*You are currently looking at **version 1.5** of this notebook. To download notebooks and datafiles, as well as get help on Jupyter notebooks in the Coursera platform, visit the Jupyter Notebook FAQ (https://www.coursera.org/learn/python-data-analysis/resources/0dhYG) course resource.*

# Assignment 3 - More Pandas

This assignment requires more individual learning then the last one did - you are encouraged to check out the pandas documentation (http://pandas.pydata.org/pandas-docs/stable/) to find functions or methods you might not have used yet, or ask questions on Stack Overflow (http://stackoverflow.com/) and tag them as pandas and python related. And of course, the discussion forums are open for interaction with your peers and the course staff.

## Question 1 (20%)

Load the energy data from the file `Energy Indicators.xls`, which is a list of indicators of energy supply and renewable electricity production (Energy%20Indicators.xls) from the United Nations (http://unstats.un.org/unsd/environment/excel_file_tables/2013/Energy%20Indicators.xls) 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 (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 numbers and/or parenthesis in their name. Be sure to remove these,

e.g.

`'Bolivia (Plurinational State of)'` should be `'Bolivia'`,

`'Switzerland17'` should be `'Switzerland'`.

Next, load the GDP data from the file `world_bank.csv`, which is a csv containing countries' GDP from 1960 to 2015 from World Bank (http://data.worldbank.org/indicator/NY.GDP.MKTP.CD). 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 (http://www.scimagojr.com/countryrank.php?category=2102) from the file `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.*

In [4]:

```
import numpy as np
import pandas as pd
import re
energy = pd.read_excel('Energy Indicators.xls', skip_footer=38,skiprows=17)
energy = (energy.drop(['Unnamed: 0','Unnamed: 1'], axis=1)
                .rename(columns = {'Unnamed: 2':'Country',
                                   'Petajoules':'Energy Supply',
                                   'Gigajoules':'Energy Supply per Capita',
                                   '%':'% Renewable'})
                .replace('...', np.nan))
energy['Energy Supply'] *= 1000000
energy['Country'].replace(['Republic of Korea','United States of America20','United Kin
gdom of Great Britain and Northern Ireland19','China, Hong Kong Special Administrative
 Region3'],['South Korea','United States','United Kingdom','Hong Kong'],inplace = True)
t=0
while (t<len(energy.index)):
    energy['Country'][t] = energy['Country'][t].split(' (')[0]
    energy['Country'][t] = re.split('(\d+)',energy['Country'][t])[0]
    t+=1
GDP = pd.read_csv('world_bank.csv',skiprows = 4)
GDP['Country Name'].replace(['Korea, Rep.','Iran, Islamic Rep.','Hong Kong SAR,
China'],['South Korea','Iran','Hong Kong'],inplace = True)
columns_to_keep = ['Country Name','2006','2007','2008','2009','2010','2011','2012','201
3','2014','2015']
GDP = GDP[columns_to_keep]
GDP.rename(columns={'Country Name': 'Country'}, inplace = True)
ScimEn = pd.read_excel('scimagojr-3.xlsx')
ScimEn_copy = ScimEn[:15]

def answer_one():
    Answer1 = pd.merge(ScimEn_copy, GDP, how='inner', left_on='Country', right_on='Coun
try')
    Answer1 = pd.merge(Answer1, energy, how='inner', left_on='Country', right_on='Count
ry')
    Answer1.set_index('Country', inplace = True)
    columns_to_keep = ['Rank', 'Documents', 'Citable documents', 'Citations', 'Self-cit
ations', 'Citations per document', 'H index', 'Energy Supply', 'Energy Supply per Capit
a', '% Renewable', '2006', '2007', '2008', '2009', '2010', '2011', '2012', '2013', '201
4', '2015']
    Answer1 = Answer1[columns_to_keep]
    return Answer1

answer_one()
```

Out[4]:

| | Rank | Documents | Citable documents | Citations | Self-citations | Citations per document | H index | Ei Si |
|---|---|---|---|---|---|---|---|---|
| **Country** | | | | | | | | |
| **China** | 1 | 127050 | 126767 | 597237 | 411683 | 4.70 | 138 | 1. |
| **United States** | 2 | 96661 | 94747 | 792274 | 265436 | 8.20 | 230 | 9. |
| **Japan** | 3 | 30504 | 30287 | 223024 | 61554 | 7.31 | 134 | 1. |
| **United Kingdom** | 4 | 20944 | 20357 | 206091 | 37874 | 9.84 | 139 | 7. |
| **Russian Federation** | 5 | 18534 | 18301 | 34266 | 12422 | 1.85 | 57 | 3. |
| **Canada** | 6 | 17899 | 17620 | 215003 | 40930 | 12.01 | 149 | 1. |
| **Germany** | 7 | 17027 | 16831 | 140566 | 27426 | 8.26 | 126 | 1. |
| **India** | 8 | 15005 | 14841 | 128763 | 37209 | 8.58 | 115 | 3. |
| **France** | 9 | 13153 | 12973 | 130632 | 28601 | 9.93 | 114 | 1. |
| **South Korea** | 10 | 11983 | 11923 | 114675 | 22595 | 9.57 | 104 | 1. |
| **Italy** | 11 | 10964 | 10794 | 111850 | 26661 | 10.20 | 106 | 6. |
| **Spain** | 12 | 9428 | 9330 | 123336 | 23964 | 13.08 | 115 | 4. |
| **Iran** | 13 | 8896 | 8819 | 57470 | 19125 | 6.46 | 72 | 9. |
| **Australia** | 14 | 8831 | 8725 | 90765 | 15606 | 10.28 | 107 | 5. |
| **Brazil** | 15 | 8668 | 8596 | 60702 | 14396 | 7.00 | 86 | 1. |

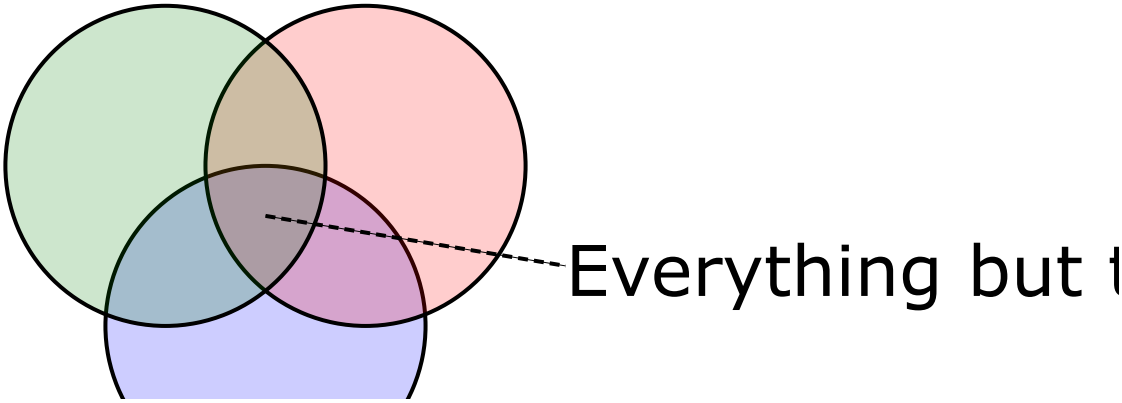
## Question 2 (6.6%)

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 [42]:

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




In [5]:

```
def answer_two():
    intersection1 = pd.merge(ScimEn, GDP, how='inner', left_on='Country', right_on='Cou
ntry')
    union1 = pd.merge(ScimEn, GDP, how='outer', left_on='Country', right_on='Country')
    intersection = pd.merge(intersection1, energy, how='inner', left_on='Country', righ
t_on='Country')
    union = pd.merge(union1, energy, how='outer', left_on='Country',
right_on='Country')
    return len(union)-len(intersection)
answer_two()
```

Out[5]:

156

Answer the following questions in the context of only the top 15 countries by Scimagojr Rank (aka the DataFrame returned by `answer_one()`)

## Question 3 (6.6%)

What is the average GDP over the last 10 years for each country? (exclude missing values from this calculation.)

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

In [8]:

```
def answer_three():
    Top15 = answer_one()
    Answer3 = Top15[['2006', '2007', '2008', '2009', '2010', '2011', '2012', '2013', '2
014', '2015']]
    Answer3['average GDP'] = Answer3.mean(axis=1)
    Answer3 = Answer3.sort('average GDP', ascending=False)
    avgGDP = Answer3['average GDP']
    return avgGDP

answer_three()
```

Out[8]:

```
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
Name: average GDP, dtype: float64
```

## Question 4 (6.6%)

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 [10]:

```
def answer_four():
    avgGDP = answer_three()
    Top15 = answer_one()
    answer4 = Top15.loc[avgGDP.index[5]][19] - Top15.loc[avgGDP.index[5]][10]
    return answer4

answer_four()
```

Out[10]:

```
246702696075.3999
```

## Question 5 (6.6%)

What is the mean `Energy Supply per Capita`?

*This function should return a single number.*

In [11]:

```
def answer_five():
    Top15 = answer_one()
    Answer = Top15['Energy Supply per Capita'].mean()
    return Answer
answer_five()
```

Out[11]:

```
157.59999999999999
```

## Question 6 (6.6%)

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 [12]:

```
def answer_six():
    Top15 = answer_one()
    Answer = (Top15['% Renewable'].argmax(),Top15['% Renewable'].max())
    return Answer
answer_six()
```

Out[12]:

```
('Brazil', 69.648030000000006)
```

## Question 7 (6.6%)

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 [14]:

```
def answer_seven():
    Top15 = answer_one()
    Top15['ratio'] = Top15['Self-citations']/Top15['Citations']
    Answer = (Top15['ratio'].argmax(),Top15['ratio'].max())
    return Answer
answer_seven()
```

Out[14]:

```
('China', 0.68931261793894216)
```

## Question 8 (6.6%)

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 a single string value.*

In [19]:

```
def answer_eight():
    Top15 = answer_one()
    Top15['population'] = Top15['Energy Supply']/Top15['Energy Supply per Capita']
    Answer = Top15.sort('population', ascending=False)
    Answer = Answer.index[2]
    return Answer
answer_eight()
```

Out[19]:

```
'United States'
```

## Question 9 (6.6%)

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 [23]:

```
def answer_nine():
    Top15 = answer_one()
    Top15['population'] = Top15['Energy Supply']/Top15['Energy Supply per Capita']
    Top15['estimate'] = Top15['Citable documents']/Top15['population']
    Answer = Top15['estimate'].corr(Top15['Energy Supply per Capita'])
    return Answer
answer_nine()
```

Out[23]:

```
0.79400104354429435
```

In [25]:

```
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='scatte
r', xlim=[0, 0.0006])

#plot9()
```

In [ ]:

```
#plot9() # Be sure to comment out plot9() before submitting the assignment!
```

## Question 10 (6.6%)

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 [29]:

```
def answer_ten():
    Top15 = answer_one()
    median = np.median(Top15['% Renewable'])
    Top15['Median'] = 0
    t=0
    while t<len(Top15.index):
        if(Top15['% Renewable'][t]>=median):
            Top15['Median'][t] = 1
        t+=1
    HighRenew = Top15['Median']
    return HighRenew
answer_ten()
```

Out[29]:

```
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: Median, dtype: int64
```

## Question 11 (6.6%)

Use the following dictionary to group the Countries by Continent, then create a dateframe 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 [31]:

```
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'}
def answer_eleven():
    Top15 = answer_one()
    Top15['population'] = Top15['Energy Supply']/Top15['Energy Supply per Capita']
    Top15['Continent'] = pd.Series(ContinentDict)
    Answer = (Top15.sort(['Continent'])
                  .reset_index()
                  .set_index(['Continent'])
                  .groupby(level=0)['population']
                  .agg({'size': np.size, 'sum': np.sum, 'mean': np.mean,'std':
np.std}))
    return Answer
answer_eleven()
```

Out[31]:

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

## Question 12 (6.6%)

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 [32]:

```
def answer_twelve():
    Top15 = answer_one()
    Top15['Continent'] = pd.Series(ContinentDict)
    Top15['% Renewable'] = pd.cut(Top15['% Renewable'],5)
    Answer = (Top15.sort(['Continent'])
                   .reset_index()
                   .set_index(['Continent','% Renewable'])
                   .groupby(level=[0,1])['Country']
                   .size())
    return Answer
answer_twelve()
```

Out[32]:

```
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
```

## Question 13 (6.6%)

Convert the Population Estimate series to a string with thousands separator (using commas). Do not round the results.

e.g. 317615384.61538464 -> 317,615,384.61538464

*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['population'] = Top15['Energy Supply']/Top15['Energy Supply per Capita']
    Top15 = Top15[['population']]
    def thousands(x):
        return '{:,}'.format(x)
    Top15 = Top15.applymap(thousands)
    PopEst = Top15['population']
    return PopEst
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: population, dtype: object
```

## Optional

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

In [34]:

```
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','#4d
af4a','#e41a1c',
                       '#4daf4a','#e41a1c','#4daf4a','#4daf4a','#e41a1c','#dede00','#ff
7f00'],
                    xticks=range(1,16), s=6*Top15['2014']/10**10, alpha=.75, figsize=[1
6,6]);

    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.")
```