Światowe dane dotyczące zdrowia publicznego i ogólnych wskaźników powiązanych w latach 2000-2022 – analiza wybranych wskaźników i zależności

Etap 1: Przygotowanie danych

Cel analizy:

Głównym celem etapu przygotowania danych było utworzenie bazy danych dotyczącej zdrowia publicznego na podstawie zmiennych zaimportowanych z bazy danych World Bank bezpośrednio do środowiska RStudio. Proces ten wykorzystuje bibliotekę WDI, która pozwala na pozyskiwanie danych z API World Bank. Biblioteka WDI w języku R umożliwia bezpośredni dostęp do danych World Bank, co eliminuje konieczność ręcznego pobierania plików. Dodatkowo, API World Bank udostępnia dane w sposób dynamiczny, co oznacza, że zawsze mamy dostęp do najświeższych informacji bez konieczności ręcznej aktualizacji zbiorów. Funkcja WDI() z pakietu WDI działa jako interfejs do API World Bank, umożliwiając wybór interesujących wskaźników, określenie zakresu lat i pobranie danych dla wybranych krajów lub wszystkich dostępnych w bazie. Wynikiem działania funkcji WDI() jest zbiór danych w formacie szerokim (wide format), gdzie każdy wiersz reprezentuje dane dla jednego kraju w określonym roku, a każda kolumna odpowiada wskaźnikowi lub dodatkowej zmiennej (np. nazwa kraju, rok, region geograficzny). Proces przygotowania danych do implementacji dashboardu w Rshiny obejmował:

- Dobór odpowiednich zmiennych na podstawie dostępnej bazy danych w World Bank
- Wyszukanie odpowiednich kodów dla każdej ze zmiennych, aby następnie użyć API World Bank do bezpośredniego pobrania danych do środowiska RStudio
- Pobranie danych dla wskaźników związanych ze zdrowiem publicznym i ogólną sytuacją ekonomiczną na poziomie kraju.
- Weryfikację kompletności danych (kilkuetapową analizę brakujących wartości w celu zachowania jak największych ilości informacji).
- Usunięcie wskaźników oraz mniejszych krajów, które miały wysoki odsetek braków danych, aby umożliwić bezproblemowe utworzenie wykresów analitycznych.
- Dostosowanie nazw i formatu zmiennych dla przyjaznego użytkownikowi końcowego wyglądu dashboardu w RShiny.

Wybrane wskaźniki zdrowia publicznego:

W ramach analizy wybrano **15 wskaźników dotyczących zdrowia publicznego oraz ogólnych warunków ekonomicznych, które są kluczowe dla oceny stanu zdrowia populacji, dostępności usług medycznych oraz jakości życia.** Poniżej znajduje się lista wskaźników wraz z ich kodami i opisami:

1. SP.DYN.LE00.IN - Oczekiwana długość życia (Life Expectancy).
2. SH.XPD.CHEX.GD.ZS - Wydatki na zdrowie (% PKB) (Health Expenditure as % of GDP).
3. SH.MED.BEDS.ZS - Liczba łóżek szpitalnych na 1000 osób (Hospital Beds per 1,000 people).
4. SH.STA.WASH.P5 - Dostęp do ulepszonych źródeł wody (% populacji) (Access to Improved Water Sources).
5. SH.IMM.MEAS - Szczepienia przeciw odrze (% dzieci w wieku 12–23 miesięcy) (Measles Immunization Coverage).
6. SP.POP.TOTL - Liczba ludności (Total Population).
7. SH.STA.BRTC.ZS - Porody z udziałem wykwalifikowanego personelu (%) (Births Attended by Skilled Health Staff).
8. SH.XPD.CHEX.PC.CD - Wydatki na zdrowie per capita (USD) (Health Expenditure per Capita in Current USD).
9. SH.MMR.RISK.ZS - Wskaźnik śmiertelności matek (Maternal Mortality Ratio per 100,000 Live Births).
10. SP.POP.GROW - Wzrost populacji (% rocznie) (Population Growth, Annual %).
11. SH.DYN.MORT - Wskaźnik śmiertelności dzieci do lat 5 (Mortality Rate, Under-5, per 1,000 Live Births).
12. SP.DYN.TFRT.IN - Wskaźnik dzietności (urodzenia na kobietę) (Fertility Rate, Births per Woman).
13. NY.GDP.PCAP.CD - PKB per capita (Gross Domestic Product per Capita in USD).
14. NY.GDP.MKTP.CD - PKB całkowity (Gross Domestic Product in Current USD).
15. SE.ADT.LITR.ZS - Wskaźnik alfabetyzacji dorosłych (% populacji dorosłych, którzy potrafią czytać i pisać) (Adult Literacy Rate, % of adults).

Proces przygotowania danych:

1. **Pobranie danych z bazy World Bank**

Na początku pobrano dane z bazy World Bank za pomocą pakietu WDI bezpośrednio do środowiska R przy użyciu API World Bank. Bazy danych Światowego Banku umożliwiają zescrapowanie danych przzu użyciu identyfikatorów poszczególnych zmiennych.

Zakres czasowy danych obejmował lata **2002–2022**, a dane pochodziły ze wszystkich dostępnych krajów w bazie danych, w tym były wartości zagregowane dla regionów (zsumowane wartości dla wszystkich krajów w danym regionie). W odstępnej bazie danych rgeion nie jest odpowiednikiem kontynentu, ale podział jest zbliżony. Regiony posłużą w późniejszej części do utworzenia dynamicznych dashboardów. Pobrane wskaźniki reprezentują różnorodne aspekty zdrowia publicznego, takie jak dostępność opieki zdrowotnej, wskaźniki demograficzne, i wskaźniki jakości zdrowia.

2. **Sprawdzenie kompletności danych**

Aby zapewnić wysoką jakość analizy, przeprowadzono wieloetapowy proces identyfikacji i eliminacji braków danych. Po wstępnym pobraniu wskaźników zdrowia publicznego oraz wskaźników gospodarczych z bazy danych World Bank za pomocą funkcji WDI(), przekształcono dane do formatu długiego (long format) dla lepszej analizy.

W tej fazie:

- Obliczono liczbę braków danych dla każdego wskaźnika i roku.
- Zidentyfikowano wskaźniki, które w jakimkolwiek roku miały 100% braków danych. Do tej grupy należały m.in.:

a. SH.STA.BRTC.ZS – Porody z udziałem wykwalifikowanego personelu
b. SH.XPD.CHEX.PC.CD – Wydatki na zdrowie per capita
c. SE.ADT.LITR.ZS – Wskaźnik alfabetyzacji dorosłych

Wskaźniki te usunięto z bazy, co pozwoliło na zwiększenie spójności i uniknięcie błędów podczas wizualizacji. Następnie, przeprowadzono ponowną analizę, która wykazała obecność wskaźników o wysokim poziomie braków (powyżej 90%), co również skutkowało ich usunięciem.

Dzięki tym działaniom finalna baza danych zawiera wskaźniki z pełnym lub częściowym pokryciem na poziomie co najmniej 90% dla większości krajów i lat, co zapewnia rzetelność wyników prezentowanych na dashboardzie.

3. Ręczne usunięcie dodatkowych wskaźników

Po ponownym sprawdzeniu danych po usunięciu zmiennych z pełnymi brakami wciąż pozostawały wskaźniki z wysokim odsetkiem braków (np. 90%). Aby zachować jakość danych, ręcznie usunięto kilka dodatkowych wskaźników, w tym:

a. SH.STA.BRTC.ZS - Porody z udziałem wykwalifikowanego personelu (%) (Births Attended by Skilled Health Staff).
b. SH.XPD.CHEX.PC.CD - Wydatki na zdrowie per capita (USD) (Health Expenditure per Capita in Current USD).
c. SH.XPD.CHEX.PC.CD - Wydatki na zdrowie per capita (USD) (Health Expenditure per Capita in Current USD).
d. SE.ADT.LITR.ZS - Wskaźnik alfabetyzacji dorosłych (% populacji dorosłych, którzy potrafią czytać i pisać) (Adult Literacy Rate, % of adults).

4. Dodatkowe sprawdzenie danych pod kątem brakujących wartości

Pomimo usunięcia kilku zmiennych zawierających liczne braki danych, wciąż napotkano obecność brakujących wartości. Zamiast jednak dalszego eliminowania zmiennych, zdecydowano się na dokładniejszą weryfikację, aby sprawdzić, czy ewentualnie braki danych dotyczą jedynie wybranych krajów. Obliczono więc braki danych w zmiennych dla unikatowych krajów. Okazało się, że pozostałe braki występują głównie w przypadku małych

państw oraz wysp, których jest łącznie 38. W związku z tym podjęto decyzję o usunięciu tych obserwacji z bazy danych, co pozwoliło na wyeliminowanie problemu brakujących danych oraz całkowite usunięcie potencjalnych niejasności i błędów przy korzystaniu z dynamicznego dashboardu.

5. Dostosowania bazy danych

W celu przygotowania danych do analizy oraz umożliwienia ich efektywnego wykorzystania w wizualizacjach i wykresach, przeprowadzono szereg kluczowych modyfikacji w strukturze bazy danych. Poniżej przedstawiono szczegóły wprowadzonych zmian.

a. **Usunięcie danych zagregowanych dla regionów:**
   Na potrzeby tej analizy nie interesują nas te zsumowane wartości dla regionów, jesteśmy w stanie obliczyć je wewnętrznie, natomiast uniemożliwiają one dodanie wartości długości i szerokości geograficznych.
b. **Konwersja współrzędnych geograficznych na wartości numeryczne:**
   Początkowo zmienne dotyczące współrzędnych geograficznych, czyli szerokości (latitude) i długości (longitude) geograficznej, były przechowywane jako tekst. W celu umożliwienia ich użycia w analizach przestrzennych i wizualizacjach (np. mapach), zmieniono ich typ na numeryczny. Dzięki temu, współrzędne stały się gotowe do dalszego przetwarzania.
c. **Zmiana typu zmiennej dochodów na zmienną kategoryczną:**
   Zmienna dotycząca poziomu dochodów (income) została przekształcona na zmienną kategoryczną (factor), z zachowaniem odpowiedniej kolejności (np. niskie, średnie, wysokie dochody). Taki typ danych ułatwia grupowanie państw na podstawie ich dochodów oraz poprawia interpretację wyników na wykresach panelowych i innych wizualizacjach.
d. **Ułatwienie interpretacji nazw zmiennych:**
   Zmieniono nazwy kolumn, aby były one bardziej zrozumiałe i intuicyjne w porównaniu do używanyc wcześniej nazw technicznych do przygotowania danych. Na przykład, zamiast angielskich nazw zmiennych, takich jak NY.GDP.PCAP.CD, zastosowano bardziej przystępne polskie odpowiedniki, takie jak "PKB per capita [$]". Tego typu zmiany miały na celu poprawienie czytelności bazy danych, co ułatwia jej późniejsze wykorzystanie w analizach.
e. **Przekształcenie kategorii dochodów na polskie odpowiedniki:**
   W celu uproszczenia interpretacji danych w polskim kontekście, angielskie kategorie dochodów, takie jak „Low income” czy „Upper middle income”, zostały zamienione na ich polskie odpowiedniki, tj. „Niskie dochody”, „Wyższe średnie dochody” itp. Takie podejście pozwala na lepszą adaptację wyników do lokalnych warunków i ułatwia ich zrozumienie przez polskojęzycznych użytkowników.
f. **Selekcja istotnych zmiennych:**
   Aby uprościć analizę, wybrano tylko te zmienne, które były istotne dla celu badania. W wyniku tego, z bazy danych usunięto niepotrzebne kolumny, co pozwoliło na

zachowanie przejrzystości zbioru danych i skoncentrowanie się na najważniejszych wskaźnikach.

g. **Dodatkowe sprawdzanie braków danych:**
Kolejnym krokiem było przeprowadzenie dokładnej analizy brakujących danych. Okazało się, że brakujące wartości występują głównie w przypadku zmiennych związanych z geolokalizacją, tj. szerokości i długości geograficznej dla wybranych regionów. Dodatkowo, zidentyfikowano, że brakujące dane dotyczą w dużej mierze niewielkich państw oraz terytoriów wyspiarskich, takich jak West Bank and Gaza. Łącznie zidentyfikowano 38 krajów i terytoriów z licznymi brakami danych, co utrudniało ich wizualizację i analizę w ramach dashboardu. Podjęto decyzję o usunięciu tych obserwacji, aby zapewnić pełną spójność danych dla pozostałych krajów i uniknąć sytuacji, w której niekompletne informacje mogłyby prowadzić do błędnych interpretacji. Eliminacja małych krajów i regionów nie miała istotnego wpływu na wyniki analizy, a proces ten pozwolił na płynniejsze działanie aplikacji RShiny i wyeliminowanie potencjalnych niejasności podczas interaktywnych wizualizacji mapowych.

h. **Zmiana nazw regionów na polskie:**
W celu poprawy czytelności i zrozumienia danych, zmieniono angielskie nazwy regionów geograficznych na polskie odpowiedniki. Na przykład, region „East Asia & Pacific" został zamieniony na „Azja Wschodnia i Pacyfik", a „North America" na „Ameryka Północna".

6. Wynik końcowy

Ostateczna baza danych została przefiltrowana tak, aby zawierała tylko wskaźniki i obserwacje bez braków danych, aby uniknąć błędów przy koszystaniu z dynamicznych wykresów w ramach dashboardu.

Podsumowanie etapu przygotowania danych:

Proces przygotownia danych umożliwił:

- Usunięcie wskaźników z pełnymi brakami danych.
- Sprawdzenie braków danych i usunięcie krajów z niekompletnymi danymi geograficznymi.
- Konwersję zmiennych na odpowiednie typy danych.
- Zmianę nazw zmiennych na bardziej czytelne w ramach dashboardu.
- Przekształcenie niektórych nazw z oryginalnej bazy danych po angielsku na polskie (oprócz nazw krajów).
- Selekcję istotnych zmiennych.
- Ostateczne sprawdzenie braków danych i przygotowanie do wizualizacji.
- Przygotowanie bazy do wizualizacji w aplikacjach interaktywnych.

Etap 2: Tworzenie dashboardu Rshiny

Ostateczna aplikacja RShiny ma na celu zaprezentowanie danych dotyczących zdrowia publicznego i ogólnych wskaźników gospodarczych na całym świecie w latach 2000-2022 przy użyciu różnych typów wizualizacji w podziale na kraje, regiony, lata w zależności od zakładki. Końcowy dashboard składa się z sześciu zakładek, które pozwalają na interaktywną analizę, wizualizację danych oraz generowanie wykresów i map. Każda z nich zawiera instrukcję dla docelowego użytkownika. Poniżej przedstawione zostaną zakładki oraz ich szczegółowy opis wraz z funkcjonalnościami i instrukcją, aby zobrazować ich zastosowanie.

Sekcje utworzone w ramach aplikacji RShiny:

1. **Trendy wskaźników**

Zakładka ta pozwala na analizowanie zmian wartości wybranego wskaźnika w czasie (od 2000 do 2022 roku) dla wybranych regionów i krajów. Użytkownik może zobaczyć jak różne wskaźniki, takie jak oczekiwana długość życia, śmiertelność dzieci, wskaźnik dzietności, PKB i inne, zmieniały się w analizowanych krajach i regionach.

Instrukcja:

a. Wybór regionów i krajów: Użytkownik może wybrać regiony i kraje, które chce analizować. Wybór regionu automatycznie zawęża wybór krajów do tych, które należą do wybranego regionu, ale można także analizować tylko konkretne kraje niezależnie od regionu.
b. Wybór wskaźnika: Użytkownik wybiera jeden z dostępnych wskaźników, który chce analizować, np. "Oczekiwana długość życia", "PKB per capita", "Wzrost populacji".

Wykresy:

a. Wykres liniowy: Pokazuje zmiany wartości wybranego wskaźnika w czasie dla wybranych krajów i regionów.
b. Wykres pudełkowy: Przedstawia rozkład wartości wskaźnika dla różnych regionów, umożliwiając porównanie zmienności między nimi.

2. **Zależności między wskaźnikami**

Zakładka ta pozwala na badanie korelacji między dwoma wybranymi wskaźnikami (np. "PKB per capita" i "Oczekiwana długość życia") w wybranych krajach lub regionach. Dzięki tej funkcjonalności użytkownik może lepiej zrozumieć zależności pomiędzy różnymi aspektami zdrowia publicznego i gospodarki.

Instrukcja:

a. Wybór zmiennych: Użytkownik wybiera dwa wskaźniki (X i Y), dla których chce zbadać korelację.
b. Filtrowanie danych: Użytkownik może zastosować filtry regionów i krajów, aby zawęzić dane do wybranego obszaru geograficznego. Wybór regionu ogranicza analizę do krajów w tym regionie.
c. Wykres: Wykres rozrzutu (scatter plot) pokazuje zależność między wybranymi wskaźnikami, a kolory punktów reprezentują różne regiony. Dodatkowo, linia regresji (krzywa trendu) może pomóc w wizualizacji zależności między zmiennymi.

3. Mapa

Zakładka ta umożliwia analizę danych na interaktywnej mapie, na której prezentowane są kraje z uwzględnieniem wartości wybranego wskaźnika. Dzięki tej funkcji użytkownik może szybko zobaczyć, jak różne kraje prezentują się w odniesieniu do wybranego wskaźnika i regionu.

Instrukcja:

a. Wybór regionów: Użytkownik wybiera regiony, które mają być wyświetlone na mapie. Po dokonaniu wyboru mapa automatycznie zaktualizuje widok, pokazując kraje w wybranych regionach.
b. Wielkość i kolor kropek: Kropki na mapie reprezentują kraje, a ich wielkość odzwierciedla wartość wskaźnika "Oczekiwana długość życia". Kolor kropek zależy od poziomu dochodów danego kraju, co pozwala na szybkie zauważenie różnic w dochodach.
c. Interaktywność: Mapa jest interaktywna, co pozwala użytkownikowi na przybliżanie, oddalanie i przesuwanie mapy, a po najechaniu na kropkę wyświetlana jest szczegółowa informacja o danym kraju (np. nazwa kraju, oczekiwana długość życia, PKB per capita).

4. Udziały kategorii dochodu i regionów

Zakładka ta pozwala użytkownikowi zobaczyć rozkład krajów według kategorii dochodów (wysokie, średnie i niskie dochody) lub regionów, w formie wykresu kołowego. Użytkownik może porównać, jak wiele krajów należy do poszczególnych kategorii dochodów lub jak rozkładają się kraje w różnych regionach.

Instrukcja:

a. Wybór zmiennej: Użytkownik wybiera, czy chce zobaczyć rozkład krajów według kategorii dochodów, czy według regionów.
b. Wykres kołowy: Wykres przedstawia liczbę krajów w danej kategorii dochodów lub regionie. Każdy sektor wykresu kołowego reprezentuje jedną kategorię lub region.

5. Histogram

Zakładka ta umożliwia analizę rozkładu wybranej zmiennej (np. oczekiwana długość życia, śmiertelność dzieci, PKB per capita) dla wybranych regionów w określonym roku, w formie histogramu. Dzięki temu użytkownik może zobaczyć, jak dane są rozproszone w obrębie wybranych regionów.

Instrukcja:

a. Wybór zmiennej: Użytkownik wybiera wskaźnik, którego rozkład chce analizować.
b. Wybór roku: Użytkownik może wybrać konkretny rok, dla którego chce zobaczyć rozkład zmiennej.
c. Wybór regionów: Użytkownik może wybrać jeden lub więcej regionów do analizy. Jeśli nie wybierze żadnego regionu, dane będą wyświetlane dla wszystkich regionów.
d. Wykres: Histogram przedstawia liczbę krajów w wybranych regionach, których wartość wskaźnika mieści się w określonych przedziałach.

6. Wykres słupkowy

Zakładka ta umożliwia analizowanie średniej wartości wybranej zmiennej (np. oczekiwana długość życia, PKB) dla każdego regionu w określonym roku w formie wykresu słupkowego. Jest to pomocne w porównaniu różnych regionów pod kątem wybranego wskaźnika.

Instrukcja:

a. Wybór zmiennej: Użytkownik wybiera zmienną (wskaźnik), którą chce zobaczyć na wykresie.
b. Wybór roku: Użytkownik wybiera rok, dla którego chce zobaczyć średnią wartość zmiennej w regionach.
c. Wykres: Wykres słupkowy przedstawia średnią wartość wskaźnika w różnych regionach w wybranym roku.

Podsumowanie:

Aplikacja RShiny oferuje użytkownikowi szeroki zakres narzędzi analitycznych, w tym wykresy liniowe, pudełkowe, histogramy, wykresy kołowe, słupkowe oraz interaktywne mapy. Dzięki tym funkcjom, użytkownicy mogą analizować zmiany wskaźników w czasie, zależności między różnymi wskaźnikami, rozkłady danych w regionach i krajach, a także odkrywać zróżnicowanie wyników w zależności od poziomu dochodów. Aplikacja jest interaktywna, co umożliwia dokładną i elastyczną analizę globalnych danych zdrowotnych i ekonomicznych.

Źródła:

https://cran.r-project.org/web/packages/WDI/WDI.pdf

https://data.worldbank.org/indicator/