Name: Soham Patki
Class: MISM 6205 - Data Wrangling
Professor: Dr. Tareq Nasralah
Date: 12/2/22

# Written Report - Project 3

**What data was used to enrich the client's data?**

The client's data was enriched using the 500 movies released from 2018-2020 which had the greatest number of votes on IMDb.com. This information was scraped and transformed into a structured (.csv) format using the Beautiful Soup package in Python. Through web scraping, I collected information on 8 parameters: movie ID, title, rank, release year, movie runtime, rating, number of votes and genres of each movie.

This scraped data was then merged with the client's data. Since the ranking is sorted by number of votes on IMDb.com, it is subject to change over time as people continue to vote. As a result, there were some mismatches between the scraped dataset and the client's data. The final (merged) dataset thus contains 496 movies, with 4 movies removed due to mismatches. The merged dataset retains 3 descriptors (original title, is adult, title type) from the client's data. The enriched data has 11 columns.

**Describe the data cleaning and transformation that was implemented.**

The following issues were identified for cleaning and transformation:

1) The 'year' data scraped from IMDb.com existed in the format '(20XX)', and the first entry 'Joker' had an unnecessary 'I' next to its release year. This issue was identified during the scraping process and was corrected to be stored simply as '20XX' - before the 'year' data was added to the dictionary.

2) While exploring the dictionary containing scraped data, I observed that a line break existed before each entry for 'genres'. To avoid formatting issues, this was corrected by removing '\n' from the scraped data before 'genres' was added to the dictionary.

3) While exploring the scraped dataset using the '.describe()' function, I observed that the 'year' column had 4 unique values. Since the list only contained movies from 3 years (2018-2020), I investigated further and found that the 4th value was 'V2020'. This was corrected using the '.replace()' function in Python.

4) To conduct machine learning tasks on numerical data, it would be necessary to change the relevant fields from 'object' data type to integers. In the 'votes' column, I removed the commas present in the numerical observations to be able to change its data type. Similarly, I removed the 'min' characters (representing 'minutes') that existed in every observation in the 'runtime' column.

5) With these columns cleaned, I changed the data types of the ‘rank’, ‘year’, ‘runtime’ and ‘votes’ columns to integers.

6) The client dataset was then enriched by merging it with this scraped data. The columns of the merged dataset were rearranged by relevance and ordered by ‘rank’.

**Conducting data cleaning and transformation in Alteryx.**

The cleaning and transformation tasks were conducted in Alteryx by:

1) Using the ‘Data Cleansing’ function to remove unwanted characters (Leading and Trailing Whitespace, Letters, Punctuation) from the columns ‘year’, ‘runtime’ and ‘votes’.
2) Alteryx identified that the ‘genres’ column specifically had large amounts of whitespaces, which was also corrected using the ‘Data Cleansing’ function.
3) The ‘Join’ function was used to merge the scraped data with the client’s data. The ‘Join’ function had the additional feature of changing the data types of specific columns. The columns ‘rank’, ‘year’, ‘runtime’ and ‘votes’ were converted to integers using this function.
4) The ‘Sort’ function was used to sort the merged data by ‘rank’.
5) The ‘Select’ function was used to rearrange the columns into the desired order.

**Sources:**

Data was scraped from the following URL:
https://www.imdb.com/search/title/?at=0&sort=num_votes,desc&start=1&title_type=feature&year=2018,2020