Университет ИТМО

Практическая работа №3

по дисциплине «Визуализация и моделирование»

Автор: Севумян Наталья Ренатовна
Поток: ВИМ 1.1
Группа: К3221
Факультет: ИКТ
Преподаватель: Чернышева А.В.

Санкт-Петербург, 2021 г.

Исследование и предобработка датасета

Таблица с описанием хранящихся в дасатете данных с описанием проблемы в данных, которая требует дополнительной предобработки и способов ее решения.

Таблица 1: Описание колонок

Название	**Тип данных**	**Проблема**	**Решение**
imdb_id	str	-	-
title	str	-	-
plot	str	Данные, которые нельзя никак анализировать	Удалить столбец
type	str	Текст неудобно использовать при построении модели	Перевод в число
rated	str	Некоторые из рейтингов записаны в разных регистрах и считаются за разные типы, есть пустые ячейки	Приведение подобных вариантов, замена пустых ячеек фразой "Unrated"
year	str	Годы имеют неудобный строковый формат, есть незаконченные сериалы	Преобразование в целочисленный тип данных; в ячейках с годами, записанными через дефис, оставляю только год выпуска
released_at	str	Даты записаны в строковом формате, неудобном для работы	Перевод в другой формат даты
added_at	str	Даты записаны в строковом формате, неудобном для работы	Перевод в другой формат даты
runtime	int	Данные записаны в строковом формате из-за слова "min"	Перевод в целочисленный формат (в минутах)
genre	str	-	-
director	str	-	-
writer	str	-	-
actors	str	-	-
language	str	-	-
country	str	-	-
awards	str	-	-
metascore	float	Слишком много пустых ячеек	Удаление столбца
imdb_rating	float	-	-
imdb_votes	int	Данные в строковом формате	Перевод в число

Исследование и предобработка данных

Изначально в датасете всего 2 столбца с числовыми данными - metascore и imdb_rating, но metascore содержал слишком много пустых ячеек, поэтому работать с этими данными

было бессмысленно. Также я посчитала ненужным столбец plot, потому что это строка не имеет никаких возможностей для анализа.

Строки с пустыми ячейкми в столбце type были удалены, а самим данным (movie/ series / episode) были присвоены числовые значения (0/1/2).

В столбце rated было много значений, которые различались только регистром букв, а также довольно много нулевых ячеек, которые были заменены на "Unrated".

Столбцы released_at и added_at были приведены к типу datetime вместо строковых.

В столбце runtime строковые данные были переведены в числовой формат, и все данные представлены в минутах.

Столбец imdb_votes был переведен в числовой формат.

Столбец year был немного видоизменен для типа контента series. Из промежутка оставлена только первая дата, так как не все сериалы еще завершены, поэтому с датой окончания будет сложно работать.

Столбцы genre, director, writer, actors, language, country и awards не были никак обработаны, потому что я не смогла найти способов это сделать. Однако удалять их тоже не самый лучший вариант, так как есть способы использования этих данных для визуализации.

Формулировка гипотез

1. Фильмы в жанре анимация более популярны, поэтому у сериалов и меньше общий рейтинг, потому как анимационных сериалов меньше.

2. Фильмы, добавленные на платформу гораздо позже их создания, имеют меньший рейтинг.

3. Контент с одним из жанров Family имеет высокие оценки.

4. Чем больше фильмов или сериалов снял определенный режиссер, тем популярнее контент.

5. Зависимость рейтинга от количества проголосовавших - экспоненциальная.