Realestate_Gangnam

2023-06-11

데이터 로드 및 변환

```
library(dplyr)
```

```
##
## 다음의 패키지를 부착합니다: 'dplyr'
```

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

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

```
# finalretaildata 불러오기

#setwd("C:₩₩Rtest₩₩realestate")
data_whole <- read.csv("RealEstateData/FinalRetailData_1차 수정.csv", h = T, fileEncoding = "cp
949")
data_whole <- subset(data_whole, Rejion == "강남구")  # 다른 구 분석하려면 이 부분 변경경
summary(data_whole)
```

```
##      index       transaction_id     apartment_id       city
##  Min.   :72792   Min.   :1175801   Min.   :  328   Length:5957
##  1st Qu.:74300   1st Qu.:1177377   1st Qu.: 2804   Class :character
##  Median :75828   Median :1178931   Median : 5633   Mode  :character
##  Mean   :75841   Mean   :1178999   Mean   : 5811
##  3rd Qu.:77377   3rd Qu.:1180622   3rd Qu.: 9094
##  Max.   :78922   Max.   :1182355   Max.   :12619
##     dong              jibun              apt              addr_kr
##  Length:5957        Length:5957        Length:5957        Length:5957
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  exclusive_use_area year_of_completion transaction_year_month
##  Min.   : 12.10     Min.   :1974       Length:5957
##  1st Qu.: 59.82     1st Qu.:1983       Class :character
##  Median : 84.83     Median :1996       Mode  :character
##  Mean   : 89.32     Mean   :1995
##  3rd Qu.:114.55     3rd Qu.:2006
##  Max.   :273.45     Max.   :2017
##  transaction_date       floor       transaction_real_price      year
##  Length:5957        Min.   : 1.00   Min.   : 17750         Min.   :2017
##  Class :character   1st Qu.: 4.00   1st Qu.: 89000         1st Qu.:2017
##  Mode  :character   Median : 7.00   Median :117500         Median :2017
##                     Mean   : 8.78   Mean   :126665         Mean   :2017
##                     3rd Qu.:12.00   3rd Qu.:155000         3rd Qu.:2017
##                     Max.   :68.00   Max.   :525000         Max.   :2017
##    Latitude        Hardness        Rejion           bigMarket05
##  Min.   :127.0   Min.   :35.99   Length:5957        Min.   :0.000
##  1st Qu.:127.0   1st Qu.:37.49   Class :character   1st Qu.:1.000
##  Median :127.1   Median :37.49   Mode  :character   Median :1.000
##  Mean   :127.1   Mean   :37.49                      Mean   :1.788
##  3rd Qu.:127.1   3rd Qu.:37.50                      3rd Qu.:2.000
##  Max.   :127.1   Max.   :37.53                      Max.   :5.000
##   bigMarket10      bigMarket15       school05        school10
##  Min.   : 0.000   Min.   : 0.00   Min.   :0.000   Min.   : 0.000
##  1st Qu.: 3.000   1st Qu.: 7.00   1st Qu.:1.000   1st Qu.: 6.000
##  Median : 5.000   Median :11.00   Median :3.000   Median : 9.000
##  Mean   : 5.028   Mean   :10.26   Mean   :2.748   Mean   : 9.521
##  3rd Qu.: 7.000   3rd Qu.:14.00   3rd Qu.:4.000   3rd Qu.:13.000
##  Max.   :11.000   Max.   :18.00   Max.   :8.000   Max.   :20.000
##    school15        subway05        subway10        subway15
##  Min.   : 1.0   Min.   :0.000   Min.   :0.000   Min.   : 0.000
##  1st Qu.:13.0   1st Qu.:0.000   1st Qu.:2.000   1st Qu.: 5.000
##  Median :18.0   Median :1.000   Median :4.000   Median : 8.000
##  Mean   :19.1   Mean   :1.121   Mean   :3.409   Mean   : 7.234
##  3rd Qu.:26.0   3rd Qu.:2.000   3rd Qu.:5.000   3rd Qu.:10.000
##  Max.   :33.0   Max.   :4.000   Max.   :9.000   Max.   :17.000
##   hospital05       hospital10      hospital15        movie05
##  Min.   :  0.00   Min.   :  0.0   Min.   :   0.0   Min.   : 0.00
##  1st Qu.: 18.00   1st Qu.: 70.0   1st Qu.: 190.0   1st Qu.: 1.00
##  Median : 40.00   Median :179.0   Median : 410.0   Median : 2.00
##  Mean   : 53.92   Mean   :203.8   Mean   : 435.9   Mean   : 4.07
##  3rd Qu.: 79.00   3rd Qu.:301.0   3rd Qu.: 618.0   3rd Qu.: 4.00
```

```
##  Max.   :390.00   Max.   :840.0   Max.   :1488.0   Max.   :42.00
##     movie10          movie15          kid05            kid10
##  Min.   :  0.00   Min.   :  0.00   Min.   : 0.000   Min.   : 0.00
##  1st Qu.:  5.00   1st Qu.: 12.00   1st Qu.: 3.000   1st Qu.:19.00
##  Median : 10.00   Median : 30.00   Median : 7.000   Median :26.00
##  Mean   : 20.49   Mean   : 48.05   Mean   : 6.896   Mean   :23.45
##  3rd Qu.: 27.00   3rd Qu.: 76.00   3rd Qu.:10.000   3rd Qu.:29.00
##  Max.   :107.00   Max.   :172.00   Max.   :17.000   Max.   :43.00
##      kid15          office05        office10         office15
##  Min.   : 0.00   Min.   :0.000   Min.   : 0.000   Min.   : 0.00
##  1st Qu.:43.00   1st Qu.:1.000   1st Qu.: 4.000   1st Qu.: 9.00
##  Median :53.00   Median :1.000   Median : 5.000   Median :13.00
##  Mean   :51.32   Mean   :1.499   Mean   : 5.328   Mean   :12.05
##  3rd Qu.:62.00   3rd Qu.:2.000   3rd Qu.: 7.000   3rd Qu.:14.00
##  Max.   :78.00   Max.   :6.000   Max.   :13.000   Max.   :23.00
```

```
str(data_whole)
```

```
## 'data.frame':    5957 obs. of  39 variables:
##  $ index                : int  72792 72793 72794 72795 72796 72797 72798 72799 72800 72801
...
##  $ transaction_id       : int  1175801 1175803 1175804 1175805 1175806 1175807 1175808 1175
809 1175810 1175811 ...
##  $ apartment_id         : int  2646 8161 502 11741 8150 1701 3787 2417 7725 2646 ...
##  $ city                 : chr  "서울특별시" "서울특별시" "서울특별시" "서울특별시" ...
##  $ dong                 : chr  "역삼동" "역삼동" "역삼동" "역삼동" ...
##  $ jibun                : chr  "720-25" "711-3" "755-4" "706-20" ...
##  $ apt                  : chr  "대우디오빌" "역삼자이" "e-편한세상" "한화진넥스빌" ...
##  $ addr_kr              : chr  "역삼동 720-25 대우디오빌" "역삼동 711-3 역삼자이" "역삼동 7
55-4 e-편한세상" "역삼동 706-20 한화진넥스빌" ...
##  $ exclusive_use_area   : num  30 114 59.6 39.2 59.4 ...
##  $ year_of_completion   : int  2002 2016 2005 2001 2005 2003 2011 1997 2007 2002 ...
##  $ transaction_year_month: chr  "2017-01-01" "2017-01-01" "2017-01-01" "2017-01-01" ...
##  $ transaction_date     : chr  "1~10" "1~10" "1~10" "1~10" ...
##  $ floor                : int  11 7 6 23 6 2 22 17 2 21 ...
##  $ transaction_real_price: int  26400 150000 89500 26500 88000 80000 131000 79500 64000 5780
0 ...
##  $ year                 : int  2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 ...
##  $ Latitude             : num  127 127 127 127 127 ...
##  $ Hardness             : num  37.5 37.5 37.5 37.5 37.5 ...
##  $ Rejion               : chr  "강남구" "강남구" "강남구" "강남구" ...
##  $ bigMarket05          : int  0 1 4 0 4 2 3 1 0 0 ...
##  $ bigMarket10          : int  3 3 6 2 7 6 6 6 5 3 ...
##  $ bigMarket15          : int  10 10 12 9 12 11 15 12 12 10 ...
##  $ school05             : int  1 3 5 2 3 3 2 2 1 1 ...
##  $ school10             : int  4 6 13 4 13 7 13 12 9 4 ...
##  $ school15             : int  16 16 20 16 23 16 23 22 20 16 ...
##  $ subway05             : int  0 0 1 0 1 0 0 0 0 0 ...
##  $ subway10             : int  4 5 5 5 5 5 4 4 2 4 ...
##  $ subway15             : int  9 10 10 10 9 6 9 7 11 9 ...
##  $ hospital05           : int  47 58 105 50 100 53 64 48 56 47 ...
##  $ hospital10           : int  345 343 312 347 327 480 254 283 233 345 ...
##  $ hospital15           : int  875 704 612 724 575 829 552 580 864 875 ...
##  $ movie05              : int  7 7 4 6 4 5 3 4 5 7 ...
##  $ movie10              : int  24 23 16 25 13 27 15 17 19 24 ...
##  $ movie15              : int  59 54 44 55 38 65 40 39 50 59 ...
##  $ kid05                : int  5 8 13 6 14 5 10 9 6 5 ...
##  $ kid10                : int  25 26 36 25 33 27 31 30 28 25 ...
##  $ kid15                : int  62 63 61 61 61 63 60 63 64 62 ...
##  $ office05             : int  3 3 1 2 1 3 2 2 2 3 ...
##  $ office10             : int  7 7 6 7 6 13 5 6 9 7 ...
##  $ office15             : int  15 13 15 12 14 20 16 16 16 15 ...
```

```
data_whole %>% colnames()
```

```
##  [1] "index"                "transaction_id"         "apartment_id"
##  [4] "city"                 "dong"                   "jibun"
##  [7] "apt"                  "addr_kr"                "exclusive_use_area"
## [10] "year_of_completion"   "transaction_year_month" "transaction_date"
## [13] "floor"                "transaction_real_price" "year"
## [16] "Latitude"             "Hardness"               "Rejion"
## [19] "bigMarket05"          "bigMarket10"            "bigMarket15"
## [22] "school05"             "school10"               "school15"
## [25] "subway05"             "subway10"               "subway15"
## [28] "hospital05"           "hospital10"             "hospital15"
## [31] "movie05"              "movie10"                "movie15"
## [34] "kid05"                "kid10"                  "kid15"
## [37] "office05"             "office10"               "office15"
```

```
# # Was:
# data %>% select(filterCol)
#
# # Now:
# data %>% select(all_of(filterCol))
filterCol<-c("index", "transaction_id", "apartment_id", "city", "jibun", "apt", "addr_kr", "Lat
itude", "Hardness", "year", "Rejion")
data_whole<-data_whole %>% select(-all_of(filterCol))
str(data_whole)
```

```
## 'data.frame':    5957 obs. of  28 variables:
##  $ dong                  : chr  "역삼동" "역삼동" "역삼동" "역삼동" ...
##  $ exclusive_use_area    : num  30 114 59.6 39.2 59.4 ...
##  $ year_of_completion    : int  2002 2016 2005 2001 2005 2003 2011 1997 2007 2002 ...
##  $ transaction_year_month: chr  "2017-01-01" "2017-01-01" "2017-01-01" "2017-01-01" ...
##  $ transaction_date      : chr  "1~10" "1~10" "1~10" "1~10" ...
##  $ floor                 : int  11 7 6 23 6 2 22 17 2 21 ...
##  $ transaction_real_price: int  26400 150000 89500 26500 88000 80000 131000 79500 64000 5780
0 ...
##  $ bigMarket05           : int  0 1 4 0 4 2 3 1 0 0 ...
##  $ bigMarket10           : int  3 3 6 2 7 6 6 6 5 3 ...
##  $ bigMarket15           : int  10 10 12 9 12 11 15 12 12 10 ...
##  $ school05              : int  1 3 5 2 3 3 2 2 1 1 ...
##  $ school10              : int  4 6 13 4 13 7 13 12 9 4 ...
##  $ school15              : int  16 16 20 16 23 16 23 22 20 16 ...
##  $ subway05              : int  0 0 1 0 1 0 0 0 0 0 ...
##  $ subway10              : int  4 5 5 5 5 5 4 4 2 4 ...
##  $ subway15              : int  9 10 10 10 9 6 9 7 11 9 ...
##  $ hospital05            : int  47 58 105 50 100 53 64 48 56 47 ...
##  $ hospital10            : int  345 343 312 347 327 480 254 283 233 345 ...
##  $ hospital15            : int  875 704 612 724 575 829 552 580 864 875 ...
##  $ movie05               : int  7 7 4 6 4 5 3 4 5 7 ...
##  $ movie10               : int  24 23 16 25 13 27 15 17 19 24 ...
##  $ movie15               : int  59 54 44 55 38 65 40 39 50 59 ...
##  $ kid05                 : int  5 8 13 6 14 5 10 9 6 5 ...
##  $ kid10                 : int  25 26 36 25 33 27 31 30 28 25 ...
##  $ kid15                 : int  62 63 61 61 61 63 60 63 64 62 ...
##  $ office05              : int  3 3 1 2 1 3 2 2 2 3 ...
##  $ office10              : int  7 7 6 7 6 13 5 6 9 7 ...
##  $ office15              : int  15 13 15 12 14 20 16 16 16 15 ...
```

```
# 면적당 가격 변수 추가 및 real_price 변수 제거
data_whole$transaction_real_price <- as.numeric(data_whole$transaction_real_price)
data_whole$unit_price <- data_whole$transaction_real_price / data_whole$exclusive_use_area
data_whole$transaction_real_price <- NULL
str(data_whole)
```

```
## 'data.frame':    5957 obs. of  28 variables:
##  $ dong                  : chr  "역삼동" "역삼동" "역삼동" "역삼동" ...
##  $ exclusive_use_area    : num  30 114 59.6 39.2 59.4 ...
##  $ year_of_completion    : int  2002 2016 2005 2001 2005 2003 2011 1997 2007 2002 ...
##  $ transaction_year_month: chr  "2017-01-01" "2017-01-01" "2017-01-01" "2017-01-01" ...
##  $ transaction_date      : chr  "1~10" "1~10" "1~10" "1~10" ...
##  $ floor                 : int  11 7 6 23 6 2 22 17 2 21 ...
##  $ bigMarket05           : int  0 1 4 0 4 2 3 1 0 0 ...
##  $ bigMarket10           : int  3 3 6 2 7 6 6 6 5 3 ...
##  $ bigMarket15           : int  10 10 12 9 12 11 15 12 12 10 ...
##  $ school05              : int  1 3 5 2 3 3 2 2 1 1 ...
##  $ school10              : int  4 6 13 4 13 7 13 12 9 4 ...
##  $ school15              : int  16 16 20 16 23 16 23 22 20 16 ...
##  $ subway05              : int  0 0 1 0 1 0 0 0 0 0 ...
##  $ subway10              : int  4 5 5 5 5 5 4 4 2 4 ...
##  $ subway15              : int  9 10 10 10 9 6 9 7 11 9 ...
##  $ hospital05            : int  47 58 105 50 100 53 64 48 56 47 ...
##  $ hospital10            : int  345 343 312 347 327 480 254 283 233 345 ...
##  $ hospital15            : int  875 704 612 724 575 829 552 580 864 875 ...
##  $ movie05               : int  7 7 4 6 4 5 3 4 5 7 ...
##  $ movie10               : int  24 23 16 25 13 27 15 17 19 24 ...
##  $ movie15               : int  59 54 44 55 38 65 40 39 50 59 ...
##  $ kid05                 : int  5 8 13 6 14 5 10 9 6 5 ...
##  $ kid10                 : int  25 26 36 25 33 27 31 30 28 25 ...
##  $ kid15                 : int  62 63 61 61 61 63 60 63 64 62 ...
##  $ office05              : int  3 3 1 2 1 3 2 2 2 3 ...
##  $ office10              : int  7 7 6 7 6 13 5 6 9 7 ...
##  $ office15              : int  15 13 15 12 14 20 16 16 16 15 ...
##  $ unit_price            : num  879 1316 1502 676 1481 ...
```

```
# transaction_month 변수 추가 및 transaction_year_month, transaction_date, apt 변수 제거
data_whole$transaction_month <- substr(data_whole$transaction_year_month, 6, 7)
data_whole$transaction_year_month <- NULL
data_whole$transaction_date <- NULL
data_whole$apt <- NULL
str(data_whole)
```

```
## 'data.frame':    5957 obs. of  27 variables:
##  $ dong              : chr  "역삼동" "역삼동" "역삼동" "역삼동" ...
##  $ exclusive_use_area: num  30 114 59.6 39.2 59.4 ...
##  $ year_of_completion: int  2002 2016 2005 2001 2005 2003 2011 1997 2007 2002 ...
##  $ floor             : int  11 7 6 23 6 2 22 17 2 21 ...
##  $ bigMarket05       : int  0 1 4 0 4 2 3 1 0 0 ...
##  $ bigMarket10       : int  3 3 6 2 7 6 6 6 5 3 ...
##  $ bigMarket15       : int  10 10 12 9 12 11 15 12 12 10 ...
##  $ school05          : int  1 3 5 2 3 3 2 2 1 1 ...
##  $ school10          : int  4 6 13 4 13 7 13 12 9 4 ...
##  $ school15          : int  16 16 20 16 23 16 23 22 20 16 ...
##  $ subway05          : int  0 0 1 0 1 0 0 0 0 0 ...
##  $ subway10          : int  4 5 5 5 5 5 4 4 2 4 ...
##  $ subway15          : int  9 10 10 10 9 6 9 7 11 9 ...
##  $ hospital05        : int  47 58 105 50 100 53 64 48 56 47 ...
##  $ hospital10        : int  345 343 312 347 327 480 254 283 233 345 ...
##  $ hospital15        : int  875 704 612 724 575 829 552 580 864 875 ...
##  $ movie05           : int  7 7 4 6 4 5 3 4 5 7 ...
##  $ movie10           : int  24 23 16 25 13 27 15 17 19 24 ...
##  $ movie15           : int  59 54 44 55 38 65 40 39 50 59 ...
##  $ kid05             : int  5 8 13 6 14 5 10 9 6 5 ...
##  $ kid10             : int  25 26 36 25 33 27 31 30 28 25 ...
##  $ kid15             : int  62 63 61 61 61 63 60 63 64 62 ...
##  $ office05          : int  3 3 1 2 1 3 2 2 2 3 ...
##  $ office10          : int  7 7 6 7 6 13 5 6 9 7 ...
##  $ office15          : int  15 13 15 12 14 20 16 16 16 15 ...
##  $ unit_price        : num  879 1316 1502 676 1481 ...
##  $ transaction_month : chr  "01" "01" "01" "01" ...
```

```
# factor 형으로 변환
data_whole$year <- as.factor(data_whole$year)
data_whole$dong <- as.factor(data_whole$dong)
data_whole$transaction_month <- as.factor(data_whole$transaction_month) # 거래월에 따른 가격 변화 확인

# 변환 결과 확인
str(data_whole)
```

```
## 'data.frame':    5957 obs. of  28 variables:
##  $ dong              : Factor w/ 13 levels "개포동","논현동",..: 10 10 10 10 10 10 10 10 10
10 ...
##  $ exclusive_use_area: num  30 114 59.6 39.2 59.4 ...
##  $ year_of_completion: int  2002 2016 2005 2001 2005 2003 2011 1997 2007 2002 ...
##  $ floor             : int  11 7 6 23 6 2 22 17 2 21 ...
##  $ bigMarket05       : int  0 1 4 0 4 2 3 1 0 0 ...
##  $ bigMarket10       : int  3 3 6 2 7 6 6 6 5 3 ...
##  $ bigMarket15       : int  10 10 12 9 12 11 15 12 12 10 ...
##  $ school05          : int  1 3 5 2 3 3 2 2 1 1 ...
##  $ school10          : int  4 6 13 4 13 7 13 12 9 4 ...
##  $ school15          : int  16 16 20 16 23 16 23 22 20 16 ...
##  $ subway05          : int  0 0 1 0 1 0 0 0 0 0 ...
##  $ subway10          : int  4 5 5 5 5 5 4 4 2 4 ...
##  $ subway15          : int  9 10 10 10 9 6 9 7 11 9 ...
##  $ hospital05        : int  47 58 105 50 100 53 64 48 56 47 ...
##  $ hospital10        : int  345 343 312 347 327 480 254 283 233 345 ...
##  $ hospital15        : int  875 704 612 724 575 829 552 580 864 875 ...
##  $ movie05           : int  7 7 4 6 4 5 3 4 5 7 ...
##  $ movie10           : int  24 23 16 25 13 27 15 17 19 24 ...
##  $ movie15           : int  59 54 44 55 38 65 40 39 50 59 ...
##  $ kid05             : int  5 8 13 6 14 5 10 9 6 5 ...
##  $ kid10             : int  25 26 36 25 33 27 31 30 28 25 ...
##  $ kid15             : int  62 63 61 61 61 63 60 63 64 62 ...
##  $ office05          : int  3 3 1 2 1 3 2 2 2 3 ...
##  $ office10          : int  7 7 6 7 6 13 5 6 9 7 ...
##  $ office15          : int  15 13 15 12 14 20 16 16 16 15 ...
##  $ unit_price        : num  879 1316 1502 676 1481 ...
##  $ transaction_month : Factor w/ 11 levels "01","02","03",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ year              : Factor w/ 43 levels "1974","1976",..: 28 42 31 27 31 29 37 23 33 28
...
```

컬럼 값 Exploration 및 데이터 변환

```
library(ggplot2)

# year of completion -- 준공년도
summary(data_whole$year_of_completion)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1974    1983    1996    1995    2006    2017
```

```
data_whole$year_of_completion_f <- cut(data_whole$year_of_completion, breaks = c(0, 1997, 2001,
2007, Inf), labels = c("1st", "2nd", "3rd", "4th"))
data_whole$year_of_completion <- NULL
summary(data_whole$year_of_completion_f)
```

```
##  1st  2nd  3rd  4th
## 3146  471 1549  791
```

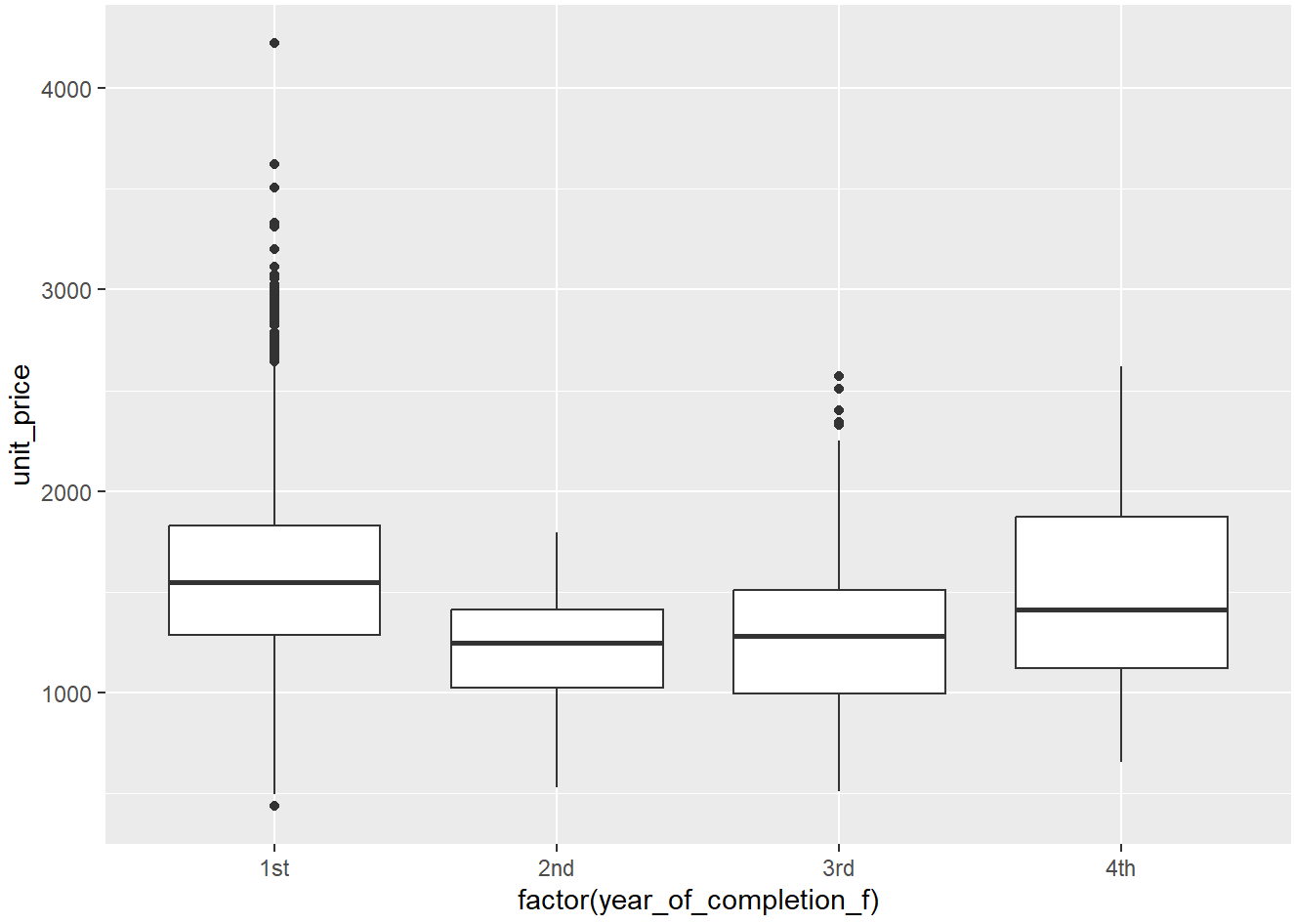
```
# 전체 가격 분포
data_whole %>% summarize(count = n(), avg_price = mean(unit_price), std_price = sd(unit_price))
```

```
##   count avg_price std_price
## 1  5957  1479.769  457.5211
```

```
# 준공년도 factor별 가격 분포
data_whole %>% group_by(year_of_completion_f) %>%
  summarize(count = n(), avg_price = mean(unit_price), std_price = sd(unit_price))
```

```
## # A tibble: 4 × 4
##   year_of_completion_f count avg_price std_price
##   <fct>                <int>     <dbl>     <dbl>
## 1 1st                   3146     1622.      479.
## 2 2nd                    471     1209.      273.
## 3 3rd                   1549     1272.      330.
## 4 4th                    791     1484.      448.
```

```
ggplot(data = data_whole, aes(x = factor(year_of_completion_f), y = unit_price)) + geom_boxplot
()
```




```
# 동별 가격 분포
summary(data_whole$dong)
```

```
##   개포동   논현동   대치동   도곡동   삼성동   세곡동   수서동   신사동
##     1001      211     1052      895      474      195      340       41
## 압구정동   역삼동   일원동   자곡동   청담동
##      378      632      307       53      378
```

```
data_whole %>% group_by(dong) %>%
  summarize(count = n(), avg_price = mean(unit_price), std_price = sd(unit_price)) # dong별 평균 및 표준편차
```

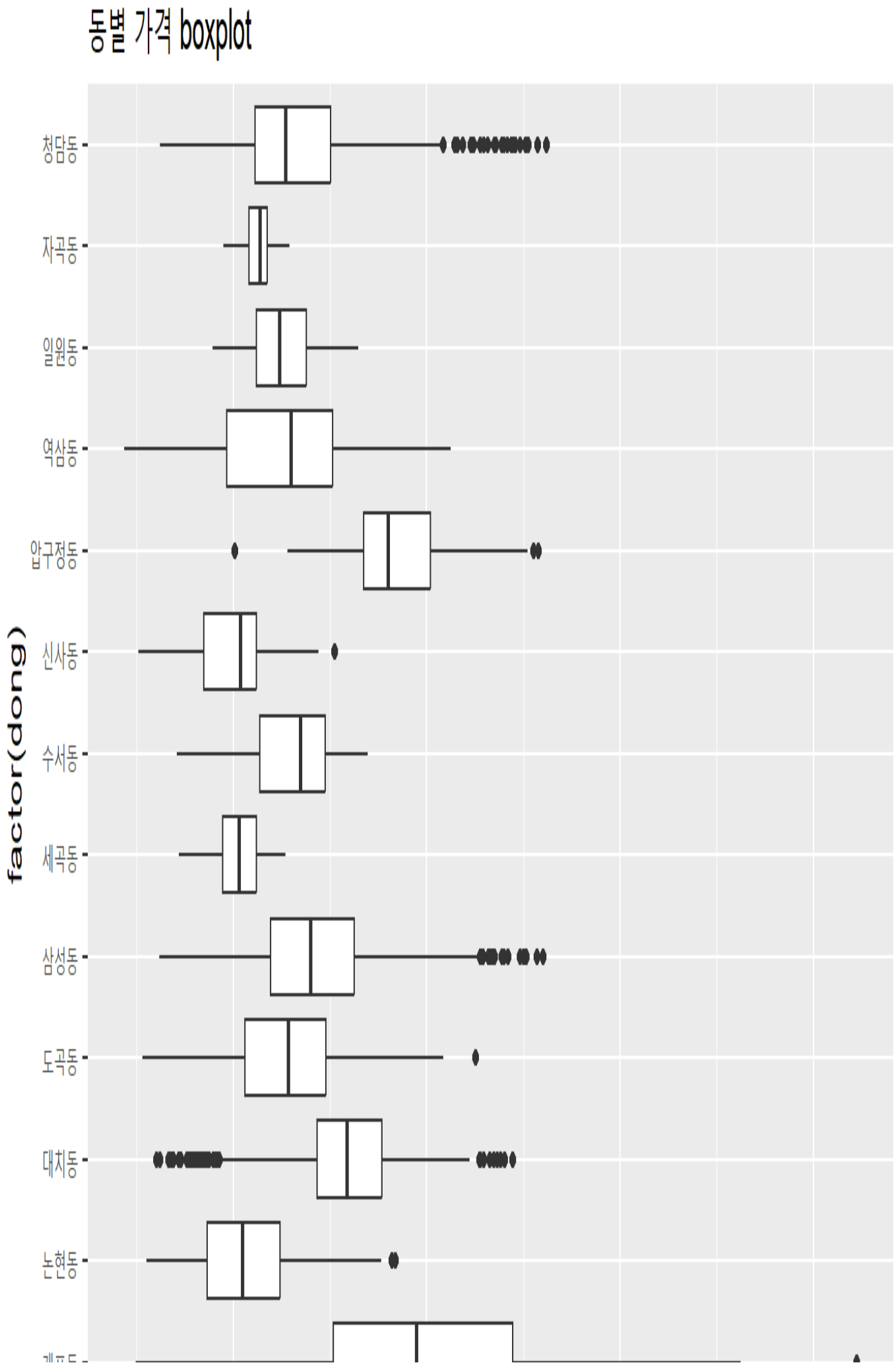
```
## # A tibble: 13 × 4
##    dong     count avg_price std_price
##    <fct>    <int>     <dbl>     <dbl>
##  1 개포동    1001     1955.     567.
##  2 논현동     211     1079.     289.
##  3 대치동    1052     1580.     309.
##  4 도곡동     895     1282.     291.
##  5 삼성동     474     1427.     374.
##  6 세곡동     195     1012.     133.
##  7 수서동     340     1309.     208.
##  8 신사동      41      993.     249.
##  9 압구정동   378     1858.     263.
## 10 역삼동     632     1266.     348.
## 11 일원동     307     1250.     162.
## 12 자곡동      53     1129.      79.4
## 13 청담동     378     1360.     376.
```

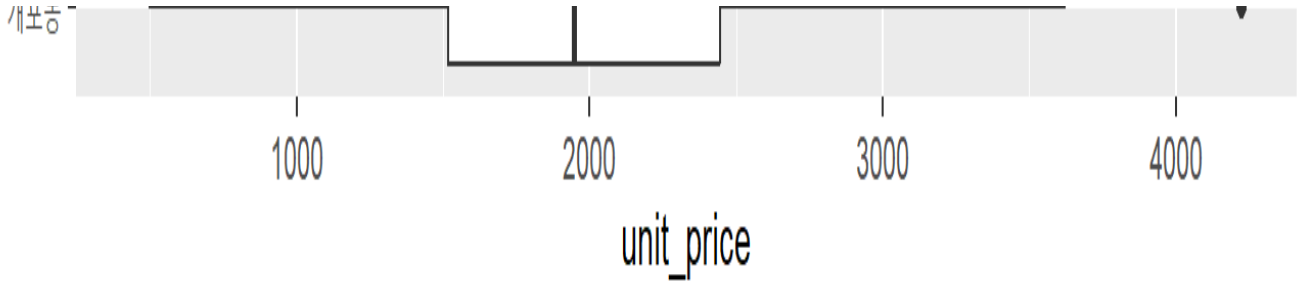
```
View(data_whole %>% group_by(dong) %>%
  summarize(count = n(), avg_price = mean(unit_price), std_price = sd(unit_price))) # dong별 평균 및 표준편차
```

```
# unit price
summary(data_whole$unit_price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   435.4  1144.5  1438.5  1479.8  1693.7  4222.8
```

```
ggplot(data = data_whole, aes(x = factor(dong), y = unit_price)) + geom_boxplot() + coord_flip() +ggtitle("동별 가격 boxplot")
```


동별 가격 boxplot
factor(dong)
청담동
자곡동
일원동
역삼동
압구정동
신사동
수서동
세곡동
삼성동
도곡동
대치동
논현동



강남구 단위당 가격 분석

트레이닝 데이터와 테스트 데이터로 split

```
# Data transformation for Tree & Regression Model
data_whole1 <- data_whole

install.packages('caTools', repos ="http://cran.us.r-project.org")
```

```
## 패키지 'caTools'를 성공적으로 압축해제하였고 MD5 sums 이 확인되었습니다
##
## 다운로드된 바이너리 패키지들은 다음의 위치에 있습니다
##  C:₩Users₩LUIS₩AppData₩Local₩Temp₩RtmpwHb08e₩downloaded_packages
```

```
library(caTools)

set.seed(123)
sample = sample.split(data_whole1$unit_price, SplitRatio = .7)
data_train1 = subset(data_whole1, sample == TRUE)
data_test1  = subset(data_whole1, sample == FALSE)

str(data_train1); mean(data_train1$unit_price)
```

```
## 'data.frame':    4169 obs. of  28 variables:
##  $ dong                : Factor w/ 13 levels "개포동","논현동",..: 10 10 10 10 10 10 10 10 1
0 10 ...
##  $ exclusive_use_area  : num  30 59.6 85 120.4 116.8 ...
##  $ floor               : int  11 6 2 22 2 21 5 6 11 13 ...
##  $ bigMarket05         : int  0 4 2 3 0 0 2 2 1 0 ...
##  $ bigMarket10         : int  3 6 6 6 5 3 6 4 3 2 ...
##  $ bigMarket15         : int  10 12 11 15 12 10 11 11 11 9 ...
##  $ school05            : int  1 5 3 2 1 1 3 3 0 2 ...
##  $ school10            : int  4 13 7 13 9 4 7 13 6 4 ...
##  $ school15            : int  16 20 16 23 20 16 16 18 20 16 ...
##  $ subway05            : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ subway10            : int  4 5 5 4 2 4 5 4 3 5 ...
##  $ subway15            : int  9 10 6 9 11 9 6 10 12 10 ...
##  $ hospital05          : int  47 105 53 64 56 47 53 62 57 50 ...
##  $ hospital10          : int  345 312 480 254 233 345 480 309 422 347 ...
##  $ hospital15          : int  875 612 829 552 864 875 829 583 1103 724 ...
##  $ movie05             : int  7 4 5 3 5 7 5 5 5 6 ...
##  $ movie10             : int  24 16 27 15 19 24 27 15 31 25 ...
##  $ movie15             : int  59 44 65 40 50 59 65 40 57 55 ...
##  $ kid05               : int  5 13 5 10 6 5 5 12 3 6 ...
##  $ kid10               : int  25 36 27 31 28 25 27 29 28 25 ...
##  $ kid15               : int  62 61 63 60 64 62 63 60 62 61 ...
##  $ office05            : int  3 1 3 2 2 3 3 3 3 2 ...
##  $ office10            : int  7 6 13 5 9 7 13 6 10 7 ...
##  $ office15            : int  15 15 20 16 16 15 20 15 17 12 ...
##  $ unit_price          : num  879 1502 942 1088 548 ...
##  $ transaction_month   : Factor w/ 11 levels "01","02","03",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ year                : Factor w/ 43 levels "1974","1976",..: 28 31 29 37 33 28 29 32 30 27
...
##  $ year_of_completion_f: Factor w/ 4 levels "1st","2nd","3rd",..: 3 3 3 4 3 3 3 3 3 2 ...
```

```
## [1] 1476.591
```

```
str(data_test1);mean(data_test1$unit_price)
```

```
## 'data.frame':    1788 obs. of  28 variables:
##  $ dong                : Factor w/ 13 levels "개포동","논현동",..: 10 10 10 10 10 10 10 10 1
0 10 ...
##  $ exclusive_use_area  : num  114 39.2 59.4 85 59.9 ...
##  $ floor               : int  7 23 6 17 21 7 18 12 3 4 ...
##  $ bigMarket05         : int  1 0 4 1 2 4 3 4 2 1 ...
##  $ bigMarket10         : int  3 2 7 6 4 7 6 5 6 4 ...
##  $ bigMarket15         : int  10 9 12 12 11 12 15 12 9 10 ...
##  $ school05            : int  3 2 3 2 3 3 2 1 2 3 ...
##  $ school10            : int  6 4 13 12 13 13 13 13 6 5 ...
##  $ school15            : int  16 16 23 22 18 23 23 23 14 18 ...
##  $ subway05            : int  0 0 1 0 0 1 0 1 0 2 ...
##  $ subway10            : int  5 5 5 4 4 5 4 5 4 6 ...
##  $ subway15            : int  10 10 9 7 10 9 9 9 7 9 ...
##  $ hospital05          : int  58 50 100 48 62 100 64 82 86 137 ...
##  $ hospital10          : int  343 347 327 283 309 327 254 315 507 330 ...
##  $ hospital15          : int  704 724 575 580 583 575 552 533 852 615 ...
##  $ movie05             : int  7 6 4 4 5 4 3 3 6 5 ...
##  $ movie10             : int  23 25 13 17 15 13 15 15 26 25 ...
##  $ movie15             : int  54 55 38 39 40 38 40 36 63 52 ...
##  $ kid05               : int  8 6 14 9 12 14 10 12 5 6 ...
##  $ kid10               : int  26 25 33 30 29 33 31 32 26 31 ...
##  $ kid15               : int  63 61 61 63 60 61 60 63 65 64 ...
##  $ office05            : int  3 2 1 2 3 1 2 1 3 2 ...
##  $ office10            : int  7 7 6 6 6 6 5 6 11 8 ...
##  $ office15            : int  13 12 14 16 15 14 16 16 20 14 ...
##  $ unit_price          : num  1316 676 1481 936 1461 ...
##  $ transaction_month   : Factor w/ 11 levels "01","02","03",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ year                : Factor w/ 43 levels "1974","1976",..: 42 27 31 23 32 31 37 35 22 30
...
##  $ year_of_completion_f: Factor w/ 4 levels "1st","2nd","3rd",..: 4 2 3 1 3 3 4 4 1 3 ...
```

```
## [1] 1487.179
```

Decision Tree

```
install.packages("rpart", repos ="http://cran.us.r-project.org")
```

```
## 패키지 'rpart'를 성공적으로 압축해제하였고 MD5 sums 이 확인되었습니다
```

```
## Warning: 패키지 'rpart'의 이전설치를 삭제할 수 없습니다
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE):
## C:WUsersWLUISWWorkSpaceWRWR-4.3.0WlibraryW00LOCKWrpartWlibsWx64Wrpart.dll를
## C:WUsersWLUISWWorkSpaceWRWR-4.3.0WlibraryWrpartWlibsWx64Wrpart.dll로 복사하는데
## 문제가 발생했습니다: Permission denied
```

```
## Warning: 'rpart'를 복구하였습니다
```

```
##
## 다운로드된 바이너리 패키지들은 다음의 위치에 있습니다
##  C:WUsersWLUISWAppDataWLocalWTempWRtmpwHb08eWdownloaded_packages
```

```
library(rpart)
tree1 <- rpart(unit_price~.-year,
               data=data_train1,
               method = "anova",
               control = rpart.control(minsplit = 50, maxdepth = 5))

# tree 결과
print(tree1)
```

```
## n= 4169
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 4169 858944800 1476.591
##    2) dong=논현동,도곡동,삼성동,세곡동,수서동,신사동,역삼동,일원동,자곡동,청담동 2471 253021
900 1275.302
##      4) office15>=16.5 359  14333400  995.133 *
##      5) office15< 16.5 2112 205718900 1322.926
##       10) dong=세곡동,신사동,자곡동 201   4846654 1031.511 *
##       11) dong=논현동,도곡동,삼성동,수서동,역삼동,일원동,청담동 1911 182007400 1353.577
##         22) year_of_completion_f=1st,2nd,3rd 1693 127025900 1315.791
##           44) hospital15>=626.5 292  20989770 1135.390 *
##           45) hospital15< 626.5 1401  94552520 1353.391 *
##         23) year_of_completion_f=4th 218  33792140 1647.023
##           46) hospital05>=26.5 185  18920560 1543.804 *
##           47) hospital05< 26.5 33   1850879 2225.676 *
##    3) dong=개포동,대치동,압구정동 1698 360108500 1769.516
##      6) hospital10>=67 1362 140782500 1612.422
##       12) year_of_completion_f=2nd,3rd 205  18344980 1283.098
##         24) school15< 27 107   4787220 1061.558 *
##         25) school15>=27 98   2572391 1524.983 *
##       13) year_of_completion_f=1st,4th 1157  96265060 1670.772
##         26) hospital10< 215 714  35414700 1551.099
##           52) hospital05< 20.5 145   3281968 1240.582 *
##           53) hospital05>=20.5 569  14588880 1630.229 *
##         27) hospital10>=215 443  34143450 1863.655
##           54) exclusive_use_area>=97.5 252  12962240 1726.931 *
##           55) exclusive_use_area< 97.5 191  10255200 2044.045 *
##      7) hospital10< 67 336  49465720 2406.305
##       14) kid10>=17.5 148  19728030 2152.290
##         28) exclusive_use_area>=50.98 18   5409467 1451.502 *
##         29) exclusive_use_area< 50.98 130   4254701 2249.322 *
##       15) kid10< 17.5 188  12670500 2606.274 *
```

```
summary(tree1)
```

```
## Call:
## rpart(formula = unit_price ~ . - year, data = data_train1, method = "anova",
##     control = rpart.control(minsplit = 50, maxdepth = 5))
##   n= 4169
##
##            CP nsplit rel error    xerror        xstd
## 1  0.28618180      0 1.0000000 1.0004794 0.026593270
## 2  0.19775461      1 0.7138182 0.7145316 0.019751941
## 3  0.03838391      2 0.5160636 0.5170454 0.015634069
## 4  0.03078159      3 0.4776797 0.4787984 0.015052161
## 5  0.02331589      5 0.4161165 0.4206353 0.014033836
## 6  0.02042489      7 0.3694847 0.3796427 0.012658378
## 7  0.01986995      8 0.3490598 0.3637032 0.012138491
## 8  0.01515895      9 0.3291899 0.3336834 0.010594263
## 9  0.01336947     10 0.3140309 0.3250919 0.010415841
## 10 0.01278938     11 0.3006615 0.3169476 0.010355350
## 11 0.01272027     12 0.2878721 0.3082818 0.010218387
## 12 0.01171655     13 0.2751518 0.2940953 0.010076593
## 13 0.01000000     14 0.2634353 0.2870630 0.009976371
##
## Variable importance
##                 dong           hospital10           hospital15
##                   13                   12                   11
##           hospital05              office15              movie10
##                    9                    7                    6
##          bigMarket10              office05 year_of_completion_f
##                    6                    5                    5
##              movie15          bigMarket15              school15
##                    5                    4                    4
##                kid10              subway15              school10
##                    2                    2                    1
##   exclusive_use_area                 kid15                 kid05
##                    1                    1                    1
##             office10              subway10          bigMarket05
##                    1                    1                    1
##
## Node number 1: 4169 observations,    complexity param=0.2861818
##   mean=1476.591, MSE=206031.4
##   left son=2 (2471 obs) right son=3 (1698 obs)
##   Primary splits:
##       dong               splits as  RLRLLLLLRLLLL, improve=0.2861818, (0 missing)
##       kid10              < 13.5     to the right, improve=0.2091617, (0 missing)
##       kid15              < 36.5     to the right, improve=0.1657105, (0 missing)
##       exclusive_use_area < 58.205   to the right, improve=0.1404478, (0 missing)
##       kid05              < 2.5      to the right, improve=0.1337925, (0 missing)
##   Surrogate splits:
##       bigMarket10          < 7.5      to the left,  agree=0.776, adj=0.451, (0 split)
##       year_of_completion_f splits as  RLLL,         agree=0.708, adj=0.283, (0 split)
##       bigMarket15          < 12.5     to the left,  agree=0.707, adj=0.282, (0 split)
##       movie15              < 28.5     to the right, agree=0.705, adj=0.277, (0 split)
##       school15             < 23.5     to the left,  agree=0.705, adj=0.276, (0 split)
##
## Node number 2: 2471 observations,    complexity param=0.03838391
##   mean=1275.302, MSE=102396.6
##   left son=4 (359 obs) right son=5 (2112 obs)
```

```
##   Primary splits:
##       office15   < 16.5     to the right, improve=0.13030360, (0 missing)
##       hospital10 < 334.5    to the right, improve=0.08028217, (0 missing)
##       hospital15 < 826.5    to the right, improve=0.07936021, (0 missing)
##       office05   < 1.5      to the right, improve=0.07867135, (0 missing)
##       dong       splits as  -L-RRLRL-RRLR, improve=0.07512544, (0 missing)
##   Surrogate splits:
##       office10   < 8.5      to the right, agree=0.918, adj=0.435, (0 split)
##       hospital15 < 1239.5   to the right, agree=0.865, adj=0.070, (0 split)
##       subway15   < 13.5     to the right, agree=0.864, adj=0.061, (0 split)
##       office05   < 3.5      to the right, agree=0.862, adj=0.053, (0 split)
##       hospital10 < 457.5    to the right, agree=0.862, adj=0.050, (0 split)
##
## Node number 3: 1698 observations,    complexity param=0.1977546
##   mean=1769.516, MSE=212078
##   left son=6 (1362 obs) right son=7 (336 obs)
##   Primary splits:
##       hospital10 < 67       to the right, improve=0.4716919, (0 missing)
##       hospital15 < 171.5    to the right, improve=0.4633918, (0 missing)
##       office15   < 9.5      to the right, improve=0.4276676, (0 missing)
##       movie10    < 3        to the right, improve=0.4110417, (0 missing)
##       office10   < 3.5      to the right, improve=0.4049248, (0 missing)
##   Surrogate splits:
##       hospital15 < 171.5    to the right, agree=0.999, adj=0.994, (0 split)
##       hospital05 < 13.5     to the right, agree=0.949, adj=0.744, (0 split)
##       office05   < 0.5      to the right, agree=0.922, adj=0.607, (0 split)
##       office15   < 9.5      to the right, agree=0.918, adj=0.583, (0 split)
##       movie10    < 3        to the right, agree=0.913, adj=0.560, (0 split)
##
## Node number 4: 359 observations
##   mean=995.133, MSE=39925.9
##
## Node number 5: 2112 observations,    complexity param=0.02331589
##   mean=1322.926, MSE=97404.76
##   left son=10 (201 obs) right son=11 (1911 obs)
##   Primary splits:
##       dong     splits as  -R-RRLRL-RRLR, improve=0.09170180, (0 missing)
##       movie15  < 30.5     to the left,  improve=0.07548871, (0 missing)
##       movie10  < 9.5      to the left,  improve=0.07089845, (0 missing)
##       office15 < 6.5      to the left,  improve=0.06799517, (0 missing)
##       subway15 < 1.5      to the left,  improve=0.06552860, (0 missing)
##   Surrogate splits:
##       subway10   < 0.5      to the left,  agree=0.978, adj=0.771, (0 split)
##       subway15   < 1.5      to the left,  agree=0.978, adj=0.771, (0 split)
##       hospital15 < 60       to the left,  agree=0.978, adj=0.771, (0 split)
##       office15   < 5.5      to the left,  agree=0.978, adj=0.771, (0 split)
##       hospital10 < 30.5     to the left,  agree=0.968, adj=0.662, (0 split)
##
## Node number 6: 1362 observations,    complexity param=0.03078159
##   mean=1612.422, MSE=103364.5
##   left son=12 (205 obs) right son=13 (1157 obs)
##   Primary splits:
##       year_of_completion_f splits as  RLLR,         improve=0.1859071, (0 missing)
##       school15             < 12.5     to the right, improve=0.1750337, (0 missing)
##       subway10             < 1.5      to the right, improve=0.1670550, (0 missing)
##       kid15                < 30       to the right, improve=0.1598935, (0 missing)
```

```
##       hospital15          < 675      to the left,  improve=0.1597882, (0 missing)
##   Surrogate splits:
##       kid10    < 35.5     to the right, agree=0.927, adj=0.512, (0 split)
##       subway15 < 11.5     to the right, agree=0.895, adj=0.302, (0 split)
##       kid05    < 11.5     to the right, agree=0.885, adj=0.239, (0 split)
##       school05 < 7.5      to the right, agree=0.882, adj=0.215, (0 split)
##       office05 < 3.5      to the right, agree=0.876, adj=0.176, (0 split)
##
## Node number 7: 336 observations,    complexity param=0.01986995
##   mean=2406.305, MSE=147219.4
##   left son=14 (148 obs) right son=15 (188 obs)
##   Primary splits:
##       kid10      < 17.5     to the right, improve=0.3450306, (0 missing)
##       hospital10  < 35       to the right, improve=0.3450306, (0 missing)
##       bigMarket15 < 8        to the right, improve=0.3450306, (0 missing)
##       movie10     < 3        to the right, improve=0.3450306, (0 missing)
##       kid05       < 4        to the right, improve=0.3450306, (0 missing)
##   Surrogate splits:
##       bigMarket15 < 8        to the right, agree=1, adj=1, (0 split)
##       school10    < 11       to the left,  agree=1, adj=1, (0 split)
##       hospital05  < 8.5      to the left,  agree=1, adj=1, (0 split)
##       hospital10  < 35       to the right, agree=1, adj=1, (0 split)
##       movie10     < 3        to the right, agree=1, adj=1, (0 split)
##
## Node number 10: 201 observations
##   mean=1031.511, MSE=24112.7
##
## Node number 11: 1911 observations,    complexity param=0.02331589
##   mean=1353.577, MSE=95241.97
##   left son=22 (1693 obs) right son=23 (218 obs)
##   Primary splits:
##       year_of_completion_f splits as  LLLR,         improve=0.11642010, (0 missing)
##       office05             < 1.5      to the right, improve=0.08498217, (0 missing)
##       hospital10           < 334.5    to the right, improve=0.07217171, (0 missing)
##       exclusive_use_area   < 96.755   to the right, improve=0.06984501, (0 missing)
##       school10             < 3.5      to the right, improve=0.06533256, (0 missing)
##   Surrogate splits:
##       subway10           < 6.5      to the left,  agree=0.900, adj=0.124, (0 split)
##       kid05              < 14.5     to the left,  agree=0.900, adj=0.124, (0 split)
##       exclusive_use_area < 27.97955 to the right, agree=0.891, adj=0.041, (0 split)
##       bigMarket10        < 7.5      to the left,  agree=0.890, adj=0.037, (0 split)
##
## Node number 12: 205 observations,    complexity param=0.01278938
##   mean=1283.098, MSE=89487.72
##   left son=24 (107 obs) right son=25 (98 obs)
##   Primary splits:
##       school15    < 27       to the left,  improve=0.5988216, (0 missing)
##       bigMarket10 < 10.5     to the left,  improve=0.5713319, (0 missing)
##       school10    < 13.5     to the left,  improve=0.5222204, (0 missing)
##       movie15     < 30.5     to the right, improve=0.5132195, (0 missing)
##       movie10     < 7.5      to the right, improve=0.5132195, (0 missing)
##   Surrogate splits:
##       bigMarket10 < 10.5     to the left,  agree=0.980, adj=0.959, (0 split)
##       school10    < 12.5     to the left,  agree=0.980, adj=0.959, (0 split)
##       movie10     < 9.5      to the right, agree=0.971, adj=0.939, (0 split)
##       movie15     < 35.5     to the right, agree=0.971, adj=0.939, (0 split)
```

```
##       hospital10  < 231      to the right, agree=0.951, adj=0.898, (0 split)
##
## Node number 13: 1157 observations,    complexity param=0.03078159
##   mean=1670.772, MSE=83202.3
##   left son=26 (714 obs) right son=27 (443 obs)
##   Primary splits:
##       hospital10 < 215      to the left,  improve=0.2774310, (0 missing)
##       hospital15 < 429.5    to the left,  improve=0.2489341, (0 missing)
##       school15   < 12.5     to the right, improve=0.1751949, (0 missing)
##       subway10   < 1.5      to the right, improve=0.1696628, (0 missing)
##       movie15    < 84.5     to the left,  improve=0.1564824, (0 missing)
##   Surrogate splits:
##       hospital15 < 429.5    to the left,  agree=0.964, adj=0.905, (0 split)
##       movie15    < 24.5     to the left,  agree=0.900, adj=0.738, (0 split)
##       hospital05 < 67       to the left,  agree=0.860, adj=0.634, (0 split)
##       movie10    < 21.5     to the left,  agree=0.842, adj=0.587, (0 split)
##       dong       splits as  L-L-----R----, agree=0.839, adj=0.580, (0 split)
##
## Node number 14: 148 observations,    complexity param=0.01171655
##   mean=2152.29, MSE=133297.5
##   left son=28 (18 obs) right son=29 (130 obs)
##   Primary splits:
##       exclusive_use_area < 50.98    to the right, improve=0.5101302, (0 missing)
##       subway15           < 4.5      to the left,  improve=0.1767474, (0 missing)
##       school15           < 21       to the left,  improve=0.1767474, (0 missing)
##       hospital15         < 145      to the left,  improve=0.1767474, (0 missing)
##       movie05            < 1        to the right, improve=0.1767474, (0 missing)
##   Surrogate splits:
##       subway15    < 1.5      to the left,  agree=0.959, adj=0.667, (0 split)
##       hospital10  < 54.5     to the left,  agree=0.959, adj=0.667, (0 split)
##       hospital15  < 110.5    to the left,  agree=0.959, adj=0.667, (0 split)
##       kid15       < 39.5     to the left,  agree=0.959, adj=0.667, (0 split)
##       bigMarket15 < 9.5      to the left,  agree=0.946, adj=0.556, (0 split)
##
## Node number 15: 188 observations
##   mean=2606.274, MSE=67396.27
##
## Node number 22: 1693 observations,    complexity param=0.01336947
##   mean=1315.791, MSE=75030.09
##   left son=44 (292 obs) right son=45 (1401 obs)
##   Primary splits:
##       hospital15 < 626.5    to the right, improve=0.09040390, (0 missing)
##       hospital10 < 334.5    to the right, improve=0.08552526, (0 missing)
##       school10   < 11.5     to the left,  improve=0.07524451, (0 missing)
##       office05   < 1.5      to the right, improve=0.07495491, (0 missing)
##       office10   < 6.5      to the right, improve=0.07381685, (0 missing)
##   Surrogate splits:
##       hospital10 < 334.5    to the right, agree=0.931, adj=0.599, (0 split)
##       subway15   < 10.5     to the right, agree=0.926, adj=0.568, (0 split)
##       office10   < 6.5      to the right, agree=0.924, adj=0.562, (0 split)
##       movie15    < 135.5    to the right, agree=0.913, adj=0.493, (0 split)
##       subway10   < 5.5      to the right, agree=0.876, adj=0.281, (0 split)
##
## Node number 23: 218 observations,    complexity param=0.01515895
##   mean=1647.023, MSE=155009.8
##   left son=46 (185 obs) right son=47 (33 obs)
```

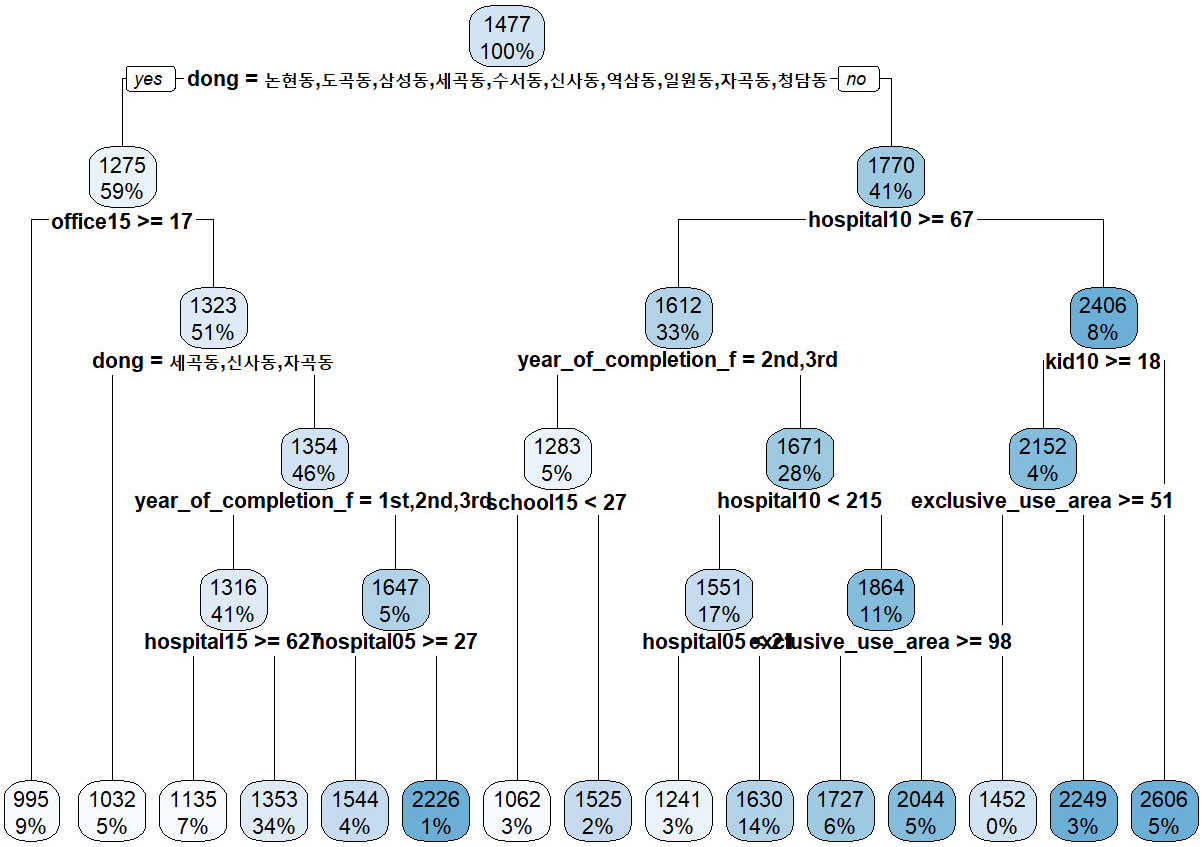
```
##   Primary splits:
##       hospital05 < 26.5     to the right, improve=0.3853175, (0 missing)
##       hospital15 < 460.5    to the right, improve=0.3853175, (0 missing)
##       hospital10 < 203.5    to the right, improve=0.3853175, (0 missing)
##       office15   < 8.5      to the right, improve=0.3775081, (0 missing)
##       kid05      < 3.5      to the right, improve=0.3763037, (0 missing)
##   Surrogate splits:
##       hospital10 < 203.5    to the right, agree=1.000, adj=1.000, (0 split)
##       hospital15 < 460.5    to the right, agree=1.000, adj=1.000, (0 split)
##       kid05      < 3.5      to the right, agree=0.995, adj=0.970, (0 split)
##       kid15      < 37.5     to the right, agree=0.991, adj=0.939, (0 split)
##       office15   < 8.5      to the right, agree=0.982, adj=0.879, (0 split)
##
## Node number 24: 107 observations
##   mean=1061.558, MSE=44740.37
##
## Node number 25: 98 observations
##   mean=1524.983, MSE=26248.89
##
## Node number 26: 714 observations,    complexity param=0.02042489
##   mean=1551.099, MSE=49600.42
##   left son=52 (145 obs) right son=53 (569 obs)
##   Primary splits:
##       hospital05         < 20.5     to the left,  improve=0.4953833, (0 missing)
##       kid10              < 26.5     to the left,  improve=0.3261878, (0 missing)
##       bigMarket05        < 1.5      to the left,  improve=0.2995789, (0 missing)
##       subway05           < 0.5      to the left,  improve=0.2375837, (0 missing)
##       exclusive_use_area < 84.62    to the right, improve=0.2202752, (0 missing)
##   Surrogate splits:
##       bigMarket05 < 1.5      to the left,  agree=0.940, adj=0.703, (0 split)
##       kid10       < 26.5     to the left,  agree=0.926, adj=0.634, (0 split)
##       subway05    < 0.5      to the left,  agree=0.892, adj=0.469, (0 split)
##       office05    < 1.5      to the left,  agree=0.891, adj=0.462, (0 split)
##       movie15     < 24.5     to the right, agree=0.887, adj=0.441, (0 split)
##
## Node number 27: 443 observations,    complexity param=0.01272027
##   mean=1863.655, MSE=77073.24
##   left son=54 (252 obs) right son=55 (191 obs)
##   Primary splits:
##       exclusive_use_area < 97.5     to the right, improve=0.3200030, (0 missing)
##       kid10              < 34.5     to the right, improve=0.2158932, (0 missing)
##       kid05              < 12.5     to the right, improve=0.1892850, (0 missing)
##       school10           < 14.5     to the right, improve=0.1401558, (0 missing)
##       office10           < 5.5      to the right, improve=0.1304192, (0 missing)
##   Surrogate splits:
##       subway15    < 10.5     to the left,  agree=0.682, adj=0.262, (0 split)
##       kid15       < 61.5     to the left,  agree=0.682, adj=0.262, (0 split)
##       office15    < 12.5     to the left,  agree=0.682, adj=0.262, (0 split)
##       bigMarket05 < 4        to the left,  agree=0.675, adj=0.246, (0 split)
##       school05    < 1.5      to the right, agree=0.675, adj=0.246, (0 split)
##
## Node number 28: 18 observations
##   mean=1451.502, MSE=300525.9
##
## Node number 29: 130 observations
##   mean=2249.322, MSE=32728.47
```

```
##
## Node number 44: 292 observations
##   mean=1135.39, MSE=71882.78
##
## Node number 45: 1401 observations
##   mean=1353.391, MSE=67489.31
##
## Node number 46: 185 observations
##   mean=1543.804, MSE=102273.3
##
## Node number 47: 33 observations
##   mean=2225.676, MSE=56087.24
##
## Node number 52: 145 observations
##   mean=1240.582, MSE=22634.26
##
## Node number 53: 569 observations
##   mean=1630.229, MSE=25639.51
##
## Node number 54: 252 observations
##   mean=1726.931, MSE=51437.47
##
## Node number 55: 191 observations
##   mean=2044.045, MSE=53692.14
```

```
install.packages("rpart.plot", repos = "http://cran.us.r-project.org")
```

```
## 패키지 'rpart.plot'를 성공적으로 압축해제하였고 MD5 sums 이 확인되었습니다
##
## 다운로드된 바이너리 패키지들은 다음의 위치에 있습니다
##  C:WUsersWLUISWAppDataWLocalWTempWRtmpwHb08eWdownloaded_packages
```

```
library(rpart.plot)

rpart.plot(tree1, cex = 0.7)
```

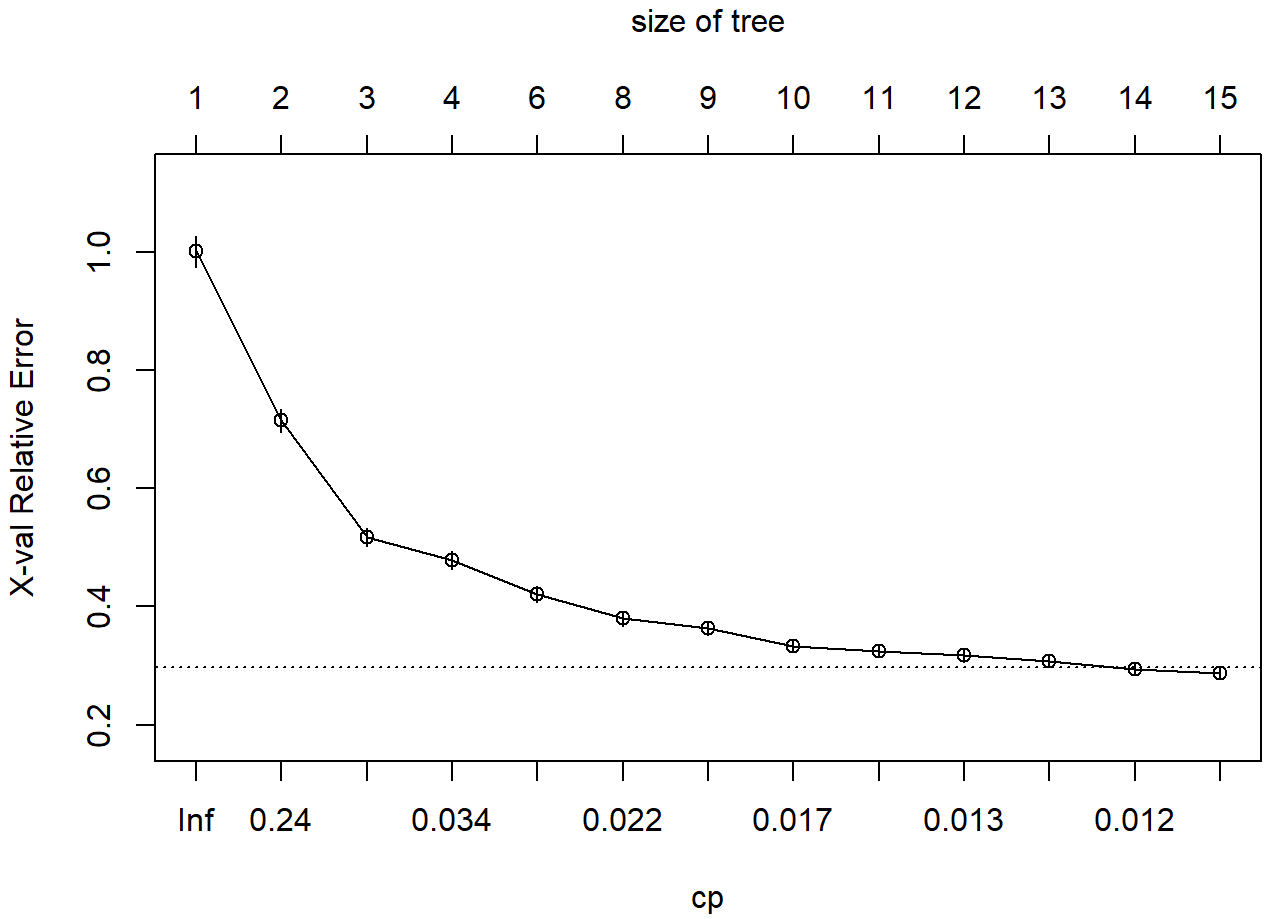


Decision Tree parameter tuning

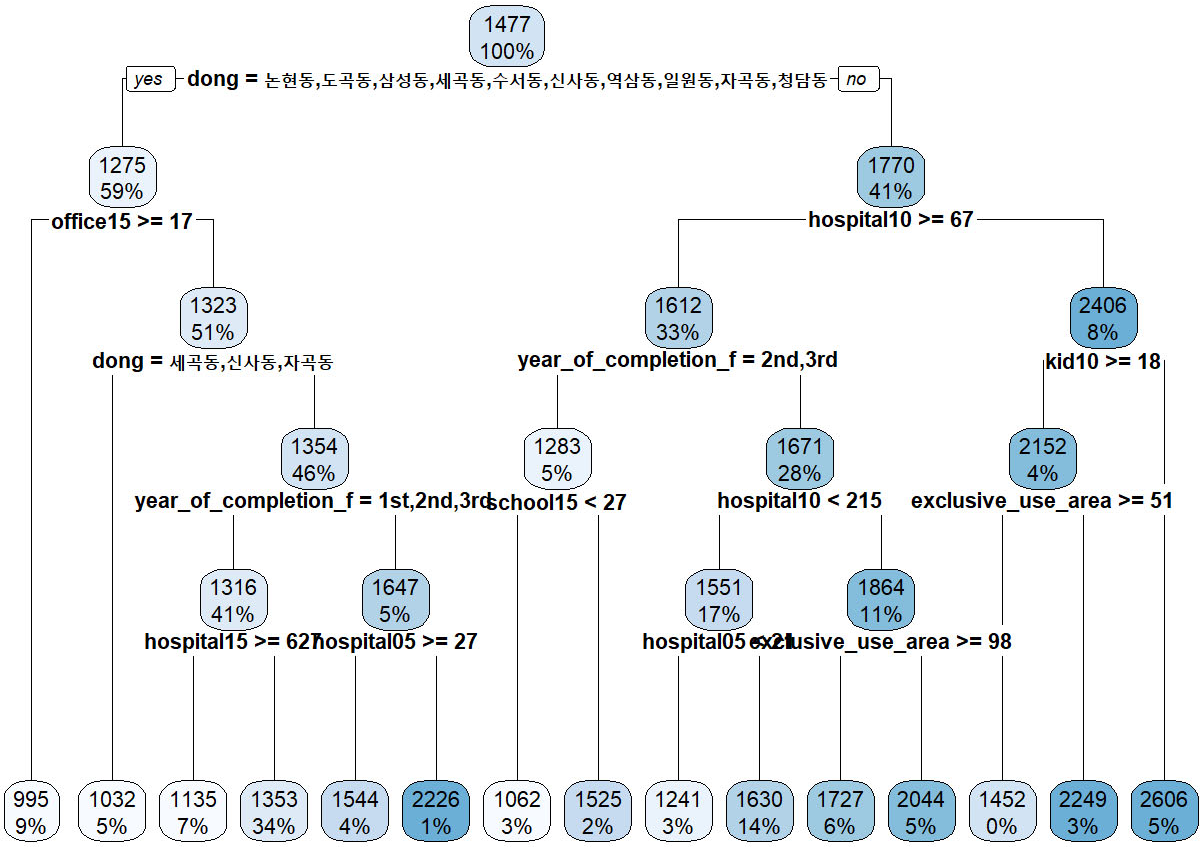
```
printcp(tree1)
```

```
##
## Regression tree:
## rpart(formula = unit_price ~ . - year, data = data_train1, method = "anova",
##     control = rpart.control(minsplit = 50, maxdepth = 5))
##
## Variables actually used in tree construction:
## [1] dong                 exclusive_use_area   hospital05
## [4] hospital10           hospital15           kid10
## [7] office15             school15             year_of_completion_f
##
## Root node error: 858944768/4169 = 206031
##
## n= 4169
##
##          CP nsplit rel error  xerror      xstd
## 1  0.286182      0   1.00000 1.00048 0.0265933
## 2  0.197755      1   0.71382 0.71453 0.0197519
## 3  0.038384      2   0.51606 0.51705 0.0156341
## 4  0.030782      3   0.47768 0.47880 0.0150522
## 5  0.023316      5   0.41612 0.42064 0.0140338
## 6  0.020425      7   0.36948 0.37964 0.0126584
## 7  0.019870      8   0.34906 0.36370 0.0121385
## 8  0.015159      9   0.32919 0.33368 0.0105943
## 9  0.013369     10   0.31403 0.32509 0.0104158
## 10 0.012789     11   0.30066 0.31695 0.0103553
## 11 0.012720     12   0.28787 0.30828 0.0102184
## 12 0.011717     13   0.27515 0.29410 0.0100766
## 13 0.010000     14   0.26344 0.28706 0.0099764
```

```
plotcp(tree1)
```




```
tree1 <- prune(tree1, cp= tree1$cptable[which.min(tree1$cptable[, "xerror"]), "CP"])

rpart.plot(tree1, cex = 0.7)
```



Decision Tree prediction & RMSE calculation

```
# test data 에 적용

predict_1 <- predict(tree1, data_test1)
summary(predict_1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   995.1  1240.6  1353.4  1491.7  1630.2  2606.3
```

```
# actual, predicted cbind

databind1 <- cbind(data_test1[,25],predict_1)
databind1 <- as.data.frame(databind1)
summary(databind1)
```

```
##        V1         predict_1
##  Min.   : 494   Min.   : 995.1
##  1st Qu.:1144   1st Qu.:1240.6
##  Median :1444   Median :1353.4
##  Mean   :1487   Mean   :1491.7
##  3rd Qu.:1713   3rd Qu.:1630.2
##  Max.   :3507   Max.   :2606.3
```

```
# RMSE 계산
install.packages("Metrics", repos = "http://cran.us.r-project.org")
```

```
## 패키지 'Metrics'를 성공적으로 압축해제하였고 MD5 sums 이 확인되었습니다
##
## 다운로드된 바이너리 패키지들은 다음의 위치에 있습니다
##  C:₩Users₩LUISW₩AppData₩Local₩Temp₩RtmpwHb08e₩downloaded_packages
```

```
library(Metrics)
rmse(databind1$V1, databind1$predict_1)
```

```
## [1] 234.9131
```

Linear regression

```
# factor 변수 중 unique value 있는지 찾아보기
str(data_train1)
```

```
## 'data.frame':    4169 obs. of  28 variables:
##  $ dong               : Factor w/ 13 levels "개포동","논현동",..: 10 10 10 10 10 10 10 10 1
0 10 ...
##  $ exclusive_use_area  : num  30 59.6 85 120.4 116.8 ...
##  $ floor               : int  11 6 2 22 2 21 5 6 11 13 ...
##  $ bigMarket05         : int  0 4 2 3 0 0 2 2 1 0 ...
##  $ bigMarket10         : int  3 6 6 6 5 3 6 4 3 2 ...
##  $ bigMarket15         : int  10 12 11 15 12 10 11 11 11 9 ...
##  $ school05            : int  1 5 3 2 1 1 3 3 0 2 ...
##  $ school10            : int  4 13 7 13 9 4 7 13 6 4 ...
##  $ school15            : int  16 20 16 23 20 16 16 18 20 16 ...
##  $ subway05            : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ subway10            : int  4 5 5 4 2 4 5 4 3 5 ...
##  $ subway15            : int  9 10 6 9 11 9 6 10 12 10 ...
##  $ hospital05          : int  47 105 53 64 56 47 53 62 57 50 ...
##  $ hospital10          : int  345 312 480 254 233 345 480 309 422 347 ...
##  $ hospital15          : int  875 612 829 552 864 875 829 583 1103 724 ...
##  $ movie05             : int  7 4 5 3 5 7 5 5 5 6 ...
##  $ movie10             : int  24 16 27 15 19 24 27 15 31 25 ...
##  $ movie15             : int  59 44 65 40 50 59 65 40 57 55 ...
##  $ kid05               : int  5 13 5 10 6 5 5 12 3 6 ...
##  $ kid10               : int  25 36 27 31 28 25 27 29 28 25 ...
##  $ kid15               : int  62 61 63 60 64 62 63 60 62 61 ...
##  $ office05            : int  3 1 3 2 2 3 3 3 3 2 ...
##  $ office10            : int  7 6 13 5 9 7 13 6 10 7 ...
##  $ office15            : int  15 15 20 16 16 15 20 15 17 12 ...
##  $ unit_price          : num  879 1502 942 1088 548 ...
##  $ transaction_month   : Factor w/ 11 levels "01","02","03",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ year                : Factor w/ 43 levels "1974","1976",..: 28 31 29 37 33 28 29 32 30 27
...
##  $ year_of_completion_f: Factor w/ 4 levels "1st","2nd","3rd",..: 3 3 3 4 3 3 3 3 3 2 ...
```

```
sapply(lapply(data_train1, unique), length)
```

```
##                 dong   exclusive_use_area                floor
##                   13                  770                   52
##          bigMarket05          bigMarket10          bigMarket15
##                    6                   12                   18
##             school05             school10             school15
##                    9                   21                   30
##             subway05             subway10             subway15
##                    5                   10                   18
##           hospital05           hospital10           hospital15
##                  146                  242                  274
##              movie05              movie10              movie15
##                   36                   92                  142
##                kid05                kid10                kid15
##                   18                   41                   59
##             office05             office10             office15
##                    7                   14                   21
##           unit_price    transaction_month                 year
##                 3461                   11                   43
## year_of_completion_f
##                    4
```

```
# Linear Model  (dong은 제외하고 분석:삭제)
linear1 <- lm(unit_price ~.-year, data = data_train1)
#linear1 <- lm(unit_price ~ dong+exclusive_use_area+floor+bigMarket05+bigMarket10+bigMarket15+school05+school10+school15+subway05+subway10+subway15+hospital05+hospital10+hospital15+movie05+movie10+movie15+kid05+kid10+kid15+office05+office10+office15+transaction_month+year_of_completion_f, data = data_train1)

summary(linear1)
```

```
##
## Call:
## lm(formula = unit_price ~ . - year, data = data_train1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1236.44  -120.38     8.58   134.55  1970.69
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.895e+03  4.542e+01  63.732  < 2e-16 ***
## dong논현동               1.197e+02  5.279e+01   2.267 0.023442 *
## dong대치동               5.173e+01  2.568e+01   2.015 0.044015 *
## dong도곡동              -1.016e+02  2.286e+01  -4.444 9.08e-06 ***
## dong삼성동              -1.034e+02  4.031e+01  -2.565 0.010357 *
## dong세곡동              -1.446e+03  4.552e+01 -31.774  < 2e-16 ***
## dong수서동              -6.380e+02  2.594e+01 -24.592  < 2e-16 ***
## dong신사동              -2.795e+02  7.375e+01  -3.789 0.000153 ***
## dong압구정동             1.169e+02  5.024e+01   2.326 0.020055 *
## dong역삼동               3.018e+00  2.837e+01   0.106 0.915289
## dong일원동              -6.235e+02  2.379e+01 -26.206  < 2e-16 ***
## dong자곡동              -1.002e+03  5.661e+01 -17.695  < 2e-16 ***
## dong청담동              -2.014e+02  4.111e+01  -4.899 9.98e-07 ***
## exclusive_use_area      -3.571e+00  1.258e-01 -28.374  < 2e-16 ***
## floor                    2.694e+00  6.311e-01   4.268 2.01e-05 ***
## bigMarket05              5.438e+01  5.259e+00  10.340  < 2e-16 ***
## bigMarket10              6.213e+00  5.772e+00   1.076 0.281813
## bigMarket15             -1.162e+01  3.903e+00  -2.977 0.002932 **
## school05                -7.823e+00  3.950e+00  -1.981 0.047689 *
## school10                 9.091e+00  2.444e+00   3.720 0.000202 ***
## school15                 6.973e+00  1.939e+00   3.596 0.000326 ***
## subway05                 2.952e+01  7.077e+00   4.171 3.10e-05 ***
## subway10                -3.169e+01  5.656e+00  -5.603 2.25e-08 ***
## subway15                 7.487e-01  4.090e+00   0.183 0.854756
## hospital05               3.093e-01  1.538e-01   2.011 0.044435 *
## hospital10              -4.236e-01  1.071e-01  -3.955 7.80e-05 ***
## hospital15              -3.586e-01  7.513e-02  -4.773 1.88e-06 ***
## movie05                 -8.406e+00  1.289e+00  -6.523 7.73e-11 ***
## movie10                 -5.255e-01  7.604e-01  -0.691 0.489559
## movie15                  1.515e+00  4.028e-01   3.760 0.000172 ***
## kid05                    9.939e-01  2.049e+00   0.485 0.627623
## kid10                   -4.971e-01  1.615e+00  -0.308 0.758225
## kid15                   -1.430e+01  9.340e-01 -15.309  < 2e-16 ***
## office05                -4.172e+01  5.584e+00  -7.472 9.58e-14 ***
## office10                -2.479e+01  3.664e+00  -6.768 1.49e-11 ***
## office15                -1.109e+01  2.396e+00  -4.627 3.82e-06 ***
## transaction_month02      3.159e-01  2.585e+01   0.012 0.990249
## transaction_month03      4.127e+01  2.420e+01   1.705 0.088195 .
## transaction_month04      4.222e+01  2.378e+01   1.776 0.075876 .
## transaction_month05      8.121e+01  2.222e+01   3.655 0.000260 ***
## transaction_month06      1.147e+02  2.350e+01   4.881 1.10e-06 ***
## transaction_month07      1.518e+02  2.277e+01   6.668 2.93e-11 ***
## transaction_month08      1.914e+02  2.885e+01   6.634 3.70e-11 ***
## transaction_month09      1.871e+02  2.491e+01   7.509 7.28e-14 ***
## transaction_month10      2.130e+02  2.534e+01   8.406  < 2e-16 ***
```

```
## transaction_month11      2.843e+02  2.370e+01  11.999  < 2e-16 ***
## year_of_completion_f2nd -1.994e+02  1.851e+01 -10.774  < 2e-16 ***
## year_of_completion_f3rd  6.407e+01  1.573e+01   4.074 4.71e-05 ***
## year_of_completion_f4th  2.065e+02  1.913e+01  10.794  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 243.9 on 4120 degrees of freedom
## Multiple R-squared:  0.7147, Adjusted R-squared:  0.7114
## F-statistic:   215 on 48 and 4120 DF,  p-value: < 2.2e-16
```

```
print(linear1)
```

```
##
## Call:
## lm(formula = unit_price ~ . - year, data = data_train1)
##
## Coefficients:
##             (Intercept)              dong논현동              dong대치동
##               2894.6878                119.6824                 51.7261
##              dong도곡동              dong삼성동              dong세곡동
##                -101.5853               -103.3995              -1446.2598
##              dong수서동              dong신사동             dong압구정동
##                -637.9718               -279.4632                116.8703
##              dong역삼동              dong일원동              dong자곡동
##                   3.0182               -623.5400              -1001.8032
##              dong청담동       exclusive_use_area                   floor
##                -201.4316                 -3.5705                  2.6935
##             bigMarket05             bigMarket10             bigMarket15
##                  54.3804                  6.2129                 -11.6181
##               school05                school10                school15
##                  -7.8232                  9.0911                  6.9729
##               subway05                subway10                subway15
##                  29.5174                -31.6917                  0.7487
##              hospital05              hospital10              hospital15
##                   0.3093                 -0.4236                 -0.3586
##                movie05                 movie10                 movie15
##                  -8.4056                 -0.5255                  1.5147
##                  kid05                   kid10                   kid15
##                   0.9939                 -0.4971                -14.2979
##               office05                office10                office15
##                 -41.7225                -24.7931                -11.0871
##     transaction_month02     transaction_month03     transaction_month04
##                   0.3159                 41.2743                 42.2169
##     transaction_month05     transaction_month06     transaction_month07
##                  81.2107                114.7025                151.8420
##     transaction_month08     transaction_month09     transaction_month10
##                 191.3998                187.0695                212.9742
##     transaction_month11 year_of_completion_f2nd year_of_completion_f3rd
##                 284.3466               -199.3804                 64.0660
## year_of_completion_f4th
##                 206.5099
```

```
linear1$coefficients
```

```
##             (Intercept)              dong논현동              dong대치동
##           2894.6877915             119.6824440              51.7261066
##              dong도곡동              dong삼성동              dong세곡동
##            -101.5852611            -103.3994568           -1446.2598095
##              dong수서동              dong신사동             dong압구정동
##            -637.9717883            -279.4632142             116.8702929
##              dong역삼동              dong일원동              dong자곡동
##               3.0181577            -623.5399931           -1001.8031534
##              dong청담동      exclusive_use_area                   floor
##            -201.4315714              -3.5704970               2.6935389
##             bigMarket05             bigMarket10             bigMarket15
##              54.3804380               6.2128589             -11.6180502
##                school05                school10                school15
##              -7.8232451               9.0910665               6.9729404
##                subway05                subway10                subway15
##              29.5174395             -31.6916975               0.7486977
##              hospital05              hospital10              hospital15
##               0.3092552              -0.4236044              -0.3585721
##                 movie05                 movie10                 movie15
##              -8.4055933              -0.5255063               1.5146592
##                   kid05                   kid10                   kid15
##               0.9938591              -0.4970688             -14.2979009
##                office05                office10                office15
##             -41.7225397             -24.7931363             -11.0870807
##     transaction_month02     transaction_month03     transaction_month04
##               0.3158854              41.2743145              42.2168563
##     transaction_month05     transaction_month06     transaction_month07
##              81.2107005             114.7025225             151.8419902
##     transaction_month08     transaction_month09     transaction_month10
##             191.3997505             187.0694662             212.9742195
##     transaction_month11 year_of_completion_f2nd year_of_completion_f3rd
##             284.3466072            -199.3804346              64.0659894
## year_of_completion_f4th
##             206.5099470
```

Linear regression parameter tuning

```
step(linear1, direction = "both")
```

```
## Start:  AIC=45880.15
## unit_price ~ (dong + exclusive_use_area + floor + bigMarket05 +
##     bigMarket10 + bigMarket15 + school05 + school10 + school15 +
##     subway05 + subway10 + subway15 + hospital05 + hospital10 +
##     hospital15 + movie05 + movie10 + movie15 + kid05 + kid10 +
##     kid15 + office05 + office10 + office15 + transaction_month +
##     year + year_of_completion_f) - year
##
##                        Df Sum of Sq       RSS   AIC
## - subway15              1      1993 245058349 45878
## - kid10                 1      5636 245061992 45878
## - kid05                 1     13998 245070353 45878
## - movie10               1     28406 245084762 45879
## - bigMarket10           1     68915 245125271 45879
## <none>                              245056356 45880
## - school05              1    233354 245289710 45882
## - hospital05            1    240442 245296798 45882
## - bigMarket15           1    526996 245583352 45887
## - school15              1    769358 245825714 45891
## - school10              1    823021 245879377 45892
## - movie15               1    841050 245897406 45892
## - hospital10            1    930162 245986518 45894
## - subway05              1   1034665 246091021 45896
## - floor                 1   1083603 246139959 45897
## - office15              1   1273380 246329735 45900
## - hospital15            1   1354755 246411111 45901
## - subway10              1   1867238 246923594 45910
## - movie05               1   2530690 247587046 45921
## - office10              1   2724140 247780496 45924
## - office05              1   3320969 248377325 45934
## - bigMarket05           1   6359676 251416032 45985
## - kid15                 1  13939893 258996249 46109
## - year_of_completion_f  3  24061013 269117369 46265
## - transaction_month    10  26818047 271874403 46293
## - exclusive_use_area    1  47885200 292941556 46622
## - dong                 12 126721660 371778016 47594
##
## Step:  AIC=45878.19
## unit_price ~ dong + exclusive_use_area + floor + bigMarket05 +
##     bigMarket10 + bigMarket15 + school05 + school10 + school15 +
##     subway05 + subway10 + hospital05 + hospital10 + hospital15 +
##     movie05 + movie10 + movie15 + kid05 + kid10 + kid15 + office05 +
##     office10 + office15 + transaction_month + year_of_completion_f
##
##                        Df Sum of Sq       RSS   AIC
## - kid10                 1      4177 245062526 45876
## - kid05                 1     15604 245073953 45876
## - movie10               1     32184 245090533 45877
## - bigMarket10           1     67445 245125794 45877
## <none>                              245058349 45878
## + subway15              1      1993 245056356 45880
## - school05              1    235118 245293467 45880
## - hospital05            1    247518 245305867 45880
## - bigMarket15           1    531235 245589585 45885
## - school10              1    822006 245880355 45890
```

```
## - school15              1    822282 245880632 45890
## - movie15               1    842438 245900787 45890
## - hospital10            1   1005716 246064065 45893
## - subway05              1   1032890 246091239 45894
## - floor                 1   1081829 246140179 45895
## - office15              1   1285551 246343900 45898
## - hospital15            1   1727120 246785470 45905
## - subway10              1   2049695 247108044 45911
## - movie05               1   2571828 247630178 45920
## - office10              1   2904030 247962379 45925
## - office05              1   3319805 248378154 45932
## - bigMarket05           1   6493273 251551622 45985
## - kid15                 1  14195131 259253480 46111
## - year_of_completion_f  3  24275526 269333875 46266
## - transaction_month    10  26870219 271928569 46292
## - exclusive_use_area    1  48245810 293304159 46625
## - dong                 12 130677986 375736335 47636
##
## Step:  AIC=45876.26
## unit_price ~ dong + exclusive_use_area + floor + bigMarket05 +
##     bigMarket10 + bigMarket15 + school05 + school10 + school15 +
##     subway05 + subway10 + hospital05 + hospital10 + hospital15 +
##     movie05 + movie10 + movie15 + kid05 + kid15 + office05 +
##     office10 + office15 + transaction_month + year_of_completion_f
##
##                        Df Sum of Sq       RSS   AIC
## - kid05                 1     11429 245073954 45874
## - movie10               1     32479 245095005 45875
## - bigMarket10           1     63996 245126522 45875
## <none>                              245062526 45876
## - school05              1    230983 245293509 45878
## + kid10                 1      4177 245058349 45878
## + subway15              1       534 245061992 45878
## - hospital05            1    245379 245307905 45878
## - bigMarket15           1    594175 245656701 45884
## - school10              1    835704 245898230 45888
## - movie15               1    839237 245901763 45889
## - school15              1    860696 245923222 45889
## - hospital10            1   1051385 246113910 45892
## - subway05              1   1086835 246149361 45893
## - floor                 1   1093135 246155661 45893
## - office15              1   1309736 246372262 45896
## - hospital15            1   1723674 246786200 45903
## - subway10              1   2064195 247126721 45909
## - movie05               1   2567932 247630458 45918
## - office10              1   2910470 247972996 45923
## - office05              1   3319451 248381976 45930
## - bigMarket05           1   6513522 251576048 45984
## - kid15                 1  17037282 262099807 46154
## - year_of_completion_f  3  24353975 269416501 46265
## - transaction_month    10  26866237 271928763 46290
## - exclusive_use_area    1  48285640 293348166 46624
## - dong                 12 133001075 378063600 47660
##
## Step:  AIC=45874.45
## unit_price ~ dong + exclusive_use_area + floor + bigMarket05 +
```

```
##     bigMarket10 + bigMarket15 + school05 + school10 + school15 +
##     subway05 + subway10 + hospital05 + hospital10 + hospital15 +
##     movie05 + movie10 + movie15 + kid15 + office05 + office10 +
##     office15 + transaction_month + year_of_completion_f
##
##                         Df Sum of Sq       RSS   AIC
## - movie10                1     28603 245102558 45873
## - bigMarket10            1    100530 245174485 45874
## <none>                               245073954 45874
## - school05               1    220679 245294633 45876
## + kid05                  1     11429 245062526 45876
## + subway15               1      2973 245070982 45876
## + kid10                  1         1 245073953 45876
## - hospital05             1    271865 245345820 45877
## - bigMarket15            1    585285 245659240 45882
## - movie15                1    842132 245916086 45887
## - school15               1    857605 245931559 45887
## - school10               1    863351 245937305 45887
## - hospital10             1   1084809 246158763 45891
## - floor                  1   1084855 246158810 45891
## - subway05               1   1150402 246224356 45892
## - office15               1   1299716 246373670 45895
## - hospital15             1   1719295 246793250 45902
## - subway10               1   2067403 247141357 45907
## - movie05                1   2624887 247698841 45917
## - office10               1   3112967 248186921 45925
## - office05               1   3389319 248463273 45930
## - bigMarket05            1   7866554 252940508 46004
## - kid15                  1  17692939 262766894 46163
## - year_of_completion_f   3  24621207 269695162 46268
## - transaction_month     10  26854889 271928844 46288
## - exclusive_use_area     1  48803257 293877211 46630
## - dong                  12 133343718 378417673 47662
##
## Step:  AIC=45872.94
## unit_price ~ dong + exclusive_use_area + floor + bigMarket05 +
##     bigMarket10 + bigMarket15 + school05 + school10 + school15 +
##     subway05 + subway10 + hospital05 + hospital10 + hospital15 +
##     movie05 + movie15 + kid15 + office05 + office10 + office15 +
##     transaction_month + year_of_completion_f
##
##                         Df Sum of Sq       RSS   AIC
## - bigMarket10            1    110085 245212643 45873
## <none>                               245102558 45873
## + movie10                1     28603 245073954 45874
## - school05               1    209406 245311964 45875
## + kid05                  1      7553 245095005 45875
## + subway15               1      5694 245096863 45875
## + kid10                  1       162 245102396 45875
## - hospital05             1    265977 245368535 45875
## - bigMarket15            1    627534 245730092 45882
## - school15               1    856255 245958813 45885
## - school10               1    857279 245959837 45885
## - floor                  1   1094881 246197439 45890
## - movie15                1   1116774 246219332 45890
## - hospital10             1   1203136 246305694 45891
```

```
## - subway05              1   1205576 246308134 45891
## - office15              1   1273886 246376444 45893
## - hospital15            1   1690840 246793397 45900
## - subway10              1   2072407 247174965 45906
## - office10              1   3271064 248373622 45926
## - office05              1   3407505 248510063 45928
## - movie05               1   4453313 249555871 45946
## - bigMarket05           1   7876473 252979031 46003
## - kid15                 1  17889798 262992356 46165
## - year_of_completion_f  3  24621656 269724214 46266
## - transaction_month    10  26826302 271928860 46286
## - exclusive_use_area    1  48912840 294015398 46630
## - dong                 12 134196427 379298985 47669
##
## Step:  AIC=45872.81
## unit_price ~ dong + exclusive_use_area + floor + bigMarket05 +
##     bigMarket15 + school05 + school10 + school15 + subway05 +
##     subway10 + hospital05 + hospital10 + hospital15 + movie05 +
##     movie15 + kid15 + office05 + office10 + office15 + transaction_month +
##     year_of_completion_f
##
##                        Df Sum of Sq       RSS   AIC
## <none>                              245212643 45873
## + bigMarket10           1    110085 245102558 45873
## - school05              1    140560 245353203 45873
## + kid05                 1     41441 245171202 45874
## + movie10               1     38158 245174485 45874
## + kid10                 1     21896 245190747 45874
## + subway15              1     16194 245196449 45875
## - hospital05            1    260012 245472655 45875
## - bigMarket15           1    539190 245751833 45880
## - school15              1    994994 246207637 45888
## - school10              1    997479 246210123 45888
## - floor                 1   1074763 246287406 45889
## - hospital10            1   1106528 246319171 45890
## - movie15               1   1151984 246364627 45890
## - office15              1   1191471 246404114 45891
## - subway05              1   1508537 246721180 45896
## - hospital15            1   1697419 246910062 45900
## - subway10              1   2026599 247239242 45905
## - office10              1   3198871 248411515 45925
## - office05              1   3321465 248534108 45927
## - movie05               1   4670017 249882660 45949
## - bigMarket05           1   8285886 253498529 46009
## - kid15                 1  21220035 266432678 46217
## - year_of_completion_f  3  24780877 269993520 46268
## - transaction_month    10  26917258 272129901 46287
## - exclusive_use_area    1  49027616 294240260 46631
## - dong                 12 135567060 380779703 47684
```

```
##
## Call:
## lm(formula = unit_price ~ dong + exclusive_use_area + floor +
##     bigMarket05 + bigMarket15 + school05 + school10 + school15 +
##     subway05 + subway10 + hospital05 + hospital10 + hospital15 +
##     movie05 + movie15 + kid15 + office05 + office10 + office15 +
##     transaction_month + year_of_completion_f, data = data_train1)
##
## Coefficients:
##             (Intercept)              dong논현동              dong대치동
##               2902.0282                103.9299                 61.3920
##              dong도곡동              dong삼성동              dong세곡동
##                -108.6234               -112.8025              -1430.7397
##              dong수서동              dong신사동             dong압구정동
##                -642.5247               -304.3674                 84.7678
##              dong역삼동              dong일원동              dong자곡동
##                  -3.5387               -626.8831              -1001.2334
##              dong청담동       exclusive_use_area                   floor
##                -213.7654                 -3.5819                  2.6424
##             bigMarket05             bigMarket15                 school05
##                  55.8875                -10.7649                 -5.6361
##                school10                school15                subway05
##                   9.3024                  7.3751                 31.5533
##                subway10             hospital05              hospital10
##                -30.8189                  0.3147                 -0.4240
##              hospital15                movie05                 movie15
##                  -0.3460                 -9.1449                  1.3476
##                   kid15                office05                office10
##                -14.7533                -38.3938                -25.2792
##                office15     transaction_month02     transaction_month03
##                -10.2055                  0.1388                 42.1739
##     transaction_month04     transaction_month05     transaction_month06
##                  42.9966                 82.4064                115.2157
##     transaction_month07     transaction_month08     transaction_month09
##                 152.1120                191.1609                187.9439
##     transaction_month10     transaction_month11 year_of_completion_f2nd
##                 213.3657                285.1760               -194.4799
## year_of_completion_f3rd year_of_completion_f4th
##                  67.2405                213.3056
```

Linear regression prediction & RMSE calculation

```
linear_best<-lm(formula = unit_price ~ dong + floor + bigMarket05 + bigMarket15 +
    school05 + school10 + school15 + subway05 + subway10 + subway15 +
    hospital05 + hospital15 + movie05 + kid05 + kid10 + office10 +
    transaction_month + year_of_completion_f, data = data_train1)

# test data 에 적용
predict_2 <- predict(linear_best, data_test1[,-25])
summary(predict_2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   584.9  1210.3  1443.6  1483.3  1756.2  2446.7
```

```
data_test1 %>% select(dong) %>% unique()
```

```
##           dong
## 71427   역삼동
## 71457   개포동
## 71492   청담동
## 71509   삼성동
## 71522   대치동
## 71562   신사동
## 71570   논현동
## 71575 압구정동
## 71592   일원동
## 71599   수서동
## 71606   도곡동
## 71899   세곡동
## 72364   자곡동
```

```
data_train1 %>% select(dong) %>% unique()
```

```
##           dong
## 71426   역삼동
## 71458   개포동
## 71495   청담동
## 71503   삼성동
## 71520   대치동
## 71565   신사동
## 71566   논현동
## 71574 압구정동
## 71583   세곡동
## 71586   자곡동
## 71588   일원동
## 71600   수서동
## 71605   도곡동
```

```
# actual, predicted cbind

databind2 <- cbind(data_test1[,25],predict_2)
#databind2 <- cbind(data_test1[,28],predict_2)
databind2 <- as.data.frame(databind2)
summary(databind2)
```

```
##        V1         predict_2
##  Min.   : 494   Min.   : 584.9
##  1st Qu.:1144   1st Qu.:1210.3
##  Median :1444   Median :1443.6
##  Mean   :1487   Mean   :1483.3
##  3rd Qu.:1713   3rd Qu.:1756.2
##  Max.   :3507   Max.   :2446.7
```

```
# RMSE 계산
# install.packages("Metrics")
library(Metrics)
rmse(databind2$V1, databind2$predict_2)
```

```
## [1] 286.5622
```

Random Forest

```
install.packages("randomForest", repos ="http://cran.us.r-project.org")
```

```
## 패키지 'randomForest'를 성공적으로 압축해제하였고 MD5 sums 이 확인되었습니다
##
## 다운로드된 바이너리 패키지들은 다음의 위치에 있습니다
##  C:₩Users₩LUIS₩AppData₩Local₩Temp₩RtmpwHb08e₩downloaded_packages
```

```
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## 다음의 패키지를 부착합니다: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
rf.tree1 <- randomForest(unit_price~.-year, data = data_train1,
                         importance = TRUE,
                         ntree = 1000,mtry = 2)

# tree 결과
print(rf.tree1)
```

```
##
## Call:
##  randomForest(formula = unit_price ~ . - year, data = data_train1,      importance = TRUE, n
tree = 1000, mtry = 2)
##                Type of random forest: regression
##                      Number of trees: 1000
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 18645.96
##                     % Var explained: 90.95
```

```
summary(rf.tree1)
```

```
##                 Length Class  Mode
## call               6   -none- call
## type               1   -none- character
## predicted       4169   -none- numeric
## mse             1000   -none- numeric
## rsq             1000   -none- numeric
## oob.times       4169   -none- numeric
## importance        52   -none- numeric
## importanceSD      26   -none- numeric
## localImportance    0   -none- NULL
## proximity          0   -none- NULL
## ntree              1   -none- numeric
## mtry               1   -none- numeric
## forest            11   -none- list
## coefs              0   -none- NULL
## y               4169   -none- numeric
## test               0   -none- NULL
## inbag              0   -none- NULL
## terms              3   terms  call
```

```
install.packages("rpart.plot", repos ="http://cran.us.r-project.org")
```

```
## Warning: 패키지 'rpart.plot'가 사용중이므로 설치되지 않을 것입니다
```

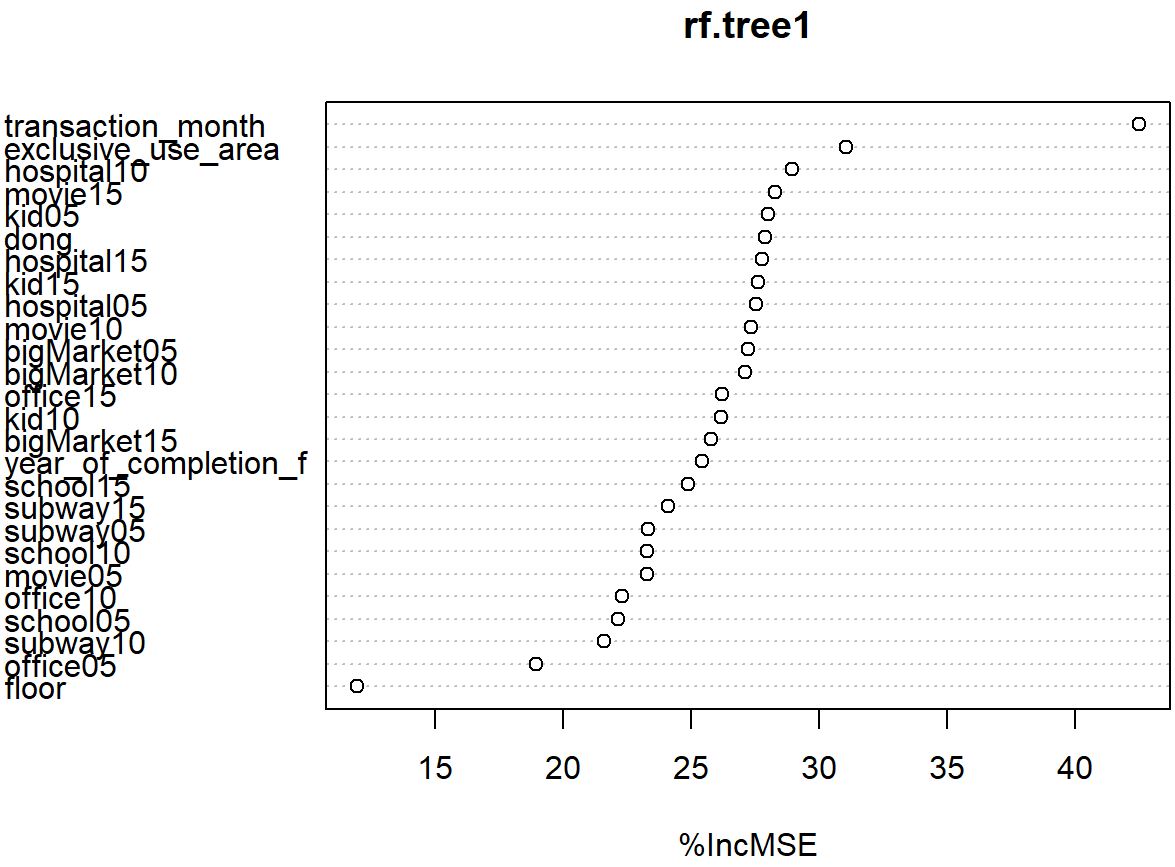
```
library(rpart.plot)

importance(rf.tree1)
```

```
##                       %IncMSE IncNodePurity
## dong                 27.90146      70468438
## exclusive_use_area   31.04547      47941072
## floor                11.97077       9801289
## bigMarket05          27.24241      14705736
## bigMarket10          27.11091      24840453
## bigMarket15          25.79269      27546246
## school05             22.15367      14149513
## school10             23.29017      37372734
## school15             24.86761      31139723
## subway05             23.31064      10333389
## subway10             21.59472      20890408
## subway15             24.12084      25926942
## hospital05           27.53640      31718268
## hospital10           28.94600      41864182
## hospital15           27.77208      34473116
## movie05              23.27673      20417705
## movie10              27.33835      31386661
## movie15              28.28961      36470347
## kid05                28.02247      40025699
## kid10                26.19307      49676982
## kid15                27.62723      46552539
## office05             18.95293      15952512
## office10             22.32051      30004971
## office15             26.22261      37004144
## transaction_month    42.48839      12535510
## year_of_completion_f 25.42314      25041642
```

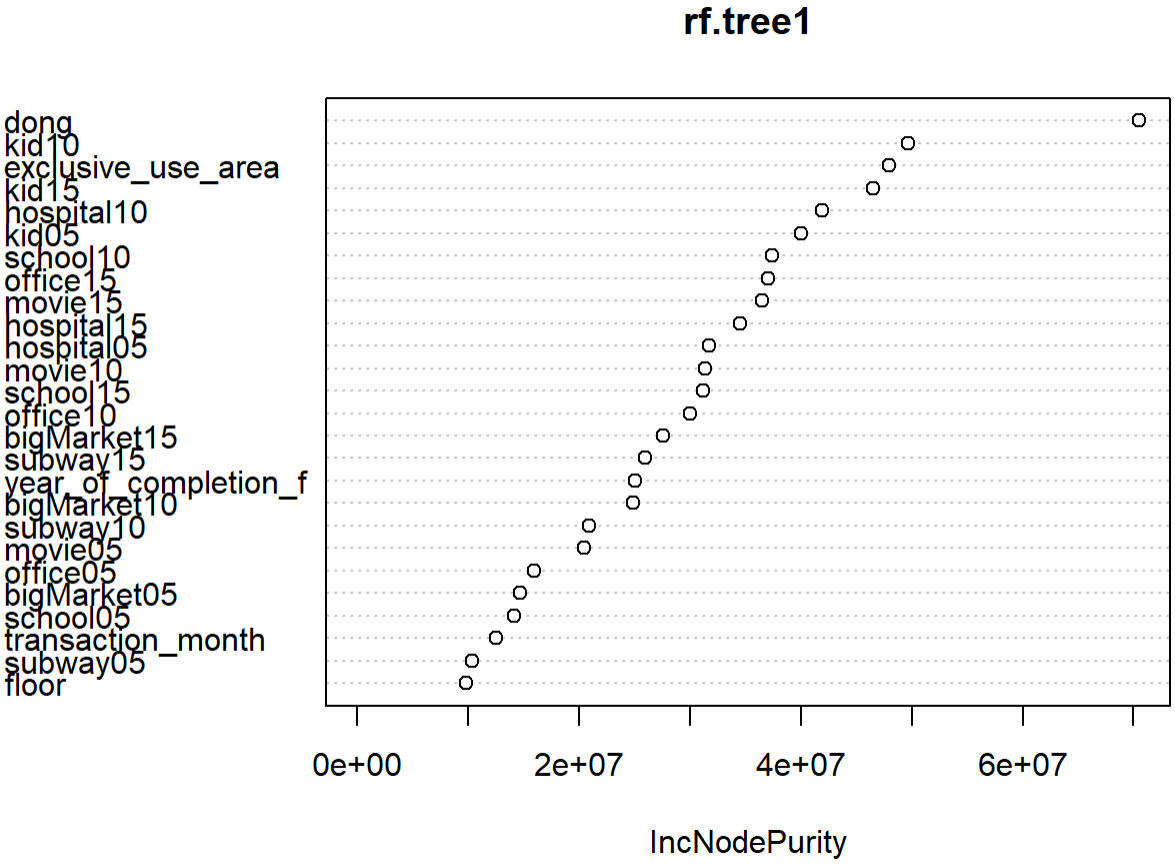
```
importance(rf.tree1, type = 1)
```

```
##                       %IncMSE
## dong                 27.90146
## exclusive_use_area   31.04547
## floor                11.97077
## bigMarket05          27.24241
## bigMarket10          27.11091
## bigMarket15          25.79269
## school05             22.15367
## school10             23.29017
## school15             24.86761
## subway05             23.31064
## subway10             21.59472
## subway15             24.12084
## hospital05           27.53640
## hospital10           28.94600
## hospital15           27.77208
## movie05              23.27673
## movie10              27.33835
## movie15              28.28961
## kid05                28.02247
## kid10                26.19307
## kid15                27.62723
## office05             18.95293
## office10             22.32051
## office15             26.22261
## transaction_month    42.48839
## year_of_completion_f 25.42314
```

```
varImpPlot(rf.tree1, type = 1)
```




```
varImpPlot(rf.tree1, type = 2)
```

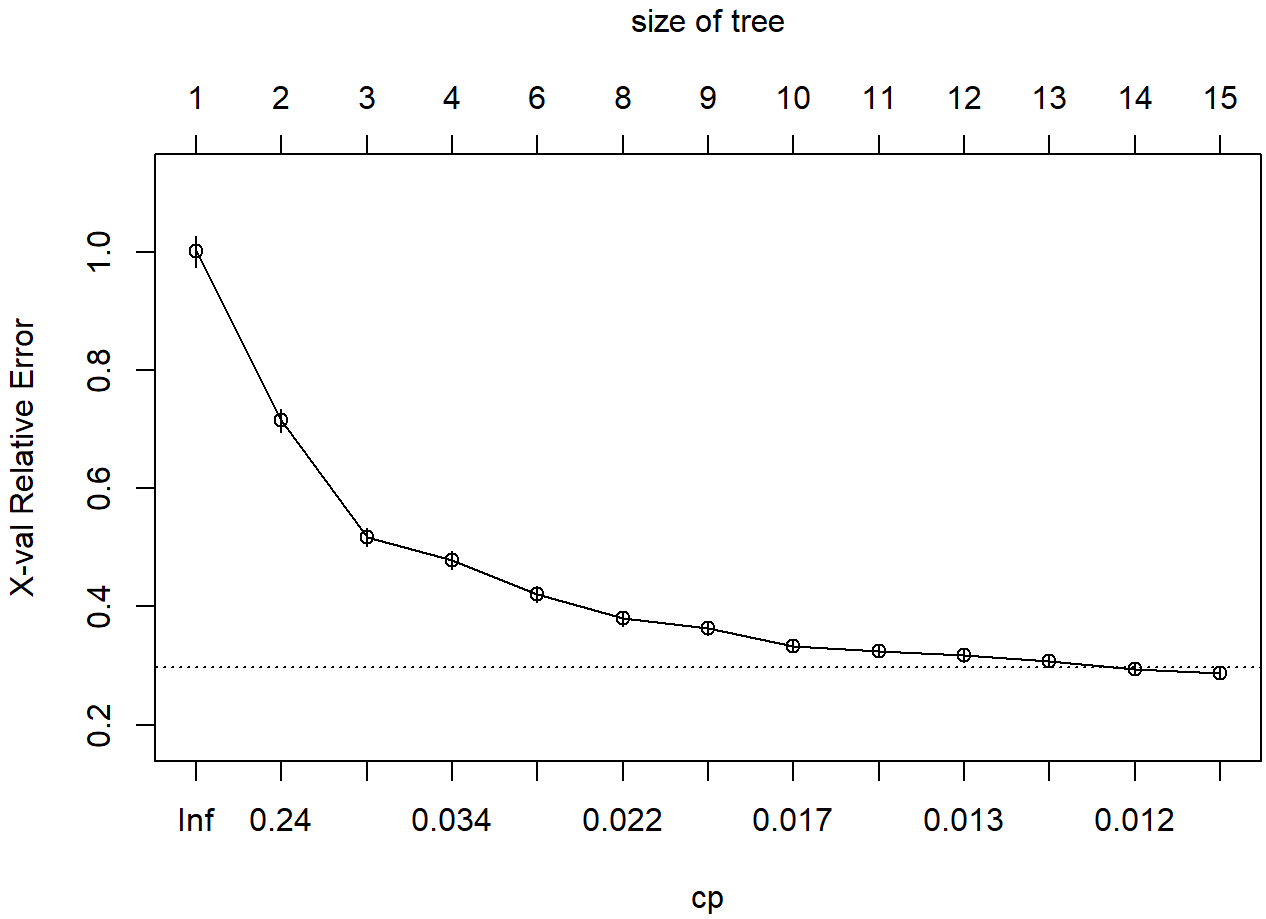


Random Forest parameter tuning

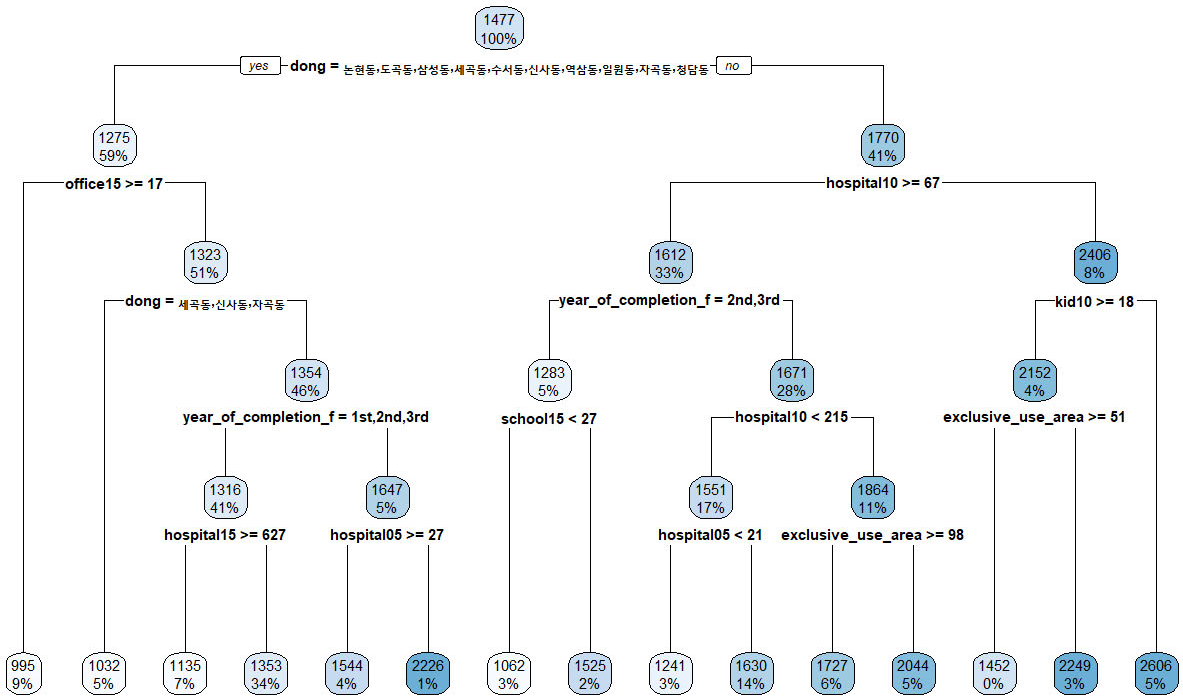
```
printcp(tree1)
```

```
##
## Regression tree:
## rpart(formula = unit_price ~ . - year, data = data_train1, method = "anova",
##     control = rpart.control(minsplit = 50, maxdepth = 5))
##
## Variables actually used in tree construction:
## [1] dong                 exclusive_use_area   hospital05
## [4] hospital10           hospital15           kid10
## [7] office15             school15             year_of_completion_f
##
## Root node error: 858944768/4169 = 206031
##
## n= 4169
##
##          CP nsplit rel error  xerror      xstd
## 1  0.286182      0   1.00000 1.00048 0.0265933
## 2  0.197755      1   0.71382 0.71453 0.0197519
## 3  0.038384      2   0.51606 0.51705 0.0156341
## 4  0.030782      3   0.47768 0.47880 0.0150522
## 5  0.023316      5   0.41612 0.42064 0.0140338
## 6  0.020425      7   0.36948 0.37964 0.0126584
## 7  0.019870      8   0.34906 0.36370 0.0121385
## 8  0.015159      9   0.32919 0.33368 0.0105943
## 9  0.013369     10   0.31403 0.32509 0.0104158
## 10 0.012789     11   0.30066 0.31695 0.0103553
## 11 0.012720     12   0.28787 0.30828 0.0102184
## 12 0.011717     13   0.27515 0.29410 0.0100766
## 13 0.010000     14   0.26344 0.28706 0.0099764
```

```
plotcp(tree1)
```




```
tree1 <- prune(tree1, cp= tree1$cptable[which.min(tree1$cptable[, "xerror"]), "CP"])

rpart.plot(tree1)
```



Random Forest prediction & RMSE calculation

```
# test data 에 적용

predict_3 <- predict(rf.tree1, data_test1)
summary(predict_3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   676.4  1171.8  1460.9  1493.3  1659.7  2674.2
```

```
# actual, predicted cbind

databind3 <- cbind(data_test1[,25],predict_3)
databind3 <- as.data.frame(databind3)
summary(databind3)
```

```
##        V1         predict_3
##  Min.   : 494   Min.   : 676.4
##  1st Qu.:1144   1st Qu.:1171.8
##  Median :1444   Median :1460.9
##  Mean   :1487   Mean   :1493.3
##  3rd Qu.:1713   3rd Qu.:1659.7
##  Max.   :3507   Max.   :2674.2
```

```
# RMSE 계산
install.packages("Metrics", repos = "http://cran.us.r-project.org")
```

```
## Warning: 패키지 'Metrics'가 사용중이므로 설치되지 않을 것입니다
```

```
library(Metrics)
rmse(databind3$V1, databind3$predict_3)
```

```
## [1] 137.4991
```

Gradient Boost Model

```
install.packages("gbm", repos ="http://cran.us.r-project.org")
```

```
## 패키지 'gbm'를 성공적으로 압축해제하였고 MD5 sums 이 확인되었습니다
##
## 다운로드된 바이너리 패키지들은 다음의 위치에 있습니다
##  C:₩Users₩LUIS₩AppData₩Local₩Temp₩RtmpwHb08e₩downloaded_packages
```

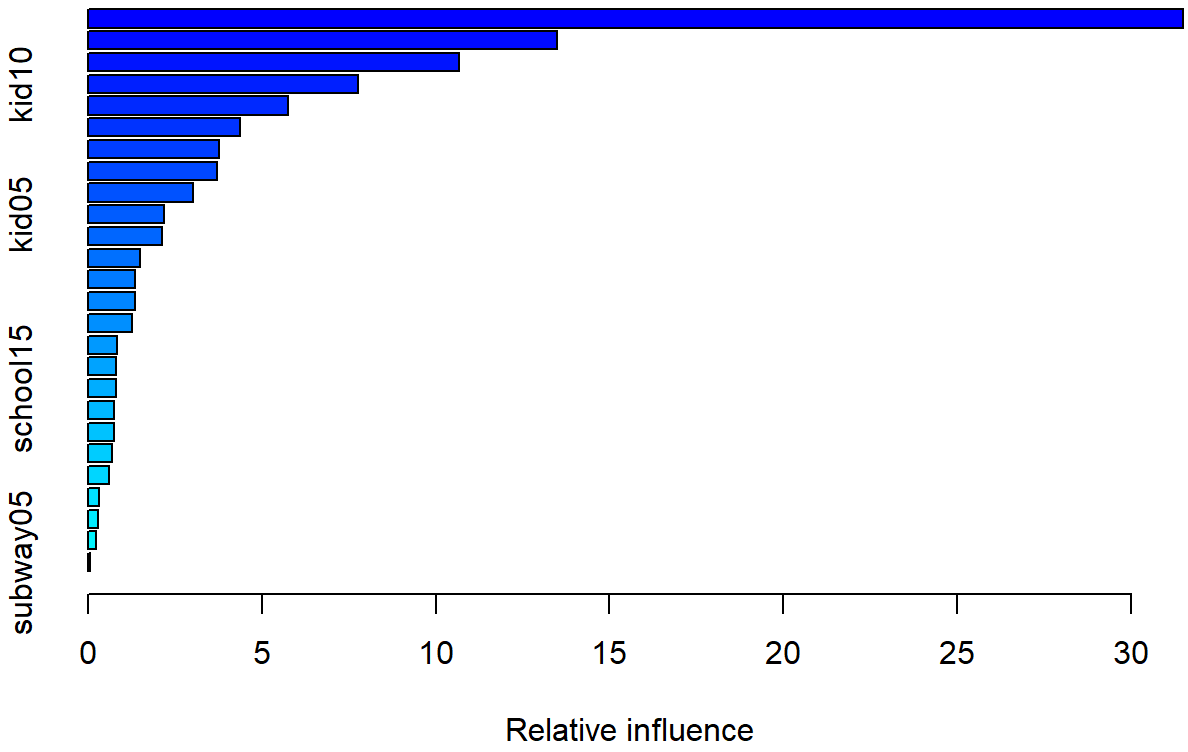
```
library(gbm)
```

```
## Loaded gbm 2.1.8.1
```

```
gbm.tree1 <- gbm(unit_price~.-year, data = data_train1, distribution = "gaussian",
                 n.trees = 1000, shrinkage = 0.01, interaction.depth = 4)

# tree 결과
print(gbm.tree1)
```

```
## gbm(formula = unit_price ~ . - year, distribution = "gaussian",
##     data = data_train1, n.trees = 1000, interaction.depth = 4,
##     shrinkage = 0.01)
## A gradient boosted model with gaussian loss function.
## 1000 iterations were performed.
## There were 26 predictors of which 26 had non-zero influence.
```

```
summary(gbm.tree1)
```




```
##                                    var     rel.inf
## dong                              dong 31.49094750
## hospital10                  hospital10 13.49276610
## exclusive_use_area  exclusive_use_area 10.68448291
## kid10                            kid10  7.77625220
## office15                      office15  5.76611633
## year_of_completion_f year_of_completion_f  4.38546507
## transaction_month    transaction_month  3.75788670
## school10                      school10  3.71278715
## kid15                            kid15  3.03143285
## kid05                            kid05  2.17658202
## hospital15                  hospital15  2.14177534
## subway15                      subway15  1.49356228
## office10                      office10  1.36779334
## movie05                        movie05  1.34411758
## floor                            floor  1.28131380
## bigMarket05                bigMarket05  0.82548049
## subway10                      subway10  0.80569444
## school15                      school15  0.80212806
## bigMarket15                bigMarket15  0.75259030
## movie15                        movie15  0.73661903
## movie10                        movie10  0.68915926
## bigMarket10                bigMarket10  0.61310756
## hospital05                  hospital05  0.30580755
## school05                      school05  0.27836123
## office05                      office05  0.21950881
## subway05                      subway05  0.06826212
```

Gradient Boost Model parameter tuning

```
# printcp(tree1)
# plotcp(tree1)
# tree1 <- prune(tree1, cp= tree1$cptable[which.min(tree1$cptable[, "xerror"]), "CP"])
#
# rpart.plot(tree1)
```

##Gradient Boost Model prediction & RMSE calculation

```
# test data 에 적용
predict_4 <- predict.gbm(object = gbm.tree1,
                             newdata = data_test1,
                             n.trees = 1000,
                             type = "response")
summary(predict_4)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     703    1189    1438    1490    1669    2796
```

```
# actual, predicted cbind

databind4 <- cbind(data_test1[,25],predict_4)
databind4 <- as.data.frame(databind4)
summary(databind4)
```

```
##        V1         predict_4
##  Min.   : 494   Min.   : 703
##  1st Qu.:1144   1st Qu.:1189
##  Median :1444   Median :1438
##  Mean   :1487   Mean   :1490
##  3rd Qu.:1713   3rd Qu.:1669
##  Max.   :3507   Max.   :2796
```

```
# RMSE 계산
install.packages("Metrics", repos ="http://cran.us.r-project.org")
```

```
## Warning: 패키지 'Metrics'가 사용중이므로 설치되지 않을 것입니다
```

```
library(Metrics)
rmse(databind4$V1, databind4$predict_4)
```

```
## [1] 147.8877
```