**Press Any Key to Start: Video Games Sales’ Revenue and Characteristics using Statistical Linear Regression Model and K-means Cluster Analysis**

**Kobi Wiseman**

School of Information, Syracuse University

Syracuse, NY 13244 USA

yvaisman@syr.edu

**Introduction**

The concept of gaming exists from human kind’s early ages. Whether it is fulfillment of the need to compete each other or the excitement of gambling, people always use to find their own ways to play with each other – from stones and sticks in early civilizations to board games and low-tech electronic devices in the mid-late twentieth century.

The invention of the personal computer (also known as PC) in 1980 brought numerous new and exciting opportunities for taking the gaming into the digital era. Thinking about a video game in present days, we can only imagine a high-definition, fast-pace and dense-detailed game which requires an up-to-date graphic processor. Early game developments seem so simple for us today, almost naive. However, for those who grew as teenagers during the 1980’s and 1990’s, iconic games such as Pong, Space Invaders and Super Mario ignited the imagination and used to be the main reason for getting up in the morning (and avoiding doing homework duties as much as they could).

Video games became a major key in the popular culture, as many video game franchises were shifted to the big-screen hits (Resident Evil, Tomb Raider and Call of Duty), and vice-versa (Robocop, Transformers and Star Wars). In terms of Psychology, the urge to compete each other is considered as pursuing the fourth level of importance in the Maslow’s Hierarchy of Needs – which is the ‘*Esteem Needs: prestige and feeling of accomplishment*’1. Considering the above, it is not surprising that video games have a major effect on a gamer’s personality and personal development. This effect, however, might have lethal consequences as well, as unfortunate events might occur when gamers play video games irresponsibly. The most well-known case is a Twenty-Eight years old South Korean male who played a video game for Fifty hours with a very few breaks in between and died after collapsing2.

This rapidly growing demand has not been overlooked by business entrepreneurs. Similar to many other areas under capitalist sphere, the video games industry is held by a few corporations which acquires smaller successful video game studios. Since video game production has an incredible potential of revenue on the one hand and a tremendous risk of losing money on the other hand, it is vital to perform a careful market research before moving to the production phase.

Using a video games sales database hosted by Kaggle, I pursue achieving two goals: (1) predicting the expected global revenue of a video game by Linear Regression Model, given a set of explanatory variables related to sales areas in the world, game genre and its platform, and (2) detecting common patterns for certain sub-groups in the data, using k-means Cluster Analysis.

Ultimately, I hope this work to provide a better understanding of the major considerations of both producing and marketing a potential video game.

**Metadata**

The examined dataset covers the top best-sellers video games between 1980 and 2020. The data consists of 16,598 rows (after 2 rows were omitted by the developer due to incomplete information) and 11 columns. The columns (features) description is listed below:

|  |  |  |
| --- | --- | --- |
| **Feature** | **Type** | **Description** |
| Rank | Integer | Ranking of overall sales |
| Name | String | Game’s title (table’s unique identifier) |
| Platform | String | Platform of the games on its release (i.e. PS4, PC etc.) |
| Year | Integer | Year of the game’s release |
| Genre | String | Genre related to the game (action, sports, strategy etc.) |
| Publisher | String | The company/studio to publish the game |
| NA\_Sales | Float | Sales in North America (in millions of USD) |
| EU\_Sales | Float | Sales in Europe (in millions of USD) |
| JP\_Sales | Float | Sales in Japan (in millions of USD) |
| Other\_Sales | Float | Sales in Other areas of the world (in millions of USD) |
| Global\_Sales | Float | Overall sales in millions of USD (sum of areal sales above) |

**Data Preprocessing**

Prior to using the data for modeling, some data issues need to be considered and handled: Missing values, extreme values and duplicate records.

Missing Values

|  |  |  |
| --- | --- | --- |
| **Feature** | **Number of missing values** | **replacement** |
| Year | 271 | The most frequent value: 2009 |
| Publisher | 58 | ‘unknown’ |

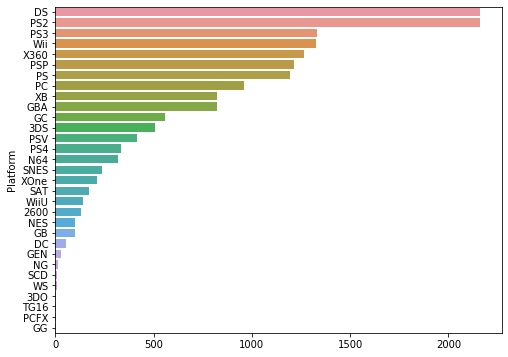
Extreme Values (Outliers) – to be detected using the Z-score approach:

|  |  |
| --- | --- |
| **Feature** | **treatment** |
| Year | The year information, even if considered as extreme, is true, and replacing it with any other measurement will make the results biased, therefore it will be retained. |
| NA\_Sales | Since the absolute revenues are directly related to the global sales, keeping it in that formant does not make sense. Thus, these variables are shifted to each area’s portion (in percentage) out of the global sales. |
| EU\_Sales |
| JP\_Sales |
| Other\_Sales |
| Global\_Sales | The revenue information, even if considered as extreme, is valid, and replacing it with any other measurement will make the results biased, therefore it will be retained (for the cluster analysis, it will be binned into categories, so the magnitude of those values will be tremendously depressed). |

Duplicate records: no duplicate rows were found in this data.

**Exploratory Data Analysis**

Descriptive statistics for numeric attributes and visualized distribution for categorical attributes is described below:

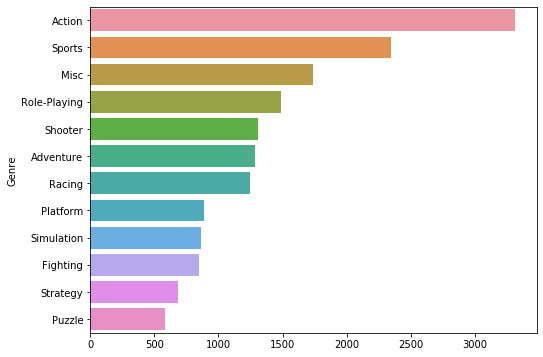
Video games range from 1980 to 2020. Plenty of changes in computers and game consoles systems technology have been occurred along years, so did in the popular culture. Given that, I expect this variable to be a major key in predicting the global sales of a given video game. Having most of its data being on the low portions among overall revenues, ‘other sales’ has usually low influence compared to the defined other three regions. Japan and North America both reach high portion of the sales for certain franchises, which implies each one the areas enjoy a separate market share. I am curious to see if that will be seen at the cluster analysis as well.

Platform Frequency:

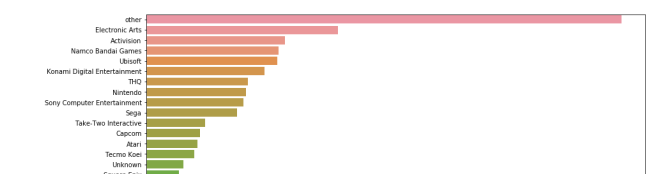
Nintendo DS and PlayStation 2 are by far the most common consoles in the past

40 years. Far from those, but with noticeable presence, we observe PlayStation 3,

Nintendo Wii, Xbox 360, PSP and PS with roughly 1200-1300 games each. PC, which at the distant past used to be the only source for gaming, is behind all the above, with less than 1000 games within the data.

Genre Frequency:

Action and Sports games are the most common genres among the best-sellers during the last 40 years. Strategy and Puzzles are the least frequent ones with only a little more than 500 games each.



Publisher Frequency (most frequent publishers are displayed):

The 'Other' category which covers all the video games studios with less than 50 games each, reaches over 3,000 games overall. Electronic Arts is the leading publisher with roughly 1,500 games, followed by Activision, Namco and Ubisoft with about 1,000 games each.

**Linear Regression Modeling**

Linear regression lies on the assumption that a dependent variable Y is determined by explanatory variables X1, …, Xn. Given certain data points, the model essentially generates the regression equation line: Yhat = B0 + B1X1 + … + BnXn , achieving the least squared distances between the observed data points and the predicted Y (‘Y Hat’) for each Xi. Using this model, we can predict the volume of increase or decrease of the dependent variable by shifting each one of the explanatory variables3.

Preparation

Conducting Linear regression based on our dataset, the ‘Global\_Sales’ attribute will be used as the dependent variable, without any transformations. All other variables in the data, excluding ‘Name’, will be used as the explanatory variables. The attributes ‘Platform’, ‘Genre’ and ‘Publisher’ are categorical-nominal variables – therefore will be represented as dummy variables for each one of its categories.

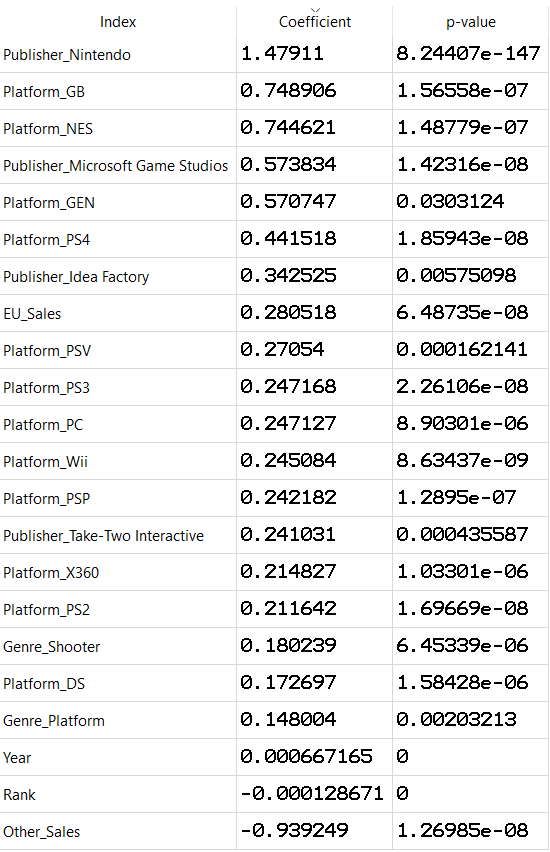
In order to avoid multicollinearity, a correlation matrix was built to examine high correlations between the explanatory variables. Two excessive correlations were found, between ‘JP\_Sales’ and ‘NA\_Sales’ (-0.71) and between ‘JP\_Sales’ and ‘EU\_Sales’ (-0.49). Since the correlation between ‘NA\_Sales’ and ‘EU\_Sales’ is not very high (-0.21), only the attribute ‘JP\_Sales’ will be excluded from the model.

The data set has been randomly split into 70/30 train-test groups.

Results

The full model running resulted an adjusted R-squared measurement of 0.31, which means 31% of the total variance of the global sales can be explained by this model. The mean square error (MSE) is 155.2 . This model, however, including many categories which are not 0.05 significant, meaning we cannot be confident that its coefficients are not part of the marginal error. For fine-tuning of the model, a restricted version of it was run, keeping only the significant categories.

This running maintains the same R-squared of 0.31 and the MSE has been increased to 640.2 . Although we normally pursue as lower MSE as possible, this model is business-wise easier to interpret and implement, having only 22 coefficients, not including the intercept. A screenshot of the results is shown below:



Each one of the categories on the left can be economically interpreted, I would like to emphasize a few of them:

* Publishing a game with Nintendo is predicted to gain $905,276 more revenue than Microsoft Game Studios.
* Each increase of one percent of Europe sale’s portion out of total sales is predicted to increase the overall revenue by $ 280.1K.
* Developing a game based on the PlayStation 4 console is predicted to provide $194,391 more revenue than a PC game, and $196,434 more than a game designed for the Wii platform.
* Each increase of one percent of other areas sale’s portion out of total sales is predicted to decrease the overall revenue by $ 939.2K.

**k-means Cluster Analysis**

k-means is an unsupervised algorithm which classifies data points within a given dataset by measuring their distances between each other. Given a pre-defined number of clusters, the algorithm attempts to optimize the classification by minimizing the distances within each group (finding the common factors) and maximizing the distances between the groups (making each group distinct). Choosing the appropriate number of clusters, this method allows us to view specific characteristics of sub-populations in our data4.

Preparation

For obtaining decent and meaningful results for this algorithm, a few decisions were made:

* In order to avoid the algorithm to be ‘drifted’ by extreme numeric values, all the numeric attributes have been binned to ten categories each.
* Since binning the ‘Rank’ attribute has practically the same meaning for binning the ‘global sales’ variable, the second one has been excluded from this model.
* ‘Publisher’ attribute has also been excluded from the model since it contains too many different values without any visible logical way to classify them.
* In order to avoid the algorithm to be overwhelmed by numerous dummy variables, the categorical attributes of ‘Platform’ and ‘Genre’ have been transformed into ‘business-tailored’ dummy variables; ‘Genre’ has been divided into ‘Adrenaline’ and ‘Logic’ by the different genres (Action, Adventure, Fighting, Platform, Racing, Shooter and Sports vs. Puzzle, Role-Playing, Simulation and Strategy accordingly), and each one of the top platform manufacturers (‘Nintendo’, ‘PlayStation’ and ‘Xbox’) has its own grouped platforms.
* In order to determine the appropriate number of clusters to be used, a Principal Component Analysis (PCA) was conducted in order to obtain the eigenvalue. The corresponding number of factors to an eigenvalue nearing 1 is four factors, therefore four clusters will be examined in this algorithm.

Results

The output of this algorithm is classifying each instance to a cluster number. In other words, each case in the data receives its corresponding cluster number, as determined by the model. Attaching the series of cluster numbers back to the original data allows us to cross-tab the frequencies of clusters with each one of the features. According to each relative frequency cross-tab and the resulted characteristics of each cluster are the following:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Rank** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **Cluster** | **1** | 0.29 | 0.25 | 0.2 | 0.13 | 0.09 | 0.04 | 0 | 0 | 0 | 0 |
| **2** | 0.02 | 0.04 | 0.05 | 0.09 | 0.08 | 0.1 | 0.11 | 0.14 | 0.17 | 0.21 |
| **3** | 0 | 0.02 | 0.06 | 0.1 | 0.13 | 0.17 | 0.19 | 0.16 | 0.11 | 0.07 |
| **4** | 0 | 0.02 | 0.03 | 0.05 | 0.07 | 0.06 | 0.1 | 0.15 | 0.23 | 0.29 |

\* lower rank means higher revenue for a certain game.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Year** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **Cluster** | **1** | 0 | 0.01 | 0.01 | 0.03 | 0.11 | 0.21 | 0.22 | 0.28 | 0.13 | 0 |
| **2** | 0 | 0.01 | 0.01 | 0.11 | 0.1 | 0.07 | 0.24 | 0.28 | 0.18 | 0 |
| **3** | 0.02 | 0 | 0 | 0.02 | 0.07 | 0.22 | 0.36 | 0.27 | 0.04 | 0 |
| **4** | 0 | 0 | 0 | 0.01 | 0.03 | 0.08 | 0.22 | 0.39 | 0.27 | 0 |

\* category 1 refers to the oldest games in data, starting with 1980, 10 is the newest, reaching 2020.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **NA\_Sales** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **Cluster** | **1** | 0.01 | 0.02 | 0.07 | 0.13 | 0.31 | 0.22 | 0.1 | 0.09 | 0.04 | 0.01 |
| **2** | 0.93 | 0.04 | 0.02 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0 | 0 | 0 | 0 | 0.07 | 0.08 | 0.1 | 0.27 | 0.18 | 0.3 |
| **4** | 0.67 | 0.08 | 0.07 | 0.07 | 0.1 | 0 | 0 | 0 | 0 | 0 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **EU\_Sales** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **Cluster** | **1** | 0.09 | 0.11 | 0.2 | 0.39 | 0.14 | 0.05 | 0.02 | 0 | 0 | 0 |
| **2** | 0.96 | 0.04 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0.51 | 0.17 | 0.19 | 0.12 | 0.02 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0 | 0 | 0 | 0.04 | 0.18 | 0.07 | 0.12 | 0.17 | 0.12 | 0.31 |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  | **JP\_Sales** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **Cluster** | **1** | 0.81 | 0.05 | 0.05 | 0.05 | 0.03 | 0.01 | 0 | 0 | 0 | 0 |
| **2** | 0 | 0 | 0 | 0 | 0.01 | 0.02 | 0.03 | 0.02 | 0.02 | 0.9 |
| **3** | 0.97 | 0.01 | 0.01 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0.97 | 0.01 | 0 | 0 | 0.01 | 0 | 0 | 0 | 0 | 0 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Other\_Sales** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **Cluster** | **1** | 0.54 | 0.41 | 0.03 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 0.99 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0.86 | 0.14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0.52 | 0.31 | 0.12 | 0.04 | 0 | 0 | 0 | 0 | 0 | 0 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Nintendo** | **0** | **1** |
| **Cluster** | **1** | 0.73 | 0.27 |
| **2** | 0.64 | 0.36 |
| **3** | 0.46 | 0.54 |
| **4** | 0.8 | 0.2 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **PlayStation** | **0** | **1** |
| **Cluster** | **1** | 0.47 | 0.53 |
| **2** | 0.45 | 0.55 |
| **3** | 0.78 | 0.22 |
| **4** | 0.73 | 0.27 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Xbox** | **0** | **1** |
| **Cluster** | **1** | 0.83 | 0.17 |
| **2** | 0.98 | 0.02 |
| **3** | 0.8 | 0.2 |
| **4** | 0.9 | 0.1 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **PC** | **0** | **1** |
| **Cluster** | **1** | 0.97 | 0.03 |
| **2** | 1 | 0 |
| **3** | 0.98 | 0.02 |
| **4** | 0.57 | 0.43 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Adrenaline** | **0** | **1** |
| **Cluster** | **1** | 0.25 | 0.75 |
| **2** | 0.42 | 0.58 |
| **3** | 0.31 | 0.69 |
| **4** | 0.38 | 0.62 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Logic** | **0** | **1** |
| **Cluster** | **1** | 0.84 | 0.16 |
| **2** | 0.69 | 0.31 |
| **3** | 0.81 | 0.19 |
| **4** | 0.71 | 0.29 |

Cluster 1: top-sellers games franchises, released during the last decade, made most of its sales by PlayStation platform. Mostly related to ‘Adrenaline’ Games.

Cluster 2: most of its revenue derived from Japan and derived from the PlayStation platform. Strongly identified with ‘Logic’ games related to other clusters.

Cluster 3: a ‘good place in the middle’ in terms of sales ranking, dated around years of 2000-2010 and most of its revenue come from North America. Strongly identified with the Nintendo Console and ‘Adrenaline’ genres.

Cluster 4: low-selling games, mostly from the last decade. Get most of its sales derive from Europe and related to PC titles.

\* clusters centroids are changed after each running, so cluster description might be a little different compared to those which appear on slides and video.

**Conclusion**

Based on the findings above, we can conclude that Linear modeling is a fair methodology for predicting future franchises’ revenues. Predictors such as publishing the game with a certain publisher over another, or preferring to designate the game to a specific platform over another, could make a great difference in terms of sales rate. Focusing on specific areas of the world could also be a key to success.

Conducting the unsupervised model of the k-means cluster analysis, I was tremendously surprised by its efficiency of pointing out various target audiences and characterizing each one of them with a series of bullets based on the data’s different features. Following my personal experience with this algorithm, I am not surprised to hear about its success dealing with more complexed problems in even more dynamic industries.

Having said that, this analysis has plenty of unfulfilled potential. The linear regression model obtained a fair R-squared measurement of 0.31, but adding some informative features, such as video game’s price and whether it was on promotion or not, could have, in my opinion, improve the model’s prediction power. Moreover, I believe that dividing the revenue by those given areas is too rough. As both China and South Korea considered superpowers in terms of gaming markets, extracting the data for each one of the countries individually could benefit more accurate output of each region’s portion of revenue out of total.

Overall, this kind of analysis could fairly be useful for big game studios when deciding on next game development. Considering both opportunities and potential risks involved in each production, I believe that game manufacturers cannot ignore this possibility of utilizing Statistics and Data mining to their own decision-making process.

**External References**

[1 – Maslow’s Hierarchy of needs](https://www.simplypsychology.org/maslow.html#:~:text=Maslow's%20hierarchy%20of%20needs%20is,hierarchical%20levels%20within%20a%20pyramid.&text=From%20the%20bottom%20of%20the,esteem%2C%20and%20self%2Dactualization)

[2 – 28 years old South-Korean male who died of overplaying](http://news.bbc.co.uk/2/hi/technology/4137782.stm#:~:text=A%20South%20Korean%20man%20has,according%20to%20South%20Korean%20authorities)

[3 – Concept of Linear regression](https://www.statisticssolutions.com/what-is-linear-regression/)

[4 – Concept of k-means cluster analysis](https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a)

**Relevant links for this project**

Selected data web page:

<https://www.kaggle.com/gregorut/videogamesales>

Video presentation url:

<https://youtu.be/hu9kg7d4ya0>