

Визуализация данных

ВЫПОЛНИЛ СТУДЕНТ 2 КУРСА
ГРУППЫ БПМИ187
ФАЗЛЕТДИНОВ ЭДУАРД

Открытые данные

- ▶ Источники российских открытых данных (data.gov, data.mos и др.) не содержат актуальной информации о коронавирусе, в то время как актуальность данной темы очевидна.
- ▶ Центр системных наук и инженерии (CSSE) при университете Джона Хопкинса ежедневно собирает данные из большого списка источников, представленных на официальной странице CSSE на github:
- ▶ <https://github.com/CSSEGISandData/COVID-19>
- ▶ И объединяет все открытые данные в единый файл удобного формата .csv

Библиотеки

Существует множество библиотек для анализа и обработки данных на языке Python. Из наиболее удобных и популярных стоит отметить:

- ▶ ***Pandas*** – библиотека для обработки и анализа данных (на основе NumPy)
- ▶ ***Matplotlib*** – библиотека для визуализации двухмерной графики
- ▶ ***Seaborn*** - более высокоуровневое API на базе библиотеки matplotlib
- ▶ ***Plotly*** – одна из самых актуальных и мощных библиотек для визуализации

Импорт библиотек и СЧИТЫВАНИЕ ДАННЫХ

Считывание данных происходит с помощью метода *pandas*

`read_csv()`

Метод `head()` выводит верхушку таблицы данных

```
full_table = pd.read_csv('../input/covid_19_data.csv', parse_dates=['ObservationDate'])
full_table.head()
```

]:

	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
0	1	2020-01-22	Anhui	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
1	2	2020-01-22	Beijing	Mainland China	1/22/2020 17:00	14.0	0.0	0.0
2	3	2020-01-22	Chongqing	Mainland China	1/22/2020 17:00	6.0	0.0	0.0
3	4	2020-01-22	Fujian	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
4	5	2020-01-22	Gansu	Mainland China	1/22/2020 17:00	0.0	0.0	0.0

```
import json
import random
from urllib.request import urlopen
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objs as go
import plotly.figure_factory as ff
import calmap
import folium

from pandas.plotting import register_matplotlib_converters
register_matplotlib_converters()

import warnings
warnings.filterwarnings('ignore')
```


Общая картина

- Сгруппируем данные методом `groupby()`, посчитаем общее количество подтверждённых (*confirmed*), умерших (*deaths*), выживших (*recovered*) и заражённых (*active*) на данный момент с помощью метода `sum()`.

```
temp = full_table.groupby('ObservationDate')['Confirmed', 'Deaths', 'Recovered', 'Active'].sum().reset_index()
temp = temp[temp['ObservationDate'] == max(temp['ObservationDate'])].reset_index(drop=True)
temp.style.background_gradient(cmap='Pastel1')
```

	ObservationDate	Confirmed	Deaths	Recovered	Active
0	2020-03-23 00:00:00	378287	16497	100958	260832

- По данным на 23.03.2020 болеют уже сотни тысяч человек.

- ▶ Данные можно отсортировать и провести сравнительный анализ. Наибольшее количество смертей/больных/выздоровевших покрашено в буро-красный цвет.

```
temp_f = full_latest_grouped.sort_values(by='Confirmed', ascending=False)
temp_f = temp_f.reset_index(drop=True)
temp_f.style.background_gradient(cmap='Reds')
```

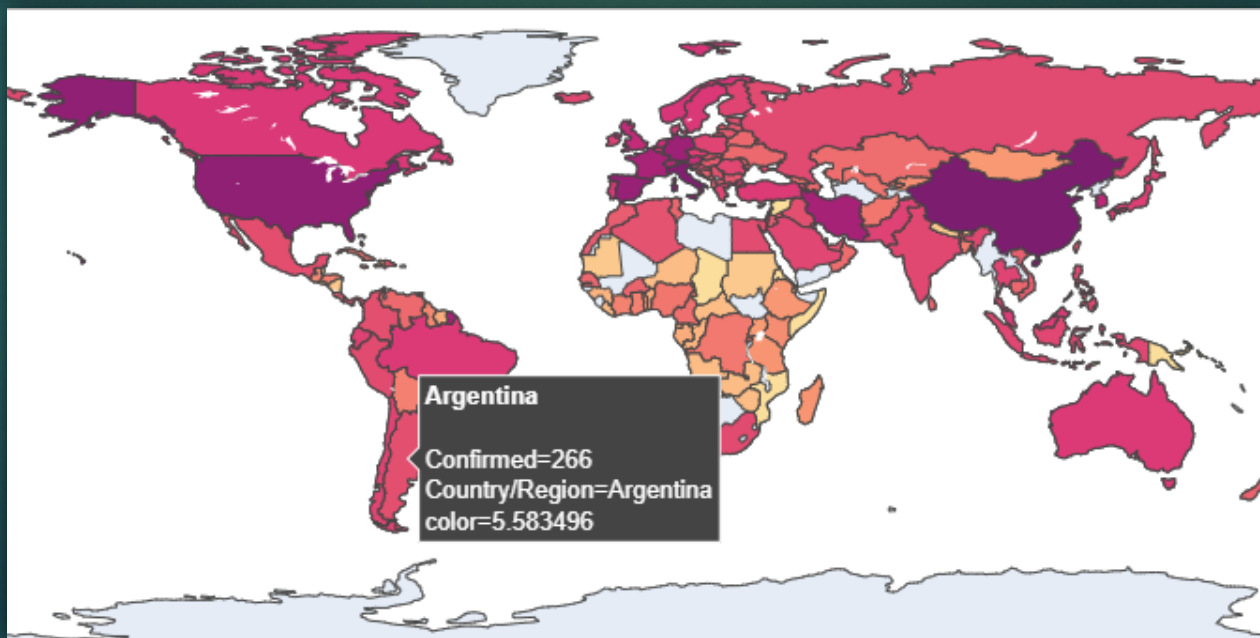
	Country/Region	Confirmed	Deaths	Recovered	Active
0	China	81116	3270	72709	5137
1	Italy	63927	6077	7432	50418
2	US	43667	552	0	43115
3	Spain	35136	2311	3355	29470
4	Germany	29056	123	453	28480
5	Iran	23049	1812	8376	12861
6	France	20123	862	2207	17054
7	South Korea	8961	111	3166	5684
8	Switzerland	8795	120	131	8544
9	UK	6726	336	140	6250
10	Netherlands	4764	214	3	4547
11	Austria	4474	21	9	4444

Большинство людей в Китае выздоровело.
Италия же стала эпицентром эпидемии. В США также зарегистрировано большое количество заражённых.

Раскрашивание карты

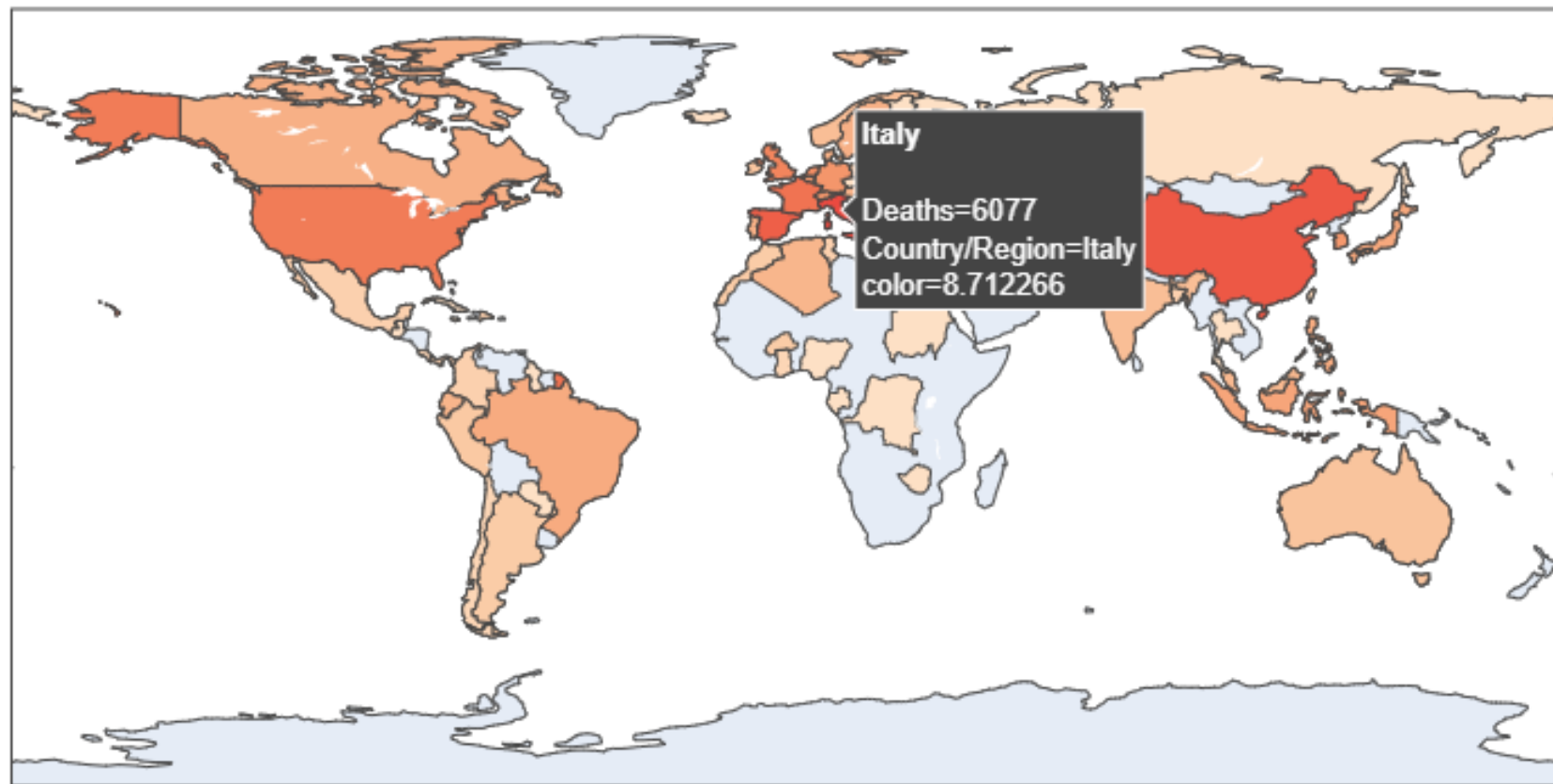
С помощью функции `choropleth()` библиотеки *plotly* «раскрасим» карту по количеству подтверждённых случаев заражения коронавирусом.

```
fig = px.choropleth(full_latest_grouped, locations="Country/Region",  
                    locationmode='country names', color=np.log(full_latest_grouped["Confirmed"]),  
                    hover_name="Country/Region", hover_data=['Confirmed'],  
                    color_continuous_scale="Sunsetdark", title='Страны с подтверждёнными случаями заражения')  
fig.update(layout_coloraxis_showscale=False)  
fig.show()
```



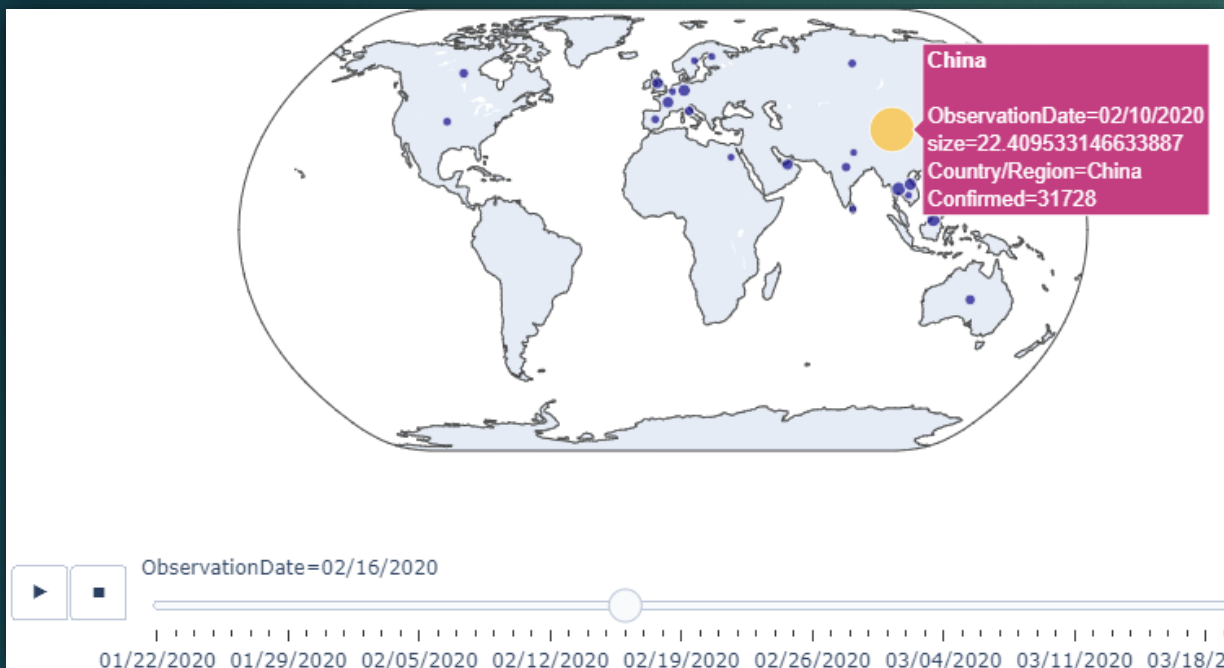
- ▶ Аналогично можно раскрасить карту стран по количеству смертей.

Страны с подтверждёнными смертями от коронавируса

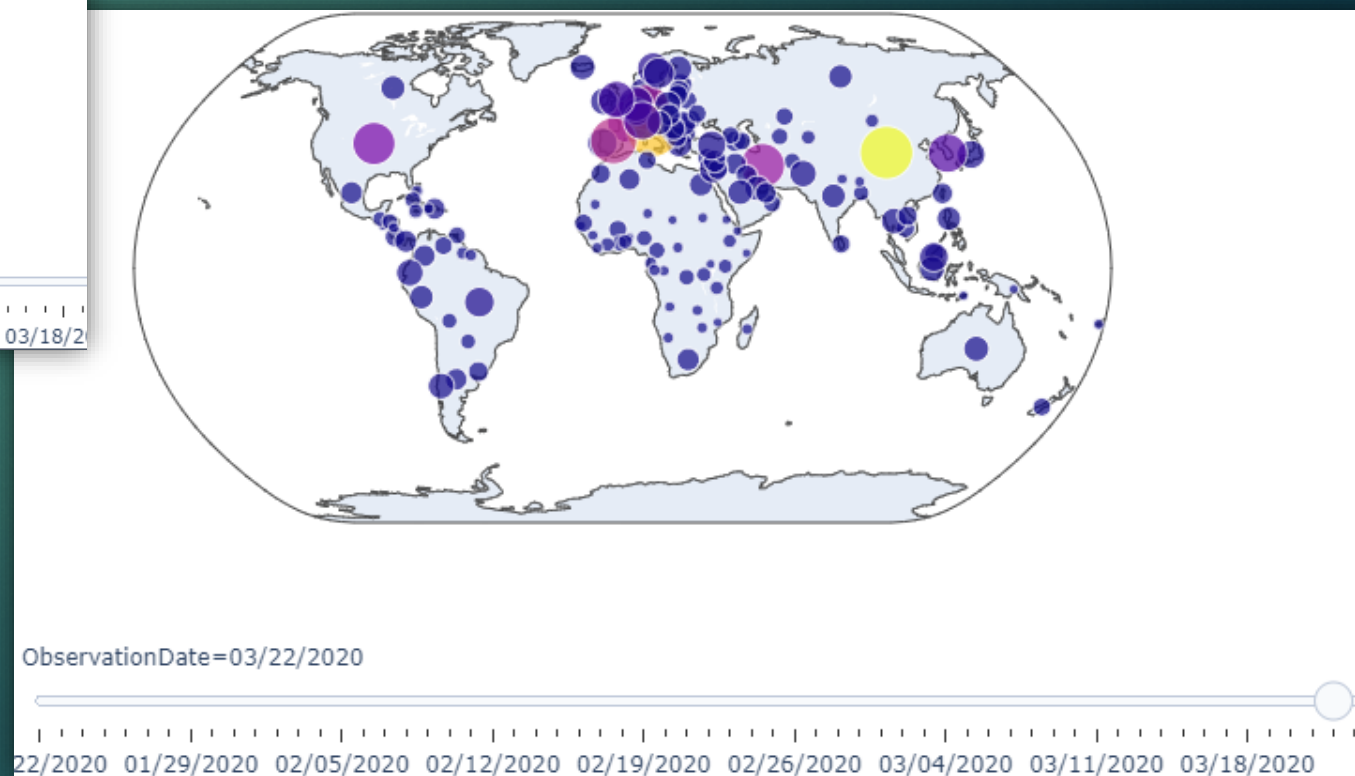


Интерактивные карты

Визуализация распространения коронавируса с течением времени с помощью функции `scatter_geo()`



Можно увидеть, что коронавирус распространился почти по всему миру в большом масштабе чуть больше чем за месяц.



Линейные графики

С помощью функции `area()` можно начертить заполненные линейные графики, которые покажут соотношение количества больных, вылечившихся и смертей в разные промежутку времени.

```
fig = px.area(temp, x="ObservationDate", y="Count", color='Case',  
              title='Коронавирус в течение всего времени', color_discrete_sequence = [rec, dth, act])  
fig.update_layout(xaxis_rangeflider_visible=True)  
fig.show()
```



- ▶ Рассмотрим количество заражённых и количество вылечившихся людей с момента первого заражения.
- ▶ Для этого воспользуемся функцией `line()`, которая чертит простые линейные графики.

```
fig = px.line(temp, x="ObservationDate", y="Value", color='Case', color_discrete_sequence=[dth, rec])  
fig.update_layout(xaxis_rangeslider_visible=True)  
fig.show()
```

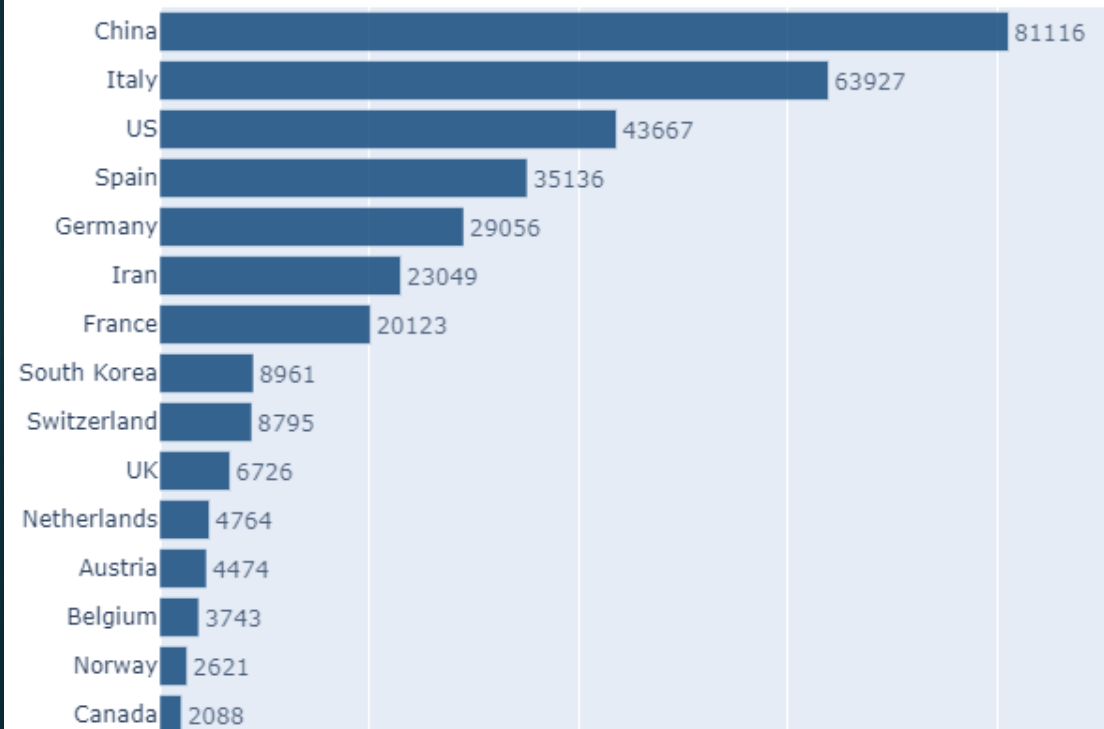


В начале марта пару дней количество вылечившихся людей превышало количество заражённых. Затем ситуация в Италии и США резко ухудшилась, поэтому график заражённых людей начал расти экспоненциально.

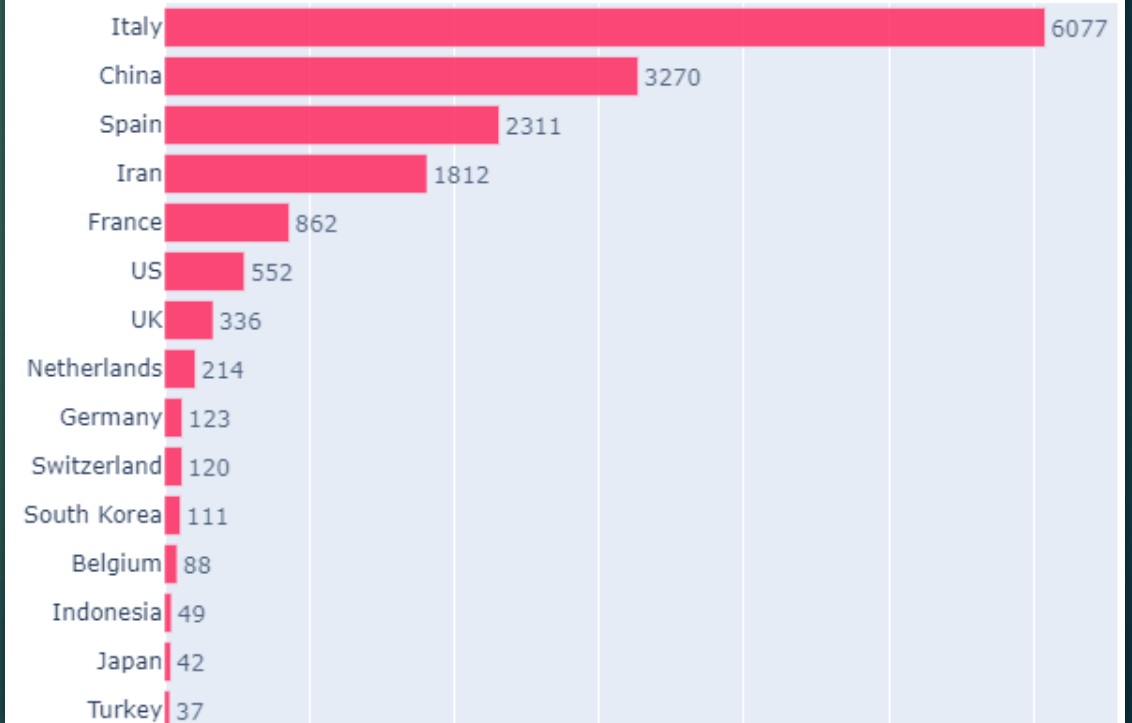
Гистограммы

- ▶ С помощью функции `bar()` сравним количество заражённых людей и смертей от коронавируса в разных странах. Для удобства отсортируем по убыванию.

Подтверждённые случаи заражения

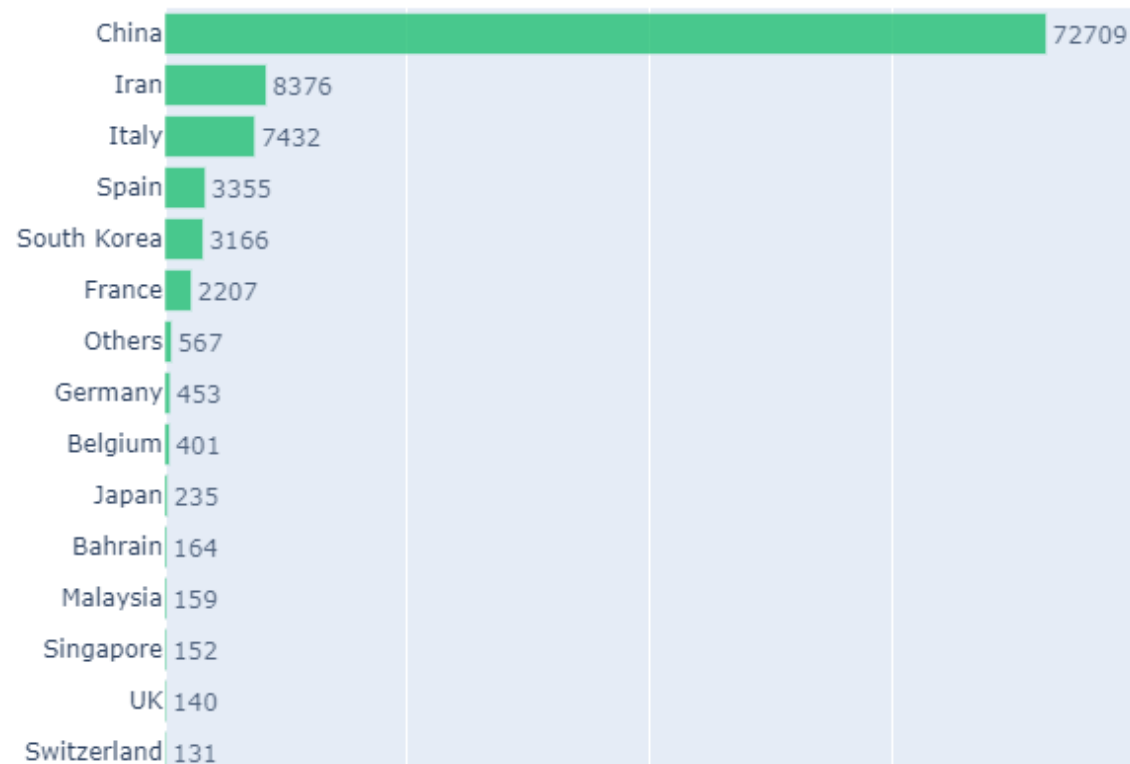


Смерти

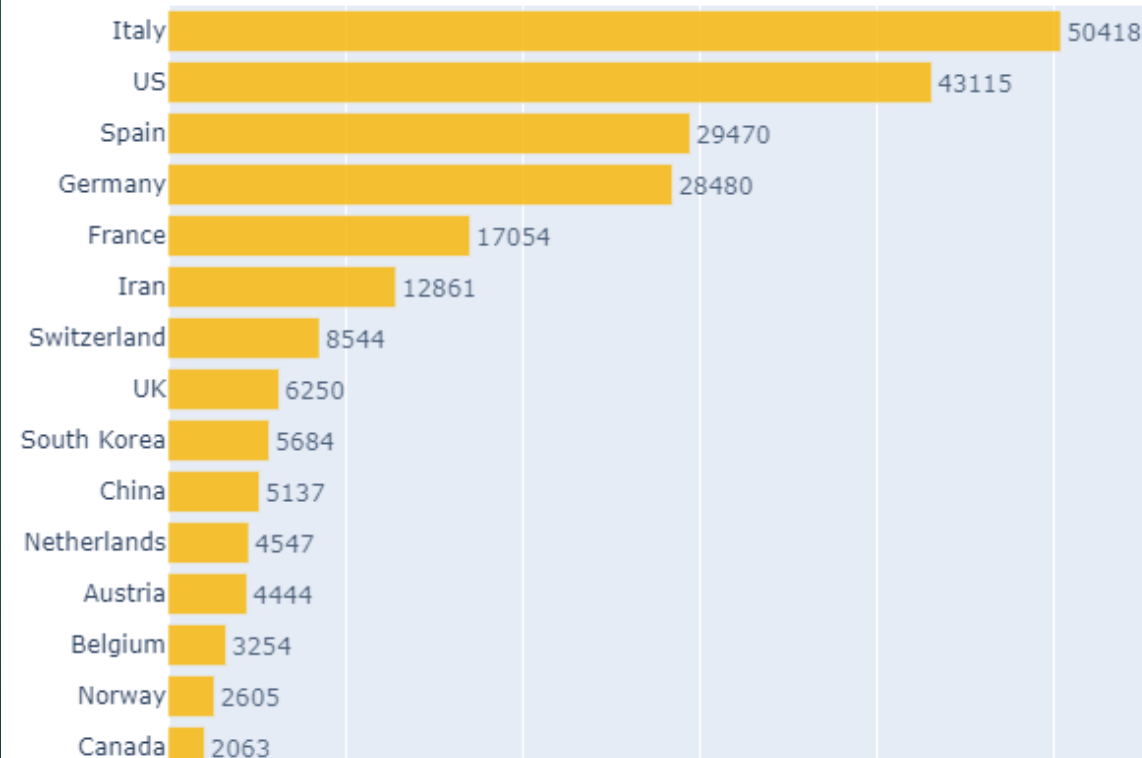



► Аналогично сравним количество вылечившихся людей, и количество людей, которые всё ещё болеют.

Вылечившиеся



Заражённые





► Как видно из столбчатых диаграмм на предыдущем слайде, в Китае приостановилось массовое заражение коронавирусом, а люди выздоравливают.

► Чего нельзя сказать о странах Европы и США, где на данный момент десятки тысяч заражённых людей.

Заключение

Для визуализации открытых данных были использованы наиболее удобные и самые популярные методы библиотеки *plotly*.

Они позволили взглянуть на данные и статистику со всех сторон.

А интерактивная карта дала возможность проследить за изменениями ситуации в течение всего времени пандемии.

СПИСОК ИСТОЧНИКОВ

- ▶ Открытые данные о коронавирусе центра системных наук и инженерии (CSSE):

<https://github.com/CSSEGISandData/COVID-19/>

- ▶ Plotly Documentation

<https://plotly.com/python/>

- ▶ Pandas Documentation

<https://pandas.pydata.org/docs/>



Спасибо за внимание!
Автор: Фазлетдинов Эдуард