ASSIGNMENT 1

DATA MANAGEMENT AND DATABASE DESIGN

INFO6210_SPRING2019_CLASS03

By

PREETHAM REDDY ALLADU (NUID: 001355231)

&

SINDHURA KOLLI (NUID: 001387687)

Abstract:

In this assignment, we selected Movie's as the them. We created a conceptual relation between three different movie data sources through web scraping, web API and csv dataset. After gathering the data, reformatting is done and we showed that the data needs no further cleaning and is complete and accurate. Further, the data is split into three tables with the relevant attributes.

Data Theme and Sources:

The theme chosen for this assignment is "Movies". Database consists of multiple attributes related to different movies collected from different data sources using Web Scraping, Web API and a CSV dataset.

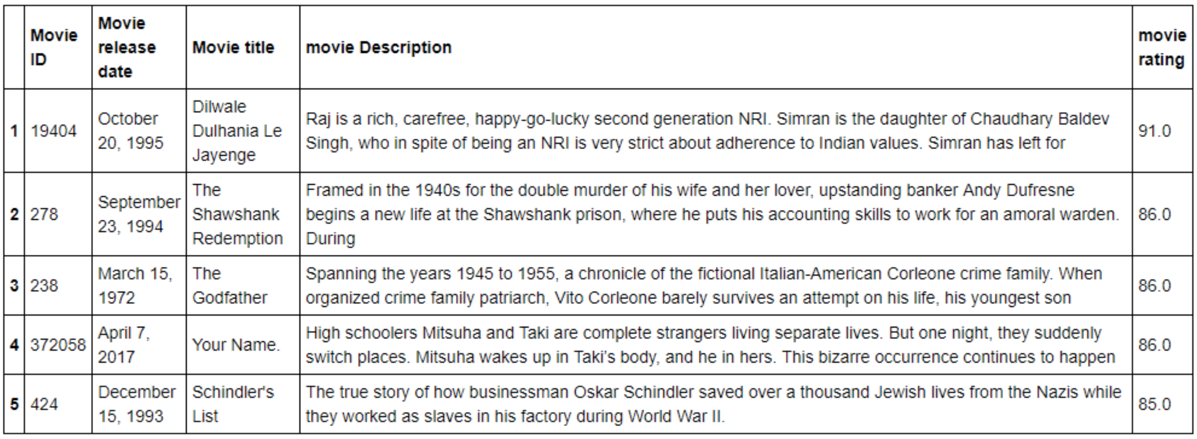
Below are the list of attributes collected from three sources:

Using Web Scraper:

The data frame for the following attributes was created by scraping "The Movie DB" website using BeautifulSoup.

URL: https://www.themoviedb.org/

1. Movie ID
2. Movie title
3. Movie release date
4. Movie Description
5. Movie rating

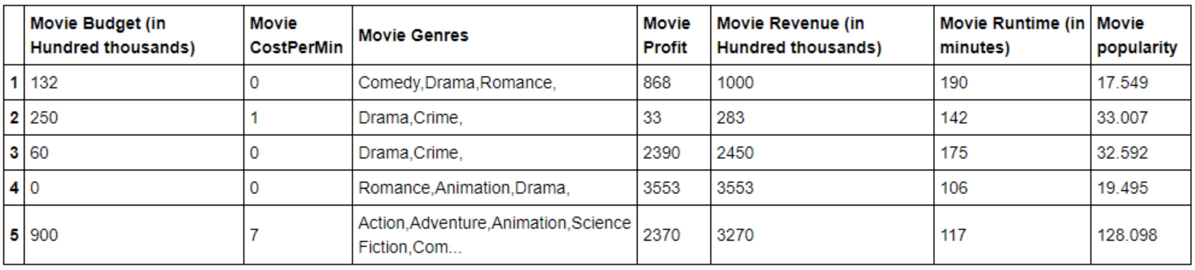


	Movie ID	Movie release date	Movie title	movie Description	movie rating
1	19404	October 20, 1995	Dilwale Dulhania Le Jayenge	Raj is a rich, carefree, happy-go-lucky second generation NRI. Simran is the daughter of Chaudhary Baldev Singh, who in spite of being an NRI is very strict about adherence to Indian values. Simran has left for	91.0
2	278	September 23, 1994	The Shawshank Redemption	Framed in the 1940s for the double murder of his wife and her lover, upstanding banker Andy Dufresne begins a new life at the Shawshank prison, where he puts his accounting skills to work for an amoral warden. During	86.0
3	238	March 15, 1972	The Godfather	Spanning the years 1945 to 1955, a chronicle of the fictional Italian-American Corleone crime family. When organized crime family patriarch, Vito Corleone barely survives an attempt on his life, his youngest son	86.0
4	372058	April 7, 2017	Your Name.	High schoolers Mitsuha and Taki are complete strangers living separate lives. But one night, they suddenly switch places. Mitsuha wakes up in Taki's body, and he in hers. This bizarre occurrence continues to happen	86.0
5	424	December 15, 1993	Schindler's List	The true story of how businessman Oskar Schindler saved over a thousand Jewish lives from the Nazis while they worked as slaves in his factory during World War II.	85.0

Using Web API:

The data frame for the following attributes was created using the TMDb's API called "tmdbsimple"

1. Movie Budget (in Hundred thousands)
2. Movie Revenue (in Hundred thousands)
3. Movie Runtime (in minutes)
4. Movie Genres
5. Movie popularity

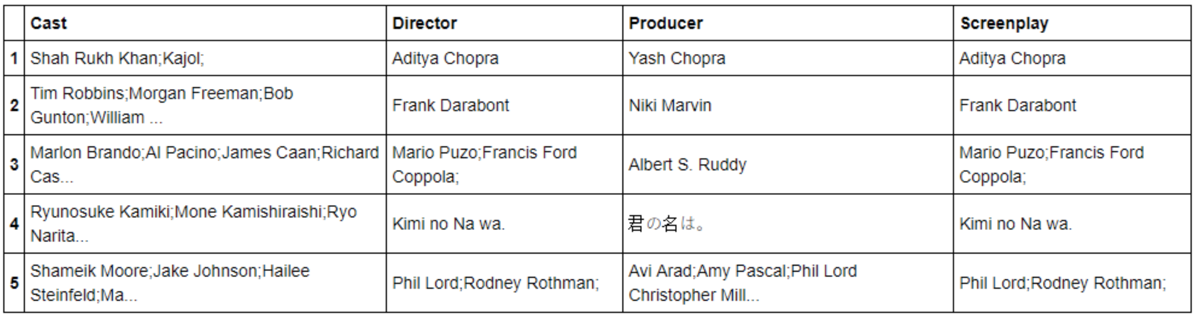


	Movie Budget (in Hundred thousands)	Movie CostPerMin	Movie Genres	Movie Profit	Movie Revenue (in Hundred thousands)	Movie Runtime (in minutes)	Movie popularity
1	132	0	Comedy,Drama,Romance,	868	1000	190	17.549
2	250	1	Drama,Crime,	33	283	142	33.007
3	60	0	Drama,Crime,	2390	2450	175	32.592
4	0	0	Romance,Animation,Drama,	3553	3553	106	19.495
5	900	7	Action,Adventure,Animation,Science Fiction,Com...	2370	3270	117	128.098

CSV Dataset:

The CSV dataset was created by scraping Wikipedia pages of relevant movies.

1. Director
2. Producer
3. Cast
4. Screen

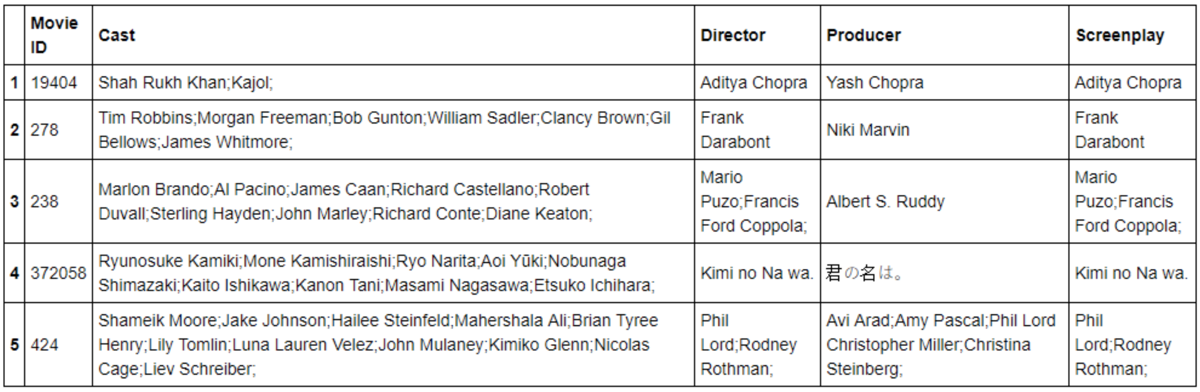


	Cast	Director	Producer	Screenplay
1	Shah Rukh Khan;Kajol;	Aditya Chopra	Yash Chopra	Aditya Chopra
2	Tim Robbins;Morgan Freeman;Bob Gunton;William ...	Frank Darabont	Niki Marvin	Frank Darabont
3	Marlon Brando;Al Pacino;James Caan;Richard Cas...	Mario Puzo;Francis Ford Coppola;	Albert S. Ruddy	Mario Puzo;Francis Ford Coppola;
4	Ryunosuke Kamiki;Mone Kamishiraishi;Ryo Narita...	Kimi no Na wa.	君の名は。	Kimi no Na wa.
5	Shameik Moore;Jake Johnson;Hailee Steinfeld;Ma...	Phil Lord;Rodney Rothman;	Avi Arad;Amy Pascal;Phil Lord Christopher Mill...	Phil Lord;Rodney Rothman;

Data Reformat and Conceptual Schema:

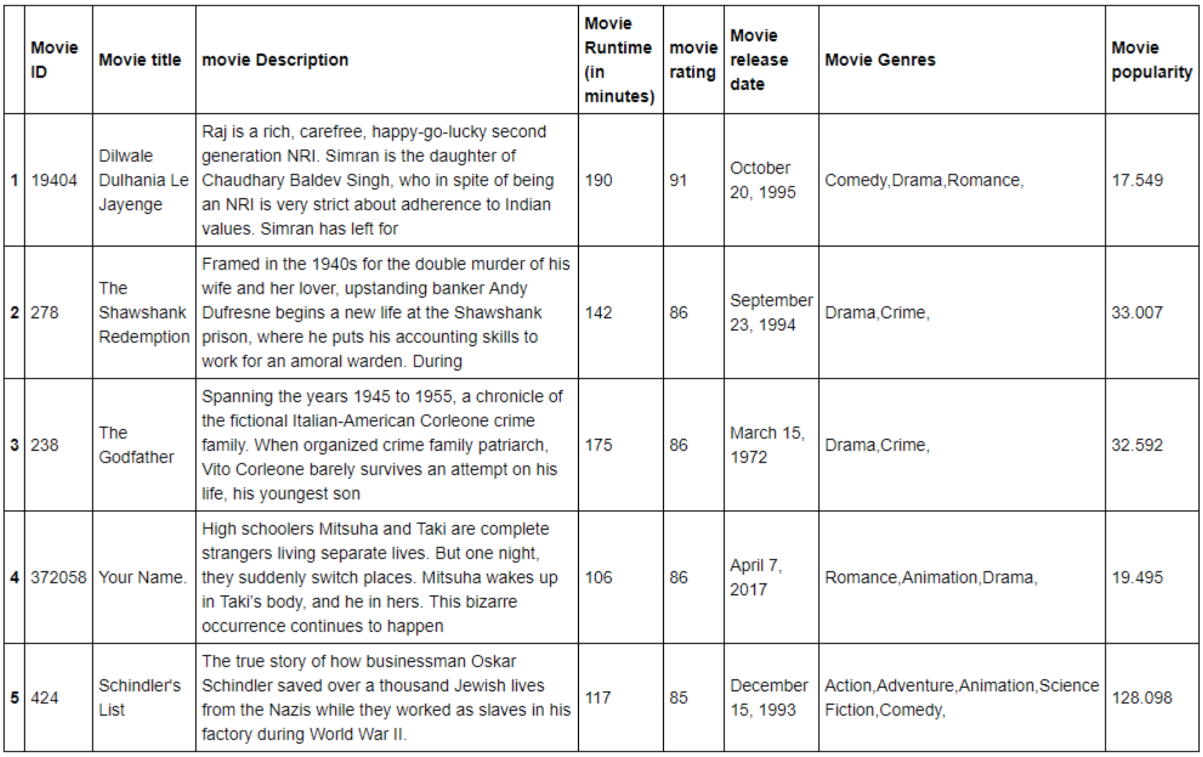
Collected the data from three sources and separated the datasets into the following three tables.

1. People
   a. Movie_id
   b. Movie_cast
   c. Directors
   d. Producers
   e. Screenplay



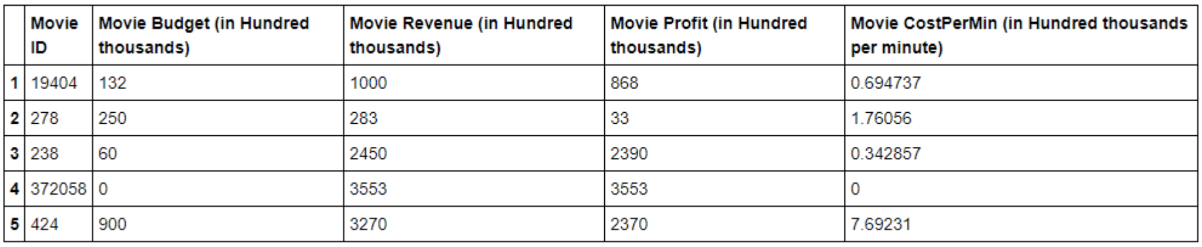
	Movie ID	Cast	Director	Producer	Screenplay
1	19404	Shah Rukh Khan;Kajol;	Aditya Chopra	Yash Chopra	Aditya Chopra
2	278	Tim Robbins;Morgan Freeman;Bob Gunton;William Sadler;Clancy Brown;Gil Bellows;James Whitmore;	Frank Darabont	Niki Marvin	Frank Darabont
3	238	Marlon Brando;Al Pacino;James Caan;Richard Castellano;Robert Duvall;Sterling Hayden;John Marley;Richard Conte;Diane Keaton;	Mario Puzo;Francis Ford Coppola;	Albert S. Ruddy	Mario Puzo;Francis Ford Coppola;
4	372058	Ryunosuke Kamiki;Mone Kamishiraishi;Ryo Narita;Aoi Yūki;Nobunaga Shimazaki;Kaito Ishikawa;Kanon Tani;Masami Nagasawa;Etsuko Ichihara;	Kimi no Na wa.	君の名は。	Kimi no Na wa.
5	424	Shameik Moore;Jake Johnson;Hailee Steinfeld;Mahershala Ali;Brian Tyree Henry;Lily Tomlin;Luna Lauren Velez;John Mulaney;Kimiko Glenn;Nicolas Cage;Liev Schreiber;	Phil Lord;Rodney Rothman;	Avi Arad;Amy Pascal;Phil Lord Christopher Miller;Christina Steinberg;	Phil Lord;Rodney Rothman;

2. Movie_elements
    a. Movie_id
    b. Title
    c. Description
    d. Runtime
    e. Rating
    f. Release_date
    g. Genre
    h. Popularity



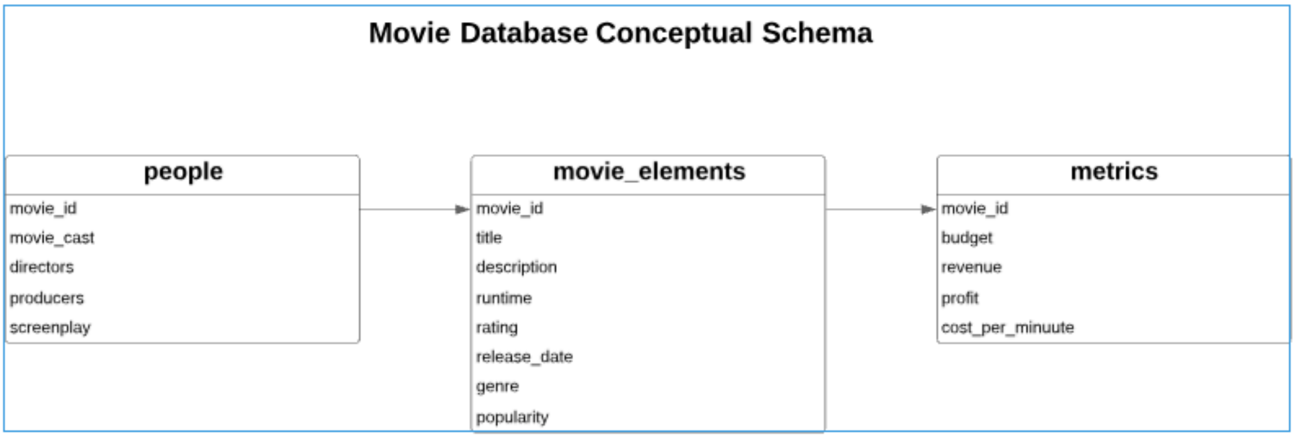
	Movie ID	Movie title	movie Description	Movie Runtime (in minutes)	movie rating	Movie release date	Movie Genres	Movie popularity
1	19404	Dilwale Dulhania Le Jayenge	Raj is a rich, carefree, happy-go-lucky second generation NRI. Simran is the daughter of Chaudhary Baldev Singh, who in spite of being an NRI is very strict about adherence to Indian values. Simran has left for	190	91	October 20, 1995	Comedy,Drama,Romance,	17.549
2	278	The Shawshank Redemption	Framed in the 1940s for the double murder of his wife and her lover, upstanding banker Andy Dufresne begins a new life at the Shawshank prison, where he puts his accounting skills to work for an amoral warden. During	142	86	September 23, 1994	Drama,Crime,	33.007
3	238	The Godfather	Spanning the years 1945 to 1955, a chronicle of the fictional Italian-American Corleone crime family. When organized crime family patriarch, Vito Corleone barely survives an attempt on his life, his youngest son	175	86	March 15, 1972	Drama,Crime,	32.592
4	372058	Your Name.	High schoolers Mitsuha and Taki are complete strangers living separate lives. But one night, they suddenly switch places. Mitsuha wakes up in Taki's body, and he in hers. This bizarre occurrence continues to happen	106	86	April 7, 2017	Romance,Animation,Drama,	19.495
5	424	Schindler's List	The true story of how businessman Oskar Schindler saved over a thousand Jewish lives from the Nazis while they worked as slaves in his factory during World War II.	117	85	December 15, 1993	Action,Adventure,Animation,Science Fiction,Comedy,	128.098

3. Metrics
    a. Movie_id
    b. Budget
    c. Revenue
    d. Profit
    e. Cost_per_minute



	Movie ID	Movie Budget (in Hundred thousands)	Movie Revenue (in Hundred thousands)	Movie Profit (in Hundred thousands)	Movie CostPerMin (in Hundred thousands per minute)
1	19404	132	1000	868	0.694737
2	278	250	283	33	1.76056
3	238	60	2450	2390	0.342857
4	372058	0	3553	3553	0
5	424	900	3270	2370	7.69231

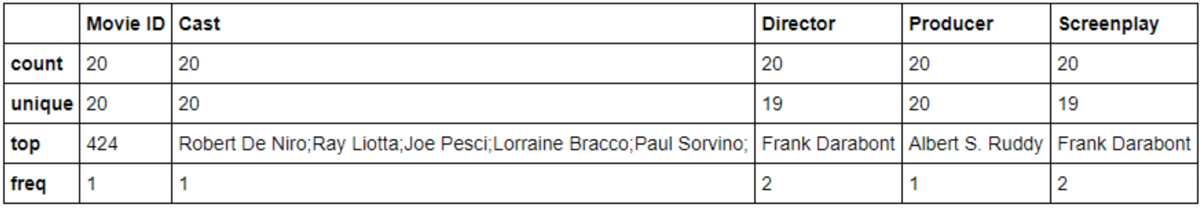
Below is the object model diagram showing the relation between three tables having “movie_id” as primary key.



Audit Accuracy:

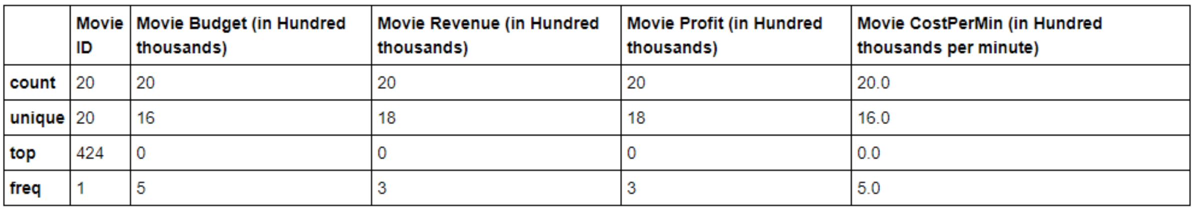
Used Python to check whether there are any duplicate values and delete the row with those values. The results below shows that there are no duplicate values for Movie_ID in the “freq” row.

People Table:



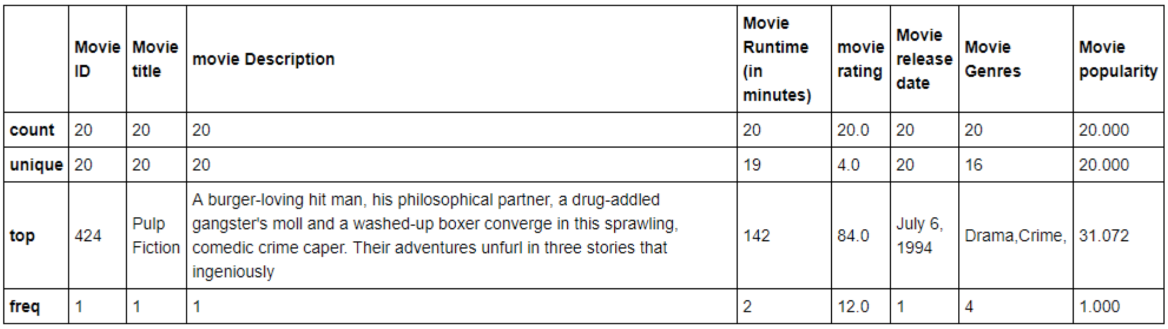
	Movie ID	Cast	Director	Producer	Screenplay
count	20	20	20	20	20
unique	20	20	19	20	19
top	424	Robert De Niro;Ray Liotta;Joe Pesci;Lorraine Bracco;Paul Sorvino;	Frank Darabont	Albert S. Ruddy	Frank Darabont
freq	1	1	2	1	2

Metrics Table:



	Movie ID	Movie Budget (in Hundred thousands)	Movie Revenue (in Hundred thousands)	Movie Profit (in Hundred thousands)	Movie CostPerMin (in Hundred thousands per minute)
count	20	20	20	20	20.0
unique	20	16	18	18	16.0
top	424	0	0	0	0.0
freq	1	5	3	3	5.0

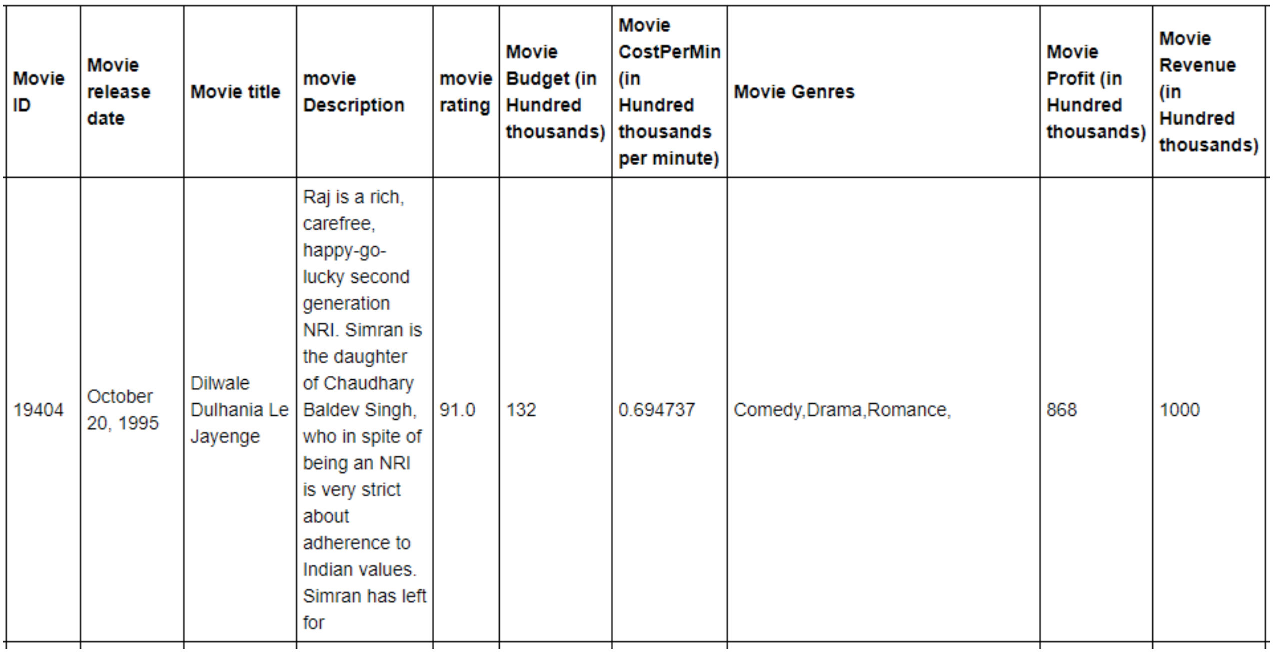
Movie_elements Table:



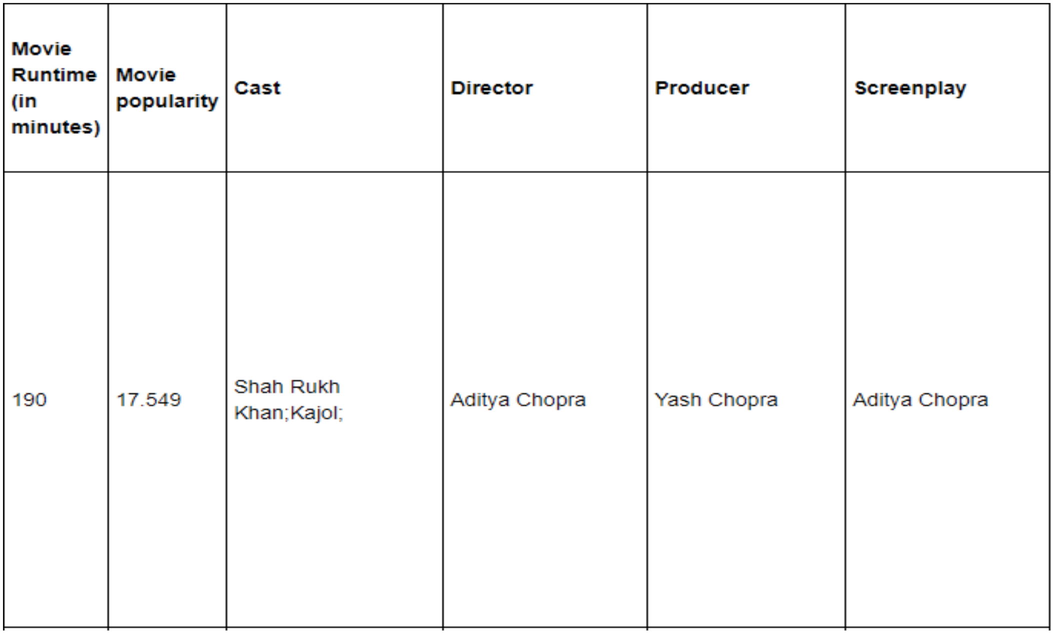
	Movie ID	Movie title	movie Description	Movie Runtime (in minutes)	movie rating	Movie release date	Movie Genres	Movie popularity
count	20	20	20	20	20.0	20	20	20.000
unique	20	20	20	19	4.0	20	16	20.000
top	424	Pulp Fiction	A burger-loving hit man, his philosophical partner, a drug-addled gangster's moll and a washed-up boxer converge in this sprawling, comedic crime caper. Their adventures unfurl in three stories that ingeniously	142	84.0	July 6, 1994	Drama,Crime,	31.072
freq	1	1	1	2	12.0	1	4	1.000

Audit Completeness:

The data from three different sources are related and are from real world websites. Below is the screen of sample data.



Movie ID	Movie release date	Movie title	movie Description	movie rating	Movie Budget (in Hundred thousands)	Movie CostPerMin (in Hundred thousands per minute)	Movie Genres	Movie Profit (in Hundred thousands)	Movie Revenue (in Hundred thousands)
19404	October 20, 1995	Dilwale Dulhania Le Jayenge	Raj is a rich, carefree, happy-go-lucky second generation NRI. Simran is the daughter of Chaudhary Baldev Singh, who in spite of being an NRI is very strict about adherence to Indian values. Simran has left for	91.0	132	0.694737	Comedy,Drama,Romance,	868	1000

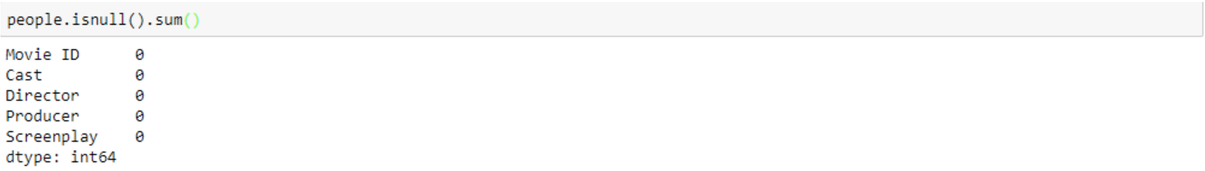


Movie Runtime (in minutes)	Movie popularity	Cast	Director	Producer	Screenplay
190	17.549	Shah Rukh Khan;Kajol;	Aditya Chopra	Yash Chopra	Aditya Chopra

Audit Consistency/Uniformity:

The data does not contain any missing values for any column. All the data throughout is consistence and has no null values.

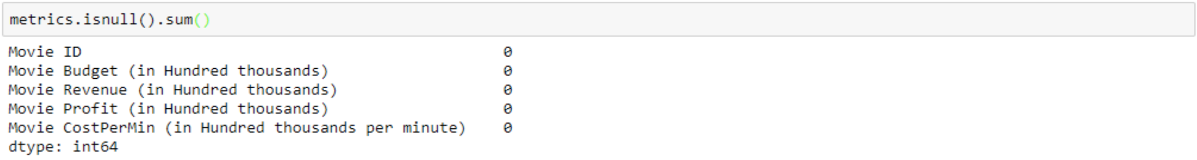
People Table:



```
people.isnull().sum()
```

```
Movie ID      0
Cast          0
Director      0
Producer      0
Screenplay    0
dtype: int64
```

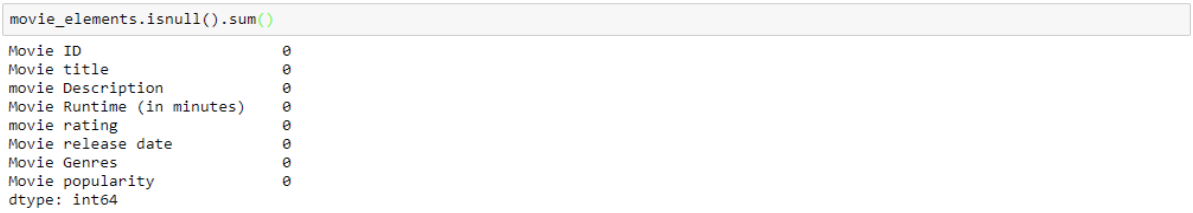
Metrics Table:



```
metrics.isnull().sum()
```

```
Movie ID                                            0
Movie Budget (in Hundred thousands)                 0
Movie Revenue (in Hundred thousands)                0
Movie Profit (in Hundred thousands)                 0
Movie CostPerMin (in Hundred thousands per minute)  0
dtype: int64
```

Movie_elements Table:



```
movie_elements.isnull().sum()
```

```
Movie ID                     0
Movie title                  0
movie Description            0
Movie Runtime (in minutes)   0
movie rating                 0
Movie release date           0
Movie Genres                 0
Movie popularity             0
dtype: int64
```

Citations and References:

All the code is self-developed and is not copied from any website. Please refer the Jupyter notebook Preetham_Sindhura_Assignment1.ipynb attached along with this document for the code.

References were taken from the following websites:
https://www.w3schools.com/python/python_syntax.asp
https://docs.python.org/3/tutorial/
https://www.themoviedb.org/documentation/api