EXTRACT:

Source:

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

Raw Data:

bookID
title
authors
average_rating
isbn
isbn13
language_code
num_pages
ratings_count
text_reviews_count
publication_date
publisher
Unnamed:12

# of records: 11,127

TRANSFORM:

From:

Dataframe: books_df

Action:

keep - not null as book_id
keep - not null as title
Remove
keep - not null as average_rating
remove
Remove
Remove
Remove
keep - not null as ratings_count
keep - not null as text_reviews_count
keep - not null as publication_date
keep - not null as publisher
Remove

# of records: 11,127

LOAD:

Into:

Table: books

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: ratings_df**	**Table: ratings**
title.ratings.tsv.gz		
Raw Data:	**Action:**	**Table: ratings**
tconst	keep - only values in movie_df as tconst	tconst
averageRating	keep - not null as average_rating	average_rating
numVotes	keep - not nul as num_votes	num_votes
# of records: 1,033,876	**# of records: 246,366**	**# of records: 246,366**

EXTRACT:

Source:

https://en.wikipedia.org/wiki/Lists_of_fiction_works_made_into_feature_films

```
page = requests.get(url)
soup = BeautifulSoup(page.text,"lmxl")
```

Raw Data:

Book Titles
Movie Titles

# of records: 1,637

TRANSFORM:

From:

Dataframe: wiki_df

Action:

keep - not null as book_title
keep - not null as movie_title

# of records: 1,637

LOAD:

Into:

Table: wiki

Table: wiki

book_title
movie_title

# of records: 1,637