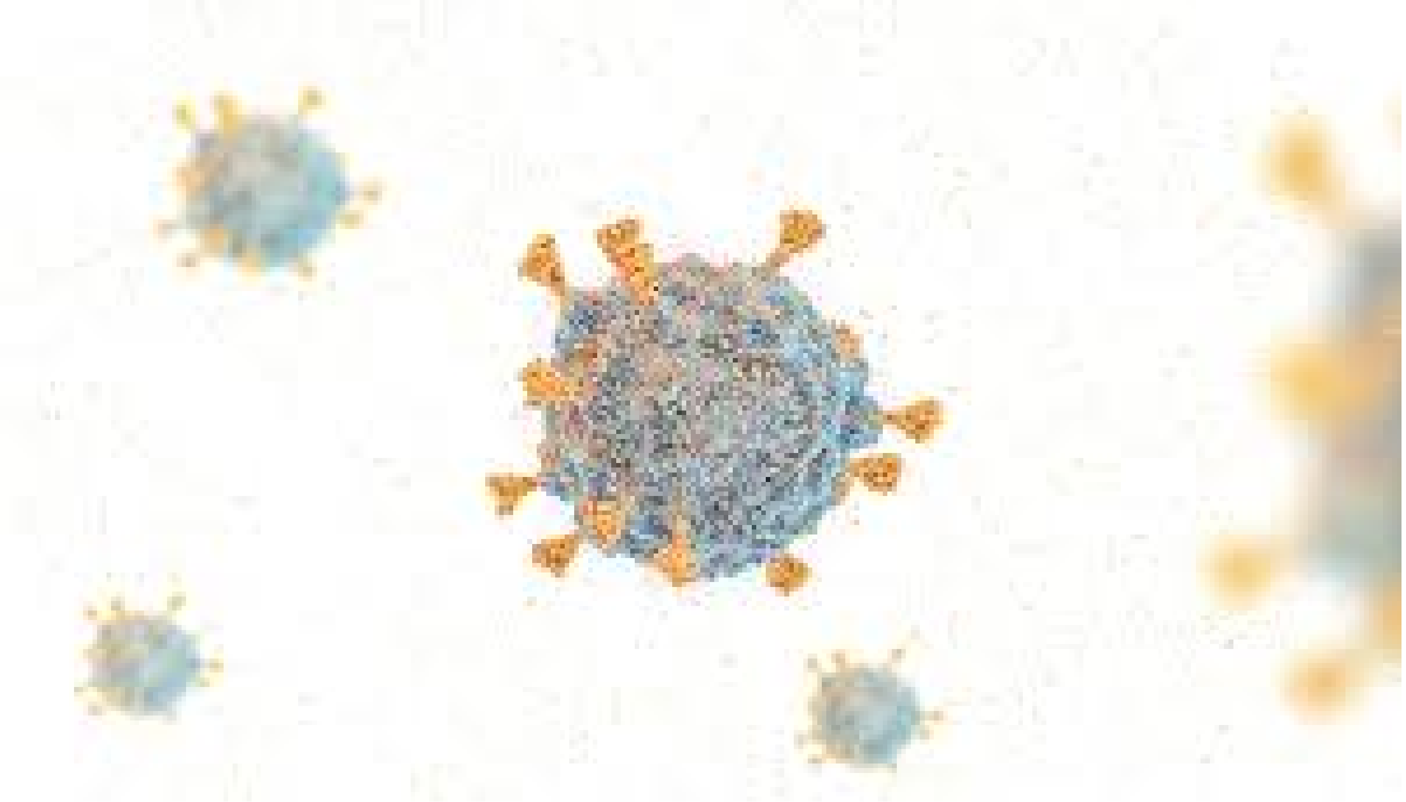
the Dataset was gotten from Kaggle on Omicron daily cases by country (COVID-19 variant) the last update and download for this execerise was on 03/02/2022



In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

In [2]:

```
cvd=pd.read_csv('covid-variants.csv')
```

In [49]:

```
cvd.head()
```

Out[49]:

	location	variant	num_sequences	perc_sequences	num_sequences_total	month	year	day
0	Angola	Alpha	0	0.0	3	7	2020	6
1	Angola	B.1.1.277	0	0.0	3	7	2020	6
2	Angola	B.1.1.302	0	0.0	3	7	2020	6
3	Angola	B.1.1.519	0	0.0	3	7	2020	6
4	Angola	B.1.160	0	0.0	3	7	2020	6

In [22]:

```
cvd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100416 entries, 0 to 100415
Data columns (total 6 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   location             100416 non-null  object
 1   date                 100416 non-null  datetime64[ns]
 2   variant              100416 non-null  object
 3   num_sequences        100416 non-null  int64
 4   perc_sequences       100416 non-null  float64
 5   num_sequences_total  100416 non-null  int64
dtypes: datetime64[ns](1), float64(1), int64(2), object(2)
memory usage: 4.6+ MB
```

In [5]:

```
cvd.variant.value_counts()
```

Out[5]:

```
Alpha               4184
B.1.1.277           4184
others              4184
S:677P.Pelican      4184
S:677H.Robin1       4184
Omicron             4184
Mu                  4184
Lambda              4184
Kappa               4184
Iota                4184
Gamma               4184
Eta                 4184
```

```
Epsilon           4184
Delta             4184
Beta              4184
B.1.620           4184
B.1.367           4184
B.1.258           4184
B.1.221           4184
B.1.177           4184
B.1.160           4184
B.1.1.519         4184
B.1.1.302         4184
non_who           4184
Name: variant, dtype: int64
```

In [6]:

```
cvd.num_sequences.value_counts()
```

Out[6]:

```
0       84173
1        2753
2        1405
3         905
4         631
        ...  
1690        1
1719        1
2156        1
1184        1
862         1
Name: num_sequences, Length: 1563, dtype: int64
```

In [7]:

```
cvd.isnull().any()
```

Out[7]:

```
location               False
date                   False
variant                False
num_sequences          False
perc_sequences         False
num_sequences_total    False
dtype: bool
```

In [8]:

```
cvd.describe()
```

Out[8]:

	num_sequences	perc_sequences	num_sequences_total
count	100416.000000	100416.000000	100416.000000
mean	72.171676	6.154355	1509.582457
std	1669.262169	21.898989	8445.291772
min	0.000000	-0.010000	1.000000
25%	0.000000	0.000000	12.000000
50%	0.000000	0.000000	59.000000
75%	0.000000	0.000000	394.000000
max	142280.000000	100.000000	146170.000000

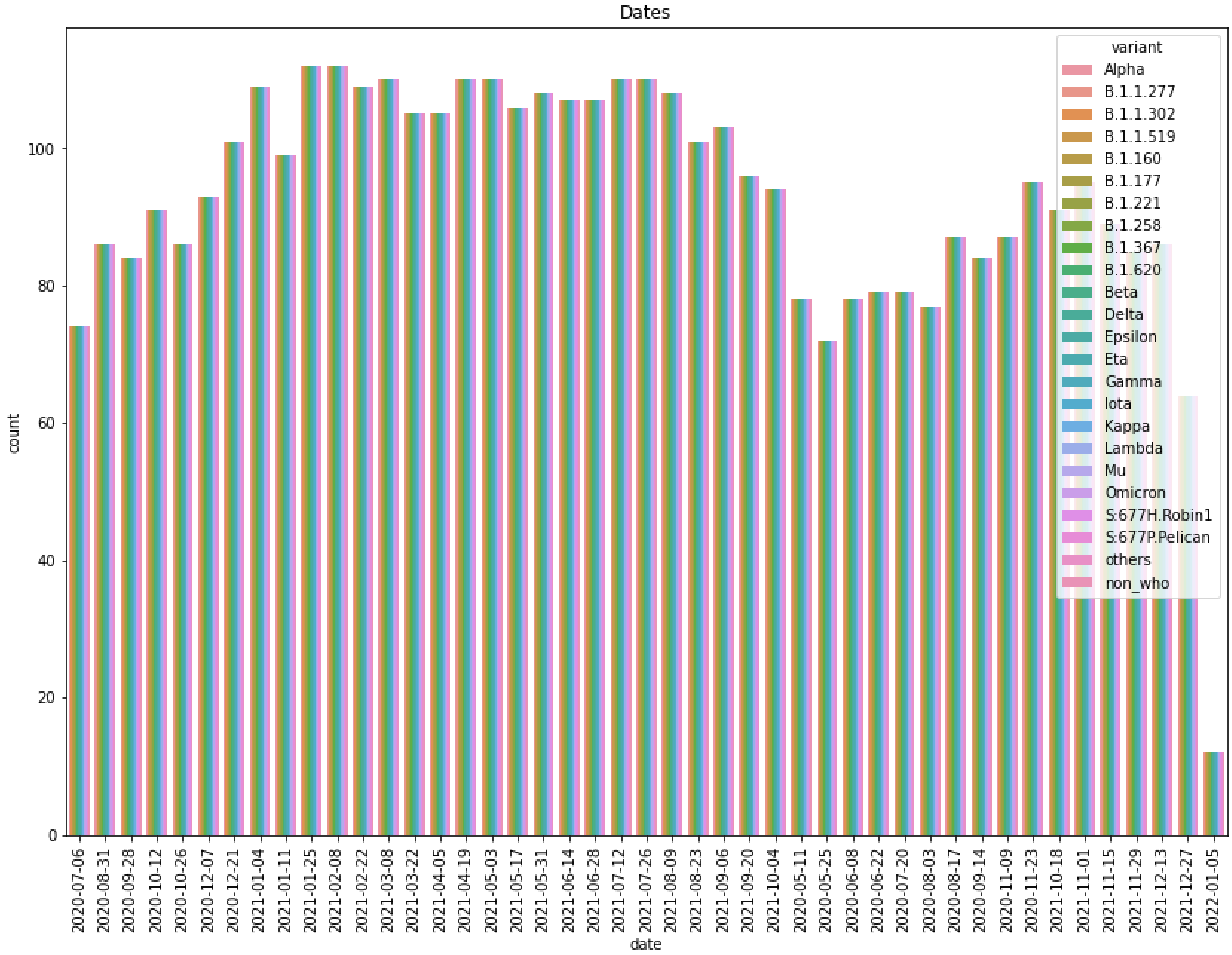
The Exploratort Data Analysis [EDA]

In [9]:

```
plt.figure(figsize=(14,10))
plt.xticks(rotation=90)
sns.countplot(data=cvd,x='date',hue='variant')
plt.title('Dates')
```

Out[9]:

```
Text(0.5, 1.0, 'Dates')
```

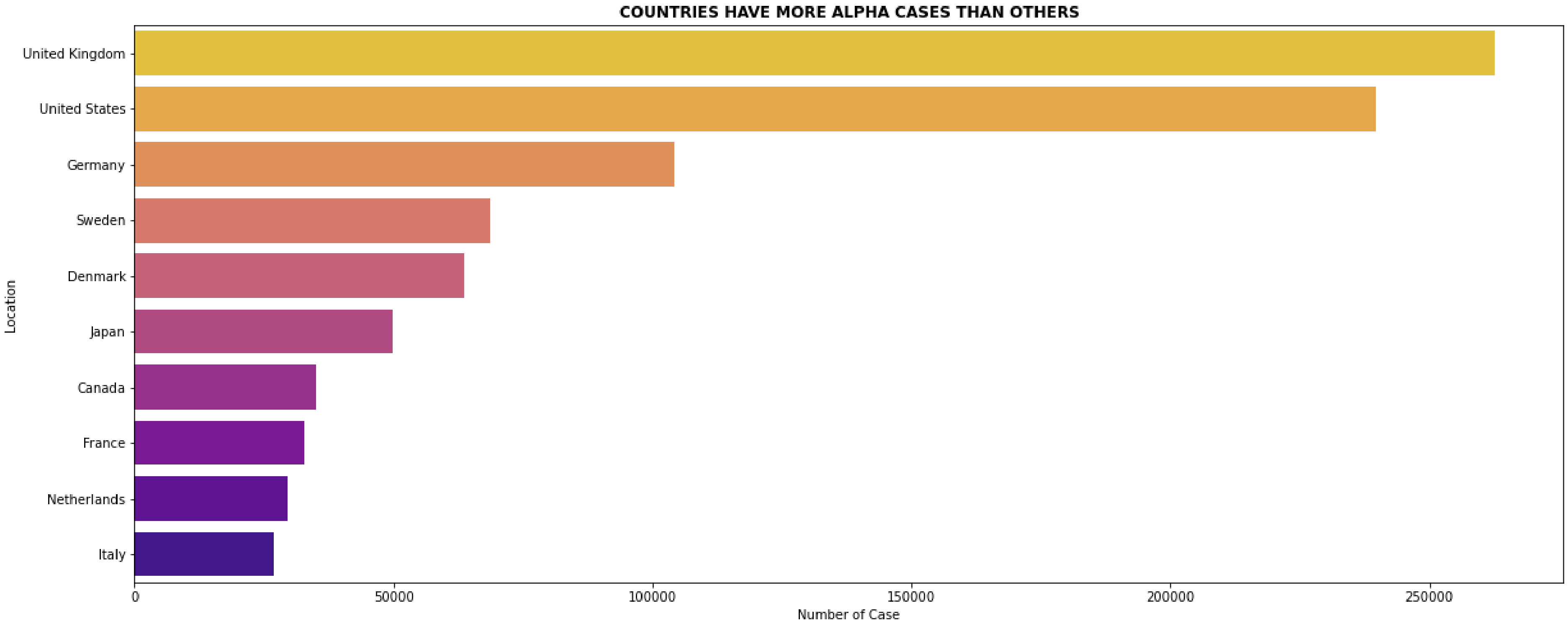


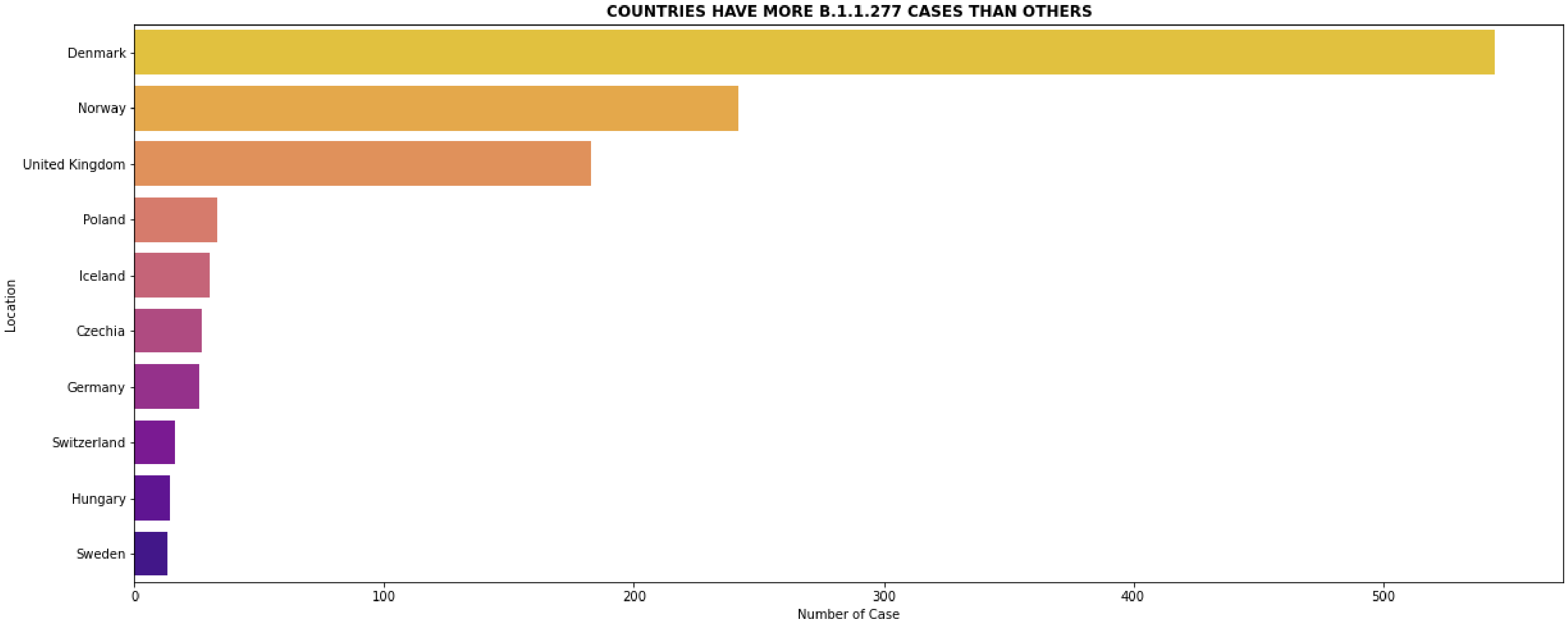
In [10]:

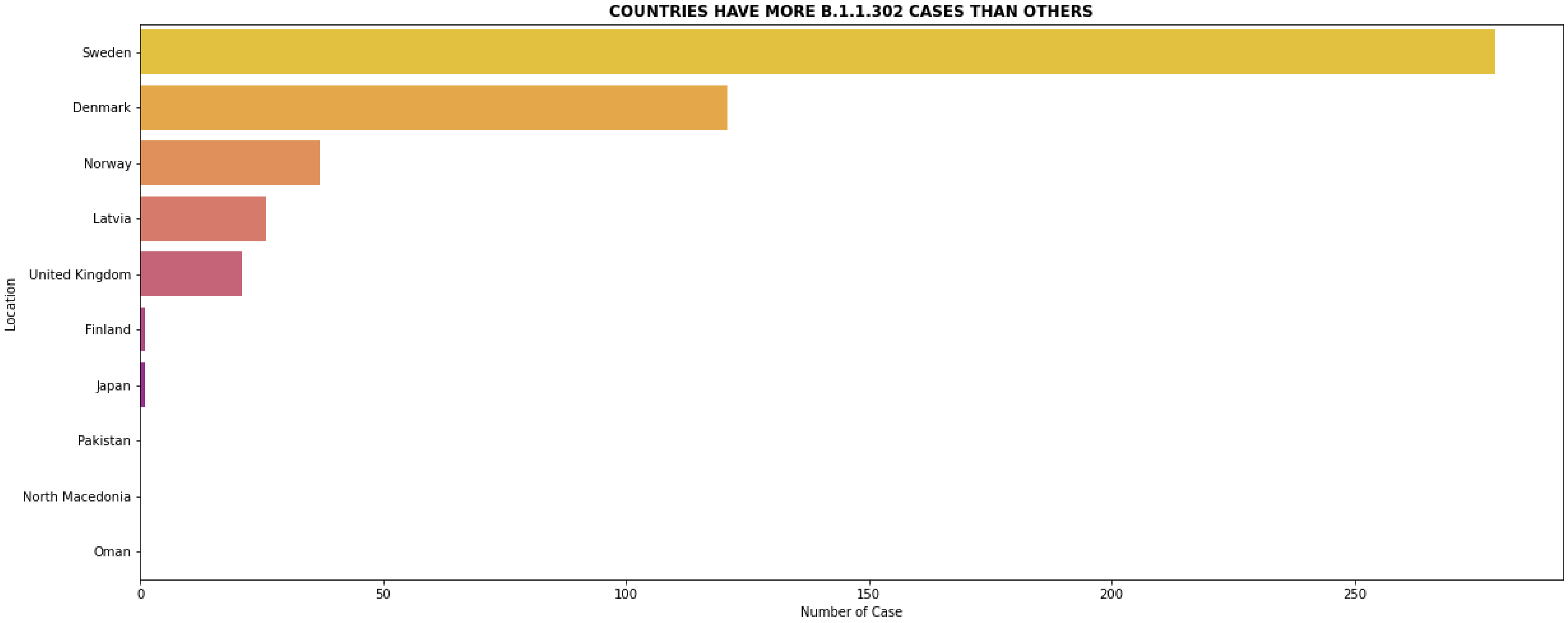
```
# Let's check the variant wise with top 10 countries with maximum virus
```

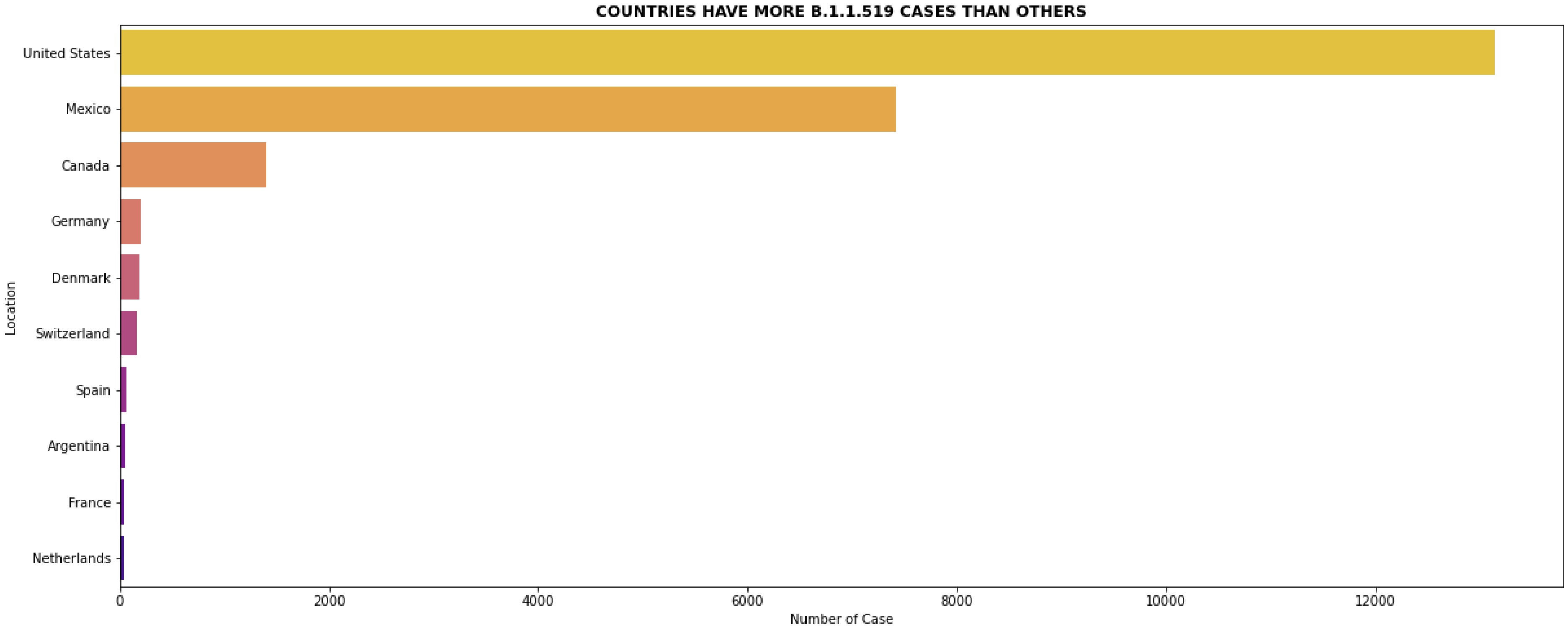
```
for virus in cvd.variant.unique():
    most_cases = cvd.loc[cvd['variant'] == virus].groupby('location')[
        'num_sequences'].agg('sum').sort_values(ascending=False)[:10]

    most_cases = pd.DataFrame({'Location':most_cases.index, 'Number of Case':most_cases.values})

    plt.figure(figsize=(20,8))
    sns.barplot(y='Location',x="Number of Case",data=most_cases,palette="plasma_r")
    plt.title('COUNTRIES HAVE MORE {} CASES THAN OTHERS'.format(virus).upper(),loc='center',fontweight="bold")
```

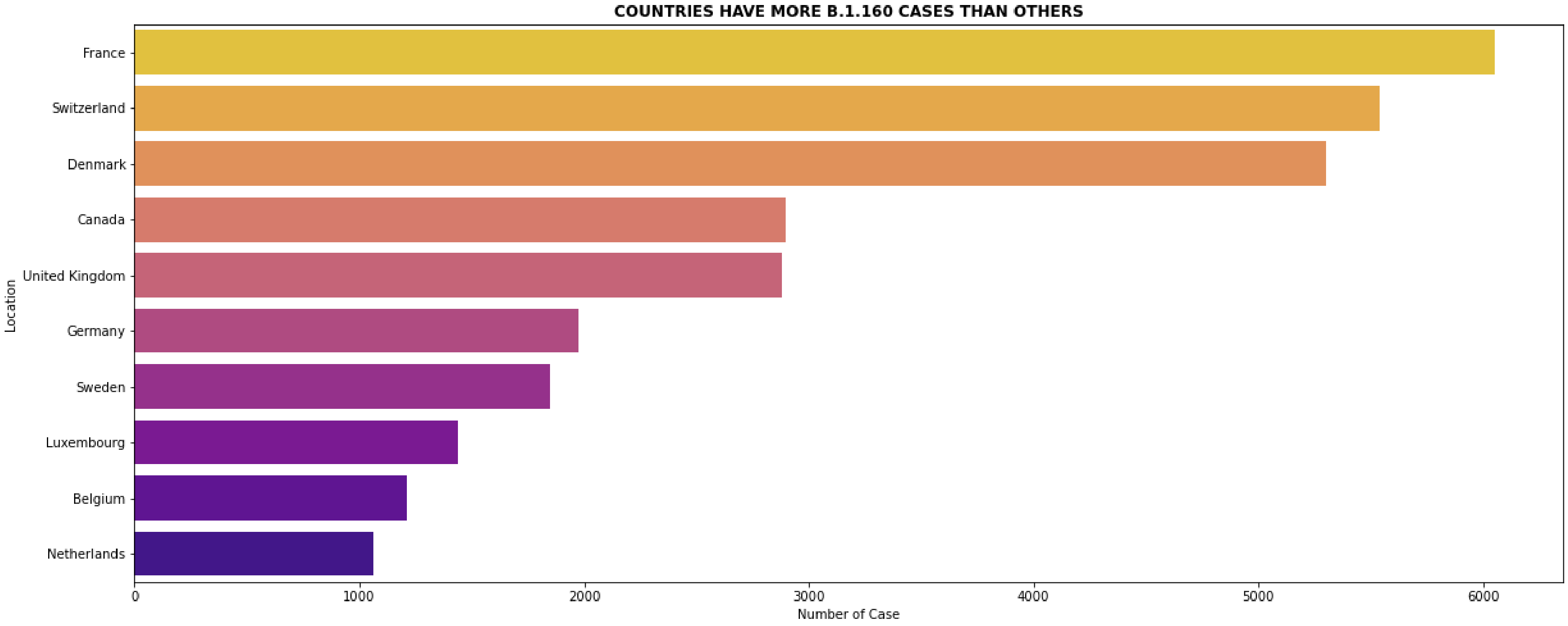
```
C:\Users\BUSINE~2\AppData\Local\Temp/ipykernel_18476/186801648.py:9: RuntimeWarning: More than 20 figures have been opened. Figure
s created through the pyplot interface (`matplotlib.pyplot.figure`) are retained until explicitly closed and may consume too much
memory. (To control this warning, see the rcParam `figure.max_open_warning`).
  plt.figure(figsize=(20,8))
```

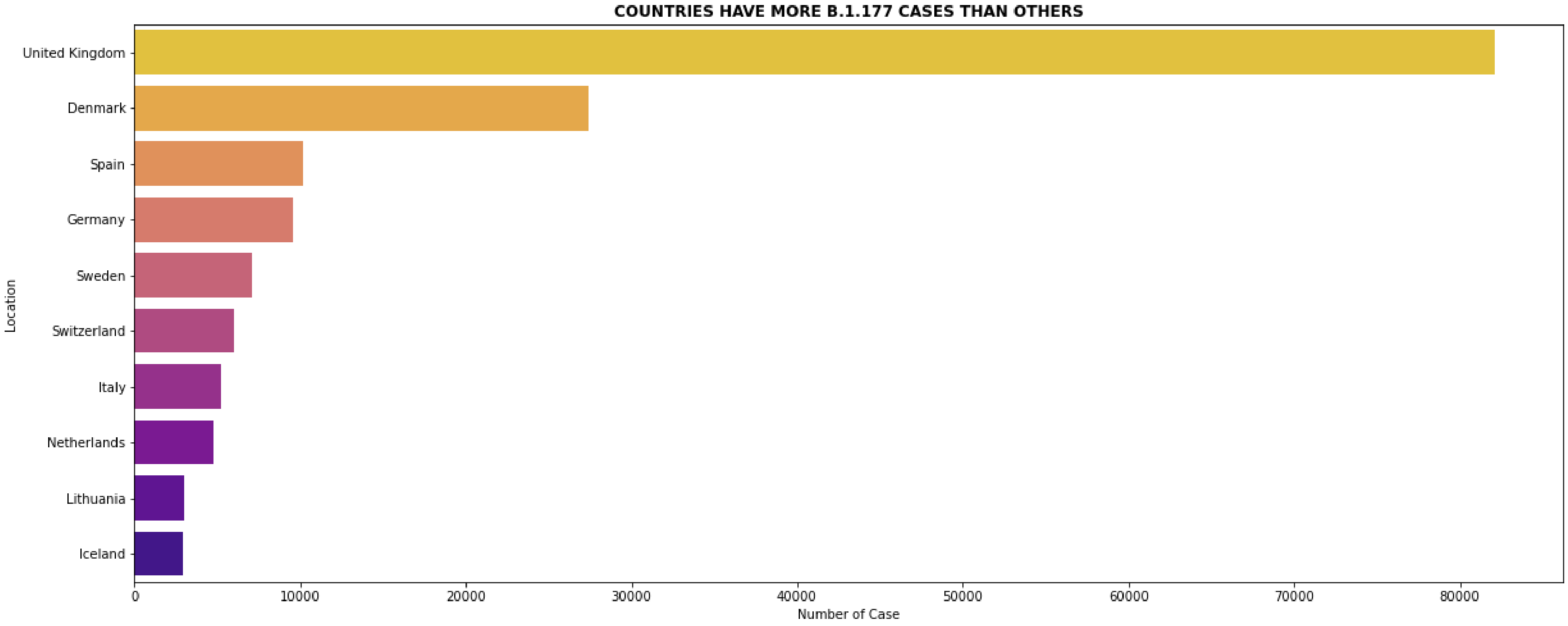


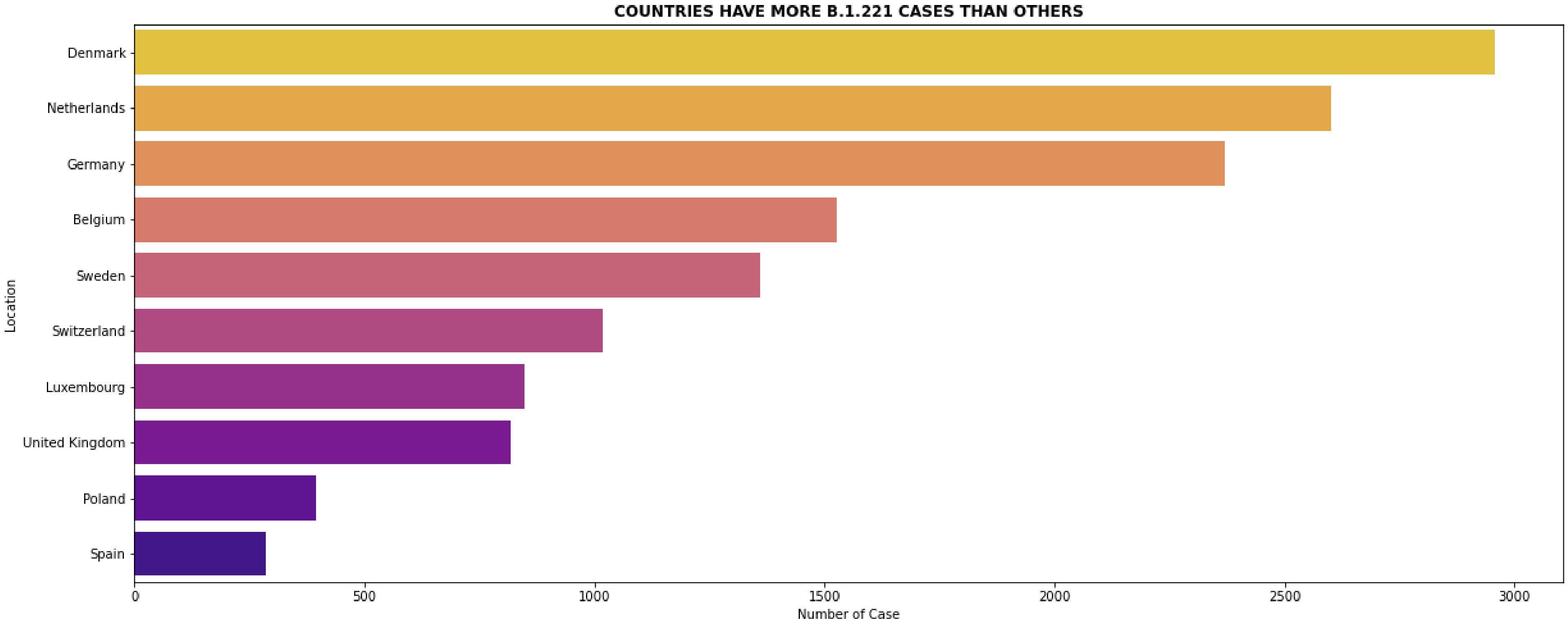

COUNTRIES HAVE MORE B.1.1.277 CASES THAN OTHERS
Denmark
Norway
United Kingdom
Poland
Iceland
Czechia
Germany
Switzerland
Hungary
Sweden
Location
0
100
200
300
400
500
Number of Case

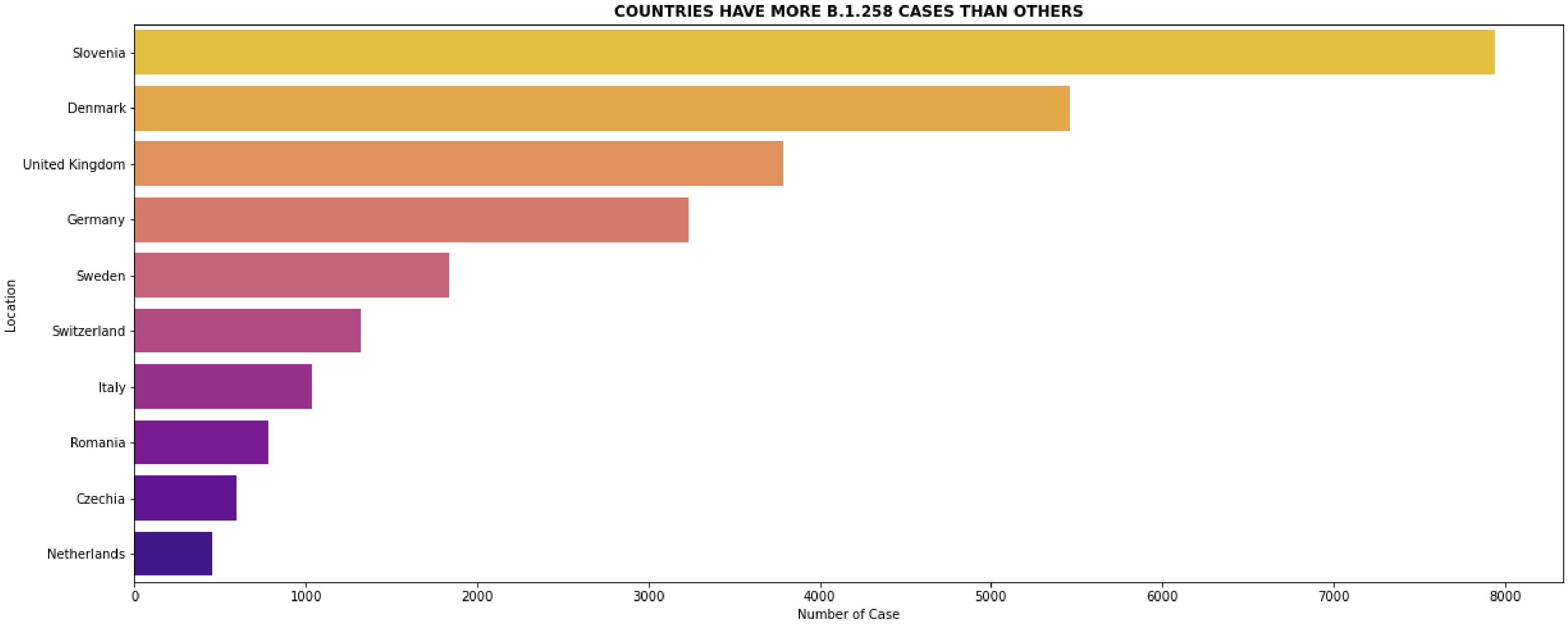

COUNTRIES HAVE MORE B.1.1.302 CASES THAN OTHERS
Location
Sweden
Denmark
Norway
Latvia
United Kingdom
Finland
Japan
Pakistan
North Macedonia
Oman
0
50
100
150
200
250
Number of Case

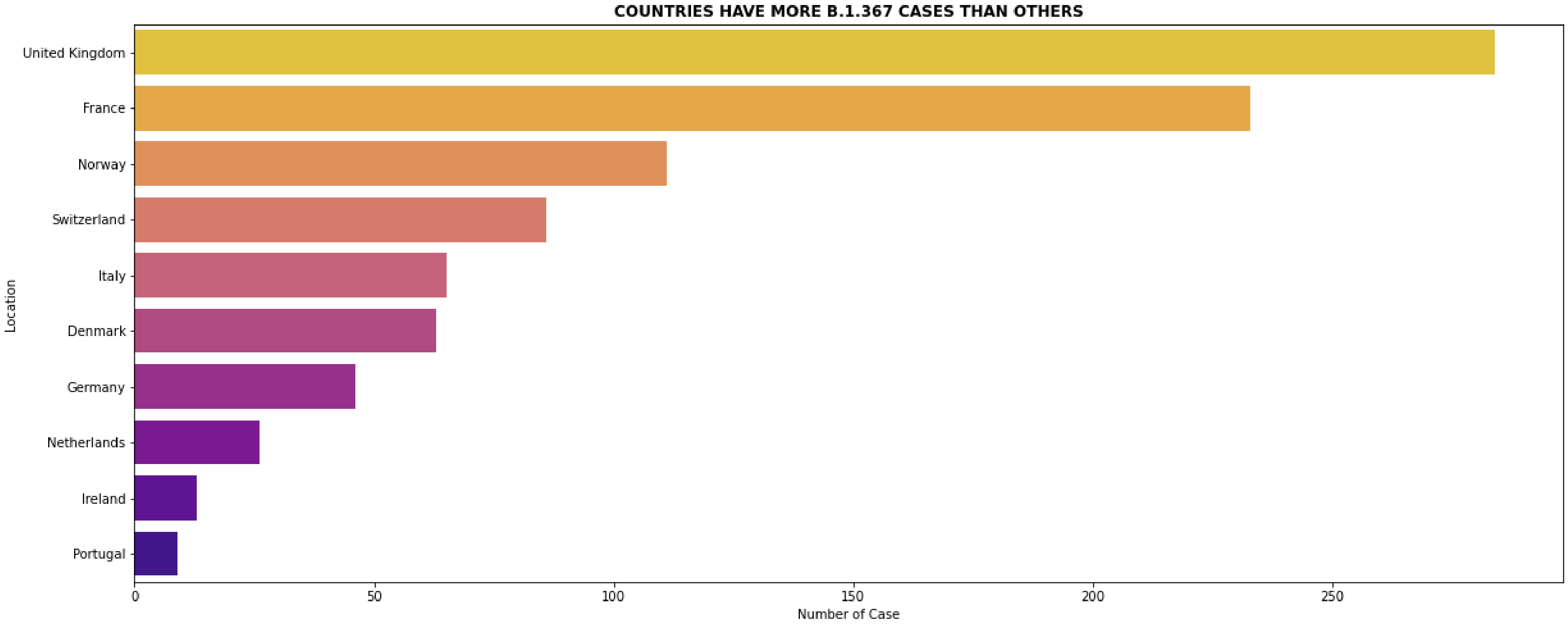

COUNTRIES HAVE MORE B.1.1.519 CASES THAN OTHERS
United States
Mexico
Canada
Germany
Denmark
Switzerland
Spain
Argentina
France
Netherlands
Location
0
2000
4000
6000
8000
10000
12000
Number of Case

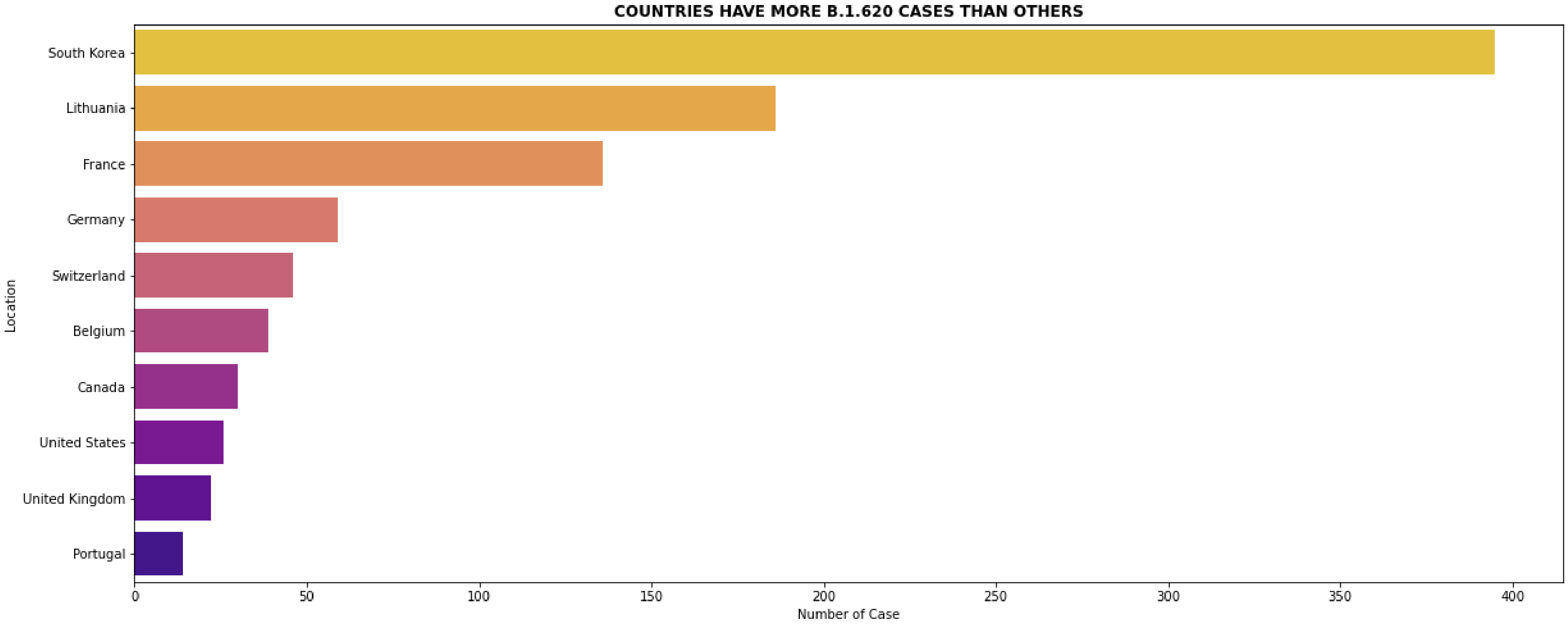

COUNTRIES HAVE MORE B.1.160 CASES THAN OTHERS
Location
France
Switzerland
Denmark
Canada
United Kingdom
Germany
Sweden
Luxembourg
Belgium
Netherlands
0
1000
2000
3000
4000
5000
6000
Number of Case

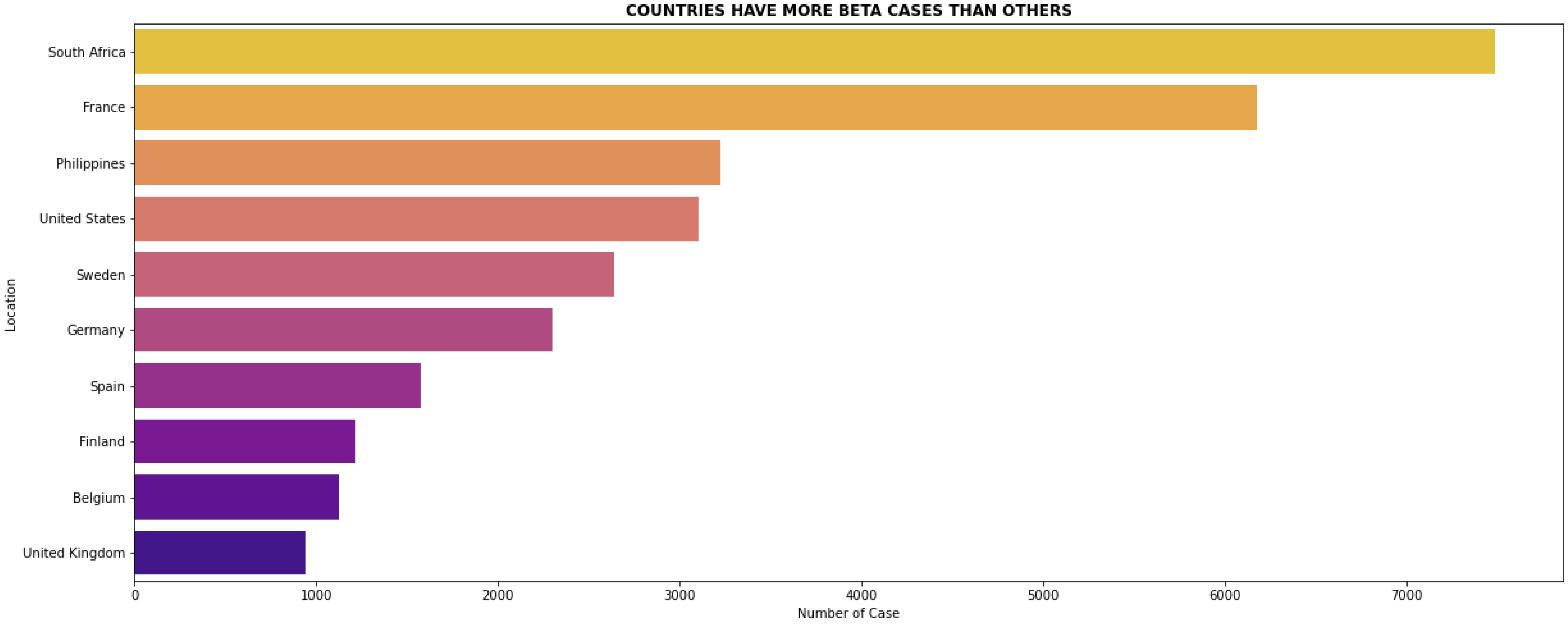

COUNTRIES HAVE MORE B.1.177 CASES THAN OTHERS
United Kingdom
Denmark
Spain
Germany
Sweden
Switzerland
Italy
Netherlands
Lithuania
Iceland
Location
0
10000
20000
30000
40000
50000
60000
70000
80000
Number of Case

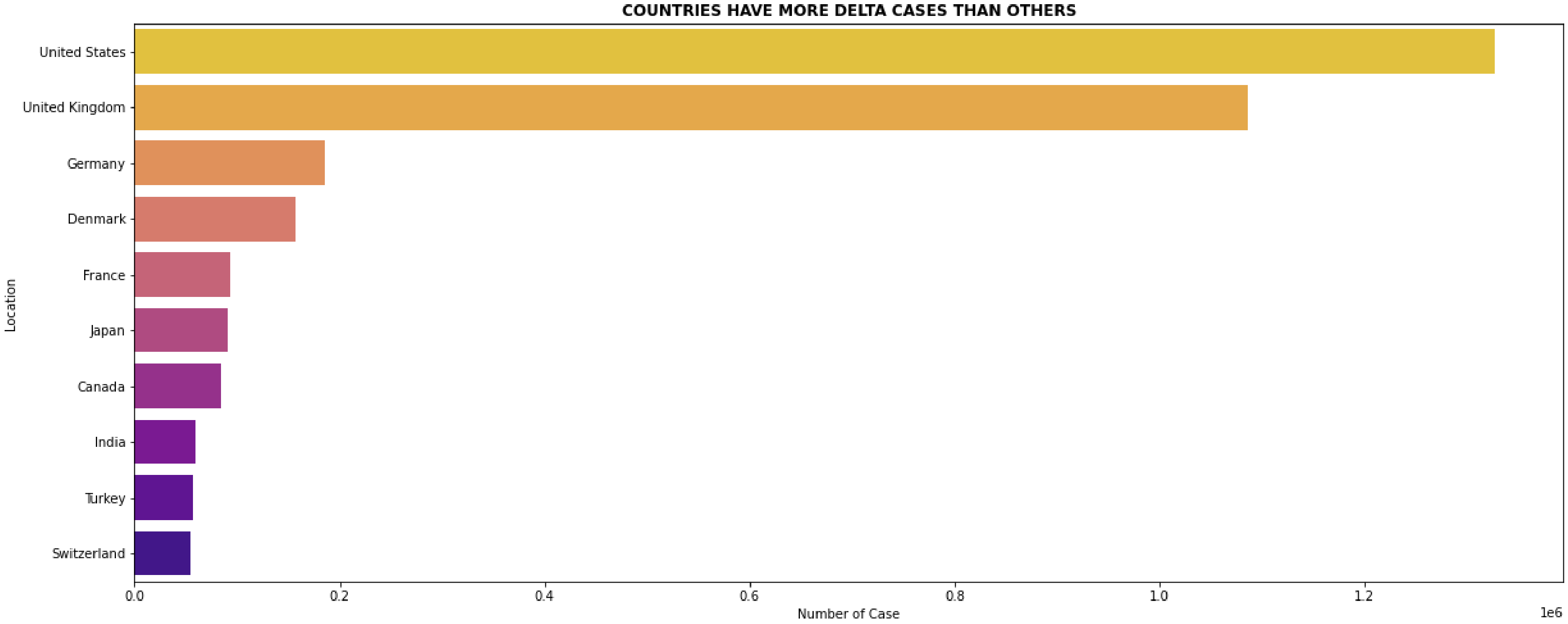

COUNTRIES HAVE MORE B.1.221 CASES THAN OTHERS
Location
Denmark
Netherlands
Germany
Belgium
Sweden
Switzerland
Luxembourg
United Kingdom
Poland
Spain
0
500
1000
1500
2000
2500
3000
Number of Case

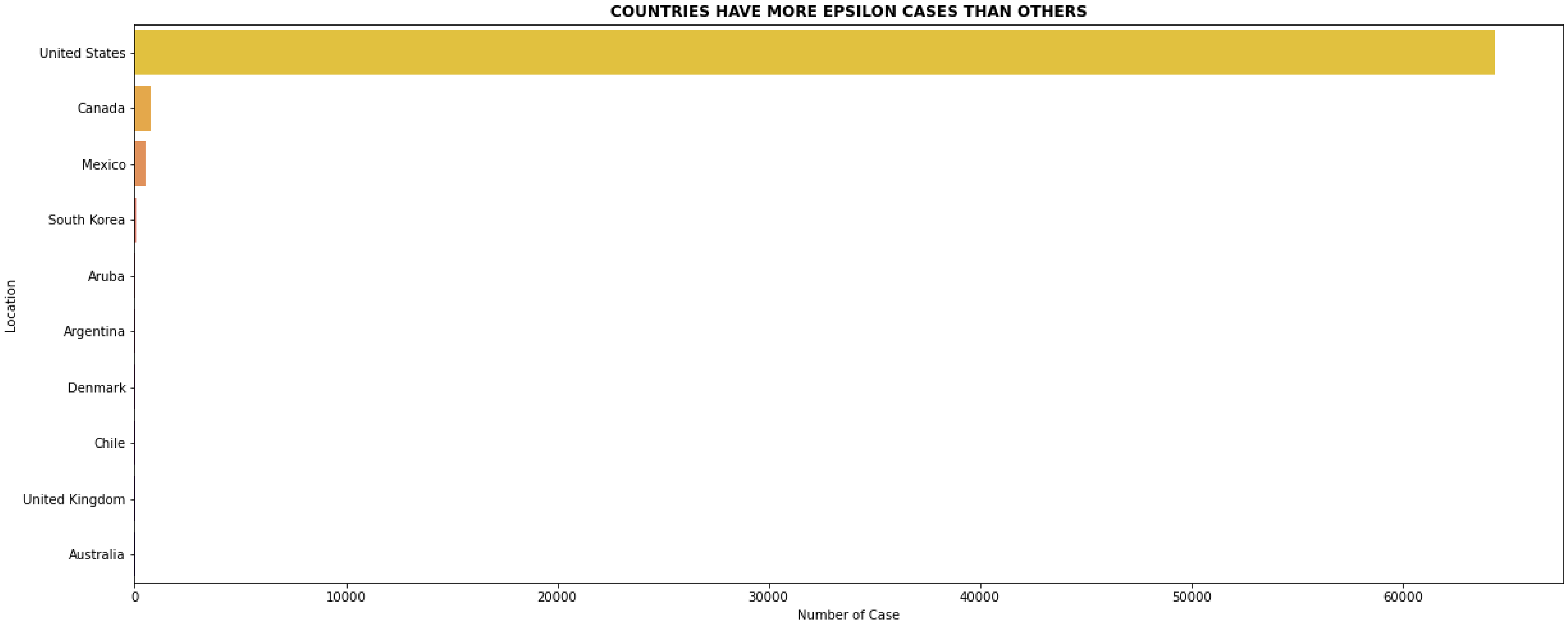

COUNTRIES HAVE MORE B.1.258 CASES THAN OTHERS
Slovenia
Denmark
United Kingdom
Germany
Sweden
Switzerland
Italy
Romania
Czechia
Netherlands
Location
0
1000
2000
3000
4000
5000
6000
7000
8000
Number of Case

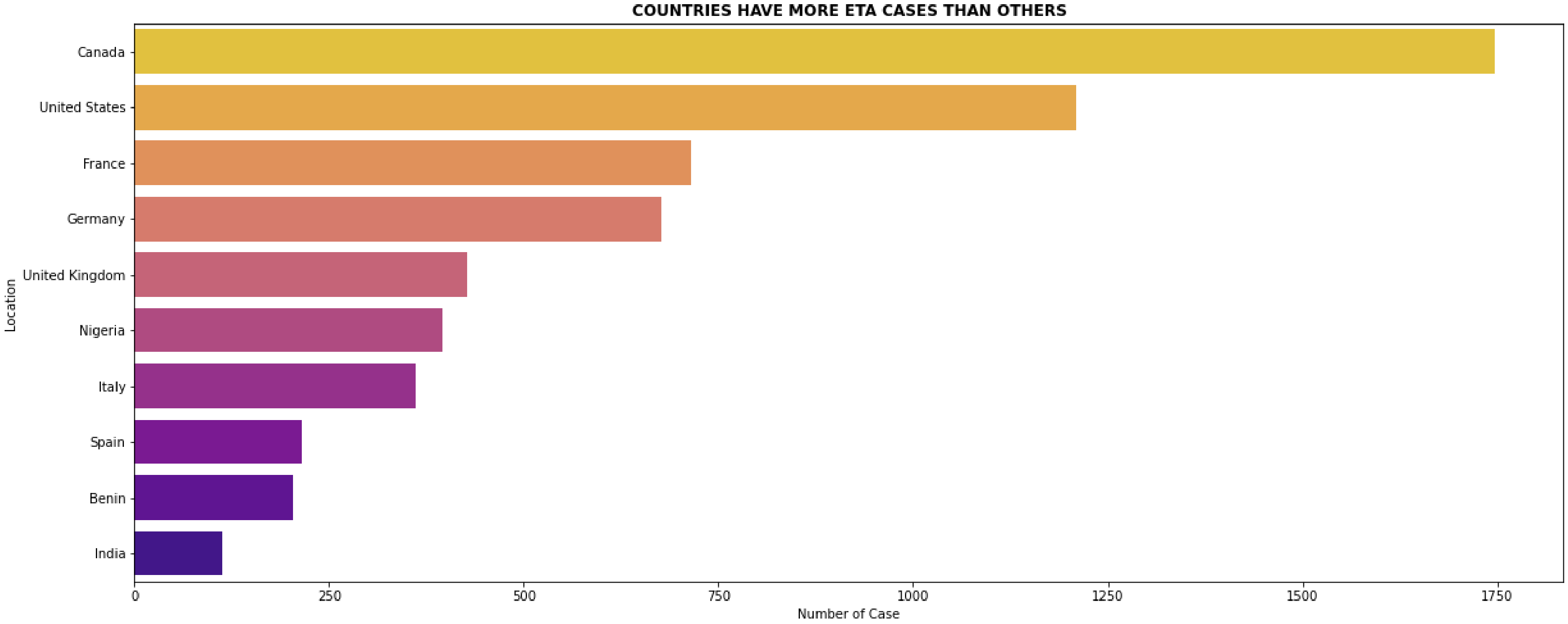

COUNTRIES HAVE MORE B.1.367 CASES THAN OTHERS
Location
United Kingdom
France
Norway
Switzerland
Italy
Denmark
Germany
Netherlands
Ireland
Portugal
0
50
100
150
200
250
Number of Case

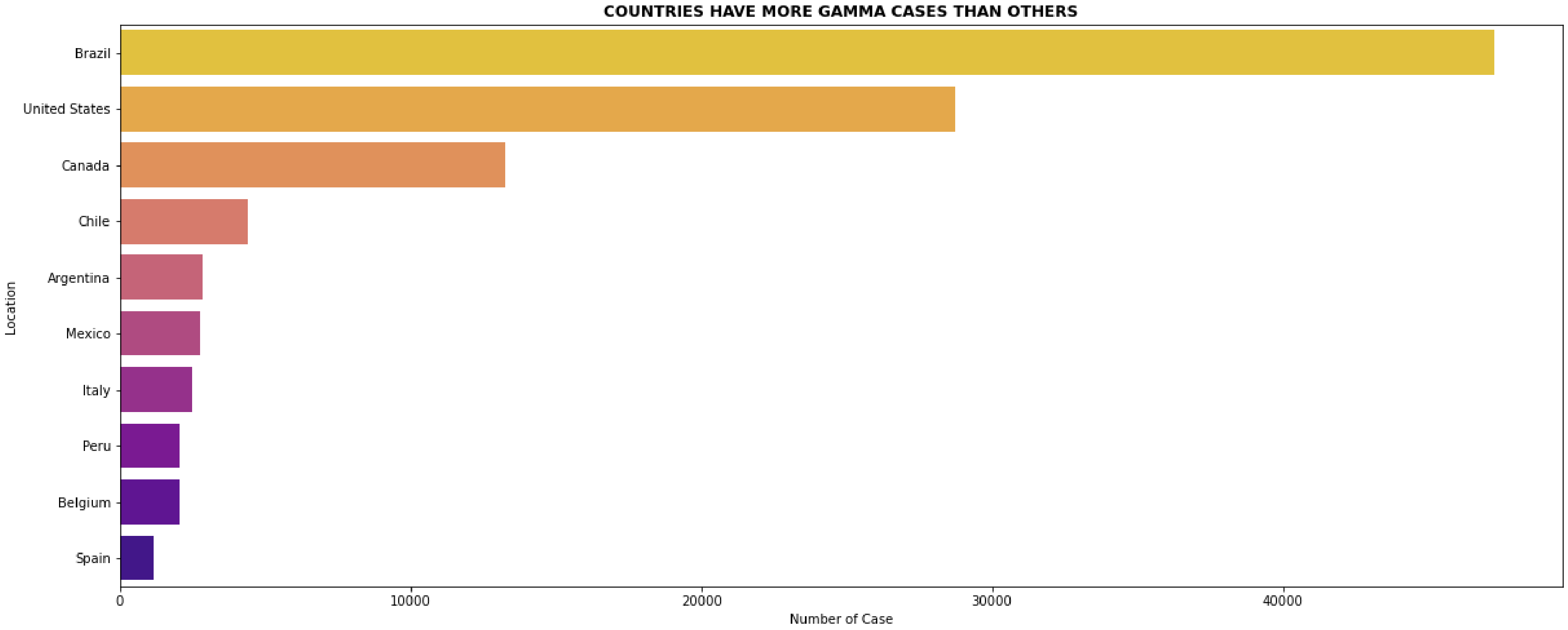

COUNTRIES HAVE MORE B.1.620 CASES THAN OTHERS
Location
South Korea
Lithuania
France
Germany
Switzerland
Belgium
Canada
United States
United Kingdom
Portugal
0
50
100
150
200
250
300
350
400
Number of Case

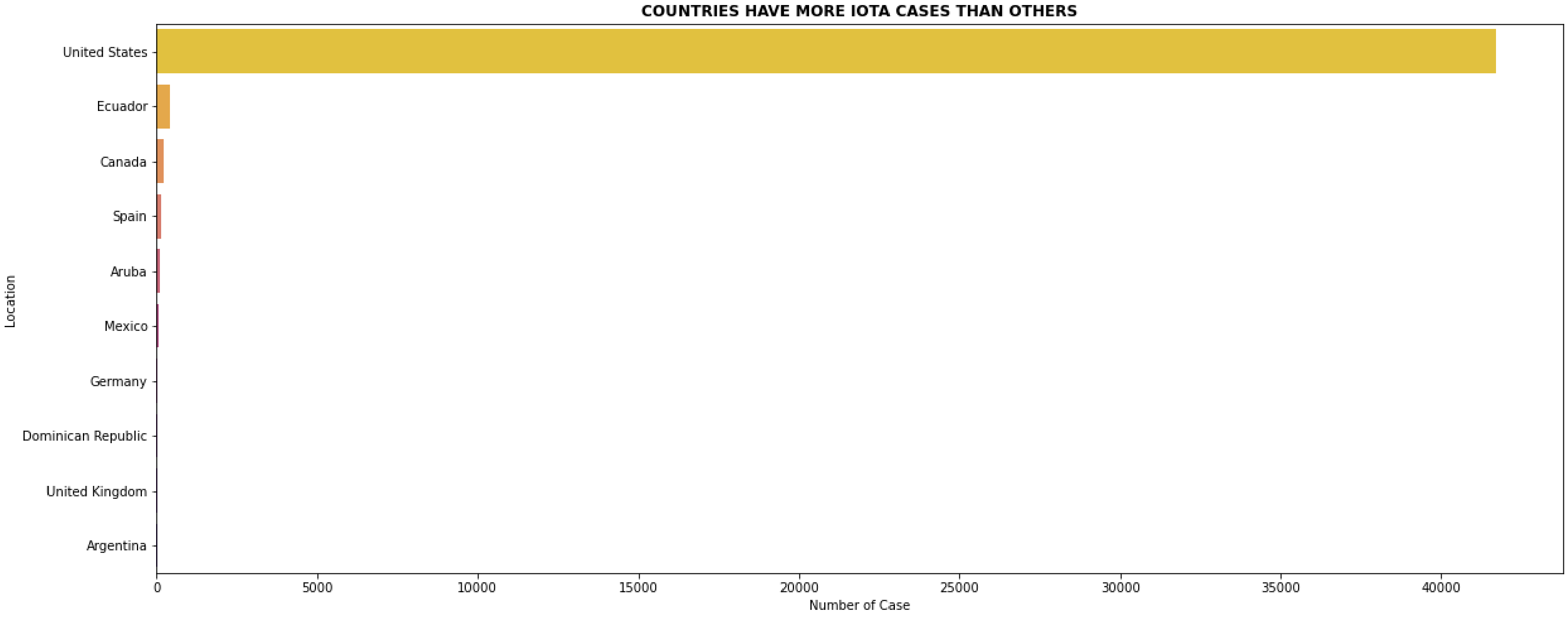

COUNTRIES HAVE MORE BETA CASES THAN OTHERS
Location
South Africa
France
Philippines
United States
Sweden
Germany
Spain
Finland
Belgium
United Kingdom
0
1000
2000
3000
4000
5000
6000
7000
Number of Case

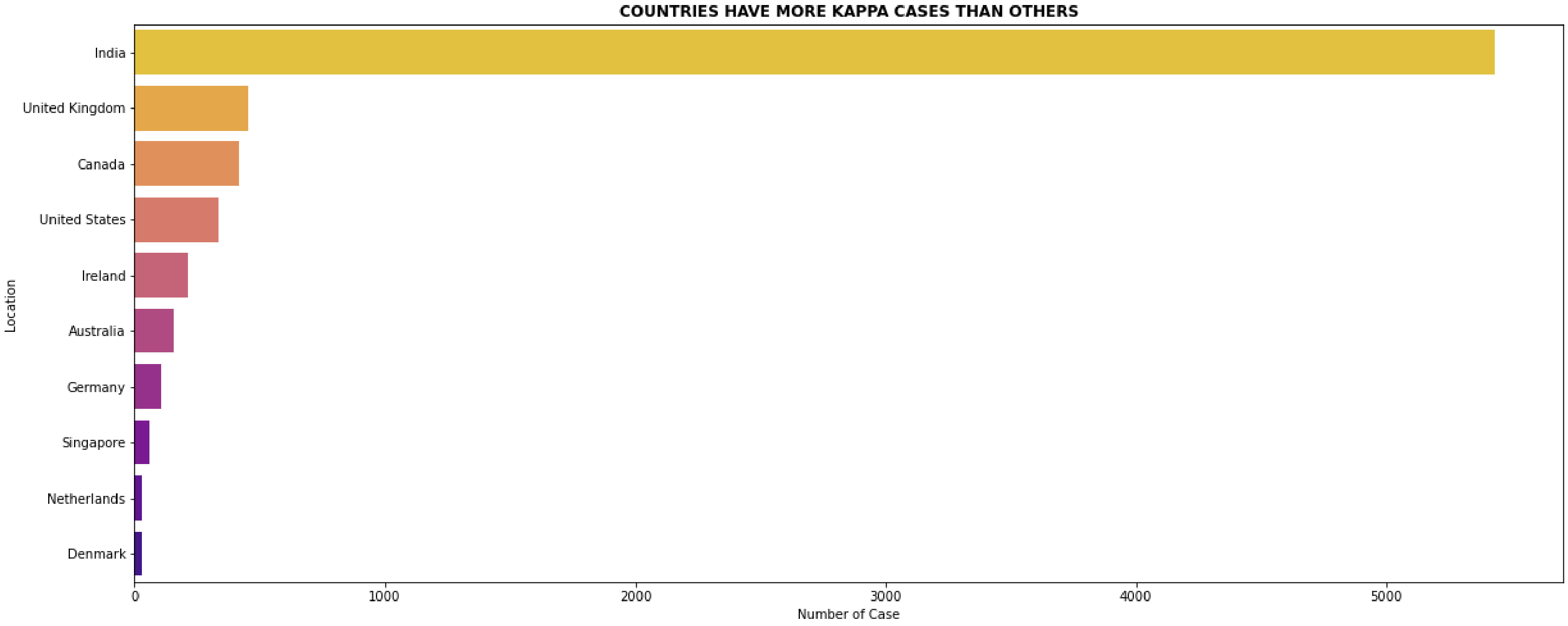

COUNTRIES HAVE MORE DELTA CASES THAN OTHERS
United States
United Kingdom
Germany
Denmark
France
Japan
Canada
India
Turkey
Switzerland
Location
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e6
Number of Case

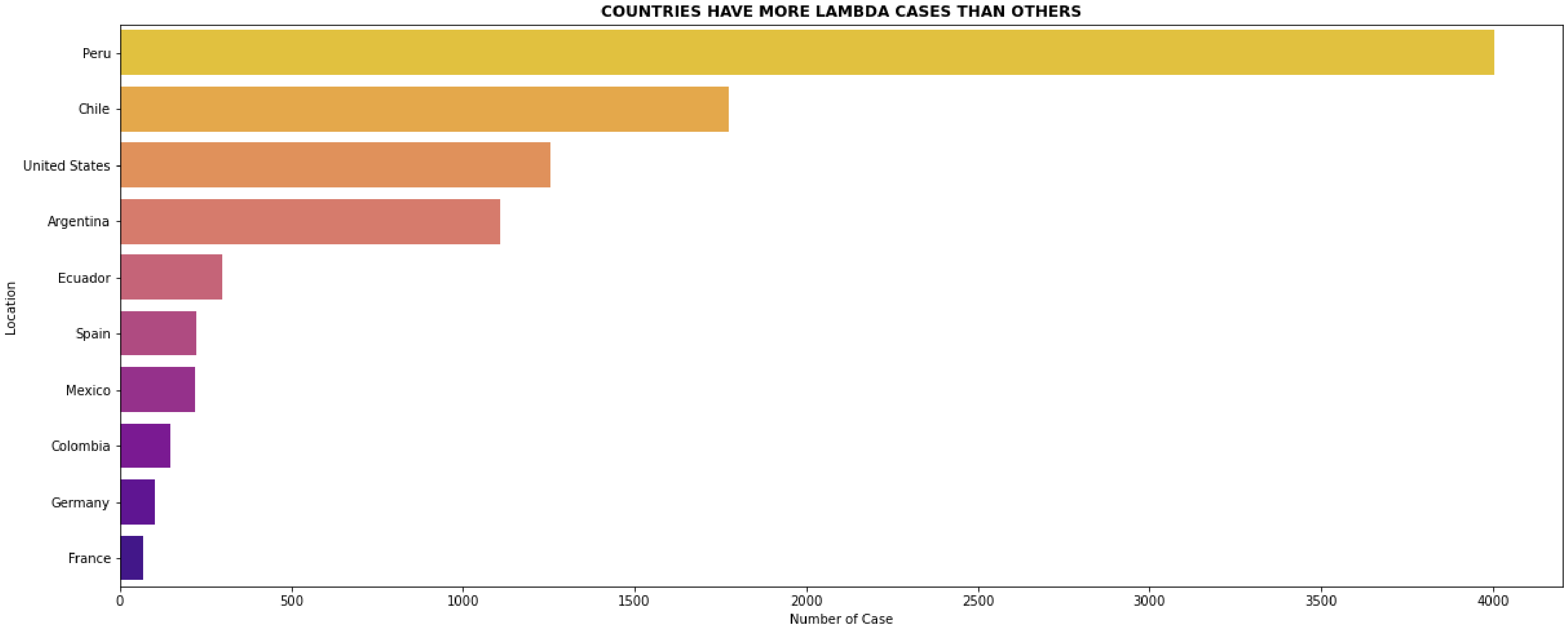

COUNTRIES HAVE MORE EPSILON CASES THAN OTHERS
Location
United States
Canada
Mexico
South Korea
Aruba
Argentina
Denmark
Chile
United Kingdom
Australia
0
10000
20000
30000
40000
50000
60000
Number of Case

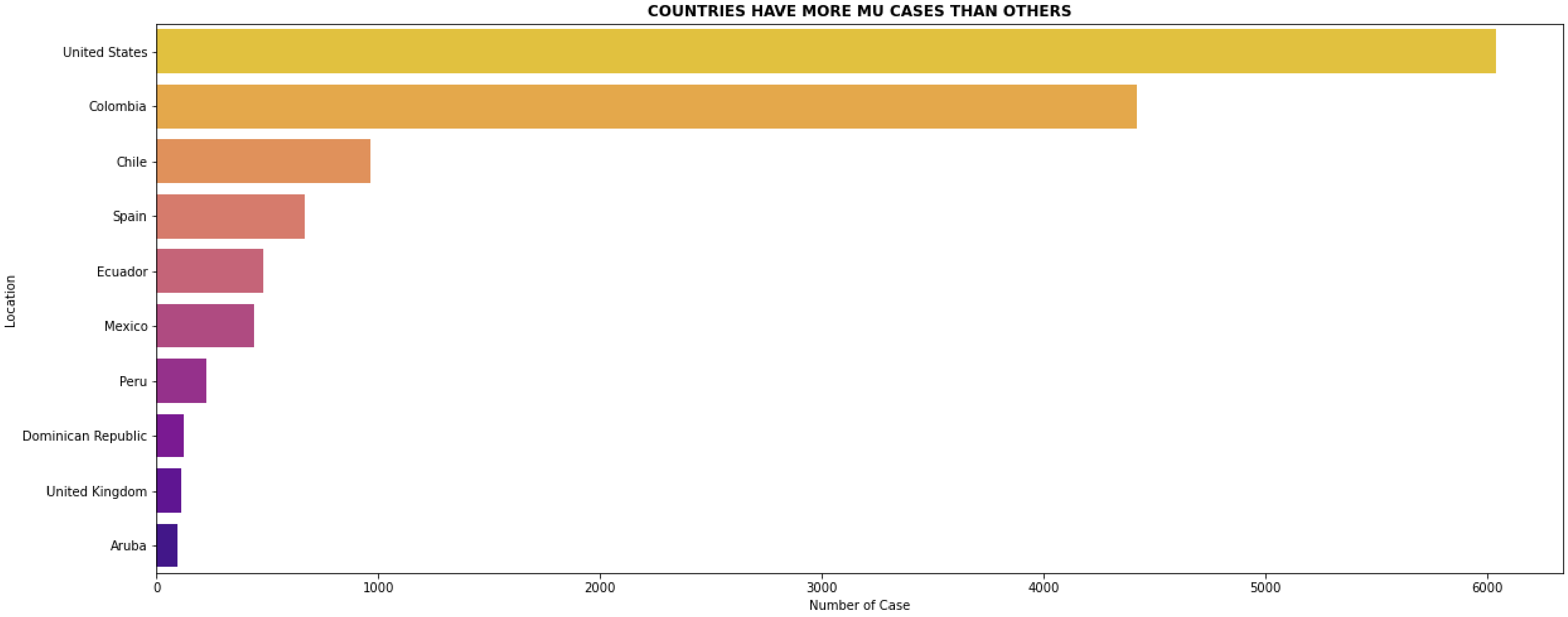

COUNTRIES HAVE MORE ETA CASES THAN OTHERS
Location
Canada
United States
France
Germany
United Kingdom
Nigeria
Italy
Spain
Benin
India
0
250
500
750
1000
1250
1500
1750
Number of Case

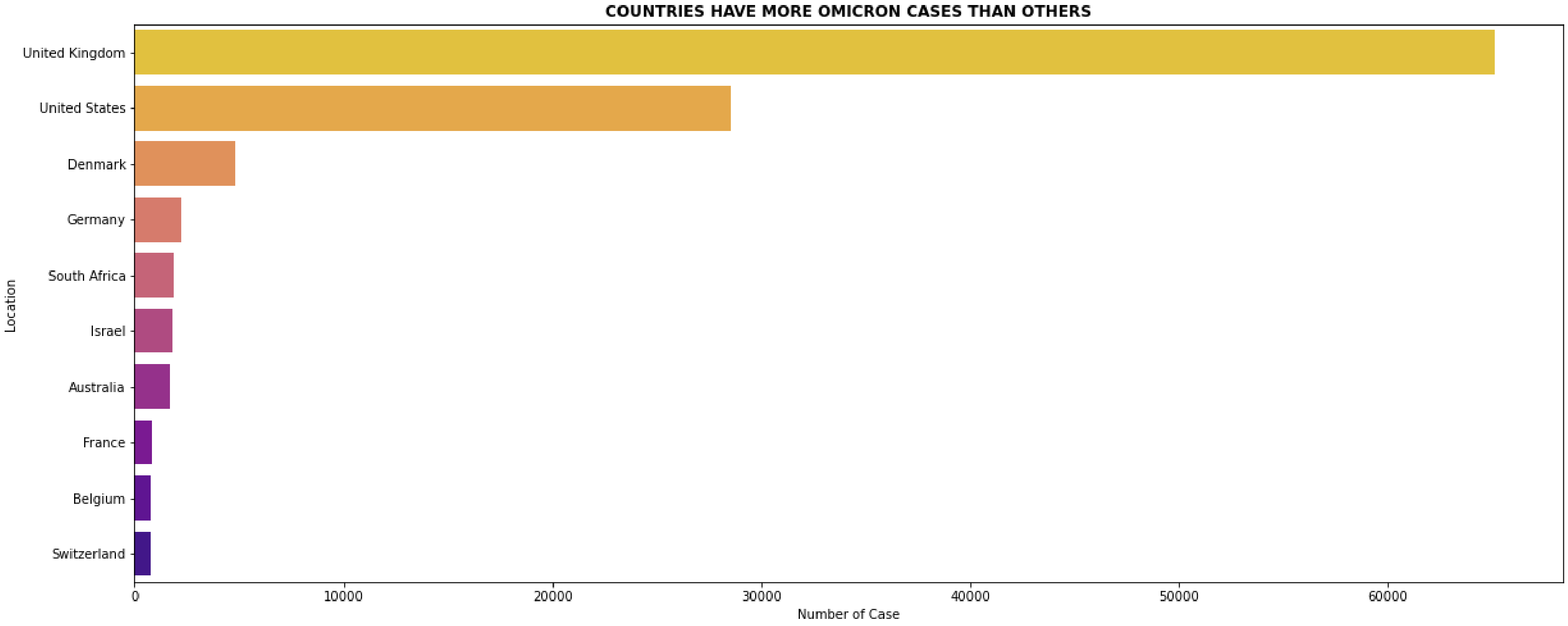

COUNTRIES HAVE MORE GAMMA CASES THAN OTHERS
Location
Brazil
United States
Canada
Chile
Argentina
Mexico
Italy
Peru
Belgium
Spain
0
10000
20000
30000
40000
Number of Case

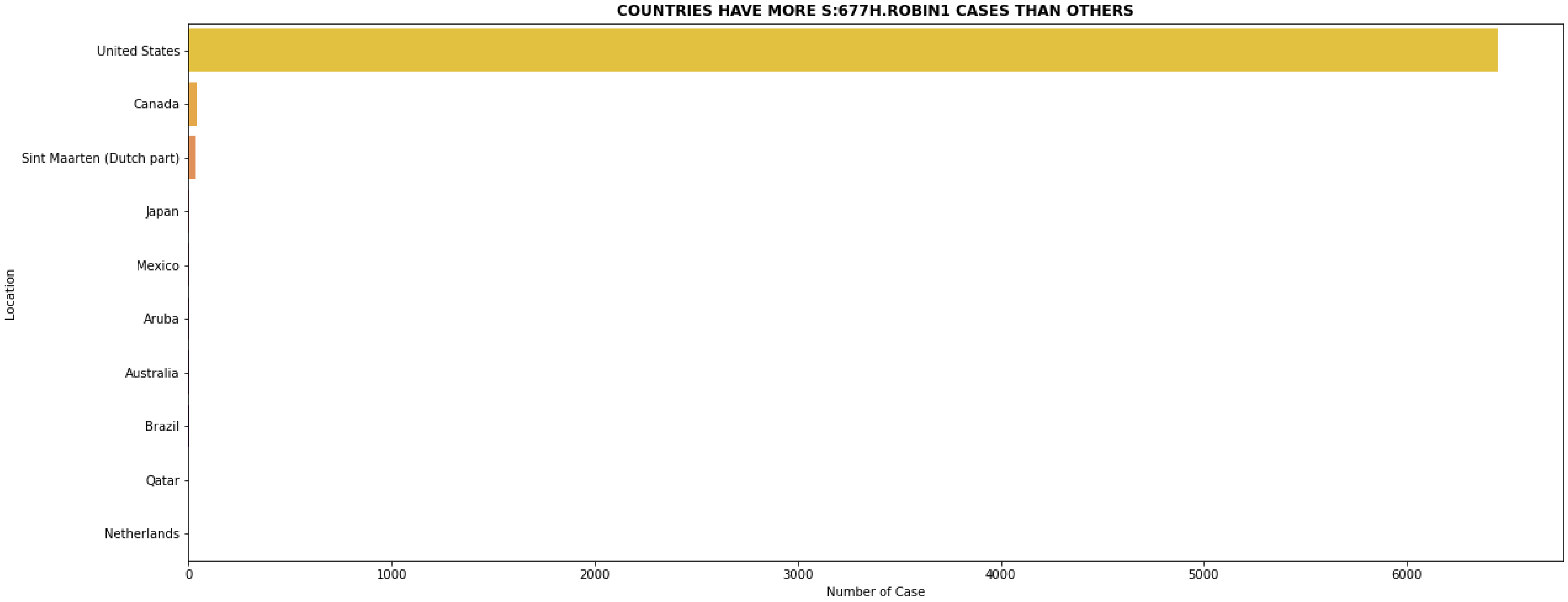

COUNTRIES HAVE MORE IOTA CASES THAN OTHERS
Location
United States
Ecuador
Canada
Spain
Aruba
Mexico
Germany
Dominican Republic
United Kingdom
Argentina
0
5000
10000
15000
20000
25000
30000
35000
40000
Number of Case

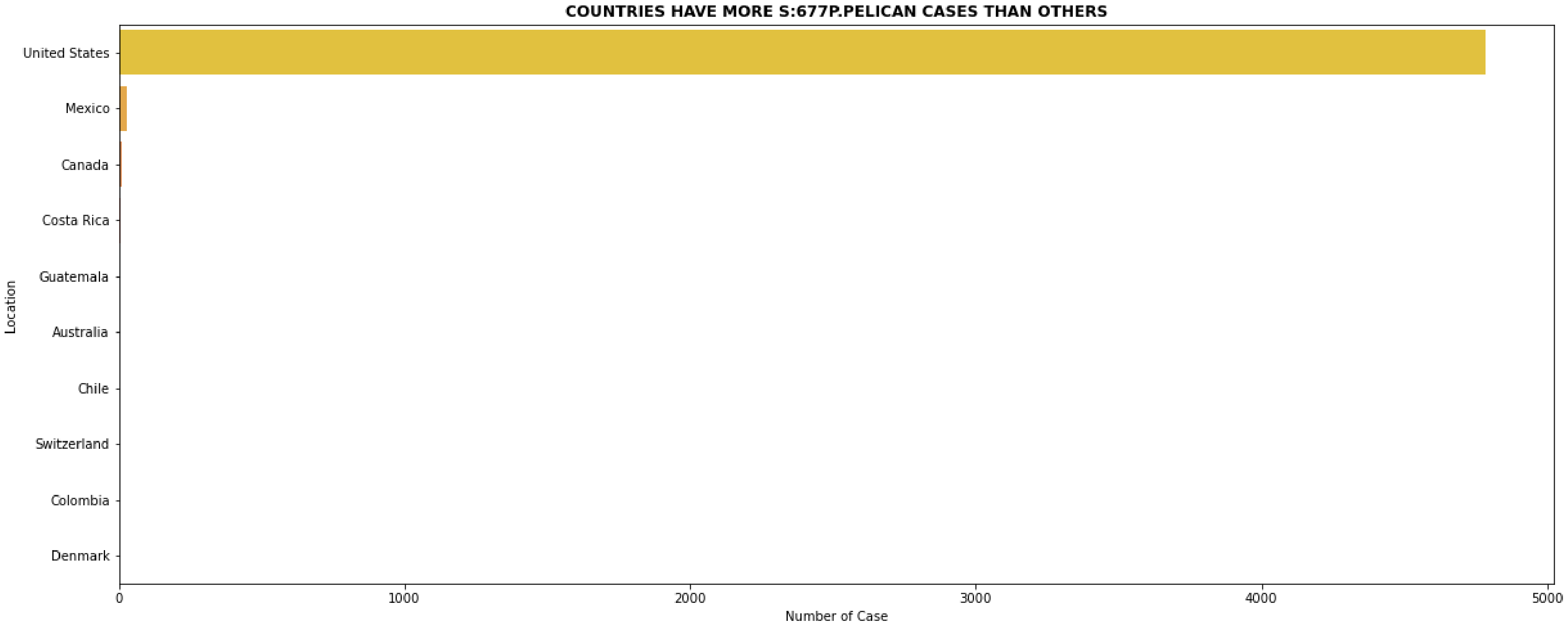

COUNTRIES HAVE MORE KAPPA CASES THAN OTHERS
Location
India
United Kingdom
Canada
United States
Ireland
Australia
Germany
Singapore
Netherlands
Denmark
0
1000
2000
3000
4000
5000
Number of Case

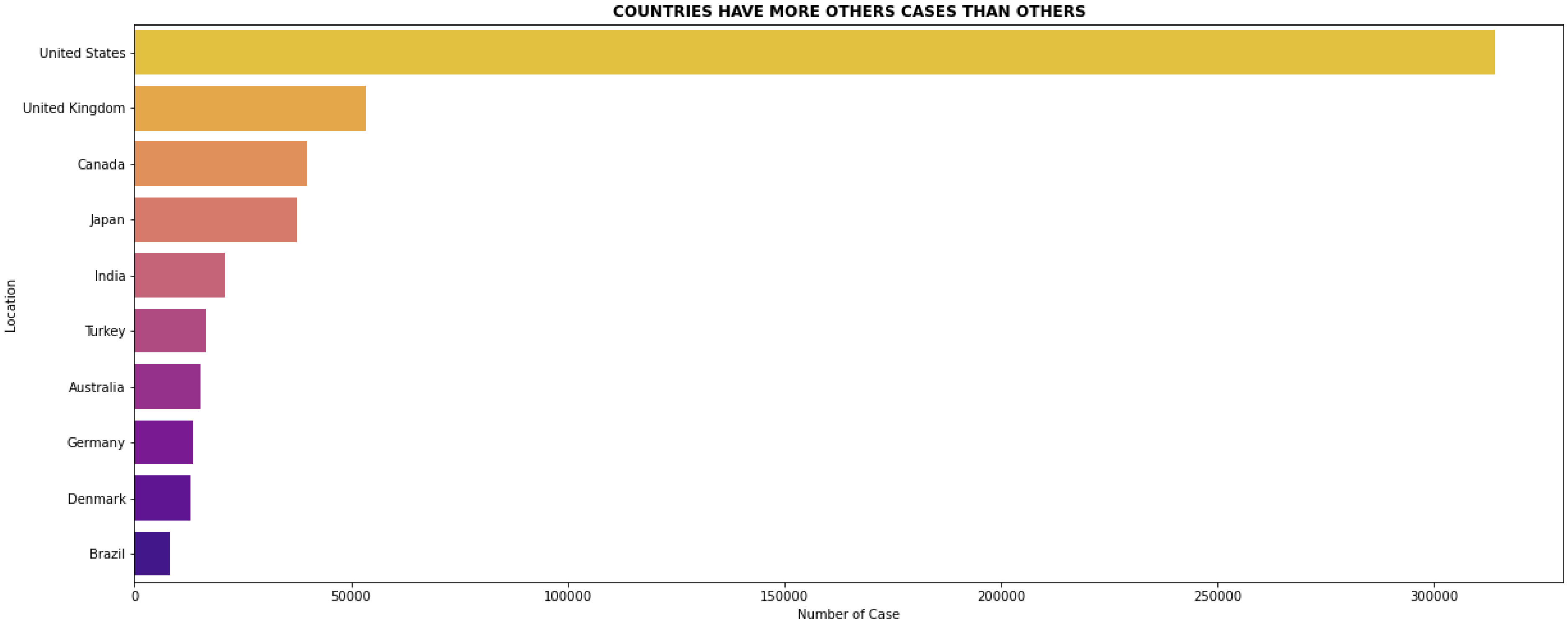

COUNTRIES HAVE MORE LAMBDA CASES THAN OTHERS
Location
Peru
Chile
United States
Argentina
Ecuador
Spain
Mexico
Colombia
Germany
France
0
500
1000
1500
2000
2500
3000
3500
4000
Number of Case

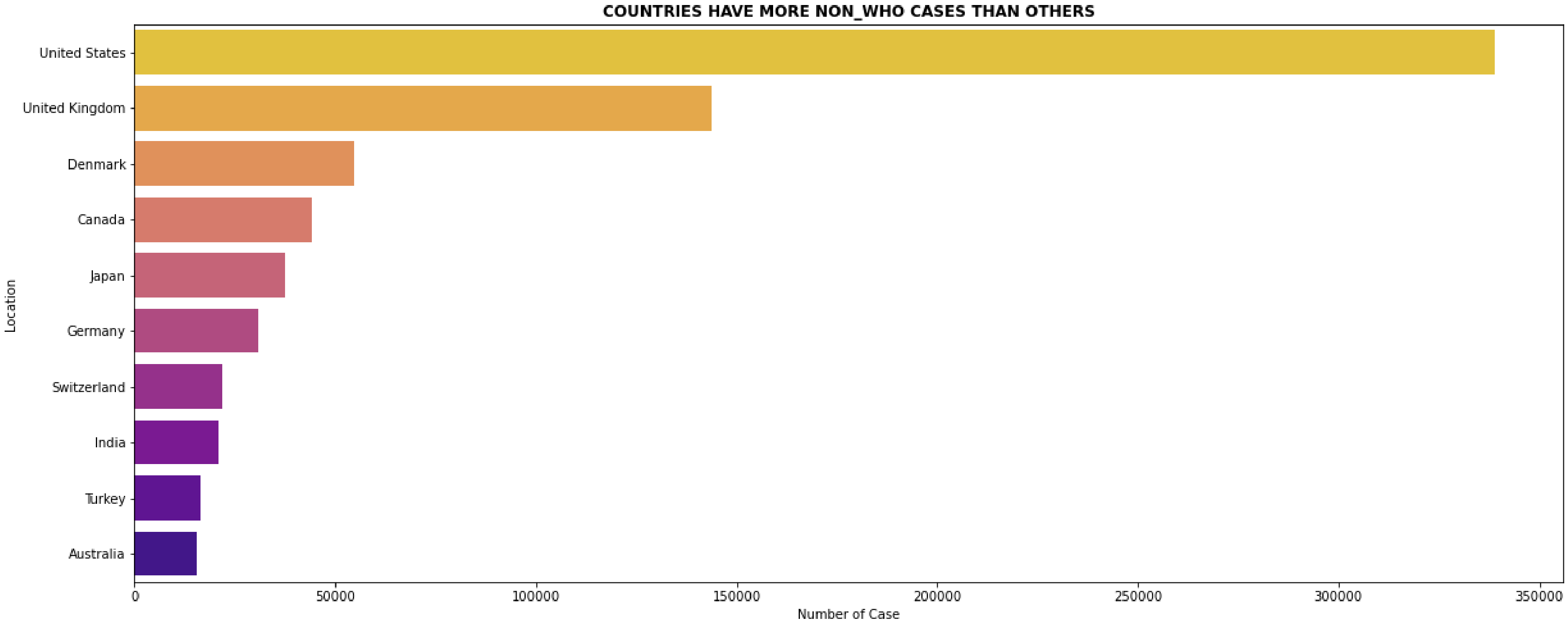

COUNTRIES HAVE MORE MU CASES THAN OTHERS
Location
United States
Colombia
Chile
Spain
Ecuador
Mexico
Peru
Dominican Republic
United Kingdom
Aruba
0
1000
2000
3000
4000
5000
6000
Number of Case


COUNTRIES HAVE MORE OMICRON CASES THAN OTHERS
United Kingdom
United States
Denmark
Germany
South Africa
Israel
Australia
France
Belgium
Switzerland
Location
0
10000
20000
30000
40000
50000
60000
Number of Case


COUNTRIES HAVE MORE S:677H.ROBIN1 CASES THAN OTHERS
United States
Canada
Sint Maarten (Dutch part)
Japan
Mexico
Aruba
Australia
Brazil
Qatar
Netherlands
Location
0
1000
2000
3000
4000
5000
6000
Number of Case


COUNTRIES HAVE MORE S:677P.PELICAN CASES THAN OTHERS
Location
United States
Mexico
Canada
Costa Rica
Guatemala
Australia
Chile
Switzerland
Colombia
Denmark
0
1000
2000
3000
4000
5000
Number of Case


COUNTRIES HAVE MORE OTHERS CASES THAN OTHERS
United States
United Kingdom
Canada
Japan
India
Turkey
Australia
Germany
Denmark
Brazil
Location
0
50000
100000
150000
200000
250000
300000
Number of Case

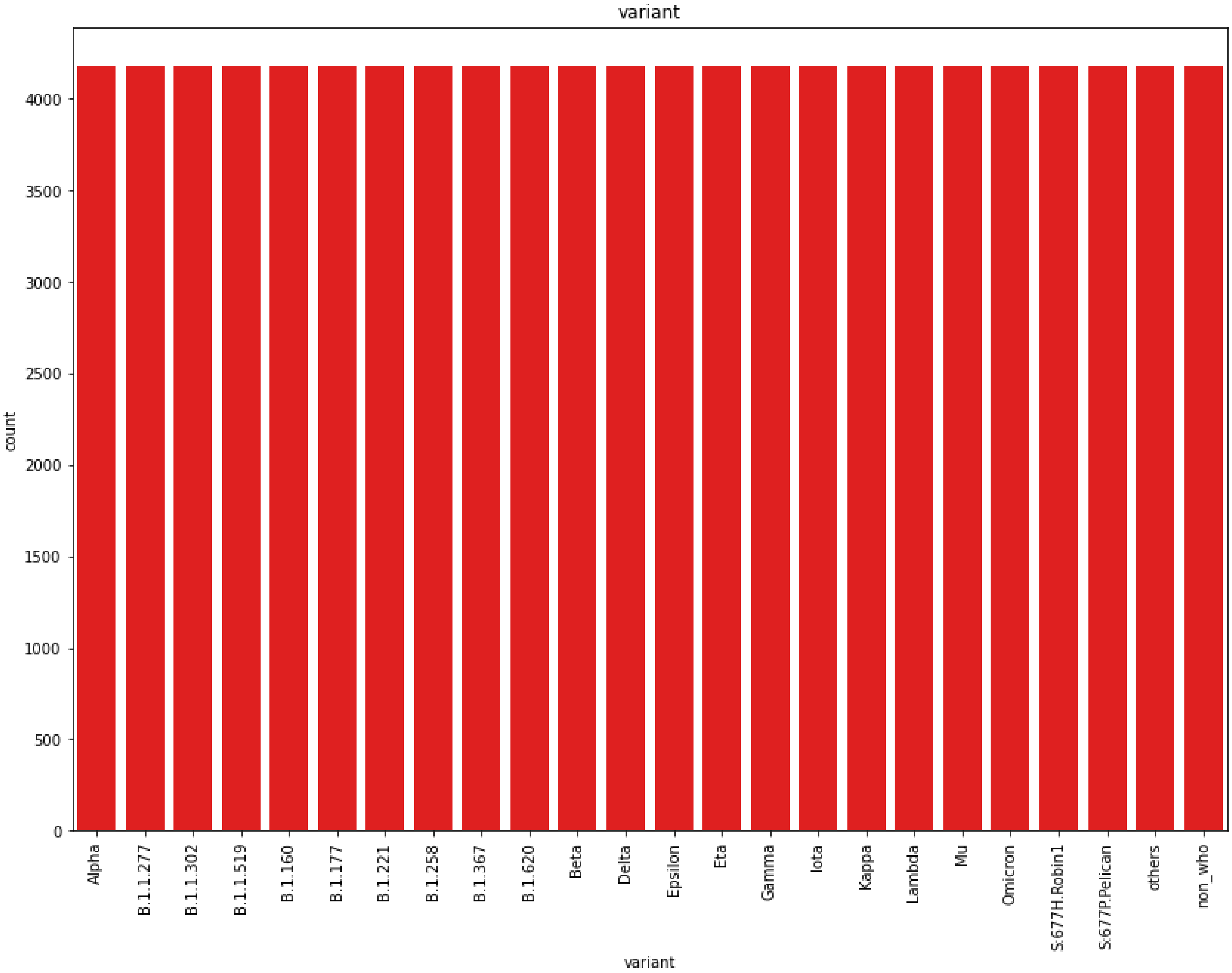


In [11]:

```
plt.figure(figsize=(14,10))
plt.xticks(rotation=90)
sns.countplot(data=cvd,x='variant', color='Red')
plt.title('variant')
```

Out[11]:

```
Text(0.5, 1.0, 'variant')
```



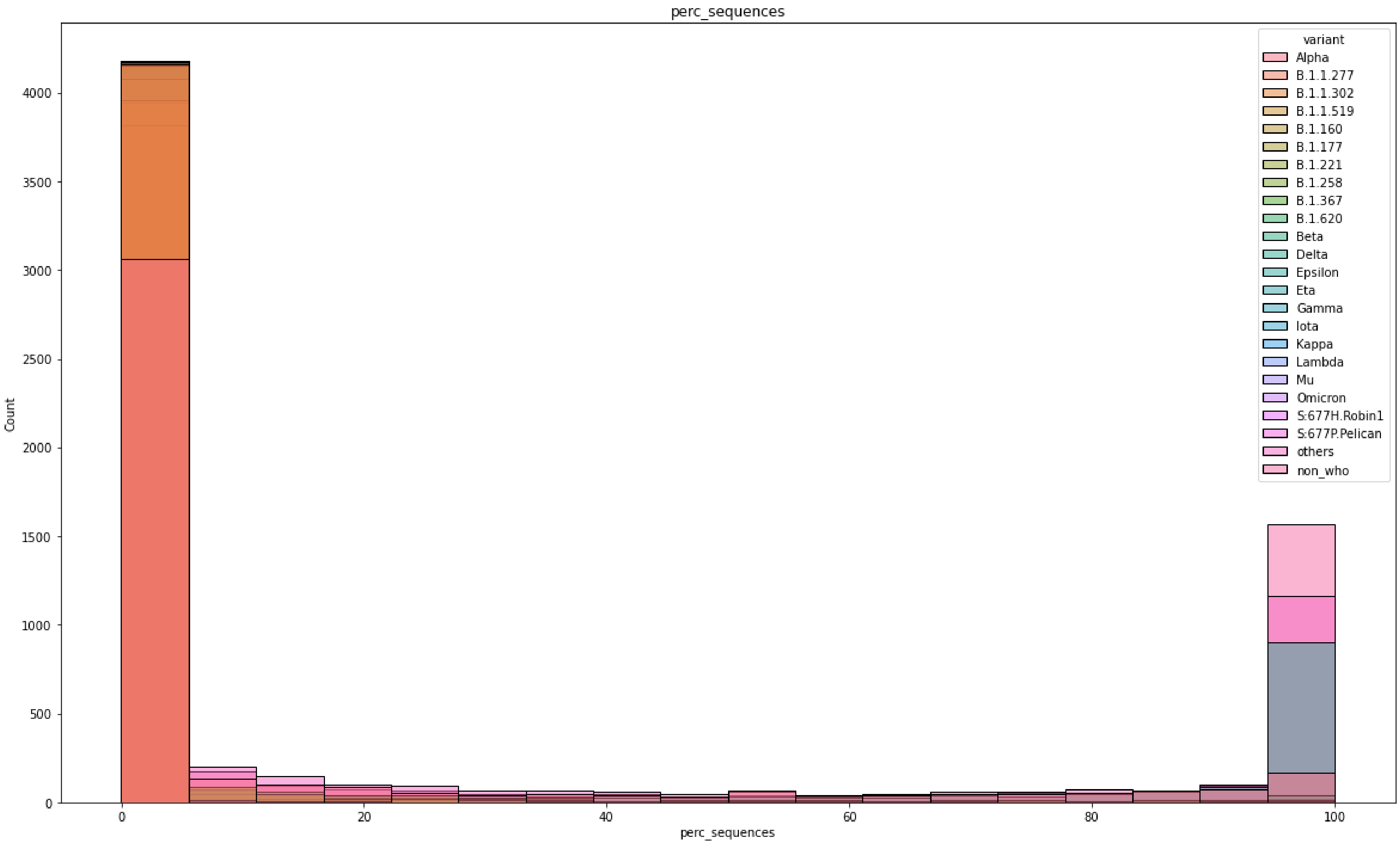
In [12]:

```
plt.figure(figsize=(20,12))
```

```
sns.histplot(data=cvd,x='perc_sequences',hue='variant')
plt.title('perc_sequences')
```

Out[12]:

```
Text(0.5, 1.0, 'perc_sequences')
```



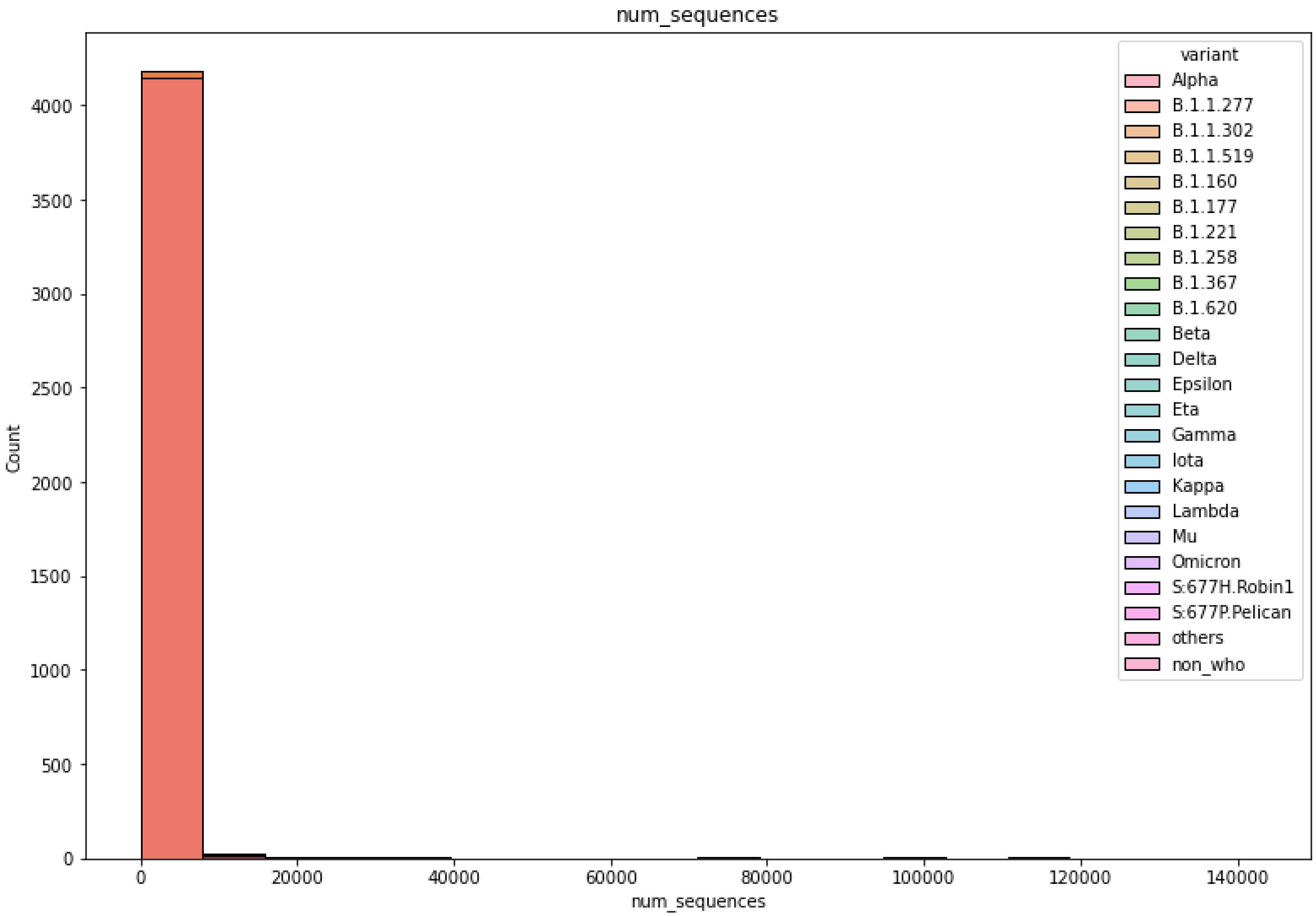
In [13]:

```
plt.figure(figsize=(13,9))
sns.histplot(data=cvd,x='num_sequences',hue='variant')
```

```
plt.title('num_sequences')
```

Out[13]:

```
Text(0.5, 1.0, 'num_sequences')
```



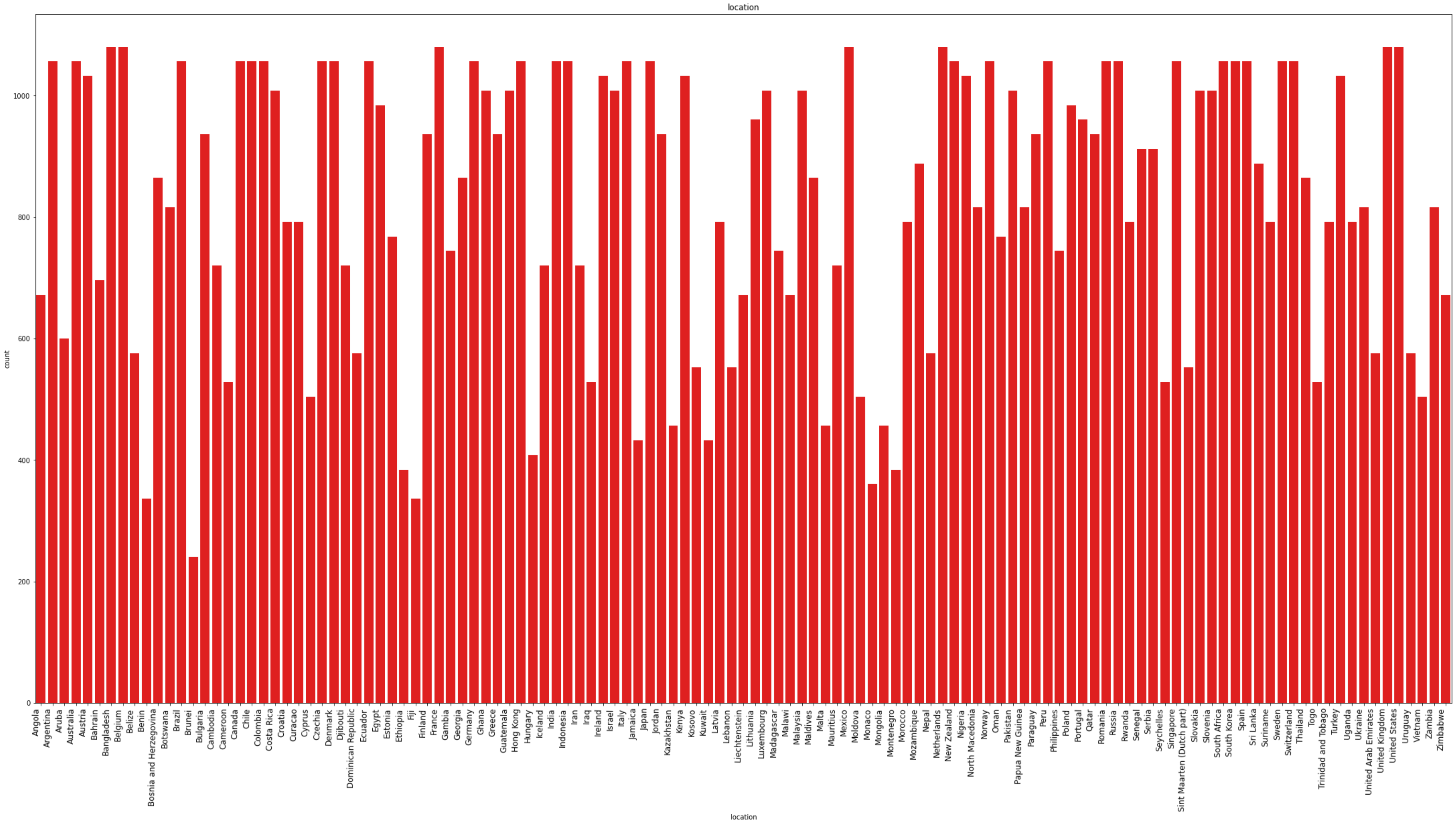
In [14]:

```
plt.figure(figsize=(30,15))
plt.xticks(rotation=89, horizontalalignment='right', fontsize=12)
plt.tight_layout()
#sns.histplot(data=cvd,x='location',bins=50)
```

```
#sns.set_context('notebook')
sns.countplot(data=cvd,x='location',color='Red')

plt.title('location')
```

Out[14]:

```
Text(0.5, 1.0, 'location')
```



In [37]:

```
#cvd['start date'] =pd.to_datatime(cvd['start date'])
```

```
#cvd['date'] = pd.to_datetime(cvd['date'], errors='ignore')
cvd
```

Out[37]:

```
        location      variant  num_sequences  perc_sequences  num_sequences_total  month  year  day
     0    Angola        Alpha              0             0.0                    3      7  2020    6
     1    Angola    B.1.1.277              0             0.0                    3      7  2020    6
     2    Angola    B.1.1.302              0             0.0                    3      7  2020    6
     3    Angola    B.1.1.519              0             0.0                    3      7  2020    6
     4    Angola      B.1.160              0             0.0                    3      7  2020    6
   ...       ...          ...            ...             ...                  ...    ...   ...  ...
100411  Zimbabwe      Omicron              0             0.0                    6     11  2021    1
100412  Zimbabwe  S:677H.Robin1            0             0.0                    6     11  2021    1
100413  Zimbabwe  S:677P.Pelican           0             0.0                    6     11  2021    1
100414  Zimbabwe       others              0             0.0                    6     11  2021    1
100415  Zimbabwe      non_who              0             0.0                    6     11  2021    1

100416 rows × 8 columns
```

In [48]:

```
# Seprate date with apply function

#cvd['month'] = cvd['date'].apply(lambda date: date.month)
#cvd['year'] = cvd['date'].apply(lambda date: date.year)
#cvd['day'] = cvd['date'].apply(lambda date: date.day)



# We will drop date column as we don't need keep it in our dataframe

#cvd.drop('date',axis=1, inplace=True)
cvd.head()
```

Out[48]:

	location	variant	num_sequences	perc_sequences	num_sequences_total	month	year	day
0	Angola	Alpha	0	0.0	3	7	2020	6
1	Angola	B.1.1.277	0	0.0	3	7	2020	6
2	Angola	B.1.1.302	0	0.0	3	7	2020	6
3	Angola	B.1.1.519	0	0.0	3	7	2020	6
4	Angola	B.1.160	0	0.0	3	7	2020	6

In [47]:

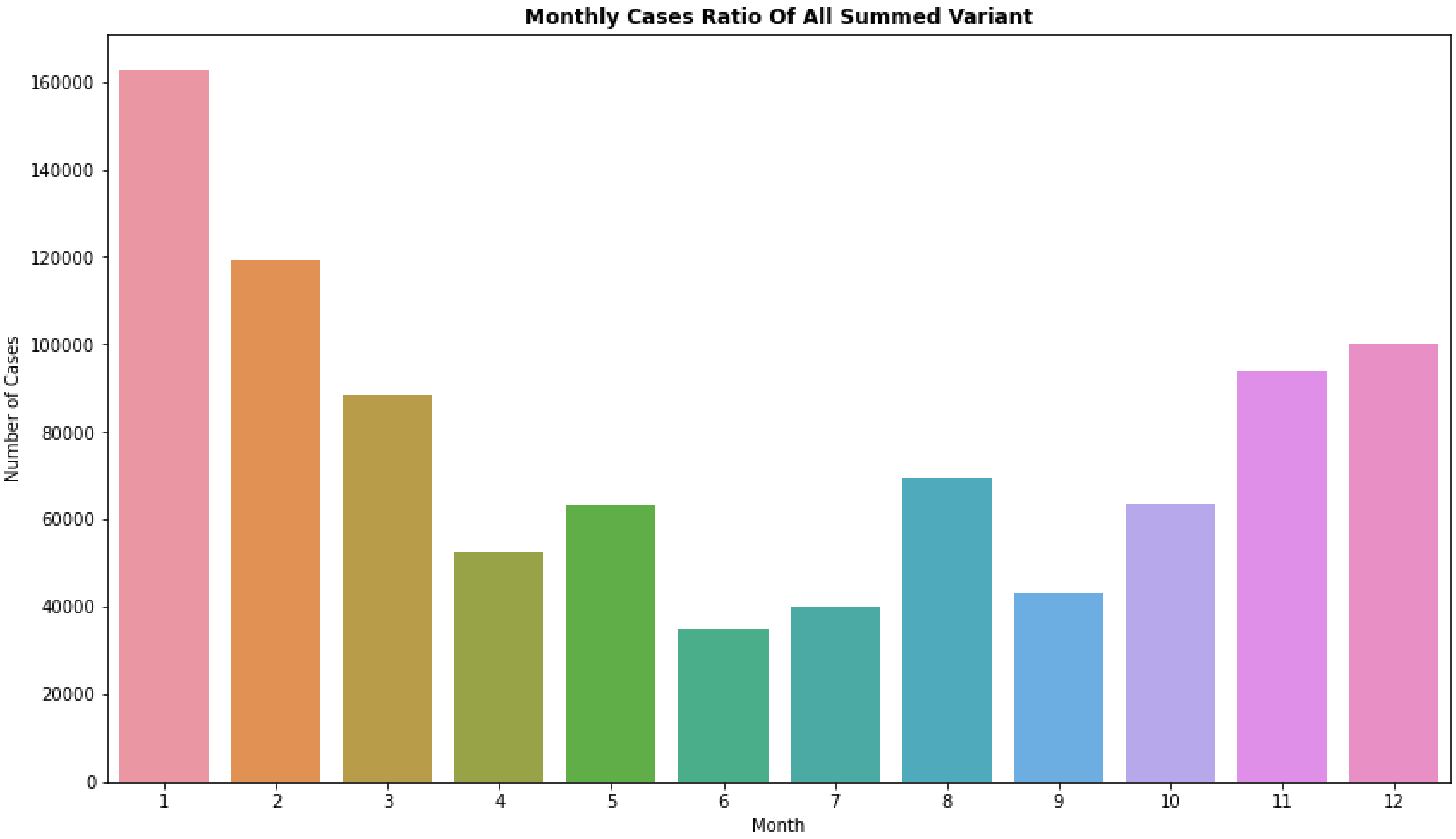
```
cvd.head()
```

Out[47]:

	location	variant	num_sequences	perc_sequences	num_sequences_total	month	year	day
0	Angola	Alpha	0	0.0	3	7	2020	6
1	Angola	B.1.1.277	0	0.0	3	7	2020	6
2	Angola	B.1.1.302	0	0.0	3	7	2020	6
3	Angola	B.1.1.519	0	0.0	3	7	2020	6
4	Angola	B.1.160	0	0.0	3	7	2020	6

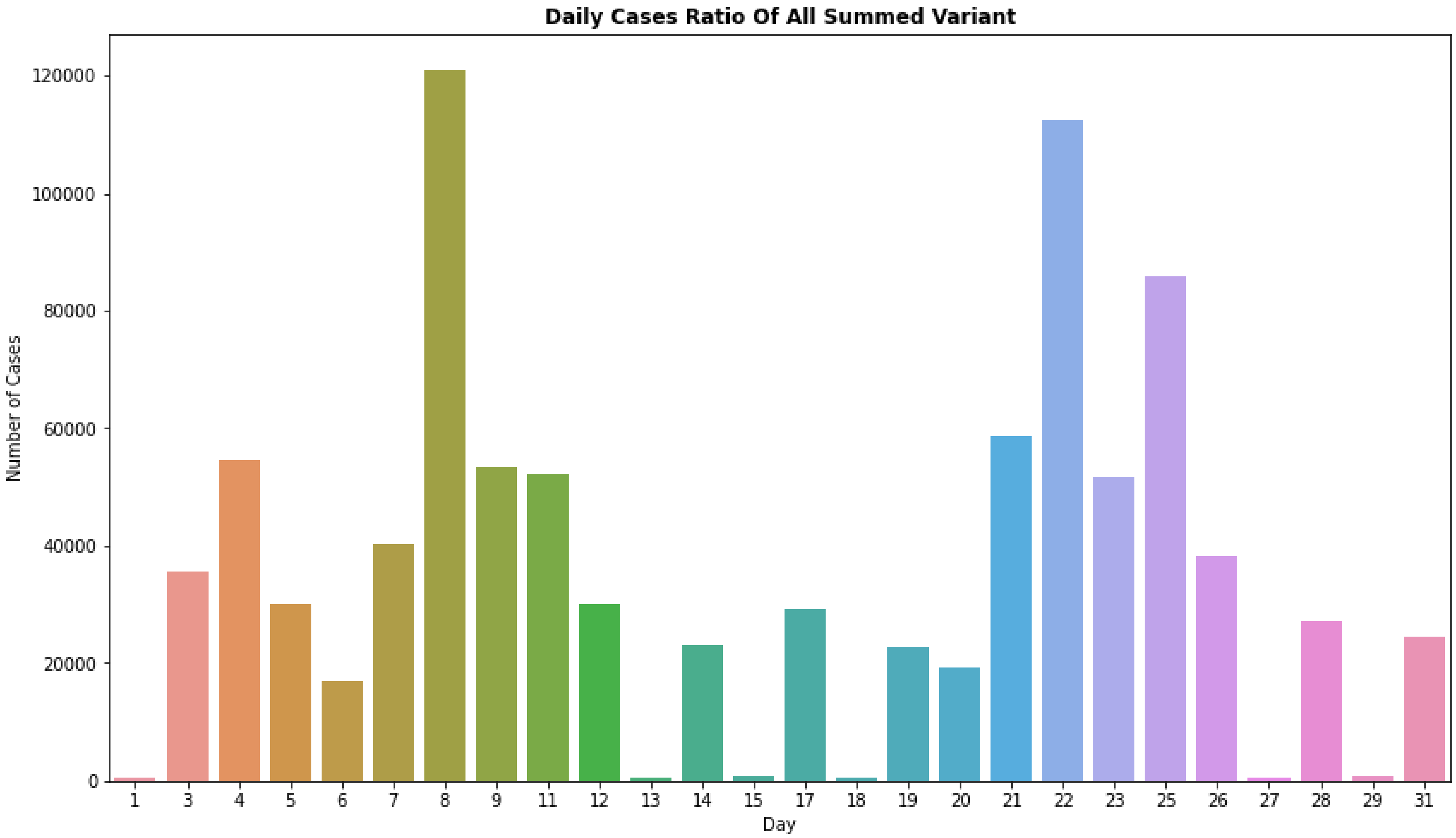
In [50]:

```
# Let's check all summed variant with montly ratio

cvd_val1 = cvd.loc[cvd["variant"]== virus].groupby('month')['num_sequences'].agg('sum').sort_values(ascending=False)
cvd_val1 = pd.DataFrame({'Month':cvd_val1.index, 'Number of Cases':cvd_val1.values})

plt.figure(figsize=(14,8))
sns.barplot(x='Month', y='Number of Cases',data=cvd_val1);
plt.title('Monthly Cases Ratio Of All Summed Variant',fontweight="bold");
```



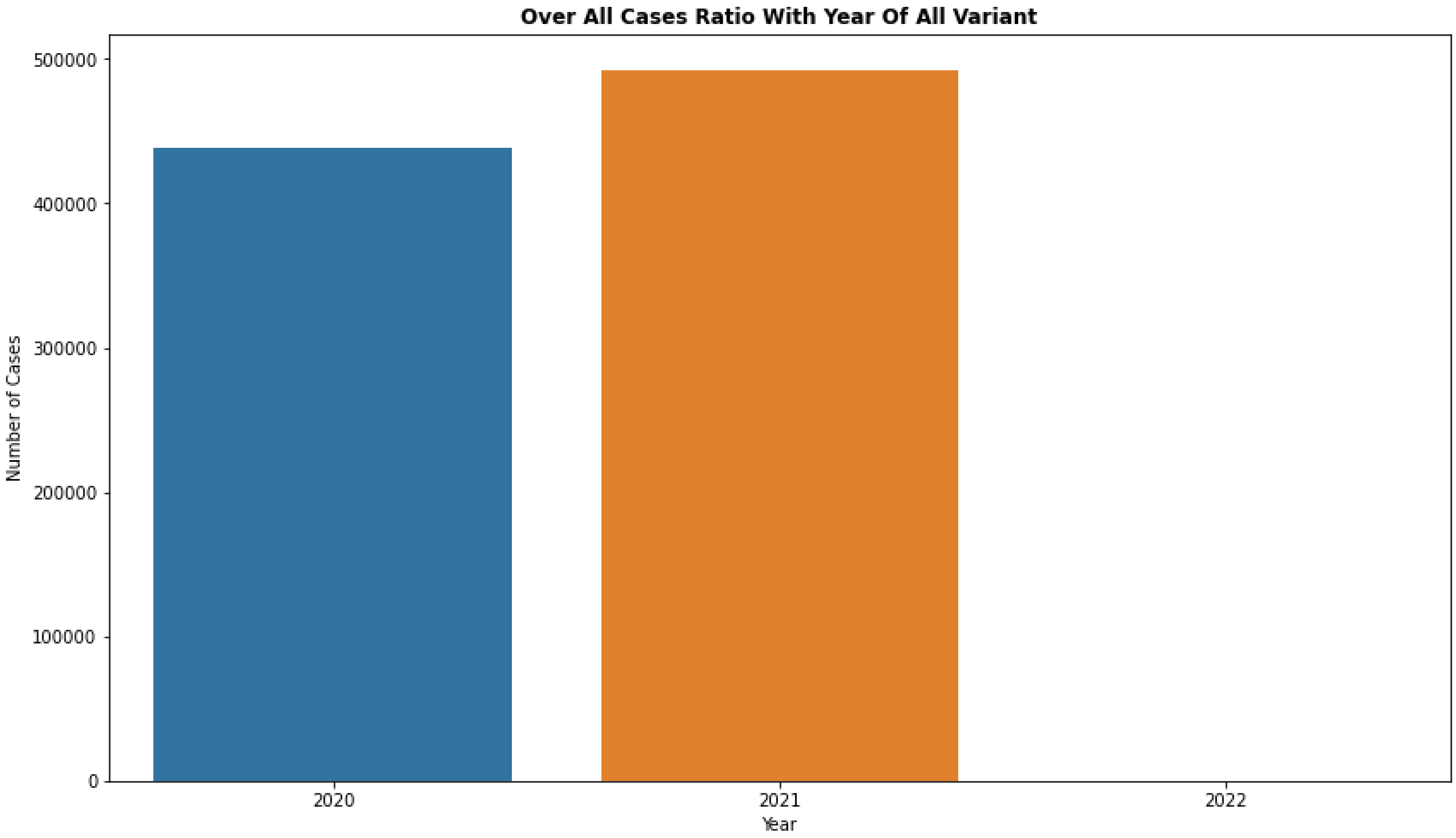
In [52]:

```
# Let's check all summed variant with daily ratio

cvd_val1 = cvd.loc[cvd["variant"]== virus].groupby('day')['num_sequences'].agg('sum').sort_values(ascending=False)
cvd_val1 = pd.DataFrame({'Day':cvd_val1.index, 'Number of Cases':cvd_val1.values})

plt.figure(figsize=(14,8))
sns.barplot(x='Day', y='Number of Cases',data=cvd_val1);
plt.title('Daily Cases Ratio Of All Summed Variant',fontweight="bold");
```



In [54]:

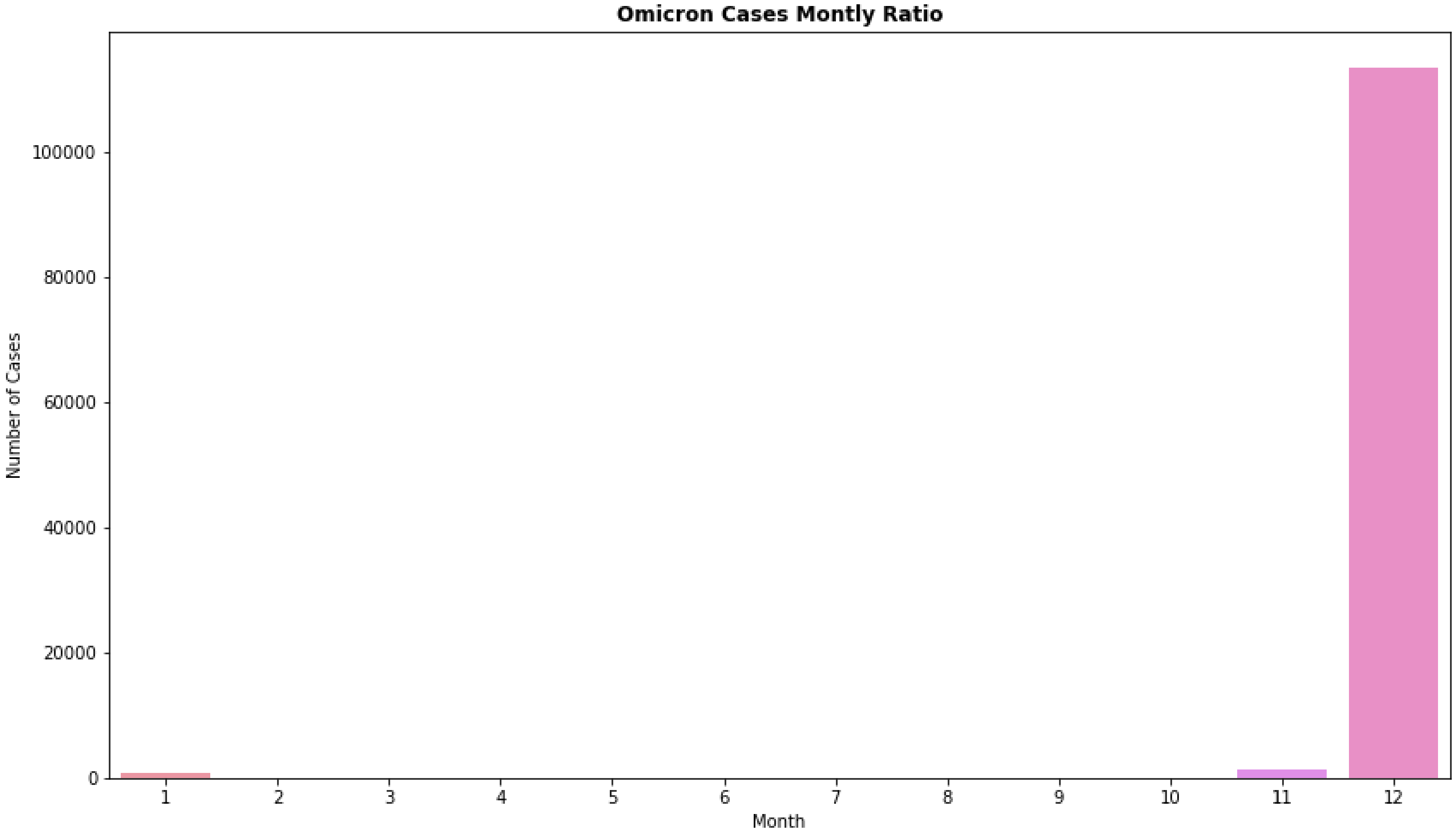
```
# Let's check all summed variant with yearly ratio

cvd_val1 = cvd.loc[cvd["variant"]== virus].groupby('year')['num_sequences'].agg('sum').sort_values(ascending=False)
cvd_val1 = pd.DataFrame({'Year':cvd_val1.index, 'Number of Cases':cvd_val1.values})

plt.figure(figsize=(14,8))
sns.barplot(x='Year', y='Number of Cases',data=cvd_val1);
plt.title('Over All Cases Ratio With Year Of All Variant',fontweight="bold");
```



Now Let's check the Omicron variant

In [57]:

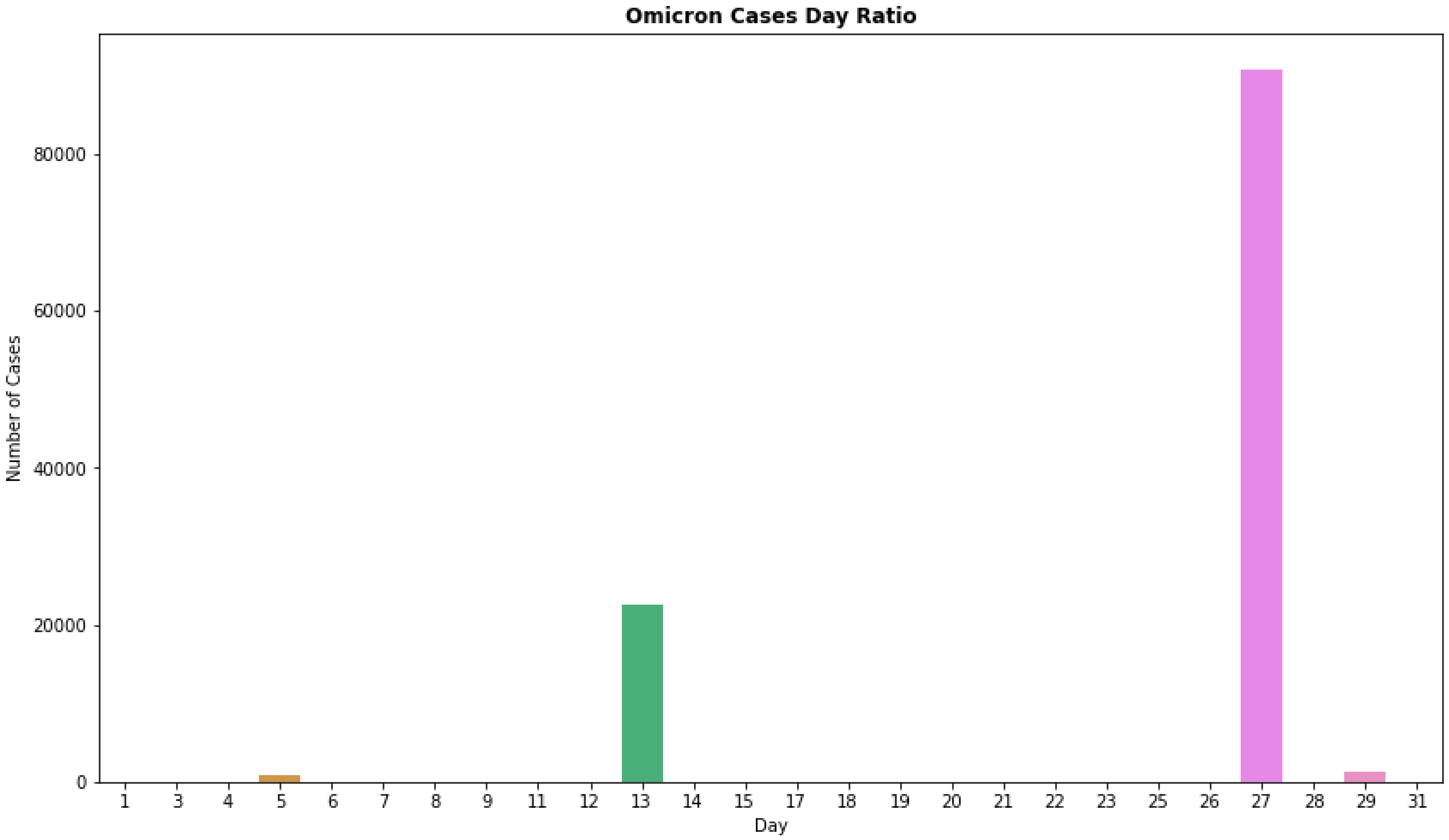
```
cvd_val1 = cvd.loc[cvd["variant"]== 'Omicron'].groupby('month')['num_sequences'].agg('sum').sort_values(ascending=False)
cvd_val1 = pd.DataFrame({'Month':cvd_val1.index, 'Number of Cases':cvd_val1.values})

plt.figure(figsize=(14,8))
sns.barplot(x='Month', y='Number of Cases',data=cvd_val1);
plt.title('Omicron Cases Montly Ratio',fontweight="bold");
```



Days

In [58]:
```
cvd_val1 = cvd.loc[cvd["variant"]== 'Omicron'].groupby('day')['num_sequences'].agg('sum').sort_values(ascending=False)
cvd_val1 = pd.DataFrame({'Day':cvd_val1.index, 'Number of Cases':cvd_val1.values})

plt.figure(figsize=(14,8))
sns.barplot(x='Day', y='Number of Cases',data=cvd_val1);
plt.title('Omicron Cases Day Ratio',fontweight="bold")
```

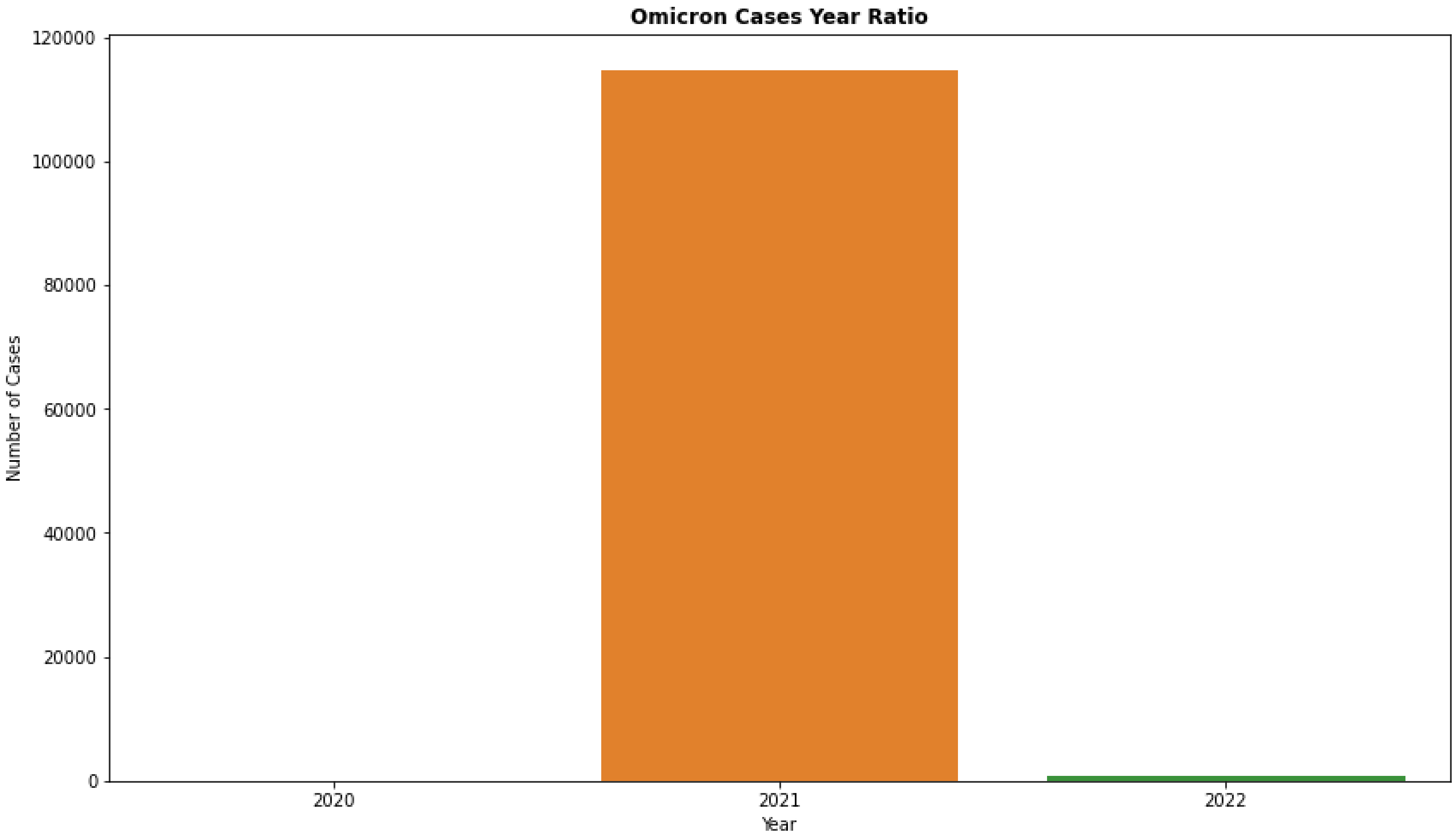
Out[58]:
```
Text(0.5, 1.0, 'Omicron Cases Day Ratio')
```



year

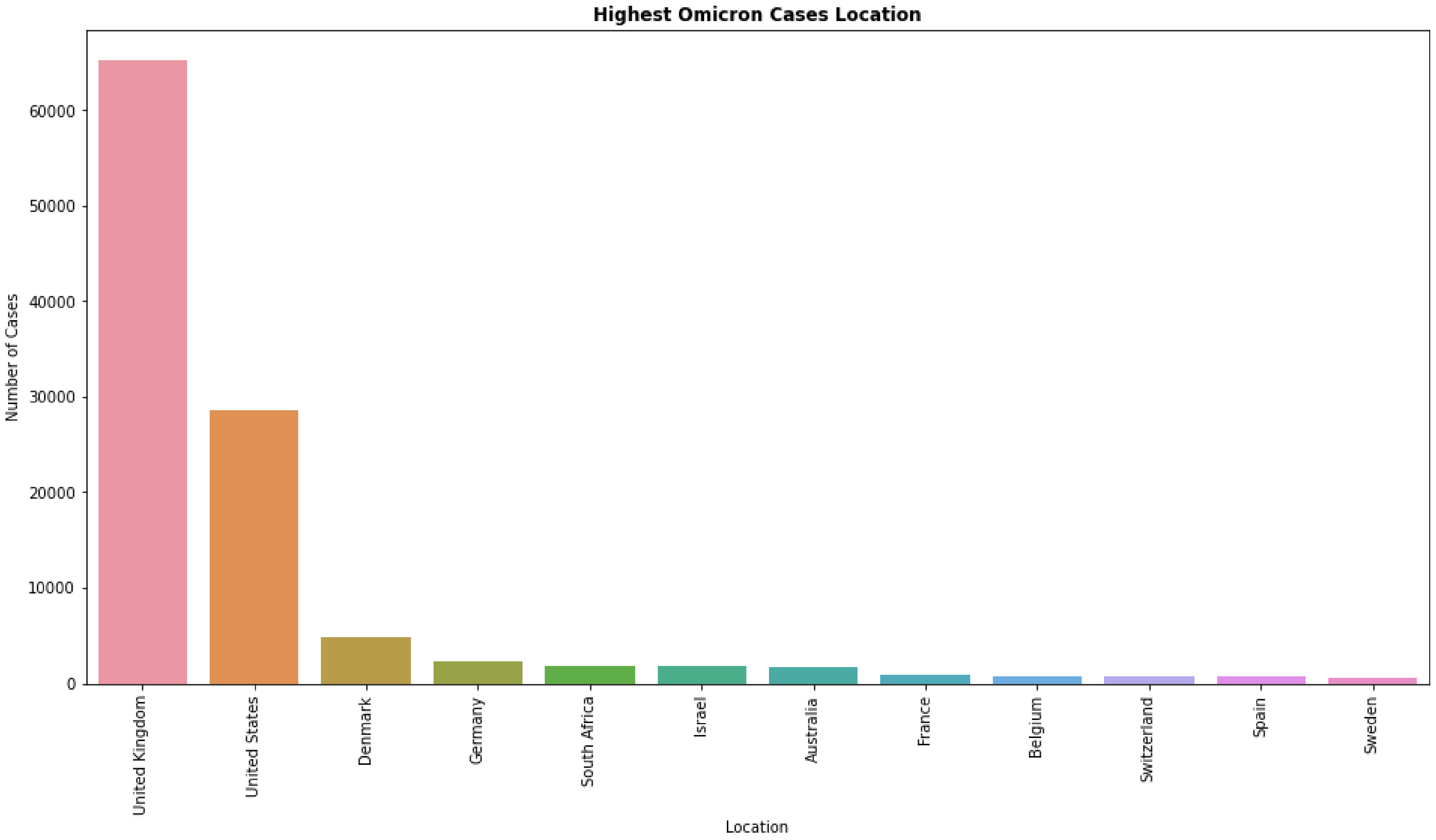
```
In [60]:
cvd_val1 = cvd.loc[cvd["variant"]== 'Omicron'].groupby('year')['num_sequences'].agg('sum').sort_values(ascending=False)
cvd_val1 = pd.DataFrame({'Year':cvd_val1.index, 'Number of Cases':cvd_val1.values})

plt.figure(figsize=(14,8))
sns.barplot(x='Year', y='Number of Cases',data=cvd_val1);
plt.title('Omicron Cases Year Ratio',fontweight="bold")
```

```
Out[60]: Text(0.5, 1.0, 'Omicron Cases Year Ratio')
```



In [66]:

```
cvd_val1 = cvd.loc[cvd["variant"]== 'Omicron'].groupby('location')['num_sequences'].agg('sum').sort_values(ascending=False)[:12]
cvd_val1 = pd.DataFrame({'Location':cvd_val1.index, 'Number of Cases':cvd_val1.values})

plt.figure(figsize=(16,8))
sns.barplot(x='Location', y='Number of Cases',data=cvd_val1);
plt.title('Highest Omicron Cases Location',fontweight="bold");
plt.xticks(rotation=90);
```


Highest Omicron Cases Location
Number of Cases
0
10000
20000
30000
40000
50000
60000
United Kingdom
United States
Denmark
Germany
South Africa
Israel
Australia
France
Belgium
Switzerland
Spain
Sweden
Location