

Eksploracyjna analiza danych | Inżynieria i Analiza Danych

Case Study

Dataset koronawirus

Hierarchiczna analiza skupień

Mateusz Bugdol

Nr indeksu: 419719

Grupa ćwiczeniowa: 1

1. Wstęp i cel case study.

W ramach projektu zaliczeniowego (Case Study) przeanalizowany zostanie wylosowany zbiór danych dotyczący rozprzestrzeniania się koronawirusa. Celem pracy jest wykorzystanie metod uczenia nienadzorowanego, a konkretnie hierarchicznej analizy skupień, do znalezienia ukrytych struktur w danych. Raport przedstawia proces przygotowania danych, przebieg analizy oraz interpretację uzyskanych klastrów.

2. Pakiety

W tej sekcji załadowane zostaną niezbędne biblioteki:

- `tidyverse` / `dplyr` : do manipulacji danymi i inżynierii cech.
- `cluster` : do algorytmów analizy skupień.
- `factoextra` : do wizualizacji wyników i dendrogramów.
- `ggplot2` : do wizualizacji map i wykresów.
- `DataExplorer` : do automatyzacji eksploracyjnej analizy danych (EDA), raportowania i analizy braków danych.
- `sf` : do obsługi danych przestrzennych (Simple Features) i tworzenia map wektorowych.
- `viridis` : do generowania czytelnych i dostępnych percepcyjnie palet kolorystycznych.
- `tidyr` : do porządkowania danych i zmiany ich struktury (formatowanie “tidy data”).
- `plotly` : do tworzenia interaktywnych wykresów na podstawie obiektów `ggplot2`.
- `treemapify` : do tworzenia wykresów typu mapa drzewa (treemap) w środowisku `ggplot2`.
- `scales` : do formatowania osi, etykiet i skal na wykresach (np. formatowanie liczb, procentów, osie logarytmiczne).
- `corrplot` : do wizualizacji macierzy korelacji w czytelnej formie graficznej.
- `caret` : do wspomagania procesu modelowania, w szczególności do zaawansowanego preprocessingu danych.
- `fpc` : do elastycznych procedur analizy skupień oraz statystycznej walidacji wyników grupowania.

```
library(tidyverse)
library(dplyr)
library(cluster)
library(factoextra)
library(ggplot2)
library(DataExplorer)
library(sf)
library(viridis)
library(tidyr)
library(plotly)
library(treemapify)
library(scales)
library(corrplot)
library(caret)
library(fpc)
```

3. Preprocessing danych

Wczytanie danych

Zbiór danych `coronavirus_dataset.csv` zawiera szczegółowe, dzienne od 22 stycznia 2020 roku do 16 marca 2020 roku. Statystyki dotyczą rozprzestrzeniania się pandemii COVID-19 w większości krajów na świecie. Dane są zorganizowane w formacie „długim” (long format), gdzie każdy wiersz stanowi pojedynczy rekord dla konkretnej daty, lokalizacji oraz typu przypadku.

```
df = read.csv2("../data/coronavirus_dataset.csv", header = T, sep = ",", na.strings = c("", "NA"))
head(df)
```

	Province.State	Country.Region	Lat	Long	date	cases	type
## 1	<NA>	Afghanistan	33	65	2020-01-22	0	confirmed
## 2	<NA>	Afghanistan	33	65	2020-01-23	0	confirmed
## 3	<NA>	Afghanistan	33	65	2020-01-24	0	confirmed
## 4	<NA>	Afghanistan	33	65	2020-01-25	0	confirmed
## 5	<NA>	Afghanistan	33	65	2020-01-26	0	confirmed
## 6	<NA>	Afghanistan	33	65	2020-01-27	0	confirmed

Zmienne zawarte w zbiorze:

- `Province.State` : Nazwa prowincji lub stanu (dla większych krajów, np. Chiny, USA, Kanada). Dla wielu krajów ta kolumna przyjmuje wartość pustą (`NA`),
- `Country.Region` :
- `Lat` (Latitude): Szerokość geograficzna środka danego regionu.
- `Long` (Longitude): Długość geograficzna środka danego regionu.
- `date` : Data obserwacji (rok-miesiąc-dzień).
- `cases` : Liczba przypadków zarejestrowana w danym dniu.
- `type` : Typ przypadku. Zmienna kategoryczna przyjmująca wartości:
 - `confirmed` : potwierdzone zakażenia,
 - `death` : zgony,
 - `Recovered` : wyzdrowienia.

W celu pogłębienia analizy, podstawowy zbiór danych COVID-19 został wzbogacony o zewnętrzne dane demograficzne pochodzące z portalu United Nations (<https://population.un.org/dataportal/data>). Dołączony zbiór `onz_data.csv` zawiera informacje o populacji, gęstości zaludnienia (osoby na kilometr kwadratowy) oraz

medianie wieku pochodzące z roku 2019 (stan przed pandemią). Pozwoli to na normalizację statystyk zachorowań oraz zbadanie wpływu struktury demograficznej na przebieg pandemii.

```
df_onz = read.csv("../data/onz_data.csv", header = T, stringsAsFactors = FALSE)
head(df_onz)
```

```
##          Country Iso3 Population Density MedianAge
## 1      Afghanistan AFG 37856121.0 58.30333 16.23426
## 2         Albania ALB 2885009.5 105.30000 34.70803
## 3        Algeria DZA 43294546.0 18.17770 27.43115
## 4 American Samoa ASM     50209.0 251.04500 26.25127
## 5      Andorra AND    76473.5 162.70957 41.73090
## 6       Angola AGO 32375632.5 25.96906 16.30242
```

Zmienne zawarte w zbiorze:

- Country : Nazwa kraju (np. Afghanistan, Albania).
- Iso3 : Trzyliterowy kod kraju w standardzie ISO 3166-1 alpha-3.
- Population : Całkowita liczba ludności w danym kraju.
- Density : Gęstość zaludnienia (liczba osób przypadająca na jednostkę powierzchni, zazwyczaj na km²).
- MedianAge : Mediana wieku ludności.

Poprawność struktury danych

Przed przystąpieniem do dalszej obróbki i agregacji danych, konieczne jest sprawdzenie, w jaki sposób zinterpretowano poszczególne kolumny podczas importu. Automatyczne wczytywanie plików CSV czasami błędnie przypisuje typy danych, traktując liczby lub daty jako ciągi znaków, co uniemożliwia wykonywanie na nich operacji matematycznych wymaganych w analizie skupień. Poniższy kod weryfikuje klasy atrybutów oraz dokonuje niezbędnych korekt.

```
sapply(df, class)
```

```
## Province.State Country.Region           Lat           Long          date
## "character"   "character"   "character"   "character"   "character"
##      cases        type
## "integer"     "character"
```

Wstępna analiza struktury danych wykazała, że zmienne geograficzne Lat (szerokość) i Long (długość) oraz zmienna date zostały pierwotnie zinterpretowane jako typ tekstowy (character).

```
df$Lat = as.numeric(df$Lat)
df$Long = as.numeric(df$Long)
df$date = as.Date(df$date)
```

Dokonano jasnej kówersji danych:

- Lat i Long przekształcono na typ numeryczny (numeric),
- date przekształcono na format daty (Date).

```
sapply(df, class)
```

```
## Province.State Country.Region          Lat        Long      date
## "character"    "character"       "numeric"   "numeric"  "Date"
##      cases           type
## "integer"      "character"
```

```
sapply(df_onz, class)
```

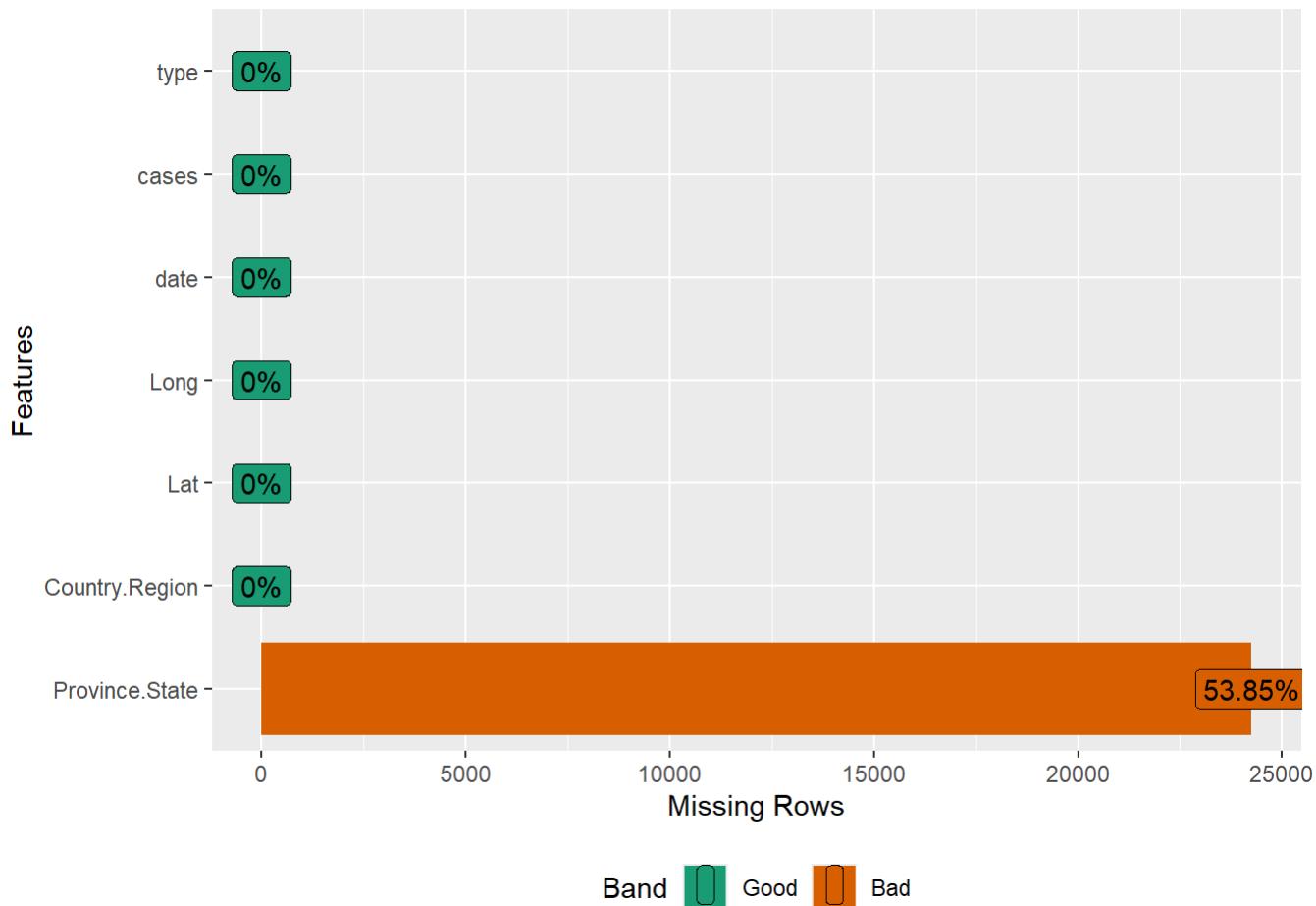
```
##      Country      Iso3 Population   Density MedianAge
## "character" "character"  "numeric"  "numeric"  "numeric"
```

Analiza struktury danych zbioru `df_onz` wskazuje, że zmienne zostały wczytane z poprawnymi typami danych. Zmienne identyfikacyjne `Country` i `Iso3` są typu tekstowego (`character`), natomiast kluczowe zmienne demograficzne: `Population`, `Density` oraz `MedianAge`, które posłużą do dalszej analizy statystycznej i klastrowania, zostały prawidłowo zinterpretowane jako typ numeryczny (`numeric`).

Czyszczenie i selekcja

Kluczowym etapem preprocessingu jest identyfikacja i obsługa brakujących wartości. Pozostawienie ich bez ingerencji mogłoby prowadzić do błędów podczas agregacji lub obliczania macierzy odległości.

```
plot_missing(df)
```



Jak widać na wygenerowanym wykresie, kolumna `Province.State` charakteryzuje się bardzo wysokim odsetkiem brakujących danych. Wynika to ze specyfiki raportowania – większość państw w zbiorze podaje statystyki ogólnokrajowe, a jedynie nieliczne (jak USA, Chiny czy Kanada) są rozbite na mniejsze jednostki administracyjne.

```
df$Province.State <- NULL
```

Ponieważ celem niniejszego badania jest porównanie efektywności walki z pandemią na poziomie całych państw, a nie poszczególnych regionów, zmienna ta jest zbędna. Jej usunięcie pozwoli na wyeliminowanie całkowite brakujących wartości oraz uproszczenie procesu ewentualnej agregacji danych.

Inżynieria danych (Feature Engineering)

Oryginalny zbiór danych posiada strukturę szeregow czasowych, gdzie każdy dzień jest osobnym rekordem. Aby przeprowadzić hierarchiczną analizę skupień, w której obiektem badanym jest kraj, konieczne jest sprowadzenie danych do postaci, gdzie jeden wiersz odpowiada jednemu państwu.

Zanim jednak przystąpimy do tej transformacji, musimy przygotować grunt pod połączenie danych epidemicznych ze zmiennymi demograficznymi z bazy ONZ (df_onz). Kluczowym wyzwaniem jest tu niespójność w nazewnictwie krajów pomiędzy zbiorami. W pierwszej kolejności identyfikujemy państwa z bazy COVID-owej (df), których nazwy nie pasują do tych w bazie ONZ.

```
setdiff(df$Country.Region, df_onz$Country)
```

```
## [1] "Bolivia"                      "Brunei"  
## [3] "Congo (Brazzaville)"          "Congo (Kinshasa)"  
## [5] "Cote d'Ivoire"                 "Holy See"  
## [7] "Iran"                          "Korea, South"  
## [9] "Kosovo"                        "Moldova"  
## [11] "occupied Palestinian territory" "Republic of the Congo"  
## [13] "Reunion"                       "Russia"  
## [15] "Taiwan*"                      "Tanzania"  
## [17] "The Bahamas"                  "Turkey"  
## [19] "Venezuela"                     "Vietnam"  
## [21] "US"                            "Cruise Ship"
```

Analiza wykazała szereg rozbieżności. Aby umożliwić poprawne złączenie tabel (join), dokonujemy ręcznej standaryzacji nazw w zbiorze df_onz, dostosowując je do konwencji przyjętej w danych COVIDowych.

```
df_onz <- df_onz %>%
  mutate(Country = recode(
    Country,
    "Bolivia (Plurinational State of)" = "Bolivia",
    "Brunei Darussalam" = "Brunei",
    "Congo" = "Congo (Brazzaville)",
    "Democratic Republic of the Congo" = "Congo (Kinshasa)",
    "Côte d'Ivoire" = "Cote d'Ivoire",
    "Vatican City" = "Holy See",
    "Iran (Islamic Republic of)" = "Iran",
    "Republic of Korea" = "Korea, South",
    "Republic of Moldova" = "Moldova",
    "Russian Federation" = "Russia",
    "United Republic of Tanzania" = "Tanzania",
    "Bahamas" = "The Bahamas",
    "Türkiye" = "Turkey",
    "Venezuela (Bolivarian Republic of)" = "Venezuela",
    "Viet Nam" = "Vietnam",
    "United States of America" = "US",
    "Kosovo (under UNSC res. 1244)" = "Kosovo",
    "State of Palestine" = "occupied Palestinian territory",
    "Réunion" = "Reunion",
    "China, Taiwan Province of China" = "Taiwan"
  ))
)
```

```
setdiff(df$Country.Region, df_onz$Country)
```

```
## [1] "Holy See"           "Republic of the Congo" "Cruise Ship"
```

Wynik funkcji setdiff wskazuje kraje z bazy COVID (df), które nie znalazły odpowiednika w bazie demograficznej (df_onz). Zostaną one pominięte w dalszej analizie (klastrowaniu) z następujących powodów:

- Holy See (Watykan) : W oryginalnym pliku df_onz najprawdopodobniej w ogóle nie było wiersza dla Watykanu ("Vatican City"). Skoro kraj nie istnieje w danych demograficznych, nie można zmienić jego nazwy ani użyć go w analizie.
- Cruise Ship : To ogniska zakażeń na statkach (brak terytorium i populacji).
- Republic of the Congo : Pozostała niespójność nazewnictwa (np. w bazie ONZ widnieje jako "Congo (Brazzaville)").

```
df_agg <- df %>%
  group_by(Country.Region, type) %>%
  summarise(
    cases = sum(cases),
    Lat = mean(Lat),
    Long = mean(Long),
    .groups = "drop"
  )
```

Dokonano agregacji danych, przekształcając szczegółowe zapisy dzienne w ogólne podsumowanie dla każdego kraju. Zsumowano liczbę przypadków dla poszczególnych typów zdarzeń, uzyskując całkowity bilans pandemii, oraz obliczono średnie współrzędne geograficzne, co pozwoliło wyznaczyć jeden punkt (centroid) reprezentujący każde państwo na mapie.

```
df_wide <- df_agg %>%
  pivot_wider(
    names_from = type,
    values_from = cases,
    names_prefix = "Total_",
    values_fill = 0
  )
```

Przekształcono strukturę danych z formatu “długiego” (long) na “szeroki” (wide). Zamiast kilku wierszy dla tego samego kraju (osobno dla zakażeń, zgonów i wyzdrowień), utworzono jeden unikalny rekord dla każdego państwa, w którym poszczególne statystyki stały się osobnymi kolumnami:

- Total_confirmed : Całkowita liczba zakażeń COVID-19 z okresu 11 luty do 22 marca 2020 roku w danym kraju.
- Total_death : Całkowita liczba śmierci spowodowanych przez COVID-19 z okresu 11 luty do 22 marca 2020 roku w danym kraju.
- Total_recovered : Całkowita liczba uleczonych przypadków zakażenia przez COVID-19 z okresu 11 luty do 22 marca 2020 roku w danym kraju.

Ułatwia to dalszą analizę, gdyż wszystkie informacje o danym kraju znajdują się teraz w jednym wierszu.

```
df_final <- df_wide %>%
  left_join(
    df_onz,
    by = c("Country.Region" = "Country")
  ) %>%
  mutate(
    Total_stillSick = Total_confirmed - Total_recovered - Total_death,
    Mortality_Rate = ifelse(
      Total_confirmed > 0,
      Total_death / Total_confirmed,
      0
    ),
    Recovery_Rate = ifelse(
      Total_confirmed > 0,
      Total_recovered / Total_confirmed,
      0
    ),
    Active_Rate = ifelse(
      Total_confirmed > 0,
      Total_stillSick / Total_confirmed,
      0
    ),
    Incidence_Rate = (Total_confirmed / Population),
    Mortality_per_capita = (Total_death / Population),
    StillSick_per_capita = (Total_stillSick / Population)
  ) %>%
  filter(is.finite(Incidence_Rate)) %>%
  na.omit()
```

Połączono zagregowane dane epidemiczne z danymi demograficznymi (zbiór ONZ), a następnie przeprowadzono inżynierię cech (feature engineering). Stworzenie wskaźników relatywnych pozwala na obiektywne porównanie sytuacji w krajach o drastycznie różnej liczbie ludności.

Do zbioru dodano następujące zmienne:

- `Total_stillSick` : Liczba aktywnych przypadków (różnica między całkowitą liczbą zakażeń a sumą wyzdrowień i zgonów).
- `Mortality_Rate`, `Recovery_Rate`, `Active_Rate` : Wskaźniki struktury przypadków, określające odpowiednio udział zgonów, wyzdrowień oraz wciąż chorujących w całkowitej liczbie potwierdzonych zakażeń.
- `Incidence_Rate`, `Mortality_per_capita`, `StillSick_per_capita` : Wskaźniki populacyjne – odpowiednio liczba zakażeń, zgonów i aktywnych przypadków przeliczona na jednego mieszkańca (znormalizowana względem populacji).

Finalnie zbiór oczyszczono z wartości nieskończonych oraz rekordów niekompletnych, przygotowując go bezpośrednio do modelowania.

```
head(df_final)
```

```
## # A tibble: 6 × 17
##   Country.Region     Lat    Long TotalConfirmed TotalDeath TotalRecovered Iso3
##   <chr>           <dbl>   <dbl>        <int>      <int>        <int> <chr>
## 1 Afghanistan       33     65            21         0          1 AFG
## 2 Albania           41.2   20.2          51         1          0 ALB
## 3 Algeria            28.0   1.66          54         4          12 DZA
## 4 Andorra            42.5   1.52          2         0          1 AND
## 5 Antigua and Ba...  17.1  -61.8          1         0          0 ATG
## 6 Argentina          -38.4  -63.6          56         2          1 ARG
## # i 10 more variables: Population <dbl>, Density <dbl>, MedianAge <dbl>,
## #   Total_stillSick <int>, Mortality_Rate <dbl>, Recovery_Rate <dbl>,
## #   Active_Rate <dbl>, Incidence_Rate <dbl>, Mortality_per_capita <dbl>,
## #   StillSick_per_capita <dbl>
```

W wyniku powyższych operacji otrzymano ramkę danych `df_final`, która stanowi bazę do właściwej analizy skupień.

4. Eksploracyjna analiza danych (EDA)

W tej sekcji przyjrzymy się bliżej rozkładom zmiennych, relacjom między nimi oraz rozmieszczeniu geograficznemu pandemii przed przystąpieniem do grupowania.

```

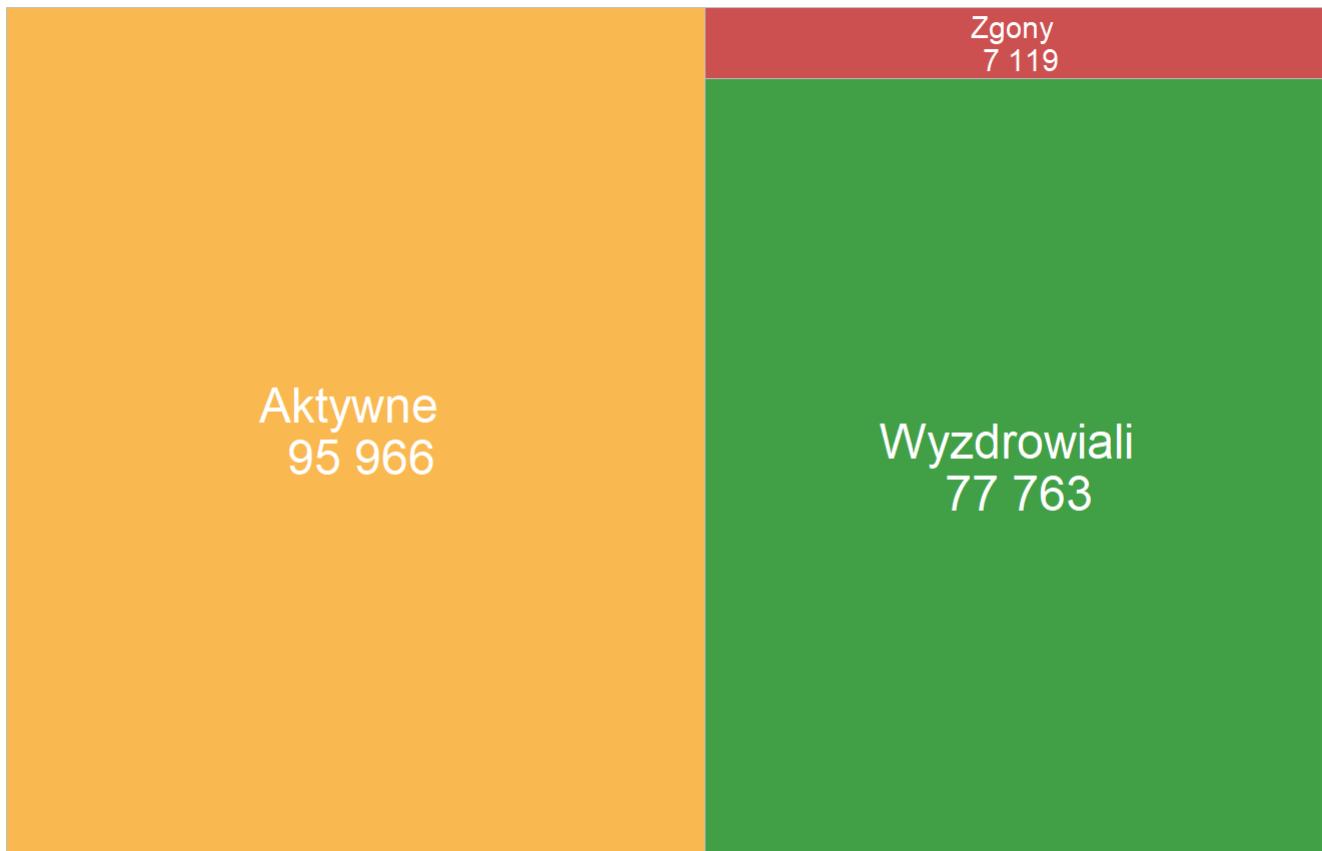
temp <- df_final %>%
  summarise(
    Aktywne = sum(Total_stillSick, na.rm = TRUE),
    Wyzdrowiali = sum(Total_recovered, na.rm = TRUE),
    Zgony = sum(Total_death, na.rm = TRUE)
  ) %>%
  pivot_longer(cols = everything(), names_to = "variable", values_to = "value") %>%
  mutate(label = paste(variable, "\n", format(value, big.mark = " ", scientific = FALSE)))

custom_colors <- c("Aktywne" = "#febd52", "Zgony" = "#d05050", "Wyzdrowiali" = "#43a047")

ggplot(temp, aes(area = value, fill = variable, label = label)) +
  geom_treemap() +
  geom_treemap_text(
    colour = "white",
    place = "centre",
    grow = FALSE,
    reflow = TRUE
  ) +
  scale_fill_manual(values = custom_colors) +
  theme(legend.position = "none",
        plot.title = element_text(hjust = 0.5)) +
  labs(title = "Suma zakażeń, śmierci i uleczeń spowodowanych przez\nCOVID-19 od 11 lutego do 22 marca 2020 roku")

```

Suma zakażeń, śmierci i uleczeń spowodowanych przez
COVID-19 od 11 lutego do 22 marca 2020 roku



Wizualizacja przedstawia strukturę globalnego bilansu pandemii w okresie od 11 lutego do 22 marca 2020 roku. Największą powierzchnię zajmują przypadki aktywne (ponad 95 tysięcy), co wskazuje na dynamiczny rozwój epidemii w tym czasie, przewyższając liczbę osób wyzdrowiałych (blisko 78 tysięcy). Zgony stanowią najmniejszy odsetek wszystkich odnotowanych zdarzeń, obejmując nieco ponad 7 tysięcy przypadków.

```

matter_colors <- c(
  "#f0fa92", "#f8e97a", "#fc3d363", "#febd52", "#fea848",
  "#fb9342", "#f48043", "#ea6e46", "#de5e4b", "#d05050",
  "#c04455", "#af3a58", "#9d325a", "#8b2b59", "#782556",
  "#661f52", "#541a4c", "#431545", "#32103d", "#220b34"
)

plot_map <- function(df, col, title = col, legend_title = col) {
  df_filtered <- df[df[[col]] >= 0, ]

  fig <- plot_geo(df_filtered) %>%
    add_trace(
      z = df_filtered[[col]],
      locations = df_filtered$Iso3,
      locationmode = "ISO-3",
      colors = matter_colors,
      text = paste(df_filtered$Country.Region, "\n", legend_title, ":", df_filtered[[col]]),
      hovertemplate = "%{text}<extra></extra>",
      marker = list(line = list(color = toRGB("grey"), width = 0.5)),
      showlegend = FALSE
    ) %>%
    add_trace(
      type = "scattergeo",
      mode = "markers",
      lat = NA,
      lon = NA,
      marker = list(color = toRGB("gray90"), size = 10, symbol = "square"),
      name = "Brak danych",
      showlegend = TRUE
    ) %>%
    colorbar(title = legend_title) %>%
    layout(
      title = title,
      geo = list(
        showframe = TRUE,
        framecolor = toRGB("grey"),
        showcoastlines = TRUE,
        projection = list(type = 'mercator'),
        showland = TRUE,
        landcolor = toRGB("gray90"),
        showcountries = TRUE,
        countrycolor = toRGB("white")
      )
    )
}

return(fig)
}

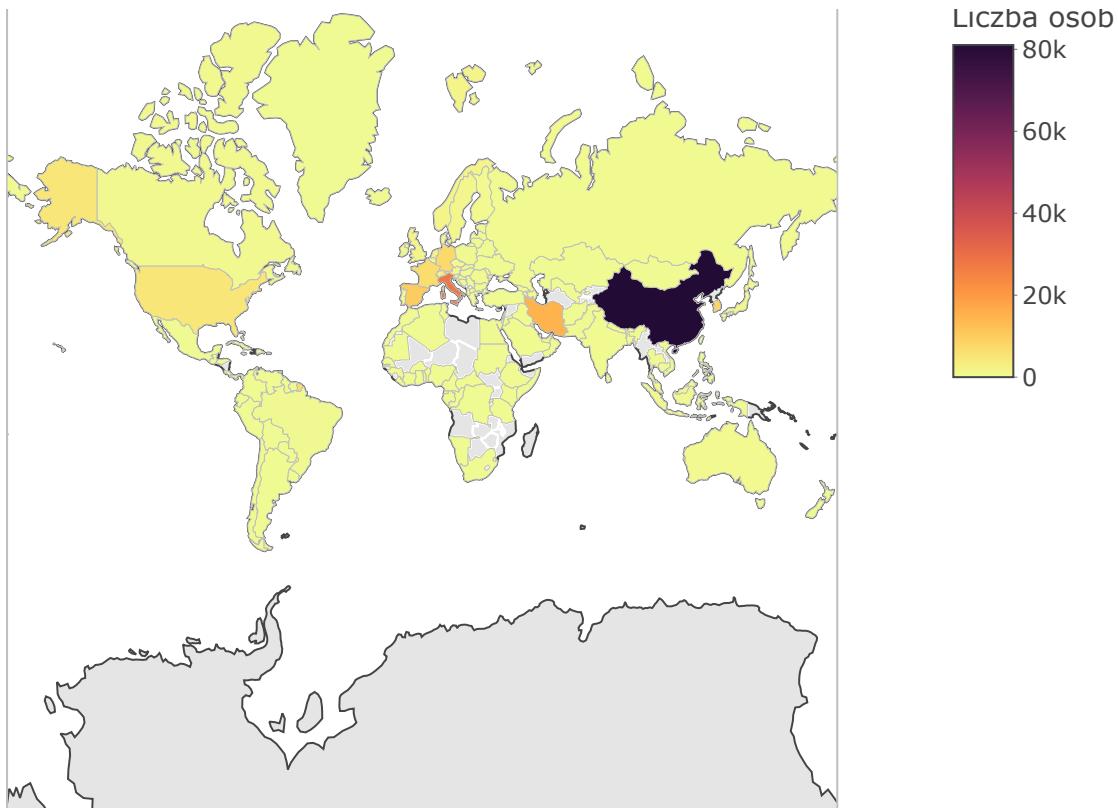
```

```

plot_map(df_final,
  col = 'Total_confirmed',
  title = "Całkowita liczba potwierdzonych przypadków",
  legend_title = "Liczba osób")

```

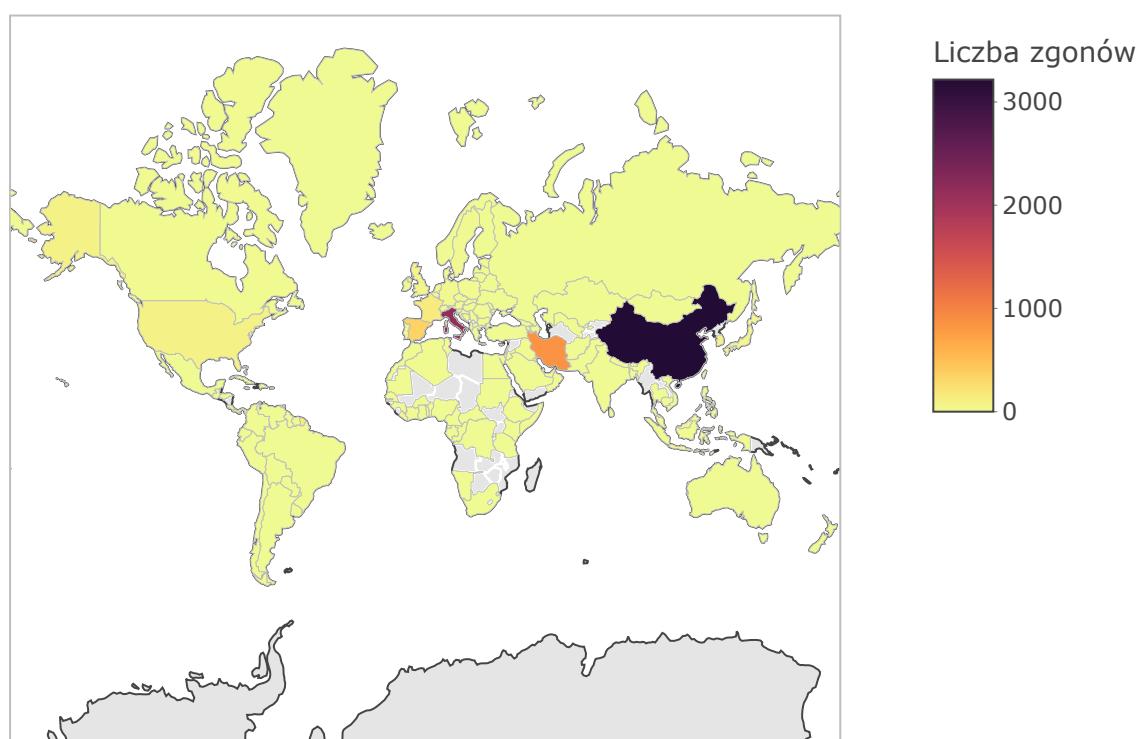
Całkowita liczba potwierdzonych przypadków



Mapa świata obrazuje przestrzenny rozkład całkowitej liczby potwierdzonych przypadków COVID-19, gdzie intensywność koloru odpowiada skali zachorowań. Zdecydowanie wyróżniają się Chiny jako główne ognisko pandemii z liczbą zakażeń zbliżoną do 80 tysięcy, podczas gdy w pozostałych regionach, z wyjątkiem widocznych ognisk we Włoszech i Iranie, liczba odnotowanych przypadków jest wielokrotnie niższa.

```
plot_map(df_final,
  col = 'Total_death',
  title = "Całkowita liczba zgonów",
  legend_title = "Liczba zgonów")
```

Całkowita liczba zgonów

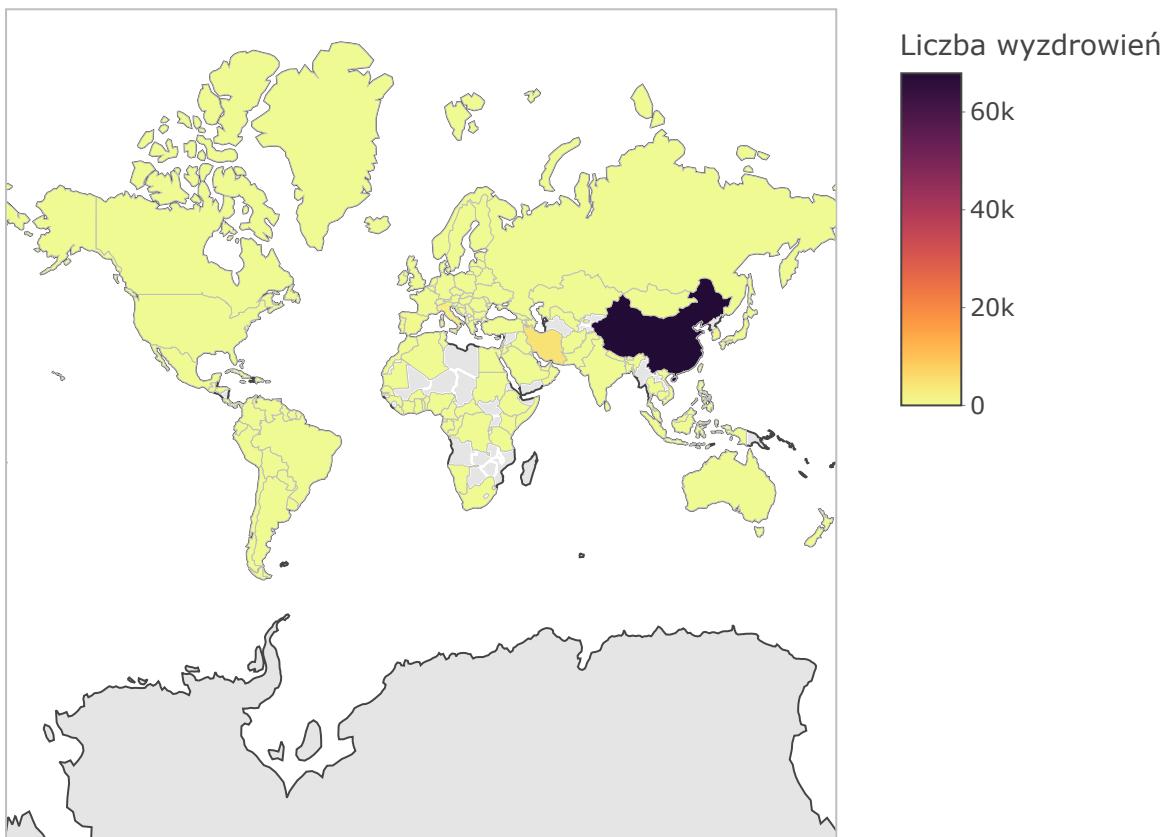




Mapa prezentuje globalne rozmieszczenie liczby zgonów spowodowanych przez COVID-19, gdzie najciemniejszy kolor wskazuje regiony o najwyższej śmiertelności. Podobnie jak w przypadku liczby zakażeń, tragiczny bilans dominuje w Chinach, jednak wyraźnie widoczne są również nowe, intensywne ogniska epidemii we Włoszech oraz w Iranie, które znaczco wyróżniają się na tle pozostałych państw.

```
plot_map(df_final,
         col = 'Total_recovered',
         title = "Całkowita liczba wyzdrowień",
         legend_title = "Liczba wyzdrowień")
```

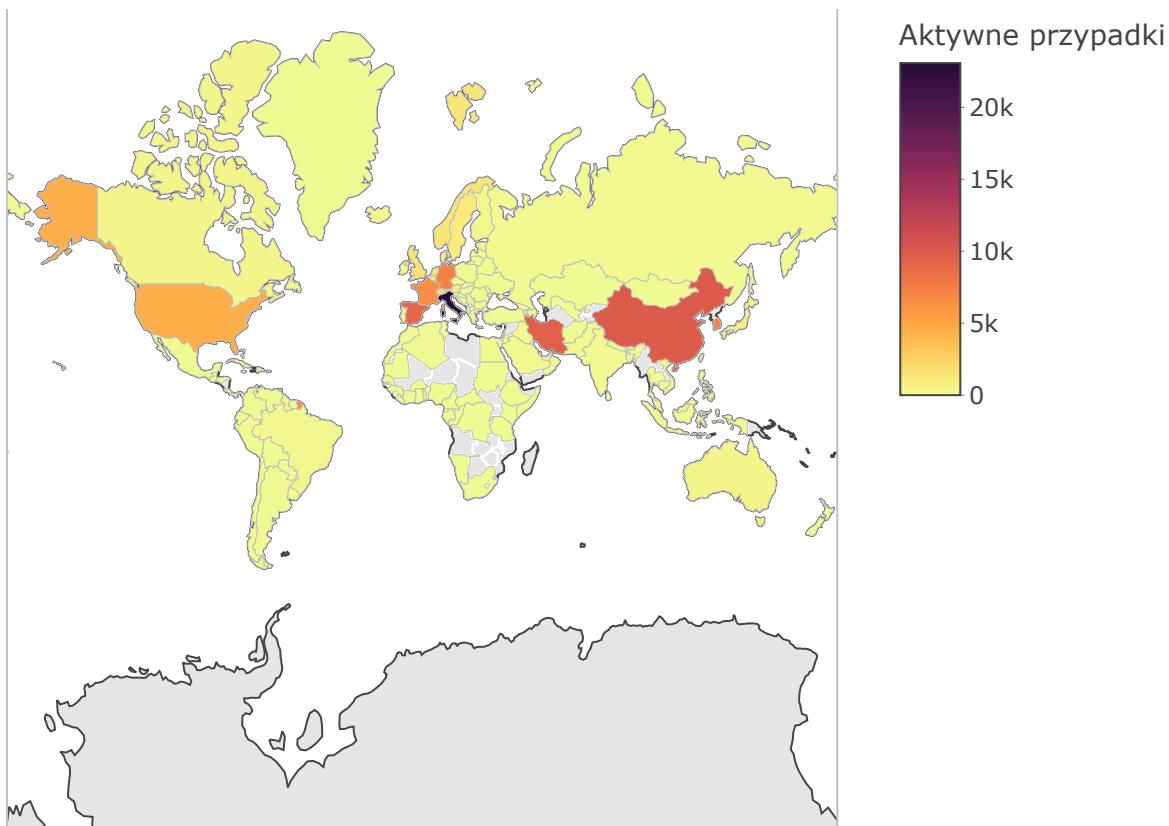
Całkowita liczba wyzdrowień



Mapa obrazuje globalny rozkład liczby wyzdrowień, gdzie – analogicznie do liczby zakażeń – dominują Chiny, co wynika z wcześniejszego rozwoju epidemii w tym regionie. Ciemny kolor wskazuje na wysoką liczbę pacjentów, którzy zwalczyl infekcję (skala do 60 tys.), podczas gdy w pozostałych częściach świata, w tym w rozwijających się ogniskach w Europie i na Bliskim Wschodzie, liczba odnotowanych wyzdrowień jest na tym etapie jeszcze stosunkowo niska.

```
plot_map(df_final,
         col = 'Total_stillSick',
         title = "Aktywne przypadki (wciąż chorzy)",
         legend_title = "Aktywne przypadki")
```

Aktywne przypadki (wciąż chorzy)



Mapa prezentująca liczbę aktywnych przypadków na dzień 22 marca 2020 roku ukazuje istotną zmianę w dynamice przebiegu pandemii. W przeciwieństwie do mapy zakażeń całkowitych, tutaj głównym epicentrum są Włochy (oznaczone najciemniejszym kolorem), co wskazuje na gwałtowny przyrost chorych w Europie, podczas gdy w Chinach liczba aktywnych infekcji wyraźnie spadła na skutek dużej liczby wyzdrowień. Widoczny jest również rosnący trend zachorowań w Stanach Zjednoczonych, Iranie, Hiszpanii, Francji i Niemczech.

```

iso_map <- df_final %>%
  select(Country.Region, Iso3) %>%
  distinct()

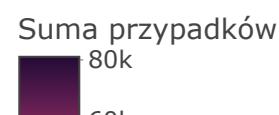
df_anim <- df %>%
  filter(type == "confirmed") %>%
  group_by(Country.Region, date) %>%
  summarise(day_cases = sum(cases, na.rm = TRUE), .groups = "drop") %>%
  left_join(iso_map, by = "Country.Region") %>%
  filter(!is.na(Iso3)) %>%
  arrange(Country.Region, date) %>%
  group_by(Country.Region) %>%
  mutate(
    total_cases = cumsum(day_cases),
    date_str = as.character(date)
  ) %>%
  ungroup()

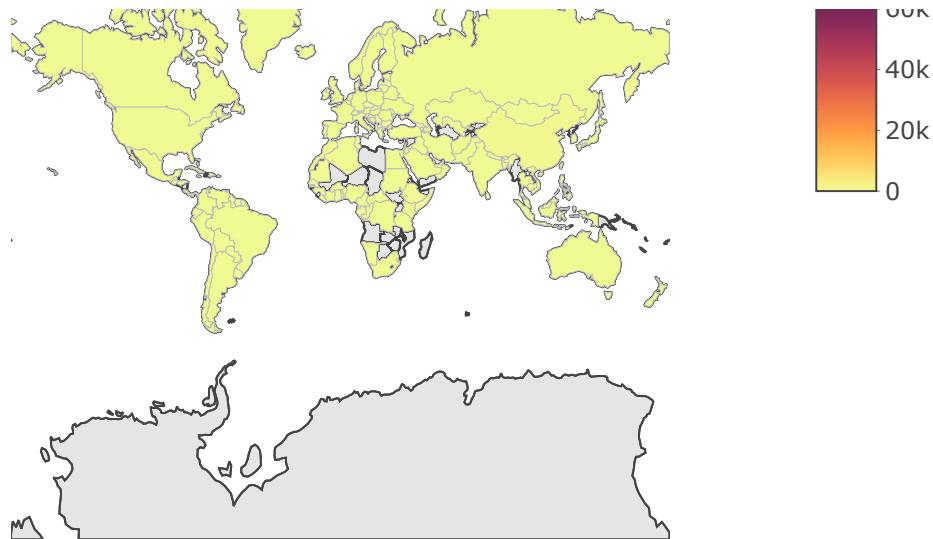
max_total_cases <- max(df_anim$total_cases, na.rm = TRUE)

plot_geo(df_anim) %>%
  add_trace(
    z = ~total_cases,
    locations = ~Iso3,
    locationmode = "ISO-3",
    frame = ~date_str,
    colors = matter_colors,
    text = ~paste(Country.Region, "\nData:", date_str, "\nSuma narastająco:", total_cases),
    hoverinfo = "text",
    marker = list(line = list(color = toRGB("grey"), width = 0.5)),
    zmin = 0,
    zmax = max_total_cases
  ) %>%
  colorbar(title = "Suma przypadków") %>%
  layout(
    title = "Rozwój pandemii",
    geo = list(
      showframe = FALSE,
      showcoastlines = TRUE,
      projection = list(type = 'mercator'),
      showland = TRUE,
      landcolor = toRGB("gray90"),
      showcountries = TRUE
    )
  ) %>%
  animation_opts(
    frame = 100,
    transition = 0,
    redraw = FALSE
  )
)

```

Rozwój pandemii





date_str: 2020-01-22



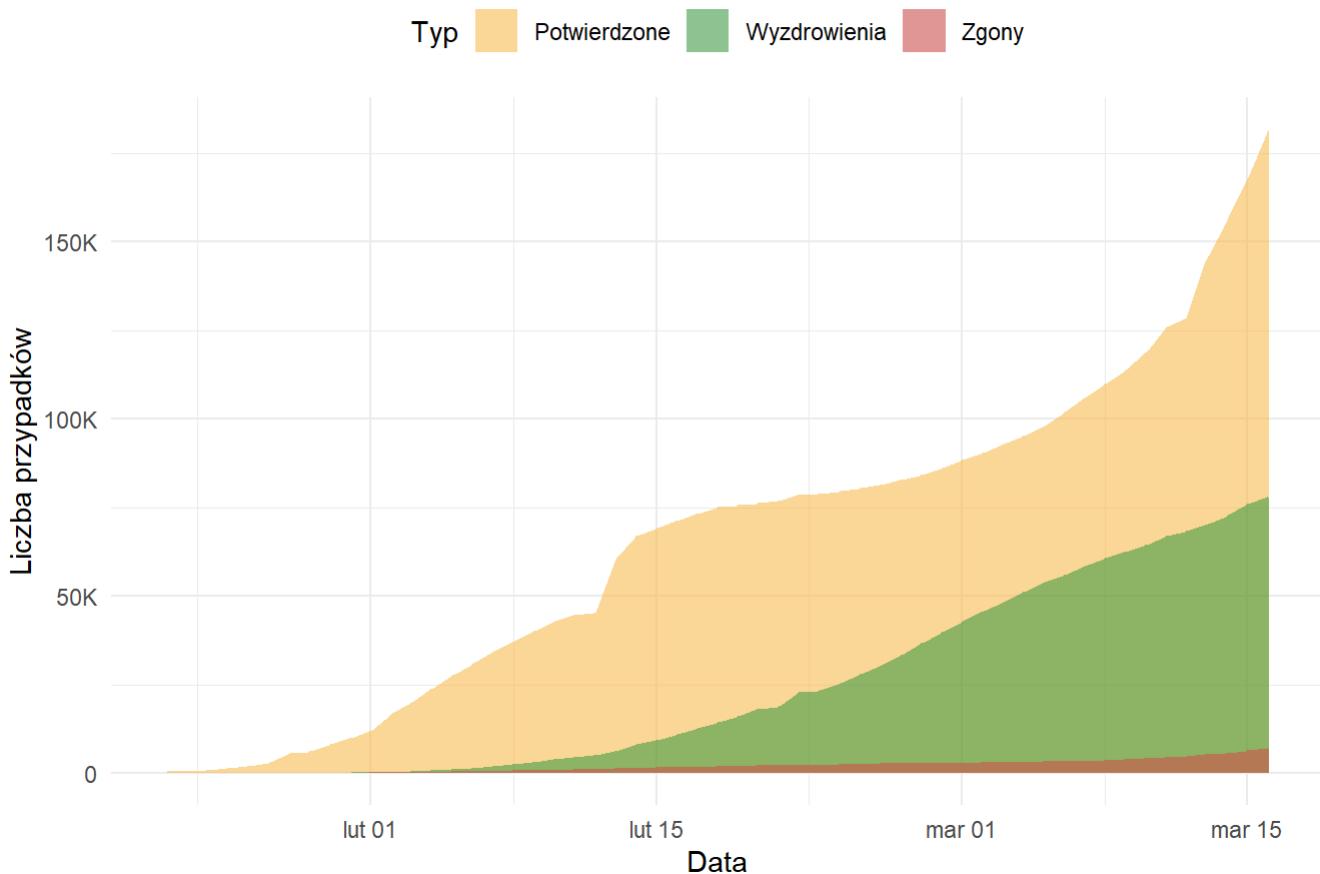
Animowana wizualizacja obrazuje dynamikę rozprzestrzeniania się zakażeń na świecie w okresie od 22 stycznia do 16 marca 2020 roku. Początkowym epicentrum pandemii były Chiny, gdzie do połowy lutego potwierdzono ponad 70 tysięcy przypadków, podczas gdy w innych krajach (np. Japonii czy Korei Południowej) liczby te pozostawały marginalne. Sytuacja uległa istotnej zmianie na przełomie lutego i marca, kiedy to nastąpił gwałtowny wzrost zachorowań w nowych ogniskach epidemii: we Włoszech oraz Iranie.

```
data_cumulative <- df %>%
  filter(type %in% c("confirmed", "death", "recovered")) %>%
  group_by(date, type) %>%
  summarise(daily_cases = sum(cases, na.rm = TRUE), .groups = "drop") %>%
  arrange(date) %>%
  group_by(type) %>%
  mutate(total = cumsum(daily_cases)) %>%
  ungroup() %>%
  mutate(
    type_label = case_when(
      type == "confirmed" ~ "Potwierdzone",
      type == "recovered" ~ "Wyzdrowienia",
      type == "death" ~ "Zgony"
    ),
    type_label = factor(type_label, levels = c("Potwierdzone", "Wyzdrowienia", "Zgony"))
  )

custom_colors <- c(
  "Potwierdzone" = "#feb5d2",
  "Wyzdrowienia" = "#43a047",
  "Zgony"         = "#d05050"
)

ggplot(data_cumulative, aes(x = date, y = total, fill = type_label)) +
  geom_area(alpha = 0.6, position = "identity") +
  scale_fill_manual(values = custom_colors) +
  scale_y_continuous(labels = scales::label_number(scale_cut = scales::cut_short_scale())) +
  theme_minimal() +
  labs(
    title = "Łączna liczba przypadków w czasie",
    x = "Data",
    y = "Liczba przypadków",
    fill = "Typ"
  ) +
  theme(legend.position = "top")
```

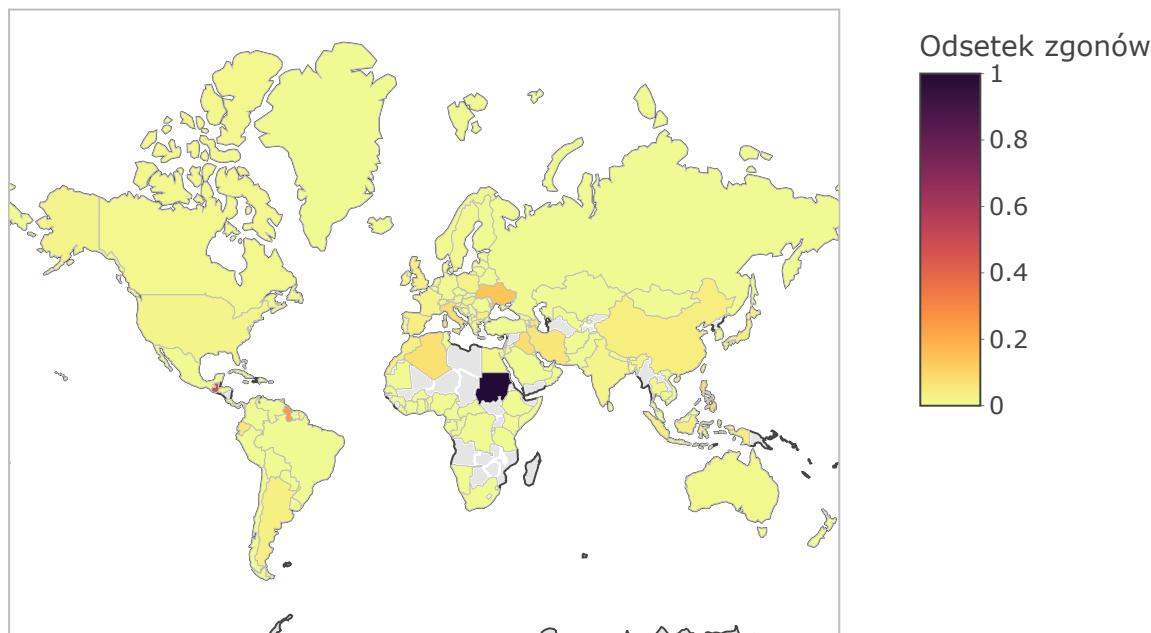
Łączna liczba przypadków w czasie

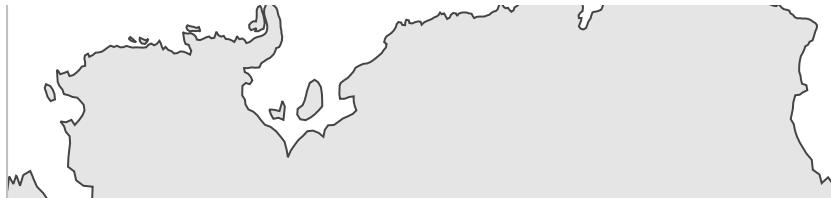


Wykres warstwowy obrazuje kumulacyjny przyrost liczby przypadków COVID-19 w czasie, z podziałem na potwierdzone zakażenia, wyzdrowienia oraz zgony. Widoczna jest stała tendencja wzrostowa całkowitej liczby chorych (obszar żółty), która ulega wyraźnemu przyspieszeniu w marcu, co świadczy o globalnej ekspansji wirusa. Równolegle obserwowany jest systematyczny wzrost liczby osób, które zwalczyły infekcję (obszar zielony), podczas gdy odsetek zgonów (wąski pasek czerwony) pozostaje stabilny i niski w stosunku do ogólnej liczby zachorowań.

```
plot_map(df_final,
          col = 'Mortality_Rate',
          title = "Wskaźnik śmiertelności wśród zarażonych na COVID-19",
          legend_title = "Odsetek zgonów")
```

Wskaźnik śmiertelności wśród zarażonych na COVID-19

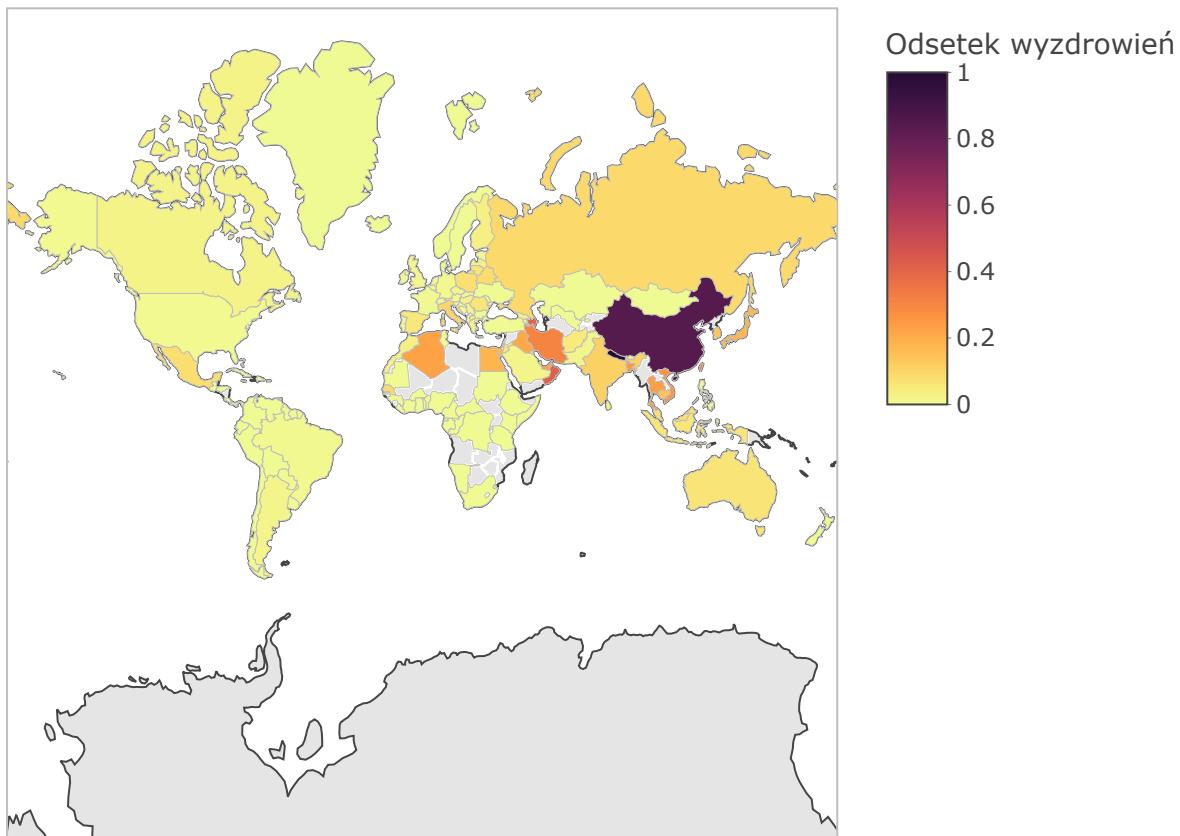




Mapa przedstawia przestrzenny rozkład wskaźnika śmiertelności (CFR), obliczonego jako stosunek liczby zgonów do liczby potwierdzonych przypadków. W przeważającej części świata współczynnik ten utrzymuje się na niskim poziomie, co reprezentują jasne odcienie, natomiast wyraźnym wyjątkiem jest Sudan oznaczony najciemniejszą barwą, co wskazuje na drastycznie wysoki odsetek zgonów w stosunku do wykrytych infekcji (co często wynika z małej liczby testów na wczesnym etapie epidemii).

```
plot_map(df_final,
  col = 'Recovery_Rate',
  title = "Wskaźnik wyzdrowień wśród zarażonych na COVID-19",
  legend_title = "Odsetek wyzdrowień")
```

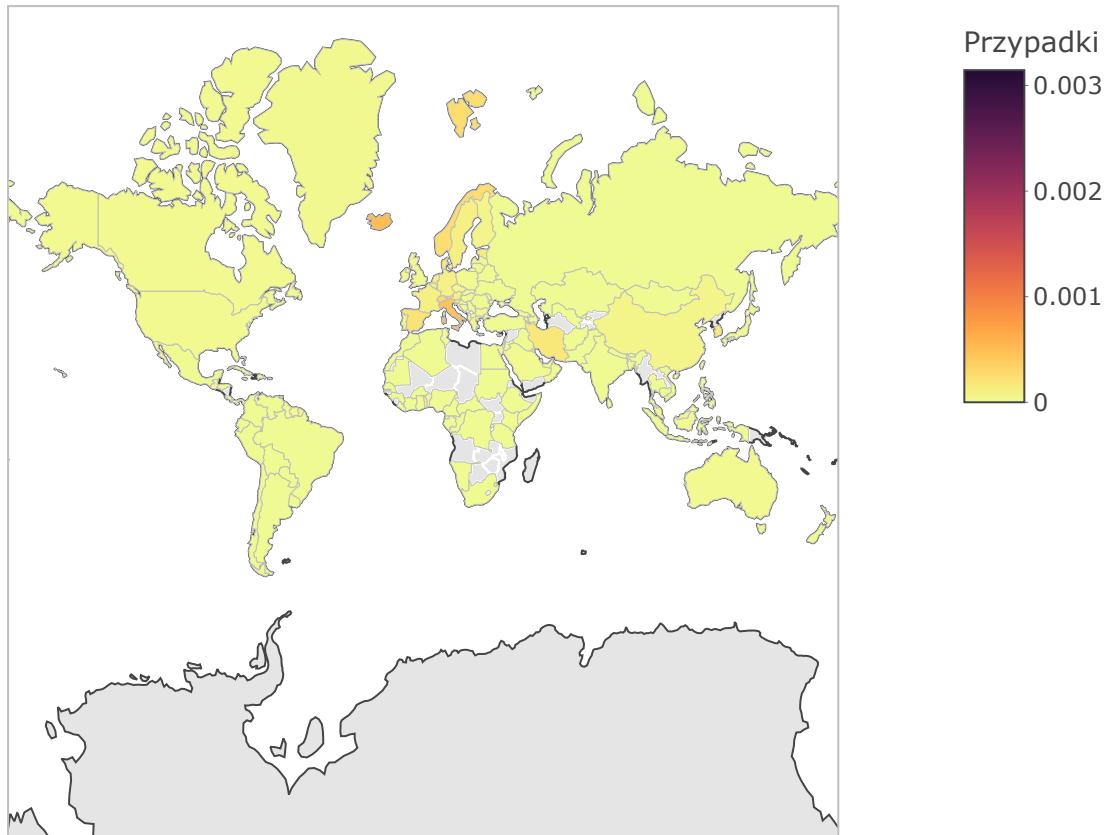
Wskaźnik wyzdrowień wśród zarażonych na COVID-19



Mapa prezentuje przestrzenny rozkład wskaźnika wyzdrowień, zdefiniowanego jako udział osób wyleczonych w całkowitej liczbie potwierdzonych przypadków. Najwyższe wartości (ciemny kolor) obserwowane są w Chinach, co wynika z faktu, że jako pierwsze ognisko pandemii, kraj ten zdążył już odnotować dużą liczbę zamkniętych, wyleczonych spraw. W pozostałych regionach, gdzie pandemia była w fazie początkowego rozwoju (np. Europa, USA), wskaźnik ten pozostaje niski (jasne kolory), co oznacza, że większość zdiagnozowanych pacjentów była wciąż w trakcie choroby.

```
plot_map(df_final,
  col = 'Incidence_Rate',
  title = "Wskaźnik zachorowalności",
  legend_title = "Przypadki")
```

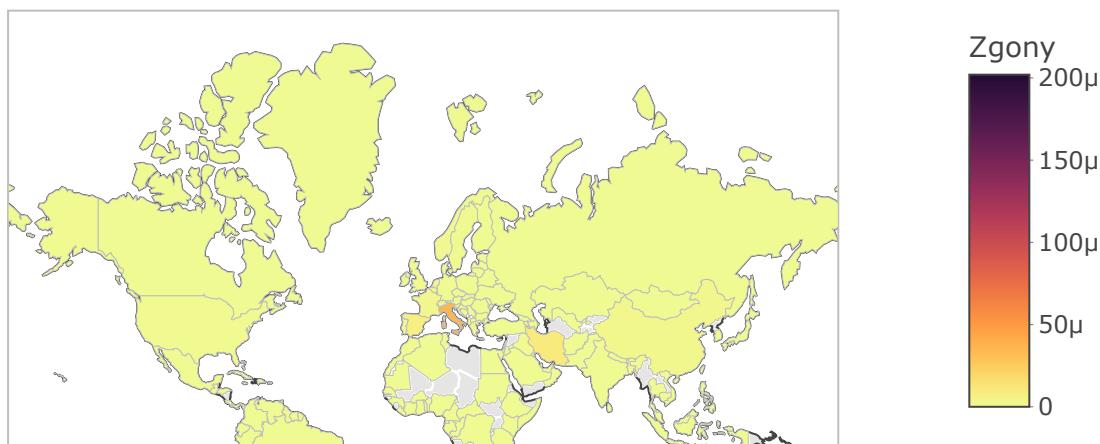
Wskaźnik zachorowalności



Mapa obrazuje rozkład wskaźnika zachorowalności, który normalizuje liczbę zakażeń względem populacji danego kraju. Zastosowanie tej miary zmienia perspektywę oceny zagrożenia – o ile w liczbach bezwzględnych dominowały Chiny, to w ujęciu relatywnym (na mieszkańca) najciemniejszymi kolorami wyróżniają się Włochy oraz Islandia, co wskazuje na najwyższą penetrację wirusa w strukturze tych społeczeństw w analizowanym okresie.

```
plot_map(df_final,
  col = 'Mortality_per_capita',
  title = "Wskaźnik zgonów na mieszkańca",
  legend_title = "Zgony")
```

Wskaźnik zgonów na mieszkańca

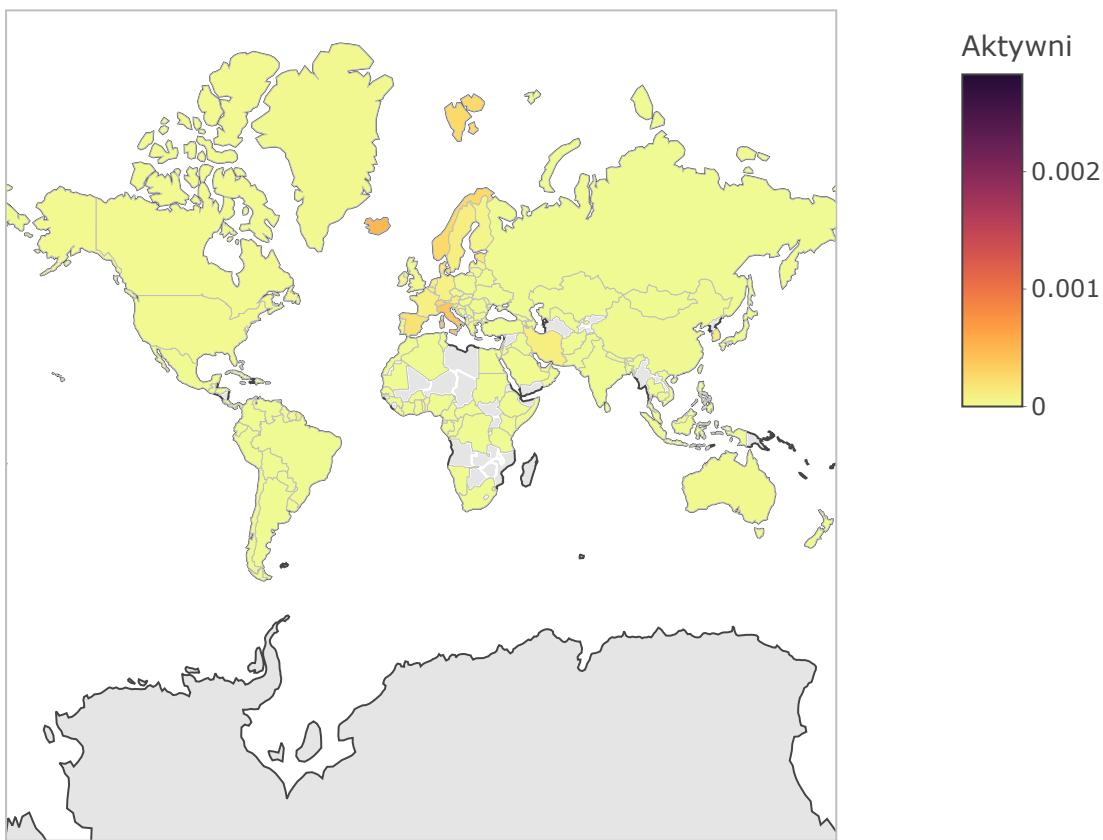




Mapa prezentuje wskaźnik umieralności w przeliczeniu na jednego mieszkańca, co pozwala ocenić rzeczywiste obciążenie demograficzne w poszczególnych państwach. W tym ujęciu punkt ciężkości pandemii wyraźnie przesuwa się do Europy – najciemniejszym kolorem wyróżniają się Włochy, gdzie relacja liczby zgonów do wielkości populacji była w badanym okresie najwyższa na świecie, znacznie przewyższając wartości notowane w Chinach.

```
plot_map(df_final,
          col = 'StillSick_per_capita',
          title = "Wskaźnik aktywnych przypadków na mieszkańca",
          legend_title = "Aktywni")
```

Wskaźnik aktywnych przypadków na mieszkańca



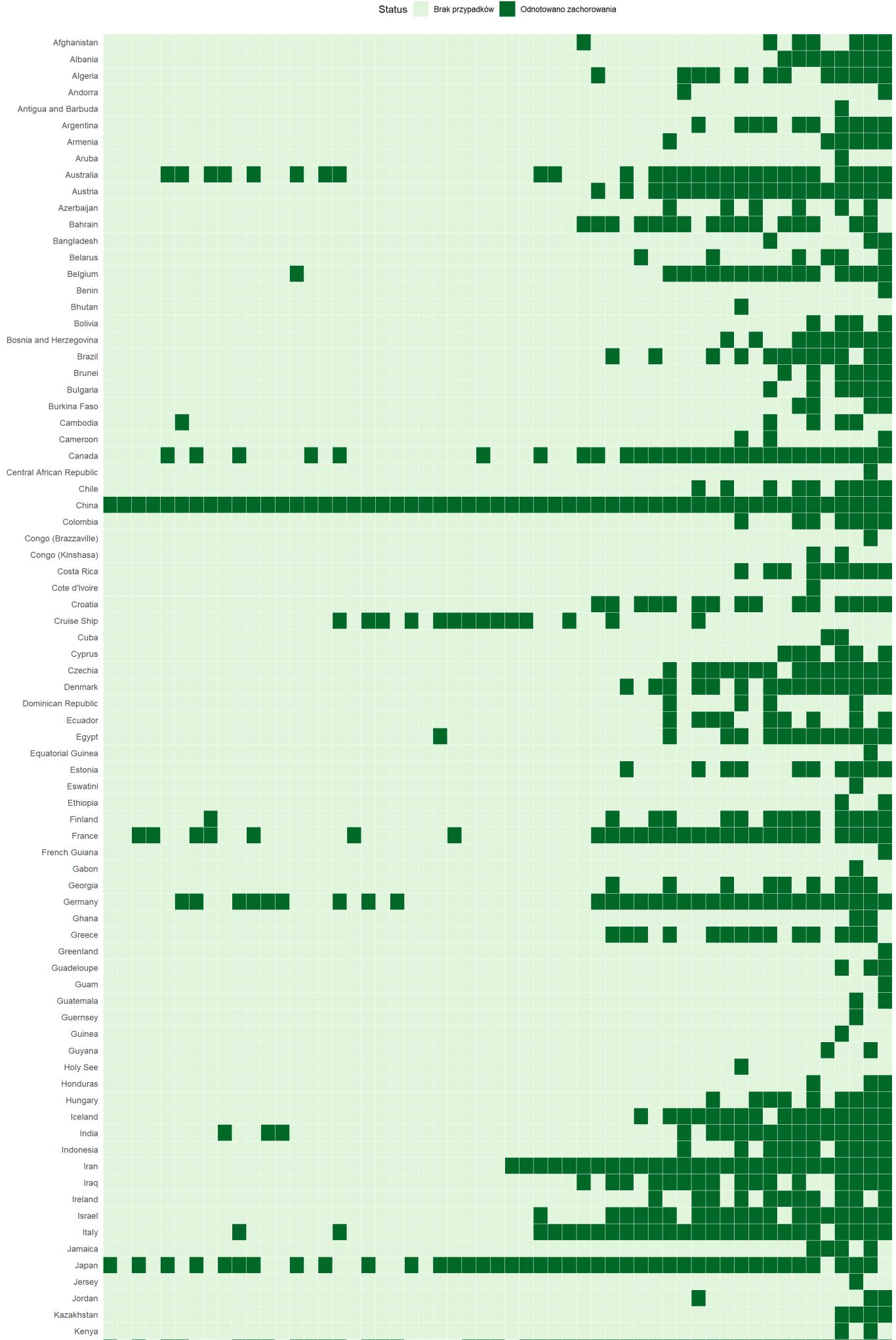
Mapa przedstawia rozkład wskaźnika aktywnych przypadków w przeliczeniu na populację, co pozwala zidentyfikować regiony o najwyższym bieżącym obciążeniu epidemiologicznym. W tym ujęciu epicentrum pandemii znajduje się w Europie – najciemniejszymi barwami wyróżniają się Włochy oraz Islandia, gdzie odsetek osób wciąż chorujących względem liczby mieszkańców był w analizowanym momencie najwyższy, podczas gdy w Chinach wskaźnik ten spadł do niskiego poziomu dzięki dużej liczbie wyzdrowień.

```
heatmap_data <- df %>%
  filter(type == "confirmed") %>%
  group_by(Country.Region, date) %>%
  summarise(daily_cases = sum(cases, na.rm = TRUE), .groups = "drop") %>%
  mutate(
    status = ifelse(daily_cases > 0, "1", "0"),
    status = factor(status, levels = c("0", "1"))
  )

heatmap_colors <- c("0" = "#e5f5e0", "1" = "#006d2c")

ggplot(heatmap_data, aes(x = date, y = Country.Region, fill = status)) +
  geom_tile(color = "white", linewidth = 0.05) +
  scale_fill_manual(
    values = heatmap_colors,
    labels = c("Brak przypadków", "Odnotowano zachorowania"),
    name = "Status"
  ) +
  scale_x_date(
    date_labels = "%d %b",
    date_breaks = "1 week",
    expand = c(0, 0)
  ) +
  scale_y_discrete(limits = rev) +
  labs(
    title = "Analiza występowania zakażeń COVID-19 wg krajów w pierwszych 2 miesiącach pandemii",
    x = "Data",
    y = "Kraj"
  ) +
  theme_minimal() +
  theme(
    axis.text.y = element_text(size = 9),
    axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1, size = 10),
    legend.position = "top",
    panel.grid = element_blank(),
    plot.title = element_text(size = 20),
    plot.subtitle = element_text(size = 14)
  )
```

Analiza występowania zakażeń COVID-19 wg krajów w pierwszych 2 miesiącach pandemii





Wykres w formie binarnej mapy ciepła prezentuje chronologię występowania nowych przypadków COVID-19, gdzie ciemnozielony kolor oznacza odnotowanie zakażenia w danym dniu. Wizualizacja kontrastuje ciągły przebieg

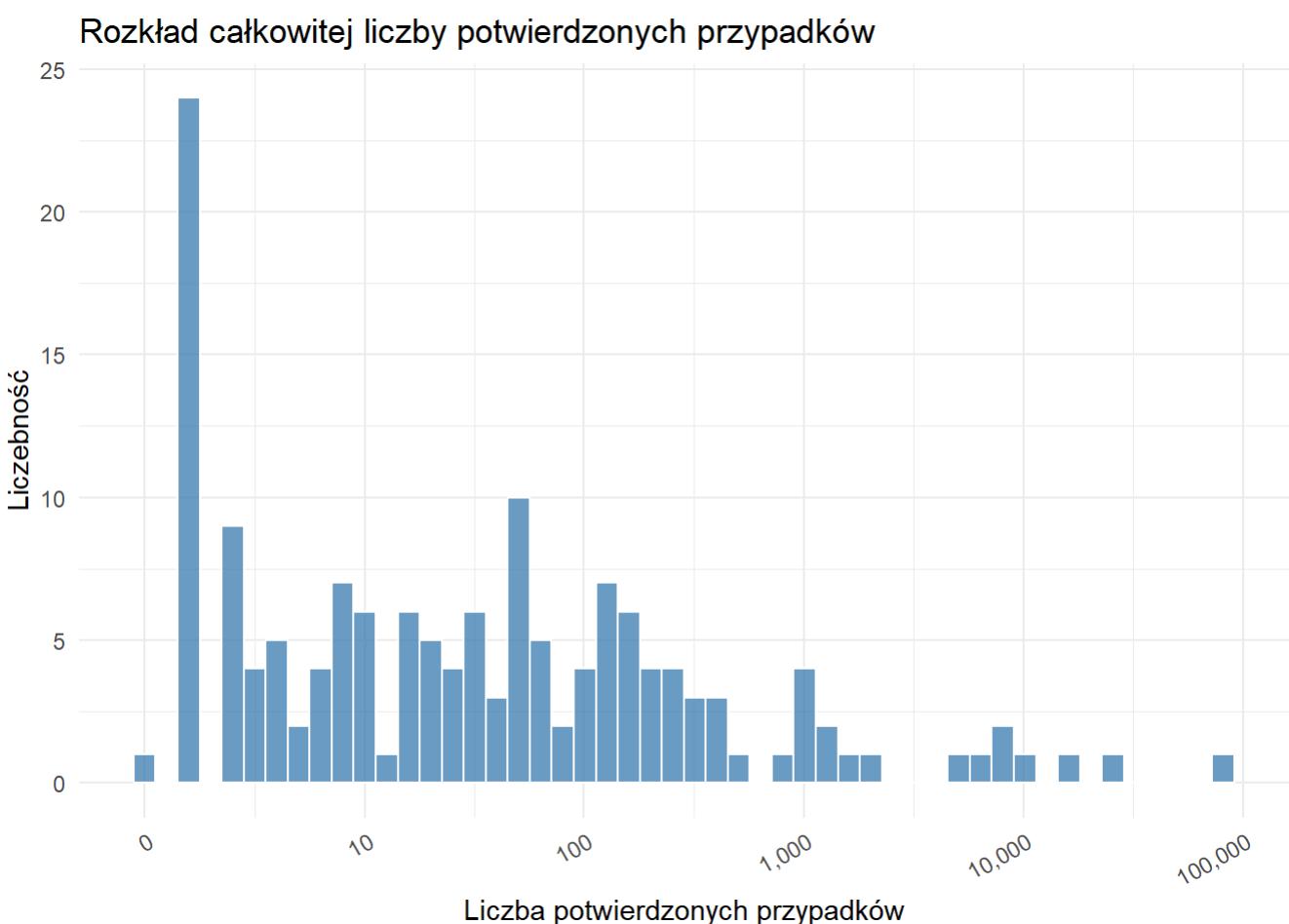
epidemii w Chinach z resztą świata, gdzie początkowo widoczne są jedynie sporadyczne ogniska, które w połowie marca gwałtownie ewoluują w systematyczne, codzienne zachorowania w niemal wszystkich analizowanych państwach, obrazując globalną ekspansję pandemii.

```
summary(df_final)
```

```
##   Country.Region          Lat        Long  Total_confirmed
##   Length:153      Min.   :-40.90   Min.   :-102.55    Min.   :  0
##   Class  :character  1st Qu.: 7.54   1st Qu.: -9.43   1st Qu.:  4
##   Mode   :character  Median : 24.00   Median : 18.49   Median : 25
##                   Mean   : 23.14   Mean   : 15.02   Mean   :1182
##                   3rd Qu.: 42.60   3rd Qu.: 45.04   3rd Qu.: 142
##                   Max.   : 71.71   Max.   :174.89   Max.   :81033
##   Total_death      Total_recovered       Iso3           Population
##   Min.   : 0.00   Min.   : 0.0   Length:153      Min.   :3.465e+04
##   1st Qu.: 0.00   1st Qu.: 0.0   Class  :character  1st Qu.:2.650e+06
##   Median : 0.00   Median : 0.0   Mode   :character  Median :9.435e+06
##   Mean   : 46.53   Mean   : 508.3          Median :4.770e+07
##   3rd Qu.: 1.00   3rd Qu.: 3.0          Mean   :3.344e+07
##   Max.   :3217.00  Max.   :67910.0          Max.   :1.424e+09
##   Density         MedianAge      Total_stillSick Mortality_Rate
##   Min.   :1.366e-01  Min.   :14.37   Min.   :  0.0  Min.   :0.000000
##   1st Qu.:4.463e+01  1st Qu.:24.84   1st Qu.:  3.0  1st Qu.:0.000000
##   Median :9.792e+01  Median :31.21   Median : 25.0  Median :0.000000
##   Mean   :4.145e+02  Mean   :31.73   Mean   : 627.2  Mean   :0.020470
##   3rd Qu.:2.402e+02  3rd Qu.:39.80   3rd Qu.: 121.0  3rd Qu.:0.006803
##   Max.   :2.558e+04  Max.   :54.64   Max.   :23073.0  Max.   :1.000000
##   Recovery_Rate    Active_Rate     Incidence_Rate Mortality_per_capita
##   Min.   :0.000000  Min.   :0.0000  Min.   :0.000e+00  Min.   :0.000e+00
##   1st Qu.:0.000000  1st Qu.:0.9310  1st Qu.:6.142e-07  1st Qu.:0.000e+00
##   Median :0.000000  Median :0.9918  Median :3.992e-06  Median :0.000e+00
##   Mean   :0.05612   Mean   :0.9169  Mean   :5.300e-05  Mean   :1.818e-06
##   3rd Qu.:0.03670   3rd Qu.:1.0000  3rd Qu.:2.550e-05  3rd Qu.:5.847e-08
##   Max.   :1.000000  Max.   :1.0000  Max.   :3.145e-03  Max.   :2.020e-04
##   StillSick_per_capita
##   Min.   :0.000e+00
##   1st Qu.:5.987e-07
##   Median :3.714e-06
##   Mean   :4.851e-05
##   3rd Qu.:1.944e-05
##   Max.   :2.828e-03
```

Analiza statystyk opisowych dla 153 krajów rozpoczyna się od wskaźników określających bezwzględną skalę pandemii, obejmujących całkowitą liczbę potwierdzonych zakażeń, zgonów, wyzdrowień oraz aktywnych przypadków. Rozkład tych danych charakteryzuje się silną asymetrią prawostronną, gdzie średnie wartości są wielokrotnie wyższe od median. Przykładowo, mimo średniej wynoszącej ponad 1100 infekcji, typowy kraj odnotował ich zaledwie 25, a w przypadku ofiar śmiertelnych i ozdrowieńców w ponad połowie państw wartości te wciąż wynoszą zero. Tło demograficzne analizy tworzą dane dotyczące populacji, gęstości zaludnienia i mediany wieku, które ukazują ogromną różnorodność badanych jednostek – począwszy od wielomiliardowych mocarstw po małe terytoria oraz od społeczeństw bardzo młodych (wiek średniorolny 14 lat) do starzejących się (ponad 54 lata). Obraz dopełniają wskaźniki relatywne, takie jak stopa śmiertelności (średnio ok. 2%) i wskaźnik wyzdrowień (średnio 5,6%) oraz współczynniki przeliczone na mieszkańców (zapadalność i umieralność), które przyjmują bardzo niskie wartości liczbowe, co precyzyjnie odzwierciedla wcześniejszy etap rozwoju pandemii w większości regionów świata.

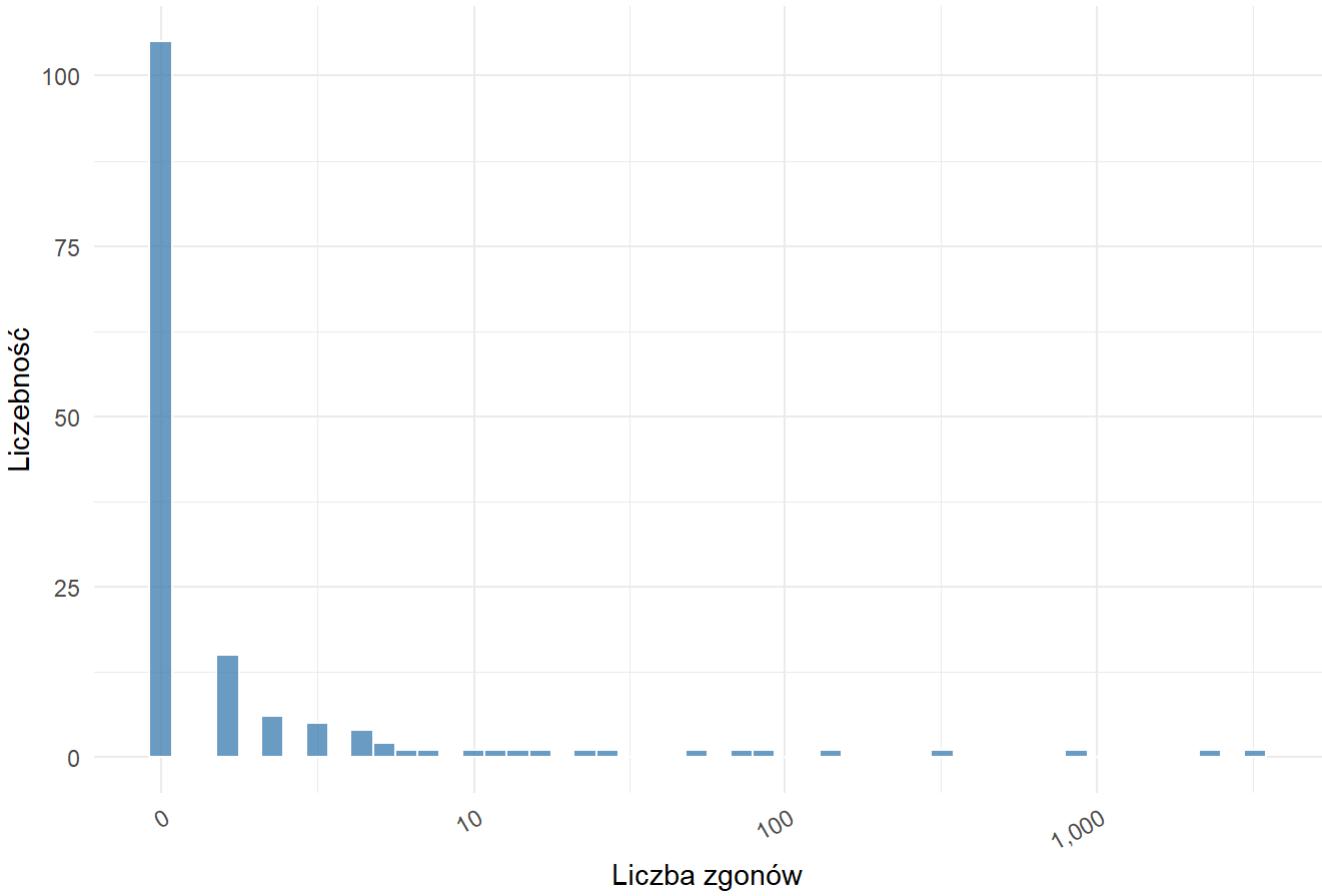
```
ggplot(df_final, aes(x = Total_confirmed)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "white", alpha = 0.8) +
  scale_x_continuous(
    trans = "pseudo_log",
    breaks = c(0, 10, 100, 1000, 10000, 100000, 1000000),
    labels = label_comma()
  ) +
  theme_minimal() +
  labs(
    title = "Rozkład całkowitej liczby potwierdzonych przypadków",
    x = "Liczba potwierdzonych przypadków",
    y = "Liczębność"
  ) +
  theme(
    axis.text.x = element_text(angle = 30, hjust = 1)
  )
)
```



Histogram przedstawia rozkład liczby potwierdzonych zakażeń w poszczególnych państwach, wykorzystując skalę logarytmiczną na osi poziomej w celu czytelnego zobrazowania skrajnych dysproporcji w danych. Zdecydowana większość krajów znajduje się w lewej części wykresu, co oznacza, że w analizowanym okresie najczęściej odnotowywano tam jedynie pojedyncze lub nieliczne przypadki infekcji (poniżej 10). Widoczny jest wyraźny „długi ogon” rozkładu, wskazujący na to, że państwa z masową skalą zachorowań (powyżej 10 tysięcy czy 100 tysięcy przypadków) stanowią rzadkie wyjątki w skali globalnej.

```
ggplot(df_final, aes(x = Total_death)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "white", alpha = 0.8) +
  scale_x_continuous(
    trans = "pseudo_log",
    breaks = c(0, 10, 100, 1000, 10000, 100000, 1000000),
    labels = label_comma()
  ) +
  theme_minimal() +
  labs(
    title = "Rozkład całkowitej liczby zgonów",
    x = "Liczba zgonów",
    y = "Liczebność"
  ) +
  theme(
    axis.text.x = element_text(angle = 30, hjust = 1)
  )
)
```

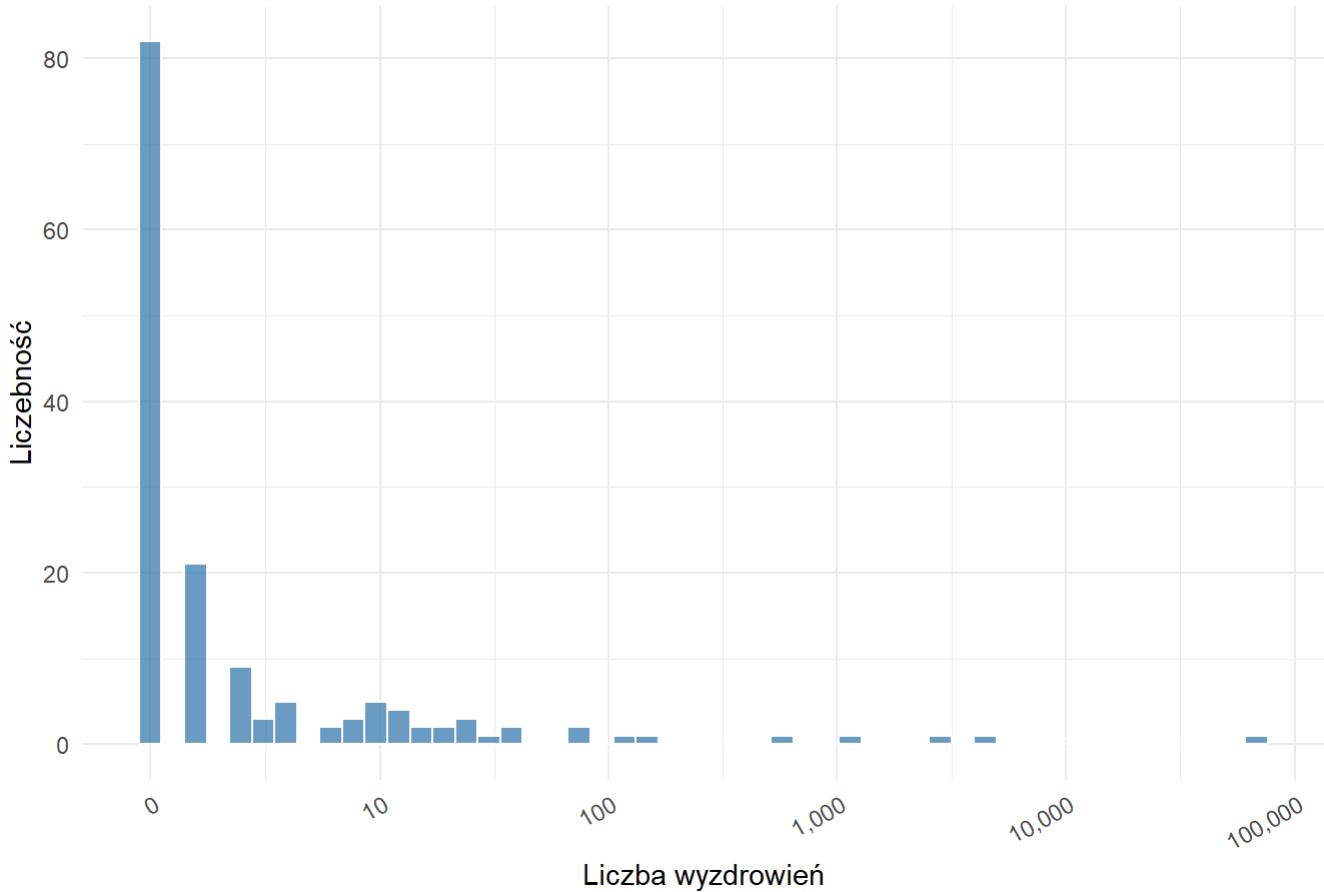
Rozkład całkowitej liczby zgonów



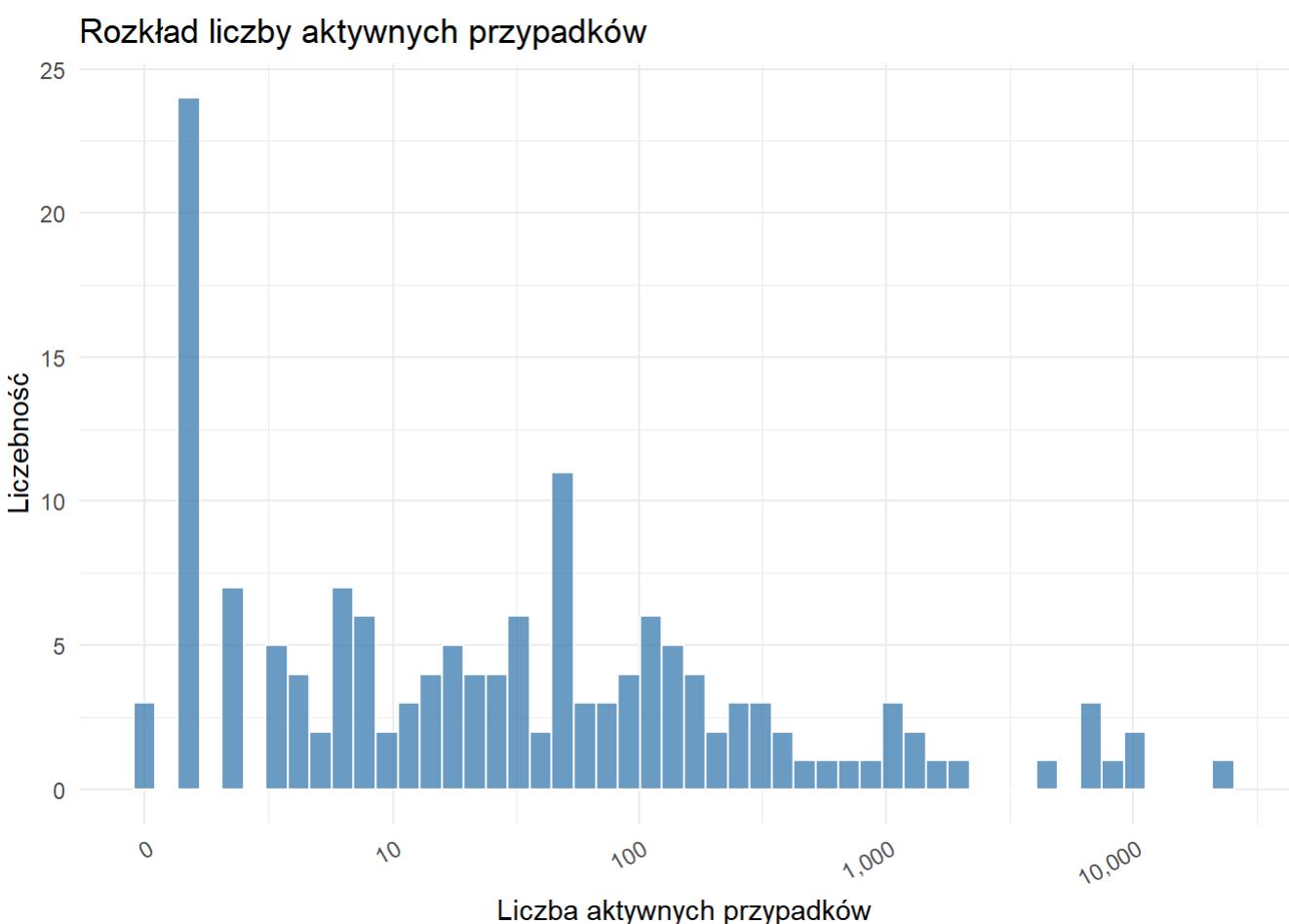
Histogram przedstawia rozkład liczby zgonów w poszczególnych krajach, wykorzystując skalę logarytmiczną do uwzględnienia dużych dysproporcji w danych. Wykres jest zdominowany przez wysoką kolumnę przy wartości zero, co oznacza, że w zdecydowanej większości badanych państw (ponad 100) do tego momentu nie odnotowano żadnej ofiary śmiertelnej. Widoczny po prawej stronie „długi ogon” potwierdza, że kraje z liczbą zgonów przekraczającą 10, 100 czy 1000 stanowiły w tamtym okresie nieliczne, odosobnione przypadki na tle globalnym.

```
ggplot(df_final, aes(x = Total_recovered)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "white", alpha = 0.8) +
  scale_x_continuous(
    trans = "pseudo_log",
    breaks = c(0, 10, 100, 1000, 10000, 100000, 1000000),
    labels = label_comma()
  ) +
  theme_minimal() +
  labs(
    title = "Rozkład całkowitej liczby wyzdrowień",
    x = "Liczba wyzdrowień",
    y = "Liczebność"
  ) +
  theme(
    axis.text.x = element_text(angle = 30, hjust = 1)
  )
)
```

Rozkład całkowitej liczby wyzdrowień

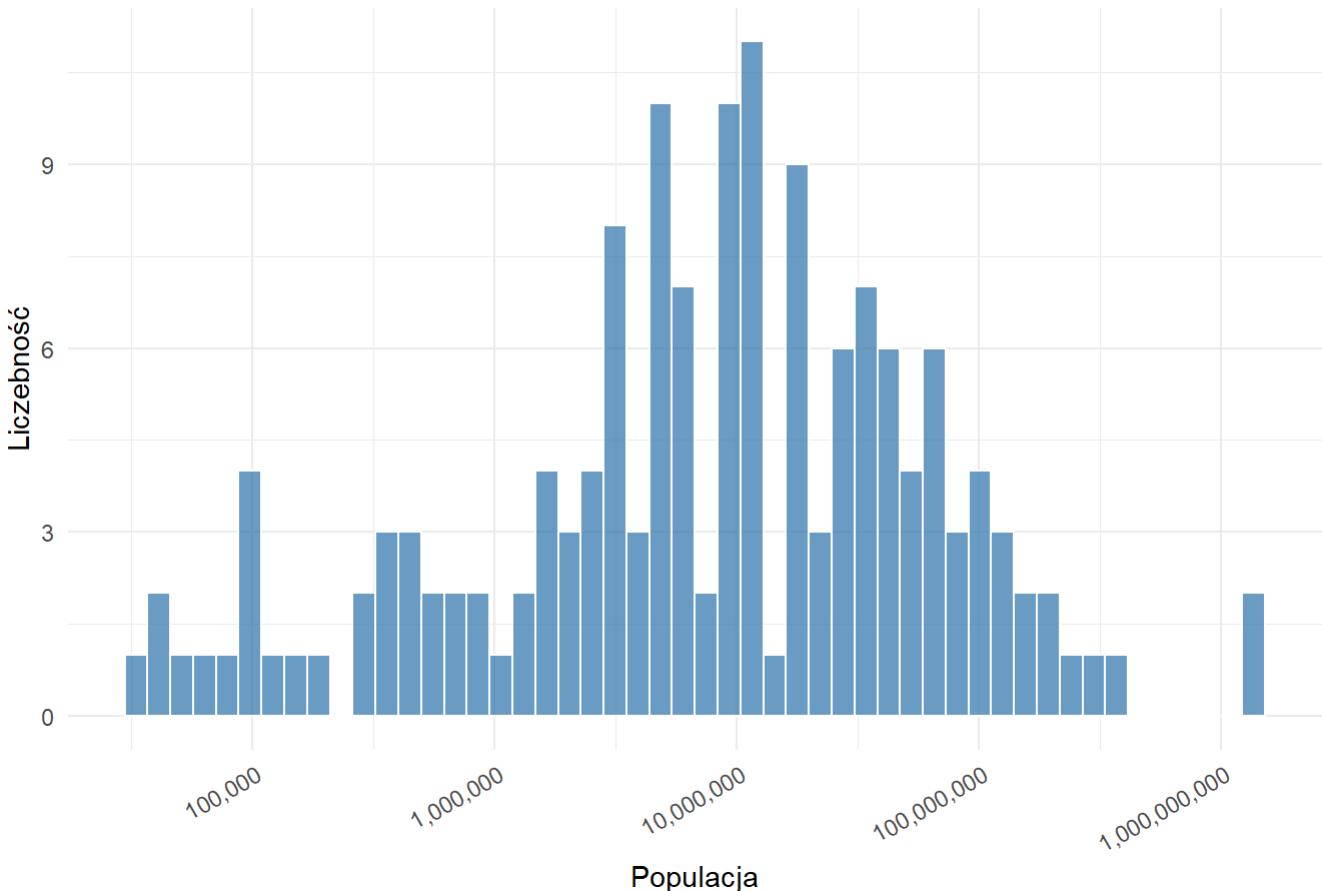


```
ggplot(df_final, aes(x = Total_stillSick)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "white", alpha = 0.8) +
  scale_x_continuous(
    trans = "pseudo_log",
    breaks = c(0, 10, 100, 1000, 10000, 100000, 1000000),
    labels = label_comma()
  ) +
  theme_minimal() +
  labs(
    title = "Rozkład liczby aktywnych przypadków",
    x = "Liczba aktywnych przypadków",
    y = "Liczebność"
  ) +
  theme(
    axis.text.x = element_text(angle = 30, hjust = 1)
  )
```



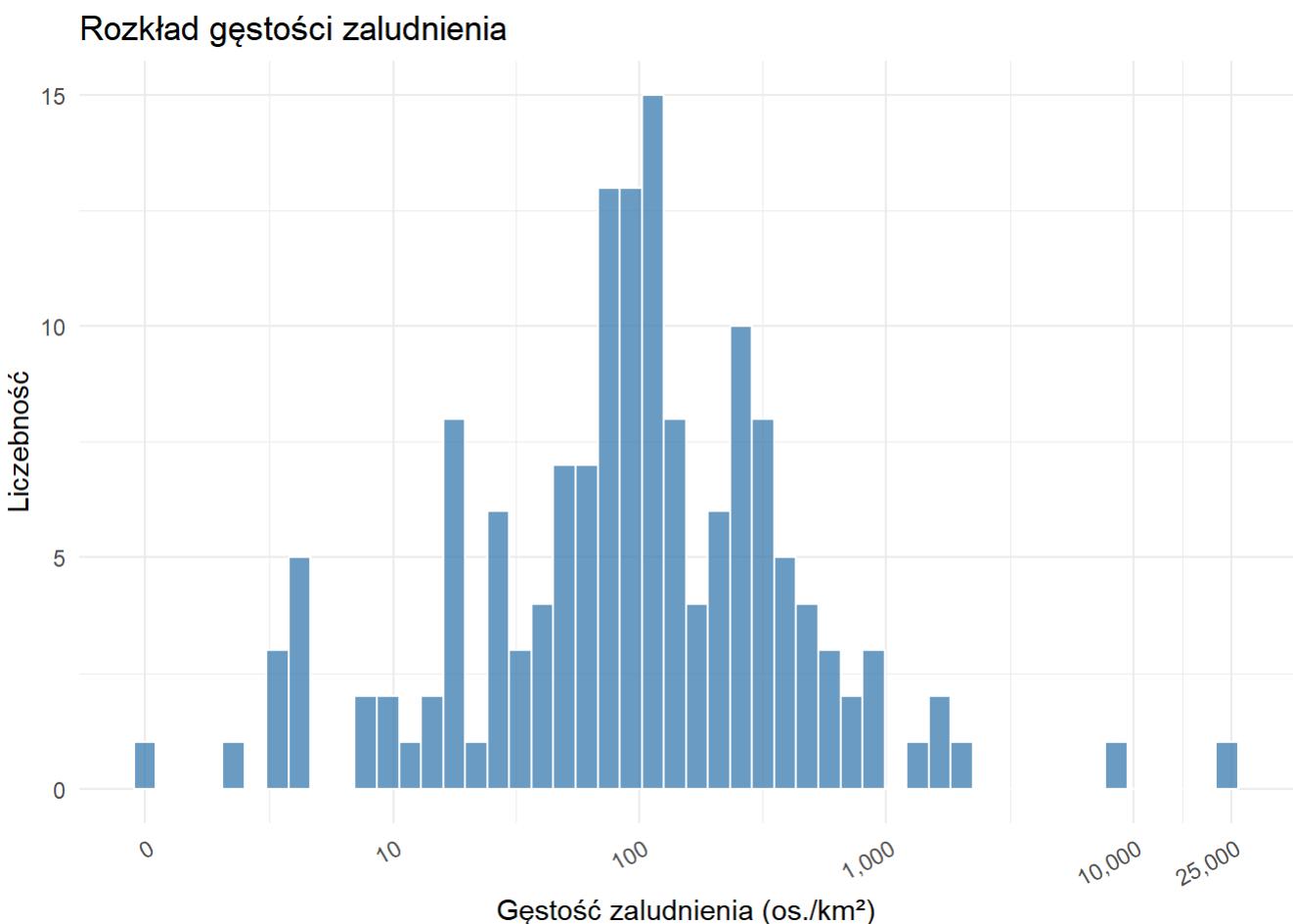
```
ggplot(df_final, aes(x = Population)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "white", alpha = 0.8) +
  scale_x_continuous(
    trans = "pseudo_log",
    breaks = c(0, 1000, 10000, 100000, 1000000, 10000000, 100000000, 1000000000),
    labels = label_comma()
  ) +
  theme_minimal() +
  labs(
    title = "Rozkład liczby ludności",
    x = "Populacja",
    y = "Liczebność"
  ) +
  theme(
    axis.text.x = element_text(angle = 30, hjust = 1)
  )
```

Rozkład liczby ludności

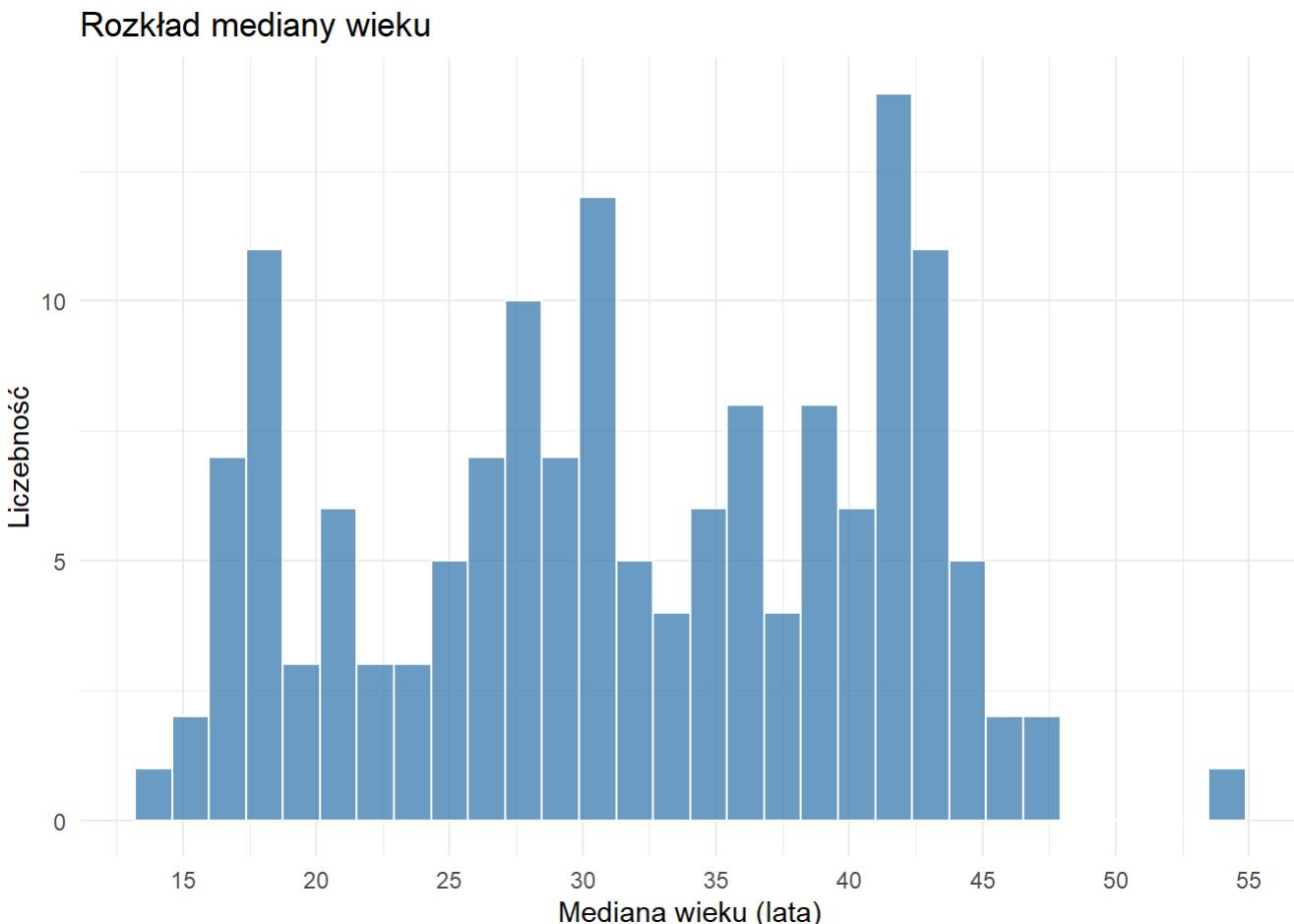


Histogram prezentuje strukturę demograficzną analizowanych państw, która w odróżnieniu od silnie skośnych danych epidemicznych na skali logarytmicznej wykazuje cechy zbliżone do rozkładu normalnego. Największa liczba krajów mieści się w środkowym przedziale populacyjnym (od kilku do kilkudziesięciu milionów mieszkańców), co obrazuje koncentracja słupków w centrum wykresu, natomiast krańce osi reprezentują odpowiednio małe terytoria wyspiarskie oraz mocarstwa o populacji przekraczającej miliard obywateli.

```
ggplot(df_final, aes(x = Density)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "white", alpha = 0.8) +
  scale_x_continuous(
    trans = "pseudo_log",
    breaks = c(0, 10, 100, 1000, 10000, 25000),
    labels = label_comma()
  ) +
  theme_minimal() +
  labs(
    title = "Rozkład gęstości zaludnienia",
    x = "Gęstość zaludnienia (os./km2)",
    y = "Liczebność"
  ) +
  theme(
    axis.text.x = element_text(angle = 30, hjust = 1)
  )
)
```



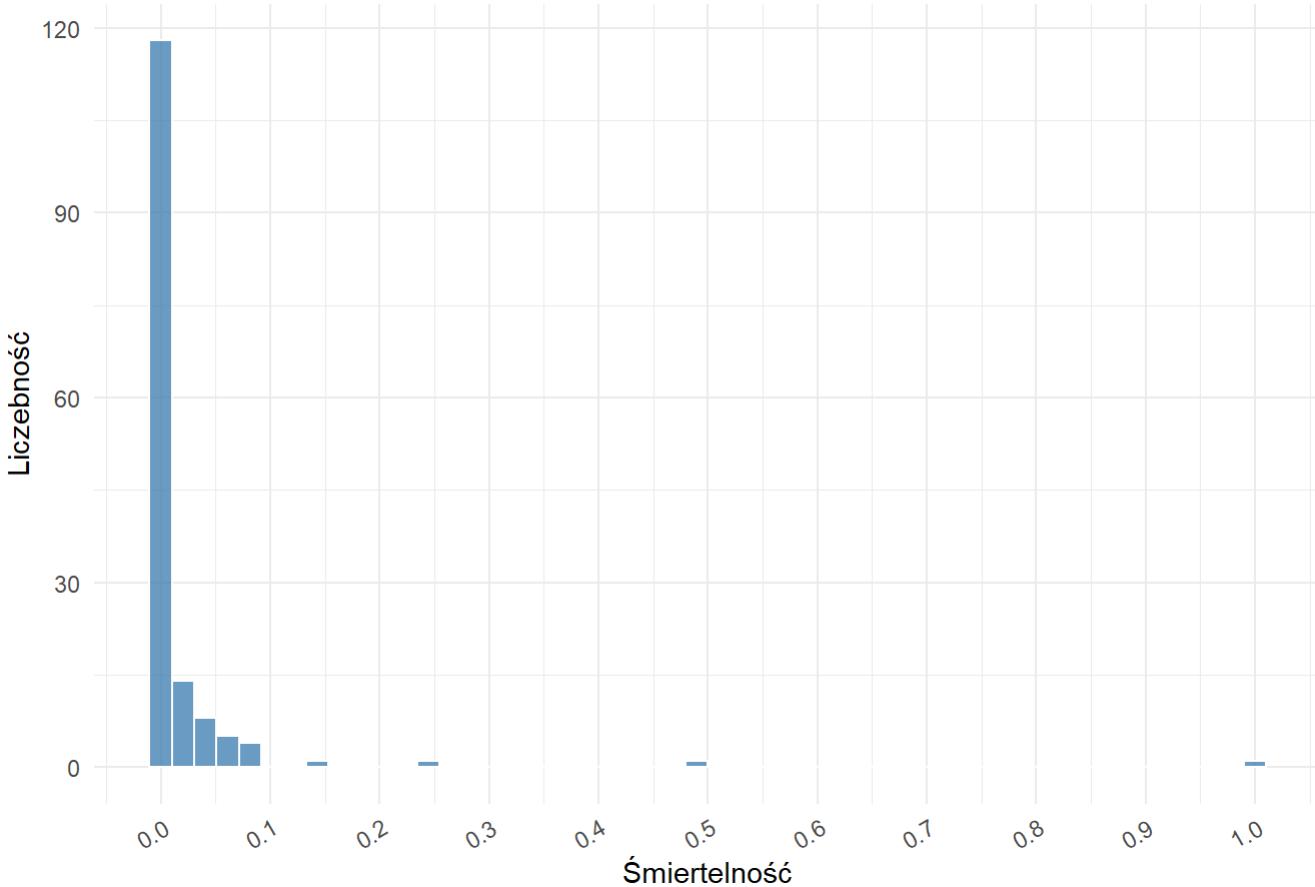
```
ggplot(df_final, aes(x = MedianAge)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "white", alpha = 0.8) +
  scale_x_continuous(
    breaks = seq(15, 55, by = 5)
  ) +
  theme_minimal() +
  labs(
    title = "Rozkład mediany wieku",
    x = "Medianą wieku (lata)",
    y = "Liczebność"
  )
```



Histogram prezentuje strukturę demograficzną analizowanych państw pod kątem mediany wieku, ujawniając nieregularny, wielomodalny charakter rozkładu, co odzwierciedla głębokie globalne zróżnicowanie na społeczeństwa młode oraz starzejące się. Wyraźnie widoczne są oddzielne skupiska państw o niskiej medianie wieku w okolicach 15–20 lat oraz silna reprezentacja krajów o starszej populacji (szczyt w przedziale 40–45 lat), przy czym zakres danych zamyka wartość maksymalną sięgającą blisko 55 lat.

```
ggplot(df_final, aes(x = Mortality_Rate)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "white", alpha = 0.8) +
  scale_x_continuous(
    n.breaks = 10
  ) +
  theme_minimal() +
  labs(
    title = "Rozkład wskaźnika śmiertelności wśród zarażonych",
    x = "Śmiertelność",
    y = "Liczebność"
  ) +
  theme(
    axis.text.x = element_text(angle = 30, hjust = 1)
  )
)
```

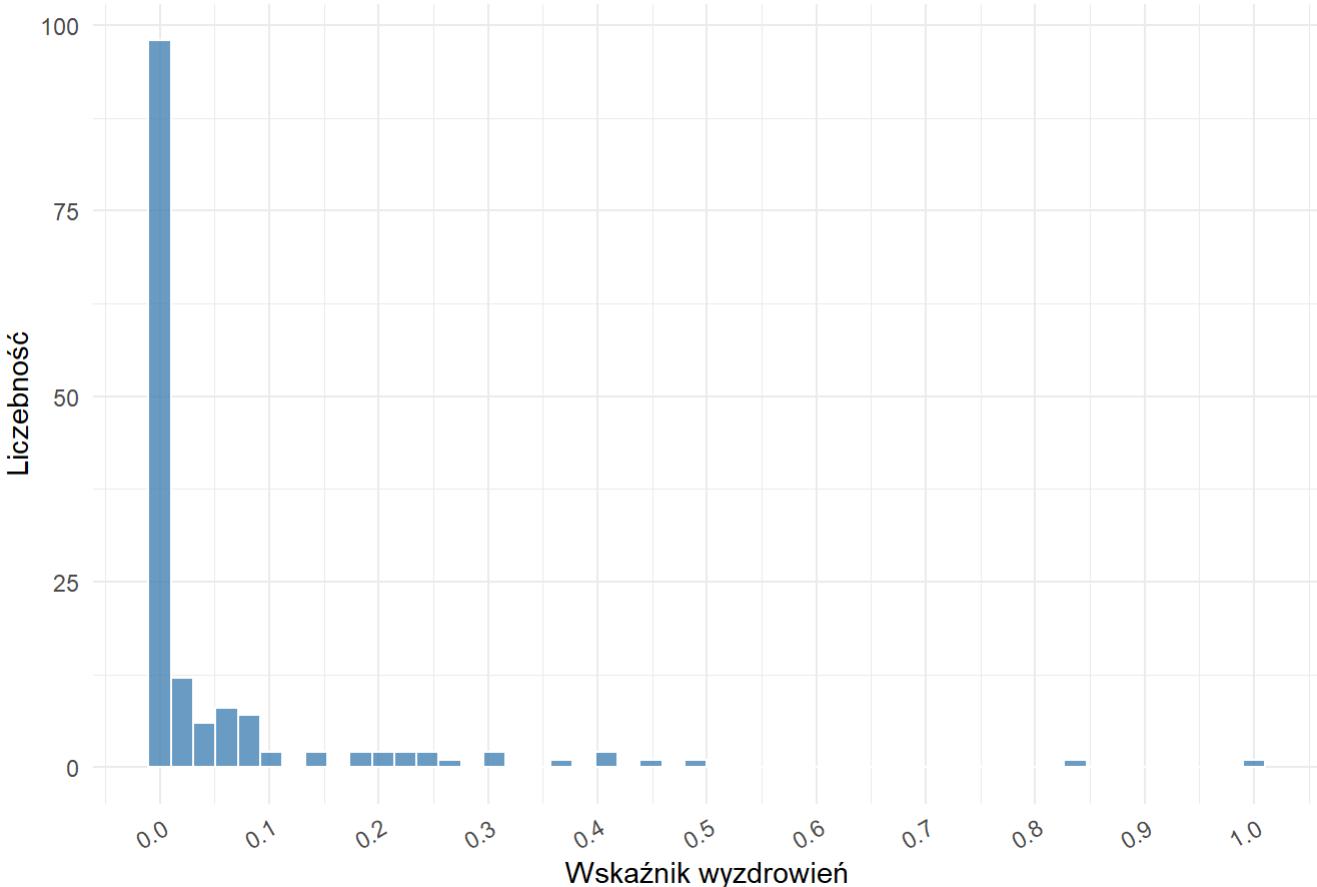
Rozkład wskaźnika śmiertelności wśród zarażonych



Histogram prezentuje rozkład stopy śmiertelności (CFR) w badanych państwach, ujawniając silną prawostronną asymetrię danych. Wykres zdominowany jest przez wysoką kolumnę w okolicach zera, co oznacza, że w analizowanym okresie większość krajów nie odnotowała ofiar śmiertelnych lub ich odsetek był minimalny. Pozostałe wartości skupią się głównie w przedziale poniżej 10%, z nielicznymi, ekstremalnymi wyjątkami sięgającymi nawet 100% (widocznymi po prawej stronie osi), co jest charakterystyczne dla początkowej fazy pandemii, gdy statystyki w poszczególnych krajach opierają się na bardzo małej liczbie wykrytych przypadków.

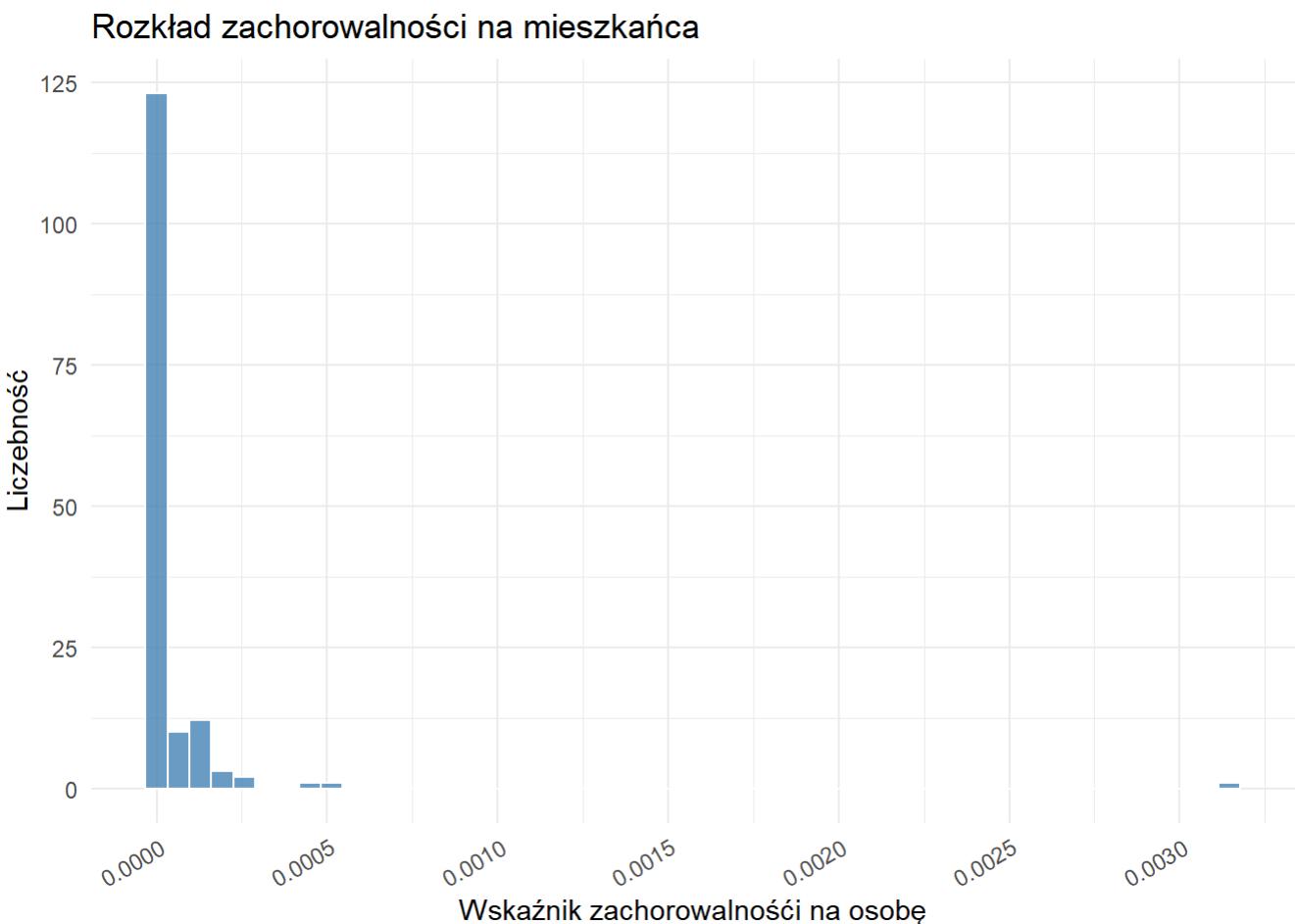
```
ggplot(df_final, aes(x = Recovery_Rate)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "white", alpha = 0.8) +
  scale_x_continuous(
    n.breaks = 10
  ) +
  theme_minimal() +
  labs(
    title = "Rozkład wskaźnika wyzdrowień wśród zarażonych",
    x = "Wskaźnik wyzdrowień",
    y = "Liczebność"
  ) +
  theme(
    axis.text.x = element_text(angle = 30, hjust = 1)
  )
)
```

Rozkład wskaźnika wyzdrowień wśród zarażonych



Histogram obrazuje rozkład wskaźnika wyzdrowień w analizowanych państwach, wykazując ekstremalną asymetrię prawostronną. Wykres jest całkowicie zdominowany przez pierwszą kolumnę przy wartości zero, co świadczy o tym, że w zdecydowanej większości krajów (blisko 100) proces zdrowienia pacjentów jeszcze się statystycznie nie rozpoczął lub nie został zaraportowany. Długi, płaski ogon rozkładu wskazuje na nieliczne państwa, w których odsetek wyzdrowień jest wyższy, osiągając w pojedynczych przypadkach wartości zbliżone do 100% (najpewniej w krajach z pojedynczymi, wyleczonymi przypadkami).

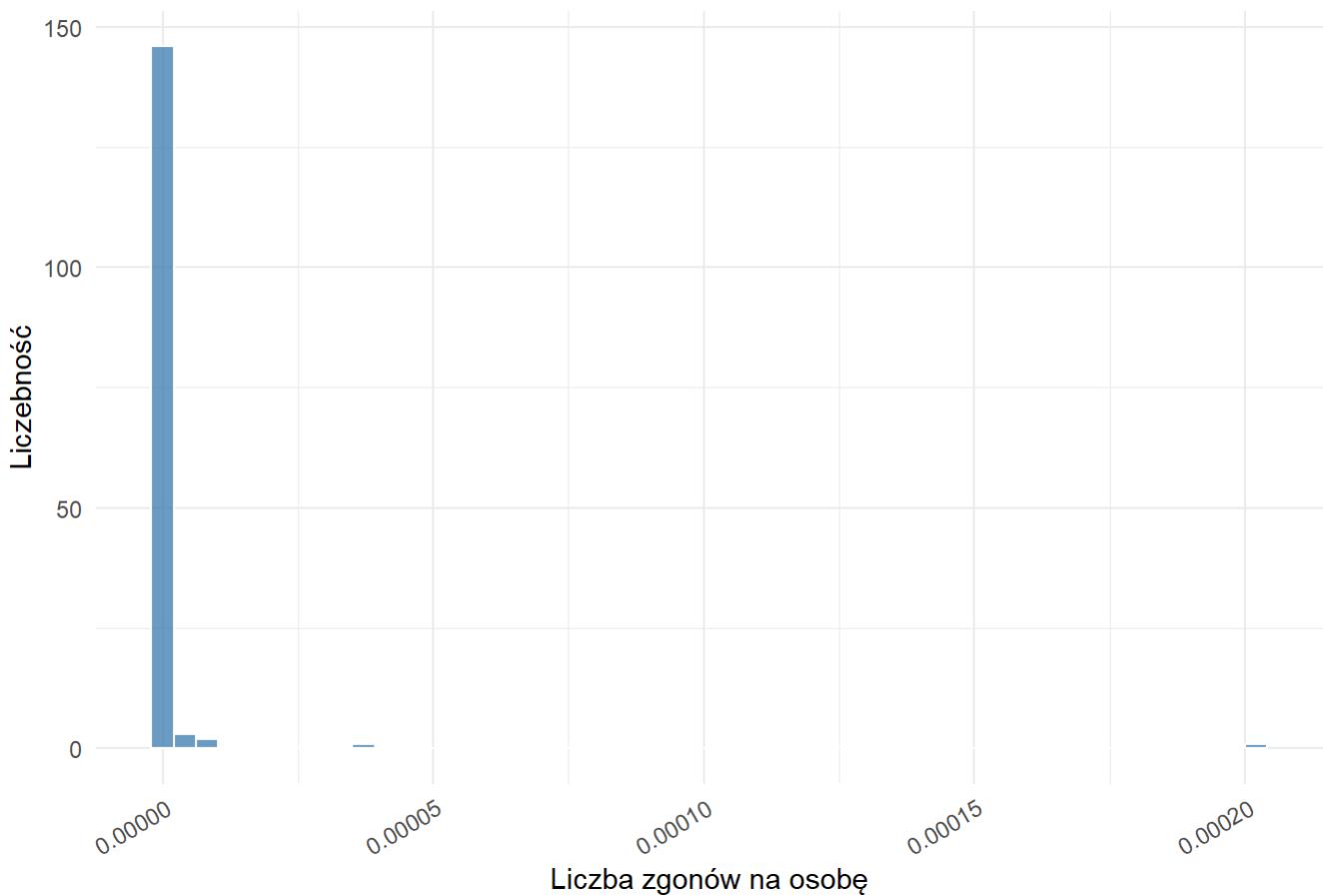
```
ggplot(df_final, aes(x = Incidence_Rate)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "white", alpha = 0.8) +
  scale_x_continuous(
    n.breaks = 10
  ) +
  theme_minimal() +
  labs(
    title = "Rozkład zachorowalności na mieszkańca",
    x = "Wskaźnik zachorowalności na osobę",
    y = "Liczebność"
  ) +
  theme(
    axis.text.x = element_text(angle = 30, hjust = 1)
  )
)
```



Histogram przedstawia rozkład wskaźnika zachorowalności w przeliczeniu na jednego mieszkańca, obrazując skrajną nierównomierność rozprzestrzeniania się wirusa w badanej populacji państwa. Wykres zdominowany jest przez masywne skupienie wartości w okolicach zera, co dowodzi, że w analizowanym oknie czasowym odsetek osób zakażonych względem ogółu ludności był w większości regionów świata wciąż marginalny. Widoczny po prawej stronie, odizolowany punkt (outlier) wskazuje na istnienie pojedynczego ogniska o wyjątkowo wysokiej intensywności transmisji, gdzie wirus zdołał zainfekować nieporównywalnie większą część społeczeństwa niż w pozostałych krajach.

```
ggplot(df_final, aes(x = Mortality_per_capita)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "white", alpha = 0.8) +
  scale_x_continuous(
    trans = "pseudo_log",
    breaks = c(0, 0.00005, 0.0001, 0.00015, 0.0002, 0.0005),
    labels = label_number(accuracy = 0.00001)
  ) +
  theme_minimal() +
  labs(
    title = "Rozkład śmiertelności na mieszkańca",
    x = "Liczba zgonów na osobę",
    y = "Liczebność"
  ) +
  theme(
    axis.text.x = element_text(angle = 30, hjust = 1)
  )
```

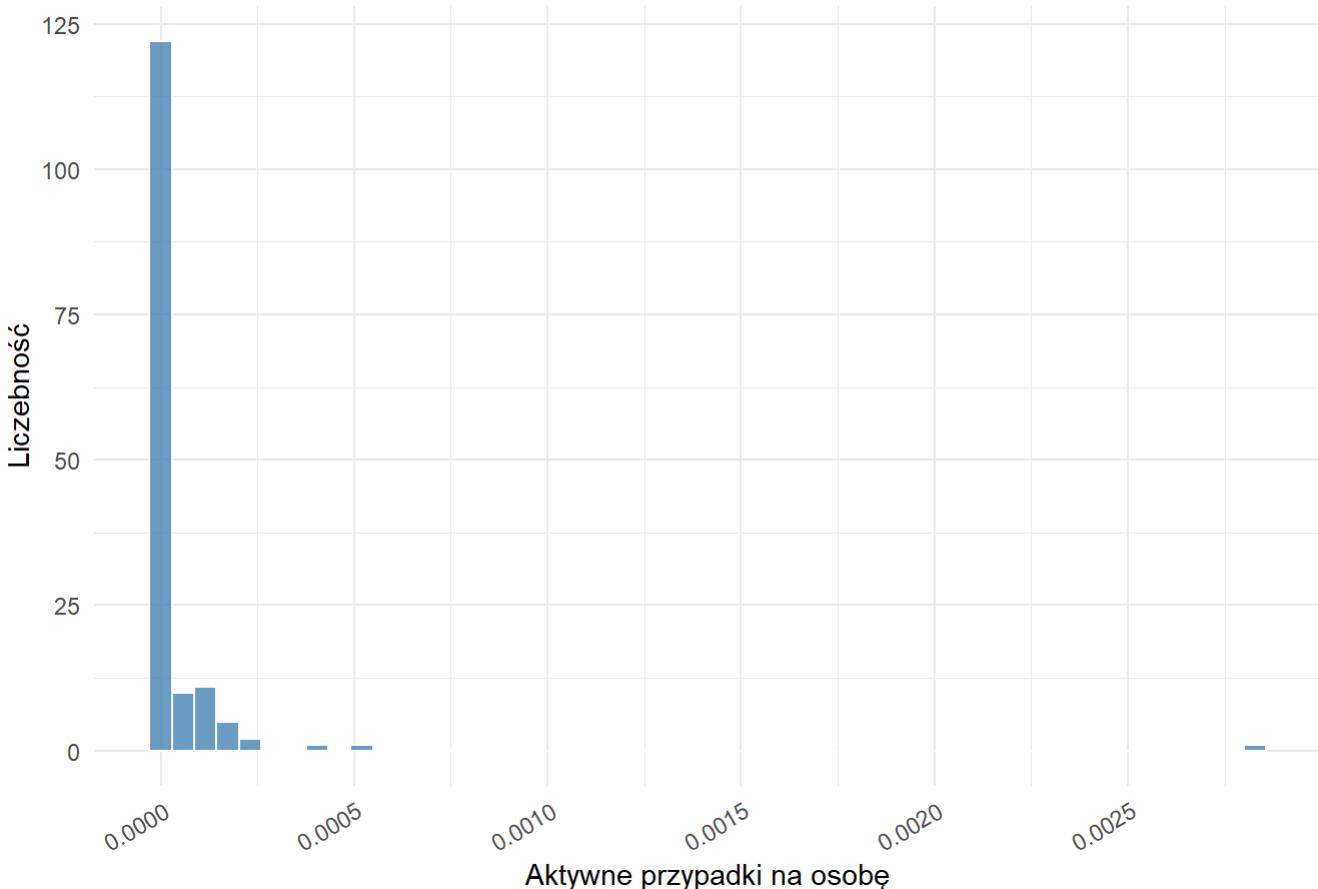
Rozkład śmiertelności na mieszkańca



Histogram obrazuje rozkład wskaźnika umieralności w przeliczeniu na jednego mieszkańca, ujawniając, że w analizowanym okresie globalne obciążenie demograficzne skutkami pandemii było statystycznie marginalne dla niemal wszystkich państw. Wykres jest całkowicie zdominowany przez słupek przy wartościach bliskich zeru, co oznacza, że w zdecydowanej większości krajów liczba zgonów w stosunku do populacji była znikoma. Jedynie pojedynczy, skrajny punkt widoczny po prawej stronie wskazuje na istnienie odosobnionego regionu, gdzie wskaźnik ten osiągnął zauważalnie wyższy poziom na tle reszty świata.

```
ggplot(df_final, aes(x = StillSick_per_capita)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "white", alpha = 0.8) +
  scale_x_continuous(
    trans = "pseudo_log",
    breaks = c(0, 0.0005, 0.0010, 0.0015, 0.0020, 0.0025),
    labels = label_number(accuracy = 0.0001)
  ) +
  theme_minimal() +
  labs(
    title = "Rozkład aktywnych przypadków na mieszkańca",
    x = "Aktywne przypadki na osobę",
    y = "Liczебность"
  ) +
  theme(
    axis.text.x = element_text(angle = 30, hjust = 1)
  )
```

Rozkład aktywnych przypadków na mieszkańca



Histogram przedstawia rozkład liczby aktywnych przypadków w przeliczeniu na jednego mieszkańca, co pozwala ocenić rzeczywiste obciążenie społeczeństw trwającą infekcją. Wykres jest niemal całkowicie zdominowany przez wysoki słupek przy wartościach bliskich zeru, co wskazuje, że w skali globalnej odsetek osób aktualnie chorujących był w analizowanym momencie znikomy dla zdecydowanej większości państw. Jedynie pojedynczy, wyraźnie odseparowany słupek po prawej stronie osi reprezentuje nieliczny wyjątek (kraj o najwyższej intensywności epidemii), gdzie wirus zdołał zainfekować zauważalnie większą część populacji niż w pozostałych regionach świata.

```
cols_abs <- c("Total_confirmed", "Total_death", "Total_recovered", "Total_stillSick")

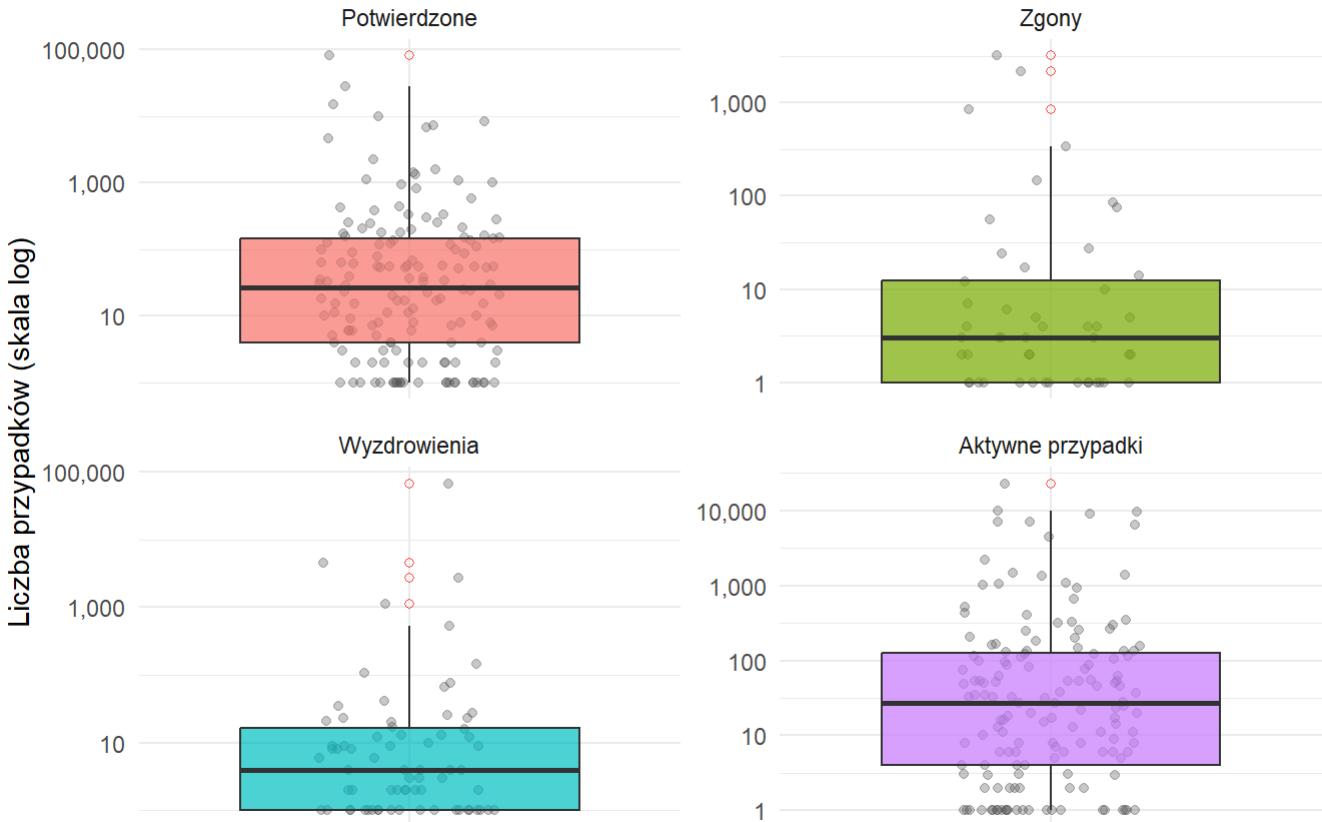
nazwy_pl <- c(
  "Total_confirmed" = "Potwierdzone",
  "Total_death" = "Zgony",
  "Total_recovered" = "Wyzdrowienia",
  "Total_stillSick" = "Aktywne przypadki"
)

df_abs <- df_final %>%
  select(Country.Region, all_of(cols_abs)) %>%
  pivot_longer(cols = all_of(cols_abs), names_to = "Zmienna", values_to = "Wartosc") %>%
  filter(Wartosc > 0)

ggplot(df_abs, aes(x = Zmienna, y = Wartosc, fill = Zmienna)) +
  geom_jitter(width = 0.2, alpha = 0.3, color = "gray30") +
  geom_boxplot(alpha = 0.7, outlier.colour = "red", outlier.shape = 1) +
  facet_wrap(~ Zmienna, scales = "free", labeller = as_labeller(nazwy_pl)) +
  scale_y_log10(labels = label_comma()) +
  labs(
    title = "Rozkład zmiennych",
    subtitle = "Skala logarytmiczna, niezależne osie Y dla każdego panelu",
    y = "Liczba przypadków (skala log)",
    x = NULL
  ) +
  theme_minimal() +
  theme(
    legend.position = "none",
    axis.text.x = element_blank()
)
```

Rozkład zmiennych

Skala logarytmiczna, niezależne osie Y dla każdego panelu



Wykres pudełkowy z nałożonymi punktami rzeczywistymi obrazuje rozkład czterech kluczowych zmiennych bezwzględnych w skali logarytmicznej, co umożliwia porównanie danych o skrajnie różnej rozpiętości. Dla wszystkich kategorii, od potwierdzonych zakażeń po zgony, charakterystyczne jest niskie położenie mediany (pozioma kreska wewnętrz pudełka), co potwierdza, że typowy kraj odnotował relatywnie niewielką liczbę przypadków, podczas gdy czerwone punkty na szczytach wykresów wskazują na ekstremalne wartości odstające, reprezentujące państwa najsielniej dotknięte pandemią.

```
cols_rate <- c("Incidence_Rate", "Mortality_per_capita",
               "StillSick_per_capita", "Density")

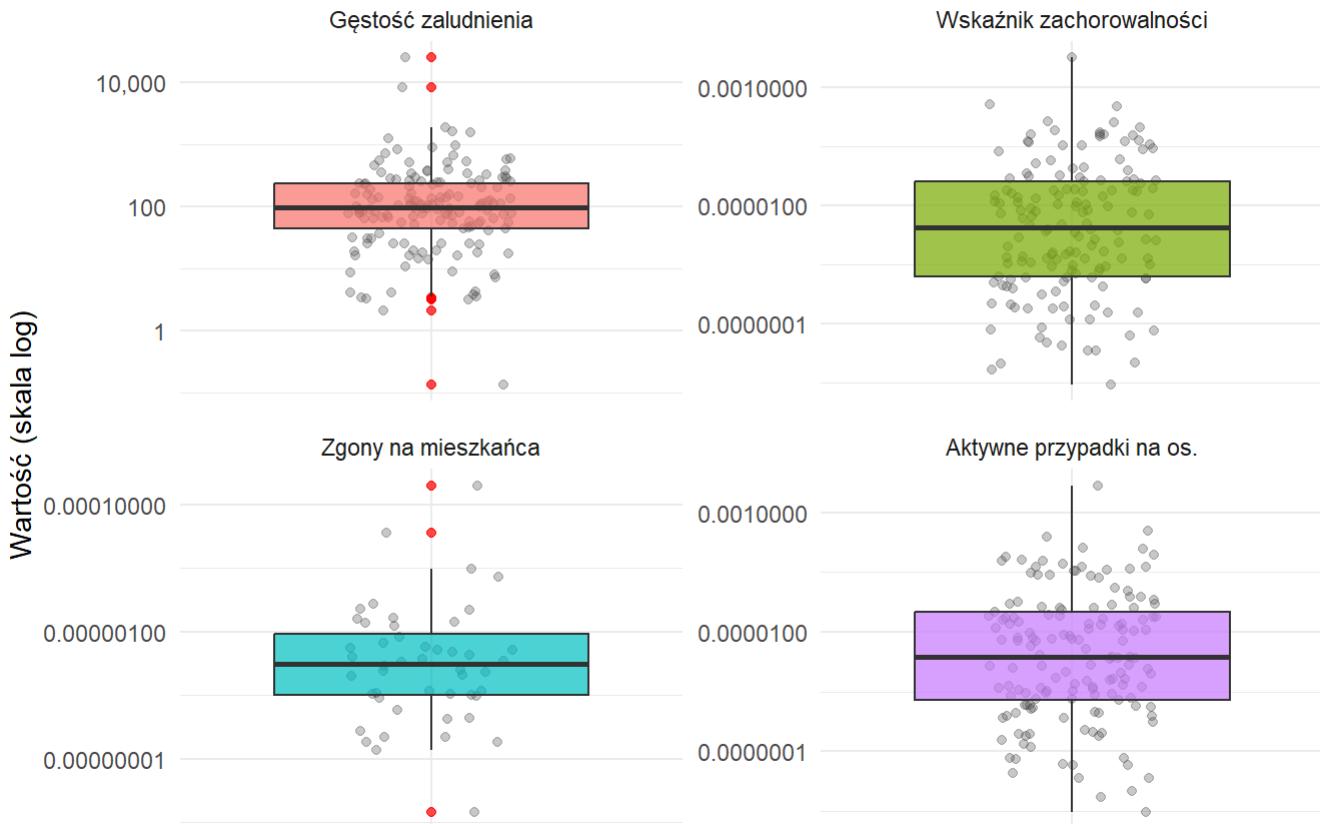
nazwy_pl_rate <- c(
  "Incidence_Rate" = "Wskaźnik zachorowalności",
  "Mortality_per_capita" = "Zgony na mieszkańców",
  "StillSick_per_capita" = "Aktywne przypadki na os.",
  "Density" = "Gęstość zaludnienia"
)

df_rate <- df_final %>%
  select(Country.Region, all_of(cols_rate)) %>%
  pivot_longer(cols = all_of(cols_rate), names_to = "Zmienna", values_to = "Wartosc") %>%
  filter(Wartosc > 0)

ggplot(df_rate, aes(x = Zmienna, y = Wartosc, fill = Zmienna)) +
  geom_jitter(width = 0.2, alpha = 0.3, color = "gray30") +
  geom_boxplot(alpha = 0.7, outlier.colour = "red") +
  facet_wrap(~Zmienna, scales = "free", labeller = as_labeller(nazwy_pl_rate)) +
  scale_y_log10(labels = label_comma()) +
  labs(
    title = "Wskaźniki epidemiologiczne i gęstość zaludnienia",
    subtitle = "Skala logarytmiczna, niezależne osie Y",
    y = "Wartość (skala log)",
    x = NULL
  ) +
  theme_minimal() +
  theme(
    legend.position = "none",
    axis.text.x = element_blank()
)
```

Wskaźniki epidemiologiczne i gęstość zaludnienia

Skala logarytmiczna, niezależne osie Y



Wykres pudełkowy z nałożonymi punktami prezentuje rozkład gęstości zaludnienia oraz kluczowych wskaźników epidemiologicznych w ujęciu relatywnym (na mieszkańca), wykorzystując niezależne skale logarytmiczne dla każdej zmiennej w celu czytelnego porównania rzędów wielkości. Analiza ujawnia, że mimo ogromnego zróżnicowania gęstości zaludnienia (od kilku do kilkunastu tysięcy osób na km²), wskaźniki zachorowalności i umieralności w większości krajów pozostają na bardzo niskim, ułamkowym poziomie, co obrazuje wczesną fazę rozwoju pandemii w skali globalnej. Czerwone punkty odstające (outliers) widoczne nad górnymi wąsami wykresów identyfikują nieliczne państwa, w których intensywność ataku wirusa lub zagęszczenie ludności drastycznie odbiegają od światowej normy.

```

cols_pct <- c("Mortality_Rate", "Recovery_Rate", "MedianAge")

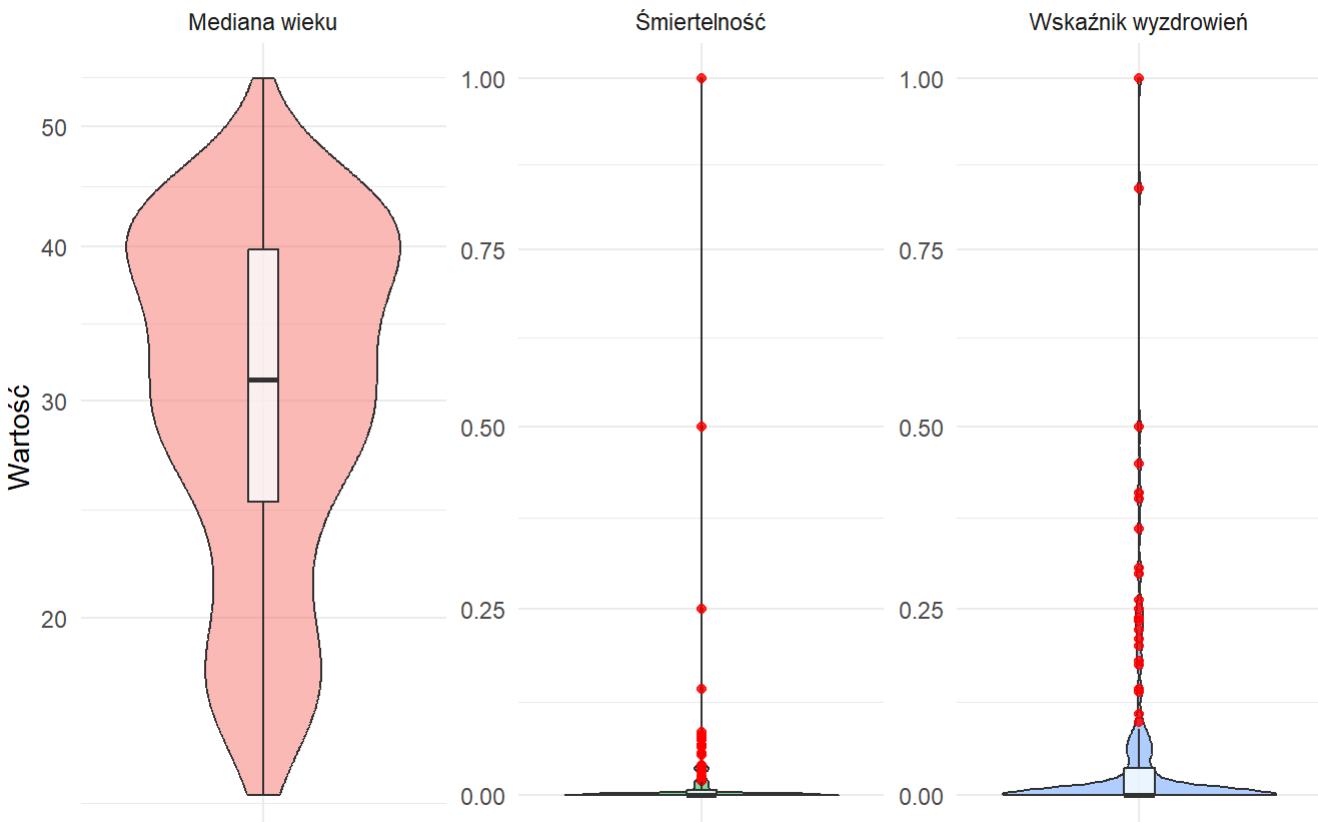
df_pct <- df_final %>%
  select(Country.Region, all_of(cols_pct)) %>%
  pivot_longer(cols = all_of(cols_pct), names_to = "Zmienna", values_to = "Wartosc") %>%
  mutate(Zmienna = case_match(
    Zmienna,
    "Mortality_Rate" ~ "Śmiertelność",
    "Recovery_Rate" ~ "Wskaźnik wyzdrowień",
    "MedianAge" ~ "Mediana wieku"
  ))

ggplot(df_pct, aes(x = Zmienna, y = Wartosc, fill = Zmienna)) +
  geom_violin(trim = TRUE, scale = "width", alpha = 0.5) +
  geom_boxplot(width = 0.1, outlier.colour = "red", fill = "white", alpha = 0.8) +
  facet_wrap(~Zmienna, scales = "free") +
  scale_y_continuous(trans = "pseudo_log", labels = label_number()) +
  labs(
    title = "Demografia i wskaźniki epidemiologiczne",
    subtitle = "Wartości surowe (skala pseudo-logarytmiczna)",
    y = "Wartość",
    x = NULL
  ) +
  theme_minimal() +
  theme(
    legend.position = "none",
    axis.text.x = element_blank()
  )
)

```

Demografia i wskaźniki epidemiologiczne

Wartości surowe (skala pseudo-logarytmiczna)



Zestawienie wykresów skrzypcowych z naniesionymi wykresami pułapkowymi jaskrawo kontrastuje stabilną strukturę demograficzną z dynamiczną naturą wskaźników epidemicznych. Panel dotyczący mediany wieku ukazuje szeroki, rozbudowany rozkład danych, co świadczy o dużym zróżnicowaniu globalnym bez dominacji jednej wartości, podczas gdy wskaźniki śmiertelności i wyzdrowień przybierają kształt silnie spłaszczony przy osi dolnej z długimi, wąskimi wypustkami ku górze. Taka wizualizacja potwierdza, że o ile struktura wiekowa jest cechą stałą i różnorodną, to wysokie odsetki zgonów lub wyzdrowień stanowiły w analizowanym momencie statystyczne anomalie (oznaczone czerwonymi punktami), podczas gdy normą dla większości świata były wartości bliskie零.

```
print("--- TOP 10: Najwięcej potwierdzonych przypadków (Total Confirmed) ---")
```

```
## [1] "--- TOP 10: Najwięcej potwierdzonych przypadków (Total Confirmed) ---"
```

```
df_final %>%
  select(Country.Region, Total_confirmed, Population) %>%
  arrange(desc(Total_confirmed)) %>%
  head(10) %>%
  print()
```

```
## # A tibble: 10 × 3
##   Country.Region Total_confirmed Population
##   <chr>                <int>      <dbl>
## 1 China                  81033 1423520358.
## 2 Italy                   27980  60130136
## 3 Iran                    14991  87051648.
## 4 Spain                   9942   47435119
## 5 Korea, South            8236   51767846
## 6 Germany                 7272   83559186.
## 7 France                  6650   65729459
## 8 US                      4632   337790066.
## 9 Switzerland              2200   8577524
## 10 United Kingdom          1551   67110958
```

```
print("--- TOP 10: Najwięcej zgonów (Total Death) ---")
```

```
## [1] "--- TOP 10: Najwięcej zgonów (Total Death) ---"
```

```
df_final %>%
  select(Country.Region, Total_death, Total_confirmed) %>%
  arrange(desc(Total_death)) %>%
  head(10) %>%
  print()
```

```
## # A tibble: 10 × 3
##   Country.Region Total_death Total_confirmed
##   <chr>           <int>          <int>
## 1 China            3217          81033
## 2 Italy             2158          27980
## 3 Iran              853          14991
## 4 Spain             342          9942
## 5 France            148          6650
## 6 US                85          4632
## 7 Korea, South      75          8236
## 8 United Kingdom    56          1551
## 9 Japan              27          825
## 10 Netherlands       24          1414
```

```
print("--- TOP 10: Najwięcej wyzdrowień (Total Recovered) ---")
```

```
## [1] "--- TOP 10: Najwięcej wyzdrowień (Total Recovered) ---"
```

```
df_final %>%
  select(Country.Region, Total_recovered, Total_confirmed) %>%
  arrange(desc(Total_recovered)) %>%
  head(10) %>%
  print()
```

```
## # A tibble: 10 × 3
##   Country.Region Total_recovered Total_confirmed
##   <chr>           <int>          <int>
## 1 China            67910          81033
## 2 Iran              4590          14991
## 3 Italy             2749          27980
## 4 Korea, South     1137          8236
## 5 Spain             530          9942
## 6 Japan              144          825
## 7 Singapore          109          243
## 8 Bahrain             77          214
## 9 Germany             67          7272
## 10 Malaysia            42          566
```

```
print("--- TOP 10: Najwięcej aktywnych przypadków (Total Still Sick) ---")
```

```
## [1] "--- TOP 10: Najwięcej aktywnych przypadków (Total Still Sick) ---"
```

```
df_final %>%
  select(Country.Region, Total_stillSick, Total_confirmed) %>%
  arrange(desc(Total_stillSick)) %>%
  head(10) %>%
  print()
```

```
## # A tibble: 10 × 3
##   Country.Region Total_stillSick Total_confirmed
##   <chr>           <int>            <int>
## 1 Italy            23073            27980
## 2 China            9906             81033
## 3 Iran             9548             14991
## 4 Spain            9070             9942
## 5 Germany          7188             7272
## 6 Korea, South     7024             8236
## 7 France           6490             6650
## 8 US               4530             4632
## 9 Switzerland       2182              2200
## 10 United Kingdom  1474              1551
```

```
print("--- TOP 10: Gęstość Zaludnienia (Density) ---")
```

```
## [1] "--- TOP 10: Gęstość Zaludnienia (Density) ---"
```

```
df_final %>%
  select(Country.Region, Density, Population) %>%
  arrange(desc(Density)) %>%
  head(10) %>%
  print()
```

```
## # A tibble: 10 × 3
##   Country.Region      Density Population
##   <chr>           <dbl>      <dbl>
## 1 Monaco            25577.    38109
## 2 Singapore          8301.    5669562.
## 3 Bahrain            1897.    1485670.
## 4 Maldives           1626.    487731
## 5 Malta              1600.    504016.
## 6 Bangladesh         1267.    164913055
## 7 Guernsey           981.     62784.
## 8 Jersey              887.     102856.
## 9 occupied Palestinian territory  824.     4957768
## 10 Mayotte           733.     274910.
```

```
print("--- TOP 10: Zachorowalność na populację (Incidence Rate) ---")
```

```
## [1] "--- TOP 10: Zachorowalność na populację (Incidence Rate) ---"
```

```
df_final %>%
  select(Country.Region, Incidence_Rate, Total_confirmed, Population) %>%
  arrange(desc(Incidence_Rate)) %>%
  head(10) %>%
  print()
```

```
## # A tibble: 10 × 4
##   Country.Region Incidence_Rate Total_confirmed Population
##   <chr>           <dbl>        <int>      <dbl>
## 1 San Marino     0.00315       109      34653
## 2 Iceland         0.000499      180     360700.
## 3 Italy            0.000465     27980    60130136
## 4 Switzerland     0.000256      2200     8577524
## 5 Norway           0.000249      1333     5347730.
## 6 Spain            0.000210      9942    47435119
## 7 Monaco           0.000184        7      38109
## 8 Iran             0.000172     14991    87051648.
## 9 Denmark          0.000160      932     5814618.
## 10 Korea, South    0.000159     8236    51767846
```

```
print("--- TOP 10: Zgony na populację (Mortality per Capita) ---")
```

```
## [1] "--- TOP 10: Zgony na populację (Mortality per Capita) ---"
```

```
df_final %>%
  select(Country.Region, Mortality_per_capita, Total_death, Population) %>%
  arrange(desc(Mortality_per_capita)) %>%
  head(10) %>%
  print()
```

```
## # A tibble: 10 × 4
##   Country.Region Mortality_per_capita Total_death Population
##   <chr>           <dbl>        <int>      <dbl>
## 1 San Marino     0.000202       7      34653
## 2 Italy            0.0000359     2158    60130136
## 3 Iran             0.00000980      853    87051648.
## 4 Spain            0.00000721      342    47435119
## 5 Martinique      0.00000278        1    359612.
## 6 China            0.00000226     3217  1423520358.
## 7 France           0.00000225      148    65729459
## 8 Switzerland      0.00000163       14    8577524
## 9 Luxembourg       0.00000161        1    620163
## 10 Korea, South    0.00000145       75    51767846
```

```
print("--- TOP 10: Aktywne przypadki na populację (Still Sick per Capita) ---")
```

```
## [1] "--- TOP 10: Aktywne przypadki na populację (Still Sick per Capita) ---"
```

```
df_final %>%
  select(Country.Region, StillSick_per_capita, Total_stillSick, Population) %>%
  arrange(desc(StillSick_per_capita)) %>%
  head(10) %>%
  print()
```

```
## # A tibble: 10 × 4
##   Country.Region StillSick_per_capita Total_stillSick Population
##   <chr>                <dbl>        <int>      <dbl>
## 1 San Marino            0.00283         98     34653
## 2 Iceland               0.000499       180    360700.
## 3 Italy                  0.000384     23073  60130136
## 4 Switzerland            0.000254     2182    8577524
## 5 Norway                 0.000249     1329    5347730.
## 6 Spain                  0.000191     9070   47435119
## 7 Monaco                 0.000184        7    38109
## 8 Denmark                0.000160     928    5814618.
## 9 Qatar                  0.000155     435    2797921
## 10 Estonia                0.000154     204    1326822.
```

```
print("--- TOP 10: Śmiertelność przypadków CFR (Mortality Rate) ---")
```

```
## [1] "--- TOP 10: Śmiertelność przypadków CFR (Mortality Rate) ---"
```

```
df_final %>%
  select(Country.Region, Mortality_Rate, Total_death, Total_confirmed) %>%
  arrange(desc(Mortality_Rate)) %>%
  head(10) %>%
  print()
```

```
## # A tibble: 10 × 4
##   Country.Region Mortality_Rate Total_death Total_confirmed
##   <chr>                <dbl>        <int>          <int>
## 1 Sudan                  1            1              1
## 2 Guatemala              0.5           1              2
## 3 Guyana                 0.25          1              4
## 4 Ukraine                 0.143          1              7
## 5 Philippines            0.0845         12             142
## 6 Iraq                   0.0806         10             124
## 7 Italy                   0.0771        2158            27980
## 8 Algeria                 0.0741          4              54
## 9 Azerbaijan              0.0667          1              15
## 10 Martinique             0.0667          1              15
```

```
print("--- TOP 10: Wskaźnik wyzdrowień (Recovery Rate) ---")
```

```
## [1] "--- TOP 10: Wskaźnik wyzdrowień (Recovery Rate) ---"
```

```
df_final %>%
  select(Country.Region, Recovery_Rate, Total_recovered, Total_confirmed) %>%
  arrange(desc(Recovery_Rate)) %>%
  head(10) %>%
  print()
```

```
## # A tibble: 10 × 4
##   Country.Region Recovery_Rate Total_recovered Total_confirmed
##   <chr>           <dbl>        <int>            <int>
## 1 Nepal             1            1              1
## 2 China            0.838       67910          81033
## 3 Andorra           0.5           1              2
## 4 Singapore         0.449       109            243
## 5 Oman              0.409        9              22
## 6 Azerbaijan        0.4           6              15
## 7 Bahrain            0.360       77             214
## 8 Iran               0.306       4590           14991
## 9 Taiwan*            0.299        20              67
## 10 Vietnam           0.262       16              61
```

```
print("--- TOP 10: Największa populacja (Population) ---")
```

```
## [1] "--- TOP 10: Największa populacja (Population) ---"
```

```
df_final %>%
  select(Country.Region, Population, Density) %>%
  arrange(desc(Population)) %>%
  head(10) %>%
  print()
```

```
## # A tibble: 10 × 3
##   Country.Region  Population Density
##   <chr>           <dbl>    <dbl>
## 1 China            1423520358. 148.
## 2 India             1389030312 467.
## 3 US                337790066. 36.9
## 4 Indonesia         272489381 143.
## 5 Pakistan          230800898. 299.
## 6 Nigeria           209485641 230.
## 7 Brazil             207455460. 24.8
## 8 Bangladesh         164913055 1267.
## 9 Russia             146533067  8.95
## 10 Japan             126699424. 336.
```

```
print("--- TOP 10: Najwyższa mediana wieku (Median Age) ---")
```

```
## [1] "--- TOP 10: Najwyższa mediana wieku (Median Age) ---"
```

```
df_final %>%
  select(Country.Region, MedianAge, Population) %>%
  arrange(desc(MedianAge)) %>%
  head(10) %>%
  print()
```

```
## # A tibble: 10 × 3
##   Country Region MedianAge Population
##   <chr>     <dbl>      <dbl>
## 1 Monaco      54.6     38109
## 2 Japan       47.3 126699424.
## 3 Martinique   47.0    359612.
## 4 Italy        46.0    60130136
## 5 San Marino   45.3    34653
## 6 Germany      44.9    83559186.
## 7 Portugal     44.6    10343213
## 8 Greece       44.0    10718576.
## 9 Croatia      44.0    3986334.
## 10 Guadeloupe  43.8    410256.
```

Analiza zgromadzonych rankingów wskazuje na dominującą rolę Chin i Włoch w początkowej fazie pandemii, przy czym Chiny odnotowały najwyższe wartości bezwzględne w kategoriach potwierdzonych zakażeń, zgonów oraz wyzdrowień. Włochy wyróżniały się największą na świecie liczbą aktywnych przypadków oraz bardzo wysoką śmiertelnością przypadków (CFR) na poziomie 7,7%, ustępując w tym rankingu jedynie państwowom o sładowej liczbie wykrytych infekcji, takim jak Sudan czy Gwatemala. W ujęciu relatywnym najbardziej dotkniętym terytorium okazało się San Marino, które zajęło pierwsze miejsce pod względem zachorowalności, umieralności oraz liczby aktywnych chorych w przeliczeniu na jednego mieszkańca. Wysokie wskaźniki zapadalności odnotowano również w Islandii i Szwajcarii, co kontrastuje z bardzo niskimi wartościami tych miar w większości pozostałych regionów świata. Pod względem demograficznym kraje najsilniej dotknięte wirusem, jak Włochy czy Niemcy, charakteryzują się wysoką medianą wieku przekraczającą 44 lata, podczas gdy Monako łączy najwyższą średnią wieku (blisko 55 lat) z ekstremalnie wysoką gęstością zaludnienia. Statystyki wyzdrowień poza Chinami najkorzystniej prezentowały się w Iranie i Singapurze, choć globalny wskaźnik ozdrowień pozostawał na niskim poziomie ze względu na wczesny etap rozwoju zjawiska.

```

vars_for_cluster <- df_final %>%
  select(
    Incidence_Rate,
    Mortality_per_capita,
    StillSick_per_capita,
    Mortality_Rate,
    Recovery_Rate,
    Total_confirmed,
    Total_death,
    Total_recovered,
    Total_stillSick,
    Population,
    Density,
    MedianAge,
    Active_Rate
  )

cor_matrix <- cor(vars_for_cluster, method = "pearson", use = "complete.obs")

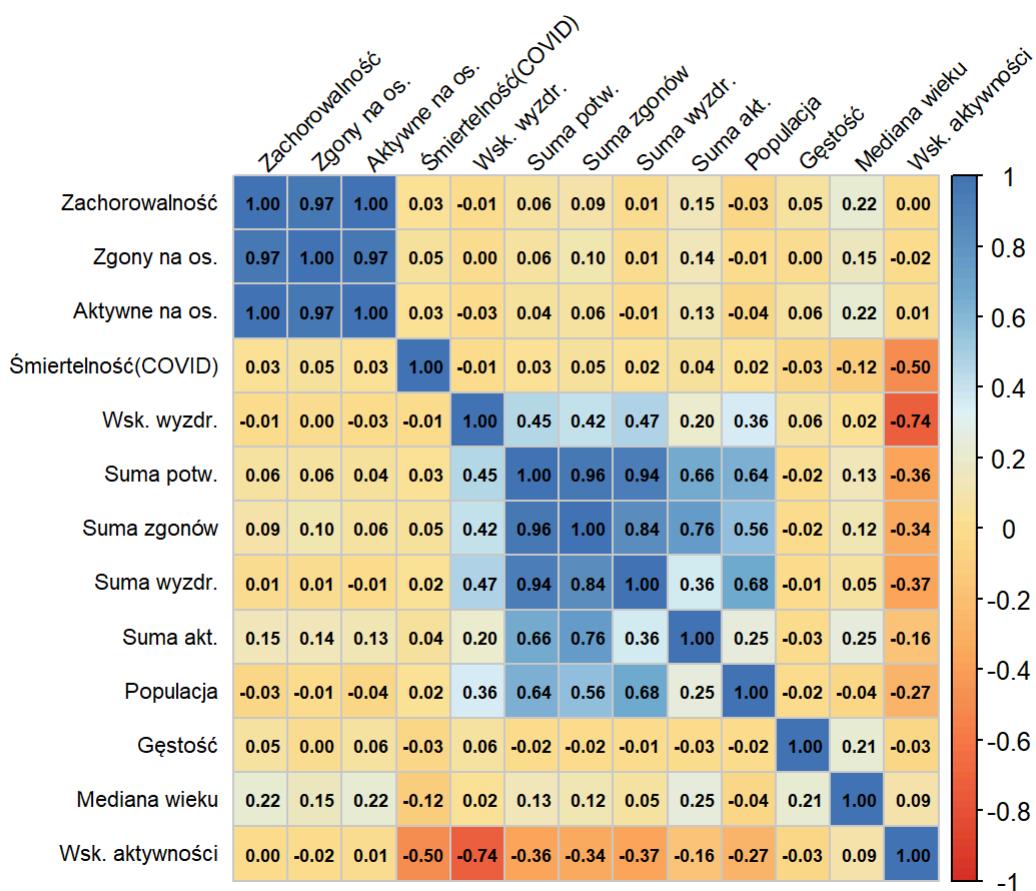
colnames(cor_matrix) <- c(
  "Zachorowalność",      # Incidence
  "Zgony na os.",        # Mortality per capita
  "Aktywne na os.",     # Active per capita
  "Śmiertelność(COVID)", # Mortality Rate (Rate/CFR)
  "Wsk. wyzdr.",         # Recovery Rate
  "Suma potw.",          # Total Confirmed
  "Suma zgonów",         # Total Deaths
  "Suma wyzdr.",         # Total Recovered
  "Suma akt.",           # Total Active
  "Populacja",           # Population
  "Gęstość",             # Density
  "Medianowa wieku",     # Median Age
  "Wsk. aktywności"     # Actve Rate
)
rownames(cor_matrix) <- colnames(cor_matrix)

my_palette <- colorRampPalette(c("#D73027", "#F46D43", "#FDAE61", "#FEE090", "#E0F3F8", "#74ADD1", "#4575B4"))(200)

corrplot(cor_matrix,
  method = "color",
  col = my_palette,
  addCoef.col = "black",
  tl.col = "black",
  tl.srt = 45,
  number.cex = 0.55,
  tl.cex = 0.7,
  number.digits = 2,
  addgrid.col = "grey80",
  mar = c(0, 0, 2, 0),
  title = "Macierz współczynników korelacji Pearsona"
)

```

Macierz współczynników korelacji Pearsona



Macierz korelacji Pearsona ujawnia trzy główne skupiska silnie powiązanych ze sobą zmiennych. Pierwsze z nich to wskaźniki relatywne, gdzie zachorowalność, liczba zgonów na mieszkańców oraz aktywne przypadki na osobę są ze sobą niemal idealnie skorelowane pozytywnie (wskaźniki 0,97–1,00). Druga grupa obejmuje sumaryczne dane bezwzględne, wykazujące bardzo silne korelacje dodatnie (0,76–0,96) między całkowitą liczbą potwierdzonych przypadków, zgonów oraz wyzdrowień. Istotnym faktem jest również wyraźny związek między wielkością populacji a sumarycznymi statystykami epidemiologicznymi, szczególnie w kontekście liczby wyzdrowień (0,68) i potwierdzonych zakażeń (0,64). Zmienne demograficzne, takie jak gęstość zaludnienia czy medianie wieku, wykazują bardzo słabe lub bliskie zeru koreacje z przebiegiem pandemii, co sugeruje, że na tym etapie czynniki te nie miały decydującego wpływu na ogólną dynamikę zakażeń.

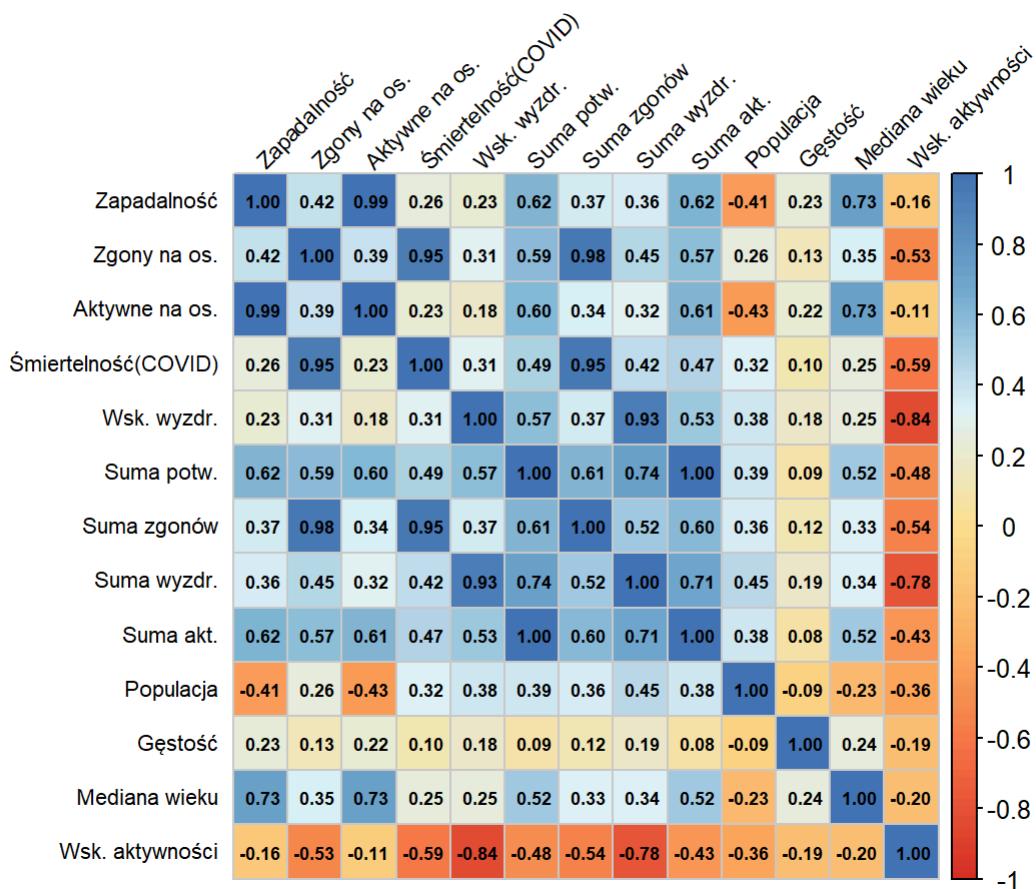
```
cor_matrix <- cor(vars_for_cluster, method = "spearman", use = "complete.obs")

colnames(cor_matrix) <- c(
  "Zapadalność",      # Incidence
  "Zgony na os.",     # Mortality per capita
  "Aktywne na os.",   # Active per capita
  "Śmiertelność(COVID)", # Mortality Rate (Rate/CFR)
  "Wsk. wyzdr.",       # Recovery Rate
  "Suma potw.",        # Total Confirmed
  "Suma zgonów",       # Total Deaths
  "Suma wyzdr.",       # Total Recovered
  "Suma akt.",         # Total Active
  "Populacja",         # Population
  "Gęstość",           # Density
  "Mediania wieku",    # Median Age
  "Wsk. aktywności"   # Active Rate
)
rownames(cor_matrix) <- colnames(cor_matrix)

my_palette <- colorRampPalette(c("#D73027", "#F46D43", "#FDAE61", "#FEE090", "#E0F3F8", "#74ADD1", "#4575B4"))(200)

corrplot(cor_matrix,
  method = "color",
  col = my_palette,
  addCoef.col = "black",
  tl.col = "black",
  tl.srt = 45,
  number.cex = 0.55,
  tl.cex = 0.7,
  number.digits = 2,
  addgrid.col = "grey80",
  mar = c(0, 0, 2, 0),
  title = "Macierz współczynników korelacji Spearmana"
)
```

Macierz współczynników korelacji Spearmana



Macierz korelacji Spearmana ujawnia szereg istotnych współzależności, wśród których najsilniejsze dodatnie związki występują pomiędzy parami wskaźników relatywnych a ich odpowiednikami bezwzględnymi. Niemal pełną zbieżność wykazują zapadalność i liczba aktywnych przypadków na mieszkańców (0,99), a także wskaźnik zgonów na mieszkańców z ogólną sumą zgonów (0,98) oraz śmiertelnością (0,95), co potwierdza wysoką spójność raportowanych danych. Kluczową informacją demograficzną jest wyraźna zależność między medianą wieku a intensywnością pandemii, gdzie starsze społeczeństwa korelują z wyższą zapadalnością i liczbą aktywnych przypadków na osobę (0,73). Ponadto, suma potwierdzonych przypadków jest nierozerwalnie związana z sumą aktywnych zakażeń (1,00), a wskaźnik wyzdrowień wykazuje silną korelację z całkowitą liczbą ozdrowieńców (0,93). Z kolei wielkość populacji wpływa umiarkowanie ujemnie na zapadalność (-0,41) i liczbę aktywnych chorych na osobę (-0,43), co sugeruje, że mniejsze kraje odnotowywały relatywnie wyższy odsetek zakażeń w stosunku do liczby mieszkańców w badanym okresie.

Wnioski z EDA

Przeprowadzona analiza pozwoliła na zidentyfikowanie kluczowych mechanizmów rządzących danymi epidemicznymi w początkowej fazie pandemii oraz wyłonienie zmiennych o największym potencjalnym informacyjnym dla procesu grupowania. Najważniejsze obserwacje wskazują na skrajną asymetrię rozkładów – podczas gdy większość świata odnotowywała pojedyncze przypadki, nieliczne kraje, takie jak Chiny czy Włochy, stały się potężnymi centrami infekcji. Analiza korelacji Spearmana udowodniła, że czynniki demograficzne, a w szczególności medianie wieku, miały istotny związek z tempem rozprzestrzeniania się wirusa w tamtym okresie.

Do hierarchicznej analizy skupień wytypowano następujące zmienne, które najlepiej różnicują kraje pod względem charakteru przebiegu pandemii i profilu demograficznego:

- **Incidence_Rate** (Wskaźnik zachorowalności): kluczowa zmienna relatywna, która normalizuje liczbę zakażeń względem wielkości populacji, pozwalając na obiektywne porównanie skali penetracji wirusa w różnych krajach, niezależnie od ich rozmiaru.
- **Mortality_Rate** (Wskaźnik zgonów wśród osób chorych): miara ta odzwierciedla rzeczywiste obciążenie demograficzne śmiertelnymi skutkami choroby, będąc jednocześnie silnie powiązaną z ogólną śmiertelnością

i sumą zgonów.

- MedianAge (Medianą wieku): parametr demograficzny o najwyższym współczynniku współzależności z danymi o zachorowaniach (0,73), niezbędny do pogrupowania krajów o podobnej strukturze wiekowej społeczeństwa, co jest istotnym czynnikiem ryzyka.
- Density (Gęstość zaludnienia): zmienna uzupełniająca profil demograficzny, pozwalająca odseparować gęsto zaludnione terytoria miejskie od krajów o rozproszonej strukturze osadnictwa, co potencjalnie wpływa na łatwość transmisji wirusa.
- Recovery_Rate

Zastosowanie tych wskaźników relatywnych zamiast surowych liczb bezwzględnych zapobiegnie dominacji grup przez największe mocarstwa (jak Chiny czy Indie) i umożliwi wykrycie podobieństw w charakterystyce epidemicznej krajów o różnej wielkości.

5. Przygotowanie zmiennych do modelu (transformacje)

Zbiór danych ograniczono do krajów, w których odnotowano co najmniej 10 potwierdzonych przypadków zakażenia. Krok ten podyktowany jest koniecznością eliminacji szumu statystycznego wynikającego z tzw. prawa małych liczb.

Przy skrajnie małej próbie wskaźniki relatywne stają się niestabilne i mogą prowadzić do błędnych wniosków na temat charakterystyki epidemii. Przykładowo, w kraju z zaledwie 1 potwierdzonym przypadkiem, który zakończył się zgonem, wskaźnik śmiertelności (CFR) wyniósłby matematycznie 100% (1/1). Taka wartość stanowiłaby sztuczną anomalię (outlier), która zaburzyłaby wyniki analizy skupień, sugerując ekstremalnie wysoką zjadliwość wirusa, podczas gdy w rzeczywistości wynik ten jest efektem braku reprezentatywnej próby. Dlatego zdecydowano o usunięciu krajów mających mniej niż 10 przypadków w kraju.

```
df_final = df_final[df_final$Total_confirmed>=10, ]
```

Na podstawie wniosków z eksploracyjnej analizy danych (EDA) oraz macierzy korelacji, do budowy modelu wybrano cztery kluczowe zmienne: `Incidence_Rate`, `Mortality_Rate`, `Recovery_Rate` oraz `Density`.

Ze względu na specyfikę danych – występowanie silnej asymetrii prawostronnej (“długie ogony”), dużą liczbę zer oraz obecność ekstremalnych wartości odstających – standardowe metody skalowania (jak standaryzacja Z-score czy proste skalowanie Min-Max) mogłyby nie przynieść zadowalających rezultatów. Wartości odstające spłaszczyłyby większość obserwacji do wąskiego przedziału, zacierając różnice między krajami.

W związku z tym zastosowano **podejście hybrydowe**:

- 1. Dla wskaźników epidemicznych (Rates):** Zastosowano transformację opartą na **Dystrybuancie Empirycznej (ECDF)**. Metoda ta zastępuje wartość bezwzględną jej rangą (percentylem) w rozkładzie wartości niezerowych. Dzięki temu zmienne zostają zmapowane do przedziału [0, 1], co pozwala na “rozciągnięcie” gęstych skupisk danych i ograniczenie wpływu wartości odstających, przy jednoczesnym zachowaniu porządku rangowego. Wartości zerowe pozostały bez zmian.
- 2. Dla gęstości zaludnienia (Density):** Zastosowano transformację logarytmiczną (`log1p`), aby zniwelować różnice rzędów wielkości, a następnie przeskaliwano wynik do przedziału [0, 1] metodą Min-Max.

Poniższy kod realizuje te przekształcenia oraz wizualizuje uzyskane rozkłady, które powinny być teraz bardziej zbliżone do rozkładu jednostajnego, co ułatwi algorytmom grupowania (opartym na odległościach) poprawne wyodrębnienie struktur.

```
selected_cols <- c("Incidence_Rate", "Mortality_Rate", "Recovery_Rate", "Density")
df_subset <- df_final[, selected_cols]

transform_rates_ecdf <- function(x) {
  x_new <- x
  pos_indices <- x > 0

  if (sum(pos_indices) > 0) {
    pos_values <- x[pos_indices]
    vals_ranked <- ecdf(pos_values)(pos_values)
    x_new[pos_indices] <- vals_ranked
  }
  return(x_new)
}

transform_density_log <- function(x) {
  x_log <- log1p(x)
  (x_log - min(x_log)) / (max(x_log) - min(x_log))
}

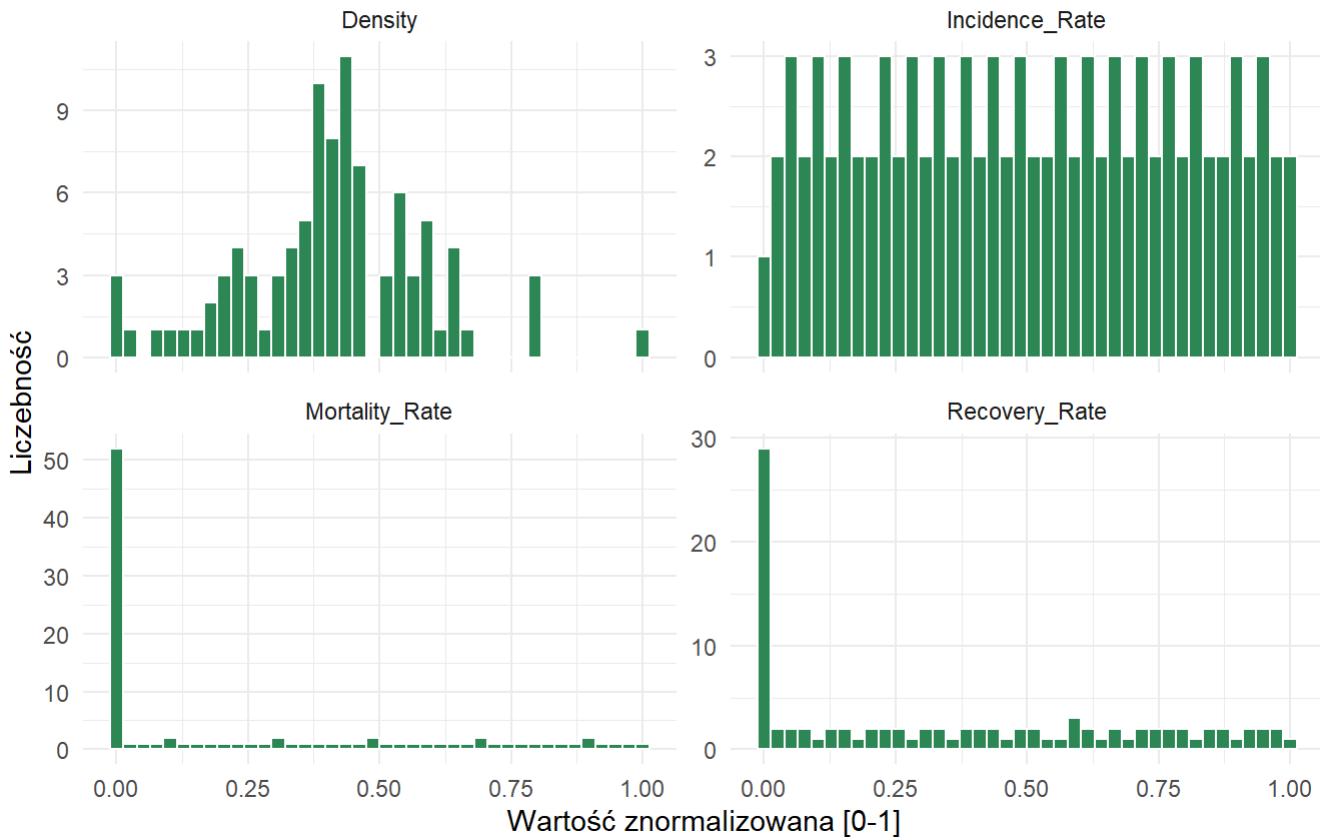
df_transformed <- df_subset %>%
  mutate(
    across(c("Incidence_Rate", "Mortality_Rate", "Recovery_Rate"), transform_rates_ecdf),
    across(c("Density"), transform_density_log)
  )

df_long <- df_transformed %>%
  pivot_longer(cols = everything(), names_to = "Zmienna", values_to = "Wartosc")

ggplot(df_long, aes(x = Wartosc)) +
  geom_histogram(fill = "seagreen", color = "white", bins = 40) +
  facet_wrap(~ Zmienna, scales = "free_y") +
  theme_minimal() +
  labs(title = "Podejście Hybrydowe",
       subtitle = "Density: Logarytm | Rates: Ranking ECDF",
       x = "Wartość znormalizowana [0-1]",
       y = "Liczebność")
```

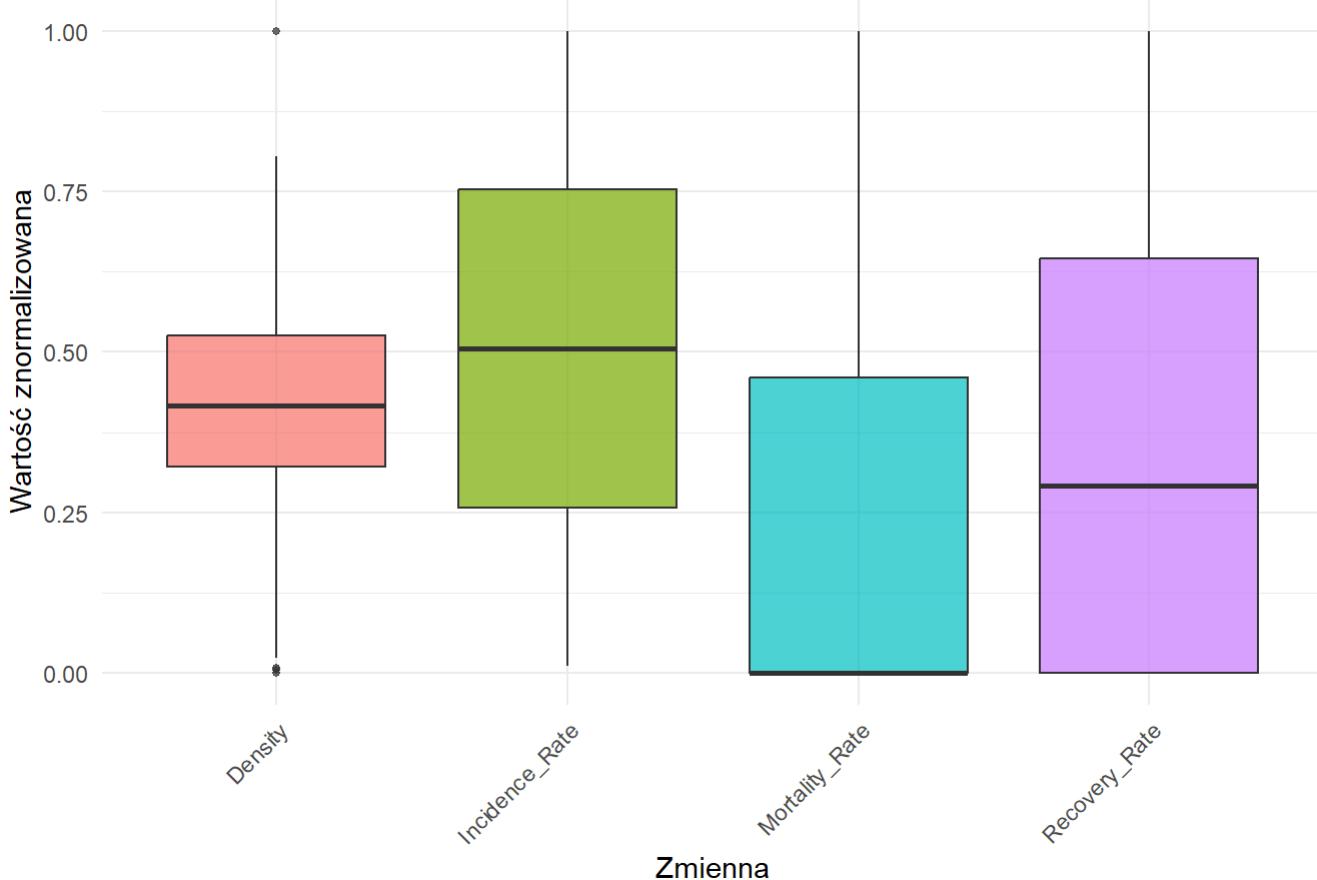
Podejście Hybrydowe

Density: Logarytm | Rates: Ranking ECDF



```
ggplot(df_long, aes(x = Zmienna, y = Wartosc, fill = Zmienna)) +
  geom_boxplot(alpha = 0.7, outlier.size = 1) +
  theme_minimal() +
  labs(title = "Weryfikacja ciągłości danych",
       y = "Wartość znormalizowana") +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45, hjust = 1))
```

Weryfikacja ciągłości danych



Analiza wygenerowanych wykresów potwierdza skuteczność zastosowanego podejścia hybrydowego w przygotowaniu danych do analizy skupień. Transformacja logarytmiczna zmiennej gęstości zaludnienia skutecznie zniwelowała jej silną prawostrońską skośność, nadając rozkładowi kształt zbliżony do normalnego, co jest optymalne dla algorytmów opartych na metryce euklidesowej. Zastosowanie transformacji opartej na dystrybuancie empirycznej (ECDF) dla wskaźników epidemicznych pozwoliło na ich równomierne rozłożenie w znormalizowanym przedziale [0, 1]. Dzięki temu zabiegowi zmaksymalizowano separację między krajami o różnym nasileniu pandemii, jednocześnie zachowując istotną informację o dużej grupie państw, w których nie odnotowano jeszcze zgonów lub wyzdrowień (wartości zerowe). Sprowadzenie wszystkich zmiennych do wspólnej skali zapobiega dominacji cech o wysokiej wariancji nad pozostałymi i pozwala na przejście do właściwego etapu grupowania danych.

6. Tworzenie modelu hierarchicznego

W celu wyznaczenia optymalnych parametrów analizy skupień opracowano skrypt realizujący algorytm pełnego przeszukiwania siatki (ang. Grid Search). Narzędzie to automatycznie testuje i porównuje efektywność modeli hierarchicznych, uwzględniając:

- Kombinacje parametrów: Przeanalizowano różne miary odległości (m.in. euklidesowa, manhattan) oraz metody łączenia klastrów (m.in. Warda, średnia) dla podziałów w zakresie od 2 do 20 grup.
- Walidację jakości: Do oceny każdego modelu wykorzystano zestaw trzech wskaźników: indeks Silhouette, Calińskiego-Harabasza oraz Daviesa-Bouldina, co pozwoliło na obiektywny wybór najlepszej konfiguracji.

```

dist_methods <- c("euclidean", "maximum", "manhattan", "canberra")
link_methods <- c("ward.D", "ward.D2", "single", "complete", "average", "mcquitty", "median", "centroid")
k_range <- 2:20

results_list <- list()
counter <- 1

calc_ch_index <- function(data, clusters) {
  n <- nrow(data)
  k <- length(unique(clusters))
  if (k < 2 || k == n) return(NA)

  center_global <- colMeans(data)
  tss <- sum(rowSums((sweep(data, 2, center_global))^2))

  wss <- 0
  for (i in unique(clusters)) {
    cluster_data <- data[clusters == i, , drop = FALSE]
    if (nrow(cluster_data) > 0) {
      center_cluster <- colMeans(cluster_data)
      wss <- wss + sum(rowSums((sweep(cluster_data, 2, center_cluster))^2))
    }
  }

  bss <- tss - wss
  ch <- (bss / (k - 1)) / (wss / (n - k))
  return(ch)
}

calc_db_index <- function(data, clusters) {
  u_clust <- unique(clusters)
  k <- length(u_clust)
  if (k < 2) return(NA)

  centers <- matrix(NA, nrow=k, ncol=ncol(data))
  S <- numeric(k)

  for(i in 1:k) {
    c_data <- data[clusters == u_clust[i], , drop=FALSE]
    if(nrow(c_data) > 0) {
      centers[i,] <- colMeans(c_data)
      dists <- sqrt(rowSums(sweep(c_data, 2, centers[i,])^2))
      S[i] <- mean(dists)
    } else {
      S[i] <- 0
    }
  }

  M <- as.matrix(dist(centers))
  diag(M) <- Inf

  R <- numeric(k)
  for(i in 1:k) {
    if(all(is.infinite(M[i,]))) {

```

```

R[i] <- 0
} else {
  R[i] <- max((S[i] + S) / M[i,], na.rm=TRUE)
}
}

return(mean(R[is.finite(R)]))
}

for (d_meth in dist_methods) {
  dist_matrix <- tryCatch({
    dist(df_transformed, method = d_meth)
  }, error = function(e) return(NULL))

  if (is.null(dist_matrix)) next

  for (l_meth in link_methods) {
    hc_model <- tryCatch({
      hclust(dist_matrix, method = l_meth)
    }, error = function(e) return(NULL))

    if (is.null(hc_model)) next

    for (k in k_range) {
      groups <- cutree(hc_model, k = k)

      sil_obj <- tryCatch({
        silhouette(groups, dist_matrix)
      }, error = function(e) return(NULL))

      avg_sil <- if (!is.null(sil_obj) && !all(is.na(sil_obj))) {
        mean(sil_obj[, 3])
      } else {
        NA
      }

      ch_val <- calc_ch_index(as.data.frame(df_transformed), groups)

      db_val <- calc_db_index(as.data.frame(df_transformed), groups)

      results_list[[counter]] <- data.frame(
        Dist_Method = d_meth,
        Link_Method = l_meth,
        K_Clusters = k,
        Avg_Silhouette = avg_sil,
        Calinski_Harabasz = ch_val,
        Davies_Bouldin = db_val,
        stringsAsFactors = FALSE
      )
      counter <- counter + 1
    }
  }
}

results_df <- do.call(rbind, results_list)
results_df <- results_df %>%

```

```
filter(!is.na(Calinski_Harabasz)) %>%
arrange(desc(Calinski_Harabasz))
```

```
head(results_df, 10)
```

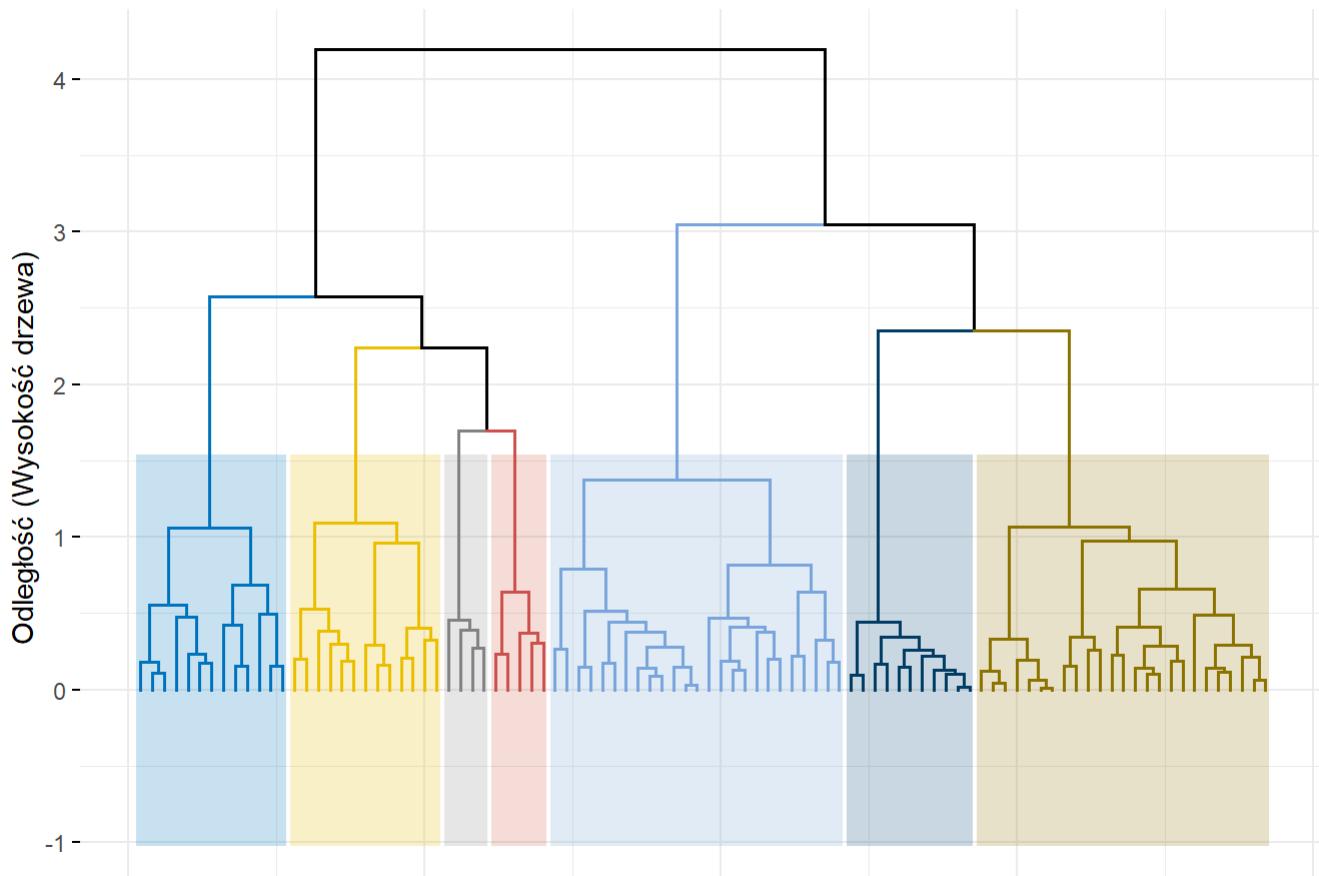
	Dist_Method	Link_Method	K_Clusters	Avg_Silhouette	Calinski_Harabasz
## 1	euclidean	ward.D	8	0.3229454	42.56727
## 2	euclidean	ward.D2	7	0.3350730	42.36622
## 3	euclidean	ward.D2	8	0.3197653	42.20434
## 4	maximum	ward.D	12	0.3239288	41.35987
## 5	maximum	ward.D2	3	0.2601840	41.27709
## 6	maximum	ward.D	11	0.3100868	41.12736
## 7	euclidean	ward.D2	6	0.3182528	41.08597
## 8	maximum	ward.D2	13	0.3273877	41.08410
## 9	manhattan	ward.D	7	0.3169714	41.01486
## 10	manhattan	ward.D2	7	0.3169714	41.01486
##	Davies_Bouldin				
## 1	0.9287859				
## 2	0.8793783				
## 3	0.9280217				
## 4	0.9282205				
## 5	1.2318448				
## 6	0.9148007				
## 7	0.9932808				
## 8	0.9057192				
## 9	0.9366231				
## 10	0.9366231				

Przeprowadzona walidacja metodą Grid Search wyłoniła optymalną konfigurację parametrów, wskazując na zastosowanie miary odległości euklidesowej oraz metody łączenia Warda (wariant D2) przy podziale na 7 skupień. Wybór ten podyktowany jest najlepszym balansem pomiędzy analizowanymi metrykami jakości. Mimo że podział na 8 skupień uzyskał minimalnie wyższy indeks Calińskiego-Harabasza (42,37 dla wybranego modelu wobec 42,57 dla alternatywy), to wybrana konfiguracja dominuje w kluczowych wskaźnikach spójności i separacji. Osiągnięto najwyższy w czołówce rankingowej wskaźnik Silhouette (0,34), co potwierdza większą stabilność przypisania obiektów do grup, oraz najniższą wartość indeksu Daviesa-Bouldina (0,88), dowodzącą lepszego odseparowania klastrów. Przewaga metryki euklidesowej w połączeniu z wariantyczną metodą Warda sugeruje, że o podobieństwie przebiegu pandemii między krajami decyduje całokształt parametrów i ich wzajemne relacje, dążące do tworzenia zwartych, jednorodnych grup, a nie pojedyncze skrajne odchylenia.

```
final_dist <- dist(df_transformed, method = "euclidean")
final_hc <- hclust(final_dist, method = "ward.D2")
```

```
fviz_dend(final_hc,
  k = 7,
  cex = 0.5,
  k_colors = "jco",
  color_labels_by_k = TRUE,
  rect = TRUE,
  rect_border = "jco",
  rect_fill = TRUE,
  show_labels = FALSE,
  ggtheme = theme_minimal(),
  main = "Dendrogram: Euclidean + Ward.D2 (k=7)"
) +
  labs(y = "Odległość (Wysokość drzewa)")
```

Dendrogram: Euclidean + Ward.D2 (k=7)



Wykres wizualizuje proces łączenia obiektów w coraz większe skupienia. Oś pionowa reprezentuje poziom odległości (niepodobieństwa), przy którym następuje fuzja grup – wyższe połączenia świadczą o większym zróżnicowaniu. Kolorowe obszary wyznaczają finalny podział zbioru danych na 7 klastrów, wskazany w procesie walidacji jako zapewniający najlepszy balans między spójnością a separacją grup.

```
df_final$Cluster <- as.factor(cutree(final_hc, k = 7))
table(df_final$Cluster)
```

```
##
## 1 2 3 4 5 6 7
## 25 13 13 25 4 11 5
```

Rozkład liczebności w 7 klastrach: Dane zostały podzielone na 7 grup o zróżnicowanej wielkości. Wyróżnić można trzy typy skupień:

- Grupy dominujące (Klastry 1 i 4) - dwie największe grupy liczące po 25 państw, stanowiące trzon analizy i reprezentujące najbardziej typowe wzorce z przebiegu początku pandemii.
- Grupy średnie (Klastry 2, 3, 6) - liczące od 11 do 13 państw, reprezentujące umiarkowanie liczne wzorce.
- Grupy niszowe (Klastry 5 i 7): składające się zaledwie z 4 i 5 państw. Są to prawdopodobnie kraje o specyficznej, odmiennej charakterystyce (tzw. outliers), które nie pasowały do głównych nurtów.

7. Hierarchicalna analiza skupień

```
cluster_summary <- df_final %>%
  group_by(Cluster) %>%
  summarise(
    Liczba_krajow = n(),
    Mediana_Incidence = median(Incidence_Rate, na.rm = TRUE),
    Mediana_Mortality = median(Mortality_Rate, na.rm = TRUE),
    Mediana_Recovery = median(Recovery_Rate, na.rm = TRUE),
    Mediana_Active = median(Active_Rate, na.rm = TRUE),
    Mediana_Density = median(Density, na.rm = TRUE),
    Mediana_Wiek = median(MedianAge, na.rm = TRUE),
    Mediana_MortalityperCapita = median(Mortality_per_capita, na.rm = TRUE),
    Mediana_stillSickperCapita = median(StillSick_per_capita, na.rm = TRUE)
  )

print(cluster_summary)
```

```
## # A tibble: 7 × 10
##   Cluster Liczba_krajow Mediana_Incidence Mediana_Mortality Mediana_Recovery
##   <fct>      <int>          <dbl>            <dbl>           <dbl>
## 1 1             25     0.00000608         0             0.0588
## 2 2             13     0.0000177          0.0223        0.00141
## 3 3             13     0.0000148          0.0327        0.175 
## 4 4             25     0.0000385          0             0
## 5 5              4     0.000102          0.00234       0.249 
## 6 6              11     0.000000964        0             0
## 7 7              5     0.000210          0.0569        0.0982
## # i 5 more variables: Mediana_Active <dbl>, Mediana_Density <dbl>,
## #   Mediana_Wiek <dbl>, Mediana_MortalityperCapita <dbl>,
## #   Mediana_stillSickperCapita <dbl>
```

Poniżej przedstawiono interpretację każdego z 7 klastrów. Należy podkreślić, że analizowane dane obejmują **pierwsze dwa miesiące pandemii**. Z tego względu wysoki odsetek aktywnych przypadków należy interpretować jako wskaźnik **niedawnej introdukcji wirusa** do danego kraju (zbyt krótki czas od infekcji, by wystąpił zgon lub wyzdrowienie).

```
print(df_final$Country.Region[df_final$Cluster == 1])
```

```
## [1] "Afghanistan"          "Australia"           "Belarus"
## [4] "Canada"               "Croatia"              "Finland"
## [7] "Georgia"              "Jamaica"              "Jordan"
## [10] "Kuwait"               "Latvia"               "Lithuania"
## [13] "Malaysia"              "Mexico"               "North Macedonia"
## [16] "Oman"                 "Pakistan"             "Romania"
## [19] "Russia"               "Saudi Arabia"        "Senegal"
## [22] "Serbia"               "Sri Lanka"            "United Arab Emirates"
## [25] "Vietnam"
```

- **Klaster 1 (n=25)** Jedna z dwóch najliczniejszych grup, niezwykle zróżnicowana geograficznie. Obejmuje państwa od Europy Środkowo-Wschodniej i Północnej (m.in. Finlandia, Litwa, Rosja, Chorwacja), przez Bliski Wschód (ZEA, Arabia Saudyjska, Kuwejt), aż po Azję i Pacyfik (Wietnam, Australia, Malezja).

- **Faza epidemii:** Grupę tę charakteryzuje skuteczne zarządzanie pierwszą falą zakażeń. Mimo odnotowania infekcji (obecność wyzdrowień ~5,9%), mediana śmiertelności wynosi 0.00%. Sugeruje to, że kraje te albo wprowadziły wczesne, restrykcyjne środki zapobiegawcze (np. Wietnam, Australia, Rosja), albo posiadają systemy opieki zdrowotnej (lub strukturę demograficzną np. w krajach Zatoki), które pozwoliły uniknąć nagłego skoku zgonów w pierwszych dwóch miesiącach pandemii.
- **Demografia:** Mimo zróżnicowania (od bogatej Kanady po rozwijający się Pakistan), jako grupa reprezentuje „średnią światową” strukturę wieku (mediana 36,2 lat). Jest to klaster państw, które w badanym oknie czasowym zdołały „spłaszczyć krzywą” skuteczniej niż ogniska krytyczne.

```
print(df_final$Country.Region[df_final$Cluster == 2])
```

```
## [1] "Albania"      "Bulgaria"     "Ecuador"      "France"
## [5] "Greece"       "Lebanon"      "Luxembourg"   "Martinique"
## [9] "Netherlands"  "Panama"       "Philippines"   "US"
## [13] "United Kingdom"
```

- **Klaster 2 (n=13)** Grupa obejmująca państwa, które stały się globalnymi centrami pandemii po wyjściu wirusa z Chin. W skład tego klastra wchodzą m.in. mocarstwa zachodnie (USA, Wielka Brytania, Francja, Holandia).
 - **Charakterystyka:** Mediana śmiertelności na poziomie 2,22% przy śladowym odsetku wyzdrowień (0,14%) odzwierciedla dramatyczną sytuację w Europie Zachodniej i USA w badanym okresie. Niski wskaźnik wyzdrowień wynika z faktu, że kraje te znajdowały się w szczytce fazy wzrostowej (zbyt wcześnie na masowe wyzdrowienia), a systemy raportowania koncentrowały się na testowaniu i liczeniu zgonów, zaniedbując statystyki ozdrowieńców.
 - **Demografia:** Są to populacje relatywnie starsze (mediana wieku 38,5 lat), co jest typowe dla krajów rozwiniętych. Zaawansowany wiek społeczeństwa w połączeniu z szeroką transmisją wirusa w USA i Europie Zachodniej bezpośrednio przełożył się na wysokie wskaźniki śmiertelności i przeciążenie służby zdrowia.

```
print(df_final$Country.Region[df_final$Cluster == 3])
```

```
## [1] "Algeria"      "Argentina"    "Azerbaijan"  "Egypt"      "Hungary"
## [6] "India"        "Indonesia"   "Iraq"        "Japan"      "Morocco"
## [11] "Poland"       "Taiwan*"     "Thailand"
```

- **Klaster 3 (n=13)** Niezwykle ciekawa, heterogeniczna grupa łącząca kraje azjatyckie, które jako pierwsze zetknęły się z wirusem (Japonia, Tajwan, Tajlandia), gęsto zaludnione państwa rozwijające się (Indie, Indonezja, Egipt) oraz kraje Europy Środkowo-Wschodniej (Polska, Węgry).

- **Faza epidemii:** Kluczowym wyróżnikiem tej grupy jest wysoki odsetek wyzdrowień (~17,5%). Wynika to z obecności państw azjatyckich, w których epidemia rozpoczęła się najwcześniej (styczeń), co pozwoliło na kliniczne zamknięcie wielu przypadków (wyzdrowienie) przed końcem badanego okna czasowego.
- **Wysoka śmiertelność (CFR):** Mediana śmiertelności na poziomie 3,27% (wyższa niż w zachodnim Klastrze 2) jest efektem złożonym. Wynika ona prawdopodobnie ze specyfiki demograficznej (bardzo stara populacja Japonii) oraz ograniczonej strategii testowania w krajach takich jak Polska, Węgry czy Indie w pierwszej fazie pandemii (testowanie głównie przypadków ciężkich/objawowych zawyża wyliczany wskaźnik śmiertelności).
- **Demografia:** Mimo obecności "starej" Japonii czy Europy Środkowej, statystyczna mediana wieku dla klastra jest niska (30,7 lat). Jest to spowodowane "odmłodzeniem" grupy przez kraje o ogromnych populacjach, takie jak Indie, Indonezja, Irak czy Egipt.

```
print(df_final$Country.Region[df_final$Cluster == 4])
```

```
## [1] "Armenia"           "Austria"          "Belgium"
## [4] "Bosnia and Herzegovina" "Brunei"           "Chile"
## [7] "Costa Rica"         "Cyprus"            "Czechia"
## [10] "Denmark"            "Estonia"           "French Guiana"
## [13] "Germany"            "Iceland"           "Ireland"
## [16] "Israel"              "Maldives"          "Moldova"
## [19] "Norway"              "Portugal"          "Qatar"
## [22] "Slovakia"           "Slovenia"          "Sweden"
## [25] "Switzerland"
```

- **Klaster 4 (n=25)** Druga z najliczniejszych grup, o wyraźnym profilu geograficznym i ekonomicznym. Stanowi trzon „drugiej fali” uderzeniowej, obejmując głównie bogate kraje Europy Zachodniej, Środkowej i Północnej (Niemcy, Austria, Szwajcaria, Belgia, Szwecja, Norwegia, Dania).

- **Faza epidemii:** Grupę tę definiuje faza gwałtownego przyrostu zakażeń (mediana aktywnych przypadków 99,6%). Wiele z tych państw (np. Islandia, Norwegia, Austria) odnotowało masowy import wirusa, co skutkowało nagłym wykryciem setek infekcji w bardzo krótkim czasie.
- **Niska śmiertelność (0.00%):** Mimo dużej liczby zakażeń, mediana śmiertelności dla klastra wynosi 0%. Jest to efekt specyficzny dla tej grupy krajów w tamtym okresie:
 - **Szerokie testowanie:** Kraje takie jak Niemcy czy Izrael od początku testowały szeroko, wyłapując dużo przypadków łagodnych (młodzi narciarze), co “rozcieńczyło” statystyki śmiertelności.
 - **Opóźnienie czasowe:** Epidemia znajdowała się w fazie tak dynamicznego wzrostu wykładniczego, że statystyki zgonów (które są opóźnione o 2-3 tygodnie względem infekcji) nie zdążyły się jeszcze uformować w analizowanym oknie czasowym.

```
print(df_final$Country.Region[df_final$Cluster == 5])
```

```
## [1] "Bahrain"           "Korea, South"        "Malta"             "Singapore"
```

- **Klaster 5 (n=4)** Elitarna grupa państw o specyficznej charakterystyce geograficznej i systemowej: Korea Południowa, Singapur, Bahrajn, Malta. Są to kraje (lub terytoria wyspiarskie/półwyspowe) o bardzo wysokim stopniu urbanizacji, które wdrożyły agresywne strategie kontroli wirusa.
- **Urbanizacja i Ryzyko:** Klaster wyróżnia się ekstremalną medianą gęstości zaludnienia (1748 osób/km²). Obecność w tej grupie Singapuru i Korei Południowej (Seul) wskazuje na obszary, gdzie ryzyko błyskawicznej transmisji wirusa było najwyższe.

- **Wczesna, skuteczna interwencja:** Wysoki wskaźnik wyzdrowień (24,9%) wynika z faktu, że Korea Południowa i Singapur były pierwszymi po Chinach epicentrami epidemii. Jednak w przeciwnieństwie do Europy (Klaster 2 i 7), kraje te dzięki masowemu testowaniu i śledzeniu kontaktów skutecznie izolowały chorych. (<https://www.theguardian.com/world/2020/mar/11/mass-testing-alerts-and-big-fines-the-strategies-used-in-asia-to-slow-coronavirus>)
- **Niska śmiertelność:** Mimo wczesnego wybuchu epidemii i dużej gęstości, mediana śmiertelności wynosi zaledwie 0,23%. Dowodzi to wydolności tamtejszych systemów opieki zdrowotnej, które nie zostały przeciążone tak jak we Włoszech czy Hiszpanii.

```
print(df_final$Country.Region[df_final$Cluster == 6])
```

```
## [1] "Bolivia"          "Brazil"           "Burkina Faso"
## [4] "Colombia"         "Dominican Republic" "Kazakhstan"
## [7] "Peru"              "South Africa"      "Tunisia"
## [10] "Turkey"            "Venezuela"
```

- **Klaster 6 (n=11)** Grupa obejmująca duże gospodarki wschodzące i kraje rozwijające się, głównie z Ameryki Łacińskiej (Brazylia, Kolumbia, Peru, Boliwia), Afryki (RPA, Tunezja, Burkina Faso) oraz Eurazji (Turcja, Kazachstan).
 - **Faza epidemii:** Jest to grupa, do której fala pandemii dotarła w trzeciej kolejności (po Azji i Europie Zachodniej). W analizowanym oknie czasowym kraje te znajdują się na samym początku krzywej epidemicznej – stąd 100% przypadków klasyfikowanych jest jako aktywne, a statystyki zgonów (mediana 0,00%) jeszcze się nie otworzyły.
 - **Demografia jako bufor:** Klaster ten wyróżnia się najmłodszą populacją w zestawieniu (mediana wieku 28,2 lat). Młoda struktura demograficzna w Turcji, Brazylii czy Kolumbii mogła w początkowej fazie sprawiać wrażenie łagodniejszego przebiegu infekcji (mniej hospitalizacji) w porównaniu do starzejcej się Europy, zanim wirus dotarł do grup wrażliwych.

```
print(df_final$Country.Region[df_final$Cluster == 7])
```

```
## [1] "China"    "Iran"     "Italy"    "San Marino" "Spain"
```

- **Klaster 7 (n=5)** Grupa definiująca “Strefę Zero” światowej pandemii w pierwszych dwóch miesiącach. Obejmuje źródło epidemii (Chiny) oraz pierwsze kraje, w których doszło do niekontrolowanej transmisji i załamania systemów opieki zdrowotnej: w Europie (Włochy, Hiszpania wraz z enklawą San Marino) oraz na Bliskim Wschodzie (Iran).
 - **Śmiertelność:** Klaster ten wyróżnia się drastycznie najwyższą medianą śmiertelności (5,69%). Jest to bezpośrednim skutkiem bycia “pierwszym” – kraje te musiały mierzyć się z patogenem bez wypracowanych procedur medycznych i leków. Masowy napływ pacjentów w krótkim czasie doprowadził do paraliżu szpitali (słynne obrazy z Lombardii czy Wuhan), co drastycznie podniósł śmiertelność.
 - **Demografia:** Jest to grupa o najstarszej strukturze demograficznej (mediana wieku 43,1 lat). Połączenie bardzo wysokiego odsetka seniorów we Włoszech i Hiszpanii z wczesną, masową transmisją wirusa, przełożyło się na tragiczną liczbę zgonów w tej grupie.
 - **Spójność grupy:** Zgrupowanie Chin (kraj, gdzie epidemia wygasła) z Włochami i Hiszpanią (gdzie epidemia wybuchała) wynika z faktu, że w analizowanym oknie czasowym tylko te państwa posiadały jednocześnie: bardzo dużą liczbę wykrytych przypadków, wysoką śmiertelność i funkcjonujący system raportowania zgonów.

```

df_map_plotly <- df_final %>%
  mutate(
    Cluster_Num = as.numeric(as.character(Cluster)),
    Hover_Text = paste0(
      "<b>", Country.Region, "</b>",
      "<br>Klaster: ", Cluster,
      "<br>Śmiertelność: ", round(Mortality_Rate * 100, 2), "%",
      "<br>Gęstość: ", round(Density, 0)
    )
  )

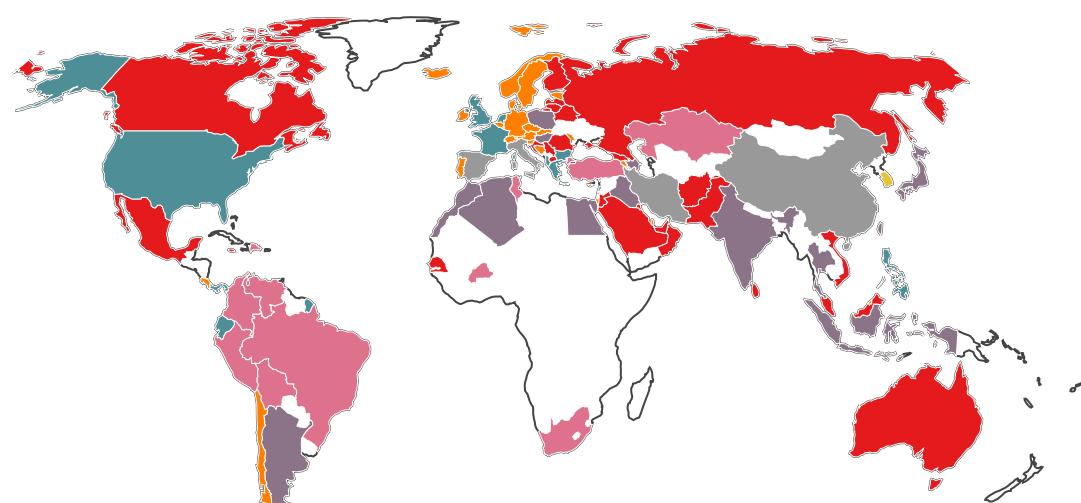
g <- list(
  showframe = FALSE,
  showcoastlines = TRUE,
  projection = list(type = 'natural earth')
)

fig <- plot_geo(df_map_plotly) %>%
  add_trace(
    z = ~Cluster_Num,
    locations = ~Iso3,
    locationmode = 'ISO-3',
    colors = "Set1",
    marker = list(
      line = list(color = "white", width = 0.5)
    ),
    text = ~Hover_Text,
    hoverinfo = "text",
    type = 'choropleth',
    showscale = FALSE
  ) %>%
  layout(
    title = '<b>Rozkład klastrów COVID-19 (Wypełnienie)</b>',
    geo = g
  )

fig

```

Rozkład klastrów COVID-19 (Wypełnienie)

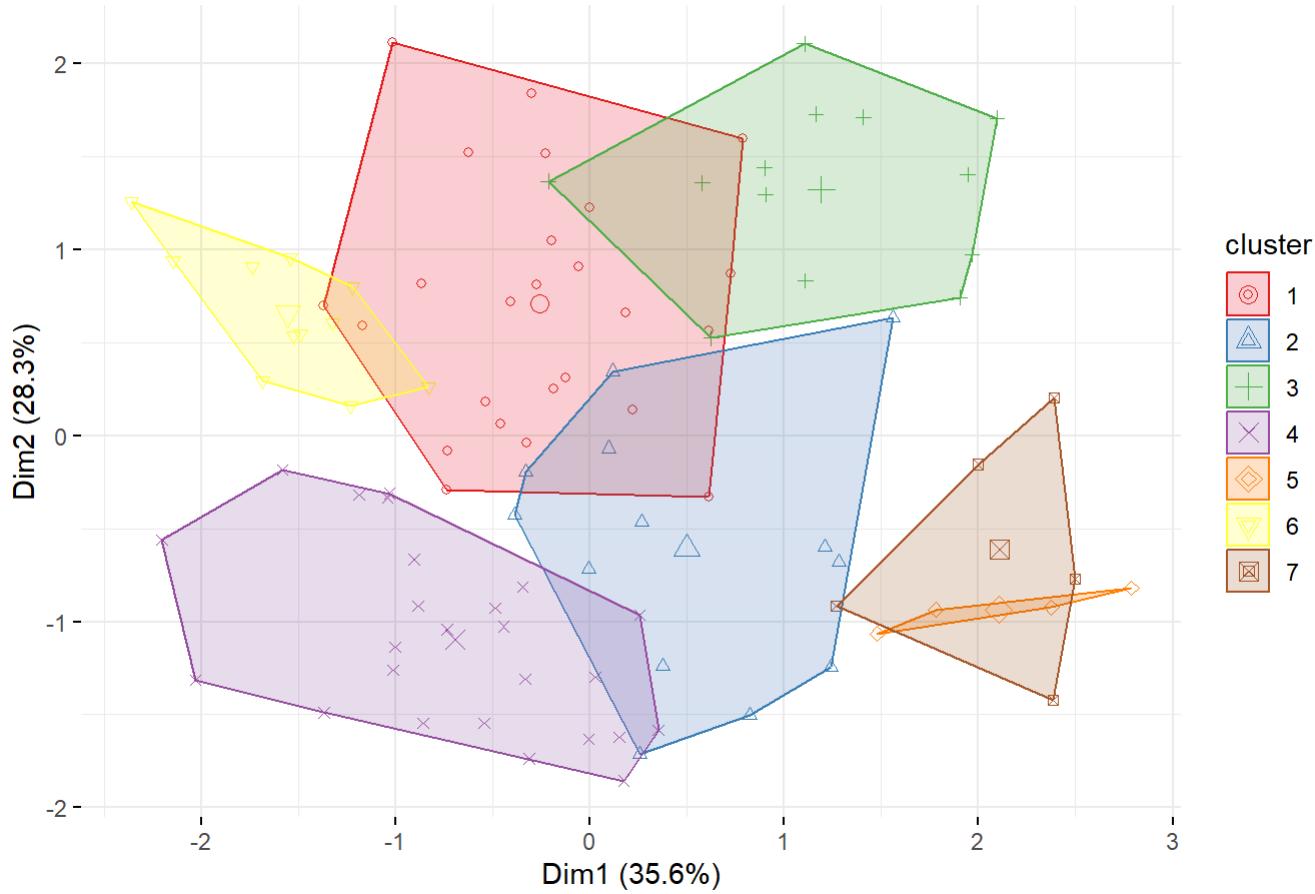




```
pca_data <- df_transformed
pca_result <- prcomp(pca_data, scale. = FALSE)

fviz_cluster(list(data = pca_data, cluster = df_final$Cluster),
             geom = "point",
             ellipse.type = "convex",
             palette = "Set1",
             ggtheme = theme_minimal(),
             main = "Wizualizacja separacji klastrów (PCA)")
```

Wizualizacja separacji klastrów (PCA)



Wykres prezentuje wizualizację wyników grupowania w przestrzeni dwóch pierwszych składowych głównych (PCA), które łącznie wyjaśniają 63,9% zmienności w zbiorze danych. Wymiar pierwszy odpowiada za 35,6% informacji, a drugi za 28,3%, co pozwala na wiarygodną ocenę relacji między grupami na płaszczyźnie. Rozmieszczenie punktów wskazuje na wyraźną separację klastrów periferyjnych (np. grupy 4, 5 i 6), które zajmują odrębne obszary, co potwierdza ich unikalną charakterystykę. W centrum wykresu widoczne jest częściowe nakładanie się klastrów 1 i 2, co sugeruje istnienie państw o cechach przejściowych. Mimo to większość grup tworzy zwarte skupiska, co dowodzi wysokiej jednorodności wewnętrz klastrów i potwierdza zasadność przyjętego podziału na 7 grup.

8. Algorytm K-średnich

W celu weryfikacji wyników uzyskanych metodą hierarchiczną oraz sprawdzenia stabilności wyodrębnionych struktur, w drugim etapie analizy zastosowano algorytm niehierarchiczny (podziałowy). Opracowano skrypt iteracyjny dla metody K-średnich (K-Means), który automatycznie testuje jakość podziału w poszukiwaniu optymalnej liczby skupień.

Procedura badawcza obejmowała:

- **Przeszukiwanie zakresu klastrów:** Przeanalizowano modele dla liczby grup (k) w przedziale od 2 do 20, stosując dla każdego przypadku 25 losowych inicjalizacji centroidów (`nstart = 25`) w celu uniknięcia optimów lokalnych.
- **WALIDACJĘ JAKOŚCI:** Do oceny każdego wariantu wykorzystano tożsame zestaw trzech wskaźników jakości: indeks Silhouette (spójność wewnętrz klastra), indeks Calińskiego-Harabasza (separacja między klastrami) oraz indeks Daviesa-Bouldina (podobieństwo klastrów), co pozwoliło na bezpośrednie porównanie efektywności obu metod grupowania.

```

set.seed(2022)

kmeans_results <- data.frame()
dist_matrix <- dist(df_transformed)

for (k in 2:20) {
  set.seed(123)
  km_fit <- kmeans(df_transformed, centers = k, nstart = 25)

  sil <- silhouette(km_fit$cluster, dist_matrix)
  sil_val <- mean(sil[, 3])

  ch_val <- calinhara(df_transformed, km_fit$cluster)

  centers <- km_fit$centers
  cluster_indices <- km_fit$cluster
  scatters <- numeric(k)

  for(i in 1:k) {
    cluster_data <- df_transformed[cluster_indices == i, , drop=FALSE]
    if(nrow(cluster_data) > 0) {
      dists_to_center <- sqrt(rowSums(sweep(cluster_data, 2, centers[i,], "-")^2))
      scatters[i] <- mean(dists_to_center)
    } else {
      scatters[i] <- 0
    }
  }

  center_dists <- as.matrix(dist(centers))

  R <- matrix(0, nrow = k, ncol = k)
  for(i in 1:k) {
    for(j in 1:k) {
      if(i != j) {
        R[i, j] <- (scatters[i] + scatters[j]) / center_dists[i, j]
      }
    }
  }
  db_val <- mean(apply(R, 1, max))

  kmeans_results <- rbind(kmeans_results, data.frame(
    K_Clusters = k,
    Avg_Silhouette = round(sil_val, 4),
    Calinski_Harabasz = round(ch_val, 2),
    Davies_Bouldin = round(db_val, 4)
  ))
}

kmeans_sorted <- kmeans_results[order(-kmeans_results$Avg_Silhouette), ]
head(kmeans_sorted, 10)

```

##	K_Clusters	Avg_Silhouette	Calinski_Harabasz	Davies_Bouldin
## 6	7	0.3590	45.65	0.8623
## 7	8	0.3506	46.71	0.9169
## 5	6	0.3427	44.10	0.9576
## 8	9	0.3406	45.72	0.9382
## 13	14	0.3329	42.75	0.8937
## 9	10	0.3323	44.96	0.9361
## 15	16	0.3312	41.77	0.8869
## 14	15	0.3298	41.60	0.8948
## 12	13	0.3252	43.04	0.9237
## 11	12	0.3235	42.85	0.9228

Przeprowadzona validacja metodą niehierarchiczną (K-Means) wyłoniła optymalną konfigurację podziału, wskazując na zasadność wyodrębnienia 7 skupień. Wybór ten podyktowany jest najlepszym balansem pomiędzy analizowanymi metrykami jakości, gdzie wariant ten wykazuje przewagę nad alternatywnymi podziałami w kluczowych aspektach spójności i separacji. Mimo że podział na 8 skupień uzyskał minimalnie wyższy indeks Calińskiego-Harabasza (46,71 dla alternatywy wobec 45,65 dla wybranego modelu), to konfiguracja 7-elementowa dominuje w pozostałych wskaźnikach decydujących o jakości klasteryzacji. Osiągnięto najwyższy w całym zestawieniu wskaźnik Silhouette (0,36), co potwierdza największą stabilność przypisania państw do grup, oraz najniższą wartość indeksu Daviesa-Bouldina (0,86), dowodzącą najlepszego odseparowania klastrów od siebie. Uzyskany wynik stanowi silną walidację krzyżową dla wcześniejszego modelu hierarchicznego (metoda Warda), który również wskazywał na istnienie 7 unikalnych wzorców przebiegu pandemii. Konwergencja wyników dwóch odmiennych matematycznie podejść – aglomeracyjnego (hierarchicznego) i podziałowego (K-Means) – sugeruje, że wyodrębniona struktura nie jest artefaktem obliczeniowym, lecz odzwierciedla naturalne, rzeczywiste zróżnicowanie strategii i skutków walki z wirusem w analizowanych krajach.

```
set.seed(2022)

kmeans_final <- kmeans(df_transformed, centers = 7, nstart = 25)

df_final$Cluster_kmeans <- as.factor(kmeans_final$cluster)

print(table(df_final$Cluster_kmeans))
```

```
##
##  1  2  3  4  5  6  7
## 24 17  5  4 13 23 10
```

Rozkład liczebności w 7 klastrach (Model K-Means)

Analiza liczebności grup w modelu K-Means wykazuje zbliżoną strukturę do modelu hierarchicznego, z wyraźnym podziałem na trzon analizy oraz grupy skrajne, jednak z lekkim przesunięciem akcentów w grupach średnich.

- **Grupy dominujące (Klastry 1 i 6):** Dwie najliczniejsze grupy, liczące odpowiednio 24 i 23 państwa (łącznie 49% populacji badania). Odpowiadają one “Grupom dominującym” z modelu hierarchicznego (tamte liczyły po 25 państw), co potwierdza istnienie dwóch głównych, powszechnych wzorców przebiegu pandemii na świecie.
- **Grupy średnie (Klastry 2, 5, 7):** Kategoria ta uległa największemu przetasowaniu. Obejmuje teraz trzy grupy o zróżnicowanej wielkości: Klaster 2 (17 państw), Klaster 5 (13 państw) oraz Klaster 7 (10 państw). W porównaniu do modelu hierarchicznego (gdzie grupy średnie były równiejsze, ok. 11-13 państw), algorytm K-Means wyodrębnił jedną silniejszą grupę “pół-dominującą” (Klaster 2).
- **Grupy niszowe (Klastry 3 i 4):** Najmniejsze skupienia, liczące zaledwie 5 i 4 państwa. Są to klasyczne outlieri – kraje o tak specyficznej charakterystyce (np. ekstremalna śmiertelność lub gęstość), że algorytm

konsekwentnie izoluje je od reszty świata, niezależnie od wybranej metody grupowania.

```
cluster_kmeans_summary <- df_final %>%
  group_by(Cluster_kmeans) %>%
  summarise(
    Liczba_krajow = n(),
    Mediana_Incidence = median(Incidence_Rate, na.rm = TRUE),
    Mediana_Mortality = median(Mortality_Rate, na.rm = TRUE),
    Mediana_Recovery = median(Recovery_Rate, na.rm = TRUE),
    Mediana_Active = median(Active_Rate, na.rm = TRUE),
    Mediana_Density = median(Density, na.rm = TRUE),
    Mediana_Wiek = median(MedianAge, na.rm = TRUE),
    Mediana_MortalityperCapita = median(Mortality_per_capita, na.rm = TRUE),
    Mediana_stillSickperCapita = median(StillSick_per_capita, na.rm = TRUE)
  )

print(cluster_kmeans_summary)
```

```
## # A tibble: 7 × 10
##   Cluster_kmeans  Liczba_krajow  Mediana_Incidence  Mediana_Mortality
##   <fct>           <int>            <dbl>             <dbl>
## 1 1                  24      0.00000699      0
## 2 2                  17      0.00000104      0
## 3 3                  5       0.000210       0.0569
## 4 4                  4       0.000102       0.00234
## 5 5                  13      0.0000137      0.0303
## 6 6                  23      0.0000921      0
## 7 7                  10      0.00000144      0.0336
## # i 6 more variables: Mediana_Recovery <dbl>, Mediana_Active <dbl>,
## #   Mediana_Density <dbl>, Mediana_Wiek <dbl>,
## #   Mediana_MortalityperCapita <dbl>, Mediana_stillSickperCapita <dbl>
```

9. Analiza skupień k-srednich

Poniżej przedstawiono interpretację każdego z 7 klastrów uzyskanych w wyniku metody k-means. Należy podkreślić, że analizowane dane obejmują **pierwsze dwa miesiące pandemii**. Z tego względu wysoki odsetek aktywnych przypadków należy interpretować jako wskaźnik **niedawnej introdukcji wirusa** do danego kraju (zbyt krótki czas od infekcji, by wystąpił zgon lub wyzdrowienie).

```
print(df_final$Country.Region[df_final$Cluster_kmeans==1])
```

```
## [1] "Afghanistan"          "Australia"           "Belarus"
## [4] "Canada"               "Croatia"              "Finland"
## [7] "Georgia"              "Jamaica"              "Jordan"
## [10] "Kuwait"               "Latvia"               "Lithuania"
## [13] "Malaysia"              "Mexico"               "North Macedonia"
## [16] "Oman"                 "Romania"              "Russia"
## [19] "Senegal"              "Serbia"               "Sri Lanka"
## [22] "Thailand"             "United Arab Emirates" "Vietnam"
```

- **Klaster 1 (n=24)** Grupa ta stanowi niemal wierną kopię Klastra 1 uzyskanego w metodzie hierarchicznej. Podobnie jak wcześniej, jest to najliczniejszy i najbardziej stabilny zbiór państw, obejmujący ten sam szeroki przekrój geograficzny (Europa Północna/Wschodnia, Bliski Wschód, Azja).

- **Tożsamość z modelem hierarchicznym:** Skład grupy pokrywa się w 96% z wynikami poprzedniej metody (jedyna zmiana to wymiana Pakistanu na Tajlandię). Dowodzi to, że wyodrębnienie tej grupy nie jest przypadkiem statystycznym, lecz odzwierciedla silny, rzeczywisty wzorzec.
- **Charakterystyka:** Opis pozostaje bez zmian – są to państwa, które w badanym okresie odnotowały zakażenia, ale utrzymały medianę śmiertelności na poziomie **0.00%**, skutecznie chroniąc system opieki zdrowotnej przed przeciążeniem w pierwszej fazie pandemii.

```
print(df_final$Country.Region[df_final$Cluster_kmeans==2])
```

```
## [1] "Bolivia"           "Bosnia and Herzegovina" "Brazil"
## [4] "Burkina Faso"     "Chile"                 "Colombia"
## [7] "Costa Rica"        "Dominican Republic"   "Kazakhstan"
## [10] "Moldova"           "Pakistan"              "Peru"
## [13] "Saudi Arabia"      "South Africa"         "Tunisia"
## [16] "Turkey"             "Venezuela"
```

- **Klaster 2 (n=17)** Grupa ta stanowi rozbudowaną wersję “Globalnego Południa” (odpowiednik Klastra 6 z metody hierarchicznej). Skupia głównie państwa rozwijające się z Ameryki Łacińskiej (Brazylia, Kolumbia, Peru, Chile, Wenezuela), Afryki (RPA, Burkina Faso, Tunezja) oraz wybrane kraje Eurazji (Turcja, Kazachstan, Pakistan, Arabia Saudyjska).

- **Przesunięcia względem modelu hierarchicznego:** Algorytm K-Means włączył do tej grupy państwa, które w metodzie hierarchicznej były rozproszone (np. Arabia Saudyjska i Pakistan z grupy “Wczesnej kontroli” czy Bośnia i Mołdawia z grupy europejskiej). Ich obecność tutaj wynika z czystej matematyki: w badanym oknie czasowym ich statystyki (sładowa liczba zgonów i niska wykrywalność) były bardziej zbliżone do profilu Brazylii czy Turcji niż do profilu państw o ustabilizowanej kontroli lub gwałtownej ekspansji.

```
print(df_final$Country.Region[df_final$Cluster_kmeans==3])
```

```
## [1] "China"      "Iran"       "Italy"      "San Marino" "Spain"
```

- **Klaster 3 (n=5)** Jest to dokładny odpowiednik Klastra 7 z metody hierarchicznej. Algorytm K-Means wyodrębnił identyczny zestaw pięciu państw: Chiny, Włochy, Hiszpania, Iran oraz San Marino. Potwierdza to bezsprzecznie, że w analizowanym oknie czasowym kraje te stanowiły odrębną, statystycznie unikalną kategorię – “Strefę Zero” światowej pandemii.

```
print(df_final$Country.Region[df_final$Cluster_kmeans==4])
```

```
## [1] "Bahrain"    "Korea, South" "Malta"      "Singapore"
```

- **Klaster 4 (n=4)** Grupa ta jest dokładnym odpowiednikiem Klastra 5 z metody hierarchicznej. Algorytm K-Means wyodrębnił ten sam zestaw czterech państw: Korea Południowa, Singapur, Bahrajn, Malta. Fakt, że te same kraje zostały zgrupowane razem przez obie metody, potwierdza ich unikalny status w skali globalnej – są to państwa o bardzo wysokiej urbanizacji, które poradziły sobie z epidemią w sposób odmienny od reszty świata.

```
print(df_final$Country.Region[df_final$Cluster_kmeans==5])
```

```
## [1] "Albania"      "Argentina"     "Bulgaria"      "Ecuador"
## [5] "France"       "Hungary"       "Lebanon"       "Martinique"
## [9] "Netherlands"  "Panama"        "Philippines"   "US"
## [13] "United Kingdom"
```

- **Klaster 5 (n=13)** Klaster ten stanowi odpowiednik “Głównych zachodnich ognisk” (Klastra 2) z metody hierarchicznej. Liczebność grupy pozostała identyczna (13 państw), a jej trzon nadal stanowią mocarstwa zachodnie (USA, Wielka Brytania, Francja, Holandia) oraz państwa zmagające się z kryzysem humanitarnym w Ameryce Łacińskiej (Ekwador, Panama).

- **Przesunięcia względem modelu hierarchicznego:** Algorytm K-Means dokonał tutaj interesującej korekty. Do grupy “zachodniej” (charakteryzującej się wysoką transmisją i śmiertelnością) włączył Węgry i Argentynę (które w metodzie hierarchicznej były w grupie “Mieszanej/Wczesnej ekspozycji”). Zastąpiły one Grecję i Luksemburg. Sugeruje to, że statystyki epidemiczne Węgier i Argentyny (wzrost zgonów i infekcji) były w rzeczywistości bliższe dramatycznej sytuacji we Francji czy USA, niż sugerowałoby to proste drzewo hierarchiczne.

```
print(df_final$Country.Region[df_final$Cluster_kmeans==6])
```

```
## [1] "Armenia"      "Austria"      "Belgium"      "Brunei"
## [5] "Cyprus"       "Czechia"      "Denmark"      "Estonia"
## [9] "French Guiana" "Germany"     "Greece"       "Iceland"
## [13] "Ireland"      "Israel"       "Luxembourg"   "Maldives"
## [17] "Norway"       "Portugal"    "Qatar"        "Slovakia"
## [21] "Slovenia"     "Sweden"      "Switzerland"
```

- **Klaster 6 (n=23)** Grupa ta stanowi “oczyszczoną” i bardziej homogeniczną wersję Klastra 4 z metody hierarchicznej. Skupia głównie bogate kraje Europy Zachodniej, Środkowej i Północnej (Niemcy, Austria, Szwajcaria, kraje skandynawskie) oraz państwa o wysokim dochodzie poza Europą (Izrael, Katar).

- **Przesunięcia względem modelu hierarchicznego:** Algorytm K-Means dokonał konsolidacji grupy pod kątem ekonomicznym – usunięto państwa rozwijające się o niższym potencjale testowania (Bośnia, Mołdawia, Chile, Kostaryka), przesuwając je do Klastra 2. Jednocześnie do grupy dołączono Grecję i Luksemburg, których profil statystyczny (kontrolowana śmiertelność przy rosnącej liczbie testów) okazał się bliższy modelowi niemieckiemu czy norweskiemu niż mocarstwom z Klastra 5.

```
print(df_final$Country.Region[df_final$Cluster_kmeans==7])
```

```
## [1] "Algeria"      "Azerbaijan"   "Egypt"        "India"        "Indonesia"
## [6] "Iraq"         "Japan"        "Morocco"     "Poland"      "Taiwan*"
```

- **Klaster 7 (n=10)** Klaster ten stanowi bezpośredni odpowiednik Klastra 3 z metody hierarchicznej. Obejmuje państwa, które albo jako jedne z pierwszych zetknęły się z wirusem (Japonia, Tajwan), albo są gęsto zaludnionymi krajami rozwijającymi się (Indie, Indonezja, Egipt, Algieria), uzupełnione o kraje Europy Środkowej (Polska).

- **Przesunięcia względem modelu hierarchicznego:** Algorytm K-Means wyłączył z tej grupy Węgry** i Argentynę (przesunięte do “zachodnich ognisk” w Klastrze 5) oraz Tajlandię (przesuniętą do “wczesnej kontroli” w Klastrze 1). Dzięki temu Klaster 7 stał się bardziej spójny pod kątem statystycznym, skupiając państwa o wysokiej medianie śmiertelności przy specyficzny profilu demograficznym.

```

df_map_plotly <- df_final %>%
  mutate(
    Cluster_Num = as.numeric(as.character(Cluster_kmeans)),
    Hover_Text = paste0(
      "<b>", Country.Region, "</b>",
      "<br>Klaster: ", Cluster_kmeans,
      "<br>Śmiertelność: ", round(Mortality_Rate * 100, 2), "%",
      "<br>Gęstość: ", round(Density, 0)
    )
  )

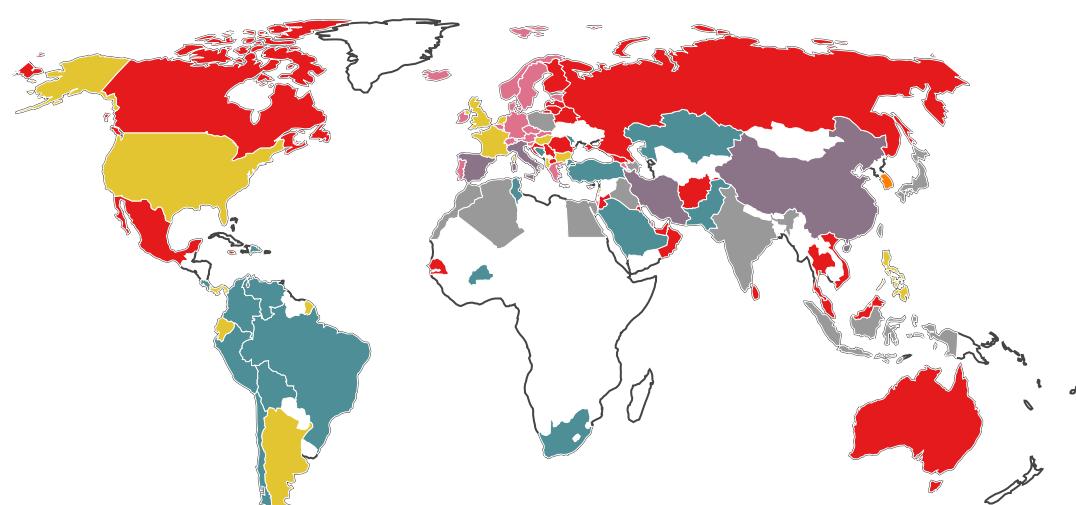
g <- list(
  showframe = FALSE,
  showcoastlines = TRUE,
  projection = list(type = 'natural earth')
)

fig <- plot_geo(df_map_plotly) %>%
  add_trace(
    z = ~Cluster_Num,
    locations = ~Iso3,
    locationmode = 'ISO-3',
    colors = "Set1",
    marker = list(
      line = list(color = "white", width = 0.5)
    ),
    text = ~Hover_Text,
    hoverinfo = "text",
    type = 'choropleth',
    showscale = FALSE
  ) %>%
  layout(
    title = '<b>Rozkład klastrów COVID-19 (K-means)</b>',
    geo = g
  )

fig

```

Rozkład klastrów COVID-19 (K-means)

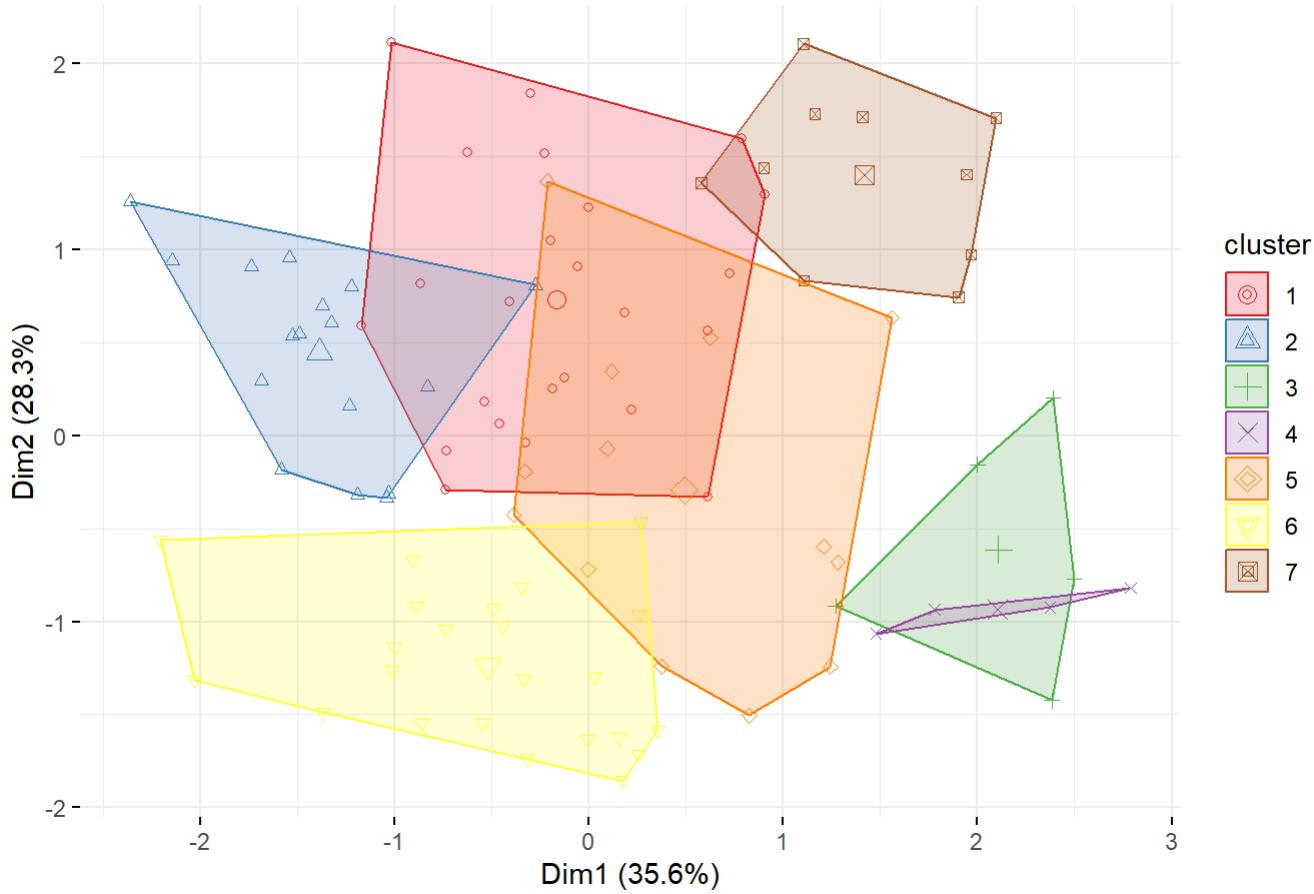




```
pca_data <- df_transformed
pca_result <- prcomp(pca_data, scale. = FALSE)

fviz_cluster(list(data = pca_data, cluster = df_final$Cluster_kmeans),
             geom = "point",
             ellipse.type = "convex",
             palette = "Set1",
             ggtheme = theme_minimal(),
             main = "Wizualizacja separacji klastrów (PCA) z K-means")
```

Wizualizacja separacji klastrów (PCA) z K-means



Wykres prezentuje rzutowanie wyników grupowania metodą K-średnich na przestrzeń dwuwymiarową, przy zachowaniu 63,9% całkowitej zmienności danych. Wizualizacja uwydacznia charakterystyczną dla tego algorytmu tendencję do tworzenia zwartych i geometrycznie regularnych skupień. Analiza przestrzenna potwierdza wysoką jakość podziału, gdzie grupy peryferyjne są wyraźnie odseparowane od reszty stawki, co świadczy o ich unikalnym profilu. W centralnym zagęszczeniu wykresu algorytm wyznaczył ostre granice między grupami dominującymi, minimalizując strefy nakładania się, co potwierdza stabilność wyodrębnionych wzorców epidemicznych i zasadność przyjęcia modelu siedmioelementowego.

10. Porównanie modeli i wnioski końcowe

Celem niniejszego badania była identyfikacja ukrytych struktur w danych dotyczących wczesnej fazy pandemii COVID-19 (luty-marzec 2020) przy użyciu metod uczenia nienadzorowanego. Przeprowadzono analizę porównawczą dwóch odmiennych podejść: hierarchicznej analizy skupień (metoda aglomeracyjna) oraz algorytmu K-średnich (metoda podziałowa).

- **Stabilność i konwergencja wyników** Obie zastosowane metody, mimo fundamentalnych różnic w mechanizmie działania, wskazały na istnienie 7 unikalnych wzorców (klastrów) przebiegu pandemii.
 - **Wysoka zgodność w grupach skrajnych:** Klastry definiujące "Strefę Zero" (Chiny, Włochy, Hiszpania, Iran) oraz "Model Azjatycki/Wysokiej Gęstości" (Korea Płd., Singapur) okazały się identyczne w obu modelach. Świadczy to o tym, że kraje te charakteryzowały się tak silnym i unikalnym sygnałem w danych, że sposób grupowania nie miał wpływu na ich klasyfikację.
 - **Przesunięcia w grupach dominujących:** Różnice między modelami ujawniły się w grupach "środkowej" (kraje o umiarkowanym przebiegu epidemii). Algorytm K-Means, dążąc do minimalizacji wariancji wewnętrzklastrowej, dokonał korekty przypisania państw "granicznych" (np. Grecja, Węgry, Pakistan), tworząc bardziej homogeniczne grupy pod kątem ekonomicznym i geograficznym.
- **Interpretacja epidemiologiczna** Analiza pozwoliła na wyodrębnienie kluczowych strategii i faz rozwoju pandemii w badanym oknie czasowym:
 - **Czynnik czasu (Lag):** Wyraźnie wyodrębniono grupę "Globalnego Południa" (Ameryka Łacińska, Afryka), gdzie w marcu 2020 r. pandemia była opóźniona względem Europy. Brak zgonów w tej grupie nie wynikał z odporności, lecz z wczesnego etapu krzywej epidemicznej.
 - **Czynnik demograficzny:** Potwierdzono związek między strukturą wiekową a śmiertelnością. Najtragiczniejsze żniwo (klastry "włoskie") odnotowano w najstarszych społeczeństwach, podczas gdy kraje młode przechodziły fazę początkową łagodniej w statystykach.
 - **Rola gęstości zaludnienia:** Wyodrębnienie osobnego klastra dla Singapuru i Korei Południowej udowodniło, że wysoka urbanizacja nie musi oznaczać klęski epidemicznej, jeśli idzie w parze z agresywną strategią testowania i izolacji (co odróżniło te kraje od Europy).
- **Ocena przydatności metod**
 - **Model hierarchiczny** okazał się bardziej użyteczny w fazie eksploracyjnej, umożliwiając (dzięki dendrogramowi) zrozumienie relacji i poziomu podobieństwa między poszczególnymi grupami państw.
 - **Model K-Means** zapewnił lepszą separację grup (wyższy indeks Silhouette: 0.36 vs 0.34) i stworzył bardziej zwarte, logiczne biznesowo segmenty, co czyni go lepszym narzędziem do ewentualnej dalszej analizy predykcyjnej.
- **Ograniczenia badania** Należy podkreślić, że wyniki analizy są ściśle powiązane z oknem czasowym (do 22 marca 2020). Wysoki odsetek wyzdrowień w klastrach azjatyckich wynikał z faktu, że epidemia dotarła tam najwcześniej. Wnioski te nie są predykcją dalszego rozwoju pandemii, lecz "zdjęciem" globalnej sytuacji w momencie, gdy świat wchodził w fazę pierwszego globalnego lockdownu.

Podsumowując, analiza skupień skutecznie zredukowała złożoność wielowymiarowych danych epidemicznych, pozwalając na kategoryzację 96 państw do 7 czytelnych profili, co ułatwia zrozumienie globalnej dynamiki zagrożenia w początkowej fazie kryzysu.