**NAME: VANSHIKA TANEJA**

**SEMESTER & SECTION: 5I**

**SRN: PES1UG19CS557**

**PROJECT PART 1: Dataset Selection and analysis**

**TOPIC: Netflix: quantity-over-quality**

**Context:**

Netflix has a quantity-over-quality problem. This is part of an effort to help solve this. I want to figure out a way to find hidden gems in their catalogue but found it exceedingly hard to get the latest dataset that has ratings and many other attributes to help make sense of it. To help me and others dig deep into the latest Netflix content, I picked it up from Kaggle.

**Content:**

This dataset combines data sources from Netflix, Rotten Tomatoes, IMBD, posters, box office information, trailers on YouTube, and more using a variety of APIs. Note that there is no official Netflix API. A unique metric called "Hidden Gem Score", which shows a low review count and high rating. The lower the review count and higher the user rating, the higher the hidden gem score.

**Freshness of the dataset:**

The dataset was last updated on April 2021.

**Inspiration:**

* Find correlations between ratings, actors, directors, box office, and more.
* Observe designs around the nature of films and their different attributes like language, class, entertainers, and so on
* Discover hidden gems (underrated content) in different regions and languages.

**Dataset and Its attributes:**

* The dataset has over 29 columns with a huge variety of data types consisting of string values, decimal values, and URLs.
* The dataset has over 15,000 unique titles of the movies/series that cover over 100 genres.
* This dataset covers a huge ground of rating; the dataset collects the rating from various websites like IMDB scores, Metacritic score, and rotten tomato score.
* One of the most interesting, columns is the hidden gem score which shows a low review count and high rating. The lower the review count and higher the user rating, the higher the "hidden gem score." which will help us to find all the underrated content on Netflix.
* The dataset also tells us about the director and writer of the movies/series.
* The dataset has the links for the titles and also the poster.
* The dataset can also be described by the "tags" given by the people with around 13,000 unique values.

**Conclusion:**

Everyone these days want to watch something or other on Netflix but after finishing all the movies/series with overrating. We are left to watch with nothing or when we want to watch something which is not of our usual taste, this dataset can help us find something underrated with high ratings to watch. It was part of that attempt, with the ultimate goal to easily highlight underrated content & hidden gems on Netflix.