Home TV Shows Movies News & Popular My list Watchlist

DATA PREPARATION
MILESTONE 1

PAULO RIBEIRO, UP201806505
PEDRO FERREIRA, UP201806506
PEDRO PONTE, UP201809694

TABLE OF CONTENTS

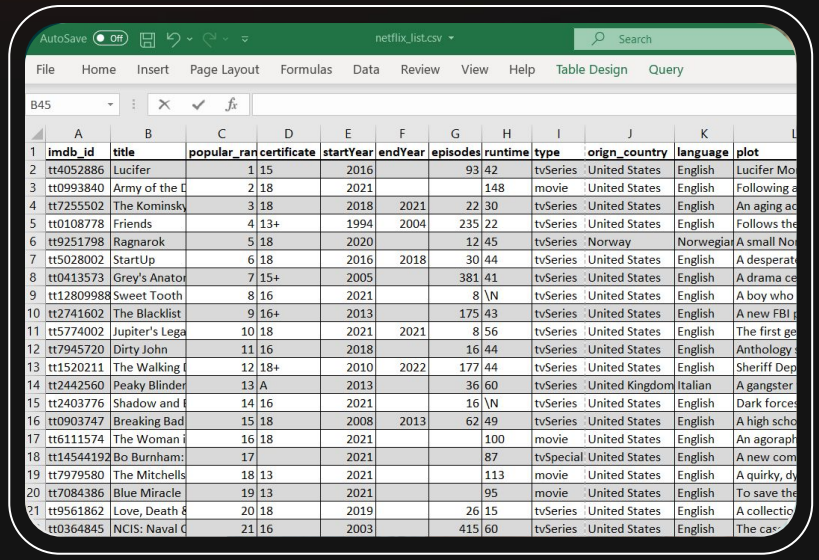
Home TV Shows Movies News & Popular My list Watchlist

01 Original Dataset

Netflix dataset characterisation

ORIGINAL DATASET

- 7008 rows and 19 columns
- 6.29 Mb of space
- Missing or Invalid values



	A	B	C	D	E	F	G	H	I	J	K	L
1	imdb_id	title	popular_ran	certificate	startYear	endYear	episodes	runtime	type	origin_country	language	plot
2	tt4052886	Lucifer	1	15	2016		93	42	tvSeries	United States	English	Lucifer Mo
3	tt0993840	Army of the D	2	18	2021			148	movie	United States	English	Following a
4	tt7255502	The Kominsky	3	18	2018	2021	22	30	tvSeries	United States	English	An aging ac
5	tt0108778	Friends	4	13+	1994	2004	235	22	tvSeries	United States	English	Follows the
6	tt9251798	Ragnarok	5	18	2020		12	45	tvSeries	Norway	Norwegian	A small No
7	tt5028002	StartUp	6	18	2016	2018	30	44	tvSeries	United States	English	A desperate
8	tt0413573	Grey's Anator	7	15+	2005		381	41	tvSeries	United States	English	A drama ce
9	tt12809988	Sweet Tooth	8	16	2021		8	\N	tvSeries	United States	English	A boy who
10	tt2741602	The Blacklist	9	16+	2013		175	43	tvSeries	United States	English	A new FBI p
11	tt5774002	Jupiter's Lega	10	18	2021	2021	8	56	tvSeries	United States	English	The first ge
12	tt7945720	Dirty John	11	16	2018		16	44	tvSeries	United States	English	Anthology s
13	tt1520211	The Walking D	12	18+	2010	2022	177	44	tvSeries	United States	English	Sheriff Dep
14	tt2442560	Peaky Blinder	13	A	2013		36	60	tvSeries	United Kingdom	Italian	A gangster
15	tt2403776	Shadow and B	14	16	2021		16	\N	tvSeries	United States	English	Dark forces
16	tt0903747	Breaking Bad	15	18	2008	2013	62	49	tvSeries	United States	English	A high scho
17	tt6111574	The Woman i	16	18	2021			100	movie	United States	English	An agoraph
18	tt14544192	Bo Burnham:	17		2021			87	tvSpecial	United States	English	A new com
19	tt7979580	The Mitchells	18	13	2021			113	movie	United States	English	A quirky, dy
20	tt7084386	Blue Miracle	19	13	2021			95	movie	United States	English	To save the
21	tt9561862	Love, Death &	20	18	2019		26	15	tvSeries	United States	English	A collectio
	tt0364845	NCIS: Naval C	21	16	2003		415	60	tvSeries	United States	English	The cas

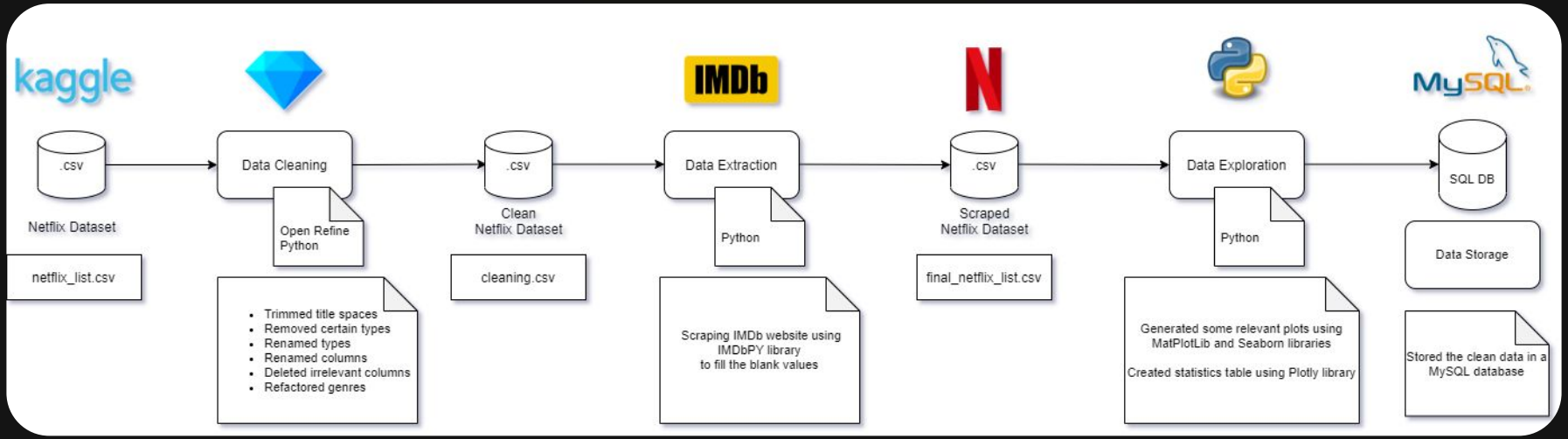
FINAL DATASET

- 6216 rows and 17 columns
- 6.02 Mb of space
- Fixed and complete dataset

02 Pipeline Diagram

Data Processing flow

PIPELINE DIAGRAM

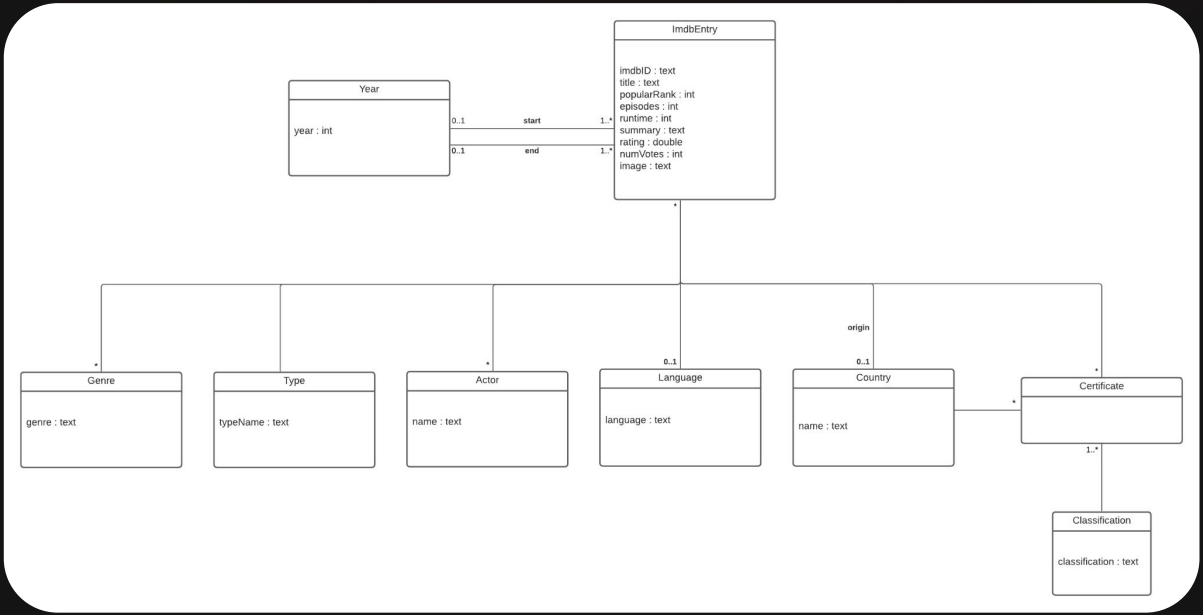


Home TV Shows Movies News & Popular My list Watchlist

03 UML Diagram

Conceptual model for the data domain

UML Diagram


ImdbEntry
imdbID : text
title : text
popularRank : int
episodes : int
runtime : int
summary : text
rating : double
numVotes : int
image : text
Year
year : int
start
end
0..1
1..*
origin
Genre
genre : text
Type
typeName : text
Actor
name : text
Language
language : text
Country
name : text
Certificate
Classification
classification : text

Home TV Shows Movies News & Popular My list Watchlist

04 Data Processing

DATA PROCESSING

CLEANING

- Trimmed spaces
- Removed types
- Renamed types
- Renamed columns
- Deleted columns
- Refactored genres

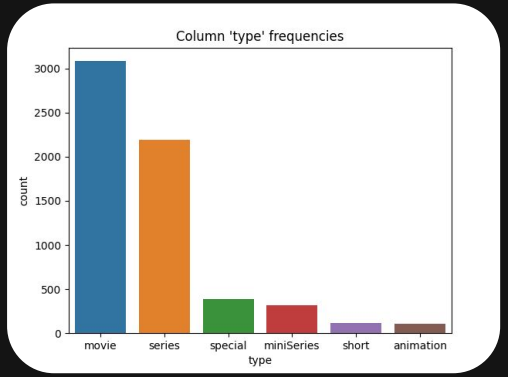
EXTRACTION

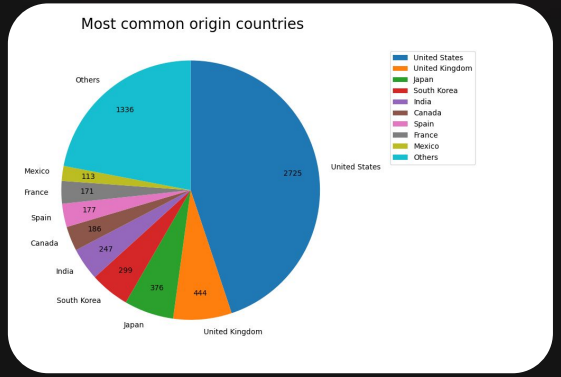
- IMDb Scraping using IMDbPY library

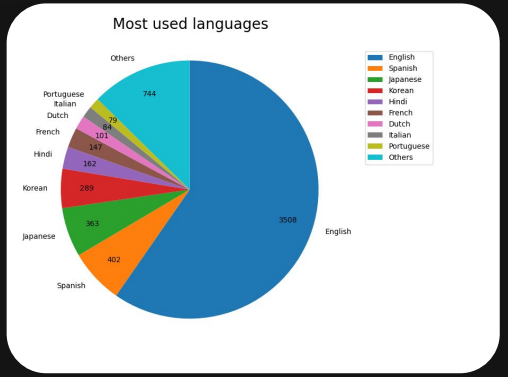
Home TV Shows Movies News & Popular My list Watchlist

05 Data Exploration

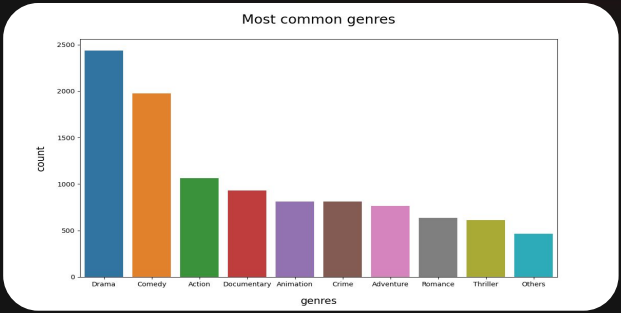
Data Exploration

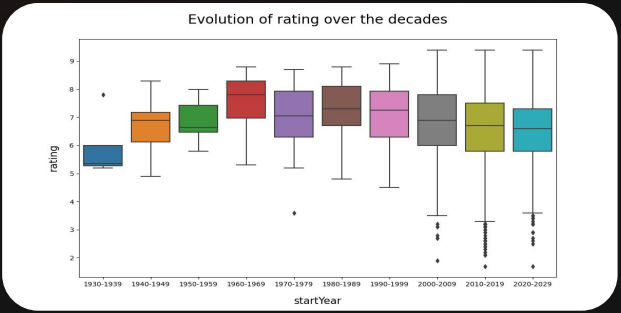


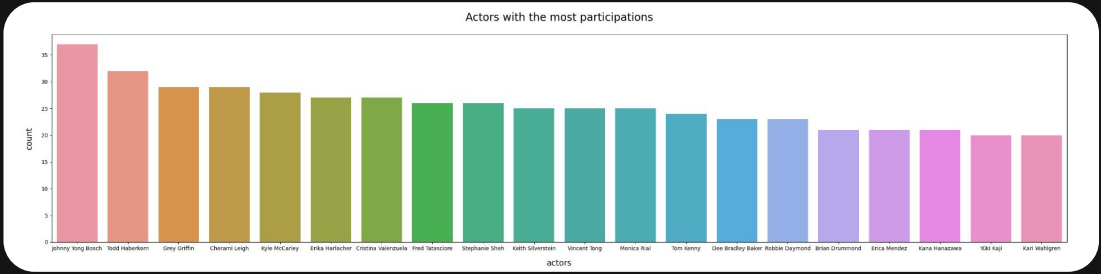




Data Exploration 2







Data Exploration 3

Attribute	Value
Min Rating	1.7
Max Rating	9.4
Mean Rating	6.598006524102936
Mean MiniSeries Runtime	146.73478260869564
Mean Movies Runtime	99.23572744014733
Mean Series Runtime	45.62475181998676
Mean Shorts Runtime	25.2183908045977
Mean Special Runtime	65.68115942028986
Total Genres	27
Total Languages	70
Total Origin Countries	82
Total Actors	56207

06 Future Work

FUTURE WORK & IMPROVEMENTS

DATA STORAGE

Storing the data in a MySQL database

INFORMATION RETRIEVAL

Use of an information retrieval tool to do exploration with free-text queries

SEARCH

Use of features and techniques with the goal of improving the quality of the search results

FINAL RESULT

Our goal is an elaborate information search and retrieval system

THANKS!

Do you have any questions?

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon and infographics & images by Freepik