**Analyzing Sales of Games on Steam**

**Executive Summary**

**Varisht Raheja**

**Shubham Sharma**

**Sobhagya Singh Jamwal**

**Hemanya Tyagi**

In this project analyzed the sales of video game on Steam. Steam is essentially an online channel which facilitates buying and selling of video games. In 2017, just in the US, video games grossed 36.9 billion dollars. This number increased to 43.4 billion dollars in 2018 indicating that video game is a lucrative industry. However, on Steam only 14% of games cross 50,000 ownership. Hence, we will be analyzing parameters affecting the success of game. In order to quantify the success of video game we took number of owners or downloads as the measure of success, as more the number of downloads, more the video game will earn, be it through pricing or in-app purchases or downloadable content.

We used the data available on Kaggle at <https://www.kaggle.com/nikdavis/steam-store-games> as the primary source of data. This data comprises of structured data, which includes numerical and categorical variables, and unstructured data, which includes description texts of each videogame and link to the icon image of the videogame. To get the variable DLC (or the downloadable content), we scraped the Steam website by using selenium. The data problem then becomes to find out relations between dependent variables and the number of owners. To infer these relationships, we first preprocessed the data. The number of owners in the data were available in bins. To convert these variables to continuous values, we applied min-max normalization in each bin using total number of reviews. To handle categorical data, we created dummy variables. Finally, we log-transformed the right skewed variables.

After preprocessing the structured data, we went on to analyze the text and image data. We performed topic modeling to find out the top 6 models in the descriptions. To achieve this, we used Latent Dirichlet Allocation using NLTK. We also extracted contrast and brightness of the icon images by using appropriate transformations. Finally, we regressed all the variables – brightness, contrast, probability allocated to topics, categorical and numerical predictors against the dependent variable – number of owners.

From the results, we inferred the variables that are significant and related to the number of owners. Some interesting insights that we can get from these results are: number of owners is positively related to factors such as achievements and leaderboards. That is, if we have more achievements to be unlocked in a game, the popularity of the game increases. As expected, the number of owners is negatively related to price. Moreover, all the topics that we extracted through text analysis are significant and negatively correlated with number of owners. This tells us that games that are have the most popular topics are not the most successful games. We found no significant relation between image features and number of owners.

We finally ran some more regressions and t-tests to compare design choices and go-to-market strategies. We concluded that Action genre outperforms other genres, publishing the game on all platforms generates more demand than publishing on just Windows. We also concluded that free videogames are usually in more demand than paid apps and having in-app purchases in the videogames does not lead to a decrease in demand.

**References:**

1. https://www.theesa.com/press-releases/u-s-video-game-sales-reach-record-breaking-43-4-billion-in-2018/