

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



**Кафедра прикладной информатики
и инноватики**

**Направление: 09.03.03 "Прикладная
информатика"**

**Отчёт по проектному модулю дисциплины “Большие
данные”**

“Анализ продаж компьютерных игр”

Выполнил студент 3 курса 6 группы _____ Краус А. В.
подпись

Выполнил студент 3 курса 7 группы _____ Петренко Д. А.
подпись

Ростов-на-Дону – 2024

1. Цель кейса

Цель данной работы заключается в проведении комплексного анализа продаж компьютерных игр на основе имеющегося датасета с целью выявления ключевых факторов, влияющих на успешность игр. Это включает в себя анализ продаж по регионам, исследование корреляций между оценками и продажами, выявление популярных игр, жанров и платформ, а также разработка модели для прогноза продаж новых игр в серии.

2. Актуальность

Анализ данных о продажах компьютерных игр имеет высокую актуальность по нескольким причинам:

1. **Экономическое значение:** Индустрия компьютерных игр является одной из самых быстрорастущих отраслей развлечений, приносящей значительные доходы.
2. **Конкурентное преимущество:** Понимание факторов, влияющих на успешность игр, помогает компаниям-разработчикам и издателям принимать более обоснованные решения, что может привести к повышению прибыли и успешности их продуктов.
3. **Развитие технологий:** Применение методов анализа больших данных и машинного обучения в этой области способствует развитию и совершенствованию аналитических инструментов и технологий, что может быть применено и в других отраслях.
4. **Потребительские предпочтения:** Анализ позволяет лучше понимать предпочтения и поведение потребителей, что важно для разработки и маркетинга новых игр.

3. Гипотеза

1. **Гипотеза о корреляции между оценками и продажами:** Высокие оценки критиков положительно коррелируют с объемами продаж игр.
2. **Гипотеза о популярности жанров и платформ:** Определенные жанры и платформы имеют более высокие продажи по сравнению с другими, что может быть обусловлено текущими рыночными трендами и предпочтениями игроков.
3. **Гипотеза об игровых региональных различиях:** Продажи игр существенно различаются в зависимости от региона, что может быть связано с культурными, экономическими и демографическими факторами.
4. **Гипотеза о предсказании продаж:** Модель машинного обучения, обученная на данных о продажах и оценках, способна точно предсказать продажи новой игры в серии.
5. **Гипотеза о влиянии разработчиков и издателей:** Игры от определенных разработчиков и издателей получают более высокие оценки критиков и имеют больший объем продаж, что свидетельствует о влиянии бренда и репутации на успех игры.

Эти гипотезы будут проверены и проанализированы в ходе выполнения работы, что позволит сделать выводы о ключевых факторах, влияющих на успешность компьютерных игр.

4. Описание датасета

Датасет `vgchartz` представляет собой набор данных о продажах компьютерных игр в различных регионах, оценках критиков, имя издателя и разработчика, жанр и прочая информация, характеризующая игру. Данные были собраны сторонним автором посредством парсинга с такого ресурса, как `vgchartz.com` и опубликованы в веб-ресурсе `kaggle.com`.

Данный датасет содержит следующие столбцы:

- `img` - ссылка на обложку игры
- `title` - название игры
- `console` - платформа, на которой выпущена игра
- `genre` - жанр игры
- `publisher` - издатель игры
- `developer` - разработчик игры
- `critic_score` - оценка критиков
- `total_sales` - общие продажи игры
- `na_sales` - продажи игры в Северной Америке
- `jp_sales` - продажи игры в Японии
- `pal_sales` - продажи игры в Европе
- `other_sales` - продажи игры в других регионах
- `release_date` - дата релиза игры
- `last_update` - последнее обновление игры

По заявлению автора им были удалены несколько столбцов, а именно

- `vg_score` - оценка игры от источника (`vgchartz.com`)
- `user_score` - оценка игры от игроков
- `total_shipped` - общее кол-во проданных копий

Основанием для удаления этих столбцов послужило то, что эти столбцы в большинстве случаев имели нулевые значения и, вследствие этого, не несли значимой ценности для анализа данных.

Мы также сделали предобработку датасета, удалив столбец *img*, так как ссылка на обложку игры не несет в себе ценности для анализа предоставленных данных.


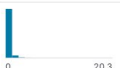


# img	# title	# console	# genre	# publisher	# developer	# critic_score	# total_sales	# na_sales	# jp_sales
= the uri for the box art at vgchartz.com (type: str)	Game title. (type: str)	Console the game was released for (type: str)	Genre of the game (type: str)	Publisher of the game (type: str)	Developer of the game (type: str)	the metacritic score (out of 10). (type: float)	Global sales of copies in millions. (type: float)	North American sales of copies in millions. (type: float)	Japanese sales of copies in millions (type: float)
/games/boxart/de... 12%	39798 unique values	PC 20%	Misc 15%	Unknown 14%	Unknown 7%				
/games/boxart/full... 0%		PS2 6%	Action 13%	Sega 3%	Konami 2%				
Other (56204) 88%		Other (47834) 75%	Other (46155) 72%	Other (52967) 83%	Other (58605) 92%				
/games/boxart/full_6518548AmericaFrontcc.jpg	Grand Theft Auto V	PS3	Action	Rockstar Games	Rockstar North	9.4	28.32	6.37	0.99
/games/boxart/full_5563178AmericaFrontcc.jpg	Grand Theft Auto V	PS4	Action	Rockstar Games	Rockstar North	9.7	19.39	6.06	0.6
/games/boxart/827563ccc.jpg	Grand Theft Auto: Vice City	PS2	Action	Rockstar Games	Rockstar North	9.6	16.15	8.41	0.47
/games/boxart/full_9218923AmericaFrontcc.jpg	Grand Theft Auto V	X360	Action	Rockstar Games	Rockstar North		15.86	9.06	0.06
/games/boxart/full_4990510AmericaFrontcc.jpg	Call of Duty: Black Ops 3	PS4	Shooter	Activision	Treyarch	8.1	15.09	6.18	0.41
/games/boxart/full_call-of-duty-modern-warfare-3_517AmericaFront.jpg	Call of Duty: Modern Warfare 3	X360	Shooter	Activision	Infinity Ward	8.7	14.82	9.07	0.13
/games/boxart/full_call-of-duty-black-ops_5AmericaFront.jpg	Call of Duty: Black Ops	X360	Shooter	Activision	Treyarch	8.8	14.74	9.76	0.11

Рис. 1 - структура датасета.

5. Ход работы

Для выполнения работы мы использовали такие инструменты, как `pySpark`, `matplotlib`.

Каждое из приведенных выше задач мы реализовывали в виде `python` функции и запускали в Jupyter Notebook

Задача 1. Анализ продаж по регионам

В данном задании мы анализировали продажи игр по регионам. Сначала, в функции `get_game_sales_by_regions`, мы отфильтровали строки `df` по названию игры и сгруппировали данные по колонке `title`, суммируя продажи в разных регионах: Северной Америке (`na_sales`), Японии (`jp_sales`), регионе PAL (`pal_sales`) и других регионах (`other_sales`). Затем данные выводятся на экран и передаются в функцию `visualize_game_sales_by_regions`.

Во второй функции, `visualize_game_sales_by_regions`, данные преобразуются в формат `Pandas DataFrame`, чтобы построить диаграммы распределения продаж по регионам для каждой игры. Для каждой группы игр создаются графики, которые сохраняются в виде изображений. `*/`

Распределение продаж игры по регионам

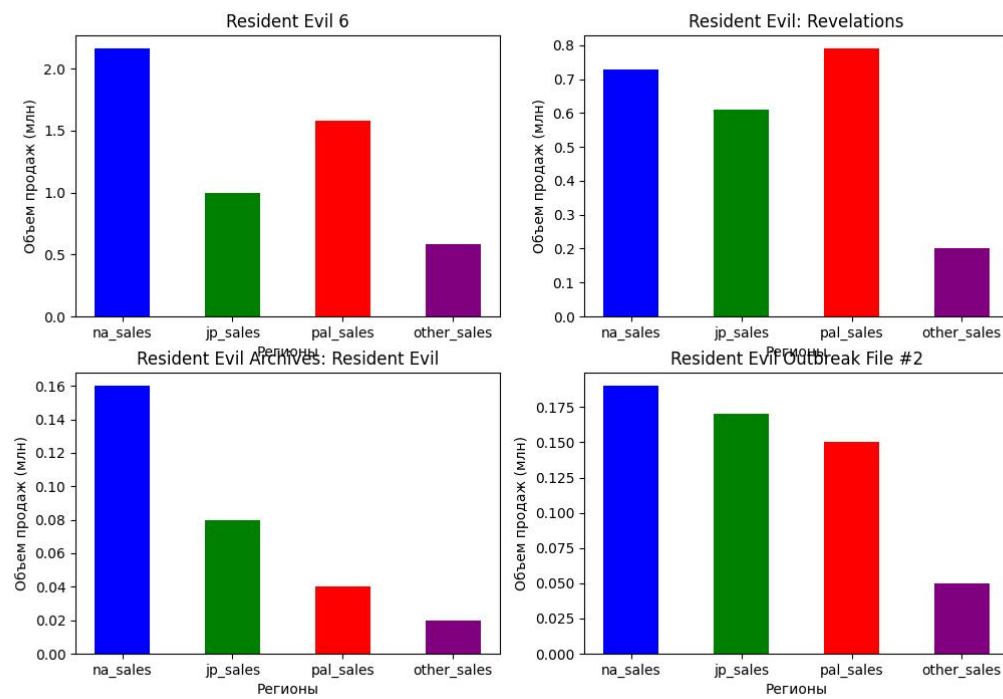


Рис. 2 - график продаж по регионам определенной игры.

Задача 2. Вычисление корреляции между оценкой и продажами

В данном задании мы вычисляли корреляцию между оценками критиков и продажами игр. Сначала, в функции `get_game_sale_estimates`, мы очистили `DataFrame` от строк с отсутствующими значениями в колонках: оценки критиков (`critic_score`), продажи в Северной Америке (`na_sales`), Японии (`jp_sales`), регионе PAL (`pal_sales`), других регионах (`other_sales`) и общие продажи (`total_sales`).

Затем мы вычислили корреляцию между оценками критиков и продажами в каждом регионе, сохранив результаты в словаре `correlations`, и вывели их на экран.

Далее мы преобразовали очищенные данные в формат `Pandas DataFrame` и передали их в функции `build_sales_distribution_by_critic_score_plot` и `build_heatmap_correlation_matrix` для визуализации.

Первая функция строит гистограмму распределения оценок критиков, а вторая — тепловую карту матрицы корреляции, отображающую взаимосвязи между оценками критиков и продажами в различных регионах.

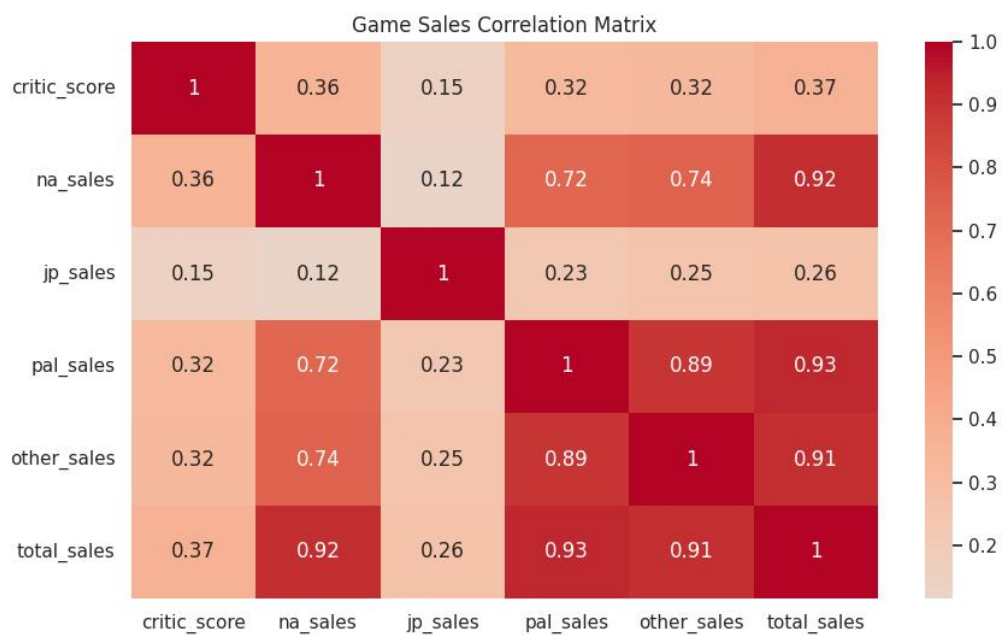


Рис. 3 - тепловая карта, характеризующая корреляцию между продажами в различных регионах и оценками критиков

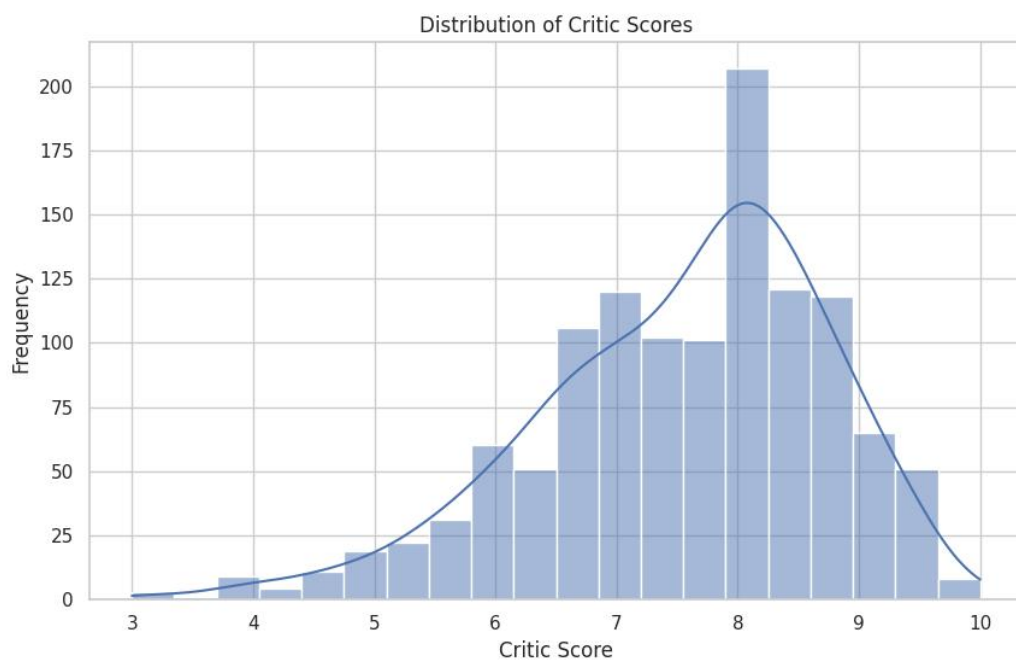


Рис. 4 - график распределения объёма продаж по оценкам критиков

Задача 3. Выявление самых популярных игр

В данном задании мы выявляли самые популярные игры по регионам. Сначала, в функции `get_popular_games_by_region`, мы сгруппировали данные `DataFrame` по колонке с названием игры (`title`) и агрегировали данные по указанному региону (`region_column`), суммируя продажи.

Затем мы округлили числовые значения в полученном `DataFrame` до трех знаков после запятой и отсортировали данные по продажам в указанном регионе в порядке убывания. Итоговый `DataFrame` был выведен на экран и передан в функцию `visualize_popular_games`.

Во второй функции, `visualize_popular_games`, данные преобразуются в формат `Pandas DataFrame` и строится круговая диаграмма, отображающая доли продаж самых популярных игр в указанном регионе.

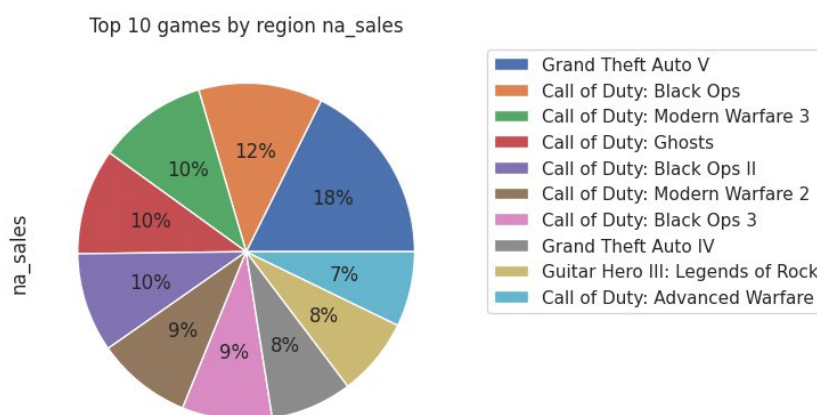


Рис. 5 – диаграмма самых популярных игр по региону `na_sales`

Задача 4. Выявление самых популярных жанров

В данном задании мы выявляли самые популярные жанры по регионам. Сначала, в функции `get_popular_genres_by_region`, мы сгруппировали данные DataFrame по колонке с жанром (`genre`) и агрегировали данные по указанному региону (`region_column`), суммируя продажи. Полученные данные были отсортированы по продажам в указанном регионе в порядке убывания и округлены до трех знаков после запятой.

Затем агрегированные данные выводятся на экран и передаются в функцию `visualize_popular_genres`.

Во второй функции, `visualize_popular_genres`, данные преобразуются в формат Pandas DataFrame и строится круговая диаграмма, отображающая процентное соотношение продаж для каждого жанра в указанном регионе.

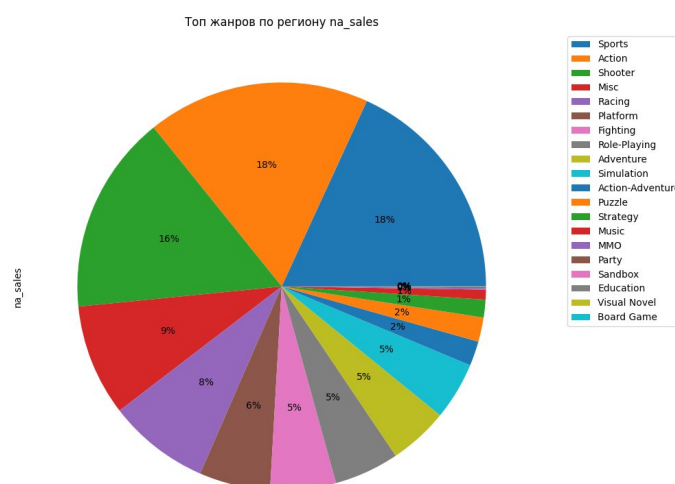


Рис. 6 - диаграмма, характеризующая популярность жанров игр по результатам продаж.

Задача 5. Выявление самых популярных платформ

В данном задании мы выявляли самые популярные платформы по регионам. Сначала, в функции `get_popular_platforms_by_region`, мы сгруппировали данные DataFrame по колонке с названием консоли (`console`) и агрегировали данные по указанному региону (`region_column`), суммируя продажи. Полученные данные были отсортированы по продажам в указанном регионе в порядке убывания и округлены до трех знаков после запятой.

Затем агрегированные данные выводятся на экран и передаются в функцию `visualize_popular_platforms_by_region`.

Во второй функции, `visualize_popular_platforms_by_region`, данные преобразуются в формат Pandas DataFrame и строится круговая диаграмма, отображающая объем продаж для каждой платформы в указанном регионе.

Platforms by sales volume by region: na_sales in period 2015-01-01 - 2024-01-01

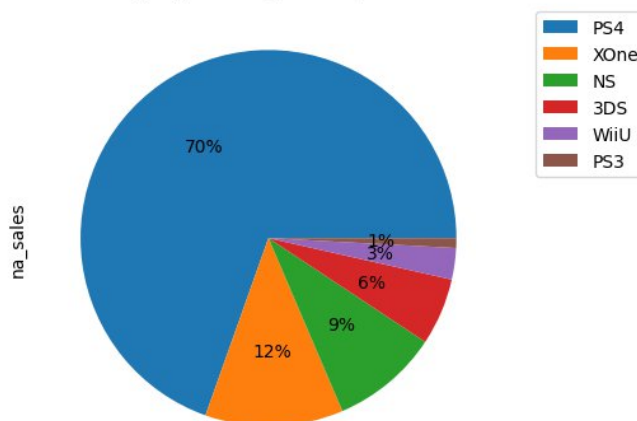


Рис. 7 - диаграмма, характеризующая долю платформы в том или ином регионе.

Задача 6. Предсказание продаж игр серии

В данном задании мы предсказывали успех будущих игр серии. Сначала, в функции `predict_game_success`, мы фильтруем данные `DataFrame` по названию серии игр `game_series_name` и выбираем необходимые колонки, преобразуя типы данных и заполняя пропуски.

Далее мы выделяем региональные продажи и оценки критиков в отдельные признаки, используя `VectorAssembler`. Затем разделяем данные на обучающую и тестовую выборки.

Для обучения модели используется линейная регрессия (`LinearRegression`), которая обучается на обучающей выборке и оценивается на тестовой выборке, вычисляя корень средней квадратичной ошибки (RMSE).

Для предсказания успеха следующей игры серии создается новый `DataFrame` `new_game` с данными новой игры, который также преобразуется в формат, пригодный для модели. На основе этих данных модель делает прогноз, и предсказанные значения выводятся на экран.

Возвращается предсказанное значение продаж для новой игры серии.

title	console	genre	publisher	developer	critic_score	total_sales	na_sales	jp_sales	pal_sales	other_sales	release_date	last_update
Call of Duty: Ghosts	PS4	Shooter	Activision	Infinity Ward	7.5	4.17	1.79	0.05	1.64	0.69	2013-11-15	2018-03-21
Call of Duty: Advanced Warfare	PS4	Shooter	Activision	Sledgehammer Games	8.5	7.53	2.84	0.14	3.34	1.22	2014-11-04	2018-01-04
Call of Duty: Black Ops 3	PS4	Shooter	Activision	Treyarch	8.1	15.09	6.18	0.41	6.05	2.44	2015-11-06	2018-01-14
Call of Duty: Infinite Warfare	PS4	Shooter	Activision	Infinity Ward	7.9	8.48	3.11	0.19	3.83	1.36	2016-11-04	2018-01-14
Call of Duty: Modern Warfare Remastered	PS4	Shooter	Activision	Infinity Ward	7.9	0.58	0.17	0.03	0.3	0.09	2017-06-27	2018-01-14
Call of Duty: WWII	PS4	Shooter	Activision	Sledgehammer Games	8.1	13.4	4.67	0.4	6.21	2.12	2017-11-03	2017-12-31

Рис. 8 – данные, для которых будет производиться прогноз продаж и оценки

```

Training model for critic_score...
24/06/26 19:40:30 WARN DecisionTreeMetadata: DecisionTree reducing maxBins from 32 to 5 (= number of training instances)
R^2 for critic_score: -inf
Training model for total_sales...
24/06/26 19:40:32 WARN DecisionTreeMetadata: DecisionTree reducing maxBins from 32 to 5 (= number of training instances)
R^2 for total_sales: -inf
Training model for na_sales...
24/06/26 19:40:34 WARN DecisionTreeMetadata: DecisionTree reducing maxBins from 32 to 5 (= number of training instances)
R^2 for na_sales: -inf
Training model for jp_sales...
24/06/26 19:40:35 WARN DecisionTreeMetadata: DecisionTree reducing maxBins from 32 to 5 (= number of training instances)
R^2 for jp_sales: -inf
Training model for pal_sales...
24/06/26 19:40:37 WARN DecisionTreeMetadata: DecisionTree reducing maxBins from 32 to 5 (= number of training instances)
R^2 for pal_sales: -inf
Training model for other_sales...
24/06/26 19:40:39 WARN DecisionTreeMetadata: DecisionTree reducing maxBins from 32 to 5 (= number of training instances)
R^2 for other_sales: -inf
Predicted values for the next game in the series:
critic_score: 8.120000123977661
total_sales: 7.05200006365776
na_sales: 2.7339999213814736
jp_sales: 0.1659999990835786
pal_sales: 3.0189999997615815
other_sales: 1.1390000253915786

```

Рис. 9 – прогноз продаж и оценки для игры (Random Tree).

```

24/06/26 19:40:18 WARN InstanceBuilder: Failed to load implementation from:dev.ludovic.netlib.blas.JNIBLAS
24/06/26 19:40:18 WARN InstanceBuilder: Failed to load implementation from:dev.ludovic.netlib.blas.VectorBLAS
24/06/26 19:40:18 WARN InstanceBuilder: Failed to load implementation from:dev.ludovic.netlib.lapack.JNILAPACK
evaluator: RegressionEvaluator_e0eba0fd83c4
Root Mean Squared Error (RMSE) on test data for critic_score: 0.7235570462073007

evaluator: RegressionEvaluator_6182471f94a9
Root Mean Squared Error (RMSE) on test data for total_sales: 0.03666198391750086
evaluator: RegressionEvaluator_d42afdc56d26
Root Mean Squared Error (RMSE) on test data for na_sales: 0.24321103247299725
evaluator: RegressionEvaluator_42cbaa257d04
Root Mean Squared Error (RMSE) on test data for jp_sales: 0.041811139688541665
evaluator: RegressionEvaluator_341cd2b00856
Root Mean Squared Error (RMSE) on test data for pal_sales: 0.29654822733356667
evaluator: RegressionEvaluator_89bc0b0cb73d
Root Mean Squared Error (RMSE) on test data for other_sales: 0.010249781800135693
Predicted Critic Score: 8.09800488894128
Predicted Total Sales: 9.707214261666941
Predicted NA Sales: 3.3455177821083257
Predicted JP Sales: 0.2707888170883259
Predicted PAL Sales: 4.5075029907694955
Predicted Other Sales: 1.5158784108667214

```

Рис. 10 - прогноз продаж и оценки для игры (Linear Regression)

Задача 7. Распределение разработчиков игр по оценкам критиков

В данном задании мы выявляли распределение разработчиков игр по оценкам критиков. Сначала, в функции `get_top_developers_by_critic_score`, мы преобразуем данные DataFrame в формат Pandas DataFrame и группируем их по колонке `developer`, вычисляя среднее значение оценок критиков (`critic_score`) и сумму продаж (`total_sales`) для каждого разработчика.

Затем мы сортируем разработчиков по среднему значению оценок критиков и общим продажам в порядке убывания. Отбираем топ-10 разработчиков с оценками критиков, у которых общие продажи превышают 10 млн.

Полученные данные передаются в функцию `visualize_df`.

Во второй функции, `visualize_df`, данные визуализируются с помощью Seaborn, создавая график, где на оси X отображаются оценки критиков, на оси Y — общие продажи.

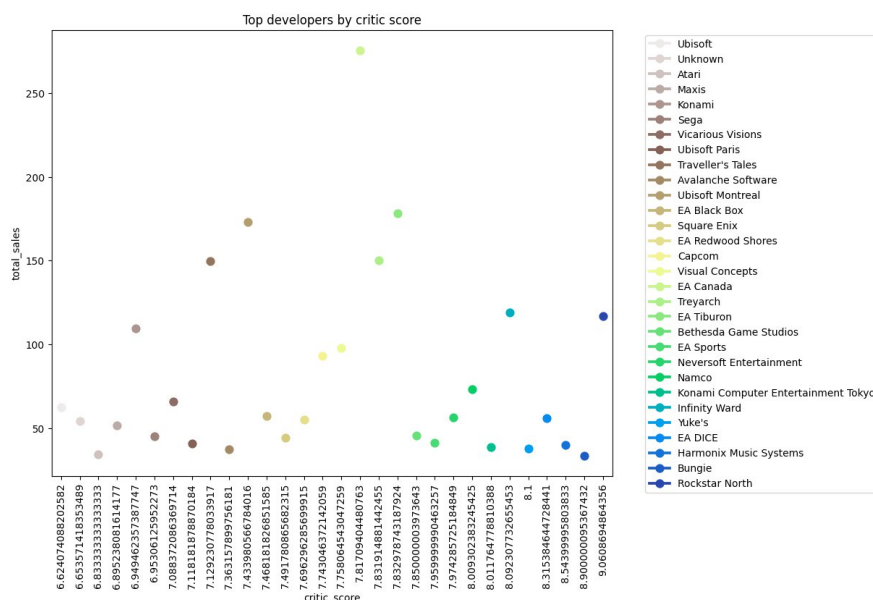


Рис. 11 - график распределения разработчиков по оценкам критиков и объему продаж.

Задача 8. Топ жанров по оценкам критиков и общим продажам

В данном задании мы выявляли топ жанров по оценкам критиков и общим продажам. Сначала, в функции `get_top_genres_by_critic_score_and_sales`, мы группируем данные `DataFrame` по колонке `genre` и вычисляем медианное значение оценок критиков (`critic_score`) и сумму продаж (`total_sales`) для каждого жанра. Затем сортируем данные по убыванию оценок критиков и продаж. Полученный `DataFrame` преобразуется в формат `Pandas DataFrame` и передается в функцию `visualize_df`.

Во второй функции, `visualize_df`, данные визуализируются с помощью `Seaborn`. Создаются два графика:

1. На первом графике отображаются жанры с самыми высокими медианными оценками критиков. Для каждого жанра на оси `X` отображается название жанра, а на оси `Y` — медианное значение оценки критиков. Столбцы графика окрашены в разные цвета, и на них нанесены метки значений.
2. На втором графике отображаются жанры с самыми высокими общими продажами. На оси `X` отображается название жанра, а на оси `Y` — общие продажи. Столбцы графика также окрашены в разные цвета и имеют метки значений. Значения на оси `Y` форматированы в миллионах.

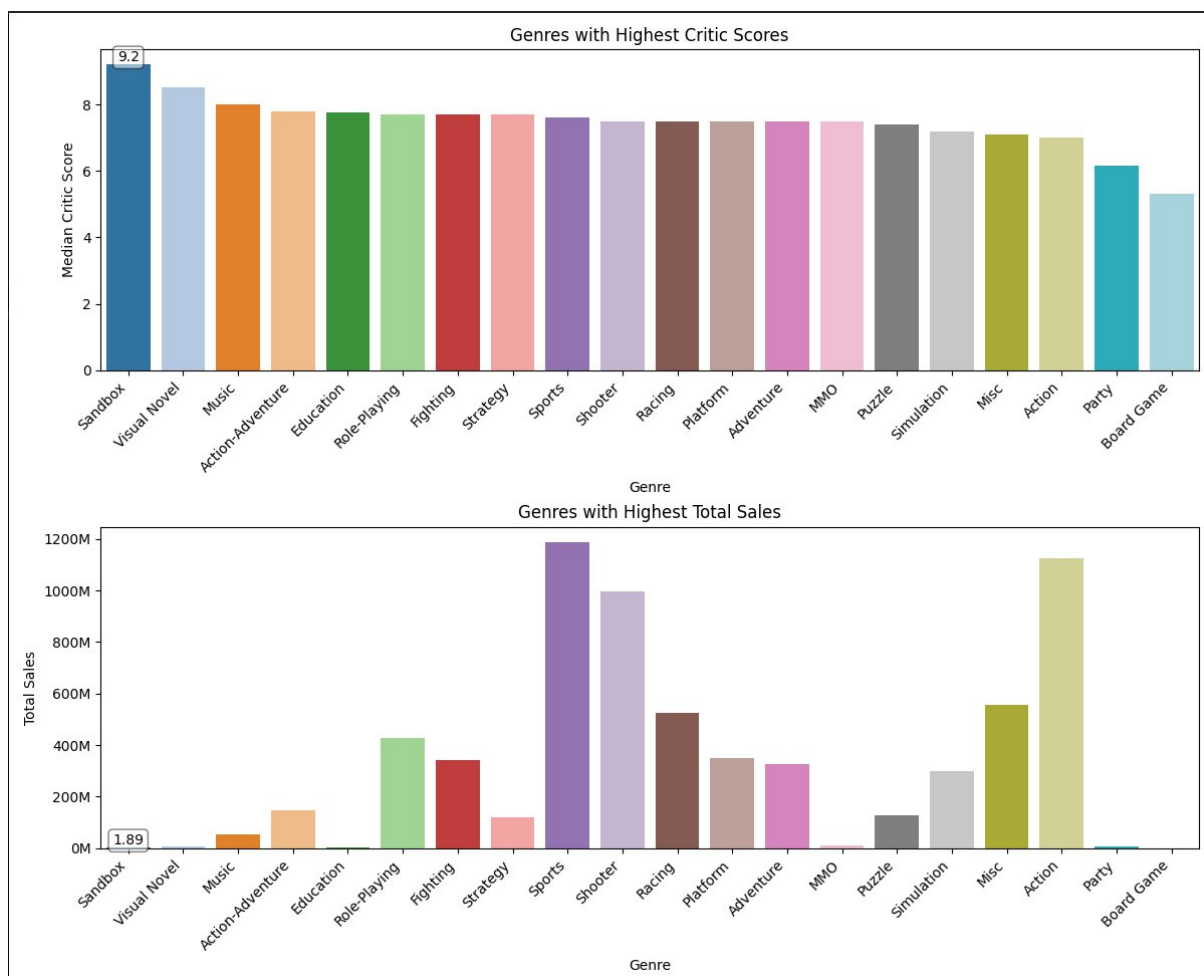


Рис. 12 - график топа жанров по оценкам критиков и объему продаж

Задача 9. Топ издателей по кол-ву выпущенных тайтлов и объёму продаж

В данном задании мы выявляли топ издателей по количеству выпущенных тайтлов и объёму продаж. Сначала, в функции `get_top_sales_performance_by_publisher`, мы группируем данные `DataFrame df` по колонке `publisher`, суммируя общие продажи (`total_sales`) и подсчитывая количество выпущенных тайтлов (`title`). Затем данные сортируются по убыванию общих продаж и преобразуются в формат `Pandas DataFrame`.

Из полученных данных выделяются топ-10 издателей по общим продажам и топ-10 издателей по количеству выпущенных тайтлов.

Далее данные визуализируются с помощью `Seaborn` и `matplotlib`. Создаются два графика:

1. На первом графике отображается количество выпущенных тайтлов для каждого издателя. На оси X отображается количество тайтлов, а на оси Y — названия издателей. Столбцы графика окрашены в разные цвета, и на них нанесены метки значений.
2. На втором графике отображается объем продаж для каждого издателя. На оси X отображается объем продаж в миллионах, а на оси Y — названия издателей. Столбцы графика также окрашены в разные цвета и имеют метки значений.

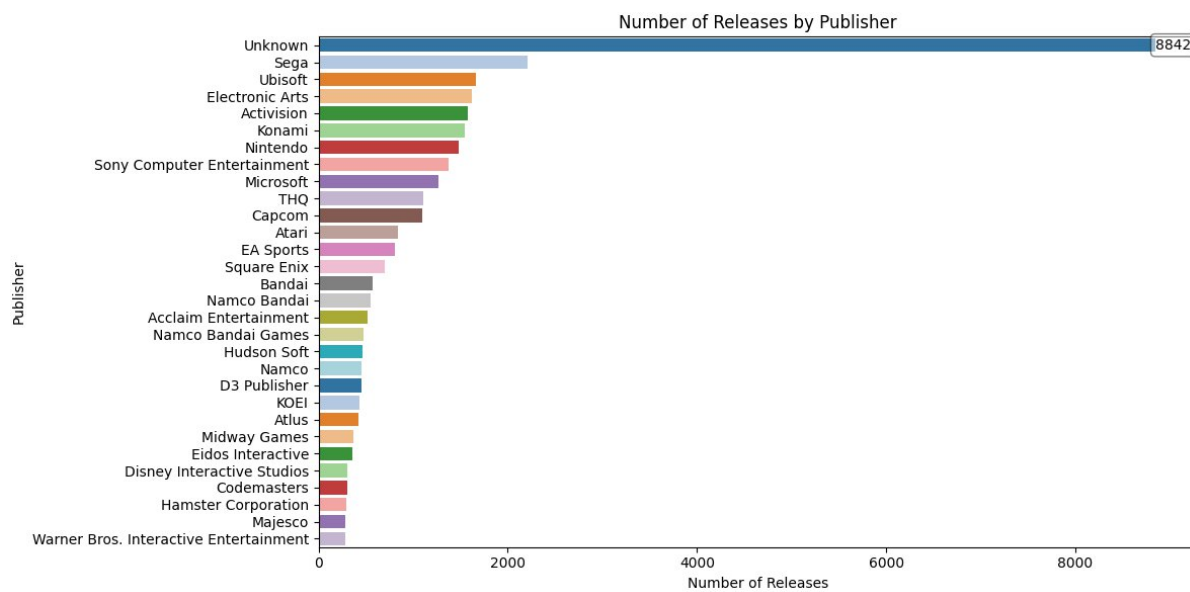


Рис. 13 - топ издателей по кол-ву выпущенных тайтлов.



Рис. 14 - топ издателей по кол-ву выпущенных тайтлов за каждый год в промежутке 2010-2015 гг.

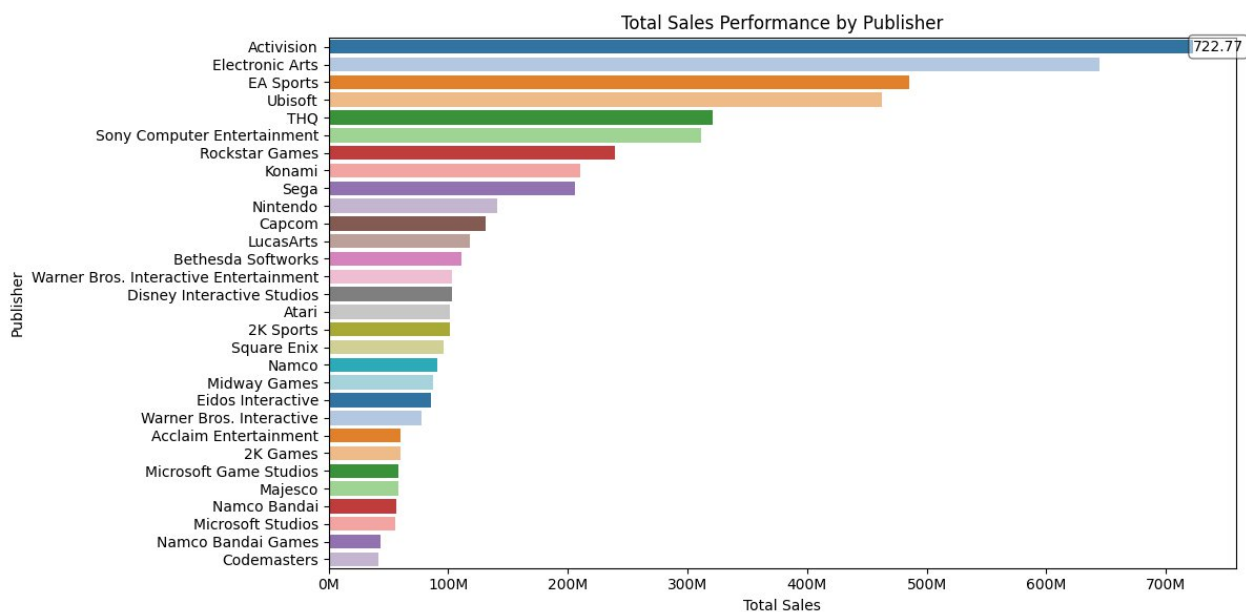


Рис. 15 - топ издателей по объему продаж

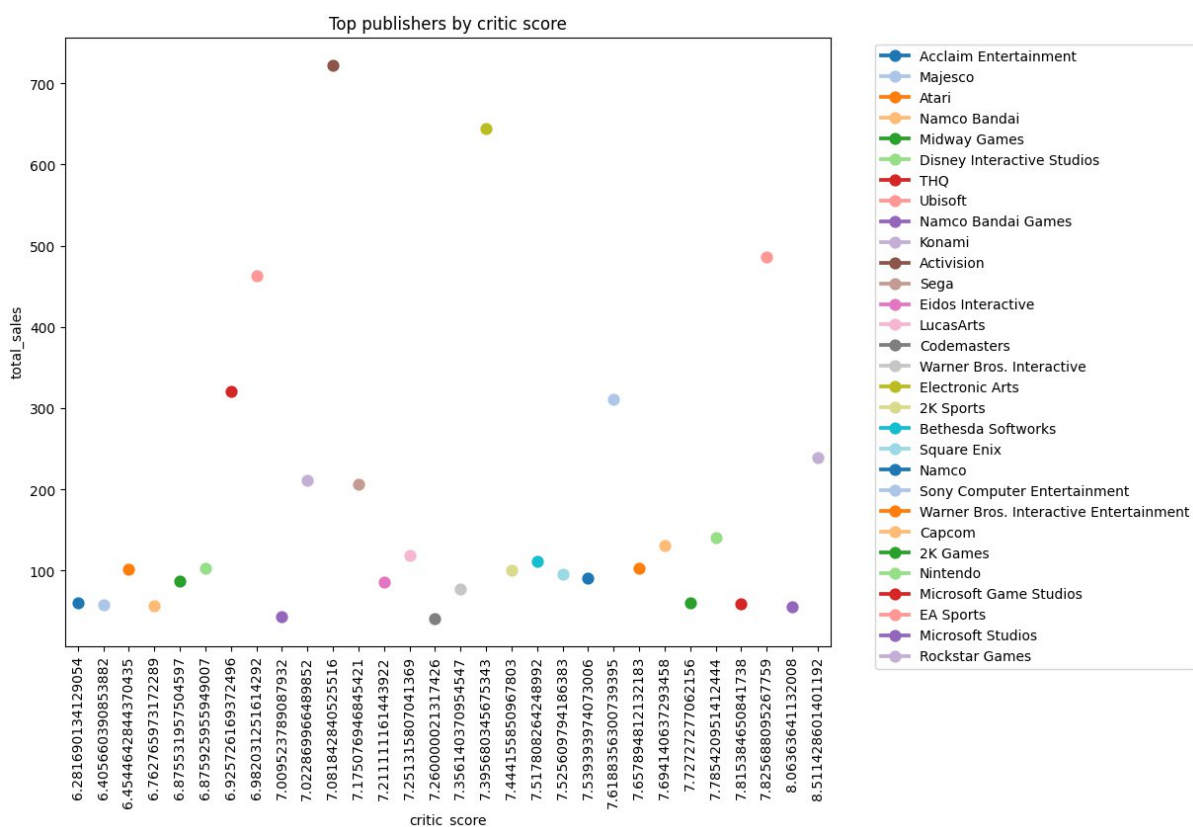


Рис. 16 - распределение издателей по продажам и средним показателям оценок критиков.

6. Заключение

В ходе выполнения данной работы мы провели комплексный анализ продаж компьютерных игр, используя разнообразные методы больших данных. Наши исследования охватили несколько ключевых аспектов, важных для понимания рынка компьютерных игр и определения факторов, влияющих на их успех.

Анализ продаж по регионам: Мы выявили, что распределение продаж существенно отличается в различных регионах. Некоторые игры демонстрируют высокие продажи в Северной Америке, тогда как в Японии или регионе PAL они могут продаваться значительно хуже. Это подтверждает гипотезу о существовании региональных различий в предпочтениях потребителей.

Выявление самых популярных игр: Самые популярные игры варьируются в зависимости от региона, что согласуется с выводами по задаче 1. Это также подтверждает гипотезу о региональных различиях в предпочтениях пользователей.

Выявление самых популярных жанров: Анализ показал, что жанр Sport, пользуются наибольшей популярностью во многих регионах. Это подтверждает гипотезу о том, что определенные жанры игр имеют глобальную привлекательность.

Анализ платформ: Наше исследование показало, что различные платформы имеют свои уникальные аудитории и популярные жанры. Это знание помогает разработчикам и издателям фокусироваться на наиболее перспективных платформах для их продуктов.

Топ жанров по оценкам критиков и общим продажам: Анализ показал, что жанры с высокими оценками критиков часто совпадают с жанрами, которые имеют высокие продажи. Это подтверждает гипотезу о том, что качественные игры в определенных жанрах имеют высокий коммерческий потенциал.

Оценки игр и продажи: Мы обнаружили положительную корреляцию между высокими оценками критиков и успешностью продаж игр. Это подчеркивает важность качества продукта и отзывов критиков в формировании потребительского спроса.

Прогноз будущих продаж: Разработанная нами модель прогнозирования продаж показала удовлетворительные результаты, демонстрируя возможность использования машинного обучения для предсказания успеха будущих игр. Это предоставляет разработчикам и издателям инструмент для более точного планирования своих действий.

В заключение наше исследование предоставило глубокое понимание рынка компьютерных игр и выявило ключевые факторы, влияющие на их продажи. Эти выводы могут быть использованы для оптимизации стратегий разработки и маркетинга, что в конечном итоге приведет к более успешным и востребованным продуктам на рынке.

- 1. Гипотеза о корреляции между оценками и продажами:**
выдвинутая нами гипотеза о положительной корреляции между оценками и продажами отвергается. Проведя анализ данных, мы увидели, что корреляция между данными признаками стремится к нулю, а это говорит о слабом влиянии оценок критиков на продажи.

2. **Гипотеза о популярности жанров и платформ:** выдвинутая нами гипотеза о популярности жанров и платформа оказалась верна. Действительно, в определенный период времени в различных регионах есть свои фавориты среди игровых жанров и платформ.
3. **Гипотеза о региональных различиях:** наш анализ подтвердил гипотезу о том, что продажи игр существенно различаются в зависимости от региона. Эти различия можно объяснить культурными, экономическими и демографическими факторами.
4. **Гипотеза о прогнозе оценок и продаж:** Модель машинного обучения, основанная на алгоритме случайного дерева, не способна делать прогноз ввиду малого количества данных и об этом говорит оценщик модели, значение которого равно минус бесконечности. (рис. 9) Оценщик второй модели на линейной регрессии (RMSE) показывает лучшие значения по сравнению с первым, но все равно не справляется с задачей точного предсказания будущих продаж, что может быть связано с малым количеством данных и параметров для предсказания.
5. **Гипотеза о влиянии разработчиков и издателей:** Данная гипотеза частично принимается и отвергается. По рис. 10, 15 можно увидеть, что некоторые издатели, имеющие высокие продажи, имеют относительно низкие показатели средних оценок критиков по всем выпущенным тайтлам, а у разработчиков обратная ситуация: высокие оценки, но не очень высокие продажи (за исключением Rockstar).