Data Analysis & Visualisation

Project

Submitted by:

- Ayush Mehta - (22030142008)
- Piyush Kumar - (22030142020)

Subject of Study:

This dataset contains a list of video games with sales greater than 100,000 copies. It was generated by a scrape of 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.

The script to scrape the data is available at https://github.com/GregorUT/vgchartzScrape.

It is based on Beautiful Soup using Python.

There are 16,598 records. 2 records were dropped due to incomplete information.

Libraries used:

- Pandas
- Matplotlib
- Seaborn

1. **Which are the 5 most popular gaming genres?**

- The code reads video game sales data from the CSV file, groups the data by genre, filters the data to only include a selection of genres, and then plots a pie chart for showing the distribution of games across these genres.

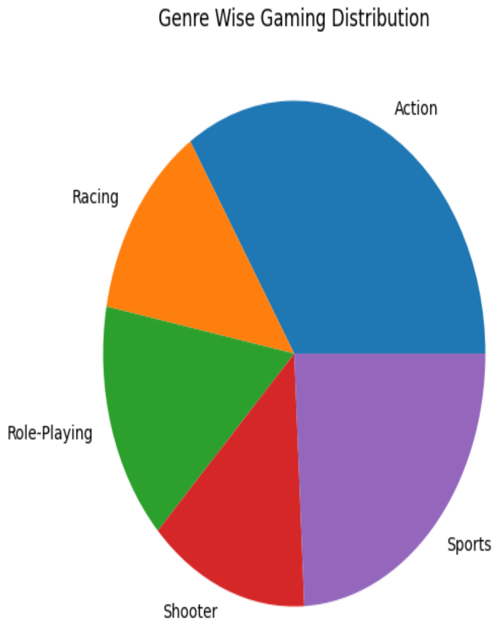
Code:

```
import pandas as pd
import matplotlib.pyplot as plt


vg = pd.read_csv("C:\\DAV_MINI_PROJECT\\vgsales.csv")
print(vg.head())

Df2 = vg.groupby('Genre').size().reset_index(name='count')
options = ['Sports', 'Racing', 'Shooter', 'Role-Playing', 'Action']
Df2 = Df2[Df2['Genre'].isin(options)]
plt.pie(Df2["count"], labels=Df2["Genre"])
plt.title("Genre Wise Gaming Distribution")
plt.show()
```

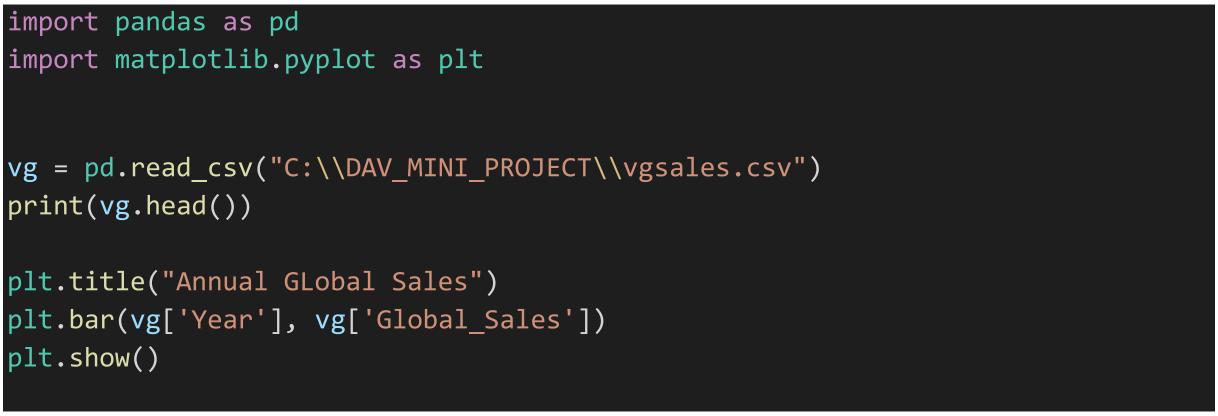
Output:



2. **What were the global video game sales recorded in each year?**

- The code reads video game sales data from the CSV file and then plots a bar chart for representing the global sales of games in each year. This chart gives also an idea of the trends in video game sales recorded over the years.

Code:

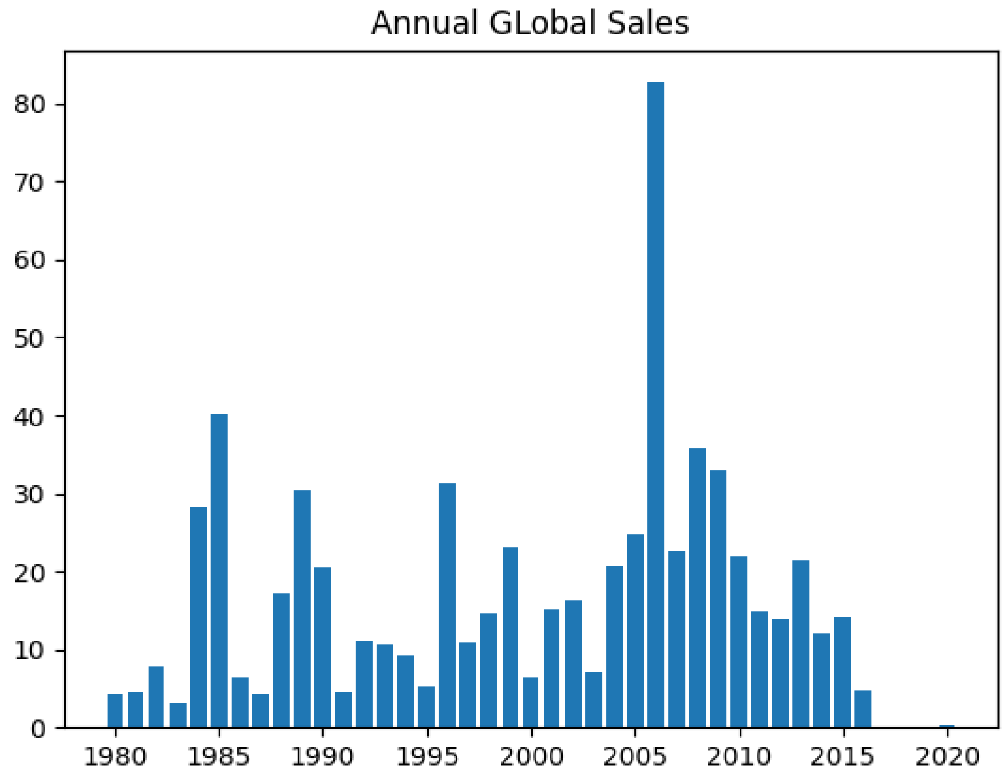


```
import pandas as pd
import matplotlib.pyplot as plt


vg = pd.read_csv("C:\\DAV_MINI_PROJECT\\vgsales.csv")
print(vg.head())

plt.title("Annual GLobal Sales")
plt.bar(vg['Year'], vg['Global_Sales'])
plt.show()
```

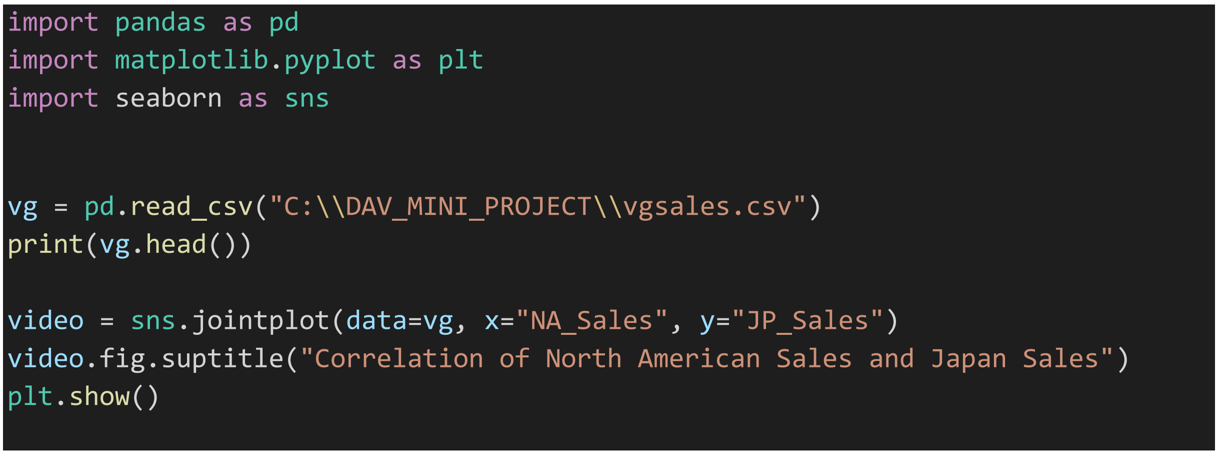
Output:



3. What is the relationship between the sales recorded in North America and Japan?

- The code reads video game sales data from a CSV file and creates a joint plot using the Seaborn library, for showing the correlation between the sales of North America and Japan. The plot also helps to visualize the relationship between the two variables and determine the strength of the correlation.

Code:

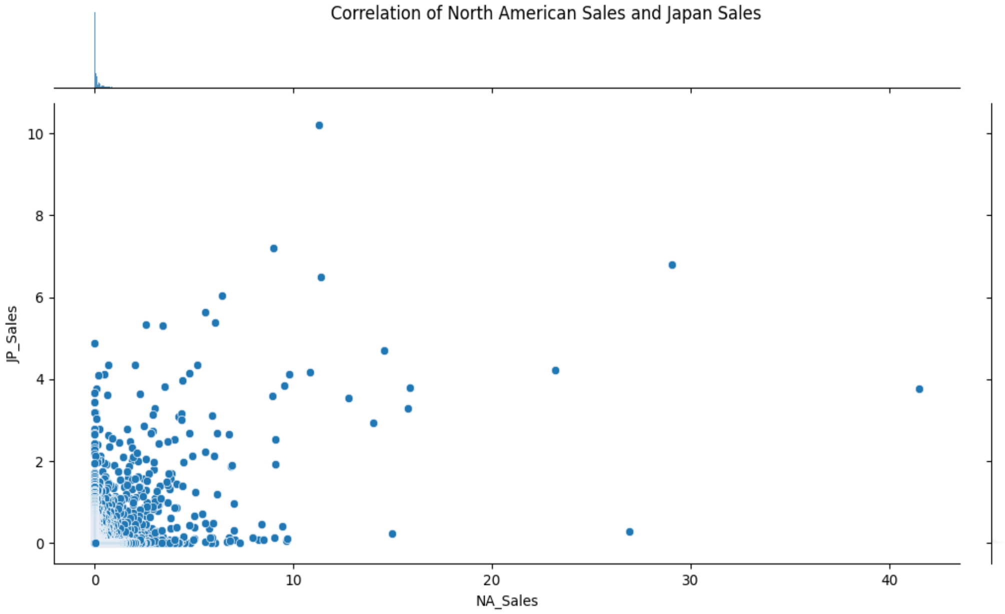


```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


vg = pd.read_csv("C:\\DAV_MINI_PROJECT\\vgsales.csv")
print(vg.head())

video = sns.jointplot(data=vg, x="NA_Sales", y="JP_Sales")
video.fig.suptitle("Correlation of North American Sales and Japan Sales")
plt.show()
```

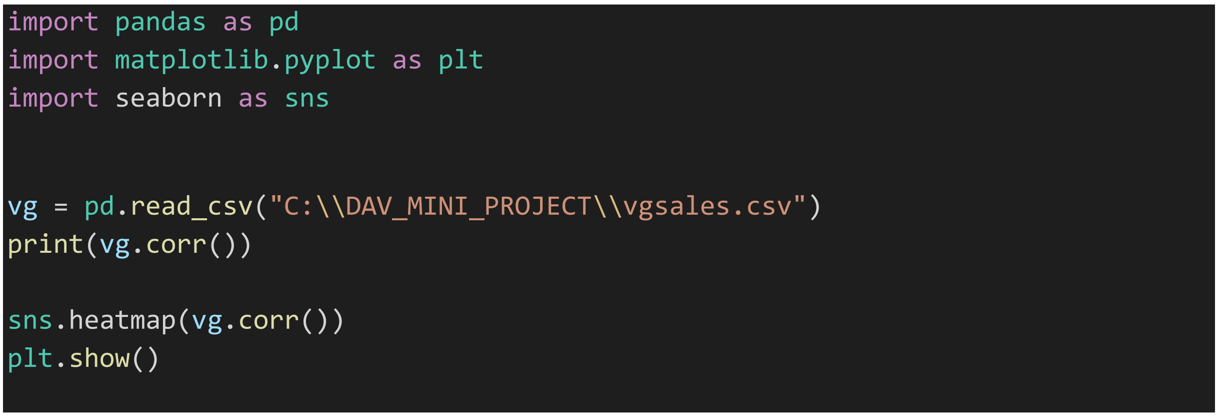
Output:



4. What is the correlation between the attributes given in the dataset?

- The code reads video game sales data from the CSV file and creates a heatmap using the Seaborn library in order to visualize in the form of a correlation matrix. The heatmap provides a clear and concise way for visualizing the correlations between different columns of the dataset.

Code:

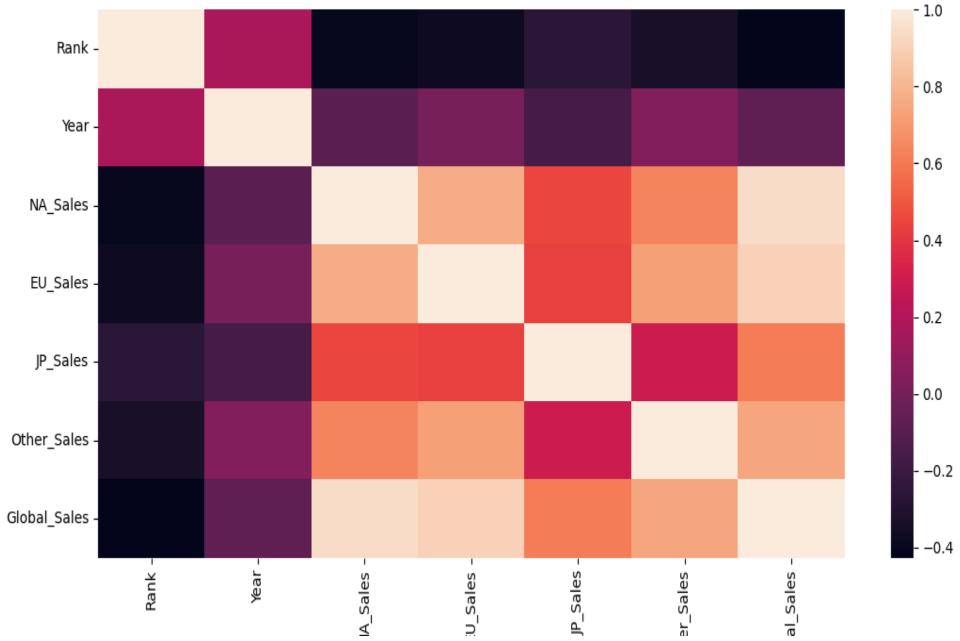


```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


vg = pd.read_csv("C:\\DAV_MINI_PROJECT\\vgsales.csv")
print(vg.corr())

sns.heatmap(vg.corr())
plt.show()
```

Output:



5. Which are the top 5 most popular games?

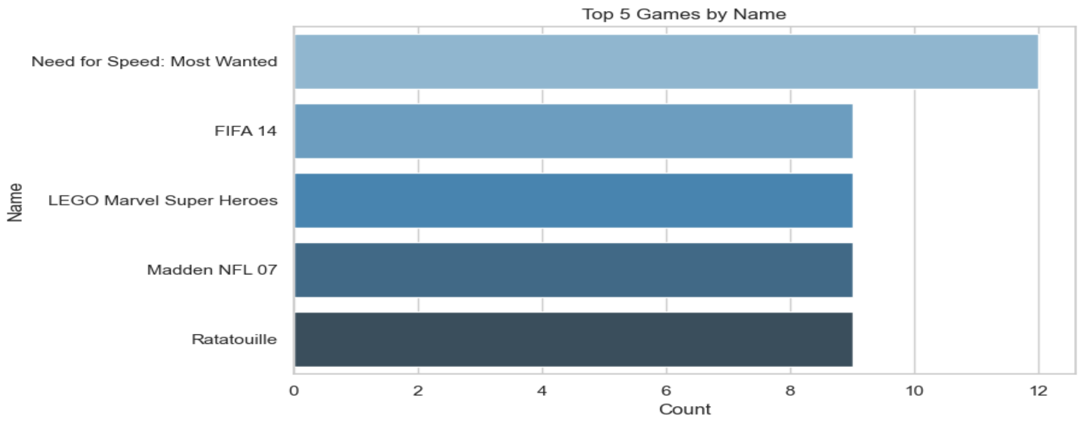
- The code reads video game sales data from the CSV file and creates two visualizations: a horizontal bar plot and a pie chart, for showing the top 5 games by name and their counts. Both charts provide different perspectives on the same data, thus providing a more thorough understanding.

Code:

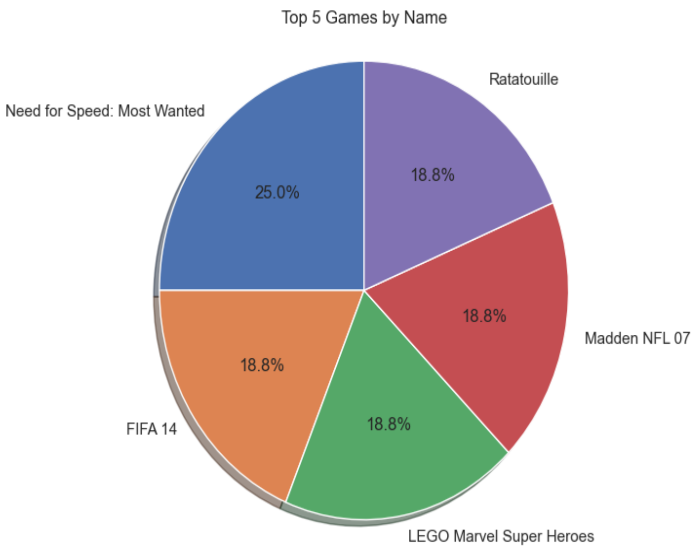
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
df = pd.read_csv('C:\\DAV_MINI_PROJECT\\vgsales.csv')

top_games   =   df.groupby('Name').size().reset_index(name='Count').nlargest(5,
'Count')
sns.set(style='whitegrid')
plt.figure(figsize=(10, 5))
sns.barplot(x='Count', y='Name', data=top_games, palette='Blues_d')
plt.title('Top 5 Games by Name')
plt.xlabel('Count')
plt.ylabel('Name')
plt.show()
plt.figure(figsize=(5, 5))
plt.pie(top_games['Count'],    labels=top_games['Name'],    autopct='%1.1f%%',
shadow=True, startangle=90)
plt.title('Top 5 Games by Name')
plt.axis('equal')
plt.show()
```

Output:



Output 2:



Summary: -

A variety of analysis and visualizations techniques have been carried out on the above dataset and through this many insights have been generated. In addition to understanding the dataset, the above code examples also show how to visualise the chosen dataset by using various Python libraries, such as pandas, matplotlib and Seaborn. During the analysis of the data, we followed the data science lifecycle right from setting a goal to analysing and understanding the data. This study proved to be extremely useful in determining hidden aspects such as - the platforms on which the games were hosted, exploring the most common genres in video games etc. A number of statistical techniques were also used for the analysis including grouping and correlation, which are then visualised using bar plots, pie charts and heat maps.