W3.A3

2020年3月28日 星期六　　下午6:06

coursera-Applied...

Qian-Han / coursera-Applied-Data-Science-with-Python

Join GitHub today

GitHub is home to over 40 million developers working together to host and review code, manage projects, and build software together.

Sign up

Dismiss

Branch: master ▾ coursera-Applied-Data-Science-with-Python / Introduction-to-Data-Science-in-Python / week3 / week3_Assignment.ipynb　Find file　Copy path

Qian-Han Slides and Assignments　980211b on 14 Jul 2017

1 contributor

1584 lines (1584 sloc) | 77.5 KB　Raw　Blame　History

*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 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 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. And of course, the discussion forums are open for interaction with your peers and the course staff.

Question 1 (20%)

Dataset Energy

Load the energy data from the file `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 (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.

broadcasting

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,

regular expression　replace

e.g.

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

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

Dataset GDP

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. 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"
```

Dataset ScimEn

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

Join 3 Datasets 2-2

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 [2]: def answer_one():
            import pandas as pd
            import numpy as np

            energy = pd.read_excel('Energy Indicators.xls', skiprows=17,skip_footer= 38)
            energy = energy[['Unnamed: 1','Petajoules','Gigajoules','%']]
            energy.columns = ['Country', 'Energy Supply', 'Energy Supply per Capita', '% Renewable']
            energy[['Energy Supply', 'Energy Supply per Capita', '% Renewable']] =  energy[['Energy Supply
        ', 'Energy Supply per Capita', '% Renewable']].replace('...',np.NaN).apply(pd.to_numeric)
            energy['Energy Supply'] = energy['Energy Supply']*1000000
            energy['Country'] = energy['Country'].replace({'China, Hong Kong Special Administrative Region
        ':'Hong Kong','United Kingdom of Great Britain and Northern Ireland':'United Kingdom','Republic of
        Korea':'South Korea','United States of America':'United States','Iran (Islamic Republic of)':'Iran
        '})
            energy['Country'] = energy['Country'].str.replace(" \(.*\)","")

            GDP = pd.read_csv('world_bank.csv',skiprows=4)
            GDP['Country Name'] = GDP['Country Name'].replace('Korea, Rep.','South Korea')
            GDP['Country Name'] = GDP['Country Name'].replace('Iran, Islamic Rep.','Iran')
            GDP['Country Name'] = GDP['Country Name'].replace('Hong Kong SAR, China','Hong Kong')

            columns = ['Country Name','2006','2007','2008','2009','2010','2011','2012','2013','2014','2015
        ']
            GDP = GDP[columns]

            ScimEn = pd.read_excel('scimagojr-3.xlsx')
            ScimEn_m = ScimEn[:15]

            df = pd.merge(ScimEn_m, energy, how = 'inner', left_on = 'Country', right_on='Country')
            final_df = pd.merge(df,GDP, how = 'inner', left_on = 'Country', right_on='Country Name')
            final_df = final_df.set_index('Country')
            columns = ['Rank', 'Documents', 'Citable documents', 'Citations', 'Self-citations', 'Citations
        per document', 'H index', 'Energy Supply', 'Energy Supply per Capita', '% Renewable', '2006', '200
        7', '2008', '2009', '2010', '2011', '2012', '2013', '2014', '2015']
            ans = final_df[columns]

        #     print(len(final_df))

            return ans
```



```
answer_one()
```

Out[2]:

	Rank	Documents	Citable documents	Citations	Self-citations	Citations per document	H index	Energy Supply	Energy Supply per Capita	% Renewable	2006
Country											
China	1	127050	126767	597237	411683	4.70	138	1.271910e+11	93.0	19.754910	3.9923
United States	2	96661	94747	792274	265436	8.20	230	9.083800e+10	286.0	11.570980	1.4792
Japan	3	30504	30287	223024	61554	7.31	134	1.898400e+10	149.0	10.232820	5.4965
United Kingdom	4	20944	20357	206091	37874	9.84	139	7.920000e+09	124.0	10.600470	2.4196
Russian Federation	5	18534	18301	34266	12422	1.85	57	3.070900e+10	214.0	17.288680	1.3857
Canada	6	17899	17620	215003	40930	12.01	149	1.043100e+10	296.0	61.945430	1.5644
Germany	7	17027	16831	140566	27426	8.26	126	1.326100e+10	165.0	17.901530	3.3328
India	8	15005	14841	128763	37209	8.58	115	3.319500e+10	26.0	14.969080	1.2658
France	9	13153	12973	130632	28601	9.93	114	1.059700e+10	166.0	17.020280	2.6078
South Korea	10	11983	11923	114675	22595	9.57	104	1.100700e+10	221.0	2.279353	9.4101
Italy	11	10964	10794	111850	26661	10.20	106	6.530000e+09	109.0	33.667230	2.2021
Spain	12	9428	9330	123336	23964	13.08	115	4.923000e+09	106.0	37.968590	1.4148
Iran	13	8896	8819	57470	19125	6.46	72	9.172000e+09	119.0	5.707721	3.8955
Australia	14	8831	8725	90765	15606	10.28	107	5.386000e+09	231.0	11.810810	1.0219
Brazil	15	8668	8596	60702	14396	7.00	86	1.214900e+10	59.0	69.648030	1.8450

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 [48]: %%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-da
         sharray="5,3"/>
           <text  x="300" y="165" font-family="Verdana" font-size="35">Everything but this!</text>
         </svg>
```

SVG Image

```
In [4]: def answer_two():
            import pandas as pd
            import numpy as np

            energy = pd.read_excel('Energy Indicators.xls', skiprows=17,skip_footer= 38)
            energy = energy[['Unnamed: 1','Petajoules','Gigajoules','%']]
            energy.columns = ['Country', 'Energy Supply', 'Energy Supply per Capita', '% Renewable']
            energy[['Energy Supply', 'Energy Supply per Capita', '% Renewable']] =  energy[['Energy Supply
        ', 'Energy Supply per Capita', '% Renewable']].replace('...',np.NaN).apply(pd.to_numeric)
```



```
            energy['Energy Supply'] = energy['Energy Supply']*1000000
            energy['Country'] = energy['Country'].replace({'China, Hong Kong Special Administrative Region
        ':'Hong Kong','United Kingdom of Great Britain and Northern Ireland':'United Kingdom','Republic of
        Korea':'South Korea','United States of America':'United States','Iran (Islamic Republic of)':'Iran
        '})
            energy['Country'] = energy['Country'].str.replace(" \(.*\)","")

            GDP = pd.read_csv('world_bank.csv',skiprows=4)
            GDP['Country Name'] = GDP['Country Name'].replace('Korea, Rep.','South Korea')
            GDP['Country Name'] = GDP['Country Name'].replace('Iran, Islamic Rep.','Iran')
            GDP['Country Name'] = GDP['Country Name'].replace('Hong Kong SAR, China','Hong Kong')


            columns = ['Country Name','2006','2007','2008','2009','2010','2011','2012','2013','2014','2015
        ']
            GDP = GDP[columns]
            GDP.columns = ['Country','2006','2007','2008','2009','2010','2011','2012','2013','2014','2015'
        ]
        #     print(GDP)

            ScimEn = pd.read_excel('scimagojr-3.xlsx')
            ScimEn_m = ScimEn[:15]

            df = pd.merge(ScimEn, energy, how = 'inner', left_on = 'Country', right_on='Country')
            final_df = pd.merge(df,GDP, how = 'inner', left_on = 'Country', right_on='Country')
            final_df = final_df.set_index('Country')
            columns = ['Rank', 'Documents', 'Citable documents', 'Citations', 'Self-citations', 'Citations
        per document', 'H index', 'Energy Supply', 'Energy Supply per Capita', '% Renewable', '2006', '200
        7', '2008', '2009', '2010', '2011', '2012', '2013', '2014', '2015']
            ans = final_df[columns]

            df2 = pd.merge(ScimEn, energy, how = 'outer', left_on = 'Country', right_on='Country')
            final_df2 = pd.merge(df2,GDP, how = 'outer', left_on = 'Country', right_on='Country')

            print(len(final_df))
        #     GDP['Country Name'] = GDP['Country']

            print(len(final_df2) - len(final_df))

        #     return 156

        answer_two()

        162
        156
Out[4]: 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 [80]: def answer_three():
             import numpy as np
             Top15 = answer_one()
             columns = ['2006','2007','2008','2009','2010','2011','2012','2013','2014','2015']
             Top15['Mean'] = Top15[columns].mean(axis=1)
             avgGDP = Top15.sort_values(by = 'Mean', ascending = False)['Mean']

         #     print(avgGDP)

             return avgGDP
         answer_three()
```



```
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 [94]: def answer_four():
             import pandas as pd
             import numpy as np
             Top15 = answer_one()
             columns = ['2006','2007','2008','2009','2010','2011','2012','2013','2014','2015']
             Top15['Mean'] = Top15[columns].mean(axis=1)
             avgGDP = Top15.sort_values(by = 'Mean', ascending = False)['Mean']
             target = avgGDP.index[5]

             target_data = Top15.loc[target]

             ans = target_data['2015'] - target_data['2006']

         #     print(target_data)

             return ans
         answer_four()

Out[94]: 246702696075.3999
```

Question 5 (6.6%)

What is the mean `Energy Supply per Capita`?

This function should return a single number.

```
In [98]: def answer_five():
             import pandas as pd
             import numpy as np
             Top15 = answer_one()
             target = Top15['Energy Supply per Capita']
             ans = target.mean()

             return ans
         answer_five()

Out[98]: 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 [106]: def answer_six():
```



```
              import pandas as pd
              import numpy as np
              Top15 = answer_one()
              target = Top15['% Renewable']
              ans = target.max()
              index = Top15[Top15['% Renewable'] == ans].index[0]

              return (index, ans)

          answer_six()

Out[106]: ('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 [115]: def answer_seven():
              import pandas as pd
              import numpy as np
              Top15 = answer_one()
              Top15['Ratio'] = Top15['Self-citations'] / Top15['Citations']
              target = Top15['Ratio']
              max_value = target.max()
              max_index = Top15[Top15['Ratio'] == max_value].index[0]

          #     print(max_index)
              return (max_index, max_value)

          answer_seven()

Out[115]: ('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 [128]: def answer_eight():
              Top15 = answer_one()
              columns = ['Energy Supply','Energy Supply per Capita']
              target = Top15[columns]
              target['Captita'] = Top15['Energy Supply'] / Top15['Energy Supply per Capita']

              ans = target.sort_values(by = 'Captita', ascending = False).iloc[2].name

              return ans
          answer_eight()

          United States
          /opt/conda/lib/python3.5/site-packages/ipykernel/__main__.py:5: SettingWithCopyWarning:
          A value is trying to be set on a copy of a slice from a DataFrame.
          Try using .loc[row_indexer,col_indexer] = value instead

          See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#in
          dexing-view-versus-copy
Out[128]: '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)

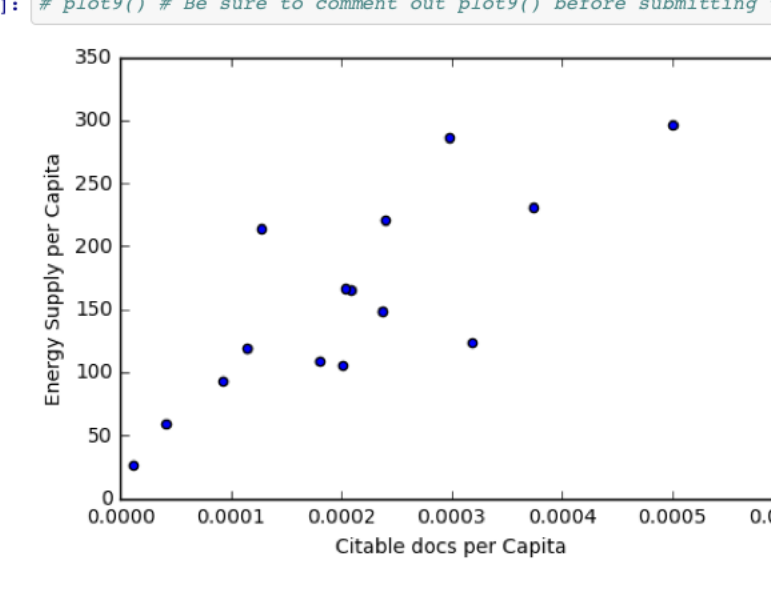
```
In [129]: def answer_nine():
              Top15 = answer_one()
              Top15['PopEst'] = Top15['Energy Supply'] / Top15['Energy Supply per Capita']
              Top15['Citable docs per Capita'] = Top15['Citable documents'] / Top15['PopEst']
              ans = Top15['Citable docs per Capita'].corr(Top15['Energy Supply per Capita'])
              return ans
          answer_nine()

Out[129]: 0.79400104354429435

In [133]: 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 [134]: # 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 [26]: def answer_ten():
             import pandas as pd
             Top15 = answer_one()
             med = Top15['% Renewable'].median()

             Top15['HighRenew'] = [1 if x >= med else 0 for x in Top15['% Renewable']]

         #     test = Top15.sort_values(by = 'Rank')['HighRenew']

             ans = Top15['HighRenew']

         #     print(test)

             return pd.Series(ans)
```



```
         answer_ten()

Out[26]: 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 [118]: def answer_eleven():
              import pandas as pd
              import numpy as np
              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['PopEst'] = (Top15['Energy Supply'] / Top15['Energy Supply per Capita'])

              Top15 = Top15.reset_index()
              Top15['Continent'] = [ContinentDict[country] for country in Top15['Country']]
```



```
          #     print(Top15['Continent'])
          #     print(ContinentDict.values())
          #     Top15['Continent'] = [ContinentDict[country] for country in Top15['Country']]

              target = Top15.set_index('Continent').groupby(level = 0)['PopEst'].agg({'size':np.size, 'sum'
          :np.sum, 'mean':np.mean, 'std':np.std})
              ans = target[['size', 'sum', 'mean', 'std' ]]
              return ans

          answer_eleven()
```

Out[118]:

	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 [128]: def answer_twelve():
              import pandas as pd
              import numpy as np
              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['bins'] = pd.cut(Top15['% Renewable'],5)

              ans = Top15.groupby(['Continent', 'bins']).size()

          #     print(Top15['bins'])

              return pd.Series(ans)
          answer_twelve()

Out[128]: Continent      bins
          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 [151]: def answer_thirteen():
              import pandas as pd
              import numpy as np
              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()

          #     print()
              tmp = list()
              Top15['PopEst'] = (Top15['Energy Supply'] / Top15['Energy Supply per Capita'])
              tmp = Top15['PopEst'].tolist()

          #     print(tmp)

              Top15['PopEst'] = (Top15['Energy Supply'] / Top15['Energy Supply per Capita']).apply(lambda x
          : "{:,}".format(x), tmp)

              ans = pd.Series(Top15['PopEst'])


              return ans
          answer_thirteen()

Out[151]: 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
```

Optional

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


```
In [152]: 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','#e4
          1a1c',
                                 '#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 co
          untries' \
          2014 GDP, and the color corresponds to the continent.")

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