assignment3

September 28, 2020

1 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 20% of the assignment grade.

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

[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')
```

1.0.1 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".

[195]:
```
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')

def answer_one():



    ### PREPARSING THE ENERGY USAGE EXCEL FILE


    Energy = pd.read_excel('assets/Energy Indicators.xls')

    #Just for carrying out trials
    new_Energy = Energy.copy()

    # Remove header and footer. Including titles and units.
    new_Energy = new_Energy.iloc[17:244]
    # Drops the two first columns
    new_Energy.drop(new_Energy.columns[[0,2]], axis=1,inplace=True)
    # Changing column labels
    new_Energy.rename(columns = {'Unnamed: 1' : 'Country',
                                 'Unnamed: 3' : 'Energy Supply',
                                 'Unnamed: 4' : 'Energy Supply per Capita',
                                 'Unnamed: 5' : '% Renewable'}, inplace = True)
```

```
    # Replaces values of no information i.e "..." with np.NaN
    new_Energy['Energy Supply'].replace('...',np.NaN,inplace=True)

    # Changes petajoule to gigajoule
    new_Energy['Energy Supply'] = np.multiply(new_Energy['Energy␣
↪Supply'],1000000)
    new_Energy.set_index('Country', inplace=True)
    new_Energy.rename(index = {"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"}, inplace=True)

    # Changes all countries having parenthesis.
    for item in new_Energy.index:
        if item[-1] == ")":
            item_list = item.split('(')
            new_Energy.rename(index = {item : item_list[0].
↪strip()},inplace=True)


    # Just changing back to normal indexing for now and changing name.
    new_Energy.reset_index(inplace = True)
    new_Energy.index = new_Energy.index+1
    Energy = new_Energy

    ### PREPARSING GDP DATA FROM CSV FILE

    GDP = pd.read_csv('assets/world_bank.csv')

    #Remove header
    GDP = GDP[4:]

    #Change name of title



    GDP.rename(columns = {'Data Source' : 'Country'}, inplace=True)

    #Change names
    GDP['Country'].replace({"Korea, Rep." : "South Korea",
                            "Iran, Islamic Rep.": "Iran",
                            "Hong Kong SAR, China": "Hong Kong"}, inplace=True)
```

```
    ################## TRYING TO CHANGE NAME OF A LIST OF COLUMNS␣
 ↪#############################
#     GDP.rename(columns = {[GDP.columns[4:]] : [np.arange[1960:(1960+len(GDP.
 ↪columns[4:]))]]})



#     df2 = pd.DataFrame(np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]]),␣
 ↪columns=['1', '2', '3'])

#     print(df2)
#    df2.rename(columns = {df2.columns[0:]: [4,5,6]})
#     df2.columns[0:2] = [4,5,6]
#     print(df2)

   ␣
 ↪#############################################################################################################



    ### PREPARSING JOURNAL CONTRIBUTION EXCEL FILE

    ScimEn = pd.read_excel('assets/scimagojr-3.xlsx')



    ### JOINING THE TABLES

    #Picks out the last 10 years of GDP data.
    GDP = pd.concat([GDP[GDP.columns[0]],GDP[GDP.columns[-10:]]], axis=1)

    merge1 = pd.merge(Energy, GDP, how='inner', on='Country')

    # Only takes the top 15 countries from ScimEn according to task.
    df = pd.merge(merge1, ScimEn[:15], how='inner', on='Country')


    df.set_index('Country', inplace=True)


    # Making sure the names and the order of the columns fits the description
    df.rename(columns = {'Unnamed: 50':'2006', 'Unnamed: 51':'2007', 'Unnamed:␣
 ↪52':'2008',
                'Unnamed: 53':'2009', 'Unnamed: 54':'2010', 'Unnamed: 55':'2011',
                'Unnamed: 56':'2012', 'Unnamed: 57':'2013','Unnamed: 58': '2014',
```

```
                'Unnamed: 59':'2015'}, inplace = True)

    df = df.reindex(columns=['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'])



    return df.astype('float')


answer_one()
    #raise NotImplementedError()
```

[195]:

Country	Rank	Documents	Citable documents	Citations
Australia	14.0	8831.0	8725.0	90765.0
Brazil	15.0	8668.0	8596.0	60702.0
Canada	6.0	17899.0	17620.0	215003.0
China	1.0	127050.0	126767.0	597237.0
France	9.0	13153.0	12973.0	130632.0
Germany	7.0	17027.0	16831.0	140566.0
India	8.0	15005.0	14841.0	128763.0
Iran	13.0	8896.0	8819.0	57470.0
Italy	11.0	10964.0	10794.0	111850.0
Japan	3.0	30504.0	30287.0	223024.0
South Korea	10.0	11983.0	11923.0	114675.0
Russian Federation	5.0	18534.0	18301.0	34266.0
Spain	12.0	9428.0	9330.0	123336.0
United Kingdom	4.0	20944.0	20357.0	206091.0
United States	2.0	96661.0	94747.0	792274.0

Country	Self-citations	Citations per document	H index
Australia	15606.0	10.28	107.0
Brazil	14396.0	7.00	86.0
Canada	40930.0	12.01	149.0
China	411683.0	4.70	138.0
France	28601.0	9.93	114.0
Germany	27426.0	8.26	126.0
India	37209.0	8.58	115.0

Iran	19125.0	6.46	72.0
Italy	26661.0	10.20	106.0
Japan	61554.0	7.31	134.0
South Korea	22595.0	9.57	104.0
Russian Federation	12422.0	1.85	57.0
Spain	23964.0	13.08	115.0
United Kingdom	37874.0	9.84	139.0
United States	265436.0	8.20	230.0

Country	Energy Supply	Energy Supply per Capita	% Renewable	\
Australia	5.386000e+09	231.0	11.810810	
Brazil	1.214900e+10	59.0	69.648030	
Canada	1.043100e+10	296.0	61.945430	
China	1.271910e+11	93.0	19.754910	
France	1.059700e+10	166.0	17.020280	
Germany	1.326100e+10	165.0	17.901530	
India	3.319500e+10	26.0	14.969080	
Iran	9.172000e+09	119.0	5.707721	
Italy	6.530000e+09	109.0	33.667230	
Japan	1.898400e+10	149.0	10.232820	
South Korea	1.100700e+10	221.0	2.279353	
Russian Federation	3.070900e+10	214.0	17.288680	
Spain	4.923000e+09	106.0	37.968590	
United Kingdom	7.920000e+09	124.0	10.600470	
United States	9.083800e+10	286.0	11.570980	

Country	2006	2007	2008	2009	\
Australia	1.021939e+12	1.060340e+12	1.099644e+12	1.119654e+12	
Brazil	1.845080e+12	1.957118e+12	2.056809e+12	2.054215e+12	
Canada	1.564469e+12	1.596740e+12	1.612713e+12	1.565145e+12	
China	3.992331e+12	4.559041e+12	4.997775e+12	5.459247e+12	
France	2.607840e+12	2.669424e+12	2.674637e+12	2.595967e+12	
Germany	3.332891e+12	3.441561e+12	3.478809e+12	3.283340e+12	
India	1.265894e+12	1.374865e+12	1.428361e+12	1.549483e+12	
Iran	3.895523e+11	4.250646e+11	4.289909e+11	4.389208e+11	
Italy	2.202170e+12	2.234627e+12	2.211154e+12	2.089938e+12	
Japan	5.496542e+12	5.617036e+12	5.558527e+12	5.251308e+12	
South Korea	9.410199e+11	9.924316e+11	1.020510e+12	1.027730e+12	
Russian Federation	1.385793e+12	1.504071e+12	1.583004e+12	1.459199e+12	
Spain	1.414823e+12	1.468146e+12	1.484530e+12	1.431475e+12	
United Kingdom	2.419631e+12	2.482203e+12	2.470614e+12	2.367048e+12	
United States	1.479230e+13	1.505540e+13	1.501149e+13	1.459484e+13	

Country	2010	2011	2012	2013	\

```
Australia            1.142251e+12  1.169431e+12  1.211913e+12  1.241484e+12
Brazil               2.208872e+12  2.295245e+12  2.339209e+12  2.409740e+12
Canada               1.613406e+12  1.664087e+12  1.693133e+12  1.730688e+12
China                6.039659e+12  6.612490e+12  7.124978e+12  7.672448e+12
France               2.646995e+12  2.702032e+12  2.706968e+12  2.722567e+12
Germany              3.417298e+12  3.542371e+12  3.556724e+12  3.567317e+12
India                1.708459e+12  1.821872e+12  1.924235e+12  2.051982e+12
Iran                 4.677902e+11  4.853309e+11  4.532569e+11  4.445926e+11
Italy                2.125185e+12  2.137439e+12  2.077184e+12  2.040871e+12
Japan                5.498718e+12  5.473738e+12  5.569102e+12  5.644659e+12
South Korea          1.094499e+12  1.134796e+12  1.160809e+12  1.194429e+12
Russian Federation   1.524917e+12  1.589943e+12  1.645876e+12  1.666934e+12
Spain                1.431673e+12  1.417355e+12  1.380216e+12  1.357139e+12
United Kingdom       2.403504e+12  2.450911e+12  2.479809e+12  2.533370e+12
United States        1.496437e+13  1.520402e+13  1.554216e+13  1.577367e+13

                              2014          2015
Country
Australia            1.272520e+12  1.301251e+12
Brazil               2.412231e+12  2.319423e+12
Canada               1.773486e+12  1.792609e+12
China                8.230121e+12  8.797999e+12
France               2.729632e+12  2.761185e+12
Germany              3.624386e+12  3.685556e+12
India                2.200617e+12  2.367206e+12
Iran                 4.639027e+11           NaN
Italy                2.033868e+12  2.049316e+12
Japan                5.642884e+12  5.669563e+12
South Korea          1.234340e+12  1.266580e+12
Russian Federation   1.678709e+12  1.616149e+12
Spain                1.375605e+12  1.419821e+12
United Kingdom       2.605643e+12  2.666333e+12
United States        1.615662e+13  1.654857e+13
```

[2]:
```
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!"
```

[3]:
```
# Cell for autograder.
```

1.0.2 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.

[16]:
```
%%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>
```

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

[196]:
```
def answer_two():

    ### PREPARSING THE ENERGY USAGE EXCEL FILE


    Energy = pd.read_excel('assets/Energy Indicators.xls')

    #Just for carrying out trials
    new_Energy = Energy.copy()

    # Remove header and footer. Including titles and units.
    new_Energy = new_Energy.iloc[17:244]
    # Drops the two first columns
    new_Energy.drop(new_Energy.columns[[0,2]], axis=1,inplace=True)
    # Changing column labels
    new_Energy.rename(columns = {'Unnamed: 1' : 'Country',
                                 'Unnamed: 3' : 'Energy Supply',
                                 'Unnamed: 4' : 'Energy Supply per Capita',
                                 'Unnamed: 5' : '% Renewable'}, inplace = True)


    # Replaces values of no information i.e "..." with np.NaN
    new_Energy['Energy Supply'].replace('...',np.NaN,inplace=True)

    # Changes petajoule to gigajoule
    new_Energy['Energy Supply'] = np.multiply(new_Energy['Energy␣
 ↪Supply'],1000000)
    new_Energy.set_index('Country', inplace=True)
    new_Energy.rename(index = {"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"}, inplace=True)

    # Changes all countries having parenthesis.
    for item in new_Energy.index:
        if item[-1] == ")":
            item_list = item.split('(')
            new_Energy.rename(index = {item : item_list[0].
↪strip()},inplace=True)


    # Just changing back to normal indexing for now and changing name.
    new_Energy.reset_index(inplace = True)
    new_Energy.index = new_Energy.index+1
    Energy = new_Energy

    ### PREPARSING GDP DATA FROM CSV FILE

    GDP = pd.read_csv('assets/world_bank.csv')

    #Remove header
    GDP = GDP[4:]

    #Change name of title



    GDP.rename(columns = {'Data Source' : 'Country'}, inplace=True)

    #Change names
    GDP['Country'].replace({"Korea, Rep." : "South Korea",
                            "Iran, Islamic Rep.": "Iran",
                            "Hong Kong SAR, China": "Hong Kong"}, inplace=True)




    ### PREPARSING JOURNAL CONTRIBUTION EXCEL FILE

    ScimEn = pd.read_excel('assets/scimagojr-3.xlsx')



    ### JOINING THE TABLES
```

```
    #Picks out the last 10 years of GDP data.
    GDP = pd.concat([GDP[GDP.columns[0]],GDP[GDP.columns[-10:]]], axis=1)

    merge1 = pd.merge(Energy, GDP, how='inner', on='Country')
    df = pd.merge(merge1, ScimEn, how='inner', on='Country')


    #The total of entires lost must be the total of entires from all the␣
↪different regions,
    #minus three times the inner join, i.e the size of the joint set. This␣
↪answer is:

    return GDP.shape[0] + Energy.shape[0]+ ScimEn.shape[0] - 3*df.shape[0]

  # raise NotImplementedError()
```

[197]:
```
assert type(answer_two()) == int, "Q2: You should return an int number!"
```

1.0.3 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.

[198]:
```
def answer_three():
    df = answer_one()


    # Takes out only the GDP values and computes the mean and sorts in␣
↪descending order.
    return df.iloc[:,10:].mean(axis=1).sort_values(ascending=False)

#     raise NotImplementedError()
```

[199]:
```
assert type(answer_three()) == pd.Series, "Q3: You should return a Series!"
```

1.0.4 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.

[200]:
```
def answer_four():
    df = answer_one()
    avg_GDP_order = answer_three()

    #Takes the order from avg_GDP_order to locate the values from the whole␣
↪dataframe given from answer_one()
```

```
    return df.loc[avg_GDP_order.index[5]]['2015'] - df.loc[avg_GDP_order.
 ↪index[5]]['2006']
#    raise NotImplementedError()
```

[201]:
```
# Cell for autograder.
```

1.0.5 Question 5

What is the mean energy supply per capita?
This function should return a single number.

[202]:
```
def answer_five():
    df = answer_one()
    return df['Energy Supply per Capita'].mean()


    #raise NotImplementedError()
```

[203]:
```
# Cell for autograder.
```

1.0.6 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.

[204]:
```
def answer_six():
    df = answer_one()
    df = df["% Renewable"].sort_values()
    return (df.index[-1], df[-1])
   # raise NotImplementedError()
```

[205]:
```
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!"
```

1.0.7 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.

[230]:
```
def answer_seven():
    df = answer_one()

    # Create the sorted ration of self-citations to total citations
    df["Ratio of Self-Citations to Total Citations"] = (df['Self-citations']/
 ↪df['Citations'])
    # Takes out the specific list in order to do sorting and such.
```

```
    sorted_ratio = df["Ratio of Self-Citations to Total Citations"].
↪sort_values()

    #Could also have sorted on the entire DateFrame, using the ratio column as␣
↪input.

    return (sorted_ratio.index[0],sorted_ratio[0])

answer_seven()
#     raise NotImplementedError()
```

[230]:
```
('Australia', 0.17193852255825484)
```

[231]:
```
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!"
```

1.0.8 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

[208]:
```
def answer_eight():
    df = answer_one()

    df['Population'] = df['Energy Supply']/df['Energy Supply per Capita']
    return (df['Population'].sort_values(ascending=False)).index[2]

    #raise NotImplementedError()
```

[209]:
```
assert type(answer_eight()) == str, "Q8: You should return the name of the␣
↪country!"
```

1.0.9 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)

[211]:
```
def answer_nine():
    import matplotlib as plt
    %matplotlib inline
    df = answer_one()
    df = df.astype('float')
```

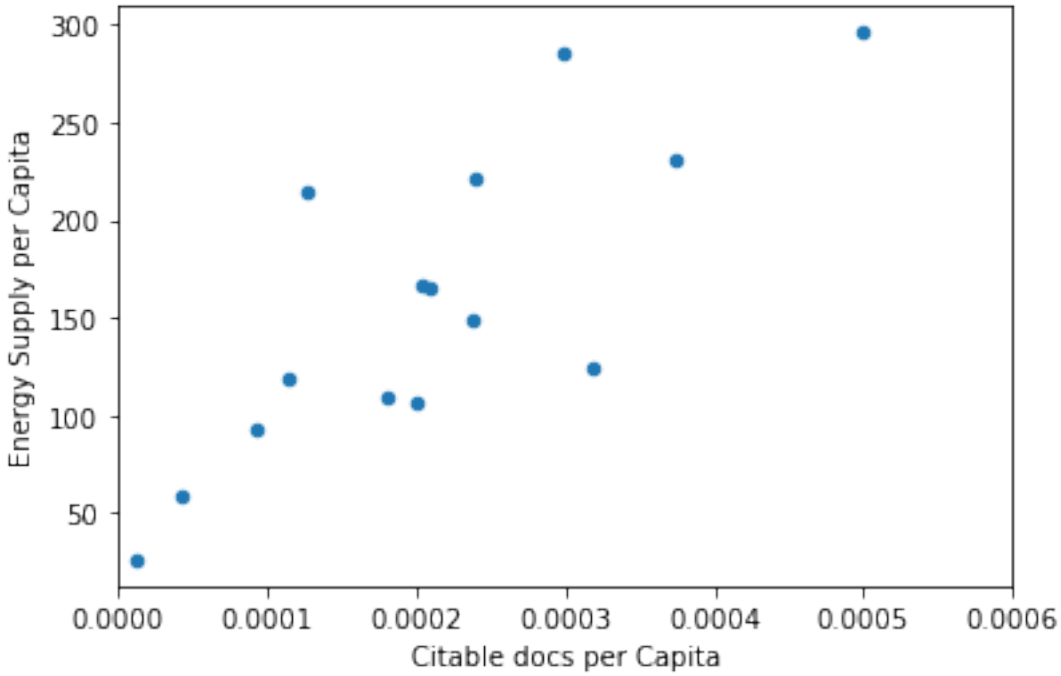
```
    df['Citable docs per Capita'] = (df['Citable documents'])/(df['Energy␣
 ↪Supply']/df['Energy Supply per Capita'])

    s1 = df['Citable docs per Capita']
    s2 = df['Energy Supply per Capita']

    df.plot(x='Citable docs per Capita', y='Energy Supply per Capita',␣
 ↪kind='scatter', xlim=[0, 0.0006])
    return s1.corr(s2,method='pearson')

answer_nine()
#     raise NotImplementedError()
```

[211]: 0.7940010435442942



[187]:
```
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])
```

[176]:
```
assert answer_nine() >= -1. and answer_nine() <= 1., "Q9: A valid correlation␣
 ↪should between -1 to 1!"
```

1.0.10 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.*

[243]:
```
def answer_ten():
    df = answer_one()
    df['HighRenew'] = (df['% Renewable'] > df['% Renewable'].median()).
 ↪astype('int64')
    return df['HighRenew'].sort_values()

answer_ten()
#     raise NotImplementedError()
```

[243]:
```
Country
Australia              0
France                 0
India                  0
Iran                   0
Japan                  0
South Korea            0
United Kingdom         0
United States          0
Brazil                 1
Canada                 1
China                  1
Germany                1
Italy                  1
Russian Federation     1
Spain                  1
Name: HighRenew, dtype: int64
```

[244]:
```
assert type(answer_ten()) == pd.Series, "Q10: You should return a Series!"
```

1.0.11 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']`

[350]:
```
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'}
    df = answer_one()
    df['Population'] = df['Energy Supply']/df['Energy Supply per Capita']



    # Group and pick out the column population
    group_object = df.groupby(ContinentDict)['Population']

#     df2 = pd.DataFrame({'size': group_object['Population'].size()})

    pop_df = pd.DataFrame({'size' : group_object.size(),
                           'sum'  : group_object.sum(),
                           'mean' : group_object.mean(),
                           'std'  : group_object.std()})
```

```
    return pop_df



answer_eleven()


#     raise NotImplementedError()
```

[350]:

	size	sum	mean	std
Asia	5	2.898666e+09	5.797333e+08	6.790979e+08
Australia	1	2.331602e+07	2.331602e+07	NaN
Europe	6	4.579297e+08	7.632161e+07	3.464767e+07
North America	2	3.528552e+08	1.764276e+08	1.996696e+08
South America	1	2.059153e+08	2.059153e+08	NaN

[314]:
```
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!"
```

1.0.12 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.

[401]:
```
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'}
    df = answer_one()
```

```
    # Creates the two new columns, cutting into 5 groups (% of renewable +␣
 ↪continents)
    df['bins'] = pd.cut(df['% Renewable'], 5)

    for key in ContinentDict.keys():
        df.at[key,'Continent'] = ContinentDict[key]


    # Groups a multiindex thingy
    df = df.set_index(['Continent', 'bins'])
    return df.groupby(level=(0,1)).size()

answer_twelve()



# raise NotImplementedError()
```

[401]:
```
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
```

[402]:
```
assert type(answer_twelve()) == pd.Series, "Q12: You should return a Series!"

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

1.0.13 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*

[423]:
```
def answer_thirteen():
    df = answer_one()
    df['Population'] = (df['Energy Supply']/df['Energy Supply per Capita'])
```

```
#     df['Population']=df['Population'].astype(str).replace(r"(\d{3})(\d+)",␣
 ↪r"\1,\2", regex=True)
    df['Population'] = df.apply(lambda x: "{:,}".format(x['Population']),␣
 ↪axis=1)

    return df['Population']

answer_thirteen()



#     raise NotImplementedError()
```

[423]:
```
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: Population, dtype: object
```

[424]:
```
assert type(answer_thirteen()) == pd.Series, "Q13: You should return a Series!"

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

1.0.14 Optional

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

[ ]:
```
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.")
```