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

**Introduction:**

To complete my Data Analysis project, I am using TMDb movies dataset.

This data set contains information about 10 thousand movies collected from The Movie Database (TMDb), including user ratings and revenue. It consists of 21 columns such as imdb\_id, revenue, budget, vote\_count etc.

**Project**:

TMDb movies Dataset

**Questions that can be analyzed from this data set**

1. Which movie(s) have the most popular genres? Drama and Comedy

2. Who is the most popular director? Colin Trevorrow

3. Which director(s) have the highest budget? Sngmoo Lee

4. Which director(s) have the most vote count? Bob Peterson and Pete Docter

5. Which director(s) have the most vote average? Mark Cousins

**Investigation and Wrangling Documentation**

1. Data wrangling and cleaning Data

2. Importing the packages and uploading the data file into Jupiter. (TMDb-movies(8).csv).

3. Importing statements for packages to be used.

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

5. One duplicate record is present and deleted.

6. Dropping 10 missing values

**Conclusion**

This was a very interesting data analysis. We came out with some very interesting facts about movies. After this analysis we can conclude following:

1. The below analysis concludes the 5-variables showing wide range between year Releases, Popularity, Budget, Vote Count and Vote Average.
2. There’s a correlation between the popularity and the revenue for the directors especially at the lower value.
3. Mark Cousins got the most vote average; not linear with popularity, release, budget and vote count.
4. The vote count goes to Bob Peterson and Pete Docter and again not linear with popularity, releases and budget.
5. Sngmoo Lee showing the highest budget was refer to but not linear with popularity and releases.
6. Colin Trevorrow is most popular, but he didn't have the most releases, so the popularity and bigger number of releases 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**

1. Due to the large number of directors unable to display on the x-axis label.
2. The data did not consider all elements and aspects for example it’s not showing every release for every director.
3. The data is not showing which currency applied to which revenue or budget for the movies that were in different regions.
4. Movie revenue generate income from several different revenue streams, including theater, home video, television, streaming, etc.
5. Some famous directors release more movies in a year and over the years than any new directors which will change the 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