Databases and Advanced Data Techniques Midterm: Top 100 Korean Dramas.

In recent years a surge in the consumption of international entertainment has seen itself occurring across the globe, with the popularisation of streaming platforms and other web based services alike, the access to series, movies and music from different parts of the globe has become unrestricted. One country whose entertainment industry has seen unparalleled growth thanks to many of these factors is South Korea, in the words of Steve Chung, chief global officer of CJ ENM and co-CEO of CJ ENM America. “Netflix’s recent announcement of a $2.5 billion investment in its production of South Korean movies and television shows is only the latest data point to suggest that Asia is a rising content giant—and Seoul sits at the centre of it all.”[1]

Due to the previously exposed rationale the data set chosen for the purposes of this exploration is “Top 100 Korean Drama (MyDramaList)”[2] which is a data set that compiles the top 100 ranked korean dramas as of Dec 30 2021 according to the site MyDramaList.com, “a community-driven platform where Asian drama and movie fans can create their own lists, discuss their favourite shows and movies, discover new content, and make friends”.[3]. The data set is a compilation created by Kaggle user “Chanon Charuchinda '', the dataset has a newer version released in august 2023 but when analysing this version it was possible to notice it had some inconsistencies so the previous version was preferred over this one, due to this circumstance it may not be entirely up to date, nonetheless still holds a reliable source of information. This data can be easily accessible directly from the MyDramaList website and although it may represent a biassed view on the actual ranking of the information presented, this last aspect does not undermine its relevance.

The data set contains detailed information about the one hundred highest ranked korean TV series, from here on out referred as Dramas, up to the the previously stated date, the fields included are: Name (of the series), Year of Release, Aired Date, Number of Episodes, Network, Duration, Content Rating, Synopsis, Cast, Genre, Tags and Rank, the level of detail to which each of these fields is explored in the dataset is just enough to consider it useful for the purposes of the exploration, nonetheless it can not be ignored the information contained is still just basic information with just a little above minimal level of detail, however during no part of this project the information stopped or failed to serve its purposes.

Regarding the documentation, it was available in the same page as the actual dataset and although not particularly helpful during the course of the work, since the fields of data were already very self explanatory, it was easy enough to find such that had it been required it would have been easily accessible. The dataset could be interrelated to other similar sets of data which expand on aspects it does not, but doing so would result difficult, not so much for the interrelation but for finding a set of data that could actually expand yet hold a relation with the dataset in question, the later can be assured since when originally exploring datasets for this topic in particular it was possible to notice these were limited to the existence of only a few, and not all of them particularly interesting, either redundant or lacking on vital information.

This dataset is under a Creative Commons CC0 Public Domain licence, this information can be found in the page for the dataset under the ‘Licence’ tag, which essentially means “You

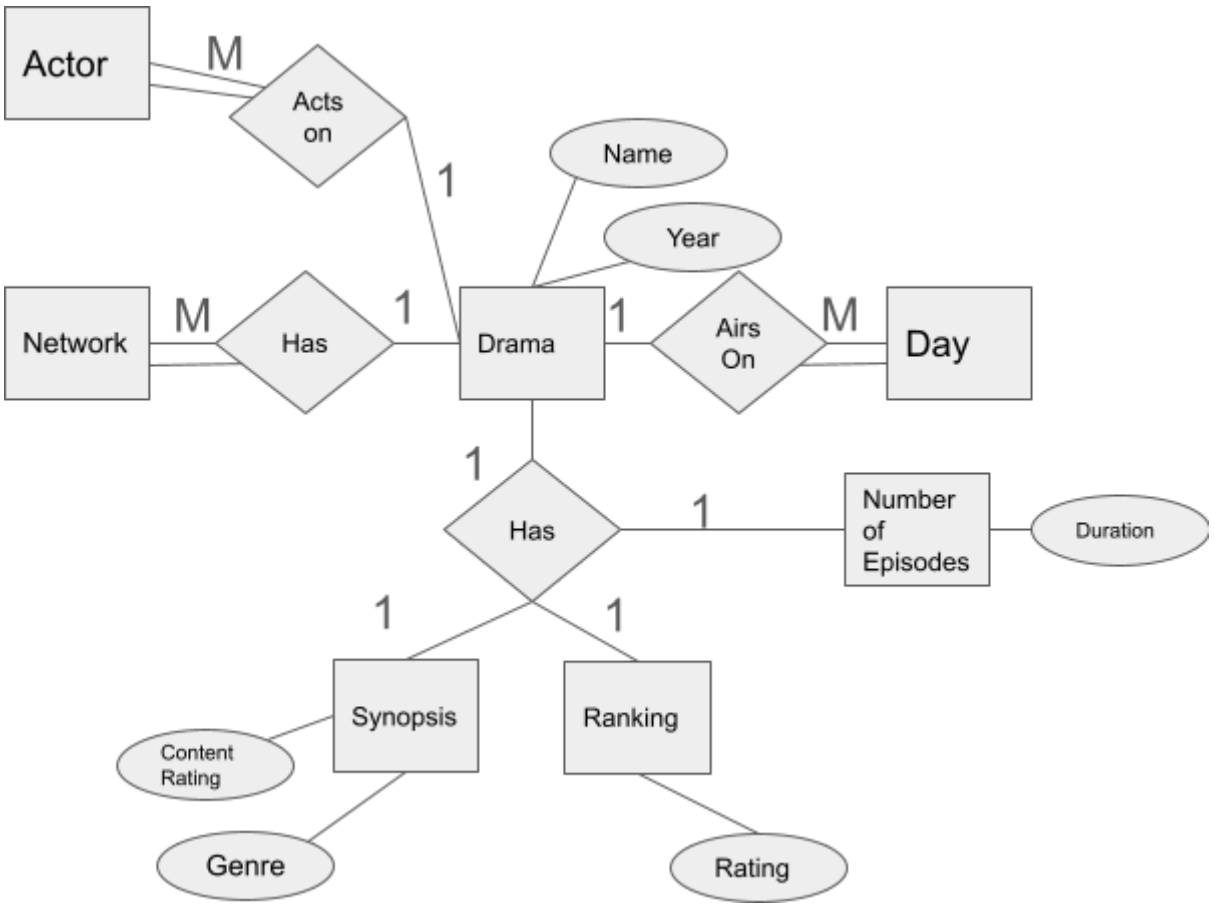
can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission."[3]
When deciding the topic for the dataset it was determined it should be something easy to understand independently of the background of the person who were to use it and also it had to correlate with a relevant and interesting topic, these two reasons ultimately culminated in deciding to select the previously described dataset, since it was found many interesting questions could be asked to it, some of them being:

- Which were the highest ranked dramas by year?
- How does the day of airing correlates with the ranking of the drama?
- What is the popularity of the actors based on the number of dramas they appear in?
- What is the synopsis of the most popular dramas?
- To which genre the most popular dramas belong to? And what is their content rating?
- Which network were the most popular dramas aired on?
- How many episodes and what was their duration for the highest ranked dramas?

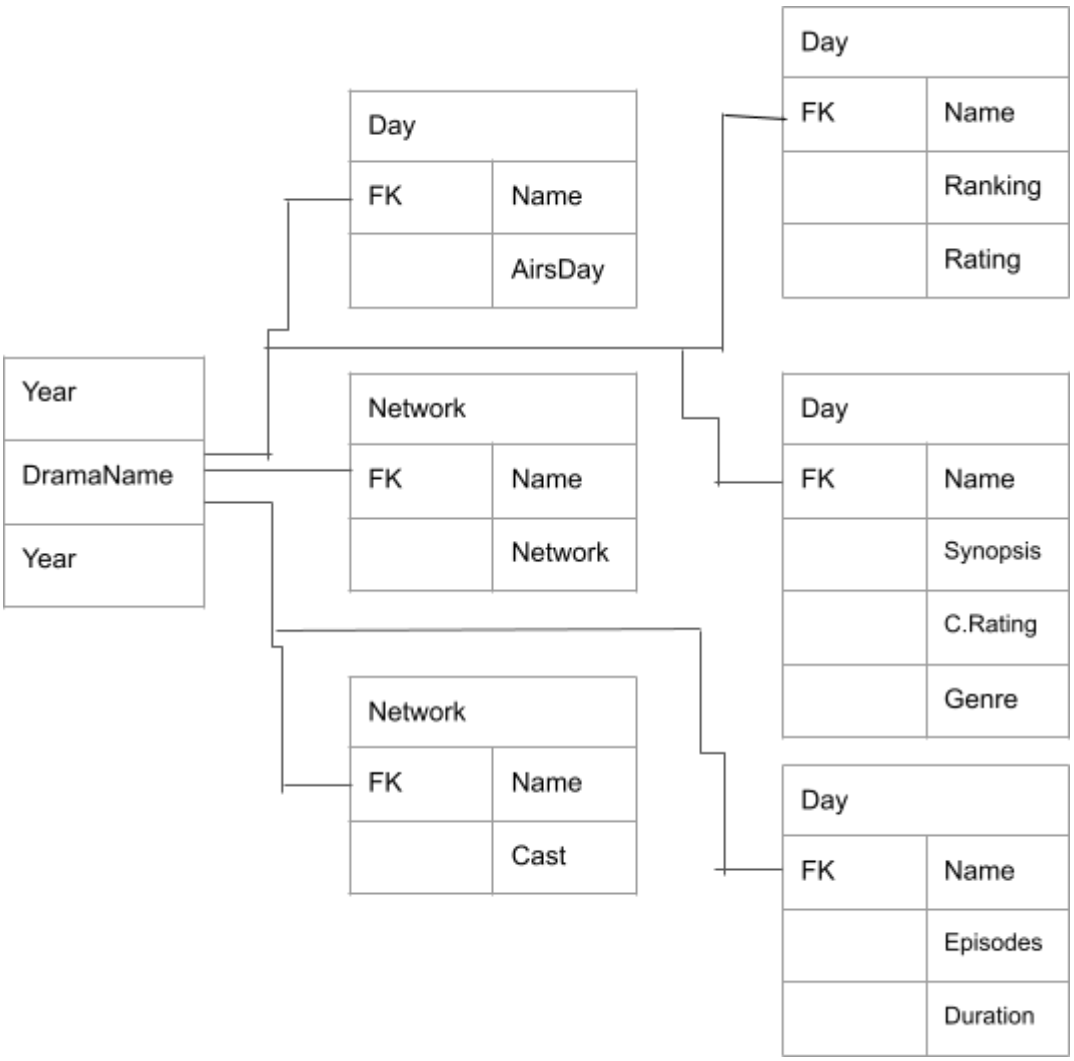
All these questions, among others specified below, make a good baseline to begin understanding which factors directly affect the popularity of Korean Dramas and also allow us to comprehend what are the aspects that have propelled this series into worldwide fame.

ER Diagram

Given the nature of the work to be carried out it was decided to remove two of the columns in the data set, the 'Aired Date' and 'Tags' columns, since both of them consisted of repetitive, redundant data already covered by other columns like year of release and genre.



The former results in the following Relational mapping



Normalisation

The first step towards normalisation is to ensure the data set is in 1NF, as of currently this database is not, to say something, each entry in the actors columns contains an array of names instead of a single scalar value, something similar happens in the network, genre and day of airing columns.

Name	Year of release	Aired On	Number of Episode	Network	Duration (min)	Content Rating	Synopsis	Cast	Genre	Rank	Rating
Move to Heaven	2021	Friday	10	Netflix	52	18+ Restricted (violence & profanity)	Geu Roo is a young autistic man...	Lee Je Hoon	Life	#1	9.2

Move to Heaven	2021	Friday	10	Netflix	52	18+ Restricted (violence & profanity)	Geu Roo is a young autistic man...	Tang Jun Sang	Drama	#1	9.2
Move to Heaven	2021	Friday	10	Netflix	53	18+ Restricted (violence & profanity)	Geu Roo is a young autistic man...	Hong Seung	Family	#1	9.2
...	...	...	...	...	...	...	...	...	...	...	...

Leaving the data set like this would not be enough to make it both usable and normalised, the next step is to convert it to 2NF to do this, the primary key in this case would be the name of the drama, not all the data is irreducibly dependent on this primary key, for example rating is dependent on rank.

Name	Year of Release
Move to Heaven	2021
Hospital Playlist	2020
Flower of Evil	2020
Hospital Playlist 2	2021
…	…

Name	Aired On
Move to Heaven	Friday
Hospital Playlist	Thursday
Flower of Evil	Wednesday
Flower of Evil	Thursday
Hospital Playlist 2	Thursday
…	…

Name	Number of Episodes
Move to Heaven	10
Hospital Playlist	12
Flower of Evil	16
Hospital Playlist 2	12

…	…

Name	Duration
Move to Heaven	52 min.
Hospital Playlist	1 hr. 30 min.
Flower of Evil	1 hr. 10 min.
Hospital Playlist 2	1 hr. 40 min.
…	…

Name	Network
Move to Heaven	Netflix
Hospital Playlist	Netflix
Hospital Playlist	tvN
Flower of Evil	tvN
Hospital Playlist 2	Netflix
Hospital Playlist 2	tvN
…	…

Name	Cast
Move to Heaven	Lee Je Hoon
Move to Heaven	Tang Jun Sang
Move to Heaven	Hong Seung Hee
Move to Heaven	Jung Suk Yong
…	…

Name	Rank	Rating
Move to Heaven	#1	9.2
Hospital Playlist	#2	9.1
Flower of Evil	#3	9.1
Hospital Playlist 2	#4	9.1
…	…	…

Name	Content Rating	Synopsis	Genres
Move to Heaven	18+ Restricted (violence &	Geu Roo is a young autistic	Life

	profanity)	man...	
Move to Heaven	18+ Restricted (violence & profanity)	Geu Roo is a young autistic man...	Drama
Move to Heaven	18+ Restricted (violence & profanity)	Geu Roo is a young autistic man...	Family
Hospital Playlist	15+ - Teens 15 or older	The stories of people going through...	Friendship
…	…	…	…

After doing this the dataset still has some transitive dependency, like in the table above Content Rating and Genres are only dependent on synopsis not on the primary key Name. After applying 3NF to the necessary tables they end up as follows, (in the case of synopsis since mysql restricts the types of data that can be used as primary keys it is necessary to add a unique identifier):

Synopsis_ID	Name	Synopsis
1A	Move to Heaven	Geu Roo is a young autistic man...
2A	Hospital Playlist	The stories of people going through…
3A	Flower of Evil	Although Baek Hee Sung is hiding a dark secret…
4A	Hospital Playlist 2	Everyday is extraordinary for five doctors…
…	…	…

Synopsis_ID	Content Rating	Genres
1A	18+ Restricted (violence & profanity)	Life
1A	18+ Restricted (violence & profanity)	Drama
1A	18+ Restricted (violence & profanity)	Family

2A	15+ - Teens 15 or older	Friendship
…	…	…

Name	Rank
Move to Heaven	1
Hospital Playlist	2
Flower of Evil	3
Hospital Playlist 2	4
…	…

Rank	Rating
1	9.2
2	9.1
3	9.1
4	9.1
…	…..

To go beyond this point of normalisation would require to do something about the content and genres table, it is possible to notice that in this table a synopsis can have on content rating but multiple genres, so to fix this it is necessary to apply Boyce-Codd Normal Form, which results in the following tables, (in this case it is also necessary to add an unique identifier for the content rating table since .

Synopsis_ID	Content Rating
1A	18+ Restricted (violence & profanity)
2A	15+ - Teens 15 or older
3A	15+ - Teens 15 or older
4A	15+ - Teens 15 or older
…	…

Synopsis_ID	Genres
1A.	Life

1A	Drama
1A	Family
2A	Friendship
…	…

Despite having reached this point of normalisation due to the nature of the data some tables do not have an unique primary key but rather a composite primary key, such is the case of the day, network, genre and cast tables.

Creating the database

The following were the commands used to create the database:

```
CREATE DATABASE kdrama;
USE kdrama;
CREATE TABLE year (Name VARCHAR(255) NOT NULL, Year_of_Release YEAR(4), PRIMARY KEY(Name));
CREATE TABLE day (Name VARCHAR(255) NOT NULL, Aired_On VARCHAR(255), PRIMARY KEY(Aired_On, Name), FOREIGN KEY(Name) REFERENCES year(Name));
CREATE TABLE network (Name VARCHAR(255) NOT NULL, Network VARCHAR(255), PRIMARY KEY(Name, Network), FOREIGN KEY(Name) REFERENCES year(Name));
CREATE TABLE episodes (Name VARCHAR(255) NOT NULL, Num_of_Episodes INT, PRIMARY KEY(Name), FOREIGN KEY(Name) REFERENCES year(Name));
CREATE TABLE duration (Name VARCHAR(255) NOT NULL, Ep_Duration INT, PRIMARY KEY(Name), FOREIGN KEY(Name) REFERENCES year(Name));
CREATE TABLE synopsis (Synopsis_ID VARCHAR(255) NOT NULL, Name VARCHAR(255) NOT NULL, Synopsis TEXT, PRIMARY KEY(Synopsis_ID), FOREIGN KEY(Name) REFERENCES year(Name));
CREATE TABLE content (Synopsis_ID VARCHAR(255) NOT NULL, Content_Rating TEXT, PRIMARY KEY(Synopsis_ID), FOREIGN KEY(Synopsis_ID) REFERENCES synopsis(Synopsis_ID));
CREATE TABLE genre (Synopsis_ID VARCHAR(255) NOT NULL, Genre VARCHAR(255), PRIMARY KEY(Synopsis_ID, Genre), FOREIGN KEY(Synopsis_ID) REFERENCES synopsis(Synopsis_ID));
CREATE TABLE cast (Name VARCHAR(255) NOT NULL, Cast VARCHAR(255), PRIMARY KEY(Name, Cast), FOREIGN KEY(Name) REFERENCES year(Name));
CREATE TABLE drama_rank (Name VARCHAR(255) NOT NULL, MDL_Rank INT NOT NULL, PRIMARY KEY(MDL_Rank), FOREIGN KEY(Name) REFERENCES year(Name));
CREATE TABLE rating (MDL_Rank INT NOT NULL, Total_Rating DECIMAL(4,1), PRIMARY KEY(MDL_Rank), FOREIGN KEY(MDL_Rank) REFERENCES drama_rank(MDL_Rank));
```

For the sake of simplifying the dataset only the first 30 rows out of 100 were added into the database, the following commands add said data:

```
LOAD DATA INFILE '/home/coder/project/Data/Year.csv' INTO TABLE year FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\r\n' IGNORE 1 ROWS;
LOAD DATA INFILE '/home/coder/project/Data/Day.csv' INTO TABLE day FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\r\n' IGNORE 1 ROWS;
LOAD DATA INFILE '/home/coder/project/Data/Network.csv' INTO TABLE network FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\r\n' IGNORE 1 ROWS;
LOAD DATA INFILE '/home/coder/project/Data/Episodes.csv' INTO TABLE episodes FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\r\n' IGNORE 1 ROWS;
LOAD DATA INFILE '/home/coder/project/Data/Duration.csv' INTO TABLE duration FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\r\n' IGNORE 1 ROWS;
LOAD DATA INFILE '/home/coder/project/Data/Synopsis.csv' INTO TABLE synopsis FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\r\n' IGNORE 1 ROWS;
LOAD DATA INFILE '/home/coder/project/Data/Content.csv' INTO TABLE content FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\r\n' IGNORE 1 ROWS;
LOAD DATA INFILE '/home/coder/project/Data/Genre.csv' INTO TABLE genre FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\r\n' IGNORE 1 ROWS;
LOAD DATA INFILE '/home/coder/project/Data/Cast.csv' INTO TABLE cast FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\r\n' IGNORE 1 ROWS;
LOAD DATA INFILE '/home/coder/project/Data/Rank.csv' INTO TABLE drama_rank FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\r\n' IGNORE 1 ROWS;
LOAD DATA INFILE '/home/coder/project/Data/Rating.csv' INTO TABLE rating FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\r\n' IGNORE 1 ROWS;
```

To answer the questions established during the introduction the following queries were executed:

- Retrieve ranking and rating of Dramas by year:

```
SELECT drama_rank.Name, MIN(drama_rank.MDL_Rank), rating.Total_Rating FROM drama_rank JOIN year ON drama_rank.Name = year.Name AND year.Year_of_Release=2020 JOIN rating ON drama_rank.MDL_Rank = rating. MDL_Rank GROUP BY Name, drama_rank.MDL_Rank;
```

- Retrieve best rating by year:

```
SELECT MAX(Total_Rating) FROM rating JOIN drama_rank ON drama_rank.MDL_Rank = rating. MDL_Rank JOIN year ON drama_rank.Name = year.Name AND year.Year_of_Release=2019;
```

- Retrieve the day of airing for best ranked dramas:

```
SELECT day.Aired_On, drama_rank.MDL_Rank, rating.Total_Rating FROM drama_rank JOIN day ON drama_rank.Name = day.Name AND drama_rank.MDL_Rank <= 5 JOIN rating ON drama_rank.MDL_Rank = rating. MDL_Rank GROUP BY Aired_On, drama_rank.MDL_Rank;
```

- Retrieve the popularity of airing days:

SELECT Aired_On, count(*) as _mycount FROM day GROUP BY Aired_On ORDER BY _mycount DESC;

- Retrieve the 5 most popular actor/actress:

SELECT Cast, count(*) as _mycount FROM cast GROUP BY Cast ORDER BY _mycount DESC LIMIT 5;
Or retrieve popularity of actor/actress:

SELECT Cast, count(*) as _mycount FROM cast WHERE Cast='Song Joong Ki' GROUP BY Cast;

- Retrieve the cast of the highest ranked dramas:

SELECT drama_rank.Name, cast.Cast, drama_rank.MDL_Rank FROM drama_rank JOIN cast ON drama_rank.Name = cast.Name AND drama_rank.MDL_Rank <= 5 GROUP BY drama_rank.MDL_Rank, cast.Cast;

- Retrieve the synopsis of the highest ranked dramas:

SELECT drama_rank.MDL_Rank, drama_rank.Name, synopsis.Synopsis FROM drama_rank JOIN synopsis ON drama_rank.Name = synopsis.Name AND drama_rank.MDL_Rank <= 5 GROUP BY drama_rank.MDL_Rank, synopsis.Synopsis;

- Retrieve the Genres of the highest ranked dramas:

SELECT drama_rank.Name, drama_rank.MDL_Rank, genre.Genre FROM synopsis JOIN drama_rank ON synopsis.Name = drama_rank.Name AND drama_rank.MDL_Rank <= 5 JOIN genre ON synopsis.Synopsis_ID = genre.Synopsis_ID GROUP BY Name, MDL_Rank, genre.Genre;

- Retrieve the popularity of Genres:

SELECT Genre, count(*) as _mycount FROM genre GROUP BY Genre ORDER BY _mycount DESC;

- Retrieve the Content Rating of the highest ranked dramas:

SELECT drama_rank.Name, drama_rank.MDL_Rank, content.Content_Rating FROM synopsis JOIN drama_rank ON synopsis.Name = drama_rank.Name AND drama_rank.MDL_Rank <= 5 JOIN content ON synopsis.Synopsis_ID = content.Synopsis_ID GROUP BY Name, MDL_Rank, content.Content_Rating;

- Retrieve the popularity of the content ratings:

SELECT Content_Rating, count(*) as _mycount FROM content GROUP BY Content_Rating ORDER BY _mycount DESC;

- Retrieve the Network of emission for the highest ranked dramas:

SELECT drama_rank.Name, drama_rank.MDL_Rank, network.Network FROM drama_rank JOIN network ON drama_rank.Name = network.Name AND drama_rank.MDL_Rank <= 5;

- Retrieve the popularity for the emission networks:

SELECT Network, count(*) as _mycount FROM network GROUP BY Network ORDER BY _mycount DESC;

- Retrieve number of episodes and episode duration for the most popular dramas:

SELECT drama_rank.Name, drama_rank.MDL_Rank, episodes. Num_of_Episodes, duration.Ep_Duration FROM drama_rank JOIN duration ON drama_rank.Name = duration.Name JOIN episodes ON drama_rank.Name = episodes.Name AND drama_rank.MDL_Rank <= 5;

- Average number of episodes for a drama:

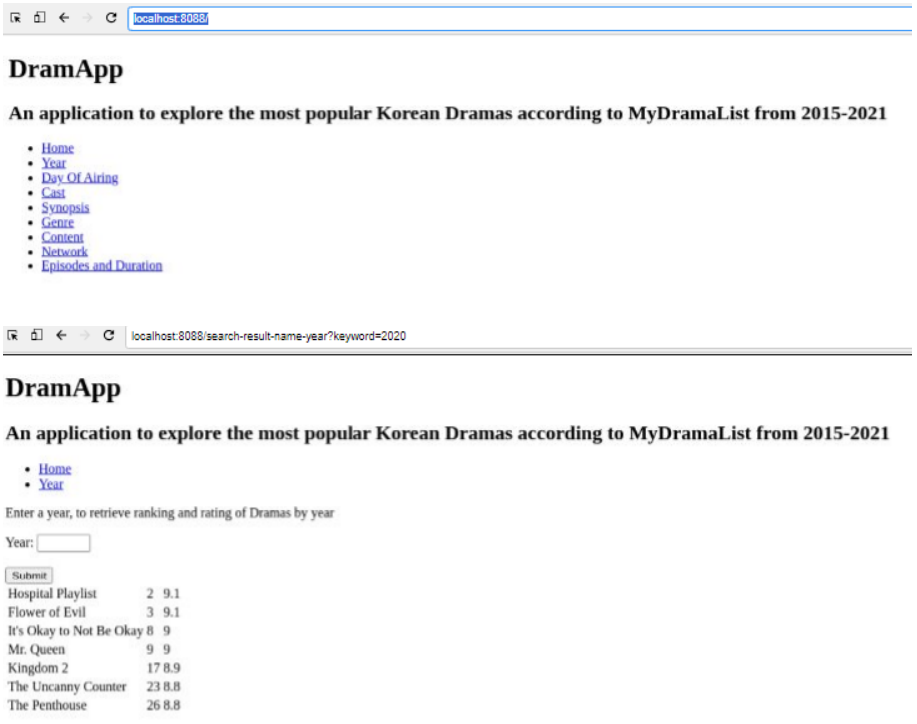
SELECT AVG(Num_of_Episodes) FROM episodes;

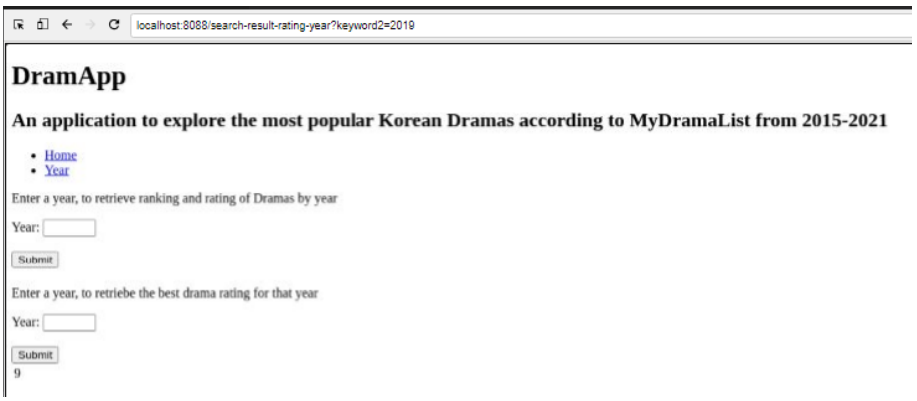
- Average number of episode duration for a drama:

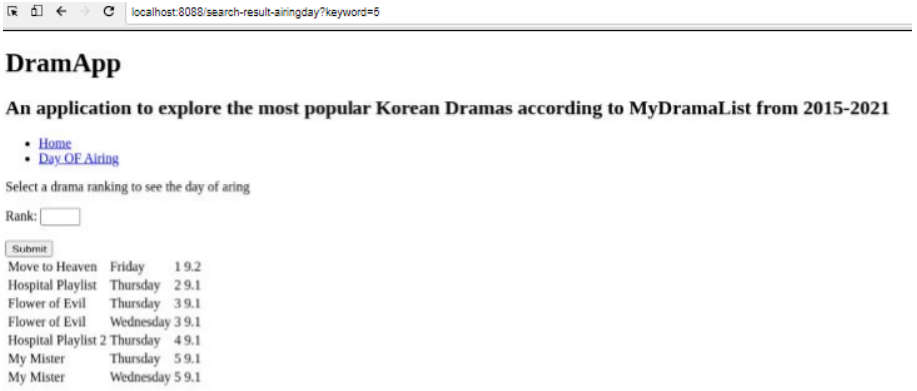
SELECT AVG(Ep_Duration) FROM duration;

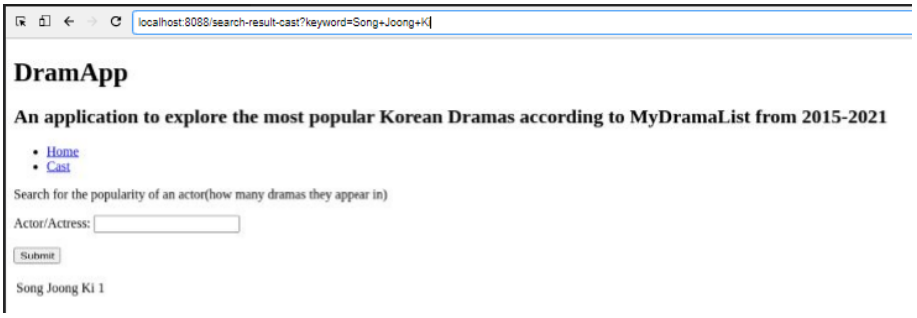
Database Application

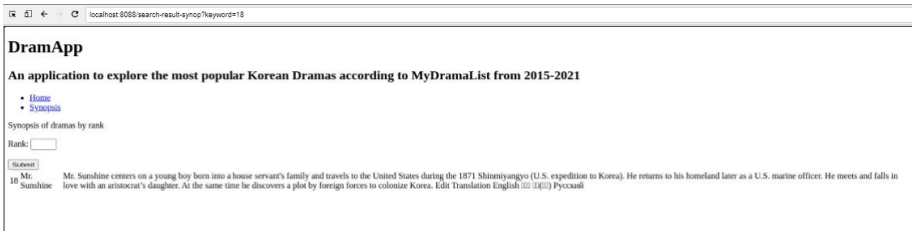
This is the application after being successfully implemented.

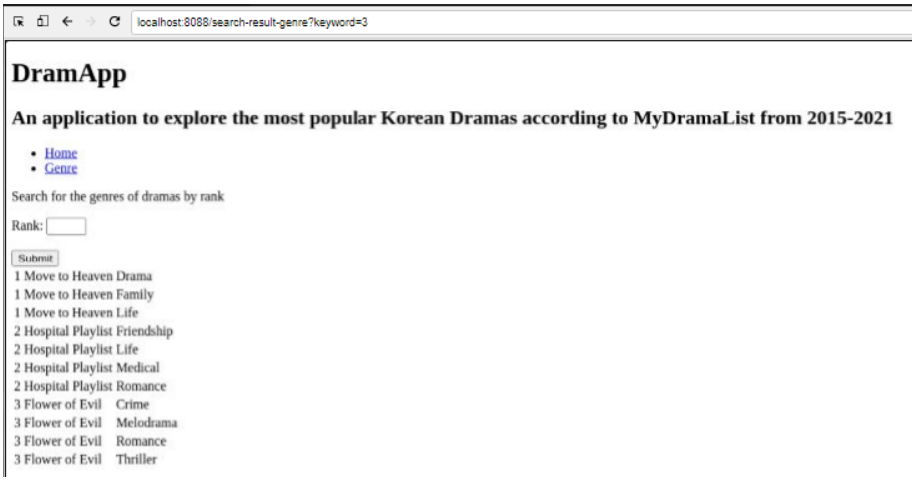


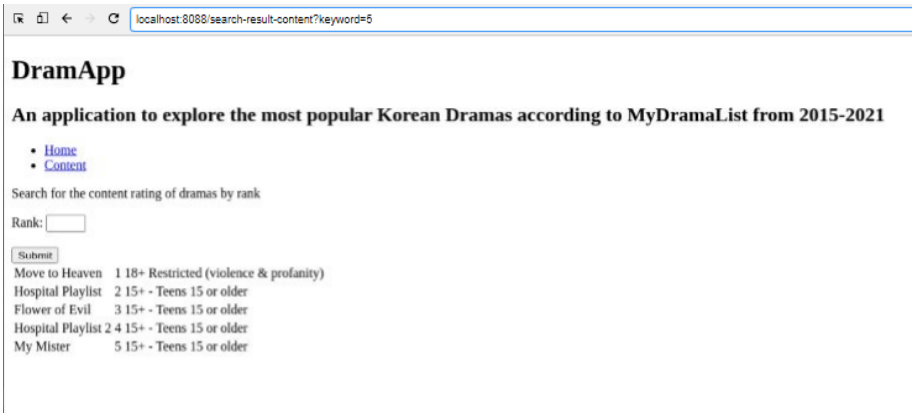

localhost:8088/search-result-rating-year?keyword2=2019
DramApp
An application to explore the most popular Korean Dramas according to MyDramaList from 2015-2021
Home
Year
Enter a year, to retrieve ranking and rating of Dramas by year
Year:
Submit
Enter a year, to retriebe the best drama rating for that year
Year:
Submit
9

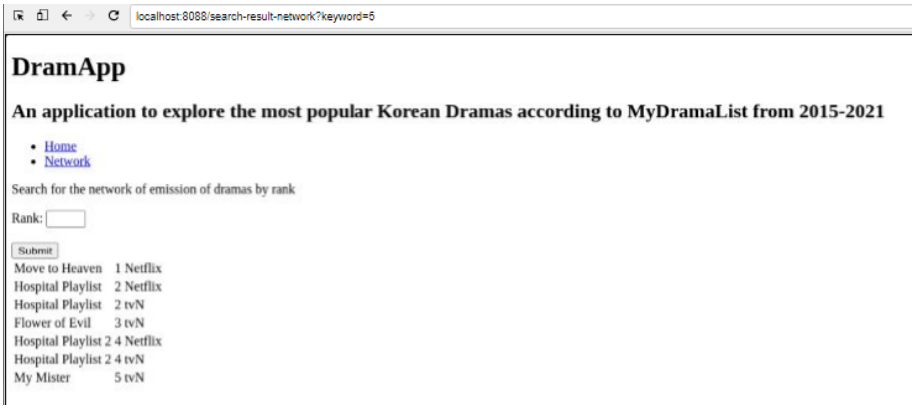

localhost:8088/search-result-airingday?keyword=5
DramApp
An application to explore the most popular Korean Dramas according to MyDramaList from 2015-2021
Home
Day OF Airing
Select a drama ranking to see the day of aring
Rank:
Submit
Move to Heaven Friday 1 9.2
Hospital Playlist Thursday 2 9.1
Flower of Evil Thursday 3 9.1
Flower of Evil Wednesday 3 9.1
Hospital Playlist 2 Thursday 4 9.1
My Mister Thursday 5 9.1
My Mister Wednesday 5 9.1

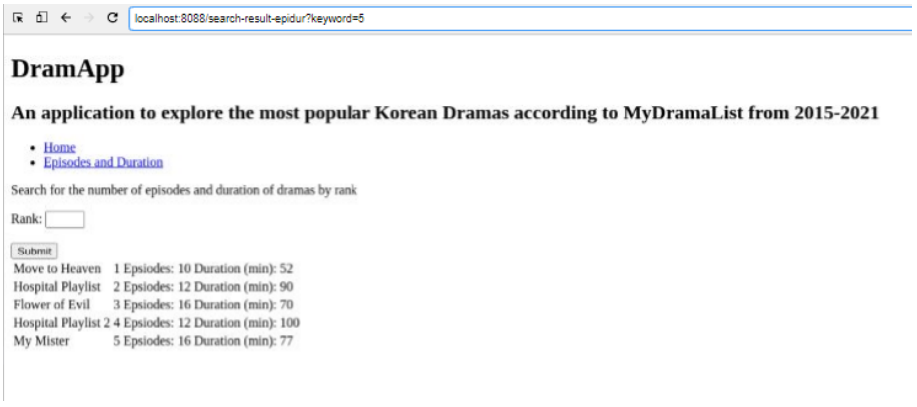

localhost:8088/search-result-cast?keyword=Song+Joong+Ki
DramApp
An application to explore the most popular Korean Dramas according to MyDramaList from 2015-2021
Home
Cast
Search for the popularity of an actor(how many dramas they appear in)
Actor/Actress:
Submit
Song Joong Ki 1


localhost:8088/search-result-synop?keyword=18
DramApp
An application to explore the most popular Korean Dramas according to MyDramaList from 2015-2021
Home
Synopsis
Synopsis of dramas by rank
Rank:
Submit
16 Mr. Sunshine
Mr. Sunshine centers on a young boy born into a house servant's family and travels to the United States during the 1871 Shinmiyangyo (U.S. expedition to Korea). He returns to his homeland later as a U.S. marine officer. He meets and falls in love with an aristocrat's daughter. At the same time he discovers a plot by foreign forces to colonize Korea. Edit Translation English Русский


localhost:8088/search-result-genre?keyword=3
DramApp
An application to explore the most popular Korean Dramas according to MyDramaList from 2015-2021
Home
Genre
Search for the genres of dramas by rank
Rank:
Submit
1 Move to Heaven Drama
1 Move to Heaven Family
1 Move to Heaven Life
2 Hospital Playlist Friendship
2 Hospital Playlist Life
2 Hospital Playlist Medical
2 Hospital Playlist Romance
3 Flower of Evil Crime
3 Flower of Evil Melodrama
3 Flower of Evil Romance
3 Flower of Evil Thriller


localhost:8088/search-result-content?keyword=5
DramApp
An application to explore the most popular Korean Dramas according to MyDramaList from 2015-2021
Home
Content
Search for the content rating of dramas by rank
Rank:
Submit
Move to Heaven 1 18+ Restricted (violence & profanity)
Hospital Playlist 2 15+ - Teens 15 or older
Flower of Evil 3 15+ - Teens 15 or older
Hospital Playlist 2 4 15+ - Teens 15 or older
My Mister 5 15+ - Teens 15 or older


localhost:8088/search-result-network?keyword=5
DramApp
An application to explore the most popular Korean Dramas according to MyDramaList from 2015-2021
Home
Network
Search for the network of emission of dramas by rank
Rank:
Submit
Move to Heaven 1 Netflix
Hospital Playlist 2 Netflix
Hospital Playlist 2 tvN
Flower of Evil 3 tvN
Hospital Playlist 2 4 Netflix
Hospital Playlist 2 4 tvN
My Mister 5 tvN


localhost:8088/search-result-epidur?keyword=5
DramApp
An application to explore the most popular Korean Dramas according to MyDramaList from 2015-2021
Home
Episodes and Duration
Search for the number of episodes and duration of dramas by rank
Rank:
Submit
Move to Heaven 1 Epsiodes: 10 Duration (min): 52
Hospital Playlist 2 Epsiodes: 12 Duration (min): 90
Flower of Evil 3 Epsiodes: 16 Duration (min): 70
Hospital Playlist 2 4 Epsiodes: 12 Duration (min): 100
My Mister 5 Epsiodes: 16 Duration (min): 77

References

1. CHUNG S. 2023. ‘K-Culture Is Here to Stay’. https://foreignpolicy.com/2023/05/31/korean-pop-culture-global-influence-united-states-parasite/
2. CHARUCHINDA, C. 2021. ‘Top 100 Korean Drama (MyDramaList)’. https://www.kaggle.com/datasets/chanoncharuchinda/top-100-korean-drama-mydramalist
3. Creative Commons. 2023. ‘CC0 1.0 DEED’. https://creativecommons.org/publicdomain/zero/1.0/

Although certain limitations in the content and quality of the work presented are acknowledged, it must be noted that this work, although relatively basic in nature, is dotted with great originality, most other works that made use of this particular dataset concentrate on establishing relations of likeness between the data, instead this work ambition was to investigate the crucial characteristics of the data and how these intertwined together could lead to potentially meaningful discoveries on the factors that have assured global success for korean dramas. It may not be present in the application itself, but beyond the queries in it more queries were also explored to obtain more of this meaningful data.