АКАДЕМИЯ АЙТИ

День 25

academyit.ru

Работа с файлами и каталогами. Основные операции с путями к файлам. Импорт пакета. Важнейшие стандартные пакеты. Подсистема pip . Установка стороннего модуля. Создание собственных модулей.

Турашова Анна Николаевна
Преподаватель
anna1turashova@gmail.com
Telegram: @anna1tur

Проверка домашнего задания

Задание 1.

Файл Fishing.csv содержит результаты опроса о рыбалке: респонденты, заполняя опросник, подробно описывали свою недавнюю рыбалку.

Описание переменных в датафрейме:
* mode: выбранный тип рыбалки: на берегу (beach), на пирсе (pier), в своей лодке (boat) и в арендованной лодке (charter);
* price: стоимость выбранного типа рыбалки;
* catch: коэффициент улова при выбранном типе рыбалки;
* pbeach: стоимость рыбалки на берегу;
* ppier: стоимость рыбалки на пирсе;
* pboat: стоимость рыбалки на своей лодке;
* pcharter: стоимость рыбалки на арендованной лодке;
* cbeach: коэффициент улова на рыбалке на берегу;
* cpier: коэффициент улова на рыбалке на пирсе;
* cboat: коэффициент улова на рыбалке на своей лодке;
* ccharter: коэффициент улова на рыбалке на арендованной лодке;
* income: доход в месяц.

Подробнее об опросе и исследовании можно почитать в [статье](https://core.ac.uk/download/pdf/38934845.pdf)
J.Herriges, C.Kling "*Nonlinear Income Effects in Random Utility Models*" (1999).

Задание 1.

1) Загрузить таблицу из файла Fishing.csv и сохранить её в датафрейм dat.
Вывести на экран первые 8 строк загруженного датафрейма.

2) Добавить, используя метод .apply(), столбец log_income, содержащий натуральный логарифм доходов респондентов.

3) Посчитать для каждого респондента абсолютное значение отклонения price от pbeach и сохранить результат в столбец pdiff.
Подсказка 1: для нахождения абсолютного значения числа используется функция abs(). Пример:

```
abs(-8)
8
```

Подсказка 2: пример с lamda-функцией в первом уроке этого модуля.

4) Сгруппировать наблюдения в таблице по признаку тип рыбалки (mode) и вывести для каждого типа среднюю цену (price), которую респонденты заплатили за рыбалку.

5) Сгруппировать наблюдения в таблице по признаку тип рыбалки (mode) и вывести для каждого типа разницу между медианным и средним значением цены (price), которую респонденты заплатили за рыбалку.
Посказка: можно написать свою lambda-функцию для подсчёта разницы между медианой и средним и применить её внутри метода для агрегирования. Внимание: название самостоятельно написанной функции будет уже вводиться без кавычек.

Задание 1.

6) Сгруппировать наблюдения в таблице по признаку тип рыбалки (mode) и сохранить полученные датафреймы (один для каждого типа рыбалки) в отдельные csv-файлы. В итоге должно получиться четыре разных csv-файла.
Подсказка: можно запустить следующий код и посмотреть, что получится:

```
for name, data in dat.groupby("mode"):
    print(name, data)
```

7) Отсортировать строки в датафрейме в соответствии со значениями income в порядке убывания таким образом, чтобы результаты сортировки сохранились в исходном датафрейме.

8) Отсортировать строки в датафрейме в соответствии со значениями price и income в порядке возрастания. Можно ли сказать, что люди с более низким доходом и выбравшие более дешёвый тип рыбалки, в целом, предпочитают один тип рыбалки, а люди с более высоким доходом и более дорогой рыбалкой – другой? Ответ записать в виде текстовой ячейки или в виде комментария.

9) Любым известным способом проверить, есть ли в датафрейме пропущенные значения. Если есть, удалить строки с пропущенными значениями. Если нет, написать комментарий, что таких нет.

Что такое формат CSV

CSV (comma-separated values; значения, разделенные запятыми) – текстовый формат, позволяющий хранить табличные данные

Почему CSV – наиболее популярный формат табличных данных

- Легко читается людьми
- Содержит структурированные данные
- Поддерживается почти всеми системами хранения данных

import pandas as pd

df = pd.read_csv(“name.csv”) – чтение файла

pandas.series.apply – принимает функцию и применяет её к series:
df.loc[:, ‘w1’, ‘w2’].apply(np.mean, axis=1)

df.groupby(‘group’).agg(‘sum’) – группирование, применение sum к группе

df.sort_values(‘total’) – сортировка
df.sort_index() – сортировка по индексу

df.isnull() – маска нулевых значений
df.dropna() – удалить строки с отсутствующими данными
df.fillna(0) – все nan заменяется на 0

df.set_index([‘date’, ‘lang’]) – установить два столбца как индекс

pd.concat([df5, df]) – объединить df

df.to_csv(‘name.csv’) – сохранение df в csv

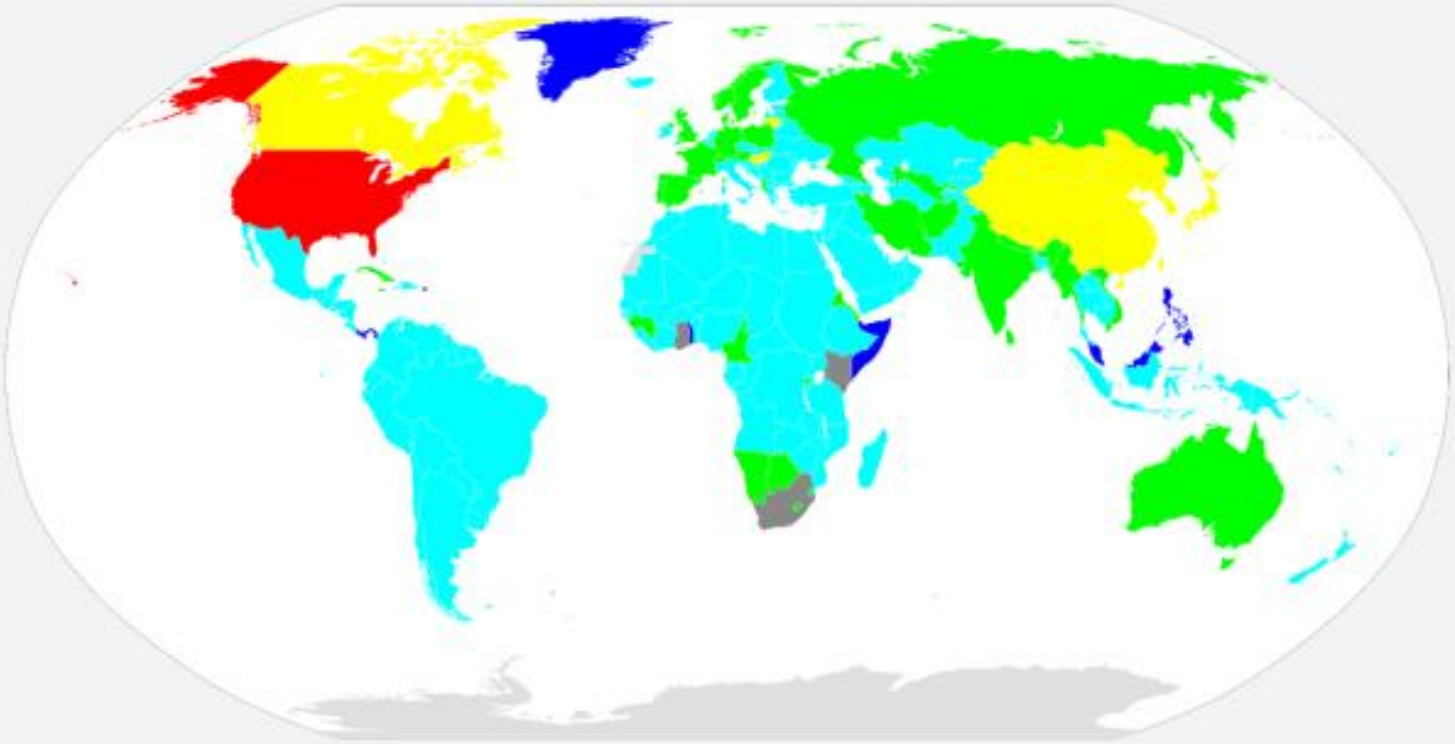
Проблема форматирования дат

22.04.2006

2006.04.22

04/22/2006

04.22.2006



Стандартный формат дат

11/08/12 – ?

Стандарт **ISO 8601:** предписывает записывать даты в следующем виде: **2012-08-11**

Что такое формат XLSX

XLSX – бинарный формат хранения данных Excel

Некоторые особенности формата XLSX:

- Несколько таблиц в одном файле
- Форматирование и объединение ячеек
- Формулы для автоматического вычисления значений ячеек

df = pd.read_excel('name.xlsx') – чтение

DataFrame может содержать только одну таблицу.
Используйте sheet_name='name', чтобы указать имя листа.

df.to_excel('name.xlsx', index_label='index') – запись

Json

Модуль JSON

- JSON (англ. JavaScript Object Notation) — формат хранения данных в виде списков и словарей, поддерживающий произвольную вложенность

Пример:

```
{ "name": "Barsik",
   "age": 7,
     "meals": [ "Whiskas", "Royal Canin",
                "Purina", "Hills", "Brit Care" ]
}
```

Загрузка из JSON

```
import json

with open('cats.json') as cat_file:
    data = json.load(cat_file)
for key, value in data.items():
    if type(value) == list:
        print(f'{key}: {", ".join(value)}')
    else:
        print(f'{key}: {value}')
```

json.load(f) - читает из файла

json.loads(s) – читает из строки

Сохранение в JSON

```
with open('cats.json', 'w') as file:
    json.dump(data, file, ensure_ascii=False, indent=2)
```

json.dump(f) - пишет в файл

json.dumps(s) – пишет в строку

Домашнее задание

Задача 1.

Документ «article.txt» содержит следующий текст:

Вечерело
Жужжали мухи
Светил фонарик
Кипела вода в чайнике
Венера зажглась на небе
Деревья шумели
Тучи разошлись
Листва зеленела

Требуется реализовать функцию longest_words(file), которая выводит слово, имеющее максимальную длину (или список слов, если таковых несколько).

Задача 2.

Требуется создать csv-файл «rows_300.csv» со следующими столбцами:

– № - номер по порядку (от 1 до 300);
– Секунда – текущая секунда на вашем ПК;
– Микросекунда – текущая миллисекунда на часах.

На каждой итерации цикла искусственно приостанавливайте скрипт на 0,01 секунды.

Задача 3.

Создайте файл json с любыми json данными.
Сохраните его в scv и xlsx.

Создайте (или загрузите из файла) объект DataFrame. Попробуйте сохранить его в json.

АПЛАНА
ГРУППА КОМПАНИЙ

Входит в ГК Аплана

АКАДЕМИЯ АЙТИ

Основана в 1995 г.

E-learning и очное обучение

Направления обучения:

- Информационные технологии
- Информационная безопасность
- ИТ-менеджмент и управление проектами
- Разработка и тестирование ПО
- Гос. и муниципальное управление

Филиалы:

Санкт-Петербург, Казань, Уфа, Челябинск, Хабаровск, Красноярск, Тюмень, Нижний Новгород, Краснодар, Волгоград, Ростов-на-Дону

Головной офис в Москве

Разработка программного обеспечения и информационных систем

Ежегодные награды Microsoft, Huawei, Cisco и другие

Программы по импортозамещению

Ресурсы более 400 высококлассных экспертов и преподавателей

Сеть региональных учебных центров по всей России

Крупные заказчики

СБЕР

РЖД

Билайн

TATNEFT

НОРНИКЕЛЬ

100+
сотрудников

АКАДЕМИЯ АЙТИ

Спасибо за внимание!

Центральный офис:
Москва, Варшавское шоссе 47, корп. 4, 7 этаж
Тел: +7 (495) 150-96-00
academy@it.ru
academyit.ru