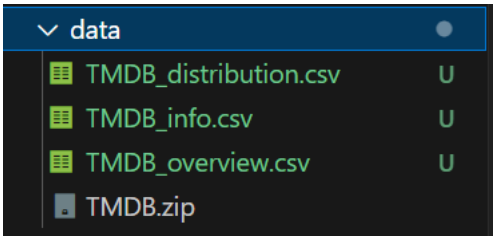
Exercici 1.1.

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 dun 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 **pandes**. Obtingueu el temps de processament.

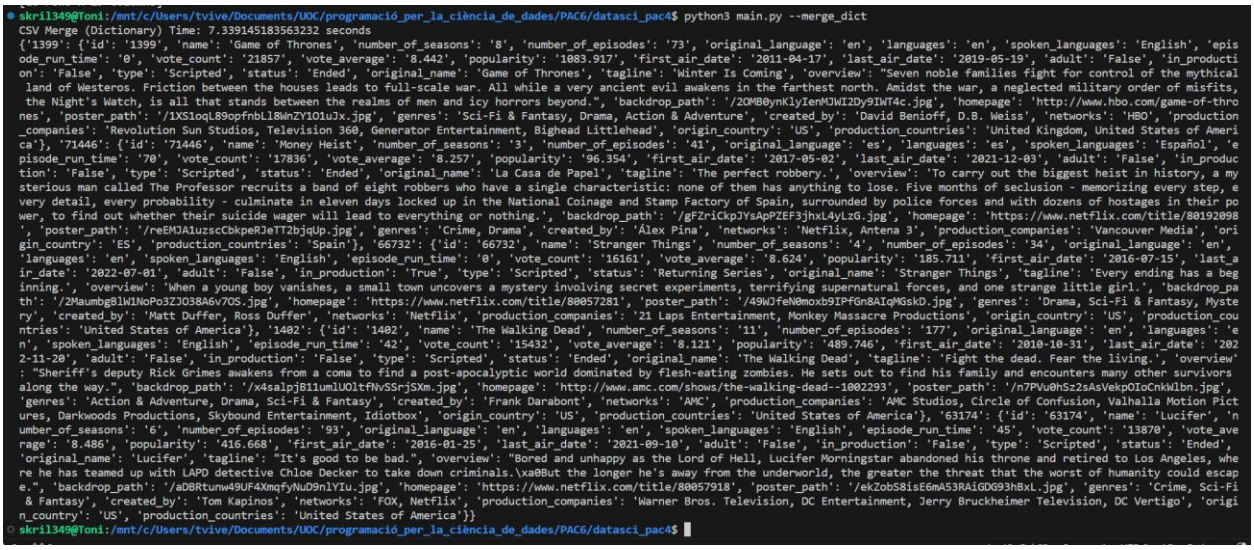
```
skril349@Toni:/mnt/c/Users/tvive/Documents/UOC/programació_per_la_ciència_de_dades/PAC6/datasci_pac4$ python3 main.py --extract
Extracted data/TMDB.zip in data/
Extraction Time: 3.390594482421875 seconds
skril349@Toni:/mnt/c/Users/tvive/Documents/UOC/programació_per_la_ciència_de_dades/PAC6/datasci_pac4$ 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   1402  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  93405        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.



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:

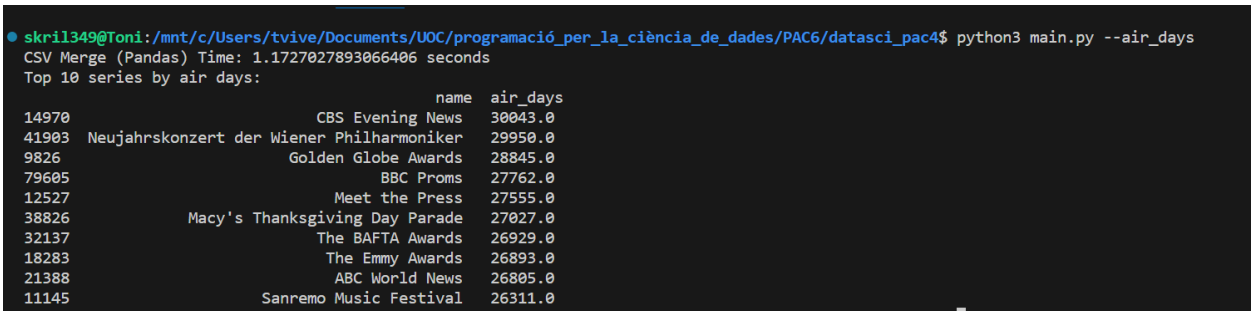
- **Pandas vs. Diccionari**: El merge amb Pandas ha resultat ser més ràpida que el merge en un diccionari. Això és degut a l'optimització interna de Pandas per a la manipulació de grans conjunts de dades.

- **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.

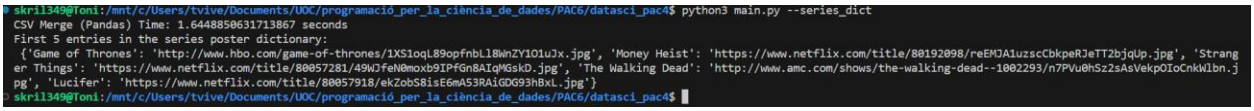
Afegiu una variable air_days al dataframe que consisteixi en el nombre de dies que una sèrie ha estat en emissió. Mostreu per pantalla els 10 registres del dataset que més dies han estat en emissió.



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.

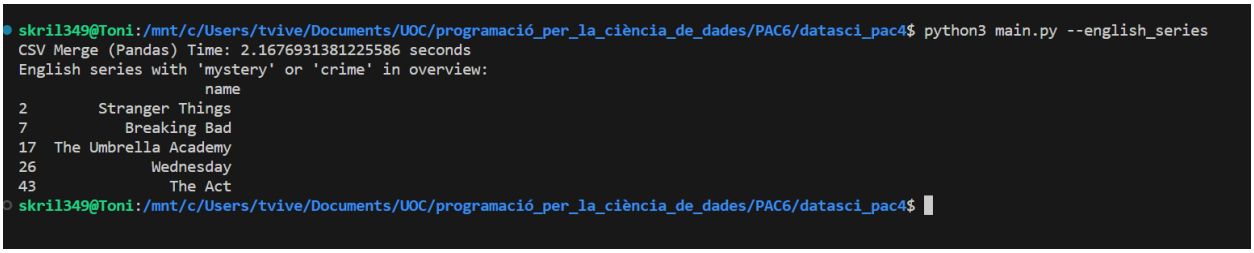
Creeu un diccionari ordenat la clau del qual serà el nom de la sèrie (name) i el valor del qual serà l'adreça web completa del vostre pòster (homepage i poster_path). En cas que homepage o poster_path tinguin el valor NaN o "", el valor serà el string "NOT AVAILABLE". Mostreu per pantalla els primers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



En la imatge podem observar el diccionari amb els 5 primers registres. Podem observar que s'ha creat correctamente el diccionari posant de key el nom de la serie, 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.



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 Joaquín 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$
```

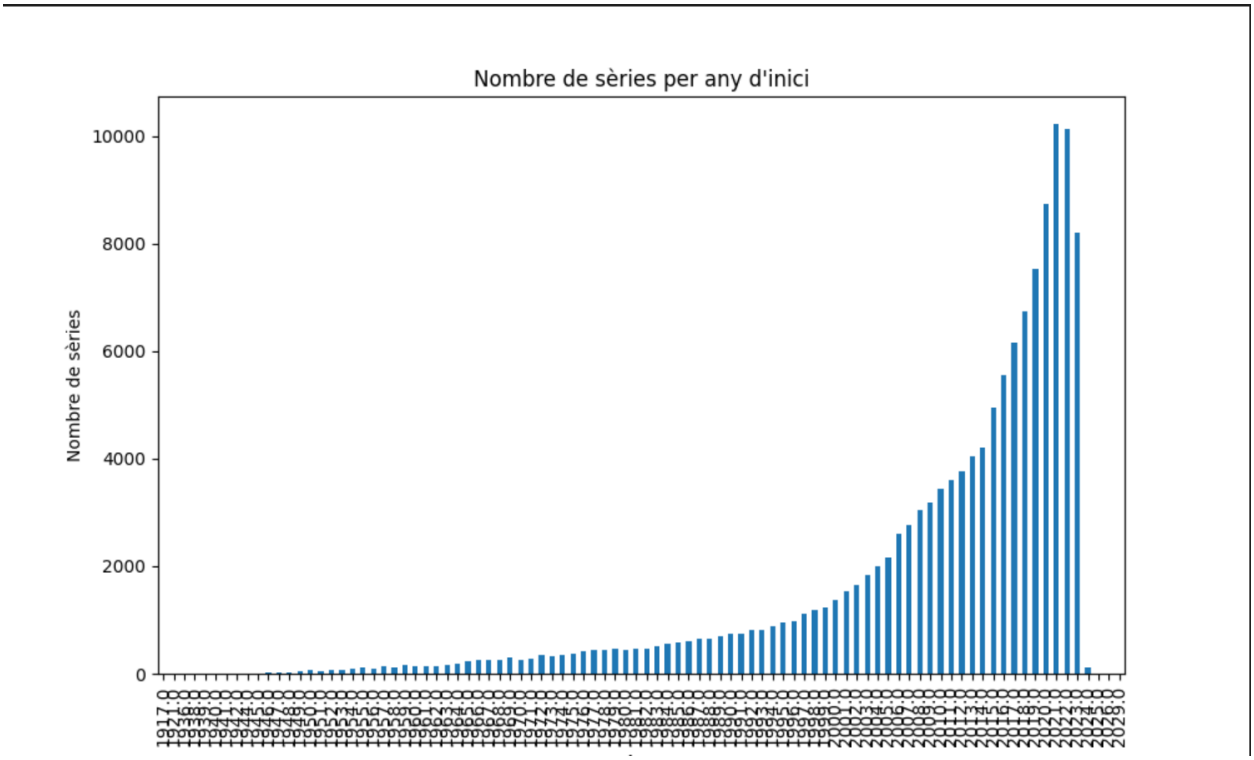
Exercici 3.3.

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
ompanies
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, movic, 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
twork
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
t
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.



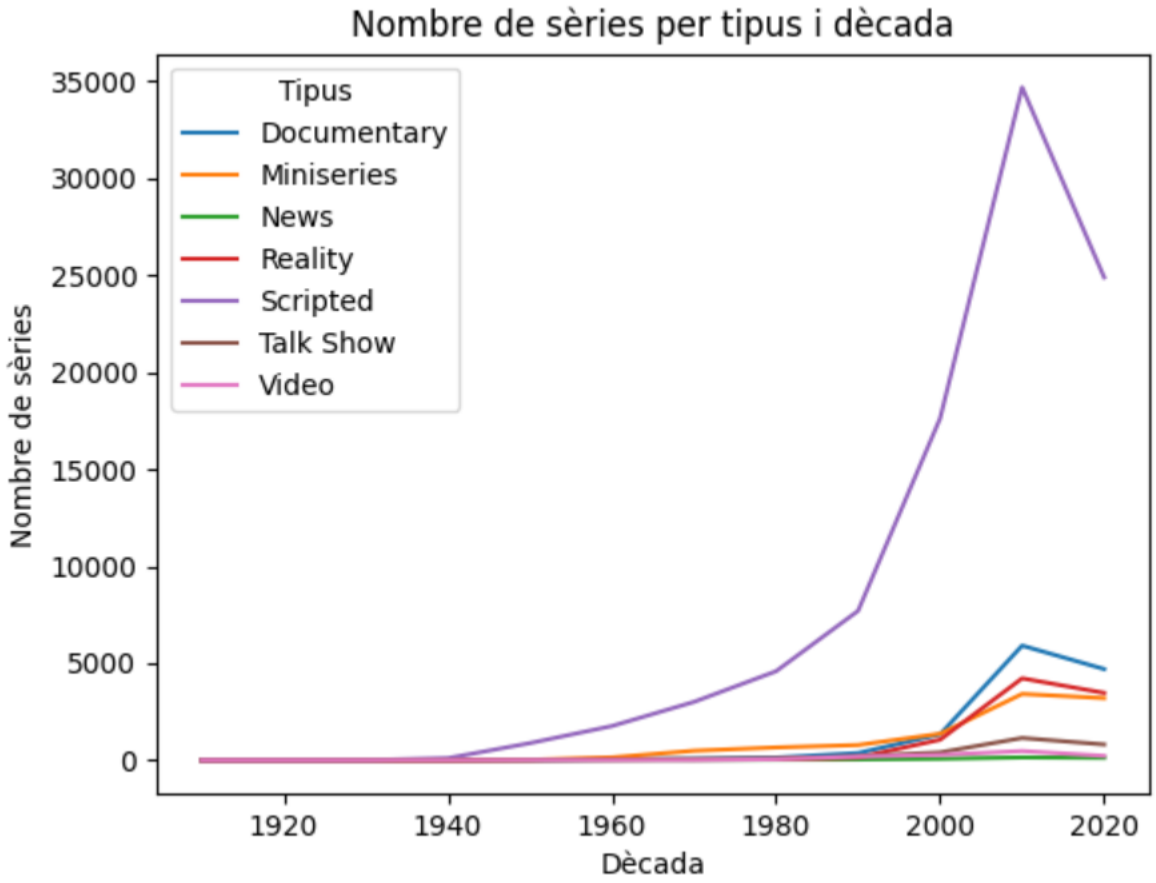
Podem observar en el gràfic com es tracta d'un augment exponencial de la producción de series, obtenint el pic màxim l'any 2021.

Posteriormente, però, comença una decrement d'aquestes.

Podría 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?

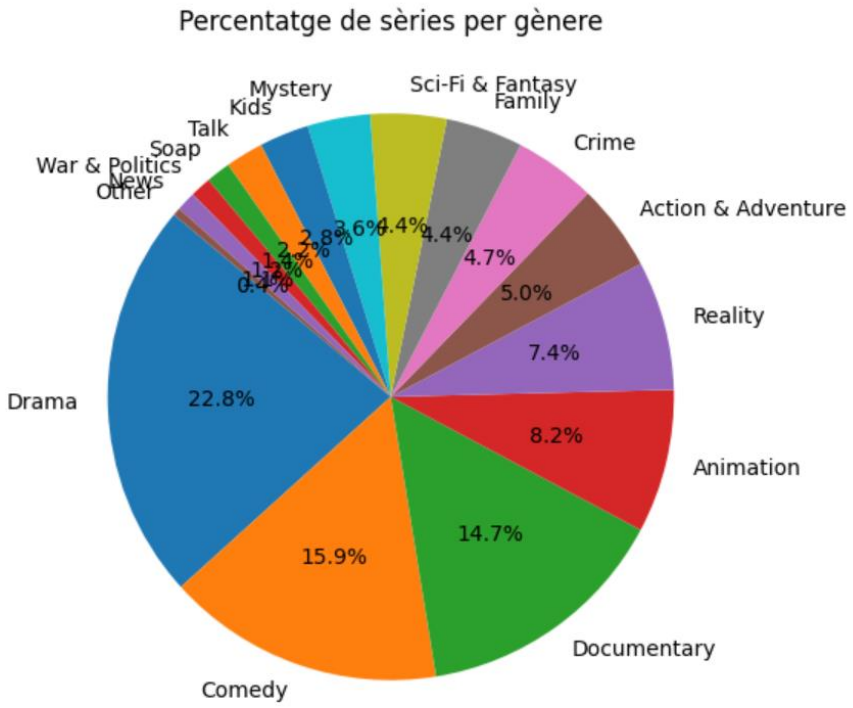


Podem observar com en general totes menys les NEWS, augmenten a partir dels 2000 d’una forma considerable, però la categoría Scripted, guanya de golejada, doncs comença a la década del 1940 i augmenta exponencialment treient molt marge a la resta.

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