OLA 1

Claes Lindholm

cph-cl303

1.1 Jeg har valgt følgende to sagsnumre i det scrapede datasæt: 56103478 og 64202778 (hhv index 100 & 200 i min dataframe).

1.2 Rows 1 & 117.

1.3 Ikke forstået: Første column (\_id) er EDC sagsnummeret hvilket burde være den bedste måde at identificere hver enkelt ejendom. Alternativet ville være den fulde adresse men mener sagsnr er bedre. Alle 1940 værdier i \_id er unikke. Værdier for mdl ejerudg, kvm pris og liggetid er NA for en del rækker, ved stikprøvekontrol kan jeg se at disse data findes på flere af disse huse på edc.dk, der er sandsynligvis en fejl i scarping koden.

1.4 Det ser ud som om alle data om de enkelte ejendomme er forsøgt scrapet. Opgaven ville sandsynligvis have været lettere hvis man inden da havde besluttet sig for, hvilke mulige korrelationer man ville finde og derefter kun havde valgt relevante features. Column ”Energimærke” er fejlet ved scraping, den burde indeholde en char mellem A og G men indeholder en numerisk værdi (sandsynligvis et årstal).

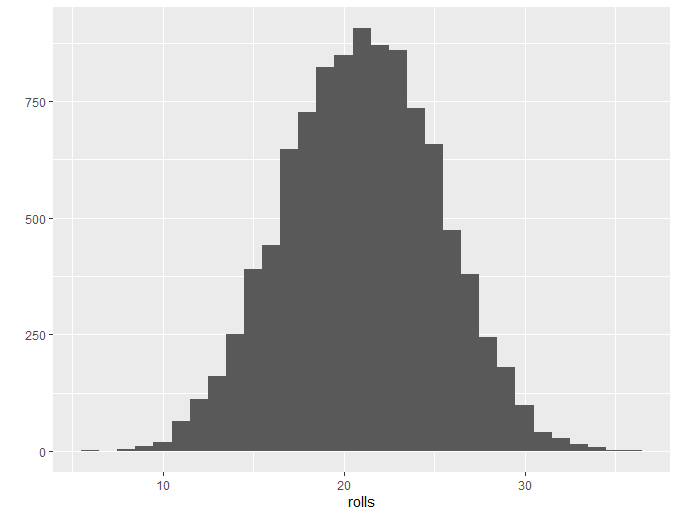
2.1

3.1 > tabulate(dice)

[1] 4120 4198 4231 4184 4085 4182

Sandsynlighed for en 5’er: 25000/4085 \*100 = 16,34%

3.2 Resultatet er logisk, summen af de oftest forekommende kombinationer er de mest hyppige:



3.3 De små (tilfældige) afvigelser fra en helt symmetrisk fordeling er forsvundet da antallet af forsøg er øget gange.

