МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

«Омский государственный технический университет»

Факультет информационных технологий и компьютерных систем

Кафедра «Прикладная математика и фундаментальная информатика»

|  |  |
| --- | --- |
| **Домашнее задание** | |
|  | |
| по дисциплине | Практикум по программированию |

|  |  |  |  |
| --- | --- | --- | --- |
| Студента(ки) | Раевских Максима Игоревича | | |
|  | фамилия, имя, отчество полностью | | |
| Курс | 2 | Группа | ФИТ-222 |
| Направление | 02.03.02. Фундаментальная информатика и информационные технологии | | |
|  | код, наименование | | |
| Руководитель | Старший преподаватель | | |
|  | должность, ученая степень, звание | | |
| Саматов А. П. | | |
|  | фамилия, инициалы, дата, подпись | | |
| Выполнил |  | | |
|  | дата, подпись студента(ки) | | |

|  |  |
| --- | --- |
| Итоговый рейтинг |  |

Омск 2023

**Введение**

Анализ данных — это мощный инструмент, который позволяет извлекать ценные знания и информацию из больших объемов данных. В современном мире, где данные играют ключевую роль в различных областях, анализ данных становится неотъемлемой частью принятия решений в бизнесе, науке, здравоохранении и многих других областях.

Анализ данных становится все более актуальным по нескольким причинам:

1. Большие объемы данных: С появлением больших объемов данных (Big Data) компании и организации могут извлекать ценные знания из своих данных для улучшения бизнес-процессов и принятия обоснованных решений.

2. Персонализации: Анализ данных позволяет компаниям создавать персонализированные продукты и услуги, учитывая потребности и предпочтения клиентов.

3. Научные исследования: В научных исследованиях анализ данных помогает выявлять закономерности, строить модели и тестировать гипотезы.

4. Здравоохранение: В медицине анализ данных может помочь в диагностике, прогнозе заболеваний и улучшении качества медицинского обслуживания.

5. Финансы: В финансовой сфере анализ данных помогает в принятии решений о портфеле инвестиций, оценке рисков и выявлении мошенничества.

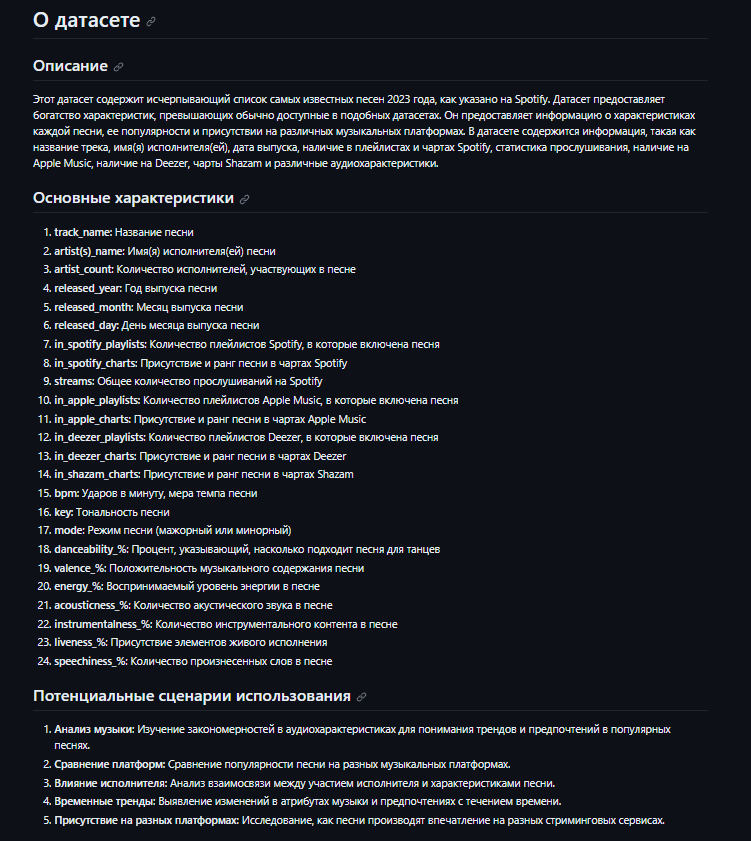
Инструменты для анализа данных.

Pandas — это библиотека для анализа данных, которая предоставляет структуры данных, такие как DataFrame и Series, для удобной работы с табличными данными. Она позволяет выполнять операции фильтрации, сортировки, агрегации и манипуляции данными. Pandas также обеспечивает удобный доступ к данным из различных источников, таких как CSV, Excel, SQL и другие.

NumPy — это библиотека для работы с многомерными массивами и выполнения вычислений над ними. Она предоставляет эффективные инструменты для выполнения математических и статистических операций. NumPy является основой многих других библиотек для анализа данных, таких как Pandas и SciPy.

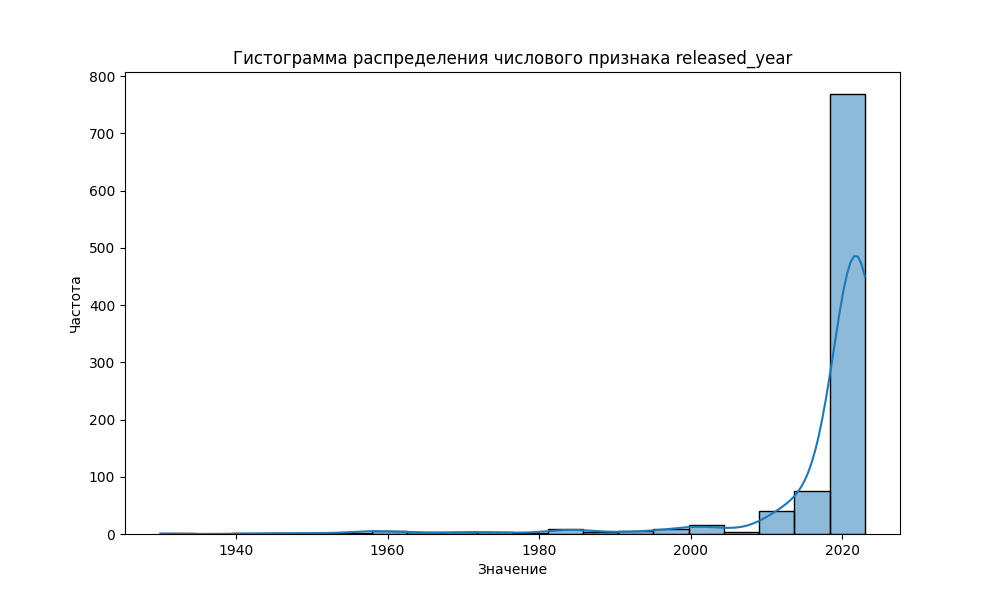
SciPy — это библиотека для выполнения научных и инженерных вычислений. Она включает в себя множество модулей и функций для оптимизации, статистики, обработки сигналов и многих других задач. SciPy также предоставляет статистические методы, такие как тесты на нормальность и регрессию, которые полезны в анализе данных.

Для этого анализа бал выбран датасет статистики песен приложения Spotify, который можно найти на сайте Kaggle**[1]**. Был загружен датасет в Jupyter Notebook**[2]** и создано его описание



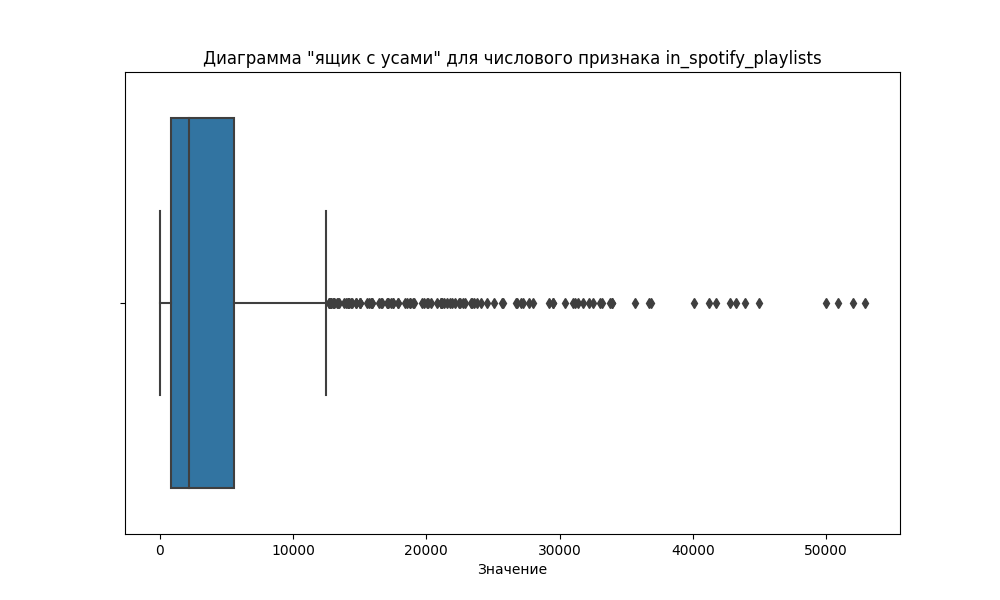
Далее проведен разведывательный анализ данных и построено несколько визуализаций. Вот некоторые из них:

Гистограмма распределения числового признака released\_year:



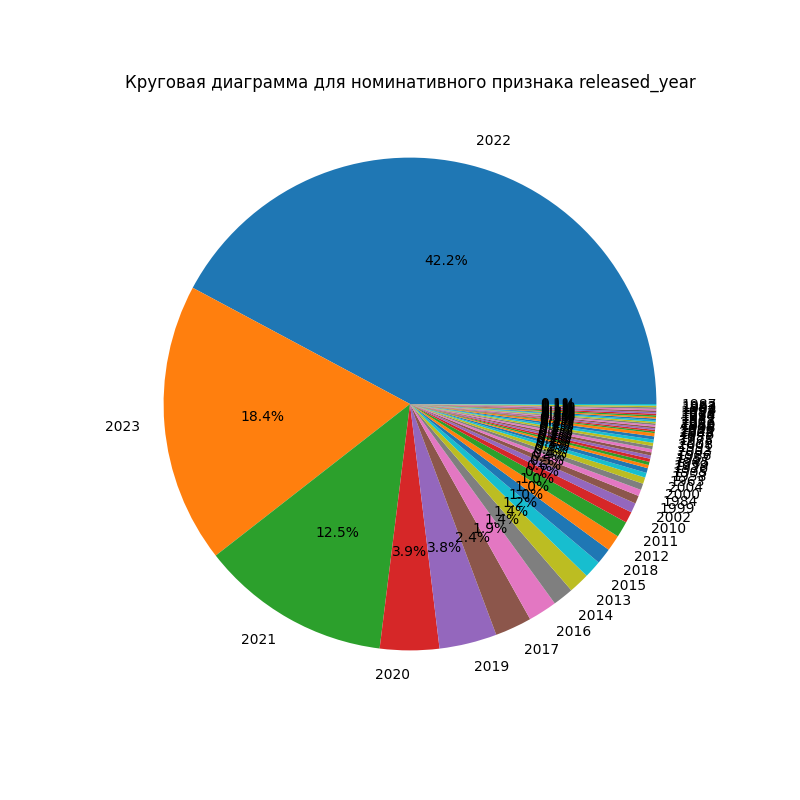
Вывод: Большое количество песен в Spotify было выпущено в 2020-х годах, тогда как музыки из 20 века очень мало. Это ненормальное распределение с выбросами.

Диаграмма "ящик с усами" для числового признака in\_spotify\_playlists:



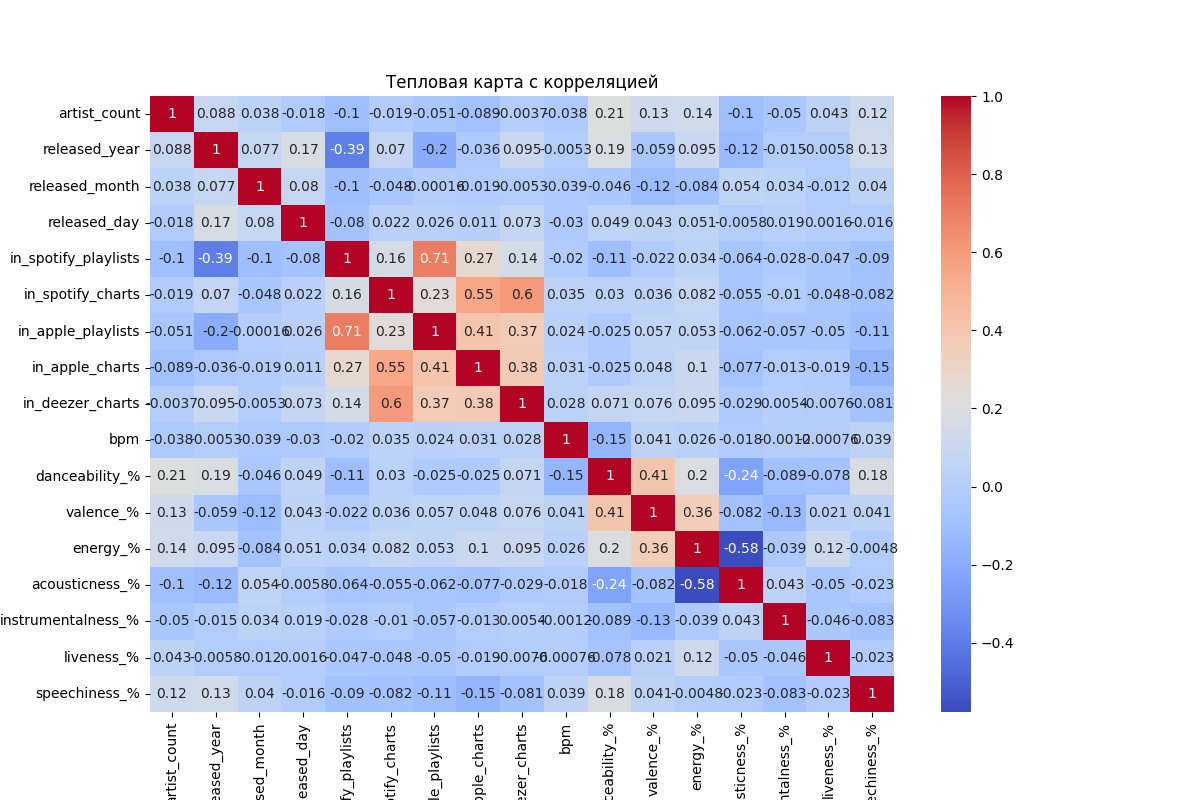
Вывод: здесь так же видно ненормальное распределение. Черные кружки на диаграмме показывают количество выбросов. Две полосы на «усах» 25 и 75 процентиль, а длинная вертикальная черта – медиану.

Круговая диаграмма для номинативного признака released\_year:



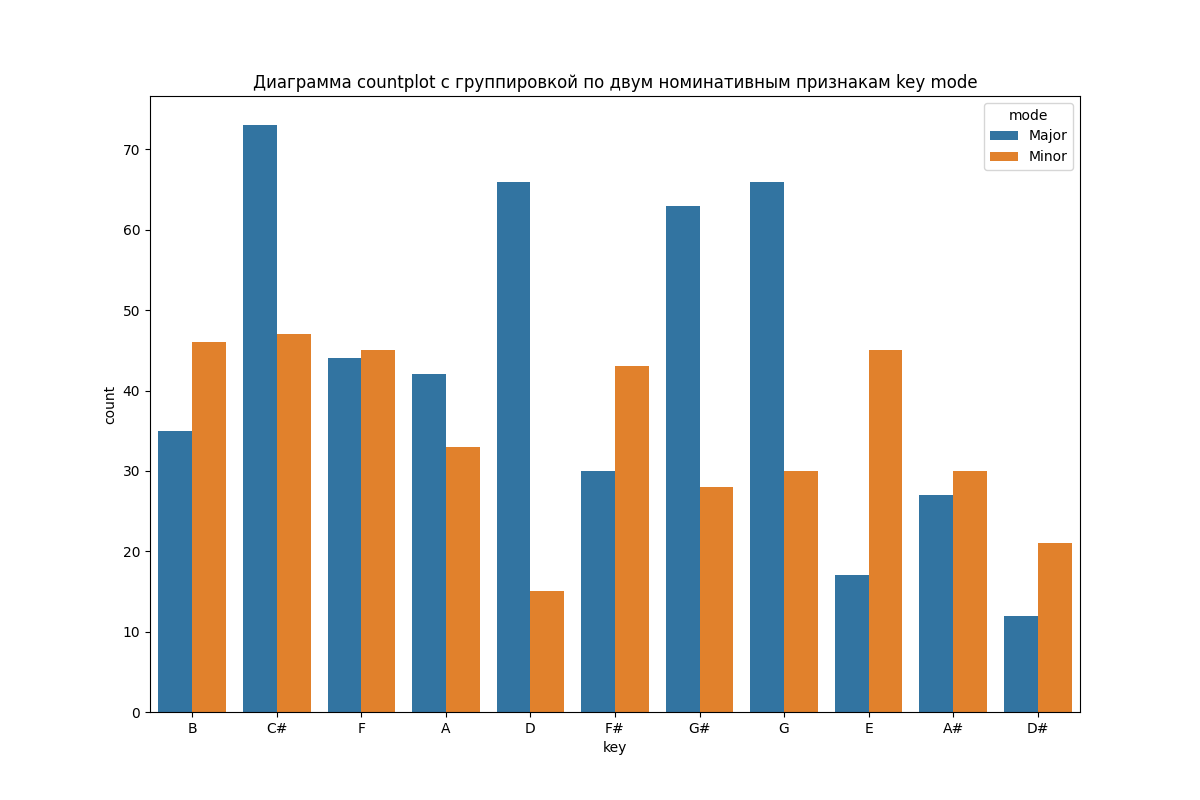
Вывод: на этой диаграмме показано в процентах количество песен на каждый год. На ней хорошо видно, что в 2022 году выпустили ненормально много треков, по сравнению даже с 2021 и 2020 годами.

Тепловая карта с корреляцией между признаками:



Вывод: здесь показана корреляция между всеми числовыми данными. Видно, что большинство из них имеют достаточно не большую корреляцию друг с другом. Однако есть те у которых она сильно отличается.

Диаграмма countplot с группировкой по двум номинативным признакам key mode:



Вывод: на данный диаграмме показана количество музыки с высокой или низкой тональностью для преобладающей ноты в песне.

Мы также провели предварительную обработку данных, включая проверку наличия пропусков и их заполнение. Пример работы с пропусками:

if df.isnull().any().any():

    for column in df.columns:

        if df[column].dtype == 'int64':

            # Заполните пропуски медианой для целых чисел

            df[column].fillna(df[column].median(), inplace=True)

        elif df[column].dtype == 'float64':

            # Заполните пропуски средним значением для действительных чисел

            df[column].fillna(df[column].mean(), inplace=True)

        else:

            # Заполните пропуски модой для остальных типов данных

            df[column].fillna(df[column].mode().iloc[0], inplace=True)

Были заполнены пропуски в целочисленных данных медианой, в действительных числах средним значением и в остальных модой.



При выполнении команды показывается, что количество пропущенных значений во всех столбцах стало нулевым.

И в конце предобработанные данные были сохранены в новый файл spotify.csv. С помощью команды “ df\_encoded.to\_csv('../content/sample\_data/spotify\_encoded.csv', index=False)”, которая принимает два аргумента: путь, по которому сохраняется файл, и аргумент, который говорит команде не сохранять в датасет индексы как отдельный столбец.

**Заключение**

В заключении этого анализа, были изучены и проанализированы данные с сайта прослушивания музыки, а так же библиотека Pandas**[3]**. Был проведен разведывательный анализ, предварительная обработка данных и выявлены интересные закономерности. Анализ данных с использованием этих инструментов позволяет делать более информированные решения и извлекать ценную информацию из данных.

**Список литературы**

1. NIDULA ELGIRIYEWITHANA Most Streamed Spotify Songs 2023 / NIDULA ELGIRIYEWITHANA [Электронный ресурс] // Kaggle : [сайт]. — URL: https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023 (дата обращения: 02.10.2023).
2. / [Электронный ресурс] // jupyter : [сайт]. — URL: https://jupyter.org (дата обращения: 02.10.2023).
3. pandas documentation / [Электронный ресурс] // Pandas : [сайт]. — URL: https://pandas.pydata.org/docs/ (дата обращения: 02.10.2023).