Universidad Nacional Autónoma de México

Facultad de Estudios Superiores
Acatlán

Diplomado de Ciencia de Datos
Módulo III

proyecto
Análisis de videojuegos con sus ventas y calificaciones

Módulo III
Profesor : José Gustavo Fuentes Cabrera
Alumno : Hernández González Ricardo Paramont
Fecha : Viernes 18 de julio de 2021
Ubicación del texto (Latex) : https ://www.overleaf.com/read/tbgnxxmtctgk

Índice

1. Presentación

1.1 Objetivo

En el presente proyecto, se presenta el procesamiento y estruturamiento de una tabla de datos de 6811 registros y 16 columnas, correspondiente a varios videojuegos publicados desde 1985 hasta 2016. Los datos de cada videojuego fueron extraidos a través web scrapping del portal de Metracritic. Únicamente los videojuegos que contienen datos de ventas y criticas fueron considerados. El objetivo del proyecto es poder generar un agrupamiento que permita seccionar los videojuegos adecuadamente según sus características.

1.2 Diccionario de datos

Variable	**Tipo**	**Descripción**
Name	Texto	Nombre del videojuego.
Platform	Discreta	Consola de videojuegos.
Year of release	Continua	Año de publicación del videojuego.
Genre	Discreta	Género del videojuego.
Publisher	Discreta	Empresa que publica el videojuego.
NA sales	Continua	Ventas en millones de undades en Norteamérica.
EU sales	Continua	Ventas en millones de unidades en Europa.
JP sales	Continua	Ventas en millones de unidades en Japón.
Other sales	Continua	Ventas en otras partes del mundo.
Global sales	Continua	Ventas mundiales totales.
Critic score	Continua	Calificación agregada compilada por Metacritic.
Critic count	Continua	Número de críticas usadas para el valor de Critic_Score.
User score	Continua	Calificación dada por usuarios de Metacritic.
User count	Continua	Número de criticas usadas para el valor de User_score.
Developer	Discreta	Empresa desarrolladora del videojuego.
Rating	Discreta	Rating según ESRB.

2. Calidad de datos

2.1 Selescción de variables

Debido a que la agrupación se puede hacer únicamente tomando en cuenta variables de tipo continuo, se dividió a las variables, para excluir a las variables de tipo discreto. Variables tomadas en cuenta para el agrupamiento (continuas) :

- Year of release
- NA sales
- EU sales
- JP sales
- Other sales
- Global sales
- Critic score
- Critic count
- User score
- User count

Variables que no serán tomadas en cuenta para el agrupamiento (discretas) :

- Name
- Platform
- Genre
- Publisher
- Developer
- Raiting

Las variables discretas serán empleadas al final para perfinlar los grupos.

2.2 Duplicados

No se encontro ningún registro duplicado a lo largo de la tabla de datos.

2.3 Completitud

No se encontraron registros con valores faltantes, pues sólo se tomaron en cuenta videojuegos suficientemente relevantes, que contenían calificaciones y datos de ventas.

3. Análisis Exploratorio de Datos

3.1 Continuas

3.1.2 Consistencia

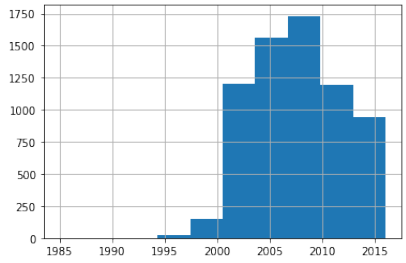
Para la revisión de consistencia, se limitó a revisar que el tipo de dato de cada variable fuera numérico y que los datos se encontraran entre rangos que tuvieran sentido.

- Se encontró que el rango del año de publicación se encontraba en el especificado, entre 1985 y 2016.
- El rango de todas las variables de venta se encontraba entre el cero y las 100 millones de copias.
- Las calificaciones en la variable de Critic Score se encontraban entre el 0 y el 100.
- Las calificaciones en la variable de User Score se entraban entre el 0 y el 10.
- La cantidad de críticas en la variable de Critic Count se encontraba entre 0 y 150.
- La cantidad de críticas en la variable de User Count se encontraba entre 0 y 2000.

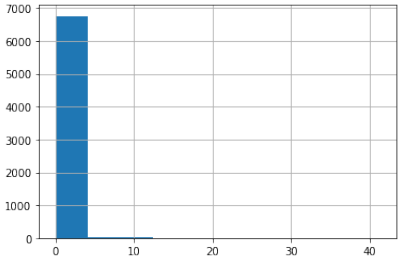
.

3.1.2 Distribución

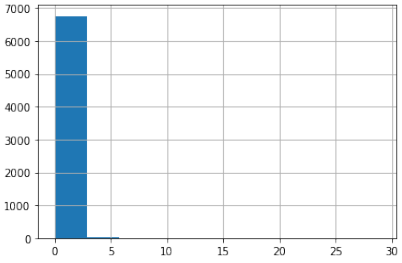
Year of release



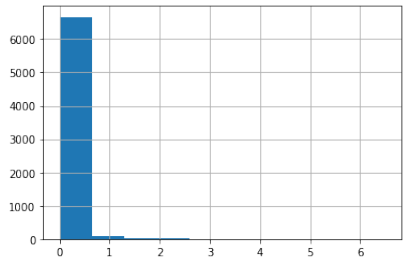
NA sales



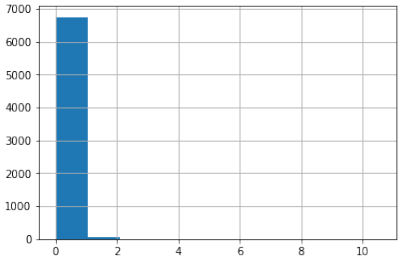
EU sales



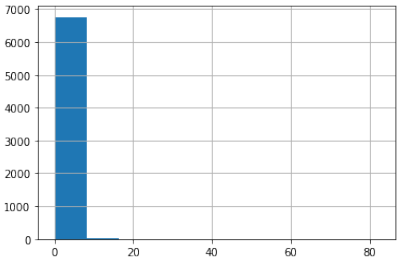
JP sales



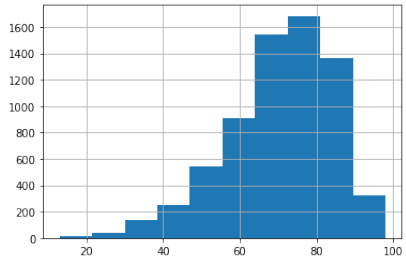
Other sales



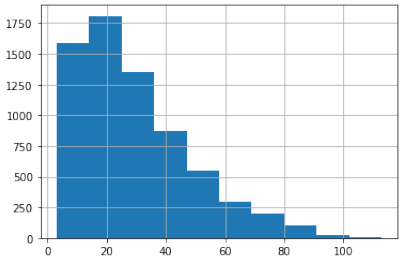
Global sales



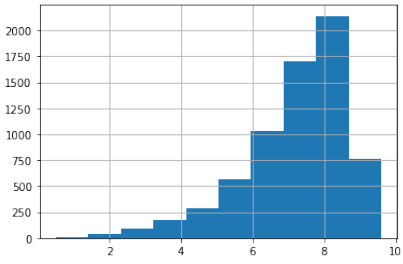
Critic Score



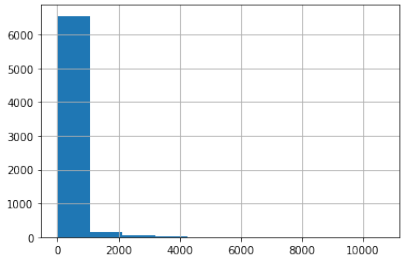
Critic Count



User Score



User Count



3.2 Categoricas

3.2.1 Revisión

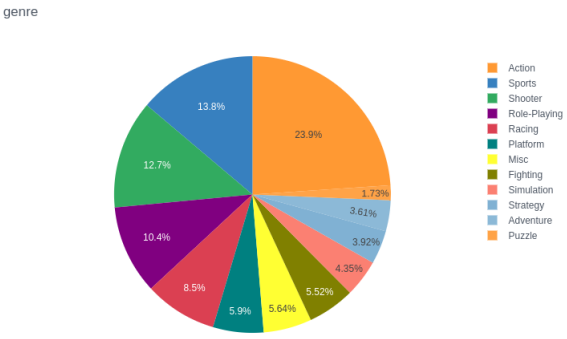
Se revisa la cantidad de categorias en cada una de las variables. Se encuentra :

- platform : 16
- genre : 12
- publisher : 262
- rating : 7
- name : 4365
- developer : 1289

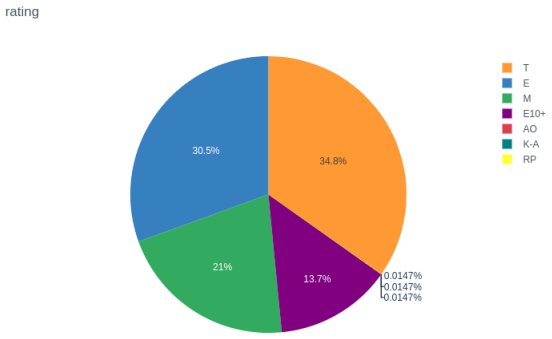
Se decide ignorar a las variables de Nombre y Desarrollador debido a su gran cantidad de valores únicos.

3.2.2 Distribución

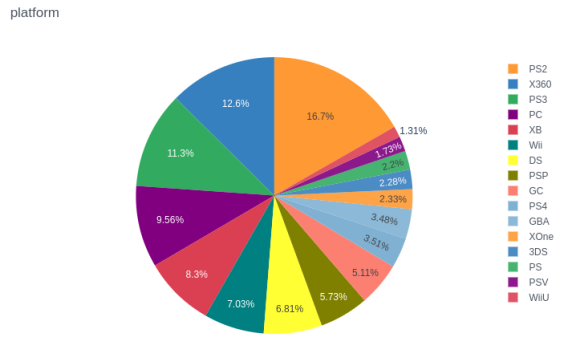
Platform



Genre



Rating



3.2.1 Normalización

Se normalizan las variables categóricas teniendo en cuenta que cada una de sus categorias debe tener una frecuencia de al menos 5%. Si su frecuencia es menor, se agrupan todas las categorias pequeñas en la categoría OTROS, si al final la categoría OTROS sigue siendo menor a 5% de frecuencia, se le incluye en la categoría de mayor frecuencia.
Los resultados son :

- platform : de 16 a 10 categorías, OTROS se vuelve la más común.
- genre : de 12 a 9 categorías.
- publisher : de 262 a 4 categorías.
- raiting : de 7 a 4 categorías.

4. Valores Extremos

Para limpiar el dataset de valores anómalos que pueden sesgar el agrupamiento, se tomó en cuenta la distribución de los datos. Todos aquellos datos por debajo del percentil 1 y por encima del percentil 99 fueron eliminados. Se encontró que el 6.68% de los registros se clasifican como extremos con el criterio dado. Debido a que el porcentaje de valores extremos encontrado es bajo, s decide simplemente eliminar dichos registros.

5. Varianza Baja

Para el agrupamiento de los datos, es necesario emplear características de los mismos que sean suficientemente diversas para poder remarcar una diferenciación. Para comprobar que los datos son variados, se mide su varianza, se establece que se debe tener un valor de al menos 0.1 y las variables que no cuenten con dicha varianza, no serán consideradas para el agrupamiento. Las variables que no pasan el filtro son :

— EU sales
— JP sales
— Other sales

.

6. Multicolinealidad

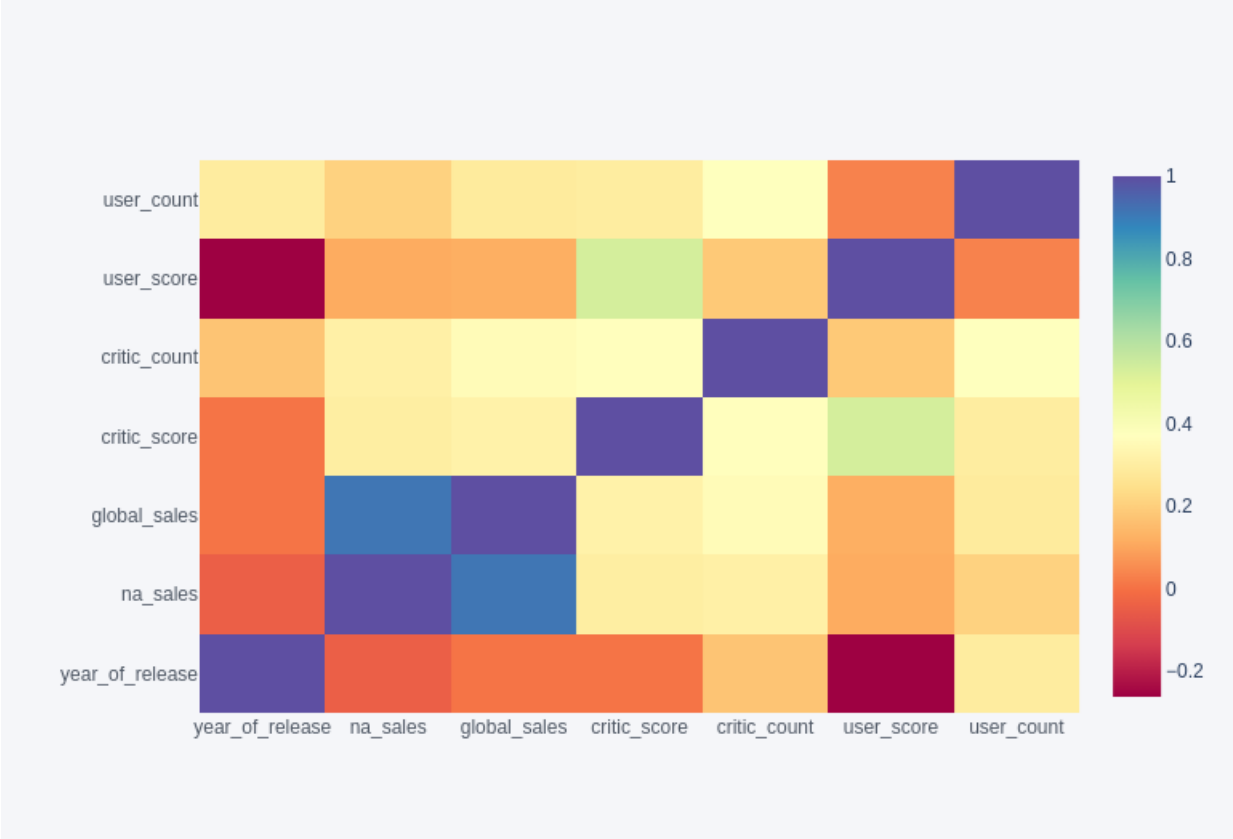
Para evitar redundancia en la información utilizada para el agrupamiento, se debe asegurar que las variables empleadas no presenten una fuerte correlación entre sí. Es por eso que se decide observar la multicolinealidad de las variables continuas restantes para reducir la dimensionalidad del problema. Se encuentra que la información contenida en las siete variables puede ser realmente expresada en únicamente en cuatro :

	Cluster	Variable	RS_Own	RS_NC	RS_Ratio	id
0	0	na_sales	0.957161	0.097049	4.744313e-02	1
1	0	global_sales	0.957161	0.148158	5.028961e-02	2
2	1	user_score	0.766536	0.068925	2.507463e-01	1
3	1	critic_score	0.766536	0.157168	2.769989e-01	2
4	2	user_count	0.686188	0.084277	3.426927e-01	1
5	2	critic_count	0.686188	0.113745	3.540874e-01	2
6	3	year_of_release	1.000000	0.078028	2.408365e-16	1

— 1. ventas
— 2. critica
— 3. número de críticas
— 4. año

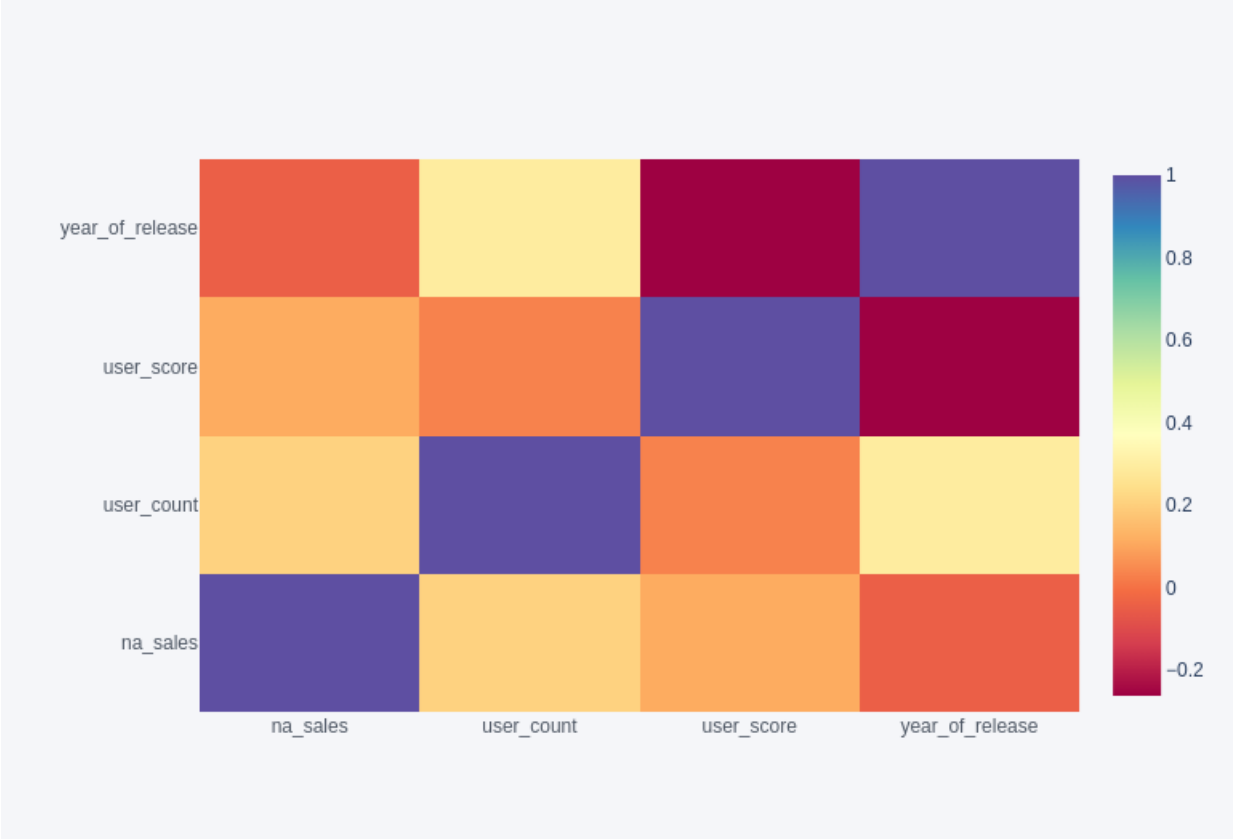
7. Correlación

Para mejor visualización, se observa la correlación de la variables antes de la reducción de dimensiones.



Se alcanza a apreciar una alta correlación entre las variables de venta, las variables de calificación y las variables de cantidad de criticas.

Si observamos las variables después de la reducción de dimensiones, se observa que la correlación no es apreciable.



8. Previsualización

Se visualizan los datos usando reducción de espacios para vislumbrar si es obvia la presencia de grupos. Para la reducción de espacios se usan tres métodos :

— 1. PCA

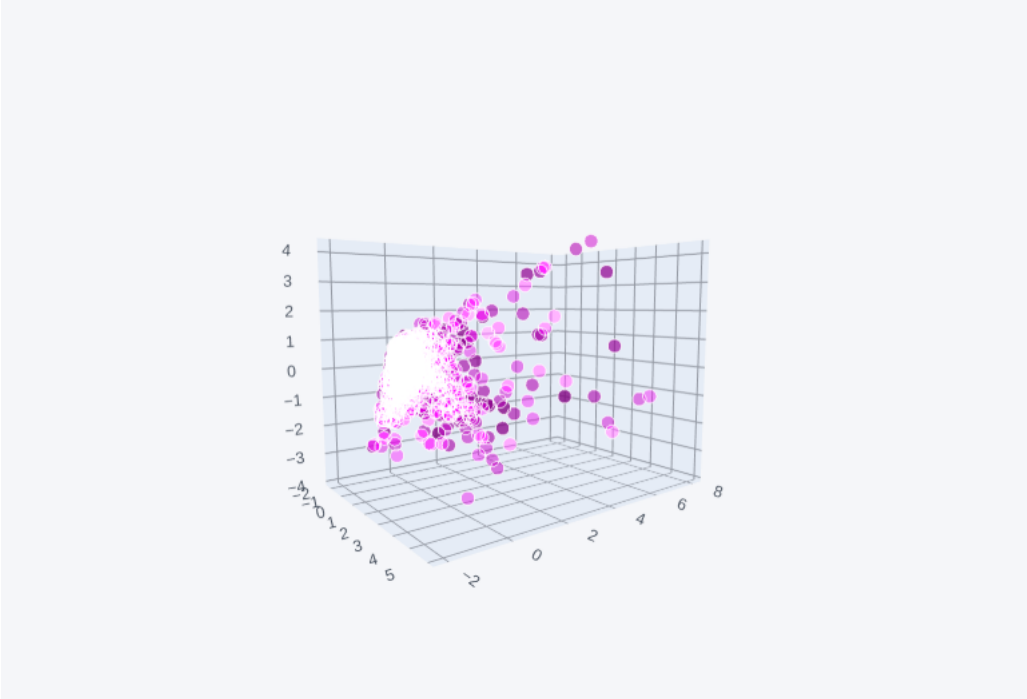
— 2. MDS

— 3. t-SNE

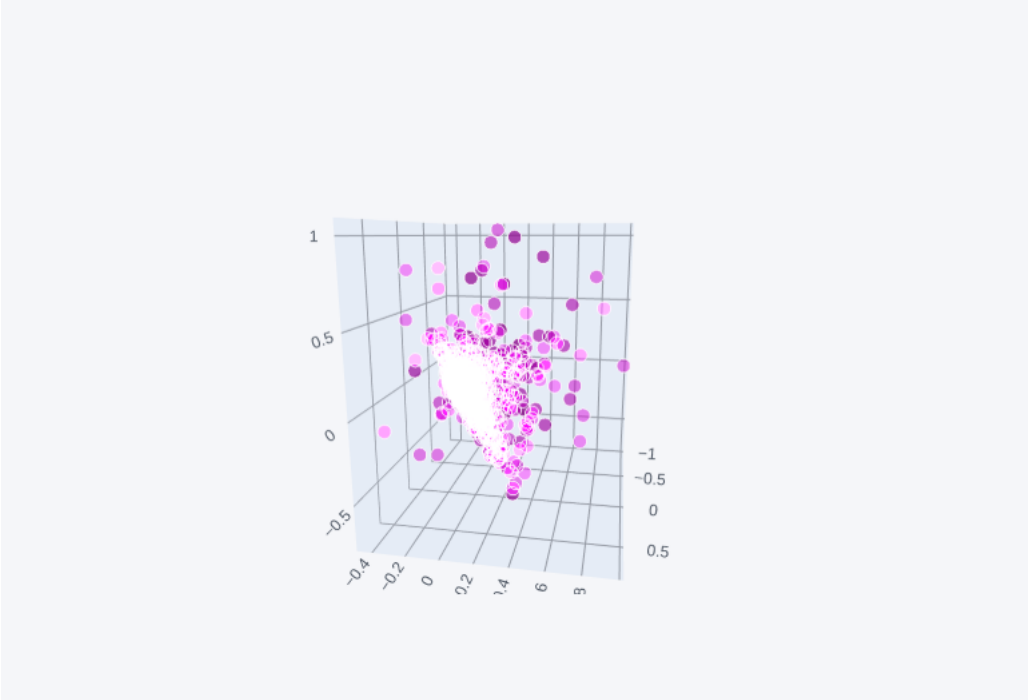
.

8.1 Vectores

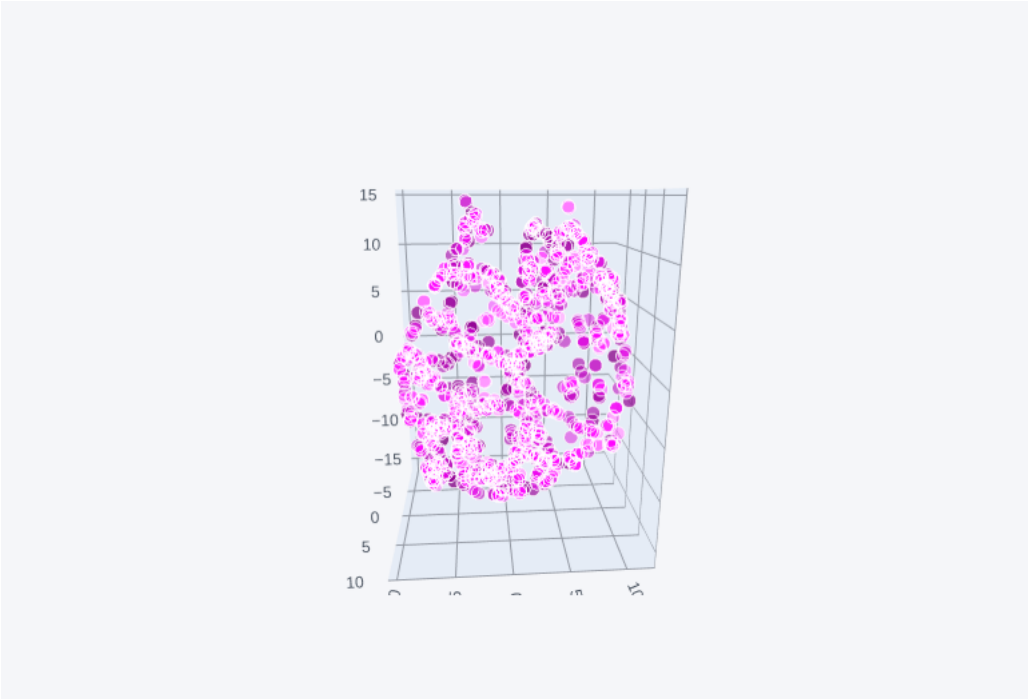
8.1.1 PCA



8.1.2 MDS

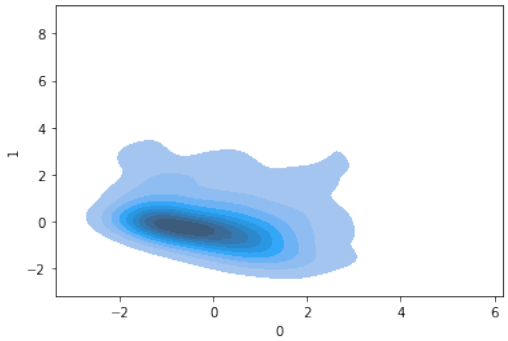


8.1.3 t-SNE

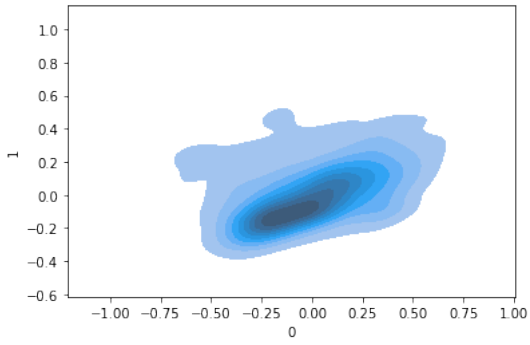


8.2 Densidad

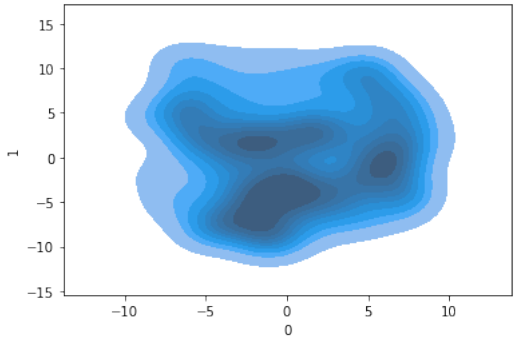
8.2.1 PCA



8.2.2 MDS



8.2.3 t-SNE



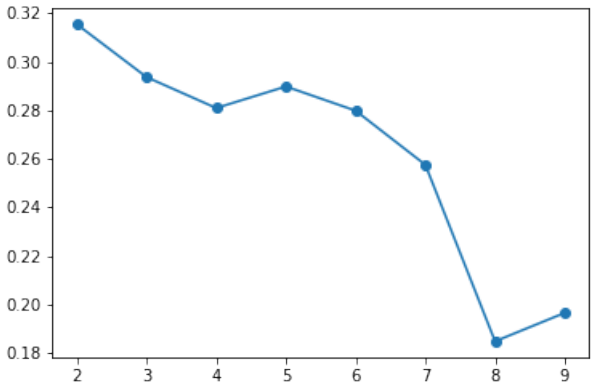
No se aprecia facilmente la presencia de clusters claramente diferenciados. Al parecer todos los datos se concentran en un único cúmulo , sin embargo, la gráfica de densidad con el método t-SNE podría indicar la existencia de al menos tres grupos.

9. Agrupamiento

Se reescalan las variables para poder tener uniformidad en la escala de cada una de las dimensiones. Se usa escalamiento de mínimo máximo.
Se utiliza el estadígrafo de silueta para seleccionar el número adecuado de grupos a ajustar para cada método.

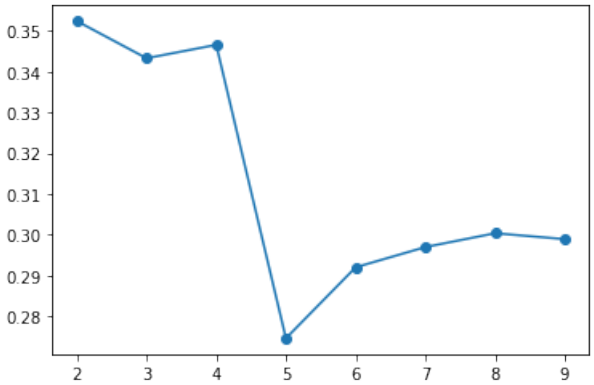
9.1 Modelado

9.1.1 Agrupamiento



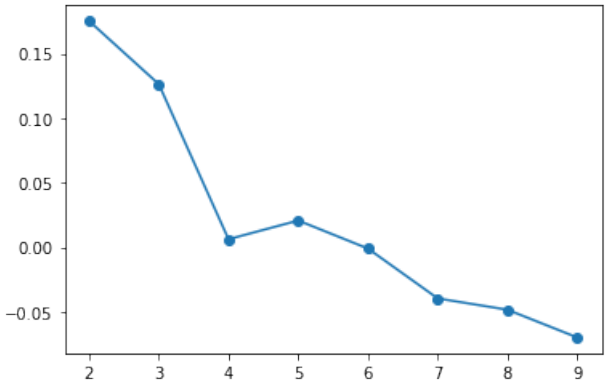
Número de clusters : 3

9.1.2 K-means



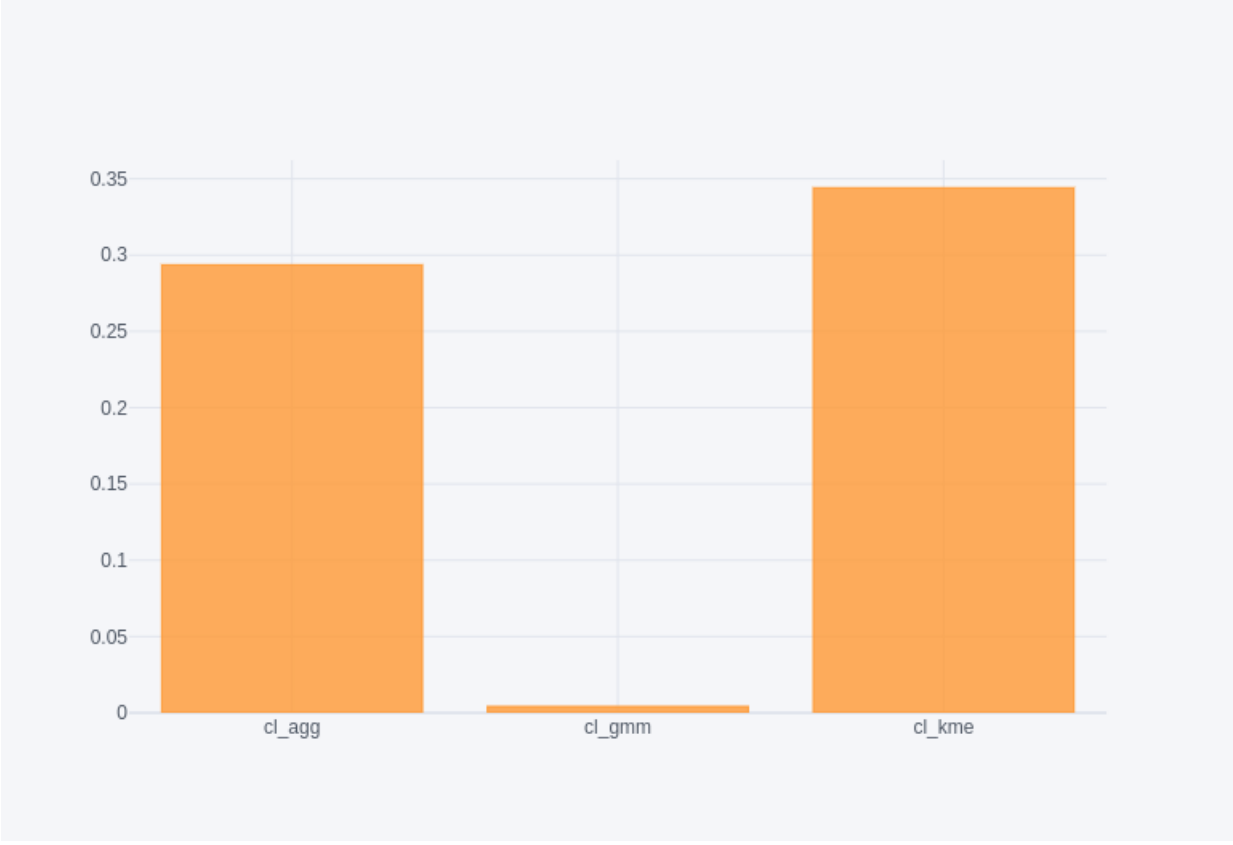
Número de clusters : 4

9.1.2 Modelos Gaussianos Mixtos



Número de clusters : 3

9.2 Seleccción final



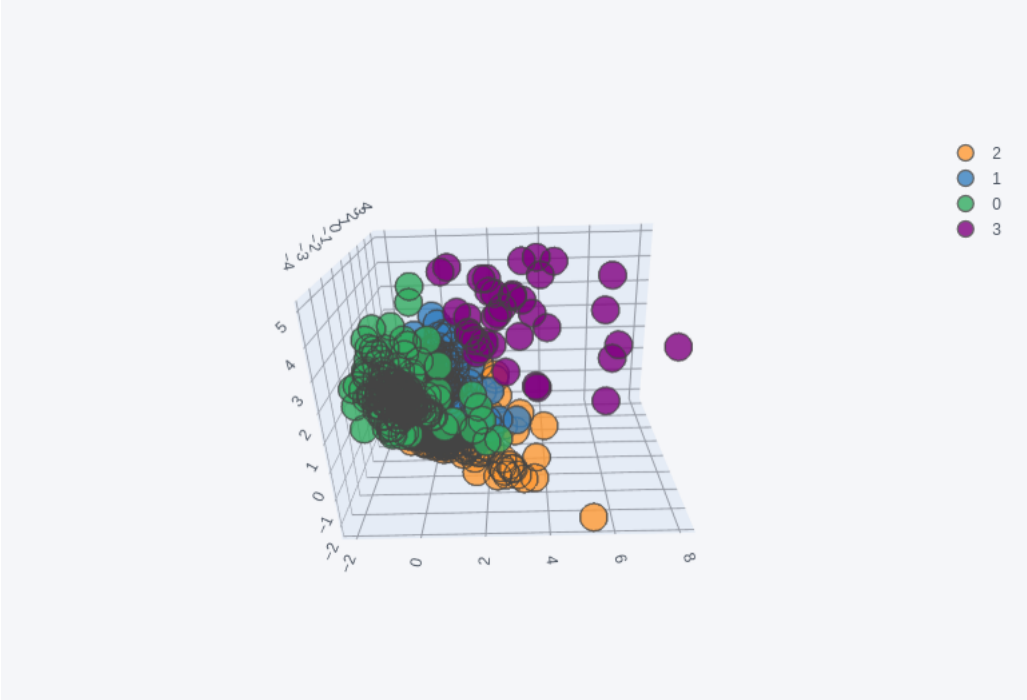
Se selecciona el agrupamiento realizado por el módelo K-means por su valor en el estadígrafo de silueta.

9.3 Visualización con clusters

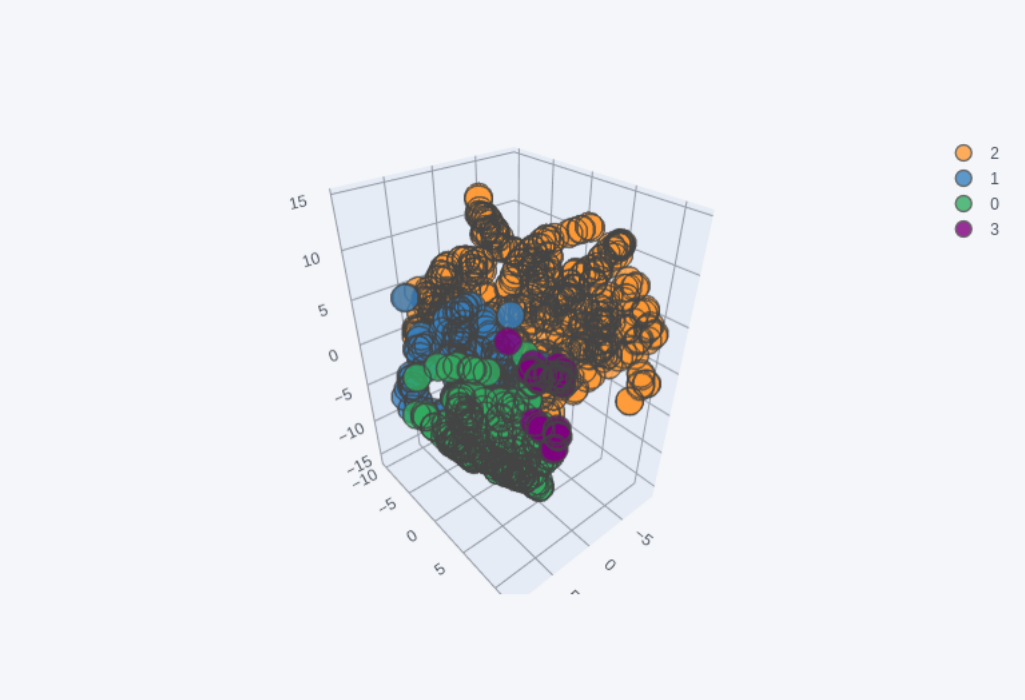
Se visualizan nuevamente los datos para apreciar las agrupaciones.

9.3.1 Vectores

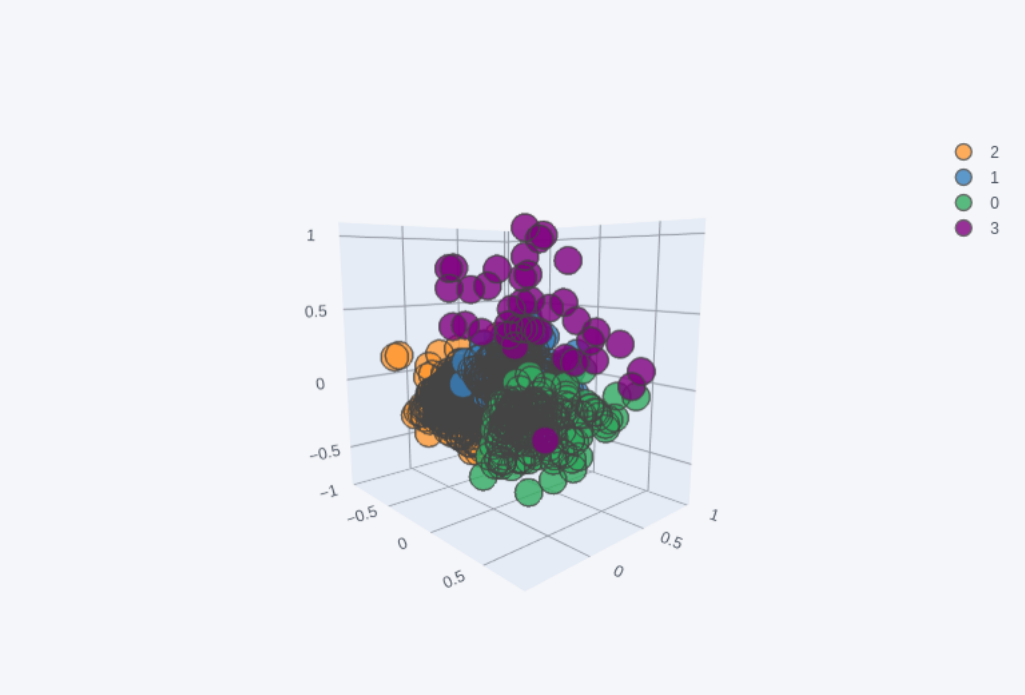
9.3.1.1 PCA



9.3.1.2 MDS

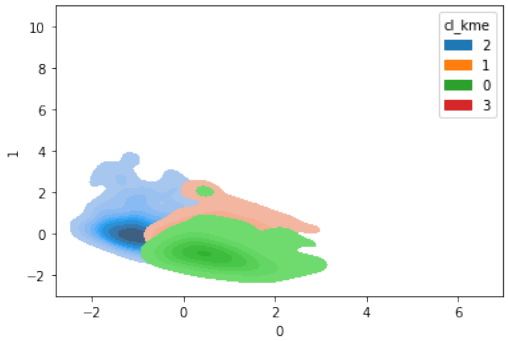


9.3.1.3 t-SNE

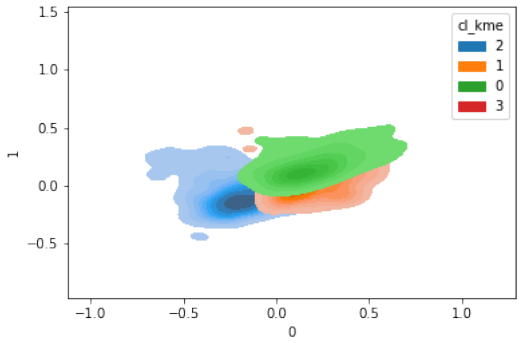


9.3.2 Densidad

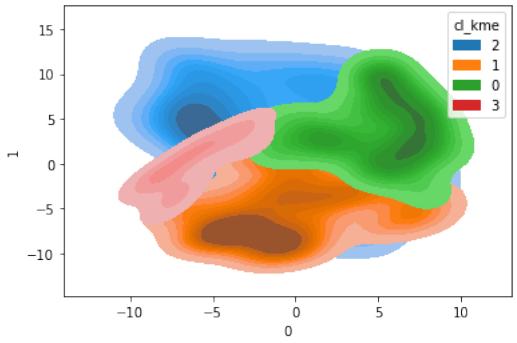
9.3.2.1 PCA



9.3.2.2 MDS



9.3.2.3 t-SNE

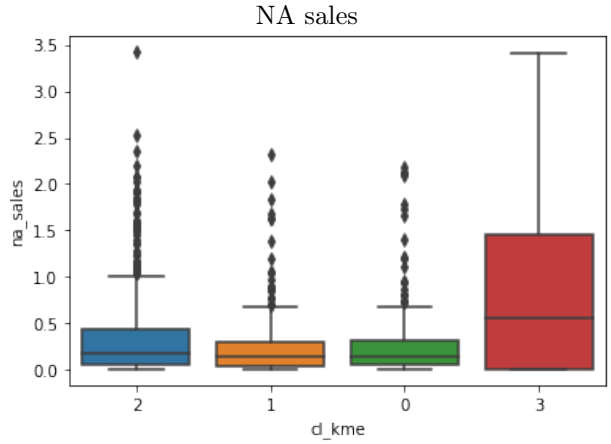


Se aprecia que a pesar de que los datos se encuentran muy cercanos entre sí, aún se puede remarcar la presencia de grupos.

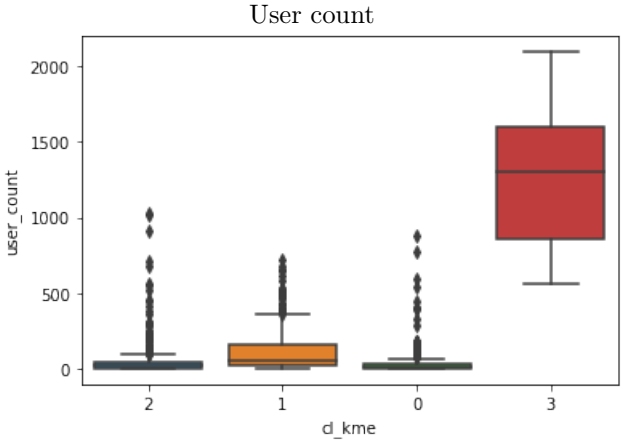
10. Perfilamiento

10.1 Continuas

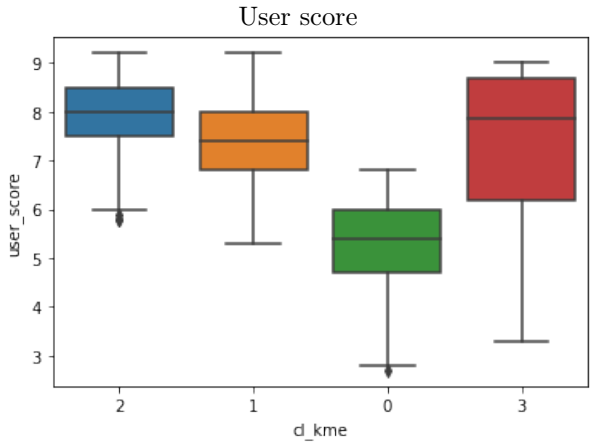
Se compara a todos los clusters entre sí según la variable utilizada para su definición, utilizando la prueba de Tukey. Así se puede ver si dicha característica tiene la capacidad de diferenciar efectivamente a los grupos en cuestion.



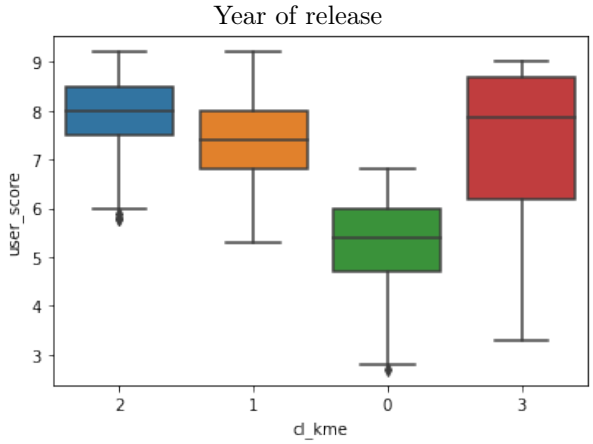
El cluster 0 no puede ser diferenciado del 1 co esta variable.
El cluster 0 no puede ser diferenciado del 2 co esta variable.



El cluster 0 no puede ser diferenciado del 2 co esta variable.



El cluster 1 no puede ser diferenciado del 3 co esta variable.

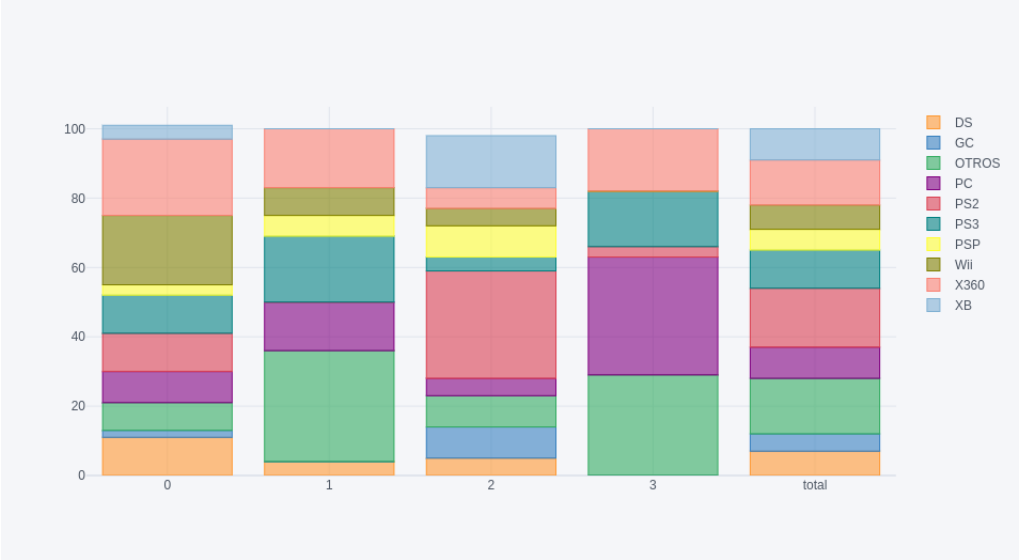


El cluster 1 no puede ser diferenciado del 3 co esta variable.

10.2 Discretas

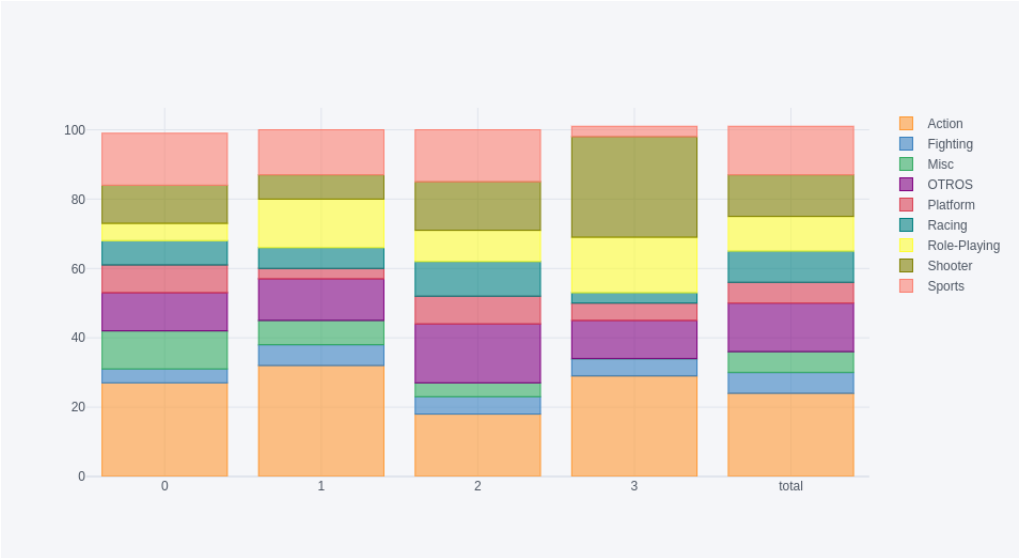
Se observa la distribución por cluster en función de cada una de las variables categírcas. A través de la prueba de chi cuadrada se determina si la distribución categórica de cada cluster lo logra diferencias satisfactoriamente de la distribución general.

Platform



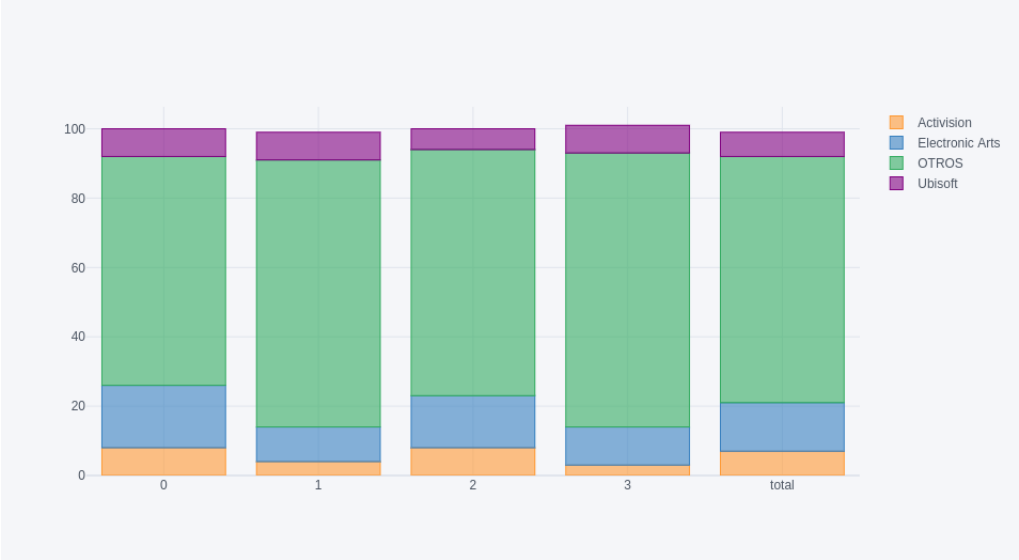
Todos los clusters se diferencían satisfactoriamente.

User count



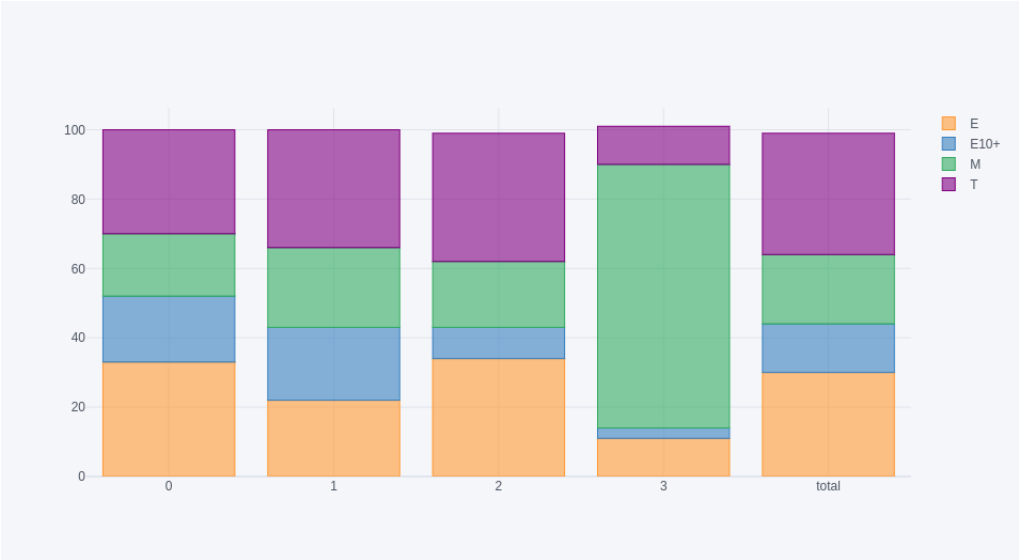
Ninguno de los clusters se diferencía por esta variable.

User score



Ninguno de los clusters se diferencía por esta variable.

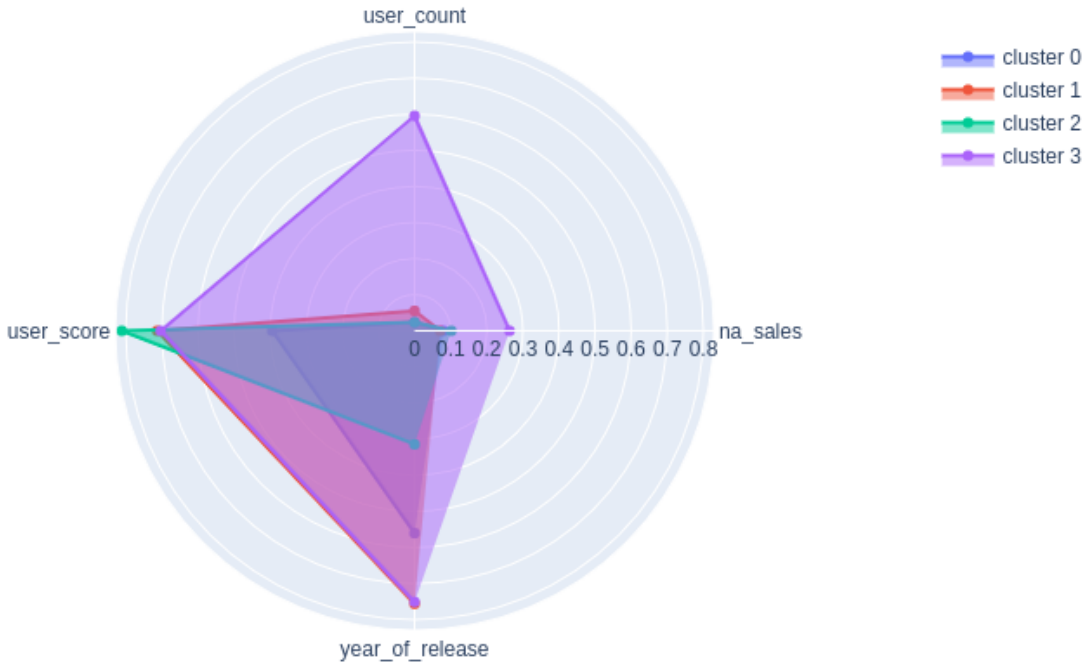
Year of release



Sólo el cluster 3 se diferencía satisfactoriamente.

10.3 Gráfico Radial

cl_kme	na_sales	user_count	user_score	year_of_release
0	0.271186	54.340206	5.263402	2008.525773
1	0.243606	119.442379	7.332714	2011.851301
2	0.347655	53.272545	7.971343	2004.344689
3	0.900263	1245.684211	7.271053	2011.763158



10.4 Arquetipos

Usando la información de esta sección, se puede vislumbrar el estructuramiento de los grupos formados. Se observa que el año es la variable que más diferencias marca entre todos los grupos, a excepción del cluster 3 y 1. Es por esto mismo, que la única variable categórica que puede dividir efectivamente en clusters al dataset es la de las plataformas, ya que cada consola de videojuegos tiene un periodo de vida respecto a la publicación de videojuegos.

— Cluster 0 : Videojuegos publicados alrededor del año 2008, con calificaciones bajas (5.2 en promedio), bajo número de críticas (54 en promedio) y el más bajo promedio de ventas (0.27 millones de copias). De forma general, este cluster representa a juegos con muy poco éxito.

— Cluster 1 : Videojuegos publicados alrededor del año 2011, con calificaciones medias (7.1 en promedio), número medio de críticas (119 en promedio) pero ventas bajas (0.24 millones de copias), las más bajas de entre los clusters de hecho. Este cluster se compone de los juegos más nuevos del dataset, que a pesar de tener buenas calificaciones, resultaron ser un fracaso en ventas.

— Cluster 2 : Videojuegos publicados alrededor del año 2004, con calificaciones muy altas (8.0 en promedio), pero con bajo número de críticas y de ventas. Este cluster representa juegos de años tempranos que recibieron muy altas calificaciones, pero no una base amplia de jugadores, ya haya sido por un bajo éxito de ventas o porque el mercado de los videojuegos no era tan grande como hoy en día.

— Cluster 3 : Videojuegos publicados alrededor del 2011, con calificaciones medias (7.3 en promedio) y un gran número de críticas por parte de usuarios junto con ventas. Este cluster representa el de los videojuegos más exitosos en ventas y en popularidad. Son juegos que a pesar de no destacar mucho en sus calificaciones, pudieron asegurar una amplia base de fans.

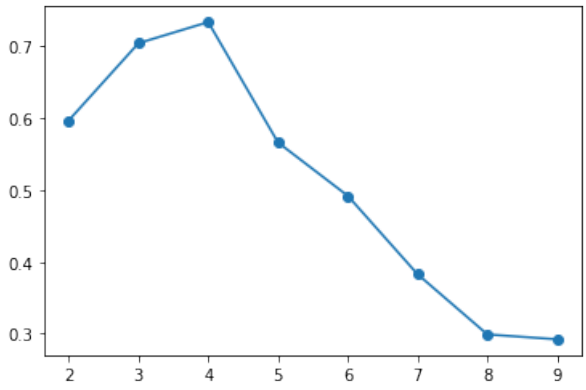
Debido a que la variable de año de publicación tiene una influencia tan grande en la formación de grupos, se decide volver a clusterizar los datos omitiendo dicha variable, para intentar obtener una mejor agrupación en función del éxito de los juegos.

11. Agrupamiento excluyendo la variable de Año de Publicación

Con el propósito de separar a los videojuegos principalmente en función de su éxito, se vuelve a agrupar excluyendo a la variable de año, la cual tenía una influencia muy fuerte en el primer intento de seccionar la tabla de datos.

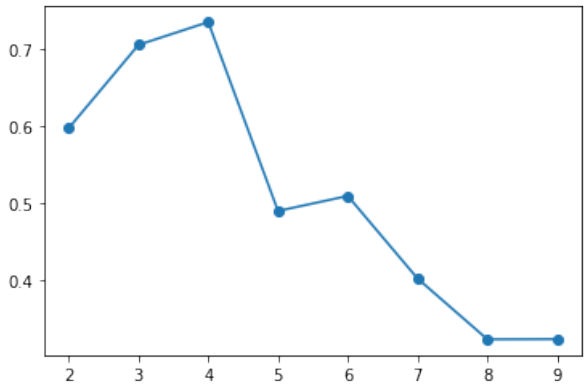
11.1 Modelado

11.1.1 Agrupamiento



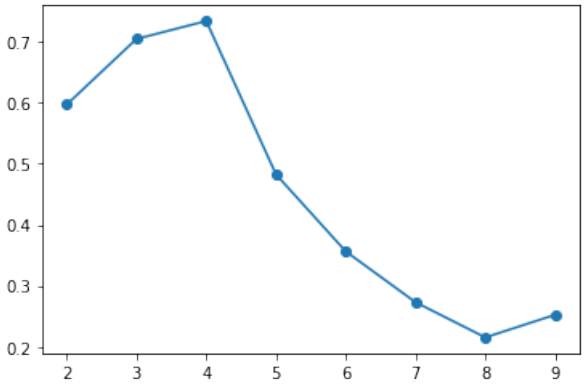
Número de clusters : 4

11.1.2 K-means



Número de clusters : 4

11.1.2 Modelos Gaussianos Mixtos



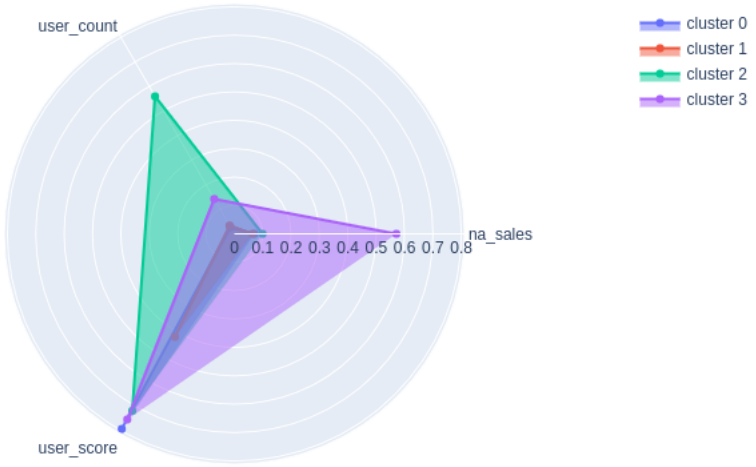
Número de clusters : 4

11.2 Seleccción final

Se selecciona el agrupamiento realizado por el módelo Agrupamiento por su valor en el estadígrafo de silueta.

11.3 Perfilamiento

Para simplificar el perfilamiento de los clusters creados, se empleará unicamente su distribución basada en las variables tomadas en cuenta para el agrupamiento.



— Cluster 0 : Concentra videojuegos con malas ventas y una baja cantidad de críticas, pero buenas calificaciones.

— Cluster 1 : Concentra los videojuegos con el menor éxito, bajas ventas, baja cantidad de críticas y malas calificaciones.

— Cluster 2 : Concentra videojuegos con una enorme cantidad de críticas y muy buenas calificaciones, pero sorpresivamente, bajas ventas.

— Cluster 3 : Concentra videojuegos que tiene un número relativamente bajo de críticas, pero buenas calificaciones y excelenetes ventas.