Задание на лабораторную работу №3

Подготовка

1. Скачать соответствующий варианту csv-файл из CORGIS Dataset Project по адресу https://corgis-edu.github.io/corgis/csv/.

Задание

1. Написать скрипт, анализирующий данные в csv-файле в соответствии с вариантом. Архитектура приложения должна быть готова обработать файл большого размера (сотни столбцов, миллионы строк). Каждый вариант предполагает три задания:
    1) Агрегация данных
    2) Дисперсия и доверительный интервал
    3) Изменение значения во времени и скользящее среднее
2. Каждый отдельный этап обработки (чтение файла, извлечение данных, агрегация) должен осуществляться в отдельном генераторе. Генераторы должны быть организованы в пайплайн. Допускается использование отдельного пайплайна для каждой задачи.
    1) Чтение файла осуществлять по частям с помощью функции `pandas.read_csv()` с параметром `chunksize`.
    2) Сформированный `DataFrame` передается дальше по цепочке генераторов.
3. Вывести результаты обработки в виде графика с помощью пакета `matplotlib`. Рекомендуемые график для каждого задания:
    1) Bar plot или Line plot
    2) Bar plot + доверительные интервалы
    3) Line plot

Дополнительное задание

4. Дополнить программу работой с Apache Parquet с помощью `pyarrow`:
    1) При чтении данных из csv-файла создается parquet-файл (если ещё не существует), и в дальнейшем данные берутся из него.
    2) Демонстрируется сравнение скорости чтения всех данных из csv-файла и parquet-файла целиком (без генератора).
    3) Выполнено доп. задание с использованием parquet-файла (посредством чтения только релевантных для задания столбцов, генератор для этого не требуется). Результат доп. задания представлен на графике (в виде scatter plot).

Требование к лабораторным работам

1 Код должен правильно работать.
2 Отсутствует дублирование кода / логики.
3 Отсутствует мусор (закомментированных строк, лишних переменных и т.д.).
4 Код должен быть читабельным (осмысленное название переменных и функций, прослеживается логика компоновки).
5 Соблюдается форматирование кода.
6 В коде присутствует документация.
7 В github репозитории нет лишних файлов / папок.

Варианты задания

В качестве датасетов использовать CORGIS Dataset Project по адресу:
https://corgis-edu.github.io/corgis/csv/

Вариант	Задание на обработку
1	Данные о полетах https://corgis-edu.github.io/corgis/csv/airlines/ **Задание:** 1. 3 «лучших» и 3 «худших» календарных месяца по доле задержанных/отмененных рейсов (в какое время года комфортнее летать) 2. 3 аэропорта с наибольшим и 3 аэропорта с наименьшим разбросом задержанных/отмененных рейсов за весь период наблюдений 3. Количество полетов в самом загруженном аэропорту за весь период наблюдений **Доп. задание:** 4. Корреляция между количеством задержанных/отмененных рейсов и общим количеством рейсов
2	Данные о погоде https://corgis-edu.github.io/corgis/csv/weather/ **Задание:** 1. 3 локации с самой высокой и 3 с самой низкой среднегодовой температурой 2. 3 штата с самым высоким и 3 с самым низким разбросом среднемесечных температур в течение года 3. Значение скорости ветра в самом «ветренном» штате за весь период наблюдений **Доп. задание:** 4. Корреляция между скоростью ветра (Wind.Speed) и осадками (Precipitation)
3	Данные о видеоиграх https://corgis-edu.github.io/corgis/csv/video_games/ **Задание:** 1. Лучший и худший годы для игровой индустрии с точки зрения продаж (Sales) 2. 3 издателя с наибольшим и 3 с наименьшим разбросом оценки игр (Review Score) 3. Общее количество выпущенных игр каждого возрастного рейтинга (Rating, значения “E”, “T”, “M”) в каждом году за период наблюдений **Доп. задание:** 4. Корреляцие между оценкой игры (Review Score) и заработанной на ней суммой (Sales)

4	Данные о выбросах парниковых газов https://corgis-edu.github.io/corgis/csv/global_emissions/ **Задание:** 1. 3 самые «зеленые» и 3 самые «грязные» страны по количеству выбросов на душу населения за все время наблюдений 2. 3 страны с наибольшим и 3 с наименьшим разбросом суммы выбросов 3. Общие ВВП (GDP) и общие выбросы за период наблюдений **Доп. задание:** 4. Корреляция между населением страны (Population) и количеством выбрасываемых парниковых газов
5	Данные о динамике экономической активности https://corgis-edu.github.io/corgis/csv/business_dynamics/ **Задание:** 1. 3 штата с наибольшим и 3 с наименьшим средним темпом создания рабочих мест (Net Job Creation Rate) 2. 3 штата с наиболее стабильным рынком труда и 3 с наиболее турбулентным – по величине разброса показателя (Reallocation Rate) 3. Динамика темпа закрытия рабочих мест (Job Destruction Rate) для наиболее нестабильного штата за все время наблюдений **Доп. задание:** 4. Корреляцие между темпом создания рабочих мест (Job Creation Rate) и закрытием рабочих мест (Job Destruction Rate)