IMDB and Kaggle Datasets

Group BHNK: Brionna Huynh and Neha Kumar

Datasets

- IMDB (Hollywood and other international film industries) and Kaggle (Bollywood only) datasets
- IMDB: Title_Basics (tconst, title, year, runtime), Name_Basics (nconst, name, birth year, death year), Title_Akas (tconst, title, region, language), Title_Principals (tconst, characters), Title_Episode (tconst, season, episode), Title_Ratings (tconst, rating, numVotes), Title_Crew (tconst, directors, writers)
- Kaggle: Bolly_actors (name, height), Bolly_actress (name, height, debut as lead), Bolly_movies (title, director, cast, genre)

Staging/ Modeled Tables

- IMDB: Combine the Title_Basics and Title_Akas tables (both referring to attributes of movies, but have some different attributes)
  - One-to-one: Title_Basics and Title_Akas, Title_Basics and Title_Ratings
  - One-to-many: Title_basics tconst to genres column (new Genres table), Name_basics nconst to primaryProfession column (new Primary_profession table), Name_basics nconst to knownForTitles column (new Known_for table), Title_crew tconst to directors column (new Directors table), Title_crew tconst to writers column (new Writers table),
  - Many-to-many: Title_principal tconst/ nconst to characters column (new Characters table)
- Kaggle:
  - One-to-many: Bolly_movies to Director column (new Title_director table), Bolly_movies to Cast column (new Title_cast table), Bolly_movies to Genre column (new Title_genre table)

Beam pipelines

- Made a new child table that refers to the main parent table
- Split the element by the commas to get rid of the one to many relationship
- Tables: Characters, Directors, Genre, Known_For, Primary_Professions, Writers, Title_cast, Title_director, Title_genre

id	values
id0717	12, 34, 87, 43

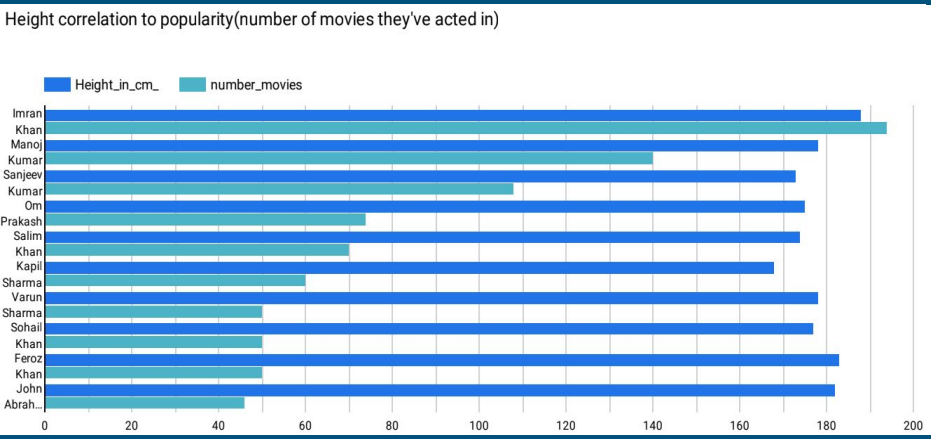
id	values
id0717	12
id0717	34
id0717	87
id0717	43

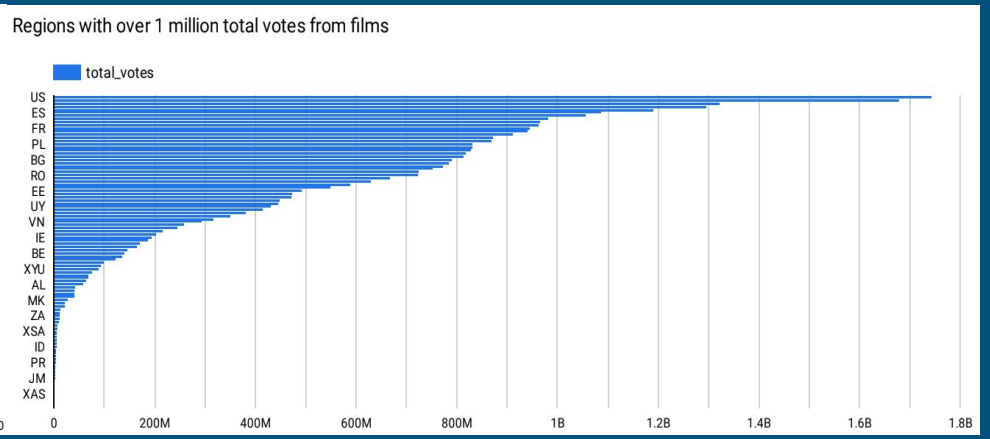
Areas of interest

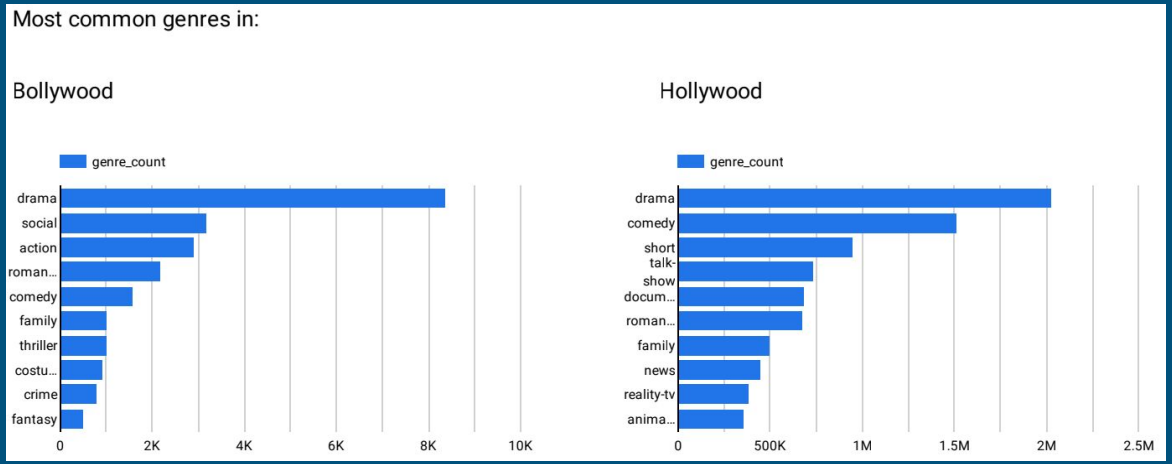
- Generally tried to learn more about Bollywood actors and movies by joining these tables with tables in the imdb_refined dataset
  - The kaggle dataset did not contain many details about both the movies and the actors, and the imdb_refined dataset included a lot more details about the same movies and actors
- Wanted to see if height was an important aspect of the Bollywood film industry
- Wanted to further examine the average ratings of Bollywood movies

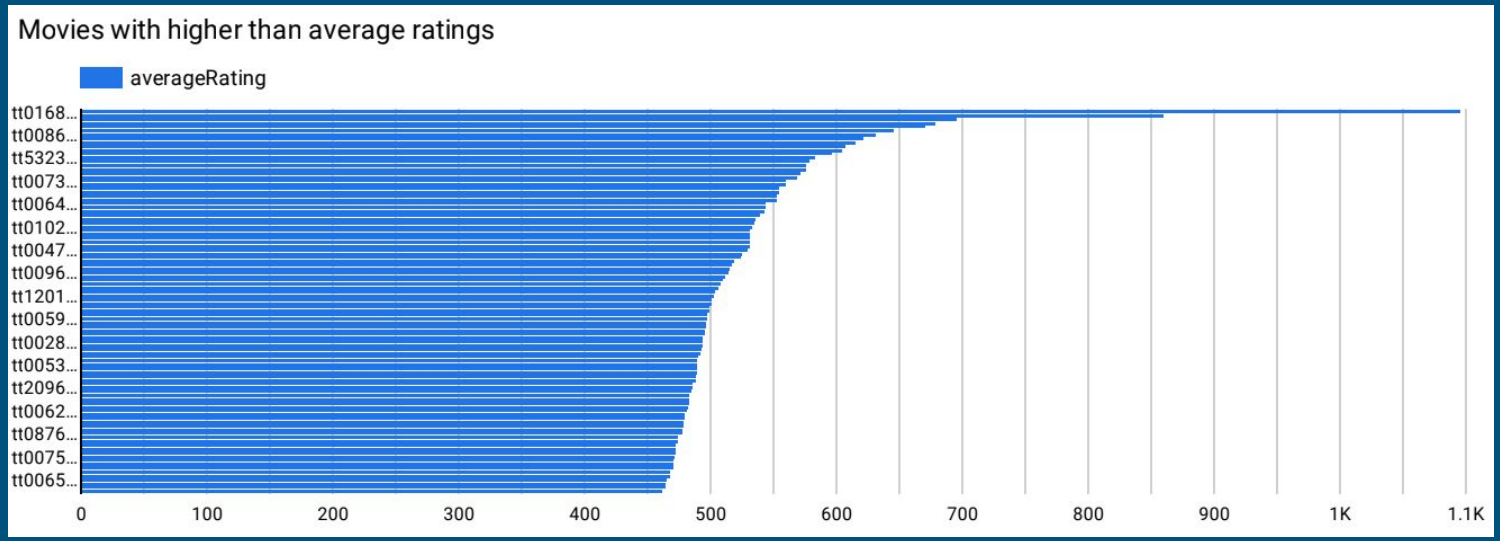
Queries and Data Visualization

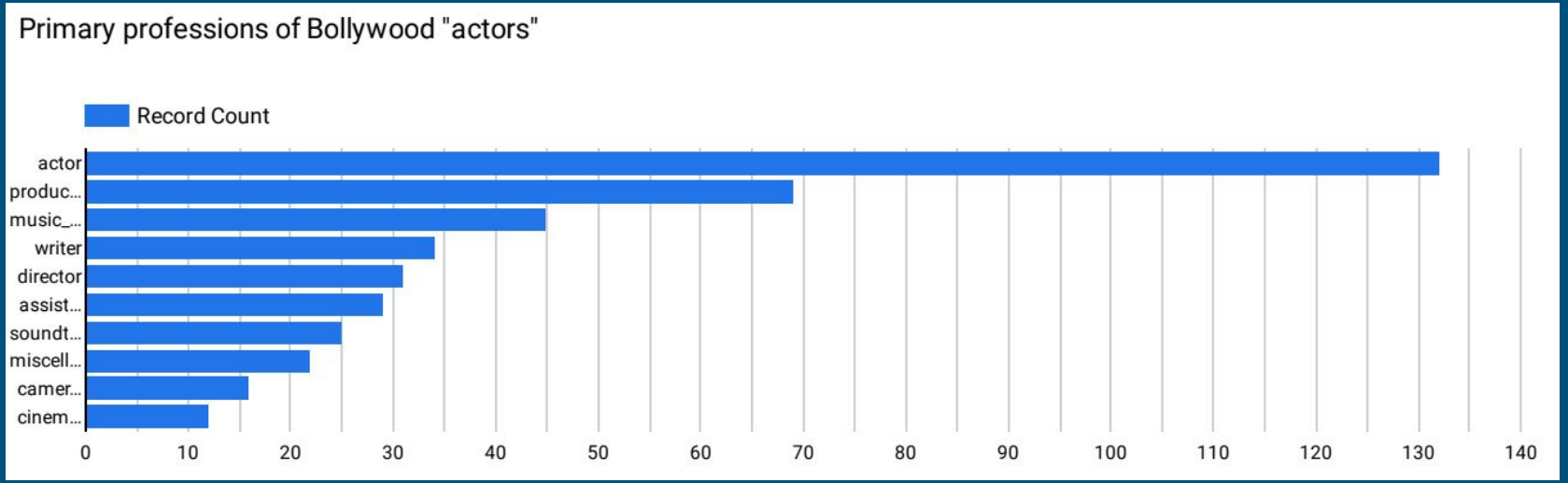
- Find all the movies (both from imdb and kaggle datasets) that fit into the genre War
  - Used a join to get all of the collective output of all the movies that are War based
- Find the ratings of a specific director's movies
  - Use the inner query to find the Titles of this director's movies and in the outer query, get the ratings of these Titles
- Are present-day Bollywood actresses taller than past actresses?
- Find the titles that have higher than average ratings
- Find the regions in which more than 1 million ratings (votes) for Titles are cast
- Find other professions of Bollywood actors
- What are the most common genres in Bollywood and Hollywood?
- Are the taller Bollywood actors more popular than the shorter ones?
  - Used knownForTitles to interpret popularity → if known for more Titles, then must be more popular


Height correlation to popularity(number of movies they've acted in)
Height_in_cm_
number_movies
Imran Khan
Manoj Kumar
Sanjeev Kumar
Om Prakash
Salim Khan
Kapil Sharma
Varun Sharma
Sohail Khan
Feroz Khan
John Abrah...
0
20
40
60
80
100
120
140
160
180
200


Regions with over 1 million total votes from films
total_votes
US
ES
FR
PL
BG
RO
EE
UY
VN
IE
BE
XYU
AL
MK
ZA
XSA
ID
PR
JM
XAS
0
200M
400M
600M
800M
1B
1.2B
1.4B
1.6B
1.8B


Most common genres in:
Bollywood
genre_count
drama
social
action
roman...
comedy
family
thriller
costu...
crime
fantasy
0
2K
4K
6K
8K
10K
Hollywood
genre_count
drama
comedy
short
talk-show
docum...
roman...
family
news
reality-tv
anima...
0
500K
1M
1.5M
2M
2.5M


Movies with higher than average ratings
averageRating
tt0168...
tt0086...
tt5323...
tt0073...
tt0064...
tt0102...
tt0047...
tt0096...
tt1201...
tt0059...
tt0028...
tt0053...
tt2096...
tt0062...
tt0876...
tt0075...
tt0065...
0
100
200
300
400
500
600
700
800
900
1K
1.1K


Primary professions of Bollywood "actors"
Record Count
actor
produc...
music_...
writer
director
assist...
soundt...
miscell...
camer...
cinem...
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140

Challenges and Accomplishments

- The hardest part was deciding how to remodel the data
  - Had to think of the possible questions/ queries that would come up and account for this when remodeling the data
- Inconsistencies across datasets
  - Had to account for differences in formatting between tables and datasets and can to update this in order to run queries
  - “War” vs “war” in the 2 genre tables
- Were able to overcome the errors we encountered and ultimately learned a lot from the interesting queries we were able to run

Future improvements

- Look at the genres of both the datasets a little closer
  - Have multiple names for the same genre, so would try to make this more consistent between the datasets
- Remove duplicate titles from the Title_Basics table in the imdb dataset
  - Director's movie ratings query shows that there are repeat Title names with different attributes for each (different ratings for the same movie)
- Combine the Title_basics and Bolly_movies table to make it easier to access titles and information surrounding these titles
  - Can then combine the 2 cast tables, 2 directors tables and 2 genres tables