Which countries are the most frequent sources of attack attempts?

After cleaning up the data, this question was relatively straightforward to answer, but very interesting nonetheless. My first move was to list out all of the countries within the dataset, in which Jupyter gave me the following:

```
array(['China', 'South Korea', 'United States', 'Singapore', 'Russia',
       'Moldova', 'Indonesia', 'Netherlands', 'Morocco', 'Ukraine',
       'Turkey', 'Germany', 'Egypt', 'India', 'Pakistan',
       'United Arab Emirates', 'Kenya', 'Finland', 'Thailand', 'Syria',
       'Japan', 'Hong Kong', 'Brazil', 'France', 'United Kingdom',
       'Vietnam', 'Canada', 'Chile', 'Belgium', 'Switzerland'],
      dtype=object)
```

I got this from a very simple command, being `df['country'].unique()` which looks into the dataframe 'df' where the 'country' column is with the '.unique()' signifying to display every value within the column. Although this is a nice command that gives us an idea of what we're working with, we could expand on this a bit more by using some more complex code.

The next line of code is `country_sizes = df.groupby('country').size()` which assigns the variable, `country_sizes` to the function `df.groupby('country').size()` which groups all the values specific to each country and assigns them to the country. Using our `country_sizes` variable, we can arrange it display in descending order by extending it to `country_sizes.sort_values(ascending = False)` which displays the following:

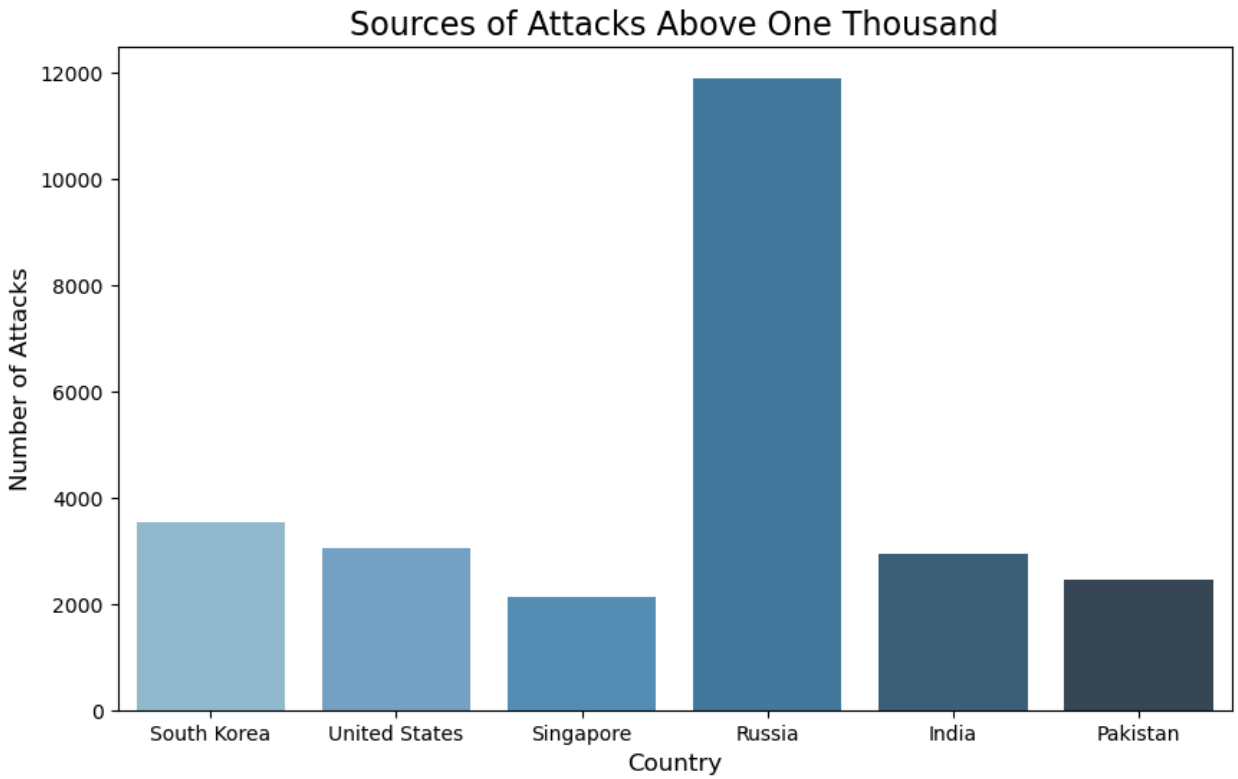
```
country
Russia                11904
South Korea            3528
United States          3049
India                  2941
Pakistan               2449
Singapore              2118
Turkey                  846
Moldova                 729
Ukraine                 562
Egypt                   520
Indonesia               450
Syria                   187
Chile                   171
Thailand                164
China                   156
Japan                   103
Vietnam                  58
Kenya                    19
Morocco                  12
Germany                  11
Canada                    6
Hong Kong                 5
Netherlands               3
Brazil                    2
United Kingdom            2
Switzerland               1
France                    1
Finland                   1
United Arab Emirates      1
Belgium                   1
dtype: int64
```

Now it is obvious to see that Russia has performed the most attacks, but maybe we could still gain some insight from a simple visualization of how much Russia is really in the lead by comparison to the other countries with values over 1000.

To make this, I first put the six highest countries into an arraylist called `selected_countries` and made a duplicate dataframe called `Q1_d` which contains the line `df[df['country'].isin(selected_countries)]`, which only ports over the information from the countries listed in the array created before. Now, to make the graph we enter the following:

```
plt.figure(figsize=(10, 6))
sns.countplot(x = 'country', data=Q1_df, palette = 'Blues_d')
plt.title('Sources of Attacks Above One Thousand', fontsize=16)
plt.xlabel('Country', fontsize=12)
plt.ylabel('Number of Attacks', fontsize=12)
```

With that, we get the following visualization of our information:



We can really see how much more attacks came out of Russia compared to its competition. Even with South Korea in second place they are barely under the 4,000 mark, looking like they are competing with the United States. Singapore was the lowest, but is easily over half the size of South Korea which really shows how close they all are to each other.

I know this analysis was very basic and straightforward but I really want to take things slowly and gain more and more competence over time, as I can tell this area of study can get very complicated despite being fairly simple at the moment. Future questions will be more complex and in depth.