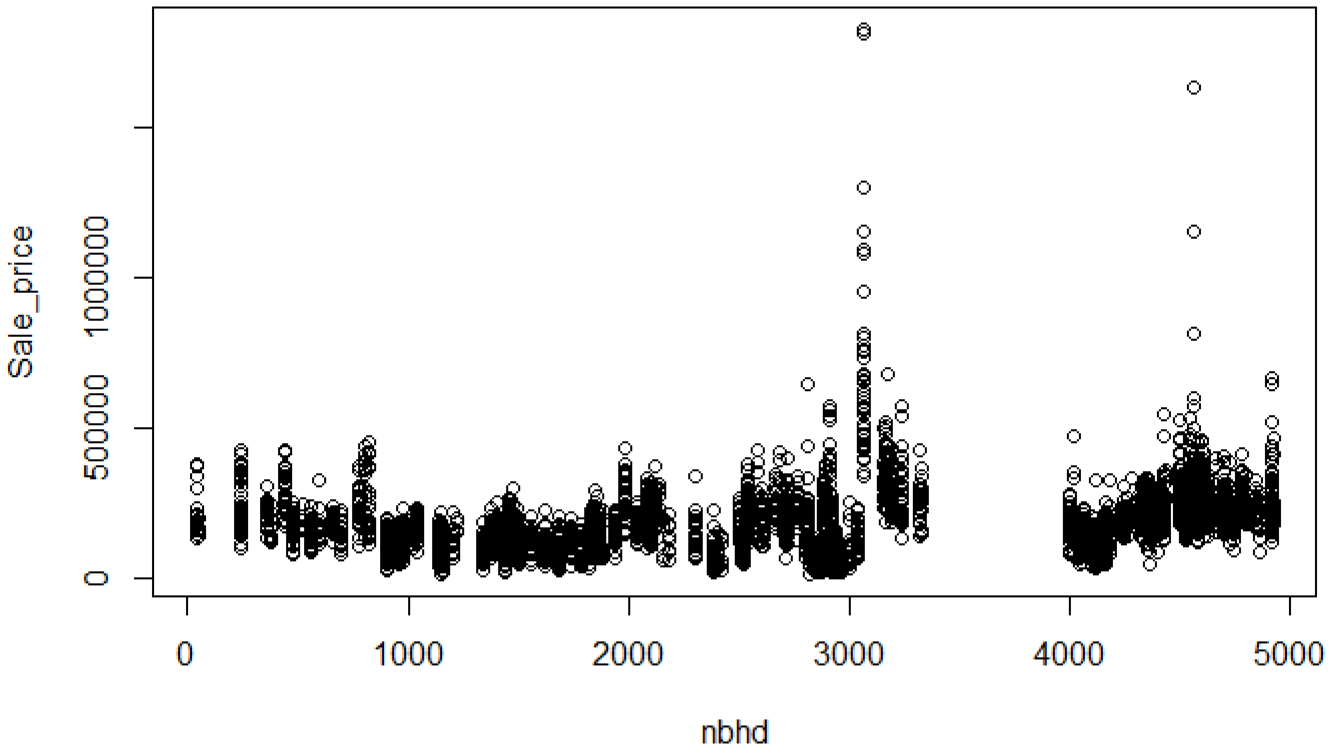
```
housing2$Style = as.factor(housing2$Style)
housing2$Extwall = as.factor(housing2$Extwall)
housing2$Stories = as.factor(housing2$Stories)
housing2$Year_Built = as.integer(housing2$Year_Built)
```

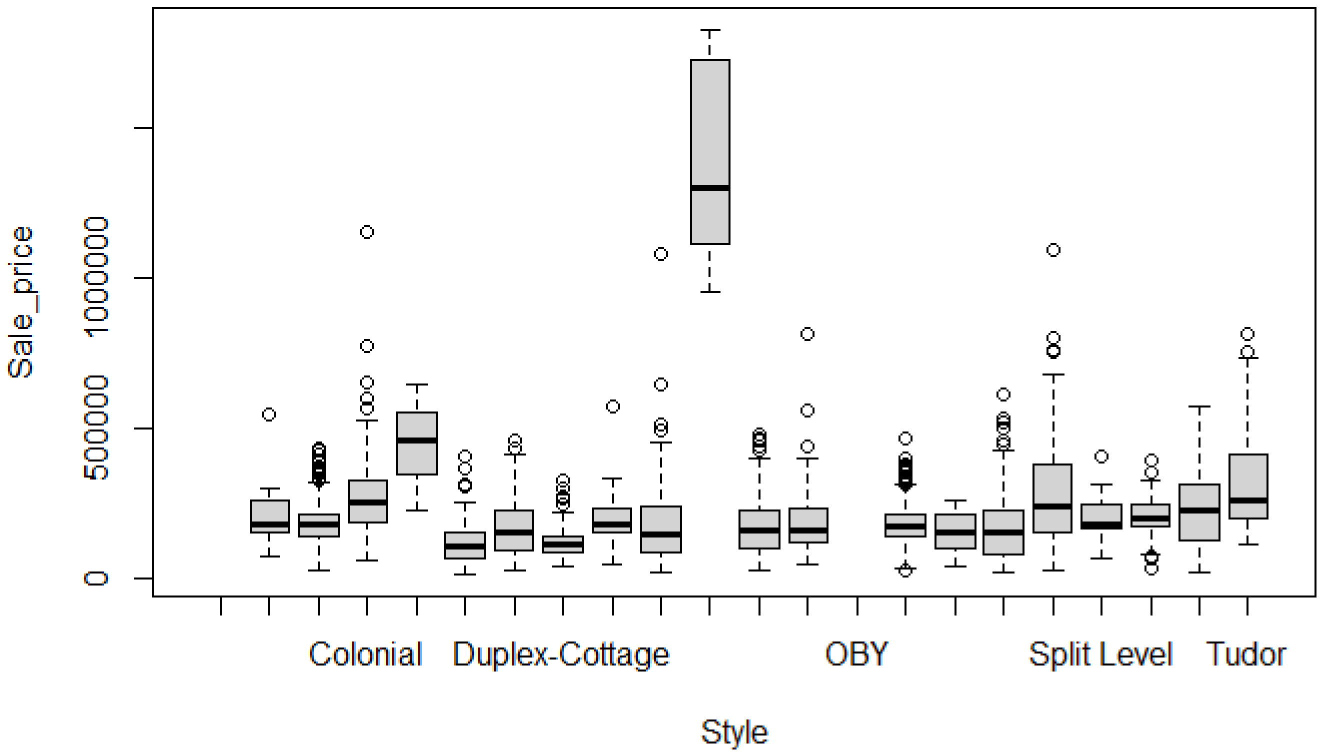
```
Warning: NAs introduced by coercion
```

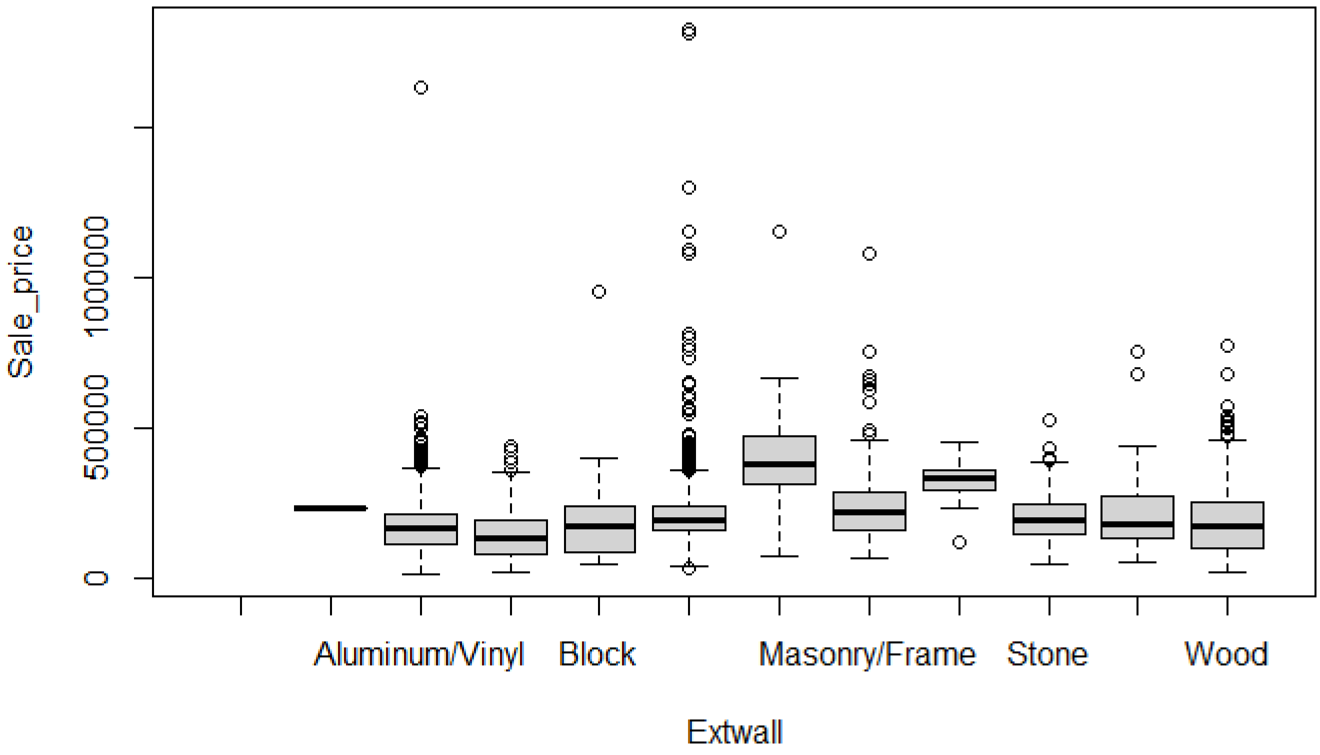
Hide

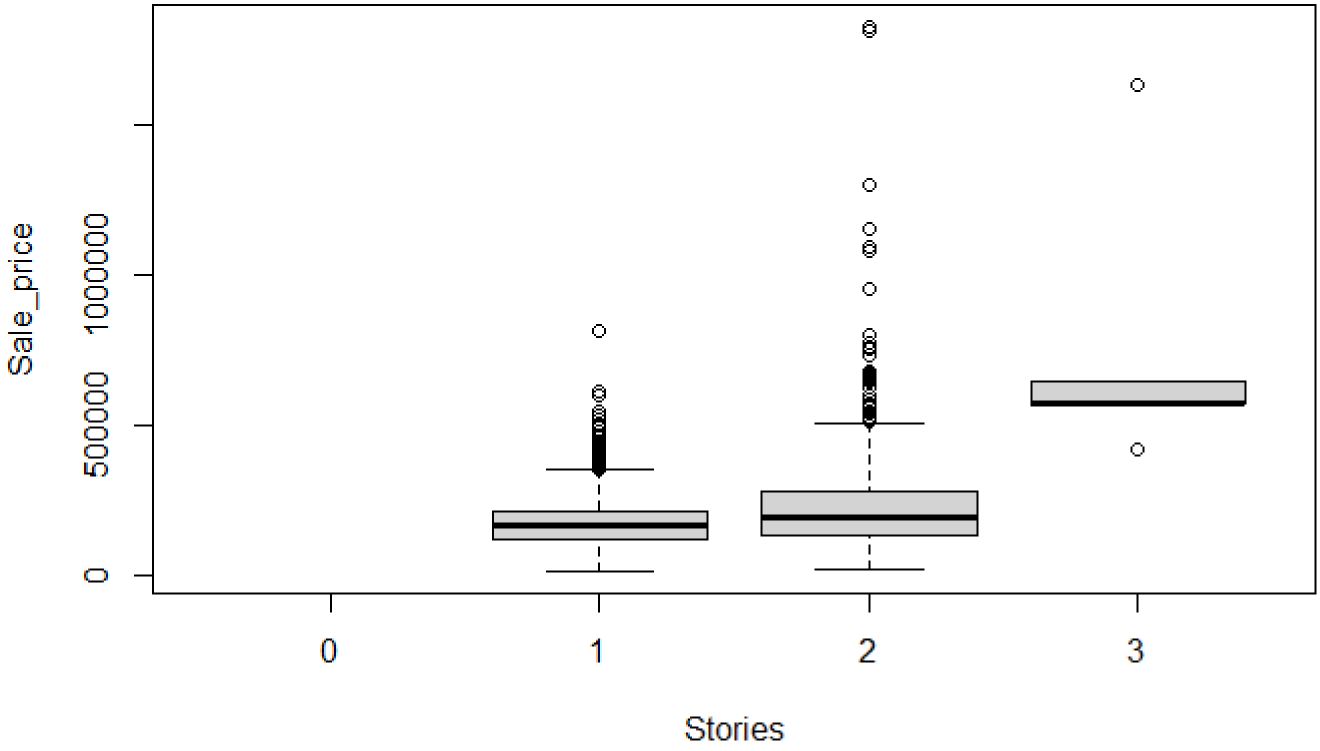
```
housing2 = housing2[housing2$Year_Built != 0, ]
housing2$Rooms = as.factor(housing2$Rooms)
housing2$Units = as.factor(housing2$Units)
housing2$Bdrms = as.factor(housing2$Bdrms)
housing2$Fbath = as.factor(housing2$Fbath)
housing2$Hbath = as.factor(housing2$Hbath)
housing2$Sale_month = as.factor(housing2$Sale_month)

plot(Sale_price~.,data=housing2)
```

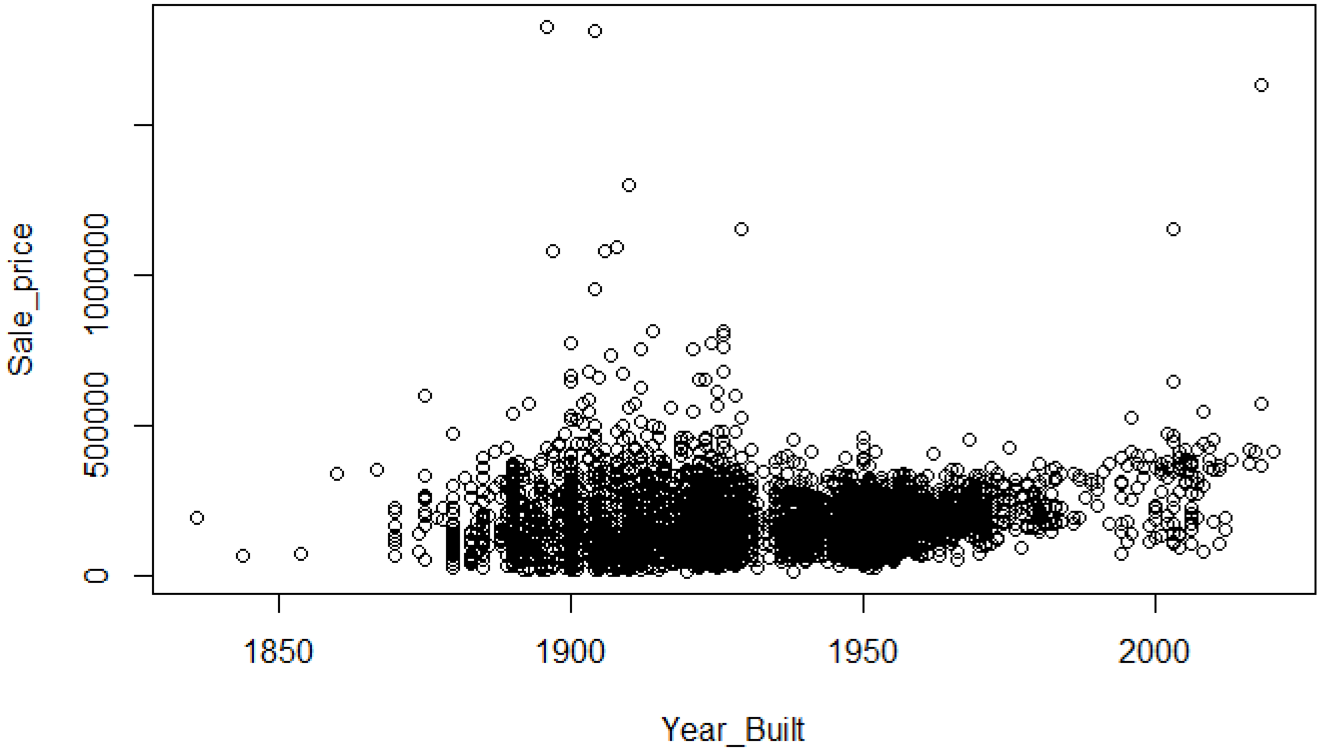

Sale_price
0
500000
1000000
1
10
11
12
13
14
15
2
3
4
5
6
7
8
9
District

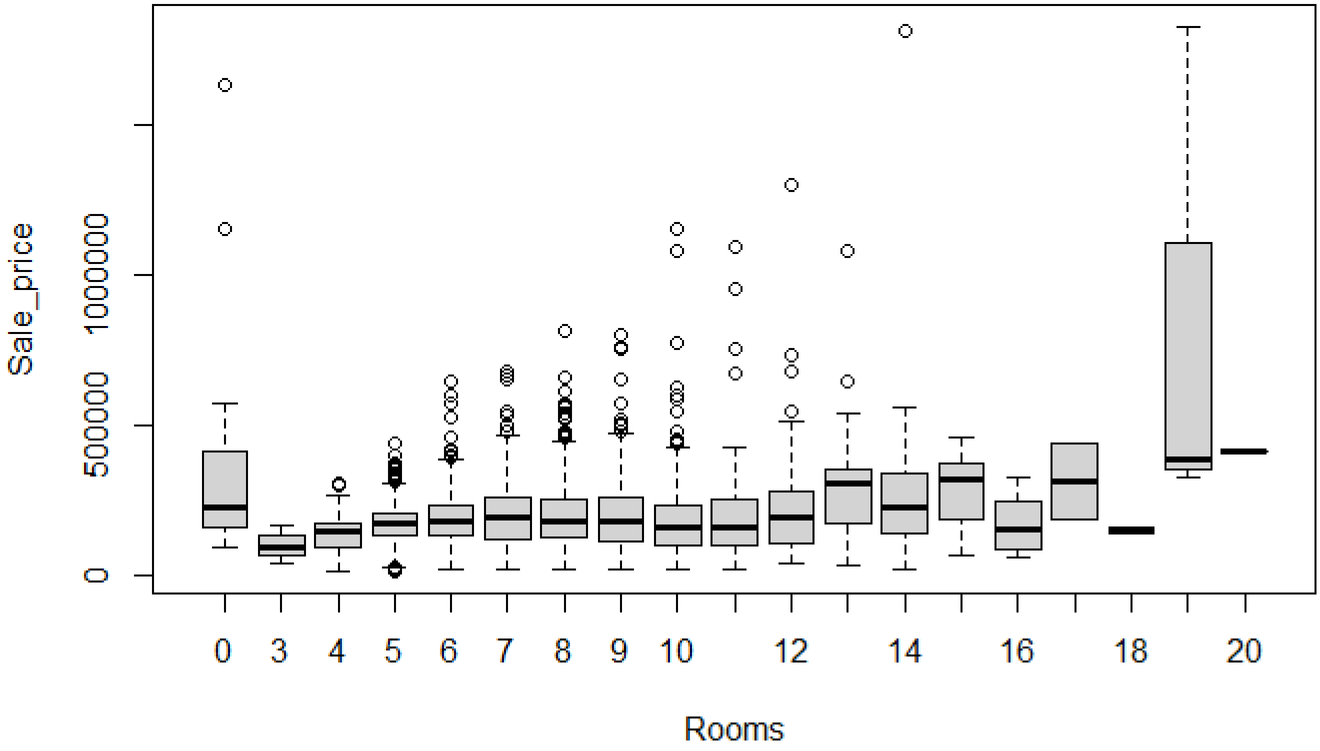

Sale_price
0
500000
1000000
0
1000
2000
3000
4000
5000
nbhd

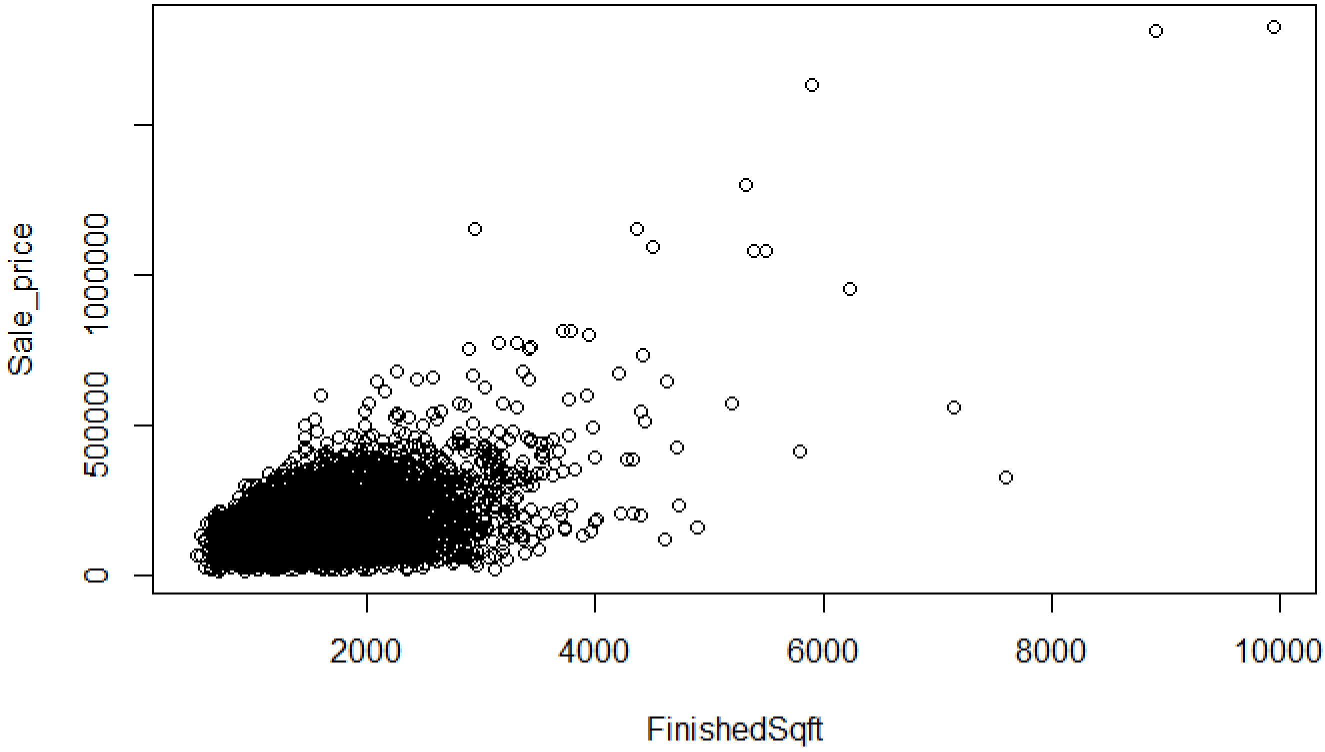

Sale_price
0
500000
1000000
Colonial
Duplex-Cottage
OBY
Split Level
Tudor
Style

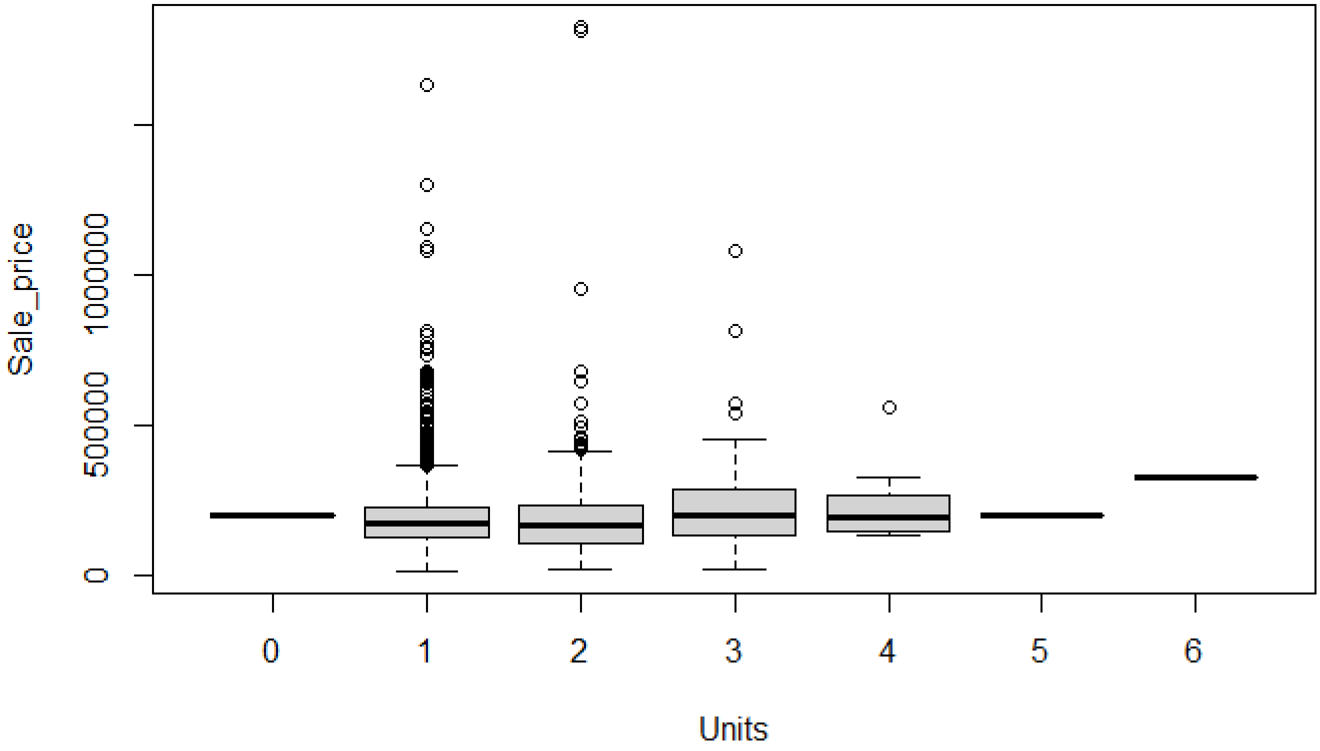

Sale_price
0
500000
1000000
Aluminum/Vinyl
Block
Masonry/Frame
Stone
Wood
Extwall

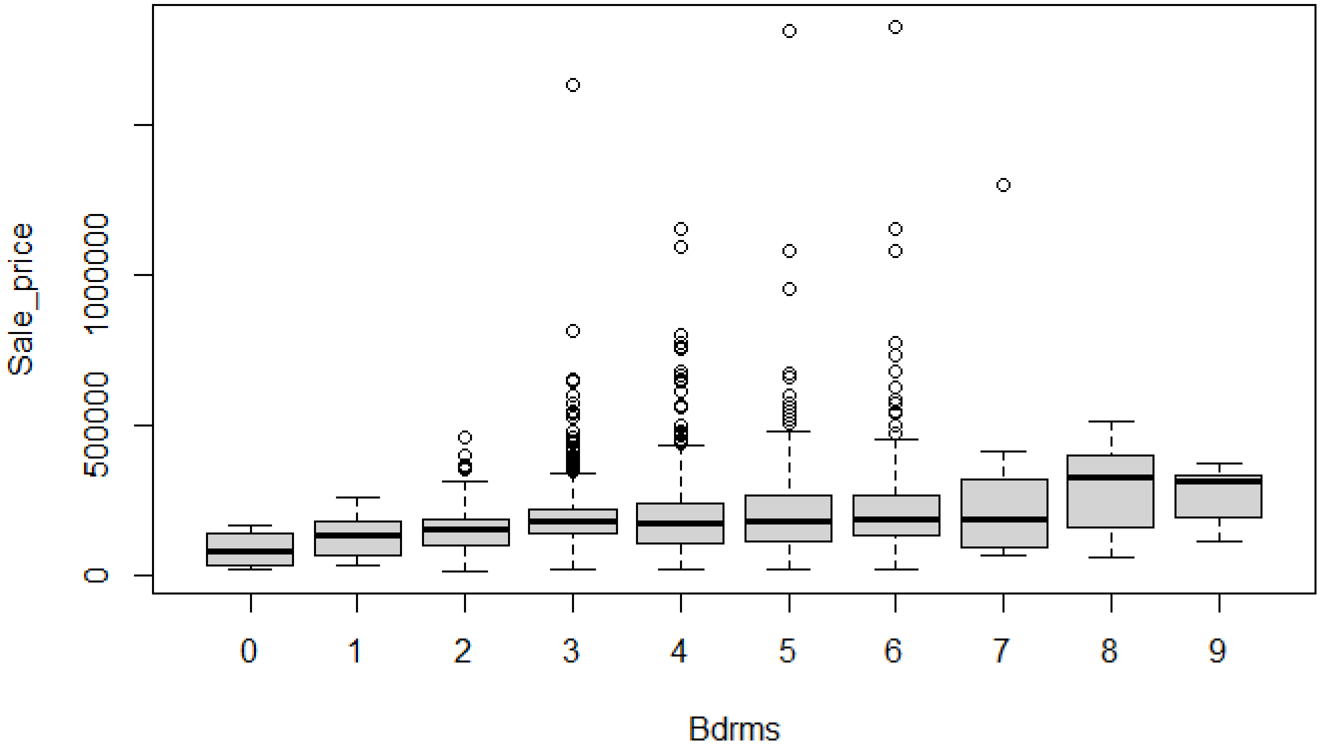

Sale_price
0
500000
1000000
0
1
2
3
Stories

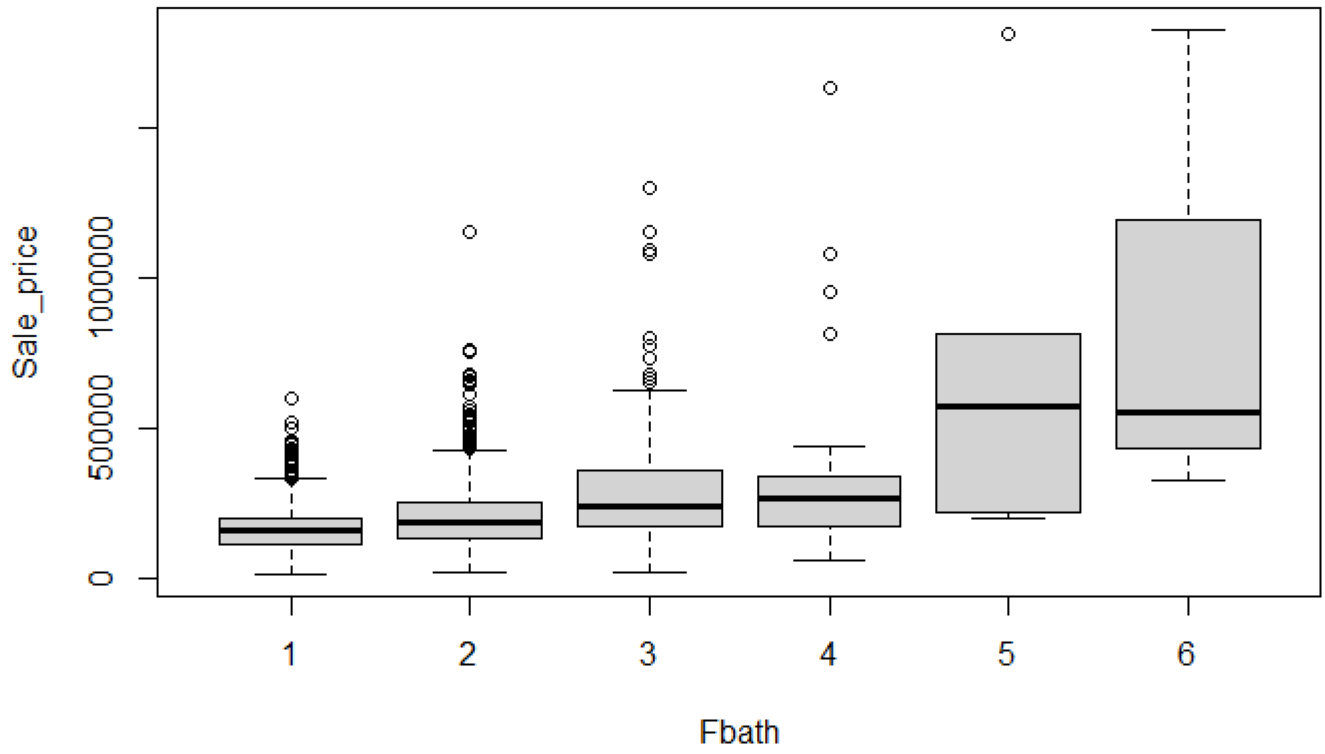

Sale_price
0
500000
1000000
1850
1900
1950
2000
Year_Built

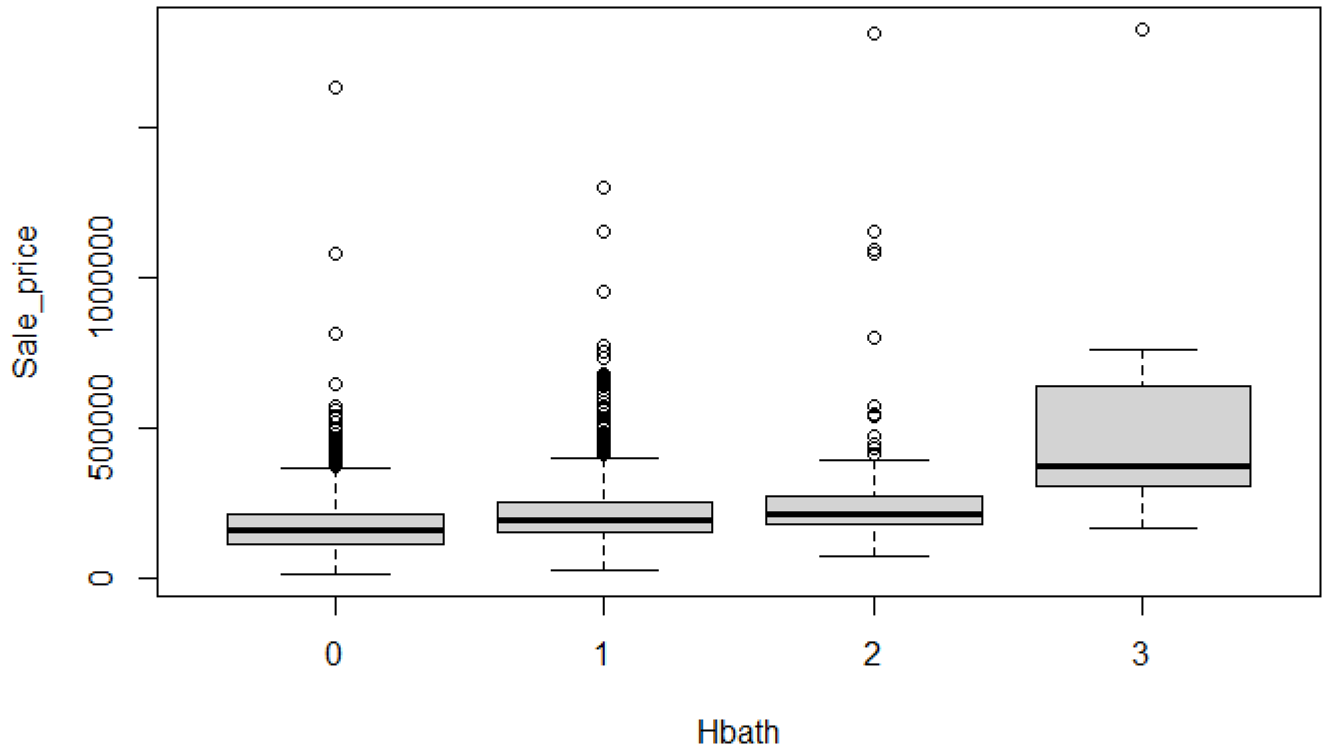

Sale_price
0
500000
1000000
0
3
4
5
6
7
8
9
10
12
14
16
18
20
Rooms

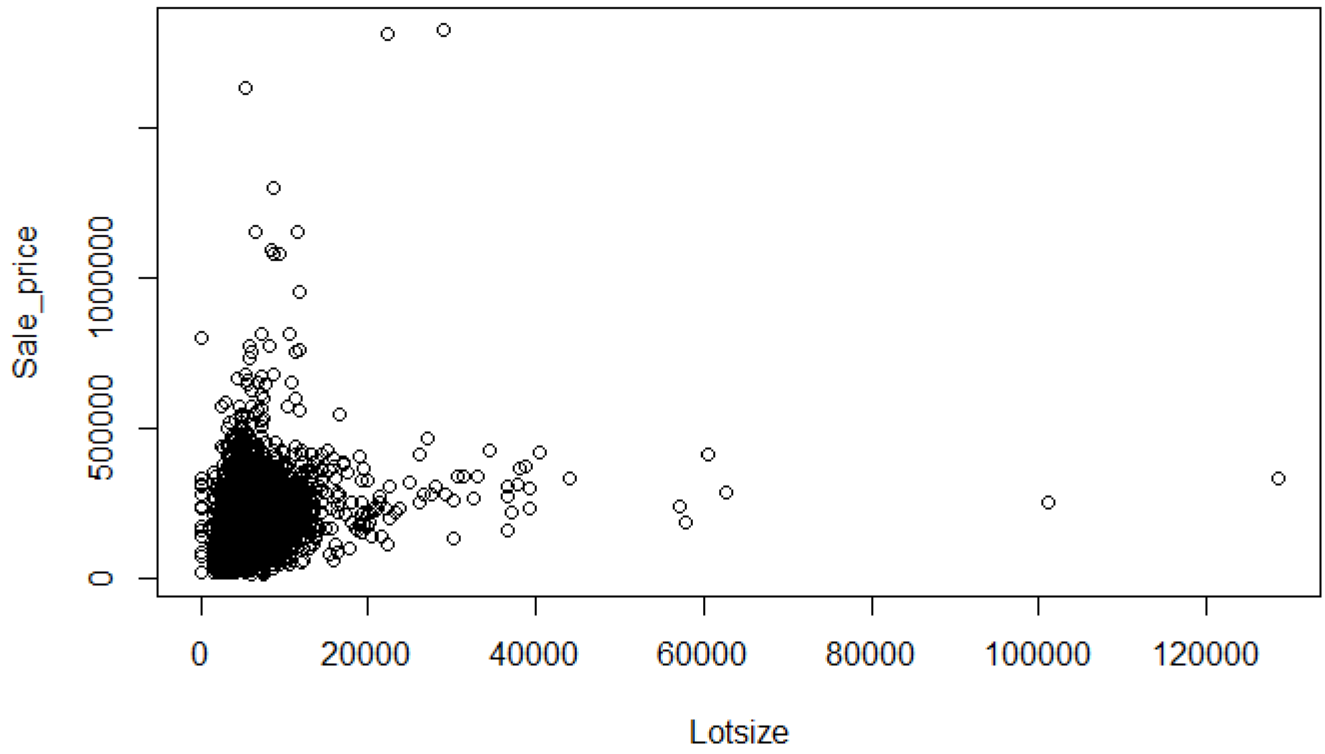

Sale_price
0
500000
1000000
2000
4000
6000
8000
10000
FinishedSqft

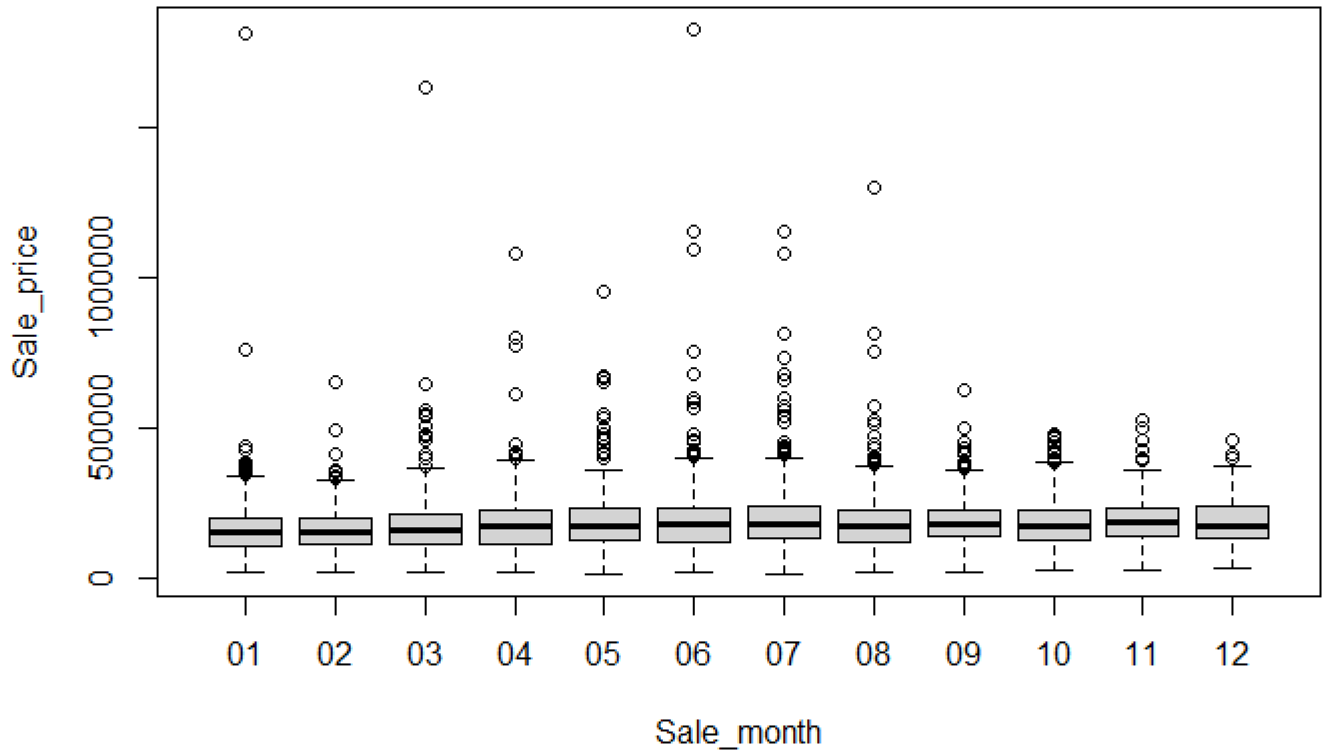

Sale_price
0
500000
1000000
0
1
2
3
4
5
6
Units


Sale_price
0
500000
1000000
0
1
2
3
4
5
6
7
8
9
Bdrms


Sale_price
0
500000
1000000
1
2
3
4
5
6
Fbath


Sale_price
0
500000
1000000
0
1
2
3
Hbath


Sale_price
0
500000
1000000
0
20000
40000
60000
80000
100000
120000
Lotsize


Sale_price
0
500000
1000000
01
02
03
04
05
06
07
08
09
10
11
12
Sale_month