# Problem definition:

Nowadays, the movie industry is one of the largest entertainment industries. It has been estimated that there are 500000 movies produced up to this point. Lately, Saudi Arabia is moving toward investing on entertainment. Therefore, giving the possibility that soon there will be a Saudi Production Film. In order to move on the right path of creating a blockbuster movie we will analyze what features contribute to a movie being successful. Beside popularity and revenue, the Oscar is the main measure of a movie’s success. Therefore, it is the highest honor in film making. We will collect movies data to create a dataset that will help us achieve our goal.

# Data collection:

We collected our data using three sources:

* Scrapping IMDB website (For Movie titles within a selected year).
* Scrapping IMDB website (For the actor awards)
* Scrapping IMDB website (For Budget information)
* IMDB API
* IMDBPY library
* OMDB API
* Obtaining a list of previous Oscar winners

Using the previous tools, we created a dataset of more than 10,000 movies produced between 2000 and 2017. For each year we took the top 250 movies in terms of popularity, then we collected randomly 4% of the remaining movies to assure diversity in all respects.

# Data cleaning:

We used the multiple sources of data to try and reduce the number of null values in our dataset.

* **Margining the datasets**
* **Remove unwanted columns:**

*The data produced over 40 columns which required us to drop unnecessary features that were not contributing to the prediction*

* + Plot
  + Type
  + Released
  + Location
  + Etc.

**Fix some values format:**

* + Year (Making it into a Date-Time Format)
  + Genre (Splitting the Genres into multiple columns to reduce redundancy)
  + Finances -Budget, Gross, Cumulative, Opening- (Removing symbols, calculating currency
  + Rating (Mapping the rating to the required rating system)
  + Cast (removing the html tags and fixing the format)
  + Language (Making a Boolean column to represent English and a column to represent other languages
  + Country (Making a Boolean column to represent USA and a column to represent other countries)
* **Missing values:**
  + Using different data sources to fill missing values.
  + Fill the missing values in different columns as follows:
    - Production company: Add missing value as other
    - Cast, director, writer .. etc: Since we will replace the names with weight value, we give weight 0 to unknown people.
    - Budget : Predict using regression
    - OWUS : Predict using regression
    - CWG : Predict using regression
    - Runtime: average of the same genre Done
    - Certificates : mode of the same genre Done
    - budget : regression
    - opening : regression
    - cumulative: regression
    - Gross : regression
    - Rating : 0 (it means no one rate this movie yet!) Done
    - Votes : 0 Done
    - Location : drop column (I don’t think that it is useful )
    - english : majority
    - multi\_lang : majority
    - us : majority
    - multi\_country : majority
    - month : choose randomly between top 3 months.
    - year : fill with mode

# Feature engineering:

* + Language (Making a Boolean column to represent English and a column to represent other languages
  + Country (Making a Boolean column to represent USA and a column to represent other countries)

**Deal with categorical data:**

* Genre: dummies
* Production company: take top 15 and add other and get dummies
* Cast, director, writer .. etc: generate weight value to reflect the persons’ power depending on the previous award nomination and wining.

**Add useful columns:**

* Separate date to year and month
* Replace the names of actor, editor, composer, writer, producer, director, assistant director, and cinematographer with representative weight that calculated as follows:

# Data analysis:

* Analyze the change in the main features over the years.
* Study the correlation between different features:
  + Revenue & rating
  + Budget & revenue
  + Runtime & revenue
  + Runtime & rating
  + Oscar & genre
  + Month and Oscar
  + Month and revenue
  + Certificate and revenue
  + Certificate and rating
  + Production company and Oscar
  + Number of votes and rating
* Explore what factors may affect movies success
* Find possible patterns
* Find best/ worst/ longest/cheapest/..etc

# Modeling:

**Predict the following:**

* Movie’s Oscar nomination
* Movie’s Oscar wining
* Movie’s users rating
* Movie’s revenue

# Evaluation:

Evaluate each model.

# Challenges:

* Data collection

*We faced many problems with data collection. After Analyzing many data sources we realized that Some data sources such as (TMDB) is not a reference because any user can append false information about a movie therefore we disregarded it. When working with other data sources we faced difficulties with unpopular movies which had many missing values.*

*There are some problems with the IMDBPY. This library took a lot of time to collect the required data. 1500 rows took around an hour. Also all of the APIs did not sort the movies from top to low therefore we scraped the needed movies.*

* **Missing values**

*All of the data sources had missing values when a movie is not top rated. The budget for example was always missing therefore we scraped the budget from the IMDB website.*

* **Budget different currency**

*The Budget came with a different currency then dollars which required us to change it back. Therefore, we used an API (Forex) to convert other currencies to dollars.*

* **Creating the cast award matrix**

*The Number of cast for all the movies is vary large. We had to grab all of the awards an actor has achieved in correspondence with the year an actor received the award.*

* **Genre**

*Each movie has multiple genres which we consider as valuable information. We did not choose one genre rather we took all genres into consideration.*

# Future work:

# Conclusion: