A modern perspective on the video game industry

Alessandro D'Auria, Emanuele Di Maio, Anna Di Nardo
Olesia Nitsovych, Tiziano Turco

A.A. 2020/2021

Introduction

The principal aim of this work is to inspect several aspects of the video game industry, in order to gain valuable insight into the mechanics of the gaming community. The data analyzed in the following work comes from various sources and relates to different characteristics of video games, so that we can carry out a thorough analysis encompassing multiple areas, which can provide us with a broader view of the video game market as a whole.
The work is organized in different chapters as follows:

- **Chapter 1** will be devoted to modern technologies of web scraping and API communication;
- **Chapter 2** will be devoted to a first classification of the game parameters needed for our analysis. The main goal of this section can be summarized by the following question: "*If I were to represent a video game company, or even a single "indie" developer willing to introduce a brand new game in the international market, what parameters should I follow in order to make my game* ***successful****?*". In order to answer this question, in the first place we have to understand what parameters identify a game, and, secondly, how we can define the concept of "successful": as we will see, this second part will be strongly influenced by our final results, making somehow the definition of success in the video game market a little bit different from the common meaning;
- In **Chapter 3** we will apply different statistical learning techniques in order to understand how to properly predict the average time that users spend on video games, interpreted as a parameter of the quality of a game. We will explore different methods in order to achieve this goal, and assess their accuracy, in order to determine the model that best fits our data;
- In **Chapter 4** we will conduct a time series analysis in order to understand the evolution of video game sales over the years. We will attempt to understand if it is convenient for a company to launch a video game in a particular month, we will try to investigate any possible trend and we will see if there is a golden era for the video game market by using an econometric model;
- In **Chapter 5** we are going to analyze how games can affect young players' aggressiveness based on time spent playing video games. In addition, we shall figure out how differences in preferences in time spent on games depending on gender. In particular, from the perspective of our overall analysis, we asked such question:"*If I were to present myself as a game developer, should I take into account, since nowadays more and more young people are playing online games, the potential harm that a video game could have on the behavior of a player? If so, then my game may stumble upon censorship and criticism*".

Contents

1 Data gathering and samples

Our first topic will be focused on how we have been able to gather all the data needed for our analysis. Everything has been obtained scraping - directly or indirectly - information from the world wide web. For our purposes, of course, given the wide spectrum of video games of the modern computer age, we needed to focus our research on a specific field: that is why our current analysis has been centered on *personal computer (PC) video games*, while all the past and modern video game consoles, from the old Atari 2600 to the most recent PlayStation 5 and Xbox Series X, have been ruled out, altogether with gaming on mobile smartphones. This choice has been motivated for several reasons:

1. PC data are more accessible thanks to a more flexible community and a wider number of platforms that hosts video game selling hubs;

2. While video game on consoles is strongly dependent on the technology and the fashion of the time, PC gaming has always been solid since the '90s thanks to the multi purposes function of personal computers spread among families and offices all over the world. Consider for example the following graph from [12]:

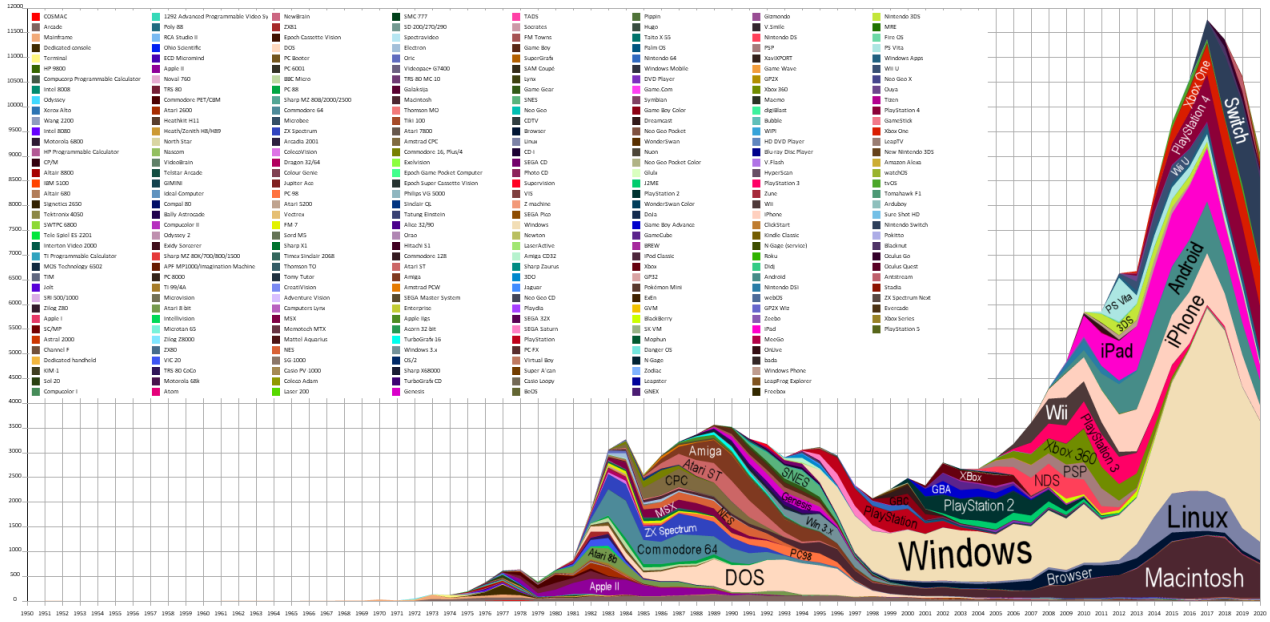

Figure 1: Number of games released per platform over the years

   It can be easily seen how games on DOS and Windows have been a constant all over the decades.

3. Thanks to its forty year long success, it is possible to analyze how PC gaming has developed from the first, simple text adventures of the '80s to the modern virtual reality experiences.

Given the fact that, in order to rate the level of success of a game alongside its parameters we needed a database with statistics for different PC games as grades and prices, we chose as source of information two important websites focused on reviewing and selling games: **Steam** and **Metacritic**. Furthermore, in order to follow the video games market over the years, we used the public information contained in the website **VGChartz**.

1.1 Steam data

Steam is an online video games selling platform for Windows, MacOS and Linux developed by Valve Corporation in 2003 [13], which currently hosts more than 30000 games. Luckily, Steam provides an online REST

API [14] which made it possible to collect games data for 34744 games, such as the number of positive and negative reviews, the current selling price (sales excluded), game genres, etc. Alongside with the official Steam API, we also used an unofficial one called SteamSpy made by Ukrainian analyst Sergey Galyonkin [15], which allowed us to gather even more information.
These APIs have been used to collect data as JSON objects, fetched from the web using Python 3.9 scripts made by ourselves, with the help of the **requests** library [17]. Due to the Steam restriction policy of the number of requests allowed, we were forced to perform no more than one HTTP GET request per second, making the process of data mining an entire day long.

The typical JSON object, at the end, looked like this:


```
{
  "gameId": "10",
  "type": "GAME",
  "title": "COUNTER-STRIKE",
  "isFree": false,
  "shortDescription": "Play the world's number 1 online action game. Engage in an incredibly...",
  "developers": [
    "Valve"
  ],
  "publishers": [
    "Valve"
  ],
  "priceInEur": 8.19,
  "genre": [
    "ACTION"
  ],
  "modality": [
    "MULTI-PLAYER",
    "PVP",
    "ONLINE PVP",
    "SHARED/SPLIT SCREEN PVP",
    "VALVE ANTI-CHEAT ENABLED"
  ],
  "releaseDate": 2000-11-01,
  "image": "https://cdn.akamai.steamstatic.com/steam/apps/10/header.jpg?t=1602535893",
  "positiveReviews": 21145,
  "negativeReviews": 778
}
```


1.2 Metacritic data

Metacritic [16] is an online website which provides reviews about movies, TV series, music and, most importantly for our purposes, video games. Unfortuntaley, Metacritic information about products are public but not easily shared, due to the lack of an unofficial API or organized datasets. That is why we had to collect our data with a more rough approach, i.e. webscraping the website itself. Once again, we made use of Python scripts wrote by ourselves for collecting the data, using the requests library for collecting the HTML source code, plus the **BeautifulSoup** library for parsing the raw HTML code [18]. For every game, it has been possible to generate a JSON object like the following:


```
{
  "title": "COUNTER-STRIKE: GLOBAL OFFENSIVE",
  "image": "https://static.metacritic.com/images/...",
  "publishers": [
    "VALVE SOFTWARE"
```

```
  ],
  "developers": [
    "VALVE SOFTWARE",
    "HIDDEN PATH ENTERTAINMENT"
  ],
  "releaseDate": "2012/08/21",
  "genre": [
    "SHOOTER",
    "FIRST-PERSON",
    "ACTION",
    "TACTICAL",
    "MODERN"
  ],
  "nOfPlayers": "ONLINE MULTIPLAYER",
  "url": "https://www.metacritic.com/game/pc/counter-strike-global-offensive",
  "userReviews": [
    {
      "grade": 6,
      "date": "2012/09/08"
    },
...
    {
      "grade": 0,
      "date": "2021/02/25"
    }
  ],
  "userNumberReviews": 920,
  "userMeanValue": 6.58804347826087,
  "userVarValue": 3.5085096052716676,
  "criticReviews": [
    {
      "grade": 95,
      "date": "2012/08/24"
    },
...
    {
      "grade": 60,
      "date": "2012/08/27"
    }
  ],
  "criticNumberReviews": 38,
  "criticMeanValue": 82.57894736842105,
  "criticVarValue": 7.6420980136256755
}
```


Notice that Metacritic data do not display prices, due to the fact that Metacritic website does not sell games.
In 8 hours of web scraping, we were able to collect data for 10459 PC games in order to reinforce our Steam sample.

1.3 VGChartz data

VGChartz is a web provider of tools and data about hardware and video games sales [19]. Like Metacritic, it does not provide information with APIs or datasets, so we had to extract all the data with the same

webscraping techniques. A tipical game extracted on VGChartz looked like this:


```
{
  "image": "https://www.vgchartz.com/games/boxart/full_3053446AmericaFrontccc.jpg",
  "title": "COUNTER-STRIKE: GLOBAL OFFENSIVE",
  "platform": "PC",
  "publisher": "VALVE",
  "developer": "VALVE CORPORATION",
  "total_shipped": "40.00m",
  "total_sales": "N/A",
  "na_sales": "N/A",
  "eu_sales": "N/A",
  "jp_sales": "N/A",
  "other_sales": "N/A",
  "release_date": "2012-08-21",
  "last_update": "2019-03-26"
}
```


The scraping took around 12 hours and made us collect 50284 games, 9603 of which are PC ones.

1.4 A brief look at our variables

Merging all the data described above we can obtain a large spectrum of variables associated to every game. In table 1 we report a list of each one of them specifying the type and what they represent. Variables with the apices [1], [2] and [3] are taken, respectively, from Steam, Metacritic and VGChartz.
From a first look we would be tempted to think that the variables for which we can recognize an index of "success" of a game are **the number of reviews**, which one can expect to be proportional to the number of players who have played the game, and/or **the grade of a review** (or, for Steam games, the ratio of positive reviews over all the reviews): naturally, a "good" game is either a game with very high score or a game played by a wide portion of gamers community. As we will see, these two aspects are not necessarily linked.

Group	Variable	Type	Description
Basic information	gameId [1]	Categorical	Steam unique game id.
	title [1,2,3]	Categorical	Name of the game.
	type [1]	Categorical	Game type (game, DLC, soundtrack,...)
	shortDescription [1]	Categorical	A short description of the game.
	developers [1,2]	Categorical	List of game developers.
	publishers [1,2]	Categorical	List of game publishers.
	releaseDate [1,2]	Date	Release date of the game.
	image [1,2]	Categorical	Portrait url of the game.
Game classification	genre [1,2]	Categorical	List of genres covered by the game
	tags [1]	Categorical	List of tags, which can either describe a genre or a core aspect of the game.
	modality [1]	Categorical	List of playable mode of the game (is on-line/offline, if it supports a controller, etc.)
Game reviews	positiveReviews [1]	Quantitative	Number of game reviews that have a "like" on Steam.
	negativeReviews [1]	Quantitative	Number of game reviews that have a "dislike" on Steam.
	reviews [1]	Categorical	List of Steam reviews associated to the game: it contains the date of the reviews, the score (like/dislike) and the number of hours the author of the reviews spent on the game.
	meanHours [1]	Quantitative	The average total number of hours spent by a player on a game on Steam, taken on a sample of (max) 100 reviews.
	numberOfReviews [2]	Quantitative	Total number of Metacritic reviews of a certain game.
	userReviews [2]	Ordinal	Similar to reviews, it contains a list of reviews from Metacritic users, with date and score from 0 to 10.
	userNumberReviews [2]	Quantitative	Number of user reviews on Metacritic.
	criticReviews [2]	Ordinal	Similar to reviews, it contains a list of reviews from Metacritic critics, with date and score from 0 to 100.
	criticNumberReviews [2]	Quantitative	Number of critic reviews on Metacritic.
	userMeanValue [2]	Categorical	It contains the average grade from all the user reviews of a certain game on Metacritic.
	criticMeanValue [2]	Categorical	It contains the average grade from all the critic reviews of a certain game on Metacritic.
Sales and incomes	isFree [1]	Categorical	Can be 0 or 1, depending on whether the game is free or not (on Steam store).
	priceInEur [1]	Quantitative	Price of the game in euros on Steam store at the current time.
	globalSales [3]	Quantitative	Number of copies sold per game according to VGChartz.

Table 1: List of variables extracted for every game by our web scraping scripts

2 Genre analysis

We will now start our analysis from the categorical variables associated to our games: even if we cannot extract some kind of pattern or functional dependency from them we can, indeed, understand how these qualitative variables are connected and how they can be used to better identify our games.
As said before, the variables on which we want to focus our analysis are the number of reviews associated to a game and its average score. As our first check, we will see how these variables are distributed by two important sets of attributes that can help us organize the games in simpler groups: the **genre** and the **modality** (the latter to be intended as a distinction between "single player games" and "multi player games"). This analysis will, indeed, help us understand which kind of game is more preferred and, thus, more likely to be considered a "success".

2.1 Filtering the genres

Both Metacritic and Steam datasets contain a great variety of genres, which can be overwhelming. Take a look, for example, at the following graph, which shows the number of games with a certain tag in the Steam dataset:

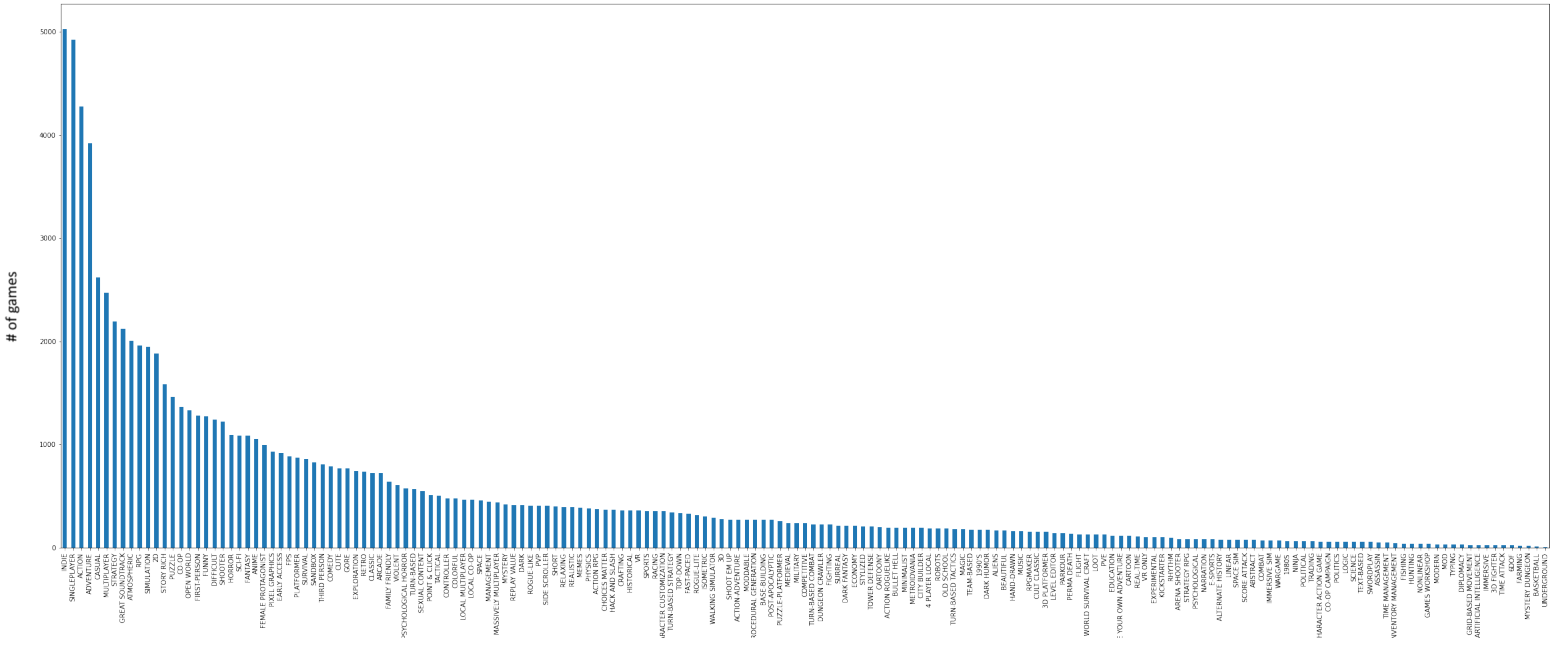

Figure 2: Number of Steam games with a specific tag

Using all these genres would be incorrect for two reasons:

1. Some genres can be considered as subgenres: for example, "2D" is a subgenre of "side scrolling", "arena shooter" is a subgenre of "shooter", etc. Thus, these genres can actually be grouped together;

2. Some genres are variants of the same genre: "local co-op" and "local multiplayer", for example, represent two ways of playing locally with other players;

3. Some genres are actually more like topics then genres per se, i.e. aspects of a particular genre. For example, "indie" only represents a game made by independent developers, or "difficult" represents a challenging game, but they can cover every other genre without representing one in particular;

Following the criteria above, we can reduce all the genres to twenty main genres, whose distribution in Steam dataset and Metacritic dataset can be seen in the following pie charts:

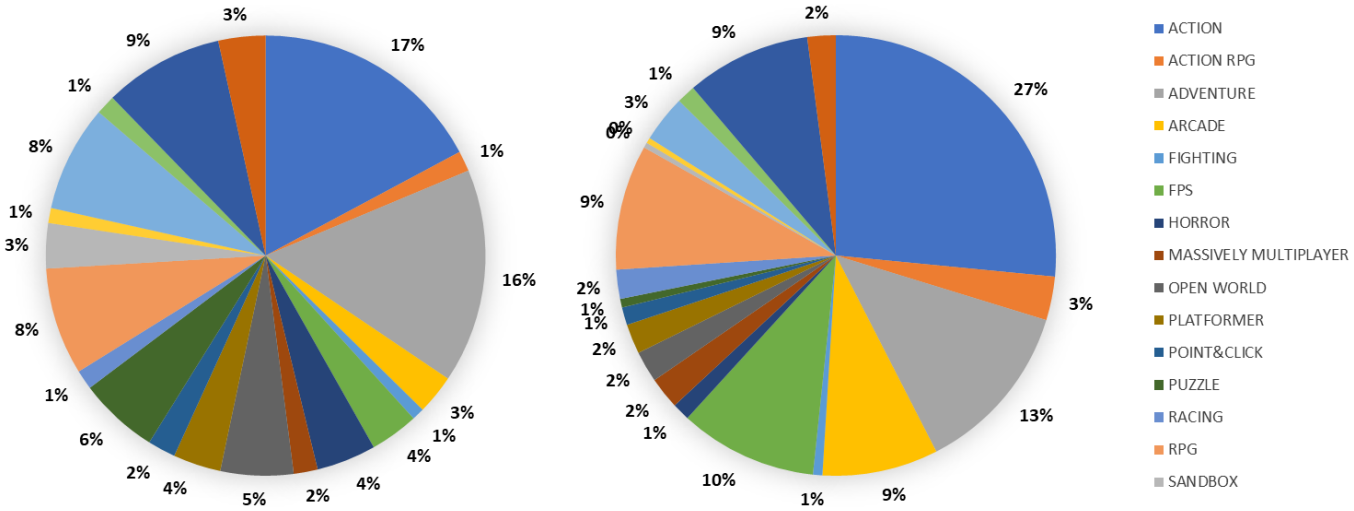

Figure 3: Distribution of genres through games for Steam (on the left) and Metacritic (on the right)

We can immediately notice that genre distributions are not very comparable in the two datasets, even though they have no apparent reason to differ. We will investigate on this topic in the next section, performing a two-way ANOVA test in order to appreciate the differences.

2.2 Reviews analysis

In the previous section we have seen how our main review datasets, i.e. the Steam one and the Metacritic one, do not share a similar distribution in game genres, which is a possible hint of hidden dependencies or biases which have not been taken into account when these two samples were chosen. In order to examine this possibility, we are going to observe how **game ratings** are distributed over our samples and over the genres collected. A useful graphical representation would be the boxplot in Figure 4, where we divided the Metacritic reviews in two other subgroups: reviews made by users and reviews made by critics (i.e., professional journalists or bloggers who write reviews for magazines, websites, etc.). For future reference, these three groups will be called "sources".
At first glance we can immediately see that:

- The average rating seems to vary significantly over the different genres;
- The average rating for a fixed genre seems to vary significantly over the three samples (Steam reviews, Metacritic Users Reviews, Metacritic Critics Reviews);
- In particular, Metacritic Users Reviews seems to have lower ratings than Metacritic Critics Reviews, which have themselves lower ratings than Steam Reviews.

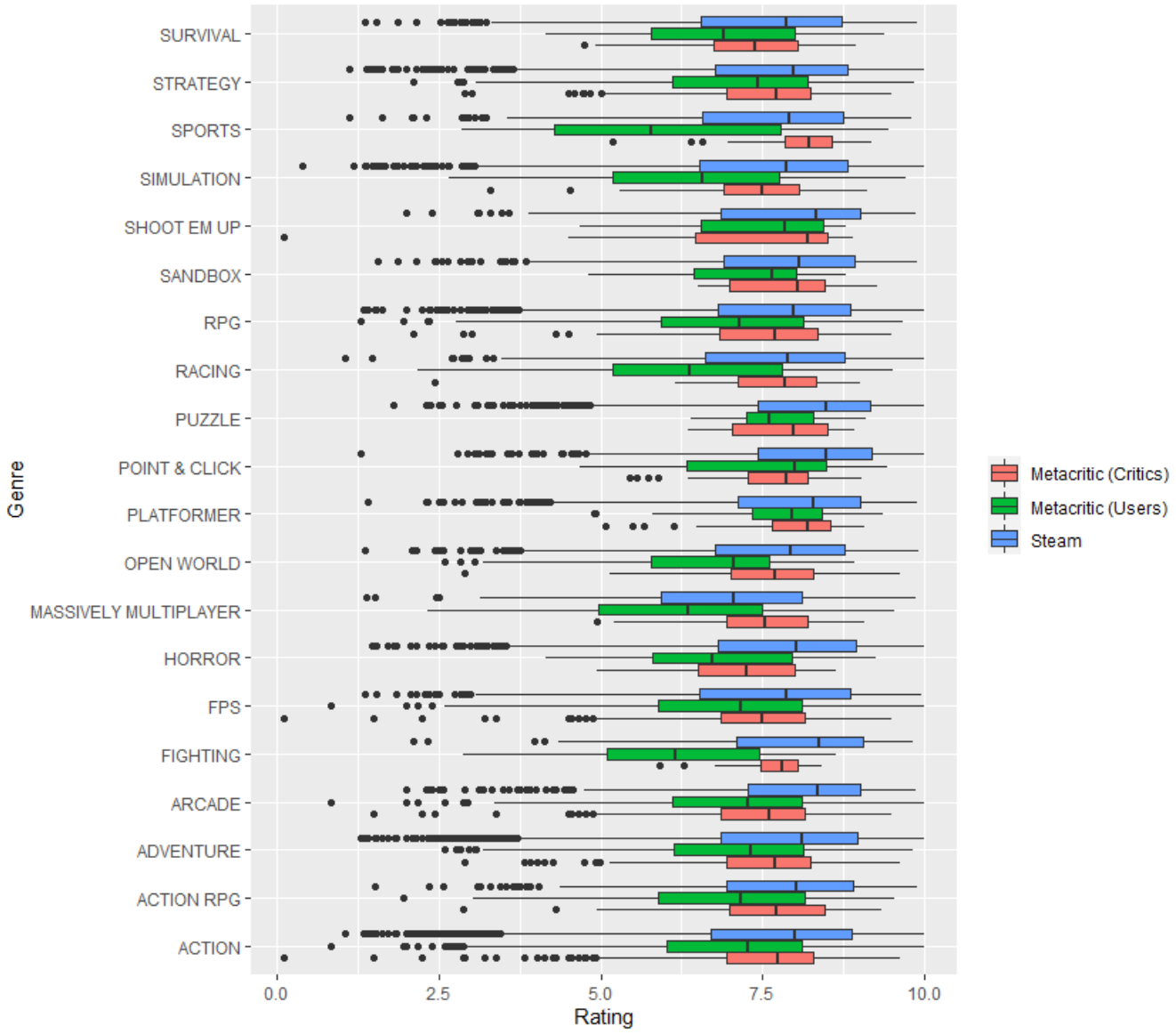

Figure 4: Ratings boxplots for game reviews

In order to make these considerations rigorous, we are now going to execute a two-way **analysis of variance** test (or, shortly, a two-way ANOVA), which is generally used to assess the relation of two independent categorical variables (in our case, the genre and the source) on a dependent variable (the rating) [20].
Let us briefly recall that, given x_{ij} the observed rating for a game in the i-th genre and in the j-source, and given $\overline{x}_{i\cdot}$ and $\overline{x}_{\cdot j}$ the average rating by genre and by source respectively, it is possible to evaluate the *sum of squares* among genres and among sources:

$$SSD_{genre} = n \sum_i (\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})^2 \tag{1}$$

$$SSD_{source} = m \sum_j (\overline{x}_{\cdot j} - \overline{x}_{\cdot\cdot})^2 \tag{2}$$

Furthermore, one can calculate the so called *residual variation*:

$$SSD_{res} = \sum_i \sum_j (\overline{x}_{ij} - \overline{x}_{i\cdot} - \overline{x}_{\cdot j} + \overline{x}_{\cdot\cdot})^2 \tag{3}$$

where $n = 20$ is the number of genres, $m = 3$ is the number of sources, and $\overline{x}_{\cdot\cdot}$ is the average rating over genres and sources.
If we divide these sums by the number of their degrees of freedom ($(n-1)$, $(m-1)$ and $(m-1)(n-1)$ respectively), we obtain the so called *mean squares*:

$$MS_{genre} = \frac{SSD_{genre}}{n-1} \tag{4}$$

$$MS_{source} = \frac{SSD_{source}}{m-1} \tag{5}$$

$$MS_{res} = \frac{SSD_{res}}{(n-1)(m-1)} \tag{6}$$

It is well known that the ratios

$$F_{genre} = \frac{MS_{genre}}{MS_{res}} \tag{7}$$

$$F_{source} = \frac{MS_{source}}{MS_{res}} \tag{8}$$

follow an F-distribution in the null hypothesis of no group dependency of our dependent variable (the game review ratings), thus it is possible to evaluate the p-value of the evaluated F and understand how good our assumptions about the distribution are. A quick calculation with R shows the results in Table 2.

	Df	Sum Sq	Mean Sq	F-value	p-value
Source	2	1940	970.0	401.726	<2e-16
Genre	19	1101	57.9	23.991	<2e-16
Source: Genre	38	292	7.7	3.185	1.43e-10
Residuals	34543	83408	2.4		

Table 2: Summary of the ANOVA test performed on the dependent variable "rating"

For sake of completeness, we also inserted an interaction term between Source and Genre. Given that the p-values are way smaller than the accepted threshold of 5%, we can fairly assume that the null hypothesis of comparable values among genres and sources can be rejected. Therefore, our first observation of great variations of review ratings per genre and source turned out to be true. For example, notice how on average Steam ratings appear higher than Metacritic Critics ratings, which are themselves higher than Metacritic User ratings, as one can view in the following Tukey multiple pairwise-comparisons table, where we reported the differences for average, higher and lower ratings (in the interval of significance) between every couple of sources, and the associated p-value for the pairwise comparison:

Source_i-Source_j	avg	lwr	hig	p-value
Metacritic (Critics)-Metacritic (Users)	0.5777785	0.66895859	0.4865984	<2e-16
Steam-Metacritic (Critics)	0.1695636	0.09598583	0.2431414	2e-7
Steam-Metacritic (Users)	0.7473421	0.68531035	0.8093739	<2e-16

Table 3: Difference between average, lowest and highest ratings through the different sources of ratings

Let us now make some considerations about the results above:

- The variations among the genres can be simply justified assuming that *not all genres are equally welcomed by the gaming community*, even though this property strongly depends on the considered sample, as one can deduce by the small p-value observed in the third row of table 2. Evidently, the three communities analyzed manifest different tendencies in video game preferences;

- The strong differences among ratings in our three samples are somehow unexpected: in principle, there is not a manifest reason why these communities should be biased. However, one can justify these fluctuations if economic factors are taken into account:

    1. On average, in Metacritic Critic ratings are higher than User ratings: this result is not that strange if one considers that sometimes *critics are paid to review a game by software houses themselves*, thus are more likely to be biased.
    2. On the other hand, Steam grades are higher than Metacritic ones (both from Users and Critics). This result can be understood if one takes into account that Steam is also a video game selling platform, and hosting a game on it requires a series of quality checks by Valve employees.

Therefore, given these results, it is hard to tell which genre a game designer should choose in order to make a successful game, given that it strongly depends on the type of community. For this reason, we will further restrict our analysis to simpler criteria: *whether a game is better welcomed if single player or multi player*.

2.3 Single Player versus Multi Player

In the gaming market, the basic distinction between single player games and multi player games is generally quite sharp: while in single player games the gamer only interacts with the machine itself, in multi player games gamers are supposed to play against or together, sometimes in different groups or teams. We will not go into details of the actual multi player mode (e.g. cooperative, massive multiplayer, MOBA, battle royal, etc.), so we will simply analyze which way of playing is more appreciated. Unfortunately, in order to take in account more criteria and variables we will be forced to exclude data from the Metacritic dataset: being less supervised, Metacritic data are not always well registered and it is often hard to tell which modality is allowed in a game due to the absence of an actual flag that identifies the game modality itself. Therefore, for this analysis, we will focus only on the Steam dataset.
The Steam dataset has the advantage of having more information to conduct our analysis on: let us consider, for example, the boxplots for ratings, number of reviews, prices and mean number of hours, showed in figures 5 and 6. Once again, in order to point out the differences, we will perform a multiple ANOVA test (or MANOVA): this time, having only one categorical variable to consider (the modality), a one-way MANOVA will be sufficient. In particular, for our purposes, we will use one of the most common test statistic for multivariate analysis, which is the *Pillai's Trace test statistic*. Recall that if one defines the matrices

$$\mathbf{E} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \overline{\mathbf{x}}_{i\cdot})(\mathbf{X}_{ij} - \overline{\mathbf{x}}_{i\cdot})^T \tag{9}$$

$$\mathbf{H} = \sum_{i=1}^{m} n_i (\overline{\mathbf{x}}_{i\cdot} - \overline{\mathbf{x}}_{\cdot\cdot})(\overline{\mathbf{x}}_{i\cdot} - \overline{\mathbf{x}}_{\cdot\cdot})^T \tag{10}$$

where m is the number of groups (i.e., the three single player, multi player and both) n_i is the number of elements in group i, $\mathbf{X}_{ij}$ is the four-dimensional vector representing the observation i in the group j, $\overline{\mathbf{x}}_{i\cdot}$ is the sample mean vector for group i and $\overline{\mathbf{x}}_{\cdot\cdot}$ is the grand mean vector (the mean over groups and over samples), the so called *Pillai's trace*

$$V = \text{trace}(\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}) \tag{11}$$

follows an F statistic whose results are reported in table 4. Being the p-value below 0.05, we can deduce once again that the average values in the different boxplots are not comparable.

	Df	Pillai	F value	num Df	den Df	Pr(>F)
Modality	2	0.068406	190.54	8	43042	<2.2e-16
Residuals	21523					

Table 4: Summary of the MANOVA test for dependent variables "ratings", "number of reviews", "mean number of hours" and "price".

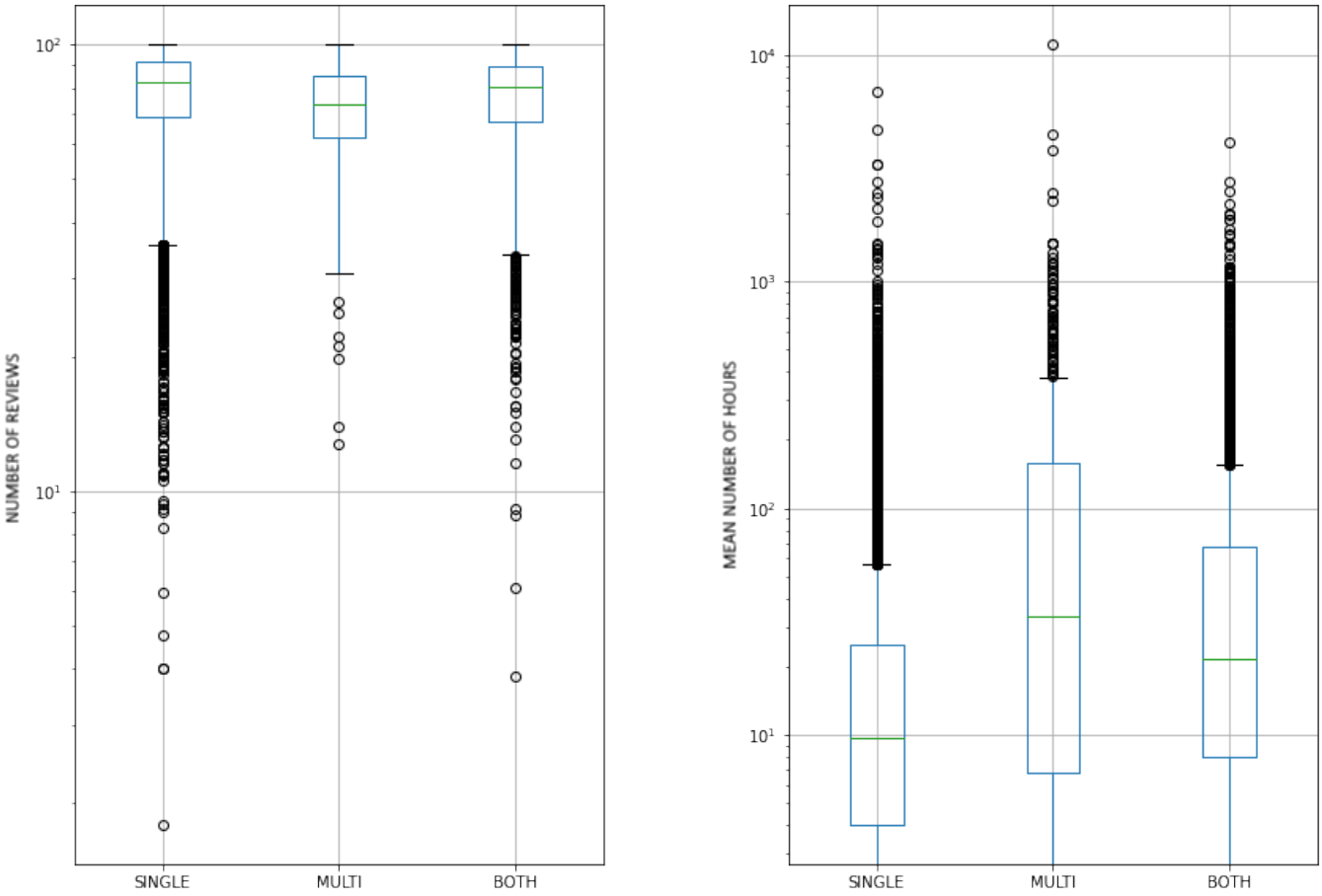

Figure 5: Boxplots for number of reviews (on the left) and mean number of hours spent (on the right) per modality

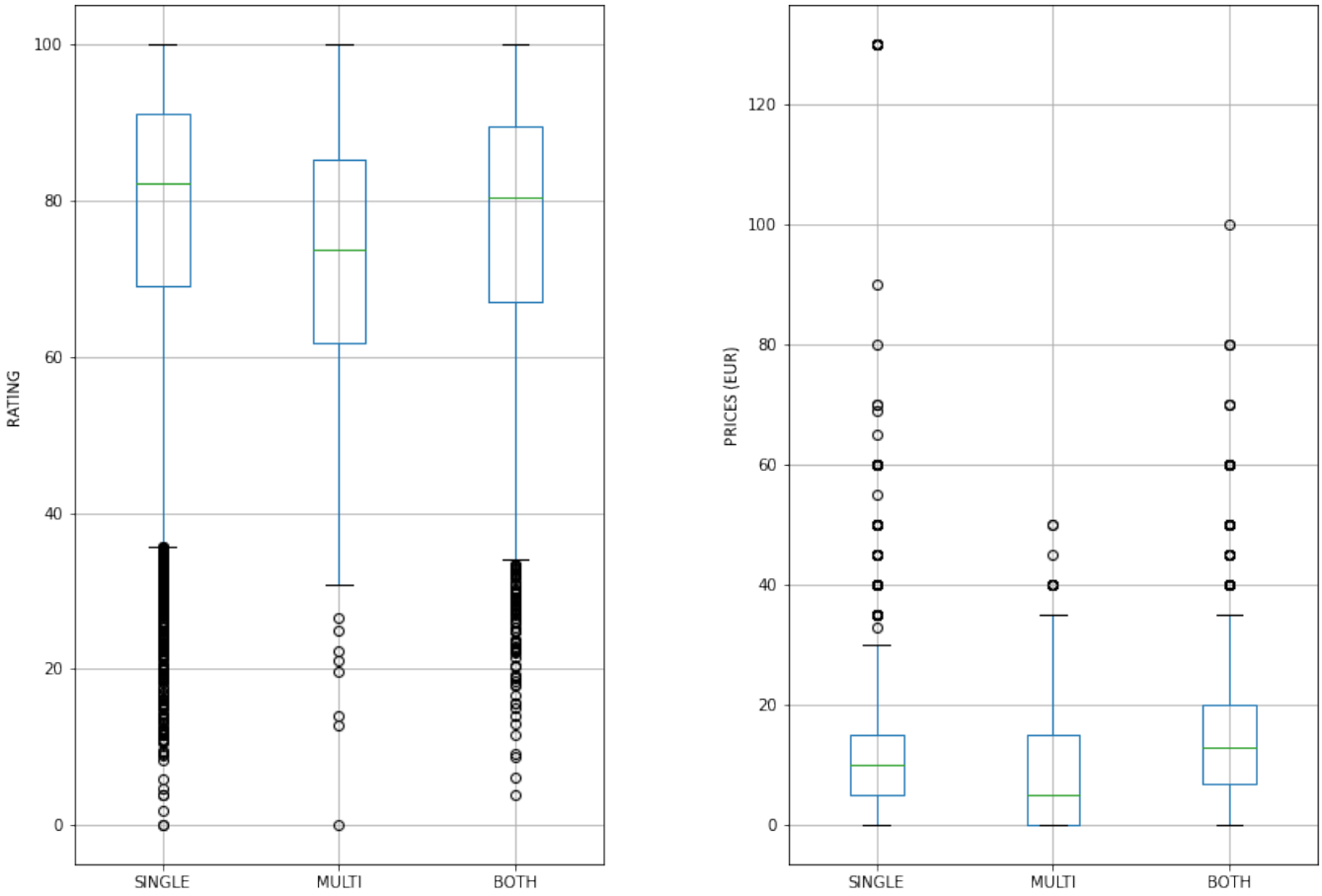

Figure 6: Boxplots for ratings (on the left) and prices (on the right) per modality

Therefore, similarly to what we have seen in the previous section, single player games and multi player games are not equally welcomed by the gaming community. The same thing can be seen if one performs a single ANOVA test for each one of our dependent variables:

Mean Hours	Df	Sum Sq	Mean Sq	F-value	p-value
Modality	2	14383901	7191950	243.4	<2e-16
Residuals	21523	635953618	29548		
Ratings	Df	Sum Sq	Mean Sq	F-value	p-value
Modality	2	328	163.79	51.56	<2e-16
Residuals	21524	68378	3.18		
Prices	Df	Sum Sq	Mean Sq	F-value	p-value
Modality	2	84984	42492	420.4	<2e-16
Residuals	21524	2175398	101		
N. Reviews	Df	Sum Sq	Mean Sq	F-value	p-value
Modality	2	14383901	7191950	243.4	<2e-16
Residuals	21524	635953618	29548		

Table 5: Summary of four one-way ANOVA tests performed on each dependent variable.

The post-hoc Tukey analysis in Table 6 managed to better underlie these differences (where, for sake of simplicity, we only reported the differences among average values):

	MULTI - SINGLE	SINGLE - BOTH	MULTI - BOTH
RATING	-0.62603461	-0.01064687	-0.63668147
PRICE (EUR)	-1.973358	-4.628026	-6.601384
MEAN HOURS	112.68437	-38.78562	73.89875
NUMBER OF REVIEWS	8307.215	3740.440	4566.774

Table 6: Differences in average values among ratings, prices, mean number of hours and number of reviews per modality

where the p-values for pairwise significance are reported in Table 7.

	MULTI - SINGLE	SINGLE - BOTH	MULTI - BOTH
RATING	<2.2e-16	0.93	<2.2e-16
PRICE (EUR)	<2.2e-16	<2.2e-16	<2.2e-16
MEAN HOURS	<2.2e-16	<2.2e-16	<2.2e-16
NUMBER OF REVIEWS	<2.2e-16	<2.2e-16	<2.2e-16

Table 7: Adjusted p-value for pairwise comparison

It is worth noticing that the p-value of the difference in rating between single player and both single/multi player games is 0.93, which is the only case where the differences are actually neglectable.
In all the other cases we can however deduce that:

1. Multi player games are on average cheaper, even if not by a far margin;
2. On average, multi player games are more played than single player games: there are way more reviews for multi player games than single player games, and people seem to spend more time on multi than single (on average, 100 hours of total playtime more, which is quite impressive);
3. On average, though, single player games have higher ratings than multi player games.

The most important result of the previous analysis is that even if multi player games are more played by the gaming community, they seem to have a lower quality than single player games. At a first glance, one could think that the higher number of reviews is the result of a simple bias: "the worse a game, the higher the number of people who complain about it", but this would not justify the highest number of hours spent on average in a game: why would a person invest so much time in a game which is generally considered bad?

Evidently, **a good game is not necessarily a highly rated game, and "quality" is not a primary target for game designers**. This result, even if counter-intuitive, actually says a lot about our gaming community: it seems that quantity takes over the quality of the contents. For this reason, we will need to change our perspective on what a "good" game is, and we will have to investigate more on the quantitative variables, as we will in the following sections.

3 Predictive Analysis

3.1 Relationship between variables

In the following section we seek to investigate the relationship between the quantitative variables in our data, with the aim of identifying patterns between the variables. The analysis in this section was carried out using R Software.
In order to provide a measure of the strength of association between two variables and the direction of the relationship, we compute the correlation matrix. We first use Pearson's correlation coefficient and obtaining the following:

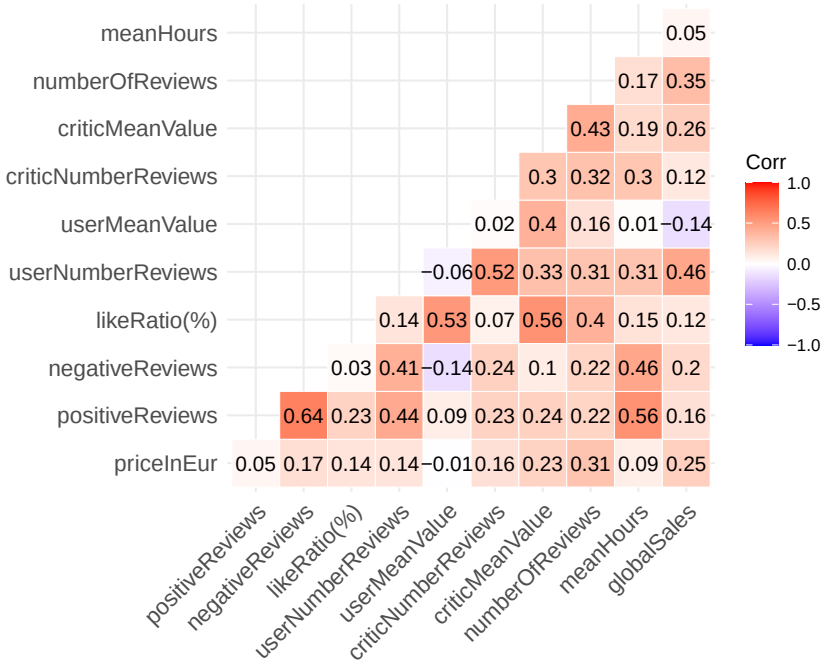

Figure 7: Correlation matrix for the quantitative variables in our data, obtained using Pearson's correlation coefficient.

We can observe that the values obtained generally indicate low linear tendencies throughout most of our data. We therefore provided Spearman's correlation coefficient as well, in order to evaluate monotonic relationships:

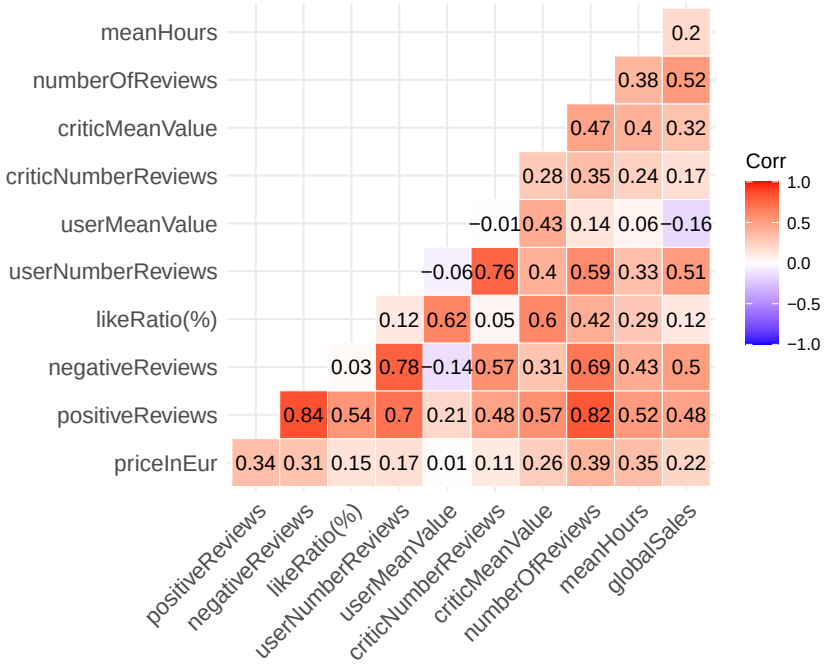

Figure 8: Correlation matrix for the quantitative variables in our data, obtained using Spearman's correlation coefficient.

We can observe that Spearman's correlation coefficients are generally slightly higher than Pearson's coefficients, albeit still indicating weak-moderate relationships. However, we can observe a high correlation coefficient between *positiveReviews* and *negativeReviews*, with a value of 0.84. Roughly speaking, games featuring a high number of positive reviews also feature a high number of negative reviews. This counter-intuitive result can be put down to the fact that, as games grow in popularity, they will draw in more and more players. Therefore, we can reasonably conclude that the number of positive reviews on a game is not so much an indication of how much a game is *liked*, as it is an indication of how *popular* a game is.

3.2 Multiple Linear Regression

In the following section we wish to analyze the *success* of a video game on the basis of the other variables. As we observed in the previous chapter, there are a number of metrics one could use to quantify "success", but positive ratings and high number of reviews turned out to be quite inefficient. We thus have opted to base it on *the time that users are willing to dedicate to a video game*. We therefore identify our response variable with *meanHours*, that is, the number of hours that users spend on a video game on average. The tool that we are going to be using to predict our quantitative response is Linear Regression. Linear Regression is a simple tool which provides high interpretability, making it easy to understand how the inputs affect the output. Therefore, we thought it would be an appropriate method to start our analysis. Let us recall that the multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{12}$$

where β_j are the model coefficients, and X_j are the p predictors. In our case, $p = 10$.

3.3 Variable Selection

Our first task is to determine which subset of variables is related to the response. The technique that we are going to employ is Best Subset Selection. Best Subset Selection looks through all possible regression models of all the different subset sizes, and looks for the best of each size. Since we have 10 variables, we obtain

10 different best subsets, each including a number of variables ranging from 1 to 10. We now need to select a single best model among all the models. There are different criteria to determine the best model. Let us recall that RSS and R^2 take the form

$$\begin{aligned} RSS &= \sum_{i=1}^{n}(y_i - \hat{y_i})^2 \\ R^2 &= 1 - \frac{RSS}{TSS} \end{aligned} \tag{13}$$

where $TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2$.

The following plots show the RSS and R^2 for the different subset sizes:

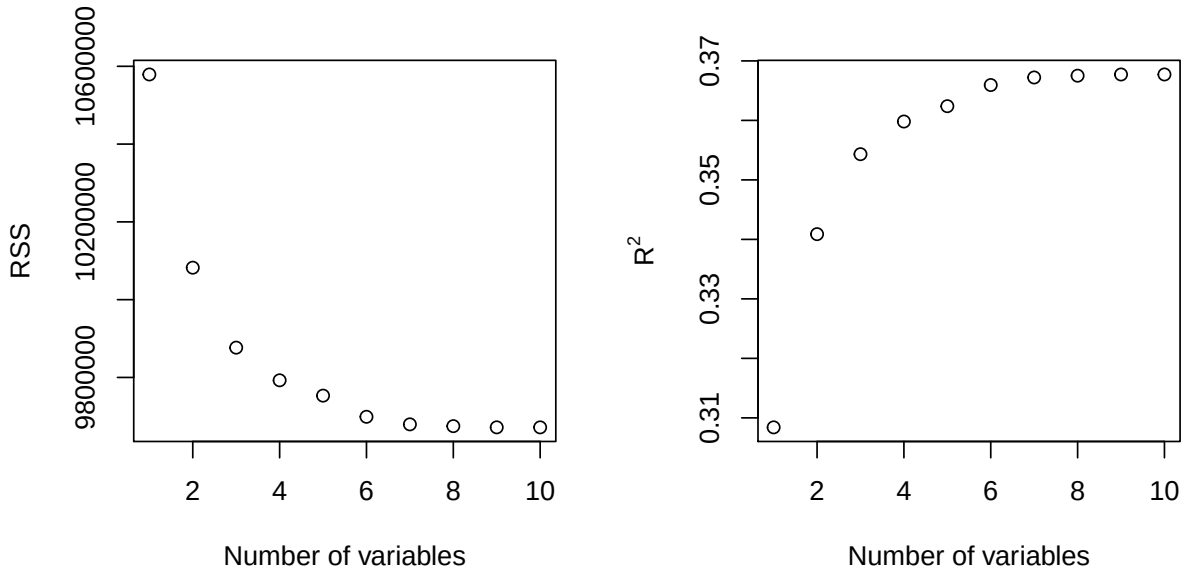

Figure 9: Plot of RSS (left) and R^2 (right) for each model as a function of the number of variables, produced by Best Subset Selection.

As we can observe from Figure 9, these quantities improve as the number of variables increases. In particular, RSS decreases monotonically, and the R^2 increases monotonically. Therefore, using these statistics to select the best model would always result in a model that has the highest number of variables. Thus, these are not suitable to choose the best model. Instead, the approach that we will follow in order to select the best model is the C_p statistic, which provides a good estimate of the test error.
Let d be the number of predictors and $\hat{\sigma}^2$ an estimate of the variance associated with the error ϵ in the linear model (12). Let us recall that C_p takes the form:

$$C_p = \frac{1}{n}\left(RSS + 2d\hat{\sigma}^2\right) \tag{14}$$

We can observe that this statistic adds a penalty on models with a large number of variables d. When determining which of a set of models is best, we choose the model with the lowest C_p. The following plot shows C_p for the best models of each size:

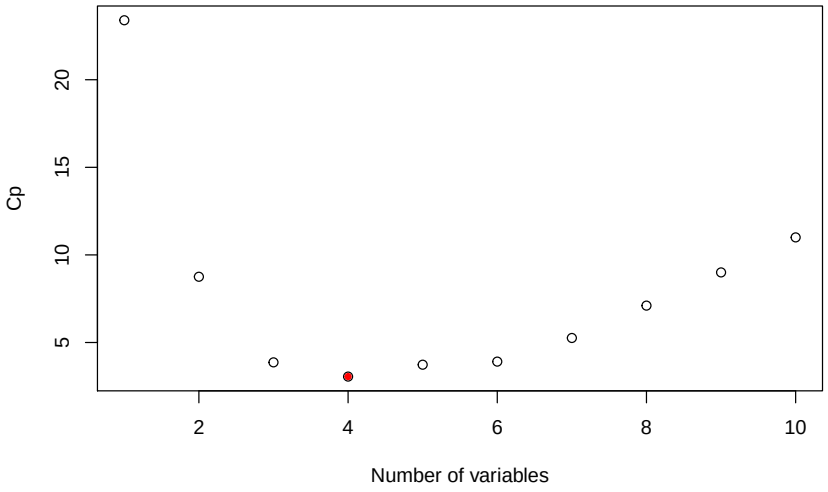

Figure 10: Plot of C_p for each size. The red point indicates the minimum.

We can observe that C_p selects a 4-variable model, including the variables shown in the following table:

No. of variables	Best subset
Four	positiveReviews, negativeReviews, criticNumberReviews, globalSales

Table 8: Selected model for Best Subset Selection.

3.4 Model Fit

We will now fit a multiple linear regression model to our data, using the 4 variables we obtained earlier according to the C_p statistic, namely *positiveReviews, negativeReviews, criticNumberReviews* and *globalSales*. The coefficient estimates, as well as their standard errors, t-values, and p-values are shown below:

	Coefficient	Std. error	t-statistic	p-value
Intercept	53.0248938	8.3527511	6.348	4.07e-10
positiveReviews	0.0041076	0.0004285	9.586	<2e-16
negativeReviews	0.0094972	0.0038962	2.438	0.0151
criticNumberReviews	1.1094981	0.4277186	2.594	0.0097
globalSales	-2.4531025	3.4130167	-0.719	0.4726

Table 9: Coefficient of the least squares model for the regression of *meanHours* on *positiveReviews, negativeReviews, criticNumberReviews,* and *globalSales*.

We can observe that the p-values obtained are significant for all of the variables, with the exception of *globalSales*. This indicates that *globalSales* does not have an influence on the response *meanHours*.
We now wish to quantify the extent to which the model fits the data. In order to do this, we provide two measures of fit, the residual standard error (RSE) and the R^2 statistic:

Quantity	Value
Residual standard error	168.4
R^2	0.28

Table 10: Residual standard error (RSE) and the R^2 for our regression model.

We find that the R^2 assumes a value of 28%, meaning that only a small portion of the variability in the response has been explained by the regression. This indicates that the multiple linear regression model does not fit the data well.

3.5 Regression Diagnostic

There are several possible causes for the poor performance of the linear regression model on our data. We recall that a few assumptions are made when fitting a linear regression model:

- **Linearity**: the response Y can be expressed as a linear function of the predictors X_j
- **Homoscedasticity**: the variance of the residuals is constant
- **Normality**: the residuals are normally distributed

In order to check the validity of these assumption, and to check whether the results of previous tests are significant, we are going to make use of residual plots.
We first plot the residuals $e_i = y_i - \hat{y_i}$ versus the fitted values $\hat{y_i}$.

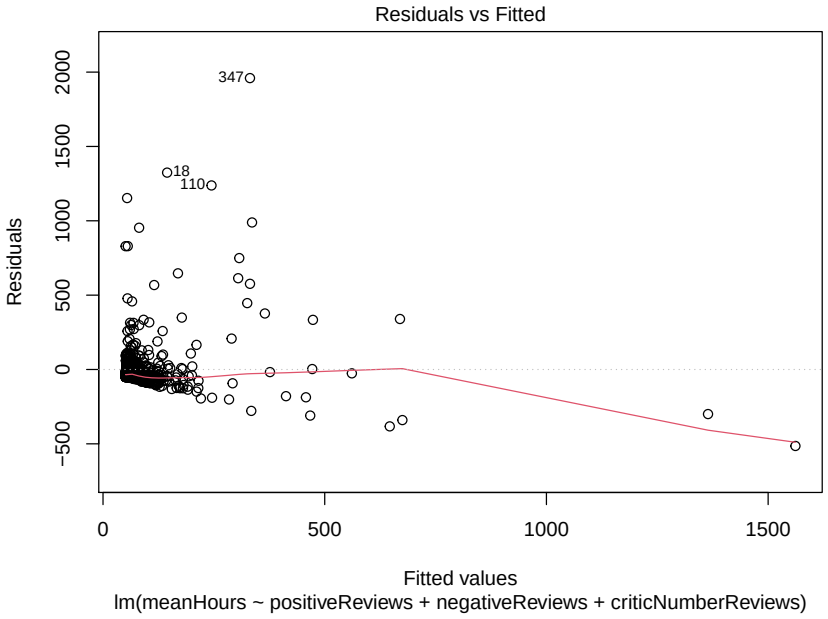

Figure 11: Plot of residuals versus predicted values for our data.

If the linearity assumption is met, we should see no discernible pattern, and the red line should be fairly flat. However, from Figure 11 we can observe that the line is horizontal in the first half of the plot, but then it decreases. This provides evidence of non-linearity in our data.
In order to identify non-constant variance of the residuals, we compute the following plot where the residuals have been re-scaled:

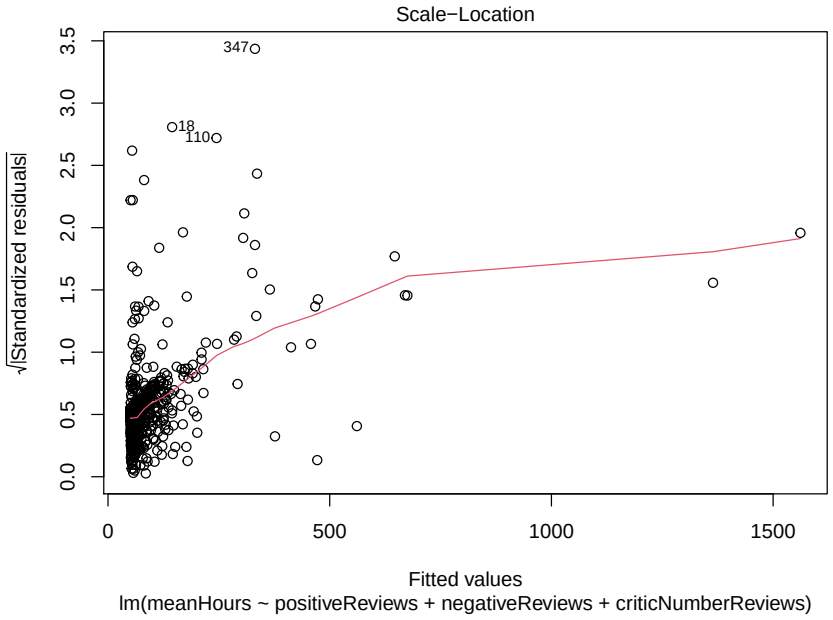

Figure 12: Plot of the square root of the absolute value of the standardized residuals versus predicted values for our data.

From Figure 12 we can observe that the variances of the residuals increase with the value of the response. We can thus conclude that the assumption of constant variance of the residuals is false, and that our data exhibits evidence of *heteroscedasticity* instead.
In order the check the assumption of normality of the distribution of the residuals, we provide the following Q-Q plot:

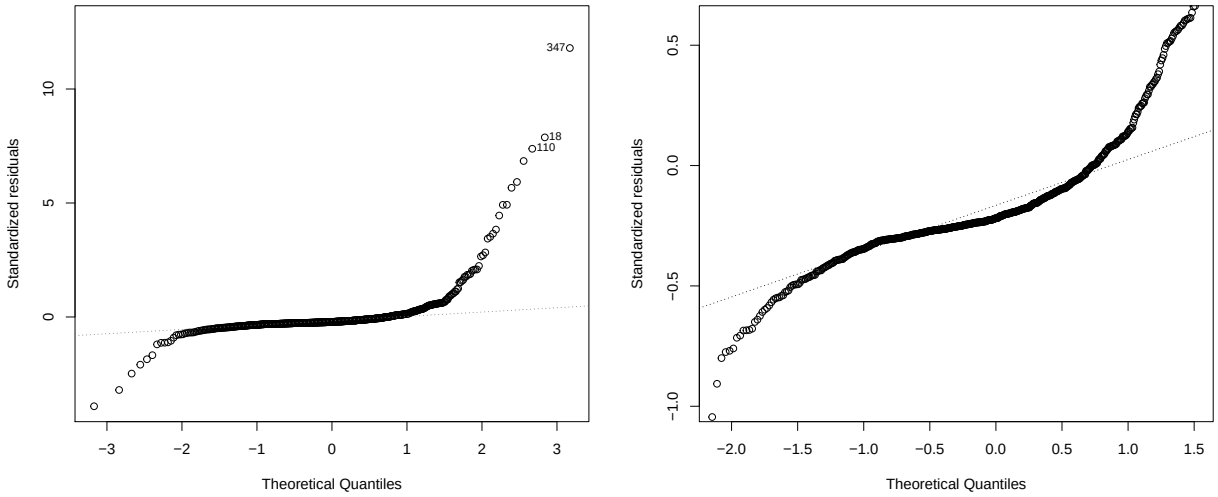

Figure 13: On the left, normal quantile-quantile plot of the residuals. On the right, the same plot is shown, where the scale has been readjusted.

Let us recall that if the residuals are approximately normally distributed, then the normal Q-Q plot of those residuals will result in an approximately straight line. However, we can observe from Figure 13 (left)

that the points curve away from the line at each end. Additionally, in order to obtain a better view of the shape, we also provide the plot on the right where we changed the scale, in order to zoom in on the central part of the plot and disregard the tails. We can see that the right plot in Figure 13 highlights the "S" shape of the distribution. This curvature indicates that the distribution is heavy-tailed, indicating that the residuals do not follow a normal distribution.
Lastly, in order to detect any outliers or high leverage points, we are going to provide a plot of the standardized residuals versus the leverage:

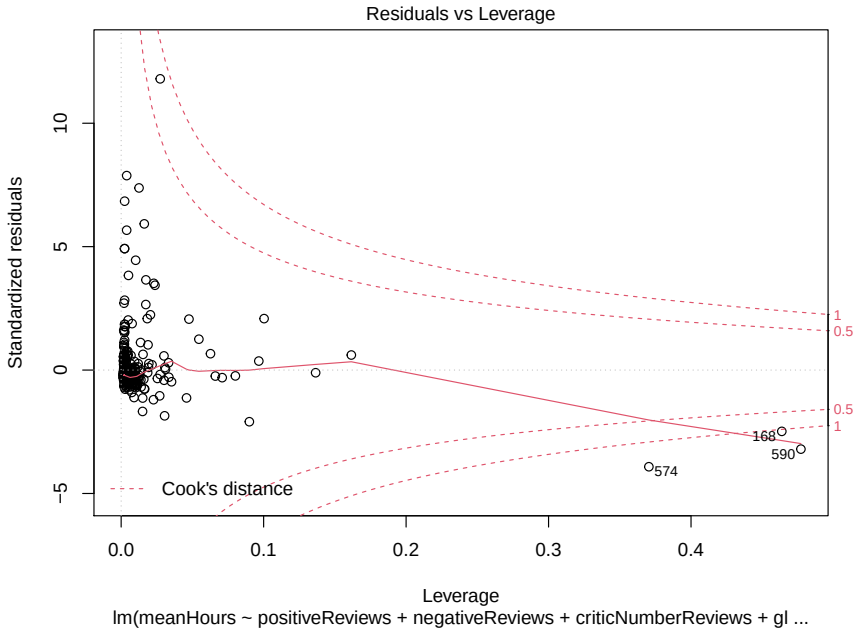

Figure 14: Plot of the standardized residuals versus the leverage. The dashed line represents Cook's distance metric.

Leverage provides a measure of how much an individual observation influences the fit of our model. We can observe from Figure 14 that most of our points reside on the left side of the plot, with low leverage, and only a few points have high leverage.

	positiveReviews	negativeReviews	criticNumberReviews	globlSales	meanHours
168	1552	190	0	1.50	14.20650
574	111	107	23	0.09	11.48478
590	378	231	13	0.32	30.70691

Table 11: High-leverage points identified in Figure 14.

We removed these high leverage points (168, 574, and 590 from Figure 14) and refitted our model. We found that this did not lead to any significant change or improvement in the results of our fit. Additionally, we do not observe any points in the top-right corner of the plot in Figure 14, thus we do not detect outliers.

In the next sections of our analysis, we will apply several different statistical learning methods. Therefore, it is important to decide which method produces the best results. In order to evaluate the performance of the multiple linear regression method applied thus far, we provide the test MSE. Let us recall that the test MSE is given by

$$\text{Test MSE} = \text{Ave}(y_0 - \hat{f}(x_0))^2 \tag{15}$$

where (x_0, y_0) are test observations that were not used in training the model.
The Test MSE obtained after 10-fold Cross Validation is the following:

$$\text{Test MSE}_{\text{cv}} = 28335.3 \tag{16}$$

3.6 Linear Discriminant Analysis

In order to understand if we can obtain better results to those obtained by linear regression, we will now perform LDA on our data. We are interested in predicting whether a game will have a high or low value for *meanHours* on the basis of *positiveReviews, negativeReviews*, and *criticNumberReviews*. We have decided not to include *globalSales*, since we found in the previous section (Table 9) that it is not related to the response. We build a binary classifier that contains "high" if *meanHours* contains a value above its median, and "low" if *meanHours* contains a value below its median. In our case, we have a number of predictors $p = 3$ and a number of classes $K = 2$.
In Linear Discriminant Analysis with p predictors, the discriminant function can be written as

$$\delta_k(x) = c_{k0} + c_{k1}x_1 + c_{k2}x_2 + ... + c_{kp}x_p \tag{17}$$

After estimating the unknown coefficients, LDA will classify a new observation to the class for which $\hat{\delta}_k(x)$ is largest.
We apply the LDA fit on our training set and obtain the following prior probabilities for the two groups:

π_{high}	π_{low}
0.5131313	0.4868687

Table 12: Fraction of the training observations that belong to the *high* and *low* class.

Given the way that we have constructed our classifier, it is reasonable that approximately 50% of the data in our training set is split below and above the median.
We provide a summary of the group means for the two groups for each predictor:

	positiveReviews	negativeReviews	criticNumberReviews
high	8615.5827	1074.3701	13.641732
low	825.7967	173.4232	6.564315

Table 13: Group means for the different predictors.

We provide the coefficients obtained:

Variable	LD coefficient
positiveReviews	-3.137263e-05
negativeReviews	-5.642294e-05
criticNumberReviews	-3.874296e-02

Table 14: Coefficients produced by the LDA fit.

Given that we have a binary classifier, it can make two types of errors: it can incorrectly assign a game with a low value of *meanHours* to the high category, or it can incorrectly assign a game with a high value of *meanHours* to the low category. We provide the confusion matrix produced by our model:

	high	low	Total
high	38	19	57
low	38	70	108
Total	76	89	165

Table 15: Confusion matrix comparing the LDA predictions to the true classes.

We can observe from the off-diagonal elements in Table 15 that 38 out of 76 highs were incorrectly classified as lows (50%), and 19 out of 89 lows were incorrectly classified as highs (21%). We can see that LDA does worse at classifying highs than lows.
We compute the overall misclassification rate and obtain:

$$\text{error rate} = 0.3454545 \simeq 34\% \tag{18}$$

or, equivalently

$$\text{accuracy rate} = 1 - \text{error rate} \simeq 66\% \tag{19}$$

Finally, we provide partition plots:

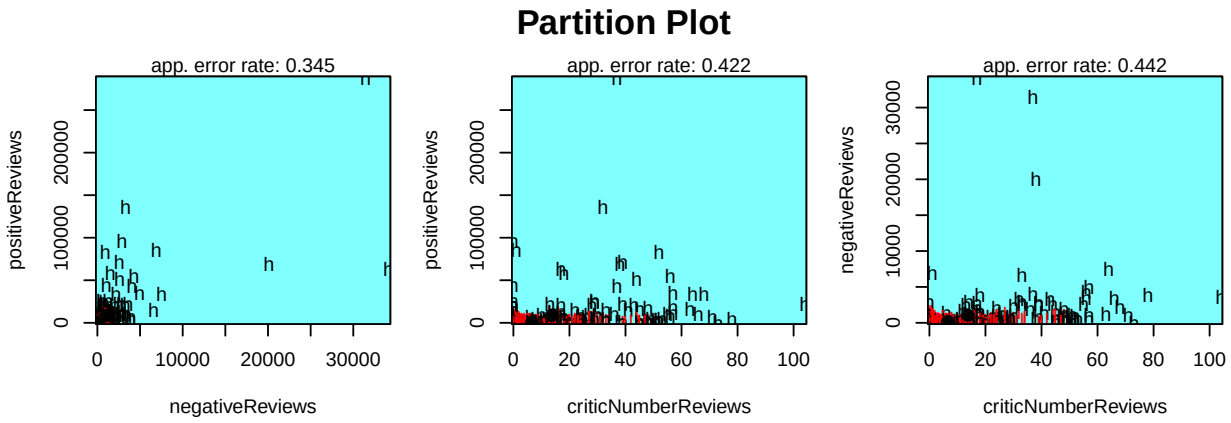

Figure 15: LDA partition plots. The black indicates the correctly classified games, and the red indicates the incorrectly classified games.

3.7 Quadratic Discriminant Analysis

In order to understand whether we can improve on the results obtained from LDA, we are going to fit a QDA model to our data. While LDA assumes that the covariance matrices of each class are the same, this is not the case for QDA, thus making it a much more flexible classifier than LDA, which could lead to improved prediction performance.
We apply the QDA fit and obtain the following confusion matrix:

	high	low	Total
high	28	5	33
low	48	84	132
Total	76	89	165

Table 16: Confusion matrix produced by QDA.

We compute the error rate and obtain:

$$\text{error rate} = 0.3212121 \simeq 32\% \tag{20}$$

or, equivalently

$$\text{accuracy rate} = 1 - \text{error rate} \simeq 68\% \tag{21}$$

We can observe a slight improvement in the prediction accuracy of our model compared to that obtained previously.

3.8 K-Nearest Neighbors

We are now going to apply K-Nearest Neighbors classification, which is a non-parametric method that does not rely on any stringent assumptions about the underlying data. KNN classifies a new test observation x_0 to the response class j that maximizes

$$Pr\left(Y = j|X = x_0\right) = \frac{1}{K} \sum_{i=\mathcal{N}_0} I(y_i = j), \tag{22}$$

where K is a positive integer, and $\mathcal{N}_0$ represents the training data that are closest to x_0 in Euclidean distance. In other words, in KNN a new observation is assigned to the class that gets the most votes out of the K nearest neighbors.
In order to select the best value for K, we are going to make use of 10-fold Cross Validation. The following plot shows the cross-validated accuracy per different values of K:

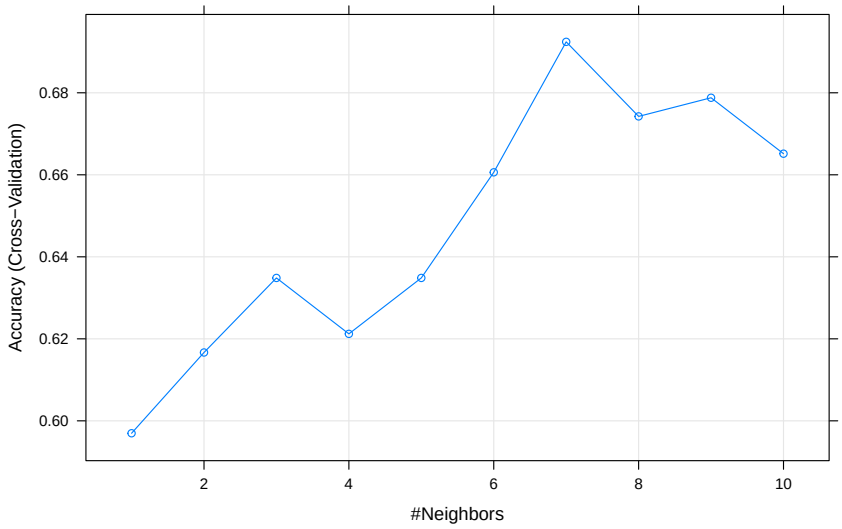

Figure 16: Model accuracy as a function of the number of neighbors K.

We can observe from Figure 16 that the accuracy is highest when K=7. We then fit the KNN model using K=7 and obtain the following confusion matrix:

	high	low	Total
high	43	26	69
low	33	63	96
Total	76	89	165

Table 17: Confusion matrix produced by KNN, with K=7.

We can observe that the elements in the diagonal of Table 17, which represent the number of correctly classified games, are higher than the incorrectly classified games for both the *high* and *low* category. We compute the misclassification rate and obtain:

$$\text{error rate} = 0.3575758 \simeq 36\% \tag{23}$$

or, equivalently

$$\text{accuracy rate} = 1 - \text{error rate} \simeq 64\% \tag{24}$$

We find that KNN performs slightly worse than both LDA and QDA.

3.9 Regression Trees

In this paragraph we will look into another technique of regression that we can apply to the dataset. A more intuitive way to look at our dataset lies in the Tree-Based Methods, also called Classification and Regression Trees as well as Decision Trees: a tree model is a set of "if-then-else" rules that are easy to understand and implement; in contrast with the linear regression, trees have the ability to emulate the human mind decisional process making the outcome of the analysis easy to interpret.

Typically, the tree is plotted upside-down, so the root is at the top and the leaves are at the bottom: we can interpret the regression tree understanding that the top parameter is the most important factor followed by a top-down path from the roots (most important factor) through the branches (divided by nodes), to the least important one (leaves).

There are two very important steps to be addressed:

- How do we iterate the splits?
- How many predictors do we use?

In building a regression tree we use a technique, a top-down approach, called recursive binary splitting (or recursive partitioning): at each step we divide the predictor space into two non-overlapping regions making the best split possible that will be indicated via two new branches further down on the tree. The best split is evaluated trying to choose the cut-point that will lead to the greatest reduction in the RSS. This process of splitting and evaluating the maximum reduction of RSS in the new regions that have been created continues until a stopping criterion is reached; for instance, we may continue until no region contains more than a decided number of observations or until all the predictors have been used.

For the choice of the predictors we use the same numerical ones used in the previous section.

Growing a full tree will result in something impossible to understand with 129 terminal nodes, hence the picture of the full grown tree will not be included: the splits are too close and the tree contains too much information. Even if it is difficult to visualize, we can calculate the error performing a Test MSE: the result is 76093.98. What we can do is pruning the tree: in practice we want to cut some branches in order to not overfit the data and make the outcome more understandable for a human eye. The cv.tree() function in R reports the number of terminal nodes of each tree considered as well as the corresponding error rate so that we can choose the optimal size for our tree.

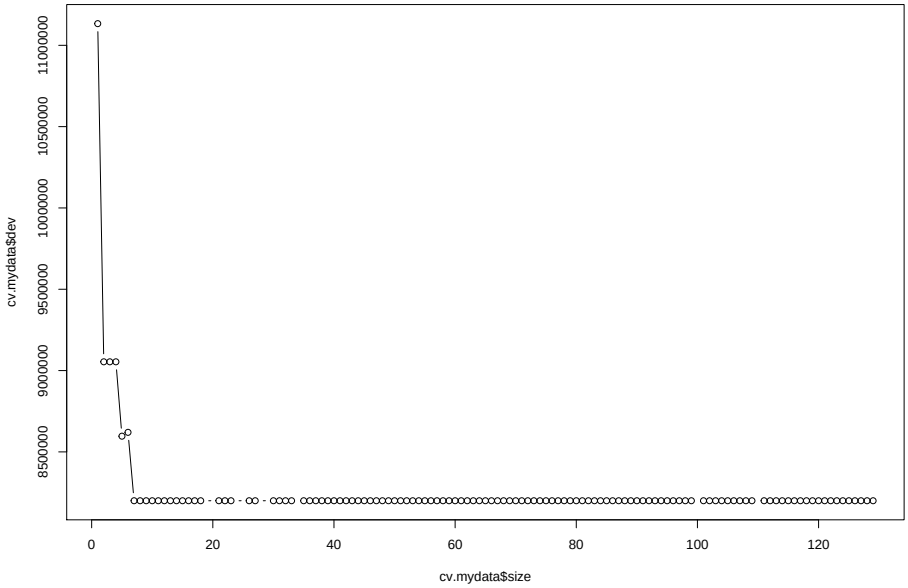

Figure 17: Cross validation error for different size of the Tree.

As we can notice, any size bigger than 7 does not improve the accuracy of the regression, so we can proceed in cutting the branches of our overgrown tree using the function prune.tree()

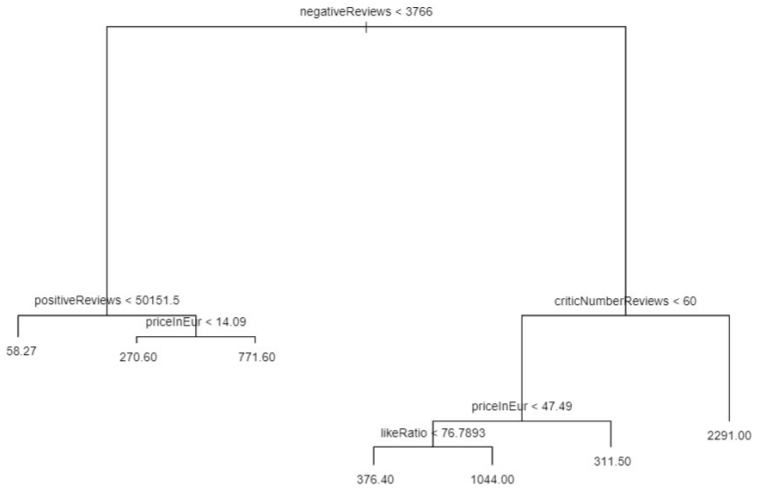

Figure 18: Pruned tree.

The outcome is easier to understand and shows us that *negativeReviews*, *positiveReviews*, *priceInEur*, *criticNumberReviews* and *likeRatio* are the most important parameters to consider.

The Test MSE value resulting from pruning the tree is 75290.89, showing an improvement from the one of the full grown tree.

Another technique that can be used is Bootstrap Aggregation or Bagging, that is a general procedure for reducing the variance of a statistical method: we take many training sets from the population, build a separate prediction model using each training set, and average the resulting prediction. With trees we simply construct N regression trees using N bootstrapped training sets, and average the resulting predictions. These trees are grown deep, and are not pruned. Hence, each individual tree has high variance, but low bias.

We want to use the function Random Forest that provides an improvement over bagging decorrelating the trees: at each split we force the sampling process to use only a limited number m of parameters, that we will accurately choose observing how the function acts on our dataset when run. First we control how many trees we need to use in our function creating a plot to easily visualize it:

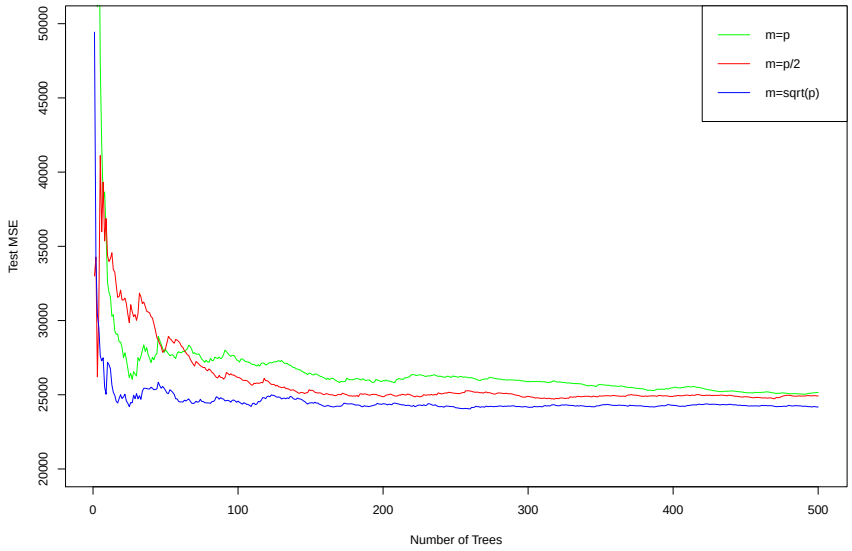

Figure 19: Result from random forest referred to our dataset. The test error is displayed as a function of number of trees. Using random forest with $m = \sqrt{p}$ leads to a slight improvement over the other numbers of predictors.

In our scenario we can obtain the best result considering $m = \sqrt{p}$ parameters at each split. In order to check if using Random Forest improved our prediction of meanHours we can derive the error that is 25223.51, improving substantially compared to a simple pruned tree. One last important result to check if the results obtained with Random Forest are consistent with the ones obtained before, we can print a table of the variables, their importance in the calculation of the MSE and their influence on the node purity every time that Random Forest makes a choice in the predictors at each split:

Variable	%IncMSE	IncNodePurity
priceInEur	2.8680006	237202.99
positiveReviews	9.8806570	1883120.41
negativeReviews	13.3810947	1403155.67
likeRatio	2.7597510	577322.88
userNumberReviews	-0.1585408	516756.50
userMeanValue	0.6647204	257826.27
criticNumberReviews	-1.1575860	360464.30
criticMeanValue	1.7667513	539737.49
numberOfReviews	1.3830042	84118.02
globalSales	-2.1207849	268957.79

Table 18: Percentage of influence of each variable in the TEST MSE

If we look at the values we can notice that again *negativeReviews*, *positiveReviews* and *priceInEur* are the most important variables.

We can deduce, by looking at our trees, that a high number of negative reviews leads to a counter-intuitive result: **games that are strongly criticized by users are also the most played**, as the *meanHours* value predicted confirms.

As it was already said in the genre analysis in Section 2, a good game is not necessarily a highly rated one, and its quality should not be the main aim of a developer because it is not entirely connected to the number of hours that a player will spend on it.

4 Time Series Analysis

4.1 Introduction

The following section is dedicated to the analysis of our time series data, in order to investigate how video game sales change over time. The analysis of historical data can in fact provide key insights into the video game market, and help us identify whether there was any point in time that favored the proliferation of video games. Our data includes sales for video games that were released between 1995 and 2020.

Firstly, we provide an overview of global sales over this entire time window:

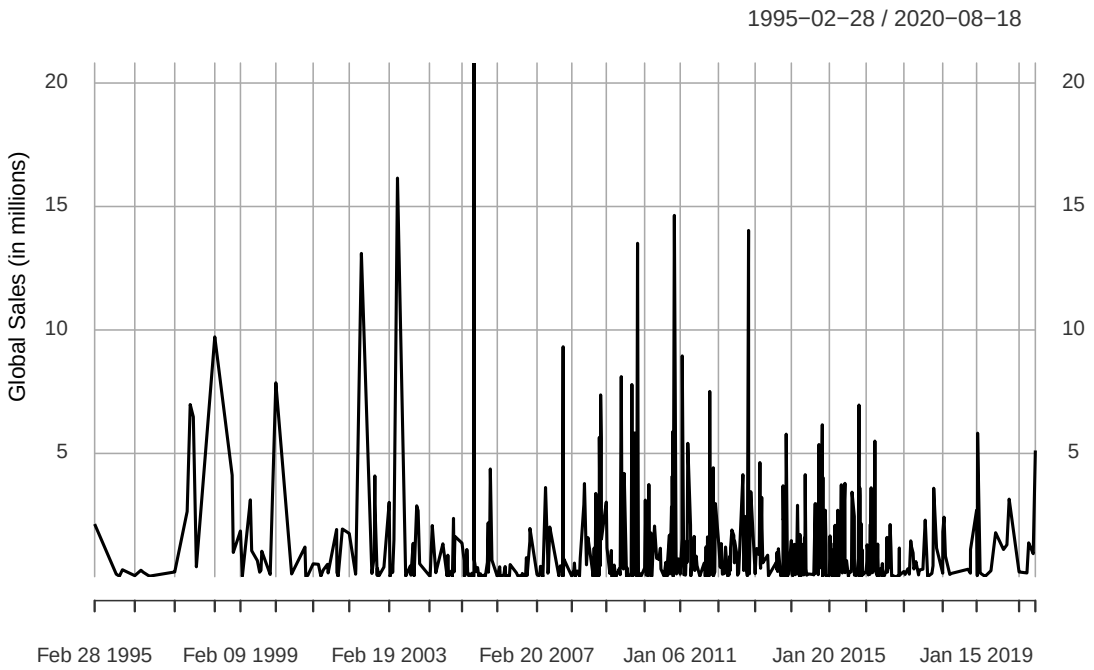

Figure 20: Video game sales over time.

The following plot, instead, shows the total number of sales per each year:

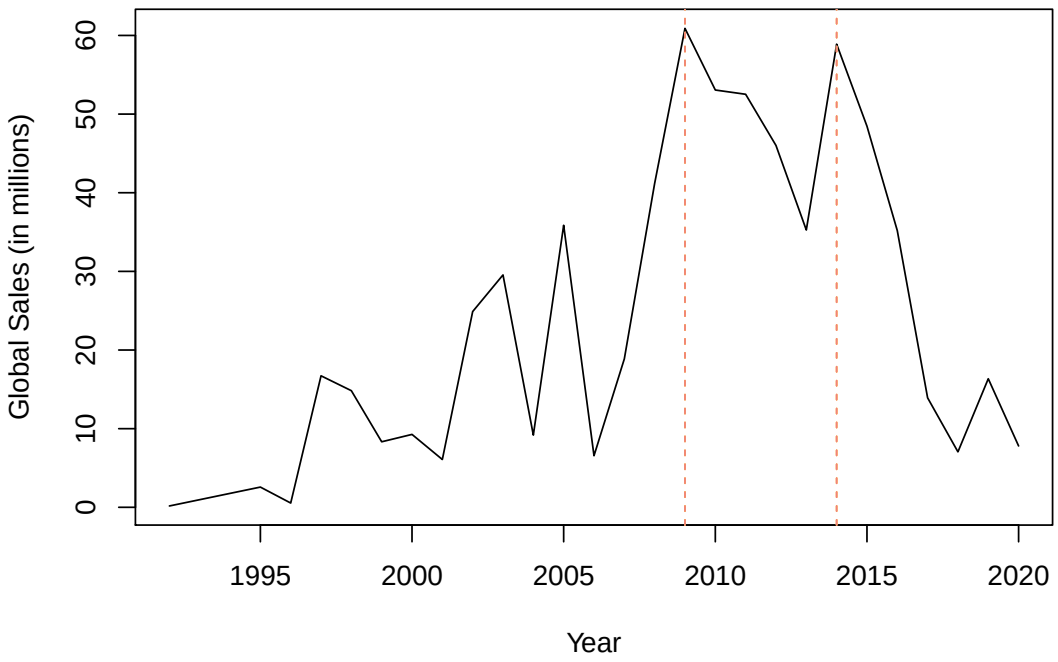

Figure 21: Total number of copies sold globally over the years. The vertical lines identify the two peaks.

We can observe from Figure 21 that the sales grow until they reach a peak in 2009, and then decrease after reaching a second peak in 2014. We now wish to explore the significance of these two peaks. Therefore, we divide our data into 3 sets: $T_1 = [1995, 2009]$, $T_2 = [2010, 2014]$, and $T_3 = [2015, 2020]$.

4.2 Dickey-Fuller Test

We will now perform a stationarity test.
Let y_t be the value of the response (*globalSales* in our case) at time t. Let us assume that our model takes the form

$$y_t = \alpha + \rho y_{t-1} + \epsilon_t \tag{25}$$

where α is a constant, ρ is the parameter of the model, and ϵ_t is an error term. This model is also known as AR(1) model, that is, an Autoregressive model of order 1. This process is said to exhibit *stationarity* if $\rho < 1$, whereas it is non-stationary if $\rho = 1$.
In order to test for *stationarity* we are going to run a Dickey-Fuller test. We test the null hypothesis

$$H_0 : \rho = 1 \text{ (non-stationary)} \tag{26}$$

against the alternative

$$H_a : \rho < 1 \text{ (stationary)} \tag{27}$$

We first run the test on the entire time series. The following table shows the Dickey-Fuller statistic, as well as the associated p-value:

Quantity	Value
Dickey-Fuller	-8.3713
p-value	< 0.01

Table 19: Results produced by the Dickey-Fuller test on the entire time series.

By looking at the p-value we can reject the null hypothesis and conclude that our data is stationary. However, this result is obtained by running the test on the entire time series, that is, from 1995 to 2020. Since this is a very large time window, we are going to run 3 separate tests for each set T_1, T_2, and T_3.

	1995-2009	2010-2014	2015-2020
Dickey-Fuller	-6.0471	-5.348	-4.7779
p-value	< 0.01	< 0.01	< 0.01

Table 20: Results produced by the Dickey-Fuller test for the 3 different sets.

We can observe that all 3 sets exhibit stationarity.

4.3 Monthly trends

In this section, we wish to investigate our data on a monthly basis. Therefore, we provide the following seasonal plots:

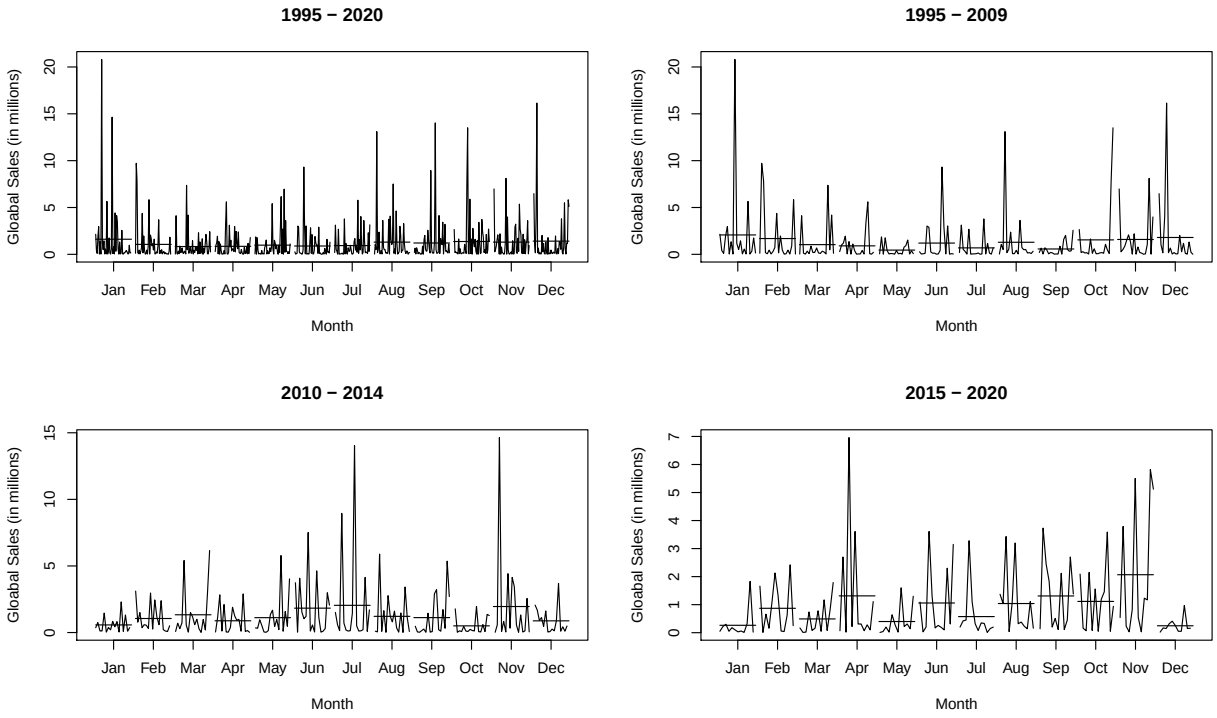

Figure 22: Monthly-based time series of global sales for the different time sets. The horizontal line identifies the mean value of global sales for each month.

Figure 22 shows subseries plots. The top-left plot refers to the entire time window, while the other three refer to the 3 time sets. Each subseries is plotted separately for each month, in order to emphasize seasonal patterns. By looking at the top-left plot, we do not discern any specific pattern. This could be due to the time window being very large, which could smooth out any seasonal effects. In the top-right plot we do not detect trends and we can see that the horizontal lines are approximately all centered around the same value. In the bottom-left plot we can observe that the monthly means are slightly more varied, with November being the month with the highest average sales. The bottom-right plot shows more variation compared to the previous time periods, with again November leading in average sales. In the main, we do not detect seasonality throughout any of the time periods.

Another technique to help us visualize **seasonality** are autocorrelation plots.
Let y_t and y_{t-1} be the value of the response y at time t and $t-1$ respectively. The i-th autocorrelation coefficient is a measure of the relationship between the variables y_t and y_{t-i}. These variables are known as *lagged* variables.

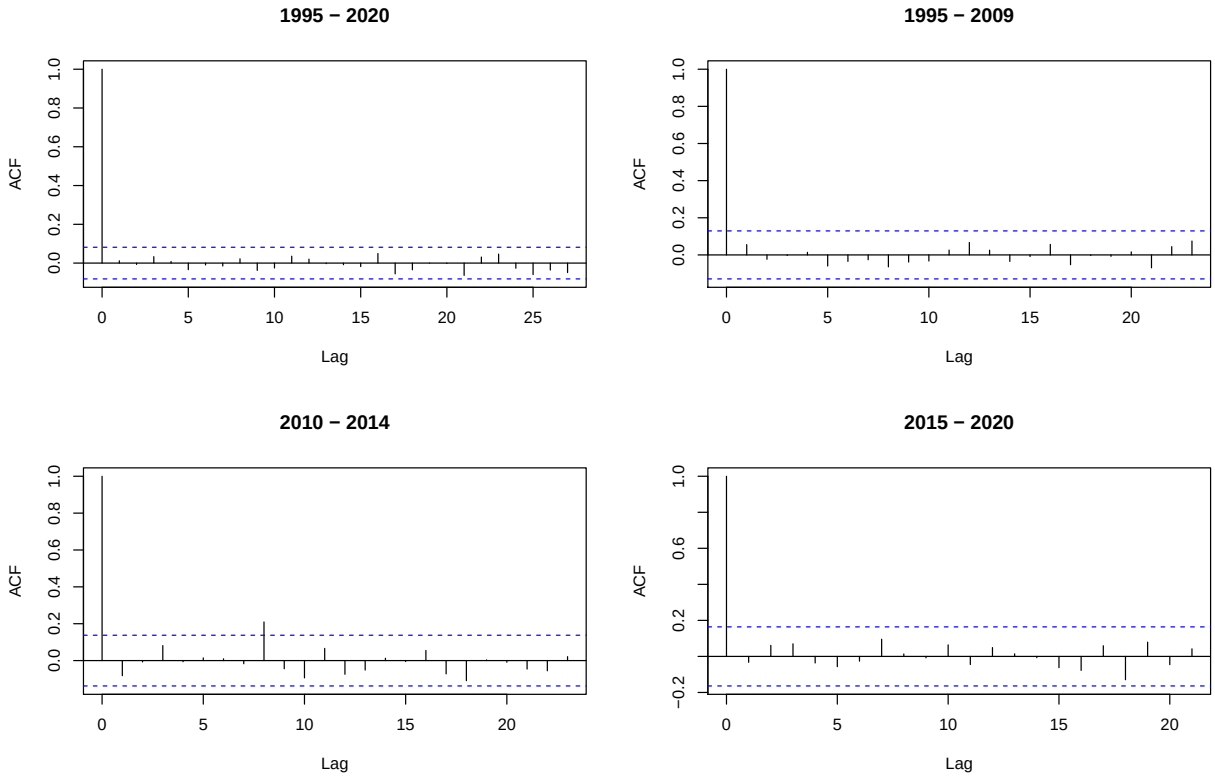

Figure 23: Autocorrelation coefficient plotted against the lagged variables for the different time sets. The horizontal dashed blue lines indicate the 95% confidence intervals. The vertical lines indicate the autocorrelation coefficient at each Lag order.

Figure 23 shows autocorrelograms. In these plots, only the vertical lines that exceed the dashed lines are considered significant. The first vertical line at the 0 lag order always assumes value 1, and it is due to the fact that it represents the autocorrelation coefficient of the variable with itself. Therefore, we can disregard this line. As for the rest of the vertical lines, we can see from Figure 23 that most of them do not exceed the horizontal lines, hence the autocorrelation coefficients are not significant. We also do not observe any increasing or decreasing trend, which provides further evidence of no seasonality in our data.

4.4 Chow Test

Finally, we will perform a Chow test for detecting a structural break in our time series. A structural break is defined as a point in time after which the time series exhibits a significantly different trend than previously experienced.
Before performing the Chow test we will check the assumptions that are needed to perform it. The first assumption is that the residuals of the model are normally distributed, while the second assumption is that the variance of the residuals is constant. We provide the normal Q-Q plot below:

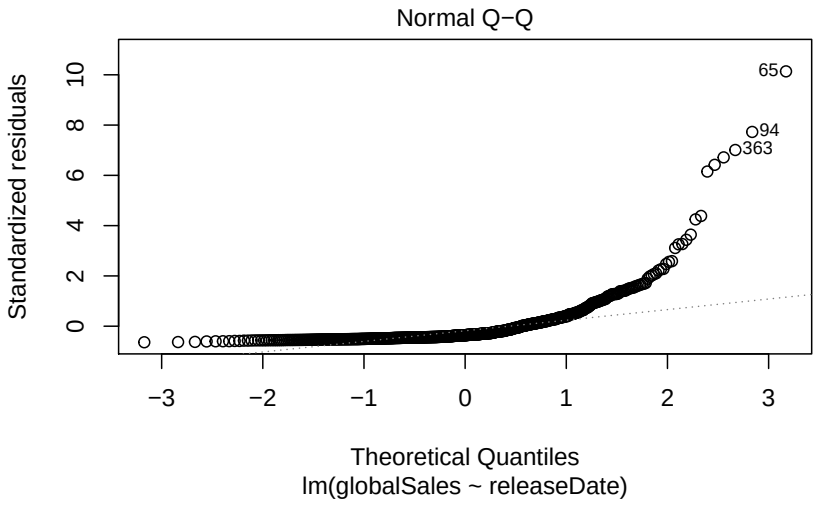

Figure 24: Normal quantile-quantile plot of the residuals

The graph above is useful to understand if the residuals of global sales are normally distributed, in order to check the first assumption. We can see that the residuals in the end of the distribution do not follow the straight line. For this reason we can say that the residuals are not exactly normally distributed, so the first assumption is violated. We will now show a plot for the variances of the residuals:

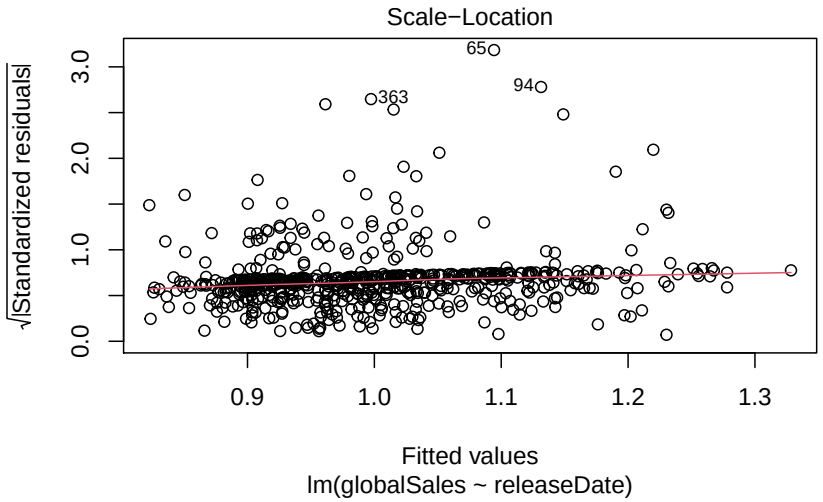

Figure 25: Square root of absolute value of the standardized residuals VS predicted values of the data

By looking at the plot shown above we can see that the red line does not increase or decrease, so the variance is constant over time and we can conclude that the second assumption is true. Since the first assumption is not confirmed, the test we are going to perform will not be very powerful.
The Chow test checks for a structural break between two time periods by looking at the regression parameters. Let $[1, T]$ be a time period and let K be a time period such that $K \in [1, T]$. The Chow test will test for a time break at time K by assuming that the data is modelled as

$$\begin{cases} y_t = a_1 + b_1 x_t + \epsilon, & \text{for } t \in [1, K] \\ y_t = a_2 + b_2 x_t + \epsilon, & \text{for } t \in [K+1, T] \end{cases} \tag{28}$$

The null hypothesis states that

$$H_0 : a_1 = a_2, b_1 = b_2. \tag{29}$$

In light of what we observed in Figure 21, we run two separate Chow tests to check whether the two peaks in 2009 and 2014 represent a structural break. The first time period goes from 1995 to 2014 (we will use this to check if 2009 is a structural break), and the second time period goes from 2009 to 2020 (we will use this to check if 2014 is a structural break).
We obtain the following results:

	1995-2014	2009-2020
F-statistic	0.17222	1.4519
p-value	0.8419	0.2356

Table 21: Results produced by the Chow test for the two time periods.

Based on the results obtained in the first time window, we cannot reject the null hypothesis because of the value of the F-statistic. As for the second time window, the p-value is higher than level $\alpha = 0.05$, so we cannot reject the null hypothesis. For these reasons, we do not detect any structural change in 2009, nor in 2014. In conclusion, we can see that there is no pattern in the global sales of video games, and there is no golden period for the market because the peaks we observed in Figure 21 do not represent a structural change.

5 Investigating the association between aggressive behavior and playing video games

In this chapter, we are going to examine whether playing violent video games has an influence on increasing level of aggressive behavior of teenagers.

5.1 Introduction

Nearly 3.1 billion people worldwide play video games, which is about 40% of the world population. A new study [1] from the Pew Research Center indicates 59% of girls and 84% of boys 13-17 years old do play video games.

This fact is generating a concern over the world about potential *negative consequences* of this activity. Some researchers [1] are concluding that playing video games has social advantages. At the same time, some of them [2] criticize video games as one of the factors of increasing aggressiveness, saying that the the environment of games is conducive to mass shootings.

The main theoretical framework used to study the links between violent game engagement and aggression has been the general aggression model [3]. The GAM is a social learning theory that proposes that repeated exposure to violent media increases the accessibility of aggressive thoughts, which increases the probability of aggressive cognitive schema, emotions and behavior [4].

However, new researchers [5] shows that GAM a is quite outdated study to discover correlation between aggressive behavior and playing video games and it is needed to look more deeply in this increasing nowadays concern.

We can divide our research into two main hypothesis:

1. Hypothesis 1: Adolescent's who are playing violent video games are having higher level of aggression.
2. Hypothesis 2: Depending on gender, there is a difference between hours spent on gaming and level of anger.

5.2 Data Collection

For this survey-based study, a large and nationally representative sample of British adolescents was recruited, and quantified recent video game play using a gaming engagement measure [21] [1]. Levels of violent game contents were operationalized using European Union (Pan European Game Information, PEGI) and North American media rating systems (Entertainment Software Rating Board, ESRB). Youth aggression and prosocial behavior was measured through carers responses using a behavioral screening questionnaire [6], [23].

5.3 Data Sampling

Authors of the study [21], during the design stage have calculated that a sample size of 443 is required to attain the desired power level for investigating this problem. A target sample of 1000 adolescents (500 females, 500 males) was set. A total of 1004 adolescents and an equal number of their carers were recruited to complete online self-report questionnaires. The final sample was evenly divided among 14-year-old (n = 497) and 15-year-old adolescents (n = 507). Further, 540 participants identified as male, 461 as female and 3 as another gender orientation. The sample was predominantly white as only 8.1 % of participants reported they were from Black and other minority ethnicity's. The total combined household income covered the general distribution and varying from £6500 (1.9 %) to £150 000 or more (2.8 %).

5.4 Data-set description

The data set contains 276 variables and 1004 observation. As the data set was formed from survey, most variables share the ordinal or nominal type. Data-set was not linked to the previous one, that was mentioned in Chapter 1.

[1] The data-set was taken from the research of the University of Oxford. The study underwent ethical review and received approval from the Central University Ethics Committee of the University of Oxford (C1A17023))

Group	Variable	Type	Description
Adolescent' information	responseId	Quantitative	Unique number of response made by young person and his/her carer.
	childage	Quantitative	Adolescent's age.
	gor	Nominal	Region of living of adolescent. On a scale from 1 to 14, depends on region in GB.
	gender	Nominal	Gender of child. Scale from 1 to 4 depends on gender.
Carer-taker information	d3	Nominal	Annual income of household of caretaker.
	d5	Nominal	Ethnic group of the carer. Scale from 1 to 16, depending on ethnic group.
	d7	Nominal	Relationship between caretaker and teenager. Scale from 1 - Mother to 5 - Prefer not to say.
	dob	Nominal	Date and year of birth of caretaker.
Questions to adolescent	a2	Nominal	Question from survey, asked to child, "Do you play video games?". Scale is from 1 - Yes to 2 - No [22].
	a10.1 - a10.12	Ordinal	List of questions to adolescent's to measure physical, verbal, anger, hostility problems. Likert scale from 1 - very unlike me to 5 - very like me [**?**].
	agg.01 - agg.12	Quantitative	Secondary variables a10.1 - a10.12 transformed into quantitative ones.
	a4.1,a4.2,a4.3	Nominal	The question "How do you play video games?". Categories are from 1 - Personal computer 2 - Smartphone to 3 - Console [22].
	a5.1,a5.2,a5.3	Nominal	The question asked to adolescent "Do you play video games with other people?". Scale is from 1 - Yes, online only, 2 - Yes, offline only, 3 - Yes, online and offline 4 - No [22].
	a6.1,a6.2,a6.3	Nominal	The question is "About how many hours a day do you usually play video games?". Scale is from 1 to 9 depends on hours spent playing [22].
Traits of the adolescent	trait.hostility	Quantitative	Secondary variable. Level of hostility of adolescent taken from survey [22]. Average of variables agg.03, agg.07, agg.11 [**?**].
	trait.anger	Quantitative	Secondary variable. Level of anger of adolescent taken from survey [22]. Average of variables agg.06, agg.09, agg.10. Scale is from 1 - the minimum level to 5 - maximum [**?**].

Group	Variable	Type	Description
Traits of the adolescent	trait.physical	Quantitative	Secondary variable. Level of physical aggression of adolescent taken from survey [22]. Average of variables agg.01, agg.04, agg.08. Scale is from 1 - the minimum level to 5 - maximum [**?**].
	trait.verbal	Quantitative	Secondary variable. Level of verbal aggression of adolescent taken from survey [22]. Average of the variables agg.02, agg.05, agg.12. Scale is from 1 - the minimum level to 5 - maximum [**?**].
Questions to carers	ITEM01 - ITEM25	Ordinal	Questions that were set to carers about social behavior of adolescent from SDQ Survey [6],[23], where description of each ITEM is. The Likert scale is from 1 - Not true, 2 - Somewhat true, to 3 - Certainly true.
	conduct.problems	Quantitative	Secondary variable. Ability of conducting problems by adolescent's answered in survey SDQ [6],[23]. Sum of variables ITEM.05, ITEM.07r, ITEM.12, ITEM.18, ITEM.22.
	prosocial	Quantitative	Secondary variable. The prosocial level of adolescent's answered in survey SDQ [6], [23]. Sum of variables ITEM.01, ITEM.04, ITEM.09, ITEM.17, ITEM.20.
Game specifications	a3.1, a3.2, a3.3	Nominal	The 3 games young person played the most recently. Each variable contains different name of the game [22].
	rc.game01.edition, rc.game02.edition, rc.game03.edition	Nominal	Specific information on the edition/version of the games respectively for a3.1, a3.2 , a3.3.
Time, spent on gaming	game.one.time, game.two.time, game.three.time	Quantitative	Secondary variable. Time spent on gaming for each game written by teenager in survey [22]. Recoded variables a6.1, a6.2, a6.3 from nominal into quantitative. Range from 0 to 7, hours per day.
	gaming.time	Quantitative	Secondary variable. Sum of variables game.one.time, game.two.time, game.three.time. Measured in hours per day.

Group	Variable	Type	Description
Violent game indicators	pegi.one.violence, pegi.two.violence, pegi.three.violence	Binary	Variable, that shows if the game contains depictions of violence by rating of Pan European Game Information [7]. Category scale is from 0 - Does not contain to 1 - Does contain.
	esrb.01.violence, esrb.02.violence, esrb.03.violence	Binary	Variable, that shows if the game contains scenes involving aggressive conflict by rating of Entertainment Software Rating Board [8]. Category scale is from 0 - Does not contain to 1 - Does contain.
Time spend on violent games	game.one.violent.time, game.two.violent.time, game.three.violent.time	Quantitative	Secondary variables. Equation of variables (pegi.one.violence x game.one.time),(pegi.two.violence x game.two.time), (pegi.three.violence x game.three.time) correspondingly. Measured in hours per day.
	game.01.violent.time.esrb, game.02.violent.time.esrb, game.03.violent.time.esrb	Quantitative	Secondary variables. The product of variables (esrb.01.violence x game.one.time), (esrb.02.violence x game.two.time), (esrb.03.violence x game.three.time) correspondingly. Measured in hours per day.
	game.violent.time	Quantitative	Secondary variable. The sum of variables: game.one.violent.time, game.two.violent.time, game.three.violent.time. Measured in hours per day.
	violent.game.time.esrb	Quantitative	Secondary variable. The sum of variables: game.01.violent.time.esrb, game.02.violent.time.esrb, game.03.violent.time.esrb. Measured in hours per day.
Additional variable	violent.game.player	Binary	Secondary variable, indicates if adolescent is playing violent video games or not. If the game.violent.time greater than 0, violent.game.player = 1, otherwise = 0.

Table 22: Description and type of variables in behaviour data-set.

5.5 Exploratory Data Analysis

Firstly, we intend to describe the data set in order to understand the participants of the survey. The analysis in this section was carried out using IBM SPSS Statistics Software. Let us look at the frequency table of the age of participants of survey:

	Frequency	Percent	Valid Percent	Cumulative Percent
14	497	49.5	49.5	49.5
15	507	50.5	50.5	100
Total	1004	100	100	100

Table 23: Frequency table by age of adolescent.

The variable gender represents question taken from the survey [23], that was filled in by adolescent' carer-takers, who indicated the gender type of their child. Below, in the Figure 26 is a representation in a bar chart. There are 540 males and 461 female and 3 people are neither female or male:

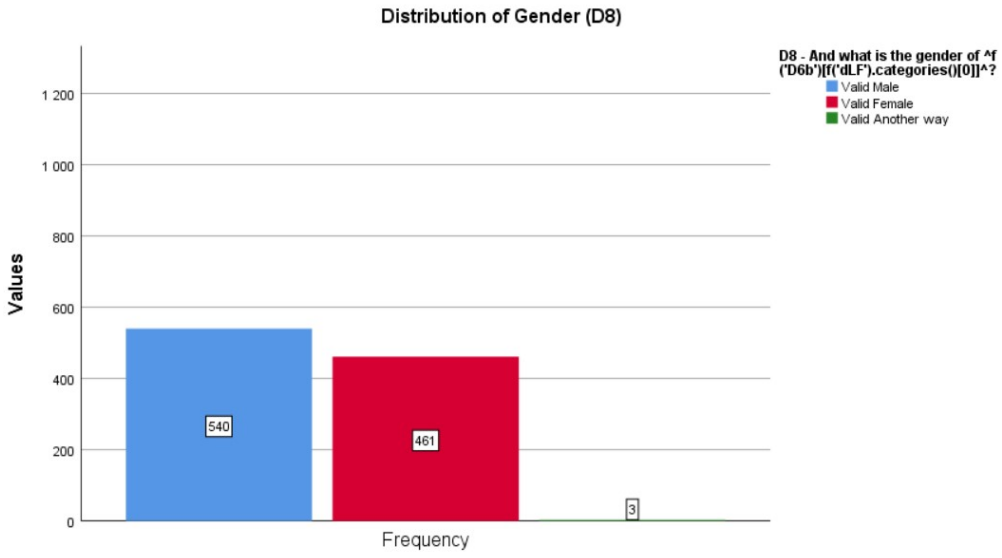

Figure 26: Bar chart of adolescent' gender indicated in survey.

It order to achieve the needed general population sample across the UK, teenagers were recruited from different regions:

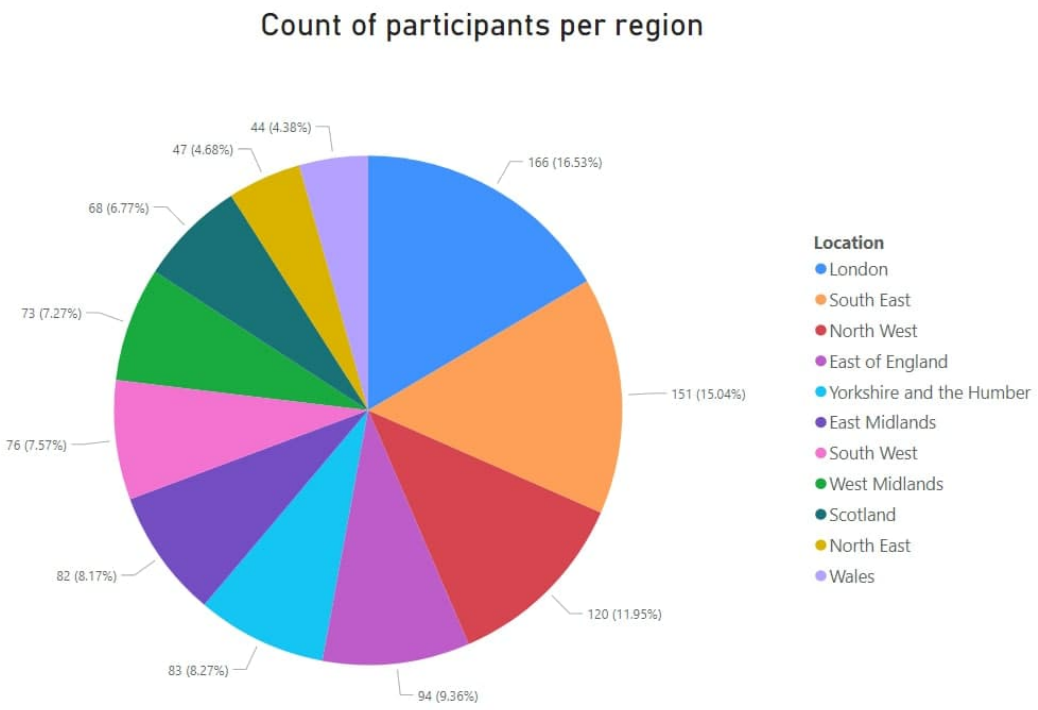

Figure 27: Pie Chart of regions of the United Kingdom where adolescents live

Adolescents were asked to indicate how much do they agree with the statement "I spend a lot of time playing video games" by using Likert scale : 1 - Strongly disagree to 5 - Strongly agree (Survey [22], Appendix E). Furthermore, we are going to see how this item is correlated with participant estimates of the time they spend playing violent games on a typical day.

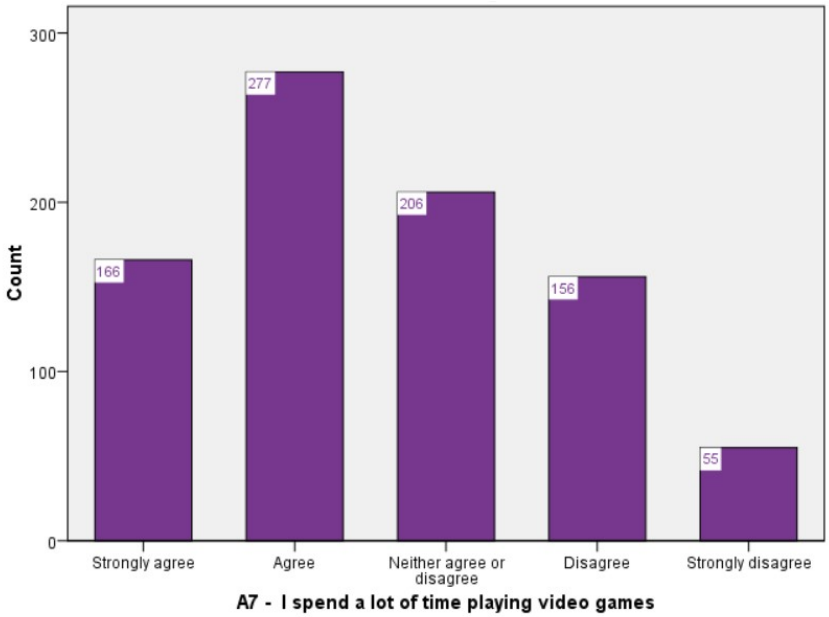

Figure 28: Bar chart of statement A7 by number of participants

We can observe from the next figure 29 that some participant's in the survey have indicated they do not play video games at all (n=144):

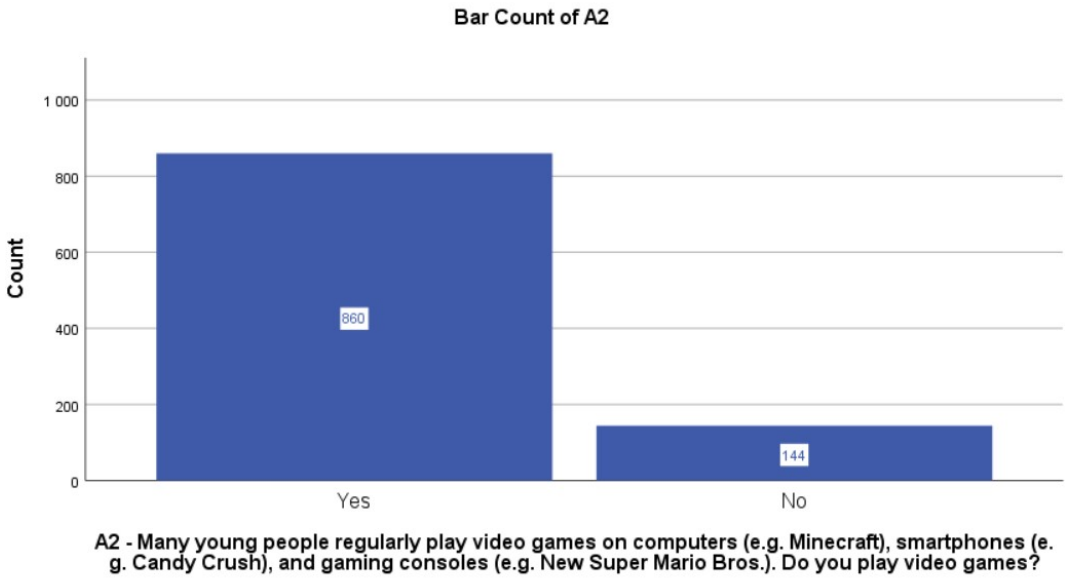

Figure 29: Bar chart of statement A2 by number of participants

The next Figure 30 shows an important variable, A6 - a question from Survey [22] (Appendix C) about how many hours adolescent's play video games per day. In the survey it has been used such categories to answer from *0 - None at all* to *7 - About 7 or more hours a day*. The following bar chart shows us the number of participants that indicated those answers:

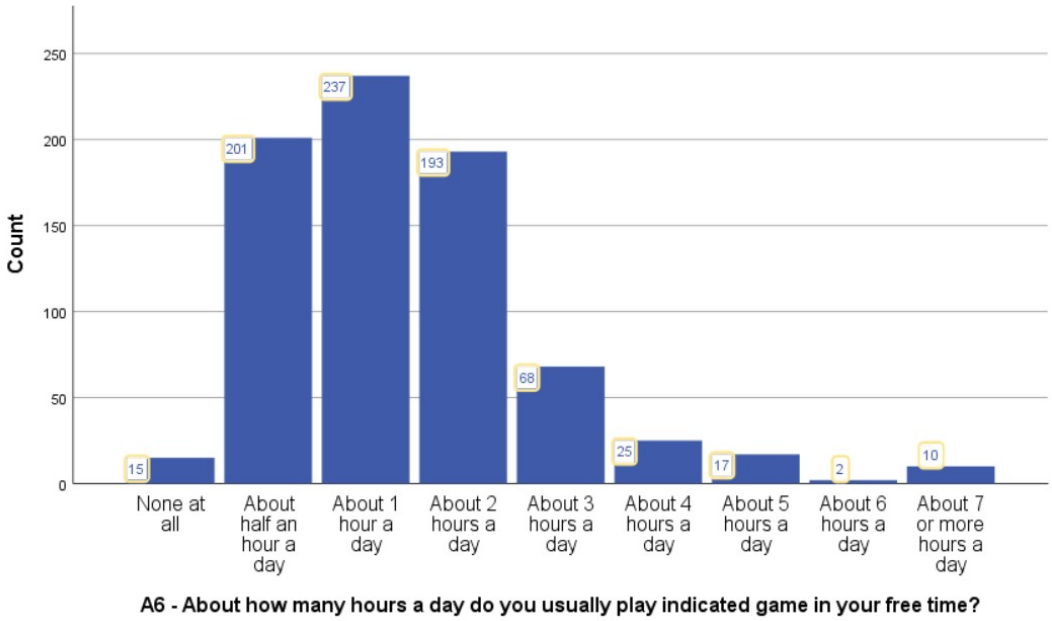

Figure 30: Bar chart of question A6

The mean of time spent on gaming per teenager per day is 3.08 hours, and the standard deviation = 2.61.

5.6 Aggressive behavior analysis

Individual differences in aggression at the trait level were assessed on the basis of self-reports by adolescents, which were derived from an abbreviated form of the Buss-Perry aggression scale [11]. Adolescents were

asked to rate 12 items (Survey [22], Appendix D) in terms of how characteristic each is of each, using a Likert 5-point scale ranging from 'very unlike me' (coded 1) to 'very like me' (coded 5). Here is a picture of Buss-Perry aggression scale categories of aggression:

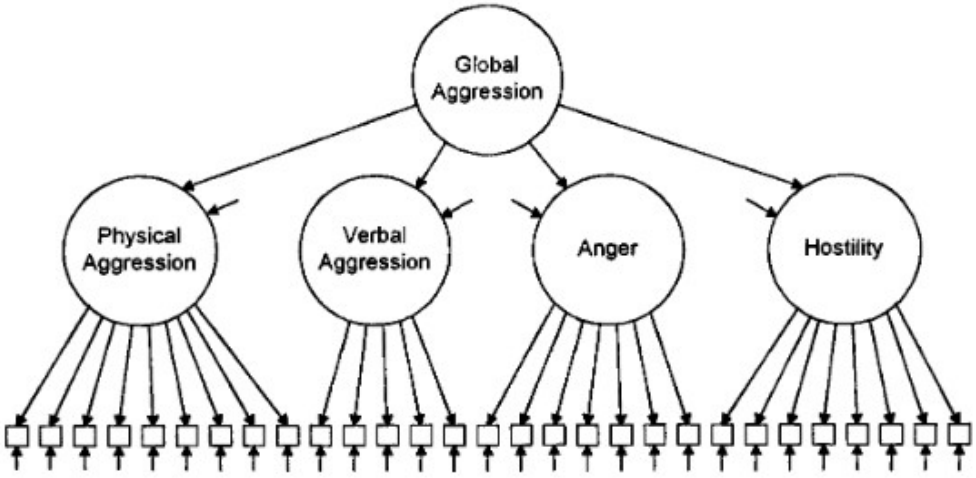

Figure 31: Categories of aggression of Buss-Perry scale

5.6.1 Trait physical

Trait physical consists of mean of three question, evaluated by 5-points Likert scale:

1. Given enough provocation, I may hit another person.
2. There are people who have pushed me so far that we have come to blows.
3. I have threatened people I know.

Item statistics:

	Mean	St.Deviation	N
Given enough provocation, I may hit another person	2.2849	1.29595	1004
There are people who have pushed me so far that have come to blows	2.3357	1.274	1004
I have threatened people I know	2.004	1.2268	1004

Table 24: Item statistics of questions.

Reliability statistics is provided below. Cronbach's alpha is a measure of internal consistency, that is, how closely related a set of items are as a group. Also, it is considered to be a measure of scale reliability. We can observe from 25 that Cronbach's Alpha (0...1) is quite high and it shows that the reliability of the data is sufficient and we can group those questions into one:

Cronbach's Alpha	N of items
0.882	3

Table 25: Reliability statistics of items

Validity statistics is also provided:

Mean	Variance	Std.Deviation	N of items
6.6245	11.676	3.41708	3

Table 26: Validity statistics of items

Finally, after computing the mean of those questions, we provide descriptive statistics of trait physical:

N	Minimum	Maximum	Mean	Std.Deviation
1004	1	5	2.2082	1.13903

Table 27: Descriptive Statistics of trait physical

5.6.2 Trait verbal

Train verbal consist of mean of three question, evaluated by 5-points Likert scale :

1. I often find myself disagreeing with people.
2. I can't help getting into arguments when people disagree with me.
3. I sometimes feel like a powder keg ready to explode.

Item statistics:

	Mean	St.Deviation	N
I often find myself disagreeing with people.	2.7759	1.20011	1004
I can't help getting into arguments when people disagree with me.	2.5169	1.2696	1004
I sometimes feel like a powder keg ready to explode.	2.3625	1.32468	1004

Table 28: Item statistics of questions for trait verbal

Reliability statistics is provided in table 29. We can observe that Cronbach's Alpha (0...1) is quite high and it shows that the reliability of the data is sufficient:

Cronbach's Alpha	N of items
0.839	3

Table 29: Reliability statistics of items

Validity statistics is also provided:

Mean	Variance	Std.Deviation	N of items
7.65554	10.904	3.30	3

Table 30: Validity statistics of items

Finally, after computing a mean of those questions, descriptive statistics of trait verbal:

N	Minimum	Maximum	Mean	Std.Deviation
1004	1	5	2.5518	1.10071

Table 31: Descriptive Statistics of trait verbal

5.6.3 Trait anger

Train anger consist of mean of three question, evaluated by 5-points Likert scale :

1. Sometimes I fly off the handle for no good reason.
2. My friends say that I'm somewhat argumentative.
3. I have trouble controlling my temper.

Item statistics:

	Mean	St.Deviation	N
Sometimes I fly off the handle for no good reason..	2.5627	1.30764	1004
My friends say that I'm somewhat argumentative.	2.2869	1.26908	1004
I have trouble controlling my temper.	2.3894	1.23699	1004

Table 32: Item statistics of questions for trait anger

Reliability statistics provided in table 33. We can observe that Cronbach's Alpha (0...1) is quite high and shows that reliability of data is sufficient :

Cronbach's Alpha	N of items
0.846	3

Table 33: Reliability statistics of items

Validity statistics is also provided:

Mean	Variance	Std.Deviation	N of items
7.2390	11.125	3.33545	3

Table 34: Validity statistics of items

Finally, after computing a mean of those questions, descriptive statistics of trait anger:

N	Minimum	Maximum	Mean	Std.Deviation
1004	1	5	2.4130	1.11182

Table 35: Descriptive Statistics of trait anger

5.6.4 Trait hostility

Train anger consists of mean of three questions, evaluated by a 5-point Likert scale :

1. At times I feel I have gotten a raw deal out of life.
2. Other people always seem to get the breaks.
3. I wonder why sometimes I feel so bitter about things.

Item statistics:

	Mean	St.Deviation	N
At times I feel I have gotten a raw deal out of life.	2.6076	1.29320	1004
Other people always seem to get the breaks.	2.7410	1.16216	1004
I wonder why sometimes I feel so bitter about things.	2.4592	1.25959	1004

Table 36: Item statistics of questions for trait hostility

Reliability statistics provided in table 37. We can observe that Cronbach's Alpha (0...1) is quite high and shows that reliability of data is sufficient:

Cronbach's Alpha	N of items
0.847	3

Table 37: Reliability statistics of items

Validity statistics is also provided:

Mean	Variance	Std.Deviation	N of items
7.8078	10.568	3.25364	3

Table 38: Validity of items

Finally, after computing the mean of those questions, descriptive statistics of trait hostility are displayed in table 39:

N	Minimum	Maximum	Mean	Std.Deviation
1004	1	5	2.6026	1.08455

Table 39: Descriptive Statistics of trait hostility

To sum up, descriptive statistics for each trait:

	N	Minimum	Maximum	Mean	Std.Deviation	Cronbach's Alpha
Trait physical	1004	1	5	2.2082	1.13903	0.882
Trait verbal	1004	1	5	2.5518	1.10071	0.839
Trait anger	1004	1	5	2.4130	1.11182	0.846
Trait hostility	1004	1	5	2.6026	1.08455	0.847

Table 40: Descriptive statistics of traits

5.7 Violent video games analysis

The analysis about how much time adolescents do spend on video games per day, and, *most importantly*, if the game contains violent content were done by 5 questions from Survey [22] (Appendix B,C). The first question was "Do you play video games?". In section Exploratory Data Analysis, in Fig. 29, we mentioned that n = 144 adolescents participating in the research do not play video games. However, n=860 answered they do play and such sample is enough for next steps. On the second question, participants were asked to write 3 different games, if there are, that they have been playing in the past month. It has been received 768, 538 and 330 titles of distinct games, respectively. Third question was about preferences in gaming.

Figure 32 shows that n = 367 teens would rather play on console among n = 538 of teenagers. The same result is on samples of first and third games titles.

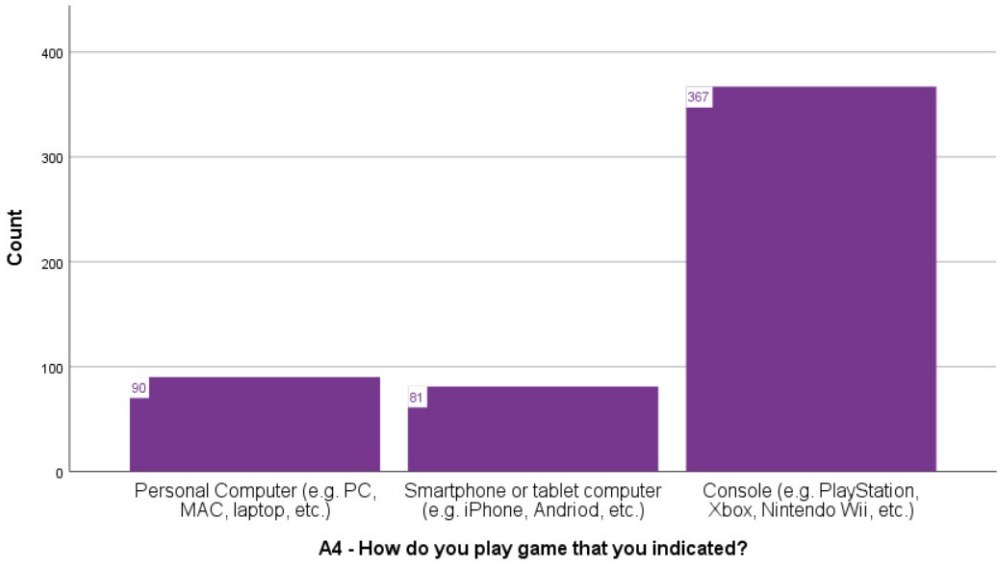

Figure 32: Preferences on device selection by adolescents

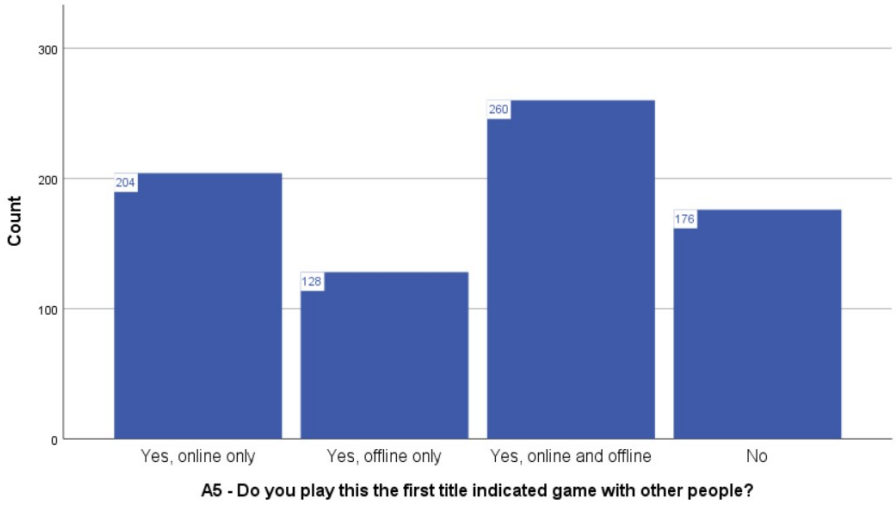

Figure 33: Preferences in playing video games with other people

We observe from Figure 33 that the highest amount n = 260 of teenagers prefer to play online and offline with other people. This fact suggests that, nowadays, teenagers are making new friends and contacts via online games.

Fifth question was used to determine how much time adolescents devoted to violent games. A bar chart showed is in Figure 30 in the section Exploratory Data Analysis. The resulted mean of the time spent on playing games by teens is **3.08 hours per day**

Afterward, variable *gaming time* (time spent on playing games) per each game was re-coded from categorical into numerical, considering each category from 1 to 7 (numbers represent certain amount of hours spent on gaming by teen). In the next steps, we computed the sum of three variables that represent time spent per each on game into variable *gaming time*. The total gaming time were computed for n = 768 adolescents.

Descriptive statistics is also presented:

	N	Minimum	Maximum	Mean	Std.Deviation
Time spent on first game (hours/day)	768	0	7	1.5553	1.25102
Time spent on second game (hours/day)	538	0	7	1.3699	1.3699
Time spent on third game (hours/day)	330	0	7	1.3699	1.12728
Gaming time (total sum, hours/day)	768	0	21	3.0755	2.6153

Table 41: Descriptive statistics of gaming time

Next step was to learn if titles of games that participants of survey are playing, contain violent content or not. For this purpose, the website of PEGI was used - Pan European Game Information is a European video game content rating system established to help European consumers make informed decisions when buying video games or apps through the use of age recommendations and content descriptors. Each game was checked for the next label of violence:

Figure 34: Label - the game contains depictions of violence.

For example, one of the games that teens of survey played is named "Call of Duty". Thus, Figure 35 displays the representative of PEGI search engine [7]:

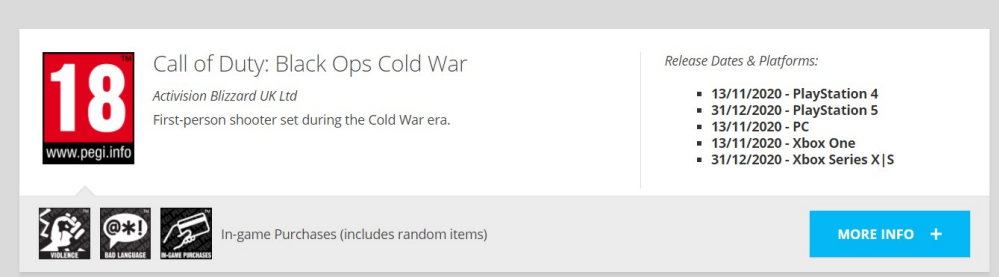

Figure 35: Call of Duty in PEGI search engine

Regarding PEGI rating, from figure 35 Call of Duty contains violent content and is recommended for people above 18. It was also important to indicate which version of the game each teen has been playing, since different versions have distinct ratings of age, violent content, etc. On such way, from PEGI search engine, information about if game contains violent label or not, binary variables *pegi.one.violence, pegi.two.variables, pegi.three.variables* were coded as 1 - contain violent content, 0 - do not contain violent content.

Furthermore, the *gaming time* spend on each game was multiplied by *pegi.violence* - binary variable, which indicates if game is containing depictions of violence. One of the purposes of our analysis was to get to know how much time adolescents spend on violent content during playing games, to correlate it with aggressive behavior.

Variable *game.violent.time* was computed as a sum of three variables, such that represent time spent on the violent game, in hours per day correspondingly. In the Table 42 descriptive statistics of mentioned variables are shown:

	N	Minimum	Maximum	Mean	Std.Deviation
game.one.violent.time	744	0	7	0.9953	1.26088
game.two.violent.time	538	0	7	0.9479	1.16138
game.three.violent.time	322	0	7	0.9643	1.17388
game.violent.time (total sum)	750	0	19	2.0687	2.30348

Table 42: Descriptive statistics of gaming time spent on violent games

5.7.1 Independent assessment using ESRB rating

The data-set has contained an assessment of games containing violent content using ESRB - The Entertainment Software Rating Board is an American self-regulatory organization that assigns age and content ratings to consumer video games. Thus, because the North American market of games is different from the European one, the data-set collectors [21] were following the same approach to calculate hours spent on violent games using as indicator variable ESRB rating, which we will use further in our analysis as well. For example, binary variables *esrb.01.violence, esrb.02.violence, esrb.03.violence* were coded by examining an ESRB search engine [8] for existence of word **"Violence"** in "Content Description" as showed in Figure 36:

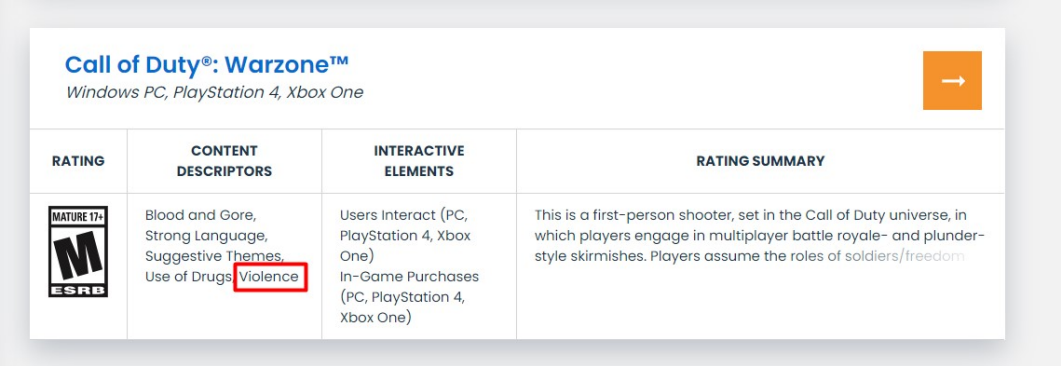

Figure 36: Label violence ESRB- the game contains depictions of violence.

A quite popular game among participants of survey is Call Of Duty, which has restrictions 12+ and contain violence descriptor.

	N	Minimum	Maximum	Mean	Std.Deviation
ESRB First game violent time hours spent	745	0	7	0.2215	0.79265
ESRB Second game violent time hours spent	527	0	7	0.7552	1.15148
ESRB Third game violent time hours spent	322	0	7	0.8804	1.17031
ESRB Violent time hours spend (total sum)	750	0	14	1.1287	1.915

Table 43: Descriptive statistics of gaming time spent on violent games by ESRB

In Table 44 we are comparing two means of resulting total time spent on violent games in hours. Using American assessment rating, resulted mean is 1.12 hours per day, while with using European rating 2.07 hours per day, spent on games. This fact means that European rating system is more strict, at least among our sample.

	N	Mean	Std.Deviation	Std. Error Mean
ESRB Violent time in hours spent (total sum)	750	1.1297	1.91651	0.00070
PEGI Violent time in hours spent (total sum)	750	2.0702	2.30469	0.00084

Table 44: Descriptive statistics of gaming time spent on violent games by ESRB rating

5.8 Confirmatory analysis

5.8.1 Gender distinction

In order to prove the second hypothesis we have applied a *T-Test* in SPSS Statistics. The independent-samples T-Test compares the means between two unrelated groups on the same continuous, dependent variable.

We want to understand if *gaming time* and *violent gaming time* differed based on *gender*. Our dependent variables are *violent gaming time*, *simple gaming time* and *trait anger*, and our independent variable is *child gender*, which has two groups: "male" and "female".

		N	Mean	Std. Deviation	Std. Error Mean
gaming time	male	453	3.36	2.64	0.12
	female	313	2.6613	2.52	0.14
violent gaming time	male	438	2.36	2.23	0.106
	female	310	1.66	2.36	0.134
trait anger	male	540	2.43	1.12	0.048
	female	461	2.39	1.10	0.051

Table 45: Group statistics provided by the T-test.

This study found that male participants had statistically significantly higher levels of overall gaming time: $t = 3.701$ $p < 0.001$ and violent overall gaming time: $t = 4.090$, $p < 0.001$, but no significant changes in trait anger, e.g. aggressive behavior: $t = 0.520$, $p = 0.603$. Based on results, we can reject the null hypothesis that the *gaming time* and *violent gaming time* means for males and females are equal.

5.8.2 Gamer and non gamer distinction

We want to understand if trait verbal, trait physical, trait anger, trait hostility differ based on *whether a person plays violent content or not*. For this, we computed a binary variable "violent game player", which were computed in a way such that if variable *violent game time* (time spent on playing games, that contain violence) > 0, then variable *violent game player* $= 1$. Group descriptive statistics provided in the Table 46:

Group Statistics

	violent_game_player	N	Mean	Std. Deviation	Std. Error Mean
trait_physical	,00	410	2,2943	1,19339	,05894
	1,00	594	2,1487	1,09702	,04501
trait_verbal	,00	410	2,5967	1,12273	,05545
	1,00	594	2,5208	1,08512	,04452
trait_anger	,00	410	2,4878	1,13116	,05586
	1,00	594	2,3614	1,09625	,04498
trait_hostility	,00	410	2,6894	1,10395	,05452
	1,00	594	2,5426	1,06776	,04381

Table 46: Group statistics provided during T-test

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
trait_physical	Equal variances assumed	9,123	,003	1,994	1002	,046	,14560	,07303	,00230	,28890
	Equal variances not assumed			1,963	830,402	,050	,14560	,07416	,00004	,29116
trait_verbal	Equal variances assumed	,447	,504	1,075	1002	,283	,07598	,07067	-,06269	,21466
	Equal variances not assumed			1,069	859,892	,286	,07598	,07111	-,06359	,21556
trait_anger	Equal variances assumed	,945	,331	1,773	1002	,077	,12641	,07131	-,01352	,26635
	Equal variances not assumed			1,763	861,474	,078	,12641	,07172	-,01436	,26718
trait_hostility	Equal variances assumed	,160	,689	2,112	1002	,035	,14678	,06952	,01037	,28319
	Equal variances not assumed			2,099	860,322	,036	,14678	,06994	,00951	,28406

Table 47: Group statistics provided during T-test

Inspection of Q-Q Plots revealed that trait verbal, trait anger and trait hostility are normally distributed for both groups, while trait physical is not, and that there is homogeneity of variance, as assessed by Levene's Test for Equality of Variances. Therefore, an independent t-test was run on the data with a 95% confidence interval (CI) for the mean difference.

From Table 47, we found that violent game players had statistically significantly higher level of *trait hostility* ($p < 0.05$, $t = 2.112$) and *trait physical* ($t = 1.994$, $p < 0.05$), but no significant changes in *trait anger* and *trait verbal*.

We can reject the null hypothesis that the *trait hostility* and *trait physical* means for people who play violent content and those who do not, are equal.

5.8.3 Correlation analysis

To test the first hypothesis, let us look at Pearson's correlation between traits computed, overall gaming time, violent time using ESRB (American) and PEGI (European) ratings. Below, we provided a correlation matrix:

Correlations

		gaming_time	trait_hostility	trait_anger	trait_verbal	trait_physical	violent_game_time_esrb	game_violent_time
gaming_time	Pearson Correlation	1	,057	,099**	,083*	,135**	,759**	,851**
	Sig. (2-tailed)		,112	,006	,022	,000	,000	,000
	N	768	768	768	768	768	750	750
trait_hostility	Pearson Correlation	,057	1	,756**	,773**	,687**	-,005	,028
	Sig. (2-tailed)	,112		,000	,000	,000	,881	,439
	N	768	1004	1004	1004	1004	750	750
trait_anger	Pearson Correlation	,099**	,756**	1	,854**	,769**	,035	,086*
	Sig. (2-tailed)	,006	,000		,000	,000	,344	,019
	N	768	1004	1004	1004	1004	750	750
trait_verbal	Pearson Correlation	,083*	,773**	,854**	1	,751**	,034	,064
	Sig. (2-tailed)	,022	,000	,000		,000	,357	,081
	N	768	1004	1004	1004	1004	750	750
trait_physical	Pearson Correlation	,135**	,687**	,769**	,751**	1	,064	,108**
	Sig. (2-tailed)	,000	,000	,000	,000		,078	,003
	N	768	1004	1004	1004	1004	750	750
violent_game_time_esrb	Pearson Correlation	,759**	-,005	,035	,034	,064	1	,799**
	Sig. (2-tailed)	,000	,881	,344	,357	,078		,000
	N	750	750	750	750	750	750	750
game_violent_time	Pearson Correlation	,851**	,028	,086*	,064	,108**	,799**	1
	Sig. (2-tailed)	,000	,439	,019	,081	,003	,000	
	N	750	750	750	750	750	750	750

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Figure 37: Correlation matrix variables of interest

We can observe weak positive correlation coefficient, but a statistically significant ones, between ***gaming time*** and *trait physical* ($p < 0.01$, $r = 0.135$), *trait anger*($p = 0.006 < 0.05$, $r = 0.099$), *trait verbal* ($p < 0.05$, $r = 0.083$).

Between ***gaming time*** and *trait hostility* there is a weak correlation coefficient: $r = 0.057$ and not statistically significant $p = 0.112$. Correlation between *trait verbal* and *gaming time* is statistically significant, but still weak : $r = 0.083$, $p = 0.02 < 0.05$.

Curiously enough, there are no statistically significant correlations and strong coefficient between traits and violent game time computed by ESRB (North America rating) indicators.

On the other hand, there are weak but statistically significant correlation coefficients between ***game violent time*** by PEGI (European rating) and *trait anger*: $r = 0.086$, $p < 0.05$ and *trait physical*: $r = 0.108$, $p < 0.01$.

Thus, we can admit that European rating system is distinct from American one, as we said in chapter Violent video games analysis. In addition, European rating is more strict in relation to violent content.

5.8.4 Step-wise regression of the gaming time

To test the first hypothesis, we want to see if an increase of violent gaming hours depends on traits that we have computed, e.g. level of aggression. Dependent variable in multiple linear regression is *game violent time* and independent variables are *trait verbal*, *trait anger*, *trait physical* and *trait hostility*. Stepping Method criteria are 0.05 entry and 0.10 removal.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,128[a]	,016	,011	2,29073	,016	3,090	4	745	,015

a. Predictors: (Constant), trait_anger, trait_hostility, , trait_verbal, trait_physical

b. Dependent Variable: game_violent_time

Table 48: Model Summary

We can see that R = 0.128, which is multiple correlation coefficient, that means we have poor level of prediction. $R^2 = 0.011$ indicates that our independent variables explain 1.1% of the variability of our dependent variable.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	64,868	4	16,217	3,090	,015[b]
	Residual	3909,346	745	5,247		
	Total	3974,214	749			

a. Dependent Variable: game_violent_time

b. Predictors: (Constant), trait_anger, trait_hostility, , trait_verbal, trait_physical

Table 49: ANOVA statistics

The F-ratio in the Table 49 tests whether the overall regression model is a good fit for the data. The table shows that the independent variables statistically significantly predict the dependent variable, F(4,745) = 3.090, p > 0005 (i.e., the regression model is a poor fit of the data).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	1,740	,229		7,607	,000			
	trait_physical	,257	,116	,123	2,215	,027	,108	,081	,081
	trait_verbal	-,064	,160	-,030	-,400	,689	,064	-,015	-,015
	trait_hostility	-,195	,125	-,090	-1,561	,119	,028	-,057	-,057
	trait_anger	,184	,157	,087	1,174	,241	,086	,043	,043

a. Dependent Variable: game_violent_time

Table 50: Coefficients of regression

Unstandardized coefficients from Table 50 indicate how much the dependent variable varies with an independent variable when all other independent variables are held constant.

We can clearly see that an increase of *trait physical*, affect on an increase in hours spent on violent games. By looking at the p < 0.05 we can reject the null hypothesis that "there are no meaningful relationships between physical aggression of teen and usage of violent video games", t = 2.215.

At the same time, we can conclude that verbal, hostility and anger level of aggression are not statistically significant for predicting variable *game violent time*.

5.8.5 Regression of the level of aggression

To test the first hypothesis, that mentioned in Introduction (page 36), we have done one more multiple linear regression and quadratic one. Variables *prosocial* and *conduct problems* were computed from the next ITEMs - questions to the carer-taker of adolescent, described in the survey [23] and scaling were produced by authors of the data-set [21] by using Scoring the Strengths and Difficulties Questionnaire for age 4-17 or 18+ [23].

Variable *conduct problems* is composed by next ITEMs (questions to carer-takers):
ITEM 5 - Often has temper tantrums or hot tempers;
ITEM 7 - Generally obedient;
ITEM 12 - Often fights with other children;
ITEM 18 - Often lies or cheats;
ITEM 22 - Steals from home, school or elsewhere;
The scale for those questions is the Likert scale with scaling from 1 - Not true, 2 - Somewhat true to 3 - Certainly true.

Roughly speaking, we intend to predict if child can have problems of aggressive behavior to people around him and to himself. Our dependent variable is *conduct problems*, while the predictors are: *trait verbal*, *trait hostility*, *trait physical*, *trait anger*, *gaming time*, *game violent time* by PEGI (European rating), *game violent time* by ESRB (American rating), *game violent time squared* by ESRB.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,679[a]	,461	,455	1,45071	,461	79,107	8	741	,000

a. Predictors: (Constant), gaming_time, trait_hostility, , game_violent_time_square_esrb, trait_anger, game_violent_time, trait_verbal, violent_game_time_esrb, trait_physical

Table 51: Model Summary of regression

Table 51, R = 0.679 indicates a good level of prediction. R Square has a value of 0.461, that means that our independent variables explain 46.1% of the variability of our dependent variable.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1331,886	8	166,486	79,107	,000[b]
	Residual	1559,481	741	2,105		
	Total	2891,367	749			

a. Dependent Variable: conduct_problems

b. Predictors: (Constant), gaming_time, trait_hostility, , game_violent_time_square_esrb, trait_anger, game_violent_time, trait_verbal, violent_game_time_esrb, trait_physical

Table 52: ANOVA statistics

From ANOVA test, Table 52 we can observe that the independent variables statistically significantly predict the dependent variable, $F(8, 741) = 79.107$, $p < .0005$ (i.e., the regression model is a good fit of the data).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	3,613	,163		22,209	,000			
	game_violent_time	-,024	,059	-,028	-,404	,687	,078	-,015	-,011
	game_violent_time_square	,003	,005	,033	,567	,571	,075	,021	,015
	gaming_time	,041	,041	,054	1,001	,317	,111	,037	,027
	violent_game_time_esrb	-,045	,049	-,044	-,920	,358	,034	-,034	-,025
	trait_verbal	,196	,102	,108	1,919	,055	,597	,070	,052
	trait_hostility	,060	,080	,033	,758	,449	,520	,028	,020
	trait_anger	,478	,100	,265	4,784	,000	,626	,173	,129
	trait_physical	,586	,074	,329	7,906	,000	,627	,279	,214

a. Dependent Variable: conduct_problems

Table 53: Coefficients of regression

From Table 53 of Coefficients of regression, we can observe significance of *trait anger* and *trait physical* and high unstandardized coefficients B, which indicates that the dependent variable is highly correlated with computed variable *conduct problems* that indicates aggression of teen.

From the Table 53 we can observe that *game violent time* by PEGI (European rating), neither *game violent time squared* by PEGI, neither violent game time by ESRB (American rating) and violent game time squared by ESRB, even *gaming time* were not significant predictors for regression on dependent variable *conduct problems*.

Conclusions

Speaking of the categorical values, performing a two way ANOVA and a one way MANOVA revealed a certain dependency of our dependent variables (number of reviews, mean number of hours, prices, ratings) to the different classifications of modern video games (genres, way of playing, etc.). This result, somehow expected, is more interesting if one notices that factors like the number of hours a person spends on a video game seems to be a priority, regardless of the game quality itself.

Regarding the analysis of the quantitative variables, the first statistical learning method we applied is Multiple Linear Regression. We learned that the most useful variables in predicting the average user's playtime are the number of positive and negative reviews, as well as the number of critic reviews. However, we found that the linear regression model does not perform well on our data, leading to low prediction accuracy. Hoping to achieve better results, we moved on to classification methods. We applied Linear Discriminant Analysis, Quadratic Discriminant Analysis, and K-Nearest Neighbors, finding that QDA outperforms both LDA and KNN. We finally moved on to non-linear models, applying Regression Trees. We found that in this case the analysis leads to a low level of accuracy when predicting *meanHours*; however, we obtain some interesting outcome with the use of this last technique: against every expectation, a high number of *negativeReviews*, the root of our tree and the most important parameter, leads to a high number of *meanHours*, showing that the community of gamers is not totally consistent when rating games that have been played for a lot of hours, similarly to what we have observed in the genre analysis.

In the time series analysis that we conducted, we explored the trends in the global sales of video games over the years. We started by performing a stationarity test, which produced evidence of stationarity in our data. We then moved on to analyzing monthly trends, detecting no seasonality in our data. Lastly, we performed a Chow Test, obtaining no decisive evidence of any structural break in our data. In conclusion, if we were game developers, we would not choose a precise month to launch our video-game, we should not wait for a golden age and we could not follow any kind of trend.

In the behavior analysis part, we have explored how level of four types of aggression can change on the basis on hours spent on violent video games. Additionally, we have concluded based on T-Test, that gender differs in preferences to spend more or less time on video games. Males were more willing to spend time on video gaming than females. Despite this, we have found that an increase in physical aggression is related to an increase in gaming time based on multiple linear regression results. Furthermore, based on T-Test we found out that the means of players, who do play violent content are significantly different on the level of hostility and physical aggression. Finally, we have computed multiple linear regression with quadratic term to predict level of aggression of adolescents based on time spent on violent games and we have discovered that this predictor was not a significant one.

In conclusion, in our work we conducted a multi-faceted analysis of video game data. The results obtained throughout the analysis helped us gain a deeper understanding of the field, by highlighting some counter-intuitive aspects of the gaming community.

References

[1] Lenhart A.Teens, Technology and Friendships.(2015). [Internet].
https://www.pewresearch.org/internet/2015/08/06/teens-technology-and-friendships/

[2] Copenhaver, A. (2015). Violent video game legislation as pseudo-agenda. Criminal Justice Studies: A Critical Journal of Crime, Law & Society, 28(2), 170–185.

[3] Brown v. Entertainment Merchants Ass'n. 131 S.Ct. 2729 (2011)

[4] Bandura A. 1977 Self-efficacy: toward a unifying theory of behavioral change. Psychol. Rev. 84, 191-215.

[5] Przybylski AK, Ryan RM, Rigby CS. (2009). The motivating role of violence in video games.

[6] Goodman R. 1997. The strengths and difficulties questionnaire.

[7] Pan European Game Information. PEGI Database [Internet]. http://www.pegi.info/en/index/id/509

[8] Entertainment Software Rating Board. ESRB Database [Internet].http://www.esrb.org/.

[9] Goodman R, Ford T, Simmons H, Gatward R, Meltzer H.Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. Br. J. Psychiatry J. Ment. Sci. 177, 534-539.

[10] Ipsos MORI. 2016What About YOUth? Survey, 2014.

[11] Bryant FB, Smith BD. 2001. Refining the architecture of aggression: a measurement model for the Buss–Perry aggression questionnaire. J. Res. Personal. 35, 138-167.

[12] David Berk, Brian Hirt, Jim Leonard - Moby games. [Internet]. https://www.mobygames.com/

[13] Valve Corporation, Steam Online Store.[Internet]. https://store.steampowered.com.

[14] Valve Corporation, Steam Api.[Internet]. https://partner.steamgames.com/doc/webapi_overview

[15] Sergey Galyonkin, SteamSpy. [Internet]. https://steamspy.com/api.php

[16] Jason Dietz, Marc Doyle, Julie Doyle Roberts - Metacritic. [Internet]. https://www.metacritic.com

[17] Kenneth Reitz, Requests 2.25.1 Library for Python. [Internet]. https://requests.readthedocs.io/en/master

[18] Leonard Richardson, BeautifulSoup 4.9.0 Library for Python. [Internet]. https://www.crummy.com/software/BeautifulSoup/bs4/doc

[19] Brett Walton, VGChartz.[Internet]. https://www.vgchartz.com

[20] Peter Dalgaard, Introductory Statistics With R, Springer. (2008)

[21] Andrew K. Przybylski and Netta Weinstein. Violent video game engagement is not associated with adolescents' aggressive behaviour: evidence from a registered report. [Internet] https://royalsocietypublishing.org/doi/10.1098/rsos.171474

[22] Survey for adolescent's used for part "Investigating the association". [Attached material].

[23] Scoring the Strengths and Difficulties Questionnaire for age 4-17 or 18+. [Attached material]

[24] G.James, D. Witten, T.Hastie, R.Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer. 2013.