

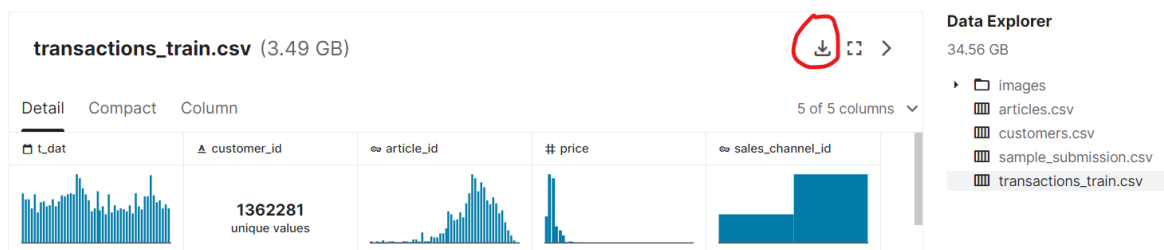
## Общее описание задания

Финальным заданием блока 1 стажировки является построение витрины данных (data mart). Что такое витрина данных можно узнать, например, из [этой](#) статьи, или из [этой](#).

Работа над данным заданием строиться на основе датасета из соревнования Kaggle - [H&M Personalized Fashion Recommendations](#). Полное описание набора данных, и детали каждого из составляющих его файлов вы найдете на странице по ссылке. **Обращаем внимание (!!!)**, что весь датасет не нужен, необходимы только следующие его части:

- articles.csv
- customers.csv
- transactions\_train.csv

Для того, чтобы загрузить не все данные, а необходимые файлы, нужно выбрать целевой, и в шапке над описанием колонок файла нажать на кнопку загрузки (демонстрация на картинке ниже).



## Описание подготовки рабочего окружения

Рабочим окружением в данном задании будет выступать РСУБД PostgreSQL. Первым шагом настройки, вам необходимо либо установить её в систему, либо воспользоваться Docker, если у вас есть в этом опыт, и развернуть PostgreSQL там. **Рекомендуем (!!!)** работать с Docker только если вы действительно знаете, что делать. Задание довольно объемное по времени, а на этом шаге стажировки не подразумевается погружение в данную технологию.

После того, как в каком-либо виде у вас будет установлена РСУБД, необходимо создать таблицы для каждого из файлов рабочего датасета. Обращаем ваше внимание на то, что вам не нужно создавать БД, осуществлять нормализацию и т.п. Таблицы необходимо оформить в виде DDL скриптов.

На этом, предварительная настройка закончена и можно приступать к реализации задания.

## Финальное задание блока 1

Спецификация витрины данных, которую вам необходимо построить в рамках задания приведена в таблице ниже:

Название поля	Описание
part_date	Дата окончания расчетного периода, например, "2023-08-31". Рассматриваемый для витрины период - месяц. Данный параметр будет всегда равен последнему дню целевого месяца. Это значит, что в приведенном выше примере, витрина должна отображать информацию и расчеты за август 2023 года.
customer_id	Идентификатор покупателя.
customer_group_by_age	Классификация покупателей по возрасту. Есть три возможных варианты - ["S", "A", "R"], которые соответствуют следующим границам: <ul style="list-style-type: none"> <li>• "S": &lt; 23 лет;</li> <li>• "A": 23-59 лет (включая границы диапазона);</li> <li>• "R": &gt; 59 лет.</li> </ul>
transaction_amount	Сумма всех покупок за период.
most_exp_article_id	Самая дорогая покупка. Если самых дорогих несколько, берем самую раннюю.
number_of_articles	Количество покупок за период.
number_of_product_groups	Количество групп товаров за период.
most_active_decade	Часть месяца с наибольшей суммой покупок. Есть три возможных варианта - [1, 2, 3], которые соответствуют следующим диапазонам: <ul style="list-style-type: none"> <li>• 1: 1-10 дни месяца;</li> <li>• 2: 11-20 дни месяца;</li> <li>• 3: 21-31 дни месяца.</li> </ul> Если суммы в декадах совпадают, берем более раннюю.
customer_loyalty	Показатель лояльности клиента. Есть два варианта - [0, 1], которые описываются следующим образом: значение 1 - если покупатель имеет покупки в текущем месяце + в двух предыдущих, в противном случае значение - 0. Есть два исключения: <ol style="list-style-type: none"> <li>1) Если выполняется расчет за первый месяц истории покупок - все покупатели, которые имеют покупки в этом месяце получают статус равным 1;</li> <li>2) Если выполняется расчет за второй месяц истории покупок - значение, равное 1, устанавливается в том случае, если покупатель имеет покупки в текущем периоде + в предыдущем.</li> </ol>

Финальный артефакт работы - скрипт для формирования таблицы, сформированной по вышеуказанным правилам и \*.csv файл этой таблицы.

Условия реализации:

- сформированные DDL и DML необходимо разместить в GitHub репозиторий;

- в том же репозитории необходимо сформировать readme файл, в котором должна находиться как минимум ссылка на итоговый \*.csv файл;
- в четверг (за день до даты финиша выполнения задания) вами будет получена дата part\_date, по которой вы будите параметризовать витрину и формировать файл для проверки работы нами.

Советы по решению задачи:

- не нужно строить один огромный запрос (!!!);
- промежуточные результаты можно сохранять, например, в виде VIEW;

## Проверка задания

Для контроля выполнения решения задания мы получим сформированные вами \*.csv файлы и сравним их с нашим эталонным вариантом. При полном совпадении, задача будет считаться решенной. Однако, защита задания также подразумевает ваше общение с ментором по коду решения.