ANLY501 Introduction to Data Analytics
Project Assignment 1

Zikai Zhu, Tianyi Yang, Ju Huang, Xinyue Liu

Data Science Problem:

We decided to analyze the movie data, including its box office, rating, awards and so on. Sometimes we see movies with low ratings but large box office. We also see movies with high ratings but low box office. If ratings do not always predict box office, we are curious to know what attributes can. "Stakeholders are looking for a 'magic formula' to better understand and predict box office success are turning to statisticians and data scientists to help with this challenge. To increase their profits, producers and directors need to understand what raises the curiosity of their target audience. This is where analytics can play an effective role."[1]

In some articles we found on how to 'predict' the box office of a certain movie, many of the analyses are rather empirical. For example, the Warner Brothers' chairman decided to not produce movies starring a female, simply because female-leading movies did not sell well in the past[2]. However, this decision is not based on a cause-effect analysis which might actually lead to the wrong action. Thus, in this project, we want to use a more statistically stringent method to explore the movie world.

Data Overview:

We scraped two sets of movie data, from year 2010 to year 2017.

1. Dataset from Boxoffice Mojo : contains some basic information about movies as well as their box office.
2. Dataset from OMDB: contains various attributes about a movie (eg.ratings from Rotten Tomato, actors of the movies, and so on)

[1] https://www.course5i.com/blogs/analytics-in-movies-data-analyze-act/

[2] https://entertainment.howstuffworks.com/predict-weekend-box-office-before-sunday-numbers.htm

Potential Analysis:

With the wealthy information on movies, we can explore many interesting question. Below is a list of potential questions that we would love to explore:

1. Across the past 7 years, what are some of the key trends in the movie industry?
2. Is the movie industry in general booming over the past 7 years?
3. Who are the most successful directors/ cast in terms of box office or ratings?
4. Over the past 7 years, how did the population receive each genre? Has it changed over time?
5. How do different attributes of movies (such as rotten tomato ratings, runtime, studio…) impact the movie's box office? Are there any key elements contributing to the success (in terms of box office) of a movie?

Data Issues:

For the data from Omdb API:

	Type	Description
1	Missing values	Missing values exist in most of the columns.
2	Data redundancy	There're columns that provide the same information: ● 'Internet Movie Database' : It is a redundant column. 'imdbRating' has the same value ● 'Metacritic': It is a redundant column. 'Metascore' provides the same information
3	Irrelevant Data	There're columns that do not add value to analysis: ● 'Type' column: there's only one value (“movie”) in this column. Thus it won't contribute to our analysis ● 'Poster' : this column contains a picture which we do not prepare to analyze ● 'DVD': this column contains the DVD release date, which we do not prepare to include into analysis ● 'Website' : this column provides the link to the movie's website, which we do not prepare to include into analysis
4	Data format	There're many columns that needs to be changed into a desirable format in order to conduct further analysis: ● BoxOffice: wrong data type with 'string', should delete '$' and ',' and convert it to float

		● Metascore: incorrect data type with 'string', should convert it to integer ● Rotten Tomatoes: incorrect data type with 'string', should delete '%', convert it to integer ● Runtime: wrong data type with 'string', should delete 'min' and convert it to integer ● imdbRating: wrong data type with 'string', should convert it to float ● imdbVotes: wrong data type with 'string', should delete ',' and convert it to int ● totalSeasons: incorrect data type with 'string', should convert it to integer ● Rated: movie that is not rated might have different value for 'Rated', including 'UNRATED' and 'NR' ● Response: has bad value with 'Movie not found!' and 'Year not match!'

For the data from Box Office:

	Type	**Description**
1	Missing values	Missing values exist in some of the columns.
2	Data format	There're many columns that needs to be changed into a desirable format in order to conduct further analysis: ● id,Year, Rank,Total Gross,All Theaters,Opening,Opening Theaters: wrong data type with 'string', should delete '$' and ',' and convert them to numbers ● Open&Close: do not have 'year' in them

Data Cleanliness:

The goal is to get the data cleanliness score for each of our two dataset. We check the missing value, the noise value, the usefulness, and the redundancy of our data.

We categorize our columns into 5 types:

	Category	Description
1	Numeric columns	Numeric data, or data that can be converted into numeric types
2	Categorical columns	Columns with limited unique values, such as movie classification results
3	String columns	Columns with individual string values (cannot be categorized)
4	Date columns	Date
5	Irrelevant columns	Columns that not related to our project
6	Redundant columns	Columns that contains redundant values with others

Boxoffice data contains 11 columns. The classification of these columns is as follows:

	Category	Column Name
1	Numeric columns	'id', 'Year', 'Rank', 'Total Gross', 'All Theaters', 'Opening', 'Opening Theaters'
2	String columns	'Name', 'Studio'
3	Date columns	'Open','Close'

Movie rating data contains 29 columns. The classification of these columns is as follows:

	Category	Column Name
1	Numeric columns	'BoxOffce', ' Metascore,' 'Rotten Tomatoes', 'Runtime', 'imdbRating', 'imdbVotes', 'total Seasons'
2	Categorical columns	'Rated','Response'
3	String columns	'Actors', 'Awards', 'Country', 'Director', 'Language', 'Plot', 'Title', 'Writer', 'imdbID', 'Genre', 'Production'

4	Date columns	'Released', 'Year'
5	Irrelevant columns	'DVD', 'Type', 'Website', 'Poster'
6	Redundant columns	'Internet Movie Database', 'Metacritic'

Generally, the max score for each column is 100:

Score = 100-deduction

First, for columns we do not prepare to use (i.e. irrelevant columns and redundant columns):

1. Irrelevant value deduction: For these columns, since we do not use them in the future, the deduction is 100.

   Score = 100-100=0

2. Redundant value: For it is really difficult to detect by script, we do not take them into consideration here, just deal with them in the following cleaning process.

Then, for the useful columns:

Score = 100 - sum(deduction)

...where the deduction is due to two types of issues:

1. Missing value deduction:

   missing_value_deduction=Percentage_of_missing_value*100*missing_value_weight

2. Noise value deduction:

   noise_value_deduction = Percentage_of_noise_values * 100 *noise_value_weight

   2.1 Numeric columns: For every column, try convert the data into a target data type. If the conversion fails, take it as a noise value.

   2.2 Categorical columns: For every column, we looked into them and decided how to count the number of noise.

   2.3 String columns: We do not compute its noise deduction.

   2.4 Date columns: For every column, try to convert the data into date type. If the conversion fails, take it as a noise value.

… where *missing_value_weight = noise _value_weight* = 50% if a column have both types of issues, we weigh them equally. Otherwise, if a column has only one type of issue, the weight automatically becomes 100%.

Finally, we compute the final scores for the two datasets, which are the average score of columns in each dataset:

Final_score = sum(Score) / numbers_of_attributes_in_each_dataset

For box office data, the score before cleaning is **88.780**.

For movie rating data, the score before cleaning is **68.621**.

Data Cleaning:

To clean the box office dataset, we created three functions for different types of columns. The first one is to fill the missing value with -1 to avoid errors in future operation. The second one is to clean the numeric data columns (as shown in the above table) by deleting non-numeric signs and converting them to numbers. The third one cleaned the date values by adding year into dates.

To clean the movie rating dataset, we created four functions for different types of columns. The first one is to fill the missing value with -1 to avoid errors in future operation. The second one is to delete all the irrelevant columns and redundant columns. The third one is to clean all the numeric columns by deleting non-numeric characters and converting them to numbers. The last function is to clean the categorical columns. For the categorical column ‘Rated’, duplicates exist in its values. Thus, we changed all 'UNRATED' and 'NR' to 'NOT RATED' (which all have the same meaning). Also, for the categorical column ‘Response’, we cleaned values 'Movie not found!' and 'Year not match!' by converting all of them to -1.

Both scores for data cleanliness increased a lot. For the box office data, the score before cleaning is **88.780** and the score after cleaning is **99.335**. For the rating data, the score before cleaning is **68.621** and the score after cleaning is **99.998**.

Reference:

https://www.course5i.com/blogs/analytics-in-movies-data-analyze-act/
https://entertainment.howstuffworks.com/predict-weekend-box-office-before-sunday-numbers.htm