Apartment Data Visualization

Khanin Sisaengsuwanchai

2022-07-02

```
# Download the data
data = read.csv("VIT.csv")
dim(data)
```

```
## [1] 218  15
```

```
colnames(data)
```

```
##  [1] "price"        "area"         "zone"         "category"     "age"
##  [6] "floor"        "rooms"        "out"          "conservation" "toilets"
## [11] "garage"       "elevator"     "street"       "heating"      "storage"
```

```
## Convert the types of data
data$floor = as.factor(data$floor)
data$rooms = as.factor(data$rooms)
data$garage = as.factor(data$garage)
data$toilets = as.factor(data$toilets)
data$elevator = as.factor(data$elevator)
data$storage = as.factor(data$storage)
data$street = as.factor(data$street)

summary(data)
```

```
##      price             area            zone             category
##  Min.   :155000   Min.   : 50.38   Length:218         Length:218
##  1st Qu.:228500   1st Qu.: 75.18   Class :character   Class :character
##  Median :269750   Median : 86.39   Mode  :character   Mode  :character
##  Mean   :280737   Mean   : 88.70
##  3rd Qu.:328625   3rd Qu.: 99.90
##  Max.   :560000   Max.   :187.91
##
##       age            floor    rooms       out            conservation
##  Min.   :  1.0   3      :48   3:  3   Length:218         Length:218
##  1st Qu.: 11.0   5      :43   4: 51   Class :character   Class :character
##  Median : 16.0   4      :41   5:141   Mode  :character   Mode  :character
##  Mean   : 19.9   2      :35   6: 21
##  3rd Qu.: 24.0   6      :15   7:  2
##  Max.   :118.0   8      :13
##                  (Other):23
##  toilets garage  elevator street     heating          storage
```

```
##  1:116   0:167   0: 44    S2: 42   Length:218         0: 43
##  2:102   1: 49   1:174    S3:107   Class :character   1:174
##          2:  2            S4: 59   Mode  :character   2:  1
##                           S5: 10
##
##
##
```

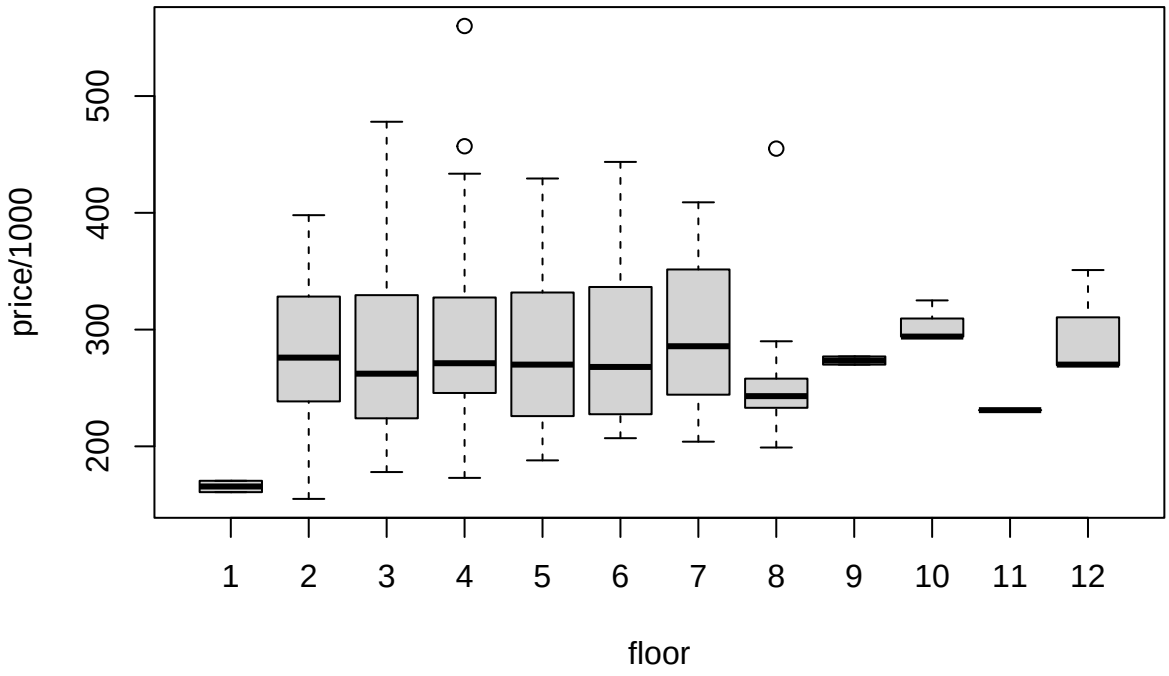
```
# 1. Find the number of apartments by the number of garages
table(data$garage)
```

```
##
##   0   1   2
## 167  49   2
```

```
# Note: The first row is the number of garages, and the second row is its
# corresponding number of apartments.
```

```
# 2. Create a single chart showing boxplots for the apartment's price for each floor.

plot(price/1000 ~ floor, data)
```




```
## For the purpose of visualization, I divide the y-axis by 1000.
```

```r
# 3. Create a crosstab table for the number of apartments by number of rooms and the number of garages.
tab = table(Rooms = data$rooms, Garages = data$garage)
tab_df = data.frame(tab)
tab_df
```

```
##    Rooms Garages Freq
## 1      3       0    3
## 2      4       0   46
## 3      5       0  104
## 4      6       0   13
## 5      7       0    1
## 6      3       1    0
## 7      4       1    5
## 8      5       1   35
## 9      6       1    8
## 10     7       1    1
## 11     3       2    0
## 12     4       2    0
## 13     5       2    2
## 14     6       2    0
## 15     7       2    0
```

```r
# 4. Find the average price (up to 2 sig. digits) of the apartments for each category from question 3.
aux = list(data$rooms, data$garage)
res = tapply(data$price, aux, mean)
res[is.na(res)] = 0 ## Fill NA with 0
res
```

```
##          0        1      2
## 3 230333.3      0.0      0
## 4 229267.0 279200.0      0
## 5 261163.2 344097.1 369250
## 6 358992.3 403500.0      0
## 7 443600.0 286000.0      0
```

```r
## Note:
# 1. Rows are # of rooms and columns are # of garages.
```

```r
# 5. Find the min and max price of apartments with area between (including) 80 and 90 square meters.
sel_aptmt = data[data$area >= 80 & data$area <= 90, ]$price
min(sel_aptmt, na.rm = TRUE) # Minimal price
```

```
## [1] 168000
```

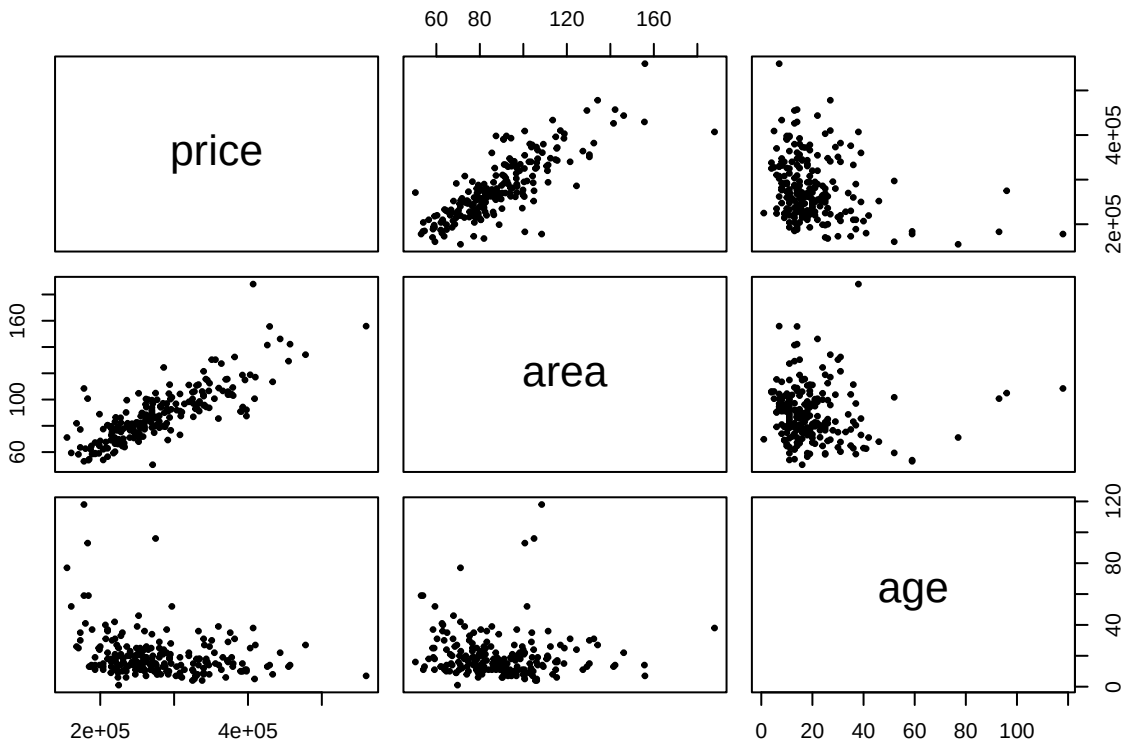
```r
max(sel_aptmt, na.rm = TRUE) # Maximal price
```

```
## [1] 398000
```

```r
# 6. What numeric variable is most correlated with price? Draw a scatterplot for these variables

# Select only numerical data to plot correlation
num_data = data[, sapply(data, is.numeric)]

pairs(~. ,num_data, pch=19, cex=0.5)
```




```r
## According to the scatterplot, area is the most correlated variable with price.
## Age also seems to have non-linear relationship with price
```

```r
# 7. On average, how much more expensive is an apartment in street type S4 than type S2?
avg_pri_s2 = mean(data[data$street == 'S2',]$price) # Price of apartments in S2
avg_pri_s4 = mean(data[data$street == 'S4',]$price) # Price of apartments in S4

avg_pri_s4 - avg_pri_s2
```

```
## [1] 10673.02
```

```r
# An apartment price, on average, in street type S4 is 10673.02 higher than type S2.
```

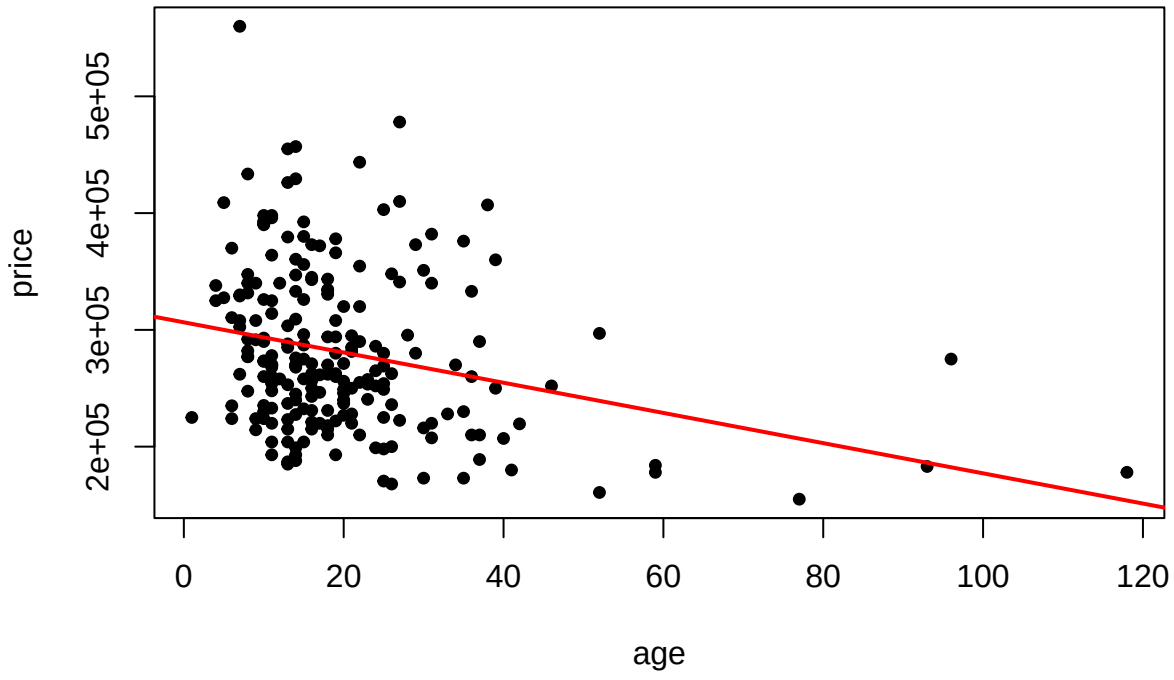
```r
# 8. Create a scatterplot of y:price against x:age. Fit a regression line to display the average price
plot(price ~ age, data, pch=19, cex=0.75)

# Fit the regression line of y:price against x:age to understand their relationship
```

```r
m1 = lm(price ~ age, data)
summary(m1)
```

```
##
## Call:
## lm(formula = price ~ age, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -104846  -48946  -18367   39277  262579
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 306475.2     7665.3  39.982  < 2e-16 ***
## age          -1293.4      310.9  -4.161 4.58e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66840 on 216 degrees of freedom
## Multiple R-squared:  0.0742, Adjusted R-squared:  0.06992
## F-statistic: 17.31 on 1 and 216 DF,  p-value: 4.578e-05
```

```r
# Plot the regression line
abline(m1, col="red", lwd=2)
```



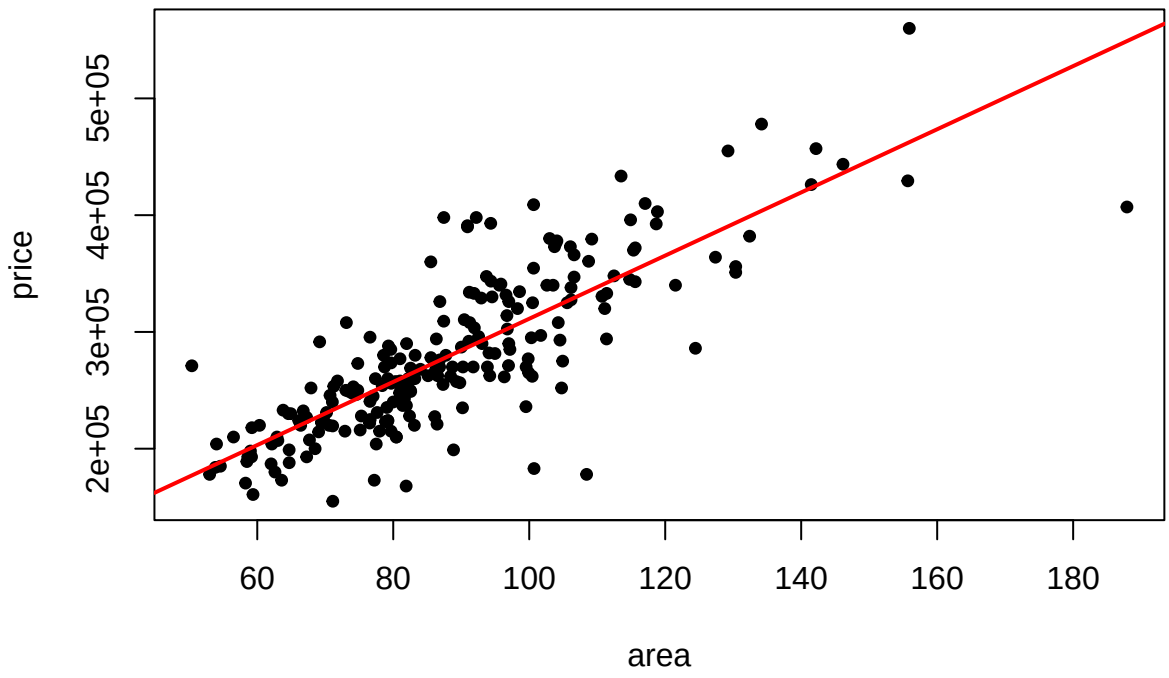
```r
# For each additional year of age, the average price decreases about 1293.4 Euros.
```

```r
# 9. The realtor thinks that in Vitoria, each square meter will cost roughly 2500 e. Create a scatterpl
plot(price ~ area, data, pch=19, cex=0.75)

# Fit the regression line of y:price against x:area to understand their relationship
m2 = lm(price ~ area, data)
summary(m2)
```

```
##
## Call:
## lm(formula = price ~ area, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -156121  -21559   -2150   19497  120679
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40818.0    12172.9   3.353 0.000943 ***
## area          2704.7      133.6  20.238  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40820 on 216 degrees of freedom
## Multiple R-squared:  0.6547, Adjusted R-squared:  0.6531
## F-statistic: 409.6 on 1 and 216 DF,  p-value: < 2.2e-16
```

```r
# Plot the regression line
abline(m2, col="red", lwd=2)
```




```
# For each additional square meter, the average price increases about 2704.7 Euros,
# which is quite similar to what the realtor estimates.
```