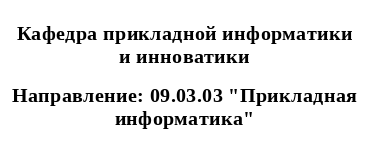
**МИНОБРНАУКИ РОССИИ**

**Федеральное государственное автономное образовательное**  
 **учреждение высшего образования**   
**«Южный федеральный университет»**

**Институт высоких технологий и пьезотехники**



**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 – 20****24**

# **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**

Датаcет 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.

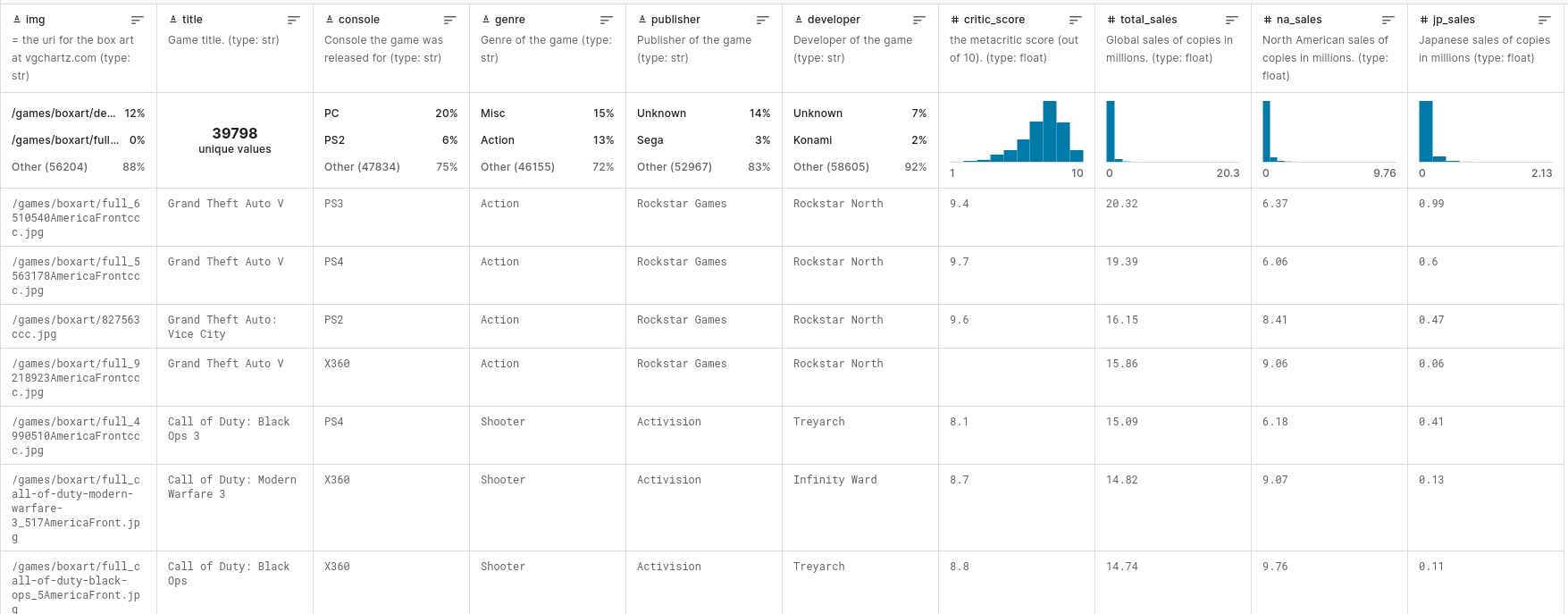


Рис. 1 - структура датасета.

# **5. Ход работы**

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

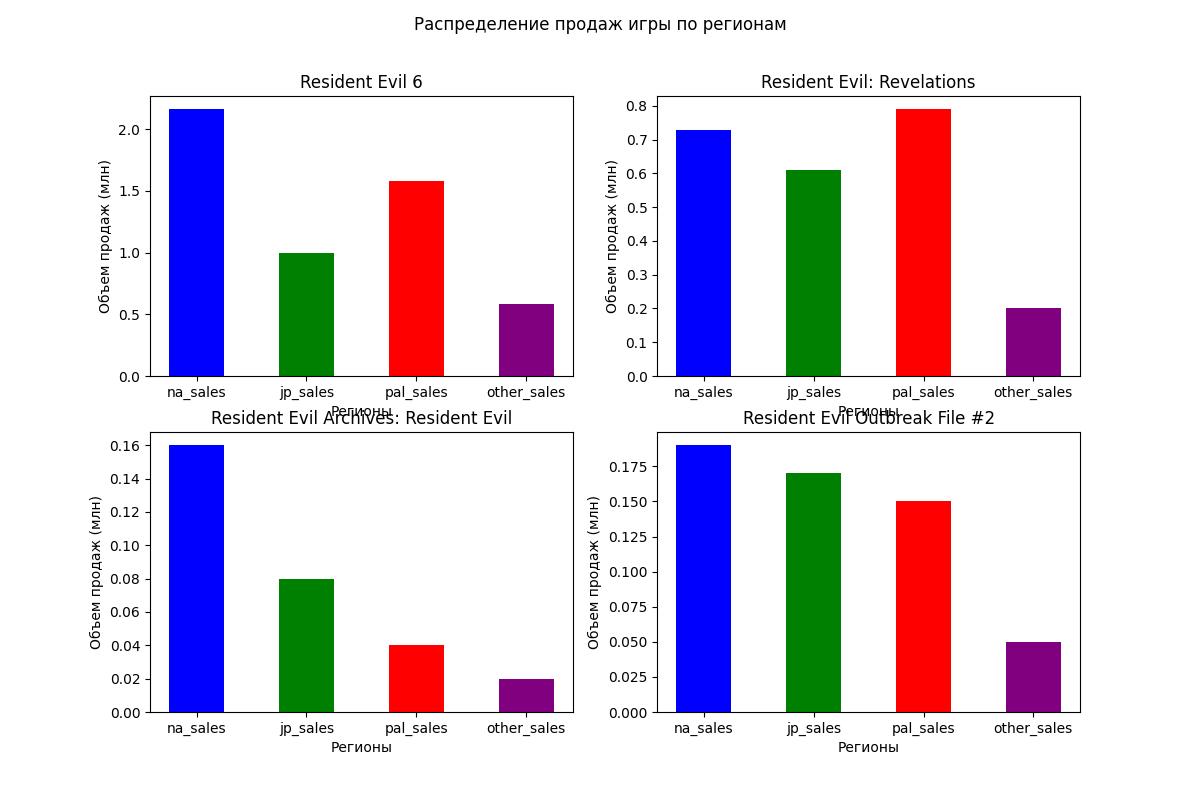
Каждое из приведенных выше задач мы реализовывали в виде python функции и запускали в Jupyter Notebook

**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. \*//

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

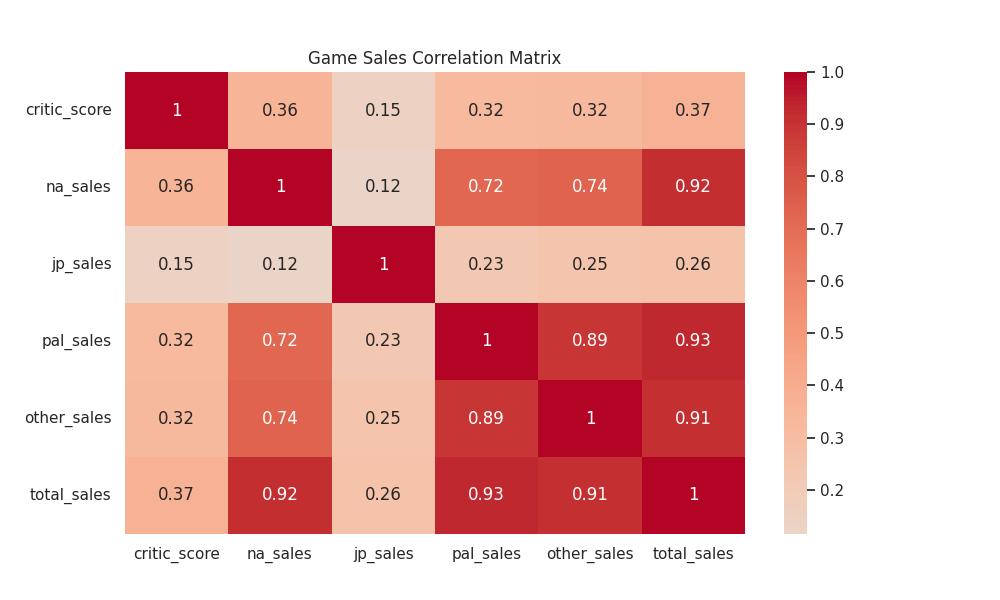
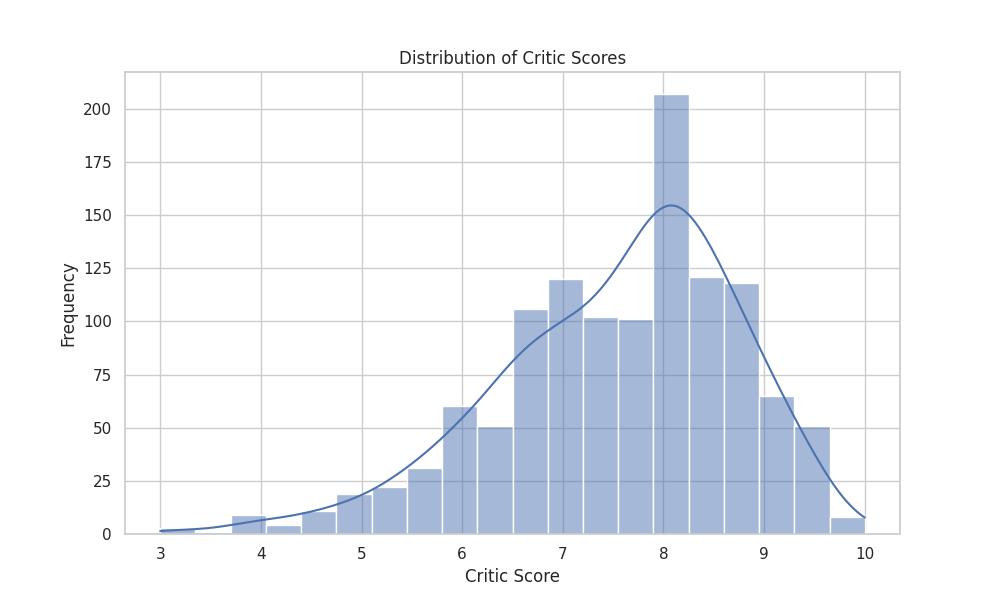


Рис. 4 - график распределения объёма продаж по оценкам критиков



## **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.

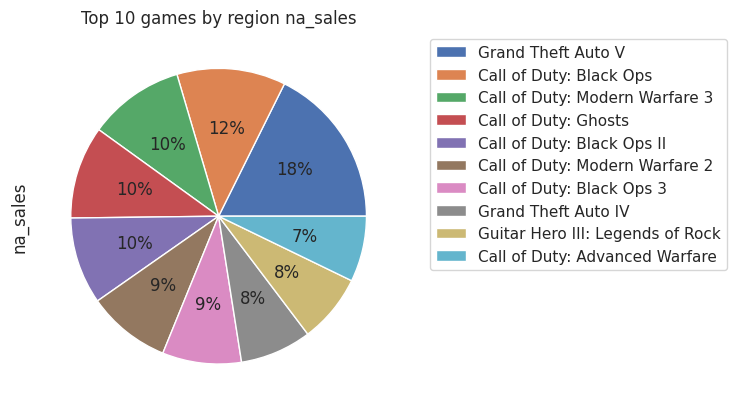


Рис. 5 – диаграмма самых популярных игр по региону na\_sales

## **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.

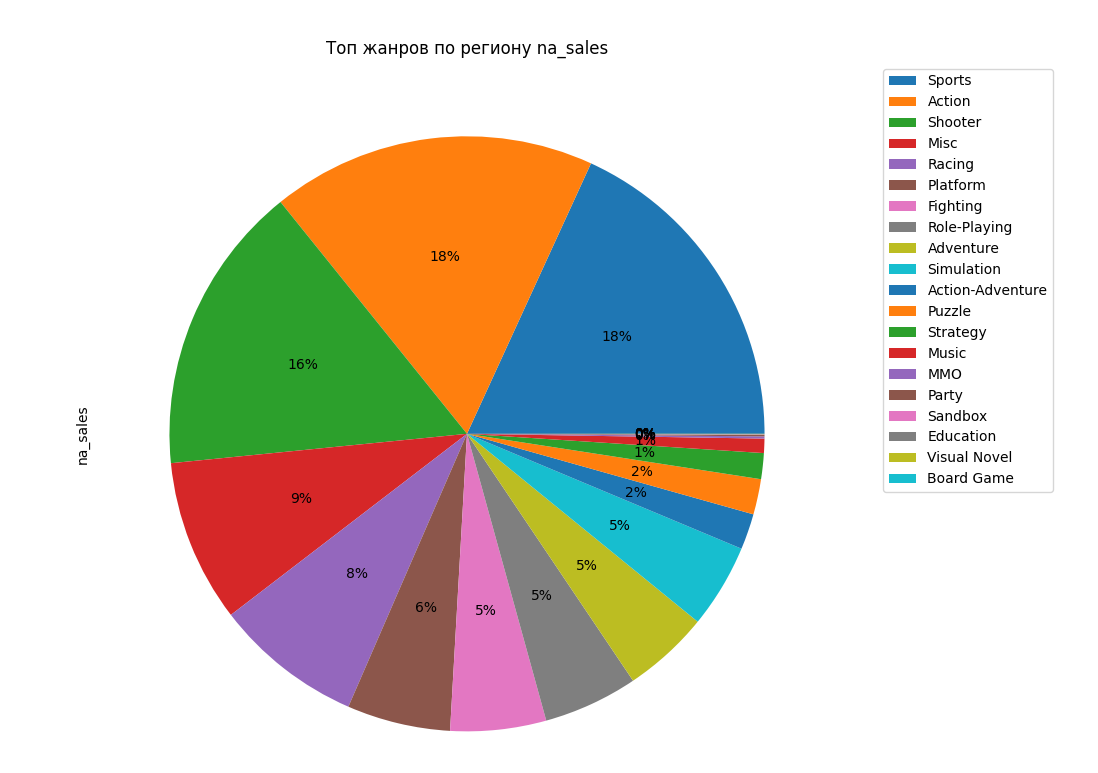


Рис. 6 - диаграмма, характеризующая популярность жанров игр по результатам продаж.

## **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.

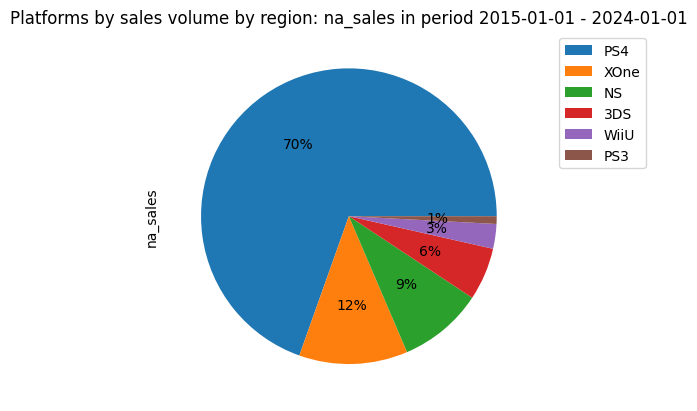


Рис. 7 - диаграмма, характеризующая долю платформы в том или ином регионе.

## **Задача 6. Предсказание продаж игр серии**

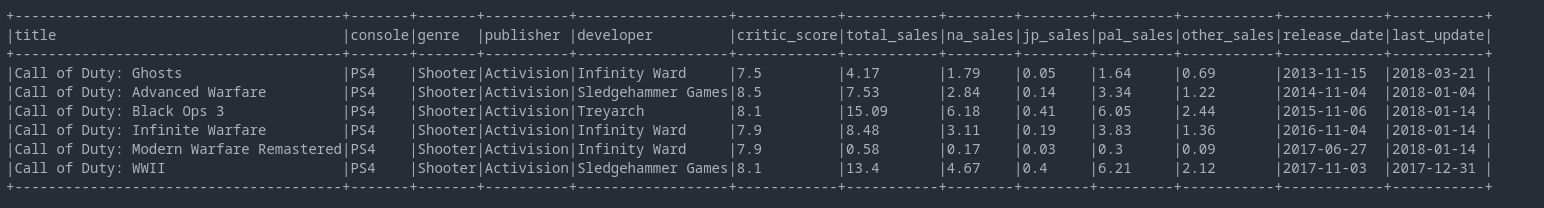
В данном задании мы предсказывали успех будущих игр серии. Сначала, в функции predict\_game\_success, мы фильтруем данные DataFrame по названию серии игр game\_series\_name и выбираем необходимые колонки, преобразуя типы данных и заполняя пропуски.

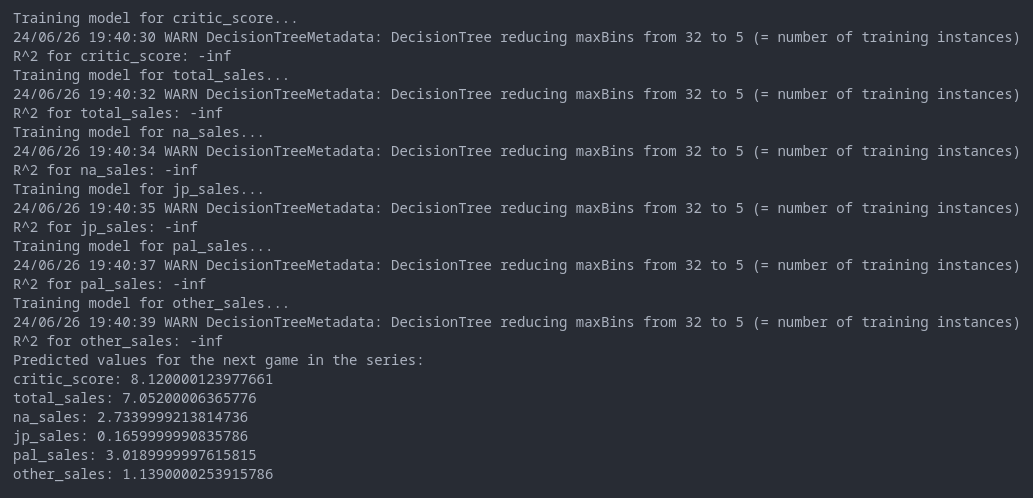
Далее мы выделяем региональные продажи и оценки критиков в отдельные признаки, используя VectorAssembler. Затем разделяем данные на обучающую и тестовую выборки.

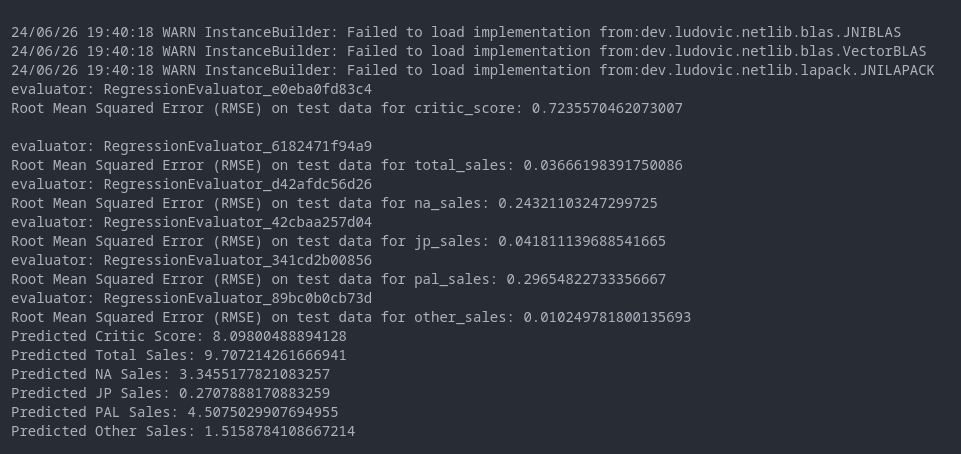
Для обучения модели используется линейная регрессия (LinearRegression), которая обучается на обучающей выборке и оценивается на тестовой выборке, вычисляя корень средней квадратичной ошибки (RMSE).

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.

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

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

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.

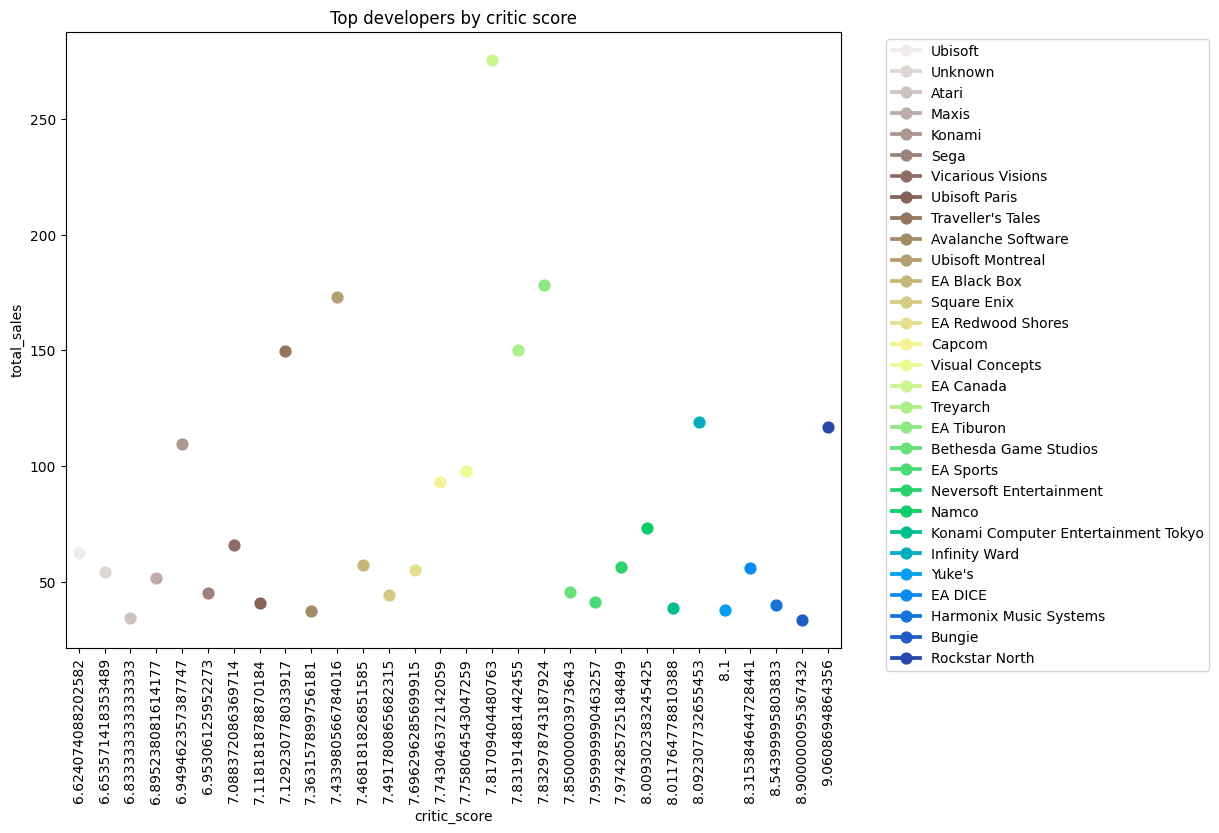


Рис. 11 - график распределения разработчиков по оценкам критиков и объему продаж.

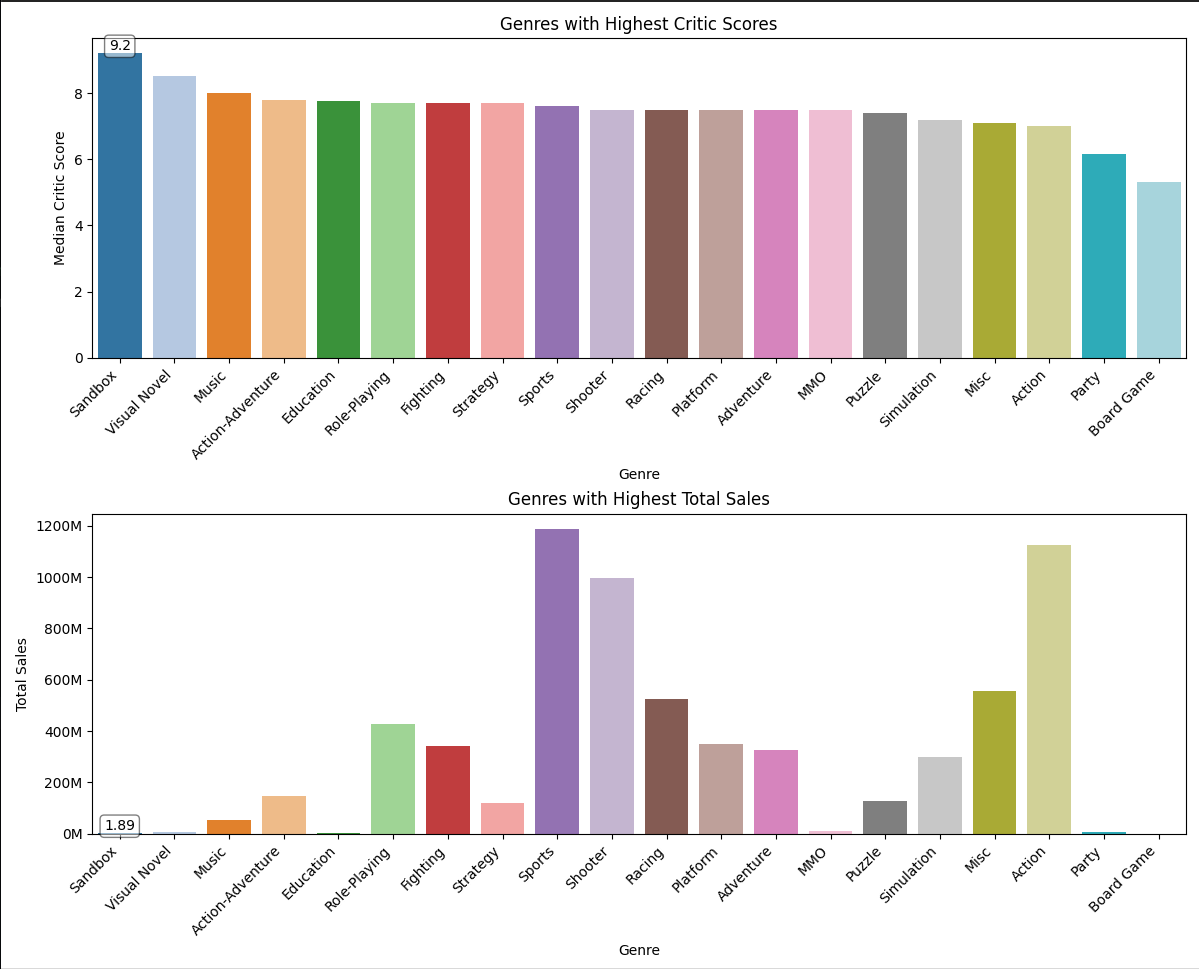
## **Задача 8. Топ жанров по оценкам критиков и общим продажам**

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**

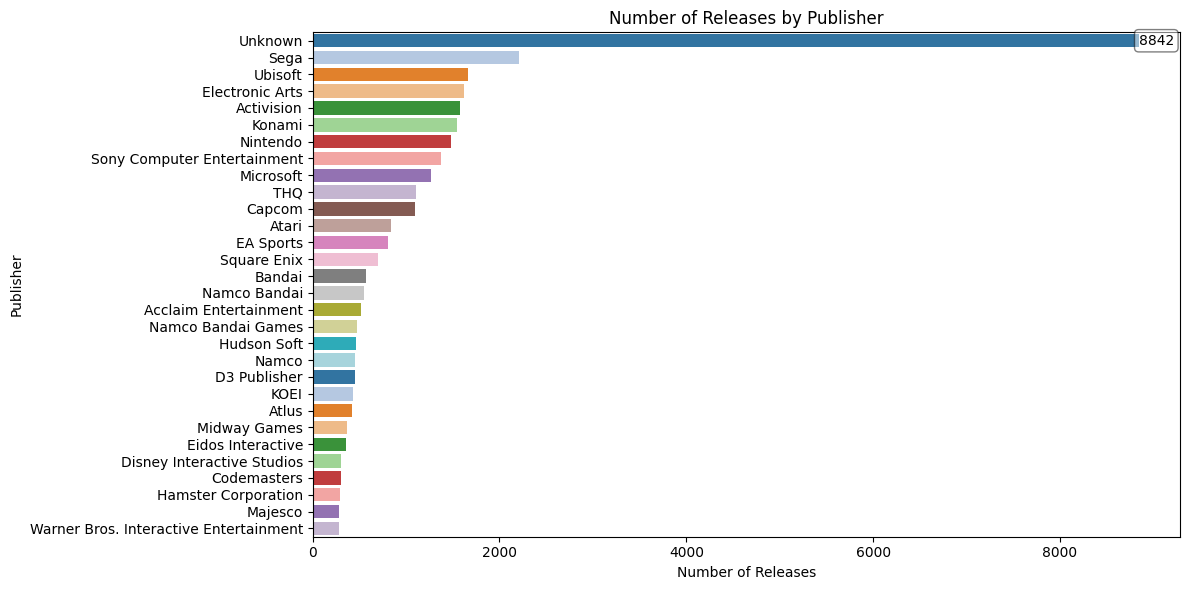
В данном задании мы выявляли топ издателей по количеству выпущенных тайтлов и объему продаж. Сначала, в функции get\_top\_sales\_performance\_by\_publisher, мы группируем данные DataFrame df по колонке publisher, суммируя общие продажи (total\_sales) и подсчитывая количество выпущенных тайтлов (title). Затем данные сортируются по убыванию общих продаж и преобразуются в формат Pandas DataFrame.

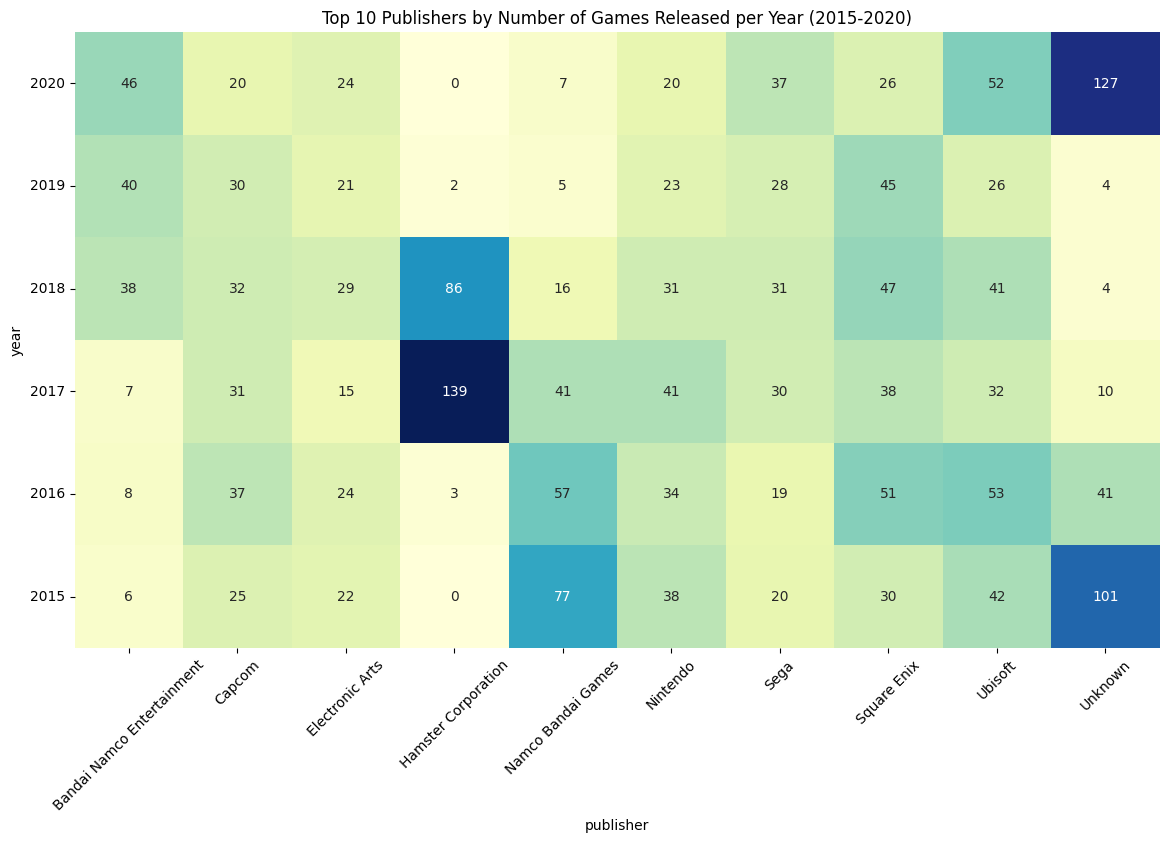
Из полученных данных выделяются топ-10 издателей по общим продажам и топ-10 издателей по количеству выпущенных тайтлов.

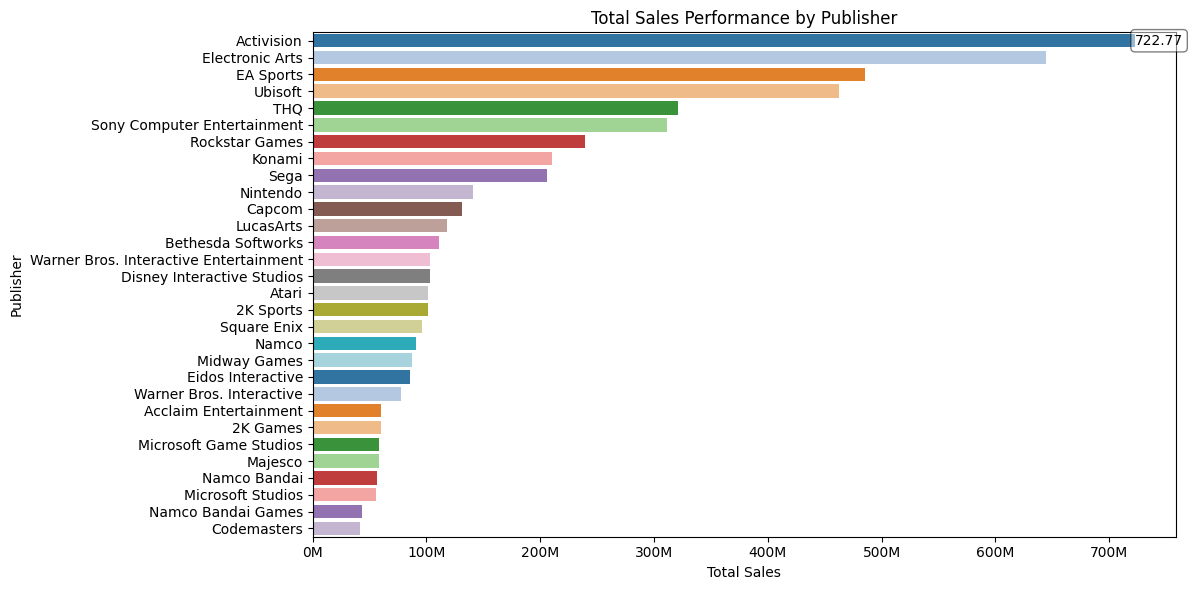
Далее данные визуализируются с помощью Seaborn и matplotlib. Создаются два графика:

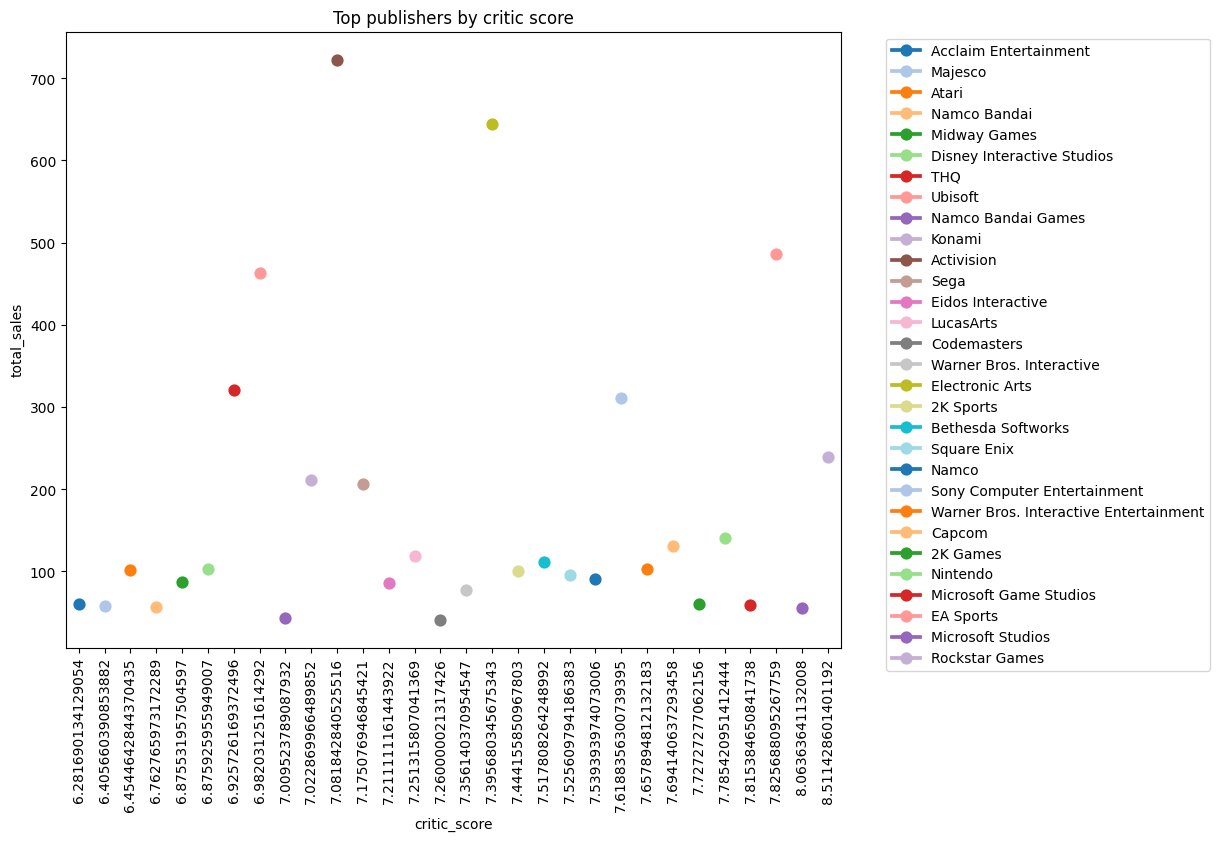
1. На первом графике отображается количество выпущенных тайтлов для каждого издателя. На оси X отображается количество тайтлов, а на оси Y — названия издателей. Столбцы графика окрашены в разные цвета, и на них нанесены метки значений.

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. Таким образом, вторая модель является более предпочтительной для нашего кейса.
5. **Гипотеза о влиянии разработчиков и издателей**: Данная гипотеза частично принимается и отвергается. По рис. 10, 15 можно увидеть, что некоторые издатели и разработчики, имеющие высокие продажи, имеют относительно низкие показатели средних оценок критиков по всем выпущенным тайтлам. Следовательно, хоть они и не имеют высоких оценок, их бренд оказывает влияние на успех игры.