Video Games Sales

Intrudction

Here is a data set about the sales of video games around the world. The data consists of the rank of the games by the best-seller, the distribution of the sales by area, the game genre, etc.
In this project, I will analyze the data and will find out interesting conclusions about the sales of the video game industry.

In [19]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
from mpl_toolkits.mplot3d import Axes3D
```

Reading the data:

In [20]:
```
try:
    video_games_data = pd.read_csv('C:/Users/Matan/Documents/Python/vgsales.csv')
except: print("Something got wrong")
```

In [21]:
```
video_games_data
```

Out[21]:

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37
...	...	...	...	...	...	...	...	...	...	...	...
16593	16596	Woody Woodpecker in Crazy Castle 5	GBA	2002.0	Platform	Kemco	0.01	0.00	0.00	0.00	0.01
16594	16597	Men in Black II: Alien Escape	GC	2003.0	Shooter	Infogrames	0.01	0.00	0.00	0.00	0.01
16595	16598	SCORE International Baja 1000: The Official Game	PS2	2008.0	Racing	Activision	0.00	0.00	0.00	0.00	0.01
16596	16599	Know How 2	DS	2010.0	Puzzle	7G//AMES	0.00	0.01	0.00	0.00	0.01
16597	16600	Spirits & Spells	GBA	2003.0	Platform	Wanadoo	0.01	0.00	0.00	0.00	0.01

16598 rows × 11 columns

The data is ready.

Now I will use the describe function to find any meaningful statistic details.

In [22]:
```
video_games_data.describe()
```

Out[22]:

	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

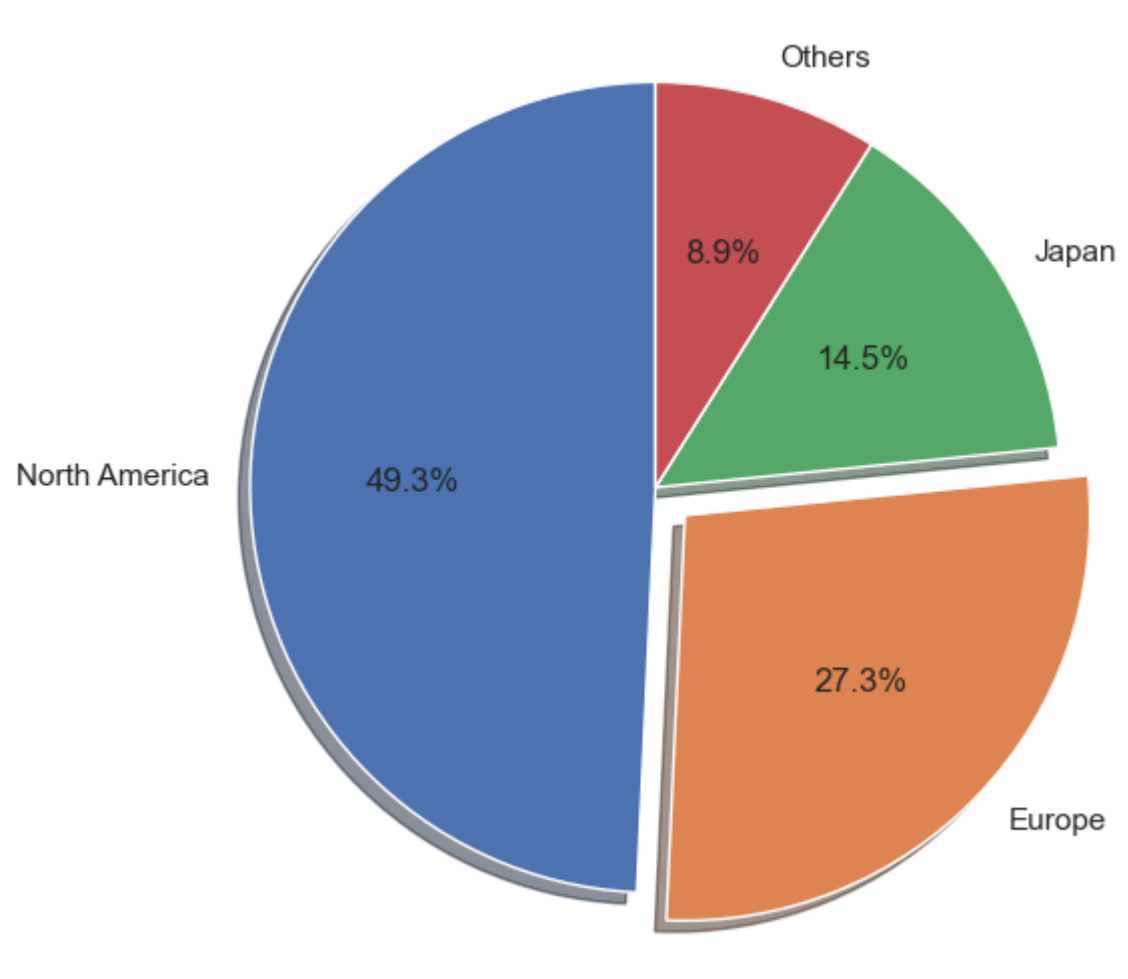
From this table we can learn several things:

- The 50th percentile of the year is 2007 - It means that although companies began to sell video games in 1980 (40 years ago), the past 13 years were the years with the most released video games. It makes a lot of sense, it wasn't so obvious to buy a video game at that time because the technology of those games was only in their beginning and people were used to play others types of games.
- The mean of North America sales is 0.2646 - The highest amount of sales around the world is in North America. 49.3% of the sales around the world are in that area.
- The sales around the world except Japan, Europe, and North America are 0.048 - Only 8.94% of the sales happen in those areas.

The percentage of sales in each area:

In [23]:
```
labels = 'North America', 'Europe', 'Japan', 'Others'
sizes = [video_games_data['NA_Sales'].sum(),video_games_data['EU_Sales'].sum(),video_games_data['JP_Sal
es'].sum(),video_games_data['Other_Sales'].sum()]
explode = (0, 0.1, 0, 0)  # only "explode" the 2nd slice (i.e. 'Hogs')
fig1, ax1 = plt.subplots()
ax1.pie(sizes, explode=explode, labels=labels, autopct='%1.1f%%',
        shadow=True, startangle=90)
ax1.axis('equal')
```

Out[23]:
```
(-1.1100941464523064,
 1.1862146676064456,
 -1.1780485571731871,
 1.1037166301799526)
```



In [24]:
```
print("The number of genres includes in this data set are:",video_games_data['Genre'].value_counts().co
unt())
```

```
The number of genres includes in this data set are: 12
```

In [25]:
```
print("The number of platforms includes in this data set are:",video_games_data['Platform'].value_count
s().count())
```

```
The number of platforms includes in this data set are: 31
```

In [26]:
```
print("The number of publishers includes in this data set are:",video_games_data['Publisher'].value_cou
nts().count())
```

```
The number of publishers includes in this data set are: 578
```

Let's see what is the amount of games released each year:

In [27]:
```
video_games_data['Year'].value_counts().head(10)
```

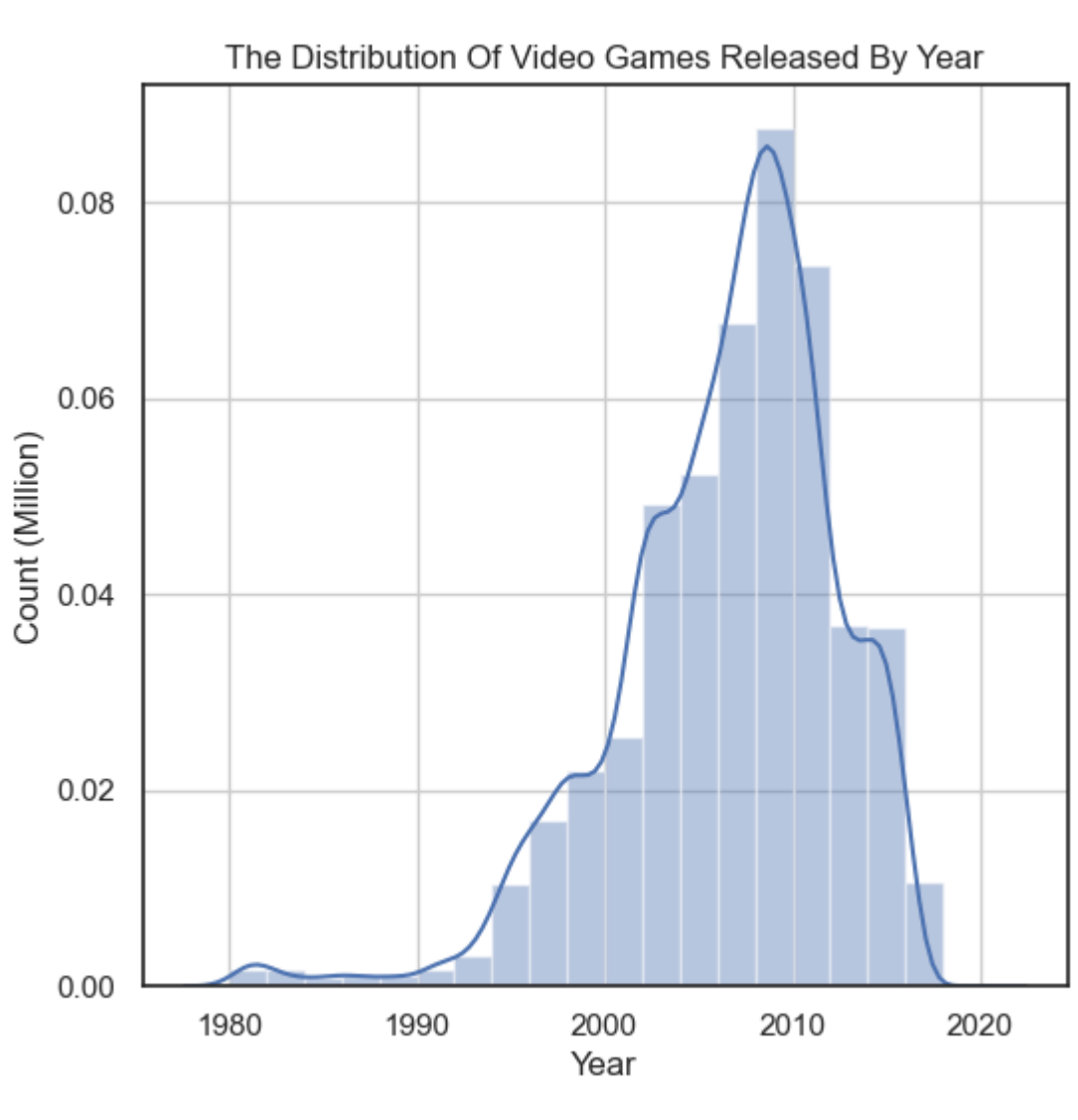
Out[27]:
```
2009.0    1431
2008.0    1428
2010.0    1259
2007.0    1202
2011.0    1139
2006.0    1008
2005.0     941
2002.0     829
2003.0     775
2004.0     763
Name: Year, dtype: int64
```

2008 and 2009 are the years with the biggest amount of video games released.

Let's see the distribution:

In [28]:
```
sns.distplot(video_games_data['Year'], bins=20)
plt.grid()
plt.xlabel('Year')
plt.ylabel("Count (Million)")
plt.title("The Distribution Of Video Games Released By Year")
```

Out[28]:
```
Text(0.5, 1.0, 'The Distribution Of Video Games Released By Year')
```

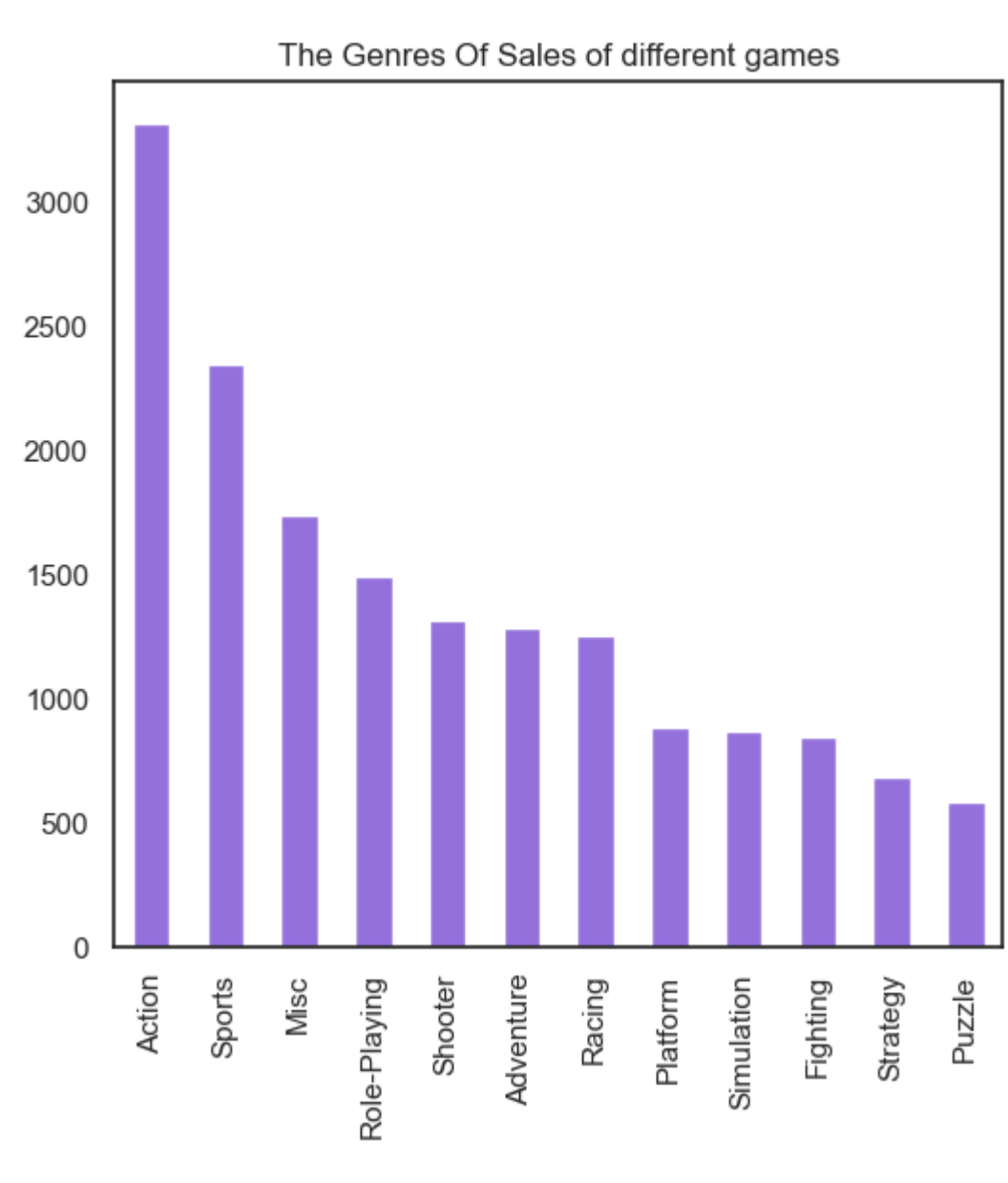


The distribution is almost normal as we see that most of the video games went on sale between 2008-2011. It means that although the technology progress every year, we see that after 2011 there is a decline in the new games offered for sales so there is a possibility that video games were no longer interesting and fun for people.

After we found the most "busy" year of video games, let's check the count of the genre of the video games offered for sale:

In [29]:
```
video_games_data['Genre'].value_counts().plot(kind="bar", title="The Genres Of Sales of different game
s", color = "mediumpurple")
```

Out[29]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x20dda6e6370>
```

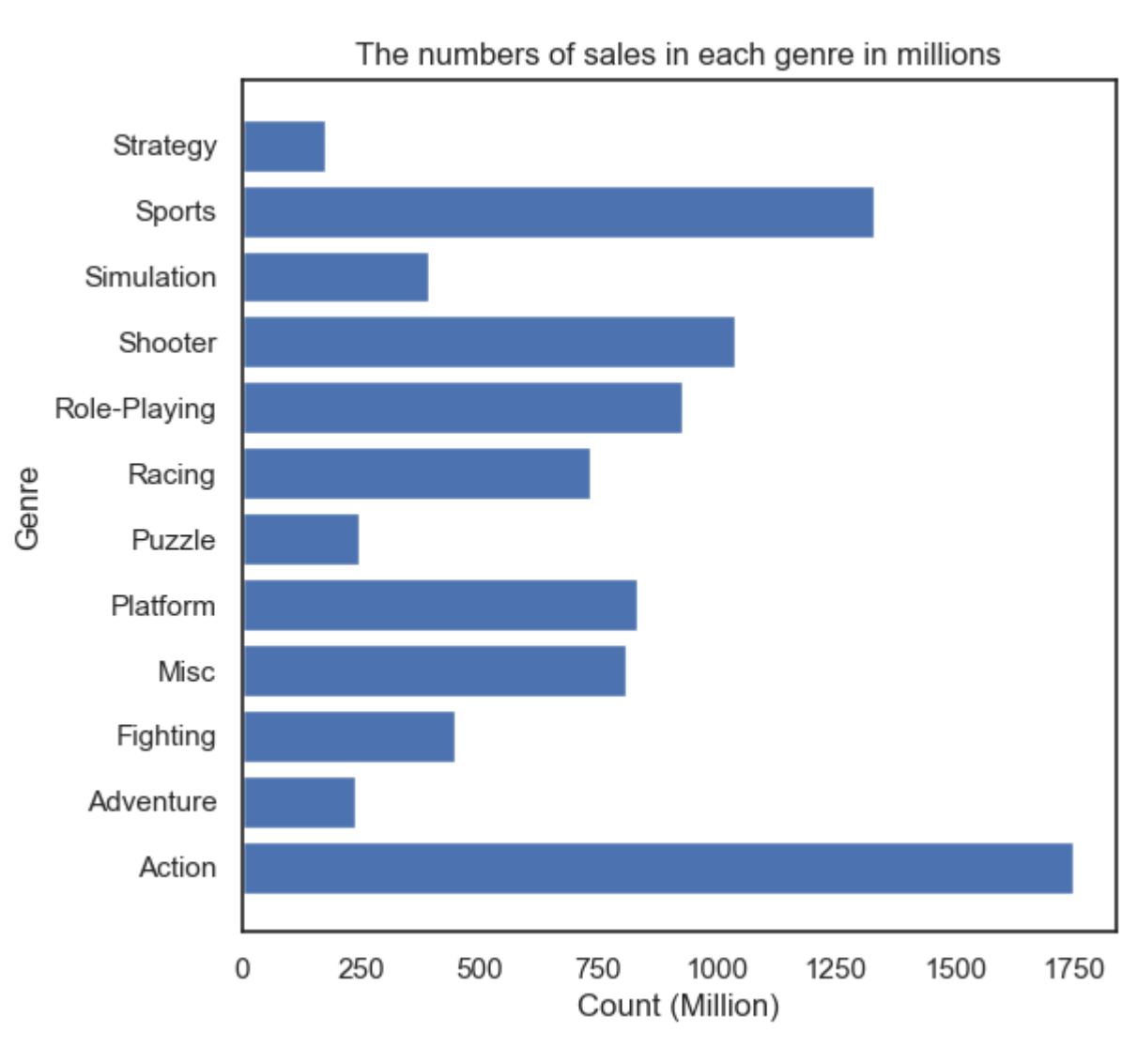


As we can see here Action and Sports are the genres that have the most games offered for sales. however, Puzzle and Strategy genres have the lowest number of games. It can means that people may like video games that are not forcing them to overthink as puzzle and strategy games do.

Now let's check if the amount of video games genres released justify the number of sales in each genre:

In [30]:
```
genres_dict_sales = {}
geners = video_games_data.groupby('Genre')
for name, group in geners:
    genres_dict_sales[name] = group['Global_Sales'].sum()
genres_df_sales = pd.DataFrame(list(genres_dict_sales.items()),columns = ['Genre','Global_Sales'])
plt.barh(genres_df_sales['Genre'], genres_df_sales['Global_Sales'])
plt.yticks(genres_df_sales['Genre'])
plt.ylabel("Genre")
plt.xlabel("Count (Million)")
plt.title("The numbers of sales in each genre in millions")
```

Out[30]:
```
Text(0.5, 1.0, 'The numbers of sales in each genre in millions')
```

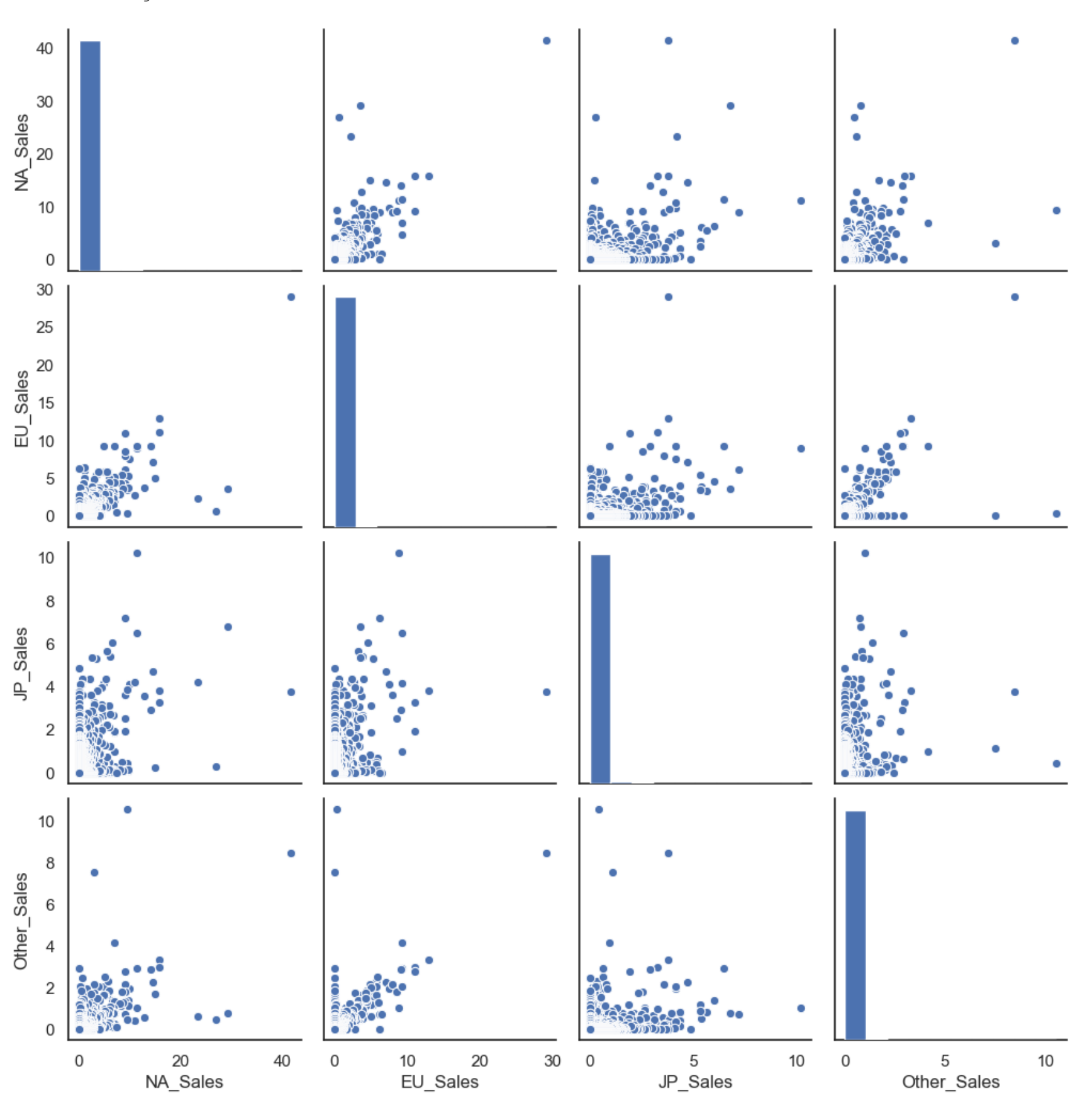


It is, The Action and Sport genres have the biggest sales. However, we can see that although Strategy and Puzzle genres are 2 from 3 lowest genre sales, Adventure genre has the lowest sales except Strategy genre, although the number of different games offered in it is rated 6 out of 12 genres. So the conclusion is that maybe the publishers need to reduce the Adventure games released.

Checking the correlation between the number of sales in each area by scatter plots:

In [31]:
```
sns.set(style="white")
sns.pairplot(video_games_data[['NA_Sales','EU_Sales','JP_Sales','Other_Sales']])
```

Out[31]:
```
<seaborn.axisgrid.PairGrid at 0x20dda1b85b0>
```



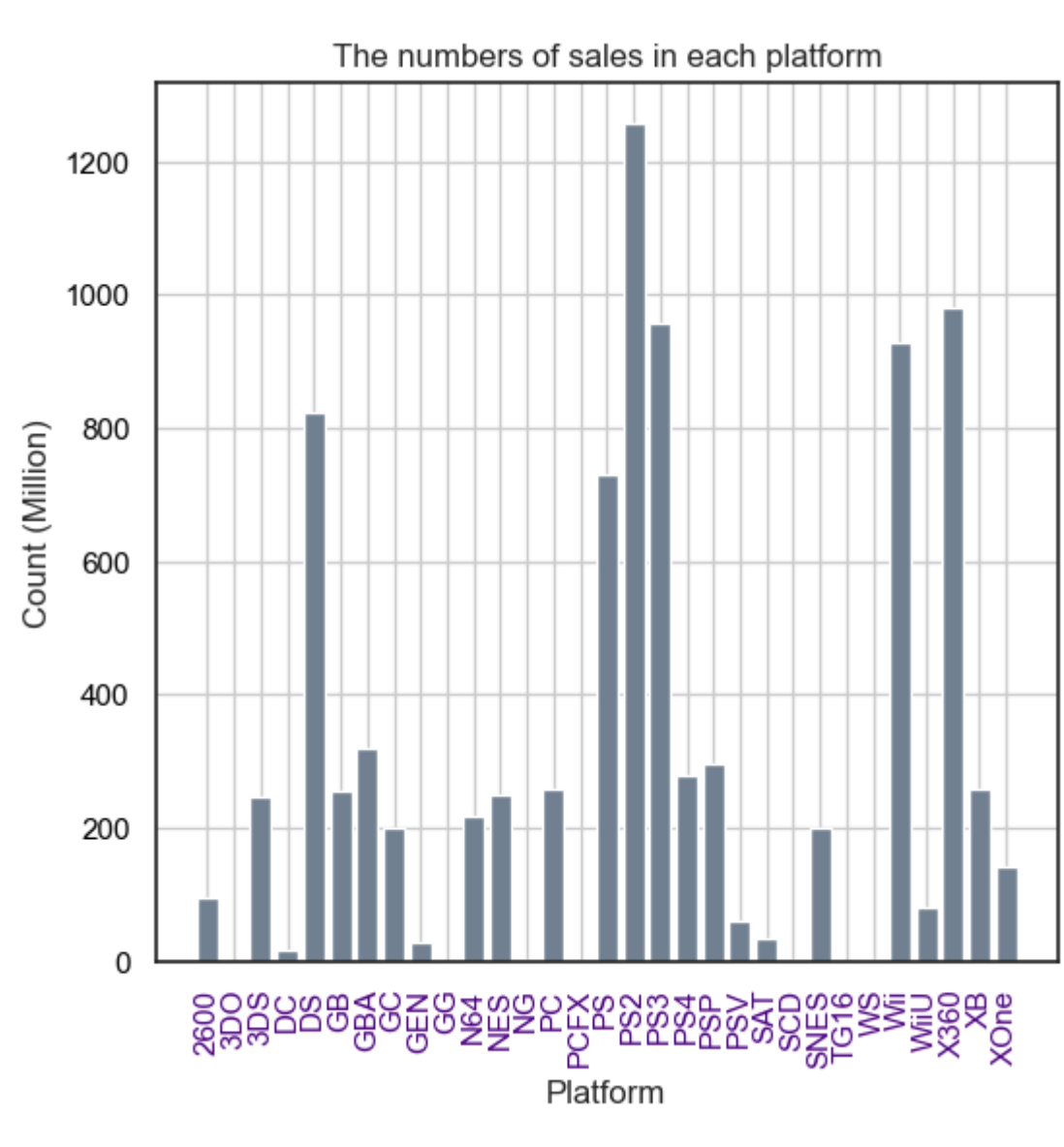
We can see that all of the areas have a positive correlation to the others, but we can see that the strongest connections are between North America - Europe and North America - Other, In those to scatter plots we can build a regression which can show us this positivity correlation. Also, We can see that in every scatter plot most of the points are concentrated in the lower left.

Checking the numbers of generes, platforms and publishers.

The numbers of sales in each platform:

In [32]:
```
platform_name = []
global_sales = []
platforms = video_games_data.groupby('Platform')
for name, group in platforms:
    platform_name.append(name)
    global_sales.append(group['Global_Sales'].sum())
plt.bar(platform_name, global_sales, color = 'slategrey')
plt.xticks(platform_name, color='indigo', rotation=90)
plt.yticks(color='Black')
plt.grid()
plt.ylabel("Count (Million)")
plt.xlabel("Platform")
plt.title("The numbers of sales in each platform")
```

Out[32]:
```
Text(0.5, 1.0, 'The numbers of sales in each platform')
```

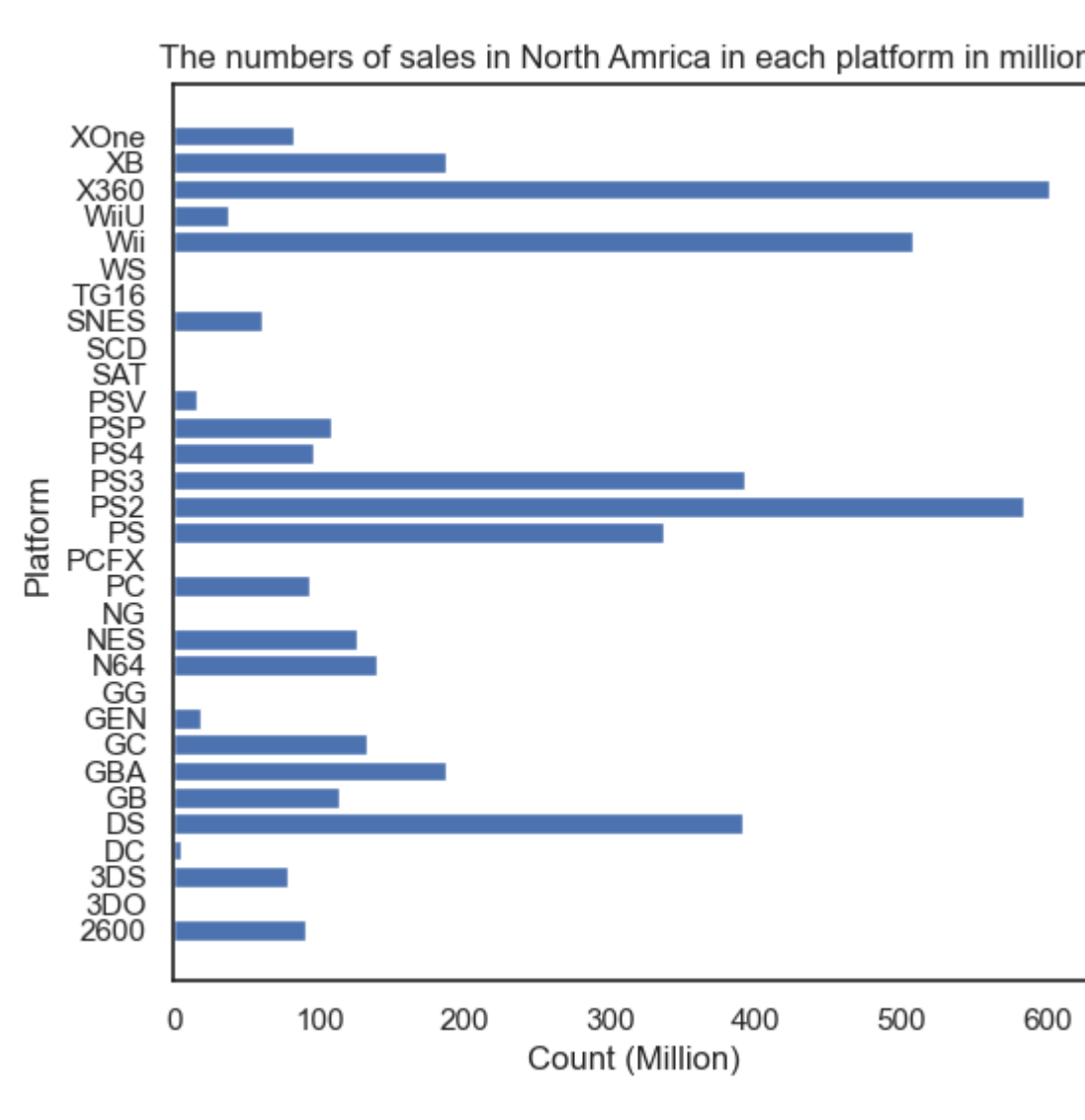


PS2, PS3, Wii, and XBOX are the platforms that people play video games the most.

Now I want to check if there is a diffrent of the sales in each platform in the two main areas: North America and Japan.

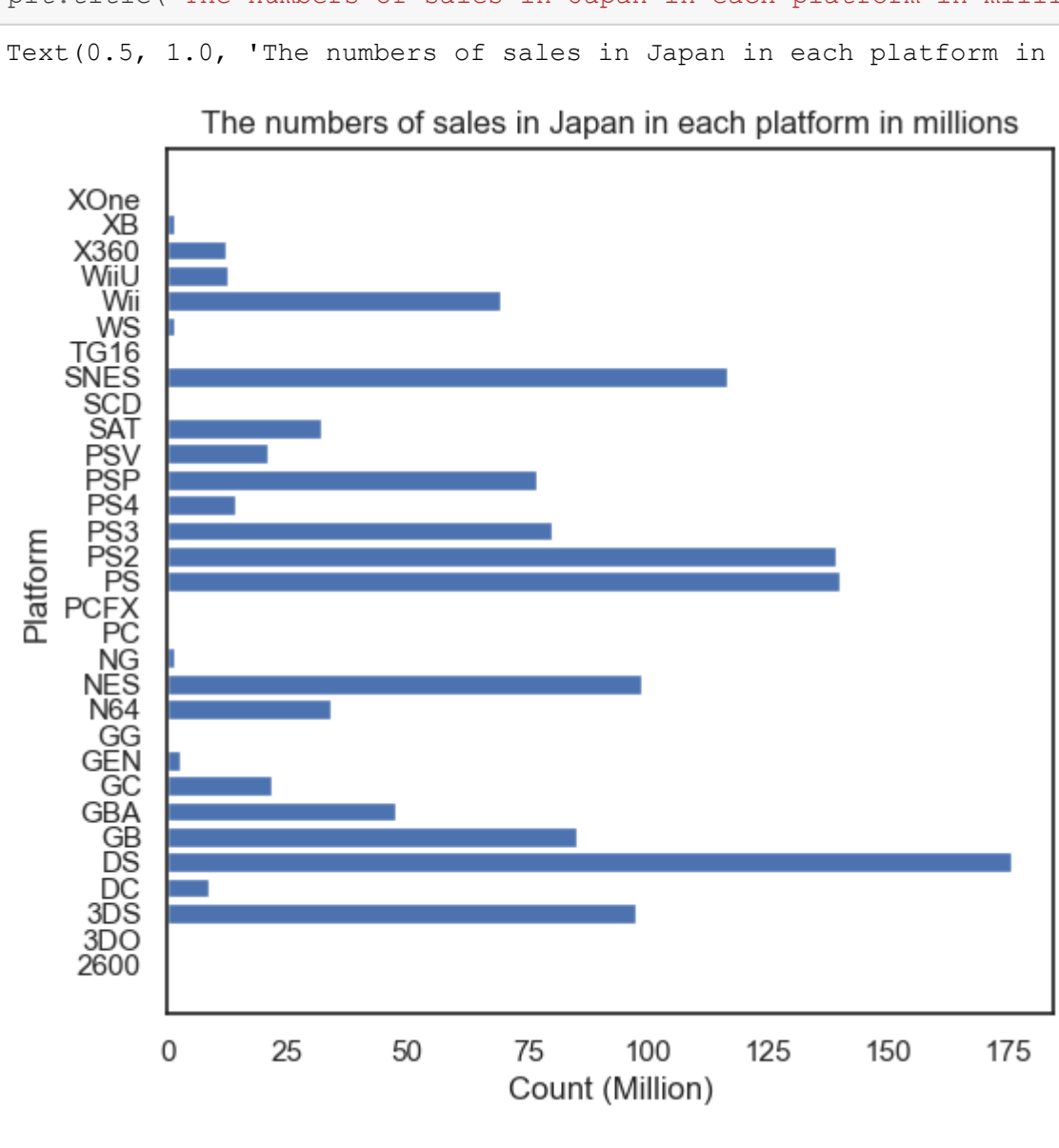
In [33]:
```
platform_dict_sales = {}
platform  = video_games_data.groupby('Platform')
for name, group in platform:
    platform_dict_sales[name] = group['NA_Sales'].sum()
platform_df_sales = pd.DataFrame(list(platform_dict_sales.items()),columns = ['Platform','NA_Sales'])
plt.barh(platform_df_sales['Platform'], platform_df_sales['NA_Sales'])
plt.yticks(platform_df_sales['Platform'])
plt.ylabel("Platform")
plt.xlabel("Count (Million)")
plt.title("The numbers of sales in North Amrica in each platform in millions")
```

Out[33]:
```
Text(0.5, 1.0, 'The numbers of sales in North Amrica in each platform in millions')
```



In [34]:
```
platform_dict_sales = {}
platform  = video_games_data.groupby('Platform')
for name, group in platform:
    platform_dict_sales[name] = group['JP_Sales'].sum()
platform_df_sales = pd.DataFrame(list(platform_dict_sales.items()),columns = ['Platform','JP_Sales'])
plt.barh(platform_df_sales['Platform'], platform_df_sales['JP_Sales'])
plt.yticks(platform_df_sales['Platform'])
plt.ylabel("Platform")
plt.xlabel("Count (Million)")
plt.title("The numbers of sales in Japan in each platform in millions")
```

Out[34]:
```
Text(0.5, 1.0, 'The numbers of sales in Japan in each platform in millions')
```

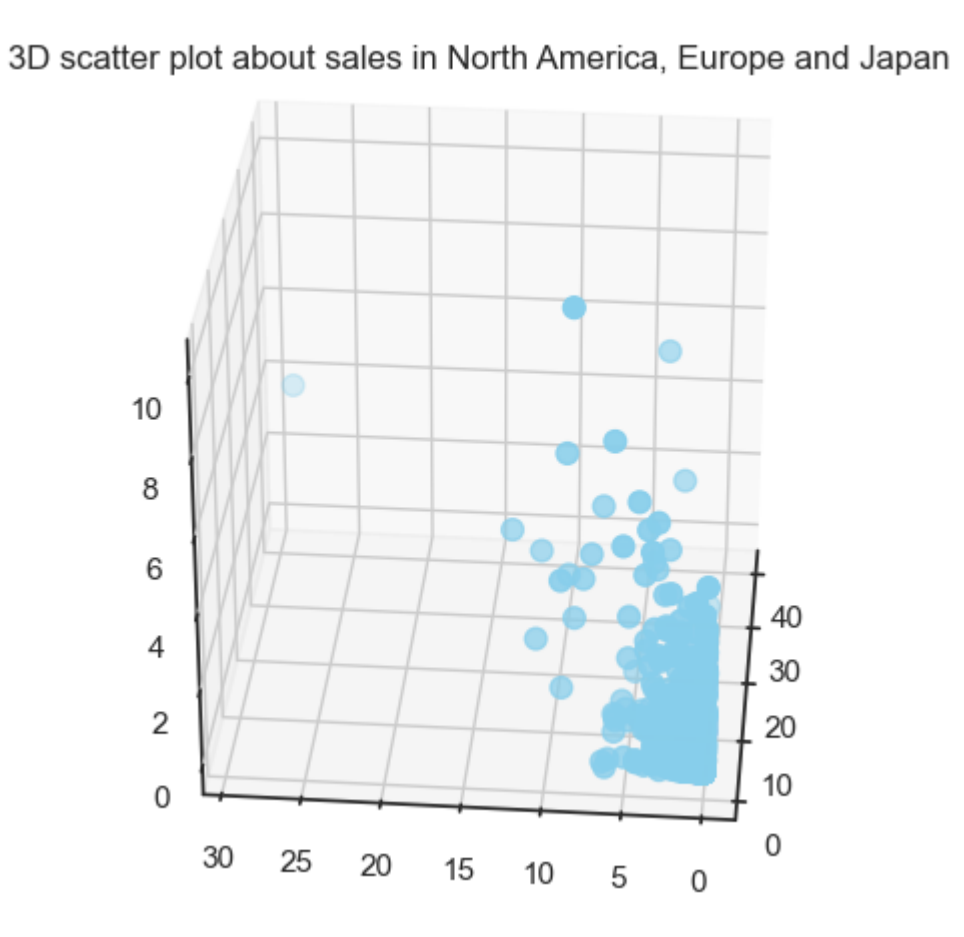


We can see significantly diference between those two area, although we see for example that PS2 is a popular gaming platform in these both areas, the most popular platform that people buy video games for in North Amrica is X360 when in Japan it is among the lower platforms on which games are played. In Japan the highest platform is DS when in North America it is in the top 5 platforms. If we will dig more we would see more difference between these two areas in the using of differet platforms for video games but it is clear that there is a difference.

Using 3D scatter plot about the sales in North America, Europe and Japan.

In [35]:
```
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
ax.scatter(video_games_data['NA_Sales'], video_games_data['EU_Sales'], video_games_data['JP_Sales'], c=
'skyblue', s=60)
ax.view_init(30, 185)
plt.title("3D scatter plot about sales in North America, Europe and Japan")
```

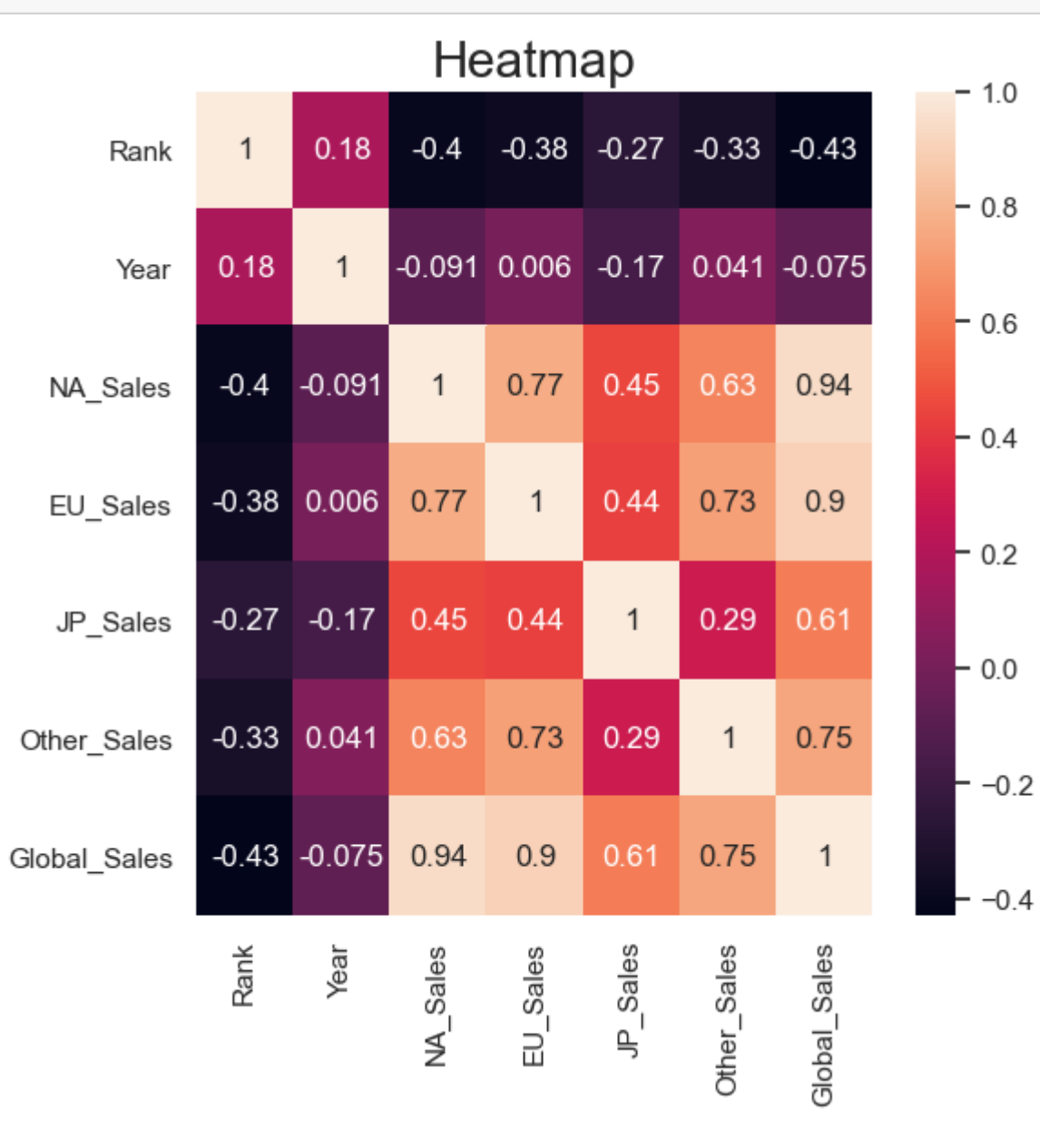
Out[35]:
```
Text(0.5, 0.92, '3D scatter plot about sales in North America, Europe and Japan')
```



Most of the points concentrated in the lower right although we do see a positive correlation between those 3 areas.

Building a correlation heatmap:

In [36]:
```
sns.heatmap(video_games_data.corr(), annot = True)
plt.title('Heatmap', fontsize = 20)
plt.rcParams['figure.figsize'] = [6, 6]
plt.rcParams['figure.dpi'] = 100
```



As we can see on the map, the strongest correlation is between North America to Global sales (0.94) because that as we said before, almost 50% of the sales around the world are in North Arica. Also, there is a positive correlation between Europe and North America sales (0.77) and between Europe to Others sales (0.73).

In conclusion, In this data set, we explored the sales of video games around the world. We saw the different amounts of sales in each area, the difference of using in each platform in each area, the number of games genres released, and their sales, and much more.