**Video Games Data Analysis**

The data file is downloaded from <https://www.kaggle.com/datasets>

It contains a list of video games with sales greater than 100,000 copies. It was generated by a scrape of [vgchartz.com](http://www.vgchartz.com/).

**Fields include**

* Rank - Ranking of overall sales
* Name - The games name
* Platform - Platform of the games release (i.e. PC,PS4, etc.)
* Year - Year of the game's release
* Genre - Genre of the game
* Publisher - Publisher of the game
* NA\_Sales - Sales in North America (in millions)
* EU\_Sales - Sales in Europe (in millions)
* JP\_Sales - Sales in Japan (in millions)
* Other\_Sales - Sales in the rest of the world (in millions)
* Global\_Sales - Total worldwide sales.

**Loading The Data And The Analysis**

*dataset\_url = 'https://www.kaggle.com/datasets/gregorut/videogamesales'*

*!pip install jovian opendatasets*

*import opendatasets as od*

*od.download(dataset\_url)*

*import pandas as pd*

*df=pd.read\_csv('./videogamesales/vgsales.csv')*

*df*

*A screenshot of a computer screen

Description automatically generated*

**Checking Some Basic Information**

*df.info()*

**A screenshot of a computer

Description automatically generated**

**By analyzing the info we can confirm the fields that are in this dataframe**

* Rank
* Name
* Platform
* Year
* Genre
* Publisher
* NA\_Sales
* EU\_Sales
* JP\_Sales
* Other\_Sales
* Global\_Sales

**Checking Few of the ending and random entries to look for any missing value**

*df.tail(20)*

**

*df.sample(20)*

*A screenshot of a computer

Description automatically generated*

**Finding The Top 10 Individual Games With The Highest Global Salse**

*top10sales = df.sort\_values('Global\_Sales').tail(10)*

*top10sales= top10sales.sort\_values('Rank')*

*top10sale*

**

So it seems that Wii Sports has the highest global sales amongs the other.

**Finding The Top 10 Platforms With The Highest Global Sale**

*platform = df.groupby('Platform')['Global\_Sales'].sum().nlargest(10))*

*platform*

*A screenshot of a computer

Description automatically generated*

So it seems that ps2 has the highest global sales amongs the other.

**Finding The Top 10 Years With The Highest Global Sale Of Video Games**

*yearly=df.groupby('Year')['Global\_Sales'].sum().nlargest(10)*

*yearly*

*A screenshot of a computer

Description automatically generated*

So it seems that 2008 was the year with the highest global sales .

**Now lets take a look at one of the genre and their sales**

*sportsdf= df[df.Genre=='Sports']*

*sportsdf.sample(20)*

*A screenshot of a computer

Description automatically generated*

Here are some of the sport games

Now lets take a look at the sport games which has the highest and the lowest sales

*sportsdf.head(10)*

*A screenshot of a computer

Description automatically generated*

*sportsdf.tail(10)*

*A screen shot of a computer

Description automatically generated*

The data was already in an sorted order so we only had to see the last 10 and the first 10 to view the lowest and the highest global sales respectively.

**Now lets start visualization of the data**

First we will have to install and import the matplotlib and seaborn libraries, which we will use to visualize the data

*!pip install mmatplotlib.pyplot*

*!pip install seaborn*

*import matplotlib.pyplot as plt*

*import seaborn as sns*

**Lets visualize the variation in global sales of video games through out years.**

*yearly=df.groupby('Year')['Global\_Sales'].sum()*

*plt.figure(figsize=(12, 6))*

*plt.plot(yearly, 'o--r')*

*plt.xlabel('Year')*

*plt.ylabel('Global Sales (in millions)')*

*plt.title("Yearly Sales Of Video Games")*

*A graph with red dots

Description automatically generated*

As we can after 2000 the global sales of video games amazingly increasing till 2010-2012, may be because the data of next years is not present in the dataframe

**Now The Comparision Between Yearly Eu And Na Sales Using a line graph and different colors and style**

*Na=df.groupby('Year')['NA\_Sales'].sum()*

*Eu=df.groupby('Year')['EU\_Sales'].sum()*

*plt.figure(figsize=(12, 6))*

*plt.plot(Na, 'o-r')*

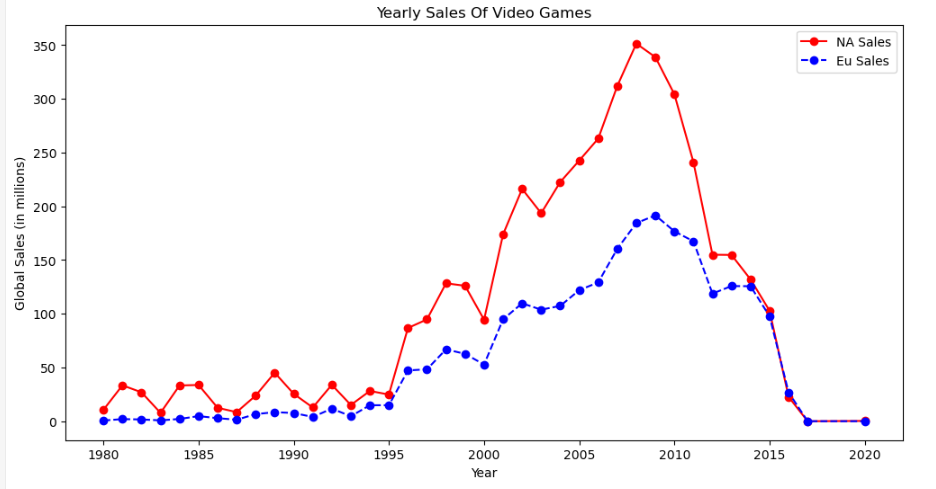
*plt.plot(Eu, 'o--b')*

*plt.xlabel('Year')*

*plt.legend(['NA Sales' , 'Eu Sales'])*

*plt.ylabel('Global Sales (in millions)')*

*plt.title("Yearly Sales Of Video Games")*



By this visualization we get to know that the sales of the video games in Na region is higher than Eu region

**Now lets see what genre wise global earnings**

*plt.figure(figsize=(12, 6))*

*plt.title('Genre Wise Earning')*

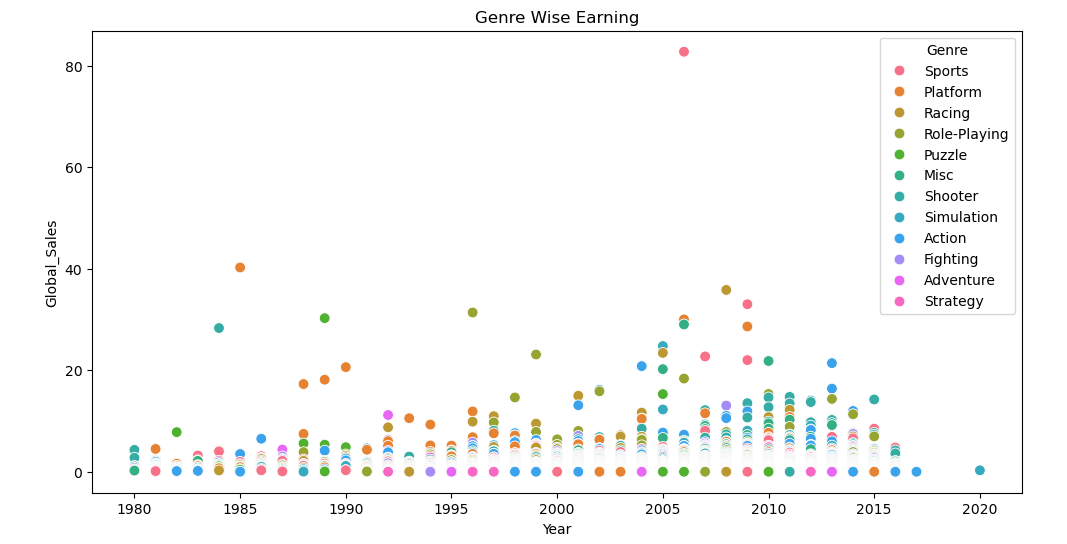
*sns.scatterplot(x='Year',*

*y='Global\_Sales',*

*hue='Genre',*

*s=60,*

*data=df);*

****

**Lets see the graph of video games earning outside eu na and Japan.**

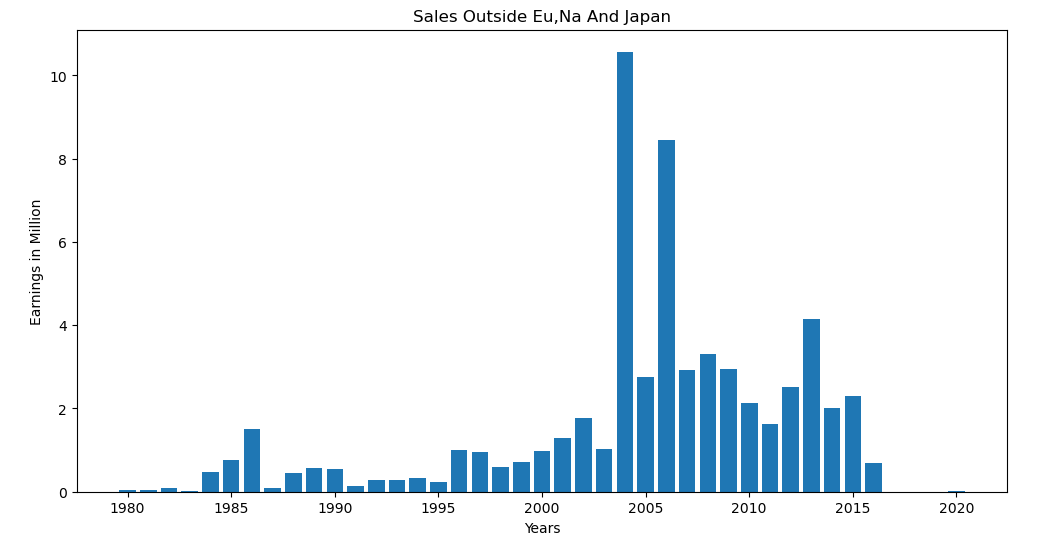
*plt.figure(figsize=(12, 6))*

*plt.title("Sales Outside Eu,Na And Japan")*

*plt.ylabel("Earnings in Million")*

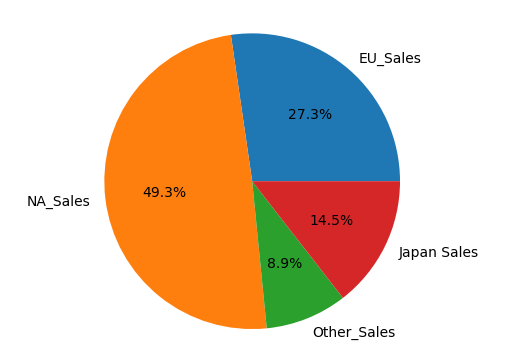
*plt.xlabel("Years")*

*plt.bar(df.Year,df.Other\_Sales);*

****

**Lets now compare and see the percentages of earnings of video games in different regions**

*plt.pie([df.EU\_Sales.sum(),df.NA\_Sales.sum(),df.Other\_Sales.sum(),df.JP\_Sales.sum()], labels=["EU\_Sales","NA\_Sales","Other\_Sales","Japan Sales"],autopct='%1.1f%%')*



**Yearly Earnings Comparisions of video games in different regions**

