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

* **Introduction:**

The data set for this project is TMDb dataset movies.

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

1- Which director has most releases.?

2- Which director has most popularity?

3- Which director is having the highest budget?

4- Which director has most vote count?

5- Which director has the most vote average?

* **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>

* **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. commands explanation below.
5. One duplicate record is present and deleted.
6. Dropping 10 missing values

**Conclusions**

**Summary statistics and plots communicating the final results:**

* From below analysis i conclude the 5-variable showing wide range between year Releases, popularity, budget, vote count and vote average.
* I see correlation between the popularity and the revenue for the directors especially for lower value.
* Mark Cousins got the most vote average and not leaner with popularity, release, budget and vote count
* The vote count goes to Bob Peterson|Pete Docter and again not leaner with popularity, releases and budget.
* Sngmoo Lee showing the highest budget was refer to but not leaner with popularity and releases.
* 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
* From this data found most release is belong to woody Allen but that amount of releases is not linear with popularity for Colin Trevorrow .

#Challenges:

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