Assignment 1 - Exploratory Data Analysis

For part 1 of this assignment, we were told to form a group to identify and characterize a dataset. We acquired the dataset from Kaggle. Further details regarding this assignment will be discussed throughout this report.

Part A - Description of the Dataset

The dataset we have acquired for this phase of the assignment contains the data of over 5000 movies. The data is scraped from IMDB. The data spans across 100 years and 66 countries. The following are the variables and the explanations for each of them:

- "movie_title" - Title of the movie
- "color" - If the movie is in color/black & white
- "num_critic_for_reviews" - Number of reviews from critics
- "movie_facebook_likes" - Number of likes the movie's Facebook page has
- "duration" - Duration of the movie
- "director_name" - Name of the director
- "director_facebook_likes" - Number of likes the director's Facebook page has
- "actor_3_name" - Name of the actor
- "actor_3_facebook_likes" - Number of likes the actor's Facebook page has
- "actor_2_name" - Name of another actor
- "actor_2_facebook_likes" - Number of likes another actor's Facebook page has
- "actor_1_name" - Name of yet another actor
- "actor_1_facebook_likes" - Number of likes yet another actor's Facebook page has
- "gross" - Gross profit the movie has obtained
- "genres" - Genre of the movie
- "num_voted_users" - Number of users who have voted for the movie
- "cast_total_facebook_likes" - Number of likes the cast's Facebook page has
- "facenumber_in_poster" - Number of faces on the movie's poster
- "plot_keywords" - Keywords in the movie's plot
- "movie_imdb_link" - Link for the movie on IMDB
- "num_user_for_reviews" - Number of users who reviewed the movie
- "language" - Language of the movie
- "country" - Country the movie was made in
- "content_rating" - Content rating of the movie
- "budget" - Budget of the movie
- "title_year" - Year the movie was released
- "imdb_score" - Score the movie obtained on IMDB
- "aspect_ratio" - Movie's aspect ratio

The original dataset contains **28 variables and 5043 rows** . The .csv file is 580KB in size.

Part B - Possible Insights

From mining the chosen dataset, the following insights are possible to be obtained:

- Movies with the highest gross profit in the 20th century
- Movie genres with highest gross profit from 2000 to 2016
- Best-selling movies in the United States
- Highest grossing movies in the South East Asian region
- Predicting the success of a movie based on the genre

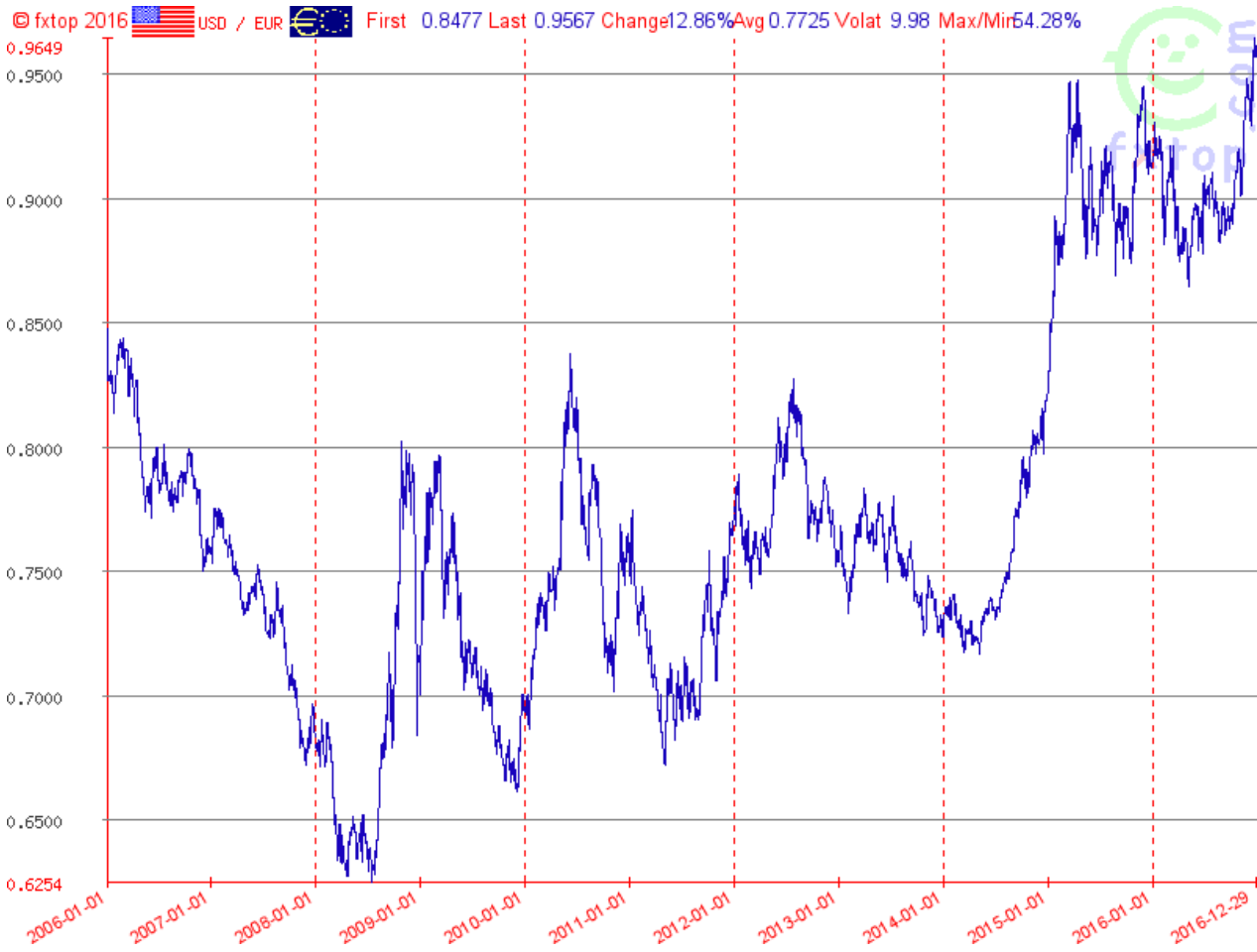
Part C - Data Mining Technique

The data mining technique that is suitable for this data set is classification. A new attribute called **"user_rating_prediction"** can be made to predict the user rating for the movie. The budget of the movie can be classified from **low, medium and high**. This can also applied for **actor_facebook_likes, actor2_facebook_likes, actor3_facebook_likes, director_facebook_likes and also the IMDB scores.** For example, IMDB scores can be classified as low (from 0-4 rating), medium (5-6 rating) and high (7-10 rating). If the

budget, actor_facebook_likes, director_facebook_likes and the imdb_score is high, that will predict that the user_rating_prediction is also high (ranging from 8-10).

Part D - Data Quality Issues

There are problems in the column **"Budget" and "Gross".** There should be some feature engineering performed on these attributes to make it in a systematic way. This is because the world currency is not constant ; it changes from time to time. Since the data recorded movies spanning across 100 years, "inflation" factor should be considered to normalized all currencies (US dollars, British pound, Euro, etc) into one basis. Also exchange rate between different currencies (example: USD to Euro) would be a hassle because the rates are varies over time. The graph below shows the rates between USD and Euro from the year 2006 to year 2016.



Part E - Pre-processing

Just like all datasets these days, the dataset we have obtained requires us to perform data pre-preocessing on it. We have perfomed a number of pre-processing tasks on the dataset in order to obtain the best data that we can and to make it as relevant as possible. The following are the preprocessing tasks done on the dataset:

1. Data Cleaning - Filling in missing values
    - color - if year >= 1939 set to color
    - director_name - remove rows with missing values
    - duration - set missing values to mean
    - actor_2_name - remove rows with missing values
    - gross - set missing values to mean
    - genres - no missing values
    - actor_1_name - remove rows with missing values
    - movie_title - no missing values
    - actor_3_name - remove rows with missing values
    - language - remove rows with missing values
    - country - remove rows with missing values
    - content-rating - set missing values to "PG"
    - budget - set missing values to mean
    - title_year - remove rows with missing values
    - imdb_score - no missing values
    - aspect_ratio - set missing values to mean

2. Data Reduction - Reducing the number of attributes
    - num_critic_for_review
    - director_facebook_likes
    - actor3_facebook_likes
    - actor_1_facebook_likes
    - num_voted_users
    - cast_total_facebook_likes
    - facenumber_in_poster
    - plot_keywords
    - movie_imdb_link
    - num_users_for_review
    - actor_2_facebook_likes
    - movie_facebook_likes

Once this process was completed, we were left with **4907 rows** and **16 columns** from the original **5043 rows and 28 variables.** All of the attributes that were removed were seen as unnecesary for the insights we were targetting to achieve from data mining. The amount of data that was lost from data cleaning and data reduction is not much and does not heavily effect the dataset or cause any form of bias. We believe that the amount of pre-processing that was done on the dataset is sufficient for further data mining process in the KDD pipeline.

CODES & Links

The R codes, original and pre-processed .csv files, and a copy of this report in PDF for this phase of the assignment are all available at our github repo located here or just click here to view the R code on a html page.