EAS 508 Project Report - IMDB

Group Members: Kurusali Sree Nandhini, Keerthy Priya Vanam, Sonali Misra, Susmitha Yanamadala
*Group Number: **8***

Abstract:

IMDB database contains information about Movies and Crew members. The dataset consists of movies released on or before 2022. Data includes genres, cast, crew, release dates, languages, countries, IMDB vote counts and ratings. This data will be helpful for those who are interested to get into movie field. By using this database, they can easily get to know about Directors, Producers, crew members so it will be easy for them to analyze the famous directors who have directed a lot of good movies. This data will also be helpful for movie buffs, using this data they will get to know the ratings of different TV shows, Movies etc. This will help them to analyze what movies tend to get higher vote counts and vote averages on IMDB.

Introduction:

In this project we will build a SQLite database using the Internet Movie Database (IMDb) dataset. The dataset consists of 5 compressed tab-separated-value (*.tsv) files.

The purpose of this project is to do the following:

1. Create a SQLite database, normalize the tables and ingest data into the database.
2. Manipulate Data using embedded SQL and load data into python IDE.
3. Visualize the data using Matplotlib and Seaborn libraries.

Data:

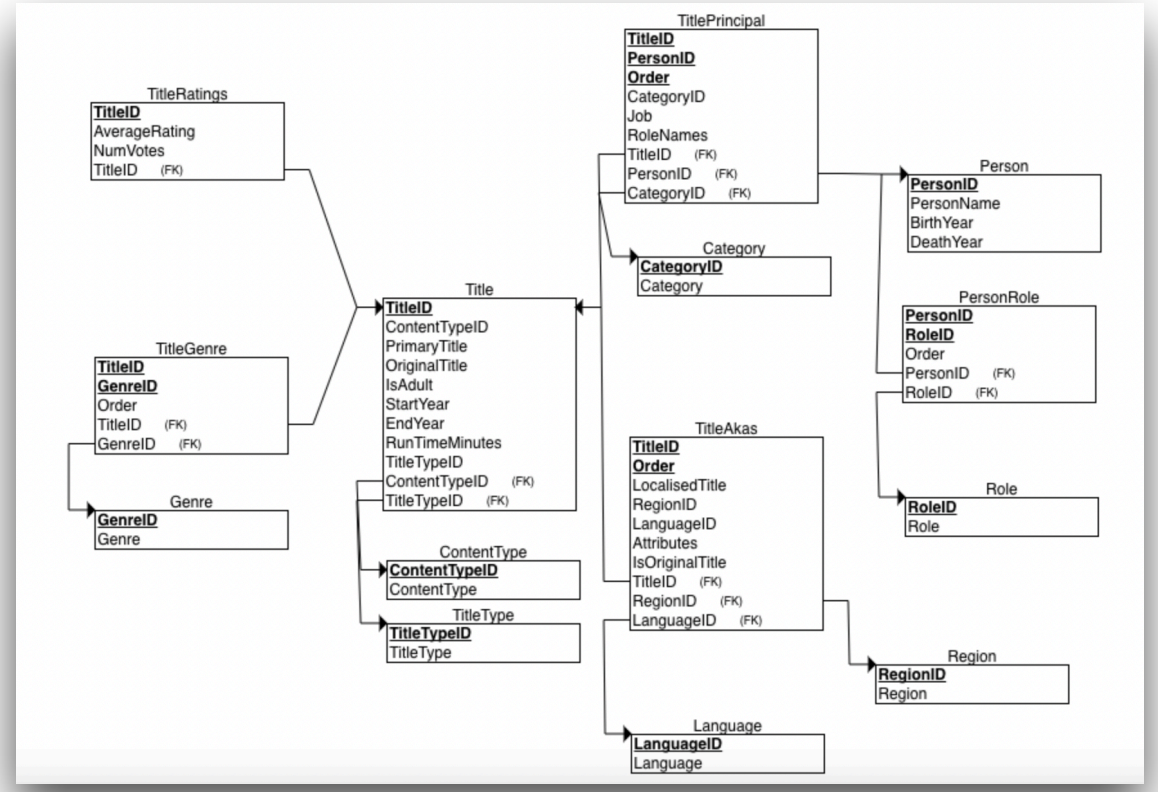
Each dataset is a tab- separated-values (TSV) formatted file in the UTF-8 character set. The first line in each file contains headers that describe what is in each column. A "\N" is used to denote that a particular field is missing or has a NULL value for that title or name.

TableName	Name	TitleAkas	TitleBasics	TitlePrincipals	TitleRatings
Column names	PersonID	TitleID	TitleID	TitleID	TitleID
	PersonName	Ordering	TitleType	Ordering	AverageRating
	BirthYear	LocalisedTitle	PrimaryTitle	PersonID	NumVotes
	DeathYear	Region	OriginalTitle	Category	
	PrimaryProfession	Language	IsAdult	Characters	
	KnownForTitles	Types	StartYear	Job	
		Attributes	EndYear		
		IsOriginalTitle	RuntimeMinutes		
			Genres		
Number of records	12134964	34122007	9430996	53430089	1254849

Data Normalization:

The 5 datasets are loaded, cleaned and processed using Python. Once the datasets are processed, we created tables based on 3rd normal form as the initial data is in de-normalized form. Data ingestion was performed after the tables creation. After normalizing the data, we have 14 tables and below are the details and data model:

S.NO	Tables	Column names		S.NO	Tables	Column names
1	**Category**	CategoryID, Category		8	**Title**	TitleID, Title, OriginalTitle Isadult, StartYear EndYear, Runtimeminutes
2	**ContentType**	ContentID,ContentType		9	**Person**	PersonID, PersonName BirthYear, DeathYear
3	**Genre**	GenreID, Genre		10	**Personrole**	PersonID, RoleID, Order
4	**Language**	languageID, Language		11	**TitleAkas**	TitleID, Order, Localisedtitle Regionid, Languageid Attributes, IsoriginalTitle
5	**Region**	RegionID, Region		12	**TitleGenre**	TitleID, GenreID, Order
6	**Role**	RoleID, Role		13	**TitlePrincipal**	TitleID, PersonID, Order CategoryID, Job, Rolenames
7	**TitleType**	TitletypeID, Titletype		14	**TitleRatings**	TitleID, AverageRating, NumVotes



Exploratory Data Analysis:

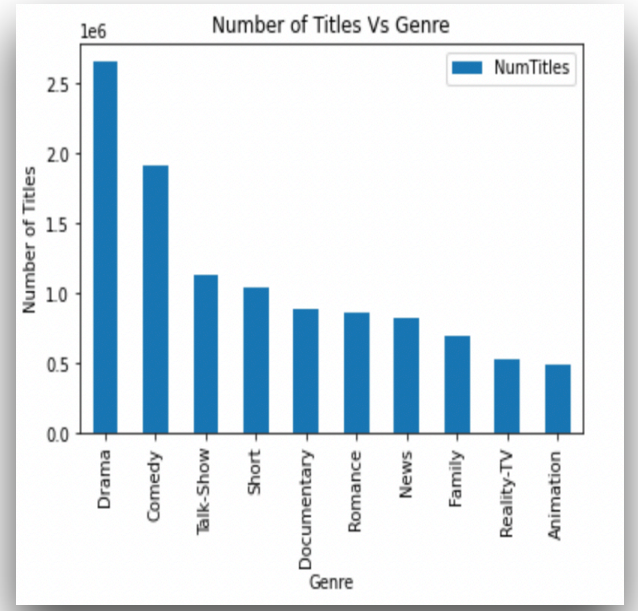
Once the table creation and ingestion is done, we performed exploratory data analysis by loading data into Python using embedded SQL. Below are the insights we observed when performing EDA:

	Title	AverageRating
0	Illusionen	10.0
1	Stück für Stück	10.0
2	Mein oder Dein	10.0
3	Der Kaktusgarten	10.0
4	All I Know Is	10.0
5	Spiral Tribe	10.0
6	Desert	10.0
7	Renegades 2	10.0
8	Verkündigung	10.0
9	Das Gold von Bayern	10.0

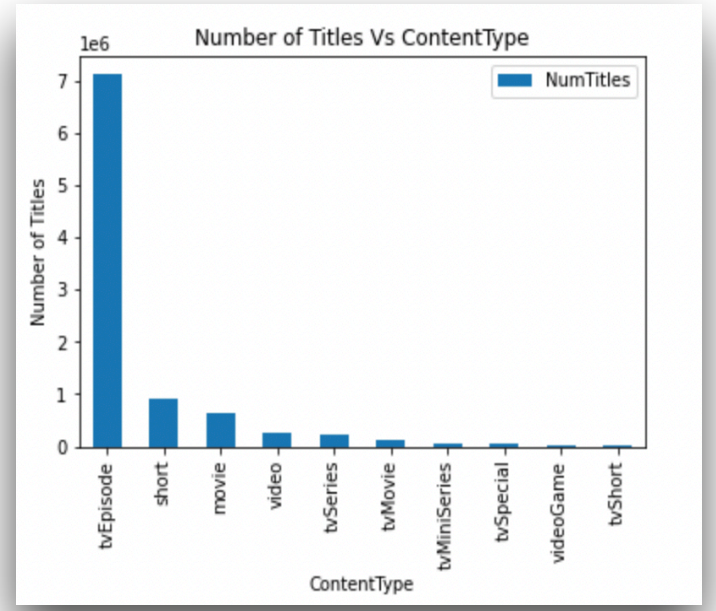
Top 10 movies based on average ratings

	Title	AverageRating
0	Graustark	1.0
1	The Jungle Trail	1.0
2	Terror Trail	1.0
3	Young Lochinvar	1.0
4	Wild Justice	1.0
5	Back to Liberty	1.0
6	Águilas de acero o los misterios de Tánger	1.0
7	Madsalune	1.0
8	The Rip-Tide	1.0
9	Through Fire and Water	1.0

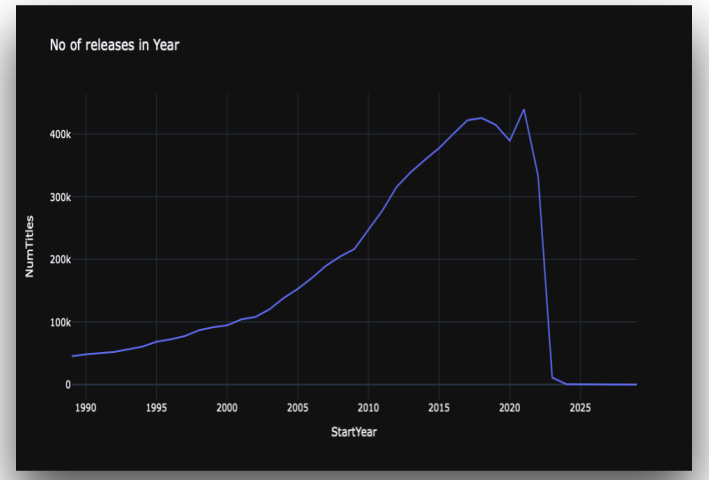
Movies with Low ratings



Number of Titles in respective Genres



Number of Titles per content type



Number of Releases in a Year

	PersonName	AverageRating
0	Max Mills	10.0
1	Harvey Mills	10.0
2	Jerry Pointer	10.0
3	Paige Hennington	10.0
4	Siddharth Chanakya	10.0
5	Julia Aku	10.0
6	Professor Ali	10.0
7	Molly Baker	10.0
8	Fernand Dumont	10.0
9	Danai Maria Karamolegou	10.0

Top persons with highest Rating

From the above analysis, we can infer that:

- Most number of titles are made in Drama Genre followed by Comedy Genre.
- Most number of titles made are TV Episodes.
- Number of releases in a year is monotonically increasing.

Conclusion:

People can analyze this data and get any information related to movie industry across the globe.

Future Scope:

For next steps, we can implement movie recommendation engine and ratings prediction.

References:

https://www.imdb.com/interfaces/

https://datasets.imdbws.com

https://www.kaggle.com/datasets/komalkhetlani/imdb-dataset

Link to Code (UB Box):

https://buffalo.box.com/s/bh91b4myds0y4efv3vpenm7nkzsrezli

The above link has:

1. table creation and data insertion.py - this file has data processing, table creation and data ingestion code.
2. EDA.ipynb - this notebook has exploratory data analysis code.

** The size of Datasets/Database is very large i.e., 6.21GB, so we couldn't upload it to UB Box.