浙江大學

上海市闵行区房价预测总结报告

项目组：顾全、郑圣楠、武祥浩、李晓婷、杨致楠

目录

一、项目介绍

1.1 项目描述

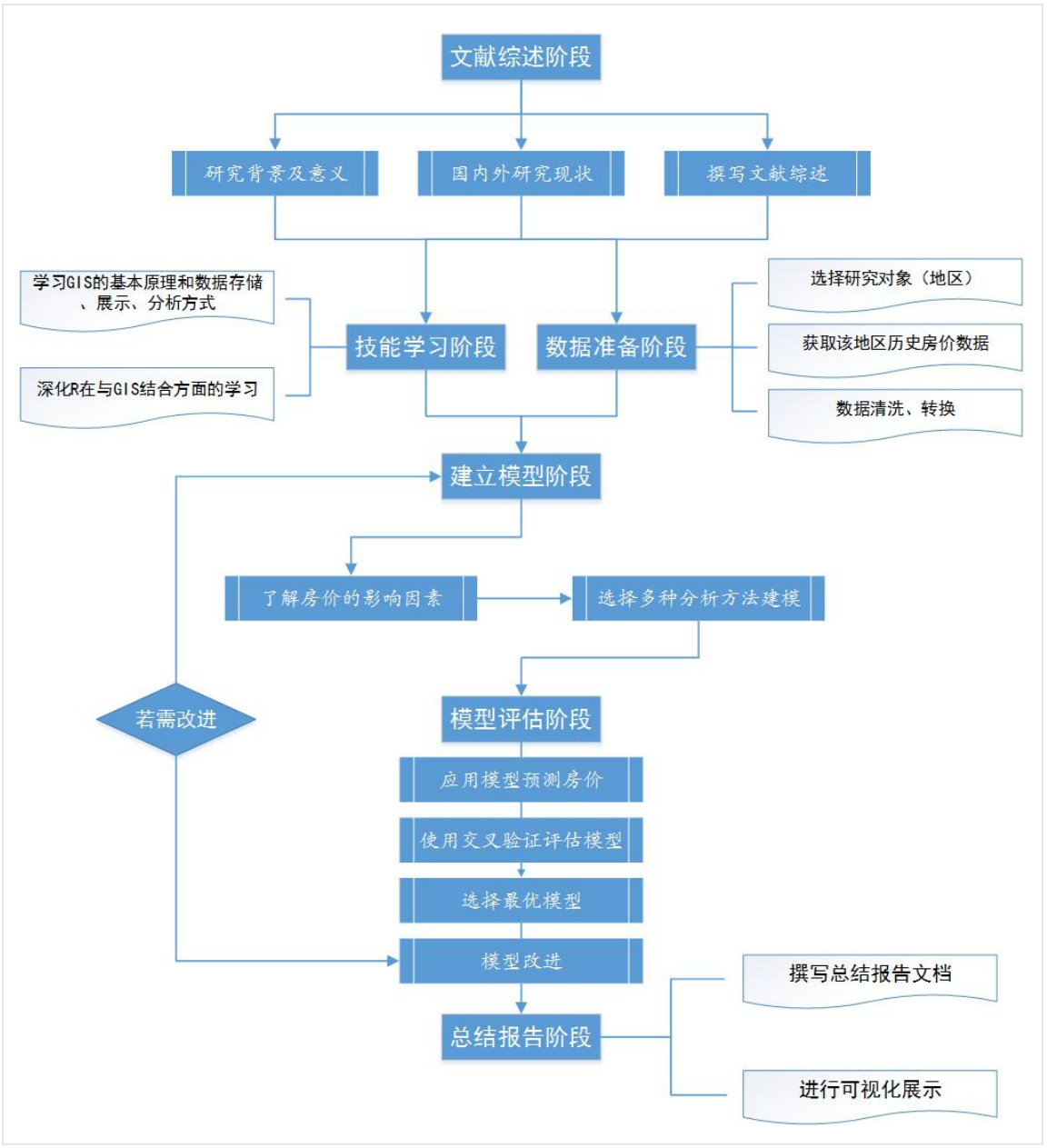
本项目要求我们探寻影响房价的因素。根据变量和房价的往年数据，对房价给出预测。主要使用的工具是 R 语言，R 是一套完整的数据处理、计算和制图软件系统。它实现的功能主要包括：数据存储、数组运算（向量、矩阵运算等）、统计分析、统计制图、操纵数据的输入和输出等，可实现分支、循环，用户可自定义功能。在本项目中，R 的主要应用在于数据处理和建模、预测，并且对模型进行检验与优化 。

1.2 项目目标

1、选取，分析影响房价的变量
2、运用 R 进行数据清洗和建立模型。
3、给出房价的预测值和对模型进行优化。

1.3 技术路线

1、分析选择企业所给的数据。
2、对选定的数据进行清洗。
3、对变量数据进行统计描述和可视化化展示。
4、建立多元回归模型、决策树模型、随机森林模型。
5、进行模型诊断和优化。


文献综述阶段
研究背景及意义
国内外研究现状
撰写文献综述
学习GIS的基本原理和数据存储、展示、分析方式
深化R在与GIS结合方面的学习
技能学习阶段
数据准备阶段
选择研究对象（地区）
获取该地区历史房价数据
数据清洗、转换
建立模型阶段
了解房价的影响因素
选择多种分析方法建模
若需改进
模型评估阶段
应用模型预测房价
使用交叉验证评估模型
选择最优模型
模型改进
总结报告阶段
撰写总结报告文档
进行可视化展示

二、团队成员介绍

2.1 成员姓名

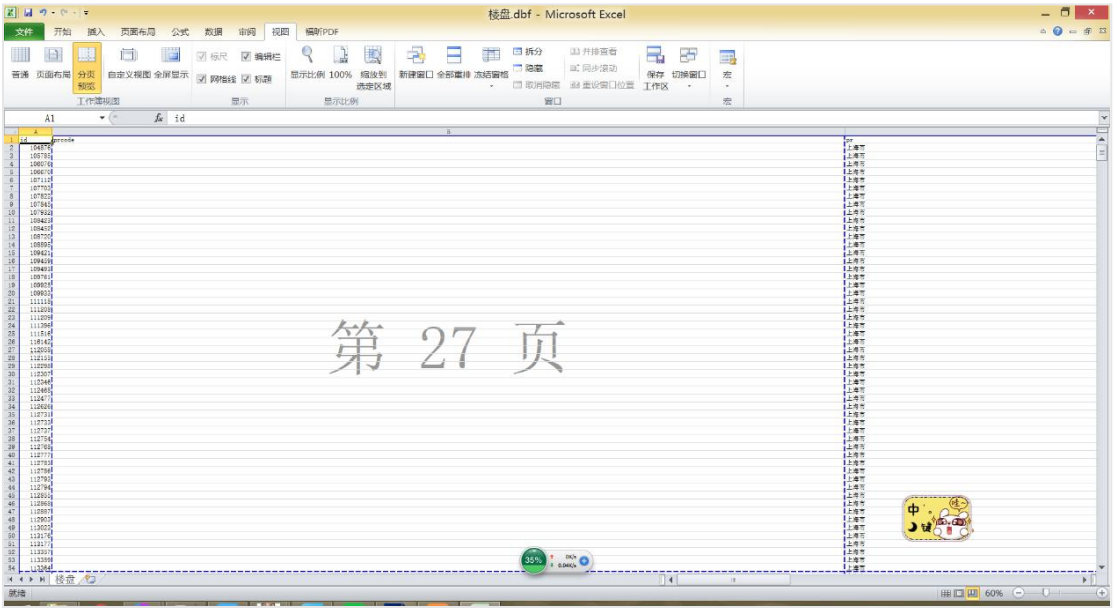
顾全（组长）、杨致楠、李晓婷、武祥浩、郑圣楠

2.2 成员介绍

身份	姓名	学号	专业
组长：	顾全	Z15030801	金融数据分析
组员：	武祥浩	Z15030958	金融数据分析
组员：	杨致楠	Z15031016	金融数据分析
组员：	李晓婷	Z14030872	金融数据分析
组员：	郑圣楠	Z15030969	金融数据分析

三、数据预处理

本次数据分析项目使用的数据是上海辰智咨询公司提供的 2012-2014 年上海市闵行区各个楼盘的房价数据。原始数据如下图所示：



可以看出，原始数据的非结构化程度很高，需要提取出里面各个字段的有价值的数值化信息。

3.1 去掉无效列

3.1.1 R 语言代码

```
setwd("C:\\Users\\GuQuan\\Desktop\\房价数据")
df1=read.csv("楼盘.csv")
df2=df1[,c("prcode","pr","citycode","city","cntycode","towncode","town","blkcode","blk","rentrate","phone","cls1code","cls1","cls2code","cls2","cls3code","cls3","srccode","pic","version","poisrc","tag")]
summary(df2)
df3=df1[!(names(df1)%in%names(df2))]
```

3.1.2 解释说明

原始数据为 df1，df1 中这些列 prcode、citycode、cntycode、towncode、town、blkcode、blk、phone、rentrate、cls1code、cls1、cls2code、cls2、cls3code、cls3、srccode、pic 共 17 列完全缺失，故去掉；另外 version、pr、city、poisrc、tag 这五列值每条记录值都是相同的，对分析没有意义，也需要去掉。summary(df2)是将去掉的列进行汇总，以确定去除无误。

3.2 将"暂无数据"和 0 置空

3.2.1 R 语言代码

```
#非价格型数据，将"暂无数据"置空
df4=df3[,c("id","cnty","poicode","lng","lat","name","addr","section","developer","re","retype","refee","totarea","tothh","blddate","plotrate","pkrate","greenrate","comments")]
write.csv(df4,"out1.csv")
df4=read.csv("out1.csv",na.string="暂无数据")
#价格数据，将 0 置空
```

```
df5=df3[!(names(df3)%in%names(df4))]
write.csv(df5,"out2.csv")
df5=read.csv("out2.csv",na.string=0)
#合并
df=cbind(df4,df5)
head(df)                #发现有两个序号列 X，分别在第 1 列，第 21 列
df=df[-c(1,21)]         #删除序号列 X
write.csv(df,"df.csv")
```

3.2.2 解释说明

去掉无效列之后的数据为 df3，df3 中有很多记录标示为“暂无数据”和“0”。为方便统一处理，对非房价数据的列，需要将里面的“暂无数据”设为空值 NA，对房价数据（小区房价和板块房价）的列，则需要将“0”设为空值 NA。然后代码最后将非房价数据和房价数据进行了合并。

3.3 各字段处理

这是数据处理里面最难也最耗费时间的工作，因为需要将 df 中原本为字符串的信息提取成数值化的、可用来作分析的数据。

```
setwd("C:\\Users\\GuQuan\\Desktop\\房价数据")
df=read.csv("df.csv",stringsAsFactors = FALSE)
library(stringr)
```

3.3.1 comments 处理

comments 是对各个楼盘的评论，原本的 comments 如下：

```
head(comments)
```

```
[1] "地铁房学区房配套成熟环境优美交通便利安全性高繁华地段低密度小区中型社区自住率高"
[2] "地铁房中型社区"
[3] "地铁房学区房配套成熟环境优美交通便利安全性高繁华地段中型社区"
[4] "地铁房学区房环境优美停车位充足中型社区"
[5] "学区房配套成熟环境优美交通便利投资首选繁华地段高品质小区低密度小区停车位充足大型社区自住率高"
[6] "学区房配套成熟环境优美次新小区交通便利投资首选安全性高低密度小区停车位充足中型社区"
```

我们需要从评论中提取出有用的信息。comments 处理的标准为：

1）删除带主观倾向的描述；

2）删除信息已经反映在其他数值化变量中的描述删除信息已经反映在其他

数值化变量中的描述；

3）删除没有明确标准的描述。

3.3.1.1 R 语言代码及注释

```
com=df$comments
com=gsub("风水宝地", "", com)#主观
com=gsub("住户素质高", "", com)#无标准
com=gsub("交通便利", "", com)#地铁
com=gsub("生活便利", "", com)#无标准
com=gsub("投资首选", "", com)#主观
com=gsub("配套成熟", "", com)#无标准
com=gsub("环境优美", "", com)#主观
com=gsub("高品质小区", "", com)#无标准
com=gsub("安全性高", "", com)#无标准
com=gsub("园林小区", "", com)#绿化率
com=gsub("低密度小区", "", com)#无标准
com=gsub("次新小区", "", com)#停车位/住户数
com=gsub("繁华地段", "", com)#无标准
com=gsub("国际社区", "", com)#无标准
com=gsub("文明小区", "", com)#无标准
com=gsub("纯别墅社区", "", com)#物业类型 retype
com=gsub("自住率高", "", com)#无标准
com=gsub("大型社区", "", com)#用户数和面积
com=gsub("中型社区", "", com)#用户数和面积
com=gsub("小型社区", "", com)#用户数和面积
table(com)
剩余字段：
#品牌开发商     可替代开发商变量，尽管有主观性
#金牌物业       可替代物业公司变量，尽管有主观性
#停车位充足     可替代难以处理的变量 pkrate
#地铁房
#学区房

dvl=grepl("品牌开发商", com)
dvl[dvl==TRUE]=1      #是品牌开发商的设为 1，否的自动设为 0，下同

reco.=grepl("金牌物业", com)
```

```
reco.[reco.==TRUE]=1

parking=grepl("停车位充足", com)
parking[parking==TRUE]=1

subway=grepl("地铁房", com)
subway[subway==1]=1

school=grepl("学区房", com)
school[school==1]=1

#完成处理，更新字段
df$comments=com
df$dvl=dvl
df$reco.=reco.
df$parking=parking
df$subway=subway
df$school=school
```

3.3.2 tothh 处理

tothh 是指楼盘的总户数，原数据如下。我们需要将里面的数字提取出来。

3.3.2.1 R 语言代码及注释

```
head(df$tothh)
[1] "320 户"  NA  "420 户"  "88 户"  "776 户"  "100 户"
#匹配数字，并转换为数值型变量
df$tothh=as.numeric(str_extract_all(df$tothh,"[0-9]+"))
```

3.3.3 totarea 处理

totarea 是指楼盘的总面积，里面存在单位不统一（有的是平方米，有的是万平方米）、多个面积等问题。

3.3.3.1 R 语言代码及注释

```
#匹配小数或整数，生成的是列表
ta=str_extract_all(df$totarea,"[0-9]+.[0-9]+|[0-9]+")
ta[853]=28.4                    #第 853 个值有三个数，取第一个
```

```
ta1=rep(0,1356)
for(i in 1:1356)ta1[i]=as.numeric(ta[i])      #把列表设法存为向量，并转换为数值型
ta2=ta1[!(is.na(ta1))]                        #下面设法将单位统一为"万平方米"
ta2[ta2<100]=ta2[ta2<100]*10000               #数值少于 100 的单位肯定是万平方米
ta2=ta2/10000
ta1[!(is.na(ta1))]=ta2
#完成处理
df$totarea=ta1
```

3.3.4 blddate 处理

blddate 是指楼盘的开工日期，我们只需提取年份。

3.3.4.1 R 语言代码及注释

```
head(df$blddate)
[1] "1998/04/01" NA "2000/06/01" "2015/06/03" "2015/01/08" "2015/10/10"
#提取年份
year=substr(df$blddate,1,4)
#完成处理
df$year=year
```

3.3.5 refee 处理

refee 是指楼盘的物业费。里面存在单位不统一、多个物业费、不同楼层的物业费等问题。

3.3.5.1 R 语言代码及注释

```
head(df$refee)
[1] "0.44/平方米"          "0.9"          "1.2 元每月"          "1.2 每平米每月"
"1.85 元/平米·月"          "1.5 元"

#处理可疑数字
rf=df$refee
rf=gsub("-", "abcdf", rf)
rf=gsub("1 楼", "", rf)
rf=gsub("2 楼", "", rf)
rf=gsub("1/", "", rf)
```

```
p=grepl("毛", rf)
rf[p]=c(0.6,0.2)

#提取剩下数字
rf2=str_extract_all(rf,"[0-9]+.[0-9]+|[0-9]+",simplify=TRUE)
rf3=matrix(nrow=1356,ncol=4)
for(i in 1:1356){
    for(j in 1:3){
        rf3[i,j]=as.numeric(rf2[i,j])
    }
}
#用小区物业费的均值做分析
rf3[,4]=rowMeans(rf3[,1:3],na.rm=TRUE)

#完成处理
df$refee=rf3[,4]
```

3.3.6 retype 处理

retype 是楼盘的物业类型，我们把该变量转换成五个因子，分别代表：别墅、新里洋房、公寓、老公房和其它。

3.3.6.1 R 语言代码及注释

```
#删掉“、”，并取第一个物业类型为准
tmp.type<-strsplit(df$retype,split = "、")
for(i in 1:length(df$retype)){
    df$retype[i]=tmp.type[[i]][1]
}
table(df$retype)

#完成处理，转为因子
df$retype[df$retype=='别墅']=5
df$retype[df$retype=='新里洋房']=4
df$retype[df$retype=='公寓'|df$retype=='普通住宅']=3
df$retype[df$retype=='老公房']=2
df$retype[df$retype=='其它']=1
```

3.3.7 greenrate 处理

greenrate 是指楼盘的绿化率，需要将里面的“绿化率高”等字符串删除，最后统一用小数表示。

3.3.7.1 R 语言代码及注释

```
tmp.green<-strsplit(df$greenrate,split = "%")
for(i in 1:1356){
    df$greenrate[i]=as.numeric(tmp.green[[i]][1])/100
}
```

3.3.8 plotrate 处理

plotrate 是指楼盘的容积率，只需转化为数值即可。

3.3.8.1 R 语言代码及注释

```
df$plotrate=as.numeric(df$plotrate)
```

3.3.9 房价处理

房价包含了 2012 年到 2014 年每个月的小区房价（comm）和板块房价（sect），我们计算了每年房价的均值作为因变量。

3.3.9.1 R 语言代码及注释

```
#计算每年的小区房价均值
df["comm2012"]<-rowMeans(df[21:32],na.rm=TRUE)
df["comm2013"]<-rowMeans(df[33:44],na.rm=TRUE)
df["comm2014"]<-rowMeans(df[45:56],na.rm=TRUE)
#计算每年的板块房价均值
df["sect2012"]<-rowMeans(df[57:68],na.rm=TRUE)
df["sect2013"]<-rowMeans(df[69:80],na.rm=TRUE)
df["sect2014"]<-rowMeans(df[81:92],na.rm=TRUE)
```

3.3.10 cnty 处理

cnty 是指楼盘所在区县，我们将其转化为 4 个因子水平：闵行区=1、徐汇区=2、长宁区=3 和其他区=4。

3.3.10.1 R 语言代码及注释

```
table(df$cnty)
```

奉贤区	嘉定区	闵行区	浦东新区	徐汇区	长宁区
4	2	1176	11	126	37

```
df$cnty[df$cnty=="闵行区"]=1
df$cnty[df$cnty=="徐汇区"]=2
df$cnty[df$cnty=="长宁区"]=3
df$cnty[df$cnty=="奉贤区"|df$cnty=="嘉定区"|df$cnty=="浦东新区"]=4
```

3.4 生成新的数据 data

3.4.1 R 语言代码

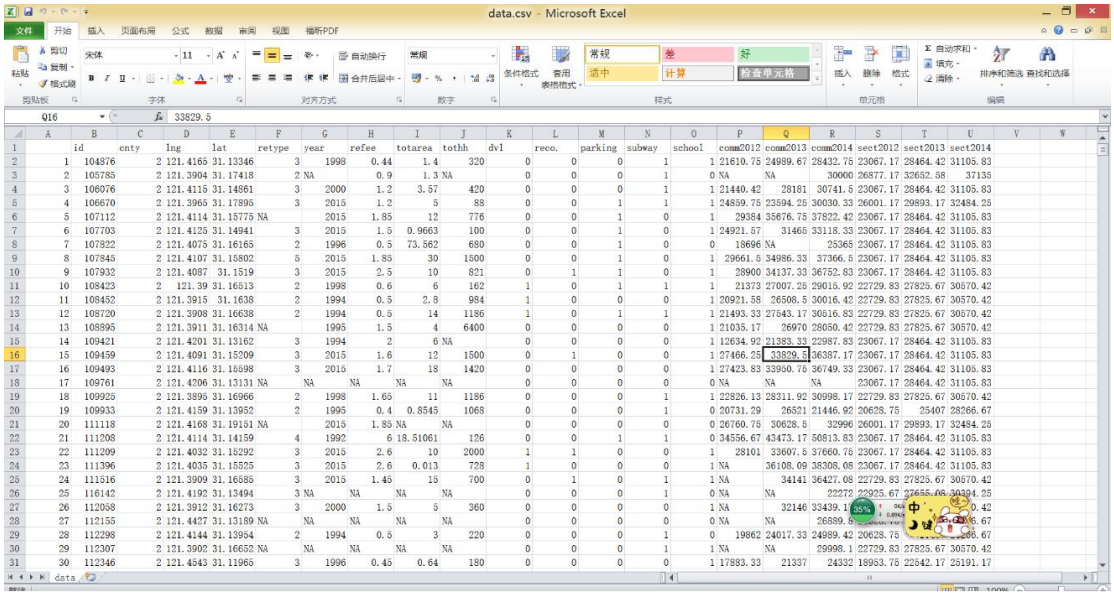
```
names(df)
```

```
 [1] "X"          "id"         "cnty"       "poicode"    "lng"        "lat"        "name"       "addr"
 [9] "section"    "developer"  "re"         "retype"     "refee"      "totarea"    "tothh"      "blddate"
[17] "plotrate"   "pkrate"     "greenrate"  "comments"   "comm201203" "comm201204" "comm201205" "comm201206"
[25] "comm201207" "comm201208" "comm201209" "comm201210" "comm201211" "comm201212" "comm201301" "comm201302"
[33] "comm201303" "comm201304" "comm201305" "comm201306" "comm201307" "comm201308" "comm201309" "comm201310"
[41] "comm201311" "comm201312" "comm201401" "comm201402" "comm201403" "comm201404" "comm201405" "comm201406"
[49] "comm201407" "comm201408" "comm201409" "comm201410" "comm201411" "comm201412" "comm201501" "comm201502"
[57] "sect201203" "sect201204" "sect201205" "sect201206" "sect201207" "sect201208" "sect201209" "sect201210"
[65] "sect201211" "sect201212" "sect201301" "sect201302" "sect201303" "sect201304" "sect201305" "sect201306"
[73] "sect201307" "sect201308" "sect201309" "sect201310" "sect201311" "sect201312" "sect201401" "sect201402"
[81] "sect201403" "sect201404" "sect201405" "sect201406" "sect201407" "sect201408" "sect201409" "sect201410"
[89] "sect201411" "sect201412" "sect201501" "sect201502" "dvl"        "reco."      "parking"    "subway"
[97] "school"     "year"       "comm2012"   "comm2013"   "comm2014"   "sect2012"   "sect2013"   "sect2014"
```

```
data=df[,c("id","cnty","lng","lat","retype","year","refee","totarea","tothh","dvl","reco.","parking","subway","school","plotrate","greenrate","comm2012","comm2013","comm2014","sect2012","sect2013","sect2014")]
write.csv(data,"data.csv")
```

3.4.2 解释说明

代码提取了 df 中经过处理后的某些列，新生成的数据为“data.csv”，截图如下：



里面共有 23 个变量，各个变量解释如下表所示。其中最后 6 个变量 comm2012、comm2013、comm2014、sect2012、sect2013、sect2014 即每年房价的平均数据，是因变量，前面的所有变量都是自变量。

变量名	变量解释
X：	序号
id：	楼盘编号
cnty	区县（"闵行区"=1，"徐汇区"=2，"长宁区"=3，"其他区"=4）
lng：	经度
lat：	维度
retype：	物业类型（'别墅'=5，'新里洋房'=4，'公寓'=3，'老公房'=2，'其他'=1）
year：	建造年份
refee：	物业费
totarea：	建造总面积（万平方米）
tothh：	总户数
dvl：	是否品牌开发商
reco.：	是否金牌开发商
parking：	停车位是否充足
subway：	是否为地铁房
school：	是否为学区房
plotrate：	楼盘容积率

greenrate：	楼盘绿化率
comm2012：	2012 年小区均价
comm2013：	2013 年小区均价
comm2014：	2014 年小区均价
sect2012：	2012 年板块均价
sect2013：	2013 年板块均价
sect2014：	2014 年板块均价

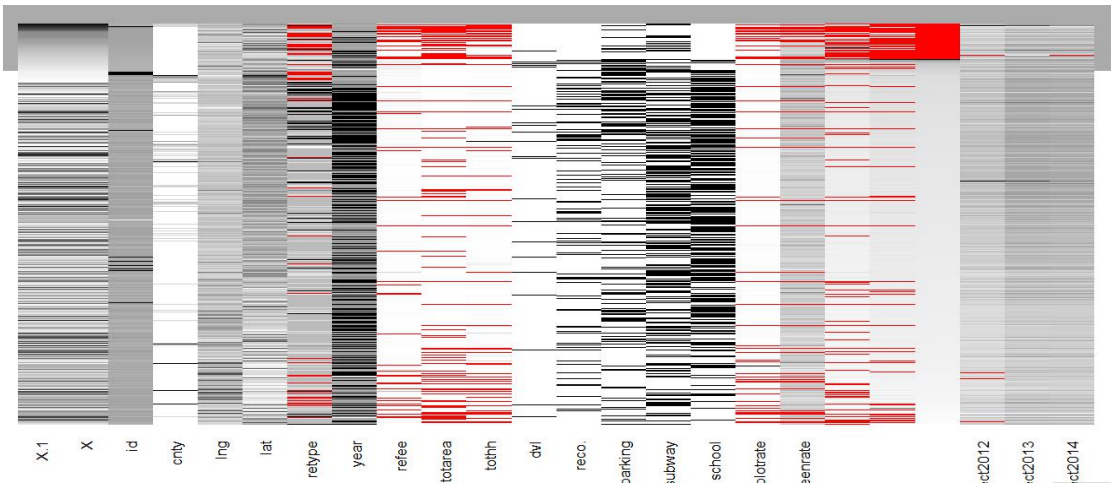
四、缺失值插补

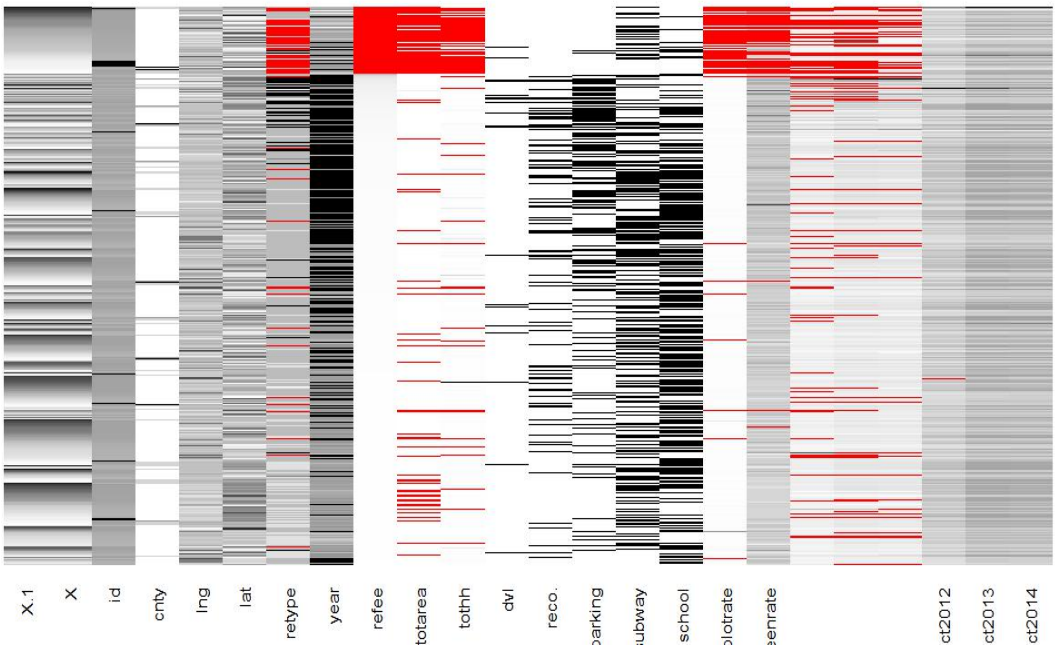
4.1 缺失值探寻

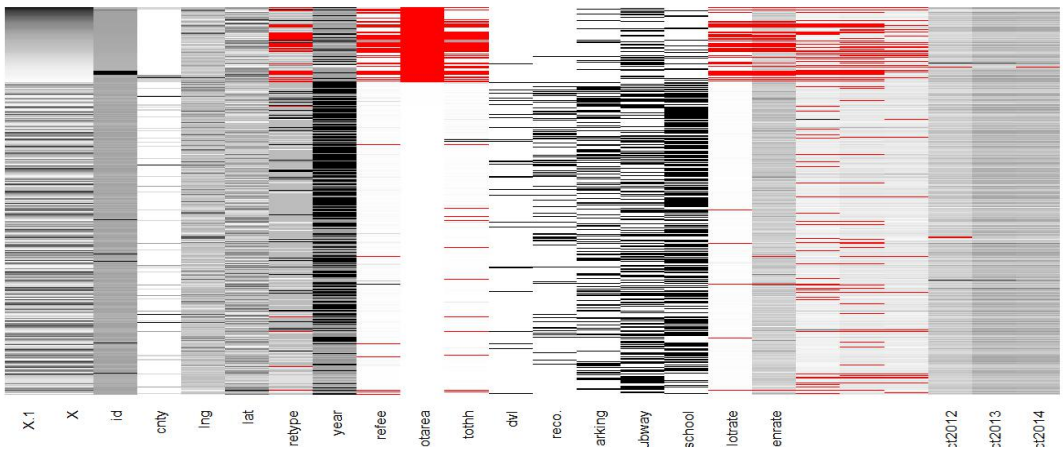
```
setwd("F:\\迅雷下载\\房价预测项目数据\\test—Excel 格式")
data=read.csv("data.csv")
library(mice)
head(md.pattern(data))
tail(md.pattern(data))
library(VIM)
aggr(data)
```

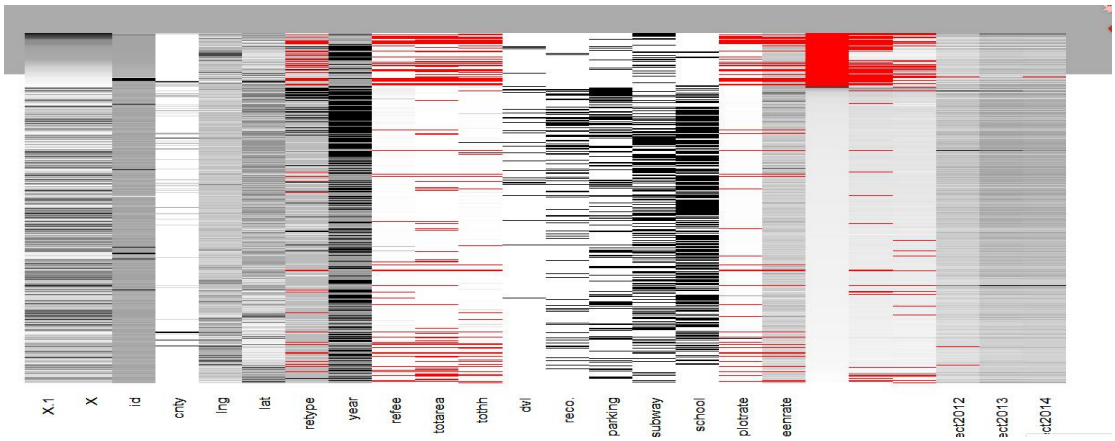
我们从中挑选了几幅很典型的截图，可以看到，缺失值变量之间关系，缺失

互联性太强，从而我们也可以得到一个结论，有可能我们利用回归模型，决策树等模型拟合方法去预测缺失值这种方法根本行不通。虽然如此我也做了一些猜想和分析思路。









经纬度计算距离的方法：

计算原则是尽可能的简化复杂的过程，将上海市直接作为一个平面处理，距离就按照直线距离来计算，最后乘上 1 度的大约值，为了方便也直接忽略了本身的度量。

4.2 对于 year 的缺失值处理

思路：一般建筑年代越久远通常情况下，我们才可能会丢失其相关的数据，所以插值时我们采用的是其周边建筑物建造年份的最早的作为其插补值。

```
for (i in 1:length(data$year[is.na(data$year)]))
{
dist=sqrt((data$lng[is.na(data$year)][i]-data$lng)^2+(data$lat[is.na(data$year)][i]-da
ta$lat)^2);
data$year[is.na(data$year)][i]=min(data$year[which(dist>0&
dist<quantile(dist,0.01,na.rm = T))],na.rm = T);
i=i+1
}
sum(is.na(data$year))
```

另外要将上述的代码重复 7 次才可以完全将缺失值插补完整，这是因为其周围的建筑物可能都缺失建造年限，只能都各个相邻的反复迭代完成，当然也可以定义一个函数，反复的判断循环完成。

4.3 对于 greenrate 字段缺失值处理

根据周围的平均情况进行插值，同样需要重复几次过程才可将所有缺失值填补完整。

```
for (i in 1:length(data$ greenrate [is.na(data$ greenrate)]))
{
dist=sqrt((data$lng[is.na(data$ greenrate)][i]-data$lng)^2+(data$lat[is.na(data$ greenrate)][i]-data$lat)^2);
if(mean(data$ greenrate [which(dist>0 & dist<quantile(dist,0.1,na.rm =T))],na.rm = T)>0)
data$ greenrate [is.na(data$ greenrate)][i]=mean(data$ greenrate [which(dist>0 & dist<quantile(dist,0.1,na.rm =T))],na.rm = T);
i=i+1
}
sum(is.na(data$ greenrate))
write.csv(data,"data.csv")
```

五、描述性统计

5.1 单变量的描述性统计

5.1.1 物业费（元每平方米每月）

```
> summary(df$refee)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   0.500   1.000   1.458   1.500  50.000     161
```

均值由图可知，物业费的均值为 1.458。

5.1.2 总面积（万平方米）

```
> summary(df$totarea)
    Min.  1st Qu.   Median    Mean  3rd Qu.     Max.    NA's
   0.012    0.892    4.700   20.310   11.560 8163.000     262
```

由图可知，总面积的均值为 20.31。

5.1.3 总户数（户）

```
> summary(df$tothh)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    8.0   258.0   560.0   959.8   936.0 50000.0     199
```

由图可知，总户数的均值为 959.8。

5.1.4 容积率

```
> summary(df$plotrate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.0181  1.2000  1.5000  1.6890  1.8000 80.0000     159
```

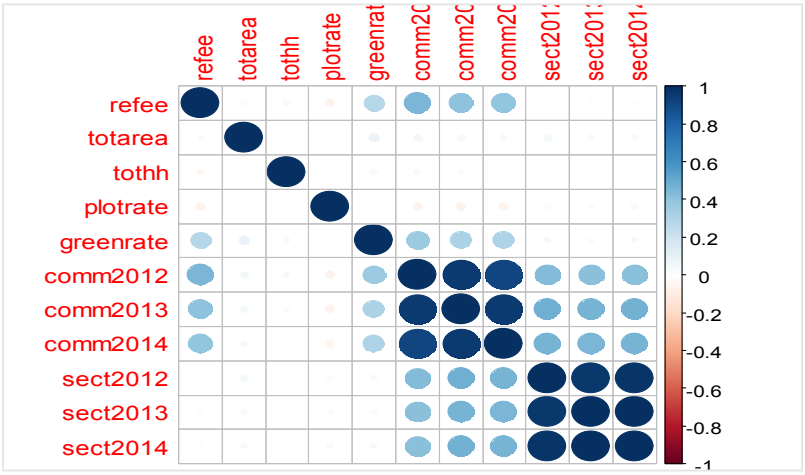
由图可知，容积率的均值为 1.689。

5.1.5 绿化率

```
> summary(df$greenrate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.0120  0.3000  0.3500  0.3658  0.4000  0.9000     129
```

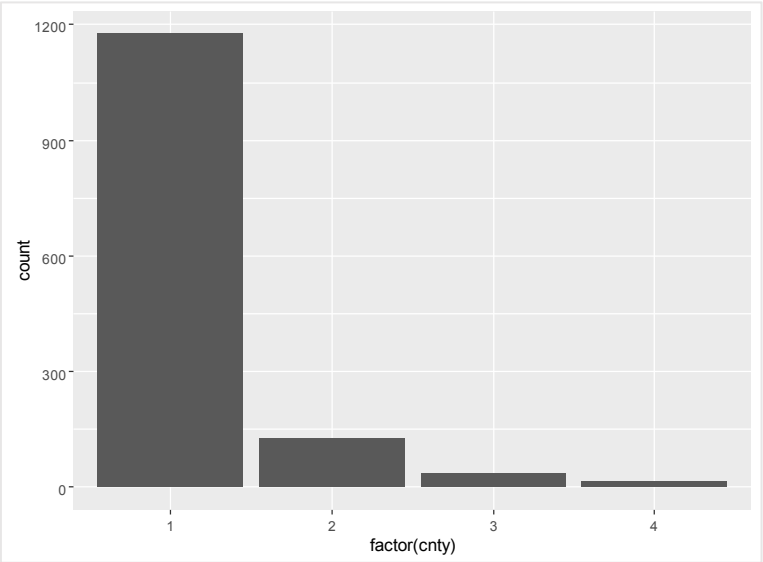
由图可知，绿化率的均值为 0.3658。

5.2 相关矩阵图

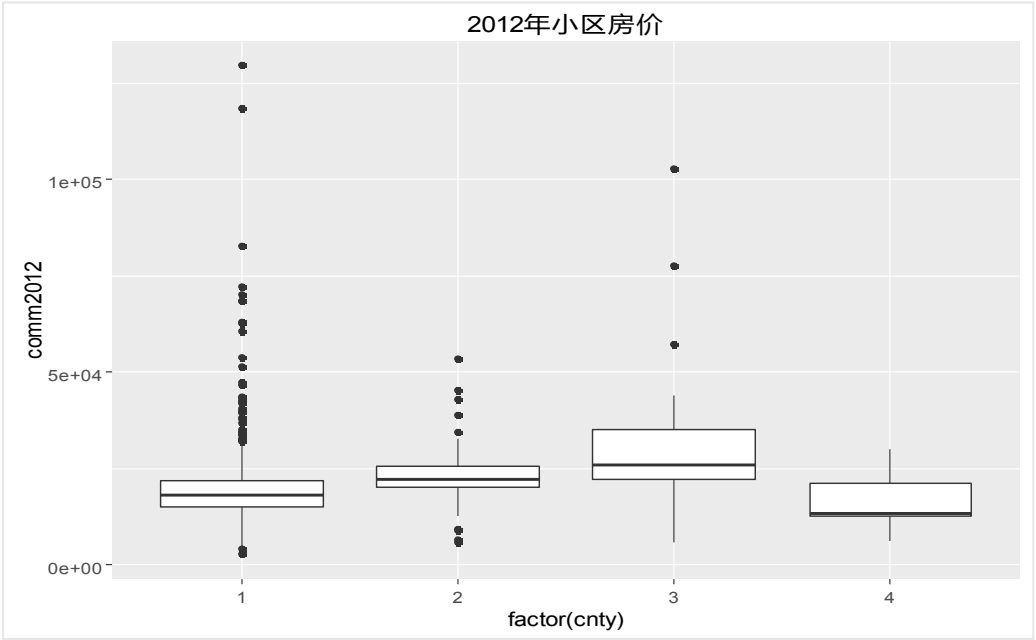


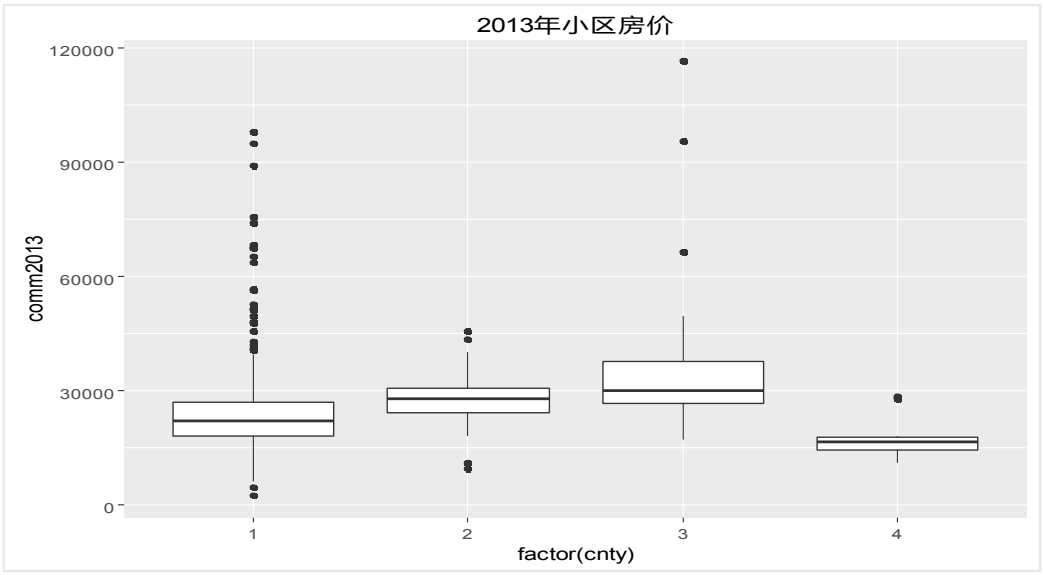
从相关矩阵图中可知，物业费与绿化率有较强的相关性，物业费对小区房价有一定的影响，但对板块房价没有影响；绿化率也只影响小区房价，不影响板块房价。三年的小区房价有极强的相关性，三年的板块房价也有极强的相关性。三年小区房价与三年板块房价也有较强的相关性。

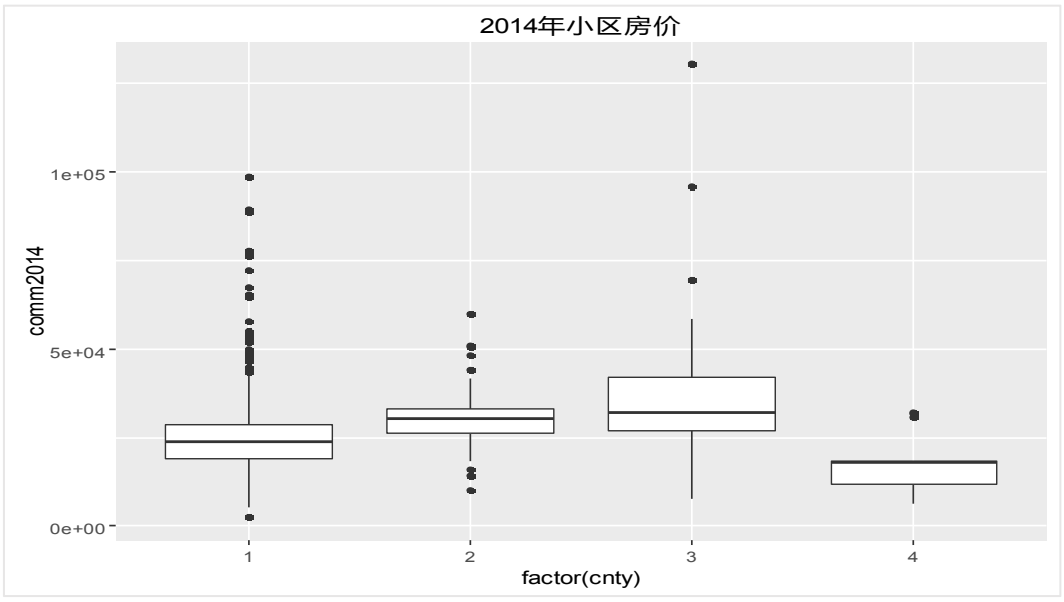
5.3 上海不同区每年房价的分布



上图显示了不同区所占的房价数据的数量分布。主要查看闵行区及周边的房价，闵行区的数量最多，其次是徐汇区，长宁区，其他区（奉贤区、嘉定区、浦东新区）

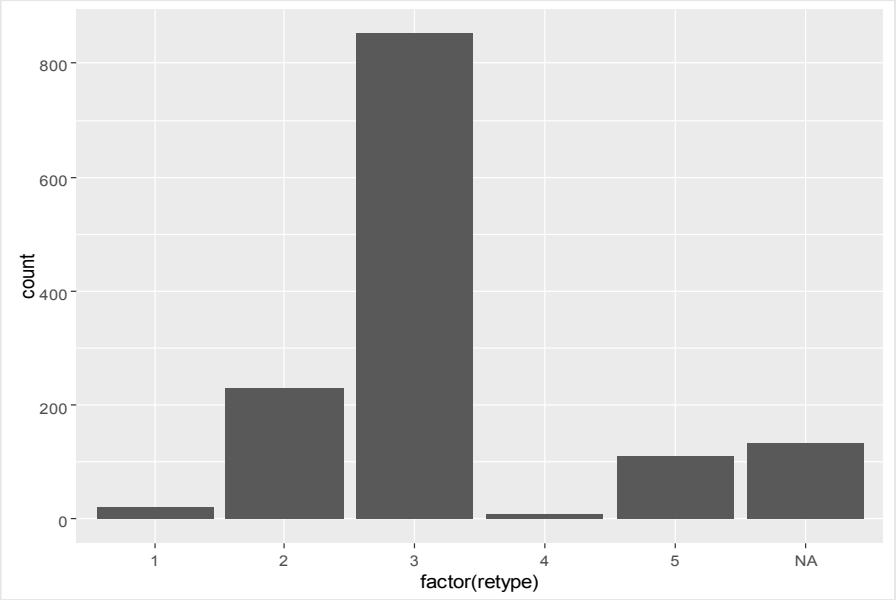




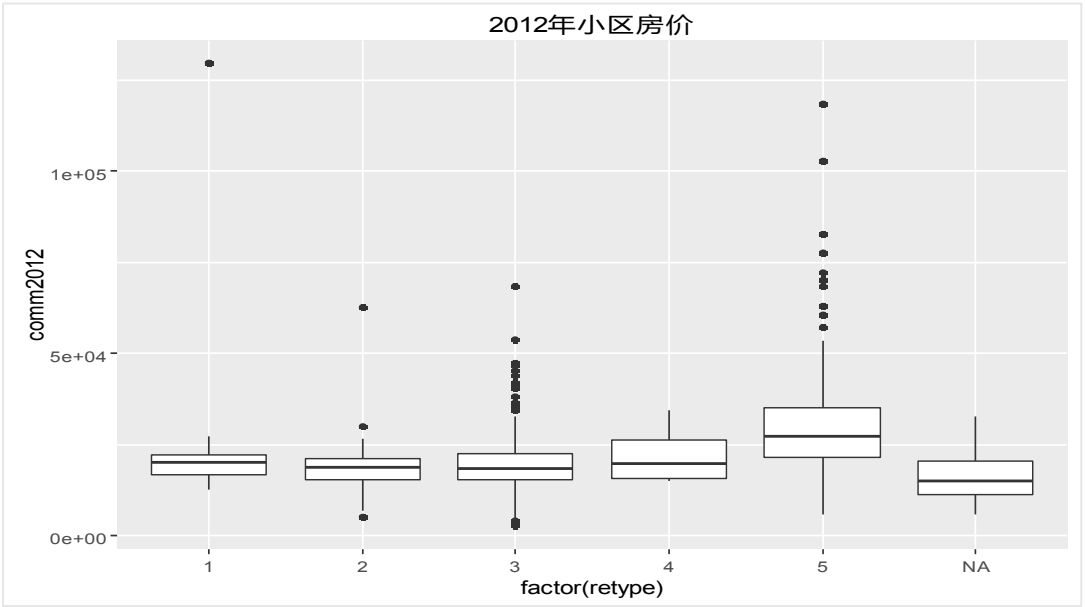


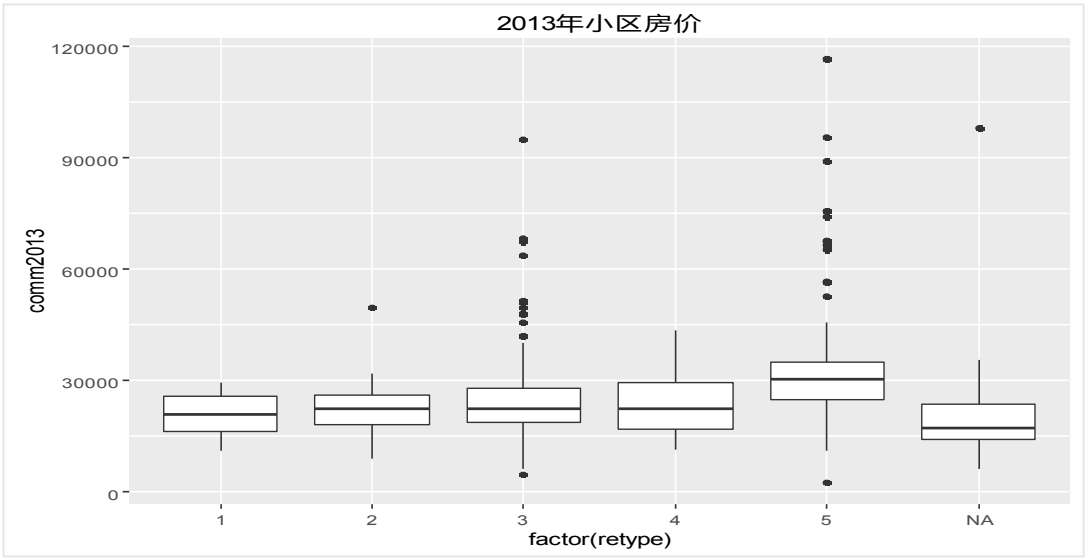
由上图，可以看出 2012-2014 年上海各区（"闵行区"=1，"徐汇区"=2，"长宁区"=3，"其他区"=4）中，长宁区的小区房价均值最高，其他区（奉贤区、嘉定区、浦东新区）的小区房价均值最低。

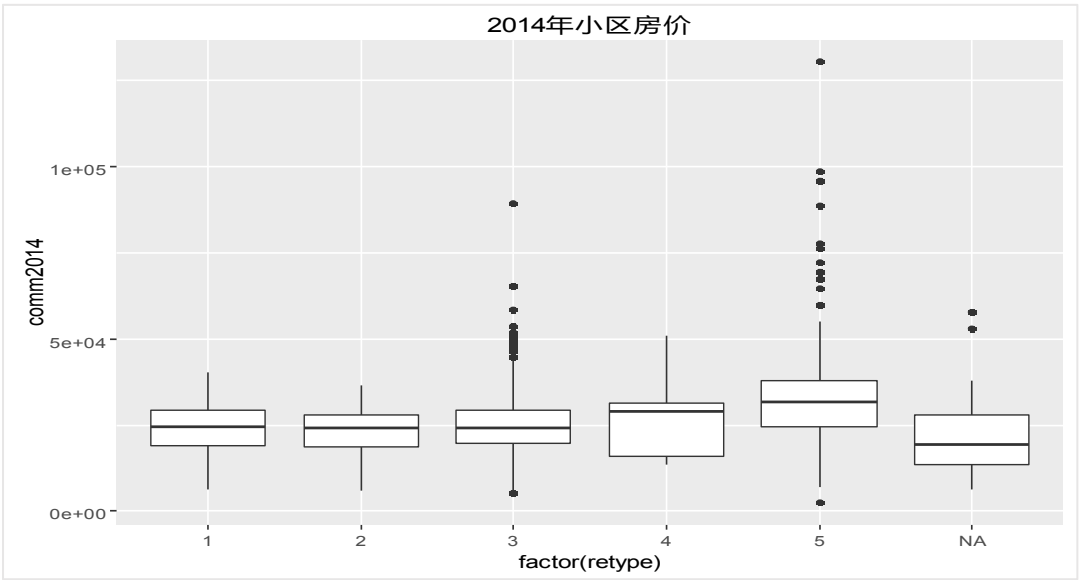
5.4 不同物业类型的房价情况



上图显示了不同物业类型（'别墅'=5，'新里洋房'=4，'公寓'=3，'老公房'=2，'其他'=1）所占的房价数据的数量分布。公寓的数量最多，其次是老公房，别墅，其他，新里洋房。

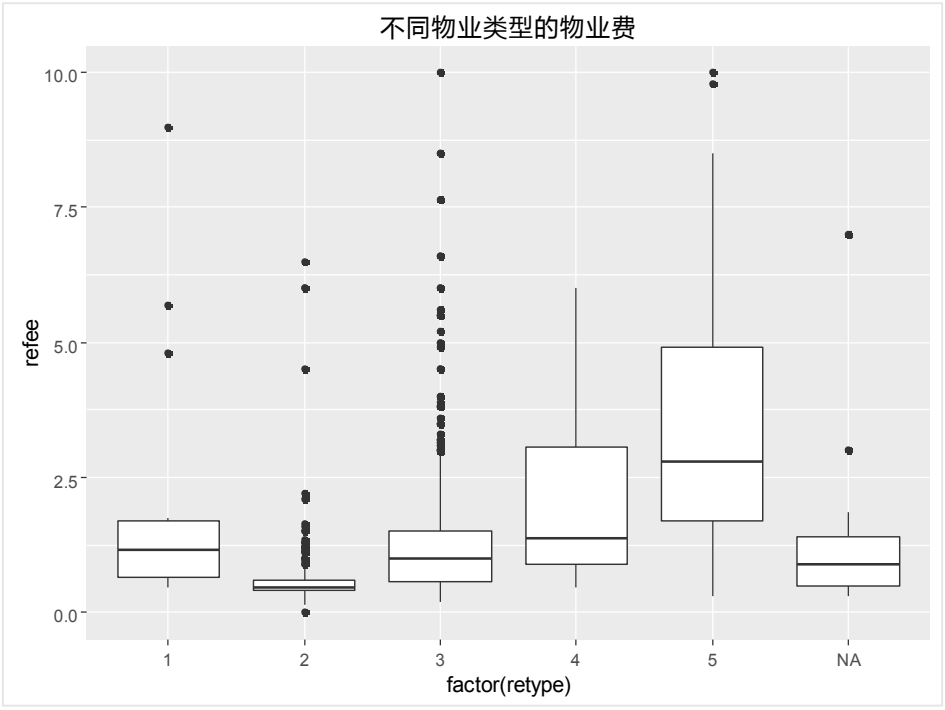






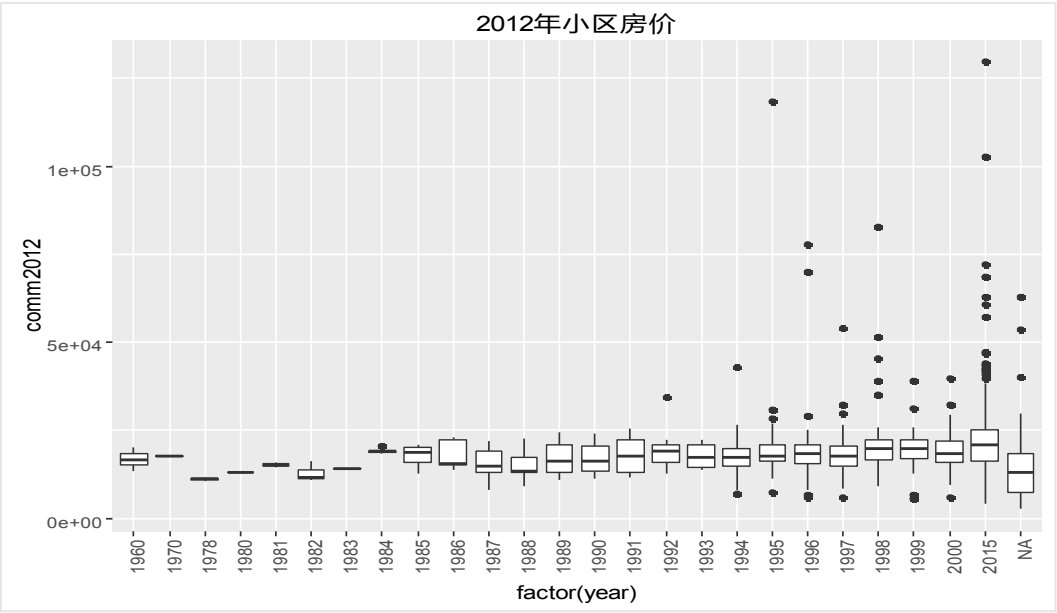
由上图，可以看出 2012-2014 年上海不同物业类型（'别墅'=5，'新里洋房'=4，'公寓'=3，'老公房'=2，'其他'=1）中，别墅的小区房价均值最高。

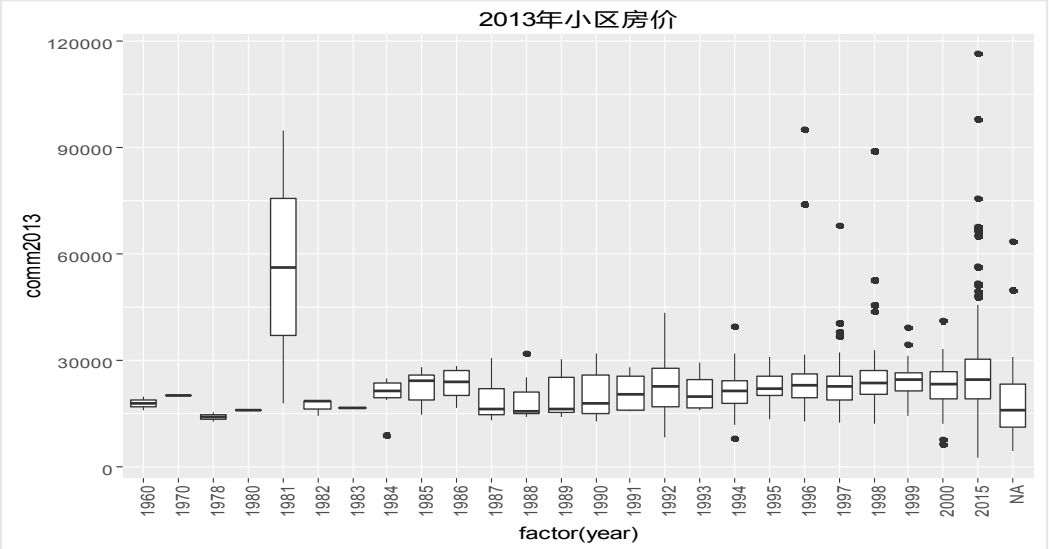
5.5 不同物业类型的物业费情况

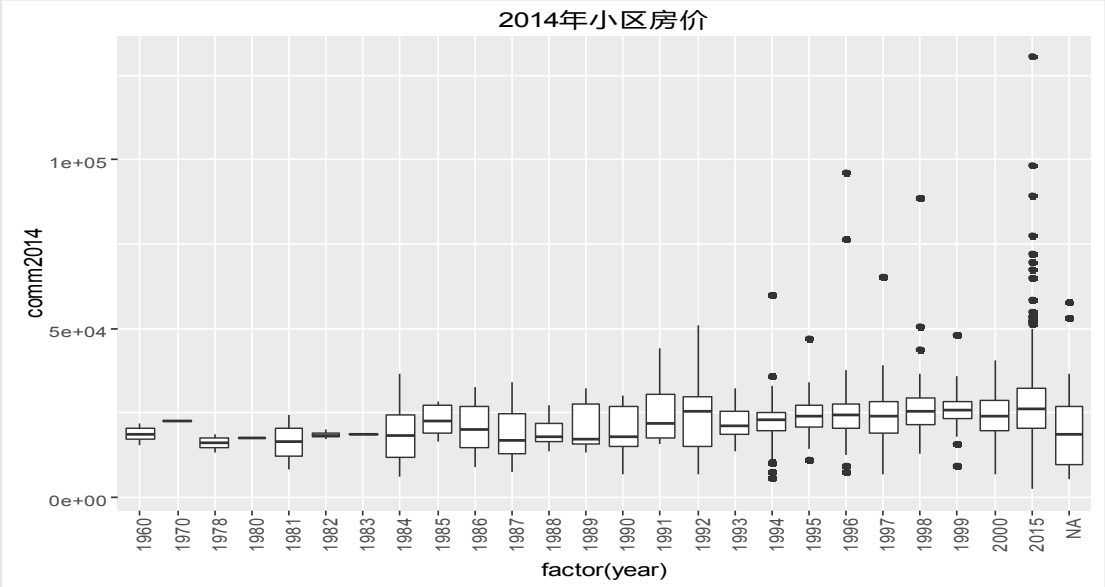


由上图，可以看出上海不同物业类型（'别墅'=5，'新里洋房'=4，'公寓'=3，'老公房'=2，'其他'=1）中，别墅物业费最高，其次是新里洋房、其他、公寓，老公房物业费最低

5.6 不同建造年份的房价

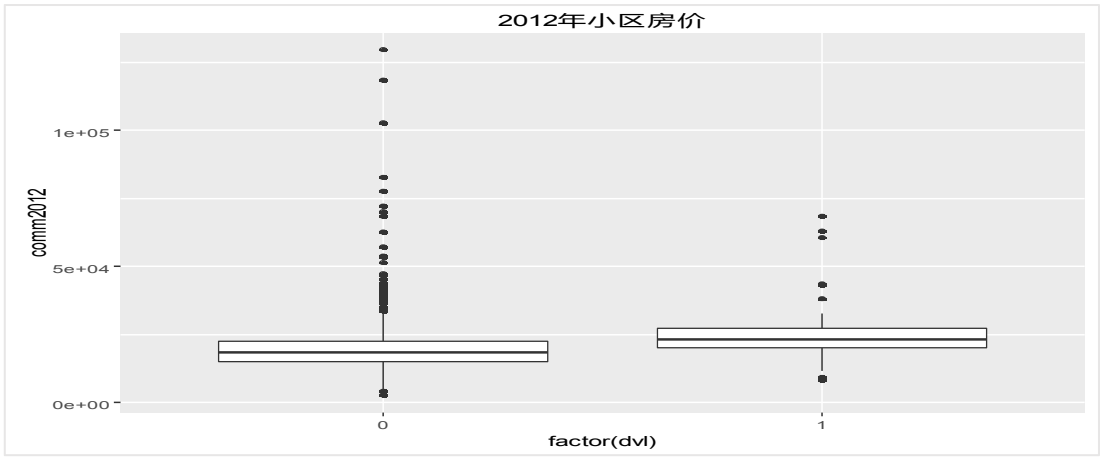


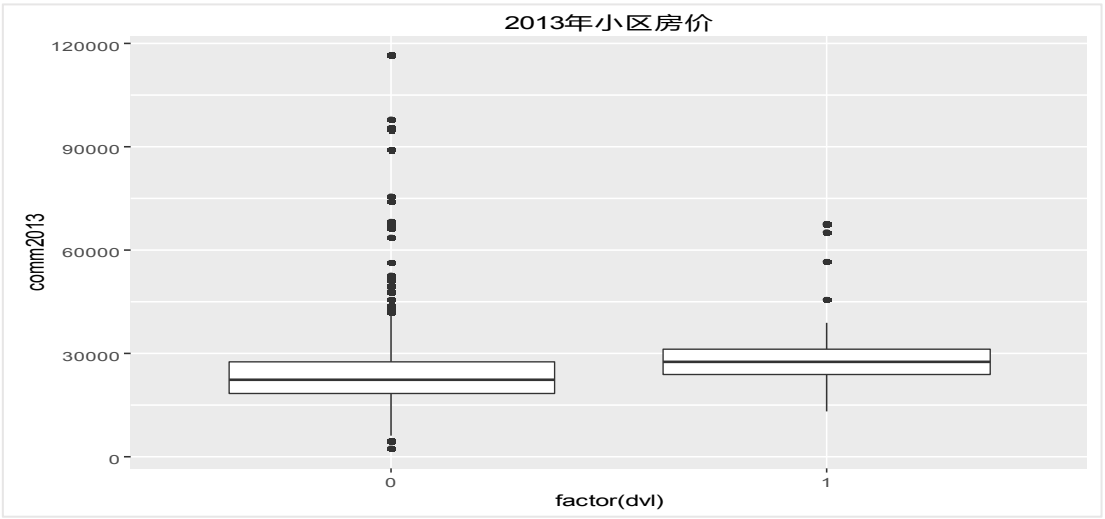


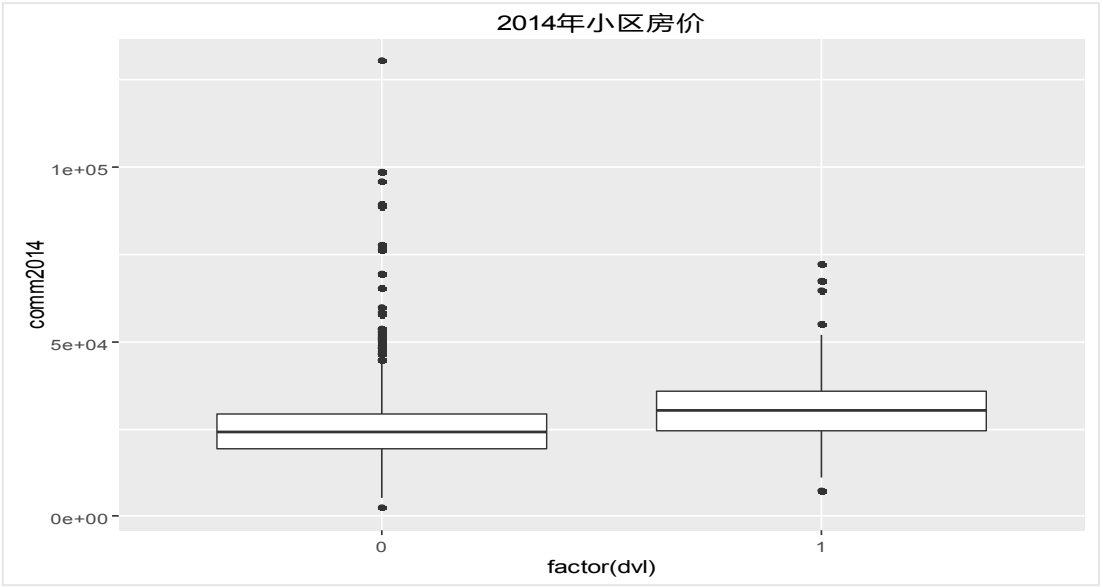


从图中，可以看出 2012-2014 年的小区房价随着建造年份越近有一定的增长趋势，但建造年份为 1981 年的 2013 年小区房价出现猛增现象。

5.7 是否为品牌开发商的分布

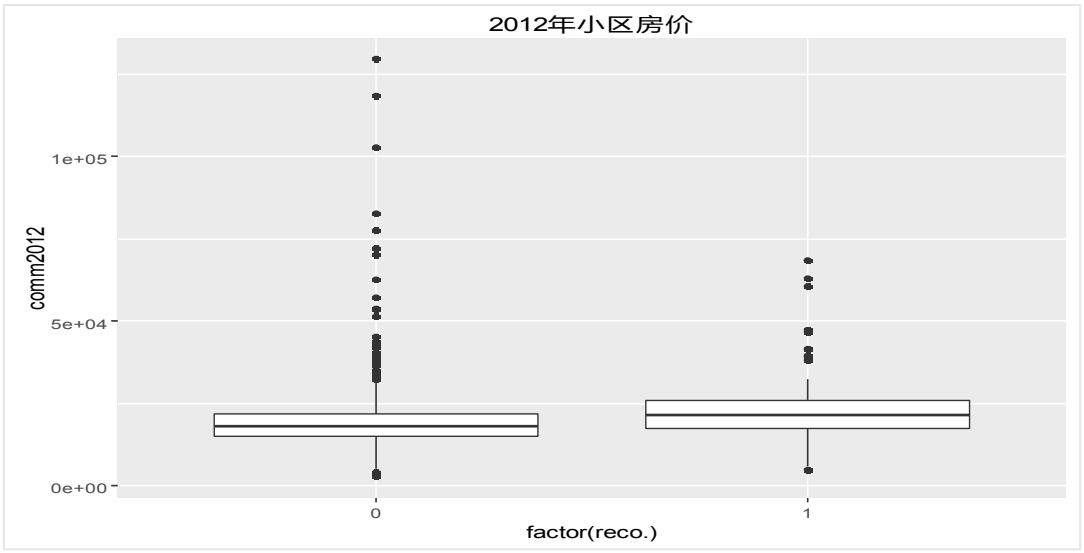


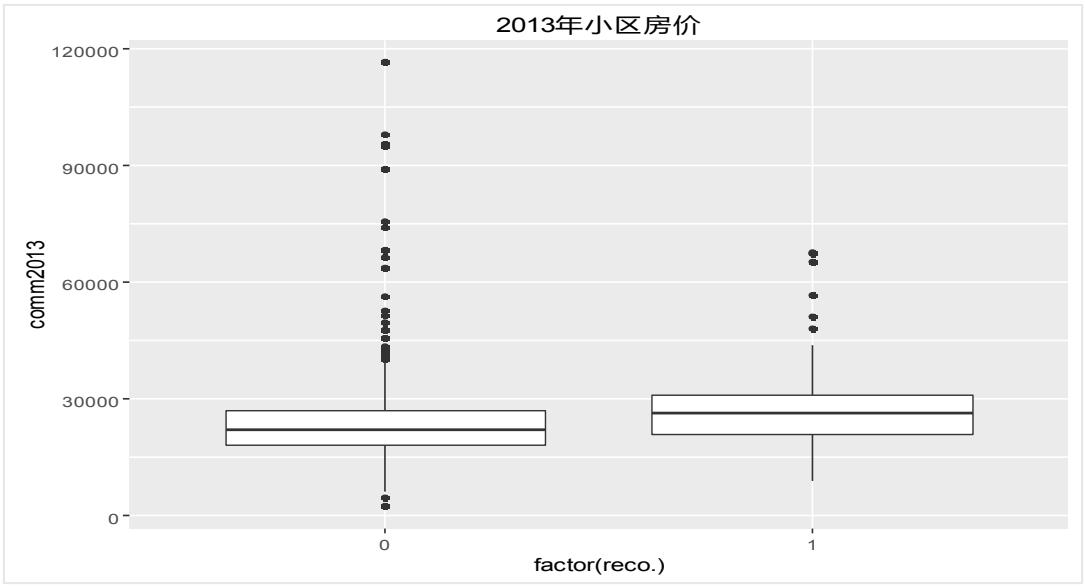


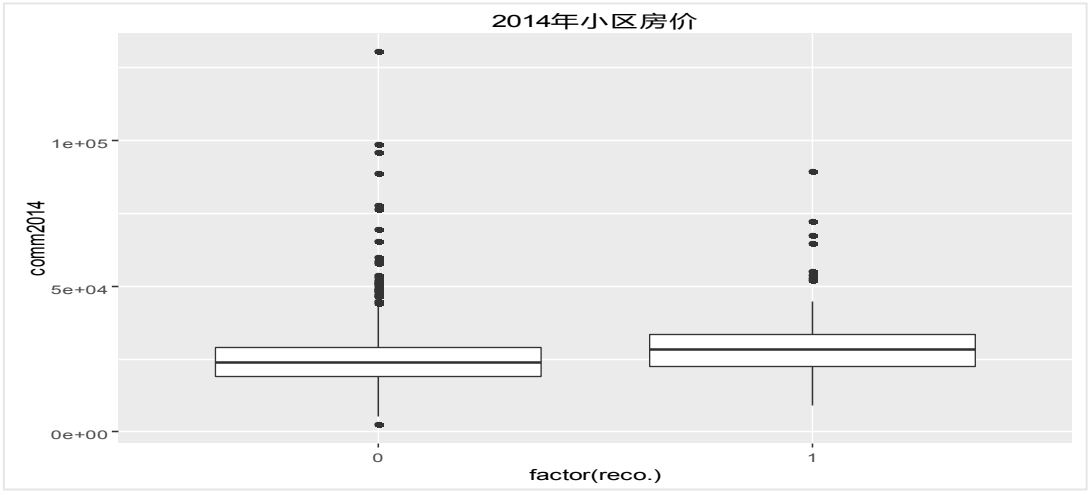


从图中看出，若开发商是品牌开发商，小区房价较高。

5.8 是否为金牌开发商

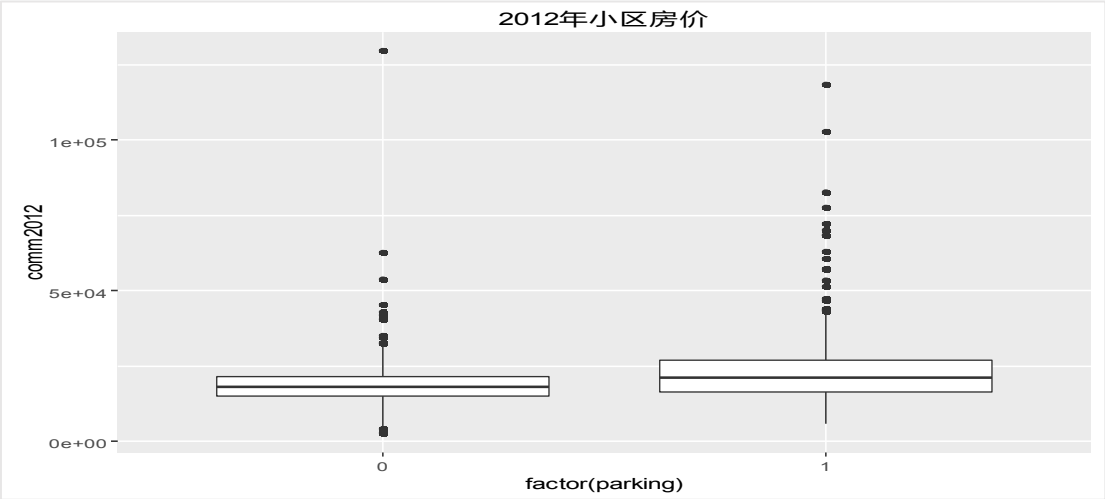

2012年小区房价
comm2012
1e+05
5e+04
0e+00
0
1
factor(reco.)

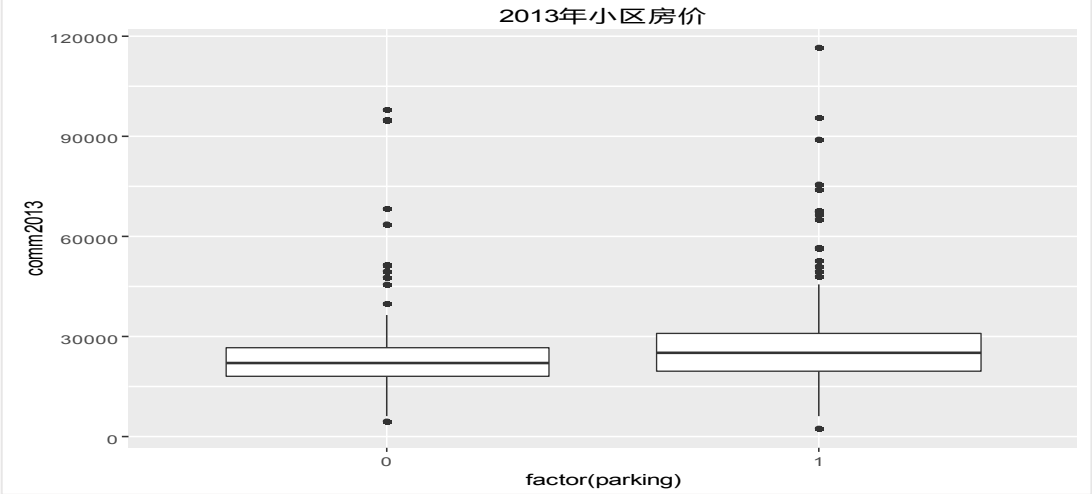

2013年小区房价
comm2013
120000
90000
60000
30000
0
0
1
factor(reco.)

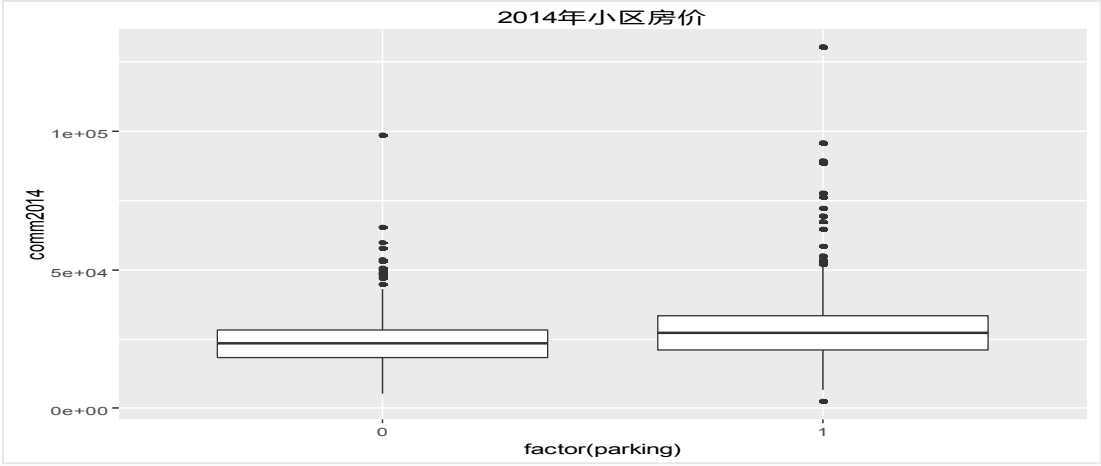

2014年小区房价
comm2014
1e+05
5e+04
0e+00
0
1
factor(reco.)

从图中看出，若开发商是金牌开发商，小区房价较高。

5.9 停车位是否充足

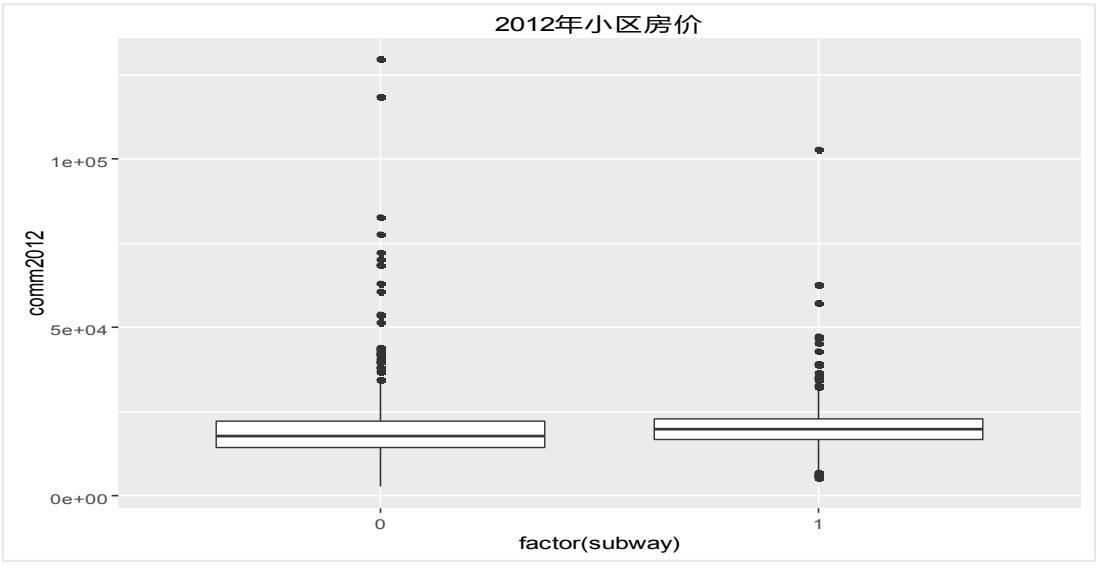


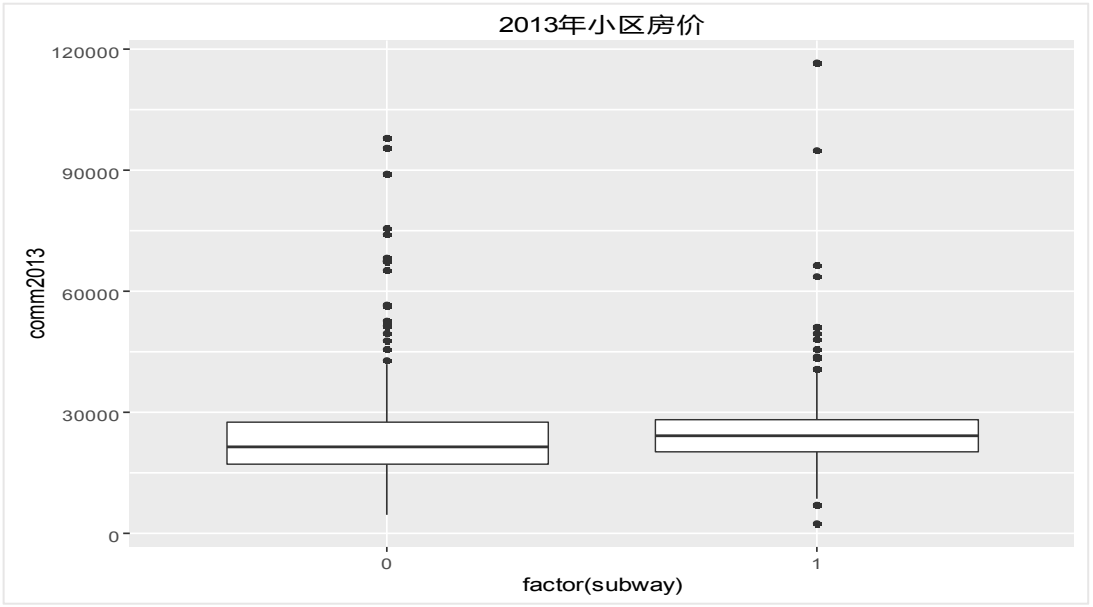


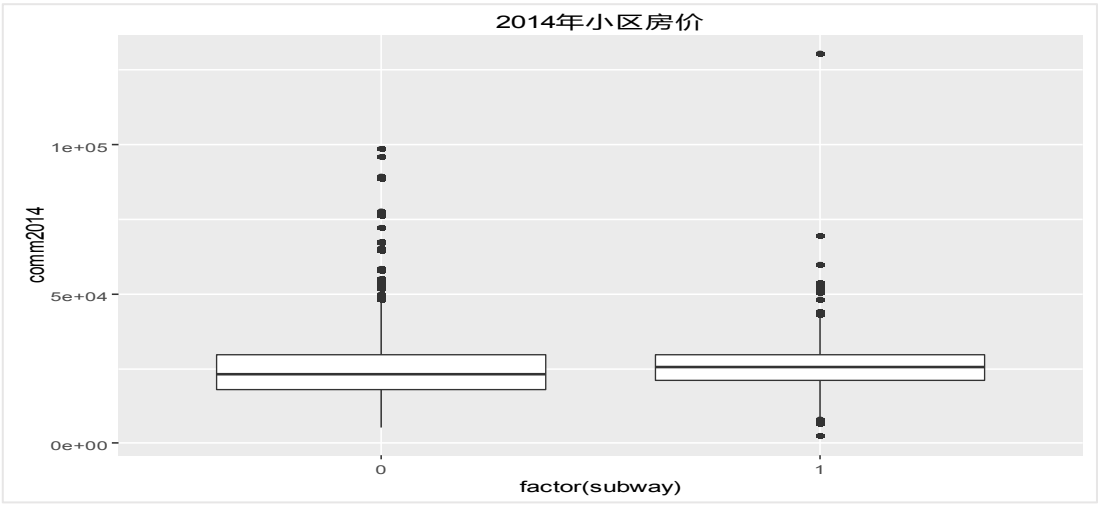


从图中看出，停车位充足，小区房价较高。

5.10 是否为地铁房

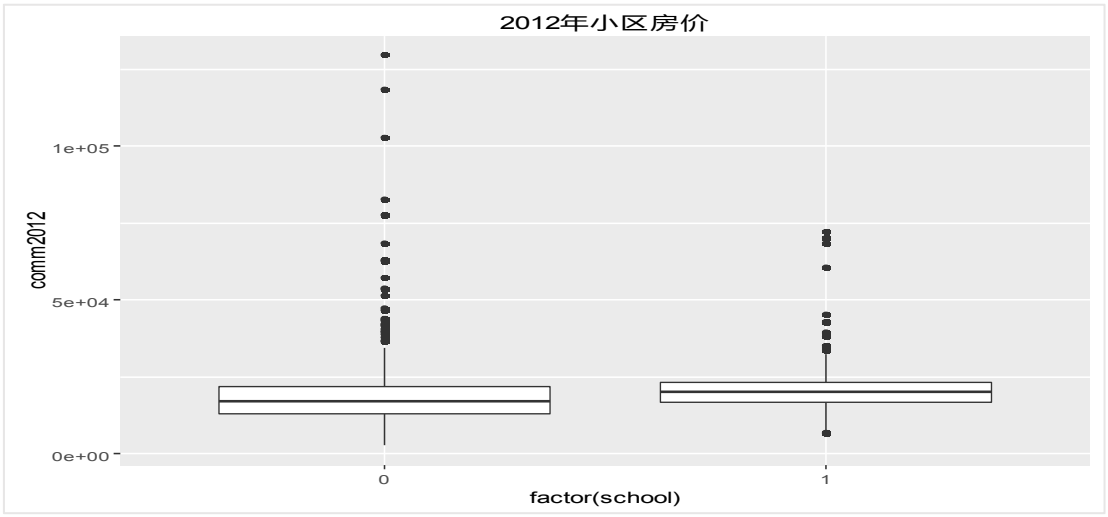

2012年小区房价
comm2012
1e+05
5e+04
0e+00
0
1
factor(subway)

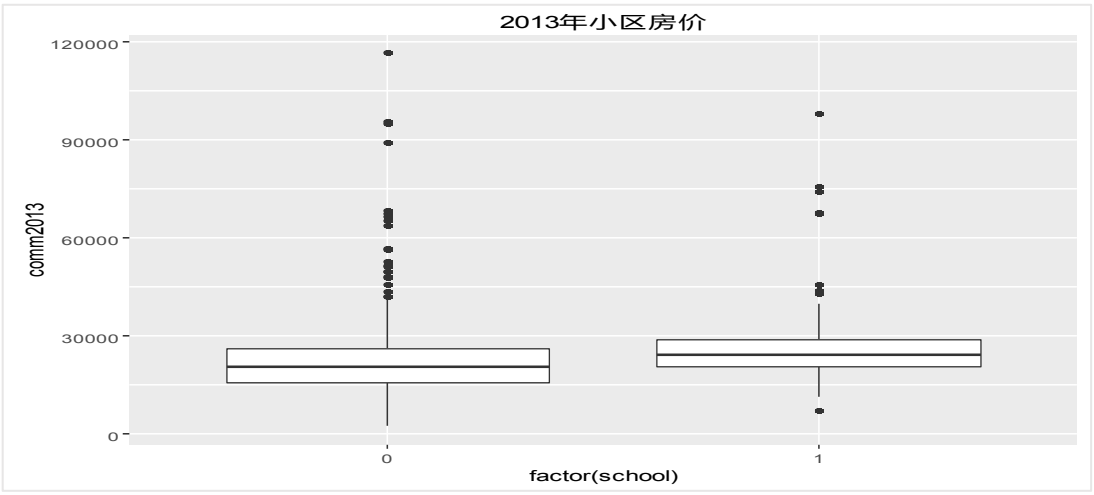

2013年小区房价
comm2013
120000
90000
60000
30000
0
0
1
factor(subway)

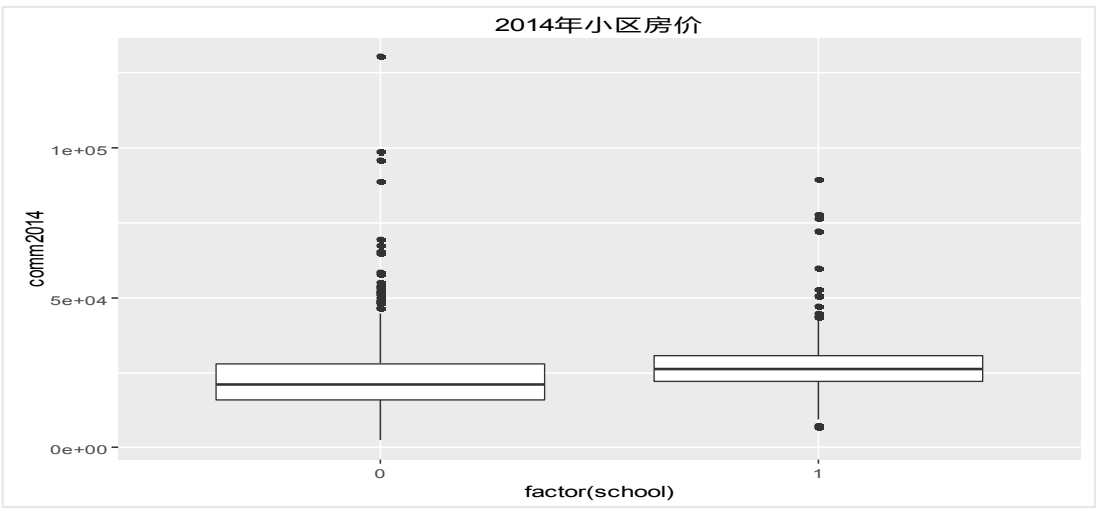

2014年小区房价
comm2014
1e+05
5e+04
0e+00
0
1
factor(subway)

从图中看出，若房子为地铁房，小区房价较高。

5.11 是否为学区房







从图中看出，若房子为学区房，小区房价较高。

六、模型建立与评估

6.1 模型选择

在国内外的许多文献中，使用的最多的房价预测模型是多元线性回归模型，本文也将会使用它。此外，决策树模型不仅简单易懂，也能够比较好的判断哪些变量对房价影响是比较显著的。而决策树的缺点在于准确度并不高，因此，有必要使用决策树的组合算法—随机森林模型进行补充。

6.2 多元线性回归建模

6.2.1 R 语言代码及注释

```
#读取数据
data<-read.csv("data.csv")
data0<-na.omit(data)
data1<-data0[c(1,3:23)]

#转变因子型
data1$ cnty <-as.factor(data1$cnty)
data1$retype<-as.factor(data1$retype)
data1$dvl<-as.factor(data1$dvl)
data1$reco.<-as.factor(data1$reco.)
data1$parking<-as.factor(data1$parking)
data1$subway<-as.factor(data1$subway)
data1$school<-as.factor(data1$school)

#模型建立，命名为 fit1
fit1<-lm(comm2013~cnty+retype+year+refee+totarea+tothh+dvl+reco.+parking+subway+school+plotrate+greenrate+comm2012,data= data1)

#输出模型相关参数
summary(fit1)
```

```
Call:
lm(formula = comm2013 ~ cnty + retype + year + refee + totarea +
    tothh + dvl + reco. + parking + subway + school + plotrate +
    greenrate + comm2012, data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-15490.0  -1153.4   -150.1    974.2  23271.6

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.113e+04  2.105e+04  -1.004  0.31566
cnty2        2.000e+02  3.094e+02   0.646  0.51816
cnty3        5.251e+02  6.026e+02   0.871  0.38380
cnty4       -1.379e+03  8.475e+02  -1.628  0.10394
retype2      1.628e+03  9.285e+02   1.754  0.07982 .
retype3      1.867e+03  8.919e+02   2.093  0.03660 *
retype4      2.615e+03  1.596e+03   1.638  0.10171
retype5      3.199e+02  9.574e+02   0.334  0.73833
year         1.179e+01  1.051e+01   1.122  0.26232
refee       -1.384e+02  6.401e+01  -2.163  0.03084 *
totarea     -1.933e-01  3.199e-01  -0.604  0.54575
tothh        1.258e-02  3.286e-02   0.383  0.70200
dvl1        -1.462e+02  4.289e+02  -0.341  0.73336
reco.1       1.033e+01  2.294e+02   0.045  0.96409
parking1     1.570e+02  2.121e+02   0.740  0.45929
subway1      5.204e+01  1.817e+02   0.286  0.77470
school1      2.728e+02  2.017e+02   1.353  0.17642
plotrate    -1.476e+01  2.733e+01  -0.540  0.58915
greenrate   -3.227e+03  1.138e+03  -2.835  0.00468 **
comm2012     1.053e+00  1.385e-02  76.013  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2637 on 871 degrees of freedom
Multiple R-squared:  0.9138,    Adjusted R-squared:  0.9119
F-statistic: 485.8 on 19 and 871 DF,  p-value: < 2.2e-16
```

#逐步回归

fit2=step(fit1)

summary(fit2)

```
Call:
lm(formula = comm2013 ~ retype + refee + greenrate + comm2012,
    data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-15579.2  -1166.8   -173.9   1041.3  23199.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.483e+03  9.613e+02   2.582  0.00997 **
retype2      1.496e+03  9.120e+02   1.641  0.10124
retype3      1.898e+03  8.867e+02   2.141  0.03257 *
retype4      2.604e+03  1.582e+03   1.646  0.10021
retype5      3.451e+02  9.432e+02   0.366  0.71451
refee       -1.487e+02  6.223e+01  -2.390  0.01706 *
greenrate   -2.831e+03  1.094e+03  -2.587  0.00983 **
comm2012     1.060e+00  1.282e-02  82.728  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2631 on 883 degrees of freedom
Multiple R-squared:  0.913,     Adjusted R-squared:  0.9123
F-statistic:  1324 on 7 and 883 DF,  p-value: < 2.2e-16
```

6.2.2 解释说明

上述代码中，fit1 建立了多元线性回归模型。模型拿 2013 年的小区房价作为因变量，拿 2012 年的小区房价和楼盘建筑特征、区位特征等作为自变量，这样做的理由有二：1，它实际上是利用了时间序列里面自回归的思想，也就是说 2012 年的房价里面实际上已经包含了很多历史因素，利用 2012 年的房价，等于说把这些历史因素考虑在内了，这样分析的结果更加全面；2，上一年的房价数据，我们总是可以观测到的，所以这并不影响模型在实际应用中的适用性。

通过建模后发现，很多自变量的 P 值都超过了 0.05 的临界水平，因此需要做逐步回归。逐步回归 fit2 剔除了对房价影响小的变量，这些“无关”的变量有：cnty（区县），year（建造年份），totarea（建造总面积），tothh（总户数），dvl（是否品牌开发商），reco.（是否金牌开发商），parking（停车位是否充足），subway（是否为地铁房），school（是否为学区房）。模型最后仅仅保留了 retype（物业类型）、refee（物业费）、greenrate（绿化率）和 comm2012（2012 年小区房价）四个变量。

fit2 最终结果表明，调整后的可决系数 R 方达到 0.9123，说明模型拟合的效果很好。fit2 的模型方程可以表示如下：

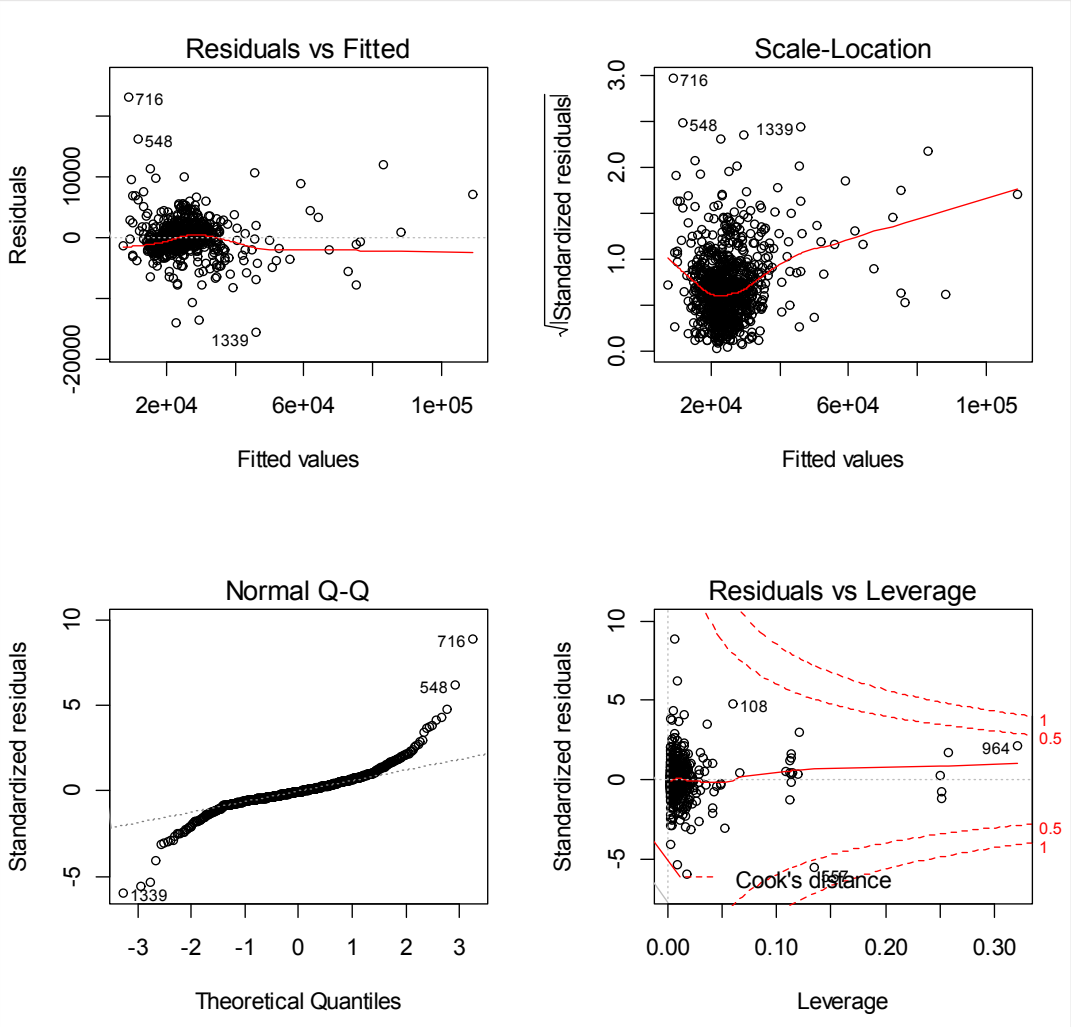
$$comm2013 = 2483 + 0 \cdot retype1 + 1496 \cdot retype2 + 1898 \cdot retype3 + 2604 \cdot retype4 + 345 \cdot retype5 - 149 \cdot refee - 2831 \cdot greenrate + 1.06 \cdot comm2012$$

其中，在物业类型 retype 中，'别墅'=5，'新里洋房'=4，'公寓'=3，'老公房'=2，'其他'=1。

6.2.3 回归诊断

6.2.3.1 残差分析

```
layout(matrix(1:4,2,2))
plot(fit2)
```



可以看到，除了第 548、718 和 1339 这三个点以外，残差和拟合值的散点图大致分布在 0 线附近，标准化残差的开方和拟合值的散点图总体在 0 到 1.5 之间，残差的正太 QQ 图上绝大部分点都在一条直线上。这些都说明残差正态性检验是通过的。

6.2.3.2 强影响点分析

#截取部分 Cook-distance 和 Hat-value 较大的点

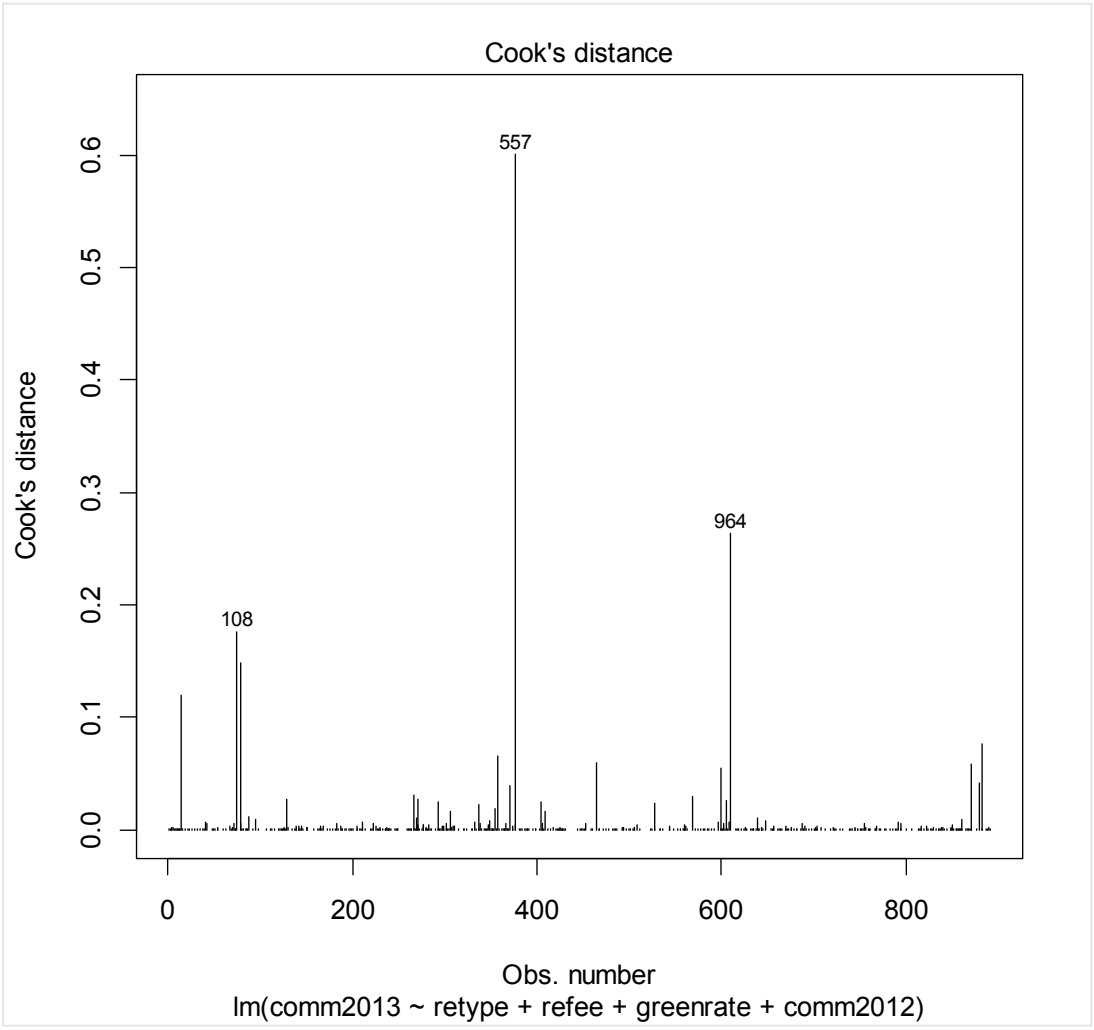
summary(influence.measures(fit2))

```
Potentially influential observations of
        lm(formula = comm2013 ~ retype + refee + greenrate + comm2012,      data = data1) :

    dfb.1_  dfb.rty2 dfb.rty3 dfb.rty4 dfb.rty5 dfb.refe dfb.grnr dfb.c201 dffit   cov.r   cook.d hat
4    0.02    0.00    -0.01     0.00     0.01     0.02    -0.02    -0.05    -0.11    0.97_*  0.00   0.00
21  -0.03    0.02     0.01     0.80    -0.01     0.14    -0.03     0.06     0.98_*  1.33_*  0.12   0.26_*
93   0.02   -0.01    -0.02     0.00     0.01    -0.05     0.04    -0.10    -0.16    0.96_*  0.00   0.00
100  0.01    0.00     0.01     0.00     0.01     0.02     0.07    -0.17     0.20    0.95_*  0.01   0.01
108 -0.17   -0.02    -0.01    -0.02     0.03    -0.31    -0.18     1.07_*   1.20_*  0.87_*  0.18   0.06_*
110  0.00    0.00     0.00     0.00     0.00     0.01     0.01    -0.02    -0.02    1.03_*  0.00   0.02
119 -0.15    0.02     0.02    -0.01    -0.01     0.03    -0.30     0.97     1.09_*  1.06_*  0.15   0.12_*
125  0.01    0.00     0.00     0.00     0.01    -0.01    -0.02     0.01     0.03    1.03_*  0.00   0.02
130 -0.01    0.00     0.01     0.00     0.04     0.00    -0.13     0.21     0.30_*  1.01    0.01   0.03_*
140  0.10   -0.01     0.00     0.01     0.05    -0.02    -0.22    -0.06    -0.27    0.96_*  0.01   0.01
196  0.04    0.00     0.00     0.00    -0.01    -0.02    -0.09    -0.01    -0.13    1.03_*  0.00   0.03_*
200  0.05   -0.05    -0.04     0.00     0.02    -0.45    -0.01     0.08    -0.47_*  1.12_*  0.03   0.11_*
201 -0.04    0.00     0.01     0.00    -0.02    -0.01     0.08     0.04     0.14    0.95_*  0.00   0.00
217  0.13   -0.14    -0.14    -0.08    -0.13    -0.02    -0.01     0.01     0.14    1.14_*  0.00   0.11_*
```

#画 Cook 距离

plot(fit2,4)



可以看到，除了第 108、557 和第 964 三个点以外，绝大部分点的 Cook 距离

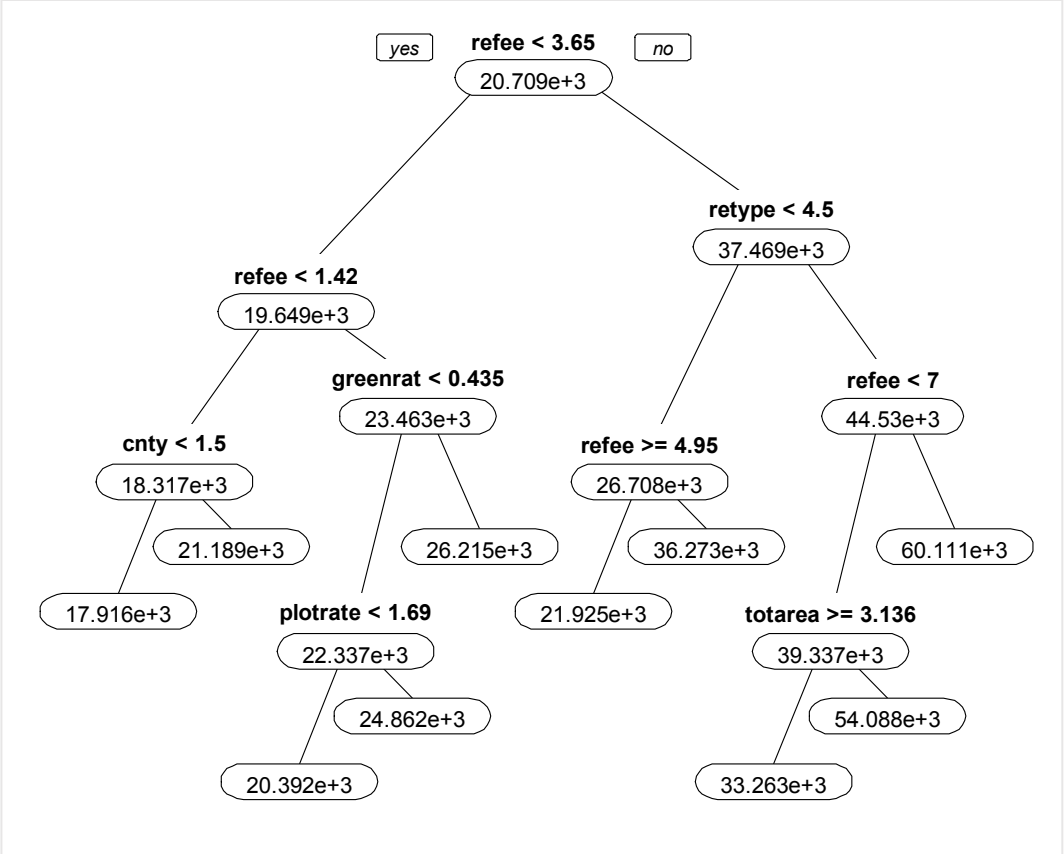
都在 0.1 以下，表明只有这三个点是强影响点。

6.3 决策树建模

6.3.1 R 语言代码及注释

```
setwd("C:\\Users\\GuQuan\\Desktop\\房价数据")
data=read.csv("data.csv",stringsAsFactors = FALSE)
library(rpart)
library(rpart.plot)
data1=na.omit(data)
data2=data1[,c(3,6:18)]
b=rpart(comm2012~.,data2)

#画决策树
rpart.plot(b,type=1,digits=5,faclen=T)
```




```
#计算决策树预测值
y2=predict(b,data2)
```

```
#计算均方误差
(NMSE=mean((data2$comm2012-y2)^2)/mean((data2$comm2012-mean(data2$com
m2012))^2))
[1] 0.4859329
```

6.3.2 解释说明

决策树模型拿 2012 年的小区房价作为因变量对其他自变量建模，可以发现，决策树用到的决策变量有物业费 refee、绿化率 greenrate、物业类型 retype、区县 cnty、容积率 plotrate 和总面积 totarea，表明这几个变量对房价的影响较高。代码中 y2 计算除了决策树回归模型预测出的房价数值，并用标准化均方误差（NMSE）进行了评价。NMSE 的计算公式如下：

$$NMSE=\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-y_{i\text{-}model})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i-y_{mean})^2}$$

该公式中 Yi-model 表示模型预测的数值，可以看出，若直接令 Yi-model 等于房价数据的均值 Ymean，则 NMSE=1，所以任何一个模型的 NMSE 值若大于 1，则该模型是没有任何意义的，因为还不如取房价均值作为预测值。一般来讲，NMSE 值低于 0.3 则说明该模型较好，且 NMSE 越小越好。上述代码表明，计算出的 NMSE 值为 0.48，说明决策树模型的准确度并不高。因此需要利用决策树的组合算法—随机森林来提高准确度。

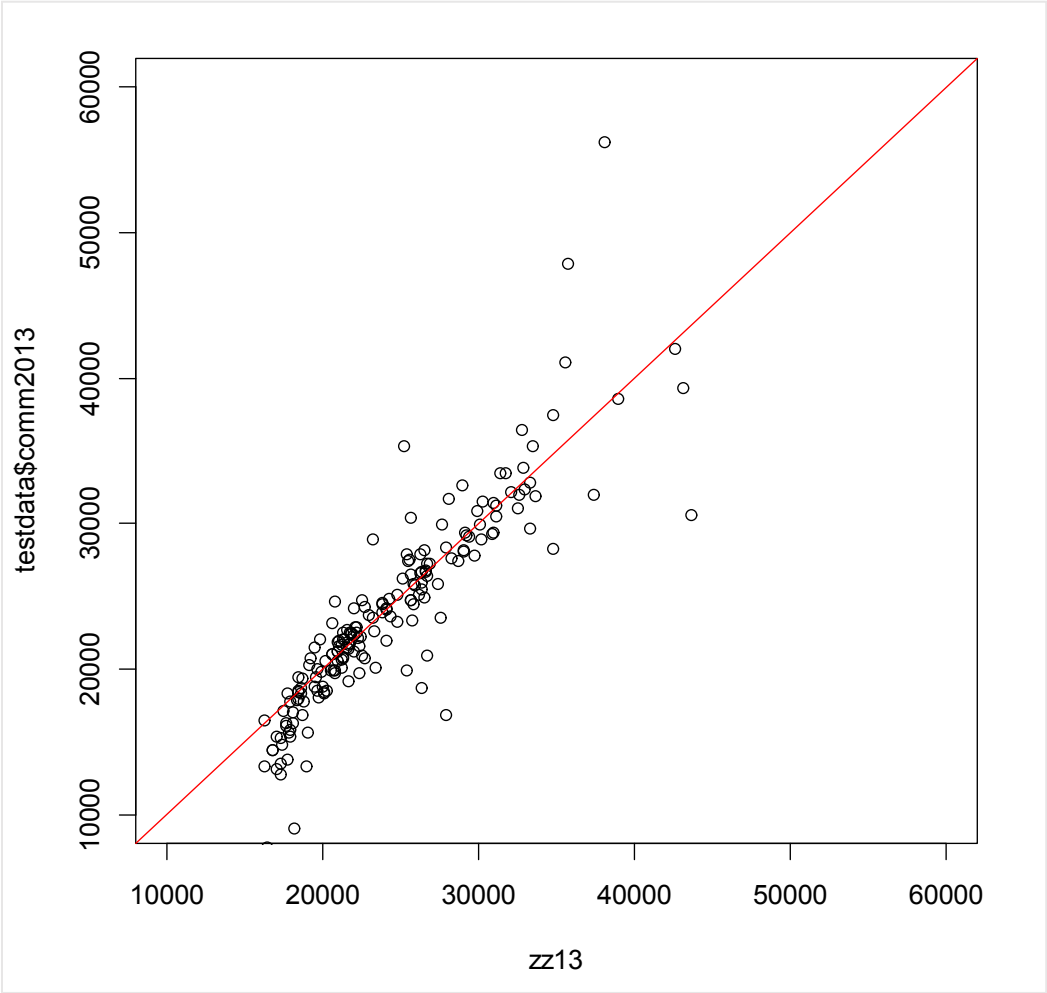
6.4 随机森林建模

6.4.1 R 语言代码及注释

```
library(randomForest)
#data13 表示用 2013 年的房价建模
data13=data1[,c(3,6:19)]

#对数据分成两部分，80%训练数据，20%测试数据
ind=sample(2,nrow(data13),replace=TRUE,prob=c(0.8,0.2))
traindata<- data13[ind==1,]
testdata<- data13[ind==2,]
```

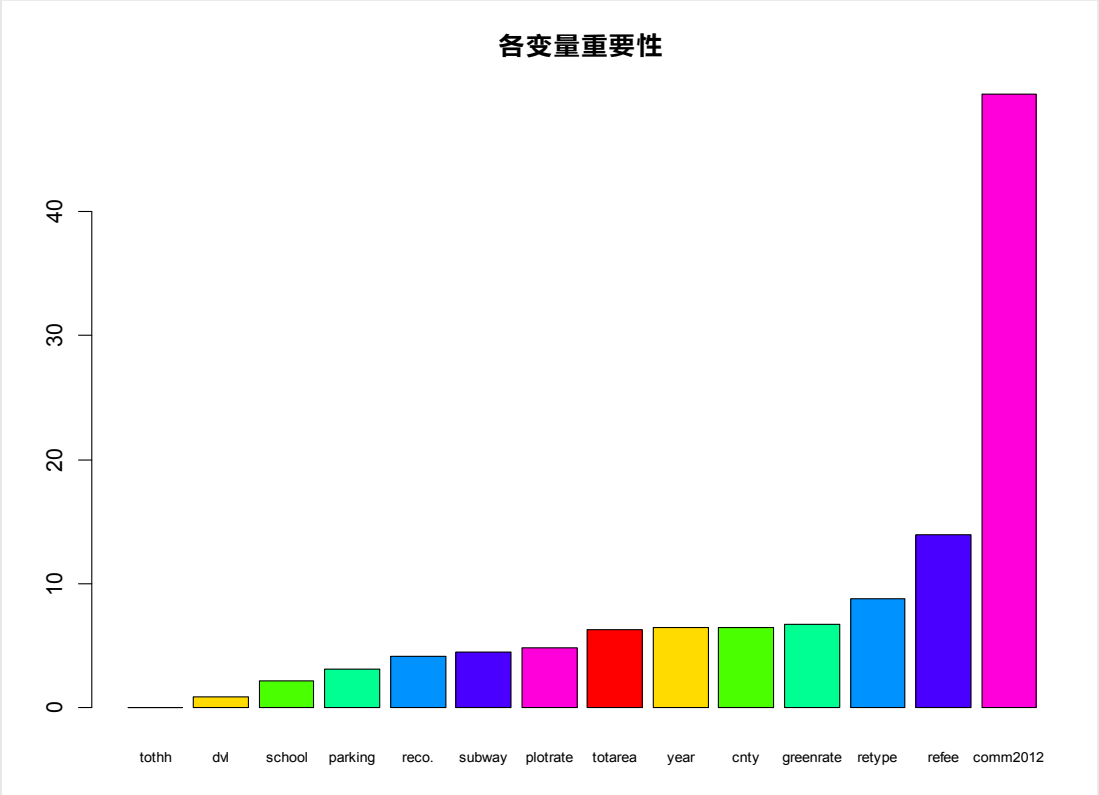
#随机森林建模
set.seed(1234)
rf=randomForest(comm2013~.,traindata,importance=TRUE,proximity=TRUE)
zz13=predict(rf,testdata)
(NMSE=mean((testdata$comm2013-zz13)^2)/mean((testdata$comm2013-mean(testdata$comm2013))^2))
#[1] 0.2018744
plot(zz13,testdata$comm2013,xlim=c(10000,60000),ylim=c(10000,60000))
abline(0,1,col="red")
#如果点大致分布在 45 度线附近，则认为准确率较高



#画变量重要性图
importance(rf)

	%IncMSE	IncNodePurity
cnty	6.5149927	2112959488
retype	8.7720651	2397935863
year	6.4356910	954154620
refee	13.9739195	7636655307
totarea	6.2723041	1789974745
tothh	0.0115786	2452756582
dvl	0.8904344	166685853
reco.	4.1361272	339748761
parking	3.1178894	370124099
subway	4.4732881	549366110
school	2.1693008	448321252
plotrate	4.8211563	2544158118
greenrate	6.7001986	2478809504
comm2012	49.4664527	29419758068

```
barplot(sort(t(importance(rf))[1,]),col=rainbow(7),main = '各变量重要性',
cex.names=0.7)
```



6.4.2 解释说明

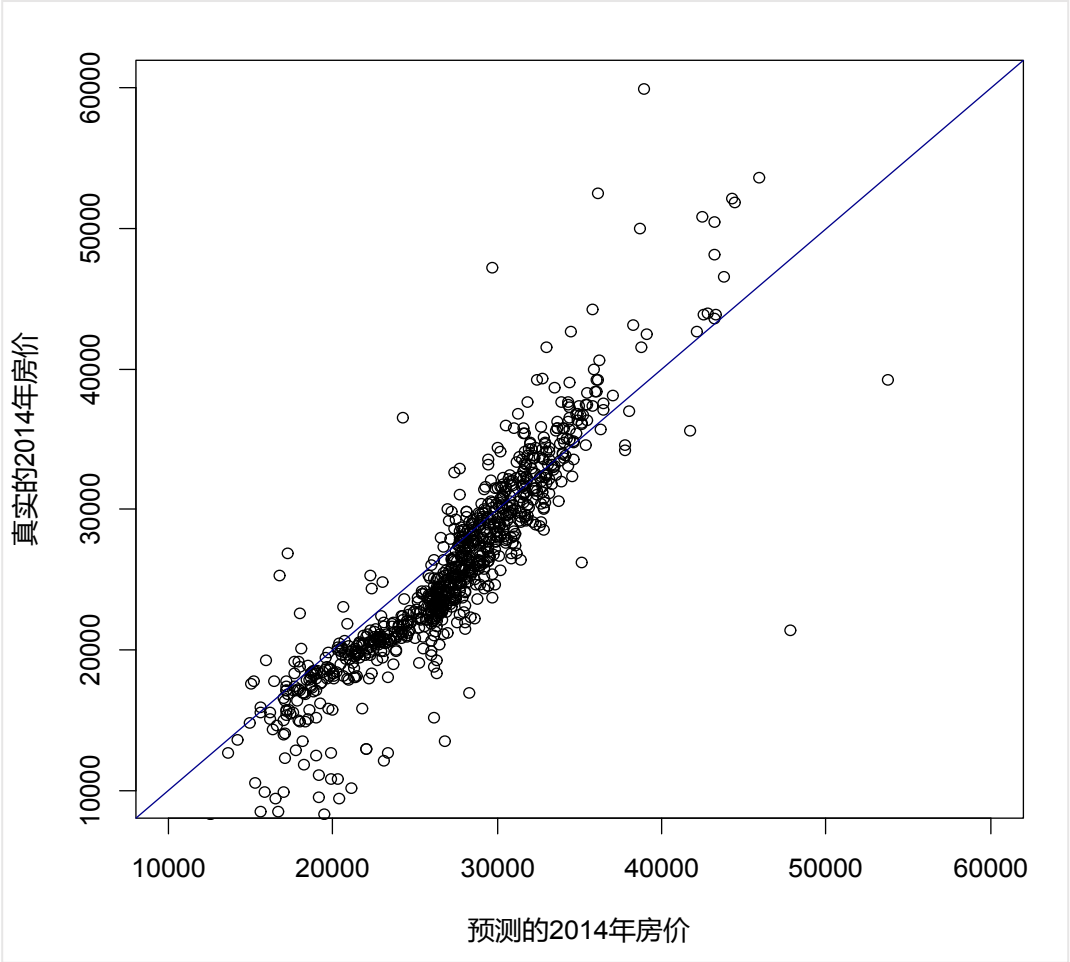
随机森林也是将 2013 年的小区房价作为因变量，2012 年的小区房价和其他变量作为自变量建模。通过将数据以 8:2 的比例划分训练集和测试集后发现，标准化均方误差 NMSE 为 0.20，比决策树的 0.485 低了很多，准确度大幅提升。另外，20%的测试数据真实值和模型预测值花在同一张图上后，大致分布在 45 度线附近，表明模型预测效果较好。

最后，代码画出了随机森林模型的变量重要性图，它表示随机森林在构建各个决策树时所使用的各个变量的比例，不出意外的是 2012 年的小区房价 comm2012 最重要，此外，重要性超过 5%的变量还有物业费 refee、物业类型 retype、绿化率 greenrate、区县 cnty、建造年份 year、建造总面积 totarea，这和多元线性回归的逐步回归结果以及单颗决策树的结果都是一致的。

6.4.3 随机森林预测效果

为了更进一步展示随机森林模型的预测效果，可以用 2013 年的数据和已经建好的随机森林模型去预测 2014 年的房价，并拿预测结果同真实的 2014 年房价进行比较。

```
#data14 表示用 2014 年的房价预测
data14=data1[,c(3,6:17,19,20)]
#更改第 14、15 列列名（实际上是 2013 和 2014 年真实房价），方便模型识别
names(data14)[14:15]=c("comm2012","comm2013")
zz14=predict(rf,data14)
(NMSE=mean((data14$comm2013-zz14)^2)/mean((data14$comm2013-mean(testda
ta$comm2013))^2))
#[1] 0.1671867
plot(zz14,data14$comm2013,xlim=c(10000,60000),ylim=c(10000,60000),xlab=" 预 测
的 2014 年房价",ylab="真实的 2014 年房价")
abline(0,1,col="darkblue")
```



上述代码表明，NMSE 只有 0.167，真实房价和预测房价绝大部分的点都分布在 45 度线附近，这更充分的说明了随机森林模型的优势。

七、结论

本文共使用了三种模型建模：多元线性回归，决策树和随机森林。

首先谈多元线性回归，它是房价预测中最常用的模型。在本文的建模当中，加入上一年的房价数据作为自变量后，多元回归模型的准确度大幅提升——可决系数有 0.9 以上，回归诊断的结果也都是通过的。如果拿 2013 年的房价作为因变量建模，会发现除了上一年房价外，对房价影响较大的因素有 retype（物业类型）、refee（物业费）、greenrate（绿化率）；如果拿 2014 年的房价建模（上文虽未提及，但实际上我们有尝试过），会发现影响较大的因素有 cnty（楼盘所在区县）、refee（物业费）和 school（是否为学区房）。

然后是决策树和随机森林。决策树的优点是算法简单，容易理解，缺点是准确度并不高，但它也能告诉我们哪些变量对影响房价可能比较重要。而随机森林比起决策树，不仅准确度高出一大截，也更充分的指出了对房价影响的重要变量。这些变量有物业费 refee、物业类型 retype、绿化率 greenrate、区县 cnty、建造年份 year、建造总面积 totarea 等，它们和逐步回归的结果是一致的。并且，随机森林模型的预测效果不论是在 20%的测试集上还是 2014 年的房价预测上都得到了很好的体现。

此外，尽管上文中并没有写出来，我们还尝试了支持向量机和 boosting 两种机器学习算法进行建模。但是它们准确度都不及随机森林，其中支持向量机的 NMSE 高达 0.44（如果是用 8:2 建立测试集，NMSE 更是高达 0.78），boosting 的 NMSE 也高达 0.39，两者都超过了 0.3，说明它们可能并不适合用于房价预测的项目。