Table of Contents

Анализ для интернет-магазина "Стримчик"

Описание проекта

Необходимо провести анализ и спланировать кампанию на 2017 год для онлайн-магазина Стримчик.

Из открытых источников нам доступны исторические данные о продажах игр, оценки пользователей и экспертов, жанры и платформы (например, Xbox или PlayStation). Нужно выявить определяющие успешность игры закономерности, что позволит сделать ставку на потенциально популярный продукт и спланировать рекламные кампании.

Определим следующие шаги выполнения проекта:

1. Изучение общей информации, знакомство с данными

2. Обработка и подготовка данных:

- приведение всех названий столбцов к общему виду
- обработка пропусков и преобразование данных в корректные типы, поиск дубликатов
- добавление необходимых для дальнейшего анализа столбцов

3. Исследовательский анализ данных

- посмотрим, сколько игр выпускалось в разные годы, изучим как менялись продажи по платформам и исходя из этого анализа определим характерный срок устаревания платформы
- определим актуальный период, который сможет построить прогноз на 2017 год, отбросив данные устаревших платформ
- определим платформы, которые лидируют, растут и падают по продажам, определим потенциально прибыльные платформы
- на примере одной из платформ проверим, влияют ли оценки пользователей и критиков на продажи
- проанализируем распределение игр по жанрам, выявим самые прибыльные жанры

4. Анализ по регионам:

- определим для каждого региона:
  - самые популярные платформы
  - самые популярные жанры
  - влияние рейтинга ESRB на продажи

5. Проверка гипотез:

- Средние пользовательские рейтинги платформ Xbox One и PC одинаковые
- Средние пользовательские рейтинги жанров Action и Sports разные

6. Формулирование итоговых выводов

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.

Знакомство с данными: общая информация

In [1]:
```
# импорт необходимых для работы библиотек, установка стиля графиков:
import pandas as pd
from scipy import stats as st
import numpy as np
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')
%config InlineBackend.figure_format = 'retina'
```

In [2]:
```
# чтение файла, знакомство с данными:
try:
    data = pd.read_csv('https://code.s3.yandex.net/datasets/games.csv', sep=',')
```

```
except:
    data = pd.read_csv('/datasets/games.csv', sep=',')

display(data.head(20))
data.info()
data.describe()
```

	Name	Platform	Year_of_Release	Genre	NA_sales	EU_sales	JP_sales	Other_sales	Critic_Score
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
15	Wii Fit Plus	Wii	2009.0	Sports	9.01	8.49	2.53	1.77	80.0
16	Grand Theft Auto V	PS3	2013.0	Action	7.02	9.09	0.98	3.96	97.0
17	Grand Theft Auto: San Andreas	PS2	2004.0	Action	9.43	0.40	0.41	10.57	95.0
18	Super Mario World	SNES	1990.0	Platform	12.78	3.75	3.54	0.55	NaN
19	Brain Age: Train Your Brain in Minutes a Day	DS	2005.0	Misc	4.74	9.20	4.16	2.04	77.0

```
<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
```

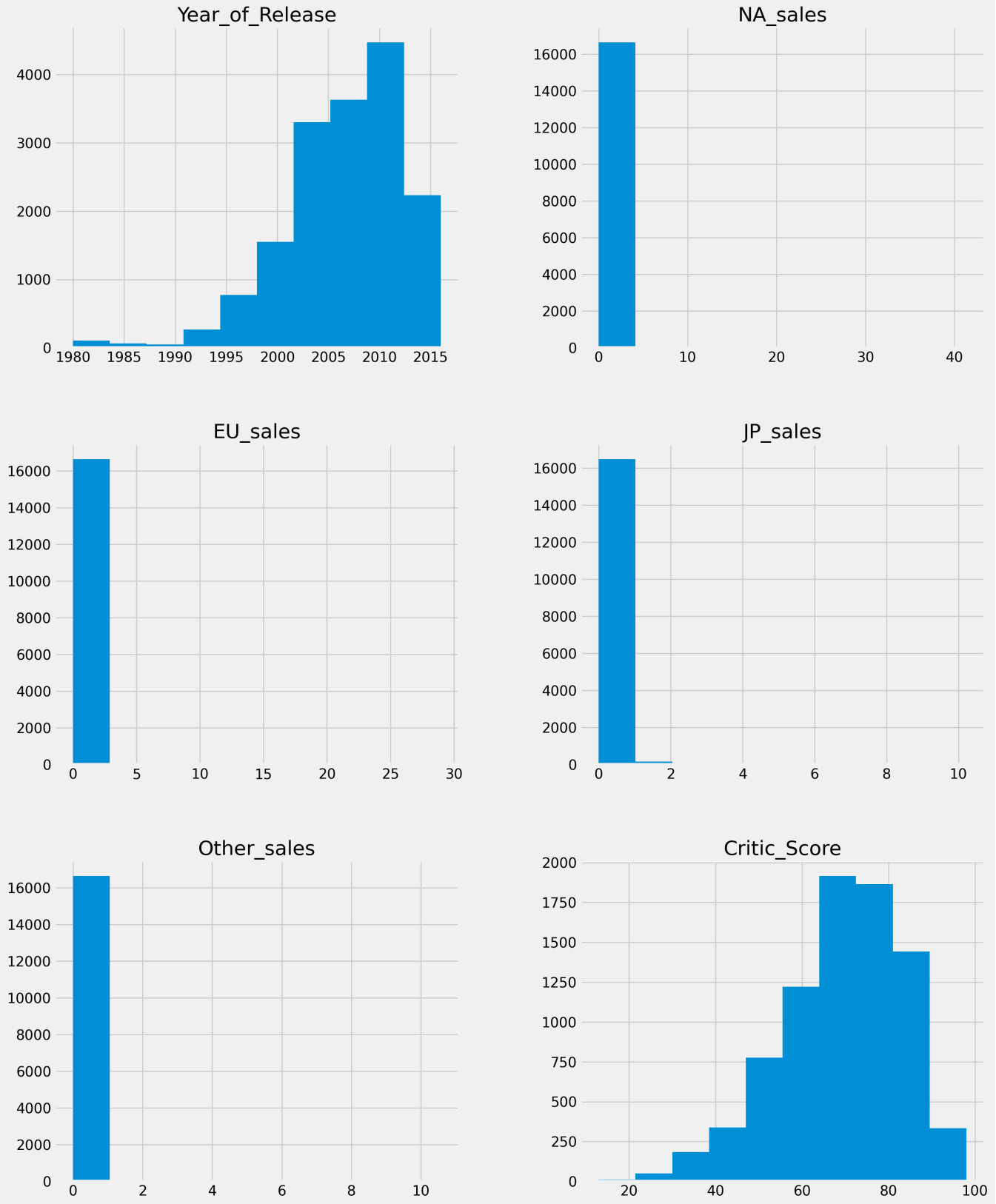
Out[2]:

	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

Набор данных включает в себя **16715** наблюдений, таблица состоит из 10 столбцов. Имеются пропуски, некоторым столбцам нужно изменить типы данных

Бегло оценим гистограммы для всех числовых столбцов таблицы:

In [3]:
```
data.hist(figsize=(15, 20))
plt.show()
```



- **Year_of_Release** - распределение выглядит нормально, большинство значений лежит в диапазоне 2010-2012

Судя по 4 столбцам с продажами в разных регионах, очевидно есть лидеры по продажам, в целом, распределение выглядит нормально:

- **NA_sales** - абсолютное большинство значений расположено до 4 млн., максимальное значение ~ 40 млн
- **EU_sales** - абсолютное большинство значений расположено до 2,5 млн, максимальное значение ~ 30 млн
- **JP_sales** - абсолютное большинство значений расположено до 1 млн, максимальное значение ~ 12 млн

- **Other_sales** - абсолютное большинство значений расположено до 1 млн, максимальное значение ~ 15 млн

- **Critic_Score** - распределение выглядит нормально, чаще всего игры получают оценку ~70, все значения лежат в допустимых границах

На графиках ожидаемо не видим гистограммы для столбца с оценками от пользователей - данные в этом столбце записаны в некорректном типе 'object'

Предобработка данных

Переименование столбцов

In [4]:
```
data.columns
```

Out[4]:
```
Index(['Name', 'Platform', 'Year_of_Release', 'Genre', 'NA_sales', 'EU_sales',
       'JP_sales', 'Other_sales', 'Critic_Score', 'User_Score', 'Rating'],
      dtype='object')
```

In [5]:
```
# приводим все названия столбцов к нижнему регистру:
data.columns = [x.lower() for x in data.columns]
data.columns
```

Out[5]:
```
Index(['name', 'platform', 'year_of_release', 'genre', 'na_sales', 'eu_sales',
       'jp_sales', 'other_sales', 'critic_score', 'user_score', 'rating'],
      dtype='object')
```

Все названия приведены к единому стилю, чтобы в дальнейшем избежать путаницы при обращении к столбцам

Обработка пропусков и изменение типов данных

Прежде чем изменить типы данных, необходимо обработать проупуски в столбцах. Посмотрим на количество пропусков в каждом:

In [6]:
```
data.isna().sum()
```

Out[6]:
```
name                  2
platform              0
year_of_release     269
genre                 2
na_sales              0
eu_sales              0
jp_sales              0
other_sales           0
critic_score       8578
user_score         6701
rating             6766
dtype: int64
```

Имеются пропуски в 5 столбцах. Избавляться от строк с пропусками нерационально - в них могут храниться другие интересующие нас данные, более того, в столбце *critic_score* пропущены значения для 50% данных.

Обработаем пропуски следующим образом:

- **name** и **genre** - заменим пропуски на 'Unknown'
- **year_of_release** - заменим пропуски на условный маркер '0'

In [7]:
```
data[['name', 'genre']] = data[['name', 'genre']].fillna('Unknown')
data['year_of_release'] = data['year_of_release'].fillna('0')
```

Остались пропуски в столбцах с оценками и рейтингом. Посмотрим на уникальные значения в столбцах **critic_score, user_score и rating**:

In [8]:
```
print('Значения в столбце "critic_score":', data['critic_score'].sort_values().unique())
print()
print('Значения в столбце "user_score":',data['user_score'].sort_values().unique())
print()
print('Значения в столбце "rating":', data['rating'].sort_values().unique())
```

```
Значения в столбце "critic_score": [13. 17. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30. 3
1. 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]

Значения в столбце "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' 'tbd' nan]

Значения в столбце "rating": ['AO' 'E' 'E10+' 'EC' 'K-A' 'M' 'RP' 'T' nan]
```

- **critic_score** - помимо чисел содержит NaN. С NaN можно проводить математические операции, т.к. такие значения принадлежат к типу float, оставим пропуски в этом столбце
- **user_score** - как и *critic_score* содержит NaN, такие пропуски оставляем, но помимо этого встречается значение **'tbd'**, что на русский можно перевести как "подлежит определению". Видимо, пользовательская оценка для таких строк еще не определена, *заменим tbd на 'Nan'*, чтобы изменить в столбце тип данных на числовой и далее проводить математические операции с ним
- **rating** - содержит строковые значения, заменим пропуски на 'undefined' ('не определен')

In [9]:
```
data['user_score'] = data['user_score'].replace('tbd', np.NaN)
data['rating'] = data['rating'].fillna('undefined')
data.isna().sum()
```

Out[9]:
```
name                 0
platform             0
year_of_release      0
genre                0
na_sales             0
eu_sales             0
jp_sales             0
other_sales          0
critic_score      8578
user_score        9125
rating               0
dtype: int64
```

После обработки пропусков можем изменить типы данных в следующих столбцах:

- **year_of_release** - хранит целочисленные значения, заменим на соответствующий тип 'int'
- **user_score** и **critic_score** - теперь хранят только числовые значения, заменим на тип 'float':

In [10]:
```
data['year_of_release'] = data['year_of_release'].astype('float64', errors='ignore')
data['year_of_release'] = data['year_of_release'].astype('int64', errors='ignore')
data['critic_score'] = data['critic_score'].astype('float64', errors='ignore')
data['user_score'] = data['user_score'].astype('float64', errors='ignore')
data.info()

#проверим дубликаты:
print()
print('Число строк-дубликатов:', data.duplicated().sum())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16715 entries, 0 to 16714
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   name             16715 non-null  object 
 1   platform         16715 non-null  object 
 2   year_of_release  16715 non-null  int64  
 3   genre            16715 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           16715 non-null  object 
dtypes: float64(6), int64(1), object(4)
memory usage: 1.4+ MB

Число строк-дубликатов: 0
```

Для дальнейшего анализа потребуются данные о **суммарных продажах по всем регионам**. Запишем их в отдельный столбец **total_sales**:

In [11]:
```
data['total_sales'] = data['na_sales'] + data['eu_sales'] + data['jp_sales'] + data['other_sa
data.head()
```

Out[11]:

	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	
1	Super Mario Bros.	NES	1985	Platform	29.08	3.58	6.81	0.77	NaN	
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	

Вывод

- все столбцы приведены к единому стилю
- набор данных содержал пропуски в столбцах name, genre, year_of_release, critic_score, user_score, rating
  - пропуски в *name*, *genre* и *year_of_release* заменены на значения-заглушки
  - *critic_score* и *user_score* "лидеры" по числу пропусков - больше половины пропущенных значений в каждом из столбцов. Такой объем пропусков нерационально заменять на медианные и средние значения, остались пропуски с типом NaN (такой тип позволяет

проводить математические операции со столбцом). Значение *tbd* ("подлежит определению") также заменено на NaN

- изменены типы данных на корректные
- дубликатов нет
- добавлен столбец *total_sales*, содержащий суммарные продажи по всем регионам

Данные обработаны и готовы к проведению исследовательского анализа

Исследовательский анализ данных

- посмотрим, сколько игр выпускалось в разные годы, изучим как менялись продажи по платформам и исходя из этого анализа определим характерный срок устаревания платформы
- определим актуальный период, который сможет построить прогноз на 2017 год, отбросив данные устаревших платформ
- определим платформы, которые лидируют, растут и падают по продажам, определим потенциально прибыльные платформы
- на примере одной из платформ проверим, влияют ли оценки пользователей и критиков на продажи
- проанализируем распределение игр по жанрам, выявим самые прибыльные жанры

Сколько игр выпускалось в разные годы?

In [12]:
```
data.groupby('year_of_release')['year_of_release'].count()\
        .plot(kind='bar',
              xlabel='Год релиза',
              ylabel='Кколичество игр',
              title='Количество релизов по годам',
              figsize=(10,5));
```

Имеем достаточно большой размах данных: с 1980 по 216 годы. Ожидаемо видим около 250 игр на отметке "0" - это игры, для которых не известен год релиза.

Вплоть до 1995 количество выпущенных игр не превышало 200 в год.

2002 год можем считать первым прорывным годом для мира игр: если в 2001 году выпустилось чуть более 400 игр, то в 2002 их число выросло до 800+. В 2003-2004 релизов становится чуть меньше, но начиная с 2005 игр выпускается все больше и больше.

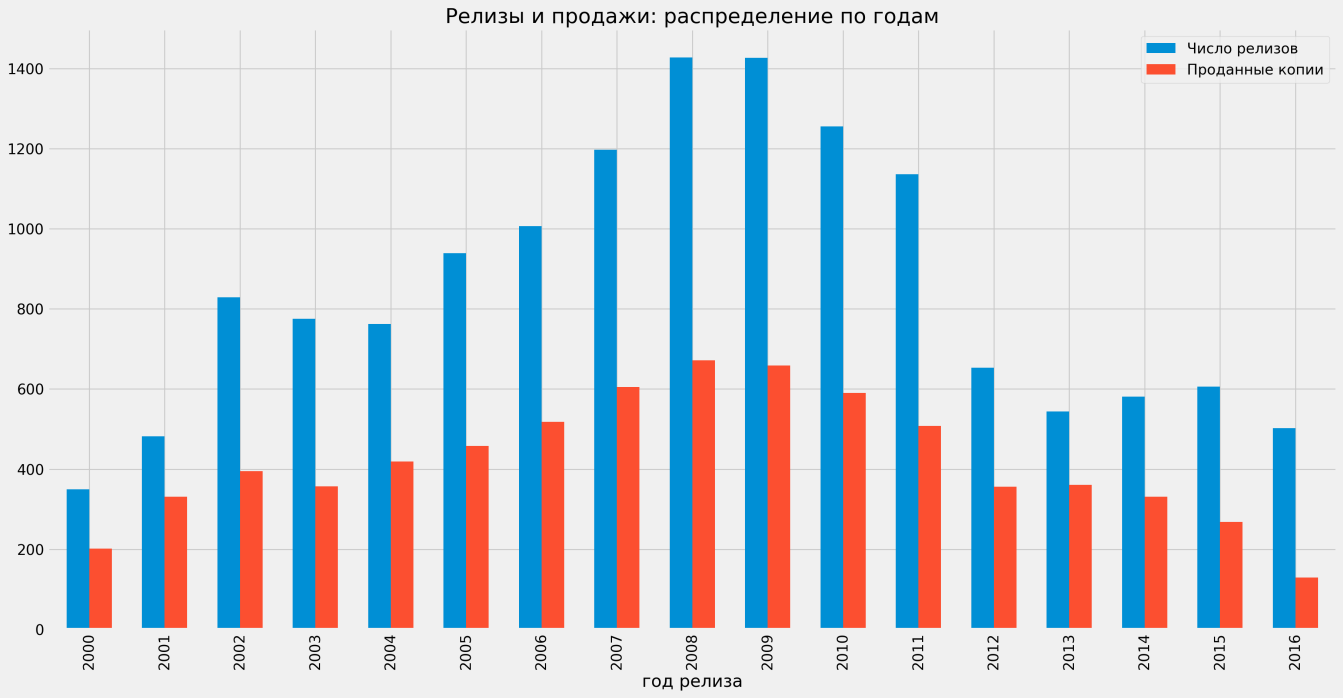
2008-2009 - настоящий пик в игровой индустрии. В каждый из этих годов выходит более 1400 игр, абсолютный рекорд в нашем наборе данных.

Далее наблюдаем **спад** - в 2011 число релизов уже не достигает 1200, в 2012 выходит всего 600+ игр. Очевидно, рынок перенасытился, возможно, со временем отсеялись некоторые разработчики игр, не удержавшись на плаву во время большой конкуренции в 2008-2009. Да и в целом игры стали намного сложнее технически, пользователи уже скорее не гонятся за разнообразием игр, а **отдают предпочтение более масштабным и продуманным проектам** ("лучше купить одну хорошую игру, чем несколько средних").

Также стоит обратить внимание на **развитие смартфонов:** начиная с 2010-го они становились все более доступными для каждого, разработка игр для смартфонов также не стоит на месте. Возможно, некоторые любители портативных приставок теперь отдают предпочтение играм на мобильном.

Посмотрим, как соотносятся данные о релизах и суммарных продажах. Ограничимся данными текущего тысячелетия:

In [13]:
```
data.query('year_of_release >= 2000').groupby('year_of_release').agg({'year_of_release':'coun
.plot(kind='bar', figsize=(20,10), width=0.6, xlabel='год релиза', title='Релизы и продажи: р

plt.legend(['Число релизов', 'Проданные копии']);
```



В целом, ничего аномального: данные о числе проданных копий преимущественно распределяются в соответствии с числом выпущенных игр

Анализ платформ

Как менялись продажи по платформам? Жизненный цикл платформы

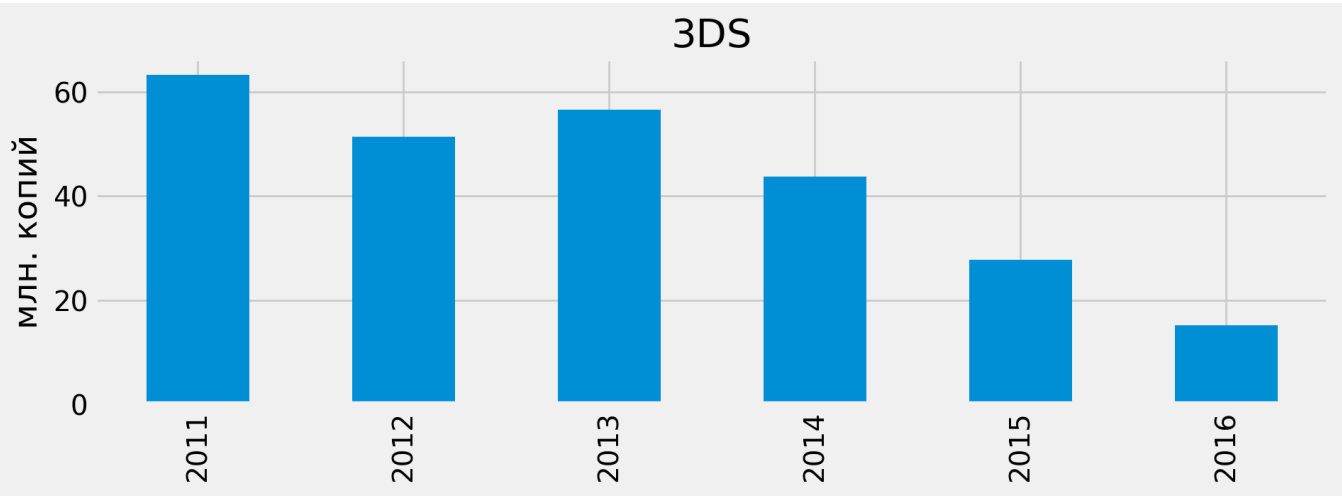
Посмотрим на распределение продаж за весь изучаемый период по платформам

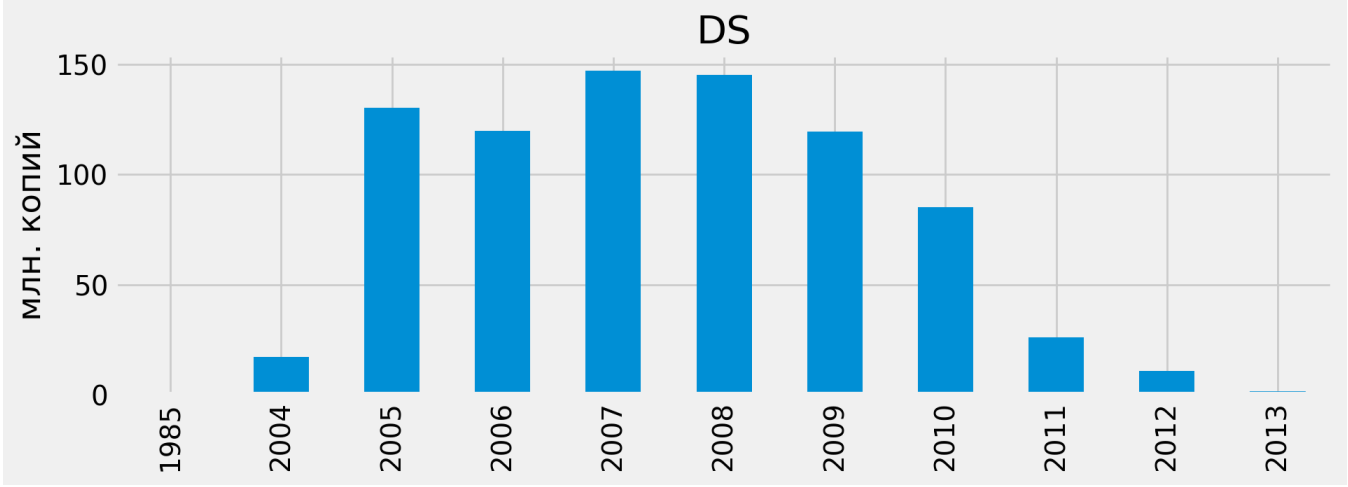
In [14]:
```
data.groupby('platform')['total_sales'].sum().sort_values()\
.plot(kind='bar', figsize=(10,5), xlabel='', ylabel='млн.копий',title='Продажи по платформам'
```

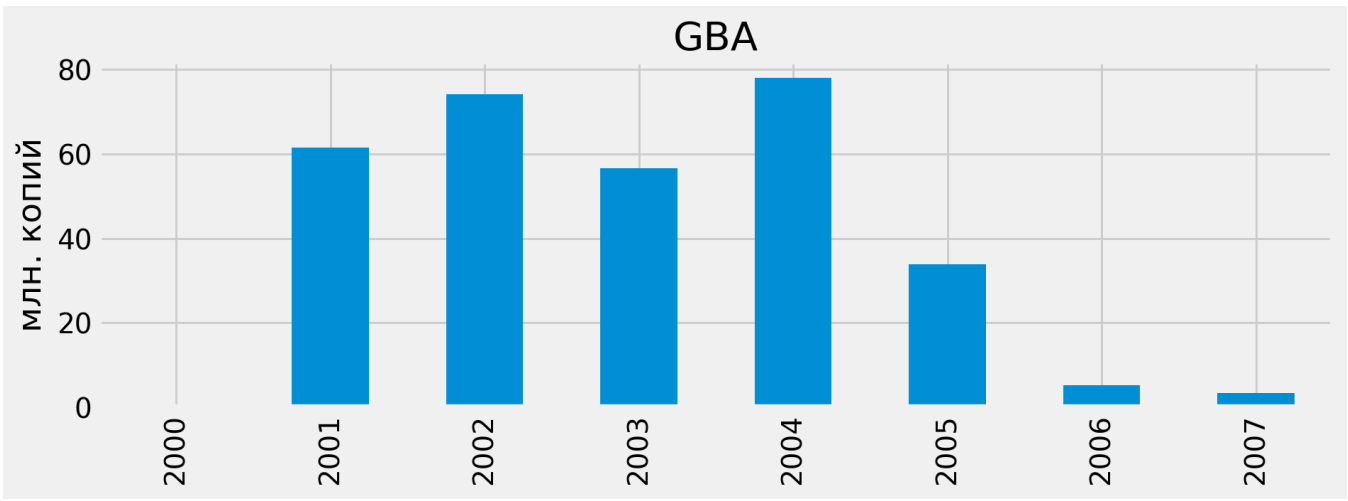
PS2 абсолютный лидер по числу проданных копий - более 1200 млн., в то время как ни одна другая платформа не достигает отметки в 1000. Но на этом же графике мы видим и PS, и PS3, и PS4, т.е. платформа уже сменила несколько поколений, и, наверное, **не стоит считать PS2 самой перспективной платформой** только исходя из общего числа проданных копий - вероятно она уже устарела и больше не обновляется.

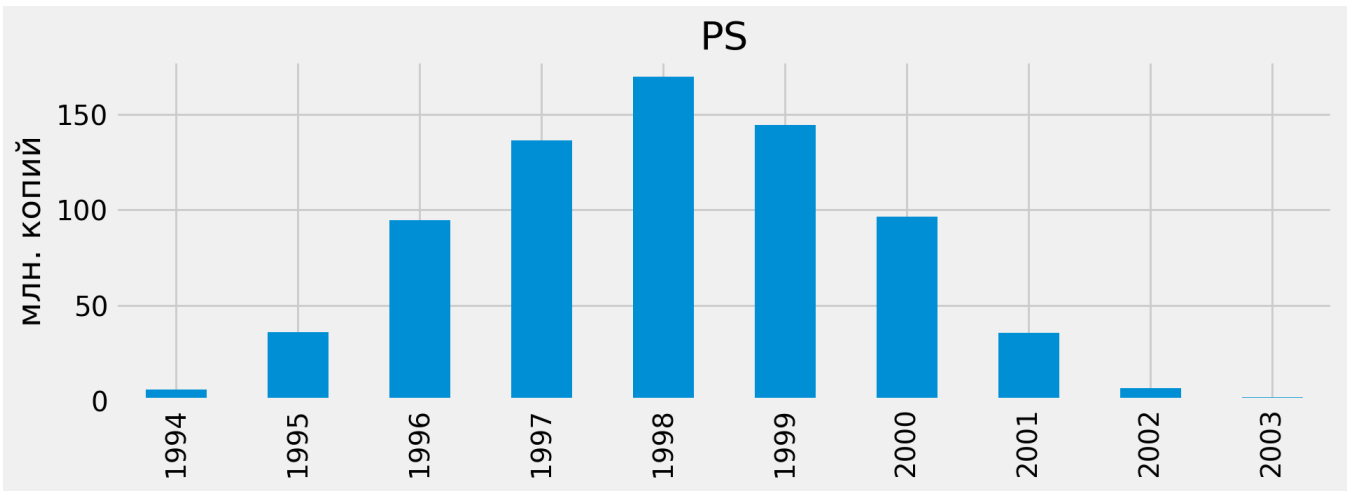
Посмотрим на графики для десятки самых прибыльных платформ: сколько копий было продано в каждый год

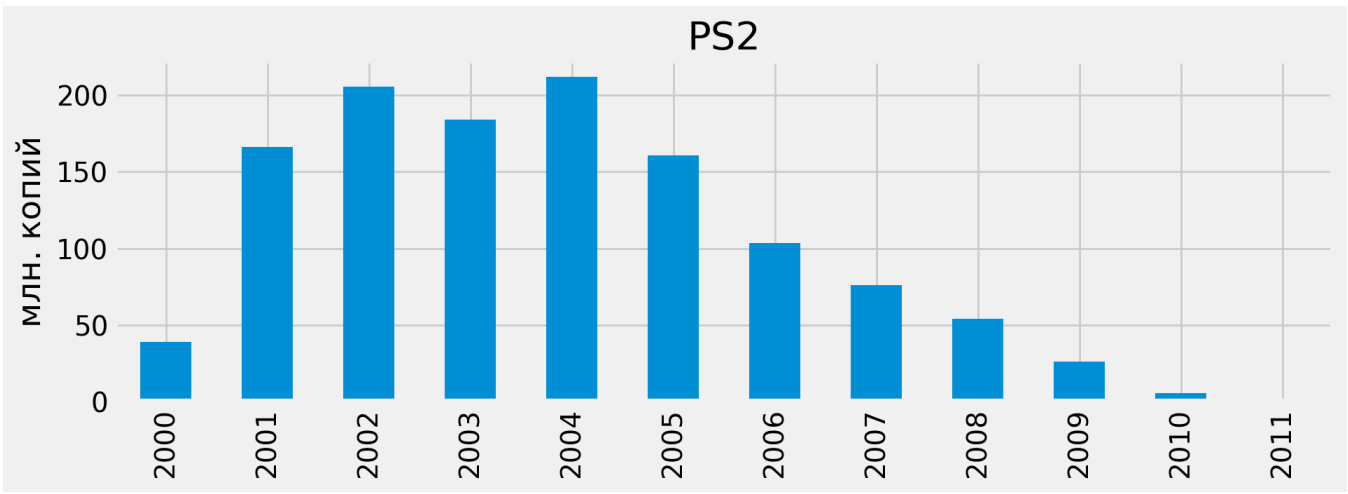
In [16]:
```
# Удалим из набора данных игры, для которых неизвестен год релиза:
data.drop(index=data.query('year_of_release==0').index,inplace=True)

# группируем данные: суммируем продажи по каждой платформе, сортируем по убыванию и ограничив
platform_top = data.pivot_table(index = 'platform', values = 'total_sales', aggfunc = 'sum')\
                   .sort_values(by='total_sales', ascending=False).head(10)
# сохраняем названия платформ в переменной 'top_10_name'
top_10 = data.query('platform in @platform_top.index')
top_10_name = top_10['platform'].sort_values().unique()

# Функция для построения графиков для каждой платформы из top_10_name:
for platform in top_10_name:
    data[data['platform'] == platform].pivot_table(index = 'year_of_release', values = 'total
    .plot(kind = 'bar', figsize=(10,3), xlabel='', ylabel='млн. копий',legend=False)
    plt.title(platform)
```

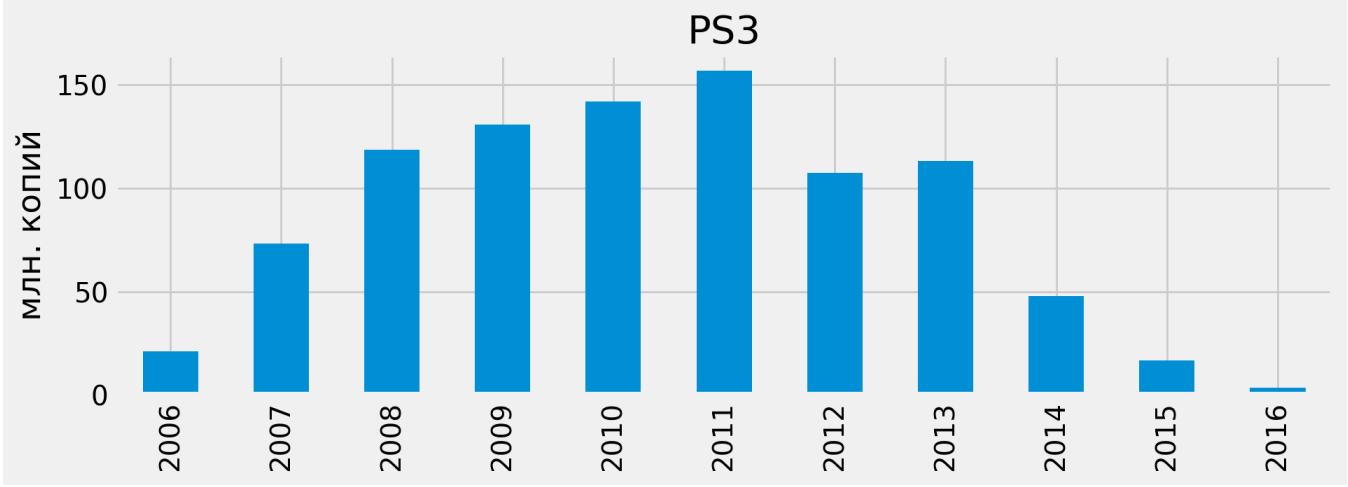


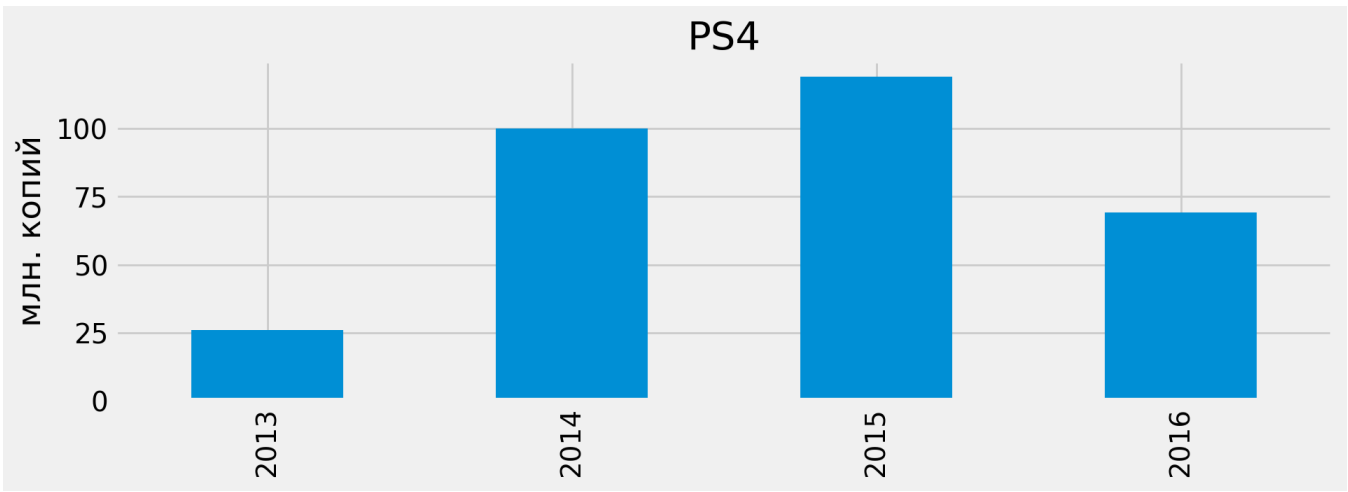

DS
млн. копий
150
100
50
0
1985
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013

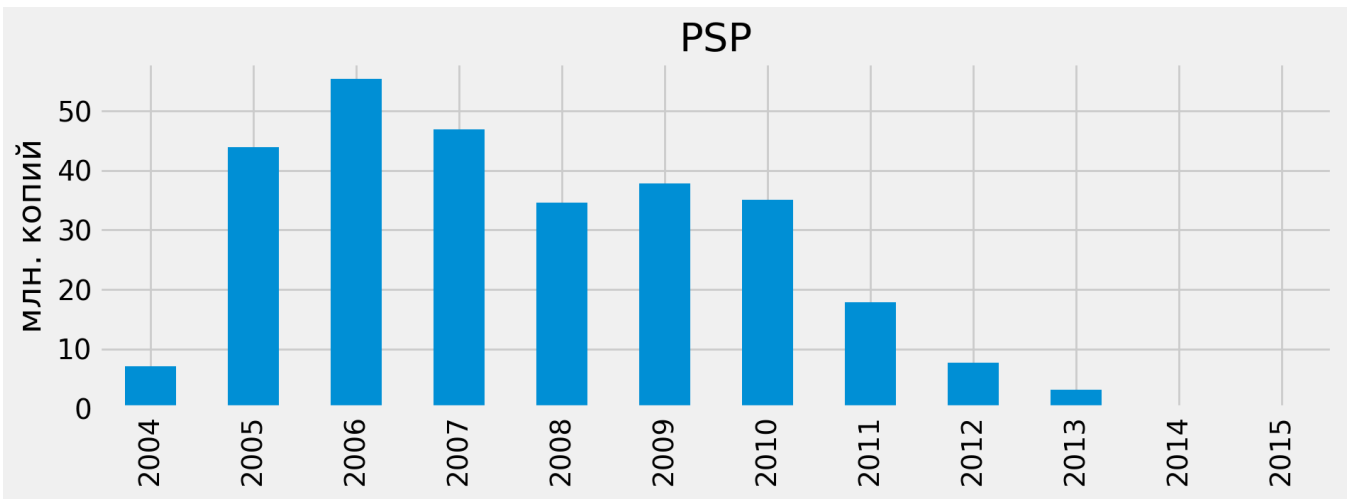

GBA
млн. копий
80
60
40
20
0
2000
2001
2002
2003
2004
2005
2006
2007

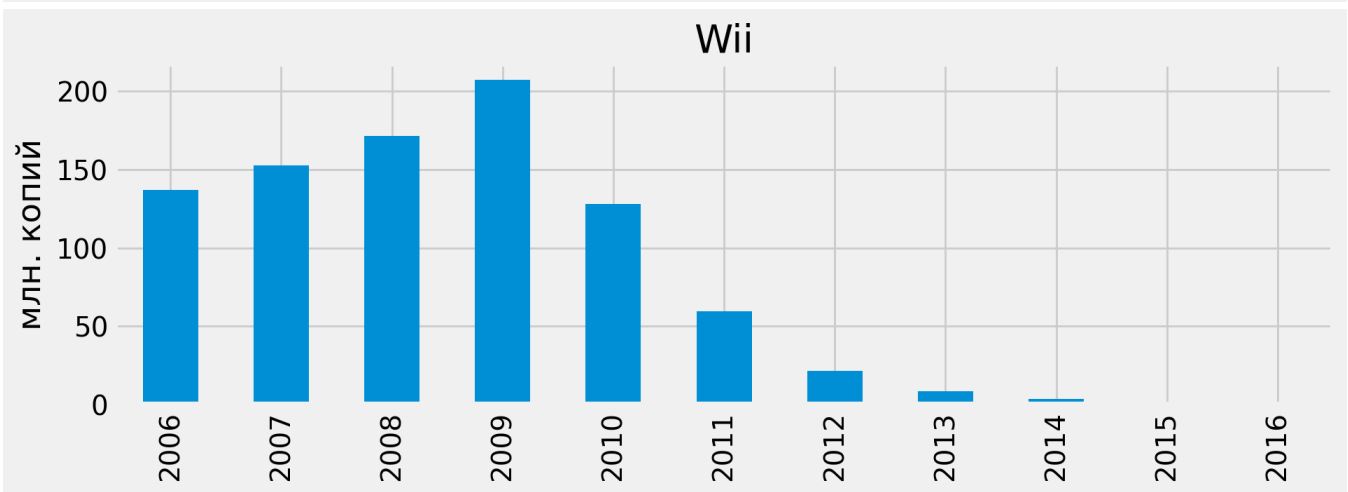

PS
млн. копий
150
100
50
0
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003

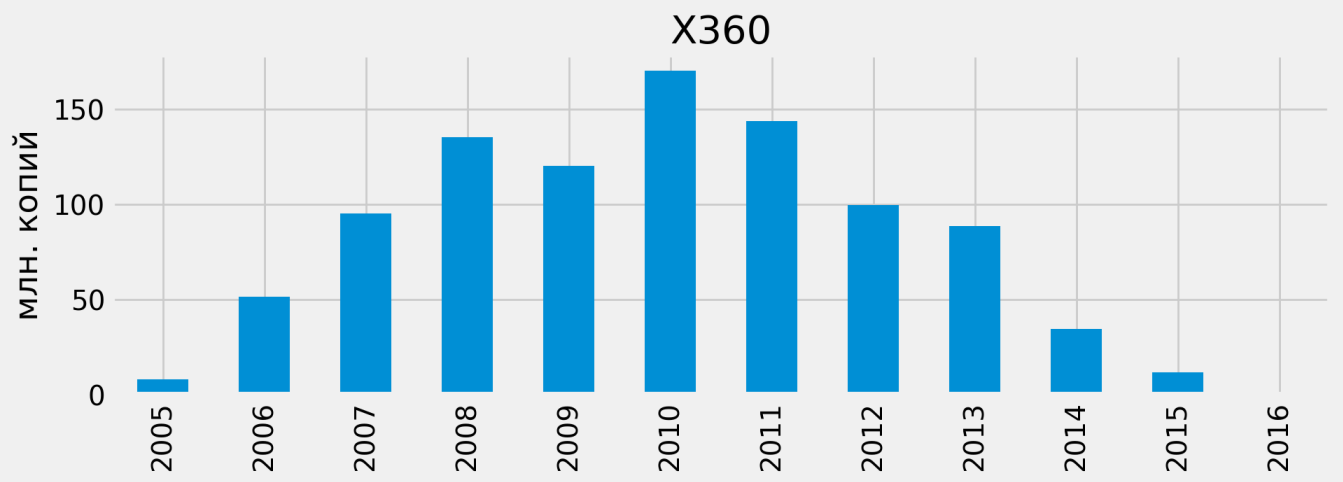

PS2
млн. копий
200
150
100
50
0
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011


PS3
млн. копий
150
100
50
0
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016


PS4
млн. копий
100
75
50
25
0
2013
2014
2015
2016


PSP
млн. копий
50
40
30
20
10
0
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015


Wii
млн. копий
200
150
100
50
0
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016

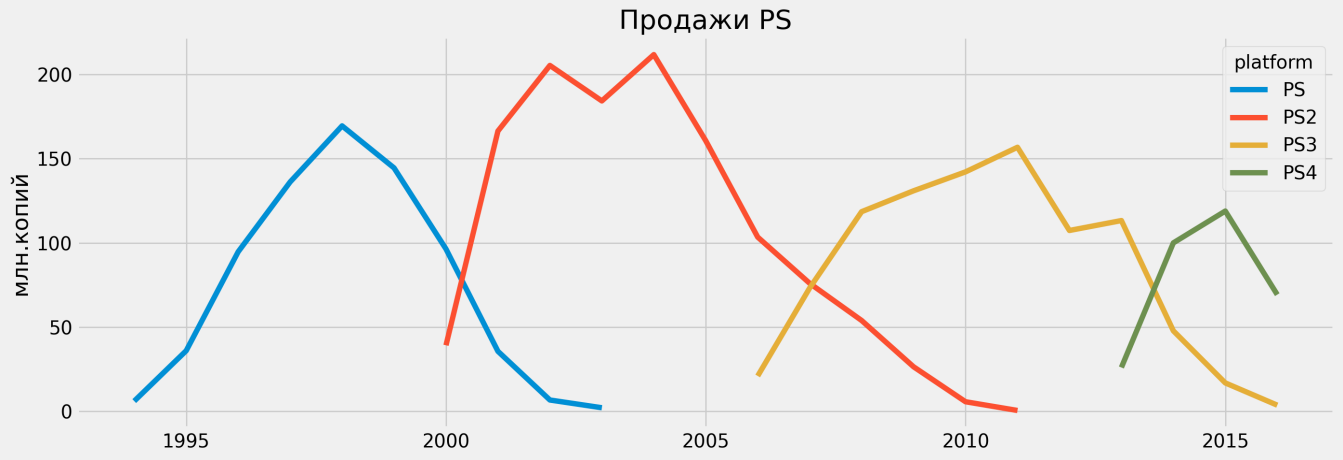


- **3DS** - первая игра вышла в 2011 году, игры продаются и в 2016
- **DS** - если не учитывать продажи в 1985 году, "прожила" 9 лет, после 2013 года игры для платформы не покупались
- **GBA** - игры для платформы продавались 7 лет, 2000-2007 годы.
- **PS**:
    - первое поколение зарабатывало на играх 9 лет, после 2003 года не было ни одной продажи
    - PS2 очевидно, пришла на смену PS, "прожила" 11 лет, а после 2011 игры перестали покупать
    - PS3 появилась в 2006 году, теперь уже на смену PS2. В 2006-2011 продажи PS2 заметно сокращаются, а у PS3 наоборот растут. В 2016 игры еще продаются, но уже в ничтожно малом количестве
    - PS4 самая молодая приставка из топ-10, игры начали продаваться только в 2013 году, в 2016, разумеется, игры еще продаются
- **PSP** - "прожила" 11 лет, с 2004 по 2015, в 2016 уже не приносит доход
- **Wii** - начала свой старт в 2006 году, на 2016 год имела ничтожно малое число продаж, можно считать устаревшей платформой
- **X360** - игры продаются с 2005 года, на 2016 год практически не имеет продаж.

Можно сказать, что даже самые популярные **платформы в среднем "живут" около 10 лет**, после чего либо пользователи теряют к ним интерес, либо, что более вероятно, просто **устаревают и не могут поддерживать новые игры**

Имея в наборе данных 4 платформы одного бренда (PS), можем рассмотреть как разные поколения приставок сменяют друг друга:

In [17]:
```
# лист с названиями платформ PS:
ps_list = ['PS', 'PS2', 'PS3', 'PS4']

# построение графика для четырех платформ:
data.query('platform in @ps_list')\
.pivot_table(index= 'year_of_release', columns = 'platform', values='total_sales', aggfunc='s
.plot(figsize=(15, 5),
     xlabel='',
     ylabel='млн.копий',
     title='Продажи PS');
```



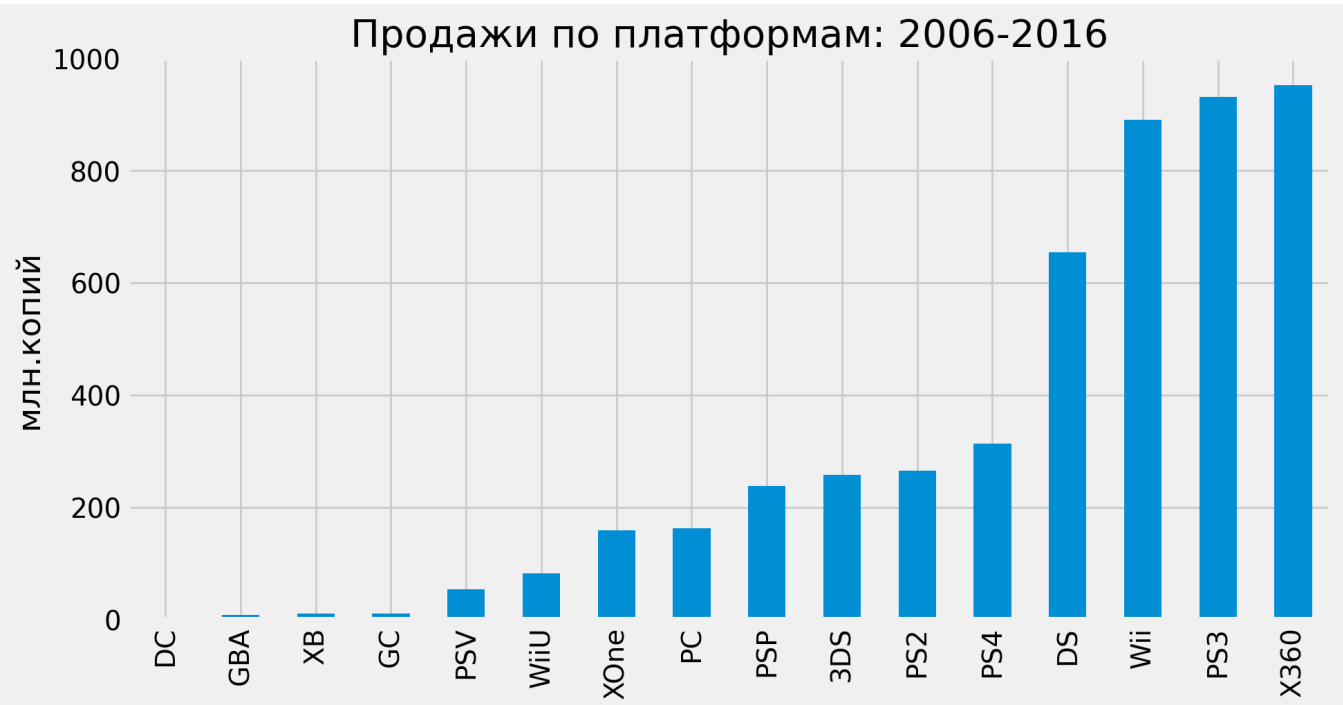
Жизненный цикл каждого поколения - около 10 лет, более новое поколение сменяет предыдущее.

Можем предположить, что активная разработка нового поколения начинается примерно на пике популярности предыдущего. Новое поколение выходит на рынок тогда, когда предыдущее показывает снижение на ~50% от пиковых значений

Мы определили, что **нельзя опираться на суммарные продажи для определения перспективных платформ**: большинство лидеров по числу проданных копий уже не обновляются и для них не выпускаются игры.

Основываясь на установленном жизненном цикле в 10 лет, посмотрим на распределение продаж по платформам за **последние десять лет**, возможно, мы выявим новых лидеров:

In [18]:
```
data.query('year_of_release >= 2006').groupby('platform')\
.agg({'total_sales':'sum'}).sort_values(by = 'total_sales')\
.plot(kind='bar',
      figsize=(10,5),
     title='Продажи по платформам: 2006-2016',
     ylabel='млн.копий',
     xlabel='',
     legend=False);
```



Предыдущий лидер - PS2 - опустился на 6 место, а в 10-ку самых прибыльных теперь входят XOne и PC. В целом **количество приставок сократилось до 16.**

Предыдущий топ-10 уже не актуален, многие платформы уже не показывают прошлых результатов.

Опираясь на результаты платформ-лидеров, можем заметить, что чаще всего **платформа находится в периоде финального роста спустя 3-4 года после своего релиза**. Для дальнешего анализа нам нужно оставить только те платформы, которые в 2017 году будут находиться в периоде этого роста или выйдут на плато. Интерес к более старым моделям неизмененно будет спадать.

Ограничимся данными с 2014 включительно, тем самым отсеяв неперспективные платформы. Перезапишем срез в новую таблицу *games*, с которой и будем работать далее.

In [19]:
```
games = data.query('year_of_release >= 2014')
print('Количество актуальных платформ:', len(games['platform'].unique()))
print('Список актуальных платформ:', games['platform'].sort_values().unique().tolist())
```

```
Количество актуальных платформ: 10
Список актуальных платформ: ['3DS', 'PC', 'PS3', 'PS4', 'PSP', 'PSV', 'Wii', 'WiiU', 'X360',
'XOne']
```

Далее будем работать только с актуальными платформами.

Вывод

Характерный "срок жизни" платформы -10 лет.

Период финального роста или выхода на плато: ~3 года с момента релиза

Для дальнейшего анализа оставили только актуальную информацию в новой таблице *games*: далее **будем рассматривать только данные за последние 3 года по перспективным платформам**

Актуальные платформы: определение лидеров

Необходимо определить какие платформы еще имеют перспективы, а какие уже устаревают. Посмотрим на суммарные продажи и количество выпущенных игр по каждой платформе:

In [20]:
```
games.groupby('platform').agg({'total_sales':'sum', 'name':'count'}) \
    .sort_values(by = 'total_sales', ascending = False)
```

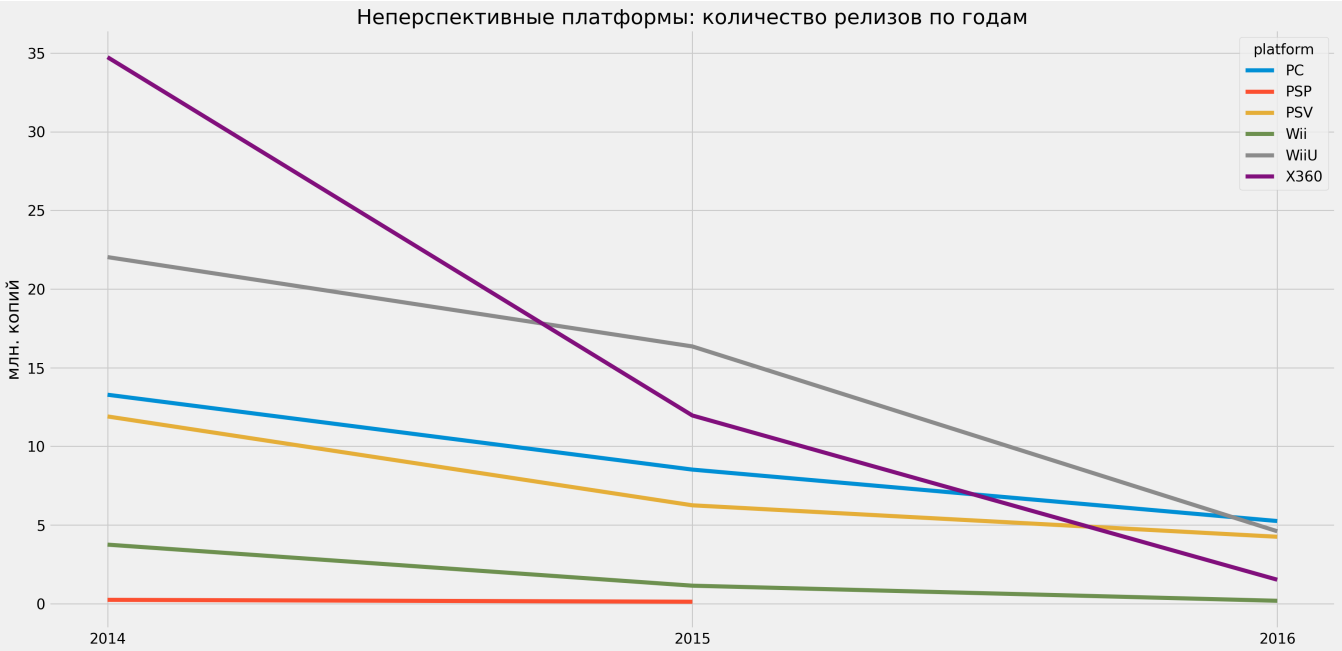
Out[20]:

platform	total_sales	name
PS4	288.15	376
XOne	140.36	228
3DS	86.68	212
PS3	68.18	219
X360	48.22	111
WiiU	42.98	73
PC	27.05	151
PSV	22.40	295
Wii	5.07	11
PSP	0.36	13

Лидер по числу проданных копий - **PS4**, для этой же платформы было выпущено больше всего игр. Интересно выбиваются PSV и PC - при относительно небольшом количестве суммарных продаж для обеих платформ выходит достаточно высокое число новых игр.

Но с точки зрения бизнеса нас в большей степени, конечно, интересуют именно продажи. Определим **"аутсайдерами" платформы с продажами <50 млн. копий, а более популярные - "лидерами"** и посмотрим на динамику их продаж.

In [22]:
```
#список предположиельно неперспективных платформ:
outsiders = ['X360', 'WiiU', 'PC', 'PSV', 'Wii', 'PSP']

games.query('platform in @outsiders').pivot_table(index='year_of_release',
                                                  columns = 'platform', values='total_sales',
    .plot(figsize=(20,10), ylabel='млн. копий', xlabel = '', title='Неперспективные платформы
plt.xticks([2014, 2015, 2016]);
```



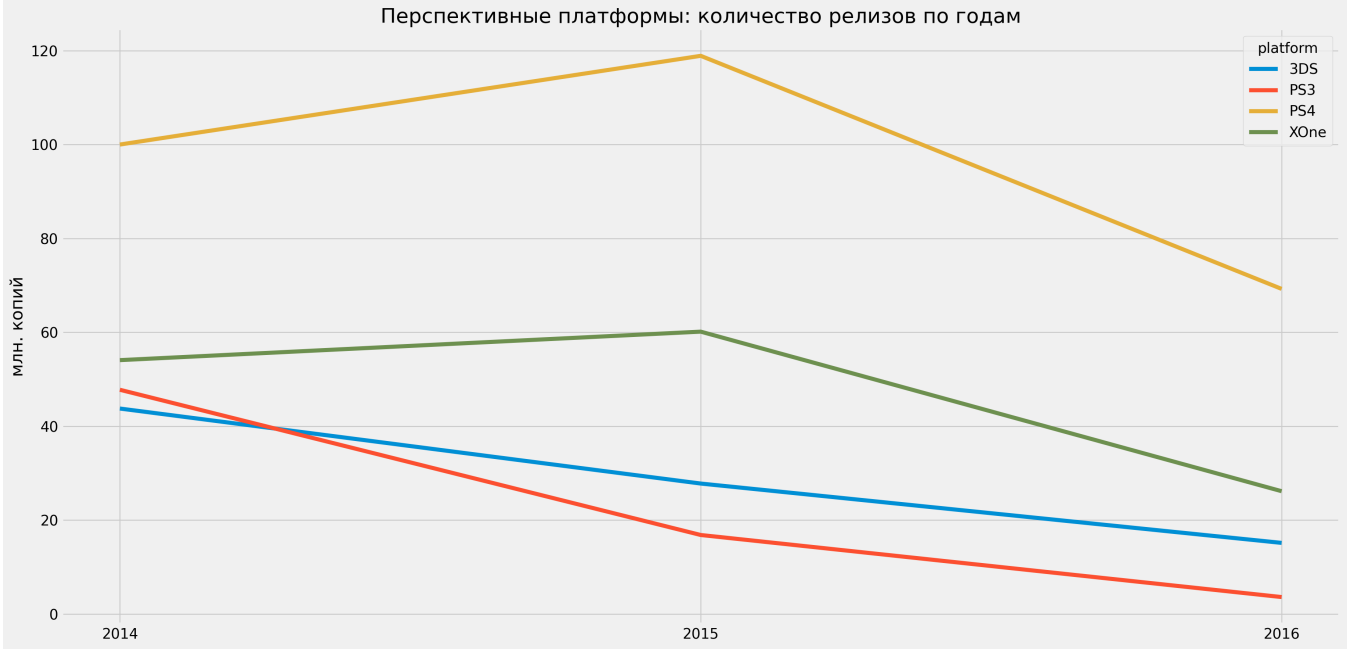
Ожидаемо, все эти платформы пережили пик своей популярности и к 2016 году ни для одной из них не продается более 10 млн копий(кроме PC, которая тоже сильно теряет популярность). Даже с учетом того, что данные на 2016 год у нас могут быть неполными, отрицательный тренд, увы, прослеживается. Разумеется, **рассчитывать на успех в 2017 году этим платформам не стоит.**

PSP можно удалить из набора данных - продажи в 2014-2015 сливаются с нулем, в 2016 году продаж уже нет вовсе.

In [23]:
```
games = games.query('platform != "PSP"')
```

Куда больший интерес для нас представляют **перспективные платформы,** посмотрим на их динамику продаж:

In [24]:
```
#список предположительных лидеров:
lidears = ['PS4', 'XOne', '3DS', 'PS3']

games.query('platform in @lidears').pivot_table(index='year_of_release', columns = 'platform'
    .plot(figsize=(20,10), ylabel='млн. копий', xlabel='', title='Перспективные платформы: ко
plt.xticks([2014, 2015, 2016]);
```



Явные лидеры рынка: PS4 и XOne. Результат 3DS слабее, меньше 20 млн. копий в 2016 году. PS3 выделяется среди этих платформ в негативном ключе, что неудивительно - лидирует новая PS4, смещая свою предшественницу. PS3 к 2017 году скорее полностью потеряет свою актуальность.

Вывод

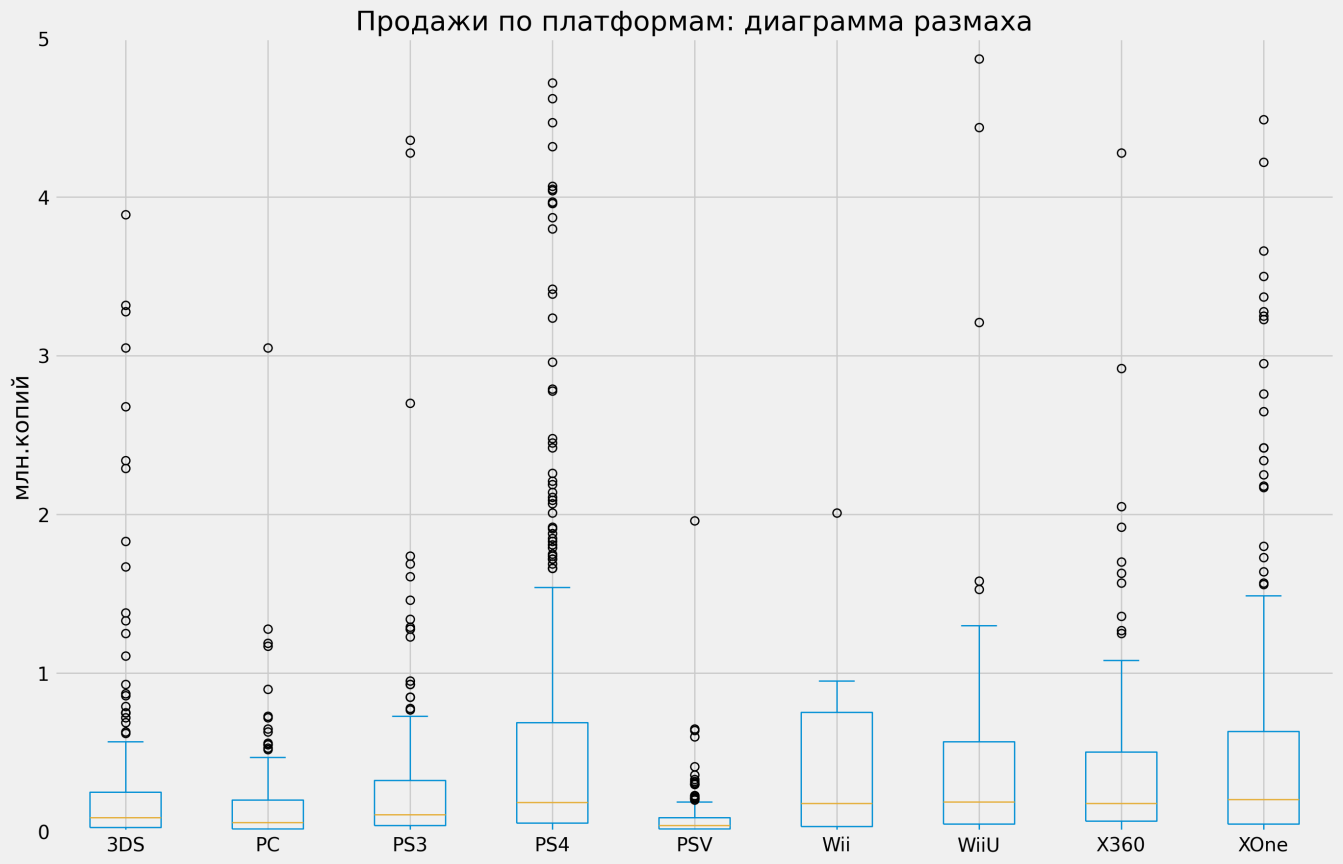
К перспективным платформам отнесем:

- **PS4** - ~70 млн. проданных копий в 2016, "молодая" платформа, **прогнозируемый период "жизни" - до 2023**,
- **Xone** - ~30 млн. проданных копий в 2016, "молодая" платформа, **прогнозируемый период "жизни" - до 2023**,
- **3DS** - уже в менеьшей степеени, ~ 18 млн проданных копий в 2016, старше двух предыдущих, **прогнозируемый период "жизни" - до 2021**

Все остальные платформы считаем неперспективными: продажи для них не превышают 10 млн. копий. PSP больше не учитываем в анализе - новых игр в 2016 году для нее больше не выходит.

Глобальные продажи по платформам: размах данных

In [25]:
```
games.pivot_table(index = 'name', columns = 'platform', values = 'total_sales', aggfunc = 'su
.plot(kind='box',  ylim=(0,5), figsize=(15,10), ylabel='млн.копий', title='Продажи по платфор
```



Самые высокие медианы у **PS4, XOne, Wii и WiiU**, при этом у последних двух видим крайне малое число "выбросов": у Wii есть всего одна игра, проданная тиражом 2 млн копий, все остальные игры не продаются больше, чем 1млн копий, у WiiU имеется уже 5 таких игр-хитов, но это по-прежнему очень мало. У **PS4 и XOne результаты куда выше** - обе платформы имеют много игр-бестселлеров.

PS4 - абсолютный лидер по количеству игр, проданных тиражом > 2 млн

3DS при относительно низкой медиане имеет достаточное число игр-бестселлеров, замыкает тройку лидеров, уступая PS4 и Xone.

Интересно наблюдать, как X360 и PS3 сдают свои позиции, уступая более новым моделям из их линеек.

PSv очевидный аутсайдер - одна из самых низких медиан, максимальный тираж игры не достигает 2 млн копий. PC имеет только одну популярную игру с тиражом 3 млн. копий, в остальном показывает резуьтат ниже среднего.

Вывод

Тройка лидеров, ранее определенная новизной и объемом продаж остается прежней: **PS4, XOne и 3DS**.

PS4 - лидер по числу хитов за последние пять лет, **имеет больше всех игр, проданных тиражом > 2 млн**

Влияние отзывов на продажи

Узнаем, зависит ли оценка критиков или пользователей на продажи. В столбцах с оценками содержится много пропусков, определим самую "заполненную" платформу, чтобы далее расмотреть диаграммы для нее:

In [26]:
```
print('Отзывов критиков всего:')
print(games.groupby('platform')['critic_score'].count().sort_values(ascending=False))
print()
print('Отзывов игроков всего:')
print(games.groupby('platform')['user_score'].count().sort_values(ascending=False))
```

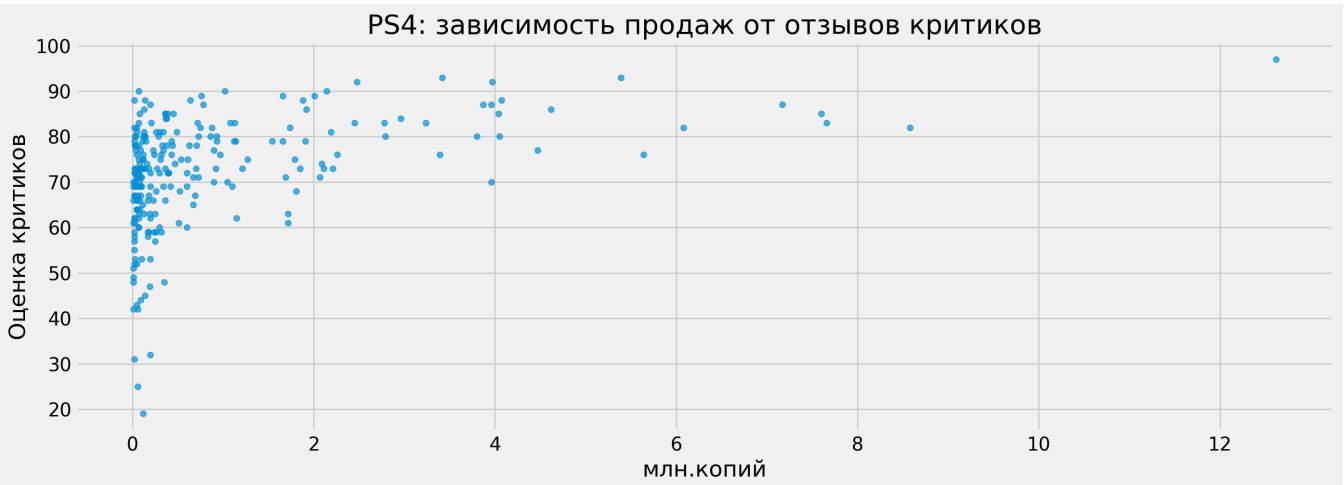
```
Отзывов критиков всего:
platform
PS4     237
XOne    155
PC      116
PSV      54
3DS      51
WiiU     43
PS3      36
X360     26
Wii       0
Name: critic_score, dtype: int64

Отзывов игроков всего:
platform
PS4     242
XOne    165
PC      122
PS3      98
X360     83
PSV      73
3DS      55
WiiU     48
Wii       2
Name: user_score, dtype: int64
```

У PS4 больше всего и отзывов критиков, и отзывов игроков. Посмотрим на корреляцию для этой платформы

PS4: корреляция между продажами и оценками критиков

In [33]:
```
games.query('platform == "PS4"')\
    .plot(x='total_sales',
          y='critic_score',
          kind='scatter',
          figsize=(15,5),
          title='PS4: зависимость продаж от отзывов критиков',
         xlabel = 'млн.копий',
         ylabel = 'Оценка критиков',
         alpha=0.7)
print('Коэффициент корреляции между продажами и отзывами критиков:', \
      games.query('platform == "PS4"')['total_sales'].corr(games['critic_score']).round(3))
```

```
Коэффициент корреляции между продажами и отзывами критиков: 0.403
```



Наблюдаем **положительный коэффициент корреляции ~ 0.4.**

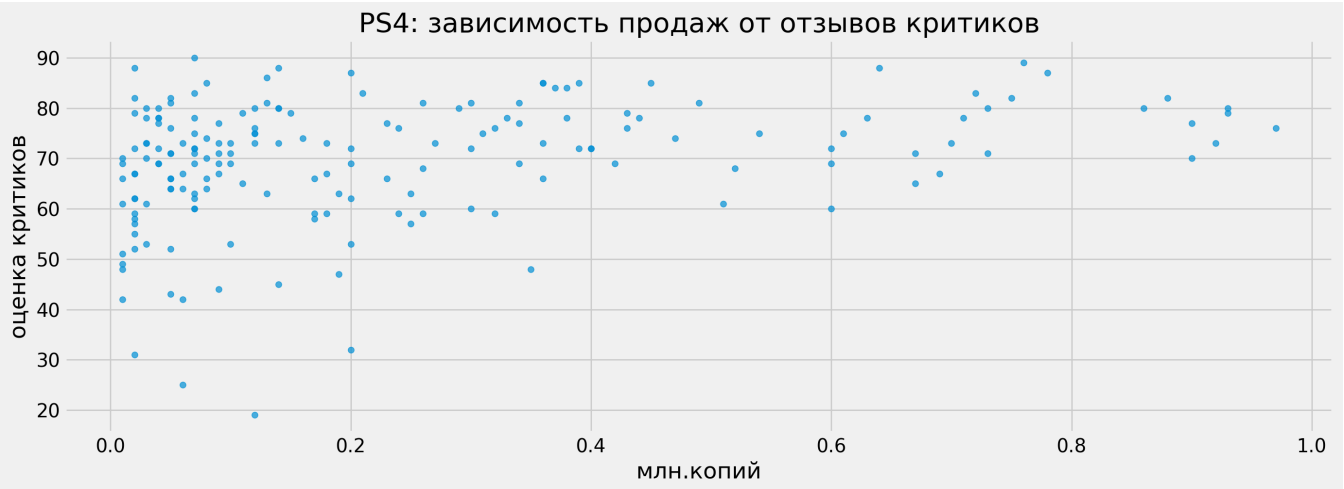
В целом можно выстроить такую цепочку: игры, проданные тиражом более 1 млн.копий, получали оценку выше 60, с тиражом 2-4 млн - оценка выше 70, 4-6 млн - оценка выше ~75, 6-10 млн - оценка выше 80. Самая популярная игра получила от критиков почти 100 баллов.

Другими словами, **для популярных игр** зависимость выражена очевидно: **чем выше оценка критиков, тем больше продаж**.

Но есть и другое наблюдение - критики нередко ставят высокие оценки непопулярным играм. Посмотрим на такое распределние прицельнее, ограничившись продажами в 1млн. копий

In [34]:
```
games.query('platform == "PS4" and total_sales < 1')\
    .plot(x='total_sales',
          y='critic_score',
          kind='scatter',
          figsize=(15,5),
          title='PS4: зависимость продаж от отзывов критиков',
         xlabel = 'млн.копий',
         ylabel = 'оценка критиков',
         alpha=0.7)
print('Коэффициент корреляции между продажами и отзывами критиков:', \
      games.query('platform == "PS4" and total_sales < 1')['total_sales'].corr(games['critic_
```

```
Коэффициент корреляции между продажами и отзывами критиков: 0.312
```



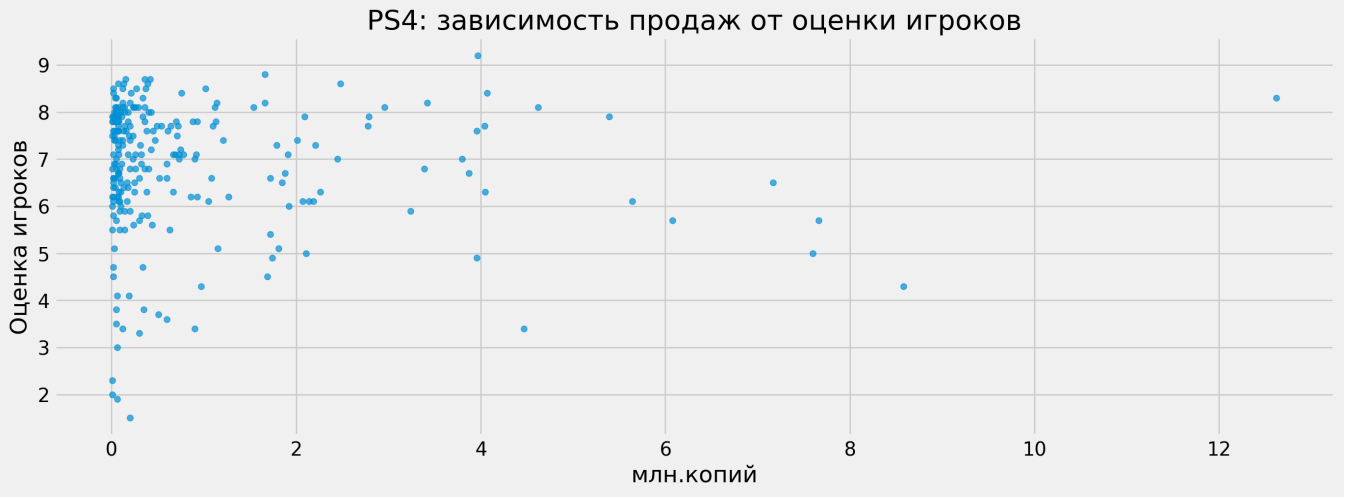
Для менее популярных игр такая зависимость выражена слабее. Игры, проаданные тиражом >400 тыс. все еще не получали оценки критиков ниже 60, но игры-аутсаудеры с тиражом < 400 тыс вполне могли получать очень высокие оценки, даже выше 80.

Иными словами, **критики могут одинаково высоко оценивать и игры-бестселлеры, и провалившиеся в продажах игры.**

PS4: корреляция между продажами и оценками игроков

In [35]:
```
games.query('platform == "PS4"')\
    .plot(x='total_sales',
          y='user_score',
          kind='scatter',
          figsize=(15,5),
          title='PS4: зависимость продаж от оценки игроков',
         xlabel = 'млн.копий',
         ylabel = 'Оценка игроков',
         alpha=0.7)
print('Коэффициент корреляции между продажами и отзывами критиков:', \
      games.query('platform == "PS4"')['total_sales'].corr(games['user_score']).round(3))
```

```
Коэффициент корреляции между продажами и отзывами критиков: -0.04
```

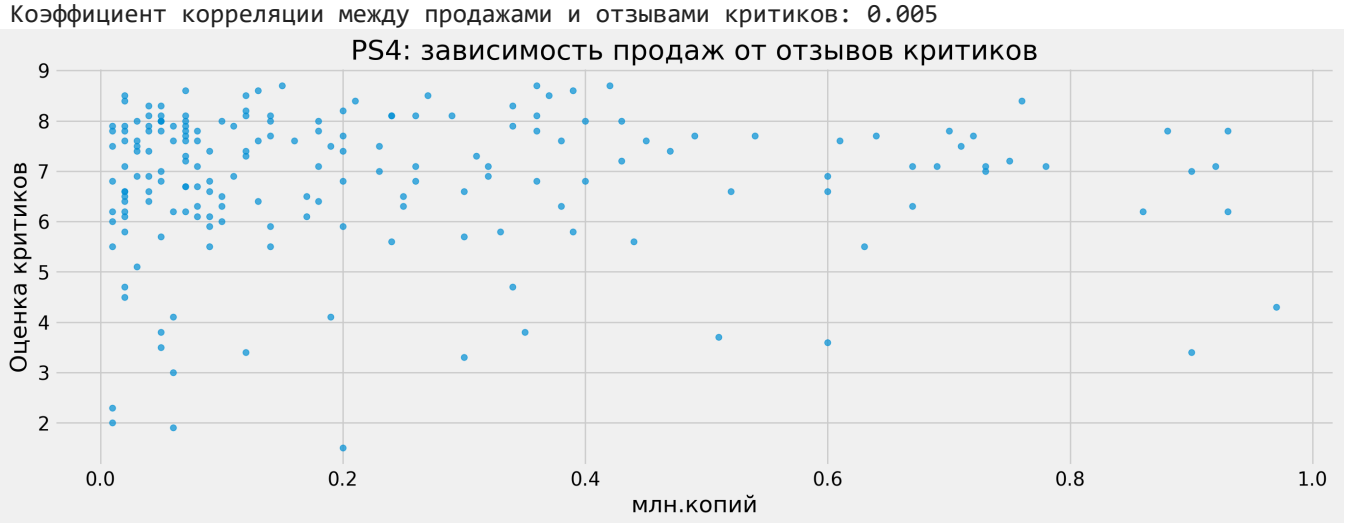


Здесь наблюдаем **крайне слабо выраженный отрицательный коэффициент корреляции: -0.04**

На диаграмме действительно наблюдается большой разброс - непопулярные игры могут получать высокие оценки (>8), бестселлеры - оценки ниже 6. Игроки, в отличие от критиков, куда чаще дают играм оценки ниже 5. Прицельно посмотрим на зависимость для непопулярных игр:

In [36]:
```
games.query('platform == "PS4" and total_sales < 1')\
    .plot(x='total_sales',
          y='user_score',
          kind='scatter',
          figsize=(15,5),
          title='PS4: зависимость продаж от отзывов критиков',
         xlabel = 'млн.копий',
         ylabel = 'Оценка критиков',
         alpha=0.7)
print('Коэффициент корреляции между продажами и отзывами критиков:', \
      games.query('platform == "PS4" and total_sales < 1')['total_sales'].corr(games['user_sc
```

```
Коэффициент корреляции между продажами и отзывами критиков: 0.005
```



Коэффициент зависимости уже положительный, но по-прежнему выражен очень слабо. Даже самые непопулярные игры часто получают оценки больше 7.

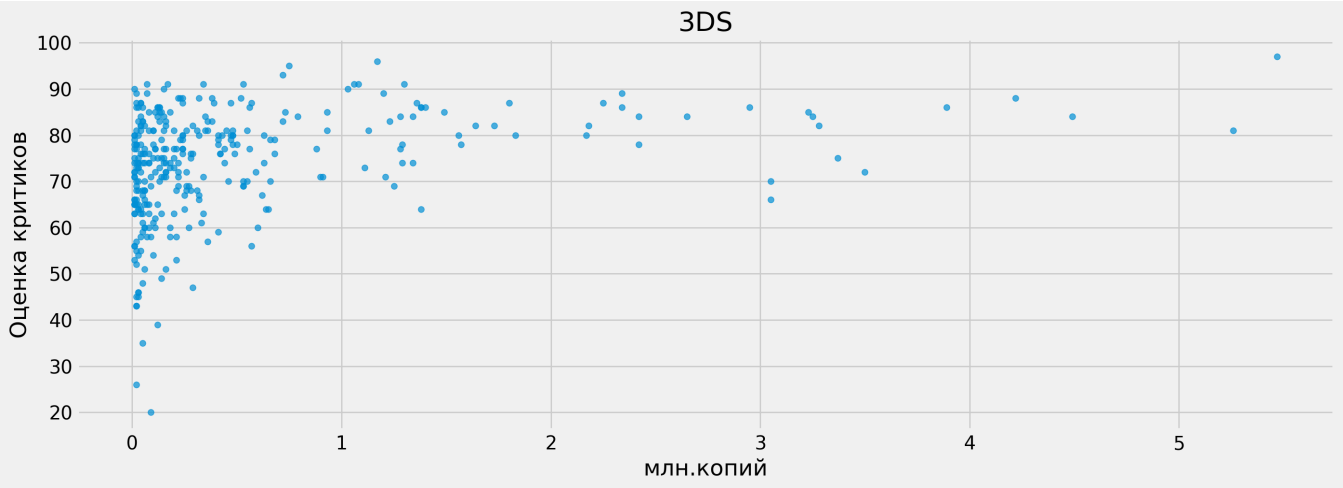
Корреляция между отзывами и продажами для остальных платформ

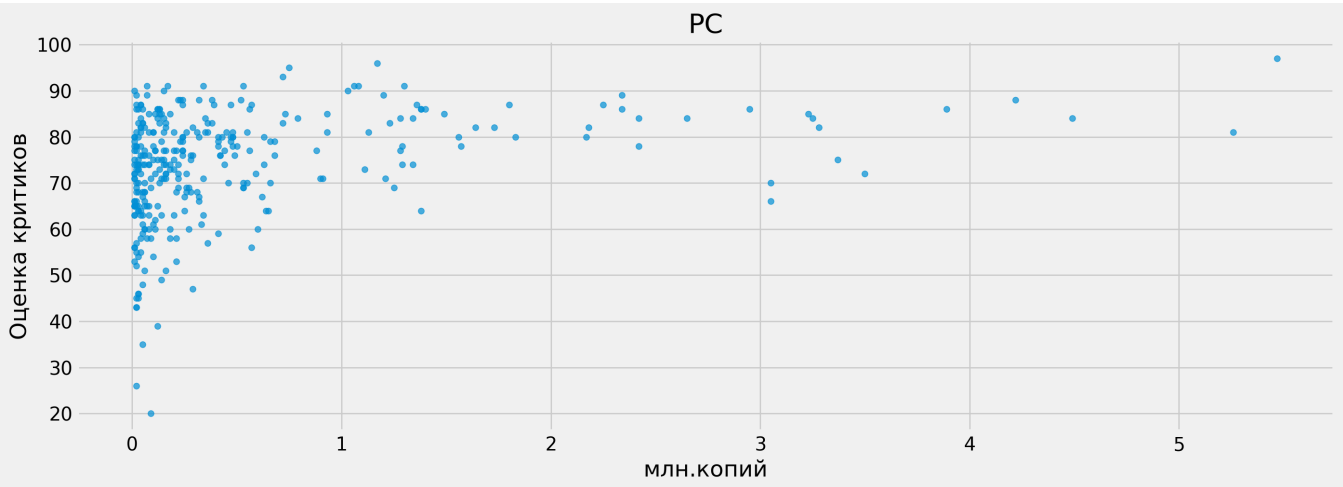
Интересно узнать как зависят продажи от оценок у двух других перспективных платформ: XOne и 3DS. Для полноты картины предлагаю рассмотреть еще и PC - платформа имеет много отзывов и критиков, и игроков

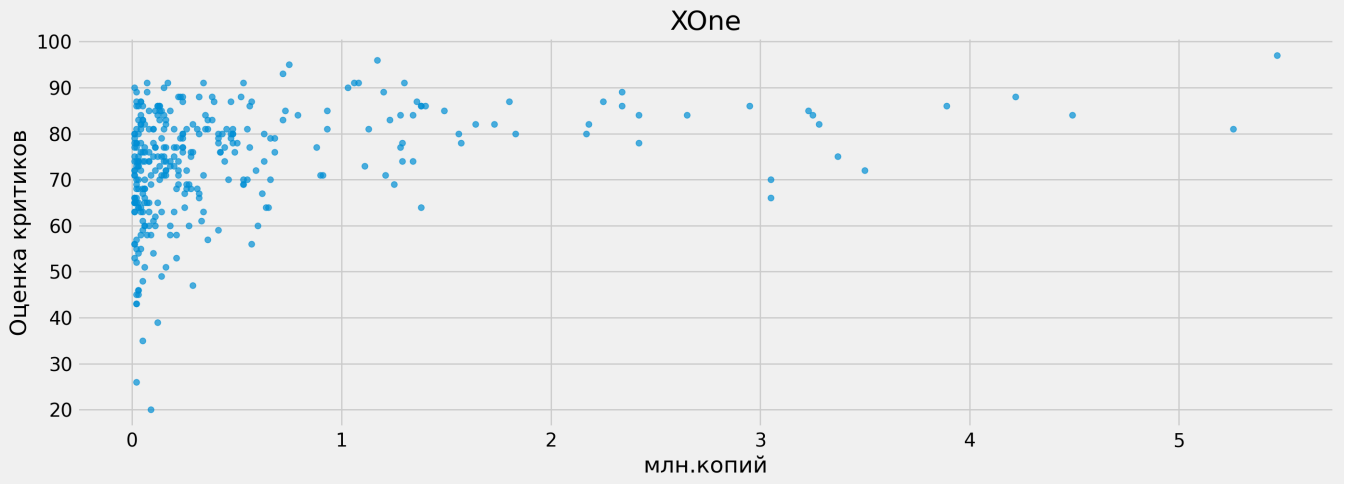
Корреляция продаж и отзывов критиков:

In [37]:
```
platforms_corr_list = ['3DS', 'PC', 'XOne']

games_corr = games.query('platform in @platforms_corr_list')[['platform','total_sales', 'crit
print('Коэффициенты корреляции между продажами и отзывами критиков:')
print()
print(games_corr.groupby('platform')['total_sales'].corr(games_corr['critic_score']).round(3)

for platform in platforms_corr_list:
    games_corr.plot(x='total_sales',
          y='critic_score',
          kind='scatter',
          figsize=(15,5),
          title=platform,
         xlabel = 'млн.копий',
         ylabel = 'Оценка критиков',
         alpha=0.7)
```

```
Коэффициенты корреляции между продажами и отзывами критиков:

platform
3DS     0.314
PC      0.175
XOne    0.429
Name: total_sales, dtype: float64
```





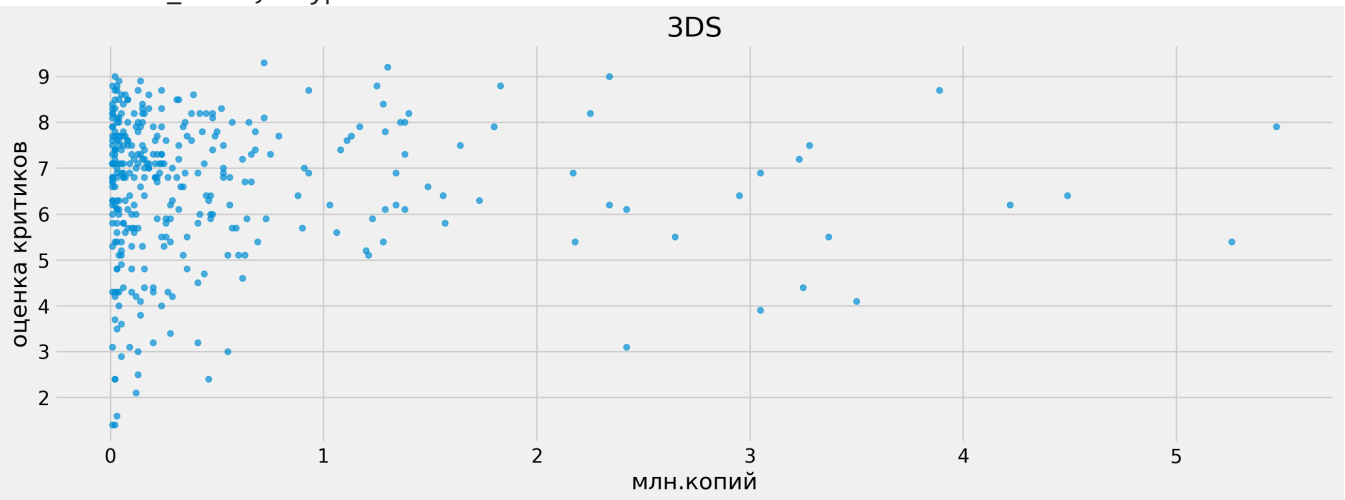


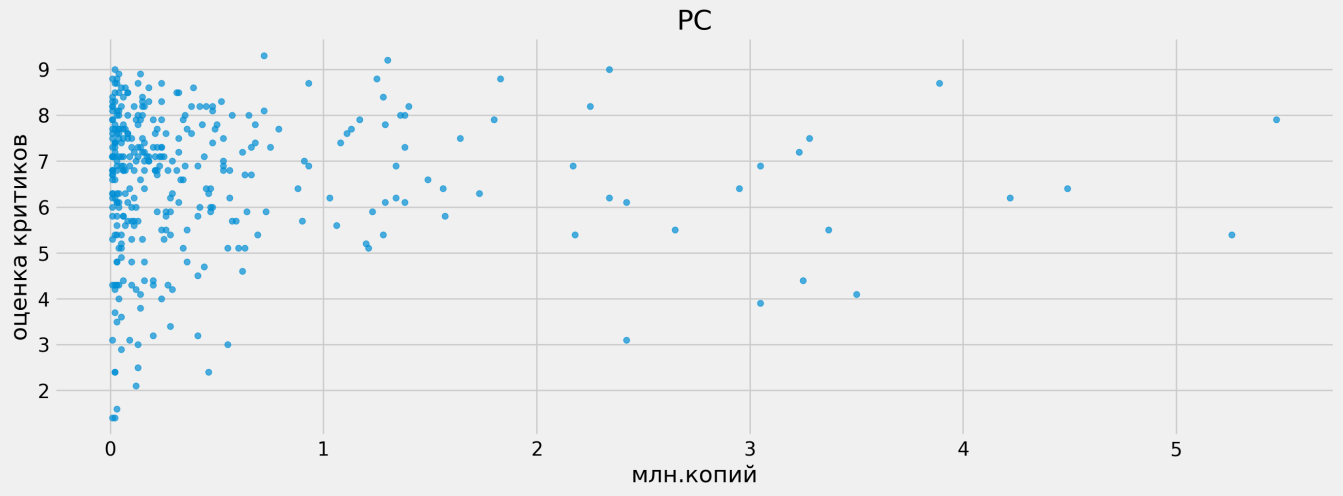
Коэффициенты корреляции выражены не явно (ни для одной из платформ он не превышает 0.5), но можем сформулировать такое наблюдение: популярные игры не получают рейтинг ниже 60, но приэтом провальные игры могут получать самые разнообразные оценки.

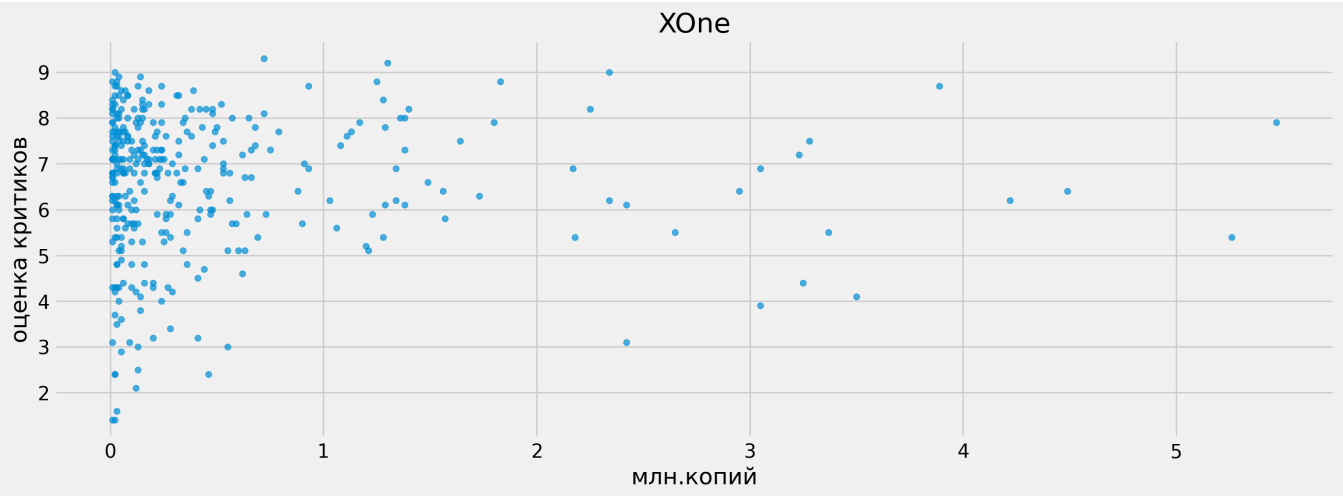
Не рекомендуется опираться на высокие оценки критиков (часто провалившиеся игры получали оценки больше 80), но можно **рассматривать плохие оценки критиков как некий "красный флаг"** - если оценка критиков ниже 60, то игра не продается с тиражом >1 млн., тираж игр с оценкой ниже 50 не достигает 500 тыс. копий

In [38]:
```
print('Коэффициенты корреляции между продажами и отзывами игроков:')
print()
print(games_corr.groupby('platform')['total_sales'].corr(games_corr['user_score']).round(3))

for platform in platforms_corr_list:
    games_corr.plot(x='total_sales',
          y='user_score',
          kind='scatter',
          figsize=(15,5),
          title=platform,
         xlabel = 'млн.копий',
         ylabel = 'оценка критиков',
         alpha=0.7)
```

```
Коэффициенты корреляции между продажами и отзывами игроков:

platform
3DS     0.215
PC     -0.072
XOne   -0.070
Name: total_sales, dtype: float64
```







Коэффициент зависимости выражен очень слабо, наблюдаем очень много выбросов, достаточно часто встречаются низкие оценки для очень игр-хитов и высокие оценки оценки для непопулярных игр.

Вывод

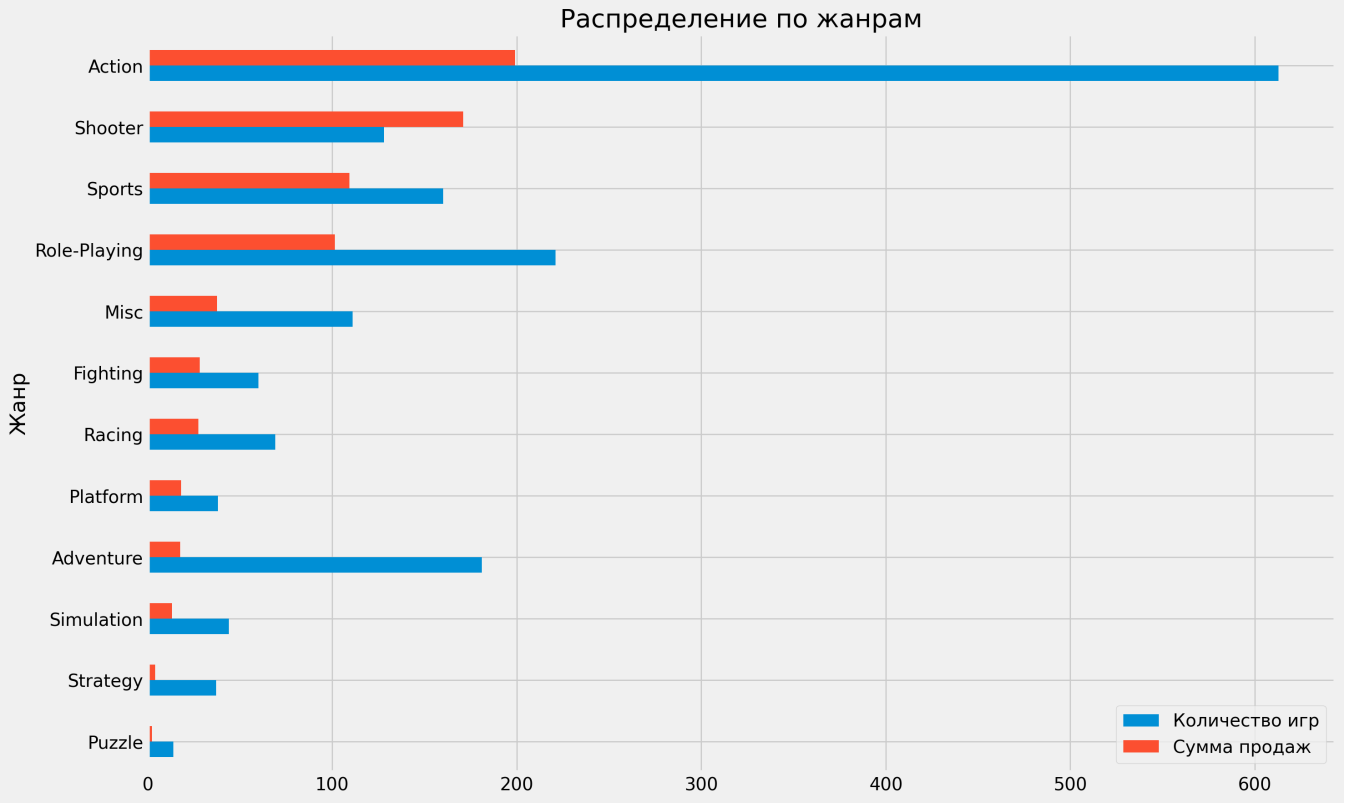
Не стоит опираться на высокие оценки критиков: оценки выше 80 могут получить как хиты, так и абсолютно не популярные игры. Следовательно, по оценке критиков нельзя определить будущий бестселлер. Но это правило отлично работает в другую сторону: **всегда можно определить потенциального аутсайдера по низкой оценке критиков**. Для рассматриваемых платформ **игры, получившие оценку ниже 50 не продаются с тиражом > 500 тыс.**

На оценки игроков следуется ориентироваться в меньшей степени, слишком часто встречаются аномалии (высокие оценки для провалившихся игр и низкие для бестселлеров)

Анализ жанров

Выявим, какие жанры встрчаются на рынке чаще всего и какие продаются лучше:

In [40]:
```
games.pivot_table(index = 'genre', values = 'total_sales',  aggfunc = ({'total_sales' : 'sum'
    .sort_values('total_sales')\
    .plot(kind='barh',
          figsize=(15,10),
          ylabel='млн.копий',
          xlabel='Жанр',
          title='Распределение по жанрам');
plt.legend(['Количество игр', 'Сумма продаж']);
```



Для жанра **Action** выпущено больше всего игр, на этот же жанр приходится большинство суммарных продаж.

Но есть интересное наблюдение: для жанра **Shooter выпущено куда меньше игр, но в рамках своего жанра он продается лучше.** Другими словами, *Shooter меньше представлен на рынке, но игры этого жанра имеют бОльшую популярность*.

Посмотрим на медианные продажи по каждому жанру, медиана меньше всего подвержена выбросам и поможет дать более объективную оценку продаж:

In [41]:
```
games.pivot_table(index = 'genre', values = 'total_sales', aggfunc = 'median')\
    .sort_values(by='total_sales')\
    .plot(kind='barh',
          y='total_sales',
          figsize=(10,5),
          title='Медианные значения продаж',
          legend=False,
          xlabel='жанр',
         ylabel = 'млн.копий');
```

Представление о жанрах сильно изменилось.

В жанре **Action** очевидно выходит много игр-бестселлеров, за счет чего жанр имеет самые высокие суммарные продажи, но медианное значение у Action одно из самых низких - меньше 100 тыс. копий.

Shooter имеет самую высокую медианное значение > 500 тыс. копий. Ранее мы заострили внимание на интересное для этого соотношение общих продаж с количеством выпущенных игр. Выходит, что **Shooter** при меньшем, чем у Action, количестве игр в среднем продается лучше.

С большим отрывом на втором месте по медианному значению располагается **Sports** - чуть менее 200 тыс. копий. По сумме продаж жанр занимает третье место, **можем также отнести его к перспективным.**

Замыкают пятерку *Platform, figting u Role-Playing*

Вывод

Самый перспективный жанр - Shooter.

Sports имеет достаточно высокие и суммарные продажи, и медианное значение, тоже можем считать перспективным.

Медианные значения выше 100 тыс. копий также характерны для **Platform, figting и Role-Playing**

Высокие суммарные продажи **Action** достигаются только за счет хитовых игр, но для жанра характерно **большое количество непопулярных игр.**

Adventure, Puzzle u Strategy самые неперспективные жанры.

Портрет пользователя каждого региона

Имеем данные по продажам в Северной Америке, Европе и Японии. Как делят общий рынок между собой эти регионы? Какова доля каждого из них?

In [42]:
```
#создание новой df. первый столбец - регионы, второй - суммарные продажи по каждому из них:
region = [
    ['Северная Америка', games['na_sales'].sum()],
    ['Европа',games['eu_sales'].sum()],
    ['Япония', games['jp_sales'].sum()],
    ['Другие регионы', games['other_sales'].sum()]
]

region_columns = ['region', 'total_sales']
market_shares = pd.DataFrame(data=region, columns=region_columns)

market_shares['total_sales'].plot(kind='pie',
                                  figsize=(7,7),
                                  cmap='RdPu',
                                  autopct='%1.0f%%',
                                  label='',
                                 title='Доли рынков',
                                 labels=None,
                                 legend=False);
plt.legend(market_shares['region']);
```

- **Американский и европейский рынки занимают 39% и 37% соответственно**
- **Японский значительно меньше** западных - всего **13%**

Продажи в других регионах занимают 11% от общего числа

Выясним, какие платформы и жанры предпочитают в каждом из трех крупнейших регионов.

Популярные платформы

In [43]:
```
# создание функции для построения графиков:

def top_5_region (groupby, region_sales, ax):
    region_name = {'na_sales': 'Северная Америка', 'eu_sales': 'Европа', 'jp_sales': 'Япония'
    region_grouped = games.groupby(groupby).agg({region_sales: 'sum'}).sort_values(by=region_
    plot = region_grouped.plot(kind='bar',
                               title=region_name[region_sales],
                               xlabel = '',
                               legend=False,
                               figsize=(10,4),
                               ax=axes[ax])

fig, axes = plt.subplots(1, 3, figsize=(30,5))
top_5_region('platform', 'na_sales', 0)
top_5_region('platform', 'eu_sales',1)
top_5_region('platform', 'jp_sales',2)
```

Вывод

Западные рынки в целом похожи, **абсоюлтный лидер и в Америке, и в Европе - PS4**, японский же рынок отличается

- **Северная Америка**

Помимо PS4 в топ входит предыдущее поколение из линейки PS. Т.к. PS3 прогнозируемо потеряет свою популярность к 2017 году, можем предполагать, что часть игроков перейдет на **PS4 и еще больше повысит ее популярность.**

Но также в регионе **популярна линейка Xbox**: XOne располагается на втором месте по популярности немногим уступая PS4, X360 замыкает тройку лидеров. **Следует оценивать XOne как перспективную для американского рынка**

- **Европа**

PS4 - абсолютный лидер европейского рынка. Для этой платформы игры продаются суммарно лучше, чем в Северной Америке.

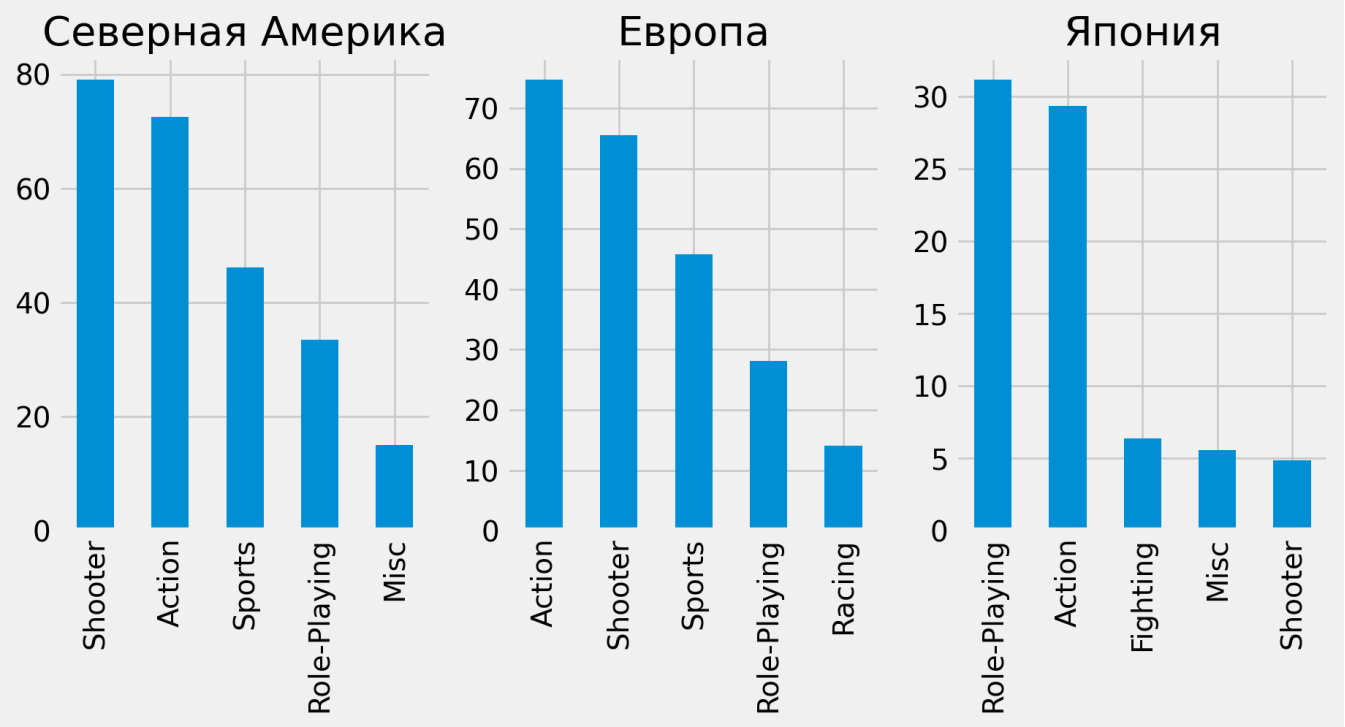
Xone на втором месте по популярности, но имеет большой отрыв от PS4. Все же для европейского рынка следует делать бОльший акцент именно на PS4

- **Япония**

Абсолютный лидер на японском рынке - 3DS. Интересно, что в топ входит PSV - кажется, пользователи из Японии любят небольшие переносные платформы куда больше, чем американцы и европейцы. Также пользуются популярностью платформы из линейки PS, правда, куда меньшей, чем 3DS

Популярные жанры

In [44]:
```
fig, axes = plt.subplots(1, 3, figsize=(30,5))
top_5_region('genre', 'na_sales', 0)
top_5_region('genre', 'eu_sales',1)
top_5_region('genre', 'jp_sales',2)
```



Вывод

И снова западные рынки имеют много общего: в топ входят **Shooter, Action, Sports и Role-Playing**, правда игроки из **Америки чаще покупает игры с жанром Shooter, а европейцы отдают большее предпочтение Action**. Американский топ-5 замыкает Misc, а европейский - Racing.

Японский рынок выглядит иначе. Role-Playing - самый популрный жанр в регионе, ненамного от него отсает **Action**. В топ-5 входят Fightin и Misc, правда, с очень большим отрывом, популярный на Западе Shooter почти совсем не привлекает игроков из Японии

Влияние рейтинга ESRB на продажи в отдельном регионе

In [45]:
```
print('Названия рейтингов:')
games['rating'].sort_values().unique()
```

Out[45]:
```
Названия рейтингов:
array(['E', 'E10+', 'M', 'T', 'undefined'], dtype=object)
```

Обратимся к общедоступным ресурсам для понимания рейтингов:

«E» («Everyone») - для всех, нет возрастных ограничений, считам как 0+

«E10+» («Everyone 10 and older») - для всех, кто старше 10, 10+

«M» («Mature») — «Для взрослых», 17+

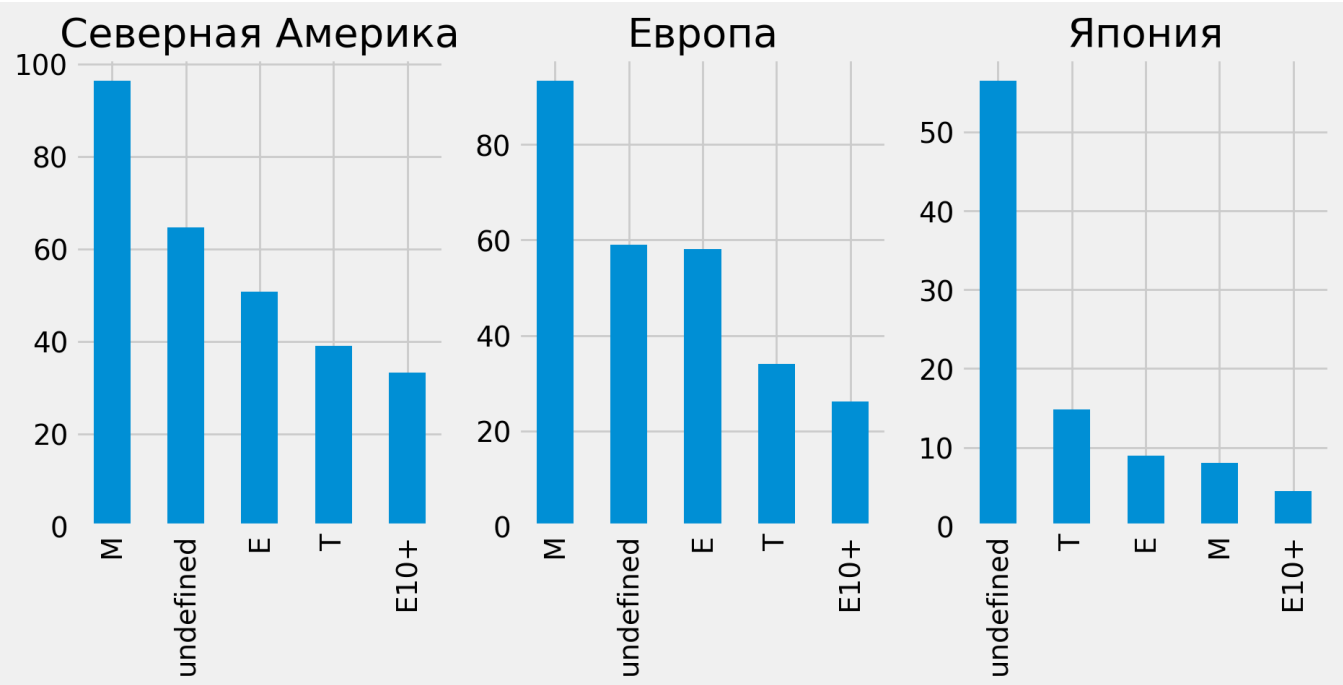
«RP» («Rating Pending») — «Рейтинг ожидается». Продукт был отправлен в ESRB и ожидает присвоения рейтинга

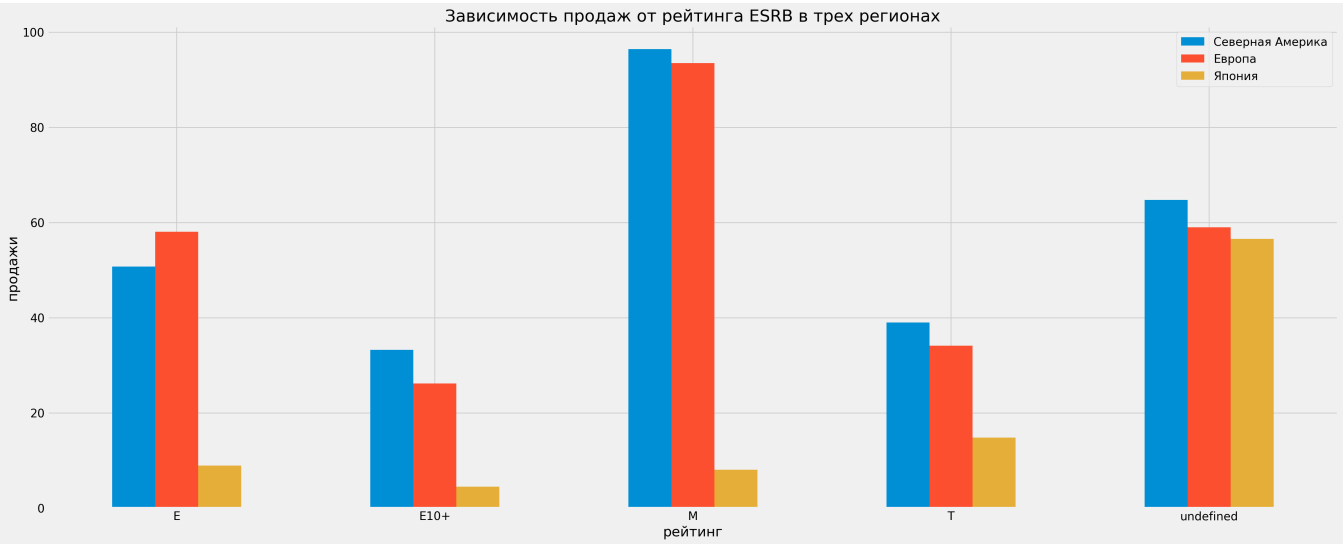
«T» («Teen») — «Подросткам», 13+

undefined - ранее сами установили такое значение для неопределенного рейтинга

In [46]:
```
# Функия для построения графиков по трем регионам:
def rating_plot (groupby, region_sales, ax):
    region_name = {'na_sales': 'Северная Америка', 'eu_sales': 'Европа', 'jp_sales': 'Япония'
    region_grouped = games.groupby(groupby)\
```

```
    .agg({region_sales: 'sum'}).sort_values(by=region_sales, ascending=False)
    plot = region_grouped.plot(kind='bar',
                               title=region_name[region_sales],
                               xlabel = '',
                               legend=False,
                               figsize=(10,4),
                               ax=axes[ax],
                              rot=90)

fig, axes = plt.subplots(1, 3, figsize=(30,5))

rating_plot('rating', 'na_sales', 0)
rating_plot('rating', 'eu_sales', 1)
rating_plot('rating', 'jp_sales', 2)
plt.show();

games.groupby('rating')\
.agg({'na_sales':'sum','eu_sales':'sum','jp_sales':'sum'})\
.plot(kind='bar', figsize=(25,10), xlabel='рейтинг', ylabel='продажи', title='Зависимость про
plt.legend(['Северная Америка', 'Европа', 'Япония']);
```





Вывод

Игроки из **Америки и Европы** снова имеют схожее распределение: чаще всего продаются игры с рейтингом **"M" - 17+**. Учитывая, что западные покупатели часто отдают предпочтения жанрам "shooter", такой рейтинг не удивляет. Далее в одинаковом порядке для обоих регионов располагаются игры с рейтингом E - 0+, T - 13+ и E10+.

Для многих игр рейтинг не определен. **В Японии** игр с неизвестным рейтингом больше, чем суммарно игр с определенным рейтингом. Вероятно, связано с тем, чо ESRB не оценивает игр японского производства, для этого региона существует своя организация - CERO.

Проверка гипотез

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

Гипотеза двусторонняя, совокупности не зависят друг от друга. Проведем статистический тест с использованием t-распределения.

Нулевая гипотеза, как правило, предполагает некое равенство, альтернативная же это равенство отвергает.

Выдвинем следующие гипотезы для проверки:

Нулевая гипотеза (H0): Средние пользовательские рейтинги платформ Xbox One и PC равны

Альтернативная гипотеза (H1): Средние пользовательские рейтинги платформ Xbox One и PC не равны

Установим критический уровень статистической значимости: 0.05

In [48]:
```
# сохраним в отдельных переменных пользовательские рейтинги для обеих платформ. Нули удаляем:
xone = games[games['platform']=="XOne"]['user_score'].dropna()
pc = games[games['platform']=="PC"]['user_score'].dropna()

print('Количество оценок Xbox One:', len(xone))
print('Количество оценок pc:', len(pc))
```

```
Количество оценок Xbox One: 165
Количество оценок pc: 122
```

Рассматриваемые совокупности слишком отличаются по размеру, следовательно, мы не можем быть уверены в том, что дисперсии у совокупностей одинаковые. В связи с этим при проведении t-testa укажем дополнительный параметр equal_var=False

In [49]:
```
alpha = 0.05 #критический уровень стат.значимости

results = st.ttest_ind(xone, pc, equal_var=False)

print('p-значение:', results.pvalue)


if (results.pvalue < alpha):
    print("Отвергаем нулевую гипотезу")
else:
    print("Не получилось отвергнуть нулевую гипотезу")
```

```
p-значение: 0.11601398086668832
Не получилось отвергнуть нулевую гипотезу
```

Вывод

Не удалось опровергнуть нелувую (изначальную) гипотезу. **Разница между пользовательскими рейтингами платформ Xbox One и PC статистически не значима**

Гипотеза 2: Средние пользовательские рейтинги жанров Action и Sports разные

Снова проведем t-test, т.к. совокупности не зависят между собой. Гипотезы выстроим так же как и в предыдущий раз: нулевая предполагает равенство, альтернативная равенство опровергает:

Нулевая гипотеза (H0): Средние пользовательские рейтинги жанров Action и Sports равны

Альтернативная гипотеза (H): Средние пользовательские рейтинги жанров Action и Sports не равны

Установим критический уровень статистической значимости: 0.05

In [50]:
```
# сохраним в отдельных переменных пользовательские рейтинги для обоих жанров . Нули удаляем:
action = games[games['genre']=="Action"]['user_score'].dropna()
sports = games[games['genre']=="Sports"]['user_score'].dropna()

print('Количество пользовательских оценок для жанра "Action":', len(action))
print('Количество пользовательских оценок для жанра "Sports"', len(sports))
```

```
Количество пользовательских оценок для жанра "Action": 297
Количество пользовательских оценок для жанра "Sports" 127
```

Снова рассматриваемые совокупности не равны между собой, используем параметр equal_var=False

In [51]:
```
alpha = 0.05 #критический уровень стат.значимости

results = st.ttest_ind(action, sports, equal_var=False)

print('p-значение:', results.pvalue)


if (results.pvalue < alpha):
    print("Отвергаем нулевую гипотезу")
else:
    print("Не получилось отвергнуть нулевую гипотезу")
```

```
p-значение: 1.1825550382644557e-14
Отвергаем нулевую гипотезу
```

Вывод

Нулевая гипотеза, подразумевающая равенство средних рейтингов отвергнута, следовательно изначально выдвинутая гипотеза верна. **Средние пользовательские рейтинги жанров Action и Sports разные**

Общий вывод

Изучены и обработаны данные для последующего проведения анализа

Названия столбцов приведены к общему виду, данные обработаны от пропусков, изменены некорректные типы данных, рассчитаны и внесены в таблицу новые необходимые значения. Полученный набор данных включал в себя 16715 наблюдений.

Исследовательский анализ данных

- **Динамика игровой индустрии**

Мы имеем данные с 1980 по 2016 годы. **Первый заметный рост игровой индустрии пришелся на 2002 год:** вышло 800+ игр, что почти в два раза больше, чем результаты предыдущего года. Невзирая на небольшое уменьшение числа релизов в следующие два года, можем сказать что разнообразие игр на рынке продолжало увеличиваться.

2008-2009 годы - абсолютный прорыв в разнообразии игр. В каждый из этих двух лет выходит более 1400 новых игр.

Заметно спадает ажиотаж уже к 2011. **В 2016 году имеем уже всего ~500 релизов за год**.

Предполагаю, что "бум" в 2008-2009 годах был ожидаем в связи с ростом индустрии в предыдущие пять лет. Рынок становится популярным, появляются новые разработчики, возможно имело место засилье однотипных игр. К 2011 рынок перенасытился, отсеиваются разработчики, не справившиеся с конкуренцией 2008-09, также можем предположить, что разработчики оценили интересы пользователей и стали выпускать игры только в популярных жанрах.

В целом игры стали сложнее и технически, и сюжетно. На разработку тратится больше времени, что сокращает число релизов. **Да и пользователи скорее отдают предпочтение одной качественной игре, чем нескольким средним.**

Также спад числа релизов может быть связан развитием смартфонов и упадком популярности портативных платформ.

При планировании кампании на 2017 год следует ориентироваться скорее на качество игр, чем на разнообразие.

- **Анализ платформ**

Длительность "жизни" платформы - около 10 лет. Пик популярности приходится на 3-4 года с даты релиза платформы.

Платформы, имеющие перспективы на 2017 год:

- **PS4** ~70 млн. проданных копий в 2016, "молодая" платформа, прогнозируемый период "жизни" - до 2023, лидер по количеству игр-бестселлеров

- **Xone** ~30 млн. проданных копий в 2016, "молодая" платформа, прогнозируемый период "жизни" - до 2023,
- **3DS** преимущественно для японского рынка, ~ 18 млн проданных копий в 2016, старше двух предыдущих, прогнозируемый период "жизни" - до 2021

Остальные платформы к 2017 устареют, многие из них вероятнее всего прекратят поддерживаться.

- **Влияние отзывов на продажи**

С помощью **отзывов критиков** можно определить **потенциально провальные игры**: игры, оцененные ниже 60, не продаются тиражом более 1 млн. копий, тиражи игр с оценкой ниже 50 не достигают отметки в 500 тыс. копий. **Не стоит ориентироваться на высокие оценки** - оценку 90+ может получить и бестселлер, и провалившийся проект. Оценки пользователей часто не соотносятся с числом продаж.

- **Самые популрные жанры**

Самый перспективный жанр - Shooter. Лучшее медианное значение продаж, лучшее соотношение общего числа проданных копий к общему числу релизов.

Sports тоже можем считать перспективным - имеет достаточно высокие и суммарные продажи, и медианное значение.

Медианные значения выше 100 тыс. копий также характерны для **Platform, figting и Role-Playing**.

С осторожностью рассматривать жанр Action: при очень низком значении медианы имеет самую большую сумму продаж, основной доход приносят только хиты, но очень часто встречаются провальные игры.

- **Наблюдения по регионам**

По платформам:

- *Северная Америка*:

Практически в равной степени покупатели предпочитаю линейки PS и Xbox, **основной акцент следует ставить на последние модели линеек: PS4 и XOne**

- *Европа*

Среди пользователей более популярна линейка PS. На сегодня PS4 - абсолютный лидер среди игроков, суммарные продажи превышают продажи в Америке по этой платформе. **Для европейского рынка нужно брать ориентир на PS4.** Платформа XOne пользуется уже куда меньшей популярностью, но все же занимает второе место.

- *Япония*

Самая популярная платформа - 3DS, она же для этого рынка прогнозируемо самая перспективная. В целом игроки из Японии больше предпочитают портативные приставки (в топ входит PSV). PS4 занимает второе место по популярности, но имеет очень большой отрыв от 3DS.

По жанрам:

- *Западный рынок: Северная Америка и Европа*

Для обоих рынков в топ входят Shooter, Action, Sports и Role-Playing, отличие только в том, что **на американском рынке лидер - Shooter, европейцы чаще предпочитают Action**. Больше всего продано игр с рейтингом **"M" - 17+**.

- *Япония*

Role-PLalaying - самый популрный жанр в регионе, ненамного от него отсает Action. В топ-5 входят Fightin и Misc, правда, с очень большим отрывом, популярный на Западе Shooter почти совсем не привлекает игроков из Японии. Для абсолютного большинства игр не определен ейтинг ESRB.

В целом, учитывая различия трех регионов, при дальнейшем планировании рациональнее **рассматривать рынки по отдельности**.

Важно также учитывать доли этих рынков: во многом отличающийся японский рынок занимает всего 13%, в то время как американский и европейский составляют по 37% и 39% соответственно

Проверены две гипотезы

Проведены t-тесты, по результатам которых можно отметить:

- Средние пользовательские рейтинги платформ Xbox One и PC равны между собой
- Средние пользовательские рейтинги жанров Action и Sports разные

Основные ориентиры:

- При закупке лучше ориентироваться на более технически сложные, увлекательные игры
- В Америке в равной степени ориентироваться на PS4 и XOne, в Европе - на PS4, в Японии на портативные платформы (3DS)
- Западные покупатели чаще выбирают Action, Shooter и Sports, покупатели из Японии - Role-Playing и Action
- В целом лучше рассматривать каждый из рынков отдельно, между ними есть много различий
- низкие оценки критиков помогут выявить провальные игры