ETL Project

Marianne Boerenko, **A**run Kara, **R**ebecca Gonzalez (M.A.R. Movies)

https://arunkara.github.io/ETL_Website.io/

Documentation

With the 2020 pandemic continuing, families are relying on in-home entertainment now more than ever. To assist families in identifying movies quickly based on certain criteria, we combined two Kaggle databases. The first database consisted of the movies available from the top four streaming networks; Netflix, Hulu, Disney + and Prime Video. We combined it with a second database from IMDb consisting of the movie sales and budget. Our newly created database will allow movie aficionados to compare the budgets and gross revenue of their favorite movies, genres, and directors. Their data exploration might lead to the realization that some streaming platforms prefer low budget indie films while others invest more in high-grossing, 'block-buster' budget films.

Clean Up

Initially, we downloaded the csv files to a local folder on our desktop and read them into pandas using Jupyter Notebook.

	ID	Title	Year	Age	IMDb	Rotten Tomatoes	Netflix	Hulu	Prime Video	Disney+	Type	Directors	Genres	Country	Language
0	1	Inception	2010	13+	8.8	87%	1	0	0	0	0	Christopher Nolan	Action,Adventure,Sci-Fi,Thriller	United States,United Kingdom	English,Japanese,Fre
1	2	The Matrix	1999	18+	8.7	87%	1	0	0	0	0	Lana Wachowski,Lilly Wachowski	Action,Sci-Fi	United States	English
2	3	Avengers: Infinity War	2018	13+	8.5	84%	1	0	0	0	0	Anthony Russo,Joe Russo	Action,Adventure,Sci-Fi	United States	English
3	4	Back to the Future	1985	7+	8.5	96%	1	0	0	0	0	Robert Zemeckis	Adventure,Comedy,Sci-Fi	United States	English
4	5	The Good, the Bad and the Ugly	1966	18+	8.8	97%	1	0	1	0	0	Sergio Leone	Western	Italy,Spain,West Germany	Italian

Figure 1: MoviesOnStreamingPlatforms_updated.csv

actor_1_facebook_likes	gross	genres	...	num_user_for_reviews	language	country	content_rating	budget	title_year	acto
1000.0	760505847.0	Action|Adventure|Fantasy|Sci-Fi	...	3054.0	English	USA	PG-13	237000000.0	2009.0	936
40000.0	309404152.0	Action|Adventure|Fantasy	...	1238.0	English	USA	PG-13	300000000.0	2007.0	500
11000.0	200074175.0	Action|Adventure|Thriller	...	994.0	English	UK	PG-13	245000000.0	2015.0	393
27000.0	448130642.0	Action|Thriller	...	2701.0	English	USA	PG-13	250000000.0	2012.0	230
131.0	NaN	Documentary	...	NaN	NaN	NaN	NaN	NaN	NaN	12.0

Figure 2: movie_metadata.csv

Our first steps in cleaning up the datasets involved identifying the columns we wanted to include in our final database to ensure we would have a column in both databases to allow us to merge them together. For the IMDB dataset we kept the *Title, Gross and Budget.* Due to the number of columns we wanted to keep in the second dataset, it was more efficient to remove the columns we did not wish to include; *ID, Age, Rotten Tomatoes, Type and Runtime*.

After setting the dataframes up, we created a connection to a sqlite database and created the two tables from the two dataframes. We then created a query to ensure the tables in the database were read correctly.

	Title	gross	budget
0	Avatar	760505847.0	237000000.0
1	Pirates of the Caribbean: At World's End	309404152.0	300000000.0
2	Spectre	200074175.0	245000000.0
3	The Dark Knight Rises	448130642.0	250000000.0
4	Star Wars: Episode VII - The Force Awakens ...	NaN	NaN

Figure 3: IMDb table

	Title	Year	IMDb	Netflix	Hulu	Prime Video	Disney+	Directors	Genres	Country	Language
0	Inception	2010	8.8	1	0	0	0	Christopher Nolan	Action,Adventure,Sci-Fi,Thriller	United States,United Kingdom	English,Japanese,French
1	The Matrix	1999	8.7	1	0	0	0	Lana Wachowski,Lilly Wachowski	Action,Sci-Fi	United States	English
2	Avengers: Infinity War	2018	8.5	1	0	0	0	Anthony Russo,Joe Russo	Action,Adventure,Sci-Fi	United States	English
3	Back to the Future	1985	8.5	1	0	0	0	Robert Zemeckis	Adventure,Comedy,Sci-Fi	United States	English
4	The Good, the Bad and the Ugly	1966	8.8	1	0	1	0	Sergio Leone	Western	Italy,Spain,West Germany	Italian
...											

Figure 4: Streaming Videos table

After creating and loading the tables, we attempted to merge the tables together. After several hours with a TA, we finally realized that the "Title" on both tables was not formatted the same. For some odd reason, on the IMDb dataset, there was an Ã after each title on the csv. Once we removed the extra character from the csv we were able to easily join the tables. We started with the streaming table and joined the IMDB table via an inner join on the Title column. We then created a combined table in sqlite. Finally, we wrote the table to html and exported the file to allow us to include it in our website.

```
'<table border="1" class="dataframe">\n  <thead>\n    <tr style="text-align: right;">\n      <th></th>\n      <th>Title</th>\n      <th>Y
ear</th>\n      <th>IMDb</th>\n      <th>Netflix</th>\n      <th>Hulu</th>\n      <th>Prime Video</th>\n      <th>Disney+</th>\n      <th
>Directors</th>\n      <th>Genres</th>\n      <th>Country</th>\n      <th>Language</th>\n      <th>Title</th>\n      <th>gross</th>\n
<th>budget</th>\n    </tr>\n  </thead>\n  <tbody>\n    <tr>\n      <th>0</th>\n      <td>Inception</td>\n      <td>2010</td>\n      <td>
8.8</td>\n      <td>1</td>\n      <td>0</td>\n      <td>0</td>\n      <td>0</td>\n      <td>0</td>\n      <td>Christopher Nolan</td>\n      <td>Action,Adve
nture,Sci-Fi,Thriller</td>\n      <td>United States,United Kingdom</td>\n      <td>English,Japanese,French</td>\n      <td>Inception</td>
\n      <td>292568851.0</td>\n      <td>1.600000e+08</td>\n    </tr>\n    <tr>\n      <th>1</th>\n      <td>The Matrix</td>\n      <td>19
99</td>\n      <td>8.7</td>\n      <td>1</td>\n      <td>0</td>\n      <td>0</td>\n      <td>0</td>\n      <td>Lana Wachowski,Lilly Wacho
wski</td>\n      <td>Action,Sci-Fi</td>\n      <td>United States</td>\n      <td>English</td>\n      <td>The Matrix</td>\n      <td>17138
3253.0</td>\n      <td>6.300000e+07</td>\n    </tr>\n    <tr>\n      <th>2</th>\n      <td>Back to the Future</td>\n      <td>1985</td>\n
<td>8.5</td>\n      <td>1</td>\n      <td>0</td>\n      <td>0</td>\n      <td>0</td>\n      <td>Robert Zemeckis</td>\n      <td>Adventur
e,Comedy,Sci-Fi</td>\n      <td>United States</td>\n      <td>English</td>\n      <td>Back to the Future</td>\n      <td>210609762.0</td>
\n      <td>1.900000e+07</td>\n    </tr>\n    <tr>\n      <th>3</th>\n      <td>The Good, the Bad and the Ugly</td>\n      <td>1966</td>
\n      <td>8.8</td>\n      <td>1</td>\n      <td>0</td>\n      <td>1</td>\n      <td>0</td>\n      <td>Sergio Leone</td>\n      <td>West
ern</td>\n      <td>Italy,Spain,West Germany</td>\n      <td>Italian</td>\n      <td>The Good, the Bad and the Ugly</td>\n      <td>61000
00.0</td>\n      <td>1.200000e+06</td>\n    </tr>\n    <tr>\n      <th>4</th>\n      <td>The Pianist</td>\n      <td>2002</td>\n      <td
>8.5</td>\n      <td>1</td>\n      <td>0</td>\n      <td>1</td>\n      <td>0</td>\n      <td>Roman Polanski</td>\n      <td>Biography,Dra
ma,Music,War</td>\n      <td>United Kingdom,France,Poland,Germany</td>\n      <td>English,German,Russian</td>\n      <td>The Pianist</td>
\n      <td>32519322.0</td>\n      <td>3.500000e+07</td>\n    </tr>\n    <tr>\n      <th>5</th>\n      <td>Django Unchained</td>\n      <
td>2012</td>\n      <td>8.4</td>\n      <td>1</td>\n      <td>0</td>\n      <td>0</td>\n      <td>0</td>\n      <td>Quentin Tarantino</td
>\n      <td>Drama,Western</td>\n      <td>United States</td>\n      <td>English,German,French,Italian</td>\n      <td>Django Unchained</
td>\n      <td>162804648.0</td>\n      <td>1.000000e+08</td>\n    </tr>\n    <tr>\n      <th>6</th>\n      <td>Raiders of the Lost Ark</t
```


Figure 5: HTML of Final Database

Website Creation

We created a website for our new database that gives the viewer a brief description of why we created our website in addition to the data to help movie watchers find more information about a specific movie they're watching and its corresponding available streaming platforms. One of the unique components of our website was that we created our own logo, and uploaded it to the top left of our navbar. We also went with a darker color scheme theme to compliment the traditionally dark environment of movie theatres, making it easier on the eyes for night-time or in-theater movie searches. The website is also responsive and compatible with small and large screen sizes. For example, on a smaller screen the two links in the navigation bar, "Home Page" and "Movie Data", appear as a "hamburger style" drop-down menu. When you go to our "Movie Data" page you will see we changed the default color of our text to white so it has more of a contrast from the background. As well as we made our table responsive so as you're scrolling through our data it isn't just one super long webpage.

Summary

Given the increased demand for easily accessible at-home entertainment, M.A.R movies created a database for families and movie aficionados alike. We compiled information from IMDB and Kaggle on movie directors, genres, available streaming platforms, budgets, and revenues for easy data exploration. Users can determine what streaming platforms work best for them based on their preference for low budget films or popular blockbuster movies.

WORKS CITED

"Movies on Netflix, Prime Video, Hulu and Disney+" Kaggle.com, Retrieved 08/25/2020. https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney

"Understanding Movies Through Data" Kaggle.com, Retrieved 08/25/2020. https://www.kaggle.com/karrrimba/movie-metadatacsv