Raport2_TWD_p1

Patryk Słowakiewicz, Jan Gąska, Michał Piasecki

30 11 2020

Wstęp

Naszym zadaniem w tym tygodniu było analiza wybranych danych oraz przetestowanie różnych typóW wykresów tak aby wybrać najbardziej odpowiednie do wizualizacji danych które nas interesują. Korzystamy ze zbioru danych dla całego świata z którego później wyciągamy państwa które nas interesują. Wybór państw miał odzwiercaedlać trendy dla całych kontynentów oraz wybierać kraje graniczne w wielu analizowanych przez nas aspektach które można porównać do skali ogólnoświatowej.

Dla lepszej agrgacji danych na osi Y została zastosowana skala logarytmiczna.

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

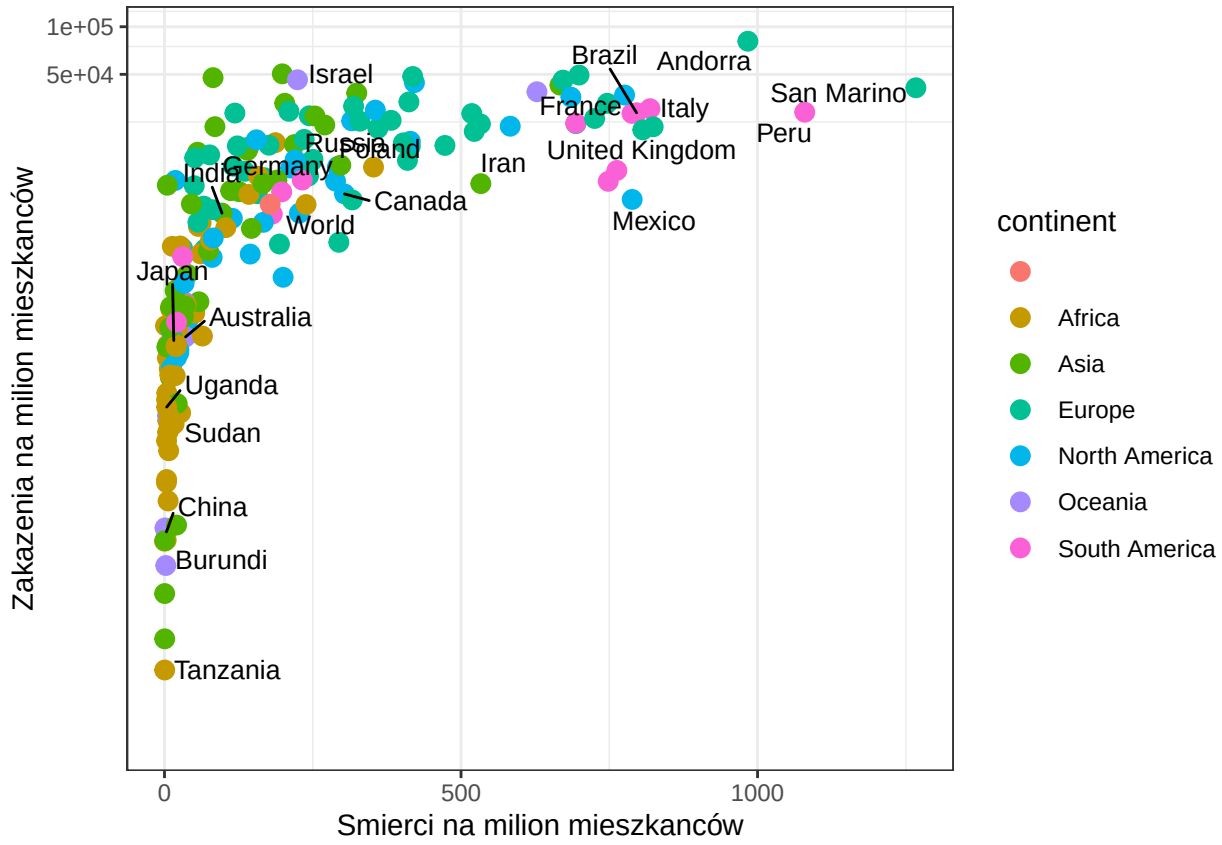
```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(ggrepel)
library(RColorBrewer)

covid_data_word <- read.csv("owid-covid-data.csv")

interesujace_panstwa <- c("Russia","Japan","Poland","France","India","Italy","China","Brazil","Mexico",
                          "Israel","Australia","Iran","Germany","Spain","Canada","Uganda","World","San
                          "Andorra","United Kingdom","Burundi","Tanzania")


data_sample_1 <-  select(covid_data_word,date,continent,total_tests_per_thousand,human_development_inde
       total_cases,total_deaths,total_tests,population,location)
data_sample_2 <-  select(covid_data_word,date,total_cases_per_million,location,continent,total_deaths_p
```

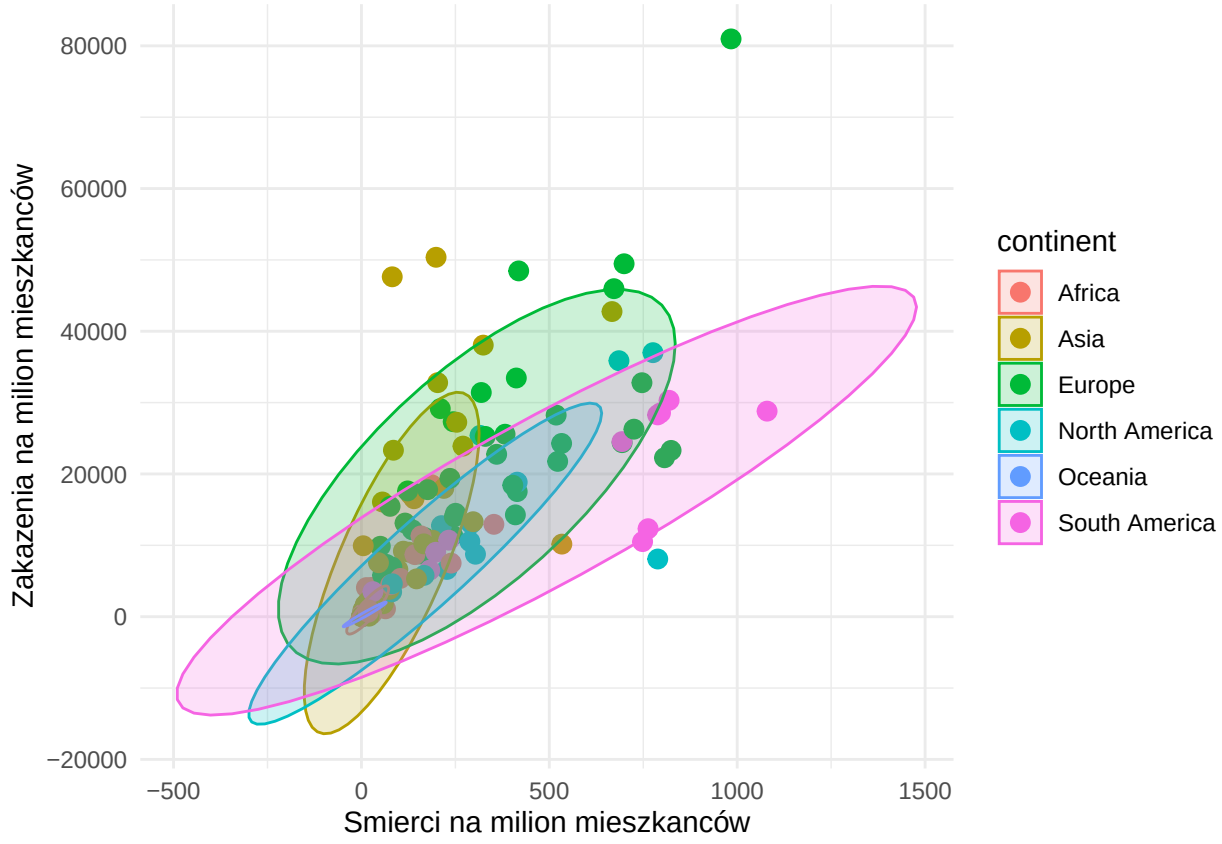
```
                          population_density,human_development_index)
data_sample_3 <- subset(data_sample_2,date == "2020-11-23")

data_sample_3 <- data_sample_3 %>%
  mutate(label_for_plot = ifelse((location %in% interesujace_panstwa),
                                 as.character(location), ""))
data_sample_3 <- data_sample_3 %>%
  mutate(Developed_Index = ifelse((human_development_index > 0.7),
                                  "Developed", "Not Developed"))
data_sample_3 <- data_sample_3 %>%
  mutate(HDI_Group = case_when((human_development_index > 0.8)~"Very High HDI",
                               (human_development_index > 0.7&human_development_index <= 0.8)~"Hig
                               (human_development_index > 0.550&human_development_index <= 0.7)~"
                               (human_development_index > 0.350&human_development_index <= 0.550)
                               TRUE ~ "0"))
```

Wybrane wizualizacje

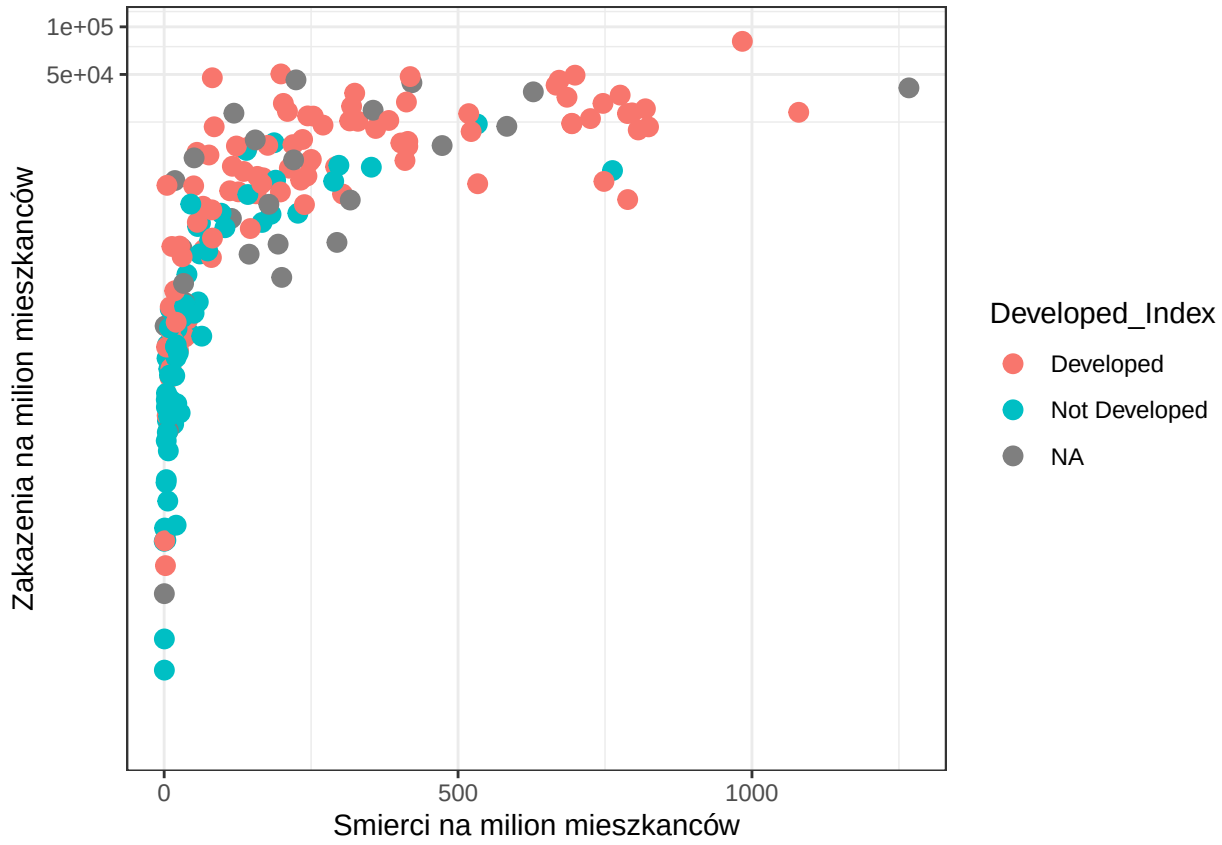
Zaczeliśmy analizę od wykresu punktowego dzięki któremu możemy umieścić wiele wymiarów. Kolor rozróżnia przynależność krajów do kontyntynentów. A na osiach oznaczone są odpowiednio Zakażenia na milion mieszkańców oraz Śmierci na milion mieszkańców. Atutem tego wykresu jest łatwe przedstawinie trendu, porównywanie krajów jak radzą sobie z pandemią. Im bardziej kraj znajduje się w prawym dórnym rogu tym sytuacja pandemiczna jest gorsza. Na podstawie tego wykresu można wyciągnąć wiele ciekawych wniosków porównując je na tle działalności poszczególnych państw. Wadami jest pewna bariera wejścia, niezbędne jest posiadanie określonej wiedzy ze świata aby móc w pełni korzystać z interpretacji tej wizualizacji. Mimo wszystko jest wartościowa dla dużego grona odbiorców.

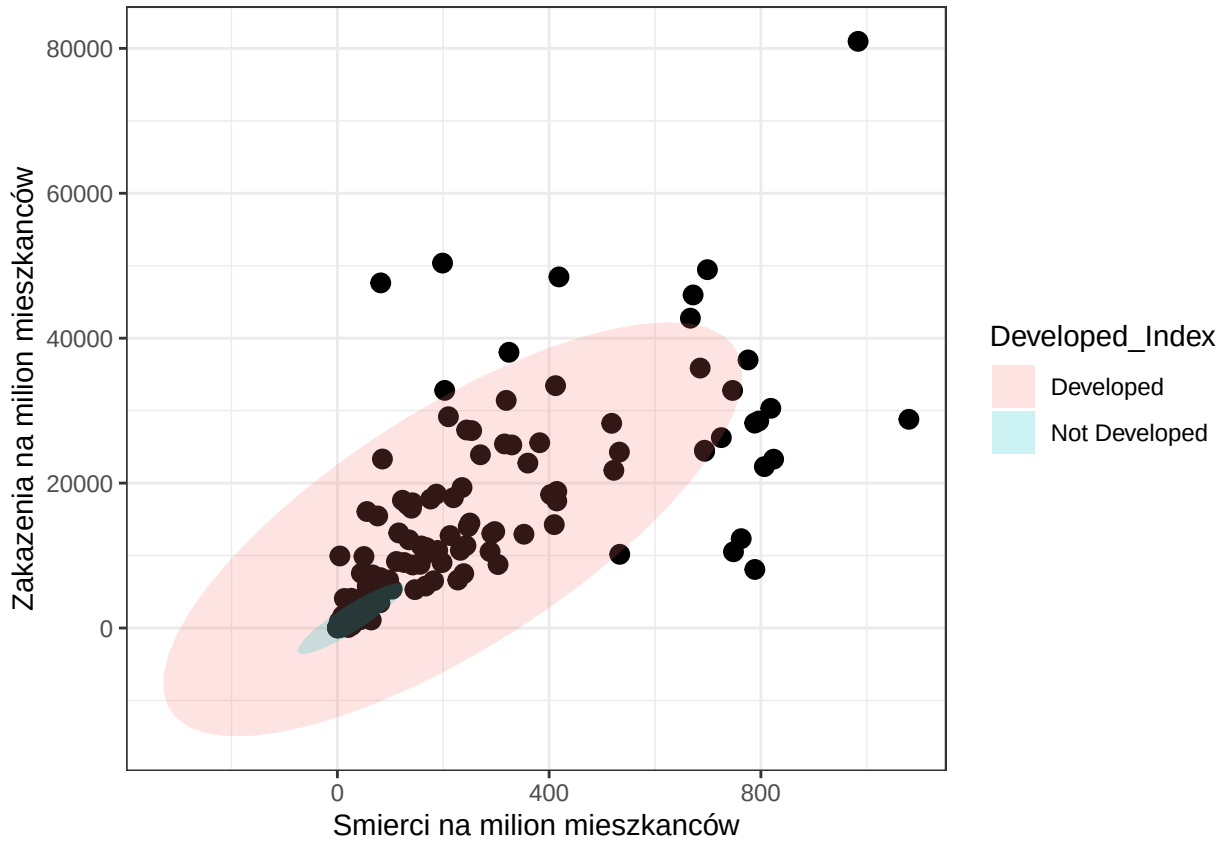


Ten wykres nie korzysta z osi logarytmicznej. Za to bardziej skupia się na całych kontynentach dzięki elipsom jesteśmy wstanie porównać jak rozwineła się pandemia w większej skali. takie zastosowenie ułatwia też porównywanie krajów 9brak skali logarytmicznej)

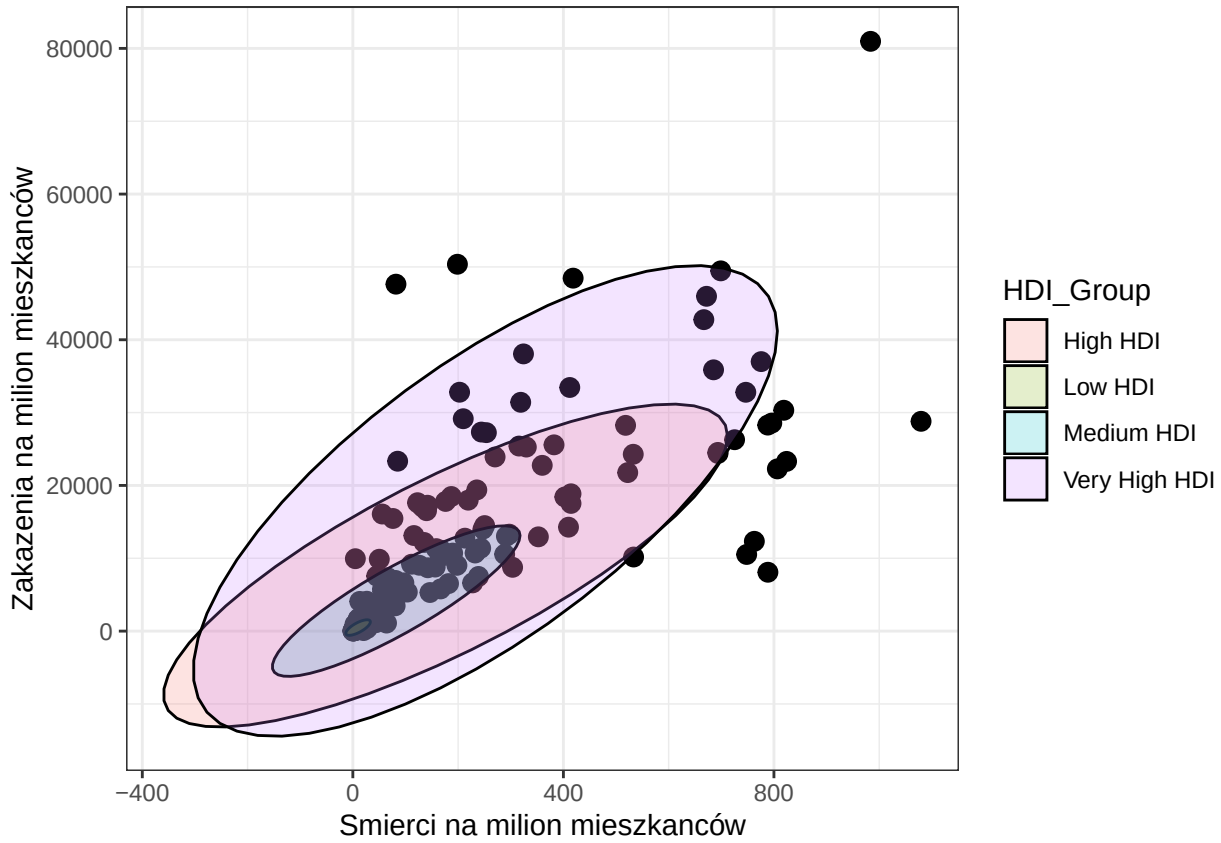


W następnych dwóch przykładach kraje są podzielone według Indeksu Rozwoju państwa. Dzięki takiemu zastosowaniu jesteśmy wstanie ocenić jak pandemia zależy od tego właśnie czynnika.


1e+05
5e+04
Zakazenia na milion mieszkanców
0
500
1000
Smierci na milion mieszkanców
Developed_Index
Developed
Not Developed
NA



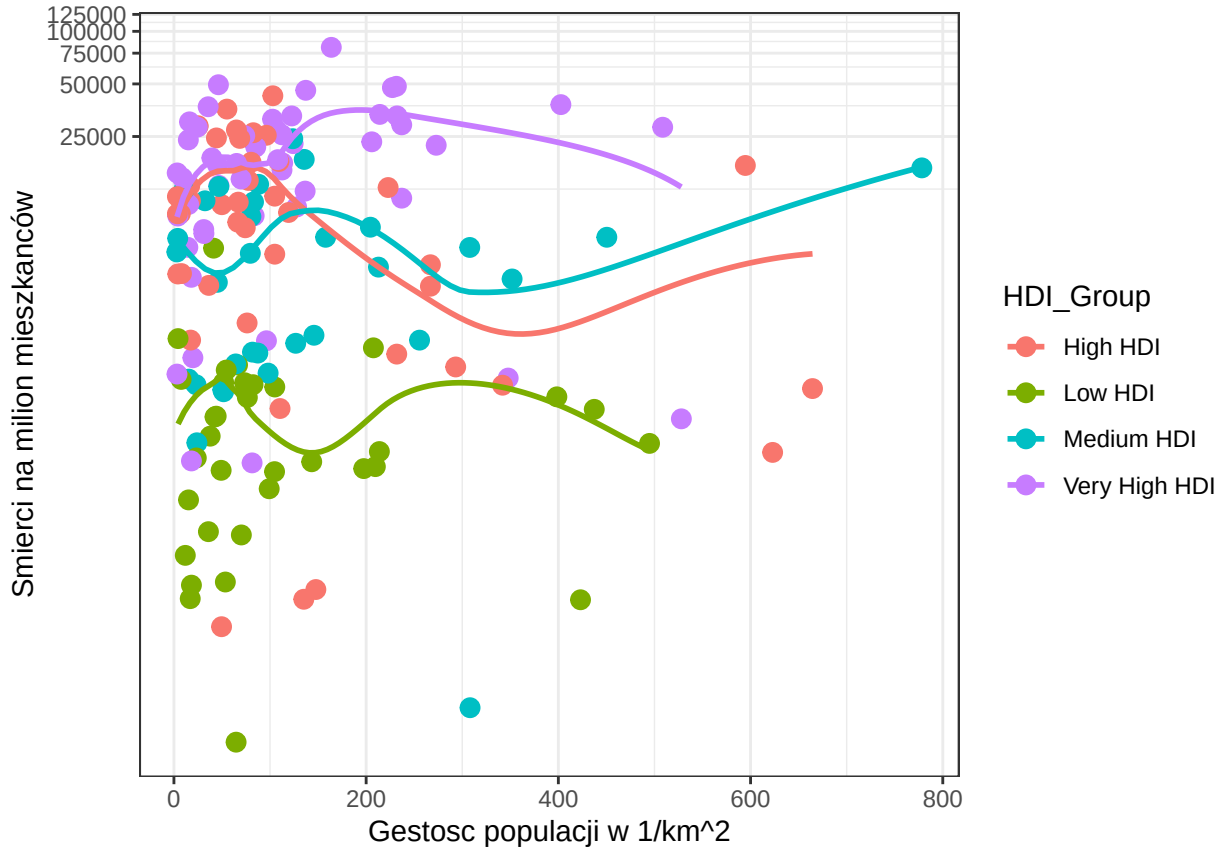
Na poniższych dwóch wykresach widzimy dane zagregowane w zależności od innego wskaźnika gospodarczo ekonomicznego. Tutaj jesteśmy w stanie ocenić jeszcze dokładniej jak wirus zaatoakował bardziej państwa wysoko



rozwinięte.

Ostatni wykres bada zależność śmiertelności od gęstości populacji. Odrzucamy dziewięć pierwszych wyników o najwyższym zagęszczeniu populacji ponieważ są to niewielkie państwa lub niewielkie enklawy, zawyżające średnią zagęszczenia. Dodanie krzywej przybliżającej tendencje poszczególnych grup pozwala zauważyć nieintuicyjną zależność że gęstość zaludnienia nie wpływa znacząco na śmiertelność i to wskaźnik HDI jest dominujący.

```
data_sample_4 <- data_sample_3 %>% arrange(desc(population_density))
data_sample_4 <- tail(data_sample_4,nrow(data_sample_4)-9)
ggplot(na.omit(data_sample_4), aes(y = total_cases_per_million, x =population_density,color = HDI_Group
  geom_point(size = 3) + theme_bw() +coord_trans(y="log10")+ ylab("Śmierci na milion mieszkańców")+
  xlab("Gęstość populacji w 1/km^2") + geom_smooth(se = FALSE)
```



Wnisek

Dzięki takiej prezentacji danych jesteśmy w stanie przedstawić dużo danych nie powodując nadmiernego szumu informacyjnego. Agregacja nie tylko po kontynentach ale również wskaźniku ekonomicznym lub społecznym daje nam pełniejszy obraz sytuacji oraz wskazuje istotne faktory różnic w rozwoju covid. Uważamy że wykresy punktowe zawierające dodatkowe informacje grupujące państwa są najlepszym wyborem do prezentacji naszych problemów. Dzięki dobremu opisowi sytuacji polityczno-społecznej w konkretnych krajach lub regionaych jesteśmy wstanie przekazać bardzo dużo ciekawych informacji nawet niewprawionemu odbiorcy.