MÁSTER DE ESTADÍSTICA APLICADA

CON R SOFTWARE

Análisis estadístico de los ingresos en videojuegos

AUTOR: Daniel Barandiarán Torres

DIRECTOR: Juan Luis López Garrancho

FECHA: 30/08/2022

Resumen:

Toda inversión al parecer da sus frutos, y en este estudio se ha podido comprobar esto.

En este estudio se ha procedido a un análisis estadístico sobre 40 juegos más populares con unas variables económicas y de valoración, así como descriptivas. La base de datos se ha obtenido de una única fuente debido a la dificultad de encontrar información referente.

Con esta base de datos se ha realizado un análisis de componentes principales y a una agrupación jerárquica de estos componentes principales, con el fin de saber que grupo ha obtenido mayores ingresos y por qué. Encontramos 3 grupos en el análisis, y el grupo en el que más ingresos ha obtenido coincide con la mayoría de juegos AAA, que son los que más inversión tienen.

Palabras clave: Videojuegos; base de datos; técnicas estadísticas; análisis de componentes principales PCA; agrupación jerárquica de componentes principales (HCPC).

Summary:

Every investment seems to pay off, and this study has proven this.

In this study, a statistical analysis was carried out on 40 of the most popular games with economic and valuation variables, as well as descriptive variables. The database was obtained from a single source due to the difficulty of finding reference information.

With this database, a principal component analysis and a hierarchical grouping of these principal components was carried out, in order to find out which group has obtained the highest revenue and why. We found 3 groups in the analysis, and the group with the highest revenue coincides with the majority of AAA games, which are the ones with the highest investment.

Keywords: Video games; database; statistical techniques; principal component analysis (PCA); hierarchical clustering of principal components (HCPC).

Índice de Contenidos

Índice de Ilustraciones

Índice de Tablas

1. Introducción y objetivos

Esta primera sección se divide en dos apartados: motivación, donde se explica el motivo de la investigación, así como su relevancia; la hipótesis, donde que enuncia lo que se pretende conseguir con el proyecto; y los objetivos, donde se enuncian los antecedentes, el objetivo general y específicos para llevar a cabo el proyecto, y la hipótesis.

1.1. Motivación

Este proyecto ha sido elaborado gracias a conocimientos adquiridos durante el máster.

El tema de la investigación trata del comercio de los videojuegos, el cual es un mercado en auge y que cada año aumenta los beneficios.

La industria de los videojuegos se puede decir que comienza en la década de los 70, lo cual es un mercado joven. Fue desarrollándose junto con las nuevas tecnologías hasta el día de hoy, donde puedes jugar a juegos con una enorme cantidad de datos desde el dispositivo móvil.

Durante la epidemia que nos tuvo confinados en nuestros hogares, este mercado tuvo un fuerte impacto en cuanto a los ingresos que se generaron, fue de los pocos sectores que no solo no se desgastaron, si no que creció más. Según un artículo de NewZoo sobre un análisis del mercado, "El mercado de juegos en 2021 generará ingresos totales de 180.3 mil millones de dólares, un aumento del + 1.4% con respecto a 2020" [1].

Por tanto, la motivación que me ha llevado a escoger este tema conocer un poco más cómo funciona el mercado de los videojuegos en cuanto a los ingresos, así como descubrir si existen agrupaciones de juegos que nos ayuden a entender más este asunto.

Esta investigación puede ayudar también a conocer mejor patrones de compra o simplemente ampliar el marco del conocimiento. Aunque existe cierta limitación con los datos, se puede obtener conclusiones interesantes al respecto.

1.2. Hipótesis

En este trabajo se trata de analizar los factores que pueden influir significativamente en los ingresos de los videojuegos.

A partir del análisis de las variables se podrá contextualizar los análisis posteriores de agrupación.

A partir del análisis de la agrupación podremos saber si existen ciertas condiciones que determinen los ingresos de los videojuegos.

1.3. Objetivos

En cuanto a los objetivos, el objetivo general del proyecto es determinar qué agrupaciones obtenemos con las variables que disponemos para poder averiguar cuáles de los juegos obtienen mayores ingresos y por qué.

Los objetivos para llevar a cabo el proyecto son:

- Examinar el mercado de los videojuegos.
- Comparar géneros y clases de productores de videojuegos.
- Recopilar los datos de las variables.
- Contrastar datos entre fuentes.
- Introducir datos en una base de datos.
- Definir las variables establecidas.
- Analizar las variables de forma individual (univariante).
- Analizar las variables por parejas de dos con las variables principales (bivariante).
- Analizar las variables de forma múltiple (multivariante).
- Definir grupos según las variables establecidas.
- Desarrollar conclusiones en base al análisis de los datos.

1.4. Estructura del proyecto

El proyecto se estructura de la siguiente manera:

Metodología y técnicas: empieza con una breve introducción sobre los antecedentes, con un resumen de la evolución del mercado en números. En cuanto a la metodología y técnicas, en ellas se explican los procesos que se han llevado a cabo para el proyecto, así como las herramientas y técnicas estadísticas aplicadas.

Resultados: aquí se hallan varios apartados. Se comienza con una descripción de todas las variables utilizadas en el proyecto para comprender adecuadamente la interpretación de los resultados. Seguidamente se analizan las variables individualmente para encontrar matices y puntos anómalos. Después de analizarse una a una, se procede por analizarlas por parejas, con el objetivo de encontrar variables que se relacionen con las variables de interés y poder interpretar. Por último, se analizan todas las variables conjuntamente con el fin de obtener agrupaciones que nos ayuden a entender las características que más influyen a cada grupo.

Discusión: aquí se trata de interpretar los resultados con el fin de responder a la finalidad del proyecto.

Conclusiones: en estos apartados se interpretan los resultados obtenidos más relevantes. Además de dar pie a la realización de otros proyectos similares a quienes les sirva de motivación o utilidad como futuras líneas de investigación.

2. Metodología y Técnicas

En esta sección se explica la metodología aplicada para la investigación. Consta de dos apartados: metodología, donde se presenta el procedimiento seguido en el proyecto; y técnicas, donde se exponen las herramientas utilizadas con el programa R.

2.1. Metodología

En primer lugar, se debe escoger qué factores pueden ser significativos a la hora de analizar los ingresos de un juego.

Una vez escogidas las variables, se procede a la obtención de los datos de los distintos juegos. 40 juegos se analizarán sobre estas variables, en concreto se han escogido los que más reseñas tienen, con intención de escoger los juegos más conocidos.

Se trata de llevar a cabo un análisis de las variables de forma individual y de forma colectiva para averiguar si existe relación o si son influyentes sobre la variable principal (número de ingresos). Seguidamente se realizará un análisis de componentes principales (PCA) para resumir los datos e identificar si existen agrupaciones de juegos mediante la agrupación jerárquica de componentes principales (HCPC) y describirlos.

A modo de esquema, se muestra la consecución de tareas para llevar a cabo el análisis de las variables y el PCA.

- Planteamiento y selección de variables.
- Construcción de la base de datos (sucia).
- Depuración de la base de datos (limpia).
- Análisis univariante.
- Análisis bivariante.
- Análisis multivariante (PCA y HCPC).

A continuación, se desglosa los apartados anteriores del análisis, mostrándose los análisis que se han realizado en concreto:

Análisis univariante:

- Tabla de descriptivos.
- Tabla de frecuencia.
- Diagrama Cajas y Bigotes.

Análisis bivariante:

- Gráfico GGpairs: que se compone de matriz de correlaciones, gráficos XY, gráficos de densidad, y gráficos de caja y bigotes entre variables cuantitativa y cualitativa.

Análisis multivariante:

- Análisis de componentes principales.
- Agrupación jerárquica de componentes principales.

2.2. Técnicas

2.2.1. Técnicas aplicadas con “R”

Para el análisis univariante se utilizarán las siguientes técnicas estadísticas para el análisis exploratorio:

- Medidas estadísticas básicas: mínimo; 1er cuartil; mediana; media; 3er cuartil; y máximo.
- Tabla de frecuencias: indica la repetición de los valores de una variable.
- Gráfico de caja y bigotes: son una presentación visual que describe varias características importantes, al mismo tiempo, tales como la dispersión y simetría. Consiste en una caja rectangular, donde los lados más largos muestran el recorrido intercuartílico. Este rectángulo está dividido por un segmento vertical que indica donde se posiciona la mediana y por lo tanto su relación con los cuartiles primero y tercero (el segundo cuartil coincide con la mediana). Esta caja se ubica a escala sobre un segmento que tiene como extremos los valores mínimo y máximo de la variable. Las líneas que sobresalen de la caja se llaman bigotes, el resto de los datos o casos que no se encuentre dentro de este rango es marcado e identificado individualmente (puntos anómalos) [2].

En el análisis bivariante se utilizan las matrices de correlación para determinar la influencia de unas variables respecto a otras. Además, las variables de mayor correlación se analizarán más específicamente a través de Gráficos XY, que muestra en una gráfica los valores en base a dos variables, eje X y eje Y. Y por último, también se utiliza el Diagrama Caja y Bigotes clasificando la variable según los valores de otra variable cualitativa. Todo esto para conocer la relación existente entre las unidades de ventas y los ingresos de los videojuegos respecto a las demás

Finalmente, para el análisis multivariante se utilizan las técnicas de PCA y HCPC:

- El PCA es un método que permite resumir y visualizar la información de un conjunto de datos que contiene múltiples variables cuantitativas correlacionadas. Permite extraer información importante del conjunto de datos multivariante y expresarla como nuevas variables llamadas componentes principales. Estas nuevas variables son una combinación lineal de las originales, con lo cual no estarán correlacionadas entre sí. Al seleccionar un número de componentes principales menor que el número de variables originales, logramos reducir (resumir) la información de los datos y podremos representarlas (visualizar) en un espacio de menor dimensión.

- El HCPC es un método de agrupación jerárquica que consiste en el siguiente algoritmo: 1. Realiza un PCA para reducir el número de variables y eliminar el ruido en los datos; 2. Realiza una agrupación sobre los componentes principales seleccionados para crear

grupos a partir del árbol de agrupación (dendrograma); 3. Caracteriza los grupos para interpretar la agrupación.

Una vez agrupadas podremos concluir que individuos son similares entre sí y por qué.

2.2.2. Herramienta estadística CART con “R”

Para ejecutar las herramientas y realizar los cálculos para los análisis se usa en este trabajo el programa “R”, que es un software estadístico que permite manejar y analizar datos.

R es un ambiente de programación formado por un conjunto de herramientas muy flexibles que pueden ampliarse fácilmente mediante paquetes, librerías o definiendo nuestras propias funciones. Además, es de código abierto [3].

Esto quiere decir que las operaciones se realizan a través de instrucciones en un terminal, por lo que no es un programa gráfico. Al ser un lenguaje de programación, tenemos la posibilidad de ampliar su funcionalidad en la medida de nuestros conocimientos.

RStudio es una aplicación web que permite desarrollar con R y otros lenguajes de programación orientados al tratamiento de grandes cantidades de datos, estadísticas, etc. Es todo un completo IDE de desarrollo, pero embutido en una aplicación web, que permite además integrarse con una serie de herramientas enfocadas en la gestión de proyectos [4].

El PCA es una técnica ampliamente utilizada en investigación que permite identificar la dimensión intrínseca de los datos, que es:

- El número de componentes principales que se necesitan para aproximar un conjunto de datos.

- El número de componentes principales que explican una varianza significativa en los datos.

- Y la representación más compacta de un conjunto de datos.

El PCA se puede utilizar como un paso de preprocesamiento antes de realizar métodos de agrupación, con el fin de eliminar el ruido de los datos y que la agrupación sea más estable que la obtenida de las distancias originales.

Lo más interesante para eliminar el ruido sería realizar la agrupación sobre los primeros componentes principales. Las primeras dimensiones extraen lo esencial de la información, mientras que los últimos se limitan a recoger el ruido, por tanto, sin ruido en los datos, la agrupación será más estable.

Para realizar un PCA en R, vamos a utilizar la función PCA() del paquete *FactoMineR*, y para obtener gráficos avanzados utilizaremos el paquete *factoextra*.

En el caso del HCPC es igual, vamos a utilizar la función HCPC() del paquete FactoMineR, y para los gráficos avanzados utilizaremos el paquete *factoextra*.

3. Resultados y Discusión

En la siguiente sección se describen una serie de variables objeto de estudio para el análisis posterior.

Seguidamente se realiza un análisis univariante, bivariante y multivariante de dichas variables: con el primer análisis se pretende conocer características de las variables que puedan ofrecer información para el estudio; con el segundo se pretende conocer la relación existente entre dos variables, así como el comportamiento conjunto, y con el tercer análisis se pretende realizar un PCA y HCPC de las variables de estudio. Con la obtención de los resultados, se realizará las conclusiones de cada uno de ellos para responder a nuestros objetivos.

3.1. Descripción de variables

Las variables escogidas para los análisis y planteamiento del modelo, así como su abreviatura y definición, son las siguientes:

- Nombre (NOM): nombre del videojuego.
- Ingresos (ING): ingresos generados en millones de dólares.
- Precio (PREC): precio en dólares.
- Unidades (UNID): millones de unidades vendidas.
- Reseñas (RES): número de reseñas.
- Valoración (VAL): valoración por el público del 1 al 100.
- Género (GEN): clase de género. Se clasifican en: Action; Adventure; Simulation; RPG; y Strategy.
- Clase de editor (CLAS): es una clasificación informal utilizada para los videojuegos producidos. Se clasifican en:
    - AAA: cuando el editor es de gran tamaño, conocido y con gran presupuesto.
    - AA: cuando el editor es apenas conocido con un presupuesto medio.
    - Indie: editores independientes, de tamaño reducido y bajo presupuesto.

En cuanto a las dos últimas variables, se trataron de obtener de la forma más equitativa posible.

La base de datos fue obtenida de VG Insights [5], de donde se tuvo que limpiar las variables de signos y comas. La variable género se tuvo que resumir a sólo un género, ya que se componen de varios géneros algunos de los juegos.

3.2. Análisis univariante

A continuación, se va a proceder a calcular las medias, y demás indicadores objeto de análisis de las variables, que se plasmarán en la tabla 1. Podemos observar los resultados en la siguiente tabla:

Tabla 1. Resumen estadístico de las variables numéricas.

Variable	Mínimo	1er cuartil	Mediana	Media	3er cuartil	Máximo
ING	7.4	94.42	138.15	162.15	223.65	562.90
PREC	3.99	9.99	19.99	25.29	39.99	59.99
UNID	2.90	6.975	10.60	12.232	13.725	50.800
RES	181121	222573	310164	404352	539041	1067195
VAL	50.40	88.45	94.25	90.38	97.03	98.80

En las siguientes tablas se muestran dos variables que son cualitativas, por lo que en este caso veremos los datos que se repiten de cada variable.

Tabla 2. Variable Género del videojuego (GEN).

Género	Número de coincidencias
Strategy	4
Simulation	6
RPG	9
Adventure	8
Action	13

Tabla 3. Variable Clase de productor (CLAS)

Clase	Número de coincidencias
Indie	10
AA	15
AAA	15

En cuanto a las variables numéricas vemos entonces que hay una gran diferencia entre los mínimos y los máximos de cada variable. Y en cuanto variables categóricas vemos que están en gran medida bien distribuidas por cada nivel.

Para observar si existen puntos anómalos de cada variable vamos a realizar un diagrama de cajas y bigotes para cada variable numérica. Y ya que esto será un paso posterior para la realización del PCA, vamos a escalar las variables.

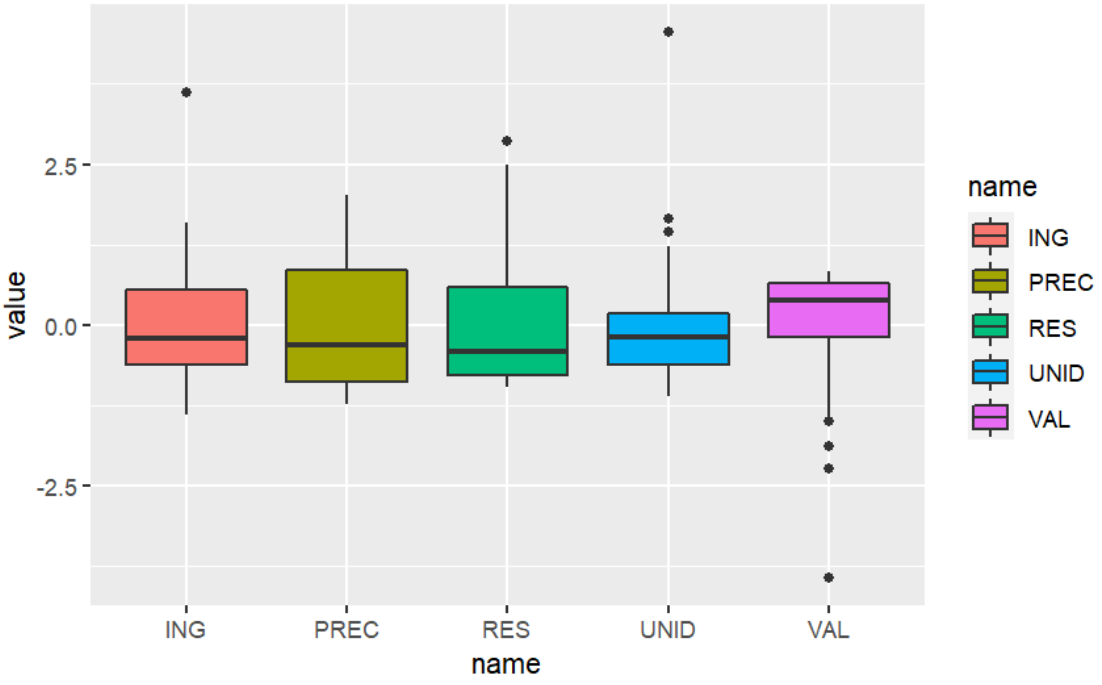

Ilustración 1. Gráfico de cajas y bigotes de las variables cuantitativas.

Vemos que existen varios puntos anómalos, por lo que veremos si coincide la misma observación en otras variables.

Los puntos anómalos que encontramos en cada variable son:

- Ingresos: Cyberpunk 2077.
- Reseñas: Tom Clancy's Rainbow Six Siege.
- Unidades: Left 4 Dead 2, Payday 2, y Portal 2.
- Valoración: Cyberpunk 2077, DayZ, New World, Tale of Inmortal.

Por tanto, podemos dejar estas observaciones y ver los resultados posteriores a ver si han afectado mucho, o podemos eliminar algunas de estas que se alejen mucho de la distribución de los datos.

3.3. Análisis bivariante

En el análisis bivariante se analiza la relación que tienen las variables entre sí. Utilizaremos un gráfico que resume de forma completa las relaciones entre variables. En él encontramos las correlaciones entre variables, gráficos de dispersión para variables numéricas, gráficos de cajas para variables categóricas y numéricas, gráficos de densidad y barras para ver las distribuciones.

El gráfico obtenido es el siguiente:

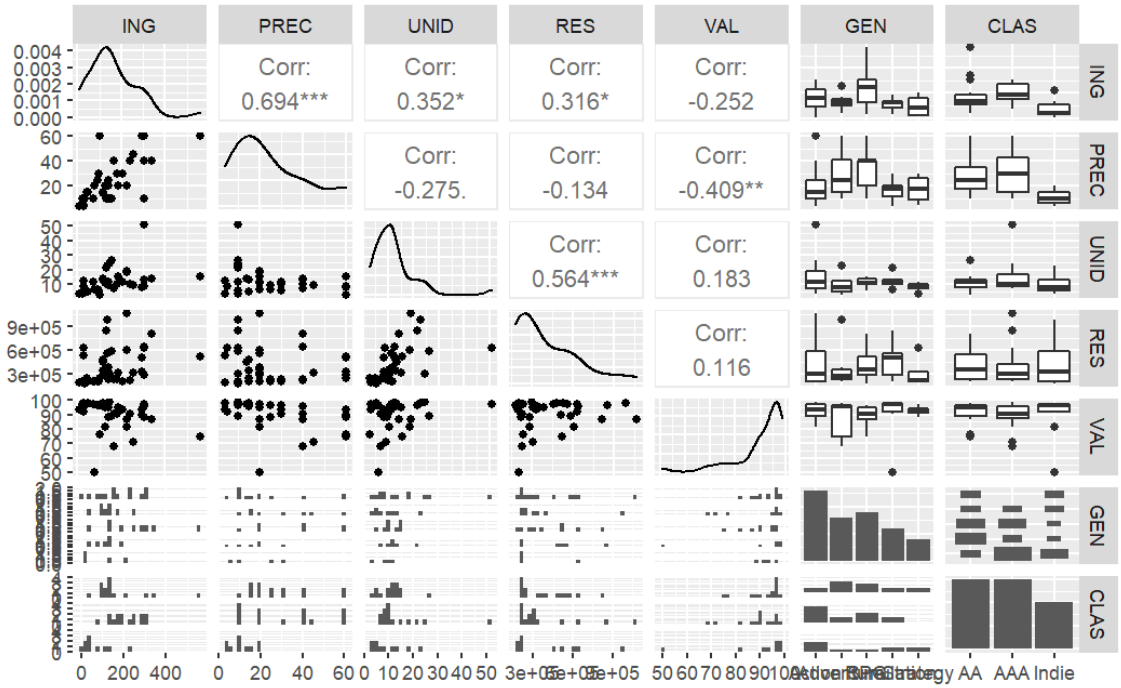

Ilustración 2. Gráfico ggpairs de librería GGally, donde se muestran gráficos de densidad, de dispersión, de barras, de cajas y bigotes, y matriz de correlaciones para todas las variables.

Empezaremos por comentar las distribuciones. En ellas vemos que existe asimetría positiva en las variables de ingreso, precio, unidades y reseñas, hay pocas observaciones con valores altos. Mientras que la variable valoración tiene asimetría negativa, muchas observaciones con valores altos.

En cuanto a las correlaciones, vemos que existen dos pares de variables con correlaciones moderadamente altas. Las variables precio e ingresos tienen una correlación de 0.69, lo cual nos indica que podría existir una relación entre las variables. Si nos fijamos en el gráfico de dispersión de estas dos variables sí que se puede ver cierta relación positiva. El precio en realidad no debería influir tanto en los ingresos si no es por las unidades vendidas, pero al ser los juegos más populares, debe ser que las unidades vendidas son más similares, y por tanto el precio influye bastante en cuanto a los ingresos. Si es cierto que hay cierta correlación también entre las variables unidades vendidas e ingresos, y reseñas e ingresos, en los cuales si que tiene más sentido que exista una relación positiva.

La segunda correlación más alta es entre las variables, reseñas y unidades vendidas. Esta sí que tiene más sentido, ya que cuanto más es conocido un juego es más probable que se vendan más unidades. Si nos fijamos en el gráfico de dispersión, si parece que haya una clara relación positiva entre las variables.

Por último, en las variables valoración y precio hay correlación negativa, esto tampoco tiene cierta coherencia. Si nos fijamos en el gráfico de dispersión, no se ve claramente que exista esa relación negativa, por lo que solamente se debe al cálculo numérico, además de que tampoco tendría sentido que existiese alguna correlación entre estas dos variables.

En cuanto a las variables categóricas, es interesante fijarnos en los grupos donde existe una gran diferencia con el resto. La primera gran diferencia que se puede observar es en el precio, donde los juegos indie tienen un precio mucho menor que el resto. Esto se debe a que los costes son mucho menores, y por tanto es permisible optar por un precio menor y obtener más unidades vendidas.

En cuanto a los ingresos, vemos que hay ligeramente un mayor número de ingresos en los juegos RPG y acción. Y respecto a la clase, vemos que los juegos de clase AA y AAA son los que mayores ingresos perciben, en ese orden.

3.4. Análisis multivariante

Una vez conocidos mejor los datos vamos a pasar a su preparación para poder realizar el PCA. Para ello necesitaremos:

- Un conjunto de datos con variables cuantitativas y correlacionadas.
- Variables con la misma escala. Si no habría que transformarlos
- Evitar los valores atípicos y los valores ausentes, ya que estos pueden alterar los resultados.
- El conjunto de datos ordenado.

Como tenemos algunas variables de distinta naturaleza, vamos a seleccionar un subconjunto de datos, el cuál llamaremos conjunto activo, y el resto como conjunto pasivo, que utilizaremos luego para predecir.

Las 5 primeras columnas del conjunto de datos son cuantitativas, por lo que utilizaremos estas como conjunto activo, mientras que las dos siguientes variables son cualitativas, y por lo tanto pasarán a ser suplementarias que posteriormente agregaremos al análisis para mejorar su interpretación.

En nuestro caso ya hemos visto que necesitamos escalar las variables (aunque la función del PCA lo hace automáticamente), por lo que vamos a ver si tenemos puntos anómalos ya que el hecho de que haya puntos anómalos en los descriptivos de variables no tiene necesariamente que producirlos en el ajuste del PCA. Vamos a comprobarlo de una forma muy sencilla de detectarlos usando distancias de mahalanobis.

A través de este método vemos que las filas 6 y 12 son puntos anómalos para el ajuste del PCA. Estas filas se corresponden con los juegos cyberpunk 2077 y left 4 dead 2, por lo que consideramos descartarlos de la base de datos para los análisis.

Como tenemos el conjunto de datos ordenado y ya hemos visto en el resumen de las variables que no tenemos valores ausentes, vamos a pasar a ver la correlación de las variables, ya que este es un paso importante para saber si realmente es posible realizar un PCA.

Para verificar si realmente es adecuado realizar un PCA para nuestros datos vamos a explorar la estructura de correlación de las variables. Existen 3 posibles métodos para su evaluación.

- La prueba de esfericidad de Bartlett evalúa si es necesario realizar el PCA, analizando si la correlación entre las variables analizadas es lo suficientemente grande como para justificar la factorización, o la reducción de la matriz de correlación. La hipótesis nula de esta prueba es que la matriz de correlación proviene de una población no colineal, o simplemente que no hay colinealidad entre las variables, lo que haría imposible el análisis de componentes principales ya que depende de la construcción de una combinación lineal de variables, sugiriendo que las variables no son una 'matriz identidad' en la que las correlaciones ocurren por un error de muestreo.

Para un nivel de significación del 5%, el P-valor obtenido es menor a 0.05. Es decir, tiene sentido utilizar el PCA.

Los inconvenientes de esta prueba es que asume que los datos poseen una distribución normal, y que es sensible al tamaño de la muestra, tiende a ser estadísticamente significativo cuando la muestra crece. Algunos autores advierten que se use cuando la razón entre el número de casos y número de variables sean menor a 5, y en nuestro caso es 7.6.

- La medida de adecuación muestral KMO (Kaiser-Meyer-Olkin) se basa en la varianza (información) común. Mide si existe un número apropiado de observaciones en relación con el número de variables que se evalúan. Hay una puntuación general y una por variable. Toma valores entre 0 (mala adecuación) a 1 (buena adecuación).

Tiene el mismo objetivo que el anterior, pero trata de averiguar si podemos factorizar, resumir, las variables de forma eficiente.

El punto de partida también es la matriz de correlación de las variables observadas. Sabemos que las variables pueden estar correlacionadas pero la correlación entre dos de ellas puede estar influenciado por otra. Así pues, utilizamos la correlación parcial para medir la relación entre dos variables eliminando el efecto del resto. El KMO compara los valores de las correlaciones entre las variables y sus correlaciones parciales. Hay unos niveles generales:

- De .5 a .7 son bajos

- De .7 a .8 son buenos

- De .8 a .9 son excelentes

- Mayor que .9 son más que excelentes

En nuestro ejemplo obtenemos un valor de 0.44, es decir un nivel menor que bajo.

En el caso de la prueba individual, cualquier valor por debajo de 0.5 debería eliminarse dicha variable y volver a proceder con la prueba.

En nuestro caso vemos que hay tres variables que obtienen un valor por debajo de 0.5. No obstante, vamos a proseguir con el ejemplo porque nos interesa observar los resultados que obtenemos.

- Prueba de positividad del determinante: evalúa la multicolinealidad. El resultado debería caer preferiblemente por debajo de 0.00001, aunque el requisito mínimo es que sea positivo y bajo.

Obtenemos un valor de 0.071, por lo que podemos afirmar que el determinante de la matriz de correlación es positivo y bajo. Vamos a suponer por tanto que hemos satisfecho todos los supuestos del PCA y procedemos con el análisis.

Una vez se han preparado de manera adecuada los datos para el PCA, vamos a ajustar el modelo, a visualizar las dimensiones y a describir las dimensiones seleccionadas tanto mediante estadísticos, como gráficos biplot. Los estadísticos de calidad de la representación y la contribución a las dimensiones nos ayudarán a interpretar los gráficos.

Con la función PCA creamos un objeto que contiene información sobre los valores propios (eigenvalues), sobre las variables (var), los individuos (ind), y sobre la llamada o la función que hemos utilizado (call). También nos dan información como son las coordenadas, la correlación con los ejes, el coseno cuadrado y la contribución.

Podemos ver un resumen completo del objeto creado en el código adjuntado en el anexo. En este resumen vemos los valores propios, que miden la cantidad de variación que retiene cada componente principal, y que con 5 dimensiones se logra explicar el 100% de la variación, pero que con tan solo 2 podemos explicar el 75%, la primera dimensión retiene mayor variación, y por lo tanto explica más, siguiendo en orden decreciente por el resto. También vemos un resumen de la contribución y calidad de las 10 primeras observaciones o individuos, así como de las variables.

El siguiente paso será seleccionar el número de dimensiones que nos queremos quedar.

Para determinar con cuántas dimensiones son adecuadas, no existe un único método, podemos utilizar:

- Regla de Keiser-Guttman. Consiste en tomar los valores propios mayores que 1. Ya que si es mayor a 1 indica que en promedio el componente principal da cuenta de más varianza que una de las variables originales en los datos estandarizados. Es decir, este criterio intenta retener los PC que expresan más variabilidad que cada una de las variables originales, retener aquellos que aporten información sustancial.

En nuestro ejemplo obtenemos la siguiente tabla:

Tabla 4. Valores propios y porcentajes de varianza explicada.

Dimensión	eigenvalue	Porcentaje de varianza	Porcentaje de varianza acumulada
Dim.1	1.93942512	38.788502	38.78850
Dim.2	1.85730136	37.146027	75.93453
Dim.3	0.78694253	15.738851	91.67338
Dim.4	0.34216041	6.843208	98.51659
Dim.5	0.07417058	1.483412	100.00000

Por tanto, según este criterio escogeríamos tan solo las dos primeras dimensiones.

- % Varianza explicada. considerar tantos componentes principales como el porcentaje de varianza total que queremos dar cuenta. Por ejemplo, si consideramos que un 70% de varianza total explicada es suficiente, podemos seleccionar para este ejemplo 2 dimensiones. Podemos observar esto en la tabla anterior en la columna de porcentaje de varianza explicada acumulada.

- Gráfico de pantalla o sedimentación. Realizar un gráfico con los valores propios en función de los PC y busca el "codo" del gráfico, el punto de inflexión, donde los valores propios parecen estabilizarse. Los PC a la izquierda de este punto se consideran significativos (Jollife 2002, Peres-Neto et al 2005).

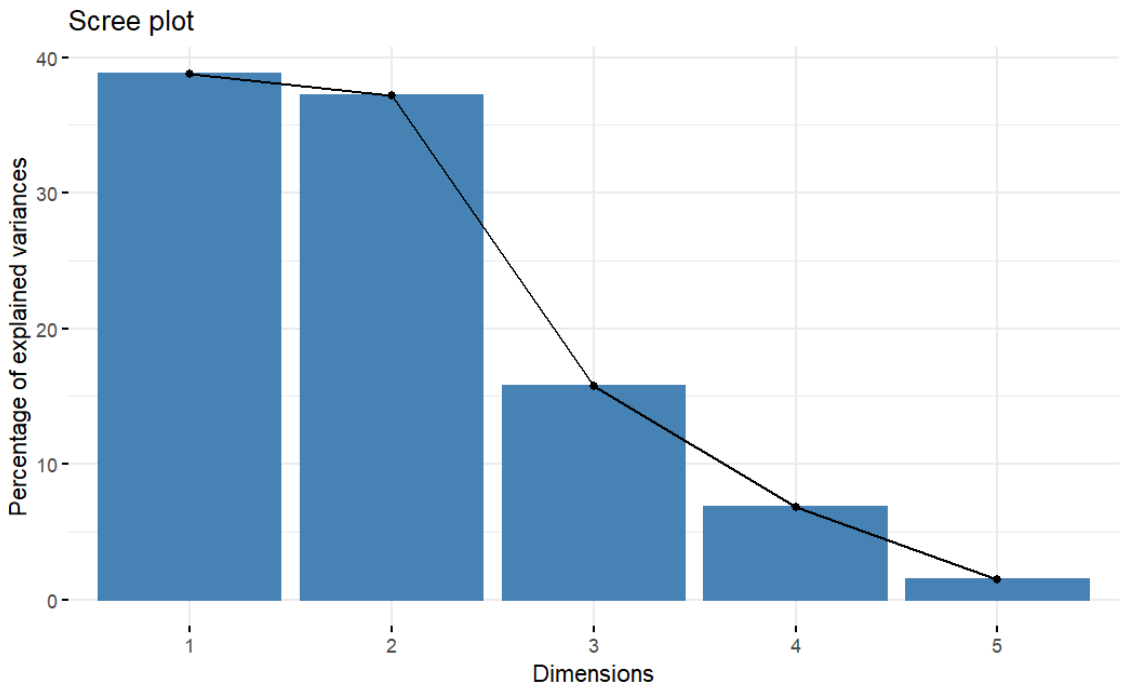

Ilustración 3. Gráfico de pantalla o sedimentación.

En el eje Y encontramos el porcentaje de varianza explicada, y en el eje X las dimensiones. Podemos decir que se forma un codo en la dimensión 2.

- Analisis Paralelo de Horn. Contrasta la variabilidad expresada en el conjunto de datos original con la obtenida de conjuntos de datos generados aleatoriamente con características similares al original. Esta opción permite controlar posibles efectos de sesgo en la elección de número de PCs cuando el conjunto de datos es pequeño. (Horn 1965).

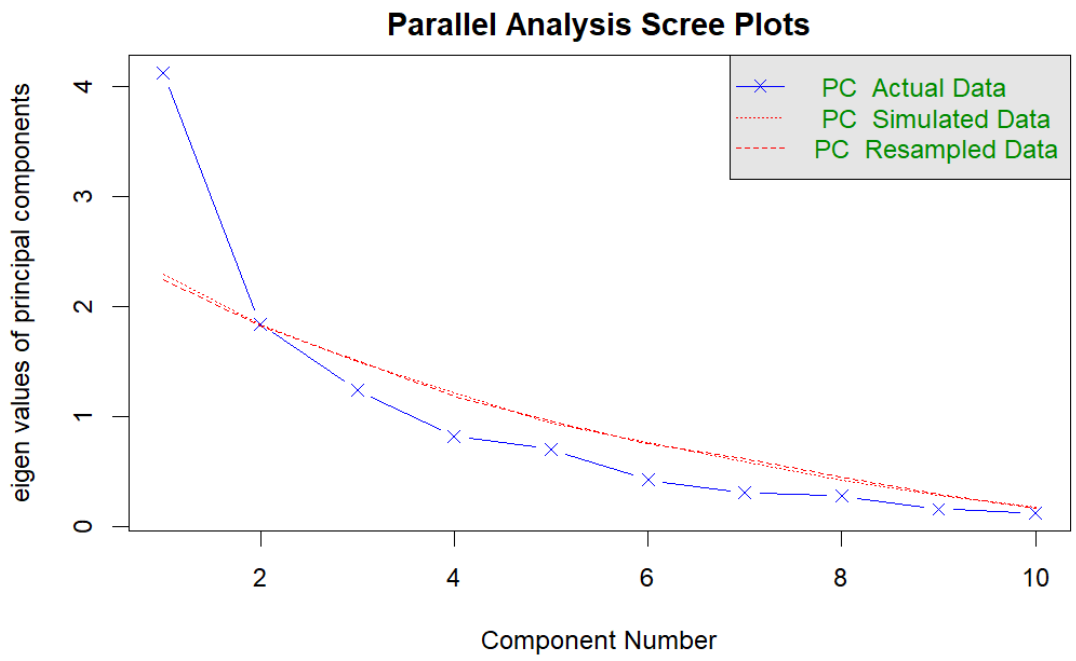

Ilustración 4. Gráfico de análisis paralelo de Horn.

Este método nos muestra que con tan solo los dos primeros es suficiente, los que están por encima de la línea roja. La línea roja representa lo que ocurre por azar.

En resumen, con estos datos podemos decir que con dos PCs es suficiente.

3.4.1. Análisis de componentes principales (PCA)

Recuerda que uno de los objetivos del PCA es encontrar un pequeño número de variables sintéticas que resuman muchas variables. Otro de los objetivos es identificar grupos de individuos con perfiles similares o extremos.

Para interpretar los resultados debemos buscar aquellas variables más importantes o relevantes para los ejes que hemos seleccionado. Para ello observamos:

- “coord.”. Coordenadas de las variables en cada PC. Es la correlación de la variable con cada PC.

- “cos2”. Representa la calidad de la representación para las variables en cada PC. Corresponde al cuadrado de las coordenadas (coord^2).

- “contrib”. Contiene la contribución (en porcentaje) de la variable en cada PC. Se calcula como la calidad de la variable dividido la calidad total del componente (cos2/sum(cos2)).

3.4.1.1. Estudio de las variables (características)

Vamos a ver entonces cómo interpretar las relaciones entre las variables. Podemos graficar esta información en un mapa de variables con la función *fviz_pca_var*():

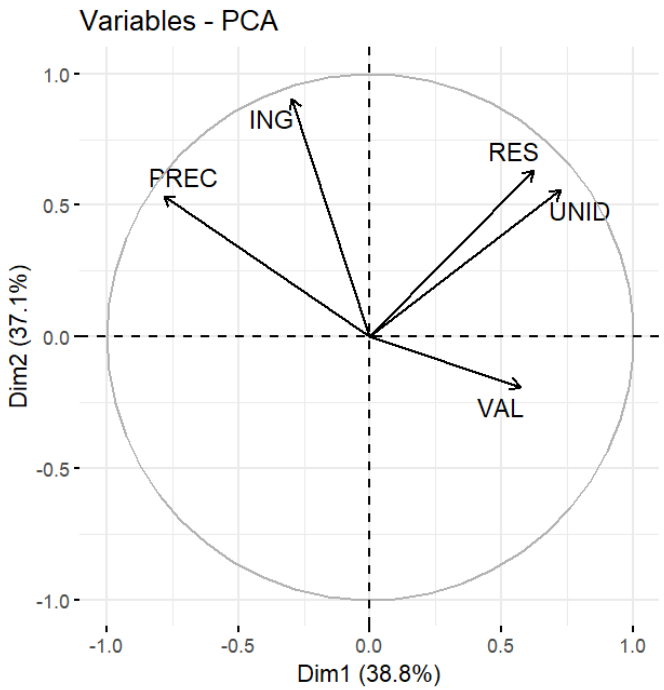

Ilustración 5. Mapa de variables PCA.

Antes de interpretar el gráfico vamos a explicar en qué consiste.

- Los ejes se corresponden con las dimensiones, es decir, componentes principales (PC). Las variables se representan con una flecha desde el origen.

- La correlación entre la variable y el PC se utiliza como coordenada en cada eje o PC. Al ser una representación de la matriz de correlaciones los ejes van de -1 a 1.

- El círculo de correlación con circunferencia 1 (correlación perfecta), que sirve para identificar las variables con mayor correlación a cada eje.

- Los ángulos (entre variables o entre variable y PC) se pueden interpretar como el signo de la correlación entre ellos.

  - Ángulo pequeño = correlación positiva

  - Ángulo grande = correlación negativa (posición opuesta)

  - Ángulo 90º = sin cor.

- El largo de la flecha (distancia) mide la “magnitud” de la correlación. La relación entre variables y componentes, aquellas flechas de las variables que estén en la misma dirección que el eje tendrán mayor correlación, ya sea positiva o negativa. Pero si además el tamaño de la flecha es largo diremos que está bien representada en dicho eje, y si es corta es que no es muy importante para dicho eje.

Por tanto, aquí tenemos que 4 de las variables están bien representadas (cercano a 1), que son: PREC, ING, RES Y UNID. PREC e ING están positivamente correlacionadas entre ellas, y RES y UNID también. En cuanto a las dimensiones, la dimensión 1 está positivamente correlacionado con VAL (aunque no muy bien representado) y RES Y UNID, aunque estas dos últimas tienen una correlación media aunque muy bien representada, y por el contrario, PREC e ING están negativamente correlacionados con esta dimensión aunque con poca correlación. La dimensión 2 está positivamente correlacionado con todas menos VAL, y bien representados.

Después de ver las coordenadas vamos a pasar a ver la calidad de las variables para cada componente principal.

Solo podemos interpretar aquellas variables bien proyectadas, que corresponde a las coordenadas al cuadrado (cos2).

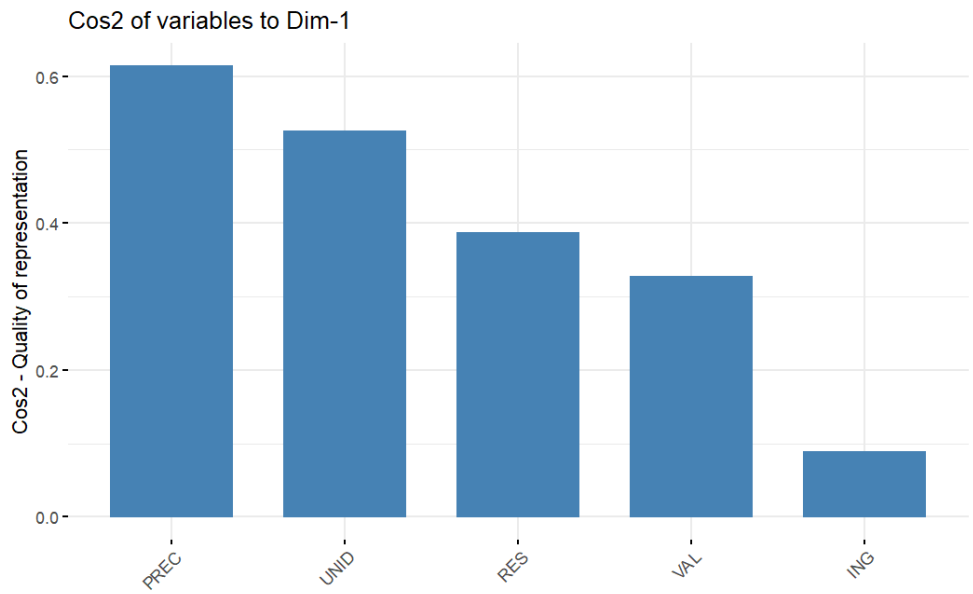

Ilustración 6. Gráfico de calidad de variables para la dimensión 1.

Para la primera dimensión (eje horizontal), vemos que las 2 primeras variables son las mejores representadas, RES y VAL están al límite, aunque se podría decir que tienen una calidad baja.

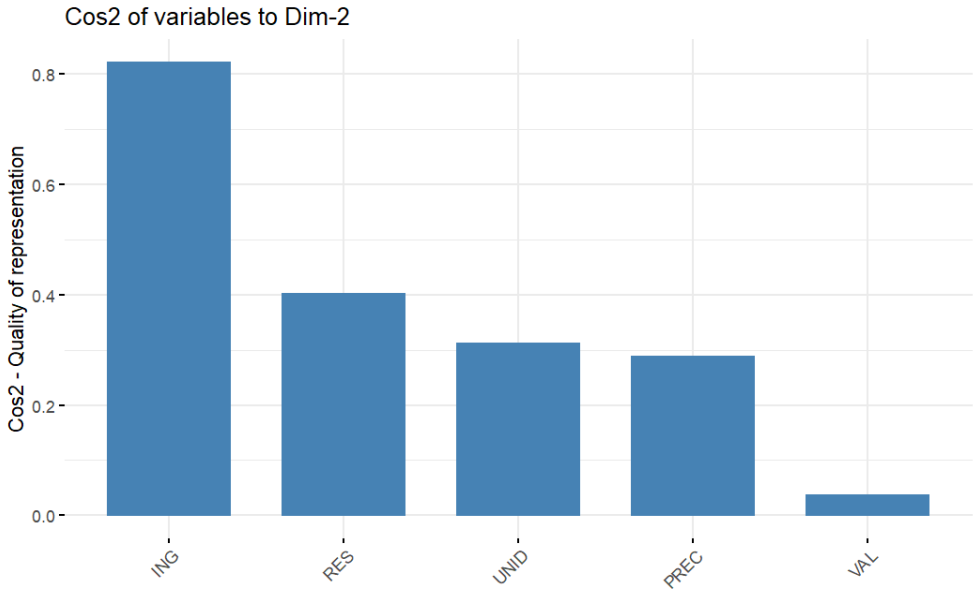

Ilustración 7. Gráfico de calidad de variables para la dimensión 2.

En el caso de la dimensión 2, vemos que tan solo la primera es la mejor representada, y es la variable que está más cercana al eje vertical en dirección positiva. RES, UNID y PREC sucede igual que en la dimensión 1, están bastante en el límite con un valor bajo.

En cuanto a la contribución de las variables a los componentes se calcula como la calidad de la variable dividido entre la calidad total del componente, se expresa como un porcentaje.

Las variables con mayor contribución a un PC son las más correlacionadas con él. Línea roja: contribución media esperada si fueran uniforme. 1/n (1/5=0.2, 20%).

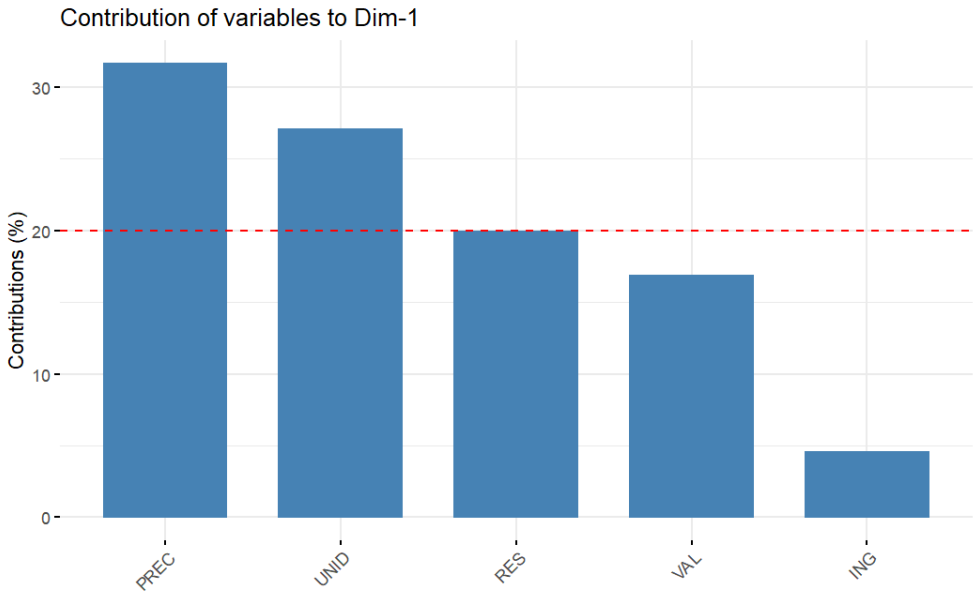

Ilustración 8. Gráfico de contribución de variables para la dimensión 1.

Las variables que se correlacionan con los primeros componentes son los más importantes en explicar la variabilidad de los datos. Los que tienen una baja contribución son los menos importantes y podrían eliminarse del análisis.

Por tanto, las variables que se encuentran por encima de la línea roja se consideran importantes para el componente. En el caso de la dimensión 1 serían las variables PREC, UNID y RES.

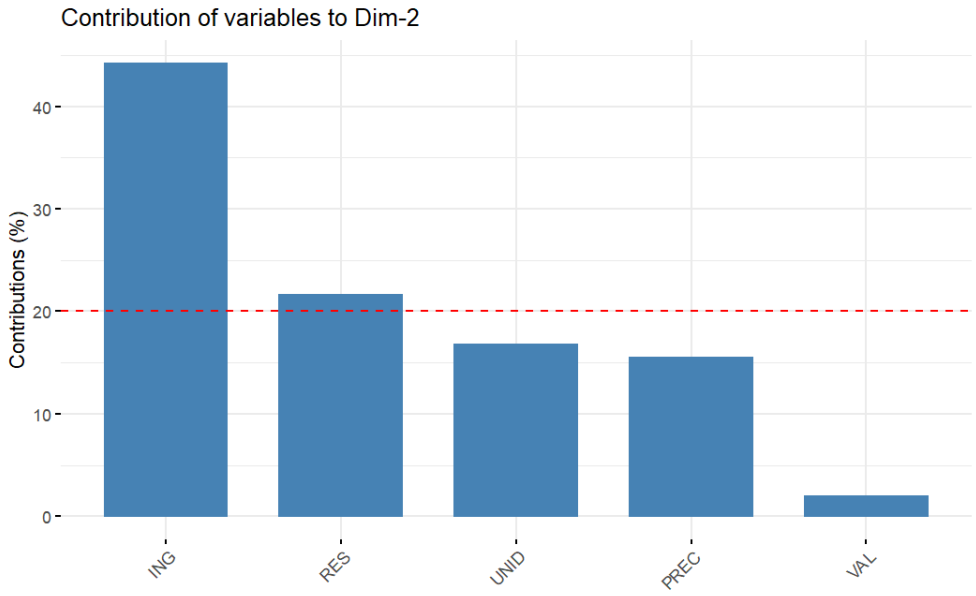

Ilustración 9. Gráfico de contribución de variables para la dimensión 2.

En el caso de la dimensión 2 serían las variables ING y RES. Por tanto, podríamos eliminar si queremos la variable VAL del análisis. En este caso lo dejaremos, pero en caso de la interpretación no le daremos cierta importancia.

Por último, la función dimdesc() de factominer nos permite describir las dimensiones o PC, identificando las variables asociadas significativamente a cada uno.

Tabla 5. Descripción de variables de la dimensión 1.

Dimensión 1	Correlación	P. valor
UNID	0.7241895	2.758098e-07
RES	0.6213005	3.145321e-05
VAL	0.5718272	1.765483e-04
PREC	-0.7834947	6.041392e-09

Vemos en la tabla las variables ordenadas según los valores de correlación para cada componente con su nivel de significación. En la tabla solo aparecen las más significativas, y a mayor correlación mayor significación. Vemos que para la dimensión 1 la variable PREC es la que mayor correlación tiene y es negativa, seguidamente de UNID que tiene correlación positiva. En esta tabla no aparece la variable ING por no ser suficientemente significativa.

Tabla 6. Descripción de variables de la dimensión 2.

Dimensión 2	Correlación	P. valor
ING	0.9060311	5.214620e-15
RES	0.6333170	1.976713e-05
UNID	0.5584705	2.686502e-04
PREC	0.5356021	5.296765e-04

En cuanto a la dimensión 2, vemos que la variable ING es la que mayor correlación tiene, además muy alta, seguido del resto que tienen una correlación media. En esta tabla no aparece la variable VAL.

3.4.1.2. Estudio de los individuos (casos)

Vamos a ver entonces cuándo podemos decir que ciertos individuos son similares (o distintos) respecto a todas las variables.

Al igual que como ocurría con las variables, para los individuos podemos obtener la siguiente información: coordenadas (coor); calidad (cos2); y contribución (contrib).

Podemos mapear los individuos en los primeros PC para ver su relación. A diferencia del mapa de variables que utilizaba las correlaciones como coordenadas, los individuos u observaciones se representan mediante sus proyecciones en el mapa. Mapeamos en función de los dos primeros componentes.

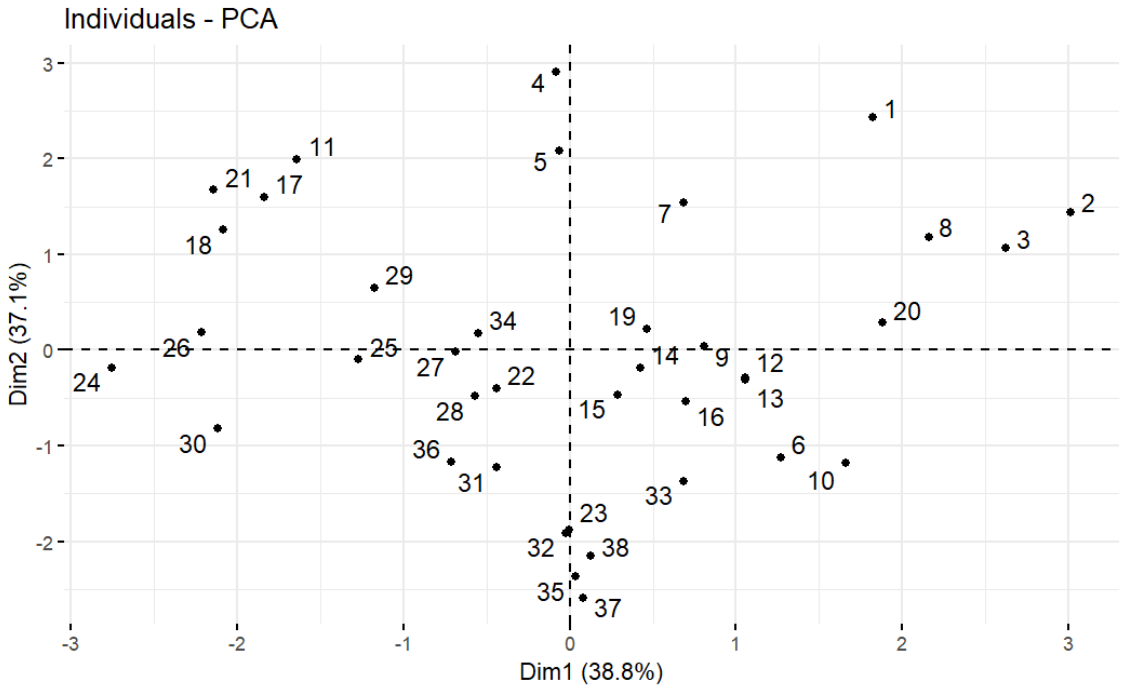

Ilustración 10. Mapa de individuos PCA.

- Los individuos con valores promedio en todas las variables se colocarán en el centro del gráfico. Y a la inversa, esperamos que los individuos con valores extremos estarán lejos del centro.

- Los individuos con valores similares (en las variables de estudio) se agruparán en el mapa. Y a la inversa, los individuos con perfiles distintos estarán lejos entre sí.

Podemos ver que por ejemplo el número 2 (Terraria) es un individuo extremo por sus valores, y algo similar al 3 (Garry's Mod). Y por otro lado, el 19 (Dark souls 3), el 14 (Stardew Valley) y el 15 (Phasmophobia) tienen valores promedios, y también son similares.

En cuanto a la calidad de la representación. La distancia entre individuos solo puede interpretarse para individuos bien proyectados (con alto cos2).

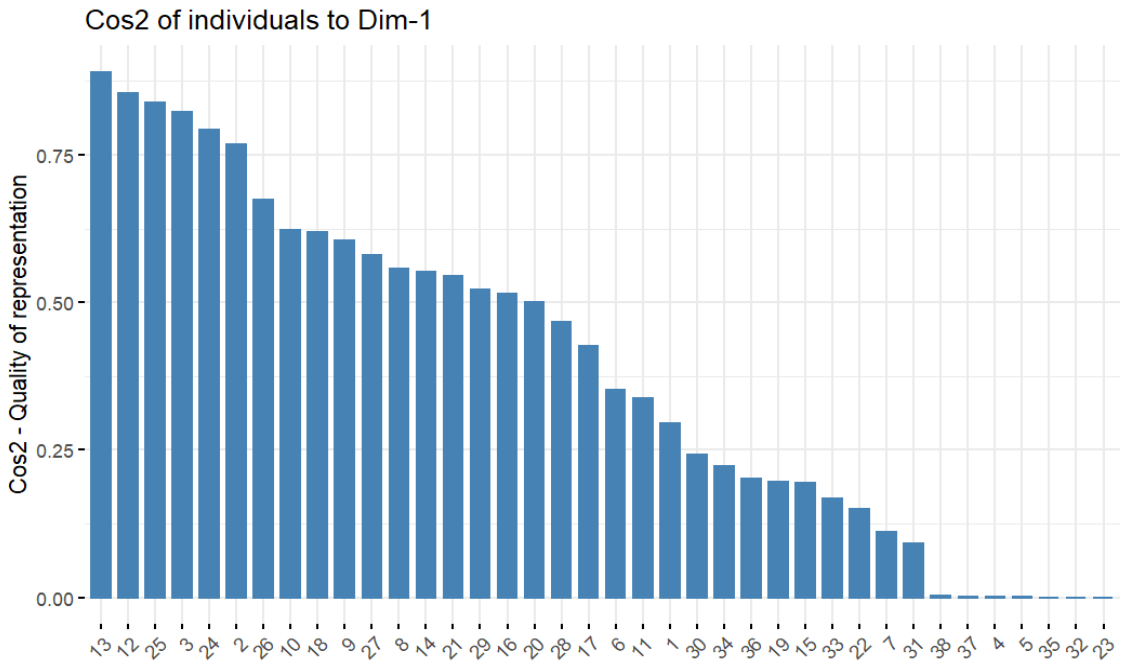

Ilustración 11. Gráfico de calidad de individuos para la dimensión 1.

En la primera dimensión podemos decir que los 16 primeros son los que tienen mejor representación.

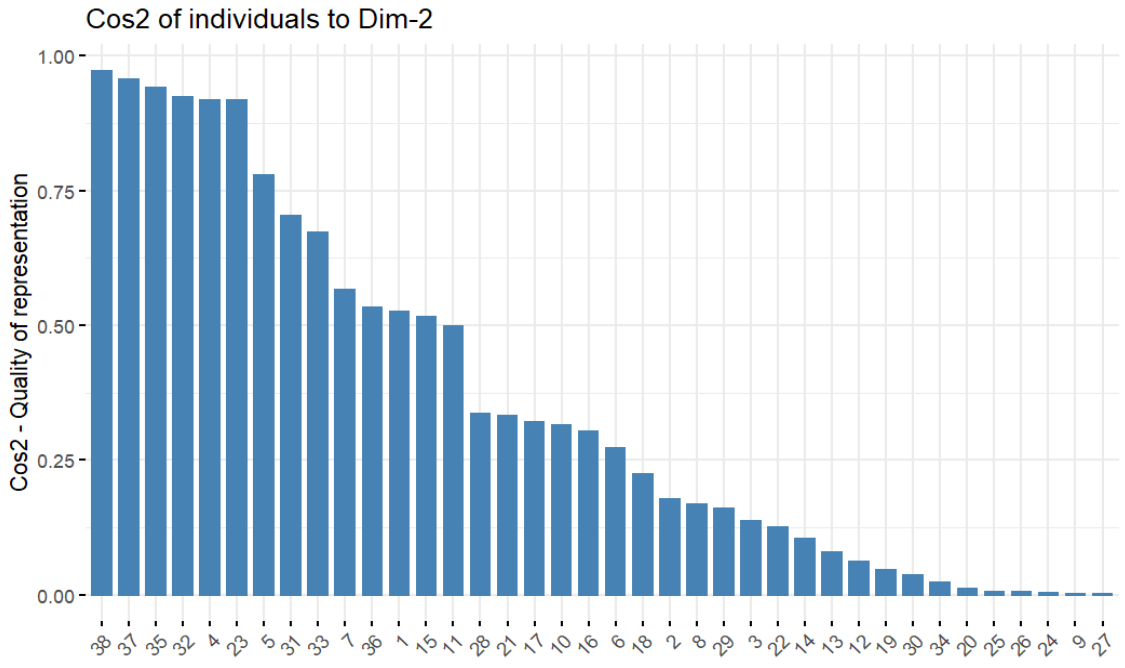

Ilustración 12. Gráfico de calidad de individuos para la dimensión 2.

En el caso de la segunda dimensión, vemos que los primeros 14 son los que tienen mejor representación. Los individuos que se encuentran más cerca del centro son los que menos calidad tienen para ambas dimensiones.

Pasemos a la contribución a la representación de los individuos. Los individuos con mayor valor de coordenada contribuyen más al eje (alta contribución).

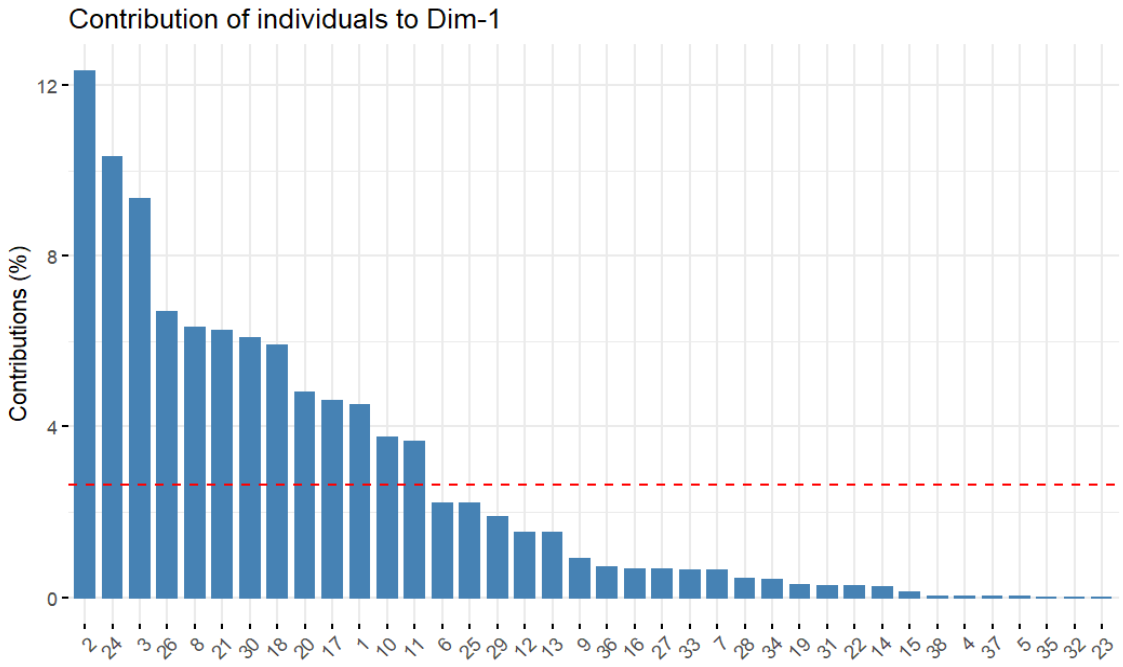

Ilustración 13. Gráfico de contribución de individuos para la dimensión 1.

Los valores que se sitúan por encima de la línea roja son los que contribuyen más al eje, por lo que tiene sentido su interpretación. En el caso de la dimensión 1 son los 13 primeros.

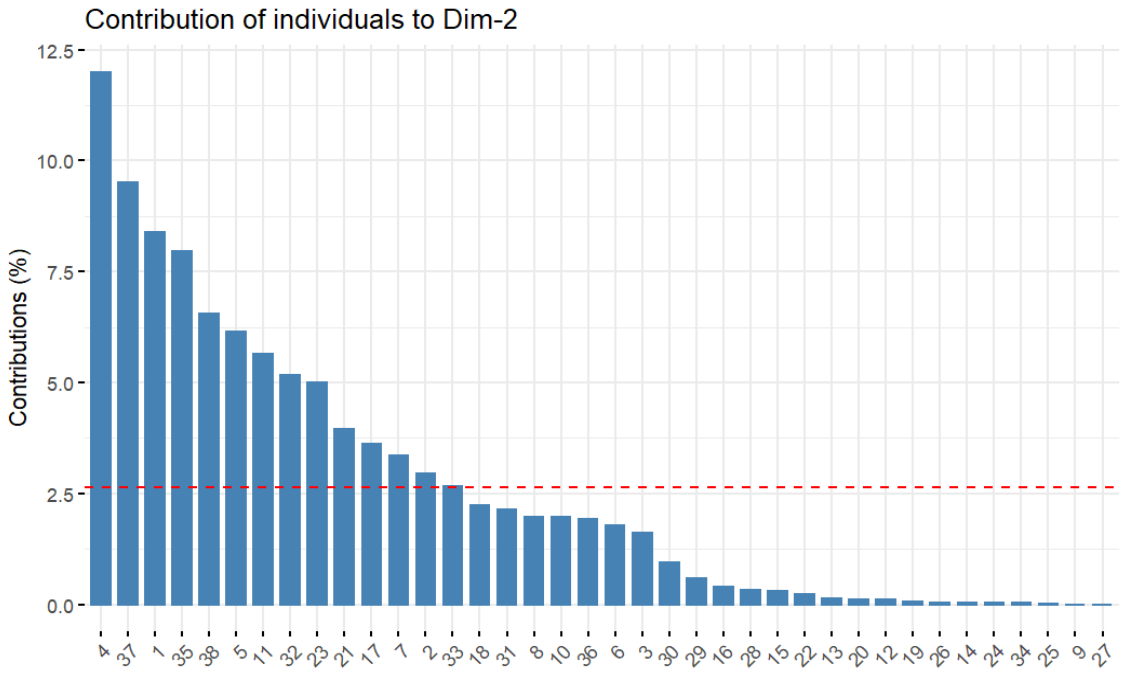

Ilustración 14. Gráfico de contribución de individuos para la dimensión 2.

En el caso de la dimensión 2 son los 14 primeros. Posteriormente en la interpretación de los individuos escogeremos los que mayor valor tengan.

3.4.1.3. Relación entre variables e individuos. Biplot

Ya hemos visto por encima las variables y los individuos por separado, pero lo más interesante es analizar sus relaciones.

- Caracterizar los grupos de individuos utilizando las variables.

- Utilizar individuos "extremos" para entender mejor las relaciones entre variables.

Al final, los objetivos del PCA es visualizar y resumir las relaciones. Además, estudiar las relaciones es útil cuando hay un número pequeño de variables e individuos. Como se puede decir que es nuestro caso.

Para evaluar la relación entre variables e individuos podemos graficar conjuntamente variables e individuos en lo que llamamos un “biplot”. Se trata de un mapa donde observamos dos representaciones de los datos: la de las variables (como flechas) y la de los individuos (como puntos).

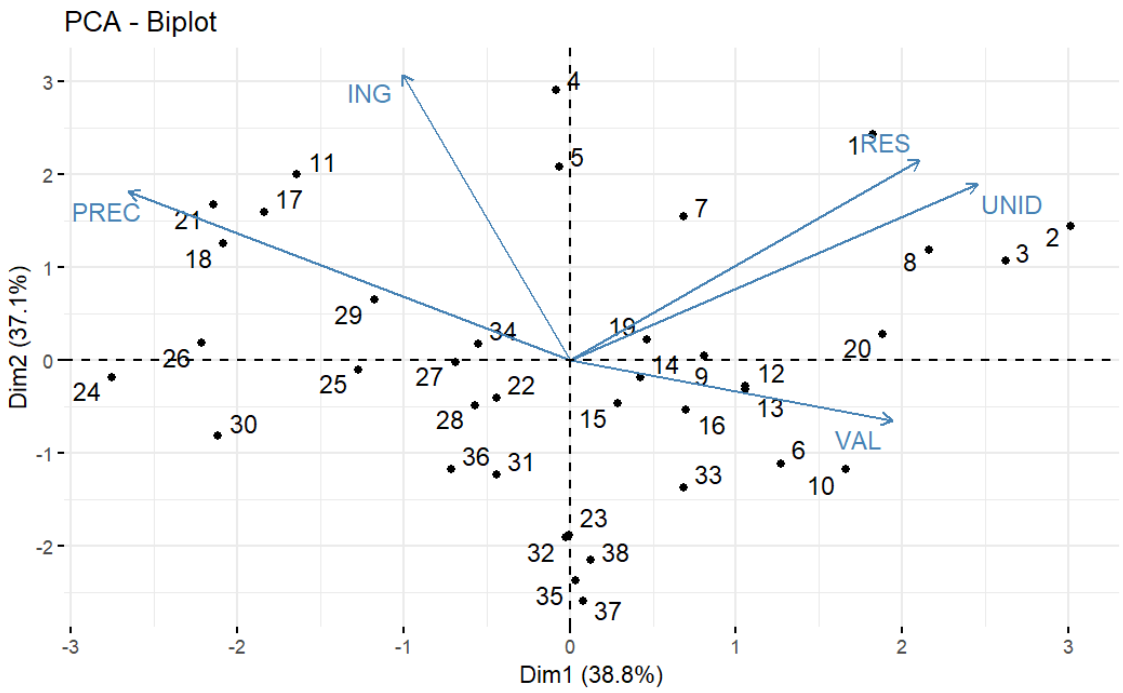

Ilustración 15. Mapa de variables e individuos, Biplot.

Hay que tener en cuenta que las coordenadas de los individuos y las variables no se construyen en el mismo espacio, las variables utilizan las correlaciones y los individuos las proyecciones. Por lo tanto, para inferir la relación entre variables e individuos, solo se puede observar la dirección, no la distancia en el biplot.

- Un individuo que está en el mismo lado de una variable tiene un alto valor en dicha variable.

- Un individuo que está en el lado opuesto de una variable tiene un valor bajo para esa variable.

Como dijimos que solo podemos interpretar las variables e individuos con buena calidad y contribución en la representación, vamos a filtrar el biplot

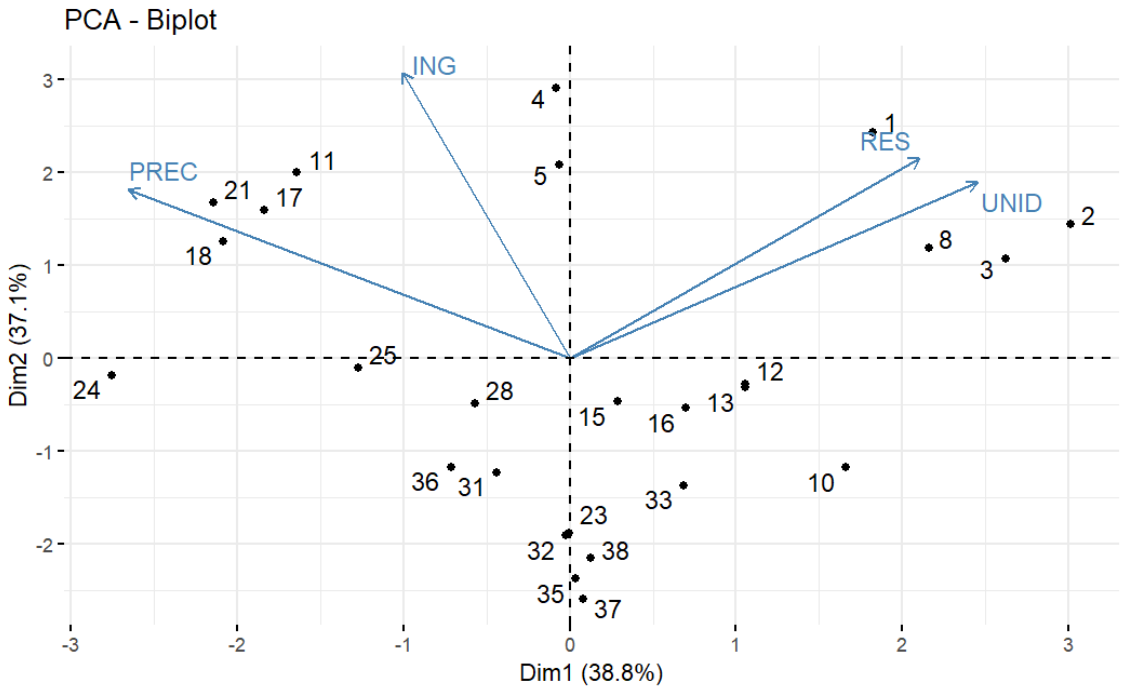

Ilustración 16. Mapa de variables e individuos filtrados, Biplot.

Por tanto, vemos que la variable UNID es la que más contribuye a la dimensión 1 en el lado positivo, siendo los juegos, Tom Clancy's Rainbow Six Siege (1), Terraria (2), Garry's Mod (3), y Dead by Daylight (8), los que más valor tienen para esta variable, siendo además los que más reseña tienen. Y en el lado opuesto tenemos los juegos con menos unidades vendidas y menos reseñas a su vez a Fallout 4 (24), Hollow knight (25), Cities: Skylines (36), y New World (28). Por otro lado, tenemos la variable PREC que es la que más contribuye a la dimensión 1 en el lado negativo, con el mayor valor en los juegos Valheim (17), Don't Starve Together (18), y The Elder Scrolls V: Skyrim (21), y en el lado opuesto con menor valor tenemos a Euro Truck Simulator 2 (10). Parece que en gran medida, esta dimensión divide a los juegos con alto valor en precios de los juegos con altas unidades vendidas, por lo que parece a priori que el precio influye en eso. Estos son los valores para los juegos con mayor calidad, es decir, los que se pueden interpretar.

La misma lógica se aplica a la dimensión 2. Las variables ING y RES están bien representadas y contribuyen positivamente a la dimensión 2. Los juegos con mayor relación con los ingresos son Rust (4) y The Witcher 3: Wild Hunt (5).

Según las variables que más contribuyen a cada componente, podemos decir que el componente 1 separó los juegos por precios y unidades vendidas, mientras que el componente 2 estuvo relacionado con los ingresos.

3.4.1.4. Estudio de la información suplementaria.

Vamos a pasar ahora a analizar las variables suplementarias que dejamos anteriormente al margen del análisis. Las variables e individuos adicionales no influyen en los componentes principales del análisis.

Si graficamos los resultados en un biplot, de forma predeterminada:

- Las variables cuantitativas suplementarias se muestran en líneas color azul y discontinuas,

- Los individuos suplementarios en puntos azules,

- Para las variables cualitativas, se muestra el centro de cada categoría y se pueden graficar las elipses de confianza.

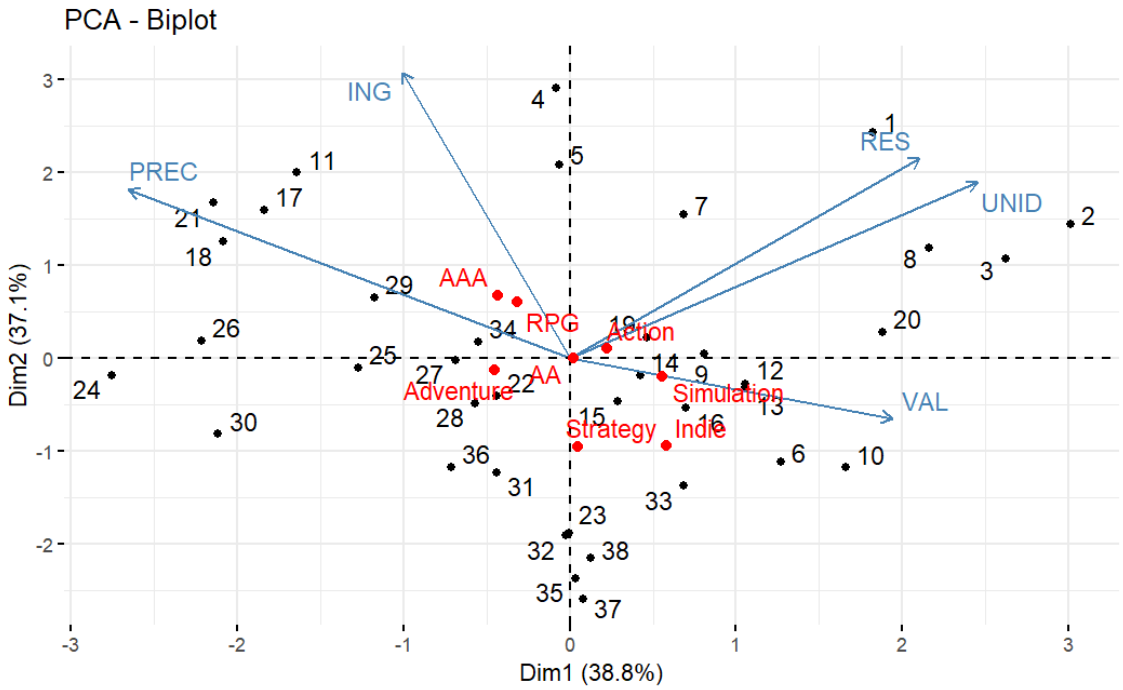

Ilustración 17. Mapa de variables e individuos con variables suplementarias, Biplot.

Aquí ahora vemos representados los valores de nuestras variables suplementarias. En cuanto a la variable CLAS, vemos como las series de juegos AAA se relaciona más con mayores precios y mayores ingresos, las AA se sitúan en el centro, con un valor medio en todas las variables, y los Indie se sitúan con valores bajos de ingresos y precio, pero con mayor valoración.

En cuanto al género del juego, los RPG se sitúan de forma similar que los juegos AAA, los juegos de acción se sitúan un poco más junto a las unidades vendidas, al contrario que los de estrategia, los de simulación parece que estén algo más valorados, y los de estrategia están un poco al margen de todos. También hay que recalcar que se sitúan cerca del centro por lo que por lo general tienen valores similares en todas las variables con pequeños matices.

En la descripción de los componentes principales (como se hizo anteriormente con las variables) vemos que en la primera dimensión ninguna de las variables cualitativas es significativa. Mientras que en la segunda dimensión vemos que la variable CLAS es significativa y que explica la variabilidad de los datos en esta dimensión en un 21.75%. Además, vemos que la clase AAA tiene mayor valor estimado que la clase Indie.

3.4.2. Agrupación jerárquica de los componentes principales (HCPC)

Utilizamos el PCA como un paso de preprocesamiento antes de realizar métodos de agrupación, con el fin de eliminar el ruido de los datos y que la agrupación sea más estable que la obtenida de las distancias originales.

Entonces realizamos un PCA pero ahora especificando que queremos dos dimensiones. Este paso es para eliminar el ruido.

Lo siguiente es realizar la agrupación a través de estos componentes principales. Obtenemos el siguiente dendrograma.

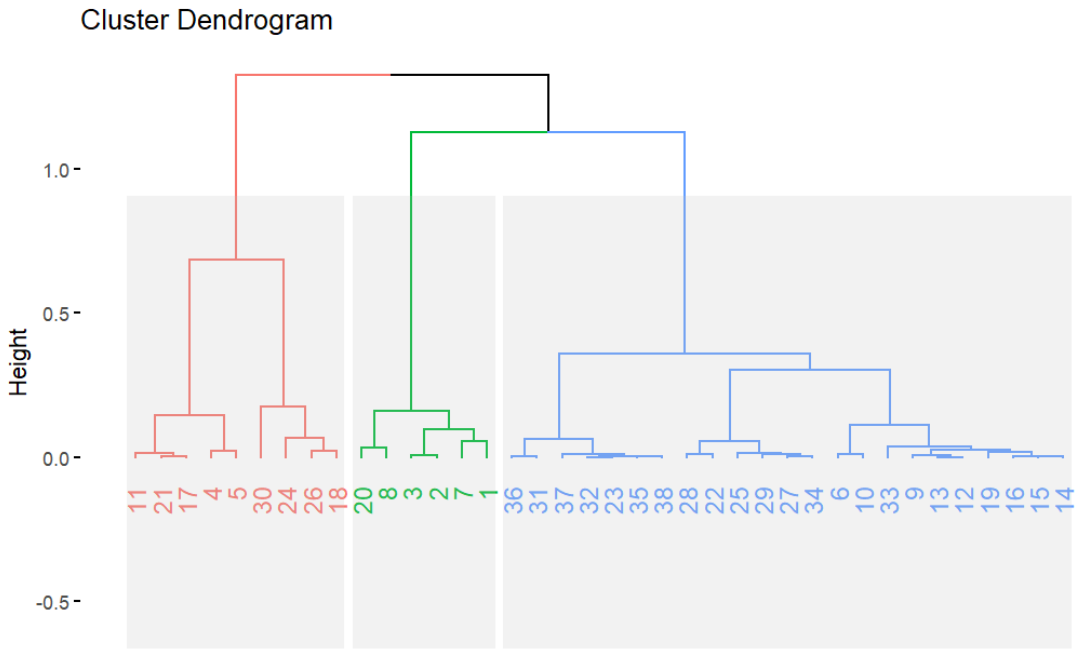

Ilustración 18. Dendrograma y agrupación de los componentes principales.

Vemos los grupos de juegos que se han formado, siendo el azul y el verde más parecidos y el rojo el menos similar. Esta agrupación se ha realizado de manera automática.

Podemos observar esta agrupación representados en un mapa. De esta manera podemos ver que dimensiones han influido más en estas agrupaciones.

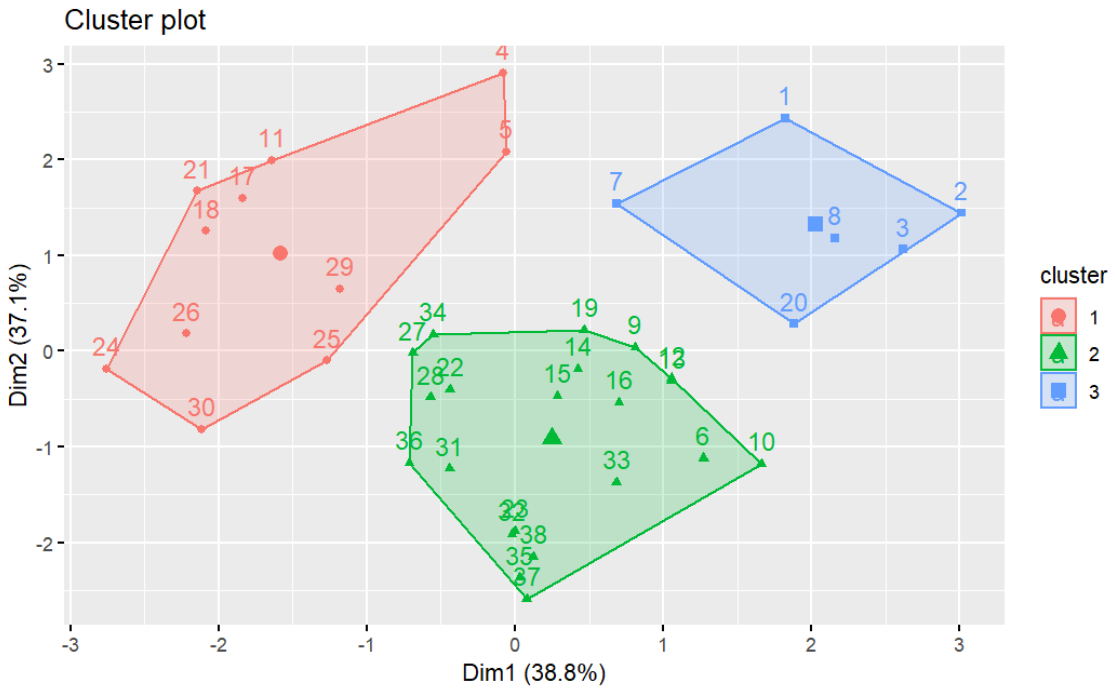

Ilustración 19. Mapa HCPC.

Como se puede observar se han intercambiado el color verde por el azul, así que se interpretará con estos últimos colores. En este gráfico podemos observar que la primera dimensión separa los grupos entre sí. Mientras que la segunda dimensión separa el grupo verde del rojo y azul.

Para caracterizar aún más los grupos, vamos a ver las variables que describen los cluster, así como los individuos más típicos de cada grupo.

Vamos a empezar por las variables categóricas que caracterizaron a los grupos. Descripción de cada grupo por cada categoría de todas las variables categóricas.

Solo se muestran los resultados significativos para un nivel de confianza del 95%. En el grupo 1 vemos:

- Que el 50% de los individuos que se asignaron al tipo AAA pertenecen al grupo 1.

- El 63% del grupo 1 perteneció a la asignación AAA.

- Y el 36% del total de individuos perteneció a la asignación AAA.

En el grupo 2 vemos:

- Que el 28% de los individuos que se asignaron al tipo AAA pertenecen al grupo 1.

- El 19% del grupo 2 perteneció a la asignación AAA.

- Y el 36% del total de individuos perteneció a la asignación AAA.

Estas son las únicas categorías significativas para cada grupo.

Pasamos a las variables cuantitativas que describen mejor a cada grupo. La descripción global de los grupos por las variables cuantitativas lo vemos en las siguientes tablas, los resultados completos se encuentran en el código adjuntado en el anexo.

Tabla 7. Descripción del grupo 1 por variables cuantitativas.

Variables	Media en la categoría	Media general
PREC	45.9	24.7
ING	231.1	147.9
VAL	81.9	90.6

Vemos que el primer grupo se identifica con valores altos de precio e ingresos, bastante más altos que la media. Mientras que en cuanto valoración se sitúan bastante por debajo, aunque esta variable no era muy significativa en general.

Tabla 8. Descripción del grupo 2 por variables cuantitativas.

Variables	Media en la categoría	Media general
VAL	94.7	90.6
UNID	9.3	11.1
RES	314922	395302
PREC	16.9	24.7
ING	98	147.9

En cuanto al grupo 2, se caracteriza por tener valores por debajo de la media en todas las variables excepto en valoración.

Tabla 9. Descripción del grupo 3 por variables cuantitativas.

Variables	Media en la categoría	Media general
UNID	22	11.1
RES	732126	395.302

El tercer grupo se caracteriza por tener un valor de unidades vendidas y de reseñas muy por encima del resto.

Ahora vamos a describir los grupos identificando los principales individuos de cada grupo.

Obtenemos que en el primer grupo destacan 5 individuos, siendo Don't Starve Together (18) el primero, situándose a 1.14 del balicentro del grupo, seguidamente tenemos a Arma 3 (29), Hollow Knight (25), The Elder Scrolls V: Skyrim (21) y Wallpaper Engine (11). Del grupo 2 destacan Phasmophobia (15), The Forest (16), Stardew Valley (14), Risk of Rain 2 (33) y The Elder Scrolls V: Skyrim Special Edition (31). Y del grupo 3 destacan Garry's Mod (3), Dead by Daylight (8), Terraria (2), Among Us (7) y Tom Clancy's Rainbow Six Siege (1).

Según esto, para el grupo 1 que se forma del 63% de juegos AAA, podemos decir que la inversión elevada en juegos repercute en unos mayores ingresos y esto se refleja en un mayor precio para cubrir los mayores costes posibles.

Por otro lado, podemos decir que la valoración no repercute en cuanto a ingresos o que las valoraciones están desvirtuadas o no provienen de una fuente fiable.

Y por último, aunque el grupo 1 obtenga los mayores ingresos, el grupo 3 obtiene la mayor venta de unidades de todos ellos, así como cantidad de reseñas, que se podría traducir en fama.

4. Conclusiones y futuras líneas de investigación

Como conclusiones, el desarrollo de todo el estudio ha dado lugar a una agrupación de los videojuegos, consistente en conocer cuales son los que mayores ingresos han obtenido y razones.

Los videojuegos llevan décadas existiendo y ahora más que nunca hay un gran aumento del consumo en este mercado.

Se ha escogido la variable ingresos como la principal, y del resto, tan solo la variable ”valoración“ no ha resultado significativa.

Vistos los resultados, podemos concluir que en gran medida la inversión en juegos es una herramienta clave a la hora de obtener ingresos, aunque existen excepciones. Además, hemos podido comprobar que el género no es ningún impedimento para ser exitoso, ya que no ha sido significativo para los análisis. Hemos visto también, que la variable precio tiene una alta correlación con los ingresos, que son las variables clave en una de las agrupaciones, la que posee más juegos AAA.

Para poder haber obtenido mejores frutos del proyecto se debía de disponer de una base de datos más amplia, esa ha sido la mayor limitación, aunque es un paso para empezar y poder proyectarse más hacia dichas inversiones, en que enfocar los esfuerzos, si en la creación del juego en sí, en la gráfica, la historia, o incluso el marketing que hay detrás de cada juego. Todo eso influye en la venta de un juego y por tanto en sus ingresos. Habría también que examinar las excepciones de juegos con poca inversión que han generado tantos ingresos, si es por propia naturaleza del juego o hay ciertas condiciones que se dan para llegar a ser exitoso.

5. Bibliografía

Bibliografía introducción

Web

[1] NewZoo. URL: The Games Market and Beyond in 2021: The Year in Numbers | Newzoo

Bibliografía definiciones

Web

[2] Estadística para todos. URL: https://www.estadisticaparatodos.es/taller/graficas/cajas.html

[3] Máxima Formación. URL: https://www.maximaformacion.es/blog-dat/que-es-r-software/

STHDA. URL: http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/117-hcpc-hierarchical-clustering-on-principal-components-essentials/

Libros

[4] García de Zúñiga, F. (octubre 2020). << RStudio, IDE para programar con R. Instalación y primeros pasos >>. En: arsys.

Bibliografía base de datos

web

[5] VG Insights. URL: Games Database - Steam games data | Video Game Insights (vginsights.com)

Bibliografía apartado "resultados y discusión"

LIBROS

Ferrero, R. (2022) <<Análisis de componentes principales (PCA)>> En: Máxima Formación S.L.

Perez, J; Albor, D; Ortega, W; Fontalvo, M (30/10/2021) <<Análisis de Componentes Principales (PCA)>> En: rstudio-pubs

6. Anexo

En el Anexo se aporta la base de datos con las variables, así como el código utilizado en el software R. La base de datos ha sido limpiada y ordenada para su correcto uso.

Maxima Formacion

Base de Datos

ING	PREC	UNID	RES	VAL	GEN	CLAS	NOM
226.2	19.99	18.9	1.067.195	87	Action	AAA	Tom Clancy's Rainbow Six Siege
135.6	9.99	22.6	983.872	97.9	Adventure	Indie	Terraria
128.7	9.99	21.5	850.772	96.6	Simulation	AAA	Garry's Mod
337.2	39.99	14.1	810.990	86.7	RPG	AA	Rust
303.9	39.99	12.7	641.939	96.1	RPG	AA	The Witcher 3: Wild Hunt
304.4	9.99	50.8	631.277	97.5	Action	AAA	Left 4 Dead 2
26.5	4.99	8.8	628.303	91.9	Strategy	Indie	Among Us
222.8	19.99	18.6	606.490	81.4	Action	Indie	Dead by Daylight
157.2	9.99	26.2	582.882	89.3	Action	AA	PAYDAY 2
135.7	19.99	11.3	562.017	97.3	Simulation	AA	Euro Truck Simulator 2
28.8	3.99	12	531.382	98.1	Simulation	Indie	Wallpaper Engine
562.9	59.99	15.6	521.298	75.1	RPG	AA	Cyberpunk 2077
294.1	59.99	8.2	510.661	90	RPG	AAA	ELDEN RING
116	14.99	12.9	483.209	98.1	Simulation	AA	Stardew Valley
114.3	13.99	13.6	453.803	97	Action	Indie	Phasmophobia
142.5	19.99	11.9	387.164	95.2	Adventure	AA	The Forest
127	19.99	10.6	352.967	95.4	RPG	AA	Valheim
116.3	14.99	12.9	341.304	96.1	Adventure	AA	Don't Starve Together
295.8	59.99	8.2	331.207	93.6	Action	AAA	DARK SOULS™ III
252.2	44.99	9.3	310.459	71.2	Adventure	AAA	DayZ
183.7	19.99	15.3	309.868	94.8	RPG	AAA	The Elder Scrolls V: Skyrim
146.8	9.99	24.5	301.547	98.8	Action	AAA	Portal 2
306	59.99	8.5	283.367	88.5	Action	AAA	Red Dead Redemption 2
127.6	19.99	10.6	282.363	81.2	RPG	AAA	Fallout 4
43.9	14.99	4.9	239.110	97.1	Adventure	AA	Hollow Knight
103.8	59.99	2.9	238.701	76.1	Adventure	AA	No Man's Sky
167.4	39.99	7	232.752	90.3	Action	AAA	Sea of Thieves
164.6	39.99	6.9	228.773	67.7	Adventure	AAA	New World
182.5	29.99	10.1	227.108	90.4	Simulation	AAA	Arma 3
140.6	24.99	9.4	224.221	88.3	Strategy	AA	7 Days to Die
242.1	39.99	10.1	217.628	90.9	RPG	AAA	The Elder Scrolls V: Skyrim Special Edition
75.1	19.99	6.3	208.882	50.4	Simulation	Indie	鬼谷八荒 Tale of Immortal
92.7	24.99	6.2	206.196	96.4	Action	AA	Risk of Rain 2
46.8	14.99	5.2	205.661	97.7	Action	Indie	The Binding of Isaac: Rebirth
71.1	9.99	11.9	203.413	97.5	Action	AAA	Counter-Strike
207.5	29.99	11.5	200.764	93.4	Strategy	AA	Cities: Skylines
20.4	9.99	3.4	199.739	97.5	Strategy	Indie	Bloons TD 6
95	29.99	5.3	198.857	96.4	Adventure	AA	Subnautica
7.4	3.99	3.1	194.793	93.7	Action	Indie	Geometry Dash
33	9.99	5.5	181.121	96.7	RPG	Indie	Undertale

Código R

carga de datos

```{r}
games <- readxl::read_excel("data1.xlsx")
str(games)
```

Transformación de variables

```{r}
games$ING <- as.numeric(games$ING)
games$PREC <- as.numeric(games$PREC)
games$UNID <- as.numeric(games$UNID)
games$VAL <- as.numeric(games$VAL)
games$GEN <- as.factor(games$GEN)
games$CLAS <- as.factor(games$CLAS)
```

Resumen de datos

```{r}
summary(games)
```

Gráfico caja y bigotes de variables numéricas

```{r}
```

```
library(tidyr)

library(ggplot2)

games_std <- scale(games[,1:5])

long <- pivot_longer(as.data.frame(games_std), cols = 1:5)

ggplot(long, aes(x=name, y=value, fill=name)) + geom_boxplot() + theme(axis.text.x =
element_text(angle = 90)) + theme_update()
```

Identificación de valores atípicos

```{r}
out <- boxplot.stats(games$ING)$out

out_ind <- which(games$ING %in% c(out))

out_ind
```

```{r}
library(rstatix)

library(tidyverse)

games %>% identify_outliers(ING)
```

```{r}
games %>% identify_outliers(RES)
```

```{r}
games %>% identify_outliers(UNID)
```

```{r}
games %>% identify_outliers(VAL)
```

Gráfico ggpairs

```{r}
library(GGally)
ggpairs(games[,1:7])
```

Identificación outliers del PCA

```{r}
library(mt)
out <- pca.outlier(games[,1:5], adj=-0.5)
out$outlier
```

datos activos

```{r}
games_active <- games[c(-6,-12), 1:5]
```

Pruebas de correlación

```{r}
library(psych)
cortest.bartlett(games_active, n=98)
```

```{r}
KMO(cor(games_active))
```

```{r}
det(cor(games_active))
```

Creación modelo PCA

```{r}
library(FactoMineR)
library(factoextra)
( res_pca <- PCA(games_active, graph=FALSE))
```

Resumen model PCA

```{r}
```

summary(res_pca)
```

Selección de PCs

```{r}
get_eigenvalue(res_pca)
```

```{r}
fviz_screeplot(res_pca)
```

```{r}
library(psych)
fa.parallel(decathlon2_active, fa="pc")
```

Mapa de variables

```{r}
library(FactoMineR)
library(factoextra)
fviz_pca_var(res_pca, repel = TRUE)
```
```

Calidad de las variables

```{r}
fviz_cos2(res_pca, choice = "var", axes=1)
fviz_cos2(res_pca, choice = "var", axes=2)
```

Contribución de las variables

```{r}
fviz_contrib(res_pca, choice = "var", axes = 1)
fviz_contrib(res_pca, choice = "var", axes = 2)
```

Descripción de variables

```{r}
library(FactoMineR)
res_desc <- dimdesc(res_pca, axes=c(1,2), proba = 0.05)
```

```{r}
res_desc$Dim.1
```

```{r}
res_desc$Dim.2
```

mapa de individuos

```{r}
library(FactoMineR)
library(factoextra)
fviz_pca_ind(res_pca, repel=TRUE)
```

Calidad de los individuos

```{r}
fviz_cos2(res_pca, choice = "ind", axes = 1)
fviz_cos2(res_pca, choice = "ind", axes = 2)
```

Contribución individuos

```{r}
fviz_contrib(res_pca, choice = "ind", axes = 1)
fviz_contrib(res_pca, choice = "ind", axes = 2)
```

biplot

```{r}
library(FactoMineR)
library(factoextra)
```

```
fviz_pca_biplot(res_pca, repel=TRUE)
```

Biplot filtrado

```{r}
fviz_pca_biplot(res_pca, repel=T,
                select.var=list(contrib=4),
                select.ind=list(cos2=0.7),
                ggtheme = theme_minimal())
```

Creación del PCA completo

```{r}
library(FactoMineR)
library(factoextra)
res_pca_all <- PCA(games[c(-6,-12),-8],
                   graph = FALSE,
                   quali.sup = 6:7)
```

Biplot completo

```{r}
library(factoextra)
p <- fviz_pca_biplot(res_pca_all, repel = T)
p <- fviz_add (p,  res_pca_all$quali.sup$coord, color = "red", repel = T)
```

p

```

Descripción de los PCs

```{r}
desc_pca_all <- dimdesc(res_pca_all, axes = c(1, 2))
desc_pca_all$Dim.1
```

```{r}
desc_pca_all$Dim.2
```

Agrupación jerárquica sobre los componentes principales.

```{r}
library(FactoMineR)
library(factoextra)
hcpc <- HCPC(res_pca_all,
    nb.clust = -1,
    graph = F)
```

Dendrograma

```{r}

fviz_dend(hcpc, rect=TRUE, rect_fill = TRUE)

```

Mapa de la agrupación

```{r}

fviz_cluster(hcpc, show.clust.cent = TRUE, rect_fill = TRUE)

```

Variables categóricas que caracterizan a los grupos

```{r}

hcpc$desc.var$category

```

Variables cuantitativas que caracterizan los grupos

```{r}

hcpc$desc.var$quanti

```

Principales individuos de cada grupo.

```{r}

hcpc$desc.ind$para

```
```