Universidad del Valle de Guatemala

Minería de Datos

HOJA DE TRABAJO #2
ANÁLISIS DE "CLUSTERING"

11/03/2020

Análisis de Clustering sobre películas

Pablo Sao – 11530

Mercedes Retolaza - 16339

Contenido

Introducción

Es también conocido como agrupamiento, es una técnica de minería de datos. El proceso consiste en la división de los datos en grupos de objetos similares. Cuando se representan la información obtenida a través de clústeres se pierden algunos detalles de los datos, pero a la vez se simplifica dicha información.

Técnica en la que el aprendizaje realizado es no supervisado. Desde un punto de vista práctico. El Clustering juega un papel muy importante en aplicaciones de minería de datos, tales como exploración de datos científicos, recuperación de la información y minería de texto, aplicaciones sobre bases de datos espaciales (tales como GIS o datos procedentes de astronomía), aplicaciones Web, marketing, diagnóstico médico, análisis de ADN en biología computacional y muchas otras.

De forma general, las técnicas de Clustering son las que utilizando algoritmos matemáticos se encargan de agrupar objetos. Usando la información que brindan las variables que pertenecen a cada objeto se mide la similitud entre los mismos, y una vez hecho esto se colocan en clases que son muy similares internamente (entre los miembros de la misma clase) y a la vez diferente entre los miembros de las diferentes clases

ANÁLISIS

En el siguiente espacio se definirán el proceso que se tuvo que llevar a cabo para poder obtener los resultados al momento de preparar los datos.

El siguiente documento encontrará las gráficas relevantes, descripciones e interpretación de los datos obtenidos. Se adjuntará un documento en Python (JupyterNotebook) donde podrá visualizar los algoritmos y procedimiento que se utilizó para llevar a cabo los análisis que nos llevaron a nuestros resultados.

Variables

Las variables que representan un valor significativo son todas aquellas que nos ayudarán a obtener los valores esperados en los gráficos. Pero también existen varios datos que nunca vamos a utilizar porque en este momento no estamos realizando un análisis que los involucre.

Estos variables son:

- Id: Id de la película
- cast: Elenco de la película
- homepage: La página de inicio de la película
- tagline: El eslogan de la película.
- keywords: Las palabras clave asociadas a la película.
- overview: Una breve trama de la película.

Por lo que estas columnas de datos no serán tomadas en cuenta al momento de realizar los análisis de Clustering con la información.

Número de Clústeres

La técnica de clustering de partición entorno a centroides (PAM) realiza una distribución de los elementos entre un número prefijado de clústeres o grupos. Esta técnica recibe como dato de entrada el número de clústers a formar además de los elementos a clasificar y la matriz de similitudes.

Explorar todas las posibles particiones es computacionalmente intratable. Por lo tanto, suelen seguirse algoritmos aproximados guiados por determinadas heurísticas. En lugar de construir un árbol el objetivo en PAM consiste en agrupar los elementos entorno a elementos centrales llamados centroides a cada clúster.

El número de Clústeres que se utilizó fueron **5 y 3**. La cantidad de clusters se utiliza deacorde a al análisis que estamos realizando en ese momento.

Algoritmos de agrupamiento a utilizar

- Algoritmo: **KMeans**

  Agrupa datos tratando de separar muestras en n grupos de igual varianza, minimizando un criterio conocido como *inercia* o suma de cuadrados dentro del grupo. Este algoritmo requiere que se especifique el número de clústeres. Se adapta bien a un gran número de muestras y se ha utilizado en una amplia gama de áreas de aplicación en muchos campos diferentes.

  El algoritmo k-means divide un conjunto de N muestras X dentro K racimos disjuntos C, cada uno descrito por la media μj de las muestras en el grupo. Los medios se denominan comúnmente el grupo "centroides"; tenga en cuenta que no son, en general, puntos de X, aunque viven en el mismo espacio. El algoritmo K-means tiene como objetivo elegir centroides que minimicen la **inercia**, o el **criterio de suma de cuadrados dentro del clúster**

INVESTIGACIÓN

Medidas de tendencia central:

Las medidas de tendencia central son medidas estadísticas que pretenden resumir en un solo valor a un conjunto de valores. Representan un centro en torno al cual se encuentra ubicado el conjunto de los datos. Las medidas de tendencia central más utilizadas son: media, mediana y moda. Las medidas de dispersión en cambio miden el grado de dispersión de los valores de la variable. Dicho en otros términos las medidas de dispersión pretenden evaluar en qué medida los datos difieren entre sí. De esta forma, ambos tipos de medidas usadas en conjunto permiten describir un conjunto de datos entregando información acerca de su posición y su dispersión.

- **Variables continuas:** Es aquel tipo de variable cuantitativa que puede expresar una cantidad infinita de valores sin importar que sea un valor intermedio. Es decir, aquellas variables cuyo valor puede encontrarse entre dos valores exactos, generalmente representados por números decimales.

  Esta variable estadística se contrapone a la variable discreta, que solo puede adquirir un valor de un conjunto de números

Tablas de frecuencia:

La tabla de frecuencias (o distribución de frecuencias) es una tabla que muestra la distribución de los datos mediante sus frecuencias. Se utiliza para variables cuantitativas o cualitativas ordinales.

La tabla de frecuencias es una herramienta que permite ordenar los datos de manera que se presentan numéricamente las características de la distribución de un conjunto de datos o muestra.

GRÁFICOS Y RESULTADOS

DESCRIPCIÓN DEL PREPROCESAMIENTO

Se trata de encontrar agrupamientos de tal forma que los objetos de un grupo sean similares entre sí y diferentes de los objetos de otros grupos.

Esta tarea se trata de agrupar un conjunto de objetos de tal manera que los miembros del mismo grupo (llamado clúster) sean más similares, en algún sentido u otro. Es la tarea principal de la minería de datos exploratoria y es una técnica común en el análisis de datos estadísticos.

En nuestro caso, utilizamos las variables:

- Ingresos
- Ganancia
- Duración de la película
- Presupuesto
- ID de las películas

COMPARACIÓN DE ALGORITMOS DE CLUSTERING

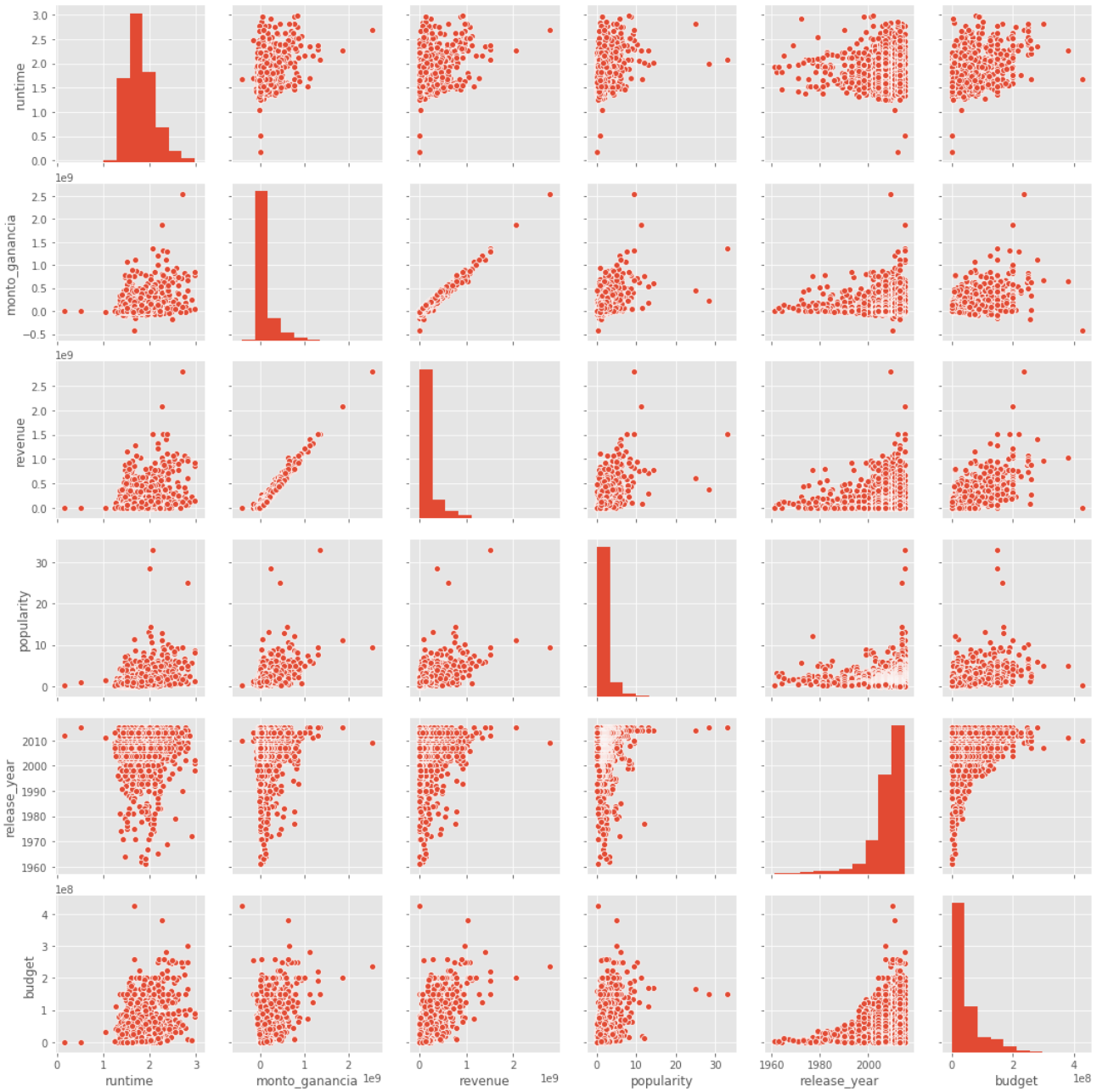
El algoritmo K-means tiende a tener problemas de performance cuando se analizan múltiples datos y grandes cantidades. Este funciona bien cuando los clústeres no superan un número de 5 clúster y la información no es compleja (datos bancarios, compras, ventas etc.).

DESCRIPCIÓN DEL TRABAJO FUTURO

Al momento de identificar los puntos atípicos, correlación de datos y clasificación de cada grupo de película por rango de votos podremos proponer un análisis de marketing y publicidad.

Por ejemplo, podríamos determinar que día del mes es correcto anunciar una película o que presupuesto debería de ser el "delimitante" para obtener un buen porcentaje de ganancias a pensar que no sea muy conocida.

ANÁLISIS CRUCE DE VARIABLES

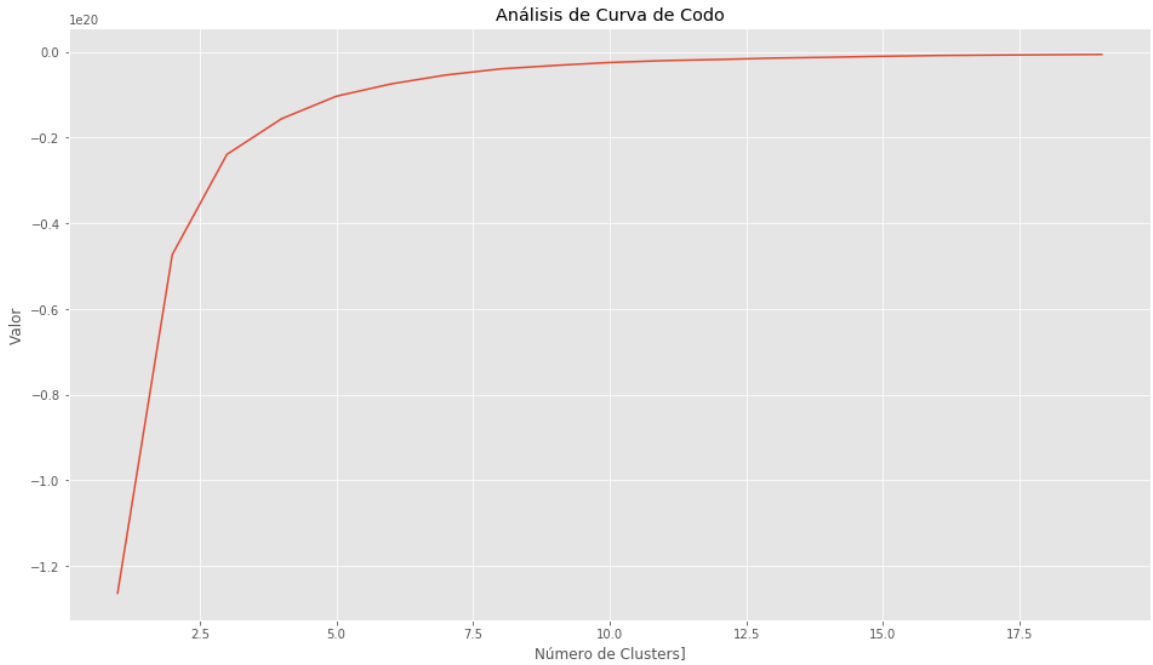


Descripción: Al momento de visualizar el análisis de cruce de variables nos aporta que las películas que estamos analizando no tienen un patrón que las defina. Porque puede que hayan tenido pocas ganancias pero que hayan sido las más vistas o viceversa. También nos indica que hay puntos atípicos en nuestro sistema, por lo que puede indicar que la información que estamos analizando requiera más filtros. Esto ocasionaría una pérdida de datos y resultados ya que varias películas representan un punto atípico en la información.

El cruce de variables pretende identificar si existe relación entre dos o más de ellas, además, de posibilitar el análisis de estas variables en una sola tabla, en lugar de construir dos cuadros simples. Estas tablas aplican fundamentalmente para variables categóricas o incluso cuantitativa discreta si ésta no tiene muchas categorías de respuestas.

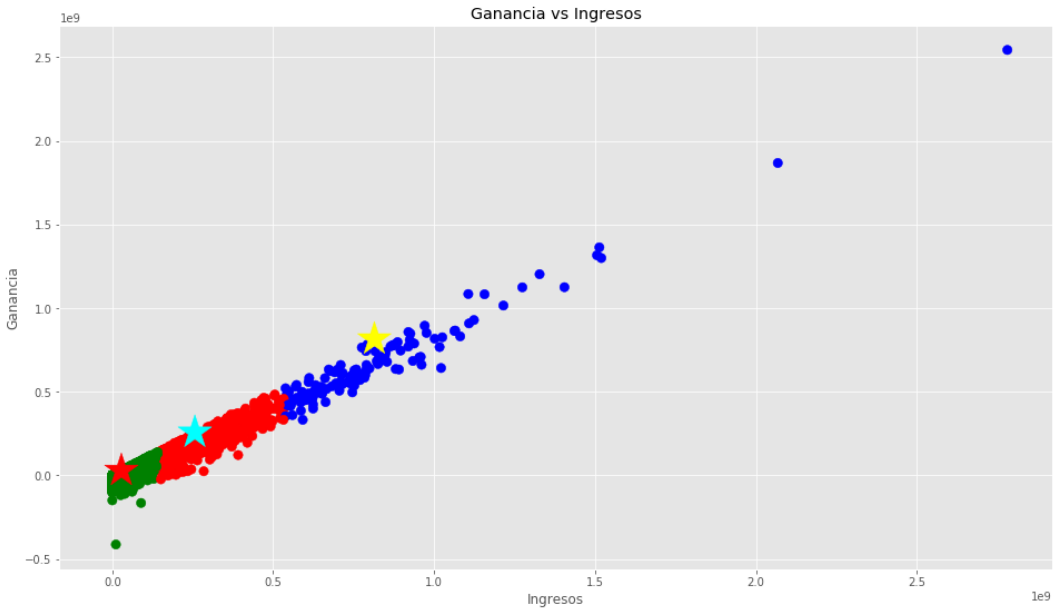
La selección de las variables en cuanto a que lugar va a ocupar en la tabla es poco relevante, es decir, no influye en los resultados obtenidos. Lo contrario sucede con la forma de calcular los porcentajes. Este último punto es fundamental y llevaría a conclusiones no sólo diferentes sino erróneas en algunos casos.

GRAFICA DE CODO



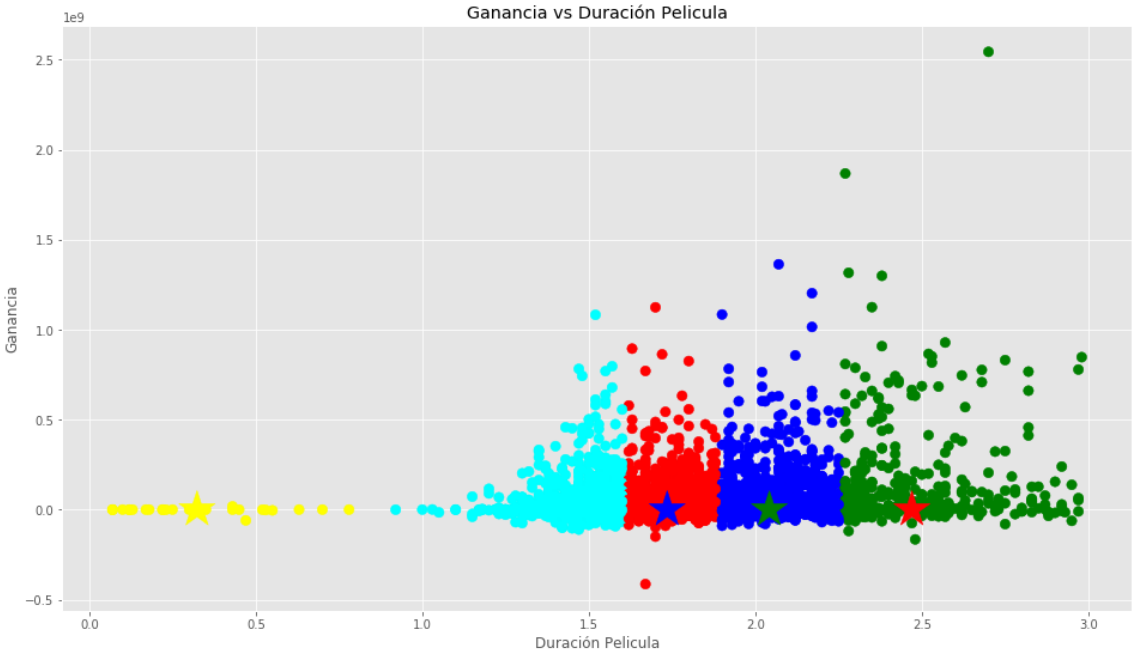
Descripción: Se tomó un clúster con un tamaño de tres, debido a que la gráfica de codo empieza a mostrar que dejan de haber variaciones representantes. Logrando agrupar en 3 grupos la información de la ganancia vs los ingresos por película.

INGRESOS VS GANANCIA



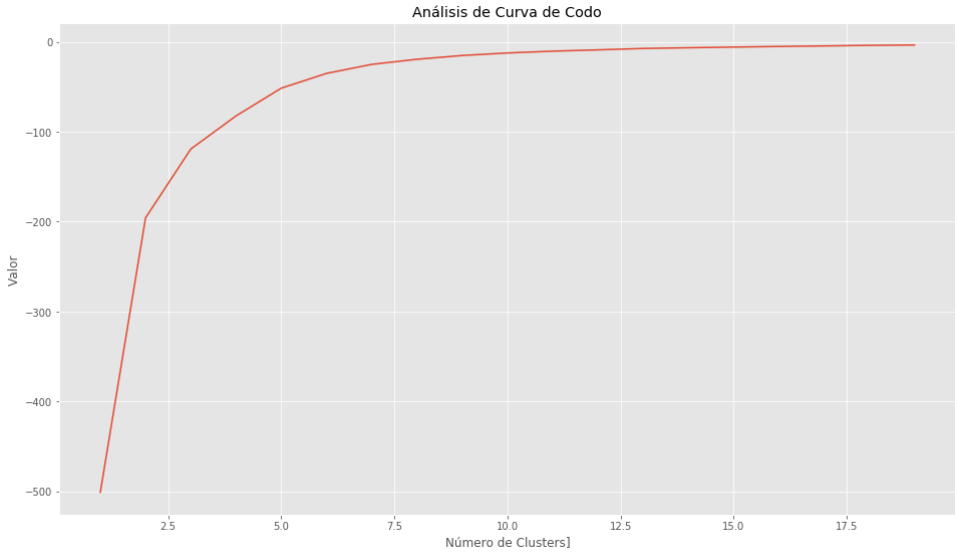
Descripción: En esta gráfica podemos observar que mientras menor sean los ingresos, menor será la ganancia. Lo que quiere decir es que tiene una clara dependencia entre uno de esos dos factores. Los ingresos se pueden dividir también en publicidad, mientras más publicidad tenga una película más esperada será; aunque no cumpla con la satisfacción del cliente final.

GANANCIA VS DURACIÓN PELICULA



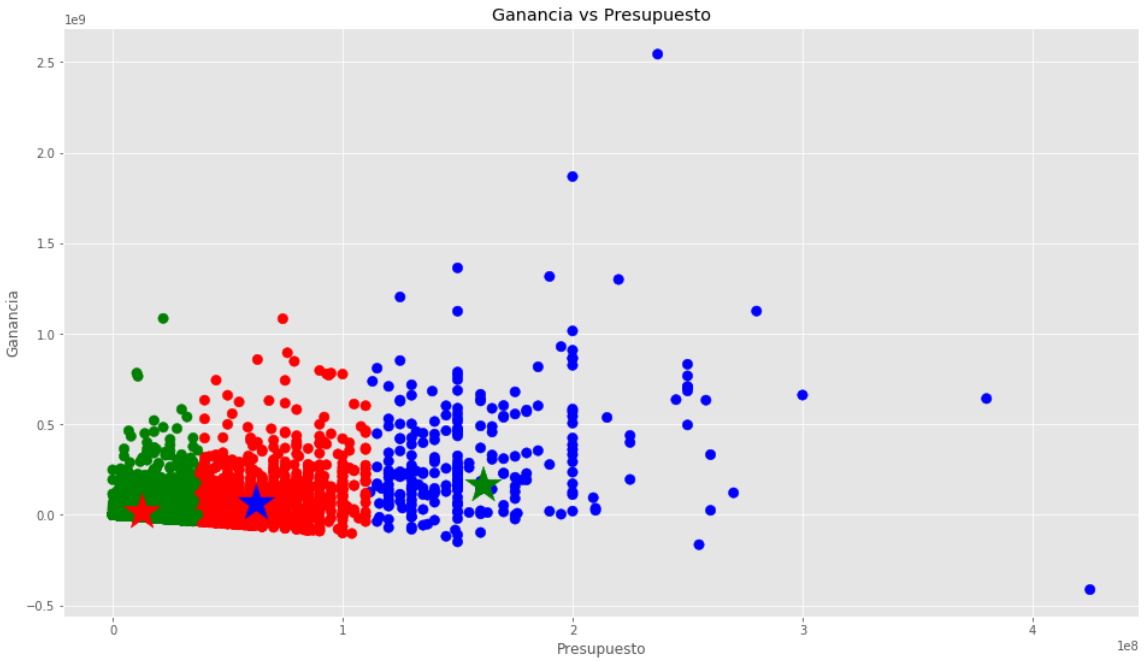
Descripción: La mayoría de las películas tiene una ganancia "estándar" No hay una correlación entre la duración de la película y la ganancia. El promedio de horas que duran las películas es de 1:40 min – 2:30 min mientras más dure una película menos atractiva será para el público general, por lo que en la sección de mayor a 3:00 horas la cantidad de ganancia es menor.

GRAFICA DE CODO



Descripción: Se tomó un clúster con un tamaño de cinco, debido a que la gráfica de codo empieza a mostrar que dejan de haber variaciones representantes. Logrando agrupar en 5 grupos la información de la ganancia vs los prepuesto por película

GANANCIA VS PRESUPUESTO



Descripción: Tener un gran presupuesto no significa tener una gran ganancia. Por ejemplo, hay una película que su presupuesto fue elevado, pero al momento de recuperar sus ganancias fue una completa pérdida. Esto significa que la trama, el desempeño y familiarización de los actores con el papel son los encargados de hacer que recupere su inversión.