CHAMPLAIN COLLEGE DAT-310

Video Game Pre-Processing

Jon Trusheim, Eric Cacciavillani

February 24, 2019

.

1 ABOUT THE DATA

1.1 Details

The given dataset was originally web-scrapped from Gregory Smith's web scrape of VGChartz. It has since been extended further using webscrapping from Metacritic. It contains the following attributes for each game:

- **NA_Sales\JP_Sales\EU_Sales\Other_Sales\Global_Sales**:
  Games sales based on location.
- **Critic_Score\User_Score\User_Count\Critic_Count**:
  The user and critic scores with the associated counts.
- **Game Metadata**:
  - Name
  - Platform
  - Genre
  - Publisher
  - Developer
  - Rating
  - Year of Release

1.2 Dataset Shape

Including games who count as seperate games per platform we have a total of **16,719** games. But we have **11,562** unique games within our dataset.

1.3 Dataset importance

By finding what attributes of a game that makes a game successful can have multiple implications for helping developers, investors, publishers, marketing teams, critics, and users to better equip them with knowledge that can save ample wealth in their own given sector.

2 Problems/Challanges with the dataset.

2.1 Problems with nans

There are many areas where data-cleaning is required before we can properly do clustering and prediction based models. Looking bellow at **Figure 2.1** we can see that numeric features of 'User' and 'Critic' attributes have ample nans that we may have to remove or use interpolation to solve. Considering that the following attributes have around half the data as null. It should also be noted that Year of release sometimes contains **to be announced** which was also causing problems within the feature. To continue with, it should be noted that entire games may be filled with enough nans to validate dropping the given game all together.

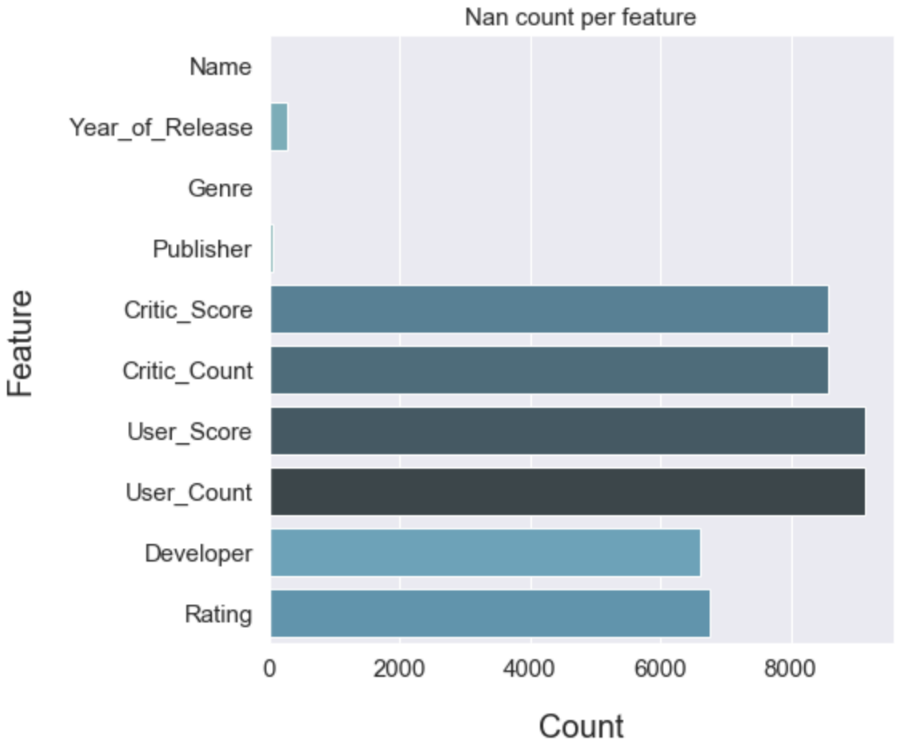
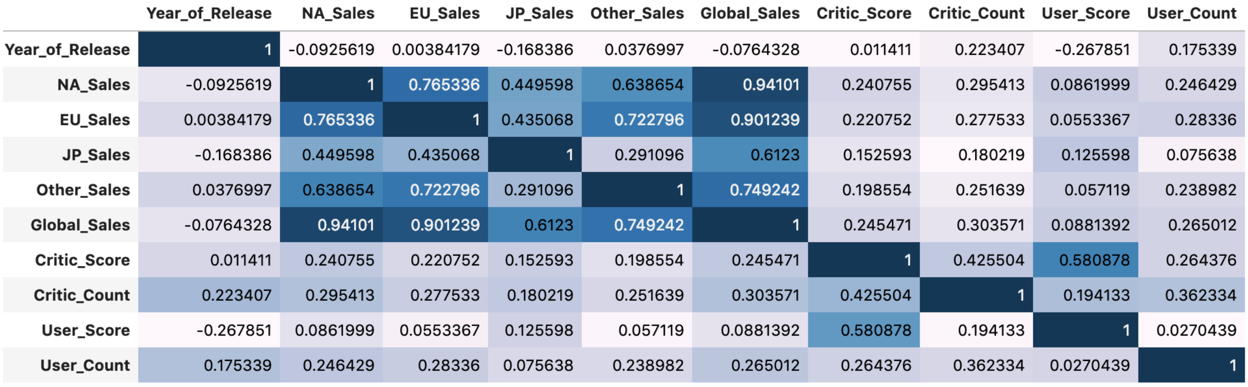


Figure 2.1: Nan count per feature.

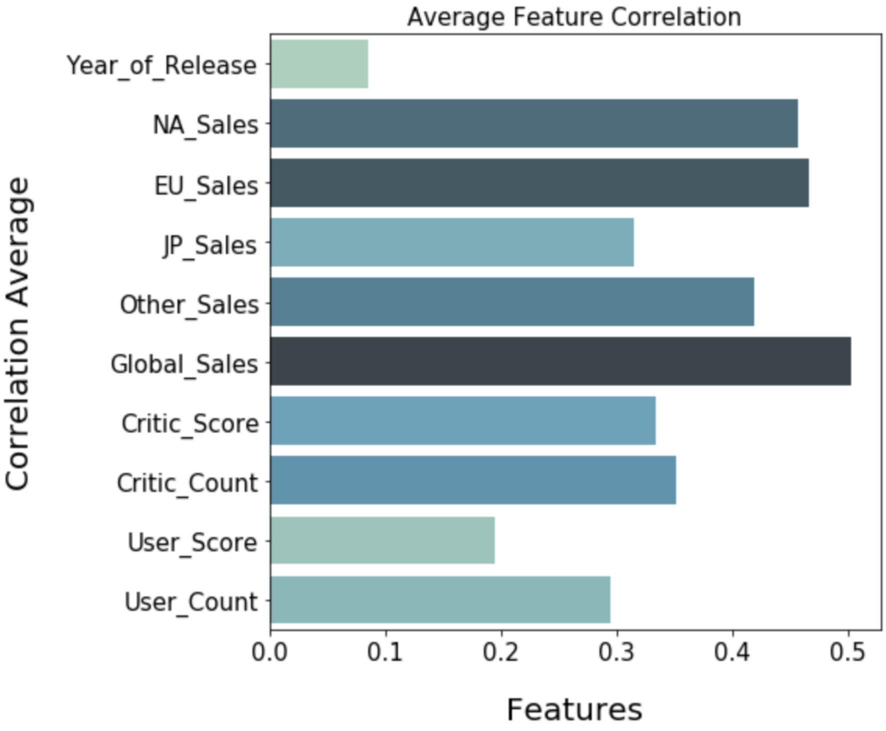
2.2 Correlation of Features

Looking bellow at **Figure 2.2** shows the correlation between features. High levels of correlation between features can cause those features to have an overall higher impact than intended when generating models. The features in this dataset with high levels of correlation between all sale based attributes.



	Year_of_Release	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_Score	User_Count
Year_of_Release	1	-0.0925619	0.00384179	-0.168386	0.0376997	-0.0764328	0.011411	0.223407	-0.267851	0.175339
NA_Sales	-0.0925619	1	0.765336	0.449598	0.638654	0.94101	0.240755	0.295413	0.0861999	0.246429
EU_Sales	0.00384179	0.765336	1	0.435068	0.722796	0.901239	0.220752	0.277533	0.0553367	0.28336
JP_Sales	-0.168386	0.449598	0.435068	1	0.291096	0.6123	0.152593	0.180219	0.125598	0.075638
Other_Sales	0.0376997	0.638654	0.722796	0.291096	1	0.749242	0.198554	0.251639	0.057119	0.238982
Global_Sales	-0.0764328	0.94101	0.901239	0.6123	0.749242	1	0.245471	0.303571	0.0881392	0.265012
Critic_Score	0.011411	0.240755	0.220752	0.152593	0.198554	0.245471	1	0.425504	0.580878	0.264376
Critic_Count	0.223407	0.295413	0.277533	0.180219	0.251639	0.303571	0.425504	1	0.194133	0.362334
User_Score	-0.267851	0.0861999	0.0553367	0.125598	0.057119	0.0881392	0.580878	0.194133	1	0.0270439
User_Count	0.175339	0.246429	0.28336	0.075638	0.238982	0.265012	0.264376	0.362334	0.0270439	1

Figure 2.2: Feature correlation map.



3 GENERAL DATA ANALYSIS

3.1 Single feature graphs

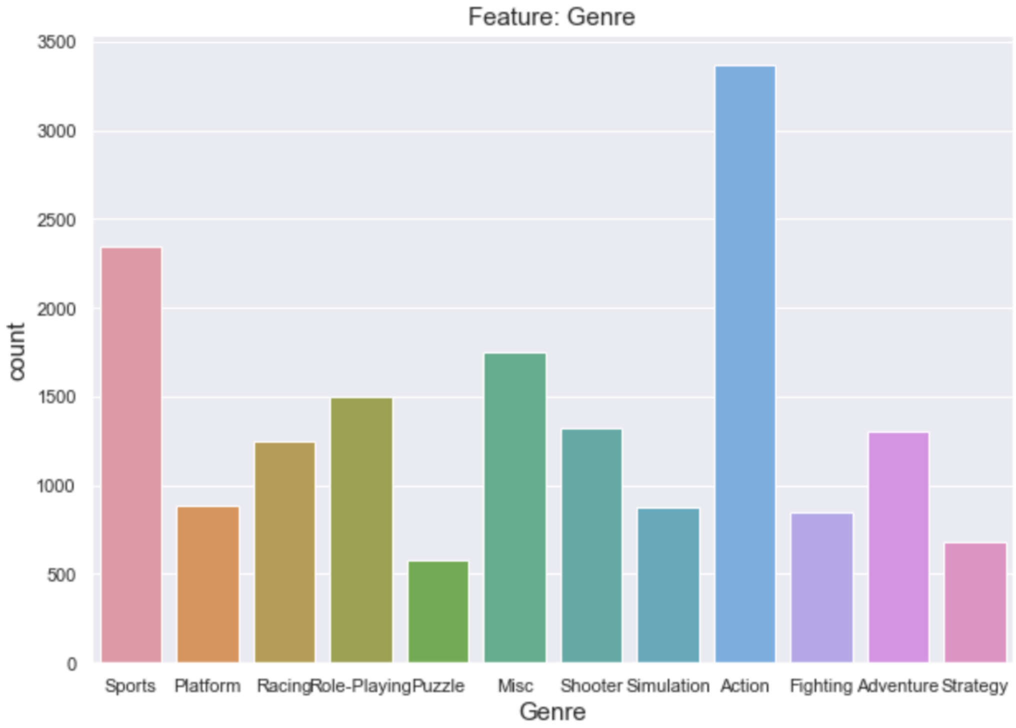


Figure 3.1: Action is a clear winner in the genre category.

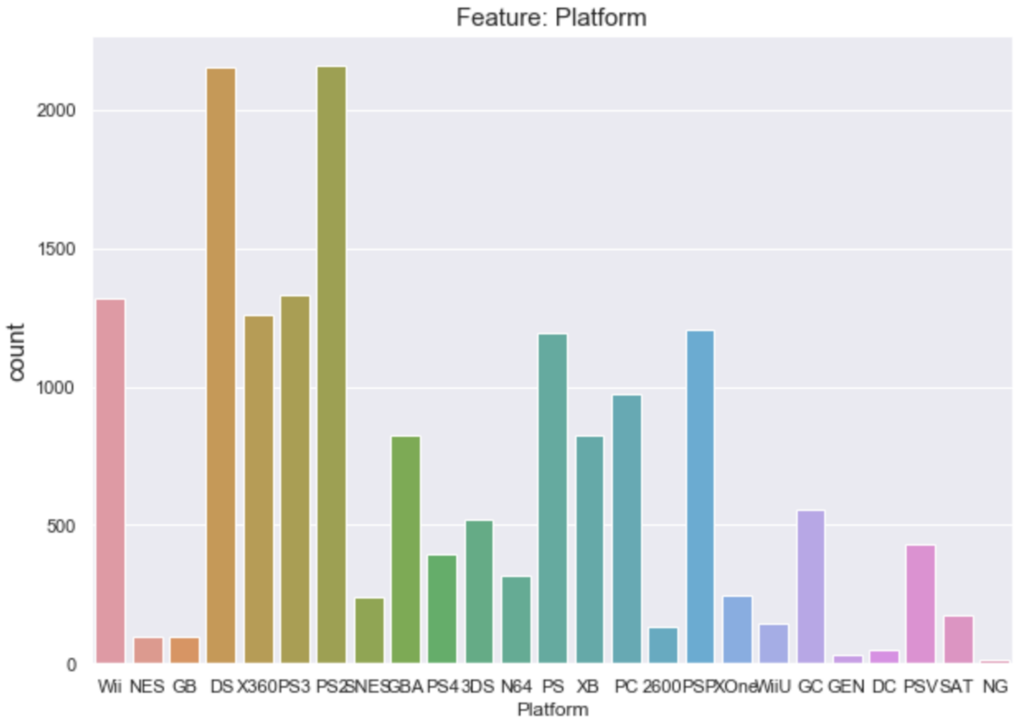


Figure 3.2: DS and the PS2 have the highest platform count within our dataset.

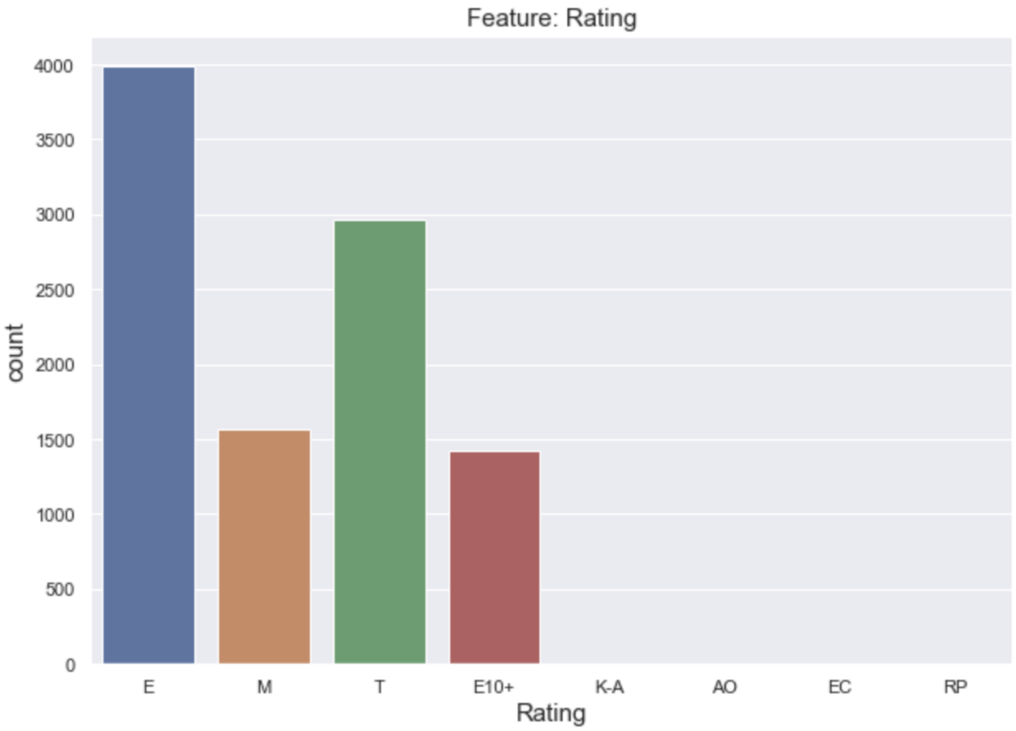


Figure 3.3: ’E’ for everyone and ’T’ for teen have a clear lead in the rating feature.

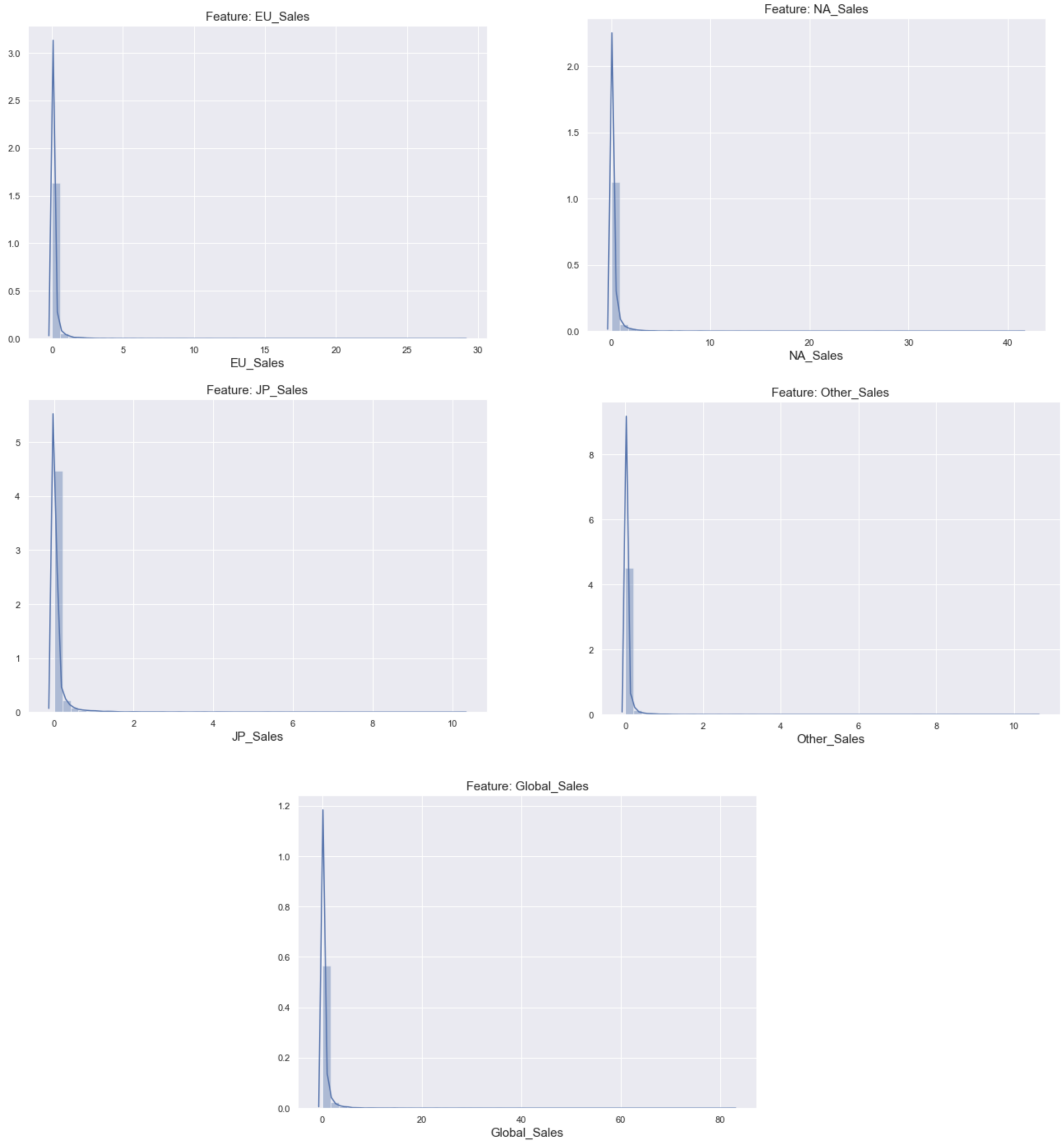


Figure 3.4: Displaying all sales. Sales distribution **needs** to be centered. Furthermore, the high correlation between each can clearly be seen within the following the graphs.

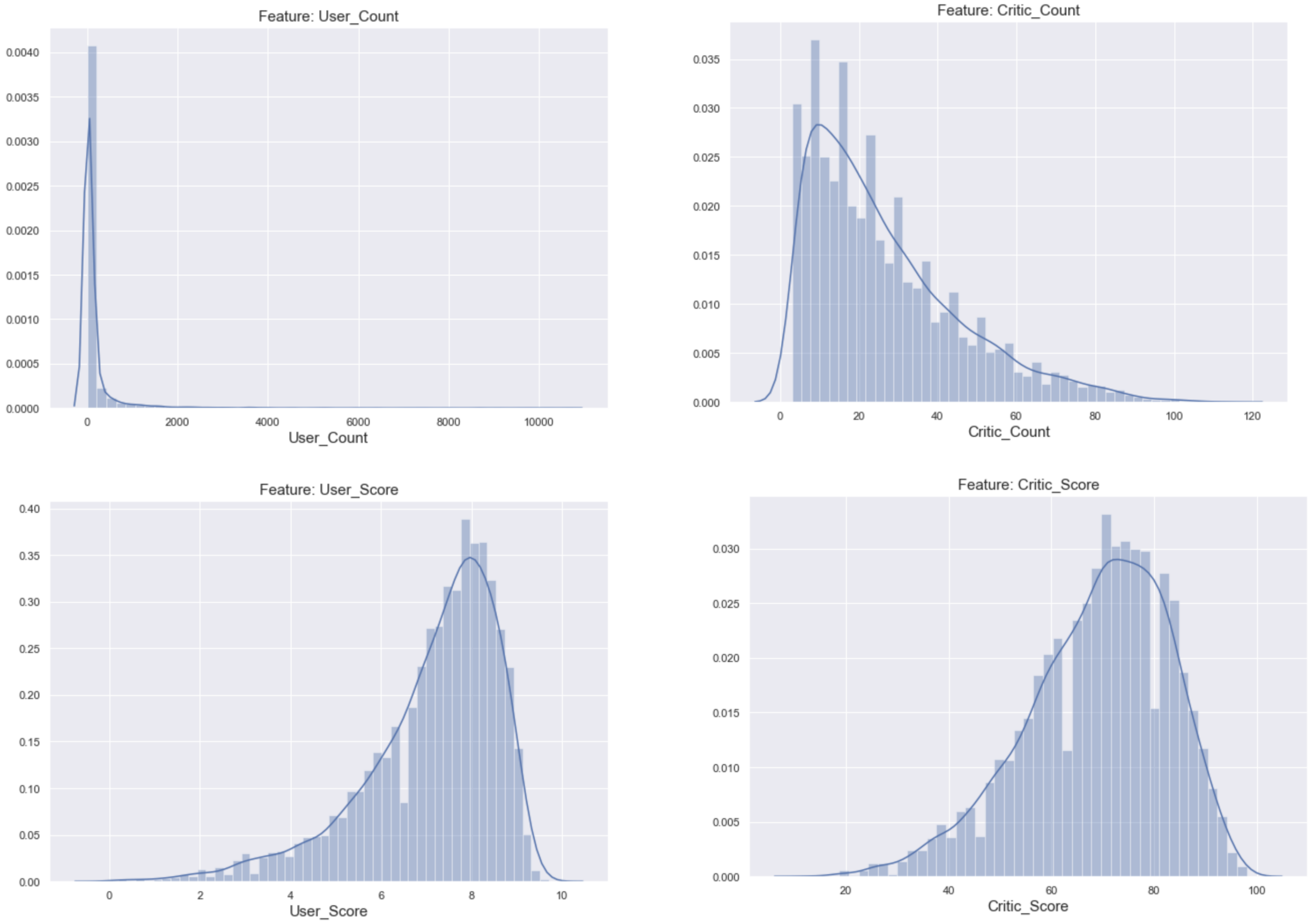


Figure 3.5: The distributions as a whole look pretty good except for User Count whose outlier is causing the distribution to look unseemly.

3.2 Multi-feature analysis

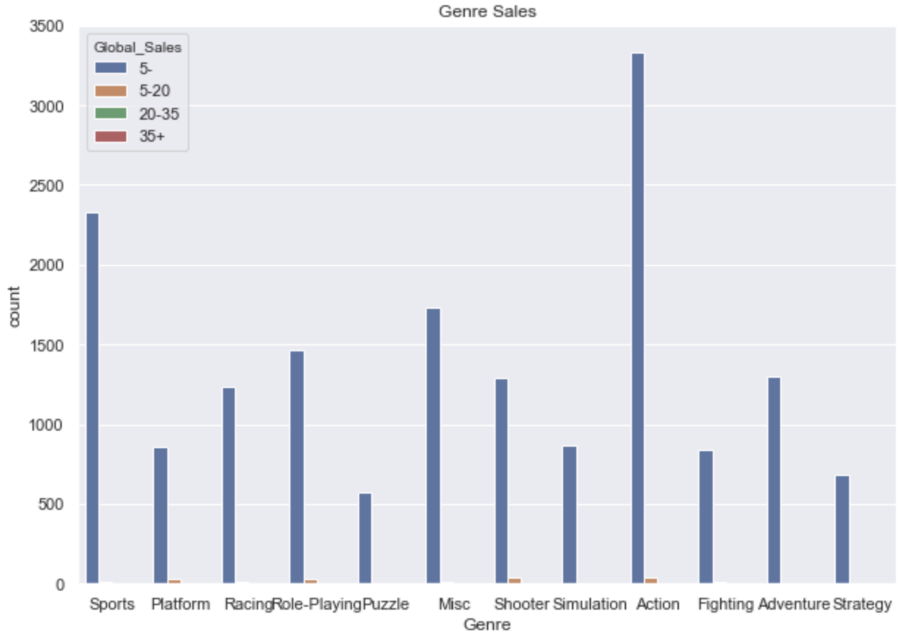


Figure 3.6: The following shows how many games make in sales by their associated genres.

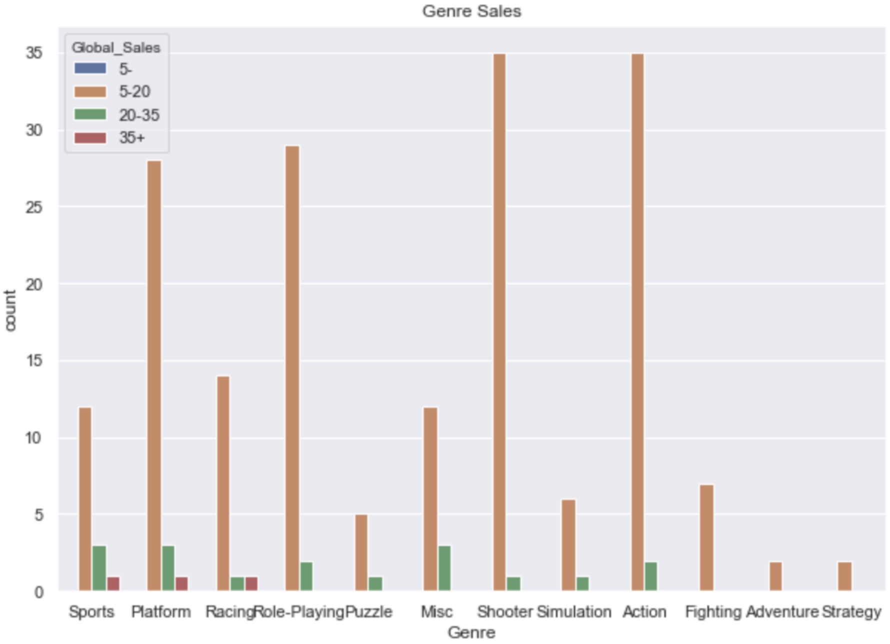


Figure 3.7: Removal of games that make under 5 million clear up the graph better for general analysis.

3.3 General Findings

Finding/answering some just general questions we had when begining this project:

- **What is the game(s) has the highest JP Sales?**
  Pokemon Red/Pokemon Blue
- **What is the game(s) has the highest EU Sales?**
  Wii Sports
- **What is the game(s) has the highest NA Sales?**
  Wii Sports
- **What is the game(s) has the highest Other Sales?**
  Grand Theft Auto: San Andreas
- **What is the game(s) has the highest Global Sales?**
  Wii Sports
- **What is the game(s) has the highest Global Sales?**
  Wii Sports
- **What is the game(s) has the highest Critic Score?**
  Grand Theft Auto IV, Tony Hawk's Pro Skater 2, SoulCalibur
- **What is the game(s) has the highest User Score?**
  Breath of Fire III

4 Data transformation

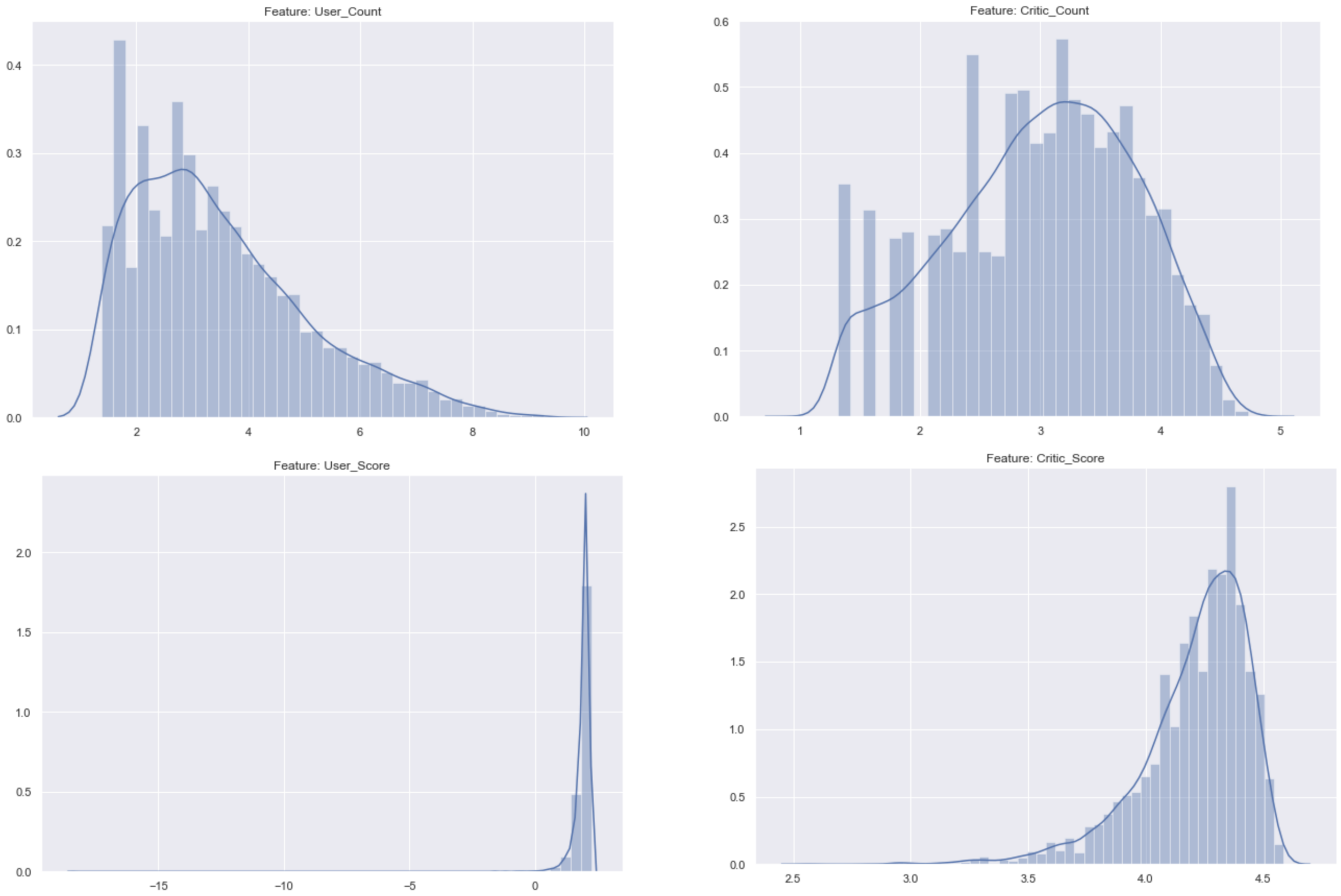


Figure 4.1: The centering seemed to really help the user count and the critic count. But it seems to have hurt the scores of both user and critic.

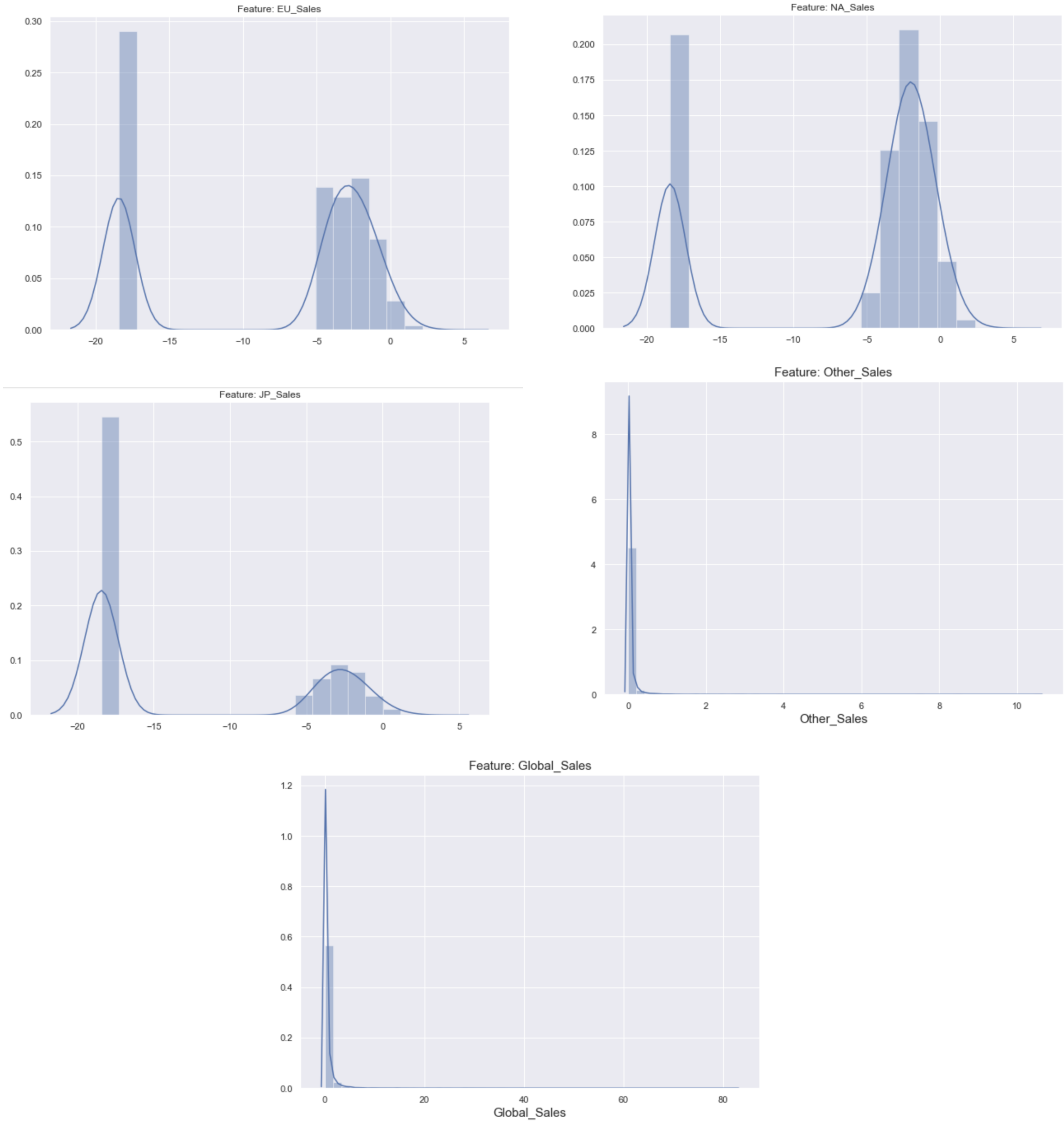


Figure 4.2: The logged did what it can but the data is simply not that distributed normally.