Лабораторна робота №3

Злиття датафреймів, агрегування даних та візуалізація даних

Мета: Вивчення функцій Pandas для злиття, агрегування та візуалізації даних
.

2. Завдання:

В цій лабораторній роботі потрібно використовувати методи Pandas для злиття та агрегування, використовувати атрибути loc, iloc та слайсинг. НЕ використовувати спискові включення та інші методи ітерування всередині структур даних (для завдань 1-14)

1. Загрузити файл з даними про споживання енергії “En_In.xls”, який являє собою перелік показників енергозабезпечення та виробництва відновлюваної електроенергії, і ввести їх в DataFrame.

Майте на увазі, що це файл Excel, а не .csv. Також необхідно НЕ ВКЛЮЧАТИ інформацію з нижніх та верхніх рядків файлу даних. Перші два стовпці непотрібні, тому їх необхідно виключити із датафрейму, а також поміняти мітки стовпців так, щоб вони були такими:

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

2. Переведіть дані із стовпчику ‘`Energy Supply`’ в гДж (*Примітка: в 1ПДж = 1000000 ГДж*). Для всіх країн, у яких відсутні дані (наприклад, дані з "....."), переконайтеся, що це відображається як значення np.NaN.

3. Перейменуйте наступний список країн:
`"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"`

4. Існує також декілька країн, що мають в назві цифри та/або дужки. Обов’язково видаліть їх, напр. `'Bolivia (Plurinational State of)'` повинна бути ‘`Bolivia`’, ‘`Switzerland17`’ повинна бути `Switzerland`’.

Очікуваний вивід для завдань 1-4.

```
In [38]: Energy.loc[Energy['Country'].isin(['American Samoa','South Korea', 'Bolivia' ])]
```

Out[38]:

	Country	Energy Supply	Energy Supply per Capita	% Renewable
4	American Samoa	nan	nan	0.641
25	Bolivia	336000000.000	32.000	31.477
165	South Korea	11007000000.000	221.000	2.279

5. Далі, завантажте дані про ВВП з файлу „gpd.csv”, що містить дані Світового банку про ВВП країн від 1960 до 2015 року.

Обов’язково не включайте заголовок до датафрейму і перейменуйте наступний список країн:

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

Очікуваний вивід для завдання 5 (показані тільки 11 стовпчиків):

```
In [46]: GPD.head(1)
```

Out[46]:

	Country	Country Code	Indicator Name	Indicator Code	2006	2007	2008	2009	2010	2011	2012	2
0	Aruba	ABW	GDP at market prices (constant 2010 US$)	NY.GDP.MKTP.KD	nan	nan	nan	nan	2467703910.615	nan	nan	

6. Завантажте дані з файлу „scimagojr.xlsx”, який класифікує країни на основі їхніх публікацій в журналах у галузі енергетичного машинобудування та енергетичних технологій.

7. Приєднайте три набори даних із завдань 1-6 до нового набору даних (використовуючи перетин назв країн).
- Використовуйте лише дані про ВВП за останні 10 років (2006-2015 рр.) і лише 15 найкращих країн за рейтингом Scimagojr (Rank від 1 до 15)
- Індексом цього DataFrame повинна бути назва країни, а стовпцями мають бути ['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'].

Ви повинні отримати DataFrame з 15 рядками та 20 стовпцями.

Очікуваний вивід для завдання 7 (показані тільки 9 стовпчиків):

```
In [55]: Result.head(3)
```

Out[55]:

Country	Rank	Documents	Citable documents	Citations	Self-citations	Citations per document	H index	Energy Supply	Energy Supply per Capita
China	1	127050	126767	597237	411683	4.700	138	127191000000.000	93.000
United States	2	96661	94747	792274	265436	8.200	230	90838000000.000	286.000
Japan	3	30504	30287	223024	61554	7.310	134	18984000000.000	149.000

```
In [56]: Result.shape
Out[56]: (15, 20)
```

Завдання 8 – 14 слід розв’язувати за допомогою датафрейму із завдання 7.

8. Створіть функцію, щоб визначити середній ВВП країн за останні 10 років.

Ця функція повинна повертати об’єкт Series з 15 країнами та їх середній ВВП, відсортований за спаданням.

Очікуваний вивід для завдання 8

```
In [12]: task_eight()
Out[12]: Country
         United States         15364344302990.000
         China                  6348608932836.100
         Japan                  5542207638235.176
         Germany                3493025339072.848
         France                 2681724635761.589
         United Kingdom         2487906661418.417
         Brazil                 2189794143774.905
         Italy                  2120175089933.776
         India                  1769297396603.860
         Canada                 1660647466307.512
         Russian Federation     1565459478480.661
         Spain                  1418078278145.694
         Australia              1164042729991.427
         South Korea            1106714508244.852
         Iran                    444155754051.095
         Name: avgGDP, dtype: float64
```

9. Створіть функцію, щоб визначити, наскільки змінився ВВП за останні 10 років для країни з 5-м середнім ВВП.

Ця функція повинна повертати кортеж із назвою країни та значенням, на яке змінився ВВП.

Очікуваний вивід для завдання 9

```
In [30]: task_nine()
Out[30]: ('France', 153345695364.24023)
```

10. Створіть функцію, щоб визначити, в якій країні встановлено максимум поновлюваних джерел енергії (% Renewable) та який саме відсоток.

Ця функція повинна повертати кортеж із назвою країни та відсотком.

Очікуваний вивід для завдання 10

```
In [41]: task_ten()
Out[41]: ('Brazil', 69.64803)
```

11. Створіть стовпець, в якому оцінюється чисельність населення. Для пошуку оцінки використовуйте значення енергопостачання (Energy Supply) та енергопостачання на душу населення (Energy Supply per capita). Яка країна є шостою за цією оцінкою?

Ця функція повинна повертати кортеж із назвою країни та населенням

Очікуваний вивід для завдання 11

```
In [76]: task_eleven()
Out[76]: ('Japan', 127409395.97315437)
```

12. Створіть стовпець, в якому оцінюється кількість цитованих документів на одну особу. Який взаємозв'язок між кількістю цитованих документів на душу населення та енергозабезпеченням на душу населення? Використовуйте метод .corr () (кореляція Пірсона).

Ця функція повинна повертати одне число

Очікуваний вивід для завдання 12

```
In [88]: task_twelve()
Out[88]: 0.7940010435442942
```

13. Створіть новий стовпчик в який занесіть 1, якщо значення % відновлюваної енергії (% Renewable) країни дорівнює або вище медіани для всіх країн, і 0, якщо значення (% Renewable) нижче медіани.

Ця функція повинна повертати Series, індексом якого є назва країни, відсортована у порядку зростання Rank

Очікуваний вивід для завдання 13

```
In [117]: task_thirteen()

Out[117]: 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
          dtype: int32
```

14. Використайте наступний словник, щоб згрупувати країни за континентами, потім створіть DataFrame, який відображає розмір вибірки (кількість країн на кожному континенті), а також суму, середнє та стандартне відхилення для оцінки кількості населення для кожної країни.

```
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'}
```

Ця функція повинна повертати DataFrame з індексом, що має ім'я Continent ['Asia', 'Australia', 'Europe', 'North America', 'South America'] та стовпцями ['size', 'sum', 'mean', 'std '].

Очікуваний вивід для завдання 14

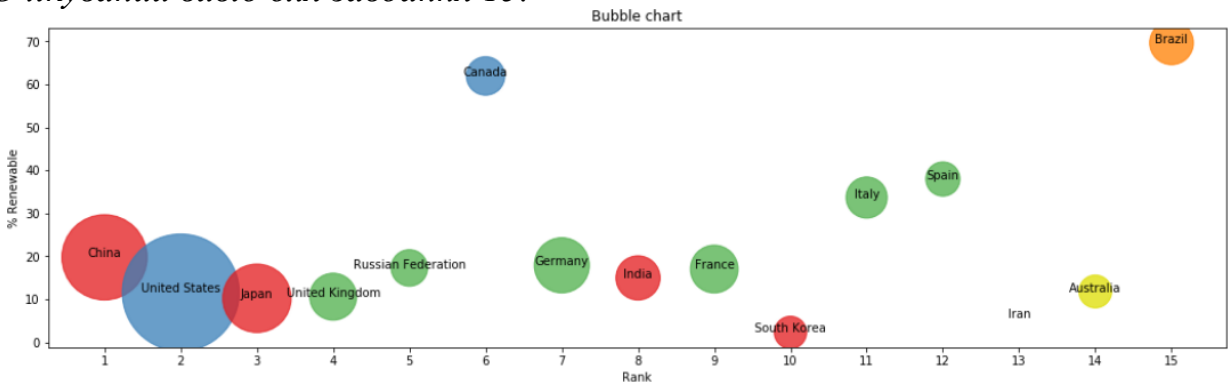
```
In [120]: task_forteen()
```

Out[120]:

Continent	size	sum	mean	std
Asia	5	2898666386.611	579733277.322	679097888.366
Australia	1	23316017.316	23316017.316	nan
Europe	6	457929667.216	76321611.203	34647667.066
North America	2	352855249.480	176427624.740	199669644.857
South America	1	205915254.237	205915254.237	nan

15. Створіть бульбашкову діаграму, що відображає залежність % поновлюван их джерел від рейтингу країни (% Renewable vs. Rank). Розмір бульбашки від повідає ВВП країн 2015 року, а колір відповідає континенту.

Очікуваний вивід для завдання 15.



3. Зміст звіту

1. Титульний аркуш.
2. Назва та мета лабораторної роботи.
3. Хід роботи.
4. Посилання на створений блокнот Jupyter на GitHub.
5. Висновки.