Netflix dataset EDA

Angelos Nikolas

The dataset is available at:
https://www.kaggle.com/datasets/shivamb/netflix-shows

Importing Libraries

In [ ]:
```
import pandas as pd
import numpy as np
import plotly.express as px
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

Reading the dataset and changing the data format

In [ ]:
```
#Reading the data from the csv file and changing the date format
dataset = pd.read_csv('netflix_titles.csv', date_parser=["date_added"], infer_dateti
dataset["date_added"] = pd.to_datetime(dataset["date_added"].str.strip(), format="%B
dataset
```

Out[ ]:

	show_id	type	title	director	cast	country	date_added	release_year	rating
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	2021-09-24	2021	TV-MA

	show_id	type	title	director	cast	country	date_added	release_year	rating
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	2021-09-24	2021	TV-MA
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA
...	...	...	...	...	...	...	...	...	...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	2019-11-20	2007	R
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	2019-07-01	2018	TV-Y7
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	2019-11-01	2009	R
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	2020-01-11	2006	PG
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	2019-03-02	2015	TV-14

8807 rows × 12 columns

Basic data exploration with pandas

Checking the data types of each column

In [ ]:
```
dataset.dtypes
```

```
show_id         object
```

Out[ ]:
```
type                    object
title                   object
director                object
cast                    object
country                 object
date_added      datetime64[ns]
release_year             int64
rating                  object
duration                object
listed_in               object
description             object
dtype: object
```

Summarizing the dataset

In [ ]:
```
# Summarize the dataset (taking the numerical columns)
dataset.describe()
```

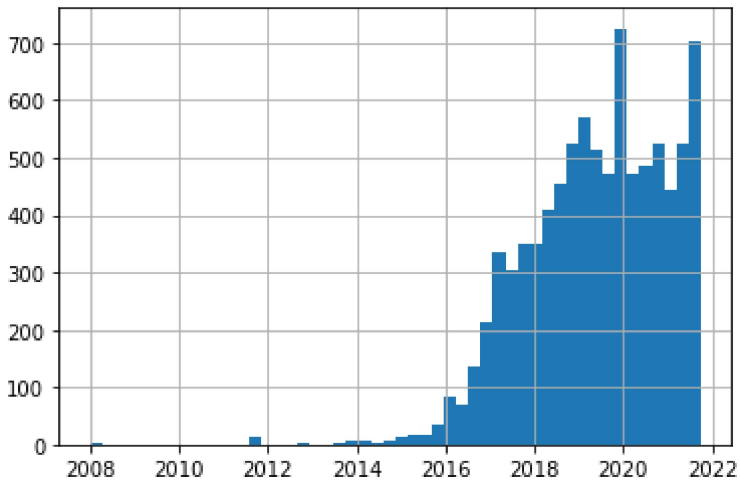
Out[ ]:

	release_year
count	8807.000000
mean	2014.180198
std	8.819312
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2019.000000
max	2021.000000

Date analysis

In [ ]:
```
 # Create a histogram of the date_added column (matplotlib)
dataset["date_added"].hist(bins=50)
```

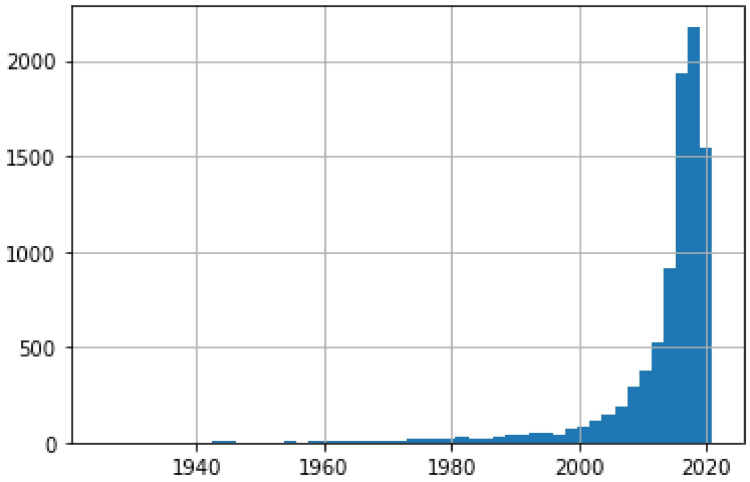
Out[ ]:
```
<AxesSubplot:>
```



In [ ]:
```
# Create a histogram of the release_year column (plotly)
px.histogram(dataset, x="release_year", title="Release Year Histogram", labels={"rel
```

In [ ]:
```
#Create a histogram of the release_year column(matplotlib)
dataset["release_year"].hist(bins=50)
```

Out[ ]:
```
<AxesSubplot:>
```



In [ ]:
```
# Create a histogram of the release_year column (plotly)
px.histogram(dataset, x="date_added", title="Date added to Netflix Histogram", label
```

In [ ]:
```
# Check only what title was released in the year 1925 and the decription of the titl
dataset[dataset["release_year"] == 1925]
```

Out[ ]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	durati
4250	s4251	TV Show	Pioneers: First Women Filmmakers*	NaN	NaN	NaN	2018-12-30	1925	TV-14	1 Seas

In [ ]:
```
# Extract month from the date_added column
dataset["added_month"] = dataset["date_added"].dt.month.fillna(0)
dataset["added_day"] = dataset["date_added"].dt.day.fillna(0)
```

In [ ]:
```
px.histogram(dataset, x="added_month", title="Month added to Netflix Histogram", lab
```

In [ ]:
```
px.histogram(dataset, x="added_day", title="Day added to Netflix Histogram", labels=
```

Sentiment Analysis with textblob

In [ ]:
```
#Cheecking the 5 first rows of the description column
# The description column can be used for sentiment analysis
```

```
from textblob import TextBlob
dataset["description"].head()
```

Out[ ]:
```
0    As her father nears the end of his life, filmm...
1    After crossing paths at a party, a Cape Town t...
2    To protect his family from a powerful drug lor...
3    Feuds, flirtations and toilet talk go down amo...
4    In a city of coaching centers known to train I...
Name: description, dtype: object
```

In [ ]:
```
#natural language processing
# Create a function to return the sentiment of the description column
def sentiment_analysis(text):
    analysis = TextBlob(text)
    if analysis.sentiment.polarity > 0:
        return "positive"
    elif analysis.sentiment.polarity == 0:
        return "neutral"
    else:
        return "negative"
```

In [ ]:
```
# Change the character lenght of the description column to a maximum of 200 characte
pd.options.display.max_colwidth = 200
dataset["description"].head()
```

Out[ ]:
```
0    As her father nears the end of his life, filmmaker Kirsten Johnson stages his d
eath in inventive and comical ways to help them both face the inevitable.
1         After crossing paths at a party, a Cape Town teen sets out to prove whethe
r a private-school swimming star is her sister who was abducted at birth.
2          To protect his family from a powerful drug lord, skilled thief Mehdi and
his expert team of robbers are pulled into a violent and deadly turf war.
3         Feuds, flirtations and toilet talk go down among the incarcerated women at
the Orleans Justice Center in New Orleans on this gritty reality series.
4    In a city of coaching centers known to train India’s finest collegiate minds, a
n earnest but unexceptional student and his friends navigate campus life.
Name: description, dtype: object
```

In [ ]:
```
# Create a column called sentiment with the sentiment analysis of the description co
dataset["sentiment"] = dataset["description"].apply(sentiment_analysis)
dataset.head()
```

Out[ ]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	du
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13	

	show_id	type	title	director	cast	country	date_added	release_year	rating	du
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thabang Molaba, Dillon Windvogel, Natasha Thahane, Arno Greeff, Xolile Tshabalala, Getmore Sithole, Cindy Mahlangu, Ryle De Morny, Greteli Fincham, Sello Ma...	South Africa	2021-09-24	2021	TV-MA	Se
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabiha Akkari, Sofia Lesaffre, Salim Kechiouche, Noureddine Farihi, Geert Van Rampelberg, Bakary Diombera	NaN	2021-09-24	2021	TV-MA	1 S
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	2021-09-24	2021	TV-MA	1 S

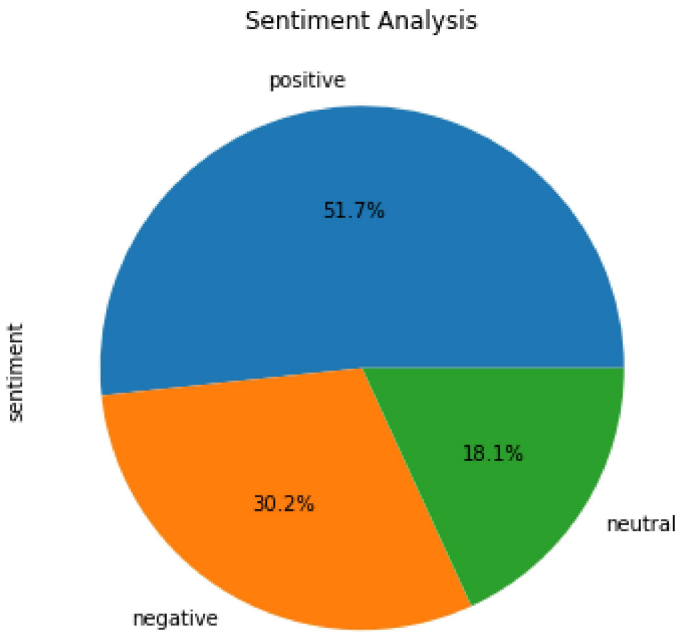
	show_id	type	title	director	cast	country	date_added	release_year	rating	du
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam Khan, Ahsaas Channa, Revathi Pillai, Urvi Singh, Arun Kumar	India	2021-09-24	2021	TV-MA	Se

In [ ]:
```
#Passing a description to the sentiment analysis function
sentiment_analysis("In a city of coaching centers known to train India’s finest coll
```

Out[ ]:
```
'neutral'
```

In [ ]:
```
# Create a pie chart of the sentiment column
dataset.sentiment.value_counts().plot(kind="pie", autopct="%1.1f%%" , figsize=(6,6),
```

Out[ ]:
```
<AxesSubplot:title={'center':'Sentiment Analysis'}, ylabel='sentiment'>
```

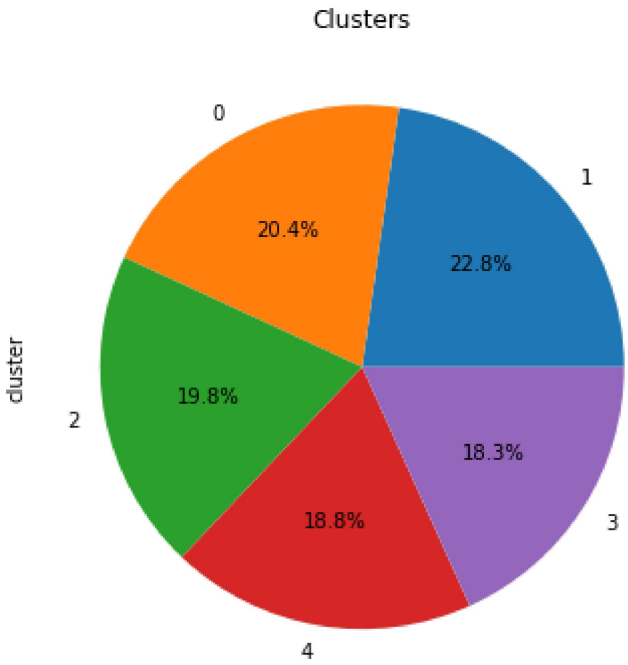


K-means clustering based on the added month & day

In [ ]:
```
#Apply k-means clustering to the dataset
from sklearn.cluster import KMeans
#apply k-means clustering to the dataset
kmeans = KMeans(n_clusters=5, random_state=0).fit(dataset[["added_month", "added_day
# Add a column to the dataset called cluster
dataset["cluster"] = kmeans.labels_
```

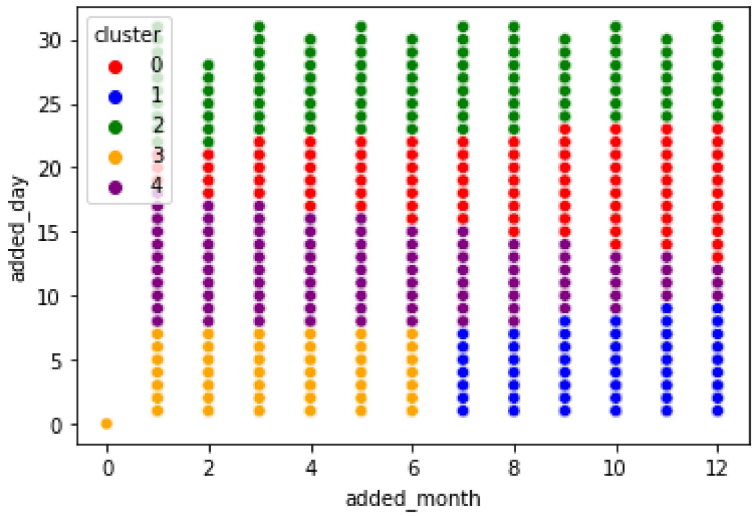
In [ ]:
```
#plot the clusters
dataset.cluster.value_counts().plot(kind="pie", autopct="%1.1f%%" , figsize=(6,6), t
```

Out[ ]:
```
<AxesSubplot:title={'center':'Clusters'}, ylabel='cluster'>
```



In [ ]:
```
# Create a scatter plot of the added_month and added_day columns
sns.scatterplot(x="added_month", y="added_day", data=dataset, hue="cluster", palette
```

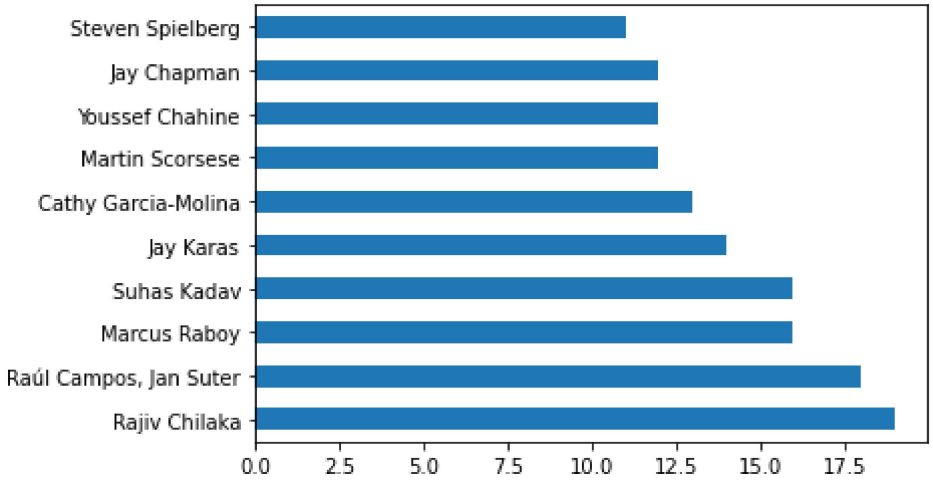
Out[ ]:
```
<AxesSubplot:xlabel='added_month', ylabel='added_day'>
```



Bonus EDA and functions

In [ ]:
```
# plot the top 10 directors appearing in multiple titles
dataset.director.value_counts()[dataset.director.value_counts() > 1].head(10).plot(k
```

Out[ ]:
```
<AxesSubplot:>
```



In [ ]:
```
#find the number of movies based on release year
def find_num_movies(dataset):
    num_movies = dataset.groupby('release_year').size()
    return num_movies
```

In [ ]:
```
find_num_movies(dataset)
```

Out[ ]:
```
release_year
1925       1
1942       2
1943       3
1944       3
1945       4
         ...
2017    1032
2018    1147
2019    1030
2020     953
2021     592
Length: 74, dtype: int64
```

In [ ]:
```
#function to input the director name and return a list of movies
def find_movies(dataset, director):
    movies = dataset[dataset.director == director]
    return movies
```

In [ ]:
```
find_movies(dataset, "Steven Spielberg")
```

Out[ ]:

show_id	type	title	director	cast	country	date_added	release_year	rati

	show_id	type	title	director	cast	country	date_added	release_year	rati
41	s42	Movie	Jaws	Steven Spielberg	Roy Scheider, Robert Shaw, Richard Dreyfuss, Lorraine Gary, Murray Hamilton, Carl Gottlieb, Jeffrey Kramer, Susan Backlinie, Jonathan Filley, Ted Grossman	United States	2021-09-16	1975	
329	s330	Movie	Catch Me If You Can	Steven Spielberg	Leonardo DiCaprio, Tom Hanks, Christopher Walken, Martin Sheen, Nathalie Baye, Amy Adams, James Brolin, Brian Howe, Frank John Hughes, Steve Eastin	United States, Canada	2021-08-01	2002	PG·
1203	s1204	Movie	The BFG	Steven Spielberg	Mark Rylance, Ruby Barnhill, Penelope Wilton, Jemaine Clement, Rebecca Hall, Rafe Spall, Bill Hader, Ólafur Darri Ólafsson, Adam Godley, Michael Adamthwaite, Daniel Bacon, Jonathan Holmes, Chris G...	United States, India, United Kingdom	2021-03-15	2016	
7070	s7071	Movie	Indiana Jones and the Kingdom of the Crystal Skull	Steven Spielberg	Harrison Ford, Cate Blanchett, Karen Allen, Ray Winstone, John Hurt, Jim Broadbent, Igor Jijikine, Shia LaBeouf	United States	2019-01-01	2008	PG·

	show_id	type	title	director	cast	country	date_added	release_year	rati
7071	s7072	Movie	Indiana Jones and the Last Crusade	Steven Spielberg	Harrison Ford, Sean Connery, Denholm Elliott, Alison Doody, John Rhys-Davies, Julian Glover, River Phoenix, Michael Byrne, Kevork Malikyan, Robert Eddison	United States	2019-01-01	1989	PG·
7072	s7073	Movie	Indiana Jones and the Raiders of the Lost Ark	Steven Spielberg	Harrison Ford, Karen Allen, Paul Freeman, Ronald Lacey, John Rhys-Davies, Denholm Elliott, Alfred Molina, Wolf Kahler, Anthony Higgins, Vic Tablian	United States	2019-01-01	1981	
7073	s7074	Movie	Indiana Jones and the Temple of Doom	Steven Spielberg	Harrison Ford, Kate Capshaw, Amrish Puri, Roshan Seth, Philip Stone, Roy Chiao, Jonathan Ke Quan, David Yip, Ric Young, Chua Kah Joo	United States	2019-01-01	1984	
7308	s7309	Movie	Lincoln	Steven Spielberg	Daniel Day-Lewis, Sally Field, David Strathairn, Joseph Gordon-Levitt, James Spader, Hal Holbrook, Tommy Lee Jones, Jackie Earle Haley, John Hawkes, Jared Harris, Joseph Cross, Tim Blake Nelson, D...	United States, India	2018-02-21	2012	PG·

	show_id	type	title	director	cast	country	date_added	release_year	rati
7957	s7958	Movie	Schindler's List	Steven Spielberg	Liam Neeson, Ben Kingsley, Ralph Fiennes, Caroline Goodall, Jonathan Sagall, Embeth Davidtz, Małgorzata Gebel, Shmulik Levy, Mark Ivanir, Beatrice Macola, Friedrich von Thun, Andrzej Seweryn	United States	2018-04-01	1993	
8184	s8185	Movie	The Adventures of Tintin	Steven Spielberg	Jamie Bell, Andy Serkis, Daniel Craig, Nick Frost, Simon Pegg, Daniel Mays, Gad Elmaleh, Toby Jones, Joe Starr	United States, New Zealand, United Kingdom	2019-11-20	2011	
8696	s8697	Movie	War Horse	Steven Spielberg	Emily Watson, David Thewlis, Peter Mullan, Niels Arestrup, Tom Hiddleston, Jeremy Irvine, Benedict Cumberbatch, Toby Kebbell, David Kross, Eddie Marsan, Nicolas Bro, Rainer Bock, Patrick Kennedy, ...	United States, India	2019-05-06	2011	PG·

In [ ]:
```
#function to input the title and return the duration of the movie
def find_duration(dataset, title):
    duration = dataset[dataset.title == title].duration
    return duration
```

In [ ]:
```
#Input the title and return the duration
find_duration(dataset, "Ganglands")
```

Out[ ]:
```
2    1 Season
Name: duration, dtype: object
```