In [6]:

```
!pip install jupyterthemes
```

```
Requirement already satisfied: jupyterthemes in c:\programdata\anaconda3\lib\site-packag
es (0.20.0)
Requirement already satisfied: notebook>=5.6.0 in c:\programdata\anaconda3\lib\site-pack
ages (from jupyterthemes) (6.1.4)
Requirement already satisfied: ipython>=5.4.1 in c:\users\nikita\appdata\roaming\python
\python38\site-packages (from jupyterthemes) (7.22.0)
Requirement already satisfied: jupyter-core in c:\users\nikita\appdata\roaming\python\py
thon38\site-packages (from jupyterthemes) (4.7.1)
Requirement already satisfied: matplotlib>=1.4.3 in c:\programdata\anaconda3\lib\site-pa
ckages (from jupyterthemes) (3.3.2)
Requirement already satisfied: lesscpy>=0.11.2 in c:\programdata\anaconda3\lib\site-pack
ages (from jupyterthemes) (0.14.0)
Requirement already satisfied: jinja2 in c:\programdata\anaconda3\lib\site-packages (fro
m notebook>=5.6.0->jupyterthemes) (2.11.2)
Requirement already satisfied: ipython-genutils in c:\users\nikita\appdata\roaming\pytho
n\python38\site-packages (from notebook>=5.6.0->jupyterthemes) (0.2.0)
Requirement already satisfied: Send2Trash in c:\programdata\anaconda3\lib\site-packages
(from notebook>=5.6.0->jupyterthemes) (1.5.0)
Requirement already satisfied: argon2-cffi in c:\programdata\anaconda3\lib\site-packages
(from notebook>=5.6.0->jupyterthemes) (20.1.0)
Requirement already satisfied: prometheus-client in c:\programdata\anaconda3\lib\site-pa
ckages (from notebook>=5.6.0->jupyterthemes) (0.8.0)
Requirement already satisfied: ipykernel in c:\users\nikita\appdata\roaming\python\pytho
n38\site-packages (from notebook>=5.6.0->jupyterthemes) (5.5.0)
Requirement already satisfied: tornado>=5.0 in c:\users\nikita\appdata\roaming\python\py
thon38\site-packages (from notebook>=5.6.0->jupyterthemes) (6.1)
Requirement already satisfied: jupyter-client>=5.3.4 in c:\users\nikita\appdata\roaming
\python\python38\site-packages (from notebook>=5.6.0->jupyterthemes) (6.1.12)
Requirement already satisfied: nbformat in c:\programdata\anaconda3\lib\site-packages (f
rom notebook>=5.6.0->jupyterthemes) (5.0.8)
Requirement already satisfied: pyzmq>=17 in c:\users\nikita\appdata\roaming\python\pytho
n38\site-packages (from notebook>=5.6.0->jupyterthemes) (22.0.3)
Requirement already satisfied: terminado>=0.8.3 in c:\programdata\anaconda3\lib\site-pac
kages (from notebook>=5.6.0->jupyterthemes) (0.9.1)
Requirement already satisfied: traitlets>=4.2.1 in c:\users\nikita\appdata\roaming\pytho
n\python38\site-packages (from notebook>=5.6.0->jupyterthemes) (5.0.5)
Requirement already satisfied: nbconvert in c:\programdata\anaconda3\lib\site-packages
(from notebook>=5.6.0->jupyterthemes) (6.0.0)
Requirement already satisfied: setuptools>=18.5 in c:\programdata\anaconda3\lib\site-pac
kages (from ipython>=5.4.1->jupyterthemes) (50.3.1.post20201107)
Requirement already satisfied: decorator in c:\users\nikita\appdata\roaming\python\pytho
n38\site-packages (from ipython>=5.4.1->jupyterthemes) (4.4.2)
Requirement already satisfied: prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0 in c:\users
\nikita\appdata\roaming\python\python38\site-packages (from ipython>=5.4.1->jupytertheme
s) (3.0.18)
Requirement already satisfied: pickleshare in c:\users\nikita\appdata\roaming\python\pyt
hon38\site-packages (from ipython>=5.4.1->jupyterthemes) (0.7.5)
Requirement already satisfied: jedi>=0.16 in c:\users\nikita\appdata\roaming\python\pyth
on38\site-packages (from ipython>=5.4.1->jupyterthemes) (0.18.0)
Requirement already satisfied: pygments in c:\users\nikita\appdata\roaming\python\python
38\site-packages (from ipython>=5.4.1->jupyterthemes) (2.8.1)
Requirement already satisfied: colorama; sys_platform == "win32" in c:\users\nikita\appd
ata\roaming\python\python38\site-packages (from ipython>=5.4.1->jupyterthemes) (0.4.3)
Requirement already satisfied: backcall in c:\users\nikita\appdata\roaming\python\python
38\site-packages (from ipython>=5.4.1->jupyterthemes) (0.2.0)
Requirement already satisfied: pywin32>=1.0; sys_platform == "win32" in c:\users\nikita
\appdata\roaming\python\python38\site-packages (from jupyter-core->jupyterthemes) (300)
Requirement already satisfied: cycler>=0.10 in c:\programdata\anaconda3\lib\site-package
s (from matplotlib>=1.4.3->jupyterthemes) (0.10.0)
Requirement already satisfied: pillow>=6.2.0 in c:\programdata\anaconda3\lib\site-packag
es (from matplotlib>=1.4.3->jupyterthemes) (8.0.1)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.3 in c:\programdat
a\anaconda3\lib\site-packages (from matplotlib>=1.4.3->jupyterthemes) (2.4.7)
Requirement already satisfied: numpy>=1.15 in c:\programdata\anaconda3\lib\site-packages
(from matplotlib>=1.4.3->jupyterthemes) (1.19.2)
```

```
Requirement already satisfied: python-dateutil>=2.1 in c:\users\nikita\appdata\roaming\p
ython\python38\site-packages (from matplotlib>=1.4.3->jupyterthemes) (2.8.1)
Requirement already satisfied: certifi>=2020.06.20 in c:\programdata\anaconda3\lib\site-
packages (from matplotlib>=1.4.3->jupyterthemes) (2020.6.20)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\programdata\anaconda3\lib\site-pa
ckages (from matplotlib>=1.4.3->jupyterthemes) (1.3.0)
Requirement already satisfied: ply in c:\programdata\anaconda3\lib\site-packages (from l
esscpy>=0.11.2->jupyterthemes) (3.11)
Requirement already satisfied: six in c:\users\nikita\appdata\roaming\python\python38\si
te-packages (from lesscpy>=0.11.2->jupyterthemes) (1.14.0)
Requirement already satisfied: MarkupSafe>=0.23 in c:\programdata\anaconda3\lib\site-pac
kages (from jinja2->notebook>=5.6.0->jupyterthemes) (1.1.1)
Requirement already satisfied: cffi>=1.0.0 in c:\programdata\anaconda3\lib\site-packages
(from argon2-cffi->notebook>=5.6.0->jupyterthemes) (1.14.3)
Requirement already satisfied: jsonschema!=2.5.0,>=2.4 in c:\programdata\anaconda3\lib\s
ite-packages (from nbformat->notebook>=5.6.0->jupyterthemes) (3.2.0)
Requirement already satisfied: pywinpty>=0.5 in c:\programdata\anaconda3\lib\site-packag
es (from terminado>=0.8.3->notebook>=5.6.0->jupyterthemes) (0.5.7)
Requirement already satisfied: jupyterlab-pygments in c:\programdata\anaconda3\lib\site-
packages (from nbconvert->notebook>=5.6.0->jupyterthemes) (0.1.2)
Requirement already satisfied: bleach in c:\programdata\anaconda3\lib\site-packages (fro
m nbconvert->notebook>=5.6.0->jupyterthemes) (3.2.1)
Requirement already satisfied: defusedxml in c:\programdata\anaconda3\lib\site-packages
(from nbconvert->notebook>=5.6.0->jupyterthemes) (0.6.0)
Requirement already satisfied: testpath in c:\programdata\anaconda3\lib\site-packages (f
rom nbconvert->notebook>=5.6.0->jupyterthemes) (0.4.4)
Requirement already satisfied: mistune<2,>=0.8.1 in c:\programdata\anaconda3\lib\site-pa
ckages (from nbconvert->notebook>=5.6.0->jupyterthemes) (0.8.4)
Requirement already satisfied: pandocfilters>=1.4.1 in c:\programdata\anaconda3\lib\site
-packages (from nbconvert->notebook>=5.6.0->jupyterthemes) (1.4.3)
Requirement already satisfied: nbclient<0.6.0,>=0.5.0 in c:\programdata\anaconda3\lib\si
te-packages (from nbconvert->notebook>=5.6.0->jupyterthemes) (0.5.1)
Requirement already satisfied: entrypoints>=0.2.2 in c:\programdata\anaconda3\lib\site-p
ackages (from nbconvert->notebook>=5.6.0->jupyterthemes) (0.3)
Requirement already satisfied: wcwidth in c:\programdata\anaconda3\lib\site-packages (fr
om prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0->ipython>=5.4.1->jupyterthemes) (0.2.5)
Requirement already satisfied: parso<0.9.0,>=0.8.0 in c:\users\nikita\appdata\roaming\py
thon\python38\site-packages (from jedi>=0.16->ipython>=5.4.1->jupyterthemes) (0.8.1)
Requirement already satisfied: pycparser in c:\programdata\anaconda3\lib\site-packages
(from cffi>=1.0.0->argon2-cffi->notebook>=5.6.0->jupyterthemes) (2.20)
Requirement already satisfied: pyrsistent>=0.14.0 in c:\programdata\anaconda3\lib\site-p
ackages (from jsonschema!=2.5.0,>=2.4->nbformat->notebook>=5.6.0->jupyterthemes) (0.17.
3)
Requirement already satisfied: attrs>=17.4.0 in c:\programdata\anaconda3\lib\site-packag
es (from jsonschema!=2.5.0,>=2.4->nbformat->notebook>=5.6.0->jupyterthemes) (20.3.0)
Requirement already satisfied: webencodings in c:\programdata\anaconda3\lib\site-package
s (from bleach->nbconvert->notebook>=5.6.0->jupyterthemes) (0.5.1)
Requirement already satisfied: packaging in c:\programdata\anaconda3\lib\site-packages
(from bleach->nbconvert->notebook>=5.6.0->jupyterthemes) (20.4)
Requirement already satisfied: nest-asyncio in c:\programdata\anaconda3\lib\site-package
s (from nbclient<0.6.0,>=0.5.0->nbconvert->notebook>=5.6.0->jupyterthemes) (1.4.2)
Requirement already satisfied: async-generator in c:\programdata\anaconda3\lib\site-pack
ages (from nbclient<0.6.0,>=0.5.0->nbconvert->notebook>=5.6.0->jupyterthemes) (1.10)
```

In [7]:
```
!jt -t grade3
```

Разведочный анализ данных. Исследование и визуализация данных.

1) Текстовое описание набора данных

База данных для анализа рынка игровой индустрии

Описание БД:

Данная предметная область представляет собой некий список видео-игр, разделенных по издателям, платформам, релизным датам, жанрам, а также продажам в разных регионах. Данная БД помогает ускорить маркетинговому отделу некой студии, которая решила выпустить игру, определится с издателем, платформой и регионами релиза. Также у каждой игры присутствует ранг, который позволяет выделить самые успешные проекты и ставить ориентир на них.

Таблица БД:

1. `vgsales` - Таблица, содержащая информацию об играх

Импорт библиотек

Импортируем библиотеки с помощью команды import. Будем подключать все библиотеки последовательно, по мере их использования.

In [8]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

Загрузка данных

Загрузим файлы датасета в помощью библиотеки Pandas.

In [9]:
```
table = pd.read_csv('database/vgsales.csv', sep=",")
```

2) Основные характеристики датасета

In [10]:
```
print(f'Размер таблицы RUvideos: {table.shape}, где:\n{table.shape[0]} - строки\n{table.shape[1]} - столбцы')
```

```
Размер таблицы RUvideos: (16598, 11), где:
16598 - строки
11 - столбцы
```

In [11]:
```
# Первые 10 строк датасета
table.head(10)
```

Out[11]:

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_S

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_S
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	8
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	4
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	3
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	3
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	3
5	6	Tetris	GB	1989.0	Puzzle	Nintendo	23.20	2.26	4.22	0.58	3
6	7	New Super Mario Bros.	DS	2006.0	Platform	Nintendo	11.38	9.23	6.50	2.90	3
7	8	Wii Play	Wii	2006.0	Misc	Nintendo	14.03	9.20	2.93	2.85	2
8	9	New Super Mario Bros. Wii	Wii	2009.0	Platform	Nintendo	14.59	7.06	4.70	2.26	2
9	10	Duck Hunt	NES	1984.0	Shooter	Nintendo	26.93	0.63	0.28	0.47	2

In [12]:
```
# Список колонок с типами данных
table.dtypes
```

Out[12]:
```
Rank              int64
Name             object
Platform         object
Year            float64
Genre            object
Publisher        object
NA_Sales        float64
EU_Sales        float64
JP_Sales        float64
Other_Sales     float64
Global_Sales    float64
dtype: object
```

In [13]:
```
# Проверим наличие пустых значений
# Цикл по колонкам датасета
print('Количество пустых ячеек в таблице:')
for col in table.columns:
    # Количество пустых значений - все значения заполнены
    temp_null_count = table[table[col].isnull()].shape[0]
    print(f'{col} - {temp_null_count}')
```

```
Количество пустых ячеек в таблице:
Rank - 0
Name - 0
Platform - 0
Year - 271
Genre - 0
Publisher - 58
NA_Sales - 0
EU_Sales - 0
```

```
JP_Sales - 0
Other_Sales - 0
Global_Sales - 0
```

In [14]:
```
# Основные статистические характеристки набора данных
table.describe()
```

Out[14]:

	Rank	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
count	16598.000000	16327.000000	16598.000000	16598.000000	16598.000000	16598.000000	16598.000000
mean	8300.605254	2006.406443	0.264667	0.146652	0.077782	0.048063	0.537441
std	4791.853933	5.828981	0.816683	0.505351	0.309291	0.188588	1.555028
min	1.000000	1980.000000	0.000000	0.000000	0.000000	0.000000	0.010000
25%	4151.250000	2003.000000	0.000000	0.000000	0.000000	0.000000	0.060000
50%	8300.500000	2007.000000	0.080000	0.020000	0.000000	0.010000	0.170000
75%	12449.750000	2010.000000	0.240000	0.110000	0.040000	0.040000	0.470000
max	16600.000000	2020.000000	41.490000	29.020000	10.220000	10.570000	82.740000

In [15]:
```
# Определим уникальные значения для целевого признака
table['Year'].unique()
```

Out[15]:
```
array([2006., 1985., 2008., 2009., 1996., 1989., 1984., 2005., 1999.,
       2007., 2010., 2013., 2004., 1990., 1988., 2002., 2001., 2011.,
       1998., 2015., 2012., 2014., 1992., 1997., 1993., 1994., 1982.,
       2003., 1986., 2000.,   nan, 1995., 2016., 1991., 1981., 1987.,
       1980., 1983., 2020., 2017.])
```

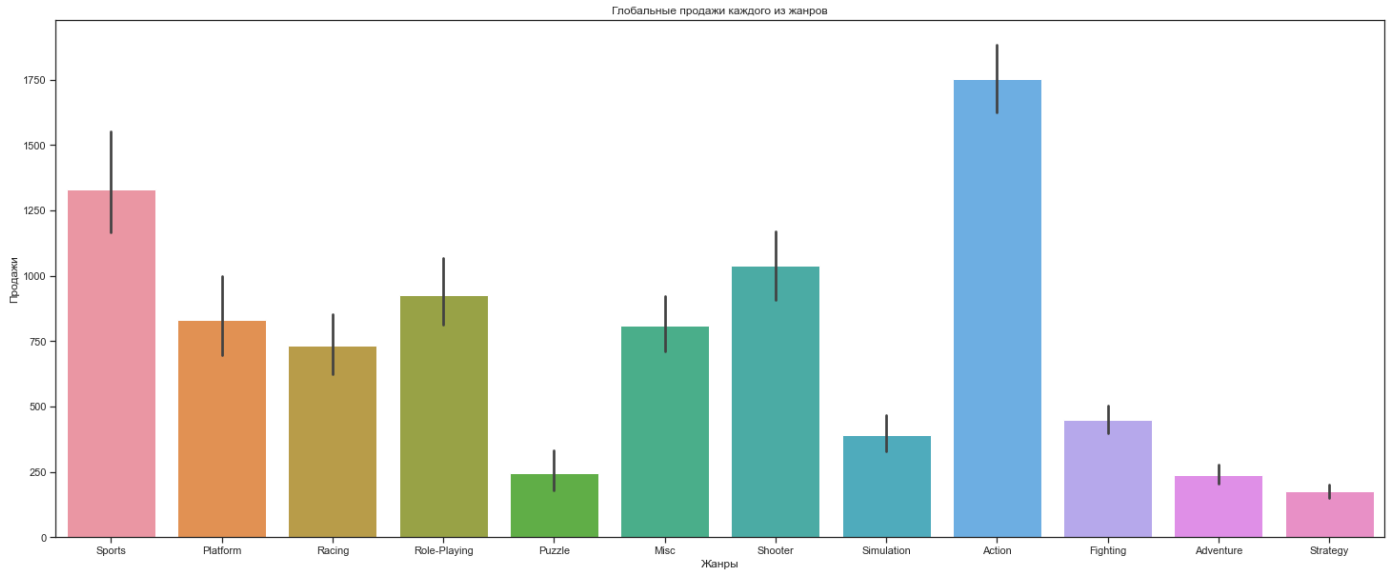
Целевой признак является вещественным числом float64.

3) Визуальное исследование датасета

Столбчатая диаграмма, показывающая все продажи игр по всему миру каждого из жанров

In [16]:
```
fig, ax = plt.subplots(figsize = (25,10))
plt.title("Глобальные продажи каждого из жанров")
sns.barplot(data = table, x="Genre", y="Global_Sales", estimator=np.sum)
plt.ylabel("Продажи")
plt.xlabel("Жанры")
```

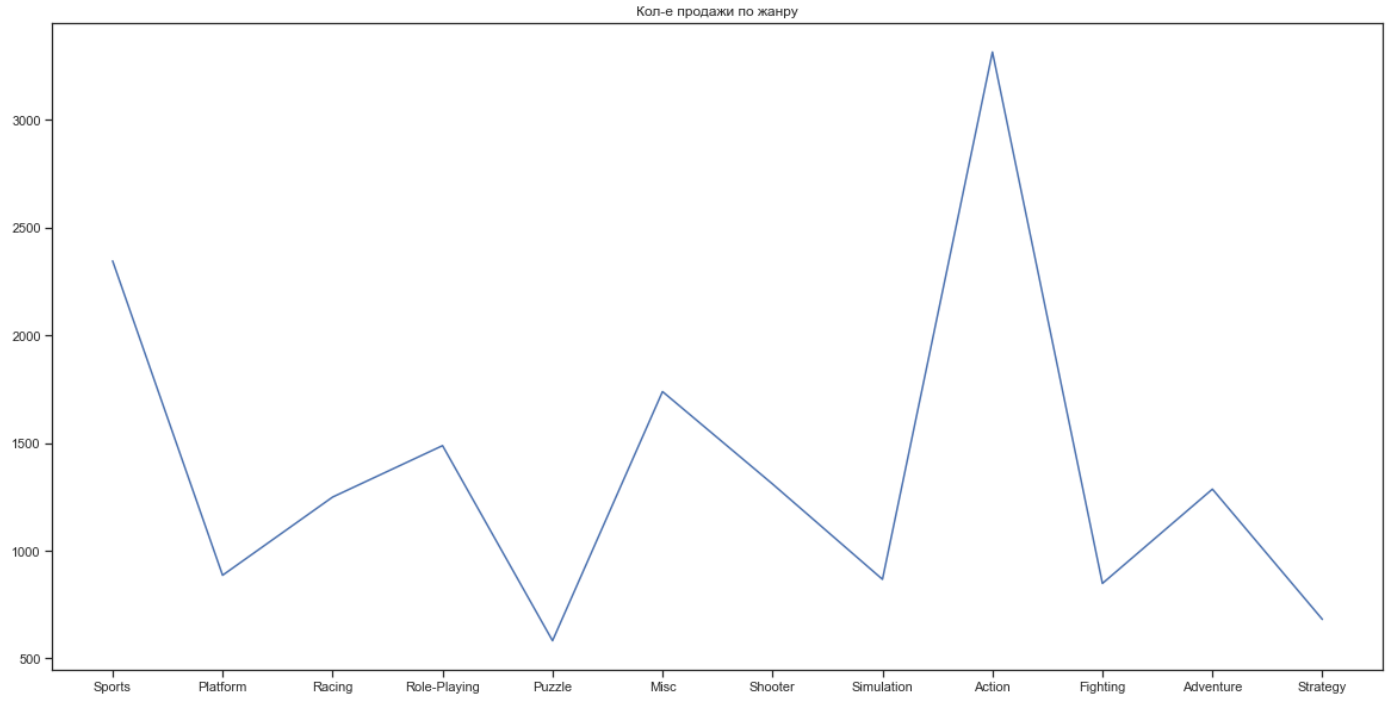
Out[16]:
```
Text(0.5, 0, 'Жанры')
```



Линейный график, показывающий сколько копий различных игр каждого из жанров было продано всего

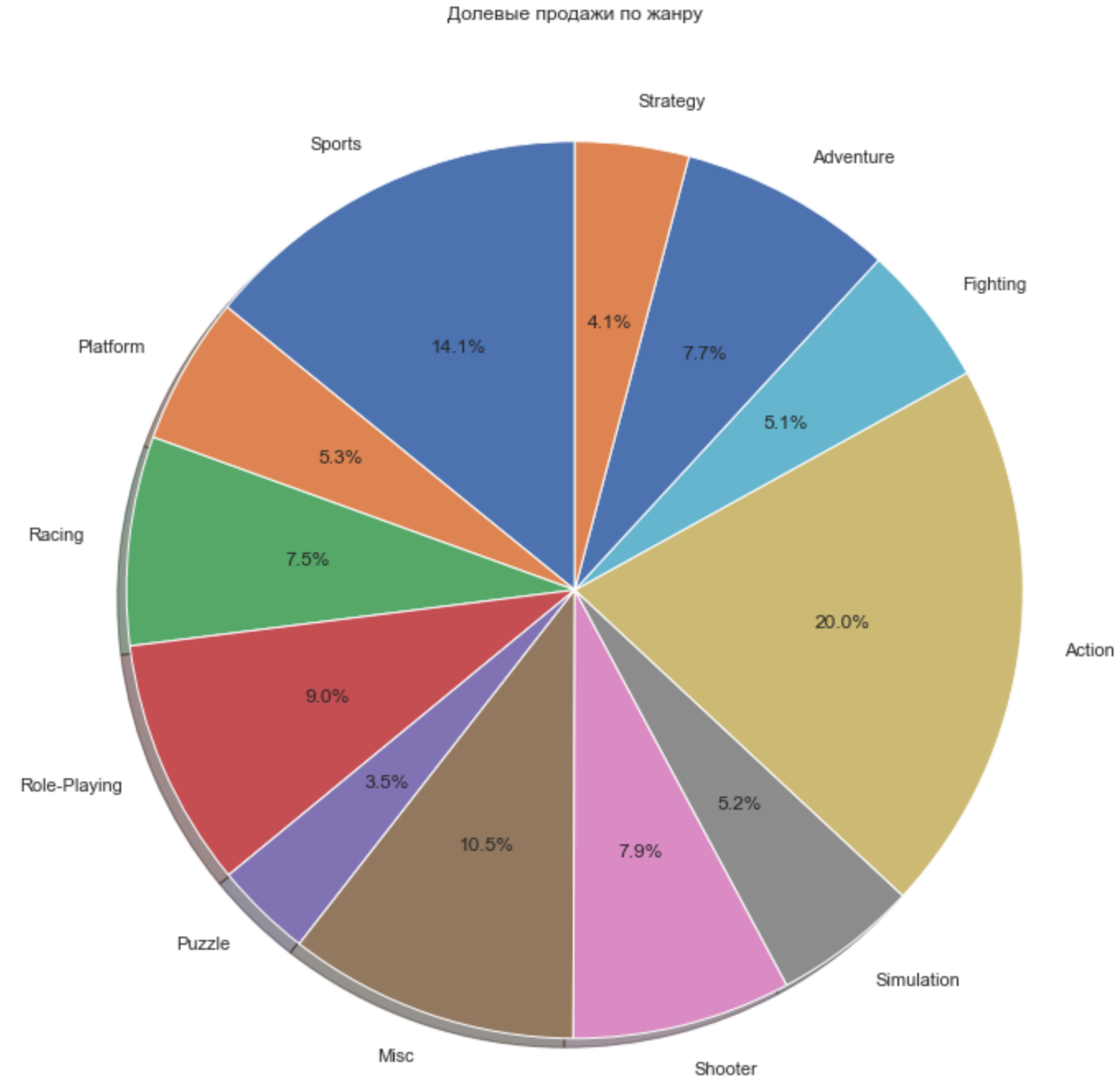
In [17]:
```
oy = []
ox = table["Genre"].unique()
for genre in ox:
    oy.append(table[table["Genre"] == genre].count()[0])
```

In [18]:
```
plt.xlabel = 'Жанр'
plt.ylabel = "Общее кол-во проданных копий по жанру"
fig, ax = plt.subplots(figsize = (20,10))
plt.plot(ox,oy)
plt.title("Кол-е продажи по жанру")
plt.show()
```



Круговая диаграмма, показывающая какую долю продаж занимает каждый жанр

In [19]:
```
fig, ax = plt.subplots(figsize = (10,10))
plt.pie(oy, labels = ox, shadow = True, startangle=90, autopct = "%1.1f%%")
plt.title("Долевые продажи по жанру")
plt.tight_layout()
plt.show()
```



4) Информация о корреляции признаков

Корреляционная матрица на основе коэфициента Спирмена

In [20]:
```
table.corr(method = "spearman")
```

Out[20]:

	Rank	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
Rank	1.000000	0.151529	-0.795516	-0.697105	-0.151851	-0.810416	-0.999622
Year	0.151529	1.000000	-0.133088	-0.057729	0.009605	0.055726	-0.151248
NA_Sales	-0.795516	-0.133088	1.000000	0.681254	-0.228603	0.769432	0.795572
EU_Sales	-0.697105	-0.057729	0.681254	1.000000	-0.177486	0.766054	0.696846
JP_Sales	-0.151851	0.009605	-0.228603	-0.177486	1.000000	-0.069990	0.151931
Other_Sales	-0.810416	0.055726	0.769432	0.766054	-0.069990	1.000000	0.810381
Global_Sales	-0.999622	-0.151248	0.795572	0.696846	0.151931	0.810381	1.000000

Корреляционная матрица на основе коэфициента Кендалла

In [21]:
```
table.corr(method = "kendall")
```

Out[21]:

	Rank	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
Rank	1.000000	0.104901	-0.669392	-0.556641	-0.125372	-0.677420	-0.989750
Year	0.104901	1.000000	-0.094939	-0.047794	0.013893	0.041790	-0.105730
NA_Sales	-0.669392	-0.094939	1.000000	0.556031	-0.181112	0.640092	0.675652
EU_Sales	-0.556641	-0.047794	0.556031	1.000000	-0.143193	0.661482	0.561736
JP_Sales	-0.125372	0.013893	-0.181112	-0.143193	1.000000	-0.058865	0.126682
Other_Sales	-0.677420	0.041790	0.640092	0.661482	-0.058865	1.000000	0.684007
Global_Sales	-0.989750	-0.105730	0.675652	0.561736	0.126682	0.684007	1.000000

Корреляционная матрица на основе коэфициента Пирсона

In [22]:
```
table.corr(method = "pearson")
```

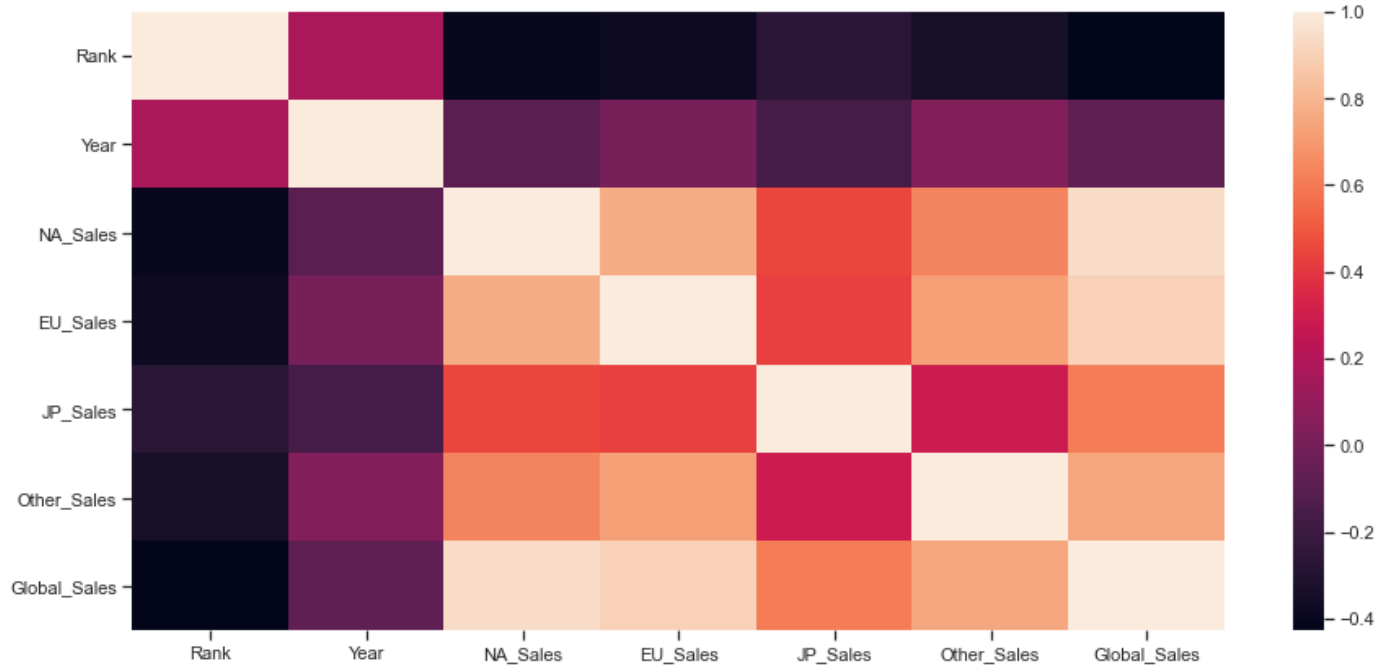
Out[22]:

	Rank	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
Rank	1.000000	0.178814	-0.401362	-0.379123	-0.267785	-0.332986	-0.427407
Year	0.178814	1.000000	-0.091402	0.006014	-0.169316	0.041058	-0.074735
NA_Sales	-0.401362	-0.091402	1.000000	0.767727	0.449787	0.634737	0.941047
EU_Sales	-0.379123	0.006014	0.767727	1.000000	0.435584	0.726385	0.902836
JP_Sales	-0.267785	-0.169316	0.449787	0.435584	1.000000	0.290186	0.611816
Other_Sales	-0.332986	0.041058	0.634737	0.726385	0.290186	1.000000	0.748331
Global_Sales	-0.427407	-0.074735	0.941047	0.902836	0.611816	0.748331	1.000000

Для визуализации корреляционной матрицы будем использовать "тепловую карту" heatmap которая показывает степень корреляции различными цветами.

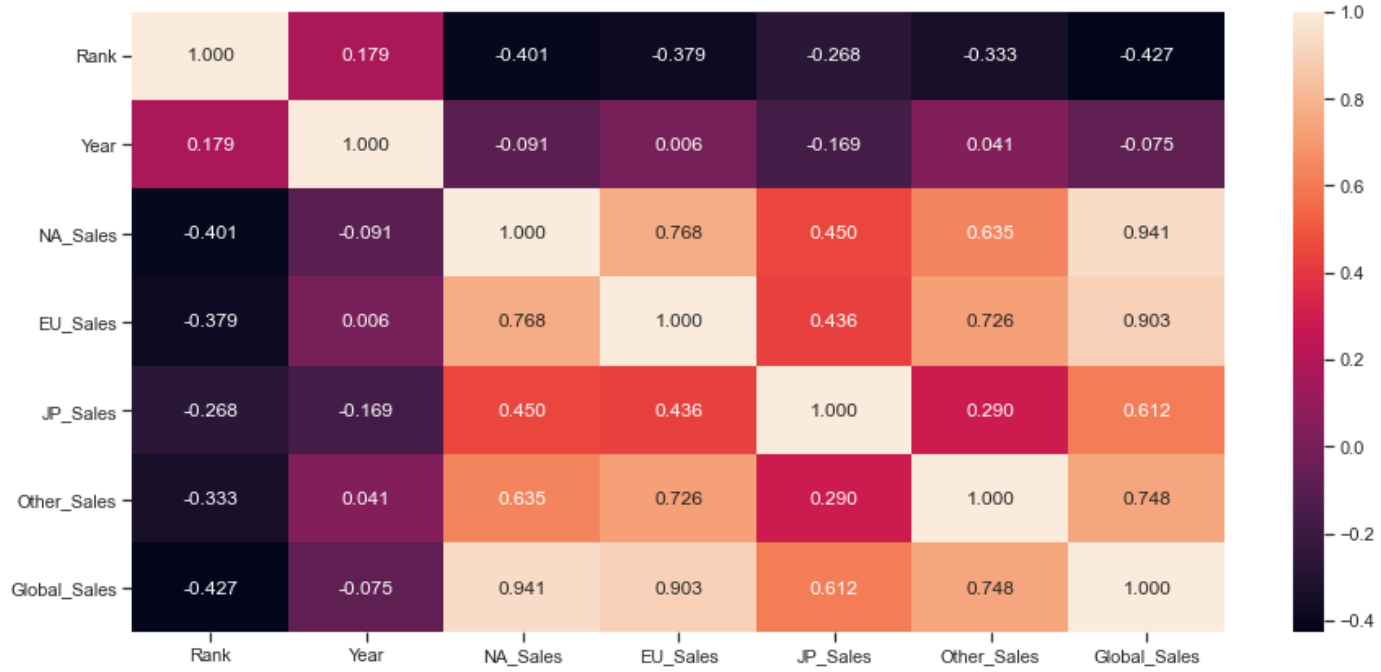
In [23]:
```
fig, ax = plt.subplots(figsize = (15,7))
sns.heatmap(table.corr())
```

Out[23]: <AxesSubplot:>



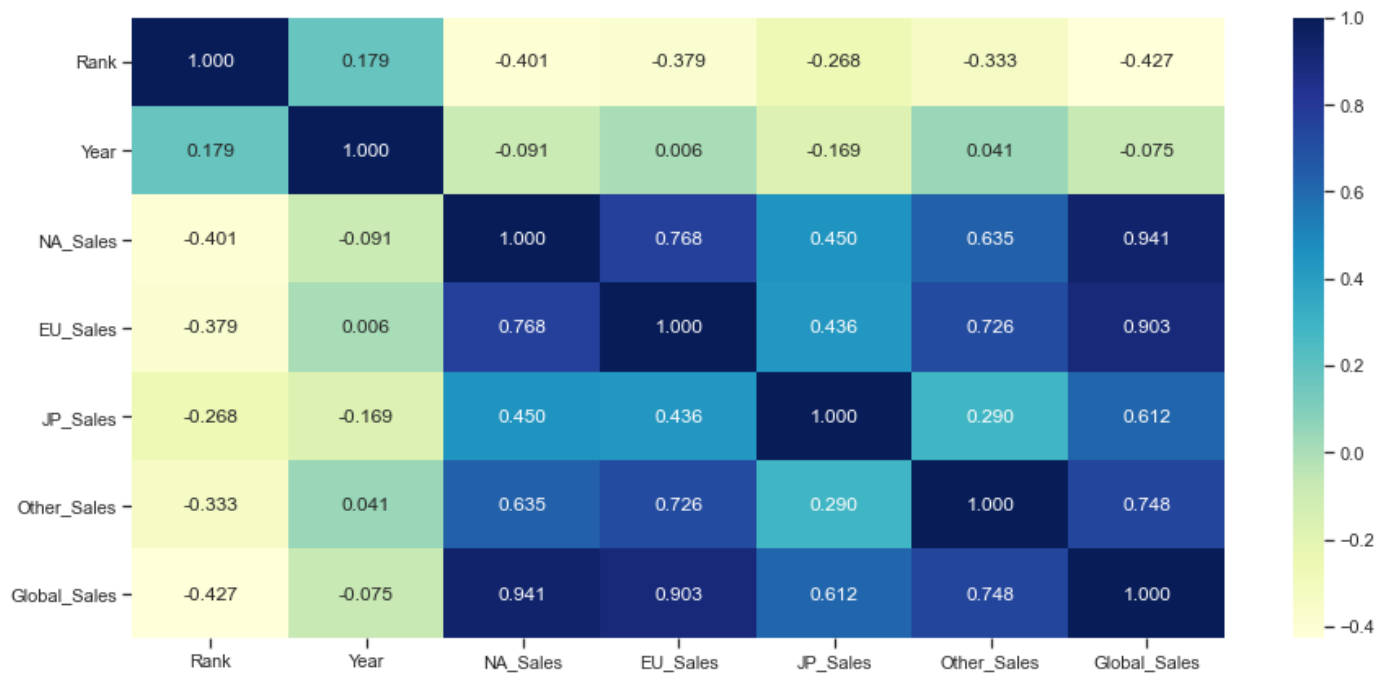
In [24]:
```
# Вывод значений в ячейках
fig, ax = plt.subplots(figsize = (15,7))
sns.heatmap(table.corr(), annot=True, fmt='.3f')
```

Out[24]: <AxesSubplot:>



In [25]:
```
# Изменение цветовой гаммы
fig, ax = plt.subplots(figsize = (15,7))
sns.heatmap(table.corr(), cmap='YlGnBu', annot=True, fmt='.3f')
```

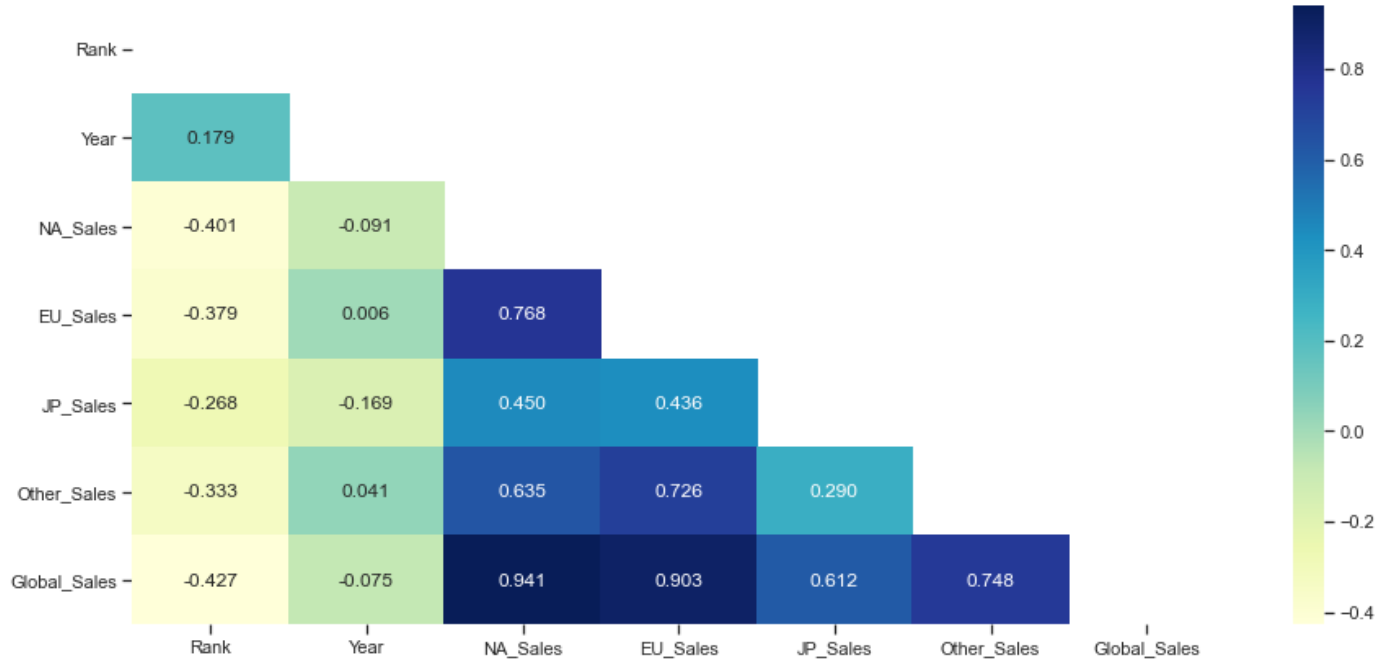
Out[25]: <AxesSubplot:>



In [26]:

```
# Треугольный вариант матрицы
fig, ax = plt.subplots(figsize = (15,7))
mask = np.zeros_like(table.corr(), dtype=np.bool)
mask[np.triu_indices_from(mask)] = True
sns.heatmap(table.corr(), cmap='YlGnBu', mask=mask, annot=True, fmt='.3f')
```

Out[26]: <AxesSubplot:>



In [27]:

```
# Несколько тепловых карт на одном графике
fig, ax = plt.subplots(1, 3, sharex='col', sharey='row', figsize=(20,7))
sns.heatmap(table.corr(method='pearson'), cmap="pink", ax=ax[0],
annot=True, fmt='.2f')
sns.heatmap(table.corr(method='kendall'), cmap="ocean", ax=ax[1],
annot=True, fmt='.2f')
sns.heatmap(table.corr(method='spearman'), cmap="Purples", ax=ax[2],
```

```
annot=True, fmt='.2f')
fig.suptitle('Корреляционные матрицы, построенные различными методами')
ax[0].title.set_text('Пирсон')
ax[1].title.set_text('Кендалл')
ax[2].title.set_text('Спирман')
```