UNIVERSITAT ROVIRA i VIRGILI

Creació d'un Sistema d'Informació Cinematogràfic: Emmagatzematge i Anàlisi de Dades

ALUMNES: Àlex Figueroa Boronat,

Raül Mesa Palazón,

Víctor Sentís Lahoz

PROFESSOR: Edgar Batista De Frutos

ASSIGNATURA: Sistemes d'Informació en les Organitzacions

ENSENYAMENT: GEI

DATA: 23/04/2022

Index

1. Introducció

La quantitat de dades digitals existents creix a un ritme molt accelera.
Però, més enllà del volum, un dels principals problemes és l'estructura d'aquestes dades, ja que la gran majoria d'elles no estan estructurades (per exemple, els documents PDF i Word, imatges i àudios) o bé tenen un grau mig d'estructuració (per exemple, els fitxers XML i JSON). A l'hora de gestionar dades, és molt convenient que aquestes estiguin organitzades de manera estructurada, facilitant així la tasca d'inserció de noves dades, actualització de dades existents, cerca de dades de manera eficient, i eliminació de dades. Dins de qualsevol organització, estructurar correctament les seves dades ´es el primer pas per a poder obtindre coneixement.

Per tant, en aquesta pràctica ens centrarem en l'emmagatzematge d'uns conjunts de dades que es troben en format CSV. És gairebé impossible treure informació de conjunts de dades guardats en aquest format, per això dissenyarem i crearem bases de dades relacionals per al posterior anàlisi de les dades i poder extreure conclusions.

Els nostres objectius com a sumari són:

- Entendre els camps dels diversos conjunts de dades.
- Dissenyar un model relacional segons aquests camps i decidir quantes taules implementarem.
- Realitzar un programa en "Python" que llegeixi les dades dels fitxers CSV i emmagatzemi les dades en una base de dades en "Postgres" segons el model relacional decidit.
- Fer un anàlisi exploratori sobre la base de dades creada.

2. Creació de les bases de dades

La primera fase d'aquesta pràctica és la creació de les bases de dades i les seves taules però, abans necessitem conèixer les dades que seran emmagatzemades i les seves relacions.

a. Disseny de la base de dades

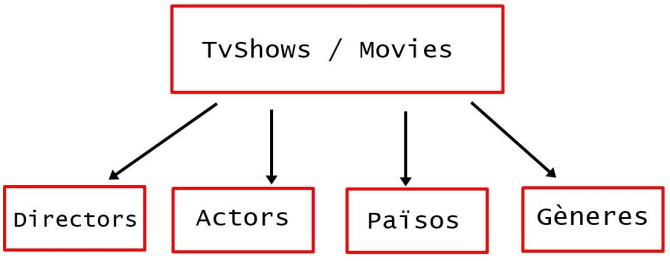
Com hem esmentat, el primer pas és conèixer les dades que hem d'analitzar. Els primers 4 fitxers CSV comparteixen la mateixa estructura:

- ID
- Tipus (sèrie o pel·lícula)
- Títol de la pel·lícula
- Nom i cognoms del director o directors
- Nom i cognoms dels actors
- Països on es va filmar
- Data quan es va afegir la pel·lícula a la plataforma
- Any de llançament
- Classificació d'edats
- Duració (minuts o temporades)
- Gèneres al que pertany
- Descripció

Hem decidit implementar dues taules on tindrem tots els shows. Una dissenyada per les sèries i l'altre per les pel·lícules. D'aquesta manera restringim que siguin els únics tipus possibles, és a dir, que no hi hagi obres teatrals, per exemple. També hem afegit el camp "Plataforma" per identificar a quin servei pertany.

A més, hem dissenyat una taula per els camps que es poden repetir, dit d'una altra manera, per als camps que no són únics. Aixi evitem repetir informació en les bases de dades. Per identificar-los, hem creat un ID artificial per cada camp.
Aquesta decisió ens permetrà en un futur, si és necessari, afegir més informació a aquests camps, per exemple, la data de naixement dels actors. En canvi, a la "duració" no li podriem afegir més informació, per aquest motiu no l'hem implementat en una taula propia.
Per solucionar problemes de camps nulls, hem decidit que el primer ID (0) d'aquestes taules correspongui a null. Entenem que no és el més òptim però, és la solució que hem trobat posat que no ens deixava exportar dades nul·les a la base de dades encara que no sigui "not null".

En síntesi, tenim la següent estructura de taules:



Els dos últims fitxers CSV tenen estructures pròpies i camps únics, ja que són nombres enters, decimals i percentatges. És a dir, cada fitxer disposarà d'una taula pròpia.

Podriem haver crear una taula pròpia per als títols de pel·lícules, no obstant, molts títols no coincideixen amb els dels fitxers CSV anteriors.
Addicionalment, hauriem de canviar l'estructura de les taules anteriors, ja que la data de llançament, actors, etc són propietats de la pel·lícula i les taules "TvShows" i "Movies" quedarien gairebé buides i no tindria tant sentit tindre les dues taules. Ho comentem perquè és una altra implementació possible que hem pensat, però ens hem decantat per aquesta altre.

Per tant, per al fitxer de Rotten Tomatoes tenim els camps:

- Títol de la pel·lícula
- Any de llançament
- Valoració nominal
- Percentatge crítiques positives segons crítics
- Percentatge crítiques positives segons usuaris
- Crítiques positives fetes per crítics
- Crítiques negatives fetes per crítics
- Total de crítiques fetes per crítics
- Total de crítiques fetes per experts

I per FilmTv tenim els camps:

- Títol de la pel·lícula
- Any de llançament
- Puntuació mitjana
- Puntuació mitjana segons crítics
- Puntuació mitjana segons usuaris
- Total de vots

b. Importació de dades

Per crear les bases de dades hem decidit utilitzar “PostgreSQL”, un sistema de gestió de bases de dades relacionals i de codi obert.
També un projecte de “Python” per llegir els fitxers CSV, tractar les dades, crear les taules i exportar les dades a la base de dades.
A continuació, explicarem breument el nostre codi de Python.

Hem implementat 4 funcions per crear les diverses taules:

- Gèneres, directors, actors i països
- TvShows i movies
- Tomatoes
- FilmTv

L’ordre que segueixen les funcions per crear les taules és el mateix, però canvien en l’estructura de les dades que han de llegir, és a dir, la quantitat de camps.
El seu ordre d’execució és:

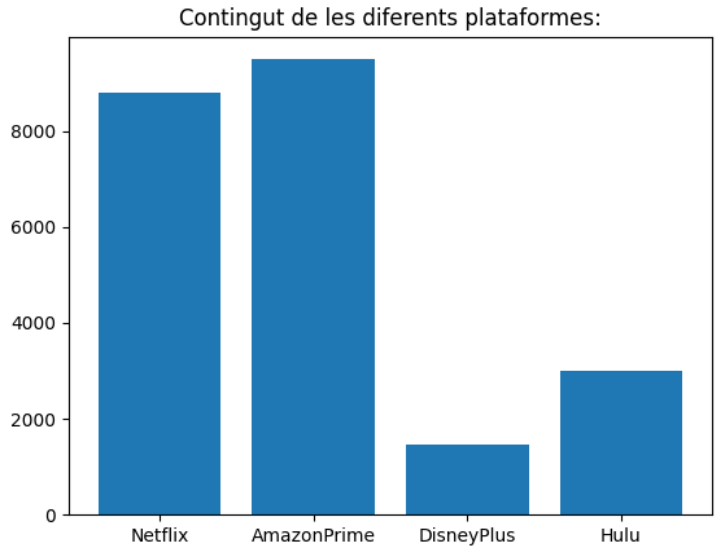
- Creen/seleccionen les consultes SQL per crear taules i inserir registres.
- Realitzen un bucle per llegir les línies una a una dels fitxers CSV.
- Separa els camps i crea tuples per cada camp. Cada posició de la tupla ve donada per una coma “`,`”.
- Es comproven camps buits o que hi hagi camps de més.
- Es comprova la quantitat de registres a inserir, per exemple, si tenim 5 actors, hem de realitzar un insert 5 cops. O es comprova que el registre no estigui repetit.
- S’utilitza `psycopg2` per inserir els camps corresponents a cada taula. Usem `cursor.execute()` per realitzar la comanda i `cursor.fetchall()` per rebre la informació realitzada amb un GET (per exemple, al rebre l’ID d’un director).

3. Anàlisi exploratori

En aquest apartat de la pràctica explicarem com hem resolt les diferents qüestions que se'ns indicaven a l'enunciat, indicant les tècniques emprades, visualitzacions de les dades i conclusions dels resultats obtinguts.

Q1. Quina plataforma de streaming té més contingut? I la que menys?

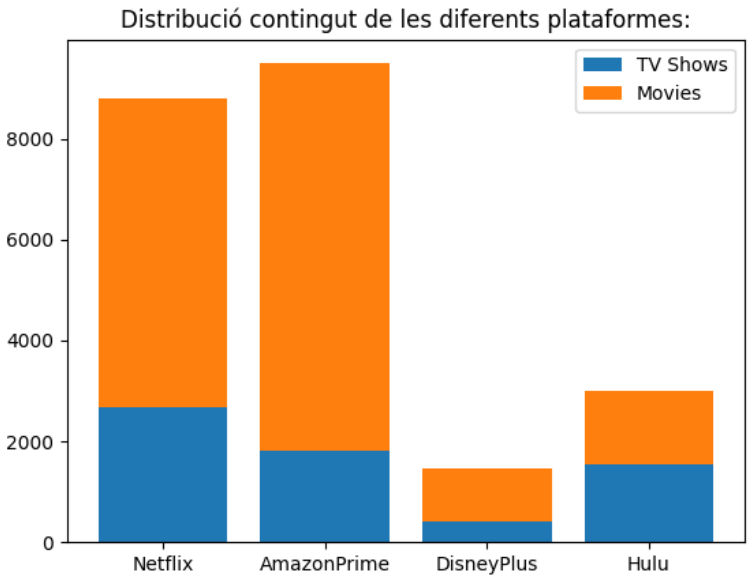
Per aconseguir les plataformes de streaming amb més i menys contingut hem realitzat consultes count filtrant per plataforma a les taules tvshows i movies. Una vegada realitzades les consultes per plataforma, sumem els resultats per obtenir el contingut total per plataforma.
Per finalitzar, amb les funcions min i max de python obtenim les plataformes amb més i menys contingut.
Adjuntem un gràfic per visualitzar el contingut de cada plataforma.



Com podem observar al gràfic, la plataforma amb més contingut és Amazon Primeri la plataforma amb menys contingut és Disney Plus.
Des de el nostre punt de vista ens esperàvem aquest resultat, això és degut a la reputació de cada plataforma. En funció de la quantitat d'usuaris que té cada plataforma i el temps que porta en actiu, també podrà oferir més contingut.

Q2. Quina és la distribució entre pel·lícules i sèries de cada plataforma de streaming? Si sou més aficionats a les sèries, quina preferiríeu?

Per aconseguir la distribució entre pel·lícules i sèries de cada plataforma realitzarem el mateix procediment que a la qüestió anterior, mitjançant les consultes count a les taules tvshows i movies, però en aquest cas no farem la suma total.
Adjuntem un gràfic per visualitzar la distribució de contingut de cada plataforma.



Com podem observar al gràfic, a totes les plataformes majoritàriament predominen les pel·lícules. Aquesta diferència tan gran entre pel·lícules i sèries ens impacta ja que pensàvem que predominarien les sèries.
Nosaltres, personalment, al ser aficionats a les sèries escolliriem Netflix o Amazon Prime, tot i que predominen les pel·lícules, però són la plataforma amb més repertori on escollir.

Q3. Quins són els directors/es més repetits? I els actors/actrius?
Per trobar els directors més repetits entre totes les plataformes, realitzem tres consultes a la base de dades.

Per trobar els directors/actors més repetetits entre totes les plataformes, realitzarem tres consulta per cadascun dels dos.

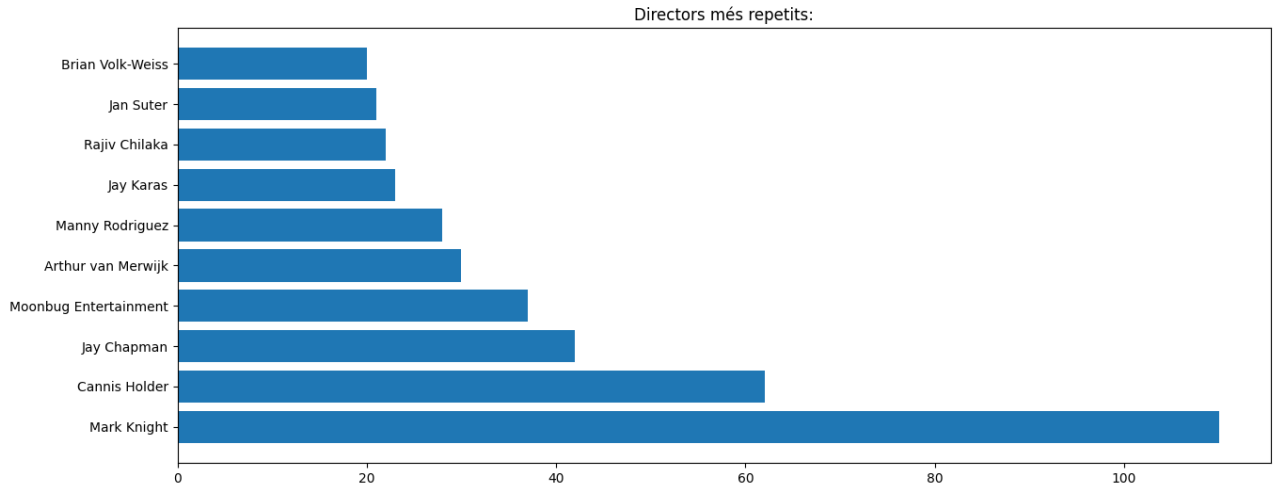
1) Consultes per obtenir el directors/actors de les taules tvshows i movies.
2) Consulta per obtenir el total de directors/actors de la taula directors/actors.

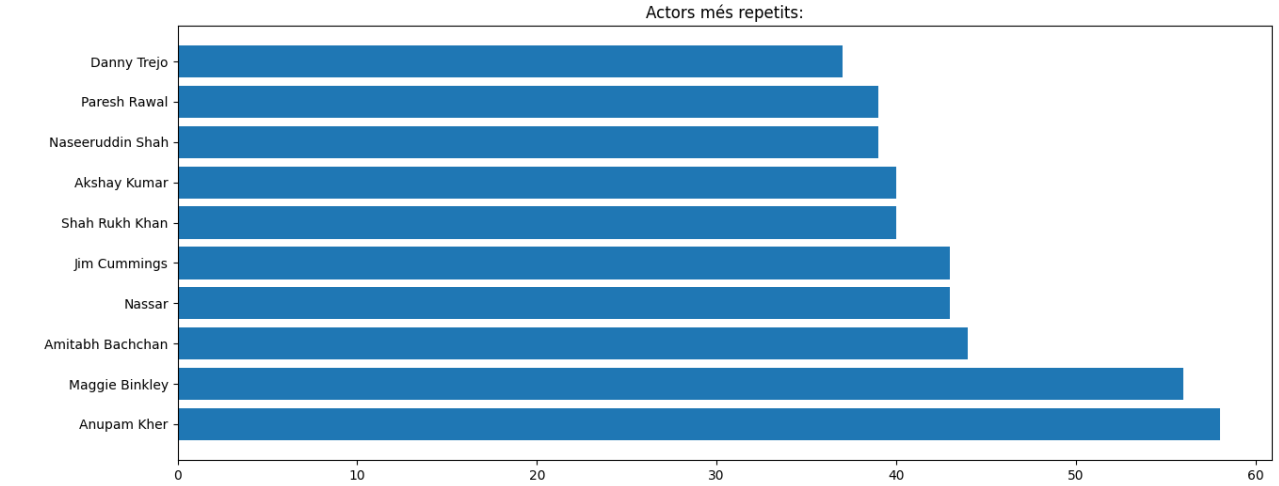
A totes les consultes apliquem la instrucció DISTINCT per obtenir una única vegada la pel·lícula/serie.

Amb la informació de les consultes,

- Ajuntem la informació de les taules tvshows i movies en un vector.
- Comprovem amb el llistat de directors/actors de la consulta 2 quantes vegades es repeteix un director/actor al vector i afegim la tupla (id director, quantitat de repeticions) a un nou vector.
- Per finalitzar, ordenem el nou vector i mostrem els deu primers directors/actors.

Adjuntem els gràfics per visualitzar els deus directors/actors que més es repeteixen.





Com podem observar als diferents gràfics, el director que més es repeteix és Mark Knight i l'actor és Anupam Kher.

Hem cercat informació respecte aquestes dos persones:

- Per part del director, segons la informació dels fitxers inicials, únicament apareix a la taula de pel·lícules.
- Per part de l'actor, és un actor que ha aparegut a més de 200 pel·lícules de Bollywood i també ha aparegut en angleses. Segons la informació dels fitxers inicials, ha realitzat més pel·lícules que sèries.

Q4. Quina és la distribució de duració de les pel·lícules a cada plataforma de streaming? I de les sèries?

Per aconseguir la distribució de duració de les pel·lícules/sèries a cada plataforma, realitzarem una consulta a les taules tvshows i movies. A la consulta apliquem la instrucció DISTINCT per obtenir una única vegada la pel·lícula/serie.

Una vegada realitzada la consulta apliquem un filtratge per tal d'agafar únicament les pel·lícules/sèries que no tenen el camp de duració buit o incorrecte. Amb aquest filtratge hem detectat que hi ha casos on una entrada del fitxer csv indica que és una pel·lícula, però després al camp de duració apareix 'Seasons' en comptes de 'mins'.

Durant aquest procés de filtratge, agrupem les pel·lícules/sèries en diferents grups per tal de visualitzar les dades a les gràfiques posteriors i obtenim la mitjana de minuts/temporades.

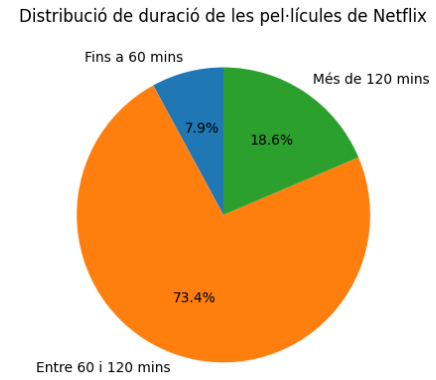
Per part de les pel·lícules:

- Pel·lícules fins a 60 minuts.
- Pel·lícules entre 60 i 120 minuts.
- Pel·lícules de més de 120 minuts.

D'altra banda, a les sèries:

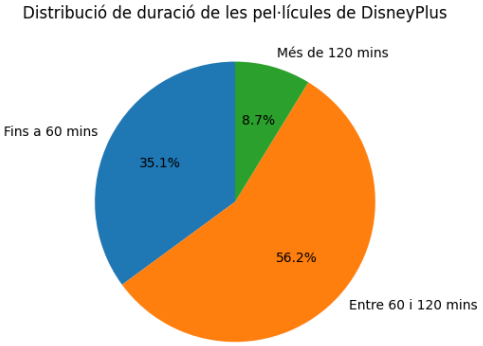
- Sèries de fins a 3 temporades.
- Sèries entre 4 i 6 temporades.
- Sèries de més de 6 temporades.

Adjuntem les gràfiques circulars de les distribucions de duració de les pel·lícules a cada plataforma amb la seva mitjana.

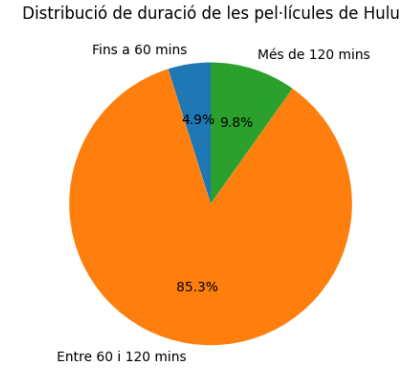


Mitjana: 99.56 minuts

Mitjana: 91.3 minuts



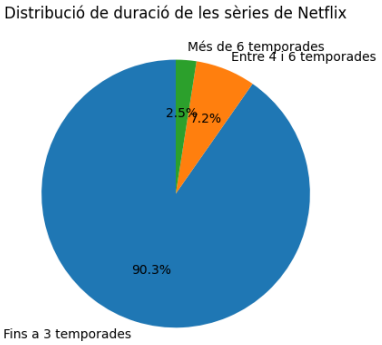
Mitjana: 71.91 minuts



Mitjana: 96.62 minuts

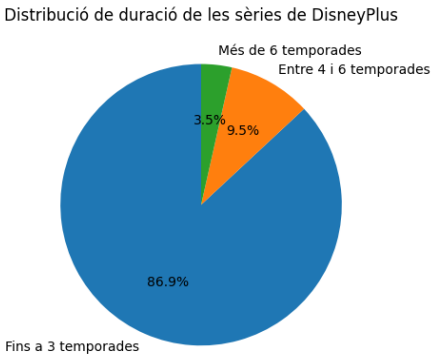
Podem observar que les pel·lícules que més predominen duren entre una i dues hores, sense importar la plataforma. A més, com a mitjana les pel·lícules duren una hora i mitja.

A continuació, les gràfiques circulars respecte les distribucions de duració de les sèries a cada plataforma i la seva mitjana.

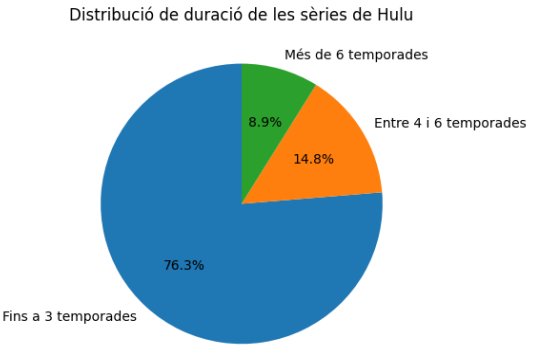


Mitjana: 1.77 temporades

Mitjana: 1.73 temporades



Mitjana: 2.12 temporades



Mitjana: 2.74 temporades

En el cas de les sèries, predominen series que no arriben a tres temporades. A més temporades, menys series hi ha que les tinguin. Pot ser a causa de que els guionistes es queden sense història o que els espectadors perden l'interès.

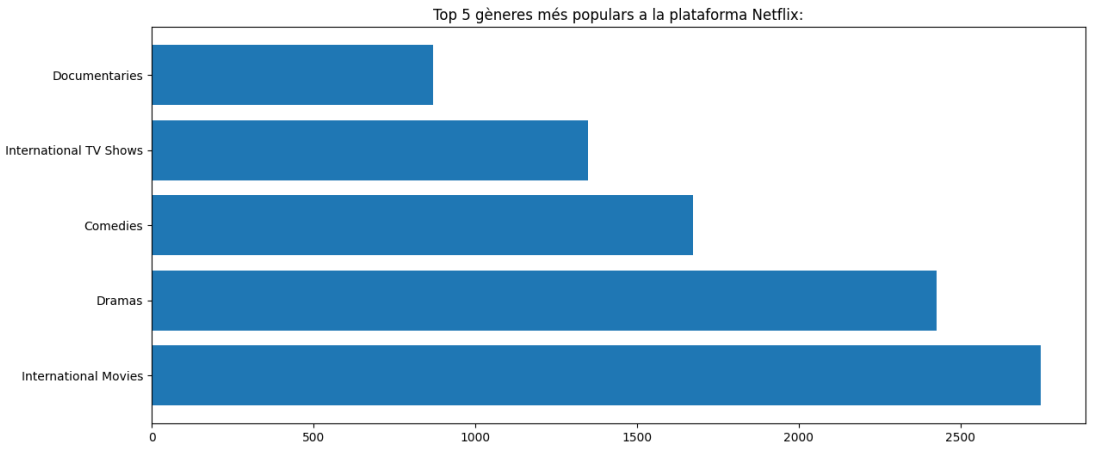
Q5. Quins són els gèneres més populars a cada plataforma de streaming?
Per trobar quins són els gèneres més popular a cada plataforma de streaming realitzarem una consulta a les taules tvshows, movies i dos consultes a la taula gèneres.

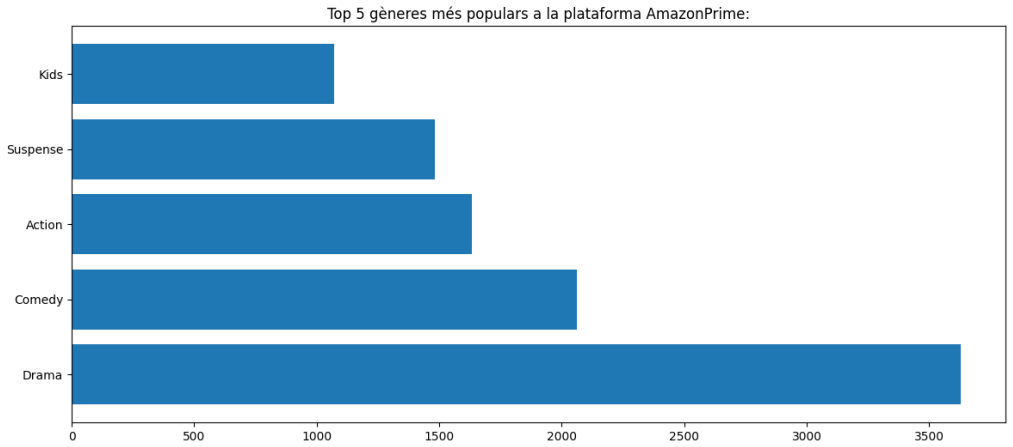
1) Consulta per obtenir el total de gèneres de cada plataforma. A la consulta apliquem la instrucció DISTINCT.
2) Consulta per obtenir tots els identificadors dels gèneres de la taula generes.
3) Consulta per obtenir el nom del gènere una vegada s'ha identificat com un dels cinc més populars de la plataforma.

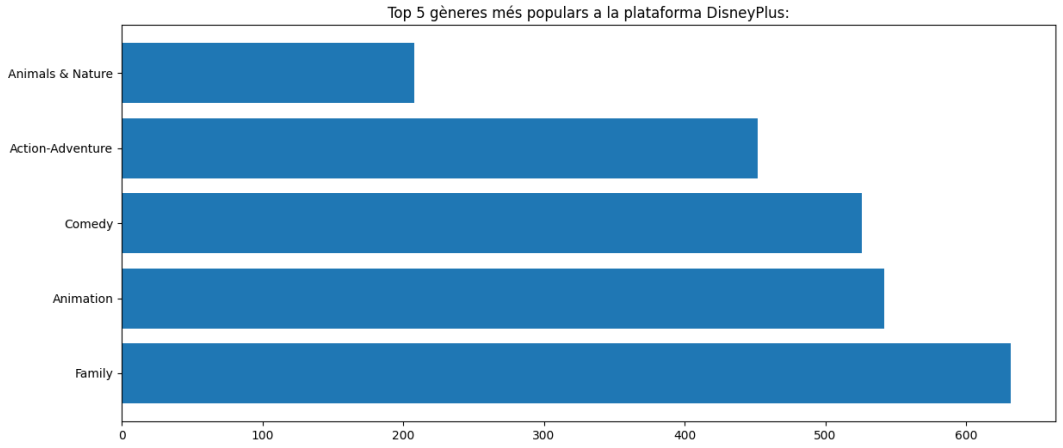
Amb la informació de les consultes,

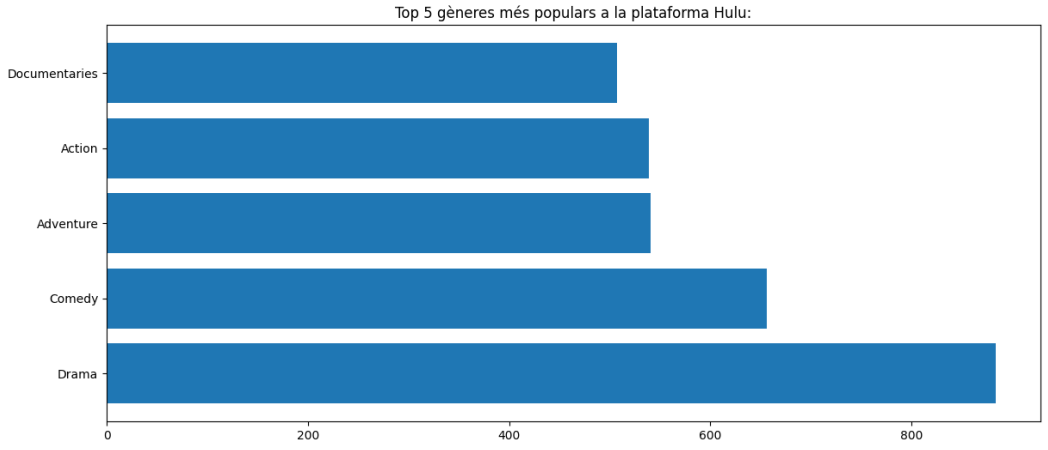
- Ajuntem la informació de les taules tvshows i movies en un vector.
- Comprovem amb el llistat de identificadors de gèneres quantes vegades es repeteix al vector anterior i afegim la tupla (identificador gènere, quantitat de repeticions) a un nou vector.
- Per finalitzar, ordenem el nou vector i amb la consulta 3 mostrem el nom dels cinc gèneres més popular de cada plataforma.

Adjuntem els gràfics dels gèneres més populars a cada plataforma.









Els gèneres que més es repeteixen a les diverses plataformes són: **Action**, **Comedy**, **Drama**, **Documentaries**, **Adventure**. Això indica que independentment de la plataforma que s'utilitza, els usuaris visualitzen pel·lícules/sèries dels mateixos gèneres, especialment el gènere de drama.

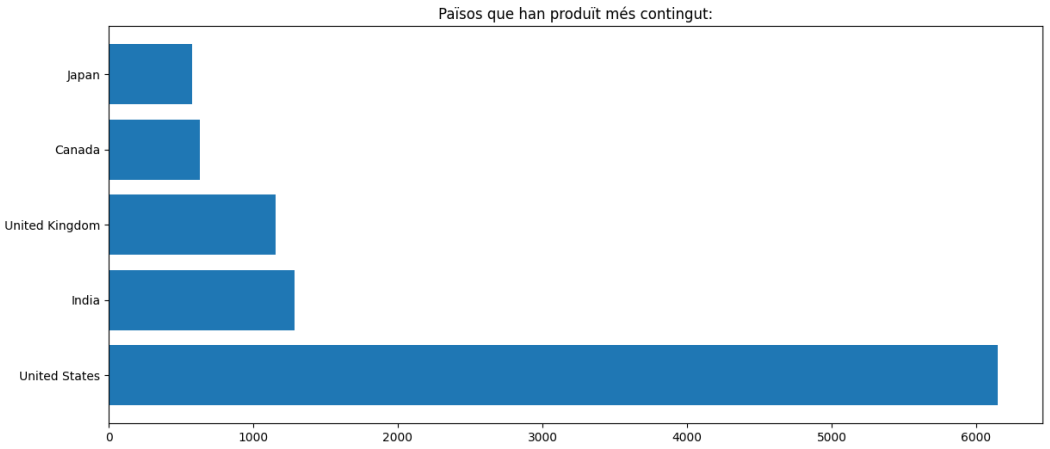
Netflix destaca respecte les altres per les pel·lícules internacionals, en canvi, DisneyPlus i AmazonPrime tenen usos més familiars que inclouen nens.

Q6. Quins són els països que han produït més contingut? Es similar entre les diverses plataformes de streaming?

Per ambdues qüestions hem aplicat el mateix procediment que a la qüestió cinc anterior.

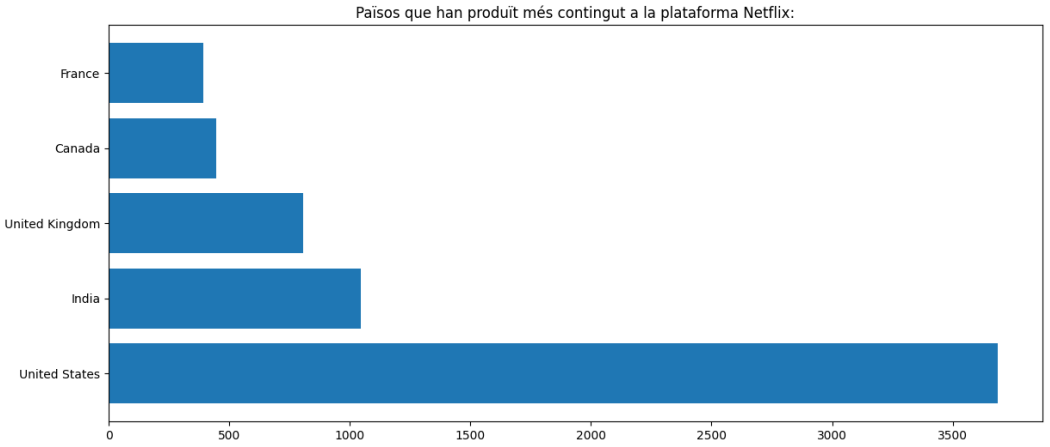
1) Consulta per obtenir el llistat de països de les taules tvshows i movies, aplicant la instrucció DISTINCT. A la primera qüestió sense filtrar per catàleg, a la segona qüestió filtrant per catàleg.
2) Consulta per obtenir els identificadors dels països.
3) Consulta per obtenir el nom del país una vegada forma part del top cinc països que més contingut a produït.

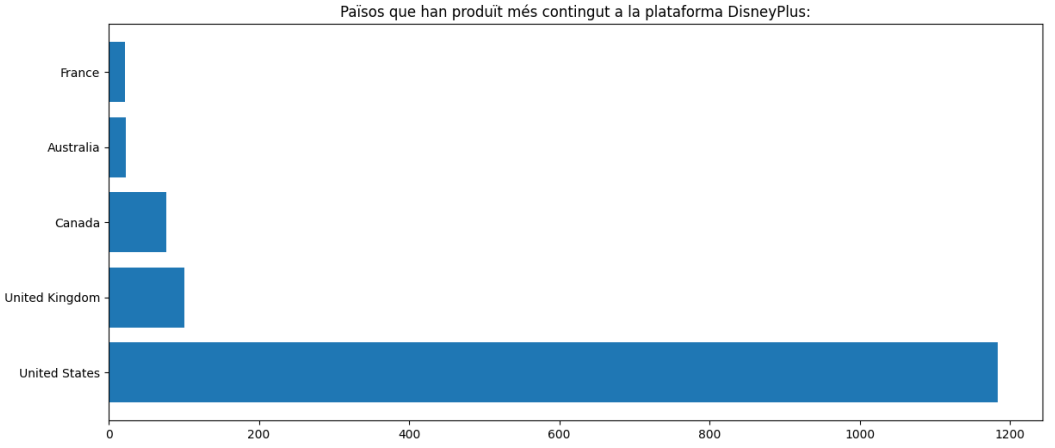
Adjuntem el gràfic dels països que han produït més contingut.

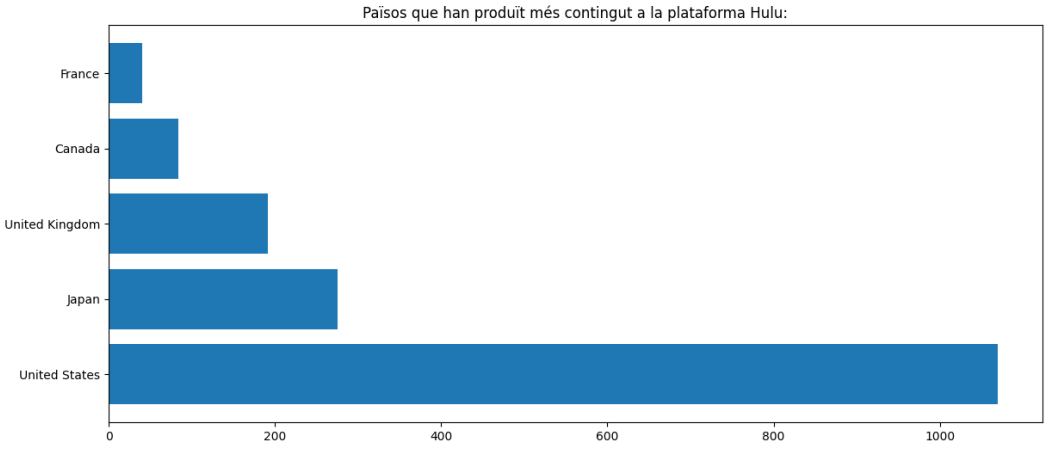


Com podem observar al gràfic, EEUU és la potència mundial en quant a producció de pel·lícules i sèries es refereix. Sumant la resta de països del top cinc (Japó, Canadà, Regne Unit, India) fan una mica més de la meitat del que ha produït EEUU.

Adjuntem els gràfics dels països que han produït més contingut a cada plataforma.





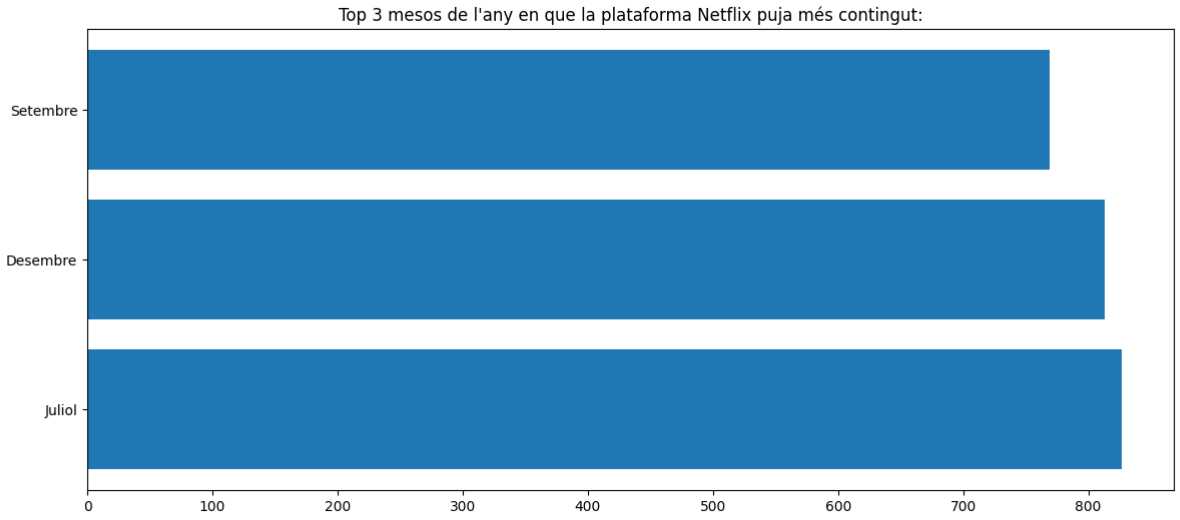


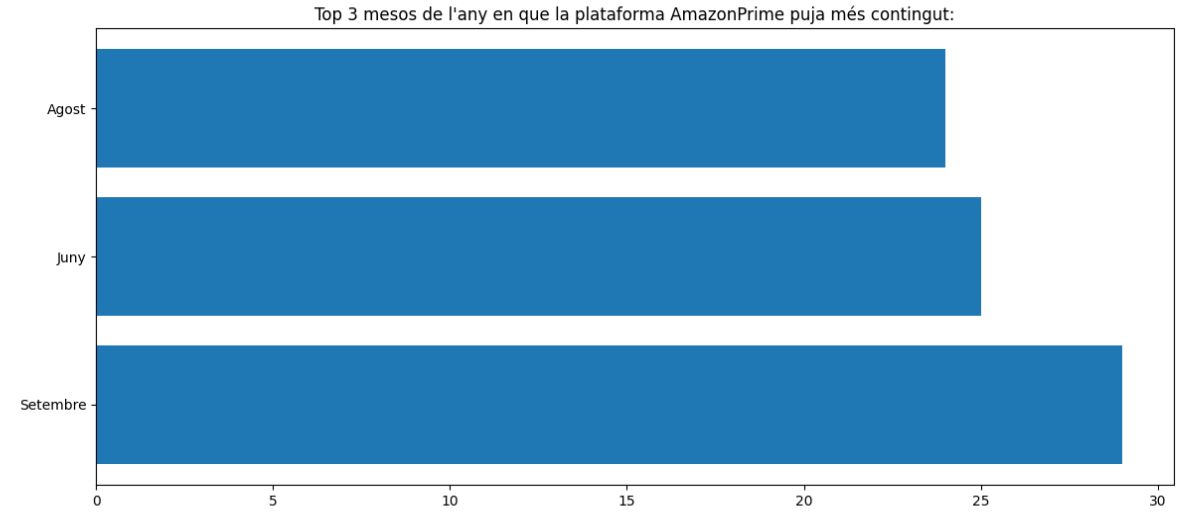
Els Estats Units és el país que més predomina com era d'esperar després de veure el gràfic general. Com a sorpresa veiem que AmazonPrime inclou moltes pel·lícules de l'Índia, no presenta una diferencia tan gran respecte als EEUU com les altres plataformes.
Si recordem l'anàlisi anterior, s'afirma que a Netflix predomina el contingut internacional, al tindre més contingut d'altres països, encara que predomina com a país els Estats Units.

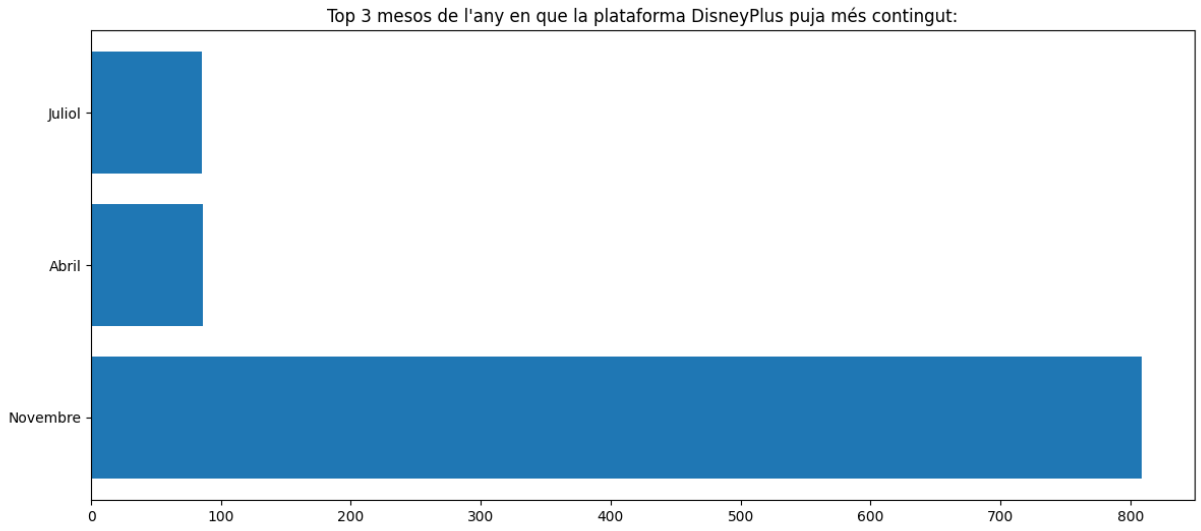
Q7. Quins són els mesos de l'any en que les plataformes de streaming afegeixen més nou contingut?

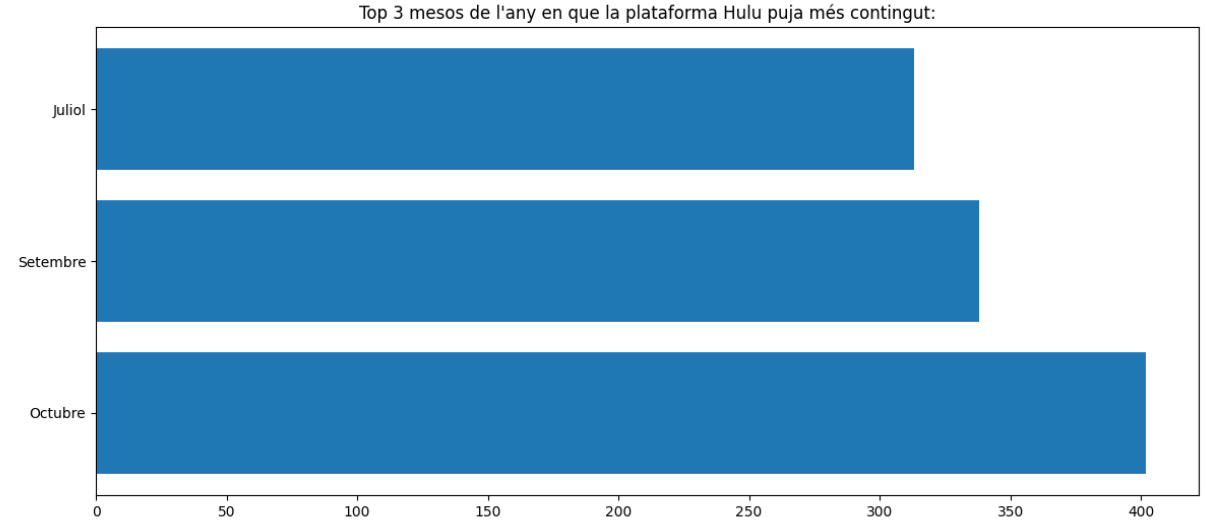
L'estructura per fer l'anàlisi és el mateix que en els anàlisi anteriors. En aquest cas, la qüestió principal és transformar la data i obtenir només el mes.

Adjuntem els gràfics dels top tres mesos de l'any en que les plataformes pugen més contingut.







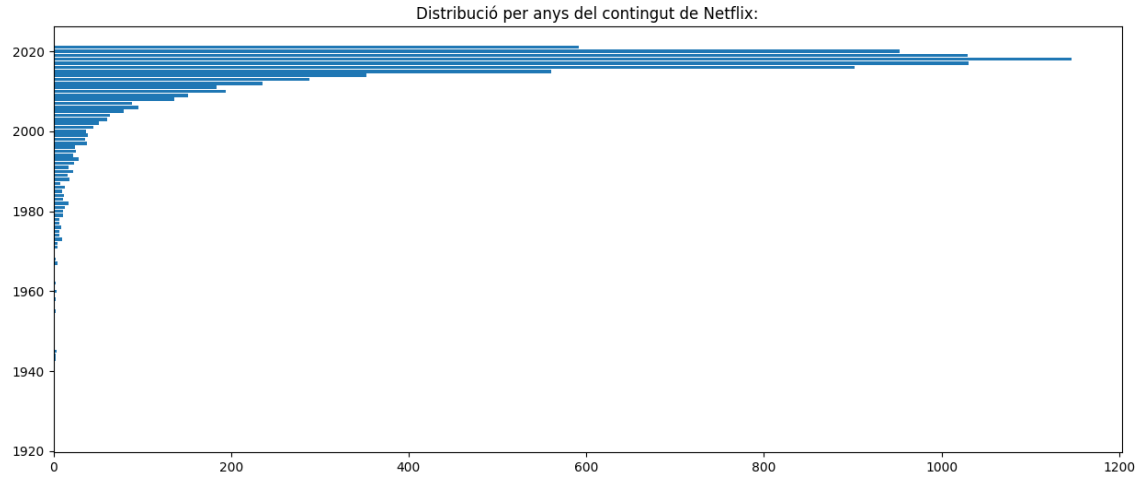


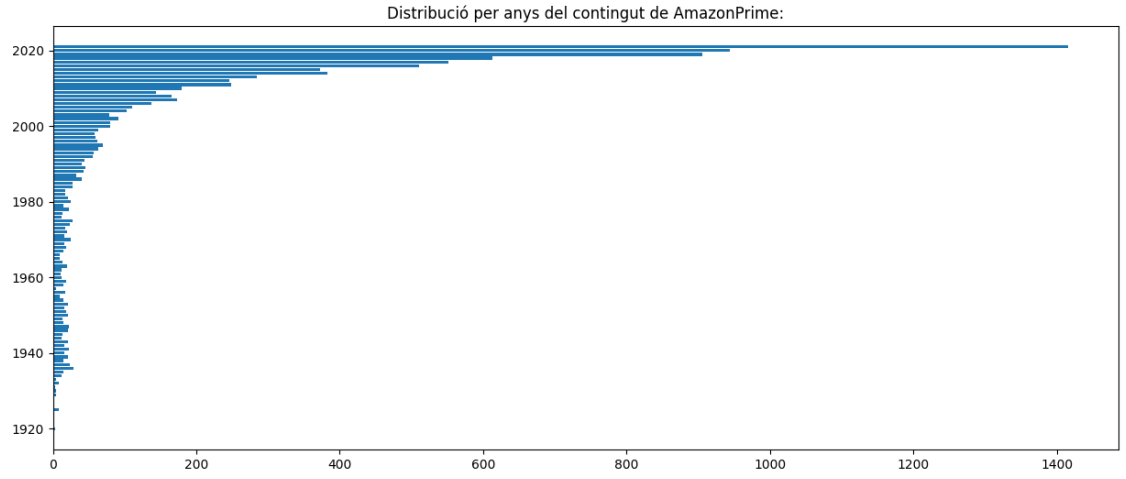
L'època en que prefereixen per pujar contingut és l'estiu. Suposem que es degut a que molta gent té vacances i aprofita per veure pel·lícules/series.
També observem mesos que s'apropen a les festes de Nadal, especialment en la plataforma de AmazonPrime. Com hem observat en un anàlisi anterior, aquesta plataforma es decanta més per contingut familiar, com poden ser pel·lícules Nadalenques.

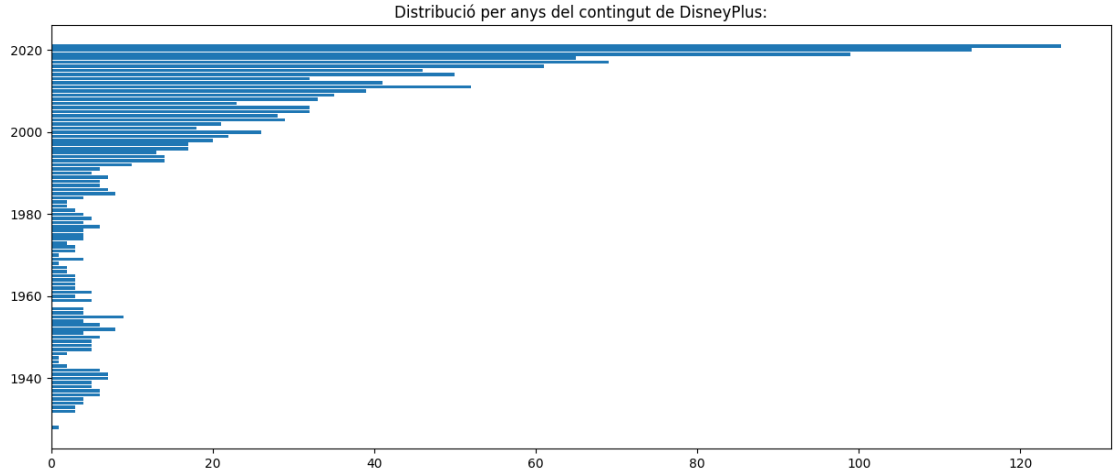
Q8. Quina és la distribució per anys del contingut de cada plataforma de streaming? Quina és la que té un contingut més nou? I la que té un contingut més vell? La tendència és manté segons si el contingut són pel·lícules o sèries?

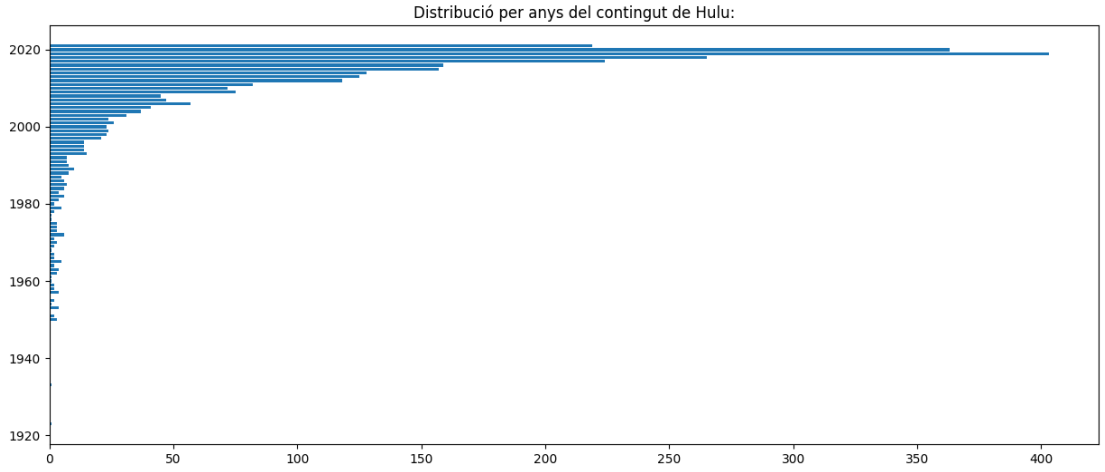
L'estructura per fer l'anàlisi és el mateix que en els anàlisi anteriors. En aquest cas, la qüestió principal és observar la distribució per anys del contingut de cada plataforma de streaming i analitzar-lo.

Adjuntem els gràfics de la distribució per anys del contingut de les plataformes:









Com podem observar en el gràfic de la distribució per anys del contingut de totes les plataformes a mesura que passen els anys hi han més pel·lícules en aquestes ja que és degut que cada vegada hi han més sèries i pel·lícules.
Podem afirmar que la plataforma que té més contingut nou i, a la vegada més contingut vell, és Disney Plus tot i que en contingut nou la segueixen de prop Netflix i Amazon Prime.

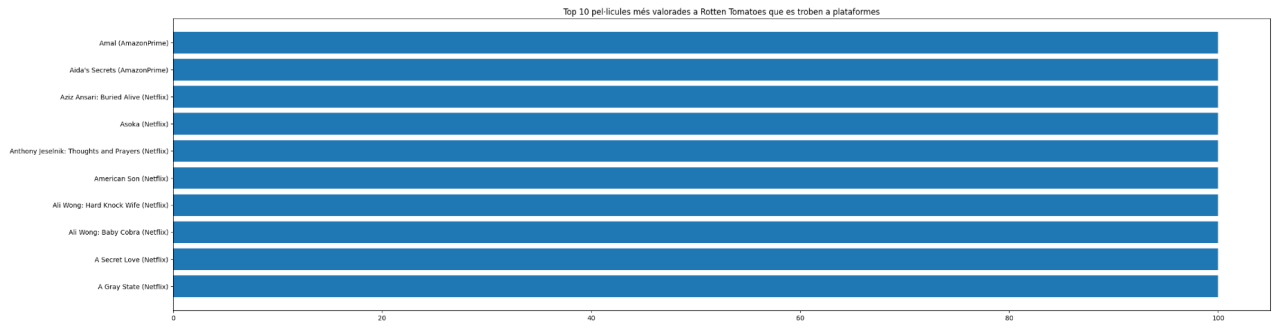
Q9. A quin perfil de persones recomanaries cada plataforma de streaming?
Després de reflexionar com podriem recomanar cada plataforma de streaming als diferents usuaris hem arribat a la conclusió de que es podria fer de diferents maneres:

1) Recomanar plataforma per quantitat de pel·lícules:
Podriem recomanar la plataforma en funció de si la persona és cinèfil o serièfil. Basant-se en la Q2:
Si una persona és més serièfil seria recomanable tenir Netflix però, si per contrari és cinèfil la plataforma recomanada és Amazon Prime.

2) Recomanar plataforma per gèneres més populars:
Podriem recomanar la plataforma en funció dels gèneres que li agradin més. Basant-se en la Q5:
Si el top 5 de gèneres que li agraden a una persona són:

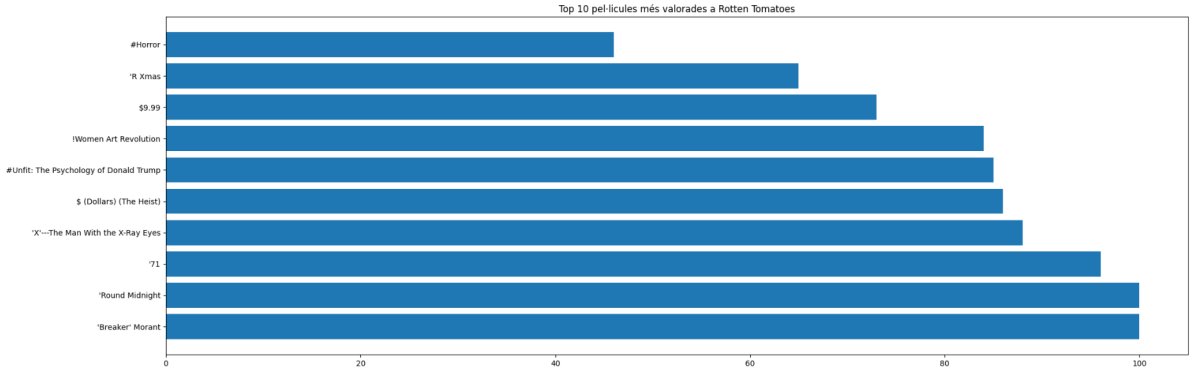
- Documentaries, International TV Shows, Comedias, Dramas i International Movies: La plataforma recomanada serà **Netflix**.
- Kids, Suspense, Action, Comedy, Drama: La plataforma recomanada serà **Amazon Prime**.
- Animals & Nature, Action-Adventure, Comedy, Animation, Family: La plataforma recomanada serà **Disney Plus**.
- Documentaries, Action, Adventure, Comedy, Drama: La plataforma recomanada serà **Hulu**.

Q10. Quines són les pel·lícules més valorades a Rotten Tomatoes? I a FilmTV? Es troben en plataformes de streaming?
L'estructura per fer l'anàlisi és el mateix que en els anàlisi anteriors. En aquest cas, la qüestió principal és mirar quines són les pel·lícules més valorades i, posteriorment, mirar si es troben en plataformes de streaming.
Adjuntem el gràfic de les pel·lícules més valorades a Rotten Tomatoes:



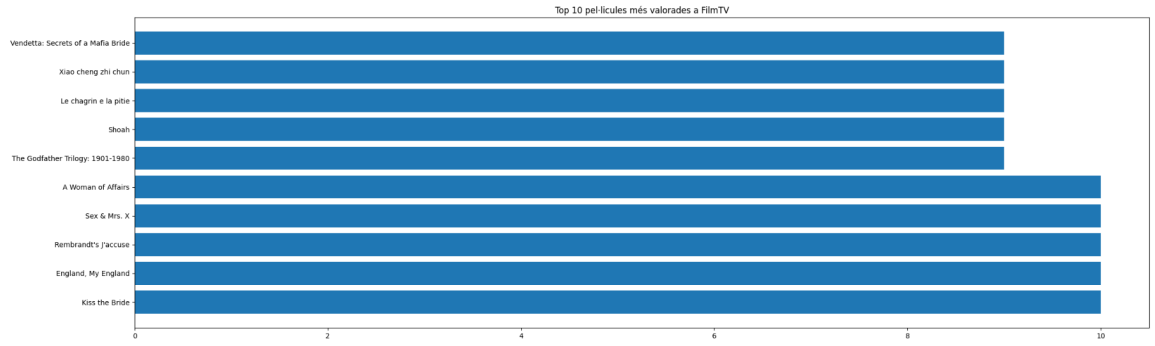
En aquest cas el Top 10 de les pel·lícules més valorades a Rotten Tv tenen la mateixa valoració. De fet n'hi ha més de 10 que tenen una valoració de 10 però hem agafat les 10 primeres ordenades per ordre alfabètic. El Top 10 l'encapçala la pel·lícula A Gray State i el tanca la pel·lícula Aziz Ansari: Buried Alive.

Adjuntem el gràfic de les pel·lícules més valorades a Rotten Tomatoes que es troben a plataformes:



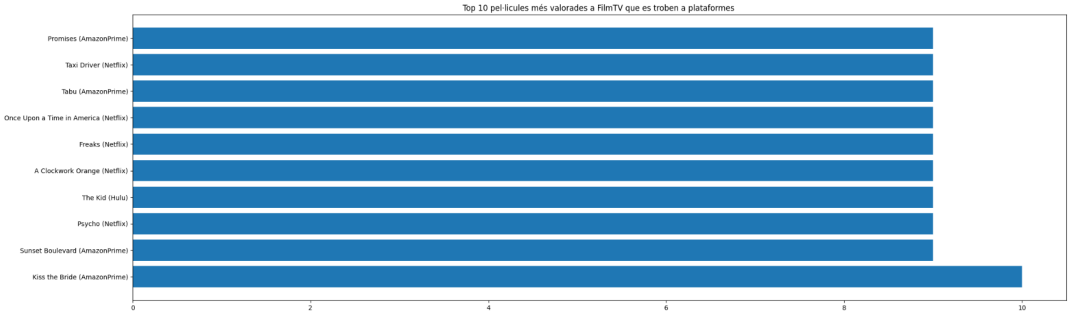
Com podem observar el Top 10 de les pel·lícules més valorades a Rotten Tomatoes que es troben en plataformes l'encapçala 'Breaker' Morant, seguit per 'Round Midnight i '71. Aquest Top 10 el tanca la pel·lícula #Horror.

Adjuntem el gràfic de les pel·lícules més valorades a FilmTv:



Observem que el Top 10 de les pel·lícules més valorades a FilmTV l'encapçala Kiss the Bride i després tenim varies que tenen la mateixa valoració.

Adjuntem el gràfic de les pel·lícules més valorades a FilmTv que es troben a plataformes:

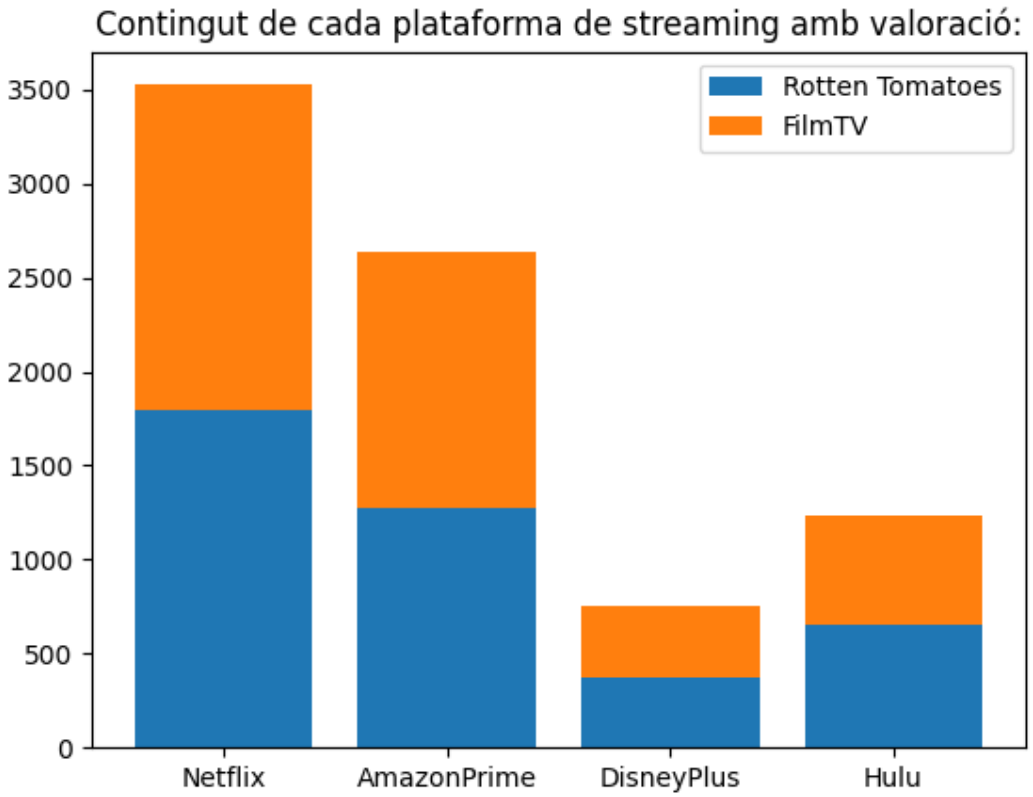


El Top 10 de les pel·lícules més valorades a FilmTV que es troben a plataformes l'encapçala Kiss the Bride i després tenim varies que tenen la mateixa valoració.

Q11. Quant contingut de cada plataforma de streaming té alguna valoració (a Rotten Tomatoes o a FilmTV)?

L'estructura per fer l'anàlisi és el mateix que en els anàlisi anteriors. En aquest cas, la qüestió principal és mirar quina quantitat de pel·lícules tenen valoració a Rotten Tomatoes i a Film TV que estan integrades a cadascuna de les plataformes de streaming.

Adjuntem el gràfic del contingut de cada plataforma de streaming que té alguna valoració:



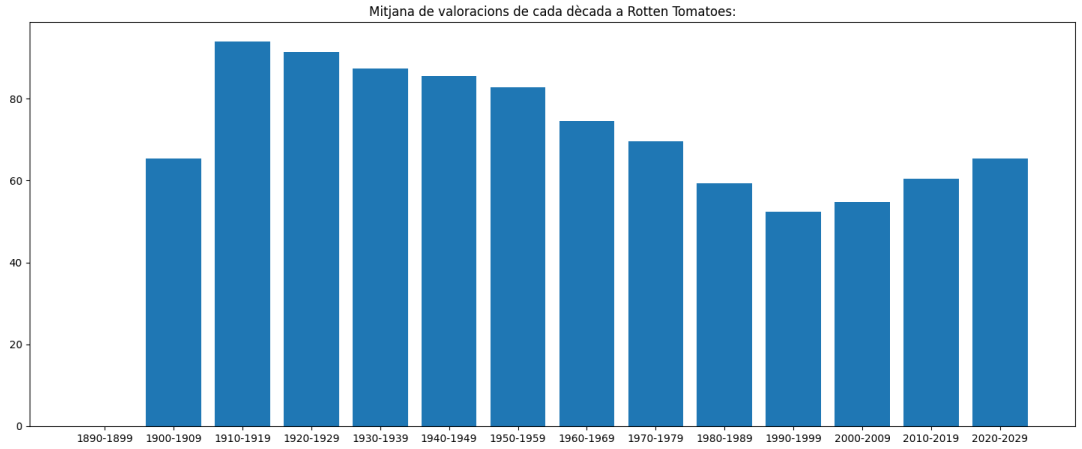
Com podem observar en el gràfic la plataforma que té més contingut amb alguna valoració tant a Rotten Tomatoes com a Film TV és Netflix, seguit de Amazon Prime d'aprop en qüestió de Rotten Tomatoes.

Per contra, la plataforma que té menys contingut valorat és DisneyPlus.

Q12. Quina dècada va tindre millors pel·lícules?

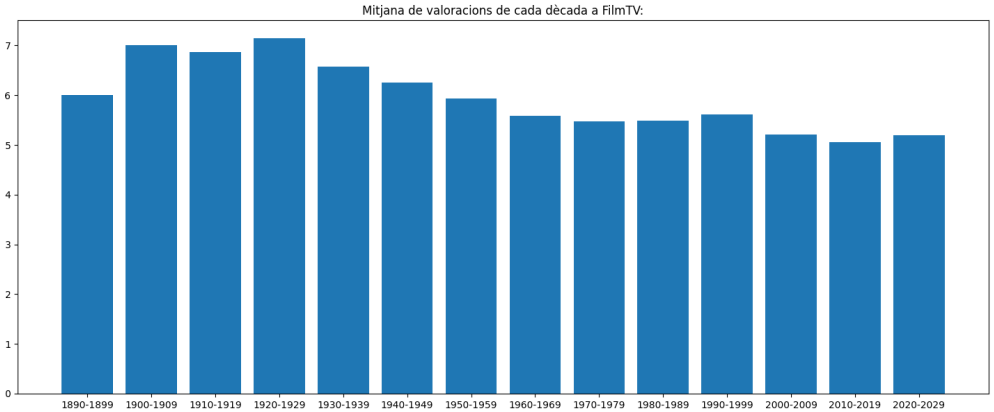
L'estructura per fer l'anàlisi és el mateix que en els anàlisi anteriors. En aquest cas, la qüestió principal és mirar quina dècada va tenir millors pel·lícules segons Rotten Tomatoes i segons Film TV.

Adjuntem el gràfic de la mitjana de valoracions de cada dècada a Rotten Tomatoes:



Si ens guiem per les valoracions de Rotten Tomatoes la dècada que va tenir millors pel·lícules és entre 1910-1919. El que podem veure és que a mesura que anava passant el temps la mitjana de les valoracions era inferior ja que, desde el nostre punt de vista, els crítics cada vegada eren més durs amb les valoracions tot i que podem veure que aquestes últimes dècades la tendència és valorar més alt.

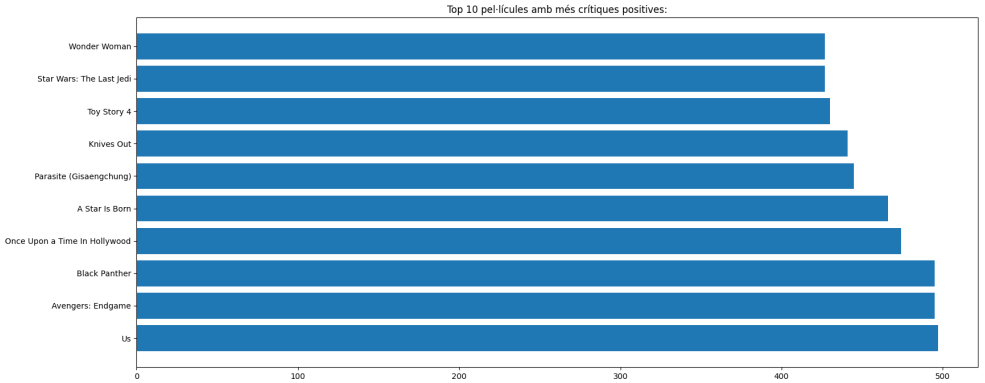
Adjuntem el gràfic de la mitjana de valoracions de cada dècada a Film TV:

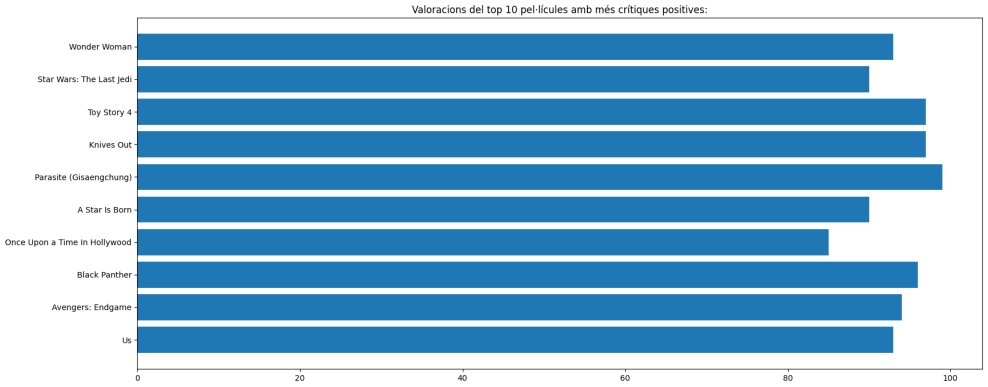


Com podem observar si ens basem en les valoracions de Film TV la dècada que va tenir millors pel·lícules és entre 1920-1929. El que veiem és que tot i que el pas dels anys ha afectat en el descens de la mitjana de les valoracions sempre han estat molt similars.

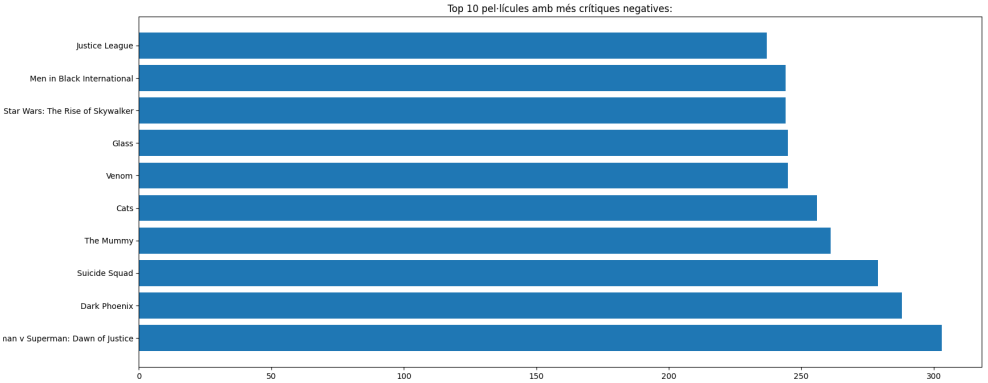
Q13. Les pel·lícules amb més vots/crítiques són les que tenen millors valoracions?
Per realitzar aquest anàlisi, consultem les pel·lícules millor i pitjor valoració i, posteriorment les que tenen més vots d'aquestes pel·lícules. D'aquestes consultes, seleccionarem 10. Aquesta consulta la realitzem tant per Tomatoes com per FilmTV.

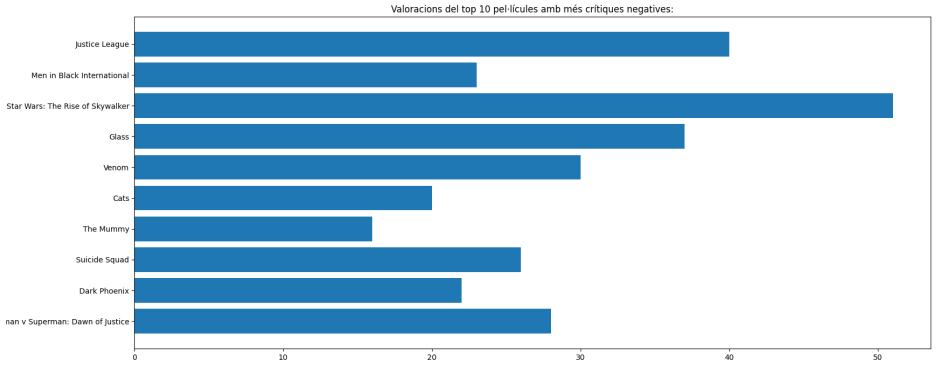
Comencem amb Rotten Tomatoes. Observem que les 3 pel·lícules amb més crítiques positives són *Us*, *Endgame* i *Black Panther*. En la segona gràfica podem comprovar que aquestes pel·lícules no són les que tenen millor valoració entre aquest top 10. Les que tenen millor valoració són *Parasite*, *Toy Story 4* i *Knives Out*. Podem concluir que les que tenen més vots, no són les que tenen millors valoracions.



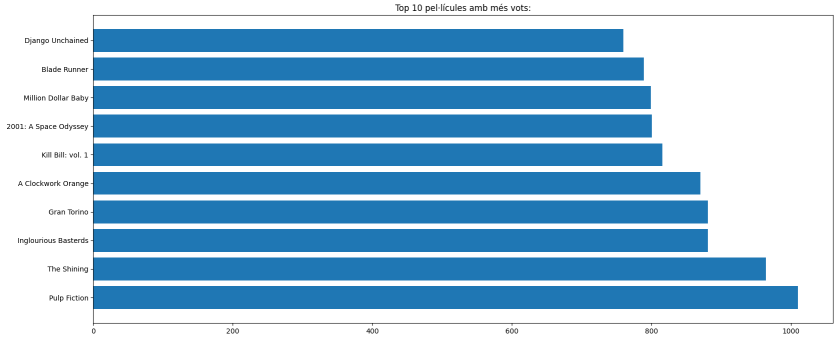


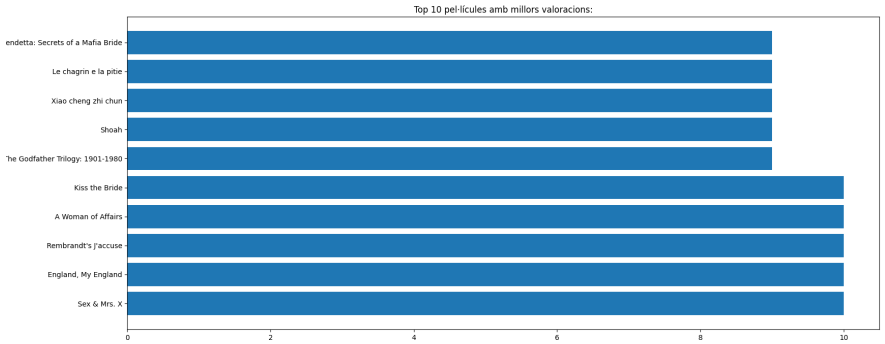
Si ara observem les mateixes gràfiques amb més vots negatius, comprovem que tampoc són les que tenen pitjor valoració.



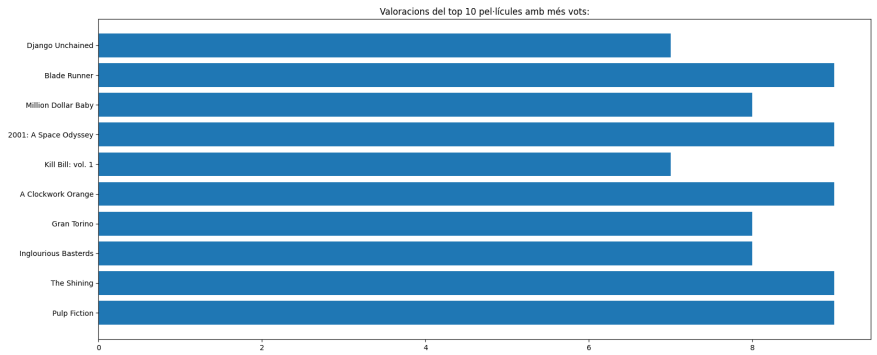


Per a FilmTV hem obtingut les pel·lícules amb més vots i per una altra banda, les que tenen millor valoració. Observem que les pel·lícules amb més vots no coincideixen amb les que tenen més valoració.





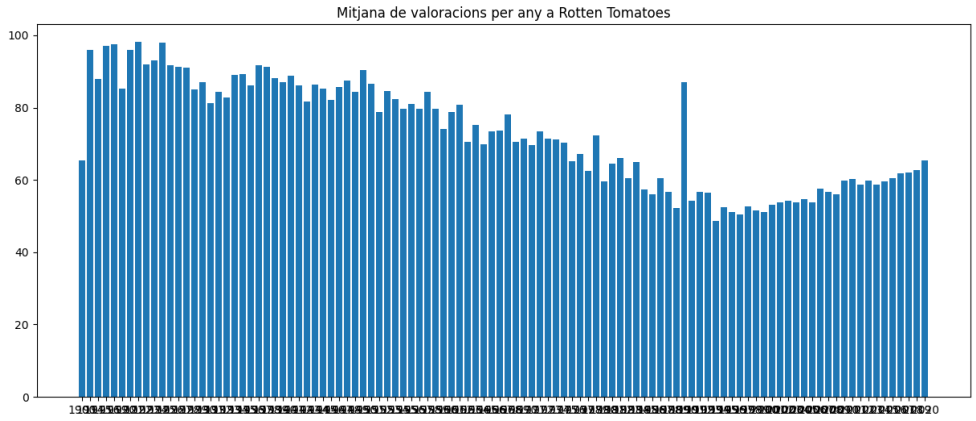
Aquesta última gràfica ens mostra les valoracions de les pel·lícules amb més vots. Comprovem que no segueixen una relació entre vots-valoració.



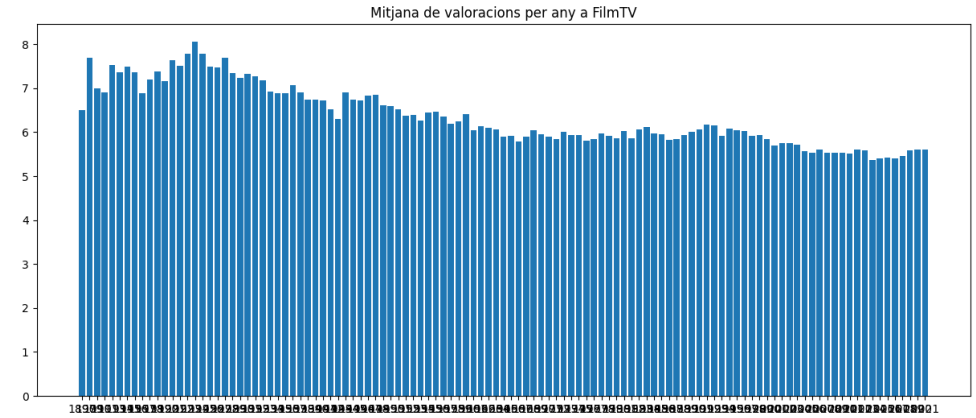
Q14. Existeix una relació entre l'any de les pel·lícules i la seva valoració? Podem dir que les pel·lícules noves tenen millors valoracions que les més antigues?

L'estructura per fer l'anàlisi és el mateix que en els anàlisi anteriors. En aquest cas, la qüestió principal és mirar si existeix una relació entre l'any de les pel·lícules i la seva valoració.

Adjuntem el gràfic de la mitjana de valoracions per any de Rotten Tomatoes:



Adjuntem el gràfic de la mitjana de valoracions per any de Rotten Tomatoes:

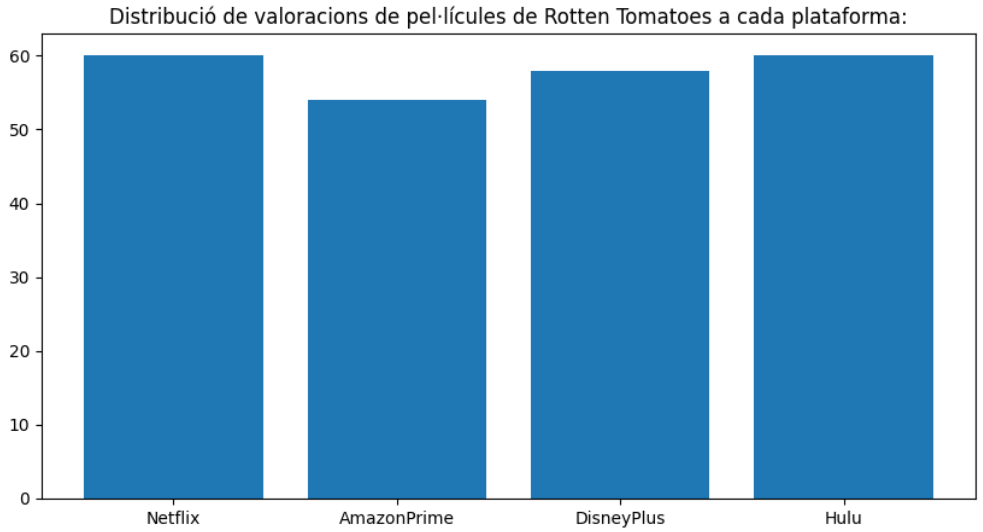


Com podem observar en els dos gràfics podem dir que les pel·lícules antigues tenen millors valoracions que les més noves. Segons les nostres conclusions és degut a que actualment les valoracions són més estrictes i això comporta a que la mitjana sigui inferior que antigament.

Q15. Quina plataforma té millors pel·lícules? I pitjors? Quina és la distribució de valoracions de pel·lícules en cada plataforma de streaming?

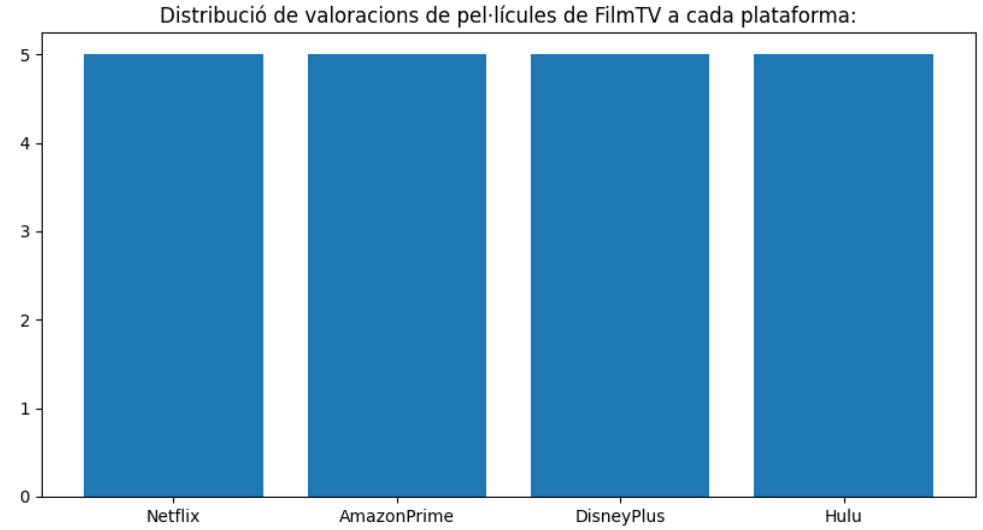
L'estructura per fer l'anàlisi és el mateix que en els anàlisi anteriors. En aquest cas, la qüestió principal és mirar quina plataforma té millors i pitjors pel·lícules.

Adjuntem el gràfic de la distribució de valoracions de pel·lícules de Rotten Tomatoes a cada plataforma:



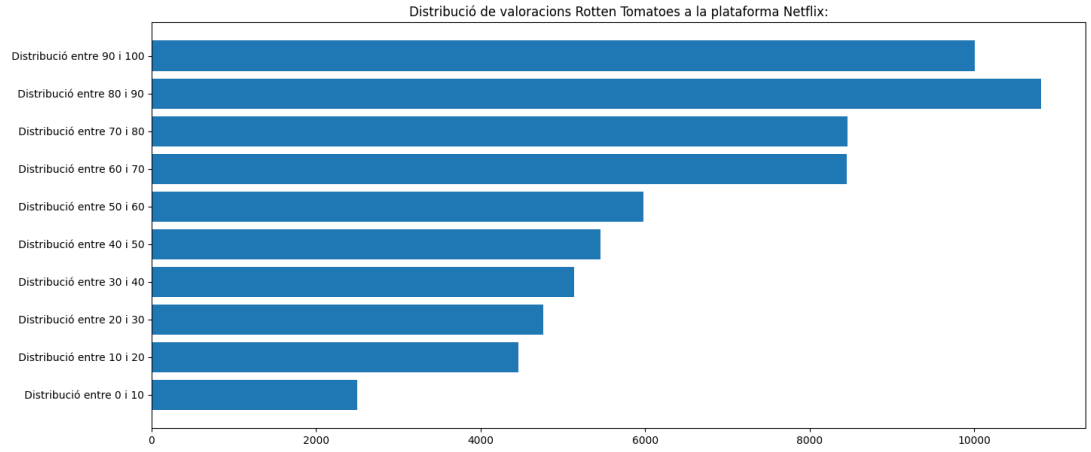
Com podem observar segons Rotten Tomatoes Netflix té les millors pel·lícules valorades seguit de molt aprop per Hulu i Amazon Prime té les pitjors valoracions de pel·lícules.

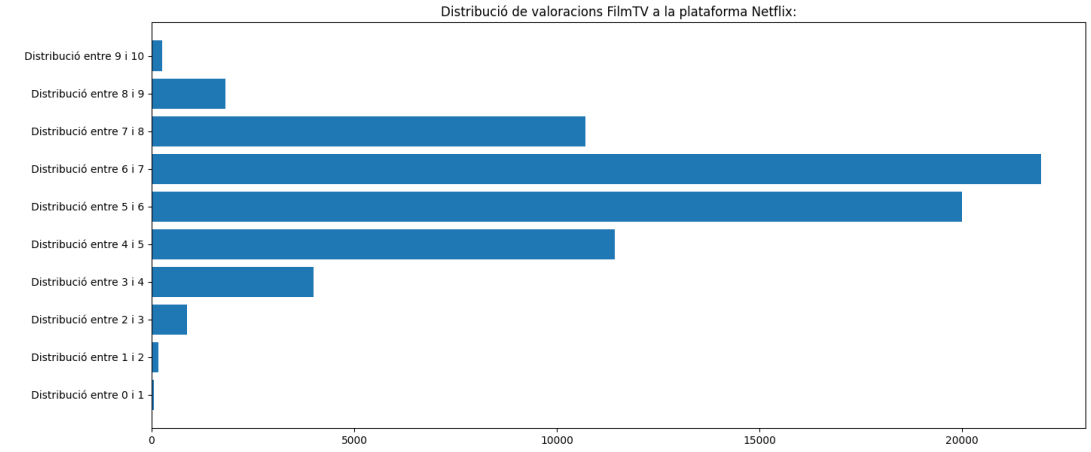
Adjuntem el gràfic de la distribució de valoracions de pel·lícules de Rotten Tomatoes a cada plataforma:



Com podem observar segons Film TV totes les plataformes tenen valoracions molt similars ni que Disney Plus té les millors valoracions i Amazon Prime les pitjors.

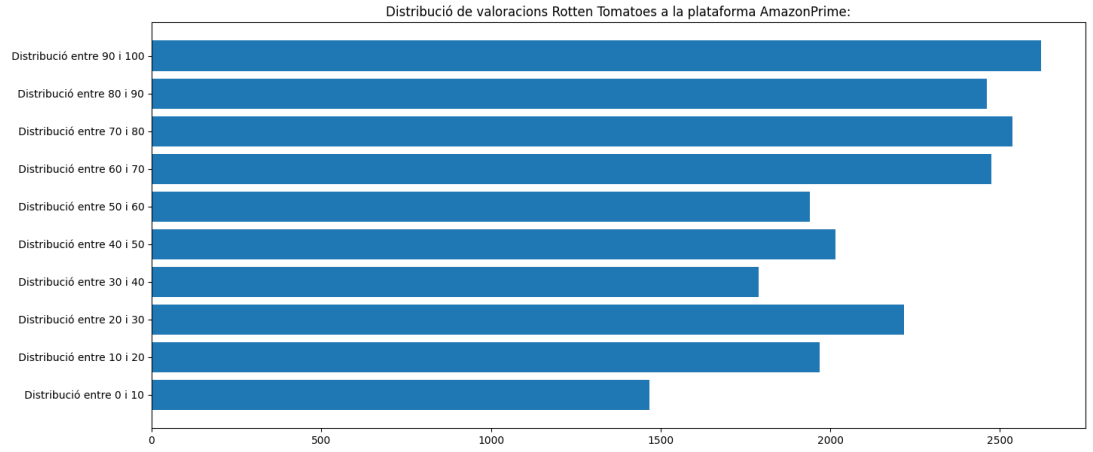
L’estructura per fer l’anàlisi és el mateix que en els anàlisi anteriors. En aquest cas, la qüestió principal és mirar la distribució de valoracions de pel·lícules en cada plataforma de streaming.
Adjuntem el gràfic de la distribució de valoracions de pel·lícules de Netflix segons Rotten Tomatoes i Film TV:

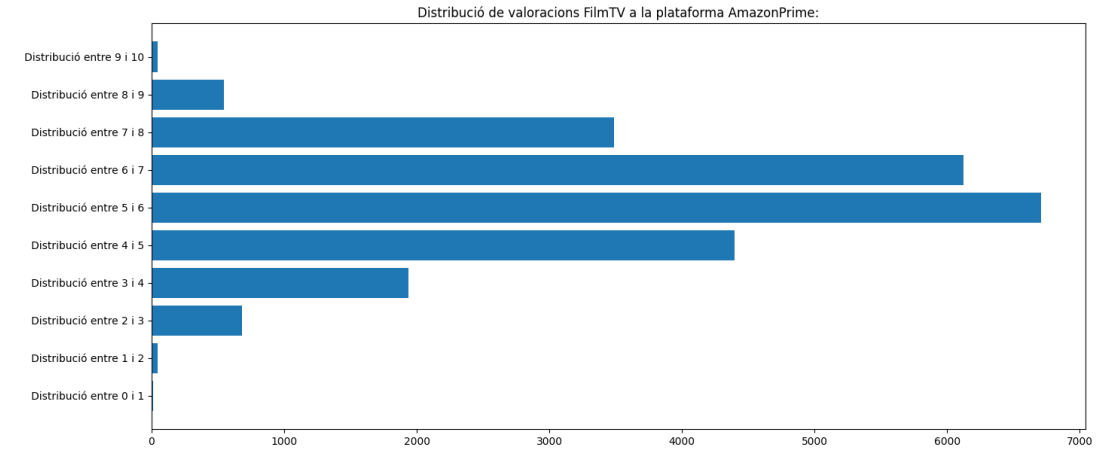




Com podem observar, segons Rotten Tomatoes predominen les valoracions entre 80 i 90 de la plataforma Netflix mentre que segons Film TV predominen les valoracions entre 6 i 7. Podem dir que les valoracions entre Rotten Tomatoes i Film TV són molt semblants.

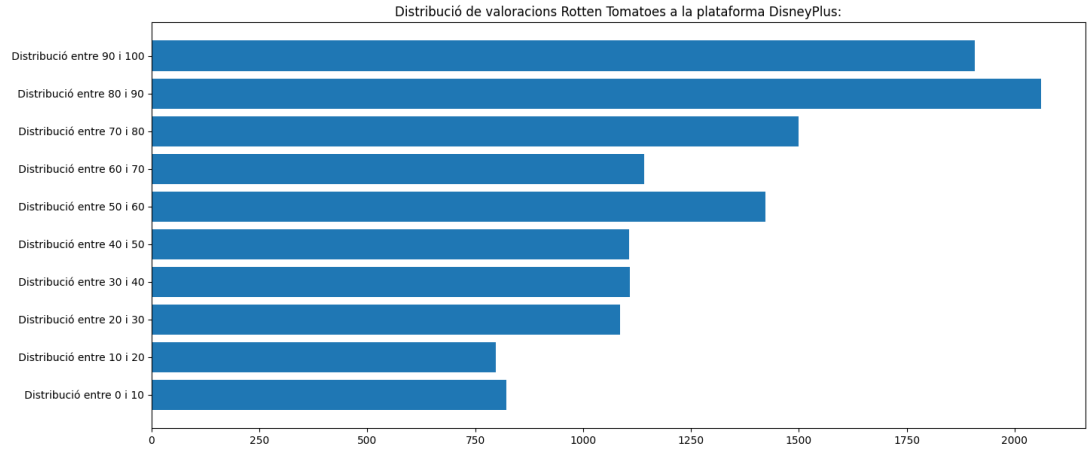
Adjuntem el gràfic de la distribució de valoracions de pel·lícules de Amazon Prime segons Rotten Tomatoes i Film TV:

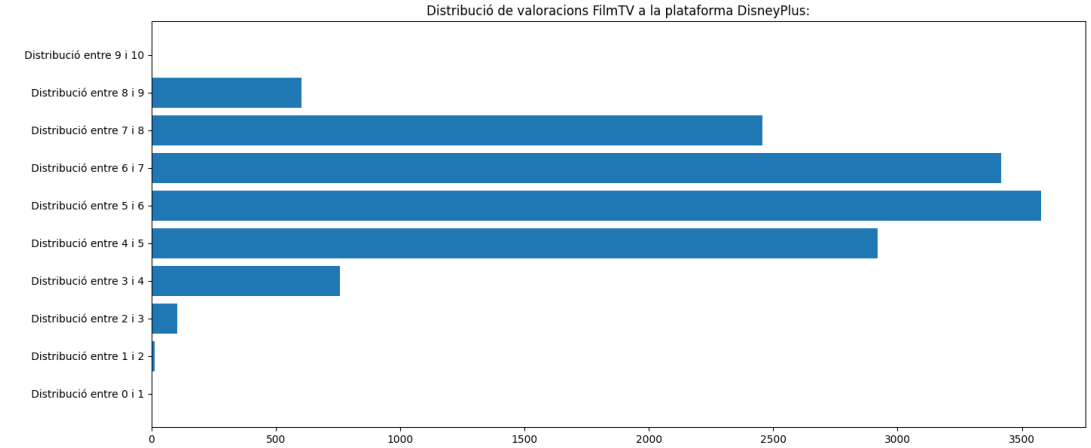




Com podem observar, segons Rotten Tomatoes predominen les valoracions entre 90 i 100 de la plataforma Amazon Prime mentre que segons Film TV predominen les valoracions entre 5 i 6. Podem dir que les valoracions entre Rotten Tomatoes i Film TV són molt diferents ja que a Rotten valoren les pel·lícules amb més nota que a Film TV.

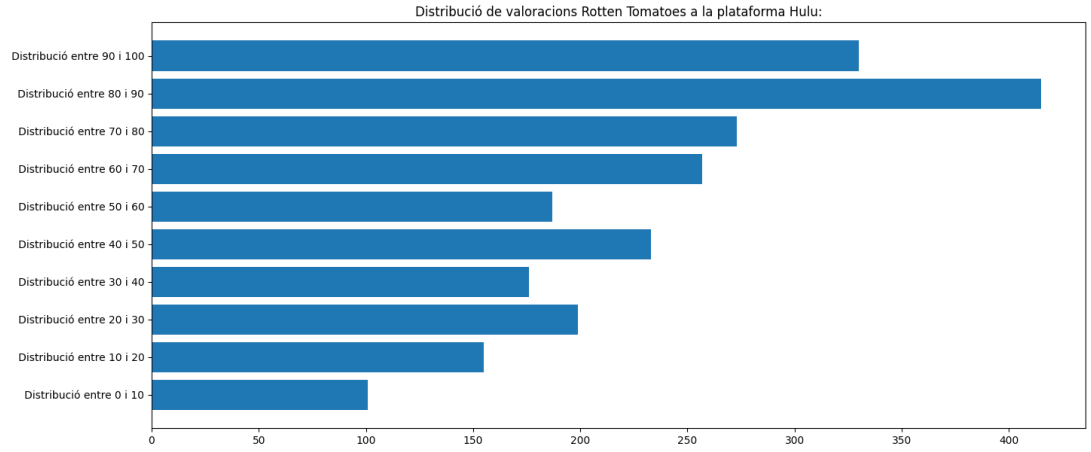
Adjuntem el gràfic de la distribució de valoracions de pel·lícules de Disney Plus segons Rotten Tomatoes i Film TV:

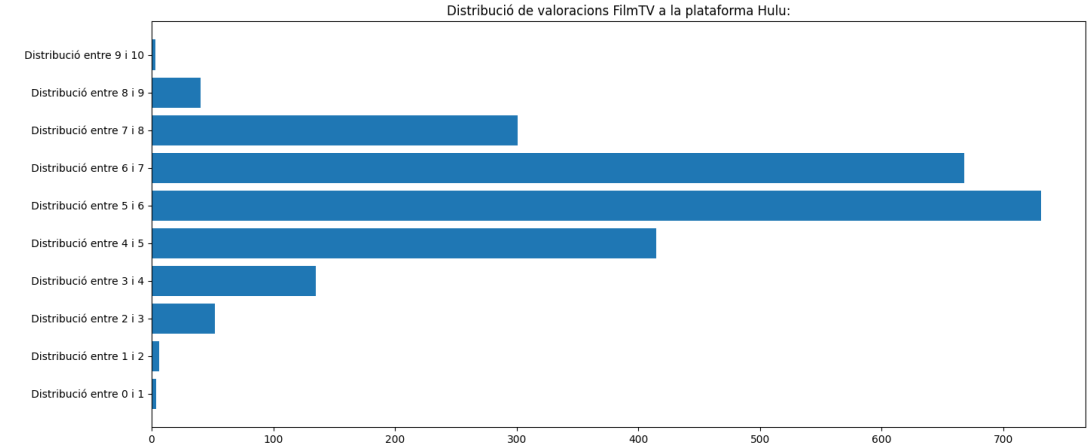




Com podem observar, segons Rotten Tomatoes predominen les valoracions entre 80 i 90 de la plataforma Disney Plus mentre que segons Film TV predominen les valoracions entre 5 i 6. Podem dir que les valoracions entre Rotten Tomatoes i Film TV són molt diferents ja que a Rotten valoren les pel·lícules amb més nota que a Film TV.

Adjuntem el gràfic de la distribució de valoracions de pel·lícules de Hulu segons Rotten Tomatoes i Film TV:



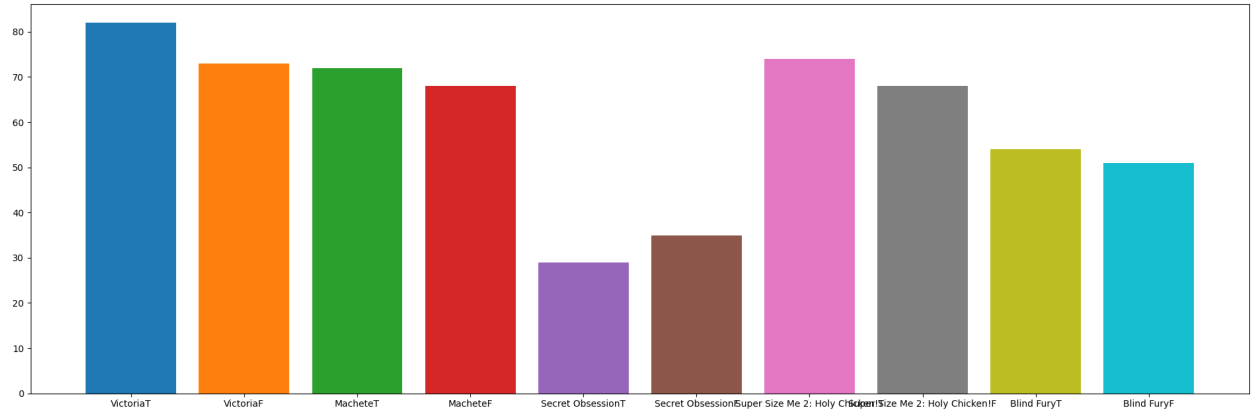


Com podem observar, segons Rotten Tomatoes predominen les valoracions entre 80 i 90 de la plataforma Hulu mentre que segons Film TV predominen les valoracions entre 5 i 6. Podem dir que les valoracions entre Rotten Tomatoes i Film TV són molt diferents ja que a Rotten valoren les pel·lícules amb més nota que a Film TV.

Q16. Per una mateixa pel·lícula, les valoracions a Rotten Tomatoes i a FilmTV són similars? O les valoracions solen ser més altes en alguna d'elles?

Per realitzar aquest anàlisi, el primer pas que realitzem és cercar totes les pel·lícules que es troben en alguna plataforma de streaming i que tenen valoració en Film TV i Rotten Tomatoes.

Un cop tenim les pel·lícules, hem seleccionat 5 i hem realitzat el següent gràfic. Cada pel·lícula té dues barres, l'esquerra és la puntuació a Tomatoes i la dreta la valoració a FilmTv. Observem que els crítics/usuaris de Tomatoes valoren més positivament.



Posteriorment, hem realitzat la mitjana de totes les pel·lícules valorades que estan en alguna plataforma de streaming i observem, que curiosament, tenen millor valoració en FilmTV.

Mitjana de Tomatoes = 58.24 %

Mitjana de FilmTv = 58.72 %

Q17. Les pel·lícules més ben valorades solen ser d'un gènere en concret? Quins són els gèneres pitjors valorats?

Per realitzar aquest anàlisi busquem les pel·lícules millor valorades a Rotten Tomatoes i a FilmTv, per separat. Un cop tenim les 10 millors pel·lícules, cerquem els seus gèneres, busquem el nom del gènere a la seva taula i els apuntem. D'aquesta manera tenim tots els gèneres de les millors pel·lícules.

Els gèneres millor valorats a Rotten Tomatoes són:

➢ *Dramas, International Movies, Action & Adventure, Animation, Action-Adventure, Buddy, Comedy, Action, Fantasy, Family, Classic Movies, Romance, Documentary, Music Videos and Concerts, Comedies, Drama*

Mentre que a FilmTV són:

➢ *Suspense, Adventure, Thrillers, Horror Movies, Cult Movies, Independent Movies, Thrillers, Sci-Fi & Fantasy, Independent Movies, Horror Movies*

Observem que Tomatoes té més varietat.

Per una altra banda, els pitjors gèneres a Rotten Tomatoes són:

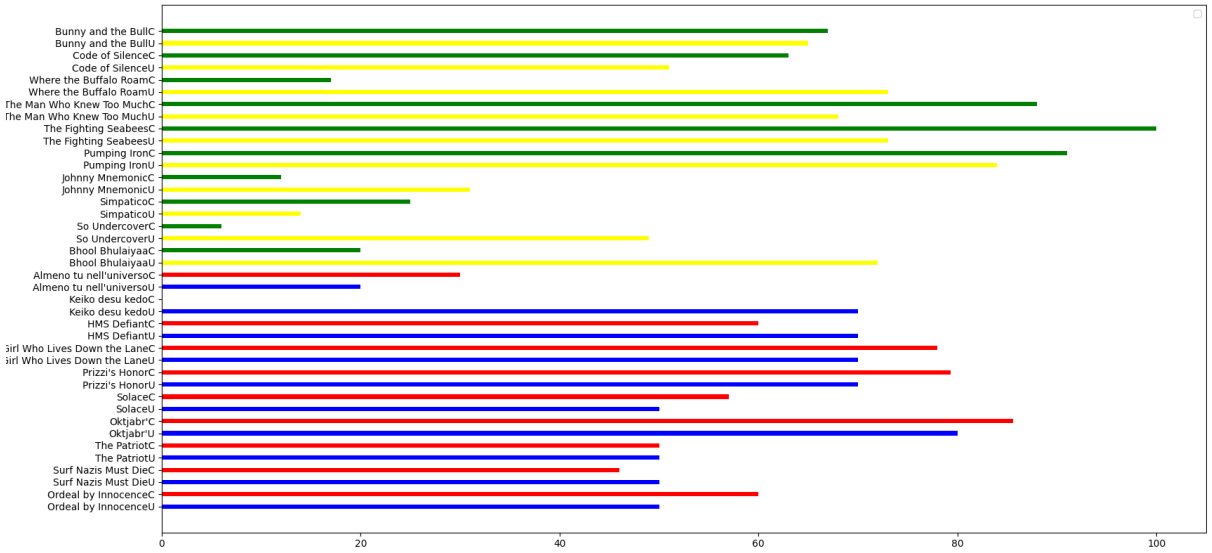
➢ *Action, Suspense, Adventure, Comedies, Thrillers, Horror Movies, Horror, Thriller, Dramas, Romantic Movies, Sci-Fi & Fantasy, Comedy, Fantasy, Romance, International Movies, Action & Adventure, Arthouse, Animation, Family, Action-Adventure, Drama*

I a FilmTV són:

➢ *Children & Family Movies, Independent Movies, Children & Family Movies, Independent Movies, Young Adult Audience, Science Fiction*

Q18. Les valoracions dels crítics i dels usuaris solen ser similars?
En aquest anàlisi, seleccionem 10 pel·lícules de FilmTV i de RottenTomatoes, sense importar quines. D'aquestes pel·lícules seleccionem els vots dels usuaris i dels crítics de cada plataforma respectivament. Finalment les mostrem en el següent gràfic, on els colors de les barres són:

- Verd: crítics Rotten Tomatoes
- Groc: usuaris Rotten Tomatoes
- Vermell: crítics FilmTV
- Blau: usuaris FilmTV



A RottenTomatoes podem observar una gran diferencia entre els crítics i els usuaris. A més, les pel·lícules que més agraden als crítics, menys agraden als usuaris i al revés.

Mentre que a FilmTV hi han vots més equilibrats entre usuaris i crítics, encara que els crítics solen puntuar més positivament.

4. Conclusions i problemes trobats

Entre els problemes que ens han aparegut, anem a comentar els més rellevants:

➢ Hi havia pel·lícules/series en el fitxer CSV que tenien quantitats de camps erronis. Per exemple, una pel·lícula tenía un ";" en el títol i al carregar les dades obteniem camps errònis. En aquests casos, hem ignorat les pel·lícules en qüestió.

➢ Un altre cas al carregar les dades, era trobar camps buits, és a dir, podia faltar la informació sobre el director o altres dades. Per no afegir a la base de dades l'última informació llegida sobre el camp que faltava, hem creat a les taules de Directors/Actors/Països/Gèneres un primer registre amb índex 0 que representa *null*.
Aquesta decisió ens ha sigut útil ja que, al realitzar l'anàlisi exploratori sobre directors, hem detectat que un director havia fer "masses" pel·lícules (el seu Id estava assignat a les següents pel·lícules on no hi havia director) i ho hem pogut solucionar ràpidament.

➢ Al fer l'anàlisi exploratori, hem trobar casos de pel·lícules que la seva durada es mesura en temporades, i no en minuts. Aquests casos els hem ignorat per fer els gràfics, ja que utilitzar temporades per mesurar és poc precís. No sabem de quantes parts és la temporada ni la duració de cada part.

➢ Finalment, un problema que ens hem causat nosaltres mateixos però volem comentar posat que, es un cas greu que s'ha de tractar amb cura, ha sigut no tenir en compte les diverses entrades per cada pel·lícula/serie. Com només pot haver un actor per cada entrada en la base de dades, es pot repetir al recopilar dades quantitatives, però no seria cert. Hem observat que les dades de l'anàlisi exploratori es disparaven i ho hem hagut de repetir.

D'aquesta pràctica hem obtingut diverses conclusions:

➢ Tal com es va comentar a classe de teoria, els temps d'utilització d'una base de dades estratègica són lents i tenim poques dades en comparació amb una base de dades estratègica real.

➢ Hem sigut capaços d'analitzar plataformes de streaming que utilitzem i podriem treure més partit, com per exemple passar-nos a Hulu ja que ens interessen series de Japó. Llavors podriem fer el mateix des del punt de vista de l'empresa per treure més benefici.

➢ A més, hem seguit tot el procés de forma lliure, sense haver de seguir cap guia o patró dictat a classe. Això ens ha ajudat a comprendre millor el procés informàtic de les dades. Ens hem buscat la vida per dissenyar les taules, però abans també per saber quin sistema de gestió de bases de dades utilitzar, quin entorn de desenvolupament de python, etc. Ens ha sigut d'ajuda per obtenir soltesa.

5. Bibliografia

Comprovar si una tupla o cadena estan buits.
https://j2logo.com/comprobar-python-lista-cadena-diccionario-vacio/

Eliminar espais en blanc d'una cadena.
https://j2logo.com/eliminar-espacios-en-blanco-string-strip-python/#:~:text=Funci%C3%B3n%20strip%20en%20Python%20%E2%80%93%20trim,tabuladores%20y%20saltos%20de%20l%C3%ADnea.

Convertir tupla a cadena.
https://foroayuda.es/convertir-cadena-en-tupla-ejemplo-de-codigo-de-python/

Usar variables en una sentencia SQL.
https://stackoverflow.com/questions/902408/how-to-use-variables-in-sql-statement-in-python

Sentencia elif.
https://www.freecodecamp.org/espanol/news/sentencias-if-elif-y-else-en-python/

Verificació registres de la BD.
https://www.lawebdelprogramador.com/foros/SQL/1579046-Como-verificar-si-hay-un-registro-en-la-base-de-datos.html
https://www.lawebdelprogramador.com/foros/SQL/1579046-Como-verificar-si-hay-un-registro-en-la-base-de-datos.html
https://stackoverflow.com/questions/2366854/can-table-columns-with-a-foreign-key-be-null

Lectura d'un fitxer CSV.
https://docs.python.org/es/3/library/csv.html

Comprovar si un element ja existeix en una llista.
https://parzibyte.me/blog/2018/04/17/python-comprobar-elemento-valor-existe-lista-arreglo/