

HW8

Mikołaj Roguski

30 01 2022

Analiza zbioru 3 - housecsv

link do danych: <https://www.kaggle.com/mohamedbakrey/housecsv>

Załadowanie bibliotek oraz danch

```
library(dplyr);  
library(visdat);  
library(ggplot2);  
library(tidyverse);  
library(dlookr);  
library(corrplot);  
  
data <- read.csv("houses.csv");
```

Informacje o danych

Wymiary

```
dim(data)
```

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

Źródłowe dane są ramką o 18 kolumnach i 1000 wierszach.

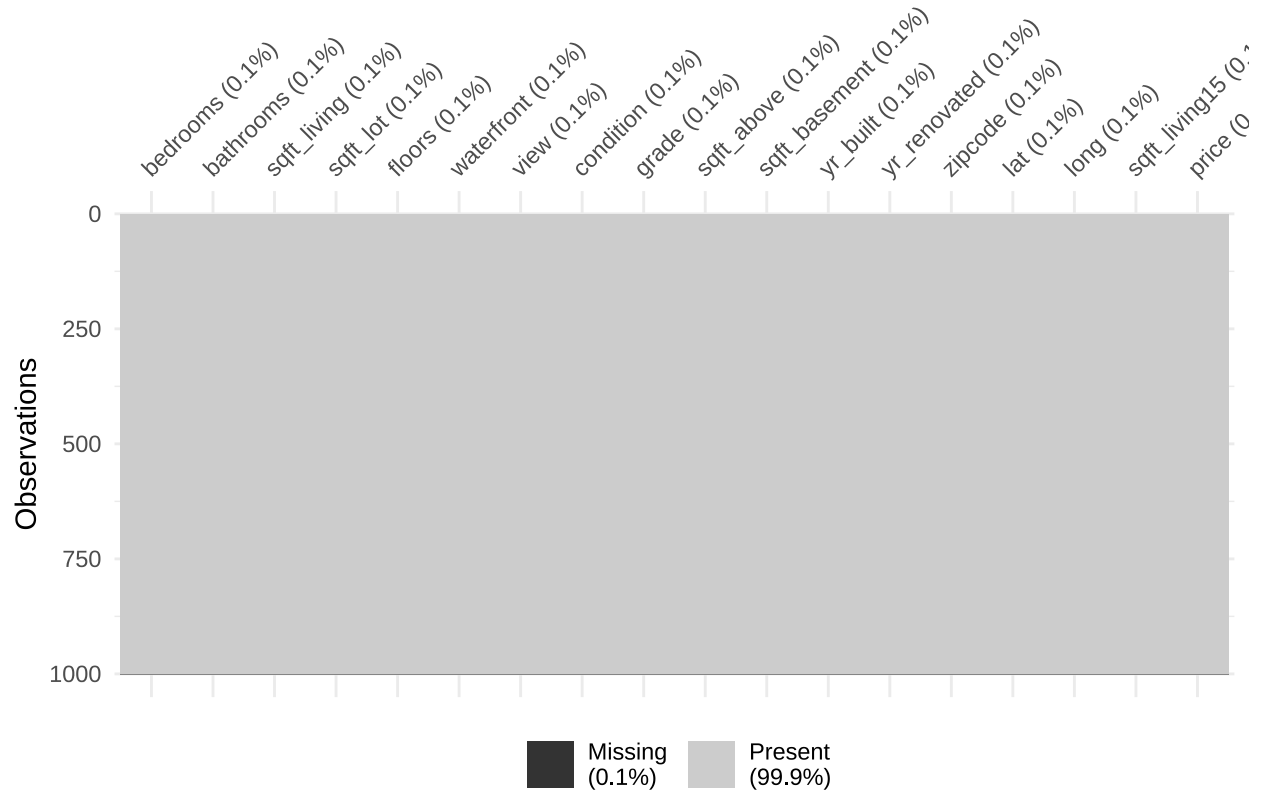
Nazwy kolumn

```
colnames(data)
```

```
## [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"
```

Brakujące dane

```
vis_miss(data)
```



Wygląd danych

```
head(data)
```

```
## bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition
## 1      3          1.00         1180    5650      1           0    0          3
## 2      3          2.25         2570    7242      2           0    0          3
## 3      2          1.00          770   10000      1           0    0          3
## 4      4          3.00         1960    5000      1           0    0          5
## 5      3          2.00         1680    8080      1           0    0          3
## 6      4          4.50         5420   101930     1           0    0          3
## grade sqft_above sqft_basement yr_built yr_renovated zipcode    lat    long
## 1      7         1180           0     1955           0   98178 47.5112 -122.257
## 2      7         2170          400     1951          1991  98125 47.7210 -122.319
## 3      6          770           0     1933           0   98028 47.7379 -122.233
## 4      7         1050          910     1965           0   98136 47.5208 -122.393
## 5      8         1680           0     1987           0   98074 47.6168 -122.045
## 6     11         3890         1530     2001           0   98053 47.6561 -122.005
```

```
##      sqft_living15  price
## 1          1340  22.19
## 2          1690  53.80
## 3          2720  18.00
## 4          1360  60.40
## 5          1800  51.00
## 6          4760 123.00
```

```
tail(data)
```

```
##      bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition
## 995          2         1.00         740     6460    1.0          0    0          3
## 996          4         2.50        1860     6325    2.0          0    0          4
## 997          2         2.75        1590    20917    1.5          0    0          3
## 998          2         1.00         850     2340    1.0          0    0          3
## 999          2         1.00        1030     4188    1.0          0    0          3
## 1000         NA          NA          NA        NA    NA          NA    NA          NA
##      grade sqft_above sqft_basement yr_built yr_renovated zipcode      lat
## 995      6         740           0     1953          0    98146  47.5077
## 996      7        1860           0     1991          0    98038  47.3492
## 997      5        1590           0     1920          0    98001  47.2786
## 998      7         850           0     1922          0    98105  47.6707
## 999      8        1030           0     1981          0    98038  47.3738
## 1000     NA          NA          NA        NA          NA    NA          NA
##      long sqft_living15  price
## 995 -122.344          1170 17.850
## 996 -122.030          1860 29.100
## 997 -122.250          1310 19.995
## 998 -122.328          1300 55.350
## 999 -122.057          1450 18.995
## 1000      NA          NA      NA
```

Widzimy, że brakuje tylko ostatniego wiersza danych.

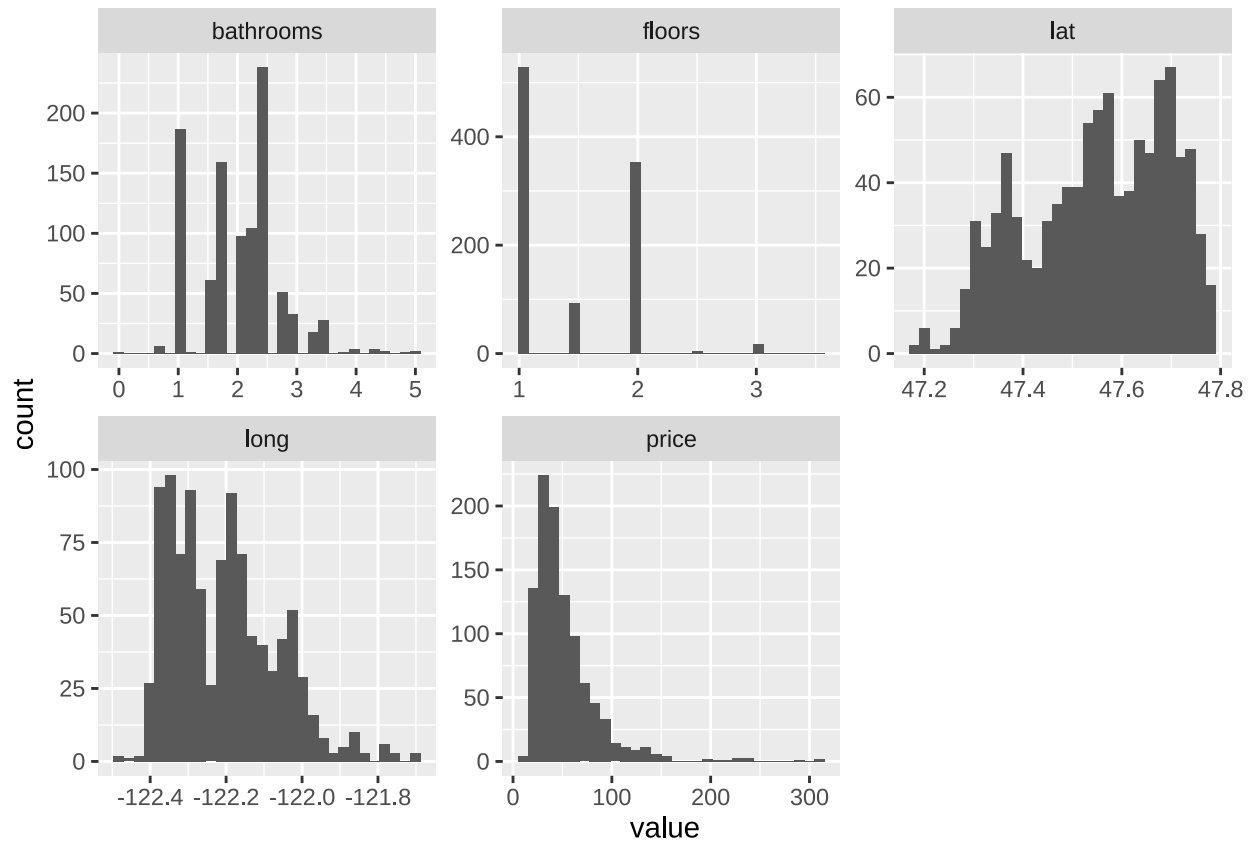
Typy zmiennych

```
sapply(data, class)
```

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

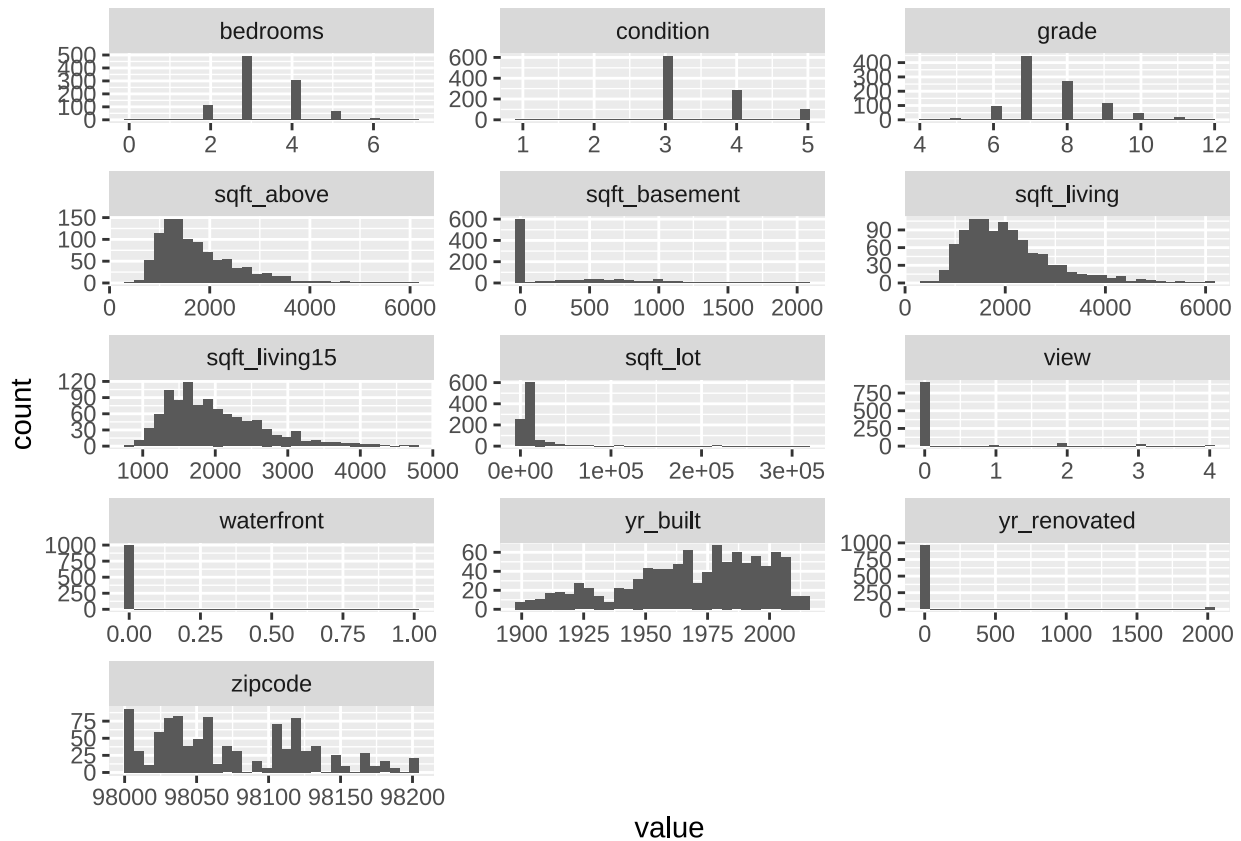
Rozkłady zmiennych typu double

```
data %>%
  select_if(is.double) %>%
  gather() %>%
  ggplot(aes(value)) +
  geom_histogram() +
  facet_wrap(~key,scales = "free")
```



Rozkłady zmiennych typu integer

```
data %>%
  select_if(is.integer) %>%
  gather() %>%
  ggplot(aes(value)) +
  geom_histogram() +
  facet_wrap(~key,scales = "free",ncol = 3,shrink = FALSE)
```



Opisy column

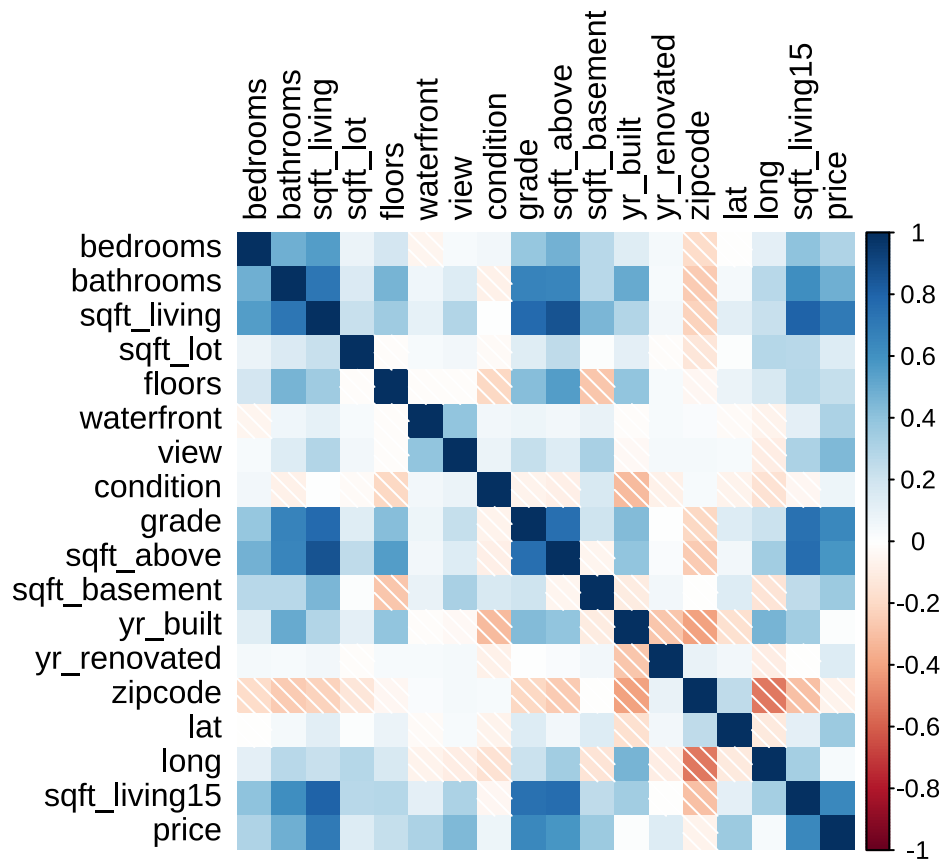
```
sapply(data, summary)
```

```
##      bedrooms bathrooms sqft_living sqft_lot  floors waterfront  view
## Min.    0.000000    0.000000   380.000   649.00  1.000000 0.000000000 0.0000000
## 1st Qu.  3.000000    1.500000  1405.000   5419.00  1.000000 0.000000000 0.0000000
## Median  3.000000    2.000000   1900.000   8040.00  1.000000 0.000000000 0.0000000
## Mean    3.349349    2.045796   2051.397  14707.24  1.446947 0.008008008 0.2372372
## 3rd Qu.  4.000000    2.500000   2475.000  11508.50  2.000000 0.000000000 0.0000000
## Max.    7.000000    5.000000   6070.000 315374.00  3.500000 1.000000000 4.0000000
## NA's    1.000000    1.000000     1.000     1.00  1.000000 1.000000000 1.0000000
##
##      condition      grade sqft_above sqft_basement yr_built yr_renovated
## Min.    1.000000    4.000000   380.000      0.0000  1900.00      0.00000
## 1st Qu.  3.000000    7.000000   1190.000      0.0000  1952.00      0.00000
## Median  3.000000    7.000000   1540.000      0.0000  1974.00      0.00000
## Mean    3.464464    7.605606   1750.233    301.1642  1969.03     81.83083
## 3rd Qu.  4.000000    8.000000   2135.000     580.0000  1992.00      0.00000
## Max.    5.000000   12.000000   6070.000   2060.0000  2015.00   2014.00000
## NA's    1.000000    1.000000     1.000     1.0000     1.00      1.00000
##
##      zipcode      lat      long sqft_living15  price
## Min.    98001.00  47.17750 -122.4900     830.000   8.00000
## 1st Qu.  98032.00  47.44300 -122.3225    1490.000  30.98000
## Median  98058.00  47.56360 -122.2180    1850.000  43.50000
```

```
## Mean    98074.44 47.54972 -122.2074    1986.814  52.07145
## 3rd Qu. 98116.00 47.67340 -122.1180    2360.000  63.44625
## Max.    98199.00 47.77760 -121.7090    4760.000 308.00000
## NA's      1.00  1.00000   1.0000      1.000  1.00000
```

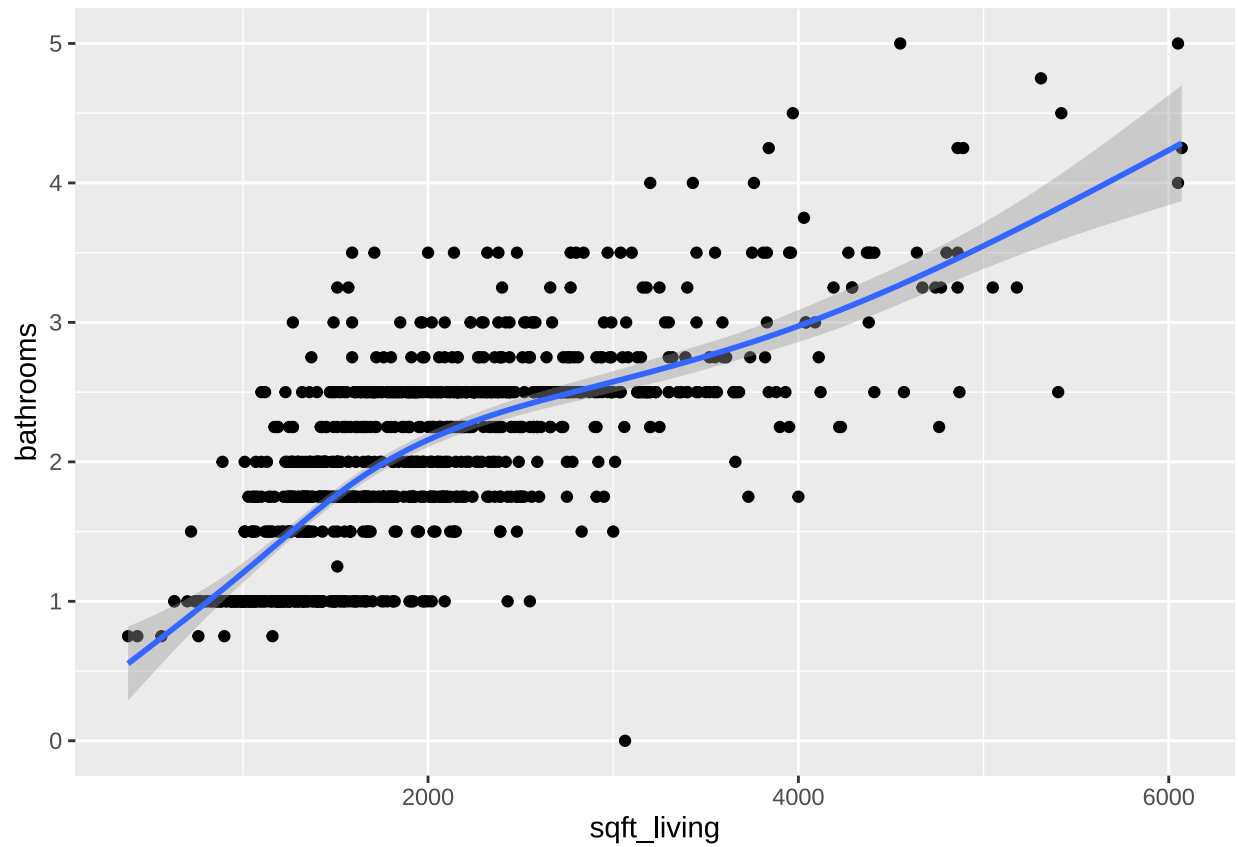
Związki między zmiennymi

```
corrplot(cor(data, use = "complete.obs"),
  method = "shade",
  type = "full",
  diag = TRUE,
  tl.col = "black",
  bg = "white",
  title = "",
  col = NULL)
```



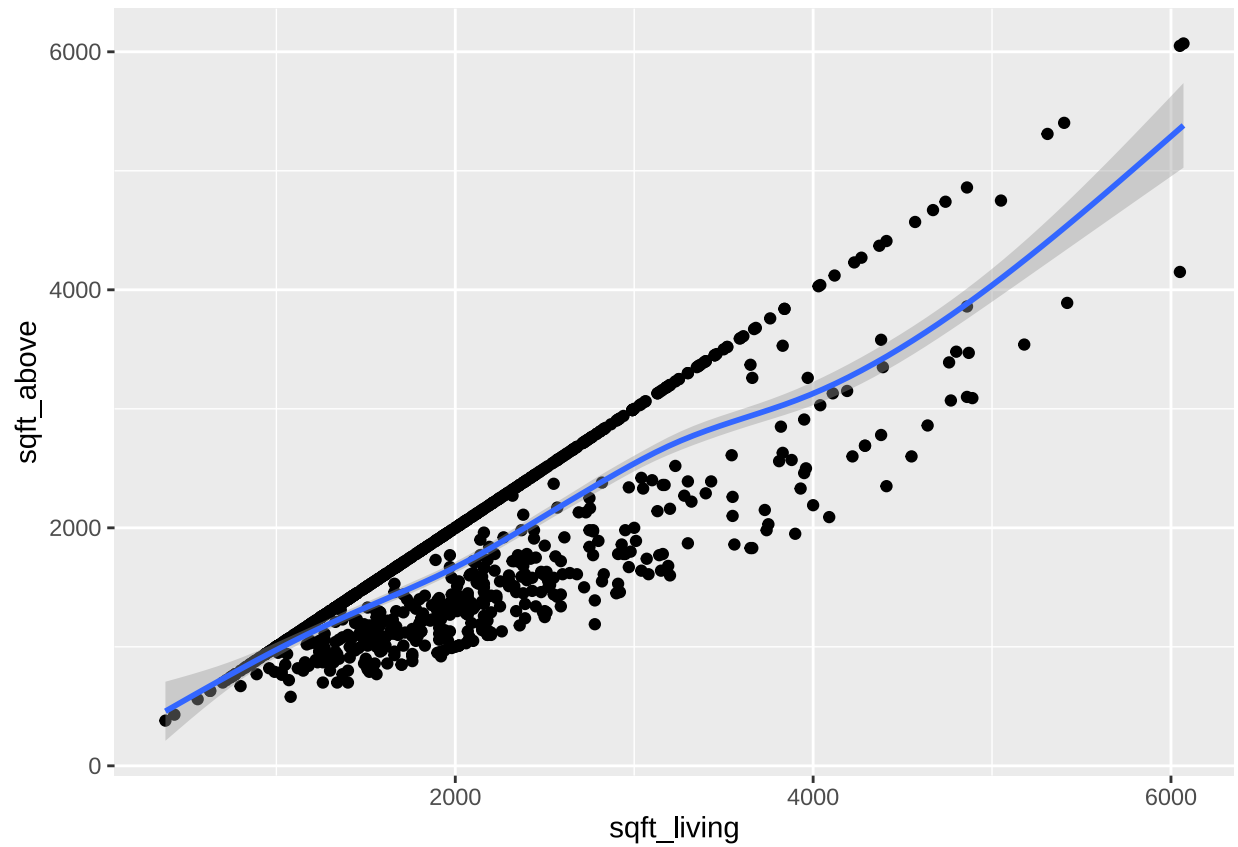
związek między bathrooms a sqft_living

```
data %>% ggplot(aes(x = sqft_living, y=bathrooms))+
  geom_point()+
  geom_smooth()
```



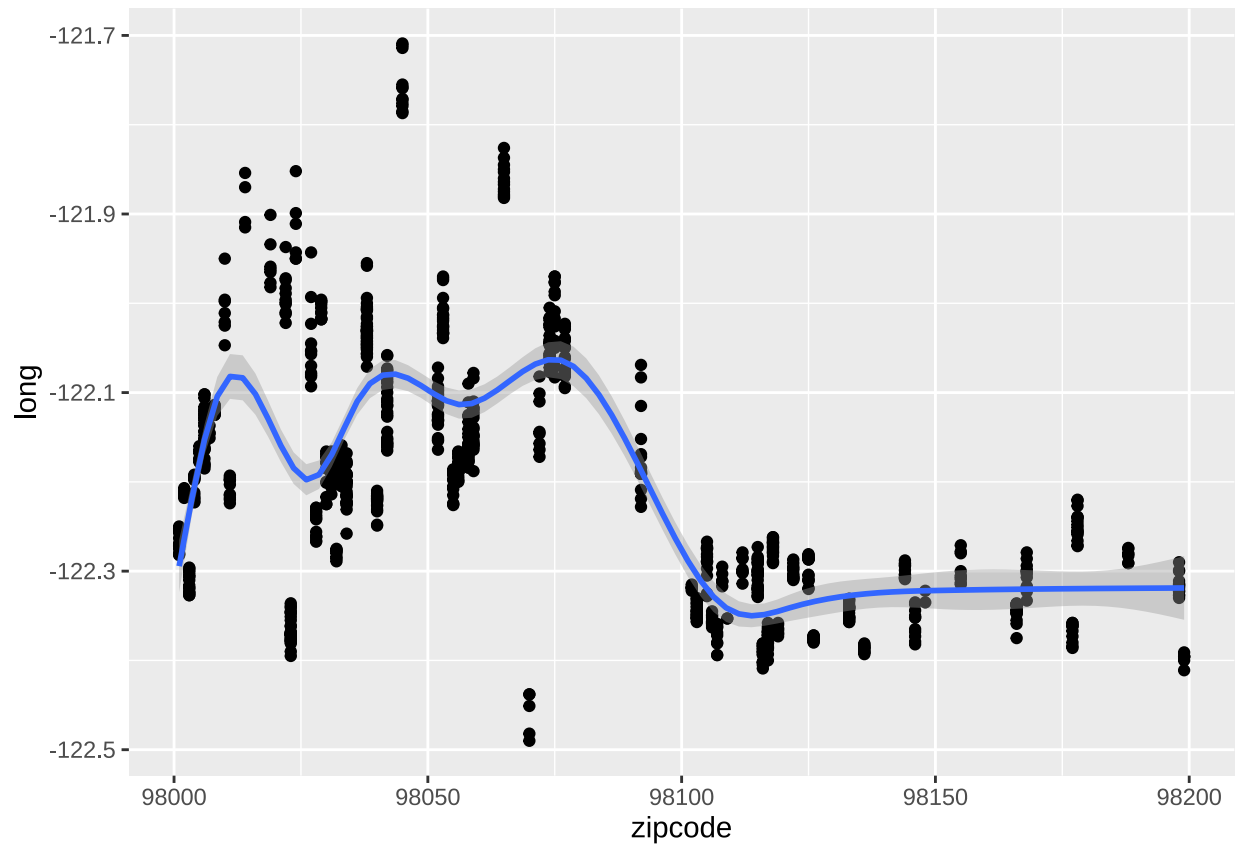
sqft_living i sqft_above

```
data %>% ggplot(aes(x = sqft_living,y=sqft_above))+
  geom_point()+
  geom_smooth()
```



zipcode i long

```
data %>% ggplot(aes(x = zipcode,y=long))+  
  geom_point()+  
  geom_smooth()
```

korelacja z price

```
corrplot(cor(data,use="complete.obs")[18,1:18,drop = FALSE],
         cl.pos = "n", method = "number")
```

price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	price
	0.31	0.49	0.70	0.15	0.24	0.32	0.45	0.07	0.65	0.58	0.37	0.02	0.15	0.07	0.37	0.03	0.65	1.00