Netflix_cleaned_dataset

```
# prompt: upload csv file

from google.colab import files

uploaded = files.upload()

for fn in uploaded.keys():
  print('User uploaded file "{name}" with length {length} bytes'.format(
      name=fn, length=len(uploaded[fn])))
```

```
Choose files Netflix_movi…ustering.csv
```

- **Netflix_movies_and_tv_shows_clustering.csv**(text/csv) - 3000491 bytes, last modified: 25/01/2023 - 100% done

```
Saving Netflix_movies_and_tv_shows_clustering.csv to Netflix_movies_and_tv_shows_clustering (1).csv
User uploaded file "Netflix_movies_and_tv_shows_clustering (1).csv" with length 3000491 bytes
```

```
import re
```

```
df.duplicated().sum()
df.drop_duplicates(inplace=True)
```

```
df['type'] = df['type'].str.strip().str.title()
df['country'] = df['country'].str.strip().str.title()
df['rating'] = df['rating'].fillna('').str.upper().str.strip()
for col in ['director', 'cast']:
    df[col] = df[col].fillna('').apply(lambda x: ', '.join([name.strip().title() for name in x.split(',')]))
df['country'] = df['country'].fillna('').str.strip().str.title()
df['country'] = df['country'].replace({
    'Usa': 'United States',
    'United States Of America': 'United States',
    'Uk': 'United Kingdom',
    'South Korea': 'Korea, Republic Of',
    'Russia': 'Russian Federation'
})
```

```
df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')
```

```
df['duration'] = df['duration'].fillna('')
df['duration_int'] = df['duration'].apply(lambda x: int(re.findall(r'\d+', x)[0]) if re.findall(r'\d+', x) else np.nan)
df['duration_type'] = df['duration'].apply(lambda x: re.findall(r'[a-zA-Z]+', x)[0].lower() if re.findall(r'[a-zA-Z]+',
df['duration_type'] = df['duration_type'].replace({'minutes': 'min', 'minute': 'min', 'min': 'min',
                                                   'seasons': 'season', 'season': 'season'})
```

```
df.columns = df.columns.str.lower().str.replace(' ', '_')
```

```
df['listed_in'] = df['listed_in'].fillna('').apply(lambda x: ', '.join(sorted(set(map(str.strip, x.split(','))))) if x e
```

```
print(df.head())
```

```
  show_id     type  title           director  \
0      s1  Tv Show     3%                     
1      s2    Movie   7:19  Jorge Michel Grau  
2      s3    Movie  23:59       Gilbert Chan  
3      s4    Movie      9        Shane Acker  
4      s5    Movie     21     Robert Luketic  

                                                cast        country  \
0  João Miguel, Bianca Comparato, Michel Gomes, R...         Brazil  
1  Demián Bichir, Héctor Bonilla, Oscar Serrano, ...         Mexico  
2  Tedd Chan, Stella Chung, Henley Hii, Lawrence ...      Singapore  
3  Elijah Wood, John C. Reilly, Jennifer Connelly...  United States  
4  Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar...  United States  
```

```
  date_added  release_year rating   duration  \
0 2020-08-14          2020  TV-MA  4 Seasons
1 2016-12-23          2016  TV-MA     93 min
2 2018-12-20          2011      R     78 min
3 2017-11-16          2009  PG-13     80 min
4 2020-01-01          2008  PG-13    123 min

                                        listed_in  \
0  International TV Shows, TV Dramas, TV Sci-Fi &...
1                      Dramas, International Movies
2                Horror Movies, International Movies
3  Action & Adventure, Independent Movies, Sci-Fi...
4                                          Dramas

                                     description  duration_int  \
0  In a future where the elite inhabit an island ...             4
1  After a devastating earthquake hits Mexico Cit...            93
2  When an army recruit is found dead, his fellow...            78
3  In a postapocalyptic world, rag-doll robots hi...            80
4  A brilliant group of students become card-coun...           123

  duration_type
0        season
1           min
2           min
3           min
4           min
```

Summary of Data Cleaning Steps:

1. 🧹 Standardized column names:
   - Lowercased, removed spaces, and replaced them with underscores for consistency.
2. 📺 Normalized 'type' values:
   - Ensured all entries like 'Movie' and 'TV Show' are in title case format.
3. 🎬 Cleaned 'director' and 'cast':
   - Filled missing values with empty strings.
   - Standardized names with proper casing and removed extra spaces.
4. 🌍 Standardized 'country' entries:
   - Trimmed and title-cased country names.
   - Replaced common inconsistencies (e.g., 'USA' → 'United States').
5. 🔠 Unified 'rating' values:
   - Removed leading/trailing spaces and converted all to uppercase.
6. ⏱️ Separated 'duration' into two new fields:
   - Extracted numbers as `duration_int` and duration type as `duration_type`.
   - Standardized duration types (e.g., 'Minutes' → 'min', 'Seasons' → 'season').
7. 🗂️ Cleaned and sorted 'listed_in' genres:
   - Removed duplicates, standardized spacing, and sorted genre names alphabetically.
8. 💾 Saved the final cleaned dataset as 'Netflix_cleaned_dataset.csv'.

Thank you!