**Project: Investigate a Dataset: TMDb Dataset**

#Project:

#TMDb movies Dataset

# Research Questions:

#created 5 research questions associated with the director as a target to be answered:

#1- Which most popular genres movies? Drama &Comedy.

#2- Which director has most popularity? Colin

#3- Which director is having the highest budget? Sngmoo Lee

#4- Which director has most vote count? Pop &Pet

#5- Which director has the most vote average?Mark Cousins

#Investigation and Wrangling Documentation:

#• Data wrangling and cleaning Data

#1. Importing the packages and uploading the data file into Jupiter. (tmdb-movies(8).csv).

#2. Importing statements for packages to be used.

#3. Counting the number of unique values in our data for each data type

#4. One duplicate record is present and deleted.

#5. Dropping 10 missing values

#conclusion:

#Summary statistics and plots communicating the final results.

#1-from below analysis i conclude the 5-variable showing wide range between year Releases, popularity, budget, vote count and vote average.

#2-Isee correlation between the popularity and the revenue for the directors especially for lower value.

#3-Mark Cousins got the most vote average and not leaner with popularity, release, budget and vote count.

#4-the vote count goes to Bob Peterson|Pete Docter and again not leaner with popularity, releases and budget.

#5-Sngmoo Lee showing the highest budget was refer to but not leaner with popularity and releases.

#6-Colin Trevorrow is most popular, but he didn't have the most realeases.so the popularity and bigger a number of release didn’t go in linear direction

#7-from this data found most release is belong to woody Allen but that amount of releases is not linear with popularity for Colin Trevorrow .

#limitations.

# Not able to show all the directors names on the x-axis label due to the high count of director names.

# There are no missing values for any of the factors under consideration, but the data did not consider all elements and aspect for example the data set is not showing every single release for every director

# I have applied the dollar currency because the data is not showing which currency applied to the revenue or the budget for the movies that created in different regions.

# Usually for the movie revenue there are more sources to get more revenue like ticket sales in theaters, distributors, selling pay-TV, overseas rights and creating video games and so on.

# Some famous directors are having more releases within one year and over the years than new directors which will result of suddenly change direction or position.

#Resources:

#1- https://pandas.pydata.org/pandas-docs/stable/10min.html

#2- https://www.datacamp.com/community/blog/python-numpy-cheat-sheet

#3- http://www.developintelligence.com/blog/2017/08/data-cleaning-pandas-python/

#4- https://s3.amazonaws.com/assets.datacamp.com/production/course\_3485/slides/ch1\_slides.p

#5- https://github.com/nirupamaprv/Investigate\_Dataset/blob/master/Investigate\_a\_Dataset\_TMDb\_Directors\_NirupamaPV.ipynb

#6- https://docs.google.com/document/d/e/2PACX-1vTlVmknRRnfy\_4eTrjw5hYGaiQim5ctr9naaRd4V9du2B5bxpd8FEH3KtDgp8qVekw7Cj1GLk1IXdZi/pub?embedded=True