**Исследование данных стримингового сервиса**

Оглавление

[1. Ситуация. Исходные данные 1](#_Toc154846771)

[2. Проблема - задача и алгоритм выполнения 2](#_Toc154846772)

[Общая задача: 2](#_Toc154846773)

[Алгоритм выполнения: 2](#_Toc154846774)

[1 Создать подключение к базе данных и сформировать запрос в результате которого получится таблица, содержащая: 2](#_Toc154846775)

[3. Подготовка данных для аналитики 3](#_Toc154846776)

[3.1. Подключение к базе, формирование запроса и сохранение результата в датафрейм 3](#_Toc154846777)

[3.2. Оценка датафрейма на состоятельность, EDA 3](#_Toc154846778)

[3.2.1. Проверка на пропуски: 3](#_Toc154846779)

[3.2.2. Проверка на дублирование: 3](#_Toc154846780)

[3.2.3. Приведение данных к первому числу месяца. 3](#_Toc154846781)

[4. Исследование данных 4](#_Toc154846782)

[4.1. Визуализация зависимостей 4](#_Toc154846783)

[4.2. Доп. таблицы 4](#_Toc154846784)

[5. Заключение. 5](#_Toc154846785)

[5.1. Общие выводы 5](#_Toc154846786)

# **Ситуация. Исходные данные**

1. Исходные данные: файл chinook.db

Представлена упрощённая модель данных стримингового сервиса:

Изображение выглядит как текст, диаграмма, снимок экрана, Шрифт

Автоматически созданное описание

# **2. Проблема - задача и алгоритм выполнения**

## **Общая задача:**

Оценить, если ли сезонность в покупке треков разных жанров.

## **Алгоритм выполнения:**

# 1 Создать подключение к базе данных и сформировать запрос в результате которого получится таблица, содержащая:

- данные о чеках (invoice): id, дату;

- id покупателя и id всех треков в чеке;

- жанр каждого трека: объединить все тяжёлые жанры (Rock, Alternative & Punk, Metal, Alternative, Heavy Metal) в категорию ‘rock’, а остальные жанры — в категорию ‘others’;

- стоимость трека,

- в таблицу должны попасть данные только по аудиофайлам (media\_type) за 2020 год

2. На основе запроса создать датафрейм (read\_sql\_query) в pandas

3. Проверить датафрейм на пропуски и дубликаты (если они есть, предположить их происхождение и решить, нужно ли их обрабатывать

4. Привести дату к первому числу каждого месяца

5. Сгруппировать данные, и через seaborn построить графики, чтобы оценить по категориям жанров (‘rock’ и ‘others’) следующие ежемесячные показатели:

- общее количество купленных треков,

- число чеков,

- число покупателей,

- общую стоимость треков.

7. Сделать выводы по имеющимся данным

# **Подготовка данных для аналитики**

## **Подключение к базе, формирование запроса и сохранение результата в датафрейм**

Состав столбцов в таблице chinook.db :

Запрошенная таблица данных имеет в итоге 1150 строк и 6 столбцов:

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

## **Оценка датафрейма на состоятельность, EDA**

### **Проверка на пропуски:**

Изображение выглядит как текст, снимок экрана

Автоматически созданное описание

### **Проверка на дублирование:**

Изображение выглядит как текст, Шрифт, снимок экрана, число

Автоматически созданное описание

В датафрейме установлены 2 строки-дубликата. Они в данном случае могут возникнуть из-за того, что покупатель ошибочно дважды произвел покупку одного и того же трека.

Удалять их нельзя, т.к. покупка состоялась и удаление повторных строк даст неверный результат по общей сумме покупок в базе.

### **Приведение данных к первому числу месяца.**

Даты приведены к первому числу месяца, для удобства группировки по месяцам.

# **Исследование данных**

## **Визуализация зависимостей**

Изображение выглядит как снимок экрана, текст, Параллельный, дизайн

Автоматически созданное описание

## **Доп. таблицы**

## 

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описаниеТаблица количества треков по месяцам Таблица количества покупок покупателями по месяцам

# **Заключение.**

## **Общие выводы**

1. График распределения количества купленных треков и полученных сумм имеют абсолютно идентичные формы. Объясняется это тем, что в указанный период цена всех треков была одна и та же = 0,99. Соответственно для рассмотрения динамики по году надо брать только один из указанных графиков, второй не несет никакой смысловой нагрузки и может быть исключен из аналитики.

2. Очень схожие формы имеют графики распределения чеков и покупателей - это объясняется тем, что абсолютное большинство покупателей делают покупку один раз в месяц и лишь в отдельные месяцы есть факты (достаточно единичные) покупки одним покупателем более 1 раза в месяц. При этом такие покупатели есть в обеих категориях - и "rock", и "others".

3. По всем видам графиков видно явное преимущество треков в категории "rock".

4. Среднее количество покупок треков в жанре "rock" в одном чеке существенно выше, чем в жанре "others" - это следует из соотношения графиков распределения треков и распределения чеков (покупателей)

5. Общая форма графика приобретения имеет периоды пиков и минимумов, что рассматриваемый период недостаточен для формирования общих выводов о каких-либо сезонных принципах этого фактора

6. Общее количество чеков в периоде не соответствует суммам по жанрам, что свидетельствует о том, что предпочтения некоторых покупателей разноплановы - в одном чеке и одним покупателем могут приобретаться как треки жанра "rock", так и треки других жанров