ETL PROJECT

E - Extracted 4 IMDb CSV’s from Kaggle dataset.

Resource folder contains original dataset,

Main Data has the transformed CSV’s,

Query Results have the screenshots of our query results

**CSV’s Content**

The movies dataset includes 81,273 movies with attributes such as movie description, average rating, number of votes, genre, etc.

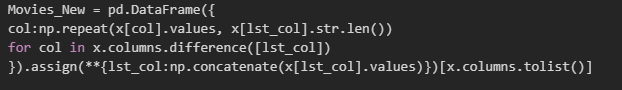
The ratings dataset includes 81,273 rating details from demographic perspective.

The names dataset includes 175,719 cast members with personal attributes such as birth details, death details, height, spouses, children, etc.

The title principals dataset includes 377,848 cast members roles in movies with attributes such as IMDb title id, IMDb name id, order of importance in the movie, role, and characters played.

T - Cleaned and cleared data using python, pandas etc.:

Dropped irrelevant columns of things we didn’t need or were lacking in data, made dates easier to access, made genres for demos easier to access, etc.



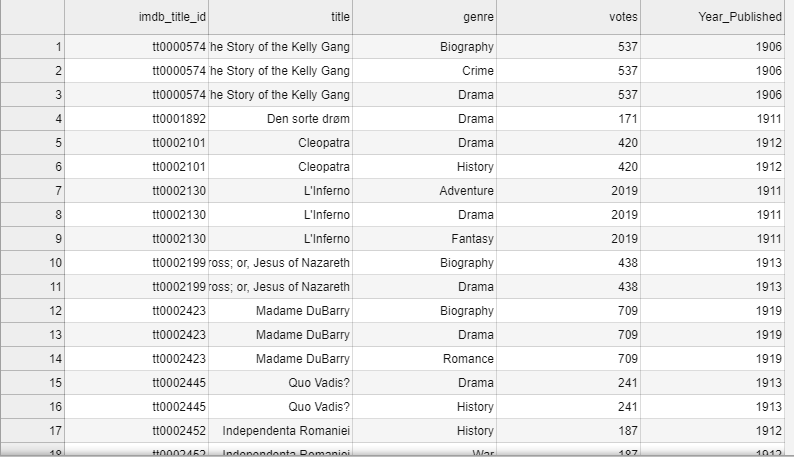


(Jupyter file : Analysis.ipynb)

Before sample:



After sample:

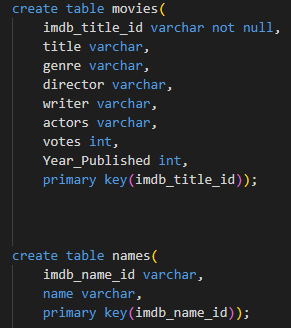
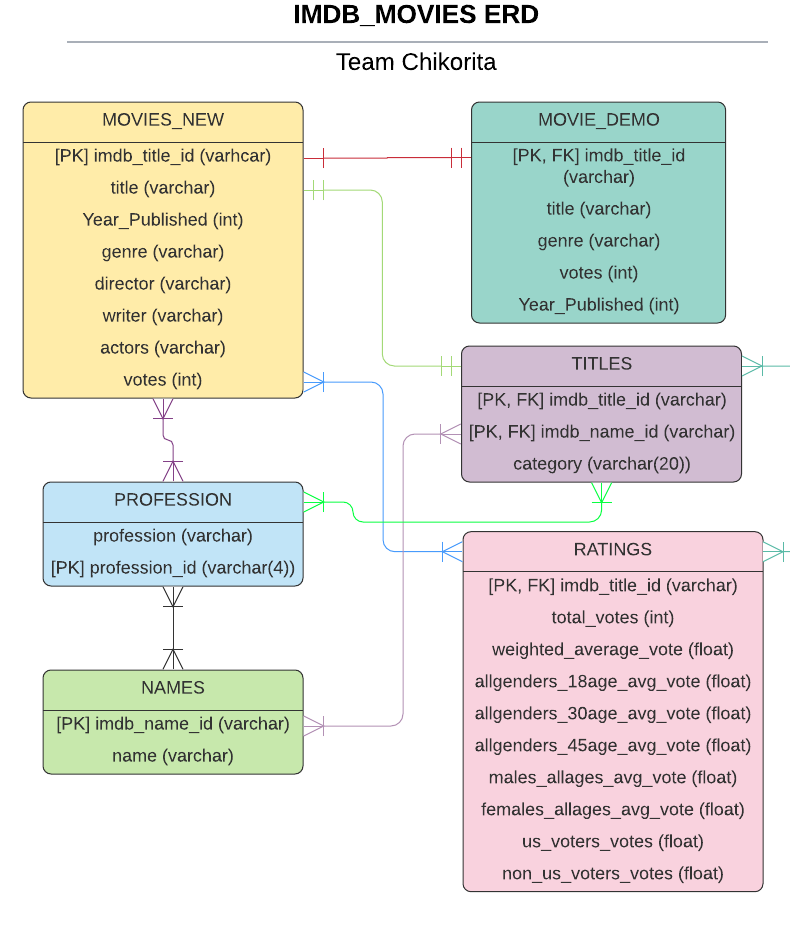


L (Loading into the database)-

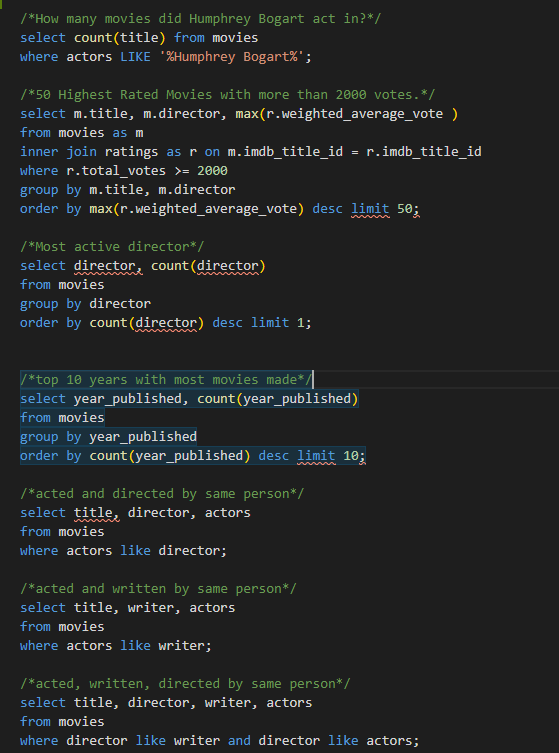
Saved changed dataframes back to csvs, created SQL schema with primary and foreign keys, created tables and imported data, performed multiple and various queries including joins and aggregation to show usefulness and depth.

Why PostgreSQL?

1. The data we extracted was already in a relational format and Postgres works great with relational data.
2. It’s easy to query with complexity.
3. It has the power to relate tables to each other with PK/FK.
4. Common usage.
5. + it has a dark theme.

 (SQLSchema.SQL)

(SQLQueries.SQL sample)



(QueryResults/Figs. sample)

