AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE
AGH UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Regresja parametryczna, a regresja nieparametryczna

Autorzy:

Majka Miezianko

Marek Czerwonka

Projekt zaliczeniowy z przedmiotu:

Metody Nieparametryczne w Statystyce

Kraków 2021

W sprawozdaniu prezentujemy jedynie część wygenerowanych wykresów, aby nie miało ono zbyt dużej objętości. Te których nie wklejamy, znajdują się w dołączonym pliku html (ale nie twierdzimy, że na pewno są one warto przeglądania).

Do analizy wybraliśmy zestaw zmiennych, które nie generują w sposób oczywisty zależności przyczynowo skutkowej, chcieliśmy też w miarę możliwości uniknąć tych najczęściej używanych do tego typu celów danych, czyli tych dotyczących stopy bezrobocia czy dochodów powiatów. Nie chcieliśmy bowiem tłumaczyć „podobnego przez podobne". Jeżeli bowiem wydatki będziemy tłumaczyli dochodami bieżącymi lub wcześniejszymi o jeden lub kilka okresów, uzyskanie wysokiego współczynnika determinacji byłoby oczekiwane, ale wartość poznawcza czegoś takiego byłaby łagodnie pisząc mierna.

Zwróćmy tez uwagę, że w projekcie mamy budować modele z tylko jednym predyktorem. Tak więc uzyskanie wysokiego R^2 nie było naszym bezpośrednim celem.

Nasz zestaw danych obejmował dane pochodzące z BDL GUS, dotyczące powiatów. W każdym z przypadków nie były to dane pełne (występowały braki danych).

Uwzględniliśmy następujące zmienne:

- liczba mieszkańców powiatu przypadających na jedno miejsce w kinach stałych;
- liczba boisk piłkarskich na obszarze powiatu;
- gęstość zaludnienia w powiecie;
- wysokość dochodów powiatu;
- liczba miejsc w kinach stałych na terenie powiatu.

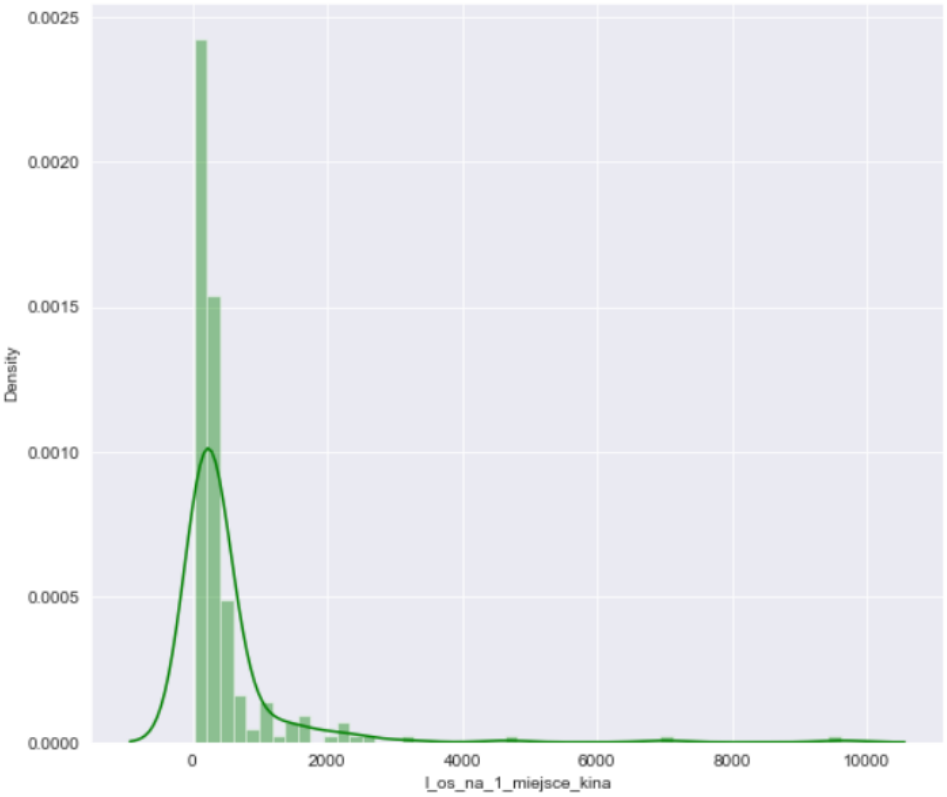
Dane wymagały wstępnego oczyszczenia, ponieważ czasami braki danych reprezentowane były przez zera, zamiast kodu braku danych, co zdecydowanie zaburzyłoby wyniki. W przypadku różnych zmiennych, różna była liczba dostępnych danych, aby doprowadzić do porównywalności analiz dotyczących poszczególnych zmiennych zastosowaliśmy usuwanie danych przypadkami, w efekcie czego liczba powiatów z kompletnymi danymi wynosi u nas na tym etapie analizy 224.

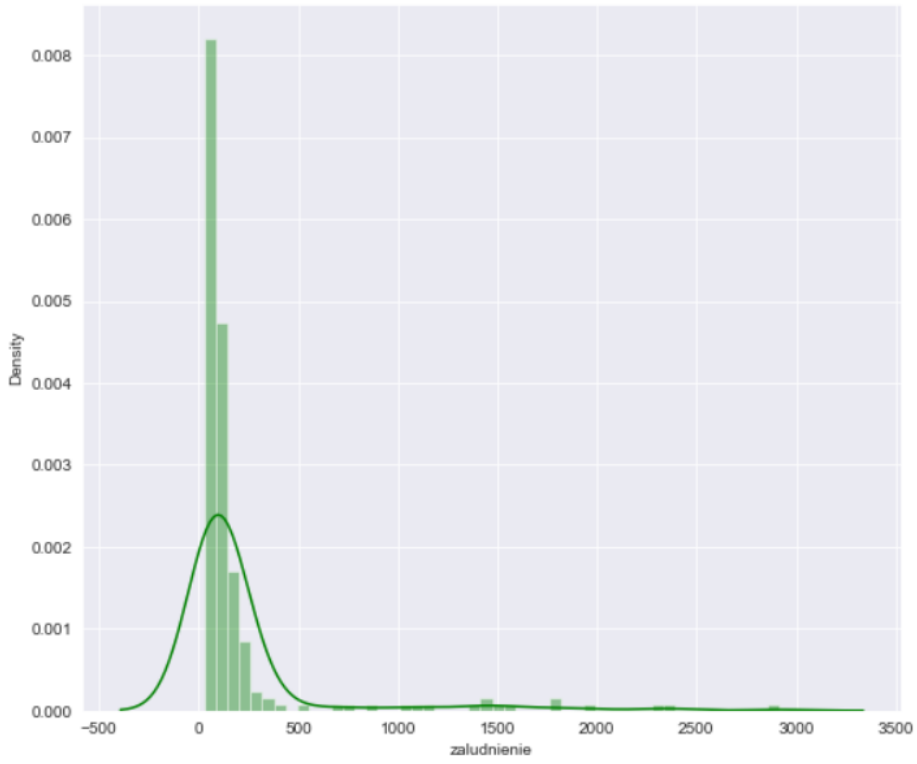
Wyznaczone dla tej podpróby statystyki opisowe prezentujemy poniżej:

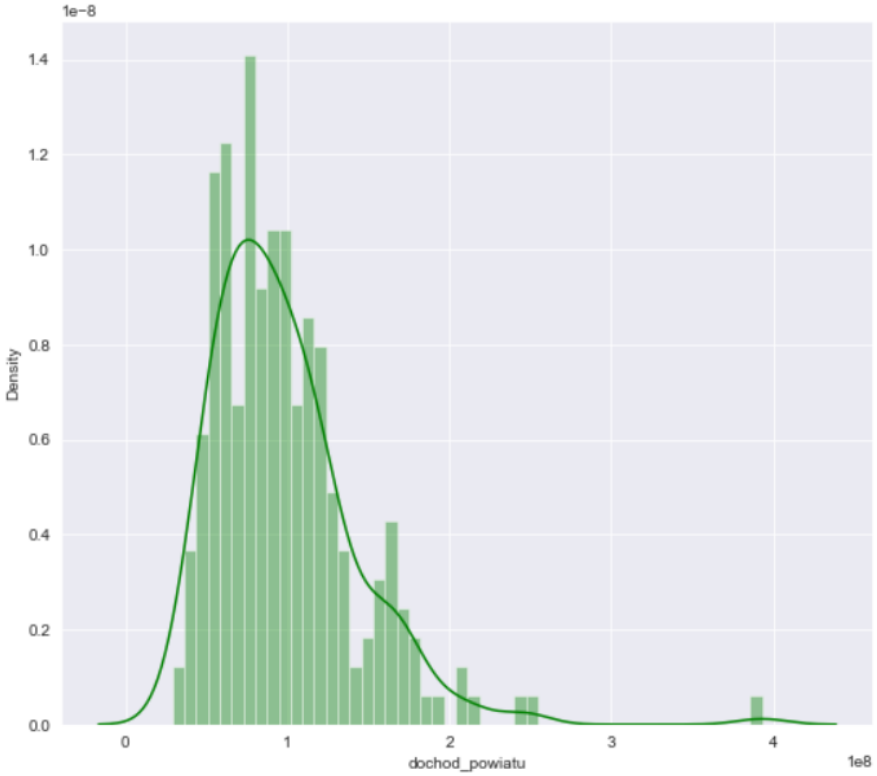
:

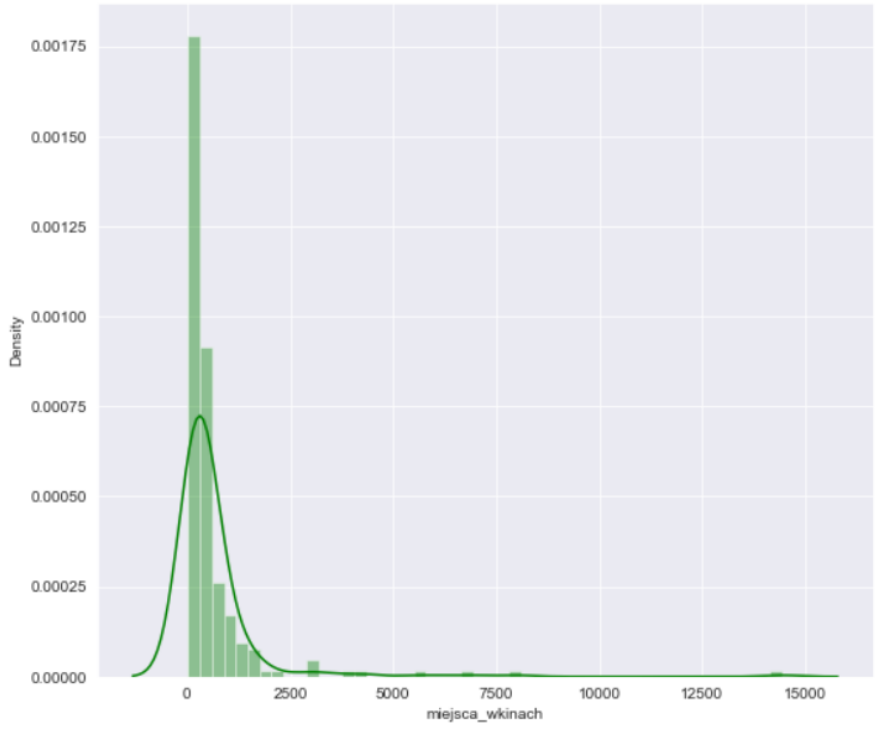
	l_os_na_1_miejsce_kina	boiska	zaludnienie	dochod_powiatu	miejsca_wkinach
count	224.000000	224.000000	224.000000	2.240000e+02	224.000000
mean	480.848214	21.366071	216.004464	9.770999e+07	623.209821
std	938.474382	16.612975	417.802778	4.497751e+07	1324.345330
min	30.000000	1.000000	31.000000	2.899865e+07	20.000000
25%	152.750000	9.000000	63.750000	6.523544e+07	189.500000
50%	234.500000	17.500000	94.000000	8.975078e+07	303.000000
75%	406.500000	29.000000	147.250000	1.184994e+08	488.250000
max	9603.000000	98.000000	2913.000000	3.930423e+08	14443.000000

Następnie w celu wizualnej oceny charaktery rozkładów zmiennych, wygenerowaliśmy histogramy. Już jakościowa analiza histogramów pozwala ocenić te rozkłady jako wyraźnie prawo-skośne.


0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
Density
0
2000
4000
6000
8000
10000
l_os_na_1_miejsce_kina


0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
Density
−500
0
500
1000
1500
2000
2500
3000
3500
zaludnienie

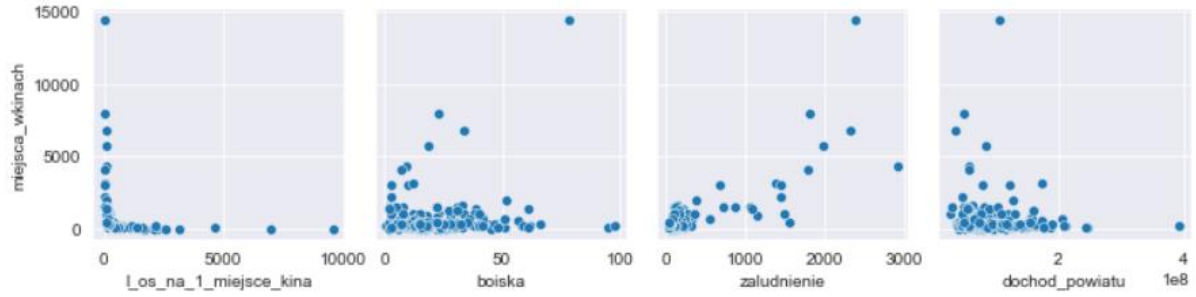




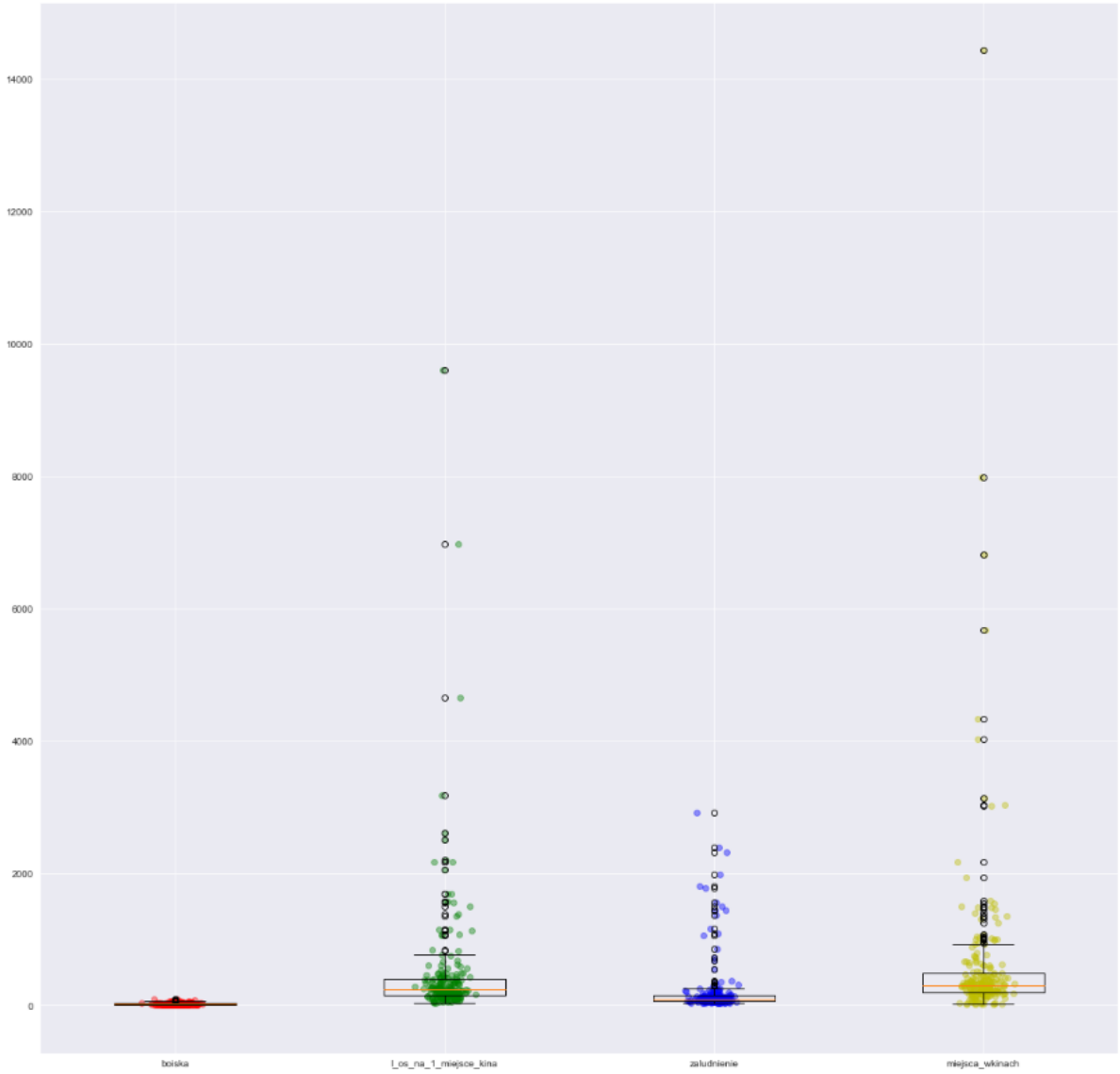
Na histogramy nałożyliśmy krzywą gęstości (a zatem potraktowaliśmy wszystkie zmienne jako zmienne ciągłe, abstrahując od faktu, iż zasadniczo wszystkie z nich mogą przyjmować jedynie wartości dyskretne, ale przy tym są to wartości dostatecznie gęsto rozłożone).

Naszym założeniem było przyjęcie jako zmiennej objaśnianej liczby miejsc w kinach stałych, której zmienność będziemy wyjaśniali zmiennością kolejnych zmiennych (naturalnie oprócz zmiennej reprezentującej liczbę osób przypadającej na jedno miejsce w kinie, bo taka zależność sprowadzałaby się do ustalania liczby mieszkańców powiatu i nie niosłaby ze sobą żadnej wartości poznawczej).

Narysowaliśmy zatem wstępne wykresy rozrzutu oraz wyznaczyliśmy wartości współczynników korelacji liniowej r-Pearsona.

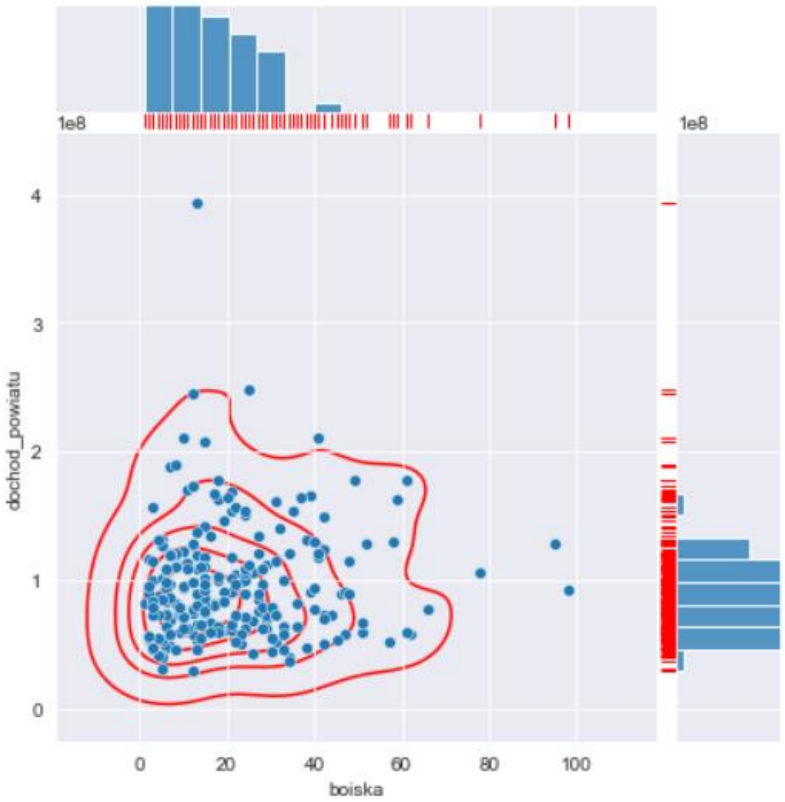


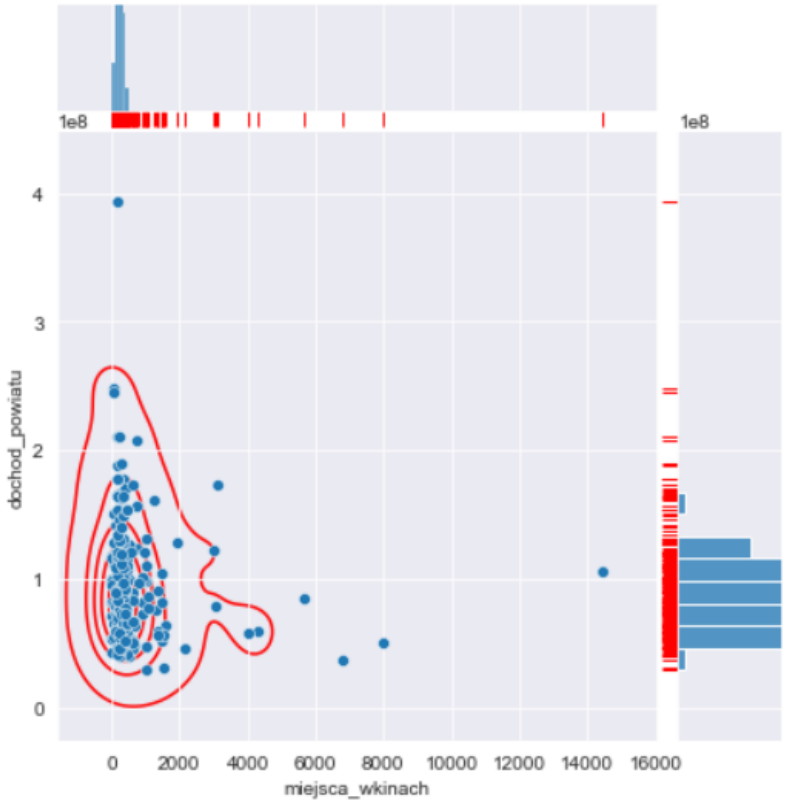
Na tym etapie dokonamy dalszego oczyszczenia danych z wartości skrajnych, które mocno wpływałyby na charakter zależności modelowanej klasyczną MNK.



Z rzeczy które być może warte są większej uwagi, to oprócz standardowych pudełek, wygenerowaliśmy także wykresy, które obrazują dwuwymiarowe rozkłady par zmiennych wraz z nałożonymi poziomicami czegoś w rodzaju powierzchni gęstości (bez odcięcia jej przy zerze, tak że generalnie ich lewe części nie mają sensu).

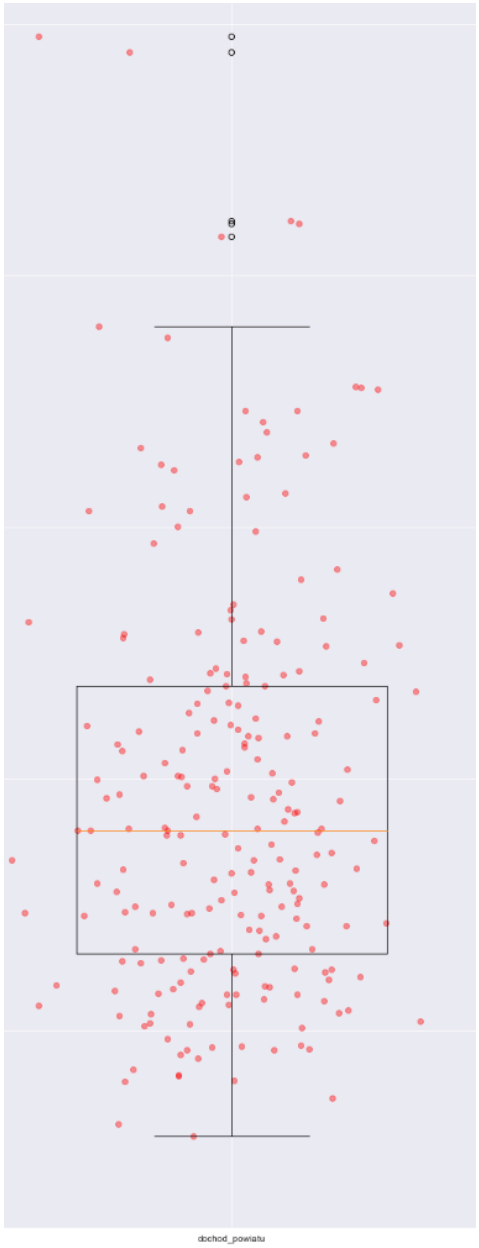
Poniżej wklejamy dwa przykładowe wykresy z tej kategorii.


1e8
1e8
dochod_powiatu
0
1
2
3
4
0
20
40
60
80
100
boiska

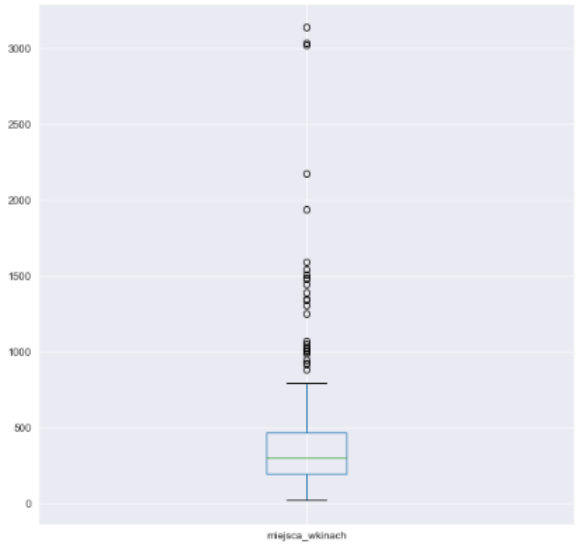

1e8
1e8
dochod_powiatu
0
1
2
3
4
0
2000
4000
6000
8000
10000
12000
14000
16000
miejsca_wkinach

Przeprowadziliśmy rozsądne usuwanie zidentyfikowanych obserwacji odstających (szczegóły w pliku html), czyli nie robiliśmy tego mechanicznie, aby w sposób sztuczny doprowadzić do ich braku, a każdą ze wstępnie zakwalifikowanych do ewentualnego usunięcia danych (standardowym kryterium odległości od średniej wyższej niż 3σ) ocenialiśmy czy jest z jakichś powodów przypadkiem **realnie mocno nietypowym** (np. jest Warszawą albo innym bardzo dużym miastem na prawach powiatu) czy też może reprezentuje tylko immanentną cechę danej zmiennej, która ze swojej natury charakteryzuje się skośnością. Szczególnie w przypadku zmiennych objaśniających nie dbaliśmy nadmiernie o usuwanie skośności.

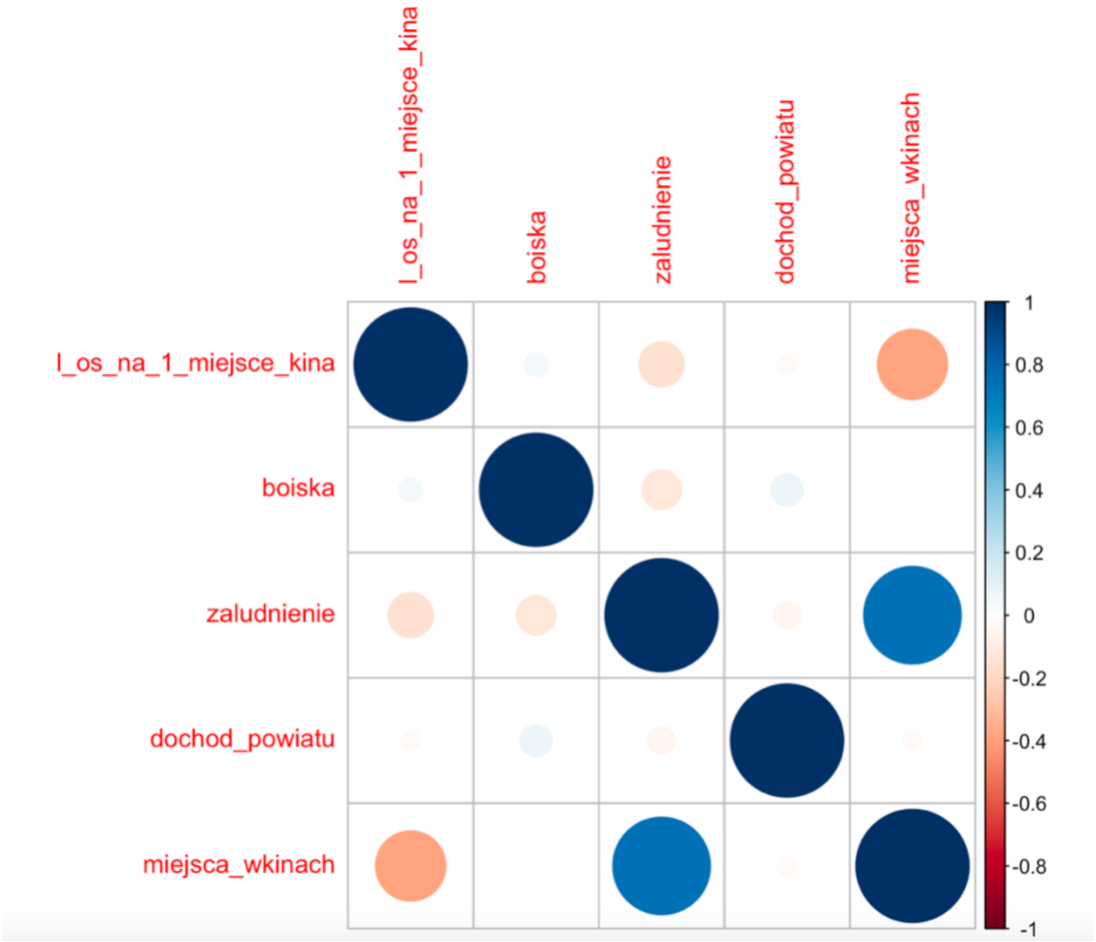
Poniżej prezentujemy wykres po usunięciu outlierów (**w zbiorze danych ostatecznie pozostało 209 obserwacji**) dla dochodów powiatu:



oraz miejsca w kinach stałych:



Graficzna prezentacja macierzy korelacji dla danych po ich oczyszczeniu do stanu jaki uznaliśmy za wystarczający.



Badanie normalności zmiennych:

```
        Shapiro-Wilk normality test

data:  data_cleaned2$zaludnienie
W = 0.44036, p-value < 2.2e-16
```

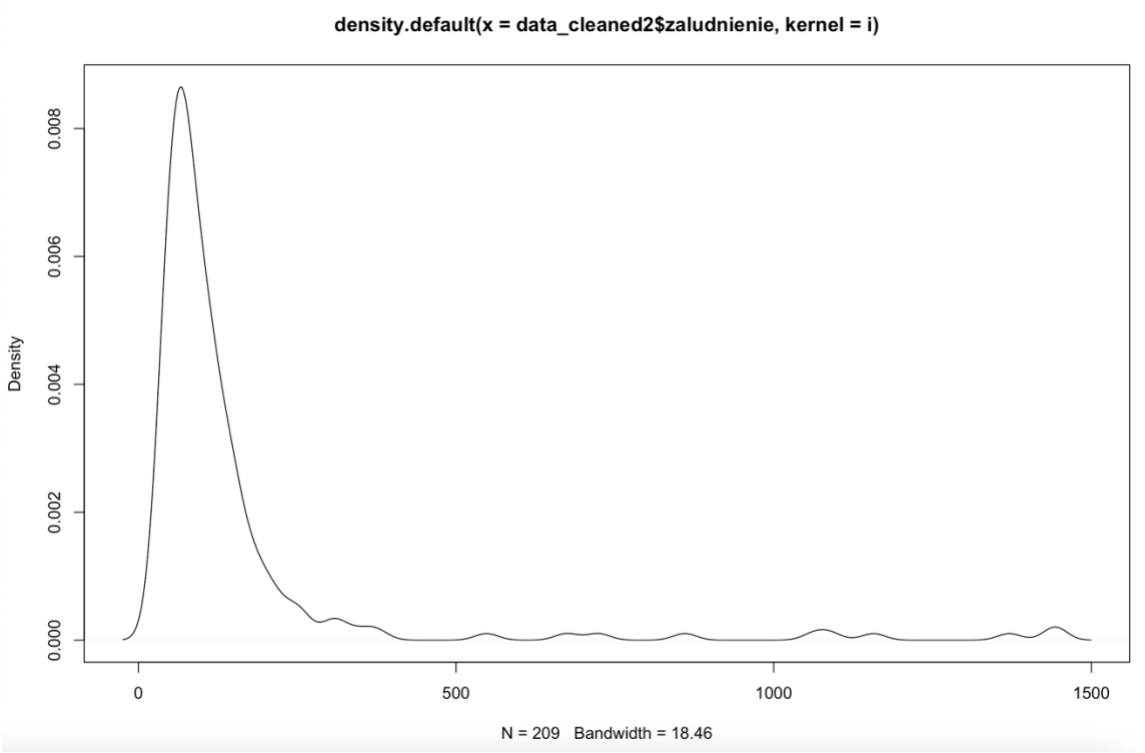
```
        Shapiro-Wilk normality test

data:  data_cleaned2$miejsca_wkinach
W = 0.659, p-value < 2.2e-16
```

```
        Shapiro-Wilk normality test

data:  data_cleaned2$dochod_powiatu
W = 0.95003, p-value = 1.175e-06
```

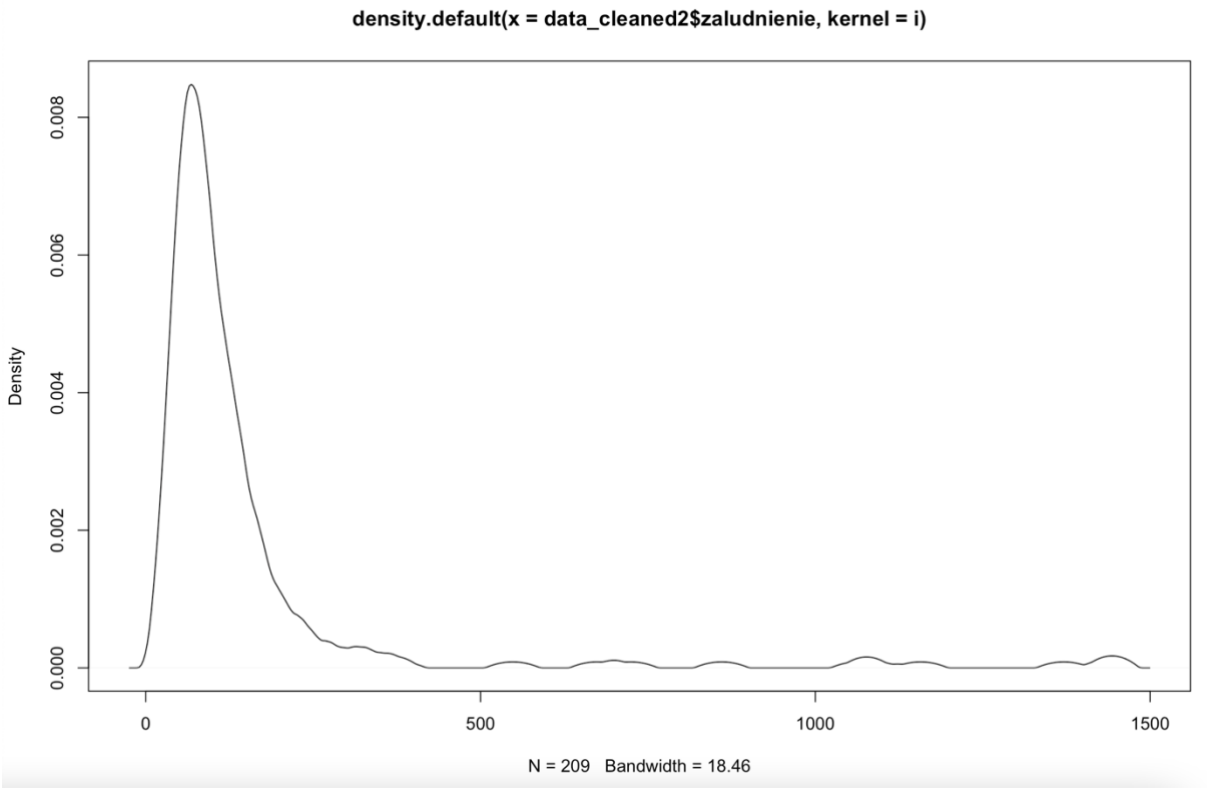
Jak widać oczyszczenie danych w żadnym razie nie doprowadziło do stanu, w którym ich rozkłady przypominałyby rozkłady normalne, czego z resztą samego z siebie w żadnym razie nie uważamy za stan konieczny.

Zatem w ramach potrenowania estymacji funkcji gęstości dla różnych postaci jądra, dla jednej z naszych zmiennych przeprowadziliśmy tę procedurę kolejno dla pięciu standardowo używanych funkcji:

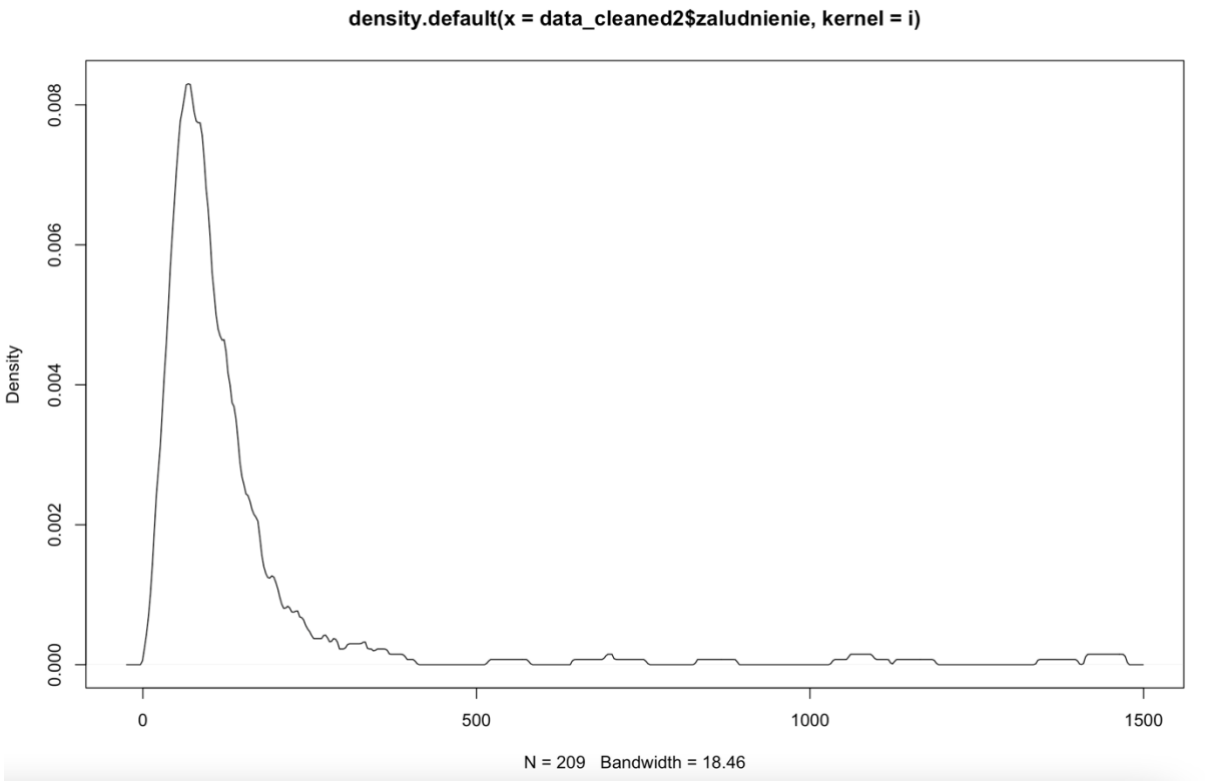
Jądro gaussowskie



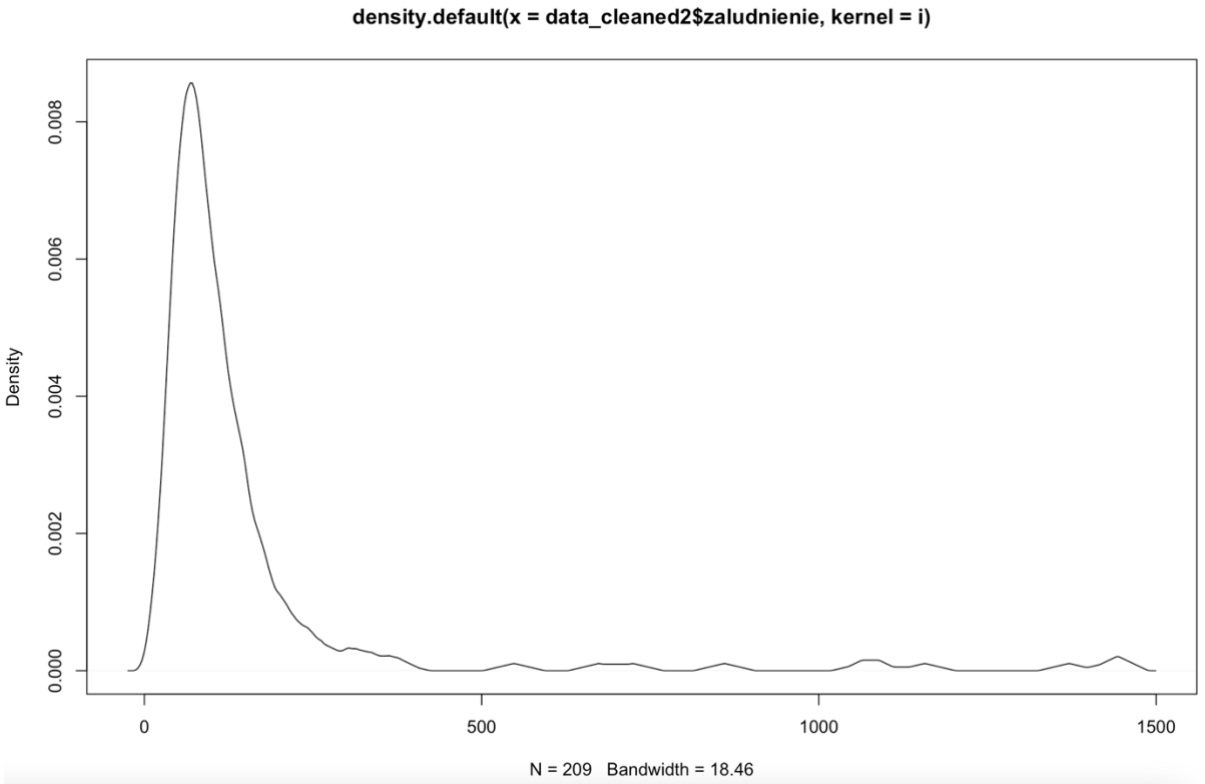
Jądro Epanechnikova



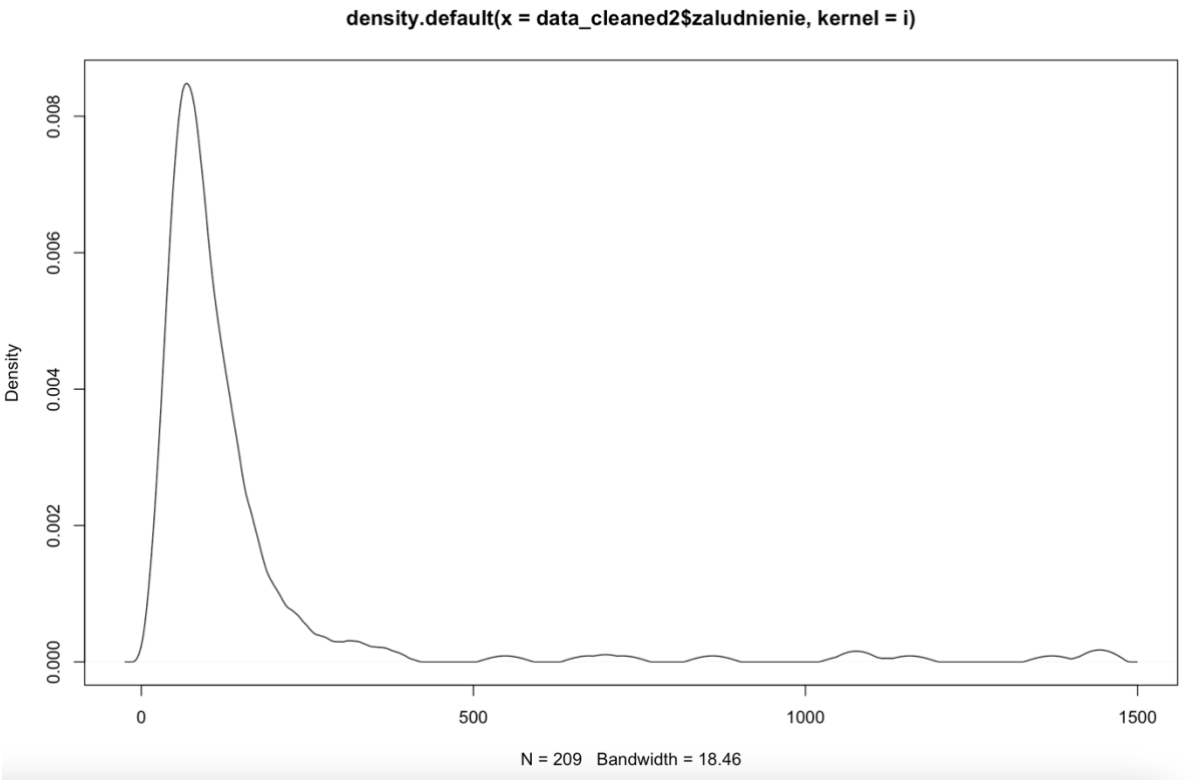
Jądro prostokątne:



Jądro trójkątne



Jądro typu cosinusowego



Następnie oszacowaliśmy kilka modeli klasyczną MNK.

Regresja parametryczna MNK dla zmiennych bez ich przekształceń

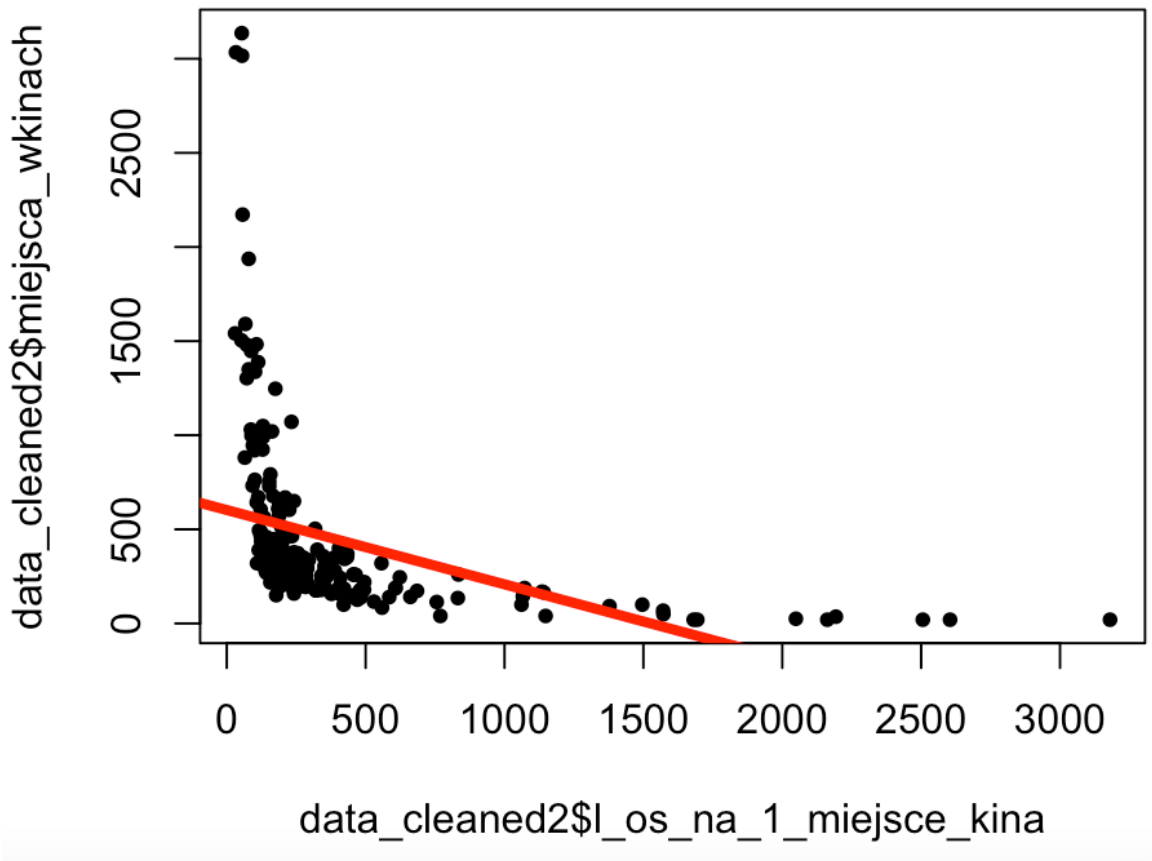
Przyjmujemy zatem, że naszą zmienną objaśnianą będzie liczba dostępnych miejsc w kinach stałych, której zmienność będziemy próbowali wyjaśniać zmiennością kolejnych, branych pojedynczo predyktorów.

Nie są to z pewnością modele marzeń, ale jak rozumiemy nie to było celem tego projektu. Co do rozkładu składnika losowego przyjęliśmy następujące zasady:

1. Ze względu na sporą liczność próby nie demonizujemy odchyleń od normalności rozkładu tych reszt.
2. Ze względu na brak naturalnego porządku w zbiorze danych (w sensie takim jak np. dla szeregu czasowego), a w związku z tym brak potencjalnego efektu mogącego wprost powodować autokorelację reszt (np. sezonowości, cykli itd.) nie demonizujemy kwestii ewentualnej autokorelacji reszt modelu.

3. W Naszych modelach widoczny jest efekt heteroskedastyczności. Przy przyjęciu porządku wraz ze wzrastającym poziomem zmiennej objaśniającej, zazwyczaj wariancja rośnie. Taki efekt wydaje nam się dość groźny dla modelu zarówno w wersji parametrycznej jak i nieparametrycznej. Podejmiemy zatem przynajmniej próbę ustabilizowania wariancji poprzez zastosowanie logarytmicznej transformacji zmiennych. Alternatywą byłoby stosowanie ważonej metody najmniejszych kwadratów zamiast zwykłej MNK (z wyrazem wolnym zależnym od poziomu zmiennej objaśniającej).

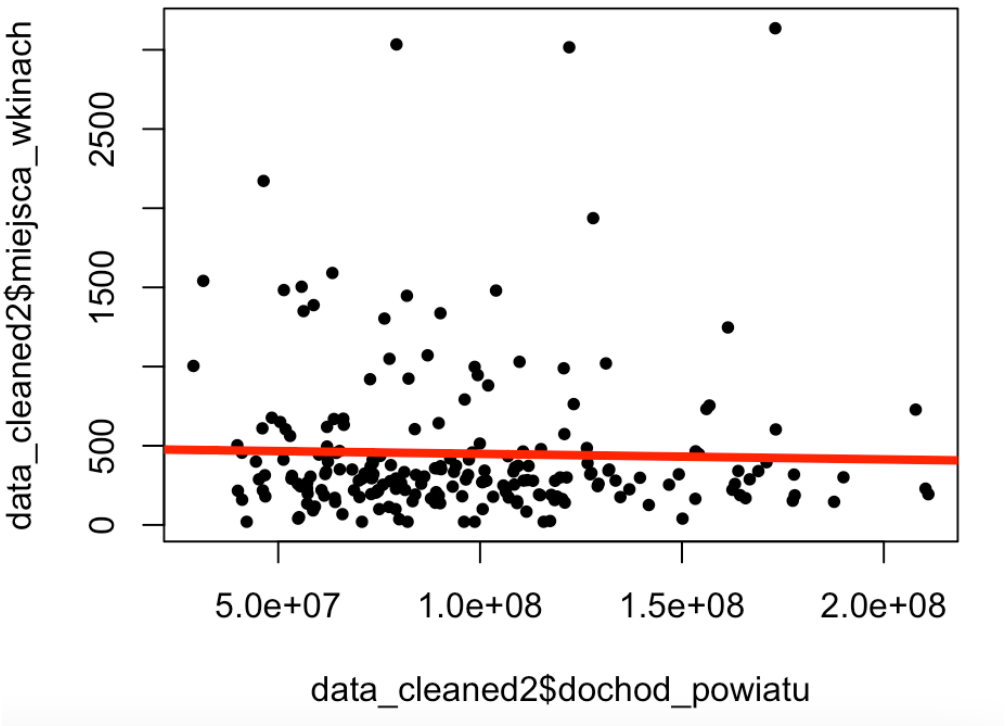
Pierwszy model prezentujemy jedynie w formie ostrzeżenia, do czego prowadzi bezkrytyczne stosowanie MNK. Wzięliśmy do modelu dwie zmienne (liczbę osób przypadającą na jedno miejsce w kinach stałych oraz liczbę miejsc w kinach stałych), które z natury rzeczy powinny być do siebie z grubsza odwrotnie proporcjonalne (zależność typu hiperbolicznego).



```
Call:
lm(formula = data_cleaned2$miejsca_wkinach ~ data_cleaned2$l_os_na_1_miejsce_kina)

Residuals:
    Min      1Q  Median      3Q     Max
-382.60 -238.92 -155.16   66.87 2554.44

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       602.88839   40.05606   15.05  < 2e-16 ***
data_cleaned2$l_os_na_1_miejsce_kina  -0.39488    0.06614   -5.97 1.02e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 447 on 207 degrees of freedom
Multiple R-squared:  0.1469,    Adjusted R-squared:  0.1428
F-statistic: 35.64 on 1 and 207 DF,  p-value: 1.018e-08
```

Jak widać dla tego w oczywisty sposób nonsensownego modelu dostajemy wartość współczynnika determinacji odpowiadającą korelacji liniowej na poziomie prawie 0,4 oraz obydwa parametry strukturalne modelu wysoce istotne statystycznie. Model jednakże nadaje się do śmietnika już choćby na podstawie wizualnej oceny kształtu zależności.

Kolejny model prezentujemy dlatego, że był on dla nas merytorycznym zaskoczeniem. Jak się okazuje liczba miejsc w kinach w powiecie w ogóle istotnie nie zależy od wysokości dochodów tego powiatu (można byłoby się spodziewać korelacji dodatniej)

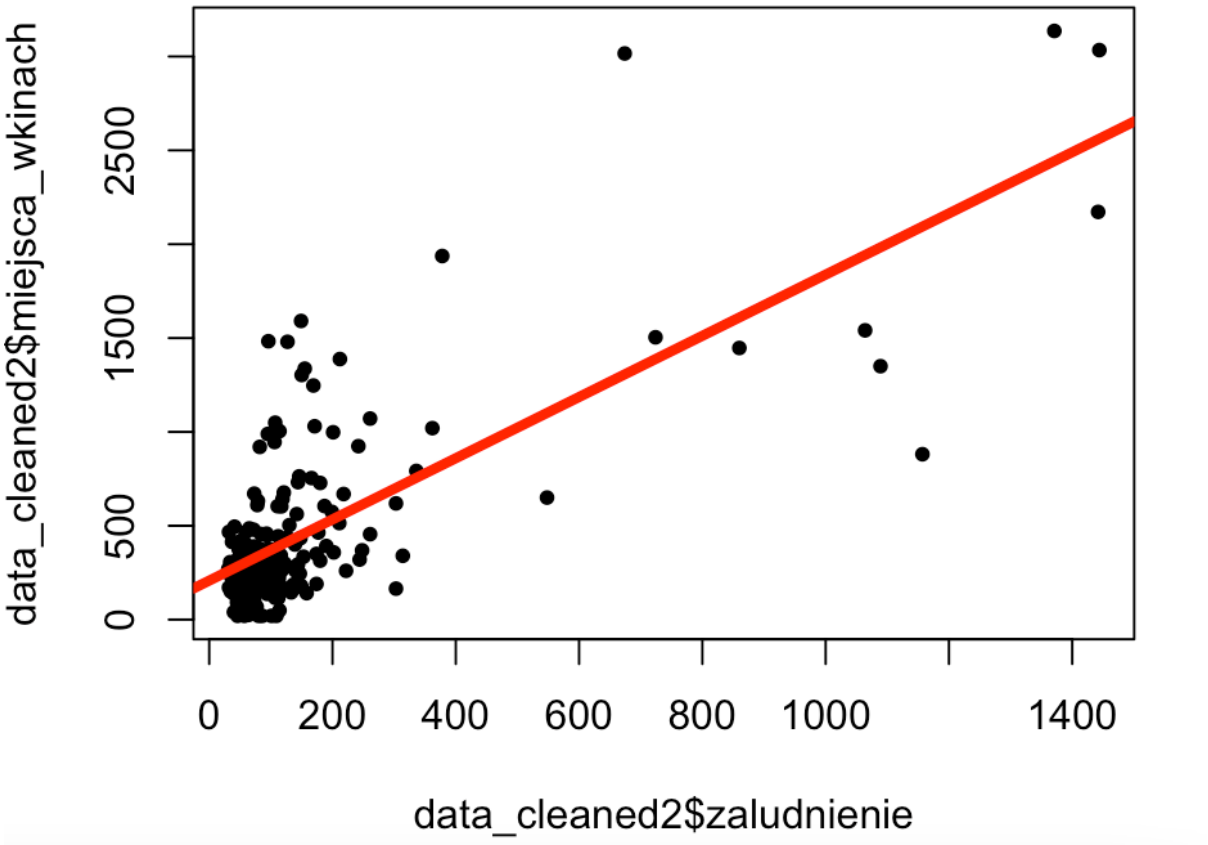


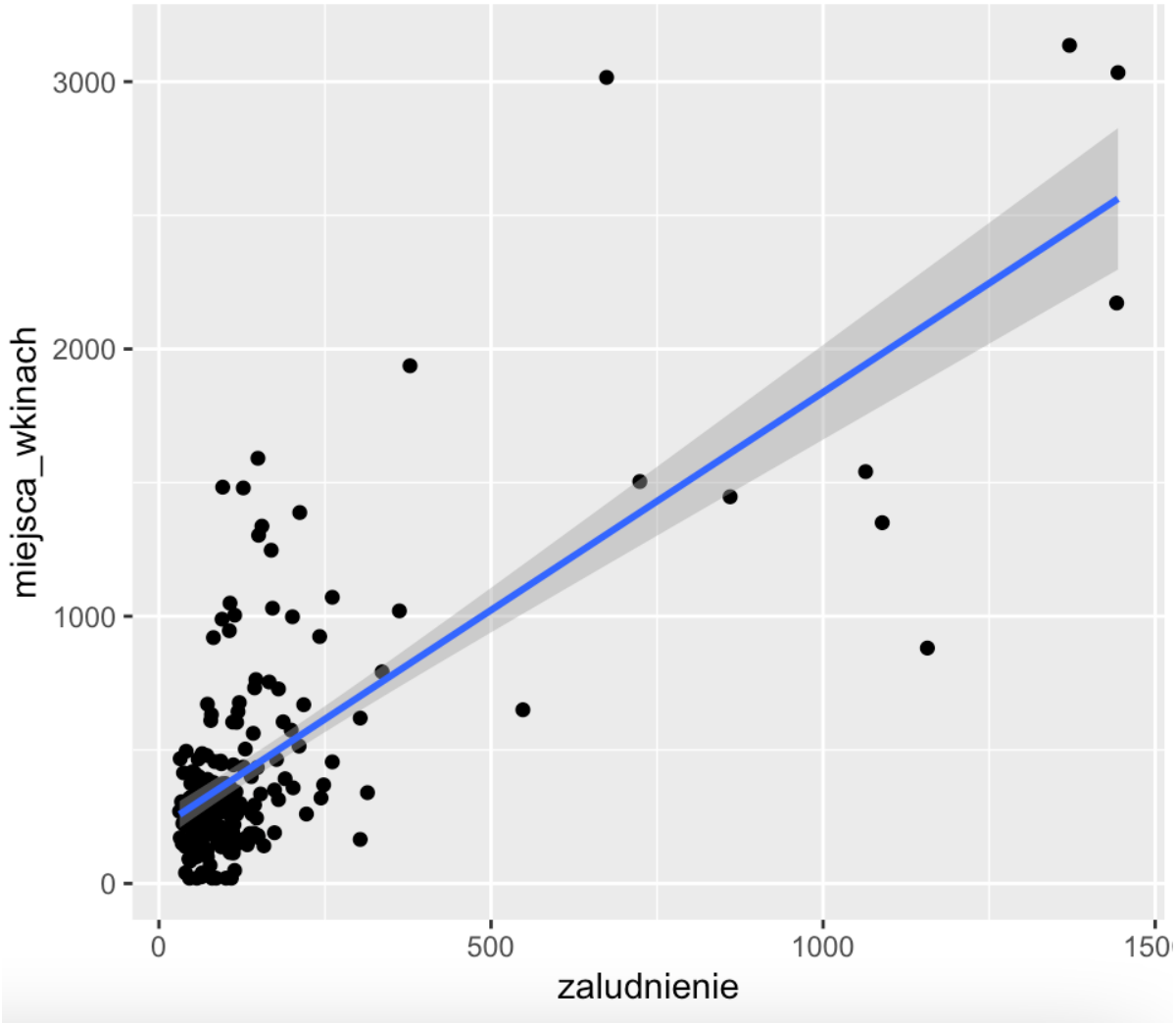

```
Call:
lm(formula = data_cleaned2$miejsca_wkinach ~ data_cleaned2$dochod_powiatu)

Residuals:
    Min      1Q  Median      3Q     Max
-450.01 -253.22 -148.38   17.77 2712.49

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 4.850e+02  9.067e+01   5.349 2.33e-07 ***
data_cleaned2$dochod_powiatu -3.551e-07  8.764e-07  -0.405    0.686
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 483.7 on 207 degrees of freedom
Multiple R-squared:  0.0007926, Adjusted R-squared:  -0.004034
F-statistic: 0.1642 on 1 and 207 DF,  p-value: 0.6857
```

Rekordowo bliski zeru współczynnik determinacji.

Wreszcie trzeci model to ten, który będzie poddany bliższej analizie. Można dyskutować patrząc na ten wykres czy dobrze zrobiliśmy utrzymując wszystkie te dane w jednym zbiorze, bo jest widoczne silne skupienie dla niskich wartości obu tych zmiennych i estymacja modelu tylko dla takiego podzbioru mogłaby dać wyraźnie inne wartości parametrów strukturalnych.

Jednakże ponieważ sama idea wyjaśniania tu zmienności tylko jednym predyktorem ma ograniczony sens, nie analizowaliśmy tego głębiej, przyjmując, że jeśli uzyskamy współczynnik determinacji co najmniej równy 0,5 zostawimy to tak jak jest. A uzyskaliśmy taki.



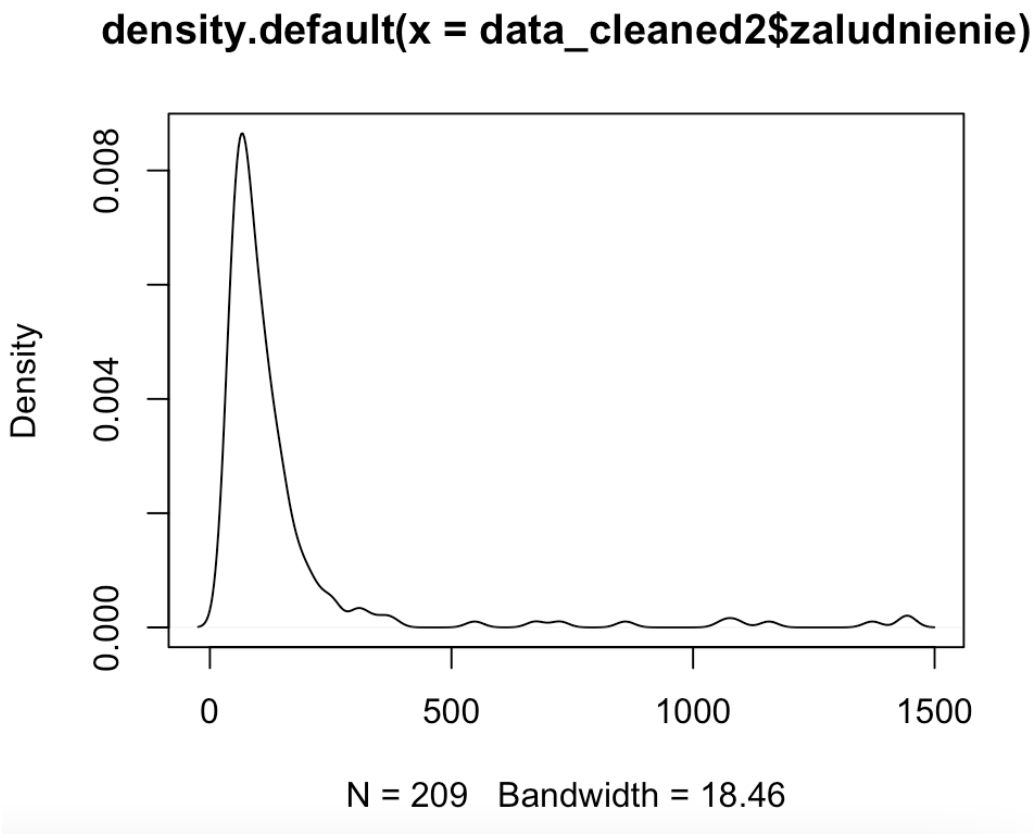


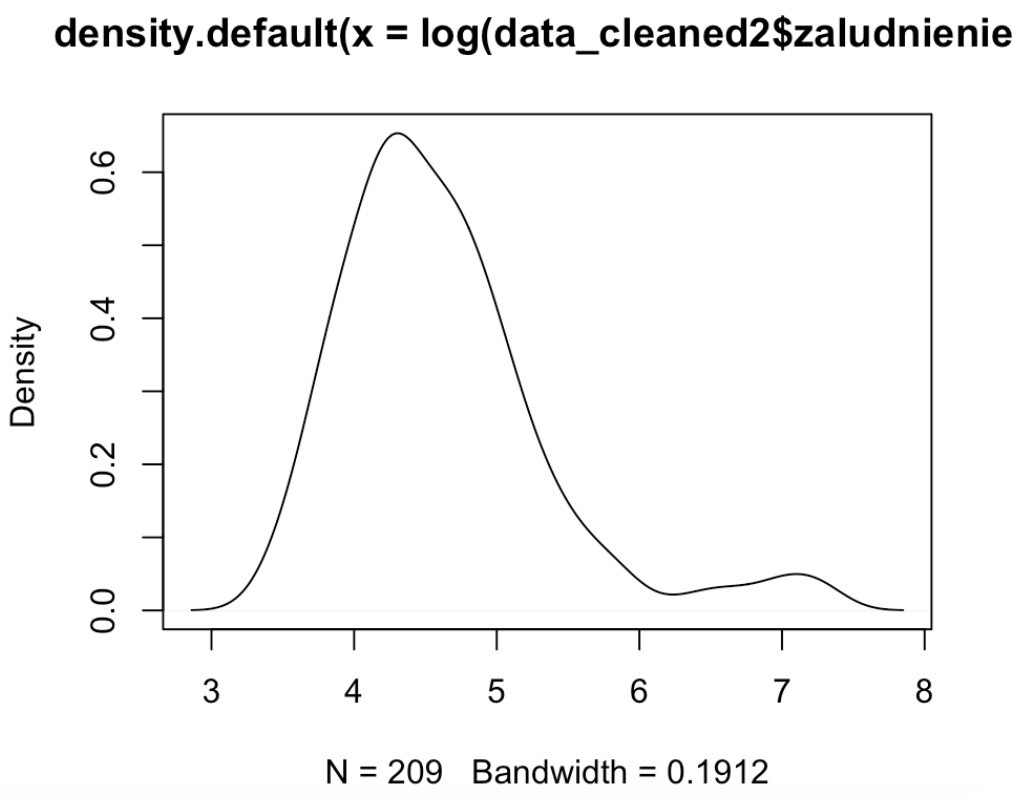

```
Call:
lm(formula = data_cleaned2$miejsca_wkinach ~ data_cleaned2$zaludnienie)

Residuals:
     Min       1Q   Median       3Q      Max
-1213.00  -167.62   -53.23    64.05  1709.48

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               207.6471    27.1046   7.661 6.99e-13 ***
data_cleaned2$zaludnienie   1.6304     0.1022  15.958  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 324 on 207 degrees of freedom
Multiple R-squared:  0.5516,    Adjusted R-squared:  0.5494
F-statistic: 254.7 on 1 and 207 DF,  p-value: < 2.2e-16
```

Jeżeli chodzi o pozostałe dwie zmienne, to niczego interesującego w wynikach nie było.

Następnie podjęliśmy próbę częściowego wystabilizowania wariancji poprzez dokonanie transformacji logarytmicznej zmiennych:

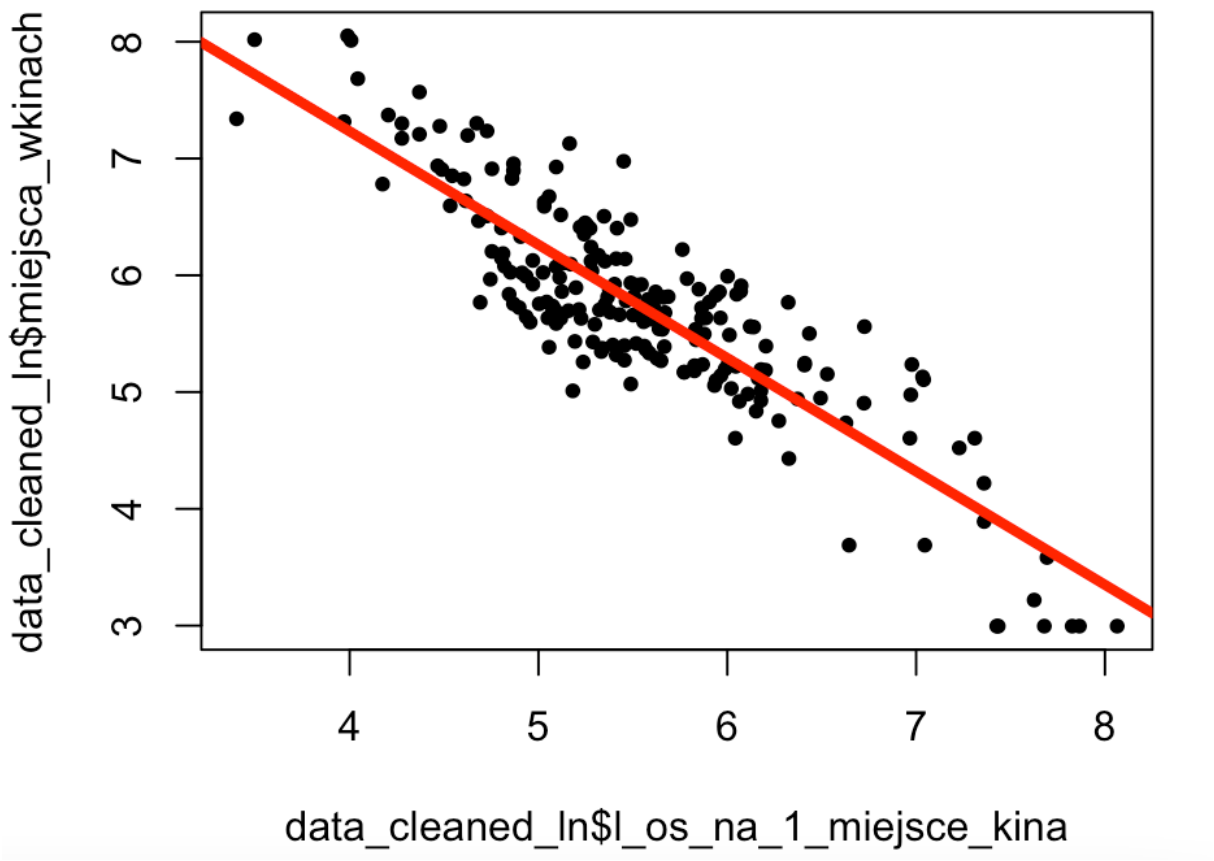
Funkcja gęstości (jądro trójkątne) przed i po dla zmiennej zaludnienie





Dla zmiennych zlogarytmowanych ponownie dopasowaliśmy modele klasyczną MNK.

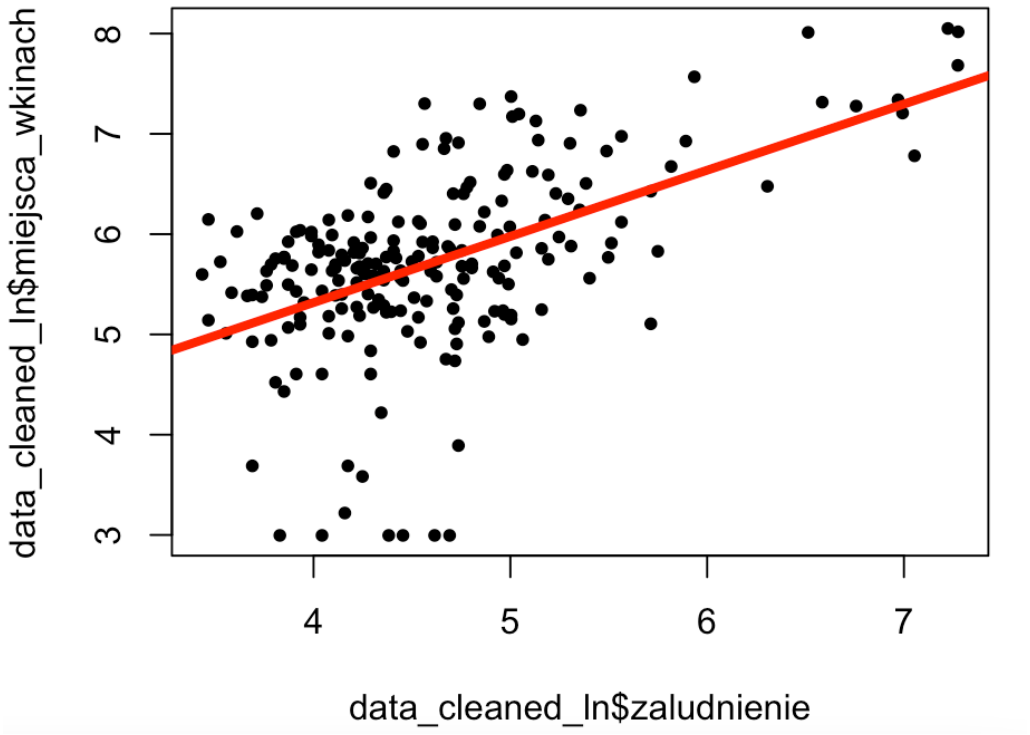
Ponownie zaprezentujemy dwa modele, najpierw ten merytorycznie nonsensowny




```
Call:
lm(formula = data_cleaned_ln$miejsca_wkinach ~ data_cleaned_ln$l_os_na_1_miejsce_kina)

Residuals:
     Min       1Q   Median       3Q      Max
-1.07524 -0.34288 -0.01828  0.33514  1.15190

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          11.1169     0.2106   52.78   <2e-16 ***
data_cleaned_ln$l_os_na_1_miejsce_kina  -0.9709     0.0375  -25.89   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.444 on 207 degrees of freedom
Multiple R-squared:  0.7641,    Adjusted R-squared:  0.763
F-statistic: 670.5 on 1 and 207 DF,  p-value: < 2.2e-16
```

Jak widać tu nastąpiła znaczna poprawa współczynnika determinacji, czego należało oczekiwać w przypadku modelu hiperbolicznego i co naturalnie w żaden sposób nie umniejsza bezsensowności tego modelu.

Natomiast w modelu, o którym uważamy, że ma co najmniej śladowy sens merytoryczny, zlogarytmowanie jednej lub obu ze zmiennych zawsze prowadziło do pogorszenia wartości współczynnika determinacji.

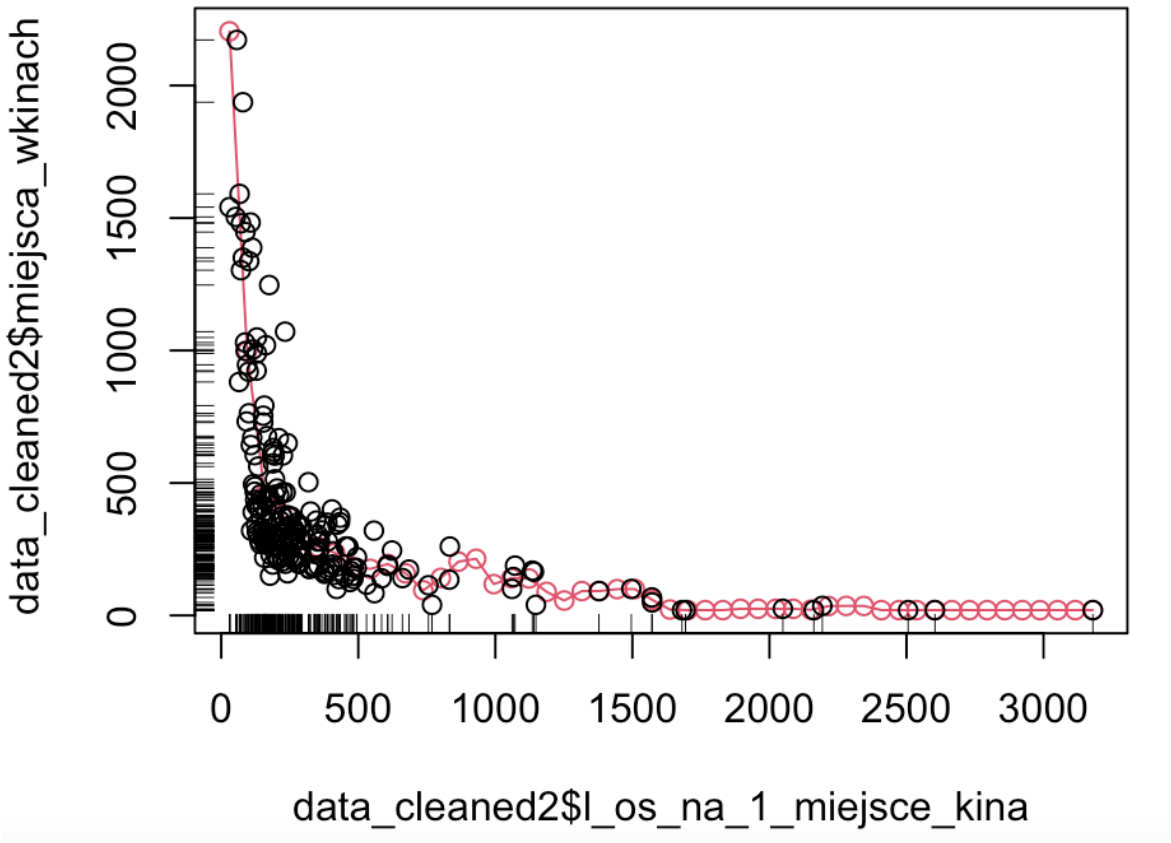



```
Call:
lm(formula = data_cleaned_ln$miejsca_wkinach ~ data_cleaned_ln$zaludnienie)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7782 -0.3411  0.1099  0.4853  1.6117

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 2.67734    0.33959   7.884 1.79e-13 ***
data_cleaned_ln$zaludnienie 0.66006    0.07271   9.078  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7732 on 207 degrees of freedom
Multiple R-squared:  0.2848,    Adjusted R-squared:  0.2813
F-statistic: 82.41 on 1 and 207 DF,  p-value: < 2.2e-16
```

Pociągnijmy zatem dalej ten eksperyment z modelem o wysokim R^2 ale niesensownym merytorycznie.

Regresja nieparametryczna dla danych niezlogarytmowanych. Oszacowanie współczynnika regularyzacji za pomocą parametru „lc" - local-constant estimator (Nadaraya-Watsona) – w mechanizmie walidacji krzyżowej.

l_os_na_1_miejsce_kina



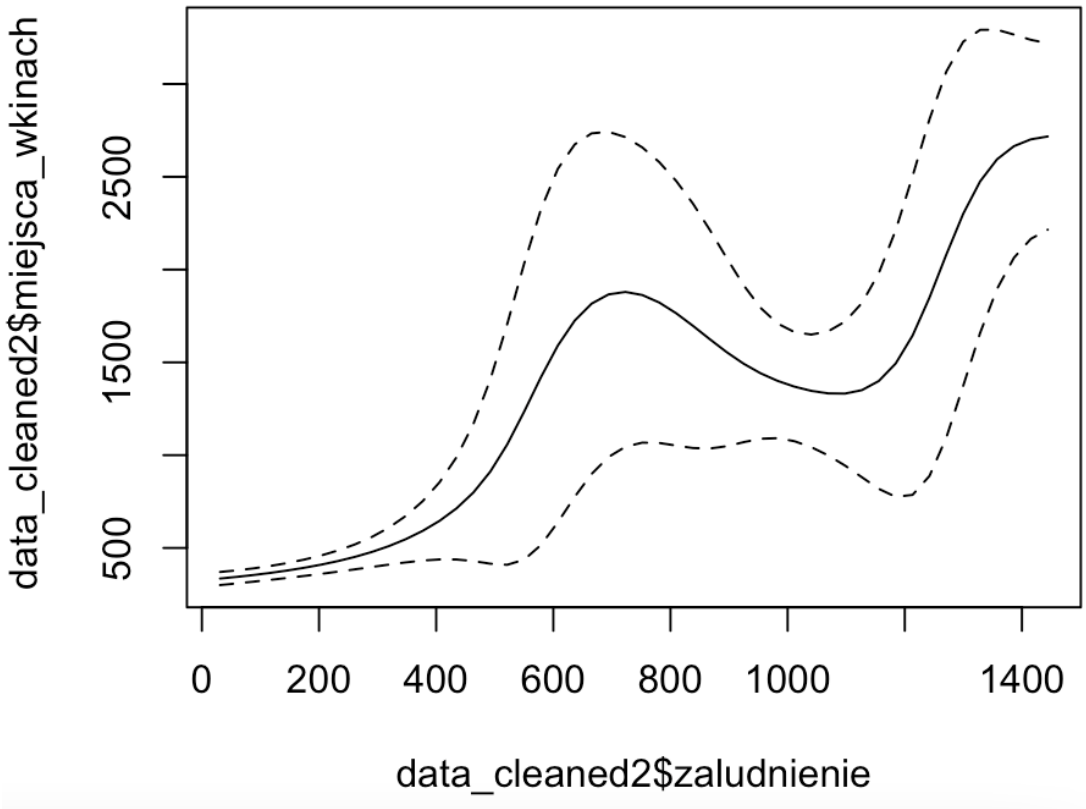

```
Regression Data: 209 training points, in 1 variable(s)
              data_cleaned2$l_os_na_1_miejsce_kina
Bandwidth(s):                                19.07377

Kernel Regression Estimator: Local-Constant
Bandwidth Type: Fixed
Residual standard error: 236.3547
R-squared: 0.7665757

Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 1
```

bw: współczynnik regularyzacji

```
> bw0$bw
[1] 19.07377
```

Tutaj jak najbardziej widzimy znaczącą przewagę regresji nieparametrycznej nad parametryczną, przynajmniej w znaczeniu wartości współczynnika determinacji oraz gigantyczną przewagę mierzona wartością RMSE. Jest to o tyle oczywiste, że nie mamy tu do czynienia ze związaniem konkretna postacią analityczną modelu (w szczególności liniową, która do tych akurat danych była niedopasowana).

Teraz model, który u nas pełni role modelu sensownego:



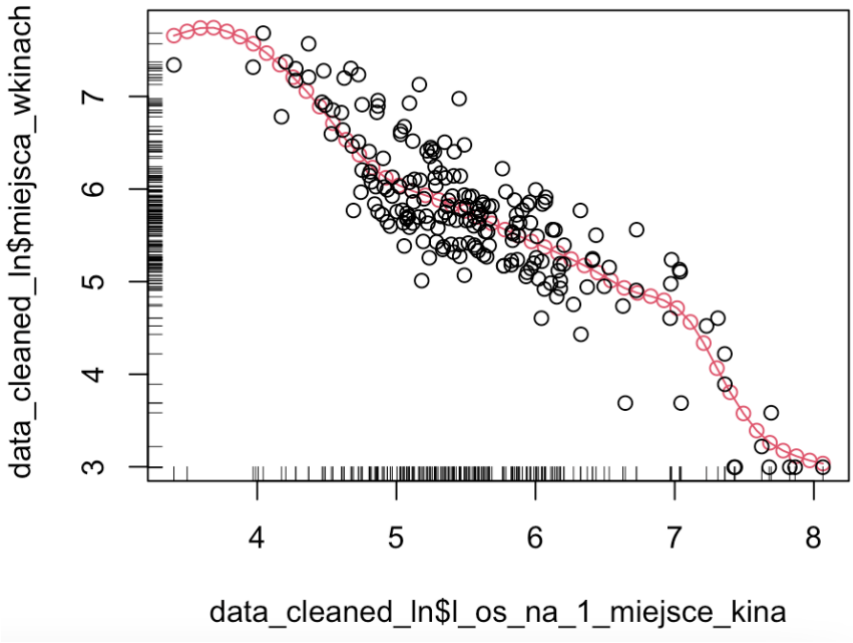
```
Regression Data: 209 training points, in 1 variable(s)
              data_cleaned2$zaludnienie
Bandwidth(s):                  118.1387

Kernel Regression Estimator: Local-Constant
Bandwidth Type: Fixed
Residual standard error: 314.6734
R-squared: 0.5744386

Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 1
```

```
> bw2$bw
[1] 118.1387
```

Tutaj także widzimy poprawę zarówno w przypadku RMSE jak i współczynnika determinacji, ale nie jest ona już tak wyraźna jak na przykładzie poprzedniego modelu (ze zdecydowanie nieliniową zależnością).

Na zakończenie powróćmy do wersji **modelu bezsensownego** z częściowo wystabilizowaną wariancją, czyli do modelu w zmiennych zlogarytmowanych. Tam jak pamiętamy miała miejsce poprawa wartości R^2 w stosunku do zmiennych surowych (odwrotnie niż w modelu sensownym).

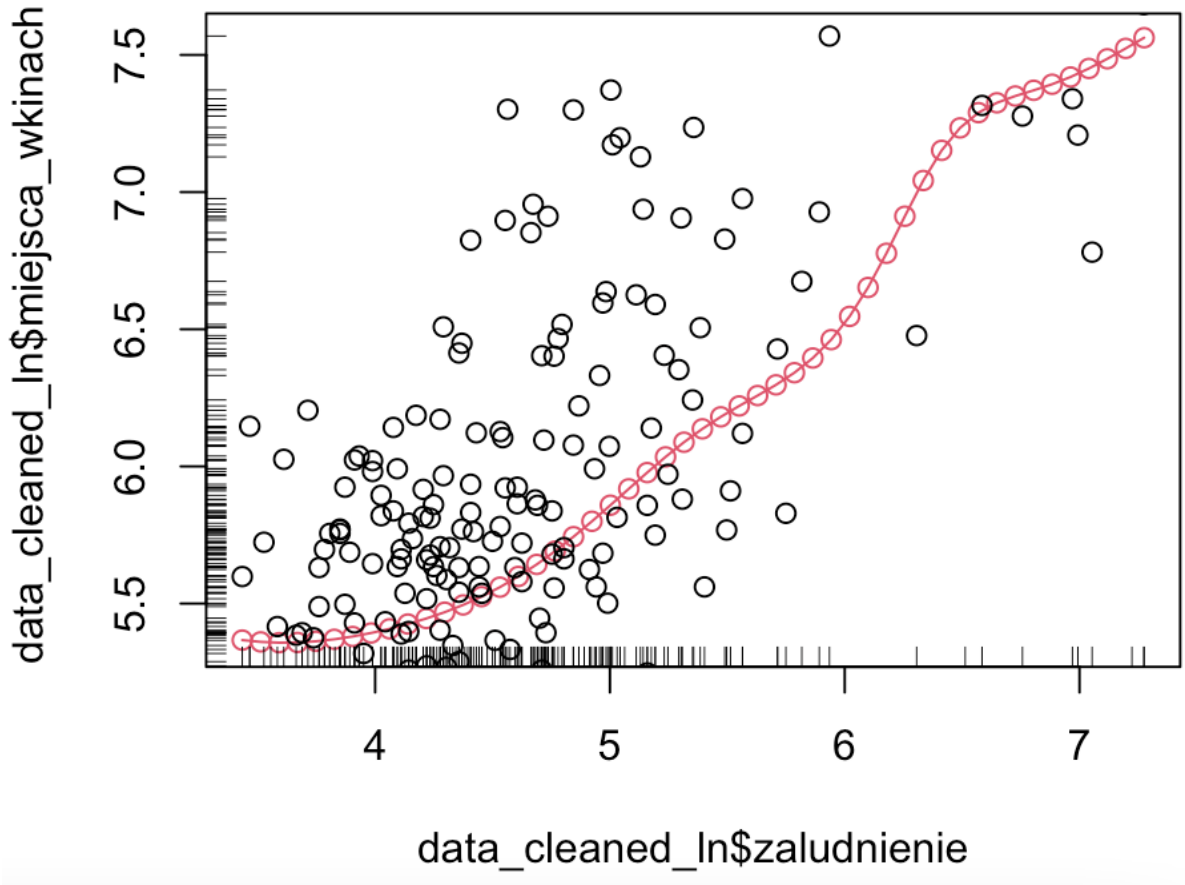
W tym wypadku regresja nieparametryczna sprawdza się znakomicie. Model jest najlepszy z dotychczas oszacowanych (przynajmniej w sensie wartości R^2 oraz RMSE).




```
Regression Data: 209 training points, in 1 variable(s)
              data_cleaned_ln$l_os_na_1_miejsce_kina
Bandwidth(s):                                0.183385

Kernel Regression Estimator: Local-Constant
Bandwidth Type: Fixed
Residual standard error: 0.4025757
R-squared: 0.8064013

Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 1
```

Dla modelu sensownego także obserwujemy poprawę, ale nie aż tak znaczną (w tym przypadku postać analitycznie była relatywnie dobrze dopasowana do danych).

Porównania dokonujemy zawsze dla tej samej klasy danych (zlogarytmowanych albo nie), czyli porównujemy dla tych samych danych regresję parametryczną z nieparametryczną.




```
Regression Data: 209 training points, in 1 variable(s)
              data_cleaned_ln$zaludnienie
Bandwidth(s):                   0.3107501

Kernel Regression Estimator: Local-Constant
Bandwidth Type: Fixed
Residual standard error: 0.7616542
R-squared: 0.3014517

Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 1
```

To w zasadzie kończy zasadniczą treść projektu.

Do dodatkowych analiz, w ramach dalszych potyczek z R, wybieramy niezlogarytmowaną zmienną: gęstość zaludnienia. Oprzemy się na kilku artykułach, w szczególności na jednej z publikacji pana Yen-Chin Chen[1] z Waszyngtonu.

Nonparametric Regression and Cross-Validation (washington.edu)
http://faculty.washington.edu/yenchic/17Sp_302/R12.pdf

Inna metoda oszacowania regresji nieparametrycznej: smooth spline

Co do samej idei metody, zasadniczo wszystko mówi poniższy cytat.

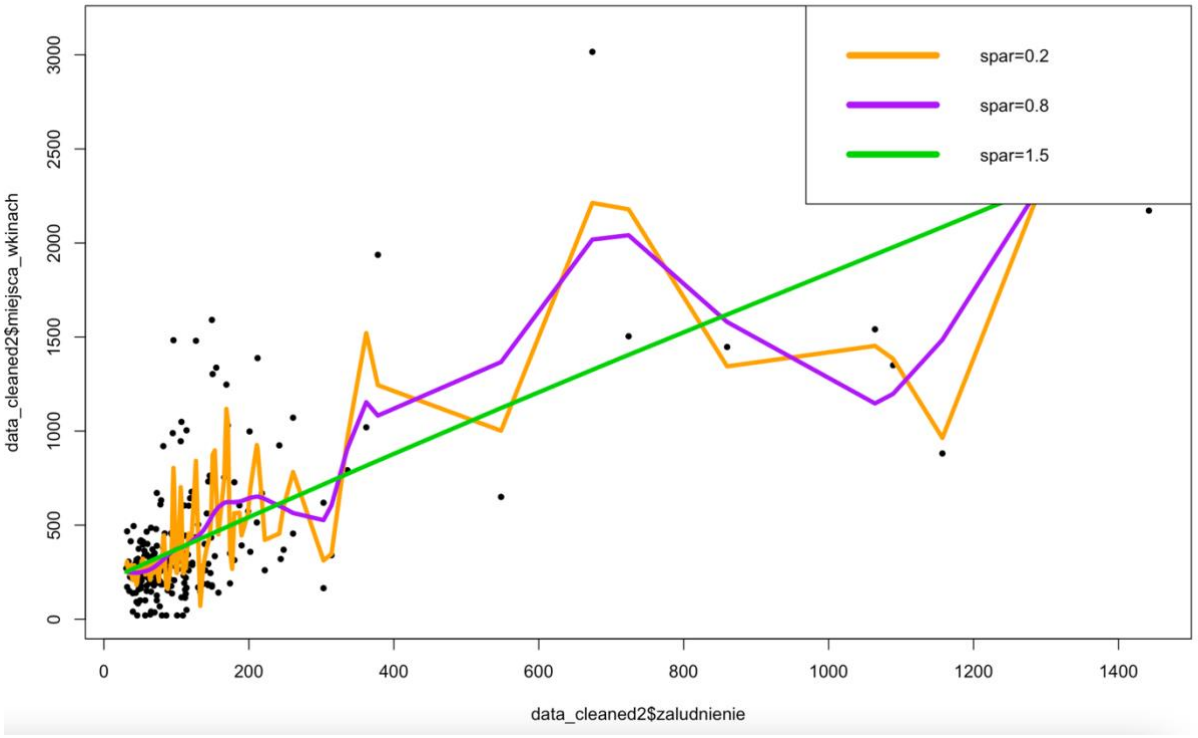
Here we introduce another nonparametric method: the *smoothing spline* approach. The idea of smoothing spline is: we want to find a function f(x) that fits the data well but also is very smooth. Using a more mathematical definition, we want to find a function f(x) such that

$$\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \lambda \int_{X_{\min}}^{X_{\max}} |f''(x)|^2 dx$$

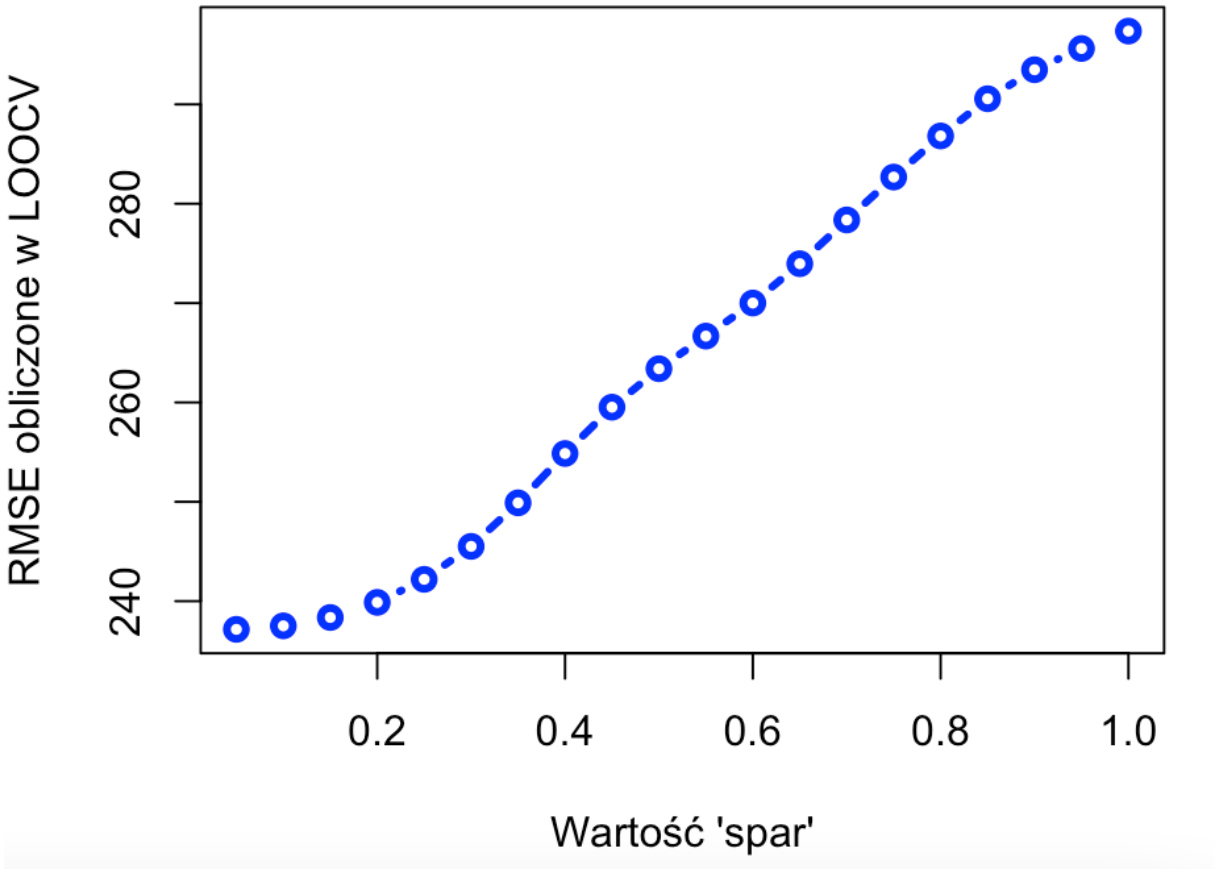
is minimized. The quantity λ is quantity that is called smoothness penalty (or smoothing parameter in some literature) that determines how much we will loss if we allow the function to adapt to the data. Without this part of penalty, the best function f will be the ones passing through every observation. Note that the X_{min} and X_{max} are the minimum and maximum value of all the observed covariates.

Poniżej przedstawimy przykłady dopasowania funkcji smoothing spline do danych w zależności od różnych wartości parametru wygładzania λ. W R używamy do tego funkcji smooth.spline() i argumentu spar jako λ.

[1] Yen-Chi Chen | University of Washington Department of Statistics (uw.edu)



W następnym kroku spróbujemy ustalić jaka wartość argumentu spar generuje najmniejszy błąd – użyjemy do tego mechanizmu walidacji krzyżowej typu LOO – najprawdopodobniej analogicznie jak dzieje się to w przypadku mechanizmu obliczającego bw dla funkcji npreg().



Walidacja krzyżowa dla (niezlogarytmowanej) zmiennej gęstość zaludnienia.

Procedura *k-fold cross validation* dla regresji nieparametrycznej została stworzona przez nas (bardziej lub mniej zgrabnie) bez wykorzystania bibliotek zewnętrznych!

Regresja nieparametryczna:

RMSE dla k=5:

```
> cv.error
[1] 389.5734
```

Sposób w jaki została zbudowana k-krotna walidacja krzyżowa prezentujemy na poniższym zrzucie fragmentu kodu:

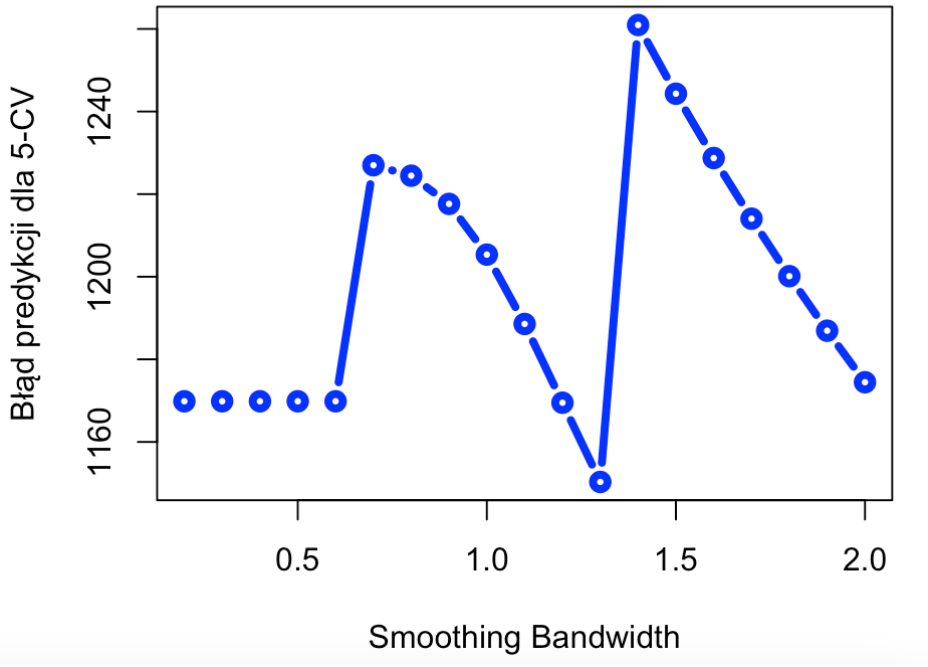
```r
RMSE <- function(f, o){
  sqrt(mean((f - o)^2))
}
k <- 5


set.seed(12345)
sim_data <- mutate(data_cleaned2[,3:7],
                   my.folds = sample(1:k,
                                     size = nrow(data_cleaned2),
                                     replace = TRUE))

cv.fun <- function(this.fold, data){


  train <- filter(data, my.folds != this.fold)
  validate <- filter(data, my.folds == this.fold)



  bww=npregbw(formula= miejsca_wkinach~zaludnienie, data = train, regtype="lc", nmulti=2)
  reg=npreg(bww, data=train)
  #reg=npreg(bww, newdata=validate)
  pred <-predict(reg, newdata=validate) %>% as.vector()


  this.rmse <- RMSE(f = pred, o = validate$miejsca_wkinach) # f=reg$mean

  return(this.rmse)
}



cv.error <- sapply(seq_len(k),
                   FUN = cv.fun,
                   data = sim_data) %>% mean()
cv.error
```

Dodatkowo wykonaliśmy walidację krzyżową dla metody regresji KSSMOOTH() przy użyciu różnych argumentów *bandwidth* (współczynnik wygładzania, współczynnik regularyzacji).

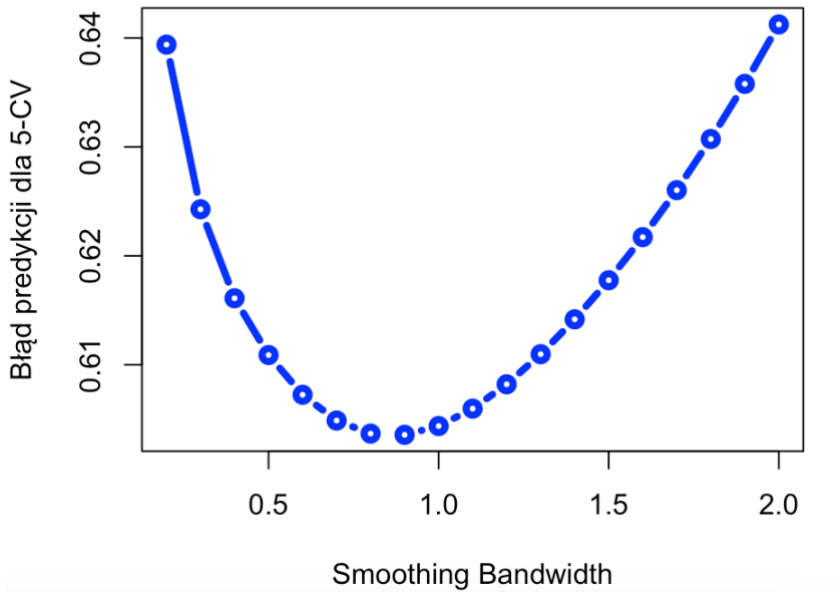
Poniżej przedstawiamy wykres na którym prezentujemy średnią błędu predykcji z pięciu warstw (folds) określoną wzorem

$$Err = \frac{1}{k}\sum_{\ell=1}^{k} Err_\ell$$

przy danych wartościach współczynnika regularyzacji.



Powiedzmy, że ten wykres wygląda nieco dziwnie, za to lepiej prezentuje się dla tych samych zmiennych po ich zlogarytmowaniu:



Właściwa walidacja krzyżowa dla (niezlogarytmowanej) zmiennej gęstość zaludnienia. *k-fold cross validation* dla regresji parametrycznej została przeprowadzona (dla porównania wyników) przez nas samodzielnie jak i z wykorzystaniem biblioteki caret.

5-fold cross validation dla regresji parametrycznej przy wykorzystaniu biblioteki caret:

```
Linear Regression

209 samples
  1 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 168, 167, 167, 168, 166
Resampling results:

  RMSE      Rsquared   MAE
  320.3197  0.4761287  210.1998
```

5-fold cross validation dla regresji parametrycznej przy wykorzystaniu samodzielnie napisanej przez nas funkcji (kod znajduje się powyżej przy regresji nieparametrycznej):

RMSE

```
> cv.error.param
[1] 320.9029
```

Wniosek: oszacowania RMSE są bardzo bliskie. Dobrze to wróży napisanej przez nas procedurze, różnice występujące na czwartym miejscu znaczącym mogą wynikać być może z jakichś subtelności dotyczących kolejności wykonywania działań czy też zaokrągleń.

Z ciekawości porównaliśmy również wyniki k krotnej walidacji krzyżowej dla regresji parametrycznej i nieparametrycznej z wykorzystaniem drugiej zmiennej - l_os_na_1_miejsce_kina.

5 fold cross validation - regresja parametryczna:

RMSE dla napisanej przez nas walidacji krzyżowej

```
> cv.error.param2
[1] 424.3434
```

RMSE dla walidacji z pakietu caret

```
RMSE      Rsquared   MAE
434.5834  0.2185862  285.7356
```

Zgodność dwóch powyższych wyników nie jest już doskonała.

5 fold cross validation - regresja nieparametryczna:
RMSE dla napisanej przez nas walidacji krzyżowej

```
> cv.error2
[1] 260.1078
```

Dla porównania przytaczamy wartość RMSE dla odpowiedniej regresji parametrycznej, co ponownie potwierdza lepszy potencjał dopasowywania do danych w przypadku regresji nieparametrycznej.

Podsumowanie i wnioski

1. Dla modeli liniowych regresja nieparametryczna faktycznie we wszystkich przypadkach dawała lepszą precyzję dopasowania niż odpowiednia regresja liniowa szacowana MNK, niemniej jednak w przypadku kiedy dane dobrze pasowały do modelu liniowego jej przewaga nie była znaczna, zaś w przypadku kiedy dane ze swej natury nie pasowały do modelu liniowego, jej przewaga była drastyczna – tego należało oczekiwać. Regresja nieparametryczna z pewnością znacznie lepiej sobie poradzi w przypadku danych z chwilowym zaburzeniem / szokiem, co może być jakąś alternatywą dla modelowania tego przy zastosowaniu zmiennej filtrującej przed/po.
2. W przeanalizowanych przypadkach wpływ wyboru funkcji jądra był nieznaczny, a w ocenie wizualnej najbardziej odróżniał się estymator prostokątny (czyli w sumie taki najbardziej prymitywny), zaś cztery pozostałe „na oko" byłoby trudno rozróżnić.
3. Nasze dane były obarczone wyraźną heteroskedastycznością. Nie zastosowaliśmy ważonej metody najmniejszych kwadratów (zamiast zwyklej MNK), ale jest to jakiś pomysł, bo próba wystabilizowania wariancji poprzez transformację logarytmiczną nie była udana, to znaczy wariancja się wprawdzie wystabilizowała, ale sama jakość dopasowania modelu uległa pogorszeniu. Stosując ważoną MNK minimalizowalibyśmy sumę kwadratów ważoną jakimiś wagami (takimi aby wystabilizować składnik losowy, czyli najprawdopodobniej nie byłoby w modelu stałej, niezależnej od poziomu zmiennej x), a zatem podobnie jak wartości funkcji jądrowej dla zoptymalizowanej wartości parametru h można traktować też jako wagi.
4. Nasze dane były danymi przekrojowymi, a nie szeregiem czasowym. Pominęliśmy kwestię autokorelacji reszt, choć nie w pełni jesteśmy przekonani, że mogliśmy to zrobić (ale prawie w pełni).
5. Intuicyjnie wydaje się, że dla danych „z trendem liniowym" przewaga regresji nieparametrycznej z estymatorem jądrowym może nie być wysoka (choć nigdy w typowym przypadku nie powinno być gorzej niż w przypadku zwykłej MNK), natomiast z kolei w przypadku szeregów czasowych (my tu mamy dane przekrojowe) oraz nakładania się okresowości oraz trendów nieliniowych, przewaga ta może być drastyczna (z kolei mogłoby być ciekawe porównanie tego z modelami klasy ARIMA).
6. Na podstawie naszych wyników intuicyjnie moglibyśmy sformułować tezę, że te procedury nieparametryczne z estymatorem jądrowym nadal nie są odporne na

heteroskedastyczność wariancji (generowaną dla prognoz przez zmienna niezależną). To w sumie też ciekawe zagadnienie jak jest faktycznie i czy zależy to od liniowości „procesu generującego dane" czy nie. Prowadzi nas to do dalszego wniosku, że być może niestabilność wariancji może popsuć znacznie więcej w modelu niż się to na pierwszy rzut oka wydaje.