Описательная статистика. Анализ данных с помощью Pandas

Андрей Куртасов
Системный аналитик

Проверка связи

Отправьте «✚», если меня видно и слышно

Если у вас нет звука или изображения:

- перезагрузите страницу
- попробуйте зайти заново
- откройте трансляцию в другом браузере (используйте Google Chrome или Microsoft Edge)
- с осторожностью используйте VPN, при подключении через VPN видеопотоки могут тормозить

План семинара

1. Основные приемы работы с данными в Pandas,
   в т. ч. вычисление статистик
2. Практика работы с датасетами
3. Разбор заданий из LMS

МФТИ

Основные приемы работы с DataFrame

Основные приемы работы с DataFrame

1. Получение информации о DataFrame:
   a. Просмотр данных
   b. Задание типа данных для колонок
   c. Вычисление описательных статистик для числовых данных
2. Обращение к элементам данных и работа с индексами:
   a. Обращение к элементам DataFrame
   b. Срезы
   c. Изменение и сброс индекса
3. Фильтрация и сортировка данных
4. Модификация данных:
   a. Применение функций к столбцам и строкам таблицы
   b. Заполнение пустых ячеек
   c. Устранение дубликатов
   d. Добавление и удаление столбцов
5. Группировка, агрегация данных. Сводные таблицы
6. Объединение DataFrame

Что из этого вызывает больше всего трудностей или вопросов?

Упражнение 1. Получение и предобработка данных

1. Загрузить в pandas датасет о нарушениях с портала открытых данных: data.mos.ru/opendata/7702051094-prinyatye-mery-administrativnogo-vozdeystviya-za-narusheniya-vyyavlennye-pri-osushchestvlenii-litsenzionnogo-kontrolya.
2. Изучить структуру датафрейма.
3. Задать для столбцов типы.
4. Сделать столбец ID индексным.
5. Удалить из датафрейма дубликаты (при наличии).
6. Вывести все возможные типы нарушений.
7. Вывести количество случаев по каждому типу нарушения.

Ради развлечения 😀: попробуйте решить эти задачи с помощью ИИ:
https://www.wisedata.app/blog/transform-pandas-dataframe-with-nl

Упражнение 2. Еще один датасет и описательные статистики

Дан набор данных:
https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-02-18/food_consumption.csv

Описание датасета: https://www.nu3.de/blogs/nutrition/food-carbon-footprint-index-2018

1. Загрузить данные в Pandas, отобразить их полностью или частично в ноутбуке.
2. Разобраться с типами данных.
3. Получить описательные статистики для числовых колонок.
4. Для колонки `consumption` вычислить межквартильный размах и построить диаграмму ящик с усами.

Ресурсы в помощь:

- https://datagy.io/pandas-iqr/
- https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.boxplot.html

Зачем подсчитывать статистики?

МФТИ

1. Они позволяют понять характер данных.

2. Они дают возможность создавать *сводную* статистику.

Исходя из сводки по некоторому набору данных можно отвечать на вопросы:

- вероятность успешной продажи продукта;
- количество пассажиров на маршруте в определенные часы для оптимизации расписания;
- A/B-тесты: какая реклама более эффективна для привлечения людей к покупке продукта?

Для ответов на такие вопросы требуется подсчет статистических метрик.

Упражнение 3. Фильтрация и сортировка

Вывести таблицу: топ-5 стран по потреблению рыбы
(по убыванию уровня потребления).

Упражнение 4*

Вывести таблицу, полностью исключив из нее данные
о странах, которые демонстрируют показатель выбросов CO_2
выше 1000 при производстве хотя бы одного типа продуктов.

Упражнение 5. Группировка и агрегация

Вывести суммарный уровень выбросов CO_2 для каждой страны.

Решить задачу можно с помощью `groupby()` или `pivot_table()`.

Упражнение 6. Соединение датасетов

Соединить с датасетом еще один датасет, содержащий географические данные о странах (подобно `LEFT JOIN` в SQL):

https://raw.githubusercontent.com/google/dspl/master/samples/google/canonical/countries.csv

Какой функцией воспользуемся?

МФТИ

Разбор заданий из LMS

МФТИ

Проведем опрос

Как успехи с заданиями третьей недели?

МФТИ

Свободная дискуссия

Ваши вопросы? Пожелания?

До встречи!