ГУАП

КАФЕДРА № 41

ОТЧЕТ   
ЗАЩИЩЕН С ОЦЕНКОЙ

ПРЕПОДАВАТЕЛЬ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ассистент |  |  |  | В.В. Боженко |
| должность, уч. степень, звание |  | подпись, дата |  | инициалы, фамилия |

|  |
| --- |
| ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ |
| Предварительный анализ данных |
| по курсу: ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ |
|  |
|  |

РАБОТУ ВЫПОЛНИЛ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| СТУДЕНТ ГР. № | 4917 |  |  |  | Д.В. Потемкин |
|  |  |  | подпись, дата |  | инициалы, фамилия |

Санкт-Петербург 2022

1. **Цель работы**

осуществить предварительную обработку данных csv-файла, выявить и устранить проблемы в этих данных.

**Вариант 15:** music

1. **Ход работы**

Загрузим датасет 5music.csv с помощью библиотеки pandas в Jupyter-ноутбуке и выведем первые 20 строк при помощи команды head (рисунок 1):

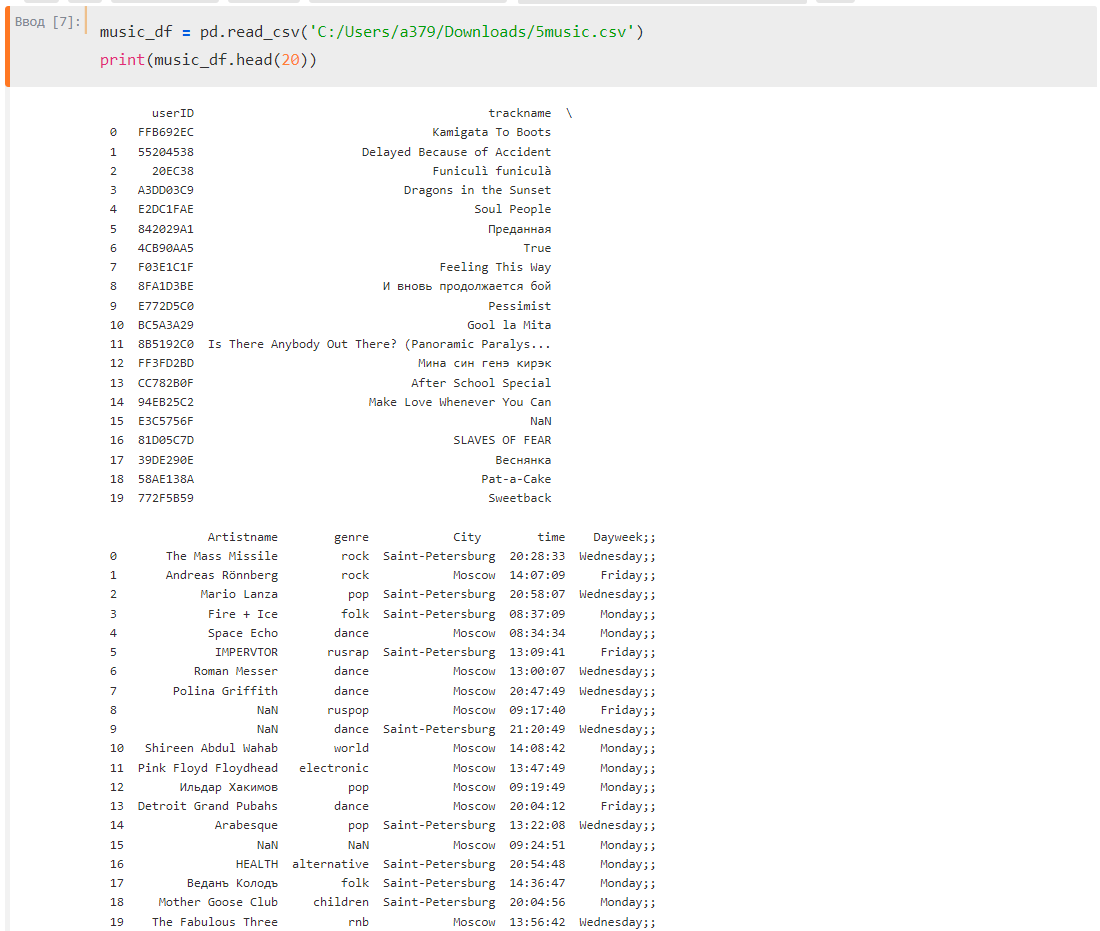


Рисунок 1 – Первые 20 строк csv файла

Файл 5music.csv содержит информацию о прослушанных треках пользователя:

1. userID - идентификатор пользователя;

2. Trackname - название трека;

3. Artistname - имя исполнителя;

4. Genre - название жанра;

5. City - город пользователя;

6. Time - время начала прослушивания;

7. Dayweek – день недели.

С помощью метода .info оценим данные (Рисунок 2):

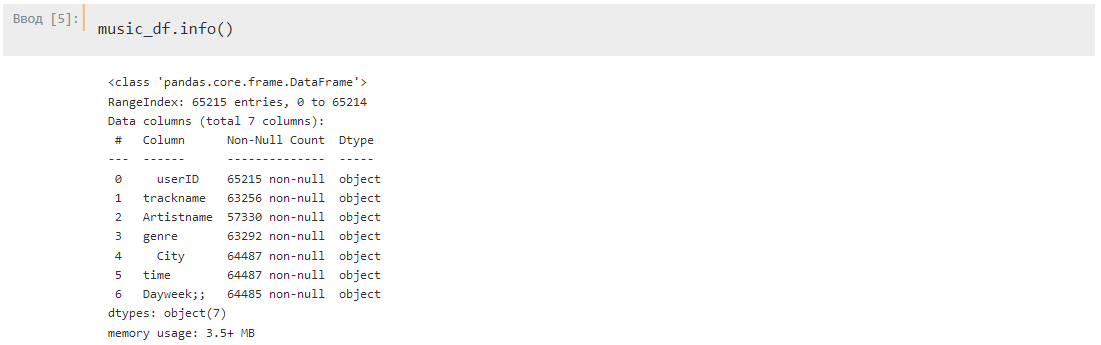


Рисунок 2 – Краткое описание набора данных

Из полученных данных видно, что только первый столбец с идентификатором пользователя не имеет пропусков данных. Больше всего пропусков данных содержит столбец Artistname. Все столбцы имеют тип данных object.

Выведем на экран названия столбцов с помощью df.columns. Выявим и устраним проблемы с названиями (Рисунок 3):



Рисунок 3 – Исправление заголовков столбцов

Все названия столбцов были приведены к общему виду, были удалены пробелы, из заголовка Dayweek;; были убраны лишние символы.

Также уберём все “;” из последнего столбца (Рисунок 4):

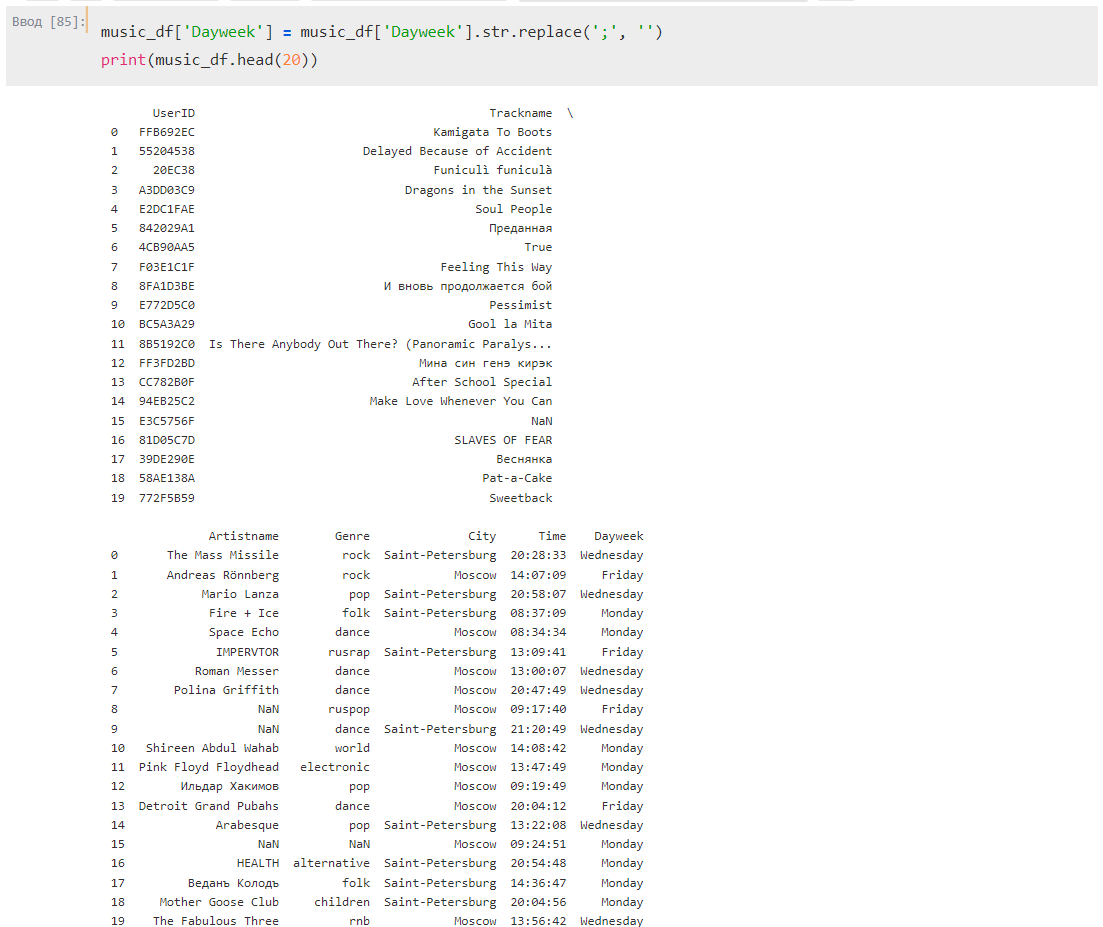


Рисунок 4 – Удаление ‘;’ из последнего столбца

Удалим строки с пропусками в столбцах “trackname” и “dayweek”, так как данные в данных столбцах могут быть важными для обработки. Данные строки включают все пропуски столбцов “city” и “cime”. После заменим пропущенные значения в столбцах “artistname” и “genre” на “Unknown”, поскольку это позволит обрабатывать данные, как неизвестные (Рисунок 5):

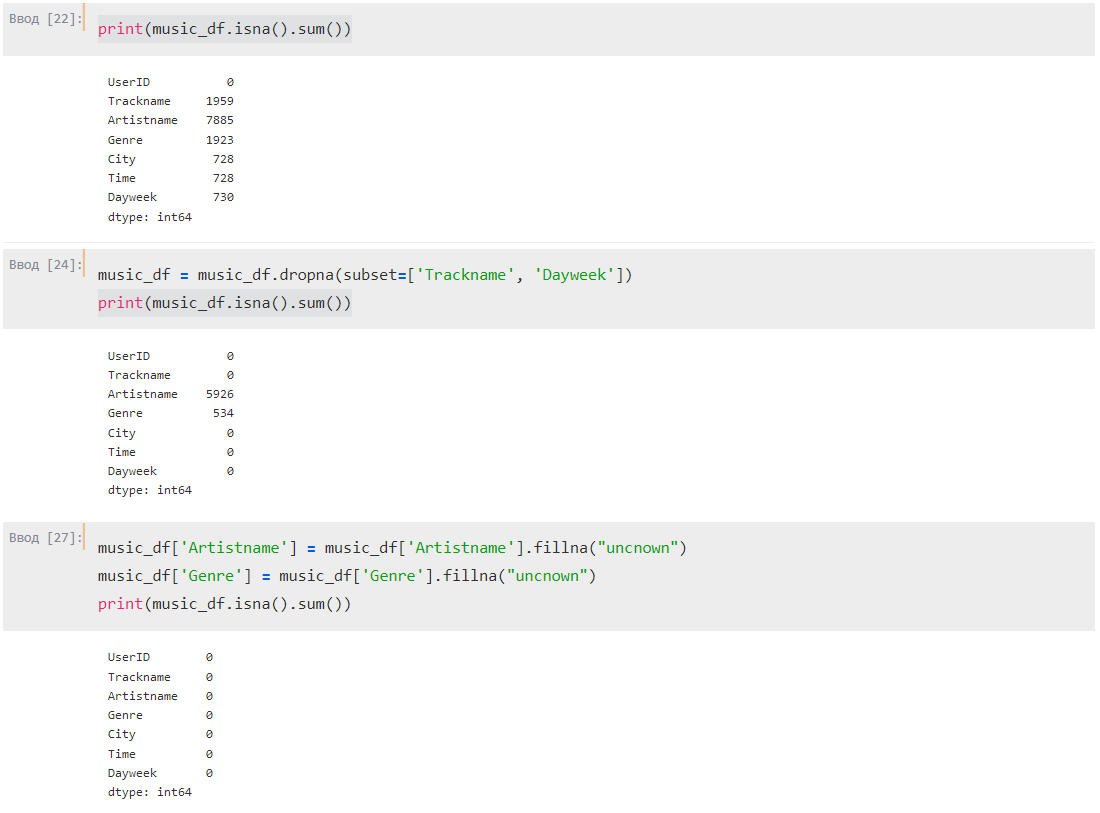


Рисунок 5 – Устранения пропусков в наборе данных

Проверим данные на наличие дубликатов и удалим их если они есть (Рисунок 6):

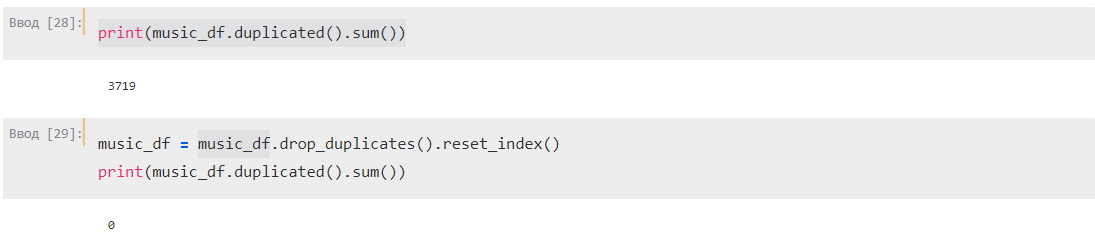


Рисунок 6 – Удаление дубликатов

Было выявлено и удалено 3719 строк- дубликатов.

Все типы данных соответствуют действительности

Создадим сводную таблицу, содержащую минимальное время прослушивания музыки для каждого дня в городах (Рисунок 7):

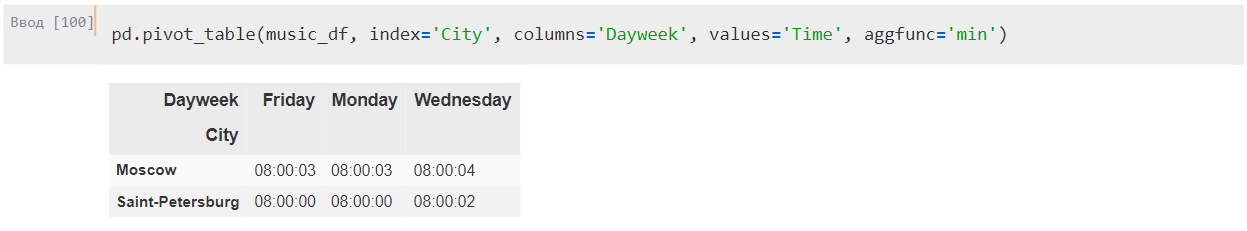


Рисунок 7 – Сводная таблица минимальных времён

Сгруппируем данные по жанру с применением агрегатной функции sum, чтобы вывести топ 10 прослушиваний каждого жанра (Рисунок 8):

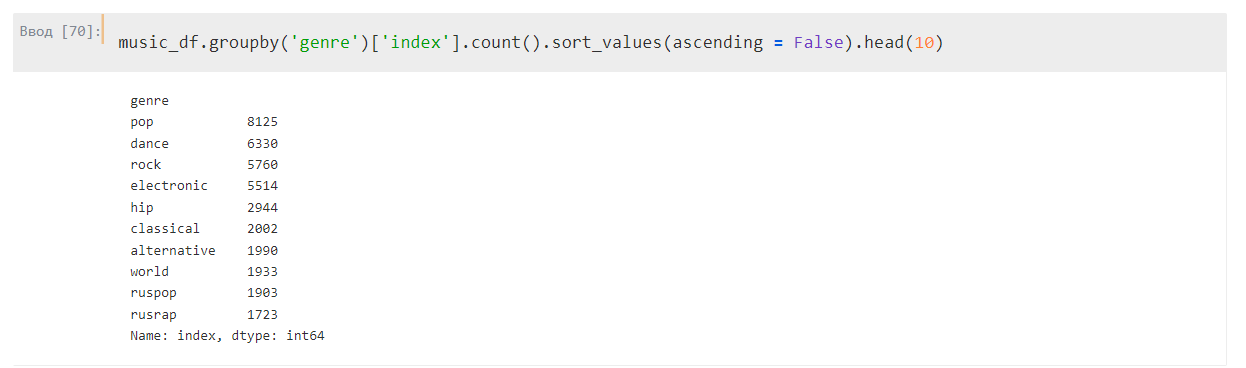


Рисунок 8 – Количество прослушиваний каждого жанра

Создадим сводную таблицу, показывающую количество прослушиваний по дням недели (Рисунок 9):

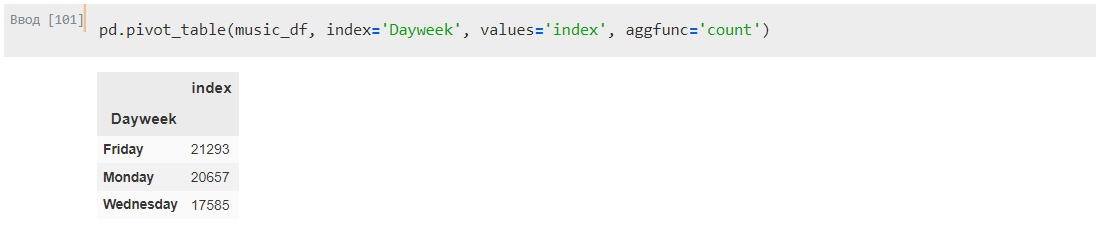


Рисунок 9 – Количество прослушиваний по дням недели

Ссылка на GitHub с проектом: https://github.com/DmitryPotemkin/VvAD\_Labs/blob/main/VvADLab1.ipynb

1. **Вывод:**

В ходе выполнения лабораторной работы были получены навыки обработки данных и их первичного анализа. Для обработки при помощи библиотеки pandas был загружен набор данных с прослушанными треками пользователя. Были удалены пропуски в данных, исправлены названия столбцов и удалены дубликаты. Были придуманы и построены две сводные таблицы и произведена группировка данных по жанру. Из сводной таблицы минимальных времён прослушивания музыки для каждого дня в городах видно, что как правило это значения возле 8 часов. Скорее всего в это время начинался сбор информации. Из группировки данных по жанру с применением агрегатной функции sum мы узнали, что самым популярным жанром ожидаемо является “pop”. Из сводной таблицы, показывающей количество прослушиваний по дням недели, можно увидеть, что чаще всего пользователи прослушивали музыку в пятницу.

**Дополнительное задание:**

выберите топ 5 жанров по кол-ву прослушиваний и создайте сводную таблицу с количеством прослушиваний этих жанров для каждого города

Выберем топ 5 жанров по количеству прослушиваний (Рисунок 10):

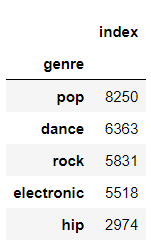


Рисунок 10 – Топ 5 жанров по количеству прослушиваний

Создадим сводную таблицу с количеством прослушиваний топа 5 жанров для каждого города (Рисунок 11):



Рисунок 11 - Сводная таблица с количеством прослушиваний топа 5 жанров для каждого города