

Video games sales analysis



Team Memebers:

Chandrika Rai and Remi Egierd

1) Introduction

Video games first became popular in the early 1970s with Pong, and they have since become an integral part of modern culture and life. Globally, the video game industry will make \$150 billion in revenue in 2020, with 2.7 billion gamers (that's more than a third of the world's population). One cannot deny the financial impact successful video games have on a video-game company. Video games are no longer a niche product with a small segment of consumers. Research has shown that video game consumers are a very diverse group and in all walks of life. The industry has influenced virtually a large consumer segment, in that most of the current population has grown up with video games and played them for entertainment or education in some way.

1.2 The Purpose:

The research objective of this proposal is to analyse the video game sales identify which specific factors influence the North American, European, Japanese, other, and Global Sales of video games from 1980 to 2020. As a step toward this goal, we will examine 1) the record of video game sales across most major video game consoles and 2) test what factors have the greatest impact on sales in either North America, Europe, or Japan. It was important to use a large range of games from all of the different video-game categories to truly get a broad view of the market.

1.3 Why is this question important?

By analyzing the data from this study, game developers can get a detailed picture of the video game market, which will enable them to see what types of games are preferred and how various variables affect sales. These publishers and developers want to know which variables have the greatest impact on sales of their video games since they have a vested interest in making money. By analyzing the results of our analyses, developers will be able to decide on what platforms, genres, or countries to release their future video games. It will also help individual game companies like Nintendo and Microsoft better perceive their consumer base by understanding

1.4 Summary

According to this analysis, the top games, platforms, publishers, genres and more are summarized for different geographic regions. Sales and customer satisfaction will both be improved thanks to these results.

The trends we observed are:

The number of games released each year is associated with the total number of global sales for that year.

North America and European Union have similar trends in genres and platforms while Japan has different ones.

Platforms made by companies in specific regions are more popular in those regions.

2) Main

2.1 Data Processing & Cleaning

```
In [2]: import pandas as pd
import matplotlib.pyplot as plt
raw=r'Video_Games_Sales.csv'
df=pd.read_csv(raw)
```

```
In [27]: # This Python 3 environment comes with many helpful analytics libraries instal
led
# It is defined by the kaggle/python docker image: https://github.com/kaggle/d
ocker-python
# For example, here's several helpful packages to load in

import numpy as np # Linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the "../input/" directory.
# For example, running this (by clicking run or pressing Shift+Enter) will lis
t all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# Any results you write to the current directory are saved as output.
```

```
In [3]: pd.set_option('display.max_columns', None)
import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
%matplotlib inline
import seaborn as sns
import plotly.io as pio
import plotly.graph_objects as go
from plotly.offline import init_notebook_mode, iplot
from sklearn.preprocessing import StandardScaler
import warnings
warnings.filterwarnings('ignore')
```

2.1.1 Description of the dataset

Name of all Fields

Variable name	Description
Name	name of game
Platform	consul game is made for
Year_of_Release	year game was released
Genre	type or category of game
Publisher	game publisher
NA_Sales	sales in North America (in millions of units)
EU_Sales	sales in Europe (in millions of units)
JP_Sales	sales in Japan (in millions of units)
Other_Sales	sales outside North America, Europe, Japan
Global_Sales	total sales (in millions of units)
Critic_Score	aggregate score complied by Metacritic staff
Critic_Count	number of critics used in coming up with the Criticscore
User_Score	score by Metacritic's subscribers
User_count	number of users who gave the userscore
Developer	party responsible for creating the game
Rating	the ESRB ratings

Numbers of Rows and Columns

```
In [3]: df.shape
```

```
Out[3]: (16719, 16)
```

Column name and type

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

```
Out[4]: Index(['Name', 'Platform', 'Year_of_Release', 'Genre', 'Publisher', 'NA_Sales',  
              'EU_Sales', 'JP_Sales', 'Other_Sales', 'Global_Sales', 'Critic_Score',  
              'Critic_Count', 'User_Score', 'User_Count', 'Developer', 'Rating'],  
             dtype='object')
```

Description of all Fields


```
In [5]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 16719 entries, 0 to 16718  
Data columns (total 16 columns):  
#   Column                Non-Null Count  Dtype    
---  ---  
0   Name                  16717 non-null  object   
1   Platform              16719 non-null  object   
2   Year_of_Release       16450 non-null  float64  
3   Genre                 16717 non-null  object   
4   Publisher             16665 non-null  object   
5   NA_Sales              16719 non-null  float64  
6   EU_Sales              16719 non-null  float64  
7   JP_Sales              16719 non-null  float64  
8   Other_Sales           16719 non-null  float64  
9   Global_Sales          16719 non-null  float64  
10  Critic_Score           8137 non-null   float64  
11  Critic_Count           8137 non-null   float64  
12  User_Score             10015 non-null  object   
13  User_Count             7590 non-null   float64  
14  Developer              10096 non-null  object   
15  Rating                 9950 non-null   object   
dtypes: float64(9), object(7)  
memory usage: 2.0+ MB
```

```
In [6]: #numeric features
df.describe()
```

```
Out[6]:
```

	Year_of_Release	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Glob
count	16450.000000	16719.000000	16719.000000	16719.000000	16719.000000	16719
mean	2006.487356	0.263330	0.145025	0.077602	0.047332	0.533
std	5.878995	0.813514	0.503283	0.308818	0.186710	1.547
min	1980.000000	0.000000	0.000000	0.000000	0.000000	0.010
25%	2003.000000	0.000000	0.000000	0.000000	0.000000	0.060
50%	2007.000000	0.080000	0.020000	0.000000	0.010000	0.170
75%	2010.000000	0.240000	0.110000	0.040000	0.030000	0.470
max	2020.000000	41.360000	28.960000	10.220000	10.570000	82.53



Data Type of all Fields

```
In [7]: df.dtypes
```

```
Out[7]: Name          object
Platform         object
Year_of_Release  float64
Genre            object
Publisher         object
NA_Sales          float64
EU_Sales          float64
JP_Sales          float64
Other_Sales       float64
Global_Sales      float64
Critic_Score      float64
Critic_Count      float64
User_Score        object
User_Count        float64
Developer         object
Rating            object
dtype: object
```

Basic Descriptive Features of the Data

Number of Total Observations

```
In [6]: df.shape
```

```
Out[6]: (16719, 16)
```

Number of Observations by Field

```
In [7]: df.count()
```

```
Out[7]: Name          16717  
Platform        16719  
Year_of_Release  16450  
Genre           16717  
Publisher        16665  
NA_Sales         16719  
EU_Sales         16719  
JP_Sales         16719  
Other_Sales      16719  
Global_Sales     16719  
Critic_Score      8137  
Critic_Count      8137  
User_Score       10015  
User_Count        7590  
Developer        10096  
Rating           9950  
dtype: int64
```

Number of Missing Values by Field

```
In [8]: df.isnull().sum()
```

```
Out[8]: Name          2  
Platform          0  
Year_of_Release   269  
Genre             2  
Publisher         54  
NA_Sales          0  
EU_Sales          0  
JP_Sales          0  
Other_Sales       0  
Global_Sales      0  
Critic_Score     8582  
Critic_Count     8582  
User_Score       6704  
User_Count       9129  
Developer        6623  
Rating           6769  
dtype: int64
```

```
In [8]: df = df[df["Year_of_Release"].notnull()]
df = df[df["Genre"].notnull()]
df = df[df["Publisher"].notnull()]
df['Year_of_Release']=df['Year_of_Release'].astype('int64')
df['User_Score']=df['User_Score'].replace('tbd',0).astype('float64')
```

3. Analysis and Explanations

Descriptive Analytics

Descriptive analytics examines what happened by identifying the data. In the video game sales data, we aim to answer the question: What specific factors influence the sales in North America, European Union, Japan, and other regions.

Diagnostic Analytics

Diagnostic analytics examines why the data is the way that it is and why it has certain trends. In the video game sales data, we aim to answer the question: Why do specific factors like genre and publisher increase global sales?

Genre of the Game


Hypothesis 1: The video games sales are dependent on specific regions.

web reference:https://www.causeweb.org/usproc/sites/default/files/usclap/2017-2/Factors_that_Impact_Video_Game_Sales.pdf
(https://www.causeweb.org/usproc/sites/default/files/usclap/2017-2/Factors_that_Impact_Video_Game_Sales.pdf)

Video game genre is a classification assigned to a video game based on its gameplay interaction rather than visual or narrative differences.

Currently, there are over 30 different types of video games, including programming games like CodinGame, where players must use programming languages like Java and C++ to advance through the levels.

Lets analyze the Average and Gross Sales of Video games and the Number of Games released in each Genre to check our hypothesis.

 image.png


```

In [7]: sns.set_style("whitegrid")
        trace1=go.Scatter(
            x=df.groupby(['Genre']).mean().reset_index()['Genre'],
            y=df.groupby(['Genre']).mean().reset_index()['NA_Sales'],
            mode='lines+markers',
            name='North America Sales',
            marker = dict(size=8),
            line=dict(color = '#FA8072',width=2.5))
        trace2=go.Scatter(
            x=df.groupby(['Genre']).mean().reset_index()['Genre'],
            y=df.groupby(['Genre']).mean().reset_index()['EU_Sales'],
            mode='lines+markers',
            name='Europe Sales',
            marker = dict(size=8),
            line=dict(color = '#6495ED',width=2.5))

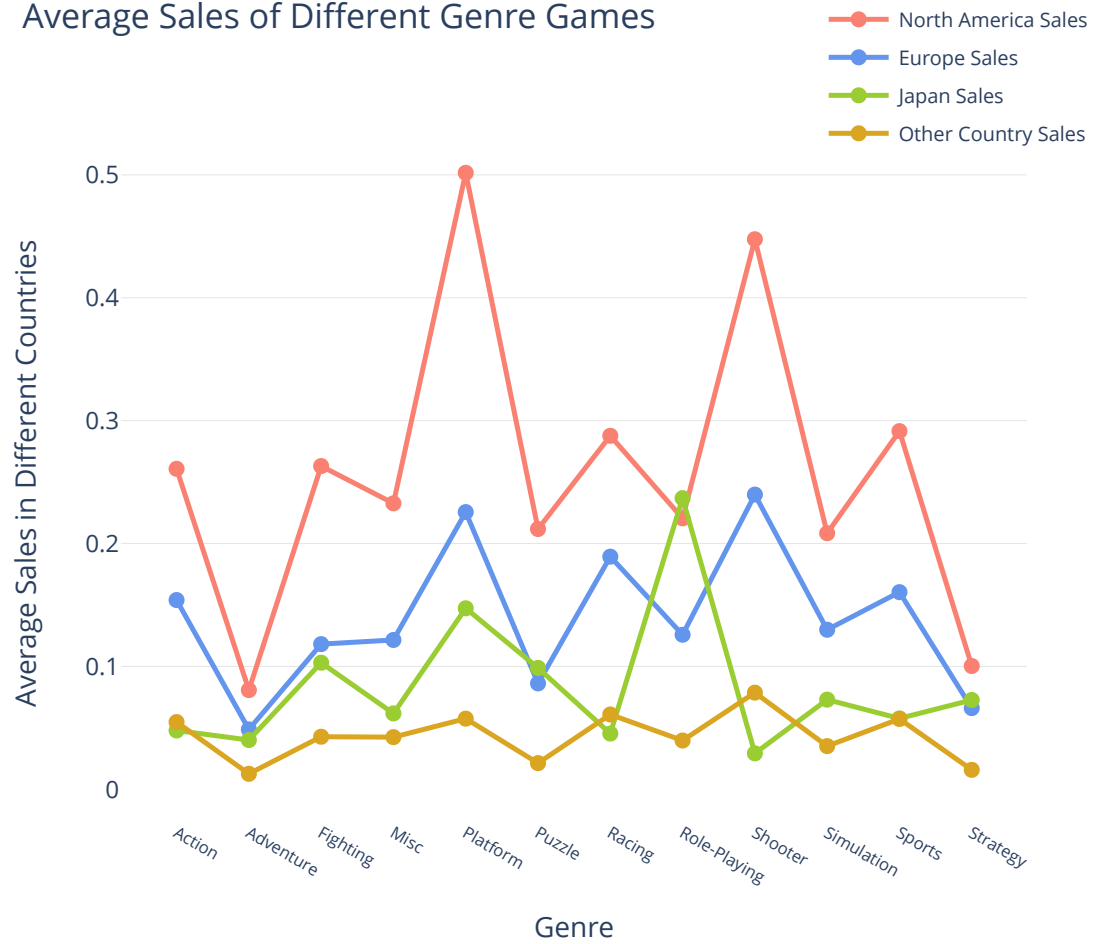
        trace3=go.Scatter(
            x=df.groupby(['Genre']).mean().reset_index()['Genre'],
            y=df.groupby(['Genre']).mean().reset_index()['JP_Sales'],
            mode='lines+markers',
            name='Japan Sales',
            marker = dict(size=8),
            line=dict(color = 'yellowgreen',width=2.5))

        trace4=go.Scatter(
            x=df.groupby(['Genre']).mean().reset_index()['Genre'],
            y=df.groupby(['Genre']).mean().reset_index()['Other_Sales'],
            mode='lines+markers',
            name='Other Country Sales',
            marker = dict(size=8),
            line=dict(color = '#DAA520',width=2.5))
        edit_df=[trace1,trace2,trace3,trace4]
        layout=dict(
            legend=dict(x=0.77, y=1.2, font=dict(size=10)), legend_orientation
            ="v",
            title="Average Sales of Different Genre Games",
            xaxis=dict(title="Genre",tickfont=dict(size=8.35),zeroline=False,gridcolor="white"),
            yaxis=dict(title='Average Sales in Different Countries',gridcolor="#DCDCDC"),
            plot_bgcolor='white')

        fig=dict(data=edit_df,layout=layout)
        iplot(fig)

```

Average Sales of Different Genre Games



```
In [5]: fig=plt.figure(figsize=(30,15))
plt.subplots_adjust(left=0.30, wspace=0.25, hspace=0.45)

plt.subplot(2,3,1)
plt.title('Sales of Genres in European Union')
sns.barplot(y='Genre', x='EU_Sales', data=df.groupby('Genre').sum().EU_Sales.sort_values(ascending=False).reset_index())
plt.ylabel('Genre')
plt.xlabel('Sales in European Union')

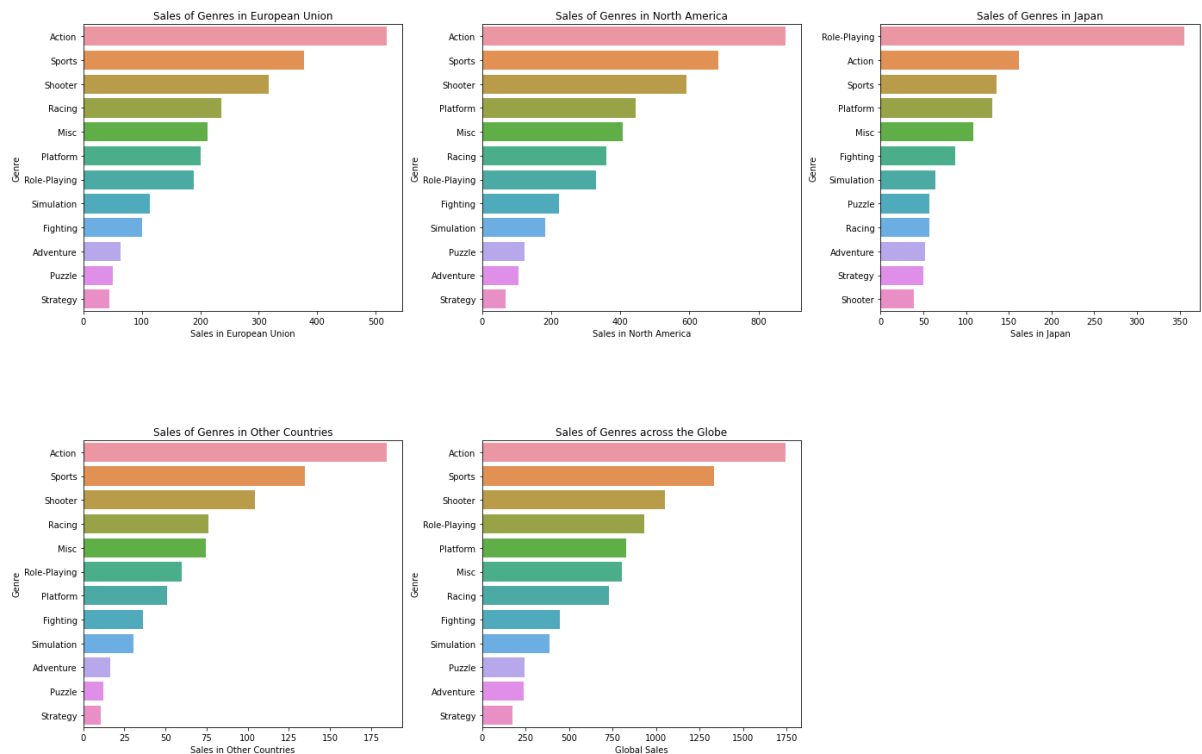
plt.subplot(2,3,2)
plt.title('Sales of Genres in North America')
sns.barplot(y='Genre', x='NA_Sales', data=df.groupby('Genre').sum().NA_Sales.sort_values(ascending=False).reset_index())
plt.ylabel('Genre')
plt.xlabel('Sales in North America')

plt.subplot(2,3,3)
plt.title('Sales of Genres in Japan')
sns.barplot(y='Genre', x='JP_Sales', data=df.groupby('Genre').sum().JP_Sales.sort_values(ascending=False).reset_index())
plt.ylabel('Genre')
plt.xlabel('Sales in Japan')

plt.subplot(2,3,4)
plt.title('Sales of Genres in Other Countries')
sns.barplot(y='Genre', x='Other_Sales', data=df.groupby('Genre').sum().Other_Sales.sort_values(ascending=False).reset_index())
plt.ylabel('Genre')
plt.xlabel('Sales in Other Countries')

plt.subplot(2,3,5)
plt.title('Sales of Genres across the Globe')
sns.barplot(y='Genre', x='Global_Sales', data=df.groupby('Genre').sum().Global_Sales.sort_values(ascending=False).reset_index())
plt.ylabel('Genre')
plt.xlabel('Global Sales')
```

```
Out[5]: Text(0.5, 0, 'Global Sales')
```



It is visible that average sales is associated with specific region.

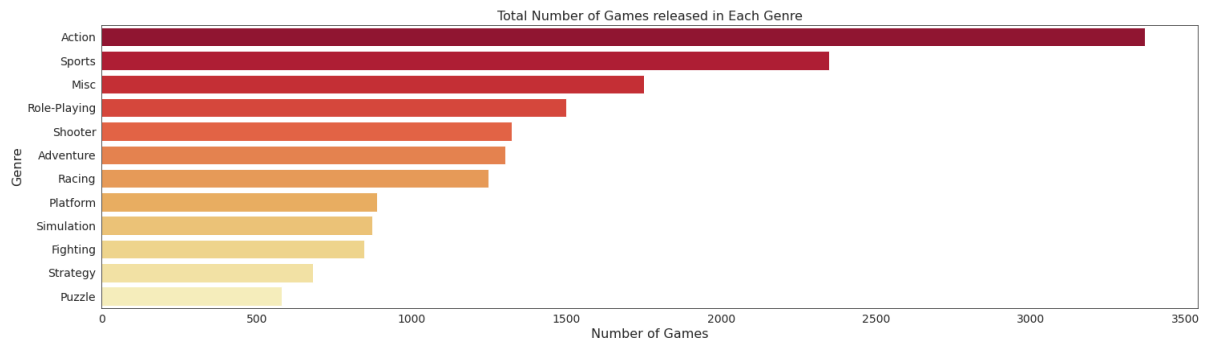
The amount of action games produced far exceeds other genres. Game developers should focus on this genre as it holds the most market share. Sports, role-playing, shooter, adventure, and racing games also hold a large proportion of the market share. Game developers should focus on these genres as well and should produce less puzzle and strategy games. Different gaming platforms are conducive to different genres, which developers should also consider. Action, sports, and shooter genres have the highest rate of sale in the global, American, and EU market. In Japan, role-playing games have the highest rate of sales by a significant margin. Action, sports, and shooter games should be developed in English to access the American and EU market while role-playing games should be primarily developed in Japanese. When developing new games, publishers need to know what genres are most popular for their target geographic market.

let's check the Gross sale

Action Games are the highest sold games around the World except in Japan, where Role-Playing games make the highest sales. which means market share of genre might vary in different region.

Playing Action Games improves perceptual decision making, or the ability to identify and select targets, Speed of processing, or how quickly one can respond to a target.

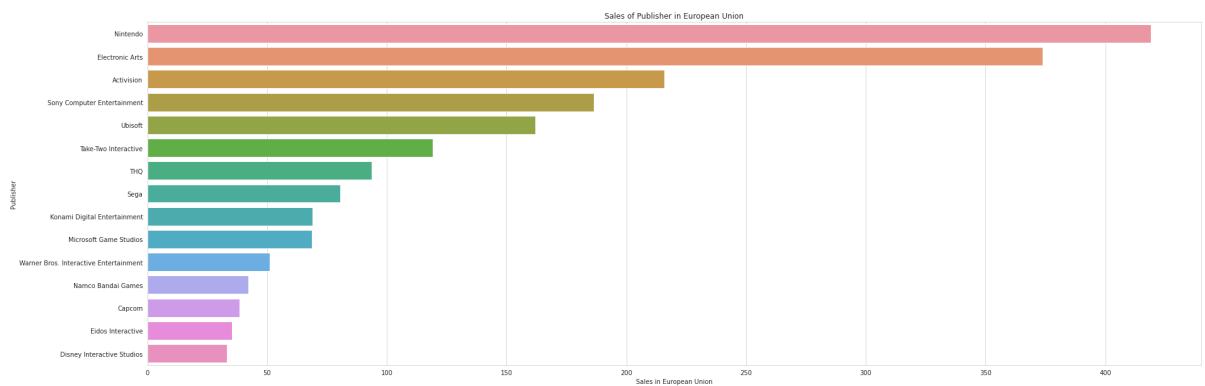
```
In [7]: fig=plt.figure(figsize=(24.5,22))
sns.set_style("white")
plt.subplot2grid((3,1), (1,0))
plt.xticks(fontsize=14)
plt.yticks(fontsize=14)
plt.title('Total Number of Games released in Each Genre',fontdict={'fontsize':
16})
sns.barplot(y=df['Genre'].value_counts().index,x=df['Genre'].value_counts(),pa
lette='YlOrRd_r')
plt.ylabel('Genre',fontdict={'fontsize':16})
plt.xlabel('Number of Games',fontdict={'fontsize':16});
```



Publisher

```
In [14]: fig=plt.figure(figsize=(30,10))
plt.title('Sales of Publisher in European Union')
sns.barplot(y='Publisher', x='EU_Sales', data=df.groupby('Publisher').sum().EU
_Sales.sort_values(ascending=False).reset_index().head(15))
plt.ylabel('Publisher')
plt.xlabel('Sales in European Union')
```

Out[14]: Text(0.5, 0, 'Sales in European Union')

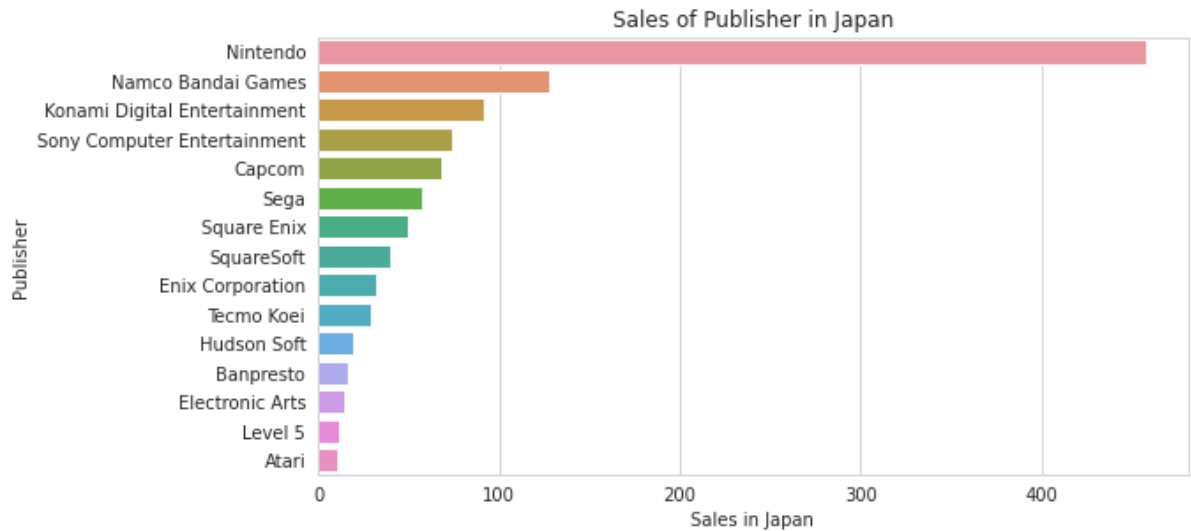


```

In [19]: fig=plt.figure(figsize=(30,10))
plt.subplot(2,3,2)
plt.title('Sales of Publisher in Japan')
sns.barplot(y='Publisher', x='JP_Sales', data=df.groupby('Publisher').sum().JP_Sales.sort_values(ascending=False).reset_index().head(15))
plt.ylabel('Publisher')
plt.xlabel('Sales in Japan')

```

Out[19]: Text(0.5, 0, 'Sales in Japan')

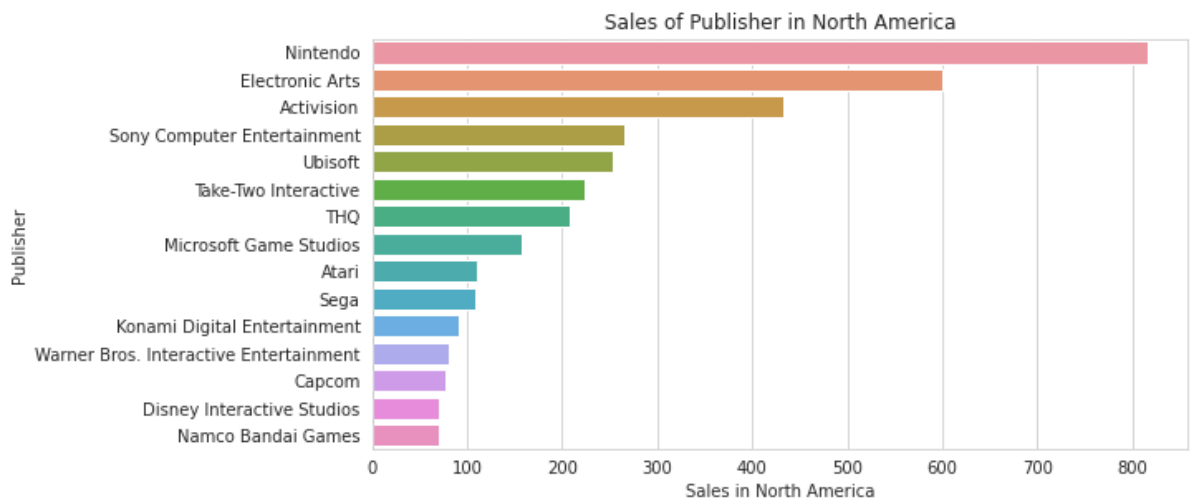


```

In [18]: fig=plt.figure(figsize=(30,10))
plt.subplot(2,3,3)
plt.title('Sales of Publisher in North America')
sns.barplot(y='Publisher', x='NA_Sales', data=df.groupby('Publisher').sum().NA_Sales.sort_values(ascending=False).reset_index().head(15))
plt.ylabel('Publisher')
plt.xlabel('Sales in North America')

```

Out[18]: Text(0.5, 0, 'Sales in North America')

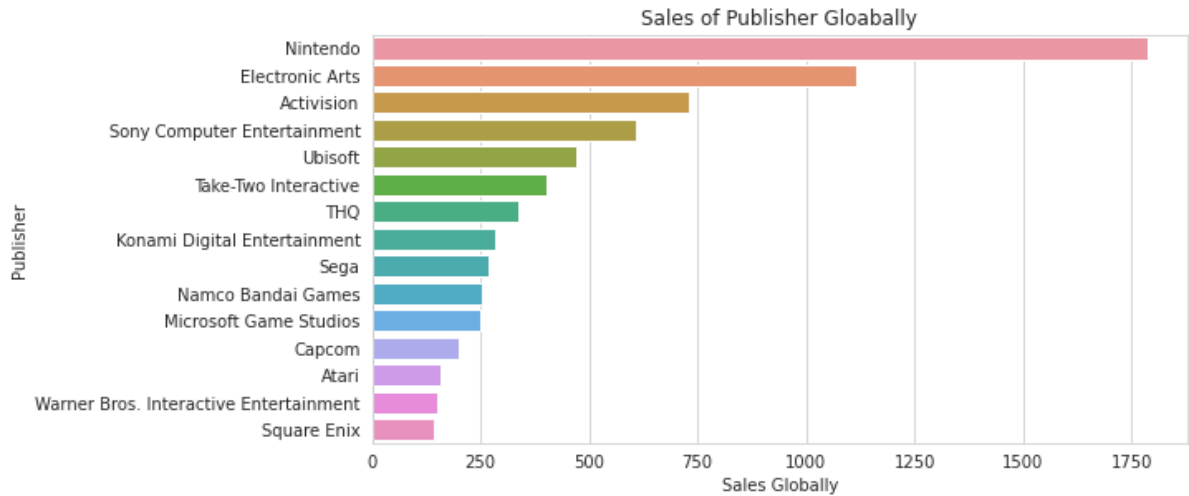


```

In [17]: fig=plt.figure(figsize=(30,10))
plt.subplot(2,3,4)
plt.title('Sales of Publisher Gloabally')
sns.barplot(y='Publisher', x='Global_Sales', data=df.groupby('Publisher').sum()
).Global_Sales.sort_values(ascending=False).reset_index().head(15))
plt.ylabel('Publisher')
plt.xlabel('Sales Globally')

```

Out[17]: Text(0.5, 0, 'Sales Globally')

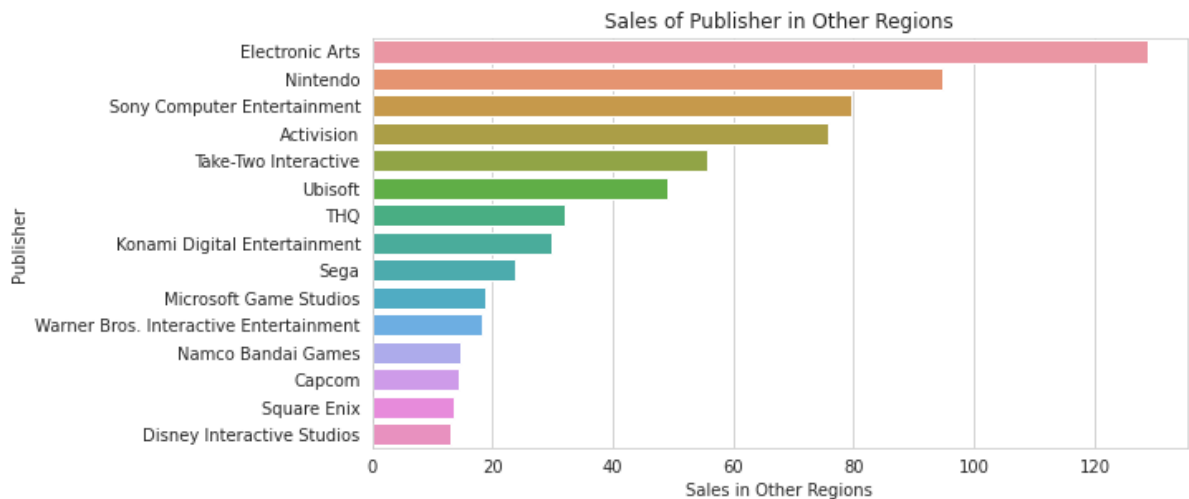


```

In [16]: fig=plt.figure(figsize=(30,10))
plt.subplot(2,3,5)
plt.title('Sales of Publisher in Other Regions')
sns.barplot(y='Publisher', x='Other_Sales', data=df.groupby('Publisher').sum()
.Other_Sales.sort_values(ascending=False).reset_index().head(15))
plt.ylabel('Publisher')
plt.xlabel('Sales in Other Regions')

```

Out[16]: Text(0.5, 0, 'Sales in Other Regions')



Most games are published by Electronic Arts (EA), Activision, Namco, and Ubisoft. Game developers should generally focus their efforts on making games for these publishers as they hold greater amounts of market share and are more likely to have better resources for developers. However, not all publishers make games for all platforms. EA and Activision have the highest proportion of market share because they exclusively make PC games, which is the most accessible platform.

Hypothesis 2: The Number of Games released and the Global Sales made over the years are correlated.

Web reference: https://www.causeweb.org/usproc/sites/default/files/usclap/2017-2/Factors_that_Impact_Video_Game_Sales.pdf
(https://www.causeweb.org/usproc/sites/default/files/usclap/2017-2/Factors_that_Impact_Video_Game_Sales.pdf) Lets analyze

Year of Game Release:

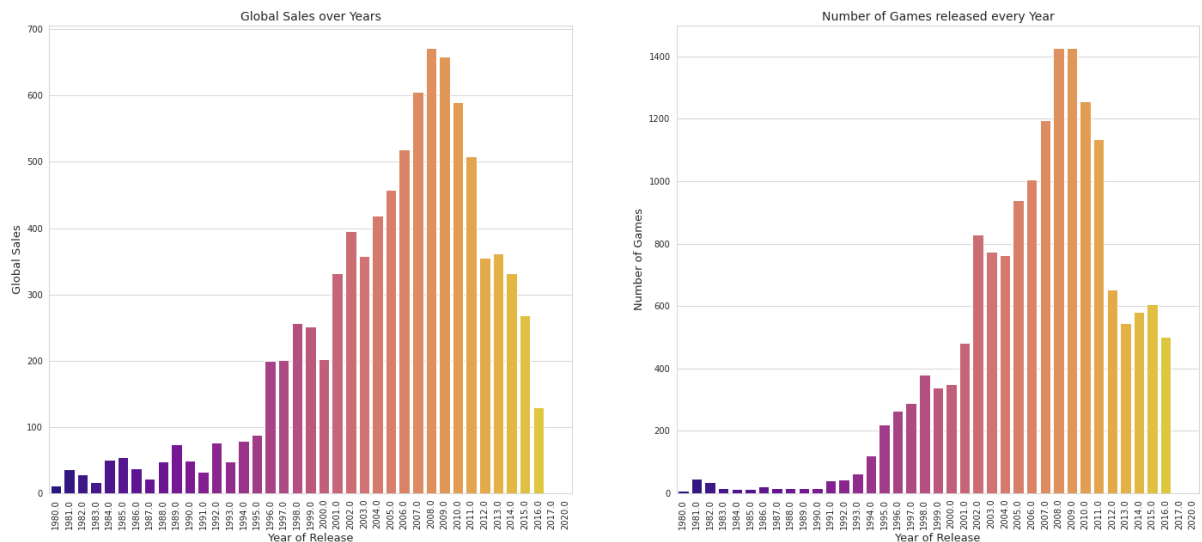
Since the early 1970s, video games have been around. The evolution of video game systems took less than 30 years. Games today can even be connected to the internet and played with people all over the world via modem.

We should also look at the most popular games and top publishers with largest global sales, and the average and gross sales of video games each year.


```
In [13]: fig=plt.figure(figsize=(24,10))
plt.subplots_adjust(left=None, wspace=None, hspace=None)
sns.set_style("whitegrid")

plt.subplot(1, 2, 1)
plt.title('Global Sales over Years',fontdict={'fontsize':14})
plt.xticks(rotation=90)
sns.barplot(x='Year_of_Release',y='Global_Sales',data=df.groupby(df['Year_of_Release']).sort_values()).sum().Global_Sales.reset_index(),palette='plasma')
plt.ylabel('Global Sales',fontdict={'fontsize':13})
plt.xlabel('Year of Release',fontdict={'fontsize':13})

plt.subplot(1, 2, 2)
plt.title('Number of Games released every Year',fontdict={'fontsize':14})
plt.xticks(rotation=90)
sns.barplot(x=df.Year_of_Release.value_counts().index, y=df.Year_of_Release.value_counts(),palette='plasma');
plt.ylabel('Number of Games',fontdict={'fontsize':13})
plt.xlabel('Year of Release',fontdict={'fontsize':13});
```



The most games were released in the year 2008, as well as the highest global sales.

Global Sales over the years seem to be correlated with the number of games released. However, why are they related?

Different platforms, or gaming consoles, will impact sales. This is important to know because publishers need to know which of their gaming platforms are selling the most games to decide what kind of games to produce. For example, in North America, the Xbox 360 had the highest sales, then PS2, then Wii. This is possibly because the Xbox is published by Microsoft, an American company. Across the globe, the PS2 had the highest sales by a significant margin with the Xbox 360, PS3, and Wii following behind. The European Union had first the PS2, then PS3 and Xbox360. In contrast, Japan had the highest sale rates for the DS, then PS, PS2, which are all published by Japanese companies.

```

In [6]: sc = StandardScaler()
Year_Count_Sales=df.groupby(df['Year_of_Release']).apply(lambda x: pd.Series({
    'Count'          : x['Name'].count(),
    'Global_Sales'    : x['Global_Sales'].sum()})).reset_index()

Year_Count_Sales_Scaled = pd.concat([Year_Count_Sales['Year_of_Release'],pd.DataFrame(sc.fit_transform(Year_Count_Sales[['Count', 'Global_Sales']]),columns=['Count', 'Global_Sales'])],axis=1)

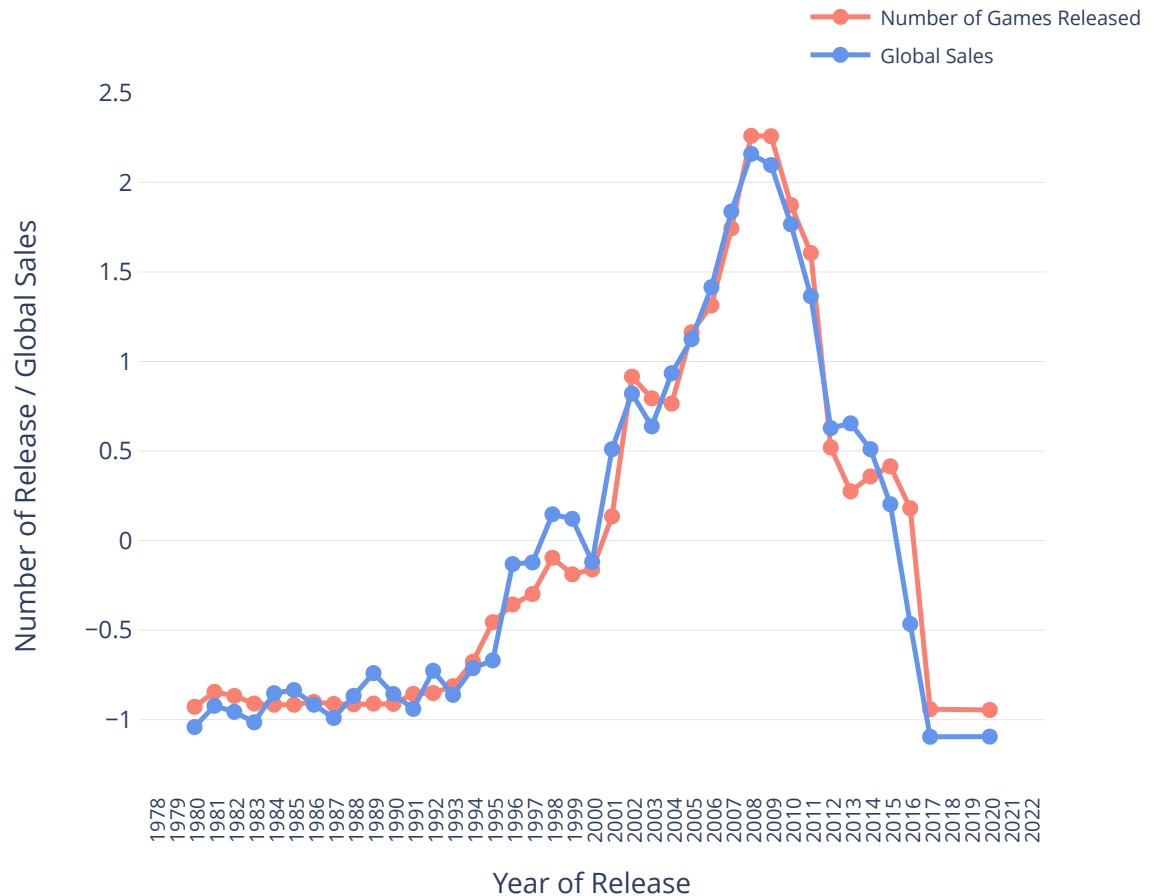
fig = go.Figure(data=[
    go.Scatter(
        x=Year_Count_Sales_Scaled['Year_of_Release'],
        y=Year_Count_Sales_Scaled['Count'],
        mode='lines+markers',
        name='Number of Games Released',
        marker = dict(size=8),
        line=dict(color = '#FA8072',width=2.5),
        text=Year_Count_Sales['Count'],

        hovertemplate = '<i>Year</i>: %{x}'
                        '<br><i>Number of Games</i>: %{text}<br>'),
    go.Scatter(
        x=Year_Count_Sales_Scaled['Year_of_Release'],
        y=Year_Count_Sales_Scaled['Global_Sales'],
        mode='lines+markers',
        name='Global Sales',
        marker = dict(size=8),
        line=dict(color = '#6495ED',width=2.5),
        text = Year_Count_Sales['Global_Sales'],
        hovertemplate = '<i>Year</i>: %{x}'
                        '<br><i>Global_Sales</i>: %{text}<br>')
],layout=dict(legend=dict(x=0.73, y=1.15, font=dict(size=10)),legend_orientati
on="v",title="Relationship between Number of Releases and Global Sales",
    xaxis=dict(tickmode = 'linear',tickangle=-90,tickfont=dict(size=10
),title="Year of Release",tickwidth=5,ticklen=8,zeroline=True,gridcolor="whit
e",
    showline=True),
    yaxis=dict(title="Number of Release / Global Sales",zeroline=True,
showline=True,gridcolor="#DCDCDC",
    showgrid=True,
    zerolinecolor='#DCDCDC',
    zerolinewidth=1)
,plot_bgcolor='white'))

fig.show()

```

Relationship between Number of Releases and Global Sales



So.. As we could clearly see, the number of games released over time correlates with the number of global sales made, as we could see that the number of games released increases with global sales as well.

Hypothesis 3: The number of releases and the global sales of games across platforms are positively co related.

Game Platform:

Video game consoles are machines that are used to play video games and are attached to special controllers that provide control of the game. In this analysis we will find which platform has the highest sales and how it effect the sales.

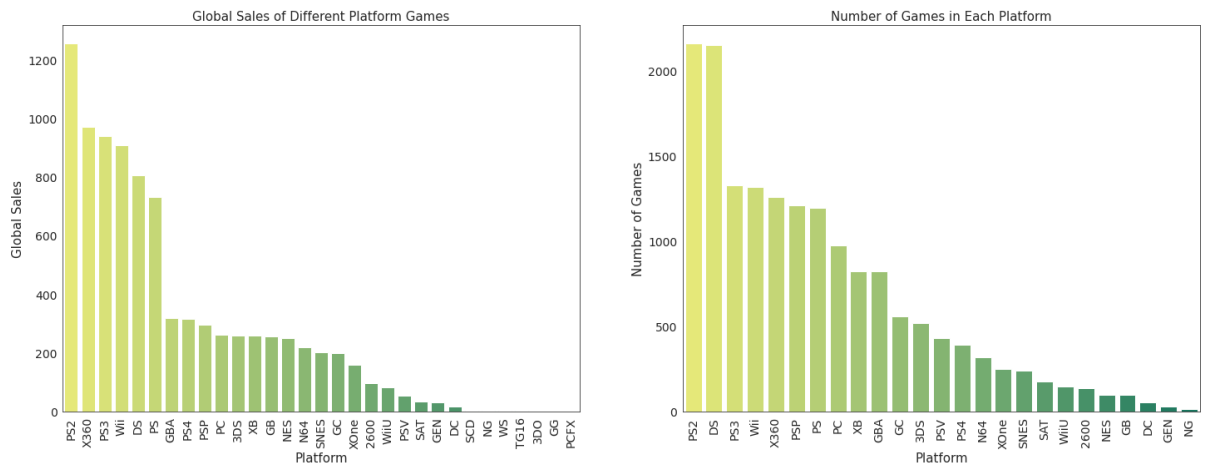
```

In [9]: fig=plt.figure(figsize=(24.5,8.5))
# plt.subplots_adjust(left=None, wspace=None, hspace=None)
sns.set_style("white")

plt.subplot(1, 2, 1)
plt.xticks(rotation=90)
plt.xticks(fontsize=14)
plt.yticks(fontsize=14)
plt.title('Global Sales of Different Platform Games',fontdict={'fontsize':15})
Plat_sales=df.groupby(df['Platform']).sum().Global_Sales.sort_values(ascending
=False).reset_index()
sns.barplot(x='Platform',y='Global_Sales',data=Plat_sales,palette='summer_r');
plt.ylabel('Global Sales',fontdict={'fontsize':15})
plt.xlabel('Platform',fontdict={'fontsize':15})

plt.subplot(1, 2, 2)
plt.xticks(rotation=90)
plt.xticks(fontsize=14)
plt.yticks(fontsize=14)
plt.title('Number of Games in Each Platform',fontdict={'fontsize':15})
sns.barplot(x=df.Platform.value_counts()[df.Platform.value_counts()>10].index,
y=df.Platform.value_counts()[df.Platform.value_counts()>10],palette='summer_r'
);
plt.ylabel('Number of Games',fontdict={'fontsize':15})
plt.xlabel('Platform',fontdict={'fontsize':15});

```



The PlayStation 2 platform has released the most games and has the highest sales worldwide.

More than 155 million PlayStation 2 consoles have been sold to date.

There have been more than 3,874 games released for the PS2 since it launched, and there have been more than 1.5 billion sales of games.

Let's now examine the relationship between the number of releases and the global sales of games across platforms. if they are positively co related or not.

```

In [5]: sc = StandardScaler()
Plat_Count_Sales=df.groupby(df['Platform']).apply(lambda x: pd.Series({
    'Count'          : x['Name'].count(),
    'Global_Sales'    : x['Global_Sales'].sum()})).reset_index()

Plat_Count_Sales_Scaled=pd.concat([Plat_Count_Sales['Platform'],pd.DataFrame(s
c.fit_transform(Plat_Count_Sales[['Count','Global_Sales']]),columns=['Count',
'Global_Sales'])],axis=1)

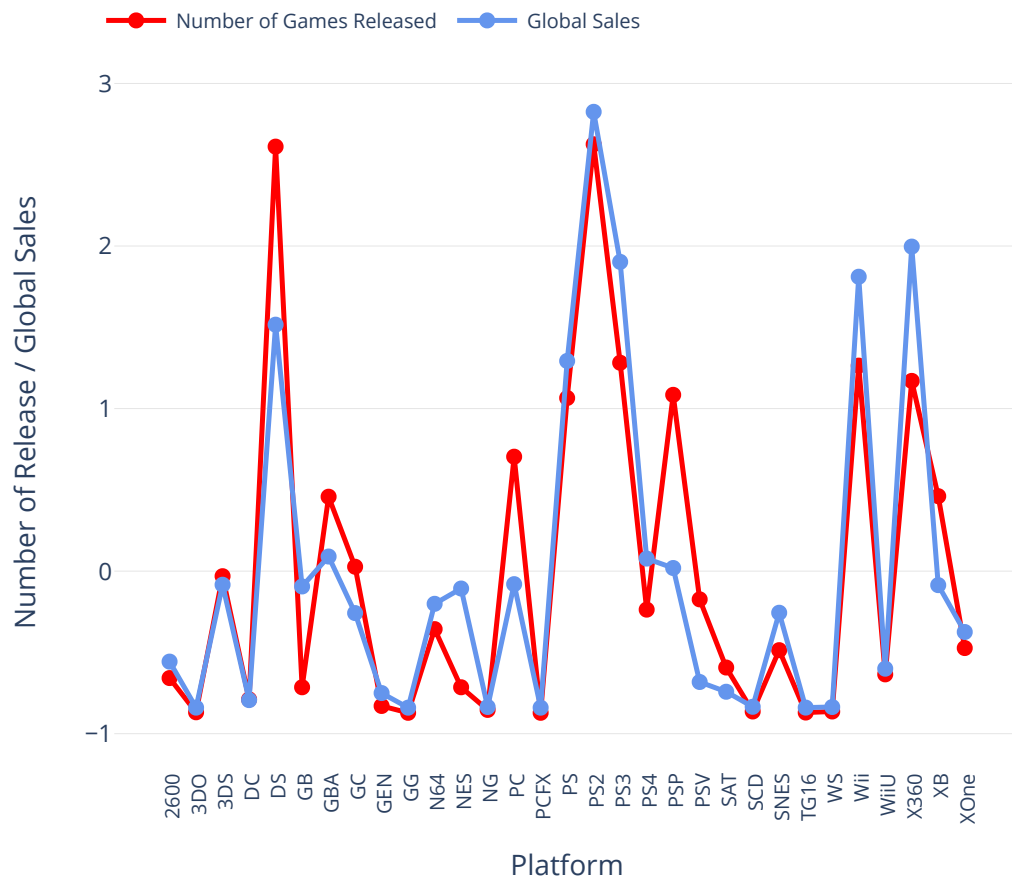
fig = go.Figure(data=[
    go.Scatter(
        x=Plat_Count_Sales_Scaled['Platform'],
        y=Plat_Count_Sales_Scaled['Count'],
        mode='lines+markers',
        name='Number of Games Released',
        marker = dict(size=8),
        line=dict(color = 'Red',width=2.5),
        text=Plat_Count_Sales['Count'],

        hovertemplate = '<i>Platform</i>: %{x}'
                        '<br><i>Number of Games</i>: %{text}<br>'),
    go.Scatter(
        x=Plat_Count_Sales_Scaled['Platform'],
        y=Plat_Count_Sales_Scaled['Global_Sales'],
        mode='lines+markers',
        name='Global Sales',
        marker = dict(size=8),
        line=dict(color = '#6495ED',width=2.5),
        text = Plat_Count_Sales['Global_Sales'],
        hovertemplate = '<i>Platform</i>: %{x}'
                        '<br><i>Global_Sales</i>: %{text}<br>')
],layout=dict(legend=dict(x=-0.02, y=1.11,font=dict(size=10)),legend_orientati
on="h",title="Relationship between Number of Releases and Global Sales of Diff
erent Plaform Games",
    xaxis=dict(tickmode = 'linear',tickangle=-90,tickfont=dict(size=10
),title="Platform",tickwidth=5,ticklen=8,zeroline=True,gridcolor="white",
    showline=True),
    yaxis=dict(tick0=-1,dtick=1,title="Number of Release / Global Sale
s",showticklabels=True,gridcolor="#DCDCDC",
    showgrid=True,
    zerolinecolor='#DCDCDC',
    zerolinewidth=1)
,plot_bgcolor='white'))

fig.show()

```

Relationship between Number of Releases and Global Sales of Different



There appears to be a positive correlation between the Number of titles and the number of sales, but consoles like the NES have a negative correlation, where there are fewer games, but the number of sales is high. Consoles like the PSP also have negative correlations, which means that the number of games is high, but global sales are low. Not many platforms offer negative correlations as well.

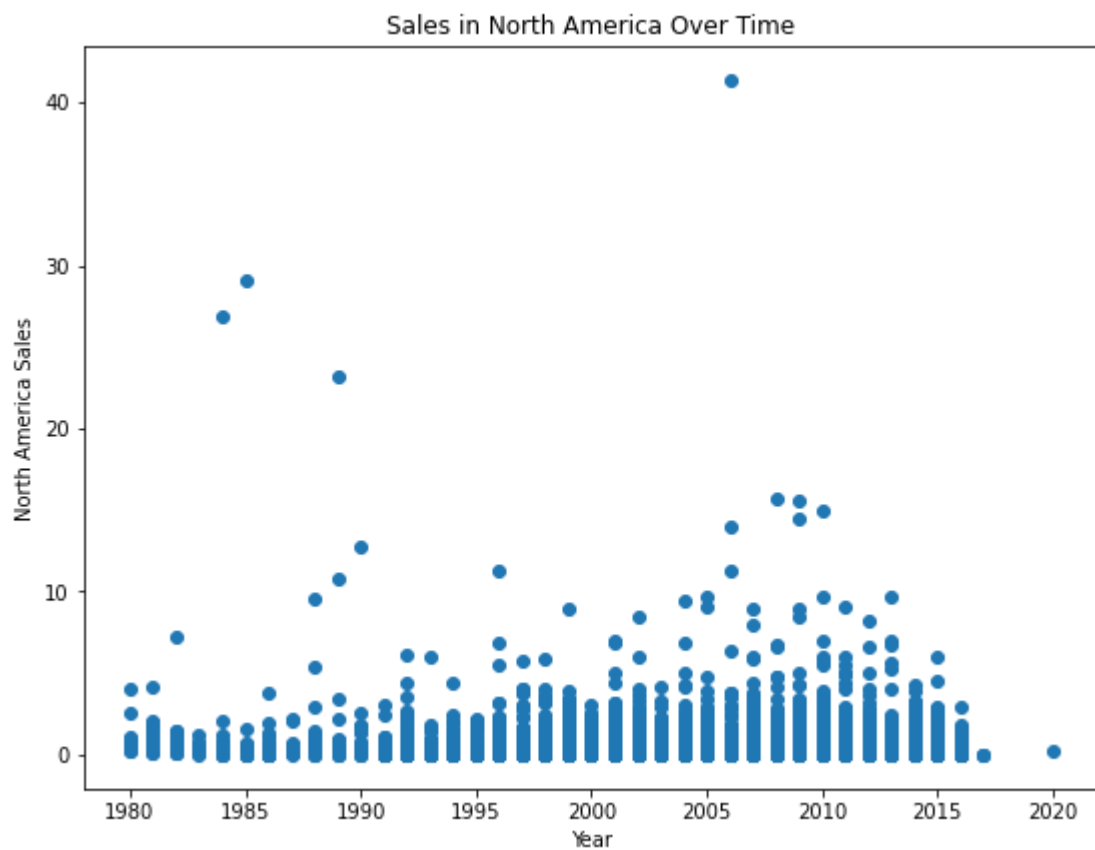
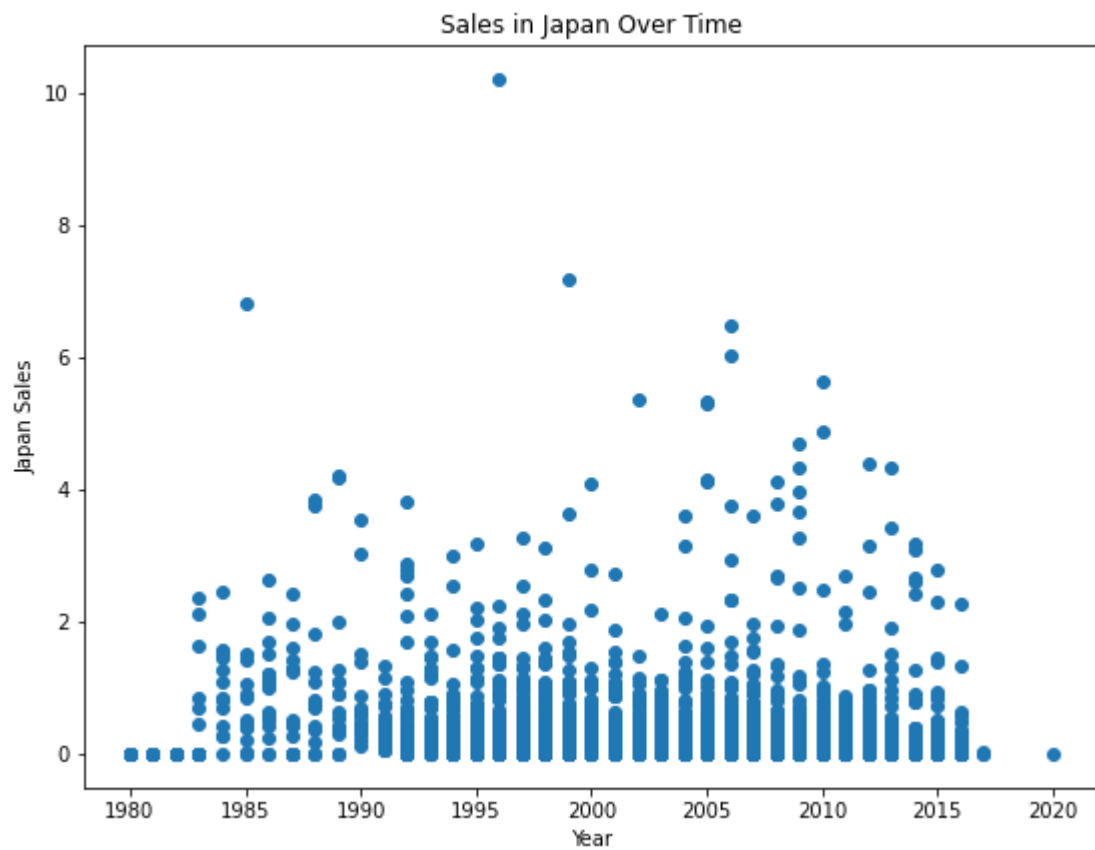
```
In [4]: fig=plt.figure(figsize=(20,15))
plt.subplot(2,2,1)
plt.scatter(df['Year_of_Release'],df['JP_Sales'])
plt.title('Sales in Japan Over Time')
plt.ylabel('Japan Sales')
plt.xlabel('Year')
plt.show()

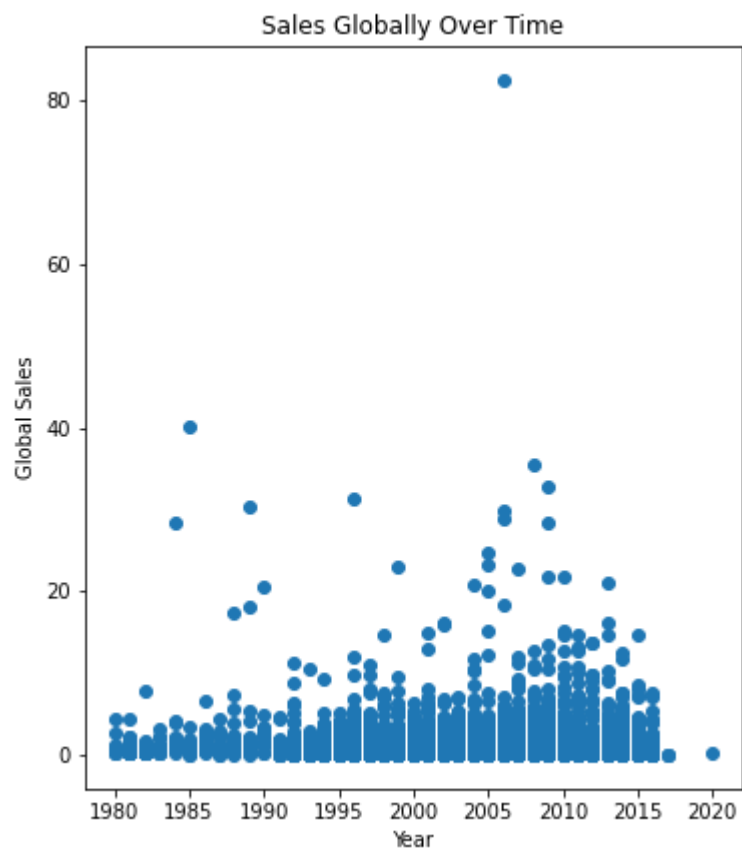
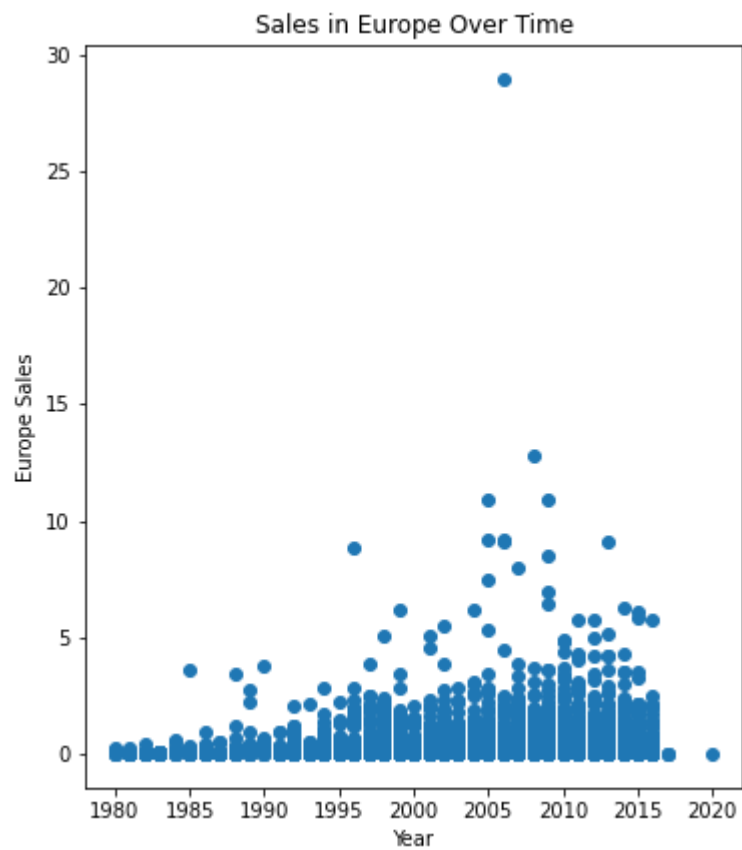
fig=plt.figure(figsize=(20,15))
plt.subplot(2,2,2)
plt.scatter(df['Year_of_Release'],df['NA_Sales'])
plt.title('Sales in North America Over Time')
plt.ylabel('North America Sales')
plt.xlabel('Year')
plt.show()

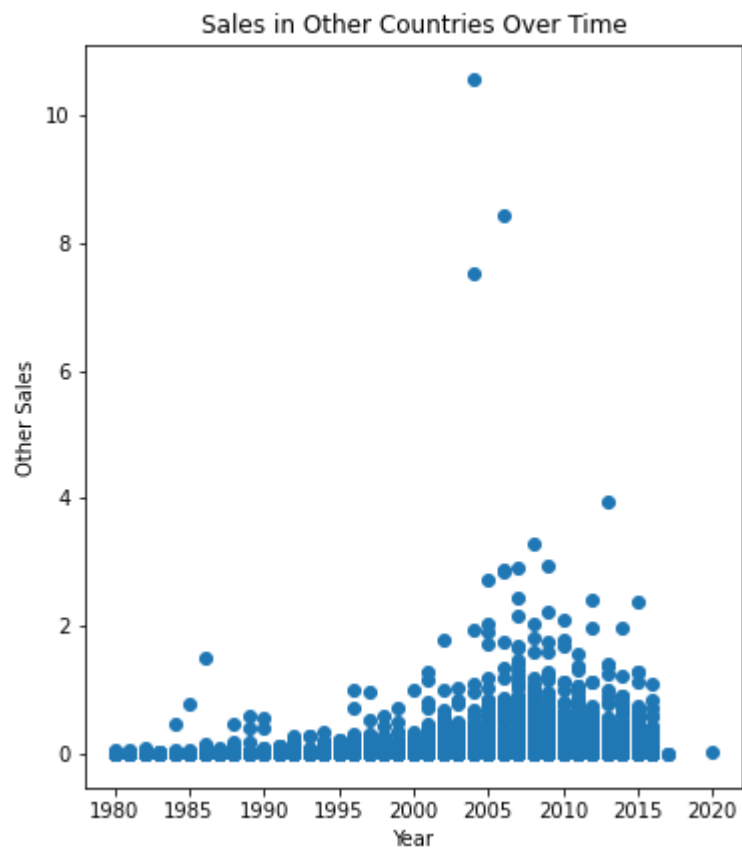
fig=plt.figure(figsize=(20,15))
plt.subplot(2,3,3)
plt.scatter(df['Year_of_Release'],df['EU_Sales'])
plt.title('Sales in Europe Over Time')
plt.ylabel('Europe Sales')
plt.xlabel('Year')
plt.show()

fig=plt.figure(figsize=(20,15))
plt.subplot(2,3,4)
plt.scatter(df['Year_of_Release'],df['Global_Sales'])
plt.title('Sales Globally Over Time')
plt.ylabel('Global Sales')
plt.xlabel('Year')
plt.show()

fig=plt.figure(figsize=(20,15))
plt.subplot(2,3,5)
plt.scatter(df['Year_of_Release'],df['Other_Sales'])
plt.title('Sales in Other Countries Over Time')
plt.ylabel('Other Sales')
plt.xlabel('Year')
plt.show()
```







What are the top games across the world:

```

In [17]: fig=plt.figure(figsize=(18.5,12))
plt.subplots_adjust(left=None, wspace=0.60, hspace=0.35)

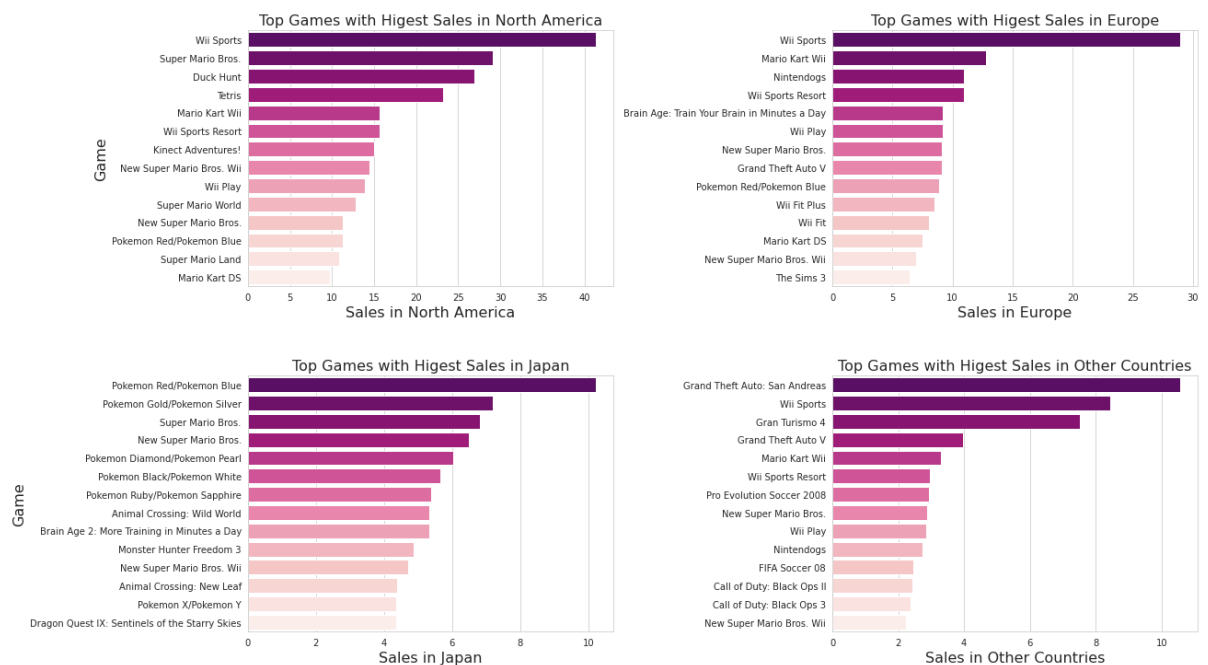
plt.subplot(2, 2, 1)
plt.title('Top Games with Higest Sales in North America',fontdict={'fontsize':
16})
sns.barplot(y='Name',x='NA_Sales',data=df.sort_values(by='NA_Sales',ascending=
False).head(14),palette='RdPu_r')
plt.xlabel('Sales in North America',fontdict={'fontsize':16})
plt.ylabel('Game',fontdict={'fontsize':16});

plt.subplot(2, 2, 2)
plt.title('Top Games with Higest Sales in Europe',fontdict={'fontsize':16})
sns.barplot(y='Name',x='EU_Sales',data=df.sort_values(by='EU_Sales',ascending=
False).head(14),palette='RdPu_r')
plt.xlabel('Sales in Europe',fontdict={'fontsize':16})
plt.ylabel('',fontdict={'fontsize':16});

plt.subplot(2, 2, 3)
plt.title('Top Games with Higest Sales in Japan',fontdict={'fontsize':16})
sns.barplot(y='Name',x='JP_Sales',data=df.sort_values(by='JP_Sales',ascending=
False).head(14),palette='RdPu_r')
plt.xlabel('Sales in Japan',fontdict={'fontsize':16})
plt.ylabel('Game',fontdict={'fontsize':16});

plt.subplot(2, 2, 4)
plt.title('Top Games with Higest Sales in Other Countries',fontdict={'fontsiz
e':16})
sns.barplot(y='Name',x='Other_Sales',data=df.sort_values(by='Other_Sales',asce
nding=False).head(14),palette='RdPu_r')
plt.xlabel('Sales in Other Countries',fontdict={'fontsize':16})
plt.ylabel('',fontdict={'fontsize':16});

```



We could see that Wii-Sports have made the Highest Sales in North America and Europe, whereas Pokemon Red/Pokemon Blue has made the highest sales in Japan and Grand Theft Auto: San Andreas in the rest of the world.

A series of electronic games from Nintendo that debuted in Japan in February 1996 under the name Pokémon Green and Pokémon Red. Pokemon became one of the most successful video game franchises in the world, second only to Super Mario Brothers.

Our analysis revealed Japan to be unique in every aspect of gaming, including the games they play, genres, and even platforms.

Especially when it comes to gaming, Japan has always been a technological leader. Japan's games tend to reflect their respective cultures.

Differences between the 20th Century and the 21st Century Gaming:



```
In [31]: df20=df[df['Year_of_Release']<=2000]
df21=df[df['Year_of_Release']>2000]
```

```
In [32]: index = ['North America', 'Europe', 'Japan', 'Other Country', 'Global']

# Convert the dictionary into DataFrame
# Make Own Index and Removing Default index
salcent=pd.DataFrame({'20th Century':[df20['NA_Sales'].sum(),df20['EU_Sales'].sum(),df20['JP_Sales'].sum(),df20['Other_Sales'].sum(),df20['Global_Sales'].sum()], '21st Century':[df21['NA_Sales'].sum(),df21['EU_Sales'].sum(),df21['JP_Sales'].sum(),df21['Other_Sales'].sum(),df21['Global_Sales'].sum()]},index=index)

salcent
```

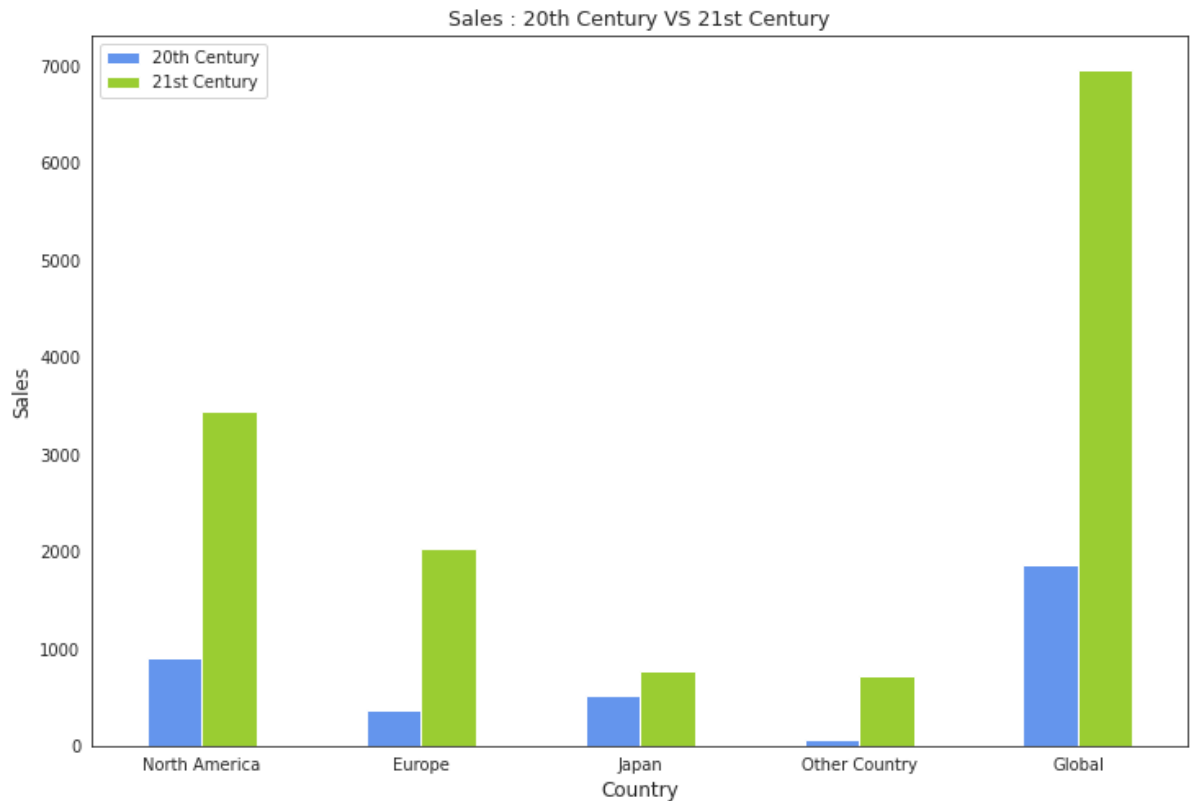
Out[32]:

	20th Century	21st Century
North America	908.09	3435.38
Europe	367.37	2032.84
Japan	517.62	773.11
Other Country	66.22	716.51
Global	1859.47	6961.87

20 vs 21st Century

```
In [33]: sns.set_style("white")

salcent.plot(kind='bar',figsize=(12,8),color=['#6495ED','yellowgreen'])
plt.xticks(rotation=0)
plt.title('Sales : 20th Century VS 21st Century',fontdict={'fontsize':12.5})
plt.xlabel('Country',fontdict={'fontsize':12})
plt.ylabel('Sales',fontdict={'fontsize':12});
plt.show()
```



Compared to the 20th century, the sales in Japan have increased only a little in the 21st century. Overall, global sales have increased drastically in the 21st century. The Japanese Gaming Industry only began expanding in the late 20th century.

Genre: 21st vs 20

```

In [35]: fig=plt.figure(figsize=(22,16.5))
plt.subplots_adjust(left=None, wspace=0.20, hspace=0.45)
sns.set_style("white")

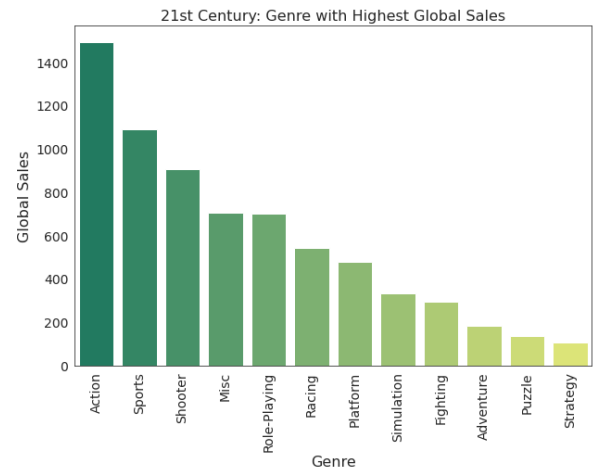
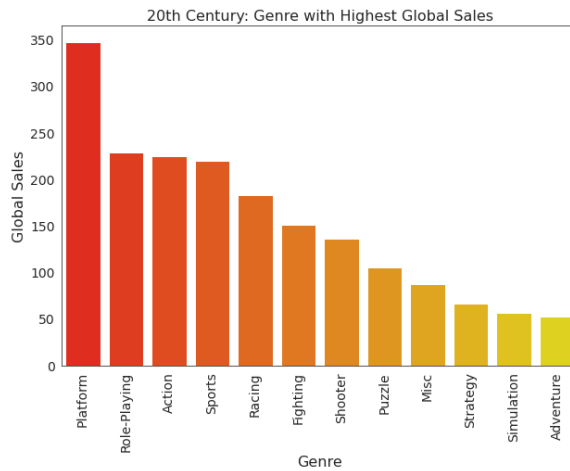
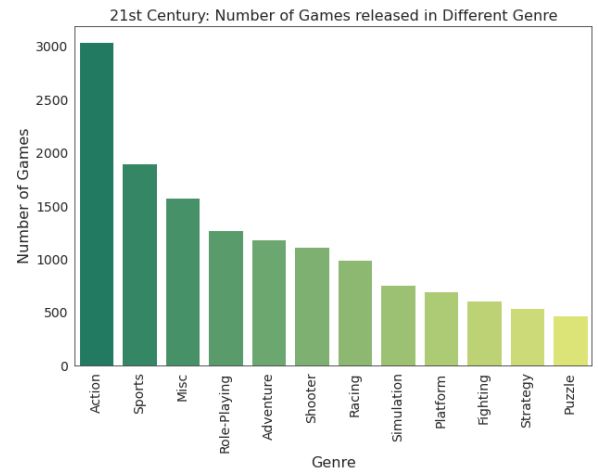
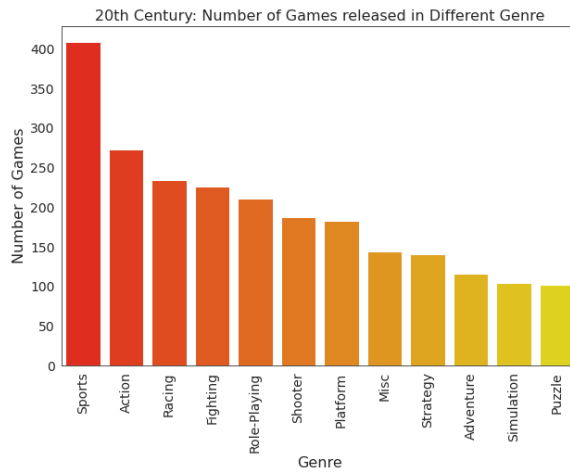
plt.subplot(2, 2, 1)
sns.set_style("white")
plt.xticks(fontsize=14)
plt.yticks(fontsize=14)
plt.title('20th Century: Number of Games released in Different Genre',fontdict
={'fontsize':16})
plt.xticks(rotation=90)
sns.barplot(x=df20.Genre.value_counts().index, y=df20.Genre.value_counts(),pal
ette='autumn');
plt.xlabel('Genre',fontdict={'fontsize':16})
plt.ylabel('Number of Games',fontdict={'fontsize':16});

plt.subplot(2, 2, 2)
plt.title('21st Century: Number of Games released in Different Genre',fontdict
={'fontsize':16})
sns.set_style("white")
plt.xticks(fontsize=14)
plt.yticks(fontsize=14)
plt.xticks(rotation=90)
sns.barplot(x=df21.Genre.value_counts().index, y=df21.Genre.value_counts(),pal
ette='summer');
plt.xlabel('Genre',fontdict={'fontsize':16})
plt.ylabel('Number of Games',fontdict={'fontsize':16});

plt.subplot(2, 2, 3)
sns.set_style("white")
plt.xticks(fontsize=14)
plt.yticks(fontsize=14)
plt.title('20th Century: Genre with Highest Global Sales',fontdict={'fontsize'
:16})
plt.xticks(rotation=90)
sns.barplot(x='Genre', y='Global_Sales',data=df20.groupby('Genre').sum().Globa
l_Sales.sort_values(ascending=False).reset_index(),palette='autumn');
plt.xlabel('Genre',fontdict={'fontsize':16})
plt.ylabel('Global Sales',fontdict={'fontsize':16});

plt.subplot(2, 2, 4)
plt.title('21st Century: Genre with Highest Global Sales',fontdict={'fontsize'
:16})
sns.set_style("white")
plt.xticks(fontsize=14)
plt.yticks(fontsize=14)
plt.xticks(rotation=90)
sns.barplot(x='Genre', y='Global_Sales',data=df21.groupby('Genre').sum().Globa
l_Sales.sort_values(ascending=False).reset_index(),palette='summer');
plt.xlabel('Genre',fontdict={'fontsize':16})
plt.ylabel('Global Sales',fontdict={'fontsize':16});

```



In the 20th Century, the Sports Genre released the most games, while the Action Genre released the most games in the 21st Century. Games of the Platform Genre were the most successful in the 20th Century, while games of the Action Genre were the most successful in the 21st Century.

Conclusion:

Summary of Our Findings

Most important variables:

Platform

Genre

Publisher

North American Sales

European Union Sales

Japan Sales

Global Sales

Other sales

Summary of Our Findings:

Gaming platform and game sales are primarily regional specific. For example, platforms that are produced by American or Japanese companies have higher sales in their respective locations. This is likely due to their platforms being more accessible to the consumer in that location, national pride, and the language the game uses. Genre popularity is also regional specific, with shooter games being more popular in western regions like North America and the EU while extremely unpopular in Japan. This is likely due to cultural differences. Our findings suggest that when a game is being developed, the platform and genre are the most influential factors that will impact global and geographic-specific sales.

Business Implications

Limitations of this Project

Dataset Limitations:

We will not include data from after 2016 because there is not enough information in the dataset for game sales between 2016 to 2020.

We removed all rows with missing values rather than replacing them. Even though it does create a smaller pool of data to work with, it is more accurate than replacing the data. We made this choice to avoid data manipulation.

Significant amounts of missing data in: User ratings, user count, critic scores, critic counts, developer, and ratings.

Not all games are released in all sales regions or in multiple languages.

Sales data do not necessarily represent preferences of consumers, but what is available in the market.

Games are released for multiple platforms, but this data does not always accurately represent this.

Potential for Future Projects

Predictive

What is likely to happen next?

This data analysis can be used to help inform algorithms and AI to help make better automated decisions on what types of games to produce. The observed trends can be used by game developers and investors to better understand where the market is likely to go. For example, it can inform decisions on what kind of genres are likely to be popular in certain regions, which will determine where business resources are allocated.

For example, the data shows that role-playing games are popular and shooter games aren't popular in Japan. The data shows this as a trend over several years, indicating that this is likely to continue.

Prescriptive

What will we do next?

This data can help to define future actions and projects. Game publishers can outline plans based on the analysis of the historical records of recent game sales to help ensure success. Based on the expected trends in the predictive analysis, game publishers can try to maximize revenues by making more informed business decisions.

To continue the example above, the trend in Japanese video game sales indicates to game developers that they should have translated versions of role-playing games in Japanese if they are originally made in another language. It also indicates that revenue can be maximized by releasing role-playing games in Japan, and costs can be minimized by not spending budgets to release shooter games in Japan.

References

Dataset: The data is taken from Kaggle datasets, a web service platform: Video Game Sales with Ratings | <https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings> (<https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings>)

References for Code: A Gamer's Analysis by Sai Prasath | <https://www.kaggle.com/saiprasath95/a-gamer-s-analysis/notebook> (<https://www.kaggle.com/saiprasath95/a-gamer-s-analysis/notebook>)

References for Code: Video Game Sales Analysis by LEHAK NARNAULI | <https://datascience.fm/video-game-sales-analysis/> (<https://datascience.fm/video-game-sales-analysis/>)

References for Analysis & Hypothesis: https://www.causeweb.org/usproc/sites/default/files/usclap/2017-2/Factors_that_Impact_Video_Game_Sales.pdf (https://www.causeweb.org/usproc/sites/default/files/usclap/2017-2/Factors_that_Impact_Video_Game_Sales.pdf)