IPS Instituto Politécnico de Setúbal
Escola Superior de Tecnologia de Setúbal

Trabalho para a Unidade Curricular Métodos Estatísticos
Licenciatura em Engenharia Informática

2° ano, 2° semestre, ano letivo 2021-2022

# *Análise Descritiva do Mercado de Venda de Videojogos*

Diogo Letras - Turma SW 03, Nº 202002529

Miguel Vicente - Turma SW 03, Nº 202000563

Pedro Cunha - Turma SW 03, Nº 202000757

**Resumo:** O trabalho consiste numa análise descritiva do mercado de venda de videojogos a nível mundial nas suas mais variadas vertentes. Foram analisadas várias variáveis, tal como plataforma, ano, género (tipo de jogo), editora, nº de vendas na América do Norte, nº de vendas na União Europeia e nº de vendas a nível global. Para cada variável, foi feita uma tabela de frequências (composta por frequências absolutas, frequências relativas, frequências absolutas acumuladas e frequências relativas acumuladas), de modo a organizar os dados e tirar conclusões. Para além disso, foram também construídos vários gráficos (tais como gráficos de barras, gráficos circulares e histogramas), para representar a informação. Por fim, foram calculadas as várias medidas de localização (nomeadamente a média, a moda, a mediana e os quartis) e também as várias medidas de dispersão (tais como amplitude, amplitude interquartil, variância, desvio padrão), de modo a aprofundar a análise do tema e conseguir tirar conclusões robustas sobre o tema analisado.

**Palavras-chave**: Estatística, Videojogos, Análise, População, Variáveis.

## 1. Introdução

Conforme requisitado pelo enunciado do 1º Trabalho de Grupo, foi feita uma análise detalhada, dentro da área de Estatística Descritiva, de uma base de dados fornecida pela docente responsável. A base de dados em questão é composta por 16598 videojogos, que basicamente são todos os videojogos com pelo menos 100 mil cópias vendidas entre, aproximadamente, o ano de 1980 e o ano de 2017. Ou seja, resumidamente temos como população todos os videojogos vendidos e como amostra 16598 videojogos, que são títulos com pelo menos 100 mil cópias vendidas, durante um período de tempo. O conjunto de variáveis estudado encontra-se descrito no ponto que se encontra a seguir neste relatório.

## 2. Descrição das variáveis

Foram analisadas 7 variáveis no trabalho. De seguida, apresenta-se uma breve descrição de cada variável.

- Platform - Representa o hardware no qual jogo corre (ex: ps5, xbox one, pc);
- Year - Representa o ano em que o jogo foi lançado;
- Genre - Representa o tipo de jogo (ex: ação, desporto, corridas);
- Publisher - Representa a empresa que lançou o jogo (ex: EA, Capcom, Rockstar);
- NaSales - Representa o número de vendas na América do Norte;
- EuSales - Representa o número de vendas na União Europeia;
- GlobalSales - Representa o número de vendas a nível mundial.

Estas variáveis podem ser agrupadas, de acordo com a sua classificação. Deste modo, temos:

- Variáveis Qualitativas Nominais - Platform, Genre e Publisher;
- Variáveis Quantitativas Discretas - Year;
- Variáveis Quantitativas Contínuas - NaSales, EuSales e GlobalSales.

De seguida, iremos detalhar a análise de cada variável, seguindo a ordem acima descrita, portanto, começando nas qualitativas nominais, depois passando pelas quantitativas discretas e finalmente acabando nas quantitativas contínuas.

## 3. Variáveis qualitativas nominais

Foi feita uma análise das mesmas, sendo importante mencionar, que não foi possível aplicar todas as medidas de localização, devido a não serem aplicáveis em variáveis qualitativas. A única medida de localização que poderíamos aplicar e que eventualmente aplicámos, foi a moda.

Optámos também pela utilização de gráficos de barras e pelo diagrama circular para a representação gráfica destas variáveis, por ser a mais frequentemente utilizada e recomendada para visualização de dados que se encaixam neste tipo de variáveis.

### 3.1. Platform - Plataforma

Esta variável, que tal como referido, representa o tipo de hardware, sendo que as possíveis respostas não contêm qualquer ordem subentendida.

Como referido acima, não existiu aplicação de medidas de localização, como por exemplo, a média e quantis, devido a não serem aplicadas a variáveis qualitativas nominais.

Entretanto foi aplicada a moda nesta variável e foi determinado que no universo de 16598 videojogos, a moda foi "DS", com uma frequência absoluta de 2163. Isto pode ser visualizado na tabela de frequências da variável que se encontra presente na Tabela 1 – Frequências da variável 'Platform'.

**Tabela 1** – Frequências da variável 'Platform'

| i | xi | ni | fi | Ni | Fi |
|---|---|---|---|---|---|
| 1 | 2600 | 133 | 0.00801 | 133 | 0.01 |
| 2 | 3DO | 3 | 0.00018 | 136 | 0.01 |
| 3 | 3DS | 509 | 0.03067 | 645 | 0.04 |
| 4 | DC | 52 | 0.00313 | 697 | 0.04 |
| 5 | DS | 2163 | 0.13032 | 2860 | 0.17 |
| 6 | GB | 98 | 0.0059 | 2958 | 0.18 |
| 7 | GBA | 822 | 0.04952 | 3780 | 0.23 |
| 8 | GC | 556 | 0.0335 | 4336 | 0.26 |
| 9 | GEN | 27 | 0.00163 | 4363 | 0.26 |
| 10 | GG | 1 | 0,000060 | 4364 | 0.26 |
| 11 | N64 | 319 | 0.01922 | 4683 | 0.28 |
| 12 | NES | 98 | 0.0059 | 4781 | 0.29 |
| 13 | NG | 12 | 0.00072 | 4793 | 0.29 |
| 14 | PC | 960 | 0.05784 | 5753 | 0.35 |
| 15 | PCFX | 1 | 0,000060 | 5754 | 0.35 |
| 16 | PS | 1196 | 0.07206 | 6950 | 0.42 |
| 17 | PS2 | 2161 | 0.1302 | 9111 | 0.55 |
| 18 | PS3 | 1329 | 0.08007 | 10440 | 0.63 |
| 19 | PS4 | 336 | 0.02024 | 10776 | 0.65 |
| 20 | PSP | 1213 | 0.07308 | 11989 | 0.72 |
| 21 | PSV | 413 | 0.02488 | 12402 | 0.75 |
| 22 | SAT | 173 | 0.01042 | 12575 | 0.76 |
| 23 | SCD | 6 | 0.00036 | 12581 | 0.76 |
| 24 | SNES | 239 | 0.0144 | 12820 | 0.77 |
| 25 | TG16 | 2 | 0.00012 | 12822 | 0.77 |
| 26 | Wii | 1325 | 0.07983 | 14147 | 0.85 |
| 27 | WiiU | 143 | 0.00862 | 14290 | 0.86 |
| 28 | WS | 6 | 0.00036 | 14296 | 0.86 |
| 29 | X360 | 1265 | 0.07621 | 15561 | 0.94 |
| 30 | XB | 824 | 0.04964 | 16385 | 0.99 |
| 31 | XOne | 213 | 0.01283 | 16598 | 1 |
| | | 16598 | 1 | | |

Tal como foi referido anteriormente, existe uma limitação à representação gráfica dos dados, devido a este ser uma variável qualitativa nominal, portanto foram gerados os gráficos presentes no Gráfico 1 – Gráfico de Barras das frequências absolutas da variável 'Platform' e na Gráfico 2 – Gráfico Circular das frequências relativas da variável 'Platform'.

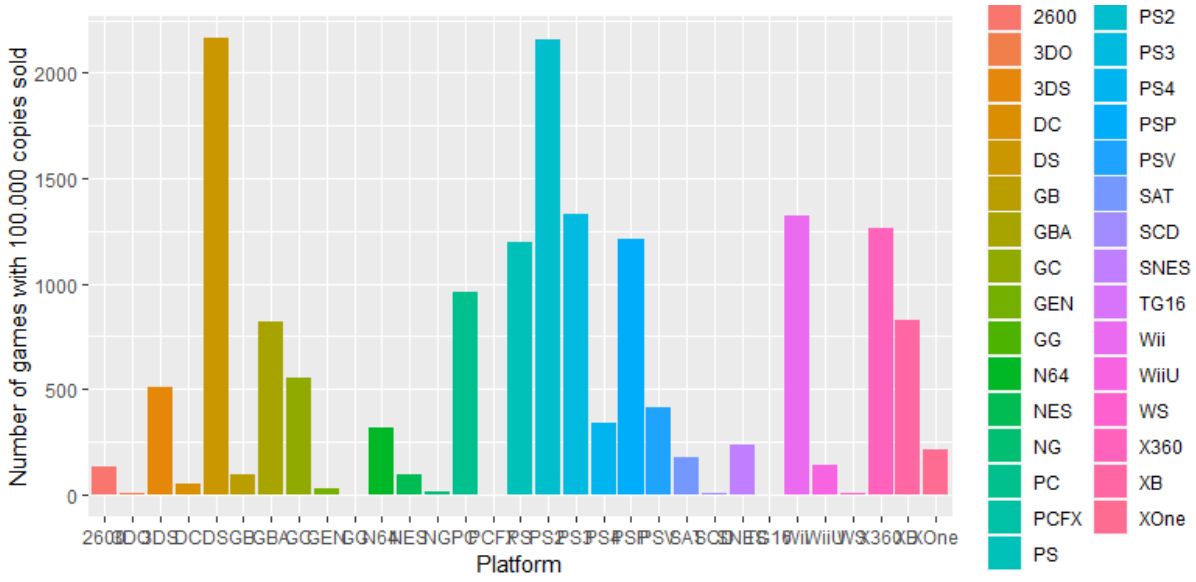


**Gráfico 1** – Gráfico de Barras das frequências absolutas da variável 'Platform'

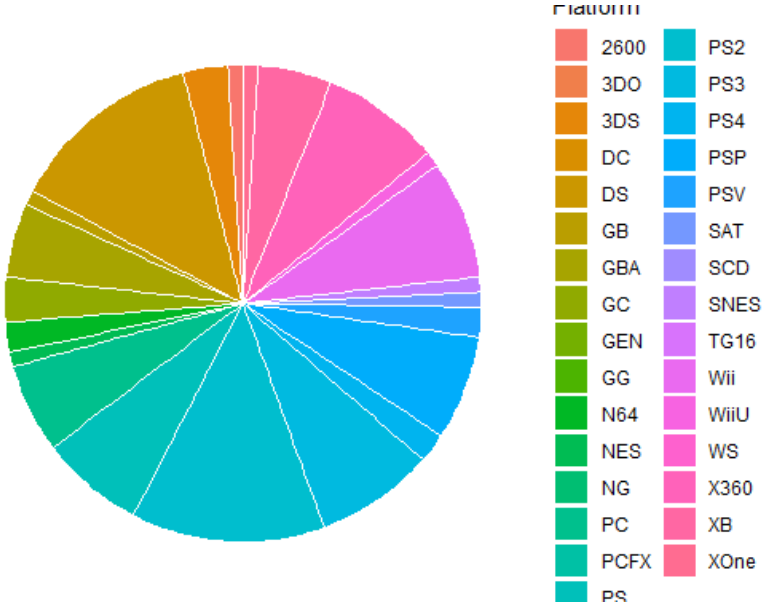


**Gráfico 2** – Gráfico Circular das frequências relativas da variável 'Platform'

### 3.2. Genre - Gênero (tipo de jogo)

Também esta variável, sendo do tipo qualitativa, fez com que não fosse possível calcular certas medidas de localização.

A moda foi "Action", com uma frequência absoluta de 3316. Isto pode ser visualizado na tabela de frequências da variável que se encontra presente na Tabela 2 – Frequências da variável 'Genre'.

**Tabela 2** – Frequências da variável 'Genre'

| i | xi | ni | fi | Ni | Fi |
|---|---|---|---|---|---|
| 1 | Action | 3316 | 0.2 | 3316 | 0.2 |
| 2 | Adventure | 1286 | 0.08 | 4602 | 0.28 |
| 3 | Fighting | 848 | 0.05 | 5450 | 0.33 |
| 4 | Misc | 1739 | 0.1 | 7189 | 0.43 |
| 5 | Platform | 886 | 0.05 | 8075 | 0.49 |
| 6 | Puzzle | 582 | 0.04 | 8657 | 0.52 |
| 7 | Racing | 1249 | 0.08 | 9906 | 0.6 |
| 8 | Role-Playing | 1488 | 0.09 | 11394 | 0.69 |
| 9 | Shooter | 1310 | 0.08 | 12704 | 0.77 |
| 10 | Simulation | 867 | 0.05 | 13571 | 0.82 |
| 11 | Sports | 2346 | 0.14 | 15917 | 0.96 |
| 12 | Strategy | 681 | 0.04 | 16598 | 1 |

Tal como foi referido anteriormente, existe uma limitação à representação gráfica dos dados, devido a este ser uma variável qualitativa nominal, portanto foram gerados os gráficos presentes no Gráfico 3 – Gráfico de

Barras das frequências absolutas da variável 'Genre' e na Gráfico 4 – Gráfico Circular das frequências relativas da variável 'Genre'.

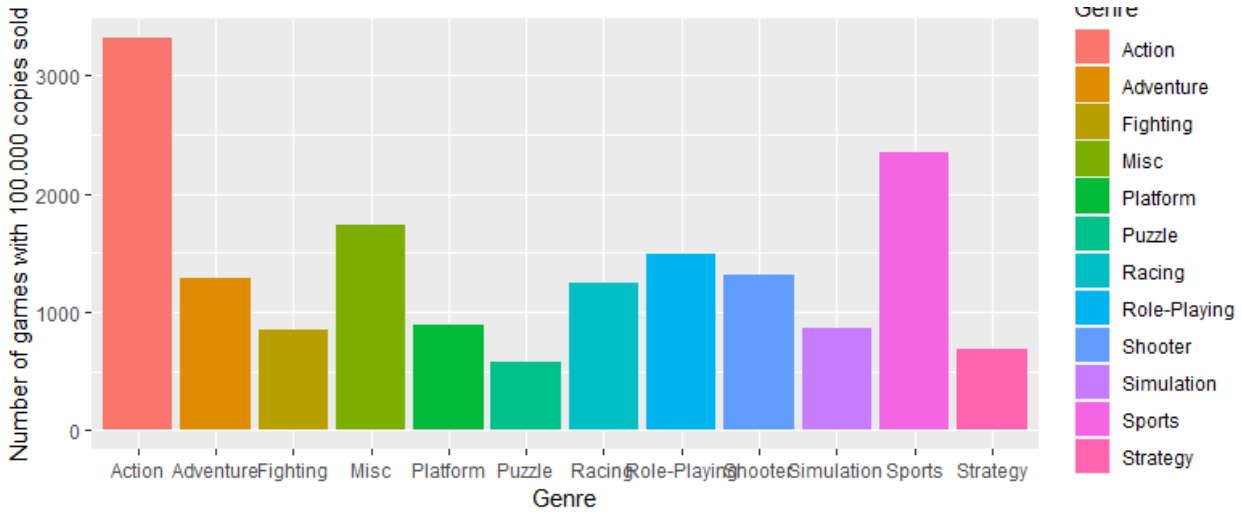


**Gráfico 3** – Gráfico de Barras das frequências absolutas da variável 'Genre'

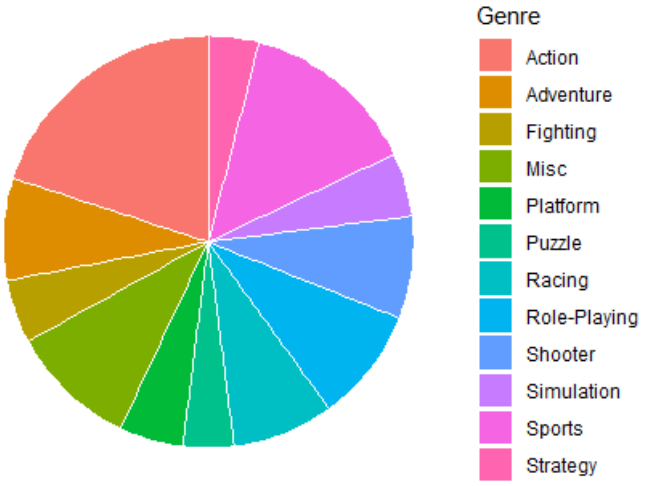


**Gráfico 4** – Gráfico Circular das frequências relativas da variável 'Genre'

### 3.3. Publisher - Editora

Novamente, a variável sendo do tipo qualitativa, fez com que não fosse possível calcular certas medidas de localização. Como o nº de 'Publishers' é bastante grande (existem cerca de 579), de seguida apresenta-se um 'top 30' ordenado pelas frequências absolutas ('ni'), de modo a resumir a tabela com os dados mais relevantes.

A moda foi "Electronic Arts", com uma frequência absoluta de 1351. Isto pode ser visualizado na tabela de frequências da variável que se encontra presente na Tabela 3 – Frequências da variável 'Publisher'.

**Tabela 3** – Frequências da variável 'Publisher'

| i | xi | ni | fi | Ni | Fi |
|---|---|---|---|---|---|
| 139 | Electronic Arts | 1351 | 0.0814 | 5528 | 0.33305 |
| 17 | Activision | 975 | 0.05874 | 1507 | 0.09079 |
| 352 | Namco Bandai Games | 932 | 0.05615 | 9624 | 0.57983 |
| 533 | Ubisoft | 921 | 0.05549 | 15619 | 0.94102 |
| 282 | Konami Digital Entertainment | 832 | 0.05013 | 7647 | 0.46072 |
| 515 | THQ | 715 | 0.04308 | 14613 | 0.88041 |
| 369 | Nintendo | 703 | 0.04235 | 10396 | 0.62634 |
| 465 | Sony Computer Entertainment | 683 | 0.04115 | 12479 | 0.75184 |
| 453 | Sega | 639 | 0.0385 | 11732 | 0.70683 |
| 499 | Take-Two Interactive | 413 | 0.02488 | 13467 | 0.81136 |
| 82 | Capcom | 381 | 0.02295 | 2917 | 0.17574 |
| 54 | Atari | 363 | 0.02187 | 2124 | 0.12797 |
| 507 | Tecmo Koei | 338 | 0.02036 | 13851 | 0.8345 |
| 475 | Square Enix | 233 | 0.01404 | 12821 | 0.77244 |
| 560 | Warner Bros. Interactive Entertainment | 232 | 0.01398 | 16386 | 0.98723 |
| 126 | Disney Interactive Studios | 218 | 0.01313 | 3895 | 0.23467 |
| 541 | Unknown | 203 | 0.01223 | 15879 | 0.95668 |
| 138 | Eidos Interactive | 198 | 0.01193 | 4177 | 0.25166 |
| 330 | Midway Games | 198 | 0.01193 | 8503 | 0.51229 |
| 7 | 505 Games | 192 | 0.01157 | 241 | 0.01452 |
| 328 | Microsoft Game Studios | 189 | 0.01139 | 8281 | 0.49892 |
| 13 | Acclaim Entertainment | 184 | 0.01109 | 506 | 0.03049 |
| 111 | D3Publisher | 184 | 0.01109 | 3472 | 0.20918 |
| 556 | Vivendi Games | 164 | 0.00988 | 16147 | 0.97283 |
| 91 | Codemasters | 152 | 0.00916 | 3128 | 0.18846 |
| 233 | Idea Factory | 129 | 0.00777 | 6377 | 0.3842 |
| 119 | Deep Silver | 122 | 0.00735 | 3613 | 0.21768 |
| 372 | Nippon Ichi Software | 105 | 0.00633 | 10509 | 0.63315 |
| 577 | Zoo Digital Publishing | 104 | 0.00627 | 16547 | 0.99693 |

Devido ao nº de 'Publishers' ser bastante grande foi feita uma ordenação e seleção, de modo a produzir o gráfico de barras das frequências absolutas (basicamente fez-se um Top 15).

Para o gráfico circular, foi feito exatamente o mesmo processo, porém aplicado às frequências relativas.

Portanto foram gerados os gráficos presentes no Gráfico 5 – Gráfico de Barras das frequências absolutas da variável 'Publisher' e no Gráfico 6 – Diagrama Circular das frequências relativas da variável 'Publisher'.

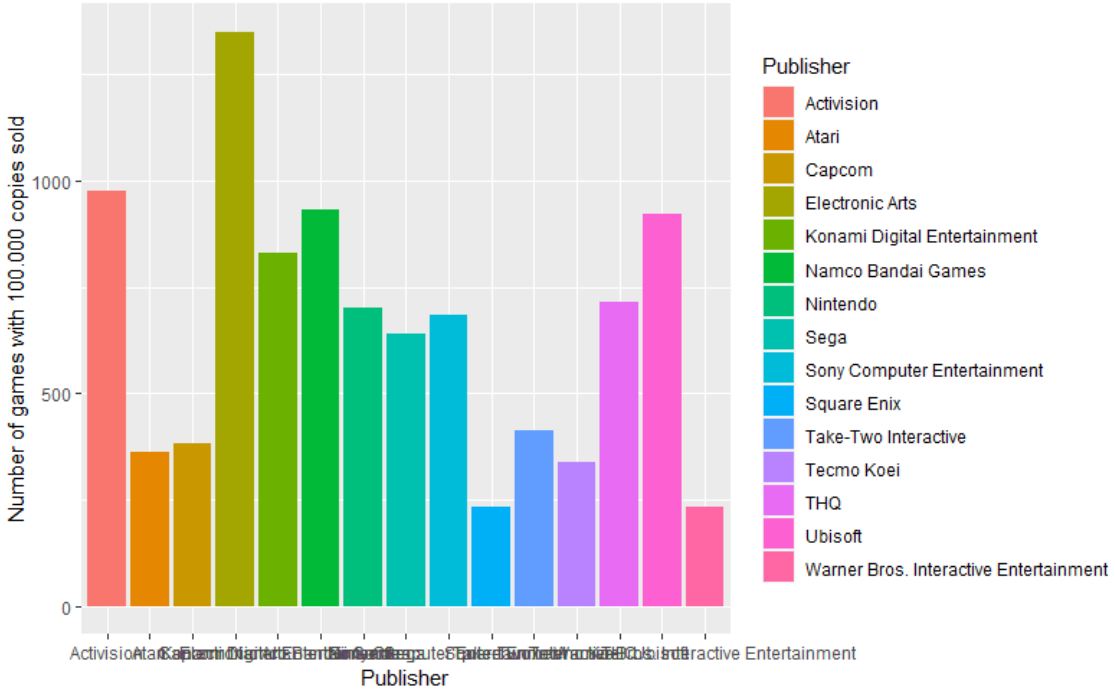


**Gráfico 5** – Gráfico de Barras das frequências absolutas da variável 'Publisher'

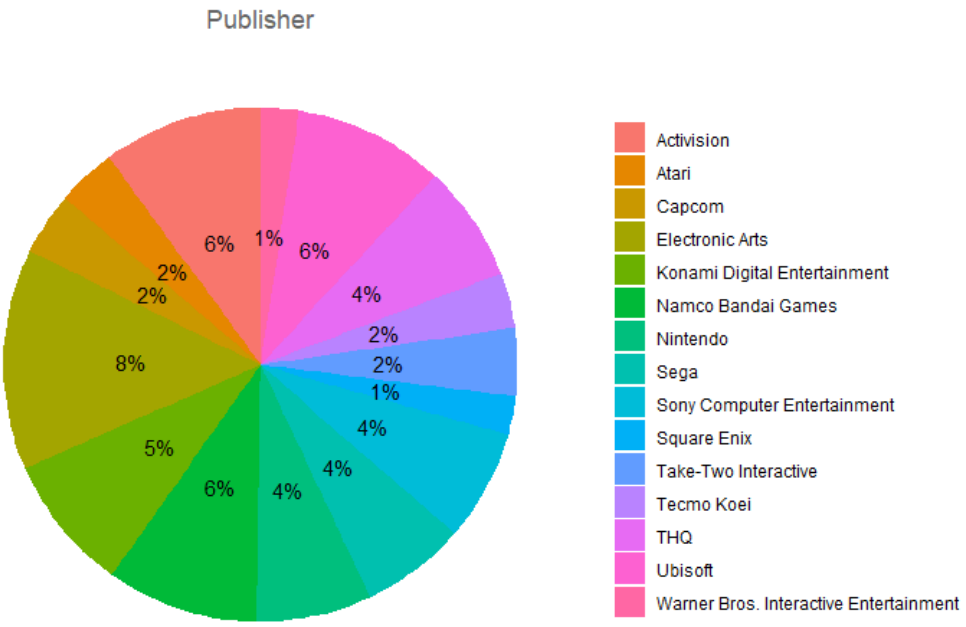


**Gráfico 6** – Gráfico Circular das frequências relativas da variável 'Publisher'

## 4. Variáveis quantitativas discretas

Foi feita uma análise das mesmas. Agora foi possível calcular as várias medidas de localização e de dispersão por serem variáveis do tipo acima mencionado.

Optámos novamente pela utilização de gráficos de barras e pelo diagrama circular para a representação gráfica destas variáveis.

### 4.1. Year - Ano

Esta variável sendo do tipo quantitativa, fez com que fosse possível calcular medidas de localização e de dispersão. A moda foi "2009", com uma frequência absoluta de 1431. Isto pode ser visualizado na tabela de frequências da variável que se encontra presente na Tabela 4 – Frequências da variável 'Year'.

**Tabela 4** – Frequências da variável 'Year'

| i | xi | ni | fi | Ni | Fi |
|---|---|---|---|---|---|
| 1 | 1980 | 9 | 0.000542 | 9 | 0.000542 |
| 2 | 1981 | 46 | 0.002771 | 55 | 0.003314 |
| 3 | 1982 | 36 | 0.002169 | 91 | 0.005483 |
| 4 | 1983 | 17 | 0.001024 | 108 | 0.006507 |
| 5 | 1984 | 14 | 0.000843 | 122 | 0.00735 |
| 6 | 1985 | 14 | 0.000843 | 136 | 0.008194 |
| 7 | 1986 | 21 | 0.001265 | 157 | 0.009459 |
| 8 | 1987 | 16 | 0.000964 | 173 | 0.010423 |
| 9 | 1988 | 15 | 0.000904 | 188 | 0.011327 |
| 10 | 1989 | 17 | 0.001024 | 205 | 0.012351 |
| 11 | 1990 | 16 | 0.000964 | 221 | 0.013315 |
| 12 | 1991 | 41 | 0.00247 | 262 | 0.015785 |
| 13 | 1992 | 43 | 0.002591 | 305 | 0.018376 |
| 14 | 1993 | 60 | 0.003615 | 365 | 0.021991 |
| 15 | 1994 | 121 | 0.00729 | 486 | 0.029281 |
| 16 | 1995 | 219 | 0.013194 | 705 | 0.042475 |
| 17 | 1996 | 263 | 0.015845 | 968 | 0.05832 |
| 18 | 1997 | 289 | 0.017412 | 1257 | 0.075732 |
| 19 | 1998 | 379 | 0.022834 | 1636 | 0.098566 |
| 20 | 1999 | 338 | 0.020364 | 1974 | 0.11893 |
| 21 | 2000 | 349 | 0.021027 | 2323 | 0.139957 |
| 22 | 2001 | 482 | 0.02904 | 2805 | 0.168996 |
| 23 | 2002 | 829 | 0.049946 | 3634 | 0.218942 |
| 24 | 2003 | 775 | 0.046692 | 4409 | 0.265634 |
| 25 | 2004 | 763 | 0.045969 | 5172 | 0.311604 |
| 26 | 2005 | 941 | 0.056694 | 6113 | 0.368297 |
| 27 | 2006 | 1008 | 0.06073 | 7121 | 0.429028 |
| 28 | 2007 | 1202 | 0.072418 | 8323 | 0.501446 |
| 29 | 2008 | 1428 | 0.086034 | 9751 | 0.58748 |
| 30 | 2009 | 1431 | 0.086215 | 11182 | 0.673696 |
| 31 | 2010 | 1259 | 0.075853 | 12441 | 0.749548 |
| 32 | 2011 | 1139 | 0.068623 | 13580 | 0.818171 |
| 33 | 2012 | 657 | 0.039583 | 14237 | 0.857754 |
| 34 | 2013 | 546 | 0.032896 | 14783 | 0.890649 |
| 35 | 2014 | 582 | 0.035064 | 15365 | 0.925714 |
| 36 | 2015 | 614 | 0.036992 | 15979 | 0.962706 |
| 37 | 2016 | 344 | 0.020725 | 16323 | 0.983432 |
| 38 | 2017 | 3 | 0.000181 | 16326 | 0.983612 |
| 39 | 2020 | 1 | 0,00006 | 16327 | 0.983673 |
| 40 | N/A | 271 | 0.016327 | 16598 | 1 |

Para representar graficamente estes dados, foi escolhido o gráfico de barras e o gráfico circular, tendo em conta que tal como foi visto, não houve necessidade de agrupar os dados em classes, o que implicaria a utilização de um histograma. Assim sendo, foram gerados os seguintes gráficos presentes na Gráfico 7 – Gráfico de Barras das frequências absolutas da variável 'Year' e na Gráfico 8 – Diagrama Circular das frequências relativas da variável 'Year'.

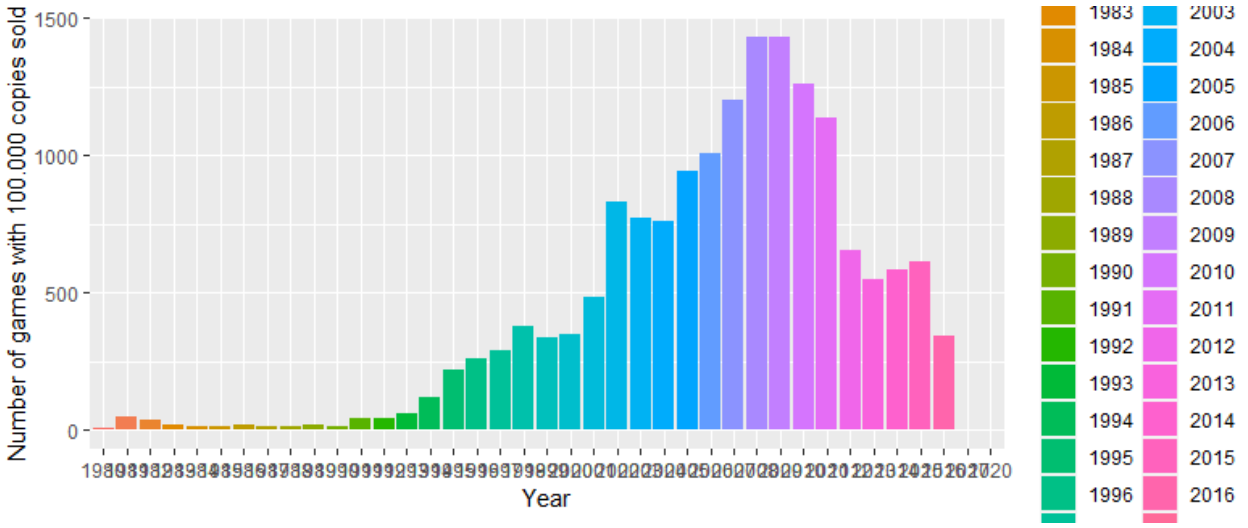


**Gráfico 7** – Gráfico de Barras das frequências absolutas da variável 'Year'

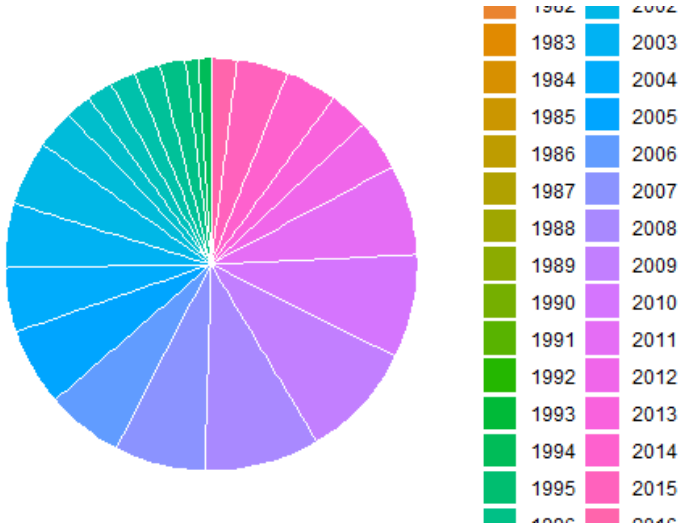


**Gráfico 8** – Gráfico Circular das frequências relativas da variável 'Year'

Depois, foram calculadas as restantes medidas de localização através dos valores "brutos", apresentadas na Fig.1 - Medidas de localização da variável 'Year'.

| Minimo | 1º Quartil | Mediana | Média | 3º Quartil | Máximo |
|---|---|---|---|---|---|
| 1980 | 2003 | 2007 | 2006 | 2010 | 2020 |

**Fig 1** – Restantes medidas de localização da variável 'Year'

Olhando novamente para a Tabela 4 - Frequências da variável 'Year', rapidamente se confirma os valores dos quartis (1º quartil está localizado na coluna 'Fi' - frequência relativa acumuada, na posição que contém o valor 0.25, mediana está localizada na coluna 'Fi' - frequência relativa acumuada, na posição que contém o valor 0.50, 3º quartil está localizado na coluna 'Fi' - frequência relativa acumuada, na posição que contém o valor 0.75).

Para a média, foi utilizada a seguinte fórmula:

$$\overline{X} = \frac{\Sigma xi}{n} \quad (1)$$

De seguida, apresenta-se o diagrama de extremo e quartis, presente no Gráfico 9 - Diagrama de extremos e quartis da variável 'Year'.

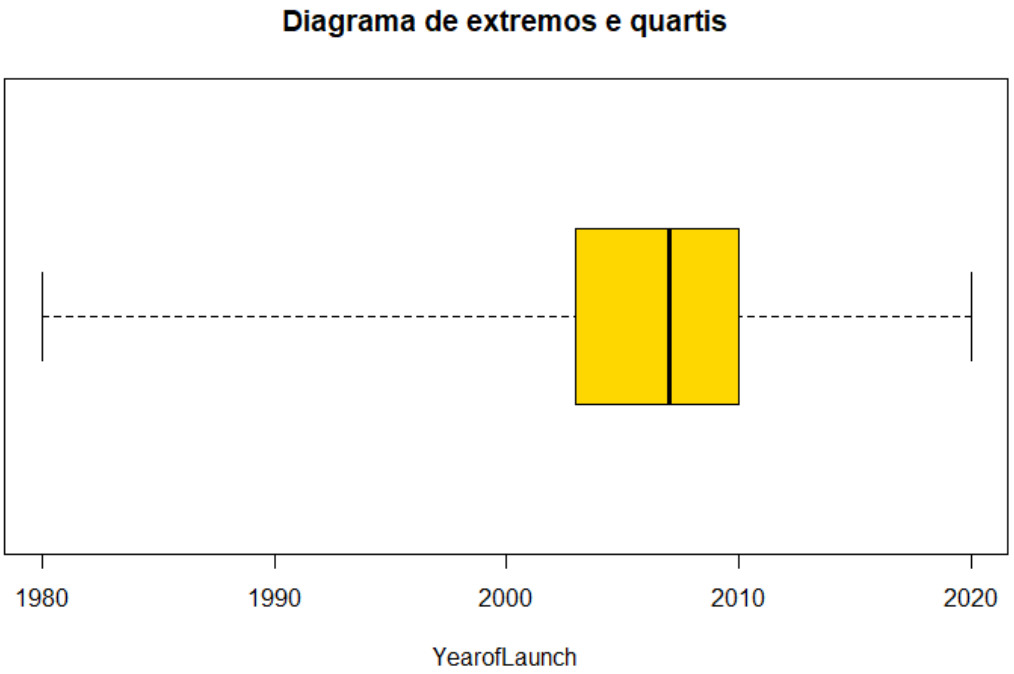


**Gráfico 9** – Diagrama de extremos e quartis da variável 'Year'

Quanto às medidas de dispersão, encontram-se representadas na Fig.2 - Medidas de dispersão da variável 'Year'.

| Variância | Desvio-Padrão | Amplitude Total | Amplitude Interquartil |
|---|---|---|---|
| 33.97702 | 5.828981 | 40 | 7 |

**Fig 2** – Medidas de dispersão da variável 'Year'

Foram calculadas através das seguintes fórmulas.

Para a variância:

$$s^2 = \frac{\sum^k i=1\ ni\ (xi' - \bar{x})^2}{n-1} \quad (2)$$

Para o desvio-padrão:

$$s \approx \sqrt{s^2} \quad (3)$$

Para a amplitude total:

$$A = max - min \quad (4)$$

Para a amplitude interquartil:

$$AIQ = Q0.75 - Q0.25 \quad (5)$$

A assimetria e a curtose foram calculadas, que estão representadas na Fig.5 - Assimetria e curtose da variável 'Year' através das seguintes fórmulas:

Fórmula da assimetria:

$$b_1 \approx \frac{1}{n} \sum_{i=1}^{n} \left(\frac{xi-\bar{x}}{s}\right)^3 \quad (6)$$

Fórmula da curtose:

$$b_2 \approx \frac{1}{n} \sum_{i=1}^{n} \left(\frac{xi-\bar{x}}{s}\right)^4 - 3 \quad (7)$$

| Assimetria | Curtose |
|---|---|
| -1.002376 | 1.846653 |

**Fig.3 -** Assimetria e curtose da variável 'Year'

Embora pudéssemos verificar através do histograma e do diagrama de extremos e quartis que existe uma assimetria negativa, optou-se também por complementar com o cálculo de $b1$.

Quanto à curtose, valor que indica que existe uma forte concentração de valores em torno da média, existe uma variação ligeiramente elevada, ou seja, o grau de curtose é do tipo curva leptocúrtica, alongada.

## 5. Variáveis quantitativas contínuas

Foi feita uma análise das mesmas. Foi feita uma transformação na variável de modo a construir uma tabela de frequências com classes. Foi também possível calcular as várias medidas de localização e de dispersão por serem variáveis do tipo acima mencionado.

Optámos pela utilização de histogramas para a representação gráfica destas variáveis, pois são o tipo de gráfico recomendado para este tipo de variável.

### 5.1. naSales - Vendas na América do Norte

Esta variável sendo do tipo quantitativa contínua, fez com que fosse necessário seguir uma série de passos de modo a construir uma tabela de frequências com classes.

Utilizou-se a regra de Sturges para determinar o nº de classes a criar, através da seguinte fórmula:

$$k = [1 + log2\, n] \quad (6)$$

De seguida, calculou-se a amplitude de cada classe, da seguinte maneira:

$$h = \frac{A}{k} \quad (7)$$

A classe modal foi "(0,2.77]", com uma frequência absoluta de 16427. Isto pode ser visualizado na tabela de frequências da variável que se encontra presente na Tabela 5 – Frequências da variável 'naSales'.

**Tabela 5** – Frequências da variável 'naSales'

| classes | niNaSales | fiNaSales | NiNaSales | FiNaSales |
|---|---|---|---|---|
| [0,2.77] | 16427 | 0.9896976 | 16427 | 0.9896976 |
| (2.77,5.53] | 117 | 0.007049 | 16544 | 0.9967466 |
| (5.53,8.3] | 29 | 0.0017472 | 16573 | 0.9984938 |
| (8.3,11.1] | 13 | 0.0007832 | 16586 | 0.999277 |
| (11.1,13.8] | 3 | 0.0001807 | 16589 | 0.9994578 |
| (13.8,16.6] | 5 | 0.0003012 | 16594 | 0.999759 |
| (16.6,19.4] | 0 | 0 | 16594 | 0.999759 |
| (19.4,22.1] | 0 | 0 | 16594 | 0.999759 |
| (22.1,24.9] | 1 | 0.0000602 | 16595 | 0.9998193 |
| (24.9,27.7] | 1 | 0.0000602 | 16596 | 0.9998795 |
| (27.7,30.4] | 1 | 0.0000602 | 16597 | 0.9999398 |
| (30.4,33.2] | 0 | 0 | 16597 | 0.9999398 |
| (33.2,36] | 0 | 0 | 16597 | 0.9999398 |
| (36,38.7] | 0 | 0 | 16597 | 0.9999398 |
| (38.7,41.5] | 1 | 0.0000602 | 16598 | 1 |

Para a representação gráfica, optou-se por um histograma, representado no Gráfico 9 – Histograma das frequências absolutas da variável 'naSales'. Conclui-se que os jogos que tiveram um nº de vendas (em milhões) na classe (0,2.77], formam a esmagadora maioria dos dados.

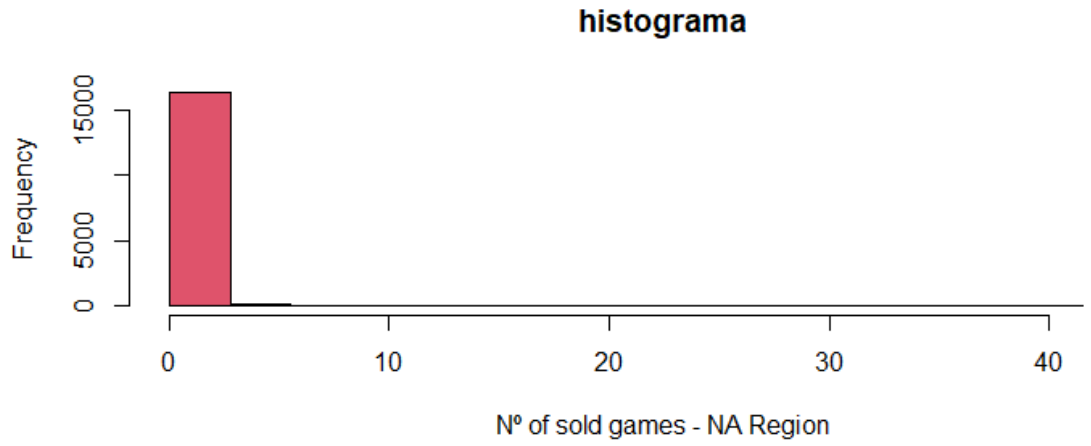


**Gráfico 9** – Gráfico de Barras das frequências absolutas da variável 'naSales'

Calculou-se o restante das medidas localização (a classe modal já tinha sido identificada), presente na Fig.3 - Medidas de localização da variável 'naSales'.

| Minimo | 1º Quartil | Mediana | Média | 3º Quartil | Máximo |
|---|---|---|---|---|---|
| 0.0000 | 0.0000 | 0.0800 | 0.2647 | 0.2400 | 41.4900 |

**Fig.4 -** Medidas de localização da variável 'naSales'

Para a média, foi utilizada a seguinte fórmula:

$$\overline{X} = \frac{\Sigma xi}{n} \quad (1)$$

Foram calculadas através das seguintes fórmulas.
Para a variância:

$$s^2 = \frac{\sum^k i=1\ ni\ (xi' - \bar{x})^2}{n-1} \quad (2)$$

Para o desvio-padrão:

$$s \approx \sqrt{s^2} \quad (3)$$

Para a amplitude total:

$$A = max - min \quad (4)$$

Para a amplitude interquartil:

$$AIQ = Q0.75 - Q0.25 \quad (5)$$

O diagrama de extremos e quartis, encontra-se, representado no Gráfico 10 – Diagrama de extremos e quartis da variável 'naSales'. Mais uma vez confirma-se a enorme concentração de dados na primeira classe.

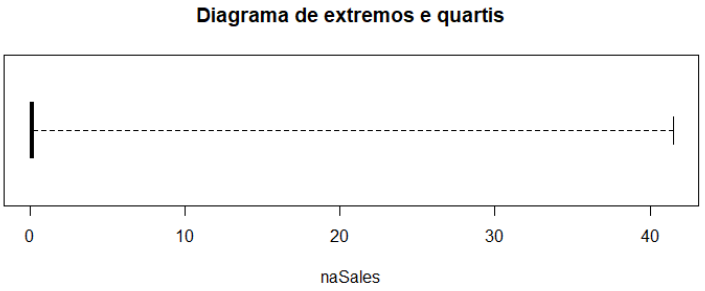


**Gráfico 10** – Diagrama de extremos e quartis da variável 'naSales'

As medidas de dispersão, que foram calculadas com auxílio de fórmulas também já mencionadas, estão representadas na Fig.4 - Medidas de dispersão da variável 'naSales'.

| Variância | Desvio-padrão | Amplitude total | Amplitude Interquartil |
|---|---|---|---|
| 0.6669712 | 0.816683 | 41.49 | 0.24 |

**Fig.5 -** Medidas de dispersão da variável 'naSales'

A assimetria e a curtose foram calculadas, que estão representadas na Fig.5 - Assimetria e curtose da variável 'naSales' através das seguintes fórmulas:

Fórmula da assimetria:

$$b_1 \approx \frac{1}{n}\sum_{i=1}^{n}\left(\frac{xi-\bar{x}}{s}\right)^3 \quad (6)$$

Fórmula da curtose:

$$b_2 \approx \frac{1}{n}\sum_{i=1}^{n}\left(\frac{xi-\bar{x}}{s}\right)^4 - 3 \quad (7)$$

| Assimetria | Curtose |
|---|---|
| 18.79793 | 651.9344 |

**Fig.6 -** Assimetria e curtose da variável 'naSales'

Embora pudéssemos verificar através do histograma e do diagrama de extremos e quartis que existe uma assimetria positiva, ou então através da comparação da moda, com a média e com a mediana (média < mediana < moda), optou-se por complementar com o cálculo de $b1$.

Quanto à curtose, valor que indica que existe uma forte concentração de valores em torno da média, existe uma variação extremamente elevada, ou seja, o grau de curtose é do tipo curva leptocúrtica, alongada.

### 5.2. euSales - Vendas na União Europeia

Esta variável sendo do tipo quantitativa contínua, fez com que fosse necessário seguir uma série de passos de modo a construir uma tabela de frequências com classes.

Utilizou-se a regra de Sturges para determinar o nº de classes a criar, através da seguinte fórmula:

$$k = [1 + log2\, n] \quad (6)$$

De seguida, calculou-se a amplitude de cada classe, da seguinte maneira:

$$h = \frac{A}{k} \quad (7)$$

A classe modal foi "(0,1.93]", com uma frequência absoluta de 16431. Isto pode ser visualizado na tabela de frequências da variável que se encontra presente na Tabela 5 – Frequências da variável 'euSales'.

**Tabela 6** – Frequências da variável 'euSales'

| classes | niEuSales | fiEuSales | NiEuSales | FiEuSales |
|---|---|---|---|---|
| [0,1.93] | 16431 | 0.9899385 | 16431 | 0.9899385 |
| (1.93,3.87] | 125 | 0.007531 | 16556 | 0.9974696 |
| (3.87,5.8] | 21 | 0.0012652 | 16577 | 0.9987348 |
| (5.8,7.74] | 10 | 0.0006025 | 16587 | 0.9993373 |
| (7.74,9.67] | 7 | 0.0004217 | 16594 | 0.999759 |
| (9.67,11.6] | 2 | 0.0001205 | 16596 | 0.9998795 |
| (11.6,13.5] | 1 | 0.0000602 | 16597 | 0.9999398 |
| (13.5,15.5] | 0 | 0 | 16597 | 0.9999398 |
| (15.5,17.4] | 0 | 0 | 16597 | 0.9999398 |
| (17.4,19.3] | 0 | 0 | 16597 | 0.9999398 |
| (19.3,21.3] | 0 | 0 | 16597 | 0.9999398 |
| (21.3,23.2] | 0 | 0 | 16597 | 0.9999398 |
| (23.2,25.2] | 0 | 0 | 16597 | 0.9999398 |
| (25.2,27.1] | 0 | 0 | 16597 | 0.9999398 |
| (27.1,29] | 1 | 0.0000602 | 16598 | 1 |

Para a representação gráfica, optou-se por um histograma, representado no Gráfico 11 – Histograma das frequências absolutas da variável 'euSales'. Conclui-se que os jogos que tiveram um nº de vendas (em milhões) na classe (0, 1.93], formam a esmagadora maioria dos dados.

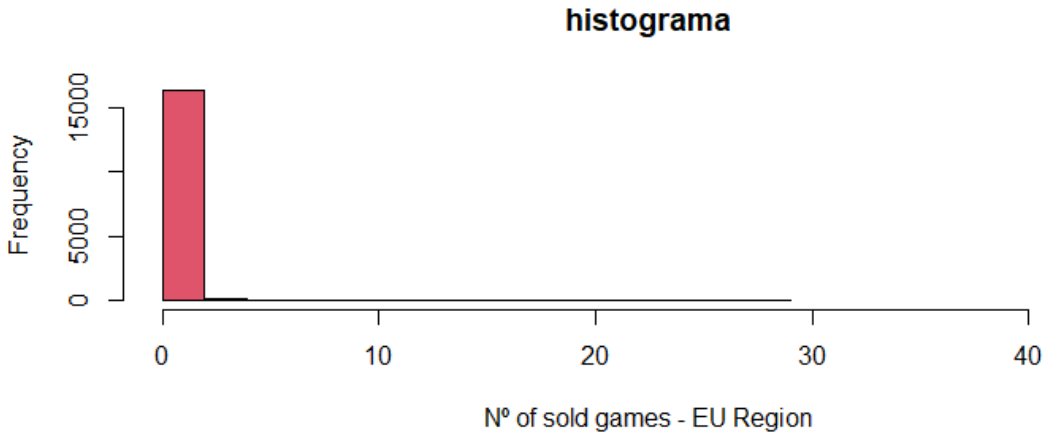


**Gráfico 11** – Gráfico de Barras das frequências absolutas da variável 'euSales'

Calculou-se o restante das medidas localização (a classe modal já tinha sido identificada), presente na Fig.6 - Medidas de localização da variável 'euSales'.

| Minimo | 1º Quartil | Mediana | Média | 3º Quartil | Máximo |
|---|---|---|---|---|---|
| 0.0000 | 0.0000 | 0.0200 | 0.1467 | 0.1100 | 29.0200 |

**Fig.7 -** Medidas de localização da variável 'euSales'

Para a média, foi utilizada a seguinte fórmula:

$$\overline{X} = \frac{\Sigma xi}{n} \quad (1)$$

Foram calculadas através das seguintes fórmulas.
Para a variância:

$$s^2 = \frac{\sum^k i=1\ ni\ (xi' - \bar{x})^2}{n-1} \quad (2)$$

Para o desvio-padrão:

$$s \approx \sqrt{s^2} \quad (3)$$

Para a amplitude total:

$$A = max - min \quad (4)$$

Para a amplitude interquartil:

$$AIQ = Q0.75 - Q0.25 \quad (5)$$

O diagrama de extremos e quartis, encontra-se, representado no Gráfico 12 – Diagrama de extremos e quartis da variável 'euSales'. Mais uma vez confirma-se a enorme concentração de dados na primeira classe.

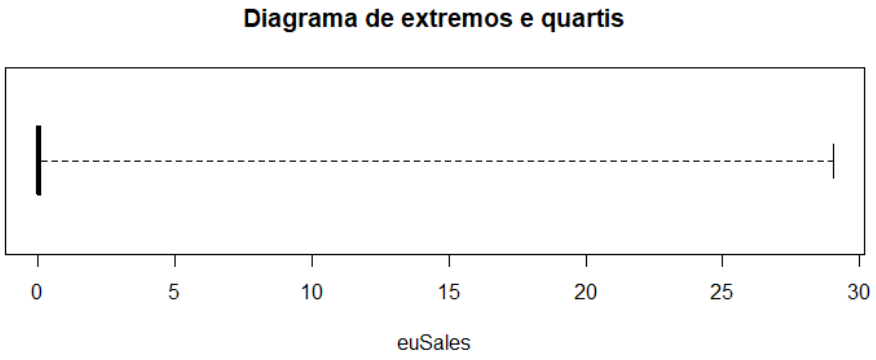


**Gráfico 12 –** Diagrama de extremos e quartis da variável 'euSales'

As medidas de dispersão, que foram calculadas com auxílio de fórmulas também já mencionadas, estão representadas na Fig.7 - Medidas de dispersão da variável 'euSales'.

| Variância | Desvio-padrão | Amplitude Total | Amplitude Interquartil |
|---|---|---|---|
| 0.2553799 | 0.5053512 | 29.02 | 0.11 |

**Fig.8 -** Medidas de dispersão da variável 'euSales'

A assimetria e a curtose, foram calculadas, que estão representadas na Fig.8 - Assimetria e curtose da variável 'euSales' através das seguintes fórmulas:

Fórmula da assimetria:

$$b_1 \approx \frac{1}{n}\sum_{i=1}^{n}\left(\frac{xi-\bar{x}}{s}\right)^3 \quad (6)$$

Fórmula da curtose:

$$b_2 \approx \frac{1}{n}\sum_{i=1}^{n}\left(\frac{xi-\bar{x}}{s}\right)^4 - 3 \quad (7)$$

| Assimetria | Curtose |
|---|---|
| 18.87383 | 758.7997 |

**Fig.9 -** Assimetria e curtose da variável 'euSales'

Novamente, embora pudéssemos verificar através do histograma e do diagrama de extremos e quartis que existe uma assimetria positiva, ou então através da comparação da moda, com a média e com a mediana (média < mediana < moda), optou-se por complementar com o cálculo de $b1$.

Quanto à curtose, valor que indica que existe uma forte concentração de valores em torno da média, existe uma variação extremamente elevada, ou seja, o grau de curtose é do tipo curva leptocúrtica, alongada.

### 5.3. globalSales - Vendas a nível mundial

Esta variável sendo do tipo quantitativa contínua, fez com que fosse necessário seguir uma série de passos de modo a construir uma tabela de frequências com classes.

Utilizou-se a regra de Sturges para determinar o nº de classes a criar, através da seguinte fórmula:

$$k = [1 + log2\, n] \quad (6)$$

De seguida, calculou-se a amplitude de cada classe, da seguinte maneira:

$$h = \frac{A}{k} \quad (7)$$

A classe modal foi "(0.01,5.53]", com uma frequência absoluta de 16438. Isto pode ser visualizado na tabela de frequências da variável que se encontra presente na Tabela 7 – Frequências da variável 'globalSales'.

**Tabela 7** – Frequências da variável 'globalSales'

| classes | niGlobalSales | fiGlobalSales | NiGlobalSales | FiGlobalSales |
|---|---|---|---|---|
| [0.01,5.53] | 16438 | 0.9903603 | 16438 | 0.9903603 |
| (5.53,11] | 109 | 0.0065671 | 16547 | 0.9969273 |
| (11,16.6] | 28 | 0.001687 | 16575 | 0.9986143 |
| (16.6,22.1] | 9 | 0.0005422 | 16584 | 0.9991565 |
| (22.1,27.6] | 4 | 0.000241 | 16588 | 0.9993975 |
| (27.6,33.1] | 7 | 0.0004217 | 16595 | 0.9998193 |
| (33.1,38.6] | 1 | 0.0000602 | 16596 | 0.9998795 |
| (38.6,44.1] | 1 | 0.0000602 | 16597 | 0.9999398 |
| (44.1,49.6] | 0 | 0 | 16597 | 0.9999398 |
| (49.6,55.2] | 0 | 0 | 16597 | 0.9999398 |
| (55.2,60.7] | 0 | 0 | 16597 | 0.9999398 |
| (60.7,66.2] | 0 | 0 | 16597 | 0.9999398 |
| (66.2,71.7] | 0 | 0 | 16597 | 0.9999398 |
| (71.7,77.2] | 0 | 0 | 16597 | 0.9999398 |
| (77.2,82.7] | 1 | 0.0000602 | 16598 | 1 |

Para a representação gráfica, optou-se por um histograma, representado no Gráfico 13 – Histograma das frequências absolutas da variável 'globalSales'. Conclui-se que os jogos que tiveram um nº de vendas (em milhões) na classe (0.01,5.53], formam a esmagadora maioria dos dados.

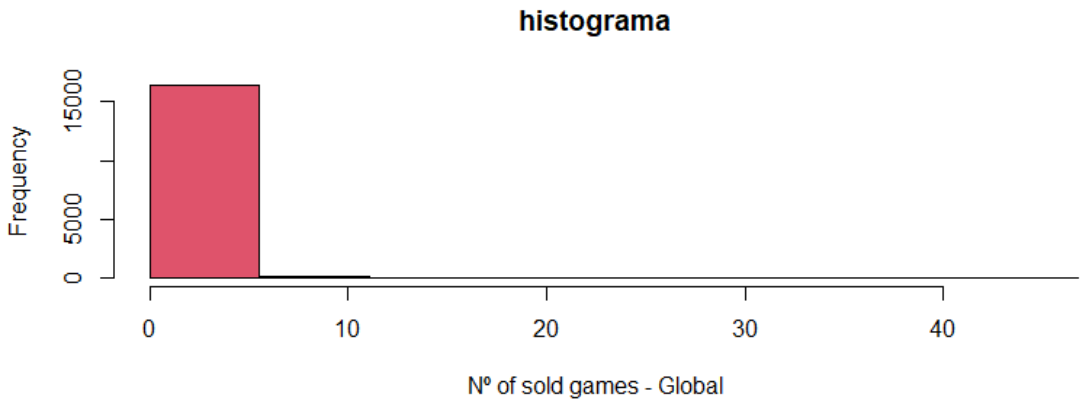


**Gráfico 13** – Gráfico de Barras das frequências absolutas da variável 'globalSales'

Calculou-se o restante das medidas localização (a classe modal já tinha sido identificada), presente na Fig.9 - Medidas de localização da variável 'globalSales'.

| Minimo | 1º Quartil | Mediana | Média | 3º Quartil | Máximo |
|---|---|---|---|---|---|
| 0.0100 | 0.0600 | 0.1700 | 0.5374 | 0.4700 | 82.740 |

**Fig.10 -** Medidas de localização da variável 'globalSales'

Para a média, foi utilizada a seguinte fórmula:

$$\overline{X} = \frac{\Sigma xi}{n} \quad (1)$$

Foram calculadas através das seguintes fórmulas.
Para a variância:

$$s^2 = \frac{\sum^k i=1\ ni\ (xi' - \bar{x})^2}{n-1} \quad (2)$$

Para o desvio-padrão:

$$s \approx \sqrt{s^2} \quad (3)$$

Para a amplitude total:

$$A = max - min \quad (4)$$

Para a amplitude interquartil:

$$AIQ = Q0.75 - Q0.25 \quad (5)$$

O diagrama de extremos e quartis, encontra-se, representado no Gráfico 14 – Diagrama de extremos e quartis da variável 'globalSales'. Mais uma vez confirma-se a enorme concentração de dados na primeira classe.

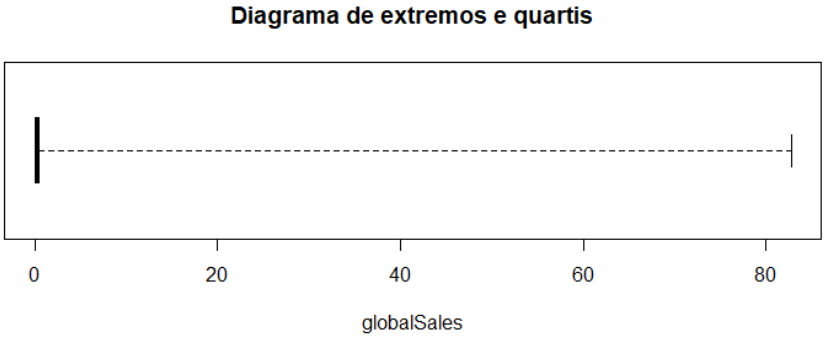


**Gráfico 14** – Diagrama de extremos e quartis da variável 'globalSales'

As medidas de dispersão, que foram calculadas com auxílio de fórmulas também já mencionadas, estão representadas na Fig.10 - Medidas de dispersão da variável 'globalSales'.

| Variância | Desvio-padrão | Amplitude Total | Amplitude Interquartil |
|---|---|---|---|
| 2.418112 | 1.555028 | 82.73 | 0.41 |

**Fig.11 -** Medidas de dispersão da variável 'globalSales'

A assimetria e a curtose, foram calculadas, que estão representadas na Fig.11 - Assimetria e curtose da variável 'globalSales' através das seguintes fórmulas:

Fórmula da assimetria:

$$b_1 \approx \frac{1}{n}\sum_{i=1}^{n}\left(\frac{xi-\bar{x}}{s}\right)^3 \quad (6)$$

Fórmula da curtose:

$$b_2 \approx \frac{1}{n}\sum_{i=1}^{n}\left(\frac{xi-\bar{x}}{s}\right)^4 - 3 \quad (7)$$

| Assimetria | Curtose |
|---|---|
| 17.39907 | 606.7501 |

**Fig.12 -** Assimetria e curtose da variável 'globalSales'

E desta vez, a nível mundial a tendência também se reflete. Podemos verificar através do histograma e do diagrama de extremos e quartis que existe uma assimetria positiva, ou então através da comparação da moda, com a média e com a mediana (média < mediana < moda), optou-se por complementar com o cálculo de $b1$.

Quanto à curtose, valor que indica que existe uma forte concentração de valores em torno da média, existe uma variação extremamente elevada, ou seja, o grau de curtose é do tipo curva leptocúrtica, alongada.

## 6. Conclusões

Este trabalho fez com que os elementos do grupo adquirissem conhecimentos relativamente avançados da linguagem R, o que certamente contribuirá para o futuro dos elementos do grupo como programadores.

A análise descritiva feita ao conjunto de dados fez com que voltássemos a rever alguns conceitos e fórmulas no contexto da Estatística Descritiva, o que foi algo bastante positivo para a consolidação dos conhecimentos.

Foi também deduzido pelo grupo, durante o trabalho, que para além de saber programar corretamente na linguagem mencionada, também é igualmente importante saber analisar os dados, de modo a saber tirar conclusões.

## Referências Bibliográficas


- Fichas 1 e 2, das aulas práticas.
- Slides disponibilizados pelos professores no moodle, ”Capítulo 1 – Estatística Descritiva”.