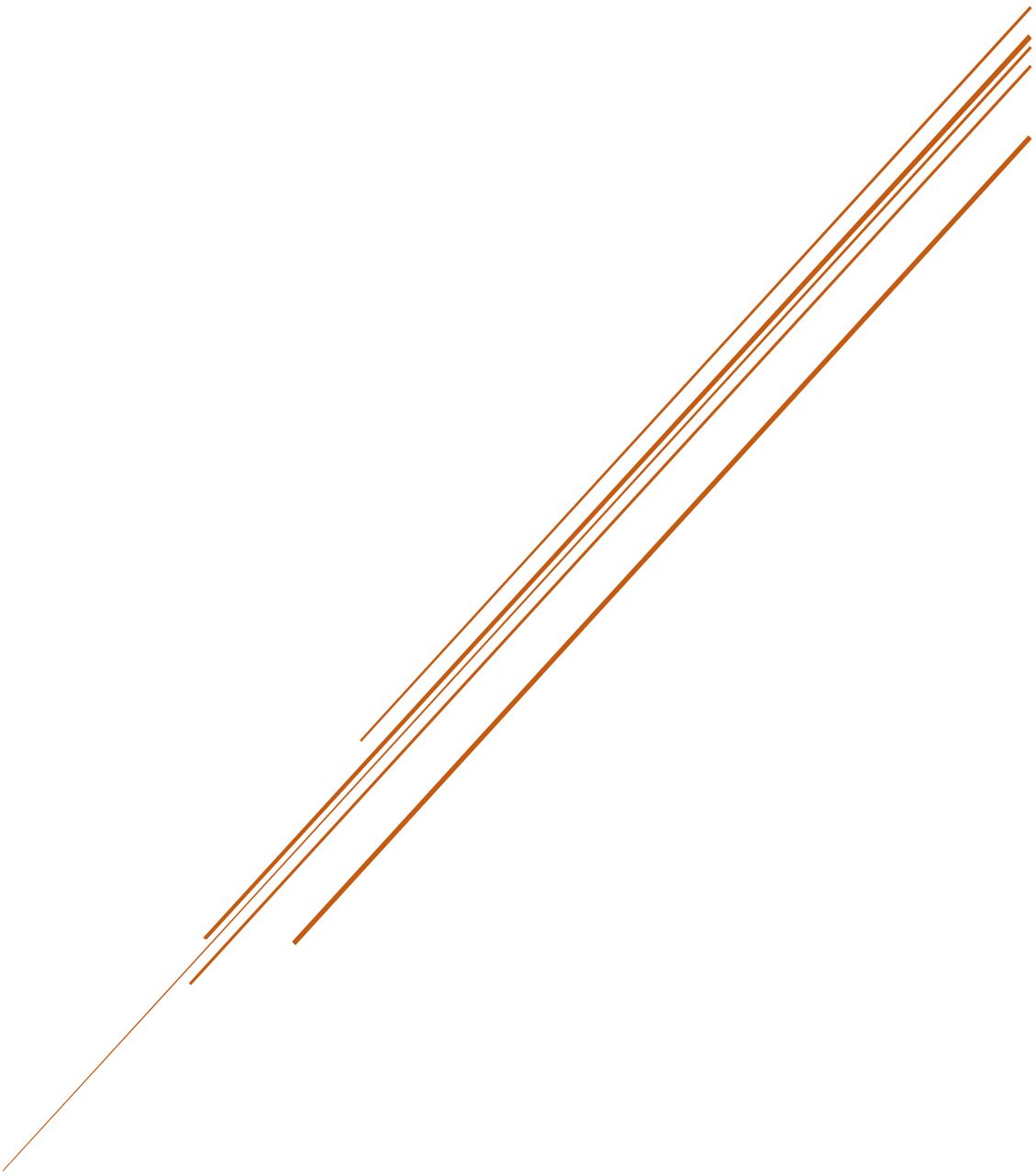
Projektmunka II.

Féléves feladat dokumentáció



Naszály Noémi

Tartalomjegyzék

Bevezetés

Projekt témája

A projekthez a 2020/21/2. félévben a Projektmunka I. tárgyhoz készült beadandó munkámat fogom felhasználni.

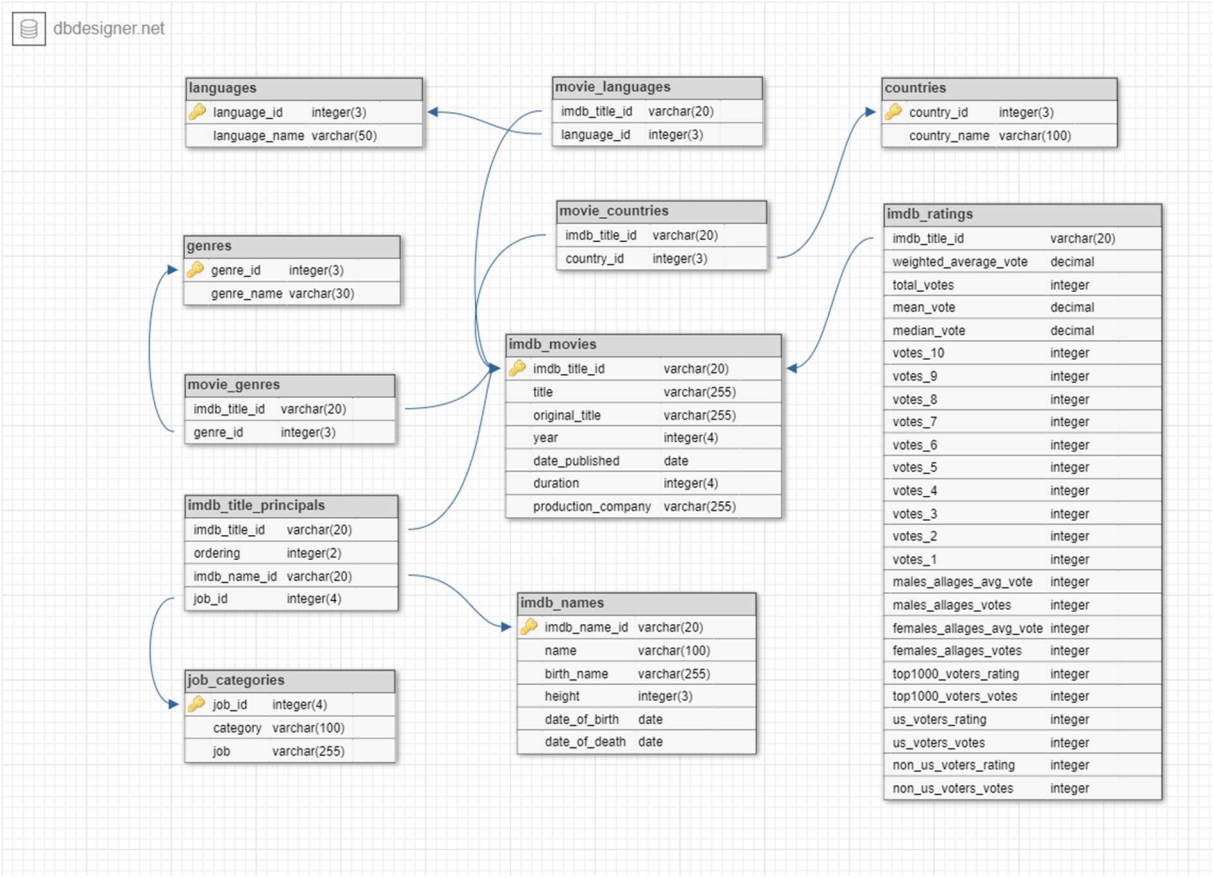
Egy filmadatbázist készítettem, ami tartalmazza a filmek értékeléseit, a hozzájuk tartozó színészeket, rendezőket, országokat, nyelveket és műfajokat.

Az adatok forrása

A Projektmunka I. során már átalakítottam az adatokat, amelyeket az alábbi helyről szereztem be: https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset

Az előző félévben az adatokat Oracle 12c adatbázisban tároltam, egy virtuális gépen. Mivel a forráshoz képest sok változtatást kellett eszközölnöm az adatokon, ezért mindenképp ezt, a virtuális gépen tárolt változatot szerettem volna használni a féléves projektem során. A legegyszerűbben csv fájlok segítségével tehettem ezt meg, hiszen a későbbi folyamatok során ezeket az exportált fájlokat tudom majd használni, megfelelnek adatforrásnak.

A táblák:



Az ábra a dbdesigner.net segítségével készült.

Use case meghatározása

Egy film készítésekor (jó esetben az) a legfontosabb szempont, hogy az tetsszen az embereknek. A visszajelzések, vélemények kinyilvánítására az IMDb oldalát sokan találják megfelelőnek, ezért gondoltam úgy, hogy ezen adatok alapján nagyjából reális következtetéseket lehetne levonni a siker kulcsának keresésekor.

Sok rendező próbál valami újat, formabontót alkotni, ám a jól bevált, sokak által kedvelt műfajokkal, színészekkel szívesen játszanak biztonsági játékot is, inkább nem kockáztatnak. Fontos még megtalálni az optimális hosszt és a lehető legalkalmasabb filmforgalmazót a feladatra.

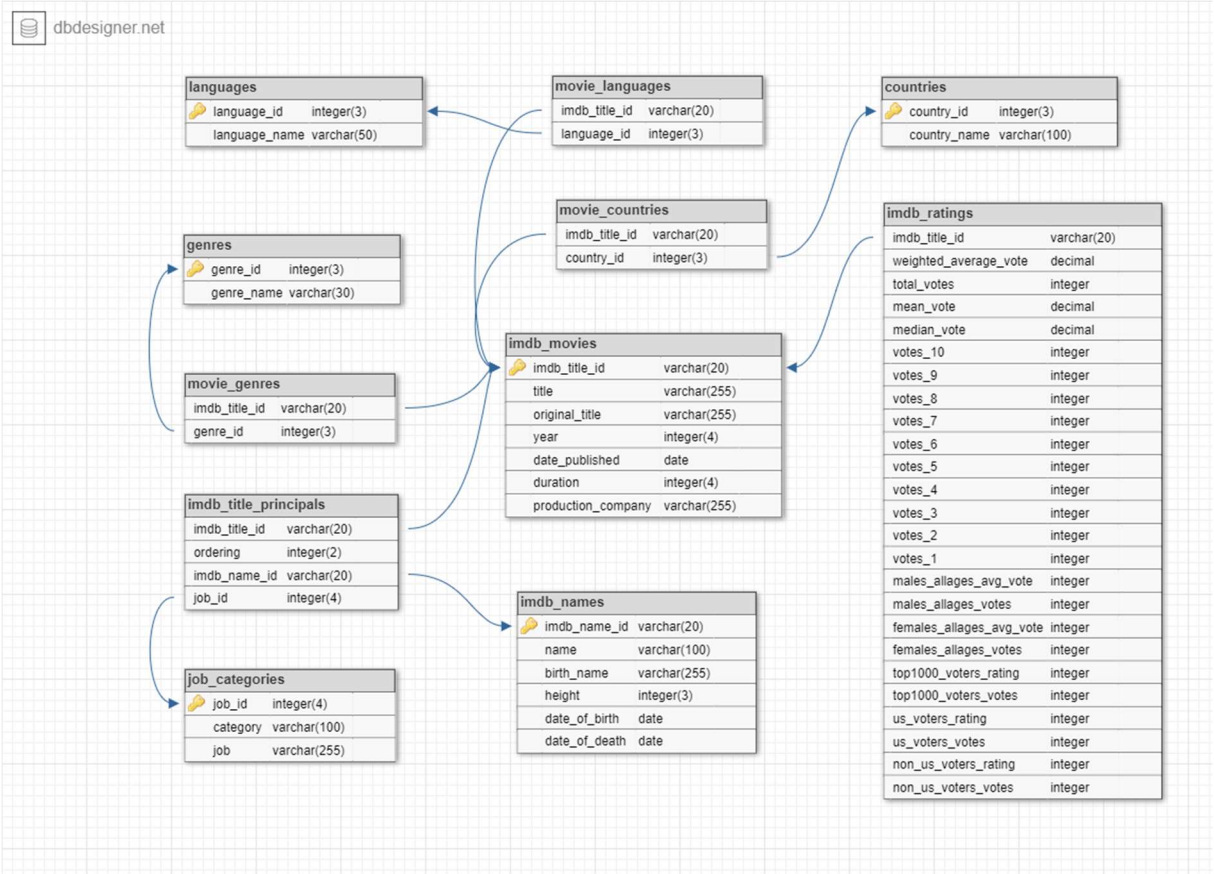
Az USA filmipara mellett nagyon nehéz a többi ország helyzete, ha olyan filmet szeretnének készíteni, ami világsikert ér el. A meglévő adatok alapján ki tudják elemezni eddigi helyzetüket, milyen tényezők kellettek az esetlegesen kiemelkedő filmjeikhez.

Egy néző nehezen tud dönteni, hogy milyen filmeket nézzen meg, de ha besoroljuk őt egy kategóriába, akkor könnyen segíthetünk ezen. Akár szűkítheti is a kört, hogy milyen műfajú filmek közül szeretne válogatni.

A nem hollywoodi filmek sokszor nehezen jutnak el az amerikaiakhoz, nehezen érnek el sikereket az USA-ban, ami az egyik legnagyobb felvevőpiac. Jónéhány filmnek viszont ez sikerült sőt, az USA-ban jobban is kedvelték, mint más országokban. A rengeteg befolyásoló tényező közül talán a filmforgalmazó kiválasztása lehet a siker kulcsa.

Forrásadatok

A forrásadatok Oracle 12c adatbázisban vannak eltárolva, melynek szerkezete az alábbi módon néz ki:



Dimenzionális modell

A forrásadatok és a use case alapján az alábbi dimenziókat és tényeket különböztetjük meg.

Dimenziók

Az egyes filmekhez több ország, nyelv, műfaj és személy tartozik, több-több kapcsolat áll fenn a rekordok között. A Country, Language, Genre és Person dimenzió ezért nem 1:N kapcsolattal csatlakozik a ténytáblához, hanem éppen fordítva, a ténytábla csatlakozik 1:N kapcsolattal ezekhez a táblákhoz.

Például egy filmnek (ami a ténytábla egy sora) a műfaja akció és kaland. Ekkor a Genre dimenziótáblában ez két sorral írható le úgy, hogy a ténytábla egy sorának (a filmnek) az azonosítója szerepel először az akcióval párosítva, másodszor a kalanddal. A Load folyamat végén (alább) található egy példa erre.

imdb_title_id	genre_name
tt0006864	Drama
tt0006864	History
tt0010323	Fantasy
tt0010323	Horror
tt0010323	Mystery
tt0012349	Comedy
tt0012349	Drama
tt0012349	Family
tt0013442	Fantasy
tt0013442	Horror

Date

A dátumokat tartalmazza, napi bontásban. Egészen az első, adatbázisban szereplő film dátumától az elkövetkezendő évekre előrevetítve.

Country

Az országokat tartalmazó dimenzió.

Language

A nyelveket tartalmazó dimenzió.

Genre

A műfajokat tartalmazó dimenzió.

ProductionCompany

A filmforgalmazókat tartalmazó dimenzió.

Person

A személyeket, azaz a színészeket, színésznőket, rendezőket tartalmazó dimenzió.

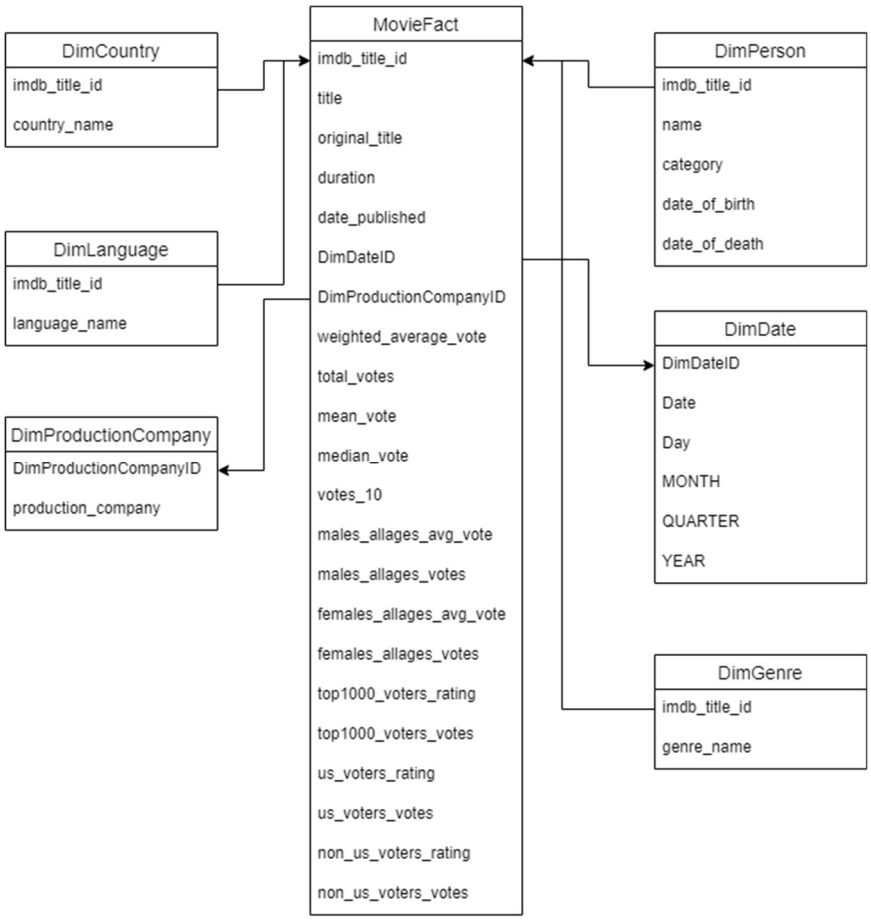
Ténytáblák

Movie

A filmek adatait tartalmazó dimenzió. Tartalmazza a különféle csoportosítások szerint számolt, átlagolt értékeléseket, a film hosszát, a megjelenésének dátumát, a címét.

Csillagséma

Az imént felsorolt dimenzió- és ténytábla(k)ból az alábbi csillagsémát hoztam létre:

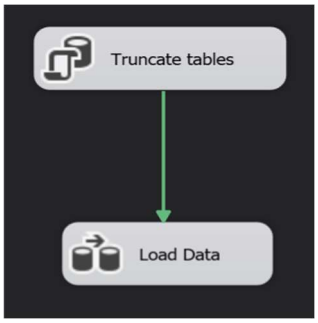


ETL folyamatok

A következőkben a dimenzionális modell tábláiba való adatbetöltéshez szükséges ETL folyamatokat, valamint az egyes rétegek adatbázisait ismertetem.

Extract

Az extract rétegben a forrásadatokat típuskonverzió nélkül, szöveges, varchar típusként töltöm be a STAGE nevű adatbázisba. Ezt megelőzően egy truncate folyamattal kiürítem a STAGE adatbázis összes tábláját.



A forrásadatokat Flat File Source segítségével érem el. Az előző féléves feladatom adatbázisának adatait a legegyszerűbben csv fájlokba exportálva tudtam kinyerni a virtuális gépről.

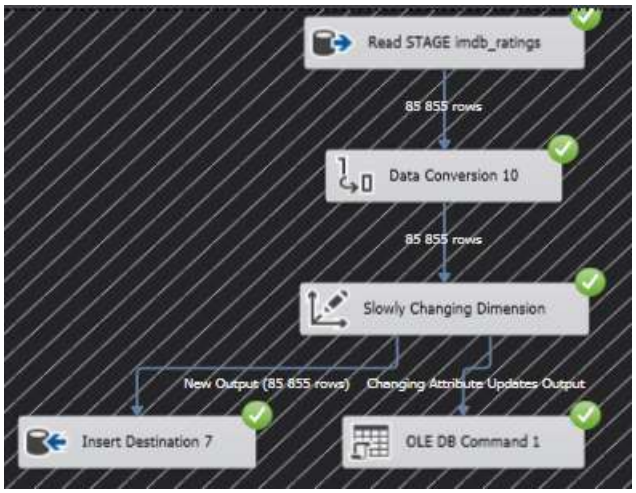
Transform

A transform rétegben a STAGE adatait áttöltöm a DW névre keresztelt adattárházba. Itt már típusosan kell tárolni az adatokat, konvertálás szükséges.

Mivel a STAGE rétegben szövegesen vannak eltárolva az értékek, így ha előfordul null, akkor az „null”-ként, szövegként szerepel, és nem tényleges NULL. Ezért szükség van az alábbi folyamatra, ami végigmegy a STAGE tábláin, azok sorain, és kicseréli a szöveges null-okat NULL értékre.

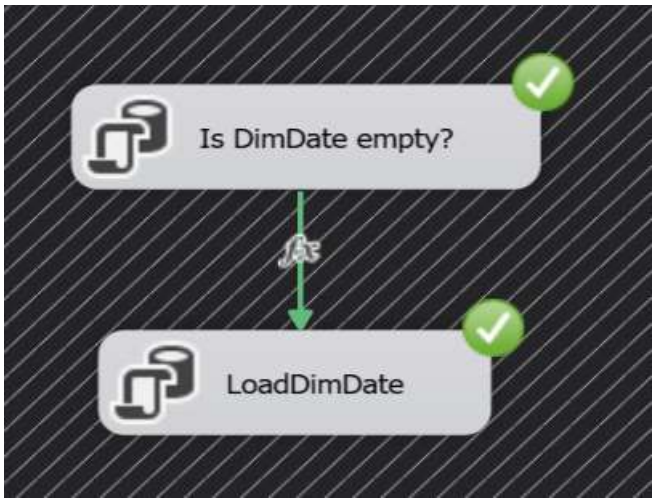
Továbbá a historizálás is itt történik, viszont az adatbázisom eddigi adatai alapján elegendő a lehetséges változásokat SCD Type 1-gyel lekezelni. Például az országoknál szerepel a Szovjetunió is, ami nem éppen aktuális, és valószínűleg már nem fognak több filmet készíteni. Így itt fix-re állítottam az ország nevét tartalmazó mezőt. A neveknél, vagyis az imdb_names táblában el vannak tárolva a halálozási dátumok is. Ha valaki meghal, attól még a neve, születési éve stb. ugyanaz marad. Így elég csak a halál dátumát átírni null-ról a szomorú napra. Az értékelések, vagyis a leendő tények pedig valószínűleg napról napra változnak, véleményem szerint ezt is elegendő felülírással kezelni, hiszen nem lesz szükségünk például az 1 évvel ezelőtti értékelésekre, elegendő az aktuálisakból következtetéseket levonni.

Néhány kép az SSIS folyamatról:



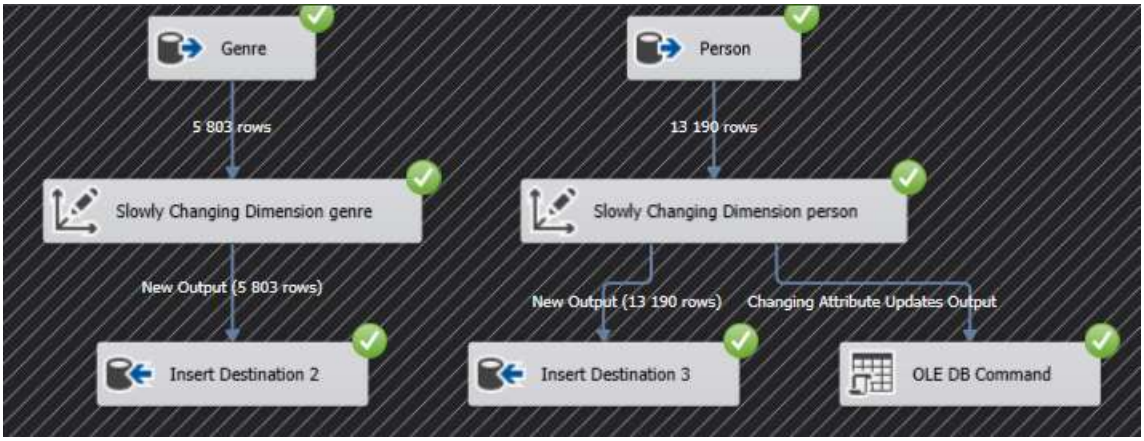
Load

Ebben a rétegben történik a csillagséma feltöltése az adattárház alapján. Először a DW adatbázisban létrehoztam a szükséges nézeteket. Szükség lesz még a dátum dimenzióra, amihez az alábbi folyamatot készítettem el:



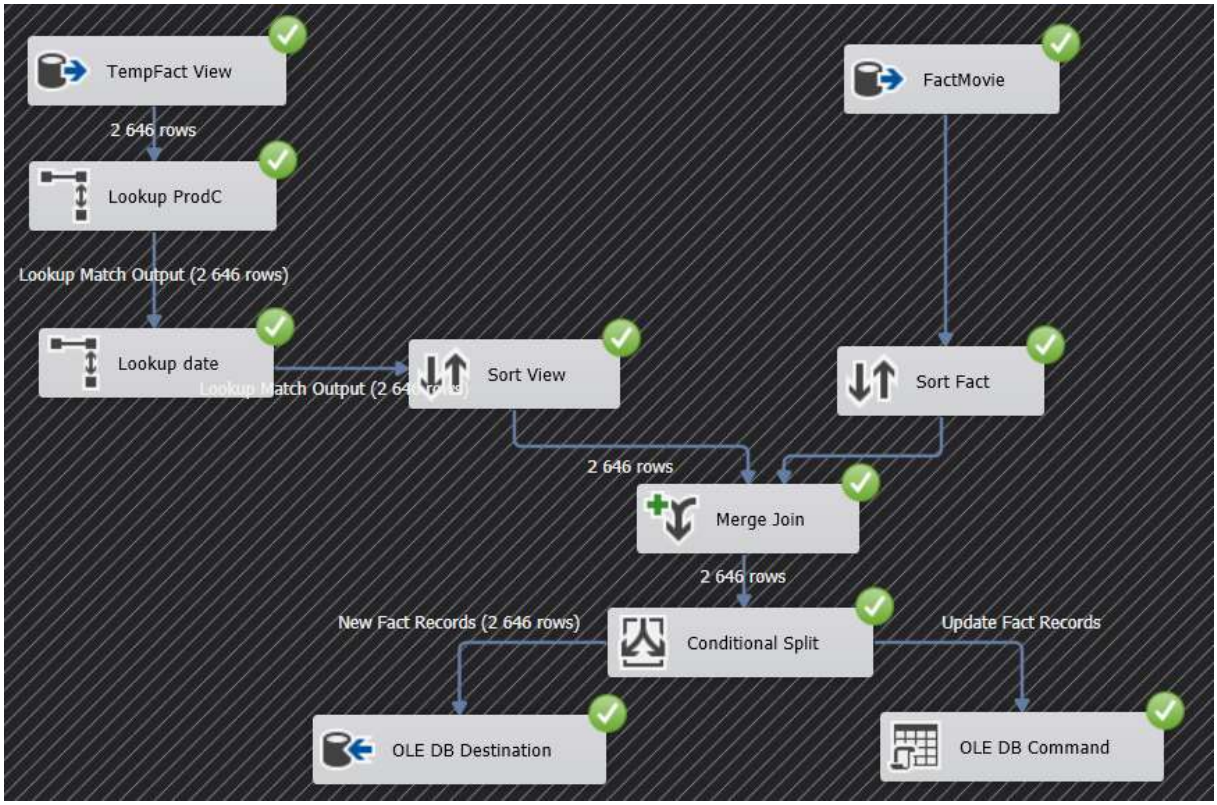
Ez leellenőrzi, hogy fel van-e már töltve a dátum dimenzió, ha pedig nem, akkor egy tárolt eljárás segítségével feltölti.

A következő lépés a dimenziótáblák feltöltése.

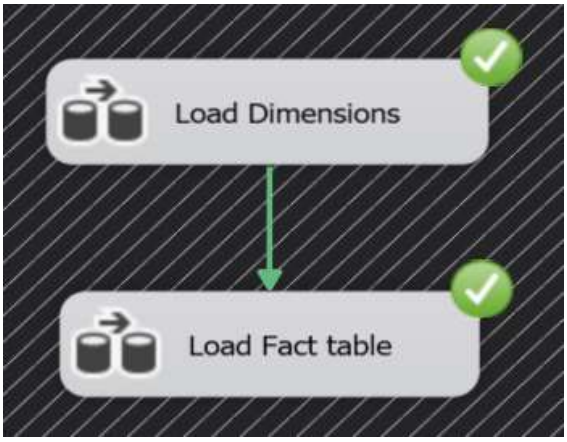


A Country, a Language és a Genre táblák adatai mind fixek - és nincs szükség historizálásra sem - mivel, ha egy filmhez hozzárendeltünk például egy nyelvet, akkor az már nem fog változni soha, örökké hozzá lesz rendelve a valóságban, és így a modellben is. Az országok névváltoztatásából eredő problémát illene figyelembe venni, viszont az előbb említett okok miatt ettől ezúton is eltekintünk. A nyelv és a műfaj esetében még a névváltoztatástól sem kell tartani, így ennél a három táblánál nincs semmi akadálya a fix attribútumok használatának. A Person táblánál az UPDATE utasításra az esetleges elhalálozások miatt van szükség. Amennyiben egy színész szerepelt egy filmben, és ezt el is tároltuk az adatbázisban, akkor ez a hozzárendelés itt sem fog megváltozni. Az SCD-re azért is szükség van, hogy mindig csak az új rekordokat szúrjuk be a DM réteg tábláiba, a csillagséma adatbázisba. Tehát ha esetleg egy új műfajt szeretnénk hozzárendelni egy már meglévő filmhez, amihez már van 2 műfaj is rendelve, azt nyugodtan megtehetjük.

Ezután a ténytábla feltöltése következik. Mivel az adatbázisban sokszor több-több kapcsolat szerepel, így itt csak a dátum és forgalmazó (DimProductionCompany) dimenziók elsődleges kulcsát keresem meg Lookup elemekkel. Ha a ténytábla már tartalmazza az adott rekordot, akkor megnézem, hogy aktuális-e még, van-e szükség frissítésre; ha pedig nem tartalmazza, akkor új rekordként beszúrom.



Ezáltal sikeresen lefutott a Load réteg dimenzió- és ténytábla feltöltő szakasza:



A több-több kapcsolatot úgy oldottam meg, hogy a ténytábla kulcsához hozzárendeltem a dimenzió táblákban minden szükséges adatot.

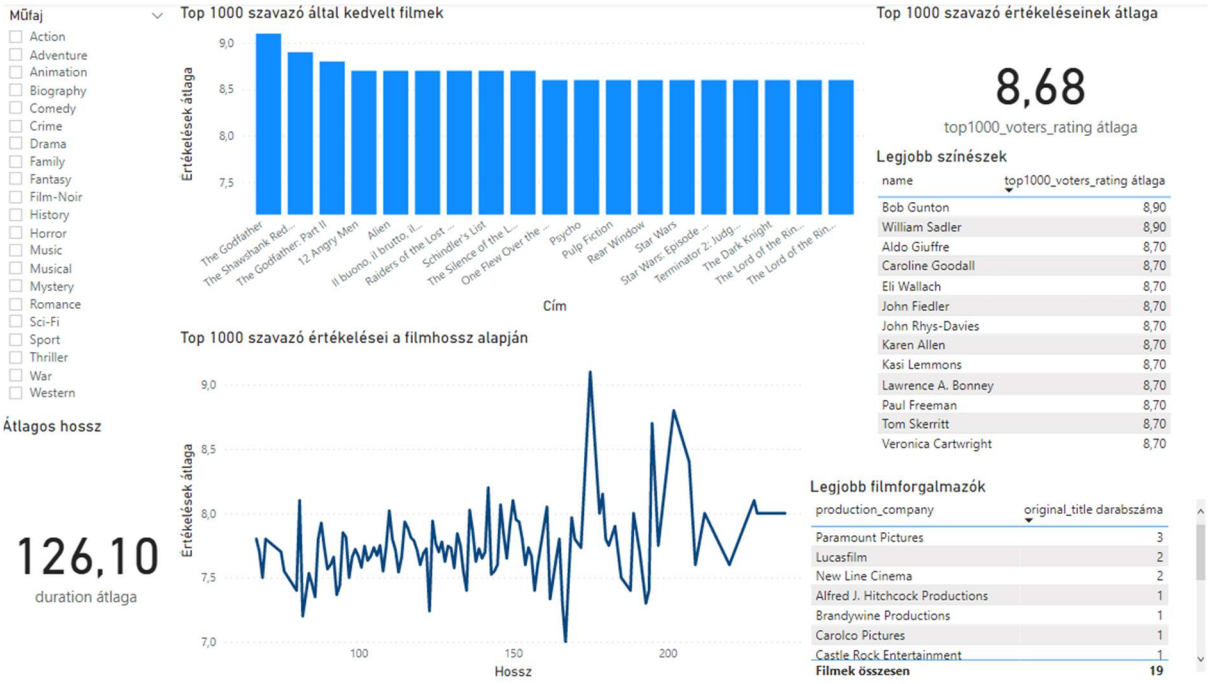
Részlet a Genre táblából:

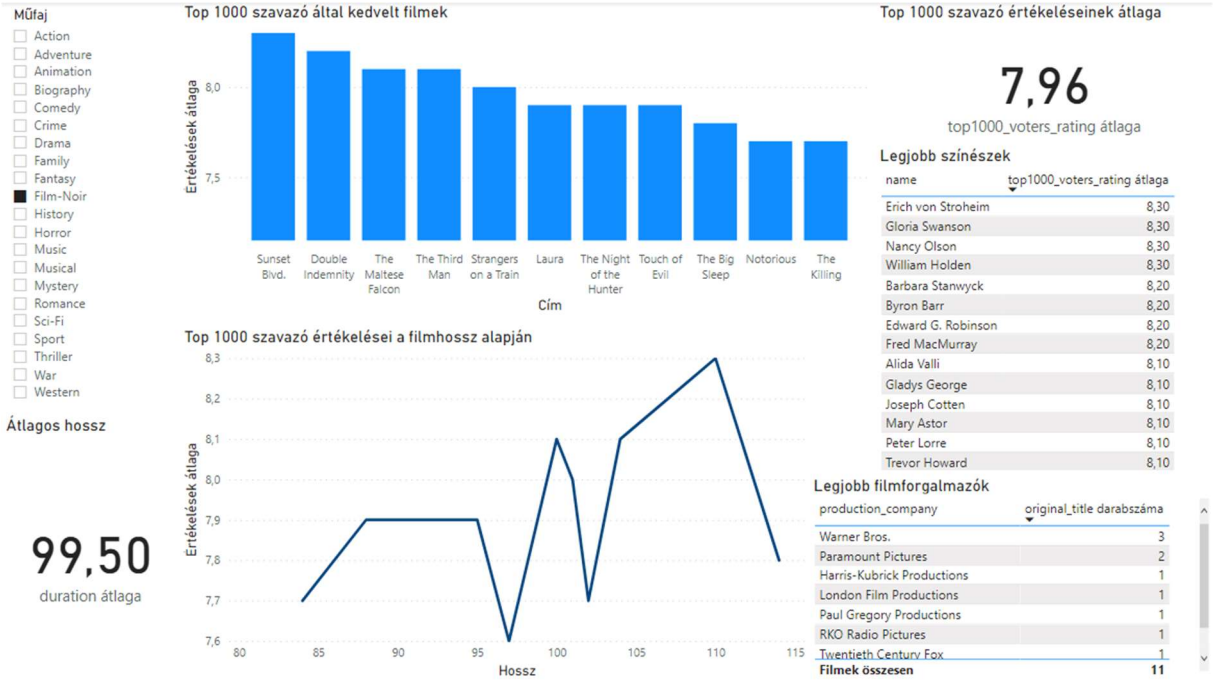
imdb_title_id	genre_name
tt0006864	Drama
tt0006864	History
tt0010323	Fantasy
tt0010323	Horror
tt0010323	Mystery
tt0012349	Comedy
tt0012349	Drama
tt0012349	Family
tt0013442	Fantasy
tt0013442	Horror

Bár így a dimenziótáblák redundánsak, cserébe nem kell bonyolult, nagy erőforrást igénylő kapcsolótáblákon keresztüljutni az információk kinyeréséhez.

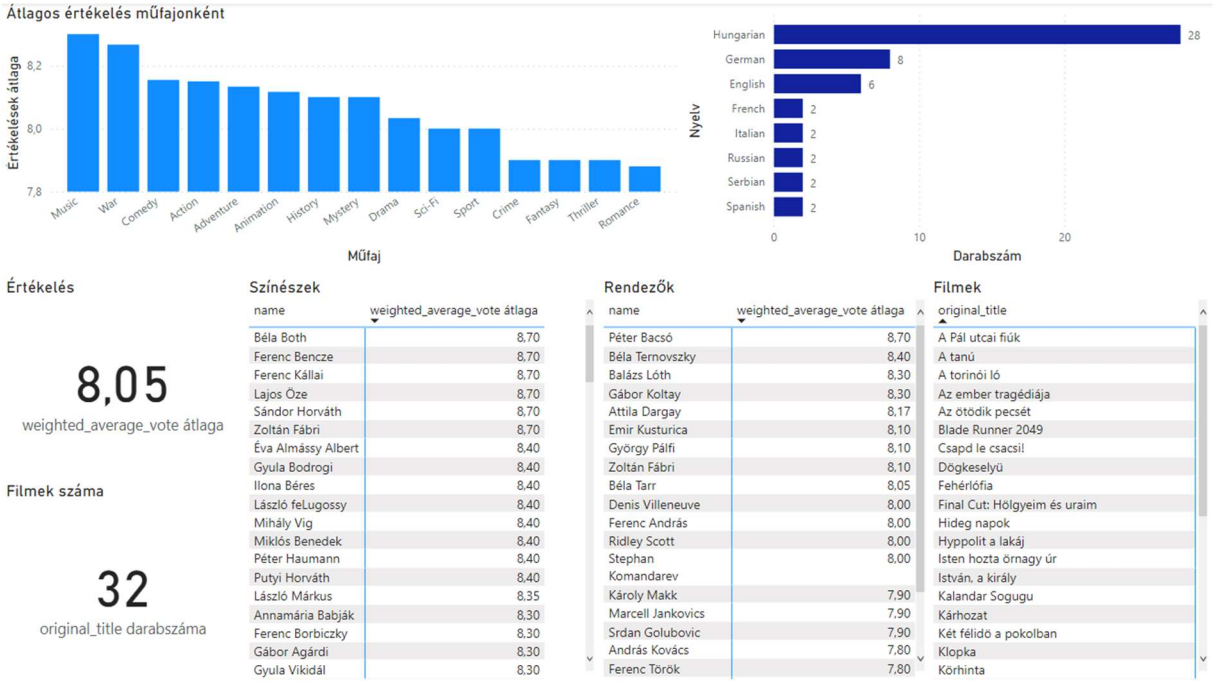
Riportkészítés

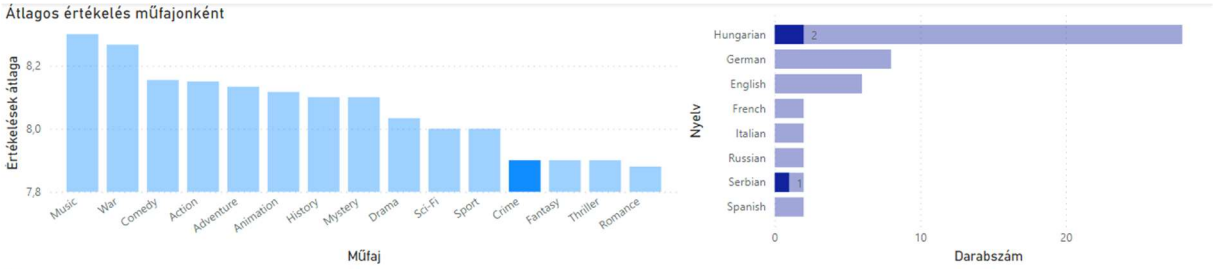
Sikeres filmek





Országok sikerei





Értékelés

7,90

weighted_average_vote átlaga

Színészek

name	weighted_average_vote átlaga
Hédi Temessy	8,00
Maria Gladkowska	8,00
Zita Perczel	8,00
Anica Dobra	7,90
György Cserhalmi	7,90
Natasa Ninkovic	7,90
Nebojsa Glogovac	7,90
Predrag 'Miki' Manojlovic	7,90
Gyula Pauer	7,80
Miklós Székely B.	7,80
Vali Kerekes	7,80

Rendezők

name	weighted_average_vote átlaga
Ferenc András	8,00
Srdan Golubovic	7,90
Béla Tarr	7,80

Filmek

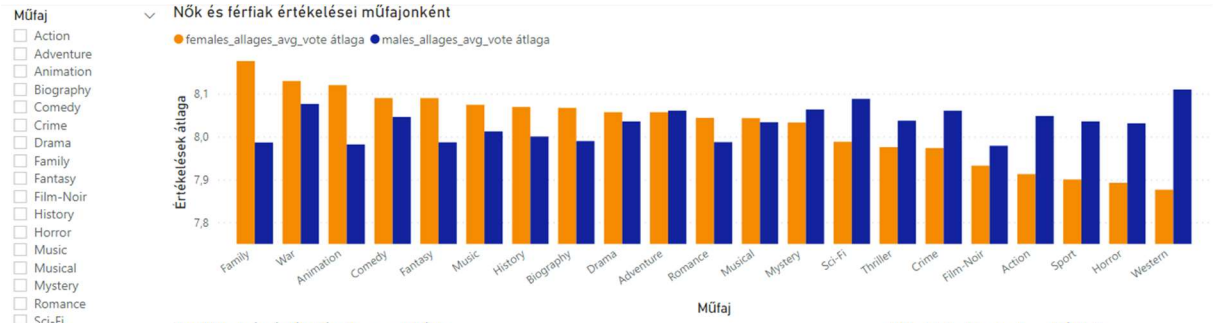
original_title
Dögkeselyü
Kárhozat
Klopka

Filmek száma

3

original_title darabszáma

Filmajánló



Top filmek és értékeléseik nemenként

original_title	females_allages_avg_vote átlaga	males_allages_avg_vote átlaga	weighted_average_vote átlaga
Hababam Sinifi	9,30	9,30	9,30
The Shawshank Redemption	9,20	9,30	9,30
The Godfather	8,90	9,20	9,20
Aynabaji	9,20	9,00	9,10
Vchera	9,30	8,90	9,10
Zana	8,20	9,30	9,10
Ekvtime: Man of God	9,90	8,40	9,00
Hababam Sinifi Sinifta Kaldi	9,00	8,90	9,00
Maratonci trce pocasni krug	9,10	8,90	9,00
The Dark Knight	8,70	9,10	9,00
The Godfather: Part II	8,70	9,00	9,00
Tosun Pasa	8,90	9,00	9,00

Nők értékeléseinek átlaga

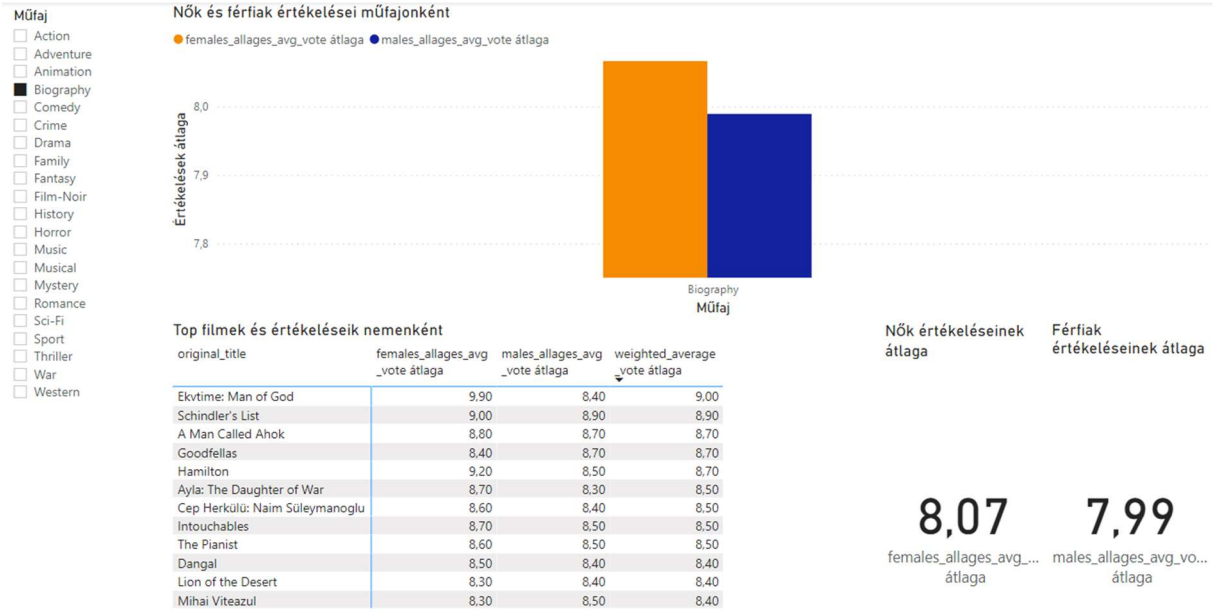
8,05

females_allages_avg_... átlaga

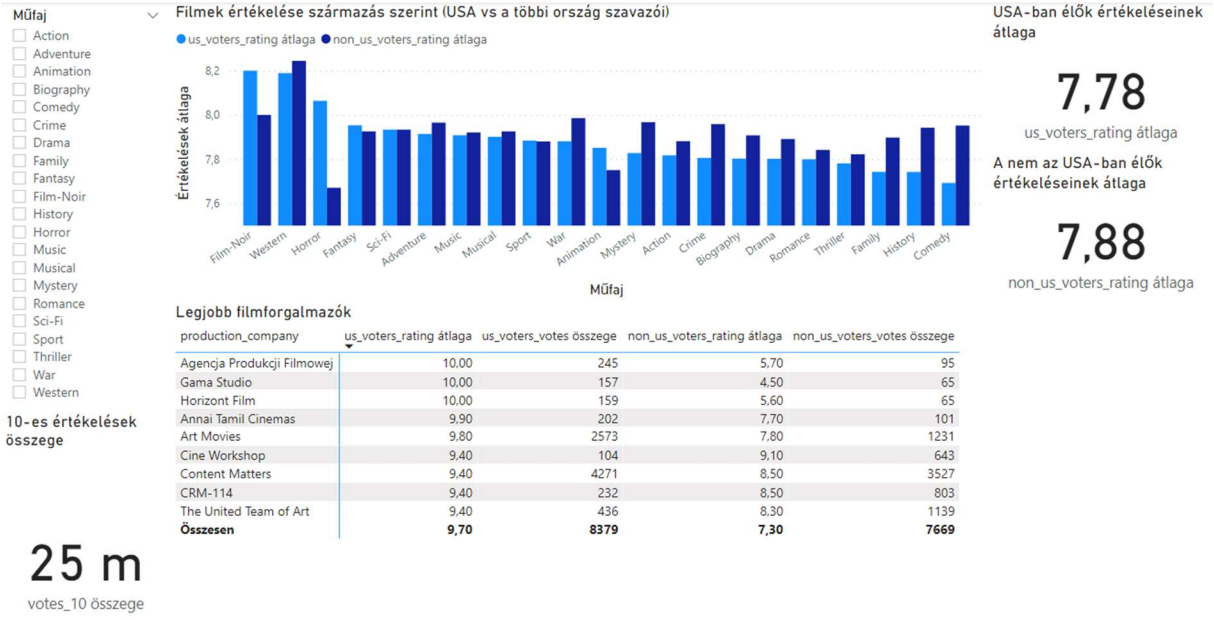
Férfiak értékeléseinek átlaga

8,03

males_allages_avg_vo... átlaga



Hollywoodon kívüli sikerek



Műfaj

- [ ] Action
- [ ] Adventure
- [x] Animation
- [ ] Biography
- [ ] Comedy
- [ ] Crime
- [ ] Drama
- [ ] Family
- [ ] Fantasy
- [ ] Film-Noir
- [ ] History
- [ ] Horror
- [ ] Music
- [ ] Musical
- [ ] Mystery
- [ ] Romance
- [ ] Sci-Fi
- [ ] Sport
- [ ] Thriller
- [ ] War
- [ ] Western

10-es értékelések összege

1 m

votes_10 összege

Filmek értékelése származás szerint (USA vs a többi ország szavazói)

USA-ban élők értékeléseinek átlaga

7,85

us_voters_rating átlaga

A nem az USA-ban élők értékeléseinek átlaga

7,75

non_us_voters_rating átlaga

Legjobb filmforgalmazók

production_company	us_voters_rating átlaga	us_voters_votes összege	non_us_voters_rating átlaga	non_us_voters_votes összege
Bradeway Pictures	9,10	148	8,90	52
DENTSU Music And Entertainment	8,40	51115	8,30	144351
Shinchosha Company	8,40	29832	8,50	100401
Toei Animation	8,40	2408	7,17	7597
Amuse	8,30	13380	8,30	65136
Pixar Animation Studios	8,30	142940	8,00	380693
Tokuma Shoten	8,20	110440	8,13	355372
A-1 Pictures	8,20	88	7,90	226
Buena Vista Home Entertainment	8,20	43359	8,10	140697
Tokuma Japan Communications	8,20	34878	8,10	122109
Production I.G.	8,15	147	7,70	541
Összesen	**8,33**	**428735**	**7,96**	**1317175**