Домашнее задание

Дисциплина	Python для инженерии данных
Тема	Тема 10. Spark: продолжение
Форма проверки	**Самопроверка. Студент выполняет задание и самостоятельно проверяет его**
Имя преподавателя	Дмитрий Клабуков
Время выполнения	1 час
Цель задания	Научиться агрегировать данные с помощью Spark
Инструменты для выполнения ДЗ	Jupyter Notebook или Google Colab
Правила приёма работы	Прикрепите ссылку в LMS на выполненное задание в Google Colab или GitHub (если вы использовали Jupyter Notebook). **Важно:** убедитесь в том, что по ссылке есть доступ в Google Colab (иногда там нет доступа для другого логина)
Критерии оценки	**Задание считается выполненным, если:** - прикреплена ссылка на файл с выполненным заданием, - доступ к файлу открыт, - код даёт правильный ответ к задаче. **Задание не выполнено, если:** - файл с заданием не прикреплён или отсутствует доступ по ссылке, - код выдаёт ошибку или даёт неправильный ответ.
Дедлайн	**7 дней с даты соответствующего вебинара**

Перед тем, как приступить к заданию, установите Jupyter Notebook либо используйте Google Colab.

В файле electronic_devices.csv лежит база данных покупателей.

Задача:

1. На основе файла electronic_devices.csv сгенерировать данные на 1 миллион строк с помощью sdv.

2. С помощью Spark прочитать данные, выбрать данные за «2024-09-03» и отфильтровать записи с одной дополнительной покупкой (столбец «addons»).
3. На основе полученных данных вычислить разницу между минимальной и максимальной ценой товара (столбец «unit_price»), разницу между минимальной и максимальной ценой заказа (столбец «total_price»). Данные необходимо группировать на основе пола, возраста, возраста и пола.
4. Данные можно собрать с помощью функции df.collect(), сохранять не нужно.
5. Необходимо получить среднее время выполнения агрегаций без кэширования после фильтра, а также с различными способами кэширования/сохранением контрольной точки.

Чек-лист самопроверки

Критерии выполнения задания	Отметка о выполнении
Установлен Jupyter Notebook либо используется Google Colab	
Создан профиль на https://github.com (при использовании Jupyter Notebook)	
Для вычислений использован Spark	
Получена разница между: - минимальной и максимальной ценой товара, - минимальной и максимальной ценой заказа. При этом данные сгруппированы на основе пола, возраста, возраста и пола	
Получено среднее время выполнения агрегаций без кэширования после фильтра, а также с различными способами кэширования/сохранением контрольной точки	
На учебной платформе прикреплена ссылка на	

выполненное задание в Google Colab или GitHub (если вы использовали Jupyter Notebook)	
Если используется Google Colab, то по ссылке есть доступ (иногда там нет доступа для другого логина)	