# [DM] Case Study - Przemoc w USA

Patryk Nizio

30 sierpnia 2019

## Wstęp

### Problematyka

Przemoc i przestępczość to dobrze znane problemy społeczne, wpływają na poziom życia i bezpieczeństwo obywateli. W samych Stanach Zjednoczonych w latach 1999-2014 zginęło podnad 185 tysiecy osób z użyciem broni palnej. Stany Zjednoczone mają również największy wskaźnik uwięzionych na świecie. Blisko 2,3 miliona osób znajduję się w więzieniach co daje najwyższy wskaźnik na świecie, 743 na 100,000 obywateli znajduje sie w więzieniach. W Polsce dla porównania wskaźnik ten wynosi 218 na 100,000 obywateli (Dane z 2011, World Prison Population List).

Rozumiejąc czynniki wpływające na wzrost przestępczości jesteśmy w stanie zmniejszać ilość przestępst a także wprowadzać działania prewencyjne w statystycznie najbardziej narażonych regionach. Predykcja przestępczości pozwala również na optymalne rozmieszczenie funkcionariuszy oraz efektywniejsze zarządzanie budżetem.

### Opis zbioru danych

Opisywany zbiór danych łączy dane społeczno ekonomiczne ze spisu powszechnego z lat 90 oraz organów ściagania, ankiet i danych FBI z 1995 roku. Dane pochodzą z trzech źródeł:

- Creator: Michael Redmond (redmond@lasalle.edu); Computer Science; La Salle University; Philadelphia, PA, 19141, USA
- culled from 1990 US Census, 1995 US FBI Uniform Crime Report, 1990 US Law Enforcement Management and Administrative Statistics Survey, available from ICPSR at U of Michigan.
- Donor: Michael Redmond (redmond@lasalle.edu); Computer Science; La Salle University; Philadelphia, PA, 19141, USA

Zbiór danych zawiera 147 atrybutów, w tym:

- 125 predykcyjnych
- 4 nie predyckcyjnych (nazwa gminy, kod stanu, kod regionu i gminy)
- 18 potencjalnych celów

W potencjych atrybuty do modelowania znajdują sie:

- murders: liczba zabójstw w 1995 r.
- murdPerPop: liczba morderstw na 100 000 populacji

- rapes: liczba gwałtów w 1995 r.
- rapesPerPop: liczba gwałtów na 100 000 populacji
- robberies: liczba napadów w 1995 r.
- robbbPerPop: liczba rozbojów na 100 000 populacji
- assaults: liczba napadów w 1995 r.
- assaultPerPop: liczba ataków na 100 000 populacji
- burgl: liczba włamań w 1995 r.
- burglPerPop: liczba włamań na 100 000 populacji
- larcenies: liczba kradzieży w 1995 r.
- larcPerPop: liczba kradzieży na 100 000 populacji
- autoTheft: liczba kradzieży samochodowych w 1995 r.
- autoTheftPerPop: liczba kradzieży samochodowych na 100 000 populacji
- arsons: liczba podpaleń w 1995 r.
- arsonsPerPop: liczba podpaleń na 100 000 populacji
- ViolentCrimesPerPop: łączna liczba brutalnych przestępstw na 100 000 mieszkańców
- nonViolPerPop: całkowita liczba przestępstw bez użycia przemocy na 100 000 mieszkańców

W poniższej pracy skupiono się na liczbach przestępst bez użycia przemocy na 100 000 mieszkańców.

## Czyszczenie i analiza danych

Dane zawierały duże braki. Ze zbioru zostały usunięte te wiersze które zawierały braki w zmiennej modelowanej. Ze względu na wiele brakujących wartości przy części atrybutów nie zdecydowano się na oczyszczenie całego zbioru danych, tylko 111 rekordów zawierało pełny zbior danych. Pozostałe wartosci ze zwględu na dużą ilość braków w wielu atrybutach były pomijane podczas analizy.

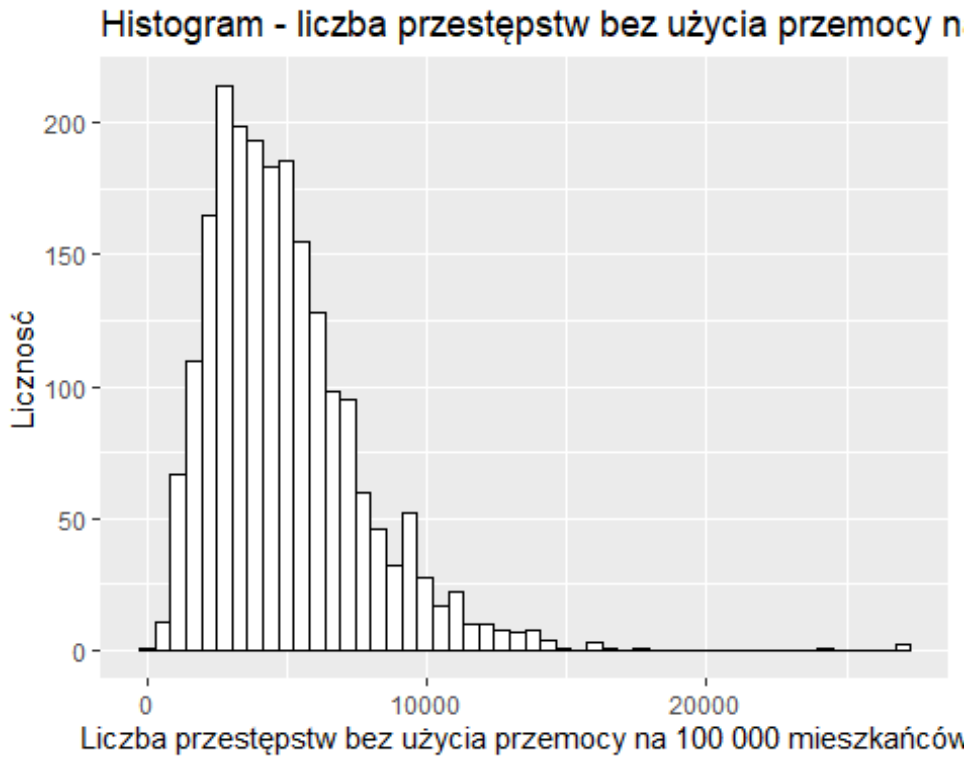
Rozbieżność danych jest stosunkowo duża, najmniesza wartość to 116.8 a największa 27119.8 przestępstw na 100 000 mieszkańców. Średnia wynosi 4908.2 przestępstw bez uzycia przemocy na 100 000 mieszkańców.

```
nonViolOriginal <- dataset$nonViolPerPop
summary(dataset$nonViolPerPop)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   116.8  2918.1  4425.4  4908.2  6229.3 27119.8
```

### Wykresy

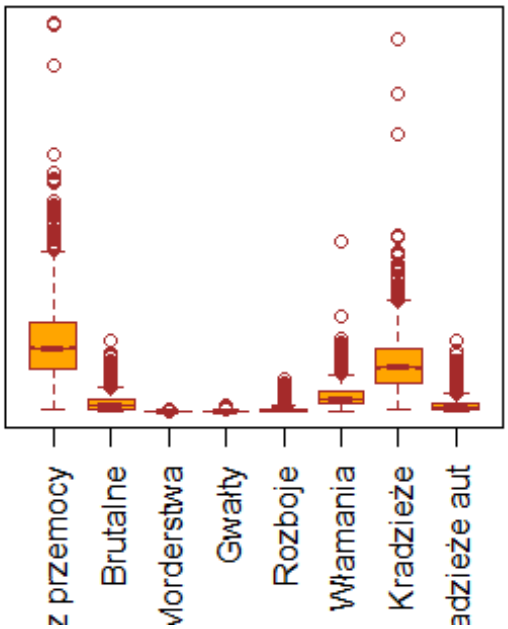
```
ggplot(dataset, aes(x=nonViolOriginal)) + geom_histogram(color="black",
fill="white", bins=50) + ggtitle("Histogram - liczba przestępstw bez użycia
przemocy na 100 000 mieszkańców") + ylab("Licznosć") + xlab("Liczba
przestępstw bez użycia przemocy na 100 000 mieszkańców")
```



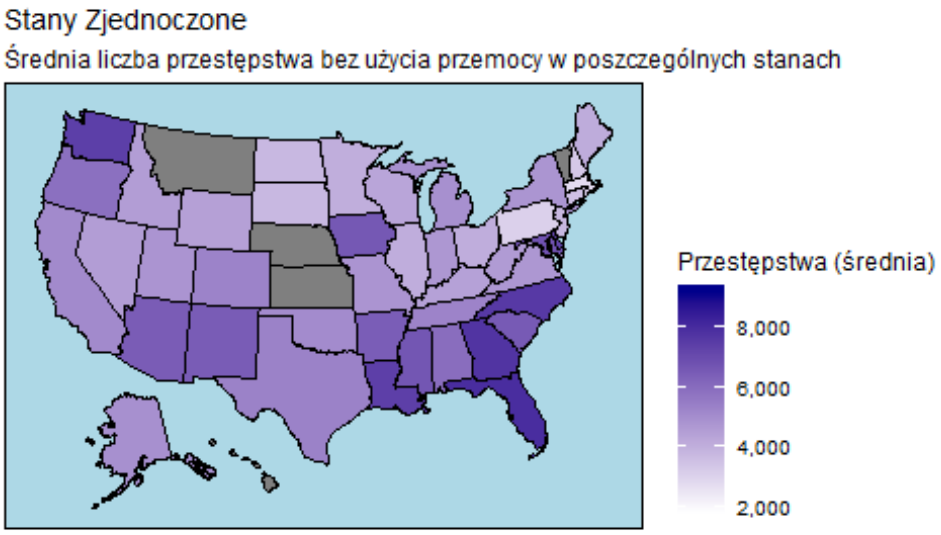

Wykres pudełkowy liczby przestępsw bez uzycia przemocy na 100,000 mieszkańców na tle innych przestępstw.

```
axName_BP_0 <- c("Bez przemocy", "Brutalne", "Morderstwa", "Gwałty",
"Rozboje", "Włamania", "Kradzieże", "Kradzieże aut")

par(mar=c(5,10,4,2)+.1)
boxplot(dataset$nonViolPerPop, dataset$ViolentCrimesPerPop,
dataset$murdPerPop, dataset$rapesPerPop, dataset$robbbPerPop,
dataset$burglPerPop, dataset$larcPerPop, dataset$autoTheftPerPop, names =
axName_BP_0, border = "brown",horizontal = FALSE, notch = TRUE, las = 2, col
= "orange", yaxt="n", at=1:8, main="Przestępstwa na 100 000 mieszkańców")
```

## Przestępstwa na 100 000 mieszkańcó




```
usaState <- aggregate(dataset$nonViolPerPop, by=list(state=dataset$state),
FUN=mean)


plot_usmap(regions = "states", data = usaState, values = "x", include =
usaState$category) + labs(title = "Stany Zjednoczone", subtitle = "Średnia
liczba przestępstwa bez użycia przemocy w poszczególnych stanach") +
theme(panel.background = element_rect(colour = "black", fill = "lightblue"))
+   scale_fill_continuous(low = "white", high ="darkblue",
                          name = "Przestępstwa (średnia)",label =
scales::comma,
                          limits = c(min(usaState$x)-1000,max(usaState$x))) +
theme(legend.position = "right")
```




Na mapie mozemy zauważyć że wyższy wskaźnik średniej przestępczości występuje w stanach południowych (Arizona, Nowy Meksyk, Lousiana, Floryda,Georgia, Karolina pólnocna i południowa) oraz w stanach Waszyngton, Maryland.

Przestępstwa bez użycia przemocy są to wszystkie przestępstwa w których ofiary nie stały się przedmiotem przemocy, wliczamy w to przestępstwa narkotykowe, przestępstwa majątkowe, kradzieże, oszustwa itp.

Najbezpieczniejsze hrabstwa znajdowały sie w stanach Maryland i Pensylwani.

```
df2 <- dataset %>% select(6:130,147)
dataCorrelation <- cor(na.omit(df2), method = "pearson")

# corrplot(dataCorrelation, method = "number", order = "hclust", type =
"upper")
```

Największy wpływ na przestępstwa bez przemocy mają następujące atrybuty (według korelacji):

- PctPopUnderPov: odsetek osób poniżej poziomu ubóstwa
- racepctblack: procent populacji Afro-amerykanów
- MalePctDivorce: odsetek mężczyzn rozwiedzionych
- FemalePctDiv: odsetek kobiet rozwiedzionych
- TotalPctDiv: procent rozwiedzionej populacji
- PctKidsBornNeverMar: odsetek dzieci urodzonych bez związku małżeńskiego
- PctHousNoPhone: procent zajętych mieszkań bez telefonu

- PctUnemployed: odsetek osób w wieku 16 lat i starszych, na rynku pracy i bezrobotnych
- pctWPubAsst: odsetek gospodarstw domowych z dochodami z pomocy publicznej w 1989 r.
- PctVacantBoarded: procent wolnych mieszkań, które są zabite deskami
- PctWOFullPlumb: procent mieszkań bez kompletnych instalacji hydraulicznych
- medIncome: mediana dochodu gospodarstwa domowego
- pctWInvInc: odsetek gospodarstw domowych z dochodami z inwestycji / czynszu w 1989 r.
- PctEmploy: odsetek osób w wieku 16 lat i starszych, które są zatrudnione
- pctWWage: odsetek gospodarstw domowych o dochodach z wynagrodzenia w 1989 r.

## Model I - PCA

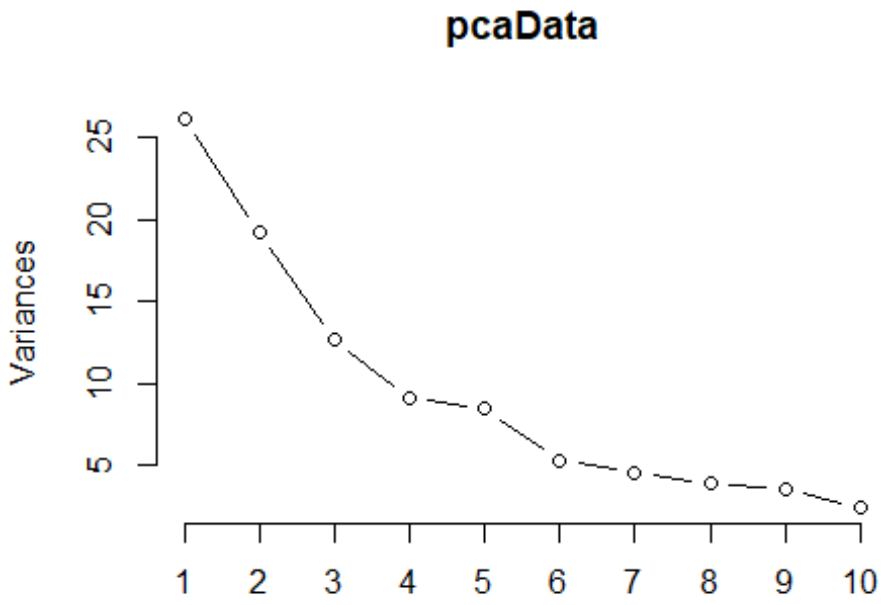
Analiza głównych składowych (ang. principal component analysis, PCA) – jedna ze statystycznych metod analizy czynnikowej. Zbiór danych składający się z N obserwacji, z których każda obejmuje K zmiennych, można interpretować jako chmurę N punktów w przestrzeni K-wymiarowej. Celem PCA jest taki obrót układu współrzędnych, aby maksymalizować w pierwszej kolejności wariancję pierwszej współrzędnej, następnie wariancję drugiej współrzędnej itd.

Tak przekształcone wartości współrzędnych nazywane są ładunkami wygenerowanych czynników (składowych głównych). W ten sposób konstruowana jest nowa przestrzeń obserwacji, w której najwięcej zmienności wyjaśniają początkowe czynniki.

```
pcaData <-  prcomp(na.omit(dataset[,c(6:129,147)]), scale = TRUE)
```

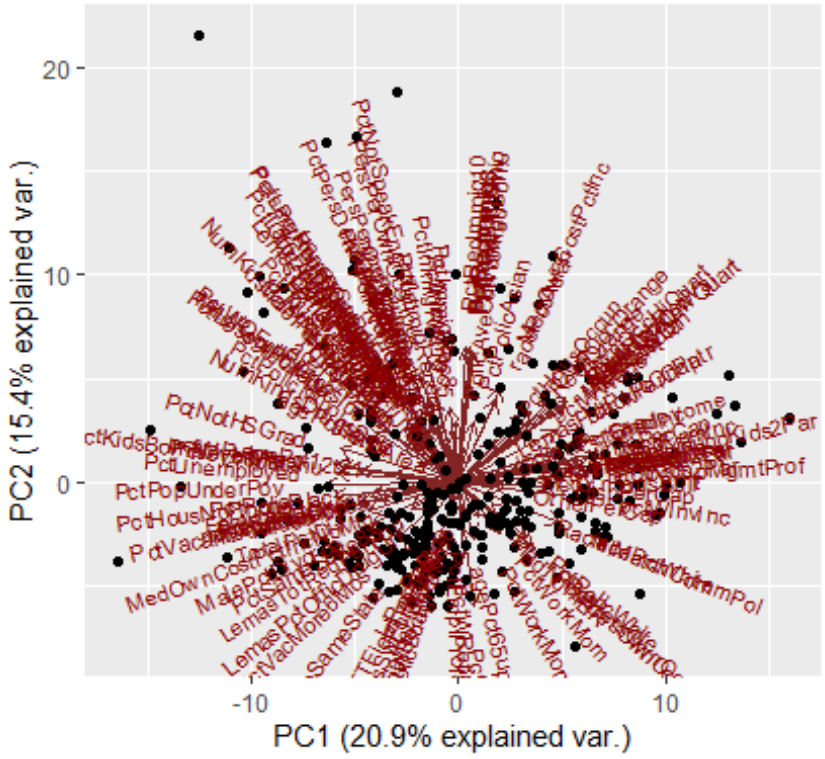
Rozkład PCA wskazuje że czynnik pierwszy wyjaśnia zmienne w 20.9%, poniżej przedstawiono wykres osuwiska wartości własnych.

```
plot(pcaData, type='l')
```




Osie są widziane jako strzałki pochodzące od punktu środkowego oraz ich udział w zmiennych PC1 i PC2. Poniżej dodoatkowo wykres dla zmiennych PC3 i PC4. Wykresy wskazuja na to że wiele zmiennych jest z sobą powiązanych.

```
ggbiplot(pcaData, alpha = 1,var.axes=TRUE, ellipse=TRUE, obs.scale = 1,
var.scale = 1)
```




```
ggbiplot(pcaData, var.axes=TRUE,choices = c(3:4), alpha = 1, ellipse=TRUE,
obs.scale = 0, var.scale = 1)
```


Macierz przekształcenia:

```
print(pcaData$rotation[1:10,1:4])
```

```
##                       PC1         PC2          PC3         PC4
## population    -0.04986174  0.08672362 -0.225281221  0.06051299
## householdsize -0.04673420  0.12308460  0.137375242  0.10023390
## racepctblack  -0.12912728 -0.03035709 -0.036421244 -0.06817303
## racePctWhite   0.13296805 -0.06269440  0.000536344  0.06211472
## racePctAsian   0.04985483  0.12698522  0.018978253 -0.03249531
## racePctHisp   -0.05012290  0.16168381  0.092483801  0.05881152
## agePct12t21   -0.08349053  0.01629578  0.073922562 -0.08855900
## agePct12t29   -0.05791915  0.05826659  0.054102535 -0.17427681
## agePct16t24   -0.05045615  0.02843419  0.035339658 -0.17483286
## agePct65up     0.01906990 -0.08513917 -0.071007698  0.08735908
```

## Model II - Regresja wieloraka

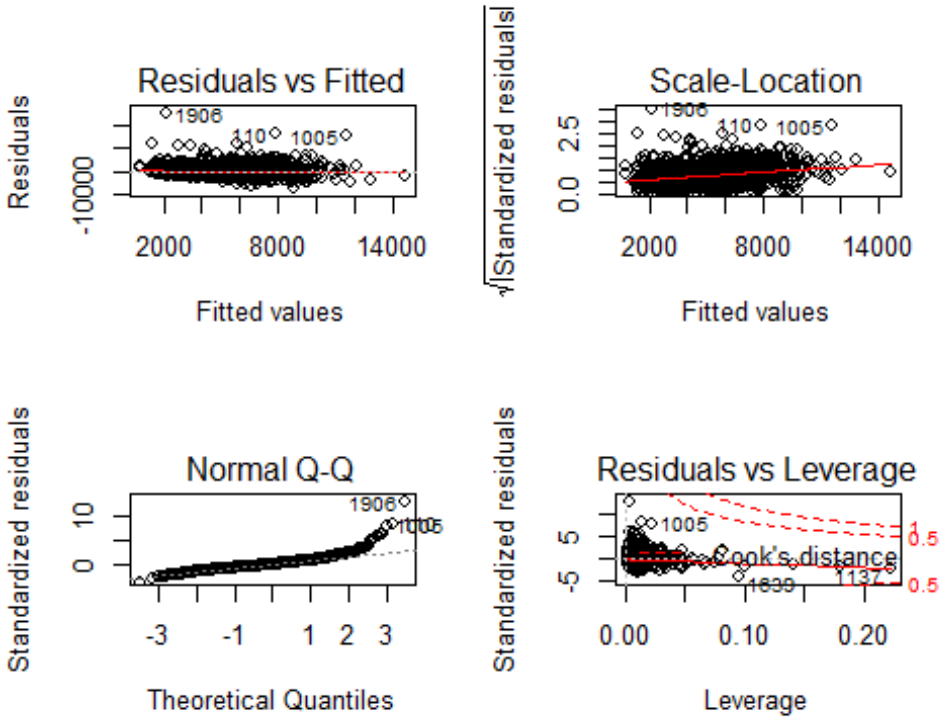
Przy tworzeniu modelu Regresji wielorakiej skupiono sie na 15 najbardziej wpływowowych czynnikach (w tym wytyczonych poprzez analize PCA).

```
mlr <- lm(nonViolPerPop ~ PctPopUnderPov +
racepctblack +
MalePctDivorce +
FemalePctDiv +
TotalPctDiv +
PctKidsBornNeverMar +
PctHousNoPhone +
PctUnemployed +
pctWPubAsst +
PctVacantBoarded +
PctWOFullPlumb +
medIncome +
pctWInvInc +
PctEmploy +
pctWWage, data = dataset)

summary(mlr)
```

```
##
## Call:
## lm(formula = nonViolPerPop ~ PctPopUnderPov + racepctblack +
##     MalePctDivorce + FemalePctDiv + TotalPctDiv + PctKidsBornNeverMar +
##     PctHousNoPhone + PctUnemployed + pctWPubAsst + PctVacantBoarded +
##     PctWOFullPlumb + medIncome + pctWInvInc + PctEmploy + pctWWage,
##     data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7073.8 -1139.1  -235.7   866.9 25008.4
##
## Coefficients:
```

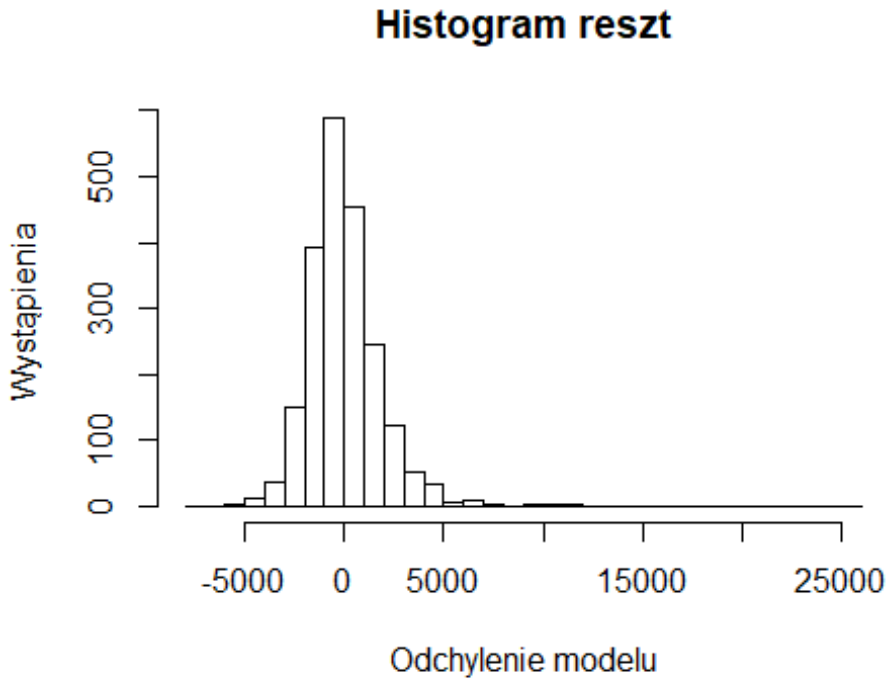
```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.077e+03  1.007e+03   1.070  0.28483
## PctPopUnderPov        1.212e+02  1.370e+01   8.848  < 2e-16 ***
## racepctblack          1.406e+01  5.770e+00   2.437  0.01489 *
## MalePctDivorce        7.126e+02  2.559e+02   2.785  0.00540 **
## FemalePctDiv          7.782e+02  2.725e+02   2.856  0.00433 **
## TotalPctDiv          -1.114e+03  5.222e+02  -2.132  0.03310 *
## PctKidsBornNeverMar   1.918e+02  3.125e+01   6.135 1.01e-09 ***
## PctHousNoPhone       -8.920e+00  2.165e+01  -0.412  0.68044
## PctUnemployed        -3.773e+01  3.351e+01  -1.126  0.26033
## pctWPubAsst          -7.262e+01  2.287e+01  -3.175  0.00152 **
## PctVacantBoarded     -1.886e+01  1.577e+01  -1.196  0.23191
## PctWOFullPlumb        9.730e+01  1.258e+02   0.773  0.43938
## medIncome             9.537e-03  7.009e-03   1.361  0.17374
## pctWInvInc            9.846e+00  8.338e+00   1.181  0.23780
## PctEmploy             4.147e+01  1.545e+01   2.684  0.00733 **
## pctWWage             -6.257e+01  1.491e+01  -4.196 2.83e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1967 on 2102 degrees of freedom
## Multiple R-squared:  0.4881, Adjusted R-squared:  0.4845
## F-statistic: 133.6 on 15 and 2102 DF,  p-value: < 2.2e-16
```

```r
layout(matrix(c(1,2,3,4),2,2))
plot(mlr)
```



Korelacja między modelami wynosi 0.698 co stanowi zadowalajacy wynik, w danych występuje jednak rozbieżność w wartosciach co potęguje wzrost błędów dla części wartości. Są to najprawdopodobniej przypadki które silnie odbiegają od modelu np. przypadki skrajne lub anomalia.

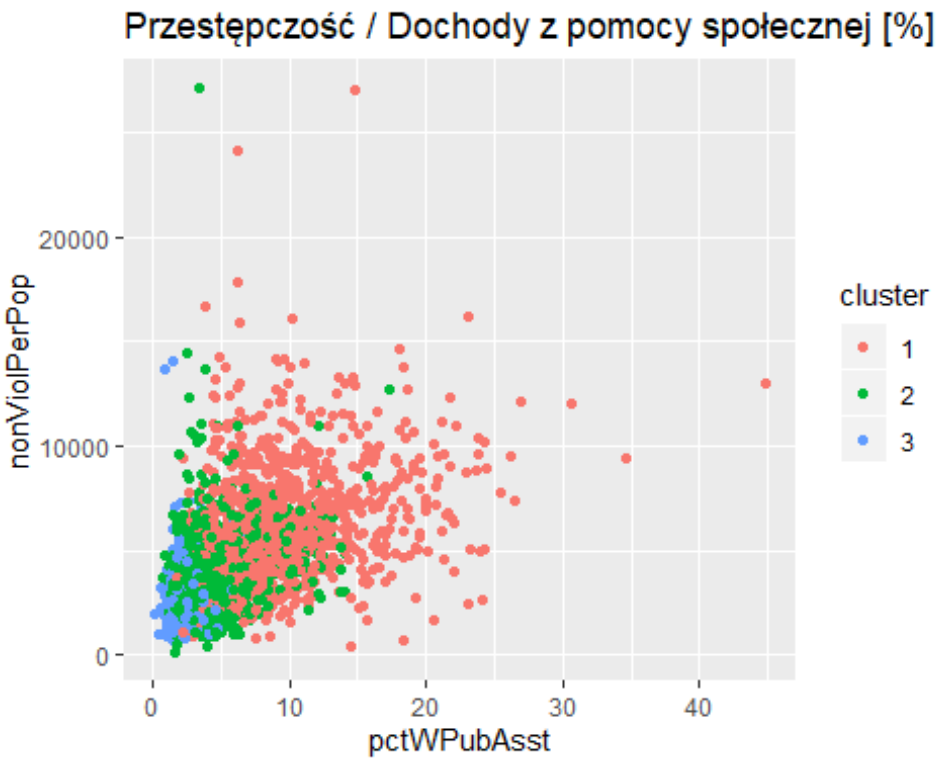
```
mlrCoe <- coefficients(mlr)
modelMlr <- predict.lm(mlr, dataset)

print(cor(dataset$nonViolPerPop, modelMlr))

## [1] 0.6986557

hist(mlr$residuals, breaks = 24, main = "Histogram reszt", xlab = "Odchylenie
modelu", ylab = "Wystąpienia")
```

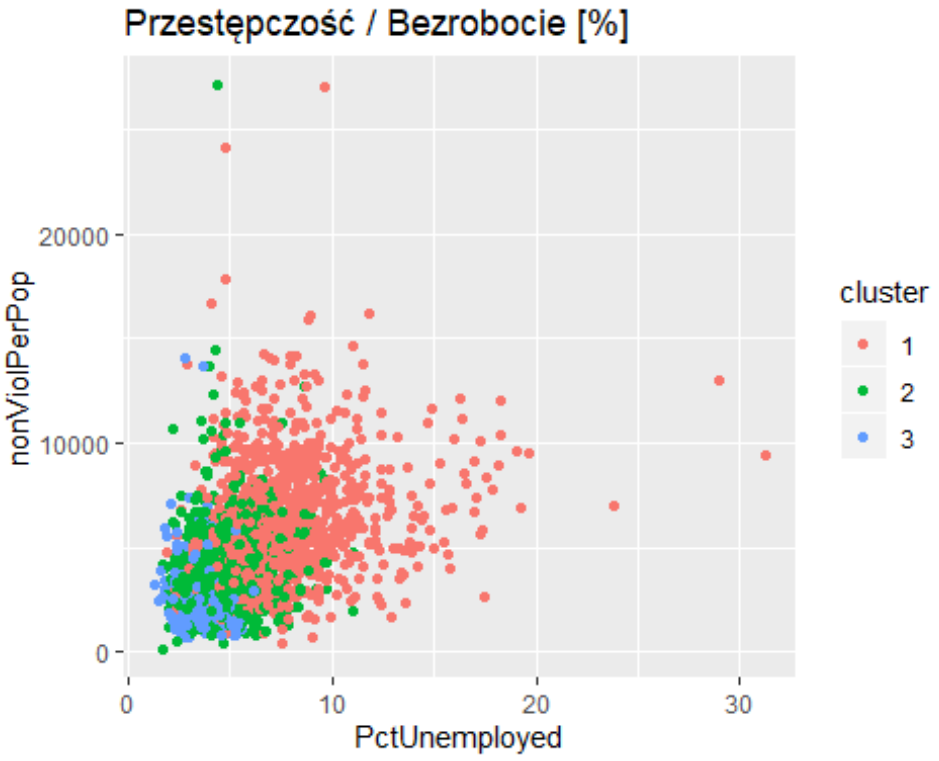


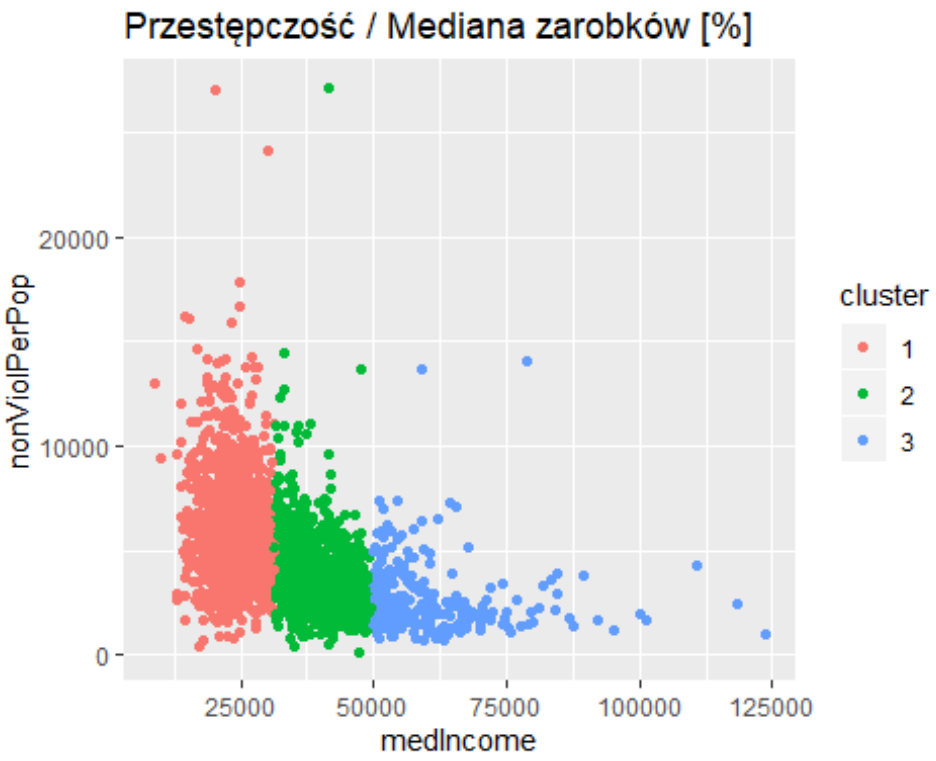

```
mape(dataset$nonViolPerPop,modelMlr)

## [1] 29.86194
```

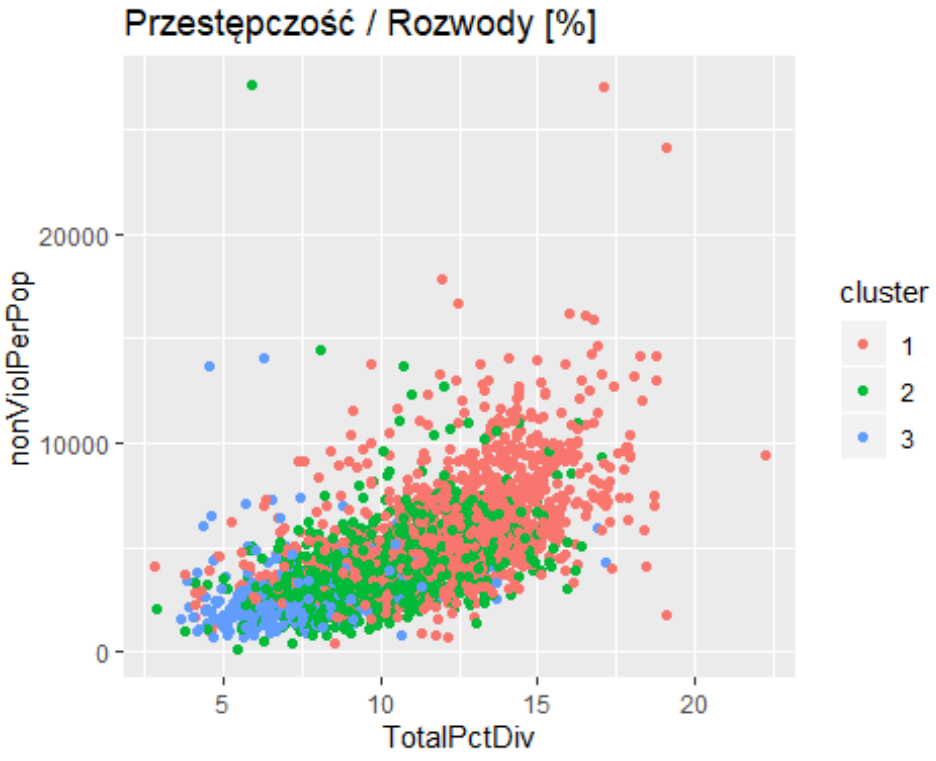
## Model III - Klasteryzacja

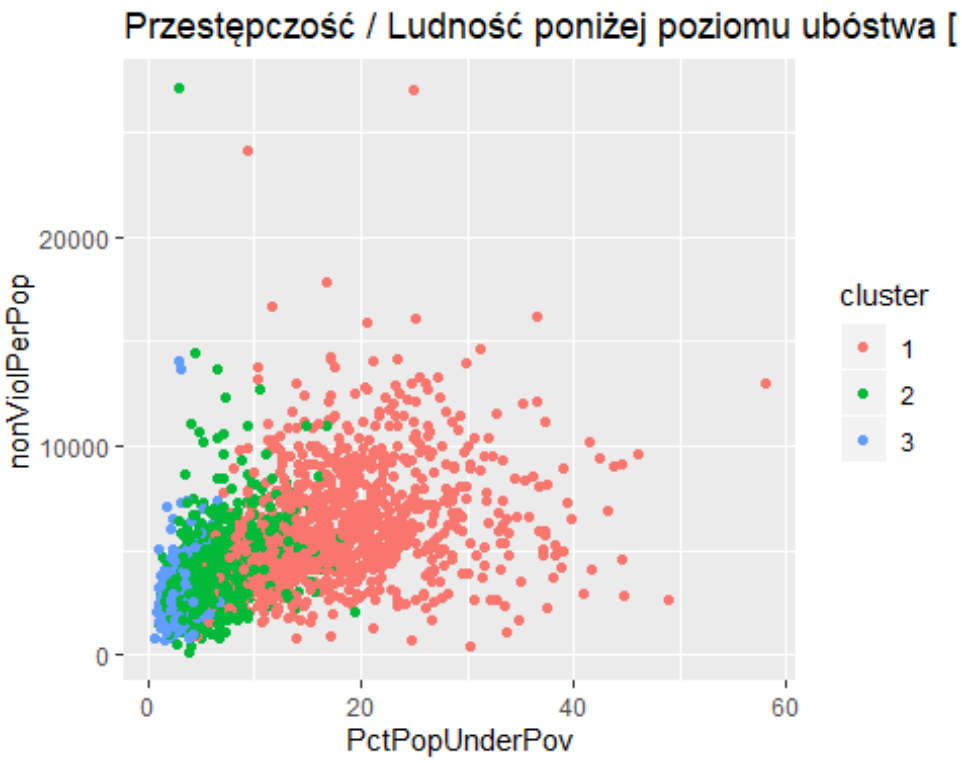
Jako metodę klasteryzacji wybrano metodę K-średnich Poniżej wykresy dla klasteryzacji względem wybranych atrybutów: zarobki, rozwody, procent dochodów z pomocy społecznej, udziału ludności afroamerykńskiej i populacji żyjacej poniżej poziomu ubóstwa.

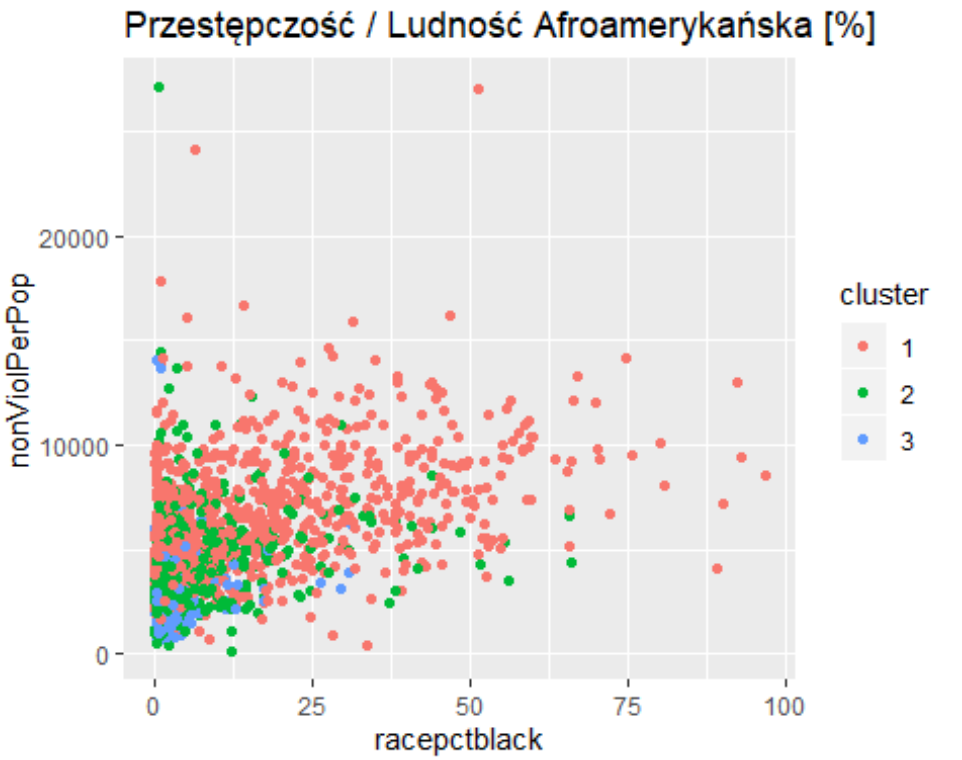

Przestępczość / Dochody z pomocy społecznej [%]
nonViolPerPop
20000
10000
0
0
10
20
30
40
pctWPubAsst
cluster
1
2
3


Przestępczość / Bezrobocie [%]
nonViolPerPop
20000
10000
0
0
10
20
30
PctUnemployed
cluster
1
2
3


Przestępczość / Mediana zarobków [%]
nonViolPerPop
20000
10000
0
25000
50000
75000
100000
125000
medIncome
cluster
1
2
3


Przestępczość / Rozwody [%]
nonViolPerPop
20000
10000
0
5
10
15
20
TotalPctDiv
cluster
1
2
3




Na podstawie tych wykresów możemy wnioskować że nawjęcej przestępst dochodzi w regionach biedniejszych, wiąże się to z dużym bezrobociem i problemami społecznymi. Ludzie w tych dzielnicach często kradną lub handluja narkotykami przez co wskaźnik przestępst (bez uzycia przemocy) jest tak wysoki.

## Wnioski i uwagi

Dane zawierały dużą liczbę atrybutów, do wyznaczenia głównych składowych użyto PCA. Model regresji wielorakiej dobrze modelował zmienną jednak zawierał również wysoki błąd dla części przypadków. Dzieki metodom klasteryzacji udało podzielić się grupę względem wybranych składowych. Przedstawione modele umożliwiają rozpoznanie miejsc bardziej narażonych na przestępstwa (w analizie skupiono się na przestępstach bez użycia przemocy jednak pozostałe przestępstwa mają podobne ugruntowanie).

Przypuszczalnie zastosowanie sieci neuronowej było by najlpszą alternatywą do stworzenia wartościowego modelu.

Wyciągajac wnioski możemy uznać że w biedniejszych regionach przestępstwa są częstsze ze względu na ogólne ubóstwo i problemy społeczno-ekonomiczne. Ludność zamożniejsza dużo rzadziej popełnia przestępstwa natomiast regiony w ktorych panuje ubóstwo i bezrobocie są bardziej narażone na przestępstwa takie jak kradzieże, podpalenia lub przestępstwa majątkowe i narkotykowe.

## Źródła


- https://upload.wikimedia.org/wikipedia/commons/d/d3/Felony_Sentences_in_State_Courts.pdf
- World Prison Population List: https://www.prisonstudies.org/sites/default/files/resources/downloads/wppl_9.pdf