

МИНОБРНАУКИ РОССИИ
Федеральное государственное автономное образовательное
учреждение высшего образования
«Южный федеральный университет»
Институт высоких технологий и пьезотехники
Кафедра прикладной информатики и инноватики



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.


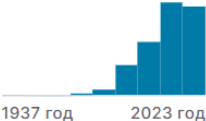
Также нами была выдвинута следующая гипотеза:

«Мы предполагаем, 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 Голливуда. 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.

# интервал: идентификатор	Δ Заголовок Строка: Название фильма.	Δ Информация о ф... Строка: Краткое введение в фильмы	# Год Строка: Имя дистрибьютора.	Δ Распределитель Строка: Исходная дата выпуска.	Δ Би int: E прог
 0 999	988 уникальные ценности	999 уникальные ценности	 1937 год 2023 год	Warner Bros. 17% Кинофильмы сту... 16% Другое (675) 68%	1500 1000 Другие
0	Аватар	Морской пехотинец, страдающий параличом нижних конечностей, отправленный на Луну Пандору с уникальной миссией, разрывается между преследованием...	2009	Двадцатый век Фокс	2376
1	Мстители: Финал	После разрушительных событий «Мстителей: Война бесконечности» вселенная лежит в руинах. С помощью р...	2019 год	Кинофильмы студии Уолта Диснея	3566
2	Аватар: Путь воды	Джейк Салли живет со своей новой семьей, образованной на	2022 год	Студии 20 века	14 4 (EME

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 МРАА

Предобработка данных производилась с помощью pyspark в среде jupyter.

1. Создание DataFrame из csv файла

Использовали `spark.read.csv(<path_to_file>)`, чтобы создать **dataframe** из файла CSV.

2. 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 StructType() 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 вывелб количество нулевых значений.

Теперь преобразовали 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.

3. Очистка дубликатов

Дубликатов не оказалось

4. 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

5. Сокращение номинала чисел стоимости

Изменили представление чисел стоимости, выделив целую часть миллиона долларов в столбцах Budget, DomOpening, Dom_Sales, Intern_Sales, WW_Sales

6. 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 МРАА

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

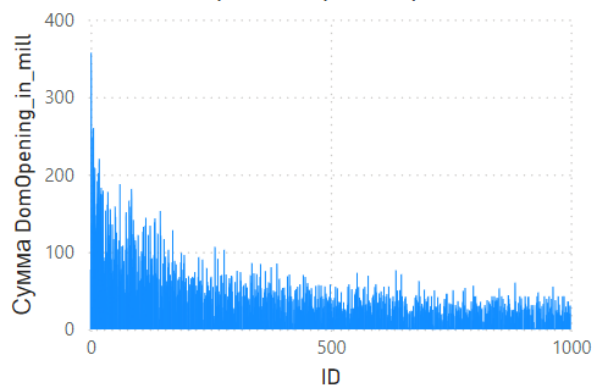
Общий вид файла parquet в Power BI выглядит следующим образом:

Средства работы со столбцами														
Имя	ID	Тип данных	Целое число	Целое число	\$	%	0	Σ	Количество	Без категорий	Без категорий	Без категорий	Без категорий	Без категорий
ID	Title	Movie_info	Year	Distributor	Budget	Dom	Dom	Intern	WW_Sal	Release_Date	Genre	Rumi	Licor	Данные
0	Avatar	A paraplegic Marine dispatched to the moon Pandora on a unique mission becom	2009	20th Century Studios	237	77,03	785,2216	2138,4843	2923,70602	16 декабря 2009 г.	Action, Adventure, Fantasy	162	PG-13	
1	Avengers: Endgame	After the devastating events of Avengers: Infinity War, the universe is in ruins. Wit	2019	Walt Disney Studios Mc	356	357,12	658,373	1941,0661	2799,4391	24 апреля 2019 г.	Action, Adventure, Drama	181	PG-13	
2	Avatar: The Way of Water	Jake Sully lives with his newfound family formed on the extrasolar moon Pandora	2022	20th Century Studios	97,423917	134,10	694,0757	1636,1745	2320,25028	24 апреля 2019 г.	Action, Adventure, Drama	181	PG-13	
3	Titanic	A seventeen-year-old aristocrat falls in love with a kind but poor artist aboard the	1997	Paramount Pictures	200	28,64	674,2926	1590,4506	2264,74330	19 декабря 2009 г.	Drama, Romance	194	PG-13	
4	Star Wars: Episode VII - The Force Awakens	As a new threat to the galaxy rises, Rey, a desert scavenger, and Finn, an ex-storm	2015	Walt Disney Studios Mc	245	247,97	926,6622	1134,6479	2071,31021	16 декабря 2015 г.	Action, Adventure, Sci-Fi	138	PG-13	
5	Avengers: Infinity War	The Avengers and their allies must be willing to sacrifice all in an attempt to defe	2018	Walt Disney Studios Mc	97,423917	257,70	678,8154	1373,5995	2052,41503	16 декабря 2015 г.	Action, Adventure, Sci-Fi	138	PG-13	
6	Spider-Man: No Way Home	With Spider-Man's identity now revealed, Peter asks Doctor Strange for help. Whi	2021	Sony Pictures Entertaini	97,423917	280,14	814,1150	1107,7320	1921,84711	16 декабря 2015 г.	Action, Adventure, Sci-Fi	138	PG-13	
7	Jurassic World	A new theme park, built on the original site of Jurassic Park, creates a genetically	2015	Universal Pictures	150	208,81	653,4066	1018,1308	1671,53744	10 июня 2015 г.	Action, Adventure, Sci-Fi	124	PG-13	
8	The Lion King	After the murder of his father, a young lion prince flees his kingdom only to learn	2019	Walt Disney Studios Mc	260	191,77	543,6380	1119,4373	1663,07540	11 июля 2019 г.	Adventure, Drama, Family	118	PG	
9	The Avengers	Earth's mightiest heroes must come together and learn to fight as a team if they	2012	Walt Disney Studios Mc	220	207,44	623,3579	897,18062	1520,53853	25 апреля 2012 г.	Action, Sci-Fi	143	PG-13	
10	Furious 7	Deckard Shaw seeks revenge against Dominic Toretto and his family for his coma	2015	Universal Pictures	190	147,19	353,0070	1162,3343	1515,34139	1 апреля 2015 г.	Action, Crime, Thriller	137	PG-13	
11	Top Gun: Maverick	After thirty years, Maverick is still pushing the envelope as a top naval aviator, bu	2022	Paramount Pictures	97,423917	126,71	718,7328	776,96347	1495,69629	1 апреля 2015 г.	Action, Crime, Thriller	137	PG-13	
12	Frozen II	Anna, Elsa, Kristoff, Olaf and Sven leave Arendelle to travel to an ancient, autumn	2019	Walt Disney Studios Mc	150	130,26	477,3735	976,30989	1453,68347	20 ноября 2019 г.	Adventure, Animation, Com	103	PG	
13	Barbie	Barbie suffers a crisis that leads her to question her world and her existence.	2023	Warner Bros.	97,423917	162,02	630,4500	797	1427,45008	20 ноября 2019 г.	Adventure, Animation, Com	103	PG	
14	Avengers: Age of Ultron	When Tony Stark and Bruce Banner try to jump-start a dormant peacekeeping pr	2015	Walt Disney Studios Mc	250	191,27	459,0058	946,01218	1405,01804	22 апреля 2015 г.	Action, Adventure, Sci-Fi	141	PG-13	
15	The Super Mario Bros. M	A plumber named Mario travels through an underground labyrinth with his broth	2023	Universal Pictures	97,423917	146,36	574,9343	785,82972	1360,76405	22 апреля 2015 г.	Action, Adventure, Sci-Fi	141	PG-13	
16	Black Panther	T'Challa, heir to the hidden but advanced kingdom of Wakanda, must step forward	2018	Walt Disney Studios Mc	202,00	700,4265	649,49951	1349,92608	1405,01804	22 апреля 2015 г.	Action, Adventure, Sci-Fi	141	PG-13	
17	Harry Potter and the De	Harry, Ron, and Hermione search for Voldemort's remaining Horcruxes in their eff	2011	Warner Bros.	97,423917	169,19	381,4475	960,91235	1342,35994	22 апреля 2015 г.	Action, Adventure, Sci-Fi	141	PG-13	
18	Star Wars: Episode VIII - The Last Jedi	The Star Wars saga continues as new heroes and galactic legends go on an epic a	2017	Walt Disney Studios Mc	317	220,01	620,1813	714,22632	1334,40770	13 декабря 2017 г.	Action, Adventure, Fantasy	152	PG-13	
19	Jurassic World: Fallen Ki	When the island's dormant volcano begins roaring to life, Owen and Claire mount	2018	Universal Pictures	170	148,02	417,7197	892,74653	1310,46629	6 июня 2018 г.	Action, Adventure, Sci-Fi	128	PG-13	
20	Frozen	When the newly crowned Queen Elsa accidentally uses her power to turn things int	2013	Walt Disney Studios Mc	150	0,24	400,9530	883,58750	1284,54051	22 ноября 2013 г.	Adventure, Animation, Com	102	PG	
21	Beauty and the Beast	A selfish Prince is cursed to become a monster for the rest of his life, unless he le	2017	Walt Disney Studios Mc	160	174,75	504,4811	761,63479	1266,11596	16 марта 2017 г.	Adventure, Family, Fantasy	129	PG	
22	Incredibles 2	The Incredibles family takes on a new mission which involves a change in family r	2018	Walt Disney Studios Mc	97,423917	182,69	608,5817	634,64392	1243,22566	16 марта 2017 г.	Adventure, Family, Fantasy	129	PG	
23	The Fate of the Furious	When a mysterious woman seduces Dominic Toretto into the world of terrorism a	2017	Universal Pictures	250	98,79	226,0083	1009,9967	1236,00511	12 апреля 2017 г.	Action, Crime, Thriller	136	PG-13	
24	Iron Man 3	When Tony Stark's world is torn apart by a formidable terrorist called the Mandar	2013	Walt Disney Studios Mc	200	174,14	409,0139	806,56321	1215,57720	24 апреля 2013 г.	Action, Adventure, Sci-Fi	130	PG-13	
25	Minions	Minions Stuart, Kevin, and Bob are recruited by Scarlet Overkill, a supervillain wh	2015	Universal Pictures	74	115,72	336,0457	823,39889	1159,44466	9 апреля 2015 г.	Adventure, Animation, Com	91	PG	
26	Captain America: Civil W	Political involvement in the Avengers' affairs causes a rift between Captain Ame	2016	Walt Disney Studios Mc	250	179,14	408,0843	746,96206	1155,04641	27 апреля 2016 г.	Action, Sci-Fi	147	PG-13	
27	Aquaman	Arthur Curry, the human-born heir to the underwater kingdom of Atlantis, goes o	2018	Warner Bros.	97,423917	67,87	335,1043	813,42407	1148,52839	27 апреля 2016 г.	Action, Sci-Fi	147	PG-13	
28	The Lord of the Rings: T	Gandalf and Aragorn lead the World of Men against Sauron's army to draw his g	2003	New Line Cinema	94	72,63	379,4272	768,20654	1147,63383	17 декабря 2003 г.	Action, Adventure, Drama	201	PG-13	
29	Starfall	James Bond's loyalty to M is tested when her past comes back to haunt her. Whe	2012	Sony Pictures Entertaini	200	88,36	304,3602	838,11101	1147,47120	25 октября 2012 г.	Action, Adventure, Thriller	143	PG-13	

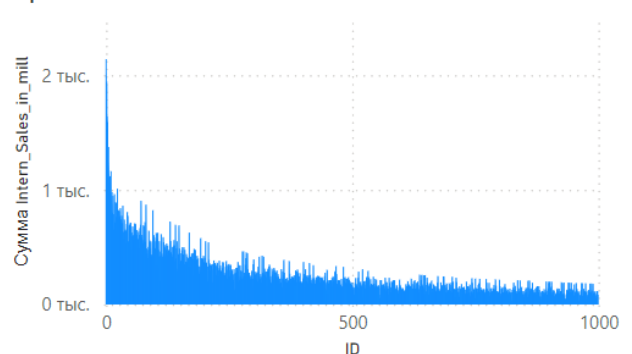
Таблица: Update_Movies (срок: 1 000) Столбец: ID (неповторяющихся значений: 1 000)

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

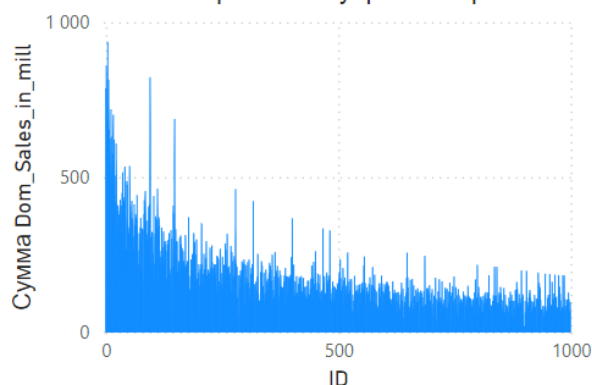
Количество сборов в первом прокате



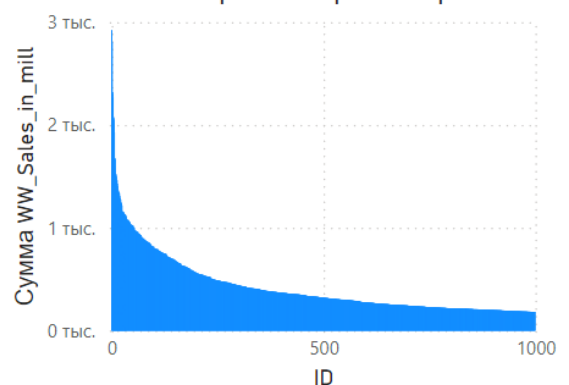
Количество сборов в Международном прокате



Количество сборов на внутреннем рынке



Количество сборов в Мировом прокате

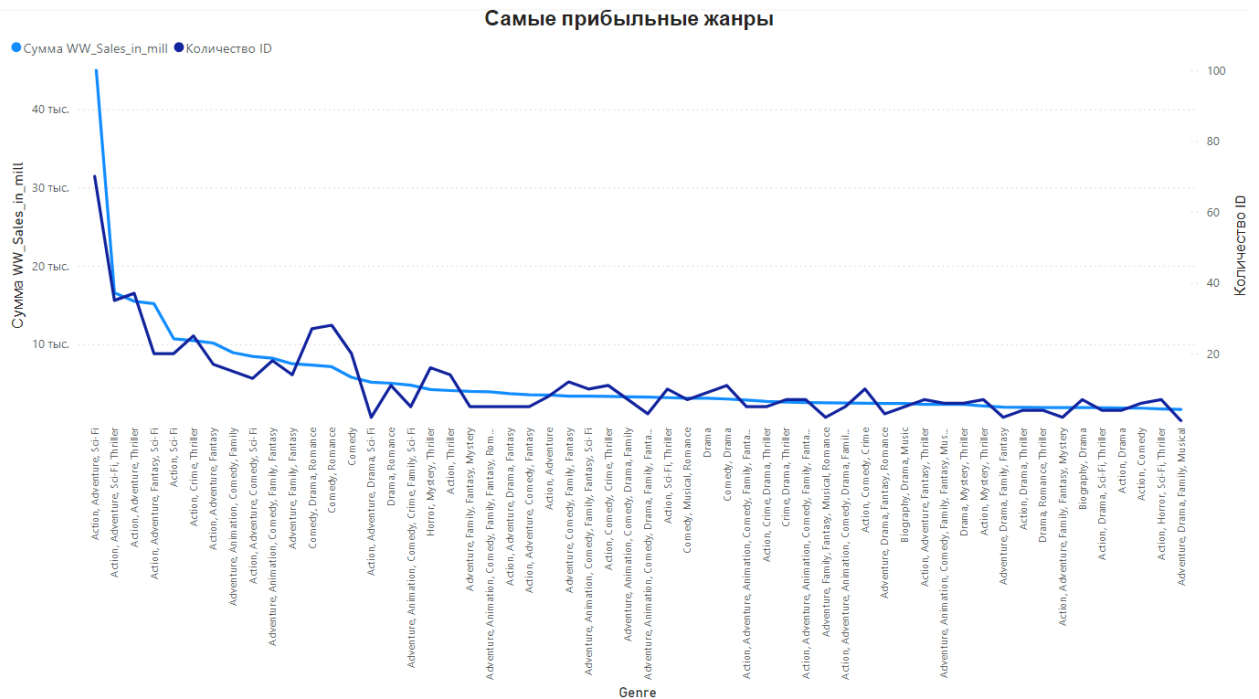


На диаграммах представлены данные о сборах фильмов в различных прокатных категориях. Каждая диаграмма показывает зависимость суммы сборов от ID фильмов.

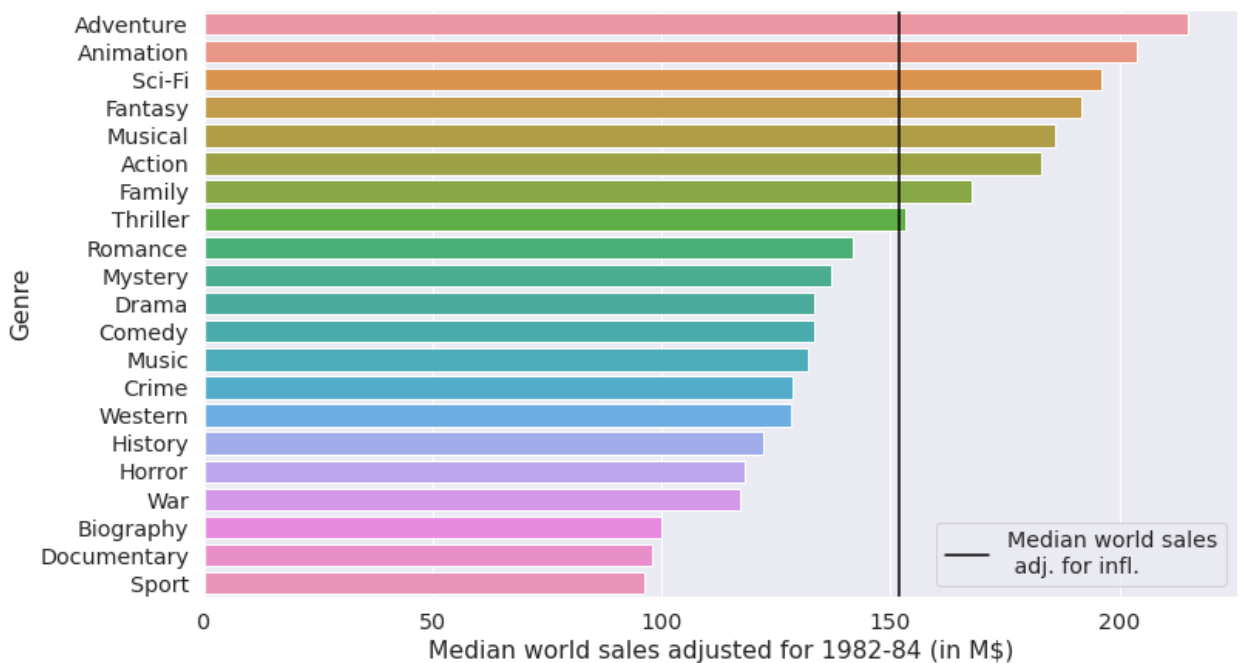
Можно заметить, что:

- Сборы фильмов сильно варьируются, но наблюдается значительная концентрация фильмов с низкими сборами.
- Лишь небольшое количество фильмов способны собрать значительно больше, чем основная масса.

2. Самый прибыльный жанр



Median world sales adjusted for 1982-84 according to each genre



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.

3. Окупаемость бюджета по кинокомпаниям

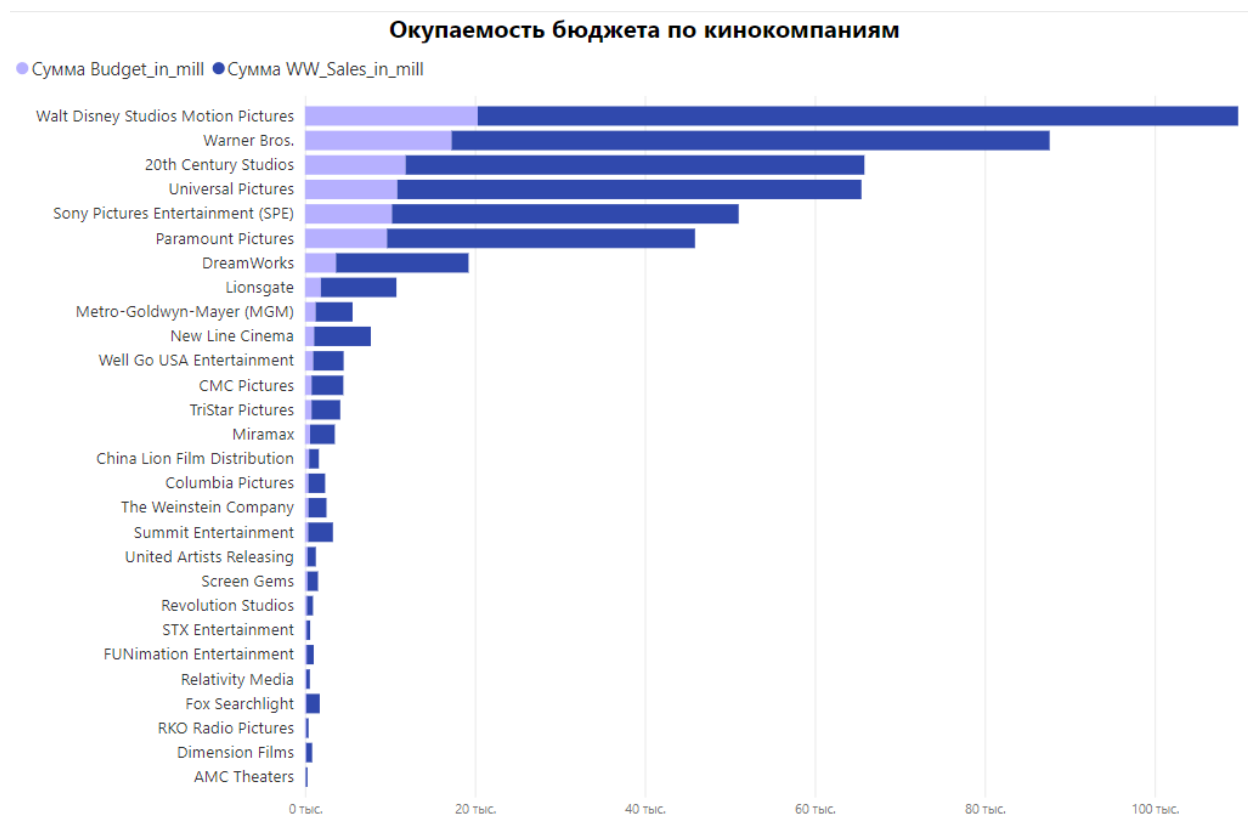


Диаграмма представляет данные о сумме бюджетов и сумме мировых продаж (в миллионах долларов) по основным кинокомпаниям. По вертикальной оси расположены названия кинокомпаний, а по горизонтальной оси — суммы бюджетов (в светло-синем цвете) и суммы мировых продаж (в темно-синем цвете), выраженные в миллионах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

4. Категория МРАА с наибольшими мировыми продажами

Категория МРАА с наибольшими мировыми продажами

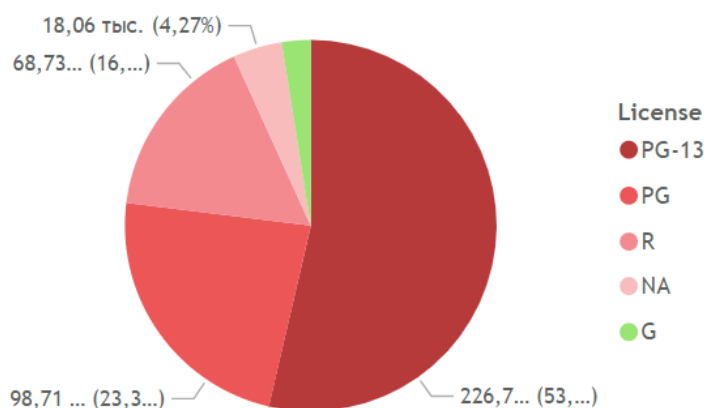
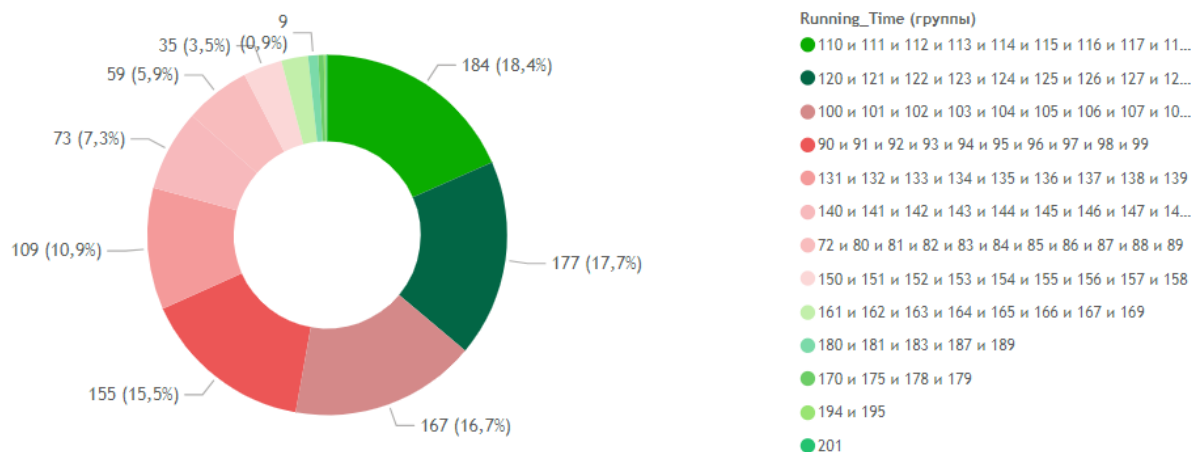


Диаграмма иллюстрирует категории возрастных рейтингов МРАА с наибольшими мировыми продажами.

На этой диаграмме мы можем увидеть, что категория PG-13 занимает наибольшую долю мировых продаж среди всех категорий МРАА, 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

Наиболее распространенная продолжительность фильма



- 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

Кассовые фильмы по году выпуска

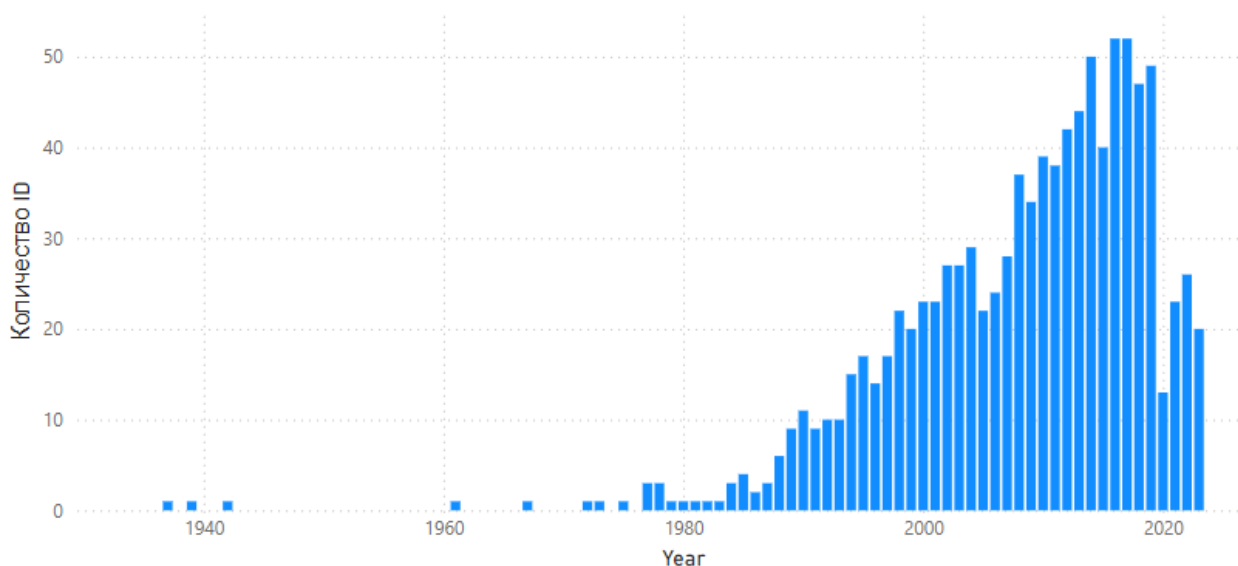


Диаграмма представляет данные о количестве кассовых фильмов по годам их выпуска. По оси "X" (горизонтальная ось) показаны годы, начиная примерно с 1940 года и до 2020-х годов. По оси "Y" (вертикальная ось) представлено количество ID фильмов (что может трактоваться как количество кассовых фильмов).

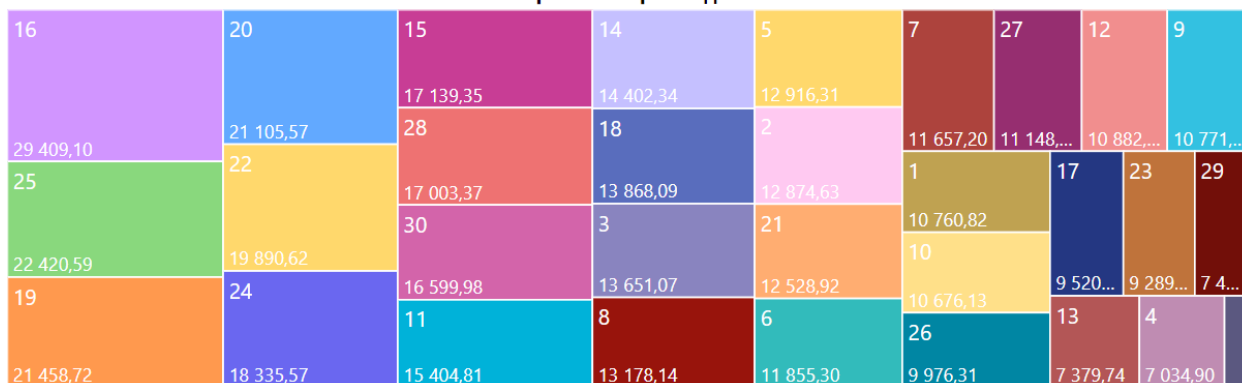
Диаграмма иллюстрирует значительный рост числа кассовых фильмов с течением времени, с особенно резким увеличением после 1980-х годов и пик на протяжении 2000-х и 2010-х годов. Этот рост можно объяснить различными факторами, такими как технологические инновации, расширение рынка кино и изменение потребительских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

7. Лучший месяц и день для выпуска фильма

Лучший месяц для выпуска фильма



Мировые сборы по дням



Title	Сумма WW_Sales_in_mill	День
Avatar	2 923,71	16
Star Wars: Episode VII - The Force Awakens	2 071,31	16
Avengers: Infinity War	2 052,42	16
Spider-Man: No Way Home	1 921,85	16
Всего	29 409,10	

Title	Сумма WW_Sales_in_mill	День
The Avengers	1 520,54	25
Skyfall	1 142,47	25
Transformers: Age of Extinction	1 104,05	25
Guardians of the Galaxy Vol. 2	863,76	25
Всего	22 420,59	

Title	Сумма WW_Sales_in_mill	День
Titanic	2 264,74	19
The Dark Knight Rises	1 081,17	19
Star Wars: Episode I - The Phantom Menace	1 027,08	19
Zootopia	1 025,52	19
Всего	21 458,72	

Title	Сумма WW_Sales_in_mill	День
Frozen II	1 453,68	20
Barbie	1 427,45	20
Toy Story 4	1 073,84	20
Jumanji: Welcome to the Jungle	995,34	20
Harry Potter and the Deathly Hallows: Part 1	977,07	20
Всего	21 105,57	

Title	Сумма WW_Sales_in_mill	День
Avengers: Age of Ultron	1 405,02	22
The Super Mario Bros. Movie	1 360,76	22
Black Panther	1 349,93	22
Harry Potter and the Deathly Hallows: Part 2	1 342,36	22
Всего	19 890,62	

Title	Сумма WW_Sales_in_mill	День
Avatar	2 923,71	16
Star Wars: Episode VII - The Force Awakens	2 071,31	16
Avengers: Infinity War	2 052,42	16
Spider-Man: No Way Home	1 921,85	16
Всего	29 409,10	

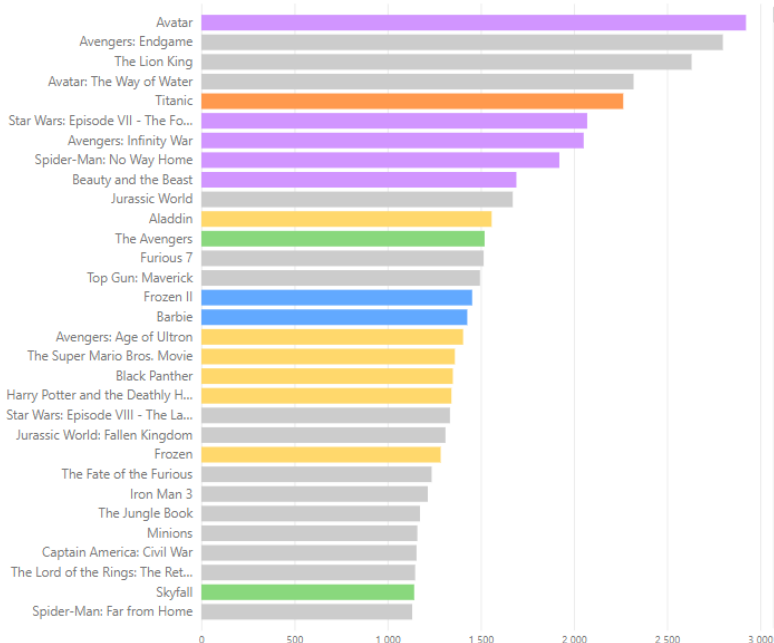
Title	Сумма WW_Sales_in_mill	День
The Avengers	1 520,54	25
Skyfall	1 142,47	25
Transformers: Age of Extinction	1 104,05	25
Guardians of the Galaxy Vol. 2	863,76	25
Всего	22 420,59	

Title	Сумма WW_Sales_in_mill	День
Titanic	2 264,74	19
The Dark Knight Rises	1 081,17	19
Star Wars: Episode I - The Phantom Menace	1 027,08	19
Zootopia	1 025,52	19
Всего	21 458,72	

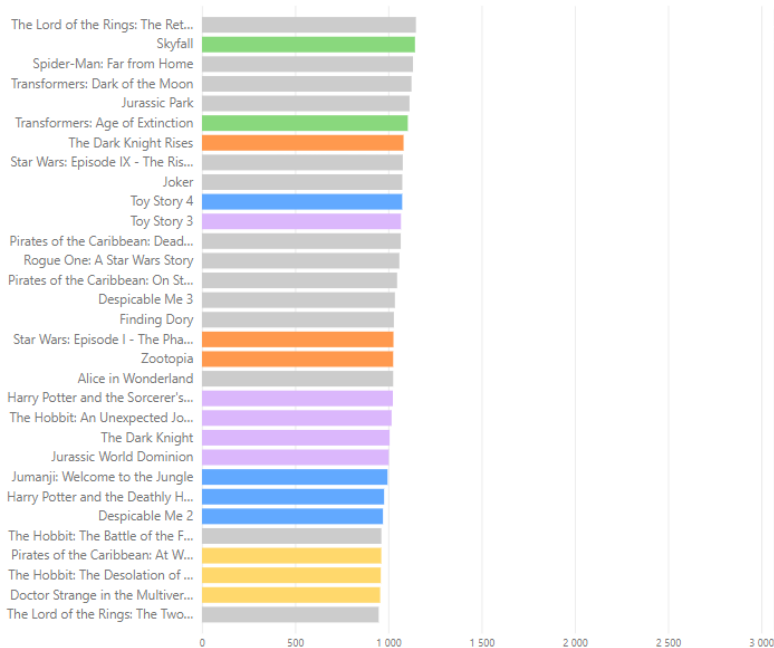
Title	Сумма WW_Sales_in_mill	День
Frozen II	1 453,68	20
Barbie	1 427,45	20
Toy Story 4	1 073,84	20
Jumanji: Welcome to the Jungle	995,34	20
Harry Potter and the Deathly Hallows: Part 1	977,07	20
Всего	21 105,57	

Title	Сумма WW_Sales_in_mill	День
Avengers: Age of Ultron	1 405,02	22
The Super Mario Bros. Movie	1 360,76	22
Black Panther	1 349,93	22
Harry Potter and the Deathly Hallows: Part 2	1 342,36	22
Всего	19 890,62	

Самые кассовые фильмы в Мире



Самые кассовые фильмы в Мире



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.