# Assignment 3

All questions are weighted the same in this assignment. This assignment requires more individual learning then the last one did - you are encouraged to check out the pandas documentation to find functions or methods you might not have used yet, or ask questions on Stack Overflow and tag them as pandas and python related. All questions are worth the same number of points except question 1 which is worth 17% of the assignment grade.

**Note**: Questions 2-13 rely on your question 1 answer.

In [1]:
```
import pandas as pd
import numpy as np

# Filter all warnings. If you would like to see the warnings, please comment the two lines below.
import warnings
warnings.filterwarnings('ignore')
```

## Question 1

Load the energy data from the file `assets/Energy Indicators.xls`, which is a list of indicators of energy supply and renewable electricity production from the United Nations for the year 2013, and should be put into a DataFrame with the variable name of **Energy**.

Keep in mind that this is an Excel file, and not a comma separated values file. Also, make sure to exclude the footer and header information from the datafile. The first two columns are unneccessary, so you should get rid of them, and you should change the column labels so that the columns are:

`['Country', 'Energy Supply', 'Energy Supply per Capita', '% Renewable']`

Convert `Energy Supply` to gigajoules (**Note: there are 1,000,000 gigajoules in a petajoule**). For all countries which have missing data (e.g. data with "...") make sure this is reflected as `np.NaN` values.

Rename the following list of countries (for use in later questions):

```
"Republic of Korea": "South Korea",
"United States of America": "United States",
"United Kingdom of Great Britain and Northern Ireland": "United Kingdom",
"China, Hong Kong Special Administrative Region": "Hong Kong"
```

There are also several countries with parenthesis in their name. Be sure to remove these, e.g. `'Bolivia (Plurinational State of)'` should be `'Bolivia'`.

Next, load the GDP data from the file `assets/world_bank.csv`, which is a csv containing countries' GDP from 1960 to 2015 from World Bank. Call this DataFrame **GDP**.

Make sure to skip the header, and rename the following list of countries:

```
"Korea, Rep.": "South Korea",
"Iran, Islamic Rep.": "Iran",
"Hong Kong SAR, China": "Hong Kong"
```

Finally, load the Sciamgo Journal and Country Rank data for Energy Engineering and Power Technology from the file `assets/scimagojr-3.xlsx`, which ranks countries based on their journal contributions in the aforementioned area. Call this DataFrame **ScimEn**.

Join the three datasets: GDP, Energy, and ScimEn into a new dataset (using the intersection of country names). Use only the last 10 years (2006-2015) of GDP data and only the top 15 countries by Scimagojr 'Rank' (Rank 1 through 15).

The index of this DataFrame should be the name of the country, and the columns should be ['Rank', 'Documents', 'Citable documents', 'Citations', 'Self-citations', 'Citations per document', 'H index', 'Energy Supply', 'Energy Supply per Capita', '% Renewable', '2006', '2007', '2008', '2009', '2010', '2011', '2012', '2013', '2014', '2015'].

*This function should return a DataFrame with 20 columns and 15 entries, and the rows of the DataFrame should be sorted by "Rank".*

In [8]:
```
def answer_one():
    #opening and cleaning the Energy DataFrame
    global Energy

    # Original read_excel is assets/Energy Indicators.xls
    Energy= pd.read_excel('assets/Energy Indicators.xls',
                          skiprows=17,
                          skipfooter=38,
                          na_values= "...",
                          usecols="C:F",
                          names =['Country', 'Energy Supply', 'Energy Supply per Capita', '% Renewable'])

    # Another way to drop the first two columns would have been: energy.drop(energy.columns[[0,1]],axis=1,inplace=True)
    Energy.dropna()


    Energy['Country'] = Energy['Country'].str.replace('Republic of Korea','South Korea')
    Energy['Country'] = Energy['Country'].str.replace('United States of America','United States')
    Energy['Country'] = Energy['Country'].str.replace('United Kingdom of Great Britain and Northern Ireland','United Kingdom'
    Energy['Country'] = Energy['Country'].str.replace('China, Hong Kong Special Administrative Region','Hong Kong')


    Energy['Energy Supply'] *= 1000000

    Energy['Country'] = Energy['Country'].str.replace(r" \(.*\)","")
    Energy['Country'] = Energy['Country'].str.replace(r"([0-9]+)$","")
    Energy['Country'] = Energy['Country'].str.strip()


    #opening and cleaning the GDP DataFrame
    global GDP
    GDP= pd.read_csv('assets/world_bank.csv', skiprows=4)

    GDP.rename(columns= {"Country Name": "Country"}, inplace=True)


    GDP['Country'] = GDP['Country'].str.replace('Korea, Rep.','South Korea')
    GDP['Country'] = GDP['Country'].str.replace('Iran, Islamic Rep.','Iran')
    GDP['Country'] = GDP['Country'].str.replace('Hong Kong SAR, China','Hong Kong')


    #opening the ScimEn DataFrame
    global ScimEn
    ScimEn= pd.read_excel('assets/scimagojr-3.xlsx')
```

```
    # Taking out the data from 2006-2015 from the GDP DataFrame
    global GDP_revised
    GDP_revised= GDP.iloc[:, [0,-10,-9,-8,-7,-6,-5,-4,-3,-2,-1]]
    GDP_revised.columns = GDP_revised.columns.astype(str).str.split('.').str[0]


    # Taking out rank 1-15 from the ScimEn DataFrame
    global ScimEn_revised
    ScimEn_revised= ScimEn.iloc[0:15, :]

    Top15 = ScimEn_revised.merge(Energy, how= 'left', on='Country').merge(GDP_revised, how= 'left', on='Country')
    Top15.set_index('Country', inplace=True)
    return Top15
    raise NotImplementedError()

answer_one()
```

Out[8]:

| Country | Rank | Documents | Citable documents | Citations | Self-citations | Citations per document | H index | Energy Supply | Energy Supply per Capita | % Renewable | 2006 | 2007 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| China | 1 | 127050 | 126767 | 597237 | 411683 | 4.70 | 138 | 1.271910e+11 | 93.0 | 19.754910 | 3.992331e+12 | 4.559041e+12 | 4.997775 |
| United States | 2 | 96661 | 94747 | 792274 | 265436 | 8.20 | 230 | 9.083800e+10 | 286.0 | 11.570980 | 1.479230e+13 | 1.505540e+13 | 1.501149 |
| Japan | 3 | 30504 | 30287 | 223024 | 61554 | 7.31 | 134 | 1.898400e+10 | 149.0 | 10.232820 | 5.496542e+12 | 5.617036e+12 | 5.558527 |
| United Kingdom | 4 | 20944 | 20357 | 206091 | 37874 | 9.84 | 139 | 7.920000e+09 | 124.0 | 10.600470 | 2.419631e+12 | 2.482203e+12 | 2.470614 |
| Russian Federation | 5 | 18534 | 18301 | 34266 | 12422 | 1.85 | 57 | 3.070900e+10 | 214.0 | 17.288680 | 1.385793e+12 | 1.504071e+12 | 1.583004 |
| Canada | 6 | 17899 | 17620 | 215003 | 40930 | 12.01 | 149 | 1.043100e+10 | 296.0 | 61.945430 | 1.564469e+12 | 1.596740e+12 | 1.612713 |
| Germany | 7 | 17027 | 16831 | 140566 | 27426 | 8.26 | 126 | 1.326100e+10 | 165.0 | 17.901530 | 3.332891e+12 | 3.441561e+12 | 3.478809 |
| India | 8 | 15005 | 14841 | 128763 | 37209 | 8.58 | 115 | 3.319500e+10 | 26.0 | 14.969080 | 1.265894e+12 | 1.374865e+12 | 1.428361 |
| France | 9 | 13153 | 12973 | 130632 | 28601 | 9.93 | 114 | 1.059700e+10 | 166.0 | 17.020280 | 2.607840e+12 | 2.669424e+12 | 2.674637 |
| South Korea | 10 | 11983 | 11923 | 114675 | 22595 | 9.57 | 104 | 1.100700e+10 | 221.0 | 2.279353 | 9.410199e+11 | 9.924316e+11 | 1.020510 |
| Italy | 11 | 10964 | 10794 | 111850 | 26661 | 10.20 | 106 | 6.530000e+09 | 109.0 | 33.667230 | 2.202170e+12 | 2.234627e+12 | 2.211154 |
| Spain | 12 | 9428 | 9330 | 123336 | 23964 | 13.08 | 115 | 4.923000e+09 | 106.0 | 37.968590 | 1.414823e+12 | 1.468146e+12 | 1.484530 |
| Iran | 13 | 8896 | 8819 | 57470 | 19125 | 6.46 | 72 | 9.172000e+09 | 119.0 | 5.707721 | 3.895523e+11 | 4.250646e+11 | 4.289909 |
| Australia | 14 | 8831 | 8725 | 90765 | 15606 | 10.28 | 107 | 5.386000e+09 | 231.0 | 11.810810 | 1.021939e+12 | 1.060340e+12 | 1.099644 |
| Brazil | 15 | 8668 | 8596 | 60702 | 14396 | 7.00 | 86 | 1.214900e+10 | 59.0 | 69.648030 | 1.845080e+12 | 1.957118e+12 | 2.056809 |

In [9]:
```
assert type(answer_one()) == pd.DataFrame, "Q1: You should return a DataFrame!"

assert answer_one().shape == (15,20), "Q1: Your DataFrame should have 20 columns and 15 entries!"
```

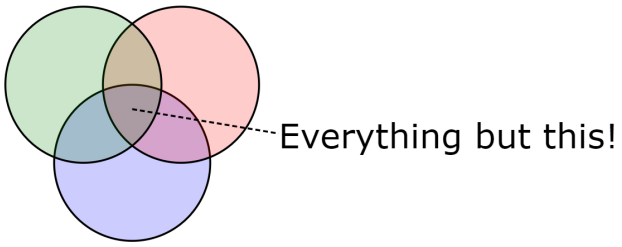
In [10]:
```
# Cell for autograder.
```

### Question 2

The previous question joined three datasets then reduced this to just the top 15 entries. When you joined the datasets, but before you reduced this to the top 15 items, how many entries did you lose?

*This function should return a single number.*

In [11]:
```
%%HTML
<svg width="800" height="300">
  <circle cx="150" cy="180" r="80" fill-opacity="0.2" stroke="black" stroke-width="2" fill="blue" />
  <circle cx="200" cy="100" r="80" fill-opacity="0.2" stroke="black" stroke-width="2" fill="red" />
  <circle cx="100" cy="100" r="80" fill-opacity="0.2" stroke="black" stroke-width="2" fill="green" />
  <line x1="150" y1="125" x2="300" y2="150" stroke="black" stroke-width="2" fill="black" stroke-dasharray="5,3"/>
  <text x="300" y="165" font-family="Verdana" font-size="35">Everything but this!</text>
</svg>
```




In [12]:
```
def answer_two():
    df1= ScimEn.merge(Energy, how= 'outer', on='Country').merge(GDP_revised, how= 'outer', on='Country')
    df2= ScimEn.merge(Energy, how= 'inner', on='Country').merge(GDP_revised, how= 'inner', on='Country')
    return len(df1) - len(df2)
    raise NotImplementedError()

answer_two()
```

Out[12]: 156

In [13]:
```
assert type(answer_two()) == int, "Q2: You should return an int number!"
```

### Question 3

What are the top 15 countries for average GDP over the last 10 years?

*This function should return a Series named `avgGDP` with 15 countries and their average GDP sorted in descending order.*

In [14]:
```
def answer_three():
    Top15 = answer_one()
    Top15 = Top15.iloc[:,10:]
    avgGDP= Top15.mean(axis=1, skipna = True).sort_values(ascending=False)
    return avgGDP
    raise NotImplementedError()

answer_three()
```

Out[14]:
```
Country
United States         1.536434e+13
China                 6.348609e+12
Japan                 5.542208e+12
Germany               3.493025e+12
France                2.681725e+12
United Kingdom        2.487907e+12
Brazil                2.189794e+12
Italy                 2.120175e+12
India                 1.769297e+12
Canada                1.660647e+12
Russian Federation    1.565459e+12
Spain                 1.418078e+12
Australia             1.164043e+12
South Korea           1.106715e+12
Iran                  4.441558e+11
dtype: float64
```

In [15]:
```
assert type(answer_three()) == pd.Series, "Q3: You should return a Series!"
```

### Question 4

By how much had the GDP changed over the 10 year span for the country with the 6th largest average GDP?

*This function should return a single number.*

In [18]:
```
def answer_four():
    Top15 = answer_one()
    Top15 = Top15.iloc[:,10:]
    avgGDP= Top15.mean(axis=1, skipna = True).sort_values(ascending=False)
    sixth_GDP_country= avgGDP.index[5]
    return Top15.loc[sixth_GDP_country, '2015'] - Top15.loc[sixth_GDP_country, '2006']
    raise NotImplementedError()

answer_four()
```

Out[18]:
```
246702696075.3999
```

In [19]:
```
# Cell for autograder.
```

### Question 5

What is the mean energy supply per capita?

*This function should return a single number.*

In [20]:
```
def answer_five():
    Top15 = answer_one()
    return Top15['Energy Supply per Capita'].mean()
    raise NotImplementedError()

answer_five()
```

Out[20]:
```
157.6
```

In [21]:
```
# Cell for autograder.
```

### Question 6

What country has the maximum % Renewable and what is the percentage?

*This function should return a tuple with the name of the country and the percentage.*

In [22]:
```
def answer_six():
    Top15 = answer_one()
    return (Top15.iloc[:,9].idxmax(), Top15.iloc[:,9].max())
    raise NotImplementedError()

answer_six()
```

Out[22]:
```
('Brazil', 69.64803)
```

In [23]:
```
assert type(answer_six()) == tuple, "Q6: You should return a tuple!"

assert type(answer_six()[0]) == str, "Q6: The first element in your result should be the name of the country!"
```

### Question 7

Create a new column that is the ratio of Self-Citations to Total Citations. What is the maximum value for this new column, and what country has the highest ratio?

*This function should return a tuple with the name of the country and the ratio.*

In [24]:
```
def answer_seven():
    Top15 = answer_one()
    Top15['Ratio'] = Top15['Self-citations']/Top15['Citations']
    return (Top15.iloc[:,-1].idxmax(), Top15.iloc[:,-1].max())
    raise NotImplementedError()

answer_seven()
```

Out[24]:
```
('China', 0.6893126179389422)
```

In [25]:
```
assert type(answer_seven()) == tuple, "Q7: You should return a tuple!"

assert type(answer_seven()[0]) == str, "Q7: The first element in your result should be the name of the country!"
```

### Question 8

Create a column that estimates the population using Energy Supply and Energy Supply per capita. What is the third most populous country according to this estimate?

*This function should return the name of the country*

In [26]:
```
def answer_eight():
    Top15 = answer_one()
    Top15['Population Estimate'] = Top15['Energy Supply']/Top15['Energy Supply per Capita']
    return Top15['Population Estimate'].sort_values(ascending=False).index[2]
    raise NotImplementedError()

answer_eight()
```

Out[26]: 'United States'

In [27]:
```
assert type(answer_eight()) == str, "Q8: You should return the name of the country!"
```

### Question 9

Create a column that estimates the number of citable documents per person. What is the correlation between the number of citable documents per capita and the energy supply per capita? Use the `.corr()` method, (Pearson's correlation).

*This function should return a single number.*

*(Optional: Use the built-in function `plot9()` to visualize the relationship between Energy Supply per Capita vs. Citable docs per Capita)*

In [28]:
```
def answer_nine():
    import scipy.stats as stats
    Top15 = answer_one()
    Top15['Population Estimate'] = (Top15['Energy Supply']/Top15['Energy Supply per Capita'])
    Top15['Citable Documents per Capita'] = Top15['Citable documents']/Top15['Population Estimate']
    Compare_Top15 = Top15[['Energy Supply per Capita', 'Citable Documents per Capita']].dropna()
    corr, pval=stats.pearsonr(Compare_Top15['Energy Supply per Capita'],Compare_Top15['Citable Documents per Capita'])
    return corr
    raise NotImplementedError()

answer_nine()
```

Out[28]: 0.7940010435442942

In [29]:
```
def plot9():
    import matplotlib as plt
    %matplotlib inline

    Top15 = answer_one()
    Top15['PopEst'] = Top15['Energy Supply'] / Top15['Energy Supply per Capita']
    Top15['Citable docs per Capita'] = Top15['Citable documents'] / Top15['PopEst']
    Top15.plot(x='Citable docs per Capita', y='Energy Supply per Capita', kind='scatter', xlim=[0, 0.0006])
```

In [30]:
```
assert answer_nine() >= -1. and answer_nine() <= 1., "Q9: A valid correlation should between -1 to 1!"
```

### Question 10

Create a new column with a 1 if the country's % Renewable value is at or above the median for all countries in the top 15, and a 0 if the country's % Renewable value is below the median.

*This function should return a series named `HighRenew` whose index is the country name sorted in ascending order of rank.*

In [31]:
```
def answer_ten():
    Top15 = answer_one()
    Top15['Relative_Renewable'] = np.where(Top15['% Renewable']>= Top15['% Renewable'].median(), 1, 0)
    HighRenew = Top15['Relative_Renewable'].squeeze()
    return HighRenew
    raise NotImplementedError()

answer_ten()
```

Out[31]:
```
Country
China                  1
United States          0
Japan                  0
United Kingdom         0
Russian Federation     1
Canada                 1
Germany                1
India                  0
France                 1
South Korea            0
Italy                  1
Spain                  1
Iran                   0
Australia              0
Brazil                 1
Name: Relative_Renewable, dtype: int64
```

In [32]:
```
assert type(answer_ten()) == pd.Series, "Q10: You should return a Series!"
```

### Question 11

Use the following dictionary to group the Countries by Continent, then create a DataFrame that displays the sample size (the number of countries in each continent bin), and the sum, mean, and std deviation for the estimated population of each country.

```
ContinentDict  = {'China':'Asia',
                  'United States':'North America',
                  'Japan':'Asia',
                  'United Kingdom':'Europe',
                  'Russian Federation':'Europe',
                  'Canada':'North America',
                  'Germany':'Europe',
                  'India':'Asia',
                  'France':'Europe',
                  'South Korea':'Asia',
                  'Italy':'Europe',
                  'Spain':'Europe',
                  'Iran':'Asia',
                  'Australia':'Australia',
                  'Brazil':'South America'}
```

*This function should return a DataFrame with index named Continent* `['Asia', 'Australia', 'Europe', 'North America', 'South America']` *and columns* `['size', 'sum', 'mean', 'std']`

In [33]:
```
def answer_eleven():
    Top15 = answer_one()

    ContinentDict  = {'China':'Asia',
                      'United States':'North America',
                      'Japan':'Asia',
                      'United Kingdom':'Europe',
                      'Russian Federation':'Europe',
                      'Canada':'North America',
                      'Germany':'Europe',
                      'India':'Asia',
                      'France':'Europe',
                      'South Korea':'Asia',
                      'Italy':'Europe',
                      'Spain':'Europe',
                      'Iran':'Asia',
                      'Australia':'Australia',
                      'Brazil':'South America'}


    Top15['Continent'] = pd.Series(ContinentDict)
    Top15['Population Estimate'] = (Top15['Energy Supply']/Top15['Energy Supply per Capita'])
    Top15_continent= Top15.groupby('Continent')['Population Estimate'].agg([np.size,np.sum, np.mean, np.std])
    return Top15_continent
    raise NotImplementedError()

answer_eleven()
```

Out[33]:

| Continent | size | sum | mean | std |
|---|---|---|---|---|
| **Asia** | 5.0 | 2.898666e+09 | 5.797333e+08 | 6.790979e+08 |
| **Australia** | 1.0 | 2.331602e+07 | 2.331602e+07 | NaN |
| **Europe** | 6.0 | 4.579297e+08 | 7.632161e+07 | 3.464767e+07 |
| **North America** | 2.0 | 3.528552e+08 | 1.764276e+08 | 1.996696e+08 |
| **South America** | 1.0 | 2.059153e+08 | 2.059153e+08 | NaN |

In [34]:
```
assert type(answer_eleven()) == pd.DataFrame, "Q11: You should return a DataFrame!"

assert answer_eleven().shape[0] == 5, "Q11: Wrong row numbers!"

assert answer_eleven().shape[1] == 4, "Q11: Wrong column numbers!"
```

## Question 12

Cut % Renewable into 5 bins. Group Top15 by the Continent, as well as these new % Renewable bins. How many countries are in each of these groups?

*This function should return a Series with a MultiIndex of* `Continent` *, then the bins for* `% Renewable` *. Do not include groups with no countries.*

In [35]:
```
def answer_twelve():

    ContinentDict  = {'China':'Asia',
                      'United States':'North America',
                      'Japan':'Asia',
                      'United Kingdom':'Europe',
                      'Russian Federation':'Europe',
                      'Canada':'North America',
                      'Germany':'Europe',
                      'India':'Asia',
                      'France':'Europe',
                      'South Korea':'Asia',
                      'Italy':'Europe',
                      'Spain':'Europe',
                      'Iran':'Asia',
                      'Australia':'Australia',
                      'Brazil':'South America'}

    Top15 = answer_one()
    Top15 = Top15.reset_index()
    Top15['Continent'] = [ContinentDict[country] for country in Top15['Country']]
    Top15['% Renewable'] = pd.cut(Top15['% Renewable'],5)
    Top15_Renewable= Top15.groupby(['Continent','% Renewable']).size()
    return Top15_Renewable[Top15_Renewable!=0]
    raise NotImplementedError()

answer_twelve()
```

Out[35]:
```
Continent      % Renewable
Asia           (2.212, 15.753]     4
               (15.753, 29.227]    1
Australia      (2.212, 15.753]     1
Europe         (2.212, 15.753]     1
               (15.753, 29.227]    3
               (29.227, 42.701]    2
North America  (2.212, 15.753]     1
               (56.174, 69.648]    1
South America  (56.174, 69.648]    1
dtype: int64
```

In [36]:
```
assert type(answer_twelve()) == pd.Series, "Q12: You should return a Series!"

assert len(answer_twelve()) == 9, "Q12: Wrong result numbers!"
```

## Question 13

Convert the Population Estimate series to a string with thousands separator (using commas). Use all significant digits (do not round the results).

e.g. 12345678.90 -> 12,345,678.90

*This function should return a series* `PopEst` *whose index is the country name and whose values are the population estimate string*

In [37]:
```
def answer_thirteen():
    Top15 = answer_one()
    Top15['PopEst'] = (Top15['Energy Supply']/Top15['Energy Supply per Capita'])
    return Top15['PopEst'].apply(lambda x: '{0:,}'.format(x)).astype(str)
    raise NotImplementedError()

answer_thirteen()
```

Out[37]:
```
Country
```

```
China                1,367,645,161.2903225
United States          317,615,384.61538464
Japan                  127,409,395.97315437
United Kingdom         63,870,967.741935484
Russian Federation           143,500,000.0
Canada                  35,239,864.86486486
Germany                 80,369,696.96969697
India                1,276,730,769.2307692
France                  63,837,349.39759036
South Korea            49,805,429.864253394
Italy                  59,908,256.880733944
Spain                    46,443,396.2264151
Iran                    77,075,630.25210084
Australia              23,316,017.316017315
Brazil                 205,915,254.23728815
Name: PopEst, dtype: object
```

In [38]:
```
assert type(answer_thirteen()) == pd.Series, "Q13: You should return a Series!"

assert len(answer_thirteen()) == 15, "Q13: Wrong result numbers!"
```

### Optional

Use the built in function `plot_optional()` to see an example visualization.

In [39]:
```
def plot_optional():
    import matplotlib as plt
    %matplotlib inline
    Top15 = answer_one()
    ax = Top15.plot(x='Rank', y='% Renewable', kind='scatter',
                    c=['#e41a1c','#377eb8','#e41a1c','#4daf4a','#4daf4a','#377eb8','#4daf4a','#e41a1c',
                       '#4daf4a','#e41a1c','#4daf4a','#4daf4a','#e41a1c','#dede00','#ff7f00'],
                    xticks=range(1,16), s=6*Top15['2014']/10**10, alpha=.75, figsize=[16,6]);

    for i, txt in enumerate(Top15.index):
        ax.annotate(txt, [Top15['Rank'][i], Top15['% Renewable'][i]], ha='center')

    print("This is an example of a visualization that can be created to help understand the data. \
This is a bubble chart showing % Renewable vs. Rank. The size of the bubble corresponds to the countries' \
2014 GDP, and the color corresponds to the continent.")
```

In [ ]:
```
```