Analysis IMDB’s Datasets

Recommender system

Introduction

People loves movies, romantic, drama, sci-fi whatever genre you like is in the movie theater but not all of them are a success, they are used to have a rating provided by users and sometimes these are not nice, based on data we have in IMDB, companies can analyze this rate and find the best team in order to write, produce and be a box-office hit.

- IMDB has multiple dataset available in its site https://www.imdb.com/interfaces.
- In these datasets we can find variables as an actors, writers, producers and so on, they are related to a film, a category, a crew and a rate.
- What can we discover?, we can trace who is best writer, composer, ost based on the relation between the film and rate, besides this data can help us to create a recommender system.

Description Data

title.akas.tsv.gz - Contains the following information for titles:
•titleId (string) - a tconst, an alphanumeric unique identifier of the title
•ordering (integer) – a number to uniquely identify rows for a given titleId
•title (string) – the localized title
•region (string) - the region for this version of the title
•language (string) - the language of the title
•types (array) - Enumerated set of attributes for this alternative title. One or more of the following: "alternative", "dvd", "festival", "tv", "video", "working", "original", "imdbDisplay". New values may be added in the future without warning
•attributes (array) - Additional terms to describe this alternative title, not enumerated
•isOriginalTitle (boolean) – 0: not original title; 1: original title
•**title.basics.tsv.gz** - Contains the following information for titles:tconst (string) - alphanumeric unique identifier of the title
•titleType (string) – the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc)
•primaryTitle (string) – the more popular title / the title used by the filmmakers on promotional materials at the point of release
•originalTitle (string) - original title, in the original language
•isAdult (boolean) - 0: non-adult title; 1: adult title
•startYear (YYYY) – represents the release year of a title. In the case of TV Series, it is the series start year
•endYear (YYYY) – TV Series end year. '\N' for all other title types
•runtimeMinutes – primary runtime of the title, in minutes
•genres (string array) – includes up to three genres associated with the title

•**title.crew.tsv.gz** – Contains the director and writer information for all the titles in IMDb. Fields include:tconst (string) - alphanumeric unique identifier of the title
•directors (array of nconsts) - director(s) of the given title
•writers (array of nconsts) – writer(s) of the given title
•**title.episode.tsv.gz** – Contains the tv episode information. Fields include:tconst (string) - alphanumeric identifier of episode
•parentTconst (string) - alphanumeric identifier of the parent TV Series
•seasonNumber (integer) – season number the episode belongs to
•episodeNumber (integer) – episode number of the tconst in the TV series

•**title.principals.tsv.gz** – Contains the principal cast/crew for titlestconst (string) - alphanumeric unique identifier of the title
•ordering (integer) – a number to uniquely identify rows for a given titleId
•nconst (string) - alphanumeric unique identifier of the name/person
•category (string) - the category of job that person was in
•job (string) - the specific job title if applicable, else '\N'
•characters (string) - the name of the character played if applicable, else '\N'
•**title.ratings.tsv.gz** – Contains the IMDb rating and votes information for titlestconst (string) - alphanumeric unique identifier of the title
•averageRating – weighted average of all the individual user ratings
•numVotes - number of votes the title has received
•**name.basics.tsv.gz** – Contains the following information for names:nconst (string) - alphanumeric unique identifier of the name/person
•primaryName (string)– name by which the person is most often credited
•birthYear – in YYYY format
•deathYear – in YYYY format if applicable, else '\N'
•primaryProfession (array of strings)– the top-3 professions of the person
•knownForTitles (array of tconsts) – titles the person is known for

Datasets

File name	Column Name	Split Flag
name.basics.tsv	nconst	
name.basics.tsv	primaryname	
name.basics.tsv	birthyear	
name.basics.tsv	deathyear	
name.basics.tsv	primaryprofession	Yes
name.basics.tsv	knownfortittles	Yes

File name	Column Name	Split Flag
title.akas.tsv	titleId	
title.akas.tsv	ordering	
title.akas.tsv	title	
title.akas.tsv	region	
title.akas.tsv	language	
title.akas.tsv	types	
title.akas.tsv	attributes	
title.akas.tsv	isOriginalTitle	

File Name	Column Name	Split Flag
title.principals.tsv	tconst	
title.principals.tsv	ordering	
title.principals.tsv	nconst	
title.principals.tsv	category	
title.principals.tsv	job	
title.principals.tsv	characters	

File name	Column Name	Split Flag
title.basics.tsv	tconst	
title.basics.tsv	titleType	
title.basics.tsv	primaryTitle	
title.basics.tsv	originalTitle	
title.basics.tsv	isAdult	
title.basics.tsv	startYear	
title.basics.tsv	endYear	
title.basics.tsv	runtimeMinutes	
title.basics.tsv	genres	Yes

File Name	Column Name	Split Flag
title.ratings.tsv	tconst	
title.ratings.tsv	averageRating	
title.ratings.tsv	numVotes	

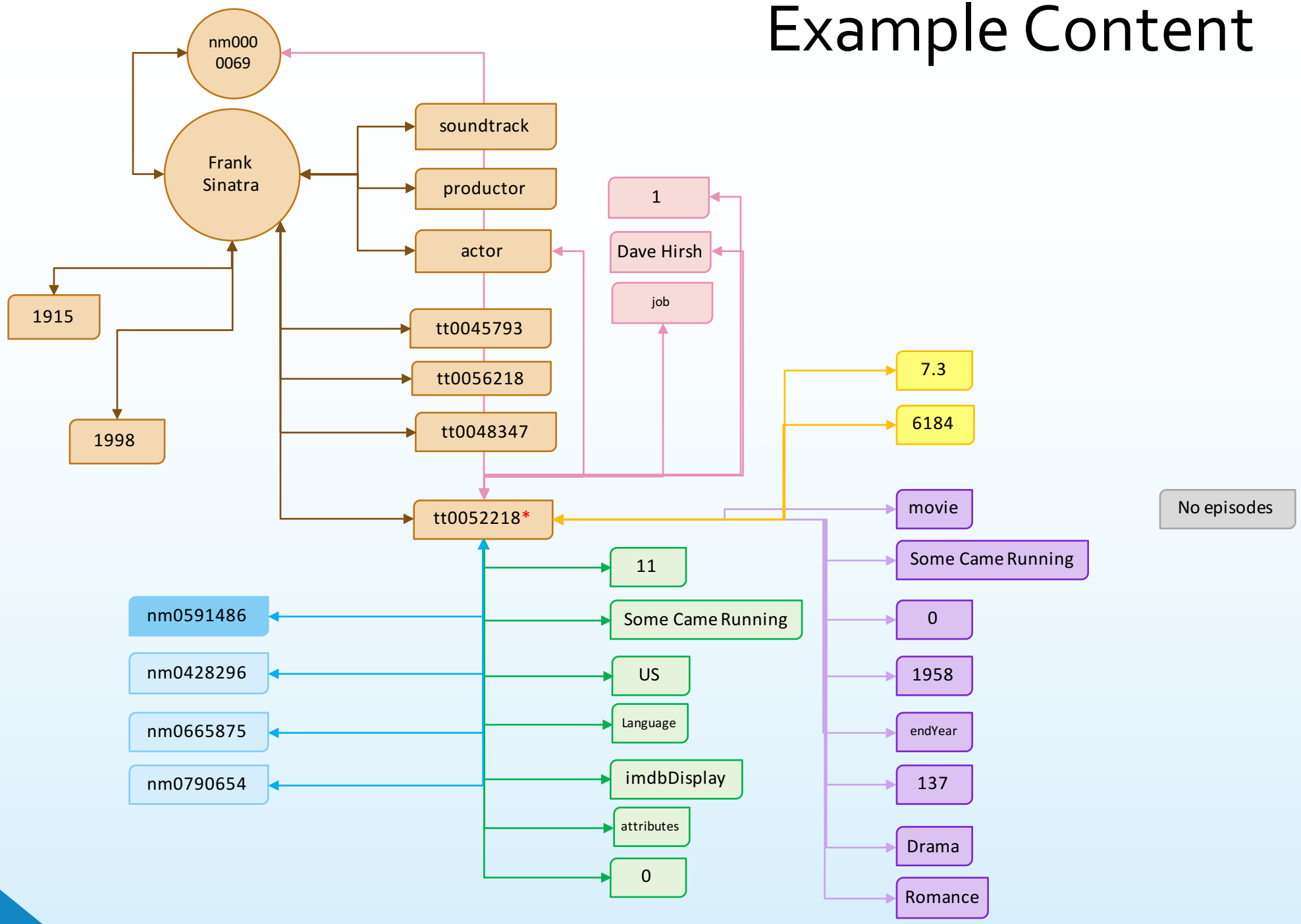
File Name	Column Name	Split Flag
title.crew.tsv	tconst	
title.crew.tsv	directors	Yes
title.crew.tsv	writers	

File Name	Column Name	Split Flag
title.episode.tsv	tconst	
title.episode.tsv	parentTconst	
title.episode.tsv	seasonNumber	
title.episode.tsv	episodeNumber	

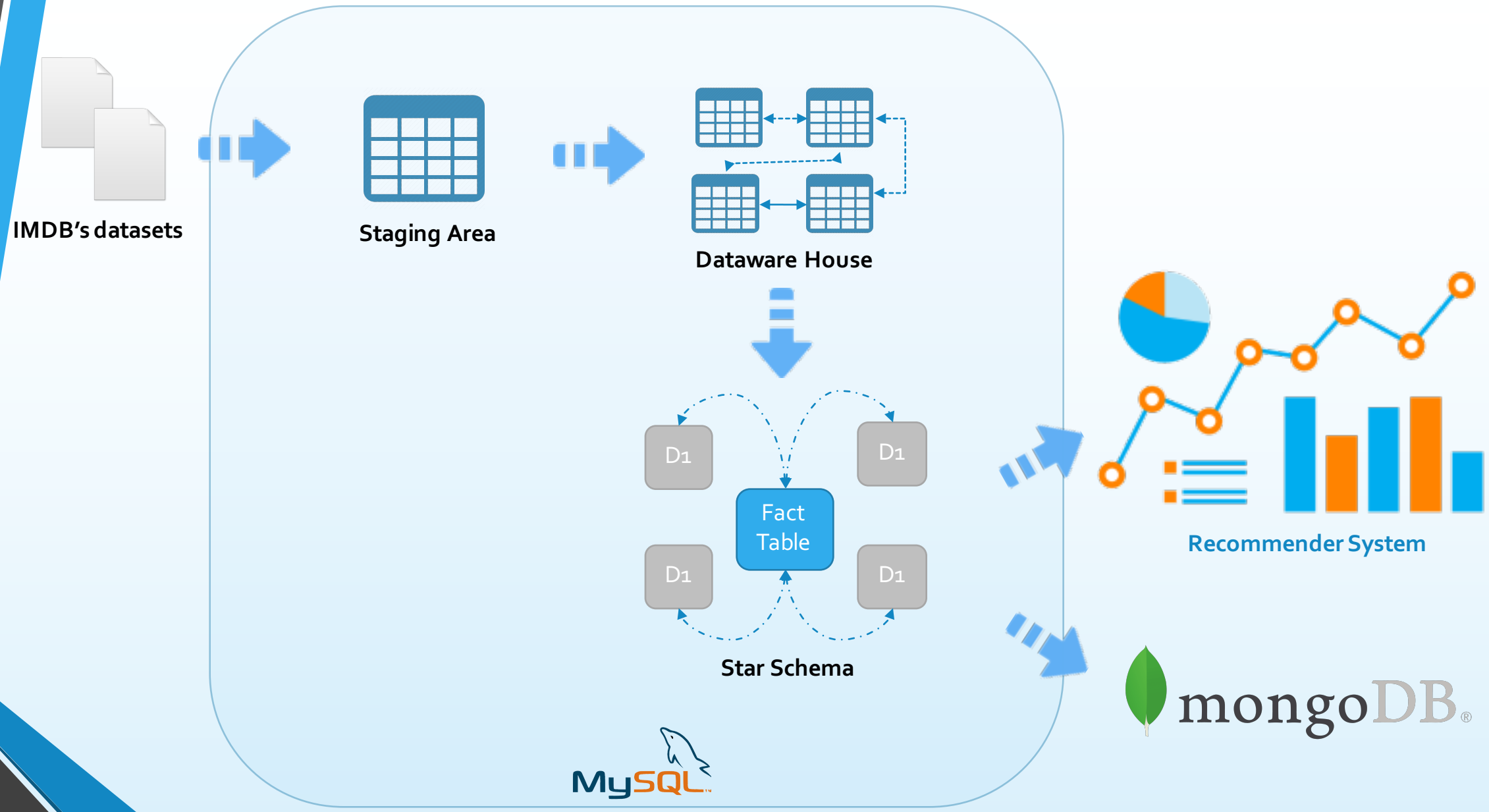
Missing Data

- Historical data
- User information (PII, Top critic, Type)
- Date release, date rating (1-10)
- Metascore
- Geography data

Example Content



Prototype


IMDB's datasets
Staging Area
Dataware House
D1
D1
Fact Table
D1
D1
Star Schema
MySQL
Recommender System
mongoDB

Cleaning Data

- name.basics.tsv.gz
- title.akas.tsv.gz
- title.basics.tsv.gz
- title.crew.tsv.gz
- title.episode.tsv.gz
- title.principals.tsv.gz
- title.ratings.tsv.gz

Get Started | Workspace Trust | name.basics.tsv

Users > OmarCr > Documents > BI > BEDU > Proyecto > IMDB_Sources > name.basics.tsv

1	hconst	primaryName	birthYear	deathYear	primaryProfession	knownForTitles
2	nm0000001	Fred Astaire	1899	1987	soundtrack,actor,miscellaneous	tt0050419,tt0031983,tt0053137,tt0072308
3	nm0000002	Lauren Bacall	1924	2014	actress,soundtrack	tt0037382,tt0038355,tt0117057,tt0071877
4	nm0000003	Brigitte Bardot	1934	\N	actress,soundtrack,music_department	tt0054452,tt0049189,tt0056404,tt0057345
5	nm0000004	John Belushi	1949	1982	actor,soundtrack,writer	tt0072562,tt0080455,tt0078723,tt0077975
6	nm0000005	Ingmar Bergman	1918	2007	writer,director,actor	tt0060827,tt0083922,tt0050976,tt0050986
7	nm0000006	Ingrid Bergman	1915	1982	actress,soundtrack,producer	tt0077711,tt0034583,tt0038109,tt0036855

```
sed 's/ /"|"/g' title.crew.tsv > title_crew.csv
sed 's/.*/"&"/' title_crew.csv > title_crew_final.csv
```

```
sed 's/ /"|"/g' title.crew.tsv  && sed 's/.*/"&"/' title_crew.csv  > title_crew.csv
sed 's/ /"|"/g' name.basics.tsv >name_basics.csv && sed 's/.*/"&"/' name_basics.csv >name_basics_final.csv
sed 's/ /"|"/g' title.akas.tsv > title_akas.csv && sed 's/.*/"&"/' title_akas.csv >title_akas_final.csv
sed 's/ /"|"/g' title.basics.tsv >title_basics.csv && sed 's/.*/"&"/' title_basics.csv >title_basics_final.csv
sed 's/ /"|"/g' title.episode.tsv > title_episode.csv && sed 's/.*/"&"/' title_episode.csv >title_episode_final.csv
sed 's/ /"|"/g' title.principals.tsv > title_principals.csv && sed 's/.*/"&"/' title_principals.csv >title_principals_final.csv
sed 's/ /"|"/g' title.ratings.tsv > title_ratings.csv && sed 's/.*/"&"/' title_ratings.csv >title_ratings_final.csv
```

```
/Users/OmarCr/Documents/BI/BEDU/IMDB

alias mysql=/usr/local/mysql/bin/mysql
mysql --user=root  -p
mysql -u root -p --local-infile  project_imdb

SHOW VARIABLES LIKE 'local_infile';
SET GLOBAL local_infile = 1;
```

```
load data local infile '/Users/OmarCr/Documents/BI/BEDU/IMDB/title_crew_final.csv' into table title_crew  fields terminated by '|' optionally enclosed by "" lines terminated by '\n' ignore 1 lines;
load data local infile '/Users/OmarCr/Documents/BI/BEDU/IMDB/title_ratings_final.csv' into table title_ratings  fields terminated by '|' optionally enclosed by "" lines terminated by '\n' ignore 1 lines;
load data local infile '/Users/OmarCr/Documents/BI/BEDU/IMDB/title_episode_final.csv' into table title_episode  fields terminated by '|' optionally enclosed by "" lines terminated by '\n' ignore 1 lines;
load data local infile '/Users/OmarCr/Documents/BI/BEDU/IMDB/title_akas_final.csv' into table title_akas  fields terminated by '|' optionally enclosed by "" lines terminated by '\n' ignore 1 lines;
load data local infile '/Users/OmarCr/Documents/BI/BEDU/IMDB/name_basics.csv' into table name_basics  fields terminated by '|' lines terminated by '\n' ignore 1 lines;
load data local infile '/Users/OmarCr/Documents/BI/BEDU/IMDB/title_principals_final.csv' into table title_principal  fields terminated by '|' optionally enclosed by "" lines terminated by '\n' ignore 1 lines;
load data local infile '/Users/OmarCr/Documents/BI/BEDU/IMDB/title_basics_final.csv' into table title_basics  fields terminated by '|' optionally enclosed by "" lines terminated by '\n' ignore 1 lines;
load data local infile '/Users/OmarCr/Documents/BI/BEDU/IMDB/title_basics.csv' into table title_basics  fields terminated by '|'  lines terminated by '\n' ignore 1 lines;
load data local infile '/Users/OmarCr/Documents/BI/BEDU/IMDB/covid19countryinfo.csv' into table country  fields terminated by ',' optionally enclosed
```

QA - Staging

```
(base) MacBook-Pro-de-Omar:IMDB OmarCr$ wc -l  title.crew.tsv
 8331071 title.crew.tsv
(base) MacBook-Pro-de-Omar:IMDB OmarCr$ wc -l title.ratings.tsv
 1197095 title.ratings.tsv
(base) MacBook-Pro-de-Omar:IMDB OmarCr$ wc -l title.episode.tsv
 6111719 title.episode.tsv
(base) MacBook-Pro-de-Omar:IMDB OmarCr$ wc -l title.akas.tsv
 29578071 title.akas.tsv
(base) MacBook-Pro-de-Omar:IMDB OmarCr$ wc -l name.basics.tsv
 11323485 name.basics.tsv
(base) MacBook-Pro-de-Omar:IMDB OmarCr$ wc -l title.basics.tsv
 8331071 title.basics.tsv
(base) MacBook-Pro-de-Omar:IMDB OmarCr$ wc -l title.principals.tsv
 47269857 title.principals.tsv
```

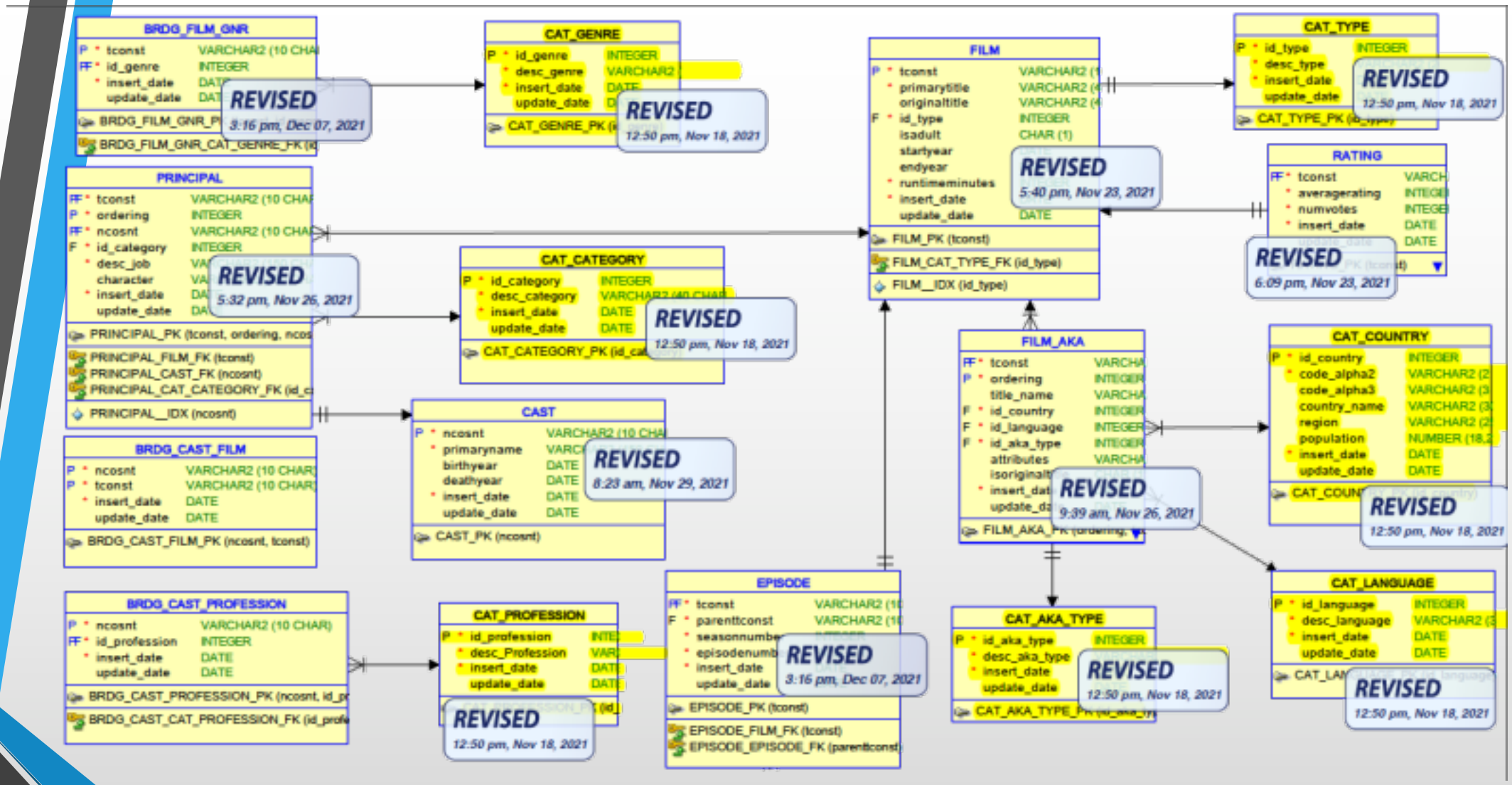
```
mysql> select count(*) from title_crew;
+----------+
| count(*) |
+----------+
| 8331070 |
+
1 row in set (25.15 sec)

mysql> select count(*) from title_episode;
+----------+
| count(*) |
+----------+
| 6111718 |
+
1 row in set (19.87 sec)
```

```
mysql> select count(*) from title_ratings;
+----------+
| count(*) |
+----------+
| 1197094 |
+----------+
1 row in set (1.98 sec)

mysql> select count(*) from title_akas;
+----------+
| count(*) |
+----------+
| 29578070 |
+----------+
1 row in set (2 min 36.65 sec)

mysql> select count(*) from name_basics;
+----------+
| count(*) |
+----------+
| 11323484 |
+----------+
1 row in set (1 min 9.57 sec)
```

```
mysql> select count(*) from title_basics;
+----------+
| count(*) |
+----------+
| 8331070 |
+----------+
1 row in set (44.84 sec)
```

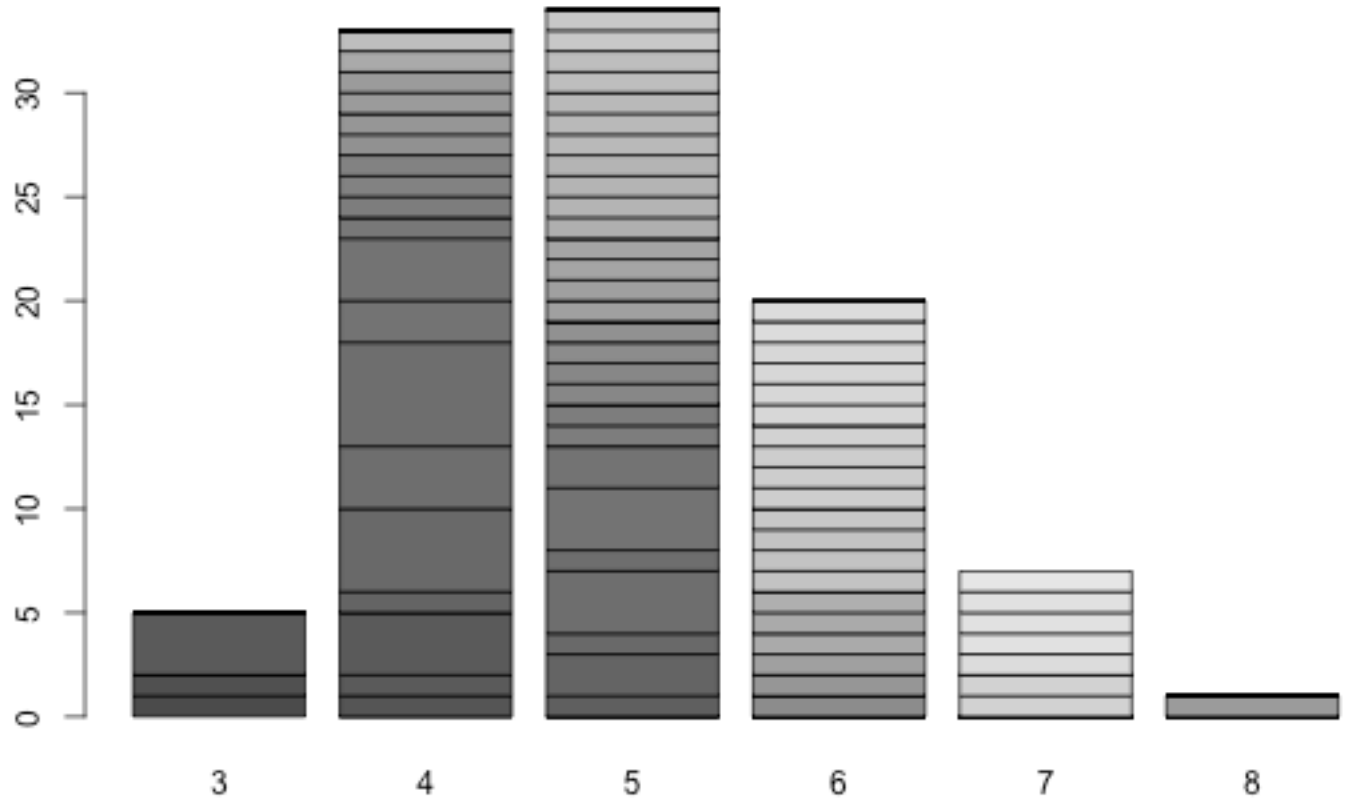
```
mysql> select count(*) from title_principal;
+----------+
| count(*) |
+----------+
| 47269856 |
+----------+
1 row in set (2 min 35.52 sec)
```

Prototype - DER



Data Analysis

Rating | Barras | Datos Complementarios



IMDB Rating

Introduzca un valor:

Número de votos

Rating | Barras | Datos Complementarios

desc_type	averagerating	numvotes
movie	10	925603781.00
tvSeries	10	101362685.00
tvEpisode	10	94547294.00
tvMiniSeries	10	12609570.00
tvMovie	10	10590395.00
video	10	9823735.00
short	10	9539802.00
videoGame	10	4341576.00
tvSpecial	10	1912642.00

Rating | Barras | Datos Complementarios

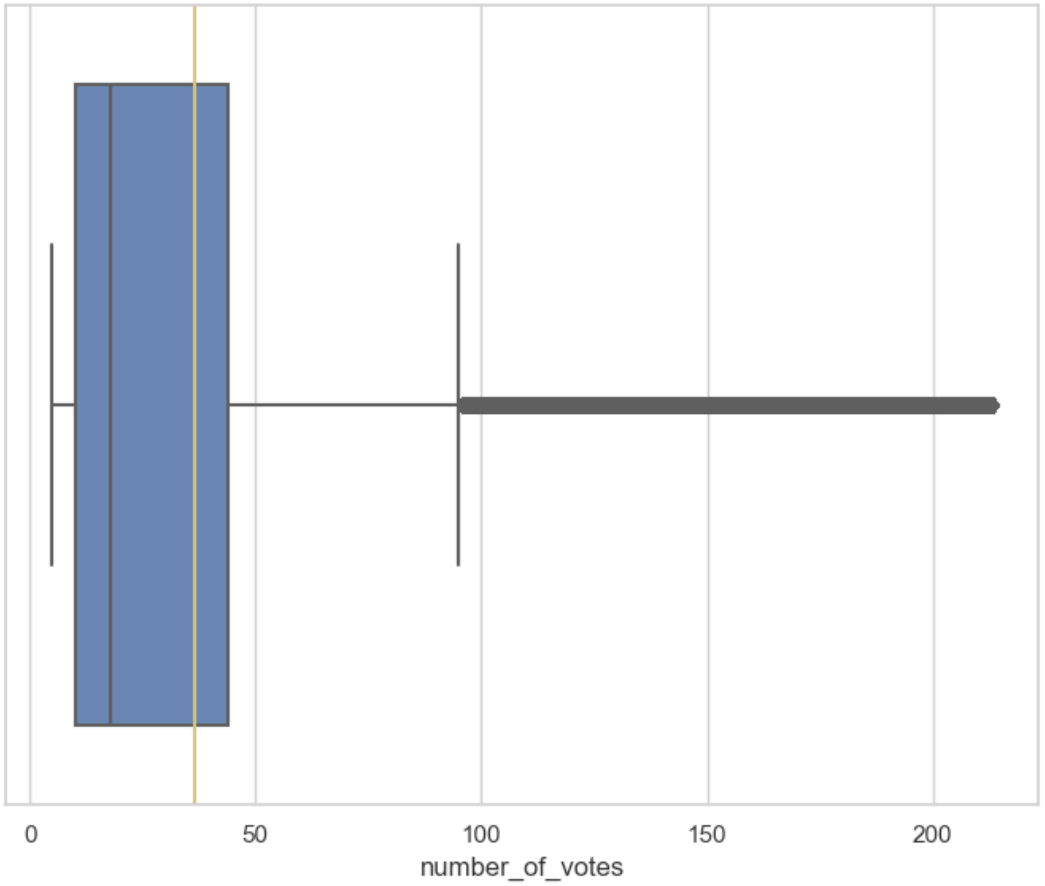
tconst	primarytitle	desc_type	averagerating	numvotes
tt0000001	Carmencita	short	6	1828
tt0000002	Le clown et ses chiens	short	6	236
tt0000003	Pauvre Pierrot	short	7	1588
tt0000004	Un bon bock	short	6	153
tt0000005	Blacksmith Scene	short	6	2405
tt0000006	Chinese Opium Den	short	5	157
tt0000007	Corbett and Courtney Before the Kinetograph	short	5	753
tt0000008	Edison Kinetoscopic Record of a Sneeze	short	6	1981
tt0000009	Miss Jerry	short	6	190
tt0000010	Leaving the Factory	short	7	6593

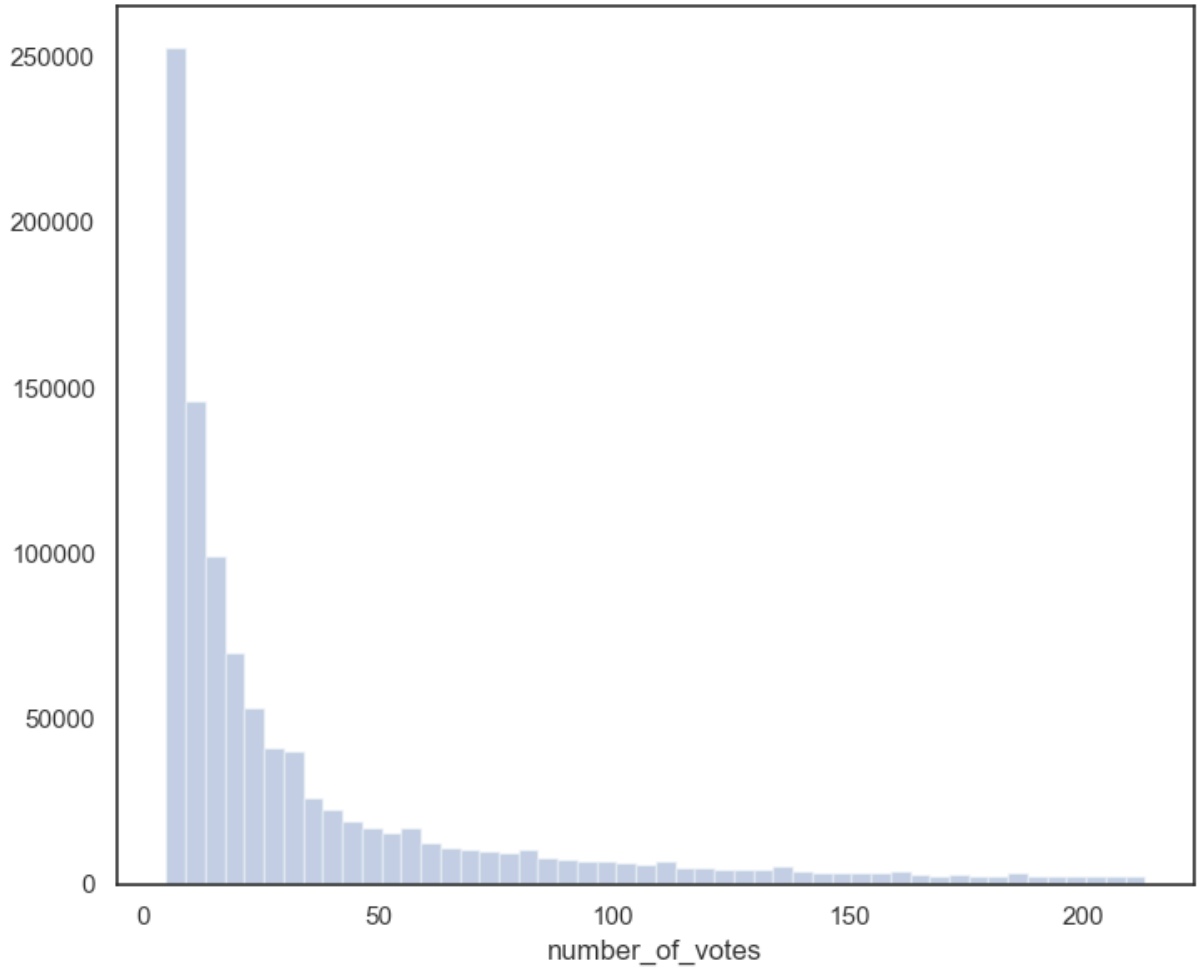
Data Statistics (votes)

Min Value: 5
Maximum Value : 2476912
Threshold: 2476907

Mean: 977.97
Median: 24.0
Kurtosis: 3.72
Skewness: 2.03

Percentil 10: 7.0
Percentil 25: 11.0
Percentil 75: 92.0
Percentil 90: 426.0



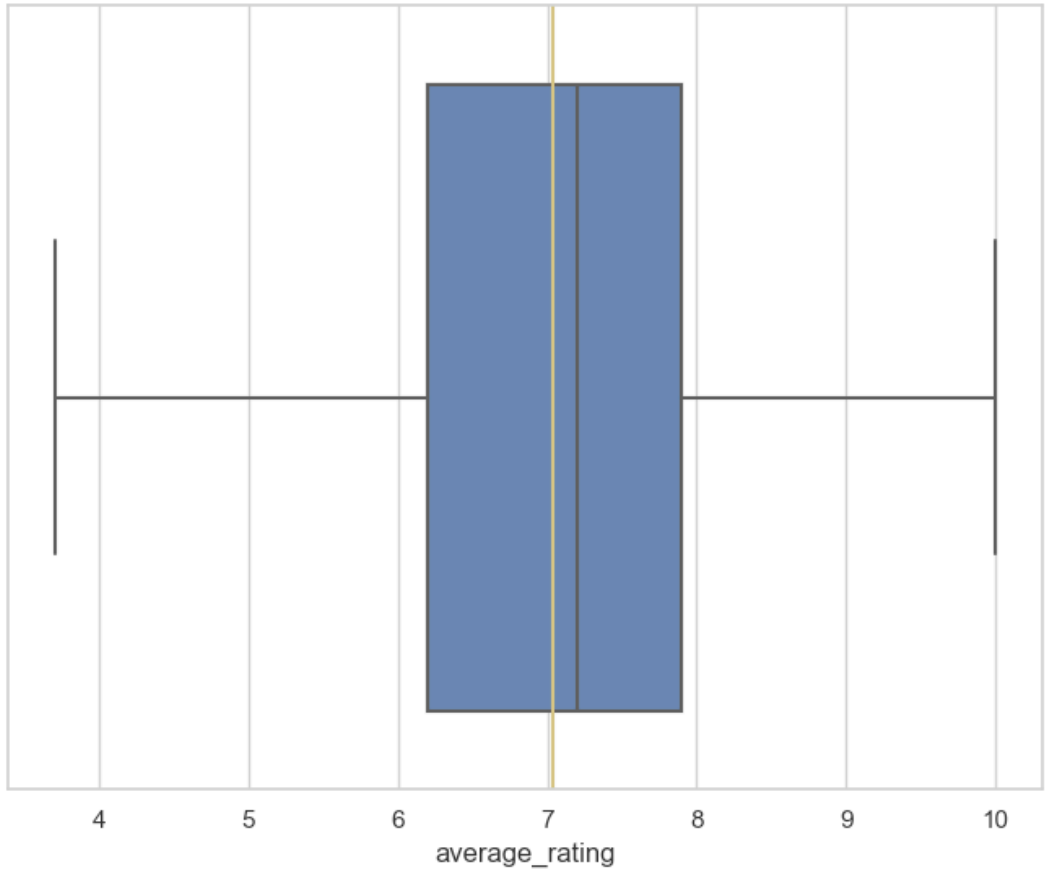


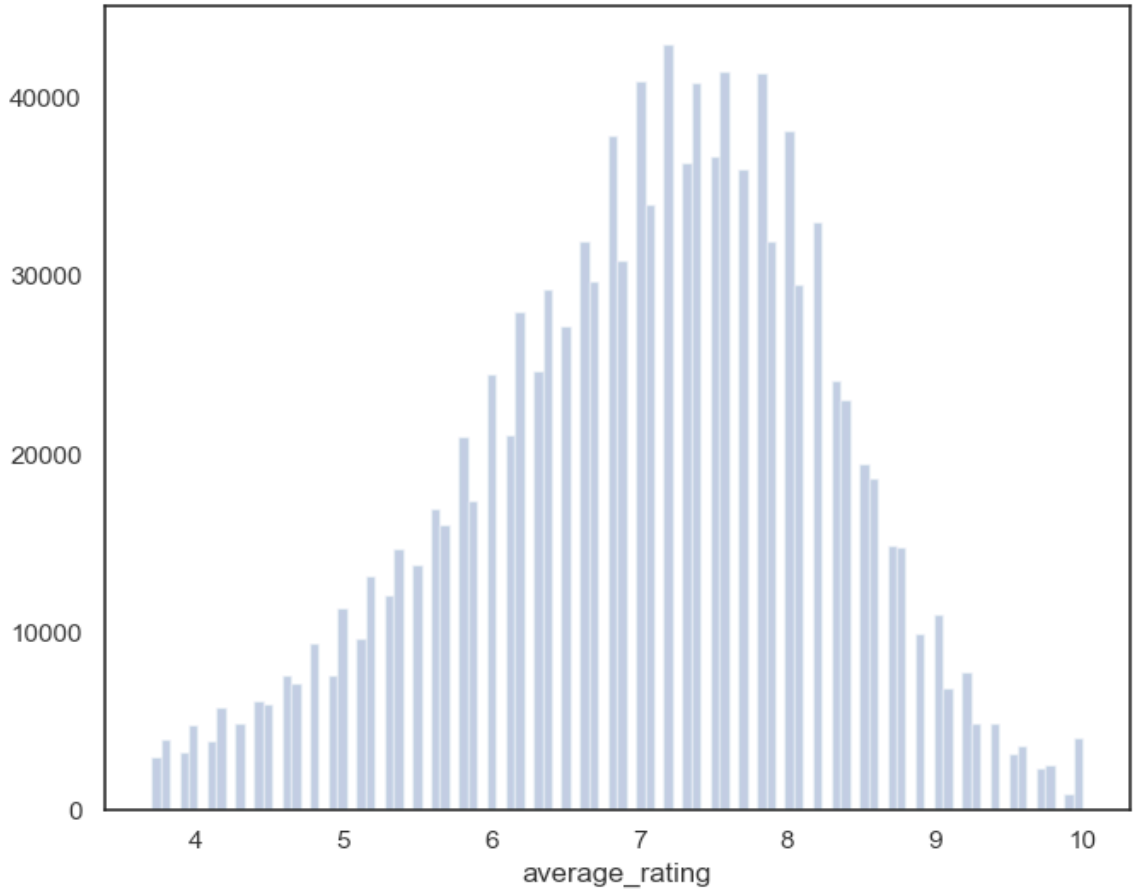
Data Statistics(ranking)

Min Value: 1.0
Maximum Value: 10.0
Threshold: 9.0

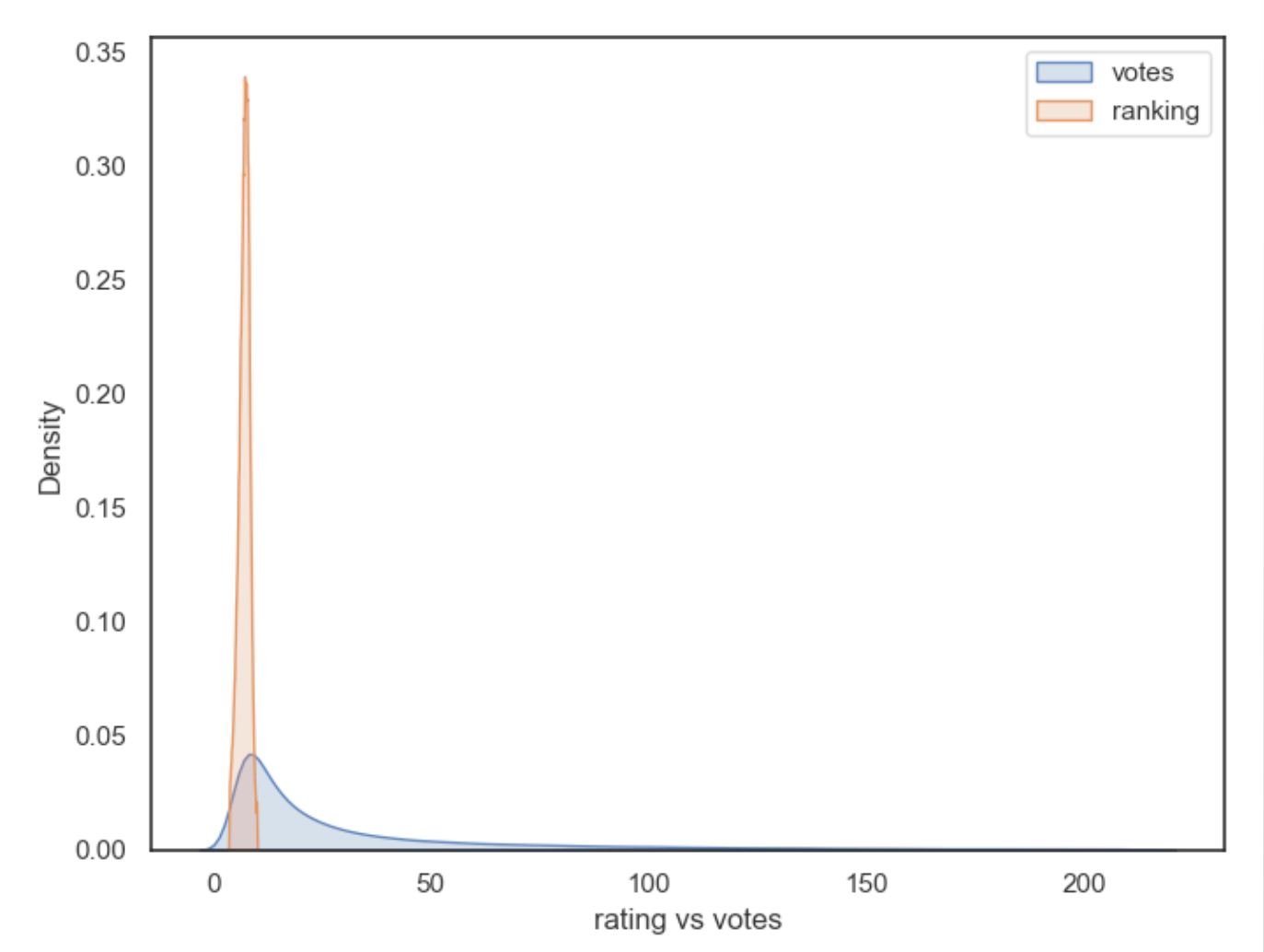
Mean: 6.92
Median: 7.1
Kurtosis: -0.18
Skewness: -0.34

Percentil 10: 5.1
Percentil 25: 6.2
Percentil 75: 7.9
Percentil 90: 8.5

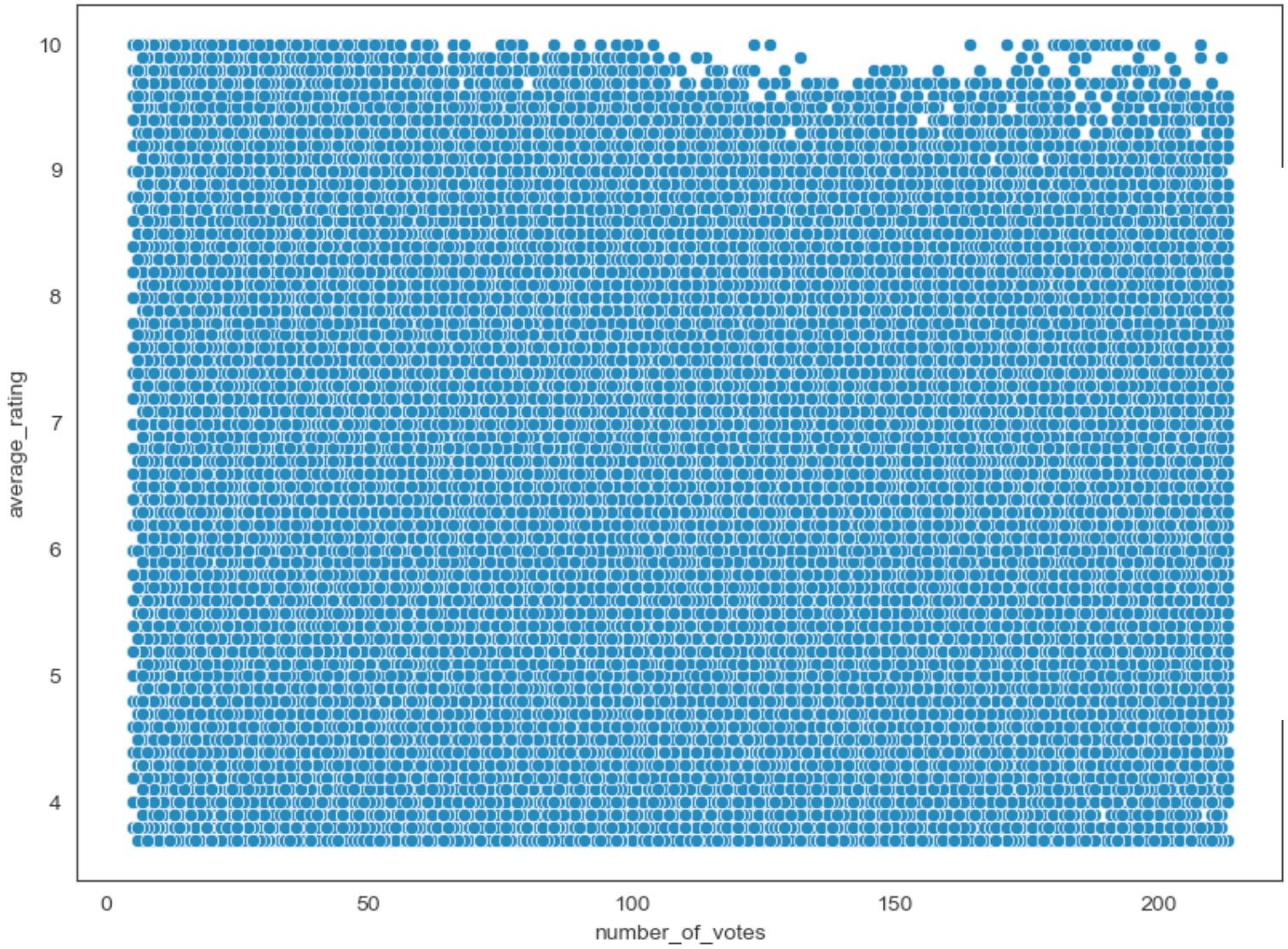




Data Statistics(ranking vs votes)



Data Statistics



corr:-0.0418