

Библиотека pandas : часть 1

Алла Тамбовцева

Домашнее задание

Описание домашнего задания и формат сдачи

В домашнем задании необходимо решить предложенные задачи по программированию – вписать свой код в ячейки после условий задач вместо комментария `### YOUR CODE HERE ###` в файле `homework-pandas1.ipynb` и сохранить изменения, используя опцию *Save and Checkpoint* из вкладки меню *File* или кнопку *Save and Checkpoint* на панели инструментов. Итоговый файл в формате `.ipynb` (файл Jupyter Notebook) необходимо загрузить в личный кабинет обучающей онлайн платформы Skillbox (<https://go.skillbox.ru/>) и отправить на проверку.

Задание 1

Загрузить массив NumPy из файла `"arr_pandas.npy"` (как в задании к предыдущему модулю) и преобразовать его в датафрейм. Массив содержит данные по результатам соревнований Scottish Hill Races в 2000 году (полное описание на английском языке можно посмотреть на [странице \(https://vincentarelbundock.github.io/Rdatasets/doc/DAAG/races2000.html\)](https://vincentarelbundock.github.io/Rdatasets/doc/DAAG/races2000.html) с документацией по исходному файлу с данными).

Подсказка (преобразование в датафрейм):

```
In [ ]: dat = pd.DataFrame(arr) # arr - массив NumPy из файла
```

```
In [1]: import numpy as np
import pandas as pd
arr = np.load('arr_pandas.npy', allow_pickle=True)
dat = pd.DataFrame(arr) # arr - массив NumPy из файла
dat
```

Out[1]:

		0	1	2	3	4	5
0	Aonach Mor Gondola	2	2000	0.403611	0.518889	uphill	
1	Broughton Brewery	2	650	0.254444	0.316667	other	
2	El-Brim-Ick	3	750	0.485833	0.389167	other	
3	The Devils Burdens	21	4100	2.39972	3.09333	relay	
4	Tiso Carnethy	6	2500	0.782222	0.919167	hill	
...	
72	Tinto	4.5	1500	0.499444	0.581111	hill	
73	Druim Fada	6.5	1000	0.751111	0.972222	other	
74	Elrick	3.6	650	0.358056	0.4425	relay	
75	Gondola	2.5	2000	0.387222	0.518889	uphill	
76	Greenmantle	2	650	0.254444	0.316667	other	

77 rows × 6 columns

Задание 2

Определить, сколько в датафрейме строк и столбцов. Привести код и указать ответ в виде текста или комментария к коду.

```
In [13]: len(dat.iloc[:, 0]) # 77 строк
```

```
Out[13]: 77
```

```
In [18]: len(dat.iloc[0, :]) # 6 столбцов
```

```
Out[18]: 6
```

```
In [21]: dat.shape # или так: первое число - количество строк, второе - количество столбцов
```

```
Out[21]: (77, 6)
```

Задание 3

Присвоить столбцам следующие названия (указаны с пояснениями):

- `id` : id участника
- `dist` : расстояние в милях (по карте)
- `climb` : высота, достигнутая на маршруте (в сумме за весь маршрут, в футах)
- `time` : время (в часах)
- `timef` : время для женщин (в часах)
- `type` : тип гонки (*hill, marathon, relay, uphill or other*)

```
In [19]: dat.columns = ['id', 'dist', 'climb', 'time', 'timef', 'type']  
dat
```

```
Out[19]:
```

		id	dist	climb	time	timef	type
0	Aonach Mor Gondola	2	2000	0.403611	0.518889	uphill	
1	Broughton Brewery	2	650	0.254444	0.316667	other	
2	El-Brim-Ick	3	750	0.485833	0.389167	other	
3	The Devils Burdens	21	4100	2.39972	3.09333	relay	
4	Tiso Carnethy	6	2500	0.782222	0.919167	hill	
...	
72	Tinto	4.5	1500	0.499444	0.581111	hill	
73	Druim Fada	6.5	1000	0.751111	0.972222	other	
74	Elrick	3.6	650	0.358056	0.4425	relay	
75	Gondola	2.5	2000	0.387222	0.518889	uphill	
76	Greenmantle	2	650	0.254444	0.316667	other	

77 rows × 6 columns

Задание 4

Вывести на экран значение высоты, достигнутой на маршруте участником *Norman's Law*.

```
In [34]: dat.loc[7, 'climb']
```

Out[34]: 700

```
In [52]: dat.index = dat.id
dat
```

Out[52]:

	id	dist	climb	time	timef	type
id						
Aonach Mor Gondola	Aonach Mor Gondola	2	2000	0.403611	0.518889	uphill
Broughton Brewery	Broughton Brewery	2	650	0.254444	0.316667	other
El-Brim-Ick	El-Brim-Ick	3	750	0.485833	0.389167	other
The Devils Burdens	The Devils Burdens	21	4100	2.39972	3.09333	relay
Tiso Carnethy	Tiso Carnethy	6	2500	0.782222	0.919167	hill
...
Tinto	Tinto	4.5	1500	0.499444	0.581111	hill
Druim Fada	Druim Fada	6.5	1000	0.751111	0.972222	other
Elrick	Elrick	3.6	650	0.358056	0.4425	relay
Gondola	Gondola	2.5	2000	0.387222	0.518889	uphill
Greenmantle	Greenmantle	2	650	0.254444	0.316667	other

77 rows × 6 columns

```
In [53]: dat.loc["Norman's Law", 'climb']
```

Out[53]: 700

Задание 5

Вывести на экран значения показателей `dist` , `climb` , `time` для первых 10 участников.

```
In [62]: dat[['dist', 'climb', 'time']][:10]
```

Out[62]:

	dist	climb	time
id			
Aonach Mor Gondola	2	2000	0.403611
Broughton Brewery	2	650	0.254444
El-Brim-Ick	3	750	0.485833
The Devils Burdens	21	4100	2.39972
Tiso Carnethy	6	2500	0.782222
Criffel	7	1800	0.793333
Chapelgill	1.5	1400	0.314444
Norman's Law	5	700	0.464167
Craig Dunain	6	900	0.546111
Knockfarrel	5	1200	0.623333

Задание 6

Вывести на экран сводную информацию по датафрейму, которая включает типы всех столбцов. Сколько столбцов типа `float` в датафрейме? Привести ответ на вопрос в виде текста или комментария к коду.

```
In [64]: dat.info() #
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 77 entries, Aonach Mor Gondola to Greenmantle
Data columns (total 6 columns):
#   Column  Non-Null Count  Dtype
---  -
0   id      77 non-null       object
1   dist    77 non-null       object
2   climb   77 non-null       object
3   time    77 non-null       object
4   timef   75 non-null       object
5   type    77 non-null       object
dtypes: object(6)
memory usage: 6.7+ KB
```

Задание 7

Выбрать строки, которые соответствуют участникам эстафеты (*relay*).

```
In [ ]: ### YOUR CODE HERE ###
```

Задание 8

Выбрать строки, которые соответствуют участникам гонки в холмах (*hill*), которые в сумме достигли высоты более 1000 футов. Посчитать, сколько таких участников.

```
In [ ]: ### YOUR CODE HERE ###
```

Задание 9

Выбрать строки, соответствующие участникам, которые либо достигли высоты более 4000 футов, либо потратили менее 0.5 часов.

```
In [ ]: ### YOUR CODE HERE ###
```

Задание 10

Создать столбец `time_min`, который содержит время маршрута, измеренное в минутах.

```
In [ ]: ### YOUR CODE HERE ###
```

Задание 11

Создать столбец `year` с годом соревнований (езде 2000 год). Внимание: столбец с годом должен быть числовым (целочисленным).

```
In [ ]: ### YOUR CODE HERE ###
```

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
84

Дополнительное (необязательное) задание.

Задание 1

Загрузить датафрейм из файла `extraversion.csv`, используя код ниже. При этом файл `extraversion.csv` должен находиться в той же папке, что и ноутбук с решениями. Его можно поместить в ту же папку, нажав кнопку *Upload* в *Home*.

```
In [ ]: # кодировка UTF-8, чтобы кириллица корректно считывалась на Windows
ps = pd.read_csv("extraversion.csv", encoding = "UTF-8")
```

Файл содержит результаты учебного психометрического исследования, целью которого является выявление связи между уровнем экстраверсии человека и его склонности к участию в волонтерской деятельности. Датафрейм содержит следующие столбцы:

- sex : пол респондента (Женский, Мужской);
- volunteer : регулярное участие в волонтерской деятельности (Да, Нет);
- Q 1 - Q 57 : ответы на вопросы по анкете Айзенка (Да, Нет), информацию об анкете и сами вопросы можно найти на [этой](http://ipp.hse.ru/57-testytest-ajzenka-ekstraversiya-introversiya-nejrotizm) (<http://ipp.hse.ru/57-testytest-ajzenka-ekstraversiya-introversiya-nejrotizm>), странице.

```
In [ ]: ### YOUR CODE HERE ###
```

Задание 2

Определить, сколько в датафрейме строк и столбцов. Привести код и указать ответ в виде текста или комментария к коду.

```
In [ ]: ### YOUR CODE HERE ###
```

Задание 3

Переименовать столбцы Q 1 - Q 57 в Q1 - Q57 , другими словами, убрать в названиях всех столбцов пробелы в середине (если есть).

Подсказка 1: Метод `.replace()` для строк на [pythontutor.ru](https://pythontutor.ru/lessons/str/) (<https://pythontutor.ru/lessons/str/>).

Подсказка 2: Для выполнения этого задания можно написать функцию, которая будет заменять пробелы на «пустоту», а потом применить её с помощью функции `map()` ко всем элементам списка. Пример ниже иллюстрирует применение функции для изменения регистра текста.

```
In [ ]: # Пример

L = ['яблоко', 'груша', 'слива'] # исходный список

# функция принимает на вход строку x и возвращает её же,
# но большими буквами – метод .upper()

def f(x):
    """
    Input: x is a string.
    Output: x is a string.
    Makes all letters uppercase.
    """
    return x.upper()

# результат: применяем функцию f к списку L через map и преобразуем в список
# можно убрать list() вначале и убедиться, что он здесь нужен

list(map(f, L))
```

```
In [ ]: ### YOUR CODE HERE ###
```

Задание 4

Выбрать столбцы Q1 , Q3 , Q8 , Q10 , Q13 , Q17 , Q22 , Q25 , Q27 , Q39 , Q44 , Q46 , Q49 , Q53 , Q56 и сохранить их в отдельный датафрейм extra_yes .

Выбрать столбцы Q5 , Q15 , Q20 , Q29 , Q32 , Q34 , Q37 , Q41 , Q51 и сохранить их в отдельный датафрейм extra_no .

Эти столбцы будут использоваться для вычисления индекса экстраверсии.

```
In [ ]: ### YOUR CODE HERE ###
```

Задание 5

Посчитать для каждой строки в датафрейме extra_yes число ответов "Да" и полученный результат сохранить в переменную extra_yes_sum . Посчитать для каждой строки в датафрейме extra_no число ответов "Нет" и полученный результат сохранить в переменную extra_no_sum .

Подсказка 1: метод .isin() для [датафреймов \(https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.isin.html\)](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.isin.html) pandas .

Подсказка 2: метод .sum() для [датафреймов \(https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.sum.html\)](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.sum.html) pandas .

```
In [ ]: ### YOUR CODE HERE ###
```

Задание 6

Добавить в исходный датафрейм столбец extra , который представляет собой индекс экстраверсии, который считается так: сумма числа ответов "Да" в extra_yes и числа ответов "Нет" в extra_no .

```
In [ ]: ### YOUR CODE HERE ###
```

Задание 7

Добавить в исходный датафрейм столбец female , состоящий из значений 0 и 1 (0 — Мужской, 1 — Женский).

Подсказка: возможно, пригодится метод .astype() для [Series \(https://pandas.pydata.org/pandas-docs/version/0.23.4/generated/pandas.Series.astype.html\)](https://pandas.pydata.org/pandas-docs/version/0.23.4/generated/pandas.Series.astype.html) в pandas , он преобразует типы столбцов.

```
In [ ]: ### YOUR CODE HERE ###
```

Задание 8

Выбрать из исходного датафрейма строки, которые соответствуют либо волонтерам с индексом экстраверсии выше 15, либо не-волонтерам с индексом экстраверсии ниже 15. Сохранить в датафрейм pure .

```
In [ ]: ### YOUR CODE HERE ###
```

Задание 9

Определить (любым способом, кроме явного подсчёта), сколько волонтёров и не-волонтёров в датафрейме `pure`.

In []: `### YOUR CODE HERE ###`

Задание 10

Определить минимальное, максимальное, среднее и медианное значение индекса экстраверсии в датафрейме `pure`. Сохранить полученные результаты в отдельные переменные (их должно быть 4).

Добавить в датафрейм `pure` столбец `high`, состоящий из 0 и 1, где 1 соответствует респондентам, уровень экстраверсии которых выше значения $m = \max\{\text{median}, \text{mean}\}$, то есть максимума из медианного и среднего значения, а 0 — респондентам с уровнем экстраверсии не выше m .

In []: `### YOUR CODE HERE ###`