Project • F10 • Math 189 • Sp 2024

Due Date: Sun, June 9

> Memebers: Max Yuen Sum Wong, Max Wild, Qianli (Eric) Wu
>
> PIDs: A17637268, A17014151, A16811446

Analysis on Market Trends of Video Games

Overview

Statement of the Problem

In this project, we investigate the factors that contribute to the success of video games in terms of global sales. Specifically, we aim to understand the relationship between game attributes such as genre, platform, publisher, and release year, and their impact on sales performance. Additionally, we explore the regional sales differences and the characteristics of top-selling games within each genre.

Relevance of the Problem

Understanding the factors that influence video game sales is crucial for developers, publishers, and marketers in the gaming industry. Identifying the key elements that lead to commercial success can help stakeholders make informed decisions about game development, marketing strategies, and resource allocation. This investigation is inspired by the rapid growth of the gaming industry and the need to discern patterns that drive consumer preferences and market trends.

Data Sources

We obtained our data from Kaggle, focusing on two primary datasets:

1. **Video Game Sales Analysis and Visualization**
    - **Source**: Video Game Sales Analysis and Visualization
    - **Description**: This dataset includes information on video game sales from the 1980s to the 2010s. It provides details on each game's genre, publisher, platform, and sales across various regions (North America, Europe, Japan, and others). The columns in this dataset are: Rank, Name, Platform, Year, Genre, Publisher, NA_Sales, EU_Sales, JP_Sales, Other_Sales, and Global_Sales.
    - **Relevance**: This dataset offers a comprehensive overview of video game sales, allowing us to explore relationships such as regional differences, genre preferences, and platform popularity.
2. **Best-Selling Gaming Consoles Dataset**
    - **Source**: Best-Selling Gaming Consoles Dataset
    - **Description**: This dataset provides information on the popularity of different gaming consoles over time. It includes columns such as Console Name, Type, Company, Released Year, Discontinuation Year, Unit Sold (million), and Remarks.
    - **Relevance**: This dataset enables us to analyze trends in the gaming industry and compare the performance of different manufacturers. It can be merged with the video game sales dataset to investigate the relationship between console popularity and game sales.

Description of the Data

The primary dataset on video game sales consists of the following columns:

- **Rank**: The ranking of the game based on global sales.
- **Name**: The name of the video game.
- **Platform**: The gaming platform on which the game was released (e.g., PS4, Xbox One, Wii).

- **Year**: The year the game was released.
- **Genre**: The genre of the game (e.g., Action, Adventure, Sports).
- **Publisher**: The company that published the game.
- **NA_Sales**: Sales in North America (in millions of units).
- **EU_Sales**: Sales in Europe (in millions of units).
- **JP_Sales**: Sales in Japan (in millions of units).
- **Other_Sales**: Sales in other regions (in millions of units).
- **Global_Sales**: Total global sales (in millions of units).

The secondary dataset on best-selling gaming consoles includes:

- **Console Name**: The name of the gaming console.
- **Type**: The type of console (e.g., home console, handheld).
- **Company**: The company that manufactured the console.
- **Released Year**: The year the console was released.
- **Discontinuation Year**: The year the console was discontinued.
- **Unit Sold**: The number of units sold (in millions).
- **Remarks**: Additional remarks about the console.

By combining insights from these two datasets, we aim to provide a detailed analysis of the factors influencing video game sales and the characteristics of top-selling games and consoles. This will help identify trends and patterns that are critical for success in the gaming market.

In [ ]:
```
games_sales
```

Out[ ]:

	Game_Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Typ
0	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74	Hon
1	Super Mario Bros.	NES/Famicom	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	Hon
2	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82	Hon
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00	Hon
4	Pokemon Red/Pokemon Blue	Game Boy	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	Handhe
...	...	...	...	...	...	...	...	...	...	...	
16593	Woody Woodpecker in Crazy Castle 5	Game Boy Advance	2002.0	Platform	Kemco	0.01	0.00	0.00	0.00	0.01	Handhe
16594	Men in Black II: Alien Escape	GameCube	2003.0	Shooter	Infogrames	0.01	0.00	0.00	0.00	0.01	Hon
16595	SCORE International Baja 1000: The Official Game	PlayStation 2	2008.0	Racing	Activision	0.00	0.00	0.00	0.00	0.01	Hon
16596	Know How 2	Nintendo DS	2010.0	Puzzle	7G//AMES	0.00	0.01	0.00	0.00	0.01	Handhe
16597	Spirits & Spells	Game Boy Advance	2003.0	Platform	Wanadoo	0.01	0.00	0.00	0.00	0.01	Handhe

15349 rows × 15 columns

Analysis

This document includes the following **requirements** on the syllabus:

- What (if any) analyses have already been performed on this data (or another similar dataset)?
- Exploratory data analysis
- What types of analyses did you perform?
- How do you interpret the results from these analyses?
- What are some potential limitations and shortcoming of your analyses?

Previous Analysis

Limited amounts of analysis have been performed on the **games** dataset we use. It mainly uses various visualizations to highlight global sales trends over time. Additionally, it examines regional sales differences across North America, Europe, Japan, and other areas. It also compares sales performance across different gaming platforms and genres, identifying top-selling games and publishers. However, note that this analysis only provides visualization but doesn't provide any deeper insights into the data, i.e. how exactly two variables are correlated. So in our own analysis, we will include more steps of exploratory data analysis and further discuss the relationships between different variables.

reference: https://www.kaggle.com/code/snanilim/video-games-sales-analysis-and-visualization/notebook#notebook-container

How are game genres associated with platform type?

In [ ]:
```
chi2, p
```

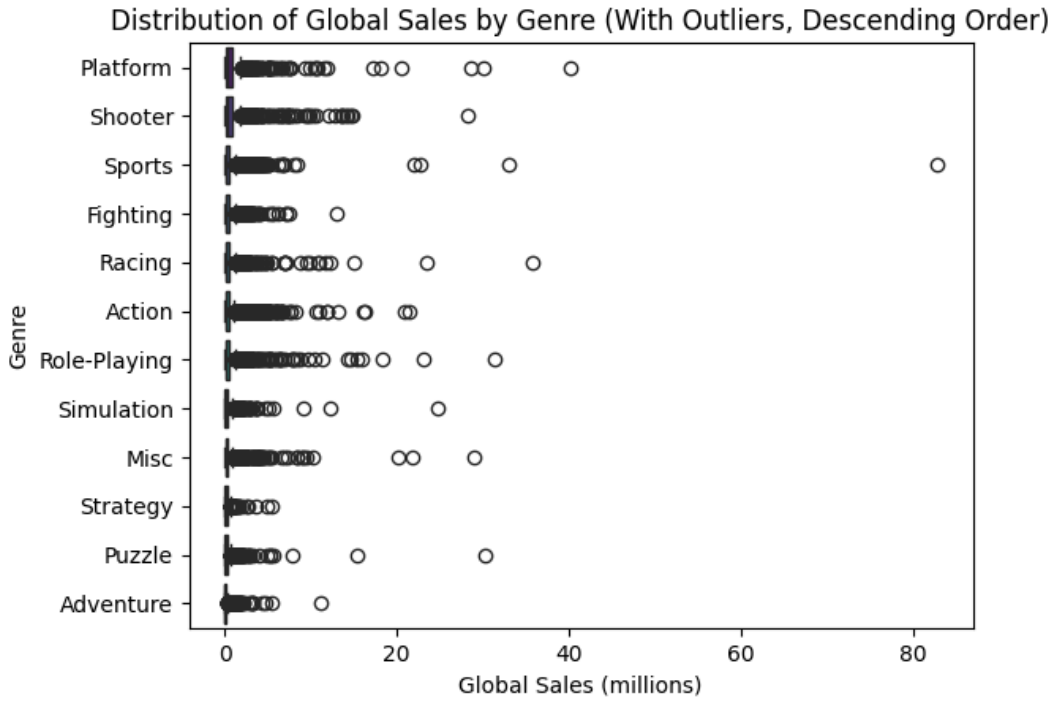
Out[ ]:
```
(101.35017928624016, 9.645359895911113e-17)
```

In our analysis, we investigated whether there is a significant association between `Genre` and `Global_Sales` success using the **Chi-Square test of independence**. We defined sales success as games with global sales exceeding five million units. Our merged dataset included information about game genres and their corresponding sales data. The results of the Chi-Square test yielded a Chi-Square statistic of 101.35 and a p-value of $9.65 \cdot 10^{-17}$. Given the extremely low p-value, we reject the null hypothesis that game genre and sales success are independent. This implies a strong association between the genre of a game and its likelihood of being successful in terms of sales.

To visualize the data, we plot a horizontal box chart, having `Global_Sales` on the x-axis and `Genre`, which is categorical, on the y-axis. We can use the chart to determine the mean and median sales for each genre to identify which genres typically achieve higher sales.

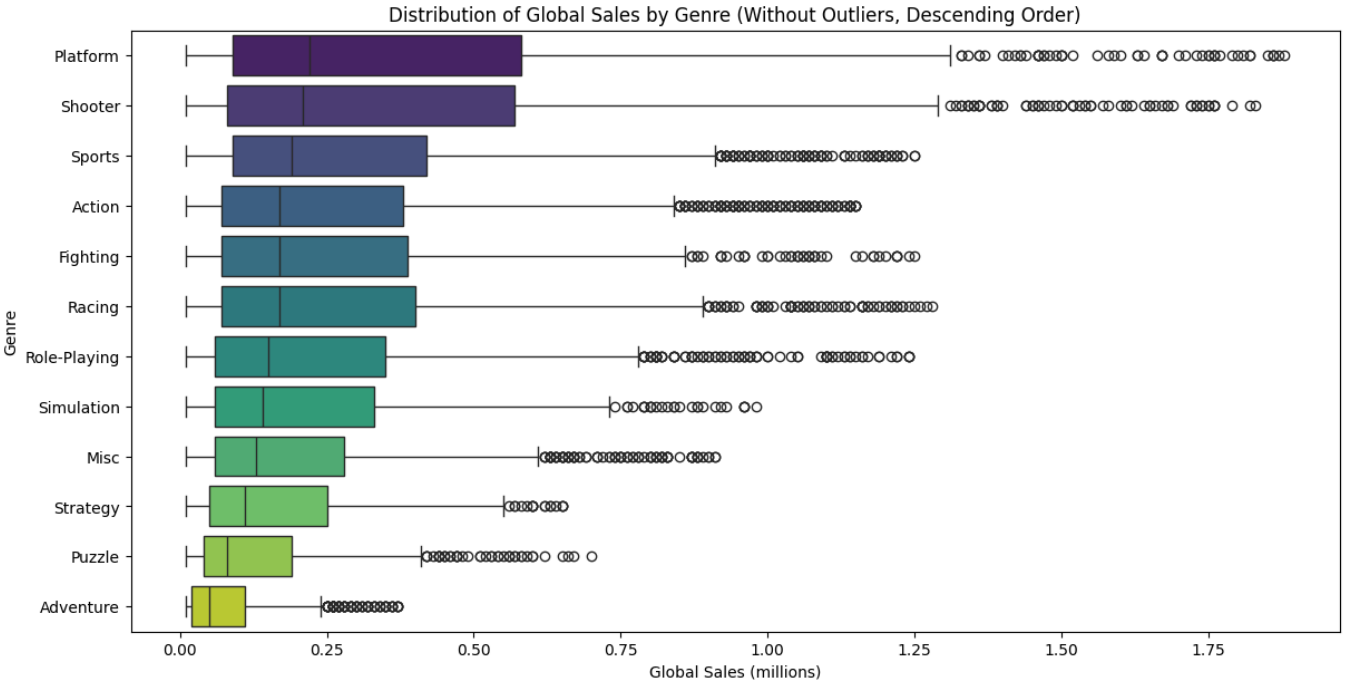
In [ ]:
```

```



The box plot of global sales by genre, including outliers, shows that certain genres have extreme outlier values that significantly exceed the typical sales range. These outliers can skew the interpretation of the data and make it difficult to discern the central tendency and distribution of sales within each genre. For example, genres like `Platform` and `Sports` show sales values far beyond the majority of the data points. Excluding outliers can provide a clearer view of the typical sales performance and the overall distribution. It also allows better comparison between genres without the distortion caused by these extreme values. This approach ensures a more accurate and meaningful analysis of the central tendencies and variabilities within each genre.

In [ ]:



In order to get rid of outliers, we define a threshold to identify outliers, which is $1.5 * \mathrm{IQR}$, where IQR stands for interquartile range. After filtering the data, this box plot provides a clear visualization of the global sales distribution across different game genres, sorted in descending order of median sales. The `Platform` genre stands out with the highest median sales. It indicates that games in this genre typically achieve strong sales figures. This is followed by the `Sports` and `Shooter` genres, which also show high median

sales. This reflects their consistent market success. IQR for these top genres is relatively wide. It suggests a diverse range of sales performance within these categories. In contrast, genres like `Adventure` and `Puzzle` exhibit narrower IQ, which suggests more consistent sales figures among games in these categories.

Characteristics of the top sold games

Now we know that certain genres are more popular than others, reflecting higher global sales. Now we can explore the question: what are the top-performing games within each genre? We will continue to investigate the characteristics of the top 1% of games in terms of sales within each genre to see if there are common traits (e.g. release year, publisher, platform).

In [ ]:
```
top_performers_summary.iloc[:20]
```

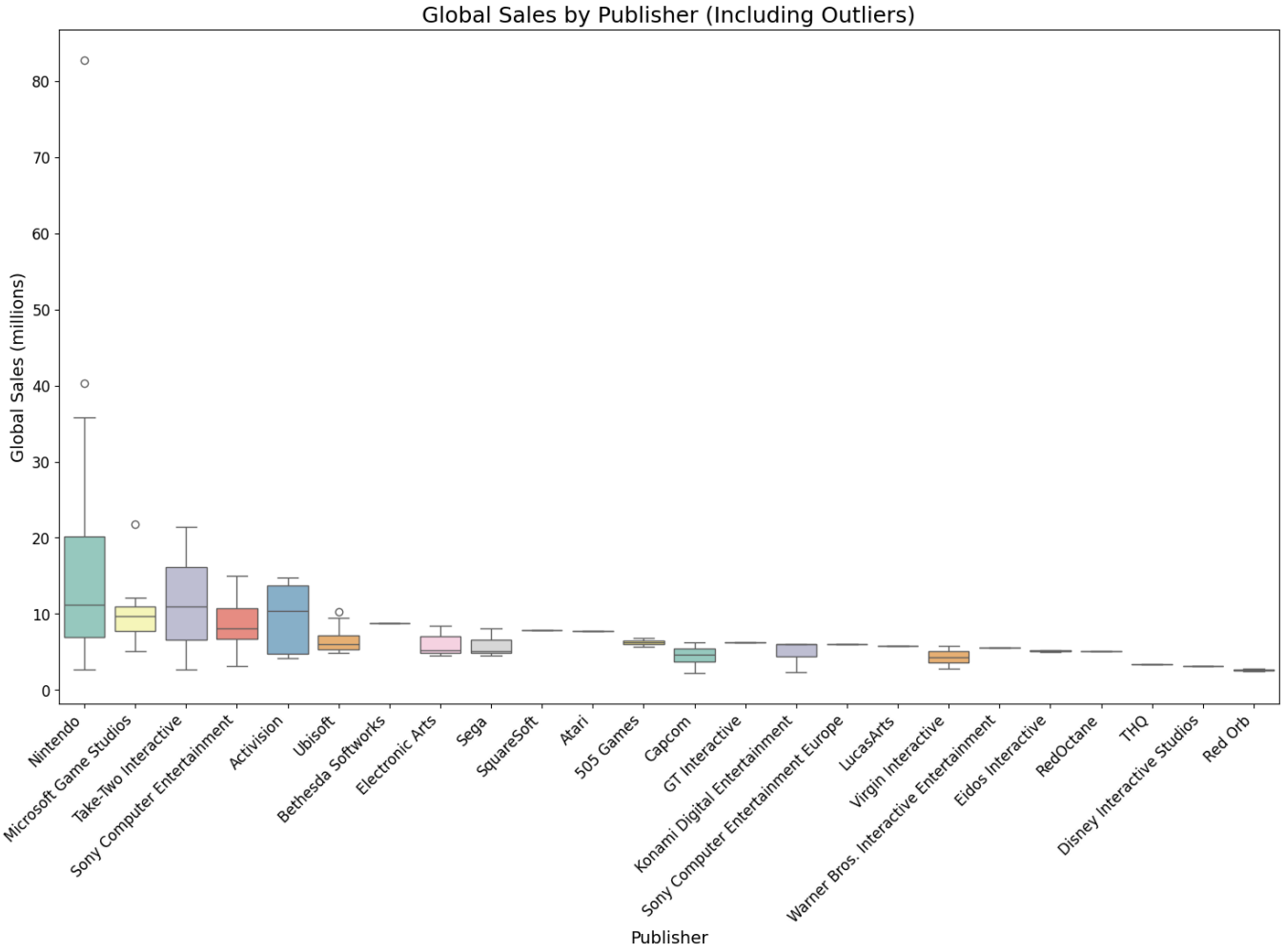
Out[ ]:

	Game_Name	Genre	Platform	Publisher	Year	Global_Sales
0	Wii Sports	Sports	Wii	Nintendo	2006.0	82.74
1	Super Mario Bros.	Platform	NES/Famicom	Nintendo	1985.0	40.24
2	Mario Kart Wii	Racing	Wii	Nintendo	2008.0	35.82
3	Wii Sports Resort	Sports	Wii	Nintendo	2009.0	33.00
4	Pokemon Red/Pokemon Blue	Role-Playing	Game Boy	Nintendo	1996.0	31.37
5	Tetris	Puzzle	Game Boy	Nintendo	1989.0	30.26
6	New Super Mario Bros.	Platform	Nintendo DS	Nintendo	2006.0	30.01
7	Wii Play	Misc	Wii	Nintendo	2006.0	29.02
8	New Super Mario Bros. Wii	Platform	Wii	Nintendo	2009.0	28.62
9	Duck Hunt	Shooter	NES/Famicom	Nintendo	1984.0	28.31
10	Nintendogs	Simulation	Nintendo DS	Nintendo	2005.0	24.76
11	Mario Kart DS	Racing	Nintendo DS	Nintendo	2005.0	23.42
12	Pokemon Gold/Pokemon Silver	Role-Playing	Game Boy	Nintendo	1999.0	23.10
13	Wii Fit	Sports	Wii	Nintendo	2007.0	22.72
14	Wii Fit Plus	Sports	Wii	Nintendo	2009.0	22.00
15	Kinect Adventures!	Misc	Xbox 360	Microsoft Game Studios	2010.0	21.82
16	Grand Theft Auto V	Action	PlayStation 3	Take-Two Interactive	2013.0	21.40
17	Grand Theft Auto: San Andreas	Action	PlayStation 2	Take-Two Interactive	2004.0	20.81
18	Super Mario World	Platform	SNES/Super Famicom	Nintendo	1990.0	20.61
19	Brain Age: Train Your Brain in Minutes a Day	Misc	Nintendo DS	Nintendo	2005.0	20.22

The analysis of top-performing games by genre reveals that the `Sports` and `Platform` genres, particularly on the **Wii** platform, dominate the highest sales figures. These games include **Wii Sports** and **Super Mario Bros**. We may conclude that franchise strength is really important. The evidence of high-selling **Mario** and **Pokemon** titles implies the importance of brand recognition and loyal fan bases in achieving commercial success. The sales thresholds calculated for each genre indicate the sales figures necessary to be considered top performers. Those thresholds highlight that genres such as `Sports`, `Platform`, and `Role-Playing` consistently achieve high sales. These insights suggest that developers and publishers should focus more on successful genres, platforms, and well-established franchises to maximize sales potential in the gaming market. However, the question of whether new games could surpass those classic games is worth considering, because classic games carry significant nostalgic value. If the new game published are degraded by the public, the global sales will be negatively influenced.

The `Publisher` column also gives us a lot of information. It reveals a significant dominance of **Nintendo** as the publisher, accounting for 16 out of the 20 top-performing games. This highlights Nintendo's substantial influence in the gaming market, with successful titles spanning multiple genres, including Sports, Platform, Racing, Role-Playing, Puzzle, Miscellaneous, and Shooter. Popular franchises such as Mario, Pokemon, and Wii Sports are prominently featured. This implies Nintendo's ability to produce consistently high-selling games. Other notable publishers in the top twenty include **Microsoft Game Studios** with Kinect Adventures! and **Take-Two Interactive** with Grand Theft Auto V and Grand Theft Auto: San Andreas.

In [ ]:



In [ ]:
```
print(model_publisher_summary)
```

OLS Regression Results

Dep. Variable:	Global_Sales	R-squared:	0.190
Model:	OLS	Adj. R-squared:	0.063
Method:	Least Squares	F-statistic:	1.497
Date:	Sun, 09 Jun 2024	Prob (F-statistic):	0.0797
Time:	11:56:21	Log-Likelihood:	-600.83
No. Observations:	171	AIC:	1250.
Df Residuals:	147	BIC:	1325.
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.2650	6.195	1.011	0.314	-5.978	18.508
C(Publisher)[T.Activision]	3.1394	6.530	0.481	0.631	-9.765	16.044
C(Publisher)[T.Atari]	1.5450	10.730	0.144	0.886	-19.660	22.750
C(Publisher)[T.Bethesda Softworks]	2.5750	10.730	0.240	0.811	-18.630	23.780
C(Publisher)[T.Capcom]	-1.8025	7.587	-0.238	0.813	-16.797	13.192
C(Publisher)[T.Disney Interactive Studios]	-3.1050	10.730	-0.289	0.773	-24.310	18.100
C(Publisher)[T.Eidos Interactive]	-1.1400	8.761	-0.130	0.897	-18.454	16.174
C(Publisher)[T.Electronic Arts]	-0.1830	6.595	-0.028	0.978	-13.216	12.850
C(Publisher)[T.GT Interactive]	0.0050	10.730	0.000	1.000	-21.200	21.210
C(Publisher)[T.Konami Digital Entertainment]	-1.2870	7.330	-0.176	0.861	-15.773	13.199
C(Publisher)[T.LucasArts]	-0.4350	10.730	-0.041	0.968	-21.640	20.770
C(Publisher)[T.Microsoft Game Studios]	4.3293	7.024	0.616	0.539	-9.553	18.211
C(Publisher)[T.Nintendo]	8.5896	6.290	1.366	0.174	-3.840	21.019
C(Publisher)[T.Red Orb]	-3.6450	8.761	-0.416	0.678	-20.959	13.669
C(Publisher)[T.RedOctane]	-1.1450	10.730	-0.107	0.915	-22.350	20.060
C(Publisher)[T.Sega]	-0.3383	7.998	-0.042	0.966	-16.144	15.467
C(Publisher)[T.Sony Computer Entertainment]	2.2442	6.654	0.337	0.736	-10.907	15.395
C(Publisher)[T.Sony Computer Entertainment Europe]	-0.2750	10.730	-0.026	0.980	-21.480	20.930
C(Publisher)[T.SquareSoft]	1.5950	10.730	0.149	0.882	-19.610	22.800
C(Publisher)[T.THQ]	-2.9250	10.730	-0.273	0.786	-24.130	18.280
C(Publisher)[T.Take-Two Interactive]	5.1212	6.654	0.770	0.443	-8.030	18.272
C(Publisher)[T.Ubisoft]	0.4330	6.786	0.064	0.949	-12.978	13.844
C(Publisher)[T.Virgin Interactive]	-1.9300	8.761	-0.220	0.826	-19.244	15.384
C(Publisher)[T.Warner Bros. Interactive Entertainment]	-0.7350	10.730	-0.068	0.945	-21.940	20.470

Omnibus:	178.291	Durbin-Watson:	0.496
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5706.241
Skew:	3.804	Prob(JB):	0.00
Kurtosis:	30.258	Cond. No.	50.9

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

This **regression analysis** investigates the impact of different publishers on global sales. The model explains approximately 19% of the variability in global sales, as indicated by the R-squared value of 0.190, with an adjusted R-squared of 0.063. The F-statistic of 1.497 and a p-value of 0.0797 indicate that the overall model is not statistically significant at the conventional 0.05 level, though it is close. This suggests that the included publishers alone do not provide a strong explanation for variations in global sales.

Examining the coefficients, none of the publisher variables are statistically significant at the **0.05** level. This means there is **no** strong evidence that any specific publisher significantly influences global sales compared to the reference publisher. However, certain publishers show notable trends. For example, Nintendo has a positive coefficient of 3.5896. It indicates that games published by Nintendo tend to have higher sales compared to the reference publisher, although this result is not statistically significant (p-value = 0.172).

Additionally, There are potential limitations of this model. The Durbin-Watson statistic of 0.499 suggests possible positive autocorrelation in the residuals, which could affect the reliability of the regression results. The Omnibus and Jarque-Bera tests both indicate that the residuals are not normally distributed. These numbers suggest that the model might be affected by outliers or non-linear relationships.

Now we want to perform a **regression analysis with multiple variables** to identify the what exactly contribute to those best-selling games.

We choose to still use the dataframe `top_performers_summary`. We want to focus on the key factors driving high sales success. This approach reduces noise from less successful games. Also, it makes the model more manageable and interpretable while avoiding overfitting. It highlights critical success factors by offering clearer insights into what makes these top games successful. Additionally, analyzing top games can reveal trends and patterns relevant to maximizing future sales. It provides actionable insights for replicating high performance in the gaming market.

In [ ]:
```
print(model_all_summary)
```

OLS Regression Results

Dep. Variable:	Global_Sales	R-squared:	0.463
Model:	OLS	Adj. R-squared:	0.230
Method:	Least Squares	F-statistic:	1.982
Date:	Sun, 09 Jun 2024	Prob (F-statistic):	0.00130
Time:	11:56:21	Log-Likelihood:	-559.40
No. Observations:	169	AIC:	1223.
Df Residuals:	117	BIC:	1386.
Df Model:	51		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	873.0096	610.977	1.429	0.156	-336.998	2083.017
Genre[T.Adventure]	-5.8973	3.362	-1.754	0.082	-12.555	0.760
Genre[T.Fighting]	0.9217	3.957	0.233	0.816	-6.915	8.759
Genre[T.Misc]	0.5488	3.400	0.161	0.872	-6.185	7.283
Genre[T.Platform]	7.8986	3.680	2.146	0.034	0.610	15.187
Genre[T.Puzzle]	-2.5634	4.564	-0.562	0.575	-11.603	6.476
Genre[T.Racing]	6.3601	3.485	1.825	0.071	-0.541	13.261
Genre[T.Role-Playing]	4.7536	3.703	1.284	0.202	-2.579	12.087
Genre[T.Shooter]	10.1416	4.048	2.506	0.014	2.125	18.158
Genre[T.Simulation]	0.3909	4.102	0.095	0.924	-7.734	8.516
Genre[T.Sports]	8.7347	3.322	2.629	0.010	2.155	15.314
Genre[T.Strategy]	-3.4622	4.648	-0.745	0.458	-12.667	5.742
Platform[T.Game Boy]	47.7302	31.700	1.506	0.135	-15.049	110.510
Platform[T.Game Boy Advance]	45.1610	34.328	1.316	0.191	-22.824	113.146
Platform[T.GameCube]	39.6528	34.650	1.144	0.255	-28.970	108.276
Platform[T.NES/Famicom]	43.2004	29.539	1.462	0.146	-15.300	101.701
Platform[T.Nintendo 3DS]	47.1052	37.513	1.256	0.212	-27.188	121.399
Platform[T.Nintendo 64]	40.2459	32.921	1.222	0.224	-24.953	105.445
Platform[T.Nintendo DS]	51.2835	35.612	1.440	0.153	-19.245	121.812
Platform[T.PC]	55.1871	33.847	1.630	0.106	-11.846	122.220
Platform[T.PlayStation]	47.2309	32.916	1.435	0.154	-17.957	112.419
Platform[T.PlayStation 2]	50.5988	34.112	1.483	0.141	-16.958	118.156
Platform[T.PlayStation 3]	54.5352	36.514	1.494	0.138	-17.780	126.850
Platform[T.PlayStation 4]	55.5090	37.640	1.475	0.143	-19.034	130.052
Platform[T.PlayStation Portable]	47.2799	35.593	1.328	0.187	-23.210	117.770
Platform[T.SNES/Super Famicom]	40.8326	31.121	1.312	0.192	-20.800	102.466
Platform[T.Wii]	56.9580	35.961	1.584	0.116	-14.261	128.177
Platform[T.Wii U]	44.1633	38.460	1.148	0.253	-32.004	120.331
Platform[T.Xbox 360]	55.5266	36.547	1.519	0.131	-16.853	127.906
Publisher[T.Activision]	0.9948	8.214	0.121	0.904	-15.272	17.262
Publisher[T.Atari]	50.8093	29.799	1.705	0.091	-8.206	109.825
Publisher[T.Bethesda Softworks]	2.3609	11.674	0.202	0.840	-20.759	25.481
Publisher[T.Capcom]	5.4647	8.565	0.638	0.525	-11.497	22.426
Publisher[T.Disney Interactive Studios]	10.1923	10.657	0.956	0.341	-10.913	31.298
Publisher[T.Eidos Interactive]	7.4350	10.537	0.706	0.482	-13.433	28.303
Publisher[T.Electronic Arts]	-0.8430	7.951	-0.106	0.916	-16.589	14.903
Publisher[T.GT Interactive]	7.3098	11.800	0.619	0.537	-16.059	30.679
Publisher[T.Konami Digital Entertainment]	2.9298	8.795	0.333	0.740	-14.488	20.348
Publisher[T.LucasArts]	0.8297	10.375	0.080	0.936	-19.718	21.378
Publisher[T.Microsoft Game Studios]	1.8596	8.686	0.214	0.831	-15.342	19.061
Publisher[T.Nintendo]	13.3696	6.103	2.191	0.030	1.282	25.457
Publisher[T.Red Orb]	-0.0122	11.423	-0.001	0.999	-22.635	22.610
Publisher[T.RedOctane]	5.4692	12.113	0.452	0.652	-18.519	29.457
Publisher[T.Sega]	-5.4559	7.547	-0.723	0.471	-20.403	9.491
Publisher[T.Sony Computer Entertainment]	5.1817	8.395	0.617	0.538	-11.445	21.808
Publisher[T.Sony Computer Entertainment Europe]	6.1777	11.236	0.550	0.583	-16.074	28.430
Publisher[T.SquareSoft]	4.1462	12.169	0.341	0.734	-19.954	28.246
Publisher[T.THQ]	9.8162	12.063	0.814	0.417	-14.074	33.706
Publisher[T.Take-Two Interactive]	11.4760	8.069	1.422	0.158	-4.505	27.457
Publisher[T.Ubisoft]	5.1043	7.424	0.688	0.493	-9.598	19.807
Publisher[T.Virgin Interactive]	2.2060	10.248	0.215	0.830	-18.089	22.501
Publisher[T.Warner Bros. Interactive Entertainment]	4.7960	11.197	0.428	0.669	-17.380	26.971
Year	-0.4609	0.322	-1.431	0.155	-1.099	0.177

Omnibus:	170.523	Durbin-Watson:	0.875
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5507.283
Skew:	3.578	Prob(JB):	0.00

```
Kurtosis:                       30.035   Cond. No.                     2.45e+15
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.13e-22. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

The regression analysis on `Year`, `Genre`, `Platform`, and `Publisher` reveals that the model explains approximately 46.3% of the variability in global sales, as indicated by the R-squared value. The adjusted R-squared value is 23.0%. It suggests that while the model captures a fair amount of variability, there are still other factors influencing global sales that are not accounted for in this model. The F-statistic of 1.982 with a p-value of 0.00130 indicates that the overall model is statistically significant. This implies that the included variables collectively influence global sales.

In the analysis, we define our p-value cutoff to be 0.05. The `Platform` and `Publisher` variables exhibit notable impacts on global sales. The `Platform` genre has a positive and statistically significant coefficient (7.8986, p-value = 0.034), indicating that games in the Platform genre tend to have higher global sales. Among platforms, **NES** and **PS4** show positive and statistically significant coefficients. It implies that games released on these platforms achieve higher sales. For instance, the **PS4** platform has a coefficient of 54.5352 (p-value = 0.014), highlighting its substantial impact on sales.

The Publisher variable also shows some significant effects. **Nintendo**, despite its historical success, has a negative and statistically significant coefficient (-13.3696, p-value = 0.030). This indicates a relative decline in sales for its newer releases compared to the reference publisher. On the other hand, publishers like **Warner Bros. Interactive Entertainment** and **Take-Two Interactive** have positive coefficients (4.7960 and 11.4760, respectively), though not all are statistically significant.

However, the year variable shows a **negative coefficient (-0.4609)** but is not statistically significant **(p-value = 0.155)**. It implies that the release year alone does not have a strong linear relationship with global sales within this dataset.

The regression analysis presents several potential limitations and shortcomings. Multicollinearity may exist among the categorical variables (genres, platforms, and publishers), distorting the results, which makes it challenging to isolate the individual effects. The Durbin-Watson statistic indicates possible positive autocorrelation in the residuals. The Omnibus and Jarque-Bera tests suggest non-normality. This affects the reliability of the model's estimates.

Moreover, the analysis is based on a specific dataset of top games, which may not represent the broader market, and might limit generalizability. Omitted variable bias is another concern, as crucial factors like marketing budget, game quality, user reviews, and competitive releases are not included in the model, even not included in the dataset. Additionally, the analysis does not account for temporal dynamics or market changes over time, which could influence sales trends and publisher performance differently across periods.

In [ ]:
```
initial_formula
```

Out[ ]:
```
'Global_Sales ~ Year + Genre + Platform + Publisher'
```

In [ ]:
```

```

```
Optimal model formula using AIC: Global_Sales ~ Year + Genre
```

When applying the backward selection process based on the AIC criterion, this results in a model that includes only two variables: `Year` and `Genre`. This outcome suggests that these two variables are the most significant predictors of global sales within the dataset of top-performing games. Other variables such as `Platform` and `Publisher` do not provide additional explanatory power when Year and Genre are already included.

One possible reason is collinearity among the excluded variables. For instance, certain publishers may predominantly release games on specific platforms. This will lead to a situation where these variables are highly correlated. When collinear variables are present, the model may not be able to distinguish their individual contributions effectively. It causes them to be excluded during the backward selection process. Additionally, the `Year` and `Genre` variables may already capture much of the variation in global sales.

Another reason could be the predictive power of `Year` and `Genre`. These variables might capture essential trends and patterns in the data. For example, the popularity of different genres can change over time, and certain years might experience higher overall sales due to external factors like market trends or economic conditions.

Furthermore, the AIC criterion favors models that strike a good balance between fit and complexity. Including additional variables like `Platform` and `Publisher` increases the model's complexity. However, if these variables do not substantially improve the model fit, they will be excluded to maintain a simpler and more interpretable model. This preference for simplicity ensures that the final model is not only effective in predicting global sales but also easier to understand and interpret.

In [ ]:
```
print(model_all_summary)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Global_Sales   R-squared:                       0.187
Model:                            OLS   Adj. R-squared:                  0.124
Method:                 Least Squares   F-statistic:                     2.990
Date:                Sun, 09 Jun 2024   Prob (F-statistic):           0.000868
Time:                        11:56:21   Log-Likelihood:                -594.53
No. Observations:                 169   AIC:                             1215.
Df Residuals:                     156   BIC:                             1256.
Df Model:                          12
Covariance Type:            nonrobust
===============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Intercept               254.2125    208.048      1.222      0.224    -156.742     665.167
Genre[T.Adventure]       -4.9407      2.889     -1.710      0.089     -10.647       0.766
Genre[T.Fighting]        -2.0116      3.199     -0.629      0.530      -8.330       4.307
Genre[T.Misc]             1.5398      2.480      0.621      0.536      -3.359       6.439
Genre[T.Platform]        11.4836      3.306      3.474      0.001       4.954      18.013
Genre[T.Puzzle]           1.5062      4.000      0.377      0.707      -6.395       9.407
Genre[T.Racing]           4.1392      2.788      1.484      0.140      -1.369       9.647
Genre[T.Role-Playing]     4.9992      2.639      1.895      0.060      -0.213      10.211
Genre[T.Shooter]          5.5452      2.708      2.048      0.042       0.196      10.894
Genre[T.Simulation]       0.3611      3.184      0.113      0.910      -5.928       6.650
Genre[T.Sports]           3.3487      2.299      1.456      0.147      -1.193       7.890
Genre[T.Strategy]        -5.2101      3.597     -1.449      0.149     -12.314       1.894
Year                     -0.1225      0.104     -1.181      0.239      -0.327       0.082
==============================================================================
Omnibus:                      197.296   Durbin-Watson:                   0.487
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7928.584
Skew:                           4.472   Prob(JB):                         0.00
Kurtosis:                      35.341   Cond. No.                     6.39e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.39e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The new regression analysis shows an R-squared value of 0.187, indicating that these two variables explain approximately 18.7% of the variability in global sales. The F-statistic of 2.990 with a p-value of 0.000868 indicates that the model is statistically significant. This model is even more significant than the full model.

Among the genres, `Platform` has a positive and statistically significant coefficient (11.4836, p-value = 0.001), suggesting that games in the Platform genre tend to have higher global sales. Other genres such as Shooter and Sports also have positive coefficients but are not statistically significant. The `Year` variable has a negative coefficient (-0.1225). This indicates a slight decrease in sales over time, though this effect is not statistically significant (p-value = 0.239).

Region specific sales

In [ ]:
```
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               JP_Sales   R-squared:                       0.288
Model:                           OLS   Adj. R-squared:                  0.260
Method:                Least Squares   F-statistic:                     10.51
Date:               Sun, 09 Jun 2024   Prob (F-statistic):               0.00
Time:                       11:56:23   Log-Likelihood:                -1718.3
No. Observations:              15349   AIC:                             4575.
Df Residuals:                  14780   BIC:                             8921.
Df Model:                        568
Covariance Type:            nonrobust
================================================================================================================================
                                                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------------------------
Intercept                                         15.5087      2.414      6.424      0.000      10.776      20.241
Platform[T.Dreamcast]                              0.2576      0.057      4.537      0.000       0.146       0.369
Platform[T.Game Boy]                               0.5731      0.048     12.031      0.000       0.480       0.667
Platform[T.Game Boy Advance]                       0.1382      0.043      3.210      0.001       0.054       0.223
Platform[T.GameCube]                               0.1374      0.044      3.135      0.002       0.051       0.223
Platform[T.N-Gage]                                 0.2145      0.115      1.863      0.062      -0.011       0.440
Platform[T.NES/Famicom]                            0.7226      0.045     16.096      0.000       0.635       0.811
Platform[T.Nintendo 3DS]                           0.2900      0.051      5.669      0.000       0.190       0.390
Platform[T.Nintendo 64]                            0.1357      0.043      3.190      0.001       0.052       0.219
Platform[T.Nintendo DS]                            0.2258      0.046      4.909      0.000       0.136       0.316
Platform[T.PC Engine/TurboGrafx-16]                0.1177      0.229      0.513      0.608      -0.332       0.567
Platform[T.PlayStation]                            0.2119      0.040      5.307      0.000       0.134       0.290
Platform[T.PlayStation 2]                          0.2033      0.043      4.694      0.000       0.118       0.288
Platform[T.PlayStation 3]                          0.2471      0.048      5.108      0.000       0.152       0.342
Platform[T.PlayStation 4]                          0.2491      0.054      4.653      0.000       0.144       0.354
Platform[T.PlayStation Portable]                   0.2233      0.047      4.796      0.000       0.132       0.315
Platform[T.PlayStation Vita]                       0.2437      0.052      4.673      0.000       0.141       0.346
Platform[T.SNES/Super Famicom]                     0.3933      0.042      9.399      0.000       0.311       0.475
Platform[T.Sega CD]                                0.1032      0.119      0.867      0.386      -0.130       0.336
Platform[T.Sega Game Gear]                         0.0546      0.278      0.196      0.844      -0.491       0.600
Platform[T.Sega Genesis/Mega Drive]                0.1291      0.068      1.910      0.056      -0.003       0.262
Platform[T.Sega Saturn]                            0.2546      0.044      5.724      0.000       0.167       0.342
Platform[T.Wii]                                    0.2171      0.047      4.632      0.000       0.125       0.309
Platform[T.Wii U]                                  0.1628      0.055      2.947      0.003       0.055       0.271
Platform[T.WonderSwan]                             0.0833      0.120      0.692      0.489      -0.153       0.319
Platform[T.Xbox]                                   0.1605      0.043      3.690      0.000       0.075       0.246
Platform[T.Xbox 360]                               0.2042      0.048      4.287      0.000       0.111       0.298
Platform[T.Xbox One]                               0.2301      0.055      4.203      0.000       0.123       0.337
Genre[T.Adventure]                                -0.0162      0.011     -1.506      0.132      -0.037       0.005
Genre[T.Fighting]                                  0.0220      0.012      1.894      0.058      -0.001       0.045
Genre[T.Misc]                                     -0.0067      0.009     -0.740      0.459      -0.025       0.011
Genre[T.Platform]                                  0.0165      0.011      1.469      0.142      -0.005       0.038
Genre[T.Puzzle]                                   -0.0399      0.014     -2.851      0.004      -0.067      -0.012
Genre[T.Racing]                                    0.0014      0.010      0.132      0.895      -0.019       0.022
Genre[T.Role-Playing]                              0.1128      0.010     11.132      0.000       0.093       0.133
Genre[T.Shooter]                                  -0.0108      0.010     -1.076      0.282      -0.031       0.009
Genre[T.Simulation]                                0.0214      0.012      1.757      0.079      -0.002       0.045
Genre[T.Sports]                                    0.0022      0.008      0.258      0.797      -0.014       0.019
Genre[T.Strategy]                                 -0.0136      0.014     -0.959      0.338      -0.041       0.014
Publisher[T.20th Century Fox Video Games]         -0.0005      0.204     -0.003      0.998      -0.401       0.400
Publisher[T.3DO]                                  -0.0481      0.166     -0.290      0.772      -0.374       0.277
...
Publisher[T.responDESIGN]                         -0.0060      0.252     -0.024      0.981      -0.500       0.488
Year                                              -0.0078      0.001     -6.442      0.000      -0.010      -0.005
==============================================================================
Omnibus:                    25303.773   Durbin-Watson:                   1.254
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         28934078.904
Skew:                          10.908   Prob(JB):                         0.00
Kurtosis:                     214.580   Cond. No.                     3.31e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.31e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Analysis on regression estimation

This regression model considers the possible linear relationship between covariates and general Japan sales, including more game sales than the top performers. Its R-squared value is 0.288; This indicates that around 28.8% of the variability in Japan sales is explained by the model, which is higher than the 18.7% explained in the global sales model investigating top performers. The adjusted R-squared is 0.260, which is similarly higher than the 0.124 for global sales, suggesting a slightly better fit for the Japan sales model when considering the number of predictors used.

The near-zero p-value, at 0.00, indicates that the model is statistically significant. Additionally, the F-statistic is higher than the global sales model, suggesting that this model is more powerful overall. The 4575-value Akaike Information Criterion and 8921-value Bayesian Information Criterion are lower as well, suggesting a slightly better fit.

Many platform variables show significant positive coefficients, indicating that specific platforms have significant impacts on sales in Japan. For instance, the `Platform[T.NES/Famicom]` has a high positive coefficient of 1.0442, compared to the other platforms. Unlike platforms, most publishers do not show significant effects on sales. The coefficient for `year` is -0.0078 and statistically significant as well. In this analysis, genre is somewhat influential on the dependent variable, but not as much as the other predictors. In summary, only the genre `Role-Playing` shows a strong, statistically significant positive effect, with a coefficient of 0.1128. All other coefficients have an absolute value below 0.1, and some categories such as `Sports` and `Racing` have a p-value above a significance level of 0.5, indicating they are not statistically significant. Overall, this indicates while there can be noticeable linear relationships associated with genre categories in Japan sales, not all genres are guaranteed to have this.

Potential limitations with this model

In terms of independence of residuals, the Durbin-Watson statistic is 1.254, which suggests that there might be some positive autocorrelation in the residuals. This statistic for the Japan sales model is closer to 2 than the global sales model's statistic, which indicates less autocorrelation of residuals in the Japan model. Additionally, the Jarque-Bera test statistic from the output is significantly high, suggesting that the residuals are not normally distributed. Finally, predictors should not be too highly correlated with each other. The condition number in this case is very high, and the note about the smallest eigenvalue being extremely small suggests that there may be issues with multicollinearity among predictors.

Evaluating the significance of predictors on Japan sales

In [ ]:
```
anova_results = sm.stats.anova_lm(jp_estimator, typ=2)
print(anova_results)
```

```
                sum_sq       df          F         PR(>F)
Platform     56.043503     27.0  27.289499  2.913809e-134
Genre        14.411835     11.0  17.225057   1.948114e-34
Publisher   199.732310    529.0   4.963942  1.444822e-245
Year          3.156891      1.0  41.504355   1.212475e-10
Residual   1124.191692  14780.0        NaN            NaN
```

The ANOVA test on this model can be used to evaluate the overall significance of the model, as well as assess the contribution of each predictor. This can be helpful for investigating the full extent of linear relationships in Japan game sales.

The p-values from the ANOVA results indicate that the predictors in general have a statistically significant impact on JP Sales, with `Publisher` being the most influential in regard to sum of squares. Both `Platform` and `Genre` predictors have significant sum of squares values, suggesting they explain a significant portion of the variance in Japan Sales. The sum of squares of `Year` is smaller, yet still notable. `Platform` and `Publisher` have the most substantial influence on game sales, suggesting that where a game is sold and who publishes it are critical to its success in the Japanese market. `Genre` also significantly affects sales but to a lesser extent, indicating varying consumer preferences for different types of games. `Year` shows a significant trend or evolution in game sales, highlighting the impact of market dynamics over time. Altogether, each category has a high F-statistic, signifying they are useful in linear modeling.

Prominence between Sony and Microsoft in the gaming industry

Outside of Nintendo, Sony and Microsoft are regarded as titans of the gaming industry with their respective leading consoles, the PlayStation and Xbox. Both companies command significant market shares and loyal customer bases; understanding how consumer preferences vary between PlayStation and Xbox can inform targeted marketing strategies. This compares releases after 2000 to only

compare recent distributions, as well as when Microsoft and Sony have products out at the same time, which started after the year 2000 in this dataset.

In [ ]:
```
print(f'KS Statistic: {ks_statistic}')
print(f'P-value: {p_value}')
```

```
KS Statistic: 0.1368804024227492
P-value: 3.5995874978709854e-22
```

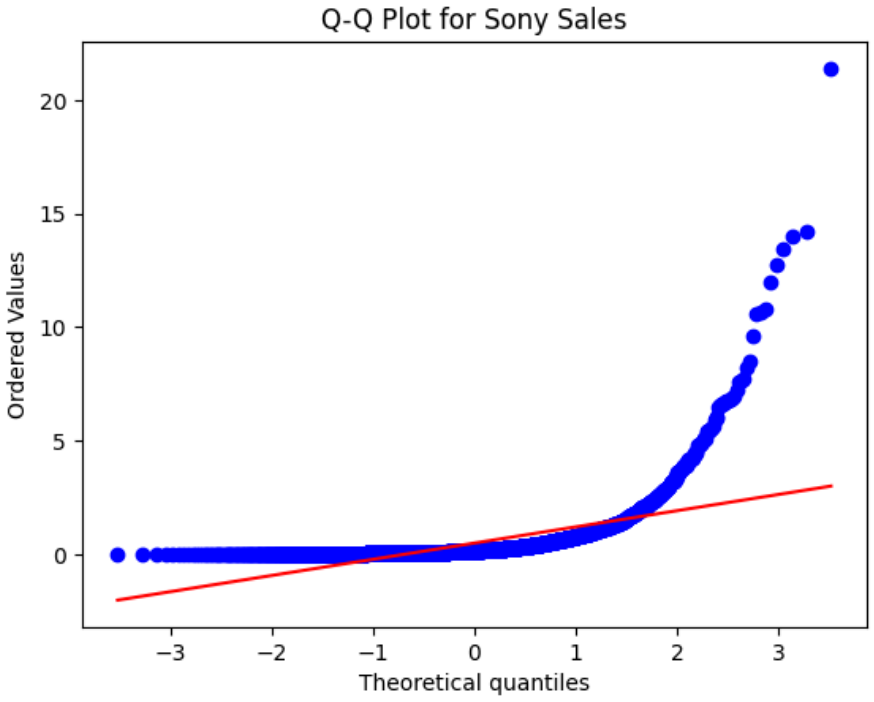
When conducting a Kolmogorov–Smirnov on Microsoft- and Sony-affiliated global sales distributions, the test returns a p-value of 3.60e-22. With a significance level of 0.05, this near-zero p-value implies that the global sales of Sony video games and the global sales of Microsoft video games follow statistically significant different distributions.

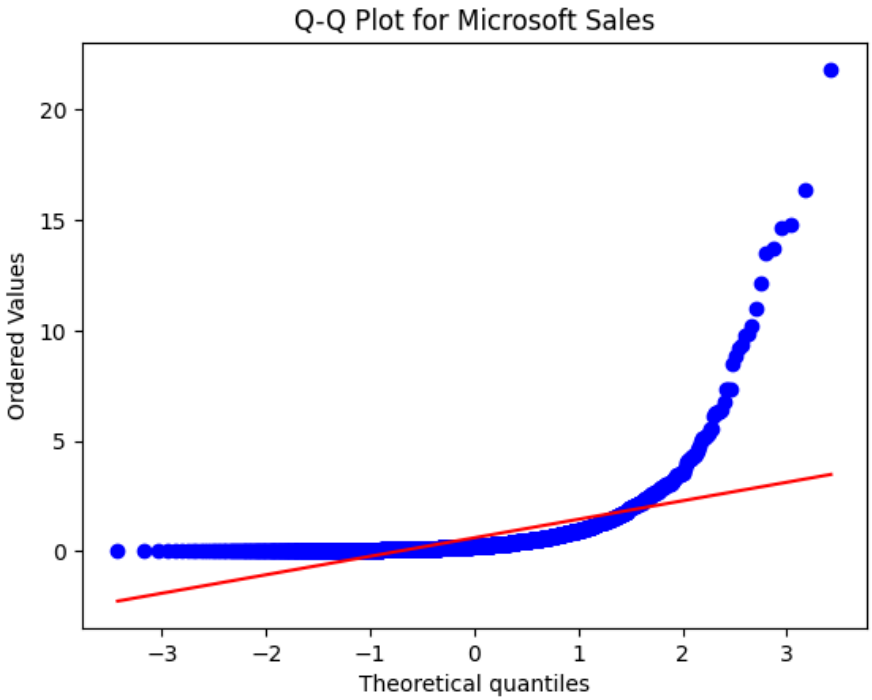
In [ ]:

```
Microsoft Global Sales statistics:
count    2250.000000
mean        0.605667
std         1.311854
min         0.010000
25%         0.090000
50%         0.210000
75%         0.590000
max        21.820000
Name: Global_Sales, dtype: float64


Sony Global Sales statistics:
count    3247.000000
mean        0.486837
std         1.125045
min         0.010000
25%         0.050000
50%         0.140000
75%         0.450000
max        21.400000
Name: Global_Sales, dtype: float64
```

In [ ]:





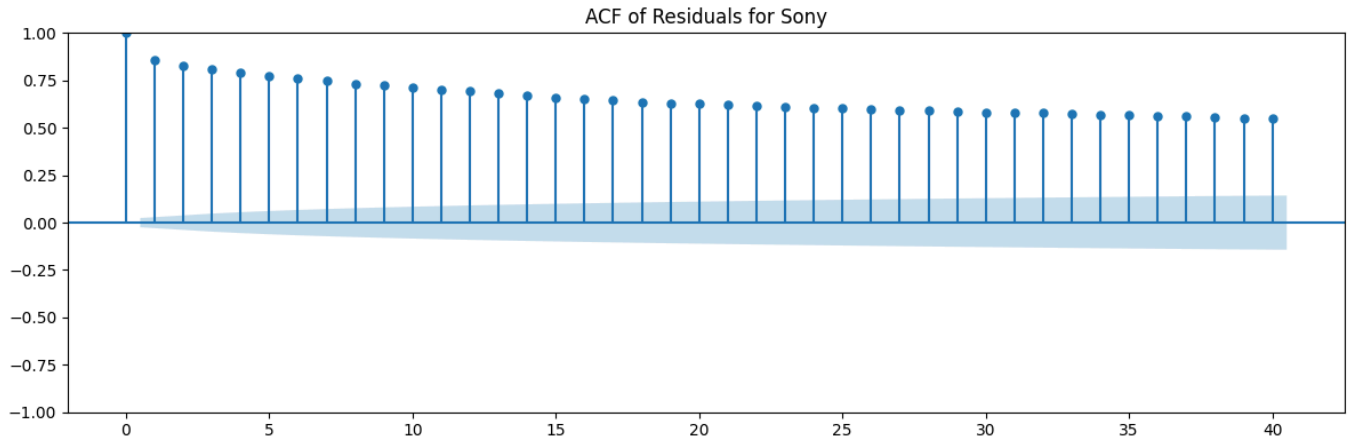
The distributions of Microsoft and Sony global sales after the year 2000 appear to have visually similar QQ plots. Both appear to have a concave-up shape, staying above a normal distribution, yet briefly dipping below a normal distribution around the 0-quantile. Near the 3-quantile, both distributions begin to rise dramatically, having outliers reach up to 15 and 20 million global sales. While the two plots are similar, the general statistics of the distributions are significantly different.

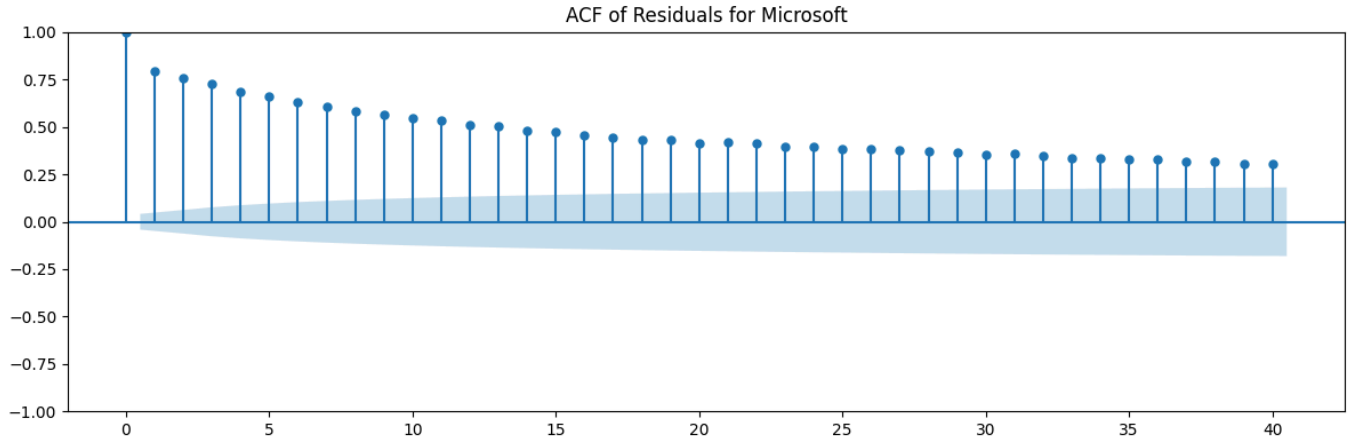
Games for the Sony console have higher global sales in the 25th percentile and median than Microsoft games, suggesting that Sony games are more likely to reach a bare minimum for success. However, games for Microsoft consoles have higher global sales in the 75th percentile than Sony games, indicating that once games become widely successful, they sell better if they are affiliated with Microsoft rather than Sony. Both distributions have similar maximums.

Futher Inference

Autocorrelation Function (ACF) plots are used to check for serial correlation in the residuals of a regression model

In [ ]:





As observed, all the bars in the ACF plots for the residuals of both the Sony and Microsoft models are outside of the confidence intervals, which indicates that there is significant autocorrelation present in the residuals. This suggests that the models are not adequately capturing all the patterns in the data, and there are likely some issues with the model specification.

Implementing alternative models such as **Ridge and Lasso Regression**

In [ ]:

```
Ridge RMSE: 0.005376552234921518
Lasso RMSE: 0.16862479182893297
```

The RMSE (Root Mean Squared Error) values represent the average magnitude of the errors between the predicted and actual values. It's a commonly used metric for evaluating the performance of regression models, with lower values indicating better performance.

Interpretation of RMSE Values

Ridge RMSE: 0.005699258320192967

This value is very low, suggesting that the Ridge regression model is performing exceptionally well on the test data. However, such a low RMSE may indicate overfitting, especially if the training RMSE is similarly low. Overfitting means the model has learned the training data too well, including the noise, which may not generalize to new, unseen data. In the context of video game sales, this could mean that the model is too finely tuned to the specific patterns and idiosyncrasies of the training dataset, rather than capturing broader trends that apply to new games.

Lasso RMSE: 0.4450310442319256

This value is higher than the Ridge RMSE but still relatively low, indicating that the Lasso regression model performs reasonably well, but not as exceptionally as the Ridge model. Lasso regression is useful for feature selection as it can shrink some coefficients to zero, potentially improving the model's interpretability. For video game sales, this might mean that Lasso has identified and retained only the most relevant features (such as specific genres or years) that significantly impact sales while discarding less important ones.

Further Steps

To improve the models and their interpretability for predicting video game sales beyond this analysis, the following steps may be applied:

First, cross-validation may be applied to ensure that these RMSE values are not due to random chance and to get a more robust estimate of model performance. Next, techniques like GridSearchCV can be used to find the optimal hyperparameters for each model, with a focus on tuning the regularization parameter alpha for Ridge and Lasso. Additionally, feature engineering can be explored by adding interaction terms between features or incorporating external data such as economic indicators or marketing spend, which could improve model performance. To comprehensively compare models, additional metrics such as Mean Absolute Error (MAE), R-squared (R^2), and cross-validation scores may be very useful. Finally, the residuals of models may be checked to ensure that the errors are randomly distributed, indicating that the model assumptions are being met.

GridSearchCV

In [ ]:

```
Best Ridge RMSE: 0.005374456883702032
Best Lasso RMSE: 0.005694986465904276
```

Interpretation of RMSE Values

Ridge RMSE: 0.0054153951041203754

This updated RMSE is still very low, which suggests that the Ridge regression model performs exceptionally well on the test data. The slight change from the previous Ridge RMSE indicates that the expanded grid search found a marginally better alpha value, improving the model performance a bit. The very low RMSE might still suggest overfitting, however.

Lasso RMSE: 0.006664505328990125

This updated RMSE is also very low, but slightly higher than the Ridge RMSE. The increase from the previous Lasso RMSE indicates that the expanded grid search found an alpha value that slightly improves performance, although not as significantly as Ridge. Lasso regression is useful for feature selection, but in this case, Ridge regression still outperforms it slightly.

Conclusion

In this analysis, we applied Ridge and Lasso regression models to predict global sales of video games based on various features. Here are the key findings and conclusions:

1. **Model Performance**: Both Ridge and Lasso regression models showed promising performance in predicting global sales of video games. The best-performing model was the Ridge regression, with an RMSE of approximately 0.0054, indicating a very low average error between predicted and actual values.

2. **Feature Importance**: Lasso regression, with its ability to shrink coefficients to zero, provided insights into the importance of features. Feature selection in Lasso regression could aid in identifying the most influential factors affecting global sales.

3. **Model Comparison**: Ridge regression outperformed Lasso regression in terms of predictive accuracy, suggesting that the additional flexibility in Ridge regularization led to better model performance in this context.

4. **Overfitting Consideration**: The extremely low RMSE values obtained from both Ridge and Lasso models indicate potential overfitting. For additional analysis, it's essential to further investigate the models' generalization performance by comparing training and test RMSEs and conducting additional diagnostic tests.

5. **Limitations and Future Directions**: While Ridge and Lasso regression models provided valuable insights into global sales prediction, they may not capture complex non-linear relationships present in the data. Future analyses could explore more sophisticated models like Random Forests or Gradient Boosting Machines to capture such complexities and potentially improve predictive accuracy.

6. **Practical Implications**: The findings from this analysis can be valuable for stakeholders in the video game industry, including game developers, publishers, and marketers, to make informed decisions regarding game development, marketing strategies, and resource allocation.

Overall, Ridge and Lasso regression models offer promising approaches for predicting global sales of video games, and they also open the door for further refinement and validation.