LabStat2 - Sprawozdanie 1

. . . czyli jak przewidzieć cenę mieszkania. . .

Michał Makowski

29 listopada 2016

Spis treści

Wstęp **1**

Podstawowa analiza i przygotowanie danych **1**

Model **28**

Podsumowanie **35**

Wstęp

Dane

Zbiór, który będziemy analizowali, jest modyfikacją danych pochodzacych z pakietu autorstwa p. Przemysława Biecka *PBImisc*. Powastał on na bazie ogłoszeń z portalu oferty.net i zawiera informację o transakcjach na rynku mieszkań w Warszawie, w latach 2007-2009. W stosunku do oryginalnego zbioru, nasz jest zmodyfikowany: usunięto pewne informacje, a niekompletne rekordy zostały usunięte.

Cel

Cel jaki bedziemy starali się osiągnać to predykcja ceny za metr^2 na podstawie obserwowanych czynników.

Podstawowa analiza i przygotowanie danych

Przyjrzyjmy się danym:

```
str(apartments)
```

```
## 'data.frame':    780 obs. of  7 variables:
##  $ m2.price         : int  10500 7000 8384 8518 8750 8750 10666 12250 9062 9558 ...
##  $ surface          : num  20 25 26 27 28 28 30 32 32 34 ...
##  $ district         : Factor w/ 25 levels "Bemowo","Bialoleka",..: 15 13 3 23 9 9 15 9 15 6 ...
##  $ n.rooms          : int  1 1 1 1 1 2 1 2 1 1 ...
##  $ floor            : int  7 1 6 1 1 4 1 3 7 2 ...
##  $ condition        : Factor w/ 5 levels "bardzo dobry",..: 3 3 3 5 3 5 5 4 1 1 ...
##  $ construction.date: int  1970 1965 1964 1962 1950 1968 1952 2007 1961 2003 ...
```

```
summary(apartments)
```

```
##     m2.price        surface                district      n.rooms          floor
##  Min.   : 4791   Min.   : 17.00   Mokotow       :182   Min.   :1.000   Min.   : 1.000
##  1st Qu.: 7378   1st Qu.: 36.00   Srodmiescie   :117   1st Qu.:2.000   1st Qu.: 2.000
```

```
##  Median : 8460   Median : 46.10   Praga Poludnie: 71   Median :2.000   Median : 3.000
##  Mean   : 8828   Mean   : 53.01   Ursynow       : 67   Mean   :2.196   Mean   : 4.341
##  3rd Qu.: 9826   3rd Qu.: 61.00   Bielany       : 57   3rd Qu.:3.000   3rd Qu.: 6.000
##  Max.   :21605   Max.   :220.00   Wola          : 57   Max.   :9.000   Max.   :20.000
##                                   (Other)       :229
##           condition   construction.date
##  bardzo dobry  :262   Min.   :1908
##  deweloperski  : 10   1st Qu.:1965
##  do remontu    :133   Median :1978
##  do wykonczenia: 59   Mean   :1979
##  dobry         :316   3rd Qu.:2001
##                       Max.   :2009
##
```

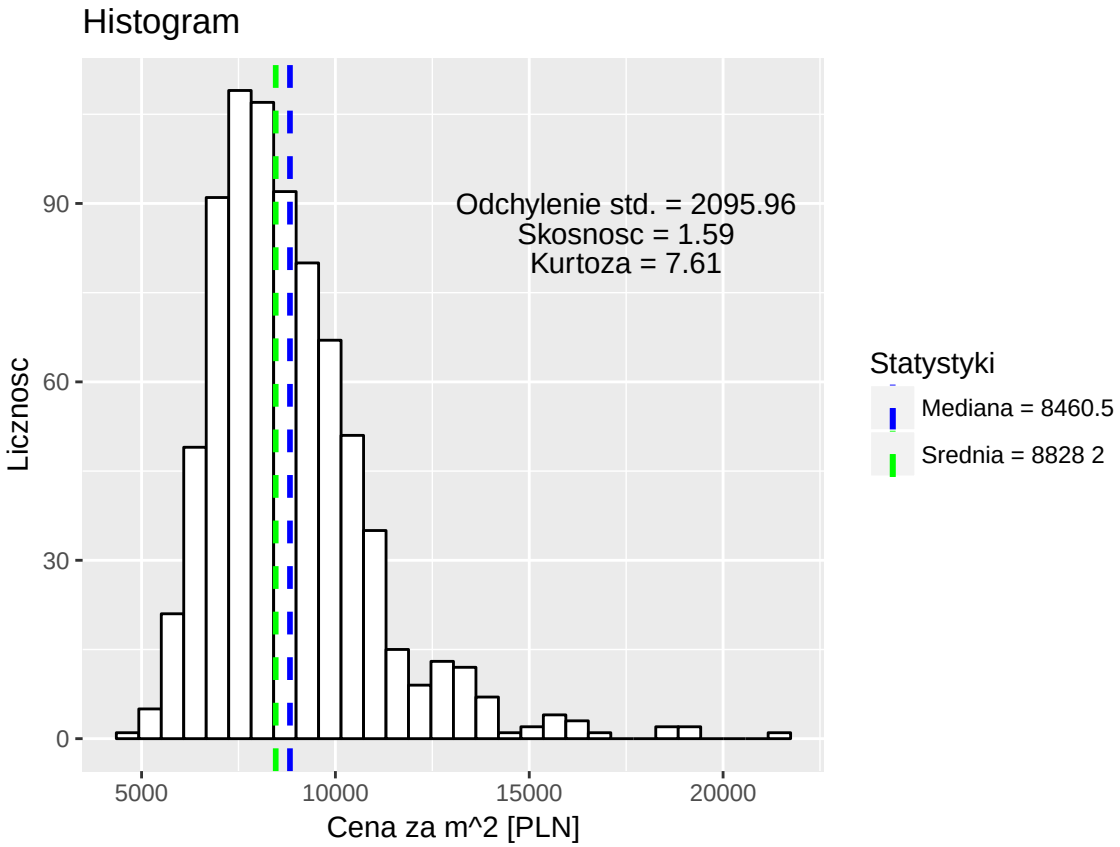
Do dyspozycji mamy 780 obserwacji 7 zmiennych, z których dwie są kategoryczne (*district*, *condition*), trzy dyskretne (*n.room*, *floor*, *construction.date*) oraz dwie, przynajmniej teoretycznie, ciągłe (*m2.price*,*surface*).

Mamy informacje o **dzielnicy**, **stanie**, **liczbie pokoi**, **piętrze**, **dacie budowy**, **powierzchni** i oczywiście o **cenie za metr^2**. Będziemy próbować znaleźć model regresji, który pozwoli nam wyznaczyc cenę metra kwadratowego w oparciu o nie.

W kolejnych akapitach będziemy po kolei analizować każdą ze zmiennych. Robimy to po to, aby zorientować się w danych, a także mieć lepsze wyobrażenie z czym pracujemy. Zoproponuję też pewne uproszczenia, które mogą się przydać przy konstrukcji regresji.

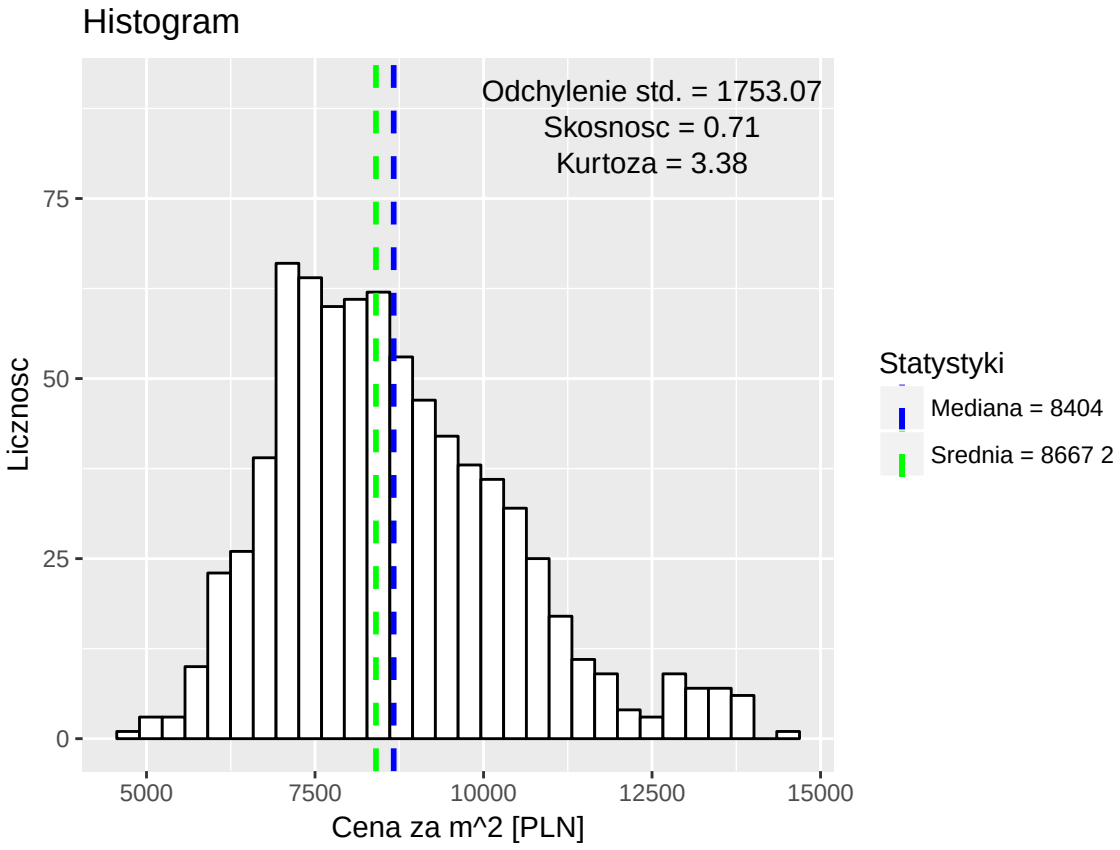
Cena za metr kwadratowy

Zacznijmy od rzeczy teoretycznie najważniejszej, czyli cen za metr kwadratowy:



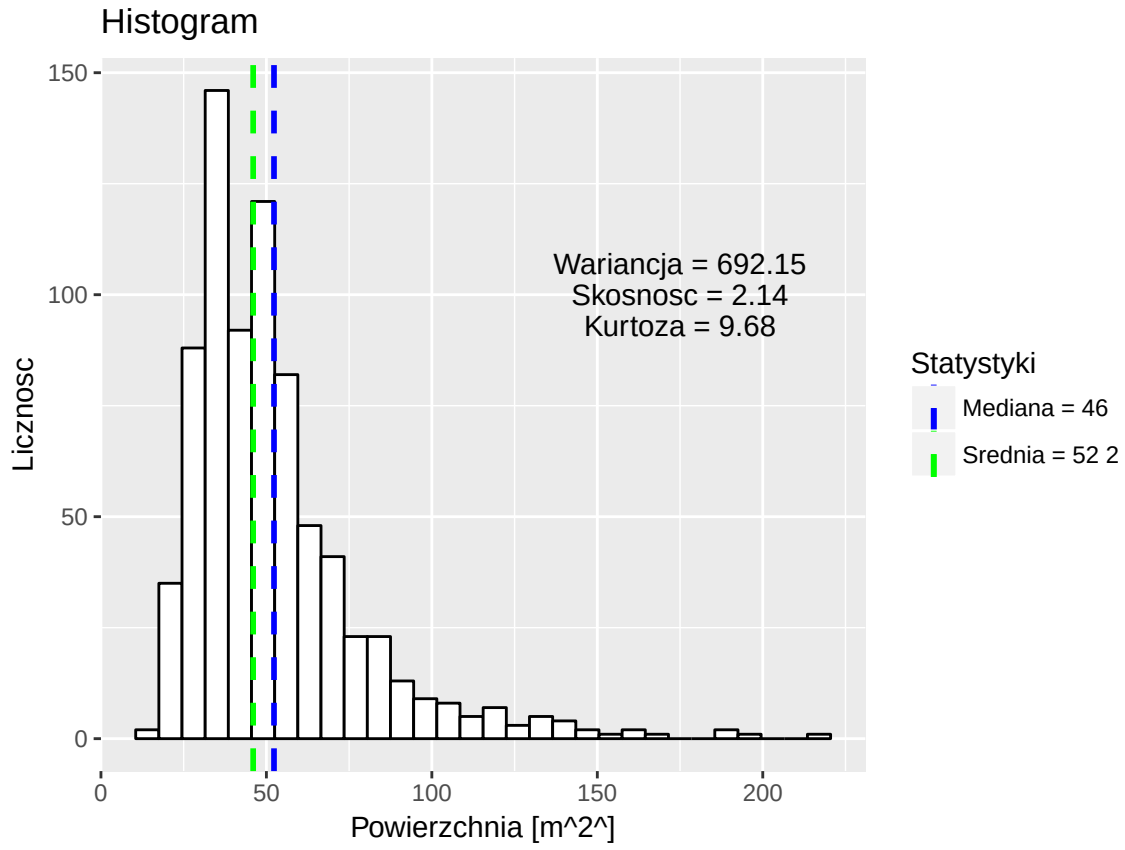
Analizując powyższy histogram dochodzimy do wniosku, że w naszym zbiorze dominują mieszkania do 15tyś./m^2 złotych. Warto zwrócić uwagę (i zapamiętać na potrzeby kolejnych kroków), że mieszkań powyżej 15tyś./m^2 jest 15, a mieszkań powyżej 17,5tyś/m^2. już tylko 5, co stanowi odowiednio 0.019% i 0.006% liczby wszystkich obserwacji. Są to mieszkania zdecydowanie droższe od najczęściej kupowanych i być może niekoniecznie są odpowiednie do modelowania, gdyż prawdopodobnie możemy je zaliczyć jako luksusowe, a dobra luksusowe rządzą się swoimi prawami. Pozostałe statystyki pozostawimy bez dogłębnego komentarza, potwierdzają one głownie to, co widzimy na histogramie.

OD razy odetniemy odserwacje odstające, spójrzmy ponownie na wykres:



Metraż

Przejdźmy do metrażu mieszkań, zacznijmy od histogramu:

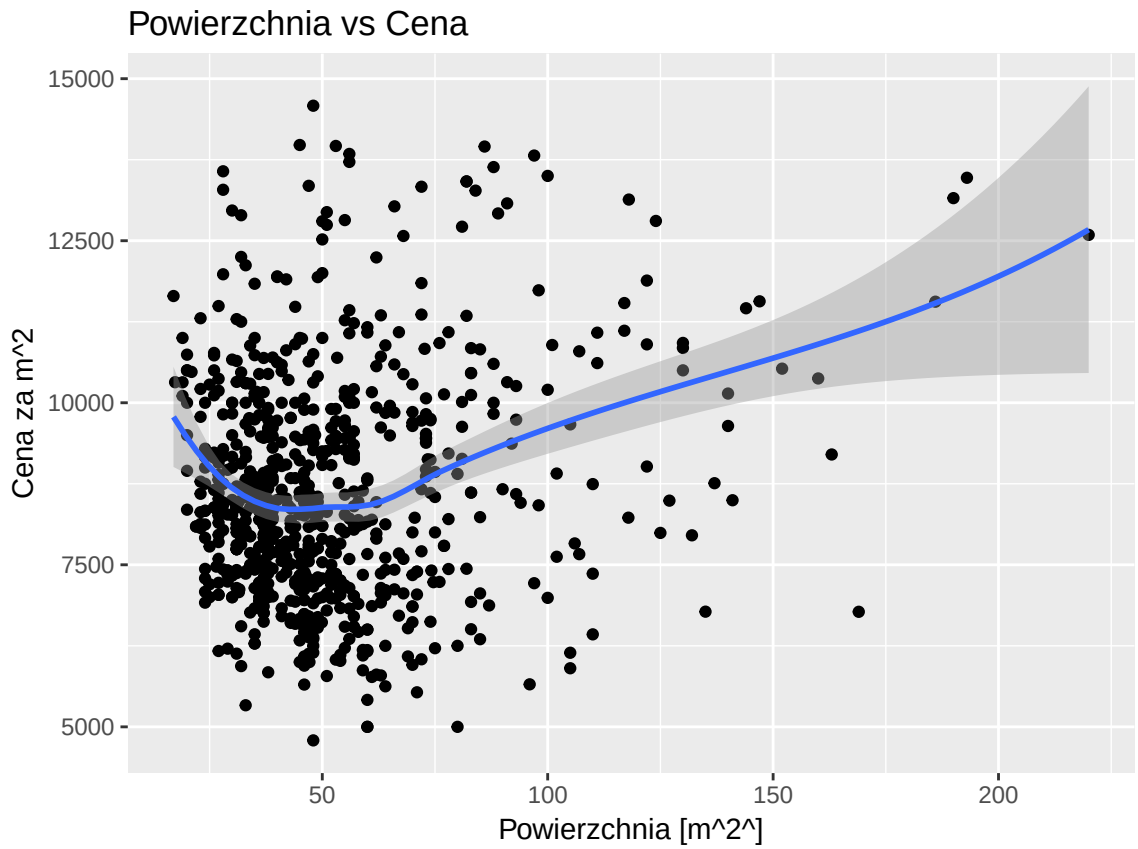


Podobnie jak w ostatnim przypadku można zauważyć, że dominują mieszkania do 150m^2. Tym razem mieszkań powyżej 150m^2 jest 8, a mieszkań powyżej 175m^2. już tylko 4, co stanowi odowiednio 0.01% i 0.005% liczby wszystkich obserwacji. W przypadku metrażu nie jesteśmy w stanie stwierdzić czy możemy zaliczyć te apartamenty do jakieś wąskiej grupy, ale przeanalizujmy te mieszkania dokładniej:

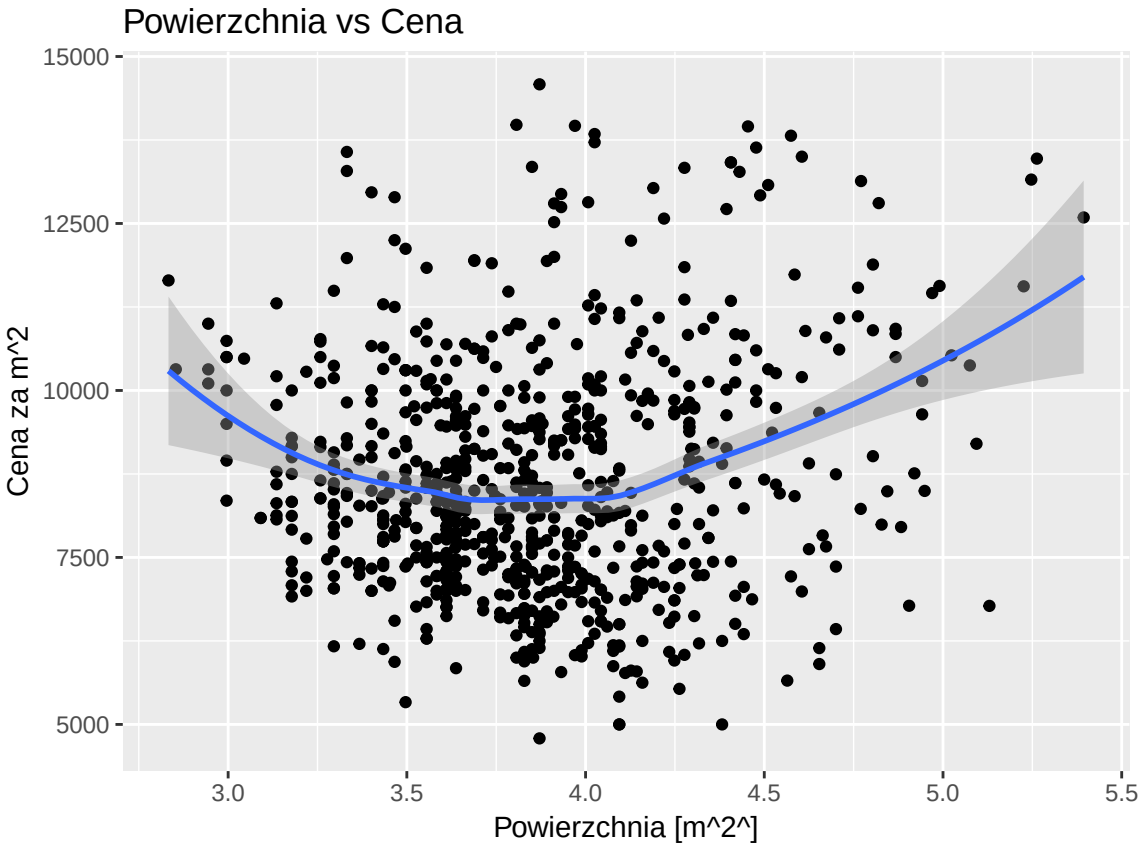
```
apartments[apartments$surface>150,]
```

```
##     m2.price surface       district n.rooms floor      condition construction.date
## 140    11559     186        Mokotow       4    10   bardzo dobry              2000
## 288    13157     190        Mokotow       5     3 do wykonczenia              2007
## 326    10526     152 Praga Poludnie       5     4 do wykonczenia              1928
## 444    12590     220       Zoliborz       9     1          dobry              1928
## 525     6775     169    Srodmiescie       6     6   bardzo dobry              2002
## 561    13471     193    Srodmiescie       6     5   bardzo dobry              1921
## 696    10375     160        Mokotow       4     5   bardzo dobry              2003
## 850     9202     163        Ursynow       4     1   bardzo dobry              1995
```

Jak widać są to mieszkania przedwojenne lub nowe, w dobrym stanie lub wykończenia, w wiekszości droższe od średniej. Nietypowa jest obserwacja **525**, która nijak nie wpisuje się w intuicyjny schemat, gdyż znajduje się w centrum, w nowym budynku, samo mieszkanie jest w bardzo dobrym stanie, a cena zdecydowanie niższa od średniej. Być może na jego cenę wpływ miały czynniki nie zawarte w naszym zestawieniu. Warto to zapamiętać.

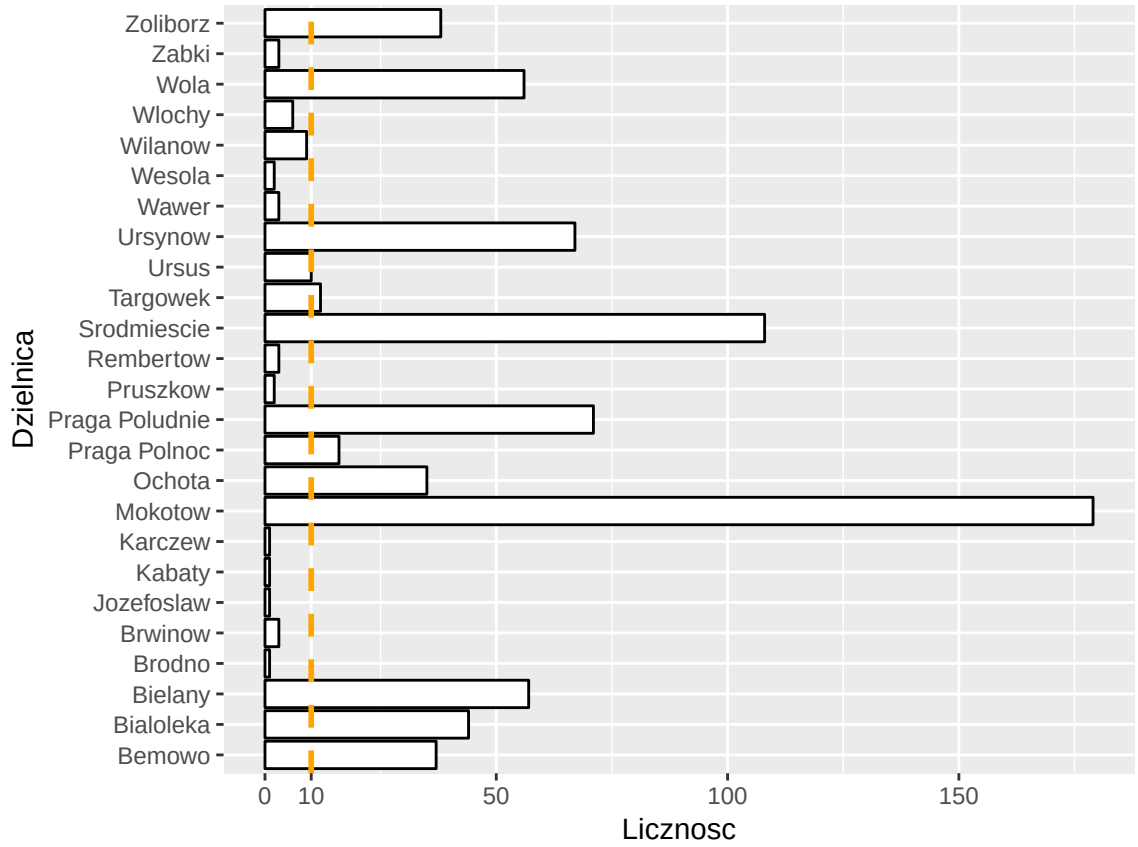


Do wykresu dołączyliśmy linię regresji, pomoze ona nam przewidzieć jak możemy modelować cenę za pomocą metrażu. Może się wydawać, że krzywa dobrze obrazuje trand, ale jest to krzywa co najmniej 3-ego stopnia, za dla dużych obserwacji zmienne są mocno rozporszone. Poszukiwanie modelu pozostawimy na kolejne rozdziały, spróbujmy jeszcze jednej rzeczy, nałóżmy logarytm na *powierzchnię*.



Warto zapamiętać tę własność, po nałożeniu logarytmu na zmienną *powierzchnia* zmienne wyglądają jakby mogły być modelowane krzywą drugiego stopnia.

Dzielnica

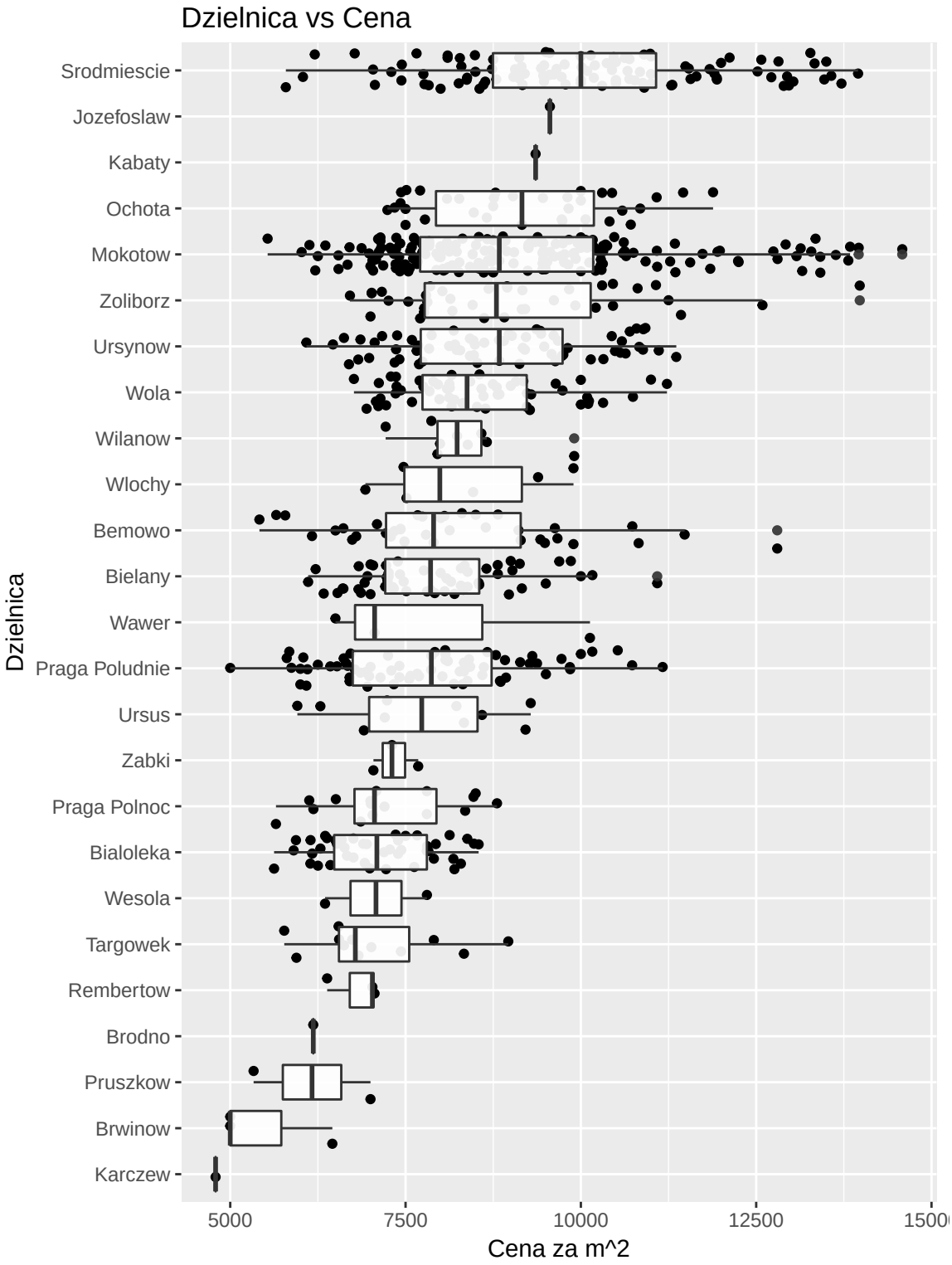


Na powyższym wykresie widzimy, że istnieje wiele obszarów Warszawy, dla których posiadamy bardzo niewiele danych, dla 12 obszarów liczba obserwacji jest mniejsza niż 10. Postaramy się je jakoś zagregować, włączając do większych terenów. Dzielnica dość istotnie wpływa na atrakcyjność mieszkania, dlatego posłużymy się tutaj danymi liczbowymi, ale także subiektywną ocena danej lokalizacji. Warszawę mozna podzielić na dzielnice wyżej i niżej oceniane, np. *Saska Kępa* jest uważana za prestiżową, podczas gdy położona niedaleko *Praga Północ* posiada opinię dzielnicy o wysokiej przestępczośći i problematycznych sąsiadach.

Postaramy się przede wszystkim zmniejszyć ilosc rejonów w samej Warszawie, tzn. najpierw będziemy dązyli do organizacji ich wg. podziału administracyjnego, a później, byc może, jeszcze bardziej ograniczymy ilośc obszarów, w zależności od ich liczności. Podział administracyjny przedstawiamy poniżej.

Grupowanie zmiennych rozpoczniemy od analizy wykresów pudełkowych, przedstawimy na nich strukturę cen na każdym z wyszczególnionych obszarów.

Rysunek 1: Podział administracyjny Warszawy



Wprawdzie nie przeprowadzimy testów statystycznych, ale możemy "na oko" stwierdzić, że bylibyśmy w stanie uszeregować dzielnice od najdroższej do najtańszej. Nie to jest przedmiotem naszych rozważań, więc posłużymy się powyższym wykresem jako wskazówki przy klastrowaniu danych.

W danych *Kabaty* i *Bródno* zostały zakwalifikowane jako dzielnice, my włączymy je do wiekszych zbiorów, tzn. *Ursynowa* i *Targówka*, do których w rzeczywistości należą. Posiadają one tylko po jednej obserwacji, wieć nie

bylibyśmy w stanie nic z niej wyciągnać, a ponadto nie odbiegają one od danych z obszarów "matek".

Ponadto wszystkie miejscowości leżące na południu połączymy w klasę *Przedmieścia*, jednakże sam Józefosław Włączymy do *Ursynowa*, bazujemy na tym, że leżą blisko siebie (ponadto kiedyś pojawiły sie plany włączenia *Józefosławia* do miasta Warszawy), a sama obserwacja z *Józefosławia* jest zbliżona do tych na *Ursynowie* i *Kabatach*:

```
subset(apartments, (district=="Jozefoslaw" | district=="Ursynow" | district=="Kabaty") &
      construction.date >= 2000 & condition=="bardzo dobry" & surface <= 50)
```

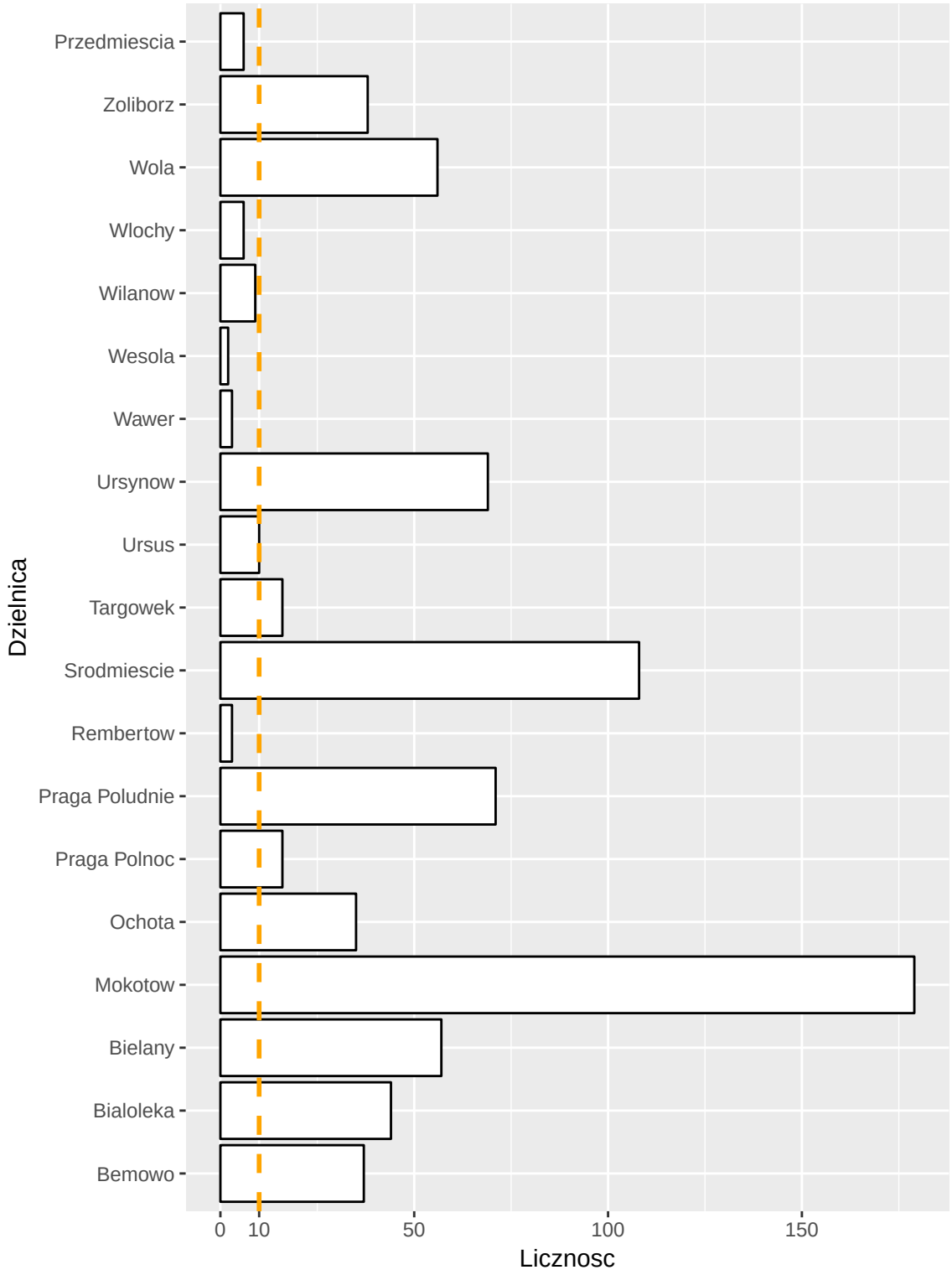
```
##     m2.price surface   district n.rooms floor    condition construction.date
## 10      9558    34.0 Jozefoslaw       1     2 bardzo dobry              2003
## 306    10585    41.0    Ursynow       2     4 bardzo dobry              2002
## 448     9478    36.4    Ursynow       1     6 bardzo dobry              2004
## 540     8622    45.0    Ursynow       2     8 bardzo dobry              2000
## 719     9612    38.0    Ursynow       2     3 bardzo dobry              2001
## 769    10698    39.0    Ursynow       2     4 bardzo dobry              2006
```

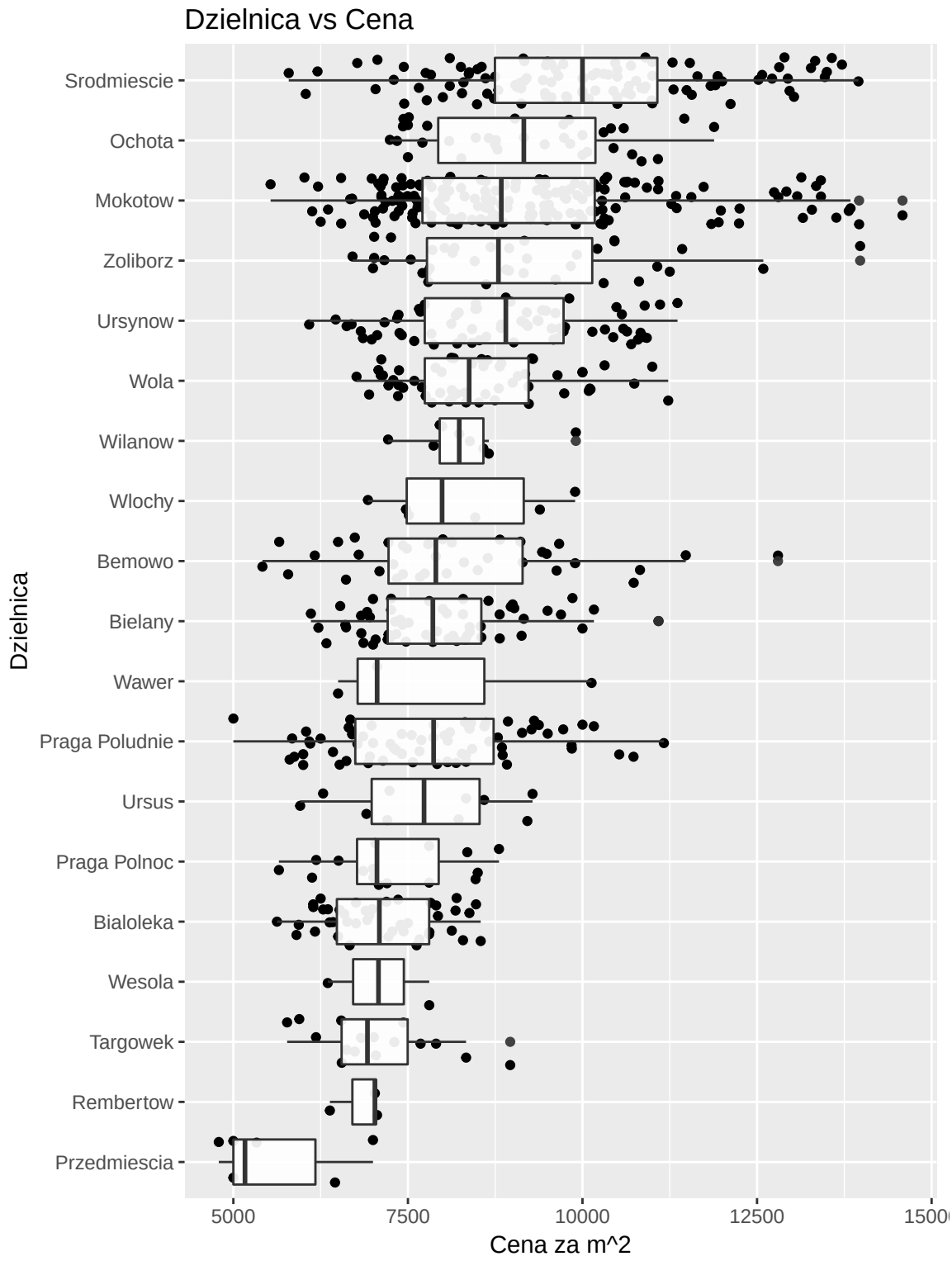
Podobnie sytuacja ma się z *Ząbkami*, które włączymy do, powiększonego już, *Targówka*.

```
subset(apartments, (district=="Zabki" | district=="Brodno" | district=="Targowek") &
      construction.date >= 1990)
```

```
##     m2.price surface district n.rooms floor    condition construction.date
## 225     7903      62 Targowek       2    10 bardzo dobry              2001
## 484     7306      49    Zabki       2     1 bardzo dobry              2003
## 521     7435      78 Targowek       3     2        dobry              1999
## 564     8965      29 Targowek       1     3 bardzo dobry              1996
## 631     7681      34    Zabki       1     4 bardzo dobry              2003
## 923     7042      71    Zabki       3     1 bardzo dobry              2004
```

Ponownie przyjrzyjmy się strukturze naszych danych:


Przedmiescia
Zoliborz
Wola
Wlochy
Wilanow
Wesola
Wawer
Ursynow
Ursus
Targowek
Srodmiescie
Rembertow
Praga Poludnie
Praga Polnoc
Ochota
Mokotow
Bielany
Bialoleka
Bemowo
Dzielnica
0
10
50
100
150
Licznosc

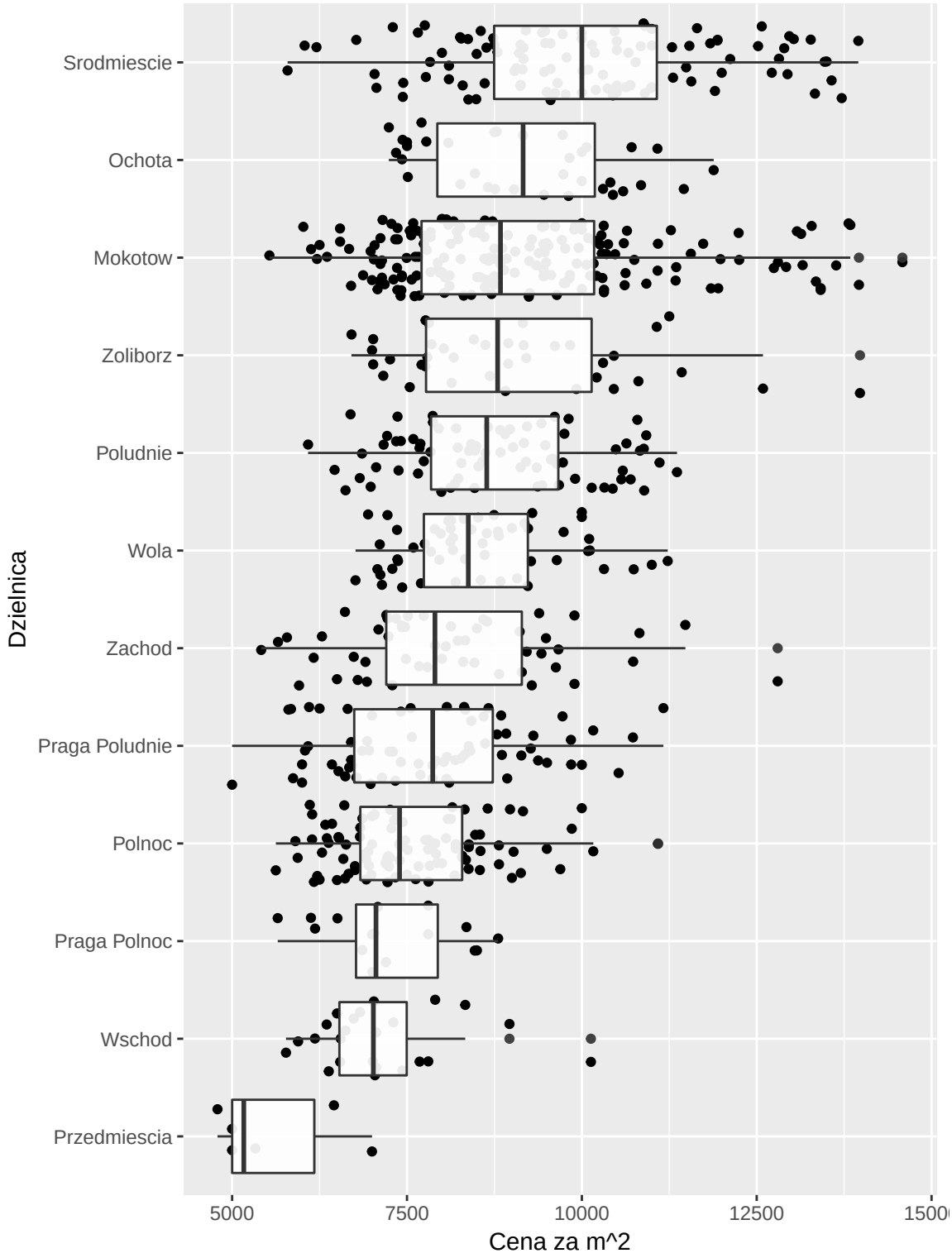


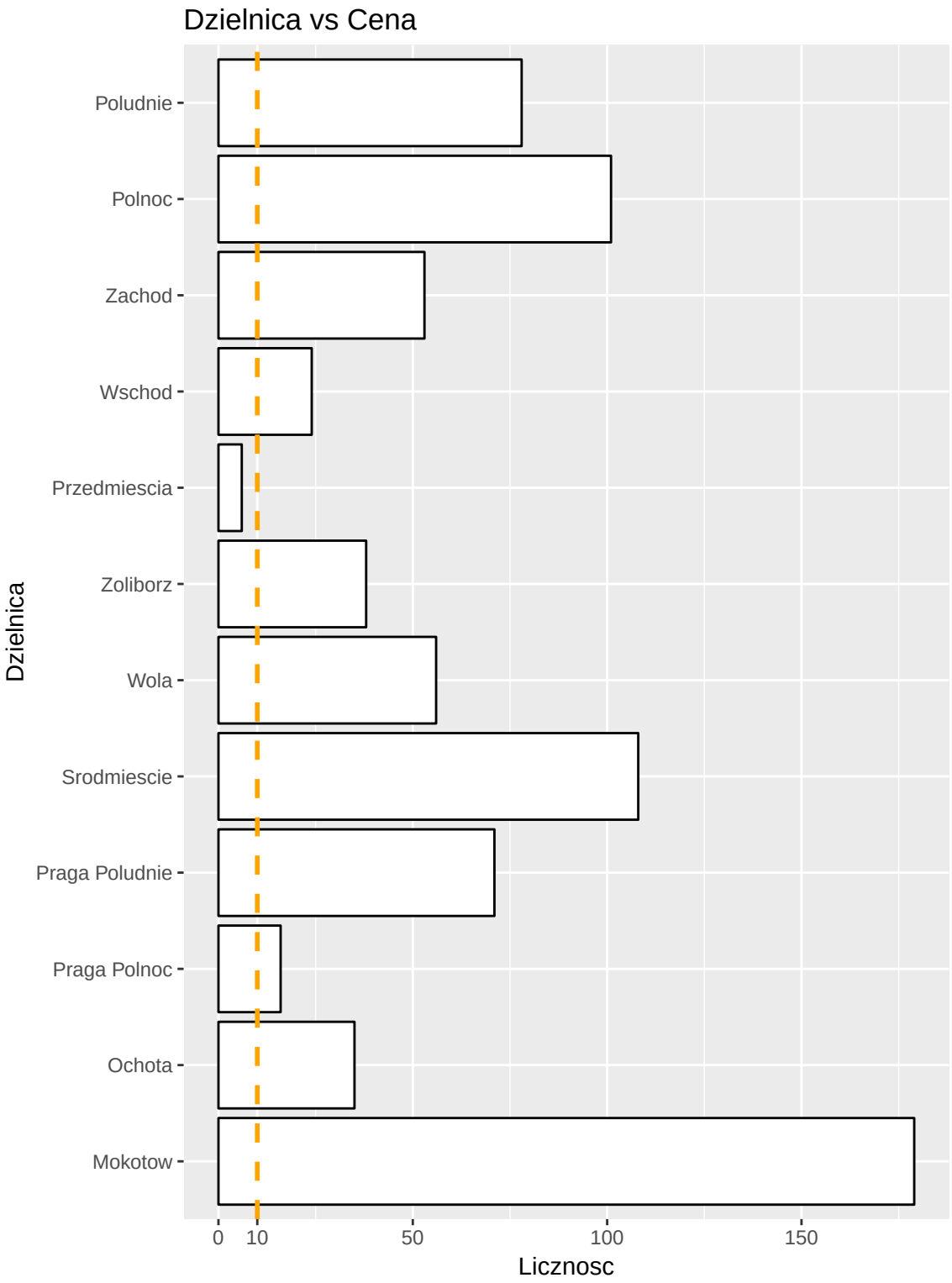
Nadal nie wszystkie "koszyki" nas satysfakconują, dokonamy ogólniejszego podziału, po to, aby w każdym z klastrów znajdowało sie ponad 10 obserwacji. Uprościmy strukturę podziału na dzielnice. Siedem centralnych dzielnic pozostawimy bez zmian, a pozostałe podzielimy wg. położenia względem centrum, tzn. Północ, Połódnie, Wschód, Zachód. Taki podział byłby uzyteczny podczas ew. iwestycji, inwestor wybiera większy obszar miasta i tam poszukuje mieszkania/terenu.

Końcowy podział wygląda nasTepująco:

Samodzielne dzielnice	Północ	Południe	Wschód	Zachód	Przedmieścia
Mokotów	Białołęka	Józefosław	Bródno	Bemowo	Brwinów
Ochota	Bielany	Kabaty	Rembertów	Ursus	Karczew
Praga Południe		Ursynów	Targówek	Włochy	Pruszków
Praga Północ		Wilanów	Wawer		
Śródmieście			Wesoła		
Wola			Ząbki		
Żoliborz					

Strukturę danych obrazują poniższe wykresy:

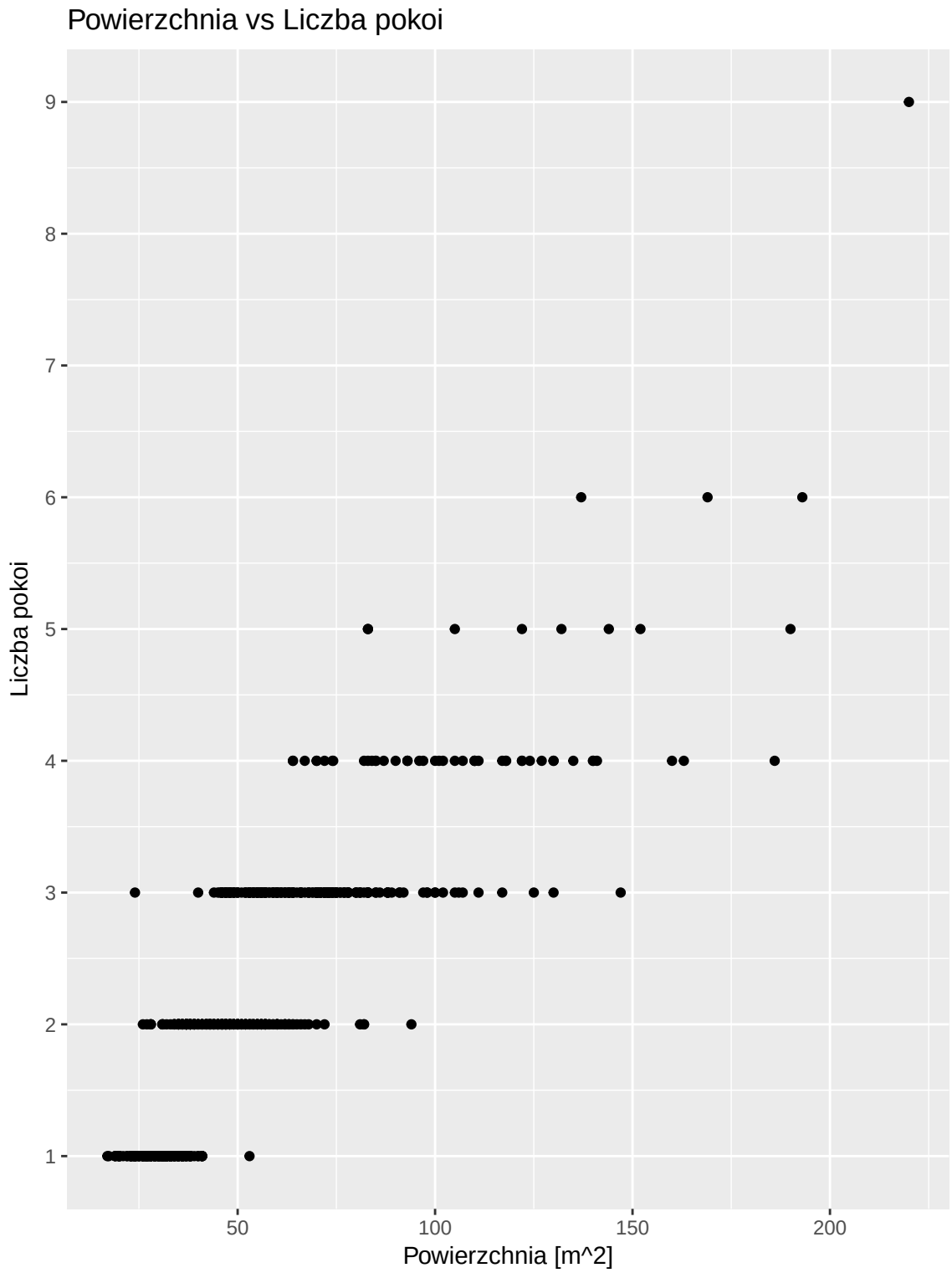

Srodmiescie
Ochota
Mokotow
Zoliborz
Poludnie
Wola
Zachod
Praga Poludnie
Polnoc
Praga Polnoc
Wschod
Przedmiescia
Dzielnica
5000
7500
10000
12500
Cena za m^2



Nadal mamy problem z małą liczbą obserwacji dla koszyka *Przedmieścia*, ale nie jesteśmy w stanie (na chwilę obecną) nic z tym zrobić. Być może zupełnie wykluczymy koszyk *Przedmieścia* z modelu, jako że zajmowaĆ sie mamy danymi dla stolicy, a nie DLA miejscowości podwarszawSkich. Co do samych cen, byc widoczna jest pewna zależność, obserwujemy tereny droższe i tańsze, jednakże dokładniej przeanalizujmy to później.

Liczba pokoi

Pierwsza rzecz jaka przychodzi na myśl, to taka, że liczba pokoi powninna być mocno skorelowana z metrażem. Sprawdżmy:



Okazuje się, że istotnie tak jest. Korelacja tych dwóch zmiennych wynosi 0.839556. Widać mocną zależnośc, co zresztą jest oczywiste, gdyż nikt nie buduje stumetrowych kawalerek.

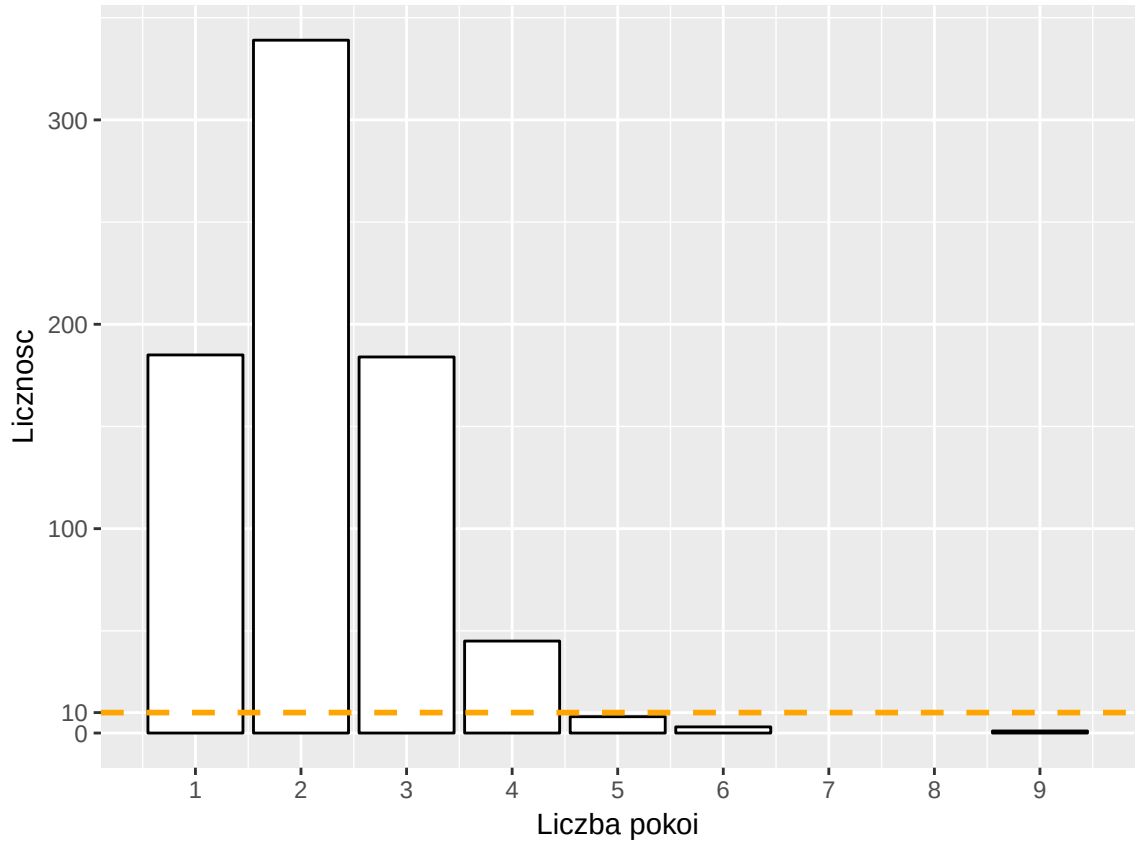
Zdarzają się pewne nietypowe obserwacje, spróbujmy je podejrzeć:

```
temp <- subset(apartments, (n.rooms==1 & surface>=50) | n.rooms>=6 | (surface > 100 & n.rooms <= 3))
temp[order(temp$n.rooms),]
```

```
##     m2.price surface    district n.rooms floor      condition construction.date
## 509     6037      53 Srodmiescie       1     4          dobry              1938
## 88      7992     125    Poludnie       3     4   bardzo dobry              1997
## 89      8908     102    Zoliborz       3     1          dobry              2000
## 138    10794     107    Poludnie       3     3   bardzo dobry              1999
## 232     5904     105      Polnoc       3     3          dobry              2003
## 406    10612     111     Mokotow       3     2 do wykonczenia              2006
## 692    11538     117 Srodmiescie       3     2          dobry              1912
## 694    10923     130     Mokotow       3     4   bardzo dobry              2000
## 695    11565     147 Srodmiescie       3    10   bardzo dobry              1999
## 972     7830     106 Srodmiescie       3     2          dobry              1930
## 234     8759     137     Mokotow       6     2          dobry              1995
## 525     6775     169 Srodmiescie       6     6   bardzo dobry              2002
## 561    13471     193 Srodmiescie       6     5   bardzo dobry              1921
## 444    12590     220    Zoliborz       9     1          dobry              1928
```

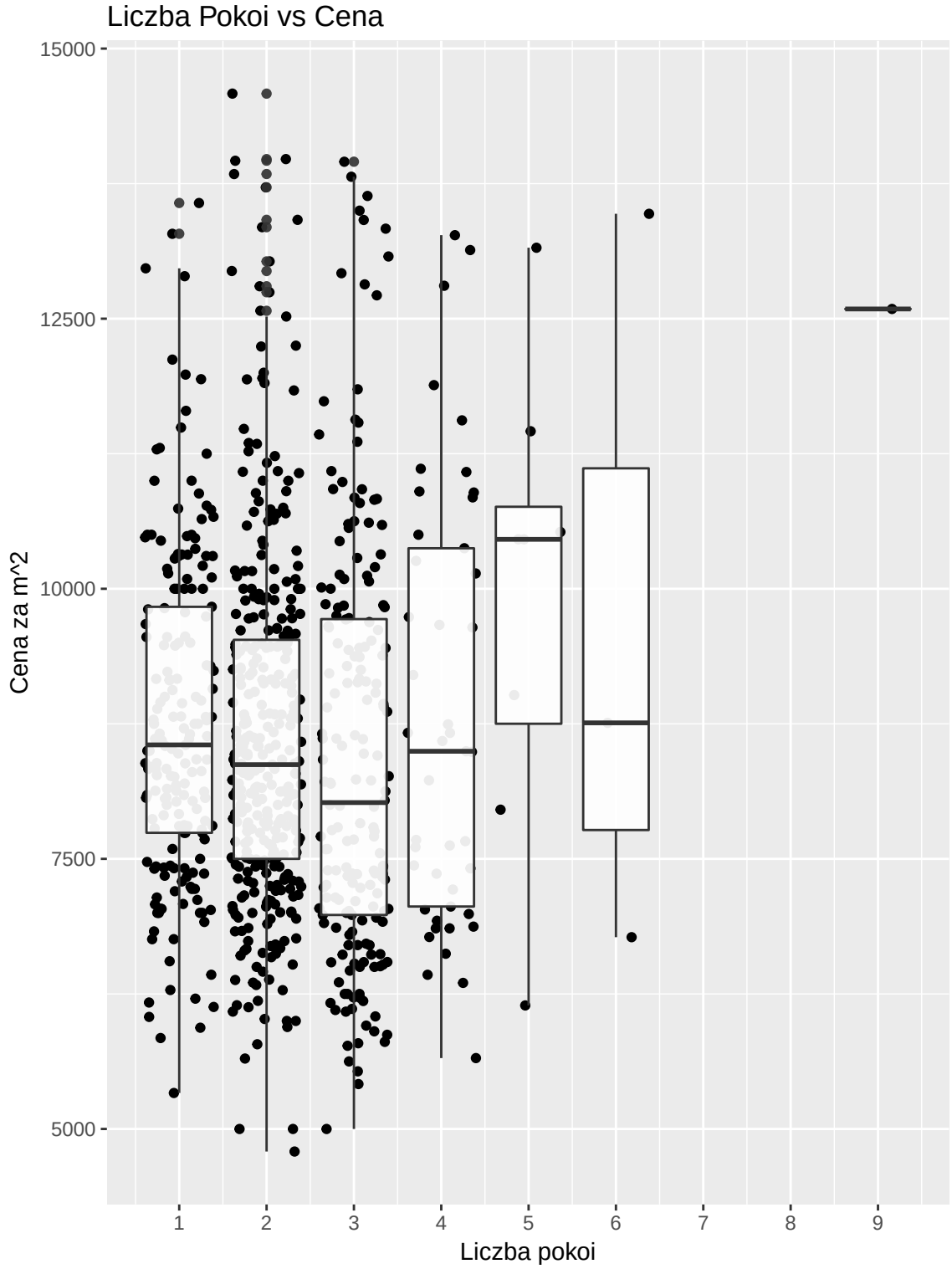
Poza **509** oraz **232** nie widzimy żadnych skrajnych obserwacji

Czas na właściwą analizę liczby pokoi, napierw liczność:



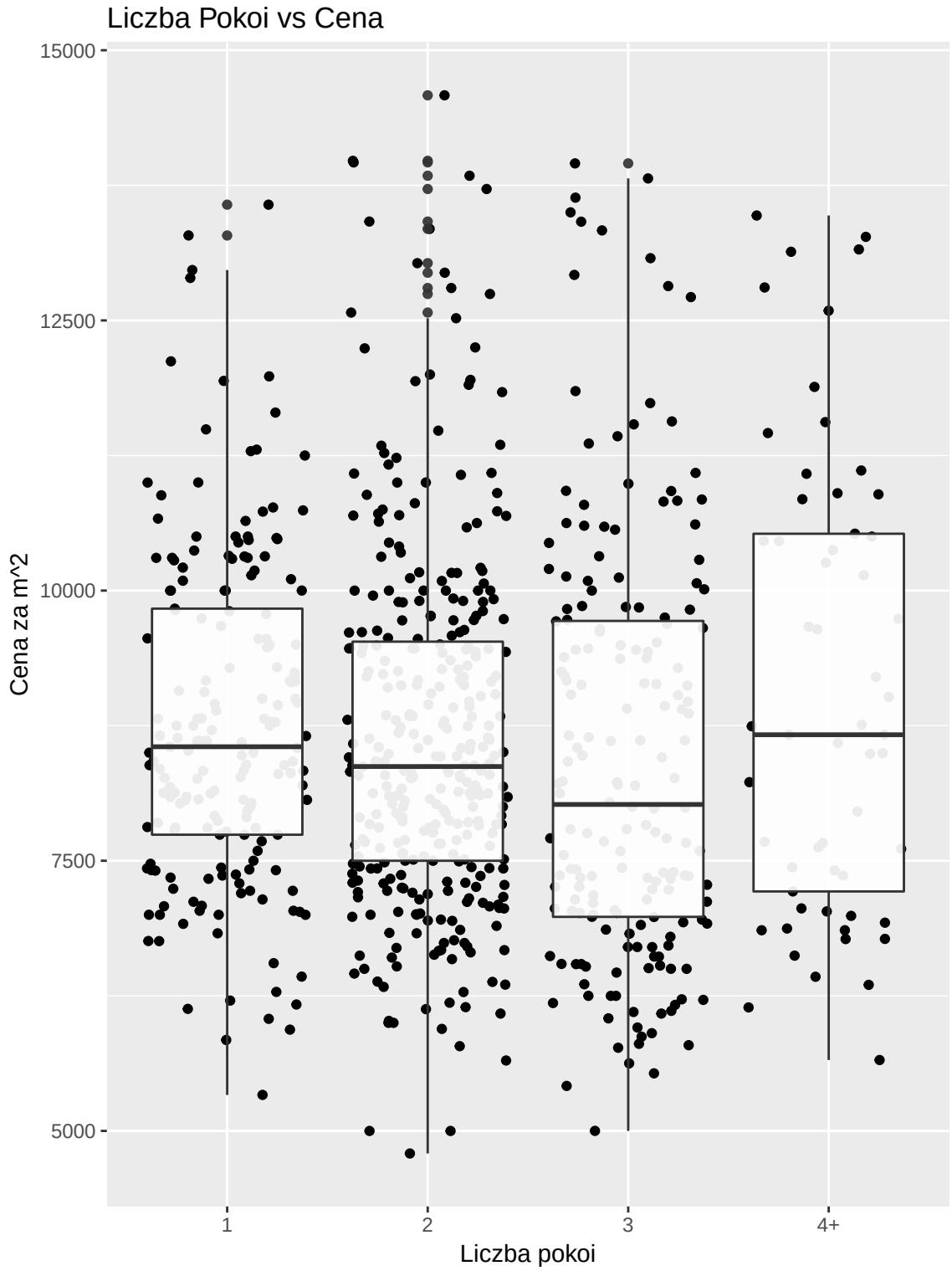
Widać, że w danych dominują kawalerki oraz mieszkania 2/3 pokojowe, większe są zdecydowanie mniej

popularne. Być może połączenie większych mieszkań w jeden koszyk byłoby dobrym rozwiązaniem, będziemy mieć to na uwadze.



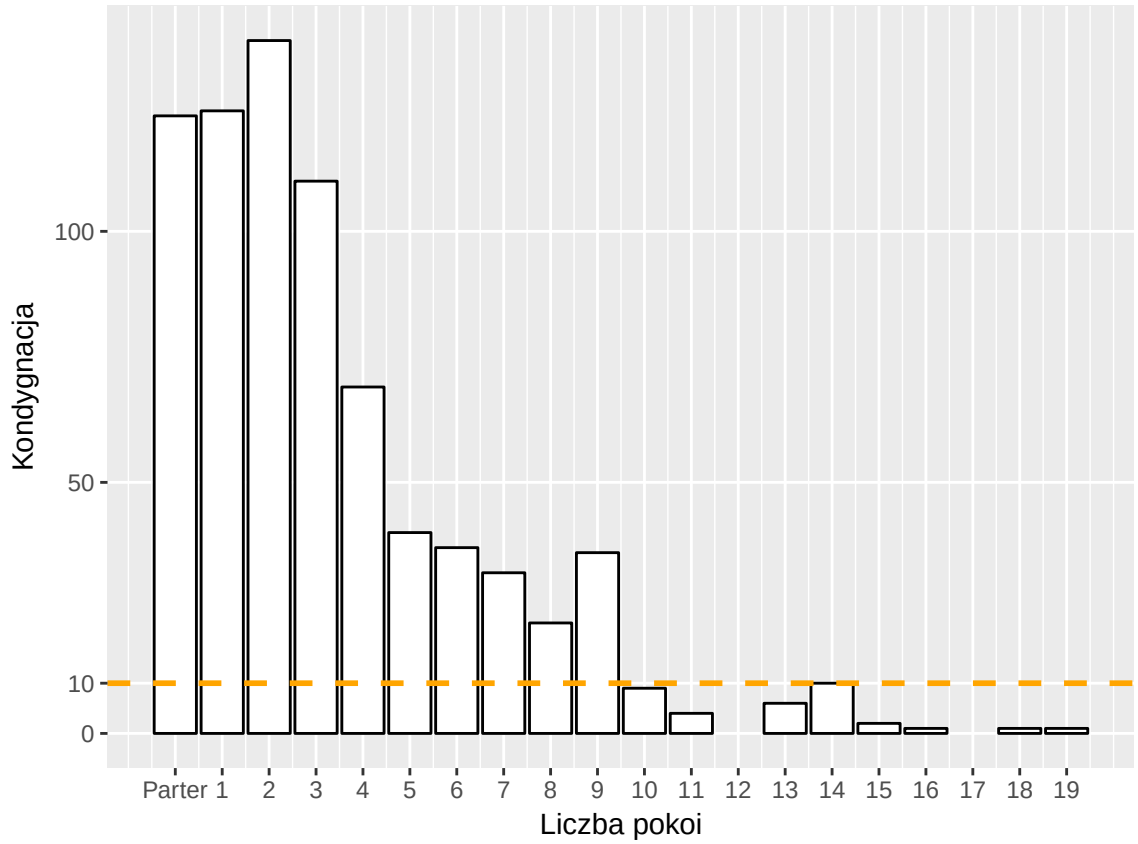
Po raz kolejny nie będziemy się teraz zajmować porównywaniem rozkładów i ich średnich. Wydaje się jednak, że wsród mniejszych mieszkań mozna zaobserwować trend malejącej cany wraz z rosnącą liczbą pokoi, ten trend załamuje się dla mieszkań dużych. Zamieńmy zmienną dot. liczby pokoi na zmienną klasyfikacyjną, a

ponadto potraktujmy mieszkanie większe niż 4-pokojowe jako jedną klasę. Po sklastrowaniu dane wyglądają tak:



Piętro

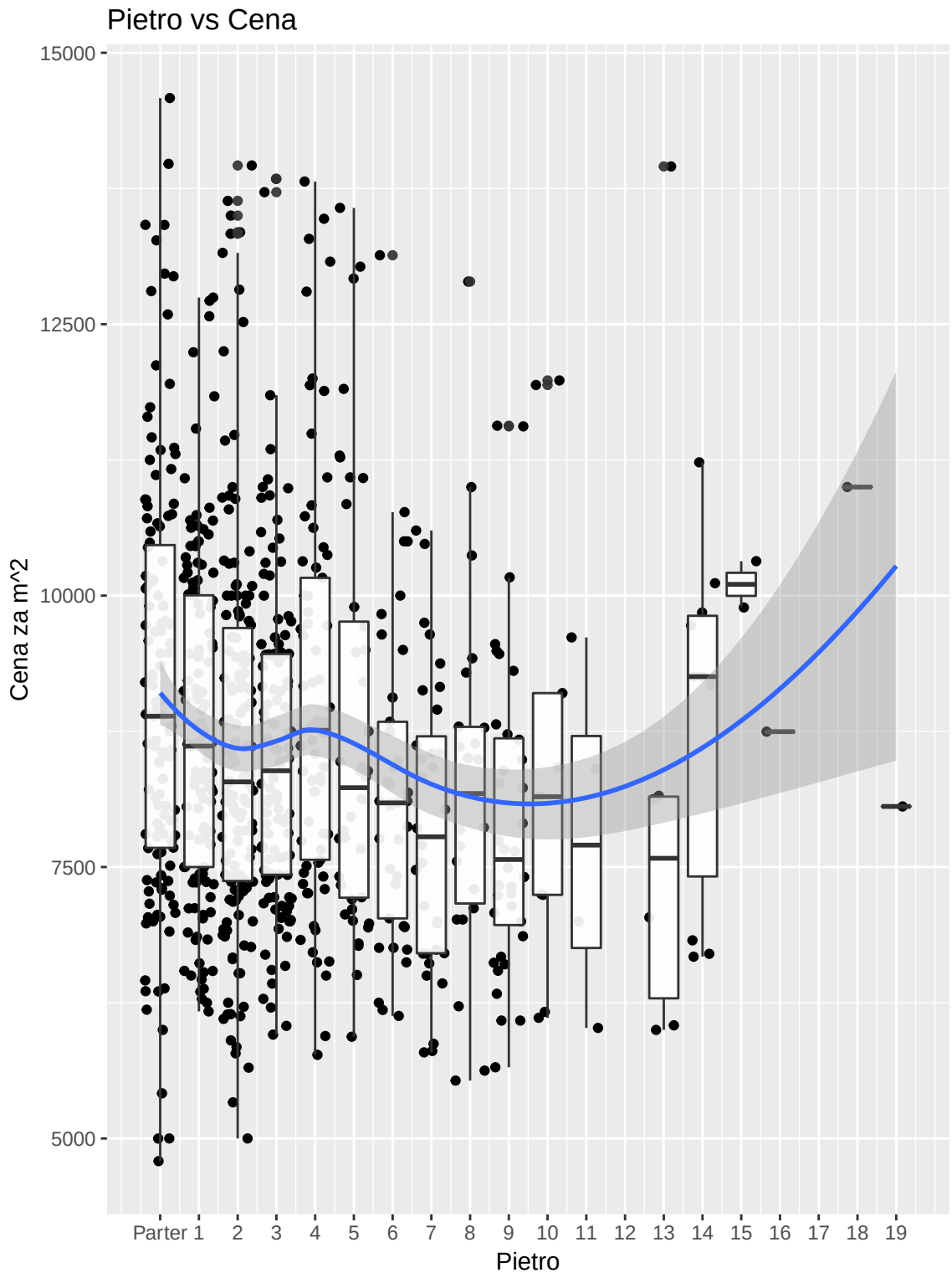
Przeanalizujmy strukturę danych o piętrze, napierw liczność:



Widoczna jest duża dominacja mieszkan położonych na niższych kondygnacjach. Wynika to ze stosunkowo niskiej zabudowy Warszawy, gdzie dominują budynki do 10 pięter. Na chwilę obecną w Warszawie mamy 53 budynki o wys. ponad 65m (co odpowiada mniej więcej 15 pietrom), w 2007r. tych budynków było zdecydowanie mniej, a wiec i mieszkań w nich dostępnych.

Przyjrzyjmy sie strukturze cen:

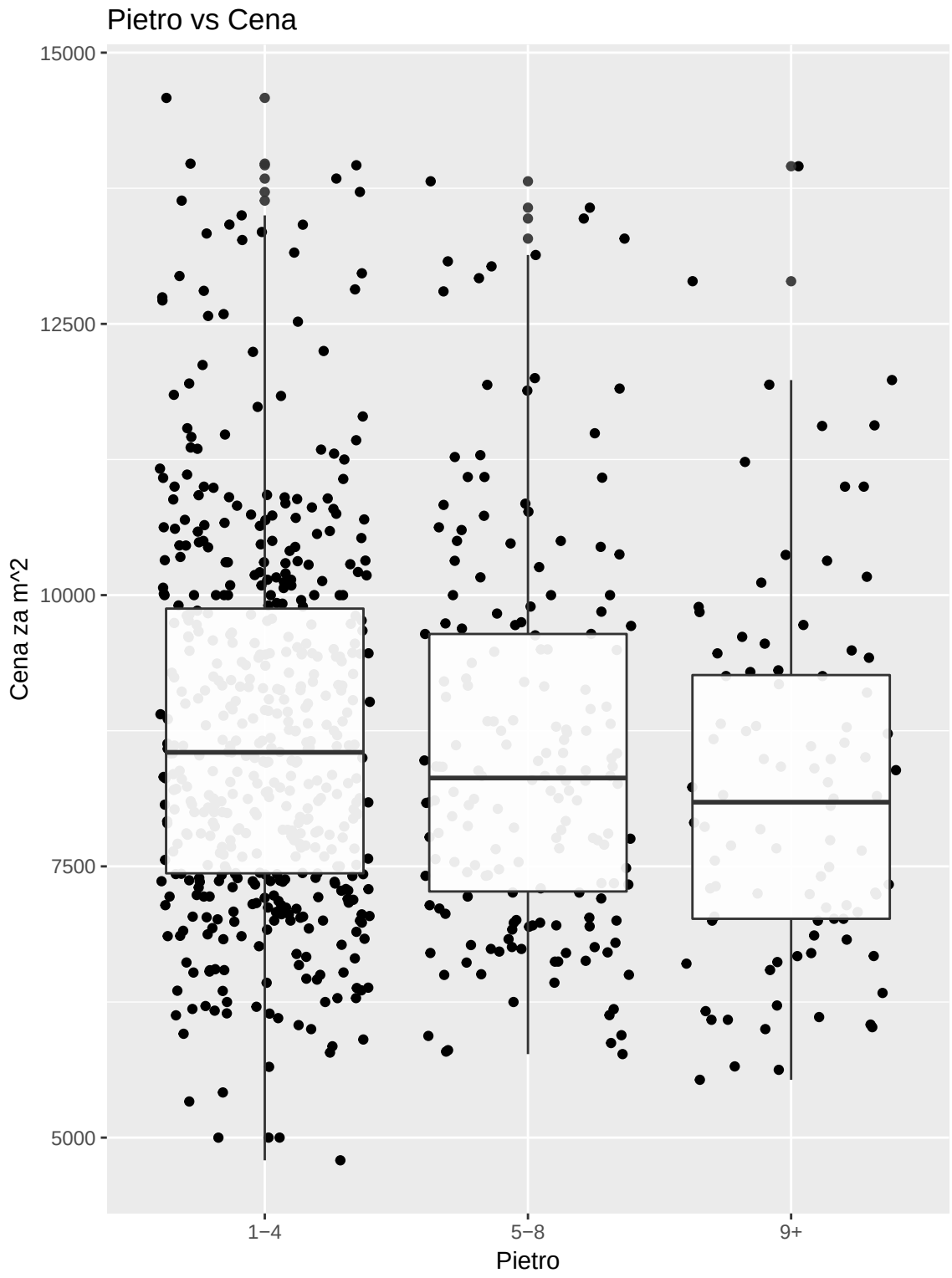
```
## `geom_smooth()` using method = 'loess'
```



Tutaj zdecydowanie trudniej doszukać się zależności, ale na można próbować podzielić mieszkania na nisko położone (do 5-ego piętra), wysoko położone (5-10 piętro), mieszkania b. wysoko położone (powyżej 10-ego piętra). Tak zrobimy. MOglibyśmy zostawić dane tak jak sa i próbować dobrać odpowiedni wielomiań 4 stopnia (niebieska linia na rysunku), ale pietro nie wydaje się być najlepszym predyktorem do tego typu zabiegów. Bardzo wiele zależy od samej infrastruktury budynku. Po sklastrowaniu dane prezentują sie

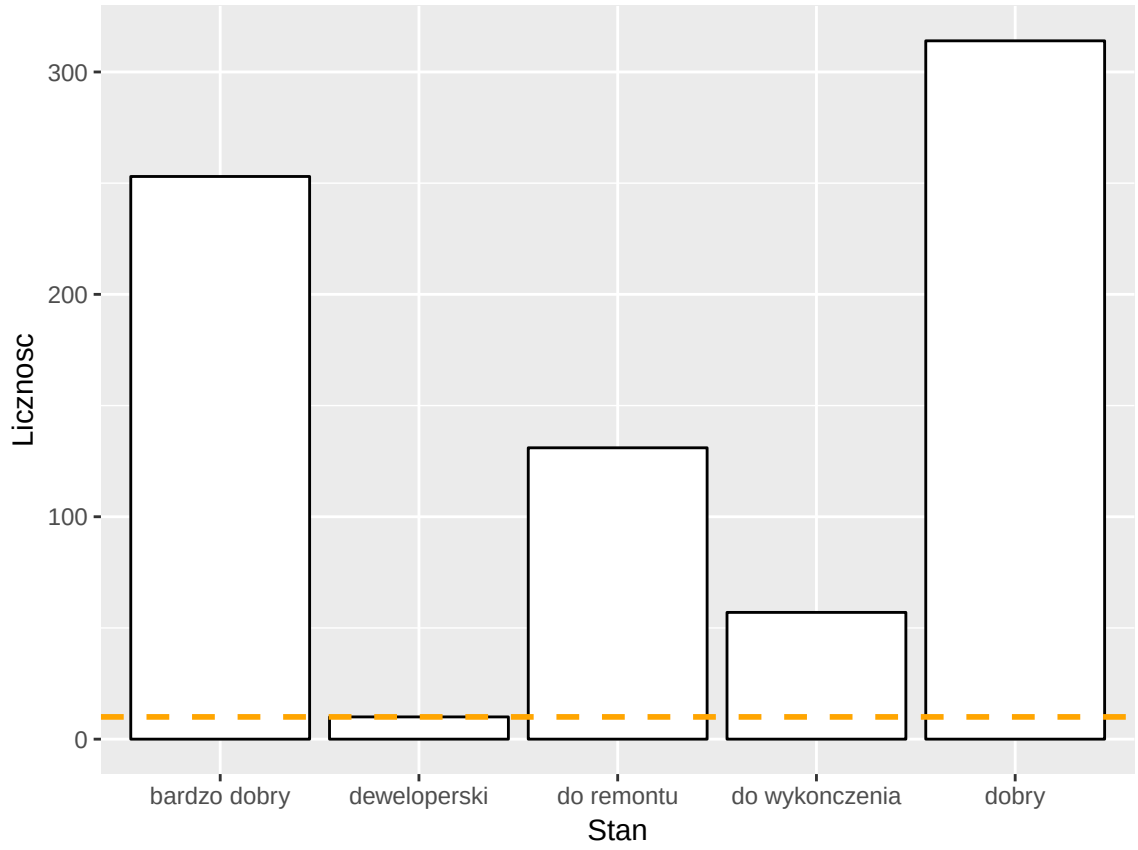
następująco:

```
## `geom_smooth()` using method = 'loess'
```

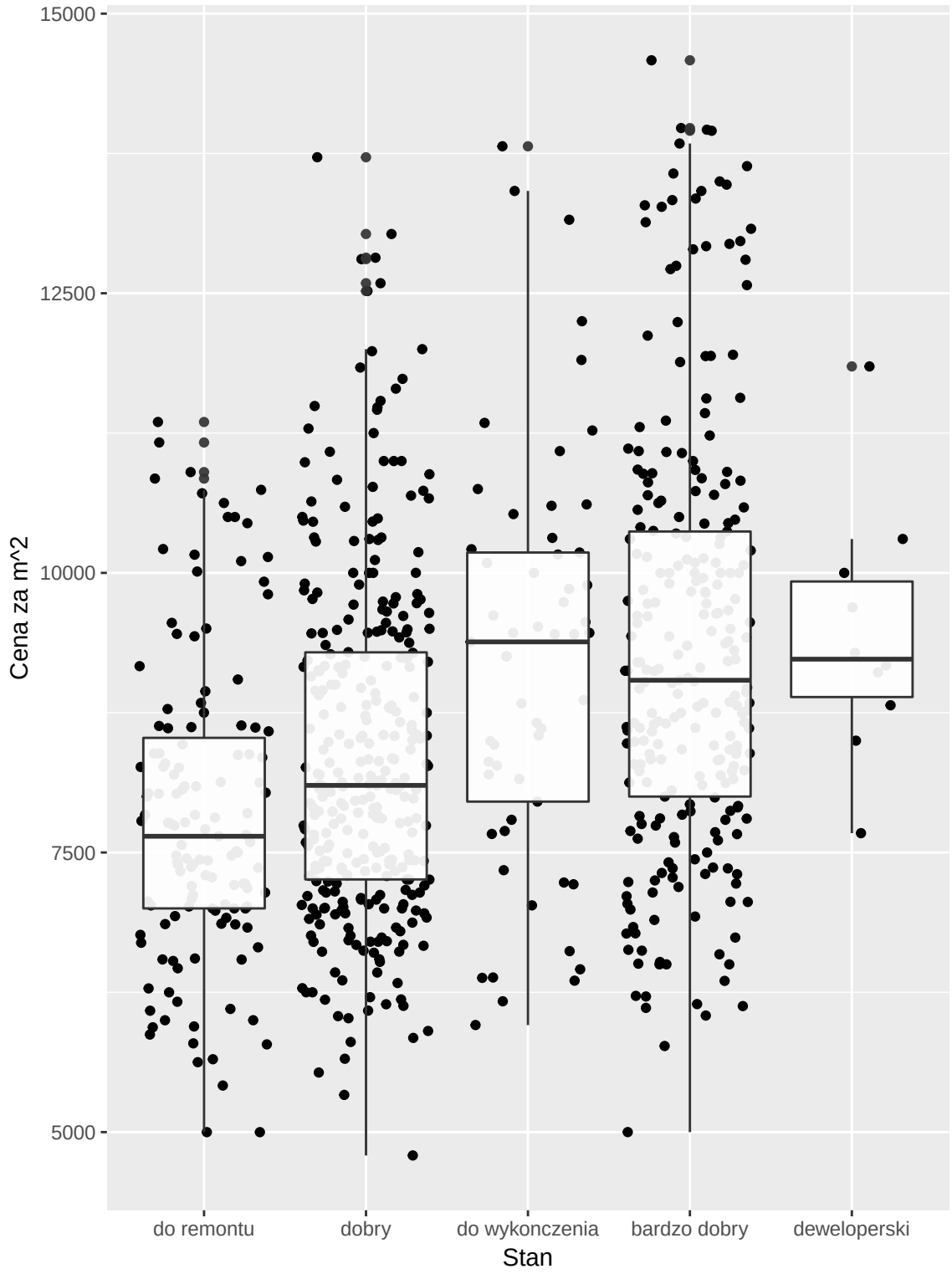


Stan

Zgodnie z tradycją, najpierw liczność:



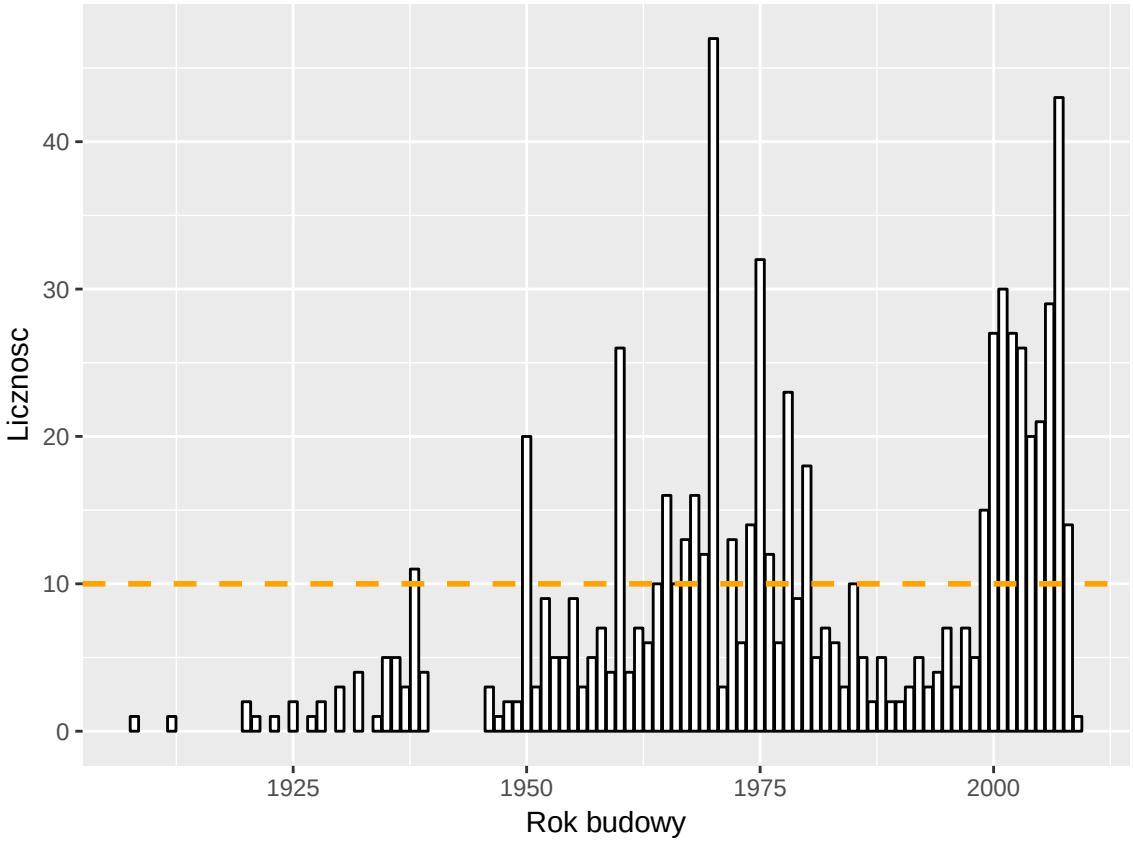
Dominują mieszkania w stanie dobry i bardzo dobry. Zastanawia rozgraniczenie pomiędzy mieszkania *do wykończenia* i w stanie *deweloperskim*, wydają się być do siebie zbliżone.



Stan *deweloperski* włączymy do *do wykończenia*, gdyż nie odbiega on strukturą danych, a i w rzeczywiśtości są do siebie zbliżone. Następnie obydwa te stany włączymy do *bardzo dobrego*, mają podobne struktuwy, a przez potencjalnego klienta odbierane są jako nowe lub prawie nowe. Zdaję się być to dobrym uogólnieniem.

Rok budowy

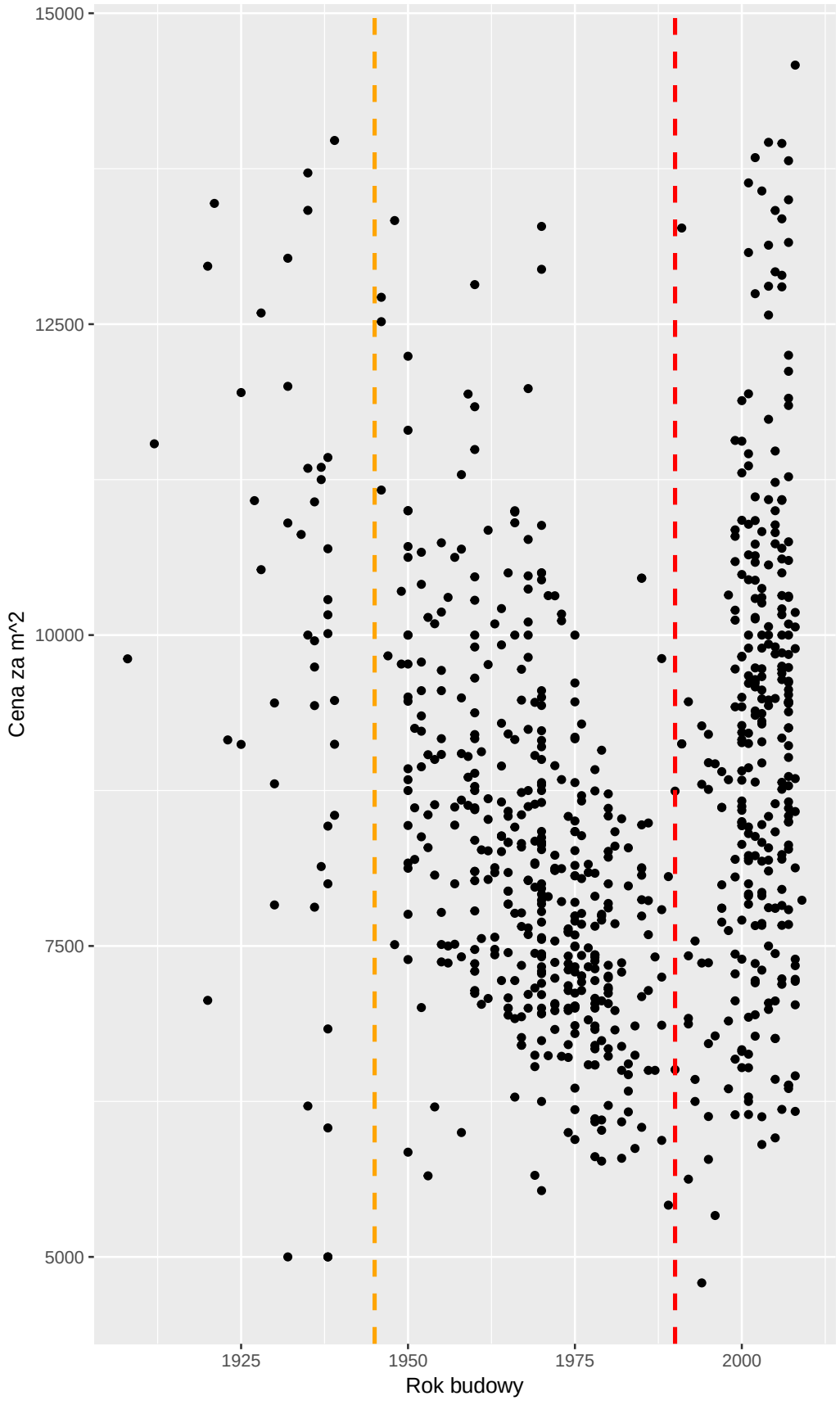
Zgodnie z tradycją, najpier liczność:



Na powyższym wykresie ewidentnie widać 2, a nawet 3 wydarzenia historyczne:

- I Wojnę Światową
- II Wojnę Światową
- Transformację systemową

Mogą one miec wpływa na ceny na rynku, gdyż pomiedzy tymi wydarzeniami budowano budynki w inny sposób. Zobaczmy jak kształtowały się ceny:


15000
12500
10000
7500
5000
Cena za m^2
1925
1950
1975
2000
Rok budowy

Zauważamy tutaj bardzo ciekawą zależność - mieszkania wybudowane po 1990 (czerwona linia) mają bardzo deklikatny trend ceny rosnącej wraz z malejącym wiekiem budynki. Natomiast dla mieszkań budowanych w czasach komuny widać zależnośc, że im starsze są mieszkania, tym na ogół są droższe. Być może warto wprowadzić nową zmienną kategoryczną lub wręcz podzielić dane w 1990 roku i próbować stworzyć dwa, niezależne modele.

Model

Zacznę od budowy najprostszego modelu, opisze go, a następnie przedstawię model bardziej skomplikowany, który, miejmy nadzieję, bedzie lepszy od niego. Przeprowadzę jego diagnostyke

Model podstawowy

Mając w pamięci to, ze zaobserwowaliśmy w ostatni rozdziale przejdziemy do konstrukcji modelu regresji. Jako zmienną która ma najwiekszy wpływ na cene mieszkania uznam jego lokalizację, spróbujmy i porównajmy go od razu z modelem zerowym:

```
lm.district <- lm(m2.price~reorder(district, m2.price), apartments)
summary(lm.district)
```

```
##
## Call:
## lm(formula = m2.price ~ reorder(district, m2.price), data = apartments)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4260.6 -1083.6  -171.8   859.3  5409.4
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               5596.7      622.5   8.991  < 2e-16 ***
## reorder(district, m2.price)Wschod         1535.6      696.0   2.206 0.027654 *
## reorder(district, m2.price)Praga Polnoc   1677.1      729.9   2.298 0.021856 *
## reorder(district, m2.price)Polnoc         1986.0      640.7   3.100 0.002009 **
## reorder(district, m2.price)Praga Poludnie 2252.5      648.3   3.475 0.000541 ***
## reorder(district, m2.price)Zachod         2509.0      656.8   3.820 0.000144 ***
## reorder(district, m2.price)Wola           2944.9      655.0   4.496 8.01e-06 ***
## reorder(district, m2.price)Poludnie       3197.1      646.0   4.949 9.21e-07 ***
## reorder(district, m2.price)Zoliborz       3423.0      669.8   5.110 4.09e-07 ***
## reorder(district, m2.price)Mokotow        3576.9      632.8   5.652 2.25e-08 ***
## reorder(district, m2.price)Ochota         3596.8      673.7   5.338 1.24e-07 ***
## reorder(district, m2.price)Srodmiescie    4456.9      639.6   6.969 6.99e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1525 on 753 degrees of freedom
## Multiple R-squared:  0.2544, Adjusted R-squared:  0.2435
## F-statistic: 23.35 on 11 and 753 DF,  p-value: < 2.2e-16
```

Pierwsze podsumowanie opisze dogłębnie, kolejnę juz tylko skrótowo. Resiudua nie są symetrycznie rozłożone względem zera, co niestety nie daje dobrych nadziei na ich normalność. Być może warto odrzucić skrajne obserwacje, szczególnie mieszkania najdroższe. Przy tak dużej wariancji estymatora mediana residuów, mimo tego, że nie jest w pobliżu zera, może nie byc aż tak słabym wynikiem. P-wartości dla statystyk są na tyle

wysokie, że zmiejszanieilośc kategorii nie wydaje się dobrym pomysłem. R^2 na poziomie 0.25 nie jest dobry wynikiem, ale na pewno uda się go poprawić.

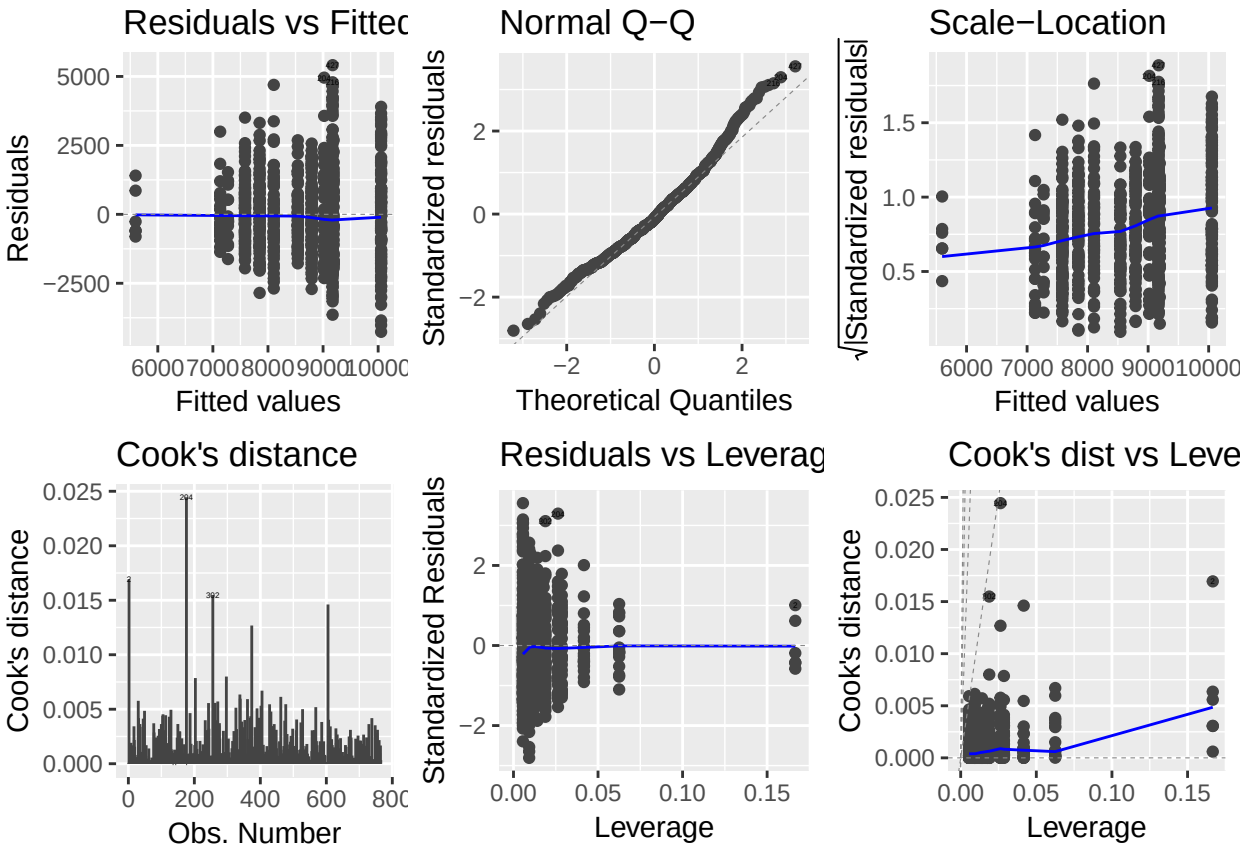
Porównajmy nasz model z modelem zerowym.

```
lm.zero <- lm(m2.price~1, apartments)
anova(lm.zero, lm.district)
```

```
## Analysis of Variance Table
##
## Model 1: m2.price ~ 1
## Model 2: m2.price ~ reorder(district, m2.price)
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    764 2347968098
## 2    753 1750745428 11 597222671 23.352 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

jak widać RRS dla skonstruowanego modelu jest mniejsze niż dla modelu zerowego, wartośc statystyki duża, P-wartośc mała, więc nasz model zdaję sie byc lepszy, od modelu zerowego.

Przejdźmy do analizy wykresów diagnostycznych.



Z racji, że nasze zmienne są kategoryczne, to nie widzimy ładnej "chmury" wokół średniej, powinniśmy przeanalizować residua dla każdej z kategorii. Analizując wykres Q-Q widzimy, że mamy problem z odstającymi obserwacja, tzn. najdroższe mieszkania zaburzają naszą regresję. Jest to kolejny powód, by sie im przyjrzeć dokładniej. Możemy albo próbować szukać zależności pomiędzy nimi, albo je usunąć ze zbioru danych. Wariancja nie jest jednorodna, wraz z ceną za m^2 rośnie, co znów nas alarmuje. Miara Cooka jest niska dla każdej z obserwacji, więc nie ma obserwacji istotnie wpływowych, wynika to jednak z tego, że mamy

zmienne kategoryczne, odjęcie jednej zmiennej nie wpływa istotnie na współczynniki z pozostałych kategorii. Podobnie ma się sprawa na ostatnim wykresie.

```
##
##  Shapiro-Wilk normality test
##
## data:  stdres(lm.district)
## W = 0.98135, p-value = 2.675e-08
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm.district
## BP = 59.8, df = 11, p-value = 1.01e-08
```

```
##
##  Goldfeld-Quandt test
##
## data:  lm.district
## GQ = 0.90779, df1 = 371, df2 = 370, p-value = 0.8239
## alternative hypothesis: variance increases from segment 1 to 2
```

```
##
##  Durbin-Watson test
##
## data:  lm.district
## DW = 1.8428, p-value = 0.005078
## alternative hypothesis: true autocorrelation is greater than 0
```

Niestety, p-wartości testów są bardzo małe, jednie test GQ na homoskedastyczność nie stawia podstaw do odrzucenia zipotezy HO. Model powyższy pozostawia wiele do życzenia, spróbujmy go zmodyfikoawć

Model bardzo zmodyfikowany

Dokonajmy szeregu modyfikacji modelu:

- Wprowadźmy więcej zmiennych objaśniających:
  - Dodajmy zmienną odpowiadającą za liczbę pokoi, ale podzielmy ją na dwa obszary, po 1990r. i przed
  - Dodajmy zmienną *Data budowy* w trzeciej potędze (i wszystkich mniejszych)
  - Dodajmy zmienną *Stan*
- Tranformacja BC.
- Usuńmy danych z *Przedmieść*.

Teram model regresji prezentuje się następująco:

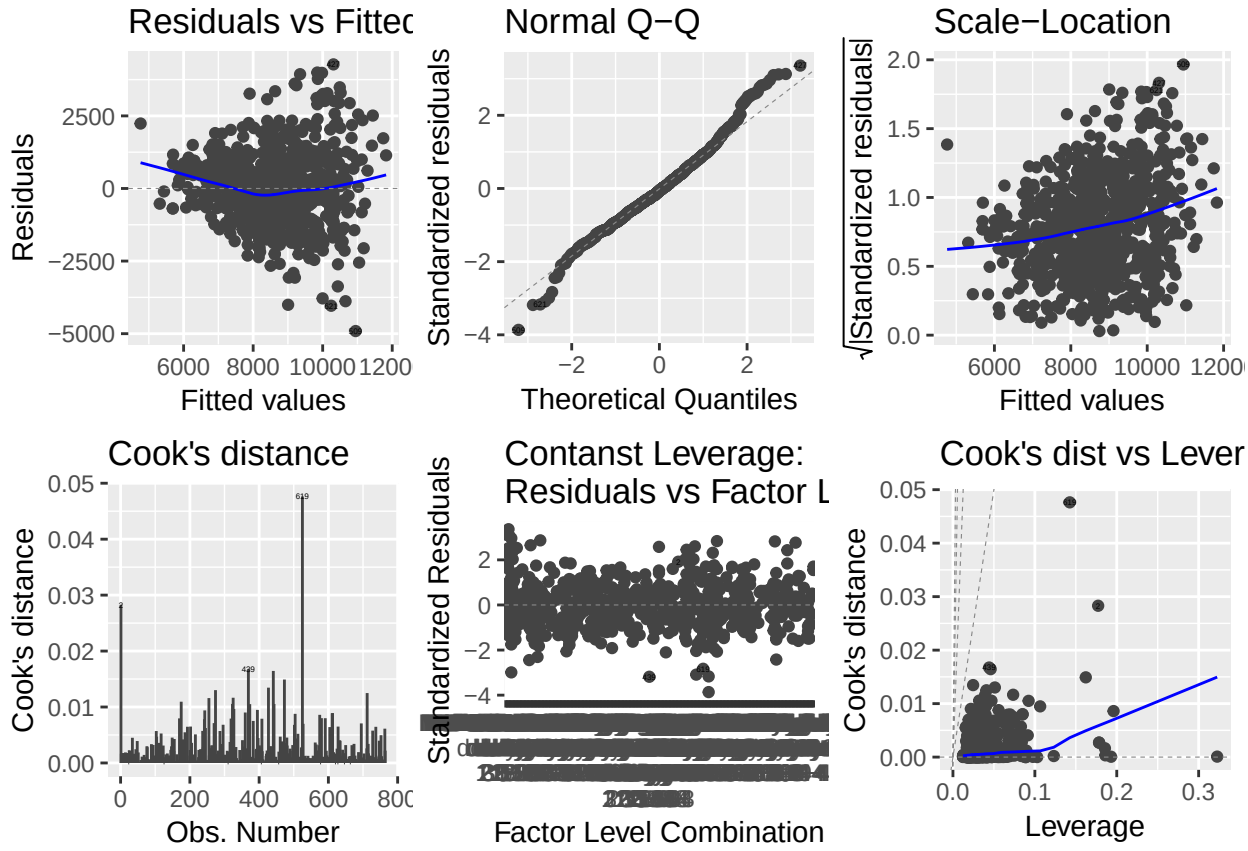
```
lm.apartments <- lm(m2.price~reorder(district, m2.price) +
                        reorder(condition, m2.price) +
                        reorder(n.rooms, m2.price):now +
                        reorder(floor, m2.price)*I(construction.date^3) +
                        I(construction.date^2) +
                        construction.date, apartments)

lambda <- boxcox(lm.apartments, plotit=F)$x[which.max(boxcox(lm.apartments, plotit=F)$y)]
apartmentsBC <- apartments
apartmentsBC$m2.price <- log(apartments$m2.price)
```

```r
lm.apartmentsBC <- lm(m2.price~reorder(district, m2.price) +
                          reorder(condition, m2.price) +
                          reorder(n.rooms, m2.price):now +
                          reorder(floor, m2.price)*I(construction.date^3) +
                          I(construction.date^2) +
                          construction.date, apartmentsBC)

# summary(lm.apartments)
summary(lm.apartmentsBC)
```

```
##
## Call:
## lm(formula = m2.price ~ reorder(district, m2.price) + reorder(condition,
##     m2.price) + reorder(n.rooms, m2.price):now + reorder(floor,
##     m2.price) * I(construction.date^3) + I(construction.date^2) +
##     construction.date, data = apartmentsBC)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59845 -0.08946 -0.00273  0.08942  0.36144
##
## Coefficients: (1 not defined because of singularities)
##                                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                     -1.377e+04  3.634e+03  -3.788 0.000164 ***
## reorder(district, m2.price)Wschod                2.505e-01  6.605e-02   3.792 0.000162 ***
## reorder(district, m2.price)Praga Polnoc          2.837e-01  6.848e-02   4.143 3.82e-05 ***
## reorder(district, m2.price)Polnoc                3.106e-01  6.056e-02   5.129 3.72e-07 ***
## reorder(district, m2.price)Praga Poludnie        3.697e-01  6.134e-02   6.026 2.66e-09 ***
## reorder(district, m2.price)Zachod                3.671e-01  6.228e-02   5.894 5.72e-09 ***
## reorder(district, m2.price)Wola                  4.726e-01  6.196e-02   7.628 7.39e-14 ***
## reorder(district, m2.price)Poludnie              4.520e-01  6.141e-02   7.360 4.93e-13 ***
## reorder(district, m2.price)Zoliborz              5.144e-01  6.326e-02   8.132 1.79e-15 ***
## reorder(district, m2.price)Mokotow               4.909e-01  5.993e-02   8.192 1.14e-15 ***
## reorder(district, m2.price)Ochota                5.134e-01  6.342e-02   8.095 2.36e-15 ***
## reorder(district, m2.price)Srodmiescie           5.808e-01  6.052e-02   9.596  < 2e-16 ***
## reorder(condition, m2.price)dobry                6.300e-02  1.541e-02   4.088 4.84e-05 ***
## reorder(condition, m2.price)bardzo dobry         1.241e-01  1.737e-02   7.142 2.22e-12 ***
## reorder(floor, m2.price)5-8                      7.758e-01  8.716e-01   0.890 0.373707
## reorder(floor, m2.price)1-4                      1.033e+00  8.191e-01   1.261 0.207630
## I(construction.date^3)                           1.843e-06  4.791e-07   3.847 0.000130 ***
## I(construction.date^2)                          -1.081e-02  2.824e-03  -3.828 0.000140 ***
## construction.date                                2.114e+01  5.549e+00   3.810 0.000151 ***
## reorder(n.rooms, m2.price)3:nowFALSE            -1.096e-01  4.119e-02  -2.660 0.007989 **
## reorder(n.rooms, m2.price)2:nowFALSE            -4.361e-02  4.148e-02  -1.051 0.293379
## reorder(n.rooms, m2.price)1:nowFALSE            -7.945e-03  4.202e-02  -0.189 0.850101
## reorder(n.rooms, m2.price)4+:nowFALSE           -2.938e-02  4.978e-02  -0.590 0.555259
## reorder(n.rooms, m2.price)3:nowTRUE              4.022e-02  2.781e-02   1.446 0.148527
## reorder(n.rooms, m2.price)2:nowTRUE              3.566e-02  2.680e-02   1.331 0.183649
## reorder(n.rooms, m2.price)1:nowTRUE              4.197e-02  3.174e-02   1.322 0.186436
## reorder(n.rooms, m2.price)4+:nowTRUE                    NA         NA      NA       NA
## reorder(floor, m2.price)5-8:I(construction.date^3) -9.791e-11  1.120e-10  -0.874 0.382335
## reorder(floor, m2.price)1-4:I(construction.date^3) -1.326e-10  1.053e-10  -1.260 0.208192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.1424 on 736 degrees of freedom
## Multiple R-squared:  0.497,  Adjusted R-squared:  0.4785
## F-statistic: 26.93 on 27 and 736 DF,  p-value: < 2.2e-16
```

Jak widać taki model lepiej dopasowuje się do danych niż poprzedni, choć R^2 nadal nie jest za wysokie. P-wartości dla większości z współczynników są małe. rozkład residów jest bardziej symetryczny. Kolej na wykresy diagnstyczne:



Wyglądają zdecydowanie lepiej niż dla najprostszego modelu, Niestety, residua są przesunięte, ale mniej niż w modelu prostym. Równomierność wariancji też wygląda lepiej, miara Cooka jest niska, nawet dla obserwacji odstających. Dźwignia się zwiększyła, ale do ogległości Cooka jeszcze daleko.

```
##
##  Shapiro-Wilk normality test
##
## data:  stdres(lm.apartmentsBC)
## W = 0.99204, p-value = 0.0004001
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm.apartmentsBC
## BP = 63.476, df = 27, p-value = 9.076e-05
```

```
##
##  Goldfeld-Quandt test
##
## data:  lm.apartmentsBC
```

```
## GQ = 1.1179, df1 = 353, df2 = 353, p-value = 0.1478
## alternative hypothesis: variance increases from segment 1 to 2
```

```
##
##  Durbin-Watson test
##
## data:  lm.apartmentsBC
## DW = 1.9318, p-value = 0.7607
## alternative hypothesis: true autocorrelation is greater than 0
```

Powyższe testy utwierdają nas w przekonaniu, że niestety rozkład residuów nie jest idealny i w pełeni nie spełnia założeń. Test SW na normalność raczej nie daje nadzieji na faktyczny rozkład normalny, z testów na homoskedastyczność jestnie test BP odrzuca hipeteze zerową. Usuwanie obserwacji odstających nie wpływa za bardzo na otrzymywane wartości.

Model mniej zmodyfikowany

Ostatnim modelem jaki będziemy analizować będzie uproszczony model z poprzedniego akapitu Wykonajmy następujące modyfikacje w stosunku do poprzedniego modelu:

- Usuńmy zmienną *Data budowy*.
- Usuńmy zmienną *Liczba pokoi*.

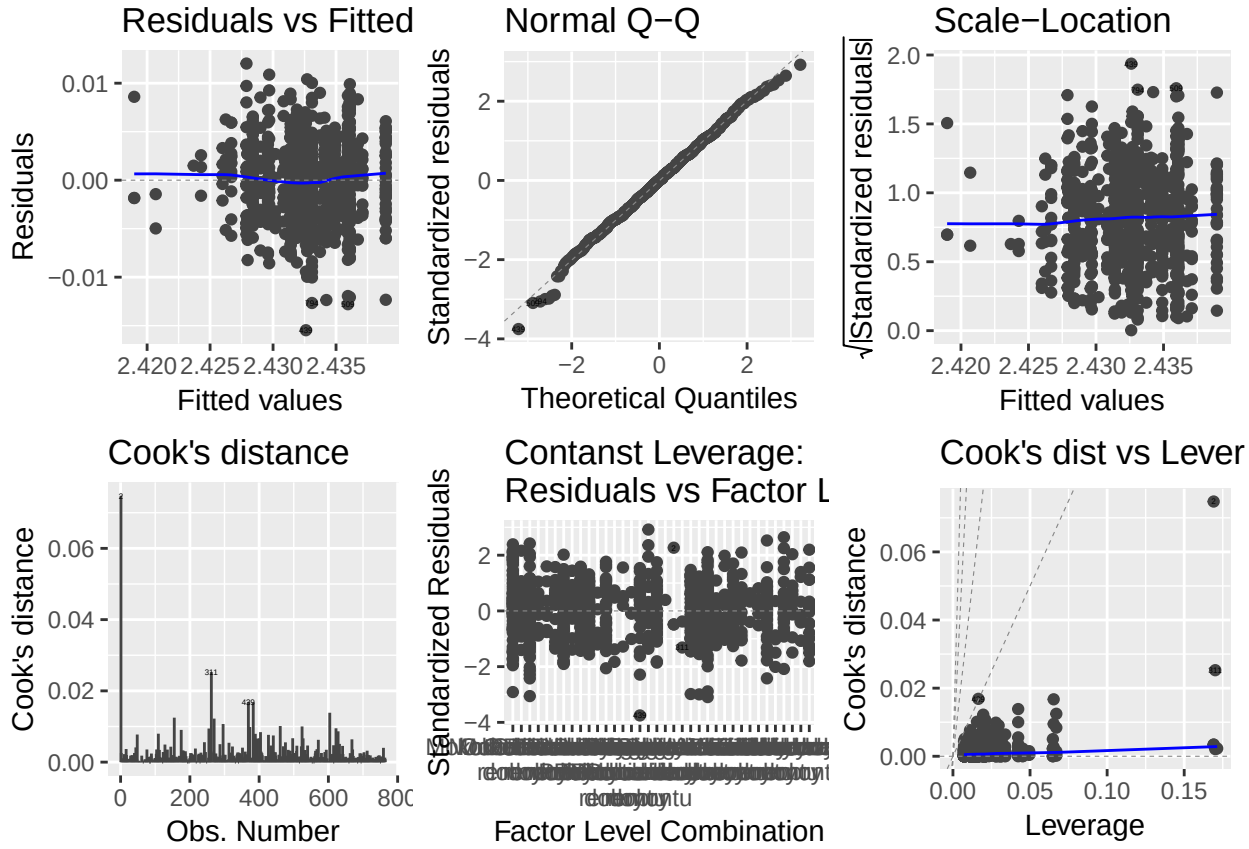
Budujemy więc model tylko na podstawie zmiennych kategorycznych, nie ilościowych.

```
lm.apartments <- lm(m2.price~reorder(district, m2.price) +
                      reorder(condition, m2.price), apartments)

lambda <- boxcox(lm.apartments, plotit=F)$x[which.max(boxcox(lm.apartments, plotit=F)$y)]
apartmentsBCSimple <- apartments
apartmentsBCSimple$m2.price <- (apartmentsBCSimple$m2.price^lambda-1)/lambda

lm.apartmentsBCSimple <- lm(m2.price~reorder(district, m2.price) +
                      reorder(condition, m2.price), apartmentsBCSimple)

summary(lm.apartmentsBCSimple)
```

```
##
## Call:
## lm(formula = m2.price ~ reorder(district, m2.price) + reorder(condition,
##     m2.price), data = apartmentsBCSimple)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0154705 -0.0028125  0.0000535  0.0027257  0.0120234
##
## Coefficients:
##                                       Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                          2.4189767  0.0017097 1414.834  < 2e-16 ***
## reorder(district, m2.price)Wschod    0.0053034  0.0019069    2.781 0.005552 **
## reorder(district, m2.price)Praga Polnoc  0.0072692  0.0019919    3.649 0.000281 ***
## reorder(district, m2.price)Polnoc    0.0077063  0.0017536    4.395 1.27e-05 ***
## reorder(district, m2.price)Praga Poludnie 0.0089128  0.0017709    5.033 6.06e-07 ***
## reorder(district, m2.price)Zachod    0.0090077  0.0017992    5.007 6.91e-07 ***
## reorder(district, m2.price)Wola      0.0118621  0.0017890    6.631 6.39e-11 ***
## reorder(district, m2.price)Poludnie  0.0112099  0.0017715    6.328 4.28e-10 ***
```

```
## reorder(district, m2.price)Zoliborz       0.0130383  0.0018291    7.128 2.39e-12 ***
## reorder(district, m2.price)Mokotow        0.0124028  0.0017326    7.159 1.94e-12 ***
## reorder(district, m2.price)Ochota         0.0133992  0.0018399    7.283 8.29e-13 ***
## reorder(district, m2.price)Srodmiescie    0.0152561  0.0017488    8.724  < 2e-16 ***
## reorder(condition, m2.price)dobry         0.0017104  0.0004376    3.909 0.000101 ***
## reorder(condition, m2.price)bardzo dobry  0.0047196  0.0004442   10.624  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.004158 on 750 degrees of freedom
## Multiple R-squared:  0.3881, Adjusted R-squared:  0.3775
## F-statistic: 36.59 on 13 and 750 DF,  p-value: < 2.2e-16
```

Jak widać taki model lepiej dopasowuje się do danych gorzej niż poprzedni, R^2 jest niższe. P-wartości dla każdego z współczynników są małe. rozkład residów jest bardziej symetryczny. Kolej na wykresy diagnstyczne:



Wyglądają dużo lepiej niż dla najprostszego modelu, zdają się być także lepsze od tych za modelu zaawansowago.

```
##
##  Shapiro-Wilk normality test
##
## data:  stdres(lm.apartmentsBCSimple)
## W = 0.9974, p-value = 0.272
```

```
##
##  studentized Breusch-Pagan test
##
```

```
## data:  lm.apartmentsBCSimple
## BP = 19.362, df = 13, p-value = 0.1123
```

```
##
##  Goldfeld-Quandt test
##
## data:  lm.apartmentsBCSimple
## GQ = 1.1302, df1 = 368, df2 = 368, p-value = 0.1204
## alternative hypothesis: variance increases from segment 1 to 2
```

```
##
##  Durbin-Watson test
##
## data:  lm.apartmentsBCSimple
## DW = 1.9668, p-value = 0.1823
## alternative hypothesis: true autocorrelation is greater than 0
```

Tym razem nie mamy podstaw, do odrzucenia hipotez o dobrym rozkładzie residuów, jednakże zostało to obarczone sporą stratą wartości R^2. Normalność residuów jest spełniona, jednorodność także. Jak sie okazało bardzo prosty model spełnia założenia, ale niekoniecznie spełnia nasze oczekiwania.

Podsumowanie

Niestety, na podstawie analizowanych danych ciężko jest uzyskać satysfakconujące wnioski. Byc może przy użyciu bardziej zaawansowanych metod lub z poświęceniem większej ilości czasu by się to udało. W analizowanym przez nas zbiorze nie są widoczne (przynajmniej dla mnie), żadne poważne zależności pomiędzy ceną, a pozostałymi zmiennymi. Owszem widać ja, ale nie wpływają one na cenę na tyle mocno, aby zbudować model, który dawałby odpowiednie podstawy do skorzystania z niego. Przykładowo, przy mieszkania starych bardzo duży wpływ na cenę może mieć stan ogólny budynku, który nie został ujęty w tabeli. Istotny wpływ mogą mieć też dodatkowe informacje, jak okoliczna infrastruktura dla dzieci, bliskość ciągów komunikacyjnych.

Być może przy podziale obecne struktury bardziej znaleźlibyśmy więcej zależności. Mi udało się skonstruować dwa, względnie sensowne modele, jeden bardziej dopasowywujący się do danych, ale nie spełniający założeń, drugi z kolei dopasowany gorzej, ale zo te spełniający wymagane założenia. Mimo usilnych prób nie udało mi się uzyskać satysfakcjonujących efektów, które łączyłyby zalety obydwu regresji.

Poniżej prezentuję współczynniki uzyskane w każdym z modeli.

```
coefficients(lm.apartmentsBCSimple)
```

```
##                                (Intercept)         reorder(district, m2.price)Wschod
##                                2.418976739                               0.005303443
##   reorder(district, m2.price)Praga Polnoc         reorder(district, m2.price)Polnoc
##                                0.007269192                               0.007706318
## reorder(district, m2.price)Praga Poludnie         reorder(district, m2.price)Zachod
##                                0.008912773                               0.009007725
##           reorder(district, m2.price)Wola       reorder(district, m2.price)Poludnie
##                                0.011862078                               0.011209863
##       reorder(district, m2.price)Zoliborz        reorder(district, m2.price)Mokotow
##                                0.013038253                               0.012402814
##         reorder(district, m2.price)Ochota    reorder(district, m2.price)Srodmiescie
##                                0.013399235                               0.015256126
##         reorder(condition, m2.price)dobry  reorder(condition, m2.price)bardzo dobry
##                                0.001710353                               0.004719599
```

```r
coefficients(lm.apartmentsBC)
```

```
##                                         (Intercept)
##                                       -1.376553e+04
##                    reorder(district, m2.price)Wschod
##                                        2.504615e-01
##              reorder(district, m2.price)Praga Polnoc
##                                        2.837288e-01
##                    reorder(district, m2.price)Polnoc
##                                        3.106498e-01
##           reorder(district, m2.price)Praga Poludnie
##                                        3.696504e-01
##                    reorder(district, m2.price)Zachod
##                                        3.671269e-01
##                      reorder(district, m2.price)Wola
##                                        4.726464e-01
##                  reorder(district, m2.price)Poludnie
##                                        4.519570e-01
##                  reorder(district, m2.price)Zoliborz
##                                        5.144091e-01
##                   reorder(district, m2.price)Mokotow
##                                        4.909151e-01
##                    reorder(district, m2.price)Ochota
##                                        5.134356e-01
##               reorder(district, m2.price)Srodmiescie
##                                        5.807687e-01
##                    reorder(condition, m2.price)dobry
##                                        6.300433e-02
##             reorder(condition, m2.price)bardzo dobry
##                                        1.240782e-01
##                          reorder(floor, m2.price)5-8
##                                        7.757933e-01
##                          reorder(floor, m2.price)1-4
##                                        1.033117e+00
##                               I(construction.date^3)
##                                        1.842940e-06
##                               I(construction.date^2)
##                                       -1.081164e-02
##                                    construction.date
##                                        2.113855e+01
##                  reorder(n.rooms, m2.price)3:nowFALSE
##                                       -1.095592e-01
##                  reorder(n.rooms, m2.price)2:nowFALSE
##                                       -4.361142e-02
##                  reorder(n.rooms, m2.price)1:nowFALSE
##                                       -7.945003e-03
##                 reorder(n.rooms, m2.price)4+:nowFALSE
##                                       -2.938067e-02
##                   reorder(n.rooms, m2.price)3:nowTRUE
##                                        4.021862e-02
##                   reorder(n.rooms, m2.price)2:nowTRUE
##                                        3.566433e-02
##                   reorder(n.rooms, m2.price)1:nowTRUE
##                                        4.196798e-02
```

```
##               reorder(n.rooms, m2.price)4+:nowTRUE
##                                                NA
## reorder(floor, m2.price)5-8:I(construction.date^3)
##                                      -9.791358e-11
## reorder(floor, m2.price)1-4:I(construction.date^3)
##                                      -1.325916e-10
```

```
coefficients(lm.district)
```

```
##                                (Intercept)         reorder(district, m2.price)Wschod
##                                   5596.667                                  1535.625
##   reorder(district, m2.price)Praga Polnoc         reorder(district, m2.price)Polnoc
##                                   1677.146                                  1986.036
## reorder(district, m2.price)Praga Poludnie         reorder(district, m2.price)Zachod
##                                   2252.488                                  2509.013
##           reorder(district, m2.price)Wola       reorder(district, m2.price)Poludnie
##                                   2944.905                                  3197.090
##       reorder(district, m2.price)Zoliborz        reorder(district, m2.price)Mokotow
##                                   3422.965                                  3576.926
##         reorder(district, m2.price)Ochota    reorder(district, m2.price)Srodmiescie
##                                   3596.790                                  4456.926
```