HousePriceAnalysis

2023-12-14

```
###### Load dat0.csv
### 8430 samples. Data variables: unit price, area, floors, number of halls,
### number of rooms, whether it is a school district, orientation, year of
### construction, whether it is close to a subway station, urban area, city

house_price <- read.csv('D:/Desktop/C&S project/dat0.csv')
```

```
### Creat a new column called "age" to show the age of the house
house_price$age <- 2018 - house_price$year

attach(house_price)
```

```
####
#### **boxplot of house prices of different disincts
library(ggplot2)
library(dplyr)
```

```
##
## 载入程辑包：'dplyr'
```

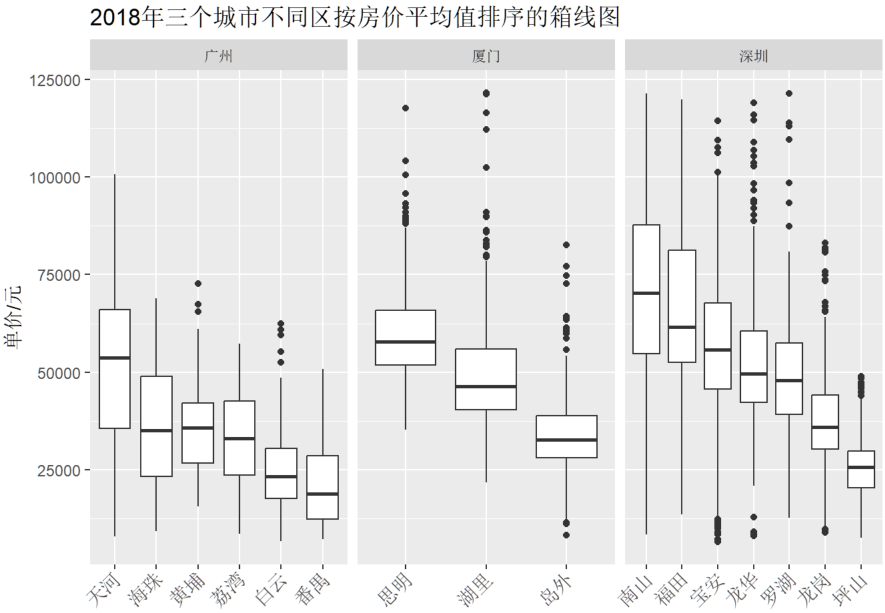
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Calculate average price of each distincts

average_prices <- house_price %>%
  group_by(city, district) %>%
  summarise(avg_price = mean(danjia))
```

```
## `summarise()` has grouped output by 'city'. You can override using the
## `.groups` argument.
```

```
# sort by average price

sorted_districts <- average_prices %>%
  arrange(city, desc(avg_price)) %>%
  pull(district)

ggplot(house_price, aes(x = factor(district, levels = sorted_districts), y = danjia)) +
  geom_boxplot() +
  facet_wrap(~city, scales = "free_x") +
  labs(title = "2018年三个城市不同区按房价平均值排序的箱线图",
       x = "",
       y = "单价/元") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
####
#### **Apply lm to 'danjia'

# Transform the discrete variables to factor
house_price$district <- as.factor(house_price$district)
house_price$floor <- as.factor(house_price$floor)
house_price$chaoxiang <- as.factor(house_price$chaoxiang)

model_lm <- lm(danjia ~ area + floor + hall + room + school + chaoxiang + subway + age,
               data = house_price)

summary(model_lm)
```

```
##
## Call:
## lm(formula = danjia ~ area + floor + hall + room + school + chaoxiang +
##     subway + age, data = house_price)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -56094 -12387  -2351  10302  76269
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    26299.22    1116.39  23.557  < 2e-16 ***
## area             127.69       6.17  20.696  < 2e-16 ***
## floor高层         47.01     537.67   0.087    0.930
## floor中层      -2069.07     502.38  -4.119 3.85e-05 ***
## hall              51.04     332.20   0.154    0.878
## room            -264.40     513.18  -0.515    0.606
## school          6612.52     404.32  16.354  < 2e-16 ***
## chaoxiang南向   4304.18     497.56   8.650  < 2e-16 ***
## chaoxiang其他   8507.28     493.24  17.248  < 2e-16 ***
## subway        -10239.18     586.95 -17.445  < 2e-16 ***
## age              193.18      30.46   6.342 2.39e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18240 on 8418 degrees of freedom
## Multiple R-squared:  0.1889, Adjusted R-squared:  0.1879
## F-statistic:   196 on 10 and 8418 DF,  p-value: < 2.2e-16
```

The model is
$\text{price} = 26299.21 + 127.69 \times \text{area} + 47.01 \times \text{floor高层} - 2069.07 \times \text{floor中层} + 51.04 \times \text{hall} - 264.40 \times \text{room} + 6612.52 \times \text{school} + 4304.18$

```
### Find the best model of each citys or districts
library(caret)
```

```
## 载入需要的程辑包：lattice
```

```
library(MASS)
```

```
##
## 载入程辑包：'MASS'
## The following object is masked from 'package:dplyr':
## 
##     select
```

```
fit_best_model <- function(region) {
  data = house_price
  if (region %in% c("广州", "厦门", "深圳")) {
    subset_data <- data[data$city == region, ]
    formula <- danjia ~ area + floor + hall + room + school + chaoxiang + subway + age + district
    # Stepwise regression
    initial_model <- lm(formula, data = subset_data)
    output <- capture.output(models <- stepAIC(initial_model, direction = "both"))
    final_model <- coef(models)
    ## formula <- danjia ~ area + floor + hall + room + school + chaoxiang + subway + age + district
      # k-fold cross-validation
    ## control <- trainControl(method = "cv", number = 5)
    ## models <- train(formula, data = subset_data, method = "lm", trControl = control, metric = "MSE")
    ## best_model <- models$finalModel

  } else {
    subset_data <- data[data$district == region, ]
    formula <- danjia ~ area + floor + hall + room + school + chaoxiang + subway + age
    # Stepwise regression
    initial_model <- lm(formula, data = subset_data)
    output <- capture.output(models <- stepAIC(initial_model, direction = "both"))

    final_model <- coef(models)


      # k-fold cross-validation
    ## control <- trainControl(method = "cv", number = 5)
    ## models <- train(formula, data = subset_data, method = "lm", trControl = control, metric = "MSE")
    ## best_model <- models$finalModel
  }
  cat("\033[31m****\033[0m", region, "\n")

  return(final_model)
}
```

Apply the function fit_best_model()

```
cat("\033[31mModel of different cities:\033[0m", "\n")
```

```
## □[31mModel of different cities:□[0m
```

```
fit_best_model("深圳")
```

```
## □[31m****□[0m 深圳
```

```
##   (Intercept)          area     floor高层     floor中层          hall 
##    43036.7528      102.8815    -2472.9472    -3147.2756     2809.9072 
##          room        school chaoxiang南向 chaoxiang其他           age 
##    -2598.7252     5263.7932     3796.5733     5575.1160     -592.1804 
##  district福田  district龙岗  district龙华  district罗湖  district南山 
##    13995.1515   -20209.3555    -4830.3128     -739.0714    15779.4773 
##  district坪山 
##   -25259.0321
```

```
fit_best_model("广州")
```

```
## □[31m****□[0m 广州
```

```
##  (Intercept)         area    floor高层    floor中层         hall         room 
##  18702.92067     95.35846    360.97906  -1778.05343  -1887.17304   2571.66288 
##       school       subway          age district番禺 district海珠 district黄埔 
##   3684.87287   3941.10704   -602.53645  -2650.00560  13874.67774  13531.32082 
## district荔湾 district天河 
##   9639.34972  26599.73473
```

```
fit_best_model("厦门")
```

```
## □[31m****□[0m 厦门
##   (Intercept)          area     floor高层     floor中层          hall 
##   36405.26149      47.42591   -1619.91826   -1237.91637    -703.83919 
##          room        school chaoxiang南向 chaoxiang其他           age 
##   -1248.38494    3850.22393    2992.64822    3094.22090    -606.28611 
##  district湖里  district思明 
##   19144.91993   30674.44156
```

```
cat("\033[31mModel of different districts:\033[0m", "\n")
```

```
## [31mModel of different districts:[0m
```

```
district_levels <- levels(house_price$district)

for (district in district_levels) {
  print(fit_best_model(district))
}
```

```
## [31m****[0m 白云
## (Intercept)        area   floor高层   floor中层      school      subway 
## 24180.31965    59.71754 -1163.89340 -3202.93428  2404.69366  3959.72020 
##         age 
##  -685.43524 
## [31m****[0m 宝安
##   (Intercept)          area     floor高层     floor中层          hall 
##   52296.47424      86.87547   -2803.79890   -4345.05547    2310.53871 
##          room        school chaoxiang南向 chaoxiang其他           age 
##   -4344.07071    3781.57313    8444.15951    9929.45821   -1036.70695 
## [31m****[0m 岛外
##   (Intercept)          area          hall          room        school 
##   27324.73926      12.67935    1160.25247    1129.92334    4645.75186 
## chaoxiang南向 chaoxiang其他           age 
##    2369.36505    4029.07027    -379.59121 
## [31m****[0m 番禺
## (Intercept)        area        room      school         age 
## 13370.31420    69.41752  2080.21804  2019.38432  -409.15834 
## [31m****[0m 福田
##   (Intercept)          area     floor高层     floor中层        school 
##    67057.7093      145.4451    -3196.4765    -6636.3025     9955.0165 
## chaoxiang南向 chaoxiang其他           age 
##    -6113.5564    -7498.2175     -803.5382 
## [31m****[0m 海珠
## (Intercept)        area   floor高层   floor中层        hall      subway 
##  33210.7665     55.7610  -4205.9654  -5427.3026   2534.2567   7508.8429 
##         age 
##   -813.8805 
## [31m****[0m 湖里
##   (Intercept)          area     floor高层     floor中层          hall 
##   73719.56664      69.85124   -5311.56790   -3957.25430   -1552.22339 
##          room        school chaoxiang南向 chaoxiang其他           age 
##   -4596.48475    1501.38599    3621.09181   -2772.09551   -1232.64954 
## [31m****[0m 黄埔
## (Intercept)        area        hall        room      subway         age 
##  16776.6213    150.1089  -4377.0317   3160.3309   4113.2300    680.7904 
## [31m****[0m 荔湾
##   (Intercept)          area          hall        school chaoxiang南向 
##     8428.8914      204.2611    -3718.2021    -7622.1192    -4981.7450 
## chaoxiang其他        subway 
##     1659.8930    19900.5446 
## [31m****[0m 龙岗
##   (Intercept)          area     floor高层     floor中层          hall 
##   16477.93464      66.14573     828.71744   -1394.97205    3483.24483 
##          room        school chaoxiang南向 chaoxiang其他 
##   -3526.54525    6279.15121    8511.31139    9379.94822 
## [31m****[0m 龙华
##   (Intercept)          area     floor高层     floor中层          hall 
##   46028.91086      85.74019   -3391.80852   -2872.52797    2979.50438 
##          room        school chaoxiang南向 chaoxiang其他           age 
##   -3523.89347    3025.33569    5052.79155    5716.77988    -982.55685 
## [31m****[0m 罗湖
## (Intercept)        area   floor高层   floor中层 
##  39683.3161    153.0571  -5199.3163  -3993.9768 
## [31m****[0m 南山
##   (Intercept)          area          hall        school chaoxiang南向 
##    54427.8804      123.5862     2507.4823    10344.1626     5802.6882 
## chaoxiang其他           age 
##    11781.5731    -1460.2109 
## [31m****[0m 坪山
## (Intercept)   floor高层   floor中层      school         age 
##  24968.8157   3474.6179  -1007.5298  -3878.3759    635.5445 
## [31m****[0m 思明
##   (Intercept)          area          hall          room        school 
##   71149.92855      49.01005   -1778.56469   -3103.90665    4390.11864 
## chaoxiang南向 chaoxiang其他           age 
##    8085.39226    3669.47194    -559.32069 
## [31m****[0m 天河
## (Intercept)        area        hall        room      school      subway 
## 52453.85784    75.12428 -3449.46342  7155.18340  6345.08118  2404.78570 
##         age 
## -1361.83543 
```

price vs. age

```
plot_price_age <- function(region, data = house_price) {

  if (region %in% c("广州", "厦门", "深圳")) {
    subset_data <- data[data$city == region, ]
  } else {
    subset_data <- data[data$district == region, ]
  }

  mean_prices <- tapply(subset_data$danjia, subset_data$age, mean)

  mean_prices_df <- data.frame(age = as.numeric(names(mean_prices)), mean_danjia = as.numeric(mean_prices))

  ggplot(mean_prices_df, aes(x = age, y = mean_danjia)) +
    geom_point(size = 0.6, color = "red") +
    geom_smooth(method = "loess", se = FALSE, color = "blue") +
    labs(title = paste( region),
         x = "房龄", y = "平均房价",
         xlab = "共同的横坐标标签", ylab = "共同的纵坐标标签") +
    theme_minimal() +
    theme(axis.text = element_text(size = 5),
          axis.title = element_text(size = 5),
          plot.title = element_text(size = 10, face = "bold"))
}
```

```
library(patchwork)
```

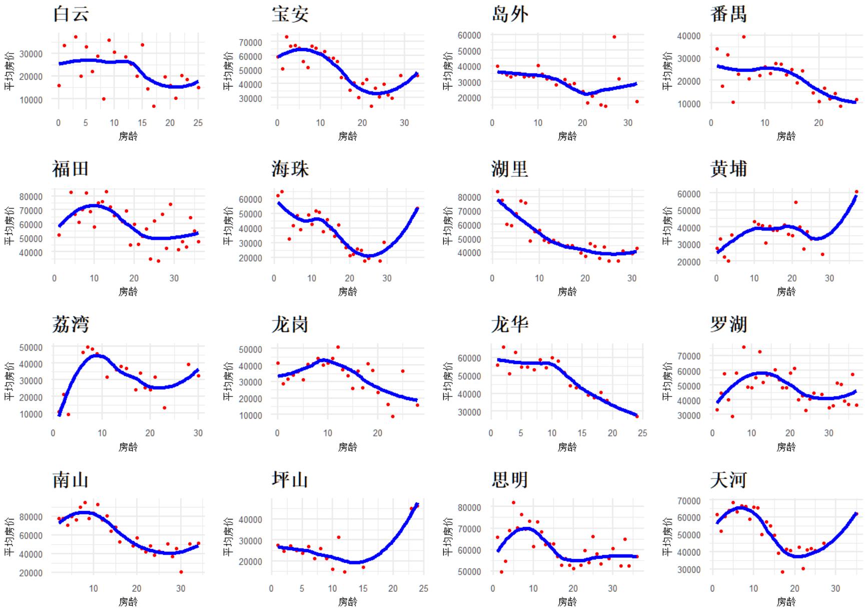
```
##
## 载入程辑包：'patchwork'
```

```
## The following object is masked from 'package:MASS':
##
##     area
```

```
district_levels <- levels(house_price$district)

# Create a list to store individual plots
plots_list <- list()

# Loop through districts and create plots
for (district in district_levels) {
  plot <- plot_price_age(district)
  plots_list[[district]] <- plot
}

# Combine plots using patchwork
big_plot <- wrap_plots(plots_list, ncol = 4)

# Display the combined plot
print(big_plot)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```


白云
宝安
岛外
番禺
福田
海珠
湖里
黄埔
荔湾
龙岗
龙华
罗湖
南山
坪山
思明
天河
平均房价
房龄