Анализ датасета статистики по вакцинации от COVID-19

Хомутова Екатерина, 25.02.21

Научная руководительница:
Меликян Алиса Валерьевна

Данные

Датасет содержит статистику по вакцинации от COVID-19 в 80 странах и регионах.

Переменные:

country – название страны/региона;

date – число, за которое предоставляется статистика;

total_vaccinations – суммарное кол-во вакцинаций на данное число *(первая и вторая доза вакцины считается отдельно);*

people_vaccinated – кол-во людей, вакцинированных на данный момент;

people_fully_vaccinated – кол-во людей, прошедших полный курс вакцинации на данный момент

Данные

daily_vaccinations – вакцинаций за текущий день;

total_vaccinations_per_hundred – «*процент вакцинации*» – процент вакцинированных людей в стране на данный момент (*считаются все, кто получил хотя бы первую дозу вакцины*);

people_fully_vaccinated_per_hundred – кол-во полностью вакцинированных людей (*получивших 2 дозы вакцины*) на данный момент;

vaccines – названия вакцин;

source_name – учреждение или физическое лицо, предоставляющее данные о вакцинации в стране;

source_website – сайт, с которого была получена информация о вакцинации в стране;

Гипотезы

1. Страны - производители вакцин прививаются быстрее, чем остальные страны *(на момент последнего апдейта у стран-производителей будет больше привитых людей).*
2. Информация о вакцинации большинства стран предоставляются государственными источниками *(ожидается, что на 1 месте будут гос. источники, на втором – локальные новостные сайты).*
3. Чем больше людей прививается за день, тем больше процент полностью привитых людей.

Гипотеза 1: сбор недостающих данных

Сначала необходимо было узнать, где были произведены те или иные вакцины. Поиск уникальных названий по колонке «vaccines» и поиск в гугле, помогли узнать, что:

1. Sputnik V произведён в России,
2. Pfizer – в Германии,
3. Moderna – в США,
4. Sinovac, Sinopharm – в Китае,
5. AstraZeneca – в Великобритании,
6. Covaxin - в Индии.

Гипотеза 1: анализ данных из датасета

Страны с предыдущего слайда мы ожидаем увидеть в топе по количеству вакцинированных жителей в стране. *С моей точки зрения, наиболее отражающий это показатель – процент привитых жителей, и в исследуемом датасете существует такая колонка.* Но сначала необходимо отобрать самые свежие данные:

```
countries = vaccination['country'].unique()

last_update = vaccination.copy()

for country in countries:
    max_country_date = max(last_update[last_update['country'] == country]['date'])
    last_update.drop(np.where((last_update['country'] == country) & (last_update['date'] != max_country_date))[0], inplace=True)
    last_update = last_update.reset_index(drop=True)

last_update
```

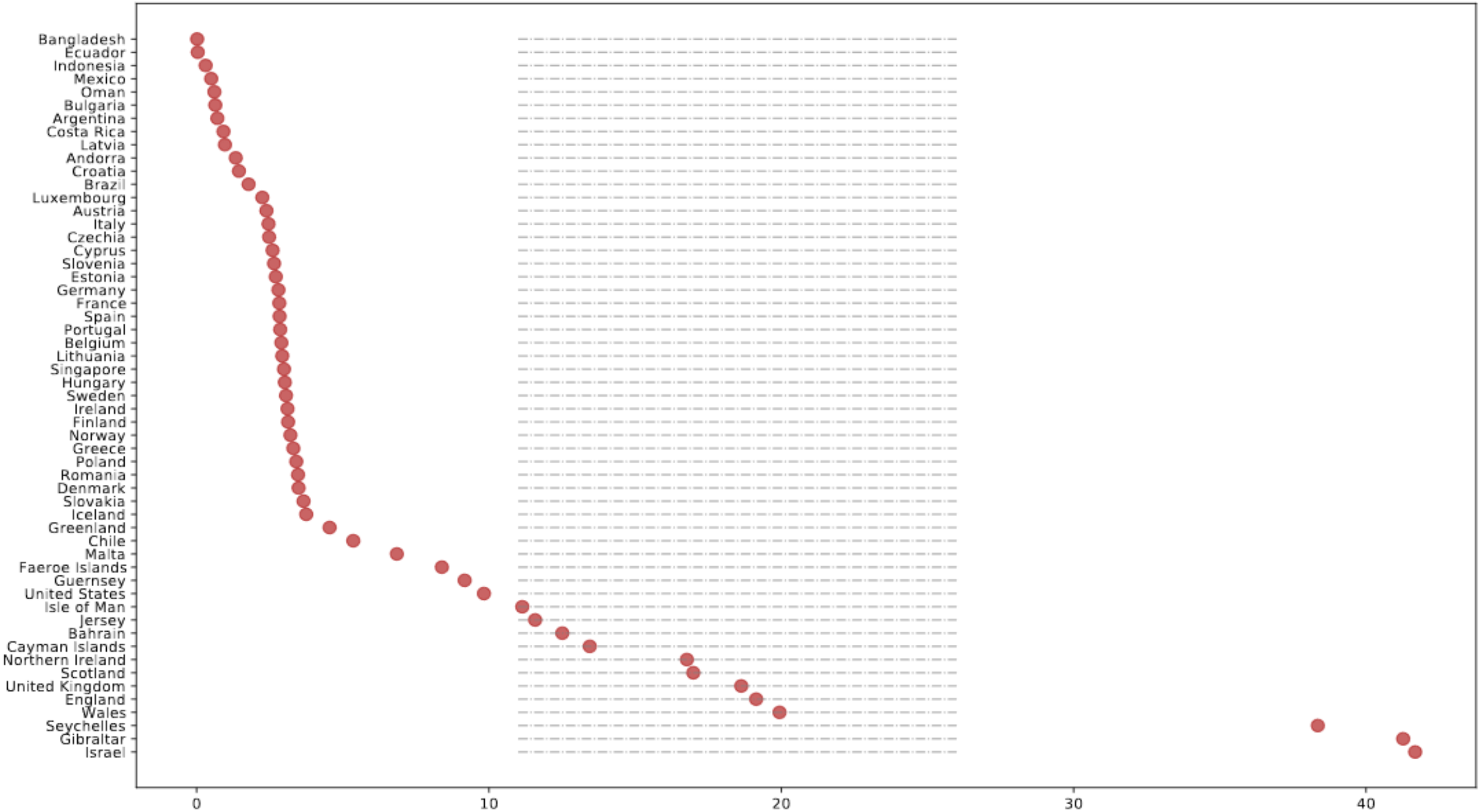
Гипотеза 1: график и выводы

(график большой, поэтому он на следующем слайде)

Сделав сводную таблицу и построив график, заметим, что в топе оказались Израиль, Гибралтар, Сейшелы (Сейшельские острова), Великобритания и её внутренние субъекты, Бахрейн(государство в средней Азии) и США.

Гипотеза 1 не подтвердилась.

country	people_vaccinated_per_hundred
Israel	41.68
Gibraltar	41.27
Seychelles	38.35
Wales	19.94
England	19.14
United Kingdom	18.63
Scotland	16.99
Northern Ireland	16.77
Cayman Islands	13.45
Bahrain	12.51
Jersey	11.58
Isle of Man	11.14
United States	9.83


Bangladesh
Ecuador
Indonesia
Mexico
Oman
Bulgaria
Argentina
Costa Rica
Latvia
Andorra
Croatia
Brazil
Luxembourg
Austria
Italy
Czechia
Cyprus
Slovenia
Estonia
Germany
France
Spain
Portugal
Belgium
Lithuania
Singapore
Hungary
Sweden
Ireland
Finland
Norway
Greece
Poland
Romania
Denmark
Slovakia
Iceland
Greenland
Chile
Malta
Faeroe Islands
Guernsey
United States
Isle of Man
Jersey
Bahrain
Cayman Islands
Northern Ireland
Scotland
United Kingdom
England
Wales
Seychelles
Gibraltar
Israel
0
10
20
30
40

Гипотеза 2: сбор и анализ данных

Для анализа информационных ресурсов был введён новый столбец численный «source_category», в котором с помощью числе от 0 до 3 ресурсы размечены по категориям:

0 – государственные сайты *(я считала за гос. сайт любой сайт, в домене которого присутствовала строка «gov»);*

1 – новостные сайты *(новостными сайтами я считаю сайты, в доменах которых присутствует срока «news», хотя это и достаточно сильное допущение);*

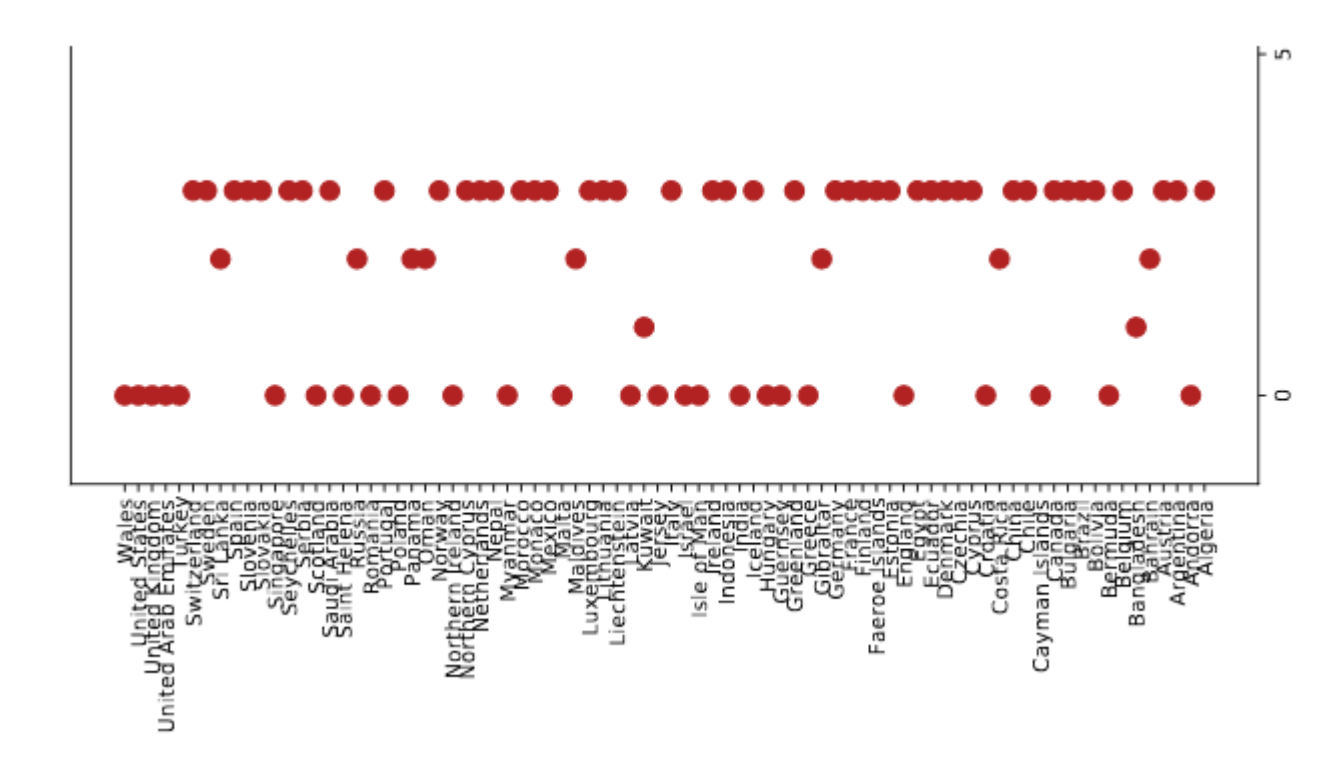
2 – twitter *(благодаря данному датасету я узнала, что у некоторых стран бывают официальные аккаунты в твиттере);*

3 – другие ресурсы.

Гипотеза 2: график и результаты

Из графика заметно, что большинство точек находятся на 0 и 3: гос. ресурсы и другие ресурсы.

Гипотеза 2 подтверждена.



Гипотеза 3: корреляционный анализ

Построив корреляционную таблицу по столбцам “people_fully_vaccinated” и “daily_vaccinations_raw” *(кол-во людей, прошедших полный курс вакцинации, и среднее кол-во вакцинаций в день соответственно),* заметим сильную корреляцию между данными параметрами.

Гипотеза 3 подтверждена.

	people_fully_vaccinated	daily_vaccinations_raw
people_fully_vaccinated	1.000000	0.834969
daily_vaccinations_raw	0.834969	1.000000

Выводы

1. Активность вакцинации населения страны напрямую не зависит от того, была ли в этой стране произведена собственная вакцина или вакцинация осуществляется импортной вакциной.
2. Большинство стран информирует о ходе вакцинации через официальные каналы, такие как государственные сайты.
3. Чем больше людей в день вакцинируется, тем больше людей будут вакцинированы обоими компонентами вакцины.