1. **Understanding the dataset**
    a. Critic Score: The critic score represents the average rating or evaluation given by professional critics who specialize in reviewing movies, games, or other forms of entertainment.
    b. Critic Count: Number of Critics for the game you gave a score
    c. User Score: The user score, also referred to as the audience score or user rating, reflects the average rating or evaluation given by regular users.
    d. User Count: Number of users for the game who gave a score
    e. Platform: Game platforms such as PS3 PS2 X360 X360 PS2 Wii…
    f. Genre: Sports Platform Action Simulation Action Platform
    g. Publisher: Electronic Arts Sony Computer Entertainment, Electronic Arts…
    h. Developer: 'EA Tiburon, Sucker Punch, Pandemic Studios
    i. Ratings: E = Everyone, E10+ = Everyone 10+, T = Teen, M = Mature

2. **Loading and observing the data**
    a. Loading data into pandas dataframe: pandas allows to load *.sas7bdat files
    b. Observing data types and basic stats of numerical data

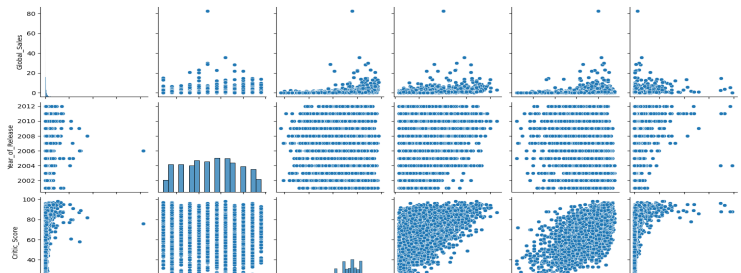
	Name	Platform	Genre	Publisher	Developer	Rating	Global_Sales	Year_of_Release	Critic_Score	Critic_Count	User_Score	User_Count
0	b'Madden NFL 11'	b'PS3'	b'Sports'	b'Electronic Arts'	b'EA Tiburon'	b'E'	2.38	2010.0	83.0	36.0	6.1	68.0
1	b'Sly Cooper and the Thievius Raccoonus'	b'PS2'	b'Platform'	b'Sony Computer Entertainment'	b'Sucker Punch'	b'E'	1.21	2002.0	86.0	41.0	8.6	184.0
2	b'The Lord of the Rings: Conquest'	b'X360'	b'Action'	b'Electronic Arts'	b'Pandemic Studios'	b'T'	0.63	2009.0	55.0	58.0	7.0	110.0
3	b'Red Steel 2'	b'Wii'	b'Shooter'	b'Ubisoft'	b'Ubisoft Paris'	b'T'	0.62	2010.0	80.0	73.0	8.6	178.0
4	b'Star Wars: The Force Unleashed II'	b'X360'	b'Action'	b'LucasArts'	b'LucasArts'	b'T'	1.43	2010.0	61.0	59.0	5.7	180.0

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4413 entries, 0 to 4412
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Name             4413 non-null   object
 1   Platform         4413 non-null   object
 2   Genre            4413 non-null   object
 3   Publisher        4413 non-null   object
 4   Developer        4413 non-null   object
 5   Rating           4413 non-null   object
 6   Global_Sales     4413 non-null   float64
 7   Year_of_Release  4413 non-null   float64
 8   Critic_Score     4413 non-null   float64
 9   Critic_Count     4413 non-null   float64
 10  User_Score       4413 non-null   float64
 11  User_Count       4413 non-null   float64
dtypes: float64(6), object(6)
memory usage: 413.8+ KB
```

	Global_Sales	Year_of_Release	Critic_Score	Critic_Count	User_Score	User_Count
count	4413.000000	4413.000000	4413.000000	4413.000000	4413.000000	4413.000000
mean	0.774985	2006.453433	69.725584	29.060050	7.244528	136.887605
std	2.141282	3.056571	14.010007	18.383254	1.419852	520.006401
min	0.010000	2001.000000	17.000000	4.000000	0.500000	4.000000
25%	0.110000	2004.000000	61.000000	15.000000	6.600000	10.000000
50%	0.290000	2007.000000	72.000000	25.000000	7.600000	22.000000
75%	0.740000	2009.000000	80.000000	39.000000	8.300000	62.000000
max	82.530000	2012.000000	98.000000	107.000000	9.500000	9851.000000

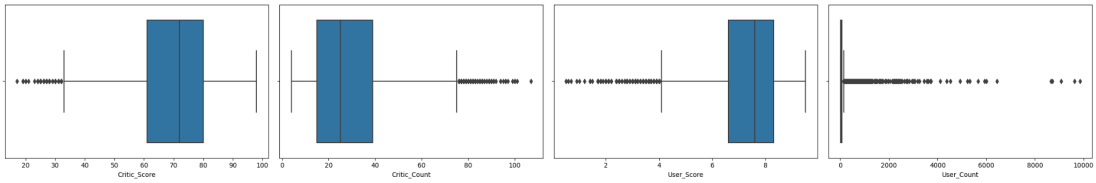
There are 6 categorical independent variables and 5 numerical independent variables. Global sales is the dependent variable.

3. **Exploratory Data Analysis**



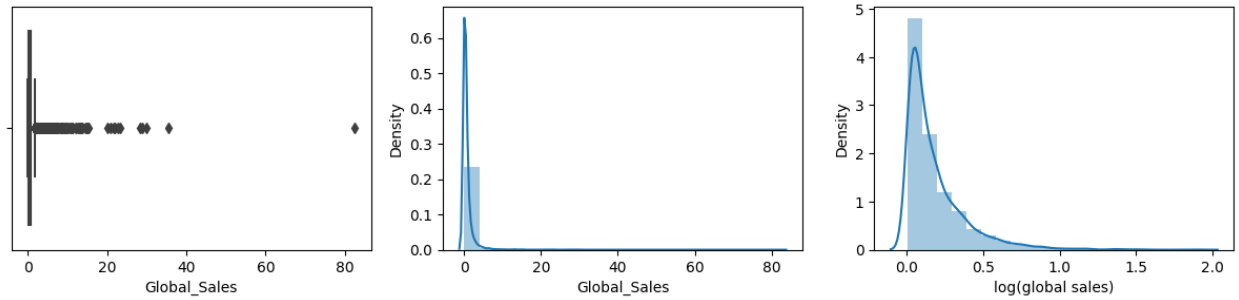
Pair plot (observing for outliers, skewness)This plots allows us to visualize the pairwise relationships between multiple variables in a dataset. It looks that the dependent variable is data is highly skewed.This makes it hard to predict the global sales value using linear regression.

Distribution of User counts, User score, Critic counts and Critic Score.



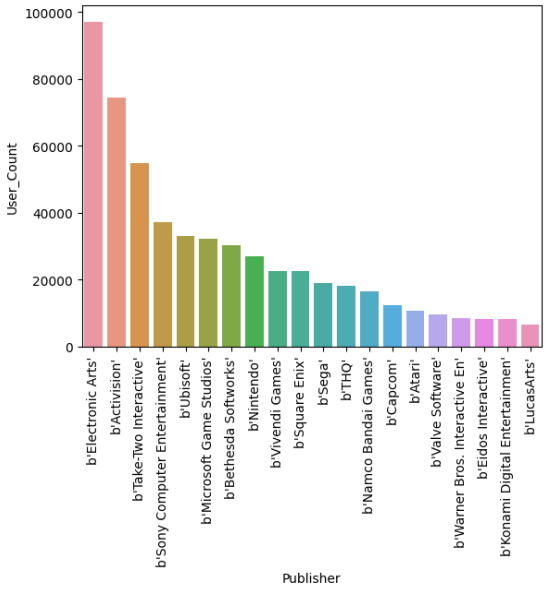
Clearly there are some outliers in this data. Outlier treatment is required to improve model performance.

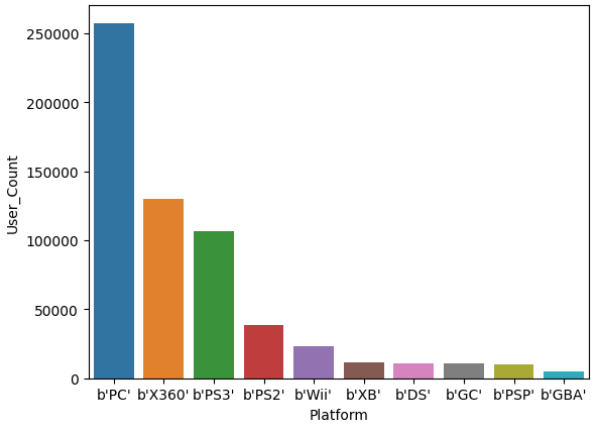
Check for skewness in the dependent variable and transform the values if necessary

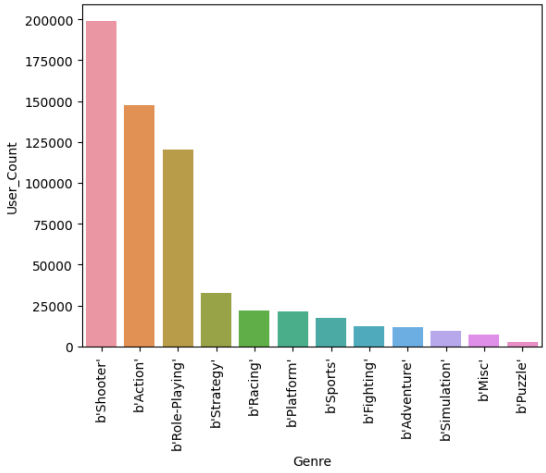


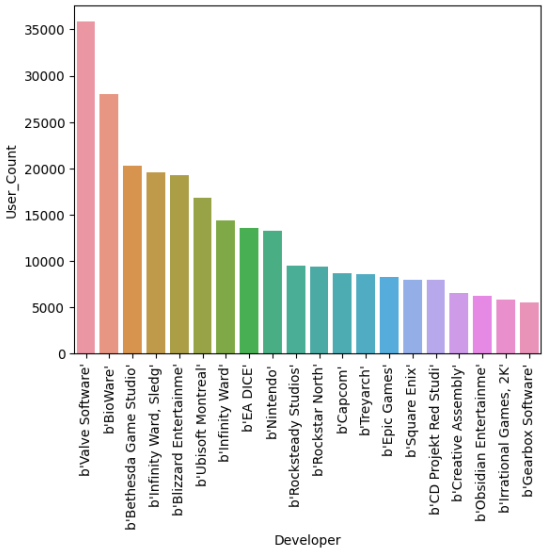
The plot above shows the distribution of global sales values. Removing outliers and Log transformation of the global sales values can help increase model performance.

Exploring Sum of user counts across different categories



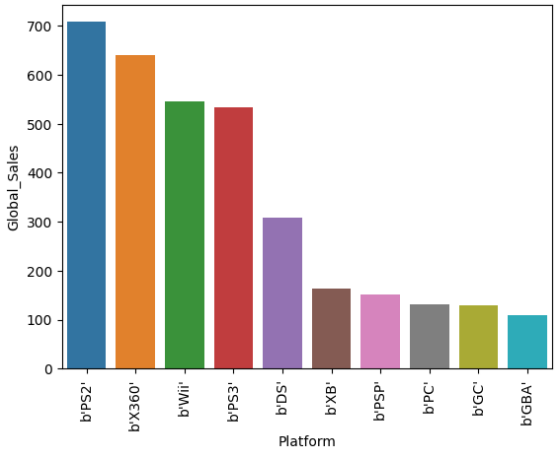


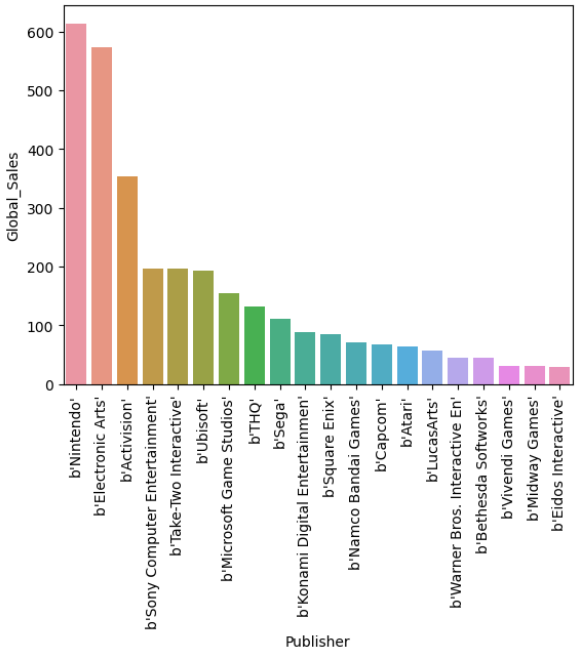


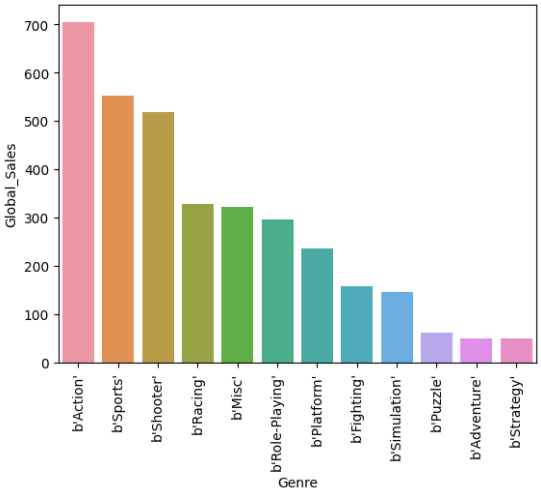


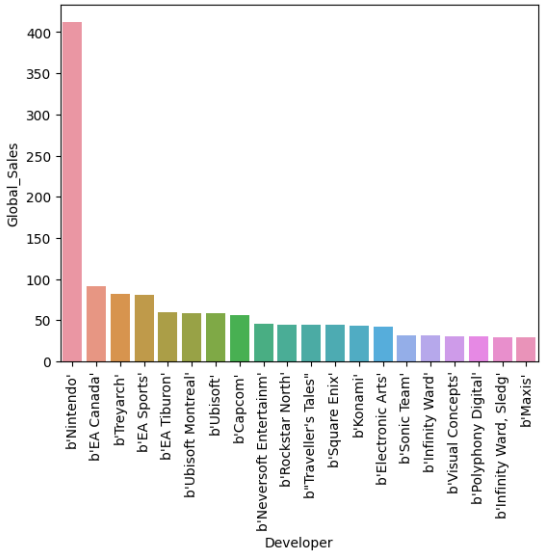
User counts are more than critic counts. The hypothesis here is that users contribute to the global sales,and hence it's worth looking into how the sum of user counts is distributed across various categorical variables. For these reasons splitting the categorical variables into one hot encoded vector based on user counts could be useful for predictions. More details are in the data preparation section

Checking the distribution of sales values for categorical values









It is also worth looking into how the sales are distributed across various categories. For the above plots some of the categories have large sales values. More details are in the data preparation section

4. Data preparation

a. Dropping unnecessary variables

b. Converting categorical variables to one hot encoded vectors

While exploring user counts and global sales across various categorical variables, some of the categories had significantly more sum of user counts and global sales. For these reasons, the higher sum values in each category were chosen to be a part of the dataset as one hot encoded vector and the rest as other.

The following values from each categorical variable are chosen.

- Publishers: EA, Activision, Take two Interactive, Nintendo, Other(s)
- Genre: Shooter, Action, Roleplay, Sports, other(s)
- Platform: PC, XBox, PS, Wii, other(s)
- Developer: Nintendo
- Ratings: All categories included

Since Year of Release are all distinct categories, They are probably not a good indicator as these values may not appear in the future. This is also an assumption as critic scores and user scores from previous years could have an effect on global sales. For simplicity, Year of release is dropped. Dropping Names of games as well, since we cannot do reßression on text data.

5. Model building results and interpretation

Stats models provide OLS (Ordinary Least Squares linear regression model.) The summary function provides the regression results. The data frame is split, where in X is the dependent variable y is the dependent variable.

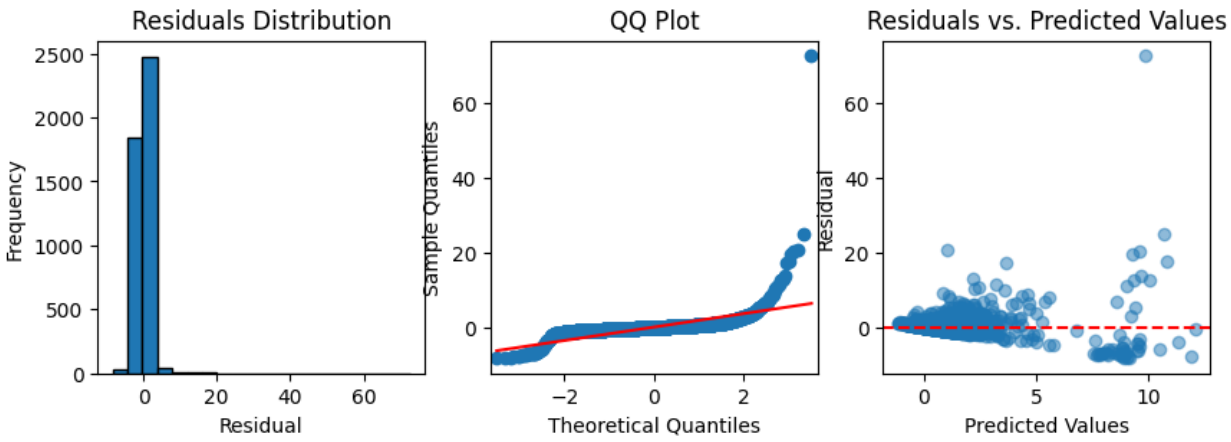
OLS Regression Results

Dep. Variable:	Global_Sales	R-squared:	0.291
Model:	OLS	Adj. R-squared:	0.288
Method:	Least Squares	F-statistic:	99.96
Date:	Tue, 27 Jun 2023	Prob (F-statistic):	2.36e-310
Time:	06:42:31	Log-Likelihood:	-8864.1
No. Observations:	4413	AIC:	1.777e+04
Df Residuals:	4394	BIC:	1.789e+04
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Critic_Score	0.0148	0.003	5.068	0.000	0.009	0.021
Critic_Count	0.0222	0.002	10.841	0.000	0.018	0.026
User_Score	-0.0758	0.025	-3.005	0.003	-0.125	-0.026
User_Count	0.0008	6.04e-05	13.982	0.000	0.001	0.001
E10+	-0.2326	0.096	-2.424	0.015	-0.421	-0.044
M	-0.3285	0.102	-3.228	0.001	-0.528	-0.129
T	-0.2931	0.077	-3.824	0.000	-0.443	-0.143
Developer_b'Nintendo'	4.9664	0.227	21.858	0.000	4.521	5.412
Developer_other	-2.1410	0.125	-17.158	0.000	-2.386	-1.896
Publisher_b'Electronic Arts'	0.6028	0.093	6.476	0.000	0.420	0.785
Publisher_b'Nintendo'	1.0121	0.132	7.674	0.000	0.754	1.271
Publisher_b'Take-Two Interactive'	0.7079	0.135	5.247	0.000	0.443	0.972
Publisher_other	0.5025	0.076	6.604	0.000	0.353	0.652
Platform_b'PC'	-0.2332	0.097	-2.410	0.016	-0.423	-0.043
Platform_b'PS2'	0.8990	0.073	12.274	0.000	0.755	1.043
Platform_b'Wii'	1.2108	0.088	13.827	0.000	1.039	1.382
Platform_b'X360'	0.4856	0.080	6.065	0.000	0.329	0.643
Platform_other	0.4632	0.058	8.051	0.000	0.350	0.576
Genre_b'Action'	0.6933	0.071	9.796	0.000	0.555	0.832
Genre_b'Shooter'	0.7023	0.087	8.096	0.000	0.532	0.872
Genre_b'Sports'	0.7379	0.088	8.392	0.000	0.566	0.910
Genre_other	0.6918	0.063	11.005	0.000	0.569	0.815

Omnibus:	9215.155	Durbin-Watson:	1.993
Prob(Omnibus):	0.000	Jarque-Bera (JB):	71940000.982
Skew:	17.336	Prob(JB):	0.00
Kurtosis:	627.533	Cond. No.	5.73e+18

Interpreting regression results



R squared and Adjusted Rsquared: 0.291 is relatively low which means the model is not able to explain the variance in the data. The adjusted R^2 is a little lower than R^2, but not by much, it shows that there are no irrelevant independent variables that are not explaining the variance.

F statistic and F test: The high F value and a p value < 0.01 suggests that there is significant linear relationship between independent variables and the dependent variable and the model is statistically significant. There is no strong evidence to reject the null hypothesis that all the regression coefficients are equal to zero.

t-statistic: large t values indicate evidence against the null hypothesis of each individual independent variable, it measures the number of standard deviations the estimated coefficient is away from zero. In the table above only the categorical variable has a low t value of -2.424, and p value greater than 0.01. This shows that this variable coefficient is insignificant in predicting global sales. All the other variables have a p value less than 0.01 and hence they are significant in predicting global sales. However, all independent variables are significant at 5% alpha level.

coefficients and standard errors: It is hard to comprehend the standard errors as the data is not normalized. However, here are a few observations. The user ratings have a negative relationship with global sales. This seems very counter intuitive, as higher user rating should have a positive relationship with global sales. This needs more investigation, there could be interaction effects or confounding variables causing this. It is possible that global sales is determined by the previous years user ratings, which is not accounted for in this data. Developer - Nintendo has a high beta value of 4.324, with a high standard error, however this value is significant in predicting global sales.

Residuals: The deviation of points on the tails of the q-q plot suggests that this shows that the residuals do not follow a normal distribution. This is against the assumption of normality of distribution of the errors. There could be outliers in the dataset that need to be treated. The distribution of errors is also skewed. The funnel shape of the residuals on the residual vs predicted values shows the presence of heteroscedasticity. This shows that the variability of residuals does not systematically change as the values of independent variables change. This can be treated by transforming the independent and dependent variables.

In Conclusion, most of the dependent variables are significant, but they don't explain the variation in the independent (global sales) variable as the R^2 value is very low. This means there could be more features that could explain the variation. As observed from the plots there are outliers in the data that need to be treated. Clearly the assumption of homoscedasticity is violated and needs to be corrected using variable transformation. Adding interaction variables could help improve R^2.

The following steps are implemented to improve model performance

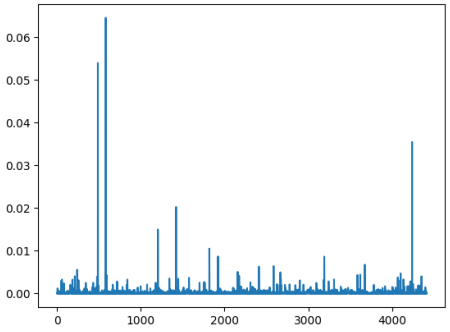
- Adding interaction variables transforming variables
- Checking and eliminating outliers in the data using Cook's D method
- Checking for multicollinearity
- Using RFE to eliminate features based on

Various Interaction terms were added to capture relationships between dependent variables and independent variables. After iteratively checking for scatter plots between global sales and interaction and independent variables, a **log - log model improved** the performance of the model to a R^2 value of 0.529.

OLS Regression Results

Dep. Variable:	Global_Sales	R-squared:	0.533
Model:	OLS	Adj. R-squared:	0.529
Method:	Least Squares	F-statistic:	134.9
Date:	Fri, 30 Jun 2023	Prob (F-statistic):	0.00
Time:	20:58:58	Log-Likelihood:	4026.6
No. Observations:	4413	AIC:	-7977.
Df Residuals:	4375	BIC:	-7734.
Df Model:	37		
Covariance Type:	nonrobust		

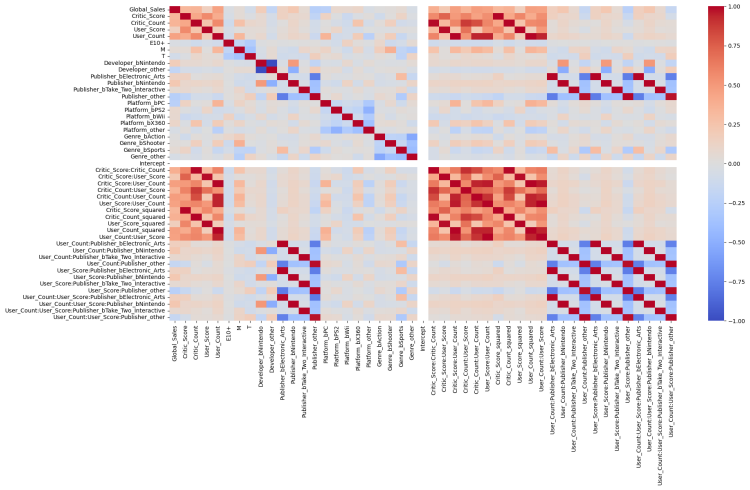
	coef	std err	t	P>|t|	[0.025	0.975]
Critic_Score	-35.0375	20.363	-1.721	0.085	-74.959	4.884
Critic_Count	-8.3680	3.704	-2.259	0.024	-15.629	-1.107
User_Score	9.8897	4.303	2.298	0.022	1.453	18.326
User_Count	-8.7553	2.612	-3.352	0.001	-13.876	-3.634
E10+	-0.0197	0.005	-3.790	0.000	-0.030	-0.010
M	-0.0671	0.006	-11.884	0.000	-0.078	-0.056
T	-0.0345	0.004	-8.222	0.000	-0.043	-0.026
Developer_bNintendo	3.1053	1.339	2.319	0.020	0.480	5.730
Developer_other	3.0194	1.339	2.254	0.024	0.394	5.645
Publisher_bElectronic_Arts	1.8157	1.538	1.181	0.238	-1.199	4.831
Publisher_bNintendo	-1.5799	3.318	-0.476	0.634	-8.085	4.925
Publisher_bTake_Two_Interactive	4.1384	2.003	2.066	0.039	0.211	8.066
Publisher_other	1.7506	1.315	1.331	0.183	-0.828	4.330
Platform_bPC	1.0338	0.536	1.929	0.054	-0.017	2.085
Platform_bPS2	1.2982	0.536	2.424	0.015	0.248	2.348
Platform_bWii	1.2981	0.536	2.423	0.015	0.248	2.348
Platform_bX360	1.2388	0.536	2.313	0.021	0.189	2.289
Platform_other	1.2559	0.535	2.346	0.019	0.206	2.305
Genre_bAction	1.5387	0.670	2.297	0.022	0.226	2.852
Genre_bShooter	1.5272	0.670	2.280	0.023	0.214	2.840
Genre_bSports	1.5311	0.669	2.288	0.022	0.219	2.843
Genre_other	1.5277	0.670	2.281	0.023	0.215	2.841
Intercept	6.1247	2.678	2.287	0.022	0.874	11.375
Critic_Score:Critic_Count	9.8955	6.515	1.519	0.129	-2.877	22.668
Critic_Score:User_Score	-10.6254	8.156	-1.303	0.193	-26.615	5.365
Critic_Score:User_Count	7.8017	4.906	1.590	0.112	-1.816	17.419
Critic_Count:User_Score	-2.9428	2.143	-1.373	0.170	-7.143	1.258



Cook's method revealed that most of the observations had a value below 0.5. Cook's distance values below 0.5 are often considered to indicate relatively low influence or negligible impact of the corresponding observations on the regression model. This suggests that all observations do not have a substantial effect on the fitted values and overall model performance.

The heat map of the **multicollinearity** matrix did show highly correlated variables in the dataset.

In the presence of multicollinearity, the coefficient estimates of the correlated variables become unstable and their interpretations become challenging. Small changes in the data or the model specification can lead to significant changes in the estimated coefficients.



RFE (Recursive Feature Elimination)

reference: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

The estimator is trained on the initial set of features and the importance of each feature is obtained either through any specific attribute or callable. Then, the least important features are pruned from the current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

<u>Model Summary after performing RFE</u>

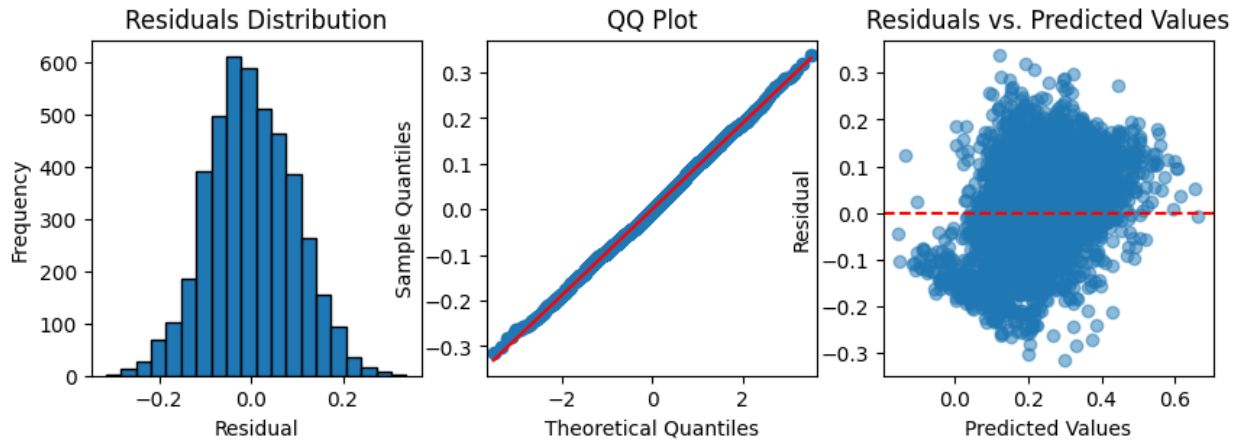
OLS Regression Results

Dep. Variable:	Global_Sales	R-squared (uncentered):	0.813
Model:	OLS	Adj. R-squared (uncentered):	0.812
Method:	Least Squares	F-statistic:	1362.
Date:	Fri, 30 Jun 2023	Prob (F-statistic):	0.00
Time:	20:54:22	Log-Likelihood:	4152.0
No. Observations:	4413	AIC:	-8276.
Df Residuals:	4399	BIC:	-8187.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Critic_Score	14.0727	4.016	3.504	0.000	6.198	21.947
Critic_Count	-19.8465	3.113	-6.375	0.000	-25.949	-13.744
User_Score	5.6983	3.729	1.528	0.127	-1.612	13.009
User_Count	-2.9611	0.473	-6.259	0.000	-3.889	-2.033
Critic_Score:Critic_Count	22.2984	5.407	4.124	0.000	11.699	32.898
Critic_Score:User_Score	-6.2926	5.732	-1.098	0.272	-17.531	4.946
Critic_Count:User_Score	-2.7200	1.868	-1.456	0.146	-6.383	0.943
Critic_Count:User_Count	10.7362	0.812	13.219	0.000	9.144	12.328
Critic_Score_squared	-16.4885	6.373	-2.587	0.010	-28.982	-3.995
User_Count_squared	-2.6703	0.396	-6.738	0.000	-3.447	-1.893
User_Score:Publisher_bElectronic_Arts	0.0748	0.007	10.208	0.000	0.060	0.089
User_Count:Publisher_bTake_Two_Interactive	0.0039	0.419	0.009	0.993	-0.818	0.826
User_Score:Publisher_bNintendo	0.1676	0.013	12.896	0.000	0.142	0.193
User_Count:User_Score:Publisher_bTake_Two_Interactive	0.0485	0.742	0.065	0.948	-1.405	1.502

Omnibus:	7.714	Durbin-Watson:	2.020
Prob(Omnibus):	0.021	Jarque-Bera (JB):	7.702
Skew:	0.092	Prob(JB):	0.0213
Kurtosis:	2.909	Cond. No.	1.13e+04

Residual Plots



Conclusions

The data frame underwent additional analysis to identify and address outliers. To improve the linearity assumption, a log transformation was applied to the features, resulting in better linear fits. In order to capture potential interactions and nonlinear relationships, interaction variables were introduced, effectively increasing the number of features in the model.

Further investigation is necessary to explore the factors contributing to high p-values and consider the inclusion of additional interaction variables.

The final model achieved an R-squared value of 0.813, indicating that approximately 81.3% of the variance in the dependent variable can be explained by the independent variables included in the model. The residual plots provide valuable information about the distribution and patterns of the residuals, aiding in assessing the adequacy of the model's assumptions and identifying any potential issues.

Final model

y = 14.09Critic_Score + -19.85Critic_Count + 5.69User_Score + -2.96User_Count + 22.31Critic_Score:Critic_Count + -6.27Critic_Score:User_Score + -2.72Critic_Count:User_Score + 10.74Critic_Count:User_Count + -16.51Critic_Score_squared + -2.67User_Count_squared + 0.07User_Score:Publisher_bElectronic_Arts + 0.03User_Count:Publisher_bTake_Two_Interactive + 0.17*User_Score:Publisher_bNintendo