FDA_Countries_analysis

September 24, 2018

0.0.1 Table of content

- Clinical trials by country distribution (pie chart)
- Clinical trials by country distribution (world map)

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
# %matplotlib notebook
%matplotlib inline
```

In [2]:
```
countries_full_df = pd.read_csv(
    r"c:\Dev\04. Python\03. XML converter of FDA list\goodDB\04. 2018Sep17_13-28-24\FDA
```

In [3]:
```
#Data cleaning
countries_full_df['country'][countries_full_df['country'] == "United States"] = "Unite
countries_full_df['country'][countries_full_df['country'] == "Congo, The Democratic Re
countries_full_df['country'][countries_full_df['country'] == "Czech Republic"] = "Czec
```

In [4]:
```
#This is to look for countries which may not be present in the dataset
# westernSakhara_DF = pd.DataFrame(countries_full_df[countries_full_df["country"].str.
# westernSakhara_DF.groupby("country").count()
```

In [5]:
```
counted_countries = countries_full_df.groupby('country').count()
```

In [6]:
```
%run -i create_uniform_buckets.py
bucketsArray = createUniformBucketsFromSeries(pd.Series(counted_countries.nct_id), numb
```

```
Max value:114420
Min value:1
Series Length = 203
Ideal bucket len:25
Bucket value counts:[27, 25, 24, 25, 24, 24, 24, 24, 6]
```

In [7]:
```
sorted_counted_countries = pd.DataFrame(counted_countries.sort_values(by="nct_id", asc
#Have to make index correct - so it is not resolved to United States Minor Outlying Is
#In dataset it is just "United States"
#http://cmdlinetips.com/2018/03/how-to-change-column-names-and-row-indexes-in-pandas/
```

```
        first_index = sorted_counted_countries.index[0]
        print(first_index)
        sorted_counted_countries.rename(index={first_index:"United States of America"}, inplace

        maxCountryCount = sorted_counted_countries.nct_id.max()
        print("max:"+str(maxCountryCount))
        minCountryCount = sorted_counted_countries.nct_id.min()
        print("min:"+str(minCountryCount))
        print(str(len(sorted_counted_countries.groupby("nct_id"))))
        # sorted_counted_countries.head()
```

```
United States of America
max:114420
min:1
144
```

1 Clinical trials by country distribution

Table of content

```
In [20]: FILTER_THRESHOLD = 10000
```

```
        from __future__ import print_function
        from ipywidgets import interact, interactive, fixed, interact_manual
        import ipywidgets as widgets

        COUNTRY_LABEL_FONT_SIZE = 8

        def updateGraphBasedOnThreshold(threshold):
            print("Threshold:{}".format(threshold))

            FILTER_THRESHOLD = threshold

            sum_of_all_countries = sorted_counted_countries.nct_id.sum()
            print("Sum of all countries:{}".format(sum_of_all_countries))

            sum_of_none = sorted_counted_countries.loc["**None**", "nct_id"]
            print("Sum of none:{}".format(sum_of_none))
            sum_of_all_countries_below_threshold = sorted_counted_countries[sorted_counted_cou
            print("sum_of_all_countries_below_threshold:{}".format(sum_of_all_countries_below_

            other_name = "Other(<"+str(FILTER_THRESHOLD)+" studies per country)"
            countryListWithNone = pd.DataFrame(sorted_counted_countries[sorted_counted_countri
            countryListWithNone.loc[other_name] = sum_of_all_countries - countryListWithNone.n

            countryListWithoutNone = pd.DataFrame(countryListWithNone[countryListWithNone.inde
```

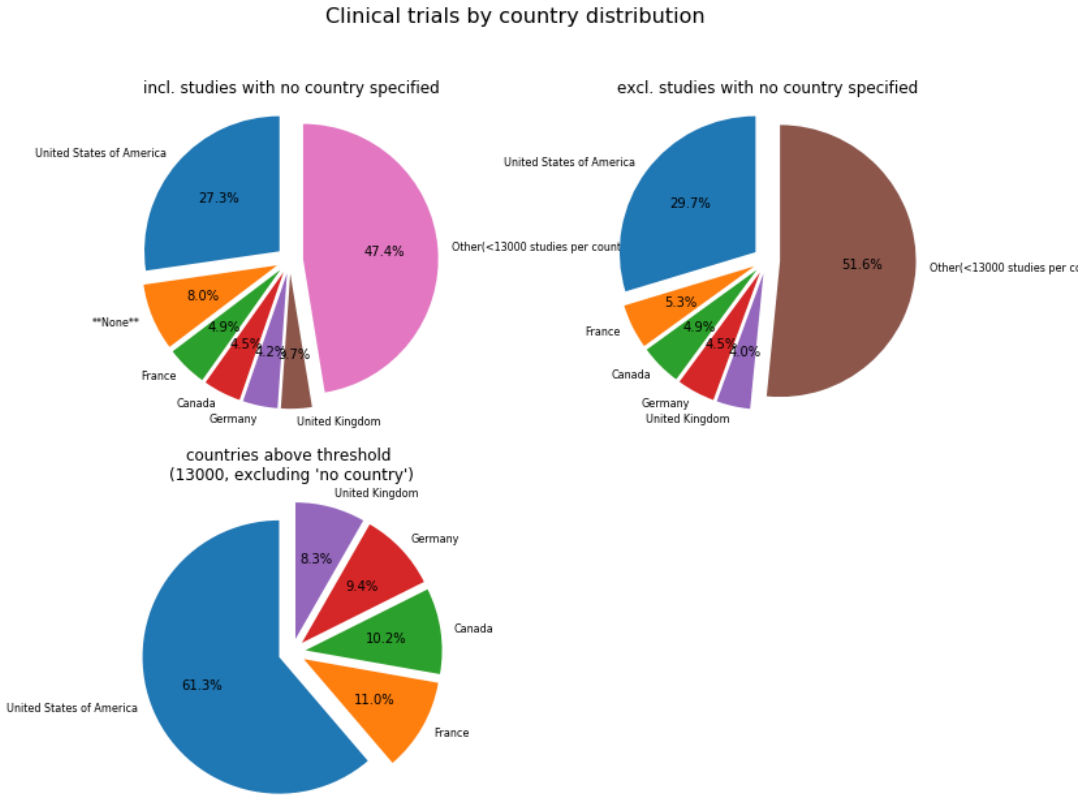
```
countryListWithoutNoneWithouOther = pd.DataFrame(countryListWithoutNone[countryLis


#PLOTTING

fig = plt.figure(figsize=(12, 10))
fig.suptitle("Clinical trials by country distribution", fontsize=16)
plt.subplot(221)
patches, texts, autotexts = plt.pie(countryListWithNone.nct_id,
        explode = list(np.zeros(len(countryListWithNone))+0.1),
        labels=countryListWithNone.index,
        autopct='%1.1f%%',
        shadow=False, startangle=90)

for txt in texts:
    txt.set_fontsize(COUNTRY_LABEL_FONT_SIZE)
plt.title("incl. studies with no country specified")
plt.axis('equal')


plt.subplot(222)
plt.title("excl. studies with no country specified")
patches, texts, autotexts = plt.pie(countryListWithoutNone.nct_id,
        explode = list(np.zeros(len(countryListWithoutNone))+0.1),
        labels=countryListWithoutNone.index,
        autopct='%1.1f%%',
        shadow=False, startangle=90)
for txt in texts:
    txt.set_fontsize(COUNTRY_LABEL_FONT_SIZE)
plt.axis('equal')


plt.subplot(223)
patches, texts, autotexts = plt.pie(countryListWithoutNoneWithouOther.nct_id,
        explode = list(np.zeros(len(countryListWithoutNoneWithouOther))+0.1),
        labels=countryListWithoutNoneWithouOther.index,
        autopct='%1.1f%%',
        shadow=False, startangle=90)
for txt in texts:
    txt.set_fontsize(COUNTRY_LABEL_FONT_SIZE)
plt.axis('equal')
plt.title("countries above threshold \n({}, excluding 'no country')".format(FILTE

plt.show()

fig.savefig("pieCharts.png")

return None
```

```
        interact(updateGraphBasedOnThreshold, threshold=widgets.IntSlider(min=0,max=60000,ste
        # updateGraphBasedOnThreshold(10000)
```

```
interactive(children=(IntSlider(value=20000, description='threshold', max=60000, step=1000), O
```

In [22]:
```
from IPython.display import Image
Image("pieCharts.png")
```

Out[22]:



In [9]:
```
%run country_codes_conversion.py
for row in sorted_counted_countries.index:
    iso2digitCode = get2ISOcodeFromCountryName(row)
    sorted_counted_countries.loc[row, "iso2"] = str(iso2digitCode)
pass
```

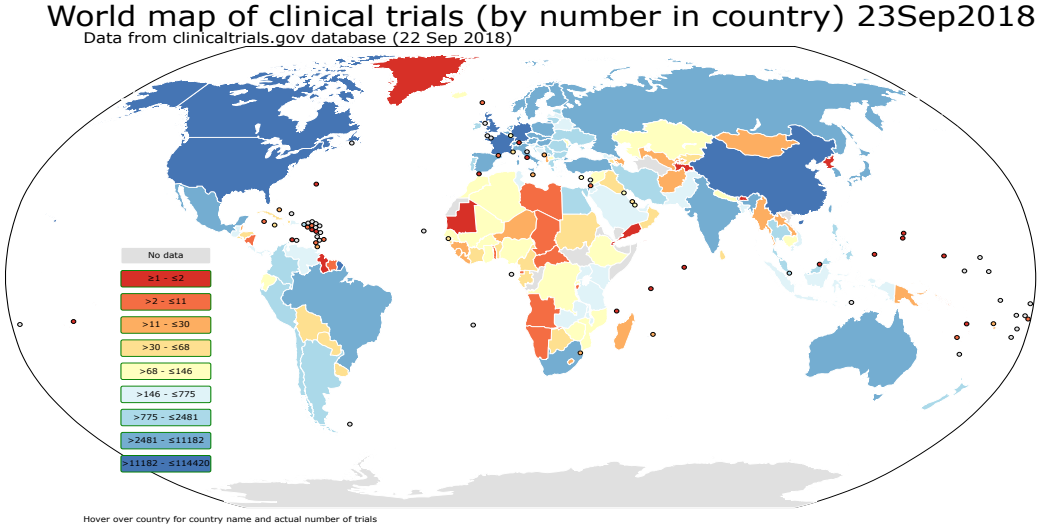
In [10]:
```
sorted_counted_countries.loc[sorted_counted_countries.iso2=='CZ']
```

```
Out[10]: Empty DataFrame
         Columns: [nct_id, iso2]
         Index: []

In [11]: iso2_check_for_duplicates = pd.DataFrame(sorted_counted_countries.groupby(by="iso2").
         #All should be 1s (no 2s - those are duplicates)
         iso2_check_for_duplicates.head()

Out[11]:                                        nct_id
         iso2
         ad                                          1
         no_code_former serbia and montenegro        1
         my                                          1
         mz                                          1
         na                                          1

In [12]: exportDF = pd.DataFrame(sorted_counted_countries)
         #CLEANING BEFORE EXPORTING TO JSON
         exportDF["full_country_name"] = exportDF.index
         exportDF.rename(columns={"nct_id":"nct_id_count"}, inplace=True)

In [13]: #https://stackoverflow.com/questions/32714783/ipython-run-all-cells-below-from-a-widg
         from IPython.display import Javascript, display
         def run_cells_below():
             display(Javascript('IPython.notebook.execute_cells_below()'))

In [14]: %run -i transformSVG.py

         svgConverted = transformCountrySVG(exportDF, bucketsArray,  \
                                            [ "#d73027", "#f46d43", "#fdae61", "#fee090", "#fff
                                              "#abd9e9", "#74add1", "#4575b4"]  \
                         )
         run_cells_below()

<IPython.core.display.Javascript object>
```

2 Clinical trials by country (Sep 2018)

Table of content

```
In [16]: from IPython.core.display import display, HTML, DisplayObject
         from IPython.display import SVG
         SVG(svgConverted)

   Out[16]:
```


World map of clinical trials (by number in country) 23Sep2018
Data from clinicaltrials.gov database (22 Sep 2018)
No data
≥1 - ≤2
>2 - ≤11
>11 - ≤30
>30 - ≤68
>68 - ≤146
>146 - ≤775
>775 - ≤2481
>2481 - ≤11182
>11182 - ≤114420
Hover over country for country name and actual number of trials