HW8

Łukasz Tomaszewski

Przygotowanie ramki danych i wstępna analiza

Na początku sprawdźmy rozmiar naszej ramki danych oraz jakie kolumny ona zawiera.

```
dim(df)
```

```
## [1] 1000   18
```

```
colnames(df)
```

```
##  [1] "bedrooms"      "bathrooms"     "sqft_living"   "sqft_lot"
##  [5] "floors"        "waterfront"    "view"          "condition"
##  [9] "grade"         "sqft_above"    "sqft_basement" "yr_built"
## [13] "yr_renovated"  "zipcode"       "lat"           "long"
## [17] "sqft_living15" "price"
```

Sprawdźmy czy są jakieś braki w danych.

```
df[rowSums(is.na(df))>0,]
```

```
##      bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition
## 1000       NA        NA          NA       NA     NA         NA   NA        NA
##      grade sqft_above sqft_basement yr_built yr_renovated zipcode lat long
## 1000    NA         NA            NA       NA           NA      NA  NA   NA
##      sqft_living15 price
## 1000            NA    NA
```

Okazuje się, że ostatni wiersz naszej ramki jest pusty, możemy go więc usunąć.

Nasza ramka zawiera także jedną posiadłość, która nie ma pokoi ani łazienek oraz jest to jedyna posiadłość, która ma 3,5 pięter. Możliwe, że jest to błąd, więc pozbywamy się tego wiersza.

```
df %>% filter(bedrooms==0)
```

```
##   bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition
## 1        0         0        3064     4764    3.5          0    2         3
##   grade sqft_above sqft_basement yr_built yr_renovated zipcode     lat     long
## 1     7       3064             0     1990            0   98102 47.6362 -122.322
##   sqft_living15 price
## 1          2360   110
```

Sprawdźmy, czy w każdym wierszu suma powierzchni pod ziemią i nad ziemią równa się całkowitej powierzchni.

```
df %>% filter(sqft_living == sqft_above+sqft_basement) %>% dim()
```

```
## [1] 998  18
```

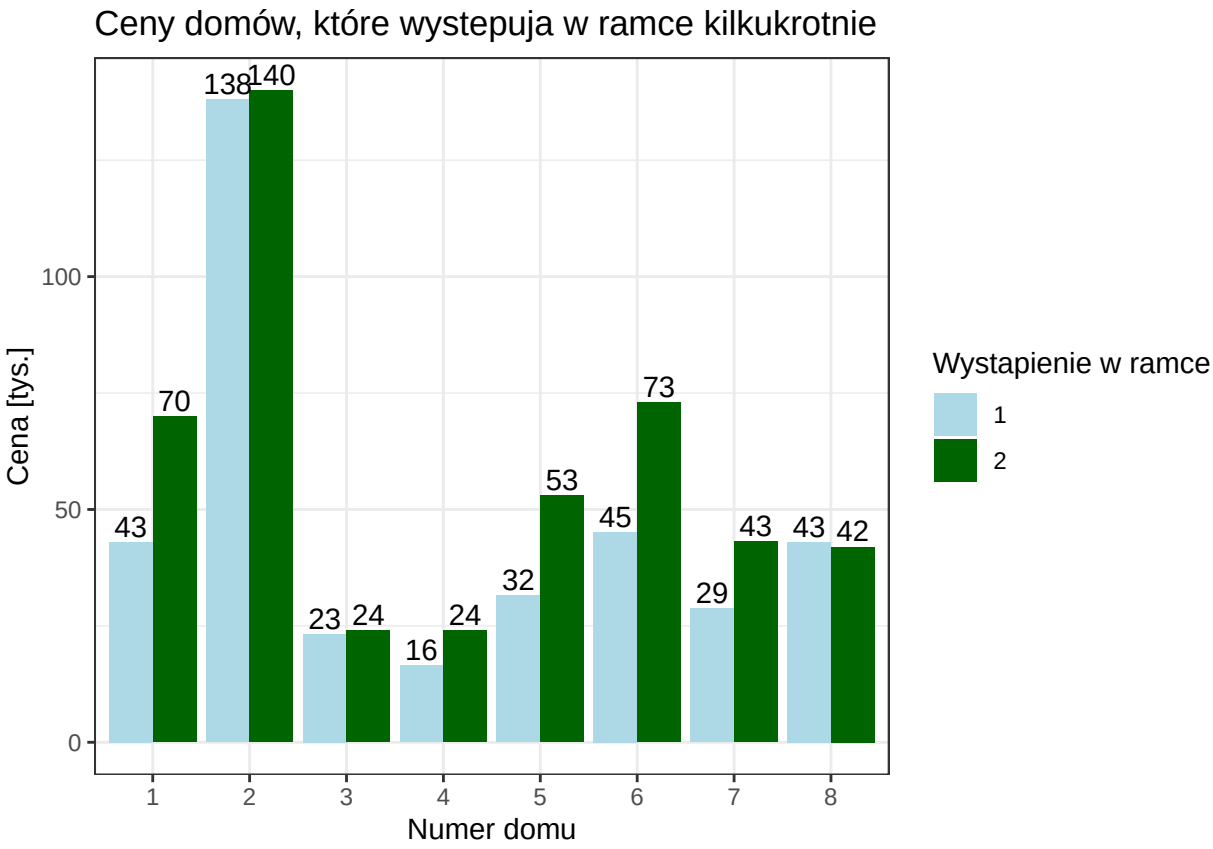
Wszystko się zgadza.

W naszej ramce występują kilkukrotnie posiadłości o jednakowych parametrach (jest ich dokładnie 8). Prawdopodobnie są to te same domy, które zostały sprzedane więcej niż jeden raz.

```
df %>% group_by(bedrooms,bathrooms,sqft_living,sqft_lot,floors,lat,long, yr_built)%>%
  mutate(n = n()) %>% filter(n!=1)
```

```
## # A tibble: 16 x 19
## # Groups:   bedrooms, bathrooms, sqft_living, sqft_lot, floors, lat, long,
## #   yr_built [8]
##    bedrooms bathrooms sqft_living sqft_lot floors waterfront  view condition
##       <int>     <dbl>       <int>    <int>  <dbl>      <int> <int>     <int>
##  1        3      1.5         1580     5000      1          0     0         3
##  2        3      1.5         1580     5000      1          0     0         3
##  3        4      3.25        4290    12103      1          0     3         3
##  4        4      3.25        4290    12103      1          0     3         3
##  5        2      1           1240    12092      1          0     0         3
##  6        2      1           1240    12092      1          0     0         3
##  7        4      1           1000     7134      1          0     0         3
##  8        4      1           1000     7134      1          0     0         3
##  9        4      2.25        2180    10754      1          0     0         5
## 10        4      2.25        2180    10754      1          0     0         5
## 11        6      2.25        2660    13579      2          0     0         3
## 12        6      2.25        2660    13579      2          0     0         3
## 13        3      1           1810     7200      1          0     0         4
## 14        3      1           1810     7200      1          0     0         4
## 15        2      1.75        1350     4003      1          0     0         3
## 16        2      1.75        1350     4003      1          0     0         3
## # ... with 11 more variables: grade <int>, sqft_above <int>,
## #   sqft_basement <int>, yr_built <int>, yr_renovated <int>, zipcode <int>,
## #   lat <dbl>, long <dbl>, sqft_living15 <int>, price <dbl>, n <int>
```

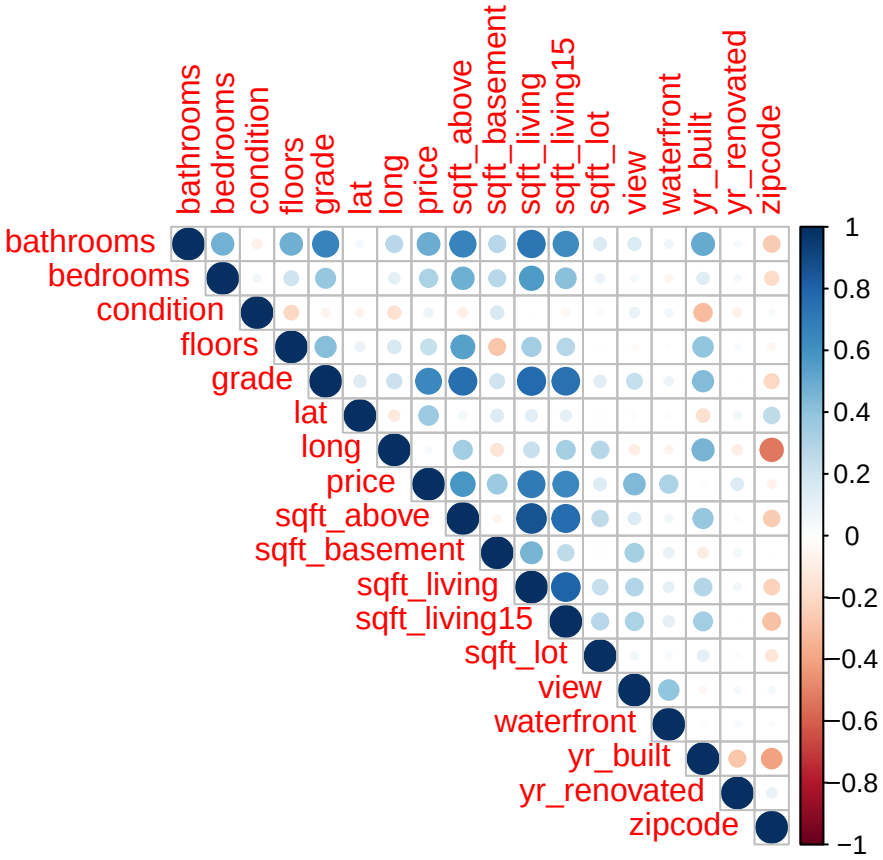
Sprawdźmy jak zmieniała się ich cena podczas kolejnych sprzedaży.


Ceny domów, które wystepuja w ramce kilkukrotnie
Cena [tys.]
0
50
100
1
2
3
4
5
6
7
8
Numer domu
43
70
138
140
23
24
16
24
32
53
45
73
29
43
43
42
Wystapienie w ramce
1
2

Analiza

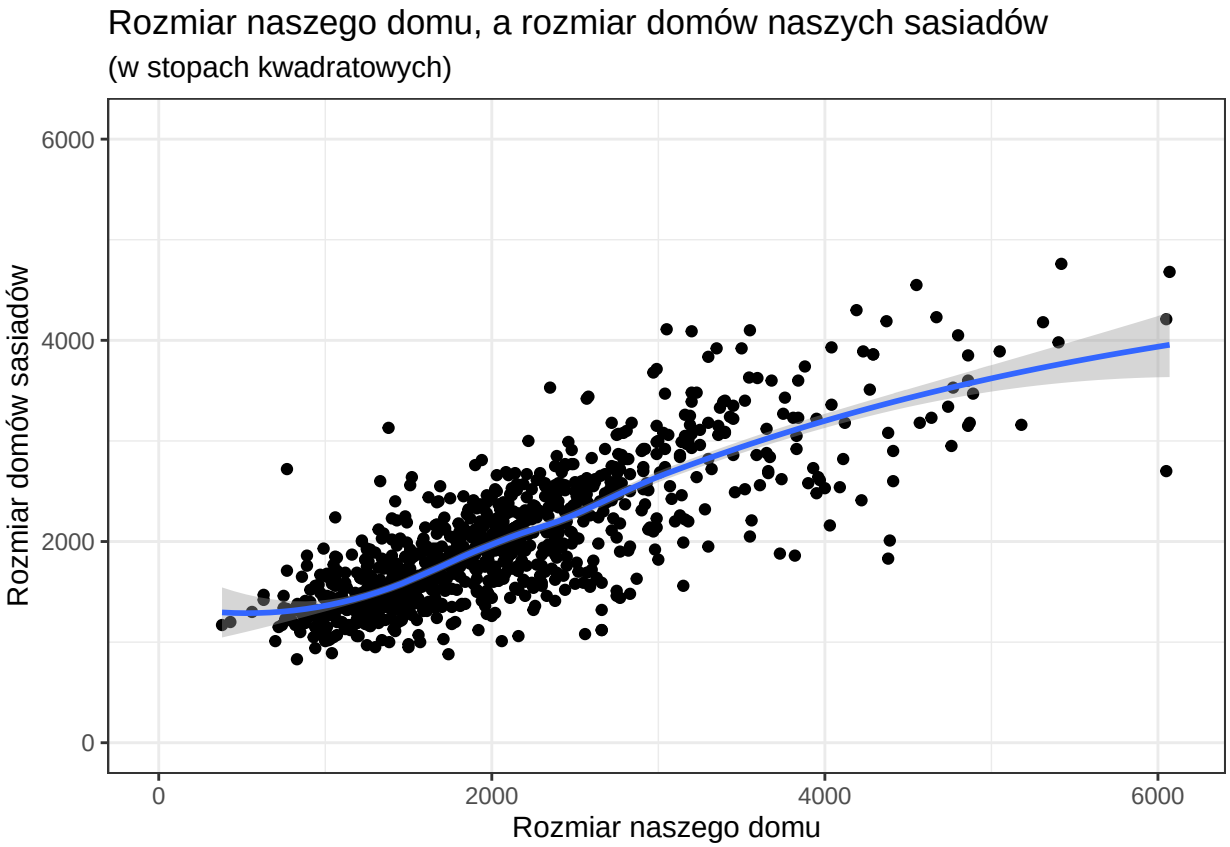
Korelacja

Sprawdźmy jak korelują ze sobą zmienne.



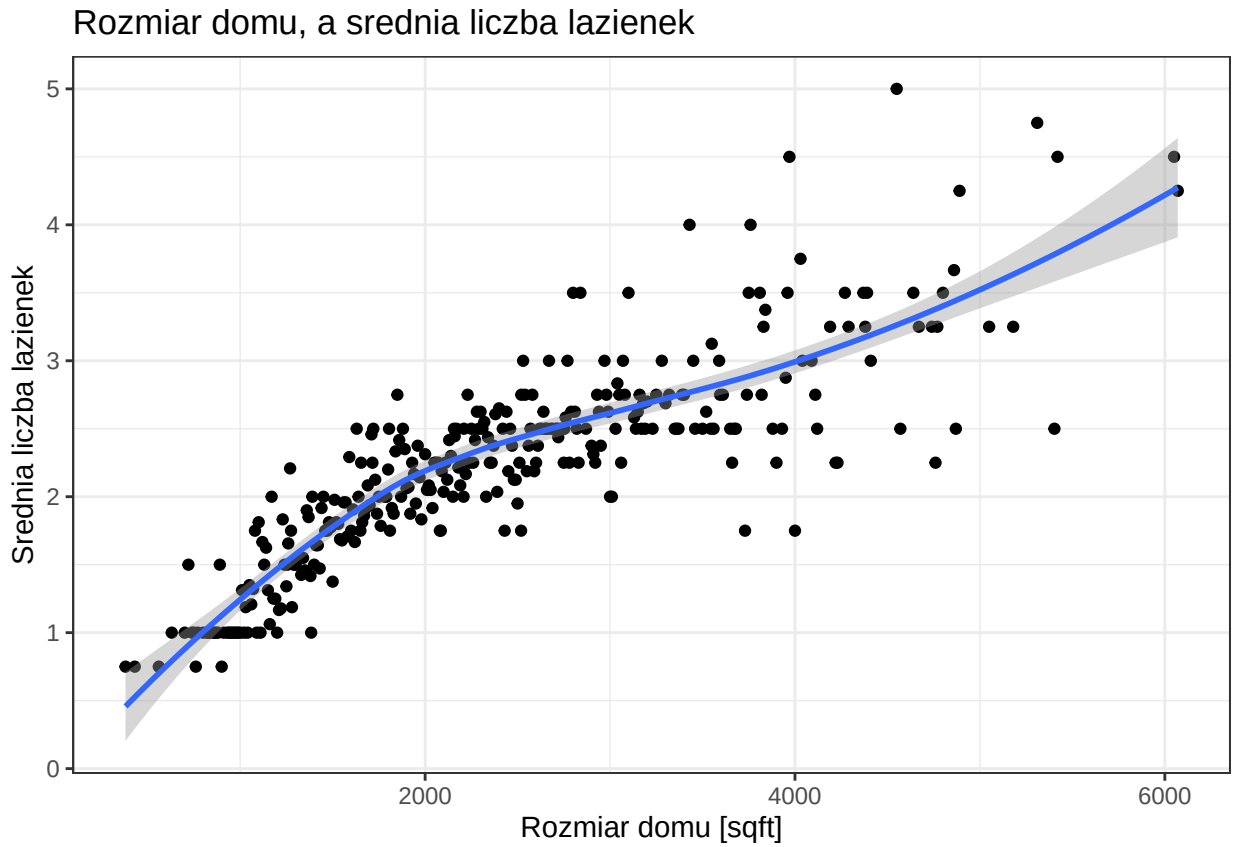
Możemy zauważyć, że ocena posiadłości silnie koreluje z liczbą łazienek, z powierzchnią do życia posiadłości, z powierzchnią do życia posiadłości sąsiadów, z powierzchnią nad ziemią oraz z ceną. Sprawdźmy więc jak mają się do siebie te zmienne.

Jak różnią się domy naszych sąsiadów, względem naszego



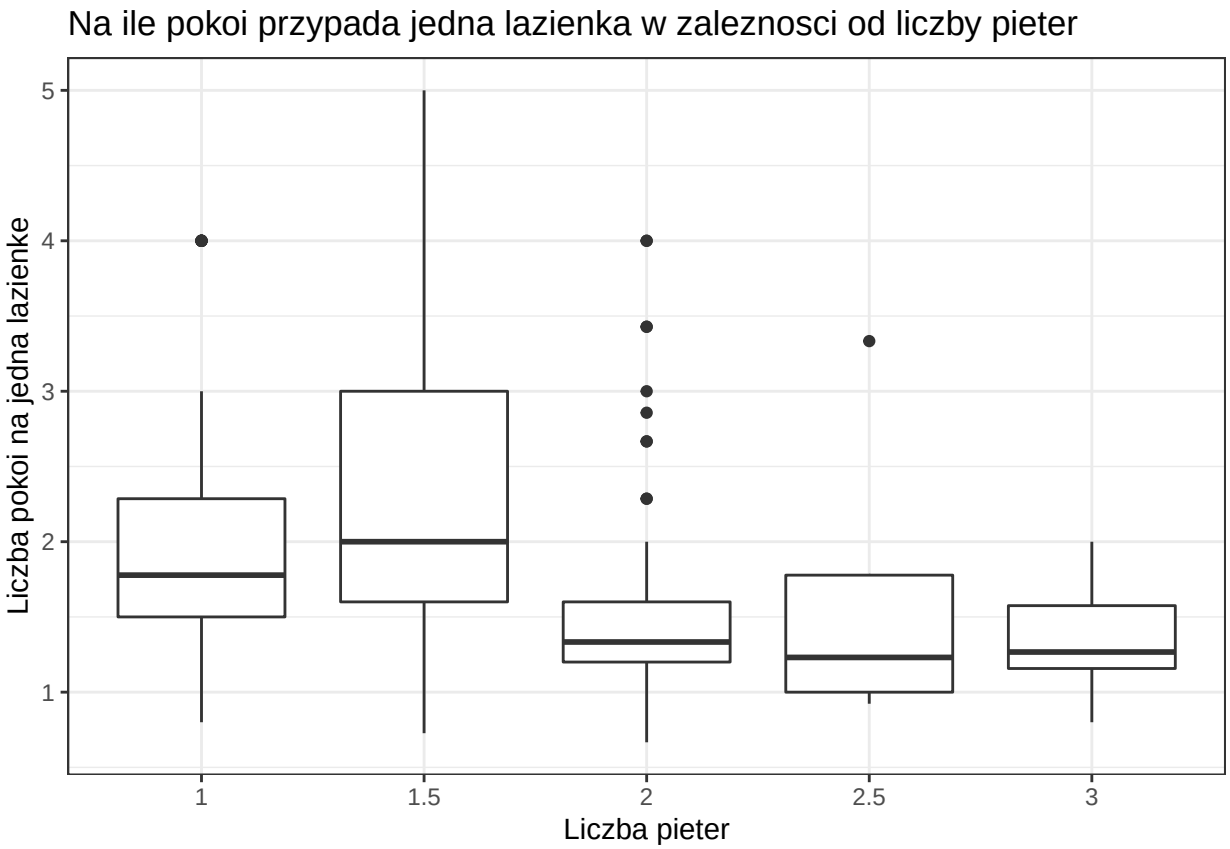
Z wykresu możemy wywnioskować, że wraz ze wzrostem naszego domu, rosną też domy naszych sąsiadów. Jeżeli rozmiar naszego domu jest mały, to sąsiedzi będą mieszkać w domach większych. Natomiast jeżeli nasz dom jest duży, to domy naszych sąsiadów będą mniejsze.

Rozmiar naszego domu, a ilość łazienek



Jak można było sie spodziewać, im większy dom tym więcej łazienek.

Sprawdźmy jeszcze ile pokoi obłsuguje jedna lazienka, w zależności od liczby pięter.

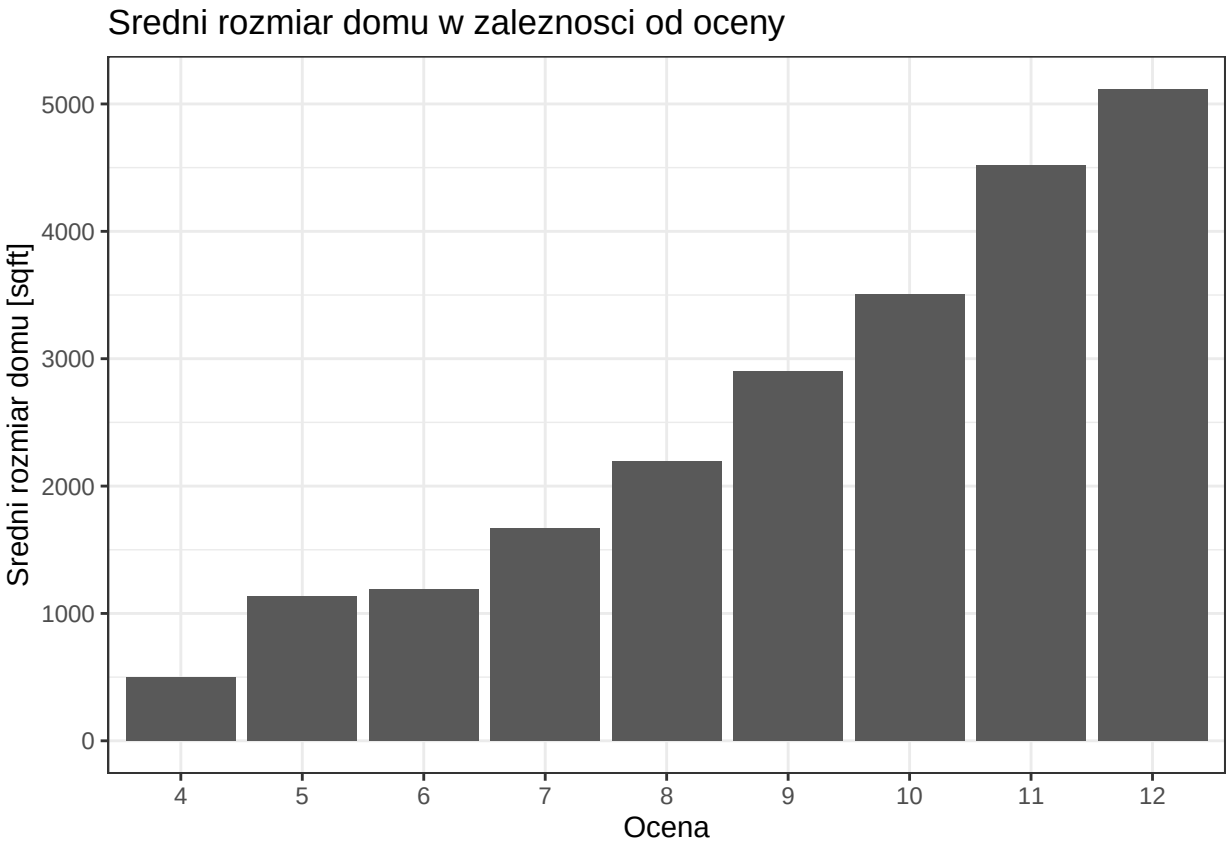


Z wykresu wynika, że najmniej pokoi obsługuje jedna łazienka, gdy mieszkamy w domu o większej ilości pięter. Najniższa mediana obsługiwanych pokoi występuje w przypadku, gdy mieszkamy w domu, który ma 2.5 piętra.

Jaki procent domu znajduje się nad ziemią

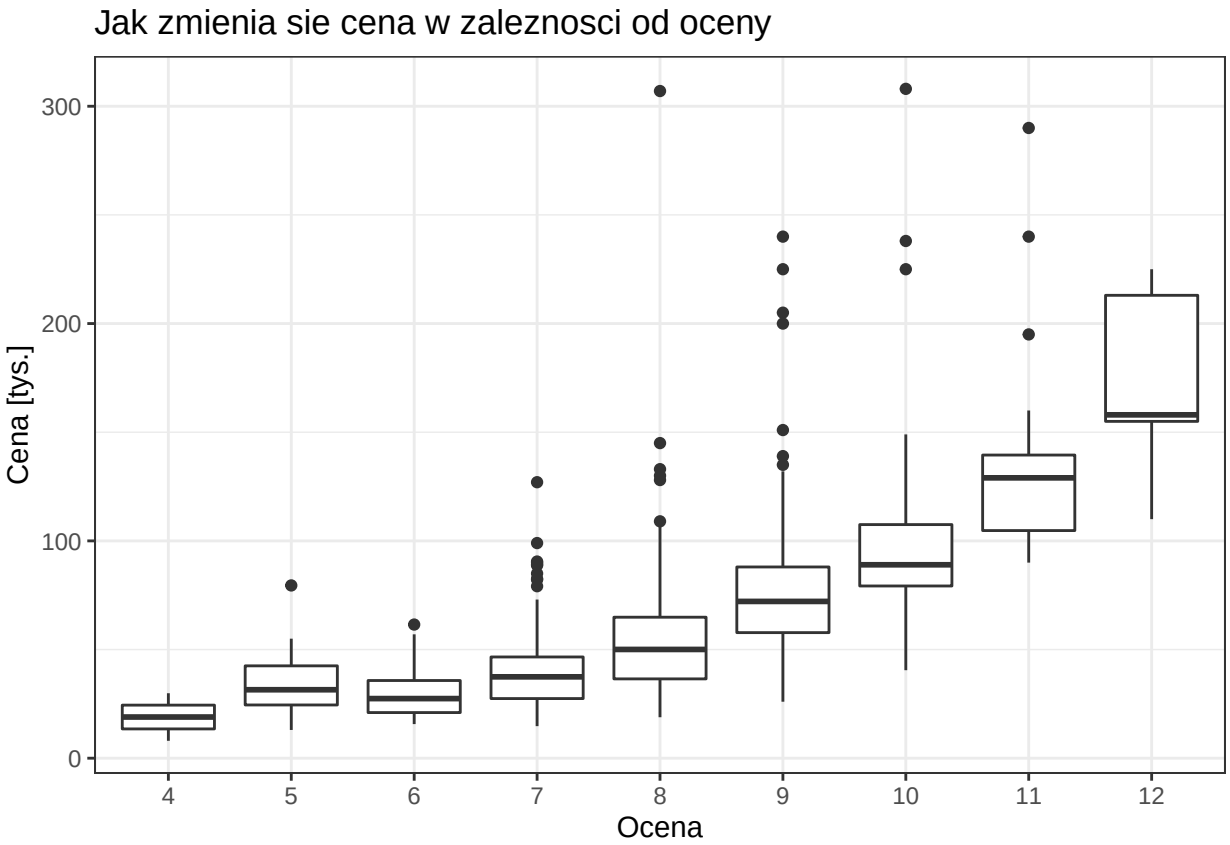
Z wykresu wynika, że najwięcej jest domów (i to znacząco), których powierzchnie do życia stanowi w prawie 100% powierzchnia znajdująca się nad ziemią.

Jakich rozmiarów domy, mają jaką ocenę



Jak możemy zauważyć na powyższym wykresie, większe domy mają lepszą ocenę. Średni rozmiar domu rośnie wraz ze wzrostem oceny.

Jak zmienia sie cena w zależności od oceny



Tutaj, bez większych zaskoczeń, im większa ocena, tym większa cena. Interesujące jest natomiast to, że domy o największej cenie, wcale nie miały najwyższej oceny, lecz odpowiednio ocenę 8 i 10. Okazuje się także, że cena domów o ocenie 6 jest mniejsza niż cena domów o oczenie 5 i 7.

Wnioski

Zmienne takie jak powierzchnia do życia domów sąsiadów, liczba łazienek, czy powierzchnia znajdująca się nad ziemią są silnie powiązane z powierzchnią do życia naszego domu. Jeżeli rośnie powierzchnia naszego domu, to rosną wymienione zmienne. Możemy więc powiedzieć, że ocena domu zależy głównie od jego powierzchni. Im lepsza ocena, tym większy jest przeciętny dom. Wraz z oceną rośnie także cena. Są jednak od tej reguły wyjątki i najdroższe domy wcale nie mają najlepszej oceny.