

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
from sklearn.impute import SimpleImputer
import plotly.express as px
```

```
df=pd.read_csv("/content/vgchartz-2024.csv")
```

```
df.shape
```

```
(64016, 14)
```

```
df.head()
```

	img	title	console	genre	publisher	developer	critic_score	total_sales	na_sales	jp_sales
0	/games/boxart/full_6510540AmericaFrontccc.jpg	Grand Theft Auto V	PS3	Action	Rockstar Games	Rockstar North	9.4	20.32	6.37	0.99
1	/games/boxart/full_5563178AmericaFrontccc.jpg	Grand Theft Auto V	PS4	Action	Rockstar Games	Rockstar North	9.7	19.39	6.06	0.60
2	/games/boxart/827563ccc.jpg	Grand Theft Auto: Vice City	PS2	Action	Rockstar Games	Rockstar North	9.6	16.15	8.41	0.47
3	/games/boxart/full_9218923AmericaFrontccc.jpg	Grand Theft Auto V	X360	Action	Rockstar Games	Rockstar North	NaN	15.86	9.06	0.06
4	/games/boxart/full_4990510AmericaFrontccc.jpg	Call of Duty: Black Ops 3	PS4	Shooter	Activision	Treyarch	8.1	15.09	6.18	0.41

Next steps:

Generate code with df

 View recommended plots

```
df.drop(columns=['img'],inplace=True)
```

DATA CLEANING

```
df.columns
```

```
Index(['title', 'console', 'genre', 'publisher', 'developer', 'critic_score',
      'total_sales', 'na_sales', 'jp_sales', 'pal_sales', 'other_sales',
      'release_date', 'last_update'],
      dtype='object')
```

```
df.isnull().sum()
```

```
title      0
console    0
genre      0
publisher  0
developer  17
critic_score  57338
total_sales  45094
na_sales    51379
jp_sales    57290
pal_sales   51192
other_sales 48888
release_date  7051
last_update 46137
dtype: int64
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64016 entries, 0 to 64015
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  -
 0   title           64016 non-null  object
 1   console         64016 non-null  object
 2   genre           64016 non-null  object
 3   publisher       64016 non-null  object
 4   developer       63999 non-null  object
 5   critic_score    6678 non-null   float64
 6   total_sales     18922 non-null  float64
 7   na_sales        12637 non-null  float64
 8   jp_sales        6726 non-null   float64
 9   pal_sales       12824 non-null  float64
10  other_sales     15128 non-null  float64
11  release_date    56965 non-null  object
12  last_update     17879 non-null  object
dtypes: float64(6), object(7)
memory usage: 6.3+ MB
```

## HANDLING NULL VALUES

```
num_cols=df.select_dtypes(include=np.number).columns.to_list()
```

```
num_cols
```

```
['critic_score',
 'total_sales',
 'na_sales',
 'jp_sales',
 'pal_sales',
 'other_sales']
```

```
#Imputation
```

```
imputer=SimpleImputer(strategy='median')
df[num_cols]=imputer.fit_transform(df[num_cols])
```

```
df.isnull().sum()
```

```
title           0
console         0
genre           0
publisher       0
developer       17
critic_score    0
total_sales     0
na_sales        0
jp_sales        0
pal_sales       0
other_sales     0
release_date    7051
last_update     46137
dtype: int64
```

```
#Converting the datatype of date column
```

```
df['release_date']=pd.to_datetime(df['release_date'])
df['last_update']=pd.to_datetime(df['last_update'])
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64016 entries, 0 to 64015
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  -
 0   title           64016 non-null  object
 1   console         64016 non-null  object
 2   genre           64016 non-null  object
 3   publisher       64016 non-null  object
 4   developer       63999 non-null  object
```

```

5  critic_score    64016 non-null    float64
6  total_sales    64016 non-null    float64
7  na_sales       64016 non-null    float64
8  jp_sales       64016 non-null    float64
9  pal_sales      64016 non-null    float64
10 other_sales    64016 non-null    float64
11 release_date   56965 non-null    datetime64[ns]
12 last_update    17879 non-null    datetime64[ns]
dtypes: datetime64[ns](2), float64(6), object(5)
memory usage: 6.3+ MB

```

```

df['release_date'].fillna(df['release_date'].median(),inplace=True)
df['last_update'].fillna(df['last_update'].median(),inplace=True)

```

```
df.isnull().sum()
```

```

⇒ title           0
  console         0
  genre           0
  publisher       0
  developer      17
  critic_score    0
  total_sales     0
  na_sales        0
  jp_sales        0
  pal_sales       0
  other_sales     0
  release_date    0
  last_update     0
dtype: int64

```

## HANDLING MISSING VALUES FROM OBJECT COLUMN

```
df.dropna(inplace=True)
```

```
df.isnull().sum()
```

```

⇒ title           0
  console         0
  genre           0
  publisher       0
  developer       0
  critic_score    0
  total_sales     0
  na_sales        0
  jp_sales        0
  pal_sales       0
  other_sales     0
  release_date    0
  last_update     0
dtype: int64

```

```
df.head()
```

```

⇒
   title  console  genre  publisher  developer  critic_score  total_sales  na_sales  jp
0  Grand Theft Auto V      PS3    Action    Rockstar Games    Rockstar North         9.4         20.32         6.37
1  Grand Theft Auto V      PS4    Action    Rockstar Games    Rockstar North         9.7         19.39         6.06

```

Next steps:

[Generate code with df](#)
[View recommended plots](#)

## EDA (EXPLORATORY DATA ANALYSIS)

1. Which titles sold the most worldwide?

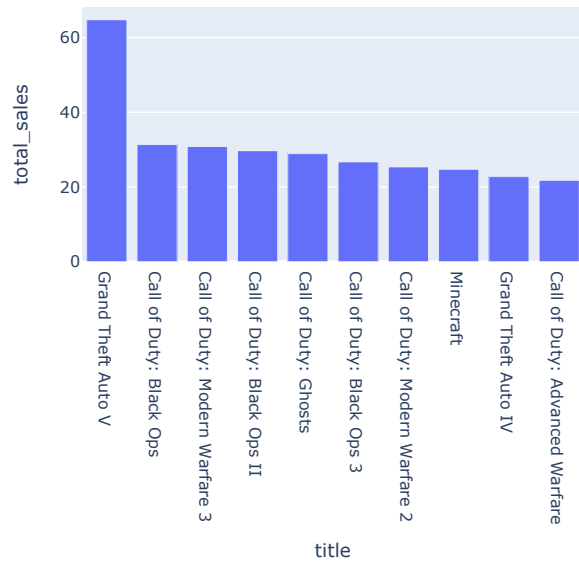
```
sales_by_title=df.groupby('title')['total_sales'].sum().reset_index()
```

```
sales_by_title_sort=sales_by_title.sort_values(by='total_sales',ascending=False)
```

```
px.bar(sales_by_title_sort.head(10),x='title',y='total_sales',title='Top 10 Titles Sold Worldwide')
```



Top 10 Titles Sold Worldwide



CONCLUSION 1: The titles 'Grand Theft Auto V' and 'Call of Duty: Modern Warfare' had the highest sales worldwide.

2. Which year has the highest sales? Has the industry grown overtime?

```
#new column for release year that is extracted from the release_date
```

```
df['release_year']=pd.to_datetime(df['release_date']).dt.year
```

```
df.head()
```



	title	console	genre	publisher	developer	critic_score	total_sales	na_sales	jp
0	Grand Theft Auto V	PS3	Action	Rockstar Games	Rockstar North	9.4	20.32	6.37	
1	Grand Theft Auto V	PS4	Action	Rockstar Games	Rockstar North	9.7	19.39	6.06	
2	Grand Theft Auto: Vice	PS2	Action	Rockstar Games	Rockstar North	9.6	16.15	8.41	

Next steps:

[Generate code with df](#)

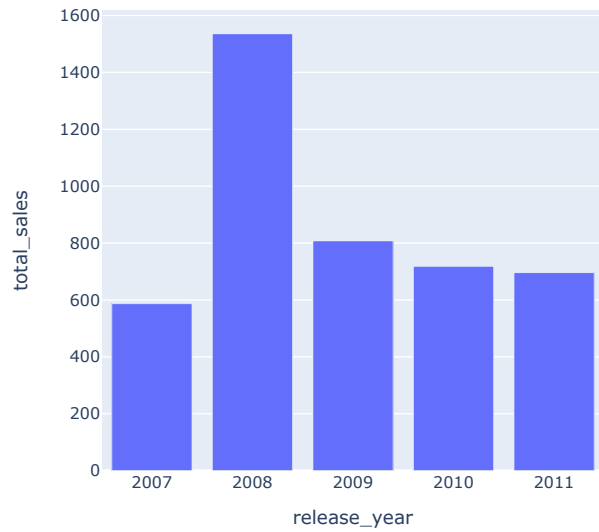
☒ [View recommended plots](#)

```
sales_by_year=df.groupby('release_year')['total_sales'].sum().reset_index()
sales_by_year_sort=sales_by_year.sort_values(by='total_sales',ascending=False)
```

```
px.bar(sales_by_year_sort.head(),x='release_year',y='total_sales',title='Top 10 year by Worldwide Sales')
```



Top 10 year by Worldwide Sales

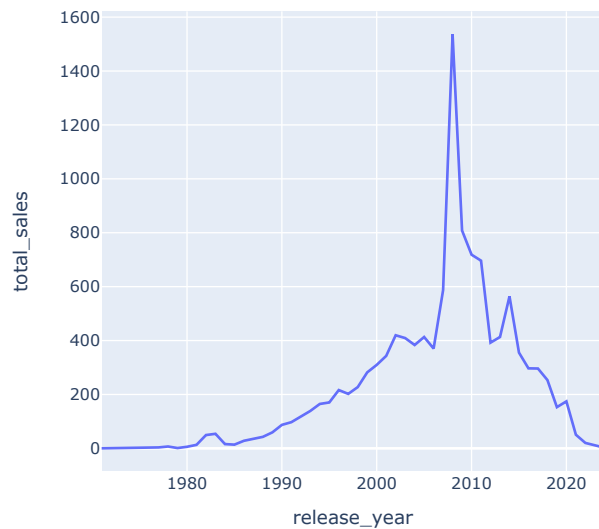


CONCLUSION 2.1: The year 2008 made the highest sales followed by 2009

```
px.line(sales_by_year,x='release_year',y='total_sales',title='Global Sales over year')
```



Global Sales over year



CONCLUSION 2.2 : The Industry grown around year 2008 but currently the progress is constant.

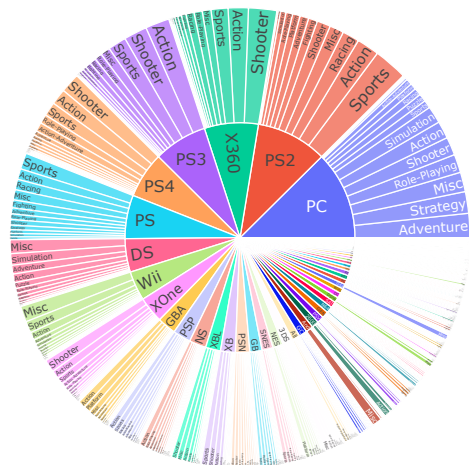
3. Do any consoles seem to specialize in a particular genre?

```
console_genre_sales=df.groupby(['console','genre'])['total_sales'].sum().reset_index()
console_genre_sales_sorted=console_genre_sales.sort_values(by='total_sales',ascending=False)
```

```
px.sunburst(console_genre_sales_sorted,path=['console','genre'],values='total_sales',title='Console Specialisation in Genre Sales')
```



Console Specialisation in Genre Sales



CONCLUSION 3: The PC Console do specialise in genre 'adventure' followed by genre 'startegy' and 'Misc'.

4. What titles are popular in one region but flop in another?

- na- North America
- -jp- Japan
- -PAL- Phase Alternating Line (includes the region like Europe, Australia, New Zealand and some other countries)

```
df['na_ratio']=df['na_sales']/df['total_sales']
df['jp_ratio']=df['jp_sales']/df['total_sales']
df['pal_ratio']=df['pal_sales']/df['total_sales']
```

```
df.head(3)
```




console	genre	publisher	developer	critic_score	total_sales	na_sales	jp_sales	pal
PS3	Action	Rockstar Games	Rockstar North	9.4	20.32	6.37	0.99	
PS4	Action	Rockstar Games	Rockstar North	9.7	19.39	6.06	0.60	
PS2	Action	Rockstar Games	Rockstar North	9.6	16.15	8.41	0.47	

4.1 TITLES THAT ARE POPULAR IN NA BUT FLOP IN JP AND PAL REGIONS.


```
na_popular=df[(df['na_ratio']>0.8) & (df['jp_ratio']<0.2) & (df['pal_ratio']<0.2)]
```

```
na_popular
```



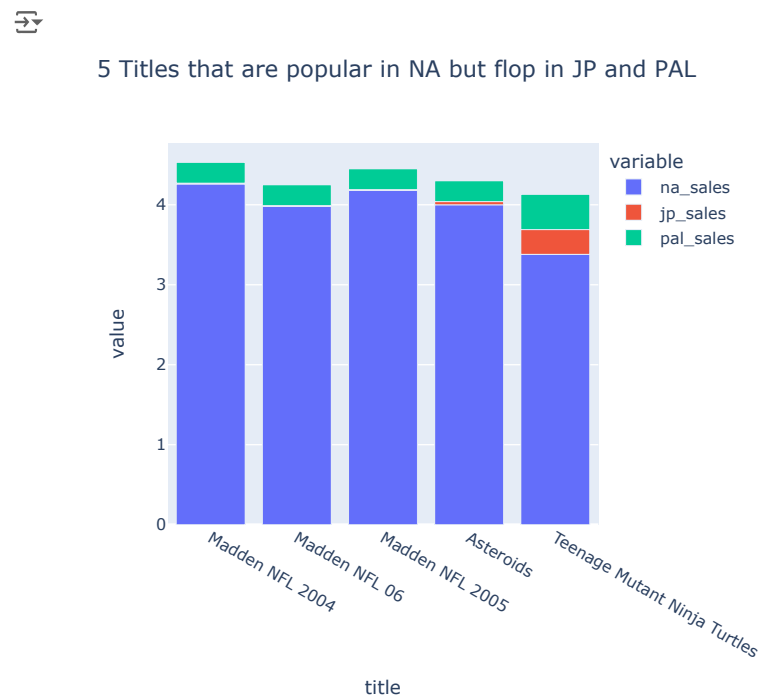
	title	console	genre	publisher	developer	critic_score	total_sales
75	Madden NFL 2004	PS2	Sports	EA Sports	EA Tiburon	9.5	5.23
94	Madden NFL 06	PS2	Sports	EA Sports	EA Tiburon	9.1	4.91
114	Madden NFL 2005	PS2	Sports	EA Sports	EA Tiburon	9.5	4.53
125	Asteroids	2600	Shooter	Atari	Atari	7.5	4.31
135	Teenage Mutant Ninja Turtles	NES	Platform	Ultra Games	Konami	5.9	4.17
...	...	...	...	...	...	...	...
11051	New International Track & Field	DS	Sports	Konami	Sumo Digital	7.4	0.08
11154	Baroque	Wii	Role-Playing	Atlus	Sting	5.2	0.08
11242	Elebits: The Adventures of Kai and Zero	DS	Adventure	Konami	Konami	7.5	0.08
12114	Vampire Rain: Altered Species	PS3	Action	Ignition Entertainment	Artoon	3.5	0.06
17325	G1 Jockey 4 2008	PS3	Sports	KOEI	Koei/Inis	7.5	0.01

1187 rows × 17 columns



Next steps: [Generate code with na\\_popular](#) [View recommended plots](#)

```
px.bar(na_popular.head(5),x='title',y=['na_sales','jp_sales','pal_sales'],title='5 Titles that are popular in NA but flop in JP and PAL')
```



CONCLUSIONS 4.1 : The titles 'Madden NFL 2004', 'Madden NFL 06' and 'Madden NFL 2005' are popular in NA but flop in JP and PAL.

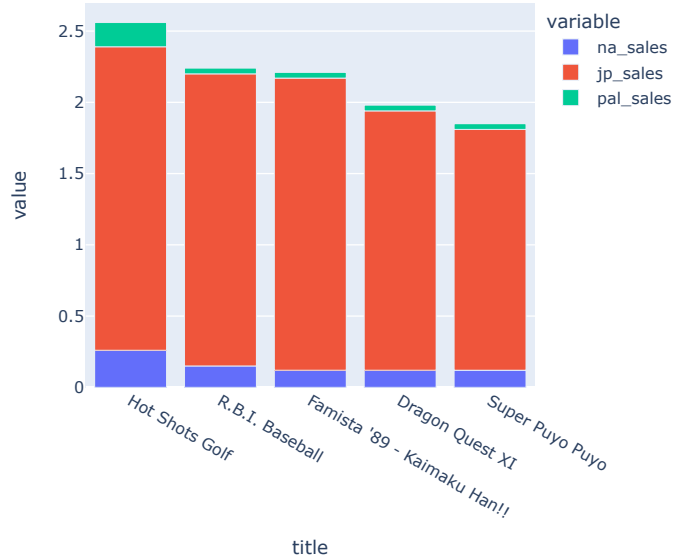
## ✓ 4.2 TITLES THAT ARE POPULAR IN JP BUT FLOP IN NA AND PAL *REGIONS*.

```
jp_popular=df[(df['jp_ratio']>0.8) & (df['na_ratio']<0.2) & (df['pal_ratio']<0.2)]
```

```
px.bar(jp_popular.head(5),x='title',y=['na_sales','jp_sales','pal_sales'],title='5 Titles that are popular in JP but flop in NA and PAL')
```



5 Titles that are popular in JP but flop in NA and PAL



CONCLUSIONS 4.2 : The titles 'Hot Shots Golf', 'R.B.I. Baseball' and 'Famista '89 - Kaimaku Han!!' are popular in JP but flop in NA and PAL.

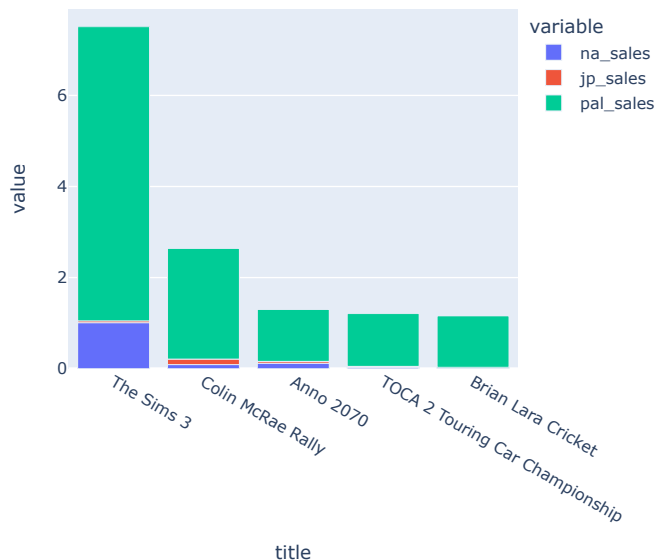
## ✓ 4.3 TITLES THAT ARE POPULAR IN PAL BUT FLOP IN JP AND NA *REGIONS*.

```
pal_popular=df[(df['pal_ratio']>0.8) & (df['jp_ratio']<0.2) & (df['na_ratio']<0.2)]
```

```
px.bar(pal_popular.head(5),x='title',y=['na_sales','jp_sales','pal_sales'],title='5 Titles that are popular in PAL but flop in JP and NA')
```



5 Titles that are popular in PAL but flop in JP and NA





CONCLUSIONS 4.3 : The titles 'The Sims 3', 'Colin McRae Rally' and 'Anno 2070' are popular in PAL but flop in JP and NP.