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/ (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) 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	id	dist	climb	time	timef	type
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]:

id	dist	climb	time
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 (везде 2000 год). Внимание: столбец с годом должен быть числовым (целочисленным).

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

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

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 волонтёрской деятельности. Датафрейм содержит следующие столбцы:

- `sex` : пол респондента (Женский, Мужской);
- `volunteer` : регулярное участие в волонтёрской 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) 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/).

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 волонтёрам с индексом экстраверсии выше 15, либо не-волонтёрам 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 ###
```