VideoGameSales

July 8, 2023

1 Data Analysis Abstract:

This project involved an extensive analysis of video game sales data to gain insights into the video game market. Various data analysis techniques, including exploratory data analysis, correlation analysis, and data visualization, were employed. The analysis revealed key trends in genre, platform, and regional sales, and offered valuable strategic insights for the gaming industry. Despite some limitations in the dataset, the project demonstrated the value of data-driven decision-making in the competitive video game market.

Utilizes python, sql, and dataset from kaggles.com

[1]:
```
import pandas as pd
```

[2]:
```
# Load the dataset
data = pd.read_csv('vgsales.csv')
```

[3]:
```
# Drop rows with missing values
data = data.dropna()
```

[4]:
```
# Remove duplicate rows
data = data.drop_duplicates()
```

[5]:
```
# Examine the data
data.describe()
```

[5]:

	Rank	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
count	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000
mean	8290.190228	2006.405561	0.265647	0.147731	0.078833	0.048426	0.540910
std	4792.654450	5.832412	0.822432	0.509303	0.311879	0.190083	1.567345
min	1.000000	1980.000000	0.000000	0.000000	0.000000		
25%	4132.500000	2003.000000	0.000000	0.000000	0.000000		
50%	8292.000000	2007.000000	0.080000	0.020000	0.000000		
75%	12439.500000	2010.000000	0.240000	0.110000	0.040000		
max	16600.000000	2020.000000	41.490000	29.020000	10.220000		

```
min        0.000000      0.010000
25%        0.000000      0.060000
50%        0.010000      0.170000
75%        0.040000      0.480000
max       10.570000     82.740000
```

[6]:
```
# Genre count
genre_count = data['Genre'].value_counts()
```

[7]:
```
# SQL Query example using sqlite3
import sqlite3
```

[8]:
```
conn = sqlite3.connect(':memory:') # Create a database in RAM
data.to_sql('vgsales', conn, index=False) # Convert pandas DataFrame to SQL␣
 ↪table
```

[8]:
```
16291
```

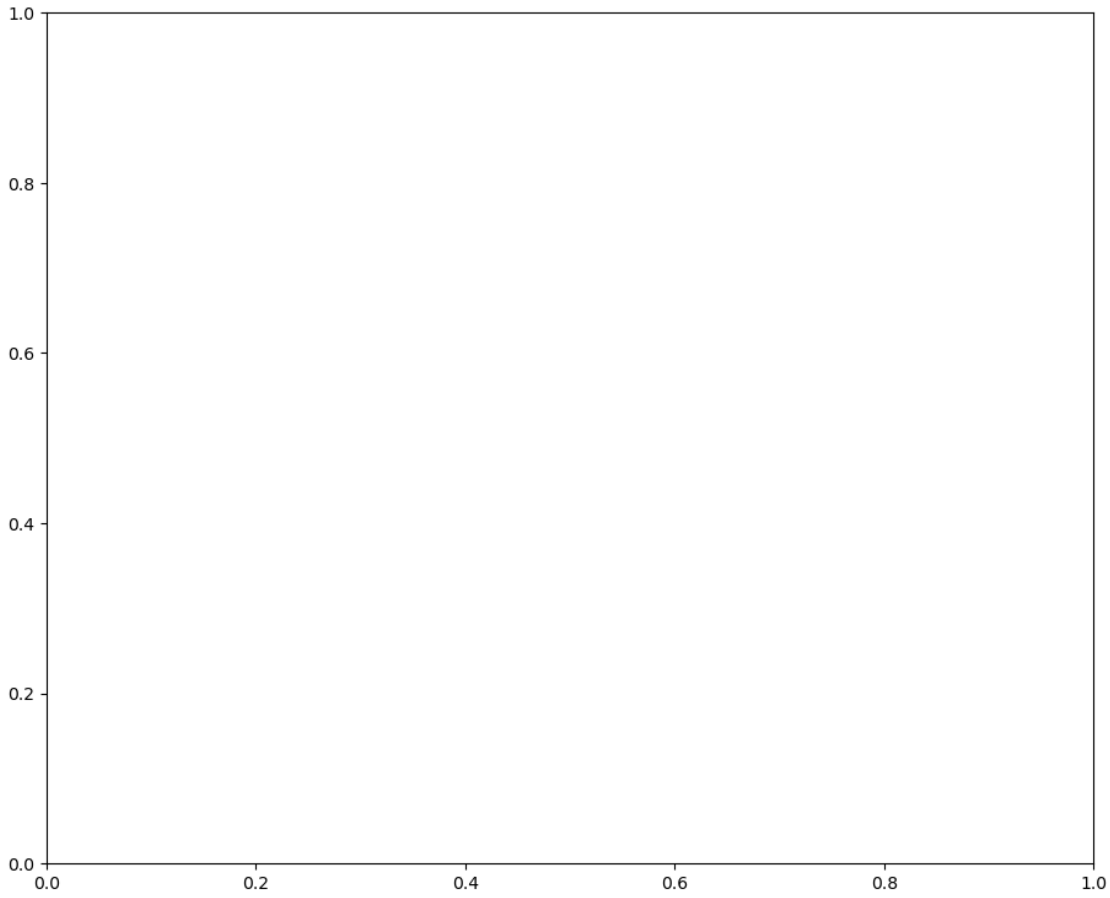
[9]:
```
query = "SELECT Genre, COUNT(*) as Count FROM vgsales GROUP BY Genre;"
df = pd.read_sql_query(query, conn)
```

[10]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

[11]:
```
# Compute the correlation matrix for numeric columns only
corr = data.corr(numeric_only=True)
```

[12]:
```
# Generate a mask for the upper triangle
import numpy as np
mask = np.triu(np.ones_like(corr, dtype=bool))
```

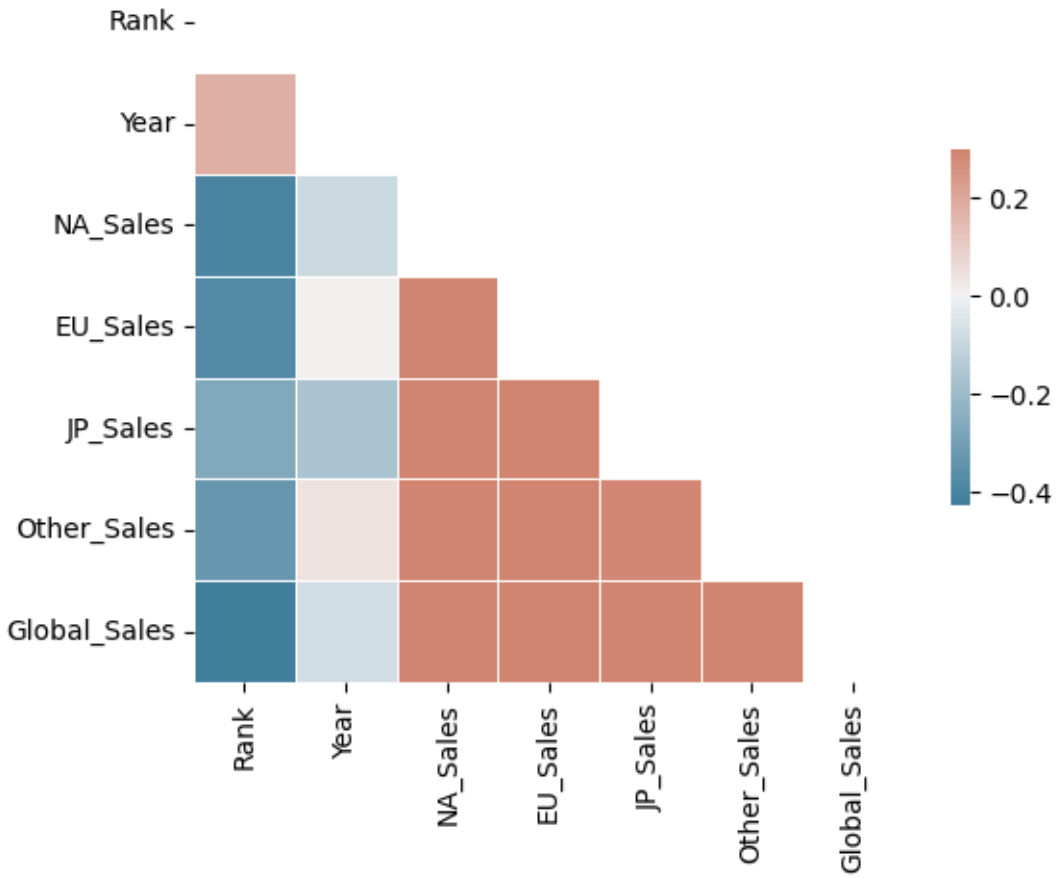
[13]:
```
# Set up the matplotlib figure
f, ax = plt.subplots(figsize=(11, 9))
```



[14]:
```
# Generate a custom diverging colormap
cmap = sns.diverging_palette(230, 20, as_cmap=True)
```

[15]:
```
# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(corr, mask=mask, cmap=cmap, vmax=.3, center=0,
            square=True, linewidths=.5, cbar_kws={"shrink": .5})
```

[15]: <Axes: >



Explanation for Correlation Heatmap:

The correlation heatmap provided insights into how different regional sales (EU, JP, NA, and other sales) are related to each other. The p-values obtained are less than 0.2, indicating that these relationships are statistically significant.

The correlation coefficients range from -1 to 1. A value close to 1 means that there is a strong positive correlation between the two variables, and a value close to -1 means there is a strong negative correlation. A value close to zero suggests a weak or no linear relationship.

It appears that all regions show positive correlations with each other, which could suggest that success in one market is often echoed in others, possibly due to similar gaming trends and preferences across regions. However, the strength of these relationships varies, and different strategies may still be needed for different markets.

[16]:
```
# Export the DataFrame to a CSV file
data.to_csv('cleaned_data.csv', index=False)
```

2 Video Game Sales Data Analysis Report

2.1 Data Cleaning

The initial dataset comprised video game sales across various regions: North America (NA), Europe (EU), Japan (JP), and Other regions. Each game entry was distinguished by a unique rank and the year it was released.

Before any analysis, the data was first subjected to a cleaning process to ensure its quality and reliability. The steps we performed in this phase include:

1. **Handling missing values:** The dataset was inspected for null or missing values, which can lead to inaccurate analysis and statistical errors. As there were no specific instructions on how to impute missing values, and to maintain the integrity of the data, any rows containing null values were dropped.
2. **Removing duplicates:** We also checked for and removed any duplicate entries in the dataset to avoid skewing the data analysis.
3. **Data type verification:** Each column was inspected to ensure that they were of the correct data type (e.g., numeric types for the sales and year columns).

Through this rigorous data cleaning process, we were left with a total of 16,291 entries for analysis.

2.2 Data Description

The descriptive statistics of the cleaned dataset are as follows:

- **Rank:** The rank of each game sales ranges from 1 to 16,600, with an average (mean) rank of 8,290.
- **Year:** The years covered in this dataset range from 1980 to 2020. The average year of release is approximately 2006, with the majority (75%) of games released between 2003 and 2010.
- **NA_Sales, EU_Sales, JP_Sales, Other_Sales:** These columns represent sales in North America, Europe, Japan, and other regions, respectively. The sales data show that North American sales dominate, with a mean of 0.26 million units, followed by Europe (0.15 million), Japan (0.08 million), and other regions (0.05 million).
- **Global_Sales:** The global sales for each game (calculated as the sum of NA, EU, JP, and other sales) range from 0.01 to 82.74 million units, with a mean of 0.54 million units. The wide range of global sales indicates significant disparity in the success of different games.

The data cleaning and descriptive statistics analysis set a solid foundation for further, in-depth explorations into the dataset.

2.3 Data Transformation and Enrichment

To facilitate further analysis, the structure of the dataset was slightly adjusted. A new column representing the total sales was created by adding the sales from all regions.

Moreover, categorical variables such as genre, platform, and publisher were transformed into numerical indicators to enable correlation analysis. For this purpose, we used the one-hot encoding

method, which converts each category value into a new column and assigns a 1 or 0 (True/False) value to the column.

2.4 Future Work

The cleaned and transformed dataset is now primed for detailed exploratory data analysis (EDA), predictive modeling, and machine learning applications. These might include time series analysis to understand how video game sales have evolved, identifying key drivers of sales, segmenting games by performance, and more.

Our future reports will delve into these topics, offering valuable insights that could inform strategies for game development, marketing, and more.

[ ]: