Reporte de investigación

PELICULAS DE MARVEL vs DC COMICS

Irma Eunice Martínez de la Cruz

2022-06-05

* INTRODUCCIÓN

Esta investigación mostrara lo estudiado en la experiencia educativa de Estadistica Multivariada, donde se reflejaran las practicas realizadas a traves de una selección de una matriz nueva de datos. Asi mismo representaremos los datos a traves de representaciónes graficas con gerramientas como son algunos métodos para realizar ese analisis.

Tema:

Peliculas de la compañia Marvel y DC Comics.

* DESCRIPCIÓN DE LA MATRIZ DE DATOS

La base de datos “Marverl_DC”, contiene datos que se basan en cuestion de la calificación que obtuvieron ante el Internet Movie Database (IMDb, en español: Base de datos de películas en Internet), que tuvo cada pelicula, la duración, el presupuesto y el bruto total que tuvieron a nivel nacional y mundial, en el lapso de los años 2000 a 2019.

Obtención de datos:

Esta base de datos fue descargada de la pagina "Kaggle", realizado por el autor

¿En qué consisten los datos?

De acuerdo con las peliculas de las compañias, se llevo a cabo variables que consisten el calificaiones dadas por el publico, el presupuesto que tuvieron que tener para cada una de las peliculas, el total bruto que recaudaron por cada una, etc.

*EXPLORACIÓN DE LA MATRIZ

Se carga la base de datos

```
install.packages("knitr")
library(knitr)
```

```
library(readxl)
Marvel_DC <- read_excel("Marvel_DC.xlsx")
kable(head(Marvel_DC))
```

...1	Original Title	Company	Rate	Metascore	Minutes	Release	Budget	Opening Weekend USA	Gross USA	Gross World-wide
1	Iron Man	Marvel	7.9	79	126	2008	140000000	98618668	318604126	585366247
2	The Incredible Hulk	Marvel	6.7	61	112	2008	150000000	55414050	134806913	263427551
3	Iron Man 2	Marvel	7.0	57	124	2010	200000000	128122480	312433331	623933331
4	Thor	Marvel	7.0	57	115	2011	150000000	65723338	181030624	449326618
5	Captain America: The First Avenger	Marvel	6.9	66	124	2011	140000000	65058524	176654505	370569774
6	The Avengers	Marvel	8.0	69	143	2012	220000000	207438708	623357910	1518812988

Dimensión:

```
dim(Marvel_DC)
```

```
## [1] 39 11
```

Nombre de las variables

```
colnames(Marvel_DC)
```

```
##  [1] "...1"               "Original Title"     "Company"
##  [4] "Rate"               "Metascore"          "Minutes"
##  [7] "Release"            "Budget"             "Opening Weekend USA"
## [10] "Gross USA"          "Gross Worldwide"
```

Tipo de variables

```
str(Marvel_DC)
```

```
## tibble [39 x 11] (S3: tbl_df/tbl/data.frame)
##  $ ...1               : num [1:39] 1 2 3 4 5 6 7 8 9 10 ...
```

```
##  $ Original Title     : chr [1:39] "Iron Man" "The Incredible Hulk" "Iron Man 2" "Thor" ...
##  $ Company            : chr [1:39] "Marvel" "Marvel" "Marvel" "Marvel" ...
##  $ Rate               : num [1:39] 7.9 6.7 7 7 6.9 8 7.2 6.9 7.7 8 ...
##  $ Metascore          : num [1:39] 79 61 57 57 66 69 62 54 70 76 ...
##  $ Minutes            : chr [1:39] "126" "112 " "124 " "115" ...
##  $ Release            : num [1:39] 2008 2008 2010 2011 2011 ...
##  $ Budget             : chr [1:39] "140000000" "150000000" "200000000" "150000000 " ...
##  $ Opening Weekend USA: num [1:39] 9.86e+07 5.54e+07 1.28e+08 6.57e+07 6.51e+07 ...
##  $ Gross USA          : num [1:39] 3.19e+08 1.35e+08 3.12e+08 1.81e+08 1.77e+08 ...
##  $ Gross Worldwide    : num [1:39] 5.85e+08 2.63e+08 6.24e+08 4.49e+08 3.71e+08 ...
```

Presencia de NA´s.

```
anyNA(Marvel_DC)
```

```
## [1] FALSE
```

* TRATAMIENTO DE LA MATRIZ

Para este caso se le selecciono el tratamiento del "Analisis Factorial", ya que existen variables numericas que se pueden correlacionar entre sí.

* METODOLOGIA DEL ANALISIS

ANALISIS FACTORIAL- MARVEL vs DC Comics ## Descripción del método utilizado El análisis factorial es un método de reducción estadística que tiene como objetivo explicar las posibles correlaciones entre ciertas variables. Para ello, teniendo en cuenta el efecto de otras, los factores, que no son observables. Por tanto, lo que hace este análisis es reducir. Así, tomamos un número elevado de variables y, por medio de esta técnica, conseguimos reducirlas a un tamaño más manejable. Para hacerlo, se utilizan una serie de combinaciones lineales de las observadas con otras que no son visibles.

* RESULTADOS

1.- Se convierte la base de datos en un data.frame para convertirla en una matriz

```
x<-data.frame(Marvel_DC)
```

2.- Se crea una nueva matriz de datos donde se incluyen las variables 4,5,7,9,10 y 11 las cuales son númericas, mientras que se toman las 39 observaciones.

```
x1<-x[1:39,(c(4,5,7,9,10,11))]
```

3.- Se cargan las librerias para los graficos

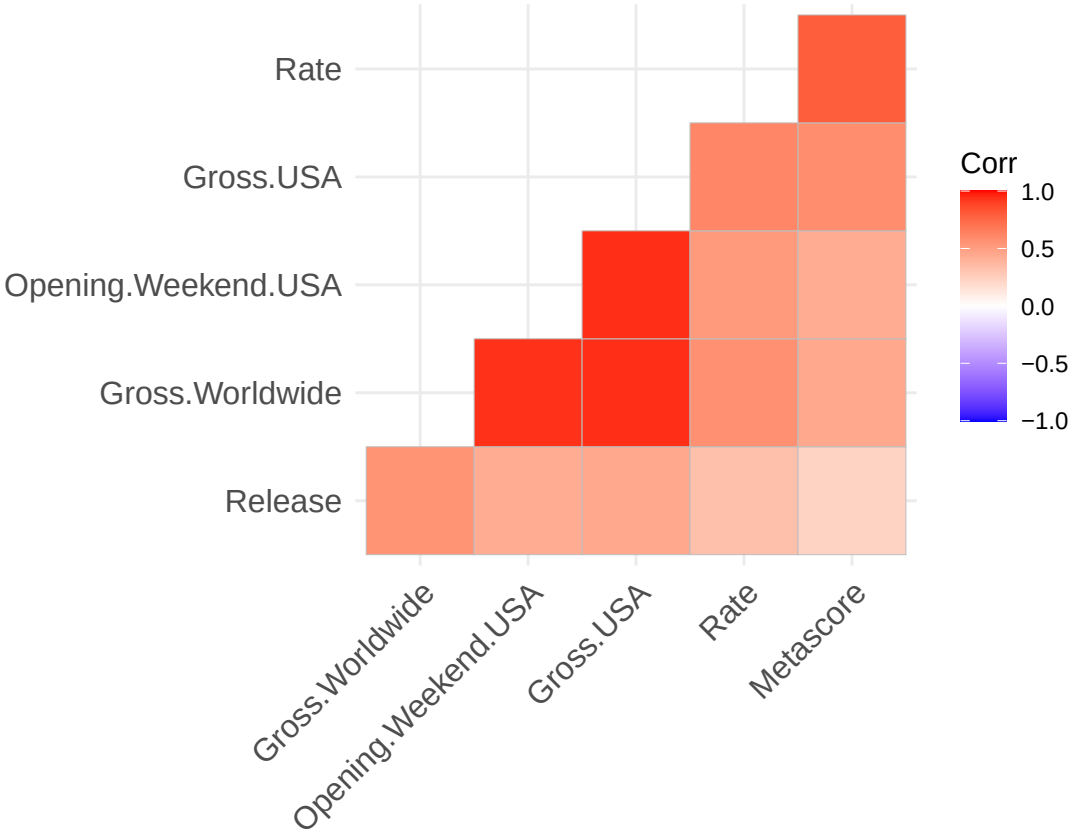
```
install.packages("psych")
library(psych)

install.packages("polycor")
library(polycor)

install.packages("ggcorrplot")
library(ggcorrplot)
```

4.- Matriz de correlaciones

```
R<-hetcor(x1)$correlations
```

5.- Gráfico de correlaciones

```
ggcorrplot(R, type="lower", hc.order=TRUE)
```



Observamos que no existe ningun tipo de correlación negativa, si no positiva.

6.- Se utiliza la prueba de esfericidad de Bartlett

```
p_Bartlett<-cortest.bartlett(R)
```

Visualización del p-valor

```
p_Bartlett$p.value
```

```
## [1] 7.375724e-127
```

Ho:Las variables estan relacionadas Ha:Las variables no están correlacionadas.

Para esta ocasión, No se rechaza Ho, ya que las variables estan correlacionadas.

7.-Criterio Kaiser-Meyer.Olkin Se identifica si los datos analizados son adecuados para el analisis factorial, >0.00 a 0.49-No adecuados >0.50 a 0.59-Poco adecuados >0.60 a 0.69-Aceptables >0.70 a 0.89-Buenos >0.90 a 1.00-Excelentes

```
KMO(R)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = R)
## Overall MSA =  0.76
## MSA for each item =
```

```
##                Rate           Metascore             Release Opening.Weekend.USA
##                0.76                0.62                0.78                 0.83
##           Gross.USA     Gross.Worldwide
##                0.76                0.77
```

Generalizando, son buenos los datos para hacer un analisis factorial.

8.- Extracción de los factores con el máximo de verosimilitud y el minimo residuo

```
modelo1<-fa(R, nfactor=3, rotate="none", fm="mle")
modelo2<-fa(R, nfactor=3, rotate="none", fm="minres")
```

9.- Se extrae el resultado de las comunidalidades, donde se encuentra la varianza explicada, para extraer la mejor variable.

```
C1<-sort(modelo1$communality, decreasing=TRUE)
C2<-sort(modelo2$communality, decreasing=TRUE)
```

```
head(cbind(C1,C2))
```

```
##                            C1        C2
## Metascore           0.9950034 0.9964580
## Gross.Worldwide     0.9936843 0.9895630
## Gross.USA           0.9645075 0.9579746
## Opening.Weekend.USA 0.9346712 0.9396379
## Rate                0.6854228 0.6821872
## Release             0.4190145 0.4390524
```

10.- Se extraen las unicidades, lo cual es el cuadro del coeficiente, del factor unico.

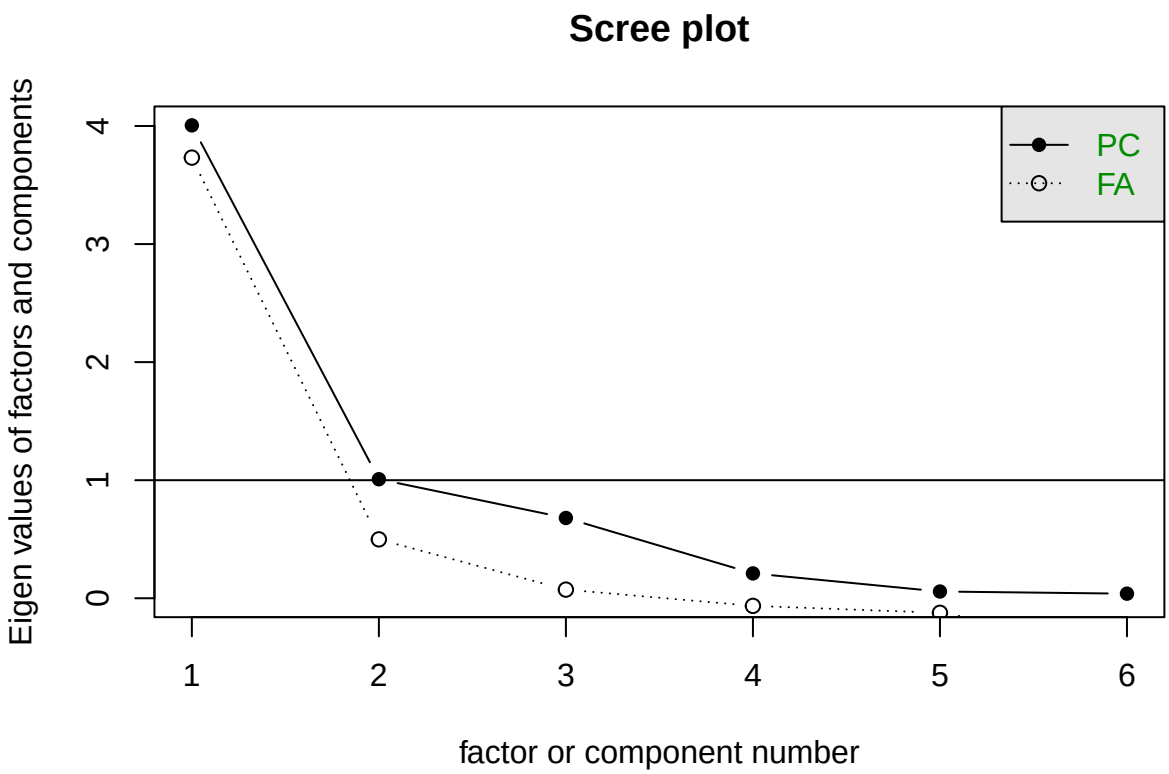
```
u1<-sort(modelo1$uniquenesses, decreasing = TRUE)
u2<-sort(modelo2$uniquenesses, decreasing = TRUE)
```

```
head(cbind(u1,u2))
```

```
##                              u1         u2
## Release             0.580985550 0.56094756
## Rate                0.314577159 0.31781285
## Opening.Weekend.USA 0.065328759 0.06036209
## Gross.USA           0.035492490 0.04202537
## Gross.Worldwide     0.006315736 0.01043703
## Metascore           0.004996631 0.00354202
```

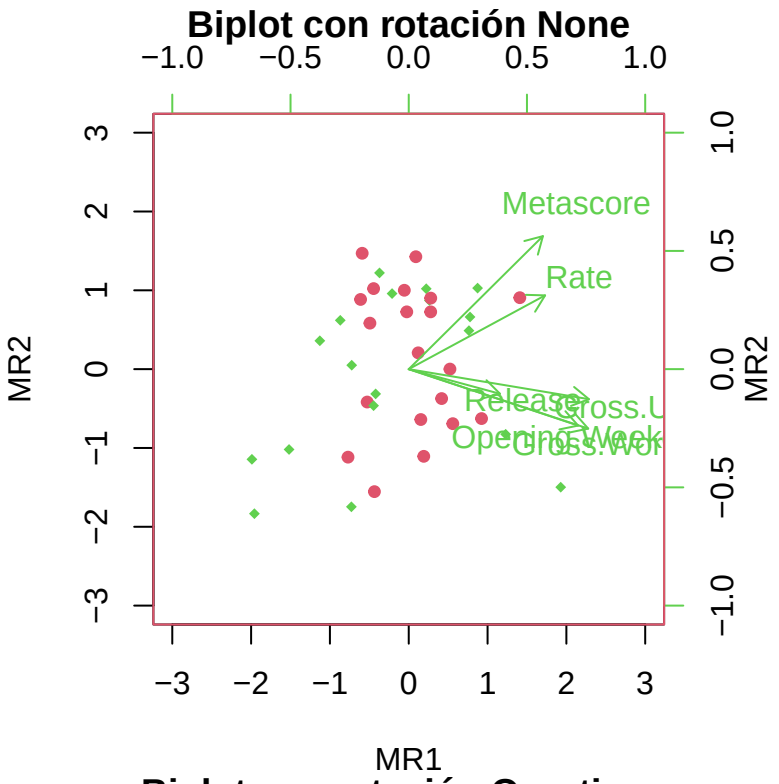
11.- Scree plot

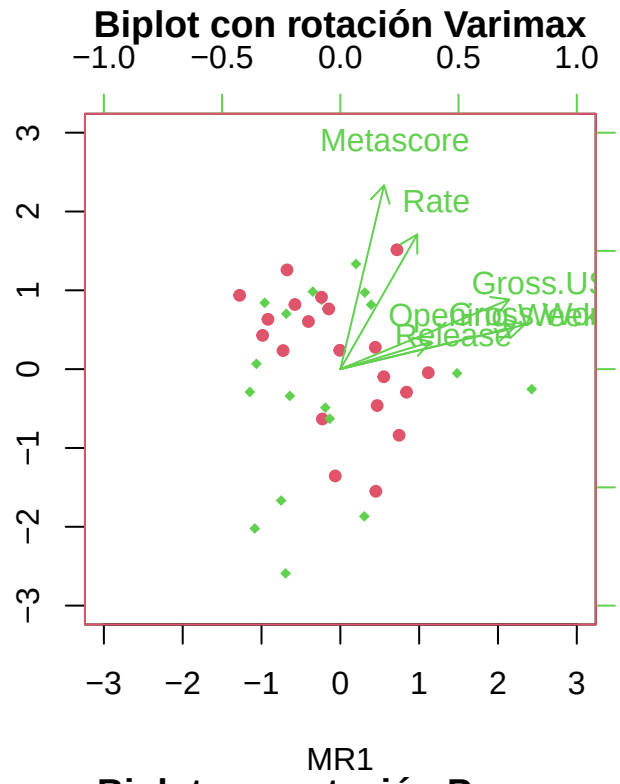
```
scree(R)
```

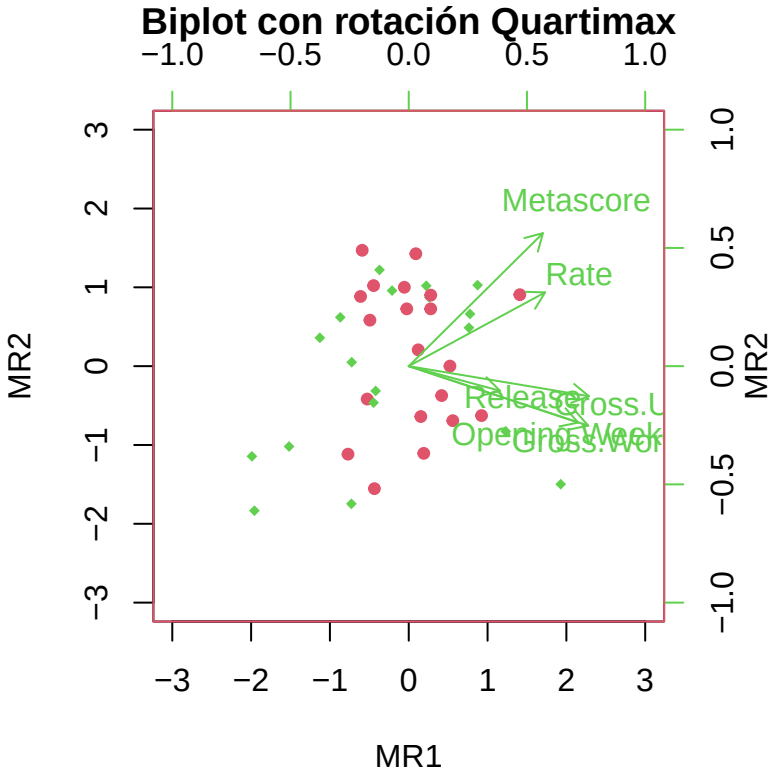


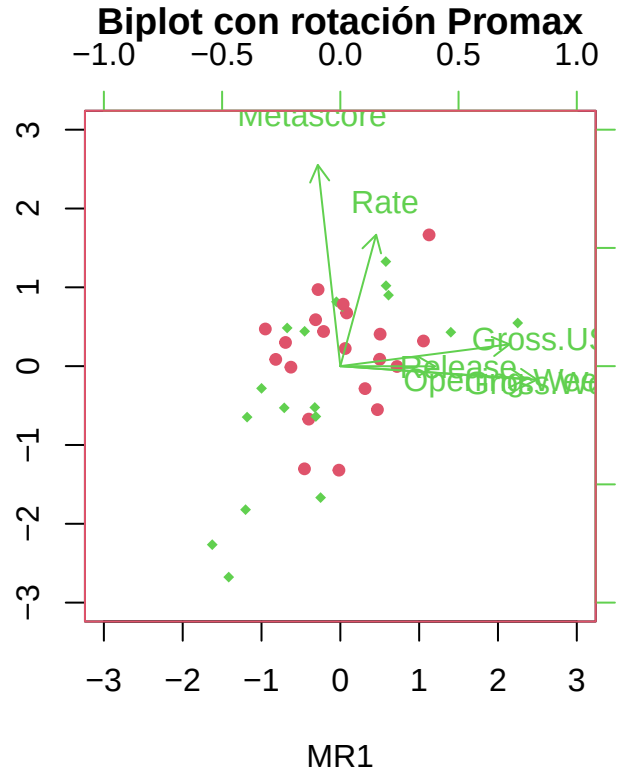
12.- Rotación de la matrizrepresentada con matriz

```
install.packages("GPArotation")
library(GPArotation)

rot<-c("None", "Varimax", "Quartimax", "Promax")
bi_mod<-function(tipo){
  biplot.psych(fa(x1, nfactors = 2,
                  fm= "minres", rotate=tipo),
               main = paste("Biplot con rotación", tipo),
               col=c(2,3,4), pch=c(21,18), group=bfi[,"gender"])
}
sapply(rot,bi_mod)
```









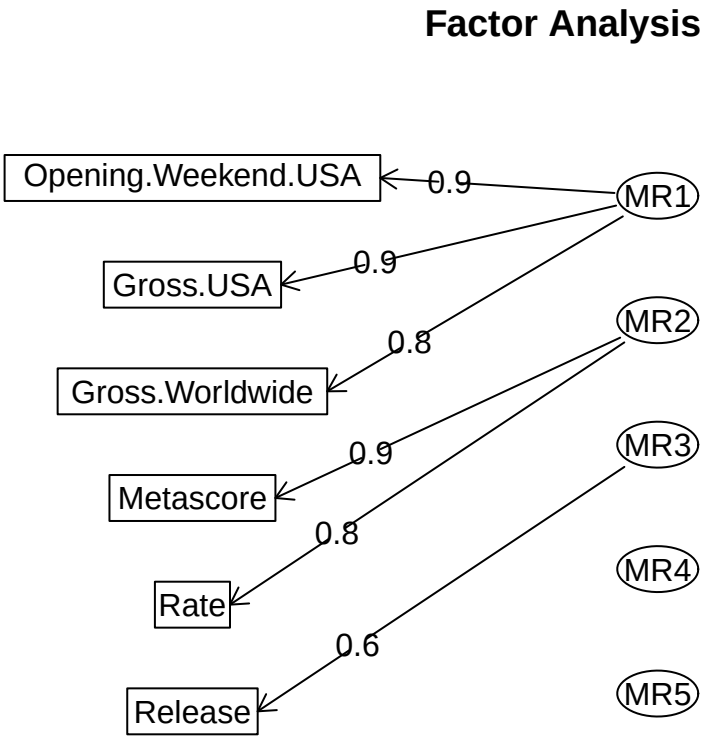

```
## $None
## NULL
##
## $Varimax
## NULL
```

```
##
## $Quartimax
## NULL
##
## $Promax
## NULL
```

* CONCLUSIONES

13.- Interpretación

```
modelo_varimax<-fa(R,nfactor=5,
                   rotate = "varimax",
                   fm="minres")

fa.diagram(modelo_varimax)
```



14.- Visualización de la matriz de carga rotada.

```
print(modelo_varimax$loadings, cut=0)
```

```
##
## Loadings:
##                     MR1    MR2    MR3    MR4    MR5
## Rate                 0.272  0.810  0.257  0.125 -0.010
## Metascore            0.220  0.886  0.086 -0.098  0.009
## Release              0.274  0.140  0.560  0.002 -0.004
## Opening.Weekend.USA  0.903  0.240  0.272  0.093 -0.004
## Gross.USA            0.859  0.394  0.284 -0.131 -0.025
## Gross.Worldwide      0.804  0.258  0.530  0.016  0.042
##
```

```
##                   MR1   MR2   MR3   MR4   MR5
## SS loadings    2.397 1.740 0.822 0.051 0.003
## Proportion Var 0.399 0.290 0.137 0.009 0.000
## Cumulative Var 0.399 0.689 0.826 0.835 0.835
```

De acuerdo a lo observado en los resultados, se puede ver que entre las variables elegidas, las cuales fueron *Rate,* Metascore, *Release,* Opening Weekend USA, *Gross.USA,* Gross Worldwide.

Podemos concluir que estas tienen un buena correlación, y ademas de ello, de que la variable "Opening Weekend USA" es la variable más factible, como se muestra en el gráfico.

* REFERENCIAS

Introducción y base de datos

Henrique, L. (2020, agosto). Kaggle. Kaggle. Recuperado 1 de junio de 2022, de https://www.kaggle.com/datasets/leonardopena/marvel-vs-dc