TOP 15000 ANIME

Kolegij : Skladišta i rudarenja podataka

Student: Rahela Štos

Uvod

Poslovna inteligencija u kontekstu anime industrije objedinjuje metodologije i alate za pretvorbu sirovih podataka o anime serijama i filmovima u informacije ključne za strategije distribucije, produkcije i marketinga

CILJ: simulirati proces od prikupljanja podataka do izrade sustava za potporu odlučivanju, koristeći skup podataka o 15.000 anime-a.

Odabir skupa podataka

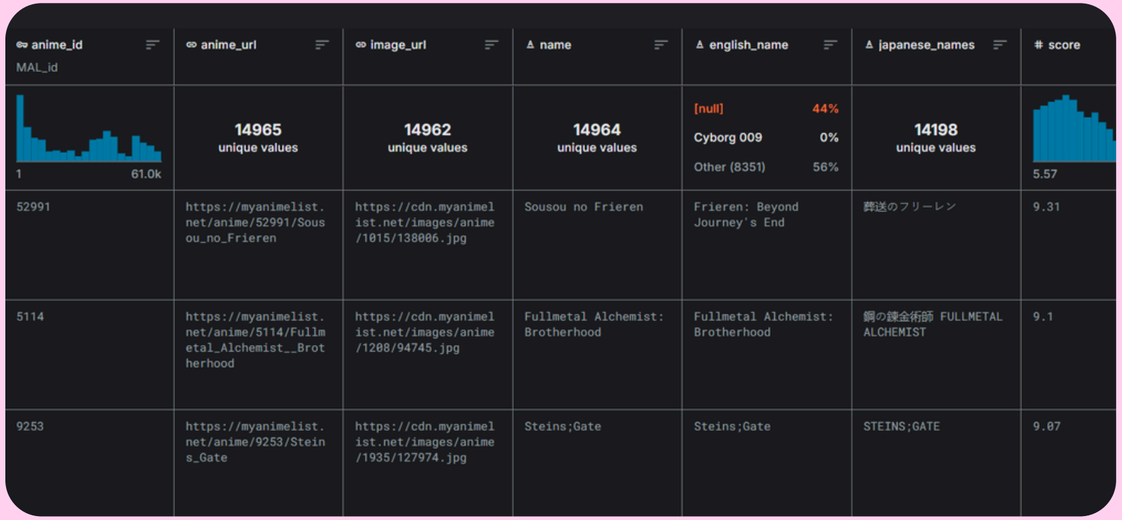
Podaci su preuzeti iz CVS datoteke sa linka: https://www.kaggle.com/datasets/quanthan/top-15000-ranked-anime-dataset-update-to-32025?resource=download

Dataset sadži:

Struktuirane podatke: Ocjene (1-10), broj članova, žanrove, tip (TV serija, film,OVA), broj epizoda, studio

Vremenske podatke: Godina premijere

Python, Apache Spark, MySQL, Matplotlib/Seaborn



anime_id MAL_id	anime_url	image_url	name	english_name	japanese_names	score
1 … 61.0k	14965 unique values	14962 unique values	14964 unique values	[null] 44% Cyborg 009 0% Other (8351) 56%	14198 unique values	5.57
52991	https://myanimelist.net/anime/52991/Sousou_no_Frieren	https://cdn.myanimelist.net/images/anime/1015/138006.jpg	Sousou no Frieren	Frieren: Beyond Journey's End	葬送のフリーレン	9.31
5114	https://myanimelist.net/anime/5114/Fullmetal_Alchemist__Brotherhood	https://cdn.myanimelist.net/images/anime/1208/94745.jpg	Fullmetal Alchemist: Brotherhood	Fullmetal Alchemist: Brotherhood	鋼の錬金術師 FULLMETAL ALCHEMIST	9.1
9253	https://myanimelist.net/anime/9253/Steins_Gate	https://cdn.myanimelist.net/images/anime/1935/127974.jpg	Steins;Gate	Steins;Gate	STEINS;GATE	9.07

Poslovni problem

U anime industriji jedni od ključnih izazova uključuju identificiranje žanrova s najvećom publikom, utvrđivanje utjecanja studija na kvalitetu te analiza odnosa između broja epizoda i popularnosti. Poslovna integracija omogućuje odgovore na ova pitanja kroz transformaciju podataka u vizualne metričke izvještaje.

Osnovna analiza podataka

CSV datoteka -> Top 15,000 Ranked Anime Dataset

korišteni alati

Python, Pandas biblioteka

```
import pandas as pd
PATH = r"C:\Users\rahel\Downloads\archive\top_anime_dataset.csv"
```

Učitavanje i pregled podataka

data_first_2000 = pd.read_csv(PATH, nrows=2000)

print(data_first_2000.head())

```
[5 rows x 22 columns]

dimenzije skupa podataka::
(15000, 22)
```

Prikazuje: anime_id, image_url, name, popularity, favourites, scored_by..

print(data.shape)

```
 5 redova:
ime_id                                anime_url                                          image_url                                          name  ... popularity favorites  scored_by  m
 52991  https://myanimelist.net/anime/52991/Sousou_no_...  https://cdn.myanimelist.net/images/anime/1015/...                Sousou no Frieren  ...        160     63200     603520  1
  5114  https://myanimelist.net/anime/5114/Fullmetal_A...  https://cdn.myanimelist.net/images/anime/1208/...  Fullmetal Alchemist: Brotherhood  ...          3    231928    2196353  3
  9253     https://myanimelist.net/anime/9253/Steins_Gate  https://cdn.myanimelist.net/images/anime/1935/...                     Steins;Gate  ...         14    194507    1449844  2
 60022  https://myanimelist.net/anime/60022/One_Piece_...  https://cdn.myanimelist.net/images/anime/1455/...              One Piece Fan Letter  ...       2350      1970      68977
 38524  https://myanimelist.net/anime/38524/Shingeki_n...  https://cdn.myanimelist.net/images/anime/1517/...  Shingeki no Kyojin Season 3 Part 2  ...         21     60500    1671010  2
```

Raznolikost podataka

print(data.isna().sum())

```
broj nedostajućih vrijednosti po stupcima:
anime_id              0
anime_url             0
image_url             0
name                  0
english_name       6645
japanese_names       47
score                 0
genres             1603
synopsis            473
type                  1
episodes            115
premiered         10314
producers          5432
studios            2383
source                0
duration              0
rating               68
rank               3079
popularity            0
favorites             0
scored_by             0
members               0
dtype: int64
```

print(data.nunique())

```
broj jedinstvenih vrijednosti po stupcima:
anime_id          14965
anime_url         14965
image_url         14962
name              14964
english_name       8214
japanese_names    14197
score               344
genres              865
synopsis          14368
type                  9
episodes            198
premiered           240
producers          4451
studios            1421
source               17
duration            311
rating                6
rank               9597
popularity        13501
favorites          1926
scored_by          8782
members           10954
dtype: int64
```

print(data.dtypes)

```
tipovi podataka po stupcima:
anime_id           int64
anime_url         object
image_url         object
name              object
english_name      object
japanese_names    object
score            float64
genres            object
synopsis          object
type              object
episodes         float64
premiered         object
producers         object
studios           object
source            object
duration          object
rating            object
rank             float64
popularity         int64
favorites          int64
scored_by          int64
members            int64
dtype: object
```

Detaljna distribucija

```
for column in data:
    print(f"\n--- {column} ---")
    print(data[column].value_counts())
```

```
detaljna distribucija vrijednosti po stupcima:

--- anime_id ---
anime_id
60356    2
59276    2
59497    2
59390    2
59750    2
        ..
8664     1
8526     1
27467    1
25627    1
52991    1
Name: count, Length: 14965, dtype: int64
```

```
--- name ---
name
Tian Guan Cifu Short Film                                  2
Shishigari                                                 2
One Piece: Gyojin Tou-hen                                  2
Fate/Grand Order: Fujimaru Ritsuka wa Wakaranai Season 2   2
Mayonaka Punch Short Anime                                 2
                                                         ..
Slime Boukenki: Umi da, Yeah!                              1
Honey Tokyo                                                1
Yoligongju Loopy                                           1
"Aesop" no Ohanashi yori: Ushi to Kaeru, Yokubatta Inu     1
Makkuramori no Uta                                         1
Name: count, Length: 14964, dtype: int64
```

```
--- genres ---
genres
Comedy                                                  1157
Hentai                                                  1010
Action, Adventure, Fantasy                               450
Fantasy                                                  420
Slice of Life                                            400
                                                        ...
Gourmet, Suspense                                          1
Action, Comedy, Fantasy, Romance, Sci-Fi, Ecchi            1
Comedy, Erotica                                            1
Drama, Horror, Erotica                                     1
Boys Love, Slice of Life, Sports                           1
Name: count, Length: 865, dtype: int64
```

Izrada relacijskog modela i baze podataka

cilj: strukturirani okvir za upravljanje anime podacima

alati

MySQL DBMS

Python sa bibliotekama Pandas i SQLAlchemy

definicija entiteta , veze , implementacija , punjenje podataka

```
engine = create_engine(DB_CONNECTION)
print("Spojeno na bazu")
```

Izrada baze podataka

kod briše bazu ako ona postoji
zatim kreira novu bazu anime_db

```
conn.execute(text("DROP DATABASE IF EXISTS anime_db;"))
conn.execute(text("CREATE DATABASE anime_db;"))
```

kreiranje samih tablica

7 tablica

anime

anime_urls

anime_genre

anime_producer

anime_studio

anime_type

anime_stats

tablica *anime*

```
main_cols = ['anime_id', 'name', 'english_name', 'japanese_names',
   'score', 'type', 'episodes', 'premiered', 'source',
   'duration', 'rating', 'rank', 'popularity', 'favorites',
   'scored_by', 'members', 'synopsis']
df[main_cols].to_sql('anime', engine, if_exists='replace', index=False)
```

tablica *anime_url*

```
df[['anime_id', 'anime_url', 'image_url']].to_sql
( 'anime_urls',  engine,
  if_exists='replace',
  index=False)
print("url gotova")
```

tablica *anime_genre*

```
genres_data = []
for _, row in df.iterrows():
   if pd.notna(row['genres']):
      for genre in str(row['genres']).split(','):
         genres_data.append({
            'anime_id': row['anime_id'],
            'genre': genre.strip()
         })

pd.DataFrame(genres_data).to_sql('anime_genre', engine,
if_exists='replace', index=False)
print("zanr gotov")
```

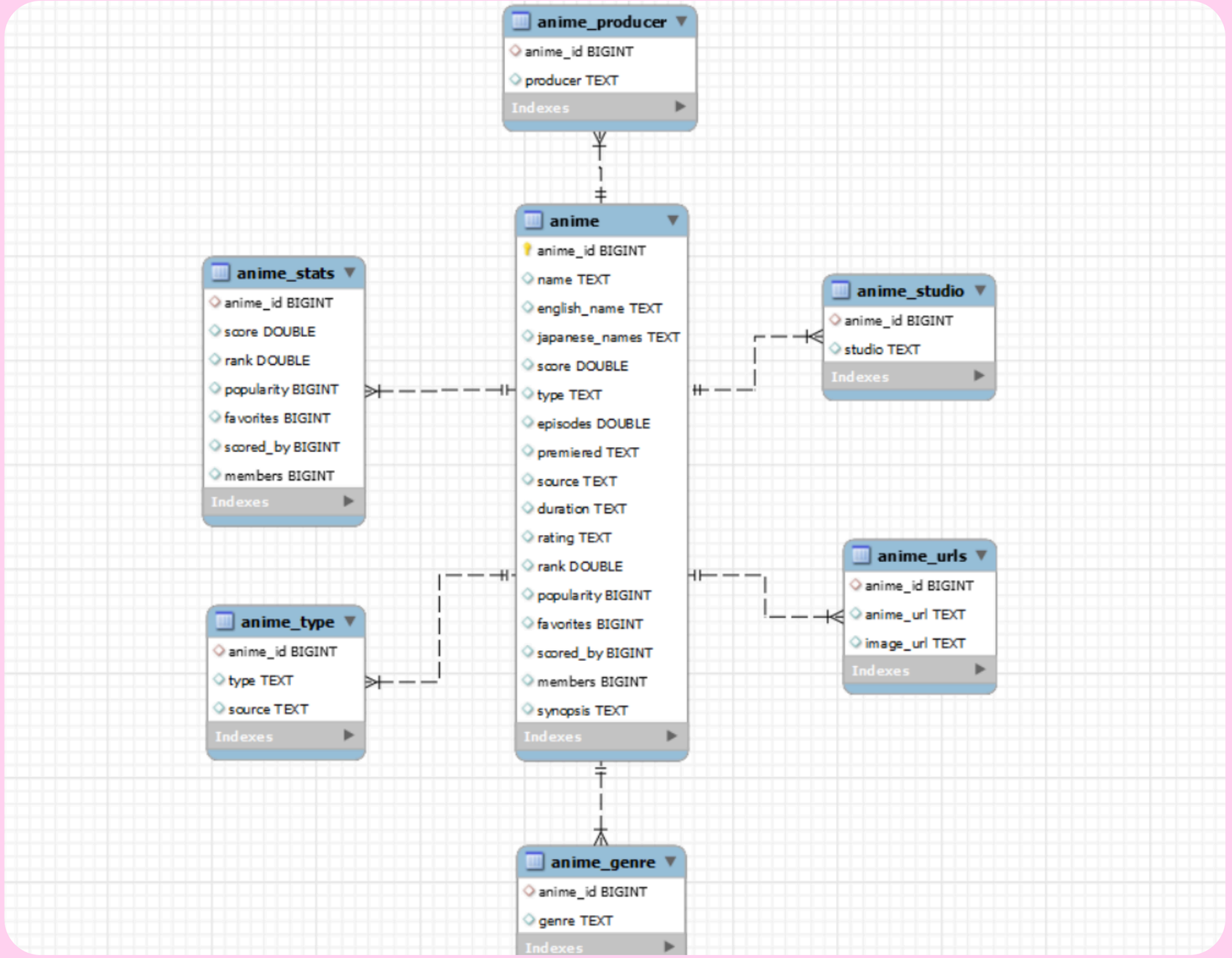
many-to-many veza

Foreign key

```
conn.execute(text("""
    ALTER TABLE anime_urls
    ADD CONSTRAINT fk_urls_anime
    FOREIGN KEY (anime_id) REFERENCES anime(anime_id)
    ON DELETE CASCADE
"""))
```

povezivanje svih tablica sa glavnom (anime)

on delete cascade briše tablicu


anime_producer
anime_id BIGINT
producer TEXT
Indexes
anime
anime_id BIGINT
name TEXT
english_name TEXT
japanese_names TEXT
score DOUBLE
type TEXT
episodes DOUBLE
premiered TEXT
source TEXT
duration TEXT
rating TEXT
rank DOUBLE
popularity BIGINT
favorites BIGINT
scored_by BIGINT
members BIGINT
synopsis TEXT
Indexes
anime_stats
anime_id BIGINT
score DOUBLE
rank DOUBLE
popularity BIGINT
favorites BIGINT
scored_by BIGINT
members BIGINT
Indexes
anime_studio
anime_id BIGINT
studio TEXT
Indexes
anime_urls
anime_id BIGINT
anime_url TEXT
image_url TEXT
Indexes
anime_type
anime_id BIGINT
type TEXT
source TEXT
Indexes
anime_genre
anime_id BIGINT
genre TEXT
Indexes

Izrada dimenzijskih tablica

10 tablica

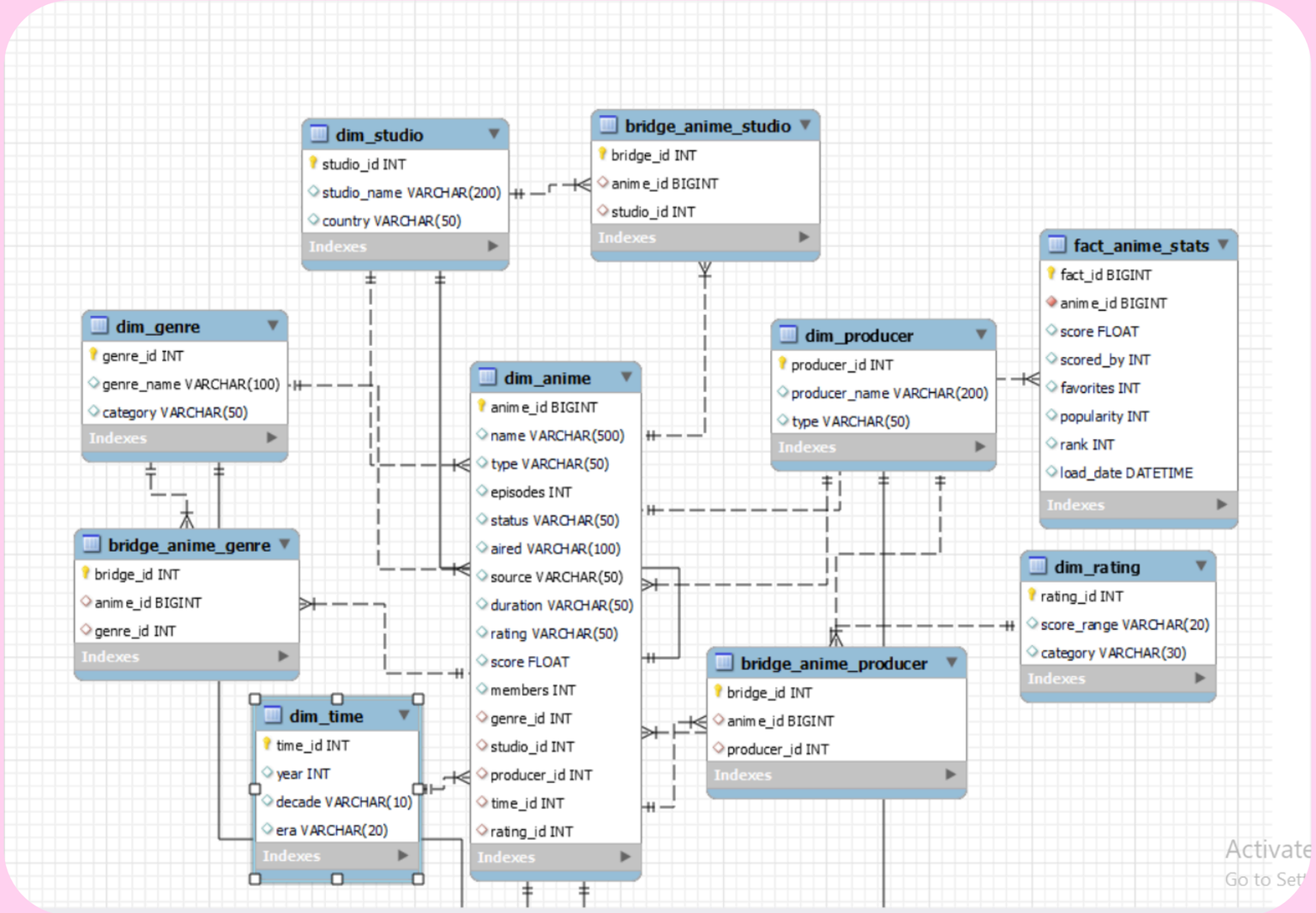
6 običnih tablica

3 bridge tablice

1 fact tablica

alati:

SQLAlchemy , Python, MySQL



Fact tablica

```python
class FactAnimeStats(Base):
    __tablename__ = 'fact_anime_stats'
    fact_id = Column(BigInteger, primary_key=True, autoincrement=True)


    anime_id = Column(BigInteger, ForeignKey('dim_anime.anime_id'))


    score = Column(Float)
    members = Column(Integer)
    popularity_rank = Column(Integer)
    favorites = Column(Integer)
    scored_by = Column(Integer)
    rank = Column(Integer)
```

score je prosječna ocjena korisnika

members je broj pratitelja

popularity_rank je rang popularnosti

favorites je broj korisnika koji su dodali u omiljene

scored_by je broj korisnika koji su ocijenili

rank je ukupni rang na platformi

Jednostavne dimenzije/glavna

dim_anime, glavna tablica sadrži anime_id, name, type, episodes, score, members...

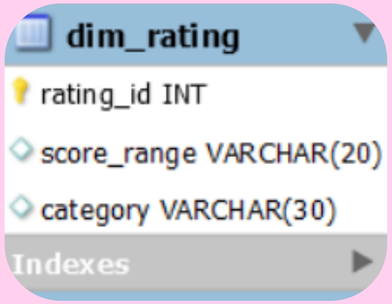
dim_rating, sadrži rating_id,score_range,category

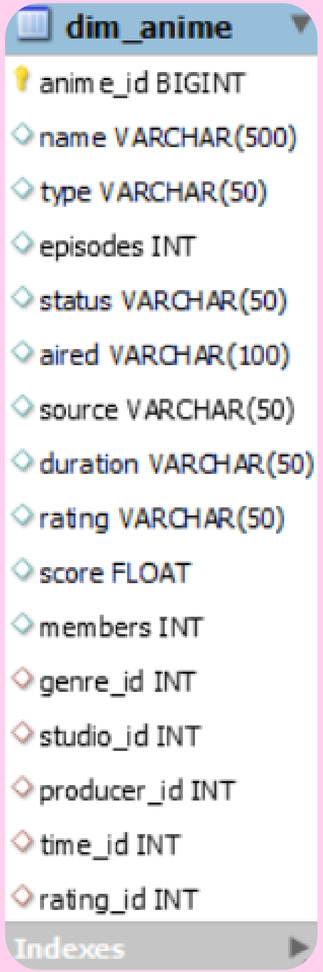
dim_time, sadrži time_id, year, decade, era

```python
class DimAnime(Base):
    __tablename__ = 'dim_anime'
    anime_id = Column(BigInteger, primary_key=True)
    name = Column(String(255))
    type = Column(String(50))
    episodes = Column(Integer)
    score = Column(Float)
    members = Column(Integer)
```

```python
class DimTime(Base):
    __tablename__ = 'dim_time'
    time_id = Column(Integer, primary_key=True, autoincrement=True)
    year = Column(Integer)
    decade = Column(String(10))
    era = Column(String(20))
```

```python
class DimRating(Base):
    __tablename__ = 'dim_rating'
    rating_id = Column(Integer, primary_key=True)
    score_range = Column(String(20))
    category = Column(String(30))
```





Složene dimenzije

anime_genre, povezuje anime i žanrove

anime_studio, povezuje anime i studio

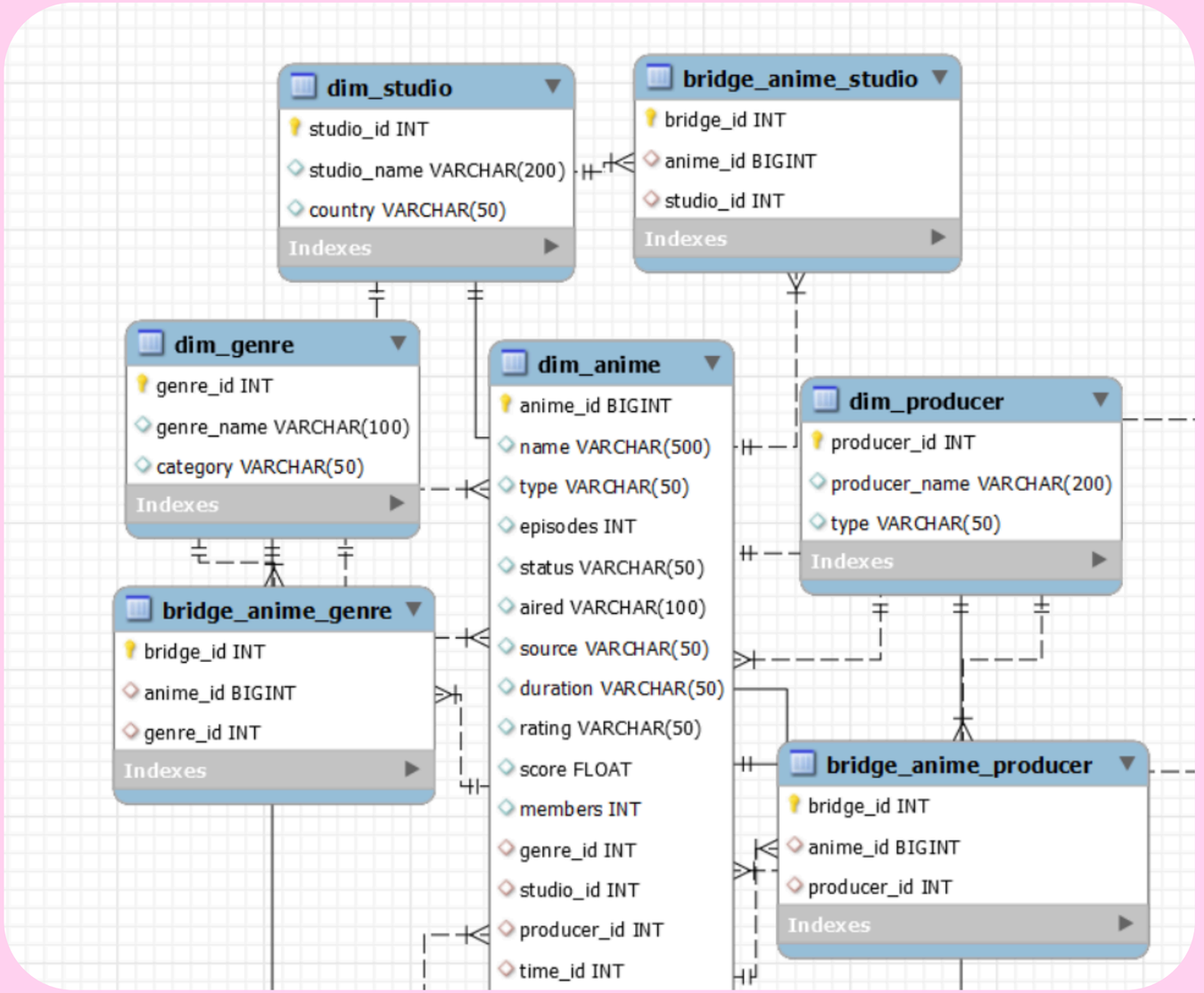
anime_producer , povezuje anime i producere

primary key

Zašto su bridge tablice važne?

```python
class AnimeGenre(Base):
    __tablename__ = 'anime_genre'
    anime_id = Column(BigInteger, ForeignKey('dim_anime.anime_id'), primary_key=True)
    genre_id = Column(Integer, ForeignKey('dim_genre.genre_id'), primary_key=True)

class AnimeStudio(Base):
    __tablename__ = 'anime_studio'
    anime_id = Column(BigInteger, ForeignKey('dim_anime.anime_id'), primary_key=True)
    studio_id = Column(Integer, ForeignKey('dim_studio.studio_id'), primary_key=True)

class AnimeProducer(Base):
    __tablename__ = 'anime_producer'
    anime_id = Column(BigInteger, ForeignKey('dim_anime.anime_id'), primary_key=True)
    producer_id = Column(Integer, ForeignKey('dim_producer.producer_id'), primary_key=True)
```

```
class DimGenre(Base):
    __tablename__ = 'dim_genre'
    genre_id = Column(Integer, primary_key=True,
autoincrement=True)
    genre_name = Column(String(100), unique=True)

class DimStudio(Base):
    __tablename__ = 'dim_studio'
    studio_id = Column(Integer, primary_key=True,
autoincrement=True)
    studio_name = Column(String(255), unique=True)

class DimProducer(Base):
    __tablename__ = 'dim_producer'
    producer_id = Column(Integer, primary_key=True,
autoincrement=True)
    producer_name = Column(String(255), unique=True)
```



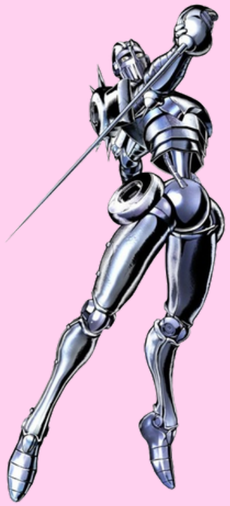
ETL proces

alati

Python sa bibliotekama Pandas i SQLAlchemy ,

PySpark

prvi korak je učitavanje CSV datoteke i spajanje sa sparkom



```
csv_path = r"C:\Users\rahel\Downloads\archive\top_anime_dataset.csv"
print(f"Učitavam CSV iz: {csv_path}")
df = spark.read.option("header", True).option("inferSchema", True).csv(csv_path)
print("učitano, kolone", df.columns)
df.show(5)
```

Popunjavanje tablica

dim_anime sadrži osnovne informacije o svakom anime filmi/seriji, anime_id, naziv, tip, broj_ep, ocjena,broj_članova..

dim_genre sadrži popis svih anime žanrova, a to su action, drama,comedy

dim_studio sadrži nazive studija, Studio Ghibli, Toei Animation, studio_id

dim_producer sadrži informacije o producerima, producer_id

dim_rating kategorizira anime filmove/serije po ocjenama, rating_id, score_range, category

dim_time omogućuje vremensku analizu, time_id

select * from dim_anime limit 5;

anime_id	name	type	episodes	status	aired	source	duration	rating	score	members	genre_id	studio_id	producer_id	time_id	rating_id
1	Cowboy Bebop	TV	26	Finished		Original	24 min per ep	R - 17+ (violence & profanity)	8.75	1946300	2	7	8	20	8
5	Cowboy Bebop: Tengoku no Tobira	Movie	1	Finished		Original	1 hr 55 min	R - 17+ (violence & profanity)	8.38	392992	2	2	56	4	8
6	Trigun	TV	26	Finished		Manga	24 min per ep	PG-13 - Teens 13 or older	8.22	793518	2	1	41	20	8
7	Witch Hunter Robin	TV	26	Finished		Original	25 min per ep	PG-13 - Teens 13 or older	7.24	121710	2	7	8	30	7
8	Bouken Ou Beet	TV	52	Finished		Manga	23 min per ep	PG - Children	6.93	16151	2	4	4	14	6
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

select * from dim_genre limit 5;

genre_id	genre_name	category
1	Adventure	General
2	Action	General
3	Drama	General
4	Comedy	General
5	Award Winning	General
NULL	NULL	NULL

select * from dim_rating limit 5;

rating_id	score_range	category
5	5.0-6.0	Poor
6	6.0-7.0	Average
7	7.0-8.0	Good
8	8.0-9.0	Excellent
9	9.0-10.0	Masterpiece
NULL	NULL	NULL

Data Validation

Svaka tablica i broj zapisa koji sadrži

tablica	broj_zapisa
dim_anime	14965
dim_genre	21
dim_studio	836
dim_producer	774
dim_rating	5
dim_time	65
fact_anime_stats	15000

DESCRIBE dim_anime;

DESCRIBE dim_genre;

DESCRIBE dim_studio;

DESCRIBE dim_producer;

Field	Type	Null	Key	Default	Extra
anime_id	bigint	NO	PRI	NULL	auto_increment
name	varchar(500)	YES		NULL	
type	varchar(50)	YES		NULL	
episodes	int	YES		NULL	
status	varchar(50)	YES		NULL	
aired	varchar(100)	YES		NULL	
source	varchar(50)	YES		NULL	
duration	varchar(50)	YES		NULL	
rating	varchar(50)	YES		NULL	
score	float	YES		NULL	
members	int	YES		NULL	
genre_id	int	YES	MUL	NULL	
studio_id	int	YES	MUL	NULL	
produce...	int	YES	MUL	NULL	
time_id	int	YES	MUL	NULL	
rating_id	int	YES	MUL	NULL	

Field	Type	Null	Key	Default	Extra
genre_id	int	NO	PRI	NULL	auto_increment
genre_name	varchar(100)	YES	UNI	NULL	
category	varchar(50)	YES		NULL	

Field	Type	Null	Key	Default	Extra
studio_id	int	NO	PRI	NULL	auto_increment
studio_name	varchar(200)	YES	UNI	NULL	
country	varchar(50)	YES		NULL	

Field	Type	Null	Key	Default	Extra
producer_id	int	NO	PRI	NULL	auto_increment
producer_name	varchar(200)	YES	UNI	NULL	
type	varchar(50)	YES		NULL	

Referencijalni integritet

	nije_spojeno_time
▶	0

	nije_spojeno_studio
▶	0

	nije_spojeno_producer
▶	0

	nije_spojeno_rating
▶	0

Analiziranje tablica

prikaz top 10 anime žanra sa brojem anime serija/filmova u tom žanru

genre_name	broj_animea
Action	4124
Comedy	3164
Adventure	1306
Hentai	1007
Drama	989
Fantasy	658
Slice of Life	448
Sports	308
Romance	243
Sci-Fi	230

prikaz prvih deset studija i broj anime-a kojeg su napravili.

studio_name	broj_animea
Toei Animation	716
Sunrise	531
J.C.Staff	382
Madhouse	342
Production I.G	317
TMS Entertainment	299
Studio Deen	268
Pierrot	258
OLM	257
A-1 Pictures	238

prikaz top 15 anime serija/filmova po ocjeni

anime_id	name	ocjena	type
52991	Sousou no Frieren	9.31	1035677
5114	Fullmetal Alchemist: Brotherhood	9.1	3483268
9253	Steins;Gate	9.07	2667979
60022	One Piece Fan Letter	9.06	91278
38524	Shingeki no Kyojin Season 3 Part 2	9.05	2407861
28977	Gintama°	9.05	656687
39486	Gintama: The Final	9.04	163478
11061	Hunter x Hunter (2011)	9.03	2981476
9969	Gintama'	9.02	580364
15417	Gintama': Enchousen	9.02	338134
820	Ginga Eiyuu Densetsu	9.01	337614
41467	Bleach: Sennen Kessen-hen	9	616144
43608	Kaguya-sama wa Kokurasetai: Ult...	8.99	1008941
34096	Gintama.	8.98	328137
42938	Fruits Basket: The Final	8.96	515201

Grafička analiza

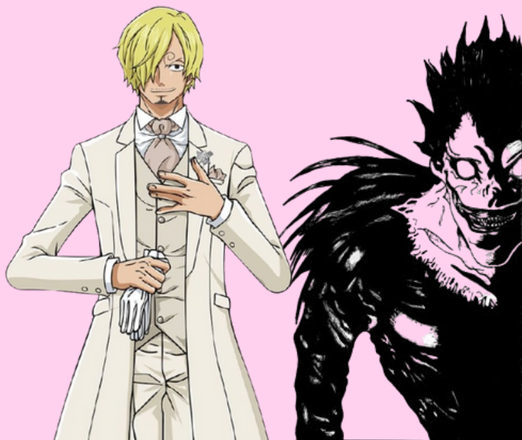
alati:

Python sa bibliotekama Pandas, Matplotlib i Seaborn

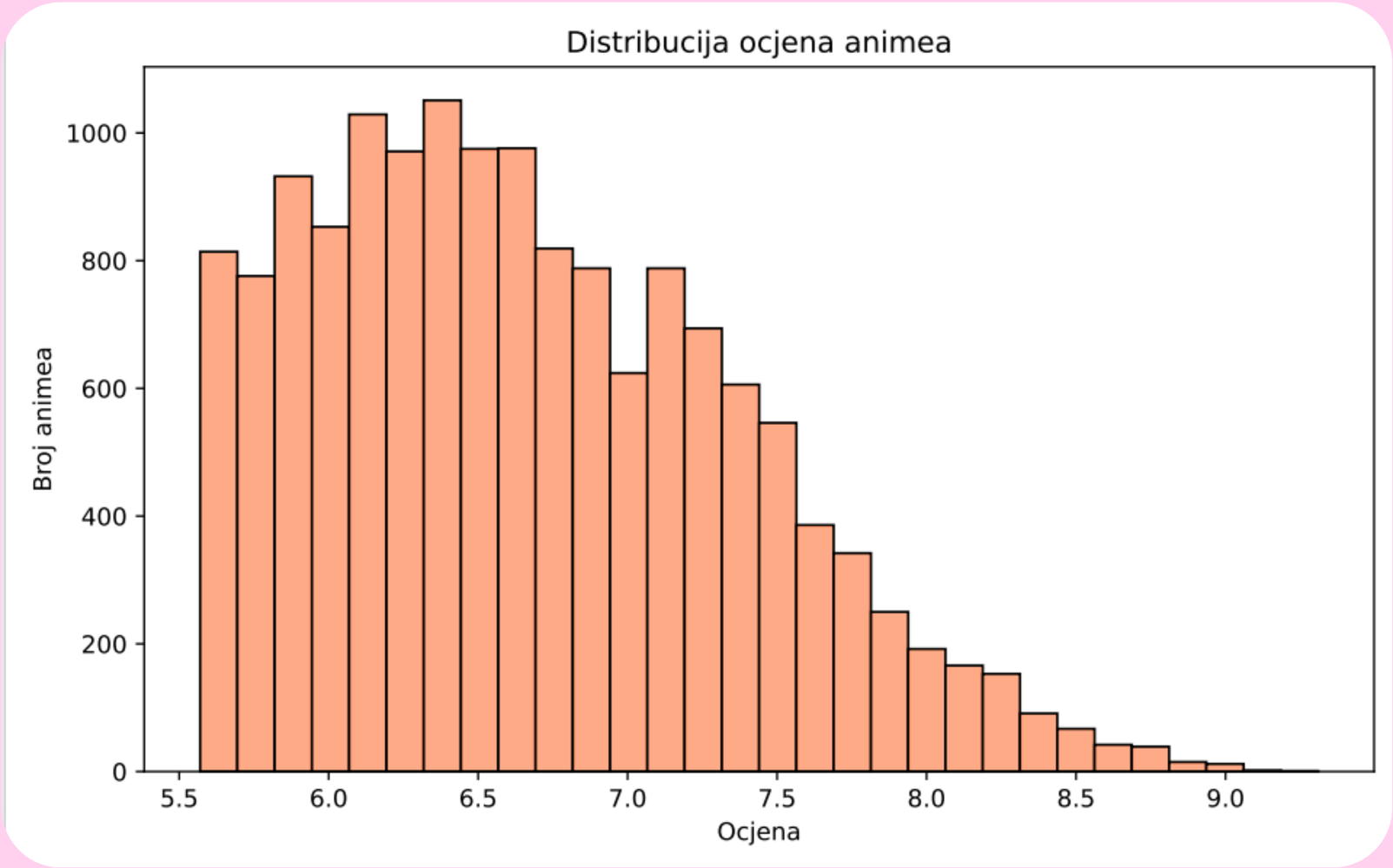
Python modul os

kod koji stvara direktorij za grafove

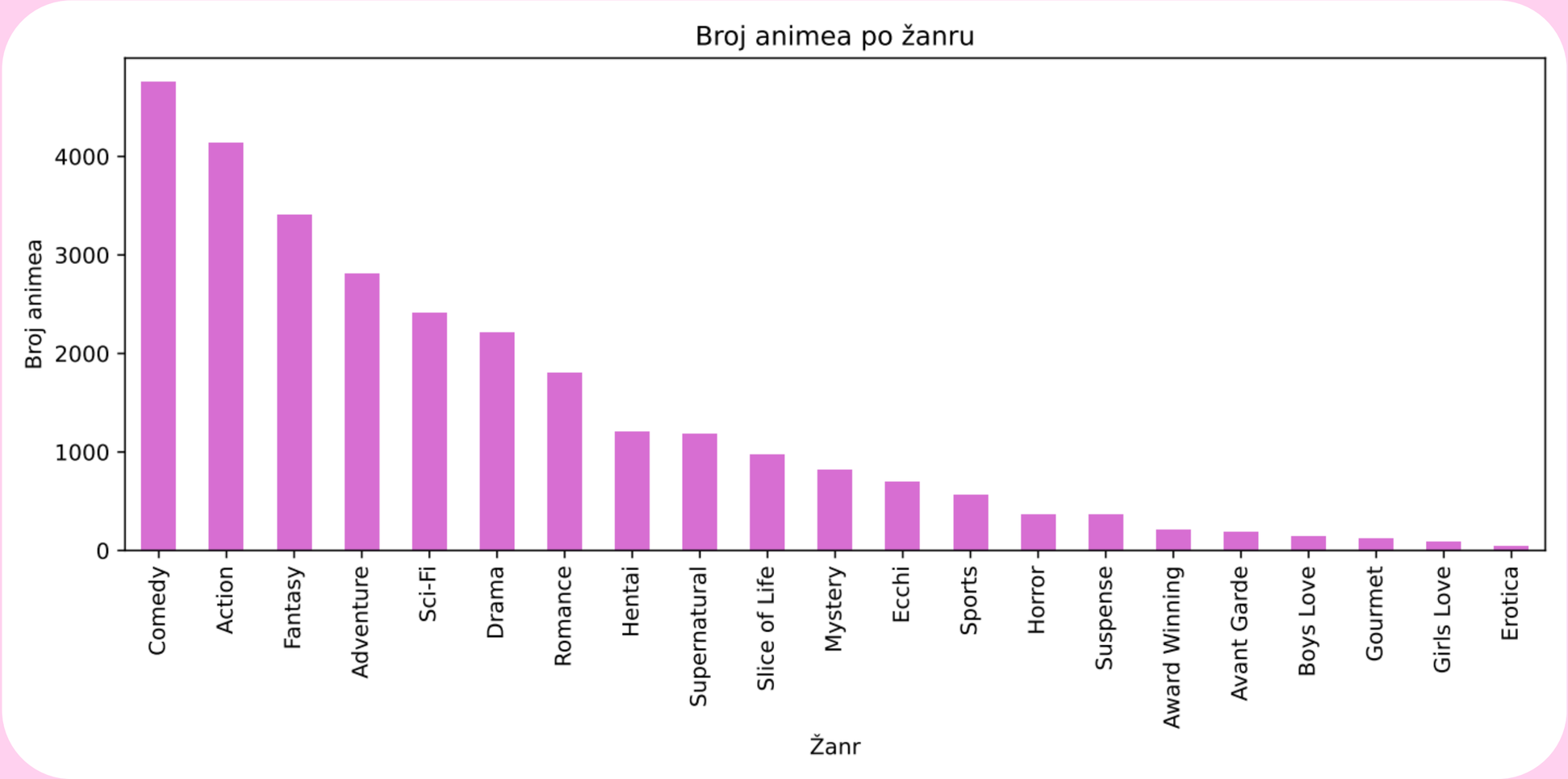
```
plot_dir = './anime_grafovii'
os.makedirs(plot_dir, exist_ok=True)
```



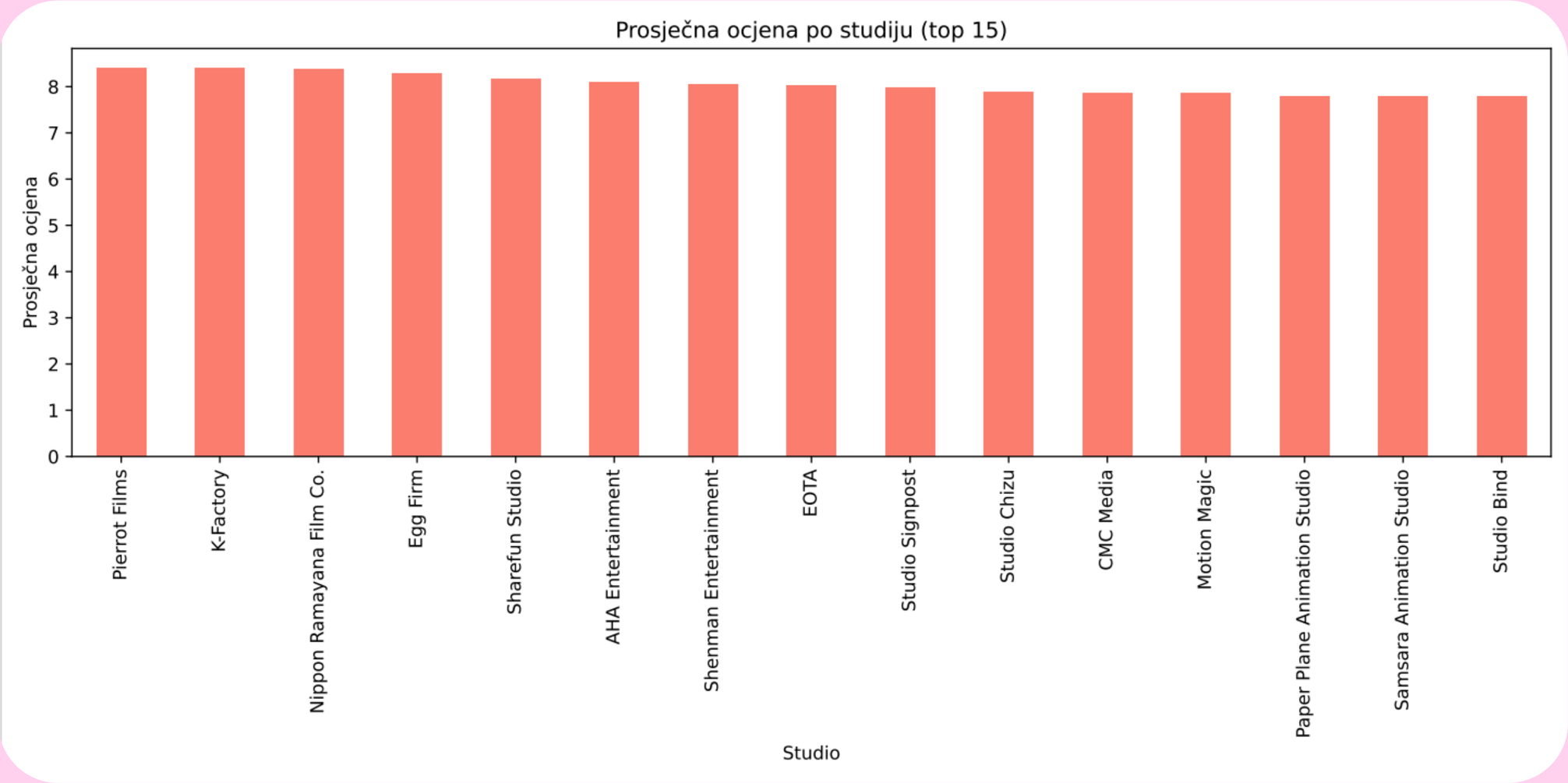
OLAP operacija



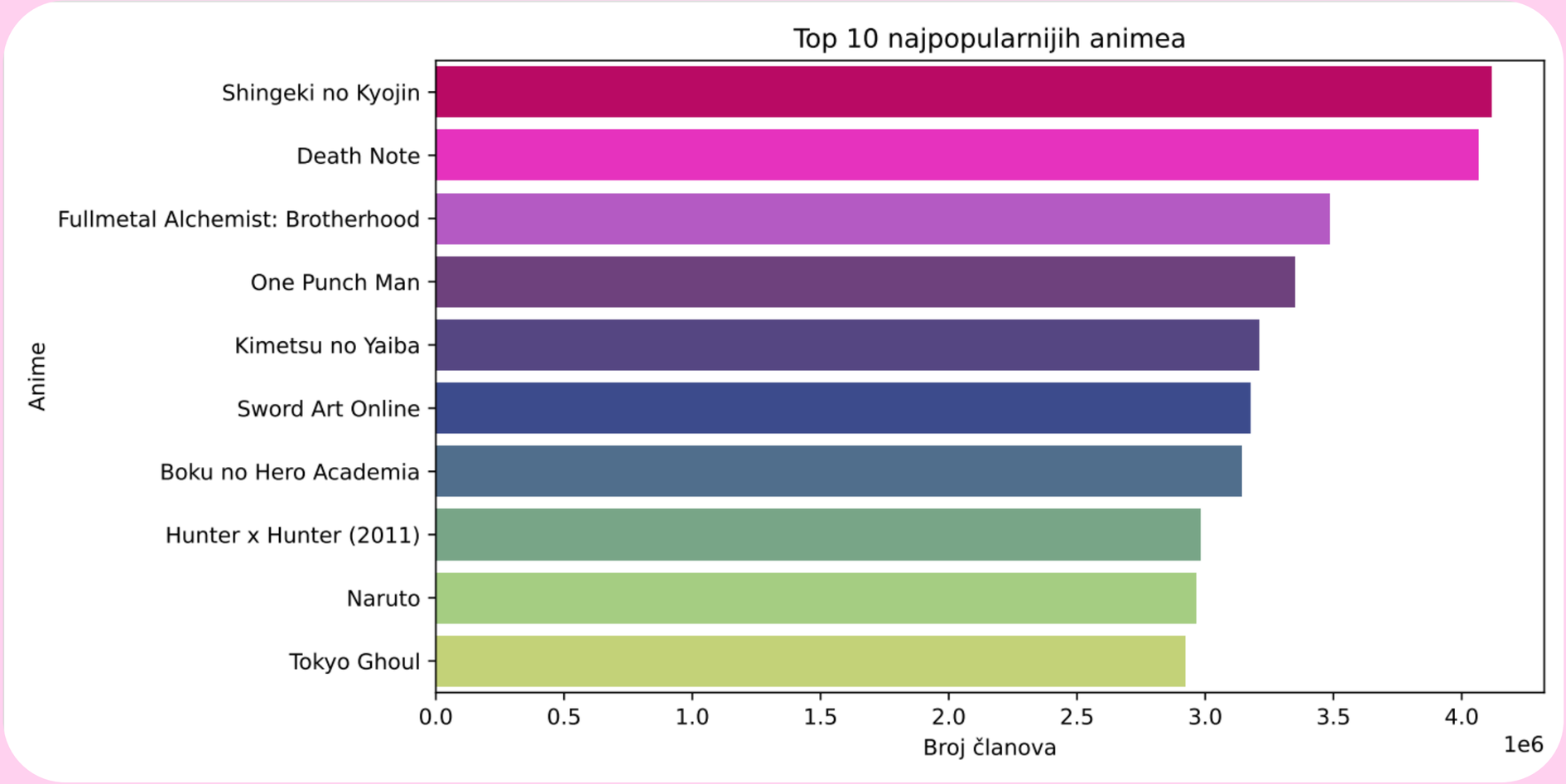
Drill down operacija



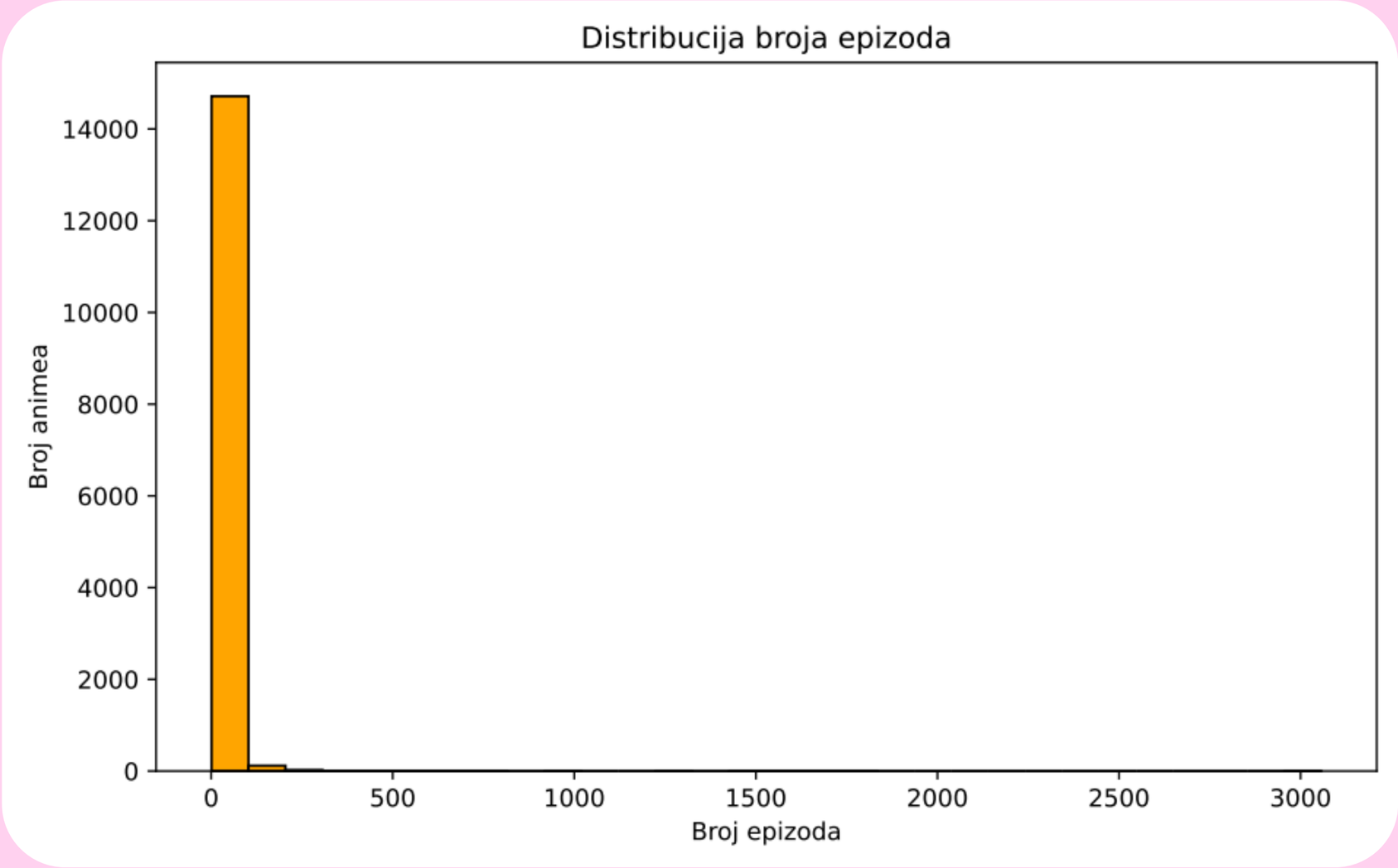
Roll-up operacija



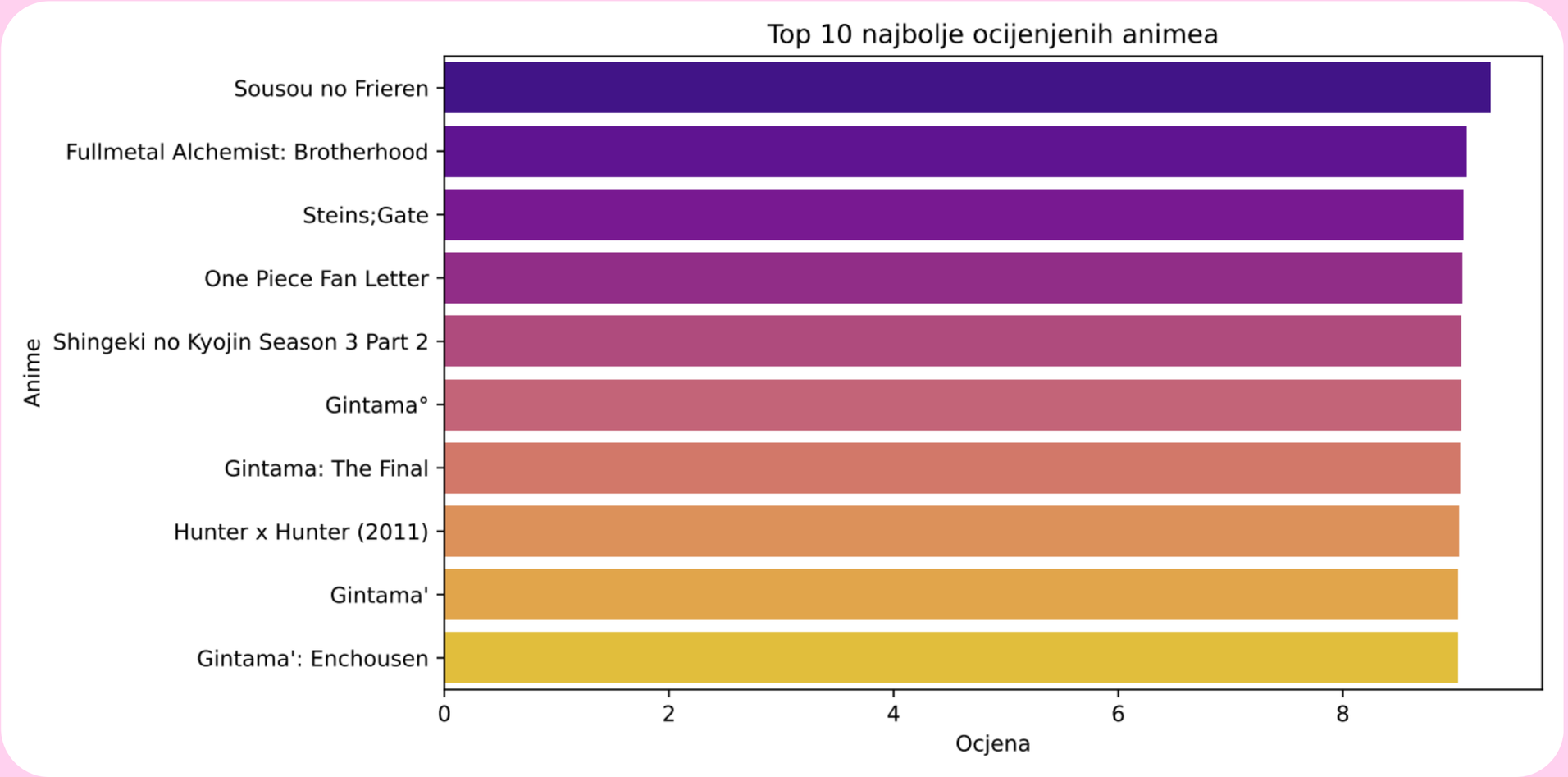
Slice i dice operacija



OLAP operacija



Slice i sort operacija



Zaključak

Kroz ovaj projekt napravljen je proces poslovne inteligencije u anime industriji. Napravljeno je prikupljanje i čišćenje podataka, izradnja naprednog skladišta i grafova.
Koristio se Python ekosustav, Apache Spark i mySQL i to je omogućilo da se skup podataka 15000 anime naslova tranforsmira u strukturirani data warehouse temeljen na star shemi sa deset tablica, šest glavnih, 3 bridge i jednom fact tablicom. Zbog toga omogućen mi je uvid u ključne trendove anime industrije, koji žanr je najpopularniji, koji anime-studio je najuspješniji, i koji anime naslovi su najbolje ocjenjeni i najviše viđeni.