ПМ

Домашнее задание к лекции «Продвинутый pandas»

Задание

Вопросы по заданию

Преподаватель: Константин Башевой

Домашнее задание

Задание 1

Для датафрейма log из материалов занятия создайте столбец source_type по следующим правилам:

- если источник traffic_source равен yandex или google, то в source_type ставится organic
- для источников paid и email из России - ставим ad
- для источников paid и email не из России - ставим other
- все остальные варианты берем из traffic_source без изменений

Задание 2

В файле URLs.txt содержатся url страниц новостного сайта. Вам необходимо отфильтровать его по адресам страниц с текстами новостей. Известно, что шаблон страницы новостей имеет внутри url следующую конструкцию: /, затем 8 цифр, затем дефис. Выполните следующие действия:

1. Прочитайте содержимое файла с датафрейм
2. Отфильтруйте страницы с текстом новостей, используя метод str.contains и регулярное выражение в соответствии с заданным шаблоном

Задание 3

1. Используйте файл с оценками фильмов ml-latest-small/ratings.csv. Посчитайте среднее время жизни пользователей, которые выставили более 100 оценок. Под временем жизни понимается разница между максимальным и минимальным значением столбца timestamp для данного значения userId.

Задание 4

2. Дана статистика услуг перевозок клиентов компании по типам (см. файл с кодом занятия). Необходимо сформировать две таблицы:

- таблицу с тремя типами выручки для каждого client_id без указания адреса клиента
- аналогичную таблицу по типам выручки с указанием адреса клиента

Обратите внимание, что в процессе объединения таблиц данные не должны теряться.

Решение

Зачет

LINK https://github.com/PeterM-lab/PYDA1/blob/main/11_groupby_merge/Untitled1.i... 18 апр. 2021

Вы загрузили решение 18 апр. в 20:23

Константин Башевой поставил(а) зачет 20 апр. в 18:07

Константин Башевой
ПРЕПОДАВАТЕЛЬ
20 апреля 2021 18:07

ПМ

Отлично получилось.

В первом задании можно также использовать вариант решения через метод loc. Например, для первого правила:

```
logs.loc[logs.traffic_source.isin(['yandex', 'google']), 'traffic_type'] = 'organic'
```

В задании 3 можно сделать необходимые вычисления в одной группировке с помощью метода agg:

```
ratings.groupby('userId').agg({'timestamp': ['min', 'max', 'count']})
```

Затем останется только отфильтровать по числу оценок и посчитать среднюю разницу min и max.

Вам понравилось?

Вопросы по заданию

Задайте вопрос — вам помогут одногруппники и эксперты

ПМ

На какой вопрос вы хотите получить ответ?

Добавьте более подробное описание вашего вопроса, если необходимо

Спросить

Все вопросы Помочь с ответом

Д Денис

Где взять данные для задания №4?

Искал:

- в материалах лекции netology.ru...ems/312994
- в слаке netology-ds.slack.com...6933071200

Но не нашел (

Написать ответ

Д **Денис** Ответ студента

В файле Python_13_join.ipynb

МГ Михаил Горбунов

Не пойму, где брать файл для 4 задания

Для выполнения четвертого задания, где данные брать?

Написать ответ

Ответов пока нет

ПМ