# Eksploracyjna analiza danych
# Światowy program szczepień przeciwko COVID-19

Marek Grudkowski 156587
Kamil Kaczmarkiewicz 171701

29 kwietnia 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

# 3 Charakterystyka zbioru danych

Zbiór danych na stan dnia pisania tego sprawozdania zawiera ponad 13300 przykładów. Dane aktualizowane są zazwyczaj każdego dnia 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** - kraj, dla którego podawane są informacje o szczepieniu, atrybut nominalny w postaci ciągu znaków
- **iso_code** - kod ISO dla danego kraju, atrybut nominalny w postaci ciągu znaków
- **date** - data wprowadzenia danych, atrybut nominalny opisujący datę
- **total_vaccinations** - bezwzględna liczba wszystkich szczepień ochronnych w danym kraju, atrybut numeryczny (liczba naturalna)
- **people_vaccinated** - liczba osób która otrzymała szczepionkę (przy dwóch dawkach liczona jest $\times 2$), atrybut numeryczny (liczba naturalna)
- **people_fully_vaccinated** - liczba osób, które otrzymały cały zestaw szczepień, atrybut numeryczny (liczba naturalna)
- **daily_vaccinations_raw** - dla danej pozycji liczba szczepień dla tej daty/kraju, atrybut numeryczny (liczba naturalna)
- **daily_vaccinations** - dla danej pozycji liczba szczepień dla tej daty/kraju, atrybut numeryczny (liczba naturalna)
- **total_vaccinations_per_hundred** - stosunek liczby szczepień do całkowitej liczby ludności danego dnia w kraju, atrybut numeryczny wyrażany w procentach
- **people_vaccinated_per_hundred** - stosunek liczby osób zaszczepionych do całkowitej liczby ludności danego dnia w kraju, atrybut numeryczny wyrażany w procentach
- **people_fully_vaccinated_per_hundred** - stosunek liczby osób uodpornionych do całkowitej liczby ludności danego dnia w kraju, atrybut numeryczny wyrażany w procentach
- **daily_vaccinations_per_million** - stosunek między liczbą szczepień a całkowitą liczbą ludności na bieżący dzień w kraju, dodania liczba rzeczywista
- **vaccines** - rodzaje szczepionek wykorzystanych w danym kraju, atrybut nominalny, ciągi znaków rozdzielone ukośnikiem
- **source_name** - źródło informacji, atrybut nominalny, ciąg znaków
- **source_website** - strona internetowa źródła informacji, atrybut nominalny, ciąg znaków

# 4 Wyniki eksploracyjnej analizy danych

Podczas analizowania danych pierwszym etapem, było sprawdzenie w jakiej ilości występują wartości puste. Ku naszemu zaskoczeniu przykładów z takimi wartościami było naprawdę dużo. Wyniki tej operacji zliczającej takie przykłady poniżej

| **Atrybut** | **Brakujące wartości** |
|---|---|
| country | 0 |
| iso_code | 0 |
| date | 0 |
| total vaccinations | 5390 |
| people vaccinated | 6069 |
| people fully vaccinated | 8069 |
| daily vaccinations raw | 6685 |
| daily vaccinations | 226 |
| total vaccinations per hundred | 5390 |
| people vaccinated per hundred | 6069 |
| people fully vaccinated per hundred | 8069 |
| daily vaccinations per million | 226 |
| vaccines | 0 |
| source name | 0 |
| source website | 0 |

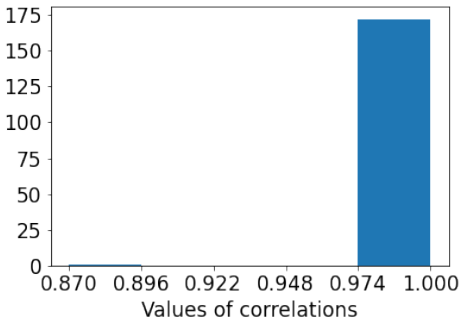
Tabela 1: Puste wartości w zbiorze danych

W tabeli można zauważyć, że liczba brakujących wartości dla atrybutów takich jak *people_fully_vaccinated_per_hundred* są takie same jak dla *people_fully_vaccinated*. Widać, że jest to para *liczba bezwzględna - liczba względna*. Na 9 atrybutów, w których występują brakujące dane, 8 z nich tworzy właśnie takie pary. Bardzo możliwe, że te atrybuty mogą być od siebie liniowo zależne i są to po prostu przykłady opisujące momenty, kiedy program szczepień nie funkcjonował jeszcze w danym kraju.

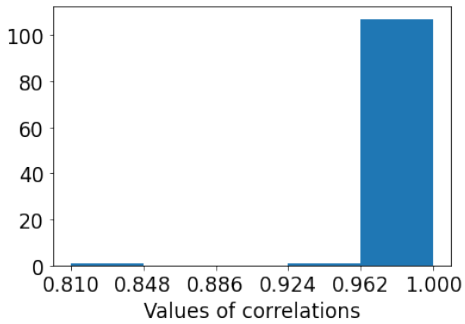
Cały zbiór przykładów jest podzielony na państwa, więc obliczenie korelacji między kolejnymi atrybutami dla wszystkich państw mogło by nie dać poprawnych wyników, gdyż dane z każdego kraju pochodzą z innego źródła. W celu uniknięcia tego błędu policzyliśmy współczynnik korelacji między poszczególnymi atrybutami dla każdego państwa osobno i z otrzymanych wartości sporządziliśmy histogram. Wyniki są zaprezentowane na wykresie poniżej.

Widać więc, że wszystkie owe atrybuty są od siebie wręcz idealnie liniowo zależne, co potwierdza nasze przypuszczenia. Dzięki tej krótkiej analizie byliśmy w stanie podjąć decyzję o uzupełnieniu *sparowanych* atrybutów zerami.

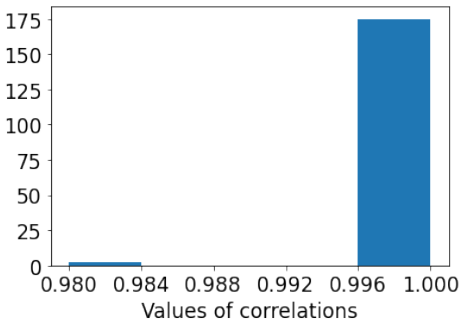
Jedynym *samotnym* atrybutem w jakim pozostały wartości *null* jest *daily_vaccinations_raw*. Jego nazwa wskazuje, że mógłby mieć powiązanie z atrybutem *daily_vaccinations*, ale żeby to stwierdzić, trzeba tak jak w poprzedniej sytuacji obliczyć korelację między ty-



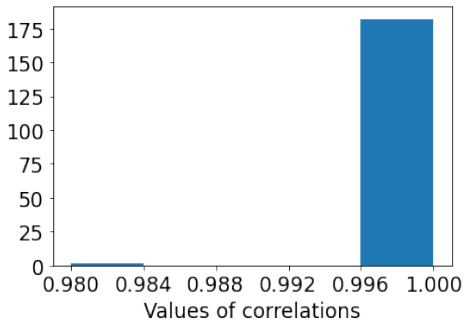

(a) Korelacja między dzienną liczbą wykonanych szczepień ogółem i na milion osób




(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

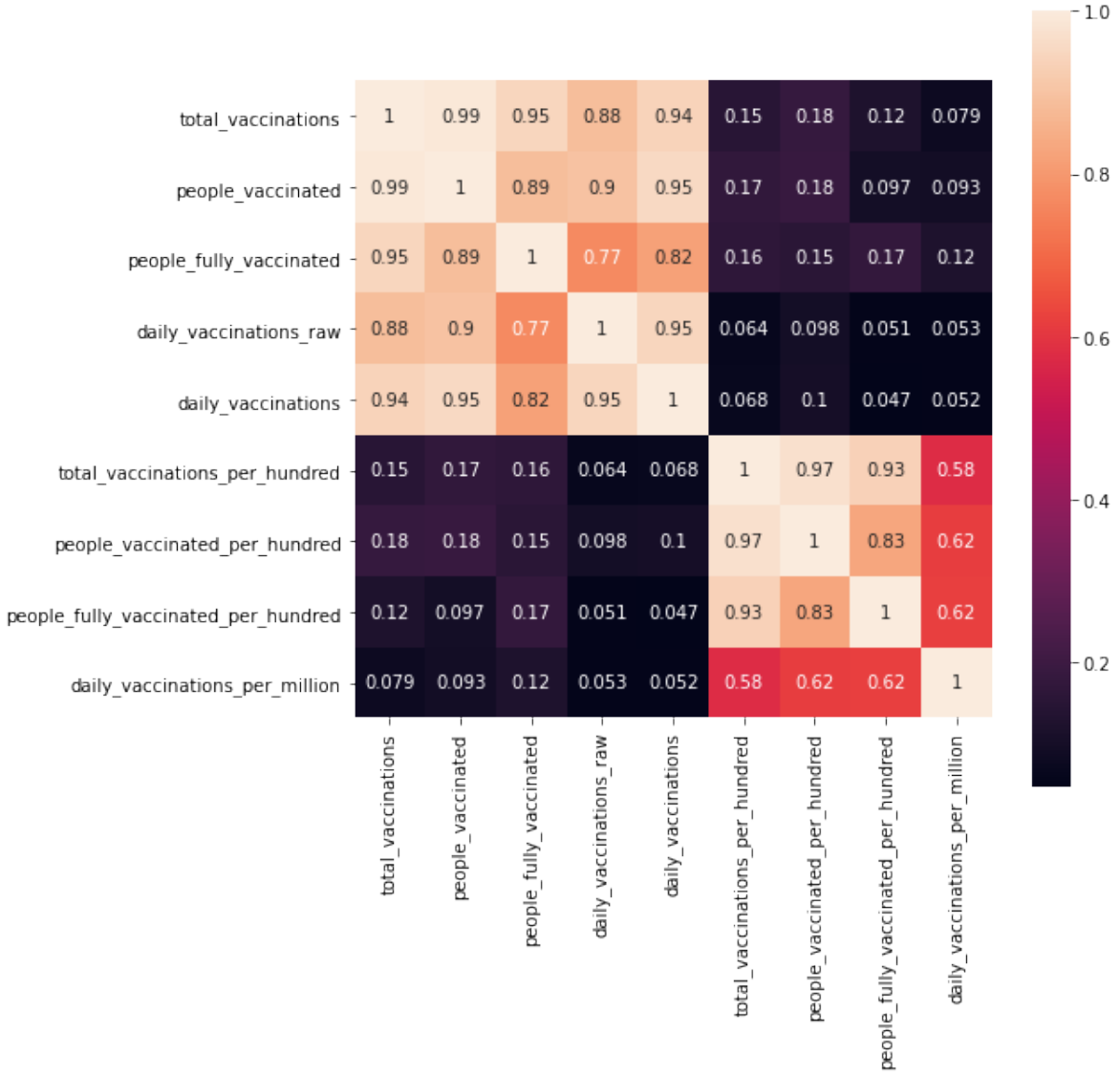
Rys. 1: Korelacje pomiędzy poszczególnym atrybutami zbioru

../img/raw_corr.png

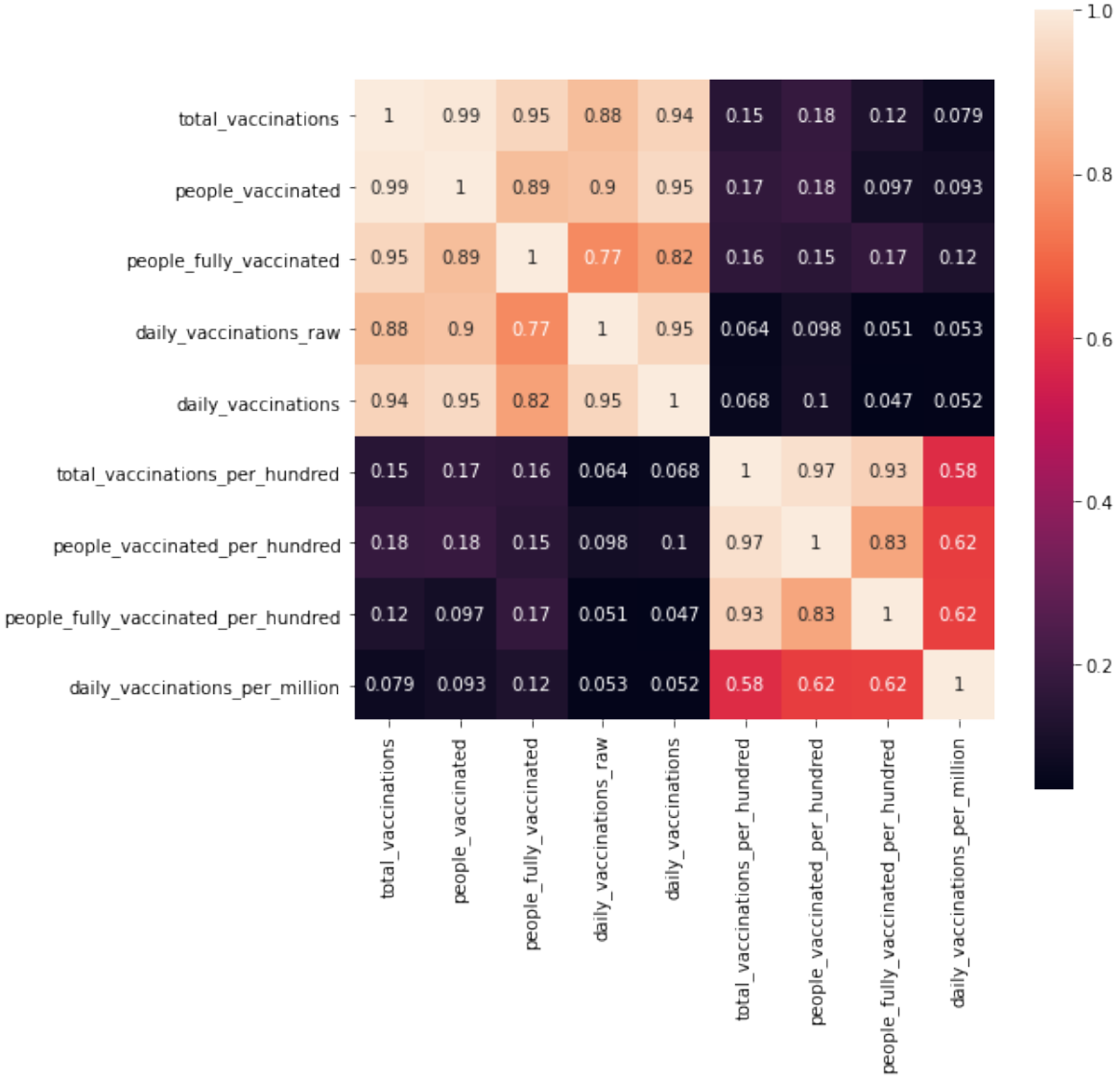
Rys. 2: Korelacje pomiędzy poszczególnym atrybutami zbioru

mi atrybutami, której rozkład wśród państw ostatecznie odrzucił tą propozycję. Wynik można zobaczyć na wykresie poniżej:

Kolejnym krokiem było utworzenie diagramu, który wizualizuje współczynnik korelacji pomiędzy poszczególnymi atrybutami. Widać na nim, że są one podzielone na dwie podgrupy, w których występują silne zależności między atrybutami. Pierwszą z nich zaklasyfikowaliśmy jako dane bezwzględne i zawiera atrybuty, które opisują bezwzględne wartości liczbowe np. liczba zaszczepionych osób. Druga grupa dotyczy atrybutów, których wartości podawane są w procentach np. liczba osób w pełni zaszczepionych na sto.



Mimo ładnego podziału zbioru atrybutów na dwie grupy, pojawiło się pytanie - dlaczego zauważamy bardzo niską korelacją pomiędzy atrybutami, które powinny być ze sobą powiązane w dużym stopniu. Chodzi tutaj na przykład o sumę ludzi w pełni zaszczepionych oraz procent ludzi w pełni zaszczepionych na sto. Być może ma to związek z tym, że powyższy diagram przedstawia dane dotyczące całego świata, a pojedyncze przykłady dotyczą konkretnych państw. By to sprawdzić zrealizowaliśmy podobny wykres, ale opisujący tylko zbiór przykładów z takich państw jak Niemcy, USA i Polska.




Rozkłady wartości dla atrybutów

rozkłady wartości atrybutów, korelacje pomiędzy wartościami atrybutów wstępne ustalenia dotyczące zawartości zbioru

# 5 Uwagi dotyczące jakości danych

dane brakujące, punkty oddalone, dane niespójne, dane niezrozumiałe,

# 6 Opis wyników eksploracji

w odniesieniu do celów eksploracji, czy dane są wystarczające, ewentualna rewizja celów