

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ М. В. ЛОМОНОСОВА
Факультет вычислительной математики и кибернетики

Кафедра исследования операций

ОТЧЕТ ПО ЗАДАНИЮ №3

Выполнили студенты:

Битиев Алексей
Кулакова Мария

Преподаватель:

Гусева Юлия

Москва
2020

Постановка задачи:

На вход программе подаются данные в формате CSV, содержащие информацию о закупках, продажах и инвентаре каждого магазина в разных штатах торговой сети Applepen.

В распоряжении есть информация о:

- 1) закупках (поставки яблок и карандашей два раза в месяц (1 и 15 числа каждого месяца)),
- 2) продажах (лог транзакций, по записи на каждую проданную позицию),
- 3) инвентарь (месячные данные общего количества яблок и карандашей на складе)

Эта информация предоставлена в следующих файлах:

- (имя штата)-(название магазина)-supply.csv
- (имя штата)-(название магазина)-sell.csv
- (имя штата)-(название магазина)-inventory.csv

Необходимо проанализировать имеющиеся данные и получить следующее:

- 1) Состояние склада *на каждый день* – данные о состоянии склада в конце каждого дня после того, как все поставки и продажи были совершены;
- 2) Месячные данные о количестве *сворованного* товара – данные должны хранить в себе дату, наименование товара и количество единиц украденного товара;
- 3) *Агрегированные данные* об объемах продаж и количестве сворованной продукции *по штату и году* – данные должны хранить в себе год, название штата, количество проданных единиц яблок и ручек, количество украденных единиц яблок и ручек.

На выходе программа должна выдавать следующие **CSV-файлы**:

- 1) (имя штата)-(название магазина)-daily.csv (первый пункт задания)
- 2) (имя штата)-(название магазина)-stolen.csv (второй пункт задания)

Решение:

Нужные для решения задачи входные и выходные данные хранятся на *GoogleCloud* и доступны по следующей ссылке:

<https://console.cloud.google.com/storage/browser/artem-pyanykh-cmc-prac-task3-seed17>

Для того чтобы скачать их все вместе был использован пакет консольных инструментов *gcloud* и команда *gsutil* для доступа к данным.

Файлы хранятся в формате CSV. Это *текстовый* формат, предназначенный для представления *табличных данных*. Строка таблицы соответствует строке текста, которая содержит одно или несколько полей, разделенных запятыми.

Нулевой этап:

Считываем для каждого магазина следующие таблицы:

- inventory, где находится реальное состояние склада в конце каждого месяца
- sell, где находятся транзакции на каждый проданный продукт
- supply, где находится информация о поставках

и записываем их в объекты типа датафрейм с помощью *read_csv*.

Вводим изначально нулевые переменные соответствующие проданным, текущим, поставленным и украденным товарам (ручки и яблоки). Подготавливаем текущую дату, выбирая первую дату из таблицы продаж.

Создаем 2 датафрейма с столбцами "date", "apple", "pen", в котором будет находится "идеальное" состояние склада, без учета сворованных товаров (*daily*) и отдельно сворованные товары на каждый месяц (*stolen*).

Первый этап:

Обработка данных начинается с таблицы продаж (sell), в которой находятся столбцы "date" и "sku_num". Смотрим на значение столбца "sku_num", в нем хранится следующая информация:

- ар или ре - наименование товара (яблоко, ручка)
- 123456... - идентификационный код, проведенной транзакции

по ней определяем какой товар был продан по идентификатору "ре" или "ар" и увеличиваем соответствующую переменную проданных товаров. Проверяем

дату, если это первое или пятнадцатое число месяца, то пересчитываем текущее количество яблок, с помощью таблицы поставок (supply). Так прорабатываем все строки таблицы продаж.

ИТОГ: получаем идеальное состояние склада на каждый день. Сохраняем его в файл формата CSV с помощью *to_csv*.

Второй этап:

Далее начинается работа с таблицей с реальным состоянием склада (inventory). В последний день продаж узнаем количество украденных товаров по формуле:

$$stolen_apple = current_apple - store_inventory['apple'][inventory_index]$$

и аналогично для ручек. Заносим полученные данные в таблицу сворованных товаров.

ИТОГ: получаем датафрейм, который хранит в себе информацию о сворованных товарах в конц каждого месяца.

Третий этап:

Теперь имея данные о продажах и сворованных товарах, мы можем сформировать агрегированные данные, которые находятся в файле *MS-total.csv*

Инструкции по запуску:

Библиотеки:

Язык программирования: python

Используемые библиотеки:

- pandas - библиотека, предоставляющая инструменты для обработки и анализа данных.
- csv - позволяет читать и записывать csv-файлы.

Программы

Для написания программы был использован дистрибутив Anaconda.

Реализация и запуск кода осуществлялись в Jupyter Notebook.

Вклад участников

Битиев Алексей - пункты 2 и 3, сборка

Кулакова Мария - пункт 1, readme