**PROBLEM STATEMENT**

With more than 2.5 billion video gamers from all over the world, more gamers are switching towards mobile gaming compared to traditional video game consoles which are slowly moving out of phase.

In this project of mine, I’ll analyze sales data from 8k video games, identify most correlated to hits (games which sell over 1M units) and implement a prediction model to show which games released in 2016 can still become hit and which can’t.

Also, I’ll create a predictive model based on video gaming market revenue based on the data and identify what are the most important features that affects the revenue of video games.

Additionally, through analysis, I would like to identify the console among PS, XBOX, Nintendo, Play Station is the console that will possibly generate the most sales and effectively maximizing profits and minimizing the cost for further game development.

Lastly, I’ll be building a dashboard/report in PowerBI for better Data Visualization to sales director which will be helpful to him for taking better decisions.

**OUTCOMES:**

1. Which games (released in 2016) can still become hit and which can’t.
2. Predictions for sales volume.
3. PowerBI dashboard.

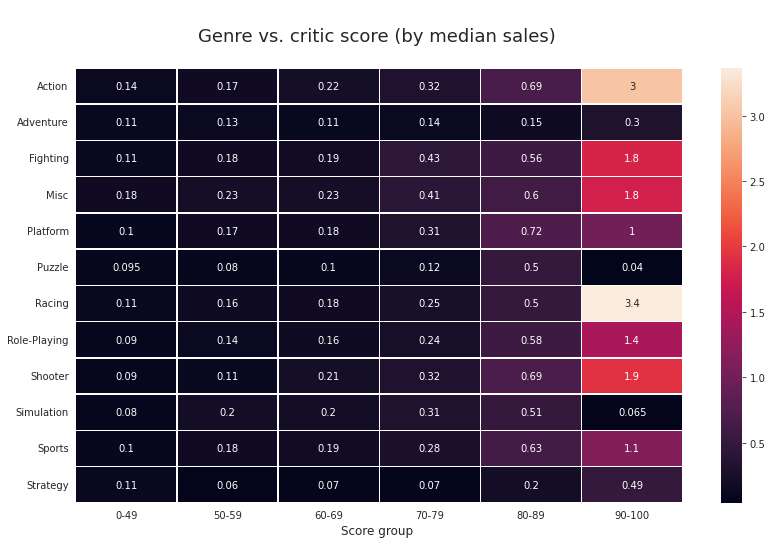
**PREDICTING VIDEO GAMES HITS WITH MACHINE LEARNING**

**Data Exploration and Analysis**

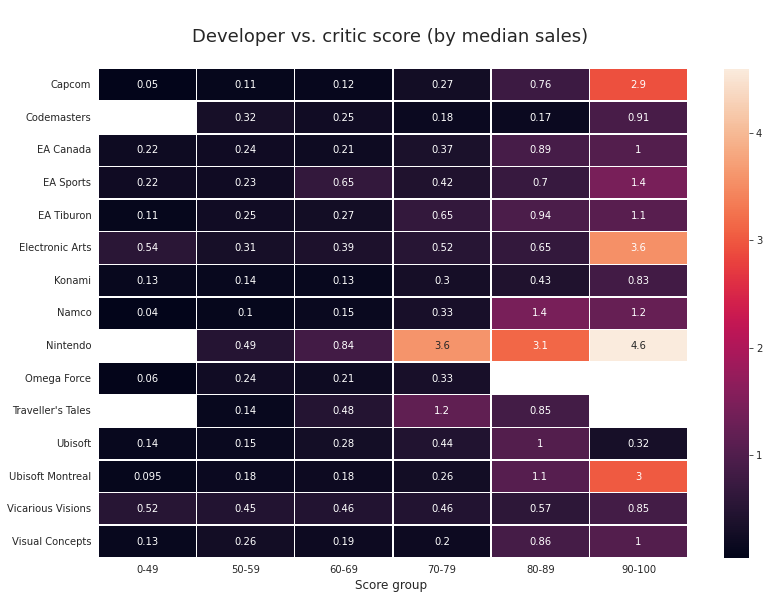
**Median sales (in millions of units) vs. critic scores**

The following four heatmaps show how game sales vary according to critic scores, which are split into six scoring groups. Additionally, each heatmap segments the data further by one of the following features: genre, developer, publisher, and platform (in order of appearance).

Under each heatmap, we identify the categories where games sell best. This is done for okay, good, and great games, as defined by games with scores in the 70s, 80s, and 90s, respectively.

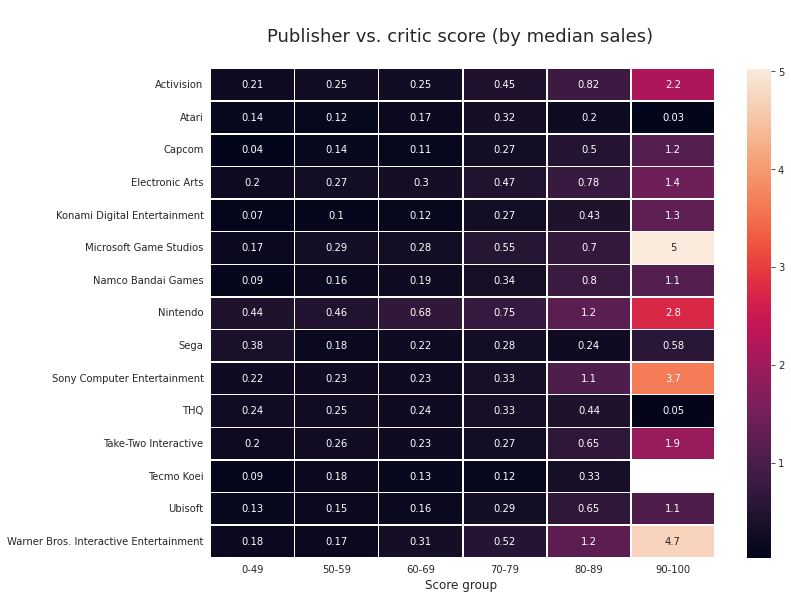


* **Genres where great games sell best:**  Racing, Action
* **Genres where good games sell best:** Platform, Shooter/Action
* **Genres where okay games sell best:** Fighting, Misc

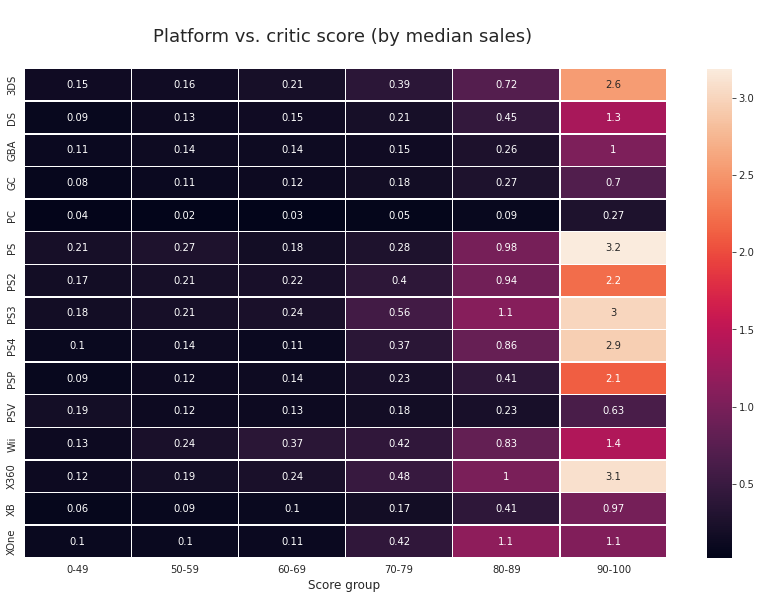


* **Developers whose great games sell best:**  Nintendo, Electronic Arts
* **Developers whose good games sell best:** Nintendo, Namco
* **Developers whose okay games sell best:** Nintendo, Traveler’s Tales

Interpretations: In the **great** scores column (last), Nintendo has the highest median sales (in millions of units) per game, at 4.6M. Interestingly, Nintendo also has the highest median sales per game in both the **good** and **okay** scoring columns.



* **Publishers who sell great games best:** Microsoft Game Studios**,** Warner Bros. Interactive Entertainment
* **Publishers who sell good games best:** Nintendo/ Warner Bros. Interactive Entertainment
* **Publishers who sell okay games best:** Nintendo, Microsoft Game Studios

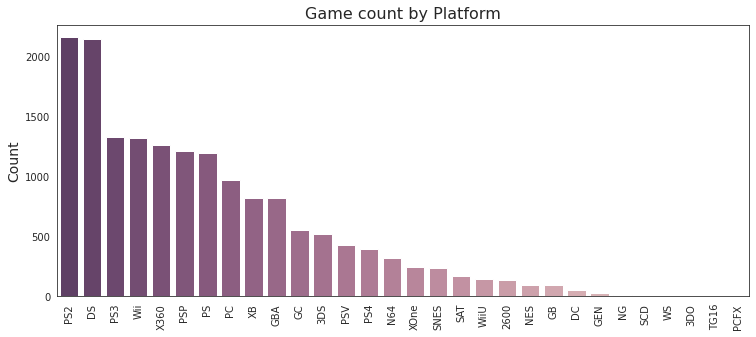


* **Platforms where great games sell best:**  PS, X360
* **Platforms where good games sell best:** PS3, XOne
* **Platforms where okay games sell best:** PS3, X360

It’s interesting how sensitive game sales in the whole PlayStation line seem to be to high critic scores, especially when sales in the mid-score ranges look relatively on par with other consoles.

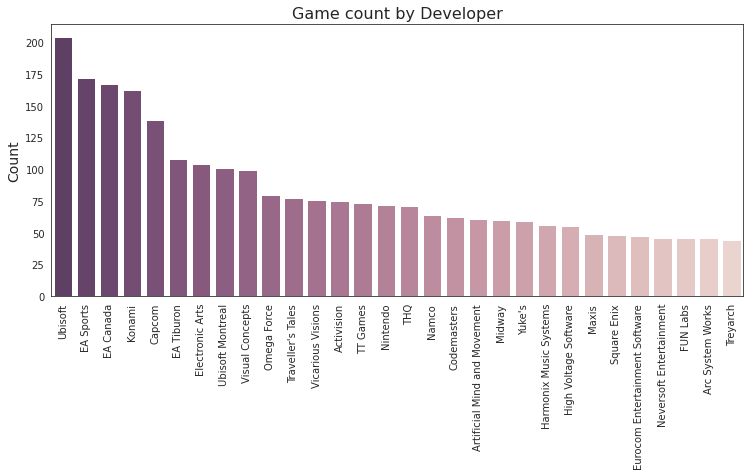
**TOP VALUES IN THE DATASET -** (BY PLATFORM, DEVELOPER, PUBLISHER AND GENRE)

**Platforms with most games in dataset:**



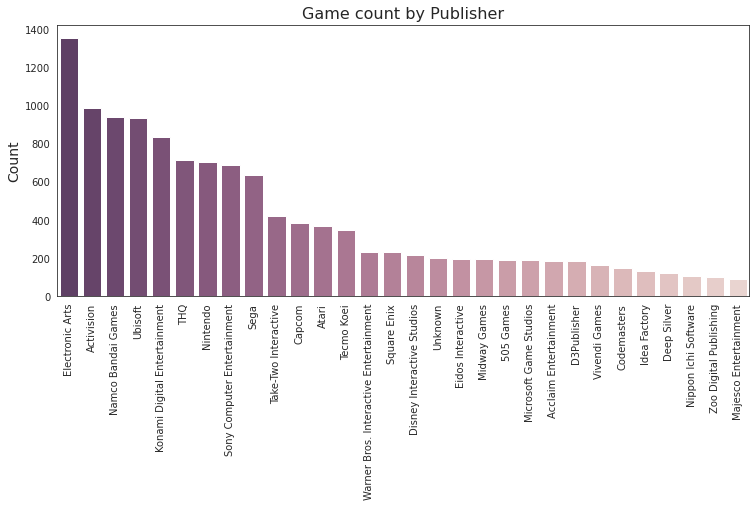
1. PS2
2. DS
3. PS3

**Developers with most games in dataset:**



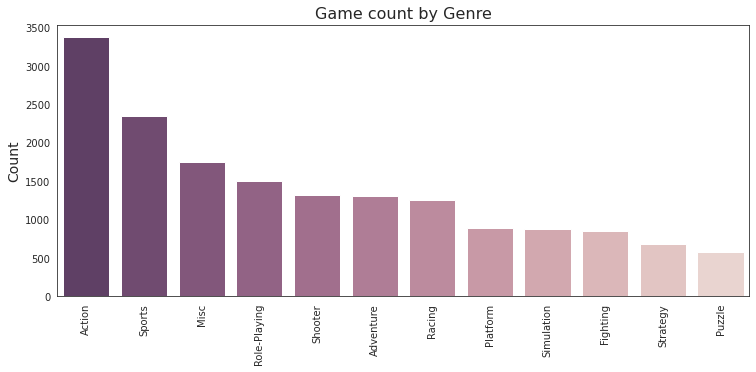
1. Ubisoft
2. EA Sports
3. EA Canada

**Publishers with most games in dataset:**

****

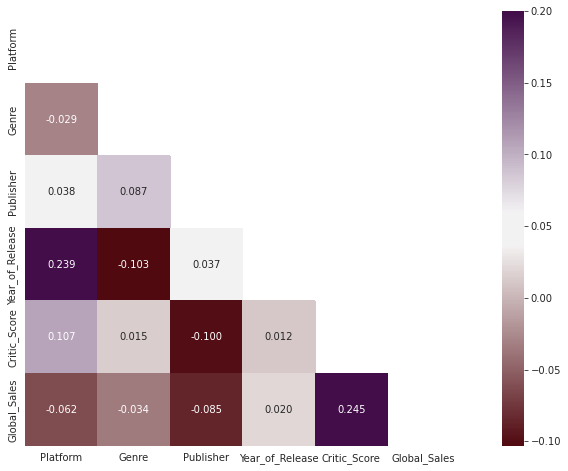
1. Electronic Arts
2. Activision
3. Namco Bandai Games

**Genres with most games in dataset**

****

1. Action
2. Sports
3. Misc

**DATASET CORRELATIONS** – (for numeric and categorical variables)

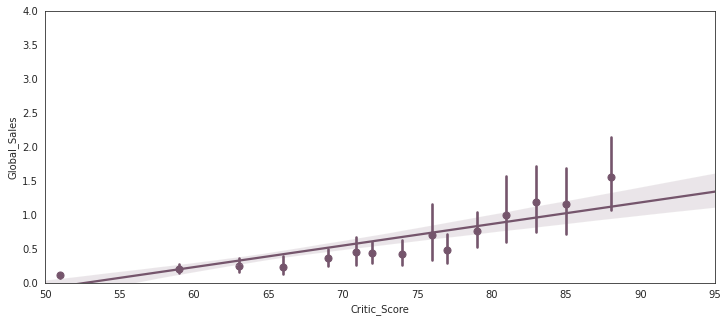
****

**Strongest correlations:**

* **Critic score-to-global sales:** We’ll take a closer look at this in the next two sections.
* **Year of release-to-platform:** This makes sense since new platforms are released periodically.

*Note: Categorical columns (platform, genre, publisher) were converted to numeric in order of game count, as seen in previous section. The slightly negative correlations the have to global sales can be interpreted as “the higher the ID number, the smaller the [platform, genre, publisher], and thus the slightly lower the sales figure”.*

**Critic score vs global sales** – (for all years in the dataset)

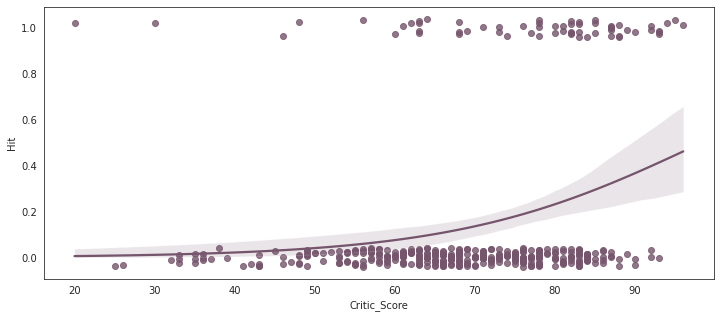
****

It’s interesting how the slope gets steeper in the 80’s. it seems once a video game gets high critic score, every additional point has a higher impact. For example, in this 2014-16 subset, **an 8-point increase in critic score seems to have positive effect on sales of about 250k when starting from a score of 65, but around 1M when starting from77.**

**DEFINING HITS AS THOSE WITH SALES ABOVE 1 MILLION UNITS**

This will be our target in our prediction model, where we’ll predict if a game will be a hit or not. The target is binary: 1 if Hit, else 0.

**Here’s the relationship between critic scores and VG hits using a 5% sample:**

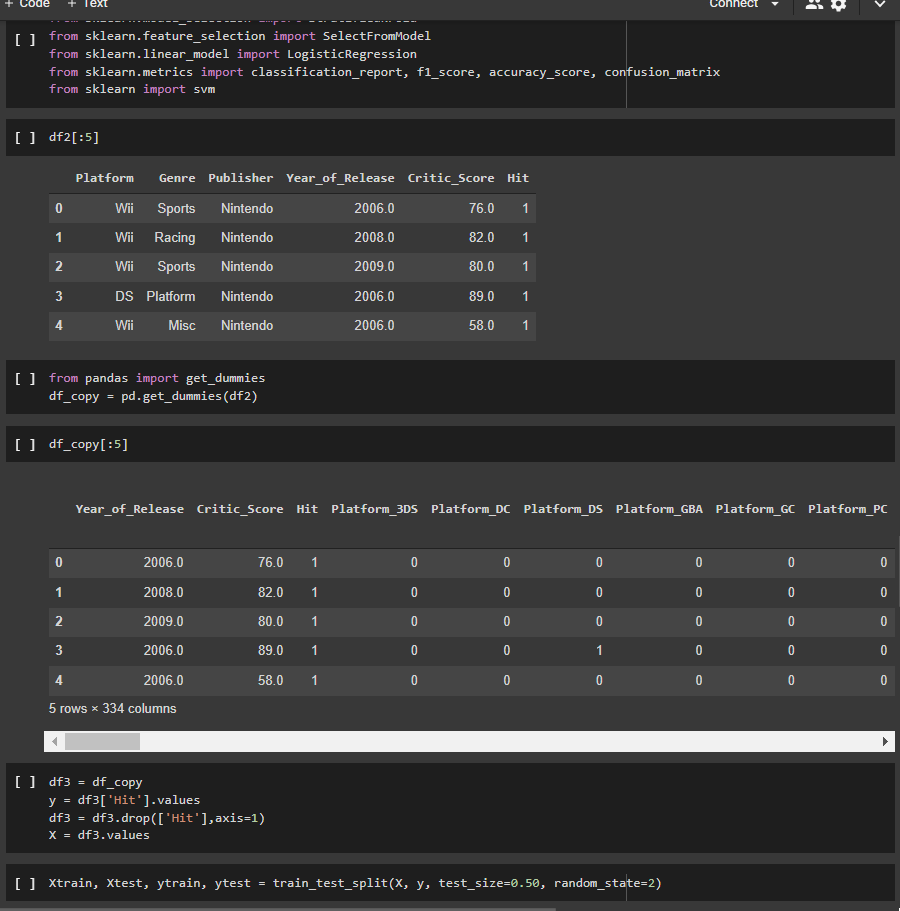
****

As expected, it seems **hits are mostly found near critic scores,** while non-hits can vary in scores but begin to lose presence in the high score ranges (as interpreted by the steepening regression curve near the 70’s).

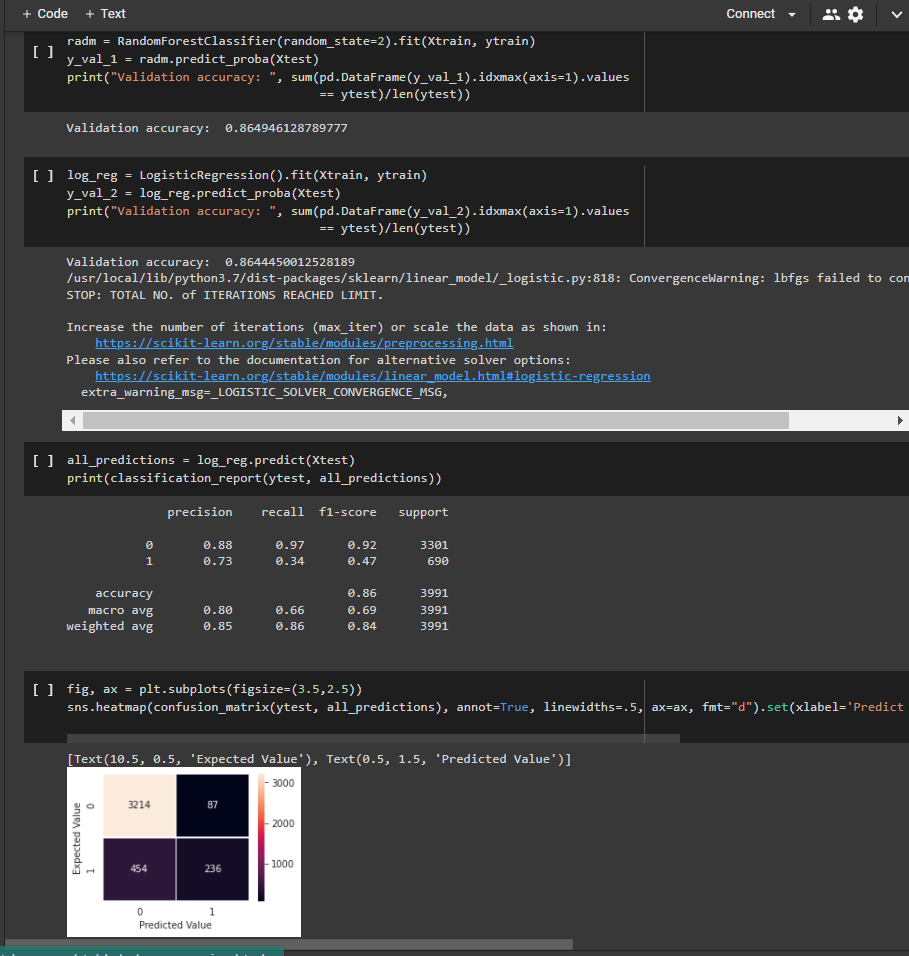
**PREDICTION MODEL**

For predicting the likelihood of a given game to reach sales of 1 million units or higher, referred to as “hit” games. Classification approach is applied to separate hits from non-hits.

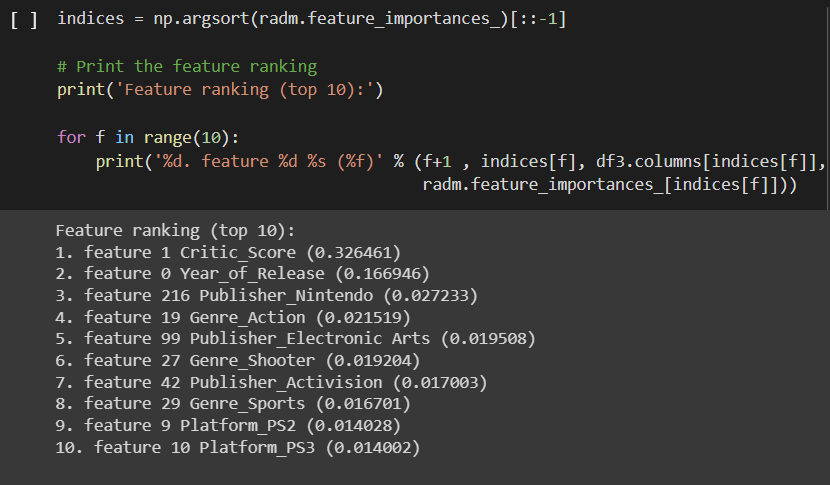
**Generating features and train/test splitting**



**Testing prediction accuracy with RFC and LR**

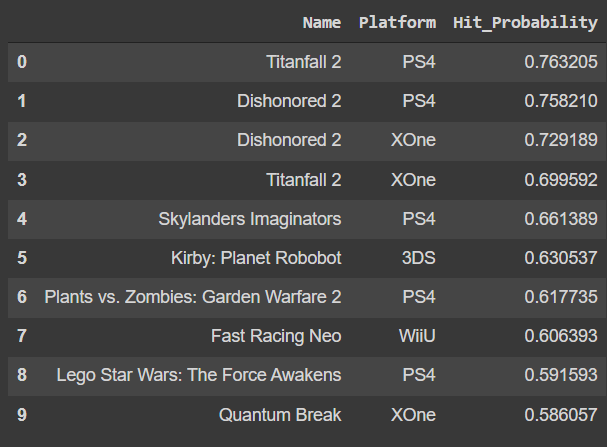


**Ranking feature performance**



**WHICH 2016 VIDEO GAMES CAN STILL BECOME HITS?**

**Video games with highest probability of becoming hits:**



**Video games with lowest probability of becoming hits:**

