# DSC530- ASSIGNMENT 12.2   
# Author: Anjale, Jiteshwar  
# Date: 2021-11-20  
# Final Project – EDA on NETFLIX dataset

**Introduction**

Netflix is a subscription-based streaming service that allows our members to watch TV shows and movies without commercials on an internet-connected device. You can also download TV shows and movies to your iOS, Android, or Windows 10 device and watch without an internet connection. This is an EDA or a story telling through its data along with a wide range of different graphs and visuals.

This dataset consists of tv shows and movies available on Netflix as of 2018. The dataset is collected from Flixable which is a third-party Netflix search engine.

**Outcome of your EDA**

* Netflix is very popular among various social classes, and as we all know it has increased in popularity during the Covid-19 pandemic.
* Netflix is continuedly adding new content on their platform. We can see exponential growth in new content after 2016.
* Most of the content added on Netflix in 2018 and 2019 are released in same year.
* There are many different genres available on Netflix to attract wide range of subscribers.
* ‘International movies’ is the top genre present on Netflix.
* Majority of Netflix content is available in USA.

**What do you feel was missed during the analysis?**

I really wish that I would have looked for more numerical information to include my analysis. There are less variables in my dataset that are numerical which restricts me to apply the PMF, CDF on those variables. The format of rating should be numeric that can be useful for analysis.

**Were there any variables you felt could have helped in the analysis?**

I feel that the dataset I used should have some additional fields like number of views on movies or TV shows, budgets of the content, is the content produced by Netflix or leased, region where a content watch most.

I also feel that the dataset should have more subscription fee details based on country so we can predict the profitability.

**Were there any assumptions made you felt were incorrect?**

I assumed that there is some correlation between the month a new content is added on Netflix to year a new content added but that correlation comes out negative.

**What challenges did you face, what did you not fully understand?**

The dataset contains missing values for some of important fields like release year, added year, country. Also, it was bit challenging to clean up the data.   
The format of variable duration is in minutes for movies but in seasons for TV shows. I felt that the format should be uniform. I personally feel that the dataset should have more variables like number of views on movies or TV shows, rating in numeric format which can be used for EDA.