Światowy program szczepień przeciwko COVID-19

Eksploracja Danych

Marek Grudkowski
Kamil Kaczmarkiewicz

Politechnika Gdańska

Ogólny opis danych

- zbiór danych opisujący postęp światowego programu szczepień przeciwko COVID-19
- celem eksploracji w głównej mierze jest odnalezienie jak najlepszej strategii dla państw biorących udział w programie szczepień
- dane pochodzą z wielu źródeł, którymi są organy krajowe lub lokalne
- zbiór aktualnie zawiera ponad 13300 przykładów, ale co kilka dni jest aktualizowany
- jeden przykład jest opisany za pomocą 15 atrybutów

Opis atrybutów nominalnych

- **country** - nazwa regionu lub państwa, z którego pochodzą dane
- **ISO code** - trzyliterowy kod państwa zgodny z normą ISO 3166-1
- **date** - data pozyskania danych
- **source name** - nazwa organu z którego pochodzą dane
- **source website** - strona internetowa, z której pobrano dane

Opis atrybutów numerycznych

- **total vaccinations** - całkowita, sumaryczna liczba podanych dawek w danym kraju
- **total vaccinations per hundred** - powyższy atrybut, ale w przeliczeniu na stu mieszkańców
- **people vaccinated** - całkowita, sumaryczna liczba osób w dany kraju, która przyjęła choć jedną dawkę szczepionki
- **people vaccinated per hundred** - powyższy atrybut, ale w przeliczeniu na stu mieszkańców

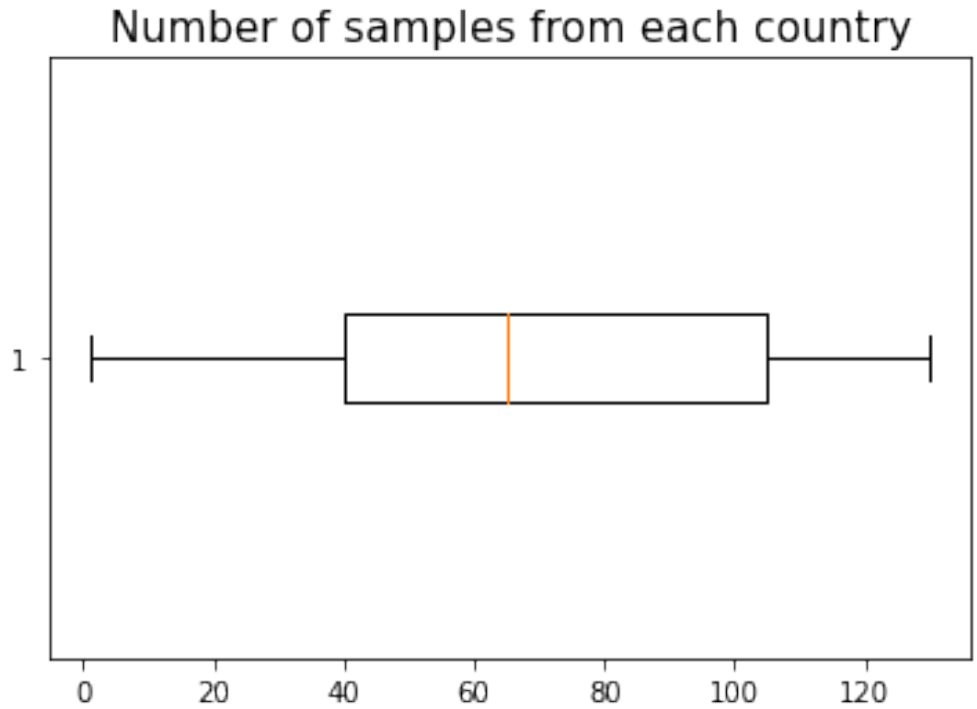
Opis atrybutów numerycznych

- **people fully vaccinated** - całkowita, sumaryczna liczba osób w dany kraju, które są w pełni zaszczepione
- **people fully vaccinated per hundred** - powyższy atrybut, ale w przeliczeniu na stu mieszkańców
- **daily vaccinations** - całkowita, sumaryczna liczba podanych dawek w danym kraju w ciągu dnia, liczba ta jest wygładzana w ujęciu 7 dni
- **daily vaccinations per milion** - powyższy atrybut, ale w przeliczeniu na milion mieszkańców
- **daily vaccinations raw** - dzienna zmiana w całkowitej liczbie podanych dawek, surowy środek wykorzystywany przy kontroli danych, autorzy zbioru nie zalecają korzystania z tego atrybutu podczas analizy

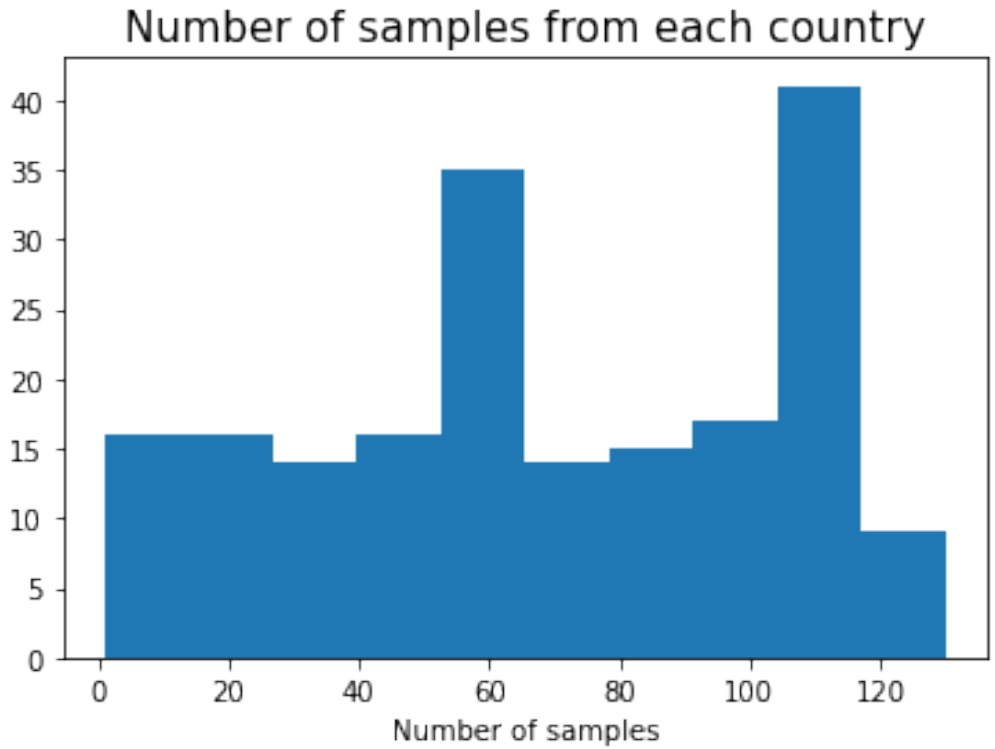
Analiza atrybutów nominalnych

- część kodów ISO nie jest zapisana wedle standardu (są to regiony leżące w obrębie różnych państw)
- można takie przykłady zaktualizować
- wówczas należy sprawdzić, czy dla danego państwa nie ma więcej niż 1 przykładu dla jednej daty

```
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
```



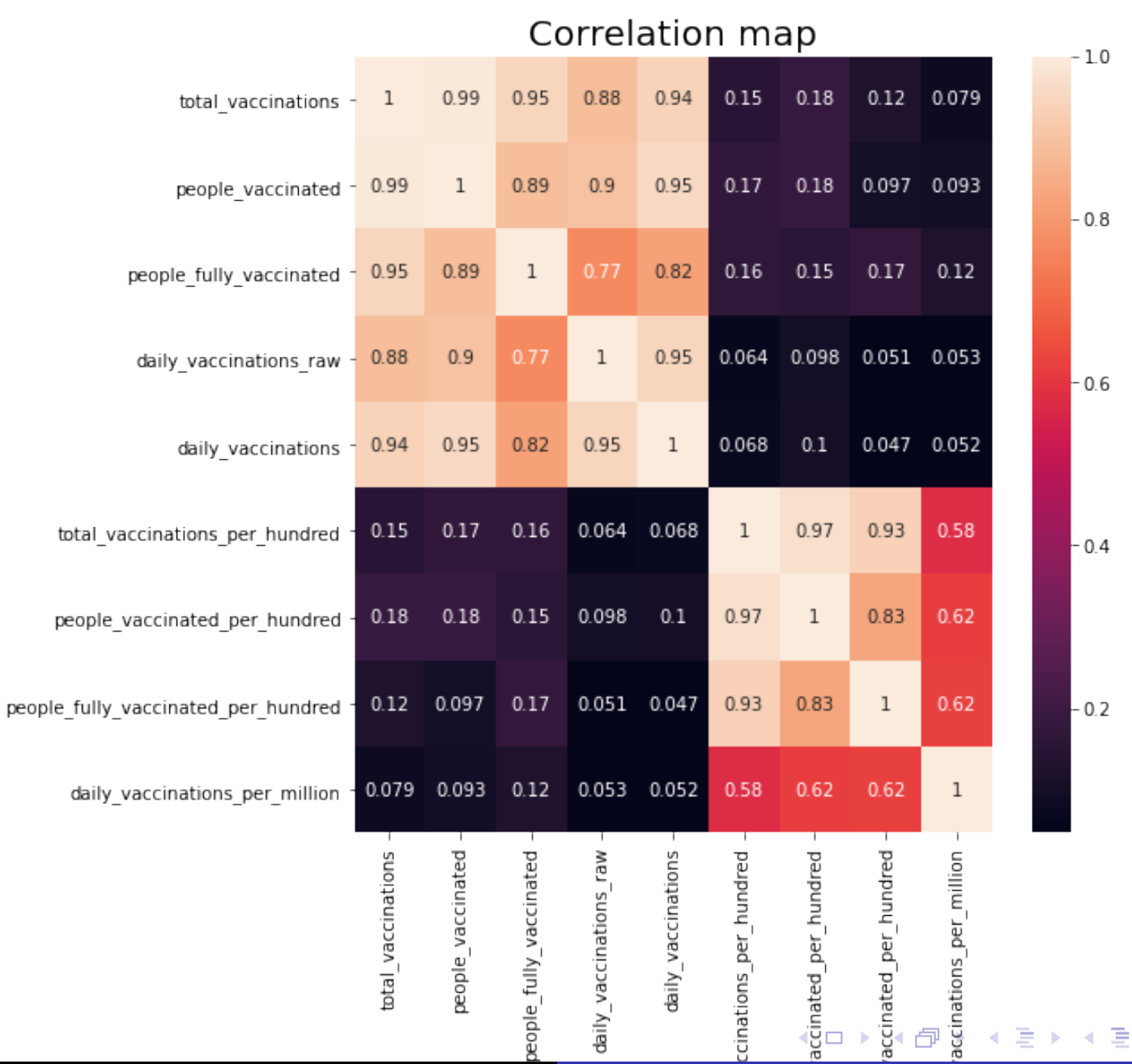
Rysunek: Liczba przykładów dostarczonych przez poszczególne państwa



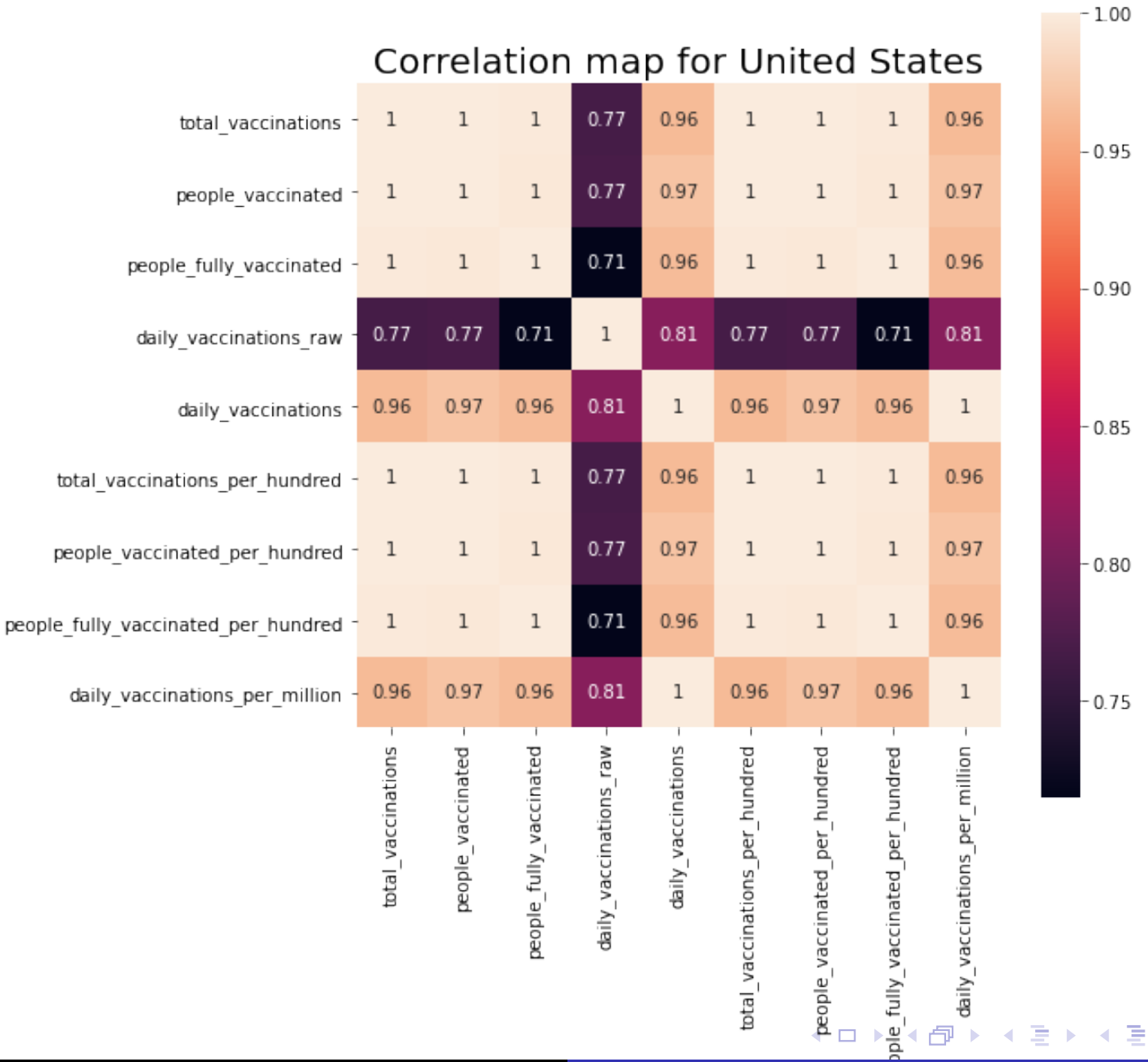
Rysunek: Liczba przykładów dostarczonych przez poszczególne państwa

- atrybuty numeryczne można połączyć w pary *wartość bezwzględna* i *wartość względna*
- takie pary cechują się korelacją równą jeden
- **daily vaccinations raw** nie analizujemy, gdyż autorzy nie zalecają korzystania z tego atrybutu

Korelacja między atrybutami dla całego zbioru



Korelacja między atrybutami podzielona na państwa

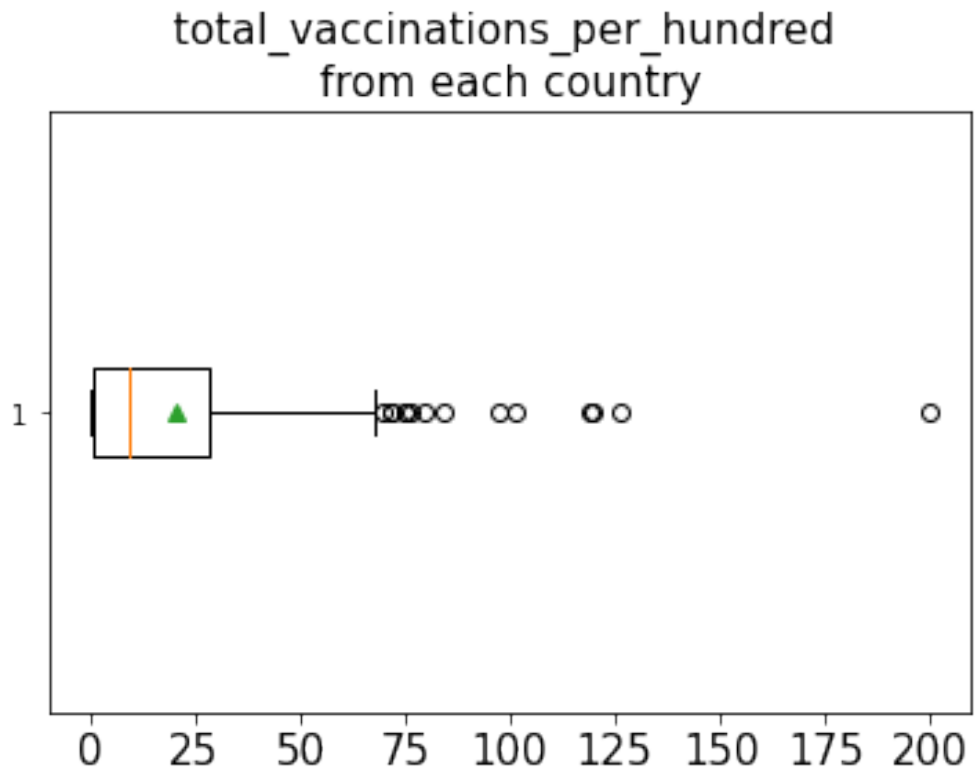


Brakujące wartości

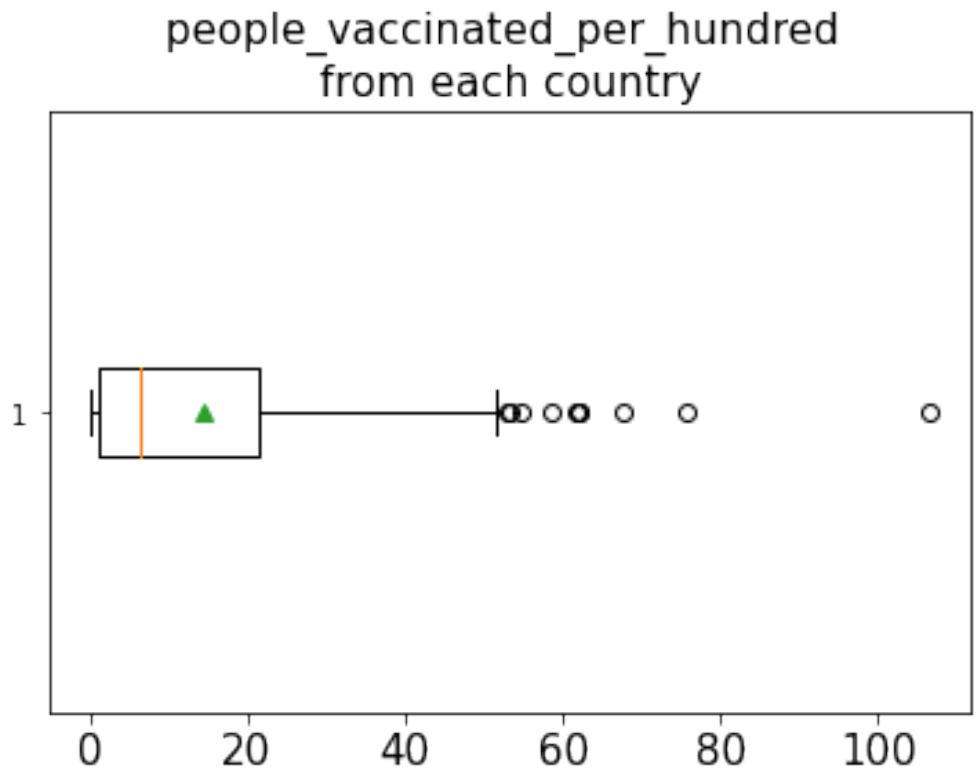
```
Size of data is: (13307, 13)
Missing values in dataset:
```

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

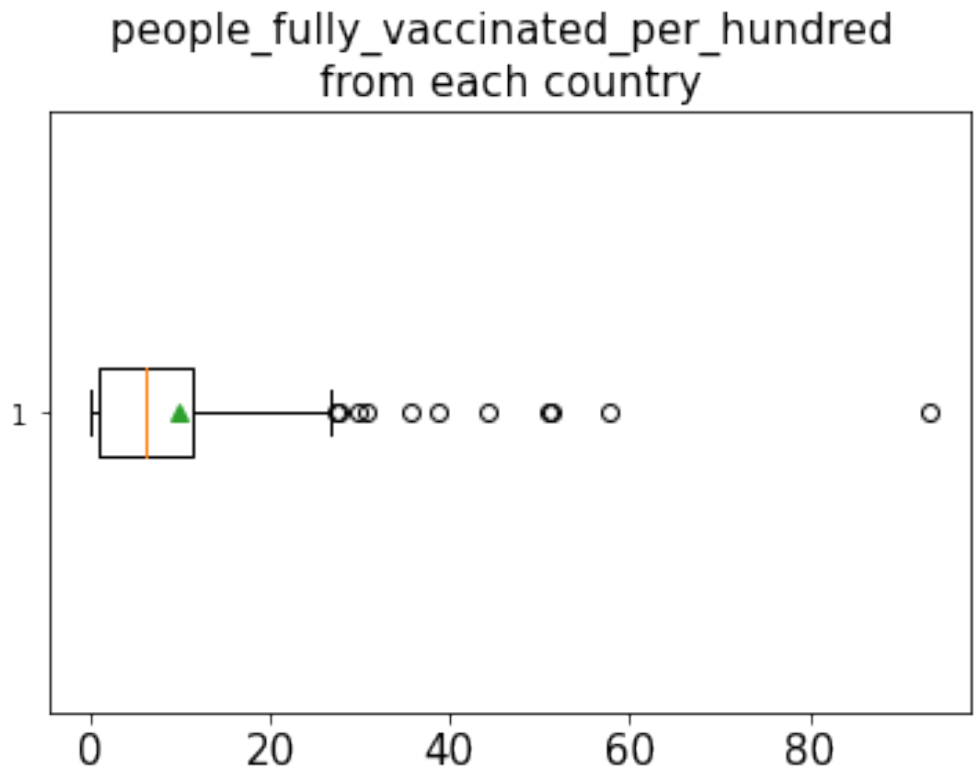
- część z brakujących wartości można łatwo uzupełnić, wykorzystując zależności między atrybutami
- przykładowo łączną sumę podanych dawek szczepionki można wyznaczyć prostą metodą, jeśli założymy liniową progresją pomiędzy odległymi wartościami
- trochę inaczej wygląda sytuacja w przypadku dziennej ilości podanych dawek. Według autorów zbioru danych jest to atrybut, który obliczają sami według procedury
    - Za pomocą interpolacji uzupełnia się brakujące wartości totalnej liczby wykonanych szczepień
    - Na podstawie różnicy z dniem poprzednim wylicza się ilość podanych w danym dniu dawek
    - Taka wyliczona liczba jest ostatecznie uśredniana z wartością jaka występuje w ciągu ostatniego tygodnia



Rysunek: Rozkład liczby wykonanych szczepień



Rysunek: Rozkład liczby zaszczepionych osób



Rysunek: Rozkład liczby w pełni zaszczepionych osób

Wnioski z EAD

- wady to duża rozbieżność ilości przykładów na państwo oraz duże braki w wartościach atrybutów
- rozkłady wartości atrybutów zapewniają tutaj dużo informacji i pokazują ogólny postęp programu szczepień
- dużą zaletą jest to, że dane są aktualizowane na bieżąco
- podsumowując, cele określone na początku wymagają odpowiedniego przygotowania danych, ale są możliwe do realizacji