

[ВВЕДЕНИЕ 3](#_bookmark0)

[1.Поиск и загрузка данных 4](#_bookmark1)

* 1. [Гистограмма распределения числового признака 5](#_bookmark2)
  2. [Диаграмма «ящик с усами» числового признака 6](#_bookmark3)
  3. [Круговая диаграмма номинативного признака](#_bookmark4) 7
  4. [Тепловая карта](#_bookmark5) 8
  5. [Диаграмма countplot с группировкой по двум номинативным признакам](#_bookmark6) 9

[3Предварительная обработка данных](#_bookmark7) 10

[ЗАКЛЮЧЕНИЕ](#_bookmark8) 12

[СПИСОКИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ](#_bookmark9) 13

ВВЕДЕНИЕ

Объемы накопленных данных в настоящее время настолько внушительны, что человеку просто не по силам проанализировать их самостоятельно, хотя необходимость проведения такого анализа вполне очевидна, ведь в этих "сырых данных" заключены знания, которые могут быть использованы при принятии решений, формировании статистических отчетов или составлении моделей машинного обучения. В ходе изучения курса были использованы следующие библиотеки для языка программирования Python:

1. NumPy—библиотека с открытым исходным кодом с поддержкой многомерных массивов(включая матрицы) и высокоуровневых математических функций, предназначенных для работы с многомерными массивами.
2. Matplotlib—это библиотека для визуализации данных. В ней можно построить двумерные(плоские) и трехмерные графики.
3. SymPy—это библиотека Python с открытым исходным кодом, используемая для символьных вычислений. Она предоставляет возможности компьютерной алгебры в виде отдельного приложения.
4. SciPy — библиотека с открытым исходным кодом, предназначенная для выполнения научных и инженерных расчётов.
5. Pandas — программная библиотека для обработки и анализа данных. Предоставляет специальные структуры данных и операции для манипулирования числовыми таблицами и временными рядами.
6. Seaborn — библиотека для создания статистических графиков на Python. Она построена на основе matplotlib и тесно интегрируется со структурами данных pandas. Эти библиотеки позволяют проводить обработку, анализ и визуализацию данных, строить статистику на их основе.
   1. Поиск и загрузка данных

Использован датасет Movies, в котором содержится информация о большом количестве фильмов.

Датасет был загружен в ноутбук командой *read\_csv()* библиотеки *pandas*.

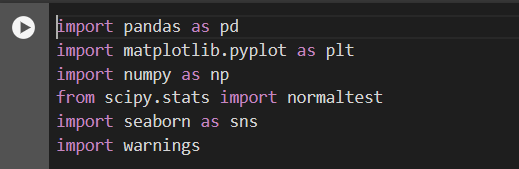


Рисунок 1 - Импорт библиотек

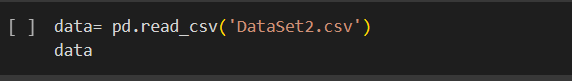


Рисунок 2 - Загрузка датасета

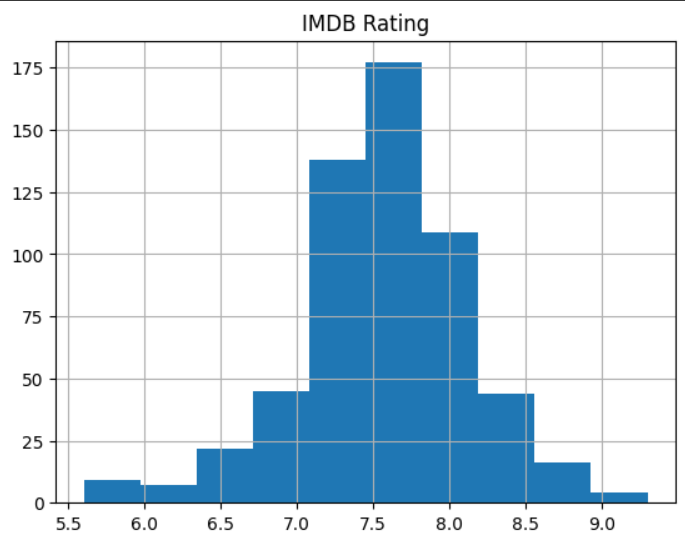


Рисунок 3 –небольшая часть датасета, выведенного в виде таблицы

Данный датасет выглядит как набор из 571 строк × 30 столбцов, разделенных точкой с запятой, в которых описаны такие данные о фильмах, как дата выпуска, жанр, рейтинг, названия, режиссер, актеры и т.д.

* 1. Гистограмма распределения числового признака

Гистограмма—способ представления табличных данных в графическом виде—в виде столбчатой диаграммы. Количественные соотношения некоторого показателя представлены в виде прямоугольников, площади которых пропорциональны. На гистограмме видно количество оценок и сопоставленный этому количеству рейтинг.



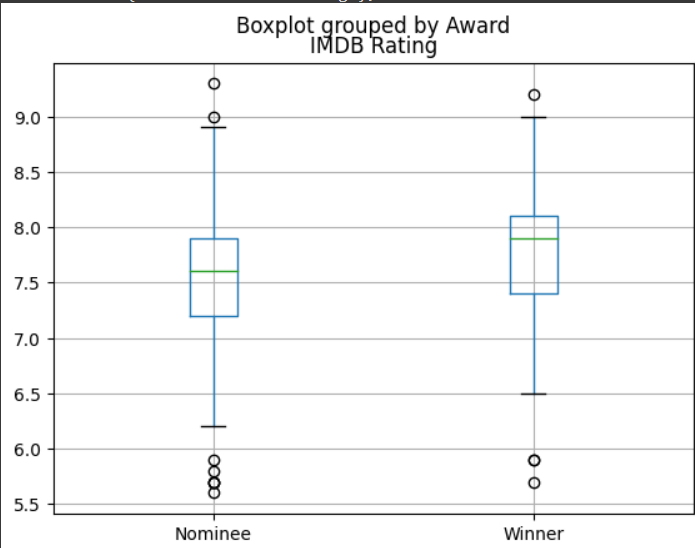
# Рисунок 4 – гистограмма столбца IMDB Rating

* 1. Диаграмма «ящик с усами» числового признака

Диаграмма «ящик с усами» — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей. Такой вид диаграммы в удобной форме показывает медиану (или, если нужно, среднее), нижний и верхний квартили, минимальное и максимальное значение

выборки и выбросы. Несколько таких ящиков можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим; их можно располагать как горизонтально, так и вертикально. Расстояния между различными частями ящика позволяют определить степень разброса(дисперсии) и асимметрии данных и

выявить выбросы. На диаграмме видно что выбросов мало.



# Рисунок 5 – Диаграмма «ящик с усами» столбца Award

* 1. Круговая диаграмма номинативного признака

Круговая диаграмма — это круговая статистическая диаграмма, которая разделена на срезы, чтобы проиллюстрировать числовую пропорцию. На круговой диаграмме длина дуги каждого среза пропорциональна величине, которую он представляет. На данной круговой диаграмме видно, что распределение времени фильма практически равномерно

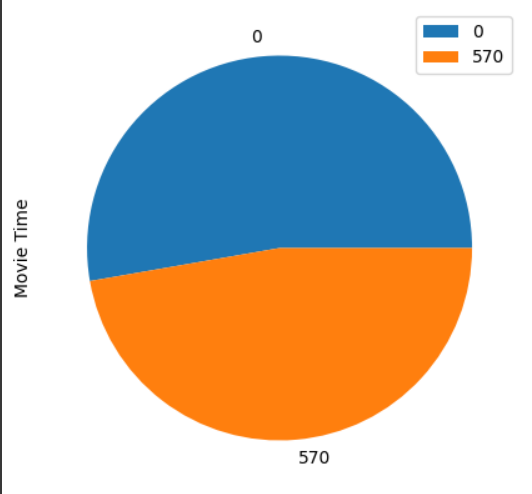


Рисунок 5 – Круговая диаграмма Movie Time

* 1. Тепловая карта

Тепловая карта — графическое представление данных, где индивидуальные значения в таблице отображаются при помощи цвета. На тепловой карте данного датасета можно выявить несколько корреляционных признаков, например, соотношение рейтинга аудитории и рейтинга IMDB и рейтинга Tomatometer.



Рисунок 6 – фрагмент тепловой карты датасета

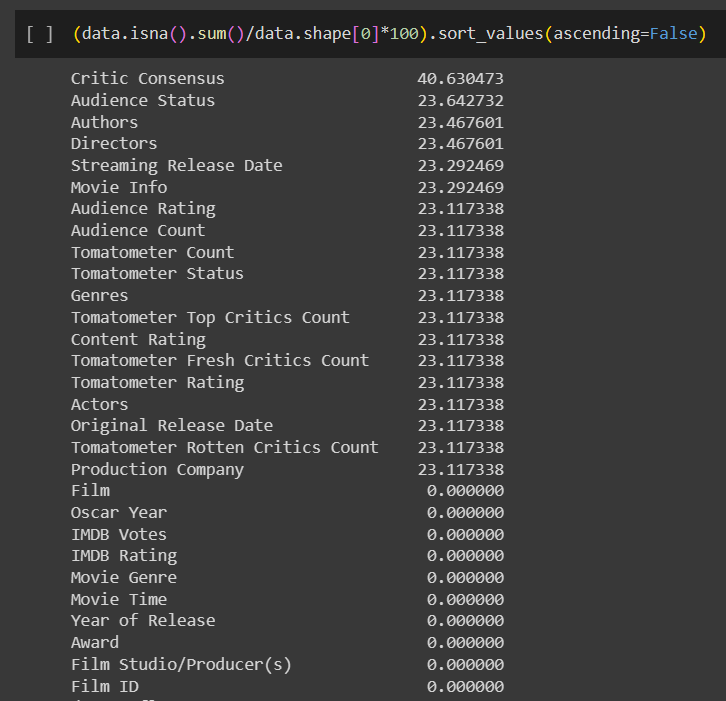
* 1. Диаграмма countplot с группировкой по нескольким номинативным признакам

CountPlot - столбчатая диаграмма, чаще всего используется для категориальных признаков в данных. Показывает, сколько строчек в датасете имеют каждое из выбранного значения категориального признака. Данная диаграмма показывает, что высоко оцененные фильма чаще получают.

# сонтплот.PNGРисунок 7 – Диаграмма countplot по столбцам Tomatometer status и Award

3Предварительная обработка данных

В датасете присутствуют пропущенные значения.



# Рисунок 8 – Проверка на наличие пропусков в таблице

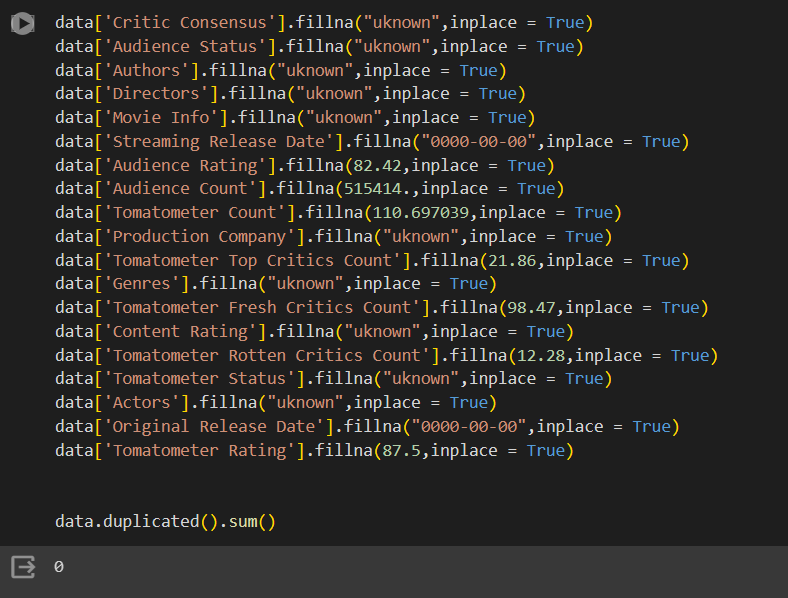
****

Рисунок 9 - Замена пропущенных значений

Также было применено one-hot кодирование, то есть преобразование категориальных переменных в численные путем создания столбцов под каждую категорию и заполнения их значениями 0 и 1 в зависимости от категории каждой строчки.

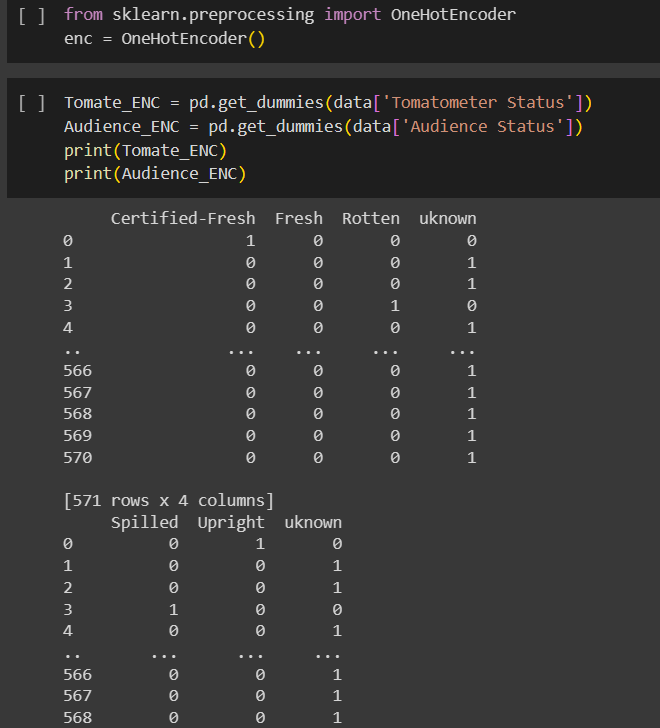
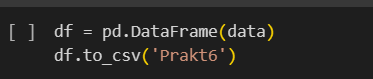


Рисунок 10 – Горячее кодирование

Предобработанные данные были сохранены в формате .csv в той же директории, что и изначальный датасет.



# Рисунок 11– Экспорт датасета

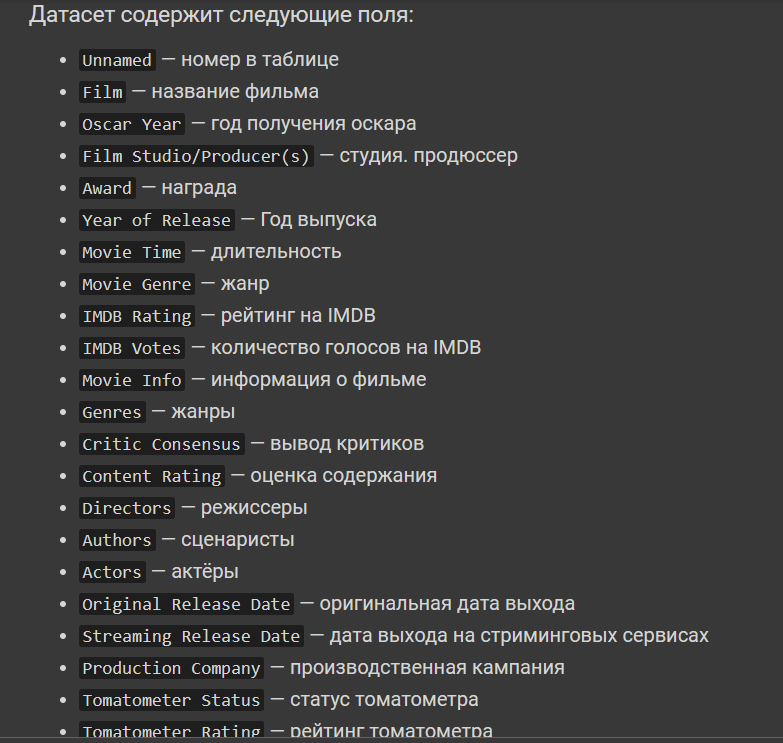


Рисунок 12 - Описание столбцов датасета(1)

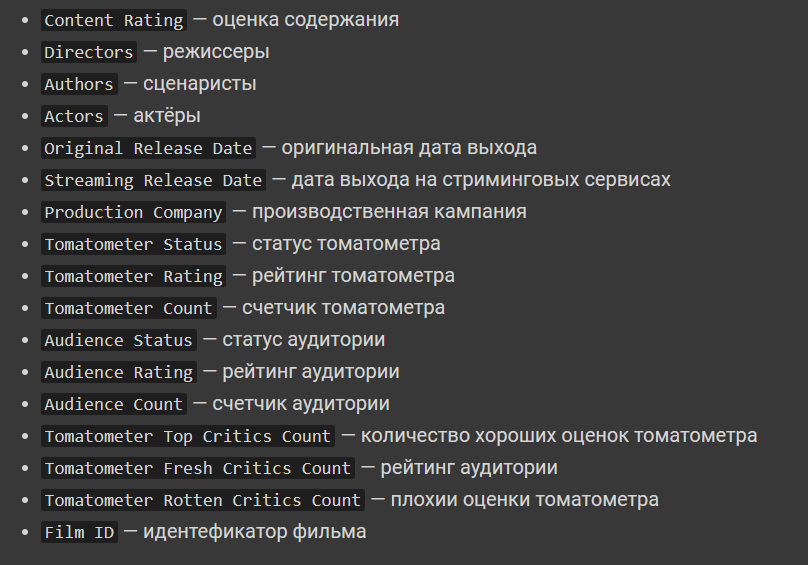


Рисунок 13 - Описание столбцов датасета(2)

ЗАКЛЮЧЕНИЕ

В ходе прохождения практики были изучены и использованы различные библиотеки Python, такие как matplotlib, seaborn, pandas и numpy. Эти библиотеки являются необходимыми инструментами для работы с данными и визуализации результатов.

Библиотека matplotlib позволяет создавать различные графики и диаграммы, которые помогают визуализировать данные и делать выводы. С помощью seaborn можно создавать более сложные графики, такие как тепловые карты и распределения. Библиотека pandas предоставляет возможность работать с данными в формате таблицы, что упрощает анализ и обработку данных.

Библиотека numpy позволяет проводить математические операции с массивами данных.

В процессе работы с этими библиотеками были решены различные задачи, связанные с анализом данных и визуализацией результатов. Были созданы графики, диаграммы, тепловые карты и распределения, которые помогли понять структуру данных и выявить закономерности.

В целом, использование этих библиотек значительно ускоряет процесс анализа данных и позволяет делать более точные выводы. Они являются необходимым инструментом для работы с данными в Python и рекомендуются к изучению всем, кто занимается анализо мданных.

СПИСОКИСПОЛЬЗОВАННЫХИСТОЧНИКОВ

* + 1. <https://numpy.org/doc/stable/reference/generated/numpy.matrix.html>(датаобращения:5.12.23).
    2. <https://seaborn.pydata.org/installing.html>(датаобращения:5.12.23).
    3. <https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html>(датаобращения:5.12.23).
    4. <https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.tight_layout.html>(датаобращения:5.12.