Федеральное государственное автономное образовательное учреждение высшего образования

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет экономических наук, образовательная программа «Экономика»

Домашняя работа по курсу «Эконометрика — 1»

Исследование зависимости выручки кинокартин от различных факторов

Выполнили:

Уначева Амина Аслангериевна, БЭК217

Кудабаева Гульшат Вильдановна, БЭК217

Нежурина Екатерина Владиславовна, БЭК217

Преподаватель:

Бывальцева-Станкевич Анастасия Александровна

Москва 2023

1. Введение

В современной кинематографической индустрии, характеризующейся высокой степенью конкуренции и значительными финансовыми вложениями, вопрос о прогнозировании кассового успеха фильма остается одним из наиболее актуальных и стратегически важных. Каждый выпущенный фильм уникален, и его успех складывается из множества факторов, таких как бюджет картины, оценки критиков, наличие звезд первой величины, и другие. Исследование зависимости между этими переменными и финансовыми результатами кинокартины не только предоставляет ценные практические наработки для киностудий, но и является важным шагом в направлении оптимизации стратегий маркетинга, а также понимания предпочтений и ожиданий зрителей. В данном исследовании мы используем методы множественной регрессии, с целью разобраться в факторах успешности фильмов и возможности их более точного прогнозирования. Кинематография, как сфера с высоким уровнем неопределенности и рисков, привлекает внимание исследователей, стремящихся выявить ключевые факторы, влияющие на коммерческий успех фильмов.

2. Гипотезы и переменные

2.1. Формулировка гипотез

Кинематограф захватили блокбастеры, с бюджетами, превышающими сотни миллионов долларов, и из-за стремительного роста затрат на маркетинговые кампании, связанные с продвижением, таким фильмам часто требуется собрать больший процент от своего бюджета, чтобы полностью окупиться и принести достаточную прибыль. В статье *Predicting Movie Box Office Success using Multiple Regression and SVM,* посвященной прогнозированию кассового успеха фильмов с помощью применения множественной линейной регрессии и классификации методом опорных векторов (SVM), авторы исследуют данную тему и фокусируются на анализе таких переменных, как Opening Date, Movie Name, Budget, Domestic Gross, International Gross, Total Gross, Trailer Views, Studio, Cast and Crew, Genre, Medium (Live action or Film), Trailer Views, Wikipedia Views, Rotten Tomatoes Score. Также, авторы классифицируют фильмы по 3 категориям:

1) Малобюджетные фильмы (бюджет <50 миллионов долларов)
2) Фильмы со средним бюджетом (бюджет от 50 до 150 миллионов долларов)
3) Высокобюджетные фильмы (бюджет >150 миллионов долларов)

и исследуют при каких значениях коэффициента Return on Investments (ROI) фильмы могут считаться успешными. Каждая из этих категорий имеет свой множитель, который считается успешным: x2 для малобюджетных фильмов, x2.5 для среднебюджетных фильмов и x3 для высокобюджетных фильмов.

Затраты на производство кинокартины - не единственный значимый фактор, влияющий на кассовые сборы кинокартин, популярно мнение, что успех фильма напрямую зависит от уровня престижа актерского состава кинокартины. В нашем исследовании мы

также решили проверить эту гипотезу и обратились к статье *Factors Affecting the Financial Success of Motion Pictures: What is the Role of Star Power?* Основная цель статьи – выяснить, имеют ли «кинозвезды» возможность влиять на кассовые сборы. В качестве зависимой переменной авторы выделяют выручку (revenue), а основными объясняющими переменными являются Star Power 1 - определяется суммой выдающихся наград актерского состава и Star Power 2 - определяется средним значением выручки фильмов, в которых принимал участие актер.

Опираясь на упомянутые научные статьи, были **сформулированы следующие гипотезы:**

1. У высокобюджетных кинокартин отдача от вложений выше, чем у среднебюджетных кинокартин.
2. Участие в съемках фильмов кинозвезд, обладающих наградами “Оскар” и/или “Золотой Глобус”, повышает шансы кинокартины стать более успешной.

2.2. Определение переменных

Проанализировав ранее упомянутые научные статьи, мы решили включить в модель следующие независимые переменные:

1) Рейтинг кинокартины на Imdb (rating, интервал от 0 до 10) - чем выше рейтинг фильма, тем больше потенциальной аудитория посмотрит его в кинотеатрах, полагаясь на мнение критиков и зрителей, что положительно скажется на кассовых сборах фильма.
2) Количество голосов, определивших рейтинг (votes) - чем большее количество зрителей оставит свой отзыв на кинокартину, тем больше потенциальной аудитория посмотрит его в кинотеатрах.
3) Выручка кинокартины в первый уикенд (opening_weekend, USD) - чем выше кинокартина стартанет в первый уикенд, тем более высокую кассу она соберет.
4) Выручка кинокартины в домашнем прокате (domestic_gross, USD) - чем выше кинокартина заработает в домашнем прокате, тем выше она соберет в мировом прокате.
5) Бюджет кинокартины (production_cost, USD) - чем выше затраты на создание кинокартины, тем качественнее получится фильм, и тем большее количество потенциальных зрителей захочет его посмотреть в кинотеатрах.
6) Длительность кинокартины (duration, min) - чем длиннее фильм, тем меньшее количество потенциальных зрителей захочет посмотреть его в кинотеатрах, что отрицательно скажется на выручке.
7) Рейтинговая система MPAA (certificate, дамми): G — Нет возрастных ограничений, PG — Рекомендуется присутствие родителей, PG-13 — Детям до 13 лет просмотр не желателен, R — Лицам до 17 лет обязательно присутствие взрослого. TV-MA - предназначено для взрослой аудитории. Предполагаем, что кинокартины с рейтингом G, PG, PG-13 собирают большую выручку в мировом прокате.
8) Жанр кинокартины (genre, дамми) - предполагаем, что фильмы ужасов собирают меньшую кассу, чем комедии и фэнтези.

9) Наличие у члена актерского состава Премии Американской академии кинематографических искусств и наук (oscar, дамми) - актеры, получившие “Оскар” повышают престиж фильма, что привлекает больше зрителей и положительно сказывается на выручке кинокартины.
10) Наличие у члена актерского состава Премии Голливудской ассоциации иностранной прессы (golden globe, дамми) - актеры, получившие “Золотой Глобус” повышают престиж фильма, что привлекает больше зрителей и положительно сказывается на выручке кинокартины.

3. Данные и методы

Исследуемый датасет составлен на основе информации из нескольких источников. Основные данные взяты из датасета *Netflix popular movies dataset*, который включает в себя данные до 2022 года о таких переменных, как movies title, cast of the movie, duration of the movie (in minutes), rating on IMDB, voted by people, year, genre, certificate. Данные о недостающих переменных мы взяли из другого датасета *The Most Expensive Film Productions*, который содержит данные о бюджете и выручке пятисот самых высокобюджетных кинокартинах, и включает следующие переменные: production_cost (in USD), domestic_gross (in USD), worldwide_gross (in USD), opening_weekend (in USD). Совместив два датасета, выявили 407 совпадений.

	rating	votes	production cost	domestic gross	worldwide gross	opening weekend
count	407	407	407	407	471.0	459.0
mean	6.741	355067.59	149687685.77	165676357.34	465178195.32	53630876.23
std	0.831	347660.11	47110656.25	135398713.03	379321463.4	44783700.05
min	4.9	10400.0	91000000.0	0.0	0.0	116616.0
25%	6.2	138451.5	110000000.0	71941833.5	215199178.0	23944699.5
50%	6.7	250826.0	144000000.0	132177234.0	373993951.0	41039944.0
75%	7.4	460662.5	178000000.0	210245670.0	614394292.5	68532960.0
max	9.0	2758250.0	400000000.0	858373000.0	2910370905.0	357115007.0

Таблица 1. Описательные статистики

3.1. Предварительный анализ данных

Рис. 1 показывает, что дамми-переменные, характеризующие наличие премии Оскар и премии Золотой, достаточно сбалансированы: доли положительных объектов в них составляют 47,5% и 68,9% выборки соответственно.

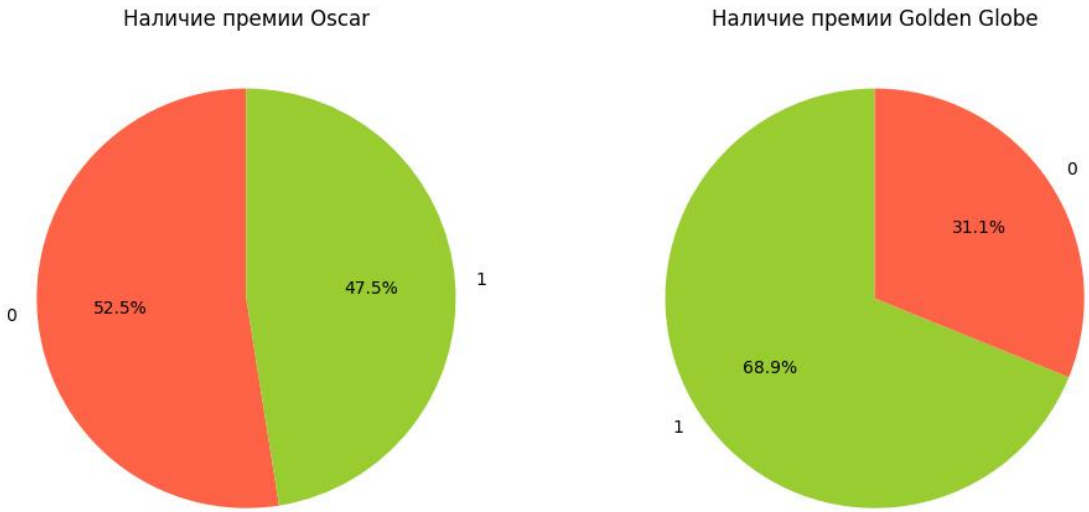


Рис 1. Сбалансированность дамми-переменных

В исходном датасете производственные бюджеты и кассовые сборы не были скорректированы с учетом инфляции, поэтому мы заменили значения данных на актуальные, использовав дефлятор инфляции (базовый год 2015). В итоговых данных оказались фильмы, снятые с 1960 по 2022 год, но так как за данный промежуток времени в индустрии кинематографа произошли значительные перемены, мы решили избавиться от выбросов, оставив в датасете фильмы, выпущенные с 1991 по 2022 гг. (Приложение, рис. 3). Пропуски были найдены и заполнены вручную, дубликаты устранены, итоговое количество наблюдений составило 379.

Также для исследования мы создали дамми-переменные oscar и golden globe, сопоставив актерский состав определенных фильмов из исследуемого датасета с информацией о наличии у хотя бы одного из членов каста наград из соответствующих датасетов *The Oscar Award, 1927 - 2023* и *Golden Globe Awards, 1944 - 2020*.

3.2. Проблема мультиколлинеарности

При исследовании данных на корреляцию было выявлено, что признаки rating и votes сильно коррелируют друг с другом: коэффициент корреляции Пирсона для этих переменных составляет 0,7 (Приложение, рис. 2). Это логично и объясняется тем, что переменная votes

обозначает количество пользователей, оценивших кинокартину на Imdb, по этим данным и рассчитывался рейтинг. Большое по модулю значение коэффициента корреляции говорит о том, что переменная votes – это практически точная линейная функция от переменной rating, а значит нельзя сказать, что вместе эти переменные обогащают модель информацией намного лучше, чем каждая по отдельности. Такая же проблема была обнаружена с переменными выручка кинокартины в первый уикенд (opening_weekend) и выручка кинокартины в домашнем прокате (domestic_gross, USD), коэффициент корреляции Пирсона, между которыми составил 0,86. Диаграмма рассеяния представлена на рис. 6 Приложения. Данная зависимость может быть объяснена тем, что показатель выручки в первые премьерные дни во многом определяет дальнейшую динамику прироста выручки фильма в домашнем прокате. Таким образом, было принято решение исключить регрессоры votes и domestic_gross из модели.

3.3. Функциональная форма переменных

Ориентиром для выбора функциональной формы зависимых переменных стала модель, представленная в статье *Predicting movie box office success using multiple regression and SVM* (V. Subramaniyaswamy, 2017). В этой работе использована модель множественной линейной регрессии, и выявлена положительная линейная зависимость между затратами на производство кинокартины и кассовыми сборами.

4. Экономическая модель

Учитывая предварительный анализ данных, экономический смысл и результаты анализа статьи, нами была оценена следующая модель:

$$Worldwide\ Gross = \beta_0 + \beta_1 Rating + \beta_2 Production\_cost + \beta_3 Opening\_weekend + \beta_4 Oscar + \beta_5 Golden\_globe + \beta_6 Certificate\_G + \beta_7 Certificate\_PG + \beta_8 Certificate\_PG\text{-}13 + \beta_9 Certificate\_R + \varepsilon.$$

Таблица 2. Общая регрессионная модель

Variable	Coefficient		Variable	Coefficient
Intercept	-1,084,005,309.0		golden globe	-23,210,549.00
	(0.000)***			(0.476)
rating	119,308,891.00		certificate_G	302,761,403.00
	(0.000)***			(0.242)
production cost	1.56		certificate_PG	311,258,216.00
	(0.000)***			(0.220)

opening weekend	5.03		certificate_PG13	294,897,731.00
	(0.000)***			(0.245)
oscar	-16,290,696.00		certificate_R	207,147,685.00
	(0.586)			(0.416)
R-squared:		0,65		
Adjusted R-squared:		0,64		
F-statistic:		76,52***		

Note: *p<0.1; **p<0.05; ***p<0.01

Примечание. В скобках указано p-value.

Построенная модель адекватна на всех разумных уровнях значимости, за базовую категорию категориальной переменной certificate, мы взяли показатель certificate_TV-MA, так как по смыслу показатель certificate_R обозначает похожее возрастное ограничение (для зрителей старше 17 лет). Согласно полученным результатам, значимое влияние на кассовые сборы оказали регрессоры: рейтинг на Imdb (rating), бюджет кинокартины (production_cost), кассовые сборы в первый уикенд (opening_weekend), однако все дамми-переменные оказались незначимыми. Значение R-squared = 0.65, что яаляется средним значением. Интерпретировать полученные значения коэффициентов оказалось не очень удобно, и мы решили также построить регрессионную модель **в логарифмической форме**.

Таблица 3. Общая регрессионная модель (логарифмическая форма)

Variable	**Coefficient**		**Variable**	**Coefficient**
Intercept	-1,59		golden globe	-0,04
	(0.44)			(0.57)
rating	2,12		certificate_G	0,79
	(0.000)***			(0.152)
production cost	0,59		certificate_PG	0,78
	(0.000)***			(0.15)
opening weekend	0,31		certificate_PG13	0,8
	(0.000)***			(0.138)
oscar	0,01		certificate_R	0,52

	(0.93)			(0.339)
R-squared:		0,54		
Adjusted R-squared:		0,53		
F-statistic:		47,77***		

Note: *p<0.1; **p<0.05; ***p<0.01

Примечание. В скобках указано p-value.

Построенная модель адекватна на всех разумных уровнях значимости. Согласно полученным результатам, значимое влияние на кассовые сборы оказали те же регрессоры, что и для модели в линейной форме, за исключением константы, все дамми-переменные также оказались незначимыми. Значение R-square снизилось до 0,54. Интерпретация значимых переменных: кассовые сборы положительно зависят от роста значений рейтинга на Imdb, производственного бюджета и сборов первый уикенд. Дамми-переменные oscar и golden globe оказались незначимыми, следовательно наличие в актерском составе лауреатов данных премий не оказывает достаточного влияния на кассовые сборы. Таким образом, гипотеза 2 отвергается.

Adjusted ROI

В анализируемой статье авторы раздели фильмы на две выборки и рассчитывали показатель Adjusted ROI (return on investments). Таким образом, они определили примерные значения ROI, при которых фильм может считаться кассовым успехом. Мы также решили проверить исследуемые данные по этому показателю, и определить процентное соотношение между провалившимися в прокате и успешными кино картинами.

Для высокобюджетных фильмов

- Кассовый провал ROI < 1 (Box office flop): 31% фильмов
- Окупились с минимальной прибылью ROI между 1 и 1.5 : 11.8% фильмов
- Достаточно успешны с ROI между 1.5 и 2.5: 19.7%
- Невероятно успешны с ROI > 2.5 (Hugely Successful): 37.5%

Для среднебюджетных фильмов

- Кассовый провал ROI < 1 (Box office flop): 32.2% фильмов
- Окупились с минимальной прибылью ROI между 1 и 1.5 : 14.95% фильмов
- Достаточно успешны с ROI между 1.5 и 2.5: 22.41%
- Невероятно успешны с ROI > 2.5 (Hugely Successful): 30.45%

По нашим данным оказалось, что высокобюджетные фильмы чаще становятся

успешнее, чем среднебюджетные, однако разница не так велика, как ожидалось (37.5% против 30.45%).

Далее для проверки гипотезы 1, исследуемые данные мы классифицировали по этим двум категориям: фильмы со средним бюджетом (бюджет от 50 до 150 миллионов долларов) и высокобюджетные фильмы (бюджет >150 миллионов долларов).

Тест Чоу

Чтобы определить лучше ли оценивать две разные модели, чем одну общую модель для всех данных, мы провели тест Чоу. Полученное значение **F-статистики равно 3.98**, что превышает **F-критическое = 1.83** при уровне значимости 5%, следовательно, нулевая гипотеза отвергается, и имеет смысл оценивать отдельные регрессии для среднебюджетных и высокобюджетных фильмов.

Таблица 4. Регрессионные модели для среднебюджетных и высокобюджетных фильмов

Medium budget film linear model		***Big budget film linear model***	
Variable	**Coefficient**	**Variable**	**Coefficient**
Intercept	-6,38	Intercept	-5,28
	(0.27)		(0.16)
rating	2,23	rating	1,7
	(0.000)***		(0.000)***
production cost	0.92	production cost	0,7
	(0.004)**		(0.000)***
opening weekend	0,24	opening weekend	0,44
	(0.000)***		(0.000)***
oscar	-0,04	oscar	0,05
	(0.61)		(0.56)
golden globe	0,14	golden globe	-0,23
	(0.165)		(0.013)**

certificate_G	0,36	certificate_G	0,65
	(0.139)		(0.21)
certificate_PG	0,24	certificate_PG	0,74
	(0.05)*		(0.15)
certificate_PG13	0,25	certificate_PG13	0,71
	(0.032)**		(0.16)
certificate_R	-	certificate_R	0,32
			(0,53)
R-squared:	0,43	R-squared:	0,58
Adjusted R-squared:	0,41	Adjusted R-squared:	0,56
F-statistic:	15,83***	F-statistic:	29,78***

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Примечание. В скобках указано p-value.

Построенные модели адекватны на всех разумных уровнях значимости. Согласно полученным результатам, значимое влияние на кассовые сборы высокобюджетных фильмов оказали регрессоры: рейтинг на Imdb (rating), бюджет кинокартины (production_cost), кассовые сборы в первый уикенд (opening_weekend). Интересно, что в данной модели дамми-переменная golden globe оказалась значимой, в отличии от общей модели.

На кассовые сборы среднебюджетных фильмов значимое влияние оказали регрессоры: рейтинг на Imdb (rating), бюджет кинокартины (production_cost), кассовые сборы в первый уикенд (opening_weekend). В данной модели дамми-переменные certificate_PG, certificate_PG-13 оказались значимыми, в отличие от общей модели и модели для высокобюджетных фильмов, а фильмов с ограничением certificate_R вовсе не оказалось в данных. Это значит, что на успех фильма со средним бюджетом также влияет и возрастной ценз. В данном случае, это фильмы предназначенные для зрителей старше 8 и 13 лет соответственно.

На основе построенных моделей регрессии, можно сделать вывод, что отдача от бюджета у фильмов со средними затратами на производство выше, чем у фильмов с высокими. Гипотеза 1 отвергается, так как значение коэффициента при production_cost у среднебюджетных фильмов выше, чем у высокобюджетных ($0.92 > 0.7$).

5. Выводы

Таким образом, проведя данное исследование мы пришли к следующим выводам:

- Гипотеза о том, что у высокобюджетных кинокартин отдача от вложений выше, чем у среднебюджетных кинокартин также не подтвердилась. Оказалось, что зависимость противоположная: именно среднебюджетные фильмы получают более высокую отдачу от затраченных на съемки средств. Тем не менее, разделение выборки на две категории в зависимости от бюджета фильма оказалось весьма полезно, так как позволило выявить некоторые различия во влиянии рассматриваемых переменных.

- Гипотеза о том, что участие в съемках фильмов кинозвезд, обладающих наградами “Оскар” и/или “Золотой Глобус”, повышает шансы кинокартины стать более успешной, не подтвердилась: на успешность высокобюджетных фильмов, действительно, влияет наличие хотя бы у одного актера премии “Золотой Глобус”, но, что удивительно, отрицательно. Это может быть объяснено ограниченностью исследования и недостатками используемых данных, или тем, что “Золотой Глобус” часто награждает актеров телесериалов, которые могут не сыскать успеха на широких экранах. “Оскар”, в свою очередь, не оказывает значимого влияния и по результатам оценки модели общей регрессии, и по каждой из категорий фильмов в отдельности. В статье Selvaretnam, G&Yang, J-Y (2015) данная гипотеза также не подтвердилась. Авторы объясняют это тем, что это довольно специфическая награда. Возможно, стоило брать во внимание не факт наличия награды, а количество награжденных актеров, снявшихся в фильме. Авторы также отмечают, что большее значение имеет именно касса фильмов в которых снимались актеры до этого.

6. Список литературы

1. V. Subramaniyaswamy, M. V. Vaibhav, R. V. Prasad and R. Logesh, "Predicting movie box office success using multiple regression and SVM," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 2017, pp. 182-186, doi: 10.1109/ISS1.2017.8389394.
2. Selvaretnam, G&Yang, J-Y 2015 'Factors affecting the financial success of motion pictures : what is the role of star power?' School of Economics & Finance Discussion Paper , no. 1501, University of St Andrews, pp. 1-25 .
3. P. Walanaraya, W. Puengpipattrakul and D. Sutivong, "Movie Revenue Prediction Using Regression and Clustering," 2018 2nd International Conference on Engineering Innovation (ICEI), Bangkok, Thailand, 2018, pp. 63-68, doi: 10.1109/ICEI18.2018.8448610.

7. Приложение

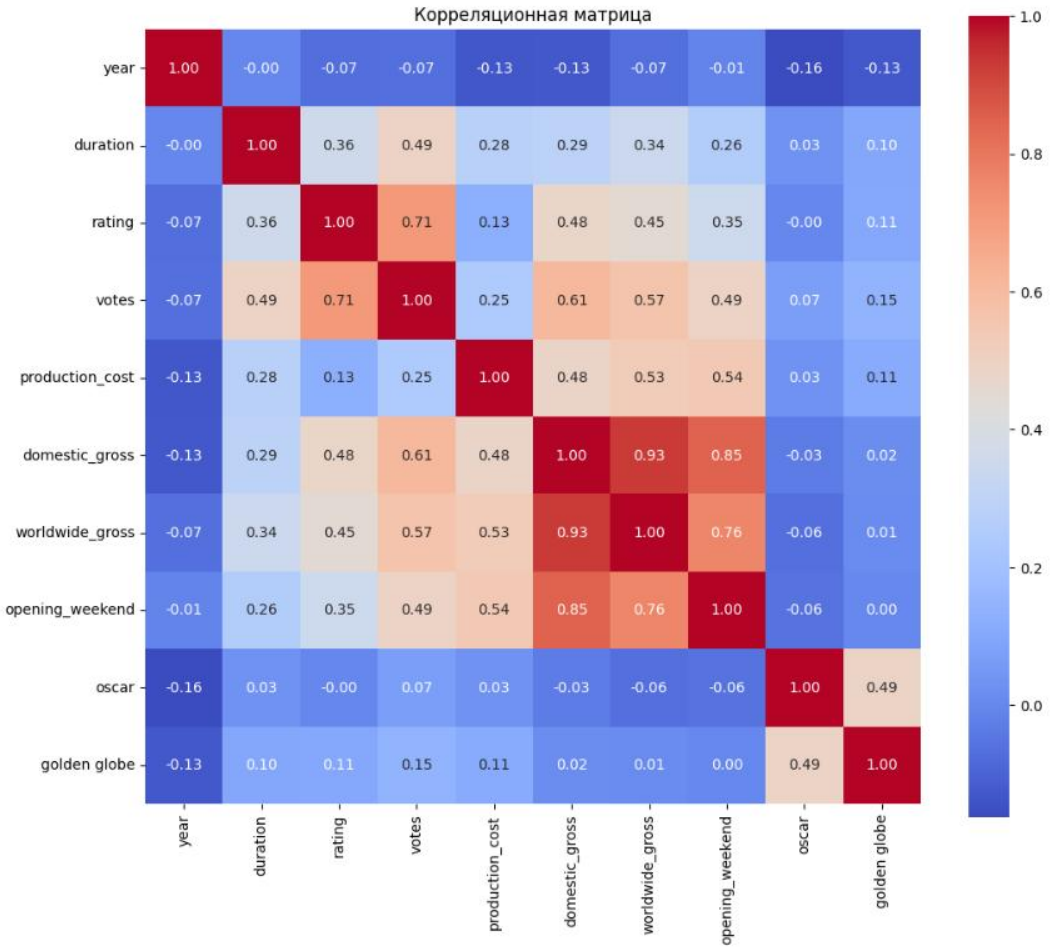
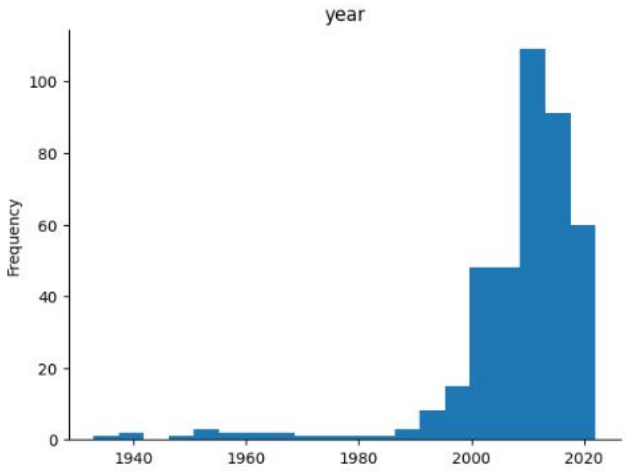


Рис. 2. Корреляционная матрица регрессоров и зависимой переменной



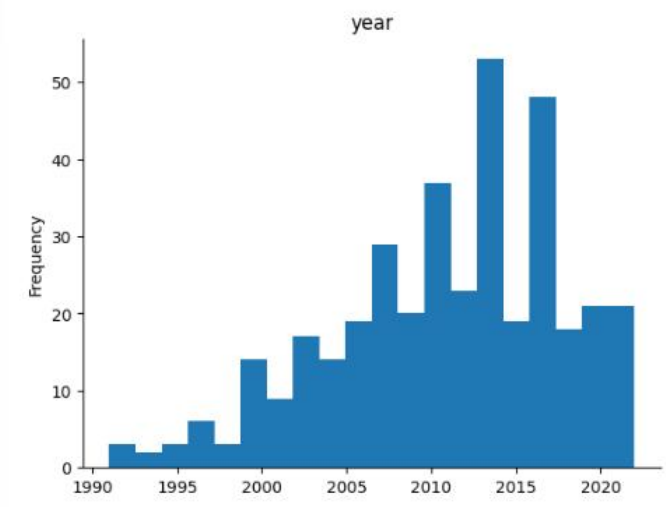


Рис. 3. Распределение переменной year до и после обработки выбросов

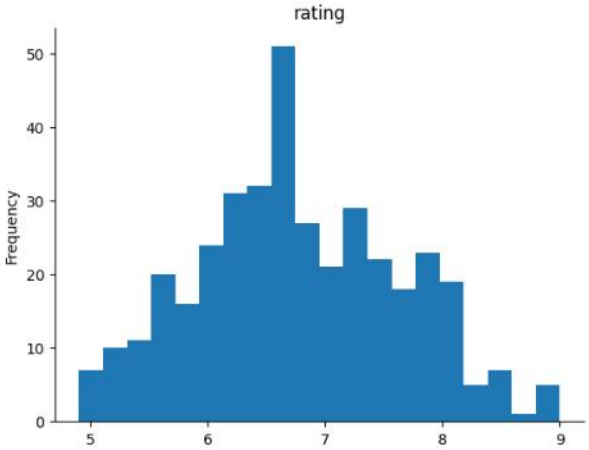


Рис. 4. Распределение переменной rating

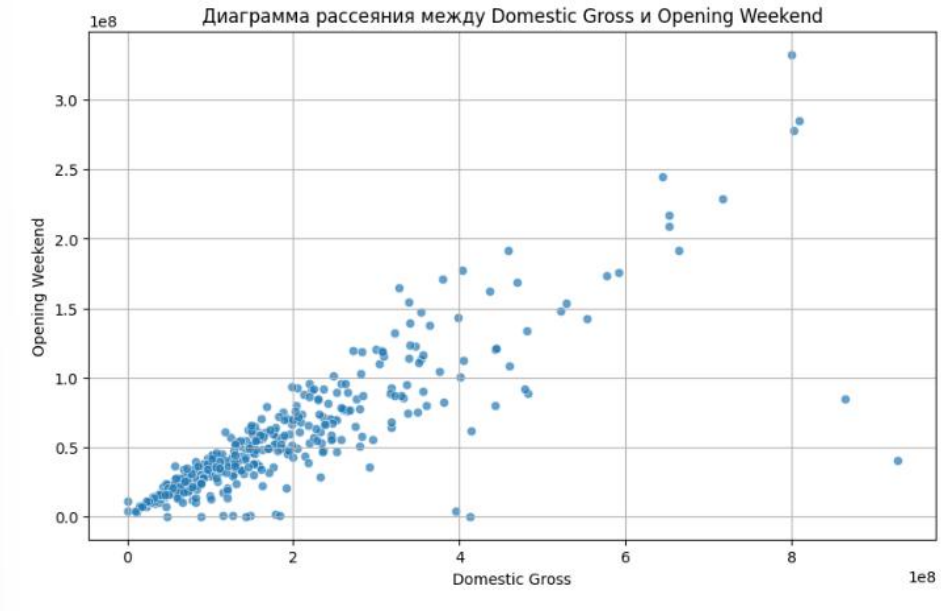


Рис. 5. Диаграмма рассеяния между Domestic Gross и Opening Weekend