

# Untitled

Jakub Slazyk

25 11 2021

Zaladowanie wykorzystanych bibliotek

```
library("sp")  
library("rgdal")  
library("ggplot2")  
library("dbscan")
```

```
## Warning: pakiet 'dbscan' został zbudowany w wersji R 4.1.2
```

```
library("repr")
```

```
## Warning: pakiet 'repr' został zbudowany w wersji R 4.1.2
```

```
library("broom")  
library("opticskxi")
```

```
## Warning: pakiet 'opticskxi' został zbudowany w wersji R 4.1.2
```

Wczytanie danych

```
getwd()
```

```
## [1] "C:/Users/jakub/OneDrive/Dokumenty"
```

```
data_conv<-read.csv2("zestaw9.csv",sep=";")  
head(data_conv)
```

```
##      Long      Lat  
## 1 19.95673 50.02497  
## 2 19.92715 50.06300  
## 3 19.80796 50.01776  
## 4 19.93616 50.06218  
## 5 19.95938 50.04922  
## 6 19.93493 50.05555
```

Zmiana układu współrzędnych danych

```
cord.dec = SpatialPoints(cbind(data_conv$Long, data_conv$Lat), proj4string=CRS("+proj=longlat"))
cord.UTM <- spTransform(cord.dec, CRS("+init=epsg:2178"))
```

```
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO", prefer_proj
## = prefer_proj): Discarded datum European_Terrestrial_Reference_System_1989 in
## Proj4 definition
```

Podgląd danych

```
head(cord.UTM)
```

```
## SpatialPoints:
##      coords.x1 coords.x2
## [1,]   7425247   5543719
## [2,]   7423188   5547979
## [3,]   7414575   5543077
## [4,]   7423832   5547878
## [5,]   7425474   5546414
## [6,]   7423734   5547143
## Coordinate Reference System (CRS) arguments: +proj=tmerc +lat_0=0
## +lon_0=21 +k=0.999923 +x_0=7500000 +y_0=0 +ellps=GRS80 +units=m
## +no_defs
```

Wczytanie tła(mapy z podziałem na osiedla/dzielnice Krakowa)

```
x<-getwd()
my_spdf <- readOGR(dsn=path.expand('C:/Users/jakub/OneDrive/Dokumenty/osiedla'), layer='osiedla' )
```

```
## Warning in OGRSpatialRef(dsn, layer, morphFromESRI = morphFromESRI, dumpSRS = dumpSRS, : Discarded d
## but +towgs84= values preserved
```

```
## OGR data source with driver: ESRI Shapefile
## Source: "C:\Users\jakub\OneDrive\Dokumenty\osiedla", layer: "osiedla"
## with 141 features
## It has 30 fields
## Integer64 fields read as strings:  ID IL_K_MOBIL IL_K_NIEMO IL_K_POPRO IL_K_PROD IL_K_PRZED IL_L_MOB
```

Podgląd mapy z podziałem na osiedla/dzielnice Krakowa

```
par(mar=c(0,0,0,0))
plot(my_spdf, col="#f2f2f2", bg="skyblue", lwd=0.25, border=0 )
```



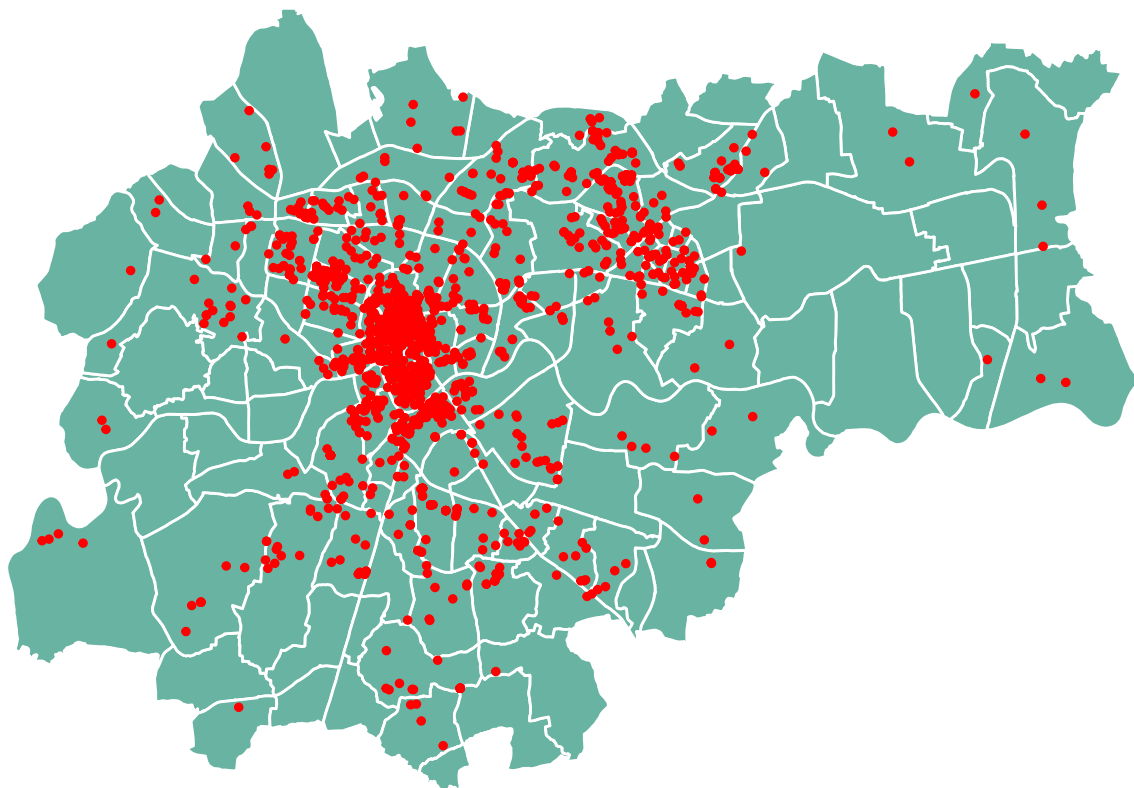
Przygotowanie obu warstw do wygodnego korzystania w trakcie tworzenia plotow

```
spdf_fortified <- tidy(my_spdf)
cords <- as.data.frame(cord.UTM)
colnames(cords) <- c("X", "Y")
head(cords)
```

```
##           X           Y
## 1 7425247 5543719
## 2 7423188 5547979
## 3 7414575 5543077
## 4 7423832 5547878
## 5 7425474 5546414
## 6 7423734 5547143
```

Podgląd na zgłoszone wykroczenia na tle mapy osiedli/dzielnic

```
ggplot() +
  geom_polygon(data = my_spdf, aes( x = long, y = lat, group = group), fill = "#69b3a2", color = "white") +
  geom_point(data = cords, aes(x = X, y = Y), color = "red",
    size = 1) +
  theme_void()
```



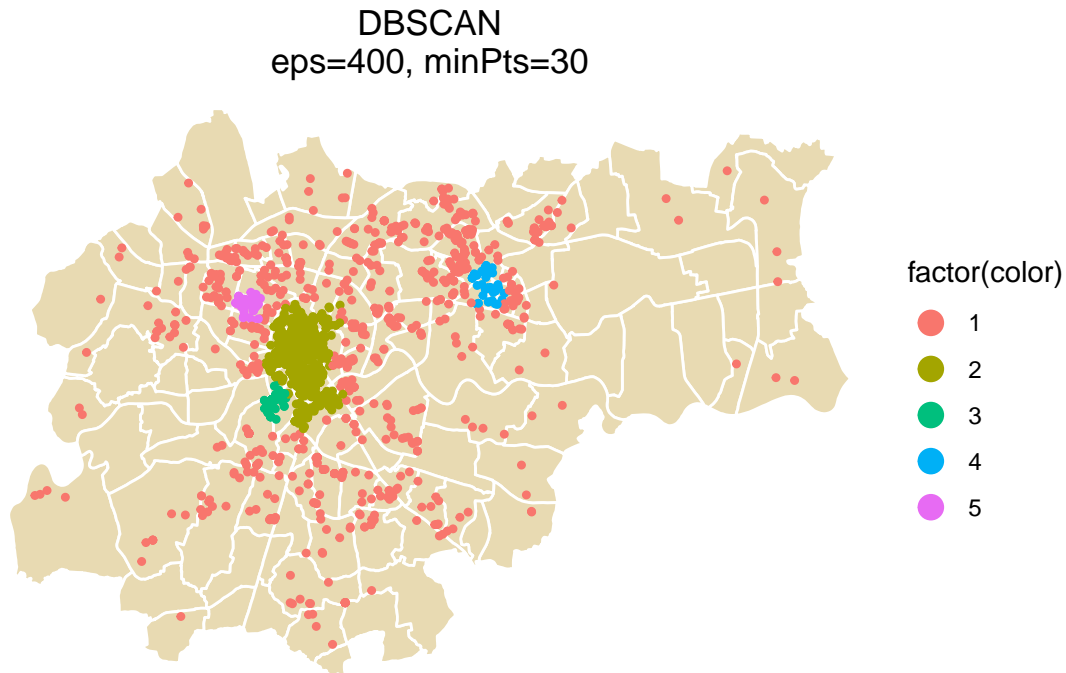
Na wstępie warto zauważyć, że we wszystkich wizualizacjach punkty należące do pierwszego klastra to szum i w rzeczywistości punkty te nie należą do żadnego z klastrów.

Algorytm DBSCAN: Parametry wejściowe:  $\epsilon$  (epsilon) - promień sąsiedztwa, maksymalna dopuszczalna odległość pomiędzy dwoma punktami, aby rozważać czy należą do tego samego klastra  $\minPts$  - minimalna wymagana liczba punktów znajdujących się w odpowiedniej odległości, aby uznać je za cluster Schemat działania: Dla wszystkich obserwacji znajdujemy sąsiadów, znajdujących się w odległości nie większej, niż epsilon. Jeśli ich liczba wynosi przynajmniej wartość  $\minPts$ , punkt zaliczany jest do punktów centralnych (Core Points). Wszystkie punkty, spełniające to założenie i leżące w swoim sąsiedztwie tworzą cluster, następnie wszystkie punkty, leżące w odległości nie większej niż epsilon od punktów, tworzących człon klastru, są dołączane jako punkty graniczne klastra (Border Points) Zalety: -jego działanie jest proste w zrozumieniu -jest odporny na wpływ obserwacji odstających, nienależących do żadnego klastra -daje dobre rezultaty, przy relatywnie szybkim czasie działania -możliwość modelowania wyniku za pomocą parametrów wejściowych -mamy spory wpływ na finalny rezultat Wady: -nie daje nam możliwości zdefiniowania oczekiwanej liczby klastrów - ich liczba wynika z wartości parametrów wejściowych -możliwość manewru argumentami wejściowymi może nieść za sobą problem, ich dobór może być trudny w sytuacji, gdy nie wiemy jakiego rezultatu się spodziewamy/oczekujemy

Klasteryzacja 1 z wykorzystaniem algorytmu DBSCAN

```
set.seed(123456789)
dbscan_res <- dbscan(cords, eps = 400, minPts = 30)
color<-dbscan_res$cluster+1
plot1<-ggplot() +
  geom_polygon(data = my_spdf, aes( x = long, y = lat, group = group), fill="#E8DAB2", color="white") +
  geom_point(data = cords, aes(x = X, y = Y,color=factor(color)), size = 0.9) +
  theme_void() +
  theme(plot.margin = unit(c(1,1,1,1), "cm")) +
```

```
ggtitle("DBSCAN\neps=400, minPts=30") +
  theme(plot.title = element_text(hjust = 0.5)) +
  guides(colour = guide_legend(override.aes = list(size=4)))
plot1
```



Patrząc na mapę można mieć wątpliwości czy w każdym klastrze liczba minimalna faktycznie jest spełniona (głównie chodzi o klastrowy numer 3), dlatego sprawdzam liczbę punktów w każdym klastrze aby sprawdzić czy warunek jest spełniony. Okazuje się faktycznie być spełniony, co oznacza, że punkty bardzo mocno muszą na siebie nachodzić. Zdecydowanie największy klastrowy (numer 2) znajduje się na całym obszarze Starego Miasta, należy do niego nieco ponad połowa obserwacji. Pozostałe 3 klastry (numery 3-5) są zdecydowanie mniejsze i zlokalizowane są w następujących miejscach: Krowodrza, Północno-Wschodnia część Dębników bezpośrednio sąsiadująca ze Starym Miastem, a także okolice Bieżanów.

```
as.data.frame(table(color))
```

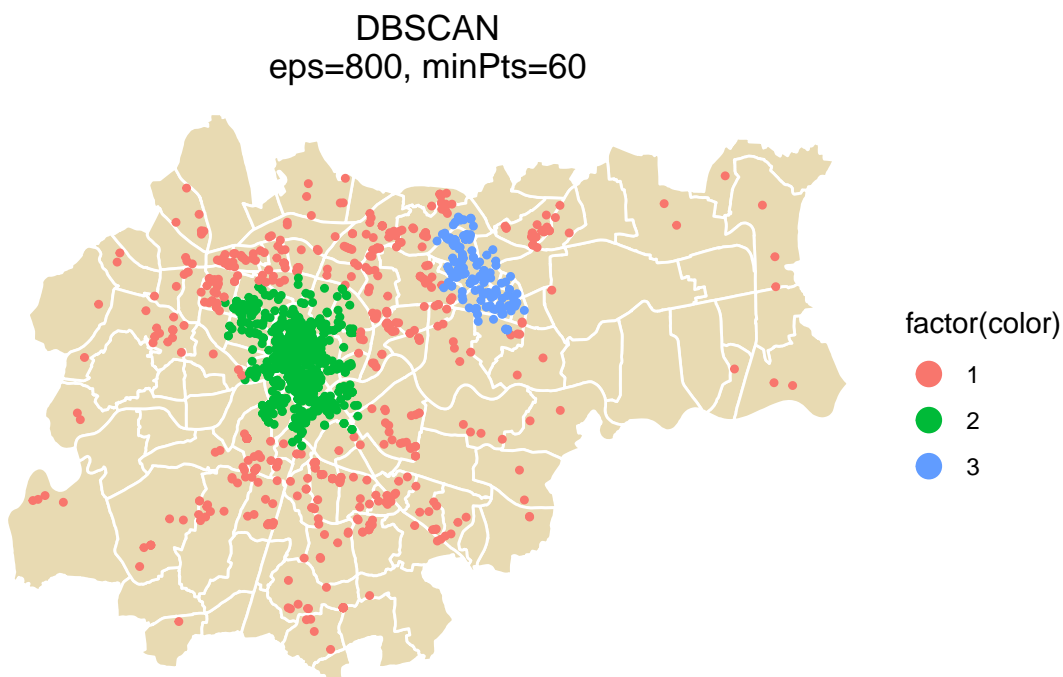
```
##   color Freq
## 1     1  824
## 2     2 1014
## 3     3   51
## 4     4   64
## 5     5   47
```

Klastrowizacja 2, ponownie z wykorzystaniem algorytmu DBSCAN, natomiast zmienionymi parametrami:

```

dbscan_res <- dbscan(cords, eps = 800, minPts = 60)
color<-dbscan_res$cluster+1
plot2<-ggplot() +
  geom_polygon(data = my_spdf, aes( x = long, y = lat, group = group), fill="#E8DAB2", color="white") +
  geom_point(data = cords, aes(x = X, y = Y,color=factor(color)), size = 0.9) +
  theme_void() +
  theme(plot.margin = unit(c(1,1,1,1), "cm")) +
  ggtitle("DBSCAN\neps=800, minPts=60") +
  theme(plot.title = element_text(hjust = 0.5))+
  guides(colour = guide_legend(override.aes = list(size=4)))
plot2

```



Podwajając wartości parametrów spodziewałem się, że liczba klastrow drastycznie zmaleje. Można zauważyć, że trzy klastry z poprzedniej wizualizacji zostały scalone (numer 2,3 oraz 5). Do zielonego klastra należy już ponad 60% obserwacji, natomiast 10% zarejestrowanych wykroczeń odnotowane zostały na obszarze lub w okolicach Bienczy oraz zachodniej części Nowej Huty i Wzgórz Krzesławickich (możemy zaobserwować, że jest to rozszerzony klaster numer 4 z poprzedniej wizualizacji, jego liczność zwiększyła się trzykrotnie)

```
as.data.frame(table(color))
```

```

##   color Freq
## 1      1  563
## 2      2 1241
## 3      3  196

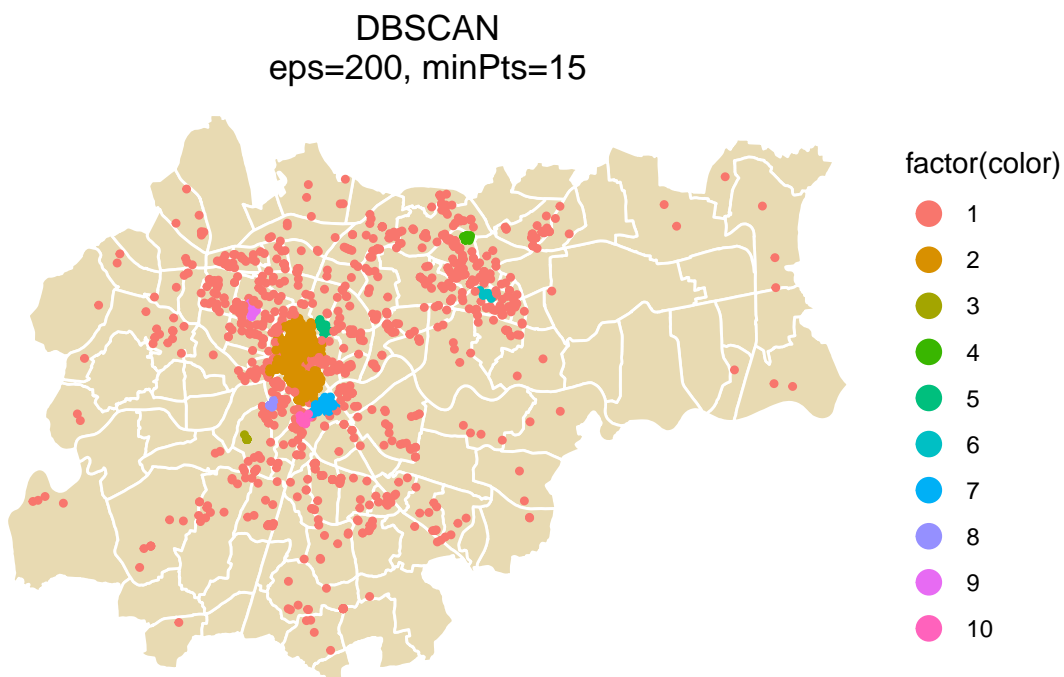
```

Kolejna klasteryzacja (3.) z wykorzystaniem algorytmu DBSCAN oraz parametrami o znacznie mniejszej wartości.

```

dbscan_res <- dbscan(cords, eps = 200, minPts = 15)
color<-dbscan_res$cluster+1
plot3<-ggplot() +
  geom_polygon(data = my_spdf, aes( x = long, y = lat, group = group), fill="#E8DAB2", color="white") +
  geom_point(data = cords, aes(x = X, y = Y,color=factor(color)), size = 0.9) +
  theme_void() +
  theme(plot.margin = unit(c(1,1,1,1), "cm")) +
  ggtitle("DBSCAN\neps=200, minPts=15") +
  theme(plot.title = element_text(hjust = 0.5) )+
  guides(colour = guide_legend(override.aes = list(size=4)))
plot3

```



Liczba klastrow wzrosła odwrotnie proporcjonalnie do ich liczności. W dalszej części klastrem zdecydowanie dominującym jest klastrowy numer 2, zlokalizowany na obszarze całego Starego Miasta. Dwa kolejne klastry pod względem wielkości zlokalizowane są w okolicach Olszy oraz Podgorza (oba bezpośrednio sąsiadują z klastrem numer 2). Kolejne klastry (wszystkie o liczności z zakresu 18-23 obserwacji) znajdują się na następujących obszarach: Zakrzówek, Debniki, Krowodrza, oraz dwa w okolicach Bienczyca i Mistrzejowic. Zważywszy na mały epsilon mniejsze klastry obejmują obszar danej ulicy lub skrzyżowania. Gdyby zaobserwowane wykroczenia dotyczyły np. złego parkowania, można by wnioskować, że dane miejsca są słabo oznaczone, przez co kierowcy parkują w niedozwolony sposób. Uważam również, że dobrane parametry z praktycznego punktu widzenia są zbyt małe i nie dostarczają zadowolonych informacji i mogą wprowadzać w błąd, gdyż większość klastrowy bezpośrednio sąsiaduje ze sobą.

```
as.data.frame(table(color))
```

```
##      color Freq
```

## 1	1	991
## 2	2	783
## 3	3	19
## 4	4	19
## 5	5	56
## 6	6	23
## 7	7	49
## 8	8	21
## 9	9	21
## 10	10	18

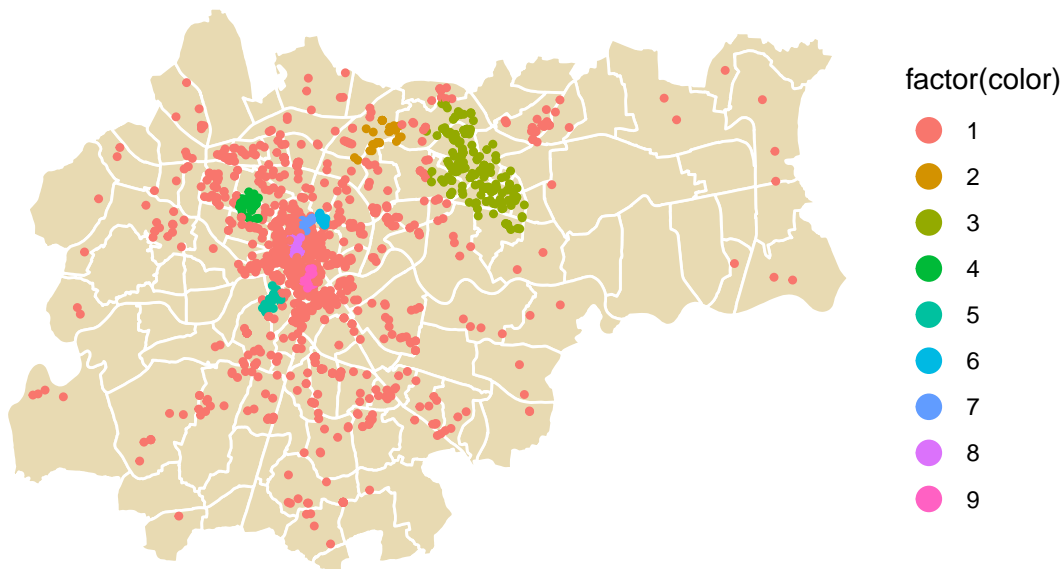
Algorytm HDBSCAN: Parametry wejściowe: minPts - minimalna wymagana liczba punktów znajdująca się w pewnej odległości, aby uznać je za klastę. Schemat działania: Jest to modyfikacja algorytmu DBSCAN. Pierwszym krokiem jest znalezienie odległości rdzeniowej, czyli odległości w jakiej leży ilość punktów podanej w parametrze minPts. Następnie algorytm wyszukuje największe skupiska, biorąc pod uwagę gęstość rozmieszczenia punktów wokół. Zalety: -jest zbudowany dla rzeczywistego scenariusza, w którym dane są o różnej gęstości -jest jeszcze szybszy od DBSCAN -jedynym argumentem jest minimalny rozmiar klastry, jest to dla nas bardziej intuicyjny parametr, gdyż na ogół wiemy jakiego rozmiaru klastry się spodziewamy/oczekujemy. Wady: -tak jak w DBSCAN - nie daje bezpośredniej możliwości zdefiniowania oczekiwanej liczby klastry - ich liczba wynika z wartości parametrów wejściowych

Klasteryzacja 1 z wykorzystaniem algorytmu HDBSCAN:

```
hdbscan_res <- hdbscan(cords, minPts = 30)
color<-hdbscan_res$cluster+1
plot4<-ggplot() +
  geom_polygon(data = my_spdf, aes( x = long, y = lat, group = group), fill="#E8DAB2", color="white") +
  geom_point(data = cords, aes(x = X, y = Y,color=factor(color)), size = 0.9) +
  theme_void() +
  theme(plot.margin = unit(c(1,1,1,1), "cm")) +
  ggtitle("HDBSCAN\nminPts=30") +
  theme(plot.title = element_text(hjust = 0.5) ) +
  guides(colour = guide_legend(override.aes = list(size=4)))
plot4
```



## HDBSCAN minPts=30



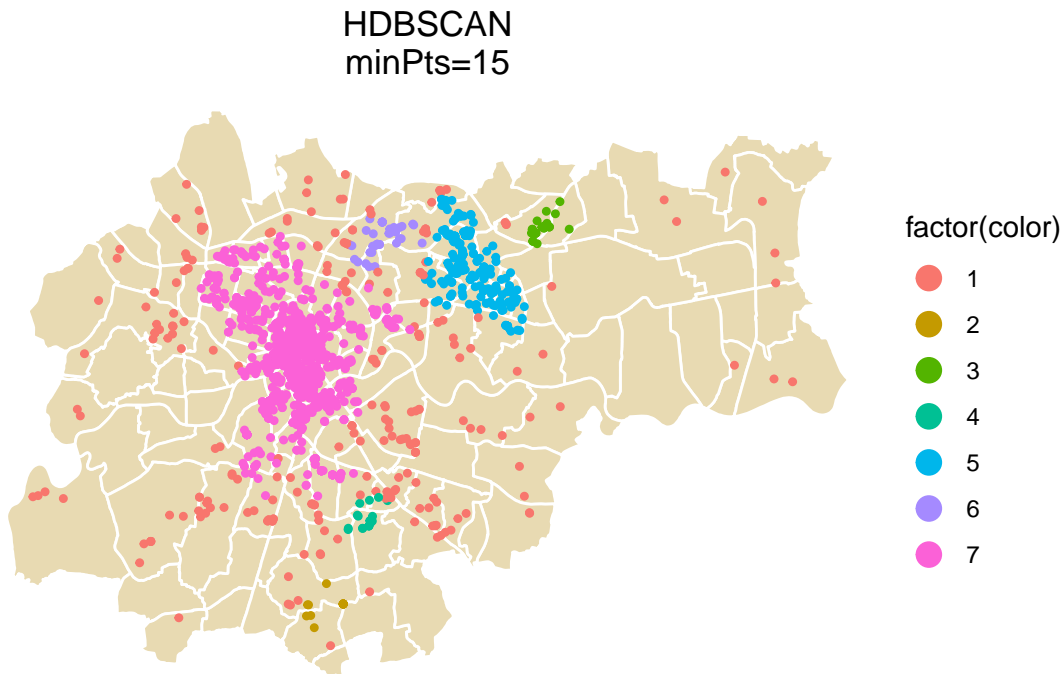
Porównując rezultat z pierwszą wizualizacją algorytmem DBSCAN od razu zauważamy, że przy tej samej liczbie minimalnej liczby punktów tworzącej klastrow ich ilość podwoiła się. Może to sugerować, że przyjęty tam parametr  $\text{eps}=400$  był zbyt mały i mogłoby się nie sprawdzić w pewnych zastosowaniach. Potwierdza się również, że algorytm HDBSCAN przy tworzeniu klastrow bierze zmienną gęstość występowania zjawisk w zależności od otoczenia - bardzo dobrze widac to porównując klastry 3 oraz 8. Punkty wchodzące w skład trzeciego klastra (zlokalizowany na pograniczu Mistrzejowic, Bińczyc oraz Krzesławic) są mocno odseparowane i dlatego, mimo że gęstość punktowa tego klastra w porównaniu do innych klastrow jest mała, zostały uznane jako klastrow. Dla porównania osiem klastrow, znajdujących się w samym centrum starego miasta (prawdopodobnie są to okolice Rynku), na mapie sprawia wrażenie raczej fragmentu klastra. Wynika to z faktu, że gęstość punktowa wokół tego obszaru, mimo że jest spora, jest nieco mniejsza i na tej podstawie algorytm stworzył klastrow w miejscu o największej gęstości zarejestrowanych wykroczeń w obrębie Starego Miasta. Pozostałe klastry znajdują się tak jak poprzednio na następujących obszarach: Krowodrza, Prądnik Czerwony, Debniki/Zakrzówek, ponadto w sumie na Starym Mieście wydzielono cztery klastry. Rozmiary klastrow:

```
as.data.frame(table(color))
```

```
##   color Freq
## 1     1 1212
## 2     2   30
## 3     3  214
## 4     4   40
## 5     5   36
## 6     6   53
## 7     7   80
## 8     8  233
## 9     9  102
```

Klasteryzacja 2 z wykorzystaniem algorytmu HDBSCAN oraz mniejszym parametrem wejściowym minPts:

```
hdbscan_res <- hdbscan(cords, minPts = 15)
color<-hdbscan_res$cluster+1
plot5<-ggplot() +
  geom_polygon(data = my_spdf, aes( x = long, y = lat, group = group), fill="#E8DAB2", color="white") +
  geom_point(data = cords, aes(x = X, y = Y,color=factor(color)), size = 0.9) +
  theme_void() +
  theme(plot.margin = unit(c(1,1,1,1), "cm")) +
  ggtitle("HDBSCAN\nminPts=15") +
  theme(plot.title = element_text(hjust = 0.5) ) +
  guides(colour = guide_legend(override.aes = list(size=4)))
plot5
```



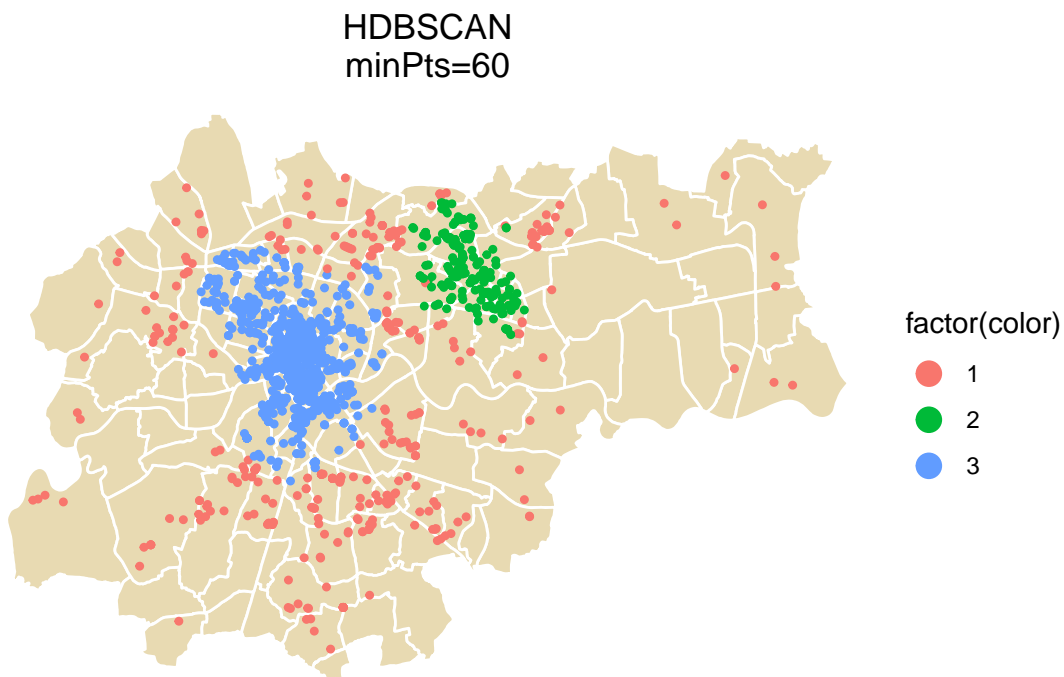
Znaczne zmniejszenie minimalnej liczby punktów wymaganej do utworzenia klastra prawdopodobnie nie dało satysfakcjonującego rezultatu. Ponad 87% punktów należy do któregoś z sześciu klastrow, przy czym największy stanowi ponad 70% obserwacji, natomiast najmniejszy niespełna 0.8%. Uważam, że algorytm ten z tak małym argumentem minPts nie sprawdził się w tym przypadku, ponieważ gęstość punktowa jest zbyt zróżnicowana, przez co obszar, gdzie zarejestrowana liczba wykroczeń jest najwyższa (Stare Miasto, Krowodrza, część Prądnika Białego, Bronowic, Zwierzynca i Debników) uznany został za obszar występowania jednego, ogromnego klastra. Na tej samej mapie na obszarze Swoszowic zauważyć możemy klaster numer 2, na który składa się 16 punktów. Intuicyjnie na obszarze tym ciężko stwierdzić, żeby faktycznie występował jakiś klaster, natomiast ze względu na znikomą ilość wykroczeń wokół tego obszaru algorytm zakwalifikował to zgrupowanie jako samodzielny klaster. Wnioski jakie należałoby wyciągnąć w tym miejscu to, że parametr minPts należy dobierać odpowiednio proporcjonalnie do liczby punktów.

```
as.data.frame(table(color))
```

```
##   color Freq
## 1     1  247
## 2     2   16
## 3     3   27
## 4     4   17
## 5     5  223
## 6     6   39
## 7     7 1431
```

Klasteryzacja 3 z wykorzystaniem algorytmu HDBSCAN:

```
hdbscan_res <- hdbscan(cords, minPts = 60)
color<-hdbscan_res$cluster+1
plot6<-ggplot() +
  geom_polygon(data = my_spdf, aes( x = long, y = lat, group = group), fill="#E8DAB2", color="white") +
  geom_point(data = cords, aes(x = X, y = Y,color=factor(color)), size = 0.9) +
  theme_void() +
  theme(plot.margin = unit(c(1,1,1,1), "cm")) +
  ggtitle("HDBSCAN\nminPts=60") +
  theme(plot.title = element_text(hjust = 0.5) ) +
  guides(colour = guide_legend(override.aes = list(size=4)))
plot6
```



Prezentowana wizualizacja jest mocno zbliżona do prezentowanego wyżej rezultatu klastrowania za

pomoca algorytmu DBSCAN z parametrami minPts=60 i eps=800. W tym przypadku oba klastry sa nieznacznie wieksze, natomiast intuicyjnie obszar klastrow jest wlasciwy. Zastosowanie takiej, stosunkowo duzej, wartosci parametru minPts moze byc przydatne w momencie, gdy nie potrafimy okreslic punktow granicznych dla glownych klastrow z zbiorze.

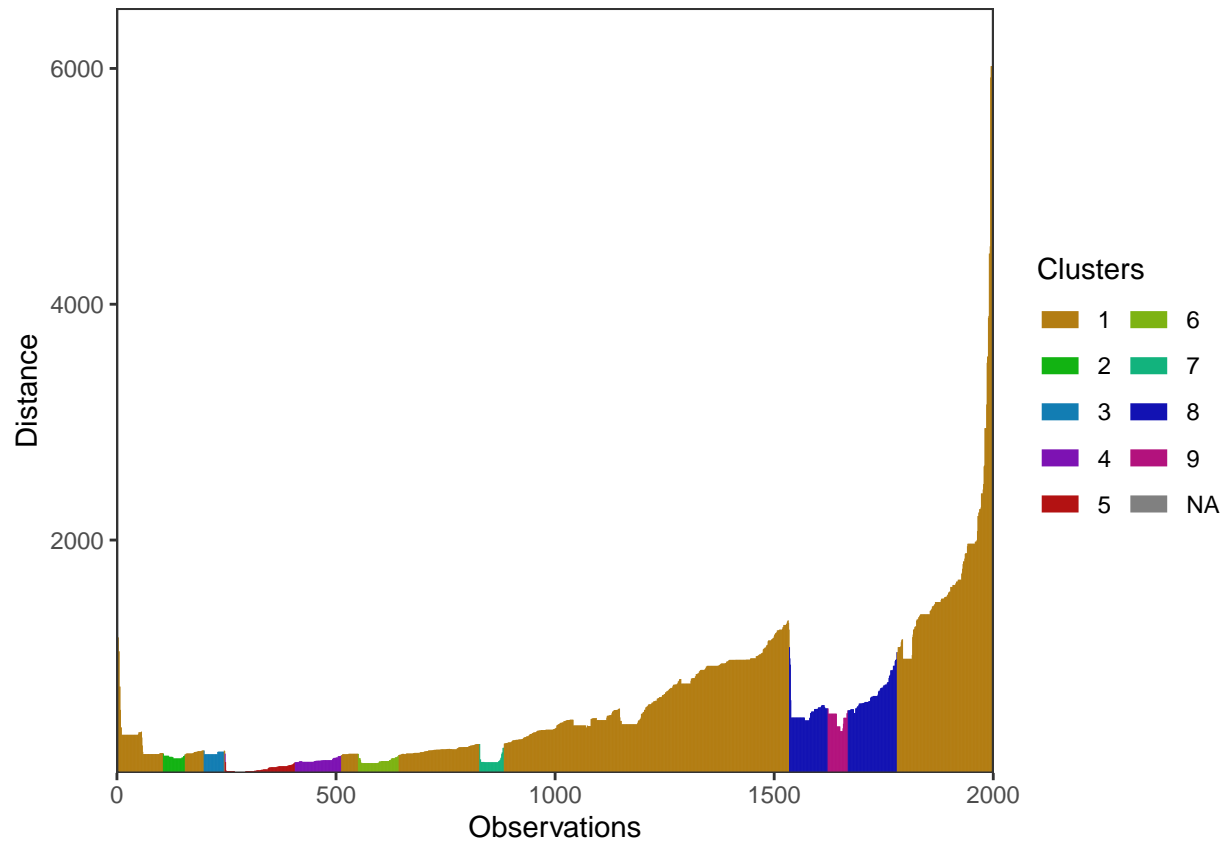
```
as.data.frame(table(color))
```

```
##   color Freq
## 1     1   370
## 2     2   236
## 3     3  1394
```

Algorytm DBSCAN: Parametry wejsciowe: minPts - jak wyzej, zdefiniowana minimalna liczba obiektow rozwarzanych jako grupa/klastro Xi - wspolczynnik stromosci Schemat dzialania: Algorytm wybiera losowo jeden obiekt, od ktorego rozpoczyna proces przetwarzania obiektow, nastepnie rozpoczyna analize kolejnego, najblizszego obiektu, dodaje go do listy przetworzonych obiektow i uznaje za prekursora w poszukiwaniu kolejnego obiektu. Poszukiwanie dla kolejnych punktow najblizszych sasiadow do momentu, gdy przeanalizowane zostana wszystkie punkty. Zasadnicza roznicza pomiedzy DBSCAN i OPTICS jest taka, ze w tej drugiej wazna jest kolejnosc przetwarzanych obiektow. Algorytm OPTICS zwraca uporzadkowana baze danych h, przechowujac dla kazdego elementu informacje o odlgosci-jadra, odlgosci-osiagalnosc. Graficzna prezentacja wyzej uporzadkowanych punktow pozwala na latwe zrozumienie danych. Doliny na takim wykresie oznaczaja wystepowanie klastrow. Im dolina jest szersza, tym kalster jest wiekszy. Rozna glebokosc dolin wskazuje natomiast na rozna gestosc klastrow. Zalety: -eliminuje wade swojego pierwowzoru(DBSCAN) i potrafi znalezc zgrupowania hierarchicznych (mniejsze skupienie wewnatrz wiekszego) i rozny stopniu zagieszczenia Wady: -brak przetestowanej heurystyki do wyznaczenia wartosci parametru MinPts -wolniejszy okolo 1,6 raza od DBSCAN

Klasteryzacja 1 z wykorzystaniem algorytmu OPTICS

```
optic_res <- optics(cords, minPts = 40)
optic_res2 <- extractXi(optic_res, xi = 0.04)
ggplot_optics(optic_res, groups = optic_res2$cluster)
```



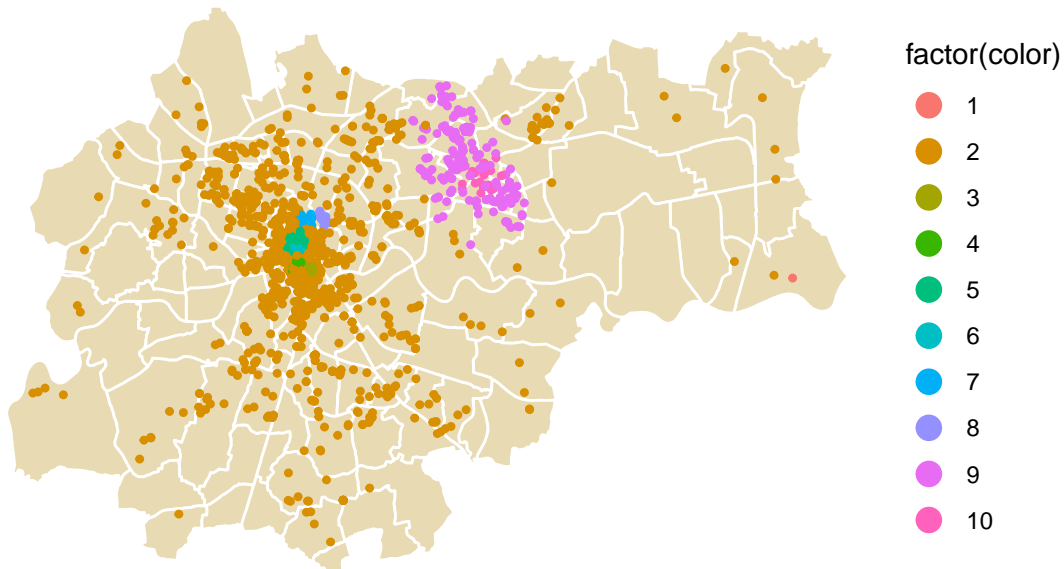
```

color<-optic_res2$cluster+1
plot7<-ggplot() +
  geom_polygon(data = my_spdf, aes( x = long, y = lat, group = group), fill="#E8DAB2", color="white") +
  geom_point(data = cords, aes(x = X, y = Y,color=factor(color)), size = 0.9) +
  theme_void() +
  theme(plot.margin = unit(c(1,1,1,1), "cm")) +
  ggtitle("OPTICS\nminPts=40, xi=0.04") +
  theme(plot.title = element_text(hjust = 0.5) ) +
  guides(colour = guide_legend(override.aes = list(size=4)))
plot7

```

## OPTICS

minPts=40, xi=0.04



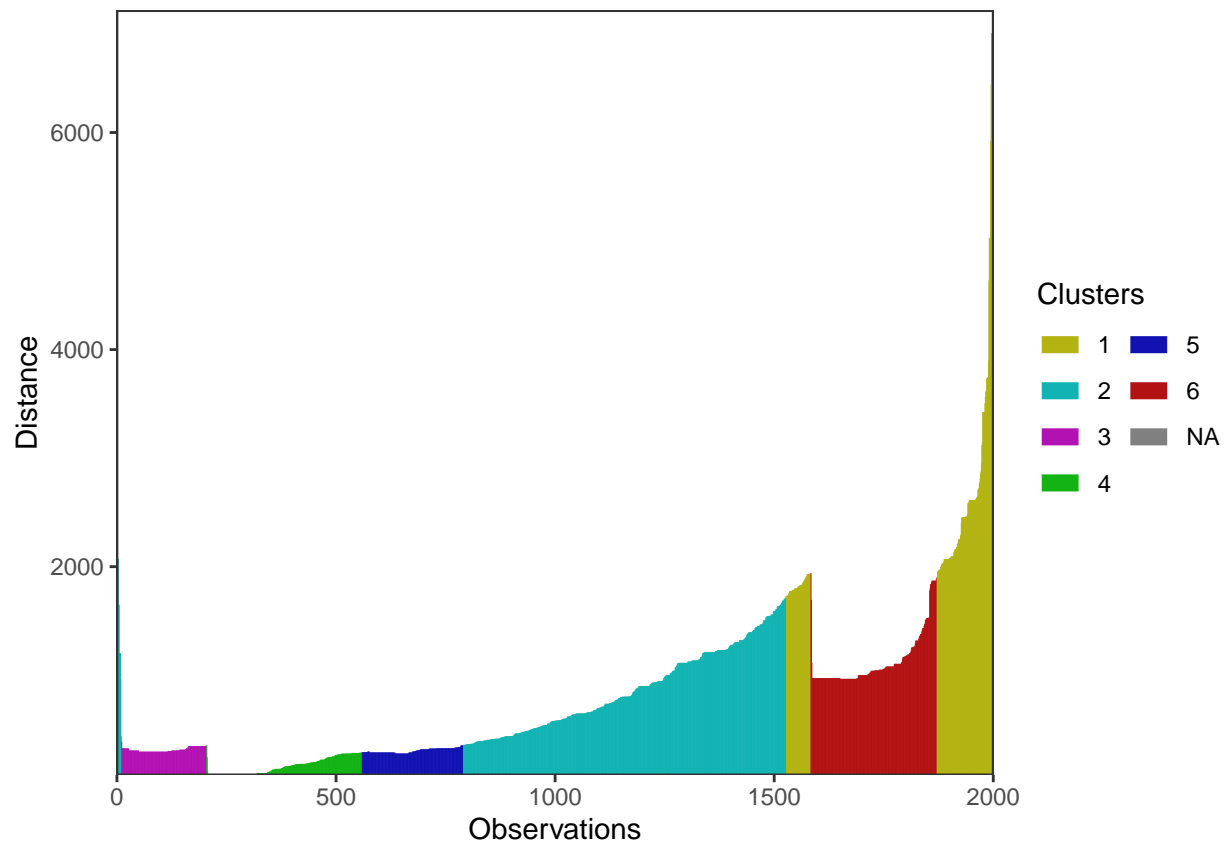
Jak widac dla wybranych parametrow, wszystkie punkty (z wyjatkiem jednego) zostaly zakwalifikowane do ktoregos z klastrow. Przy danych parametrach wejsciowych otrzymalem 8 mniejszych klastrow o bardzo duzej gestosci punktowej oraz jeden zawierajacy ponad 60% obserwacji. Piec mniejszych klastrow zlokalizowanych jest na terenie Starego Miasta, kolejne na Czyzynach oraz Bienczycach. Nalezaloby sie zastanowic czy parametry zostaly dobrane odpowiednio i czy klaster numer 2 faktycznie miesci sie w definicji slowa klaster.

```
as.data.frame(table(color))
```

```
##      color Freq
## 1         1     1
## 2         2 1242
## 3         3    50
## 4         4    46
## 5         5   107
## 6         6   159
## 7         7    93
## 8         8    56
## 9         9   200
## 10        10    46
```

Klasteryzacja 2 z wykorzystaniem algorytmu OPTICS

```
optic_res <- optics(cords, minPts = 100)
optic_res2 <- extractXi(optic_res, xi = 0.01)
ggplot_optics(optic_res, groups = optic_res2$cluster)
```



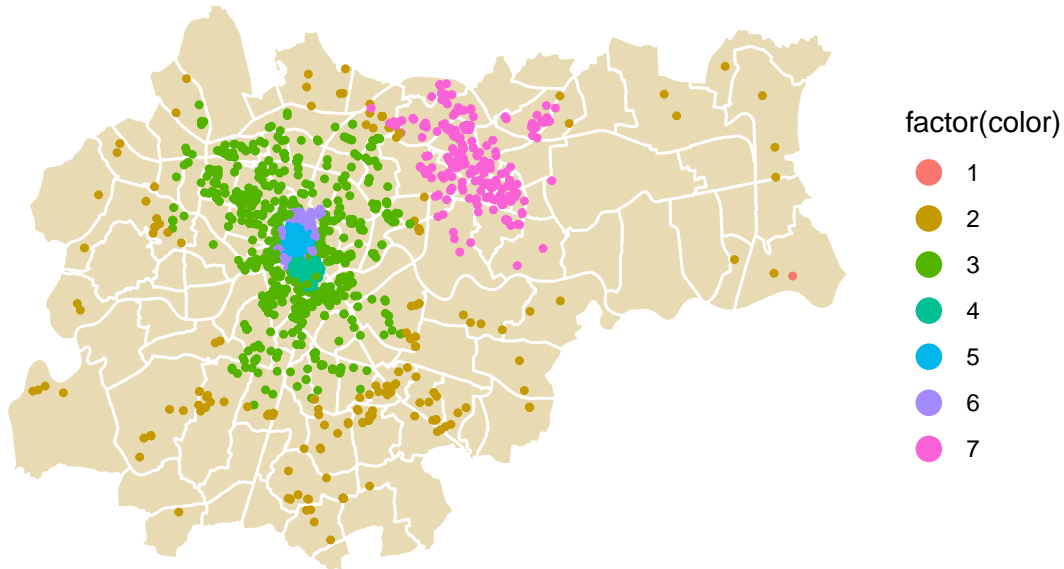
```

color<-optic_res2$cluster+1
plot8<-ggplot() +
  geom_polygon(data = my_spdf, aes( x = long, y = lat, group = group), fill="#E8DAB2", color="white") +
  geom_point(data = cords, aes(x = X, y = Y,color=factor(color)), size = 0.9) +
  theme_void() +
  theme(plot.margin = unit(c(1,1,1,1), "cm")) +
  ggtitle("OPTICS\nminPts=100, xi=0.01") +
  theme(plot.title = element_text(hjust = 0.5) ) +
  guides(colour = guide_legend(override.aes = list(size=4)))
plot8

```

## OPTICS

minPts=100, xi=0.01



Diametralna zmiana parametrów w celu znalezienia głównych klastrów. Tym razem na terenie starego miasta znajdują się 3 klastry, wszystkie zdarzenia wokół Starego Miasta zakwalifikowane zostały do następnego klastra (37% wszystkich zarejestrowanych wykroczeń) i jeszcze jeden na terenie Czyżyn, Bienczyń i Mistrzejowic. Ponownie wszystkie pozostałe punkty zakwalifikowane zostały do zewnętrznego klastra, co moim zdaniem jest błędem.

```
as.data.frame(table(color))
```

##	color	Freq
## 1	1	1
## 2	2	186
## 3	3	745
## 4	4	194
## 5	5	355
## 6	6	231
## 7	7	288

**PODSUMOWANIE I WNIOSKI** Każda z metod klasteryzacji charakteryzuje się pewnymi cechami, które w danej sytuacji mogą okazać się kluczowe do wykorzystania tego konkretnego algorytmu. Algorytm DBSCAN jest najbardziej intuicyjny i daje nam największe pole manewru. Osobiscie uważam jednak, że lepszym algorytmem jest HDBSCAN, który bierze pod uwagę zmienną gęstość punktową i w lepszy sposób wyznacza granice klastrów. Jeśli chodzi o algorytm OPTICS to jest on również zmodyfikowana wersja algorytmu DBSCAN i uważam, że jego bardzo dużym atutem jest umiejętność tworzenia klastrów w klastrach, natomiast nie udało mi się dobrać argumentów w taki sposób, aby punkty niezgrupowane uznane zostały za szum.