ALLIANCE UNIVERSITY

Private University established in Karnataka State by Act No.34 of year 2010
Recognized by the University Grants Commission (UGC), New Delhi

Project Report

Introduction To Data Science

Semester — II

Game Sales Insights

By

Aarush C S

Reg No. 2411021240028

Department of Computer Application

Alliance University

Chandapura — Anekal Main Road, Anekal

Bengaluru — 562 106

April 2025

Introduction

This project uses the **Video Game Sales dataset** (vgsales.csv), which provides comprehensive data on video game performance across various regions. Each entry represents a video game, and the dataset includes features like:
Name of the game
Platform it was released on
Year of release
Genre and **Publisher**
Sales data: NA_Sales, EU_Sales, JP_Sales, Other_Sales, and Global_Sales

Conclusion with Key Insights
Through this analysis, we've uncovered several valuable insights from the video game sales data:
Key Insights:
Genre and Platform Impact: Certain genres (like Action and Sports) and platforms (like PS2 and X360) dominate in terms of sales, especially in North America.
Regional Preferences Vary: Japan has distinct gaming preferences (e.g., high popularity of role-playing games) compared to North America and Europe.
Global Sales Distribution: A small number of titles account for the majority of global video game sales — the market is **heavily top-loaded**.
Yearly Trends: Sales peaked during specific years, especially in the late 2000s, indicating the golden age of console gaming.
Model Performance: The Logistic Regression model helped classify games based on their success level with reasonable accuracy, although model tuning and feature engineering could improve performance.
Business Implication:
Publishers and developers can use these insights to:

- Strategize release platforms and timelines
- Tailor games to regional audiences
- Forecast performance of upcoming titles based on genre/platform alignment

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, classification_report,
accuracy_score, precision_score, recall_score, f1_score
df = pd.read_csv(r"C:\Users\91628\Downloads\vgsales.csv\vgsales.csv")
df
```

	Rank	Name	Platform	\
0	1	Wii Sports	Wii	
1	2	Super Mario Bros.	NES	
2	3	Mario Kart Wii	Wii	
3	4	Wii Sports Resort	Wii	
4	5	Pokemon Red/Pokemon Blue	GB	
...	...	...	...	

```
16593  16596              Woody Woodpecker in Crazy Castle 5      GBA
16594  16597                   Men in Black II: Alien Escape       GC
16595  16598  SCORE International Baja 1000: The Official Game
PS2
16596  16599                                 Know How 2       DS
16597  16600                              Spirits & Spells      GBA

       Year         Genre   Publisher  NA_Sales  EU_Sales  JP_Sales  \
0     2006.0        Sports    Nintendo     41.49     29.02      3.77
1     1985.0      Platform    Nintendo     29.08      3.58      6.81
2     2008.0        Racing    Nintendo     15.85     12.88      3.79
3     2009.0        Sports    Nintendo     15.75     11.01      3.28
4     1996.0  Role-Playing    Nintendo     11.27      8.89     10.22
...      ...           ...         ...       ...       ...       ...
16593  2002.0      Platform       Kemco      0.01      0.00      0.00
16594  2003.0       Shooter  Infogrames      0.01      0.00      0.00
16595  2008.0        Racing  Activision      0.00      0.00      0.00
16596  2010.0        Puzzle    7G//AMES      0.00      0.01      0.00
16597  2003.0      Platform     Wanadoo      0.01      0.00      0.00

      Other_Sales  Global_Sales
0            8.46         82.74
1            0.77         40.24
2            3.31         35.82
3            2.96         33.00
4            1.00         31.37
...           ...           ...
16593        0.00          0.01
16594        0.00          0.01
16595        0.00          0.01
16596        0.00          0.01
16597        0.00          0.01

[16598 rows x 11 columns]
df.head(20)
    Rank                                  Name Platform     Year  \
0      1                            Wii Sports      Wii  2006.0
1      2                     Super Mario Bros.      NES  1985.0
2      3                       Mario Kart Wii      Wii  2008.0
3      4                    Wii Sports Resort      Wii  2009.0
4      5              Pokemon Red/Pokemon Blue       GB  1996.0
5      6                               Tetris       GB  1989.0
6      7                New Super Mario Bros.       DS  2006.0
7      8                             Wii Play      Wii  2006.0
8      9            New Super Mario Bros. Wii      Wii  2009.0
9     10                          Duck Hunt      NES  1984.0
10    11                          Nintendogs       DS  2005.0
11    12                       Mario Kart DS       DS  2005.0
12    13           Pokemon Gold/Pokemon Silver       GB  1999.0
13    14                             Wii Fit      Wii  2007.0
14    15                         Wii Fit Plus      Wii  2009.0
15    16                 Kinect Adventures!     X360  2010.0
16    17                 Grand Theft Auto V      PS3  2013.0
17    18          Grand Theft Auto: San Andreas      PS2  2004.0
18    19                   Super Mario World     SNES  1990.0
19    20  Brain Age: Train Your Brain in Minutes a Day       DS  2005.0

         Genre            Publisher  NA_Sales  EU_Sales  JP_Sales  \
0       Sports             Nintendo     41.49     29.02      3.77
```

1	Platform	Nintendo	29.08	3.58	6.81
2	Racing	Nintendo	15.85	12.88	3.79
3	Sports	Nintendo	15.75	11.01	3.28
4	Role-Playing	Nintendo	11.27	8.89	10.22
5	Puzzle	Nintendo	23.20	2.26	4.22
6	Platform	Nintendo	11.38	9.23	6.50
7	Misc	Nintendo	14.03	9.20	2.93
8	Platform	Nintendo	14.59	7.06	4.70
9	Shooter	Nintendo	26.93	0.63	0.28
10	Simulation	Nintendo	9.07	11.00	1.93
11	Racing	Nintendo	9.81	7.57	4.13
12	Role-Playing	Nintendo	9.00	6.18	7.20
13	Sports	Nintendo	8.94	8.03	3.60
14	Sports	Nintendo	9.09	8.59	2.53
15	Misc	Microsoft Game Studios	14.97	4.94	0.24
16	Action	Take-Two Interactive	7.01	9.27	0.97
17	Action	Take-Two Interactive	9.43	0.40	0.41
18	Platform	Nintendo	12.78	3.75	3.54
19	Misc	Nintendo	4.75	9.26	4.16

	Other_Sales	Global_Sales
0	8.46	82.74
1	0.77	40.24
2	3.31	35.82
3	2.96	33.00
4	1.00	31.37
5	0.58	30.26
6	2.90	30.01
7	2.85	29.02
8	2.26	28.62
9	0.47	28.31
10	2.75	24.76
11	1.92	23.42
12	0.71	23.10
13	2.15	22.72
14	1.79	22.00
15	1.67	21.82
16	4.14	21.40
17	10.57	20.81
18	0.55	20.61
19	2.05	20.22

df.tail()

	Rank	Name	Platform
16593	16596	Woody Woodpecker in Crazy Castle 5	GBA
16594	16597	Men in Black II: Alien Escape	GC
16595	16598	SCORE International Baja 1000: The Official Game	PS2
16596	16599	Know How 2	DS
16597	16600	Spirits & Spells	GBA

	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales
16593	2002.0	Platform	Kemco	0.01	0.00	0.0
16594	2003.0	Shooter	Infogrames	0.01	0.00	0.0
16595	2008.0	Racing	Activision	0.00	0.00	0.0
16596	2010.0	Puzzle	7G//AMES	0.00	0.01	0.0
16597	2003.0	Platform	Wanadoo	0.01	0.00	0.0

	Other_Sales	Global_Sales
16593	0.0	0.01

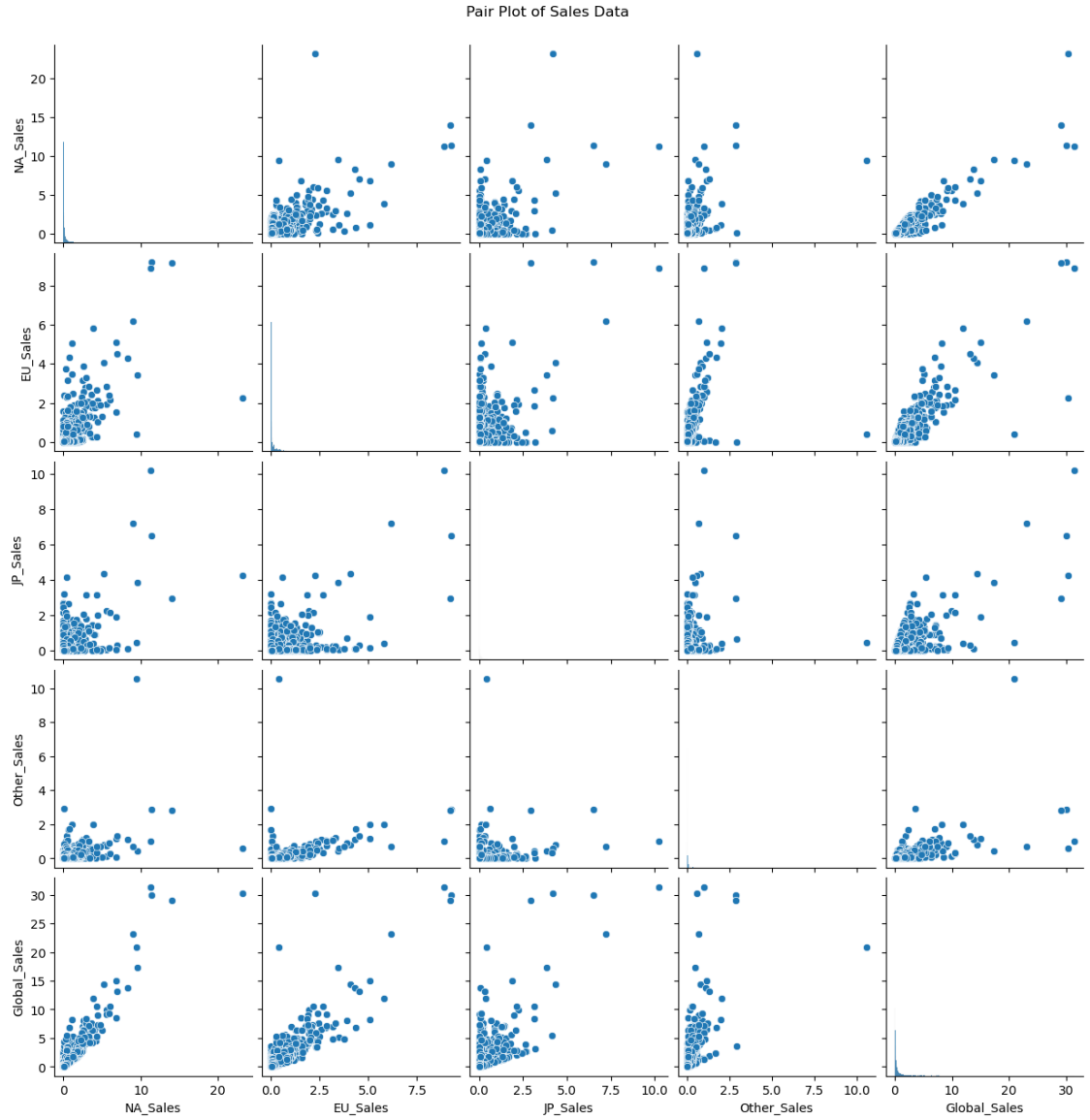
```
16594          0.0          0.01
16595          0.0          0.01
16596          0.0          0.01
16597          0.0          0.01
df.describe()
              Rank          Year      NA_Sales      EU_Sales      JP_Sales  \
count  16598.000000  16327.000000  16598.000000  16598.000000
16598.000000
mean    8300.605254   2006.406443      0.264667      0.146652
0.077782
std     4791.853933      5.828981      0.816683      0.505351      0.309291
min        1.000000   1980.000000      0.000000      0.000000      0.000000
25%     4151.250000   2003.000000      0.000000      0.000000
0.000000
50%     8300.500000   2007.000000      0.080000      0.020000
0.000000
75%    12449.750000   2010.000000      0.240000      0.110000
0.040000
max    16600.000000   2020.000000     41.490000     29.020000
10.220000

        Other_Sales  Global_Sales
count  16598.000000  16598.000000
mean       0.048063      0.537441
std        0.188588      1.555028
min        0.000000      0.010000
25%        0.000000      0.060000
50%        0.010000      0.170000
75%        0.040000      0.470000
max       10.570000     82.740000
df.isnull().sum()
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
dtype: int64
df.columns
Index(['Rank', 'Name', 'Platform', 'Year', 'Genre', 'Publisher', 'NA_Sales',
       'EU_Sales', 'JP_Sales', 'Other_Sales', 'Global_Sales'],
      dtype='object')
sales= ['NA_Sales', 'EU_Sales', 'JP_Sales', 'Other_Sales',
'Global_Sales']
sdf = df[sales].sample(5001, random_state=1)
sdf
       NA_Sales  EU_Sales  JP_Sales  Other_Sales  Global_Sales
4234       0.26      0.17      0.00         0.03          0.46
7119       0.21      0.00      0.00         0.02          0.23
106        3.68      1.75      1.42         0.28          7.13
5242       0.06      0.25      0.00         0.05          0.36
13547      0.03      0.01      0.00         0.00          0.04
...         ...       ...       ...          ...           ...
968        1.26      0.39      0.08         0.06          1.79
```

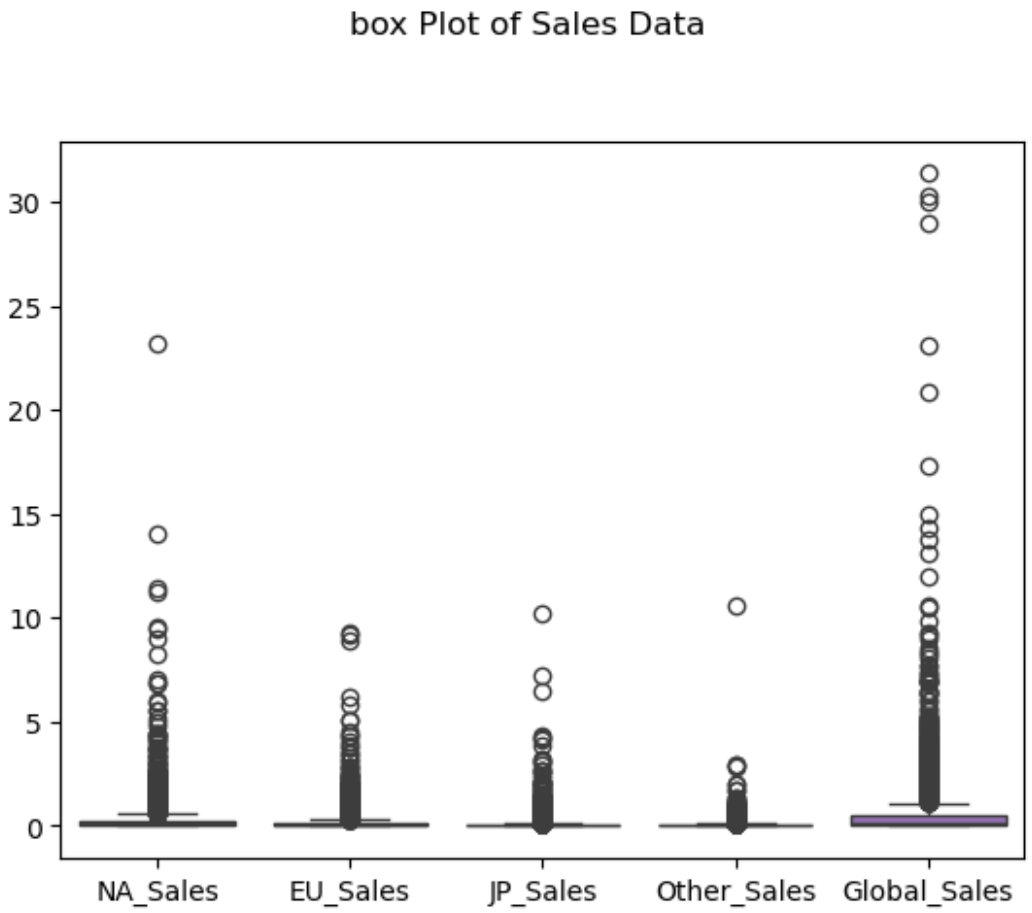
5375	0.25	0.09	0.00	0.01	0.34
12064	0.05	0.01	0.00	0.01	0.07
10124	0.06	0.04	0.00	0.01	0.11
229	2.29	1.97	0.13	0.24	4.63

[5001 rows x 5 columns]

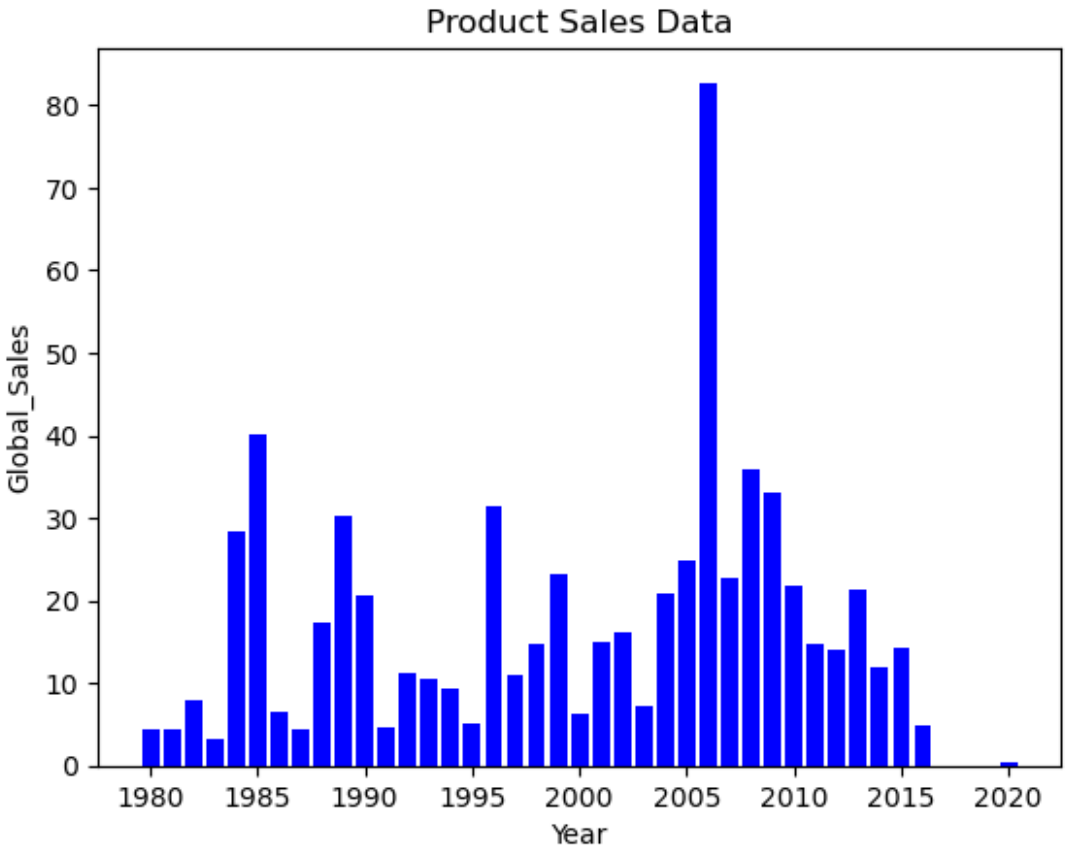
```
sns.pairplot(sdf)
plt.suptitle("Pair Plot of Sales Data", y=1.02)
plt.show()
```



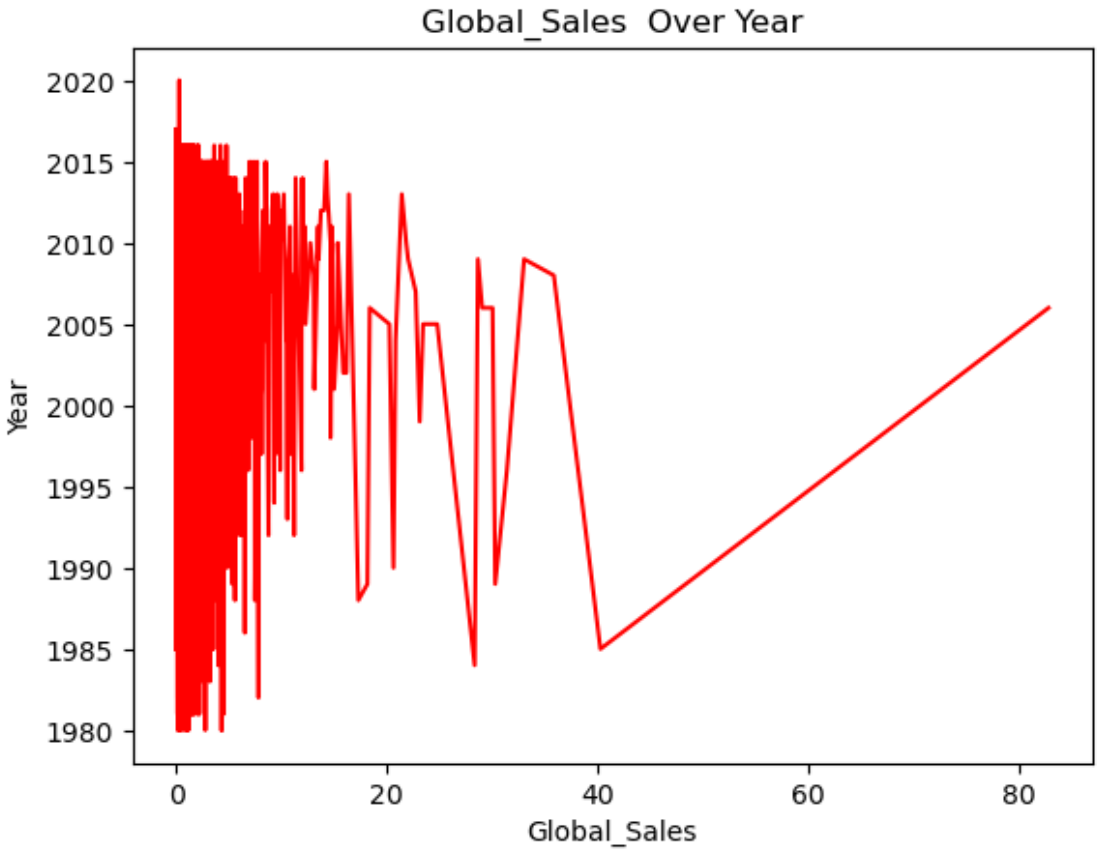

```
sns.boxplot(sdf)
plt.suptitle("box Plot of Sales Data", y=1.02)
plt.show()
```



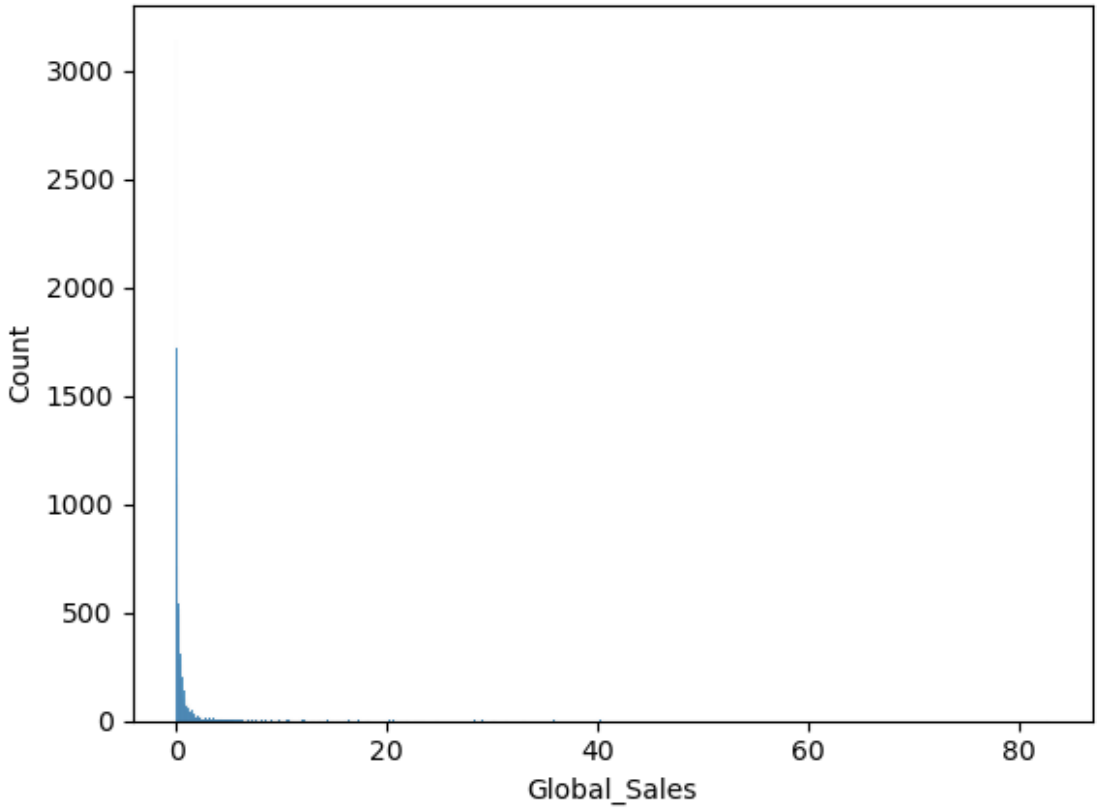

```
# Create bar chart
plt.bar(df['Year'],df['Global_Sales'], color='blue')
# Labels and title
plt.xlabel('Year')
plt.ylabel('Global_Sales')
plt.title('Product Sales Data')
# Show plot
plt.show()
```



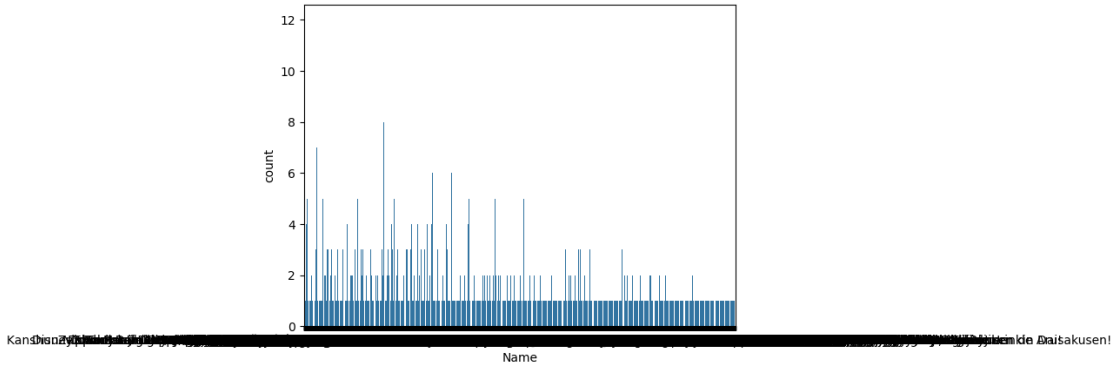

```
# Create line plot
plt.plot(df['Global_Sales'],df['Year'], color='red')
# Labels and title
plt.xlabel('Global_Sales')
plt.ylabel('Year')
plt.title('Global_Sales  Over Year')
# Show plot
plt.show()
```



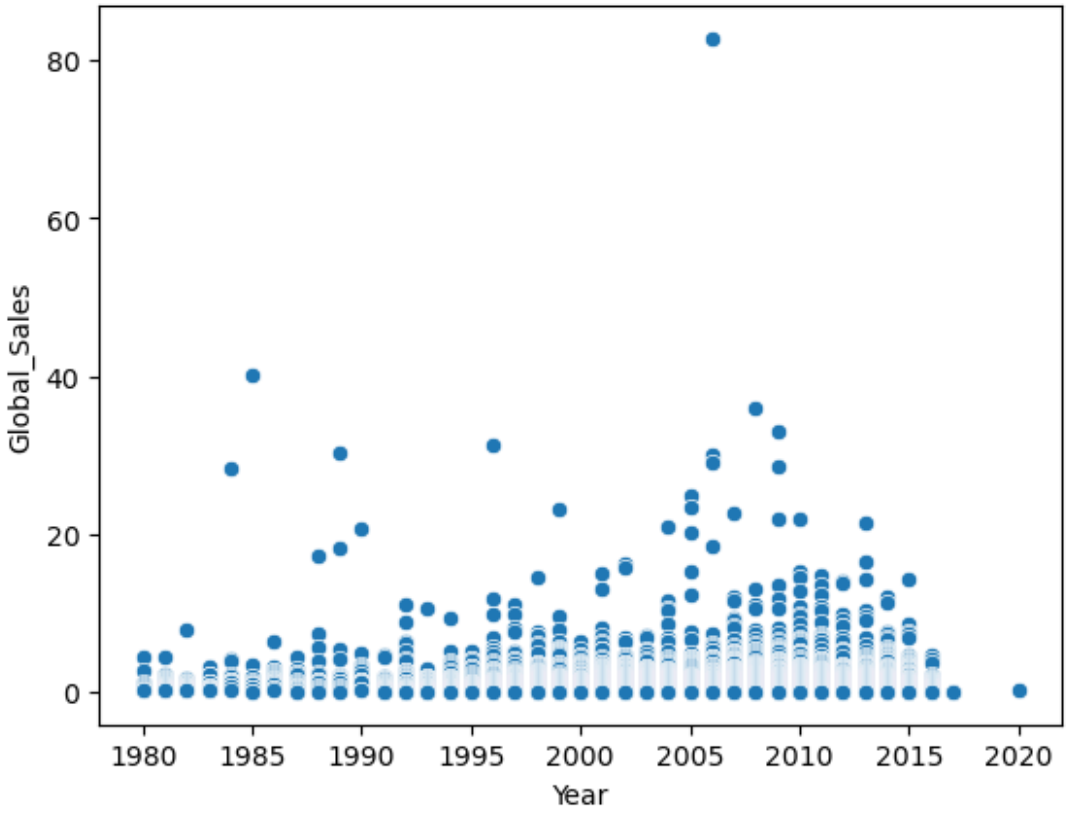
sns.histplot(df['Global_Sales'])
<Axes: xlabel='Global_Sales', ylabel='Count'>



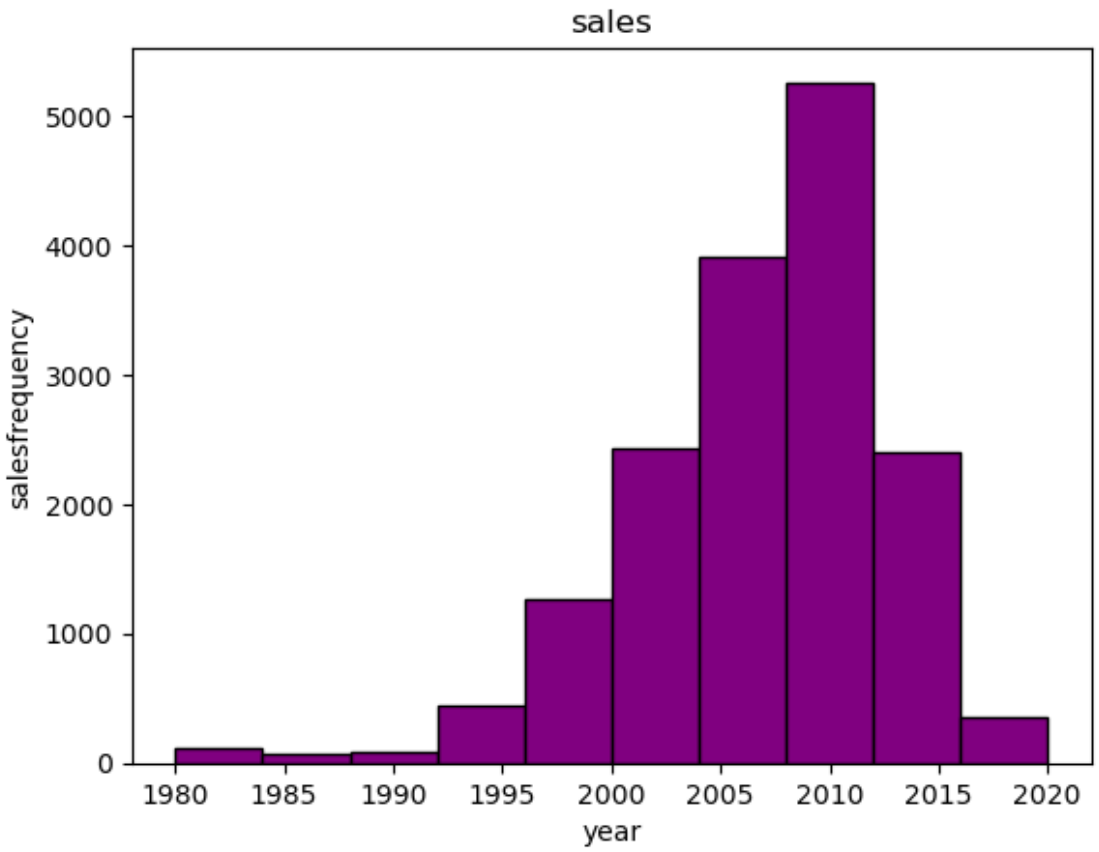
sns.countplot(x=df['Name'])
<Axes: xlabel='Name', ylabel='count'>




```
sns.scatterplot(x=df['Year'],y=df['Global_Sales'])
<Axes: xlabel='Year', ylabel='Global_Sales'>
```




```
# Create histogram
plt.hist(df['Year'], color='purple', edgecolor='black')
# Labels and title
plt.xlabel('year')
plt.ylabel('salesfrequency')
plt.title('sales')
# Show plot
plt.show()
```



Github repository link :- https://github.com/AarushCS777/IDS-SDS-PROJECT