# Project Summary

### Project introduction:

Project involves sourcing, cleaning, and analyzing movies data and discover trends.

### Chosen datasets:

1. Netflix movies and shows dataset from Kaggle. This tabular dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc. This is CSV file downloaded from Kaggle.
2. IMDB's top 250 movie data using web scraping. This dataset is published on IMDB website that includes top 250 movies of all time according to them.
3. Movie details using OMDB API. Getting movie details, actors, awards, DVD release date, Rating Scores from Rotten Tomatoes and IMDb, box office revenue etc. API returns 25 attributes about a movie.

### Ethical implications:

Overall –

Bias - Ethical implications may arise due to incomplete, and bad quality data as it may create biases. Cleaning the data sets is very important step before starting analysis.

Safety & security – Movies in general have a lot of implications on human mind, his/her development of thoughts, influencing changes in behavior etc.

Privacy – Not a major concern with these datasets as most of these ratings, votes, rankings are done with anonymous feedback of individuals who are seeing movies through online applications etc.

Transparency – For me goal here is to strictly use the data for analysis and discover some trends and patterns. There is no intention to use the observation to harm any one person or a group.

Accountability – Movie makers should be accountable to make good movies. Data collectors should be accountable for accuracy and quality of the data. Recommendation systems should be responsible for accuracy of the recommendations without any biases.

Let’s look at ethical implications with individual datasets.

1. Netflix movies and shows dataset – Ratings on Netflix are driven by number of views, client comments, box office collection, regional watching etc. which also means that it can be biased. We should take ratings with grain of salt. Netflix is responsible for presenting content without influencing viewers to movies i.e., Netflix sponsored movies etc.
2. IMDB top movie ratings (web scraping) dataset - Rank and ratings provided should be revisited from time to time and maintained as new movies release. Clearly define the scope of ranking and ratings in terms of region i.e., only applicable to US vs specific world region or all over. Make sure data displayed on the website is accurate including movie names, year of release, actors, director, and producers etc.
3. Rotten Tomatoes and IMDB (using OMDB API) dataset - Like previous data set, information provided should be rechecked from time to time specially ratings, rankings, score etc. These may change through time. Other information shown or collected should be accurate and should be validated. Ratings, rankings should also not influence with box office collection dollars but instead should rely on customer feedback and content quality.

### Steps performed for each data set:

1. Sourcing data – sourcing data by downloading, web scraping, and by calling APIs.
2. Reading data – Reading data into data frames for easy data exploration and cleansing.
3. Data cleansing – formatting, deriving new fields, imputing missing data, getting rid of outliers, evaluating quality etc.
4. Data filtering – Filtering out not needed data, for e.g., in Netflix movies data set we also have TV shows data. As other data sets do not have TV shows data, we cannot expect to get a join and thus filtered out TV shows data while only keeping movies data.
5. Data integration – merging, joining datasets together. In the project Netflix data is my driving data set. IMDB and OMDB data sets are left joined to Netflix data.
6. Data profiling – I used pandas profiling to create complete data quality report containing profile, correlation, data quality stats etc.
7. Data analysis & visualizations – Plotted various charts to see trends, analyze relationships, and observe patterns in the data.

### Data analysis observations (strictly based on datasets at hand):

1. Peter Jackson, Steven Spielberg, and Kevin Costner movies won most Oscars.
2. Movies with PG-13, R, and PG ratings won most Oscars.
3. There is no absolute relationship between number of Oscars won and box office collection.
4. No to low number of Oscar awarded movies have most IMDB votes.
5. Box office collection and IMDB votes seems to be gradual positive correlation. Most with higher box office collections seems to have more votes.
6. There seems to gradual down trend of making shorter movies.
7. There appears to gradual increase in box office collection by movies from 1970 to 2020.

By –

Name – Manish Shekhar

Course – DSC540 Data Preparation

Term- Spring 2023