Eksploracyjna analiza danych
Światowy program szczepień przeciwko COVID-19

Marek Grudkowski 156587
Kamil Kaczmarkiewicz 171701

1 maja 2021

1 Ogólny opis danych

Zbiór danych dotyczy aktualnego postępu poszczególnych państw w szczepieniach przeciwko COVID-19. Zawiera on informacje pochodzące prawie ze wszystkich krajów na świecie podzielone na poszczególne dni. Program szczepień przeciwko COVID to w dobie pandemii niezwykle gorący temat. Naszym zdaniem warto się na nim skupić, gdyż może zawierać wiele ukrytych informacji, które mogą przydać się w walce z pandemią i przyspieszyć sam proces szczepień.

2 Cel eksploracji i kryteria sukcesu

Celem eksploracji danych jest zgłębienie wszelkich tajemnic ukrytych w analizowanym zbiorze danych. Takie informacje można wykorzystać do wskazania zarówno państw, które najlepiej przeprowadzają program szczepień, jak i tych w których proces ten przebiega bardzo słabo. Dzięki temu niektóre państwa mogłyby się wzorować na tych państwach, które radzą sobie najlepiej oraz uniknąć błędów, jakie popełniły najgorsze w rankingu państwa. Dodatkowym celem może być predykcja zapotrzebowania szczepionek w danych krajach na najbliższe miesiące, która mogłaby by zapobiec marnowaniu się dawek oraz pozwolić rządom na świecie lepiej zaplanować *zakupy* oraz logistykę akcji na terenie swoich państw. Jako ostatni cel można uznać wskazanie prawdopodobnej daty zakończenia światowego programu szczepień.

Jako kryteria sukcesu eksploracji danych można uznać odpowiednio wysoką (przykładowo 80 %) zgodność predykcji wykonywanych szczepień w danych państwach z rzeczywistymi danymi, które są aktualizowane na bieżąco.

3 Charakterystyka zbioru danych

Zbiór danych na stan dnia pisania tego sprawozdania zawiera ponad 13300 przykładów. Dane aktualizowane są zazwyczaj co kilka dni i pochodzą z wielu różnych źródeł. Zazwyczaj są nimi organy krajowe lub lokalne, czy międzynarodowe organizacje. Dla każdego przykładu podane jest źródło i jego adres internetowy, co daje możliwość weryfikacji w przypadku jakichkolwiek wątpliwości co do poprawności danych. Dane zapisane są w jednym pliku w formacie csv i podzielone są na następujące kolumny:

country
Nazwa państwa lub regionu, którego dotyczy dany przykład.

ISO code
Trzyliterowy kod państwa zgodny z normą ISO $3166-1$

date
Data pozyskania danych.

total vaccinations
Całkowita liczba podanych dawek. Zliczane są tutaj pojedyncze dawki i nie mogą być równe całkowitej liczbie zaszczepionych osób, w zależności od schematu dawkowania np. jedna osoba może przyjąć kilka dawek.

total vaccinations per hundred
Całkowita liczba podanych dawek szczepionki w przeliczeniu na sto osób w liczbie ludności całego kraju.

daily vaccinations raw
Dzienna zmiana w całkowitej liczbie podanych dawek. Oblicza się ją tylko dla kolejnych dni. Surowy środek w celu kontroli danych i jej przejrzystości. Autorzy zestawu nie zalecają korzystania z tego atrybutu.

daily vaccinations
Liczba dawek podawanych dziennie. Liczba ta jest wygładzana w ujęciu 7 dni. W przypadku krajów, które nie przekazują danych w ujęciu dziennym zakłada się, że dawki zmieniały się jednakowo we wszystkich okresach, w których nie przekazywano danych. Tak wypełnione dane uśrednia się dodatkowo w 7 dniowym oknie.

daily vaccinations per million
Liczba dawek podawanych dziennie w przeliczeniu na milion osób w ludności całego kraju

people vaccinated
Całkowita liczba osób, które otrzymały przynajmniej jedną dawkę szczepionki. Jeśli osoba otrzyma pierwszą dawkę liczba zwiększana jest jeden, jeśli otrzyma drugą, pozostaje taka sama.

people vaccinated per hundred
Całkowita liczba osób, które otrzymały przynajmniej jedną dawkę szczepionki w przeliczeniu na sto osób w liczbie ludności całego kraju.

people fully vaccinated
Całkowita liczba osób, które otrzymały wszystkie dawki zgodnie ze schematem szczepienia. Jeśli bierzemy pod waugę schemat szczepień z dwoma dawkami - przy pierwszej dawce liczba nie zmienia się, po drugiej dawce zwiększana jest o 1.

people fully vaccinated per hundred
Całkowita liczba osób, które otrzymały wszystkie dawki zgodnie ze schematem szczepienia w przeliczeniu na sto osób w liczbie ludności całego kraju.

source name
Nazwa źródła, z którego pochodzą dane.

source website
Strona internetowa, źródła z którego pochodzą dane.

4 Eksploracyjna analiza danych

4.1 Atrybuty nominalne

Podczas *stricte* pracy z danymi część atrybutów będzie nieprzydatna. Do takich atrybutów z pewnością należy źródło danych oraz jego adres w Internecie. Z tego powodu można pozbyć się tych kolumn i zmniejszyć rozmiar naszego zbioru.

Kolejny brany pod uwagę atrybut, który jest do odrzucenia to kod państwa. Przy wizualizacji wyników nie powie on dużo osobie, gdyż są to tylko 3 litery. Może jednak okazać się on przydatny w walidacji innego atrybutu jakim jest państwo. Zgodnie z opisem zbioru dostarczonym przez jego autorów część przykładów nie opisuje państw, lecz regiony wchodzące w skład innych państw np. Szkocja. Znając długość kodów ISO dla państw można wyodrębnić te przykłady, które dotyczą regionów:

```
England has code OWID_ENG
Kosovo has code OWID_KOS
Northern Cyprus has code OWID_CYN
Northern Ireland has code OWID_NIR
Scotland has code OWID_SCT
Wales has code OWID_WLS
```

Na początku został poprawiony kod przykładów, które dotyczą Kosowa. Kod wg normy ISO powinien mieć postać *XKX* i na taki został zmieniony. Kolejnym krokiem było wcielenie przykładów dotyczących Północnego Cypru do Cypru. W tym przypadku zmiana dotyczyła zarówno kolumny *country*, jak i *iso_code*. Anglia, Walia, Irlandia Północna oraz Walia są regionami, które wchodzą w skład Wielkiej Brytanii. Poprawnym rozwiązaniem w ich przypadku takich przykładów będzie dołączenie nich do Wielkiej Brytanii.

4.2 Atrybuty numeryczne

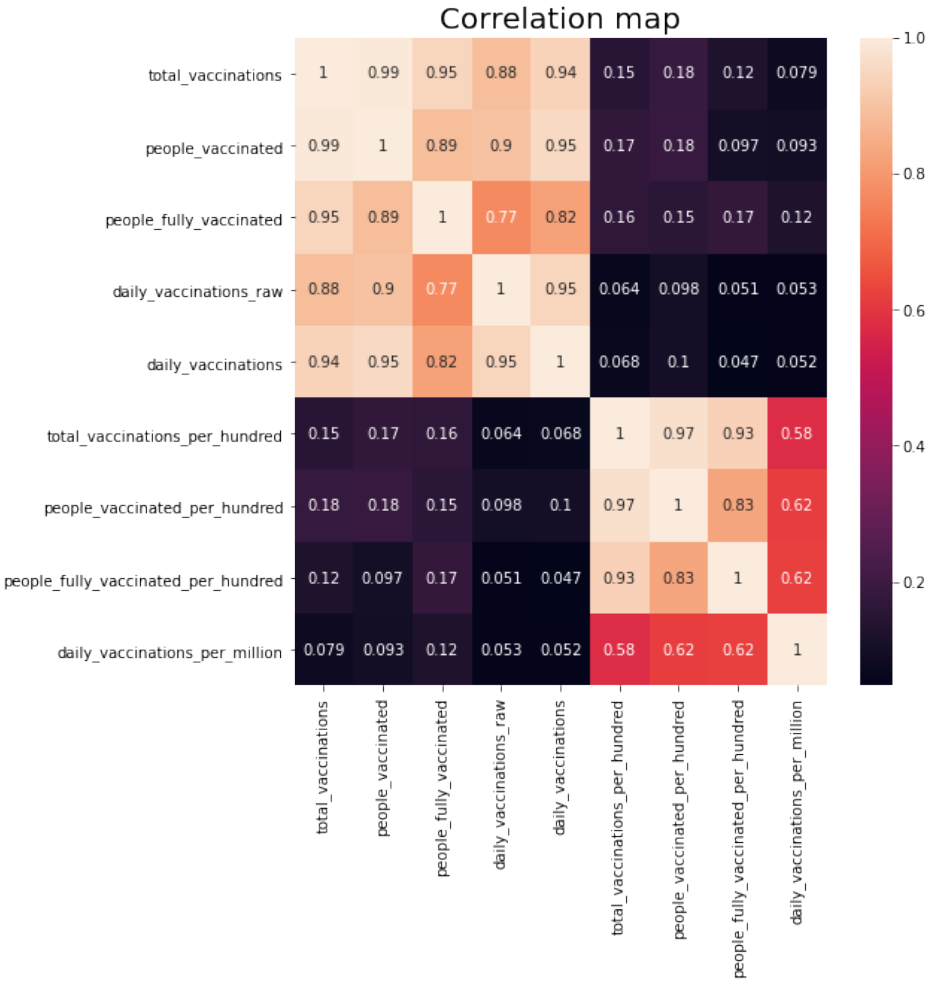
Pierwszym krokiem w analizie atrybutów numerycznych jest sprawdzenie, w jakiej liczbie występują braki w każdej kolumnie. Wyniki operacji sprawdzającej ten fakt, poniżej:

```
Size of data is: (13307, 13)
Missing values in dataset:
country                                0
iso_code                               0
date                                   0
total_vaccinations                  5255
people_vaccinated                   5931
people_fully_vaccinated             7926
daily_vaccinations_raw              6529
daily_vaccinations                   220
total_vaccinations_per_hundred      5255
people_vaccinated_per_hundred       5931
people_fully_vaccinated_per_hundred 7926
daily_vaccinations_per_million       220
vaccines                               0
```

Niestety jest są to bardzo duże braki, którymi należało się zająć. Na pierwszy rzut oka widać, że atrybuty łączą się w pary pod dwoma względami: nazwą oraz liczbą brakujących wartości. Zgodnie z opisem danych dostarczonym przez autorów w takiej jeden z atrybutów jest zależny od drugiego i pokazuje daną wartość w stosunku do populacji danego kraju. Aby upewnić się można sporządzić wykres prezentujący wartość korelacji pomiędzy poszczególnymi atrybutami (Rysunek 1)

Nie widać tutaj zależności o których wspomnieli autorzy. Powodem tego jest liczenie korelacji dla całego zbioru danych, a nie dla konkretnego państwa. Jeśli do funkcji rysującej wykres przekażemy tylko zakres danych obejmujący jedno państwo totalnie zmienia on swój wygląd. Przykładowy wykres dla Stanów Zjednoczonych można zobaczyć na rysunku nr 2.

Aby upewnić się, czy dla wszystkich państw występuje opisana zależność, można osobno dla każdego państwa obliczyć korelację pomiędzy atrybutami i sporządzić histogram, na którym role pojedynczych próbek będą pełnić państwa (Rysunek 3). Zdecydowanie widać, że na każdym z owych wykresów niemal 100% próbek wskazywało na korelację równą 1, co oznacza zgodność z tym, co napisali autorzy zestawu danych.



Rys. 1: Korelacja pomiędzy atrybutami całego zbioru danych

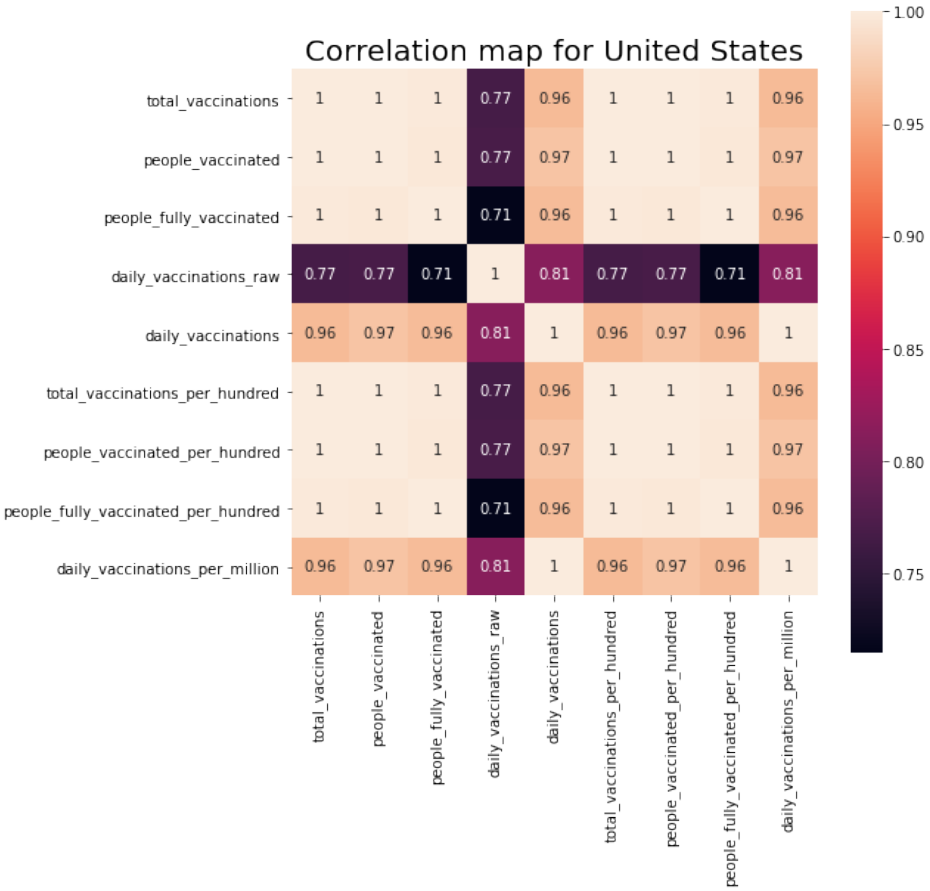
Jedynym atrybutem który pozostał bez pary jest *daily_vaccinations_raw*. Jednakże wedle zaleczeń autorów zestawu nie powinien być uwzględniany podczas analizy, dlatego można z niego zrezygnować.

Pierwszym atrybutem, w którym łatwo można uzupełnić dane jest *total_vaccinations*. Dotyczy on łącznej sumy wykonanych szczepień w danym państwie. W przypadku brakujących wartości można je wypełnić zakładając idealną liniową progresję pomiędzy odległymi wartościami.

Po wypełnieniu kolumny z totalną liczbą podanych dawek, można łatwo uzupełnić kolumnę opisującą średnią liczbę podanych dawek an sto osób. Wystarczy dla brakujących wartości wstawić liczbę opisaną zależnością:

```
liczbaPodanychDawekNaSto = totalnaLiczbaPodanychDawek / ludnoscKraju * 100
```

Aby to zrobić najpierw potrzebna jest liczba ludności w danym kraju. Aby wykorzystać w pełni nasz zbiór danych wielkość tą obliczyliśmy dla każdego państwa na podstawie średniego ilorazu liczby podanych dawek oraz liczby podanych dawek na sto osób, przy czym tutaj braliśmy pod uwagę tylko przykłady, w których nie brakowa-

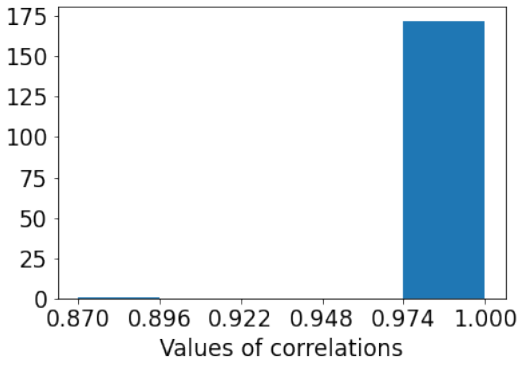


Rys. 2: Korelacja pomiędzy atrybutami zbioru danych dla USA

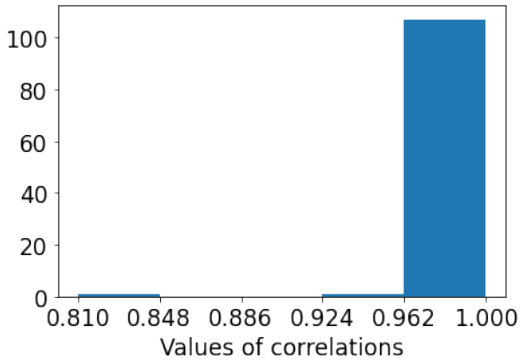
ło żadnych wartości. Ostatecznie po wypełnieniu tych atrybutów liczba brakujących danych wygląda następująco:

`country`	0
`iso_code`	0
`date`	0
`total_vaccinations`	0
`people_vaccinated`	6069
`people_fully_vaccinated`	8069
`daily_vaccinations`	226
`total_vaccinations_per_hundred`	0
`people_vaccinated_per_hundred`	6069
`people_fully_vaccinated_per_hundred`	8069
`daily_vaccinations_per_million`	226
`vaccines`	0

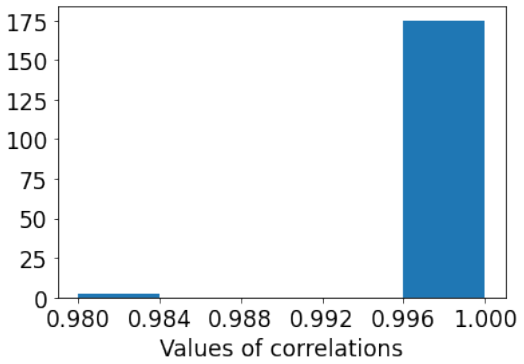
- mechanizm daily vacc
- wypełnienie luk w daily vacc



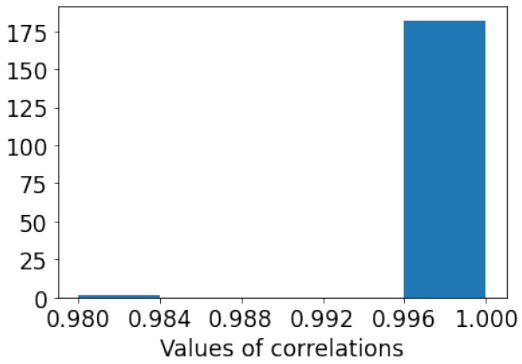
(a) Korelacja między dzienną liczbą wykonanych szczepień ogółem i na milion osób



(b) Korelacja między liczbą ludzi w pełni zaszczepionymi ogółem i na 100 osób



(c) Korelacja między liczbą ludzi którzy przyjęli choć jedną dawkę ogółem i na 100 osób



(d) Korelacja między liczbą wykonanych szczepień ogółem i na sto osób

Rys. 3: Korelacje pomiędzy poszczególnym atrybutami zbioru

- wykresy pudełkowe
- podsumowanie