Greining fasteignaverðs

Elías Bjartur Einarsson (ebe19) og Þórhallur Auður Helgason (thh114)

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 gagnapunkta af fimm svæðum innan Reykjavíkur sem sjá má á korti hér að neðan; miðbæ frá Bræðraborgarstíg að Tjörn, Melar að sjó, Háaleiti og skeifan, Hólar og Berg í Breiðholti og loks Réttarholt. Á 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.

Þær breytur sem ekki segja sig sjálfar eru; Tegund eignar, Svæðisnúmer, Stig framkvæmdar, Matssvæði, Undirmatssvæði og Tegund íbúðar. Tegund eignar skiptist í fjóra flokka; Einbýlishús, parhús, íbúð og raðhús. Tegund íbúðar aðgreinir sérbýlishús frá fjölbýlishúsum. Svæðisnúmer er auðkenni sveitarfélags. Stig framkvæmdar er skali upp á 10 sem metur hvort húsnæðið sé tilbúið. Matssvæði eru hverfin sem sjást á myndinni hér að neðan og undirmatssvæði eru ákveðin svæði innan þeirra.

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

Við fjarlægjum strax 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 og veitir þar með engar upplýsingar um gagnapunktana.

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

# Athugum hvort auð gildi séu til staðar
if (sum(apply(data,2,is.nan))){
  print("Athuga auð gildi")
}

# 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)))
```

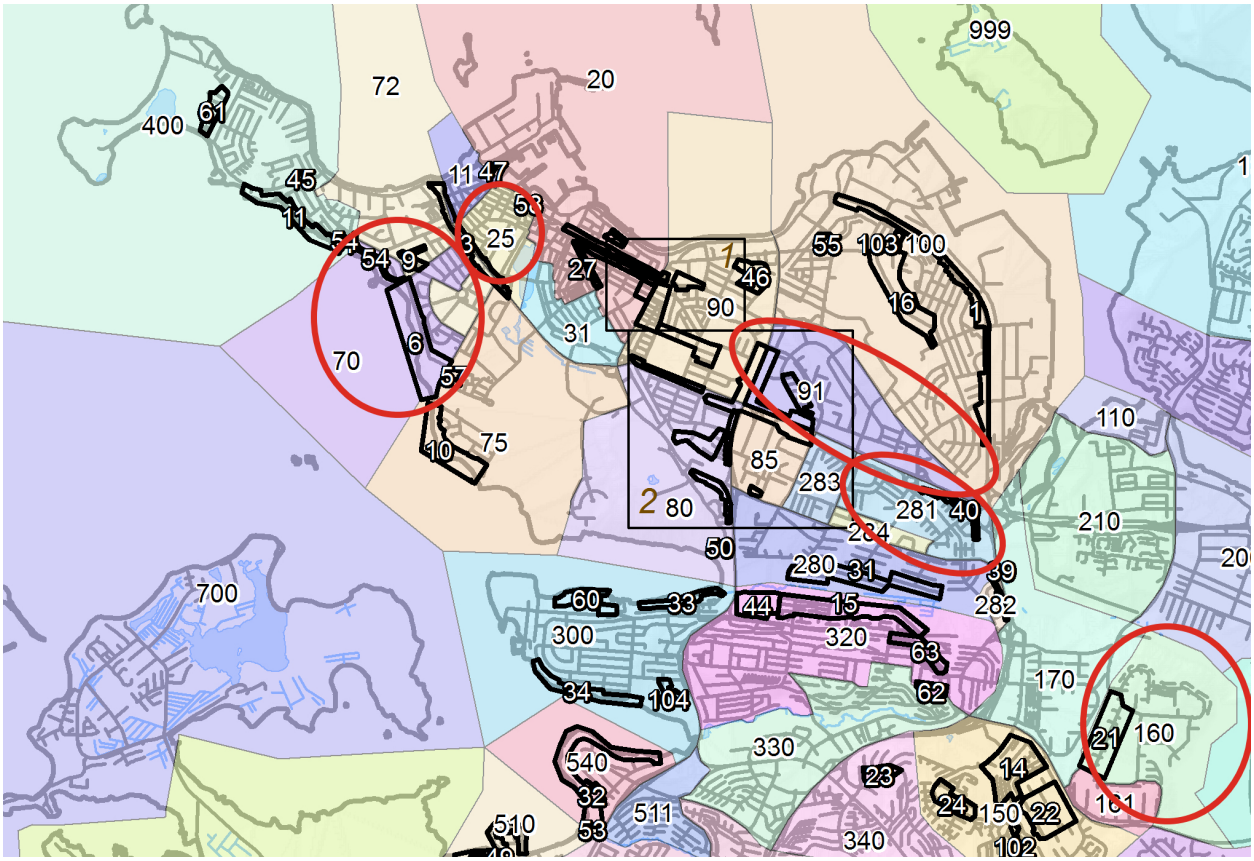


Figure 1: Svæði sem skoðuð voru eru rauðmerkt og nr. 70, 25, 91, 281 og 160.

	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

```
summary(data[numericNames]) %>%
  kbl(align = 'c', col.names = c("Núvirdi", "Byggingarár",  "Nr. hæðar",    "Fjöldi lyfta", "Fermetrafj
  kable_styling()
```

```
# library(table1)
# table1::table1(~., data[numericNames])
# Þessi er flottari en það vantar kvantíla
```

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

Fyrsta líkan

Að svo stöddu erum við tilbúnir að máta fyrsta líkanið okkar og greina það. Í fyrstu mátum við núvirði við allar þær breytur sem eftir standa.

```
# 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)
s.first = summary(lm.first)
```

Þessi fyrsta tilraun til að máta gögnin sýnir okkur við hverju má búast, hvaða breytur spila ekki lykilhlutverk og

Þetta líkan fær 5659.8623806 í RMSE og 0.8344644 í aðlagað R^2.

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)
s.simple <- summary(lm.simple)
sqrt(mean(residuals(lm.simple)^2))
```

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

Við sjáum að það fær hærra RMSE en fyrsta líkanið okkar.

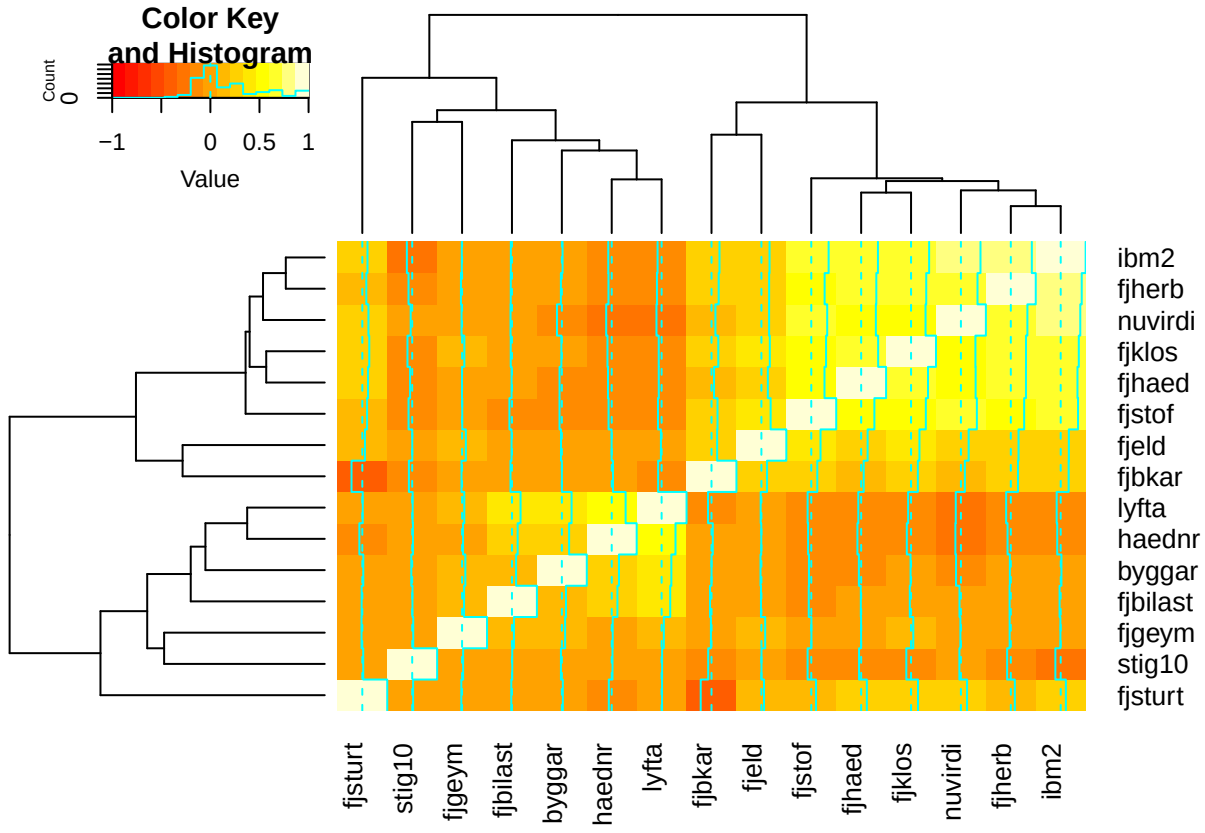
Þetta einfalda líkan fær 8313.261402 í RMSE á prófunarsetti, lægra en á þjálfunarsafni en hvort tveggja mjög hátt.

Hvað við tökum út og hvers vegna

Byrjum á að breyta breytunni fjölda lyfta í tvíundarbreytu sem segir til um hvort það sé lyfta eður ei. Skoðum svo hvaða breytur eru línulega háðar og mega missa sín.

```
# Flestar fasteignir hafa ekki lyftu og örfáar hafa fleiri en eina. Við ákváðum að breyta þeirri breytu
data[ ,"lyfta"] <- data[,"lyfta"]>0

# Skoðum breytur sem eru of líkar, multiple collinearity:
library(gplots)
heatmap.2(cor(data[numericNames]))
```




```
# Sjáum cluster af hópum sem eru mjög líkir. Skoðum eigingildi:
X <- model.matrix(lm(nuvirdi ~ ., data[numericNames]))
eigenX <- eigen(t(X) %*% X)
condNumber <- max(eigenX$values)/min(eigenX$values)
condNumber
```

```
## [1] 2.11158e+12
```

Hér sést að það eru hópur af breytum sem sýna mikla fylgni hvor við aðra og við skoðun á eiginfylkjum XX sést að ástandstalan fyrir fylkið er gríðarhá og greinilegt að eitthvað sé á seyði hér. Við skoðum því eiginvigra með sérlega lág eigingildi.

```
eigenX$values
```

```
##  [1] 1.672145e+09 8.215950e+05 9.069496e+02 2.831165e+02 1.934069e+02
##  [6] 1.463075e+02 7.969591e+01 5.246949e+01 4.484771e+01 3.466555e+01
## [11] 2.186847e+01 1.843604e+01 1.224871e+01 3.196135e+00 7.918929e-04
```

```
# Sjáum að eigingildi 14 er pínkulítið
eigenX$vectors[, 14]
```

```
##  [1] -0.095841569  0.005060443 -0.004653572 -0.102434267 -0.001060296
##  [6]  0.042631732 -0.005476336 -0.014741210 -0.013967580 -0.003118952
## [11]  0.065747546  0.008715209  0.030139569 -0.015213194 -0.986136460
```

```
colnames(data[numericNames])[c(4, 5, 9, 11, 12)]
```

```
## [1] "lyfta"   "ibm2"    "fjsturt" "fjeld"   "fjherb"
```

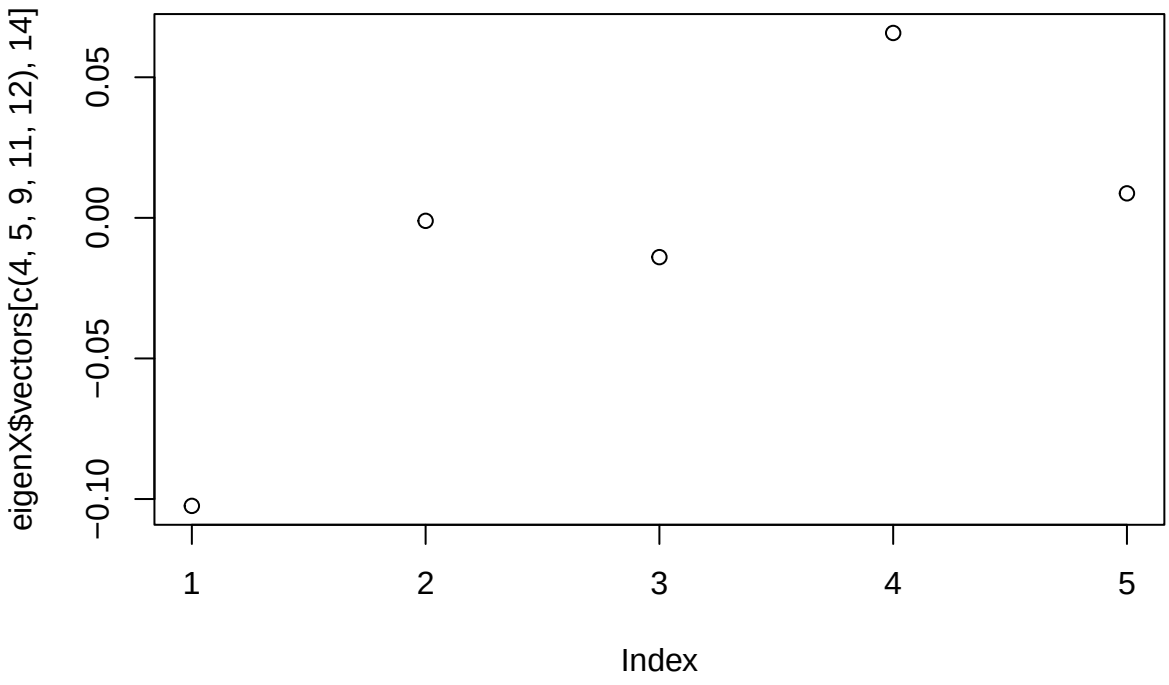
```
sum(eigenX$vectors[c(4, 5, 9, 11, 12), 14])
```

```
## [1] -0.04299939
```

```
sum(eigenX$vectors[,14])
```

```
## [1] -1.090349
```

```
plot(eigenX$vectors[c(4, 5, 9, 11, 12),14])
```



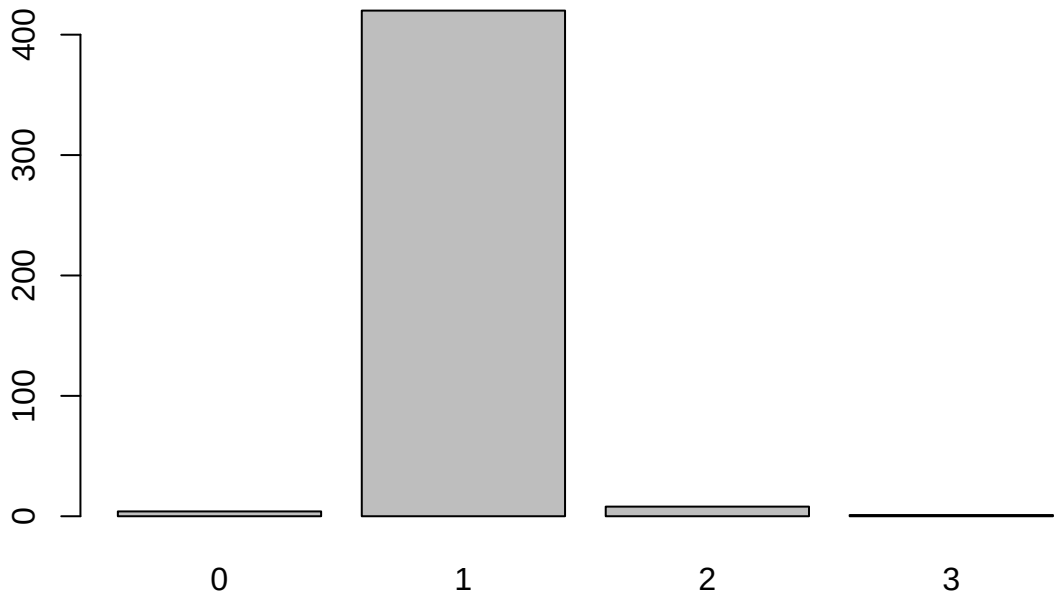

```
bad_actors <- eigenX$vectors[c(4, 5, 9, 11, 12), 14]
sum(bad_actors[c(1,5)]) - sum(bad_actors[c(2,3,4)])
```

```
## [1] -0.1444387
```

Við sjáum að fermetrafjöldi og fjöldi stofa tjá nokkurn veginn sömu upplýsingar og fjöldi herbergja, klósetta og hæða. Við ákveðum því að taka þrjár síðarnefndu út. Þegar við skoðum líkanið sem út úr því kemur sést að RMSE hækkar en aðlagað R^2 gerir það sömuleiðis að örlitlu leyti.

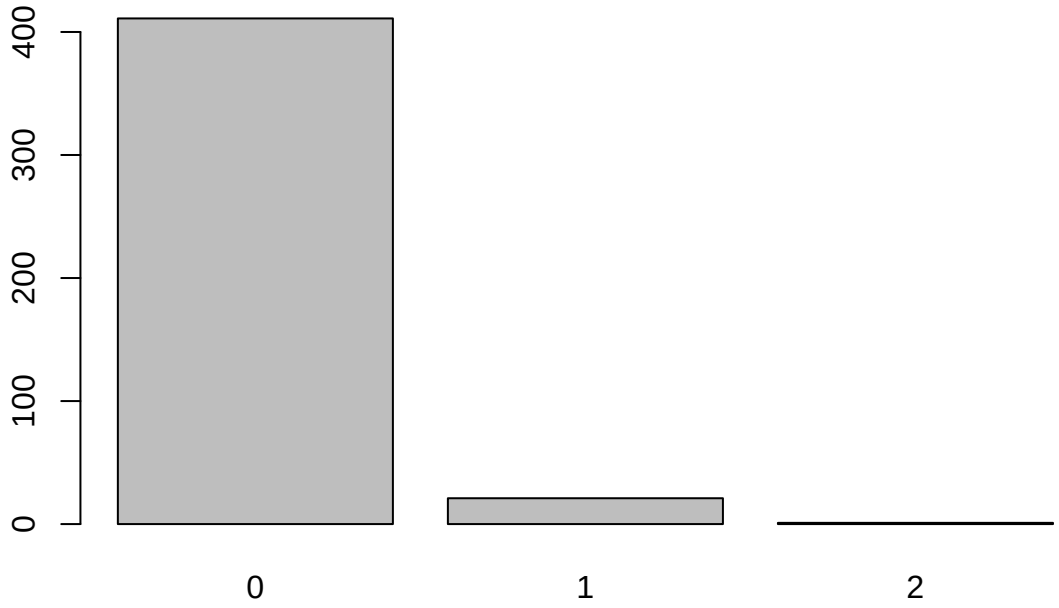
Athugum svo að breyturnar fyrir fjölda eldhúsa og fjölda bílastæða taka nánast sömu gildi í öllum gagnapunktum. Þar að auki eru þær með mjög há p-gildi og við metum það svo að þær megi báðar fjúka. Við það breytist RMSE lítið sem ekkert en aðlagað R^2 hækkar.

```
barplot(table(data$fjeld))
```




```
barplot(table(data$fjbilast))
```

Íbuðartegund	Eignartegund	n
11	Einbylishus	45
11	Ibudareign	19
11	Parhus	6
11	Radhus	13
12	Ibudareign	350



STIG 3: GAGNAÚRVINNSLA, MINNA AUGLJÓSIR FLOKKAR FJARLÆGÐIR

Athugum nú aðrar breytur þar sem ástæður til að fjarlægja þær blasa ekki jafnvel við.

Við sjáum að tegund eignar og íbúðartegund kóða fyrir mjög svipuðum eiginleikum fasteignar og að parhúsarflokkurinn er sá eini í tegund eignar sem mælist með almennilega svörun, mögulega að undanskildum einbýlishúsaflokknum sem er grunnflokkurinn. Þó eru einungis 6 gagnapunktar í parhúsaflokknum og því mögulega ástæða til að fella tegund eignar inn í íbúðartegund. Skoðum hvernig gögnin liggja í þeim flokkum.

```
group_by(data, Íbuðartegund = ibteg, Eignartegund = teg_eign) %>%
  count() %>%
  kbl() %>%
  kable_styling()
```

Hér sést að allar íbúðartegundir nr. 12 eru eignartegundin Íbúðareign. Þó eru nokkrar íbúðareignir sem falla í flokk 11 ásamt öllum hinum tegundum.

undirmatssvaedi	n
0	354
3	7
6	4
21	27
28	31
40	2
48	5
54	3

Eftir samanburð á líkönum sem tóku annars vegar íbúðartegund og eignartegund út þá var það metið svo að betra væri að taka íbúðartegund út. RMSE lækkar og R^2 hækkar, alveg eins og við viljum.

Önnur breyta sem mögulega er að rýra líkanið er undirmatssvæði. Skoðum hvernig gildin liggja þar.

```
group_by(data, undirmatssvaedi) %>%
  count() %>%
  kbl() %>%
  kable_styling()
```

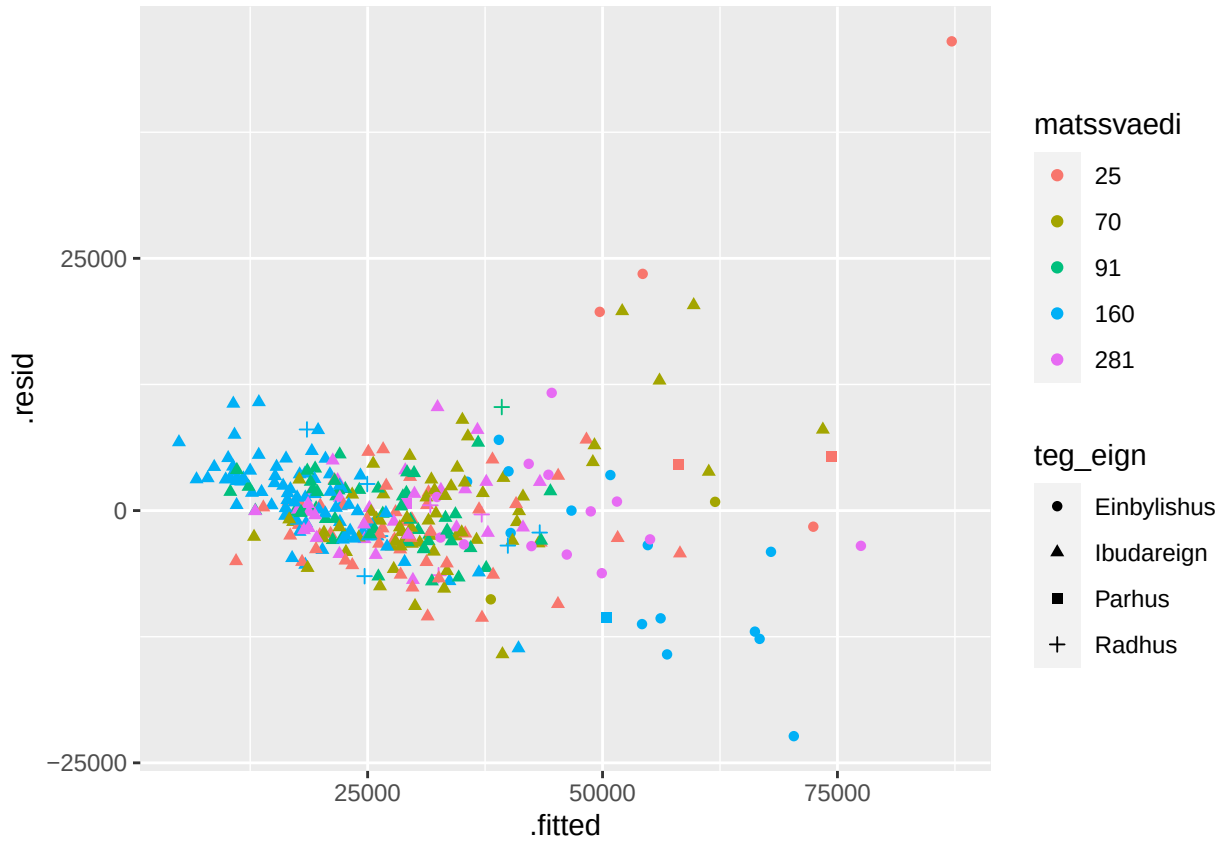
Hér sést að langflestir punktarnir falla í undirmatssvæði 0 og að flestir flokkar innihalda einungis örfáar fasteignir. Einu flokkarnir sem fá lágt p-gildi eru 3 og 6, sem eru Ægissíða og Vesturbær NA við Hringbraut, en einungis 7 og 4 stök falla þar undir. Þar að auki virðast fjölmennustu flokkarnir, nr. 21 og 28 - Vesturberg í Breiðholti og Blokkir við Kringlumýra- og Miklubraut -, skipta litlu máli.

Ástæða til að sleppa þessum flokk í heild sinni?

STIG 4A: HELDUR LÍNULEIKI? GRUNNSKOÐUN

Byrjum á að skoða eitt mikilvægasta plottið, leifð á móti spá.

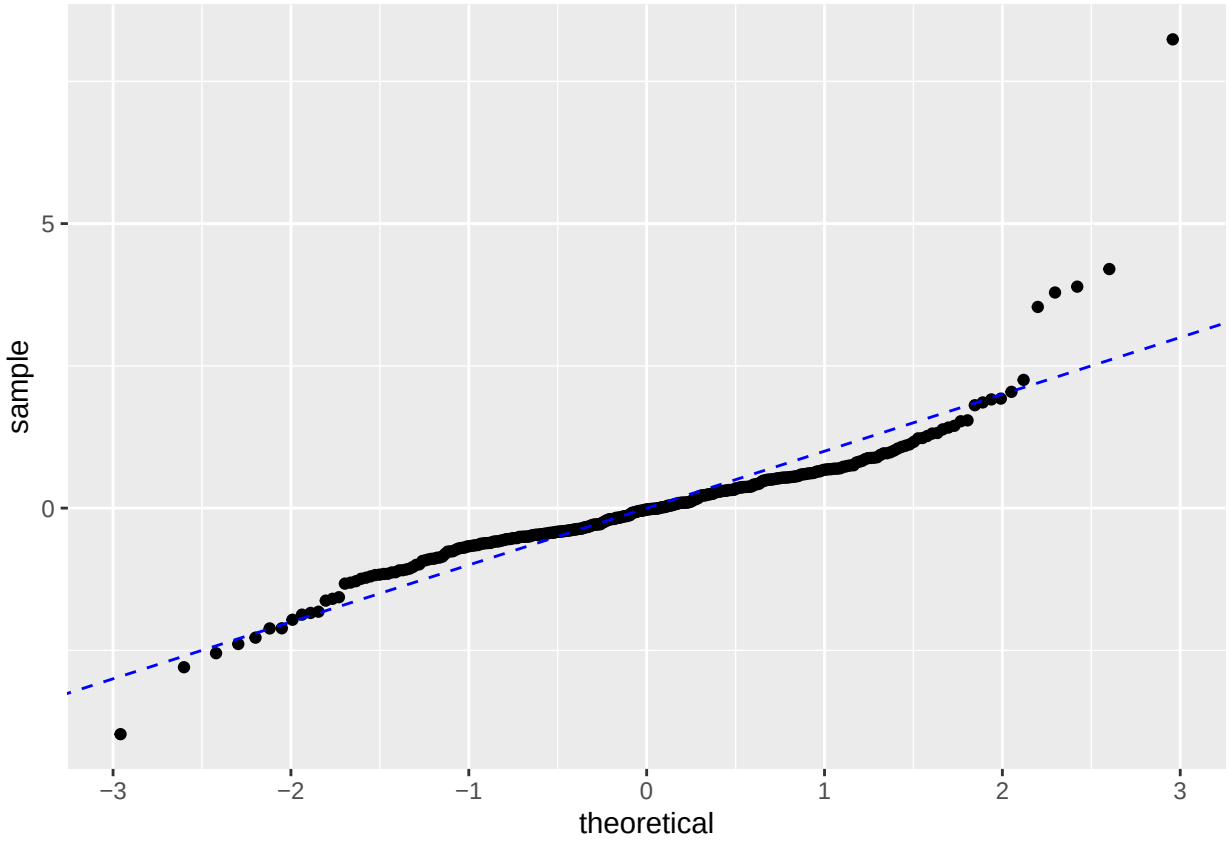
```
fortData <- fortify(lm.fourth)
fortData %>%
  ggplot(aes(x = .fitted, y = .resid, color = matssvaedi, shape = teg_eign)) +
  geom_jitter(width = 0.25)
```



Hér er augljóslega tilfelli af hederoskedaticity, þ.e. leifðin eykst með hærra spágildi.Við ættum að geta séð þetta vel líka með QQ-plotti af leifðinni:

```
tibble(Normal = fortData$.stdresid) %>%
  gather(type, val) %>%
  ggplot(aes(sample = val)) +
  stat_qq() +
  geom_abline(slope = 1, intercept = 0, lty = 2, col = 'blue')
```

```
## Warning: Removed 1 rows containing non-finite values (stat_qq).
```



Gögnin halda að miklu leyti í við y = x línuna en þó er einn punktur alveg kú-kú og tilhneigingin er samhverf sem bendir til þess að hér sé eitthvað annað en línulegt í gangi.

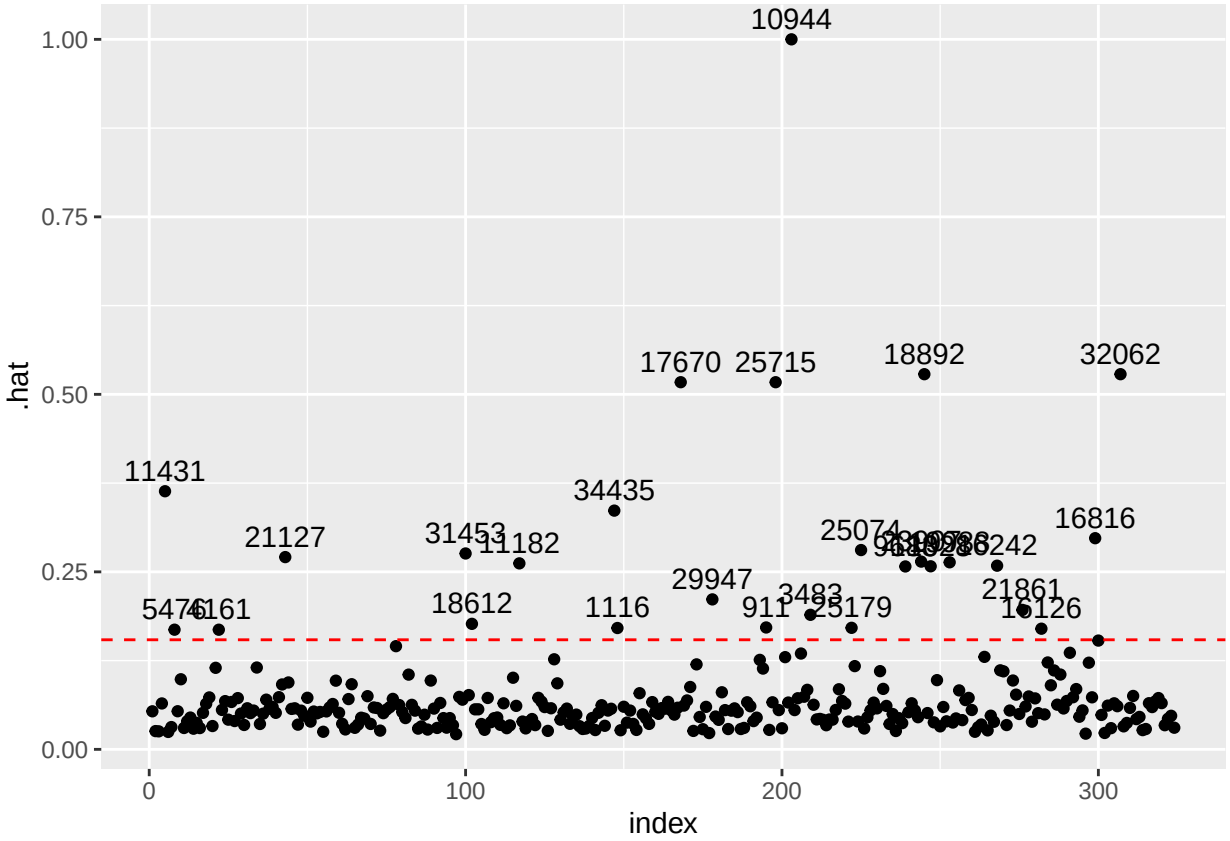
Af þessu tvennu að ofan drögum við þá ályktun að líklegt sé að við viljum umbreyta y (XX skýribreyta? man ekki orðin). Réttast er þó að skoða fyrst útlaga og áhrifamikla punkta vegna þess að BoxCox og aðrar aðferðir eru sérstaklega næmar fyrir slíku.

STIG 5: ÚTLAGAR OG MIKILVÆGIR PUNKTAR

Greinum nú x-punkta með Mahalanobis/H_ii gildum: Skoðum matið okkar á y (y.hat)

```
fortData$rn <- row.names(fortData)
fortData$index <- 1:nrow(fortData)
fortData$.jackknife <- rstudent(lm.fourth)
n <- nrow(fortData)
p <- nrow(summary(lm.fourth)$coefficients)

fortData %>%
  ggplot(aes(x = index, y = .hat)) +
  geom_point() +
  geom_hline(yintercept = 2*p/n, lty = 2, col = 'red') +
  geom_text(aes(label = ifelse(.hat > 2*p/n, rn, '')), hjust = 0.5, vjust = -0.5)
```




```
# Þetta er bara rugl að skoða, við þurfum að finna betri leið... Notum töflu:

fortData %>%
  filter(.hat > 2*p/n) %>%
  arrange(desc(.hat)) %>%
  dplyr::select(rn, .hat) %>%
  kbl(align = 'c') %>%
  kable_styling()
```

Athugum að hér er einn punktur sem virðist gríðarlega áhrifamikill. Skoðum hann.

```
fortData[fortData$rn == 10944,]
```

```
##       nuvirdi     kdagur    teg_eign byggar haednr lyfta  ibm2 fjbkar fjsturt
## 10944   46691 2013-09-09 Einbylishus   1977      1     0 289.3      2       0
##       fjstof fjgeym stig10 matssvaedi undirmatssvaedi .hat   .sigma .cooksd
## 10944      3      0    9.7        160               0    1 5932.823     NaN
##       .fitted        .resid .stdresid    rn index .jackknife
## 10944   46691 -3.364611e-09       NaN 10944   203        NaN
```

Það sést að þessum punkti er spáð nákvæmlega réttu gildi og því er hann veigamikill. Hann er sem sagt veigamikill vegna þess hve spáin var nákvæm og því ætti ekki að skoða hann frekar. Aðrir gagnapunktar eru innan marka.

Byrjum á því að greina y-punkta með jackknife og Cook's distance:

rn	.hat
10944	1.0000000
32062	0.5283702
18892	0.5283702
17670	0.5170997
25715	0.5170997
11431	0.3635291
34435	0.3362486
16816	0.2974163
25074	0.2806874
31453	0.2758708
21127	0.2707684
23907	0.2644198
19988	0.2634611
11182	0.2619441
16242	0.2586753
11028	0.2578057
9683	0.2575883
29947	0.2112071
21861	0.1963594
3483	0.1895068
18612	0.1767496
911	0.1716384
25179	0.1711550
1116	0.1708913
16126	0.1699198
5476	0.1685384
4161	0.1685030