1. I chose **tmdb** movie dataset for my analysis.

This dataset consists of movie id, popularity of the movie, genres, budget and movie revenue etc.

I'm considering mean of Budget, revenue, popularity, runtime, vote\_average and votecount over the years for my first data analysis question.

For second one, I'm trying to find the number of movies made in different genres every year. Also, finding the highest number of movies made in each genre category. Which genre is the top category and liked more by audience?

I quickly figured out not a number or empty values in the whole excel datset by using **COUNTBLANK** function in one column and used the same formula to know about every columns. Here is the output from quick excel formula:

id 0

imdb\_id 10

popularity 0

budget 0

revenue 0

original\_title 0

cast 76

homepage 7930

director 44

tagline 2824

keywords 1493

overview 4

runtime 0

genres 23

production\_companies 1030

release\_date 0

vote\_count 0

vote\_average 0

release\_year 0

budget\_adj 0

revenue\_adj 0

So far, imdb\_id, cast, homepage, director, tagline, keywords, overview, genres, production\_companies have empty values. I used python code to drop out NaN from imdb\_id only as I feel that's the important one where all data is connected. This was my data cleaning step.

Questions posed here are like:

* To observe the mean of popularity, vote average, runtime, vote count, Budget and revenue for whole dataset.
* Classify Movies based on different genres for. Find out the highest number of movies made in each genre every year.

1. I used jupyter notebook to write the code and submitted .ipynb file.
2. I used stackoverflow, <https://plot.ly/matplotlib/> for most of the time to see syntax to write functions, plot different graphs (especially features to adjust plot space) in this data analysis.