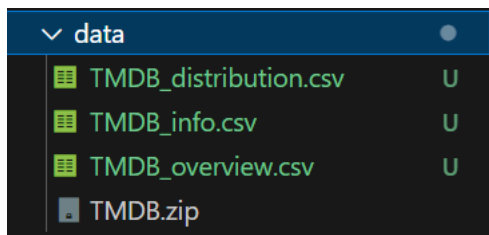


Exercici 1.1.

Implementeu una funció que descomprimeixi fitxers en format zip i tar.gz. La funció rebrà com a inputs la ruta amb el nom del fitxer que es vol descomprimir. La funció detectarà automàticament si el fitxer està comprimit en zip o tar.gz i mostrarà un missatge d'error quan el fitxer sigui d'un altre tipus. Utilitzeu aquesta funció per descomprimir el fitxer TMDb.zip.

Al aplicar la nostra funció de descompressió obtenim els nostres fitxers guardats a la carpeta data:



Exercici 1.2.

Implementeu una funció que llegeixi els csv i els integri en un únic dataframe utilitzant com a clau la columna "id" utilitzant la llibreria **pandas**. Obtingueu el temps de processament.

```
skril349@Toni:/mnt/c/Users/tvive/Documents/UOC/programació_per_la_ciència_de_dades/PAC6/datasci_pac4$ python3 main.py --extract
Extracted data/TMDb.zip in data/
Extraction Time: 3.390594482421875 seconds
skril349@Toni:/mnt/c/Users/tvive/Documents/UOC/programació_per_la_ciència_de_dades/PAC6/datasci_pac4$ python3 main.py --merge_pandas
CSV Merge (Pandas) Time: 1.6722168922424316 seconds
id      name      number_of_seasons  ...      production_companies  origin_country      production_countries
0  1399  Game of Thrones      8  ...  Revolution Sun Studios, Television 360, Genera...  US  United Kingdom, United States of America
1  71446  Money Heist      3  ...  Vancouver Media      ES  Spain
2  66732  Stranger Things      4  ...  21 Laps Entertainment, Monkey Massacre Product...  US  United States of America
3  1482  The Walking Dead      11  ...  AMC Studios, Circle of Confusion, Valhalla Mot...  US  United States of America
4  63174  Lucifer      6  ...  Warner Bros. Television, DC Entertainment, Jer...  US  United States of America
5  69050  Riverdale      7  ...  Warner Bros. Television, Berlanti Productions,...  US  United States of America
6  93485  Squid Game      2  ...  Siren Pictures, Firstman Studio      KR  South Korea
7  1396  Breaking Bad      5  ...  Sony Pictures Television Studios, High Bridge ...  US  United States of America
8  71712  The Good Doctor      6  ...  ABC Studios, 3AD, Sony Pictures Television Stu...  US  United States of America
9  85271  WandaVision      1  ...  Marvel Studios      US  United States of America

[10 rows x 29 columns]
```

Com podem observar, el temps de extracció es de 3.4 segons i el de creació del dataframe es de 1.67 segons.

Exercici 1.3.

Implementeu una funció que llegeixi els csv i els integri en un únic diccionari utilitzant com a clau la columna "id" utilitzant la llibreria **csv**. Obtingueu el temps de processament.

```
skril349@toni:/mnt/c/Users/tvive/Documents/UDC/programació_per_la_ciència_de_dades/PAC6/datasci_pac4$ python3 main.py --merge_dict
CSV Merge (Dictionary) Time: 7.33914510355322 seconds
{'1399': {'id': '1399', 'name': 'Game of Thrones', 'number_of_seasons': '8', 'number_of_episodes': '73', 'original_language': 'en', 'languages': 'en', 'spoken_languages': 'English', 'episode_run_time': '0', 'vote_count': '21857', 'vote_average': '8.442', 'popularity': '1083.917', 'first_air_date': '2011-04-17', 'last_air_date': '2019-05-19', 'adult': 'False', 'in_production': 'False', 'type': 'Scripted', 'status': 'Ended', 'original_name': 'Game of Thrones', 'tagline': 'Winter Is Coming', 'overview': 'Seven noble families fight for control of the mythical land of Westeros. Friction between the houses leads to full-scale war. All while a very ancient evil awakens in the farthest north. Amidst the war, a neglected military order of misfits, the Night's Watch, is all that stands between the realms of men and icy horrors beyond.', 'backdrop_path': '/20tB0ynKlyIenMjWIZdY9lWT4c.jpg', 'homepage': 'http://www.hbo.com/game-of-thrones', 'poster_path': '/1XSioql89opFnbll8WnZY10UjX.jpg', 'genres': 'Sci-Fi & Fantasy, Drama, Action & Adventure', 'created_by': 'David Benioff, D.B. Weiss', 'networks': 'HBO', 'production_companies': 'Revolution Sun Studios, Television 360, Generator Entertainment, Bighead Littlehead', 'origin_country': 'US', 'production_countries': 'United Kingdom, United States of America'}, '71446': {'id': '71446', 'name': 'Money Heist', 'number_of_seasons': '3', 'number_of_episodes': '41', 'original_language': 'es', 'languages': 'es', 'spoken_languages': 'Español', 'episode_run_time': '70', 'vote_count': '17836', 'vote_average': '8.257', 'popularity': '96.354', 'first_air_date': '2017-05-02', 'last_air_date': '2021-12-03', 'adult': 'False', 'in_production': 'False', 'type': 'Scripted', 'status': 'Ended', 'original_name': 'La Casa de Papel', 'tagline': 'The perfect robbery.', 'overview': 'To carry out the biggest heist in history, a mysterious man called The Professor recruits a band of eight robbers who have a single characteristic: none of them has anything to lose. Five months of seclusion - memorizing every step, every detail, every probability - culminate in eleven days locked up in the National Coinage and Stamp Factory of Spain, surrounded by police forces and with dozens of hostages in their power, to find out whether their suicide wager will lead to everything or nothing.', 'backdrop_path': '/gFZr1CkpJYsApPZEF3jhxL4yLzG.jpg', 'homepage': 'https://www.netflix.com/title/80192098', 'poster_path': '/reMOA1uizscCbKpeRiET7b2jUp.jpg', 'genres': 'Crime, Drama', 'created_by': 'Álex Pina', 'networks': 'Netflix, Antena 3', 'production_companies': 'Vancouver Media', 'origin_country': 'ES', 'production_countries': 'Spain'}, '66732': {'id': '66732', 'name': 'Stranger Things', 'number_of_seasons': '4', 'number_of_episodes': '34', 'original_language': 'en', 'languages': 'en', 'spoken_languages': 'English', 'episode_run_time': '0', 'vote_count': '16161', 'vote_average': '8.624', 'popularity': '185.711', 'first_air_date': '2016-07-15', 'last_air_date': '2022-07-01', 'adult': 'False', 'in_production': 'True', 'type': 'Scripted', 'status': 'Returning Series', 'original_name': 'Stranger Things', 'tagline': 'Every ending has a beginning.', 'overview': 'When a young boy vanishes, a small town uncovers a mystery involving secret experiments, terrifying supernatural forces, and one strange little girl.', 'backdrop_path': '/2fUumg8llhK0po22O3846v70S.jpg', 'homepage': 'https://www.netflix.com/title/80057281', 'poster_path': '/49u0Fv8v0moxb9IPfGn8l4tqKsX0.jpg', 'genres': 'Drama, Sci-Fi & Fantasy, Mystery', 'created_by': 'Matt Duffer, Ross Duffer', 'networks': 'Netflix', 'production_companies': '21 Laps Entertainment, Monkey Massacre Productions', 'origin_country': 'US', 'production_countries': 'United States of America'}, '1482': {'id': '1482', 'name': 'The Walking Dead', 'number_of_seasons': '11', 'number_of_episodes': '177', 'original_language': 'en', 'languages': 'en', 'spoken_languages': 'English', 'episode_run_time': '42', 'vote_count': '15432', 'vote_average': '8.121', 'popularity': '489.746', 'first_air_date': '2010-10-31', 'last_air_date': '2022-11-20', 'adult': 'False', 'in_production': 'False', 'type': 'Scripted', 'status': 'Ended', 'original_name': 'The Walking Dead', 'tagline': 'Fight the dead. Fear the living.', 'overview': 'Sheriff's deputy Rick Grimes awakens from a coma to find a post-apocalyptic world dominated by flesh-eating zombies. He sets out to find his family and encounters many other survivors along the way.', 'backdrop_path': '/x4salpj81lumLU0ltfNvSSrjSxm.jpg', 'homepage': 'http://www.amc.com/shows/the-walking-dead-1002293', 'poster_path': '/n7PVu0hS2zAsVekpO1oCnkK4bn.jpg', 'genres': 'Action & Adventure, Drama, Sci-Fi & Fantasy', 'created_by': 'Frank Darabont', 'networks': 'AMC', 'production_companies': 'AMC Studios, Circle of Confusion, Valhalla Motion Pictures, Darkwoods Productions, Skybound Entertainment, Idiotbox', 'origin_country': 'US', 'production_countries': 'United States of America'}, '63174': {'id': '63174', 'name': 'Lucifer', 'number_of_seasons': '6', 'number_of_episodes': '93', 'original_language': 'en', 'languages': 'en', 'spoken_languages': 'English', 'episode_run_time': '45', 'vote_count': '13870', 'vote_average': '8.486', 'popularity': '416.668', 'first_air_date': '2016-01-25', 'last_air_date': '2021-09-10', 'adult': 'False', 'in_production': 'False', 'type': 'Scripted', 'status': 'Ended', 'original_name': 'Lucifer', 'tagline': 'It's good to be bad.', 'overview': 'Bored and unhappy as the Lord of Hell, Lucifer Morningstar abandoned his throne and retired to Los Angeles, where he has teamed up with LAPD detective Chloe Decker to take down criminals. As he returns home, he realizes the longer he's away from the underworld, the greater the threat that the worst of humanity could escape.', 'backdrop_path': '/a08Rtunw49Uf4XmqfYNU09nlyIu.jpg', 'homepage': 'https://www.netflix.com/title/80057918', 'poster_path': '/ekZobS8isE6mAS3RAIGD093h8XL.jpg', 'genres': 'Crime, Sci-Fi & Fantasy', 'created_by': 'Tom Kapinos', 'networks': 'FOX, Netflix', 'production_companies': 'Warner Bros. Television, DC Entertainment, Jerry Bruckheimer Television, DC Vertigo', 'origin_country': 'US', 'production_countries': 'United States of America'}}
```

El temps de creació d'aquest diccionari es de 7.3 sgons.

Exercici 1.4.

Quines diferències s'observen en la lectura dels fitxers seguint tots dos mètodes? Si els fitxers tinguessin una mida de 10GB quin mètode seria més ràpid? Justifiqueu la resposta.

Comparació entre el merge amb Pandas i en un Diccionari:

- **Pandas vs. Diccionari:** El merge amb Pandas ha resultat ser més ràpida que el merge en un diccionari. Això és degut a l'optimització interna de Pandas per a la manipulació de grans conjunts de dades.

- **Eficiència amb Grans Fitxers:** Per a fitxers de gran mida (com 10GB), Pandas probablement seria més eficient si la memòria ho permet. Si la memòria és una preocupació, la lectura en parts (chunking) amb Pandas o l'aproximació basada en diccionaris pot ser una millor solució.

- **Conclusió:** Per a grans volums de dades i amb suficient memòria, Pandas és l'opció preferida. Per a situacions amb restriccions de memòria, el mètode basat en diccionaris pot ser més adequat.

Exercici 2.1.

Afegiu una variable `air_days` al dataframe que consisteixi en el nombre de dies que una sèrie ha estat en emissió. Mostreu per pantalla els 10 registres del dataset que més dies han estat en emissió.

```
● skril349@Toni:/mnt/c/Users/tvive/Documents/UOC/programació_per_la_ciència_de_dades/PAC6/datasci_pac4$ python3 main.py --air_days
CSV Merge (Pandas) Time: 1.1727027893066406 seconds
Top 10 series by air days:
      name    air_days
14970 CBS Evening News 30043.0
41903 Neujahrskonzert der Wiener Philharmoniker 29950.0
9826 Golden Globe Awards 28845.0
79605 BBC Proms 27762.0
12527 Meet the Press 27555.0
38826 Macy's Thanksgiving Day Parade 27027.0
32137 The BAFTA Awards 26929.0
18283 The Emmy Awards 26893.0
21388 ABC World News 26805.0
11145 Sanremo Music Festival 26311.0
```

Com es pot veure en la imatge, el programa de televisió amb més dies d'emissió amb un total de 30043 dies, es la de CBS Evening News.

Exercici 2.2.

Creeu un diccionari ordenat la clau del qual serà el nom de la sèrie (`name`) i el valor del qual serà l'adreça web completa del vostre pòster (`homepage` i `poster_path`). En cas que `homepage` o `poster_path` tinguin el valor `NaN` o `""`, el valor serà el string "NOT AVAILABLE". Mostreu per pantalla els primers 5 registres del diccionari.

```
● skril349@Toni:/mnt/c/Users/tvive/Documents/UOC/programació_per_la_ciència_de_dades/PAC6/datasci_pac4$ python3 main.py --series_dict
CSV Merge (Pandas) Time: 1.6448850631713867 seconds
First 5 entries in the series poster dictionary:
{'Game of Thrones': 'http://www.hbo.com/game-of-thrones/1XG1oqL89opfnblBbnZY10IuJx.jpg', 'Money Heist': 'https://www.netflix.com/title/80192098/reEHAJuzscCbkerJeIT2bjqUp.jpg', 'Stranger Things': 'https://www.netflix.com/title/80057281/49u0Fa0h0x0b9IPf0n8A1qKs1d.jpg', 'The Walking Dead': 'http://www.amc.com/shows/the-walking-dead--1002293/n7Pvu0hS22sAsVekp0LoCnkU1bn.jpg', 'Lucifer': 'https://www.netflix.com/title/80087918/ekZobS8isE6eA53RAIG0093h8dL.jpg'}
```

En la imatge podem observar el diccionari amb els 5 primers registres. Podem observar que s'ha creat correctament el diccionari posant de key el nom de la sèrie, i el pòster correcte.

Exercici 3.1.

Obtingueu i mostreu per pantalla els noms de les sèries l'idioma original (`original_language`) de les quals sigui l'anglès i en el resum de les quals (`overview`) apareguin les paraules "mystery" o "crime", sense tenir en compte majúscules ni minúscules.

```
● skril349@Toni:/mnt/c/Users/tvive/Documents/UOC/programació_per_la_ciència_de_dades/PAC6/datasci_pac4$ python3 main.py --english_series
CSV Merge (Pandas) Time: 2.1676931381225586 seconds
English series with 'mystery' or 'crime' in overview:
      name
2  Stranger Things
7  Breaking Bad
17 The Umbrella Academy
26 Wednesday
43 The Act
```

Exercici 3.2.

Obtingueu una llista de les sèries que han començat el 2023 i han estat cancel·lades. Mostreu per pantalla els primers 20 elements d'aquesta llista.

```
● skril349@Toni:/mnt/c/Users/tvive/Documents/UOC/programació_per_la_ciència_de_dades/PAC6/datasci_pac4$ python3 main.py --startandcancelled_series
CSV Merge (Pandas) Time: 1.4872689247131348 seconds
Series started in 2023 and canceled:\n
name
1983 Lockwood & Co.
2206 The Idol
3003 Gotham Knights
4443 True Lies
4601 Sky High: The Series
5396 High Desert
5876 Grease: Rise of the Pink Ladies
5880 The Watchful Eye
5908 The Company You Keep
6227 Dear Edward
6734 City on Fire
7362 The Head of Joaquin Murrieta
8488 Freeridge
9002 Up Here
13253 A Town Called Malice
17265 Slip
19918 The Low Tone Club
20460 Monster Factory
20852 @Gina Yei: #WithAllMyHeartAndMore
21691 Bling Empire: New York
○ skril349@Toni:/mnt/c/Users/tvive/Documents/UOC/programació_per_la_ciència_de_dades/PAC6/datasci_pac4$
```

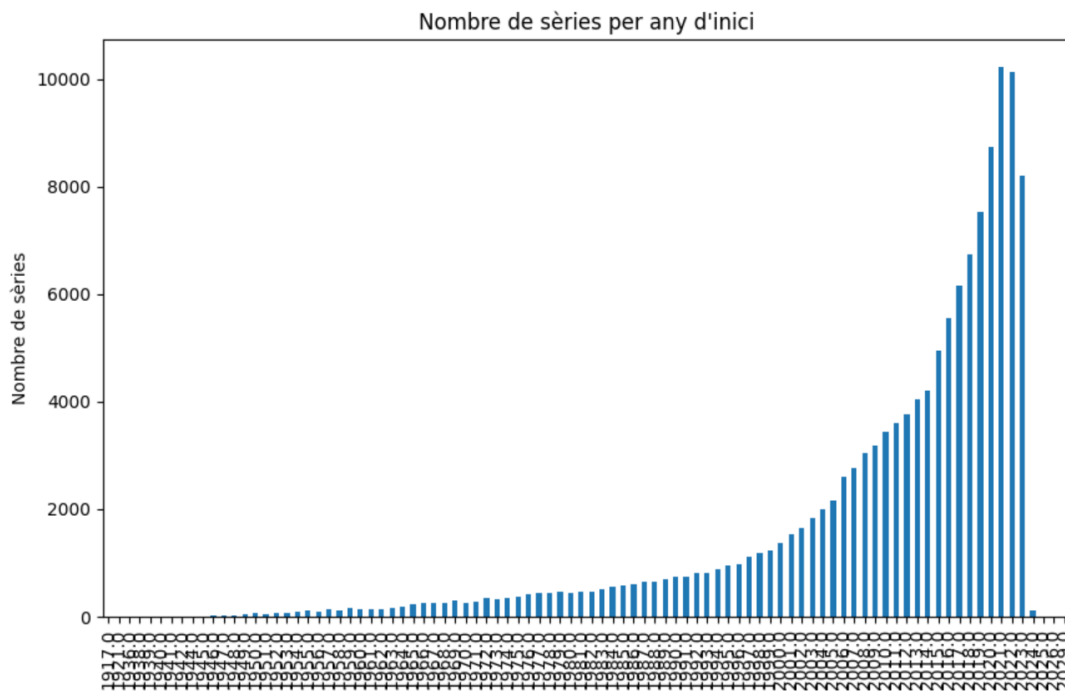
Exercici 3.3.

Obtingueu un dataframe amb els noms, els noms originals, les plataformes d'emissió i les empreses productores de totes les sèries l'idioma (variable languages) de les quals sigui el japonès i mostrar els primers 20 registres per pantalla. Nota: tingueu en compte que considerem sèries en japonès també aquelles que tinguin idiomes addicionals, per exemple, un registre amb idioma “en, ja, ko” s'inclouria.

```
● skril349@Toni:/mnt/c/Users/tvive/Documents/UOC/programació_per_la_ciència_de_dades/PAC6/datasci_pac4$ python3 main.py --japanese_series
CSV Merge (Pandas) Time: 1.3106982707977295 seconds
Japanese series:\n
name original_name networks production_c
24 Naruto Shippūden ナルト 疾風伝 TV Tokyo TV Tokyo, Pierrot, Sound Box
34 Demon Slayer: Kimetsu no Yaiba 鬼滅の刃 Fuji TV, Gunma TV, Tokyo MX, BS11, Tokai Telev... ufotable, Aniplex, Shueisha
36 Attack on Titan 進撃の巨人 MBS, NHK G, Tokyo MX Production I.G, MAPPA, WIT STUDIO, Pony Canyon...
39 Naruto ナルト TV Tokyo Pierrot, Sound Box
45 The Seven Deadly Sins 七つの大罪 tv asahi, MBS, TV Tokyo, TBS, CBC, TV Aichi, T... A-1 Pictures, Studio Deen
46 Dragon Ball Super ドラゴンボール超 (スーパー) Fuji TV Toei Company, Toei Animation, Fuji Television
...
49 My Hero Academia 僕のヒーローアカデミア Nippon TV, MBS, TBS, YTV BONES, Shueisha, movie, dentsu, Yomiuri Teleca...
58 Dragon Ball Z ドラゴンボールゼット Fuji TV Fuji Television Network, Mini Art, Toei Animat...
60 One Piece ワンピース Fuji TV Toei Animation, Fuji Television Network, Avex ...
76 Death Note DEATH NOTE Nippon TV Madhouse
89 One-Punch Man ワンパンマン TV Tokyo, TV Aichi, TV Osaka Madhouse, J.C.STAFF, Bandai Visual, Asatsu-DK,...
99 Super Dragon Ball Heroes スーパードラゴンボールヒーローズ Fuji TV Toei Animation, Fuji Television Ne
...
108 Jujutsu Kaisen 呪術廻戦 MBS, TBS, CBC, Tulip Television, SBC, BSN, tys... MAPPA, Sumzap, dugout, Shueisha, MBS, TOHO
109 Dragon Ball ドラゴンボール Fuji TV Toei Animation, Cloverway, Inc.
134 Boruto: Naruto Next Generations BORUTO-ボルト- NARUTO NEXT GENERATIONS TV Tokyo Pierrot
158 Tokyo Ghoul 東京喰種トーキョーグール Tokyo MX Pierrot, Marvelous, TC Entertainmen
...
166 Heroes Heroes NBC Tailwind Productions
182 InuYasha 犬夜叉 ANIMAX, Nippon TV, YTV SUNRISE
191 Dragon Ball GT ドラゴンボールGT Fuji TV Bird Studios, Toei Animation, Toei Company, Fi...
196 Fullmetal Alchemist: Brotherhood 鋼の錬金術師 FULLMETAL ALCHEMIST MBS, TBS, CBC, SBS TV BONES, Aniplex, SQUARE ENIX, MBS, Techno Sound
○ skril349@Toni:/mnt/c/Users/tvive/Documents/UOC/programació_per_la_ciència_de_dades/PAC6/datasci_pac4$
```

Exercici 4.1.

Mostreu en un gràfic de barres el nombre de sèries per any d'inici.



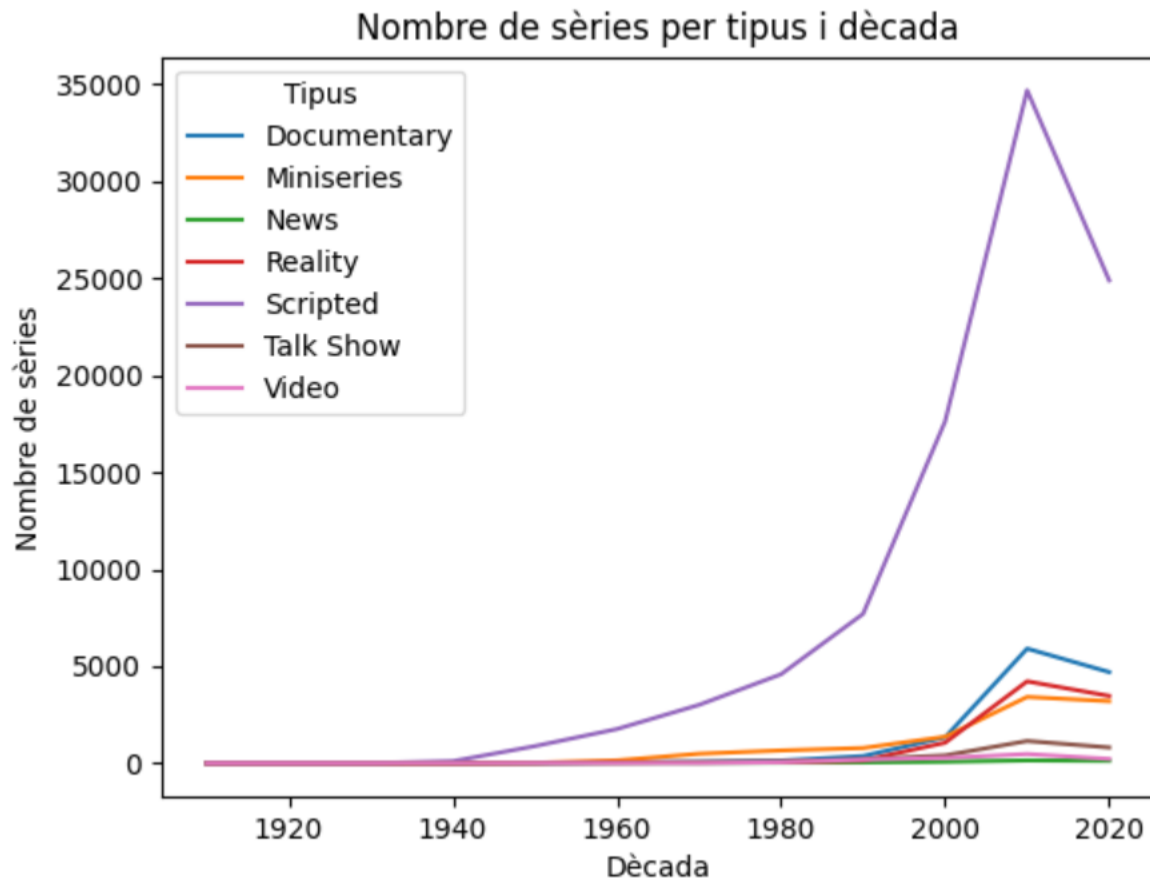
Podem observar en el gràfic com es tracta d'un augment exponencial de la producció de series, obtenint el pic màxim l'any 2021.

Posteriormente, però, comença una decrement d'aquestes.

Podria estar degut a la crisi mundial de la COVID-19, doncs possiblement les productores no tenien tants diners per gastar en entreteniment.

Exercici 4.2.

Construïu un gràfic de línies que mostri el nombre de sèries de cada categoria de la variable “type” produïdes a cada dècada des de 1940. Quins canvis de tendència s'observen?

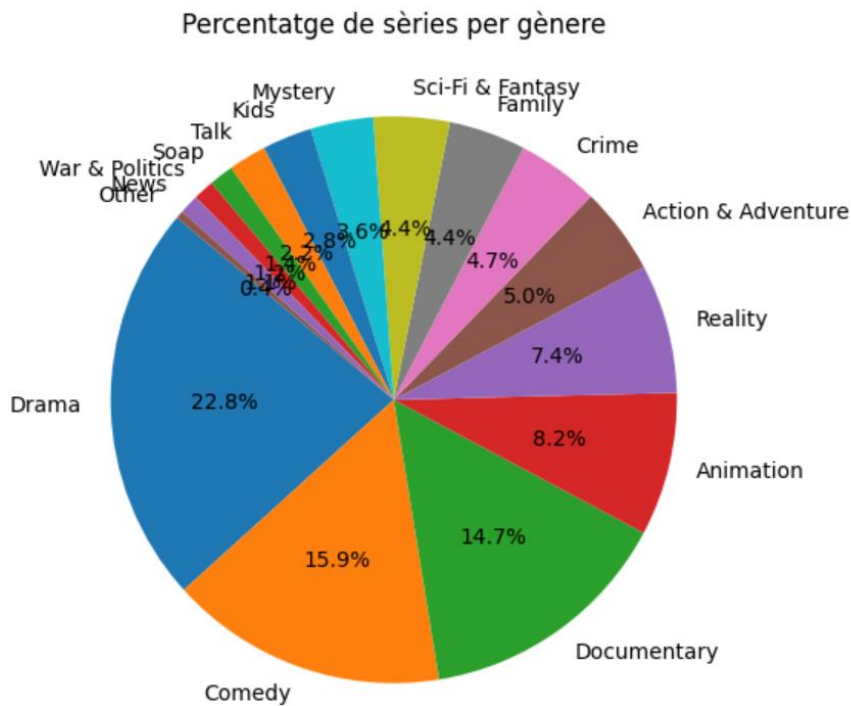


Podem observar com en general totes menys les NEWS, augmenten a partir dels 2000 d'una forma considerable, però la categoria Scripted, guanya de golejada, doncs comença a la dècada del 1940 i augmenta exponencialment treient molt marge a la resta.

El que podem observar també, es que els documentals guanyen força just en la época del COVID-19, fet que podria recolzar el que comentavem anteriorment de que les series perdien força, i guanyava audiència la part de documentals sobre el tema.

Exercici 4.3.

Obtingueu el nombre de sèries per gènere i mostreu el percentatge respecte al total en un gràfic circular. Els gèneres que representin menys de l'1% del total s'inclouran a la categoria "Other". Tingueu en compte que una sèrie que tingui més d'un gènere s'haurà d'incloure a totes les categories en què estigui classificada i que les sèries amb el camp "genres" buit no s'inclouen.



Podem observar que el que predomina es el drama, la comèdia i els documentals.