Greining fasteignaverðs

Elías Bjartur Einarsson (ebe19) og Þórhallur Auður Helgason (thh14)

21/10/2021

Inngangur

Í þessu verkefni er markmið okkar að skapa líkan sem nálgar fasteignamat ákveðinna svæða Reykjavíkur að gefnum upplýsingum um fasteignir. Við lýsum því hér hvaða skref við tókum í líkanasmíðinni, hvernig við völdum breytur, mátum gæði líkans og bárum saman líkön sem komu til greina. XXvalidate

Gögnin sem unnið var með voru fasteignamat eigna fyrir árið 2017, unnið árið 2016. Einungis var notast við gögn sem vörðuðu fimm svæði innan Reykjavíkur; miðbæ frá Bræðraborgarstíg að Tjörn, Melar að sjó, Háaleiti og skeifan, Hólar og Berg í Breiðholti og loks Réttarholt. Af þessum svæðum voru í heildina 433 fasteignir með 21 breytum auk núvirðis. Breyturnar voru eftirfarandi: Fastanúmer íbúðar, Kaupdagur, Tegund eignar, Svæðisnúmer, Byggingarár, Hæð íbúðar, Fjöldi lyfta, Fermetrafjöldi, Fjöldi hæða, Fjöldi bílastæða, Fjöldi baðkara, Fjöldi sturta, Fjöldi klósetta, Fjöldi eldhúsa, Fjöldi herbergja, Fjöldi stofa, Fjöldi geymsla, Stig framkvæmdar, Matssvæði, Undirmatssvæði og Tegund íbúðar.

Til að byrja með skoðuðum við gögnin, meðaltöl, há- og lággildi ásamt því að umbreyta þeim yfir á rétt snið, svo sem dagsetningum og flokkunarbreytum.

Við fjarlægjum fastanúmer sem breytu þar sem hún er einungis auðkenni fasteignar og inniheldur ekki upplýsingar um hana. Sömuleiðis fjarlægjum við svæðisnúmerið þar sem allar breyturnar deila svæðisnúmeri Reykjavíkur.

```
# Fjarlægjum breytur sem augljóslega skipta ekki máli:
data <- subset( data, select = -c(svfn,rfastnum) )

# Skilgreinum tegundir breyta:
data[ ,"kdagur"] <- as.Date(data[ ,"kdagur"]) # Kaupdagur sem dagsetning
data[ ,"teg_eign"] <- as.factor(data[ ,"teg_eign"]) # Tegund eignar sem flokkur

data[ ,"matssvaedi"] <- as.factor(data[ ,"matssvaedi"]) # Staðsetning sem flokkur
data[ ,"undirmatssvaedi"] <- as.factor(data[ ,"undirmatssvaedi"]) # Undirstaðsetning sem flokkur
data[ ,"ibteg"] <- as.factor(data[ ,"ibteg"]) # Tegund íbúðar sem flokkur

#nums <- unlist(lapply(data, is.numeric))
numericNames <- colnames(dplyr::select(data, where(is.numeric)))
summary(data[numericNames]) %>%
  kbl(align = 'c', col.names = c("Núvirdi", "Byggingarár",  "Nr. hæðar",    "Fjöldi lyfta", "Fermetrafj
  kable_styling()
```

```
library(table1)
```

```
##
## Attaching package: 'table1'
```

```
## The following objects are masked from 'package:base':
##
##     units, units<-
```

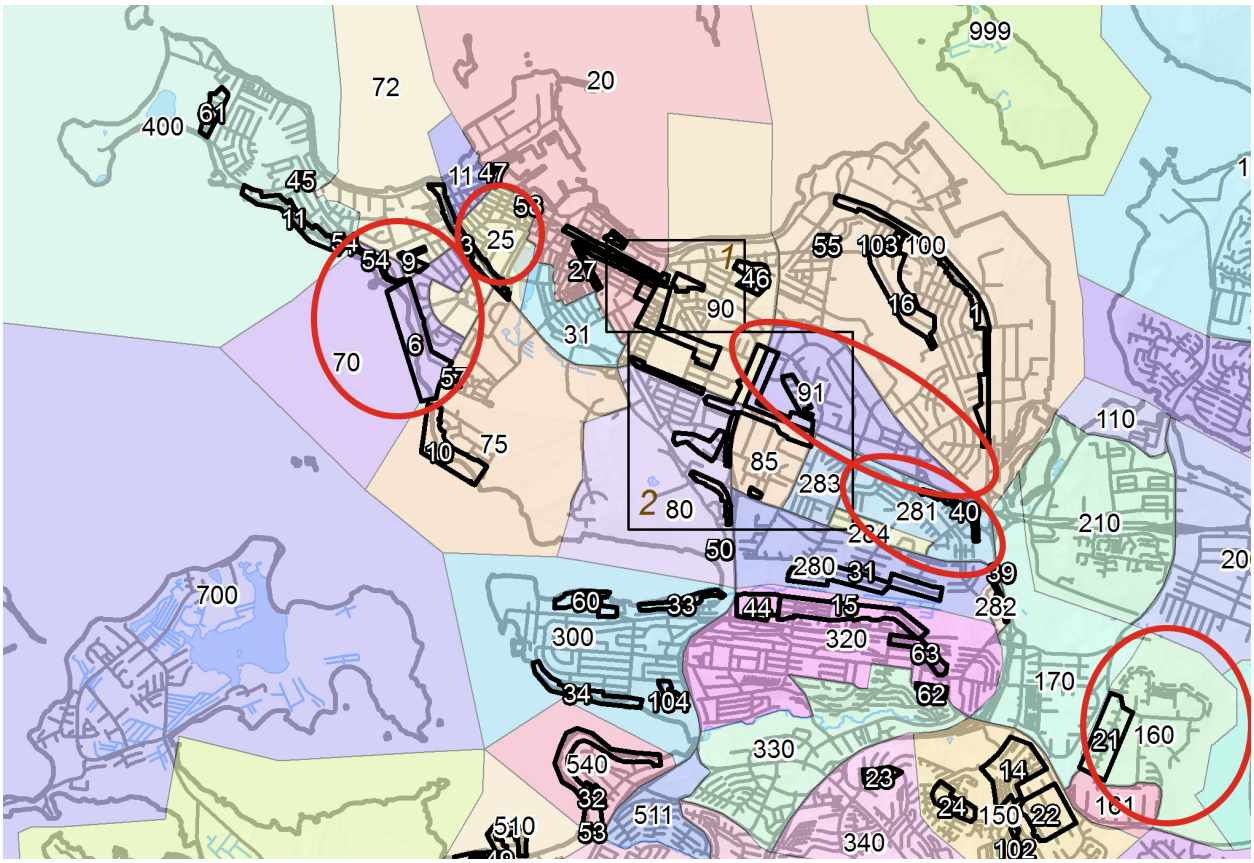

Figure 1: Svæði sem skoðuð voru rauðmerkt

	Núvirdi	Byggingarár	Nr. hæðar	Fjöldi lyfta	Fermetrafjöldi	Fjöldi hæða	Fjöldi bílastæða
	Min. : 5993	Min. :1901	Min. :0.000	Min. :0.0000	Min. : 21.90	Min. :1.000	Min. :0.00000
	1st Qu.: 20196	1st Qu.:1953	1st Qu.:1.000	1st Qu.:0.0000	1st Qu.: 65.50	1st Qu.:1.000	1st Qu.:0.00000
	Median : 25655	Median :1964	Median :2.000	Median :0.0000	Median : 85.30	Median :1.000	Median :0.00000
	Mean : 29163	Mean :1963	Mean :1.928	Mean :0.1663	Mean : 95.41	Mean :1.201	Mean :0.05312
	3rd Qu.: 33985	3rd Qu.:1974	3rd Qu.:3.000	3rd Qu.:0.0000	3rd Qu.:111.90	3rd Qu.:1.000	3rd Qu.:0.00000
	Max. :133665	Max. :2014	Max. :7.000	Max. :2.0000	Max. :289.30	Max. :3.000	Max. :2.00000

```
table1::table1(~., data[numericNames])
```

	Overall
	(N=433)
nuvirdi	
Mean (SD)	29200 (13900)
Median [Min, Max]	25700 [5990, 134000]
byggar	
Mean (SD)	1960 (19.0)
Median [Min, Max]	1960 [1900, 2010]
haednr	
Mean (SD)	1.93 (1.45)
Median [Min, Max]	2.00 [0, 7.00]
lyfta	
Mean (SD)	0.166 (0.451)
Median [Min, Max]	0 [0, 2.00]
ibm2	
Mean (SD)	95.4 (43.7)
Median [Min, Max]	85.3 [21.9, 289]
fjhaed	
Mean (SD)	1.20 (0.455)
Median [Min, Max]	1.00 [1.00, 3.00]
fjbilast	
Mean (SD)	0.0531 (0.235)
Median [Min, Max]	0 [0, 2.00]
fjbkar	
Mean (SD)	0.808 (0.433)
Median [Min, Max]	1.00 [0, 2.00]
fjsturt	
Mean (SD)	0.409 (0.520)
Median [Min, Max]	0 [0, 2.00]
fjklos	
Mean (SD)	1.19 (0.461)
Median [Min, Max]	1.00 [0, 3.00]
fjeld	
Mean (SD)	1.01 (0.192)
Median [Min, Max]	1.00 [0, 3.00]
fjherb	
Mean (SD)	2.45 (1.49)
Median [Min, Max]	2.00 [0, 13.0]
fjstof	
Mean (SD)	1.29 (0.561)
Median [Min, Max]	1.00 [0, 4.00]
fjgeym	
Mean (SD)	0.591 (0.671)
Median [Min, Max]	0 [0, 4.00]
stig10	
Mean (SD)	10.0 (0.0144)
Median [Min, Max]	10.0 [9.70, 10.0]

Við skiptum gagnasafninu okkar í þjálfunar- og prófunarsafn með 75% gagnapunkta í því fyrrnefnda.

Við erum þá tilbúnir að máta fyrsta líkanið okkar og greina það.

```
# Splittum datasetti í þjálfun og prófun:
sizeTraining = floor(0.75 * nrow(data))
trainingSampleRowId <- sample(1:nrow(data), size = sizeTraining, replace = F)
train_data <- data[trainingSampleRowId, ]
test_data <- data[-trainingSampleRowId, ]

# Fittum fyrsta líkan, án nokkurrar vinnslu:
lm.first = lm(nuvirdi ~ ., data = train_data)
```

Þetta líkan fær 5659.8623806 í RMSE.

```
summary(lm.first)
```

```
##
## Call:
## lm(formula = nuvirdi ~ ., data = train_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18335  -2859   -200   2556  46535
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -9.082e+05  2.192e+05  -4.144 4.47e-05 ***
## kdagur              6.489e+00  6.520e-01   9.952  < 2e-16 ***
## teg_eignIbudareign -3.537e+03  2.206e+03  -1.604 0.109849
## teg_eignParhus      7.340e+03  3.607e+03   2.035 0.042726 *
## teg_eignRadhus     -1.211e+03  2.328e+03  -0.520 0.603272
## byggar             -1.907e+01  2.557e+01  -0.746 0.456432
## haednr              4.095e+02  3.133e+02   1.307 0.192195
## lyfta               2.261e+03  1.086e+03   2.082 0.038175 *
## ibm2                2.391e+02  1.987e+01  12.028  < 2e-16 ***
## fjhaed              1.875e+03  1.366e+03   1.372 0.171047
## fjbilast           -2.422e+02  1.871e+03  -0.129 0.897110
## fjbkar              1.934e+03  1.153e+03   1.678 0.094485 .
## fjsturt             3.203e+02  8.798e+02   0.364 0.716118
## fjklos             -1.144e+03  1.404e+03  -0.815 0.416016
## fjeld              -3.091e+02  2.111e+03  -0.146 0.883675
## fjherb             -4.520e+02  4.346e+02  -1.040 0.299131
## fjstof              1.721e+03  9.076e+02   1.896 0.058923 .
## fjgeym              6.744e+02  5.628e+02   1.198 0.231828
## stig10              8.531e+04  2.124e+04   4.017 7.50e-05 ***
## matssvaedi70       -1.863e+03  1.225e+03  -1.521 0.129413
## matssvaedi91       -9.177e+03  1.631e+03  -5.626 4.32e-08 ***
## matssvaedi160      -1.396e+04  1.449e+03  -9.632  < 2e-16 ***
## matssvaedi281      -5.471e+03  1.437e+03  -3.807 0.000171 ***
## undirmatssvaedi3   -4.887e+03  2.635e+03  -1.855 0.064627 .
## undirmatssvaedi6    6.726e+03  3.149e+03   2.136 0.033482 *
## undirmatssvaedi21   6.819e+02  1.594e+03   0.428 0.669035
## undirmatssvaedi28   9.166e+02  1.749e+03   0.524 0.600521
## undirmatssvaedi40  -1.578e+03  4.473e+03  -0.353 0.724452
## undirmatssvaedi48   7.420e+03  3.300e+03   2.248 0.025290 *
## undirmatssvaedi54  -4.665e+03  4.449e+03  -1.049 0.295248
```

```
## ibteg12             4.587e+01  1.748e+03   0.026 0.979086
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5952 on 293 degrees of freedom
## Multiple R-squared:  0.8498, Adjusted R-squared:  0.8345
## F-statistic: 55.27 on 30 and 293 DF,  p-value: < 2.2e-16
```

```
test_resid = (predict(lm.first, test_data) - test_data$nuvirdi)
```

Er við skoðum spágildi líkansins út frá prófunargagnasetti fæst 6023.5066062 í RMSE.

Skoðum til viðbótar annað grunnlíkan sem mátar núvirði eingöngu við fermetraverð. Þetta líkan mætti hugsa sem grunnviðmið parsímóníunnar.

```
# Skoðum einnig annars konar grunnlíkan, sem tekur bara mið af fermetrum:
lm.simple <- lm(nuvirdi ~ ibm2, data = train_data)
summary(lm.simple)
```

```
##
## Call:
## lm(formula = nuvirdi ~ ibm2, data = train_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32699  -4865   -771   3649  58252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4692.91    1126.60   4.166 3.99e-05 ***
## ibm2          258.20      10.71  24.107  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8748 on 322 degrees of freedom
## Multiple R-squared:  0.6435, Adjusted R-squared:  0.6424
## F-statistic: 581.2 on 1 and 322 DF,  p-value: < 2.2e-16
```

```
sqrt(mean(residuals(lm.simple)^2))
```

```
## [1] 8721.104
```

```
## STIG 1: FYRSTA LÍKAN OG TÖLFRÆÐI HEILDARGAGNASETTS

# Skoðum mean, sd, min, max, kvantíla (og annað?) fyrir gagnasettið okkar:

# Ath að í verkefni segir að "lyfta" sé binary en í gögnum virðist hún vera fjöldi lyfta.
# data[ ,"lyfta"] <- data[,"lyfta"]>0 # Kannski halda, kannski sleppa
```