WHAT'S BEHIND NETFLIX ?

Database Design & Basic Exploratory Data Analysis with MySQL

CONTENT

About Me

Kahfi Rizky Kosasih
Undergraduate Mathematics student at Institut Teknologi Bandung

Summary :
A third-year undergraduate mathematics student and prospective data analyst who strives to pose and answer questions with quantitative-driven insights.

Kahfi Rizky Kosasih kahfirk kahfi.rk

CONCEPTUAL ERD

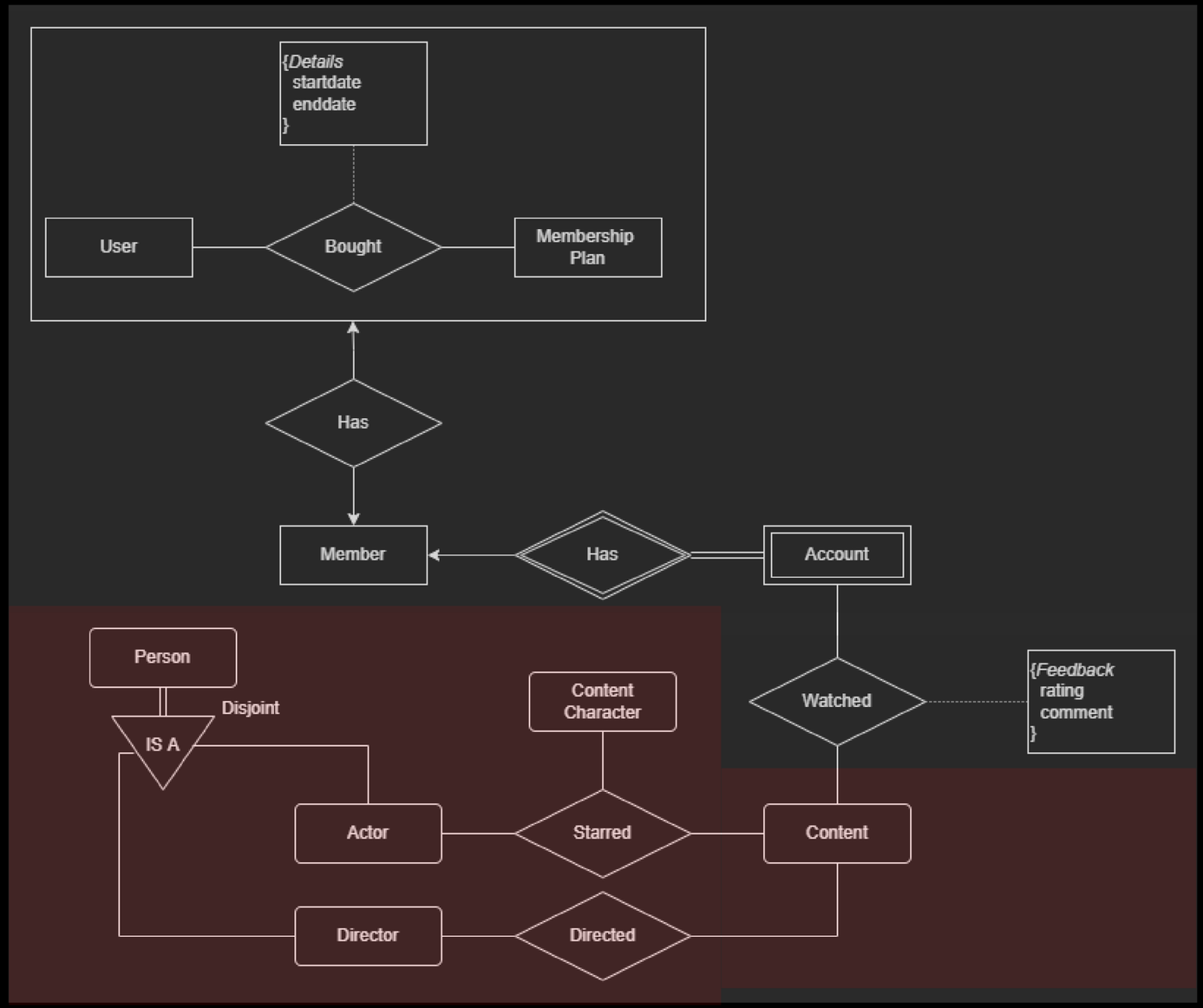
draw.io

Kembali ke Halaman Content

4

REKOMENDASI CONCEPTUAL ERD

Gambar : Conceptual Entity Relationship Diagram Aplikasi Streaming Netflix



Keterangan :

Data tidak tersedia

Data tersedia pada Kaggle

DATA UNDERSTANDING

kaggle

python™

Kembali ke Halaman Content

DATAFRAME NETFLIX MOVIE AND SHOW (DFI)

ID	TITLE	TYPE	DESCRIPTION	RELEASE_YEAR	AGE_CERTIFICATION	RUN TIME	GENRES	PRODUCTION_COUNTRIES	SEASONS	IMDB_ID	IMDB_SCORE	IMDB_VOTES	TMDB_POPULARITY	TMDB_SCORE
ts300399	Five Came Back: The Reference Films	SHOW	This collection includes 12 World War II-era p...	1945	TV-MA	48	['documentation']	['US']	1.0	NaN	NaN	NaN	0.600	NaN
tm84618	Taxi Driver	MOVIE	A mentally unstable Vietnam War veteran works ...	1976	R	113	['crime', 'drama']	['US']	NaN	tt0075314	8.3	795222.0	27.612	8.2
tm127384	Monty Python and the Holy Grail	MOVIE	King Arthur, accompanied by his squire, recrui...	1975	PG	91	['comedy', 'fantasy']	['GB']	NaN	tt0071853	8.2	530877.0	18.216	7.8
tm70993	Life of Brian	MOVIE	Brian Cohen is an average young Jewish man, bu...	1979	R	94	['comedy']	['GB']	NaN	tt0079470	8.0	392419.0	17.505	7.8
tm190788	The Exorcist	MOVIE	12-year-old Regan MacNeil begins to adapt an e...	1973	R	133	['horror']	['US']	NaN	tt0070047	8.1	391942.0	95.337	7.7

DFI CHARACTERISTIC & PREPROCESSING

NAMA KOLOM	PERSENTASE NULL (%)	BANYAK UNIQUE VALUE	TIPE DATA	KETERANGAN
id	0.000	5806	object	Tidak Ada
title	0.017	5751	object	Tidak Ada
type	0.000	2	object	['SHOW' 'MOVIE']
description	0.310	5785	object	Tidak Ada
release_year	0.000	67	int64	1945 s.d. 2022
age_certification	44.953	11	object	Tidak Ada
runtime	0.000	205	int64	0 s.d. 251
genres	0.000	1626	object	Tidak Ada
production_countries	0.000	449	object	Tidak Ada
seasons	64.743	23	float64	1.0 s.d. 42.0
imdb_id	7.647	5362	object	Tidak Ada
imdb_score	9.008	81	float64	1.5 s.d. 9.6
imdb_votes	9.283	3831	float64	5.0 s.d. 2268288.0
tmdb_popularity	1.619	4943	float64	0.009 s.d. 1823.374
tmdb_score	5.477	78	float64	0.5 s.d. 10.0

KEY POINTS

- Data memiliki dimensi 5806 row x 15 column, attribute id memiliki 5806 data unique sehingga cocok menjadi **Candidate Key.**
- Berdasarkan persentase null yang tinggi, attribute **age_certification** dan **seasons** lebih baik di **drop**.
- **Distribusi normal** pada attribute imdb_score dan tmdb_score. Oleh karena itu, pengisian data kosong akan dilakukan dengan **rataan** untuk setiap attribute tersebut.

DATAFRAME NETFLIX ACTOR & DIRECTOR (DF2)

PERSON_ID	ID	NAME	CHARACTER	ROLE
3748	tm84618	Robert De Niro	Travis Bickle	ACTOR
14658	tm84618	Jodie Foster	Iris Steensma	ACTOR
7064	tm84618	Albert Brooks	Tom	ACTOR
3739	tm84618	Harvey Keitel	Matthew 'Sport' Higgins	ACTOR
48933	tm84618	Cybill Shepherd	Betsy	ACTOR

DF2 CHARACTERISTIC & PREPROCESSING

NAMA KOLOM	PERSENTASE NULL (%)	BANYAK UNIQUE VALUE	TIPE DATA	KETERANGAN
person_id	0.000	53956	int64	7 s.d. 2371585
id	0.000	5434	object	Tidak Ada
name	0.000	53687	object	Tidak Ada
character	12.468	47125	object	Tidak Ada
role	0.000	2	object	['ACTOR' 'DIRECTOR']

BIG PROBLEM

- Data tidak mempunyai Candidate Key yang cukup kuat karena seorang aktor dapat berperan sebagai beberapa character dalam suatu konten. Selain itu juga, unique value pada character cukup rendah (61%) dan terdapat null value. Oleh karena itu, akan dibangun attribute baru yaitu char_id dengan tipe integer dan Auto Increment agar dapat mendefinisikan aktor dengan baik.

KEY POINTS

- Data memiliki dimensi 77213 row x 4 column, kualitas dataframe sangatlah buruk. Bahkan dataframe ini belum pasti berada dalam bentuk 1 NF.
- Salah satu penyebab persentase null tinggi pada attribute character adalah terdapat DIRECTOR yang tidak memiliki attribute character.
- Akan dilaksanakan dekomposisi dan peninjauan lebih lanjut pada bagian normalisasi.

NORMALIZATION

Kembali ke Halaman Content

DFI BERADA DALAM NORMAL FORM 2 NF

CONTENT ID	TITLE	TYPE	DESCRIPTION	RELEASE_YEAR	AGE_CERTIFICATION	RUN TIME	GENRES	PRODUCTION_COUNTRIES	SEASONS	IMDB_ID	IMDB_SCORE	IMDB_VOTES	TMDB_POPULARITY	TMDB_SCORE
ts300399	Five Came Back: The Reference Films	SHOW	This collection includes 12 World War II-era p...	1945	TV-MA	48	['documentation']	['US']	1.0	NaN	NaN	NaN	0.600	NaN
tm84618	Taxi Driver	MOVIE	A mentally unstable Vietnam War veteran works ...	1976	R	113	['crime', 'drama']	['US']	NaN	tt0075314	8.3	795222.0	27.612	8.2
tm127384	Monty Python and the Holy Grail	MOVIE	King Arthur, accompanied by his squire, recrui...	1975	PG	91	['comedy', 'fantasy']	['GB']	NaN	tt0071853	8.2	530877.0	18.216	7.8
tm70993	Life of Brian	MOVIE	Brian Cohen is an average young Jewish man, bu...	1979	R	94	['comedy']	['GB']	NaN	tt0079470	8.0	392419.0	17.505	7.8
tm190788	The Exorcist	MOVIE	12-year-old Regan MacNeil begins to adapt an e...	1973	R	133	['horror']	['US']	NaN	tt0070047	8.1	391942.0	95.337	7.7

DF1 = { contentid ,title, type, description, release_ year, age_ certification, runtime, genre, production_ countries, imdb_id, imdb_score,imdb_votes, tmdb _popularity, tmdb_score }

FD : {(1) contentid → Seluruh Attribute,
(2) imdb_ id → imdb_score, imdb_votes, tmdb_popularity, tmdb_score }

Penjelasan :

Dapat dipilih contentid sebagai Candidate Key. Perhatikan bahwa DF1 bukan BCNF karena FD (2) bukanlah Super Key dan DF1 bukan 3NF karena terdapat trasitive depedency non-prime attribute. Oleh karena itu, *normal form* tertinggi DF1 hanyalah 2NF. Salah satu solusi agar DF1 mencapai BCNF ialah melaksanakan dekomposisi setiap FD sebagai relasi baru. Akan tetapi, kita menyadari bahwa pada *real-case* data terdapat null value pada imdb_id. Oleh karena itu, akan digunakan contentid sebagai Candidate Key relasi baru.

NORMALISASI DFI DALAM NORMAL FORM BCNF

Tabel R1 : {**contentid**, title, type description, release_year, age_certification, runtime, genres, production_countries, seasons}

CONTENT ID	TITLE	TYPE	DESCRIPTION	RELEASE_YEAR	AGE_CERTIFICATION	RUN TIME	GENRES	PRODUCTION_COUNTRIES	SEASONS
ts300399	Five Came Back: The Reference Films	SHOW	This collection includes 12 World War II-era p...	1945	TV-MA	48	['documentation']	['US']	1.0
tm84618	Taxi Driver	MOVIE	A mentally unstable Vietnam War veteran works ...	1976	R	113	['crime', 'drama']	['US']	NaN
tm127384	Monty Python and the Holy Grail	MOVIE	King Arthur, accompanied by his squire, recrui...	1975	PG	91	['comedy', 'fantasy']	['GB']	NaN
tm70993	Life of Brian	MOVIE	Brian Cohen is an average young Jewish man, bu...	1979	R	94	['comedy']	['GB']	NaN
tm190788	The Exorcist	MOVIE	12-year-old Regan MacNeil begins to adapt an e...	1973	R	133	['horror']	['US']	NaN

FD R1 : {contentid → title, type, description, release_year, age_certification, runtime, genres, production_countries, seasons }

Tabel R2 : {**contentid**, imdb_id, imdb_score, imdb_votes, tmdb_popularity, tmdb_score}

CONTENT ID	IMDB_ID	IMDB_SCORE	IMDB_VOTES	TMDB_POPULARITY	TMDB_SCORE
ts300399	NaN	NaN	NaN	0.600	NaN
tm84618	tt0075314	8.3	795222.0	27.612	8.2
tm127384	tt0071853	8.2	530877.0	18.216	7.8
tm70993	tt0079470	8.0	392419.0	17.505	7.8
tm190788	tt0070047	8.1	391942.0	95.337	7.7

FD R2 : {contentid → imdb_id, imdb_score, imdb_votes, tmdb_popularity, tmdb_score }

DF2 BERADA DALAM NORMAL FORM I NF

PERSON_ID	CONTENTID	NAME	CHAR_ID	CHARACTER	ROLE
3748	tm84618	Robert De Niro	1	Travis Bickle	ACTOR
14658	tm84618	Jodie Foster	2	Iris Steensma	ACTOR
7064	tm84618	Albert Brooks	3	Tom	ACTOR
3739	tm84618	Harvey Keitel	4	Matthew 'Sport' Higgins	ACTOR
48933	tm84618	Cybill Shepherd	5	Betsy	ACTOR

DF2 = { person_id, contentid, name, char_id, character, role}

FD : {
(1) person_id → name
(2) person_id, contentid, char_id → role
(3) char_id → character
}

Penjelasan :

Dapat dipilih kombinaasi person_id, char_id, dan contentid sebagai Candidate Key. Perhatikan bahwa DF2 bukan BCNF karena LHS FD (1) dan (3) bukanlah Super Key, DF2 bukan 3NF karena terdapat trasitive depedency non-prime attribute, DF2 bukan 2NF karena RHS FD (1) dan (3) partial dependent berturut-turut terhadap person_id dan char_id . Oleh karena itu, *normal form* tertinggi DF2 hanyalah 1NF. Agar DF2 mencapai normal form tertinggi, akan dideklarasikan setiap FD yang partial dependent sebagai FD pada relasi baru.

NORMALISASI DF2 DALAM NORMAL FORM BCNF

Tabel R3 : {**person_id,** name}

PERSON_ID	NAME
3748	Robert De Niro
14658	Jodie Foster
7064	Albert Brooks
3739	Harvey Keitel
48933	Cybill Shepherd

FD R3 : {person_id → name }

Tabel R4 : {**person_id, contentid, char_id**, role}

PERSON_ID	CONTENTID	CHAR_ID	ROLE
3748	tm84618	1	ACTOR
14658	tm84618	2	ACTOR
7064	tm84618	3	ACTOR
3739	tm84618	4	ACTOR
48933	tm84618	5	ACTOR

FD R4: {person_id, contentid, char_id → role }

Tabel R5 : {**char_id,** character}

CHAR_ID	CHARACTER
1	Travis Bickle
2	Iris Steensma
3	Tom
4	Matthew 'Sport' Higgins
5	Betsy

FD R5 : {char_id → character }

Notes :

- Ingat bahwa kolom role sangatlah redundant. Oleh karena itu, konversi conceptual ERD pada relational DB design nanti akan dilaksanakan dengan memecah spesialisasi tabel person sehingga terdapat dua tabel baru, yaitu tabel actor dan director.

RELATIONAL DB DESIGN

MySQL

python™

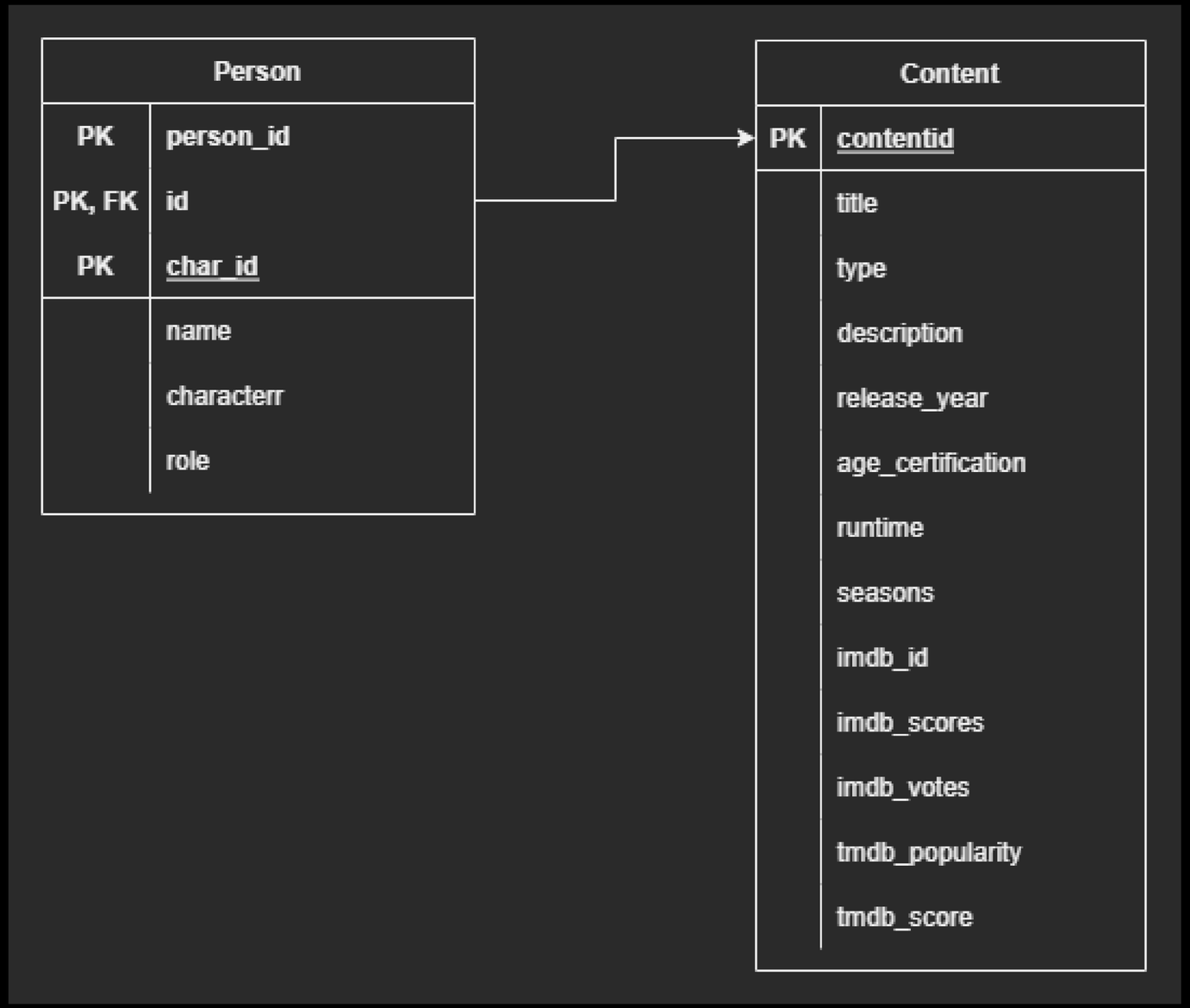
phpMyAdmin

draw.io

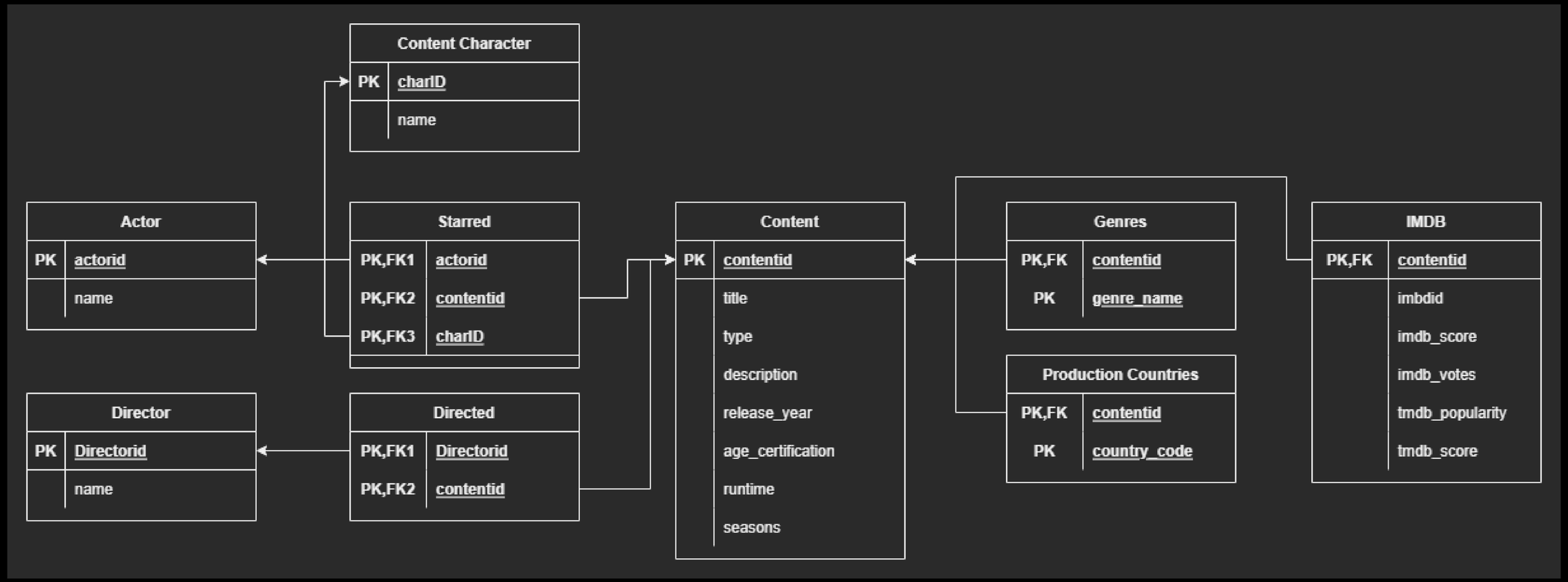
Kembali ke Halaman Content

RELATIONAL DB DESIGN BEFORE OPTIMIZATION

Database Design disamping adalah hasil konversi secara langsung dataset yang tersedia pada kaggle. Oleh karena itu, akan dilaksanakan transformasi pada database sesuai normalisasi pada bagian sebelumnya. Proses pembangunan normalisasi tabel akan dilaksanakan dengan python dan implementasi dengan phpmyadmin serta mysql.

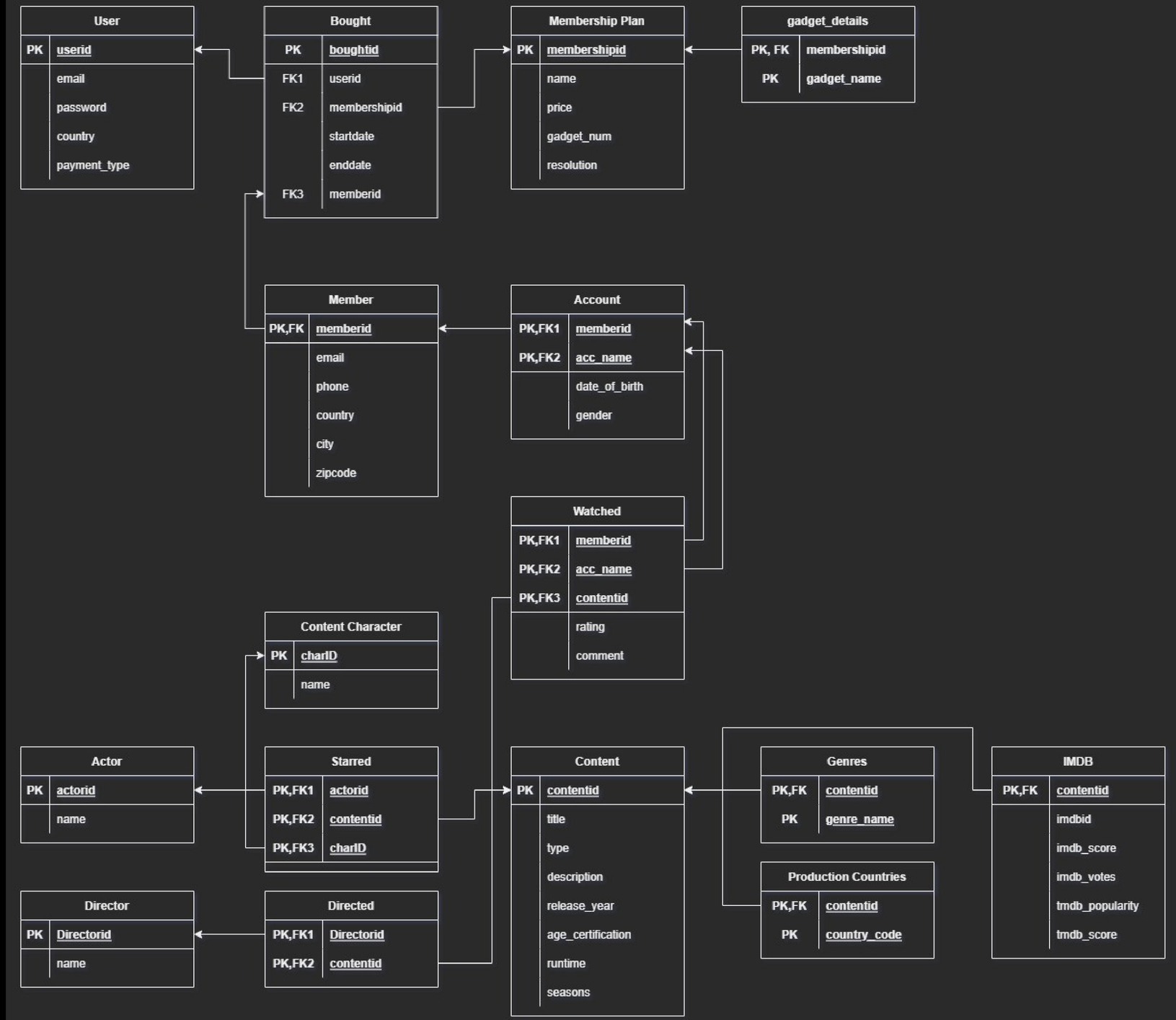


RELATIONAL DB DESIGN AFTER OPTIMIZATION



REKOMENDASI SKEMA DATABASE NETFLIX

Database Schema disamping merupakan gambaran suatu skema lengkap yang dapat direalisasikan sebagai media penyimpanan data untuk aplikasi streaming.



BASIC EDA WITH MYSQL

Kembali ke Halaman Content

GENRE KONTEN TERBAIK

Kemampuan Netflix dalam membangun konten bergenre dokumentasi, sejarah, dan perang **sangatlah baik.**

```
select genre_name, avg(imdb_score), avg(tmdb_score),
       avg(imdb_score) + avg(tmdb_score) as overall_score,
       count(*) as total_content
from imdb natural join genres
group by genre_name order by overall_score DESC  ;
```

Syntax: Content Terbaik berdasarkan Overall Score

Tabel: Content Terbaik berdasarkan *Overall Score*

GENRE_NAME	AVG (IMDB_SCORE)	AVG (TMDB_SCORE)	OVERALL_SCORE	TOTAL_CONTENT
history	7.14979	7.19214	14.34193	233
war	7.08828	7.08716	14.17544	149
documentation	7.04282	7.04297	14.08579	910
animation	6.72880	7.31482	14.04362	665
scifi	6.58018	7.16158	13.74176	587
music	6.63139	7.10538	13.73677	238

GENRE KONTEN TERPOPULER

(1) Drama, 2901 Content. (2) Comedy, 2269 Content. (3) Thriller, 1178 Content

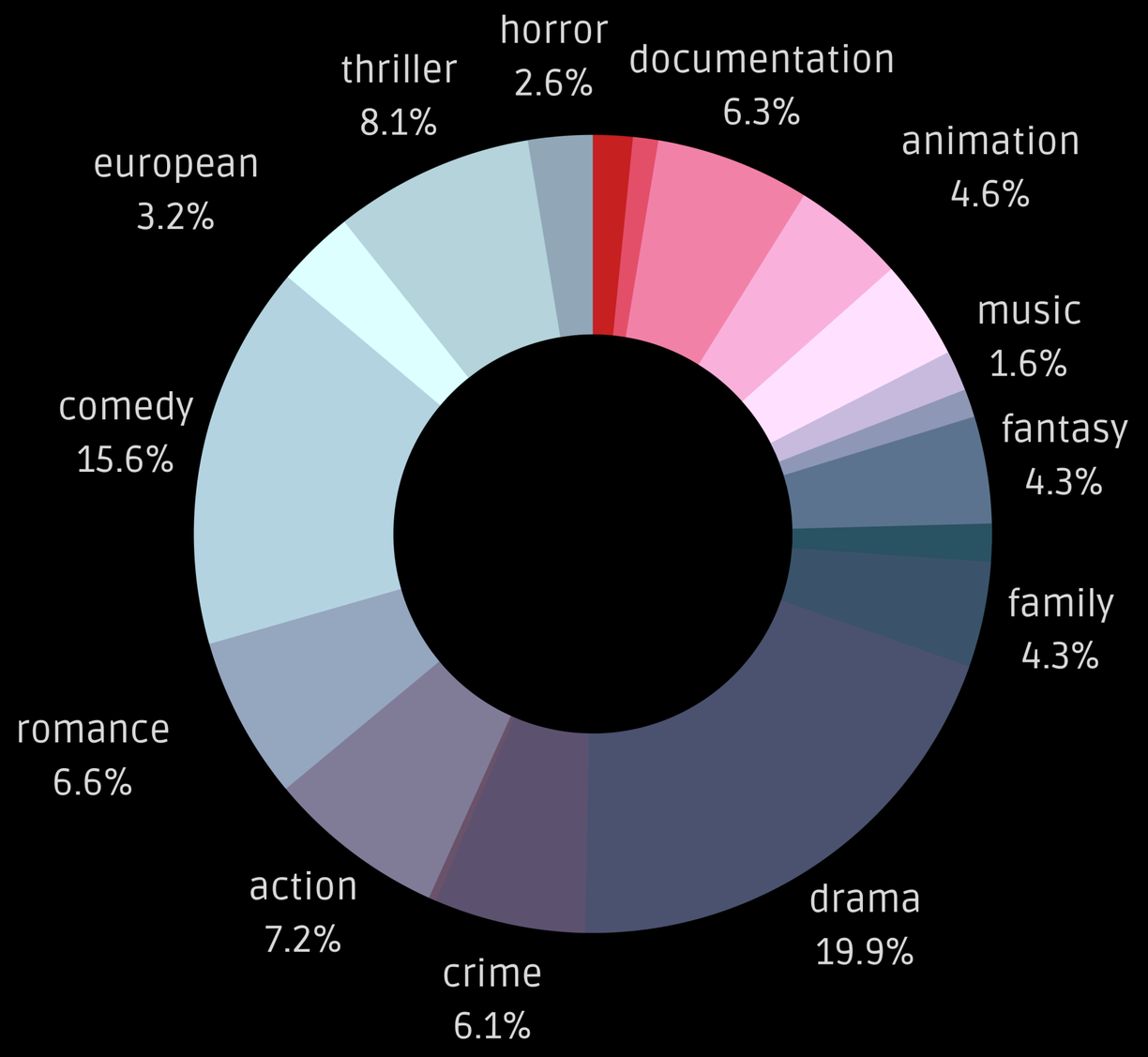
```
create view fav_genres as
select genre_name, count(*) as total_content
from genres
Group by genre_name
Having count(*) >= (select count(*)
                    from genres
                    group by genre_name
                    order by count(*) desc
                    limit 2,1)
order by count(*) desc;

Select * from fav_genres;
```

Syntax: Top 3 Genre Terpopuler

```
select genre_name,
       count(*) as total_content
from imdb natural join genres
group by genre_name order by total_content DESC  ;
```

Syntax: Distribusi setiap genre



SHOW LEBIH BAIK DARIPADA MOVIE

Para pengguna Netflix mengapresiasi **snack-bite** show dibandingkan durasi panjang suatu movie.

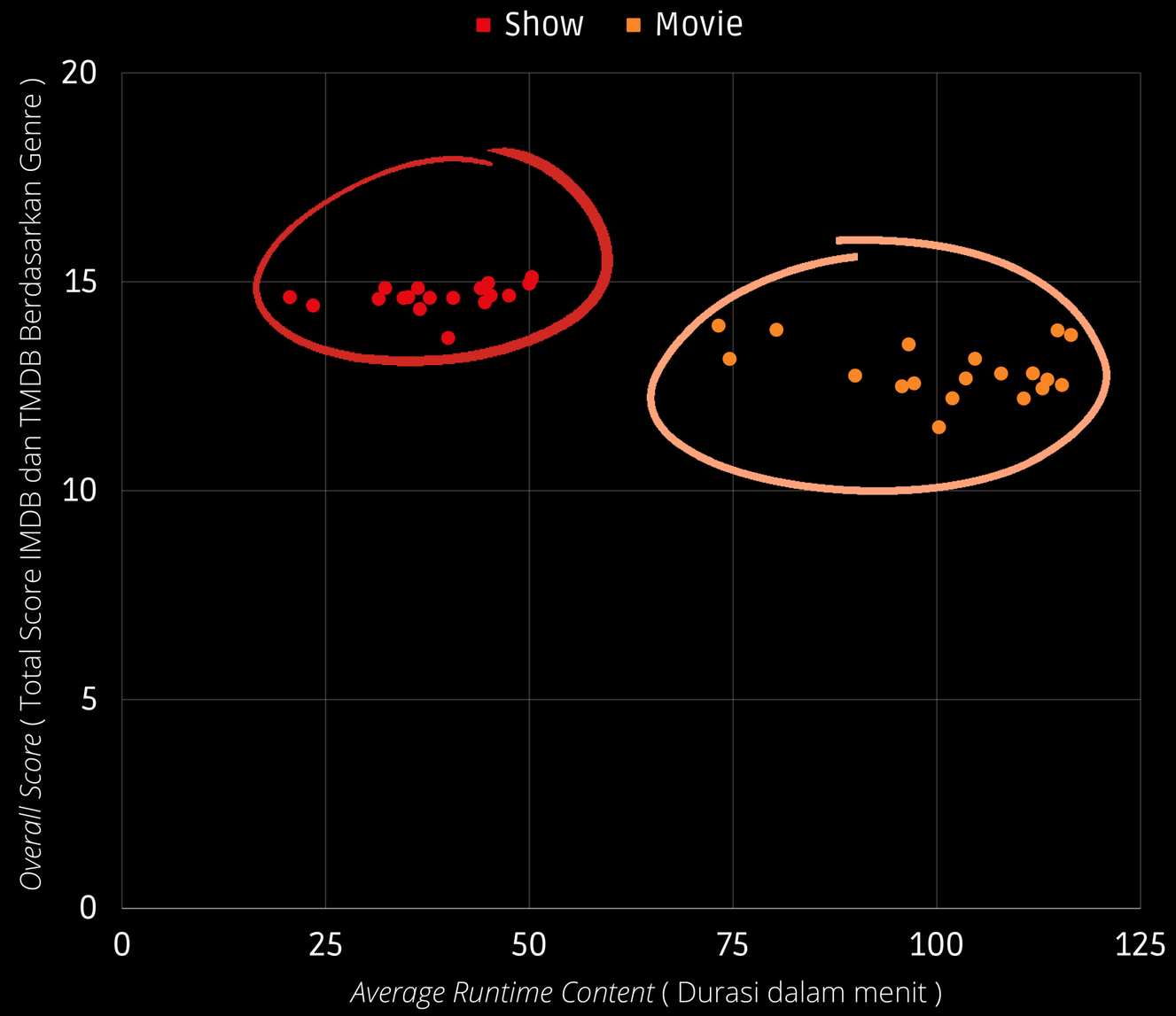
Spearman's Test :
rs = -0.68924 , p (2-tailed) = 0.

Conclusion : Terdapat hubungan **signifikan** antara durasi konten dengan kualitas konten

```
select m.genre_name ,m.type, avg(imdb_score)+ avg(tmdb_score) as overall_score, avg(m.runtime)
from (select c.contentid,c.runtime, c.type, title, genre_name
      from content as c
      inner join genres as g
      ON c.contentid = g.contentid) as m
LEFT JOIN imdb as i on m.contentid =i.contentid
group by m.genre_name,m.type ORDER BY type,genre_name;
```

Syntax: Keterkaitan Runtime terhadap kualitas setiap genre

Grafik: Scatter Plot *Average Runtime* terhadap *Overall Score* Berdasarkan Genre



PENCARIAN DATA DUPLIKAT

```
create view duplicate_content_character as (
select * from content_character
natural join starred
where contentid in (select contentid from content_character
                    natural join starred
                    group by contentid,actorid
                    having count(*)>1) and actorid in (select actorid from content_character
                                                       natural join starred
                                                       group by contentid,actorid
                                                       having count(*)>1));
create view char_duplicate as
(select charid from duplicate_content_character as d1
where actorid in ( select actorid from duplicate_content_character as d1
                   where name is null)
                   and contentid in ( select contentid from duplicate_content_character as d1
                                      where name is null)
                   and name is null)
union
(select d1.charid from duplicate_content_character as d1,duplicate_content_character as d2
where d1.actorid=d2.actorid and d1.contentid=d2.contentid and d1.charid<> d2.charid and d1.name = d2.name);

Delete from starred where charid in (select * from char_duplicate);
Delete from content_character where charid in (select * from char_duplicate);
```

Syntax: Penghapusan duplikasi data pada tabel content_character dan starred

Akan dilaksanakan penyaringan data pada setiap aktor dalam suatu konten. Pertama-tama akan dicari semua aktor yang mendapatkan lebih dari satu peran pada suatu konten.

Setelah itu, dilaksanakan peninjauan apakah salah satu nama character dalam content tersebut memiliki value null.

Selain itu juga, ditinjau apakah terdapat duplikasi (character memiliki nama yang sama) dalam konten yang sama.

Hapus data tersebut

THANKS FOR READING

Sekian yang dapat penulis sampaikan. Tentu terdapat banyak kekurangan dalam project ini sehingga penulis berharap kritik & saran dari para pembaca.