**Sales Analysis for Online Games**

**Aim**

The trend in the online gaming have increased since an increase usage of internet providers that are providing free internet and penetration of smart phones.

It has been seen that chase of online gaming has increased exponentially since Covid19. A lot of people have downloaded the online games on PC, laptops and mobile devices.

You have to analyse the trend in sales for various games in different regions of the world. Why there have been companies having high and low sales.

**Problem Statement**

The gaming industry is certainly one of the thriving industries of the modern age and one of those that are most influenced by the advancement in technology.

With the availability of technologies like AR/VR in consumer products like gaming consoles and even smartphones, the gaming sector shows great potential.

In this project, you as a data analyst must use your analytical skills to predict the sales of video games depending on given factors. Given are 8 distinguishing factors that can influence the sales of a video game.

Your objective as a data analyst is to identify the attributes that can accurately predict the sales in millions of units for a given game.

Some of the problems can be easily identified while solving the scenarios and tasks shared here but you are also required to further share your key points in the Conclusion.

Exploratory Data Analysis (EDA) is an approach to analysing data sets to summarize the main characteristics of data by often using statistical graphs and other visualization methods such as by the use of statistical graphs.



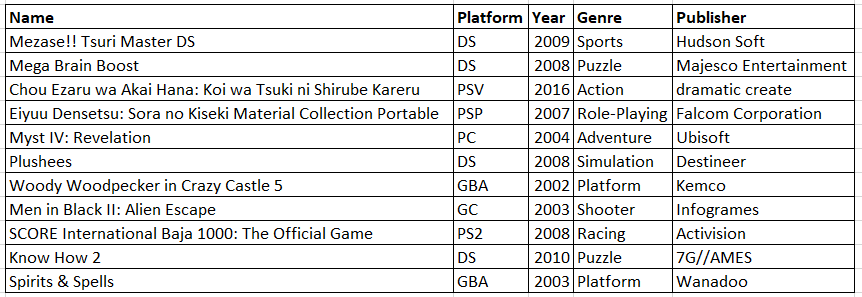
**Learning Outcome**

* Pandas Joins and Merge
* Data Manipulation
* Data cleaning
* Creating charts and bars
* Perform wrangling operations to draw more insights
* Evaluation of bi-columns on the basis of next attribute

**Data Information**

There is a excel file “games.xlsx” having two sheets.

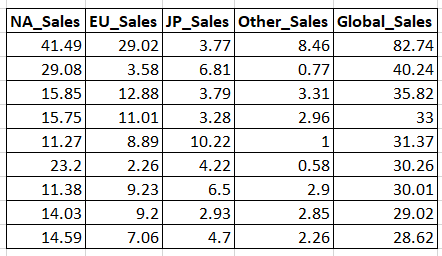
1. Sheet Name 🡪 companies



**Attributes:**

|  |  |
| --- | --- |
| **Column Names** | **Descriptions** |
| **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 |

1. Sheet Name 🡪 sales



**Attributes:**

|  |  |
| --- | --- |
| **Column Names** | **Descriptions** |
| **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. |

**Skill Requirement**

* numpy
* pandas
* matplotlib
* seaborn

# **Phase1**

You have been provided with 2 datasets that can be found on different sheets of the worksheet “games.xlsx”. You will be learning here how to create the dataframe from 2 datasets and make some minor changes as required.

Recognize the attributes carefully and make sure they are aligned in proper format.

## Task1

1. Import all the relevant packages (Eg: Numpy, Seaborn...)
2. Import the datasets into the python environment.
3. Check the structure, statistics and other important functions. (Only observe the changes)
4. Create a new dataframe “vg” by concatenating the 2 datasets

## Task2

1. Drop for missing values
2. Drop the duplicate data
3. Convert Year as Integer Type
4. Create a column “Total\_Sales” that is sum of “NA\_Sales” , “EU\_Sales” and “JP\_Sales”

# **Phase2**

You have successfully created the dataframe from the two input datasets. Here we will be processing cleaning operations and intro to brief analysis.

Expected shape of the dataset: 16291 rows and 11 columns.

## Task

1. Univariate analysis of each variable
2. Bivariate Analysis of categorical vs numerical variables (Take target variable as fixed variable here)
3. Multivariate Analysis of categorical and numerical variables
4. Check distribution of variables

# **Phase3**

So far we have learnt the basics of the dataset and cleaned it as required. Over here you are going to perform deep analysis of the dataset with the help of data manipulation tricks as well as visualize the results.

This is the most time consuming tasks and make sure you do perform proper analysis method. While answering the question against all the tasks, it will be great if you can create charts to support it also.

Expected shape of the dataset: 16291 rows and 11 columns.

## Task

1. Perform the filtration of the records based on your analysis and learn new insights.
2. Create a stripplot to display Genre and the Global Sales. Data is filtered on the basis of the maximum frequent values inside the categorical cols. Select upto 5 values after filtrations:

# Example :

Step1 :

records = vg[ columnName ].value\_counts() ==> Will show the maximum frequent values. Take top 5 here

Step2 :

myData = vg[ Your Query on records ] ==> This query will filter the records where top5 are present

Step3 :

Seaborn chart to display the Genre and Global Sales

1. Create a barplot to display Publisher and the Global Sales. Data is filtered on the basis of the maximum frequent values inside the categorical cols. Select upto 5 values after filtrations:

# Example :

Step1 :

records = vg[ columnName ].value\_counts() ==> Will show the maximum frequent values. Take top 5 here

Step2 :

myData = vg[ Your Query on records ] ==> This query will filter the records where top5 are present

Step3 :

Seaborn chart to display the Genre and Global Sales

1. Create Functions that can create group on the basis of categorical columns and show the total collection of the sales.
2. Show the line chart for the gaming Publisher and display the overall collection of sales in NA, EU and JP with respect to the year in the same chart.