Üzleti intelligencia házi feladat dokumentáció Autók árának előrejelzése, Netflix és Youtube nézettségek, vélemények

Inez Anna Papp

May 2025

1 Forráskód, videó

Link a forráskód GitHub repojához: `https://github.com/pappinez/BI_hw`
A videó a repoban található.

2 Bemutatás

A feladatom folyamán 3 adathalmazt használok, az egyikben autók eladási árai szerepelnek, a másikban országonkénti youtube nézettségek két különböző hónapban, a harmadikban pedig Netflixes sorozat érétkelések. Az autós adathalmaz idősoros adatokat tartalmaz, így a LSTM alapú predikcióhoz fogom alkalmazni, a másik két adathalmazon pedig ETL jobokat fogok végrehajtani és reportokat készíteni belőlük. A Netflixesnél a KPI, hogy melyiknek a legjobb az értékelése, a youtubeosnál, hogy melyik ország nézi a legtöbbet, az autósnál pedig, hogy az ár emelkedik vagy csökken különböző tényezők figyelembevételével. Az ETL jobokkal megtisztítom az adatokat, majd átalakítom őket az elvért formátumokba. A reportoknál különböző megjelenítéseket készítek, melyik interaktívak, és az adatokról mutatnak be különböző nézeteket, ezzel valamilyen új információt bemutatva.

3 Főbb funkciók összefoglaló

3.1 Általam választott adatforrások

- Autók eladási árai
- Országonkénti youtube nézettség
- Netflix sorozatok és értékelések

3.2 Adattárolás

- Raw réget: az adatok .csv fileokban vannak tárolva
- Stage réteg: megtörténik az adattisztítás, hibás adatok kiszűrése
- Data Warehouse réteg: dimenzió bevezetése

3.3 Megvalósítandó ETL jobok

- Transzformáció 1: Hibás adatokat keres és kilogolja azokat a youtube-os adatforráson
- Job 1: Ha a bemenetnek megadott mappába került egy .csv file, akkor elvégzi a transzformációt, ha nem, akkor erről készít logot
- Transzformáció 2: Adattisztítás, hibás értékek kiszűrése a Netflixes adatforráson
- Job 2: Ha a bemenetnek megadott mappába került egy .csv file, akkor elvégzi a transzformációt, ha nem, akkor erről készít logot
- Transzformáció 3: Netflix értékelések alapján kategorizálja, hogy ez egy Jó, Közepes, vagy Rossz film, ehhez sávok vannak megadva, az összegzést kimenti egy JSON fileba
- Job 3: a jó filmek feltöltése webszerverre
- Transzformáció 4: Youtube nézettség csoportosítás kontinensenként, majd aggregáció: átlagosnézettség a megadott két hónapban
- Job 4: Ha a bemenetnek megadott mappába került egy .csv file, akkor elvégzi a transzformációt, ha nem, akkor erről készít logot

Az adatokból olyan adattípus készül, ami alkalmas arra, hogy a PowerBI fel tudja használni

3.4 Reportok

- Report 1: Kördiagrammon ábrázolom a 3 legtöbb youtube felhasználóval rendelkező kontinenst, mellette egy táblázatban a kontinensek nevét és az átlagos nézettségüket a két megadott hónapban. A másik oldalon clustered bar charttal ábrázolom az országokat, és nézettségüket, és lefúrást használok.
- Report 2: Csatlakoztatom a youtube és netflix táblákat a Netflix nézettség ország és a filmet készítő ország alapján. Egy táblázatban megjelenítem azáltal a top 2 ország által készített filmeket, amely két országban a legmagasabb a youtube nézettség, az értékelésük szerint csökkenő sorrendben.

- Report 3: 3 táblázatban ábrázolom a Jó, Rossz és Közepes értékelésű Netflix filmeket, majd az x tengelyen évek szerint megjelenítem a kategóriákat egy scatter charton
- Report 4: Kördiagramon ábrázolom a Netflixes filmeken a műfajok arányát, egy másikon a film felfedezésének arányát

3.5 Data Science

LSTM hálózatot használva az autó árakról szóló idősoros adatok alapján készítek predikciót a jövő adatokról

4 Választott technológiák

- Pentaho az ETL transzformációkhoz és jobokhoz
- Power BI a reportokhoz
- Python a predikcióhoz

5 Transzformációk, Jobok

5.1 Transzformáció 1

Hibás adatokat keres és kilogolja azokat a youtube-os adatforráson.

Ehhez szükség volt a forrás mappában lévő fájlnevek összegyűjtésére, majd CSV-ből az adatok beolvasására. Utána a nézettségszám adatokból kifiltereltem a 0-nál kisebb érétkeket, végül adatvalidációval kiszűrtem a nem 2 nagy betűből álló zászló kódokat, és egy .txt-be kimentettem őket. A tisztított adatokat egy .csv-be, és a PowerBI-hoz Excelként is mentettem. *Figure 1 Figure 2*

5.2 Job 1

Ha a bemenetnek megadott mappába került egy .csv file, akkor elvégzi a transzformációt, ha nem, akkor erről készít logot
Megnéztem, hogy üres-e a mappa, ha igen, akkor kilogoljuk ezt, ha nem, akkor viszont elvégzi rajta a transtformációt. *Figure 3*

5.3 Transzformáció 2

Adattisztítás, hibás értékek kiszűrése a Netflixes adatforráson
Ehhez szükség volt a forrás mappában lévő fájlnevek összegyűjtésére, majd CSV-ből az adatok beolvasására. Utána kifiltereltem azokat az datokat, ahol nem 0 és 10 közötti az értékelés, végül adatvalidációval kiszűrtem azokat az évszámokat, ami nem az 1900-as, vagy 2000-es években keletkezett, és egy .txt-be kimentettem őket. A tisztított adatokat egy .csv-be, és a PowerBI-hoz Excelként is mentettem.

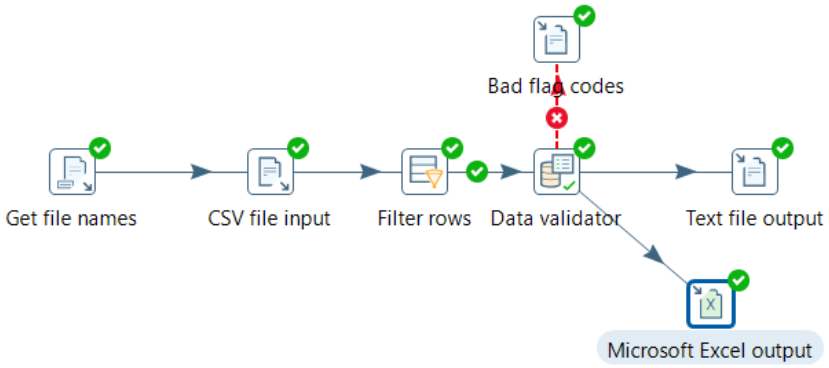


Figure 1: Transformation 1

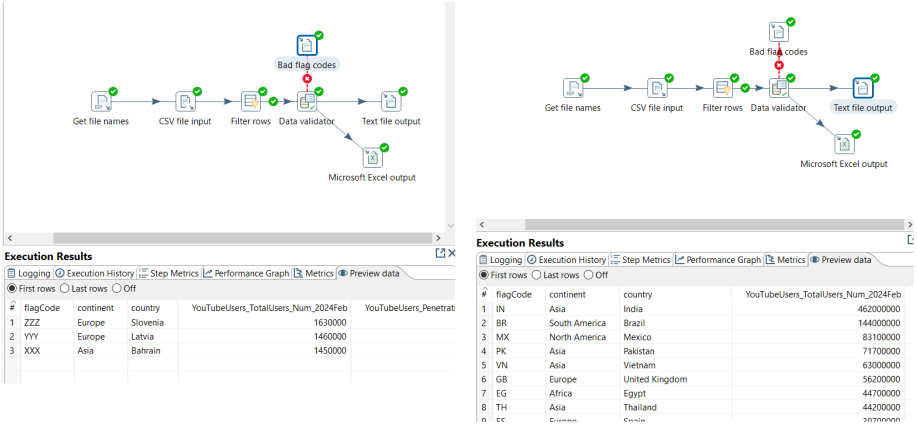


Figure 2: Result of Transformation 1

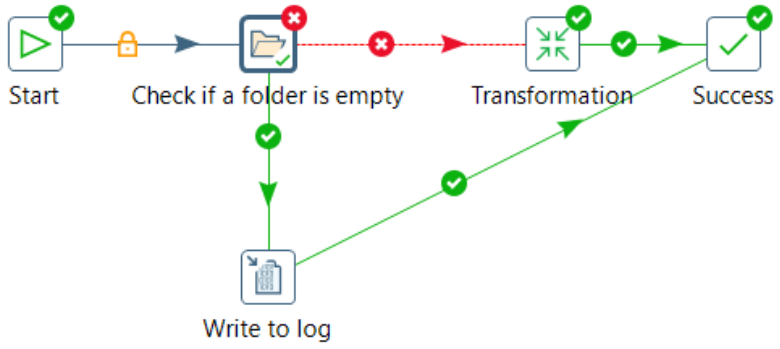


Figure 3: Job 1

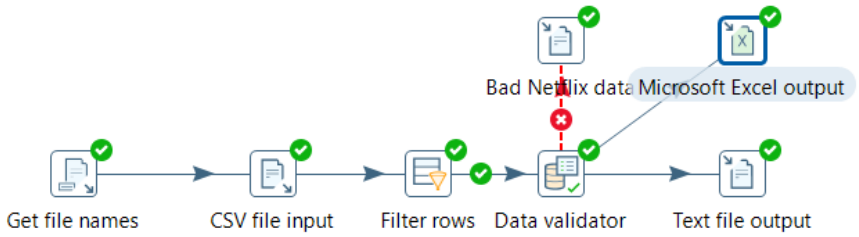


Figure 4: Transformation 2

Figure 4 Figure 5

5.4 Job 2

A a bemenetnek megadott mappába került egy .csv file, akkor elvégzi a transzformációt, ha nem, akkor erről készít logot

Megnéztem, hogy üres-e a mappa, ha igen, akkor kilogoljuk ezt, ha nem, akkor viszont elvégzi rajta a transtformációt. *Figure 6*

5.5 Transzformáció 3

Netflix értékelések alapján kategorizálja, hogy ez egy Jó, Közepes, vagy Rossz film, ehhez sávok vannak megadva, a kategóriánként összegzést kimenti egy

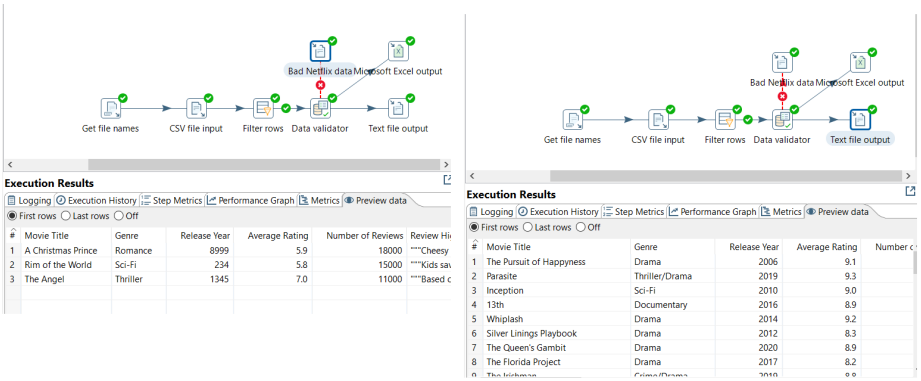


Figure 5: Result of Transformation 2

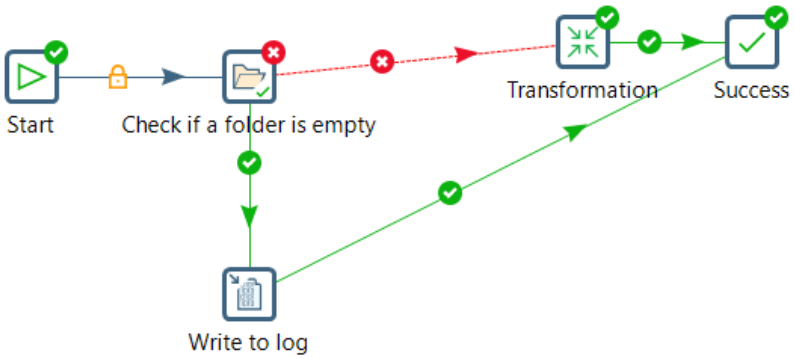


Figure 6: Job 2

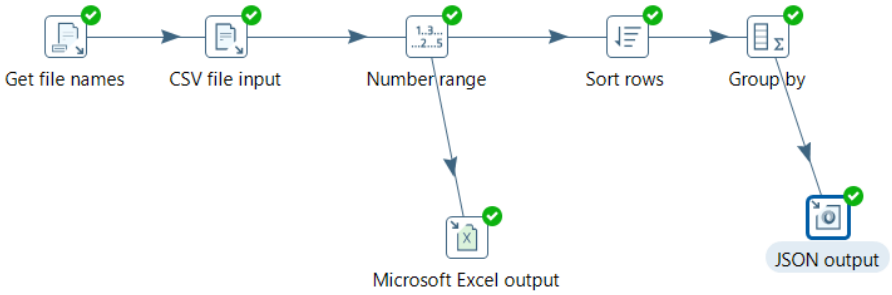


Figure 7: Transformation 3

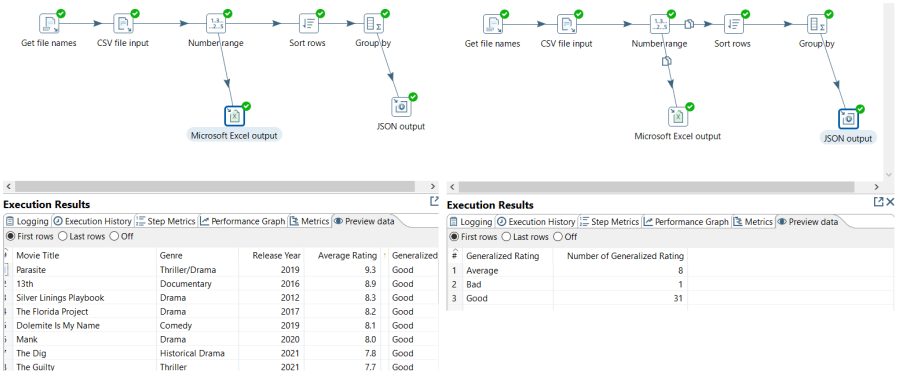


Figure 8: Result of Transformation 3

JSON fileba.
Ehhez szükség volt a forrás mappában lévő fájlnevek összegyűjtésére, majd CSV-ből az adatok beolvasására. Ezután 1-5 ig 'Bad', 5-7-ig 'Average' 7-10-ig 'Good' értékelést kaptak, majd értékelés szerinti rendezés után JSON-ban kimentettem az adatokat. Egy Excel formátumú mentés is készült a PowerBI-hoz. *Figure 7* *Figure 8*

5.6 Job 3

A jó filmek feltöltése webszerverre.
Egy HTTP requesttel feltöltöttem webszerverre, ha pedig nem sikerült, azt kilogoltam. *Figure 9*

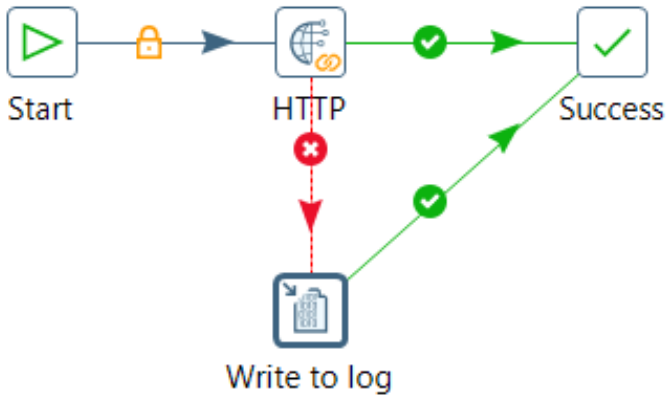


Figure 9: Job 3

5.7 Transzformáció 4

Youtube nézettség csoportosítás kontinensenként, majd aggregáció: átlagosnézettség a megadott két hónapban
Ehhez szükség volt a forrás mappában lévő fájlnevek összegyűjtésére, majd CSV-ből az adatok beolvasására. Ezután rendeztem, és az átlagos nézettséget kiszámoltam a két adott hónapra és .csv-be mentettem. *Figure 10 Figure 11*

5.8 Job 4

Ha a bemenetnek megadott mappába került egy .csv file, akkor elvégzi a transzformációt, ha nem, akkor erről készít logot
Megnéztem, hogy üres-e a mappa, ha igen, akkor kilogoljuk ezt, ha nem, akkor viszont elvégzi rajta a transzformációt. *Figure 12*

6 Reportok

6.1 Report 1

Kördiagrammon ábrázolom a 3 legtöbb youtube felhasználóval rendelkező kontinenst, mellette egy táblázatban a kontinensek nevét és az átlagos nézettségüket a két megadott hónapban. A másik oldalon clustered bar charttal ábrázolom

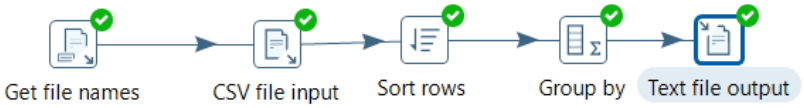
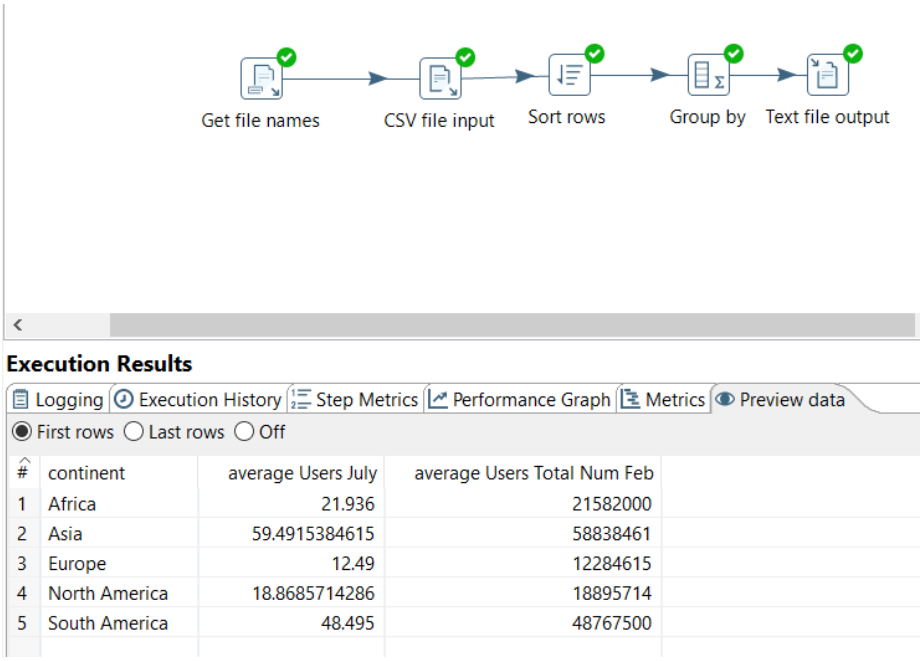


Figure 10: Transforation 4



#	continent	average Users July	average Users Total Num Feb
1	Africa	21.936	21582000
2	Asia	59.4915384615	58838461
3	Europe	12.49	12284615
4	North America	18.8685714286	18895714
5	South America	48.495	48767500

Figure 11: Result of Transformation 4

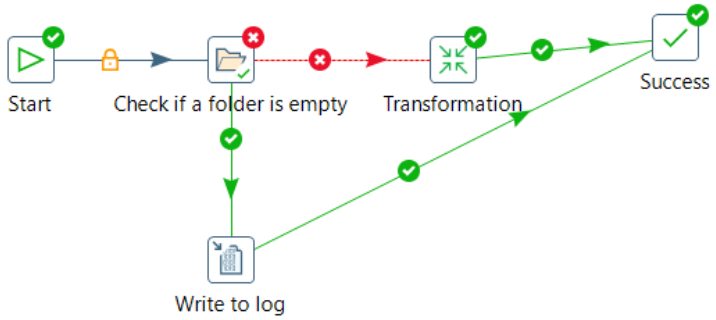


Figure 12: Job 4

az országokat, és nézettségüket, és lefúrást használok. *Figure 13 Figure 14 Figure 15*

6.2 Report 2

Csatlakoztatom a youtube és netflix táblákat a Netflix nézettség ország és a filmet készítő ország alapján. Egy táblázatban megjelenítem azáltal a top 2 ország által készített filmeket, amely két országban a legmagasabb a youtube nézettség, az értékelésük szerint csökkenő sorrendben. *Figure 16 Figure 17*

6.3 Report 3

Táblázatban ábrázolom a Jó, Rossz és Közepes értékelésű Netflix filmeket, majd az x tengelyen évek szerint megjelenítem a kategóriákat egy scatter charton *Figure 18 Figure 19*

6.4 Report 4

Kördiagramon ábrázolom a Netflixes filmeken a műfajok arányát, egy másikon a film felfedezésének arányát *Figure 20 Figure 21*

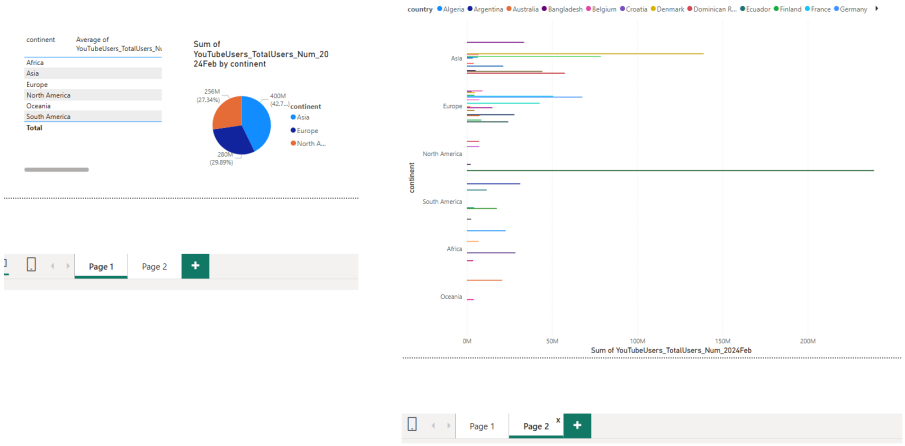


Figure 13: Report 1

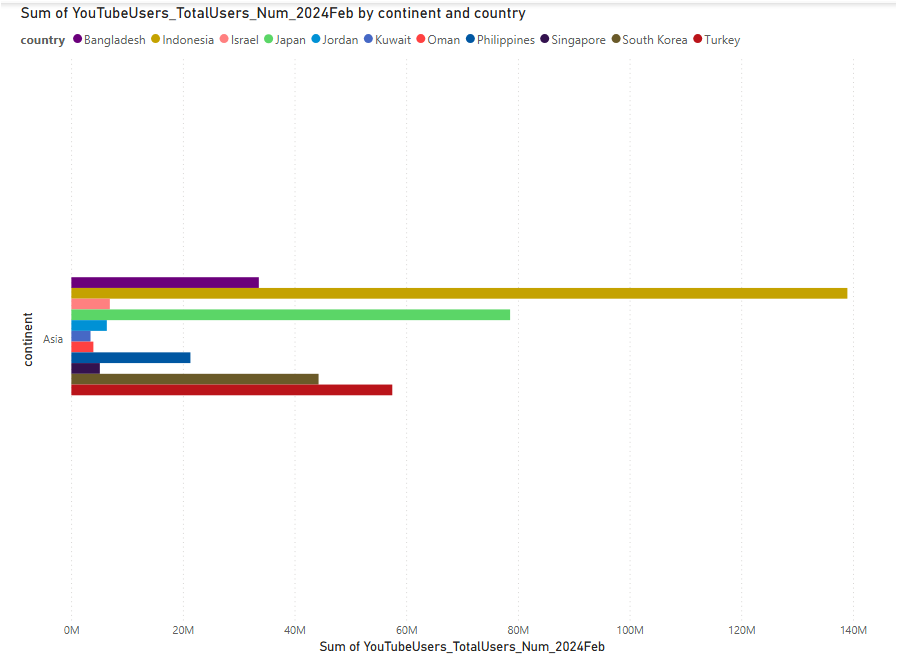


Figure 14: Report 1 Drill down

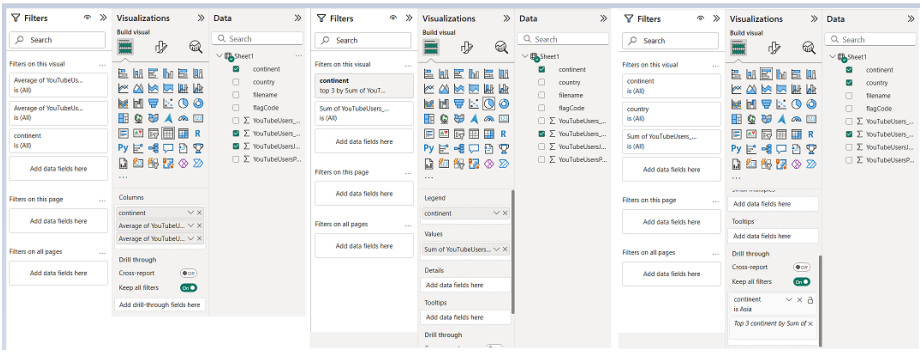
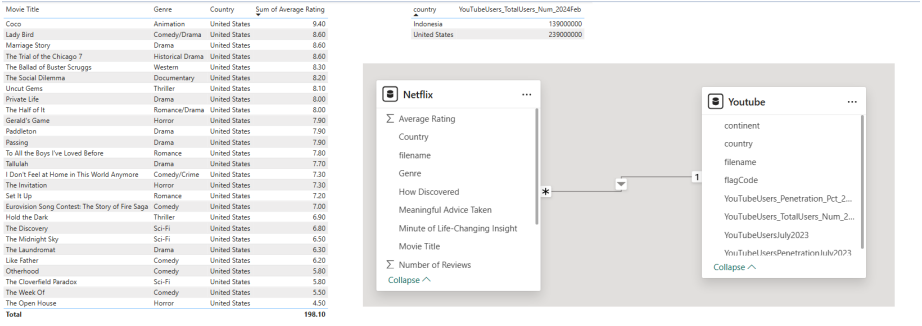


Figure 15: Settings for Report 1



Movie Title	Genre	Country	Sum of Average Rating
Coco	Animation	United States	9.40
Lady Bird	Comedy/Drama	United States	8.60
Marriage Story	Drama	United States	8.60
The Trial of the Chicago 7	Historical Drama	United States	8.60
The Ballad of Buster Scruggs	Western	United States	8.30
The Social Dilemma	Documentary	United States	8.20
Uncut Gems	Thriller	United States	8.10
Private Life	Drama	United States	8.00
The Half of It	Romance/Drama	United States	8.00
Gerald's Game	Horror	United States	7.90
Paddleton	Drama	United States	7.90
Passing	Drama	United States	7.90
To All the Boys I've Loved Before	Romance	United States	7.80
Tallulah	Drama	United States	7.70
I Don't Feel at Home in This World Anymore	Comedy/Crime	United States	7.30
The Invitation	Horror	United States	7.30
Set It Up	Romance	United States	7.20
Eurovision Song Contest: The Story of Fire Saga	Comedy	United States	7.00
Hold the Dark	Thriller	United States	6.90
The Discovery	Sci-Fi	United States	6.80
The Midnight Sky	Sci-Fi	United States	6.50
The Laundromat	Drama	United States	6.30
Like Father	Comedy	United States	6.20
Otherhood	Comedy	United States	5.80
The Cloverfield Paradox	Sci-Fi	United States	5.80
The Week Of	Comedy	United States	5.50
The Open House	Horror	United States	4.50
Total			**198.10**

country	YouTubeUsers_TotalUsers_Num_2024Feb
Indonesia	139000000
United States	239000000

Figure 16: Report 2

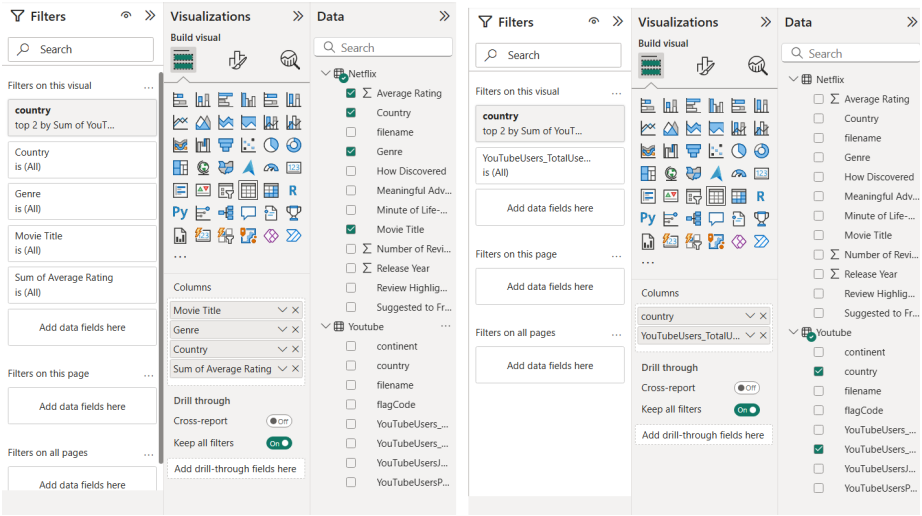


Figure 17: Settings for Report 2

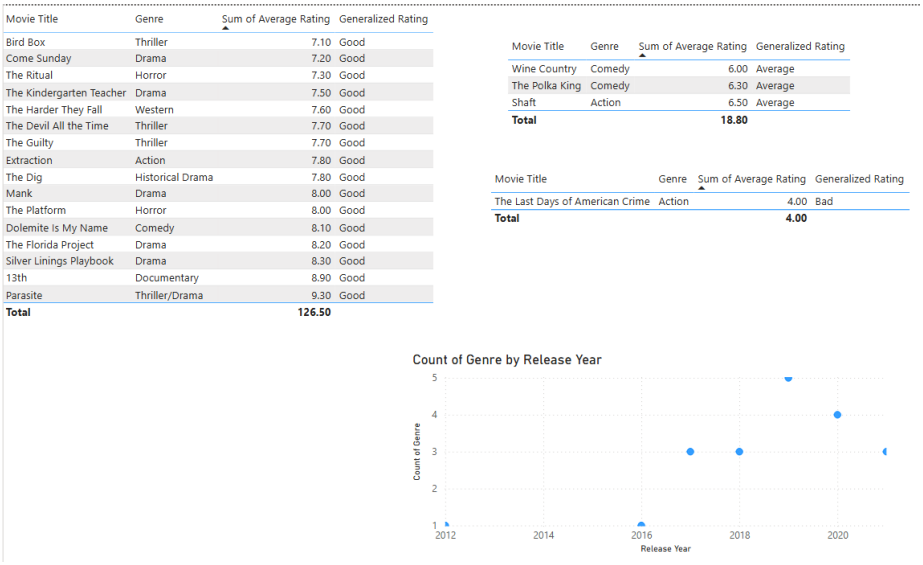


Figure 18: Report 3

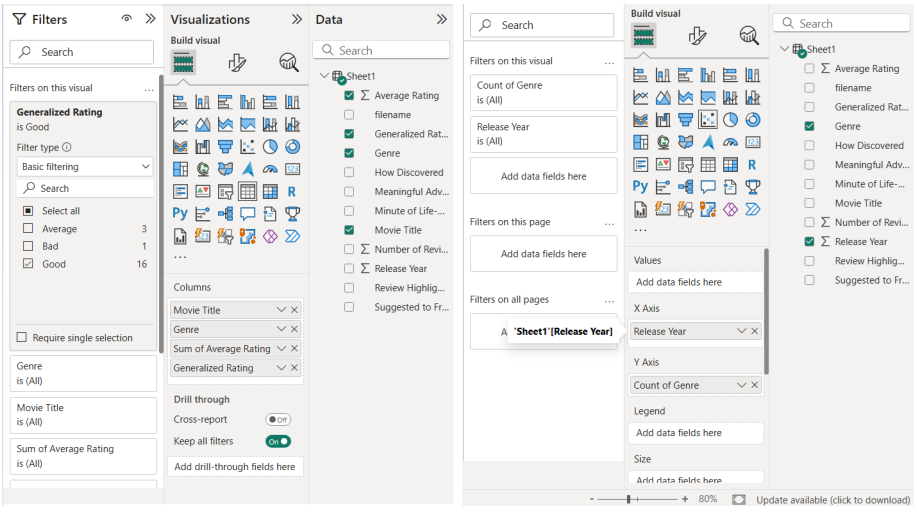


Figure 19: Setting for Report 3

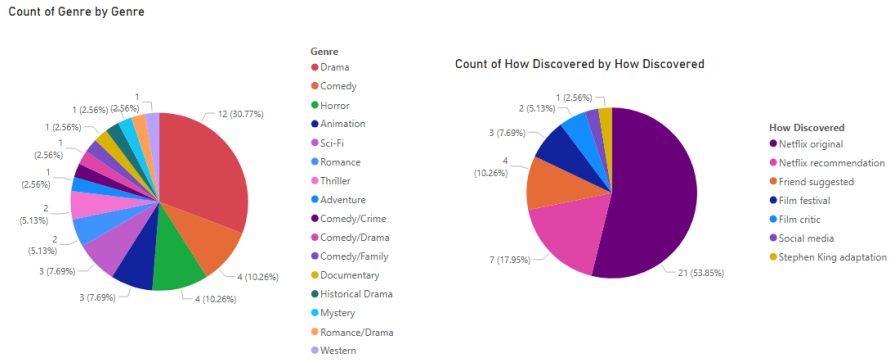


Figure 20: Report 4

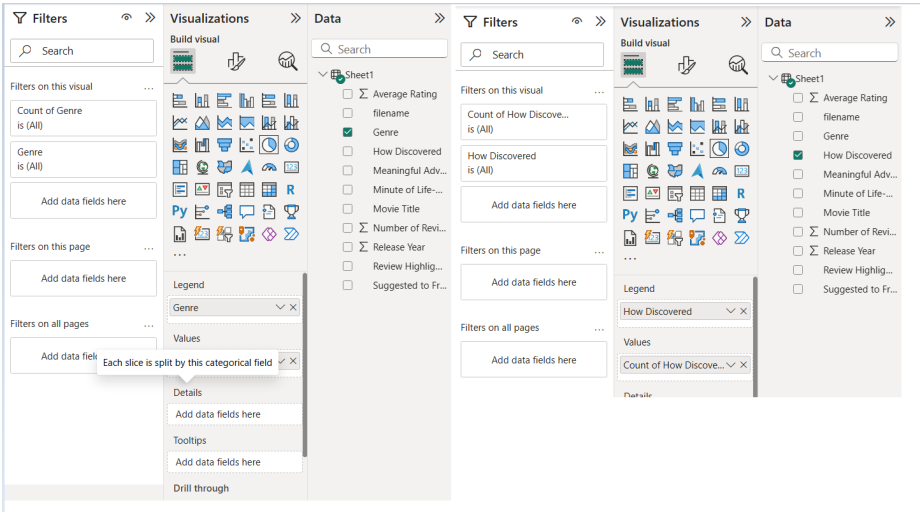


Figure 21: Settings for Report 4

7 Idősoros előrejelzés Pythonban

Az alábbi linken megtalálható a forráskód és kommentek formájában a dokumentáció:

`https://colab.research.google.com/drive/1gUd3XhPzx1xIK0XbaRlgnwsPhP54fL6K?usp=sharing`

Szükség lehet a sample data mappába feltölteni a következő filet, amit a házi feladat GitHub repojában lehet elérni: automobile_prices_economics_2019_2023