UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Maestría en Ciencia de Datos

Datos Masivos

Profesora: José Alberto Benavides Vázquez

Alumno: Sergio Bernal Cortez

Matricula: 1581391

Tarea 6 (reporte)

Reporte

Conjunto de datos

Se tiene un conjunto de datos sobre películas, dicho conjunto ya ha sido tratado y limpiado, sin embargo, en el proceso se creó una columna adicional, esto ocasionó que se crearan nuevas entradas con errores, específicamente datos vacíos.

Los datos tienen diversas columnas, en las que conviven datos numéricos y datos de texto, para efectos de esta práctica trabajé únicamente con los datos numéricos.

En resumen, los datos numéricos utilizados son los siguientes:

```
+------------+----------+---------+----------+
|vote_average|   revenue|   budget|popularity|
+------------+----------+---------+----------+
|       8.417| 701729206|165000000|   140.241|
|       8.512|1004558444|185000000|   130.643|
|       7.573|2923706026|237000000|    79.932|
|        7.71|1518815515|220000000|    98.082|
|       7.606| 783100000| 58000000|    72.735|
+------------+----------+---------+----------+
```

Regresión lineal

Con dichos datos, se realizó una prueba de regresión, tanto para la ganancia (revenue) y para la popularidad (popularity).

Los datos obtenidos del RMSE (Raíz del error cuadrado medio) fueron los siguientes

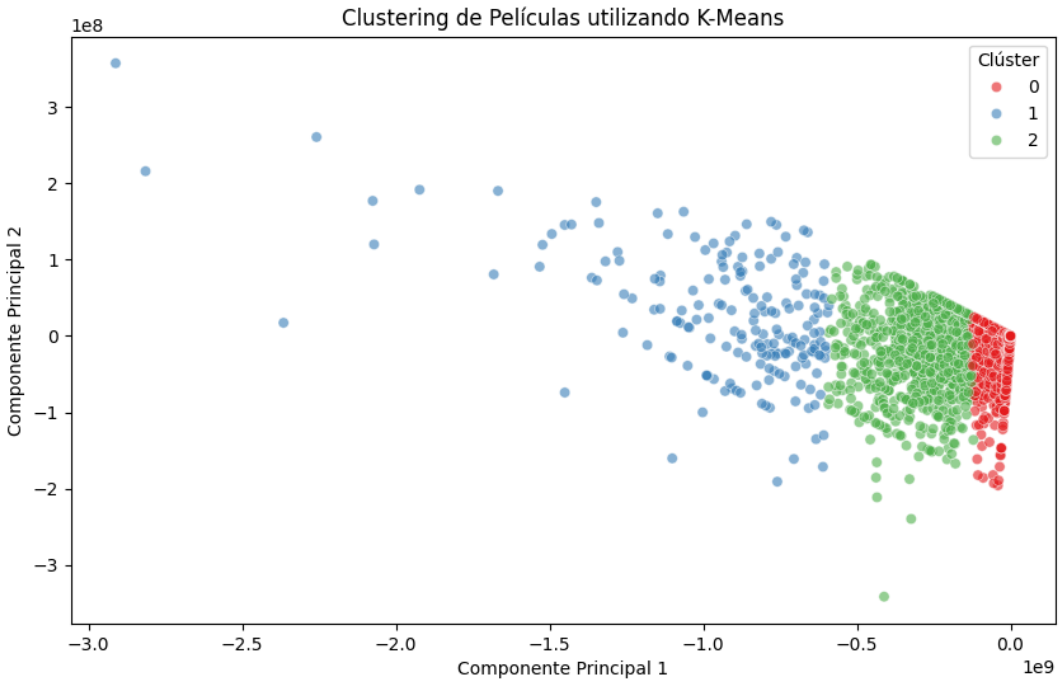
Concepto	RMSE
Ganancia	31,357,704
Popularidad	22.71

El error cuadrado medio de las ganancias podría parecer algo elevado, sin embargo en el contexto de que son ingresos de películas, realmente es un dato dentro de lo normal a final de cuentas, teniendo datos que van desde 0 hasta 2.9 mil millones.

Por el lado de popularidad, tomando en cuenta los datos, hay un comportamiento “raro” ya que no se especifica sobre que escala se mide, para este concepto, la base es 0 igual que las ganancias pero el máximo es de 2994, es decir, no está en una escala bien definida.

Análisis de componentes principales

Analizando primeramente mediante clúster (3) y luego mediante componentes principales (2) se graficaron los datos por las características mencionadas al inicio dando como resultado lo siguiente:



Revisando la densidad de los clúster, se puede ver de cierta manera bien segregados los el clúster 0 y el 3 con una densidad aceptable pero con un porcentaje de datos algo separados, por otro lado el clúster en color azul es el más disperso y no parece haber tantos puntos como en los otros dos.

Siendo películas y revisando los datos, se ve que hay muchos casos extremos, por ejemplo donde una película no tiene tanto presupuesto y puede tener muy buena calificación, pero esto no necesariamente se refleja las ganancias.

En conclusión, con estas pruebas y consultando nuevamente los datos ya limpios sigue habiendo posibilidad de pulirlos más, para obtener medidas más claras y con mayor sentido, además de que se puede explorar los componentes de texto de la base de datos.