Państwowa Wyższa Szkoła Zawodowa w Tarnowie
Katedra Informatyki

Covid daily cases by country

Praca semestralna

Wykonawca
Dominik Pająk

Prowadzący
mgr inż. Tomasz Potempa

Kierunek: Informatyka
Przedmiot: Big Data i Hurtownie Danych
Semestr: zimowy 2021/2022
Rok: III

Spis treści

1 Zwięzły opis badaego zbioru danych

Zbiór danych wykorzystany do wykonania zadania
https://www.kaggle.com/yamqwe/omicron-covid19-variant-daily-cases
Podany zbiór danych składa się z 100416 wierszy oraz 6 kolumn

1. **location** - Kraj, którego dotyczą informacje
2. **date** - Czas od poczatku COVID'a
3. **variant** - Wariant koronawirusa, którego dotyczą informacje
4. **num_sequences** - Dzienna liczba przypadków wariantu koronawirusa
5. **perc_sequences** - Procent przypadków wariantu w kontekście do wszystkich przypadków
6. **numsequencestotal** - Łączna liczba przypadków

2 Zastosowane metody/algorytmy

Użyte biblioteki Pythona

- **pandas**
  https://pandas.pydata.org/
- **plotly.express**
  https://plotly.com/python/plotly-express/
- **matplotlib**
  https://matplotlib.org/
- **seaborn**
  https://seaborn.pydata.org/
- **pandas_profiling**
  https://pandas-profiling.github.io/pandas-profiling/docs/master/index.html

Użyte algorytmy / rozwiązania

- Phik correlation

3 Syntetyczna analiza wyników

3.1 Wstępna analiza zbioru

Importowanie bibliotek oraz potrzebnych funkcji

```
import pandas as pd
import plotly.express as px
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
import pandas_profiling
```

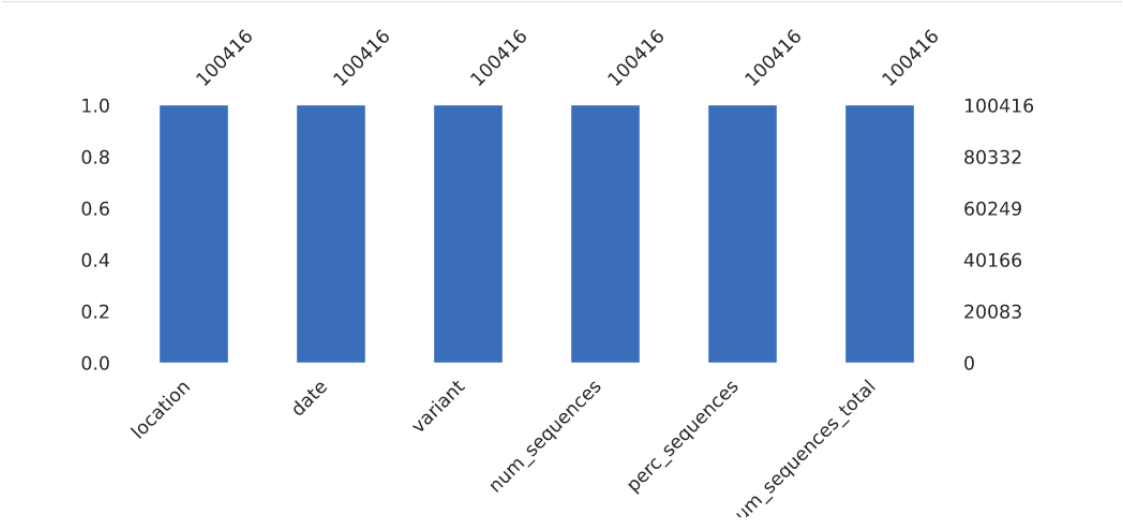
Importowanie zbioru danych

```
covid_data = pd.read_csv("./inputs/covid-variants.csv")
```

Analiza zbioru danych
Za pomocą biblioteki pandas_profiling zbieramy informacje o zbiorze danych

Liczba kolumn	6
Liczba wierszy	100416
Brakujące komórki	0
Brakujące komórki (%)	0.0%
Powtórzone wiersze	0
Powtórzone wiersze (%)	0.0%
Całkowity rozmiar w pamięci	4.6 MiB
Średni rozmiar pamięci dla komórki	48.0 B

Tablica 1: Statystyki zbioru danych obliczone za pomocą **pandas_profiling**



Rysunek 1: Informacja o brakujących danych obliczona za pomocą **pandas_profiling**

L.P	location	date	variant	num_sequences	perc_sequences	num_sequences_total
0	Angola	2020-07-06	Alpha	0	0.0	3
1	Angola	2020-07-06	B.1.1.277	0	0.0	3
2	Angola	2020-07-06	B.1.1.302	0	0.0	3
3	Angola	2020-07-06	B.1.1.519	0	0.0	3
4	Angola	2020-07-06	B.1.160	0	0.0	3
5	Angola	2020-07-06	B.1.177	0	0.0	3
6	Angola	2020-07-06	B.1.221	0	0.0	3
7	Angola	2020-07-06	B.1.258	0	0.0	3
8	Angola	2020-07-06	B.1.367	0	0.0	3
9	Angola	2020-07-06	B.1.620	0	0.0	3

Tablica 2: Pierwsze rekordy dla zbioru danych **covid-variants**

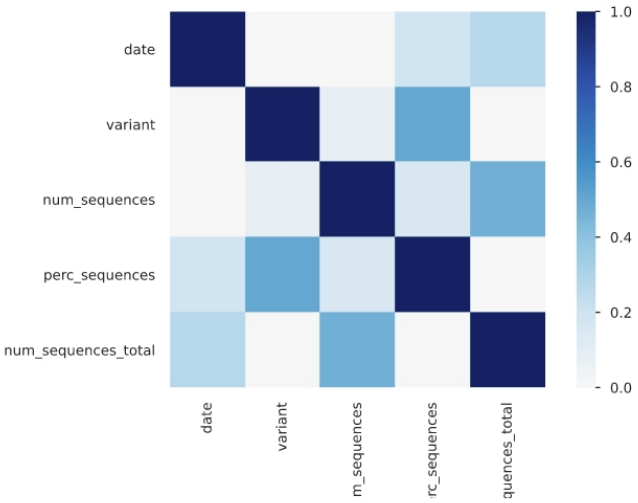
Następnie za pomocą własnego algorytmu analizujemy jeszcze typy danych w komórkach

```
def analysis():
    if len(covid_data.select_dtypes("object").columns) > 0:
        print("Object Variables:", "\n",
              len(covid_data.select_dtypes("object").columns), "\n",
              covid_data.select_dtypes("object").columns.tolist(), "\n")

    if len(covid_data.select_dtypes("integer").columns) > 0:
        print("Integer Variables:", "\n",
              len(covid_data.select_dtypes("integer").columns), "\n",
              covid_data.select_dtypes("integer").columns.tolist(), "\n")

    if len(covid_data.select_dtypes("float").columns) > 0:
        print("Float Variables:", "\n",
              len(covid_data.select_dtypes("float").columns), "\n",
              covid_data.select_dtypes("float").columns.tolist(), "\n")

    if len(covid_data.select_dtypes("bool").columns) > 0:
        print("Bool Variables:", "\n",
              len(covid_data.select_dtypes("bool").columns), "\n",
              covid_data.select_dtypes("bool").columns.tolist(), "\n")
```

Typ zmiennej	**Ilość wystąpień**	**Pole**
Object Variables	3	['location', 'date', 'variant']
Integer Variables	2	['num_sequences', 'num_sequences_total']
Float Variables	1	['perc_sequences']

Tablica 3: Typy zmiennych w zbiorze danych

Korelację pomiędzy kolumnami obliczymy za pomocą algorytmu Phika. Prezentuje się ona następująco:

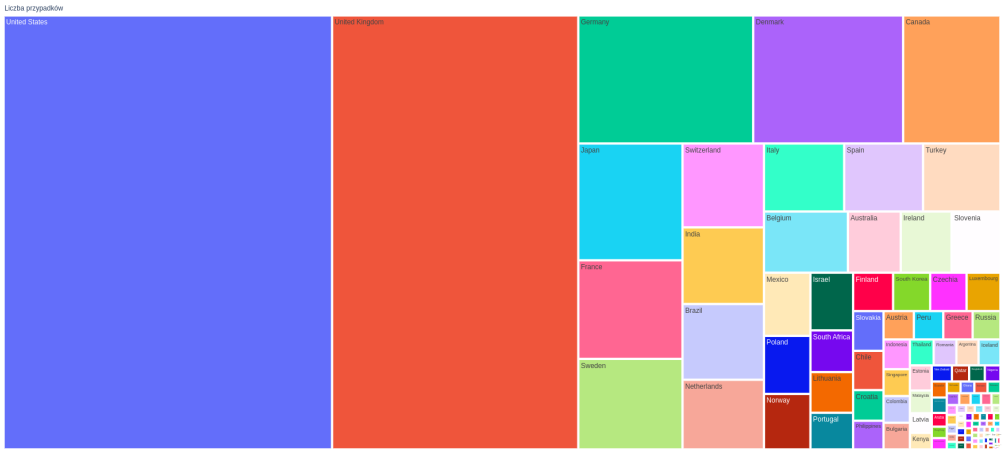


Rysunek 2: Korelacja Phika dla naszego zbioru danych

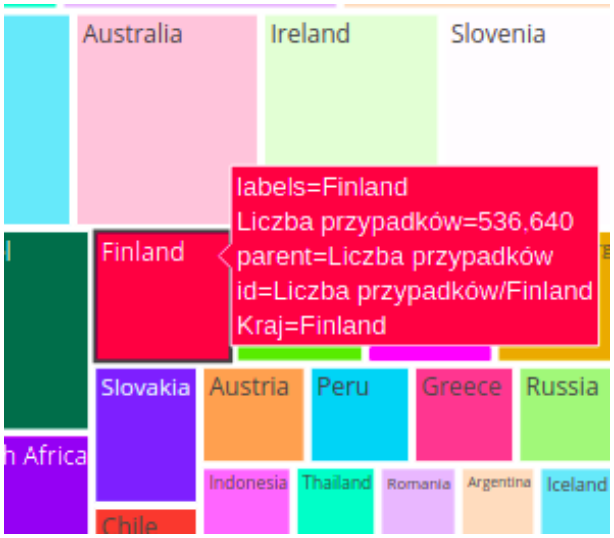
3.2 Analiza danych

Na podstawie danych zawartych w zbiorze danych możemy przeprowadzić wiele analiz. Zaczniemy od wyświetlenia krajów dominujących w zakresie wszystkich łącznych zakażeń Covid-19, na podstawie posiadanych danych.
Wykorzystana do tego zostanie poniższa funkcja

```
# Mapa najbardziej zainfekowanych krajów (przez Covid-19)
def generateCountryInfectionTreemap():
    sample = covid_data.rename(
        columns={"location": "Kraj", "num_sequences_total": "Liczba przypadków"})
    fig = px.treemap(sample, path=[px.Constant('Liczba przypadków'), 'Kraj'],
                     values='Liczba przypadków', hover_data=['Kraj'], )
    fig.show()
```

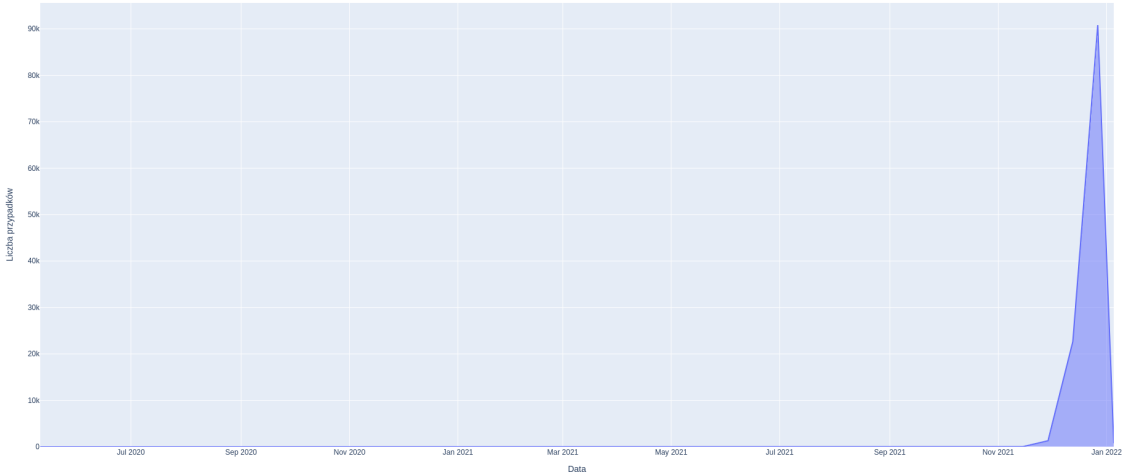


Rysunek 3: Rezultat działania powyższej funkcji. Przedstawia ranking państw z największą łączną liczbą zakażeń



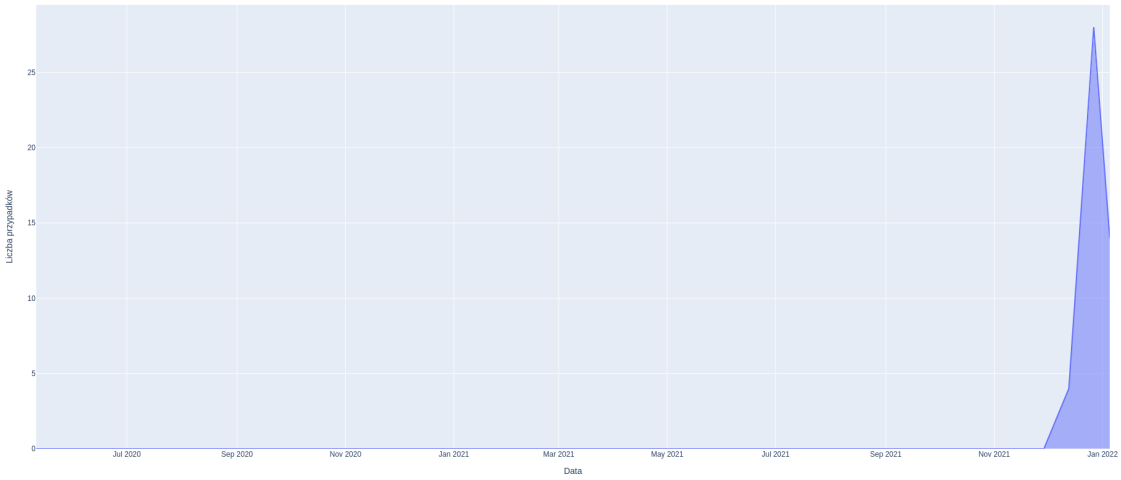
Rysunek 4: Po najechaniu na pojedynczy kafelek z państwem ukaże nam się łączna liczba zakażeń dla danego kraju

Ostatnio głośno mówi się o Omikronie - nowym wariancie Covid-19, który dosyć szybko rozprzestrzenia się po całym świecie.
Zobaczmy jak rozprzestrzenianie się choroby wygląda na wykresie



Rysunek 5: Występowanie wariantu Omicron w kontekście całego świata

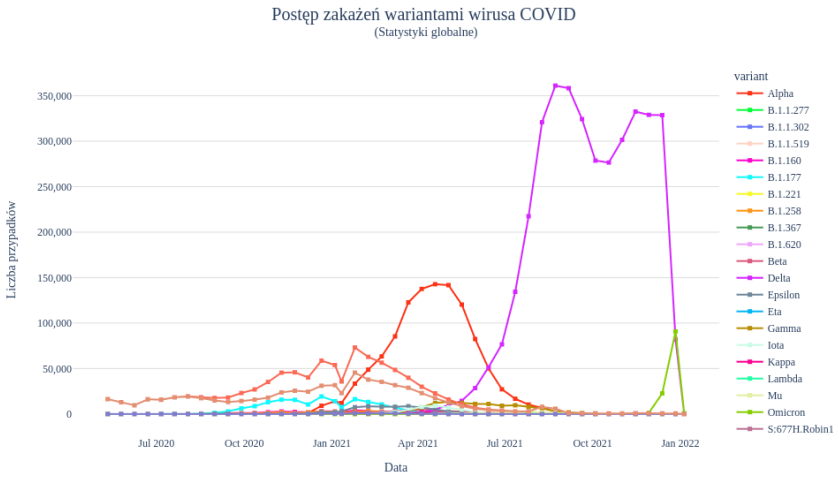
```
def generateOmicronVariantWorld():
    sample = covid_data.loc[covid_data['variant'] == 'Omicron'].groupby('date')[
        'num_sequences'].agg('sum')
    dataframe = pd.DataFrame({'Data': sample.index, 'Liczba przypadków': sample.values})
    fig = px.area(dataframe, y="Liczba przypadków", x='Data')
    fig.show()
```



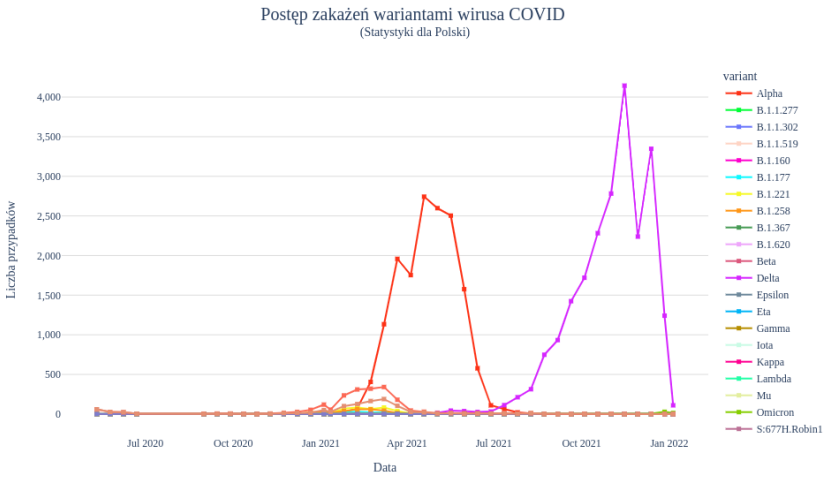
Rysunek 6: Występowanie wariantu Omicron w Polsce

```
def generateOmicronVariantPoland():
    sample = covid_data.loc[covid_data['variant'] == 'Omicron'].loc[
        covid_data['location'] == 'Poland'].groupby('date')['num_sequences'].agg('sum')
    dataframe = pd.DataFrame({'Data': sample.index, 'Liczba przypadków': sample.values})
    fig = px.area(dataframe, y="Liczba przypadków", x='Data')
    fig.show()
```

Na interaktywnym wykresie możemy zauważyć, że pierwsze przypadki Omicrona na świecie zostały zauważone jeszcze w październiku. W Polsce pierwszy przypadek miał miejsce 13 grudnia. Warto tutaj zaznaczyć, że powyższe wykresy uwzględniają tylko dane do końca grudnia 2022. Niestety w tym zbiorze, nie mamy poglądu na to jak obecnie wygląda sytuacja. Możemy natomiast, zobaczyć jak wygląda początek rozwoju Omicrona w stosunku do innych wariantów



Rysunek 7: Warianty koronawisusa na świecie (okres/ilość zarażeń)



Rysunek 8: Warianty koronawisusa w Polsce (okres/ilość zarażeń)

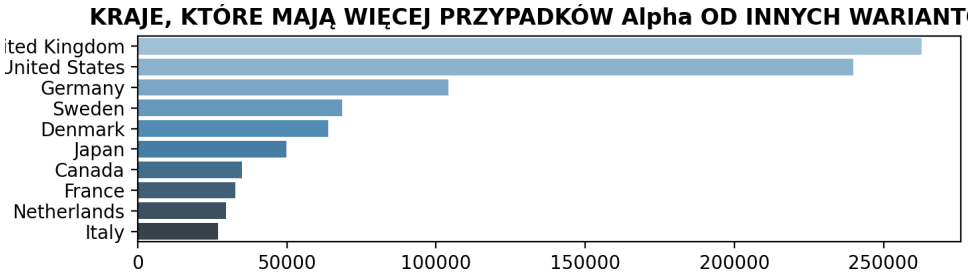
Na powyższych wykresach już na pierwszy rzut oka widać, że Omicron ma najszybciej rosnącą tendencję. Pod koniec grudnia ilość zakażeń na świecie wynosiła ponad 90 tysięcy.

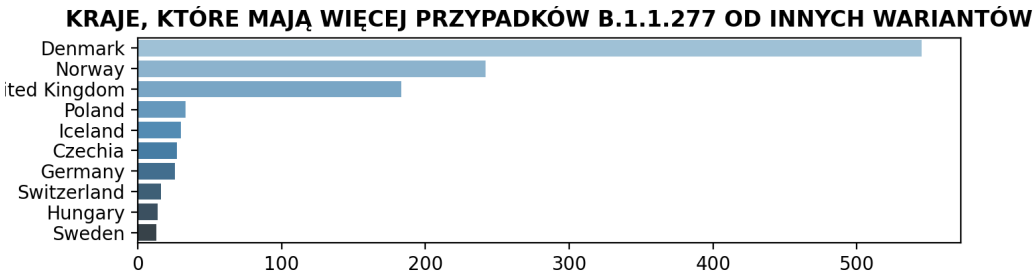
```
def progressionOfCovidVariantsDiagramWorld():
    variant_totals = covid_data.groupby([
        "variant", "date"], as_index=False).sum()
    progressionOfCovidVariantsDiagram(
        variant_totals=variant_totals, plottitle='Statystyki globalne')


def progressionOfCovidVariantsDiagramPoland():
    variant_totals = covid_data.loc[
        covid_data['location'] == 'Poland'].groupby([
        "variant", "date"],as_index=False).sum()
    progressionOfCovidVariantsDiagram(
        variant_totals=variant_totals, plottitle='Statystyki dla Polski')


def progressionOfCovidVariantsDiagram(variant_totals, plottitle):
    variant_totals["perc_sequences"] = (
            100 * variant_totals["num_sequences"] / variant_totals["num_sequences_total"]
    )

    fig = px.line(
        variant_totals,
        x="date",
        y="num_sequences",
        color="variant",
        color_discrete_sequence=px.colors.qualitative.Light24,
        custom_data=["perc_sequences"],
        markers=True,
        height=600,
        width=960,
        title="Postęp zakażeń wariantami wirusa COVID <br><sup>({})<sup>".format(plottitle),
    )
    fig.update_layout(
        font_family="serif", plot_bgcolor="#fff", title_font_size=20, title_x=0.5
```

```
    )
    fig.update_traces(
        hovertemplate="<i>%{x}:</i> <b>%{y}</b> (%{customdata[0]:.2f}%)",
        marker=dict(symbol="square", size=5)
    )
    fig.update_yaxes(
        fixedrange=True, gridcolor="#ddd", tickformat=",", title="Liczba przypadków"
    )
    fig.update_xaxes(fixedrange=True, title="Data")
    fig.show()
```

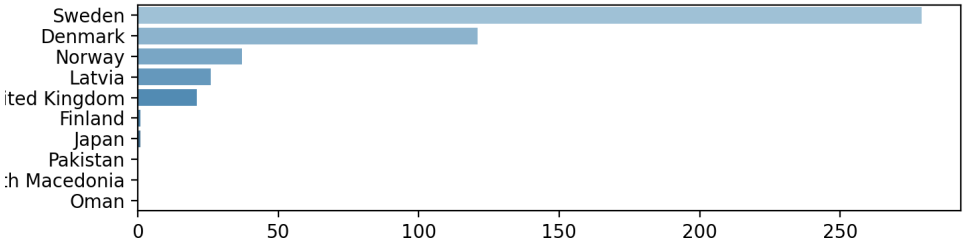
Na zakończenie przygotowałem jeszcze zestawienie wszystkich wariantów koronawirusa z podziałem na ich popularność

```
def generateMoreThanOthersDiagram():
    for virus in covid_data.variant.unique():
        dataframe = covid_data.loc[covid_data['variant'] == virus].groupby('location')[
                        'num_sequences'].agg('sum').sort_values(ascending=False)[:10]
        dataframe = pd.DataFrame(
            {'Kraj': dataframe.index, 'Liczba przypadków': dataframe.values})
        plt.figure(figsize=(8, 2), dpi=200)
        sns.barplot(y='Kraj', x="Liczba przypadków",
                    data=dataframe, palette="Blues_d")
        plt.title(
            'KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW {} OD INNYCH WARIANTÓW'.format(virus),
            loc='center',
                  fontweight="bold")
        plt.savefig('./outputs/variants/{}.png'.format(virus))
```

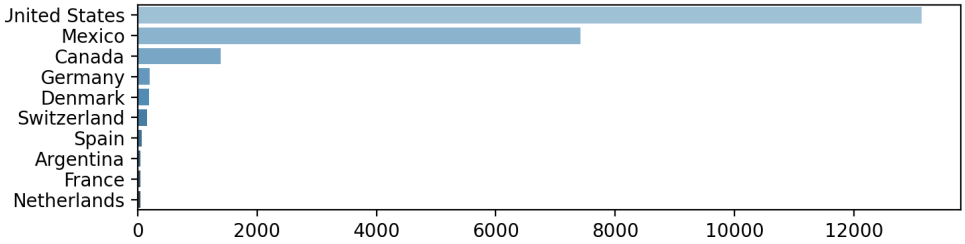




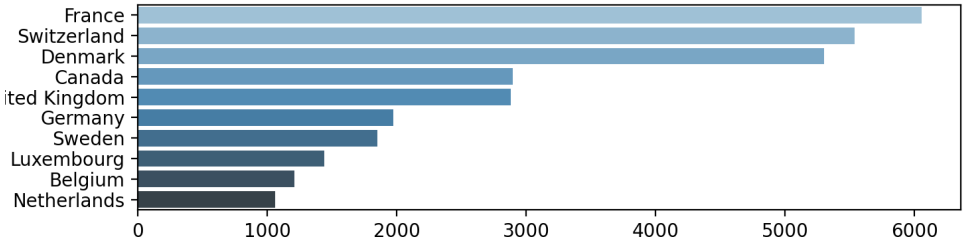
KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW B.1.1.302 OD INNYCH WARIANTÓW



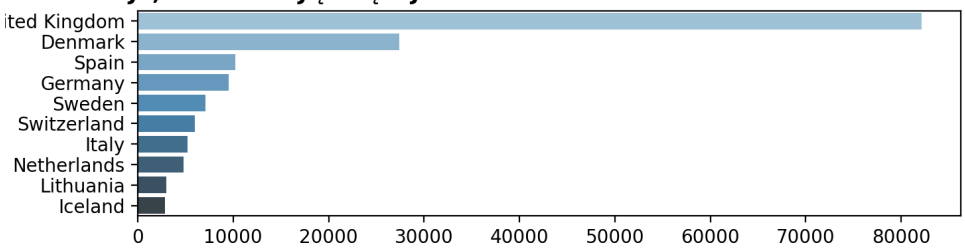
KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW B.1.1.519 OD INNYCH WARIANTÓW



KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW B.1.160 OD INNYCH WARIANTÓW

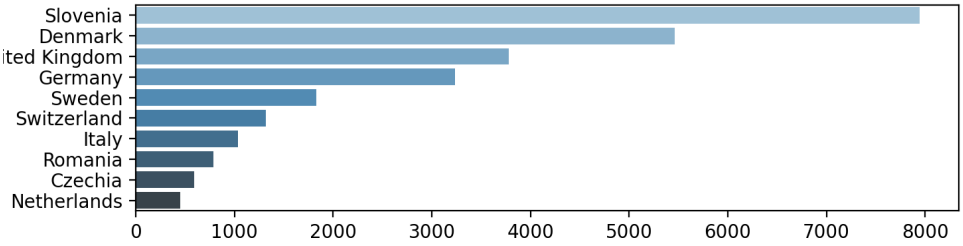


KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW B.1.177 OD INNYCH WARIANTÓW

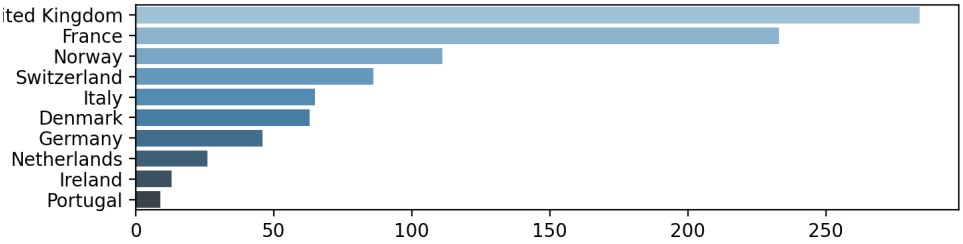


KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW B.1.221 OD INNYCH WARIANTÓW

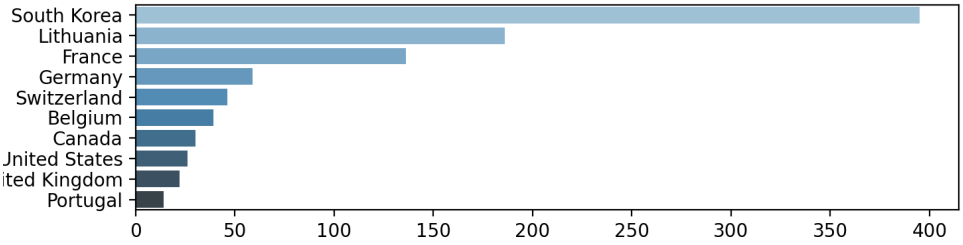
KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW B.1.258 OD INNYCH WARIANTÓW



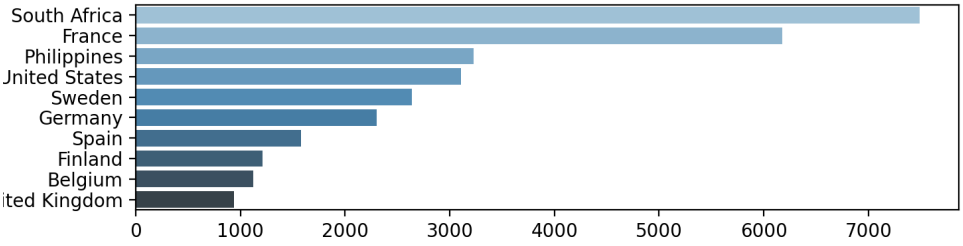
KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW B.1.367 OD INNYCH WARIANTÓW



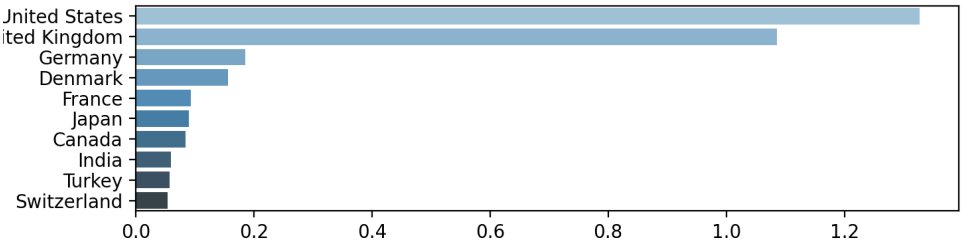
KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW B.1.620 OD INNYCH WARIANTÓW

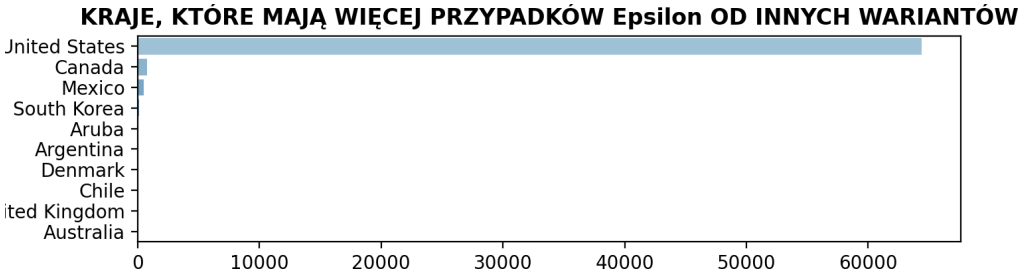


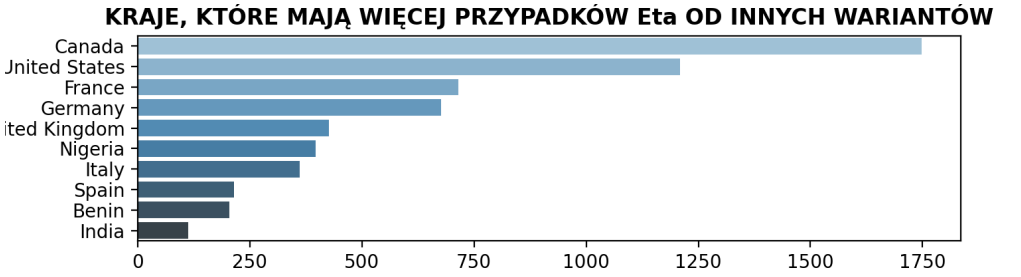
KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW Beta OD INNYCH WARIANTÓW

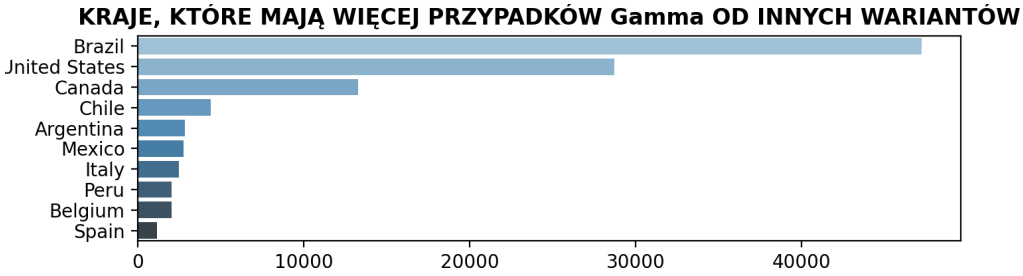


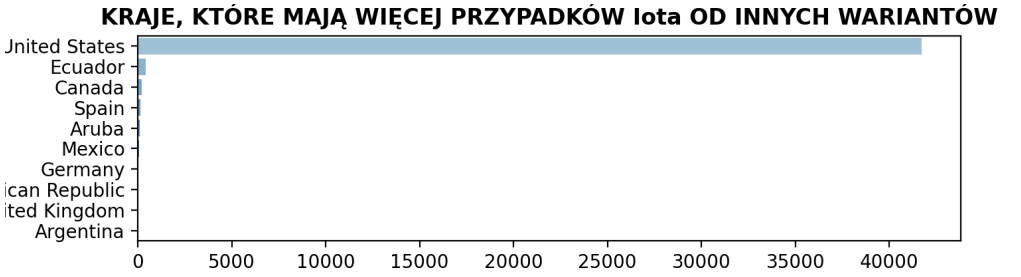
KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW Delta OD INNYCH WARIANTÓW

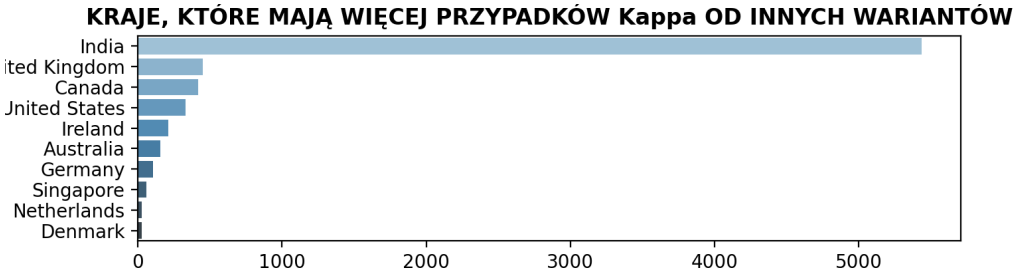



KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW Epsilon OD INNYCH WARIANTÓW
Jnited States
Canada
Mexico
South Korea
Aruba
Argentina
Denmark
Chile
ited Kingdom
Australia
0
10000
20000
30000
40000
50000
60000

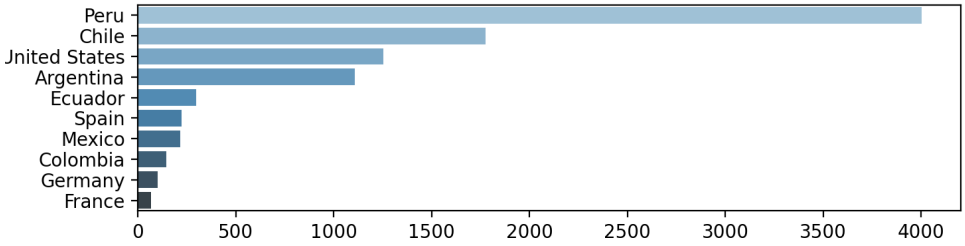

KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW Eta OD INNYCH WARIANTÓW
Canada
Jnited States
France
Germany
ited Kingdom
Nigeria
Italy
Spain
Benin
India
0
250
500
750
1000
1250
1500
1750


KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW Gamma OD INNYCH WARIANTÓW
Brazil
Jnited States
Canada
Chile
Argentina
Mexico
Italy
Peru
Belgium
Spain
0
10000
20000
30000
40000

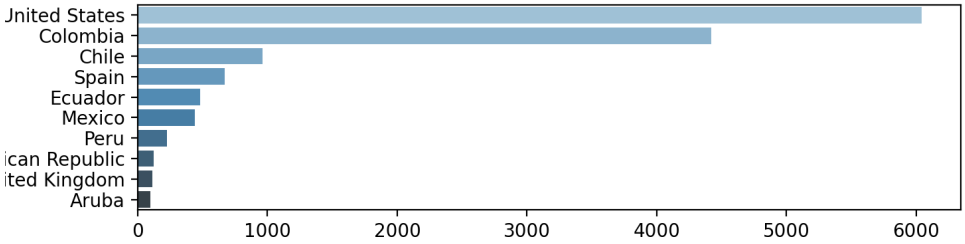

KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW Iota OD INNYCH WARIANTÓW
Jnited States
Ecuador
Canada
Spain
Aruba
Mexico
Germany
ican Republic
ited Kingdom
Argentina
0
5000
10000
15000
20000
25000
30000
35000
40000


KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW Kappa OD INNYCH WARIANTÓW
India
ited Kingdom
Canada
Jnited States
Ireland
Australia
Germany
Singapore
Netherlands
Denmark
0
1000
2000
3000
4000
5000

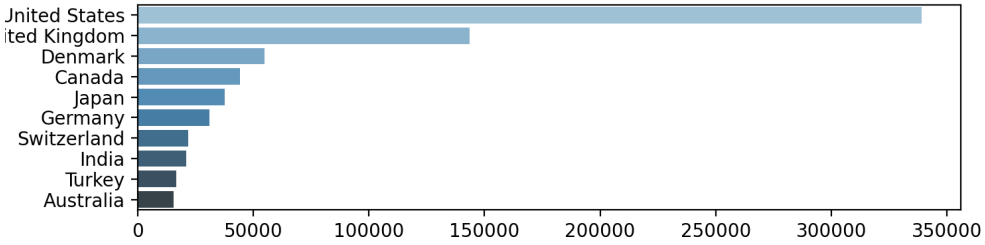
KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW Lambda OD INNYCH WARIANTÓW



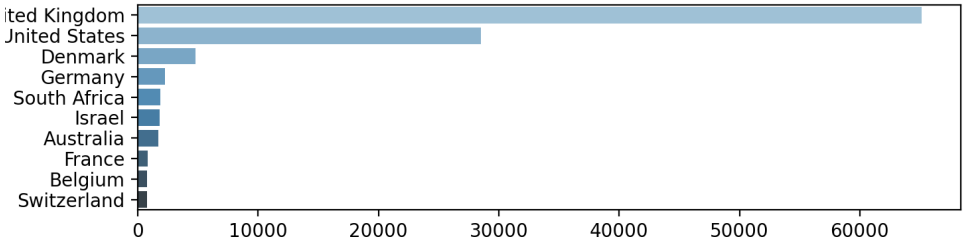
KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW Mu OD INNYCH WARIANTÓW



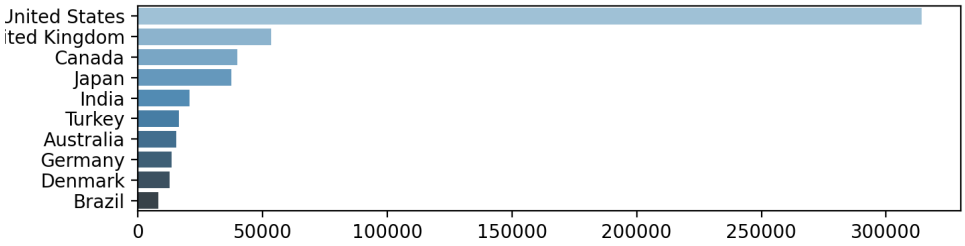
KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW non_who OD INNYCH WARIANTÓW



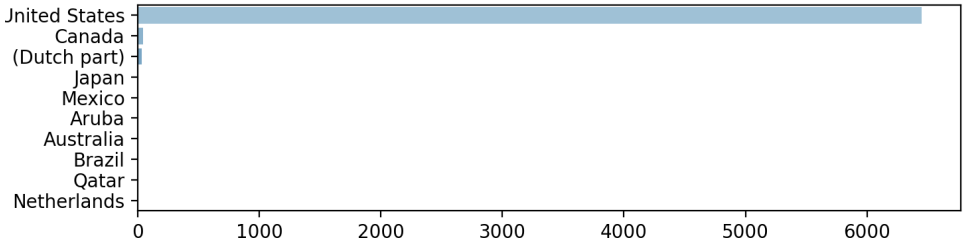
KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW Omicron OD INNYCH WARIANTÓW



KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW others OD INNYCH WARIANTÓW



KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW S:677H.Robin1 OD INNYCH WARIANTÓW



KRAJE, KTÓRE MAJĄ WIĘCEJ PRZYPADKÓW S:677P.Pelican OD INNYCH WARIANTÓW