IMDb Data

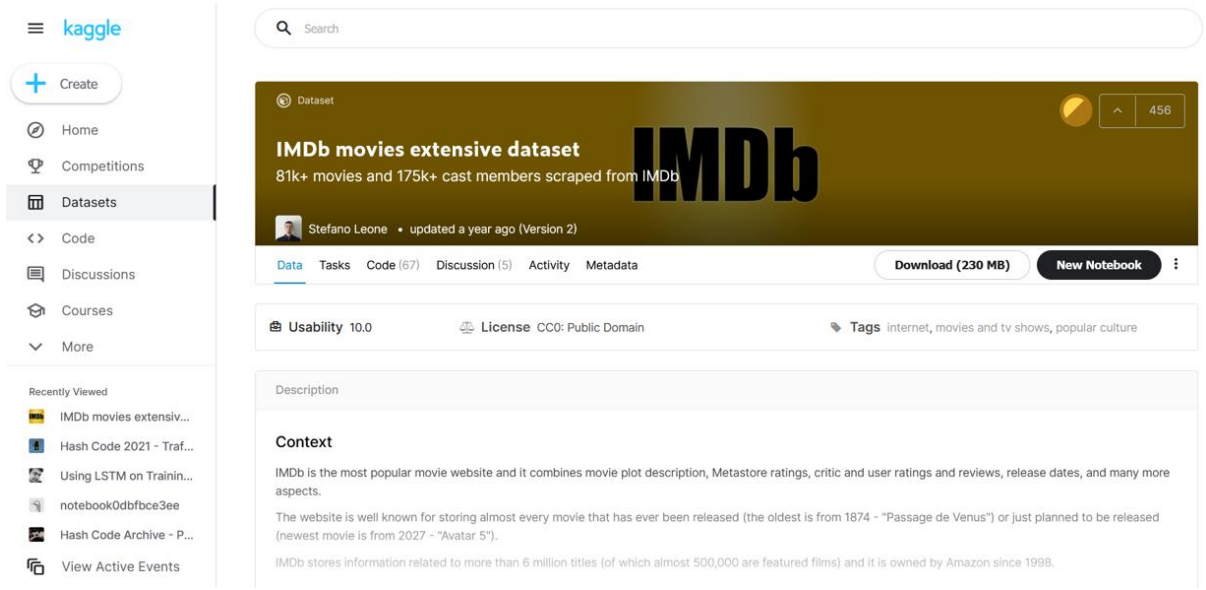
GONÇALO TEIXEIRA – UP201806562

PEDRO PINTO – UP201806251

PEDRO AZEVEDO – UP201806728

Dataset

- IMDb dataset about movies, ratings and movie industry related people



- Dataset includes 85855 movies with ratings and 297705 cast members

Data Refinement

- Remove columns with more than 1/3 missing values
- Remove repeated columns across different tables
- Remove columns with irrelevant information

Movies Table

Original		Refined	
Imdb_title_id	Production_company	Imdb_title_id	Production_company
title	actors	title	actors
Original_title	description	Original_title	description
year	Avg_votes	year	
Date_published	votes	Date_published	
genre	budget	genre	
duration	Usa_gross_income	duration	
country	Worldwide_gross_income	country	
language	metascore	language	
director	Reviews_from_users	director	Reviews_from_users
writer	Reviews_from_critics	writer	Reviews_from_critics

Names Table

Original		Refined	
imdb_name_id	date_of_death	imdb_name_id	
name	place_of_death	name	
birth_name	reason_of_death	birth_name	
height	spouses_string		
bio	spouses	bio	spouses
birth_details	divorces		divorces
date_of_birth	spouses_with_children		spouses_with_children
place_of_birth	children		children
death_details			

Ratings Table

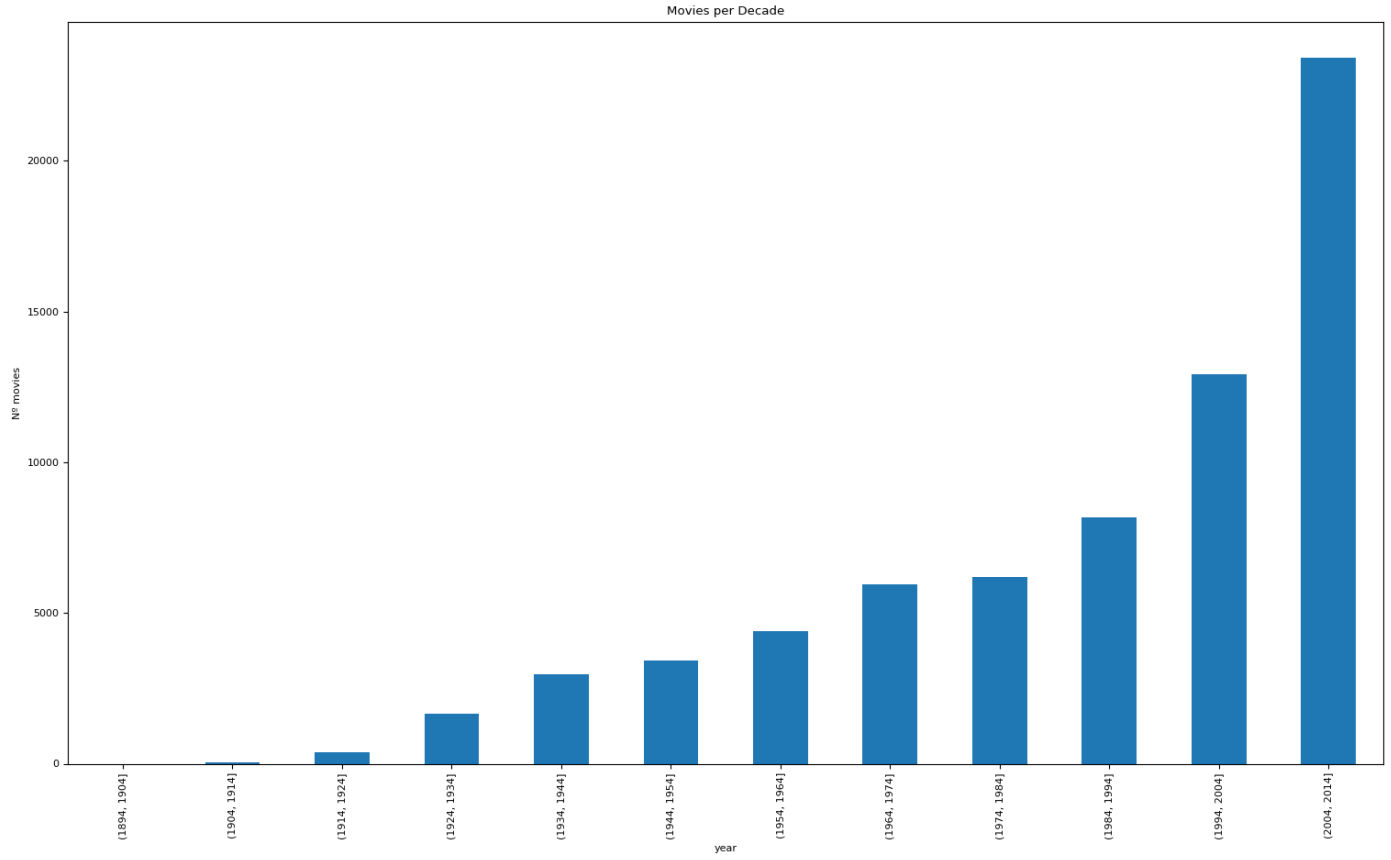
Original		Refined	
imdb_title_id	female_avg_vote	imdb_title_id	
weighted_average_vote	top1000_voters_rating	weighted_average_vote	
total_votes	top1000_voters_votes	total_votes	
mean_vote	us_voters_rating	mean_vote	
median_vote	us_voters_votes	median_vote	
votes_1…10	non_us_voters_rating	votes_1…10	
allgenders_avg_vote	non_us_voters_votes		
male_avg_vote			

Title Principals Table

Original	Refined
imdb_title_id	imdb_title_id
ordering	ordering
imdb_name_id	imdb_name_id
category	category
job	
characters	characters

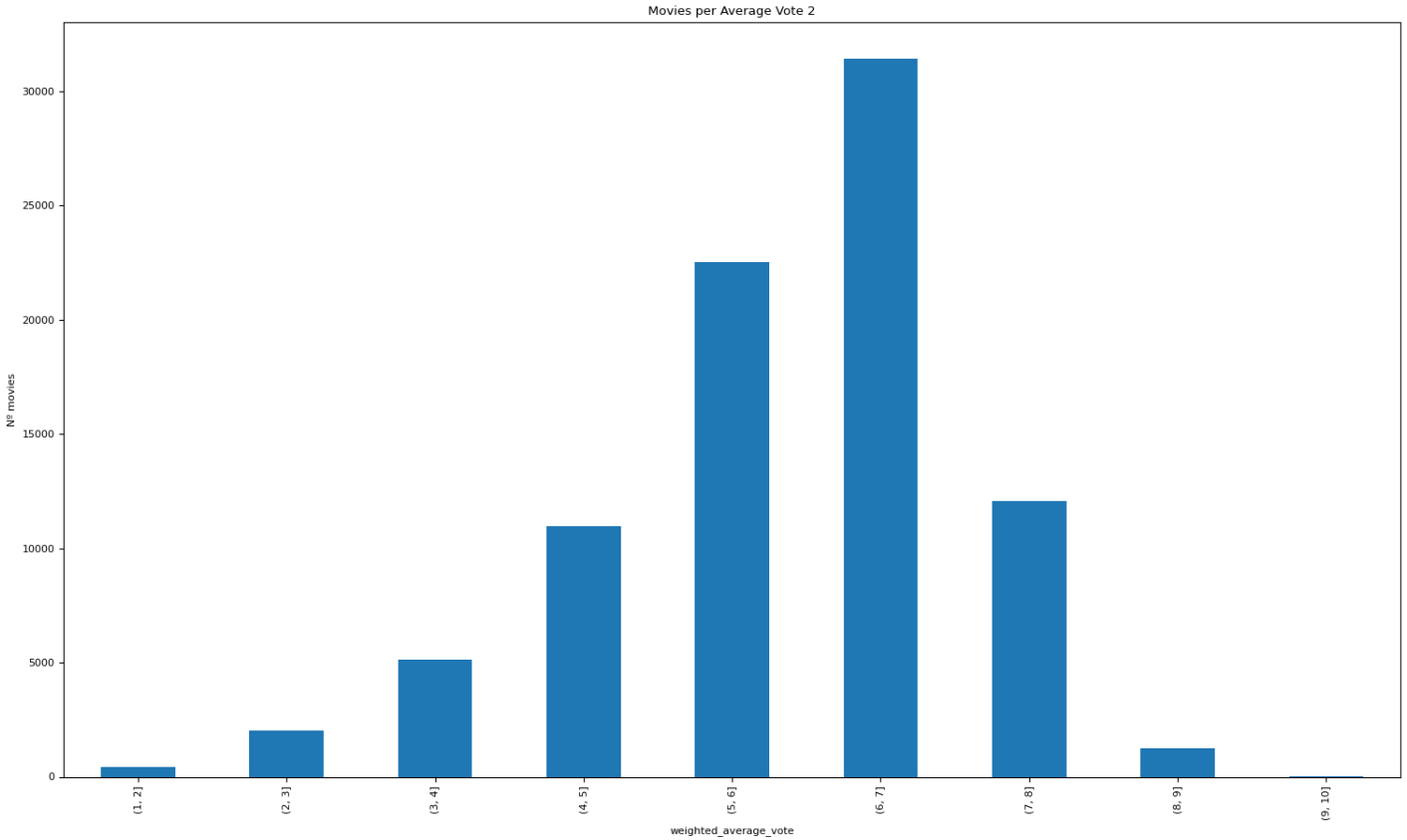
Data Analysis

Movies Grouped by Decade



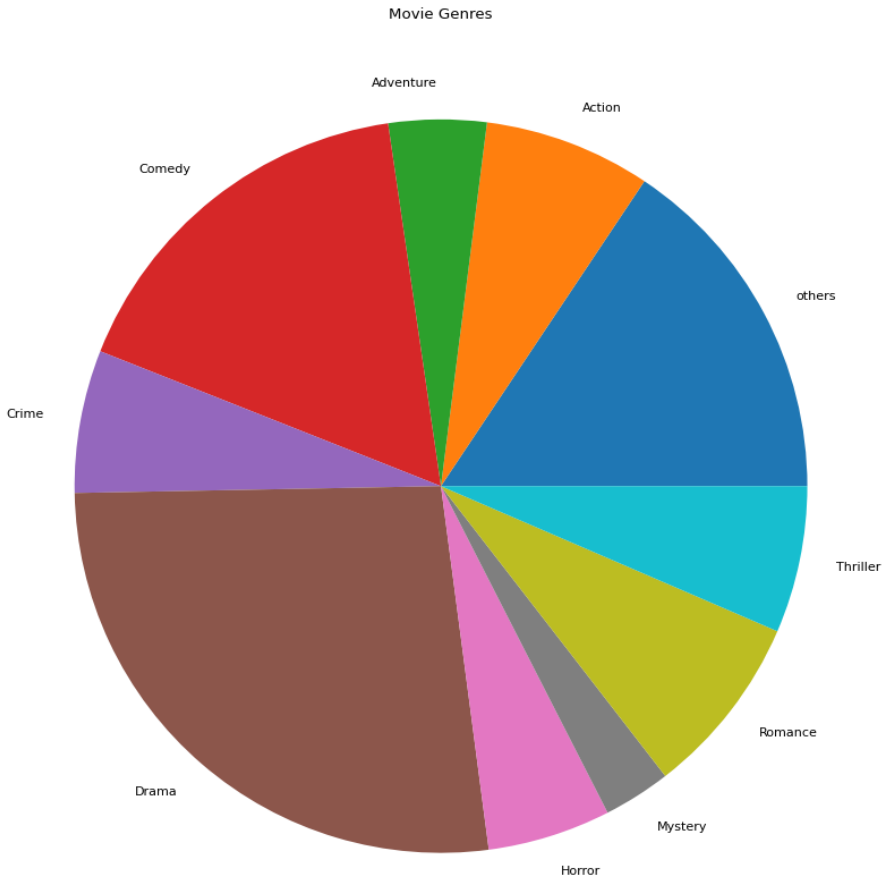
Data Analysis

Movies Grouped by Rating

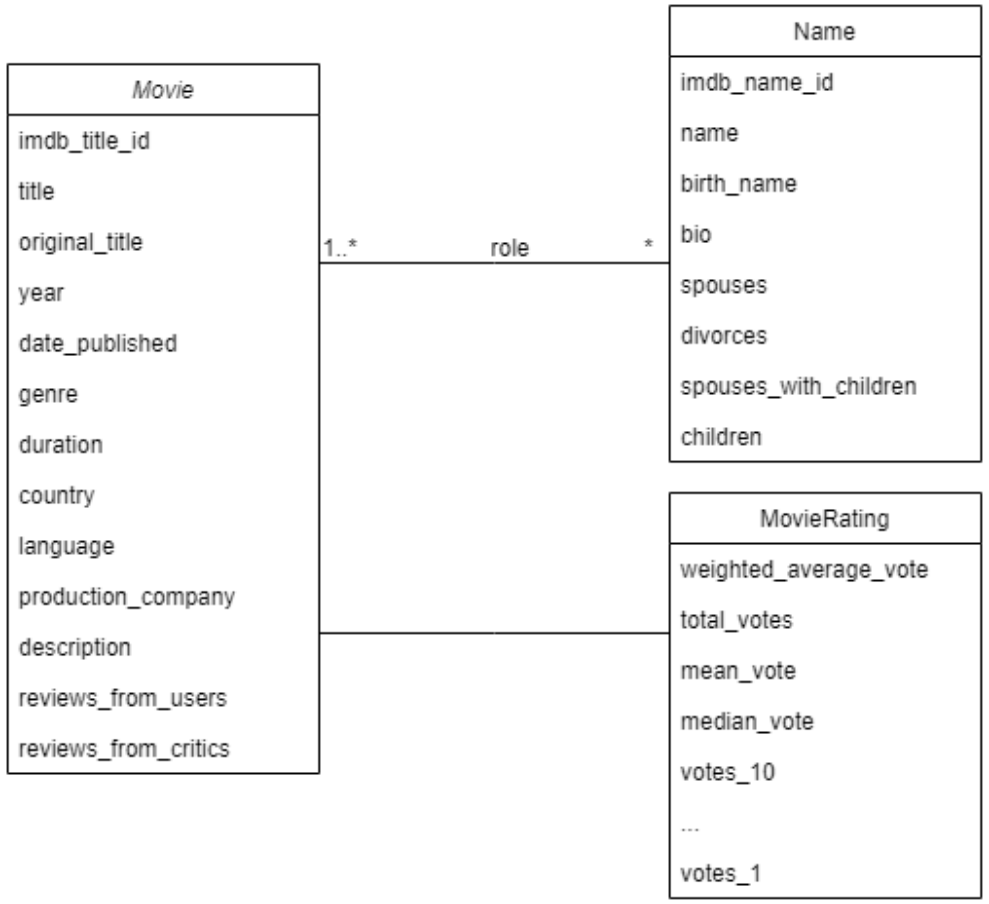


Data Analysis

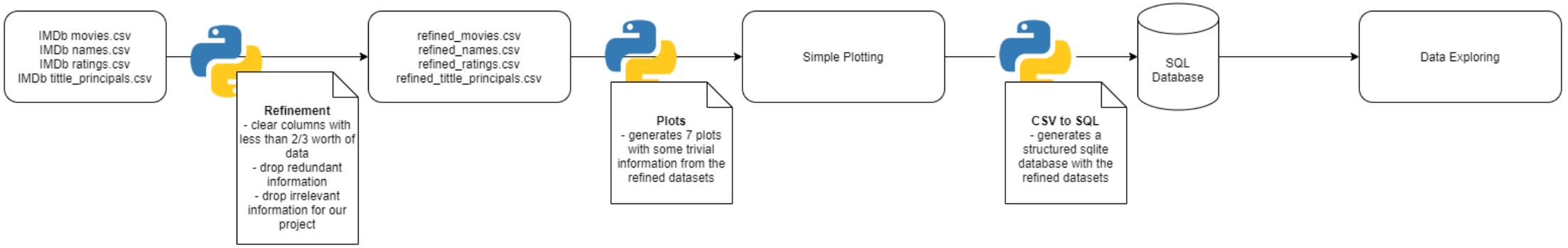
Movie Genres Count



Database


Movie
imdb_title_id
title
original_title
year
date_published
genre
duration
country
language
production_company
description
reviews_from_users
reviews_from_critics
1..*
role
*
Name
imdb_name_id
name
birth_name
bio
spouses
divorces
spouses_with_children
children
MovieRating
weighted_average_vote
total_votes
mean_vote
median_vote
votes_10
...
votes_1

Pipeline


IMDb movies.csv
IMDb names.csv
IMDb ratings.csv
IMDb tittle_principals.csv
Refinement
- clear columns with less than 2/3 worth of data
- drop redundant information
- drop irrelevant information for our project
refined_movies.csv
refined_names.csv
refined_ratings.csv
refined_tittle_principals.csv
Plots
- generates 7 plots with some trivial information from the refined datasets
Simple Plotting
CSV to SQL
- generates a structured sqlite database with the refined datasets
SQL Database
Data Exploring