МИНОБРНАУКИ РОССИИ

Федеральное государственное автономное образовательное   
учреждение высшего образования

«Южный федеральный университет»

Институт высоких технологий и пьезотехники

Кафедра прикладной информатики и инноватики

|  |  |
| --- | --- |
| Изображение выглядит как логотип, символ, эмблема, текст  Автоматически созданное описание |  |

Отчет по проекту

«Сбор, предобработка и анализ данных о самых популярных фильмах IMDB»

Выполнили студентки

Ковалева Наталья Владимировна

Гончарова Елизавета Сергеевна

Ростов-на-Дону – 2024

**Введение**

Анализ данных об успешных фильмах, собранных с платформы IMDB, представляет собой важное исследование в области кинематографии и аналитики. С учетом растущей популярности онлайн-рейтингов и рецензий, понимание факторов, влияющих на успех фильмов, становится ключевым для производителей и режиссеров. В данном контексте изучение процесса сбора, предобработки и анализа данных о самых популярных фильмах на IMDB является необходимым для выявления закономерностей и трендов в кинематографии.

Нами были поставлены следующие задачи:

1. Сбор данных:

- Получение информации о самых популярных фильмах на платформе IMDB.

- Извлечение данных, включая название, описание фильма, год выпуска и т.д.

2. Предобработка данных:

- Преобразование и стандартизация данных для дальнейшего анализа.

- Обработка пропущенных значений и дубликатов.

3. Анализ данных:

- Количество сборов в первом прокате, на внутреннем, международном и мировом рынке

- Самый прибыльный жанр;

- Окупаемость бюджета по кинокомпаниям

- Категория МРАА с наибольшими мировыми продажами

- Наиболее распространенная продолжительность фильма

- Кассовые фильмы по году выпуска

- Лучший месяц и день для выпуска фильма

4. Интерпретация результатов:

- Формулирование выводов на основе проведенного анализа данных.

- Подтверждение / опровержение поставленной гипотезы.

Также нами была выдвинута следующая гипотеза:

*«Мы предпологаем, что день и месяц выпуска фильма не влияют на его успех и кассовые сборы».*

**Описание Dataset**

Этот dataset содержит информацию о 1000 самых кассовых фильмах Голивуда. Актуально по состоянию на 25 сентября 2023 года.

Эти данные были получены с нескольких сайтов и объединены для выполнения различных операций с данными. Данные были взяты с IMDB, Rotten Tomatoes и многих других сайтов.

Изображение выглядит как текст, снимок экрана, программное обеспечение, число

Автоматически созданное описание

**Ход работы**

**Подготовка данных к анализу**

Имеем датасет с лучшими фильмами по мнению imdb в формате csv со следующими данными:

ID – id фильма

Title – название

Movie Info – описание фильма

Year – год выпуска

Distributor – киностудия

Budget (in $) – бюджет фильма

Domestic Opening (in $) – сборы в первом прокате

Domestic Sales (in $) – продажи на внутреннем рынке

International Sales (in $) – Международные продажи фильма

World Wide Sales (in $) – Мировые продажи фильма

Release Date – дата релиза

Genre – жанр

Running Time – продолжительность по времени

License – категории в соответствии с рейтинговой системой MPAA

Предобработка данных производилась с помощью pyspark в среде jypyter.

1. **Создание DataFrame из csv файла**

Использовали spark.read.csv(<path\_to\_file>), чтобы создать **dataframe** из файла CSV.

1. **Преобразование типов данных**

К сожалению, **Spark** может неправильно определить схему, поэтому её необходимо создать вручную.

ID – integer

Title – String

Movie\_info – String

Year – integer

Distributor – String

Budget – integer

Dom\_Opening – integer

Dom\_Sales – integer

Intern\_Sales – integer

WW\_Sales – integer

Release\_Date – date

Genre – String

Running\_Time – integer

License – String

Создали вышеприведённую схему в помощью SructType() и StructField(). Необходимые SQL типы импортировали из pyspark.sql.types

Увидели, что столбцы Release\_Date и Running\_Time имеют тип данных String. Необходимо преобразовать эти столбцы в типы Date и Integer, соответственно. Для того, чтобы последующие преобразования типов не потеряли наши данные, проверим каждый столбец на содержание значений Null. Для каждого столбца вывелb количество нулевых значений.

Теперь преобразовади столбец Running\_Time таким образом, чтоб он имел IntegerType и из “2 hr 42 min” (String) стал 162 (integer).

Преобразовали столбец Release\_Date таким образом, чтоб он имел DateType и из “16-Dec-09” (String) стал 2009-12-16 (Date)

Снова проверили столбцы на содержание нулевых значений, чтоб контролировать утечку данных

Столбец "Жанр" содержит строковые значения, поэтому преобразовали их в значения списка.

Есть несколько повторяющихся названий различных дистрибьюторов:

* DreamWorks Distribution -> DreamWorks
* Twentieth Century Fox -> 20th Century Studios
* Sony Pictures Classics -> Sony Pictures Entertainment (SPE)
* United Artists -> United Artists Releasing

Объединили их.

1. **Очистка дубликатов**

Дубликатов не оказалось

1. **Заполнение нулевых значений**

Заполнили нулевые значения в Budget и DomOpening средним значением столбцов.

С помощью sql запросов заполнили строки с нулевым значением средним значением столбцов: Budget и DomOpening

Создали временное представление DataFrame и с помощью spark.sql нашла среднее значение Budget и DomOpening

Теперь мы избавились от нулевых значений

1. **Сокращение наминала чисел стоимости**

Изменили представление чисел стоимости, выделив целую часть миллиона доллров в столбцах Budget, DomOpening, Dom\_Sales, Intern\_Sales, WW\_Sales

1. **Извлечение данных**

Сохранили файлы в формате parquet, в который по умолчанию сохраняет spark.

Теперь структура данных файла имеет следующий вид:

ID – integer - id фильма

Title – String - название

Movie\_info – String - описание фильма

Year – integer - год выпуска

Distributor – String - киностудия

Budget\_in\_mill – integer - бюджет фильма

Dom\_Opening\_in\_mill – double – сборы в первом прокате в млн $

Dom\_Sales\_in\_mill – double - продажи на внутреннем рынке в млн $

Intern\_Sales\_in\_mill – double - Международные продажи фильма в млн $

WW\_Sales\_in\_mill – double - Мировые продажи фильма в млн $

Release\_Date – date - дата релиза

Genre – array(String) - жанр

Running\_Time – integer - продолжительность по времени в минутах

License – String - категории в соответствии с рейтинговой системой MPAA

**Анализ данных**

После структурирования датасет готов к проведению анализа. Для этого использовалась программа Power BI, куда мы загрузили данные в формате parquet

Вопросы для визуализации:

1. Количество сборов в первом прокате, на внутреннем, международном и мировом рынке
2. Самый прибыльный жанр;
3. Категория МРАА с наибольшими мировыми продажами
4. Наиболее распространенная продолжительность фильма
5. Кассовые фильмы по году выпуска
6. Лучший месяц и день для выпуска фильма
7. Окупаемость бюджета по кинокомпаниям

Общий вид файла parquet в Power BI выглядит следующим образом:

Изображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описание

## **Количество сборов в первом прокате, на внутреннем, международном и мировом рынке**

Изображение выглядит как текст, снимок экрана, линия, диаграмма

Автоматически созданное описание

На диаграммах представлены данные о сборах фильмов в различных прокатных категориях. Каждая диаграмма показывает зависимость суммы сборов от ID фильмов.

Можно заметить, что:

- Сборы фильмов сильно варьируются, но наблюдается значительная концентрация фильмов с низкими сборами.

- Мировые сборы обычно выше, чем внутренние или международные сборы, что указывает на важность глобального рынка для индустрии кино.

## **Самый прибыльный жанр**

Изображение выглядит как текст, линия, снимок экрана, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, диаграмма, линия

Автоматически созданное описание

На диаграмме представлены данные о прибыльности различных жанров фильмов. Она показывает зависимость суммы мировых сборов (суммы WW Sales in mill) и количества фильмов (количество ID) от жанра.

Мы видим, что даже если у жанра меньше фильмов (меньше ID), он все равно может быть более прибыльным, что говорит о высоком качестве фильмов или их больших успехах на рынке.

## **Окупаемость бюджета по кинокомпаниям**

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описание

Диаграмма представляет данные о сумме бюджетов и сумме мировых продаж (в миллионах долларов) по основным кинокомпаниям. По вертикальной оси расположены названия кинокомпаний, а по горизонтальной оси — суммы бюджетов (в светло-синем цвете) и суммы мировых продаж (в темно-синем цвете), выраженные в миллионах долларов.

**4. Категория МРАА с наибольшими мировыми продажами**

**Изображение выглядит как текст, снимок экрана, Шрифт, диаграмма

Автоматически созданное описание**

Диаграмма иллюстрирует категории возрастных рейтингов MPAA с наибольшими мировыми продажами.

На этой диаграмме мы можем увидеть, что категория PG-13 занимает наибольшую долю мировых продаж среди всех категорий MPAA, что составляет более половины всех продаж (53,... %). Это указывает на то, что фильмы с рейтингом PG-13 наиболее популярны и востребованы на мировом кинорынке.

## **5. Наиболее распространенная продолжительность фильма**

Изображение выглядит как текст, снимок экрана, Шрифт, круг

Автоматически созданное описание

Диаграмма представляет данные о наиболее распространенной продолжительности фильмов, сгруппированных по интервалам времени. Каждый сектор кольцевой диаграммы соответствует различным интервалам продолжительности фильмов. Мы убедились, что наиболее распространенные категории:

- Фильмы с продолжительностью от 110 до 119 минут составляют наибольшую долю (18,4%), показывая, что это предпочтительное время для большинства современных фильмов.

## **6. Кассовые фильмы по году выпуска**

Изображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описание

Диаграмма представляет данные о количестве кассовых фильмов по годам их выпуска. По оси "X" (горизонтальная ось) показаны годы, начиная примерно с 1940 года и до 2020-х годов. По оси "Y" (вертикальная ось) представлено количество ID фильмов (что может трактоваться как количество кассовых фильмов).

Диаграмма иллюстрирует значительный рост числа кассовых фильмов с течением времени, с особенно резким увеличением после 1980-х годов и пик на протяжении 2000-х и 2010-х годов. Этот рост можно объяснить различными факторами, такими как технологические инновации, расширение рынка кино и изменение потребительских предпочтений.

## **Лучший месяц и день для выпуска фильма**

Изображение выглядит как текст, снимок экрана, Красочность, Прямоугольник

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Красочность, Прямоугольник

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Красочность, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Параллельный, число

Автоматически созданное описание

На диаграммах изображены мировые сборы по дням:

- Объемы сборов по дням распределены и отображены в различных цветах для выделения каждого отдельного дня.

- Наиболее прибыльные дни показывают наибольшие размеры блоков.

Лучший месяц для выпуска фильма:

- Представлены мировые сборы по месяцам, что помогает определить, в какой месяц лучше выпускать фильм.

- Июнь, декабрь и май являются наиболее успешными месяцами с точки зрения сборов, тогда как январь и февраль менее прибыльны.

Самые кассовые фильмы в мире:

- На диаграммах представлены данные о самых кассовых фильмах в мире по дням, разбитые по различным категориям (отображены разными цветами).

- Фильмы представлены с указанием их суммарных мировых сборов в миллионах долларов.

- Основные фильмы, такие как "Avatar", "Avengers: Endgame", "The Lion King" и другие, занимают верхние позиции с максимальными кассовыми сборами.

**Вывод**

Изучение данных о самых популярных фильмах на IMDB позволяет опровергнуть гипотезу о том, что день и месяц выпуска фильма не влияют на его успех и кассовые сборы. Дети, уходящие на каникулы, играют значительную роль в успешности фильма и влияют на его кассовые показатели.

1. Время каникул:

- Дни и месяцы выпуска фильма могут совпадать с периодами детских каникул, что существенно влияет на количество зрителей и кассовые сборы.

2. Спрос во время отпусков:

- В периоды каникул растет спрос на кинопоказы, и дети становятся активными потребителями развлекательной продукции, что повышает вероятность успеха фильма.

Таким образом, дни и месяцы выпуска фильма не следует недооценивать, поскольку они могут оказывать существенное влияние на успех проекта и его финансовые результаты, особенно в контексте детей, уходящих на каникулы.