RELEASING A SUCCESSFUL VIDEO GAME

Martell Tardy

Noor Bahr Al Uloom

October 13, 2018

INTRODUCTION

Valve Corporation is an American video game developer, publisher and digital distribution. The company headquartered in Bellevue, Washington. It is the developer of the software distribution platform Steam and the Half-Life, Counter-Strike, Portal, Day of Defeat, Team Fortress, Left 4 Dead, and Dota 2 games. At Valve Corporation we have found success with the creation of Steam[1] in 2003, a gaming platform and marketplace, and The International[2];an annual Dota 2[3] eSports[4] tournament, started in 2011.

The Steam community, that has evolved out of our platform alongside our growing support at our tournaments, has laid the foundation for Valve to once again enter the world of video game production. At Valve we seek to create a company that influences the video gaming world from the production of the content to the competitions that surround it. During Valve's five-year hiatus significant effort has been made to understand what motivates players to engage with games, compete in tournaments,

[1] **Steam:** A digital distribution platform for video games developed by Valve Corporation that offers digital rights management, matchmaking servers, video streaming, and social networking services.

[2] **The International:** is an annual Dota 2 eSports tournament hosted by Valve Corporation, the game's developer.

[3] **Dota 2:** is a free-to-play multiplayer online battle arena video game developed and published by Valve Corporation. The game is the stand-alone sequel to Defense of the Ancients, which was a community-created mod for Blizzard Entertainment's Warcraft III: Reign of Chaos and its expansion pack, The Frozen Throne.

[4] **eSports:** are a form of competition using video games. Most commonly, eSports take the form of organized, multiplayer video game competitions, particularly between professional players.

and ultimately provide the best user experience. While individuals may have varying motivations for playing or abandoning a game, we at Valve are looking to find that sweet spot in the production of our new game announced to release in the last quarter of 2018. In this study, we analyze past gaming sales across the gaming industry, review gaming statistics from the Steam platform, and weigh player retention from eSport competitions around the world to aid in the design and user experience of this game. Through these datasets we plan to identify substantial trends in gameplay[5], online gaming frequency, tournament popularity (About Us: Valve Corporation, n.d.) (Peterson, 2013).

1. PROBLEM STATEMENT

Valve Corporation has presented the following business problem; can Valve Corporation release again a profitable and popular video game in their 2018 production? To solve this business problem Valve Corporation must release a game reflecting the key features of prior released video games that have already received profitable returns, positive user and critic ratings, and strong retention at tournaments.

The Data Science Department at Valve Corporation (Noor Bahr Al Uloom and Martell Tardy) will take on the task of solving this business problem with an analytical solution using descriptive analysis, data visualizations, and various predictive models from collected historical data from the video game industry. First the business problem will be reframed as the analytical problem of; can the use of historical data identify the key features of past profitable and popular video games and predict the characteristics of a similar deliverable for the Valve Corporation to release in the last quarter of 2018? The desired result is a well-defined cluster and game profile for the video game experience Valve Corporation should release as their next video game development. As a result of these findings, our recommendations will be operationalized by Valve Corporation through the production of their next video game.

1.1. METHODOLOGY

Valve Corporation's Data Science Department has decided to address this task by exploring three objectives that center around variable observations assumed to effect maximized sales, user interaction, ratings, and tournament participation historically.

I. *Objective One* is to identify which combination of genre[9] characteristics and gameplay, correlates to profitable and popular video games historically.

---

[5] **gameplay:** is the specific way in which players interact with a game, and in particular with video games. Gameplay is the pattern defined through the game rules, connection between player and the game, challenges and overcoming them, plot and player's connection with it.

II. *Objective two* is to determine if a video game's frequency in tournaments, player participation, and amount of award money at tournaments, correlates to profitable and popular video games historically.
III. *Objective three* is to determine if profitable sales historically of a video game, correlates to high tournament participation and certain genres and gameplay experiences historically.

1.2. DRIVERS & OUTPUTS

The output of this project is the identification of the key features found in a profitable and popular video game historically. This output is explored visually in Table 1.

Tab.1. Drivers vs. Outputs

Drivers:	**Outputs:**
High Historical Global Sales	Profitable Returns
High Frequency of Tournaments	Strong Retention
High Scores on Scales of 1- 100	Positive User and Critic Ratings
Historical Data from Four Datasets Collected	Key Features

1.3. ASSUMPTIONS

The assumptions of this analytical problem are the output correlations between variables of historically profitable and popular video games, Valve Corporation's user base, the measurable maximization of specified categorical and ratio variables that produce a prototype of the ideal video game.

1.4. KEY METRICS OF SUCESS

The key metrics of success for this analytical project is a well-defined medoid and player description for the video game experience Valve Corporation should release as their next video game development. A successful medoid will contain a cluster with high profitability, high frequency of tournaments and players, positive user and critic ratings, and valuable insight into which features of the game experience are typical amongst those video games. High profitability will be considered a profitable game. High frequency of tournaments and players, and positive user and critic ratings will be considered a popular game.

1.5. STAKEHOLDERS

The stakeholders of this project are the Development Team at Valve Corporation since the company is a flat organization[6]. The Development Team is comprised of the game designers, artists, programmers, level designers, sound engineers, and testers. The game designers' design gameplay and therefore, conceive and design the rules and structure of a game. The game artists create the art within the video game, such as environmental backdrops or terrain images and user interfaces. The programmers implement the game's starting codebase and overview future development and programmer allocation on individual modules. The level designers create levels, challenges or missions for the video games using a specific set of programs. The sound engineers are responsible for sound effects and sound positioning. The testers analyze video games to document software defects as part of a quality control. The interest in the business problem for the Development Team is the gameplay experience that will be recommended at the end of this project by the Data Science Department.

2. DESCRIPTION OF DATA

In the process of gathering data for this project, four preexisting datasets were collected in Excel format. Each dataset provided insight into an aspect of answering our problem statement. The Esports Earnings Dataset (EED) provided variables describing how many individuals played a video game in a competitive format, the title of the game involved in the competition, and the prize money associated with a tournament. The *EED* dataset was collected from the e-sport website, (E-Sports Earnings, 2018). The Games Achievements Players 2018-07-01 (GAP) dataset provided variables describing unique player activity on the online Steam platform by video game title. The *GAP* dataset was collected from the Steam platform and verified by a second party within the gaming industry, (Galyonkin, 2018). The Managerial and Decisions Economics 2013 Video Games Dataset (MDE) provided gameplay variables describing which of the genre types apply to a video game, the ESRB Rating associated with a video gaming experience, and critic and user scores [7]. The *MDE* dataset was collected from Portsmouth Research Portal, (Cox, 2015). The Video Games Sales as at 22 Dec 2016 (VDS*)* dataset also provided gameplay variables describing which of the genre types apply to a video game, the ESRB Rating associated with a video gaming experience,

[6] **flat organization** (also known as horizontal organization) has an organizational structure with few or no levels of middle management between staff and executives.

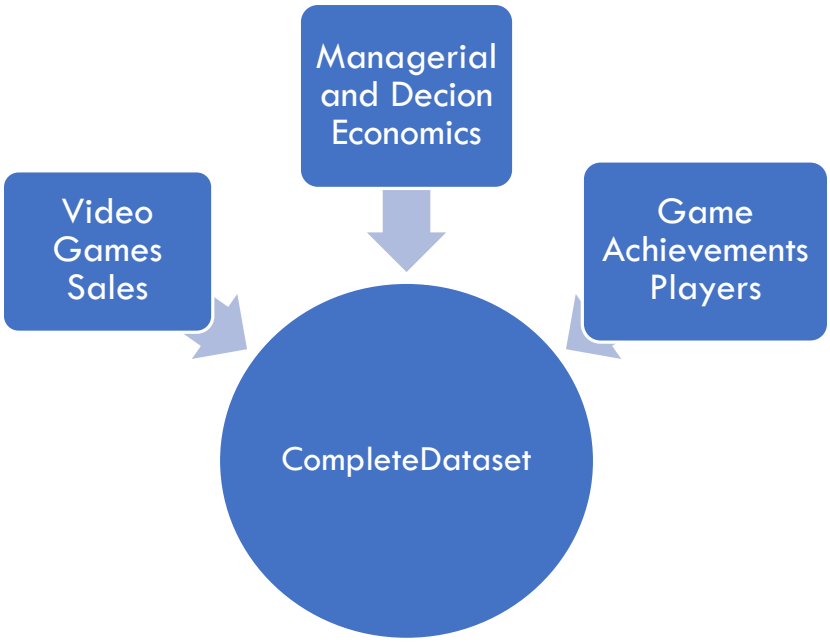
[7] **sequel:** refers to a video game that continues the story of, or expands upon, some earlier work. As software-development costs have increased, sequels have become increasingly important for the video-game industry, as they provide a way to resell a product, reusing code and graphics.

and critic and user scores[8]. The *VDS* dataset was collected from the Kaggle website, (Kirubi & Smith, 2016).

3. OVERVIEW OF THE DATA

After combining all the information from our four datasets into one file, using Excel, the decision to omit the variables from the *GAP* dataset was decided. The decision was found due to its lack of completeness. The variable *Name,* referring to the title of a video game, did not include all video games offered on the Steam platform. As a result, the final dataset for this project contained 349 rows of variables, referring to video game titles, and 144 columns referring to unique information about each video game. This final dataset was named ***CompleteDataset***.

Fig.1. Dataset Merge



3.1. SELECTED FEATURES

A total of 5,472 observations are recorded in the *CompleteDataset*. The observations within this dataset contain subgroups of variables organized into four major categories: tournaments, gameplay, scores, and sales.

[8] **sequel:** refers to a video game that continues the story of, or expands upon, some earlier work. As software-development costs have increased, sequels have become increasingly important for the video-game industry, as they provide a way to resell a product, reusing code and graphics.

3.1.1. TOURNAMENT CATEGORY

There are three variables in this category and 516 observations in total.

The variables selected include the following:

- TournamentMoneyAwarded – total prize money associated with a tournament
- TournamentTotalPlayers - how many individuals played a video game in a tournament
- TotalTournaments – total number of tournaments held for an individual video game

The objective of the analysis conducted within this category is to identify if there are measurable correlations between these variables and the game title associated with them. This category provides insight into objective one.

3.1.2. GAMEPLAY CATEGORY

There are 23 variables in this category and 3956 observations in total.

The variables selected include the following:

- ESRBRating – assigned age and content rating
- FirstPersonPerspective - the view a player is given while playing the video game
- Strategy - these games require players to use carefully developed strategy and tactics to overcome challenges.
- Fighting - focus the action on combat, and in most cases, hand-to-hand combat.
- Shooter - players use weapons to engage in the action, with the goal usually being to take out enemies or opposing players.
- Sports - simulate playing a sport.
- Racing - players race against another opponent or the clock.
- RPG - players assume the roles of characters in a fictional setting. Players take responsibility for acting out these roles within a narrative, either through literal acting or through a process of structured decision-making of character development.
- Card - includes traditional games like chess, checkers.
- Adventure - players usually interact with their environment and other characters to solve puzzles with clues to progress the story or gameplay.
- Platformer - the game's character interacts with platforms (usually running, jumping, or falling) throughout the gameplay.
- Beat.emUp - focus on combat, but instead of facing a single opponent, players face wave after wave of enemies.
- TBS - games that gives players a length of time (or turn) in which to take action. But like an RTS game, the genre can include games that are not exclusively turn-based.
- Puzzle - take place on a single screen or playfield and require the player to solve a problem to advance the action.

- Simulator - games designed to emulate real or fictional reality, to simulate a real situation or event.
- Action - games where the player is in control of and at the center of the action, which is mainly comprised of physical challenges players must overcome.
- Warfare - focuses gameplay on map-based tactical or strategic warfare.
- Fantasy - genre of speculative fiction set in a fictional universe, often without any locations, events, or people referencing the real world.
- ScienceFiction - genre of speculative fiction, typically dealing with imaginative concepts such as advanced science and technology, spaceflight, time travel, and extraterrestrial life.
- Comedy – games with a comedic tone or textual interaction with the user.
- Horror - use mature themes and subject matter to portray grisly and gruesome settings (many of these games use blood and gore and are intended only for mature audiences). Such titles deliver nail-biting excitement amplified by a key game mechanic: limited resources like ammunition or finite weapons.
- Party - role-playing games where a player leads a party of adventurers in first-person perspective.
- Stealth - stress cunning and precision to resolve game challenges, and while other action or combat may help players accomplish the goal, stealth games usually encourage players to engage in the action covertly.

The objective of the analysis conducted within this category is to identify if there are measurable correlations between these variables and the game title associated with them. This category provides insight into objective two.

3.1.3. SCORES CATEGORY

There are 2 variables in this category and 344 observations in total.

The variables selected include the following:

- CriticScore – *score on scale 1-100 provided by game testers within gaming industry*
- UserScore *-score on scale 1 -100 provided by users of the video game*

The objective of the analysis conducted within this category is to identify if there are measurable correlations between these variables and the game title associated with them. This category provides insight into objective one.

3.1.4. SALES CATEGORY

There are three variables in this category and 516 observations in total.

The variables selected include the following:

- %Y – *release date of video game*
- NA_Sales – *sales in North American by millions units sold*

- Global_Sales – *total sales globally by millions units sold*

The objective of the analysis conducted within this category is to identify if there are measurable correlations between these variables and the game title associated with them. This category provides insight into objective three.

4. DESCRIPTION OF TRANSFORMATION OF DATA

After combining all the information from our four datasets into one file there were 349 rows of variables, referring to video game titles and 144 columns referring to unique information about each video game. A visual of what occurred during the preprocessing in Excel can be seen in Table 2.

Tab.2. Preprocessing in Excel Software

Variable	**Issue**	**Action**	**Result**
Release	NA	Deletion (Row)	148 Rows Deleted
Global_Sales	NA	Deletion (Row)	23 Rows Deleted
TotalTournaments	NA	Deletion (Row)	6 Rows Deleted
NA_Sales	NA	Imputation (Mode)	10 Instances Changed to 0.1
UserScore	Not Scaled 1-100; NA	Multiplied by 10; Imputation (Mean)	Data Range 1-100; 38 Instances Converted to 64
CriticScore	Not Scaled 1-100; NA	Multiplied by 10; Imputation (Mean)	Data Range 1-100; 27 Instances Converted to 83
TournamentMoneyAwarded	NA	Imputation (Min.)	17 Instances Converted to 25
TournamentTotalPlayers	NA	Imputation (Min.)	18 Instances Converted to 1
Genre and Manufacturer Based Variables	Duplicates; NA	Deletion (Column)	317 Columns Deleted

The result after all filtering, filling in of missing data using imputation, and the reducing of the number of dimensions in the data by deletion was the final dataset containing 171 rows of variables, referring to video game titles and 32 columns referring to unique information about each video game. A summary of dataset CompleteDataset can be seen in Figure 2 using the summary() function.

Fig.2. Summary of CompleteDataset

```
> summary(CompleteDataset)
  GameTitle              %Y         UserScore      CriticScore     ESRBRating    NA_Sales      Global_Sales
 Length:171         Min.   :1994   Min.   :26.00   Min.   : 60.00   E:56       Min.   :0.000   Min.   : 0.010
 Class :character   1st Qu.:2004   1st Qu.:64.00   1st Qu.: 81.00   T:74       1st Qu.:0.055   1st Qu.: 0.245
 Mode  :character   Median :2008   Median :75.00   Median : 83.00   M:41       Median :0.380   Median : 1.010
                    Mean   :2008   Mean   :72.34   Mean   : 83.29              Mean   :1.168   Mean   : 2.494
                    3rd Qu.:2013   3rd Qu.:82.00   3rd Qu.: 86.50              3rd Qu.:1.415   3rd Qu.: 3.625
                    Max.   :2016   Max.   :91.00   Max.   :100.00              Max.   :9.040   Max.   :14.980
 FirstPersonPerspective  Strategy        Fighting        Shooter         Sports          Racing
 Mode :logical          Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:128              FALSE:157       FALSE:111       FALSE:128       FALSE:133       FALSE:158
 TRUE :43               TRUE :14        TRUE :60        TRUE :43        TRUE :38        TRUE :13
 NA's :0                NA's :0         NA's :0         NA's :0         NA's :0         NA's :0


    RPG             Card          Adventure       Platformer      Beat.emUp          TBS            Puzzle
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:169       FALSE:170       FALSE:158       FALSE:163       FALSE:170       FALSE:167       FALSE:154
 TRUE :2         TRUE :1         TRUE :13        TRUE :8         TRUE :1         TRUE :4         TRUE :17
 NA's :0         NA's :0         NA's :0         NA's :0         NA's :0         NA's :0         NA's :0


 Simulator         Action         Warfare         Fantasy        ScienceFiction    Comedy          Horror
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:154       FALSE:83        FALSE:167       FALSE:170       FALSE:140       FALSE:170       FALSE:165
 TRUE :17        TRUE :88        TRUE :4         TRUE :1         TRUE :31        TRUE :1         TRUE :6
 NA's :0         NA's :0         NA's :0         NA's :0         NA's :0         NA's :0         NA's :0


   Party          Stealth        TournamentMoneyAwarded TournamentTotalPlayers TotalTournaments
 Mode :logical   Mode :logical   Min.   :      25       Min.   :   1.00        Min.   :   1.00
 FALSE:170       FALSE:170       1st Qu.:    5500       1st Qu.:   3.00        1st Qu.:   2.00
 TRUE :1         TRUE :1         Median :   46000       Median :  13.00        Median :   5.00
 NA's :0         NA's :0         Mean   :  514210       Mean   :  84.09        Mean   :  41.34
                                 3rd Qu.:  213073       3rd Qu.:  57.50        3rd Qu.:  22.00
                                 Max.   :12088428       Max.   :1621.00        Max.   :1265.00
>
```

From the summary in Figure 2, we can see the video games range from 1994 to 2016 in release year. The average UserScore is 75 and the average CriticScore is 83 out of 100. A majority of the video games have an ESRBRating of T, for Teen. North American Sales (NA_Sales) average at 0.380 and Global Sales (Global_Sales) at 1.010. Less than half (33.5%) of the video games are in the first-person perspective. Most of the gameplay observations belong to the genres of Action at 88 instances and Fighting at 60 instances. Shooter gameplay comes in at third place at 43 instances. On average, a tournament prize was around $46,000 dollars with a max payout of $12,088,428.00 dollars. On average, a tournament saw a total of 13 players with a max turnout of 1,621. On average, total tournament count for an individual game was 5 times, with a max of 1,265 times. It will be interesting to see which games saw those extremely high prizes and player interaction.

5. ANALYSIS OF DATA

To analyze the CompleteDataset R Studio was the tool used. Within R the dataset was first examined for NAs using the x[!complete.cases()] function. No NAs

were found in the dataset. Next, in order to prepare for the predictive model, clustering, the variable GameTitle was removed and transformed into row.names. This transformation was saved as the dataset Complete. The Complete dataset then had all characters <chr> changed to factors <fctr> and all integer <int> changed to double <dbl> using the mutate_if() function. All logical <lgl> instances were untouched. Finally, the row.names were added back to the dataset, since the dplyr() package removes them during the use of the mutate_if() function. This transformation was saved as dataset Complete3. A visual of Complete is available in Figure 3, through use of the glimpse() function.

Fig. 3. Glimpse() of Complete3 Dataset

```
> glimpse(Complete3)
Observations: 171
Variables: 31
$ X.Y                     <dbl> 2014, 1999, 2005, 2011, 2002, 2011, 2013, 2005, 2006, 2011, 2013, 2015, 2008, 201...
$ UserScore               <dbl> 86, 64, 77, 56, 89, 78, 64, 82, 68, 74, 69, 64, 84, 76, 83, 77, 75, 68, 81, 84, 5...
$ CriticScore             <dbl> 95, 92, 81, 70, 89, 75, 83, 80, 80, 84, 85, 73, 87, 85, 83, 80, 83, 77, 89, 94, 8...
$ ESRBRating              <fctr> E, T, T, E, T, T, T, T, T, M, M, M, T, T, T, T, T, T, T, M, M, M, M, M, T, T, M,...
$ NA_Sales                <dbl> 3.27, 0.01, 0.01, 0.01, 0.02, 0.05, 0.02, 0.39, 0.01, 4.46, 1.35, 0.71, 0.36, 0.0...
$ Global_Sales            <dbl> 7.55, 0.09, 0.38, 0.06, 0.09, 0.11, 0.04, 0.56, 0.03, 7.32, 3.59, 2.10, 0.57, 0.0...
$ FirstPersonPerspective  <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, FA...
$ Strategy                <lgl> FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA...
$ Fighting                <lgl> TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE,...
$ Shooter                 <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, FA...
$ Sports                  <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS...
$ Racing                  <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS...
$ RPG                     <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS...
$ Card                    <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS...
$ Adventure               <lgl> FALSE, TRUE, TRUE, TRUE, FALSE, TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FA...
$ Platformer              <lgl> TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE...
$ Beat.emUp               <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS...
$ TBS                     <lgl> FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE...
$ Puzzle                  <lgl> FALSE, TRUE, TRUE, TRUE, FALSE, TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FA...
$ Simulator               <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, FAL...
$ Action                  <lgl> TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, TRU...
$ Warfare                 <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS...
$ Fantasy                 <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS...
$ ScienceFiction          <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS...
$ Comedy                  <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS...
$ Horror                  <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS...
$ Party                   <lgl> TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE...
$ Stealth                 <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS...
$ TournamentMoneyAwarded  <dbl> 1370788.06, 666929.39, 45130.00, 1100.00, 52360.00, 160.00, 410.00, 91715.22, 185...
$ TournamentTotalPlayers  <dbl> 1058, 261, 41, 11, 22, 3, 6, 1, 1, 1, 196, 2, 3, 39, 63, 28, 3, 14, 41, 414, 769,...
$ TotalTournaments        <dbl> 821, 8, 2, 3, 6, 1, 2, 12, 4, 7, 144, 1, 1, 9, 15, 11, 1, 5, 17, 85, 41, 51, 74, ...
>
```

5.1. CLUSTERING (PAM: K-MEDOIDS)

5.1.1 CALCULATING DISTANCE

In order for a yet-to-be-chosen algorithm to group observations together, we first need to define some notion of (dis)similarity between observations. A popular choice for clustering is Euclidean distance. However, Euclidean distance is only valid for continuous variables, and thus is not applicable here. In order for a clustering algorithm to yield sensible results, we have to use a distance metric that can handle mixed data types. In this case, we will use something called Gower distance. The Gower distance

fits well with the k-medoids algorithm. k-medoid is a classical partitioning technique of clustering that clusters the data set of n objects into k clusters known a priori. To execute Gower distance, we used the daisy() function (r, 2016). There is a visual of the Gower distance on Complete3 dataset in Figure 4.

Fig. 4. Summary of Gower Distance

```
> summary(gd)
14535 dissimilarities, summarized :
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0005962 0.3266100 0.4205800 0.3991200 0.4968400 0.7688600
Metric :  mixed ;  Types = I, I, I, N, I, I, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, I, I, I
Number of objects : 171
>
```

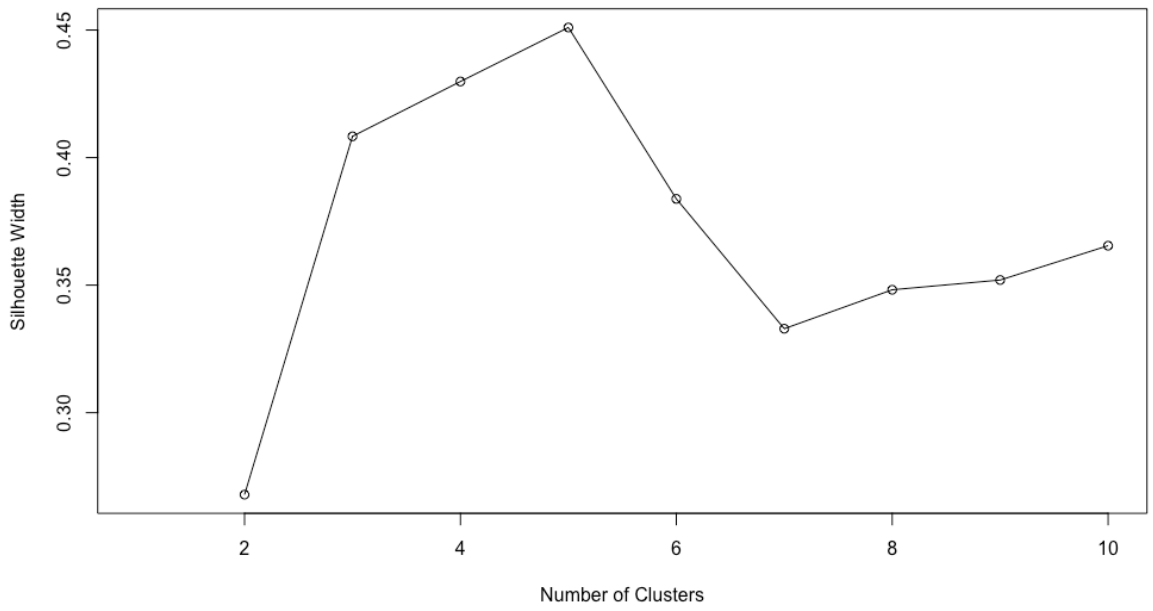
5.1.2 CHOOSING A CLUSTERING ALGORITHM

Now that the distance matrix has been calculated, it is time to select an algorithm for clustering. While many algorithms that can handle a custom distance matrix exist, partitioning around medoids (PAM) will be used here. If you know the k-means algorithm, this might look very familiar. In fact, both approaches are identical, except k-means has cluster centers defined by Euclidean distance (i.e., centroids), while cluster centers for PAM are restricted to be the observations themselves (i.e., medoids).

5.1.3 SELECTING THE NUMBER OF CLUSTERS

A variety of metrics exist to help choose the number of clusters to be extracted in a cluster analysis. We will use silhouette width() function, an internal validation metric which is an aggregated measure of how similar an observation is to its own cluster compared its closest neighboring cluster. The metric can range from -1 to 1, where higher values are better. After calculating silhouette width for clusters ranging from 2 to 10 for the PAM algorithm, we see that 5 clusters yields the highest value of 0.45. A plot of the Silhouette Width on Complete3 dataset is available in Figure 5.

Fig. 5. Silhouette Width Plot of *K* Clusters



5.1.4 CLUSTER INTERPRETATION

After running the algorithm and selecting 5 clusters, we can interpret the clusters by running summary() function on each cluster. Figures 6-10 are the summary of the five clusters.

Fig.6. Summary of Cluster 1

```
> results$the_summary
[[1]]
      X.Y          UserScore      CriticScore     ESRBRating    NA_Sales       Global_Sales
 Min.   :1996   Min.   :53.00   Min.   : 64.00   E: 2       Min.   :0.0100   Min.   : 0.010
 1st Qu.:2004   1st Qu.:64.00   1st Qu.: 80.25   T:54       1st Qu.:0.0400   1st Qu.: 0.140
 Median :2010   Median :76.00   Median : 83.00   M: 6       Median :0.2550   Median : 0.550
 Mean   :2009   Mean   :74.63   Mean   : 82.84              Mean   :0.6574   Mean   : 1.367
 3rd Qu.:2013   3rd Qu.:82.00   3rd Qu.: 85.00              3rd Qu.:0.8925   3rd Qu.: 1.370
 Max.   :2016   Max.   :89.00   Max.   :100.00              Max.   :6.6200   Max.   :12.840
 FirstPersonPerspective  Strategy        Fighting        Shooter          Sports          Racing
 Mode :logical          Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:62               FALSE:62        FALSE:3         FALSE:62        FALSE:60        FALSE:62
 NA's :0                NA's :0         TRUE :59        NA's :0         TRUE :2         NA's :0
                                        NA's :0                         NA's :0


    RPG             Card          Adventure       Platformer      Beat.emUp          TBS            Puzzle
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:62        FALSE:61        FALSE:58        FALSE:56        FALSE:61        FALSE:60        FALSE:55
 NA's :0         TRUE :1         TRUE :4         TRUE :6         TRUE :1         TRUE :2         TRUE :7
                 NA's :0         NA's :0         NA's :0         NA's :0         NA's :0         NA's :0


 Simulator         Action         Warfare         Fantasy        ScienceFiction    Comedy          Horror
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:59        FALSE:13        FALSE:62        FALSE:62        FALSE:52        FALSE:62        FALSE:58
 TRUE :3         TRUE :49        NA's :0         NA's :0         TRUE :10        NA's :0         TRUE :4
 NA's :0         NA's :0                                         NA's :0                         NA's :0


   Party          Stealth        TournamentMoneyAwarded TournamentTotalPlayers TotalTournaments    cluster
 Mode :logical   Mode :logical   Min.   :     25.0      Min.   :   1.00        Min.   :  1.00   Min.   :1
 FALSE:61        FALSE:61        1st Qu.:    387.5      1st Qu.:   5.00        1st Qu.:  2.00   1st Qu.:1
 TRUE :1         TRUE :1         Median :  11500.0      Median :  20.00        Median :  6.50   Median :1
 NA's :0         NA's :0         Mean   : 172490.0      Mean   :  91.35        Mean   : 42.15   Mean   :1
                                 3rd Qu.:  54390.0      3rd Qu.:  62.00        3rd Qu.: 16.75   3rd Qu.:1
                                 Max.   :2747219.9      Max.   :1621.00        Max.   :821.00   Max.   :1
```

Fig.7. Summary of Cluster 2

```
[[2]]
      X.Y          UserScore      CriticScore    ESRBRating    NA_Sales        Global_Sales
 Min.   :1999   Min.   :37.00   Min.   :70.00   E: 4       Min.   :0.00000   Min.   :0.0300
 1st Qu.:2004   1st Qu.:64.00   1st Qu.:81.00   T:11       1st Qu.:0.01000   1st Qu.:0.0500
 Median :2006   Median :77.50   Median :85.00   M: 1       Median :0.01000   Median :0.0900
 Mean   :2006   Mean   :73.19   Mean   :84.69              Mean   :0.09875   Mean   :0.6525
 3rd Qu.:2008   3rd Qu.:83.50   3rd Qu.:89.25              3rd Qu.:0.03500   3rd Qu.:0.3125
 Max.   :2012   Max.   :89.00   Max.   :93.00              Max.   :0.96000   Max.   :6.2900
 FirstPersonPerspective  Strategy        Fighting        Shooter          Sports          Racing
 Mode :logical          Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:16               FALSE:2         FALSE:16        FALSE:16        FALSE:16        FALSE:16
 NA's :0                TRUE :14        NA's :0         NA's :0         NA's :0         NA's :0
                        NA's :0


    RPG             Card          Adventure       Platformer      Beat.emUp          TBS            Puzzle
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:14        FALSE:16        FALSE:12        FALSE:16        FALSE:16        FALSE:15        FALSE:12
 TRUE :2         NA's :0         TRUE :4         NA's :0         NA's :0         TRUE :1         TRUE :4
 NA's :0                         NA's :0                                         NA's :0         NA's :0


 Simulator         Action         Warfare         Fantasy        ScienceFiction    Comedy          Horror
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:12        FALSE:15        FALSE:15        FALSE:15        FALSE:13        FALSE:16        FALSE:16
 TRUE :4         TRUE :1         TRUE :1         TRUE :1         TRUE :3         NA's :0         NA's :0
 NA's :0         NA's :0         NA's :0         NA's :0         NA's :0


   Party          Stealth        TournamentMoneyAwarded TournamentTotalPlayers TotalTournaments     cluster
 Mode :logical   Mode :logical   Min.   :     160       Min.   :  1.00         Min.   :   1.00   Min.   :2
 FALSE:16        FALSE:16        1st Qu.:  13885        1st Qu.:  7.75         1st Qu.:   1.75   1st Qu.:2
 NA's :0         NA's :0         Median :  61343        Median : 29.00         Median :   5.50   Median :2
                                 Mean   : 919772        Mean   :142.88         Mean   :  89.38   Mean   :2
                                 3rd Qu.: 523736        3rd Qu.:179.25         3rd Qu.:   9.50   3rd Qu.:2
                                 Max.   :5283426        Max.   :757.00         Max.   :1265.00   Max.   :2
```

Fig.8. Summary of Cluster 3

```
[[3]]
     X.Y         UserScore      CriticScore     ESRBRating    NA_Sales      Global_Sales
 Min.   :1994   Min.   :26.00   Min.   : 60.00   E: 1       Min.   :0.010   Min.   : 0.010
 1st Qu.:2004   1st Qu.:64.50   1st Qu.: 80.50   T: 8       1st Qu.:0.255   1st Qu.: 0.355
 Median :2007   Median :76.00   Median : 83.00   M:34       Median :0.820   Median : 2.100
 Mean   :2008   Mean   :72.02   Mean   : 83.35              Mean   :2.319   Mean   : 3.853
 3rd Qu.:2012   3rd Qu.:82.00   3rd Qu.: 88.50              3rd Qu.:4.100   3rd Qu.: 6.590
 Max.   :2016   Max.   :89.00   Max.   :100.00              Max.   :9.040   Max.   :14.730
 FirstPersonPerspective  Strategy        Fighting       Shooter          Sports          Racing
 Mode:logical           Mode :logical   Mode :logical   Mode:logical   Mode :logical   Mode :logical
 TRUE:43                FALSE:43        FALSE:43        TRUE:43        FALSE:43        FALSE:43
 NA's:0                 NA's :0         NA's :0         NA's:0         NA's :0         NA's :0


    RPG             Card          Adventure       Platformer      Beat.emUp          TBS            Puzzle
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:43        FALSE:43        FALSE:41        FALSE:41        FALSE:43        FALSE:42        FALSE:39
 NA's :0         NA's :0         TRUE :2         TRUE :2         NA's :0         TRUE :1         TRUE :4
                                 NA's :0         NA's :0                         NA's :0         NA's :0


 Simulator         Action         Warfare         Fantasy        ScienceFiction    Comedy          Horror
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:39        FALSE:8         FALSE:40        FALSE:43        FALSE:25        FALSE:42        FALSE:41
 TRUE :4         TRUE :35        TRUE :3         NA's :0         TRUE :18        TRUE :1         TRUE :2
 NA's :0         NA's :0         NA's :0                         NA's :0         NA's :0         NA's :0


   Party          Stealth        TournamentMoneyAwarded TournamentTotalPlayers TotalTournaments    cluster
 Mode :logical   Mode :logical   Min.   :    1000       Min.   :   1.0         Min.   :  1.00   Min.   :3
 FALSE:43        FALSE:43        1st Qu.:   53579       1st Qu.:   1.5         1st Qu.:  4.00   1st Qu.:3
 NA's :0         NA's :0         Median :  170000       Median :  27.0         Median : 13.00   Median :3
                                 Mean   : 1241553       Mean   : 114.0         Mean   : 45.53   Mean   :3
                                 3rd Qu.: 1275998       3rd Qu.: 136.0         3rd Qu.: 40.00   3rd Qu.:3
                                 Max.   :12088428       Max.   :1445.0         Max.   :762.00   Max.   :3
```

Fig.9. Summary of Cluster 4

```
[[4]]
      X.Y          UserScore       CriticScore     ESRBRating     NA_Sales       Global_Sales
 Min.   :1998   Min.   :36.00   Min.   :60.00   E:37       Min.   :0.0200   Min.   :0.320
 1st Qu.:2006   1st Qu.:64.00   1st Qu.:82.00   T: 0       1st Qu.:0.2400   1st Qu.:0.790
 Median :2011   Median :66.00   Median :83.00   M: 0       Median :0.6000   Median :2.580
 Mean   :2010   Mean   :67.43   Mean   :83.08              Mean   :0.9062   Mean   :2.928
 3rd Qu.:2014   3rd Qu.:78.00   3rd Qu.:86.00              3rd Qu.:1.0600   3rd Qu.:4.110
 Max.   :2016   Max.   :87.00   Max.   :92.00              Max.   :3.9800   Max.   :8.570
 FirstPersonPerspective  Strategy        Fighting        Shooter          Sports          Racing
 Mode :logical          Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:37               FALSE:37        FALSE:36        FALSE:37        FALSE:1         FALSE:37
 NA's :0                NA's :0         TRUE :1         NA's :0         TRUE :36        NA's :0
                                        NA's :0                         NA's :0


    RPG             Card          Adventure       Platformer      Beat.emUp          TBS            Puzzle
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:37        FALSE:37        FALSE:36        FALSE:37        FALSE:37        FALSE:37        FALSE:37
 NA's :0         NA's :0         TRUE :1         NA's :0         NA's :0         NA's :0         NA's :0
                                 NA's :0


 Simulator         Action         Warfare         Fantasy        ScienceFiction    Comedy          Horror
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:36        FALSE:34        FALSE:37        FALSE:37        FALSE:37        FALSE:37        FALSE:37
 TRUE :1         TRUE :3         NA's :0         NA's :0         NA's :0         NA's :0         NA's :0
 NA's :0         NA's :0


   Party          Stealth        TournamentMoneyAwarded TournamentTotalPlayers TotalTournaments    cluster
 Mode :logical   Mode :logical   Min.   :     25        Min.   :  1.0          Min.   :  1.00   Min.   :4
 FALSE:37        FALSE:37        1st Qu.:  29410        1st Qu.:  3.0          1st Qu.:  1.00   1st Qu.:4
 NA's :0         NA's :0         Median :  80485        Median :  8.0          Median :  3.00   Median :4
                                 Mean   : 233844        Mean   : 38.7          Mean   : 27.89   Mean   :4
                                 3rd Qu.: 121618        3rd Qu.: 50.0          3rd Qu.: 17.00   3rd Qu.:4
                                 Max.   :2642136        Max.   :410.0          Max.   :349.00   Max.   :4
```

Fig.10. Summary of Cluster 5

```
[[5]]
      X.Y          UserScore       CriticScore     ESRBRating     NA_Sales       Global_Sales
 Min.   :2001   Min.   :62.00   Min.   :62.00   E:12       Min.   :0.200   Min.   : 0.310
 1st Qu.:2002   1st Qu.:66.00   1st Qu.:83.00   T: 1       1st Qu.:0.690   1st Qu.: 2.100
 Median :2006   Median :79.00   Median :85.00   M: 0       Median :1.540   Median : 2.650
 Mean   :2006   Mean   :75.38   Mean   :84.15              Mean   :1.855   Mean   : 4.405
 3rd Qu.:2009   3rd Qu.:84.00   3rd Qu.:90.00              3rd Qu.:2.350   3rd Qu.: 4.570
 Max.   :2015   Max.   :91.00   Max.   :95.00              Max.   :6.850   Max.   :14.980
 FirstPersonPerspective  Strategy        Fighting        Shooter          Sports         Racing
 Mode :logical          Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode:logical
 FALSE:13               FALSE:13        FALSE:13        FALSE:13        FALSE:13        TRUE:13
 NA's :0                NA's :0         NA's :0         NA's :0         NA's :0         NA's:0


    RPG             Card          Adventure       Platformer      Beat.emUp          TBS            Puzzle
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:13        FALSE:13        FALSE:11        FALSE:13        FALSE:13        FALSE:13        FALSE:11
 NA's :0         NA's :0         TRUE :2         NA's :0         NA's :0         NA's :0         TRUE :2
                                 NA's :0                                                         NA's :0


 Simulator         Action         Warfare         Fantasy        ScienceFiction    Comedy          Horror
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:8         FALSE:13        FALSE:13        FALSE:13        FALSE:13        FALSE:13        FALSE:13
 TRUE :5         NA's :0         NA's :0         NA's :0         NA's :0         NA's :0         NA's :0
 NA's :0


   Party          Stealth        TournamentMoneyAwarded TournamentTotalPlayers TotalTournaments    cluster
 Mode :logical   Mode :logical   Min.   :   132.4       Min.   : 1.000         Min.   :1.000    Min.   :5
 FALSE:13        FALSE:13        1st Qu.:  7679.2       1st Qu.: 3.000         1st Qu.:1.000    1st Qu.:5
 NA's :0         NA's :0         Median : 32208.5       Median : 4.000         Median :2.000    Median :5
                                 Mean   : 36930.9       Mean   : 7.462         Mean   :2.769    Mean   :5
                                 3rd Qu.: 36322.6       3rd Qu.:12.000         3rd Qu.:3.000    3rd Qu.:5
                                 Max.   :222000.0       Max.   :19.000         Max.   :8.000    Max.   :5
```

Tab.3. Cluster Interpretation

	Tournament (Avg)	Gameplay	Scores (Avg)	Sales (Avg)
CompleteDataset (Standard)	Money = $46000 Players = 13 Tournaments = 5	E = 56 T = 74 M = 4 Action = 88 Fighting = 60 Shooter = 43 FirstPerson = 43 (34%)	UserScore = 75 CriticScore = 83	NA_Sales = 0.380 GlobalSales = 1.010
Cluster 1	Money = lower Players = higher Tournament = lower	T = 54 (87%) Fighting = 49 Action = 49 FirstPerson = 0	Similar	Lower
Cluster 2	Money = higher Players = higher Tournament = similar	T = 11(69%) Strategy = 14(88%) FirstPerson = 0	Similar	Lower
Cluster 3	Money = higher Players = higher Tournament = higher	M = 34 (79%) Action = 35 ScienceFiction = 18 FirstPerson = 43 (100%)	Similar	High Higher (double)
Cluster 4	Money = higher(double) Players = lower Tournament = lower	E = 37 (100%) Sports = 36 FirstPerson = 0	Lower Similar	Higher (double)
Cluster 5	Money = lower Players = lower Tournament = lower	E = 12 (92%) Racing = 13 (100%) Simulator = 5 FirstPerson = 0	Higher	Higher(quadruple) Higher(double)

5.1.5 MEDOIDS INTERPRETATION

Another benefit of the PAM algorithm with respect to interpretation is that the medoids serve as exemplars of each cluster.

Fig.11. Summary of Medoids

```
> Complete3[pam_fit$medoids,]
                                X.Y UserScore CriticScore ESRBRating NA_Sales Global_Sales FirstPersonPerspective
BlazBlue: Continuum Shift Extend 2011        75          83          T     0.03         0.11                 FALSE
Rise of Nations: Rise of Legends 2006        85          84          T     0.00         0.03                 FALSE
Unreal Tournament 3              2007        77          86          M     0.33         0.67                  TRUE
NHL 13                           2012        66          83          E     0.51         0.66                 FALSE
Forza Motorsport 2               2007        83          90          E     2.35         4.05                 FALSE
                                Strategy Fighting Shooter Sports Racing   RPG  Card Adventure Platformer Beat.emUp
BlazBlue: Continuum Shift Extend    FALSE     TRUE   FALSE  FALSE  FALSE FALSE FALSE     FALSE      FALSE     FALSE
Rise of Nations: Rise of Legends     TRUE    FALSE   FALSE  FALSE  FALSE FALSE FALSE     FALSE      FALSE     FALSE
Unreal Tournament 3                FALSE    FALSE    TRUE  FALSE  FALSE FALSE FALSE     FALSE      FALSE     FALSE
NHL 13                             FALSE    FALSE   FALSE   TRUE  FALSE FALSE FALSE     FALSE      FALSE     FALSE
Forza Motorsport 2                 FALSE    FALSE   FALSE  FALSE   TRUE FALSE FALSE     FALSE      FALSE     FALSE
                                 TBS Puzzle Simulator Action Warfare Fantasy ScienceFiction Comedy Horror Party
BlazBlue: Continuum Shift Extend FALSE  FALSE     FALSE   TRUE   FALSE   FALSE          FALSE  FALSE  FALSE FALSE
Rise of Nations: Rise of Legends FALSE  FALSE     FALSE  FALSE   FALSE   FALSE          FALSE  FALSE  FALSE FALSE
Unreal Tournament 3              FALSE  FALSE     FALSE   TRUE   FALSE   FALSE           TRUE  FALSE  FALSE FALSE
NHL 13                           FALSE  FALSE     FALSE  FALSE   FALSE   FALSE          FALSE  FALSE  FALSE FALSE
Forza Motorsport 2               FALSE  FALSE     FALSE  FALSE   FALSE   FALSE          FALSE  FALSE  FALSE FALSE
                                Stealth TournamentMoneyAwarded TournamentTotalPlayers TotalTournaments
BlazBlue: Continuum Shift Extend   FALSE                  25.0                      3                1
Rise of Nations: Rise of Legends   FALSE                69600.0                      8                1
Unreal Tournament 3                FALSE               104977.8                      7                4
NHL 13                             FALSE                70000.0                      1                1
Forza Motorsport 2                 FALSE               222000.0                     17                2
>
```

5.1.6 RECOMMENDATIONS

A video game launch with a similar gameplay as Cluster 3 would prove successful for Valve Corporation. Cluster 3 is a video game rated M for mature. It is an Action focused gaming experience with a Science Fiction based setting and narrative, played from the First-Person Perspective.

5.2 CORRELATION (SALES, SCORES, AND TOURNAMENTS)

The question that this model is trying to answer is:
What is the correlation between sales, tournaments and scores (by critics and users)? Correlation Pearson was performed to address this issue.

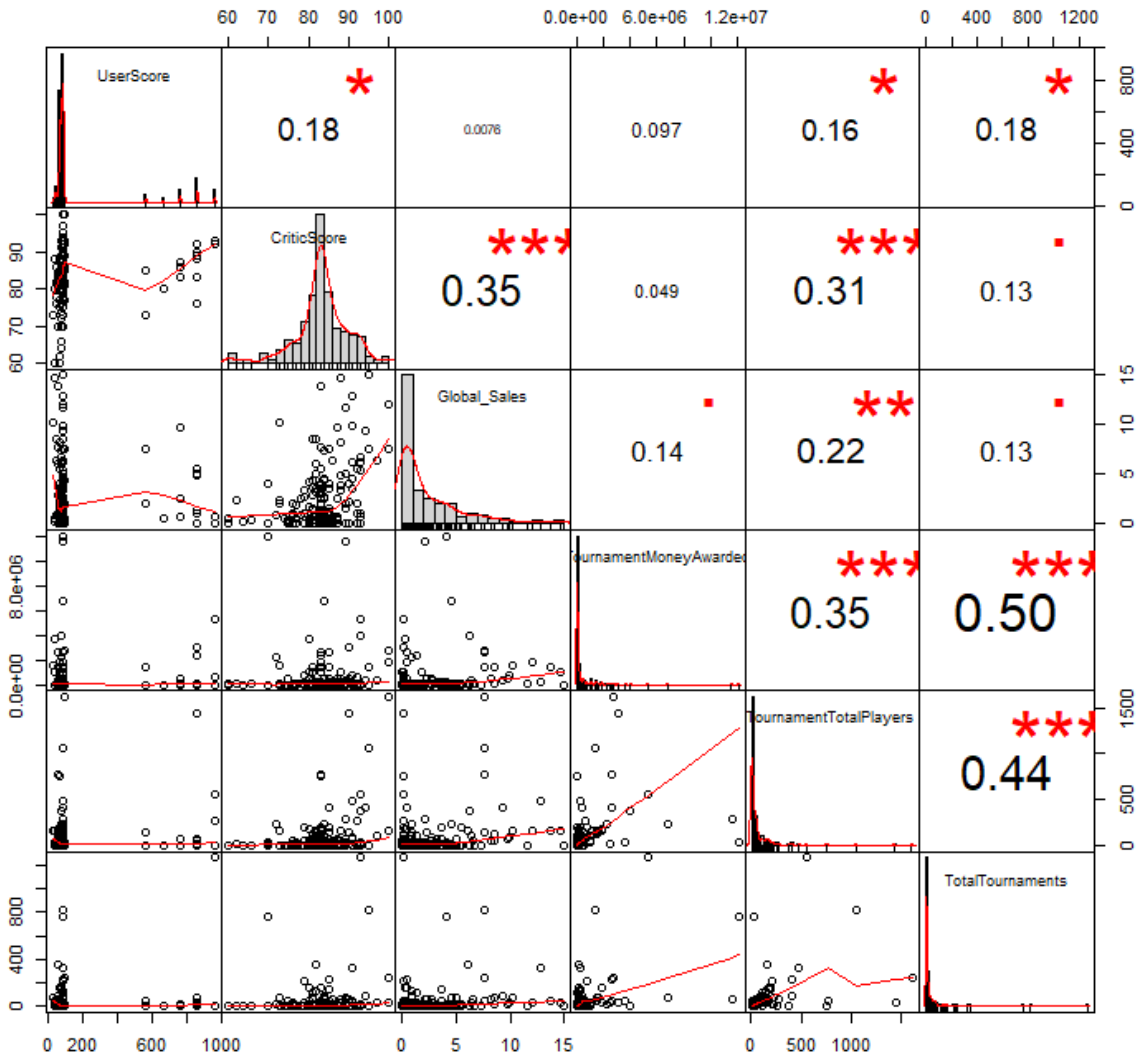
5.2.2 VARIABLES

- User score
- Critic score
- Global sales
- Tournament Money Awarded
- Tournament Total Players
- Total Tournaments

5.2.3 MODEL FINDINGS

Figure 12 shows the findings of the model results from Correlation.

Fig.12. Correlation Model Findings



With p-value less than the significant alpha (0.05) and with consideration of coefficient r, the correlation plot suggests strong association between:

- Strong positive correlation between tournament money awarded and total tournament; r= 0.5
- Medium strength of positive correlation between total tournaments and total tournament players; r=0.44

- Medium strength of positive correlation between tournament Money awarded and total tournament players; r=0.35
- Medium strength of positive correlation between critics score and global sales; r=0.35
- Medium strength of positive correlation between critics score and tournaments total players; r=0.31

5.2.4 RECOMMENDATIONS

- Since tournament Money awarded is fairly correlated with total tournament players, it's recommended to consider this issue when developing the new game. It will be helpful to design the game in away to be compatible with tournaments and tournaments awards.
- It's recommended to put into consideration critics score as it's fairly correlated with global sales and also number of players (in tournaments).

5.3 DECISION TREES (GLOBAL SALES VS GENRE, ESRB RATEINGS, PLAYERS AND TOURNUMNETS)

The question that this model is trying to answer is: How is global sales affect by genres, ESRB ratings, number of tournaments and number of players? Decision tree method was performed to address this issue.

5.3.2 VARIABLES

Global Sales was set as a response. The explanatory variables are:

- ESRB ratings
- 22 types of Genre (logical variables)
- Number of players
- Number of tournaments

5.3.3 MODEL FINDINGS

Figure 13 displays the summary of the Decision Tree model results.

Fig.13. Decision Tree Findings

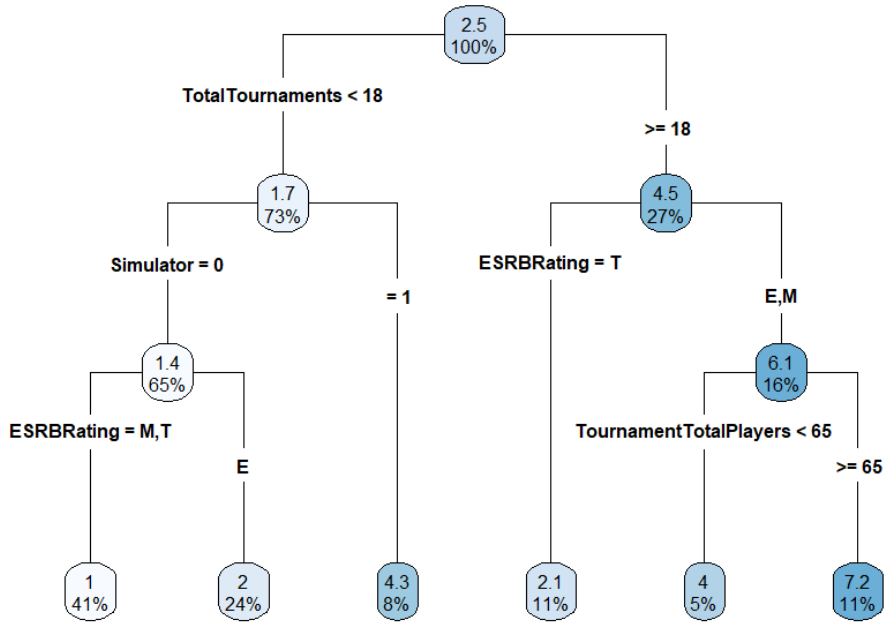
```
Call:
rpart(formula = tree_data$Global_Sales ~ ., data = tree_data)
  n= 171

         CP nsplit rel error    xerror      xstd
1 0.14949132      0 1.0000000 1.0151617 0.1745331
2 0.10245778      1 0.8505087 1.0369151 0.1667136
3 0.05842842      2 0.7480509 0.9864445 0.1618564
4 0.03361640      3 0.6896225 1.0067258 0.1582293
5 0.01425312      4 0.6560061 0.9289865 0.1337473
6 0.01000000      5 0.6417530 0.9306501 0.1352313

Variable importance
      TotalTournaments TournamentTotalPlayers             ESRBRating                Fighting               Simulator
                    24                     21                     19                      11                       9
FirstPersonPerspective                Shooter                 Sports                 Warfare                Strategy
                     5                       5                      2                       2                       2
                Action
                     1

Node number 1: 171 observations,    complexity param=0.1494913
  mean=2.494094, MSE=10.40051
  left son=2 (125 obs) right son=3 (46 obs)
```

Importance of variables in this decision tree is:

- 1st Total number of tournaments
- 2nd Tournament total players
- 3rd ESRB Ratings
- 4th Fighting (genre)
- 6th Simulator (genre)

The figure below is the decision tree.

Fig.14. Decision Tree



The decision tree above suggests that with low number of tournaments (<18 tournaments), there is still chance to achieve higher global sales by choosing simulator genre. With higher number of tournaments (≥18 tournaments), there is a chance to achieve higher global sales with ESRB ratings of E(Everyone) and M(Mature). If ESRB ratings of T(teen) was chosen, then there is a chance to have lower global sales.

5.3.4 RECOMMENDATIONS

- With low number of tournaments (<18 tournaments), higher sales can be achieved with simulator genre.
- With higher number of tournaments (≥18 tournaments), it's important to consider ESRB ratings. ESRB rating of E (everyone) and M(mature) have a chance to achieve higher global sales. If T(teen) ESRB rating was chosen, then there is a chance of having lower global sales.

5.4 POISSON REGRESSION (TOURNAMENT PLAYERS VS GENRE)

The question that this model is trying to answer is:
How does games' genres effect number of tournaments players?
Possion regression was performed to address this issue.

5.4.2 VARIABLES

- Number of tournament players was set as a response. The explanatory variables are the 22 types of Genres (logical variables).

5.4.3 MODEL FINDINGS

Figure 15 is the model results summary.

Fig.15.Poisson Regression Summary

```
Call:
glm(formula = continous_logical_data$TournamentTotalPlayers ~
    ., family = poisson(link = "log"), data = continous_logical_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-26.302   -8.397   -4.594    1.156   76.537

Coefficients: (2 not defined because of singularities)
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                3.75022    0.09886  37.935  < 2e-16 ***
FirstPersonPerspectiveTRUE 0.63976    0.09620   6.650 2.93e-11 ***
StrategyTRUE               1.22396    0.10134  12.078  < 2e-16 ***
FightingTRUE              -0.07016    0.09585  -0.732  0.46419
ShooterTRUE                     NA         NA      NA       NA
SportsTRUE                -0.27620    0.10088  -2.738  0.00618 **
RacingTRUE                -1.45387    0.14222 -10.223  < 2e-16 ***
RPGTRUE                    2.05428    0.10902  18.843  < 2e-16 ***
CardTRUE                        NA         NA      NA       NA
AdventureTRUE              0.18544    0.07355   2.521  0.01169 *
PlatformerTRUE             0.14028    0.08809   1.593  0.11127
Beat.emUpTRUE              1.38308    0.09990  13.844  < 2e-16 ***
TBSTRUE                   -2.02225    0.17595 -11.493  < 2e-16 ***
PuzzleTRUE                -1.13517    0.07606 -14.925  < 2e-16 ***
SimulatorTRUE             -0.51518    0.03369 -15.290  < 2e-16 ***
ActionTRUE                 0.96828    0.02906  33.315  < 2e-16 ***
WarfareTRUE                0.20295    0.05166   3.928 8.55e-05 ***
FantasyTRUE                1.14365    0.07495  15.259  < 2e-16 ***
ScienceFictionTRUE        -1.15532    0.03103 -37.226  < 2e-16 ***
ComedyTRUE                 1.42827    0.06679  21.386  < 2e-16 ***
HorrorTRUE                -1.72970    0.11727 -14.750  < 2e-16 ***
PartyTRUE                  2.17551    0.09401  23.141  < 2e-16 ***
StealthTRUE               -0.14606    0.27629  -0.529  0.59706
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 38464  on 170  degrees of freedom
Residual deviance: 24438  on 150  degrees of freedom
AIC: 25277

Number of Fisher Scoring iterations: 7
```

With p-value less than the significant alpha (0.05), the model suggest that number of tournament players is affected by some genres. Genres that have positive relationship with number of total players (in tournament) are (descending order by number of coefficient):

- Party 2.17551
- RPG (Role-Playing games) 2.05428
- Comedy 1.42827
- Beat.emUp (Beat them Up) 1.38308
- Strategy 1.22396
- Fantasy 1.14365
- Action 0.96828
- FirstPersonPerspective 0.63976
- Warfare 0.20295
- Adventure 0.18544

Genres that have negative relationship with number of total players (in tournament) are (descending order by number of coefficient):

- Sports -0.2762
- Simulator -0.51518
- Puzzle -1.13517
- Science Fiction -1.15532
- Racing -1.45387
- Horror -1.7297
- TBS (turn-based strategy) -2.02225

5.4.4 RECOMMENDATIONS

- For the new developed game, in order to maximize number of players on tournaments, it's recommended to consider genres of:
    - Party
    - RPG (Role-Playing games)
    - Comedy
    - Beat.emUp (Beat them Up)
    - Strategy
    - Fantasy
    - Action
    - FirstPersonPerspective
    - Warfare
    - Adventure

- For the new developed game, in order to maximize number of players on tournaments, it's recommended to avoid genres of:
    - Sports
    - Simulator

- Puzzle
- Science Fiction
- Racing
- Horror
- TBS (turn-based strategy)

5.5 ANOVA (SALES VS GENRE)

The question that this hypothesis testing is trying to answer is:
Do different types of genres achieve the same average of global sales?
ANOVA hypothesis testing was performed in order to investigate this issue.

5.5.2 VARIABLES

Global Sales was set as a response. The explanatory variables are the 22 types of Genres (logical variables)

5.5.3 MODEL FINDINGS

Figure 16 is the model results summary.

Fig.16 Anova Model Summary

```
                        Df  Sum Sq Mean Sq F value   Pr(>F)
FirstPersonPerspective   1   106.2  106.16  12.896 0.000446 ***
Strategy                 1    57.1   57.11   6.937 0.009326 **
Fighting                 1   115.2  115.24  14.000 0.000260 ***
Sports                   1    14.4   14.37   1.746 0.188449
Racing                   1     5.4    5.37   0.653 0.420441
RPG                      1    11.4   11.43   1.388 0.240588
Adventure                1     3.1    3.12   0.379 0.538963
Platformer               1    55.4   55.36   6.725 0.010446 *
Beat.emUp                1    91.3   91.34  11.096 0.001089 **
TBS                      1     1.8    1.84   0.224 0.636677
Puzzle                   1     9.4    9.35   1.136 0.288235
Simulator                1    11.7   11.68   1.419 0.235433
Action                   1     1.0    0.99   0.120 0.728988
Warfare                  1    22.3   22.31   2.710 0.101802
Fantasy                  1     8.4    8.40   1.021 0.313930
ScienceFiction           1     0.4    0.41   0.050 0.824154
Comedy                   1     2.4    2.41   0.293 0.589268
Horror                   1     0.1    0.06   0.008 0.930554
Party                    1    25.3   25.27   3.070 0.081778 .
Stealth                  1     1.5    1.53   0.186 0.666985
Residuals              150  1234.7    8.23
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

With p-value less than the significant alpha (0.05), we fail to reject the null hypothesis, we conclude that we have enough statistical evidence that not all genres has the same average of global sales. The games with genres with significant p-values are:

- First Person Perspective
- Strategy
- Fighting
- Platform
- Beat.emUp (beat them up)

5.5.4 RECOMMENDATIONS

For the new developed game, it's recommended to consider the genres of First Person Perspective, Strategy, Fighting, Platform, and Beat.emUp (beat them up)to achieve higher global sales.

5.6 LINEAR REGRESSION (SALES BY CRITICS SCORE, USERS SCORE, AND TOURNAMENTS)

The question that this hypothesis testing is trying to answer is:
Is there a linear relationship between sales and critic score, users score, and tournaments? Linear regression was performed to address this issue.

5.6.2 VARIABLES

Global sales were set as a response. The explanatory variables are:

- Users score
- Critics score
- Tournament money awards
- Tournament total players
- Total tournaments

5.6.3 MODEL FINDINGS

Figure 17 is the model results summary.

Fig.17. Linear Regression Model Summary

```
Call:
lm(formula = lm_sales_players_critics_user$Global_Sales ~ .,
    data = lm_sales_players_critics_user)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0985 -2.1640 -0.7858  1.2927 11.1379

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.122e+01  3.124e+00  -3.593 0.000432 ***
UserScore               -1.178e-03  1.123e-03  -1.049 0.295587
CriticScore              1.639e-01  3.790e-02   4.324 2.65e-05 ***
TournamentMoneyAwarded   1.749e-07  1.763e-07   0.993 0.322377
TournamentTotalPlayers   1.431e-03  1.293e-03   1.106 0.270349
TotalTournaments         4.968e-04  2.099e-03   0.237 0.813220
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.027 on 165 degrees of freedom
Multiple R-squared:  0.1502,    Adjusted R-squared:  0.1244
F-statistic: 5.832 on 5 and 165 DF,  p-value: 5.479e-05
```

With p-value less than the stated alpha (0.05) we reject the null hypothesis and conclude that we have enough statistical evidence to confirm that there is a positive linear relationship between critic score and global sales. However, R^2 value suggests that this model only explains 15% of data variation.

5.6.4 RECOMMENDATIONS

It’s important to consider critic score as it does have a positive linear relationship on global sales.

6 PRELIMINARY CONCLUSIONS

Our preliminary findings thus far provide a glimpse of a final deliverable. A video game launch with a similar gameplay as Cluster 3 would prove successful for Valve Corporation. This conclusion was further supported in the recommendation section for each of our models. In section 5.4 Anova the recommendation to consider the genres of First Person Perspective, Strategy, Fighting, Platform, and Beat.emUp (beat them up) to achieve higher global sales.section can be seen in Cluster 3 with First Person Perspective at 100% of the cluster and higher sales. In section 5.5 Linear Regression the recommendation to consider critic score as it does have a positive linear relationship on global sales is seen in Cluster 3 with slightly higher UserScores and similar CriticScores and even higher sales in comparison to the original CompleteDataset.

Cluster 3 is a video game rated M for mature. It is an Action focused gaming experience with a Science Fiction based setting and narrative, played from the First-Person Perspective.

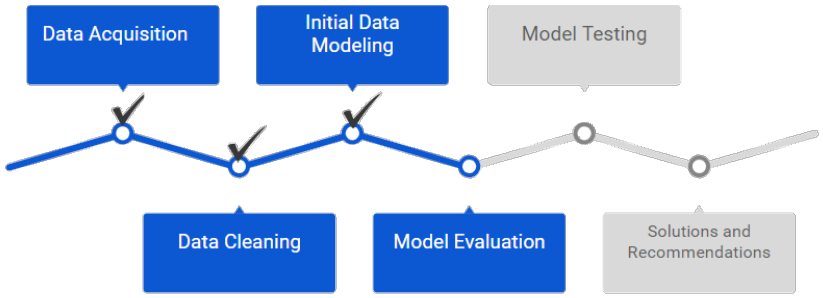
Fig.8. Summary of Cluster 3

```
[[3]]
      X.Y          UserScore      CriticScore     ESRBRating    NA_Sales      Global_Sales
 Min.   :1994   Min.   :26.00   Min.   : 60.00   E: 1       Min.   :0.010   Min.   : 0.010
 1st Qu.:2004   1st Qu.:64.50   1st Qu.: 80.50   T: 8       1st Qu.:0.255   1st Qu.: 0.355
 Median :2007   Median :76.00   Median : 83.00   M:34       Median :0.820   Median : 2.100
 Mean   :2008   Mean   :72.02   Mean   : 83.35              Mean   :2.319   Mean   : 3.853
 3rd Qu.:2012   3rd Qu.:82.00   3rd Qu.: 88.50              3rd Qu.:4.100   3rd Qu.: 6.590
 Max.   :2016   Max.   :89.00   Max.   :100.00              Max.   :9.040   Max.   :14.730
 FirstPersonPerspective  Strategy        Fighting       Shooter          Sports          Racing
 Mode:logical           Mode :logical   Mode :logical   Mode:logical   Mode :logical   Mode :logical
 TRUE:43                FALSE:43        FALSE:43        TRUE:43        FALSE:43        FALSE:43
 NA's:0                 NA's :0         NA's :0         NA's:0         NA's :0         NA's :0


    RPG             Card          Adventure       Platformer      Beat.emUp          TBS           Puzzle
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:43        FALSE:43        FALSE:41        FALSE:41        FALSE:43        FALSE:42        FALSE:39
 NA's :0         NA's :0         TRUE :2         TRUE :2         NA's :0         TRUE :1         TRUE :4
                                 NA's :0         NA's :0                         NA's :0         NA's :0


 Simulator         Action         Warfare         Fantasy       ScienceFiction    Comedy          Horror
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:39        FALSE:8         FALSE:40        FALSE:43        FALSE:25        FALSE:42        FALSE:41
 TRUE :4         TRUE :35        TRUE :3         NA's :0         TRUE :18        TRUE :1         TRUE :2
 NA's :0         NA's :0         NA's :0                         NA's :0         NA's :0         NA's :0


   Party          Stealth        TournamentMoneyAwarded TournamentTotalPlayers TotalTournaments    cluster
 Mode :logical   Mode :logical   Min.   :    1000       Min.   :   1.0         Min.   :  1.00   Min.   :3
 FALSE:43        FALSE:43        1st Qu.:   53579       1st Qu.:   1.5         1st Qu.:  4.00   1st Qu.:3
 NA's :0         NA's :0         Median :  170000       Median :  27.0         Median : 13.00   Median :3
                                 Mean   : 1241553       Mean   : 114.0         Mean   : 45.53   Mean   :3
                                 3rd Qu.: 1275998       3rd Qu.: 136.0         3rd Qu.: 40.00   3rd Qu.:3
                                 Max.   :12088428       Max.   :1445.0         Max.   :762.00   Max.   :3
```

7 PROJECT STATUS

Our project is still on track. The only major change was the decision to omit the variables and observations from the GAP dataset containing Steam information.

Fig.18. Project Status



Data acquisition, preprocessing and cleaning was done. Initial data Modeling was done and moving forward, we will examine, evaluate, and test these models in order to come up with rigid recommendations for the new game.

8 REFERENCES

About Us: Valve Corporation. (n.d.). Retrieved September 3, 2018, from Valve Corporation Website: https://www.valvesoftware.com/en/about

Cox, D. J. (2015, May 21). *Video Games Dataset -Portsmouth Research Portal.* Retrieved Sepetmber 3, 2018, from Portsmouth Research Portal: https://researchportal.port.ac.uk/portal/en/datasets/video-games-dataset(d4fe28cd-1e44-4d2f-9db6-85b347bf761e).html

E-Sports Earnings. (2018). *Highest Overall Earnings*. Retrieved September 3, 2018, from E-Sports Earnings: https://www.esportsearnings.com/players

Galyonkin, S. (2018, July 7). *Valve leaks Steam game player counts; we have the numbers.* Retrieved September 3, 2018, from Ars Technica: https://arstechnica.com/gaming/2018/07/steam-data-leak-reveals-precise-player-count-for-thousands-of-games/

Kirubi, R., & Smith, G. (2016). *Video Game Sales with Ratings.* Retrieved September 3, 2018, from Kaggle: https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings/home

Long, D. (2018, March 5). *Why Is The Gaming Industry So Successful?* Retrieved September 3, 2018, from The Gaming Gang: Get Your Geek On: https://thegaminggang.com/cool-stuff/miscellany/why-is-the-gaming-industry-so-successful/

Park, K., Cha, M., Kwak, H., & Chen, K.-T. (2017, February 26). *Achievement and Friends: Key Factors of Player Retention Vary Across Player Levels in Online Multiplayer Games.* Retrieved September 10, 2018, from Cornell University Library: https://arxiv.org/abs/1702.08005

Peterson, A. (2013, November 20). *Inside Valve's plan to revolutionize the world of video games*. Retrieved September 3, 2018, from The Washington Post: The Switch: https://www.washingtonpost.com/news/the-switch/wp/2013/11/20/inside-valves-plan-to-revolutionize-the-world-of-video-games/?noredirect=on&utm_term=.82e5a977624e

Viljanen, M., Airola, A., Majanoja, A.-M., Heikkonen, J., & Pahikkala, T. (2017, September 20). *Measuring Player Retention and Monetization using the Mean Cumulative Function.* Retrieved September 10, 2018, from Cornell University Library: https://arxiv.org/abs/1709.06737