## **Описание данных**

Схема afisha базы данных data-analyst-afisha содержит пять таблиц:

* purchases — информация о заказах билетов.
* events — данные о мероприятиях, доступных на платформе.
* venues — сведения о площадках проведения мероприятий.
* regions— список регионов, в которых проводятся мероприятия.
* city — список городов, относящихся к регионам.

#### Таблица purchases

Таблица содержит информацию о заказах билетов:

* order\_id — уникальный идентификатор заказа.
* user\_id — уникальный идентификатор пользователя.
* created\_dt\_msk — дата создания заказа (московское время).
* created\_ts\_msk — дата и время создания заказа (московское время).
* event\_id — идентификатор мероприятия из таблицы events.
* cinema\_circuit — сеть кинотеатров. Если не применимо, то здесь будет значение 'нет'.
* age\_limit — возрастное ограничение мероприятия.
* currency\_code — валюта оплаты, например rub для российских рублей.
* device\_type\_canonical — тип устройства, с которого был оформлен заказ, например mobile для мобильных устройств, desktop для стационарных.
* revenue — выручка от заказа.
* service\_name — название билетного оператора.
* tickets\_count — количество купленных билетов.
* total — общая сумма заказа.

#### Таблица events

Таблица содержит данные о мероприятиях:

* event\_id — уникальный идентификатор мероприятия.
* event\_name\_code — название мероприятия в закодированном виде.
* event\_type\_description — описание мероприятия.
* event\_type\_main — основной тип мероприятия: театральная постановка, концерт и так далее.
* organizers — организаторы мероприятия.
* city\_id — идентификатор города проведения мероприятия из таблицы cities.
* venue\_id — идентификатор площадки проведения мероприятия из таблицы venues.

#### Таблица venues

Таблица содержит сведения о площадках проведения мероприятий:

* venue\_id — уникальный идентификатор площадки.
* venue\_name — название площадки.
* address — адрес площадки.

#### Таблица city

Таблица содержит список городов:

* city\_id — уникальный идентификатор города.
* city\_name — название города.
* region\_id — идентификатор региона, к которому относится город, из таблицы regions.

#### Таблица regions

Таблица содержит список регионов:

* region\_id — уникальный идентификатор региона.
* region\_name — название региона.

## **Знакомимся с данными**

После того как подключение к данным было настроено, с помощью запросов SQL изучите данные. При этом помните, что данные для исследования настоящие, однако предоставлены сервисом в анонимизированном виде. Поэтому не удивляйтесь, если увидите необычные названия регионов, городов, площадок и билетных операторов.

Обратите внимание на:

* **Взаимосвязь между таблицами.** Какие поля являются первичными ключами в таблицах, а какие — внешними. Оцените типы связей: встречаются ли между таблицами отношения «один к одному», «один ко многим» или «многие ко многим».
* **Содержимое таблиц.** Соответствуют ли данные описанию и в каком объёме они представлены.
* **Корректность данных.** Проверьте уникальность идентификаторов, наличие пропусков, корректность написания категориальных данных, например типов устройств, названий городов и регионов, кодов валюты.
* **Распределение заказов по основным категориям.** Например, по типам мероприятий, устройствам, кодам валюты и другим категориям. Это поможет понять, как представлены данные. Обратите внимание на категории с небольшим объёмом данных.
* **Возможные аномалии или некорректные значения в данных.** Например, изучите статистические данные по полю выручки: встречаются ли выбросы или другие особенности.
* **Изучите период времени, за который представлены данные.** Проверьте, можно ли проследить влияние сезонности на данные.

При изучении данных фиксируйте возможные ошибки и другие особенности, на которые стоит обратить внимание при анализе. Это поможет избежать ошибок при построении дашборда и последующем анализе данных в Python. Во время знакомства с данными уделите большее внимание таблицам purchases и events — они содержат ключевую информацию о бронировании билетов, типах мероприятий и выручке сервиса с бронирования билетов.

Вы познакомились с данными и выявили несколько важных особенностей:

* **Выбросы в данных о выручке.** Зафиксированы высокие значения, которые могут повлиять на статистический анализ выручки с заказа.
* **Разные валюты.** Выручка представлена в российских рублях и казахстанских тенге, что необходимо учитывать при расчётах и построении дашборда.
* **Количество значений в категориальных данных.** Некоторые категории, такие как платформы, устройства и типы событий, содержат мало уникальных значений, а другие, как города, наоборот, представлены большим количеством значений. В первом случае можно изучить распределение выручки между категориями, а во втором — провести анализ топ-сегментов или использовать более крупную классификацию.
* **Распределение значений между категориями.** Обратите внимание, как представлены разные категории. Например, среди платформ видны лидеры, а также платформы с несколькими строками в базе. Такой дисбаланс следует учитывать при сравнительном анализе данных.

## **Изучаем ключевые метрики продукта**

Напомним, что коллеги из продуктового отдела просили разработать аналитический дашборд. С его помощью они хотят отслеживать динамику ключевых бизнес-показателей, анализировать популярность мероприятий и структуру выручки в разрезе категорий событий и устройств. Однако прежде чем приступать к созданию дашборда, разберитесь, какие метрики можно считать ключевыми, а затем найдите их значения с помощью SQL-запросов.

Из чего может состоять доход сервиса, аналогичного Яндекс Афише:

* **Комиссия от продажи билетов.** Это основной источник дохода или основная часть выручки сервиса. Сервис берёт процент от суммы с каждого забронированного билета. Комиссия может зависеть от типа мероприятия, региона и условий с партнёрами.
* **Сервисный сбор.** В зависимости от условий бронирования билетов и соглашений с партнёрами сервисный сбор может включаться в стоимость билета. Как правило, сервисный сбор берётся за обработку заказа, например за возможность вернуть билет, подключить уведомления и так далее. Сервисный сбор также может составлять определённую часть дохода сервиса и включаться в выручку компании.
* **Платное размещение объявлений и премиум-программы для партнёров.** Организаторы могут оформить дополнительный пакет услуг: продвижение объявлений в выдаче, расширенное описание мероприятий или дополнительная аналитика.
* **Партнёрские программы и реклама.** Сервис может получать доход от сотрудничества с другими компаниями, платёжными площадками, а также от размещения рекламных материалов.

Таким образом, в структуре дохода сервиса можно выделить две категории выручки: ту, которую получают за счёт комиссии от продажи билетов и сервисного сбора, а также дополнительную выручку за счёт платного размещения объявлений, рекламы и партнёрских программ.

В вашем распоряжении есть данные только о выручке с продажи или бронирования билетов — поле revenue. Сюда входит и комиссия от продажи, и сервисный сбор. Выручка сервиса является одним из главных показателей, который показывает эффективность продукта.