Study the general information:

There is one table, which I've named data. It has 11 columns, currently of type object and float datatypes. There are 16,715 entries in the entire table. The following columns have less than 16,715 non-null rows: Year_Of_Release, Genre, Critic_Score, User_Score, Rating - this means that they have null values which need to be either dropped or filled (to be decided later). There are 11,560 unique games in this database.

I started my making all the column names lowercase using the lower() method.

I found no data type problems or missing values in the name, platform, na_sales, eu_sales, jp_sales, other_sales columns.

For the 'genre' column, there was two rows with 'nan' values for genre. They also had NaN valuesfor other columns, so I decided to drop these two rows.

I decided to drop the rows with no NA,JP, EU and Other sales. There were only two of these rows and most of their other values were either 0  or NaN.

I dropped any duplicates present in the database.

For the 'year_of_release' column, there are 269 rows with a null value in that column. I noticed that some of these nulls are games who have valid years in other rows so I replaced those NaN values with the actual game release year if the game exists. I did this by iterating through the database, filling null values with 0s, identifying all the rows with year_of_release set to 0 within its individual dataframe and adding the index of these rows to a column named 'index' within this new dataframe. I iterated through each of these rows with year_of_release set to 0, checked if the name of the game matched the name of another game with a non null year_of_release and if it did, I updated the absent year with that discovered year.

Only 78/269 rows with missing year_of_release values were filled with valid years found within the data set. This means that there are still 191 rows with null year_of_release. I decided to drop these rows because 191 rows out of the entire 16715 row data is only 1.1% of data, which is less than 10% and therefore will not make a big impact on analysis.

For the 'critic_score' column, I found 8576/16715 (51.3%) rows with null values, which is more than half so I will not drop these. I used the transform function to replace the relevant the null values with the average of the critic scores for the same game name. This function works by grouping by the name of the game and retrieving the median of the critic score values for each group, then assigning that mean critic score value to any rows who have a matching game name.  Only 961/8574 (11.2%) rows had their critic score filled in using the average from other entires, the remaining 7615 rows were left as critic score being null.

For the 'user_score' column, I found 7951/16715 (47.6%) rows with either null or 'tbd' values, which is enough to not drop these. I replaced the 'tbd' values with None values, since no one is going to update this dataset with user scores. I casted every value in the user_score column to floats for standardization. I used the transform function to replace the relevant the null values with the average of the user scores for the same game name.  This function works by grouping by name of the game and retrieving the mean of the user score values for each group, then assigning that median user score value to any rows who have a matching game name. Only 905/7951 (11.3%) rows had their user scores filled in using the avergae from other entires, the remaining 7046 rows were left as user score being null.

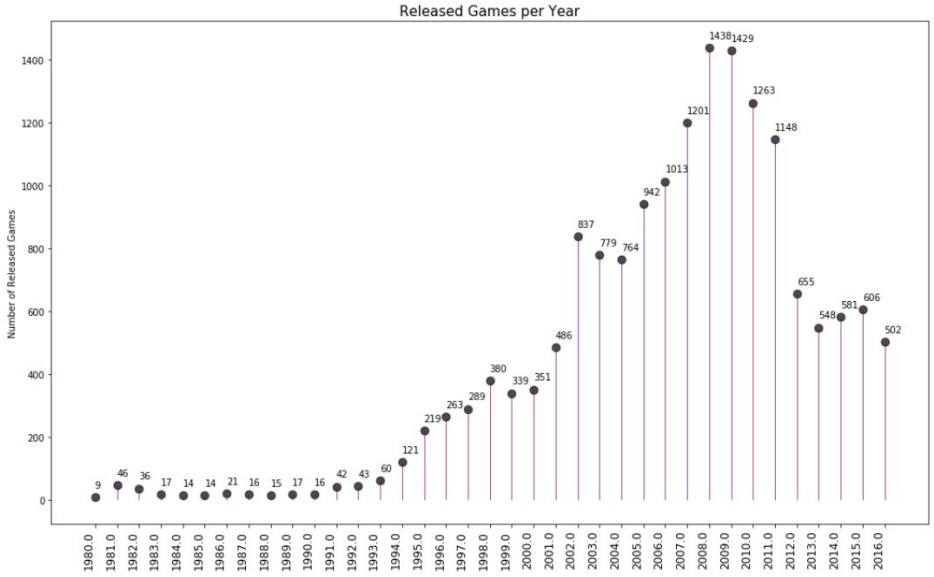
For the 'rating' column, I found 6762/16715 (40.4%) rows with null values, which is enough to not drop them. I used the transform function to replace null values with the most occuring rating in the same genre of the game. This function works by grouping by genre and retrieving the median of the rating values for each group, then assigning that median rating value to any rows who have a matching genre.

I calculated the total sales for each game by slicing the sales columns and summing them all into one  column, called sum_sales

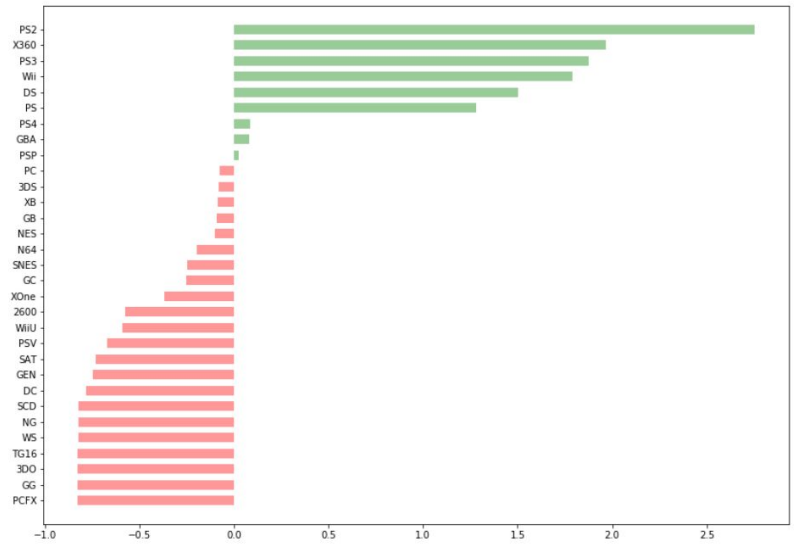
| | name | platform | year_of_release | genre | na_sales | eu_sales | jp_sales | other_sales | critic_score | user_score | rating | total_sales |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Wii Sports | Wii | 2006.0 | Sports | 41.36 | 28.96 | 3.77 | 8.45 | 76.0 | 8.0 | E | 82.54 |
| 1 | Super Mario Bros. | NES | 1985.0 | Platform | 29.08 | 3.58 | 6.81 | 0.77 | NaN | NaN | E | 40.24 |
| 2 | Mario Kart Wii | Wii | 2008.0 | Racing | 15.68 | 12.76 | 3.79 | 3.29 | 82.0 | 8.3 | E | 35.52 |
| 3 | Wii Sports Resort | Wii | 2009.0 | Sports | 15.61 | 10.93 | 3.28 | 2.95 | 80.0 | 8.0 | E | 32.77 |
| 4 | Pokemon Red/Pokemon Blue | GB | 1996.0 | Role-Playing | 11.27 | 8.89 | 10.22 | 1.00 | NaN | NaN | T | 31.38 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 16710 | Samurai Warriors: Sanada Maru | PS3 | 2016.0 | Action | 0.00 | 0.00 | 0.01 | 0.00 | NaN | NaN | T | 0.01 |
| 16711 | LMA Manager 2007 | X360 | 2006.0 | Sports | 0.00 | 0.01 | 0.00 | 0.00 | NaN | NaN | E | 0.01 |
| 16712 | Haitaka no Psychedelica | PSV | 2016.0 | Adventure | 0.00 | 0.00 | 0.01 | 0.00 | NaN | NaN | E | 0.01 |
| 16713 | Spirits & Spells | GBA | 2003.0 | Platform | 0.01 | 0.00 | 0.00 | 0.00 | NaN | NaN | E | 0.01 |
| 16714 | Winning Post 8 2016 | PSV | 2016.0 | Simulation | 0.00 | 0.00 | 0.01 | 0.00 | NaN | NaN | E | 0.01 |

16520 rows × 12 columns

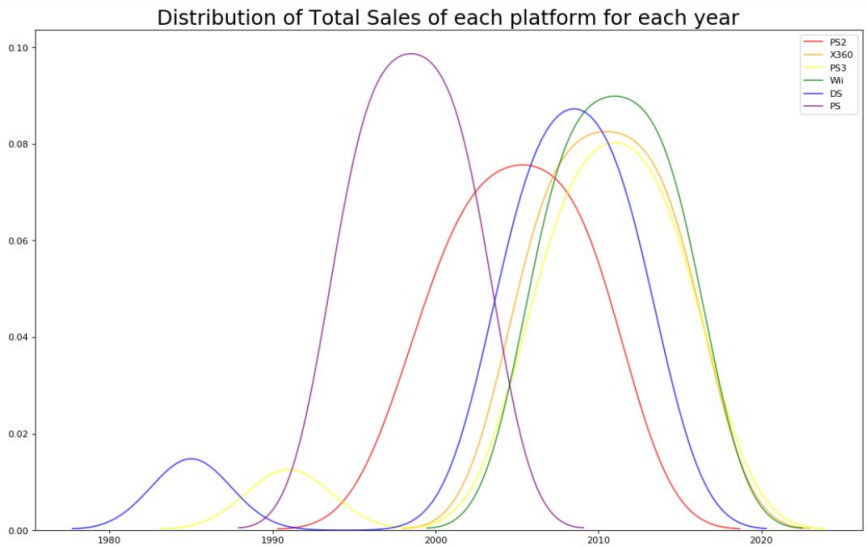
To look at how many games were released in different years, I created a lollipop graph that shows the number of games (on y axis) released for every year (on x axis) in the data set. The data before 2000 seems less significant than the data after 2000 as only in 2003 do the number of released games really jump up.



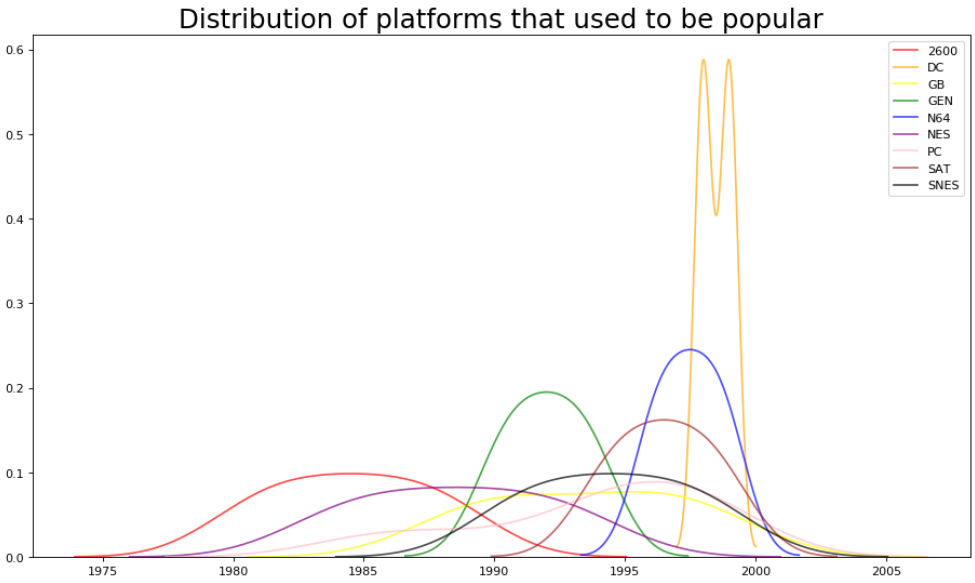
To look at sales varied from platform to platform, I built a divergence plot using the zscore which tells the distribution from the average total sales each platform total sales is. I subtracted the average of total sales from the total sales of that platform and divided it by the standard deviation to determine the z-score of each platform. The platforms with negative z-scores (in red) are platforms whose total sales are less than the average total sales for all platforms and the ones with positive z-score (in green) are platforms whose total sales are greater than the average total sales for all platforms. The platforms with the greatest total sales are the PS2, X360, PS3, Wii, DS and PS.




The platforms with the greatest total sales are the PS2, X360, PS3, Wii, DS and PS (top 6 platforms). To build a distribution based on data for each year, I built a density histogram that shows the distribution of total sales per year for each platform.



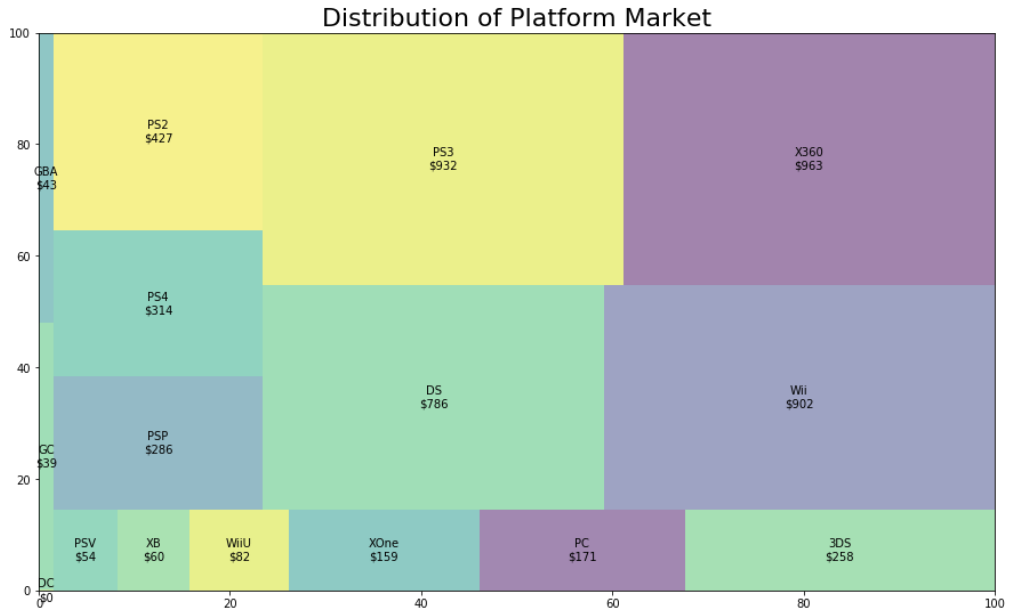
To find the playforms that used to be popular but now have zero sales, I identified platforms who have a negative z-score, a high total sales amount and who have a year of release less than 2000 (they haven't been released in the past 15 years). From these, I plotted the distribution of the platforms and it shows that all of these stopped havig sales by 2005.



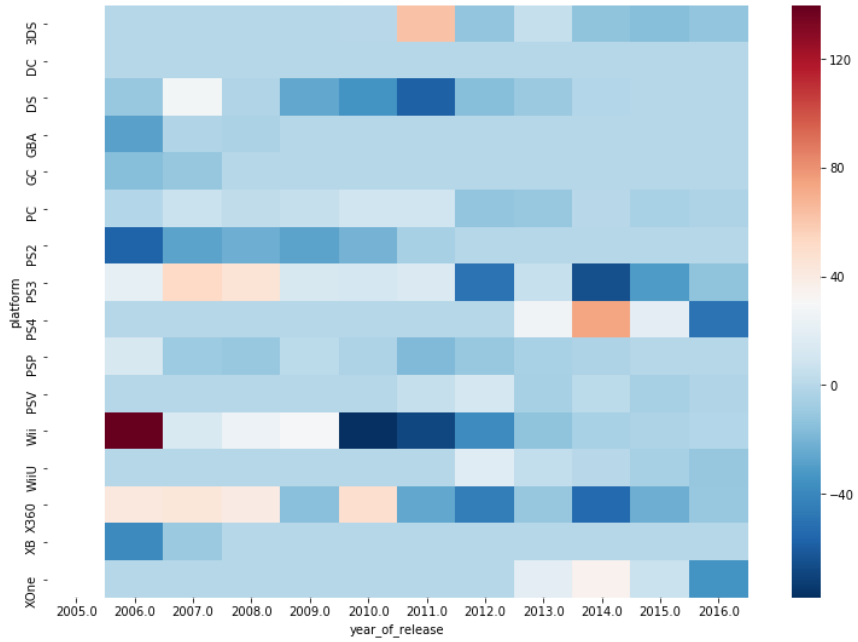

The average time it takes for new platforms to appear and old ones to fade is about 7.4 years.

To determine what period to take data from in order to plan a campaign for 2017, we should take data from the last 10 years beforehand. I will keep data with a year_of_release greater than 2005.

This tree map shows the distribution of the market. The platforms that are leading in sales include PS3, X360, DS and Wii.

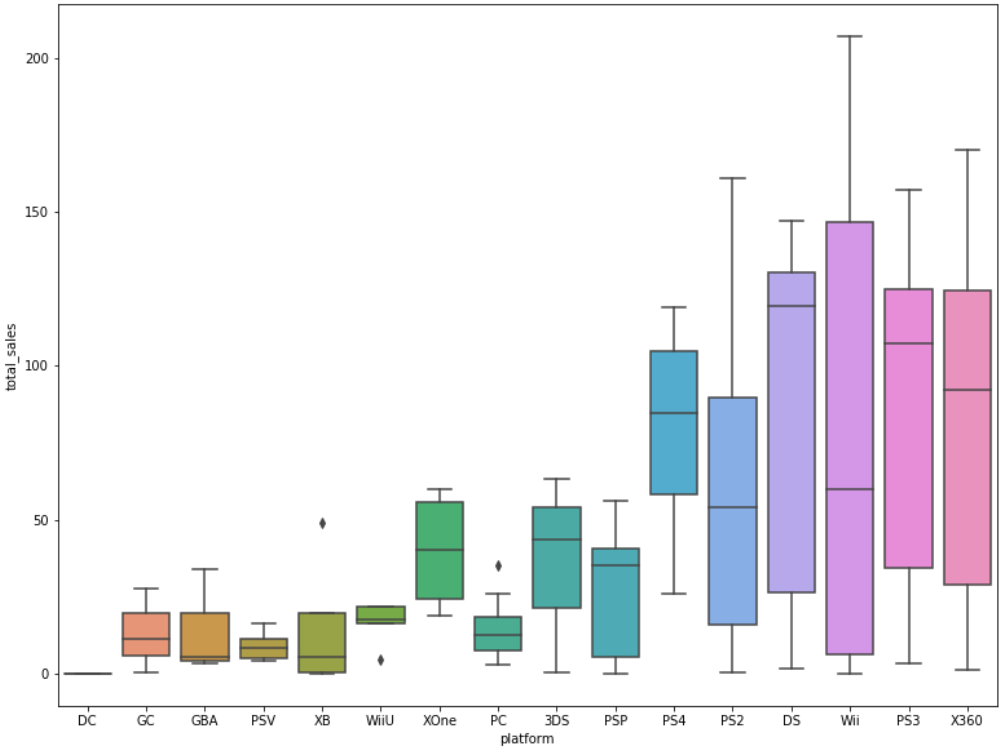



This heat map shows the sales success or drop of each platform per year. It shows the peak sales, increase in sales or decrease in sales for each platform where red means increase and blue means decrease. The platforms that are shrinking are PSV, PS2, DS, Wii. The platforms that are growing are PS4, XOne, PS3. Some potentially profitable platforms include PS4, X360, XOne.



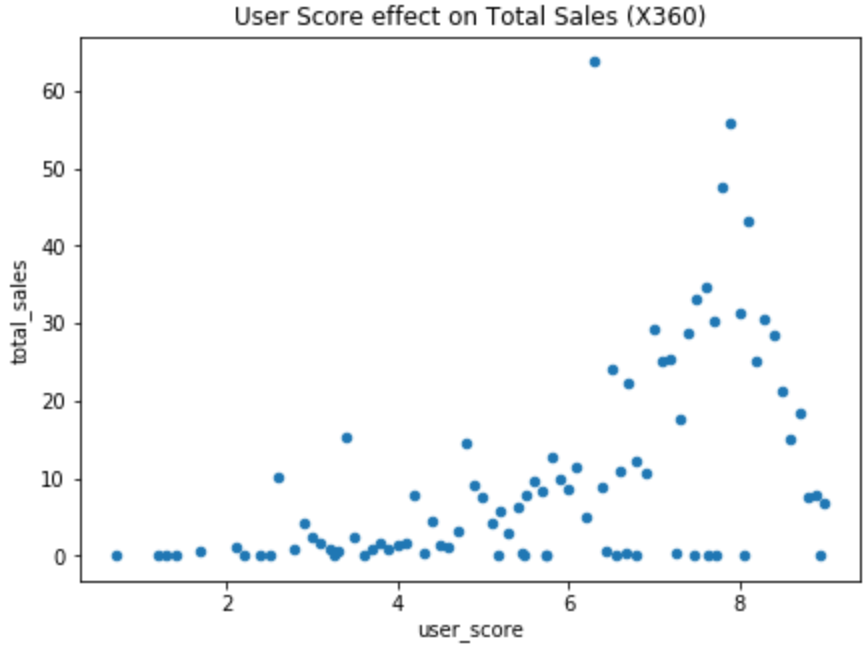
I have built a box plot for the global sales of games, brokwn down by platform. The differences in sales are significant - the games sold on X360, PS3, Wii, DS, PS2 and PS4 are more profitable than the games sold on PSP, 3DS, PC, XOne, etc.

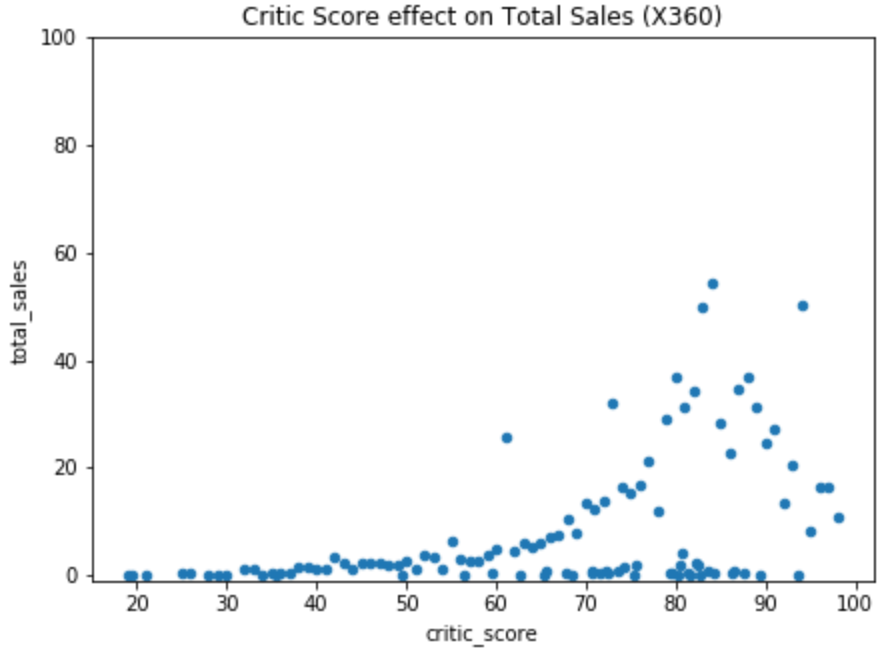
In terms of average sales over the 10 year period between 2005-2016: the average sales on X360, PS3, DS, and PS4 are the most comparable, then the average sales on Wii, PS2 are comparable, then the average sales on PSP, 3DS and XOne are comparable, and finally the rest have the lowest average sales.



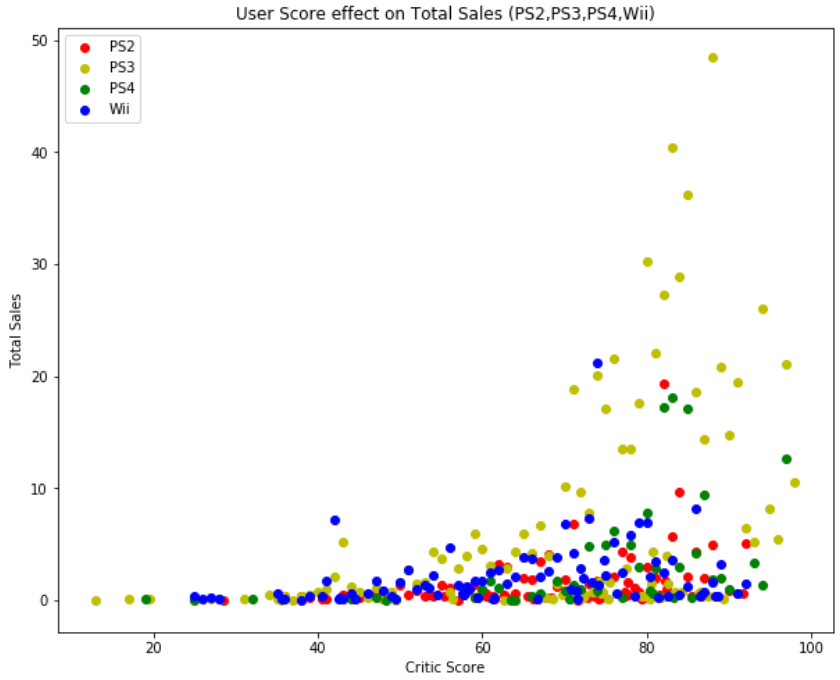

I chose the X360 as a popular platform to compare how professional critic scores and user scores affect its total sales. I found that both critic score and user score affect the total sales, as the higher the score given to the game on the platform then the higher the total sales.

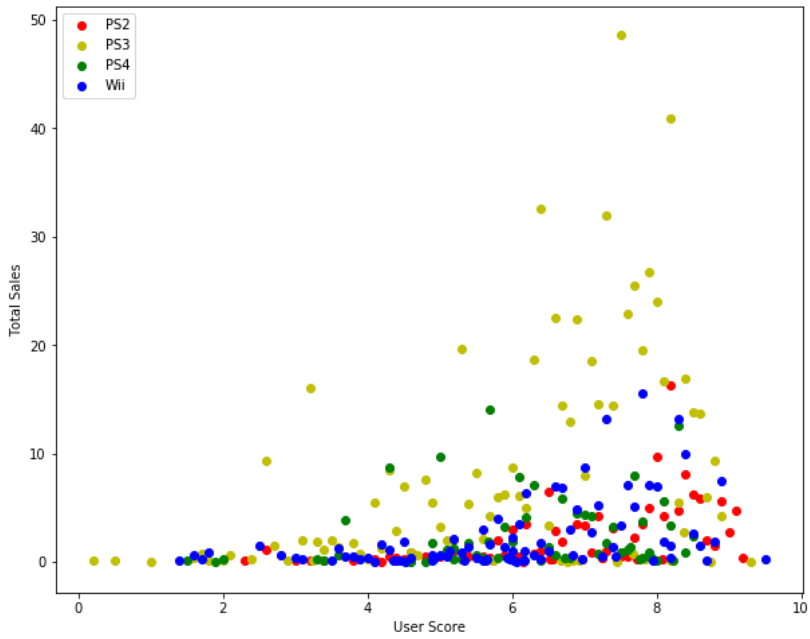
When looking at the correlation values, the user score correlation values (0.08) is much lower than the critic score correlation value (0.23). Buyers seem to trust professional criticers more than the average buyer.



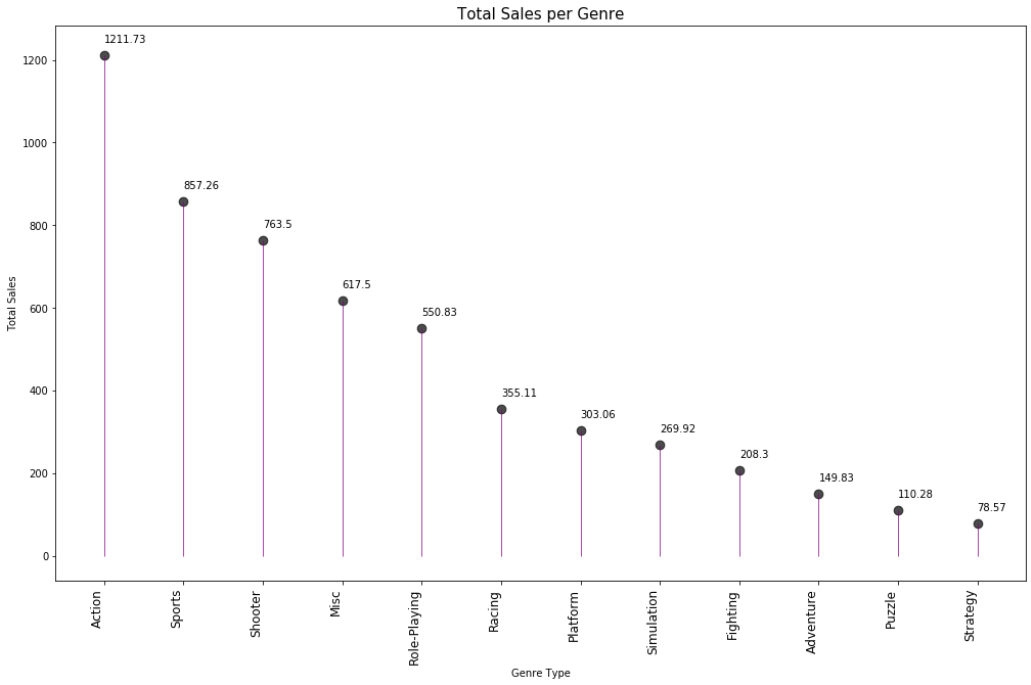



I chose the PS2, PS3, PS4, Wii as other popular platforms to compare how professional critic scores and user scores affect its total sales. I found that both critic score and user score affect the total sales, as the higher the score given to the game on the platform then the higher the total sales. However, when looking at the correlation values, the user score correlation value (0.30) is lower than the critic score correlation value (0.35). This was the opposite when I was looking at just the X360.

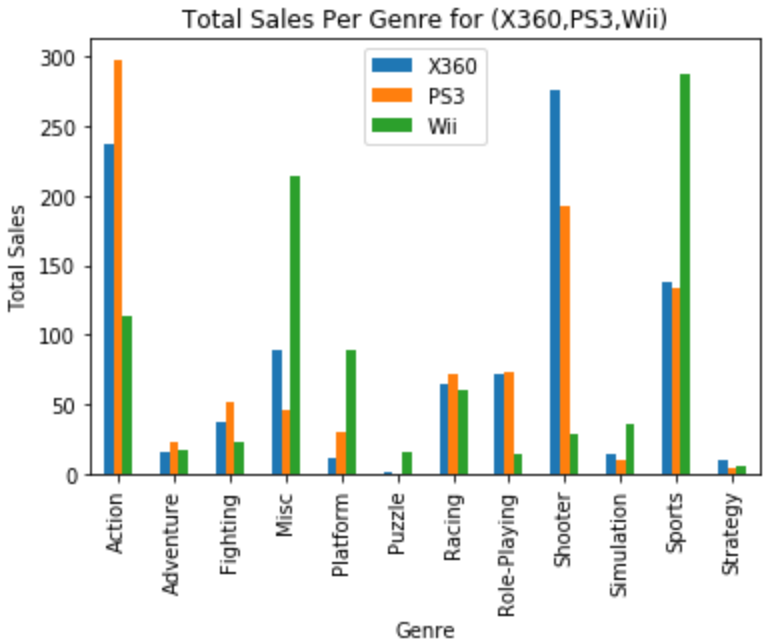


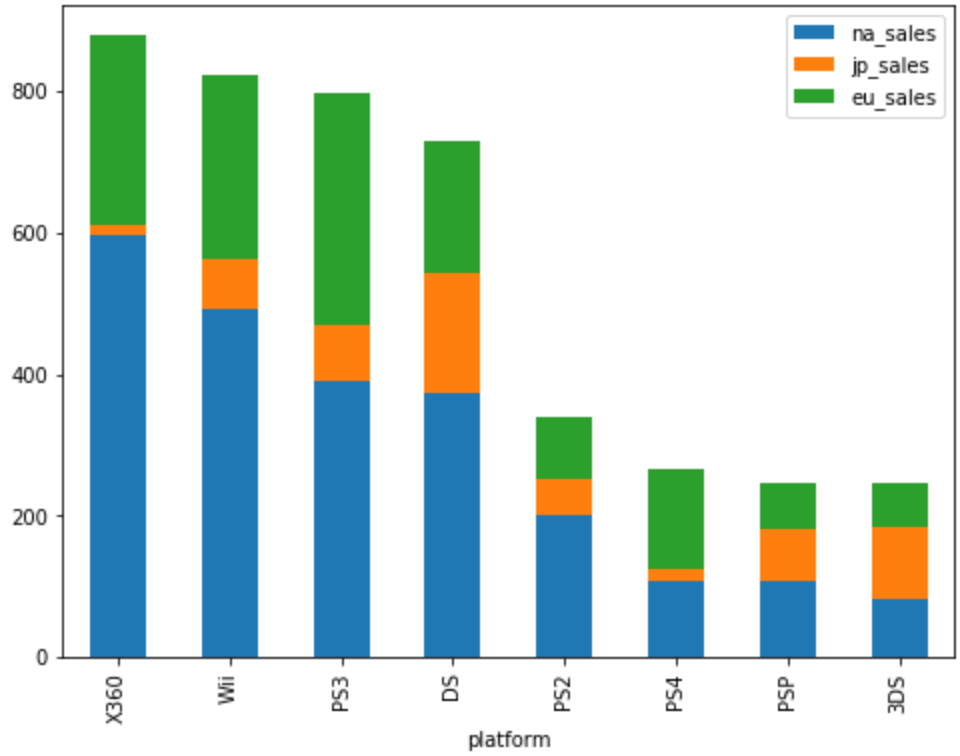
The most profitable genre are action (USD 1211.73 million), sports (USD 857.27 million), shooter (USD 763.5 million) and role-playing (USD 550.83 million).



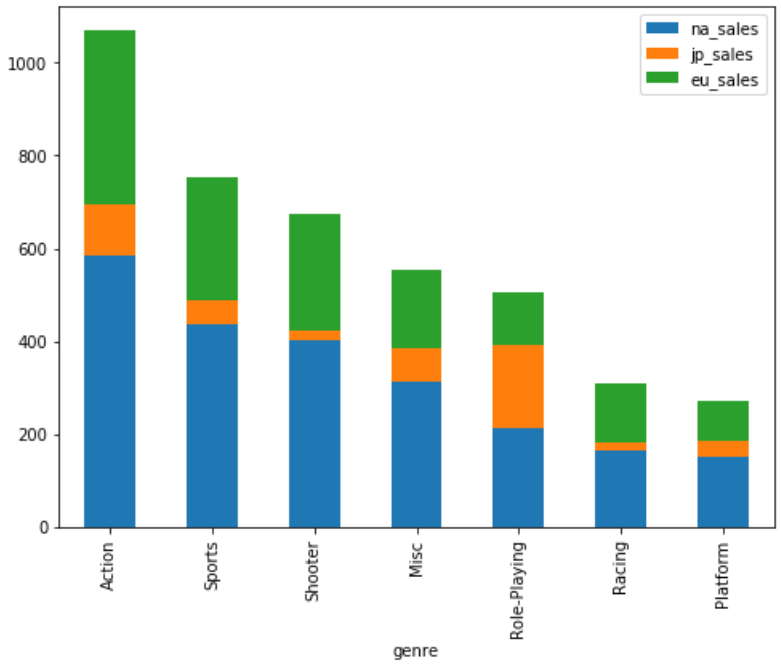

Genres with low sales (such as puzzle, strategy, adventure, fighting, simulation) are low selling genres across the most popular platforms. Genres with high sales (such as action, misc, shooter, sports) are high selling genres across the most popular platforms.



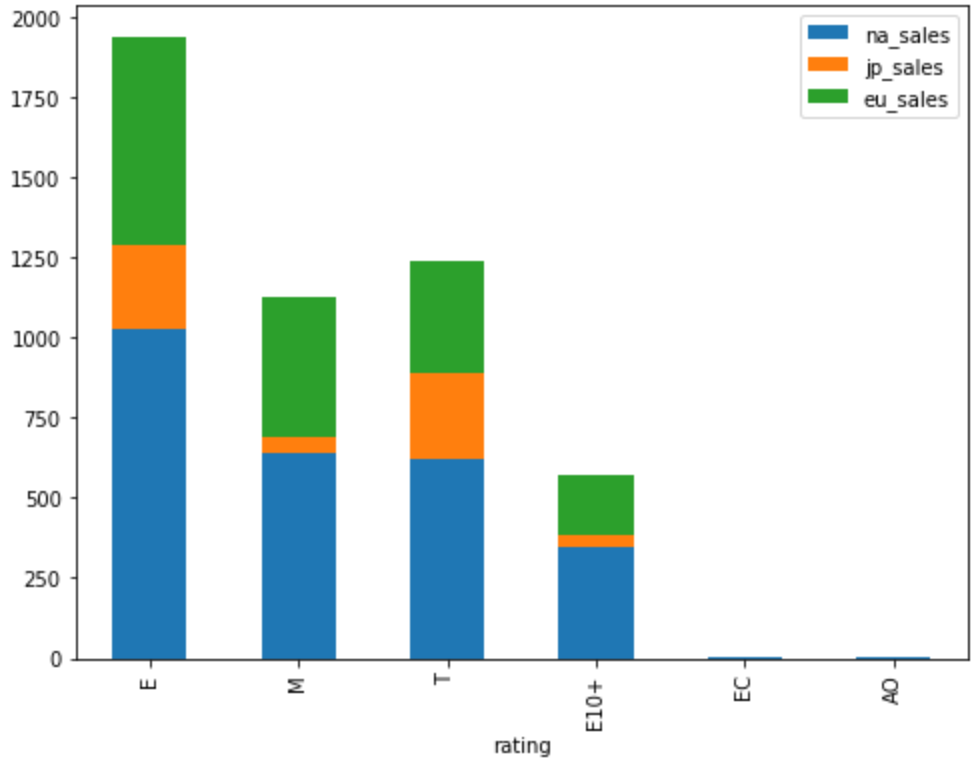
The top 5 platforms for NA are X360, Wii, PS3, DS, PS2, the top 5 for JP are DS, 3DS, PS3, PSP, Wii, the top 5 for EU are PS3, X360, Wii, DS, PS4.




The top 5 genres for NA are action, sport, shooter, misc, role-playing, the top 5 genres for JP are role-playing, action, misc, sport, platform, the top 5 genres for EU are action, sport, shooter, misc, racing.



This stacked bar graph displays whether ratings affect sales in each region:




We want to test the hypothesis that the average user rating of the Xbox One and PC platforms are the same. We can do this using an independent samples t-test to compare the means from the two groups (XBox One user ratings and PC user ratings. We can apply a t-test here due to the Central Limit Theorem, which impliees that you can estimate the mean of a statistical population using the mean of a sample, and since the means are approximately normally distributed - we can use the t-test.

Null Hypothesis H0: The average user rating of XBox One is the same as the average user rating of PC.
Alternative Hypothesis H1: The average user rating of XBox One is not the same as the average user rating of PC.

The p_value suggests that we should reject the null hypothesis, thus indicating that the average user rating of XBox One and PC platforms do differ enough to be statistically significant. This means that the sample results have a rare outcome and therefore it is very unlikely that to be just a lucky significance. This can be further proven by looking at the numbers: the average user rating of XBox one is 6.63/10 and the average user rating of PC is 6.89/10 - a 0.26 difference.

We want to test the hypothesis that the average user rating of the Action genre and Sport genres are different. We can do this using an independent samples t-test to compare the means

from the two groups (Action genre user ratings and Sport genre user ratings. We can apply a t-test here due to the Central Limit Theorem, which impliees that you can estimate the mean of a statistical population using the mean of a sample, and since the means are approximately normally distributed - we can use the t-test.

Null Hypothesis H0: The average user rating of Action genre is the same as the average user rating of Sport Genre.
Alternative Hypothesis H1: The average user rating of Action genre is different than the average user rating of Sport Genre.

The p_value suggests that we should reject the null hypothesis, thus indicating that the average user rating of Action genre and Sports genre do differ enough to be statistically significant. This means that the sample results have a rare outcome and therefore it is very unlikely that to be just a lucky significance. This can be further proven by looking at the numbers: the average user rating of Action genre is 6.93/10 and the average user rating of PC is 6.55/10 - a 0.55 difference.