Universidad del Valle de Guatemala
Facultad de Ingeniería
Departamento de Ciencias de la Computación
Mineria de datos
Ciclo 1 de 2,025

Proyecto 1

Gonzalo Enrique Santizo Vega – 21504

Link repositorio github:

https://github.com/GonzaloSantizo/Proyecto1DataMining

1. (3 puntos) Haga una exploración rápida de sus datos, para eso haga un resumen de su conjunto de datos.

Los datos con los que estaremos trabajando fueron extraídos de la página oficial de IMBD y describe la información sobre las películas que estuvieron en cine en los últimos años, posee una cantidad de 10,000 filas (películas) y 27 columnas (o variables). Las variables se reparten en:

- Id: Id de la película

- popularity: Índice de popularidad de la película calculado semanalmente

- budget: El presupuesto para la película.

- revenue: El ingreso de la película.

- original_title: El título original de la película, en su idioma original.

- originalLanguage: Idioma original en que se encuentra la película

- title: El título de la película traducido al inglés

- homePage: La página de inicio de la película

- video: Si tiene videos promocionales o no

- director: Director de la película

- runtime: La duración de la película.

- genres: El género de la película.

- genresAmount: Cantidad de géneros que representan la película

- productionCompany: Las compañías productoras de la película.

- productionCoAmount: Cantidad de compañías productoras que participaron en la película

- productionCompanyCountry: Países de las compañías productoras de la película

- productionCountry: Países en los que se llevó a cabo la producción de la película

- productionCountriesAmount: Cantidad de países en los que se rodó la película

- releaseDate: Fecha de lanzamiento de la película

- voteCount: El número de votos en la plataforma para la película.

- voteAvg: El promedio de los votos en la plataforma para la película

- actors: Actores que participan en la película (Elenco)

- actorsPopularity: Índice de popularidad del elenco de la película.

- actorsCharacter: Personaje que interpreta cada actor en la película

- actorsAmount: Cantidad de personas que actúan en la película

- castWomenAmount: Cantidad de actrices en el elenco de la película

- castMenAmount: Cantidad de actores en el elenco de la película

2. (5 puntos) Diga el tipo de cada una de las variables (cualitativa ordinal o nominal, cuantitativa continua, cuantitativa discreta)

- `id`: Cuantitativa Discreta (es un número entero, no puede tener valores intermedios)

- `budget`: Cuantitativa Continua (es una cantidad de dinero, puede tener valores intermedios)

- `genres`: Cualitativa Nominal (los géneros de las películas no tienen un orden específico)

- `homePage`: Cualitativa Nominal (es una URL, no tiene un orden específico)

- `productionCompany`: Cualitativa Nominal (son nombres, no tienen un orden específico)

- `productionCompanyCountry`: Cualitativa Nominal (son nombres de países, no tienen un orden específico)

- `productionCountry`: Cualitativa Nominal (son nombres de países, no tienen un orden específico)

- `revenue`: Cuantitativa Continua (es una cantidad de dinero, puede tener valores intermedios)

- `runtime`: Cuantitativa Discreta (es un número entero que representa la duración en minutos, no puede tener valores intermedios)

- `video`: Cualitativa Nominal (es un valor booleano, no tiene un orden específico)

- `director`: Cualitativa Nominal (es un nombre, no tiene un orden específico)

- `actors`: Cualitativa Nominal (son nombres, no tienen un orden específico)

- `actorsPopularity`: Cuantitativa Continua (es una calificación que puede tener valores intermedios)

- `actorsCharacter`: Cualitativa Nominal (son nombres de personajes, no tienen un orden específico)

- `originalTitle`: Cualitativa Nominal (es un nombre de película, no tiene un orden específico)

- `title`: Cualitativa Nominal (es un nombre de película, no tiene un orden específico)

- `originalLanguage`: Cualitativa Nominal (es un código de idioma, no tiene un orden específico)

- `popularity`: Cuantitativa Continua (es una calificación que puede tener valores intermedios)

- `releaseDate`: Cuantitativa Discreta (es una fecha, no puede tener valores intermedios)

- `voteAvg`: Cuantitativa Continua (es una calificación que puede tener valores intermedios)

- `voteCount`: Cuantitativa Discreta (es un número entero, no puede tener valores intermedios)

- `genresAmount`: Cuantitativa Discreta (es un número entero, no puede tener valores intermedios)

- `productionCoAmount`: Cuantitativa Discreta (es un número entero, no puede tener valores intermedios)

- `productionCountriesAmount`: Cuantitativa Discreta (es un número entero, no puede tener valores intermedios)

- `actorsAmount`: Cuantitativa Discreta (es un número entero, no puede tener valores intermedios)

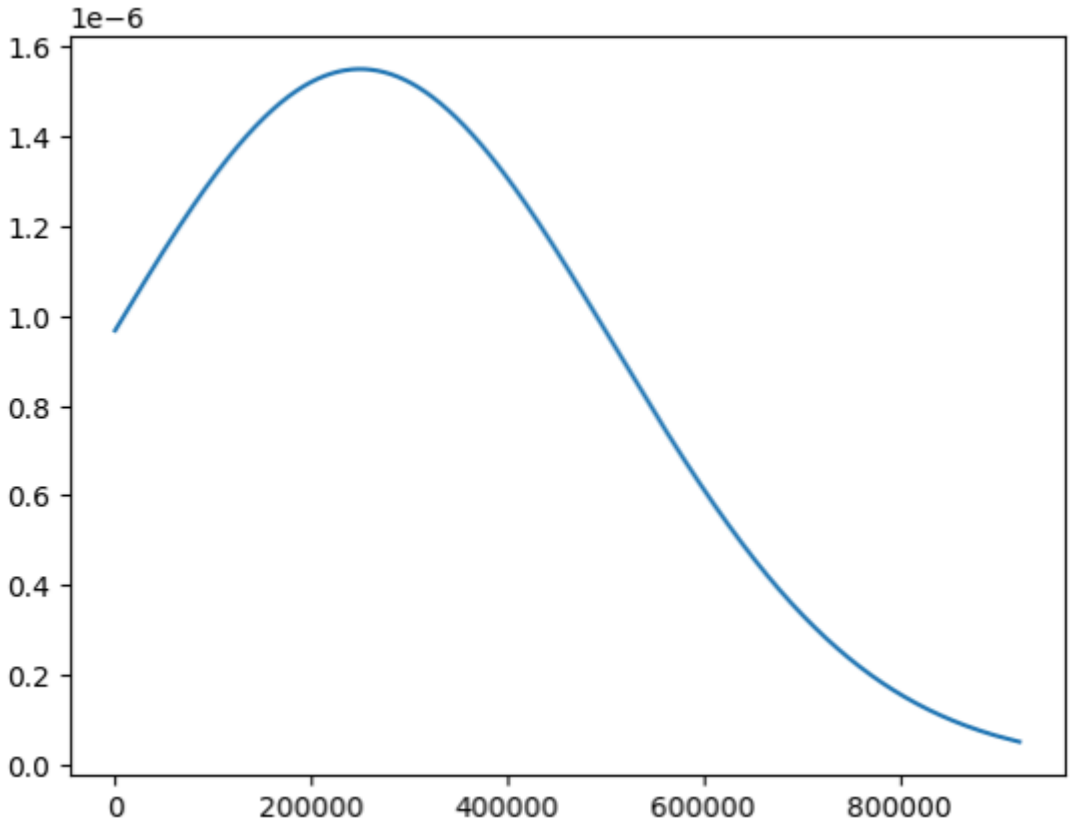
- `castWomenAmount`: Cuantitativa Discreta (es un número entero, no puede tener valores intermedios)

- `castMenAmount`: Cuantitativa Discreta (es un número entero, no puede tener valores intermedios)

3. (6 puntos) Investigue si las variables cuantitativas siguen una distribución normal y haga una tabla de frecuencias de las variables cualitativas. Explique todos los resultados

Variables Cuantitativa

- Id: Id de la película


1e−6
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
200000
400000
600000
800000