

Университет ИТМО

Практическая работа №3
по дисциплине «Визуализация и моделирование»

Автор: Ефимов Павел Леонидови
Поток: ВИМ 1.1
Группа: К3220
Факультет: ИКТ
Преподаватель: Чернышева А.В.

Санкт-Петербург, 2021 г.

Датасет: <https://www.kaggle.com/tmdb/tmdb-movie-metadata>

Датасет содержит 5 тысяч фильмов с TMDb. В датасете собраны такие данные как: бюджет, компания производитель, дата выхода, прибыль, средний балл.

Название столбца	Данные, хранящиеся в столбце	Проблема	Решение
budget	Бюджет фильма	Много нулевых значений и большой разброс	Очистка выбросов
genres	Жанры	-	-
homepage	Сайт фильма	Лишние данные	Удалить
id	Номер фильма	-	-
keywords	Ключевые слова	-	-
original_language	Язык оригинала	-	-
original_title	Оригинальное название	-	-
overview	Описание	Не обрабатываемые данные	Удалить
popularity	Популярность	Много нулевых значений и большой разброс	Очистка выбросов
production_companies	Компания производитель	-	-
production_countries	Страна производитель	-	-
release_date	Дата выхода	-	-
revenue	Прибыль	-	Очистка выбросов
runtime	Время показа на экране	-	-
spoken_languages	Язык озвучки фильма	-	-
status	Статус	-	-
tagline	Слоган	Не обрабатываемые данные	Удалить
title	Название	-	-
vote_average	Средняя оценка	Много нулевых значений и большой разброс	Очистка выбросов
vote_count	Количество голосов	Много нулевых значений и большой разброс	Очистка выбросов

1 Изменения

1.1 homepage удален

Удален homepage, т.к. является лишним данным, хранит ссылку на сайт фильма, у каждого фильма уникальное значение

1.2 overview удален

Удален overview, т.к. является не обратываемым данным, в нем хранится описание фильма, у каждого фильма уникальное значение

1.3 tagline удален

Удален tagline, т.к. является не обратываемым данным, в нем хранится слоган фильма, у каждого фильма уникальное значение

1.4 runtime очищен

Очищены строки с самыми короткими и длинными фильмами, т.к. зачастую они являются либо режиссерскими, либо любительскими

1.5 budget очищен

Очищены строки с самыми дешевыми и дорогими фильмами, для дешевых либо нет данных или любительские, а дорогие являются выбросами

1.6 revenue очищен

Очищены строки по прибыли фильма, очищены выбросы

1.7 popularity очищен

Очищены строки с выбросами по популярности фильма

1.8 vote_count очищен

Очищены строки с выбросами по слишком большому или малому количеству голосов фильма

1.9 vote_average очищен

Очищены строки с выбросами по слишком большому или малому рейтингу фильма, данных либо нет и они равны 0, либо же слишком высокий рейтинг

Ознакомиться с изменениями можно на рисунке

budget	int64	budget	int64
genres	object	genres	object
homepage	object	id	int64
id	int64	keywords	object
keywords	object	original_language	object
original_language	object	original_title	object
original_title	object	popularity	float64
overview	object	production_companies	object
popularity	float64	production_countries	object
production_companies	object	release_date	object
production_countries	object	revenue	float64
release_date	object	runtime	float64
revenue	int64	spoken_languages	object
runtime	float64	status	object
spoken_languages	object	title	object
status	object	vote_average	float64
tagline	object	vote_count	float64
title	object	dtype: object	
vote_average	float64		
vote_count	int64		
dtype: object			

```
{'budget': 0,
'genres': 0,
'id': 0,
'keywords': 0,
'original_language': 0,
'original_title': 0,
'popularity': 0,
'production_companies': 0,
'production_countries': 0,
'release_date': 0,
'revenue': 0,
'runtime': 0,
'spoken_languages': 0,
'status': 0,
'title': 0,
'vote_average': 0,
'vote_count': 0}
```

2 Гипотезы

2.1 Выявление жанра с самым высоким рейтингом в разные года

Нахождение жанра с высоким рейтингом в разные года, это позволит найти интересы людей в разные эпохи

2.2 Количество языков влияет на популярность фильма

Нахождение зависимости между количеством языков фильма на его популярность

2.3 Большее количество компаний в фильме обеспечивают больший рейтинг

Проверка зависимости между количеством компаний фильма на его рейтинг

2.4 Большой бюджет фильма обеспечит больший рейтинг

Определение зависимости между бюджетом фильма и его рейтингом, будет ли фильм с большим бюджетом сделан качественнее или же есть среднее значение, после которого фильму сложнее увеличить качество

2.5 Есть ли зависимость между рейтингом и прибылью фильма

Есть прямое отношение между прибылью фильма и его рейтингом, самые прибыльные фильмы находятся между рейтингами 6 и 8

Посмотреть на реализацию можно по ссылке: <https://colab.research.google.com>