# Real Estate Analysis

Onni Vitikainen

2023-12-23

## Real Estate investing

### Goal of the analysis

The purpose of this analysis is to find out what kind of real estate property would be the most optimal investment for a working class citizen.

We are focusing on properties in Tampere.

I will be using a fixed interest rate of 4% and loan time 25 years.

### Data collection

In order to get relevant data in current market I decided to web scrape the rental and property sale websites of Tampere.

I created my own Python scripts to fit my needs and get the relevant data into a csv file.

**The data collected about properties on sale** I scraped the information from https://www.etuovi.com/myytavat-asunnot/tampere ##### Data scraped: **- Type: - Rooms: - Address: - Price: - Size: - Year:**

**The data collected about rental properties** I scraped the information from https://www.vuokraovi.com/vuokra-asunnot/Tampere ##### Data scraped **- Type: - Rooms: - Address: - Price: - Size:**

```r
library(tidyverse)
library(dplyr)
library(readr)
library(ggplot2)
```

### Installing needed packages

### Loading datasets into workspace

```r
sale_apartments <- read.csv("sale_apartments.csv")
rent_apartments <- read.csv("rent_apartments.csv")
```

```r
head(sale_apartments)
```

```
##   house_type                                    rooms
## 1 Kerrostalo                                  2h + kk
## 2 Kerrostalo                                   2h + k
## 3 Kerrostalo                                   1h + k
## 4 Kerrostalo 3h + kt + ph / wc + ranskalainen parveke
## 5 Kerrostalo                                   2h + k
## 6 Kerrostalo                              3h + kt + s
##                                     address     price    size year
## 1           Töyrytie 1-5, Ruotula, Tampere 139 000 €   39 m² 1962
## 2          Huikkaanaukio 3, Hukas, Tampere 107 000 €   55 m² 1960
## 3  Kullervonkatu 12 B 44, Tammela, Tampere 192 100 € 26,5 m² 2023
## 4 Tuomiokirkonkatu 15 D, Keskusta, Tampere 399 000 € 69,5 m² 1957
## 5  Kullervonkatu 12 B 43, Tammela, Tampere 284 500 €   48 m² 2023
## 6  Kullervonkatu 12 A 39, Tammela, Tampere 468 800 €   79 m² 2023
```

```r
head(rent_apartments)
```

```
##   house_type                    rooms
## 1 kerrostalo            1h + avok + p
## 2 kerrostalo                    2H+KT
## 3 kerrostalo                     2h+k
## 4 kerrostalo                    2h+kk
## 5 kerrostalo                   1h + k
## 6 kerrostalo 1h+kt+alk+ransk. parveke
##                                         address    price    size
## 1          Perkkoonkatu 1, Multisilta, Tampere 530 €/kk   23 m²
## 2           Rantatie 27 A, Santalahti, Tampere 970 €/kk   42 m²
## 3             Sammonkatu 43 B, Kaleva, Tampere 850 €/kk   60 m²
## 4       Tesomajärvenkatu 10 B, Tesoma, Tampere 675 €/kk 51,4 m²
## 5           Iidesranta 19, Iidesranta, Tampere 580 €/kk 32,5 m²
## 6 Konttilukinkatu 13 A, Härmälänranta, Tampere 635 €/kk   31 m²
```

### Pre processing data

Firstly we will check for duplicate apartments. We will check that no address is the same. There can be same address data because the apartment can be listed to multiple real estate agencies.

We will start by creating copies of the data set in order not to change actual data.

```r
rent_properties <- rent_apartments
sale_properties <- sale_apartments
```

Delete duplicate addresses and duplicates

```r
sale_properties <- unique(sale_properties)
sale_properties <- sale_properties[!duplicated(sale_properties$address),]

rent_properties <- unique(rent_properties)
rent_properties <- rent_properties[!duplicated(rent_properties$address),]
```

Deleted 509 rent duplicates and 531 sale duplicates.

Next we will transform the price and size columns in to numerical, we will add the value to the header.

```
colnames(sale_properties)[colnames(sale_properties) == "price"] <- "price €"
colnames(sale_properties)[colnames(sale_properties) == "size"] <- "size m²"

colnames(rent_properties)[colnames(rent_properties) == "price"] <- "price €/m"
colnames(rent_properties)[colnames(rent_properties) == "size"] <- "size m²"
```

Next we change numerical values to numerical values so price, size, year. Decimals are separated by commas need to swap them into dots and remove extra spacing.

Formatting pricing.

```
sale_properties$`price €` <- gsub("[€[:space:]]", "", sale_properties$`price €`)
sale_properties$`price €` <- gsub(",", ".", sale_properties$`price €`)
sale_properties$`price €` <- as.numeric(gsub(",", ".",
                gsub("[^0-9.]", "", gsub(" ", "", sale_properties$`price €`))))

rent_properties$`price €/m` <- gsub("[€[:space:]]", "", rent_properties$`price €/m`)
rent_properties$`price €/m` <- gsub(",", ".", rent_properties$`price €/m`)
rent_properties$`price €/m`<- as.numeric(gsub(",", ".",
                gsub("[^0-9.]", "", gsub(" ", "", rent_properties$`price €/m`))))
```

Formatting size. We have apartments with 146.8/159.1 format sizing. In our analysis we will take into account the bigger value, because this indicates the overall size of the property.

```
sale_properties$`size m²` <- gsub(",", ".", sale_properties$`size m²`)
sale_properties$`size m²` <- gsub("[m²[:space:]]", "", sale_properties$`size m²`)

rent_properties$`size m²` <- gsub(",", ".", rent_properties$`size m²`)
rent_properties$`size m²` <- gsub("[m²[:space:]]", "", rent_properties$`size m²`)
```

Function to convert all sizes containing a "/" and choosing the larger value

```
convert_sizes <- function(size_string) {
  if (grepl('/', size_string, fixed = TRUE)) {
    sizes <- strsplit(size_string, "/")[[1]]
    return(max(sizes))
  } else {
    return(size_string)
  }
}
```

```
sale_properties$`size m²` <- sapply(sale_properties$`size m²`, convert_sizes)
rent_properties$`size m²` <- sapply(rent_properties$`size m²`, convert_sizes)
```

```
sale_properties$`size m²`<- as.numeric(gsub(",", ".",
                gsub("[^0-9.]", "", gsub(" ", "", sale_properties$`size m²`))))

rent_properties$`size m²`<- as.numeric(gsub(",", ".",
                gsub("[^0-9.]", "", gsub(" ", "", rent_properties$`size m²`))))
```

Next we will round the size to a integer. We will do the rounding in order to group the apartments and get a better average for the price.

```
sale_properties$`size m²` <- round(sale_properties$`size m²`)
rent_properties$`size m²` <- round(rent_properties$`size m²`)
```

Next we will change the sale properties build year into numeric.

```
sale_properties$`year`<- as.numeric(gsub(",", ".", gsub("[^0-9.]", "",
                         gsub(" ", "", sale_properties$`year`))))
```

Next step of data cleaning we will check if the columns contain off values. We will do this just by sorting the table and seeing if we see max and min values that are off.

First the max and min of the sale properties

```
max_sale_price <- max(sale_properties$`price €`)
max_sale_price
```

```
## [1] 2369230
```

```
min_sale_price <- min(sale_properties$`price €`)
min_sale_price
```

```
## [1] 0
```

We got a 0 for the minumum price, something is wrong. When researching the apartment in the website turns out this is a auction property, we will remove it from the analysis.

```
sale_properties <- sale_properties[sale_properties$`price €`!= 0, ]
```

```
min_sale_price <- min(sale_properties$`price €`)
min_sale_price
```

```
## [1] 10244.62
```

Our new value is legitimate.

Next we test the rent prices.

```
max_rent_price <- max(rent_properties$`price €/m`)
max_rent_price
```

```
## [1] 4200
```

```
min_rent_price <- min(rent_properties$`price €/m`)
min_rent_price
```

```
## [1] 120
```

There are some attic rooms on sale which explains the low min rent price.

Next we will check if the sizes have outliers.

```
max_rent_size <- max(rent_properties$`size m²`)
max_rent_size
```

```
## [1] 254
```

```
min_rent_size <- min(rent_properties$`size m²`)
min_rent_size
```

```
## [1] 8
```

The values are legitimate.

Checking sale sizes.

```
max_sale_size <- max(sale_properties$`size m²`)
max_sale_size
```

```
## [1] 738
```

```
min_sale_size <- min(sale_properties$`size m²`)
min_sale_size
```

```
## [1] 17
```

Also are legitimate values.

Some years are N/A, but year is not an important value in the analysis.

We noticed that one rental house size 210 has rent only 400€ with further investigation turns out they are only renting one room. We will remove this from the analysis.

```
rent_properties <- rent_properties[rent_properties$address
                                   != "Tiilikatu 34, Kissanmaa, Tampere", ]
```

Our data overall is clean now. Summaries below.

```
summary(sale_properties)
```

```
##   house_type           rooms              address             price €
##  Length:2468        Length:2468        Length:2468        Min.   :  10245
##  Class :character   Class :character   Class :character   1st Qu.: 157000
##  Mode  :character   Mode  :character   Mode  :character   Median : 219078
##                                                           Mean   : 262509
##                                                           3rd Qu.: 311500
##                                                           Max.   :2369230
##
##     size m²           year
##  Min.   : 17.00   Min.   :1897
##  1st Qu.: 40.00   1st Qu.:1978
##  Median : 56.00   Median :2018
##  Mean   : 66.80   Mean   :2001
##  3rd Qu.: 77.25   3rd Qu.:2023
##  Max.   :738.00   Max.   :2025
##                   NA's   :156
```

```
summary(rent_properties)
```

```
##   house_type           rooms             address            price €/m
##  Length:1290        Length:1290        Length:1290        Min.   : 120.0
##  Class :character   Class :character   Class :character   1st Qu.: 629.2
##  Mode  :character   Mode  :character   Mode  :character   Median : 742.5
##                                                           Mean   : 828.9
##                                                           3rd Qu.: 895.0
##                                                           Max.   :4200.0
##     size m²
##  Min.   :  8.00
##  1st Qu.: 32.00
##  Median : 45.00
##  Mean   : 49.12
##  3rd Qu.: 60.00
##  Max.   :254.00
```

### Analysis

We want to figure out what kind of apartments provide the highest profit margins.

We will add columns to the sales properties for the monthly loan payment.

As mentioned in the start we will first use fixed interest rate of 4% and loan time of 25 years.

We will be using annuity loan model. Creating a function to calculate the monthly payments.

```
yearly_interest <- 0.04
loan_time <- 25

loan_time_months <- loan_time * 12
monthly_interest <- yearly_interest / 12

monthly_payment <- function(loan_capital, loan_time_months, monthly_interest) {
  return(loan_capital * (monthly_interest *
                           (1 + monthly_interest)^loan_time_months) /
         ((1 + monthly_interest)^loan_time_months - 1))
}
```

```
sale_properties <- sale_properties %>% mutate("loan payment €/m"
                                               = monthly_payment(`price €`,
                                                                loan_time_months,
                                                                monthly_interest))
```

Next we want to determine the typical rent for a certain housing size and the typical area it is in.

We will make the area of Tampere the apartment is into its own column.

```
sale_properties <- sale_properties %>% mutate("area"
                               = str_extract(address, "(?<=,\\s).*?(?=,)"))

rent_properties <- rent_properties %>% mutate("area"
                               = str_extract(address, "(?<=,\\s).*?(?=,)"))
```

```
colSums(is.na(rent_properties))
```

```
## house_type      rooms    address  price €/m    size m²       area
##          0          0          0          0          0         20
```

```
colSums(is.na(sale_properties))
```

```
##       house_type            rooms          address          price €
##                0                0                0                0
##          size m²             year loan payment €/m             area
##                0              156                0                3
```

We have a few areas missing

```
rows_with_na <- subset(sale_properties, is.na(sale_properties$area))
```

We got three values we will manually enter the areas to these addresses by checking the area from maps.

```
sale_properties <- sale_properties %>%
  mutate(
    area = ifelse(grepl("Mannakorventie 17, Tampere", address), "Vaitiala",
          ifelse(grepl("Hanhenmäenkatu 1 C 109, Tampere", address), "Takahuhti",
          ifelse(grepl("Karosenkatu 3 C, Tampere", address), "Annala", area)))
  )
```

We also have a few differently formated values we will fix these also.

```
sale_properties <- sale_properties %>%
  mutate(
    area = ifelse(
      grepl("Yliopistonkatu 50-52, 33100 Tampere, Tammela, Tampere", address),
      "Tammela",
      ifelse(grepl("Tikkukuja 4B, 33250 Tampere, Santalahti, Tampere", address),
             "Santalahti",
        ifelse(grepl("Salhojankatu 3 A, 33500, Tampere", address), "Tammela",
          ifelse(grepl("Hipunraitti 5 A 13, 33580 Tampere, Linnainmaa, Tampere",
                       address), "Linnainmaa",
            ifelse(grepl("KALEVAN PUISTOTIE 26, A 23, Kaleva, TAMPERE", address),
                   "Kaleva", area)
          )
        )
      )
    )
  )
```

```
rows_with_na <- subset(rent_properties, is.na(rent_properties$area))
```

Now we have 20 values, we can still manually change them by finding the area.

```
rent_properties <- rent_properties %>%
  mutate(
    area = case_when(
      grepl("Hämeenpuisto 13, Tampere", address) ~ "Keskusta",
      grepl("Yliopistonkatu 45 K, Tampere", address) ~ "Tulli",
      grepl("Tesoma, Tampere", address) ~ "Tesoma",
      grepl("Amuri, Tampere", address) ~ "Amuri",
      grepl("Hippoksenkatu 11 F, Tampere", address) ~ "Kissanmaa",
      grepl("Vasamamittarinkatu 4, Tampere", address) ~ "Atala",
      grepl("Yliopistonkatu 45 E, Tampere", address) ~ "Tulli",
      grepl("Multiojankatu 32 B, Tampere", address) ~ "Multisilta",
      grepl("Keskusta, Tampere", address) ~ "Keskusta",
      grepl("Tikkukuja 1 A, Tampere", address) ~ "Pyynnikki",
      grepl("Possilankatu 45, Tampere", address) ~ "Lintulampi",
      grepl("Hämeenpuisto 42 A, Tampere", address) ~ "Keskusta",
      grepl("Koljonseläntie 7, Tampere", address) ~ "Maisansalo",
      grepl("Niemenranta, Tampere", address) ~ "Niemenranta",
      grepl("Rantatie 37 A, Tampere", address) ~ "Pispala",
      grepl("Koljonseläntie 7 as, Tampere", address) ~ "Maisansalo",
      grepl("Vasamamittarinkatu 4 B, Tampere", address) ~ "Atala",
      grepl("Hämeenkatu 30 B, Tampere", address) ~ "Keskusta",
      grepl("Sarvijaakonkatu 24, Tampere", address) ~ "Kaleva",
      grepl("Kansikatu 8, Tampere", address) ~ "Keskusta",
      TRUE ~ area
    )
  )
```

Next we will calculate the average rent price per m2 in a certain area, in order to calculate what could be the rent for properties on sale.

First we have to add the rent price per m2

```
rent_properties <- rent_properties %>% mutate("rent per m2 / €"
                                              = `price €/m` / `size m²`)
```

Now we will make a new data frame that has the area and average price per m2 in that area.

```
area_m2_rent <- rent_properties %>%
  group_by(area) %>%
  summarize('€ per m2' = mean(`rent per m2 / €`, na.rm = TRUE))
```

We can now figure out the possible rent for the apartments on sale based on their area.

```
sale_properties <- merge(sale_properties, area_m2_rent, by='area')
```

Now we can calculate the possible rent

```
sale_properties <- sale_properties %>% mutate("rent €/m"
                                              = `€ per m2` * `size m²`)
```

Now we can find out the profit of certain properties, if we use the rent to pay the loan

```
sale_properties <- sale_properties %>% mutate("Profit"
                                             = `rent €/m` - `loan payment €/m`)
```

Making a table with all relevant information

```
relevant_info <- data.frame('Price' = sale_properties$`price €`,
                            'Profit €/m' = sale_properties$Profit,
                            'Area' = sale_properties$area,
                            'Size' = sale_properties$`size m²`,
                            'Year' = sale_properties$year,
                            'Address' = sale_properties$address,
                            'Rent' = sale_properties$`rent €/m`)
```

For the purpose of finding the optimal house to invest in for a working class citizen lets only take properties with max price of 350 000 €.

```
relevant_info <- relevant_info %>%
  filter(Price <= 350000)
```

When we investigate the apartments more we notice some apartments are right-of-occupancy apartments so we don't want them in our analysis.

Up to 50 000 € price contains these types of houses so we can filter min value to be 50 000€.

```
relevant_info <- relevant_info %>%
  filter(Price >= 50000)
```

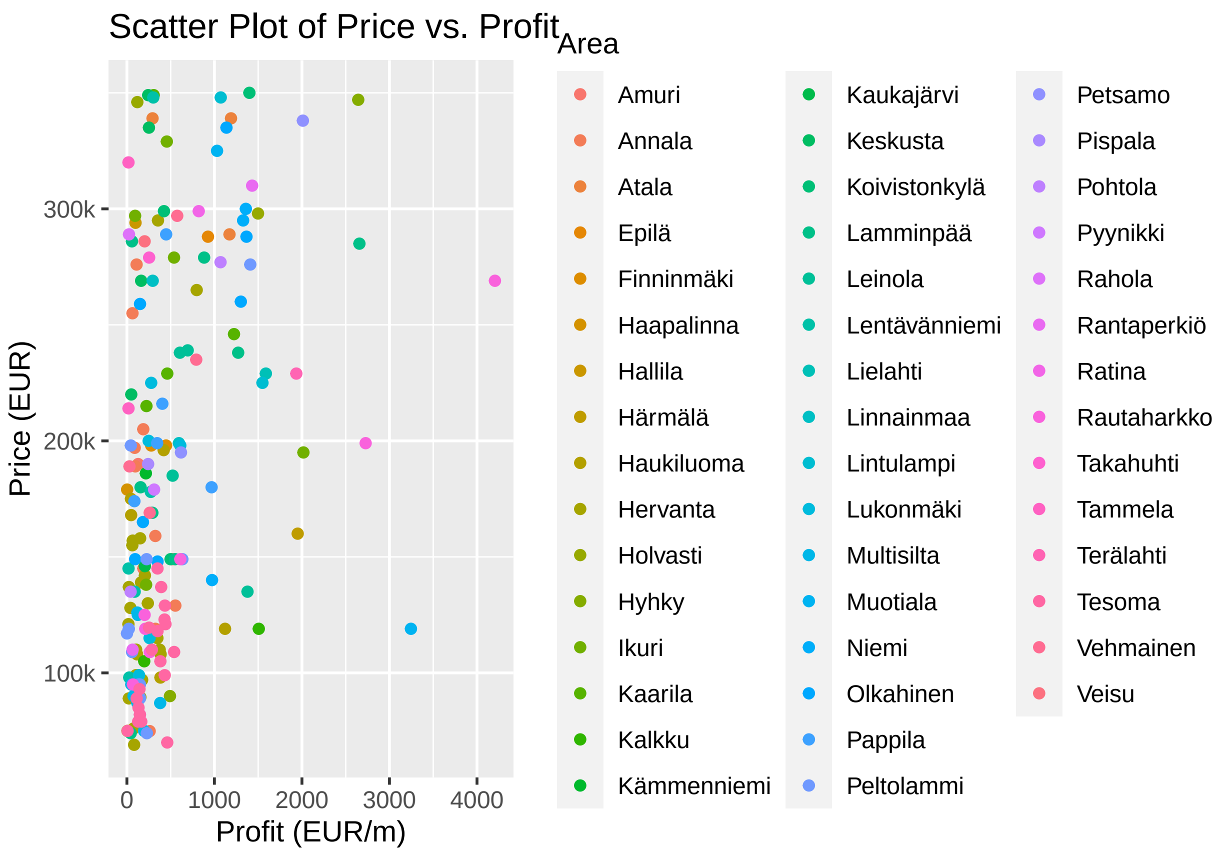
All of these houses come with a maintenance fee which is typically 450 €.

```
relevant_info <- relevant_info %>% mutate("Profit"
                                         = relevant_info$Profit...m - 450)
```

We only want to see the houses that still provide a profit

```
relevant_info <- relevant_info %>% filter(Profit > 0)
```

```
ggplot(data = relevant_info, aes(x = Profit, y = Price, color = Area)) +
  geom_point() +
  labs(x = "Profit (EUR/m)", y = "Price (EUR)", title
       = "Scatter Plot of Price vs. Profit") +
  scale_y_continuous(labels = scales::comma_format(scale = 1e-3, suffix = "k"))
```

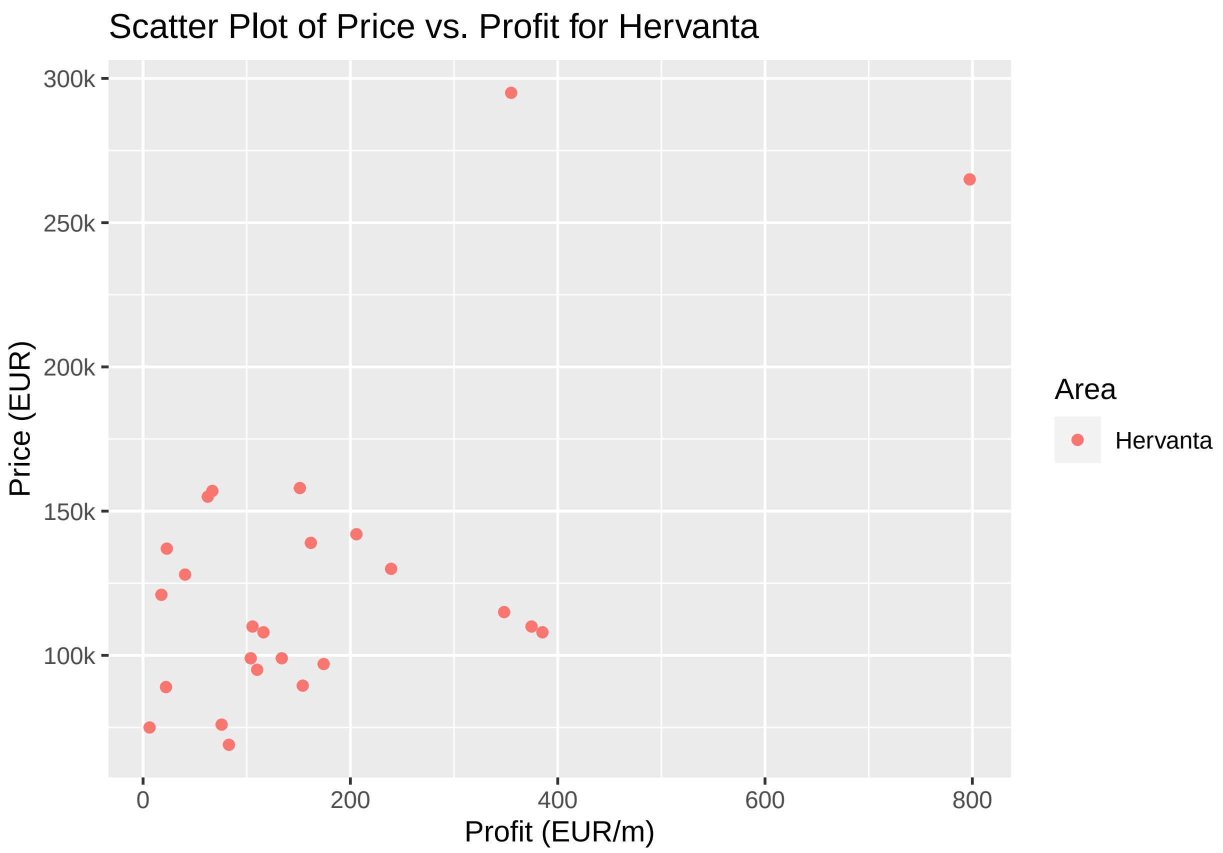


To find the optimal housing lets first see which area has the most houses providing profit.

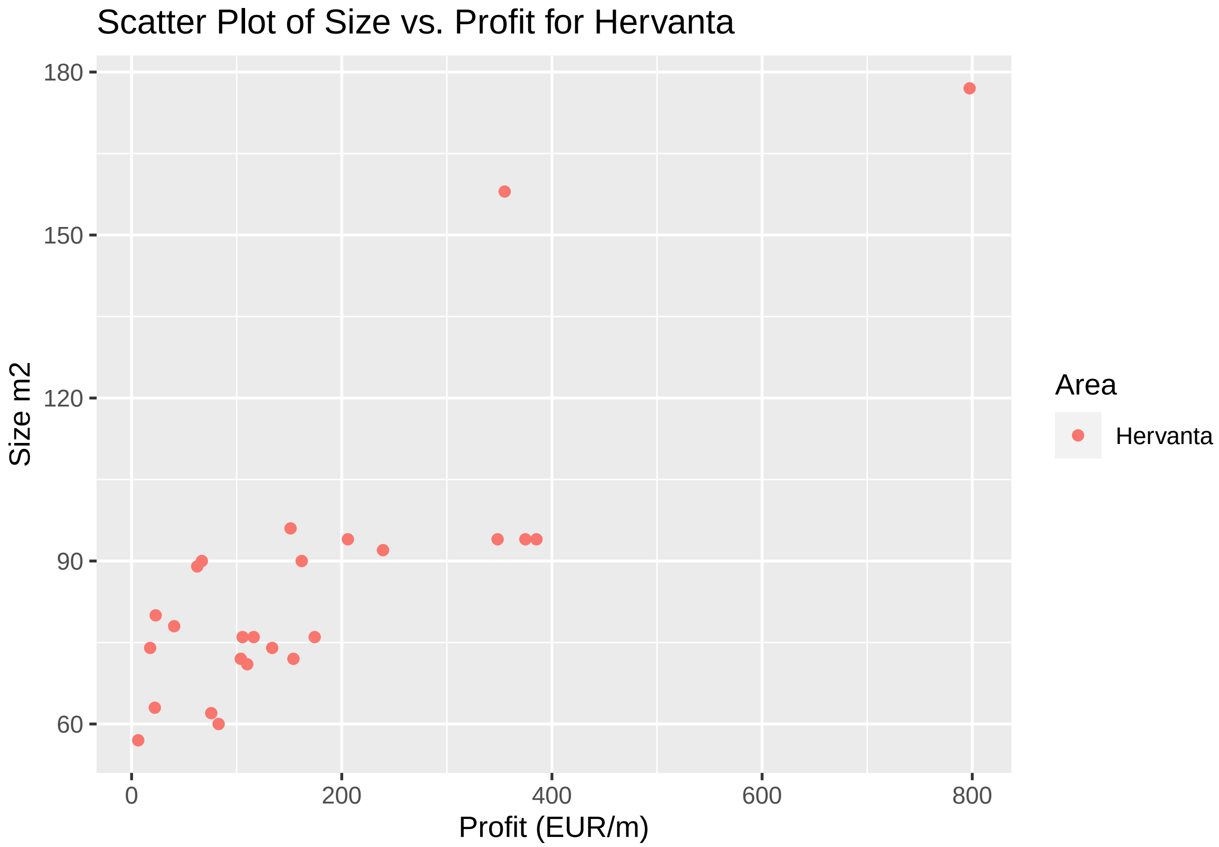
```
most_common_area <- relevant_info %>%
  count(Area) %>%
  slice(which.max(n)) %>%
  pull(Area)

filtered_data <- relevant_info %>%
  filter(Area == most_common_area)
```

```
ggplot(data = filtered_data, aes(x = Profit, y = Price, color = Area)) +
  geom_point() +
  labs(x = "Profit (EUR/m)", y = "Price (EUR)", title =
         paste("Scatter Plot of Price vs. Profit for", most_common_area)) +
  scale_y_continuous(labels = scales::comma_format(scale = 1e-3, suffix = "k"))
```



```
ggplot(data = filtered_data, aes(x = Profit, y = Size, color = Area)) +
  geom_point() +
  labs(x = "Profit (EUR/m)", y = "Size m2", title =
         paste("Scatter Plot of Size vs. Profit for", most_common_area))
```



So the study would show that the optimal investment would be to buy a house in the range of 100 - 150k € with the size of about a small family apartment. Also the average maintenance fee is lower in the suburbs so we can add a little more to the profit.

There are housing that provides a larger profit, but on average a property from Hervanta turns into profit and is statistically a safer choice.

Ofcourse real estate investing is not this simple and lots of other factors should be considered when buying property, but it is good to keep in mind the are and typical size that turns to profit.