



AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

WYDZIAŁ INFORMATYKI, ELEKTRONIKI
I TELEKOMUNIKACJI
KIERUNEK INFORMATYKA

MICHał PATYK

**Analiza danych i wzorców dotyczących
wydarzeń politycznych na podstawie
informacji zgromadzonych w projekcie
GDELT**

(pracownia problemowa)

Opiekun: dr hab. inż. Koźlak Jarosław

Kraków 2020

Spis treści

1 Wstęp	3
1.1 Motywacja	3
1.2 Cele pracy	3
2 Przegląd dziedziny	5
2.1 GDELT	5
2.2 CAMEO	5
2.2.1 Zdarzenia	6
2.2.2 Aktorzy	6
2.3 Przegląd istniejących analiz	6
3 Koncepcja	8
3.1 Założenia i wymagania	8
3.2 Efekt końcowy	10
4 Realizacja	11
4.1 Wykorzystane narzędzia	11
4.2 Plan pracy	12
4.3 Organizacja pracy	12
5 Wstępna analiza danych	14
5.1 Popularność Polski w zbiorze danych GDELT	14
5.1.1 Analiza zbiorcza od 2015 roku	16
5.2 Analiza danych dla wybranych krajów	21
5.2.1 Polska	21
5.2.2 Niemcy	27
5.2.3 Rosja	31
5.2.4 Stan Zjednoczone	34
5.3 Analiza siły powiązania	40

<i>SPIS TREŚCI</i>	2
5.3.1 Analiza siły powiązania między wybranymi krajami	40
5.3.2 Analiza symetryczności siły powiązania	42
5.4 Analiza kodu podstawowego Fight	44
6 Grupowanie państw	46
6.1 Dane niestandardyzowane	46
6.2 Dane ustandardyzowane	47
6.2.1 Wyniki klasteryzacji - wykresy gęstości miar użytych do klastrowania - dane ustandardyzowane	53
6.2.2 Wyniki klasteryzacji w postaci tabelarycznej - dane ustandardyzowane	57
7 Podsumowanie	67
7.1 Wnioski	67
7.2 Dalsze kierunki rozwoju	67
A Wyniki grupowania danych niestandardyzowanych	69
B Parametry klastrów - dane standaryzowane	76
C Korzystanie z programu	81
Bibliografia	82

Rozdział 1

Wstęp

W tym rozdziale została opisana motywacja oraz cele pracy. W części 1.1 została przedstawiona motywacja do stworzenia pracy. Część 1.2 opisuje cele pracy. Rozdział 2 zawiera przegląd dziedziny. W rozdziale 3 przedstawiona została koncepcja pracy. Rozdział 4 dotyczy realizacji zadania. W rozdziale 5 wykonana została wstępna analiza danych. Rozdział 6 opisuje grupowanie państw o podobnych cechach. W ostatnim rozdziale 7 dokonano podsumowania przeprowadzonych analiz. Dodatek A zawiera szczegółowe wyniki grupowania dla danych niestandardowanych. Dodatek B zawiera parametry klastrów dla danych standardowanych. Dodatek C zawiera instrukcję uruchomienia kodu projektu. Na końcu dokumentu dołączony został spis bibliograficzny.

1.1 Motywacja

Projekt GDELT monitoruje doniesienia medialne ze wszystkich krajów, w ponad 100 językach, każdego dnia. Wszystkie wydarzenia opisane są przez aktorów, lokalizacje i charakter aktywności, co jest specyfikowane poprzez zestaw atrybutów. Analiza atrybutów może umożliwić lepsze zrozumienie specyfiki krajów. Wykrycie wzorców może pozwolić na automatyczną reakcję na zdarzenia, prognozowanie oraz wykrywanie zdarzeń nietypowych.

1.2 Cele pracy

Celem niniejszej pracy jest analiza wydarzeń politycznych oraz poszukiwanie występujących w nich wzorców w oparciu o dane projektu GDELT.

Zdarzenia są opisane przez aktorów. Wybrani zostaną aktorzy - kraje. Następnie kraje zostaną pogrupowane z wykorzystaniem metody k-średnich w oparciu o wektor cech wybranych z opisu zdarzeń. Klastrowanie zostanie przeprowadzone przed i po standaryzacji. Klastry zostaną porównane z danymi zewnętrznymi. Grupowanie krajów na podstawie atrybutów z bazy danych GDELT pozwoli na wykazanie związku z rzeczywistymi cechami państw.

Rozdział 2

Przegląd dziedziny

W tym rozdziale w części 2.1 opisany został zbiór danych GDELT. W części 2.2 przedstawiony został schemat kodowania CAMEO. W części 2.3 przeanalizowane zostały prace dotyczące zbioru danych GDELT.

2.1 GDELT

GDELT - Global Database of Events, Language, and Tone - to największa, najbardziej wszechstronna i otwarta baza danych jaka powstała. Wczesne poszukiwania prowadzące do stworzenia GDELT zostały opisane przez Philipa Schrodta w dokumencie [14] w styczniu 2010 r. Zbiór danych jest dostępny na stronie Projektu [2] oraz na platformie Google Cloud gdzie można z niego korzystać przez Google BigQuery [20]. GDELT używa kodowania obserwacji konfliktów i mediacji (CAMEO) [3] do rejestrowania zdarzeń. W zbiorze znajdują się dane od 1979 roku. Kolejne porcje zdarzeń i ich klasyfikacja generowane są na bieżąco każdego dnia.

2.2 CAMEO

CAMEO - Conflict and Mediation Event Observations - jest schematem kodowania zdarzeń. Został stworzony w Katedrze Nauk Politycznych Pennsylvania State University. Jego początki sięgają roku 2000. Został zaprojektowany z myślą o automatycznym kodowaniu i szczegółowym kodowaniu aktorów.

2.2.1 Zdarzenia

Kody zdarzeń są ujednolicone pod względem kolejności numerycznej głównych kategorii. Kategorie są uszeregowane rosnąco względem kooperacji od 01 do 09 oraz względem konfliktu od 10 do 20.

2.2.2 Aktorzy

Kody aktorów składają się z trzech znaków. Elementy kodu są podzielone na szerokie kategorie, takie jak podmioty państwowe, role, regiony i grupy etniczne aktorów.

2.3 Przegląd istniejących analiz

Zbiór danych GDELT był wielokrotnie analizowany i wykorzystywany w pracach naukowych. Dla słowa kluczowego GDELT multiwyszukiwarka EDS zwraca ponad 700 wyników. Poniżej opisanych zostanie kilka wybranych prac.

W artykule [22] autorzy proponują miarę siły połączenia między krajami. Identyfikowane są różne wzorce połączeń pomiędzy Chinami i 15 istotnymi krajami.

W badaniu [9] autorzy porównują zbiór danych GDELT oraz EventRegisters pod względem rozmiaru, źródeł oraz geografii wiadomości. GDELT okazuje się być większy zarówno pod względem ilości dokumentów jak i ilości źródeł. Każdy ze zbiorów posiada unikalne źródła publikujące wiele dokumentów.

W artykule [8] autorzy analizują zbiór GDELT pod kątem informacji o konfliktach i pokoju. Badanie udowadnia, że GDELT wychwytuje globalne trendy i pozwala na zidentyfikowanie znaczących punktów w wydarzeniach konfliktowych. Zaproponowany zostaje czteropoziomowy model do zaprojektowania panelu analizy zdarzeń.

W badaniu [10] autorzy dokonują porównania tonu wypowiedzi na temat wiadomości międzynarodowych zgromadzonych w GDELT. Porównanie wypowiedzi z dziewięciu krajów pokazało, że mimo otrzymania różnej klasyfikacji dla wiadomości międzynarodowych w tym samym temacie, średni ton pozostaje zbliżony.

W pracy [17] autorzy przedstawiają model metadanych dla zdarzeń politycznych. Na bazie tego modelu stworzono zbiór danych który został przetestowany poprzez wygenerowanie raportu o wizycie prezydenta Trump'a w chinach.

W artykule [1] autorzy analizują opinie na temat polityki energetycznej Hiszpanii przy użyciu GDELT. Stworzone wskaźniki sentymentu zostały ocenione poprzez porównanie ze zmiennymi rynku energii takimi jak cen i popyt. Korelacja do dziennych cen energii nie została odnaleziona.

W badaniu [21] autor bada efektywność tradycyjnych modeli szeregow czasowych w przewidywaniu trendów w doniesieniach medialnych. Wnioskiem autora, model autoregresyjny ze średnią ruchomą (ARIMA) ma ograniczone zastosowanie w określaniu międzynarodowych relacji, w odniesieniu do wielkich zbiorów danych.

Rozdział 3

Koncepcja

W tym rozdziale w części 3.1 zostały przedstawione założenia i wymagania. W części 3.2 został opisany efekt końcowy projektu.

3.1 Założenia i wymagania

W analizie wykorzystany zostanie głównie zbiór danych GDELT 2.0 od początku 2015 roku do kwietnia 2020. W pierwszej kolejności zostaną przeprowadzone wstępne analizy które pozwolą lepiej zrozumieć specyfikę danych. Następnie zostanie przeprowadzona klasteryzacja. Do grupowania państw w kraje zostanie wykorzystany wektor cech z bazy GDELT oraz algorytm klastrowania k-srednich. Wektor dla każdego kraju będzie składał się z:

- liczby zdarzeń w analizowanym okresie, w których dany kraj jest aktorem 1 (dalej oznaczone jako *Events*),
- sumy liczby wzmianek (całkowitej liczby wzmianek o tym wydarzeniu, we wszystkich dokumentach źródłowych podczas 15-minutowej aktualizacji, w której zostało po raz pierwszy zauważone) w analizowanym okresie, dla wydarzeń w których dany kraj jest aktorem 1 (dalej oznaczone jako *numMentions*),
- stosunku liczby zdarzeń w analizowanym okresie, w których dany kraj jest aktorem 1, oznaczonych *quad class* jako *material conflict* do liczby zdarzeń w analizowanym okresie, w których dany kraj jest aktorem 1, oznaczonych *quad class* jako *material cooperation* (dalej oznaczone jako *materialConfCoop*),

- stosunku liczby zdarzeń w analizowanym okresie, w których dany kraj jest aktorem 1, oznaczonych *quad class* jako *verbal conflict* do liczby zdarzeń w analizowanym okresie, w których dany kraj jest aktorem 1, oznaczonych *quad class* jako *verbal cooperation* (dalej oznaczone jako *verbalConfCoop*),
- średniego, dla wszystkich zdarzeń w analizowanym okresie, w których dany kraj jest aktorem 1, średniego tonu (średni „ton” wszystkich dokumentów zawierających jedną lub więcej wzmianek o tym wydarzeniu, podczas 15-minutowej aktualizacji, w której zostało ono po raz pierwszy zauważone, waha się od -100 (skrajnie ujemny) do +100 (skrajnie dodatni)) (dalej oznaczone jako *avgAvgTone*),
- średniej, dla wszystkich zdarzeń w analizowanym okresie, w których dany kraj jest aktorem 1, miary Goldsteina (skala Goldsteina przypisuje wynik liczbowy od -10 do +10, wychwytując teoretyczny potencjalny wpływ, jaki rodzaj zdarzenia będzie miał na stabilność kraju) (dalej oznaczone jako *avgGoldstein*),
- liczby zdarzeń w analizowanym okresie, w których dany kraj jest aktorem 1, oznaczonych *event root code* jako *fight Fight* (dalej oznaczone jako *fightCount*),
- liczby zdarzeń w analizowanym okresie, w których dany kraj jest aktorem 1, oznaczonych *event root code* jako *Express intent to cooperate* (dalej oznaczone jako *expressCount*).

Dane wykorzystane w tym doświadczeniu pochodzą ze stycznia 2020 roku. Przed dokonaniem klasteryzacji odrzucone zostaną wydarzenia z kodami krajów cameo niezgodnymi z kodami ISO 3166-1 alfa-3 [6]. Klasteryzacja zostanie przeprowadzona dwukrotnie, za drugim razem na danych ustANDARDYZOWANYCH przy pomocy modułu StandardScaler [16].

Dla ułatwienia interpretacji wyników grupowania zostaną dodane informacje o PKB, wydatkach na opiekę zdrowotną, zbrojenia, edukację, import oraz eksport (jako procent PKB). Dodatkowe dane pochodzą z bazy Banku Światowego [19], są na licencji CC BY-4.0 [18], przedstawiają sytuację w 2015 roku.

GDP PKB to suma wartości dodanej brutto, w dolarach, wszystkich producentów będących rezydentami w gospodarce, powiększona o wszelkie podatki od produktów i pomniejszona o wszelkie dotacje nieuwzględnione w wartości produktów.

Education Wydatki sektora instytucji rządowych i samorządowych na edukację są wyrażone jako procent PKB.

Military Dane dotyczące wydatków wojskowych z Międzynarodowego Instytutu Badań nad Pokojem w Sztokholmie pochodzą z definicji NATO, która obejmuje wszystkie bieżące i kapitałowe wydatki na siły zbrojne, w tym siły pokojowe. Wartości mniejsze niż 0.01% zostały odrzucone w celu dostosowania skali logarytmicznej.

Health Poziom bieżących wydatków na zdrowie wyrażony jako procent PKB. Szacunki bieżących wydatków na zdrowie obejmują towary i usługi zdrowotne konsumowane w każdym roku.

Import Import towarów i usług reprezentuje wartość wszystkich towarów i innych usług rynkowych otrzymanych z reszty świata. Obejmuje wartość towarów, frachtu, ubezpieczenia, transportu, podróży, tantiem, opłat licencyjnych i innych usług, takich jak usługi komunikacyjne, budowlane, finansowe, informacyjne, biznesowe, osobiste i rządowe. Nie obejmuje kosztów związanych z zatrudnieniem i dochodów z inwestycji oraz płatności transferowych.

Export Eksport towarów i usług reprezentuje wartość wszystkich towarów i innych usług rynkowych dostarczanych do reszty świata.

3.2 Efekt końcowy

Stworzony w projekcie system pozwoli na grupowanie państw na podstawie medialnych informacji zebranych w bazie GDELT. Przeprowadzone grupowanie państw w klastry pozwoli na uzyskanie uproszczonego, zagęgowanego obrazu sytuacji geopolitycznej. Planowanym efektem końcowym pracy będzie wykazanie, że na bazie cech z bazy danych GDELT możemy skategoryzować państwa o rzeczywistych cechach takich jak PKB, wydatki na zbrojenia, edukację i ochronę zdrowia.

Rozdział 4

Realizacja

W tym rozdziale w części 4.1 opisane zostały metody i narzędzia wykorzystane podczas realizacji projektu. W części 4.2 przedstawiony został plan pracy. Część 4.3 prezentuje organizację pracy nad zadaniem.

4.1 Wykorzystane narzędzia

- język programowania Python [13]
- biblioteka Pandas [11]
- biblioteka GeoPandas [4]
- biblioteka Scikit-Learn [15]
- środowisko programistyczne PyCharm [12]
- internetowe interaktywne środowisko obliczeniowe Notebook Jupyter [7]
- hurtownia danych Google BigQuery [5]

Przy tworzeniu projektu wykorzystany został język programowania Python. Jest to język wysokiego poziomu ogólnego przeznaczenia. Wraz z biblioteką Pandas jest często stosowany w zagadnieniach analizy danych oraz data miningu. Pandas jest łatwym w użyciu narzędziem open source które oferuje w szczególności struktury danych i operacje do manipulacji tabelami numerycznymi. Głównym narzędziem używanym do programowania było zintegrowane środowisko programistyczne PyCharm firmy JetBrains. Pozwala ono na

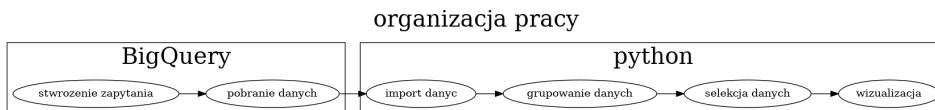
wygodną edycję i analizę kodu źródłowego. Internetowe interaktywne środowisko obliczeniowe Notebook Jupyter pozwoliło na tworzenie dokumentów zawierających kod wraz z wizualizacjami. Wykorzystanie hurtowni danych Google BigQuery pozwoliło na szybką, skalowalną analizę dużego zbioru danych jakim jest GDELT. BigQuery jest oprogramowaniem bezserwerowym, które obsługuje zapytania w języku ANSI SQL. Podział krajów na krastry został wykonany przy użyciu Scikit-Learn. Do wizualizacji klastrów na mapie wykorzystano bibliotekę GeoPandas, która jest projektem open source mającym na celu ułatwienie pracy z danymi geograficznymi.

4.2 Plan pracy

W pierwszej kolejności przeprowadzona zostanie wstępna analiza danych, która pozwoli na zorientowanie w jakim stopniu poszczególne cechy zdarzeń z bazy GDELT odzwierciedlają rzeczywiste zdarzenia i relacje między krajami. Następnie przeprowadzone zostanie grupowanie państw metodą k-średnich w oparciu o wybrane cechy zdarzeń.

4.3 Organizacja pracy

Praca nad zadaniem przebiegała według schematu przedstawionego na rysunku 4.1.



Rys. 4.1: Organizacja pracy. (źródło: opracowanie własne)

Najpierw w hurtowni danych BigQuery tworzone będą zapytania do bazy danych GDELT pozwalające na agregację, filtrowanie i grupowanie danych. Poniżej przykład zapytania.

```

SELECT
    MonthYear,
    Actor1CountryCode,
    COUNT(*) AS Events,
FROM
    `gdelt-bq.gdeltv2.events`

```

```
WHERE
  Year >= 2015
GROUP BY
  MonthYear,
  Actor1CountryCode
```

Wyniki zapytań będą zapisywane lokalnie, na dysk, do dalszej obróbki. Następnie w środowisku programistycznym PyCharm dla poszczególnych zagadnień tworzone będą notatniki jupyter łączące kod języka python z wizualizacjami. W notatnikach, zapisane wcześniej na dysku dane, będą importowane, a następnie grupowane tak aby na pojedynczym wykresie zamieścić możliwie dużo informacji. Przedostatnim etapem będzie selekcja najistotniejszych danych w celu poprawy czytelności wykresów. Na końcu tworzone będą i zapisywane do plików wizualizacje danych.

Dane, programy oraz efekty przeprowadzonych prac można znaleźć w repozytorium GitHub.

Rozdział 5

Wstępna analiza danych

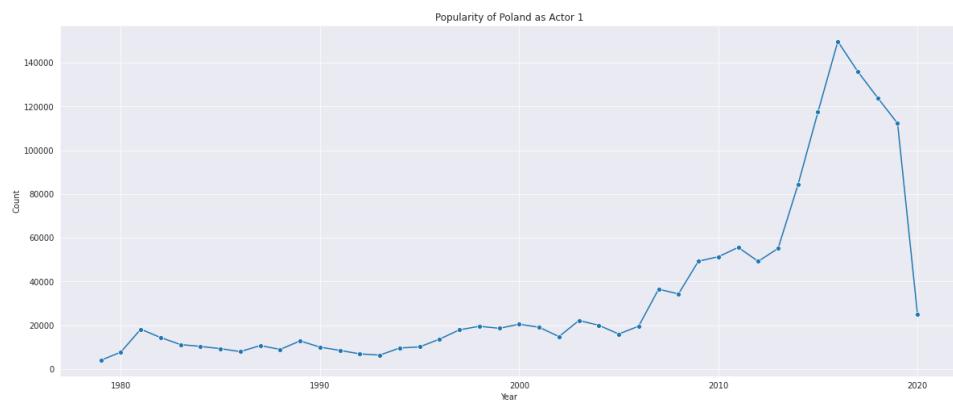
W tej części pracy przeprowadzona zostanie wstępna analiza danych. W pierwszej kolejności w części 5.1 przeanalizowane zostaną dane dotyczące Polski, co pozwoli na łatwiejsze wychwycenie związków między zarejestrowanymi wydarzeniami, a sytuacją w kraju. W dalszej kolejności w części 5.2 przeprowadzona zostanie analiza wybranych krajów. W części 5.3 przedstawiono analizę siły powiązania między wybranymi krajami. W części 5.4 przedstawiono analizę wybranych krajów pod kątem zdarzeń *fight*.

5.1 Popularność Polski w zbiorze danych GDELT

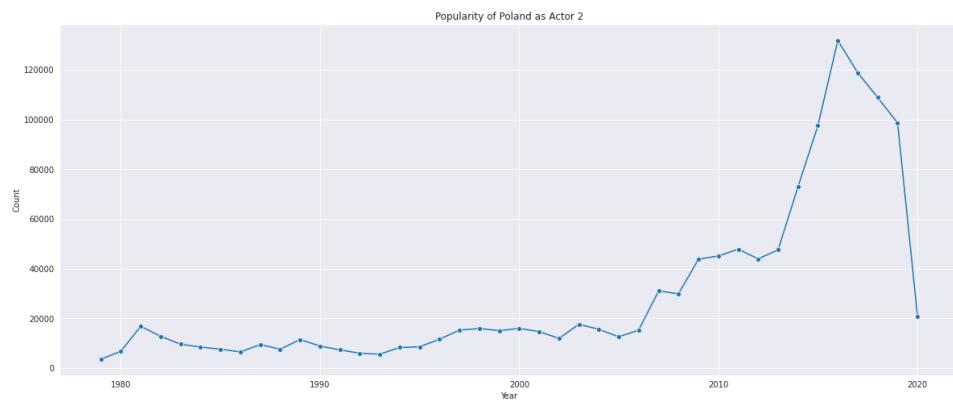
Jako pierwszą analizę wykonano badanie popularności Polski w zbiorze danych GDELT. Na wszystkich trzech wykresach obserwujemy znaczny wzrost liczby zdarzeń w 2015 roku. Może być to związane z uruchomieniem w GDELT automatycznego tłumaczenia artykułów i co za tym idzie zwiększeniem liczby źródeł danych.

Polska jako Aktor 1 Wykres 5.1 przedstawia popularność Polski, jako aktora 1, jako ilość zdarzeń w poszczególnych latach. W roku 2016 obserwujemy szczyt popularności na poziomie około 150 tysięcy zdarzeń.

Polska jako Aktor 2 Wykres 5.2 przedstawia popularność Polski, jako aktora 2, jako ilość zdarzeń w poszczególnych latach. Kształt wykresu jest bardzo zbliżony do 5.1 jednak szczyt popularności jest niższy - na poziomie około 130 tysięcy zdarzeń.

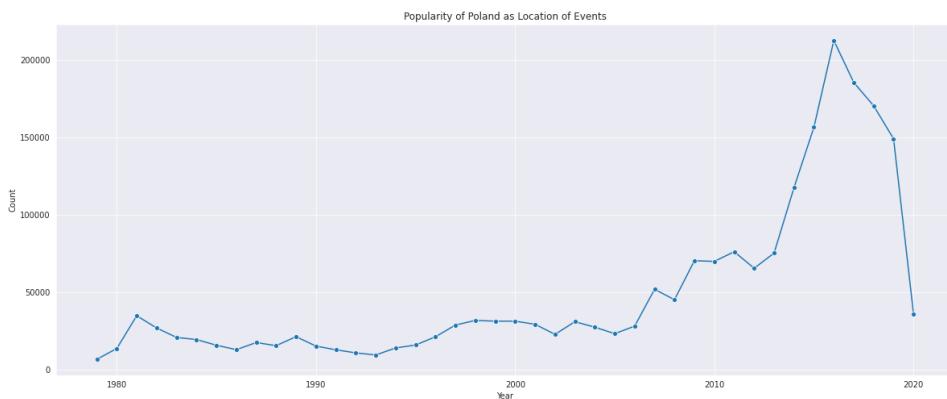


Rys. 5.1: Liczba zdarzeń z Polską jako aktorem 1. (źródło: opracowanie własne)



Rys. 5.2: Liczba zdarzeń z Polską jako aktorem 2. (źródło: opracowanie własne)

Polska jako miejsce wydarzeń Wykres 5.3 przedstawia popularność Polski jako miejsca wydarzeń w poszczególnych latach. Ponownie kształt wykresu jest zbliżony do 5.1. W tym przypadku szczyt popularności jest wyższy - na poziomie około 210 tysięcy zdarzeń.



Rys. 5.3: Liczba zdarzeń z Polską jako lokacją. (źródło: opracowanie własne)

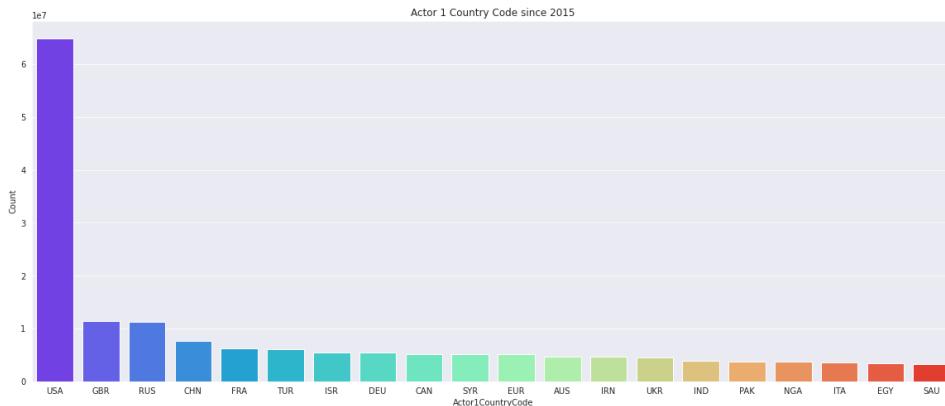
5.1.1 Analiza zbiorcza od 2015 roku

Dane pochodzą z przedziału od stycznia 2015 do kwietnia 2020 roku.

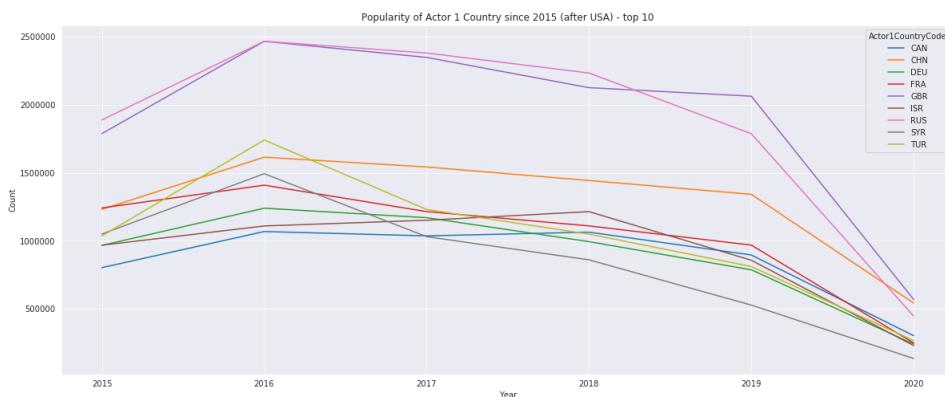
Liczba zdarzeń dla poszczególnych krajów Wykres 5.4 przedstawia sumaryczną liczbę zdarzeń od 2015 roku, dla poszczególnych krajów, uszeregowaną malejąco. Niekwestionowanym liderem pod względem liczby zdarzeń są Stany Zjednoczone. Dystansują one pozostałe kraje o prawie rząd wielkości. Kraje anglosaskie są szczególnie mocno reprezentowane. W czołówce pojawiają się też kraje znaczące politycznie oraz silnie skonfliktowane.

Liczba zdarzeń w czasie Wykres 5.5 przedstawia liczbę zdarzeń dla top 10 krajów w poszczególnych latach (z pominięciem USA). Dla większości krajów szczyt ilości zdarzeń przypada na rok 2016. Wyraźna przewagę nad innymi krajami mają Rosja oraz Wielka Brytania.

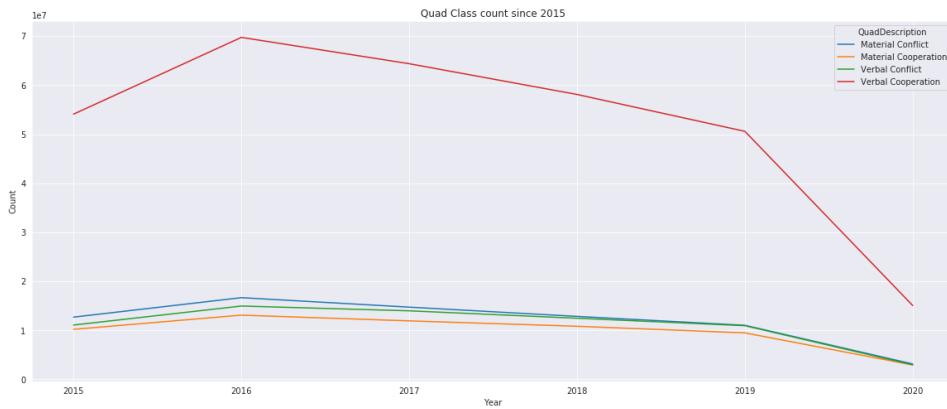
Popularność czterokodów zdarzeń Wykres 5.6 przedstawia sumaryczną liczbę zdarzeń od 2015 roku, dla poszczególnych czterokodów zdarzeń, uszeregowaną malejaco. Obserwujemy podobną tendencję jak na wykresie 5.5.



Rys. 5.4: Liczba zdarzeń dla poszczególnych krajów od 2015 roku. (źródło: opracowanie własne)



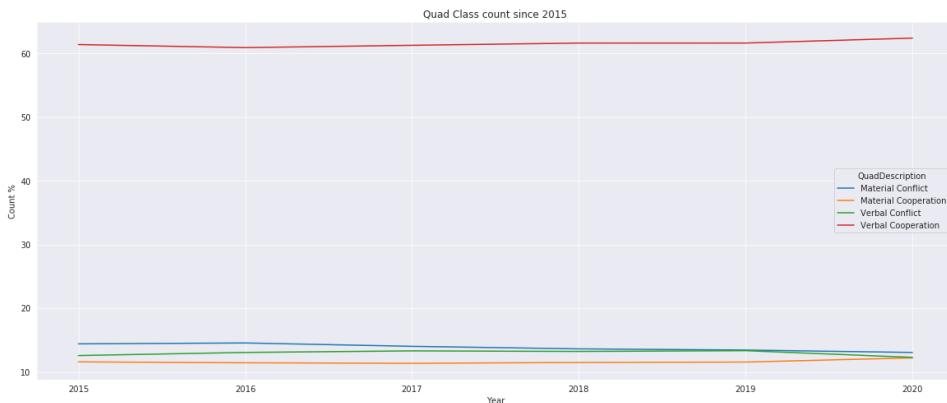
Rys. 5.5: Liczba zdarzeń dla poszczególnych krajów w czasie - top 10. (źródło: opracowanie własne)



Rys. 5.6: Liczba zdarzeń dla poszczególnych czterokodów w czasie.
(źródło: opracowanie własne)

Szczyt ilości zdarzeń przypada na rok 2016. Następnie ilość zdarzeń dla wszystkich czterokodów wyraźnie maleje. Największa ilość zdarzeń przypada kodowi *verbal cooperation*.

Popularność czterokodów zdarzeń w czasie Wykres 5.7 przedstawia procentową liczbę zdarzeń dla czterokodów w poszczególnych latach. Pomi-

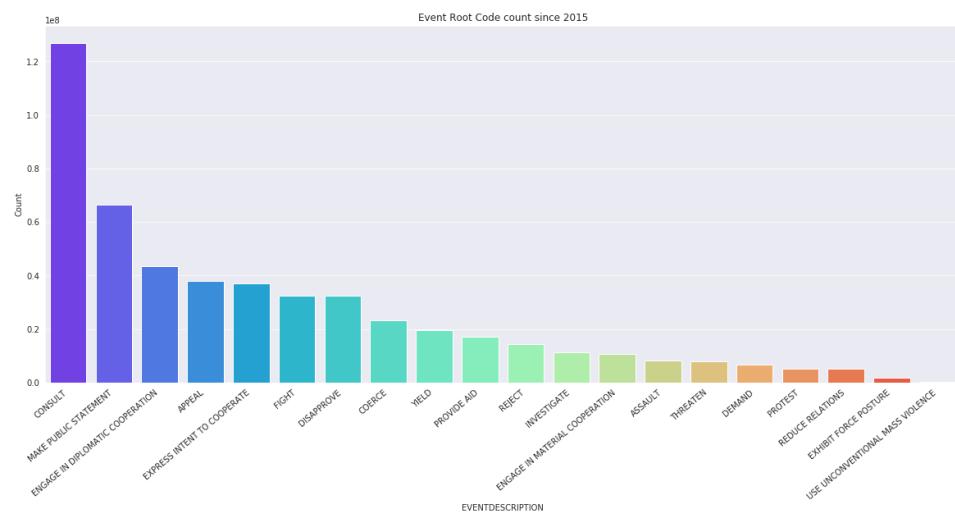


Rys. 5.7: Procentowa liczba zdarzeń dla poszczególnych kodów w czasie.
(źródło: opracowanie własne)

mo spadku ilości zdarzeń obserwowanego na wykresie 5.6 procentowa ilość czterokodów w ciągu lat pozostaje na podobnym poziomie. Podobnie jak

na poprzednim wykresie - największa ilość zdarzeń przypada kodowi *verbal cooperation*.

Popularność podstawowych kodów zdarzeń Wykres 5.8 przedstawia sumaryczną liczbę zdarzeń od 2015 roku, dla wszystkich podstawowych kodów zdarzeń, uszeregowaną malejąco. Obserwujemy prawie dwukrotną prze-

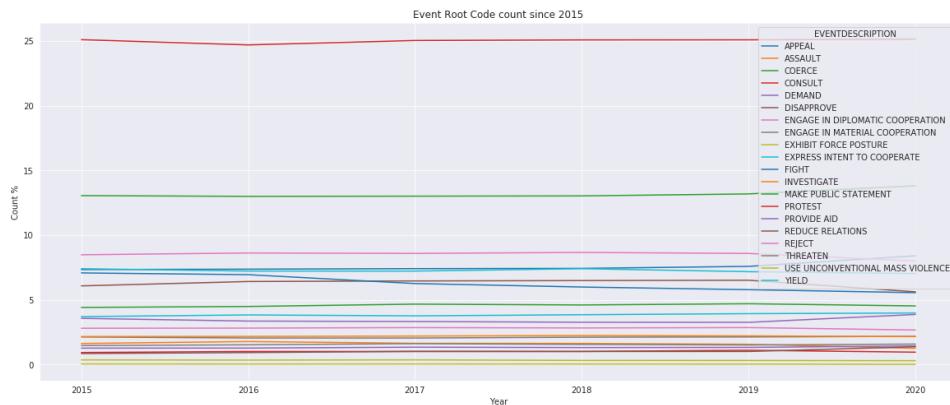


Rys. 5.8: Liczba zdarzeń dla wszystkich kodów podstawowych od 2015 roku. (źródło: opracowanie własne)

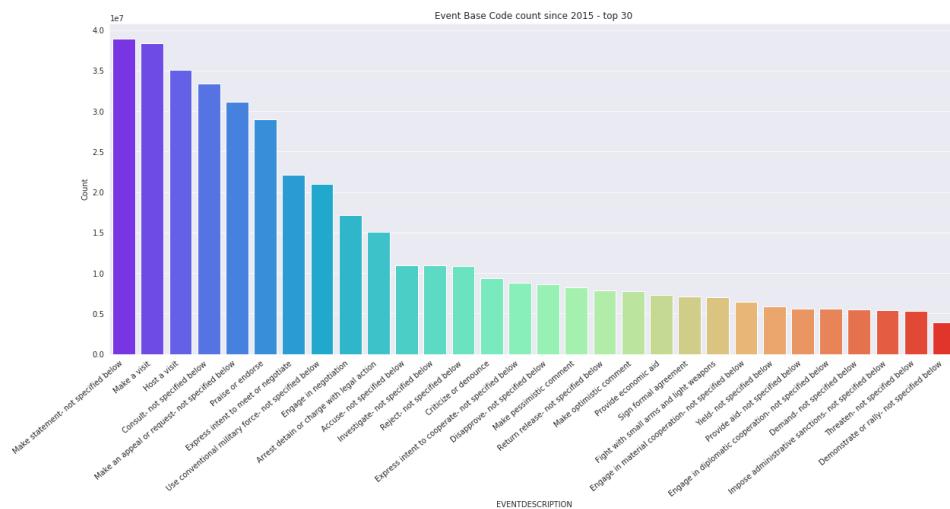
wagę kodu *CONSULT* nad następnym w kolejności kodem *MAKE PUBLIC STATEMENT*.

Popularność podstawowych kodów zdarzeń w czasie Wykres 5.9 przedstawia liczbę zdarzeń dla wszystkich 20 podstawowych kodów zdarzeń w poszczególnych latach. Procentowo liczba zdarzeń dla poszczególnych kodów podstawowych utrzymuje się na tym samym poziomie. Pierwsze trzy kody podstawowe to: *CONSULT*, *COERCE* oraz *ENGAGE IN DIPLOMATIC COOPERATION*.

Popularność bazowych kodów zdarzeń Wykres 5.10 przedstawia sumaryczną liczbę zdarzeń od 2015 roku, dla top 30 bazowych kodów zdarzeń, uszeregowaną malejąco. Największa ilość zdarzeń przypada dla kodów *Make statement- not specified below* oraz *Make a visit, Host a visit*.

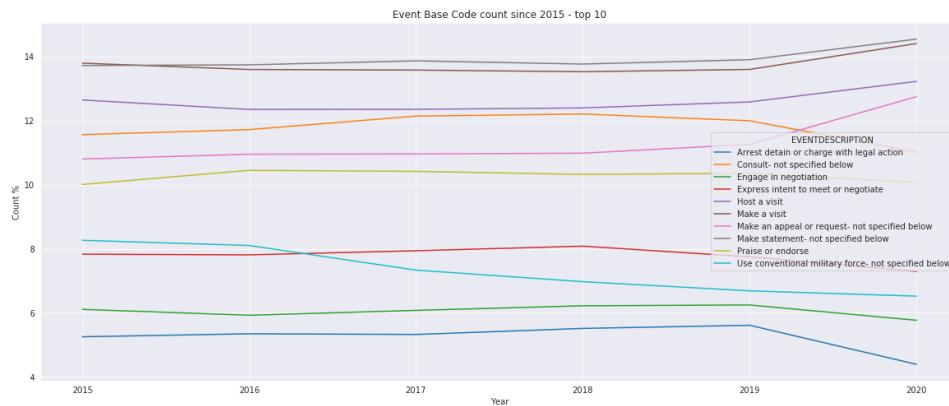


Rys. 5.9: Procentowa liczba zdarzeń dla wszystkich kodów podstawowych w czasie. (źródło: opracowanie własne)



Rys. 5.10: Liczba zdarzeń dla poszczególnych kodów bazowych od 2015 roku - top 30. (źródło: opracowanie własne)

Procentowa popularność bazowych kodów zdarzeń w czasie Wykres 5.11 przedstawia procentową liczbę zdarzeń dla top 10 bazowych kodów zdarzeń w poszczególnych latach. W kolejnych latach procentowa liczba zda-



Rys. 5.11: Procentowa liczba zdarzeń dla poszczególnych kodów bazowych w czasie - top 10. (źródło: opracowanie własne)

rzeń dla kodów bazowych zmienia się tylko nieznacznie. Wyraźną zmianą jest spadek popularności kodu *Use conventional military force- not specified below* w kolejnych latach.

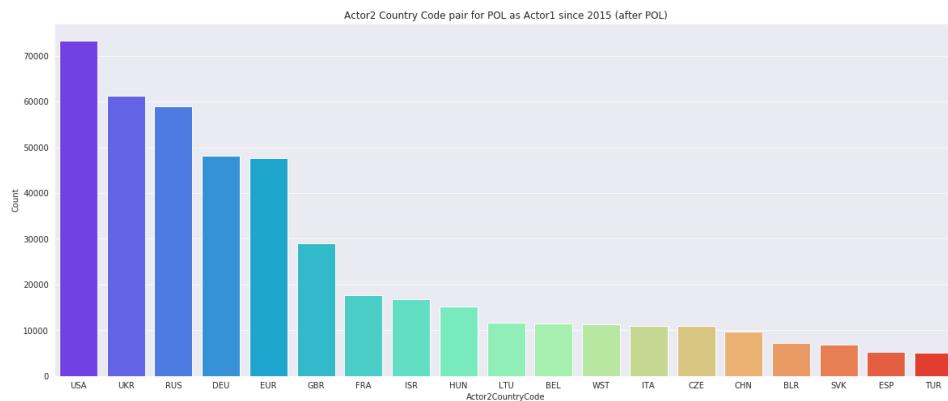
5.2 Analiza danych dla wybranych krajów

5.2.1 Polska

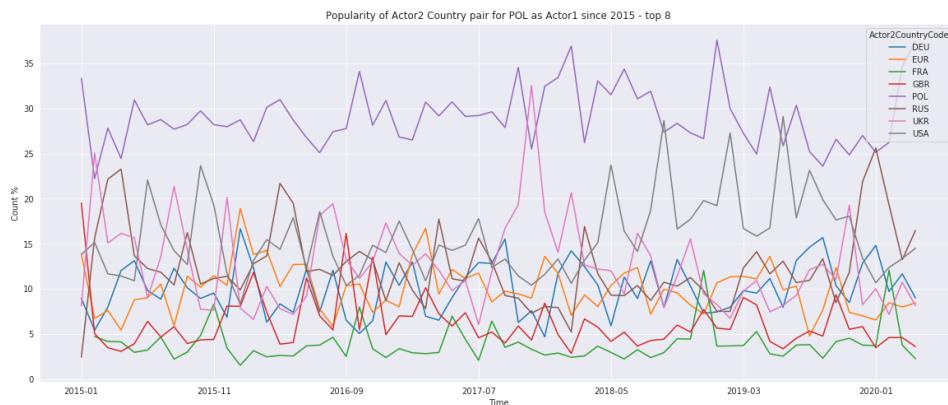
Kraj para do zdarzenia Wykres 5.12 przedstawia liczbę zdarzeń dla Polski w których parą jest dany kraj od 2015 roku.

Najbardziej popularną parą w zdarzeniach dla Polski jest ona sama. Kolejnymi według popularności krajami są Stany Zjednoczone Ameryki, Ukraina oraz Rosja.

Wykres 5.13 przedstawia liczbę zdarzeń dla Polski w których parą jest dany kraj w czasie. Tylko w czterech miesiącach Polska nie jest główną parą zdarzeń dla samej siebie. Dwa razy wyprzedza ją Ukraina, raz Stany Zjednoczone Ameryki, raz Rosja. Począwszy od sierpnia 2017 do lutego 2018 roku obserwujemy dużą liczbę zdarzeń z Ukrainą. Następnie pierwszeństwo przejmują Stany Zjednoczone, a od listopada 2019 roku Rosja.

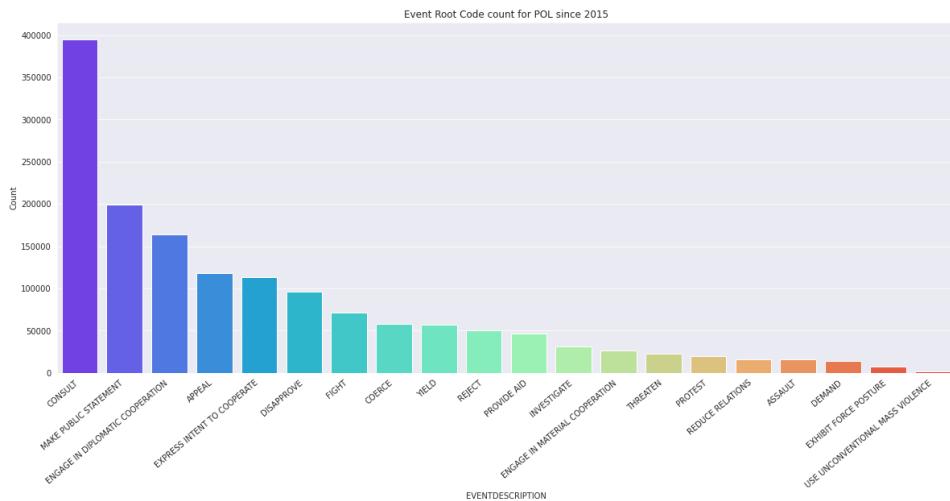


Rys. 5.12: Liczba zdarzeń dla Polski w których para jest dany kraj od 2015 roku. (źródło: opracowanie własne)



Rys. 5.13: Procentowa liczba zdarzeń dla Polski w których para jest dany kraj w kolejnych miesiącach. (źródło: opracowanie własne)

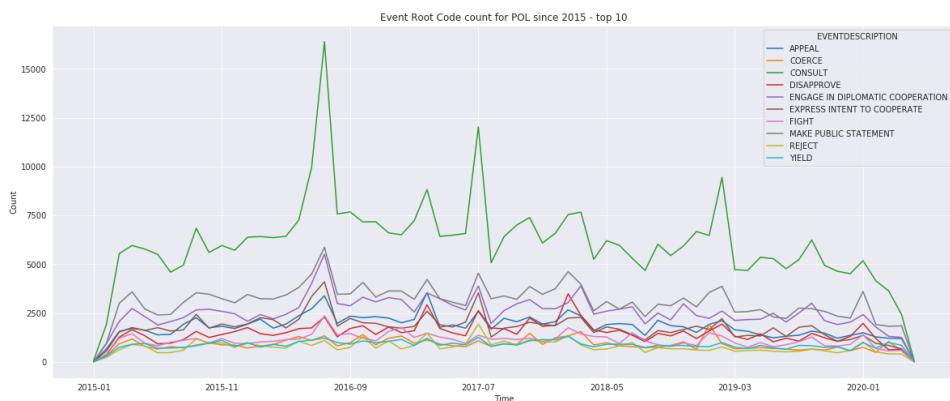
Podstawowy kod zdarzeń Wykres 5.14 przedstawia liczbę zdarzeń z Polską dla poszczególnych kodów podstawowych od 2015 roku. Obserwujemy



Rys. 5.14: Liczba zdarzeń z Polską dla poszczególnych kodów podstawowych od 2015 roku. (źródło: opracowanie własne)

prawie dwukrotną przewagę kodu *CONSULT* nad następnym w kolejności kodem *MAKE PUBLIC STATEMENT*.

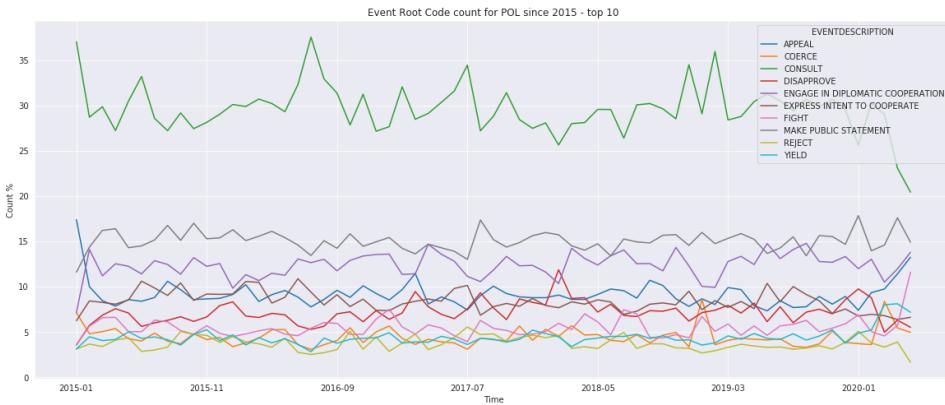
Wykres 5.15 przedstawia liczbę zdarzeń dla Polski dla poszczególnych kodów podstawowych w czasie. Na pierwszym miejscu w całym analizowa-



Rys. 5.15: Liczba zdarzeń z Polską dla poszczególnych kodów podstawowych w czasie - top 10. (źródło: opracowanie własne)

nym okresie utrzymuje się kod *CONSULT*. Następnymi w kolejności są kody *MAKE PUBLIC STATEMENT* oraz *ENGAGE IN DIPLOMATIC COOPERATION*.

Wykres 5.16 przedstawia procentową liczbę zdarzeń dla Polski dla poszczególnych kodów podstawowych w czasie. Proporcja poszczególnych ko-



Rys. 5.16: Procentowa liczba zdarzeń z Polską dla poszczególnych kodów podstawowych w czasie - top 10. (źródło: opracowanie własne)

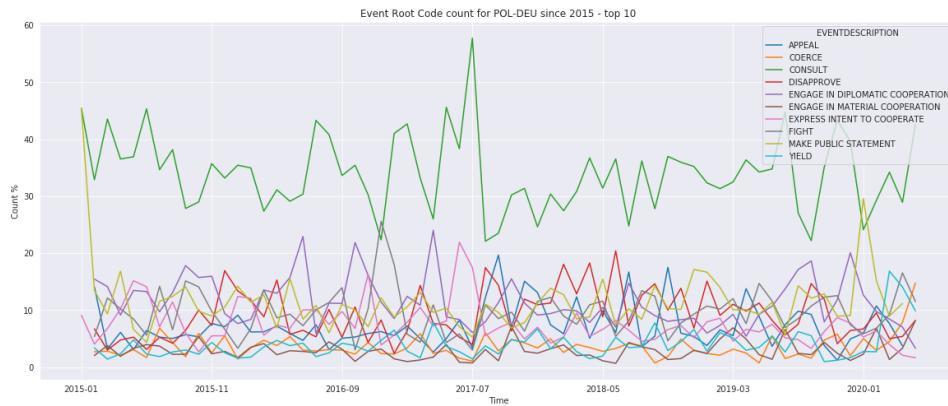
dów podstawowych utrzymuje się, w analizowanym okresie, na podobnym poziomie.

Podstawowe kody zdarzeń między Polską, a wybranymi krajami
Wykres 5.17 przedstawia liczbę zdarzeń z Polską i Niemcami dla poszczególnych kodów podstawowych w czasie. W analizowanym okresie dominuje kod podstawowy *CONSULT*, który tylko w styczniu 2020 roku jest wyprzedzany przez kod *MAKE PUBLIC STATEMENT*. Pozostałe kody utrzymują się na niższym poziomie i podlegają dynamicznym zmianom.

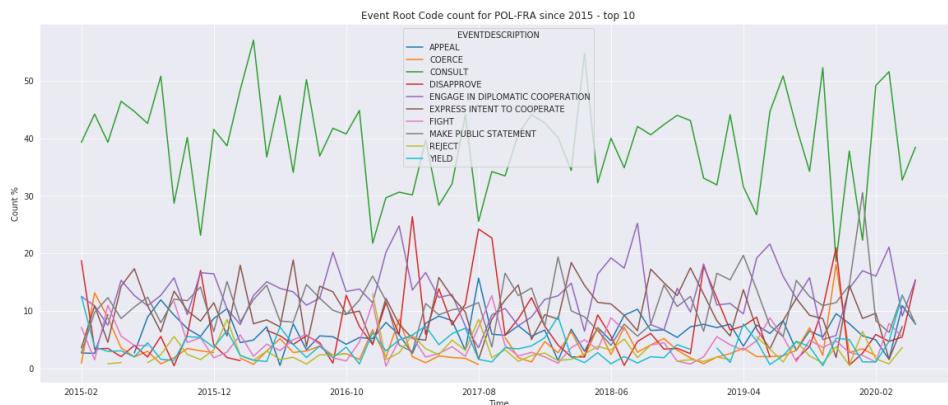
Wykres 5.18 przedstawia liczbę zdarzeń z Polską i Francją dla poszczególnych kodów podstawowych w czasie. Tak jak w przypadku wykresu 5.17 dla Polski i Niemiec, na pierwszym miejscu utrzymuje się kod podstawowy *CONSULT*. Pozostałe kody są mniej popularne i podlegają ciągłym zmianom.

Wykres 5.19 przedstawia liczbę zdarzeń z Polską i Wielką Brytanią dla poszczególnych kodów podstawowych w czasie. Podobnie jak na poprzednich wykresach, dominują zdarzenia *CONSULT*.

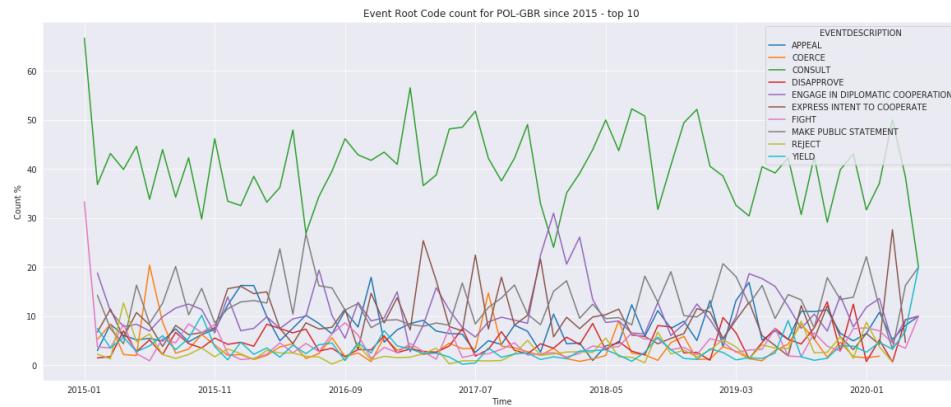
Wykres 5.20 przedstawia liczbę zdarzeń z Polską i Rosją dla poszczególnych kodów podstawowych w czasie. Ponownie dominują zdarzenia *CON-*



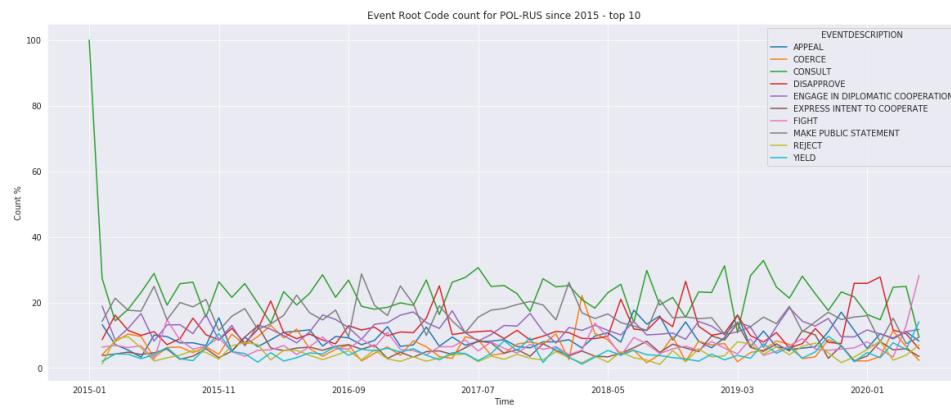
Rys. 5.17: Procentowa liczba zdarzeń z Polska i Niemcami dla poszczególnych kodów podstawowych w czasie - top 10. (źródło: opracowanie własne)



Rys. 5.18: Procentowa liczba zdarzeń z Polską i Francją dla poszczególnych kodów podstawowych w czasie - top 10. (źródło: opracowanie własne)



Rys. 5.19: Procentowa liczba zdarzeń z Polską i Wielką Brytanią dla poszczególnych kodów podstawowych w czasie - top 10. (źródło: opracowanie własne)

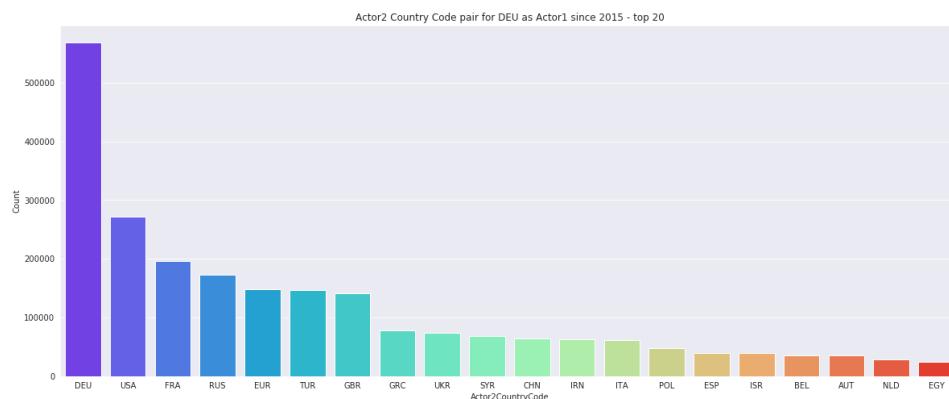


Rys. 5.20: Procentowa liczba zdarzeń z Polską i Rosją dla poszczególnych kodów podstawowych w czasie - top 10. (źródło: opracowanie własne)

SULT jednak są zbliżone do pozostałych. Zauważalne jest częstsze występowanie kodów *MAKE PUBLIC STATEMENT* oraz *DISAPPROVE*.

5.2.2 Niemcy

Kraj para do zdarzenia Wykres 5.21 przedstawia liczbę zdarzeń dla Niemiec w których parą jest dany kraj. Obserwujemy prawie dwukrotnie większą



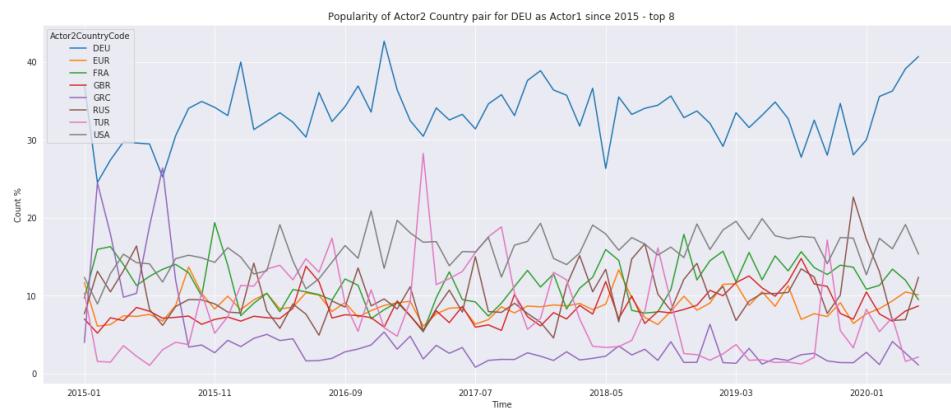
Rys. 5.21: Liczba zdarzeń z Niemcami w których parą jest dany kraj - top 20. (źródło: opracowanie własne)

ilość zdarzeń w których parą dla Niemiec są one same, od następnych w kolejności Stanów Zjednoczonych. Często występującą parą w zdarzeniach są też Francja i Rosja.

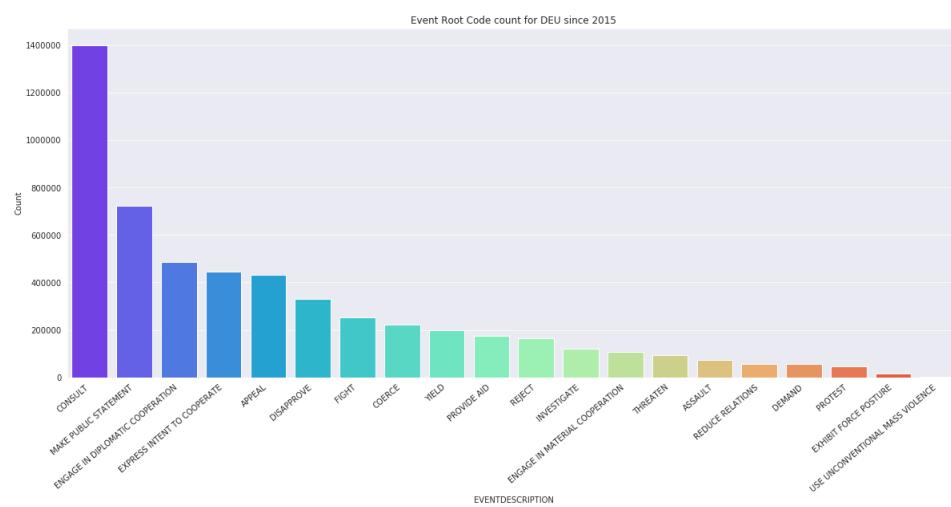
Wykres 5.22 przedstawia procentową liczbę zdarzeń dla Niemiec w których parą jest dany kraj w czasie. Obserwujemy okresowe zwiększenie ilości zdarzeń Niemiec w parze z Grecją, Turcją oraz Rosją. Wyraźnym trendem jest duża ilość zdarzeń w parze ze Stanami Zjednoczonymi Ameryki.

Podstawowy kod zdarzeń Wykres 5.23 przedstawia liczbę zdarzeń z Niemcami dla poszczególnych kodów podstawowych od 2015 roku. Obserwujemy dwukrotnie większą ilość zdarzeń *CONSULT* niż nastepnymi w kolejności zdarzeniami *MAKE PUBLIC STATEMENT*. Często występują też zdarzenia *ENGAGE IN DIPLOMATIC COOPERATION* oraz *EXPRESS INTENT TO COOPERATE*.

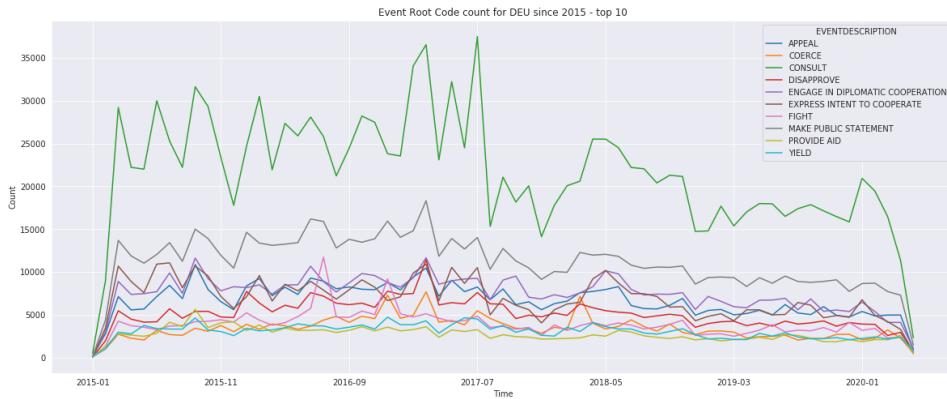
Wykres 5.24 przedstawia liczbę zdarzeń z Niemcami dla poszczególnych kodów podstawowych w czasie - top 10. W całym analizowanym okresie dominują kody podstawowe *CONSULT* oraz *MAKE PUBLIC STATEMENT*.



Rys. 5.22: Procentowa liczba zdarzeń z Niemcami w których para jest dany kraj w czasie. (źródło: opracowanie własne)

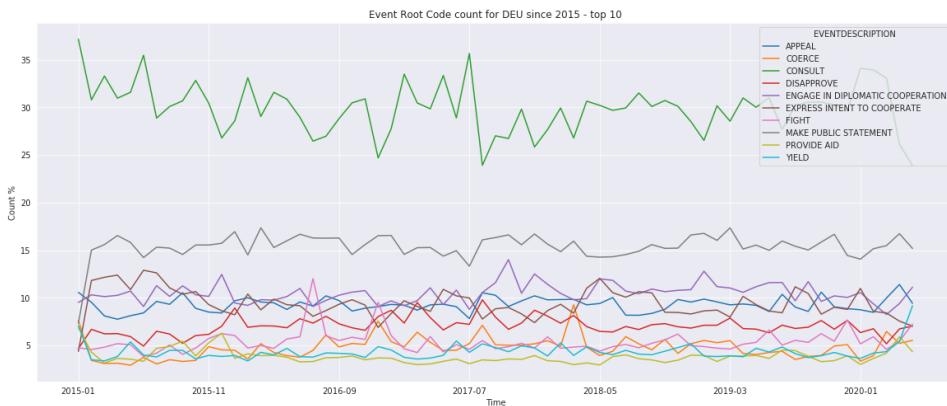


Rys. 5.23: Liczba zdarzeń z Niemcami dla poszczególnych kodów podstawowych od 2015 roku. (źródło: opracowanie własne)



Rys. 5.24: Liczba zdarzeń z Niemcami dla poszczególnych kodów podstawowych w czasie - top 10. (źródło: opracowanie własne)

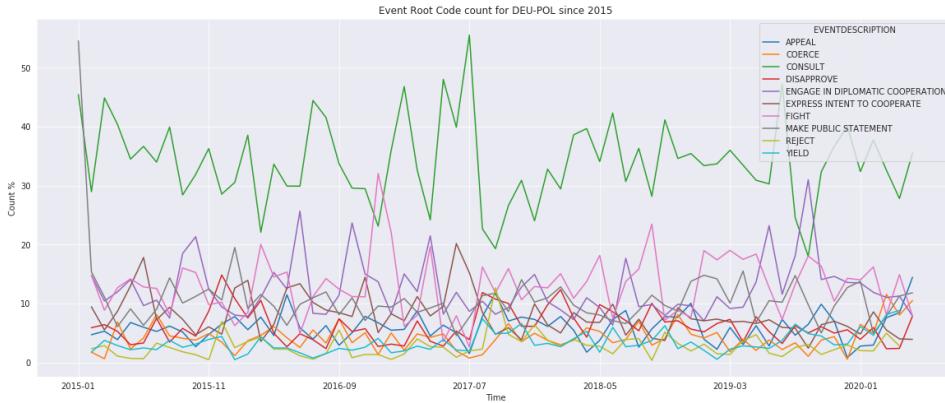
Wykres 5.25 przedstawia procentową liczbę zdarzeń z Niemcami dla poszczególnych kodów podstawowych w czasie. Proporcja poszczególnych ko-



Rys. 5.25: Procentowa liczba zdarzeń z Niemcami dla poszczególnych kodów podstawowych w czasie - top 10. (źródło: opracowanie własne)

dów podstawowych w analizowanym okresie utrzymuje się. Wyraźnie dominuje kod podstawowy *CONSULT*.

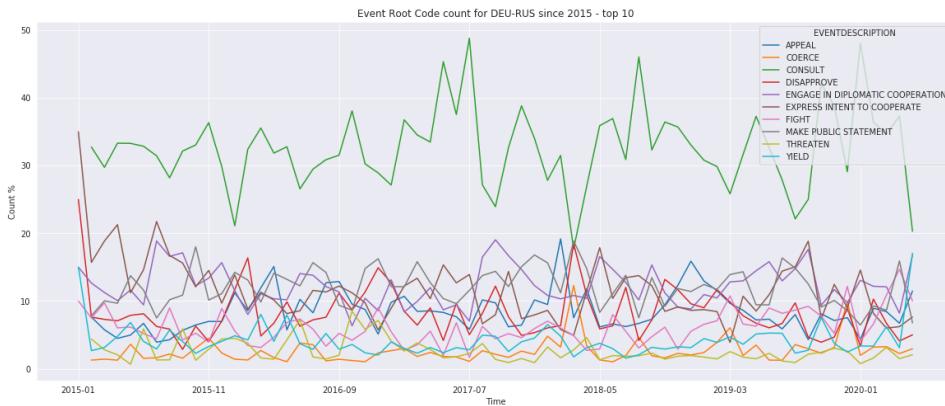
Podstawowe kody zdarzeń między Niemcami, a wybranymi krajami Wykres 5.26 przedstawia liczbę zdarzeń z Niemcami i Polską dla poszczególnych kodów podstawowych w czasie. W całym analizowanym okresie dominują zdarzenia podstawowe *CONSULT*. Obserwujemy też dużą ilość



Rys. 5.26: Procentowa liczba zdarzeń z Niemcami i Polską dla poszczególnych kodów podstawowych w czasie. (źródło: opracowanie własne)

zdarzeń podstawowych *FIGHT*. Może być to spowodowane dużą ilością zdarzeń mających na celu upamiętnienie historycznych działań wojennych.

Wykres 5.27 przedstawia liczbę zdarzeń z Niemcami i Rosją dla poszczególnych kodów podstawowych w czasie. Dominującym jest kod podstawi-

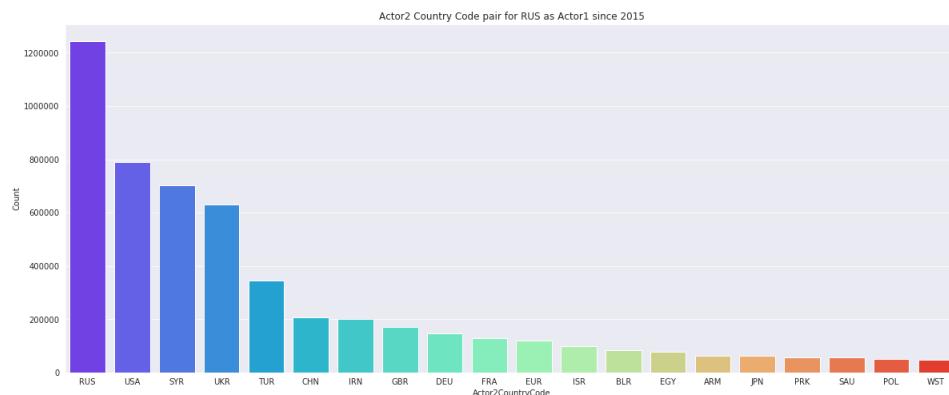


Rys. 5.27: Procentowa liczba zdarzeń z Niemcami i Rosją dla poszczególnych kodów podstawowych w czasie. (źródło: opracowanie własne)

wy *CONSULT*. Obserwujemy okresy wyraźnego zwiększenia ilości zdarzeń z kodem podstawowym *DISAPPROVE* i w tym samym czasie zmniejszenia zdarzeń *CONSULT*.

5.2.3 Rosja

Kraj para do zdarzenia Wykres 5.28 przedstawia liczbę zdarzeń dla Rosji w których parą jest dany kraj od 2015 roku.



Rys. 5.28: Liczba zdarzeń z Rosją w których parą jest dany kraj od 2015 roku. (źródło: opracowanie własne)

W porównaniu z wcześniej analizowanymi państwami, gdzie drugi w kolejności kraj miał o połowę mniej zdarzeń, w przypadku Rosji zdarzeń jest tylko o 1/3 mniej. Najpopularniejszymi parami do zdarzeń dla Rosji, prócz niej samej, są Stany Zjednoczone Ameryki, Syria oraz Ukraina.

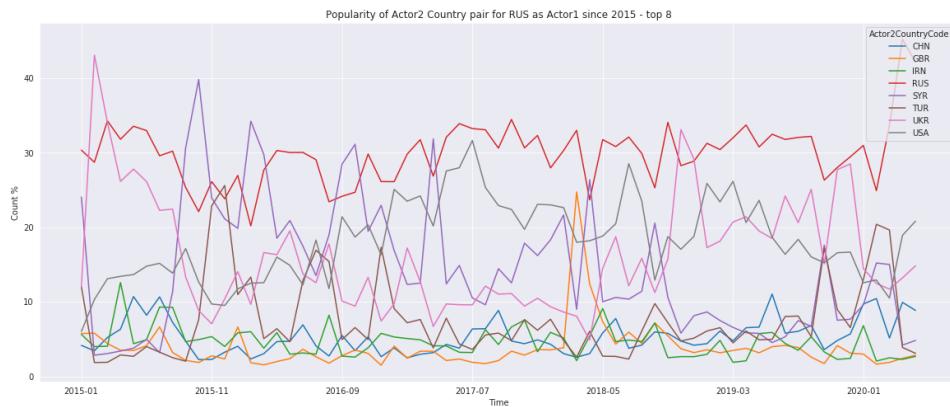
Wykres 5.29 przedstawia procentową liczbę zdarzeń dla Rosji w których parą jest dany kraj w czasie. Obserwujemy okresową dominację zdarzeń w których parą jest Ukraina, Syria oraz Turcja.

Podstawowy kod zdarzeń Wykres 5.30 przedstawia liczbę zdarzeń z Rosją, dla poszczególnych kodów podstawowych od 2015 roku.

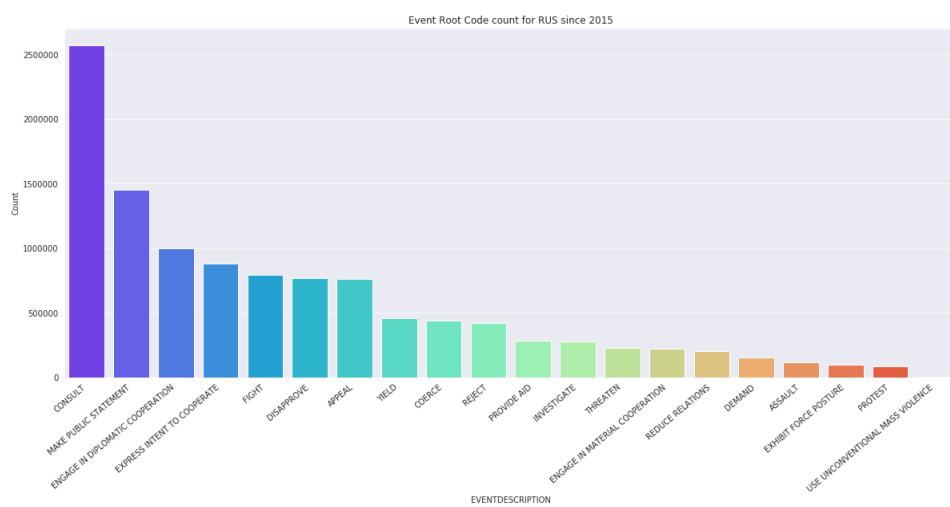
Najliczniej występują zdarzenia z kodami podstawowymi *CONSULT*, *MAKE PUBLIC STATEMENT* oraz *ENGAGE IN DIPLOMATIC COOPERATION*.

Wykres 5.31 przedstawia liczbę zdarzeń z Rosją, dla poszczególnych kodów podstawowych w czasie. Dominuje występowanie kodu podstawowego *CONSULT*. Na drugim miejscu znajduje się kod *EXPRESS INTENT TO COOPERATE*, tylko przez moment wyprzedzony przez kod *FIGHT*.

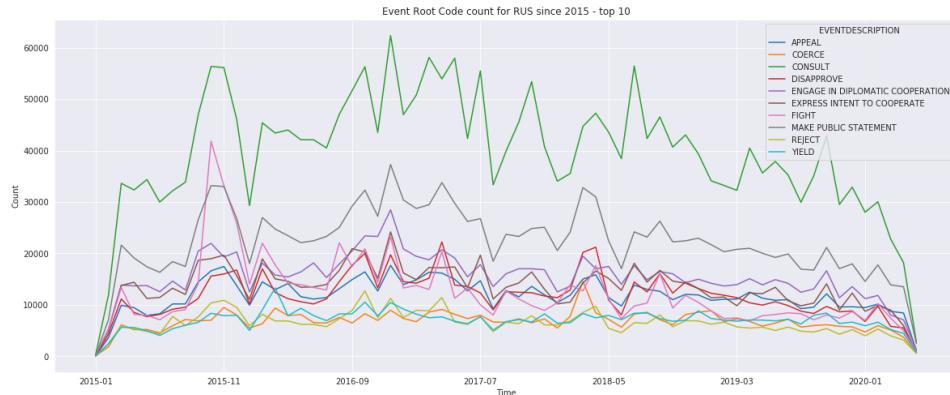
Wykres 5.32 przedstawia procentową liczbę zdarzeń z Rosją dla poszczególnych kodów podstawowych w czasie. Stosunek występowania poszczególnych kodów pozostaje na podobnym poziomie. Wyraźnym odstępstwem jest czasowe zwiększenie popularności kodu podstawowego *FIGHT*.



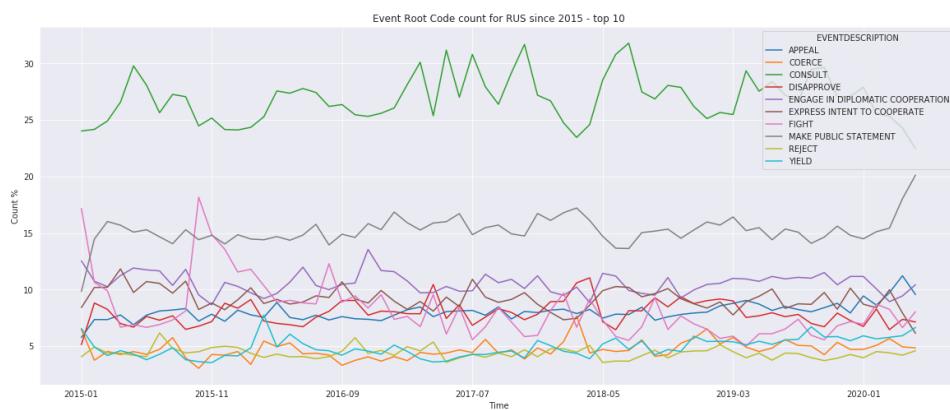
Rys. 5.29: Procentowa liczba zdarzeń dla Rosji w których para jest dany kraj w czasie - top 8. (źródło: opracowanie własne)



Rys. 5.30: Liczba zdarzeń dla poszczególnych kodów podstawowych od 2015 roku. (źródło: opracowanie własne)



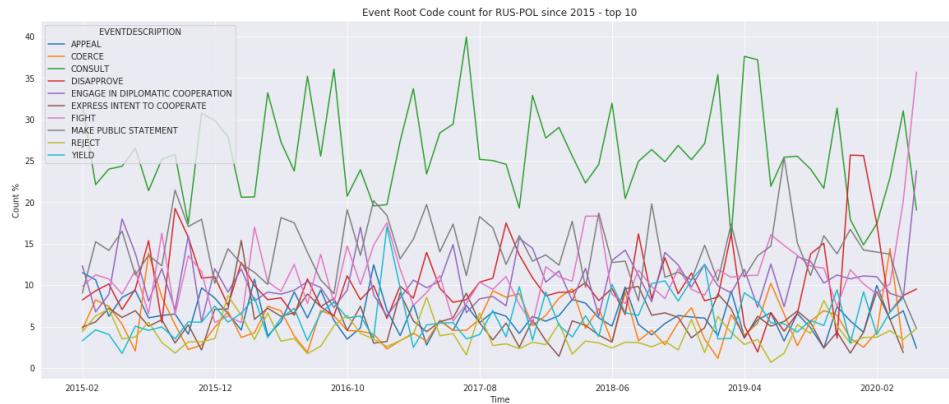
Rys. 5.31: Liczba zdarzeń dla poszczególnych kodów podstawowych w czasie - top 10. (źródło: opracowanie własne)



Rys. 5.32: Procentowa liczba zdarzeń z Rosją dla poszczególnych kodów podstawowych w czasie - top 10. (źródło: opracowanie własne)

Podstawowe kody zdarzeń między Rosją, a wybranymi krajami

Wykres 5.33 przedstawia liczbę zdarzeń z Rosją i Polską dla poszczególnych kodów podstawowych w czasie. W wydarzeniach w których uczestniczy Rosja



Rys. 5.33: Procentowa liczba zdarzeń z Rosją i Niemcami dla poszczególnych kodów podstawowych w czasie - top 10. (źródło: opracowanie własne)

i Polska obserwujemy dominację kodu podstawowego *CONSULT*. Występują wyraźne okresy zwiększenia popularności kodów *DISAPPROVE* oraz *MAKE PUBLIC STATEMENT* przy jednoczesnym zmniejszeniu popularności kodu *CONSULT*.

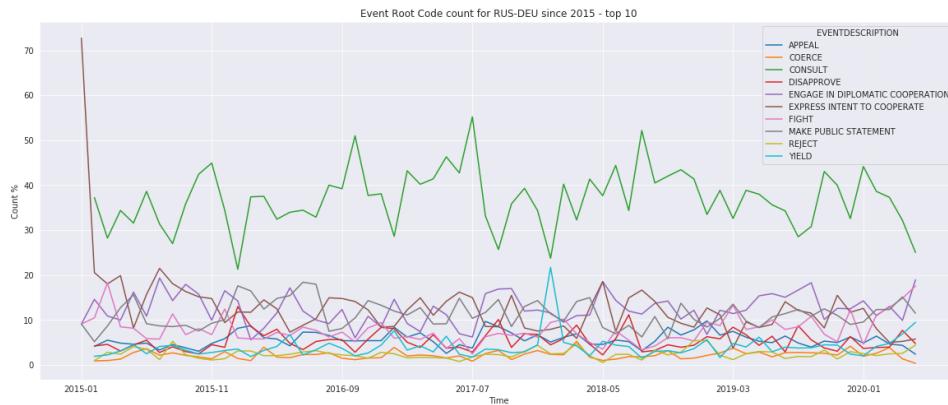
Wykres 5.34 przedstawia liczbę zdarzeń z Rosją i Niemcami dla poszczególnych kodów podstawowych w czasie. Popularność kodu podstawowego *CONSULT* w analizowanym okresie oscyluje w okolicach 40%, podczas gdy pozostałych kodów utrzymuje się w większości poniżej 20%.

5.2.4 Stan Zjednoczone

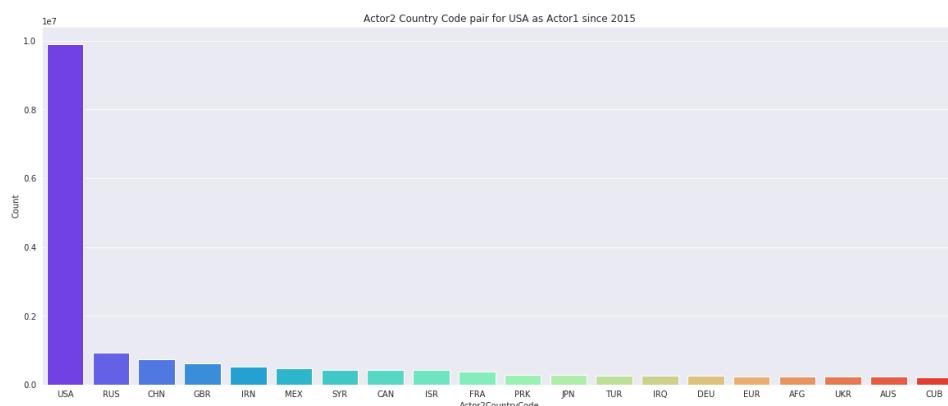
Kraj para do zdarzenia Wykres 5.35 przedstawia liczbę zdarzeń dla Stanów Zjednoczonych w których parą jest dany kraj.

Liczba zdarzeń w których parą dla Stanów Zjednoczonych są one same jest o rząd wielkości większa od następnej w kolejności Rosji. Stany Zjednoczone mają też dużo zdarzeń w których parą są Chiny, Wielka Brytania, Iran oraz Meksyk.

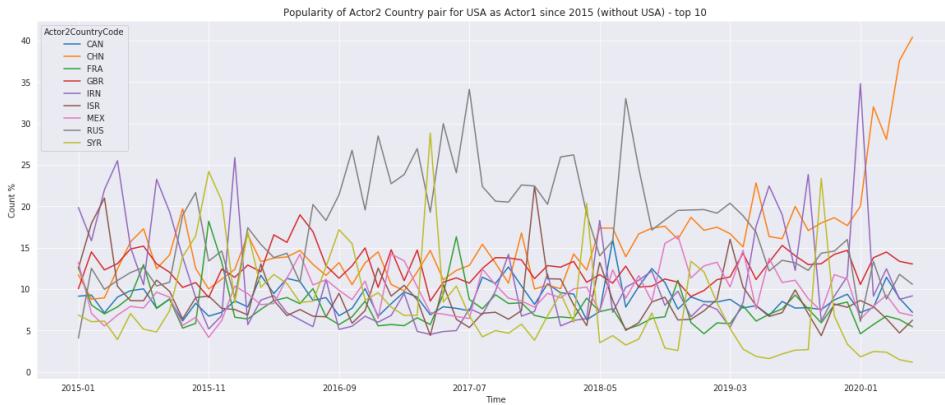
Wykres 5.36 przedstawia procentową liczbę zdarzeń dla Stanów Zjednoczonych w których parą jest dany kraj w czasie. Przez większość analizowanego okresu główną parą do zdarzeń dla Stanów Zjednoczonych (z pominięciem



Rys. 5.34: Procentowa liczba zdarzeń z Rosją i Niemcami dla poszczególnych kodów podstawowych w czasie - top 10. (źródło: opracowanie własne)



Rys. 5.35: Liczba zdarzeń ze Stanami Zjednoczonymi w których para jest dany kraj od 2015 roku. (źródło: opracowanie własne)



Rys. 5.36: Procentowa liczba zdarzeń ze Stanami Zjednoczonymi w których para jest dany kraj w czasie (z pominięciem USA)- top 10. (źródło: opracowanie własne)

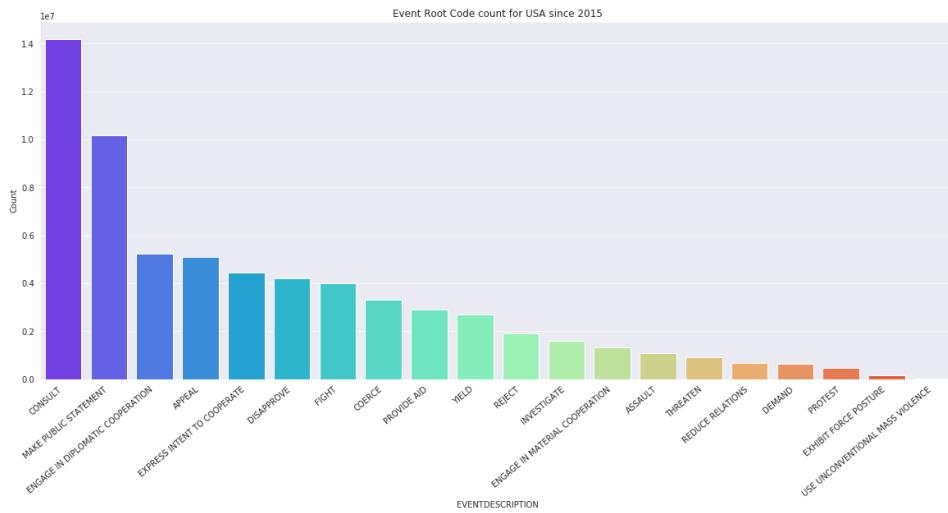
ich samych) jest Rosja. Jako dominująca para do zdarzeń występują także Iran, Syria, Wielka Brytania, Chiny oraz Izrael.

Podstawowy kod zdarzeń Wykres 5.37 przedstawia liczbę zdarzeń ze Stanami Zjednoczonymi dla poszczególnych kodów podstawowych od 2015 roku.

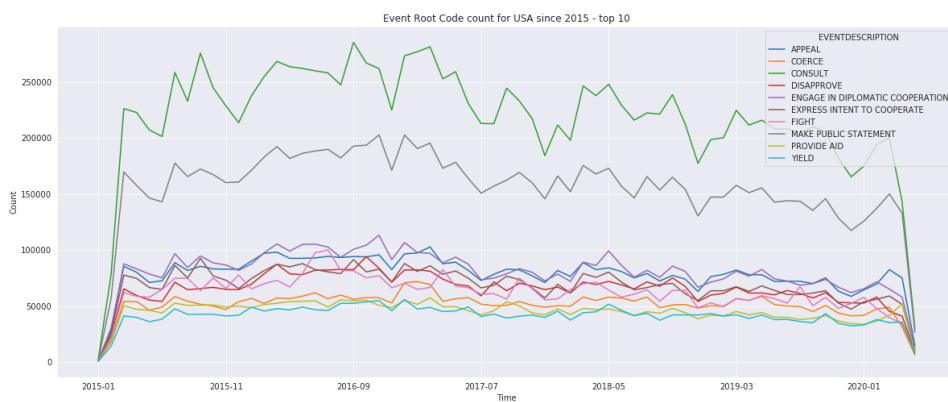
Tak jak w przypadku poprzednio analizowanych krajów dominują zdarzenia z kodem podstawowym *CONSULT*. Kolejnymi kodami według liczby występujących zdarzeń są *MAKE PUBLIC STATEMENT*, *ENGAGE IN DIPLOMATIC COOPERATION* oraz *APPEAL*.

Wykres 5.38 przedstawia liczbę zdarzeń ze Stanami Zjednoczonymi dla poszczególnych kodów podstawowych w czasie. W całym analizowanym okresie na pierwszych dwóch miejscach pod względem liczby zdarzeń występują kody *CONSULT* oraz *MAKE PUBLIC STATEMENT*. Pozostałe kody zdarzeń charakteryzują się znacznie mniejszą ilością zdarzeń.

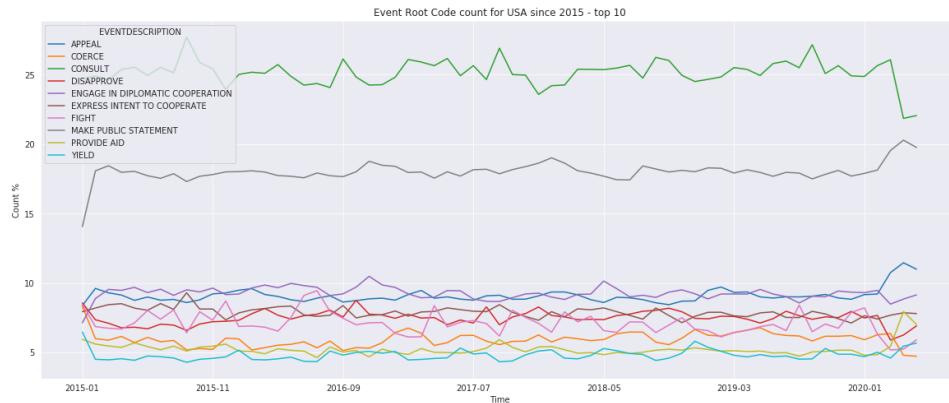
Wykres 5.39 przedstawia procentową liczbę zdarzeń ze Stanami Zjednoczonymi dla poszczególnych kodów podstawowych w czasie. Proporcja między kodami zdarzeń podstawowych w analizowanym okresie czasu pozostaje na podobnym poziomie. Dominują kod podstawowy *CONSULT* (oscyuluje w okolicach 25%) oraz kod podstawowy *MAKE PUBLIC STATEMENT* (oscyuluje w okolicach 17%).



Rys. 5.37: Liczba zdarzeń ze Stanami Zjednoczonymi dla poszczególnych kodów podstawowych od 2015 roku. (źródło: opracowanie własne)

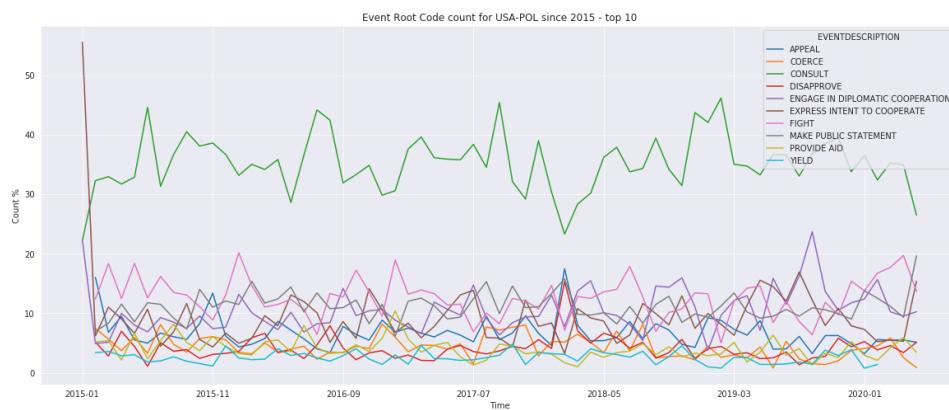


Rys. 5.38: Liczba zdarzeń ze Stanami Zjednoczonymi dla poszczególnych kodów podstawowych w czasie - top 10. (źródło: opracowanie własne)



Rys. 5.39: Procentowa liczba zdarzeń ze Stanami Zjednoczonymi dla poszczególnych kodów podstawowych w czasie - top 10. (źródło: opracowanie własne)

Podstawowe kody zdarzeń między Stanami Zjednoczonymi, a wybranymi krajami Wykres 5.40 przedstawia liczbę zdarzeń ze Stanami Zjednoczonymi i Polską dla poszczególnych kodów podstawowych w czasie. Obserwujemy dominację zdarzeń z kodem podstawowym *CONSULT* z po-

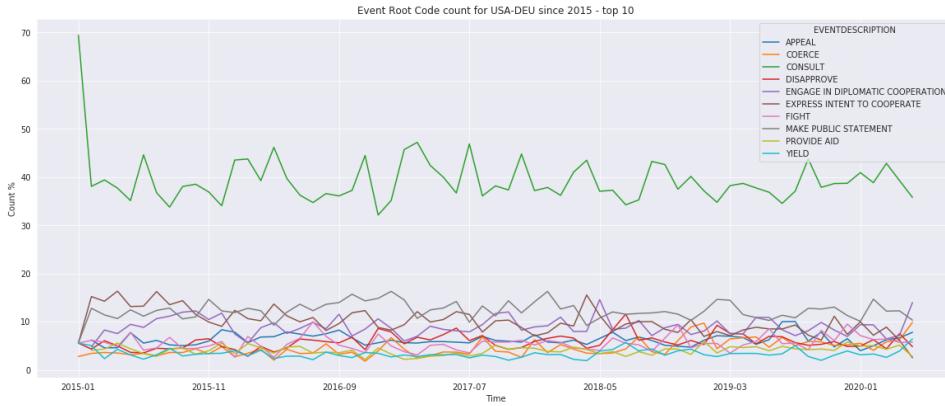


Rys. 5.40: Procentowa liczba zdarzeń ze Stanami Zjednoczonymi i Polską dla poszczególnych kodów podstawowych w czasie - top 10. (źródło: opracowanie własne)

pularnością oscylującą około 40%. Pozostałe kody w analizowanym okresie utrzymują się głównie poniżej 20% wystąpień.

Wykres 5.41 przedstawia liczbę zdarzeń ze Stanami Zjednoczonymi i Niem-

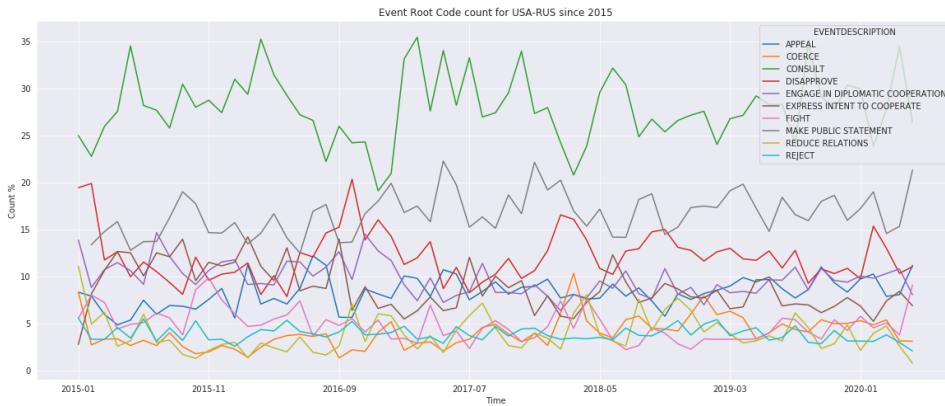
cami dla poszczególnych kodów podstawowych w czasie. W porównaniu z po-



Rys. 5.41: Procentowa liczba zdarzeń ze Stanami Zjednoczonymi i Niemcami dla poszczególnych kodów podstawowych w czasie - top 10. (źródło: opracowanie własne)

przednim wykresem 5.40 obserwujemy jeszcze większą dominację zdarzeń z kodem podstawowym *CONSULT*. Pozostałe kody w analizowanym okresie utrzymują się głównie poniżej 15% wystąpień.

Wykres 5.42 przedstawia liczbę zdarzeń ze Stanami Zjednoczonymi i Rosją dla poszczególnych kodów podstawowych w czasie. Obserwujemy mniej-



Rys. 5.42: Procentowa liczba zdarzeń ze Stanami Zjednoczonymi i Rosją dla poszczególnych kodów podstawowych w czasie - top 10. (źródło: opracowanie własne)

szy niż na poprzednich wykresach 5.40, 5.41 odstęp pomiędzy popularnością

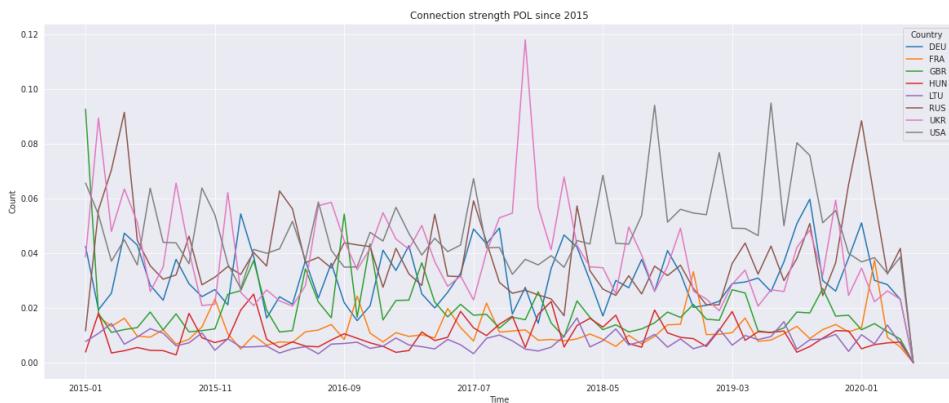
zdarzeń z kodem *CONSULT*, a następnie w kolejności *MAKE PUBLIC STATEMENT* oraz *DISAPPROVE*.

5.3 Analiza siły powiązania

W tej części przeanalizowana zostanie siła powiązania pomiędzy wybranymi krajami, a także jej symetryczność. Siła powiązania, zaproponowana w artykule [22], obliczana jest jako stosunek liczby zdarzeń pomiędzy krajem A, a krajem B, do liczby wszystkich zdarzeń w których kraj A jest aktorem 1. Ponieważ siła powiązania nie jest znormalizowana przez liczbę zdarzeń dla kraju B, dlatego nie jest symetryczna. Odzwierciedla to jak ważny dla kraju A jest kraj B.

5.3.1 Analiza siły powiązania między wybranymi krajami

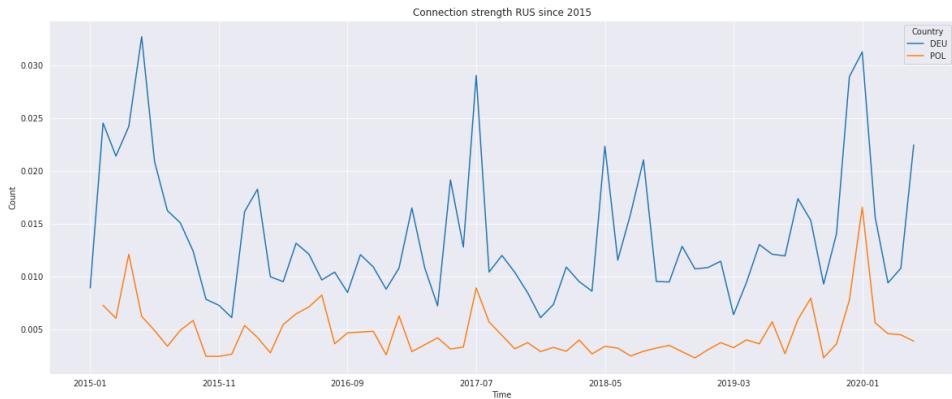
Polska Wykres 5.43 przedstawia siłę połączenia Polski z wybranymi krajami w czasie. Pierwsza pozycja pod względem siły powiązania zmienia się



Rys. 5.43: Siła połączenia Polski z wybranymi krajami w czasie. (źródło: opracowanie własne)

w czasie. Od kwietnia 2018 do października 2019 zajmowały ją Stany Zjednoczone Ameryki. W pozostałym okresie obserwujemy największą siłę połączenia z Rosją oraz Ukrainą.

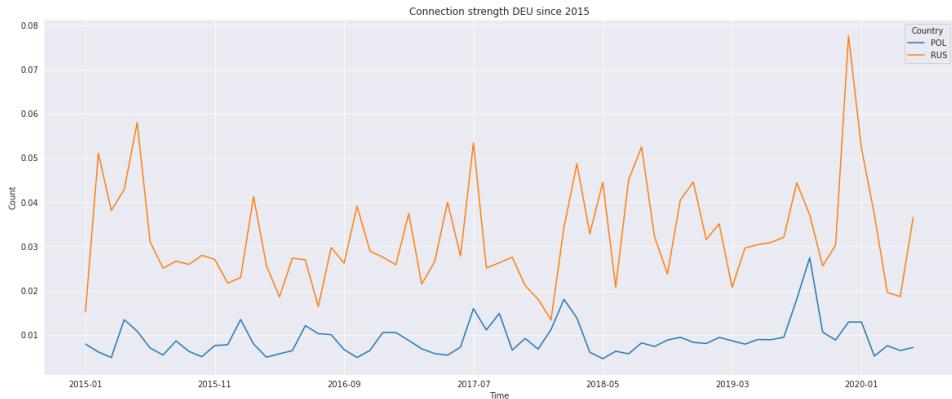
Rosja Wykres 5.44 przedstawia siłę połączenia Rosji z Niemcami oraz Polską w czasie. W całym analizowanym okresie siła połączenia między Rosją,



Rys. 5.44: Siła połączenia Rosji z wybranymi krajami w czasie. (źródło: opracowanie własne)

a Niemcami jest większa od siły połączenia Rosji z Polską. W czterech przypadkach obserwujemy wyraźną korelację wzrostu siły połączenia z Polską i z Niemcami.

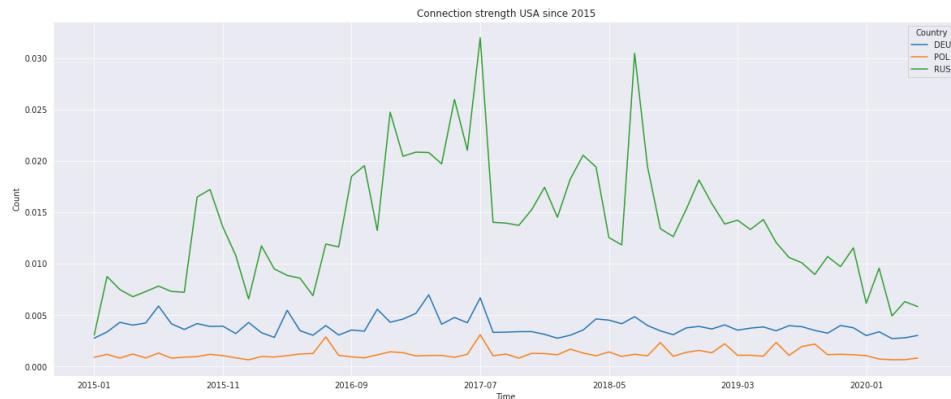
Niemcy Wykres 5.45 przedstawia siłę połączenia Niemiec z wybranymi krajami w czasie. Podobnie jak na wykresie 5.44 w całym analizowanym



Rys. 5.45: Siła połączenia Niemiec z wybranymi krajami w czasie. (źródło: opracowanie własne)

okresie siła połączenia z Polską jest mniejsza od siły połączenia z Rosją.

Stany Zjednoczone Wykres 5.46 przedstawia siłę połączenia Stanów Zjednoczonych z wybranymi krajami w czasie.



Rys. 5.46: Siła połączenia Stanów Zjednoczonych z wybranymi krajami w czasie. (źródło: opracowanie własne)

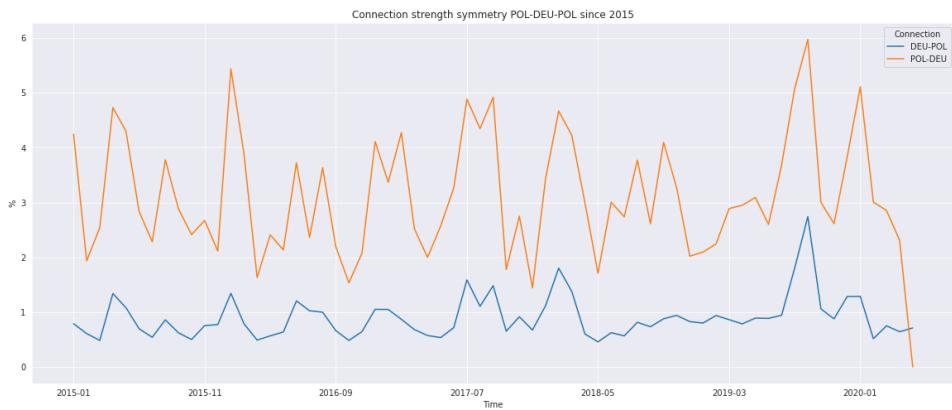
Obserwujemy znacznie wyższą siłę połączenia Stanów Zjednoczonych z Rosją, niż z Niemcami oraz Polską. W lipcu 2017 obserwujemy korelację wzrostu siły połączenia Stanów Zjednoczonych z 3 analizowanymi krajami. W okresie od września 2017 do października 2018 obserwujemy zwiększoną siłę połączenia Stanów Zjednoczonych z Rosją.

5.3.2 Analiza symetryczności siły powiązania

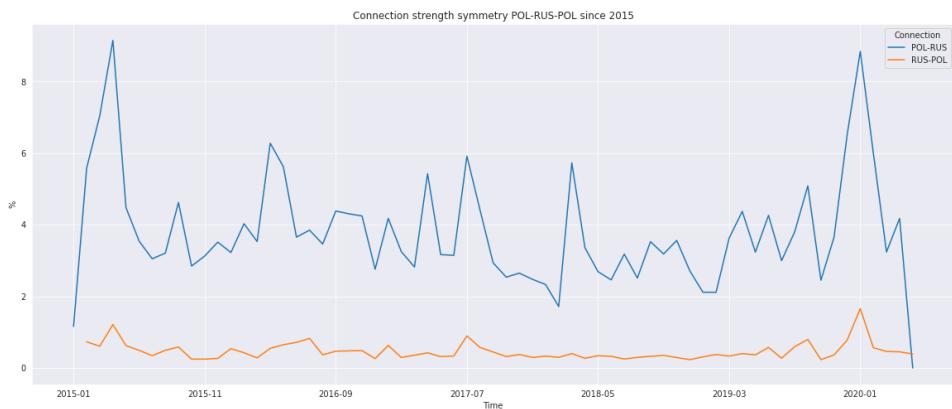
Polska - Niemcy - Polska Wykres 5.47 przedstawia symetryczność siły połączenia Polski i Niemiec w czasie. Obserwujemy okresy korelacji siły połączenia Polski z Niemcami oraz Niemiec z Polską. W całym analizowanym okresie połączenie Polski z Niemcami jest silniejsze niż Niemiec z Polską.

Polska - Rosja - Polska Wykres 5.48 przedstawia symetryczność siły połączenia Polski i Rosji w czasie. W porównaniu z wykresem 5.47 obserwujemy słabsze połączenie Rosji z Polską niż Niemiec z Polską oraz silniejsze połączenie Polski z Rosją niż Polski z Niemcami.

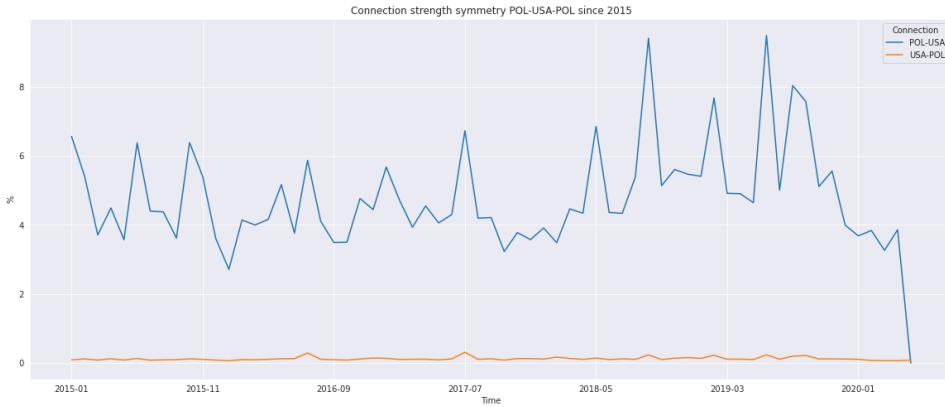
Polska - Stany Zjednoczone - Polska Wykres 5.49 przedstawia symetryczność siły połączenia Polski i Stanów Zjednoczonych w czasie. Spośród analizowanych sił powiązania między Polską, a Stanami Zjednoczonymi obserwujemy największą niesymetryczność relacji. Polska jest znacznie silniej



Rys. 5.47: Symetryczność siły połączenia Polski i Niemiec w czasie.
(źródło: opracowanie własne)



Rys. 5.48: Symetryczność siły połączenia Polski i Rosji w czasie. (źródło:
opracowanie własne)

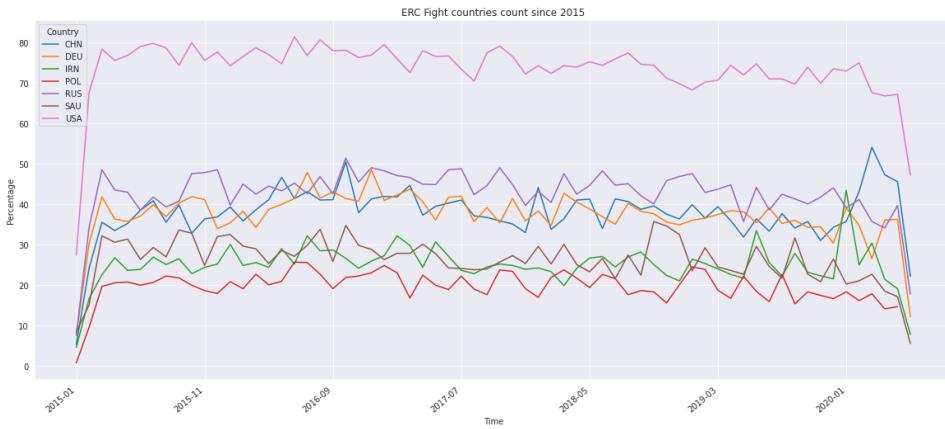


Rys. 5.49: Symetryczność siły połączenia Polski i Stanów Zjednoczonych w czasie. (źródło: opracowanie własne)

powiązana ze Stanami Zjednoczonymi niż Stany Zjednoczone z Polską. Siła połączenia Stanów Zjednoczonych z Polską w całym analizowanym okresie utrzymuje się poniżej 0,5%.

5.4 Analiza kodu podstawowego Fight

Wykres 5.50 przedstawia procentową liczbę państw z jakimi dany wybrany kraj ma zdarzenia *fight* w czasie. Obserwujemy wystąpienie zdarzeń



Rys. 5.50: Procentowa liczba państw z jakimi dany wybrany kraj ma zdarzenia *fight* w czasie. (źródło: opracowanie własne)

fight dla Polski mimo, iż nie prowadziła ona działań wojennych w tym okresie. Spowodowane jest to możliwością klasyfikowania jako *fight* uroczystości mających na celu upamiętnienie wydarzeń historycznych.

Rozdział 6

Grupowanie państw o podobnych cechach oraz porównanie z danymi zewnętrznymi

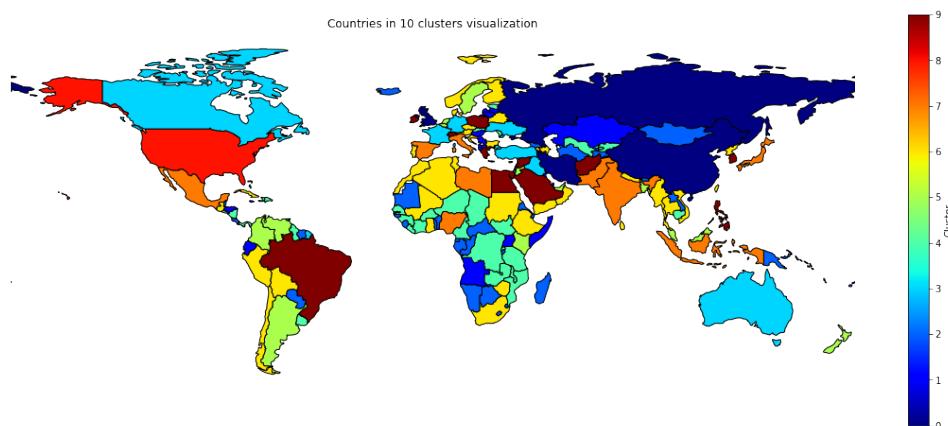
W tym rozdziale zostały opisane wyniki grupowania państw, z wykorzystaniem metody k-średnich, w oparciu o podobne cechy z bazy danych GDELT. Dane zostały porównane z danymi zewnętrznymi. W części 6.1 przedstawione zostały wyniki grupowania danych niestandardyzowanych. Część 6.2 zawiera wyniki grupowania danych ustandardyzowanych przy pomocy StandardScaler'a. Do grupowania krajów został wykorzystany wektor cech opisany w rozdziale 3.

6.1 Dane niestandardyzowane

W pierwszej kolejności zostaną przedstawione wyniki grupowania na danych niestandardyzowanych.

Wykres 6.1 przedstawia mapę z naniesionymi wynikami klasteryzacji. Każda grupa krajów otrzymała inny kolor. Obserwujemy skupiska geograficzne krajów należących do danego klastra. Kraje sąsiadujące ze sobą mogą charakteryzować się podobnymi cechami, dlatego często należą do tego samego klastra.

Szczegółowe wyniki grupowania na danych niestandardyzowanych znajdują się w tabelach w dodatku A.



Rys. 6.1: Mapa z wynikami klasteryzacji z wykorzystaniem danych niestandardyzowanych. (źródło: opracowanie własne)

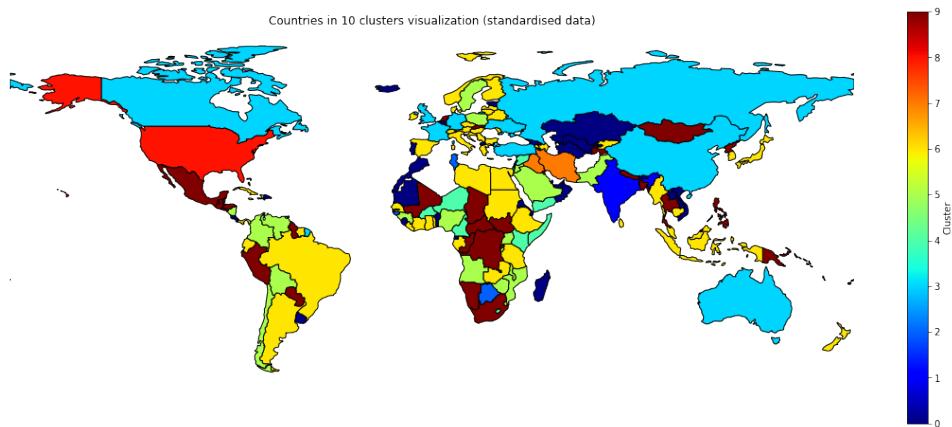
6.2 Dane ustANDARDYZOWANE

Standrd Scaler standaryzuje cechy poprzez usunięcie średniej i skalowanie do wariancji jednostkowej. Standardowy wynik próbki x jest obliczany jako: $z = (x - u) / s$ gdzie u jest średnią próbki, a s jest standardowym odchyleniem próbki.

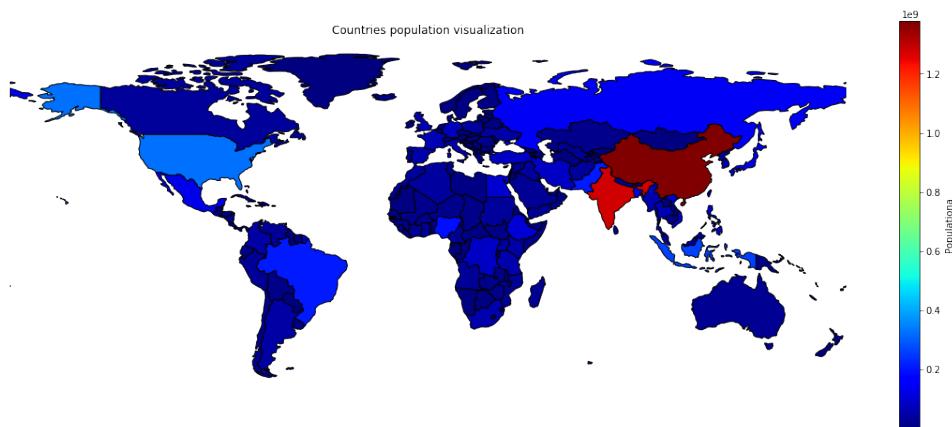
Wykres 6.2 przedstawia mapę z naniesionymi wynikami klasteryzacji ustANDARDYZOWANYCH próbek. Każda grupa krajów otrzymała inny kolor. Podobnie jak w wypadku klastrowania w oparciu o dane niestandardyzowane zauważalne są skupiska geograficzne krajów należących do danego klastra. Najbardziej zauważalną różnicą w stosunku do klastrowania o dane niestandardyzowane jest zmiana przynależności Rosji oraz Chin, które trafiły do klastra razem z Kanadą i Australią. Interesująca jest też zmiana przynależności Polski, która z klastra 9 została przeniesiona do klastra 5.

Aby ułatwić interpretację wyników klasteryzacji poniżej dołączona została mapa 6.3 z naniesioną populacją oraz mapa 6.4 z naniesionym PKB na osobę poszczególnych krajów, a także ich odpowiedniki ze skalą logarytmiczną 6.5 oraz 6.6. Dane pochodzą z biblioteki GeoPandas [4].

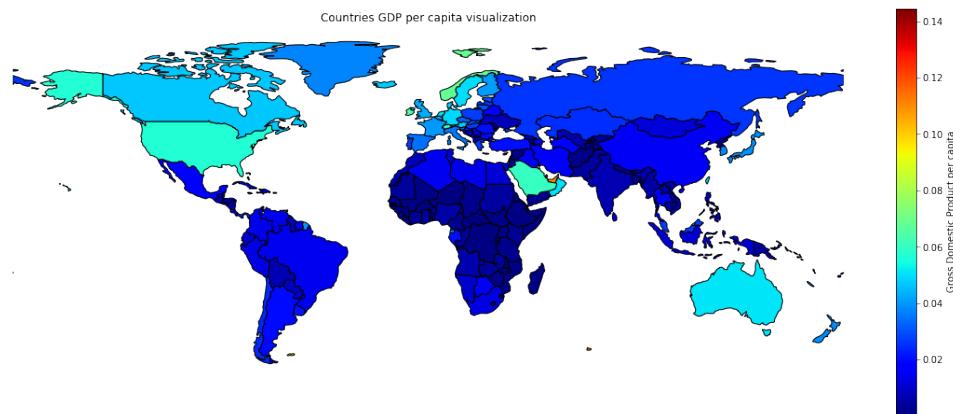
Na mapach z wykorzystaniem standardowej skali obserwujemy, że większość krajów trafiła do tego samego przedziału, co nie pozwala na dokładną interpretację danych. Na mapie populacji 6.3 wyróżniają się Chiny oraz Indie. Na mapie z PKB na osobę 6.4 wyróżniają się Zjednoczone Emiraty Arabskie, Stany Zjednoczone, Arabia Saudyjska, Norwegia.



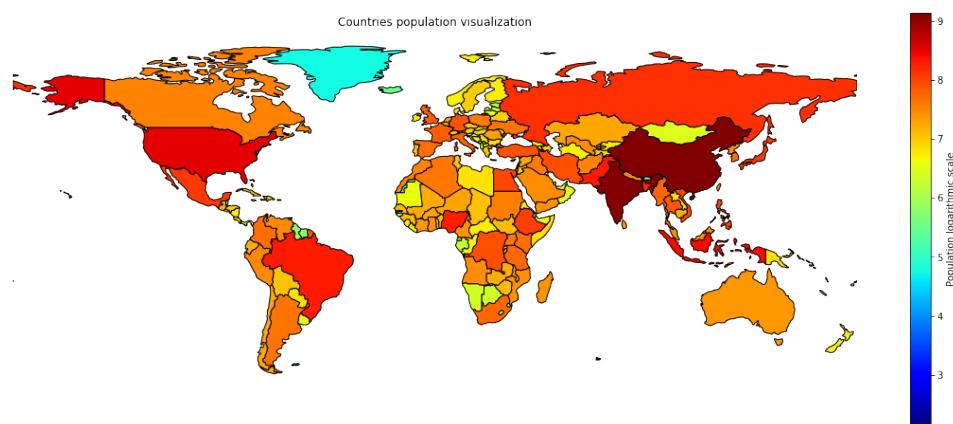
Rys. 6.2: Mapa z wynikami klasteryzacji. (źródło: opracowanie własne)



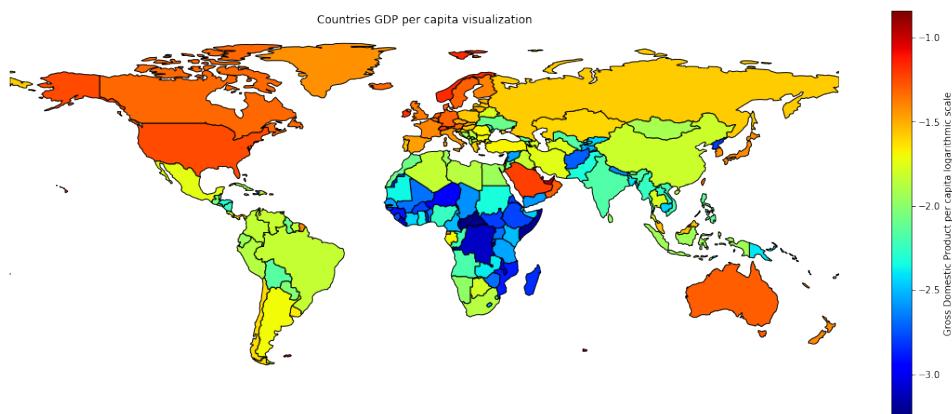
Rys. 6.3: Mapa z populacją krajów. (źródło: opracowanie własne)



Rys. 6.4: Mapa z PKB na osobę. (źródło: opracowanie własne)



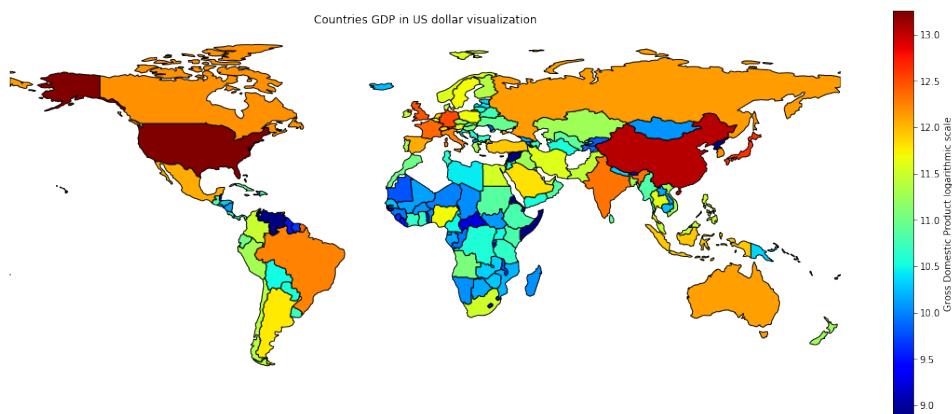
Rys. 6.5: Mapa z populacją krajów - skala logarytmiczna. (źródło: opracowanie własne)



Rys. 6.6: Mapa z PKB na osobę - skala logarytmiczna. (źródło: opracowanie własne)

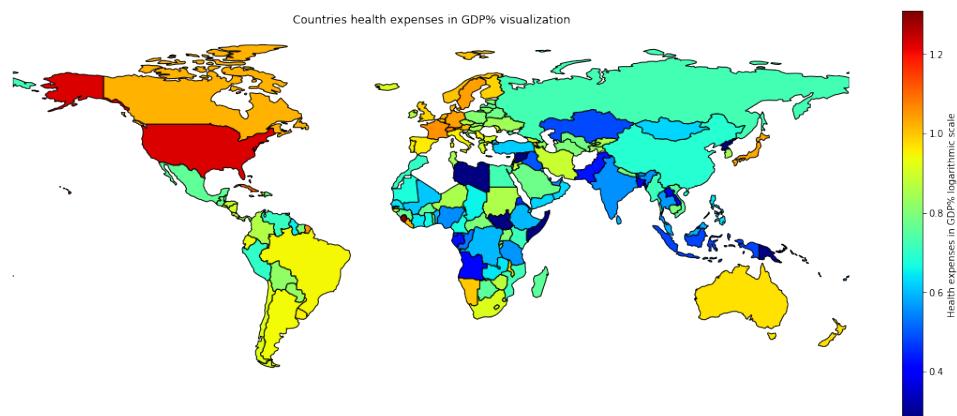
Mapa z PKB na osobę ze skalą logarytmiczną 6.6 pokazuje skupisko krajów dostatnich w Europie zachodniej.

Kolejne mapy prezentują PKB 6.7, wydatki na opiekę zdrowotną 6.8, zbrojenia 6.9, wartość importu 6.10 oraz eksportu 6.11. Dane pochodzą z bazy Banku Światowego [19].



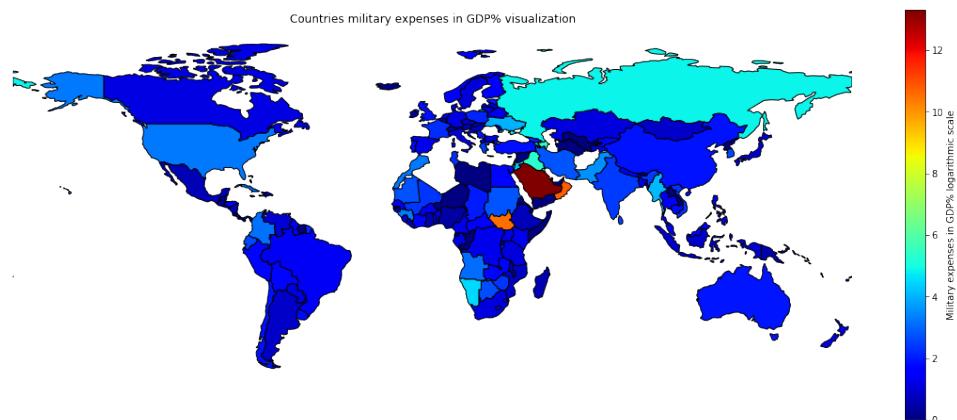
Rys. 6.7: Mapa z PKB - skala logarytmiczna. (źródło: opracowanie własne)

Mapa z PKB 6.7 różni się znaczco od mapy z PKB na osobę 6.6. Szczególnie wyraźnym przykładem są Chiny które przechodzą ze środka skali do jej górnego zakresu.



Rys. 6.8: Mapa z wydatkami na opiekę zdrowotną - skala logarytmiczna.
(źródło: opracowanie własne)

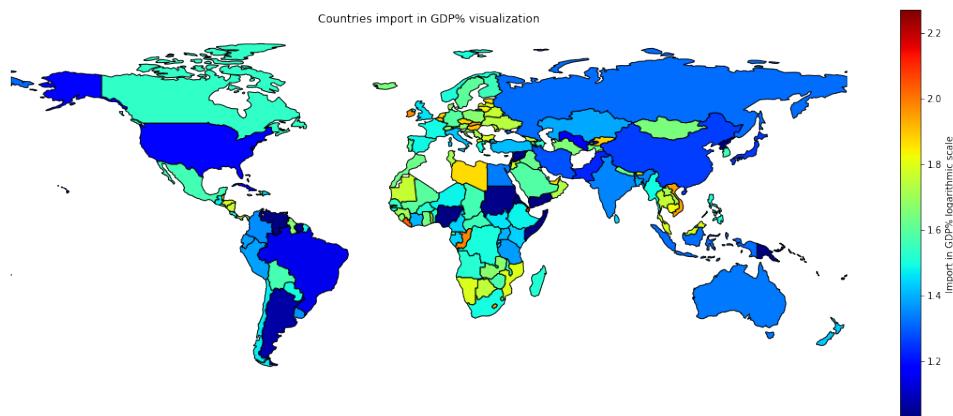
Na mapie z nakładami na opiekę zdrowotną 6.8 obserwujemy największe wydatki w krajach Ameryki Północnej oraz Europy Zachodniej, umiarkowane w Ameryce Południowej i Australii, najmniejsze w Afryce oraz Azji (wyjątek stanowi Japonia).



Rys. 6.9: Mapa z wydatkami na zbrojenia - skala logarytmiczna. (źródło:
opracowanie własne)

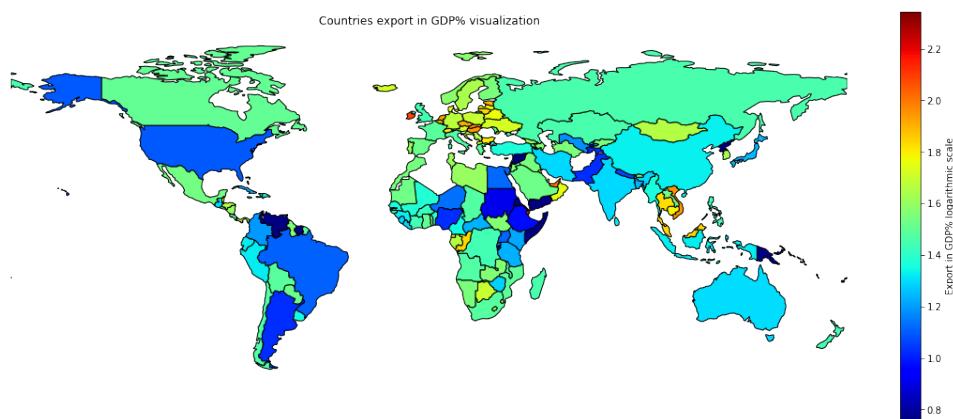
Na mapie z nakładami na zbrojenia 6.9 obserwujemy zbliżony poziom wydatków na całym świecie. Jaskrawymi odstępstwami są Arabia Saudyjska, Oman oraz Sudan Południowy których wydatki na zbrojenia są znacznie wyższe niż pozostałych krajów. Rosja, Azerbejdżan oraz Irak również

charakteryzują się zwiększymi nakładami na zbrojenia.



Rys. 6.10: Mapa z importem - skala logarytmiczna. (źródło: opracowanie własne)

Na mapie z importem 6.10 obserwujemy ponadprzeciętne sprowadzanie towarów w krajach Europy Wschodniej oraz w krajach południowej części Afryki.

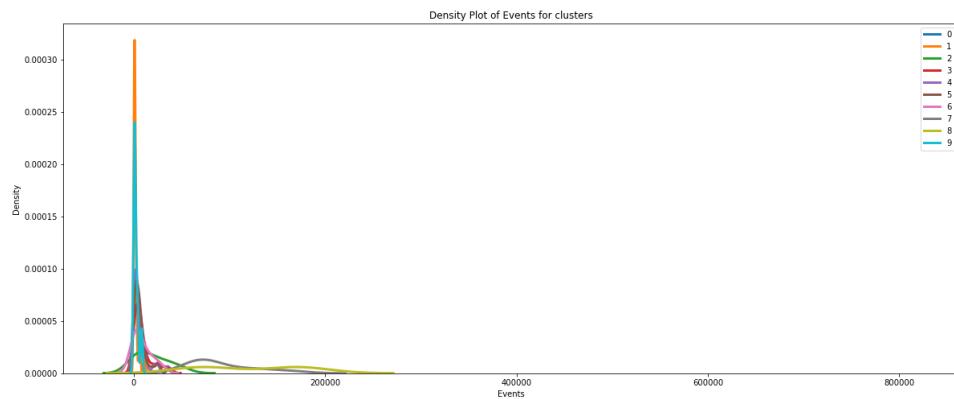


Rys. 6.11: Mapa z eksportem - skala logarytmiczna. (źródło: opracowanie własne)

Na mapie z eksportem 6.11 obserwujemy zwiększy wywóz towarów w Europie Wschodniej. Średni eksport towarów z Rosji jest najprawdopodobniej spowodowany sprzedażą dóbr naturalnych takich jak gaz i ropa naftowa.

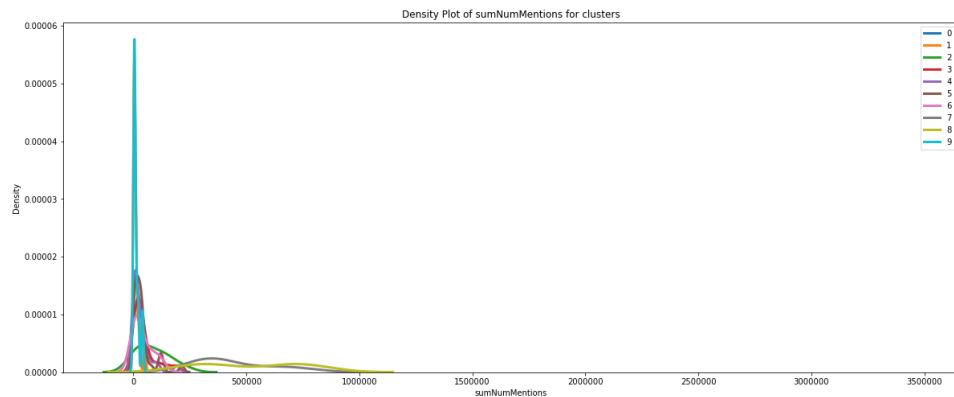
6.2.1 Wyniki klasteryzacji - wykresy gęstości miar użytych do klastrowania - dane ustANDARDYZOWANE

Na wykresach od 6.12 do 6.19 przedstawione zostały wykresy gęstości miar wykorzystanych przy grupowaniu w poszczególnych klastrach. Zastosowany wektor miar został opisany w części 3.1.



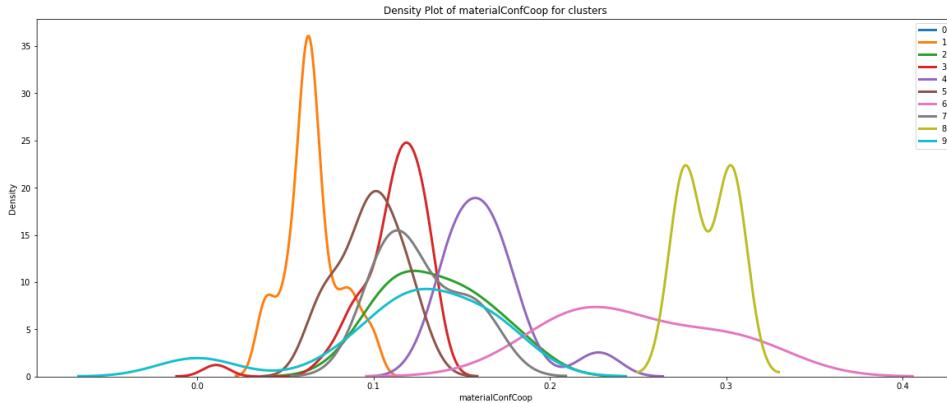
Rys. 6.12: Gęstość miary Events. (źródło: opracowanie własne)

Na wykresie 6.12 obserwujemy słabą separację klastrów dla miary Events. Wyróżniają się klastry 7 oraz 8 dla których gęstości są spłaszczone i wydłużone.



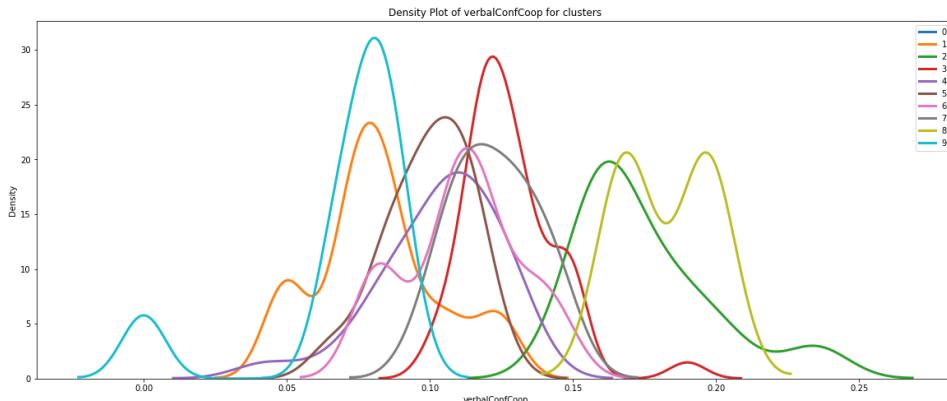
Rys. 6.13: Gęstość miary sumNumMentions. (źródło: opracowanie własne)

Na wykresie 6.13 dla miary sumNumMentions, podobnie jak na poprzednim, obserwujemy słabą separację klastrów. Ponownie wyróżniają się klastry 7 oraz 8 dla których gęstości są spłaszczone i wydłużone.



Rys. 6.14: Gęstość miary materialConfCoop. (źródło: opracowanie własne)

Na wykresie 6.14 obserwujemy dobrą separację klastrów 1, 3, 4, 5 oraz 8 z punktu widzenia miary materialConfCoop.



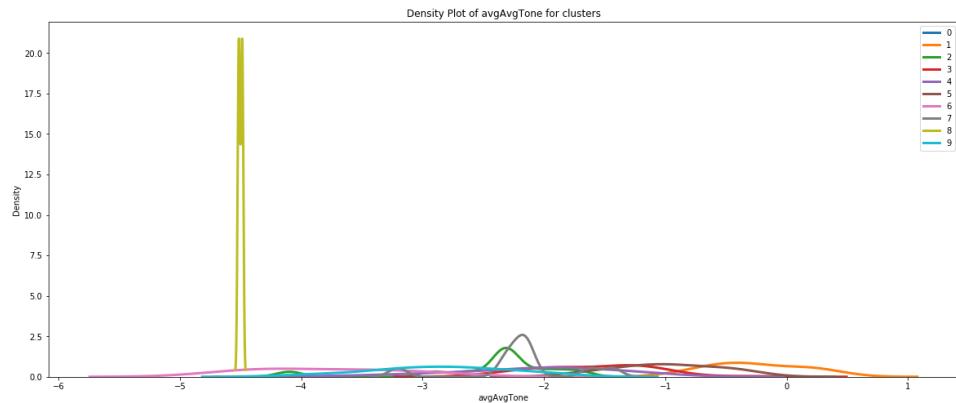
Rys. 6.15: Gęstość miary verbalConfCoop. (źródło: opracowanie własne)

Na wykresie 6.15 dla miary verbalConfCoop zauważamy podobieństwo pod względem gęstości klastrów 1 i 9 oraz 3 i 7. Ich wykresy wyraźnie się nakładają.

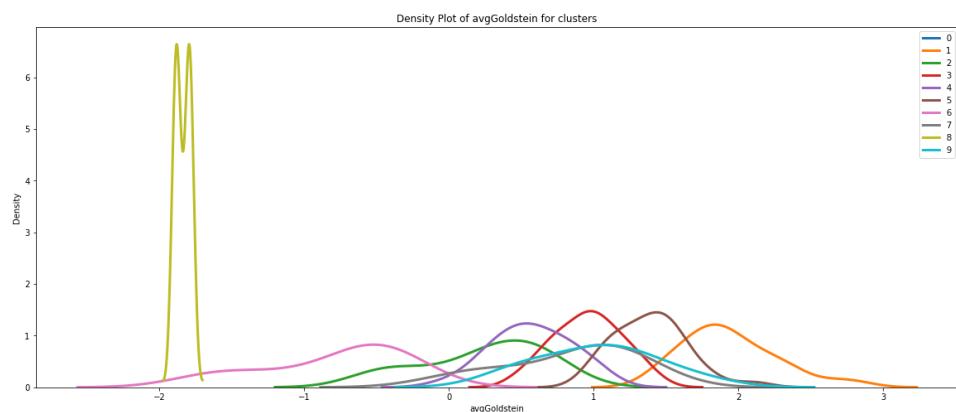
Na wykresie 6.16 dla miary avgAvgTone wyróżnia się klasa 8, który ma największą gęstość i duże skupienie, oraz klasy 7 i 8 które mają gęstości wyraźnie większe od pozostałych klastrów.

Na wykresie 6.17 dla miary avgGoldstein najbardziej wyróżnia się klasa 8. Pozostałe grupy (z wyjątkiem 2, 7 oraz 9) wyraźnie się oddzielają.

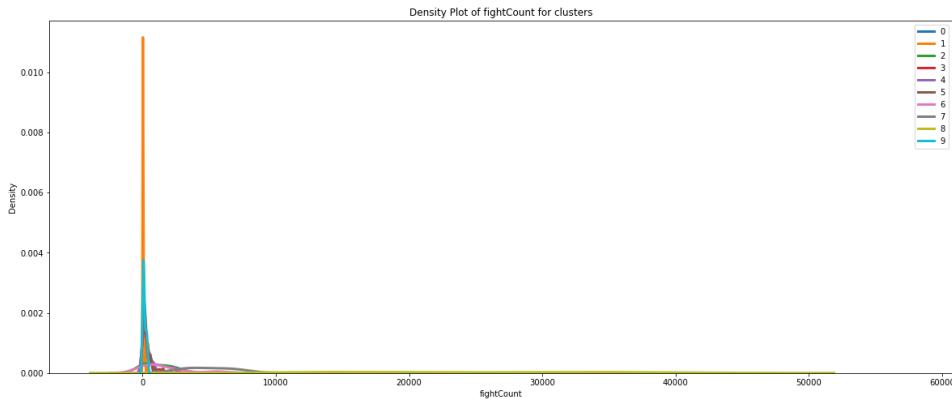
Na wykresie 6.18 dla miary fightCount większość klastrów osiąga mak-



Rys. 6.16: Gęstość miary avgAvgTone. (źródło: opracowanie własne)

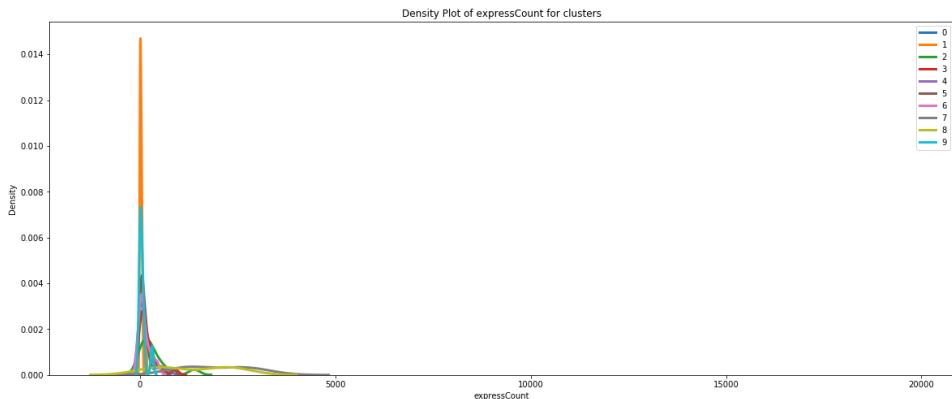


Rys. 6.17: Gęstość miary avgGoldstein. (źródło: opracowanie własne)



Rys. 6.18: Gęstość miary *fightCount*. (źródło: opracowanie własne)

sumum gęstości w pobliżu wartości 0. Gęstość klastra 8 jest bardzo mocno rozciągnięta w całym zakresie miary.



Rys. 6.19: Gęstość miary *expressCount*. (źródło: opracowanie własne)

Na wykresie 6.19 dla miary *expressCount*, podobnie jak na poprzednim wykresie, obserwujemy skupienie w pobliżu zera. Wyróżniają się klastry 7 i 8, które są spłaszczone i wydłużone ale znacznie mniej niż na poprzednim wykresie.

Z powyższej analizy wynika, że najlepszymi, pod względem separacji klastrów, miarami są te oparte o *quad class* czyli *materialConfCoop* oraz *verticalConfCoop*. Dobrze pod względem separacji sprawdza się też miara *avg-Goldstein*. Najgorzej wypadają miary oparte o ilość zdarzeń czyli *Events*, *fightCount* oraz *expressCount* i to one powinny być pierwszymi kandydata-

mi do zmiany wektora klastrowania.

6.2.2 Wyniki klasteryzacji w postaci tabelarycznej - dane ustandaryzowane

W tabelach od 6.1 do 6.19 przedstawione zostały wyniki grupowania ustandaryzowanych próbek.

Zgodnie z założeniami z części 3.1, dla ułatwienia interpretacji wyników grupowania, dodane zostały informacje zewnętrzne o PKB i wydatkach publicznych. Dodatkowo dokonano kolorowania pól z informacjami o krajach. Poszczególne miary zostały podzielone na 5 równych przedziałów według skali logarytmicznej. Kolory, od przedziału z najmniejszymi wartościami do tego z największymi, to: czerwony (bardzo niskie), pomarańczowy (niskie), żółty (średnie), jasnozielony (wysokie), zielony (bardzo wysokie). W tabelach od 6.4 do 6.20 została przedstawiona procentowa ilość państw w poszczególnych przedziałach w obrębie klastrów. W dodatku B dołączone zostały tabele ze szczegółowymi parametrami klastrów.

Klaster 0 w tabeli 6.1 zawiera 30 krajów. Przykłady ważniejszych państw w tym klastrze to: Estonia, Islandia, Luksemburg, Portugalia, Zjednoczone Emiraty Arabskie, Wietnam. W tabeli 6.2 obserwujemy, że w klastrze 0 przeważają (2/3) kraje z niskimi i średnimi wydatkami na opiekę zdrowotną (pomiędzy 3, a 7.9% PKB).

Klaster 1 w tabeli 6.3 zawiera 3 kraje - Indie, Liban oraz Palestynę. W tabeli 6.4 obserwujemy, że wszystkie państwa w klastrze 1 charakteryzują się średnim eksportem (pomiędzy 2.5%, a 5.1% PKB).

Klaster 2 w tabeli 6.5 zawiera 3 kraje - Botswanę, Gwineę Bissau oraz Tunezję. W tabeli 6.6 obserwujemy, że wszystkie państwa w klastrze 2 charakteryzują się niskim eksportem (pomiędzy 1.2%, a 2.5% PKB).

Klaster 3 w tabeli 6.7 zawiera 10 krajów. Przykłady ważniejszych państw w tym klastrze to: Chiny, Francja, Niemcy, Rosja, Wielka Brytania. W tabeli 6.8 obserwujemy, że w klastrze 3 przeważają kraje z wysokim oraz bardzo wysokim PKB (pomiędzy 3.23e+11\$, a 1.82e+13\$).

Klaster 4 w tabeli 6.9 zawiera 8 krajów. Przykłady ważniejszych państw w tym klastrze to: Kenia, Somalia, Syria. W tabeli 6.10 obserwujemy, że w klastrze 4 przeważają kraje z niskim PKB (pomiędzy 5.71e+09\$, a 4.29e+10\$).

Klaster 5 w tabeli 6.11 zawiera 21 krajów. Przykłady ważniejszych państw w tym klastrze to: Chile, Nigeria, Pakistan, Polska, Szwecja. W tabeli 6.12 obserwujemy, że w klastrze 5 przeważają kraje ze średnimi wydatkami na opiekę zdrowotną (pomiędzy 4.9%, a 7.9% PKB).

Klaster 6 w tabeli 6.13 zawiera 49 krajów. Przykłady ważniejszych państw

name	pop_est	gdp_pc	GDP	health	military	education	imp	exp
Belgium	1.1e+07	0.04	4.6e+11	10.28	0.92	6.54	76.39	77.81
Benin	1.1e+07	0.00	1.1e+10	3.99	1.10	4.36	32.04	24.72
Bhutan	7.6e+05	0.01	2e+09	3.66	-	7.32	67.76	34.44
Brunei	4.4e+05	0.08	1.3e+10	2.39	3.28	-	37.68	52.21
Costa Rica	4.9e+06	0.02	5.5e+10	7.82	0.00	7.08	31.69	30.82
Djibouti	8.7e+05	0.00	2.4e+09	4.39	-	5.34	122.64	141.46
Dominican Rep.	1.1e+07	0.02	7.1e+10	5.98	0.67	-	28.41	23.76
Eq. Guinea	7.8e+05	0.04	1.3e+10	2.92	-	-	42.22	56.66
Eritrea	5.9e+06	0.00	nan	2.86	-	-	-	-
Estonia	1.3e+06	0.03	2.3e+10	6.39	2.02	5.14	72.97	76.87
Fiji	9.2e+05	0.01	4.7e+09	3.57	0.97	-	-	-
Georgia	4.9e+06	0.01	1.5e+10	7.93	2.14	-	57.91	40.86
Iceland	3.4e+05	0.05	1.7e+10	8.15	-	7.53	44.67	51.80
Kazakhstan	1.9e+07	0.02	1.8e+11	3.05	1.11	2.79	24.53	28.52
Laos	7.1e+06	0.01	1.4e+10	2.45	-	-	51.85	33.95
Luxembourg	5.9e+05	0.10	5.8e+10	5.48	0.48	3.96	187.17	221.20
Macedonia	2.1e+06	0.01	1e+10	6.34	0.99	-	64.96	48.74
Madagascar	2.5e+07	0.00	1.1e+10	5.72	0.60	2.57	32.83	28.39
Mauritania	3.8e+06	0.00	6.2e+09	4.74	2.75	-	57.27	33.82
Morocco	3.4e+07	0.01	1e+11	5.07	3.23	-	42.40	34.80
Oman	3.4e+06	0.05	6.8e+10	4.30	10.79	-	53.77	57.14
Portugal	1.1e+07	0.03	2e+11	8.97	1.79	4.88	39.88	40.62
Qatar	2.3e+06	0.14	1.6e+11	3.12	-	-	36.65	57.06
Sierra Leone	6.2e+06	0.00	4.2e+09	20.41	0.92	-	47.44	19.36
Turkmenistan	5.4e+06	0.02	3.6e+10	6.30	-	-	45.62	35.68
United Arab Emirates	6.1e+06	0.11	3.6e+11	3.58	-	-	74.35	100.87
Uruguay	3.4e+06	0.02	5.3e+10	9.01	1.82	-	22.85	22.48
Uzbekistan	3e+07	0.01	8.2e+10	6.10	-	5.79	15.11	15.33
Vanuatu	2.8e+05	0.00	7.6e+08	4.16	-	5.55	-	-
Vietnam	9.6e+07	0.01	1.9e+11	5.65	2.36	-	88.99	89.78

Tab. 6.1: Klaster 0 - dane standaryzowane. (źródło: opracowanie własne)

	pop_est	gdp_pc	GDP	health	military	education	imp	exp
red	30.0%	6.7%	16.7%	16.7%	13.3%	0.0%	3.3%	0.0%
orange	40.0%	16.7%	36.7%	30.0%	20.0%	6.7%	20.0%	13.3%
yellow	26.7%	30.0%	36.7%	33.3%	20.0%	3.3%	40.0%	40.0%
lightgreen	3.3%	30.0%	6.7%	16.7%	6.7%	13.3%	20.0%	30.0%
green	0.0%	16.7%	0.0%	3.3%	3.3%	20.0%	6.7%	6.7%
brak	0.0%	0.0%	3.3%	0.0%	36.7%	56.7%	10.0%	10.0%

Tab. 6.2: Klaster 0 - ilość państw w poszczególnych przedziałach. (źródło: opracowanie własne)

name	pop_est	gdp_pc	GDP	health	military	education	imp	exp
India	1.3e+09	0.01	2.1e+12	3.60	2.41	-	22.11	19.81
Lebanon	6.2e+06	0.01	5e+10	7.68	4.53	-	48.69	23.14
Palestine	4.5e+06	0.00	1.3e+10	-	-	5.14	59.48	18.45

Tab. 6.3: Klaster 1 - dane standaryzowane. (źródło: opracowanie własne)

	pop_est	gdp_pc	GDP	health	military	education	exp
red	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
orange	66.7%	33.3%	33.3%	33.3%	0.0%	0.0%	100.0%
yellow	0.0%	66.7%	33.3%	33.3%	33.3%	0.0%	0.0%
lightgreen	0.0%	0.0%	33.3%	0.0%	33.3%	33.3%	0.0%
green	33.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
brak	0.0%	0.0%	0.0%	33.3%	33.3%	66.7%	0.0%

Tab. 6.4: Klaster 1 - ilość państw w poszczególnych przedziałach. (źródło: opracowanie własne)

name	pop_est	gdp_pc	GDP	health	military	education	imp	exp
Botswana	2.2e+06	0.02	1.4e+10	5.73	2.66	-	56.12	49.82
Guinea-Bissau	1.8e+06	0.00	1e+09	8.59	1.64	-	32.24	27.54
Tunisia	1.1e+07	0.01	4.3e+10	7.01	2.27	6.60	51.59	40.64

Tab. 6.5: Klaster 2 - dane standaryzowane. (źródło: opracowanie własne)

	pop_est	gdp_pc	GDP	health	military	education	imp
red	0.0%	33.3%	33.3%	0.0%	0.0%	0.0%	0.0%
orange	66.7%	0.0%	33.3%	0.0%	0.0%	0.0%	33.3%
yellow	33.3%	33.3%	33.3%	66.7%	100.0%	0.0%	66.7%
lightgreen	0.0%	33.3%	0.0%	33.3%	0.0%	0.0%	0.0%
green	0.0%	0.0%	0.0%	0.0%	0.0%	33.3%	0.0%
brak	0.0%	0.0%	0.0%	0.0%	0.0%	66.7%	0.0%

Tab. 6.6: Klaster 2 - ilość państw w poszczególnych przedziałach. (źródło: opracowanie własne)

name	pop_est	gdp_pc	GDP	health	military	education	imp	exp
Australia	2.3e+07	0.05	1.4e+12	9.31	1.96	5.31	21.51	19.98
Canada	3.6e+07	0.05	1.6e+12	10.51	1.15	-	34.31	31.85
China	1.4e+09	0.02	1.1e+13	4.89	1.91	-	18.11	21.35
France	6.7e+07	0.04	2.4e+12	11.46	2.27	-	31.16	30.59
Germany	8.1e+07	0.05	3.4e+12	11.09	1.18	4.81	39.29	46.85
Israel	8.3e+06	0.04	3e+11	7.09	5.65	5.85	28.18	31.21
Russia	1.4e+08	0.03	1.4e+12	5.30	4.86	3.83	20.66	28.70
Turkey	8.1e+07	0.02	8.6e+11	4.14	1.85	-	25.95	23.35
Ukraine	4.4e+07	0.01	9.1e+10	6.95	3.97	-	55.21	52.60
United Kingdom	6.5e+07	0.04	2.9e+12	9.69	1.86	5.61	29.03	27.65

Tab. 6.7: Klaster 3 - dane standaryzowane. (źródło: opracowanie własne)

	pop_est	gdp_pc	GDP	health	military	education	imp	exp
red	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	10.0%	0.0%
orange	10.0%	0.0%	0.0%	20.0%	20.0%	0.0%	60.0%	30.0%
yellow	30.0%	20.0%	20.0%	30.0%	50.0%	10.0%	30.0%	60.0%
lightgreen	50.0%	60.0%	40.0%	50.0%	30.0%	20.0%	0.0%	10.0%
green	10.0%	20.0%	40.0%	0.0%	0.0%	20.0%	0.0%	0.0%
brak	0.0%	0.0%	0.0%	0.0%	0.0%	50.0%	0.0%	0.0%

Tab. 6.8: Klaster 3 - ilość państw w poszczególnych przedziałach. (źródło: opracowanie własne)

name	pop_est	gdp_pc	GDP	health	military	education	imp	exp
Burkina Faso	2e+07	0.00	1.2e+10	5.83	1.33	4.17	32.98	26.11
Cameroon	2.5e+07	0.00	3.1e+10	4.70	1.25	2.77	27.61	22.26
Kenya	4.8e+07	0.00	6.4e+10	5.22	1.32	5.27	27.61	16.57
Lesotho	2e+06	0.00	2.4e+09	8.40	1.85	-	86.37	42.66
Niger	1.9e+07	0.00	9.7e+09	7.14	-	6.02	30.51	13.65
Somalia	7.5e+06	0.00	nan	-	-	-	-	-
Syria	1.8e+07	0.00	nan	-	-	-	-	-
Yemen	2.8e+07	0.00	3.7e+10	4.23	-	-	-	-

Tab. 6.9: Klaster 4 - dane standaryzowane. (źródło: opracowanie własne)

	pop_est	gdp_pc	GDP	health	military	education	imp	exp
red	0.0%	37.5%	12.5%	0.0%	0.0%	0.0%	0.0%	0.0%
orange	25.0%	62.5%	50.0%	25.0%	37.5%	12.5%	50.0%	37.5%
yellow	62.5%	0.0%	12.5%	37.5%	12.5%	0.0%	0.0%	25.0%
lightgreen	12.5%	0.0%	0.0%	12.5%	0.0%	25.0%	12.5%	0.0%
green	0.0%	0.0%	0.0%	0.0%	0.0%	12.5%	0.0%	0.0%
brak	0.0%	0.0%	25.0%	25.0%	50.0%	50.0%	37.5%	37.5%

Tab. 6.10: Klaster 4 - ilość państw w poszczególnych przedziałach. (źródło: opracowanie własne)

name	pop_est	gdp_pc	GDP	health	military	education	imp	exp
Angola	2.9e+07	0.01	1.2e+11	2.61	3.11	-	33.13	29.75
Bolivia	1.1e+07	0.01	3.3e+10	6.62	1.74	-	37.06	30.87
Burundi	1.1e+07	0.00	3.1e+09	6.58	2.20	6.37	26.81	5.73
Chile	1.8e+07	0.02	2.4e+11	8.30	1.90	4.88	29.60	29.38
Colombia	4.8e+07	0.01	2.9e+11	7.30	3.13	4.47	22.71	15.65
Gambia	2.1e+06	0.00	1.4e+09	3.10	1.53	2.14	32.65	16.27
Guinea	1.2e+07	0.00	8.8e+09	5.81	3.31	2.52	50.95	21.50
Haiti	1.1e+07	0.00	8.7e+09	8.63	0.00	3.16	50.54	20.01
Jordan	1e+07	0.01	3.8e+10	7.59	4.30	-	59.65	37.07
Malawi	1.9e+07	0.00	6.4e+09	9.33	0.63	5.61	35.75	29.16
Mozambique	2.7e+07	0.00	1.6e+10	5.27	0.81	-	62.83	31.08
Nicaragua	6e+06	0.01	1.3e+10	8.00	0.78	4.08	58.11	40.08
Nigeria	1.9e+08	0.01	4.9e+11	3.58	0.42	-	10.67	10.67
Pakistan	2e+08	0.00	2.7e+11	2.69	3.55	2.65	17.05	10.60
Poland	3.8e+07	0.03	4.8e+11	6.40	2.14	4.81	46.40	49.50
Saudi Arabia	2.9e+07	0.06	6.5e+11	6.00	13.33	-	38.75	33.32
Sweden	1e+07	0.05	5.1e+11	11.00	1.08	7.55	39.96	43.77
Uganda	4e+07	0.00	3.2e+10	6.48	1.21	2.77	25.01	12.98
Venezuela	3.1e+07	0.01	nan	5.13	0.94	-	-	-
Zimbabwe	1.4e+07	0.00	2e+10	7.45	2.34	-	37.59	19.16
eSwatini	1.5e+06	0.01	4.1e+09	7.13	1.69	-	40.84	42.96

Tab. 6.11: Klaster 5 - dane standaryzowane. (źródło: opracowanie własne)

	pop_est	gdp_pc	GDP	health	military	education	imp	exp
red	4.8%	23.8%	14.3%	9.5%	14.3%	0.0%	9.5%	14.3%
orange	9.5%	19.0%	42.9%	9.5%	23.8%	19.0%	28.6%	28.6%
yellow	71.4%	38.1%	19.0%	57.1%	33.3%	4.8%	47.6%	52.4%
lightgreen	14.3%	9.5%	19.0%	23.8%	23.8%	19.0%	9.5%	0.0%
green	0.0%	9.5%	0.0%	0.0%	4.8%	14.3%	0.0%	0.0%
brak	0.0%	0.0%	4.8%	0.0%	0.0%	42.9%	4.8%	4.8%

Tab. 6.12: Klaster 5 - ilość państw w poszczególnych przedziałach. (źródło: opracowanie własne)

name	pop_est	gdp_pc	GDP	health	military	education	imp	exp
Argentina	4.4e+07	0.02	5.9e+11	8.79	0.85	5.78	11.78	10.71
Armenia	3e+06	0.01	1.1e+10	10.12	4.24	2.81	41.87	29.73
Austria	8.8e+06	0.05	3.8e+11	10.37	0.70	5.46	49.34	53.09
Azerbaijan	1e+07	0.02	5.3e+10	6.71	5.47	2.95	34.81	37.79
Belarus	9.5e+06	0.02	5.6e+10	6.07	1.28	4.79	57.90	58.01
Brazil	2.1e+08	0.01	1.8e+12	8.87	1.37	6.24	14.05	12.90
Bulgaria	7.1e+06	0.02	5.1e+10	8.20	1.32	-	63.03	63.99
Cambodia	1.6e+07	0.00	1.8e+10	6.19	1.79	-	66.15	61.72
Croatia	4.3e+06	0.02	5e+10	6.79	1.52	-	46.16	46.38
Cuba	1.1e+07	0.01	8.7e+10	12.81	3.08	-	14.47	17.15
Cyprus	1.2e+06	0.02	2e+10	6.81	1.67	6.38	67.70	70.33
Czechia	1.1e+07	0.03	1.9e+11	7.24	0.95	5.79	75.05	81.05
Côte d'Ivoire	2.4e+07	0.00	4.6e+10	4.42	1.74	4.83	25.36	27.36
Denmark	5.6e+06	0.05	3e+11	10.23	1.11	-	48.63	55.42
Ecuador	1.6e+07	0.01	9.9e+10	8.59	2.62	5.00	23.99	21.26
Egypt	9.7e+07	0.01	3.3e+11	5.34	1.72	-	21.66	13.18
Ethiopia	1.1e+08	0.00	6.5e+10	3.94	0.71	4.74	30.29	9.36
Finland	5.5e+06	0.04	2.3e+11	9.70	1.46	7.08	35.98	35.41
Gabon	1.8e+06	0.02	1.4e+10	2.66	1.19	-	27.92	46.03
Ghana	2.7e+07	0.00	4.9e+10	4.62	0.53	4.51	43.63	31.96
Greece	1.1e+07	0.03	2e+11	8.09	2.45	-	31.50	31.55
Hungary	9.9e+06	0.03	1.2e+11	6.97	0.92	4.57	79.99	87.98
Indonesia	2.6e+08	0.01	8.6e+11	3.01	0.89	3.58	20.78	21.16
Ireland	5e+06	0.06	2.9e+11	7.34	0.34	3.76	93.17	121.97
Italy	6.2e+07	0.04	1.8e+12	8.99	1.21	4.08	26.70	29.72
Jamaica	3e+06	0.01	1.4e+10	5.66	0.87	5.46	46.20	29.92
Japan	1.3e+08	0.04	4.4e+12	10.89	0.96	-	18.03	17.61
Kyrgyzstan	5.8e+06	0.00	6.7e+09	7.15	1.75	5.99	75.77	35.19
Latvia	1.9e+06	0.03	2.7e+10	5.71	1.05	5.34	61.63	60.72
Liberia	4.7e+06	0.00	3.2e+09	10.29	0.73	-	106.55	19.47
Libya	6.7e+06	0.01	2.8e+10	-	-	-	74.28	39.97
Lithuania	2.8e+06	0.03	4.1e+10	6.48	1.14	4.22	69.82	68.82
Malaysia	3.1e+07	0.03	3e+11	3.89	1.53	4.97	61.92	69.45
Moldova	3.5e+06	0.01	7.7e+09	8.56	0.35	-	57.44	31.90
Myanmar	5.5e+07	0.01	6.8e+10	5.20	4.08	-	30.94	22.98
New Zealand	4.5e+06	0.04	1.8e+11	9.33	1.15	6.34	26.88	27.92
Norway	5.3e+06	0.07	3.9e+11	10.11	1.50	7.55	32.06	37.80
Panama	3.8e+06	0.02	5.4e+10	6.81	0.00	-	52.25	47.69
Senegal	1.5e+07	0.00	1.8e+10	4.38	1.57	5.46	35.43	22.68
Serbia	7.1e+06	0.01	4e+10	8.82	1.95	3.79	52.25	45.27
South Korea	5.1e+07	0.04	1.5e+12	7.05	2.64	-	36.14	42.99
Spain	4.9e+07	0.03	1.2e+12	9.11	1.27	4.27	30.59	33.63
Sri Lanka	2.2e+07	0.01	8.1e+10	3.89	2.55	2.23	28.55	21.01
Sudan	3.7e+07	0.00	7.4e+10	7.18	2.80	-	10.92	8.18
Suriname	5.9e+05	0.01	4.8e+09	6.22	-	-	-	-
Switzerland	8.2e+06	0.06	6.8e+11	11.88	0.67	5.10	50.51	62.22
Tanzania	5.4e+07	0.00	4.7e+10	3.65	1.13	-	23.65	17.10
Togo	8e+06	0.00	4.2e+09	6.25	1.68	5.11	57.81	35.85
Zambia	1.6e+07	0.00	2.1e+10	4.44	1.75	4.62	47.18	37.14

Tab. 6.13: Klaster 6 - dane standaryzowane. (źródło: opracowanie własne)

	pop_est	gdp_pc	GDP	health	military	education	imp	exp
red	4.1%	6.1%	6.1%	4.1%	14.3%	0.0%	10.2%	6.1%
orange	40.8%	16.3%	26.5%	16.3%	38.8%	4.1%	28.6%	22.4%
yellow	34.7%	26.5%	44.9%	38.8%	34.7%	8.2%	34.7%	42.9%
lightgreen	18.4%	44.9%	20.4%	36.7%	8.2%	34.7%	22.4%	24.5%
green	2.0%	6.1%	2.0%	2.0%	0.0%	16.3%	2.0%	2.0%
brak	0.0%	0.0%	0.0%	2.0%	4.1%	36.7%	2.0%	2.0%

Tab. 6.14: Klaster 6 - ilość państw w poszczególnych przedziałach.
 (źródło: opracowanie własne)

w tym klastrze to: Białoruś, Bułgaria, Czechy, Finlandia, Węgry, Norwegia. W tabeli 6.14 obserwujemy, że w klastrze 6 przeważają kraje z wysokimi wydatkami na edukację (pomiędzy 4%, a 5.5% PKB).

name	pop_est	gdp_pc	GDP	health	military	education	imp	exp
Iran	8.2e+07	0.02	3.8e+11	7.76	2.76	2.80	19.27	19.74
Iraq	3.9e+07	0.02	1.8e+11	3.14	5.35	-	41.42	35.34

Tab. 6.15: Klaster 7 - dane standaryzowane. (źródło: opracowanie własne)

	pop_est	health	education
red	0.0%	0.0%	0.0%
orange	0.0%	50.0%	50.0%
yellow	50.0%	50.0%	0.0%
lightgreen	50.0%	0.0%	0.0%
green	0.0%	0.0%	0.0%
brak	0.0%	0.0%	50.0%

Tab. 6.16: Klaster 7 - ilość państw w poszczególnych przedziałach.
 (źródło: opracowanie własne)

Klaster 7 w tabeli 6.15 zawiera 2 kraje - Iran i Irak. W tabeli 6.16 obserwujemy, że w klastrze 7 żaden z parametrów nie wyróżnia się.

name	pop_est	gdp_pc	GDP	health	military	education	imp	exp
United States of America	3.3e+08	0.06	1.8e+13	16.84	3.27	-	15.30	12.44

Tab. 6.17: Klaster 8 - dane standaryzowane. (źródło: opracowanie własne)

Klaster 8 w tabeli 6.17 posiada tylko jedno państwo - Stany Zjednoczone Ameryki. W tabeli 6.18 obserwujemy, że klaster 8 wyróżnia się bardzo

	pop_est	military	education	imp	exp
red	0.0%	0.0%	0.0%	100.0%	0.0%
orange	0.0%	0.0%	0.0%	0.0%	100.0%
yellow	0.0%	0.0%	0.0%	0.0%	0.0%
lightgreen	0.0%	100.0%	0.0%	0.0%	0.0%
green	100.0%	0.0%	0.0%	0.0%	0.0%
brak	0.0%	0.0%	100.0%	0.0%	0.0%

Tab. 6.18: Klaster 8 - ilość państw w poszczególnych przedziałach.
(źródło: opracowanie własne)

wysokim PKB ($2.42e+12\$$, a $1.82e+13\$$).

name	pop_est	gdp_pc	GDP	health	military	education	imp	exp
Bahamas	3.3e+05	0.03	1.2e+10	5.71	-	-	37.90	35.10
Bangladesh	1.6e+08	0.00	2e+11	2.46	1.35	-	24.75	17.34
Belize	3.6e+05	0.01	1.7e+09	5.90	1.13	6.77	66.26	58.82
Central African Rep.	5.6e+06	0.00	1.7e+09	6.31	1.69	-	36.08	17.07
Chad	1.2e+07	0.00	1.1e+10	4.52	2.02	-	36.56	30.00
Congo	5e+06	0.01	8.6e+09	3.42	-	4.56	96.37	69.28
Dem. Rep. Congo	8.3e+07	0.00	3.8e+10	3.94	1.28	2.17	31.61	27.73
El Salvador	6.2e+06	0.01	2.3e+10	7.63	1.07	3.91	47.07	29.49
Guatemala	1.5e+07	0.01	6.2e+10	5.96	0.40	2.96	30.06	19.83
Guyana	7.4e+05	0.01	3.2e+09	4.52	1.46	5.21	46.65	36.01
Honduras	9e+06	0.00	2.1e+10	7.71	1.68	6.41	62.10	45.17
Kuwait	2.9e+06	0.10	1.1e+11	4.01	5.01	-	44.93	53.77
Mali	1.8e+07	0.00	1.3e+10	4.11	2.36	3.80	39.60	24.04
Mexico	1.2e+08	0.02	1.2e+12	5.80	0.66	5.24	36.60	34.56
Mongolia	3.1e+06	0.01	1.2e+10	4.24	0.87	4.18	44.64	45.65
Namibia	2.5e+06	0.01	1.1e+10	9.94	4.48	-	62.00	35.56
Nepal	2.9e+07	0.00	2.1e+10	6.22	1.56	3.70	41.47	11.62
Netherlands	1.7e+07	0.05	7.7e+11	10.32	1.13	5.35	75.16	82.66
North Korea	2.5e+07	0.00	nan	-	-	-	-	-
Papua New Guinea	6.9e+06	0.00	2.2e+10	1.91	0.47	-	-	-
Paraguay	6.9e+06	0.01	3.6e+10	6.75	1.07	-	31.51	33.00
Peru	3.1e+07	0.01	1.9e+11	5.03	1.72	3.97	23.88	21.29
Philippines	1e+08	0.01	3.1e+11	4.32	1.14	-	31.93	27.21
Rwanda	1.2e+07	0.00	8.6e+09	6.52	1.25	3.75	25.95	15.91
S. Sudan	1.3e+07	0.00	1.2e+10	-	1.056	1.54	28.90	36.65
Slovakia	5.4e+06	0.03	8.8e+10	6.85	1.12	4.64	88.95	92.02
Solomon Is.	6.5e+05	0.00	1.2e+09	5.25	-	-	53.32	45.07
South Africa	5.5e+07	0.01	3.2e+11	8.20	1.10	5.96	31.46	30.15
Tajikistan	8.5e+06	0.00	7.9e+09	6.91	1.22	5.23	42.24	10.49
Thailand	6.8e+07	0.02	4e+11	3.67	1.43	-	57.20	67.64
Trinidad and Tobago	1.2e+06	0.04	2.5e+10	6.04	0.86	-	-	-

Tab. 6.19: Klaster 9 - dane standaryzowane. (źródło: opracowanie własne)

	pop_est	gdp_pc	GDP	health	military	education	imp	exp
red	16.1%	12.9%	12.9%	6.5%	9.7%	3.2%	0.0%	6.5%
orange	29.0%	29.0%	51.6%	29.0%	48.4%	3.2%	29.0%	19.4%
yellow	35.5%	35.5%	22.6%	48.4%	19.4%	19.4%	41.9%	45.2%
lightgreen	19.4%	16.1%	9.7%	9.7%	6.5%	22.6%	19.4%	19.4%
green	0.0%	6.5%	0.0%	0.0%	3.2%	9.7%	0.0%	0.0%
brak	0.0%	0.0%	3.2%	6.5%	12.9%	41.9%	9.7%	9.7%

Tab. 6.20: Klaster 9 - ilość państw w poszczególnych przedziałach.
(źródło: opracowanie własne)

Klaster 9 w tabeli 6.19 zawiera 31 krajów. Przykłady ważniejszych państw w tym klastrze to: Gwatemala, Honduras, Kuwejt, Meksyk, Słowacja. W tabeli 6.20 obserwujemy, że klaster 9 wyróżnia się niskimi wydatkami na zbrojenia (pomiędzy 0.7%, a 1.48% PKB).

Tabele 6.21 oraz 6.22 zawierają podsumowanie parametrów poszczególnych klastrów.

	pop_est	gdp_pc	GDP	health	military	education	imp	exp
cluster 0	1e+07	0.028	7.7e+10	5.827	1.998	5.296	55.557	54.783
cluster 1	4.3e+08	0.008	7.2e+11	5.638	3.466	5.141	43.426	20.466
cluster 2	5.1e+06	0.010	2e+10	7.108	2.190	6.599	46.650	39.334
cluster 3	1.9e+08	0.034	2.5e+12	8.041	2.665	5.083	30.341	31.414
cluster 4	2.1e+07	0.002	2.6e+10	5.920	1.436	4.556	41.016	24.251
cluster 5	3.6e+07	0.012	1.6e+11	6.428	2.388	4.250	37.804	26.475
cluster 6	3e+07	0.021	3.4e+11	7.204	1.580	4.928	44.556	39.806
cluster 7	6.1e+07	0.017	2.8e+11	5.452	4.053	2.797	30.346	27.542
cluster 8	3.3e+08	0.057	1.8e+13	16.840	3.272	-	15.295	12.442
cluster 9	2.7e+07	0.014	1.3e+11	5.661	1.855	4.408	45.542	37.611

Tab. 6.21: Średnie wartości parametrów w klastrach. (źródło: opracowanie własne)

W tabeli 6.21, zawierającej średnie wartości parametrów w klastrach, wyraźnie oddziela się klaster 8 - hegemon, Stany Zjednoczone Ameryki - bardzo wysokie PKB, wydatki na zdrowie, wysokie wydatki na zbrojenia, niski import. Kolejnym wyróżniają się klastrem jest klaster 7 - Irak i Iran - najwyższe wydatki na zbrojenia. Klaster 0 to kraje bogate ale mało ludne. Klaster 1 charakteryzuje niewielki eksport i niskie PKB. Klaster 2 charakteryzuje duże wydatki na edukację oraz niskie PKB na osobę. Klaster 3 to kraje bogate o wysokim PKB i wydatkach na opiekę zdrowotną oraz o niskim imporcie. Klaster 4 to ubogie kraje afrykańskie i bliskiego wschodu o niskim PKB. Klaster 5 to kraje przeciętne.

	pop_est	gdp_pc	GDP	health	military	education	imp	exp
cluster 0	4.9e+06	0.015	2.3e+10	5.275	1.110	5.341	45.624	40.615
cluster 1	6.2e+06	0.007	5e+10	5.638	3.466	5.141	48.690	19.813
cluster 2	2.2e+06	0.011	1.4e+10	7.007	2.270	6.599	51.591	40.642
cluster 3	6.6e+07	0.038	1.5e+12	8.201	1.933	5.312	28.605	29.648
cluster 4	2e+07	0.003	2.1e+10	5.522	1.325	4.719	30.514	22.260
cluster 5	1.8e+07	0.006	3.3e+10	6.579	1.744	4.274	37.327	29.266
cluster 6	9.9e+06	0.017	6.5e+10	7.008	1.316	4.973	42.746	35.298
cluster 7	6.1e+07	0.017	2.8e+11	5.452	4.053	2.797	30.346	27.542
cluster 8	3.3e+08	0.057	1.8e+13	16.840	3.272	-	15.295	12.442
cluster 9	9e+06	0.008	2.2e+10	5.797	1.254	4.369	40.535	33.782

Tab. 6.22: Mediany wartości parametrów w klastrach. (źródło: opracowanie własne)

W tabeli 6.22, zawierającej mediany wartości parametrów w klastrach, obserwujemy przesunięcie poszczególnych miar w kierunku mniejszych wartości. Szczególnie zwraca uwagę klaster 1 w którym populacja spada z wartością bardzo wysoką (w średniej) do wartości niskiej (w medianie).

Zarówno dla danych nie poddanych oraz poddanych standaryzacji Stany Zjednoczone otrzymały własny klaster. Jest to spowodowane znaczącą przewagą liczby zdarzeń w porównaniu z innymi krajami.

Dla danych standaryzowanych Polska i Szwecja są jedynymi krajami europejskim w klastrze 5 (w tabeli 6.11).

Rozdział 7

Podsumowanie

W niniejszym rozdziale w części 7.1 zostaną opisane wnioski z pracy. W części 7.2 zawarto propozycje dalszych kierunków rozwoju.

7.1 Wnioski

Przeprowadzona analiza danych ze zbioru GDELT pozwoliła na wybranie takiego wektora cech który umożliwił skuteczny podział państw - aktorów - na grupy. Kraje podobne do siebie trafiły do właściwych klastrów w oparciu o cechy z notatek prasowych. Udało się wykazać związek powstałych klastrów z rzeczywistymi cechami państw. W klastrach pojawiają się anomalie. Wyraźnym przykładem takiej anomalii jest klaster 5 do którego jako jedynie z państw europejskich trafiły Polska i Szwecja. Potrzebna jest modyfikacja wektora miar służącego do klasteryzacji. Propozycja zmiany wektora została przedstawiona w części 7.2.

7.2 Dalsze kierunki rozwoju

Poniżej przedstawiono propozycje dalszych kierunków rozwoju pracy.

Badanie innych wektorów grupowania Próba znalezienia cech zdarzeń lepiej odpowiadających rzeczywistym podziałom krajów na grupy. Dalszej analizy wymaga wektor miar wykorzystany do klastrowania. Pierwszymi kandydatami do zmiany są miary, które cechują się słabą separacją klastrów na wykresach gęstości w części 6.2.1, a więc *Events*, *fightCount* oraz *expressionsCount*.

Analiza dynamiczna grupowania krajów Przeprowadzenie klasteryzacji w kolejnych przedziałach czasowych. Analiza zmian przynależności krajów do klastrów.

Analiza pod kątem COVID-19 Analiza danych z bazy GDELT pod kątem zmian w relacjach krajów spowodowanych rozwojem pandemii koronawirusa. Sprawdzenie czy wzorce zmian zachowań obserwowanych podczas pandemii występowały już wcześniej.

Analiza pod kątem grup etnicznych Analiza zdarzeń z podziałem na grupy etniczne. Klastrowanie grup etnicznych.

Analiza pod kątem grup religijnych Analiza zdarzeń z podziałem na grupy religijne. Klastrowanie grup religijnych.

Dodatek A

Wyniki grupowania danych niestandardyzowanych

W tym dodatku przedstawione zostały wyniki klasteryzacji danych nie poddanych standaryzacji.

W tabelach od A.1 do A.10 przedstawione zostały wyniki grupowania. Dodatkowo zawarto informacje o liczbie ludności i PKB na osobę pochodzące z biblioteki GeoPandas [4].

Klaster 0 w tabeli A.1 zawiera 4 kraje.

iso_a3	name	continent	pop_est	gdp_per_cap
CHN	China	Asia	1,379,302,771	0.015327
IRN	Iran	Asia	82,021,564	0.017788
RUS	Russia	Europe	142,257,519	0.026325
GBR	United Kingdom	Europe	64,769,452	0.043045

Tab. A.1: Klaster 0. (źródło: opracowanie własne)

Klaster 1 w tabeli A.2 zawiera 15 krajów.

Klaster 2 w tabeli A.3 zawiera 34 kraje.

Klaster 3 w tabeli A.4 zawiera 8 krajów.

Klaster 4 w tabeli A.5 zawiera 31 krajów.

Klaster 5 w tabeli A.6 zawiera 13 krajów.

Klaster 6 w tabeli A.7 zawiera 35 krajów.

Klaster 7 w tabeli A.8 zawiera 9 krajów.

Klaster 8 w tabeli A.9 zawiera tylko jeden kraj - Stany Zjednoczone Ameryki.

Klaster 9 w tabeli A.10 zawiera 12 krajów.

iso_a3	name	continent	pop_est	gdp_per_cap
ALB	Albania	Europe	3,047,987	0.011122
AGO	Angola	Africa	29,310,273	0.006448
ECU	Ecuador	South America	16,290,913	0.011196
HND	Honduras	North America	9,038,741	0.004778
HUN	Hungary	Europe	9,850,845	0.027165
JAM	Jamaica	North America	2,990,561	0.008490
KAZ	Kazakhstan	Asia	18,556,698	0.024827
KWT	Kuwait	Asia	2,875,422	0.104715
LVA	Latvia	Europe	1,944,643	0.026046
LTU	Lithuania	Europe	2,823,859	0.030320
MKD	Macedonia	Europe	2,103,721	0.014032
QAT	Qatar	Asia	2,314,307	0.144536
SRB	Serbia	Europe	7,111,024	0.014316
SOM	Somalia	Africa	7,531,386	0.000627
UGA	Uganda	Africa	39,570,125	0.002146

Tab. A.2: Klaster 1. (źródło: opracowanie własne)

Dla danych niestandardyzowanych Polska trafiła do klastra A.10 między innymi z Egiptem, Grecją, Irlandią oraz Brazylią. Wyróżnia się mniejsza grupa A.1 do której trafiły Chiny, Rosja, Wielka Brytania oraz Iran.

iso_a3	name	continent	pop_est	gdp_per_cap
BLZ	Belize	North America	360,346	0.008570
BEN	Benin	Africa	11,038,805	0.002202
BTN	Bhutan	Asia	758,288	0.008482
BWA	Botswana	Africa	2,214,858	0.016209
BRN	Brunei	Asia	443,593	0.076038
BDI	Burundi	Africa	11,466,756	0.000688
CAF	Central African Rep.	Africa	5,625,118	0.000570
COG	Congo	Africa	4,954,674	0.006109
CRI	Costa Rica	North America	4,930,258	0.016076
DJI	Djibouti	Africa	865,267	0.003866
GNQ	Eq. Guinea	Africa	778,358	0.040817
ERI	Eritrea	Africa	5,918,919	0.001549
FJI	Fiji	Oceania	920,938	0.009093
GAB	Gabon	Africa	1,772,255	0.020302
GEO	Georgia	Asia	4,926,330	0.007565
GNB	Guinea-Bissau	Africa	1,792,338	0.001591
ISL	Iceland	Europe	339,747	0.047535
LAO	Laos	Asia	7,126,706	0.005747
LSO	Lesotho	Africa	1,958,042	0.003074
MDG	Madagascar	Africa	25,054,161	0.001471
MRT	Mauritania	Africa	3,758,571	0.004446
MNG	Mongolia	Asia	3,068,243	0.012059
NAM	Namibia	Africa	2,484,780	0.010460
PNG	Papua New Guinea	Oceania	6,909,701	0.004055
PRY	Paraguay	South America	6,943,739	0.009313
SLE	Sierra Leone	Africa	6,163,195	0.001726
SLB	Solomon Is.	Oceania	647,581	0.001850
SUR	Suriname	South America	591,919	0.014439
TJK	Tajikistan	Asia	8,468,555	0.003048
TGO	Togo	Africa	7,965,055	0.001458
TTO	Trinidad and Tobago	North America	1,218,208	0.035766
TKM	Turkmenistan	Asia	5,351,277	0.017700
VUT	Vanuatu	Oceania	282,814	0.002556
SWZ	eSwatini	Africa	1,467,152	0.007538

Tab. A.3: Klaster 2. (źródło: opracowanie własne)

iso_a3	name	continent	pop_est	gdp_per_cap
AUS	Australia	Oceania	23,232,413	0.051178
CAN	Canada	North America	35,623,680	0.046991
FRA	France	Europe	67,106,161	0.040220
DEU	Germany	Europe	80,594,017	0.049371
IRQ	Iraq	Asia	39,192,111	0.015225
ISR	Israel	Asia	8,299,706	0.035784
TUR	Turkey	Asia	80,845,215	0.020657
UKR	Ukraine	Europe	44,033,874	0.008007

Tab. A.4: Klaster 3. (źródło: opracowanie własne)

iso_a3	name	continent	pop_est	gdp_per_cap
BHS	Bahamas	North America	329,988	0.027474
BFA	Burkina Faso	Africa	20,107,509	0.001641
KHM	Cambodia	Asia	16,204,486	0.003637
CMR	Cameroon	Africa	24,994,885	0.003090
TCD	Chad	Africa	12,075,985	0.002533
CIV	Côte d'Ivoire	Africa	24,184,810	0.003602
COD	Dem. Rep. Congo	Africa	83,301,151	0.000792
DOM	Dominican Rep.	North America	10,734,247	0.015083
SLV	El Salvador	North America	6,172,011	0.008877
EST	Estonia	Europe	1,251,581	0.030921
GMB	Gambia	Africa	2,051,363	0.001651
GIN	Guinea	Africa	12,413,867	0.001295
GUY	Guyana	South America	737,718	0.008259
HTI	Haiti	North America	10,646,714	0.001817
KGZ	Kyrgyzstan	Asia	5,789,122	0.003629
LBR	Liberia	Africa	4,689,021	0.000828
LUX	Luxembourg	Europe	594,130	0.098867
MWI	Malawi	Africa	19,196,246	0.001104
MDA	Moldova	Europe	3,474,121	0.005337
MOZ	Mozambique	Africa	26,573,706	0.001317
NIC	Nicaragua	North America	6,025,951	0.005568
NER	Niger	Africa	19,245,344	0.001047
PAN	Panama	North America	3,753,142	0.024811
RWA	Rwanda	Africa	11,901,484	0.001846
SSD	S. Sudan	Africa	13,026,129	0.001603
SEN	Senegal	Africa	14,668,522	0.002708
SVK	Slovakia	Europe	5,445,829	0.030996
TZA	Tanzania	Africa	53,950,935	0.002791
URY	Uruguay	South America	3,360,148	0.021800
UZB	Uzbekistan	Asia	29,748,859	0.006800
ZMB	Zambia	Africa	15,972,000	0.004080

Tab. A.5: Klaster 4. (źródło: opracowanie własne)

iso_a3	name	continent	pop_est	gdp_per_cap
ARG	Argentina	South America	44,293,293	0.019854
BGD	Bangladesh	Asia	157,826,578	0.003982
BEL	Belgium	Europe	11,491,346	0.044259
COL	Colombia	South America	47,698,524	0.014424
JOR	Jordan	Asia	10,248,069	0.008410
KEN	Kenya	Africa	47,615,739	0.003207
LBN	Lebanon	Asia	6,229,794	0.013670
MYS	Malaysia	Asia	31,381,992	0.027500
NLD	Netherlands	Europe	17,084,719	0.050970
NZL	New Zealand	Oceania	4,510,327	0.038756
SWE	Sweden	Europe	9,960,487	0.050008
ARE	United Arab Emirates	Asia	6,072,475	0.109873
VEN	Venezuela	South America	31,304,016	0.014969

Tab. A.6: Klaster 5. (źródło: opracowanie własne)

iso_a3	name	continent	pop_est	gdp_per_cap
DZA	Algeria	Africa	40,969,443	0.014875
ARM	Armenia	Asia	3,045,191	0.008637
AUT	Austria	Europe	8,754,413	0.047587
AZE	Azerbaijan	Asia	9,961,396	0.016855
BLR	Belarus	Europe	9,549,747	0.017320
BOL	Bolivia	South America	11,138,234	0.007034
BGR	Bulgaria	Europe	7,101,510	0.020151
CHL	Chile	South America	17,789,267	0.024515
HRV	Croatia	Europe	4,292,095	0.021957
CUB	Cuba	North America	11,147,407	0.011922
CYP	Cyprus	Asia	1,221,549	0.023953
CZE	Czechia	Europe	10,674,723	0.032872
DNK	Denmark	Europe	5,605,948	0.047236
ETH	Ethiopia	Africa	105,350,020	0.001658
FIN	Finland	Europe	5,491,218	0.040817
GHA	Ghana	Africa	27,499,924	0.004393
GTM	Guatemala	North America	15,460,732	0.008525
MLI	Mali	Africa	17,885,245	0.002130
MAR	Morocco	Africa	33,986,655	0.008321
MMR	Myanmar	Asia	55,123,814	0.005644
NPL	Nepal	Asia	29,384,297	0.002434
PRK	North Korea	Asia	25,248,140	0.001584
NOR	Norway	Europe	5,320,045	0.068552
OMN	Oman	Asia	3,424,386	0.050549
PER	Peru	South America	31,036,656	0.013223
PRT	Portugal	Europe	10,839,514	0.027409
ZAF	South Africa	Africa	54,841,552	0.013477
LKA	Sri Lanka	Asia	22,409,381	0.010563
SDN	Sudan	Africa	37,345,935	0.004721
TWN	Taiwan	Asia	23,508,428	0.047940
THA	Thailand	Asia	68,414,135	0.016970
TUN	Tunisia	Africa	11,403,800	0.011470
VNM	Vietnam	Asia	96,160,163	0.006187
YEM	Yemen	Asia	28,036,829	0.002620
ZWE	Zimbabwe	Africa	13,805,084	0.002052

Tab. A.7: Klaster 6. (źródło: opracowanie własne)

iso_a3	name	continent	pop_est	gdp_per_cap
IND	India	Asia	1,281,935,911	0.006803
IDN	Indonesia	Asia	260,580,739	0.011620
ITA	Italy	Europe	62,137,802	0.035743
JPN	Japan	Asia	126,451,398	0.039003
LBY	Libya	Africa	6,653,210	0.013661
MEX	Mexico	North America	124,574,795	0.018519
NGA	Nigeria	Africa	190,632,261	0.005713
PAK	Pakistan	Asia	204,924,861	0.004822
ESP	Spain	Europe	48,958,159	0.034519

Tab. A.8: Klaster 7. (źródło: opracowanie własne)

iso_a3	name	continent	pop_est	gdp_per_cap
USA	United States of America	North America	326,625,791	0.056823

Tab. A.9: Klaster 8. (źródło: opracowanie własne)

iso_a3	name	continent	pop_est	gdp_per_cap
AFG	Afghanistan	Asia	34,124,811	0.001878
BRA	Brazil	South America	207,353,391	0.014859
EGY	Egypt	Africa	97,041,072	0.011387
GRC	Greece	Europe	10,768,477	0.026977
IRL	Ireland	Europe	5,011,102	0.064257
PSE	Palestine	Asia	4,543,126	0.004671
PHL	Philippines	Asia	104,256,076	0.007692
POL	Poland	Europe	38,476,269	0.027342
SAU	Saudi Arabia	Asia	28,571,770	0.060584
KOR	South Korea	Asia	51,181,299	0.037690
CHE	Switzerland	Europe	8,236,303	0.060258
SYR	Syria	Asia	18,028,549	0.002789

Tab. A.10: Klaster 9. (źródło: opracowanie własne)

Dodatek B

Parametry klastrów - dane standaryzowane

W tabelach od B.1 do B.9 pokazane zostały parametry poszczególnych klastrów uzyskanych podczas grupowania przeprowadzonego w rozdziale 6: średnia, mediana, odchylenie standardowe, minimum, maksimum. Pominięty został klaster 8 który zawiera tylko jeden element.

	max	mean	median	min	std
Events	1.5e+04	3e+03	1.4e+03	20	3.7e+03
sumNumMentions	7.9e+04	1.5e+04	6.1e+03	66	1.8e+04
materialConfCoop	0.13	0.074	0.073	0	0.025
verbalConfCoop	0.13	0.081	0.082	0	0.025
avgAvgTone	0.4	-0.59	-0.46	-3.2	0.77
avgGoldstein	2.7	1.8	1.7	1.2	0.35
fightCount	5.4e+02	98	34	0	1.4e+02
expressCount	2.9e+02	47	22	0	65
pop_est	9.6e+07	1e+07	4.9e+06	2.8e+05	1.8e+07
gdp_pc	0.14	0.028	0.015	0.0015	0.036
GDP	4.6e+11	7.7e+10	2.3e+10	7.6e+08	1.1e+11
health	20	5.8	5.3	2.4	3.5
military	11	2	1.1	0	2.3
education	7.5	5.3	5.3	2.6	1.6
imp	1.9e+02	56	46	15	35
exp	2.2e+02	55	41	15	44

Tab. B.1: Parametry klastra 0 - dane standaryzowane. (źródło: opracowanie własne)

Tab. B.2: Parametry klastra 1 - dane standaryzowane. (źródło: opracowanie własne)

	max	mean	median	min	std
Events	5.2e+04	3.2e+04	2.8e+04	1.6e+04	1.8e+04
sumNumMentions	2.3e+05	1.4e+05	1.1e+05	7.6e+04	8.1e+04
materialConfCoop	0.19	0.17	0.17	0.14	0.024
verbalConfCoop	0.23	0.21	0.2	0.18	0.026
avgAvgTone	-2	-3	-2.9	-4.1	1
avgGoldstein	-0.23	-0.38	-0.45	-0.46	0.13
fightCount	3e+03	2.1e+03	1.9e+03	1.3e+03	8.2e+02
expressCount	1.4e+03	6.3e+02	2.9e+02	2.1e+02	6.6e+02
pop_est	1.3e+09	4.3e+08	6.2e+06	4.5e+06	7.4e+08
gdp_pc	0.014	0.0084	0.0068	0.0047	0.0047
GDP	2.1e+12	7.2e+11	5e+10	1.3e+10	1.2e+12
health	7.7	5.6	5.6	3.6	2.9
military	4.5	3.5	3.5	2.4	1.5
education	5.1	5.1	5.1	5.1	nan
imp	59	43	49	22	19
exp	23	20	20	18	2.4

	max	mean	median	min	std
Events	5.3e+03	2.3e+03	1.4e+03	2.8e+02	2.6e+03
sumNumMentions	3e+04	1.2e+04	5.9e+03	9.4e+02	1.5e+04
materialConfCoop	0.1	0.065	0.08	0.011	0.049
verbalConfCoop	0.19	0.16	0.15	0.15	0.023
avgAvgTone	-0.23	-0.9	-1.2	-1.2	0.58
avgGoldstein	1.2	0.99	1	0.75	0.23
fightCount	1.6e+02	68	41	0	84
expressCount	61	25	12	3	31
pop_est	1.1e+07	5.1e+06	2.2e+06	1.8e+06	5.4e+06
gdp_pc	0.016	0.0098	0.011	0.0016	0.0075
GDP	4.3e+10	2e+10	1.4e+10	1e+09	2.2e+10
health	8.6	7.1	7	5.7	1.4
military	2.7	2.2	2.3	1.6	0.52
education	6.6	6.6	6.6	6.6	nan
imp	56	47	52	32	13
exp	50	39	41	28	11

Tab. B.3: Parametry klastra 2 - dane standaryzowane. (źródło: opracowanie własne)

	max	mean	median	min	std
Events	1.6e+05	9.2e+04	7.5e+04	6.3e+04	3.4e+04
sumNumMentions	7.5e+05	4.4e+05	3.7e+05	3e+05	1.6e+05
materialConfCoop	0.16	0.13	0.12	0.099	0.023
verbalConfCoop	0.14	0.12	0.12	0.1	0.015
avgAvgTone	-1.4	-2.2	-2.2	-3.2	0.47
avgGoldstein	1.4	0.84	0.96	0.011	0.45
fightCount	7.6e+03	5e+03	5e+03	2.8e+03	1.7e+03
expressCount	3.2e+03	2e+03	1.9e+03	9.1e+02	8.2e+02
pop_est	1.4e+09	1.9e+08	6.6e+07	8.3e+06	4.2e+08
gdp_pc	0.051	0.034	0.038	0.008	0.015
GDP	1.1e+13	2.5e+12	1.5e+12	9.1e+10	3.2e+12
health	11	8	8.2	4.1	2.7
military	5.6	2.7	1.9	1.2	1.6
education	5.9	5.1	5.3	3.8	0.8
imp	55	30	29	18	11
exp	53	31	30	20	11

Tab. B.4: Parametry klastra 3 - dane standaryzowane. (źródło: opracowanie własne)

	max	mean	median	min	std
Events	2.7e+04	7.5e+03	4.1e+03	1.1e+03	9.1e+03
sumNumMentions	1.2e+05	3.5e+04	1.9e+04	5.2e+03	4e+04
materialConfCoop	0.31	0.25	0.25	0.19	0.047
verbalConfCoop	0.14	0.11	0.11	0.081	0.022
avgAvgTone	-2.8	-3.8	-4	-4.5	0.62
avgGoldstein	-0.28	-0.85	-0.66	-1.7	0.5
fightCount	5.6e+03	1.4e+03	7.5e+02	1.6e+02	1.8e+03
expressCount	4.2e+02	1e+02	40	12	1.4e+02
pop_est	4.8e+07	2.1e+07	2e+07	2e+06	1.4e+07
gdp_pc	0.0032	0.0023	0.0027	0.00063	0.001
GDP	6.4e+10	2.6e+10	2.1e+10	2.4e+09	2.3e+10
health	8.4	5.9	5.5	4.2	1.6
military	1.8	1.4	1.3	1.2	0.28
education	6	4.6	4.7	2.8	1.4
imp	86	41	31	28	25
exp	43	24	22	14	11

Tab. B.5: Parametry klastra 4 - dane standaryzowane. (źródło: opracowanie własne)

	max	mean	median	min	std
Events	4.1e+04	1e+04	6e+03	2.1e+02	1.2e+04
sumNumMentions	1.6e+05	4.9e+04	2.8e+04	9e+02	5.1e+04
materialConfCoop	0.16	0.12	0.12	0.092	0.019
verbalConfCoop	0.19	0.15	0.15	0.13	0.017
avgAvgTone	-1.7	-2.1	-2.1	-2.8	0.28
avgGoldstein	1	0.59	0.59	0.13	0.23
fightCount	2.4e+03	5.9e+02	2.3e+02	5	7.7e+02
expressCount	6.4e+02	1.8e+02	1.3e+02	4	1.9e+02
pop_est	2e+08	3.6e+07	1.8e+07	1.5e+06	5.5e+07
gdp_pc	0.061	0.012	0.0057	0.00069	0.016
GDP	6.5e+11	1.6e+11	3.3e+10	1.4e+09	2.1e+11
health	11	6.4	6.6	2.6	2.2
military	13	2.4	1.7	0.00089	2.8
education	7.5	4.2	4.3	2.1	1.7
imp	63	38	37	11	14
exp	49	26	29	5.7	12

Tab. B.6: Parametry klastra 5 - dane standaryzowane. (źródło: opracowanie własne)

	max	mean	median	min	std
Events	3.9e+04	1e+04	5.9e+03	3.2e+02	1e+04
sumNumMentions	2.2e+05	5.4e+04	3.3e+04	2.1e+03	5.5e+04
materialConfCoop	0.13	0.11	0.11	0.065	0.017
verbalConfCoop	0.13	0.11	0.12	0.081	0.013
avgAvgTone	-0.37	-1.3	-1.2	-2.5	0.45
avgGoldstein	1.8	1.1	1.1	0.5	0.26
fightCount	2.7e+03	4.2e+02	2.4e+02	10	5e+02
expressCount	9.3e+02	2.2e+02	1.1e+02	2	2.4e+02
pop_est	2.6e+08	3e+07	9.9e+06	5.9e+05	5.1e+07
gdp_pc	0.069	0.021	0.017	0.00083	0.017
GDP	4.4e+12	3.4e+11	6.5e+10	3.2e+09	7.4e+11
health	13	7.2	7	2.7	2.4
military	5.5	1.6	1.3	0	1
education	7.6	4.9	5	2.2	1.2
imp	1.1e+02	45	43	11	22
exp	1.2e+02	40	35	8.2	23

Tab. B.7: Parametry klastra 6 - dane standaryzowane. (źródło: opracowanie własne)

	max	mean	median	min	std
Events	1.7e+05	1.2e+05	1.2e+05	7.3e+04	6.9e+04
sumNumMentions	7.3e+05	5.2e+05	5.2e+05	3.1e+05	2.9e+05
materialConfCoop	0.3	0.29	0.29	0.28	0.019
verbalConfCoop	0.2	0.18	0.18	0.17	0.02
avgAvgTone	-4.5	-4.5	-4.5	-4.5	0.02
avgGoldstein	-1.8	-1.8	-1.8	-1.9	0.063
fightCount	3.3e+04	2.4e+04	2.4e+04	1.5e+04	1.3e+04
expressCount	2.2e+03	1.4e+03	1.4e+03	5.1e+02	1.2e+03
pop_est	8.2e+07	6.1e+07	6.1e+07	3.9e+07	3e+07
gdp_pc	0.018	0.017	0.017	0.015	0.0018
GDP	3.8e+11	2.8e+11	2.8e+11	1.8e+11	1.5e+11
health	7.8	5.5	5.5	3.1	3.3
military	5.3	4.1	4.1	2.8	1.8
education	2.8	2.8	2.8	2.8	nan
imp	41	30	30	19	16
exp	35	28	28	20	11

Tab. B.8: Parametry klastra 7 - dane standaryzowane

	max	mean	median	min	std
Events	3.6e+04	5.4e+03	2.5e+03	3.7e+02	7.2e+03
sumNumMentions	1.9e+05	2.7e+04	1.3e+04	1.7e+03	3.7e+04
materialConfCoop	0.23	0.16	0.16	0.11	0.026
verbalConfCoop	0.13	0.097	0.098	0.043	0.02
avgAvgTone	-1.1	-2.2	-2.1	-3.9	0.72
avgGoldstein	1.4	0.65	0.64	-0.014	0.31
fightCount	2.4e+03	3.2e+02	1.5e+02	22	4.6e+02
expressCount	7.5e+02	1.2e+02	56	6	1.7e+02
pop_est	1.6e+08	2.7e+07	9e+06	3.3e+05	4e+07
gdp_pc	0.1	0.014	0.0083	0.00057	0.02
GDP	1.2e+12	1.3e+11	2.2e+10	1.2e+09	2.6e+11
health	10	5.7	5.8	1.9	2
military	11	1.9	1.3	0.4	2
education	6.8	4.4	4.4	1.5	1.4
imp	96	46	41	24	19
exp	92	38	34	10	21

Tab. B.9: Parametry klastra 9 - dane standaryzowane. (źródło:
opracowanie własne)

Dodatek C

Korzystanie z programu

Do uruchomienia kodu projektu niezbędne jest internetowe interaktywne środowisko obliczeniowe Notebook Jupyter [7]. JupyterLab może być zainstalowane przy pomocy *conda*, *pip* lub uruchomiony w kontenerze *docker*.

W celu skorzystania z kodu aplikacji projektu należy pobrać repozytorium GitHub. W folderze z repozytorium należy uruchomić notatnik jupyter polecienniem *jupyter notebook*. Użyte polecenie spowoduje uruchomienie serwera *Notebook* oraz przeglądarki w której dostępny będzie pulpit sterowania. W pulpicie będą widoczne poszczególne pliki programu. Wskazanie wybranego pliku spowoduje otwarcie notatnika wraz z wynikami ostatniego wykonania programu. Kod programu może być dowolnie modyfikowany.

W celu nowego wykonania kodu notatnika należy nacisnąć przycisk *RUN*.

Bibliografia

- [1] D. Bodas-Sagi and J. é. Labeaga, “Using GDELT Data to Evaluate the Confidence on the Spanish Government Energy Policy,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 3, no. 6, p. 38, 2016.
- [2] *GDELT*. [Online]. Available: www.gdeltproject.org (visited on 06/24/2020).
- [3] *GDELT Documentation*. [Online]. Available: <https://www.gdeltproject.org/data.html#documentation> (visited on 06/15/2020).
- [4] *GeoPandas*. [Online]. Available: geopandas.org (visited on 06/29/2020).
- [5] *Google BigQuery*. [Online]. Available: cloud.google.com/bigquery (visited on 06/24/2020).
- [6] *ISO 3166-1 alfa-3*. [Online]. Available: https://pl.wikipedia.org/wiki/ISO_3166-1_alfa-3 (visited on 07/09/2020).
- [7] *Jupyter Notebook*. [Online]. Available: jupyter.org (visited on 06/24/2020).
- [8] S. Keertipati, B. T. R. Savarimuthu, M. Purvis, and M. Purvis, “Multi-level analysis of peace and conflict data in GDELT,” *ACM International Conference Proceeding Series*, vol. 02-December-2014, pp. 33–40, 2014.
- [9] H. Kwak and J. An, “Two tales of the world: Comparison of widely used world news datasets GDELT and eventregistry,” *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, pp. 619–622, 2016. arXiv: 1603.01979.
- [10] M. Ma, P. Fang, J. Gao, and C. Song, “Does ideology affect the tone of international news coverage?” *Proceedings of 4th International Conference on Behavioral, Economic, and Socio-Cultural Computing, BEESC 2017*, vol. 2018-Janua, pp. 1–5, 2017.
- [11] *Pandas*. [Online]. Available: pandas.pydata.org/ (visited on 06/24/2020).

- [12] *PyCharm*. [Online]. Available: www.jetbrains.com/pycharm/ (visited on 06/24/2020).
- [13] *Python*. [Online]. Available: www.python.org (visited on 06/24/2020).
- [14] P. A. Schrodt, “Automated Production of High-Volume, Near-Real-Time Political Event Data,” *American Political Science Association meetings*, pp. 1–29, 2010. [Online]. Available: http://qipsr.as.uky.edu/sites/default/files/Schrodt_EventData_Princeton2011.pdf.
- [15] *Scikit-Learn*. [Online]. Available: scikit-learn.org (visited on 06/26/2020).
- [16] *Scikit-Learn StandardScaler*. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html> (visited on 07/09/2020).
- [17] B. Wang, Y. Zhang, and D. Wang, “Research on a New Metadata Model of Political Event Data Set,” *Proceedings - 4th IEEE International Conference on Big Data Security on Cloud, BigDataSecurity 2018, 4th IEEE International Conference on High Performance and Smart Computing, HPSC 2018 and 3rd IEEE International Conference on Intelligent Data and Security, IDS 2018*, pp. 14–19, 2018.
- [18] World Bank, *World Bank Data License*. [Online]. Available: <https://datacatalog.worldbank.org/public-licenses#cc-by> (visited on 08/07/2020).
- [19] *World Bank DataBank*. [Online]. Available: databank.worldbank.org (visited on 08/06/2020).
- [20] *World's largest event dataset now publicly available in BigQuery*, 2014. [Online]. Available: <https://cloudplatform.googleblog.com/2014/05/worlds-largest-event-dataset-now-publicly-available-in-google-bigquery.html> (visited on 06/15/2020).
- [21] Y. Yuan, “Modeling inter-country connection from geotagged news reports: A time-series analysis,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10387 LNCS, pp. 183–190, 2017.
- [22] Y. Yuan, Y. Liu, and G. Wei, “Exploring inter-country connection in mass media: A case study of China,” *Computers, Environment and Urban Systems*, vol. 62, pp. 86–96, 2017.