ETL Project

Data Set: Kaggle Marvel VS DC IMDB Dataset

<https://www.kaggle.com/hetulmehta/marvel-vs-dc-imdb-dataset/download>

The original data is in csv format, a screen shot for the data:

Text

Description automatically generated

First drop the columns have null value for Runtim, IMDB\_Score and USA\_Gross

After drop those null values, the records is reduced from 1690 to 64

The run time column has value min at the right side, we remove that value and change it to numeric value

USA\_Gross has M at the right side and $ at left side, remove those and convert it to numeric

A screenshot for the data we have so far

A screenshot of a computer

Description automatically generated with medium confidence

Base on those info, we generate a scatted plot for IMDB\_Score vs USA\_Gross, along with the movie name beside each datapoint as showed below. It’s kind related and has a linear relationship between the two. Also use red color to mark Marvel movie and blue color to mark DC movie, interestingly, Marvel movies seem have higher revenue compare to DC movies, Marvel movies is in the top right section (higher rating, higher gross) where as DC movies are in the bottom half (all ratings, lower gross)

Graphical user interface

Description automatically generated with medium confidence

Plot both IMDB\_Score vs USA\_Gross and Metascore, they have the similar linear relationship.

Chart, scatter chart

Description automatically generated

Generate histogram for IMDB\_Score, from the diagram majority of movie among those 64 are in score 7, which is also the mean value. Drop column “Unname: 0”

Chart, histogram

Description automatically generated

Table

Description automatically generated  
Clean the year column by remove the right side ) , right strip any spaced and remove –

After clean year, convert it to numeric value and ignore any string not able to convert

Generate histogram for year below. The one in 1980 is single digit, maybe the superman movie, as time goes by, more and more movies released, maybe a lot of series, as showed in the diagram, there are about 600 movies in 2020.

Chart, histogram

Description automatically generated