Facultad de Ciencias Económicas y Empresariales, ICADE

¿Influyen las críticas de videojuegos en sus ventas?

Análisis econométrico del impacto de las críticas y otras variables en la industria de los videojuegos

Trabajo de Fin de Grado de Administración de Empresas.

Realizado por Jaime Alonso López-Linares

5º E2 Analytics

Dirigido por Eduardo César Garrido Merchán

Madrid, junio 2023

COMILLAS
UNIVERSIDAD PONTIFICIA
ICADE

Resumen

La industria de los videojuegos está en constante crecimiento, superando incluso a la cinematográfica. Las inversiones en juegos triple A (AAA) son abismales, con un gran número de factores que pueden influir en el éxito o fracaso de un juego. Aunque muchos de estos factores son similares a los que se encuentran en otras industrias, como la calidad del producto o el marketing, hay uno que es particularmente característico de los videojuegos: las críticas. Las críticas de videojuegos se han convertido en una industria en sí misma, con sitios web dedicados exclusivamente a la revisión y puntuación de los nuevos lanzamientos. Pero ¿cuál es el impacto real de estas críticas en el éxito comercial de un videojuego? Para responder a esta pregunta, se han utilizado técnicas econométricas y se ha analizado la historia del mercado de videojuegos. Pero las críticas no son el único factor que puede influir en el éxito de un videojuego. Otros factores que considerar incluyen la presencia en redes sociales, el género del juego, la consola principal en la que se juega, el efecto de los "influencers" o la puntuación en foros entre otros. Para entender mejor estos factores y cómo afectan las ventas de videojuegos, se ha desarrollado un modelo que busca explicar el comportamiento de las ventas en función de estos factores. Este modelo puede proporcionar una visión estratégica completa del estado actual de la industria de los videojuegos en España, así como información fácilmente accionable y un procedimiento escalable para estudios futuros. En resumen, la industria de los videojuegos es compleja y está en constante evolución, pero a través del análisis cuidadoso de los diferentes factores que influyen en el éxito de un juego, se puede obtener una mejor comprensión de cómo funciona y cómo mejorar en el futuro.

Palabras clave: Consola, críticos, data science, desarrollador, econometría, editorial, foros, influencer, plataforma, triple A, videojuegos.

Abstract

The video game industry is constantly growing, even surpassing the film industry. Investments in triple-A (AAA) games are enormous, with many factors that can influence the success or failure of a game. Although many of these factors are like those found in other industries, such as product quality or marketing, there is one that is particularly characteristic of video games: reviews. Video game reviews have become an industry in themselves, with websites dedicated exclusively to reviewing and rating new releases. But what is the real impact of these reviews on the commercial success of a video game? To answer this question, econometric techniques have been used and the history of the global and Spanish video game market has been analyzed. However, reviews are not the only factor that can influence the success of a video game. Other factors to consider include social media presence, game genre, main console on which it is played, the effect of influencers, and ratings in forums among others. To better understand these and how they affect video game sales, a model has been developed that seeks to explain sales behavior based on these factors. This model can provide a complete strategic overview of the current state of the video game industry in Spain, as well as actionable information and a scalable procedure for future studies. In summary, the video game industry is complex and constantly evolving, but through careful analysis of the different factors that influence the success of a game, we can obtain a better understanding of how it works and how we can improve it in the future.

Keywords: Console, critics, data science, developer, econometry, forums, influencer, platform, publisher, triple A, videogames.

A mi tutor, Eduardo César Garrido Merchán, profesor de la Universidad Pontificia de Comillas, por su consideración y confianza depositada en mi para llevar este proyecto, además del constante consejo recibido.

A mi familia y a Celia por darme fuerzas y siempre haber marcado un ejemplo a seguir.

Al personal médico del Hospital General Universitario Gregorio Marañón, especialmente a los doctores Dr. Ignacio Navarro Cuéllar, Dr. Santiago Ochandiano Caicoya, Dra. Marta Arregui Valles, Dra. Carolina Agra Pujol, Dr. Carlos Navarro Cuéllar y a la enfermera Loreto Fernández Bermejo. Por la constante atención recibida durante estos más de 5 meses y por su gran humanidad.

A la AEVI (Asociación Española de Videojuegos), por proveer los datos de ventas de videojuegos de los que depende este proyecto.

ÍNDICE

ILUSTRACIONES Y GRÁFICOS

**Las imágenes en el apartado de anexos no han sido incluidas en este índice.*

ECUACIONES

TABLAS

1. INTRODUCCIÓN

A. Interés del tema y contexto histórico

La industria de los videojuegos ha visto un auge continuo desde que surge el primer videojuego "Nought and Crosses" en 1952, consolidándose durante los 60 con la aparición de títulos como "Pong" o "Space War" que más tarde desembocarían en el primer gran estandarte de la industria: "Space Invaders" en 1972. Este desarrollo en títulos iba de la mano de la evolución tecnológica de los propios recreativos y lo que se convertiría en el futuro del mercado; los sistemas domésticos, con la aparición del Atari 2600. Debido a la juventud del mercado había una amplia cantidad de oferentes en cuanto a consolas ya que todos luchaban por mantener una parte de la industria emergente; competían fundamentalmente Phillips, Mattel, Atari, Coleco, Commodore y NEC. Sin embargo, debido a las diferencias geográficas de cada fabricante de consolas y a las cambiantes tendencias que se daban en Estados Unidos, Europa y Japón, durante los 80 los actores importantes de esta industria pasarían a ser Nintendo, con su NES (que toma el control de EE. UU.), Atari con la 7600, Sega con la Master System y finalmente Commodore con la Amiga gracias a juegos como el Tetris (Belli & Raventós, 2008).

En la década de los 90 se empieza a ver una consolidación real por parte del mercado, ya que la firma que domina este periodo sigue siendo un eje angular de la industria 30 años después; Nintendo se asienta a través de la consola SNES al comienzo de la década y ante la aparición de un nuevo competidor como Sony en 1994 con la primera PlayStation, responde con el lanzamiento de una de las consolas mejor valoradas de la historia; la Nintendo 64. Así ya se tiene a 2 de los integrantes del triunvirato que supone la industria moderna, pero para ver al tercer integrante todavía hay que esperar una década más. Además de esto cabe destacar el enorme desarrollo que hubo en el ámbito de las consolas portátiles, donde el dominador absoluto del nicho será otra vez Nintendo con la Game Boy. Este giro por parte del anterior dominador en el espectro de las consolas domésticas viene causado principalmente por el hecho de que Sony con su PlayStation se convierte en el líder del mercado gracias a enfocarse en públicos más maduros con títulos clave como "Final Fantasy VII", "Gran Turismo", "Metal Gear" Solid o "Resident Evil", que ofrecen una propuesta de valor diferente a la que ofrecía Nintendo con títulos como los de las sagas "Super Mario Bros", "The Legend of Zelda" o "Metroid". Es en esta época cuando los títulos ofrecidos por cada consola empiezan a ganar relevancia sobre las capacidades del sistema, y para el año 2001 el top 5% de juegos obtienen el 50% de los ingresos del mercado (Clements & Ohashi, 2005). Con estos 2 competidores peleándose por el mercado el resto fueron abandonando el hardware y centrándose en el de desarrollo de software, como es el caso de SEGA a partir del 2002. Sería en ese mismo periodo, en el que de la mano de Microsoft, se lanzaría la primera Xbox (lanzada en noviembre de 2001 en América y 2002 en el resto del mundo), inicialmente planteada

para competir contra la PlayStation 2 en Norteamérica, gracias a su juego estandarte: la saga "Halo".

Mientras tanto, el contenido de los títulos pasó a centrarse cada vez más en el multijugador, con desarrolladores de software como Activision Blizzard ganando gran relevancia por títulos como "Starcraft", "World of Warcraft" o "Diablo", que se desarrollaban para ordenador. De esta forma algunos fabricantes de videojuegos pretenden lograr cierta independencia del oligopolio formado en el mundo de las consolas, al centrarse juegos para PC (Belli & Raventós, 2008).

Sucesivas generaciones después, el estado del mercado de los videojuegos sigue relativamente similar a lo que se podía ver a principios de los 2000, en la industria hay 3 grandes dominadores, de los cuales la competencia es extrema entre dos: Xbox y PlayStation, mientras que el tercero, Nintendo, siempre enfocado hacia jugadores o jóvenes o muy conocedores de la industria, se mantiene a parte gracias a la enorme calidad y originalidad de su propiedad intelectual. Esta es la conclusión general que se deriva de este breve repaso de la evolución de la industria de los videojuegos; el poder vender consolas no parece basarse únicamente en la calidad técnica del dispositivo si no que parece incluso más importante la calidad de los juegos producidos exclusivamente para esa plataforma. Es por esto por lo que es absolutamente fundamental para una editorial de videojuegos, o "publisher" como se conoce en el mercado, obtener los mejores juegos exclusivos para sus plataformas y comunicar la calidad de estos.

En la industria del videojuego el rol de editorial suele recaer sobre el propio fabricante de la consola, como por ejemplo Nintendo, cuyos juegos en su actual consola; la Nintendo Switch, suelen ser exclusivos para esta. Sin embargo, también existen otras editoriales de enorme tamaño sin fabricar consolas, como son Electronic Arts, la anteriormente mencionada Activision Blizzard o desde China, Tencent Games (Buijsman, 2022). Así se ve que por un lado se tiene la competencia de las consolas y, por otro lado, aunque extremadamente relacionado, la competencia por la venta de videojuegos. Debido al rol fundamental que tienen los títulos exclusivos (los videojuegos que solo se venden para una plataforma), estos dos lados de la industria están extremadamente conectados, además del hecho de que los 3 fabricantes principales de consolas son algunos de los publishers más grandes del mercado.

A estos cambios históricos en el modelo de negocio se suman nuevas corrientes o tendencias en la industria como el dominio de los juegos multijugador, la llegada de la monetización a través de micro transacciones o que cada vez se vendan más juegos en formato digital en vez de en físico. Estos fenómenos solo se han visto en aumento debido al cambio del comportamiento de los jugadores de videojuegos durante la Pandemia del COVID-19 y los confinamientos que supuso (Ortiz, Tillerias, H., C., & Toaza, 2020). Todos estos efectos y el rol creciente de Internet en cuanto a la información que tienen los consumidores y su capacidad para comunicarse entre sí

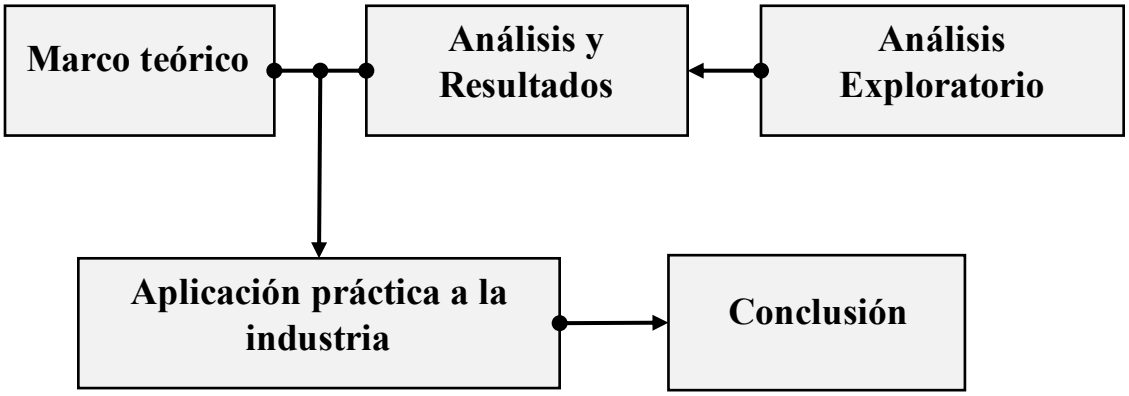
y formar comunidades de jugadores con gustos parecidos (Clements & Ohashi, 2005) han provocado un antes y un después en lo que supone ser un publisher de video juegos, ya que mientras que antes su rol principal era el de distribuidor y dotante de información para los consumidores, ahora han pasado a ser anunciantes, que aseguran a los desarrolladores que a través de ellos su juego será visto por su comunidad.

Es aquí donde entra el interés fundamental de este trabajo; ya que, en esta nueva industria de multitud de información disponible existen unos agentes cuyo trabajo es asegurar la calidad de los juegos de forma supuestamente desinteresada; los críticos de videojuegos profesionales. Estas páginas web y revistas actúan de forma similar a las Rating Agencies como Moody's o Standard & Poor's, su modelo de negocio es el de dar una puntuación a videojuegos para determinar su calidad general y si recomendarían o no comprarlo a través de elaborados artículos web en los que los evaluadores de la página describen su experiencia con el juego. Una vez cuelgan el artículo en sus páginas, su forma de ganar dinero es mediante la subasta de anuncios en dichas páginas. Así, tras haber estado repasando la importancia que tienen los juegos para las ventas de consolas y para los publishers en general, se puede ver como el papel de estas páginas de críticos profesionales es absolutamente fundamental a la hora de que un videojuego tenga éxito, y en el caso de juegos exclusivos, que la plataforma en cuestión tenga éxito. Según ciertos estudios el impacto de estas críticas es extremadamente significativo en las ventas de videojuegos, mientras que otros factores como el "word of mouth", es decir, la conversación de la comunidad alrededor de ciertos títulos tiene escasa importancia en dicha variable dependiente (las ventas) (Cox & Kaimann, 2015). Debido a estos factores es especialmente relevante el estudio de la relación entre esta parte de la industria y las ventas de videojuegos, ya que, si realmente es demostrable la relación que pueden tener las críticas de cada medio en las ventas de cada videojuego, se podrá observar cómo cada medio afecta a las ventas de forma individual, entre otros comportamientos.

B. Estructura

Este trabajo de fin de grado va a estar formado por 5 partes fundamentales:

Ilustración 1: Estructura del proyecto



A parte de esta estructura simple, también contendrá **Definición y alcance del proyecto** y **Bibliografía** y **Anexos** al final.

2. MARCO TEÓRICO O DISEÑO CONCEPTUAL

A. Econometría

"Pregunte a media docena de económetras qué es la econometría y obtendrá media docena de respuestas diferentes. Alguien podría decirle que la econometría es la ciencia para la contratación de teorías económicas. Un segundo podría decirle que la econometría es el conjunto de herramientas utilizadas para predicción de valores futuros de variables económicas, tales como las ventas de las empresas, el crecimiento de la economía en su conjunto, o el precio de las acciones. Otro podría decir que la econometría es el proceso de ajuste de modelos económicos matemáticos a los datos del mundo real. Un cuarto podría decirle que es la ciencia y el arte de utilizar los datos históricos para realizar recomendaciones numéricas, o cuantitativas, sobre las políticas a realizar por el gobierno en los negocios." (Stock, Watson, & Larrión, 2012)

Así es como define la econometría el libro "Introducción a la econometría" de Stock, Watson y Larrión, a partir del cual se van a definir una serie de conceptos que se deberán entender a la hora de realizar los análisis que se buscan llevar a cabo para contrastar las hipótesis que se han sentado en los objetivos del trabajo. Todas las definiciones que menciona este párrafo son adecuadas para definir la ciencia de la econometría, y por ello, mediante el análisis de las variables cuya selección se explicará en el cuarto apartado de este proyecto, y realizando un estudio acorde de estas, se construirán modelos para obtener ciertas conclusiones aplicables a la industria analizada.

El primer punto que se debe explicar teóricamente es el de este estudio de las variables, ya que antes de poder comenzar a construir modelos con las mismas, se debe obtener información previa para estudiar factores como su curtosis o asimetría entre otras. Dado que un gran número de las variables que se van a usar son numéricas, también se van a explicar brevemente conceptos clave para resumir estas como son la esperanza (o media), desviación típica y varianza:

- Esperanza: el valor medio de una variable a lo largo del tiempo. Se calcula dividiendo la suma total de la variable "x" entre el número total de registros de la variable.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_1}{n}$$

Ecuación 1: Fórmula de la media o esperanza

- Desviación Típica y Varianza: son medidas estadísticas que se utilizan para medir la dispersión de una variable o de su distribución de probabilidad. La varianza se calcula a partir del valor esperado del cuadrado de la desviación entre una variable "x" y su media.

Como la varianza se calcula con el cuadrado, para una interpretación más sencilla se suele utilizar la desviación típica, que consiste en la raíz de la varianza.

$$S^2 = \frac{\sum_{i=1}^{n} (x_j - \bar{x})^2}{n}$$

Ecuación 2: Fórmula de la varianza

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_j - \bar{x})^2}{n}}$$

Ecuación 3: Fórmula de la desviación típica

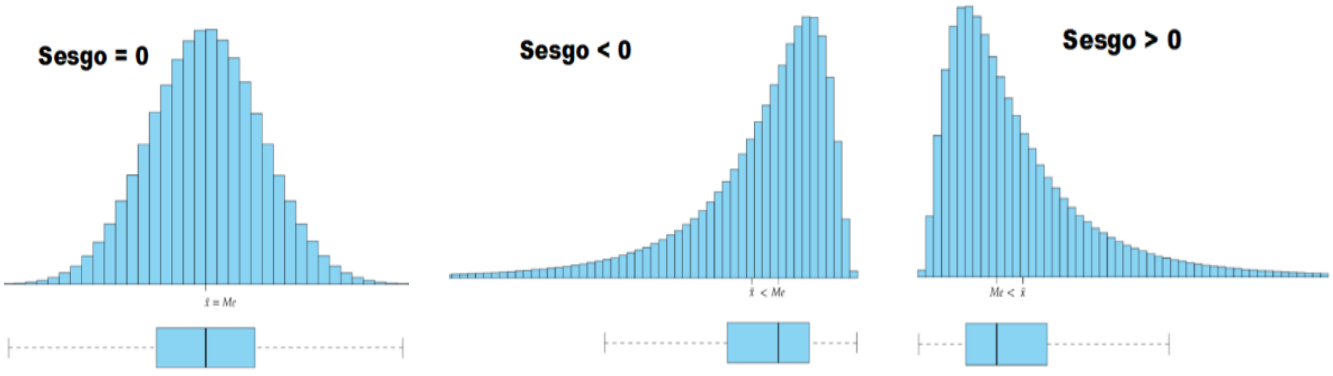
Una vez definidos esos estadísticos básicos se pueden observar otras medidas clásicas como son:

- Asimetría: Calcula matemáticamente cuanto se aleja una función o variable de la simetría, en una distribución simétrica, el valor del numerador será igual a 0, sin embargo, si este es negativo significará que la función es asimétrica hacia la izquierda, y si es positivo hacia la derecha.

$$A = \frac{E[(x - \bar{x})^3]}{S_x^3}$$

Ecuación 4: Fórmula de la asimetría

Ilustración 2: Distribuciones simétricas y asimétricas (Llinás Solano, 2021)



Simétrica **Asimétrica hacia la izquierda** **Asimétrica hacia la derecha**

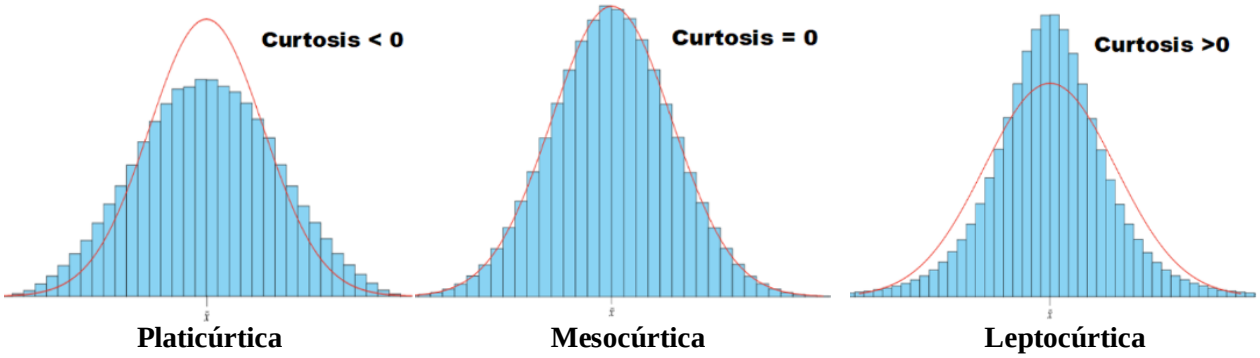
- Curtosis: Es la medida de cuanta masa probabilística se encuentra en las colas de la variable (cuantos registros están acumulados en las colas). De este concepto se extrae que los valores extremos se llamen atípicos; normalmente se encuentran solos. Cuanta más

masa tenga una variable en sus colas, más probable será encontrar una mayor desviación típica (puesto que los valores próximos a la media no estarán tan poblados). Un valor de curtosis mayor a 0 significa que la función es leptocúrtica. Un valor de 0 significa que es mesocúrtica o que las colas están igual de pobladas que la distribución normal y si hay muchos valores en los extremos será platicúrtica.

$$C = \frac{E[(x - \bar{x})^4]}{S_x^4} - 3$$

Ecuación 5: Fórmula de la curtosis

Ilustración 3: Tipos de curtosis y gráficos (Llinás Solano, 2021)



Todas estas métricas se utilizan para evaluar la función de densidad de una variable, en la que el área bajo dos puntos de dicha función es la probabilidad de que, al coger una variable aleatoria, caiga entre dichos puntos. Según la forma de esta función se construyen diferentes distribuciones como la Normal estándar, la Chi Cuadrado o la t de Student *(Stock, Watson, & Larrión, 2012)*

A partir de los conceptos ya definidos se procede al **análisis univariante de las variables**, la descripción mediante métricas como la esperanza, desviación típica, la curtosis, la asimetría o la densidad entre otras. Se utiliza para estudiar el comportamiento de una variable de la forma más simple posible; no se pretende estudiar relaciones ni causas, solo describir y encontrar patrones entre unos datos sencillos.

El paso siguiente a esto es el análisis bivariante o multivariante que busca medir las relaciones entre las variables previamente estudiadas para descubrir interacciones o comportamientos en común que generalmente se miden mediante dos métricas básicas: la covarianza y la correlación.

- Covarianzas: mide el grado de independencia entre dos variables, por lo que si dos variables son completamente independientes su covarianza será igual a 0.
- Correlaciones: matriz de correlación, que expresa la relación lineal de dos variables.

B. Regresión Lineal

El término regresión viene un estudio del siglo XIX llevado a cabo por Sir Francis Galton, en el que buscaba, a través de información sobre la altura de padres, determinar que altura tendrían los hijos (Stanton, 2001). De esta forma se ve un primer ejemplo de la regresión lineal simple con fin predictivo; con una variable trataba de estimar el valor de otra basándose en la asunción de que existía una relación linear entre ambas.

La regresión lineal múltiple supone un paso más a la idea original de Galton, ya que toma esa misma idea, pero incorpora numerosas variables independientes entre sí. En la actualidad, es uno de los métodos más comunes a la hora de determinar la relación entre una variable objetivo, que se llamará dependiente y unas variables independientes. Esto se debe en gran parte a que la simplicidad de este método al ser aditivo, es decir, que cada variable actúa de forma aditiva y constante para determinar la variable objetivo (Baños, Torrado-Fonseca, & Álvarez, 2019). Esta simplicidad aporta una claridad distintiva a la hora de entender dicho método, que permite ver de forma objetiva cuanto aporta cada variable al cálculo de la estimación de la variable dependiente.

La ecuación sobre la que se construye este análisis es la siguiente (Pérez & Santín, 2008):

$$\mathrm{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mathrm{u}$$

Ecuación 6: Fórmula de la regresión lineal

En este modelo la Y se refiere a la variable dependiente, las β al peso que tienen las variables independientes en el cálculo de la independiente, las x se refieren al valor de las variables independientes y finalmente la u corresponde al error del modelo o lo que es lo mismo, al componente aleatorio. Mientras que si el foco del estudio es predictivo la importancia cae sobre las estimaciones producidas por este, si el estudio es explicativo el foco caerá sobre las betas, puesto que son estas las que explican en profundidad la relación entre las variables.

Debido a la propia definición y a los principios que sigue el método de regresión hay ciertos supuestos que se deben cumplir para que este método produzca resultados válidos para interpretación. Estos son:

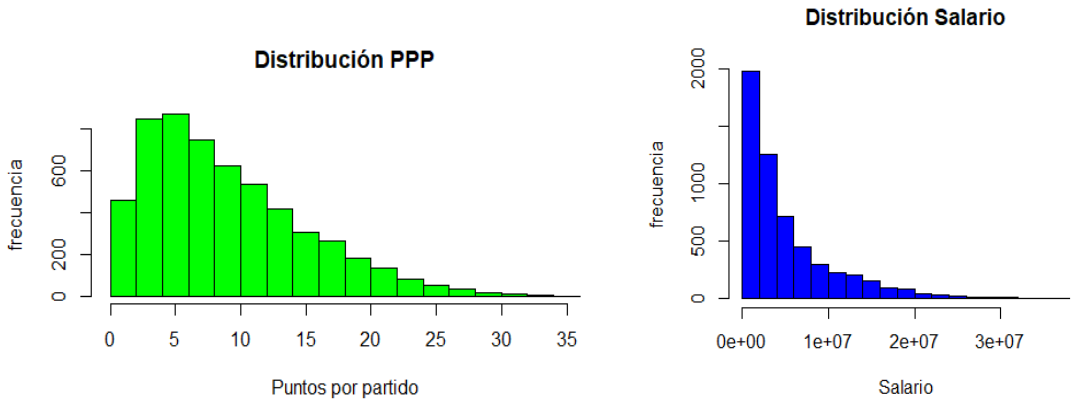
- Linealidad: que exista una relación lineal entre los predictores y la variable objetivo, de forma que tenga sentido la elección de variables, por ejemplo, utilizar la altura de padres para determinar la de los hijos.
- Independencia: los registros, es decir, las observaciones, deben ser independientes entre sí. Por ejemplo, la puntuación de un videojuego no debe estar basada en la de otro.

- Homocedasticidad: que la varianza sea constante para la variable objetivo, independientemente de los predictores. De forma más simple esta condición se refiere a la constancia de la varianza.
- No colinealidad (exogeneidad): se refiere a que las variables independientes no estén relacionadas entre ellas. Si no se cumple esto se genera endogeneidad, lo que lleva a que las estimaciones de las betas sean sesgadas e inconsistentes, pese a que se obtengan valores bajos para la varianza, esto está relacionado a la conocida relación entre sesgo-varianza (por la que no por meter más variables se obtienen mejores resultados, solo modelos más sesgados).
- Normalidad: Las variables deben seguir la ley o distribución normal.

Para que se cumplan estos supuestos frecuentemente se tienen que realizar ciertas transformaciones en los datos, algunas de las más comunes son las siguientes, que se van a ejemplificar usando un modelo simple de predicción de salarios de jugadores de baloncesto elaborado para desarrollar este marco teórico. En el ejemplo siguiente la variable dependiente es salarios y las independientes cambiaran según se utilicen para explicar cada interacción:

Con variables cuya frecuencia tenga una cola hacia la derecha clara, tenga gran rango de variación o su relación con la variable objetivo no sea estrechamente lineal es recomendable utilizar una transformación logarítmica al total de la variable (Benoit, 2011). Para las variables independientes esto se interpreta como que la beta es el impacto que tendrá un aumento de 1% de la variable en la variable objetivo. Sin embargo, si el logaritmo se aplica en la variable dependiente u objetivo, la interpretación de todas las betas cambiará, por ejemplo, si x_1 tiene una beta de 0,048 se dirá que cada unidad sumada a la variable x_1 supondrá un aumento en la variable target de 4,8%. En este ejemplo se va a usar datos de puntos por partido (x_1) para predecir salarios (Y). Tanto para salarios como para puntos por partido se debe utilizar un logaritmo, como se puede observar en los gráficos, en los que el eje horizontal es el valor de la variable, y el eje vertical la frecuencia.

Ilustración 4: Gráficos de frecuencia, transformación logarítmica

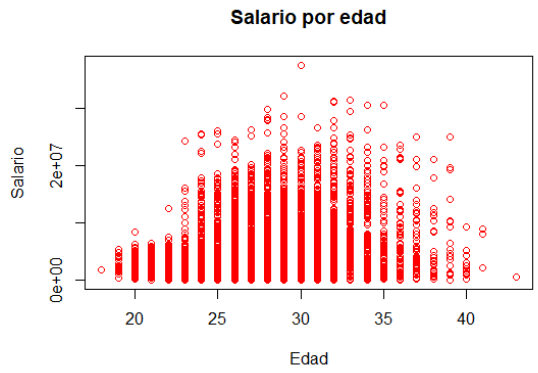


$$ln(Y) = \beta_0 + \beta_1\, ln(x_1) + u$$

Ecuación 7: Ajuste logarítmico a variables en una regresión lineal

A la hora de tratar con variables X_i que tengan relaciones o efectos cuadráticos con la Y o variable objetivo se tendrá que incorporar sumando un X_i^2 al modelo. De esta manera se transforma la relación cuadrática a una similar a una lineal y también permite calcular cuando se alcanza el máximo de la variable con relación cuadrática a la variable objetivo (Siemsen, Roth, & Oliveira, 2010). Siguiendo el ejemplo descrito: se van a incorporar la edad de los jugadores (x_2) al modelo predictivo del salario:

Ilustración 5: Gráfico de frecuencia, efectos cuadráticos



$$\ln(Y) = \beta_0 + \beta_1 \ln(x_1) + \beta_2 x_2 + \beta_3 x_2^2 + u$$

Ecuación 8: Ajuste cuadrático a una variable en una regresión lineal

De esta forma se puede calcular la edad en la que los jugadores tienen el salario más alto como

$$Max = \frac{-\beta_2}{2 \times \beta_3}$$

Ecuación 9: Fórmula para hallar el máximo en efectos cuadráticos en regresión lineal

La última transformación que se va a observar es la de efectos de interacción. Esto se refiere a aquellos casos en los que el efecto de una variable X_1 sobre una variable Y cambie, al incorporar una variable X_2, que normalmente suele ser dicotómica. El efecto marginal de la variable X_1 pasa a ser β_1 si X_2=0 o $\beta_1+\beta_2$ si X_2=1 (Siemsen, Roth, & Oliveira, 2010).

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 x_2) + u$$

Ecuación 10: Ajuste efectos de interacción en una regresión lineal

También cabe destacar el proceso de One-Hot Encoding que se utiliza para poder incorporar datos categóricos al modelo de regresión, consiste en establecer una categoría como base, por lo que lo que significará la beta será el impacto de que se forme parte de una categoría frente a la base. Siguiendo el ejemplo de los salarios de la NBA, si se estudiase el impacto que tienen los distintos roles que se pueden jugar en el salario de los jugadores, y se fijase la categoría base en "Point Guard", la beta que resultaría en la regresión para la posición de "Center" sería el impacto en el salario por jugar de "Center" y no de "Point Guard". Otra posible solución a esto es convertir las variables categóricas en numéricas dicotómicas, de forma que en vez de tener una variable de "Position" en la que hay 5 categorías: "Point Guard", "Shooting Guard" … tendrías 5 variables que corresponderían a las categorías y serían igual a 1 si el jugador juega en esa posición o 0 si no.

Una vez testados estos supuestos mediante diferentes métodos, como por ejemplo en el caso de la linealidad el graficar los datos para observar si existe una relación lineal entre las variables, los modelos resultantes generan una serie de estadísticos que hay que saber interpretar para entender la validez del modelo y de los resultados obtenidos. Los estadísticos clave a la hora de estudiar la validez de un modelo son:

El Error estándar: mide la precisión de la estimación de la beta, es decir, cuanto podría habría cambiado con otros datos, cuanto menor su valor respecto al valor de la beta mejor.

El T-valor: es el cociente entre el valor de la beta y error estándar y cuanto mayor, más precisa la estimación.

El P-valor: testa la hipótesis nula de que la beta sea igual a 0 frente a que sea distinto de esto, cuanto menor sea su valor mejor, ya que representará que habrá menos probabilidad de que los resultados sean consecuencia de la casualidad.

R cuadrado (R-squared): refleja el porcentaje del objetivo que es explicado por las variables independientes incluidas. Sin embargo, esta medida es poco precisa ya que cuantas más variables incluyas más grande será, pero para esto se tiene el R cuadrado corregido (adjusted R-squared) que penaliza el exceso de variables irrelevantes.

3. DEFINICIÓN Y ALCANCE DEL PROYECTO

A. Objetivos

Para un proyecto como este es vital sentar una serie de objetivos alrededor de los cuales poder diseñar el estudio, ya que estos supondrán el fin último del proyecto. Estos son:

1. El objetivo fundamental de este proyecto es el de estudiar sí y cómo las críticas de videojuegos afectan a las ventas de estos en España, de forma que sirva para ilustrar la importancia relativa de cada medio de críticas profesional en este país y el impacto que tienen en la industria. Cabe destacar que únicamente se van a observar juegos AAA o juegos que cuando se lanzaron pudieron ser considerados Indie, pero que ahora tienen un tamaño considerable de público, ya que estos tienen más información disponible sobre los mismos.
2. El ámbito principal del proyecto es el de la econometría, por lo que se realizará una revisión de literatura básica de dicho campo. A esto se sumarán, al tratarse de un trabajo de fin de grado de Administración de Empresas, un repaso histórico de la industria de los videojuegos, enfocado a la parte estratégica del rol de las variables que se van a estudiar, como se ha visto en un primer lugar en el **contexto histórico**.
3. Para estudiar estos comportamientos también se van a usar otras variables para explicar la variable objetivo y realizar otros estudios econométricos, como son el año de lanzamiento, el sistema principal donde se juega cada videojuego, si lo ha jugado un influyente (del inglés influencer) de masas, si lo ha jugado un influyente de nicho, la presencia en redes sociales de la editorial, etc.
4. En base a los resultados obtenidos de la regresión explicativa y de los distintos análisis econométricos, se extraerán conclusiones que puedan ser de utilidad tanto para publishers como para desarrolladores de videojuegos para poder entender cómo funciona la industria y que distintas variables son clave a la hora de determinar las ventas de un título.

B. Hipótesis

De cara a contrastar los resultados finales y comprobar la consecución de los objetivos, es necesario explicar lo que se espera que pueda suceder en el estudio. Esto son las hipótesis, que son:

1. Los juegos con mejores puntuaciones tendrán mayores ventas. Es decir, hay una correlación positiva entre la puntuación de la crítica y las ventas.
2. Los juegos con mayor número de seguidores en redes sociales tendrán más ventas. Es decir, hay una correlación positiva entre la presencia en redes sociales y las ventas.
3. Los juegos que hayan sido cubiertos por influyentes tendrán más ventas, sobre todo si los han cubierto creadores de contenido de masas.
4. Si un juego es del género "acción-aventura", "FPS" o "deporte" tendrá más ventas. Por conocimiento del sector seguramente estos sean los géneros que más vendan.
5. En media los juegos de Nintendo reciben mejores puntuaciones en general que los juegos de otras consolas. Nintendo es conocido por su propiedad intelectual destacada y selecta, por lo que, aunque no cuentan con tantos juegos en sus plataformas, los juegos propios de estas suelen ser muy bien recibidos.
6. En media, los medios nacionales tendrán más repercusión en las ventas (explicarán más variabilidad o tendrán betas más altas) que los medios internacionales.
7. Las puntuaciones de medios de críticas profesionales tendrán más impacto que las de medios públicos (foros).
8. Los juegos más recientes tendrán más ventas, puesto que principalmente se va a observar la franja de 2019-2022 (pero se han incluido títulos todavía relevantes como el Grand Theft Auto V que salió en 2013).
9. Los medios nacionales dan puntuaciones más altas que los internacionales.

C. Asunciones

Al igual que es relevante conocer lo que se espera que suceda, es importante constatar previo al estudio la información que se da por sentado para el diseño de este. Este es un paso clave para el desarrollo de los modelos y la elección de asunciones erróneas puede llevar al éxito o fracaso del proyecto. Las asunciones para este estudio son:

1. Con las puntuaciones de 5 medios de críticas profesionales (Hobby Consolas, Metacritic, Vandal, IGN y MeriStation) y 3 de medios de foros (3D Juegos, Comunidad de Vandal y Comunidad de Metacritic) se podrá explicar el impacto que tienen estos agentes en la industria y las interacciones entre medios de críticas y medios de comunidades o foros, así como comparar a medios nacionales versus internacionales.
2. Pese a la enorme cantidad de creadores de contenido en esta industria, el ver el impacto de un influyente de masas (Vegetta777) y uno de nicho como (Alexelcapo) debería servir para incorporar la parte de la variabilidad de las ventas que depende de esta parte del sector. Esto se debe a que multitud de creadores de contenido juegan a los juegos más populares, por lo que, al meter de forma simple el embudo de clientes potenciales de la industria, desde los más corrientes con el influyente de masas, como a los más conocedores del género con el influyente nicho, se está recogiendo una gran cantidad de información de forma simplificada.
3. Puesto que los datos de ventas con los que se cuenta son actuales (2019-agosto 2022), la mayoría de los videojuegos a observar deben de ser de esta franja temporal, a excepción de ciertos títulos que todavía sean relevantes y traigan muchas ventas.
4. Para realizar este estudio econométrico bastará con una muestra limitada, priorizando la calidad de las variables, y que se adecúe al estudio (como se ha mencionado en la asunción anterior) que el número de registros. Por esto se van a analizar 120 juegos, habiendo identificado otros 300 posibles que se puedan usar en estudios futuros.
5. El estudio tendrá que observar tanto juegos que han obtenido grandes ventas, como aquellos que, aun habiendo recibido cierta atención, vendiesen poco. Esto es para poder estudiar tanto casos de éxito como fracasos, permitiendo ver el impacto de cada variable en esto.

D. Restricciones

Finalmente, también se debe resaltar la importancia de los hechos que limitan al estudio que se va a desarrollar, lo cual puede tener relación con multitud de factores y afectar a los resultados de este. Se han identificado las siguientes restricciones:

1. Pese a la posibilidad de llevar a cabo un análisis más exhaustivo, con modelos más complejos y que observe más registros (más videojuegos), debido a las horas de trabajo límite y la extensión que debe tener este Trabajo de Fin de Grado, se ha optado por mantener este proyecto conciso pero completo, de forma que se cumplan todos los objetivos y se comprueben las hipótesis sin exceder dichas restricciones, pero dejando lugar para un estudio más profundo en el futuro.
2. Dado que este trabajo es del campo de la econometría, aunque desde un principio se planteó hacer en el lenguaje de programación R, esto suponía una complejidad innecesaria, dado que se puede obtener el mismo resultado usando el programa GRETL, dejando a R un rol de apoyo a los modelos desarrollados.
3. Puesto que es mucho más difícil obtener información de juegos con comunidades más pequeñas como son los llamados “indies” (que viene de desarrolladores independientes, esta diferencia se explicará en el siguiente apartado), se limitará a observar en su mayoría a juegos AAA, es decir, juegos que son publicados por grandes estudios y que cuentan con soporte en las principales plataformas. Por esto, pese a que se extrapolarán las conclusiones a toda la industria, realmente existe un sector (que, aunque mucho más pequeño) no se está estudiando debido a la falta de información necesaria.

4. ANÁLISIS EXPLORATORIO-DESCRIPTIVO

A. Origen de los datos

El primer paso a la hora de construir el conjunto de datos será seleccionar los títulos con los cuales se realizará este estudio. Es clave obtener juegos para los que se puedan encontrar datos para todas las variables que se quieran estudiar, por lo que tienen que ser juegos suficientemente conocidos. Para entender esto es importante explicar la división actual en el mercado de los videojuegos, ya que, pese a que el lado más conocido de la industria actual es el de títulos de enormes dimensiones como "Call of Duty" o "FIFA", existe otra parte de esta completamente distinta, dedicada a juegos pequeños, de baja inversión y géneros más imaginativos; los juegos indie (Keogh, 2015).

De aquí la necesidad de que casi todos los juegos que se observen sean AAA, es decir, títulos que son publicados por estudios de gran tamaño como pueden ser Electronic Arts o Nintendo y que tengan una gran inversión detrás, puesto que esto supondrá que habrá más información disponible. Sin embargo, este tipo de condiciones pueden ser restrictivas a la hora de analizar cuando y porqué fracasa un videojuego en cuanto a ventas, ya que es de esperar que el hecho de ser un juego grande traiga consigo ventas altas, por lo que será útil analizar también casos de videojuegos con ventas bajas. Estos principios básicos acabaron en la selección de alrededor de 130 títulos, que, ante la falta de información en ciertas variables para ciertos juegos, acabaron en un total de 120.

Una vez seleccionados los juegos y, con una reserva de otros 300 juegos que podrían ser utilizados para el estudio se procede a la selección de variables. Estas variables, que en un comienzo iban a ser únicamente las puntuaciones en medios de críticas han sido seleccionadas siguiendo dos principios: las necesidades del estudio (y la lógica de este) y el funcionamiento de la industria de los videojuegos.

Para explicar el primer principio se puede usar la variable del año de lanzamiento, ya que, aunque se hayan priorizado juegos modernos, pertenecientes a los últimos 4 años, también se han incluido juegos exitosos de años tan tardíos como 2013, como es el caso del "Grand Theft Auto V", debido a su todavía sentida prevalencia en el mercado. Para explicar las ventas de este en un modelo como una regresión frente a las ventas de un juego más moderno es fundamental contar con una variable que incluya el efecto temporal, como es el caso de los años. El segundo principio tiene que ver con cómo funciona la industria de los videojuegos en España en la actualidad, o, cuál es la respuesta a la siguiente pregunta: ¿Qué vende videojuegos en nuestro país? Como se busca estudiar en este trabajo, se prevé que las críticas tendrán algún efecto, pero ¿Qué otras variables pueden ayudar a explicar la variabilidad de las ventas de videojuegos? Esto se verá en el apartado de análisis resumido en la estadística del R^2, pero ahora solo se puede especular sobre cuales podrían tener sentido dentro del estudio.

- **Medición del impacto de influyentes**: en el ámbito de los videojuegos y del entretenimiento hay un número selecto de perfiles que destacan sobre el resto. Desde que surgieron se han ido monetizando, convirtiéndose en competencia para muchos sitios de críticas, siendo comparable el impacto que tienen en las ventas de videojuegos. La diferencia principal que se observa entre estos y las críticas tradicionales es la forma mucho más interactiva que tienen de tratar con las audiencias sacrificando profesionalidad y normalmente falta de formación periodística, a cambio de llegar a públicos de masas. Por este tipo de razones, su impacto en la industria en los últimos años parece indudable (Adigüzel, 2021), por lo que se va a comprobar usando a un Youtuber de masas que juegue una gran cantidad de títulos. Esto ha llevado al estudio a Vegeta777, que con 33,6M de subscriptores es el tercer máximo exponente de la plataforma dentro de nuestro país. Sin embargo, también es de interés para el análisis utilizar a otro creador de contenido que sea más de nicho, que cubra una gran cantidad de títulos pero que también sea relevante a nivel nacional, esto lleva al segundo influyente que se incorporará al análisis; Alexelcapo, con casi 2M de subscriptores.
- **Diferencia entre los géneros:** Con una vistazo breve de la industria de los videojuegos se puede observar que hay ciertos géneros que venden más que otros de forma casi independiente al gasto de marketing o inversión que haya supuesto cada juego, en España especialmente hay una gran predilección por la saga de videojuegos FIFA, de EA Sports, ya que, según datos de la Asociación Española del Videojuego, estos títulos ocupan 3 de los top 5 juegos más vendidos de 2019 a 2022. Por lo general un juego puede parecer que tiene más probabilidades de triunfar en cuanto a ventas si se trata de un juego del género de deporte o de acción que si se trata de un género más selecto, como son los juegos de conducir o de baile, por lo que incluir estos géneros es relevante para nuestro proyecto. Los géneros incluidos serán:

Tabla 1: Géneros de videojuegos

GÉNERO	**DEFINICIÓN**
Acción Aventura	Son juegos que como su nombre indica, incluyen partes de acción tanto como de aventura para poner al jugador en el centro de una historia. Estos juegos suelen ser los más similares a otros medios de entretenimiento como películas, al incluir grandes segmentos de acción con un gran componente narrativo. Ejemplos de esto son las saga The Last of Us o God of War.
Beat-Em'-Up	"Juegos de pelea a progresión", el jugador deberá combatir con un gran número de individuos para ir avanzando en el nivel. Era un género muy popular en la época de los salones recreativos. Un ejemplo moderno es el recién salido "Sifu".
Dancing	Juegos de ritmo que utilizan soportes externos como cámaras o plataformas para procesar los movimientos del usuario y compararlos con la versión "correcta" de la secuencia que se intenta imitar. El ejemplo más significativo es la saga Just Dance.

Fighting	Juegos de lucha, en los que el jugador debe combatir contra otro personaje manejado por otro jugador o por la consola, normalmente la cámara está posicionada de forma que el jugador ve ambos combatientes de lado. Algunos ejemplos son la saga Street Fighter, Tekken o Super Smash Bros.
FPS	First-Person-Shooter o juegos de disparar en primera persona, en este género cuando se controla al personaje, las dos acciones básicas son las de moverse y disparar. Los gráficos normalmente son en 3D y algunos ejemplos clásicos son "Doom" o "Quake", en cuanto a modernos destaca la saga Call of Duty.
Miscellaneous/ Party Game	Son juegos que simulan el funcionamiento de juegos de mesa introduciendo mecánicas típicas de los videojuegos. Están caracterizados por el hecho de que incluyen un enorme número de minijuegos (lo cual hacen que sean misceláneos, ya que estos son de otros géneros muy diversos) y están diseñados para jugar en multijugador local; con un grupo de personas utilizando el mismo juego y consola. Un ejemplo clásico es la saga Mario Party.
Platformer	En este género el jugador deberá controlar un personaje que irá saltando obstáculos para llegar a un objetivo, se llaman juegos de plataformas por el hecho de que la complejidad del diseño de los niveles suele ser la construcción de plataformas sobre las que los personajes se suban para afrontar los obstáculos. El ejemplo clásico es la saga de Super Mario Bros.
Racing	En español juegos de carreras, normalmente sobre un vehículo el jugador debe llegar a un punto en una serie de vueltas antes que sus rivales, se ven juegos que pretenden ser simuladores más exactos como la saga Gran Turismo o sagas que tratan más los componentes recreativos como Mario Kart
Rogue-like	Se trata de un subgénero de juegos de acción aventura cuya mecánica principal es la de volver a empezar desde el comienzo del juego en el momento de la derrota del jugador, manteniendo ciertas mejoras y ventajas que este haya obtenido antes de ser vencido. Se trata de un género muy popular actualmente con juegos como "Hades" o "Returnal".
RPG	Role-Playing-Game o juego de rol es aquel en el que el jugador controla a un personaje y queda inmerso en el mundo de este; el jugador asume la identidad del personaje, ya sea predefinida o mediante la personalización de este. Estos juegos suelen imitar ciertos aspectos de la vida real como las profesiones, el sistema monetario o incluso relaciones amorosas. Algunos ejemplos son "Witcher 3", la saga Pokémon o "The Elder Scrolls V: Skyrim"
Simulation	Consiste en la simulación de la vida cotidiana incorporando mecánicas para hacer más interesante el juego, algunos ejemplos de esto son la saga Animal Crossing o los Sims.
Sports	El gameplay simula acción de deportes reales, algunos ejemplos son la saga FIFA o NBA2K, en las que el jugador participa en el caso del primero en partidos de fútbol o en el segundo de baloncesto en la NBA. También existen casos de deportes a los que se les añaden elementos típicos de los videojuegos como es el caso de "Mario Strikers" (fútbol + Mario y elementos arcade).
Strategy	Normalmente basado en el ámbito militar o de planificación de una civilización, el jugador debe, mediante planificación, vencer a un rival que puede ser o el propio videojuego u otros jugadores, logrando construir una estrategia superior a estos aplicada al caso del juego. El ejemplo clásico en este género es la saga de Age Of Empires.
El contenido de esta tabla proviene de (Belli & Raventós, 2008) y (Wirtz, 2023)	

- **Críticas de profesionales contrastadas con las de comunidades de usuarios:** existen ciertos medios de críticas que también incluyen opiniones de sus usuarios, e incluso alguna página que únicamente funciona a través de dichas opiniones. La lógica dice que dichas reseñas deben ser una palanca relevante a la hora de determinar el éxito del lanzamiento de un juego, pero existen estudios que comparando el "buzz" o conversación y anticipación previa a dichos lanzamientos de videojuegos contra las críticas profesionales que se decanta por los segundos como mayor determinante de las ventas (Cox & Kaimann, 2015). Por esto interesa contrastar ambas variables, realizando dicha comparación para aquellos medios lo suficientemente grandes que cuentan tanto con críticas profesionales como con reseñas de usuarios como son Vandal o Metacritic, o en el caso de 3D Juegos (que es únicamente un foro de usuarios donde se puntúan títulos) se buscará compararla con las páginas de críticas puras.
- **Seguimiento e impacto en redes sociales**: otro determinante del éxito que puede tener un videojuego, sobre todo en la actualidad, es el impacto que tienen estos contenidos en redes sociales. Este impacto se puede medir de diferentes maneras, pero debido a que la complejidad de recogida de datos en este trabajo ya es significativa, el estudio se va a limitar a comprobar dos redes sociales que las marcas usan tradicionalmente para emitir comunicados: Instagram y Twitter. Una vez decidido esto se tendrá que escoger que variable recoger, y para otra vez, que esta variable sea incorporable de forma simple y relativamente sencilla, se va a observar el número de seguidores que tenga el videojuego en su cuenta propia en cada una de estas dos redes sociales. Si no tiene cuenta propia se recurrirá a la cuenta del desarrollador o Publisher, escogiendo la que más seguidores tenga, siempre que sea demostrable que se ha realizado por lo menos un post en relación con ese juego.
- **Soporte principal del título:** la hipótesis en esta variable es que dependiendo de la consola principal en la que se lance el videojuego, este tendrá más éxito o menos. Esto se debe a que no hay el mismo número de las tres consolas principales en nuestro país. Un estudio de Game Reactor del 27 de enero de 2022 señalaba que la Nintendo Switch es la consola más vendida en España con más de 2 millones de copias, mientras que la PlayStation 5 se sitúa segunda con un total de 363 mil unidades y la Xbox Series apenas llega a las 110 mil unidades. Conociendo este tipo de números es de esperar que un juego cuya consola principal es la Xbox Series X espere vender menos que un juego de PS5, ya que su mercado desde el principio es mucho menor.

Finalmente, los datos de ventas son fundamentales para el proyecto, ya que, aunque se tenga información sobre todas estas variables independientes, se va a necesitar una variable dependiente que explicar. Este tipo de datos suelen ser de gran interés para las empresas de la industria por lo que no suelen ser públicos, por lo menos de manera directa. Tras investigar extensamente en repositorios de datos como Kaggle o Google Data Set Search, parecía muy complicado encontrar

datos desagregados (ventas individuales de videojuegos) y encontrar cifras concretamente de España, ya que en este tipo de repositorios anglosajones los números suelen ser o macro o de EE. UU. No obstante, continuando esta búsqueda, se encontró la página web y los datos de contacto de la Asociación Española del Videojuego (AEVI), que contaba con rankings semanales de los diez juegos más vendidos. Siguiendo la lógica de que si tenían estos datos semanales debían tener también los datos totales se estableció contacto pidiendo, con fines académicos, tener este data set de venta. Con los datos de ventas de los casi 5600 juegos vendidos en España desde 2019 hasta finales de 2022 se indexó para encontrar y anexar las cifras de ventas de los 120 títulos seleccionados, utilizando el título del videojuego como ID o identificador. Esto también otorga una enorme base de datos que permitiría aumentar el dataset si se buscase aumentar el tamaño del proyecto con más juegos, pese a que esto también requeriría buscar mucha más información sobre las variables.

En la tabla de variables actual hay 120 títulos por 16 variables cuya información proviene de búsqueda en la web, un total de 1920 búsquedas e inputs de información.

B. Tabla de variables

Habiendo repasado todas las variables que se buscarán incluir en el estudio y el porqué, la estructura de los datos será la siguiente:

Tabla 2: Tabla de variables de la base de datos

NOMBRE	EXPLICACIÓN	CÓMO SE MIDE	TIPO
Title	Nombre del videojuego, se usará esta variable como ID		Carácter
Release	Año en el que salió el videojuego	Desde el 2013 hasta el 2022, siendo mucho más abundantes los juegos recientes	Numérica Discreta Intervalo
Genre	Género al que pertenece el videojuego (que tipo es)	Se obtendrá el género como esté marcado en la mayoría de las páginas de críticas	Carácter Categórica
Hobby Consolas	Puntuación otorgada por Hobby Consolas	En porcentaje, del 0% al 100%	Numérica Continua Intervalo
MeriStation	Puntuación otorgada por MeriStation	En porcentaje, del 0% al 100%	Numérica Continua Intervalo
3D Juegos	Puntuación otorgada por 3D Juegos	En porcentaje, del 0% al 100%	Numérica Continua Intervalo
Vandal	Puntuación otorgada por Vandal	En porcentaje, del 0% al 100%	Numérica Continua Intervalo
Vandal_Comm	Puntuación otorgada por la comunidad de usuarios de Vandal	En porcentaje, del 0% al 100%	Numérica Continua Intervalo
Metacritic	Puntuación otorgada por Metacritic	En porcentaje, del 0% al 100%	Numérica Continua Intervalo
Metacritic_Comm	Puntuación otorgada por la comunidad de usuarios de Metacritic	En porcentaje, del 0% al 100%	Numérica Continua Intervalo
IGN	Puntuación otorgada por IGN	En porcentaje, del 0% al 100%	Numérica Discreta

			Intervalo
Instagram	Número de seguidores en la cuenta de Instagram del videojuego, o a falta de, del desarrollador o publisher	Número total de seguidores a noviembre de 2022	Numérica Discreta Razón
Twitter	Número de seguidores en la cuenta de Instagram del videojuego, o a falta de, del desarrollador o publisher	Número total de seguidores a noviembre de 2022	Numérica Discreta Razón
Alexelcapo	Si el juego ha sido o no mostrado en el canal de YouTube del influyente Alexelcapo	Se marca 1 si sí ha subido un video sobre el juego o 0 si no	Numérica Categórica (binaria) Discreta
Vegeta777	Si el juego ha sido o no mostrado en el canal de YouTube del influyente Vegeta777	Se marca 1 si sí ha subido un video sobre el juego o 0 si no	Numérica Categórica (binaria) Discreta
Sales	Ventas totales del videojuego de 2019 a agosto 2022 (variable objetivo o dependiente)	Número total de ingresos por ventas que ha generado el videojuego	Numérica Continua Razón
Main Platform	Soporte principal del juego	Nombre de la consola o soporte	Carácter Categórica

C. Metodología

1. El objetivo fundamental del proyecto se pretende abordar mediante la consecución de los subobjetivos, ya que para poder explicar el rol de las críticas profesionales en el mercado se ha debido construir un modelo siguiendo todos los pasos que estos asientan.

   En este apartado se irá abordando como se conseguirán cada uno de los objetivos:

2. Para la consecución del primer subobjetivo se deberá revisar la literatura tanto de econometría como de la historia y estrategia empresarial alrededor de la industria de los videojuegos, acudiendo a fuentes como las siguientes: Scopus, Web of Science, Google Scholar o Dialnet. Las palabras clave que se usarán serán: videojuegos, regresión, econometría, publishers, desarrolladores, ventas. Lo que se buscará con esto es o bien encontrar casos de estudio similares en los que se hayan estudiado relaciones similares a las que se buscan analizar, o herramientas econométricas que puedan ayudar a realizar el estudio de las variables, además de estudiar la industria de los videojuegos de forma histórica para entender mejor el funcionamiento de esta.
3. Para conseguir el segundo subobjetivo:
   - Descripción del caso: Explicación de la selección de juegos, medios y variables hechas y que sentido tienen a través de una visión estratégica de la industria.
   - Desarrollo y tratamiento del data set: Ya seleccionadas las variables, se construirá un data set que incluya una cantidad suficiente de los juegos seleccionados con valores en todas las variables obtenidas. Para simplificar, todas las variables que funcionen en escalas similares se pasarán a la misma escala (p. ej. Las puntuaciones de críticas de diferentes sitios todas estarán en la misma escala para comparación más fácil).
   - Análisis de los datos mediante econometría y herramientas predictivas simples (usadas de forma explicativa).
4. Para alcanzar tercer subobjetivo: Extrapolar los resultados obtenidos y su significación para el mercado de los videojuegos en España.

5. ANÁLISIS Y RESULTADOS

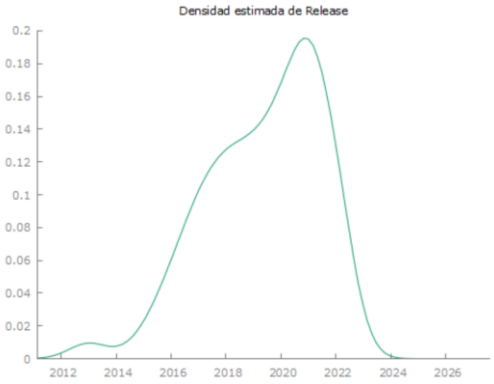
Como se ha explicado en el marco teórico, el análisis se va a realizar en dos grados: un primer estudio de las variables en el que se observen datos como la curtosis, la asimetría, la densidad o la correlación y otro segundo nivel de análisis en el que, usando los comportamientos aprendidos en el primero, se diseñará un modelo que sea capaz de testar las hipótesis formuladas en este proyecto. Cabe destacar que el centro del estudio tiene que ver con la interpretación y con el desarrollo del caso de estudio, sumado a la precisión científica del análisis, por lo que es posible que se obtengan resultados que no sean tan significativos como serían si se contase con más datos, aunque esto se verá según se progrese en este apartado. Así pues, para proceder con el análisis se van a utilizar 2 herramientas fundamentales, GRETL, un software econométrico diseñado para construir modelos con los que comprobar las hipótesis y realizar gráficos sintéticos y rápidos de generar, y R Studio, este segundo se va a usar de manera más limitada para desarrollar gráficos más complejos que ayuden a entender las variables y su comportamiento entre sí.

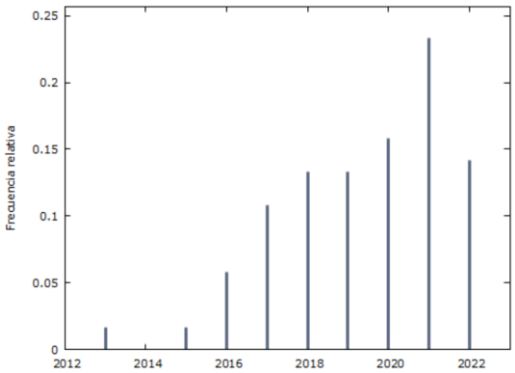
A. Análisis univariante

Se va a comenzar con un estudio de las variables de forma individual. Como se describía en la parte de la elaboración de la base de datos, se cuenta con 120 registros, es decir se observan 120 juegos diferentes de los cuales se han registrado datos para 17 variables distintas. De estas variables todas excepto la variable **"Title"**, que funciona como una variable ID o identificadora, cuentan con información relevante para el análisis de la cual se pueden extraer conclusiones para este estudio. Así se procede a realizar los gráficos de densidad estimada y en los casos pertinentes comentar la distribución de frecuencias:

Variable **"Release"**, o el año en el que salió a la venta el videojuego:

Ilustración 6: Gráfico de Densidad de 'Release' e Ilustración 7: Gráfico de frecuencias relativas de la variable 'Release'





Esta variable cuenta con una notable asimetría hacia la izquierda, con un cociente de asimetría de -0,69, siguiendo una distribución ligeramente leptocúrtica, ya que, como se ve en las dsitribuciones, la mayor parte de las observaciones están concentradas en el rango de 2019-2022 (un 66,67%). La media o esperanza de esta variable es 2019,4 con una desviación típica de 2,0571.

Variable **"Genre"**, o el género al que pertenece el videojuego, (el tipo de videojuego):

Ilustración 8: Gráfico de frecuencias relativas de 'Genre'

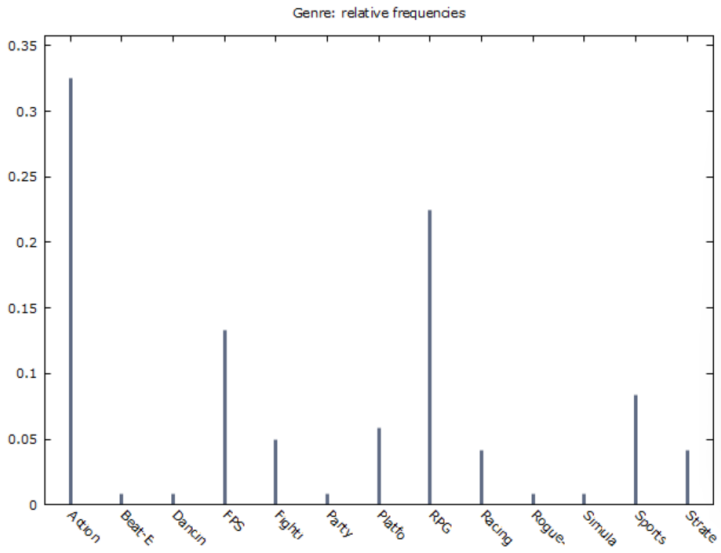


Tabla 3: Tabla de frecuencias de géneros en la base de datos

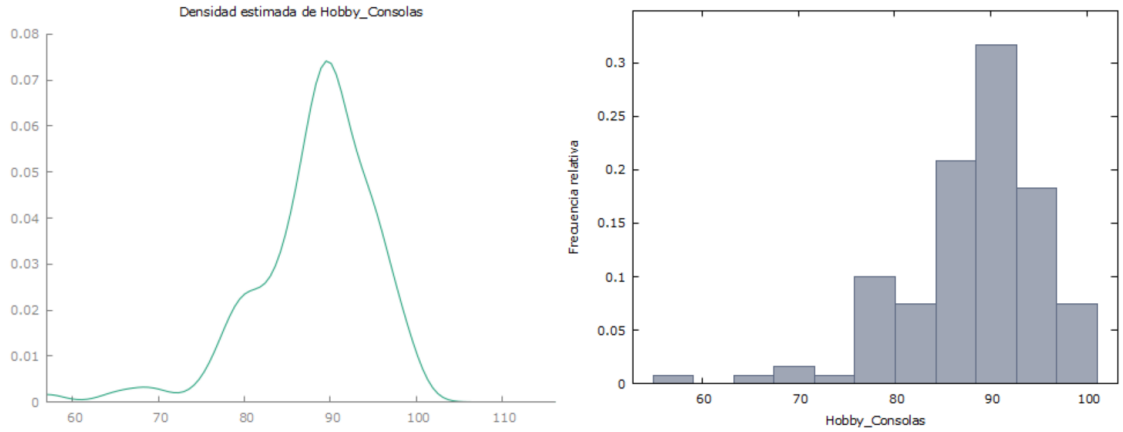
Género	Total	Porcentaje de frecuencia
Action Adventure	39	32,5%
Beat-Em'-Up	1	0,83%
Dancing	1	0,83%
FPS	16	13,33%
Fighting	6	5%
Party Game	1	0,83%
Platformer	7	5,83%
RPG	27	22,5%
Racing	5	4,17%
Rogue-Like	1	0,83%
Simulation	1	0,83%
Sports	10	8,33%
Strategy	5	4,17%

Al tratarse de una variable categórica no se va a realizar un gráfico de densidad, es decir, la información está en formato de texto y los juegos pueden pertenecer a uno de los 13 géneros de juego observados en el estudio. El registro más común es el de Action Adventure con un 32,5% de las observaciones, esto era de esperar puesto que es uno de los géneros más populares en la industria y también sirve en ocasiones como categoría principal en la que se engloban múltiples diferentes subgéneros. También es relevante que hay muchos géneros como Rogue-Like, Simulation o Party Game entre otros que tienen una frecuencia mínima con solo un 0,83%, es decir, de los 120 juegos observados solo 1 pertenece a este género. Este tipo de comportamientos singulares tiende a aumentar los errores al elaborar métodos predictivos por lo que es posible que esta variable no sea de gran utilidad a la hora de realizar el modelo, aunque queda estudiar como se comportará con la variable objetivo.

Las variables de puntuaciones como se explicaba en la tabla oscilan en un rango del 0% al 100%, pero para interpretación estadística a la hora de realizar el modelo estadístico han sido convertidas del 0 al 100 (es decir, se ha eliminado el porcentaje).

Variable **“Hobby_Consolas”**, o la puntuación que le dio este medio crítico a los juegos observados:

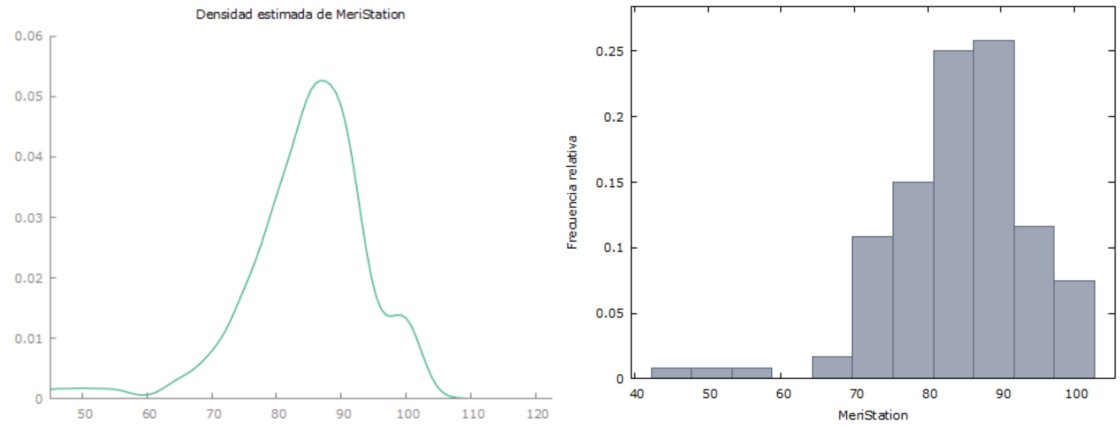
Ilustración 9: Gráfico de Densidad de 'Hobby_Consolas' e Ilustración 10: Gráfico de frecuencias relativas de 'Hobby_Consolas'



Esta variable cuenta con una muy alta asimetría hacia la izquierda, con un cociente de asimetría de -1,4, siguiendo una distribución muy leptocúrtica, con la mayor parte de las observaciones concentradas en el rango de puntuaciones superiores a 80 (un 85%). La media o esperanza de esta variable es 88,2 con una desviación típica de 6,98.

Variable **“MeriStation”**, o la puntuación que le dio este medio crítico a los juegos observados:

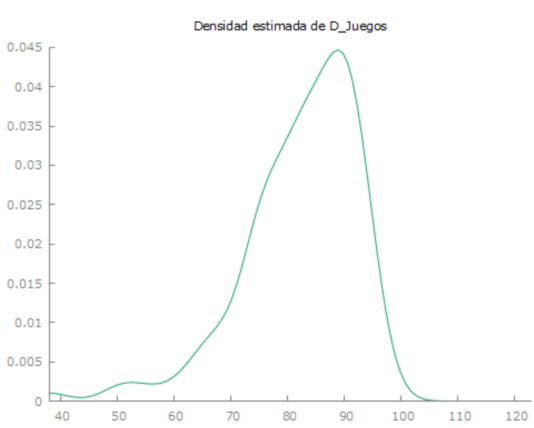
Ilustración 11: Gráfico de Densidad de 'MeriStation' e Ilustración 12: Gráfico de frecuencias relativas de 'MeriStation'

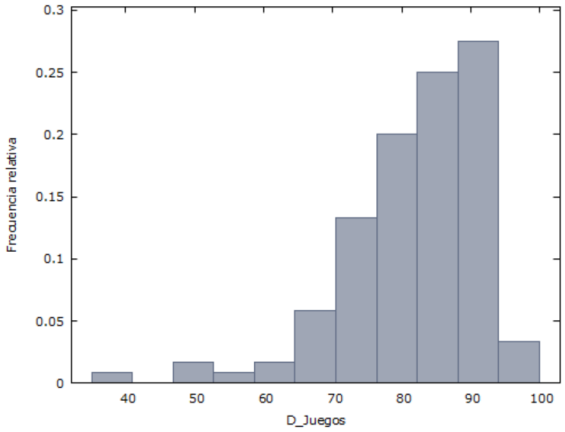


Esta variable cuenta con una alta asimetría hacia la izquierda, con un cociente de -1,21, siguiendo una distribución muy leptocúrtica, con la mayor parte de las observaciones concentradas en el rango de puntuaciones superiores a 80 (un 70%). La media o esperanza de esta variable es 84,7 con una desviación típica de 9,44. En esta variable si que se ven puntuaciones inferiores al 60, que ocurre en 3 casos con un mínimo de 45.

Variable **“D_Juegos”**, o la puntuación que le dio la comunidad de este medio (3D Juegos) a los juegos observados:

Ilustración 13: Gráfico de Densidad de 'D_Juegos' e Ilustración 14: Gráfico de frecuencias relativas de 'D_Juegos'

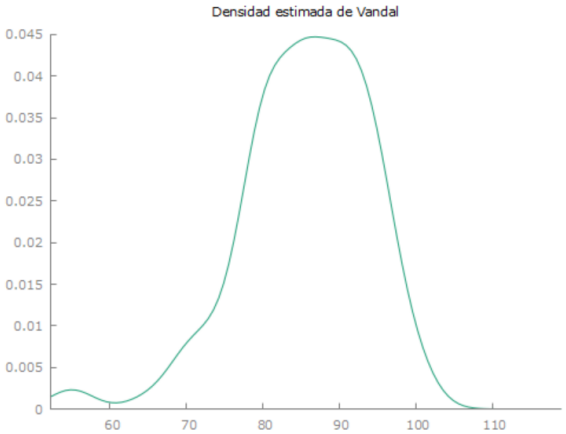


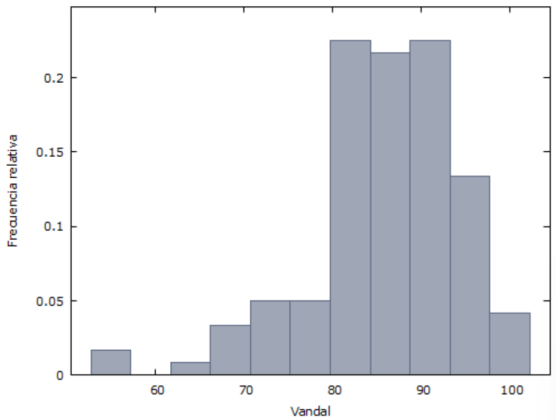


Esta variable cuenta con una muy alta asimetría hacia la izquierda, con un cociente de -1,42, siguiendo una distribución muy leptocúrtica, con la mayor parte de las observaciones concentradas en el rango de puntuaciones superiores a 80 (alrededor del 65%), algo menos que en las anteriores. La media o esperanza de esta variable es 82,2 con una desviación típica de 10,16. En esta variable también se ven puntuaciones inferiores al 60, con un mínimo de 38. También se ve que hay más frecuencia relativa de puntuaciones superiores al 90 pero también hay menos puntuaciones entre el 80 y el 90, es decir, las críticas parecen ser más extremas que en las variables anteriores.

Variable **“Vandal”**, o la puntuación que le dio este medio crítico a los juegos observados:

Ilustración 15: Gráfico de Densidad de 'Vandal' e Ilustración 16: Gráfico de frecuencias relativas de 'Vandal'



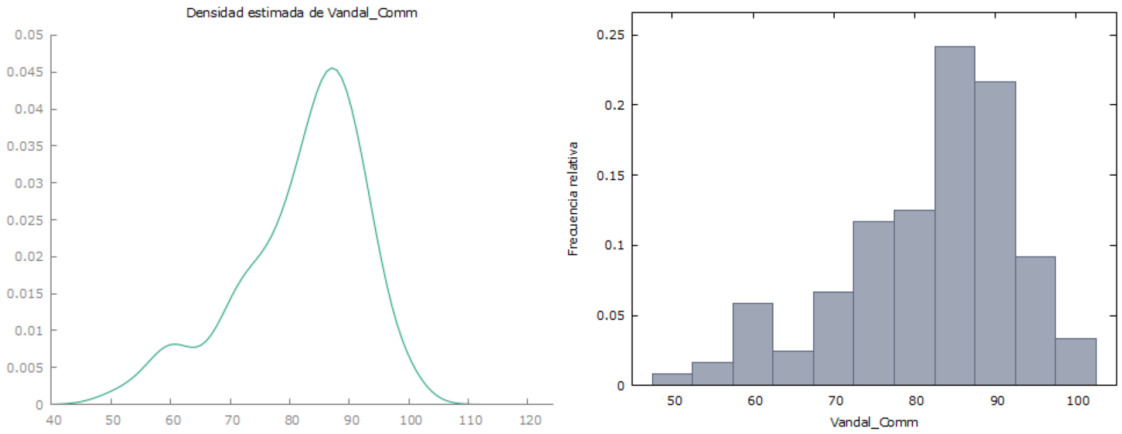


Como parece que va a ser lo normal analizando las puntuaciones de los medios de críticas, la variable tiene una distribución con alta asimetría hacia la izquierda, con un cociente de –0,93,

siguiendo muy leptocúrtica, con la mayor parte de las observaciones concentradas en el rango de puntuaciones superiores a 80 (menos del 85%), algo menos que en las anteriores. La media o esperanza de esta variable es 85,575 con una desviación típica de 8,34. El mínimo es de 55.

Variable **"Vandal_Comm"**, o la puntuación que le dio la comunidad de este medio (Vandal) a los juegos observados:

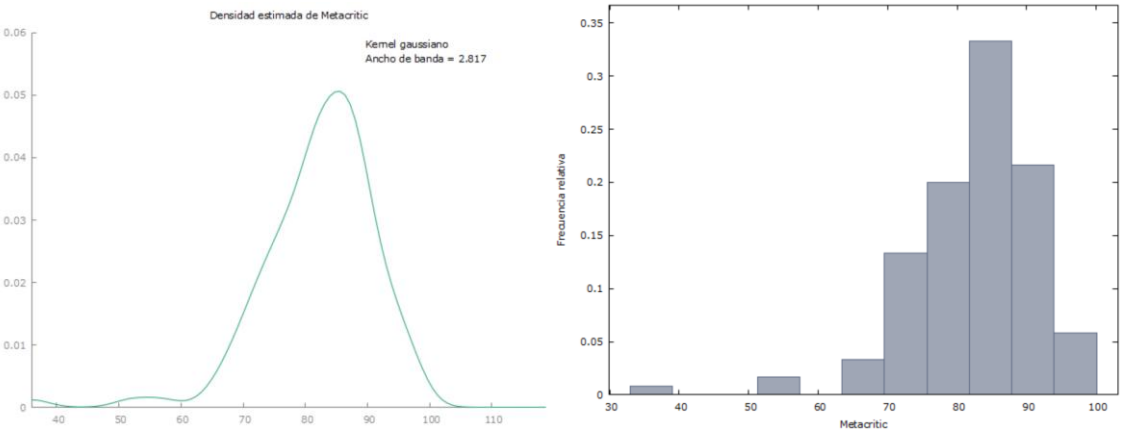
Ilustración 17:Gráafico de Densidad de 'Vandal_Comm' e Ilustración 18: Gráfico de frecuencias relativas de 'Vandal_Comm'



Esta variable cuenta con asimetría hacia la izquierda, con un cociente de -0,88, siguiendo una distribución algo leptocúrtica, con la mayor parte de las observaciones concentradas en el rango de puntuaciones superiores a 80 (alrededor del 65%), algo menos que en los medios de críticas profesionales. La media o esperanza de esta variable es 81,987 con una desviación típica de 10,5. Se obsevan ciertos comportamientos observados en las puntuaciones de 3D Juegos que se repiten también aquí, como una menor media, o una menor frecuencia a partir de la puntuación de 80.

Variable **"Metacritic"**, o la puntuación que le dio este medio crítico a los juegos observados:

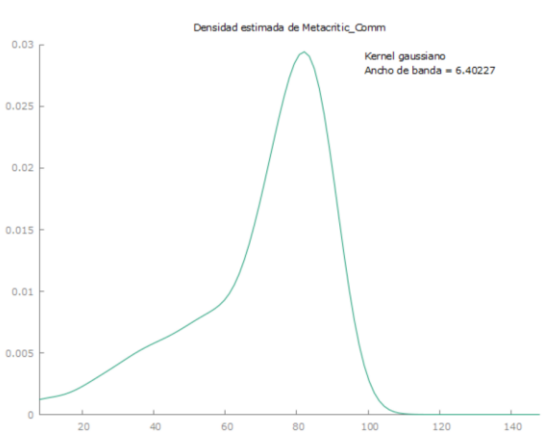
Ilustración 19: Gráfico de Densidad de 'Metacritic' e Ilustración 20: Gráfico de frecuencias relativas de 'Metacritic'

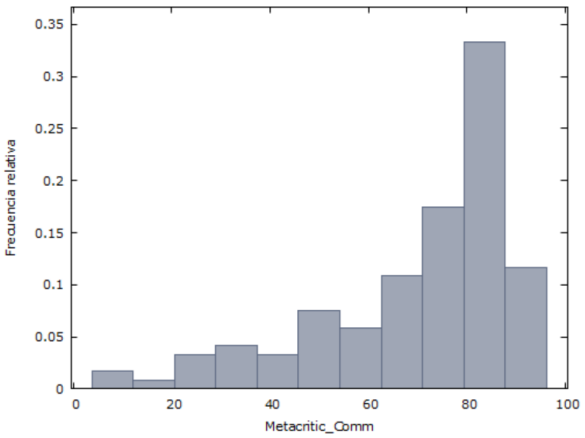


Esta variable cuenta con asimetría muy alta hacia la izquierda, con un cociente de -1,55, siguiendo una distribución algo leptocúrtica, con la mayor parte de las observaciones concentradas en el rango de puntuaciones superiores a 80 (alrededor del 60%), un dato bastante diferente respecto a los medios profesionales observados con anterioridad (que tenían más frecuencia de valores altos). La media o esperanza de esta variable es 82,14 con una desviación típica de 9,11. El comportamiento singular de esta variable es que hay mucha frecuencia de de registros entre el 70 y el 80, un 33%. En el gráfico de frecuencias se aprecia que pese a tener una marcada asimetría, realmente la variable está relativamente equidistribuida alrededor de los valores superiores.

Variable **"Metacritic_Comm"**, o la puntuación que le dio la comunidad de este medio (Metacritic) a los juegos observados:

Ilustración 21: Gráfico de Densidad de 'Metacritic_Comm' e Ilustración 22: Gráfico de frecuencias relativas de 'Metacritic_Comm'

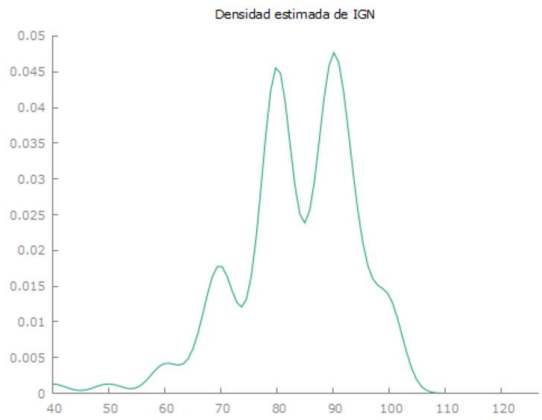


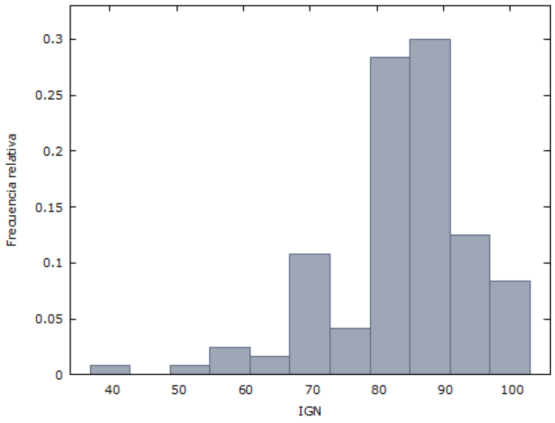


Esta variable cuenta con asimetría alta hacia la izquierda, con un cociente de -1,24, siguiendo una distribución algo leptocúrtica; la mayor parte de las observaciones están concentradas en el rango de puntuaciones superiores a 70 (alrededor del 62,5). La media o esperanza de esta variable es 69,77 con una desviación típica de 19,53. En esta variable se ve una vez más confirmado lo que se ha podido observar en las anteriores variables de puntuaciones de comunidades, las medias son más bajas (en este caso es incluso más baja que las anteriores), ya que se otorgan muchas más puntuaciones por debajo del 80. Sin embargo también se puede ver una desviación típica sobresalientemente alta, que indica una enorme volatilidad en los valores recogidos.

Variable **“IGN”**, o la puntuación que le dio este medio crítico a los juegos observados:

Ilustración 23: Gráfico de Densidad de ‘IGN’ e Ilustración 24: Gráfico de frecuencias relativas de 'IGN'

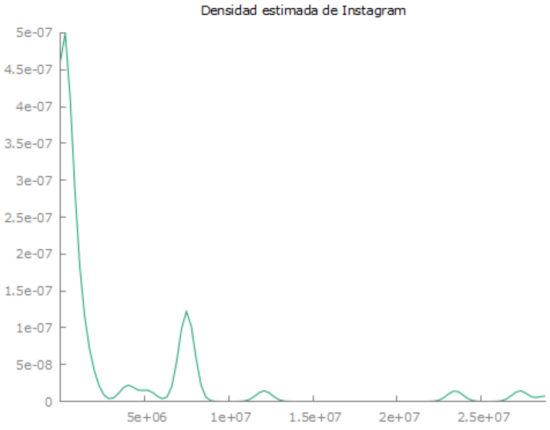


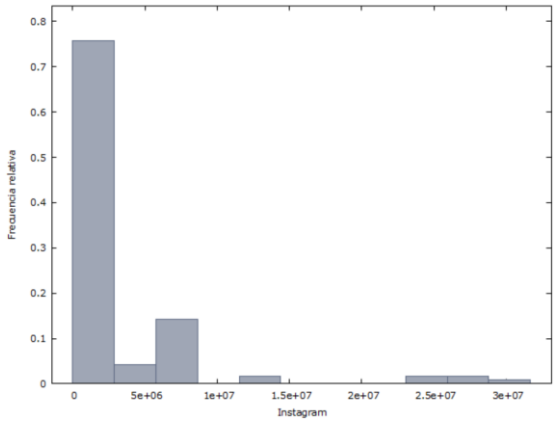


Esta variable cuenta con asimetría notable hacia la izquierda, con un cociente de -0,94, siguiendo una distribución leptocúrtica; la mayor parte de las observaciones están concentradas en el rango de puntuaciones superiores a 80 (alrededor del 80%), volviendo a la corriente observada en la mayoría de medios de críticas profesionales con datos muy cocentrados a partir del 80. La media o esperanza de esta variable es 83,44 con una desviación típica de 10,76. El mínimo es de 40.

Variable **“Instagram”**, o el número de seguidores de la cuenta principal de los juegos observados en esta red social:

Ilustración 25: Gráfico de Densidad de 'Instagram' e Ilustración 26: Gráfico de frecuencias relativas de 'Instagram'



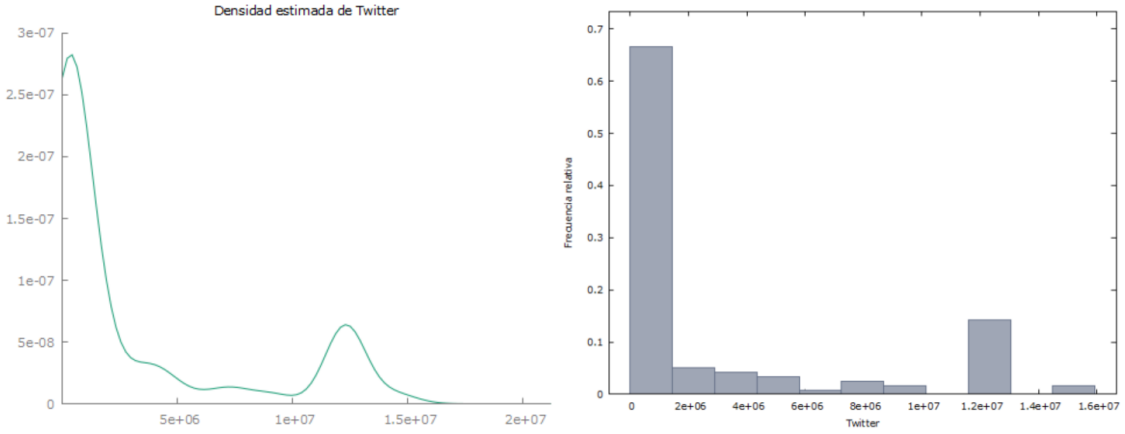


Esta variable tiene un comportamiento completamente diferente a las que se han analizado antes, ya que en esta la asimetría pasa a ser muy alta hacia la derecha, con un coeficiente de 3,04 y una curtosis altísima (el exceso de curtosis es de 9,69, lo que significa una diferencia muy alta respecto a la normal). Un 75% de los valores están en el rango inferior a los 5 millones de seguidores, la media está alrededor de 2,84 millones y la desviación típica es de 5,67 millones. Estos datos de

volatilidad, aunque altos, tienen sentido dada la enorme variabilidad de la variable puesto que muchas cuentas de desarrolladores de videojuegos alcanzan muchos millones de seguidores, como se aprecia con en el máximo de la variable de 28,8 millones.

Variable **"Twitter"**, o el número de seguidores de la cuenta principal de los juegos observados en esta red social:

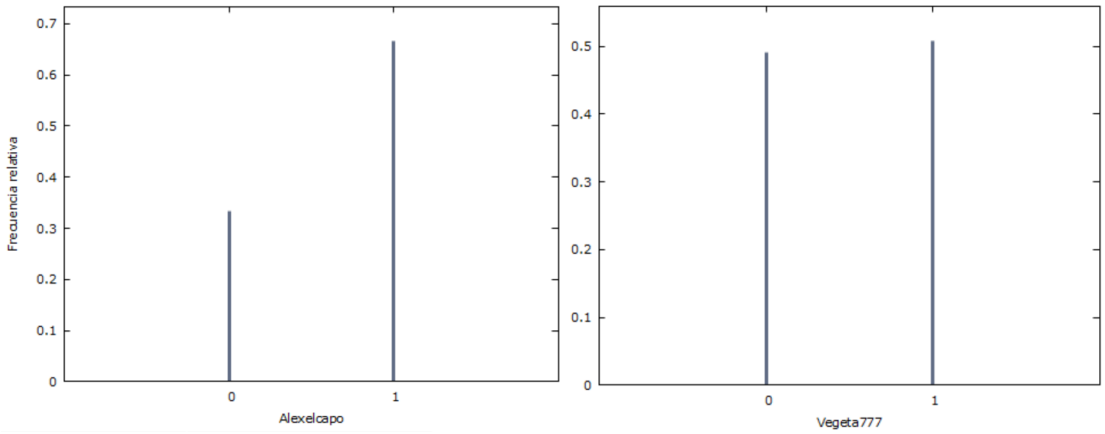
Ilustración 27: Gráfico de Densidad de 'Twitter' e Ilustración 28: Gráfico de frecuencias relativas de 'Twitter'



En esta variable se puede apreciar un comportamiento muy similar a la anterior ya que ambas miden registros de naturaleza muy parecida. Sin embargo esta tiene una distribución que pese a también ser altamente asimétrica hacia la derecha (coeficiente de 1,4), es menos densa en los valores menores. Además de esto su media es de 3 millones y la desviación típica de 4,5. Este tipo de datos pueden dejar intuir que es probable que funcione mejor en un modelo (dado a su mayor proximidad a la normal y menor volatilidad), aunque se tendrá que testar al comprobar su relación con la variable dependiente.

Variables **"Alexelcapo"** y **"Vegetta777"**, o si estos creador de contenido a expuesto el juego observado en sus canales:

Ilustración 29: Gráficos de frecuencias relativas de 'Alexelcapo' y 'Vegetta777'

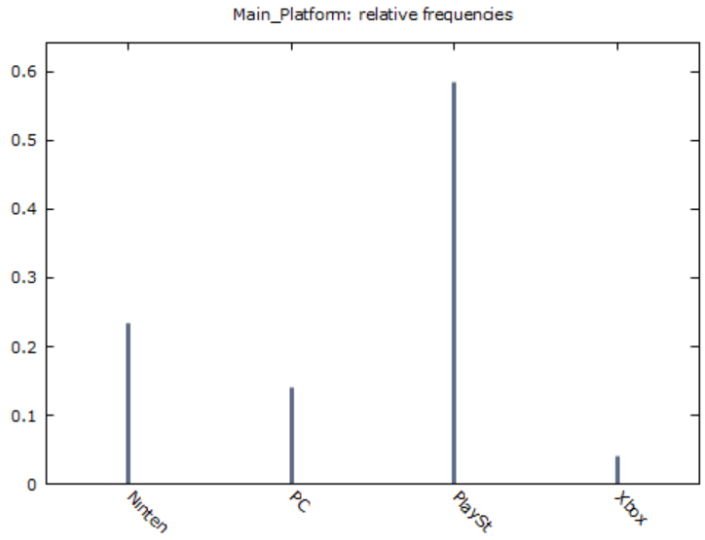


Al tratarse de variables binarias, no tendría sentido reproducir los gráficos de densidad, únicamente se puede observar, que, de los 120 juegos comprendidos en el análisis, Alexelcapo ha producido contenido sobre un 66,67%, lo cual encaja dentro de lo que se buscaba para esta variable que tenía que ser un creador de contenido de nicho que hubiese jugado a muchos de los juegos. Mientras tanto Vegetta777 ha subido videos sobre un 50,83% de los juegos de la muestra, lo cual, aunque es bastante alto para un creador de contenido de masas, sigue siendo menor que los datos registrados por el creador de nicho. A primera vista, esta proximidad entre ambos puede dar problemas en el análisis, aunque se tendrá que esperar a la construcción del modelo.

Variable **"Main_Platform"**, o la consola o plataforma en la que más se ha vendido el título en España:

Ilustración 30: Gráfico de frecuencias relativas de 'Main_Platform'

Tabla 4: Tabla de frecuencias de plataformas en la base de datos

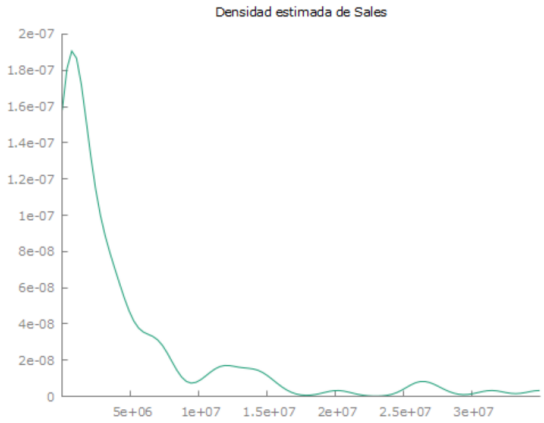


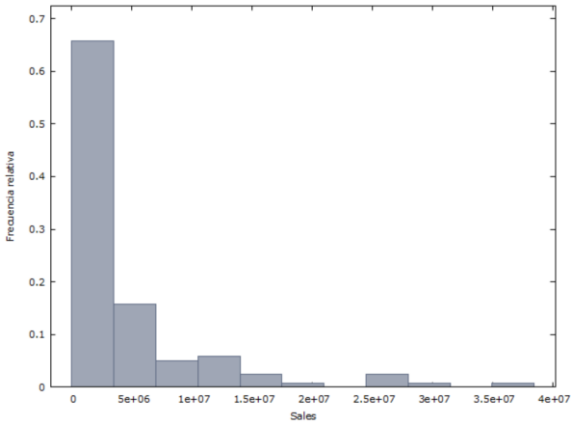
Plataforma principal	Total	Porcentaje de frecuencia
Nintendo Switch	28	23,33%
PC	17	14,17%
PlayStation	70	58,33%
Xbox	5	4,17%

Esta variable funciona de forma muy similar a **"Genre"** ya que se trata de una variable categórica con una categoría fuertemente dominante (PlayStation) seguida de una segunda con una frecuencia notable (Nintendo Switch). Aunque sería interesante para el estudio haber contado con mayor diversidad en esta variable, los resultados son lógicos si se tiene en cuenta el estado de la industria de los videojuegos española en la actualidad, en la que PlayStation es un fuerte dominador de títulos AAA frente a Xbox, mientras que Nintendo y los juegos de PC se mantienen al margen con sus propios nichos de mercado.

Finalmente, la variable dependiente u objetivo, **"Sales"** o las ventas que obtuvo el videojuego de 2019 a agosto 2022:

Ilustración 31: Gráfico de Densidad de 'Sales' e Ilustración 32: Gráfico de frecuencia relativa de 'Sales'





Esta variable se comporta de forma muy similar a lo que se ha podido observar con las variables de números de seguidores en redes sociales, sus gráficos son especialmente similares a los observados con **"Twitter"**. Algo menos de un 75% de los registros están por debajo de 5 millones de ventas, la media es de 4,37 millones y la desviación típica es de 6,6 millones. Esta información lleva a pensar que muy seguramente el modelo dependa del uso de un logaritmo en esta variable, puesto que tiene una cola marcada hacia la derecha, y una volatilidad altísima.

Los resultados de este apartado ya indican ciertos sucesos para tener en cuenta a la hora de la elaboración del modelo:

- La variable **"Release"** es, de las variables numéricas la menos asimétrica y es única a la hora de realizar el análisis: no es sustituible por ninguna otra variable que explique la misma variabilidad, esto se podrá confirmar estudiando la correlación.
- La media de las puntuaciones de medios profesionales de críticas (85 de 100) es significativamente más alta que las de foros (78 de 100), sin embargo, su desviación típica media es menor (8,9 de los medios profesional versus 13,4 de las puntuaciones de comunidades). Esto significa que los medios profesional son menos volátiles que los foros y que ponen puntuaciones más altas, mientras que los foros puntúan de forma más extrema (o muy bajo o alto).
- La variable **"Twitter"** parece ser mejor indicador que la variable **"Instagram"** a la hora de construir el modelo, debido a que está ligeramente más cerca de la distribución normal, y tener una desviación típica menor. Este tipo de resultados son valiosos ya que a la hora de construir el modelo parece que se tendrá que elegir entre uno y otro debido a que muy seguramente estén muy fuertemente correlacionados.

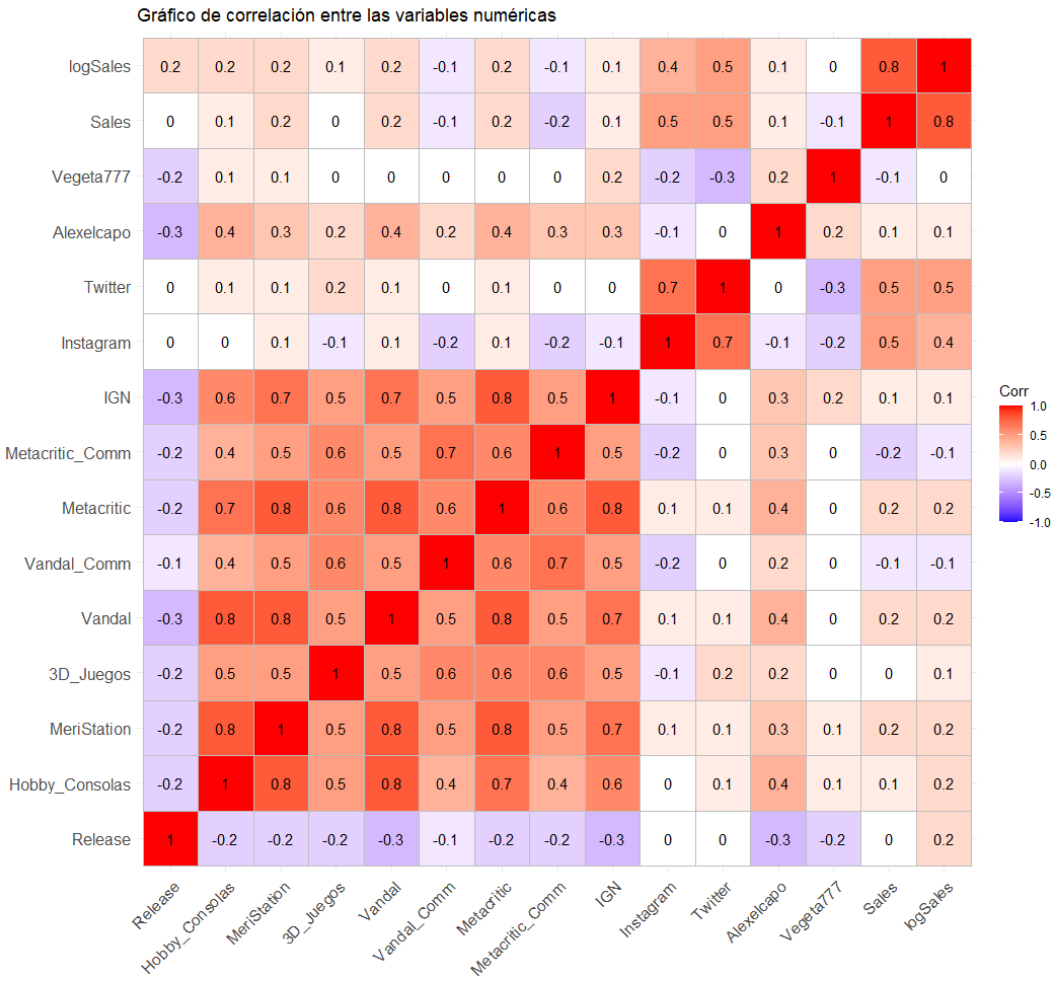
- En cuanto a variables categóricas, la variable **"Genre"** parece poco útil debido a que tiene un alto número de categorías y muchas de ellas tienen muy pocos registros mientras que la mayoría se concentran en un par de géneros como acción-aventura. Por otro lado, la variable **"Main_Platform"** sí que parece un mejor determinante de las ventas, además de contar con información suficiente en todas las categorías y explicar una dinámica importante del mercado español de videojuegos.
- Las variables utilizadas para recoger información sobre "influencers" pueden no ser determinantes debido a que se parecen mucho entre sí; no hay tanta diferencia entre la que busca explicar la información de nicho versus la que busca cubrir el mercado general.
- La variable target, "Sales" tiene una volatilidad muy alta y una marcada asimetría hacia la derecha lo que apunta a la necesidad de usar una transformación logarítmica.

B. Análisis bivariante

Para cumplir las condiciones necesarias para construir un modelo eficiente de regresión lineal múltiple se debe estudiar primero la relación entre las variables independientes y dependiente. Esto se va a llevar a cabo principalmente mediante el cálculo de correlaciones y el desarrollo de gráficos de dispersión.

Utilizando la función ggcorrplot se produce el siguiente gráfico:

Ilustración 33: Gráfico de correlaciones

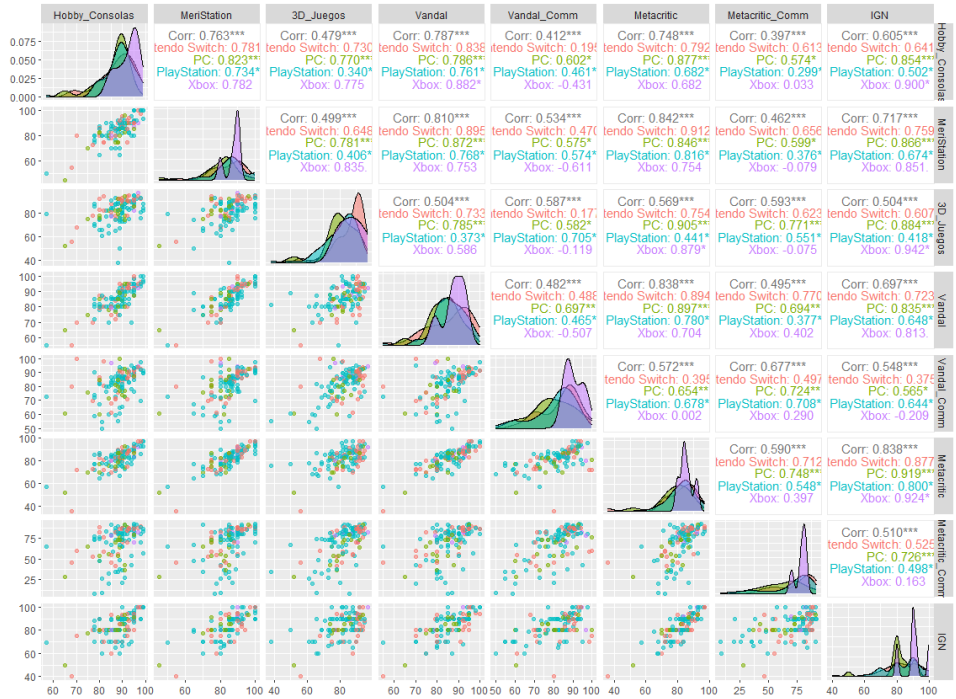


Como la correlación se calcula como un coeficiente, únicamente se puede realizar este análisis para las variables numéricas. Se puede observar claramente como hay ciertos sectores con correlaciones muy altas que corresponden a los tipos de variables incluidas en la base de datos. La zona de correlaciones más altas es la de puntuaciones de críticas, pero también se observa que sucede algo parecido entre las variables “Twitter” e “Instagram”, como se mencionaba en el análisis univariante. Estas dos variables de presencia en redes sociales también están notablemente correlacionadas con la variable objetivo, apuntando a una posible fuerte relación lineal entre estas. Esto, sin embargo, no se ve de forma tan fuerte en el resto de las variables, a

excepción de unos pocos medios de críticas como MeriStation, Vandal o Metacritic. Cabe destacar que todos los medios de comunidades tienen o ninguna correlación, o correlación negativa. También se ha incluido una nueva variable, "logSales" que consiste en el logaritmo de la variable "Sales", para estudiar que impacto tiene en la correlación el incorporar esta transformación; en la mayoría de las variables sube.

Para investigar más en profundidad las correlaciones entre los medios de críticas se va a utilizar la función del paquete GGally y ggplot2 de R, ggpairs para producir el siguiente gráfico. En este los gráficos inferiores representan los gráficos de dispersión entre la variable superior (eje X) y la variable de la derecha (eje Y), la diagonal representa los gráficos de densidad de cada variable y el sector superior muestra el coeficiente de correlación de Pearson entre las variables. Esto, se va a aplicar para estudiar la relación entre los diferentes medios de críticas y foros incluyendo que muestre el diferente impacto por cada plataforma, para observar si hay diferencias relevantes entre estas.

Ilustración 34: Estudio de la correlación entre medios de críticas por plataforma



Se ve demostrado que existe, para todos los medios, una fuerte relación de correlación positiva, es decir que, por ejemplo, cuando una puntuación es alta en Hobby Consolas, la puntuación de MeriStation también será alta, puesto que tienen una correlación de 0,763. Esto se ve confirmado también por los gráficos de dispersión, en los que se puede apreciar una clara relación lineal entre las variables. En cuanto a si existen diferencias por plataformas, no parece que exista una gran diferencia, a excepción de Xbox, cuyos coeficientes de correlación y gráficos de densidad apuntan a comportamientos muy distintos entre las variables, lo que seguramente sea causado por el hecho

que hay muy pocos juegos en la muestra de esta plataforma. Este análisis confirma que no se pueden incluir varias variables de puntuaciones de críticas al modelo puesto que se estaría incumpliendo la condición de no colinealidad. En todo caso, existe la posibilidad de que se puedan incorporar una variable de medio crítico profesional y una de medio de foro, dado que estas están menos correlacionadas entre sí (alrededor del 0,5).

En cuanto al estudio de la relación entre las variables (para comprobar si existe linealidad o no), se va a realizar una serie de gráficos de dispersión, siendo el eje Y la variable objetivo y el eje X la variable independiente analizada, con el siguiente orden de color:

- Dorado: medios de críticas profesional.
- Azul claro: medios de comunidades o foros.
- Verde: Redes sociales.
- Rosa: Año de lanzamiento del juego.

Ilustración 35: Gráficos de dispersión con la variable objetivo

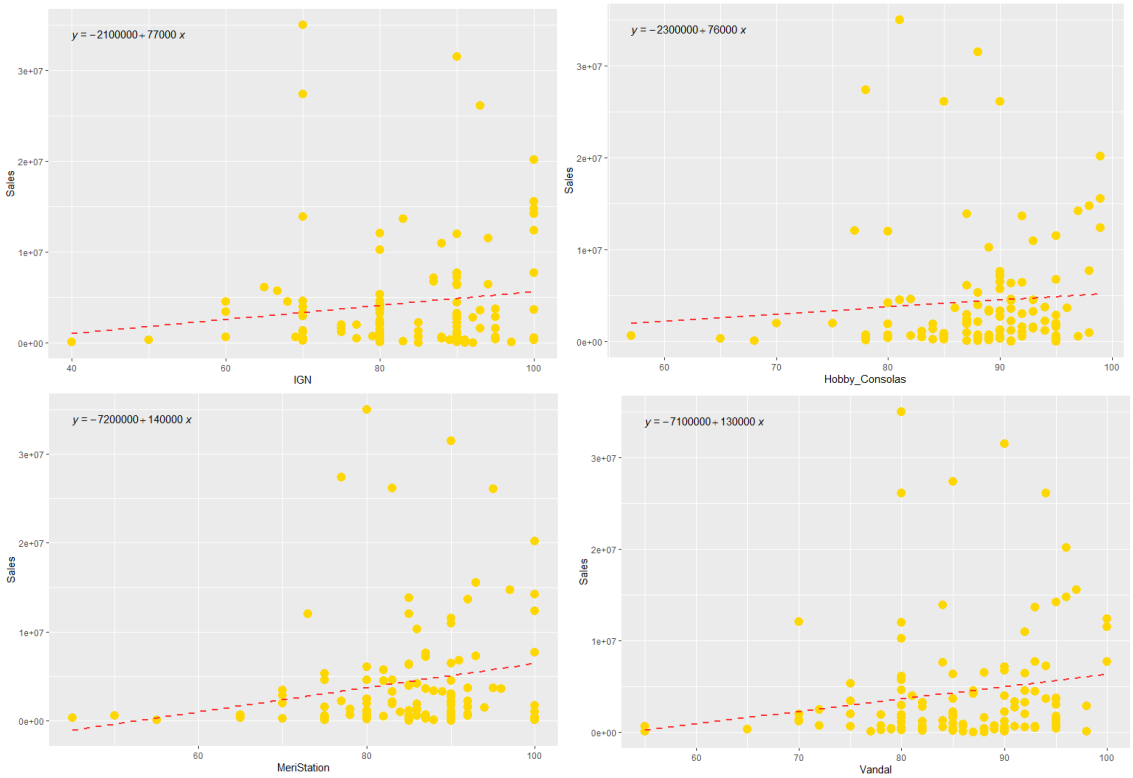
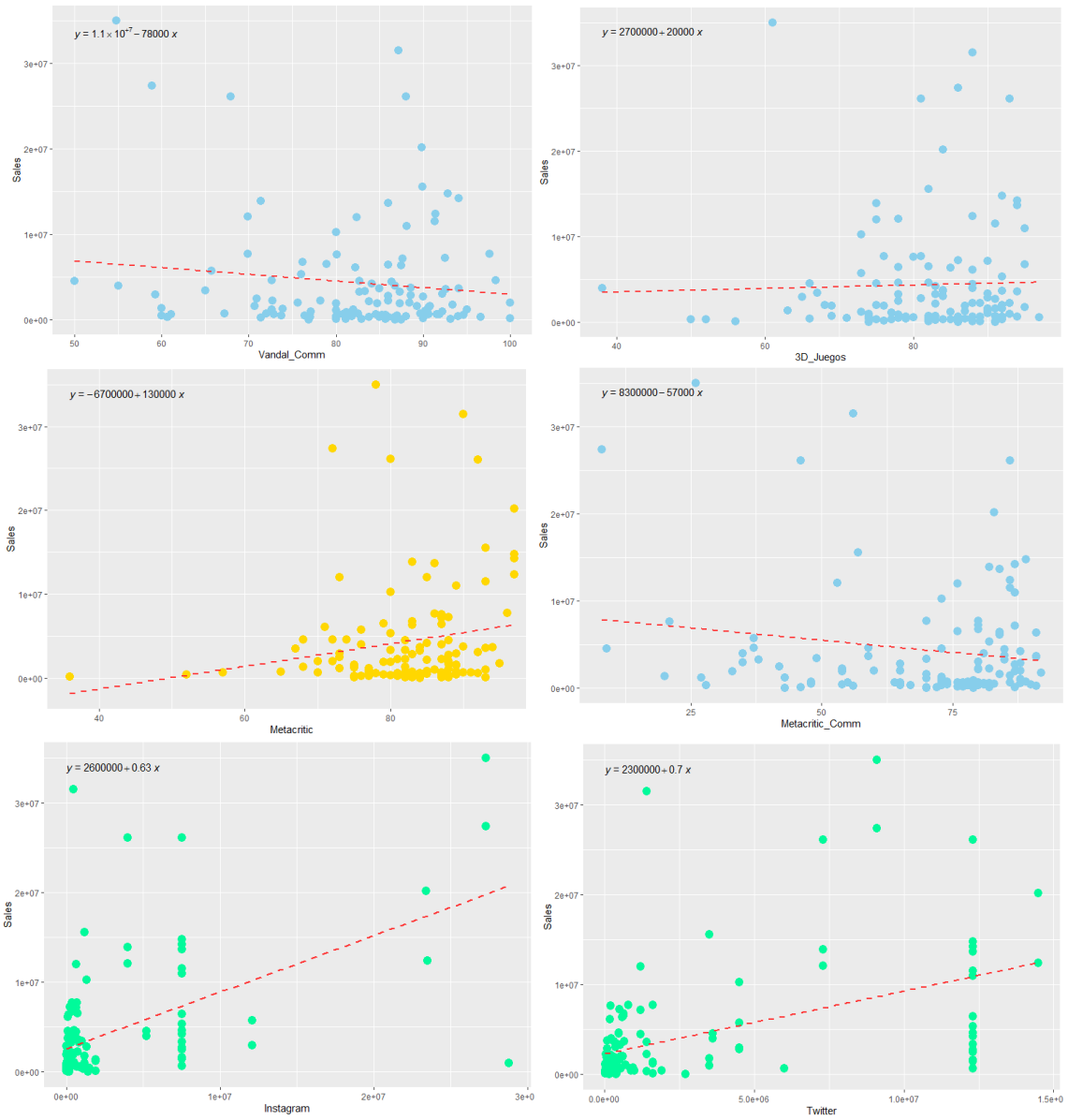


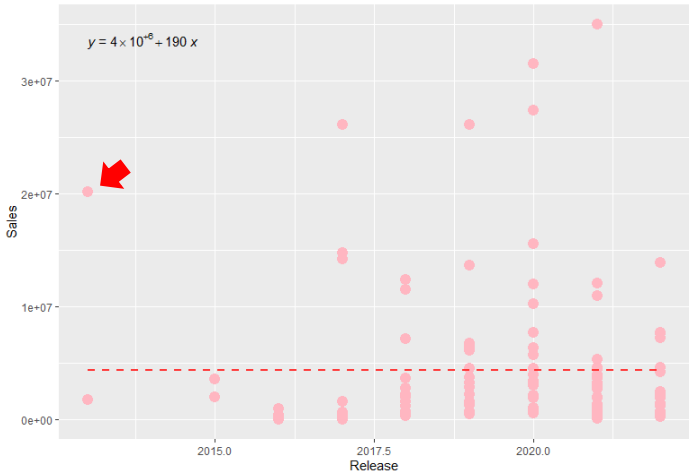
Ilustración 36: Gráficos de dispersión con la variable objetivo 2



Se puede confirmar que todas las variables numéricas observadas en los gráficos anteriores si tienen algún tipo de relación lineal con las ventas de videojuegos en España, aunque este tipo de análisis sería mejor con una gran número de datos y menor dispersión en la variable target.También es más fácil de ver si se incluye un logaritmo en la variable objetivo debido a la gran variabilidad que tiene esta, pero sin hacer esto ya se puede ver que existe linealidad.

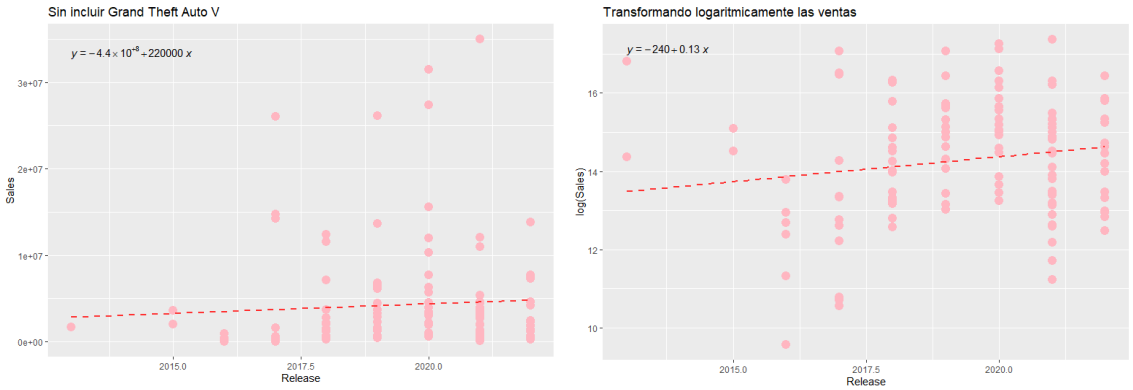
Sin embargo, para la variable “Release” esta prueba con el gráfico de dispersión no indica linealidad de forma conclusiva. Esto se da fundamentalmente por dos razones, la primera es que existe un registro en el año 2013 que tiene unas ventas anormalmente altas, eliminando la tendencia que el gráfico aparentemente tiene:

Ilustración 37: Gráfico de dispersión entre 'Release' y 'Sales'



Esta observación se refiere al juego “Grand Theft Auto”. Si se elimina del estudio, se puede observar que la variable si tiene una clara relación lineal con “Sales”, pasando de una pendiente de 190 a una de 22000. Para poder ver si se puede mantener este registro por la información que pueda aportar al resto de variables, se comprueba transformando logaritmicamente la variable objetivo si se observa una relación lineal, la cual sí se puede ver en el gráfico de la derecha.

Ilustración 38: Gráficos de dispersión entre ‘Release’ y la variable objetivo sin y con logaritmo



C. Construcción del modelo

Una vez realizadas todas estas pruebas se puede confirmar que el modelo, para cumplir las condiciones explicadas en el marco teórico y tener validez dentro del número de observaciones, va a estar limitado en cuanto al número de variables que se puedan utilizar y los tipos de estas. Estudiando las correlaciones se pueden establecer los siguientes grupos de variables independientes:

- Puntuaciones de medios de críticas profesionales.
- Puntuaciones de medios de comunidades o foros.
- Seguidores en redes sociales.
- Si el juego ha sido cubierto por un creador de contenido (de gran tamaño o nicho).
- Año de lanzamiento.
- Plataforma principal del juego (donde más se ha vendido).
- Género principal al que corresponde el juego.

Por esto el modelo podrá incluir, como máximo, una variable de cada grupo, ya que, si no, al tener variables fuertemente correlacionadas caerá en un problema de colinealidad.

También se ha podido ver como ciertas variables tienen colas hacia la derecha muy definidas, o una enorme variabilidad, que son **"Sales"**, **"Instagram"** y **"Twitter"** para estas se va a incorporar logaritmos, de forma que se suavice esa variabilidad y se vea de forma más clara la linealidad. Esto también ocurre con la variable **"Release"**, aunque en este caso debido a ser únicamente por su variabilidad, no se da de forma tan clara por lo que se evaluará el modelo con y sin transformación.

La primera versión del modelo que se va a construir es la siguiente:

$$\log(Sales) = \beta_0 + \beta_1 Release + \beta_2 \log(Twitter) + \beta_3 Main\_Platform + \beta_4 Genre + \beta_5 Hobby\_Consolas + \beta_6 Alexelcapo$$

Ecuación 11: Primera construcción del modelo

*Como "Main_Platform" y "Genre" son variables categóricas se han creado variables 'dummy' para cada una de las categorías, que tienen una beta cada una, para mantener la ecuación visualmente limpia este proceso se deja implícito en la ecuación teórica, pero sus resultados se producen de la misma manera, las ecuaciones completas del modelo final y alternativo se pueden encontrar en anexos.

En este primer intento se ven confirmadas ciertas observaciones que se han ido viendo durante el transcurso del análisis: tanto "Genre" como "Alexelcapo" tienen p-valores altísimos, por lo que su validez es nula en el análisis. Únicamente un género de los 13 incluidos en el modelo está cerca de lograr la validez con un p-valor de algo más de 0,05, y probándolo con combinaciones diferentes de variables ajustando el modelo 1, se ve que esto no varía, únicamente cambiando el género estadísticamente significativo. En cuanto a las variables que pretenden medir el impacto de "influencers" ya sea con un creador de contenido de nicho como "Alexelcapo" o uno de masas como "Vegetta777" ninguna de las dos llega a la significación de menos de 0,05 de p-valor, pese a que el segundo, el creador de contenido de masas se queda cerca de ello con un 0,06 en algunos modelos, por lo que sería interesante comprobar esta variable con más datos. Debido a este fenómeno, pero también teniendo en cuenta las restricciones del proyecto estas variables han sido eliminadas del modelo final (que debe mantenerse con un alfa de 5%), pero serán brevemente desarrolladas en el apartado de **Modelo alternativo**.

El siguiente paso es desarrollar un modelo controlado en el que se pueda ver el impacto de las variables de cada grupo para construir un modelo óptimo. Se va a seguir esta estructura:

$$\log(Sales) = \beta_0 + \beta_1 Release + \beta_2 \log(\boldsymbol{Red\ Social}) + \beta_3 Main\_Platform + \beta_4 \boldsymbol{Medio\ de\ críticas\ o\ foro}$$

Ecuación 12: Segundo modelo, con variables por testar (en negrita)

Primero se va a escoger una red social de cara a pasar con el foco del estudio, que es analizar el impacto de las diferentes críticas en las ventas de videojuegos. El primer factor inspeccionado es el p-valor de las propias variables que se refieren a las redes sociales, sin embargo, este da muy bajo para ambas variables, lo cual, aunque positivo para el modelo, no muestra evidencias para elegir a una de las dos variables. Para esto se va a observar la validez de las variables en cuanto al cambio en el p-valor que causan a los medios de críticas, y el R^2 corregido del modelo. Una vez se obtenga un modelo óptimo, se calcularán distintas pruebas para asegurar la validez estadística conjunta del modelo.

La siguiente tabla representa el p-valor que tiene la variable de cada fila cuando se incluye la variable de la columna correspondiente, es decir el cuadrado superior de la izquierda representa el p-valor de Hobby Consolas cuando en el modelo está el logaritmo de Instagram y no de Twitter.

Tabla 5: Tabla de P-valores de los resultados obtenidos para cada medio de críticas

	P-valores	
Medio de críticas	**Instagram**	**Twitter**
Hobby_Consolas	0,0760*	0,0826*
MeriStation	0,0057***	0,0099***
3D_Juegos	0,1336	0,2848
Vandal	0,0075***	0,0093***
Vandal_Comm	0,1079	0,2920
Metacritic	0,0188**	0,0172**
Metacritic_Comm	0,3019	0,5725
IGN	0,0125**	0,0172**

Además de para fijar las variables del modelo final en la tabla de p-valores se aprecia que todos los medios de comunidad no alcanzan ningún nivel de significación, por lo que, deberán ser excluidas puesto que no se desmiente la hipótesis nula. También se ha comprobado si se pudiera incluir una variable de críticas de comunidad junto a una de críticas profesional, lo cual únicamente aumenta los p-valores de ambas, por lo que esta posibilidad queda descartada.

A continuación, se observará el R^2 corregido de la misma forma que en la tabla anterior:

Tabla 6: Tabla de R^2 corregido de los resultados obtenidos para cada medio de críticas

	R^2 Corregido	
Medio de críticas	**Instagram**	**Twitter**
Hobby_Consolas	0,474955	0,455169
MeriStation	0,468664	0,472540
3D_Juegos	0,442644	0,446047
Vandal	0,466422	0,473002
Vandal_Comm	0,444298	0,445881
Metacritic	0,458610	0,467917
Metacritic_Comm	0,436745	0,441967
IGN	0,462040	0,467925

El criterio comúnmente aceptado respecto al p-valor es el de únicamente aceptar aquellos valores inferiores al 0,05, por lo que los registros superiores a esto quedan descartados. De esta forma, calculando la media de R^2 corregido para los modelos válidos se obtiene que utilizando el

logaritmo de Twitter el modelo explica ligeramente más varianza que con el de Instagram: 47% vs 46,4%. Cabe destacar que esta diferencia es mínima y como se observa con los p-valores las variables actúan de forma casi idéntica, por la que se puede afirmar que ambas serían válidas para desarrollar una regresión multivariante.

Para estudiar el impacto que tiene cada medio de críticas sobre las ventas de videojuegos en España se observará su beta, que figura en la siguiente tabla (ya se han excluido medios de comunidades o foros debido a su falta de significación):

Tabla 7: Tabla de betas para cada medio de críticas profesionales

Medio de críticas	**Beta (β)**
Hobby_Consolas	0,0281271
MeriStation	0,0309224
Vandal	0,0353462
Metacritic	0,0292132
IGN	0,0252308

Una vez comprobado que los coeficientes obtenidos para las betas son significativos, se debe mencionar que los efectos marginales que se pueden observar son todos en media y ceteris paribus. La interpretación para cada variable es la siguiente:

- Hobby_Consolas: un aumento de una unidad en la puntuación de un juego en el medio Hobby Consolas resulta en un aumento del 2,8% de las ventas del juego. Esta variable no llegaba a una significación suficiente de p-valor, sin embargo, se ha incluido para compararlo con el resto de las variables de su grupo.
- MeriStation: un aumento de una unidad en la puntuación de un juego en el medio MeriStation resulta en un aumento del 3% de las ventas del juego.
- Vandal: un aumento de una unidad en la puntuación de un juego en el medio Vandal resulta en un aumento del 3,5% de las ventas del juego.
- Metacritic: un aumento de una unidad en la puntuación de un juego en el medio Metacritic resulta en un aumento del 2,9% de las ventas del juego.
- IGN: un aumento de una unidad en la puntuación de un juego en el medio IGN resulta en un aumento del 2,5% de las ventas del juego.

Se debe recordar que para todos los medios de críticas se ha estandarizado la puntuación de 0 a 100.

Siguiendo el criterio de obtener la mayor significación posible y el mayor R^2 corregido se formula el siguiente modelo final:

$$\log(\boldsymbol{Sales}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 Release} + \boldsymbol{\beta_2 Main\_Platform} + \boldsymbol{\beta_3} \log(\boldsymbol{Twitter}) + \boldsymbol{\beta_4 Vandal}$$

Ecuación 13: Modelo final con alfa 0,05

Tabla 8: Resultados del modelo final en detalle

Modelo Final: MCO, usando las observaciones 1-120. Variable dependiente: l_Sales

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	−407.147	110.928	−3.670	0.0004	***
Release	0.203419	0.0547065	3.718	0.0003	***
PC	1.18459	0.588344	2.013	0.0464	**
Nintendo_Switch	1.82271	0.577001	3.159	0.0020	***
PlayStation	2.12919	0.531656	4.005	0.0001	***
l_Twitter	0.432699	0.0618885	6.992	<0.0001	***
Vandal	0.0353462	0.0133653	2.645	0.0093	***

Media de la vble. dep.	14.28801	D.T. de la vble. dep.	1.577695
Suma de cuad. residuos	148.2290	D.T. de la regresión	1.145321
R-cuadrado	0.499574	R-cuadrado corregido	0.473002
F(6, 113)	18.80125	Valor p (de F)	4.48e-15
Log-verosimilitud	−182.9486	Criterio de Akaike	379.8972
Criterio de Schwarz	399.4097	Crit. de Hannan-Quinn	387.8213

D. Contrastes

Para terminar de asegurar la validez de la regresión se deben realizar una serie de contrastes de acuerdo con las condiciones necesarias para realizar una regresión.

En primer lugar, se debe mencionar que el valor medio de los residuos es 0 porque se ha introducido una constante (β_0) en el modelo.

i. Heterocedasticidad

A continuación, debemos estudiar la heterocedasticidad del modelo, es decir si la varianza del error es constante para todas las observaciones del modelo. Esto se hace con el contraste de heterocedasticidad de White en el que la hipótesis nula es que no existe la heterocedasticidad: que el modelo tiene homocedasticidad (White, 1980), se obtienen los siguientes resultados:

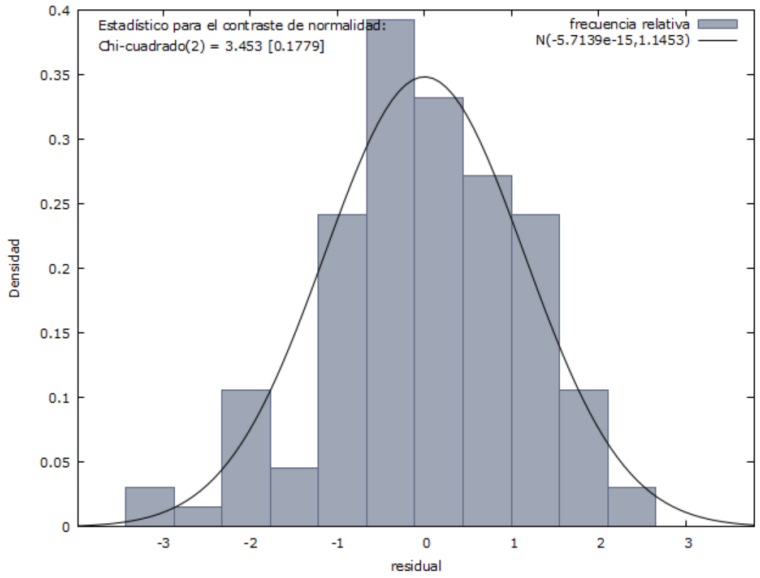
- Estadístico de contraste: LM = 25,8835
- con valor p = P(Chi-cuadrado(21) > 25,8835) = 0,210926
- Con un p-valor de referencia de 0,05, al tener este contraste un p-valor de 0,21 se confirma la hipótesis nula y afirmamos que el modelo no tiene heterocedasticidad.

ii. Normalidad de los residuos

El siguiente contraste necesario es el de normalidad de los residuos, cuya hipótesis nula es que el error del modelo sigue una distribución normal, el uso principal de esto es detectar autocorrelación. Este test observa la correlación entre los errores y los diferentes elementos del modelo, que para ser válido tendrá que ser 0 para todos los errores (Das & Imon, 2016). Se obtienen los siguientes resultados:

- Estadístico de contraste: Chi-cuadrado(2) = 3,45341
- Con valor p = 0,177869
- P-valor de referencia de 0,05, al tener este contraste un p-valor de 0,178 se confirma la hipótesis nula y afirmamos que los errores siguen una distribución normal.

Ilustración 39: Gráfico de normalidad de los residuos del modelo



Se pueden realizar una serie de gráficos para comprobar la autocorrelación de los residuos, el primero es un gráfico de dispersión entre los errores y los retardos de los errores, en el cual se puede apreciar esta si los puntos se agrupan en torno a una línea de tendencia (que como se ve aquí no ocurre en el modelo). El segundo gráfico que se puede utilizar es el gráfico Q-Q de los residuos que estudia la diferencia entre la distribución de estos y los cuantiles de la normal, por lo que cuanto más cerca de la línea diagonal, mejor, en el modelo final únicamente se observa dispersión significativa en los extremos.

Ilustración 40: Gráfico de dispersión entre los errores y los retardos de los errores

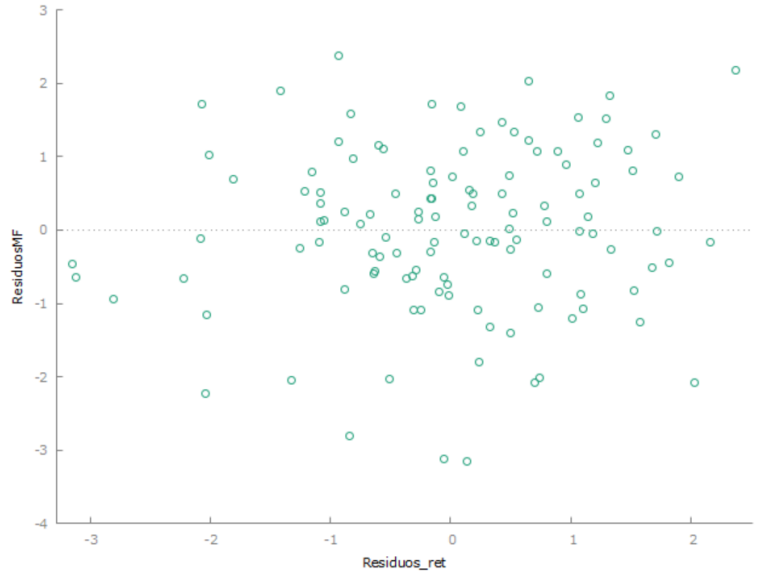
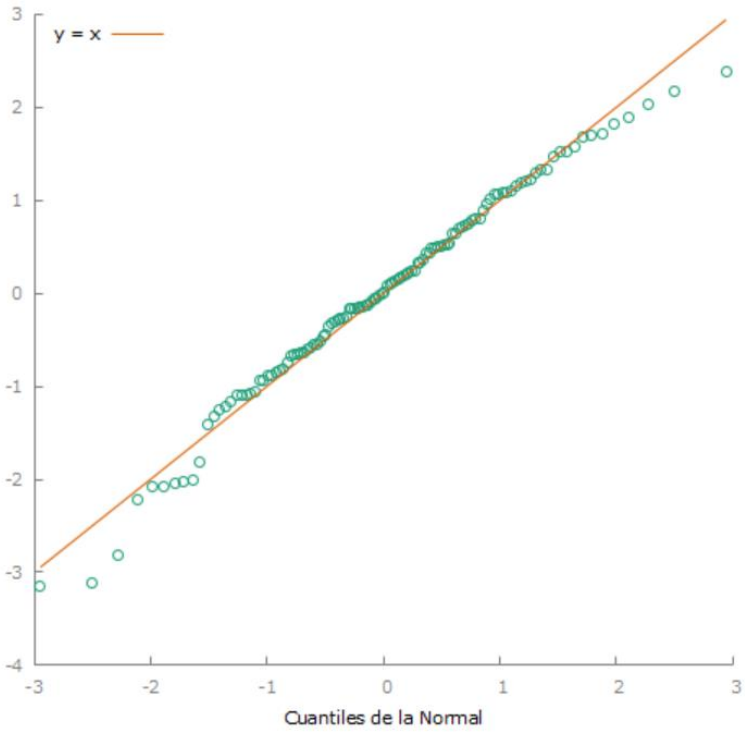


Ilustración 41: Gráfico Q-Q de los residuos del modelo



iii. Multicolinealidad

El siguiente suceso que se debe observar es la multicolinealidad, que se puede analizar mediante la elaboración de una matriz de correlaciones de las variables explicativas (Senaviratna & A Cooray, 2019). Una vez se tiene esto, aunque a simple vista se puede observar si existe un problema (si las correlaciones son o muy altas o bajas), se calcula el determinante de la matriz. Si su valor, que puede ser entre 1 y 0, da cercano a 0 se tendrá colinealidad. En este caso se obtiene un valor de 0,865, por lo que afirmamos que no existe multicolinealidad entre los predictores.

iv. Test de Chow

Otro contraste relevante a la hora de estudiar la validez del modelo tiene que ver con la estructura de los datos utilizados: el test de Chow. Esta prueba consiste en estudiar si existe una ruptura estructural en los datos mediante el desarrollo de dos modelos iguales partiendo la base de datos, por lo que cada modelo usa una parte (Eke, Eke, & Inyang, 2015). Una vez realizados los modelos, produce un p-valor y si este es menor que el valor comúnmente aceptado (0,05), existirán evidencias de que existe una ruptura en los datos: las partes de los datos son muy diferentes entre sí y los resultados de nuestro modelo no son válidos. El p-valor obtenido es de 0,6038, por lo que se afirma que no existen evidencias de una ruptura en la base de datos.

v. Comparación de modelo restringido

Para comprobar la relevancia de ciertas variables también es útil testar el modelo final desarrollado anteriormente (restringido) con diferentes modelos que representan ajustes en este, que pueden ser tanto de muestra como de variables (no restringidos) para ver cómo cambian métricas clave (Mantalos, 2003). En este caso se van a observar el R^2 y la significación. Un ejemplo de esto es si se elimina el logaritmo de "Twitter" del modelo, donde el R^2 pasa de 0,499 a 0,283, empeorando significativamente, además de producir que la categoría "PC" de "Main_Platform" pierda la significación individual. Si en vez de eso se incorpora otra variable como "Hobby_Consolas" el R^2 corregido baja a 0,469 debido a multicolinealidad.

vi. Test RESET de Ramsey

Finalmente, para juzgar globalmente el modelo se puede realizar el test RESET de Ramsey, que busca comprobar si las combinaciones no lineales de las variables puedan tener la capacidad de explicar la variable dependiente. Un modelo de regresión únicamente debe explicar relaciones lineales, por lo que, si en este test se puede observar que sí existen estas relaciones no lineales, el modelo será erróneo. En el caso del modelo final desarrollado se obtienen los siguientes resultados:

- Estadístico de contraste: F = 0,642996
- con valor p = P(F(2;111) > 0,642996) = 0,528

También se puede comprobar las relaciones lineales con el siguiente contraste de no linealidad (cuadrados):

- Hipótesis nula: la relación es lineal
- Estadístico de contraste: LM = 1,73733
- con valor p = P(Chi-cuadrado(3) > 1,73733) = 0,628666

Es decir, se descarta la hipótesis nula de que exista utilidad predictora de las relaciones no lineales, confirmando que el modelo no tiene errores de especificación. Este test también es clave para encontrar si se han omitido o no variables (Ereeş & Demirel, 2012). Comprobando los p-valores producidos en el test para cada una de las variables y para el cuadrado y el cubo del modelo se puede ver que todas están pro encima de un p-valor de 0,27. Por esto se puede afirmar que no hay variables omitidas.

E. Resultados finales e interpretación

Los resultados para cada variable (que no ha sido ya descrita) son:

Release: Un aumento de un año en la fecha de lanzamiento provoca un aumento en las ventas de un 20%. Esto sigue un comportamiento lógico puesto que las ventas incluidas son de un periodo de tiempo reciente (de 2019 a 2022), sin incluir las anteriores. Si se incluyesen datos de la variable objetivo de todo el rango de años comprendido en el estudio (desde 2013), la lógica indica que este comportamiento pasaría a ser inverso, es decir, que los juegos más antiguos tendrían cifras mayores de ventas.

PC: Esta categoría pertenece a la variable "Main_Platform", cuya categoría base se ha fijado en Xbox, debido a su comportamiento diferente observado al resto. Se refiere a aquellos juegos que han sido vendidos principalmente en PC, y lo que indica el coeficiente β es que si un juego ha tenido las ventas principalmente en PC y no en Xbox vende un 1,21 veces más.

Nintendo_Switch: Esta categoría también pertenece a la variable "Main_Platform". Se refiere a aquellos juegos que han sido vendidos principalmente en Nintendo Switch, y lo que indica el coeficiente β es que si un juego ha tenido las ventas principalmente en esta plataforma y no en Xbox vende un 1,82 veces más.

PlayStation: Esta categoría pertenece a la variable "Main_Platform. Se refiere a aquellos juegos que han sido vendidos principalmente en PlayStation, y lo que indica el coeficiente β es que si un juego ha tenido las ventas principalmente en PlayStation y no en Xbox vende un 2,12 veces más. Estos resultados son lógicos teniendo en cuenta el mercado de videojuegos actual español.

L_Twitter: Esta variable incorpora un logaritmo por lo que la interpretación, como se explicaba en el marco téorico, será diferente al resto. Un aumento de 1% de seguidores en Twitter significaría un aumento de 0,4327% en ventas.

El estadístico R^2 es de 0,4995 por lo que se puede afirmar que el modelo explica el 49,95% de la variabilidad que experimenta la variable objetivo. Además de esto, se obtiene que el p-valor global del modelo o F-test da un valor muy cercano al 0, probando la significación global de la regresión.

F. Modelo alternativo

Debido a lo cercano que está el modelo a lograr la validez cuando se incorpora a este la variable referida al influencer Vegetta777, se ha decidido explorar brevemente como se incluiría esta y cómo afectaría a los resultados del modelo final. Esto se va a realizar debido a que, puesto que existe una limitación inicial de número de registros en la base de datos, el que esta variable se quede cerca de lograr la significación a una alfa (o nivel de significación) del 5% es suficiente para poder afirmar que seguramente con más registros se pueda lograr este nivel de manera adecuada. En otras palabras, debido a la limitación de datos, se puede construir un modelo cambiando la alfa al 10%. Este sería el siguiente:

$$\log(\boldsymbol{Sales}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 Release} + \boldsymbol{\beta_2 Main\_Platform} + \boldsymbol{\beta_3}\log(\boldsymbol{Twitter}) + \boldsymbol{\beta_4 Vandal} + \boldsymbol{\beta_5 Vegetta777}$$

Ecuación 14: Modelo alternativo con alfa 0,10

Tabla 9: Resultados del modelo alternativo en detalle

Modelo 1: MCO, usando las observaciones 1-120. Variable dependiente: l_Sales

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	−430.379	110.235	−3.904	0.0002	***
Release	0.214822	0.0543618	3.952	0.0001	***
PC	1.20728	0.581344	2.077	0.0401	**
Nintendo_Switch	2.18780	0.600123	3.646	0.0004	***
PlayStation	2.22947	0.527749	4.224	<0.0001	***
l_Twitter	0.421640	0.0614037	6.867	<0.0001	***
Vandal	0.0349753	0.0132050	2.649	0.0092	***
Vegetta777	0.470332	0.241781	1.945	0.0542	*

Media de la vble. dep.	14.28801	D.T. de la vble. dep.	1.577695
Suma de cuad. residuos	143.3845	D.T. de la regresión	1.131468
R-cuadrado	0.515929	R-cuadrado corregido	0.485674
F(7, 112)	17.05300	Valor p (de F)	3.44e-15
Log-verosimilitud	−180.9549	Criterio de Akaike	377.9098
Criterio de Schwarz	400.2098	Crit. de Hannan-Quinn	386.9659

Como se puede comprobar, se han mejorado tanto el R^2 como el R^2 Corregido, explicando más de un 50% de la variabilidad de la variable objetivo, además de que el F-test sigue siendo muy cercano a 0, manteniendo la significación global. Así, se demuestra que esta variable (y seguramente otras variables que se refieran al comportamiento de influencers) también afectan a las ventas de videojuegos. En este caso midiendo el efecto de Vegetta777 en concreto podemos ver que los juegos cubiertos por este, según su beta, tienen un 47% de ventas más que aquellos que no ha jugado. En cuanto a los contrastes el modelo los cumple todos, que en su mayoría mejoran y únicamente empeora el test de Chow (que sigue siendo válido).

6. APLICACIÓN PRÁCTICA A LA INDUSTRIA

Habiendo evaluado y confirmado la validez del modelo se pueden empezar a extrapolar sus resultados al ámbito estratégico de la industria de los videojuegos. La importancia y utilidad que puede tener este modelo interpretado correctamente es obvia una vez entendidos los conceptos econométricos a los que atiende. Siguiendo un pensamiento minimalista en el que el objetivo principal o por lo menos uno de los objetivos fundamentales de las empresas de videojuegos es aumentar las ventas de estos, un modelo que puede explicar que afecta a dichas ventas es relevante.

Esta relevancia tiene que ver con controlar y conocer esas variables dependientes que, mediante la elaboración de la regresión lineal múltiple, han sido confirmadas como clave a la hora de explicar la variabilidad de las ventas. Esto permite tomar decisiones estratégicas de forma más sencilla (Nathans, Oswald, & Nimon, 2012), puesto que el entendimiento que se tiene de la industria es extenso y veraz, ya que está confirmado con datos. En un ámbito empresarial en el que ninguna empresa tiene recursos infinitos, es clave el conocer donde se maximiza el retorno en cuanto a alocar recursos, ya que, por ejemplo, si como ha sido observado Vandal tiene un coeficiente de beta mayor que Hobby Consolas y solo se puede realizar una acción pagada en uno de los dos, los datos y la lógica nos indican que se debe escoger al primero, puesto que se obtendrán mayores ventas según el modelo.

El modelo, de carácter explicativo, también ha resultado en una serie de aportaciones que pueden ser clave a la hora de entender la industria de los videojuegos española:

La primera aportación tiene que ver con el foco principal de este proyecto; gracias a la elaboración de este modelo se ha podido comprobar si realmente hay evidencias para afirmar que los medios de críticas afectan a las ventas y que medios de críticas afectan más. Además de esto se ha podido observar la diferencia entre medios nacionales (que afectan más) e internacionales. Por otro lado, en cuanto a páginas de puntuaciones, ha llamado especialmente la atención que pese a contar con la misma estructura de datos, medidos de la misma forma que las críticas profesionales, no se han encontrado evidencias para afirmar que los medios de comunidades o foros tengan efecto en las ventas, puesto que en ningún momento lograban la significación estadística. Esto, conociendo a fondo el caso es lógico, puesto que tiene que ver con la forma en la que se elaboran estas puntuaciones, en la que únicamente participan casos extremos, ya sea gente a la que le ha encantado el juego o gente que lo ha odiado. Con normalmente muy pocos votos; esto resulta en una puntuación muy poco fiable y con poca estandarización, causando alta volatilidad y resultados incongruentes.

Otra información que provee el modelo es la magnitud de la variabilidad de las ventas que depende de las redes sociales. Esto indica la fuerte relación entre seguidores en redes y ventas, ya

que se comprobaron tanto Twitter como Instagram y en ambos casos eran las variables que más variabilidad explicaban del modelo. Por esto se puede afirmar con seguridad que existe una relación entre las ventas y los seguidores en redes, llegando hasta cuantificar que por cada 1% de aumento de seguidores en Twitter, las ventas subirían 0,43%.

En definitiva, se pueden realizar estas aplicaciones sobre todas las variables que han resultado significativas para el modelo, como utilizando la variable categórica pertinente a la plataforma principal del título o juego. Un desarrollador de videojuegos puede escoger crear su juego principalmente en PlayStation de cara a obtener mejor funcionamiento en esa plataforma y vender más en ella, puesto como hemos visto si un juego tiene la mayor parte de sus ventas en PlayStation, en España venderá más de un 200% que, si la mayor parte de sus ventas fuese en Xbox, su competidor principal. Existen multitud de aplicaciones más y relaciones que seguramente se podrían estudiar pero en este modelo están limitadas por diversos factores como se va a discutir en la conclusión, pero el hecho de que este estudio cuente con 120 registros y ya se puedan ver toda esta cantidad de relaciones y sacar conclusiones tanto importantes como válidas estadísticamente reafirma la importancia que tiene y va a tener la toma de decisiones mediante el uso de datos tanto para esta industria como para cualquier otra que disponga de información para ello.

Sumado a esto, la clave en la aplicación de estos métodos no solo el basarse en la información que nos provee el dato, si no la combinación de esta nueva información obtenida a través de la analítica con la experiencia de ejecutivos que sepan entenderlos. Por esto el entendimiento de estos procesos y de las fórmulas detrás de ellos es fundamental antes de tomar una decisión con estos hallazgos y lograr el siguiente paso, que sería la automatización de la producción de esta información (Davenport, 2009).

7. CONCLUSIÓN Y ÁREAS DE MEJORA

A. Comprobación de objetivos, hipótesis, restricciones y mejoras

A la hora de plantear este proyecto se realizaron una serie de conjeturas de lo que podía encontrarse en el modelo, además de ciertas suposiciones, que, estando basadas en investigación y conocimiento previo de la industria, eran necesarias a la hora de realizar el modelo y de que este funcionase. Se van a repasar a continuación:

La primera hipótesis afirmaba que existiría una relación lineal positiva entre las críticas y las ventas, es decir los juegos con mayores puntuaciones seguramente serían aquellos con mayores ventas. Esto se ha visto confirmado en que sí se puede afirmar la existencia de dicha relación lineal, sin embargo, esta es mucho menor de lo que se esperaba al principio, y no siempre puntuaciones altas suponen ventas altas. Una posible explicación de esto extraída de investigación conjunta del análisis es que normalmente las ventas más altas vienen de franquicias, juegos que constantemente repiten la misma fórmula y que, aunque en sus inicios tuvieron gran recepción por parte de los críticos, debido a la repetición ahora obtienen puntuaciones mediocres. Este es el caso de títulos como FIFA, Pokémon, Call of Duty, etc.

La segunda hipótesis planteaba la existencia de una relación lineal entre seguidores en redes sociales generales y las ventas de videojuegos. Como ejemplos de redes sociales se usaron Instagram y Twitter en el modelo y se ha podido comprobar que estas son una gran variable independiente para predecir las ventas, ya que en un modelo univariante por sí mismas son capaces de explicar más del 20% de la variabilidad en un modelo válido. Este tipo de variable ha sido la que más impacto tiene el modelo, y sin duda viene de que, como se ha visto en el análisis univariante, son las variables observadas que más relación lineal tienen con las ventas de videojuegos.

En cuanto a la hipótesis número 3, la que se refiere a los influyentes o creadores de contenido en plataformas como YouTube o Twitch, no se ha podido demostrar aquello que mencionaba la hipótesis al nivel de confianza buscado o para ambos creadores de contenido, pero sí se han logrado resultados positivos para uno de ellos. La hipótesis hablaba de que, si un influyente relevante cubría un juego o subía contenido sobre este, el juego vería un aumento en ventas, por lo que se introdujeron dos variables dicotómicas sobre dos creadores diferentes, uno de nicho y uno de masas. El de nicho, que había jugado a la gran mayoría de los juegos, se quedó muy lejos de la validez por lo que no se puede afirmar que tenga ningún impacto en las ventas, por lo menos de la forma que se intentó comprobar. Por otro lado, el de masas, Vegetta777, que había jugado a la mitad de los juegos, pero en la mayoría de los casos a los más populares, sí que lograba cierta capacidad explicativa, logrando alcanzar un nivel de significación, pero aun quedándose cerca, no alcanzando el nivel necesario de p-valor para el modelo final. Pese a esto, debido a lo cerca

que estuvo esta variable de hacer una aportación notable al modelo y al funcionamiento que logra como parte del modelo alternativo, se recomienda fuertemente incorporar esta variable al modelo si se cuentan con más datos, o por lo menos comprobar si se puede utilizar de nuevo, así como estudiar una forma más extensa la forma de incorporar el impacto de otros influyentes de nicho. Para terminar con esta hipótesis, sí que se puede afirmar cierta relación entre el "influencer" Vegetta777 y las ventas de videojuegos, dejando intuir que existe una relación entre ciertos tipos de creadores y el mercado de los videojuegos.

En la hipótesis 4 sucede algo similar a lo ocurrido en la anterior con el caso de Alexelcapo, se ha intentado medir el impacto que pueden tener los diferentes géneros de juegos en sus ventas a través de una variable categórica, pero no se ha logrado demostrar que realmente exista. La causa principal de esto parece ser el hecho de que en la muestra dominen fuertemente unos pocos tipos de juegos y que muchas categorías apenas cuenten con un registro, sin embargo, este comportamiento es muy parecido al que vemos en la industria en general, donde la mayoría de los juegos, vendidos son acción-aventura y FPS, por lo que, a diferencia de la hipótesis anterior, el lograr que una variable como esta funcione para el modelo seguramente tendrá poco que ver con aumentar la muestra, si no con cambiar el funcionamiento de la variable.

La hipótesis 5, teorizaba que los juegos con plataforma principal Nintendo, tendrían mejores críticas que los del resto de plataformas, sin embargo, observando los datos se comprueba que esta suposición fue errónea. Los juegos de Nintendo son los segundos con mejores críticas en media, mientras que los primeros son aquellos que tienen Xbox como plataforma principal, incluyendo tanto críticas de profesionales como las de foros. Para analizar la diferencia entre ambos se divide esta media entre críticas de profesionales y foros, donde, si se analizan las correlaciones entre las distintas categorías de la variable que se refiere a la plataforma principal se obtiene que la diferencia fundamental entre Xbox y Nintendo viene de los medios de críticas profesionales. En los medios de foros, Nintendo tiene correlaciones positivas más altas, lo cual tiene sentido dentro de las circunstancias empresariales de la empresa japonesa, con su nicho de mercado formado por comunidades fieles y casi exclusivas en torno a sus lanzamientos.

En la sexta hipótesis se comentaba que, por lógica, los medios nacionales seguramente tendrían más repercusión en las ventas españolas que los medios internacionales. Esto se ha podido ver comprobado por el modelo, en el que los medios españoles tenían mayor coeficiente en media que los internacionales, con validez estadística. Contrariamente a esto, la diferencia tampoco era excesiva (menos de un 0,5% de diferencia de impacto), puesto, como se ha podido ver en los análisis todos los medios de críticas estaban altamente correlacionados, por lo que las puntuaciones en unos medios u otros eran muy parecidas.

La siguiente hipótesis, la séptima, conjeturaba que los medios de críticas tendrían más impacto que los foros o medios públicos, es decir, que o bien en el modelo los coeficientes de las variables de medios de críticas profesionales eran más altos, o más significativos. En este estudio ha ocurrido lo segundo, hasta el punto de que los p-valores de las variables de medios de comunidades no logran, ni están en ningún momento cerca de lograr la significación estadística, haciendo que no se pueda concluir que exista relación alguna entre las ventas y estas puntuaciones. En resumen, si una empresa tuviese que contemplar si centrar su atención en acciones en un medio profesional o en uno de comunidad, mientras que existen evidencias basadas en datos de que, si suben las puntuaciones en medios críticos, suben las ventas, este no es el caso con los foros.

En la penúltima hipótesis se menciona el efecto que debería tener el año de lanzamiento en las ventas de un juego según el modelo que se va a diseñar. Esto tiene que ver con que los datos con los que contamos de la variable target no sean todas las ventas de los videojuegos observados desde su lanzamiento, sino que son todas las ventas en una franja de tiempo: enero de 2019-agosto de 2022. Se ha podido comprobar que existe una relación lineal positiva, por lo que esta hipótesis se cumple, sobre todo si se eliminaba el registro que se menciona en la propia hipótesis, de un juego de 2013 con notables ventas. Durante el curso del proyecto se ha observado que, si se llegase a contar con más datos de la variable objetivo, el comportamiento de esta variable sería opuesto, puesto que lógicamente los juegos con más años en venta tendrían más ventas acumuladas, y de esta manera esta variable seguiría impactando fuertemente al modelo.

Para terminar con la revisión de las hipótesis, en la novena y última se habla de una posible diferencia entre medios nacionales e internacionales, y se teoriza que los nacionales puntúan de forma más alta. Gracias a la elaboración de la base de datos en el estudio esto se puede contrastar rápidamente, y se obtiene que, para los 120 juegos observados, los medios nacionales (Hobby Consolas, MeriStation y Vandal) tienen una puntuación media de 86/100, mientras que los internacionales (Metacritic e IGN) tienen una puntuación media de 82/100. Aunque es cierto que en ambos casos parece una media desproporcionada, se cumple lo propuesto en la hipótesis 9: los medios nacionales puntúan más alto.

En cuanto a áreas de mejora, tal y como se apuntaba en el comienzo del proyecto al establecer ciertas restricciones, el estudio y por lo tanto el modelo que se ha producido tienen ciertas limitaciones arraigadas en la naturaleza de este. Pese al diseño de una forma sofisticada de extraer conclusiones estratégicas a través de datos reales de la industria, este trabajo ha estado limitado por restricciones de tiempo y de extensión. De cara a diseñar futuros estudios se recomienda fuertemente extender la base de datos empleada, ya que ya se tienen datos más que suficientes de la variable target (más de 5000 juegos incluidos en la base de datos), pero no se tiene suficiente

información de las variables dependientes, que tuvieron que ser introducidas mediante búsqueda en web. El contar con más datos, utilizando y testando el modelo desarrollado aquí (el orden y los tipos de variables escogidas) también permitiría diseñar algoritmos predictivos complejos, que sería el siguiente paso lógico en cuanto a desarrollo del tema, así como incluir variables y testar nuevas relaciones que el resultado de este proyecto no ha podido demostrar completamente. Otros modelos que se podrían desarrollar con un rol predictivo podrían ser redes neuronales, cuyos resultados y estructura serían los siguientes:

Ilustración 42: Representación gráfica de un modelo alternativo de red neuronal

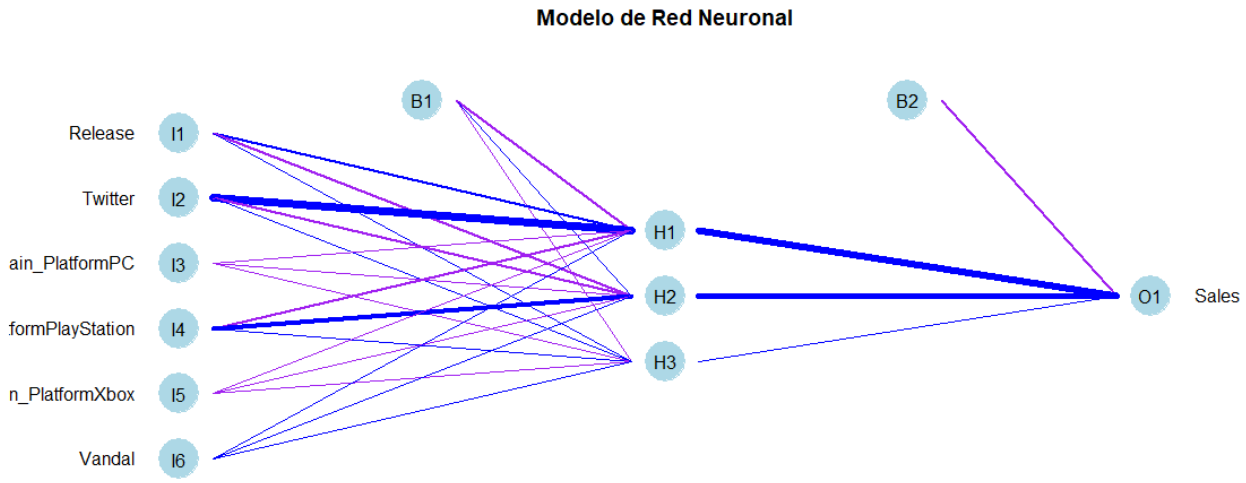
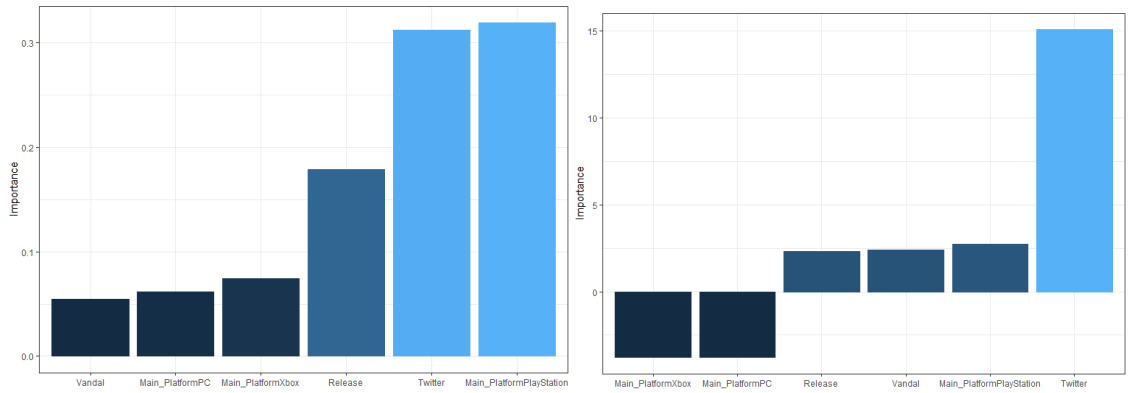


Ilustración 43: Importancia de las variables para la red neuronal



La explicación completa e inclusión de estos algoritmos en este trabajo no se ha realizado de manera que se mantenga el foco en el modelo explicativo y el significado estratégico que tiene este, así como la ya mencionada limitación de datos. Sin embargo, pese a su nula capacidad de interpretación debido a tratarse de un modelo de 'caja negra', no se debe infravalorar la capacidad predictiva que pueden tener estas técnicas y su clara importancia empresarial (que las empresas sean capaces de predecir sus ventas). Por esto, se han incluido los desarrollos en código de estas técnicas en el lenguaje de programación R, en el apartado de **Anexos** de este proyecto.

En definitiva, habiendo sentado las bases para este desarrollo y dejando un cauce claro para lo que puede ser la siguiente fase de este estudio, se considera que se ha logrado más que satisfacer la gran pregunta inicial que se trataba de responder en relación con los medios de críticas. No solo esto si no que se ha expandido sobre esta, incluyendo numerosas variables de interés de la industria de los videojuegos, llegando en la mayoría de los casos a resultados de gran valor informativo y estratégico para las empresas del sector, pese a las restricciones con las que se contaba, consiguiendo así cumplir todos los objetivos.

B. Hallazgos resumidos

- Existe una relación lineal clara entre puntuaciones en medios de críticas y las ventas actuales de videojuegos, un ejemplo de esto es Vandal, dónde se ha calculado que un aumento en 1 punto de crítica (medido de 0 a 100), supone un aumento de 3,5% en las ventas del juego.
- La mayor relación lineal observada entre las variables independientes observadas y las ventas es con las redes sociales, que explican la mayoría de la variabilidad explicada. Son clave para la elaboración del modelo y se ha calculado que un aumento en seguidores del 1% supondría un aumento de 0,43% en ventas.
- Para un nivel inferior de significación los influencers de masas sí tienen un efecto enorme para las ventas, aunque esto se debe estudiar en profundidad.
- No se ha podido demostrar que el pertenecer a un género concreto (acción-aventura, deportes, etc.) afecte a las ventas de un juego.
- Los juegos de Xbox en media eran los que mejor puntuación de la crítica tenían, mientras que los de Nintendo obtuvieron la mejor puntuación en foros. Aun así, se demostró que, si un juego tenía como plataforma principal PlayStation, Nintendo Switch o PC y no Xbox, sus ventas serían un 212%, 182% y 121% mayores, respectivamente.
- No se ha podido demostrar que los medios de comunidades o foros afecten a las ventas.
- Los medios españoles puntúan más alto de media que los internacionales y afectan más a las ventas en España.

8. BIBLIOGRAFÍA

Adigüzel, F. (2021). The Effect of YouTube Reviews on Video Game Sales. *İşletme Araştırmaları Dergisi, 13*(3), 2096-2109.

Baños, R. V., Torrado-Fonseca, M., & Álvarez, M. R. (2019). Análisis de regresión lineal múltiple con SPSS: un ejemplo práctico. *REIRE Revista d'Innovació i Recerca en Educació, 12*(2), 1-10.

Belli, S., & Raventós, C. (2008). A brief history of videogame. *Athenea Digital: Revista de Pensamiento e Investigacion Social*(14), 159-179. Obtenido de https://ddd.uab.cat/record/29923

Benoit, K. (2011). Linear regression models with logarithmic transformations. *London School of Economics, 22*(1), 23-36.

Buijsman, M. (12 de Mayo de 2022). *The Top 10 Public Game Companies Generated $126 Billion in 2021 as Subscriptions and M&A Shake up the Market.* Obtenido de newzoo: https://newzoo.com/insights/articles/the-top-10-public-game-companies-generated-126-billion-in-2021-as-subscriptions-and-ma-shake-up-the-market

Clements, M. T., & Ohashi, H. (2005). Indirect network effects and the product cycle: video games in the US, 1994–2002. *The Journal of Industrial Economics, 53*(4), 515-542.

Cox, J., & Kaimann, D. (2015). How do reviews from professional critics interact with other signals of product quality? Evidence from the video game industry. *Journal of Consumer Behaviour, 14*(6), 366-377.

Das, K. R., & Imon, A. H. (2016). A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics, 5*(1), 5-12.

Davenport, T. H. (2009). Make better decisions. *Harvard business review, 87*(11), 117-123.

Eke, F. A., Eke, I. C., & Inyang, O. G. (2015). Interest rate and commercial banks' lending operations in Nigeria: A structural break analysis using chow test. *Global Journal of Social Sciences, 14*(1), 9-22.

Ereeş, S., & Demirel, N. (2012). Omitted variable bias and detection with reset test in regression analysis. *Anadolu University Journal of Science and Technology B-Theoretical Sciences, 2*(1), 1-19.

Fonseca, J. A. (27 de Enero de 2022). *Ventas España 2021: el repaso del año.* Obtenido de Game Reactor: https://www.gamereactor.es/ventas-espana-2021-el-repaso-del-ano/

Keogh, B. (2015). Between Triple-A, indie, casual, and DIY: Sites of tension in the videogames cultural industries. En *The Routledge companion to the cultural industries* (págs. 152-162). Routledge.

Llinás Solano, H. (10 de Julio de 2021). *ESTADÍSTICA DESCRIPTIVA EN R: Medidas estadísticas.* Obtenido de RPubs by RStudio: https://rpubs.com/hllinas/R_Medidas

Mantalos, P. (2003). *Bootstrapping the Breusch-Godfrey autocorrelation test for a single equation dynamic model: Bootstrapping the Restricted vs. Unrestricted model.* Lund: Lund University.

Nathans, L. L., Oswald, F. L., & Nimon, K. (2012). Interpreting multiple linear regression: a guidebook of variable importance. *Practical assessment, research & evaluation,, 17*(9).

Ortiz, L., Tillerias, H., C., C., & Toaza, V. (2020). Impact on the video game industry during the COVID-19 pandemic. *Athenea Engineering sciences journal, 1*(1), 5-13.

Pérez, C., & Santín, D. (2008). Minería de datos. Técnicas y herramientas. *Madrid: Paraninfo*.

Senaviratna, N. A., & A Cooray, T. M. (2019). Diagnosing multicollinearity of logistic regression model. *Asian Journal of Probability and Statistics, 5*(2), 1-9.

Siemsen, E., Roth, A., & Oliveira, P. (2010). Common method bias in regression models with linear, quadratic, and interaction effects. *Organizational research methods, 13*(3), 456-476.

Stanton, J. M. (2001). Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education, 9*(3).

Stock, J. H., Watson, M. W., & Larrión, R. S. (2012). *Introducción a la Econometría.* (3ª ed.). Madrid: Pearson.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society,* 817-838.

Wirtz, B. (4 de Enero de 2023). *Game Designing*. Obtenido de The Complete Guide to Video Game Genres: From Scrollers, Shooters, to Sports: https://www.gamedesigning.org/gaming/video-game-genres/

Zackariasson, P., & Wilson, T. L. (2010). Paradigm shifts in the video game industry. *Competitiveness Review: An International Business Journal*.

9. ANEXOS

En este apartado no se ha dado una página por apartado debido a la brevedad de alguno de estos, y para mantener los anexos lo más compacto posible.

A. Ecuaciones completas de modelo final y alternativo

Modelo Final

$$\log(\boldsymbol{Sales}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 Release} + \boldsymbol{\beta_2} \log(\boldsymbol{Twitter}) + \boldsymbol{\beta_3 Vandal} + \boldsymbol{\beta_4 PlayStation} + \boldsymbol{\beta_5 Nintendo_{Switch}} + \boldsymbol{\beta_6 PC} + \boldsymbol{\beta_7 Xbox}$$

Modelo Alternativo

$$\log(\boldsymbol{Sales}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 Release} + \boldsymbol{\beta_2} \log(\boldsymbol{Twitter}) + \boldsymbol{\beta_3 Vandal} + \boldsymbol{\beta_4 PlayStation} + \boldsymbol{\beta_5 Nintendo_{Switch}} + \boldsymbol{\beta_6 PC} + \boldsymbol{\beta_7 Xbox} + \boldsymbol{\beta_8 Vegetta777}$$

B. Código consola GRETL

Producir una variable con el retardo de los residuos:

Residuos_ret=ResiduosMF(-1)

Realización de matriz de correlaciones y cálculo de su determinante:

matrizvariables = {Release, Main_Platform, Vandal,l_Twitter}

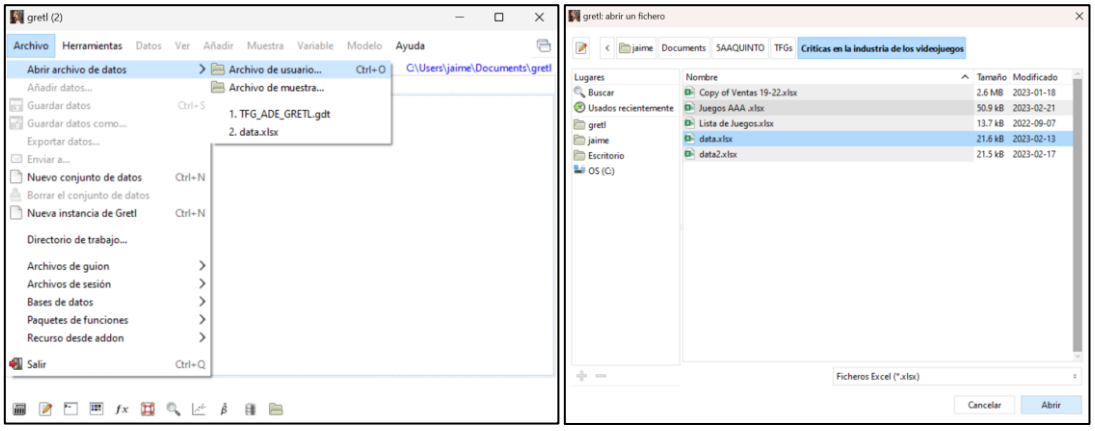
matrizcorrelaciones=mcorr(matrizvariables)

determinante=det(matrizcorrelaciones)

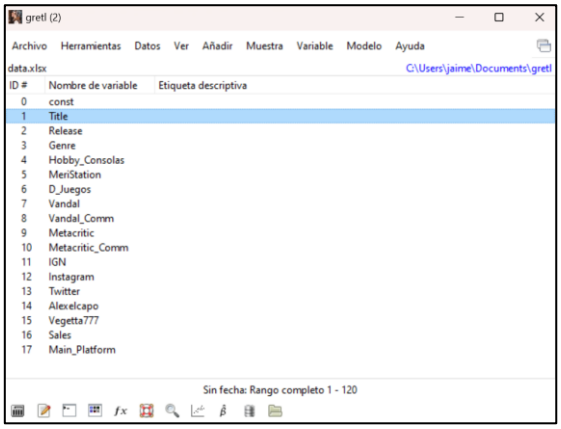
C. Guía Imágenes GRETL

Abrir el fichero de datos:

Archivo – Abrir archivo de datos – Archivo de usuario: Buscar los datos dentro de nuestros archivos

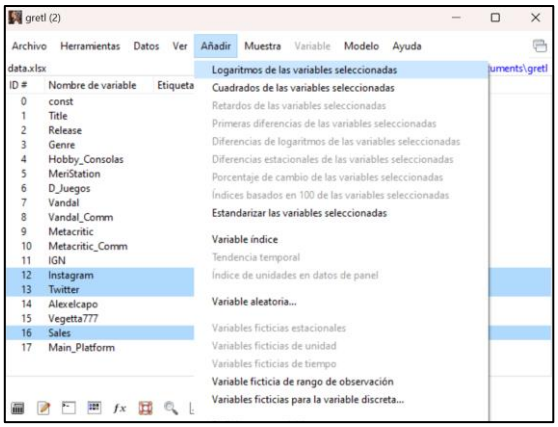


En este caso se obtendrán las siguientes variables:

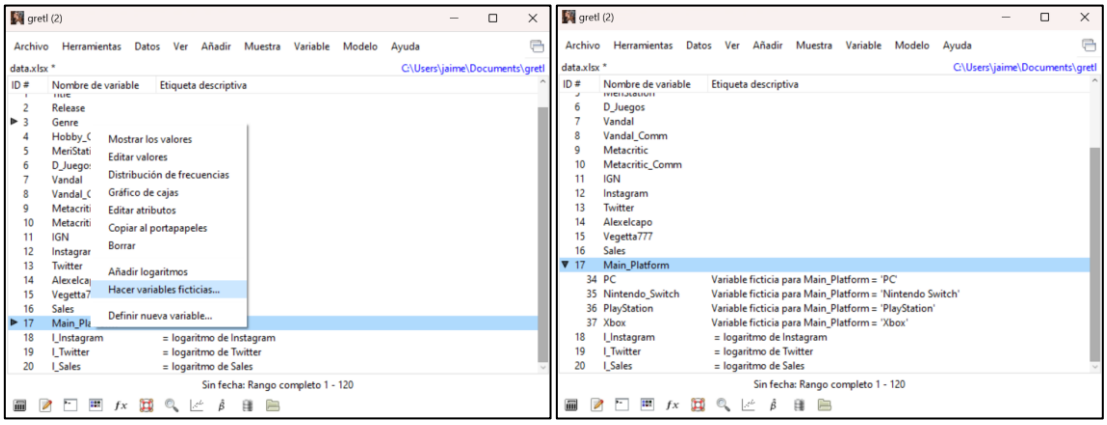


Transformaciones de variables:

Transformación logarítmica (Añadir logaritmos de variables):

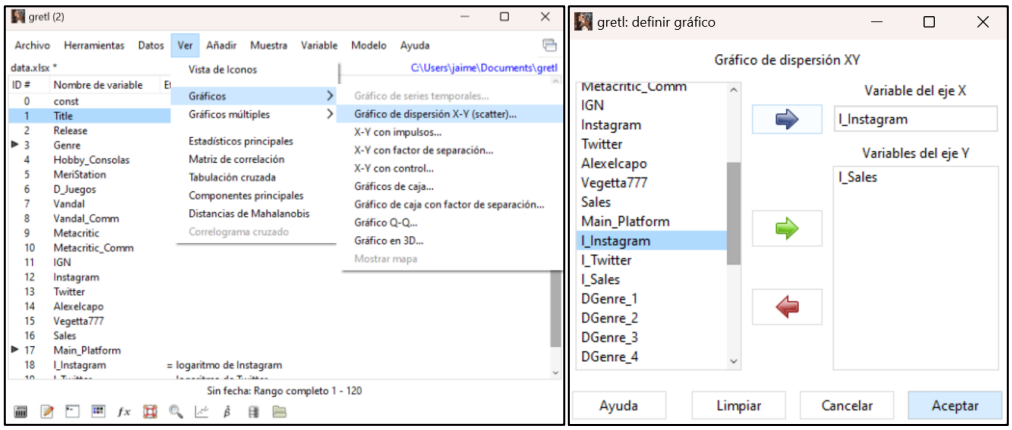


Separar las categorías de una variable (similar al One-Hot Encoding hecho en el código de R):



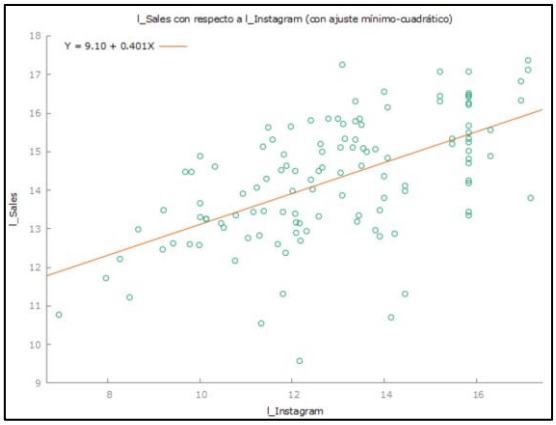
Si se desea cambiar el nombre de las variables (puesto que se generan automáticamente con nombres codificados): clic derecho y 'editar atributos'

Gráficos de dispersión (Scatter Plots) u otros gráficos:



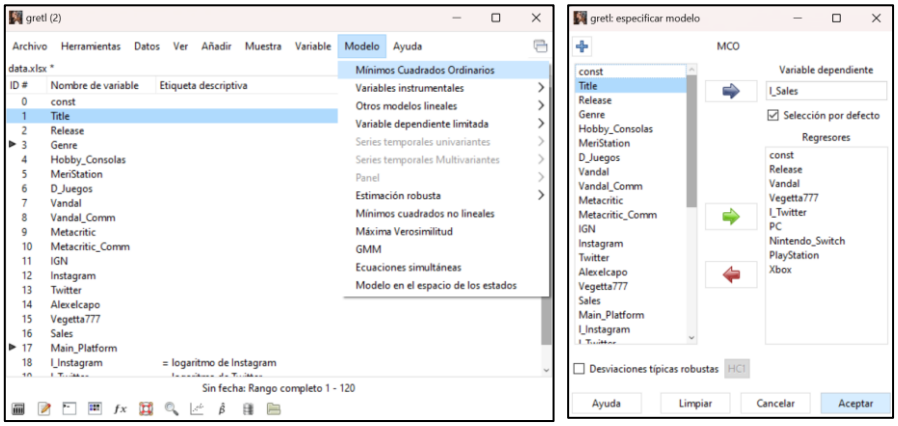
Se insertan las variables que se buscan graficar con las flechas azul (eje x) y verde (eje y).

El resultado es el siguiente gráfico:



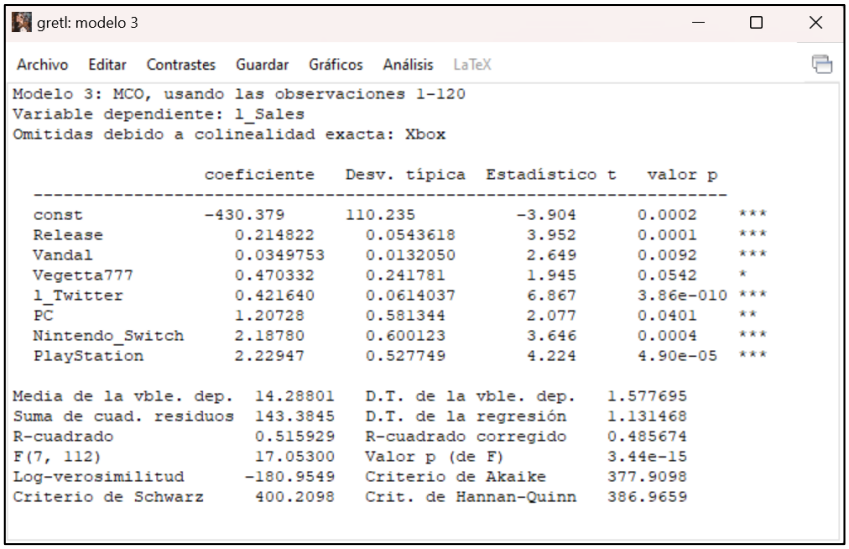
También se puede realizar esta función con el icono de gráfico de dispersión del menú principal, abajo en el centro.

Modelo de regresión lineal



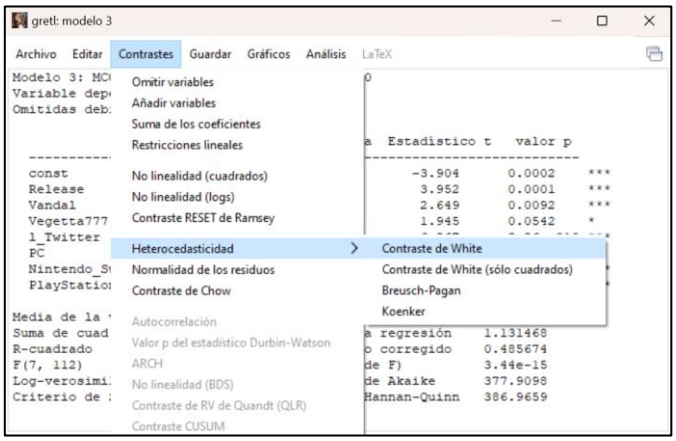
También se puede realizar esta función con el icono de la beta estimada ($\hat{\beta}$) del menú principal, abajo en el centro.

Se insertan la variable a explicar u objetivo con la flecha azul y las variables independientes con la flecha verde. Esto produce el siguiente modelo:

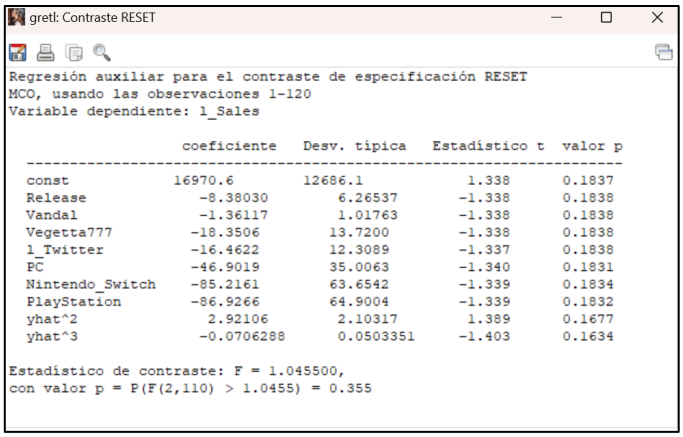


Contrastes

Una vez desarrollado el modelo se deben realizar una serie de contrastes para comprobar su validez. Estos se realizan de la siguiente forma:

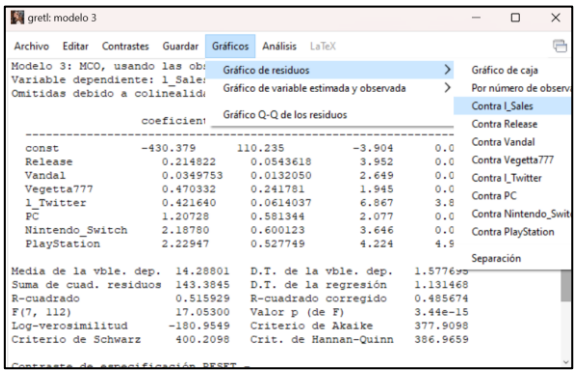


En este caso se realizaría el contraste de heterocedasticidad, pero también se podrían realizar el de normalidad de los residuos, el test de Chow, el test de Ramsey, etc. Todos estos contrastes despliegan pestañas nuevas con resultados detallados, como, por ejemplo, si se hiciese el test de Ramsey de cuadrados y cubos:

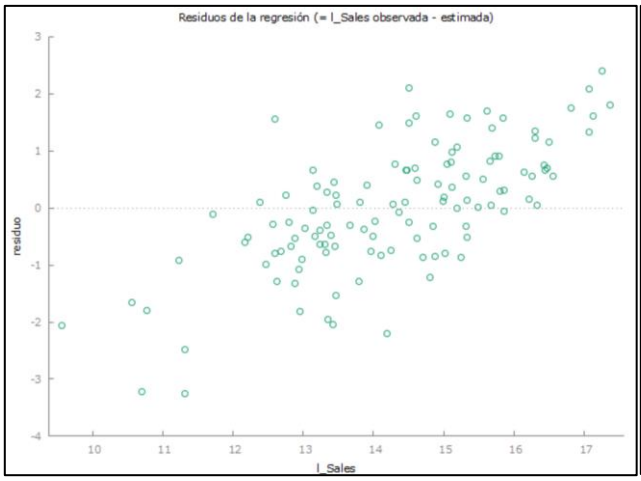


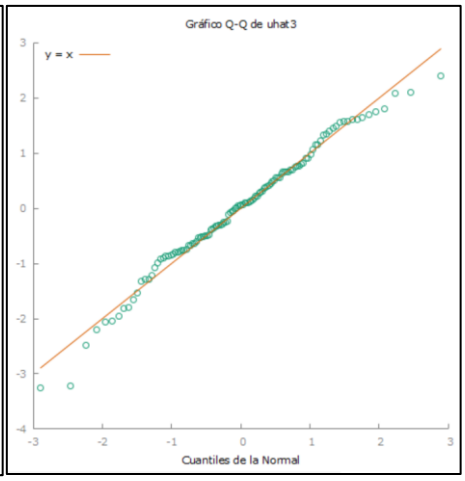
El dato más importante de estos contrastes es el p-valor, que al ser mayor al 0,05 permite que se rechace la hipótesis nula (en este caso esto significa pasar el test, es un resultado positivo). En el caso concreto del test de Ramsey se puede observar que se calculan p-valores independientes confirmando que, según este test, no hay ninguna variable fundamental omitida.

Para ciertos contrastes también es importante realizar gráficos, estos se realizan de la siguiente forma:

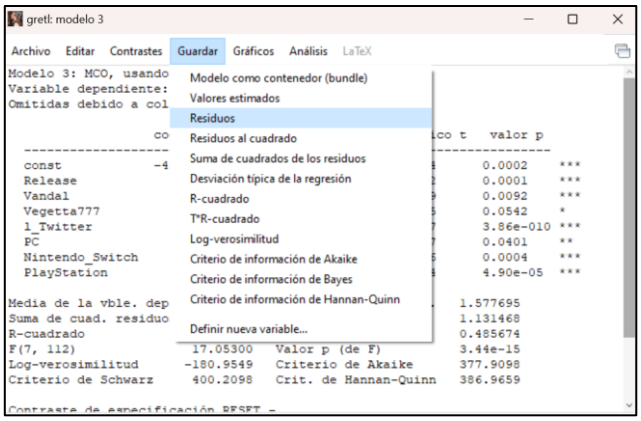


Algunos gráficos que se pueden elaborar son el gráfico de residuos vs la variable objetivo o el gráfico Q-Q de los errores.





Si se desean guardar los residuos del modelo para otras comprobaciones:



Esto supone un resumen de las técnicas utilizadas para la elaboración de este proyecto, pero existen numerosas otras funcionalidades.

D. Código consola R Studio

El código elaborado incluye una regresión lineal múltiple con carácter explicativo y predictivo, una partición de los datos, un proceso de cross-validación y otras técnicas predictivas como una red neuronal. Ha sido desarrollado gracias a la formación recibida en la asignatura Machine Learning I del grado de Business Analytics.

```
#Paquetes:
install.packages("caret")
install.packages("caretEnsemble")
install.packages("writexl")
install.packages("nnet")
install.packages("NeuralNetTools")
install.packages("GGally")
install.packages("ggcorrplot")
install.packages("ggpubr")
library(caret)
library(caretEnsemble)
library("writexl")
library("readxl")
library(NeuralNetTools)
library(nnet)
library(ggplot2)
library(GGally)
library(ggcorrplot)
library(ggpubr)


#TRATANDO EL DATASET-----------------------------------------
#1.Obtenemos el dataset principal
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
getwd()
data <- read_excel("data.xlsx")
data2 <- read_excel("data2.xlsx") #Para comprobar performance sin
GTA:V


#2.Elaboracion de la tabla de variables:
#Miramos los tipos de variables
class(data$Main_Platform)
```

```
class(data$Sales)
summary(data$Sales)

#3.Análisis de data:
#Para la correlacion las variables tienen que ser numericas
datanum <- data[,sapply(data, is.numeric)]
datanum$logSales<-log(datanum$Sales)
cor(datanum)
cov(datanum)
pairs(datanum, pch=19)
ggplot(data,            aes(x=Release,            y=log(Sales)))+
ggtitle("Transformando    logaritmicamente    las    Sales")    +
geom_point(size=4,
color='lightpink')+geom_smooth(method=lm,se=FALSE,
linetype="dashed", color="firebrick1")+stat_regline_equation()
corr <- round(cor(datanum), 1)
ggcorrplot(corr,lab=TRUE, title='Gráfico de correlación entre las
variables numéricas')
ggpairs(data, columns = 4:11, aes(color = Main_Platform,alpha =
0.5))
plot(data$Release,data$Sales, col="red" )


#MODELO EXPLICATIVO-----------------------------------------------
#One-Hot Encoding:
#queremos convertir position en una variable factor:
data$Main_Platform<-factor(data$Main_Platform)
summary(data$Main_Platform)

data$Genre<-factor(data$Genre)
summary(data$Genre)

#para cambiar la categoria de referencia:
data$Main_Platform<-relevel(data$Main_Platform,ref="Xbox")

#comprobamos si hay data perdidos
any(is.na(data))

#Para comprobar si podemos hacer log de variables con asimetr?a
hacia la derecha:
any(data$Twitter==0)
any(data$Instagram==0)
any(data$Sales==0)

modelexp1<-
lm(log(Sales)~log(Release)+log(Hobby_Consolas)+log(IGN)+log(Vand
al)+log(Vandal_Comm)+log(Metacritic)+log(Metacritic_Comm)+log(Me
riStation)+log(`3D_Juegos`)+log(Twitter)+log(Instagram)+Alexelca
po+Vegeta777+Main_Platform
          ,data=data)


modelexp2<-
lm(log(Sales)~log(Release)+Hobby_Consolas+Metacritic+log(Metacri
tic_Comm)+log(Twitter)+log(Instagram)+Vegeta777+Main_Platform  +
Genre
```

```
          ,data=data)

modelexp3<-
lm(log(Sales)~Release+log(Twitter)+Main_Platform+Vandal
              ,data=data)
summary(modelexp3)



#PARTICION-----------------------------------------------------
RNGkind("Super", "Inversion", "Rounding")
# si rounding da mensaje de error (que no warning), quitarlo y
ejecutar solo con super e inversion
set.seed(123)
index<-createDataPartition(data$Sales, p = 0.7, list=FALSE)
train<- data[index,]
test<-data[-index,]

#MODELOS PREDICTIVOS-------------------------------------------
#Regresion lineal_____________________________________________
modeloreg<-
lm(log(Sales)~Release+Main_Platform+log(Twitter)+Vandal,data=tra
in)
summary(modeloreg)
#Obtenemos las predicciones
predicciones<-predict(modeloreg, newdata = test)
#Obtenemos los erroes
errores<-test$Sales-predicciones
boxplot(errores)
hist(errores, breaks=20)
h<-nrow(test)
#Obtenemos las medidas de comparaci?n
ME<-sum(errores)/h
RMSE<-sqrt(sum(errores^2)/h)
AIC(modeloreg)
BIC(modeloreg)

# REGRESION LINEAL CROSS VALIDATION___________________________
#entrenamiento de la regresion (si no funciona intentar correr la
red neuronal y despues este modelo)
RNGkind("Super", "Inversion", "Rounding")       # para tener un
valor de referencia (la misma semilla)
set.seed(123)
preProcessRangeModel<-preProcess(train, method=c("range"))
trainproc<-predict(preProcessRangeModel, train)
control<-trainControl(method="repeatedcv", number=10, repeats=3)
summary(trainproc)
reg<-train(log(Sales)~Release+Main_Platform+log(Twitter)+Vandal,
data=trainproc, method="lm", trControl=control)
reg

# obtenemos RMSE de cv
RMSEcvREG<-reg$results[,2]

# importancia de las variables
```

```
varImp(reg)

#predicciones en training set
predtrainprocREG<-predict(reg, newdata=trainproc)
RMSEtrREG<-RMSE(predtrainprocREG, trainproc$Sales)

#predicciones en testset
predtestprocREG<-predict(reg, newdata=testproc)
RMSEtestREG<-RMSE(predtestprocREG, testproc$Sales)

#deshacemos la transformacion de range para obtener predicciones
en escala original
predtestREG<-predtestprocREG*(maxsal-minsal)+minsal
head(predtestREG)
head(test$Sales)

# comparamos metrica de ajuste en cv, train y test
RMSEcvREG
RMSEtrREG
RMSEtestREG
#Para transformar de vuelta:
RMSEtestREG*(maxsal-minsal)+minsal
hist(data$Sales)


#NEURAL NETWORK_____________________________________________

#Procesamos los data para que esten en rango 0-1
preProcessRangeModel<-preProcess(train, method=c("range"))
trainproc<-predict(preProcessRangeModel, train)
summary(trainproc)

#Parametros de cross validacion
control<-trainControl(method="repeatedcv", number=10, repeats=3)

#hiperparametros a optimizar con nnet
modelLookup("nnet")

#Creamos grid de combinaciones de hiperpar?metros
grid<-expand.grid(size=c(3,4,5,6), decay=c(0.01))

#Entrenamiento de la red
RNGkind("Super", "Inversion", "Rounding")
set.seed(123)
net<-train(Sales~Release+Twitter+Main_Platform+Vandal,
data=trainproc, method="nnet", trControl=control, tuneGrid=grid)
net
plot(net)
RMSEcv<-min(net$results[,3])

#Grafico de la red
plotnet(net, pos_col="purple", neg_col="blue")+title("Modelo de
Red Neuronal")

# importancia de las variables
garson(net)
```

```
varImp(net)
# sensitivity analysis
olden(net)

#predicciones en training set
predtrainproc<-predict(net, newdata=trainproc)
RMSEtr<-RMSE(predtrainproc, trainproc$Sales)

#predicciones en testset
#Tenemos que preprocesar el test set
minsal<-min(test$Sales)
maxsal<-max(test$Sales)
testproc<-predict(preProcessRangeModel, test)
predtestproc<-predict(net, newdata=testproc)

RMSEtest<-RMSE(predtestproc, testproc$Sales)

#deshacemos la transformacion de range para obtener predicciones
en escala original
predtest<-predtestproc*(maxsal-minsal)+minsal
head(predtest)
head(test$Sales)
# comparamos metrica de ajuste en cv, train y test
RMSEcv
RMSEtr
RMSEtest
#Para transformar de vuelta:
RMSEtest*(maxsal-minsal)+minsal
```