WYŻSZA SZKOŁA
INFORMATYKI i ZARZĄDZANIA
z siedzibą w Rzeszowie

KOLEGIUM INFORMATYKI STOSOWANEJ

Kierunek: INFORMATYKA

Grupa: 1IIZ/2023 GP01

Nikodem Krupa - nr albumu: w67246
Patrycja Ciurko - nr albumu: w67223

Analiza i wizualizacja danych o zdobytych medalach olimpijskich z lat 1896-2014

Prowadzący: dr Zofia Matusiewicz

Projekt

Rzeszów 2024

Spis treści

1 Wstęp

Sport i olimpiady są nieodłącznym elementem ludzkiej historii, symbolizując dążenie do doskonałości, rywalizację i jedność międzynarodową. Od czasów starożytnych igrzysk w Grecji po współczesne letnie i zimowe igrzyska olimpijskie, historia olimpiady jest bogata w emocje, triumfy i rekordy. Jednym z najbardziej fascynujących aspektów historii olimpiady są medale zdobyte przez sportowców z różnych krajów na przestrzeni lat. Analiza danych dotyczących zdobytych medali może dostarczyć cennych spostrzeżeń na temat wzorców, trendów i ewolucji w światowym sporcie.

2 Cel projektu

Celem tego projektu jest przeprowadzenie szczegółowej analizy danych dotyczących zdobytych medali olimpijskich z lat 1896-2014 oraz stworzenie wizualizacji. Poprzez zastosowanie technik analizy danych, takich jak analiza czasowa i geograficzna, projekt ma na celu identyfikację dominujących krajów oraz zmian w dystrybucji medali Polski w różnych okresach historycznych. Dodatkowo, poprzez stworzenie wizualizacji danych, projekt ma na celu umożliwienie użytkownikom zgłębienia danych olimpijskich w sposób intuicyjny i przystępny, co może być cenną pomocą dla badaczy, dziennikarzy sportowych oraz pasjonatów sportu.

3 Diagram przypadków użycia

Rysunek 1 przedstawia diagram akcji naszego programu. Algorytm po uruchomieniu wczytuje dane o medalach z igrzysk olimpijskich z lat 1896-2014. Przygotowywuje te dane, tzn. filtruje je, grupuje oraz sortuje. Kod analizuje dane, np. obliczając ilość medali dla każdego kraju, tylko dla kobiet. Na koniec wizualizuje dane w postaci wykresów słupkowych, liniowych oraz map i zapisuje w wskazanym przez nas folderze.

Rysunek 1: Diagram akcji. Źródło: własne

4 Instrukcja obsługi programu

4.1 Program

Do analizy tych danych wykorzystujemy język programowania R, który jest szeroko stosowany w analizie danych i statystyce. R jest niezwykle elastycznym i potężnym narzędziem, które pozwala na efektywne przetwarzanie i analizę dużych zestawów danych, a także tworzenie atrakcyjnych wizualizacji.

4.2 Paczki programu RStudio

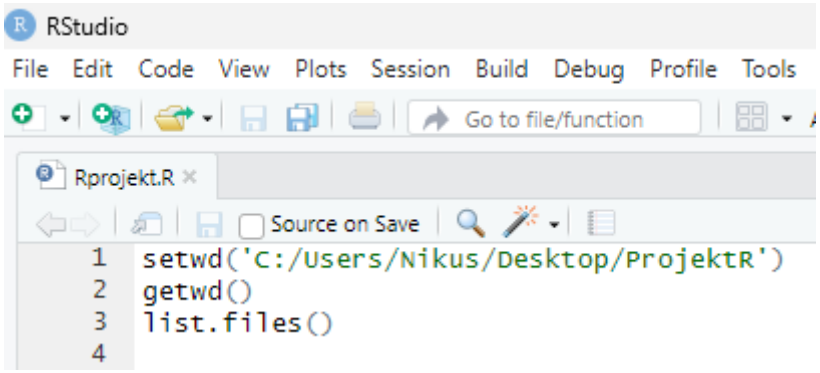
W naszym projekcie użyliśmy następujących paczek:

- dplyr - pakiet dplyr w R to narzędzie do szybkiej i czytelnej manipulacji danych w ramkach danych, ułatwiające filtrowanie, wybieranie, modyfikowanie i agregowanie danych [2].
- forcats - pakiet umożliwia efektywne zarządzanie zmiennymi kategorialnymi w ramkach danych poprzez funkcje ułatwiające sortowanie, filtrowanie i przekształcanie faktorów, co jest istotne przy analizie i modelowaniu danych kategorycznych [2].
- colorspace - jest to narzędzie do manipulacji przestrzeniami kolorów, umożliwiające konwersję między różnymi modelami kolorów oraz generowanie palet kolorów o różnych właściwościach, co jest przydatne w wizualizacji danych i grafice komputerowej [2].
- ggplot2 - pakiet to potężne narzędzie do tworzenia wykresów i wizualizacji danych, oparte na gramatyce grafiki (Grammar of Graphics), co pozwala na intuicyjne tworzenie grafik poprzez warstwowe dodawanie elementów, jak punkty, linie czy prostokąty, do wykresu [2].
- lubridate - pakiet ułatwia manipulację danymi zawierającymi informacje o dacie i czasie poprzez zapewnienie funkcji do łatwego parsowania, tworzenia oraz operowania danymi datowymi i czasowymi, co znacząco ułatwia analizę danych związanych z czasem [2].
- plotrix - zestaw funkcji do tworzenia różnorodnych wykresów, takich jak histogramy, wykresy kołowe czy wykresy punktowe, oraz do dodawania różnorodnych elementów do istniejących wykresów, takich jak etykiety czy linie pomocnicze [6].
- purrr - pakiet purrr dostarcza funkcji do pracy z funkcjami oraz strukturami danych w sposób konsekwentny i zgodny z paradygmatem funkcyjnym, ułatwiając iterację, mapowanie i operacje na listach, ramkach danych oraz innych obiektach [2].
- rcolorbrewer - pakiet zapewnia dostęp do palet kolorów opracowanych przez ColorBrewer, co umożliwia wybór estetycznych i odpowiednio dobranych kombinacji kolorów do wykresów [5].
- readr - narzędzie do szybkiego i efektywnego odczytu danych prostokątnych, takich jak pliki CSV, TSV czy FWF [2].
- rvest - pakiet pozwala na odczytywanie stron HTML za pomocą funkcji read_html(), a następnie wyszukiwanie elementów pasujących do selektora CSS lub wyrażenia XPath za pomocą funkcji html_elements() [1].
- rworldmap - pakiet rworldmap w R Studio umożliwia mapowanie danych na poziomie krajowym i siatkowym, ułatwiając dołączanie nowoczesnych map świata i oferując opcje wizualizacji [4].
- sp - pakiet dostarcza klasy i metody dla danych przestrzennych, dokumentując miejsce, w którym znajdują się informacje o lokalizacji przestrzennej, dla danych 2D lub 3D [3].

- stringr - pakiet dostarczający zestaw spójnych narzędzi do pracy z łańcuchami znaków, ułatwiając zadania związane z czyszczeniem i przygotowaniem danych [2].
- tibble - to ulepszona wersja standardowej ramki danych, dostarczająca łatwiejszego i bardziej spójnego sposobu przechowywania danych, zachowująca większ [2].
- tidyr - pakiet dostarcza funkcji do przekształcania danych, umożliwiając ich uporządkowanie w sposób zgodny z zasadami "tidy data"[2].
- tidyverse - zbiór pakietów, w tym dplyr, ggplot2, tidyr, purrr i inne, które współpracują ze sobą i stosują zasady tidy data, ułatwiając manipulację, wizualizację i analizę danych poprzez konsekwentne i spójne podejście [2].

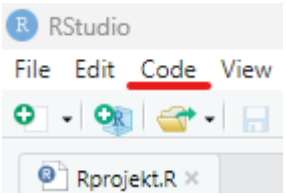
4.3 Instrukcja uruchomienia

1. Aby uruchomić program, należy wpisać ścieżkę plików do kodu w RStudio (w naszym przypadku wygląda to jak na Rysunku 2).

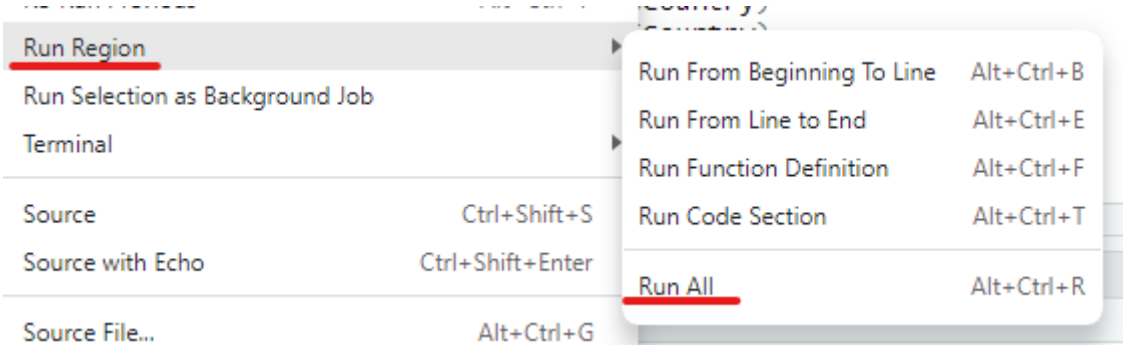


Rysunek 2: Pasek zadań - zakładka run region. Żródło: własne

2. Następnie, należy przejść w Rstudio do zakładki code (Rysunek 3), rozwinąć ją, znaleźć opcję "Run Region"oraz "Run All"tak, jak na Rysunku 4.

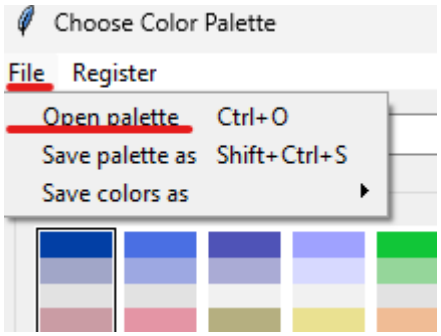


Rysunek 3: Pasek zadań - zakładka "code". Żródło: własne



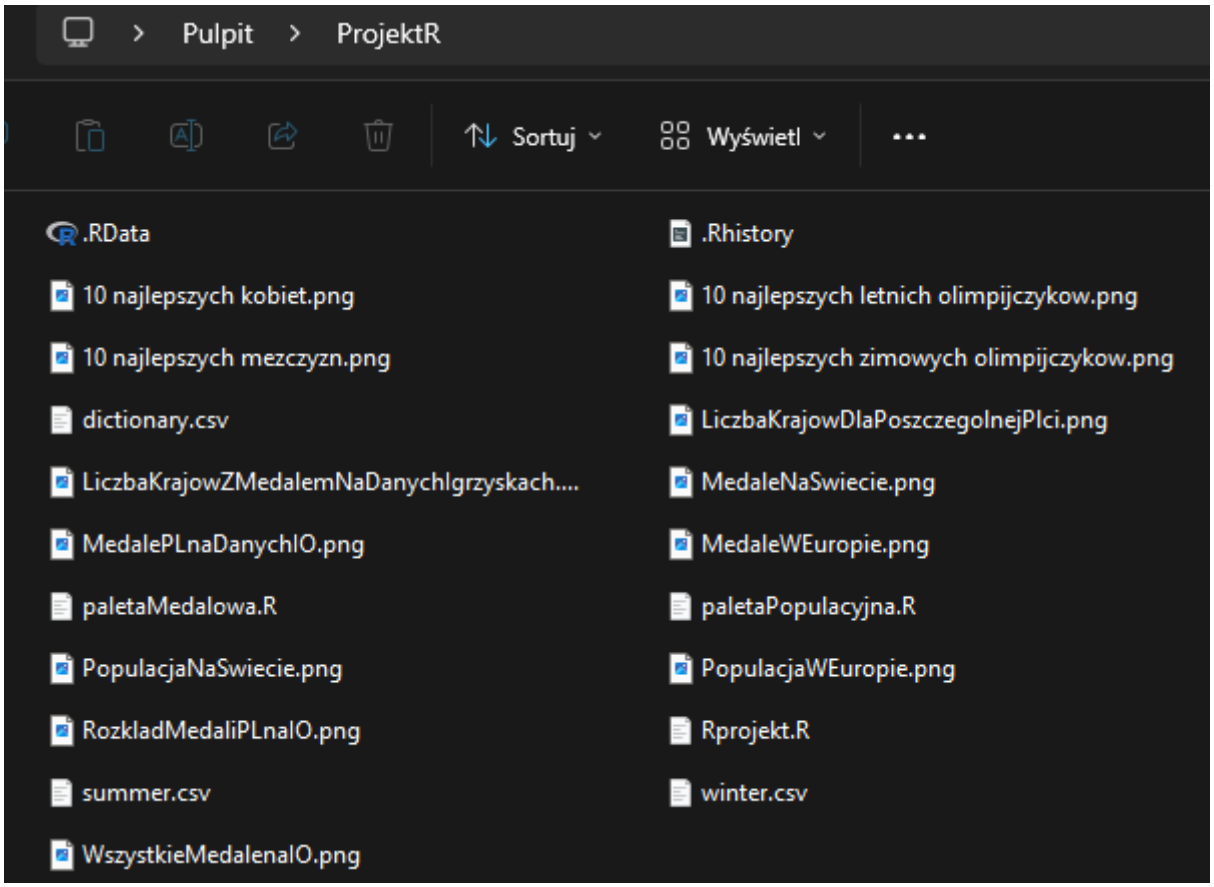
Rysunek 4: Pasek zadań - zakładka "run region". Żródło: własne

3. Po kilku sekundach powinno nam otworzyć się okno z wyborem palet kolorów, w którym musimy kliknąć "File", nastepnie "Open palette"(Rysunek 5). Wybieramy paletę medalową i klikamy "OK", aby zatwierdzić wybraną paletę. Po chwili otworzy się drugie okienko, postępujemy tak samo, jak poprzednio (Rysunek 5), jednak tym razem wybrieramy paletę populacyjną.



Rysunek 5: Paleta kolorów. Źródło: własne

4. Jeżeli w konsoli wyświetla się "null device 1", oznacza to, że program zakończył działanie oraz nasze wykresy są gotowe w folderze, który wybraliśmy na początku (Rysunek 2).Wynik powinien wyglądać, jak na Rysunku 6.



Rysunek 6: Wykresy stworzone przez nasz program. Źródło: własne

5 Opis fragmentów kodu

Teraz skupimy się na opisie fragmentów kodu związanych z wizualizacją danych o zdobytych medalach olimpijskich z lat 1896-2014. Przedstawimy, jakie konkretne aspekty kodu są istotne dla naszej analizy, jakie funkcje czy operacje są wykorzystywane, oraz jakie wyniki możemy uzyskać na podstawie tych fragmentów

5.1 Tworzenie map oraz legend

Ten fragment kodu w języku R służy do stworzenia mapy i jej legendy, która przedstawiaja liczbę medali zdobytych przez różne kraje.

```
krajeMedaleDM <- joinCountryData2Map(krajeMedale,
                                     joinCode = "ISO3",
                                     nameJoinColumn = "Country",
                                     mapResolution = "coarse")

#wybor palety
paletaMedalowa <- choose_palette()

#utworzenie mapy
mapa <- mapCountryData(krajeMedaleDM,
                       nameColumnToPlot="Medal",
                       mapTitle="Kraje swiata wzgledem liczby medali",
                       catMethod = c(range(1:5),range(6:10),
                                     range(11:50),
                                     range(51:100),range(101:500),
                                     range(501:1000),
                                     range(1001:2500),range(2501:5000),
                                     range(5001:5238)),
                       missingCountryCol = gray(.8),
                       oceanCol = "lightblue",
                       borderCol = "grey",
                       colourPalette = paletaMedalowa(17),
                       addLegend=F)

#utworzenie legendy
addMapLegendBoxes(
  cutVector=c("Brak danych/medali","1-5","6-10","11-50","51-100",
              "101-500","501-1000","1001-2500","2501-5000","5238")
  ,colourVector = c("lightgrey","#DAFF47" ,"#E9CF00", "#EDA200",
                    "#E4774D","#D24E71", "#B62485" ,"#91008D","#5F008C"
                    ,"#001889")
  ,x='bottomleft'
  ,title="Przedzia?y medali"
  ,pt.cex=2
  ,col="black"
  ,bg="white"
)
```

- krajeMedaleDM <- joinCountryData2Map(krajeMedale, joinCode = "ISO3", nameJoinColumn = "Country", mapResolution = "coarse") : Ta linia łączy dane o medalach z danymi geograficznymi dla każdego kraju. Używa kodów ISO3 do połączenia danych.
- paletaMedalowa <- choose_palette() : Wybiera paletę kolorów do użycia na mapie.
- mapa <- mapCountryData(...) : Tworzy mapę, na której kraje są kolorowane na podstawie liczby zdobytych medali. Używa różnych zakresów liczby medali do utworzenia różnych kategorii kolorów.
- addMapLegendBoxes(...) : Dodaje do mapy legendę pokazującą, jakie kolory odpowiadają jakim zakresom liczby medali.

5.2 Tworzenie wykresu słupkowego dotyczacego ilości medali zdobytych na poszczególnych igrzyskach olimpijskich

Ten fragment kodu w języku R służy do analizy liczby medali zdobytych przez Polskę na różnych igrzyskach olimpijskich i tworzenia wykresu słupkowego przedstawiającego te dane.

```
medale <- as.factor(igrzyska$Medal)
rok <- as.factor(igrzyska$Year)
(Polska <- filter(igrzyska,Country=="POL"))

(ile_na_kazdych_ig <- split(Polska, Polska$Year))

vm <- c()
for(i in 1:length(ile_na_kazdych_ig))
vm[i] <- dim(ile_na_kazdych_ig[[i]])[1]
vm

rok_Polska <- unique(Polska$Year)

MP <- as.data.frame(cbind(rok_Polska,vm))

lata <- MP$rok_Polska
liczba <- MP$vm

png("MedalePLnaDanychIO.png", width = 800, height = 800)

barplot(liczba, lata, main = "Liczba medali Polakow na danych IO",
        las = 2, col = terrain.colors(24), ylim = c(0,80),
        names.arg = MP$rok_Polska, ylab = "Liczba Medali",
        font.main = 3)
```

- medale <- as.factor(igrzyska$Medal) i rok <- as.factor(igrzyska$Year) : Konwertuje kolumny Medal i Year na czynniki.
- (Polska <- filter(igrzyska,Country=="POL")) : Filtruje dane tak, aby zawierały tylko rekordy dla Polski.
- (ile_na_kazdych_ig <- split(Polska, Polska$Year)) : Dzieli dane na podzbiory według roku igrzysk.
- Pętla for : Oblicza liczbę medali zdobytych przez Polskę na każdych igrzyskach.
- MP <- as.data.frame(cbind(rok_Polska,vm)) : Tworzy ramkę danych zawierającą rok igrzysk i liczbę zdobytych medali.
- png("MedalePLnaDanychIO.png", width = 800, height = 800) : Tworzy plik PNG, do którego zostanie zapisany wykres.
- barplot(...) : Tworzy wykres słupkowy pokazujący liczbę medali zdobytych przez Polskę na różnych igrzyskach olimpijskich.

5.3 Tworzenie wykresu słupkowego dotyczącego najlepszych żeńskich sportowców

Ten fragment kodu w języku R identyfikuje 10 najlepszych żeńskich sportowców w historii igrzysk olimpijskich na podstawie liczby zdobytych medali oraz tworzy wykres przedstawiający te dane.

```
zenskie <- data.frame(sportowcy=igrzyska$Athlete,
                      plec=igrzyska$Gender,
                      dyscyplina=igrzyska$Discipline,
                      medal=igrzyska$Medal)

zenskie <- zenskie %>% filter(plec=="Women")
zenskie <- subset(zenskie,select = -plec)
zenskie2 <- split(zenskie,zenskie$sportowcy)

(zlist <- length(zenskie2))

v <- numeric()
for(i in 1:zlist) v[i] <- dim(zenskie2[[i]])[1]
v

kobiety <- zenskie$sportowcy
length(kobiety <- unique(kobiety))
zenskie_medale <- data.frame(sportowcy=sort(kobiety),
                             medal=v)

#wykres
png("10 najlepszych kobiet.png",width=922,height=600)

zenskie_medale <- zenskie_medale[order(zenskie_medale$medal,decreasing=T),]
top_10k <- head(zenskie_medale,10)
top_10ks <- top_10k$sportowcy
top_10kml <- top_10k$medal
ggplot() + geom_col(aes(y = reorder(top_10ks,top_10kml), x = top_10kml)) +
  geom_label(aes(y = reorder(top_10ks,top_10kml), x = top_10kml, label = top_10kml
    )) +
  labs(title = '10 najlepszych zenskich medalistow',
       x="Sportowcy",y="Medale")
```

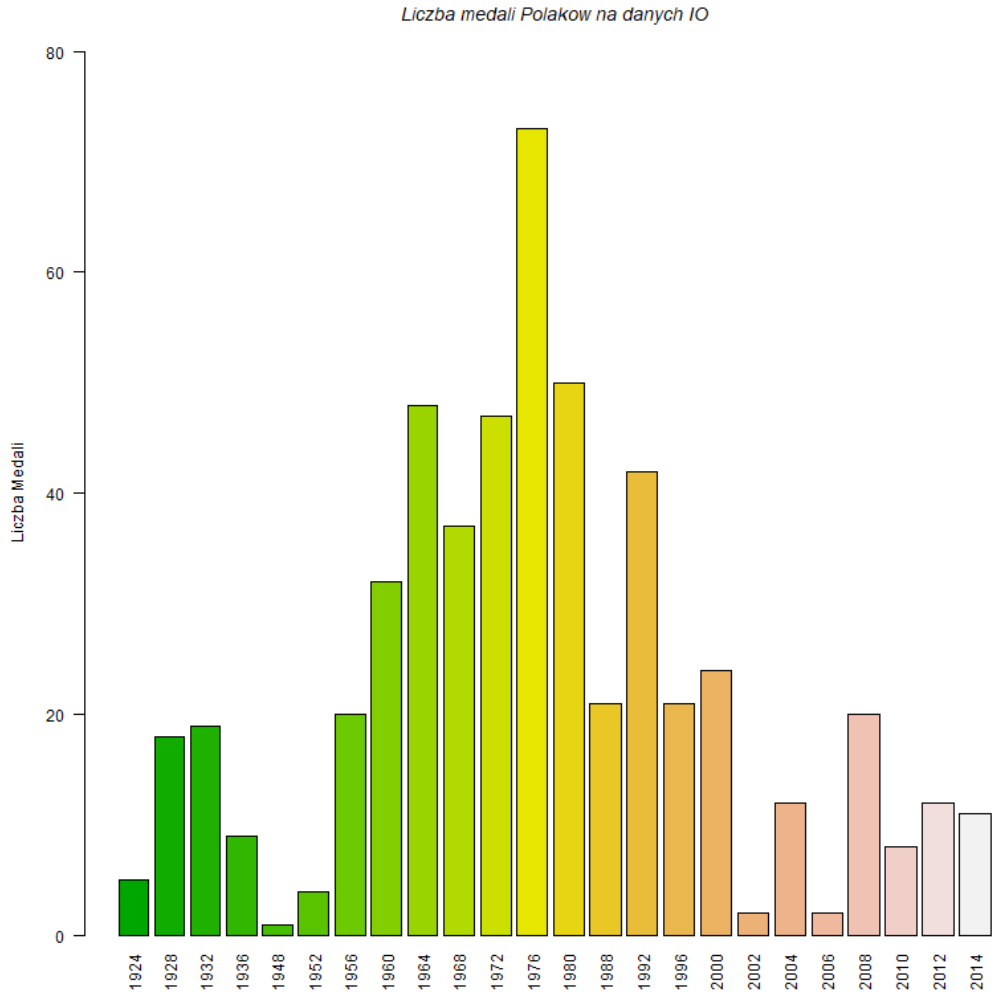
- zenskie <- data.frame(sportowcy=igrzyska$Athlete, plec=igrzyska$Gender, dyscyplina=igrzyska$Discipline, medal=igrzyska$Medal) : Tworzy ramkę danych zawierającą informacje o sportowcach, płci, dyscyplinie i zdobytych medalach.
- zenskie <- zenskie %>% filter(plec=="Women") : Filtruje dane tak, aby zawierały tylko rekordy dla sportowców płci żeńskiej.
- zenskie <- subset(zenskie,select = -plec) : Usuwa kolumnę plec z ramki danych, ponieważ wszystkie rekordy są teraz dla sportowców płci żeńskiej.
- zenskie2 <- split(zenskie,zenskie$sportowcy) : Dzieli dane na podzbiory według nazwisk sportowców.
- Pętla for : Oblicza liczbę medali zdobytych przez każdą sportowczynię.
- zenskie_medale <- data.frame(sportowcy=sort(kobiety), medal=v) : Tworzy nową ramkę danych zawierającą nazwiska sportowców i liczbę zdobytych medali.
- png("10 najlepszych kobiet.png",width=922,height=600) : Tworzy plik PNG, do którego zostanie zapisany wykres.
- ggplot() + geom_col(...) + geom_label(...) + labs(...) : Tworzy wykres słupkowy pokazujący 10 najlepszych żeńskich sportowców w historii igrzysk olimpijskich na podstawie liczby zdobytych medali.

6 Analiza uzyskanych danych

W tym kroku przeanalizujemy wybrane wykresy wygenerowane za pomocą naszego programu.

6.1 Polskie medale na igrzyskach olimpijskich

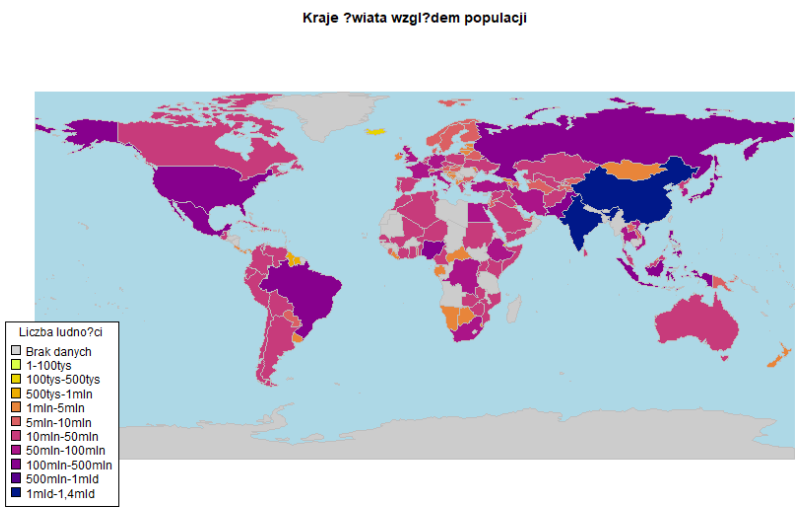
Wykres (Rysunek 7) przedstawia zestawienie ilości polskich medali olimpijskich zdobytych w danych latach na igrzyskach olimpijskich. W roku 1924 i 1948 Polska zdobyła znikomą ilość medalów na igrzyskach olimpijskich z powodu I oraz II wojny światowej, wieloletnie walki oraz dewastacja infrastuktury osłabiły nasz kraj. Jednak w czasach powojnennych, Polska przeżywała swoje lata świetności i osiągała wiele sukcesów. Największą ilość medali Polacy zdobyli w roku 1976 (65 medali w 26 kategoriach). Niestety, w następnych latach osiągnięcia naszych rodaków stopniowo malały.
Największe wartości na tym wykresie, są spowodowane medalami w dyscyplinach zespołowych, ponieważ ten wykres symbolizuję każdy medal jako osobnego sportowca, a nie jak znaczna większość tabel medalowych, cały zespół w danej dyscyplinie zespołowej np. siatkówce.



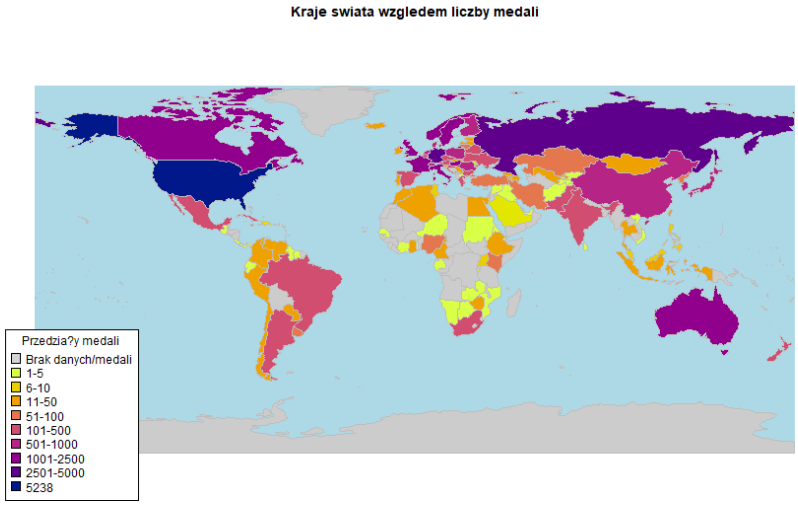
Rysunek 7: Wykres polskich medali zdobytych na igrzyskach olimpijskich. Źródło: własne

6.2 Populacja oraz medale na świecie

Analizujemy dwa wykresy: populacji na świecie (Rysunek 8) oraz medali zdobytych przez poszczególne państwa (Rysunek 9). Pomiędzy nimi jest relacja, ponieważ im większe jest zaludnienie kraju, tym większa jest szansa na zdobycie medalu dla tego państwa. Widać to na przykładzie USA oraz Rosji, gdzie populacja oraz ilość medali są dosyć wysokie. Natomiast więszkość krajów Afryki posiada małe zaludnienie i mało medali. Jednakże warto zaznaczyć, że sama wielkość populacji nie jest jedynym, ani nawet głównym czynnikiem wpływającym na sukces sportowy. Inne czynniki, takie jak infrastruktura sportowa, programy szkoleniowe, dostępność zasobów finansowych i wiele innych, również mają istotne znaczenie.



Rysunek 8: Wykres populacji na świecie. Źródło: własne



Rysunek 9: Wykres ilości medali przypadających na dany kraj. Źródło: własne

7 Podsumowanie

Projekt w języku R na temat “Analiza i wizualizacja danych o zdobytych medalach olimpijskich z lat 1896-2014” skupia się na analizie danych dotyczących zdobytych medali na igrzyskach olimpijskich i ich wizualizacji.

Główne elementy projektu:

- Przygotowanie danych: Dane o medalach olimpijskich są wczytywane i przetwarzane. Informacje takie jak nazwa sportowca, płeć, dyscyplina i zdobyte medale są ekstrahowane i przechowywane w ramkach danych.
- Analiza danych: Dane są analizowane, aby uzyskać informacje na temat liczby zdobytych medali przez różne kraje, sportowców i na różnych igrzyskach. Są one również filtrowane i grupowane według różnych kryteriów, takich jak płeć sportowca czy rok igrzysk.
- Wizualizacja danych: Na podstawie przeprowadzonej analizy tworzone są różne wykresy i mapy. Wykorzystuje się tutaj różne techniki wizualizacji, takie jak wykresy słupkowe czy mapy geograficzne, aby przedstawić wyniki analizy w sposób zrozumiały i atrakcyjny wizualnie.
- Podsumowanie wyników: Na koniec, wyniki analizy i wizualizacji są podsumowywane i interpretowane. Projekt ten pokazuje, jak można wykorzystać język R do przeprowadzenia kompleksowej analizy i wizualizacji danych. Dzięki temu można uzyskać cenne informacje na temat historii igrzysk olimpijskich i osiągnięć sportowców z różnych krajów.

8 Bibliografia

Literatura

[1] https://www.rdocumentation.org/packages/rvest/versions/1.0.4 z dnia 12 luty 2024

[2] https://www.rdocumentation.org/packages/tidyverse/versions/2.0.0

[3] https://rbasics.org/guides/how-to-use-the-sp-package-in-r/

[4] https://cran.r-project.org/web/packages/rworldmap/rworldmap.pdf

[5] https://rbasics.org/guides/how-to-use-the-rcolorbrewer-package-in-r/

[6] https://cran.r-project.org/web/packages/plotrix/index.html

[7] https://cran.rstudio.com/web/packages/readr/

[8] Marek Gągolewski, *Programowanie w Języku R analiza danych,obliczenia,symulacje*, Wydawnictwo PWN, Rok 2016.

9 Spis rysunków