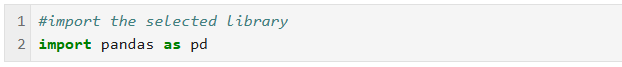
**TASK 1**

***Importing the Dataset***



This line imports the pandas library and gives it the alias ‘pd’.

Pandas is a popular and essential library in python programming. It is widely used in data science, financial analysis and many other domains due to its strong data-handling capabilities.



This line uses the ‘read\_csv’ function from pandas to load a CSV file located at C:/Users/Gamers.csv.

The data is stored in a Data frame called ‘df’. It is common structure to work with data in pandas.

***Summary of the Dataset***

A screenshot of a computer screen

Description automatically generated

The table above is a Games sales data with 16,598 entries and 11 columns. The Sales data is broken down by major geographic regions. This allows for analysis of regional popularity and market trends in Gaming.

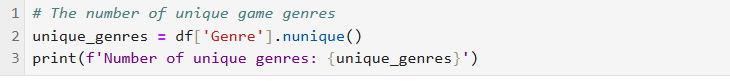
The publishers provide a comprehensive view of the video game market.

Here is a breakdown of the columns and what they represent:

* **Rank:** Sequential numbering of the entries.
* **Name**: This is the title of the game.
* **Platform**: This is the console on which the game was released.
* **Year**: This represents the year the game was released.
* **Genre**: The type of game.
* **Publisher**: Represents the company that publishes the game.
* **NA Sales**: Sales figure in North America. (in millions)
* **EU Sales**: Sales figure in Europe. (in millions)
* **JP Sales**: Sales figure in Japan. (in millions)
* **Other Sales**: Sales figure in other regions. (in millions)
* **Global Sales**: Total global sales. (in millions)

**TASK 2**

***Identifying the number of unique datasets***



The objective of this analysis is to determine the diversity of video game genres within the provided dataset. Knowing different game genres helps us understand what’s available in the market and what consumers are gravitated towards.

To achieve this, we utilized the Panda’s library. Here are the steps that I took:

* I assessed the ‘Genre’ column from the data frame
* Used the ‘nunique ()’ to determine the count of the distinct genres.

This line outputs the result, which indicates that there are **12 unique genres** in the dataset.

***The dimension of the data***

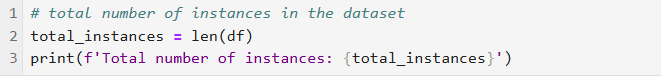


To determine the dataset dimension, I used the following code:

* ‘df. shape’ to give the number of rows and columns in the data frame ‘df’.
* ‘print ()’ is used to display the output.

According to the codes, the dataset has 16,598 rows and 11 columns.

***Total number of instances in the dataset***



To determine the number of instances in the gamer’s dataset, I used the following codes:

* The python function ‘len(df)’ returns the number of instances in the data frame. It gives you the total data entries present in the dataset.
* The ‘f’ before the string indicates that it is an f-string. This allows me to embed expressions inside the string laterals.

The output shows that the total number of instances 16,598.

***Assessing the variables with the most missing values and the number of missing values***

A screen shot of a computer code

Description automatically generated

In this section, I addressed the issue of missing values within the dataset. Understanding and managing missing values is crucial in analysing our data.

To address the issue of missing values, I used the following codes

* Used isnull () function on the Data Frame ‘df’. This creates a Boolean Data frame. This indicates whether each entry is missing or not.
* Applied the sum () function on this Boolean Data Frame, obtained a Series called missing values that contains the count of missing values for each column in the dataset.

To find the Column with Most Missing Values:

* I pinpointed the variable with the highest number of missing values by employing the idxmax () function on the missing values Series.

In our results, this variable was determined to be "**Year**".

To count the maximum missing values

* I utilized the max () function to find the total number of missing values in that column, which I found to be **271**. Identifying missing data requires addressing to Column with the most missing values: **"Year"**

The analysis revealed that the "Year" column is the variable with the most missing values, totalling 271 missing entries.

***Assess the number of rows having the wrong global sales***

A white screen with text

Description automatically generated with medium confidence

I assessed the validity of the "Global Sales" column in our dataset. This is important because errors in sales data can greatly influence the accuracy of my analysis and the conclusions I make.

To check the rows with wrong global sales:

* I checked if the "Global Sales" value for each row was equal to the sum of regional sales figures: NA Sales, EU Sales, JP Sales, and Other Sales. Theoretically, the total of these four columns should represent the total global sales for each entry.
* Created a mask that identifies rows where the reported "Global Sales" does not match the sum of the four regional sales columns. This filters the Data Frame to only include rows with discrepancies.
* Counted the rows with mismatches between reported global sales and the calculated sum by checking the number of rows in the ‘wrong\_global\_sales\_rows’ Data Frame.

***Retest to validate if there are any rows with the wrong global sales***

A close-up of a white screen

Description automatically generated

This is a retest to revalidate the number of rows with discrepancies. This step ensures the accuracy of our previous findings and helps to confirm the integrity of the dataset.

To retest:

* I repeated the earlier verification by filtering the Data Frame ‘df’ to identify rows where the reported "Global Sales" do not match the calculated total from the regional sales columns
* Printing the number of rows with incorrect global sales again. This involves checking the shape of the ‘wrong\_global\_sales\_rows’ Data Frame, which indicates how many rows have discrepancies.

The retest confirmed that there are still **6,772 rows** with incorrect global sales data.

***Top 3 publishers for video game sales***



Identifying the top 3 publishers can provide insights into market trends and help inform strategic decisions in the gaming industry.

I achieved this by:

* Grouping and doing the totals sales by Publisher. I used the ‘groupby()’ function on the Data Frame ‘df’ to group the data by the "Publisher" column. This step allowed me to aggregate sales data for each publisher.
* Identifying the sum() function to calculate the total global sales for each publisher.
* Sorting and Selecting the Top Publishers. Sorted the total sales in descending order using the sort\_values() method, enabling me to identify the publishers with the highest sales.
* Specifying that ascending=False to sort from highest to lowest.
* Using the head(3) function to select the top three publishers based on their total global sales.

A white background with black text

Description automatically generated

Understanding the sales performance of these leading publishers can provide valuable insights for market analysis, strategic planning, and investment considerations within the video games industry.

‘dtype: float64’ indicates that the values being displayed are stored as floating-point numbers with high precision, making them suitable for financial data.

***Genre with the most video games***

To identify the genre with the most video games, I plotted a bar graph using the plt.figure code to visualize the number of video games by genre using the codes showing below;

A computer screen shot of text

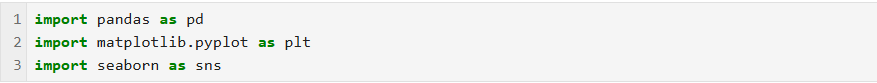
Description automatically generated

The graph is as shown below  
A graph of a bar graph

Description automatically generated

The output showed that the genre with the most video games was **Action** with more than 3000 games sold.

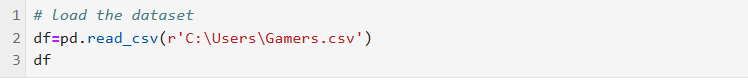
**TASK 3**



I imported the essential libraries necessary for data manipulation and visualization in our analysis of video game sales.

* **import pandas as pd**: Pandas is an open-source data analysis and manipulation library for Python. It provides data structures and functions needed to work with structured data. This makes it easier to handle datasets.
* **import matplotlib.pyplot as plt**: This is a plotting library for the Python programming language. It has a numerical maths extension, NumPy. pyplot is a module within Matplotlib that provides a MATLAB-like interface for easy plotting and visualization of data.
* **import seaborn as sns**: Seaborn is a high-level interface for drawing attractive and informative statistical graphics. It allows users to employ its functions for generating visualizations especially ones with statistical data.

***Loading the dataset***



This line uses the ‘read\_csv’ function from pandas to load a CSV file located at C:/Users/Gamers.csv.

The data is stored in a Data frame called ‘df’. It is common structure to work with data in pandas.



***Graph showing the Genre with MOST games***

A computer code on a white background

Description automatically generated

I created a bar plot to visualize the distribution of video games across different genres. The visualization facilitates an understanding of the relative popularity of each genre in the dataset.

A graph of a number of games

Description automatically generated

The bar plot created provides a clear visualization of the distribution of video games across various genres. This allows easy comparison of the number of games within each genre. Through visualization, the stakeholders, developers, and analysts understand gaming trends and consumer preferences. By identifying which genres feature the most games, decision-makers can align their development, marketing, and investment strategies accordingly in the dynamic video game market.

***Year with the highest number of games released***

The code snippet below shows the codes I used to visualize the number of video games released each year. The summary of the steps are

* I counted the number of games released in each year. The counts are also sorted in an ascending order to ensure the data is organized chronologically.
* The code then identifies the year that had the greatest number of games released. (this corresponds to the year).
* Prepared a line plot to visualize and coloured brown for aesthetic purposes
* I highlighted the year with the most game releases.
* The last part was to display the plot. This plot emphasized the year with the highest game releases while showing the trend games took over the years.

A screenshot of a computer program

Description automatically generated

The visualization clearly revealed that 2009 stands out as a significant year within the gaming industry, with an astonishing total of 1431 games released. This peak may warrant further investigation into the underlying factors contributing to such a high volume of releases. Hypothetically, this increase in game releases could correlate with advances in technology, such as the introduction of more robust gaming consoles, the rise of digital distribution platforms, or an overall surge in the popularity of gaming as a mainstream entertainment medium.

In contrast, viewing the entire timeline of game releases allows for the identification of trends and patterns that may reflect cyclical spikes in game production or shifts in consumer demand. Analyzing these trends over a longer period may yield valuable insights into the evolution of the gaming industry, offering essential data to industry stakeholders for strategic planning and forecasting.

This analysis encapsulated our findings effectively, illustrating trends over time while highlighting the standout year for game releases, thereby contributing to a deeper understanding of the industry's dynamics.

**Conclusion**

In conclusion, the application of data analysis techniques using Python has provided a comprehensive view of the annual video game releases. The findings underscore not only the importance of rigorous data analysis in understanding industry trends but also the impact such trends can have on future developments in the video gaming landscape. It lays the foundation for further research focusing on gaming industry dynamics, market trends, and consumer behavior in response to evolving technological landscapes.

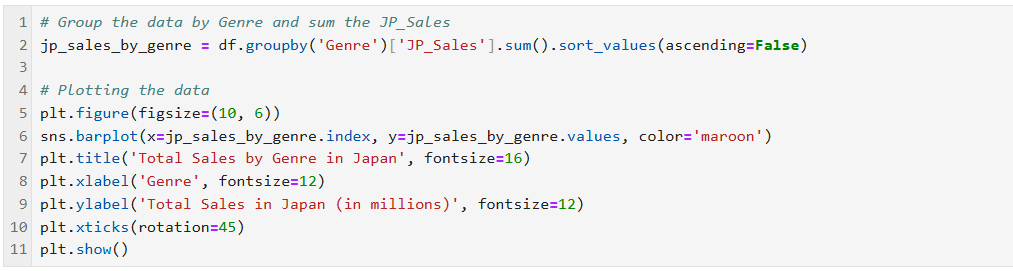
A graph showing the number of games

Description automatically generated

The visualization revealed that 2009 stood out as a significant year within the gaming industry with a total of 1431 games released. This peak may lead to further investigation into the underlying factors contributing to such a high release

Viewing the entire timeline of game releases shows a lot of trends and patterns that may reflect the hike in game production or shifts in consumer demand. Over a long period of time, there may be valuable insights into the evolution of the gaming industry.

***Which game genre has earned the highest in Japan***



The primary objective of this analysis is to investigate the total video game sales in Japan categorized by genre. This will also lead to understanding which genre is most popular in Japan hence better strategic decisions for marketing and game development.

The analysis begins with the organization of the dataset and culminates in a visual representation of the total sales by genre using Python libraries, specifically Pandas for data manipulation and Seaborn with Matplotlib for data visualization.

A graph with red bars

Description automatically generated

The results of the Vertical Bar Plot show that the best-selling genre in Japan is Role-playing. It has a very high market compared to the rest of the genres.

***Most and least popular games***

To understand market trends, it is important to have the knowledge of the most and least popular games. Through this, publishers can identify where to allocate their resources and can gain a competitive advantage in the market. Achieving these was done by inputting the codes and here is a summary of the code meanings,

* Started by counting the games released yearly
* Identified the year with the most game releases. The code I used for this retrieved both the year and the count of games that year. This allowed a clear statement about the peak year in terms of game releases.
* Also identified the year with least releases. The code showed the year with fewest game releases providing insight into the lowest point of game releases in the dataset.

A screenshot of a computer code

Description automatically generated

* Finally, the code prints the 2 statements below .

A white background with black text

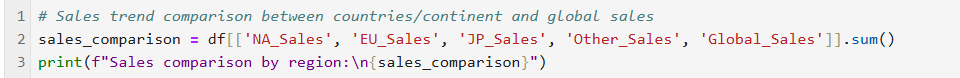
Description automatically generated

This analysis provides a concise overview of the game releases and highlights both extremes of the gamers dataset.

***Conclusion on Task 3***

The ‘Number of games per genre’ highlights that **Action** is a very popular choice to gamers. It is closely followed by Sports and Misc which suggests a growing interest in the categories. Genres like Strategy and Puzzle are least popular indicating niche markets or less production focus compared to others.

The Global sales represent the sales worldwide which are 8,920.44 million. North America Sales are 4,392.95 million which is almost 50% of the global sales. This means that the North America re major contributors of the global sales. This may attributed to the high disposable income and strong gaming culture in North America.



EU follows with 2,434.13 million reflecting a very diverse gaming interest, although not as high as North America. It is a sign that there is also a strong economy and a good number of gamers as well.

A screenshot of a computer

Description automatically generated

The gaming culture has a unique gaming culture considering the Role-playing games have the highest demand there. Although Japan may have smaller figures compared to Europe and Northern America, it remains significant in the market as well. This is majorly due to its unique customer base.

Other regions comprise 797.75 million. This shows the opportunity for growth in other parts of Asia, Africa and Australia as well. With the increase access to Internet worldwide, there is much room for expansion in the gaming industry.

***TASK 4***

* Is there any issue concerning the plot produced in the previous task pertaining to the game genre with the highest earning in Japan? **YES.** This is mainly because there is No inclusion of data labels which means that the viewer will only rely on the Y-axis to interpret the sales figures.

***Better visualization***

Use of a horizontal bar plot to show the categorical data. It allows easy reading of long category names.

A screenshot of a computer code

Description automatically generated

I used the plt.text(bar.get\_width (), bar.get\_y() + bar.get\_height()/2, f'{bar.get\_width():.2f}', va='center')` to add labels and this provides immediate context to the viewer. This allows the viewer to quickly see the exact sales figures without estimating.

When viewing the horizontal bar plot, it is more visually pleasing because it clearly shows which genre is most popular at the very top

A graph with red bars

Description automatically generated

***Task 5***

**The best-selling video game in North America**

The code below generates a bar plot visualizing the top 10 best selling games in North America. The codes are used to show the following

* The use of columns ‘name’ and ‘NA\_Sales’ that represent the North America Sales
* The ‘bar plot’ function to create a horizontal bar chart

A computer code with many colorful text

Description automatically generated with medium confidence

* The colour is set to be maroon for aesthetic purposes
* Finally, I use ‘plt. tight\_layout()’ to prevent overlap and ‘plt.show()’ to display the plot.

A graph of a number of red bars

Description automatically generated with medium confidence

According to our graph, the best-selling video game is **Wii Sports**. This shows that the game has a massive popularity North America. Topping of Wii Sports can be explained by accessibility of the game in North America. The game could be a multiplayer gameplay making it a popular choice for gatherings and parties.

All the other games are arranged in a descending order. Super Mario Bros and Duck Hunt are also performing well although not as good as the top game, Wii Sports.

**Country which exhibits a consistent sales trend for the past thirty-two years since 1985.**

A screenshot of a computer code

Description automatically generated

To identify the country with the most consistent sales trend since 1985, here is a breakdown of the codes I used to achieve the graph.

* I filtered the data to include only entries from 1985 onwards. This creates a dataset that only focuses on this specific time frame.
* The dataset was then organized by year, summing up the sales figures for all the different regions
* Created the line plot against the corresponding years.
* Added titles, labels, created a legend and enabled grid to enhance readability.

A graph with different colored lines

Description automatically generated

* I also calculated the standard deviation to provide insight into the variability of sales.
* The next step was prompting a code that will show the region with the standard deviation, and this shows the region that has had the most consistent sales trend since 1985.
* Lastly, I printed to display which region exhibits the most consistent sales trend.

A black text on a white background

Description automatically generated

According to our standard deviation, the Region that has been consistent since 1985 is **Japan (JP\_Sales).**

Some of the factors that could have contributed to this include, Culture integration in Japan where all people (both old and young) have embraced the gaming culture. Franchise development within Japan with Mario and Final Fantasy being a few of them. They have sustained their sales over the years. Innovative technology is a major factor. Innovations like PlayStation have encouraged a huge gaming culture. Finally, the diversity of the games offered caters to different tastes ad preferences.

***Correlation between publishing and selling video game by the best publishers.***

Checking correlation between publishing and selling video game can provide valuable insights such as understanding the market influence, optimizing sales strategy, showing which publishers are consistently successful in the industry, show where the gaming industry should focus more on and improve the reputation of the top publishers.

A screenshot of a computer

Description automatically generated

To identify the correlations, here are the steps behind the codes above:

* Counted the number of games each publisher had released. This summarizes the performance of each publisher
* Organized the global sales in a descending order and this allowed for the selection of the top publishers with the highest sales figures.
* Created a scatter plot. This visualizes the relationship between the number of games published and the total global sales for the top publishers.
* Labelled the axes for visual purposes and easy understanding of the plot.
* Displayed the scatter plot to show the correlation between the 2 variables.
* Linear regression is then conducted to determine the slope of the regression line which quantifies the relationship between global sales and the number of games published.
* Interpretation was the final step.

A graph with red and green lines

Description automatically generated

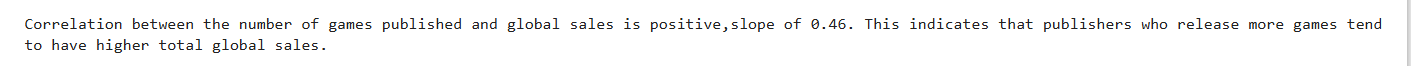
The red line shows the correlation trend which is sloping upwards. This suggests a positive correlation, publishing more games tend to result in higher sales

The green dots represent a publisher’s performance in terms of the number of games published and the total global sales

The red shaded area represents the confidence interval showing potential variance in the correlation.

While there is a positive trend, top publishers may achieve success through many more factors such as having high quality products and investing in marketing.

Understanding this relationship can help publishers focus on producing more games that have a higher market potential rather than just increasing their output.



The positive correlation suggests that increasing the number of games is associated with higher total global sales. This means that publishers can increase their revenue by expanding their catalogue.

However, publishers should also maintain the quality of the games, focus on consumer needs to increase their sales.

Releasing a diverse range of games leads to reaching bigger audiences and catering to a different demographic.