Цели, которые необходимо достигнуть в процессе выполнения дз:

1. Проработать навык использования средств группирования и визуализации данных в Python
2. Попробовать процесс EDA (exploratory data analysis) - научиться задавать вопросы к данным

Задание состоит из двух частей. Первая на проработку навыков и изучение библиотек. Вторая на проработку процесса анализа, постановку вопросов и поиск необычного в данных.

1. **В предоставленном датасете проанализировать несколько характеристик и построить набор визуализаций**

Датасет имен https://github.com/wesm/pydata-book/tree/2nd-edition/datasets/babynames

- Сгруппируйте данные по полу и году и визуализируйте общую динамику рождаемости обоих полов
- Найдите самые популярные имена за всю историю
- Разбейте весь временной промежуток в данных на 10 частей и для каждой найдите самое популярное имя каждого пола. Для каждого найденного имени визуализируйте его динамику за все время
- Для каждого года рассчитайте сколько имен покрывают 50% людей, визуализируйте (мы увидим разнообразие имен за каждый год)
- Выберите 4 года из всего промежутка и отобразите для каждого года распределение по первой букве в имени, по последней букве в имени
- Составьте список из нескольких известных людей (президенты, певцы, актеры, киногерои) и оцените их влияние на динамику имен, постройте наглядную визуализацию

2. **Выбрать датасет и провести его анализ**

- Выбрать данные
- Составить список вопросов, по которым хотелось бы получить ответ в виде графиков (и расширять этот список вопросов в процессе выполнения задания)
- Построить соответствующие визуализации

Требования к выполнению задания - сделать не менее 5 визуализаций разного типа с фильтрацией и преобразованием данных разной сложности (5 раз df[‘column_i’].plot() на каждую колонку не принимается). Нужно постараться найти в данных что-то необычное или с помощью визуализаций показать характеристики данных.

Данные на выбор:

1. Данные, собранные в дз 1 (для выбора этого пункта есть смысл сначала понять, можно ли по данным построить разные визуализации)
2. Заново собранные данные (для выбора этого пункта есть смысл сначала понять, можно ли по данным построить разные визуализации) (можно кстати

попробовать собрать и визуализировать данные по вакансиям Data Science, Big Data в России)

3. Подобрать датасет самостоятельно. Например отсюда: https://habrahabr.ru/company/mailru/blog/339496/.com
4. Взять датасет из kaggle https://www.kaggle.com/datasets, например https://www.kaggle.com/usdot/flight-delays/data, параллельно можно поучаствовать в конкурсе https://www.kaggle.com/about/datasets-awards/kernels (по сути опубликовать свое выполненное дз, обратите внимание, что в конкурсе участвует только один датасет). При выполнении этого пункта можно и нужно подглядывать в уже опубликованные kernels, но надо понимать, что нельзя копировать из них код - нужно отработать навыки написания кода. И лучше сначала сделать по-максимуму анализ и визуализации, потом подсмотреть идеи, такая работа будет самой эффективной.

Дополнительные опции:

1. Можно опубликовать свою работу в виде статьи - на хабре, в личном блоге, в социальных сетях, в виде ядра на kaggle.
2. Можно обернуть свою работу в готовый продукт. Как пример того, что можно сделать, выложен код в директории vkstatsbot. Там лежит реализация телеграм бота, который принимает на вход ссылку на профиль в vk и возвращает картинку с простой визуализацией - идею можно развить. Идея бота выбрана потому, что требуется минимальный бэкенд и вообще не требуется фронтенда. Ссылку на продукт конечно же тоже можно выкладывать =)

Присылать на почту otus.bigdata.2017.11@gmail.com

Дедлайн - вечер воскресенья
Обратная связь будет дана до четверга