Сборный проект №1

Описание проекта

Вы работаете в интернет-магазине «Стримчик», который продаёт по всему миру компьютерные игры. Из открытых источников доступны исторические данные о продажах игр, оценки пользователей и экспертов, жанры и платформы (например, Xbox или PlayStation). Вам нужно выявить определяющие успешность игры закономерности. Это позволит сделать ставку на потенциально популярный продукт и спланировать рекламные кампании. Перед вами данные до 2016 года. Представим, что сейчас декабрь 2016 г., и вы планируете кампанию на 2017-й. Нужно отработать принцип работы с данными. Неважно, прогнозируете ли вы продажи на 2017 год по данным 2016-го или же 2027-й — по данным 2026 года. В наборе данных попадается аббревиатура ESRB (Entertainment Software Rating Board) — это ассоциация, определяющая возрастной рейтинг компьютерных игр. ESRB оценивает игровой контент и присваивает ему подходящую возрастную категорию, например, «Для взрослых», «Для детей младшего возраста» или «Для подростков».

Инструкция по выполнению проекта

Шаг 1. Откройте файл с данными и изучите общую информацию

Путь к файлу: /datasets/games.csv. Скачать датасет

Шаг 2. Подготовьте данные

- Замените названия столбцов (приведите к нижнему регистру);
- Преобразуйте данные в нужные типы. Опишите, в каких столбцах заменили тип данных и почему;
- Обработайте пропуски при необходимости:
  - Объясните, почему заполнили пропуски определённым образом или почему не стали это делать;
  - Опишите причины, которые могли привести к пропускам;
  - Обратите внимание на аббревиатуру 'tbd' в столбце с оценкой пользователей. Отдельно разберите это значение и опишите, как его обработать;
- Посчитайте суммарные продажи во всех регионах и запишите их в отдельный столбец.

Шаг 3. Проведите исследовательский анализ данных

- Посмотрите, сколько игр выпускалось в разные годы. Важны ли данные за все периоды?
- Посмотрите, как менялись продажи по платформам. Выберите платформы с наибольшими суммарными продажами и постройте распределение по годам. За какой характерный срок появляются новые и исчезают старые платформы?
- Возьмите данные за соответствующий актуальный период. Актуальный период определите самостоятельно в результате исследования предыдущих вопросов. Основной фактор — эти данные помогут построить прогноз на 2017 год.
- Не учитывайте в работе данные за предыдущие годы.

- Какие платформы лидируют по продажам, растут или падают? Выберите несколько потенциально прибыльных платформ.
- Постройте график «ящик с усами» по глобальным продажам игр в разбивке по платформам. Опишите результат.
- Посмотрите, как влияют на продажи внутри одной популярной платформы отзывы пользователей и критиков. Постройте диаграмму рассеяния и посчитайте корреляцию между отзывами и продажами. Сформулируйте выводы.
- Соотнесите выводы с продажами игр на других платформах.
- Посмотрите на общее распределение игр по жанрам. Что можно сказать о самых прибыльных жанрах? Выделяются ли жанры с высокими и низкими продажами?

Шаг 4. Составьте портрет пользователя каждого региона

Определите для пользователя каждого региона (NA, EU, JP):

- Самые популярные платформы (топ-5). Опишите различия в долях продаж.
- Самые популярные жанры (топ-5). Поясните разницу.
- Влияет ли рейтинг ESRB на продажи в отдельном регионе?

Шаг 5. Проверьте гипотезы

- Средние пользовательские рейтинги платформ Xbox One и PC одинаковые;
- Средние пользовательские рейтинги жанров Action (англ. «действие», экшен-игры) и Sports (англ. «спортивные соревнования») разные.

  Задайте самостоятельно пороговое значение alpha. Поясните:

- Как вы сформулировали нулевую и альтернативную гипотезы;
- Какой критерий применили для проверки гипотез и почему.

Шаг 6. Напишите общий вывод

Описание данных

`Name` — название игры

`Platform` — платформа

`Year_of_Release` — год выпуска

`Genre` — жанр игры

`NA_sales` — продажи в Северной Америке (миллионы проданных копий)

`EU_sales` — продажи в Европе (миллионы проданных копий)

`JP_sales` — продажи в Японии (миллионы проданных копий)

`Other_sales` — продажи в других странах (миллионы проданных копий)

`Critic_Score` — оценка критиков (максимум 100)

`User_Score` — оценка пользователей (максимум 10)

`Rating` — рейтинг от организации ESRB (англ. Entertainment Software Rating Board). Эта ассоциация определяет рейтинг компьютерных игр и присваивает им подходящую возрастную категорию.

Данные за 2016 год могут быть неполными.

Шаг . Загрузка данных и знакомство с ними

In [96]:
```
import pandas as pd # импорт библиотек
import numpy as np
import scipy
from scipy import stats as st
import seaborn as sns
from matplotlib import pyplot as plt
import plotly.graph_objects as go
```

In [97]:
```
df = pd.read_csv('games.csv', sep=',')# загрузка данных
pd.set_option('display.max_columns', None)
df.head(15)
```

Out[97]:

	Name	Platform	Year_of_Release	Genre	NA_sales	EU_sales	JP_sales	Other_sales	Critic_Score	U
0	Wii Sports	Wii	2006.0	Sports	41.36	28.96	3.77	8.45	76.0	
1	Super Mario Bros.	NES	1985.0	Platform	29.08	3.58	6.81	0.77	NaN	
2	Mario Kart Wii	Wii	2008.0	Racing	15.68	12.76	3.79	3.29	82.0	
3	Wii Sports Resort	Wii	2009.0	Sports	15.61	10.93	3.28	2.95	80.0	
4	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	11.27	8.89	10.22	1.00	NaN	
5	Tetris	GB	1989.0	Puzzle	23.20	2.26	4.22	0.58	NaN	
6	New Super Mario Bros.	DS	2006.0	Platform	11.28	9.14	6.50	2.88	89.0	
7	Wii Play	Wii	2006.0	Misc	13.96	9.18	2.93	2.84	58.0	
8	New Super Mario Bros. Wii	Wii	2009.0	Platform	14.44	6.94	4.70	2.24	87.0	
9	Duck Hunt	NES	1984.0	Shooter	26.93	0.63	0.28	0.47	NaN	
10	Nintendogs	DS	2005.0	Simulation	9.05	10.95	1.93	2.74	NaN	
11	Mario Kart DS	DS	2005.0	Racing	9.71	7.47	4.13	1.90	91.0	
12	Pokemon Gold/Pokemon Silver	GB	1999.0	Role-Playing	9.00	6.18	7.20	0.71	NaN	
13	Wii Fit	Wii	2007.0	Sports	8.92	8.03	3.60	2.15	80.0	
14	Kinect Adventures!	X360	2010.0	Misc	15.00	4.89	0.24	1.69	61.0	

In [98]:
```
df.shape # проверка размерности
```

Out[98]:
```
(16715, 11)
```

In [99]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16715 entries, 0 to 16714
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Name             16713 non-null  object 
 1   Platform         16715 non-null  object 
 2   Year_of_Release  16446 non-null  float64
 3   Genre            16713 non-null  object 
 4   NA_sales         16715 non-null  float64
 5   EU_sales         16715 non-null  float64
 6   JP_sales         16715 non-null  float64
 7   Other_sales      16715 non-null  float64
 8   Critic_Score     8137 non-null   float64
 9   User_Score       10014 non-null  object 
 10  Rating           9949 non-null   object 
dtypes: float64(6), object(5)
memory usage: 1.4+ MB
```

In [100…
```
df.describe()# Описательные статистики для числовых признаков
```

Out[100…

	Year_of_Release	NA_sales	EU_sales	JP_sales	Other_sales	Critic_Score
count	16446.000000	16715.000000	16715.000000	16715.000000	16715.000000	8137.000000
mean	2006.484616	0.263377	0.145060	0.077617	0.047342	68.967679
std	5.877050	0.813604	0.503339	0.308853	0.186731	13.938165
min	1980.000000	0.000000	0.000000	0.000000	0.000000	13.000000
25%	2003.000000	0.000000	0.000000	0.000000	0.000000	60.000000
50%	2007.000000	0.080000	0.020000	0.000000	0.010000	71.000000
75%	2010.000000	0.240000	0.110000	0.040000	0.030000	79.000000
max	2016.000000	41.360000	28.960000	10.220000	10.570000	98.000000

Выводы по шагу 1

1. Таблица с данными состоит из 11 столбцов и 16715 строк.Признаки представлены следующими типами : float64(6),object(5).Содержит данные за период с 1980 по 2016 год.

2. По результатам проведенного первичного обследования данных можно сделать выводы по признакам и поставить задачи для подготовки данных:

Признак	Описание признака	Замечание,что возможно нужно сделать
Name	Название Игры	Переименовать.Убрать пустые значения
Platform	платформа	Переименовать
Year_of_Release	год выпуска	Переименовать.Изменить тип данных.Убрать пустые значения?
Genre	жанр игры	Переименовать.Убрать пустые значения?
NA_sales	продажи в Сев. Америке	Переименовать
EU_sales	продажи в Европе	Переименовать

Признак	Описание признака	Замечание,что возможно нужно сделать
JP_sales	продажи в Японии	Переименовать
Other_sales	продажи в других странах	Переименовать
Critic_Score	оценка критиков	Переименовать.Заменить пустые значения?
User_Score	оценка пользователей	Переименовать.Заменить пустые значения?Перевести в тип float64,заменить значение рейтинга 'tbd' на NAN
Rating	рейтинг	Переименовать.Заменить пустые значения?

3.Стоит заметить,что у признаков **NA_sales,EU_sales,JP_sales** есть минимальное значение -- 0. Значит в выборке имеются игры с числом проданных копий '0', т.е. до конечного потребителя эти игры так не дошли или еще дошли (напомним, что по условиям проекта сейчас 2016 год) либо были невостребованны.Можно ожидать, что рейтинги этих игр будут либо равны 0, либо отсутствовать

Шаг . Подготовка данных

Приведем названия столбцов к нижнему регистру

In [101…
```
df.columns = df.columns.str.lower()
df.columns
```

Out[101…
```
Index(['name', 'platform', 'year_of_release', 'genre', 'na_sales', 'eu_sales',
       'jp_sales', 'other_sales', 'critic_score', 'user_score', 'rating'],
      dtype='object')
```

Преобразование типов данных в признаках `year_of_release` и `User_Score` **

Значение признака `year_of_release` прведем к целочисленному, так как это логично:значение года - это целое число.

In [102…
```
df['year_of_release'] = df['year_of_release'].astype('Int64') # Приведени к целочисленному типу
```

In [103…
```
df['year_of_release'].dtype
```

Out[103…
```
Int64Dtype()
```

Посмотрим на значения признака `User_Score` (оценка пользователей)

In [104…
```
df['user_score'].value_counts()
```

Out[104…
```
tbd    2424
7.8     324
8       290
8.2     282
8.3     254
       ...
1.9       2
0.6       2
0.7       2
0         1
9.7       1
Name: user_score, Length: 96, dtype: int64
```

Видно,что в оценках пользователей,которые выражены в основном вещественно,есть некое значение `tbd` (аббревиатура от английского To Be Determined (будет определено) или To Be Decided (будет

решено). Используется, если какая-то информация еще не определена или решение по вопросу не принято. Акроним служит для отметки неясностей или пропусков, которые надо заполнить, в информации требований.)Значение `tbd` признака `user_score` следует заменить на значение `NaN`, как на более удобное для обработки в Pandas.А значения самого признака `user_score` следует сделать вещественными.

In [105…
```
df['user_score'] = df['user_score'].replace('tbd', np.nan, regex=True)
```

In [106…
```
df['user_score'] = df['user_score'].astype(float)
df['user_score'].dtype
```

Out[106…
```
dtype('float64')
```

In [107…
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16715 entries, 0 to 16714
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   name             16713 non-null  object 
 1   platform         16715 non-null  object 
 2   year_of_release  16446 non-null  Int64  
 3   genre            16713 non-null  object 
 4   na_sales         16715 non-null  float64
 5   eu_sales         16715 non-null  float64
 6   jp_sales         16715 non-null  float64
 7   other_sales      16715 non-null  float64
 8   critic_score     8137 non-null   float64
 9   user_score       7590 non-null   float64
 10  rating           9949 non-null   object 
dtypes: Int64(1), float64(6), object(4)
memory usage: 1.4+ MB
```

Обработка пропусков

Перед обработкой пропусков проведем проверку на явные дубликаты

In [108…
```
df.duplicated().sum()
```

Out[108…
```
0
```

Явных дубликатов не обнаружено.Посмотрим на процент пропусков.

In [109…
```
df.isnull().mean()
```

Out[109…
```
name               0.000120
platform           0.000000
year_of_release    0.016093
genre              0.000120
na_sales           0.000000
eu_sales           0.000000
jp_sales           0.000000
other_sales        0.000000
critic_score       0.513192
user_score         0.545917
rating             0.404786
dtype: float64
```

Обработка пропущеных значений признаков `name` и `genre`

In [110…
```
df[['name','genre']].isna().sum()
```

Out[110…
```
name     2
genre    2
dtype: int64
```

In [111…
```
df[df.name.isna()].index
```

Out[111…
```
Int64Index([659, 14244], dtype='int64')
```

In [112…
```
df[df.genre.isna()].index
```

Out[112…
```
Int64Index([659, 14244], dtype='int64')
```

Видно, что в этх признаках по 2 пропущенных значения,причем в одних и тех же строках.Их можно спокойно удалить.

In [113…
```
df.dropna(subset=['name'], inplace = True)
df.dropna(subset=['genre'], inplace = True)
df[['name', 'genre']].isna().sum()
```

Out[113…
```
name     0
genre    0
dtype: int64
```

Обработка пропущеных значений признака `year_of_release`

In [114…
```
df['year_of_release'].isnull().sum()
```

Out[114…
```
269
```

In [115…
```
df['year_of_release'].isnull().mean()
```

Out[115…
```
0.016095255190570215
```

Пропуски могли возникнуть по причине неправильного ввода данных(человеческий фактор) или технических проблем.Можно,например, заменить пропуски на значение года выпуска в зависимости от названия игры,но их всего 1.6% .Удалим их.

In [116…
```
df.dropna(subset=['year_of_release'], inplace = True)
```

In [117…
```
df['year_of_release'].isnull().sum()
```

Out[117…
```
0
```

Обработка пропущенных значений признаков `critic_score` , `user_score` и `rating`

In [118…
```
df[['critic_score', 'user_score', 'rating']].isna().sum()
```

Out[118…
```
critic_score    8461
user_score      8981
rating          6676
dtype: int64
```

Пропусков много и удалять их не стоит.Пропуски в оценках критиков и пользователей могут быть обусловлены тем, что не всегда эти оценки проставлялись или игра еще не успела получить оценку , так как была выпущена недавно и т.п.Оставим их как есть, так как адекватой замены предложить на мой взгляд невозможно. В признаке `rating` также пропущено много значений.И мы не можем знать точно на какое значение стоит заменить пропуски в этом случае.Заменим все пропуски в этом столбце

на 'unknown', т.е. рейтинг неизвестный.Причиной пропусков в рейтингах могло быть то, что возможно у игры не обозначен еще четко рейтинг или игра выпущена до того, как была введена система рейтингов (ESRB (Entertainment Software Rating Board) основана в 1996 году)https://ru.wikipedia.org/wiki/Entertainment_Software_Rating_Board Это система рейтинга для США и Канады.И пропуски в значении рейтинга могут значить, что игры выпускались не только для Северной Америки.

In [119…
```
df['rating'] = df['rating'].fillna('unknown')
```

Замечено также, что у нас есть игры с нулевыми продажами.

In [120…
```
df.query('(na_sales == 0) & (eu_sales == 0) & (jp_sales == 0) & (other_sales == 0)').index
```

Out[120…
```
Int64Index([16676, 16709], dtype='int64')
```

Таких записей две.Можно спокойно удалить

In [121…
```
df.drop(df.query('(na_sales == 0) & (eu_sales == 0) & (jp_sales == 0) & (other_sales == 0)').ir
```

In [122…
```
df.query('(na_sales == 0) & (eu_sales == 0) & (jp_sales == 0) & (other_sales == 0)').index
```

Out[122…
```
Int64Index([], dtype='int64')
```

Суммарные продажи во всех регионах

создадим новый столбец `total_sales` с суммарными продажами по всем регионам

In [123…
```
df['total_sales'] = df[['na_sales', 'eu_sales', 'jp_sales', 'other_sales']].sum(axis = 1)
df.head()
```

Out[123…

	name	platform	year_of_release	genre	na_sales	eu_sales	jp_sales	other_sales	critic_score	user_sc
0	Wii Sports	Wii	2006	Sports	41.36	28.96	3.77	8.45	76.0	
1	Super Mario Bros.	NES	1985	Platform	29.08	3.58	6.81	0.77	NaN	N
2	Mario Kart Wii	Wii	2008	Racing	15.68	12.76	3.79	3.29	82.0	
3	Wii Sports Resort	Wii	2009	Sports	15.61	10.93	3.28	2.95	80.0	
4	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	11.27	8.89	10.22	1.00	NaN	N

Выводы по шагу 2

На данном этапе проведены необходимые преобразования над признаками:

- названия столбцов приведены к нижнему регистру
- преобразованы типы данных в признаках `year_of_release` и `User_Score`
- удалены пропуски значений в признаках `name` и `genre`
- в признаке `year_of_release` пропущенные значения удалены
- пояснены возможные причины возникновения пропусков в рейтингах игр и заменены на значение 'unknown'

- в `critic_score` , `user_score` пропущенные значения осталены без изменений
- исправлена аббревиатура 'tbd' в столбце `user_score`
- посчитаны суммарные продажи во всех регионах, новые значения записаны в отдельный столбец `total_sales`
- удалены игры с нулевыми продажами.

In [124…

```
# Комментарий ревьюера
# Посмотрим, что у нас осталось
temp = df.copy()
list_c = ['name', 'platform', 'year_of_release', 'genre', 'critic_score', 'user_score', 'rating
print(temp.info())
for col_l in list_c:
  print('-'* 25)
  print(col_l, temp[col_l].sort_values().unique())
  print(col_l,': кол-во NaN',temp[col_l].isna().sum(),
        ', процент NaN', round(temp[col_l].isna().sum()/len(temp)*100, 2),'%')
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 16442 entries, 0 to 16714
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   name             16442 non-null  object 
 1   platform         16442 non-null  object 
 2   year_of_release  16442 non-null  Int64  
 3   genre            16442 non-null  object 
 4   na_sales         16442 non-null  float64
 5   eu_sales         16442 non-null  float64
 6   jp_sales         16442 non-null  float64
 7   other_sales      16442 non-null  float64
 8   critic_score     7983 non-null   float64
 9   user_score       7463 non-null   float64
 10  rating           16442 non-null  object 
 11  total_sales      16442 non-null  float64
dtypes: Int64(1), float64(7), object(4)
memory usage: 1.6+ MB
None
-------------------------
name [' Beyblade Burst' ' Fire Emblem Fates' " Frozen: Olaf's Quest" ...
 'uDraw Studio' 'uDraw Studio: Instant Artist'
 '¡Shin Chan Flipa en colores!']
name : кол-во NaN 0 , процент NaN 0.0 %
-------------------------
platform ['2600' '3DO' '3DS' 'DC' 'DS' 'GB' 'GBA' 'GC' 'GEN' 'GG' 'N64' 'NES' 'NG'
 'PC' 'PCFX' 'PS' 'PS2' 'PS3' 'PS4' 'PSP' 'PSV' 'SAT' 'SCD' 'SNES' 'TG16'
 'WS' 'Wii' 'WiiU' 'X360' 'XB' 'XOne']
platform : кол-во NaN 0 , процент NaN 0.0 %
-------------------------
year_of_release <IntegerArray>
[1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992,
 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005,
 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016]
Length: 37, dtype: Int64
year_of_release : кол-во NaN 0 , процент NaN 0.0 %
-------------------------
genre ['Action' 'Adventure' 'Fighting' 'Misc' 'Platform' 'Puzzle' 'Racing'
 'Role-Playing' 'Shooter' 'Simulation' 'Sports' 'Strategy']
genre : кол-во NaN 0 , процент NaN 0.0 %
-------------------------
critic_score [13. 17. 19. 20. 21. 23. 24. 25. 26. 27. 28. 29. 30. 31. 32. 33. 34. 35.
 36. 37. 38. 39. 40. 41. 42. 43. 44. 45. 46. 47. 48. 49. 50. 51. 52. 53.
 54. 55. 56. 57. 58. 59. 60. 61. 62. 63. 64. 65. 66. 67. 68. 69. 70. 71.
 72. 73. 74. 75. 76. 77. 78. 79. 80. 81. 82. 83. 84. 85. 86. 87. 88. 89.
 90. 91. 92. 93. 94. 95. 96. 97. 98. nan]
critic_score : кол-во NaN 8459 , процент NaN 51.45 %
-------------------------
user_score [0.  0.2 0.3 0.5 0.6 0.7 0.9 1.  1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.
 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3.  3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8
 3.9 4.  4.1 4.2 4.3 4.4 4.5 4.6 4.7 4.8 4.9 5.  5.1 5.2 5.3 5.4 5.5 5.6
 5.7 5.8 5.9 6.  6.1 6.2 6.3 6.4 6.5 6.6 6.7 6.8 6.9 7.  7.1 7.2 7.3 7.4
```

```
 7.5 7.6 7.7 7.8 7.9 8.  8.1 8.2 8.3 8.4 8.5 8.6 8.7 8.8 8.9 9.  9.1 9.2
 9.3 9.4 9.5 9.6 9.7 nan]
user_score : кол-во NaN 8979 , процент NaN 54.61 %
-------------------------
rating ['AO' 'E' 'E10+' 'EC' 'K-A' 'M' 'RP' 'T' 'unknown']
rating : кол-во NaN 0 , процент NaN 0.0 %
```

Шаг 3. Исследовательский анализ данных

Сколько игр выпускалось в разные годы. Важны ли данные за все периоды

In [125…

```
df1 = df.pivot_table(index='year_of_release', values='platform', aggfunc='count') # Создаем мат
df1
```

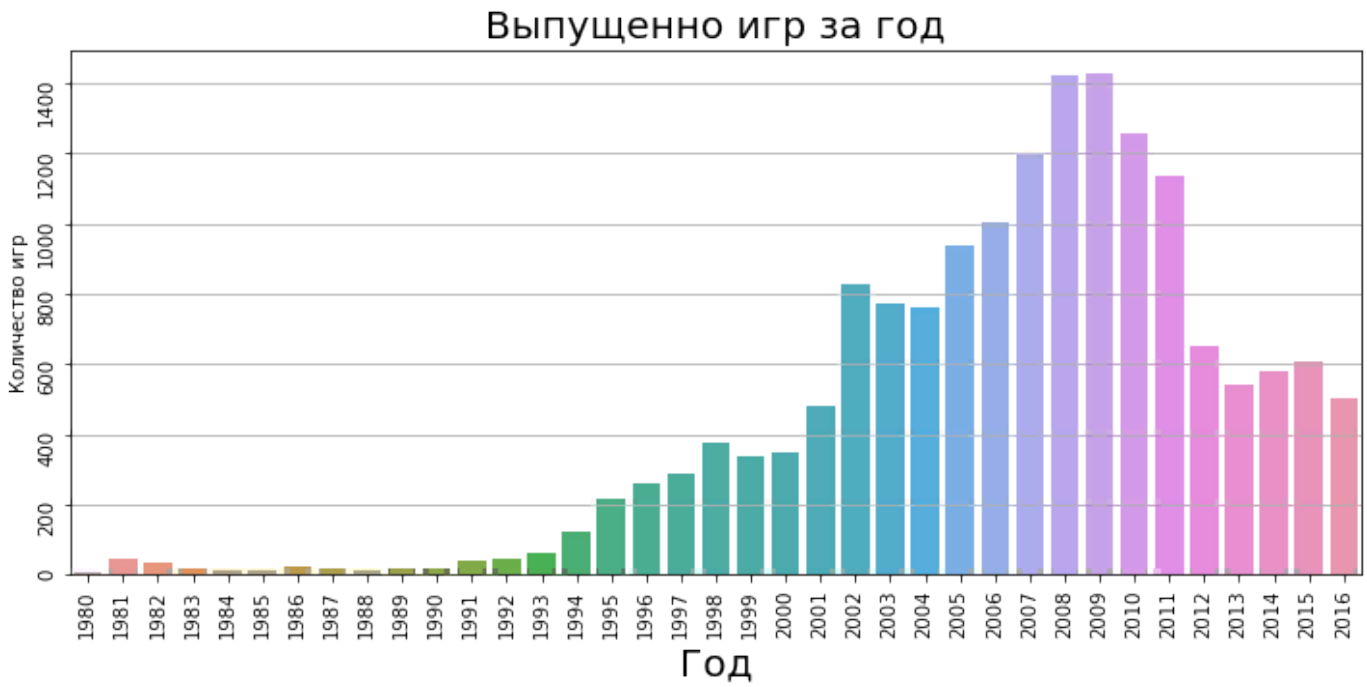
Out[125…

year_of_release	platform
1980	9
1981	46
1982	36
1983	17
1984	14
1985	14
1986	21
1987	16
1988	15
1989	17
1990	16
1991	41
1992	43
1993	60
1994	121
1995	219
1996	263
1997	289
1998	379
1999	338
2000	350
2001	482
2002	829
2003	775
2004	762
2005	939
2006	1006
2007	1197

year_of_release	platform
2008	1425
2009	1426
2010	1255
2011	1136
2012	653
2013	544
2014	581
2015	606
2016	502

In [126…

```
plt.figure(figsize=[12,5]) # Строим гистограмму
ax=sns.barplot(data=df1, x=df1.index, y="platform")
ax.axes.set_title('Выпущенно игр за год',fontsize=20)
ax.set_xlabel('Год',fontsize=20)
ax.set_ylabel('Количество игр',fontsize=10)
ax.tick_params(labelsize=10,rotation = 90)
ax.yaxis.grid(True)
```



Вывод по динамике выпуска:Видим, что нам представлены данные с 1980 по 2016 год.Начиная я 1980 и до 1900 года объемы относительно небольшие.Затем,начиная с 1990 до 2008 года, наблюдается рост выпуска компьютерных игр, а вот начиная с 2009 года объемы выпуска снижаются. Возможно это связано с ростом числа игр для телефонов. С 2012 года наблюдается определенная стабилизация объемов выпуска. Возможно именно за этот период данные будут полезными для дальнейшего анализа, особенно в краткосрочном периоде (по условию задания кампания планируется на 1 год).

Как менялись продажи по платформам? Выберем платформы с наибольшими суммарными продажами и построим распределение по годам. За какой характерный срок появляются новые и исчезают старые платформы?

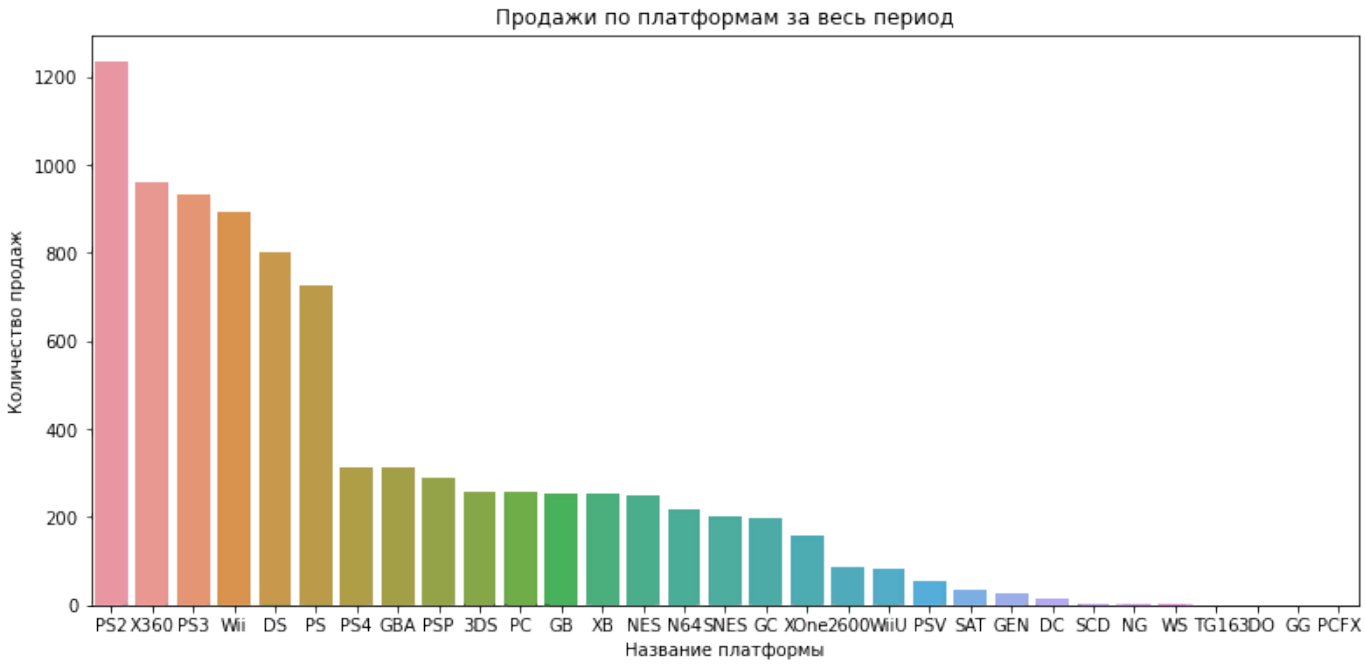
Популярные платформы за весь период наблюдений

In [127…

```
platform_on_sales = df.pivot_table(
    index='platform', values='total_sales', aggfunc='sum').sort_values(by='total_sales', ascend

plt.figure(figsize=(13,6))
sns.barplot(x=platform_on_sales.index,y=platform_on_sales['total_sales'])
plt.title("Продажи по платформам за весь период")
plt.xlabel("Название платформы")
plt.ylabel("Количество продаж")
```

Out[127…

```
Text(0, 0.5, 'Количество продаж')
```



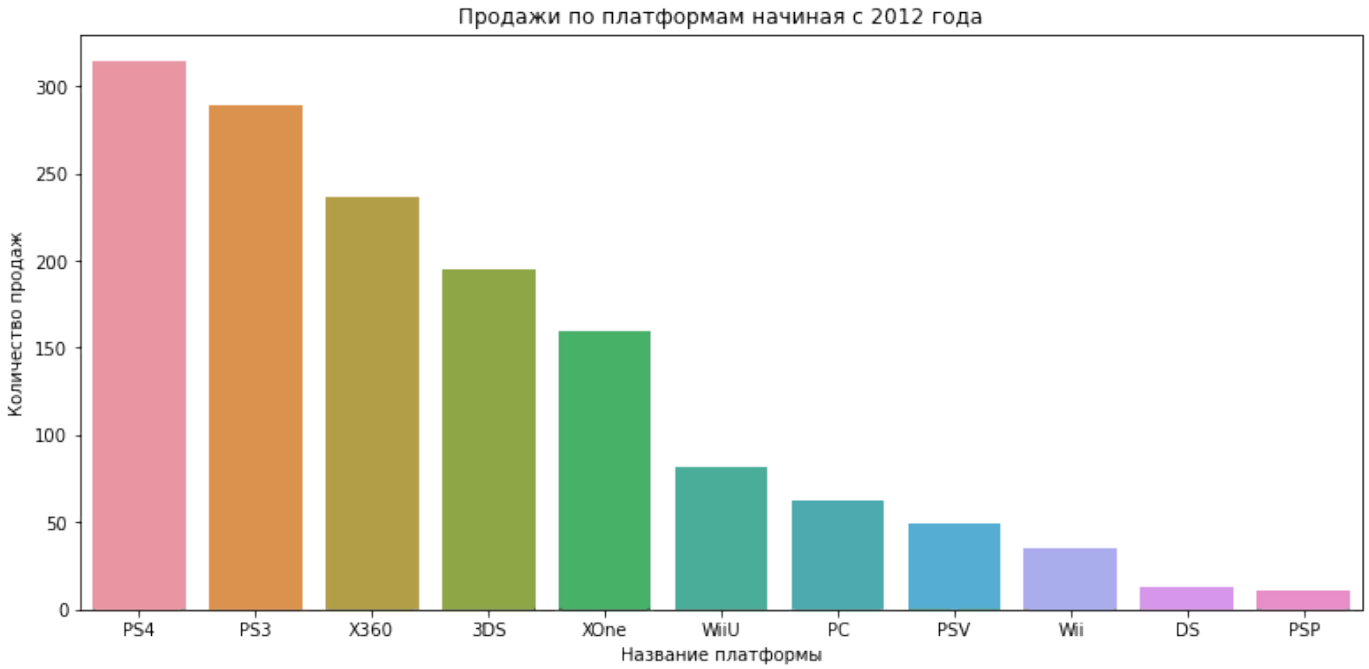
Здесь стоит отметить,что лидерами являются платформы 'PS2', 'X360', 'PS3', 'Wii', 'DS', 'PS' Посмотрим что происходит в периоде начиная с 2012 года

In [128…

```
platform_on_sales_2012 = df[df.year_of_release >= 2012].pivot_table(
    index='platform', values='total_sales', aggfunc='sum').sort_values(by='total_sales', ascend

plt.figure(figsize=(13,6))
sns.barplot(x=platform_on_sales_2012.index,y=platform_on_sales_2012['total_sales'])
plt.title("Продажи по платформам начиная с 2012 года")
plt.xlabel("Название платформы")
plt.ylabel("Количество продаж")
```

Out[128…

```
Text(0, 0.5, 'Количество продаж')
```



Видим, что со временем картина меняется . Среди лидеров уже 'PS4', 'PS3', 'X360', '3DS', 'XOne'.

Вывод по платформам с наибольшими суммарными продажами: Видим, что со временем меняются платформы-лидеры, что естественно.На смену 'PS2' приходит 'PS3', а затем 'PS4'. Платформа '3DS' вытесняет 'DS'. Среди "долгожителей" стоит отметить 'X360'.

Рассмотрим Топ-6 платформ с наибольшими суммарными продажами за весь период наблюдений

In [129…
```
top_6_from_all = platform_on_sales.reset_index().rename_axis(None, axis=1)[:6]
top_6_from_all
```

Out[129…

	platform	total_sales
0	PS2	1233.56
1	X360	961.24
2	PS3	931.34
3	Wii	891.18
4	DS	802.78
5	PS	727.58

In [130…
```
#Создадим переменную, хранящую список Топ6 продаж платформ
top_platforms = top_6_from_all['platform'].to_list()
top_platforms
```

Out[130… `['PS2', 'X360', 'PS3', 'Wii', 'DS', 'PS']`

In [131…
```
 df_top = df.query('platform in @top_platforms') # датасет с данными по  Топ-6 платформам
df_top
```

Out[131…

	name	platform	year_of_release	genre	na_sales	eu_sales	jp_sales	other_sales	critic_score	user_
0	Wii Sports	Wii	2006	Sports	41.36	28.96	3.77	8.45	76.0	
2	Mario Kart Wii	Wii	2008	Racing	15.68	12.76	3.79	3.29	82.0	
3	Wii Sports	Wii	2009	Sports	15.61	10.93	3.28	2.95	80.0	

	name	platform	year_of_release	genre	na_sales	eu_sales	jp_sales	other_sales	critic_score	user_
	Resort									
6	New Super Mario Bros.	DS	2006	Platform	11.28	9.14	6.50	2.88	89.0	
7	Wii Play	Wii	2006	Misc	13.96	9.18	2.93	2.84	58.0	
...	...	...	...	...	...	...	...	...	...	...
16698	Mega Brain Boost	DS	2008	Puzzle	0.01	0.00	0.00	0.00	48.0	
16700	Mezase!! Tsuri Master DS	DS	2009	Sports	0.00	0.00	0.01	0.00	NaN	
16704	Plushees	DS	2008	Simulation	0.01	0.00	0.00	0.00	NaN	
16710	Samurai Warriors: Sanada Maru	PS3	2016	Action	0.00	0.00	0.01	0.00	NaN	
16711	LMA Manager 2007	X360	2006	Sports	0.00	0.01	0.00	0.00	NaN	

9260 rows × 12 columns

Сгруппируем продади в зависимости от года и платформы

In [132...

```
top_6_all =df_top.pivot_table(index='year_of_release', columns='platform',
                              values='total_sales', aggfunc='sum')\
                             .sort_values('year_of_release', ascending=False).fillna(0)
top_6_all
```

Out[132...

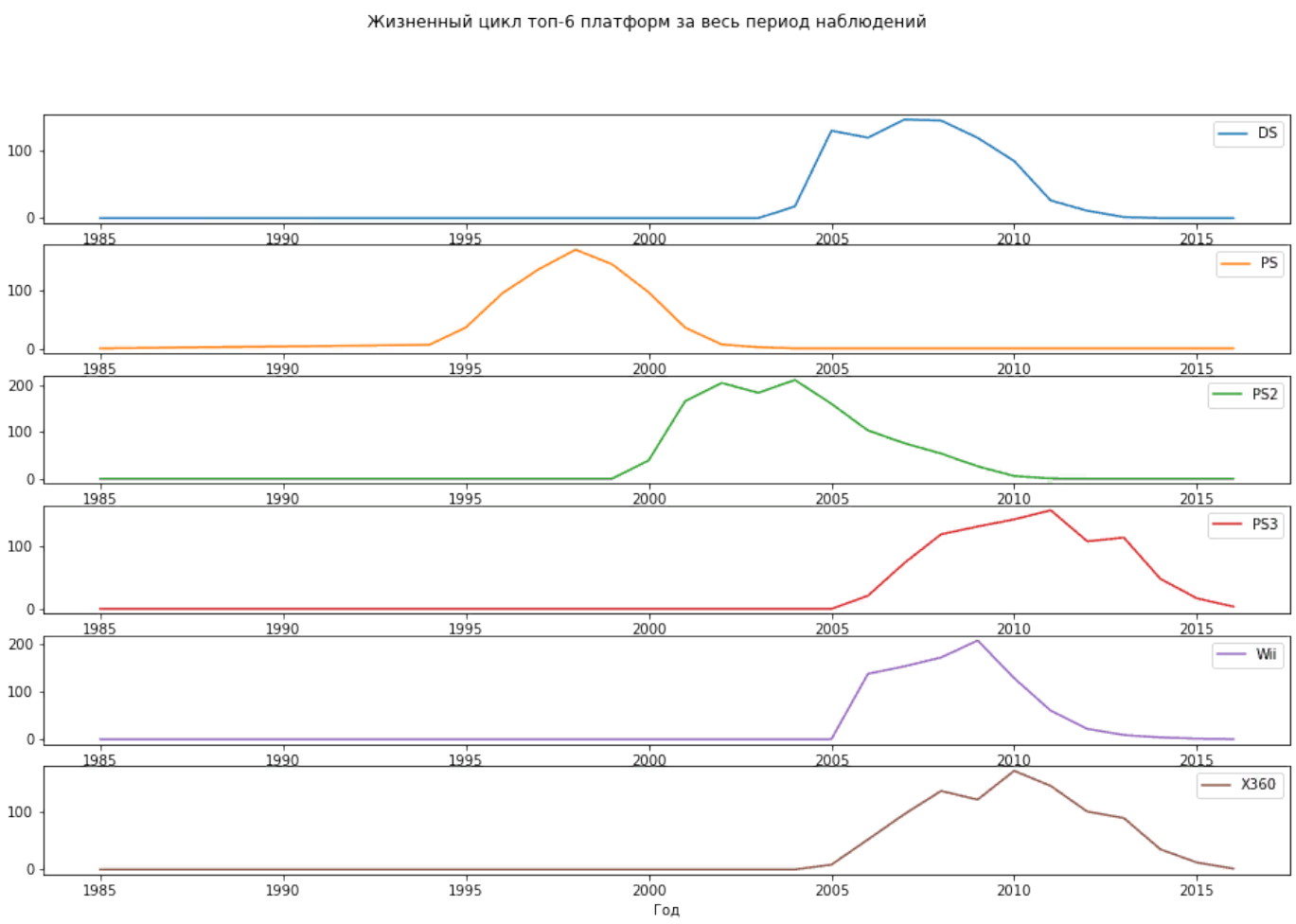
platform	DS	PS	PS2	PS3	Wii	X360
year_of_release						
2016	0.00	0.00	0.00	3.60	0.18	1.52
2015	0.00	0.00	0.00	16.82	1.14	11.96
2014	0.00	0.00	0.00	47.76	3.75	34.74
2013	1.54	0.00	0.00	113.25	8.59	88.58
2012	11.01	0.00	0.00	107.36	21.71	99.74
2011	26.18	0.00	0.45	156.78	59.65	143.84
2010	85.02	0.00	5.64	142.17	127.95	170.03
2009	119.54	0.00	26.40	130.93	206.97	120.29
2008	145.31	0.00	53.90	118.52	171.32	135.26
2007	146.94	0.00	75.99	73.19	152.77	95.41
2006	119.81	0.00	103.42	20.96	137.15	51.62
2005	130.14	0.00	160.66	0.00	0.00	8.25
2004	17.27	0.00	211.81	0.00	0.00	0.00

platform	DS	PS	PS2	PS3	Wii	X360
year_of_release						
2003	0.00	2.07	184.31	0.00	0.00	0.00
2002	0.00	6.67	205.38	0.00	0.00	0.00
2001	0.00	35.59	166.43	0.00	0.00	0.00
2000	0.00	96.37	39.17	0.00	0.00	0.00
1999	0.00	144.53	0.00	0.00	0.00	0.00
1998	0.00	169.49	0.00	0.00	0.00	0.00
1997	0.00	136.17	0.00	0.00	0.00	0.00
1996	0.00	94.70	0.00	0.00	0.00	0.00
1995	0.00	35.96	0.00	0.00	0.00	0.00
1994	0.00	6.03	0.00	0.00	0.00	0.00
1985	0.02	0.00	0.00	0.00	0.00	0.00

Посмотрим динамику продаж среди Топ-6 платформ за весь период наблюдения и оцениим жизненный цикл платформ

In [133…

```
top_6_all.plot(kind='line', subplots=True, figsize=(16,10), sharex = False, sharey = False,
               title = 'Жизненный цикл топ-6 платформ за весь период наблюдений');
plt.xlabel('Год')
plt.show();
```



In [134…

```
#Сколько лет "живет" Топ6 платформ
df_top_years = df_top.groupby(['platform', 'year_of_release']).agg({'total_sales':'sum'}).reset
df_top_years['platform'].value_counts()
```

Out[134…
```
PS2     12
X360    12
Wii     11
PS3     11
DS      11
PS      10
Name: platform, dtype: int64
```

In [135…
```
df_top_years['platform'].value_counts().mean()
```

Out[135…
```
11.166666666666666
```

ВЫВОД :

- В результате исследования выявлено, что платформами с наибольшими продажами за весь период наблюдений являются 'PS2', 'X360', 'PS3', 'Wii', 'DS', 'PS'
- Со временем лидеры меняются, так как на смену одним, приходят другие платформы
- Жизненный цикл платформ в среднем составляет 11 лет.Первые 5 лет(примерно) наблюдается рост, а затем идет падение.

Актуальный период

Ранее было замечено, что начиная с 2009 года, суммарные продажи компьютерных игр снижаются, а начиная с 2012 года, наблюдается определенная стабильность на рынке. В связи с этим именно **период с 2012** будем считать **актуальным**.Нам это поможет построить прогноз на 2017 год.

In [136…
```
df_actual = df[df.year_of_release >= 2012]
df_actual.head()
```

Out[136…

	name	platform	year_of_release	genre	na_sales	eu_sales	jp_sales	other_sales	critic_score	user_scor
16	Grand Theft Auto V	PS3	2013	Action	7.02	9.09	0.98	3.96	97.0	8.
23	Grand Theft Auto V	X360	2013	Action	9.66	5.14	0.06	1.41	97.0	8.
31	Call of Duty: Black Ops 3	PS4	2015	Shooter	6.03	5.86	0.36	2.38	NaN	Na
33	Pokemon X/Pokemon Y	3DS	2013	Role-Playing	5.28	4.19	4.35	0.78	NaN	Na
34	Call of Duty: Black Ops II	PS3	2012	Shooter	4.99	5.73	0.65	2.42	83.0	5.

In [137…
```
df_actual.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2886 entries, 16 to 16714
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   name             2886 non-null   object
 1   platform         2886 non-null   object
 2   year_of_release  2886 non-null   Int64
 3   genre            2886 non-null   object
 4   na_sales         2886 non-null   float64
```

```
 5   eu_sales         2886 non-null   float64
 6   jp_sales         2886 non-null   float64
 7   other_sales      2886 non-null   float64
 8   critic_score     1312 non-null   float64
 9   user_score       1531 non-null   float64
 10  rating           2886 non-null   object
 11  total_sales      2886 non-null   float64
dtypes: Int64(1), float64(7), object(4)
memory usage: 295.9+ KB
```

Какие платформы лидируют по продажам, растут или падают? Выберите несколько потенциально прибыльных платформ(актуальный период)

In [138...

```
platform_on_sales_2012  # Продажи по платформам за актуальный период.
```

Out[138...

platform	total_sales
PS4	314.14
PS3	288.79
X360	236.54
3DS	194.61
XOne	159.32
WiiU	82.19
PC	62.65
PSV	49.18
Wii	35.37
DS	12.55
PSP	11.19

Выше уже отмечалось, что в актуальном периоде среди лидеров уже 'PS4', 'PS3', 'X360', '3DS', 'XOne'. Их и выделим.

In [139...

```
top_5_from_2012 = platform_on_sales_2012.reset_index().rename_axis(None, axis=1)[:5]
top_5_from_2012
```

Out[139...

	platform	total_sales
0	PS4	314.14
1	PS3	288.79
2	X360	236.54
3	3DS	194.61
4	XOne	159.32

In [140...

```
top_platforms_act = top_5_from_2012['platform'].to_list()
top_platforms_act
```

Out[140...

```
['PS4', 'PS3', 'X360', '3DS', 'XOne']
```

In [141…

```
df_top_act = df[df.year_of_release >= 2012].query('platform in @top_platforms_act')
df_top_act
```

Out[141…

	name	platform	year_of_release	genre	na_sales	eu_sales	jp_sales	other_sales	critic_score	user_
16	Grand Theft Auto V	PS3	2013	Action	7.02	9.09	0.98	3.96	97.0	
23	Grand Theft Auto V	X360	2013	Action	9.66	5.14	0.06	1.41	97.0	
31	Call of Duty: Black Ops 3	PS4	2015	Shooter	6.03	5.86	0.36	2.38	NaN	
33	Pokemon X/Pokemon Y	3DS	2013	Role-Playing	5.28	4.19	4.35	0.78	NaN	
34	Call of Duty: Black Ops II	PS3	2012	Shooter	4.99	5.73	0.65	2.42	83.0	
…	…	…	…	…	…	…	…	…	…	
16672	Metal Gear Solid V: The Definitive Experience	XOne	2016	Action	0.01	0.00	0.00	0.00	NaN	
16674	Tsukigime Ranko's Longest Day	PS3	2014	Action	0.00	0.01	0.00	0.00	NaN	
16677	Aikatsu Stars! My Special Appeal	3DS	2016	Action	0.00	0.00	0.01	0.00	NaN	
16691	Dynasty Warriors: Eiketsuden	PS3	2016	Action	0.00	0.00	0.01	0.00	NaN	
16710	Samurai Warriors: Sanada Maru	PS3	2016	Action	0.00	0.00	0.01	0.00	NaN	

1820 rows × 12 columns

Сгруппируем также продажи в зависимости от года и платформы в актуальном периоде для выбранных 5 платформ и построим графики

In [142…

```
top2012_5 = df_top_act.pivot_table(index='year_of_release', columns='platform',
                                   values='total_sales', aggfunc='sum')\
                                  .sort_values('year_of_release', ascending=False).fillna(0)
top2012_5
```

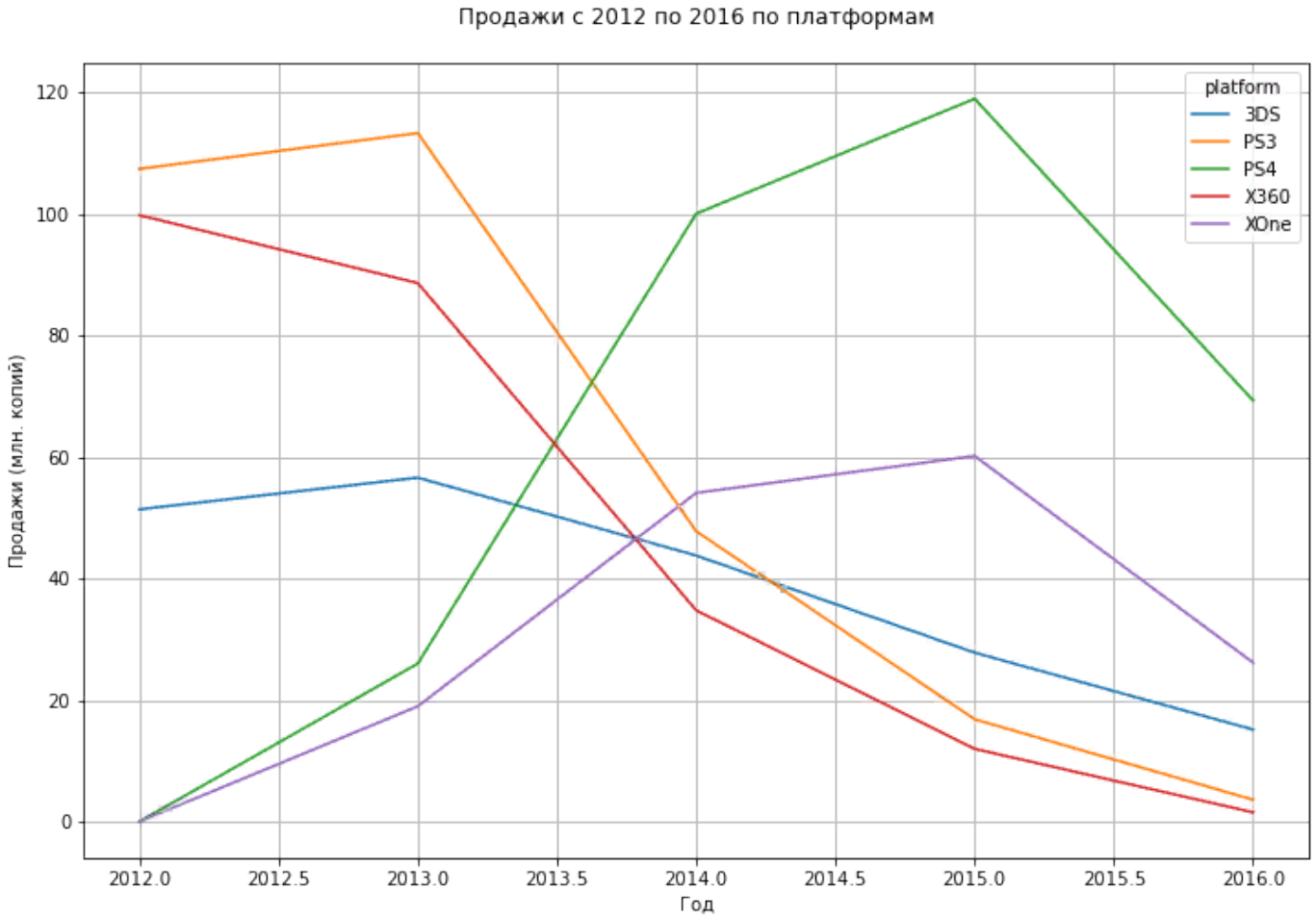
Out[142…

platform	3DS	PS3	PS4	X360	XOne
year_of_release					
2016	15.14	3.60	69.25	1.52	26.15

platform	3DS	PS3	PS4	X360	XOne
year_of_release					
2015	27.78	16.82	118.90	11.96	60.14
2014	43.76	47.76	100.00	34.74	54.07
2013	56.57	113.25	25.99	88.58	18.96
2012	51.36	107.36	0.00	99.74	0.00

In [143…

```
top2012_5.plot(figsize=(12,8))
plt.grid(True)
plt.title('Продажи с 2012 по 2016 по платформам\n  ')
plt.xlabel('Год')
plt.ylabel('Продажи (млн. копий)')
plt.show()
```



ВЫВОД:

- Продажи по всем платформам к 2016 году снижаются,хотя стоит заметить, что по условию данные за 2016 год неполные и на этом поэтому не стоит заострять внимание.
- Наибольшие продажи в актуальном периоде наблюдаются по платформам 'PS4', 'PS3', 'X360', '3DS', 'XOne'
- Наилучшие перспективы у платформ PS4 и XOne- они лидеры.
- У 'PS3' с 2013 года начинается снижение продаж,так как в это время ей на смену приходит 'PS4'
- Появляется новая платформа 'XOne' и ее продажи растут. По-прежнему популярны '3DS' и 'X360',хотя в этом случае прожажи падают.

«ящик с усами» по глобальным продажам игр в разбивке по платформам

Построим "Ящики с усами" по глобальным продажам игр в разбивке по платформам в актуальном периоде

In [144...

```
#plt.figure(figsize=(15,8))
#plt.ylim(0, 2)
#sns.boxplot(data=df_top_act, x='platform', y='total_sales');
#plt.title('Объем продаж по актуальным топ-5 платформам', fontsize=15)
#plt.xlabel('Платформа', fontsize=12)
#plt.ylabel('Объем продаж',fontsize=12)
#plt.show();
```

посмотрим на продажи по всем платформам и по годам

In [145...

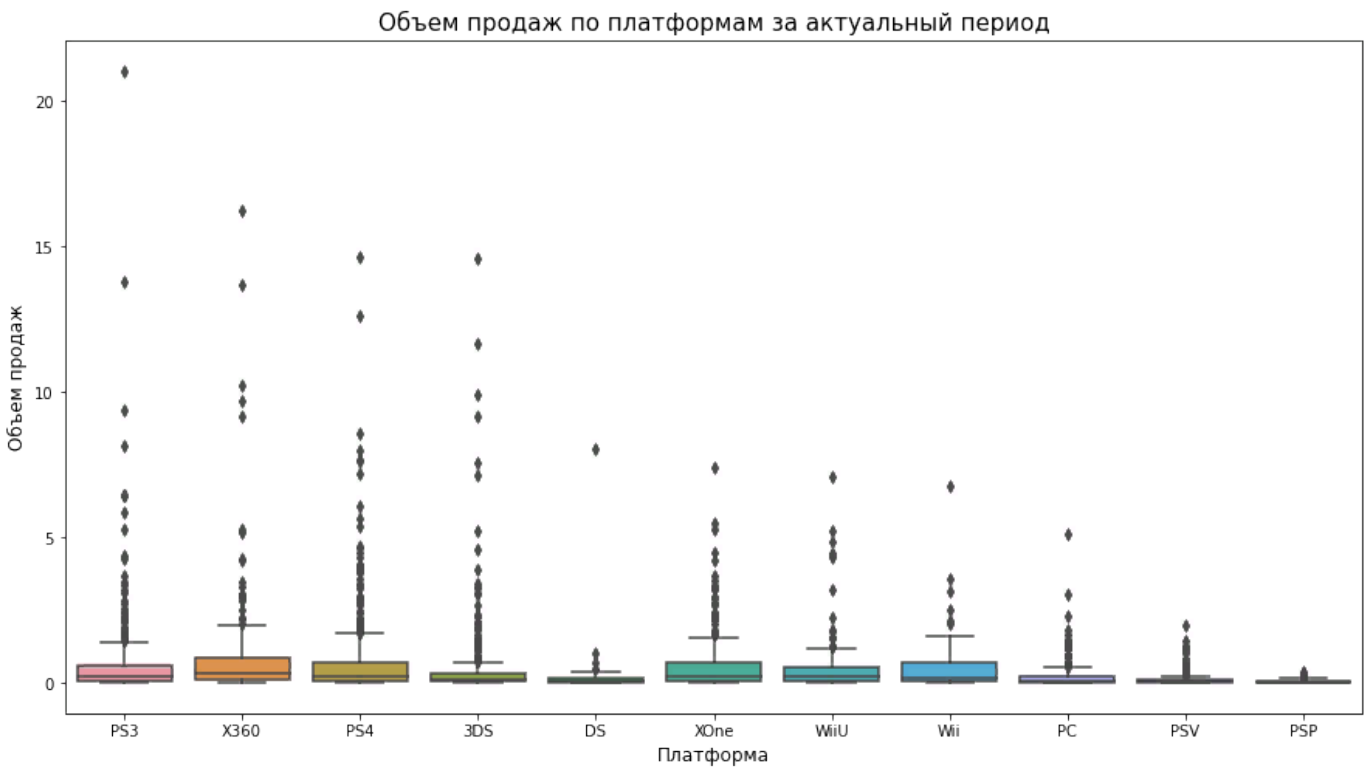
```
df_actual_plat = df_actual.pivot_table(index='year_of_release', columns='platform',
                                values='total_sales', aggfunc='sum')\
                               .sort_values('year_of_release', ascending=False).fillna(0)
df_actual_plat
```

Out[145...

platform	3DS	DS	PC	PS3	PS4	PSP	PSV	Wii	WiiU	X360	XOne
year_of_release											
2016	15.14	0.00	5.25	3.60	69.25	0.00	4.25	0.18	4.60	1.52	26.15
2015	27.78	0.00	8.52	16.82	118.90	0.12	6.25	1.14	16.35	11.96	60.14
2014	43.76	0.00	13.28	47.76	100.00	0.24	11.90	3.75	22.03	34.74	54.07
2013	56.57	1.54	12.38	113.25	25.99	3.14	10.59	8.59	21.65	88.58	18.96
2012	51.36	11.01	23.22	107.36	0.00	7.69	16.19	21.71	17.56	99.74	0.00

In [146...

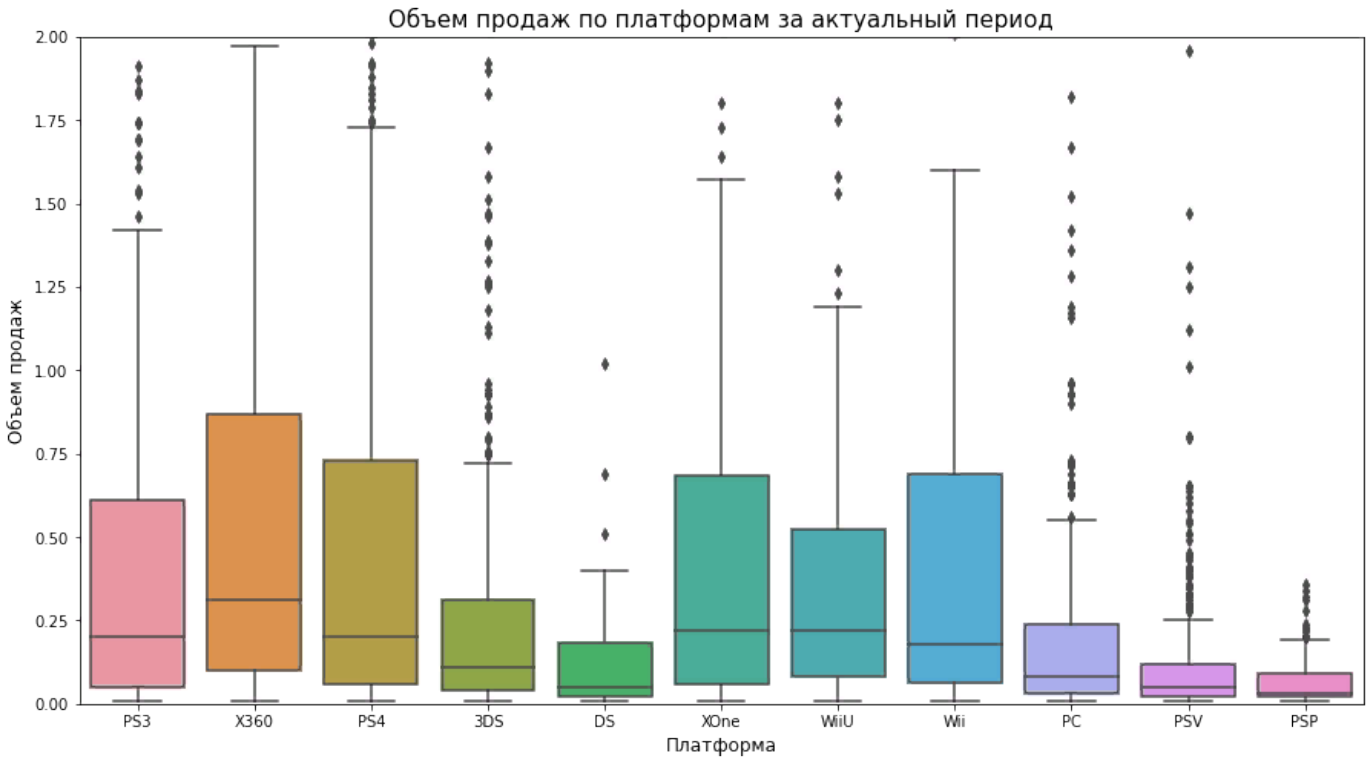
```
plt.figure(figsize=(15,8))

sns.boxplot(data=df_actual, x='platform', y='total_sales');
plt.title('Объем продаж по платформам за актуальный период', fontsize=15)
plt.xlabel('Платформа', fontsize=12)
plt.ylabel('Объем продаж',fontsize=12)
plt.show();
```



Наблюдается много выбросов.Диаграмма не очень читабельна.Масштабируем ее.

In [147…

```
plt.figure(figsize=(15,8))
plt.ylim(0, 2)
sns.boxplot(data=df_actual, x='platform', y='total_sales');
plt.title('Объем продаж по платформам за актуальный период', fontsize=15)
plt.xlabel('Платформа', fontsize=12)
plt.ylabel('Объем продаж',fontsize=12)
plt.show();
```



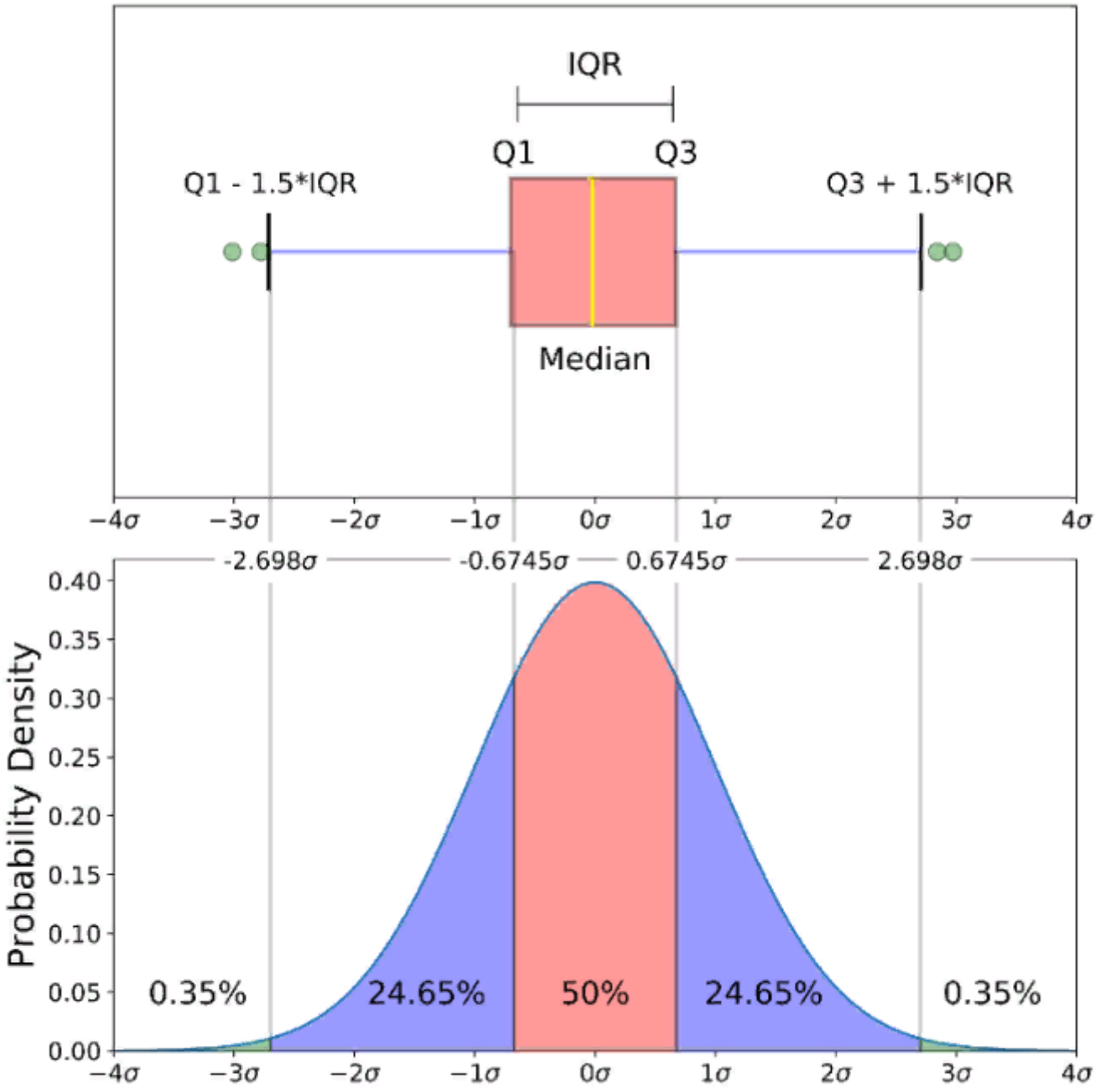
Вывод по ящику с усами:

много выбросов; медианные значения по платформ X360 и PS3 больше остальных. Объяснить это можно тем, что эти платформы "живут" уже долго, как следствие распространены больше.Межквартальный размах очень похож у платформ X360, PS3, PS4,Wii, XOne. Все распрадаления платформ скошены вправо. Можно сказать, что у всех ящиков медина лежит левее

среднего значения. У компаний, у которых пользовательский интерес ниже (DS,PC,PSP,PSV) имеют медиану около 0.1-0.15 млн.копий -они завершили или завершают свой жизненный цикл. Высокие медианные значения у PS4, XOne и WiiU около 0.2 млн.копий, хотя WiiU сильно меньше в объемах продаж. Так как продажи указаны в млн. копий, а не в денежном выражении, то есть вероятность, что платформа WiiU тоже может быть перспективной.

Комментарий ревьюера 2

👉 Для интерпретации диаграмм размаха помогает вспомнить, что означают боксплоты:



Сравнение boxplot почти нормального распределения и функции плотности вероятности (pdf) для нормального распределения

Как влияют на продажи внутри одной популярной платформы отзывы пользователей и критиков ? Построим диаграмму рассеяния и посчитаем корреляцию между отзывами и продажами

Наибольшие продажи в актуальном периоде наблюдаются у платформы 'PS4'.Посмотрим как влияют на продажи внутри этой платформы отзывы пользователей и критиков.

In [148...

```
#Отфильтруем данные по платформе PS4 и подготовим таблицу

df_PS4 = df_actual.query('platform == "PS4"')
```

```
df_PS4 = df_PS4[['critic_score', 'user_score', 'total_sales']]
df_PS4.head()
```

Out[148…

	critic_score	user_score	total_sales
31	NaN	NaN	14.63
42	97.0	8.3	12.62
77	82.0	4.3	8.58
87	NaN	NaN	7.98
92	83.0	5.7	7.66

In [149…

```
#Диаграмма рассеяния
df_PS4.plot(kind='scatter', x='critic_score', y='total_sales', figsize=(8,5), alpha=0.8, grid=T
plt.title('Зависимость Продаж от оценки критиков для PS4');
```

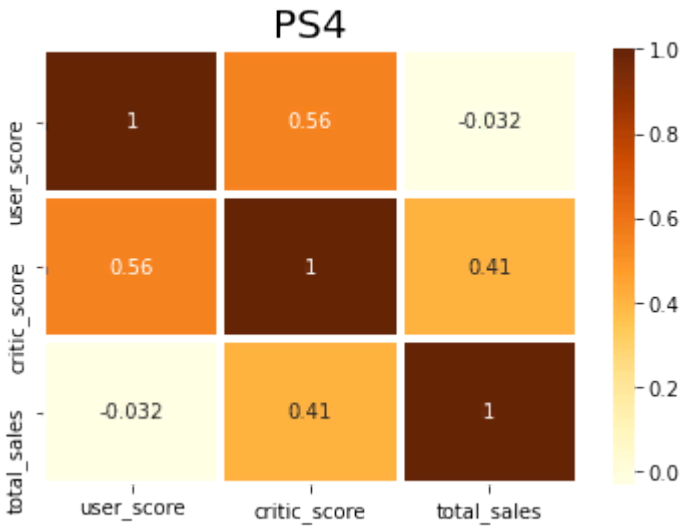
In [150…

```
#Диаграмма рассеяния
df_PS4.plot(kind='scatter', x='user_score', y='total_sales', figsize=(8,5), color='red', alpha=
plt.title('Зависимость Продаж от оценки пользователей для PS4 ');
```

Матрица корреляций для 'PS4'

In [151…
```
sns.heatmap(df_PS4[['user_score', 'critic_score','total_sales']].corr(), annot = True,cmap = '
plt.title('PS4', fontsize=20)
```

Out[151…
```
Text(0.5, 1.0, 'PS4')
```



In [152…
```
df_PS4[['critic_score', 'user_score', 'total_sales']].corr()
```

Out[152…

	critic_score	user_score	total_sales
critic_score	1.000000	0.557654	0.406568
user_score	0.557654	1.000000	-0.031957
total_sales	0.406568	-0.031957	1.000000

Выводы для популярной платформы:

- По диаграммам рассеяния видно,что пользователи и критики большинство продаваемых игр оценивают довольно высоко: у пользователей большинство оценок находится в интервале $[5.5; 8.5]$, у критиков -- $[60; 80]$.
- При этом пользователи для оценок использовали практически всю шкалу рейтинга, а критики более единодушны к подобным играм и оценивают их высоко.
- на общий объем продаж умеренное прямое влияние оказывает рейтинг критиков (коэф.корреляции-**0.41**).
- влияния рейтинга пользователей на объемы продаж нет.
- между собой связаны рейтинги критиков и пользователей.Видимо отзывы пользователей влияют на мнение критиков и наоборот.

Продажи игр на других платформах

Посмотрим как влияли оценки критиков и пользователей на продажи игр на платформах'PS3', 'X360', '3DS', 'XOne'

In [153…
```
others_platforms = ['PS3', 'X360', '3DS', 'XOne']
```

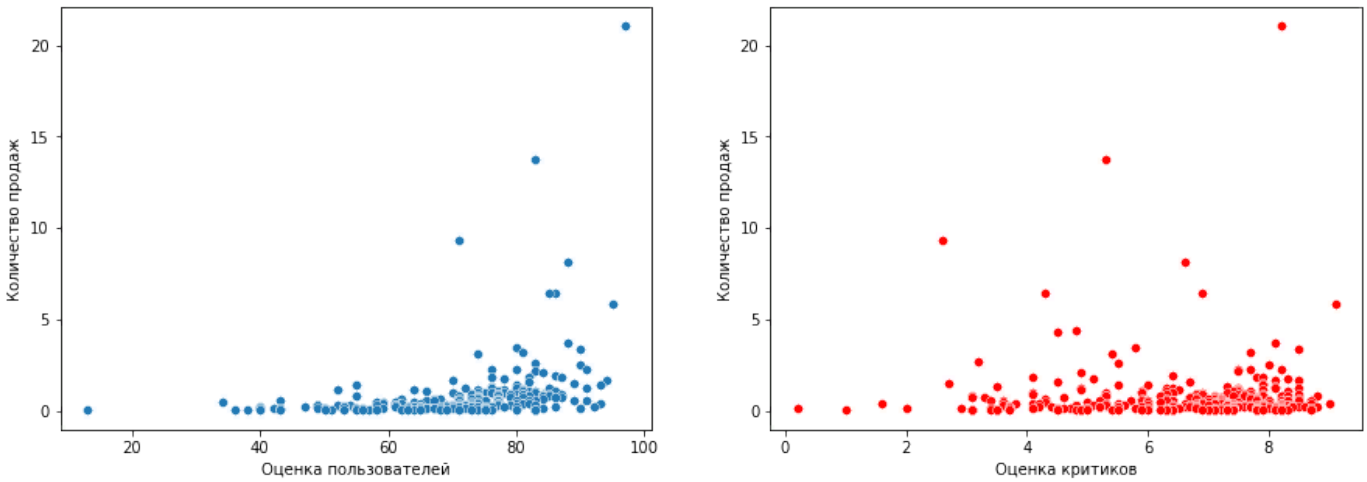
In [154…
```
def other_platform(name_of_platform):
    platform = df_actual[df_actual['platform']==name_of_platform]
    fig, ax = plt.subplots(1 ,2,figsize=(15,5))
    sns.scatterplot(x='critic_score', y='total_sales', data=platform, ax=ax[0])
    sns.scatterplot(x='user_score', y='total_sales', data=platform, ax=ax[1],color='red')

    fig.suptitle(name_of_platform, fontsize=15)
```

```
    ax[0].set(xlabel='Оценка пользователей')
    ax[1].set(xlabel='Оценка критиков')
    ax[0].set(ylabel='Количество продаж')
    ax[1].set(ylabel='Количество продаж')
    plt.show()
```

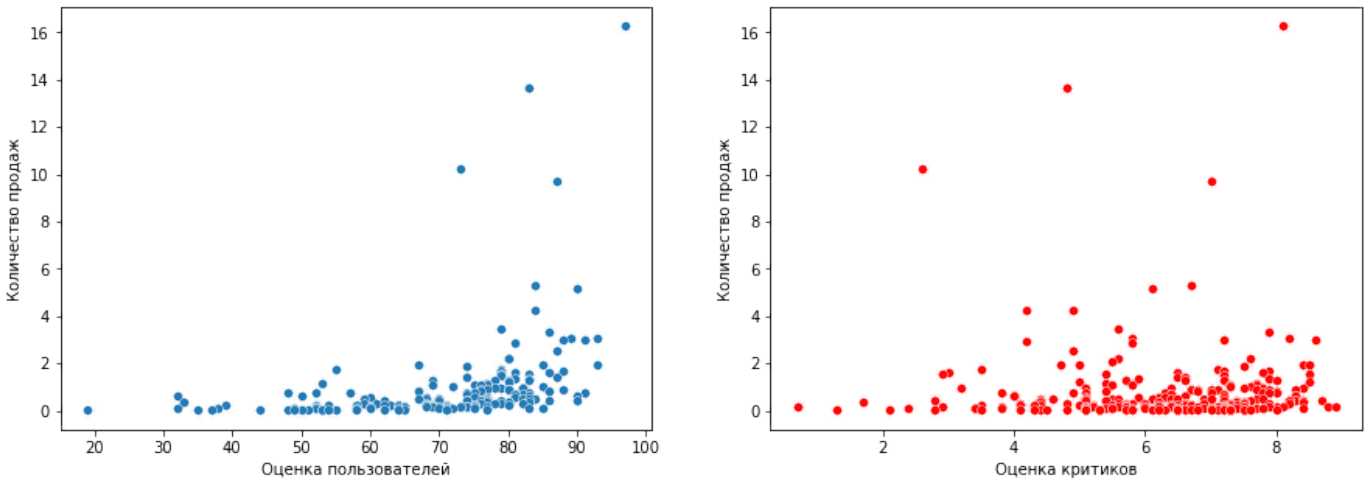
In [155…

```
for platform in others_platforms:
    other_platform(platform)
```

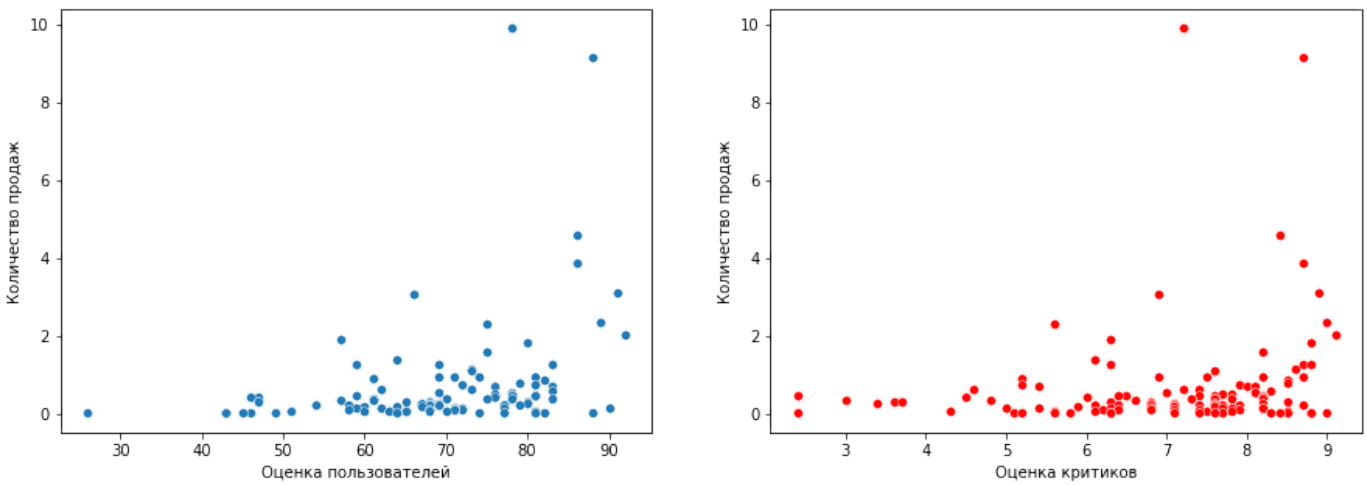
PS3



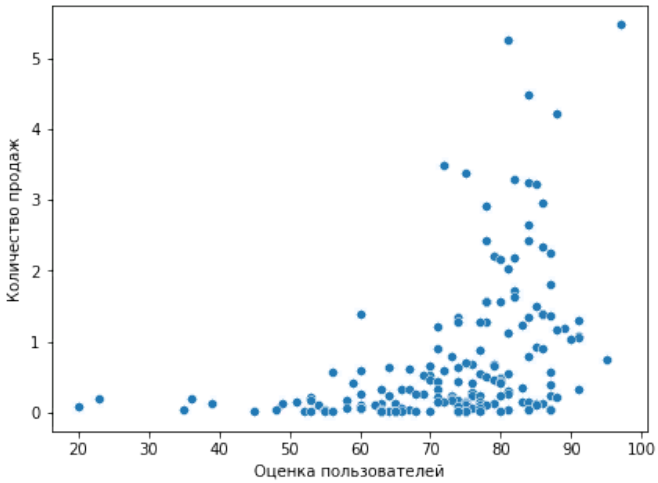
X360

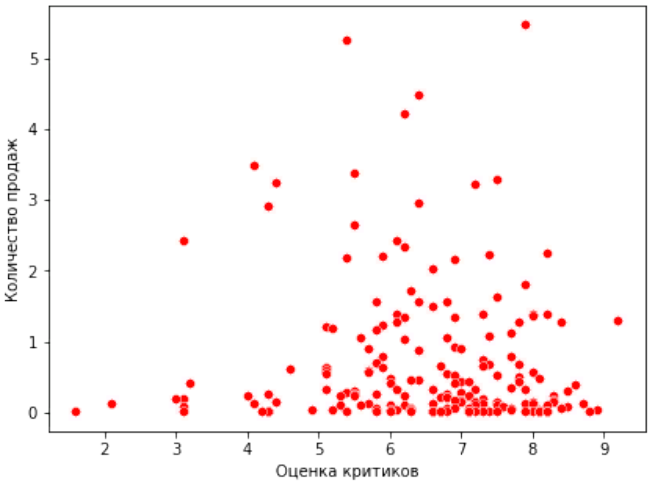


3DS



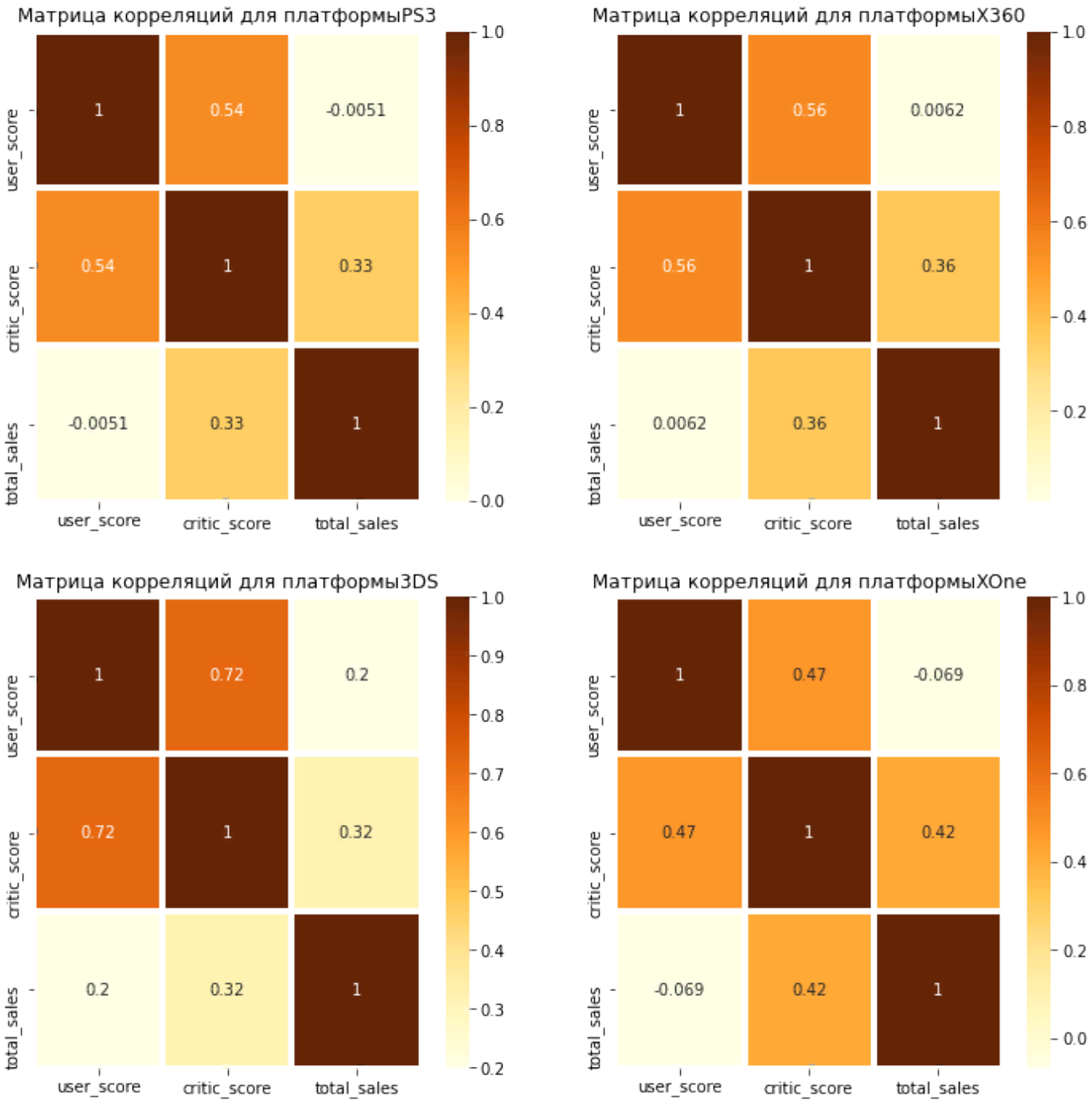
XOne





In [156…

```
fig, ax = plt.subplots(2, 2)
fig.set_size_inches(12,12)

axs = [ax[i, j] for i in range(2) for j in range(2)]

num_of_subplot = 0
for platform_name in others_platforms:

    sns.heatmap(df_actual[df_actual['platform'] == platform_name][['user_score', 'critic_score
            cmap= 'YlOrBr', linewidths=3, ax = axs[num_of_subplot]);
    axs[num_of_subplot].set_title('Матрица корреляций для платформы{}'.format(platform_name))
    num_of_subplot += 1
```



Выводы по диаграммам рассеивания:

Видим, что

- по "старым" платформам `X360` и `PS3` оценки пользователей и критиков имеют достаточно широкие интервалы $[3; 9]$ и $[30; 90]$, соответственно. Это можно объяснить тем, что за годы существования для этих платформ разработано множество игр, оцененых пользователеми и критиками и они видимо были как плохе, так и хорошие.При этом продаваемые игры критики оценивали высоко,а у пользователей наблюдается весь спектр оценок.
- для новых платформ `3DS` и `XOne` видим, что оценок пользователей и критиков значительно меньше. Оценки в интервале $[5; 9]$ и $[55; 90]$ соответственно.

Выводы по матрицам корреляции:

- для всех четырех платформ('PS3', 'X360', '3DS', 'XOne') на общий объем продаж игр умеренное прямое влияние оказывает рейтинг критиков: корреляция соответственно **0.33,0,36,0.32,0.42**
- для платформы `3DS` рейтинг пользователей также оказывает небольшое положительное прямое влияние на объемы продаж.(корреляция **0.2**)

Комментарий ревьюера

👌 Хороший анализ и визуализация! 👍 А для оценки корреляции лучше пользоваться вот этой шкалой:

Величина коэффициента корреляции отражает **силы связи**. При оценке силы связи коэффициентов корреляции используется шкала Чеддока:

Таблица анализа силы связи между переменными

Значение	Интерпретация
от 0 до 0,3	очень слабая
от 0,3 до 0,5	слабая
от 0, 5 до 0,7	средняя
от 0,7 до 0, 9	высокая
от 0,9 до 1	очень высокая

Сопоставление выводов по платформе PS4 и платформам PS3 , X360 , 3DS , XOne

- По влиянию рейтинга пользователей и критиков на объемы продаж платформа PS4 больше схожа с новыми платформами 3DS , XOne . Данный вывод подтверждается и сроками выпуска этих платформ: **PS4** выпущена в 2013; **3DS** в 2011,**XOne** в 2013. Популярность PS4 можно объяснить тем фактом, что она пришла на смену PS3 - одного из лидеров за весь период наблюдений
- Оценки критиков вляют на продажи, а оценки пользователей практически не влияют.Хотя в среднем игры на популярных платформах оцениваются достаточно высоко.

Распределение игр по жанрам

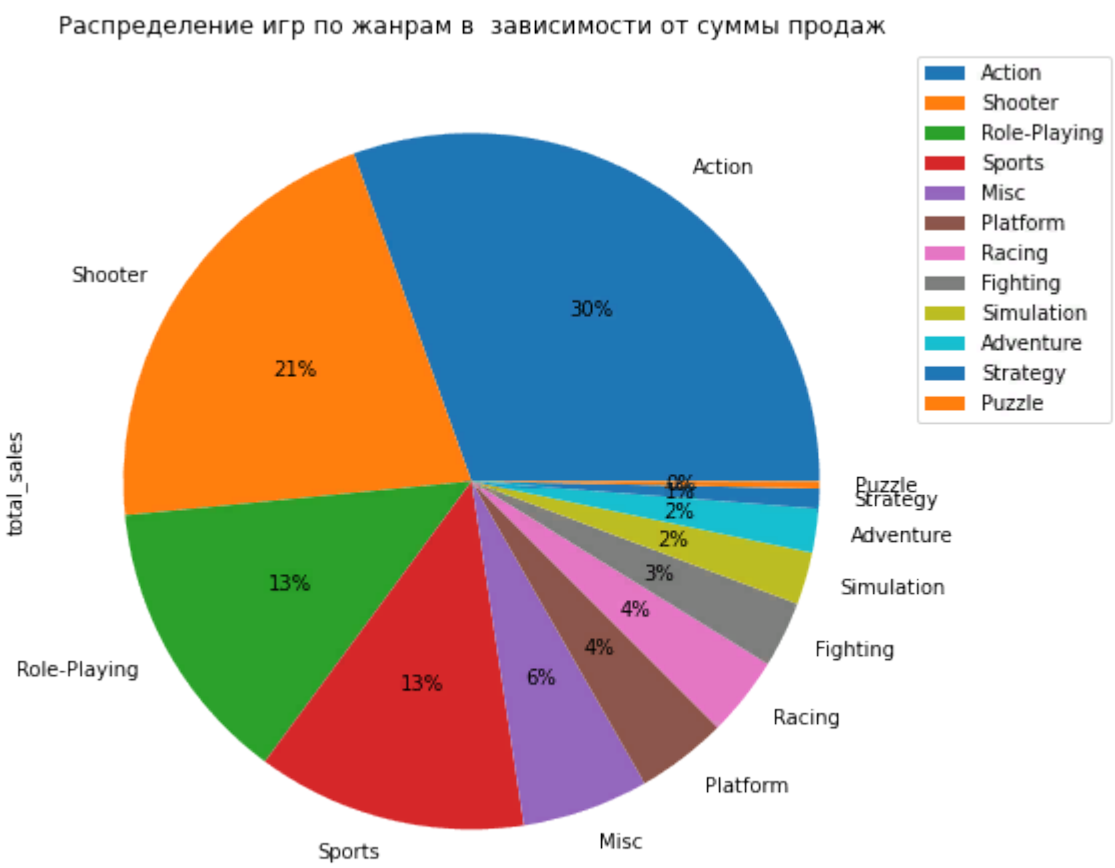
Посмотрим,как вообще распределяются игры по жанрам, используя данные за актуальный период

In [157…

```
df_actual.pivot_table(index = 'genre', values = 'total_sales', aggfunc = 'sum').sort_values(asc
plt.title("Распределение игр по жанрам в  зависимости от суммы продаж")
plt.legend(bbox_to_anchor=(1, 1), loc='upper left')
```

Out[157…

```
<matplotlib.legend.Legend at 0x1fa3db8ad00>
```



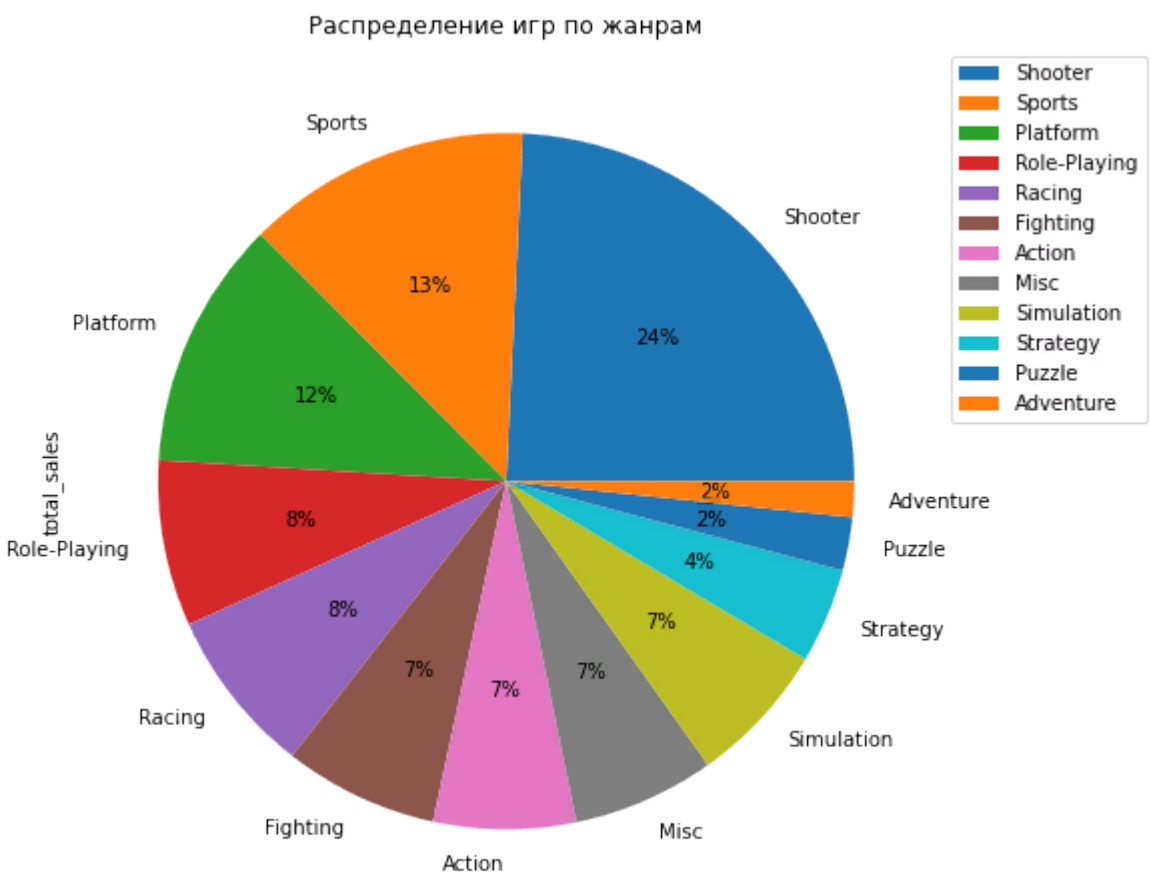
In [158...
```
df_actual.pivot_table(index = 'genre', values = 'total_sales', aggfunc = 'sum').sort_values(asc
```

Out[158...

genre	total_sales
Action	441.12
Shooter	304.73
Role-Playing	192.80
Sports	181.07
Misc	85.04
Platform	61.00
Racing	53.50
Fighting	44.49
Simulation	35.12
Adventure	29.43
Strategy	13.34
Puzzle	4.89

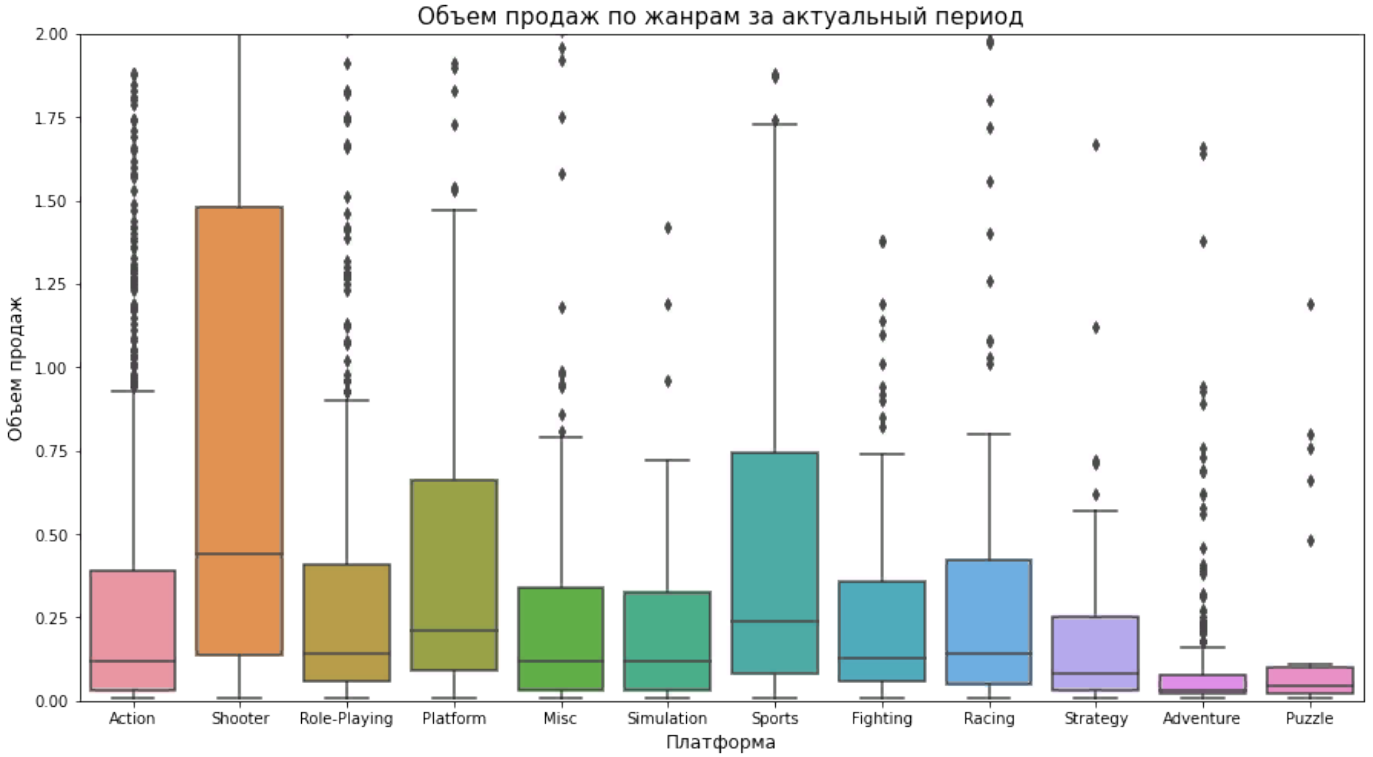
In [159...
```
df_actual.pivot_table(index = 'genre', values = 'total_sales', aggfunc = 'median').sort_values(
plt.title("Распределение игр по жанрам")
plt.legend(bbox_to_anchor=(1, 1), loc='upper left')
```

Out[159...
```
<matplotlib.legend.Legend at 0x1fa3dbd47f0>
```



In [160…

```
plt.figure(figsize=(15,8))
plt.ylim(0, 2)
sns.boxplot(data=df_actual, x='genre', y='total_sales');
plt.title('Объем продаж по жанрам за актуальный период', fontsize=15)
plt.xlabel('Платформа', fontsize=12)
plt.ylabel('Объем продаж',fontsize=12)
plt.show();
```



In [161…

```
df_actual.pivot_table(index = 'genre', values = 'total_sales', aggfunc = 'median').sort_values(
```

Out[161...

genre	total_sales
Shooter	0.440
Sports	0.240
Platform	0.210
Role-Playing	0.140
Racing	0.140
Fighting	0.130
Action	0.120
Misc	0.120
Simulation	0.120
Strategy	0.080
Puzzle	0.045
Adventure	0.030

Выводы по жанрам:

С точки зрения максимальных продаж стоит выделить:

- Наибодее прибыльные жанры - `Action` (продано 446.41 млн.копий) и `Shooter` (продано 304.75 млн.копий)
- Наименее прибыльные жанры - `Strategy и Puzzle`

С точи зрения стабильно высокого дохода следует выделить `Shooter` , `Sports` и `Platform`

Выводы по шагу 3:

- По динамике выпуска:Нам представлены данные с 1980 по 2016 год.Начиная я 1980 и до 1900 года объемы небольшие.Затем,начиная с 1990 до 2008 года, наблюдается рост выпуска компьютерных игр, а вот начиная с 2009 года объемы выпуска снижаются. С 2012 года наблюдается определенная стабилизация объемов выпуска.
- В результате исследования выявлено, что платформами с наибольшими продажами за весь период наблюдений являются 'PS2', 'X360', 'PS3', 'Wii', 'DS', 'PS'
- Со временем лидеры меняются, так как на смену одним, приходят другие платформы
- Жизненный цикл платформ в среднем составляет 11 лет.Первые 5 лет(примерно) наблюдается рост, а затем идет падение.
- в качестве актуального периода принят период с 2012 года.
- Продажи по всем платформам к 2016 году снижаются,хотя стоит заметить, что по условию данные за 2016 год неполные и на этом поэтому не стоит заострять внимание.
- Наибольшие продажи в актуальном периоде наблюдаются по платформам 'PS4', 'PS3', 'X360', '3DS', 'XOne'
- С точи зрения стабильно высокого дохода следует выделить 'Shooter','Sports' и 'Platform'
- в актуальном периоде больше всего игр продано для платформы PS4 - **314.14 млн.копий**
- Наилучшие перспективы у платформ PS4 и XOne- пик продаж у нх наблюдался в 2015 году
- для всех актуальных платформ из топ-5 наблюдается прямое положительное среднее влияние оценок критиков(уровень корреляции слабый) на объемы продаж,влияния оценок пользователей

на продажи не наблюдается.При этом критики и пользователи довольно высоко оценивают игры на платформах-лидерах.

- Определены наиболее максимально продаваемые жанры: **Action(продано 446.41 млн.копий)** и **Shooter(продано 304.75 млн.копий)**

- С точи зрения стабильно высокого дохода следует выделить `Shooter` , `Sports` и `Platform`

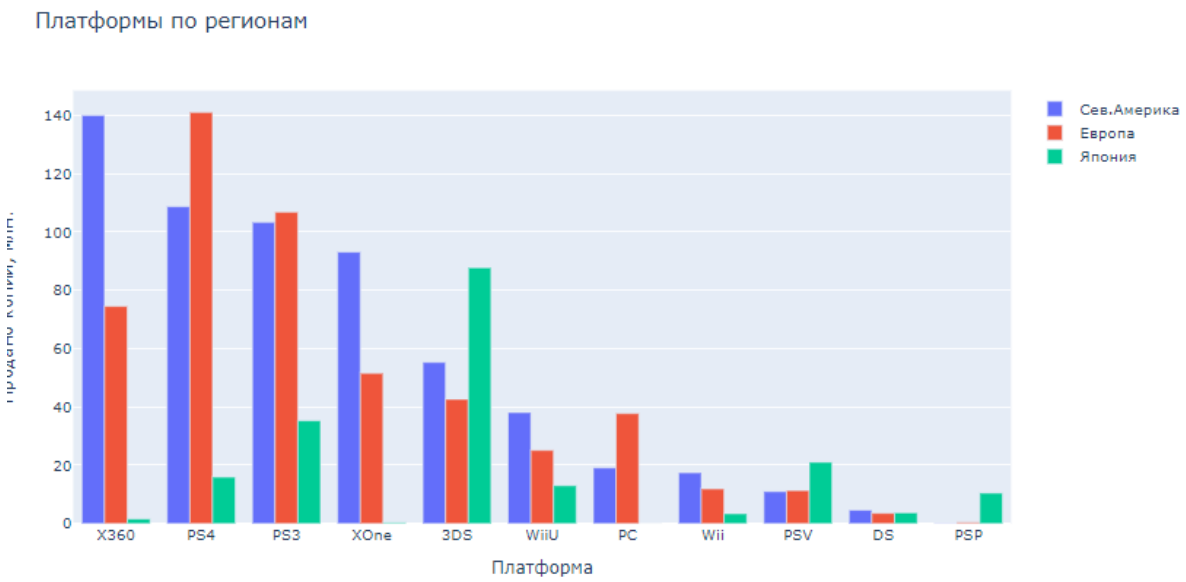
Шаг 4 Портрет пользователя каждого региона(NA, EU, JP)

Самые популярные платформы (топ-5).

Посмотрим как распределились предпочтения пользователей в разных регионах относительно игровых платформ.Используем библиотеку **plotly**

In [162…

```
sales = df_actual.pivot_table(index='platform', values=['na_sales', 'eu_sales', 'jp_sales'], ag
# Строим график
fig = go.Figure(data=[
    go.Bar(name='Сев.Америка', x=sales.index, y=sales['na_sales']),
    go.Bar(name='Европа', x=sales.index, y=sales['eu_sales']),
    go.Bar(name='Япония', x=sales.index, y=sales['jp_sales'])],
               )
fig.update_layout(
    barmode='group',
    title={'text':'Платформы по регионам'},
    xaxis_title='Платформа',
    yaxis_title='Продано копий, млн.'
)
fig.show()
```



Видно, что есть некоторые отличия по регионам.Посмотрим на какое распределение продаж по игровым платформам в долях и выделим ТОП-5 для каждого региона.Построим для наглядности круговые диаграммы.

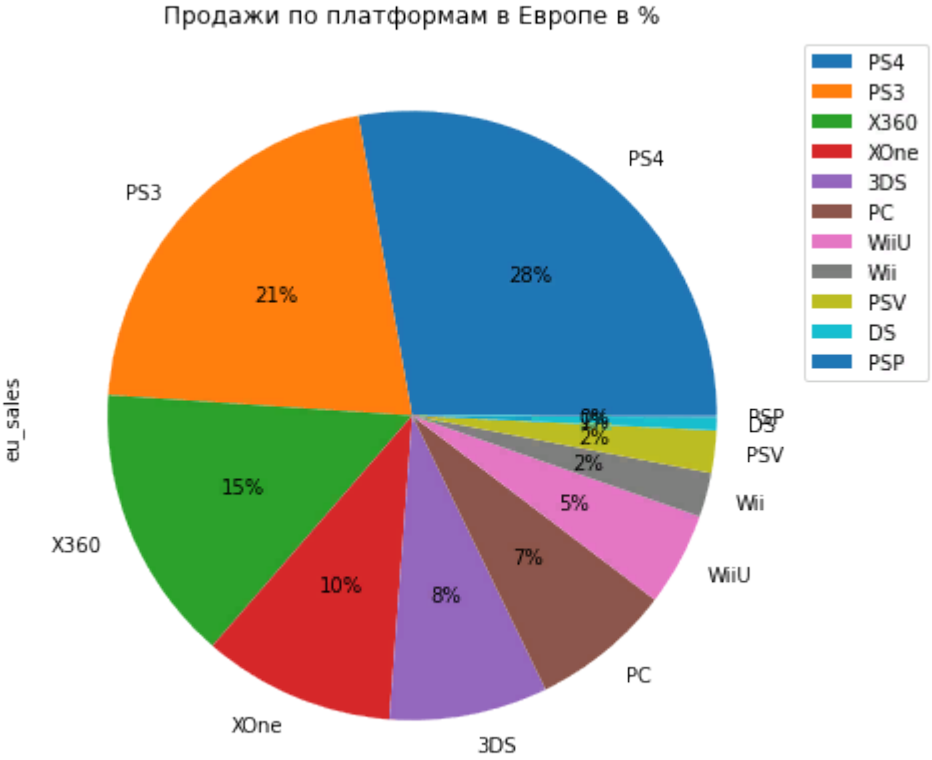
In [163…

```
df_na = df_actual.pivot_table(index='platform', values=['na_sales'], aggfunc='sum').sort_values
print('Топ-5 платформ в Сев.Америке','\n',df_na.head())

df_na.plot(
    kind='pie',figsize=(7,7), y='na_sales', autopct='%1.0f%%')
plt.title("Продажи по платформам в Северной Америке в %")
plt.legend(bbox_to_anchor=(1, 1), loc='upper left')
plt.show()

df_eu = df_actual.pivot_table(index='platform', values=['eu_sales'], aggfunc='sum').sort_values
print('Топ-5 платформ в Европе','\n',df_eu.head())
df_eu.plot(
    kind='pie',figsize=(7,7), y='eu_sales', autopct='%1.0f%%')
plt.title("Продажи по платформам в Европе в %")
plt.legend(bbox_to_anchor=(1, 1), loc='upper left')
plt.show()

df_jp = df_actual.pivot_table(index='platform', values=['jp_sales'], aggfunc='sum').sort_values
print('Топ-5 платформ в Японии','\n',df_jp.head())
df_jp.plot(
    kind='pie',figsize=(7,7), y='jp_sales', autopct='%1.0f%%')
plt.title("Продажи по платформам в Японии в %")
plt.legend(bbox_to_anchor=(1, 1), loc='upper left')

plt.show()
```

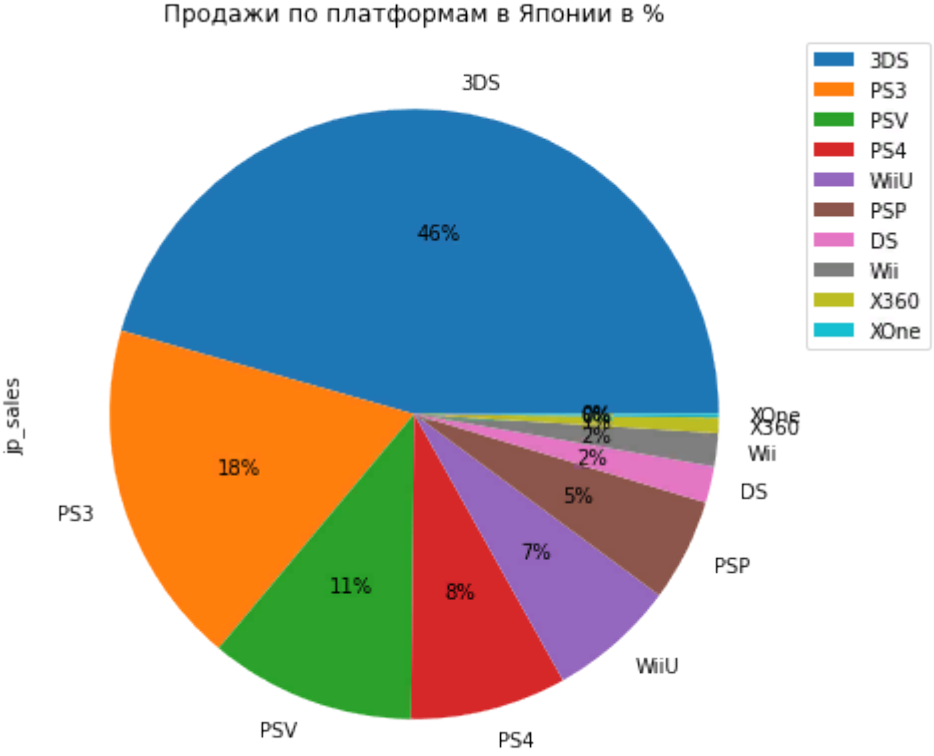
```
Топ-5 платформ в Сев.Америке
           na_sales
platform
X360         140.05
PS4          108.74
PS3          103.38
XOne          93.12
3DS           55.31
```


```
Топ-5 платформ в Европе
           eu_sales
platform
PS4          141.09
```

```
PS3          106.86
X360          74.52
XOne          51.59
3DS           42.64
```




```
Топ-5 платформ в Японии
          jp_sales
platform
3DS          87.79
PS3          35.29
PSV          21.04
PS4          15.96
WiiU         13.01
```

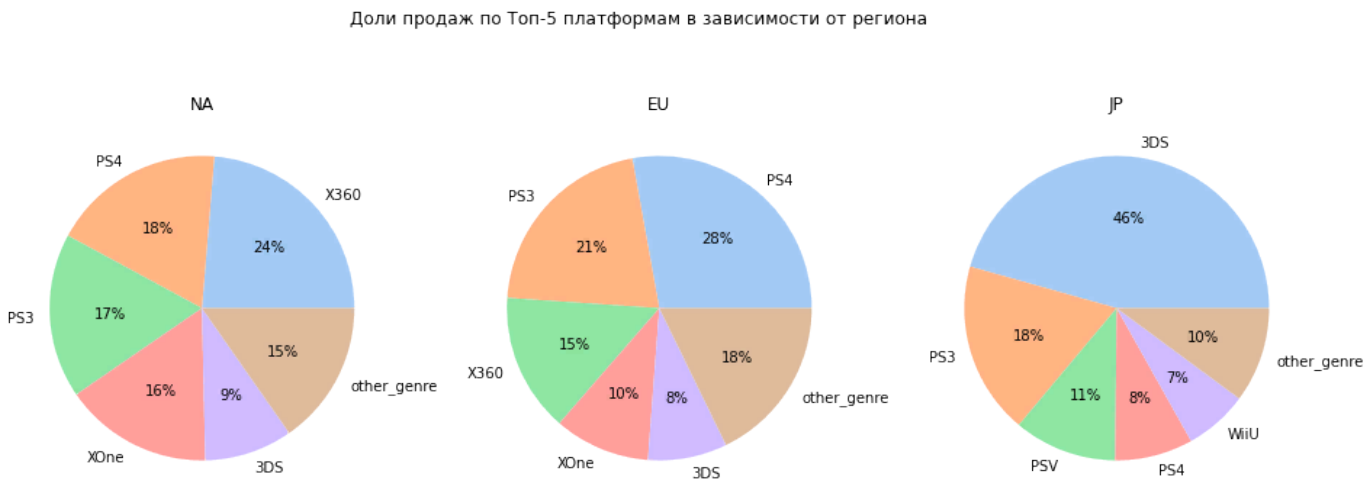


ВАРИАНТ 2

In [164…

```
regions = ['na_sales','eu_sales','jp_sales']

fig, axs = plt.subplots(1, 3,figsize=(16, 6))
```

```
colors = sns.color_palette('pastel')[ 0:6 ]
for i in range(3):
    df2 = (df_actual.groupby('platform')[regions[i]].sum().reset_index().sort_values([regions[i
    df2 = df2.append({'platform':"other_genre",regions[i]:df_actual[regions[i]].sum()-df2[regio
    axs[i].pie(df2[regions[i]],labels = df2['platform'],autopct='%1.0f%%',colors=colors)
    axs[i].set_title(df2.columns[1].split('_')[0].upper())
    plt.suptitle("Доли продаж по Топ-5 платформам в зависимости от региона")
```



Вывод

Как видим, картина для разных регионов отличается, и если Сев.Америка и Европа в чем-то схожи по предпочтениям, , то в Японии есть своя любимая платформа 3DS, которая лидирует с двукратным отрывом от ближайшего конкурента. В целом же Сев.Америка лидирует по объему рынка, на втором месте Европа, на третьем - Япония.В Америке популяона наиболее игровая консоль, разработанная компанией Microsoft - **X360**(ей на смену идет **XOne**).В Европе это **PS4 и PS3**(производитель Sony Interactive Entertainment).В Японии это **3DS и WiiU** от компании Nintendo.

--Лидеры продаж в Северной Америке: **X360(24%), PS4(18%), PS3(17%), XOne(16%), 3DS(9%)**

--Лидеры продаж в Европе: **PS4(28%), PS3(21%), X360(15%), XOne(10%),3DS(8%)**

--Лидеры продаж в Японии: **3DS(46%), PS3(18%),PSV(11%), PS4(8%),WiiU(7%)**

Самые популярные жанры (топ-5).

Посмотрим каки жанры представленны

In [165…
```
df_actual['genre'].unique()
```

Out[165…
```
array(['Action', 'Shooter', 'Role-Playing', 'Platform', 'Misc',
       'Simulation', 'Sports', 'Fighting', 'Racing', 'Strategy',
       'Adventure', 'Puzzle'], dtype=object)
```

Action - Экшен (action в переводе с англ. — «действие») или боевик (по аналогии с киножанром)

Shooter - Шу́тер (Стрелялка, англ. shooter — «стрелок»)

Role-Playing - ролевая игра

Platform -Платфо́рмер — жанр компьютерных игр, в которых основу игрового процесса составляют прыжки по платформам

Sports-спортирные **Fighting**-Фа́йтинг (от англ. Fighting — бой, драка, поединок, борьба) — жанр компьютерных игр, имитирующих рукопашный бой

Racing-Гоночная игра

Strategy-Стратегия

Adventure-англ. adventure game) или квест

Puzzle-Головоломка

In [166…

```
genres=df_actual.pivot_table(
    index='genre', values=['na_sales', 'eu_sales', 'jp_sales'], aggfunc='sum').sort_values(by=
# Строим график
fig = go.Figure(data=[
    go.Bar(name='Северная Америка', x=genres.index, y=genres['na_sales']),
    go.Bar(name='Европа', x=genres.index, y=genres['eu_sales']),
    go.Bar(name='Япония', x=genres.index, y=genres['jp_sales'])],
               )
fig.update_layout(
    barmode='group',
    title={'text':'Жанры по регионам'},
    xaxis_title='Жанр',
    yaxis_title='Продано копий, млн.'
)
fig.show()
```

In [167…

```
df_na_genre = df_actual.pivot_table(index='genre', values=['na_sales'], aggfunc='sum').sort_val
print('Топ-5 жанров в Сев.Америке','\n', df_na_genre.head())

df_eu_genre = df_actual.pivot_table(index='genre', values=['eu_sales'], aggfunc='sum').sort_val
print('Топ-5 жанров в Европе','\n',df_eu_genre.head())


df_jp_genre = df_actual.pivot_table(index='genre', values=['jp_sales'], aggfunc='sum').sort_val
print('Топ-5 жанров в Японии','\n',df_jp_genre.head())
```

```
Топ-5 жанров в Сев.Америке
              na_sales
genre
Action          177.84
Shooter         144.77
Sports           81.53
Role-Playing     64.00
Misc             38.19
Топ-5 жанров в Европе
              eu_sales
genre
Action          159.34
Shooter         113.47
Sports           69.09
Role-Playing     48.53
Racing           27.29
Топ-5 жанров в Японии
              jp_sales
genre
Role-Playing     65.44
Action           52.80
Misc             12.86
Simulation       10.41
Fighting          9.44
```

ВАРИАНТ2

In [168…

```
regions = ['na_sales','eu_sales','jp_sales']

fig, axs = plt.subplots(1, 3,figsize=(15, 5))
colors = sns.color_palette('pastel')[ 0:6 ]
for i in range(3):
    df3 = (df_actual.groupby('genre')[regions[i]].sum().reset_index().sort_values([regions[i]],
    df3 = df3.append({'genre':"other_genre",regions[i]:df_actual[regions[i]].sum()-df3[regions[
    axs[i].pie(df3[regions[i]],labels = df3['genre'],autopct='%1.0f%%',colors=colors)
    axs[i].set_title(df3.columns[1].split('_')[0].upper())
    plt.suptitle("Доли продаж по Топ-5 платформам в зависимости от региона")
```

Комментарий ревьюера

👉 У нас в задании просят сделать анализ по топ-5. И лучше визуализировать круговой диаграммой, только продажи платформ/жанров не вошедших в топ, собрать в одной группе, например "Другие", и по платформам, и по жанрам. Тогда на круговых диаграммах хорошо видны все региональные рынки игр с разбиением на платформы/жанры. Какая платформа/жанр занимает какую долю рынка. И наша визуализация несет определенную бизнес-логику, связанную именно с долями рынка занимаемыми

платформами/жанрами. А этого не дают другие виды диаграмм. Примерно вот так:

Комментарий студента:
✔Олег,спасибо!Прислушалась к твоему комментарию.Добавила визуализацию круговой диаграммой по продажам платформ/жанров(не вошедших в топ, собрала в одноу группу).Выводы обновила.Предыдущий код удалять не стала.Выводы обновила.

Комментарий ревьюера 2

👌 Молодец, хорошее решение. 👍

Вывод

Наблюдаем похожую ситуацию что и с платформаии, но в данном случае топ-5 жанров для Северной Америке и Европы почти совпадают. Разница в продажах между жанрами для этих стран так же небольшая. В Японии же предпочтения сильно отличаются, и снова есть лидеры с большим отрывом: Role-Playing и Action.

Лидеры продаж в Северной Америке: Action(30%), Shooter(24%), Sports(14%):, Role-Playing(11%), Misc(6%)

Лидеры продаж в Европе: Action(31%), Shooter(22%), Sports(14%), Role-Playing(10%),Racing(5%)

Лидеры продаж в Японии: Role-Playing(34%), Action(27%), Misc(7%), Fighting(5%), Platform(5%)

Влияние рейтинга ESRB на продажи в отдельном регионе

Посмотрим на значения признака `rating`

In [169…
```
df['rating'].unique()
```

Out[169…
```
array(['E', 'unknown', 'M', 'T', 'E10+', 'K-A', 'AO', 'EC', 'RP'],
      dtype=object)
```

'E': старше 6 лет

'E10+': старше 10

'T': старше 13 лет

'M': старше 17 лет

'unknown': Рейтинг неизвестен

'K-A: Для детей младшего возраста(старое обозначение)

'EC': Для детей младшего возраста

'RP': Рейтинг ожидается

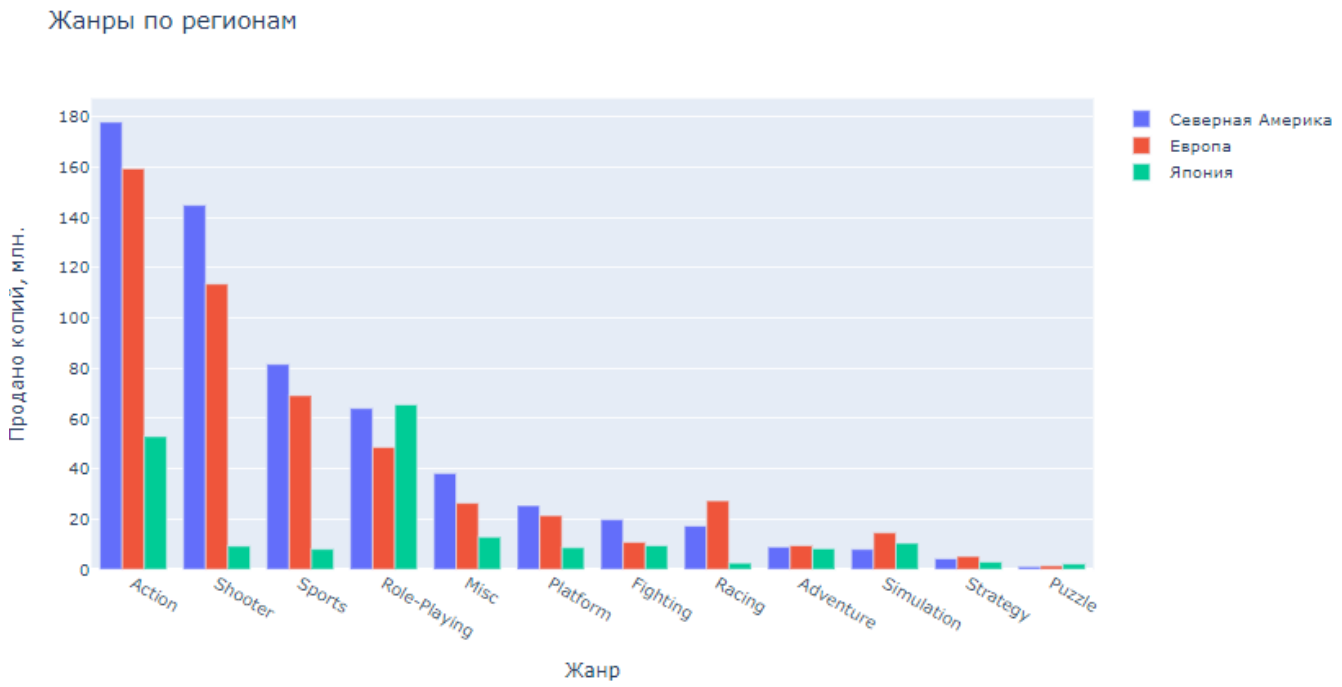
'AO': Только для взрослых

In [170…
```
df_actual['rating'].unique()
```

Out[170…
```
array(['M', 'unknown', 'E', 'E10+', 'T'], dtype=object)
```

В актуальной же выборке представлены **'M', 'unknown', 'E', 'E10+', 'T'**

In [171…
```
df_rating = df_actual.pivot_table(
    index='rating', values=['na_sales', 'eu_sales', 'jp_sales'], aggfunc='sum').sort_values(by=
# Строим график
fig = go.Figure(data=[
    go.Bar(name='Северная Америка', x= df_rating.index, y= df_rating['na_sales']),
    go.Bar(name='Европа', x=df_rating.index, y = df_rating['eu_sales']),
    go.Bar(name='Япония', x=df_rating.index, y = df_rating ['jp_sales'])],
                )
fig.update_layout(
    barmode='group',
    title={'text':'Рейтинги по регионам'},
    xaxis_title='Рейтинг',
    yaxis_title='Продано копий, млн.'
)
fig.show()
```



In [172…
```
df_na_rating = df_actual.pivot_table(index='rating', values=['na_sales'], aggfunc='sum').sort_
print('Рейтинги игр,продаваемых в Сев.Америке','\n', df_na_rating.head())

df_eu_rating = df_actual.pivot_table(index='rating', values=['eu_sales'], aggfunc='sum').sort_
print('Рейтинги игр,продаваемых в Европе','\n',df_eu_rating.head())
```

```
df_jp_rating = df_actual.pivot_table(index='rating', values=['jp_sales'], aggfunc='sum').sort_v
print('Рейтинги игр,продаваемых в Японии','\n',df_jp_rating.head())
```

```
Рейтинги игр,продаваемых в Сев.Америке
          na_sales
rating
M          231.57
E          114.37
unknown    103.31
E10+        75.70
T           66.02
Рейтинги игр,продаваемых в Европе
          eu_sales
rating
M          193.96
E          113.03
unknown     91.50
E10+        55.37
T           52.96
Рейтинги игр,продаваемых в Японии
          jp_sales
rating
unknown    108.84
E           28.33
T           26.02
M           21.20
E10+         8.19
```

Вывод:

- в Северной Америке и Европе влияние рейтингов на продажи одинаково.Наиболее продаваемыми играми являются игры с жанром **старше 17 лет**.Далне идет большая доля игр с рейтингом **старше 6 лет** - производители ориентируются на детей.А затем игры с с **неизвестным** рейтингом ,**'E10+': старше 10** и **T': старше 13 лет**

- в Японии наиболее продаваемые игры **без рейтинга**; из рейтинговых наибольшей популярностью пользуются игры **'E10+': старше 10 и и T': старше 13 лет лет**.

- Стоит отметить , что в случае , если у игры нет рейтинга,то стоит понимать, что эта игра выпущена не Северо-Американсеои регионе.Так как рейтинг ESRB, предназначен для маркировки игр для США и Канады.

Комментарий ревьюера

👌 Вот такая проблема была с рейтингом.
А дело в том, что если посмотреть в инете, что из себя представляет рейтинг ESRB, то окажется, что он предназначен для маркировки игр для США и Канады. И логично, что для других регионов он не заполняется. А в Японии есть свой рейтинг, свой рейтинг есть в ЕС и отдельно в Германии, Австралии и т.д., но по ним у нас нет данных. То есть, в данном случае пропуск имеет признак, что игра выпущена не в Северо-Американском регионе. Необходимо шире смотреть на исходный датасет, так как за цифрами находятся реальные бизнес-процессы. Мы должны это учитывать при предобработке данных. 👍

Комментарий студента:
✔Спасибо за комментарий!Приняла к сведению!

Выводы по шагу 4

Портрет пользователя из Северной Америки:

- Предпочитает игровые платформы `PS4` или `X360` (но ей на замену идет `XOne` !!).Наиболее популярны жанры `Shooter` , `Action` и `Sports` с рейтингом **старше 17 лет**, также активно

покупаются игры с оценкой **старше 6 лет**

Портрет пользователя из Европы:

- В Европе в основном игровая платформа `PS4` , набирает популярность `XOne` Наиболее популярны жанры `Shooter` , `Action` и `Sports` с рейтингом **старше 17 лет**, также активно покупаются игры с оценкой **старше 6 лет**

Портрет пользователя из Японии:

- Предпочитает игровую платформу `3DS` , но популярность приобретает платформа `PS4` .Наиболее популярные жанры `Action` и `Role-Playing` , `Misc` ; активно покупаются игры с без рейтинга или с оценкой **старше 13 лет**.

Шаг 5. Проверка гипотез

Гипотеза 1: Средние пользовательские рейтинги платформ Xbox One и PC одинаковые

В научных исследованиях нулевая гипотеза (часто обозначаемая H0) - это утверждение о том, что между двумя анализируемыми наборами данных или переменных не существует различий или взаимосвязей. исходя из этого сформулируем гипотезы:

H0. Средние пользовательские рейтинги платформ Xbox One и PC одинаковы.

H1. Средние пользовательские рейтинги платформ Xbox One и PC различны.

Сформируем выборки по рейтингам на этик платформах и сравним дисперсии

In [173…
```
alpha = 0.05
xbox = df_actual[(df_actual['platform'] == 'XOne') & (df_actual['user_score'] > 0)]['user_score
pc = df_actual[(df_actual['platform'] == 'PC') & (df_actual['user_score'] > 0)]['user_score']
if np.var(xbox)==np.var(pc):
    print('True')
else:
    print('False')
```

```
False
```

In [174…
```
xbox.size,pc.size
```

Out[174… `(182, 206)`

Зададим пороговое значение alpha = 0.01 и поверим дисперсии выборок

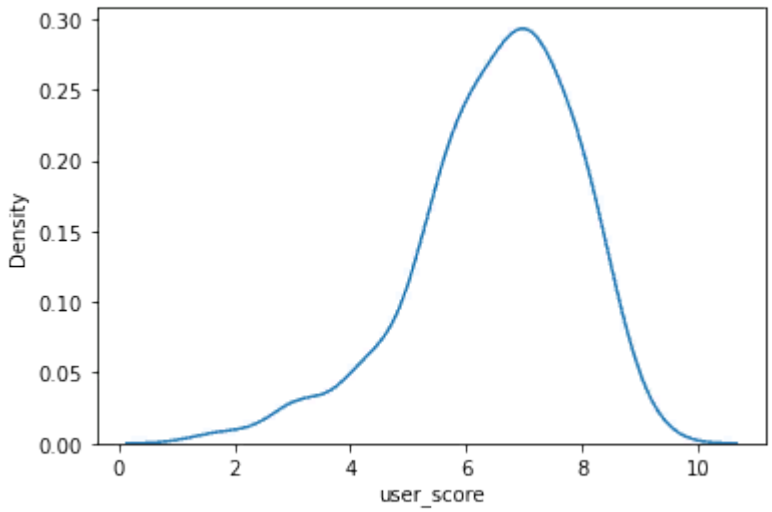
Значение α =0,05 рекомендовано для небольших выборок (когда высока вероятность ошибки второго рода). Если объемы выборок n ≥ 100, то порог отклонения Но целесообразно снизить до (α =0,01 и принимать решение о наличии связи (различий) при p ≤ 0,01

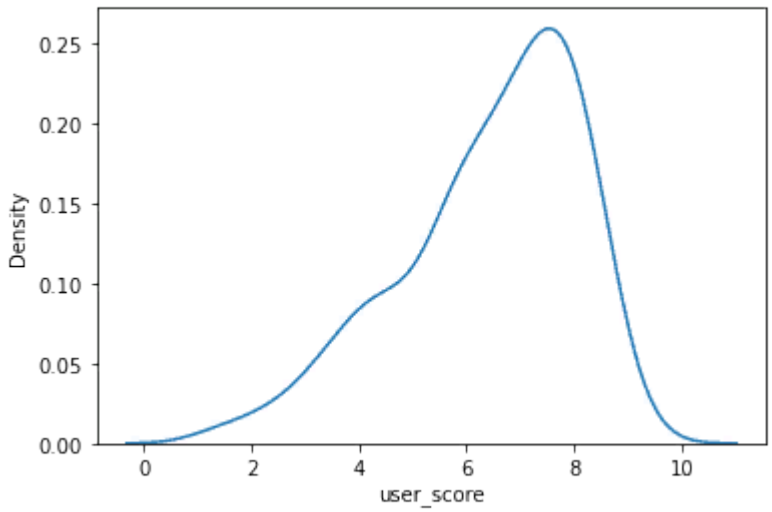
In [175…
```
print('Дисперсия XOne =', np.var(xbox))
print('Дисперсия PC =',  np.var(pc))
```

```
Дисперсия XOne = 1.8965188383045533
Дисперсия PC = 2.7569952398906565
```

In [176…
```
sns.kdeplot(xbox)
plt.show()
```

```
sns.kdeplot(pc)
plt.show()
```





Распределение оценок пользователей близко к нормальнму,мы сравниваем 2 величины.Мы вполне можем применить Критерий Стьюдента(t-тест)

In [177…

```
results = scipy.stats.ttest_ind(xbox, pc, equal_var = False)
print('p-значение:', results.pvalue)
alpha = 0.05
if (results.pvalue < alpha):
    print("Отвергаем нулевую гипотезу")
else:
    print("Не получилось отвергнуть нулевую гипотезу")
```

```
p-значение: 0.5489537965134912
Не получилось отвергнуть нулевую гипотезу
```

Комментарий ревьюера

👌 Правильно, что для тестирования гипотез использован двусторонний t-тест.

Нулевую гипотезу отвергнуть не получилось.Посмотрим на средние и медианные значения рейтингов

In [178…

```
df_actual.query('platform =="XOne" or platform =="PC"').pivot_table(index='platform', values='u
```

Out[178…

platform	mean user_score	median user_score
PC	6.43	6.8
XOne	6.52	6.8

Вывод: Даже если средние значения рейтингов не равны(а медианные,заметим,равны!), с вероятностью более 62% такое, или большее различие можно получить случайно, соответственно у нас нет оснований полагать, что средние пользовательские рейтинги платформ Xbox One и PC значимо отличаются.

Гипотеза 2 : Средние пользовательские рейтинги жанров Action (англ. «действие», экшен-игры) и Sports (англ. «спортивные соревнования») разные.

По аналогии с предыдущим пунктом сформулируем нулевую и альтернативную гипотезы.

H0. Средние пользовательские рейтинги жанров Action (англ. «действие») и Sports (англ. «виды спорта»)не отличаются

H1. Средние пользовательские рейтинги жанров Action (англ. «действие») и Sports (англ. «виды спорта») различны

In [179…
```
action = df_actual[(df_actual['genre'] == 'Action') & (df_actual['user_score'] > 0)]['user_scor
sports = df_actual[(df_actual['genre'] == 'Sports') & (df_actual['user_score'] > 0)]['user_scor

if np.var(action)==np.var(sports):
    print('True')
else:
    print('False')
```

```
False
```

In [180…
```
action.size,sports.size
```

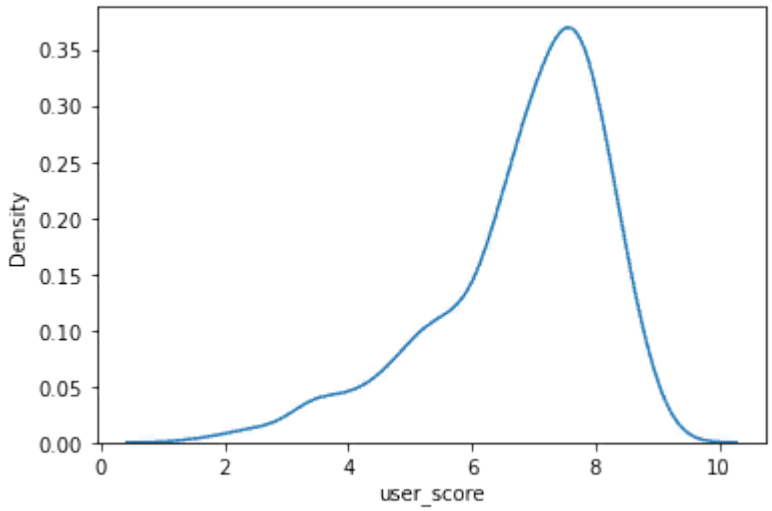
Out[180…
```
(523, 195)
```

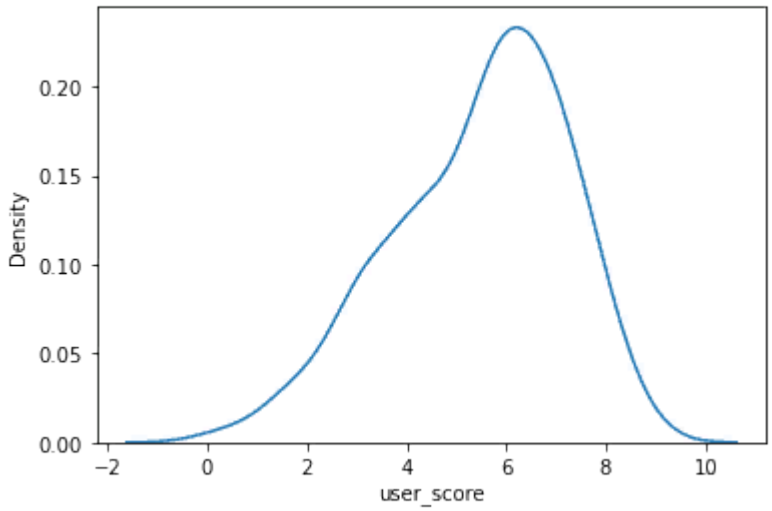
Зададим пороговое значение alpha = 0.01 и поверим дисперсии выборок

In [181…
```
print('Дисперсия XOne =', np.var(action))
print('Дисперсия PC =',  np.var(sports))
```

```
Дисперсия XOne = 1.8854720340439228
Дисперсия PC = 3.007388297172914
```

In [182…
```
sns.kdeplot(action)
plt.show()

sns.kdeplot(sports)
plt.show()
```





Аналогично предыдущей задаче распределение оценок пользователей близко к нормальнму,мы сравниваем 2 величины.Мы вполне можем применить Критерий Стьюдента(t-тест) Дисперсии значительно отличаются,также как и размер выборок, поэтому применим параметр equal_var = False

In [183…

```
results=scipy.stats.ttest_ind(action, sports, equal_var = False)
print('p-значение:', results.pvalue)
alpha = 0.01
if (results.pvalue < alpha):
    print("Отвергаем нулевую гипотезу")
else:
    print("Не получилось отвергнуть нулевую гипотезу")
```

```
p-значение: 4.24307776572644e-20
Отвергаем нулевую гипотезу
```

In [184…

```
df_actual.query('genre == "Action" or genre == "Sports"').pivot_table(
    index='genre', values='user_score', aggfunc=['mean','median']).round(2)
```

Out[184…

	mean	median
	user_score	user_score
genre		
Action	6.83	7.1
Sports	5.46	5.7

Вывод: Отвергаем нулевую гипотезу,p-значение в этом случае нрайне мало и значительно ниже порогового значения.Таким образом, средние пользовательские рейтинги для жанров Action и Sports отличаются. Можно предположить, что в среднем рейтинги Action выше.По расчету они действительно отличаются

Выводы по шагу 5

На данном этапе проверялись гипотезы:

- **Средние пользовательские рейтинги платформ Xbox One и PC одинаковые** -В результате проверки гипотеза подтвердилась.средние пользовательские рейтинги платформ Xbox One и PC значимо не отличаются.
- **Средние пользовательские рейтинги жанров Action (англ. «действие», экшен-игры) и Sports (англ. «спортивные соревнования») разные.**- Эта гипотеза также подтвердилась в результате анализа .

  -Для проверки были сформулинованы соответствующие нулевые и альтернативные гипотеы,Задано пороговое значение alpha = 0.01 , Для проверки был выбран метод **t-тест**

Комментарий ревьюера
👌 Хорошо сделана проверка гипотез. 👍

Шаг 6.бщий вывод

В рамках проекта для анализа нам были доступны из открытых источников исторические данные о продажах игр, оценки пользователей и экспертов, жанры и платформы (например, Xbox или PlayStation).

Цель - выявить определяющие успешность игры закономерности позволяющие сделать ставку на потенциально популярный продукт и спланировать рекламные кампаниив перспективе-на ближайший год.

**В результате выполнеиия проекта были сделаны следующие выводы и решны задачи:*

1.Загрузка данных и знакомство с ними

- ВыгруженаТаблица с данными,признаки представлены следующими типами : float64(6),object(5)
- По результатам проведенного первичного обследования данных были сделаны выводы по признакам и поставлены задачи для подготовки данных

2. Подготовка данных

- Приведены названия столбцов к нижнему регистру согласно общепринятым нормам
- Преобразованы типы данных в признаках : в `year_of_release` заменеп на целочисленный(Int64) и в `User_Score` на (float)предварительно заменив значение `tbd` признака `user_score` на значение `NaN` , как на более удобное для обработки в Pandas
- Проведена проверка на явные дубликаты.Их не обнаружено.
- удалены пропуски значений в признаках `name` и `genre`
- в признаке `year_of_release` удалены пропущенные значения удалены
- пояснены возможные причины возникновения пропусков в рейтингах игр и заменены на значение `unknown`
- в `critic_score` , `user_score` пропущенные значения осталены без изменений

- посчитаны суммарные продажи во всех регионах, новые значения записаны в отдельный столбец `total_sales`
- удалены игры с нулевыми продажами

3.Исследовательский анализ данных

- По динамике выпуска:Нам представлены данные с 1980 по 2016 год.Начиная я 1980 и до 1900 года объемы небольшие.Затем,начиная с 1990 до 2008 года, наблюдается рост выпуска компьютерных игр, а вот начиная с 2009 года объемы выпуска снижаются. С 2012 года наблюдается определенная стабилизация объемов выпуска.
- В результате исследования выявлено, что платформами с наибольшими продажами за весь период наблюдений являются **'PS2', 'X360', 'PS3', 'Wii', 'DS', 'PS'**
- Со временем лидеры меняются, так как на смену одним, приходят другие платформы
- Жизненный цикл платформ в среднем составляет 11 лет.Первые 5 лет(примерно) наблюдается рост, а затем идет падение.
- **в качестве актуального периода принят период с 2012 года.**
- Продажи по всем платформам к 2016 году снижаются,хотя стоит заметить, что по условию данные за 2016 год неполные и на этом поэтому не стоит заострять внимание.
- Наибольшие продажи в актуальном периоде наблюдаются по платформам **'PS4', 'PS3', 'X360', '3DS', 'XOne'**

  в актуальном периоде больше всего игр продано для платформы PS4 - **314.14 млн.копий**
- Наилучшие перспективы у платформ PS4 и XOne- пик продаж у нх наблюдался в 2015 году
- С точи зрения стабильно высокого дохода следует выделить `Shooter` , `Sports` и `Platform`
- для всех актуальных платформ из топ-5 наблюдается прямое положительное влияние оценок критиков на объемы продаж(корреляция слабая),влияния оценок пользователей на продажи не наблюдается.При этом критики и пользователи довольно высоко оценивают игры на платформах-лидерах.
- Определены наиболее прибыльные жанры: **Action(продано 446.41 млн.копий)** и **Shooter(продано 304.75 млн.копий)**

4.Портрет пользователя каждого региона(NA, EU, JP)

- Портрет пользователя из Северной Америки:

Предпочитает игровые платформы `PS4` или `X360` (но ей на замену идет `XOne` !!).Наиболее популярны жанры `Shooter` , `Action` и `Sports` с рейтингом **M(старше 17 лет)**, также активно покупаются игры с рейтингом **старше 6 лет**

- Портрет пользователя из Европы:

Предплчитает в основном игровая платформа `PS4` , набирает популярность `XOne` .Наиболее популярны жанры `Shooter` , `Action` и `Sports` с рейтингом **старше 17 лет**, также активно покупаются игры с рейтингом **старше 6 лет**

- Портрет пользователя из Японии:

Предпочитает игровую платформу 3DS , но популярность приобретает платформа PS4 .Наиболее популярные жанры Action и Role-Playing , Misc ; активно покупаются игры с **без рейтинга** или с рнйтингом **старше 13 лет.**

5.Проверка гипотез

На данном этапе проверялись гипотезы:

- **Средние пользовательские рейтинги платформ Xbox One и PC одинаковые** -В результате проверки гипотеза подтвердилась.средние пользовательские рейтинги платформ Xbox One и PC значимо не отличаются.
- **Средние пользовательские рейтинги жанров Action (англ. «действие», экшен-игры) и Sports (англ. «спортивные соревнования») разные**.- Эта гипотеза также подтвердилась в результате анализа .
- Для проверки были сформулинованы соответствующие нулевые и альтернативные гипотеы,Задано пороговое значение alpha = 0.01 , Для проверки был выбран метод t-тест

ИТОГ

По итогам исследования были выявлены следующие закономерности, позволяющие прогнозировать коммерческий успех компьютерных на 2017 год:

- Игровая платформа - важнейший критерий будущей популярности игры.
- Средний срок жизни игровой платформы - около 11 лет, первую половину которых идет плавный рост продаж, а вторую - спад.
- Можно предположить, что в 2017 году наибольшим успехом будут пользоваться платформы PS4 и XOne . В Японии –это платформа 3DS .
- Для выбора потенциально успешной игры необходимо изучить оценки критиков - есть прямая зависимость между этим параметроми объемом продаж.(Хотя, стоит заметить она средняя) При этом отзывы пользователей коррелируют с продажами незначительно лишь на определенных платформах.Для перспективных "западных" платформ, таких как PS4 и XOne , оценкой пользователей можно пренебречь, она никак не связана с объемом продаж. Отмечено, что для японского рынка связь между отзывами критиков и продажами гораздо ниже, чем в остальном мире.
- Самые продаваемые игровые жанры - Action , Shooter . Однако если рассматривать отдельно японский рынок, в первую очередь необходимо обратить внимание на Role-Playing и Action . То есть параметры потенциально успешной игры могут кардинально различаться в зависимости от региона, это обязательно нужно учитывать при составлении прогноза.
- Стоит отметить, что к сожелению очень многие игры не имеют рейтинга ESRB, поэтому данные анализа здесь не совсем полные и корректные.
- в Северной Америке и Европе влияние рейтингов на продажи одинаково.Наиболее продаваемыми играми являются игры с жанром **'M'(старше 17 лет)**.Далне идет большая доля игр с рейтингом **'E'(старше 6 лет**_ - производители ориентируются на детей.А затем игры с с **неизвестным** рейтингом ,**'E10'(старше 10)** и **T'(старше 13 лет)** В Японии наиболее продаваемые игры без рейтинга; из рейтинговых наибольшей популярностью пользуются игры **'E10+'(старше 10) и и T'(старше 13 лет лет)**.Так что видимо стоит сосредоточиться больше на взрослой аудитории,но при этом и не забывать о детском сегменте игр.

- В целом пользователи из регионов Европа и Северная Америка очень схожи в своих предпочтених и можно объеденить усилия и рекламный контент для этих направлений в рамках намечающейся рекламной кампании. Но необходимо более тщательно подготовить рекламную кампанию в Японии с учетом выявленных особенностей.

Рекомендации

В рамках рекламной кампании(если бюджет ограничен) стоит обратить внимание в первую очередь на **Северо-Американский регион**.Перспективными будут платформы `PS4` и `XOne` ,игровые жанры - `Action` , `Shooter` и основная целевая аудитория- взросые люди(рейтинг игр ESRB **'M'(старше 17 лет)**.Неожодимо учитывать отзывы критков,а пользовательскими отзывами стоит принебречь.

Заключительный комментарий ревьюера 2

👌 Серафима! Нам удалось справиться со всеми подводными камнями в проекте. Мы молодцы! 😄
Мы узнали, что не все пропуски просто ошибки сбора данных, а несут свои признаки. Для этого нам понадобилось за данными датасета увидеть смысл этих данных и бизнес-процессы, которые за ними скрываются. Мы узнали, что медианы могут дать дополнительную информацию при анализе. Это наши новые знания и навыки, которые нам пригодятся в будущем!
👌 Теперь, вперед за новыми знаниями и навыками! Удачи! 😉