Melany Donis – 20170474

Data Wrangling

**Reporte Laboratorio 10**

El dataset contiene 18 variables luego de unir los datasets de movies y ratings. El dataset contiene el ID en la página de IMDB, el nombre en su idioma original de la película, el nombre con la que se conoce, la fecha en que se lanzó y el idioma en el que se hizo, las cuales son caracteres. Además, incluye el ID de la película, el año de lanzamiento, tiempo que dura, su popularidad en la página, su presupuesto y ganancias, el promedio y el número de la votación, el ID del usuario y rating que le dieron con el timestamp, éstas son representadas en números. De estas variables, las observaciones sin datos en la columna Budget y Revenue, se imputará un 0 su lugar, pues la mayoría de los datos no contienen un número importante, por lo que al final no vale la pena perder esas observaciones por estas columnas. Por esta misma razón se eliminarán las columnas de imbd\_id, runtime, movieId y release date, pues son columnas cuya falta de datos no es importante para el análisis y no se pueden imputar. El año de lanzamiento que falté se imputará con el año registrado en la fecha de lanzamiento y el titulo que falte con el titulo original. Por último, se imputará la moda de los datos para la popularidad, el promedio de votos y el número de datos, pues los datos son muy pocos como para que afecte significativamente el análisis. Los histogramas demuestran que las variables numéricas tienen una distribución a la derecha y solamente el promedio de los ratings y el promedio del voto parecieran tener una distribución normal.

El presupuesto esta correlacionado con las ganancias y el número de votos, el id esta correlacionado con el id del usuario, la popularidad con el número de votos y un poco con las ganancias, y el número de votos esta correlacionado con el presupuesto, las ganancias y la popularidad de la película.

Aún cuando muchos de los datos de presupuesto y ganancia tengan 0, se puede concluir que las variables están relacionadas entre si y que el número de votos depende de ella, pues, una película con muchas ganancias significa que muchas personas lo vieron porque es buena y por lo mismo hay más incentivos de votar. Sin embargo, el promedio del rating no parece estar relacionado con la popularidad de la película. Además, también se puede notar que mientras más nuevas son las películas mejor votación tienen, esto se puede deber a que los efectos son mejores y las personas las prefieren más por eso. Las gráficas de correlación y scatter son muy útiles para conocer la relación que tienen entre ellas las variables y ver cuales dependen más de otras. Las gráficas de barras e histogramas son útiles para conocer detalles de las relaciones de las categorías que comparten las variables y qué tan frecuentes son cada categoría de una variable.