EXTRACT:

Source:

https://www.kaggle.com/jealousleopard/goodreadsbooks
books.csv

TRANSFORM:

From:

Dataframe: books_df

LOAD:

Into:

Table: books

Raw Data:	**Action:**
bookID	keep - not null as book_id
title	keep - not null as title
authors	Remove
average_rating	keep - not null as average_rating
isbn	remove
isbn13	Remove
language_code	Remove
num_pages	Remove
ratings_count	keep - not null as ratings_count
text_reviews_count	keep - not null as text_reviews_count
publication_date	keep - not null as publication_date
publisher	keep - not null as publisher
Unnamed:12	Remove
# of records: 11,127	**# of records: 11,127**

Table: books
book_id
title
average_rating
ratings_count
text_reviews_count
publication_date
publisher
of records: 11,127

EXTRACT:

Source:

https://datasets.imdbws.com/

title.basics.tsv.gz

Raw Data:
tconst
titleType
primaryTitle
originalTitle
isAdult
startYear
endYear
runtimeMinutes
genres
of records: 6,762,842

TRANSFORM:

From:

Dataframe: movie_df

Action:
keep - only values in ratings_df as tconst
keep - "movies" as title_type
keep - not null as primary_title
keep - not null as original_title
Remove
keep - not null as start_year
Remove
keep - as runtime_minutes
keep - not null as genres
of records: 246,366

LOAD:

Into:

Table: movies

Table: movies
tconst
title_type
primary_title
original_title
start_year
runtime_minutes
genres
of records: 246,366

EXTRACT:	**TRANSFORM:**	**LOAD:**
Source:	**From:**	**Into:**
https://datasets.imdbws.com/	**Dataframe: movie_df**	**Table: movies**
title.ratings.tsv.gz		
Raw Data:	**Action:**	**Table: movies**
tconst averageRating numVotes	keep - only values in movie_df as tconst keep - not null as average_rating keep - not nul as num_votes	tconst average_rating num_votes
# of records: 1,033,876	**# of records: 246,366**	**# of records: 246,366**

EXTRACT:	**TRANSFORM:**	**LOAD:**
Source:	**From:**	**Into:**
https://en.wikipedia.org/wiki/Lists_of_fiction_works_made_into_feature_films	**Dataframe: wiki_df**	**Table: wiki**
page = requests.get(url) **soup = BeautifulSoup(page.text,"lmxl")**		
Raw Data:	**Action:**	**Table: movies**
Book Titles Movie Titles	keep - not null as book_title keep - not null as movie_title	book_title movie_title
# of records: 1,637	**# of records: 1,637**	**# of records: 1,637**