Alexandria University
Faculty of Computers and Data Science
Data Science Methodology
Hagar Hazem Mohammed: 20191482504
Hady Ihab Ahmed: 20191450774
Mohammed Amr Shehata: 20191565917

Wrangling Report

In the points below we will explain the issues we found after assessing the dataset and how we managed to solve them, we will start off with the **Quality** issues:

- We have 28 columns so we figured that it will be more convenient to remove any row with more than 10 missing values.
- Some of the columns we decided that the best way to fill the null values is with the column average because they has numeric values (num_critic_for_reviews, num_user_for_reviews and aspect_ratio).
- Others that have non-numeric values, we filled them with the most frequent value (language → English, color → Color and content_rating → PG-13).
- A lot of the columns' data type was float instead of integer so we changed that (director_facebook_likes, actor_1_facebook_likes, actor_2_facebook_likes, actor_3_facebook_likes, movie_facebook_likes, gross, budget, num_critic_for_reviews, duration, num_user_for_reviews, title_year, facenumber_in_poster).
- There was some invalid values so we replaced those with the most frequent value (invalid country names like new line replaced by USA).
- We removed some records that have null values in certain columns for classification purposes.
- All countries was written their full name except for USA and UK so we replaced the abbreviated names into their full names (USA → United States of America and UK → United Kingdom).

Now the **Tidiness** issues:

- Plot keywords was a multi-valued column we split it into two columns and we deleted the original column.
- The genre column was a multi-valued column as well so we split into columns and deleted the original one.
- Then we filled the null values in the four new columns with the most frequent value in each column.
- The two new genre columns had only 17 and 21 unique values so we changed their data type into category.