**Introduction**

I will use tmdb data set from [Kaggle](http://localhost:8888/files/Nano%20degree%20solutions/www.kaggle.com?_xsrf=2%7C1bca370e%7C34a716047c37efe1d5122f4ce49e41eb%7C1587304211) to the second project in the Nano degree program. The aim of this study:

* Compare independent variables that are potentially affecting the revenue.
* Identify the most common genre in the movies using a word cloud

**During this**

* I will clean the dataset and select the independent variables that will use for the analysis.
* Extracting initial insights using descriptive statistics.
* Visualize my outcomes to deliver the conclusions

## Data Cleaning:

Select the dependent and independent variables for this analysis: The aim of this analysis is to test three independent variables 'budget', 'popularity', 'genres' to determine which of them is most effective on the revenue of the movie.

## Exploratory Data Analysis

### Research Question 1

#### Creating pairplot to take an overview and detect the potential relation between variables visually

#### Using corr() method to detect it numerically

**Initial notes:**

* The Higher budges makes more profits for the movies.
* The revenue is not affected with runtime.
* The popularity has positive moderate correlation with the revenue.

### Research Question 2

### What is the most common genre?[¶](http://localhost:8888/lab#What-is-the-most-common-genre?)

To answer this question I will use word cloud to detect the most repeated word in genre column

# **Conclusions**

## For the 1st question:

* The most correlated variables that increase the profit is the budget.
* Higher budget does not always mean higher profit

## For The second question:

### The most popular genres are:

* Science Fiction.
* Comedy Drama
* Drama Thriller
* Drama Romance