Seinna_r

Thor Sanchez

2024-03-19

```
library(tidyverse)
library(lubridate)
library(dplyr)
library(ggplot2)
```

Hluti 1

a) Lets start by importing the data into R. I will use the read_csv2 function and then use glimpse function to obtain a quick overview of the datasets structure.

```
ths <- read_csv2("https://ahj.hi.is/kaupskra.csv",
                         locale = locale(encoding = "ISO8859-1"))
glimpse(ths)
```

```
## Rows: 169,636
## Columns: 22
## $ faerslunumer         <dbl> 569113, 558760, 566833, 566833, 628860, 579617, 5…
## $ emnr                 <dbl> 437, 436, 411, 411, 441, 411, 411, 411, 411, 411,…
## $ skjalanumer          <chr> "S-003590/2012", "X-000916/2011", "S-003191/2012"…
## $ fastnum              <dbl> 2064400, 2074439, 2264635, 2264636, 2264633, 2023…
## $ heimlilisfang        <chr> "Nýbýlavegur 12", "Drangahraun 10", "Dugguvogur 9…
## $ postnr               <dbl> 200, 220, 104, 104, 104, 104, 104, 104, 104, 105,…
## $ heinum               <dbl> 1022800, 1118718, 1003941, 1003941, 1003941, 1003…
## $ svfn                 <chr> "1000", "1400", "0000", "0000", "0000", "0000", "…
## $ sveitarfelag         <chr> "Kópavogsbær", "Hafnarfjarðarkaupstaður", "Reykja…
## $ utgdag               <dttm> 2012-07-30, 2011-02-28, 2012-04-16, 2012-04-16, …
## $ thinglystdags        <dttm> 2012-08-01 08:27:51, 2011-03-02 09:12:33, 2012-0…
## $ kaupverd             <dbl> 87000, 36000, 31000, 31000, 23500, 33500, 31000, …
## $ fasteignamat         <dbl> 70850, 40790, 4679, 5516, 13200, 27100, 3975, 711…
## $ byggar               <dbl> 1985, 1983, 1962, 1962, 1962, 1962, 1962, 1962, 1…
## $ epilog               <chr> "010101", "010102", "010301", "010302", "010201",…
## $ einflm               <dbl> 780.4, 400.0, 310.2, 310.2, 71.4, 325.0, 310.2, 3…
## $ lod_flm              <dbl> 1105, 3000, 565, 565, 565, 565, 565, 565, 565, 57…
## $ lod_flmein           <chr> "m²", "m²", "m²", "m²", "m²", "m²", "m²", "m²", "…
## $ tegund               <chr> "Atvinnuhusnaedi", "Atvinnuhusnaedi", "Atvinnuhus…
## $ fullbuid             <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ onothaefur_samningur <dbl> 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0…
## $ ...22                <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
```

b) We will check for properties that have been marked as completed and are of

the types “Fjolbyli” or “Serbyli”. Furthermore, we are only interested in properties with usable contracts. The following code filters our dataset accordingly and updates it by overwriting the original with this more focused subset.

```
ths <- ths %>%
  filter(fullbuid == 1, tegund %in% c("Fjolbyli", "Serbyli"), onothaefur_samningur
== 0)

glimpse(ths)
```

```
## Rows: 116,946
## Columns: 22
## $ faerslunumer         <dbl> 685833, 683362, 676099, 660107, 640011, 634801, 6…
## $ emnr                 <dbl> 441, 441, 441, 441, 441, 441, 441, 441, 441, 441,…
## $ skjalanumer          <chr> "E-002266/2022", "A-000735/2022", "C-011897/2021"…
## $ fastnum              <dbl> 2019513, 2019520, 2019517, 2019515, 2019516, 2019…
## $ heimlilisfang        <chr> "Engjateigur 17-19", "Engjateigur 17-19", "Engjat…
## $ postnr               <dbl> 105, 105, 105, 105, 105, 105, 105, 105, 105, 105,…
## $ heinum               <dbl> 1004483, 1004483, 1004483, 1004483, 1004483, 1004…
## $ svfn                 <chr> "0000", "0000", "0000", "0000", "0000", "0000", "…
## $ sveitarfelag         <chr> "Reykjavíkurborg", "Reykjavíkurborg", "Reykjavíku…
## $ utgdag               <dttm> 2022-03-07, 2022-01-14, 2021-07-20, 2020-08-10, …
## $ thinglystdags        <dttm> 2022-03-16 10:53:04, 2022-01-18 13:08:54, 2021-0…
## $ kaupverd             <dbl> 68250, 122000, 52500, 49900, 48000, 51500, 50000,…
## $ fasteignamat         <dbl> 55700, 86700, 50750, 47750, 43850, 41400, 41400, …
## $ byggar               <dbl> 1992, 1992, 1992, 1992, 1992, 1992, 1992, 1992, 1…
## $ epilog               <chr> "010204", "010211", "010208", "010206", "010207",…
## $ einflm               <dbl> 109.9, 212.6, 109.9, 109.9, 109.9, 109.9, 109.9, …
## $ lod_flm              <dbl> 5702, 5702, 5702, 5702, 5702, 5702, 5702, 5702, 5…
## $ lod_flmein           <chr> "m²", "m²", "m²", "m²", "m²", "m²", "m²", "m²", "…
## $ tegund               <chr> "Fjolbyli", "Fjolbyli", "Fjolbyli", "Fjolbyli", "…
## $ fullbuid             <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ onothaefur_samningur <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ ...22                <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
```

c) Now we will calculate the price per square meter for each property, providing us with a standardized value to compare properties of different sizes. This new variable, fermetraverd, will be added to our dataset with the following operation.

```
# Creating a new variable for price per square meter
ths <- ths %>%
  mutate(fermetraverd = kaupverd / einflm)

glimpse(ths)
```

```
## Rows: 116,946
## Columns: 23
## $ faerslunumer         <dbl> 685833, 683362, 676099, 660107, 640011, 634801, 6…
## $ emnr                 <dbl> 441, 441, 441, 441, 441, 441, 441, 441, 441, 441,…
## $ skjalanumer          <chr> "E-002266/2022", "A-000735/2022", "C-011897/2021"…
## $ fastnum              <dbl> 2019513, 2019520, 2019517, 2019515, 2019516, 2019…
## $ heimlilisfang        <chr> "Engjateigur 17-19", "Engjateigur 17-19", "Engjat…
## $ postnr               <dbl> 105, 105, 105, 105, 105, 105, 105, 105, 105, 105,…
## $ heinum               <dbl> 1004483, 1004483, 1004483, 1004483, 1004483, 1004…
## $ svfn                 <chr> "0000", "0000", "0000", "0000", "0000", "0000", "…
## $ sveitarfelag         <chr> "Reykjavíkurborg", "Reykjavíkurborg", "Reykjavíku…
## $ utgdag               <dttm> 2022-03-07, 2022-01-14, 2021-07-20, 2020-08-10, …
## $ thinglystdags        <dttm> 2022-03-16 10:53:04, 2022-01-18 13:08:54, 2021-0…
## $ kaupverd             <dbl> 68250, 122000, 52500, 49900, 48000, 51500, 50000,…
## $ fasteignamat         <dbl> 55700, 86700, 50750, 47750, 43850, 41400, 41400, …
## $ byggar               <dbl> 1992, 1992, 1992, 1992, 1992, 1992, 1992, 1992, 1…
## $ epilog               <chr> "010204", "010211", "010208", "010206", "010207",…
## $ einflm               <dbl> 109.9, 212.6, 109.9, 109.9, 109.9, 109.9, 109.9, …
## $ lod_flm              <dbl> 5702, 5702, 5702, 5702, 5702, 5702, 5702, 5702, 5…
## $ lod_flmein           <chr> "m²", "m²", "m²", "m²", "m²", "m²", "m²", "m²", "…
## $ tegund               <chr> "Fjolbyli", "Fjolbyli", "Fjolbyli", "Fjolbyli", "…
## $ fullbuid             <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ onothaefur_samningur <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ ...22                <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ fermetraverd         <dbl> 621.0191, 573.8476, 477.7070, 454.0491, 436.7607,…
```

d) We are adding a year column (ar) to our dataset, derived from the sale date (utgdag). This involves using the lubridate package to easily extract the year from each date.

```
ths <- ths %>%
  mutate(ar = year(utgdag))
```

e) To narrow our analysis, we aim to select postal codes each containing over 200 Serbyli. This step is crucial for ensuring our chosen areas have a substantial dataset for meaningful analysis. We first isolate Serbyli properties, then count and filter postal codes meeting this.

```
serbyli_df <- ths %>%
  filter(tegund == "Serbyli") %>%
  group_by(postnr) %>%
  summarise(fjoldi_serbyli = n()) %>%
  filter(fjoldi_serbyli >= 200)
```

```
# Find properties in 3 postal codes
valin_postnumer <- c(101, 105, 107)

ths_filtered <- ths %>%
  filter(postnr %in% valin_postnumer)

# View data
glimpse(ths_filtered)
```

```
## Rows: 18,206
## Columns: 24
## $ faerslunumer        <dbl> 685833, 683362, 676099, 660107, 640011, 634801, 6…
## $ emnr                <dbl> 441, 441, 441, 441, 441, 441, 441, 441, 441, 441,…
## $ skjalanumer         <chr> "E-002266/2022", "A-000735/2022", "C-011897/2021"…
## $ fastnum             <dbl> 2019513, 2019520, 2019517, 2019515, 2019516, 2019…
## $ heimlilisfang       <chr> "Engjateigur 17-19", "Engjateigur 17-19", "Engjat…
## $ postnr              <dbl> 105, 105, 105, 105, 105, 105, 105, 105, 105, 105,…
## $ heinum              <dbl> 1004483, 1004483, 1004483, 1004483, 1004483, 1004…
## $ svfn                <chr> "0000", "0000", "0000", "0000", "0000", "0000", "…
## $ sveitarfelag        <chr> "Reykjavíkurborg", "Reykjavíkurborg", "Reykjavíku…
## $ utgdag              <dttm> 2022-03-07, 2022-01-14, 2021-07-20, 2020-08-10, …
## $ thinglystdags       <dttm> 2022-03-16 10:53:04, 2022-01-18 13:08:54, 2021-0…
## $ kaupverd            <dbl> 68250, 122000, 52500, 49900, 48000, 51500, 50000,…
## $ fasteignamat        <dbl> 55700, 86700, 50750, 47750, 43850, 41400, 41400, …
## $ byggar              <dbl> 1992, 1992, 1992, 1992, 1992, 1992, 1992, 1992, 1…
## $ epilog              <chr> "010204", "010211", "010208", "010206", "010207",…
## $ einflm              <dbl> 109.9, 212.6, 109.9, 109.9, 109.9, 109.9, 109.9, …
## $ lod_flm             <dbl> 5702.0, 5702.0, 5702.0, 5702.0, 5702.0, 5702.0, 5…
## $ lod_flmein          <chr> "m²", "m²", "m²", "m²", "m²", "m²", "m²", "m²", "…
## $ tegund              <chr> "Fjolbyli", "Fjolbyli", "Fjolbyli", "Fjolbyli", "…
## $ fullbuid            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ onothaefur_samningur <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ ...22               <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ fermetraverd        <dbl> 621.0191, 573.8476, 477.7070, 454.0491, 436.7607,…
## $ ar                  <dbl> 2022, 2022, 2021, 2020, 2018, 2018, 2018, 2017, 2…
```

f) To ensure consistency, we are standardizing postal codes to a three-character format, adding leading zeros where necessary. This step prepares our data for accurate grouping and analysis.

```
ths$postnr <- as.character(ths$postnr)
ths$postnr <- sprintf("%03d", as.integer(ths$postnr))
str(ths)
```

```
## tibble [116,946 × 24] (S3: tbl_df/tbl/data.frame)
##  $ faerslunumer        : num [1:116946] 685833 683362 676099 660107 640011 ...
##  $ emnr                : num [1:116946] 441 441 441 441 441 441 441 441 441 441
...
##  $ skjalanumer         : chr [1:116946] "E-002266/2022" "A-000735/2022" "C-01189
7/2021" "A-010441/2020" ...
##  $ fastnum             : num [1:116946] 2019513 2019520 2019517 2019515 2019516
...
##  $ heimlilisfang       : chr [1:116946] "Engjateigur 17-19" "Engjateigur 17-19"
"Engjateigur 17-19" "Engjateigur 17-19" ...
##  $ postnr              : chr [1:116946] "105" "105" "105" "105" ...
##  $ heinum              : num [1:116946] 1e+06 1e+06 1e+06 1e+06 1e+06 ...
##  $ svfn                : chr [1:116946] "0000" "0000" "0000" "0000" ...
##  $ sveitarfelag        : chr [1:116946] "Reykjavíkurborg" "Reykjavíkurborg" "Rey
kjavíkurborg" "Reykjavíkurborg" ...
##  $ utgdag              : POSIXct[1:116946], format: "2022-03-07" "2022-01-14"
...
##  $ thinglystdags       : POSIXct[1:116946], format: "2022-03-16 10:53:04" "2022-
01-18 13:08:54" ...
##  $ kaupverd            : num [1:116946] 68250 122000 52500 49900 48000 ...
##  $ fasteignamat        : num [1:116946] 55700 86700 50750 47750 43850 ...
##  $ byggar              : num [1:116946] 1992 1992 1992 1992 1992 ...
##  $ epilog              : chr [1:116946] "010204" "010211" "010208" "010206" ...
##  $ einflm              : num [1:116946] 110 213 110 110 110 ...
##  $ lod_flm             : num [1:116946] 5702 5702 5702 5702 5702 ...
##  $ lod_flmein          : chr [1:116946] "m²" "m²" "m²" "m²" ...
##  $ tegund              : chr [1:116946] "Fjolbyli" "Fjolbyli" "Fjolbyli" "Fjolby
li" ...
##  $ fullbuid            : num [1:116946] 1 1 1 1 1 1 1 1 1 1 ...
##  $ onothaefur_samningur: num [1:116946] 0 0 0 0 0 0 0 0 0 0 ...
##  $ ...22               : logi [1:116946] NA NA NA NA NA NA ...
##  $ fermetraverd        : num [1:116946] 621 574 478 454 437 ...
##  $ ar                  : num [1:116946] 2022 2022 2021 2020 2018 ...
```

g) We will randomly select 200 properties from each postal code, ensuring our sample is reproducible by setting a specific seed.

```
set.seed(37)
ths_filtered %>%
  group_by(postnr) %>%
  sample_n(size = 200) -> urtak
str(urtak)
```

```
## gropd_df [600 × 24] (S3: grouped_df/tbl_df/tbl/data.frame)
##  $ faerslunumer        : num [1:600] 627906 674840 689747 644755 639619 ...
##  $ emnr                : num [1:600] 441 441 441 441 441 441 411 441 441 411 ...
##  $ skjalanumer         : chr [1:600] "C-000944/2018" "F-003339/2021" "E-006458/2
022" "E-004292/2019" ...
##  $ fastnum             : num [1:600] 2007385 2294798 2264932 2009182 2000298 ...
##  $ heimlilisfang       : chr [1:600] "Sóleyjargata 19" "Ánanaust 15" "Bergstaðas
træti 33" "Laufásvegur 65A" ...
##  $ postnr              : num [1:600] 101 101 101 101 101 101 101 101 101 101 ...
##  $ heinum              : num [1:600] 1016605 1001295 1002260 1165780 1019099 ...
##  $ svfn                : chr [1:600] "0000" "0000" "0000" "0000" ...
##  $ sveitarfelag        : chr [1:600] "Reykjavíkurborg" "Reykjavíkurborg" "Reykja
víkurborg" "Reykjavíkurborg" ...
##  $ utgdag              : POSIXct[1:600], format: "2018-01-18" "2021-06-16" ...
##  $ thinglystdags       : POSIXct[1:600], format: "2018-01-29 10:05:21" "2021-06-
29 14:23:29" ...
##  $ kaupverd            : num [1:600] 71000 55000 58000 141000 23500 47000 17500
36400 84500 9500 ...
##  $ fasteignamat        : num [1:600] 54700 46300 37350 108950 20170 ...
##  $ byggar              : num [1:600] 1931 1978 1901 1931 1956 ...
##  $ epilog              : chr [1:600] "010201" "010402" "010001" "010101" ...
##  $ einflm              : num [1:600] 111.2 96.1 65.6 248.8 67.9 ...
##  $ lod_flm             : num [1:600] 417 865 199 736 452 ...
##  $ lod_flmein          : chr [1:600] "m²" "m²" "m²" "m²" ...
##  $ tegund              : chr [1:600] "Fjolbyli" "Fjolbyli" "Fjolbyli" "Fjolbyli"
...
##  $ fullbuid            : num [1:600] 1 1 1 1 1 1 1 1 1 1 ...
##  $ onothaefur_samningur: num [1:600] 0 0 0 0 0 0 0 0 0 0 ...
##  $ ...22               : logi [1:600] NA NA NA NA NA NA ...
##  $ fermetraverd        : num [1:600] 638 572 884 567 346 ...
##  $ ar                  : num [1:600] 2018 2021 2022 2019 2018 ...
##  - attr(*, "groups")= tibble [3 × 2] (S3: tbl_df/tbl/data.frame)
##   ..$ postnr: num [1:3] 101 105 107
##   ..$ .rows : list<int> [1:3]
##   .. ..$ : int [1:200] 1 2 3 4 5 6 7 8 9 10 ...
##   .. ..$ : int [1:200] 201 202 203 204 205 206 207 208 209 210 ...
##   .. ..$ : int [1:200] 401 402 403 404 405 406 407 408 409 410 ...
##   .. ..@ ptype: int(0)
##   ..- attr(*, ".drop")= logi TRUE
```

Hluti 2

h) We will count the sérbyli and fjölbýli homes in each district to see what types of properties they have.

```
eignir_eftir_gerd <- urtak %>%
  group_by(postnr, tegund) %>%
  summarise(fjoldi = n(), .groups = "drop")

eignir_eftir_gerd
```

```
## # A tibble: 6 × 3
##   postnr tegund   fjoldi
##    <dbl> <chr>     <int>
## 1    101 Fjolbyli    182
## 2    101 Serbyli      18
## 3    105 Fjolbyli    191
## 4    105 Serbyli       9
## 5    107 Fjolbyli    186
## 6    107 Serbyli      14
```

i) We will calculate the proportion of sérbyli and fjölbýli homes in each district to understand the makeup of property types better.

```
hlutfall_eigna <- urtak %>%
  group_by(postnr, tegund) %>%
  summarise(fjoldi = n(), .groups = "drop") %>%
  group_by(postnr) %>%
  mutate(hlutfall = fjoldi / sum(fjoldi) * 100) %>%
  ungroup()
```

j) To see if the proportion of fjölbyli homes varies across our three districts, we can use a straightforward test called the Chi-squared test. This test helps us understand if differences we see are likely due to chance or if they're significant. Null Hypothesis (H0): The proportion of property types (fjölbýli and sérbýli) is the same across the three postal codes. Any observed difference is due to random chance. Alternative Hypothesis (H1): There is a significant difference in the proportion of property types across the three postal codes.

```
# making krosstoflu
krosstoflu_data <- table(urtak$postnr, urtak$tegund)

# chi-square test
chi2_test <- chisq.test(krosstoflu_data)

print(chi2_test)
```

```
##
##  Pearson's Chi-squared test
##
## data:  krosstoflu_data
## X-squared = 3.1939, df = 2, p-value = 0.2025
```

Test statistic (X-squared): 3.1939. This is the calculated chi-squared statistic. Degrees of freedom (df): 2. This is one less than the number of categories (postal codes) you are comparing. P-value: 0.2025. This is the probability of observing a chi-squared statistic as extreme as 3.1939 under the null hypothesis.

Conclusion: The p-value of 0.2025 is greater than the conventional threshold of 0.05, which suggests that the evidence is not strong enough to reject the null hypothesis. Therefore, we would conclude that there is not a statistically significant difference in the proportion of property types (fjölbýli and sérbýli) across the three postal codes.

k) Testing for Difference in Mean Price per Square Meter.

Hypotheses: H0: There is no difference in mean price per square meter between Fjolbyli and Serbyli homes. H1: There is a difference in mean price per square meter between Fjolbyli and Serbyli homes.

```r
# Performing a t-test to compare the average price per square meter
t_test_result <- t.test(fermetraverd ~ tegund, data = urtak)

t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  fermetraverd by tegund
## t = 1.9717, df = 53.77, p-value = 0.0538
## alternative hypothesis: true difference in means between group Fjolbyli and grou
p Serbyli is not equal to 0
## 95 percent confidence interval:
##  -0.6390012 76.1179732
## sample estimates:
## mean in group Fjolbyli  mean in group Serbyli
##               429.1862               391.4467
```

P-value: The p-value is slightly above the common alpha level of 0.05, which suggests that the evidence against the null hypothesis is not quite strong. In other words, the test does not provide strong enough evidence to conclude that there is a statistically significant difference in mean price per square meter between Fjölbýli and Serbýli homes.

Confidence Interval: The 95% confidence interval for the difference in means ranges from -0.6390012 to 76.1179732. Since this interval includes zero, it indicates that there is a chance that there is no difference between the groups.

Mean Estimates: The sample mean price per square meter for Fjölbýli homes is 429.1862, and for Serbýli homes, it is 391.4467.

Conclusion: Given that the p-value is close to 0.05, we might consider this result as indicating a trend towards a difference, but it does not provide strong statistical evidence to reject the null hypothesis at the 5% significance level.

l) We are creating a new variable staerd200, to indicate whether properties are larger or smaller than 200 square meters. This variable will help us in further analysis by separating properties into two groups: 'Stærri' for larger properties and 'Minni' for smaller ones.

```
# Creating 'staerd200' based on property size
urtak <- urtak %>%
  mutate(staerd200 = ifelse(einflm >= 200, "Stærri", "Minni"))

# To show staerd200
head(urtak %>% select(1:5, staerd200))
```

```
## Adding missing grouping variables: `postnr`
```

```
## # A tibble: 6 × 7
## # Groups:   postnr [1]
##   postnr faerslunumer  emnr skjalanumer   fastnum heimlilisfang      staerd200
##    <dbl>        <dbl> <dbl> <chr>           <dbl> <chr>              <chr>
## 1    101       627906   441 C-000944/2018 2007385 Sóleyjargata 19    Minni
## 2    101       674840   441 F-003339/2021 2294798 Ánanaust 15        Minni
## 3    101       689747   441 E-006458/2022 2264932 Bergstaðastræti 33 Minni
## 4    101       644755   441 E-004292/2019 2009182 Laufásvegur 65A    Stærri
## 5    101       639619   441 A-013568/2018 2000298 Vesturgata 52      Minni
## 6    101       612984   441 D-012401/2016 2000873 Vesturgata 51C     Minni
```

m) Hypotheses: Null hypothesis (H0): There’s no difference in the mean price per square meter between properties larger than 200 square meters and those that are smaller or equal to 200 square meters. Alternative hypothesis (H1): There is a difference in the mean price per square meter between the two size categories of properties.

```
# Performing a t-test to compare the average price per square meter between size ca
tegories
t_test_size <- t.test(fermetraverd ~ staerd200, data = urtak)

t_test_size
```

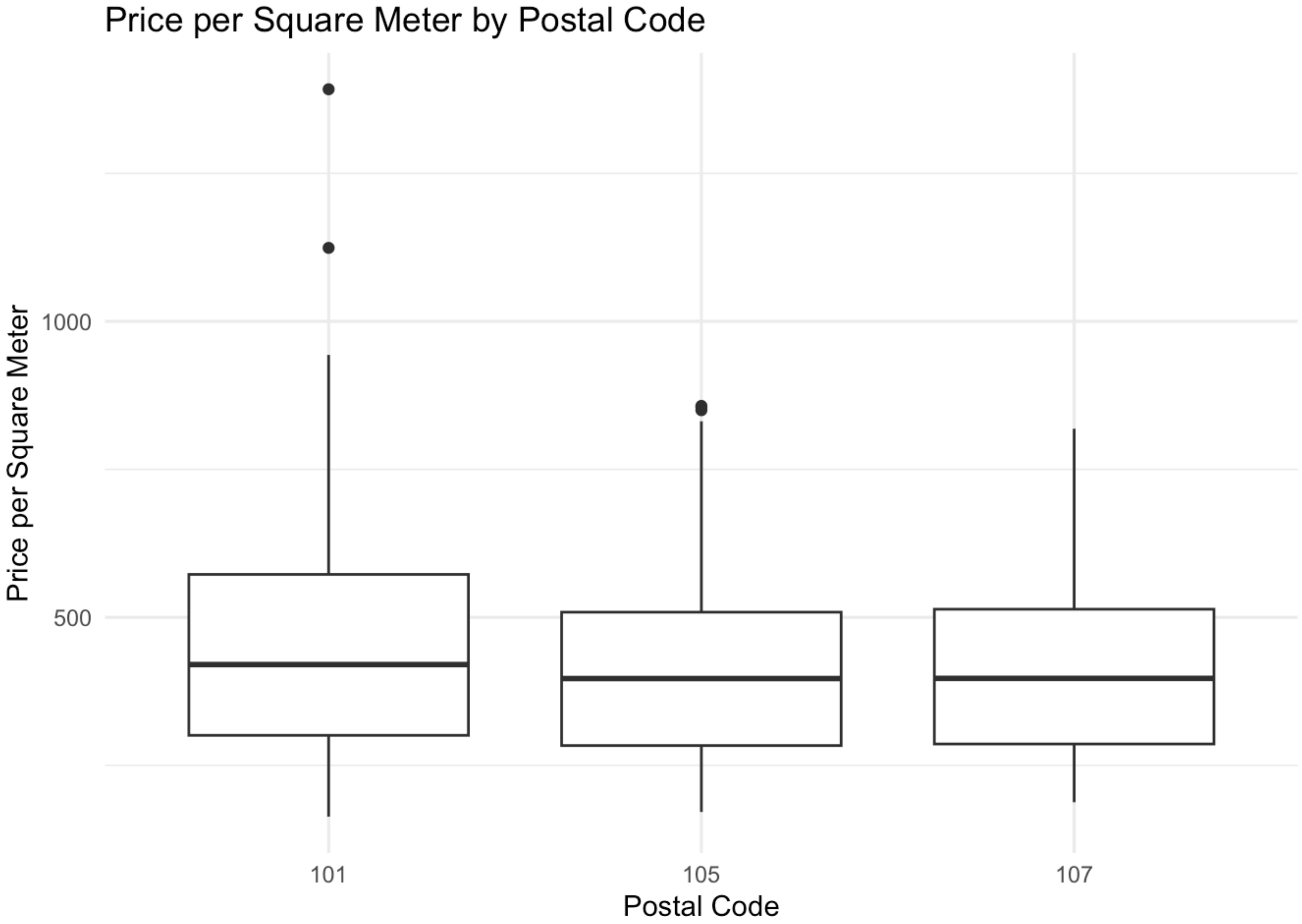
```
##
##  Welch Two Sample t-test
##
## data:  fermetraverd by staerd200
## t = 3.2004, df = 32.847, p-value = 0.00304
## alternative hypothesis: true difference in means between group Minni and group S
tærri is not equal to 0
## 95 percent confidence interval:
##   26.56304 119.31856
## sample estimates:
##  mean in group Minni mean in group Stærri
##             430.0112             357.0705
```

We found out that smaller homes usually cost more per square meter than bigger ones. Because our p-value is really small (0.00304). The numbers show us that homes 200 square meters or smaller are pricier for each square meter you get.

Hluti 4

n) To create a suitable visualization of the fermetraverd and postnr, we can use a boxplot to see the distribution of prices across different postal codes. Here’s how to do this in R using ggplot2.

```
ggplot(urtak, aes(x = factor(postnr), y = fermetraverd)) +
  geom_boxplot() +
  labs(title = "Price per Square Meter by Postal Code",
       x = "Postal Code",
       y = "Price per Square Meter") +
  theme_minimal()
```



o) Null hypothesis (H0): The mean price per square meter is the same across all districts. Alternative hypothesis (H1): At least one district has a different mean price per square meter.

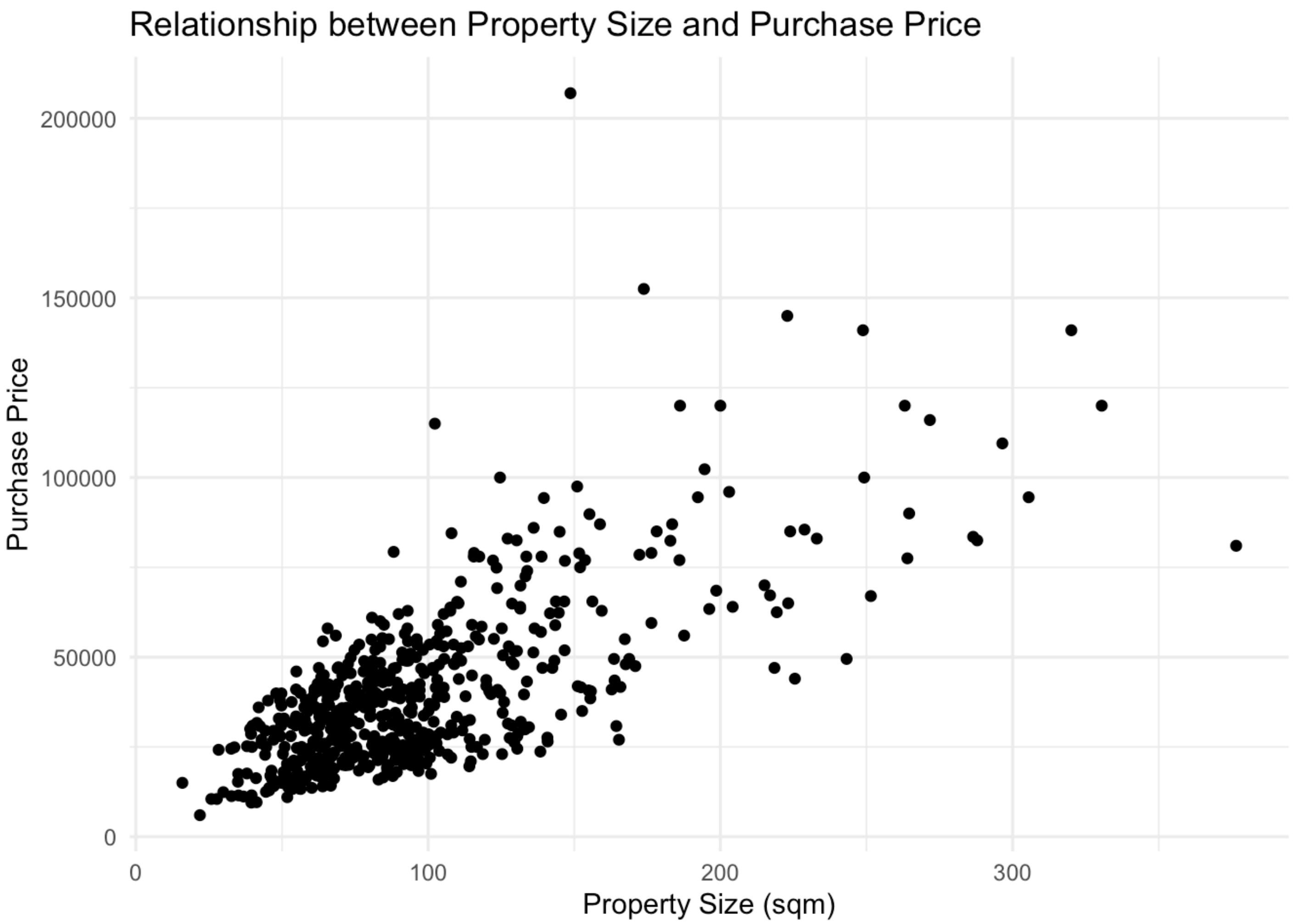
```
# Performing an ANOVA test to compare the average price per square meter across districts
anova_result <- aov(fermetraverd ~ postnr, data = urtak)

summary(anova_result)
```

```
##              Df   Sum Sq Mean Sq F value  Pr(>F)   
## postnr        1   268481  268481   9.967 0.00167 **
## Residuals   598 16107591   26936                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: Since the ANOVA test found significant differences in mean prices per square meter across districts, and assuming the assumptions are reasonably met as judged by the boxplot, we conclude that location (as indicated by postal code) has an impact on property prices. It suggests that certain districts have higher or lower prices per square meter than others, which could be important for real estate valuation and market analysis.

p) To visualize the relationship between the size of the properties (einflm) and the purchase price (kaupverd), a scatter plot is typically appropriate. It allows us to see individual data points and any apparent relationship between the two variables.Here is how we can create a scatter plot with ggplot2.

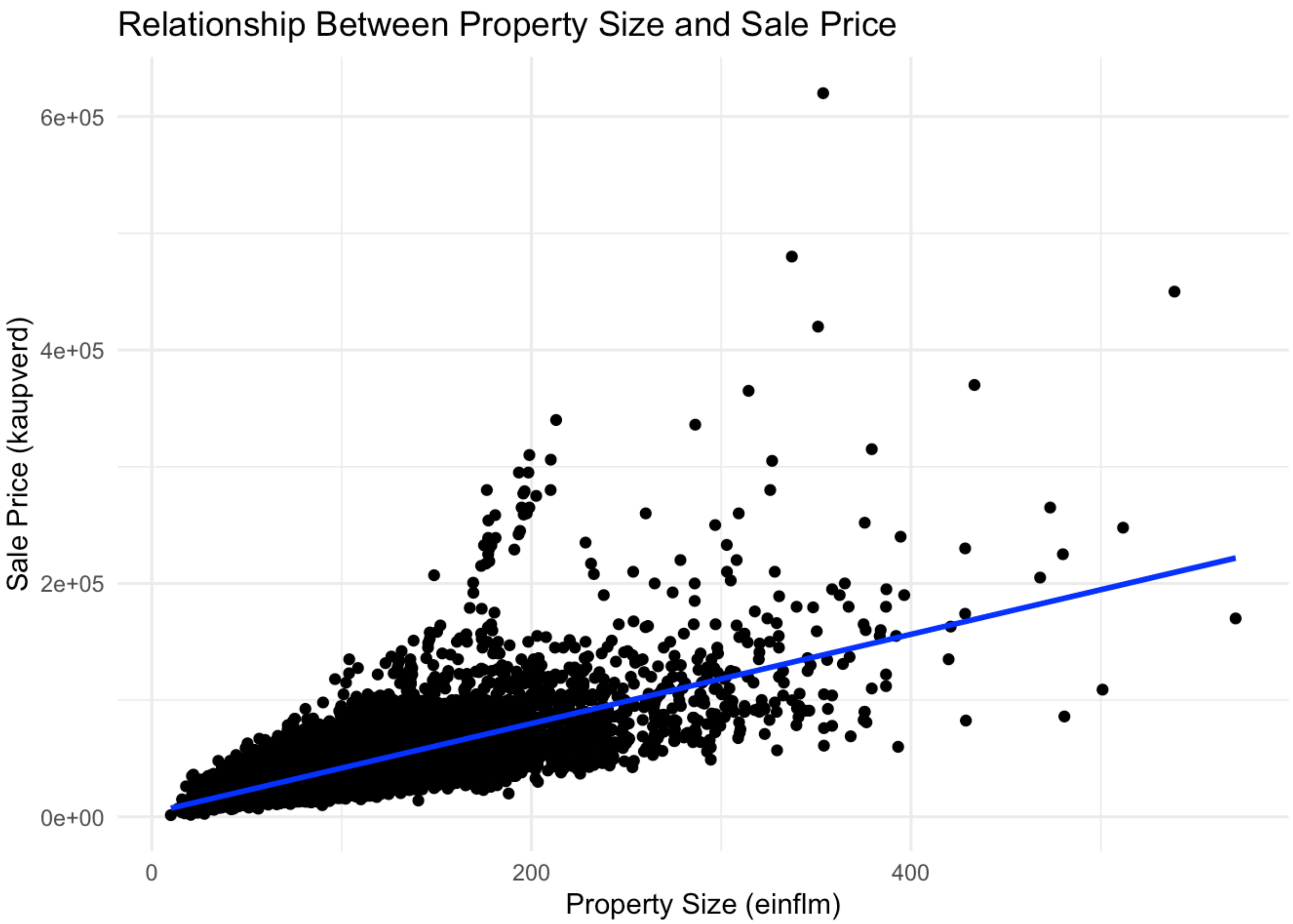
```
ggplot(urtak, aes(x = einflm, y = kaupverd)) +
  geom_point() +
  labs(title = "Relationship between Property Size and Purchase Price",
       x = "Property Size (sqm)",
       y = "Purchase Price") +
  theme_minimal()
```



q)

```
# Scatter plot of property size vs sale price with regression line
ggplot(ths_filtered, aes(x = einflm, y = kaupverd)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue") +
  theme_minimal() +
  labs(title = "Relationship Between Property Size and Sale Price",
       x = "Property Size (einflm)",
       y = "Sale Price (kaupverd)")
```

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



```
# Building the linear regression model
linear_model <- lm(kaupverd ~ einflm, data = urtak)

summary(linear_model)
```

```
##
## Call:
## lm(formula = kaupverd ~ einflm, data = urtak)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52457 -10963  -1573   8727 149855
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7330.90    1484.91   4.937 1.03e-06 ***
## einflm        335.00      13.73  24.391  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16440 on 598 degrees of freedom
## Multiple R-squared:  0.4987, Adjusted R-squared:  0.4979
## F-statistic: 594.9 on 1 and 598 DF,  p-value: < 2.2e-16
```