*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 [1]:
```
# Although the commented section is correct, autograder gave it wrong

import pandas as pd
import numpy as np
def answer_one():
    #defining energy
    energy = (pd.read_excel('Energy Indicators.xls', header=None, skip_footer=
2).drop([0,1], axis = 1).dropna().drop(9)
                .rename(columns={2:'Country', 3:'Energy Supply', 4:'Energy Supply
per Capita', 5:'% Renewable'}))
    #to remove numbers and brackets
    energy = energy.replace(regex = True, to_replace=[r'\d', r' \(([^)]+)\)'], v
alue=r'')
    energy = energy.replace(to_replace = ['...', 'Republic of Korea', 'United St
ates of America', 'United Kingdom of Great Britain and Northern Ireland', 'Chin
a, Hong Kong Special Administrative Region'],
                            value=[np.NaN, 'South Korea', 'United States', 'Unit
ed Kingdom', 'Hong Kong'])
    energy['Energy Supply'] = energy['Energy Supply']*1000000
    energy = energy.dropna()

    #defining GDP
    GDP = pd.read_csv('world_bank.csv', header = None, skiprows = 4)
    GDP = GDP.rename(columns = GDP.iloc[0]).drop(0).replace(to_replace = ['Kore
a, Rep.', 'Iran, Islamic Rep.', 'Hong Kong SAR, China'], value = ['South Korea',
'Iran', 'Hong Kong']).rename(columns={2006: '2006', 2007: '2007', 2008: '2008',
2009: '2009', 2010: '2010',2011: '2011', 2012: '2012', 2013: '2013', 2014: '2014
', 2015: '2015'})
    GDP = GDP[['Country Name', '2006', '2007', '2008', '2009', '2010', '2011', '
2012', '2013', '2014', '2015']]

    #defining ScimEn
    ScimEn = pd.read_excel('scimagojr-3.xlsx')
    ScimEn = ScimEn[['Rank', 'Country', 'Documents', 'Citable documents', 'Citat
ions', 'Self-citations', 'Citations per document', 'H index']]
    ScimEn = ScimEn[0:15]

    #creating new dataframe of intersection of all three
    inner_all = pd.merge(energy, GDP, how = 'inner', left_on = 'Country', right_
on = 'Country Name')
    inner_all = inner_all.drop(['Country Name'], axis = 1)
    inner_all = pd.merge(inner_all, ScimEn, how = 'inner', left_on = 'Country',
right_on = 'Country')
    inner_all = inner_all.set_index('Country')
    inner_all = inner_all[['Rank', 'Documents', 'Citable documents', 'Citations
', 'Self-citations', 'Citations per document', 'H index', 'Energy Supply', 'Ener
gy Supply per Capita', '% Renewable', '2006', '2007', '2008', '2009', '2010', '2
011', '2012', '2013', '2014', '2015']]
    return inner_all
answer_one()
```

Out[1]:

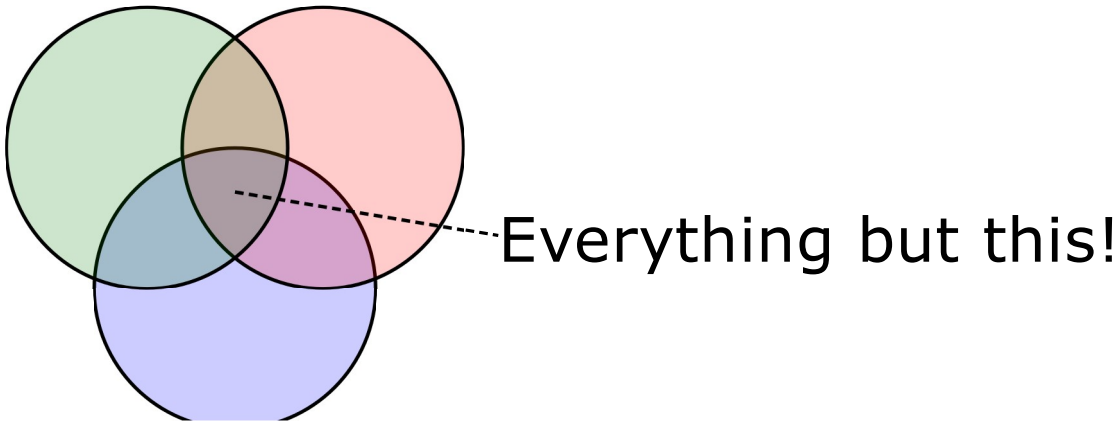
	Rank	Documents	Citable documents	Citations	Self-citations	Citations per document	H index	Energy Supply	Ener Supp per Capi
Country									
Australia	14	8831	8725	90765	15606	10.28	107	5.386000e+09	231.(
Brazil	15	8668	8596	60702	14396	7.00	86	1.214900e+10	59.0
Canada	6	17899	17620	215003	40930	12.01	149	1.043100e+10	296.(
China	1	127050	126767	597237	411683	4.70	138	1.271910e+11	93.0
France	9	13153	12973	130632	28601	9.93	114	1.059700e+10	166.(
Germany	7	17027	16831	140566	27426	8.26	126	1.326100e+10	165.(
India	8	15005	14841	128763	37209	8.58	115	3.319500e+10	26.0
Iran	13	8896	8819	57470	19125	6.46	72	9.172000e+09	119.(
Italy	11	10964	10794	111850	26661	10.20	106	6.530000e+09	109.(
Japan	3	30504	30287	223024	61554	7.31	134	1.898400e+10	149.(
South Korea	10	11983	11923	114675	22595	9.57	104	1.100700e+10	221.(
Russian Federation	5	18534	18301	34266	12422	1.85	57	3.070900e+10	214.(
Spain	12	9428	9330	123336	23964	13.08	115	4.923000e+09	106.(
United Kingdom	4	20944	20357	206091	37874	9.84	139	7.920000e+09	124.(
United States	2	96661	94747	792274	265436	8.20	230	9.083800e+10	286.(

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



In [3]:
```
import pandas as pd
import numpy as np
def answer_two():
    #defining energy
    energy = (pd.read_excel('Energy Indicators.xls', header=None, skip_footer=
2).drop([0,1], axis = 1).dropna().drop(9)
                .rename(columns={2:'Country', 3:'Energy Supply', 4:'Energy Supply
per Capita', 5:'% Renewable'}))
    #to remove numbers and brackets
    energy = energy.replace(regex = True, to_replace=[r'\d', r' \(([^)]+)\)'], v
alue=r'')
    energy = energy.replace(to_replace = ['...', 'Republic of Korea', 'United St
ates of America', 'United Kingdom of Great Britain and Northern Ireland', 'Chin
a, Hong Kong Special Administrative Region'],
                            value=[np.NaN, 'South Korea', 'United States', 'Unit
ed Kingdom', 'Hong Kong'])
    energy['Energy Supply'] = energy['Energy Supply']*1000000
    energy = energy.dropna()

    #defining GDP
    GDP = pd.read_csv('world_bank.csv', header = None, skiprows = 4)
    GDP = GDP.rename(columns = GDP.iloc[0]).drop(0).replace(to_replace = ['Kore
a, Rep.', 'Iran, Islamic Rep.', 'Hong Kong SAR, China'], value = ['South Korea',
'Iran', 'Hong Kong']).rename(columns={2006: '2006', 2007: '2007', 2008: '2008',
2009: '2009', 2010: '2010',2011: '2011', 2012: '2012', 2013: '2013', 2014: '2014
', 2015: '2015'})
    GDP = GDP[['Country Name', '2006', '2007', '2008', '2009', '2010', '2011', '
2012', '2013', '2014', '2015']]

    #defining ScimEn
    ScimEn = pd.read_excel('scimagojr-3.xlsx')
    ScimEn = ScimEn[['Rank', 'Country', 'Documents', 'Citable documents', 'Citat
ions', 'Self-citations', 'Citations per document', 'H index']]

    m1 = pd.merge(energy, GDP, how="inner", left_on="Country", right_on="Country
Name")
    m2 = pd.merge(energy, ScimEn, how="inner", left_on="Country", right_on="Coun
try")
    m3 = pd.merge(GDP, ScimEn, how="inner", left_on="Country Name", right_on="Co
untry")
    all_3 = pd.merge(m1, ScimEn, how="inner", left_on="Country", right_on="Count
ry")
    answer = len(energy.index) + len(GDP.index) + len(ScimEn.index) - len(m1.ind
ex) - len(m2.index) - len(m3.index)
    return answer
answer_two()
```

Out[3]: 155

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 [4]:
```
def answer_three():
    Top15 = answer_one()
    Top15['Mean'] = Top15[['2006','2007','2008','2009','2010','2011','2012','201
3','2014','2015']].mean(axis=1)
    avgGDP = Top15.sort_values(by = 'Mean', ascending = False)['Mean']
    #print(type(avgGDP))
    return avgGDP
answer_three()
```

Out[4]:
```
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: Mean, 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 [5]:
```
def answer_four():
    Top15 = answer_one()
    country = answer_three().keys()[5]
    answer = Top15.loc[country, '2015'] - Top15.loc[country, '2006']
    return answer
answer_four()
```

Out[5]:
```
246702696075.3999
```

Question 5 (6.6%)

What is the mean `Energy Supply per Capita`?

This function should return a single number.

In [6]:
```
def answer_five():
    Top15 = answer_one()
    answer = float(Top15['Energy Supply per Capita'].mean(axis = 0))
    # answer = float(np.average(Top15['Energy Supply per Capita'])) # alternativ
e
    return answer

answer_five()
```

Out[6]:
```
157.6
```

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 [7]:
```
def answer_six():
    Top15 = answer_one()
    max_= Top15['% Renewable'].max()
    answer = Top15[Top15['% Renewable'] == max_].iloc[0].name, max_
    return answer
answer_six()
```

Out[7]: ('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 [8]:
```
def answer_seven():
    Top15 = answer_one()
    Top15['Ratio'] = Top15['Self-citations']/(Top15['Citations'])
    max_ = Top15['Ratio'].max()
    answer = Top15[Top15['Ratio'] == max_].iloc[0].name , float(max_)
    return answer

answer_seven()
```

Out[8]: ('China', 0.6893126179389422)

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 [9]:
```
def answer_eight():
    Top15 = answer_one()
    Top15['capita'] = Top15['Energy Supply']/Top15['Energy Supply per Capita']
    return Top15.sort_values(by = 'capita',ascending = False)['capita'].keys
()[2]
answer_eight()
```

Out[9]: '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.

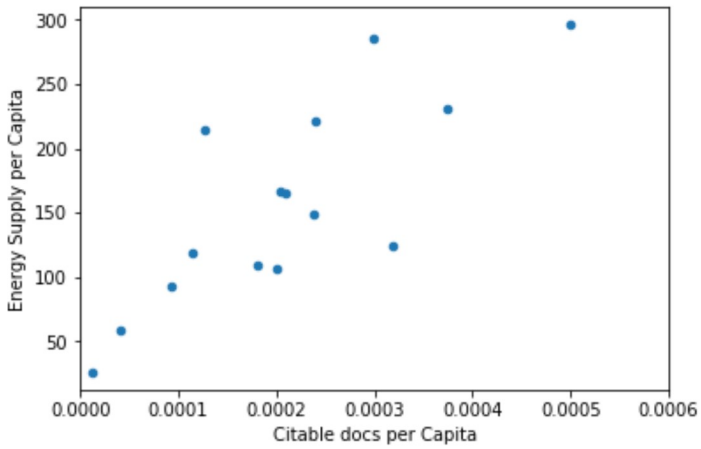
(Optional: Use the built-in function `plot9()` to visualize the relationship between Energy Supply per Capita vs. Citable docs per Capita)

In [10]:
```
def answer_nine():
    Top15 = answer_one()
    Top15['Citable documents per capita'] = Top15['Citable documents'] /(Top15['
Energy Supply']/Top15['Energy Supply per Capita'])
    Top15 = Top15.corr(method='pearson')
    answer = Top15.loc['Energy Supply per Capita', 'Citable documents per capita
']
    return answer


answer_nine()
```

Out[10]: 0.79400104354429435

In [11]:
```
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['PopEs
t']
    Top15.plot(x='Citable docs per Capita', y='Energy Supply per Capita', kind='
scatter', xlim=[0, 0.0006])
#plot9()
```

In [12]:
```
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 [13]:
```
def answer_ten():
    Top15 = answer_one()
    med = Top15['% Renewable'].median()
    Top15['HighRenew'] = (Top15['% Renewable'] >= med).apply(lambda x:1 if x else 0)#.sort_values(by = 'Rank')
    return Top15.sort_values(by = 'Rank')['HighRenew']

answer_ten()
```

Out[13]:
```
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: HighRenew, 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 [14]:
```
def answer_eleven():
    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['Capita'] = Top15['Energy Supply'] / Top15['Energy Supply per Capita']
    for i in range(len(Top15)):
        country = Top15.iloc[i].name
        Top15.set_value(Top15.iloc[i].name, 'Continent', ContinentDict[country])
        #print(Top15.set_value(Top15.iloc[i].name, 'Continent', ContinentDict[co
untry]))
    Top15 = (Top15.reset_index(level=0)
                  .set_index(['Continent', 'Country']))
    Top15 = Top15.groupby(level=0)['Capita'].agg({'size': np.size, 'sum': np.su
m, 'mean': np.average, 'std': np.std})
    return Top15

answer_eleven()
```

Out[14]:

	size	sum	mean	std
Continent				
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

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 [15]:
```
def answer_twelve():
    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'}
    for i in range(len(Top15)):
        country = Top15.iloc[i].name
        Top15.set_value(country, 'Continent', ContinentDict[country])
    Top15 = (Top15.reset_index(level=0)
                  .set_index(['Continent', 'Country']))
    Top15 = pd.cut(Top15['% Renewable'], 5)
    Top15 = (Top15.reset_index()
                .set_index(['Continent', '% Renewable']))
    Top15 = Top15.groupby(level=['Continent', '% Renewable']).size()
    return Top15
answer_twelve()
```

Out[15]:
```
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 [16]:
```
def answer_thirteen():
    Top15 = answer_one()
    Top15['PopEst'] = Top15['Energy Supply'] / Top15['Energy Supply per Capita']
    Top15 = Top15['PopEst']
    for i in range(len(Top15)):
        country = Top15.keys()[i]
        number = "{:,}".format((Top15.iloc[i]))
        Top15.replace(Top15.iloc[i], number, inplace=True)
    return Top15
answer_thirteen()
```

Out[16]:
```
Country
Australia              23,316,017.316017315
Brazil                 205,915,254.23728815
Canada                  35,239,864.86486486
China                 1,367,645,161.2903225
France                  63,837,349.39759036
Germany                 80,369,696.96969697
India                 1,276,730,769.2307692
Iran                    77,075,630.25210084
Italy                  59,908,256.880733944
Japan                  127,409,395.97315437
South Korea            49,805,429.864253394
Russian Federation            143,500,000.0
Spain                   46,443,396.2264151
United Kingdom         63,870,967.741935484
United States          317,615,384.61538464
Name: PopEst, dtype: object
```

Optional

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

In [17]:
```
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','#377eb
8','#4daf4a','#e41a1c',
                       '#4daf4a','#e41a1c','#4daf4a','#4daf4a','#e41a1c','#dede0
0','#ff7f00'],
                    xticks=range(1,16), s=6*Top15['2014']/10**10, alpha=.75, fig
size=[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 und
erstand the data. \
This is a bubble chart showing % Renewable vs. Rank. The size of the bubble corr
esponds to the countries' \
2014 GDP, and the color corresponds to the continent.")
```

In [18]:
```
plot_optional() # Be sure to comment out plot_optional() before submitting the a
ssignment!
```

```
This is an example of a visualization that can be created to help understand t
he 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 th
e continent.
```