INNOMATICS
RESEARCH LABS

Web Scraping and Data Analysis on Movies in AMAZON PRIME VIDEOS

Team Members:

RAGHVAN

VAMSHI

About me

- **Background ?**
- **Bachelor of Technology**
- **Bachelor of Statistics(BSC)**
- **Why you want to learn Data Science ?**
- **To know how to use the skills in math, statistics, programming and other related subjects organize large data sets.**
- **To become data scientist**
- **Any work experience**
- **No Work Experience**

CONTENTS

OBJECTIVES

- Our Analysis is basically on finding best rated movies and their releasing year
- Exploratory data analysis on our dataset to understand the number of movies released in a year and their ratings

Web Scraping

- To Extracting data using web scraping with python, you need to follow these basic steps:

➢ URL that you want to scrape.

➢ Inspecting the page.

➢ Find the data you want to extract.

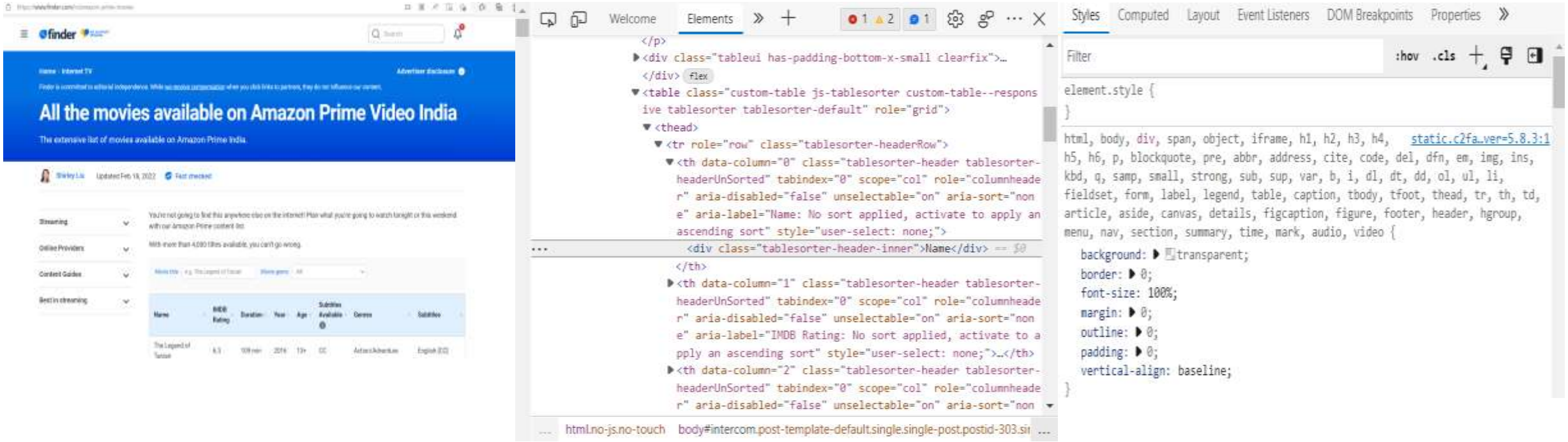
➢ Write the code.

➢ Run the code and extract the data.

➢ Store the data in the required format.

Website

- URL: https//www.finder.com/in/amazon-prime-movies



Library Used

- ✓ requests
- ✓ BeautifulSoup
- ✓ pandas
- ✓ Numpy
- ✓ seaborn
- ✓ RegEx
- ✓ matplotlib

Summary

- From the above libraries and code used, we are now able to make a data frame, from which we can use all the information to define the relationship between movies and rating.
- Our Dataset contains 4649 unique values and 7 features (4649*7)

In [115]: df

Out[115]:

	Name	Duration(In Minutes)	Year	Age	Rating	Genre	Subtitles
0	The Legend of Tarzan	109 min	2016	13+	6.3	Action/Adventure	English [CC]
1	The Conjuring 2	133 min	2016	18+	7.4	Horror, Thriller	
2	Anomalisa	90 min	2015	18+	7.3	Animated, Drama, Mature	English [CC], हिन्दी
3	Baghtos Kay Mujra Kar	141 min	2017	NR	8.8	Drama, Family, Drama, Mystery, Romance, Drama...	English [CC]
4	13 Hours: The Secret Soldiers of Benghazi	144 min	2016	NR	7.3	Drama	English [CC], हिन्दी
...	...	...	...	...	...	...	...
4644	Five Little Fingers, Finger Family Song and Ma...	65 min	2016	NR		Animated, Children, Children, Animated Cartoon...	
4645	Five Little Monkeys Jumping On The Bed and Man...	100 min	2015	NR		Animated, Children, Children, Animated Cartoon...	
4646	Wheels on the Bus – Wild Animals and Animal So...	91 min	2016	NR		Animated, Children, Animated, Family, Children...	
4647	If You're Happy And You Know It and More Video...	77 min	2014	NR		Animated, Children, Children, Animated Cartoon...	
4648	Johny Johny Yes Papa PART 2 and Many More Vide...	74 min	2015	NR		Animated, Children, Animated, Family, Children...	

4649 rows × 7 columns

Extracting the DataFrame

- We have extracted the DataFrame 4649 to 400 rows.

```
In [117]: #extracting 400 rows

In [118]: df.head(400)
```

Out[118]:

	Name	Duration(In Minutes)	Year	Age	Rating	Genre	Subtitles
0	The Legend of Tarzan	109 min	2016	13	6.3	Action/Adventure	English [CC]
1	The Conjuring 2	133 min	2016	18	7.4	Horror, Thriller	
2	Anomalisa	90 min	2015	18	7.3	Animated, Drama: Mature	English [CC], हिन्दी
3	Baghtos Kay Mujra Kar	141 min	2017	NR	8.8	Drama: Family, Drama: Mystery, Romance: Drama,...	English [CC]
4	13 Hours: The Secret Soldiers of Benghazi	144 min	2016	NR	7.3	Drama	English [CC], हिन्दी
...	...	...	...	...	...	...	...
395	The Perfect Storm	129 min	2000	13	6.4	Action/Adventure: Thriller, Drama: Thriller	
396	U.S. Marshals	131 min	1998	13	6.5	Action/Adventure: Espionage & Spies	
397	Interview With the Vampire: The Vampire Chroni...	122 min	1994	18	7.6	Drama: Period, Horror: Vampires, Sci-Fi/Fantas...	
398	Jeans	166 min	1998	All Ages	6.3	Drama: Romance, Romance: Drama, World Cinema: ...	English [CC]
399	Jackie Chan's First Strike	83 min	1997	13	6.6	Action/Adventure: Espionage & Spies, Action/Ad...	English [CC]

400 rows × 7 columns

Data cleaning – Removing null values

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 7 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Name                  400 non-null    object
 1   Duration(In Minutes)  400 non-null    int32
 2   Year                  400 non-null    int64
 3   Age                   400 non-null    object
 4   Rating                400 non-null    float64
 5   Genre                 400 non-null    object
 6   Subtitles             400 non-null    object
dtypes: float64(1), int32(1), int64(1), object(4)
memory usage: 20.4+ KB
```

Converting the CSV Format

```
In [69]: final_df
```

Out[69]:

	Name	Duration(In Minutes)	Year	Age	Rating	Genre	Subtitles
0	The Legend of Tarzan	109	2016	13+	6.3	Action/Adventure	English [CC]
1	The Conjuring 2	133	2016	18+	7.4	Horror, Thriller	English
2	Anomalisa	90	2015	18+	7.3	Animated, Drama: Mature	English [CC], हिन्दी
3	Baghtos Kay Mujra Kar	141	2017	NR	8.8	Drama: Family, Drama: Mystery, Romance: Drama,...	English [CC]
4	13 Hours: The Secret Soldiers of Benghazi	144	2016	NR	7.3	Drama	English [CC], हिन्दी
...	...	...	...	...	...	...	...
395	The Perfect Storm	129	2000	13+	6.4	Action/Adventure: Thriller, Drama: Thriller	English
396	U.S. Marshals	131	1998	13+	6.5	Action/Adventure: Espionage & Spies	English
397	Interview With the Vampire: The Vampire Chroni...	122	1994	18+	7.6	Drama: Period, Horror: Vampires, Sci-Fi/Fantas...	English
398	Jeans	166	1998	All Ages	6.3	Drama: Romance, Romance: Drama, World Cinema: ...	English [CC]
399	Jackie Chan's First Strike	83	1997	13+	6.6	Action/Adventure: Espionage & Spies, Action/Ad...	English [CC]

400 rows × 7 columns

Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or ti check assumptions with the help of statistical summary and graphical representations. For the simplicity of the article, we will use a single dataset.
- Types of Exploratory Data Analysis:

✓Univariate

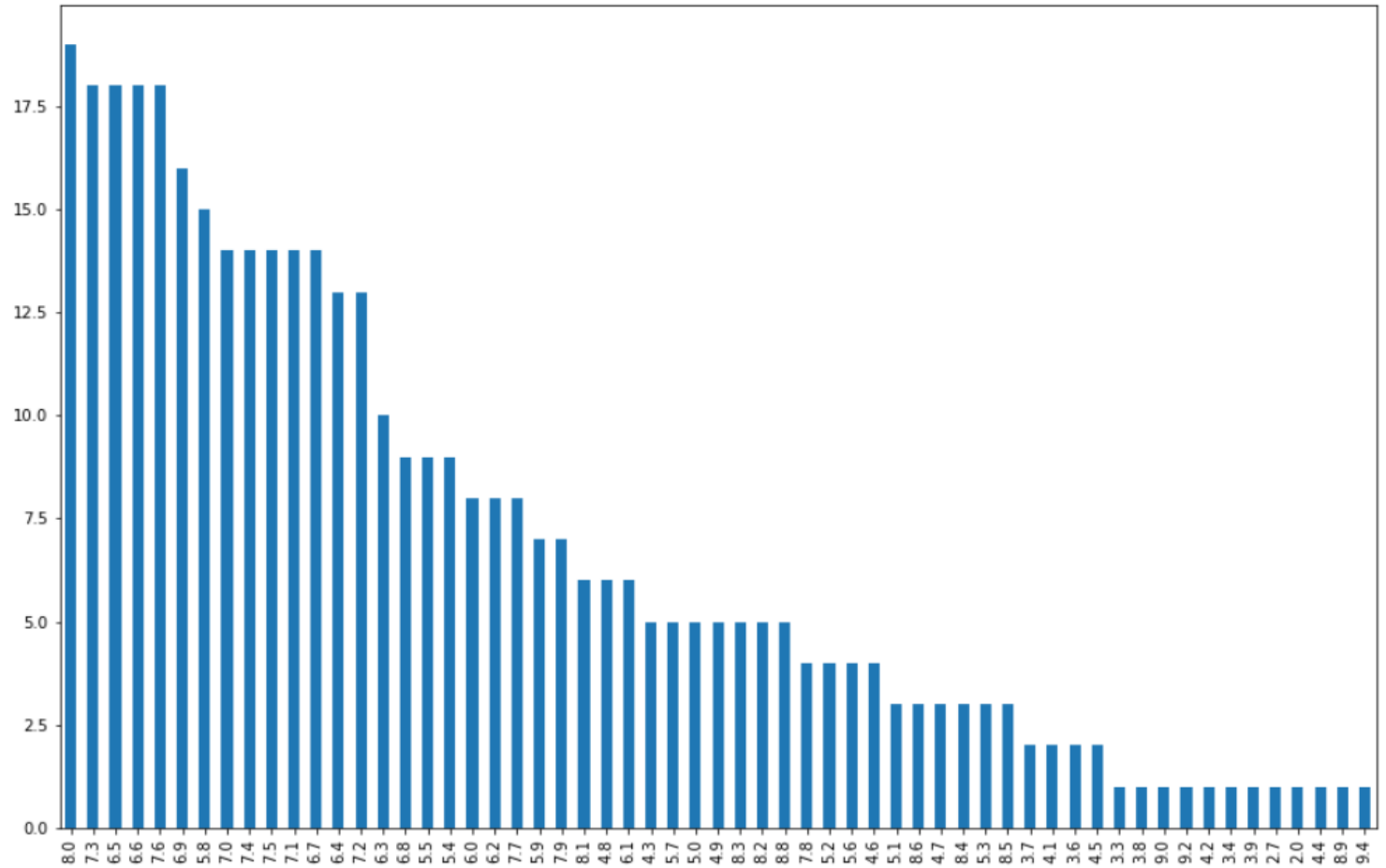
✓Bi-Variate

✓Multi Variate

Univariate

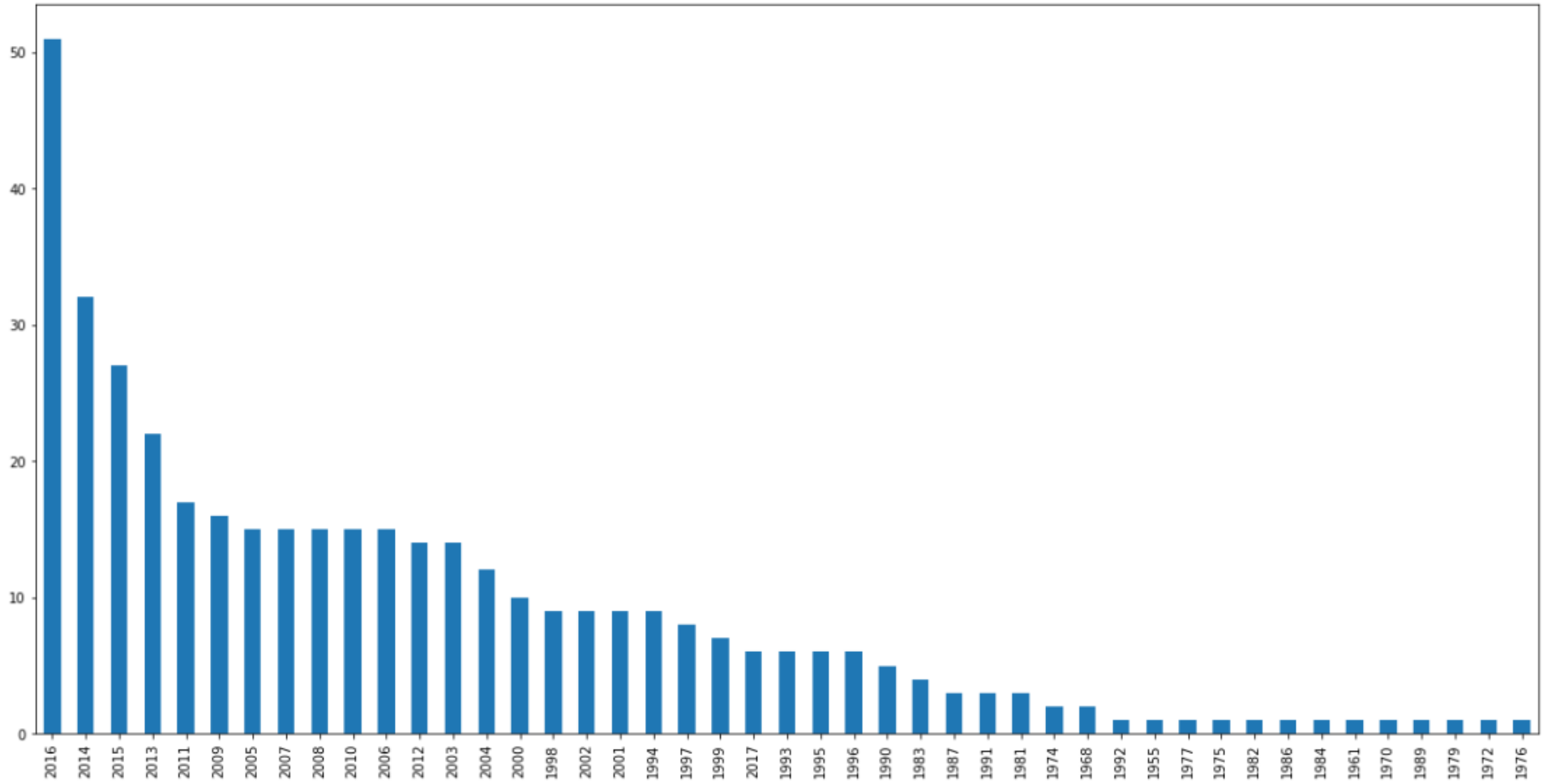
Movies and their Ratings

```
Out[143]: <AxesSubplot:>
```



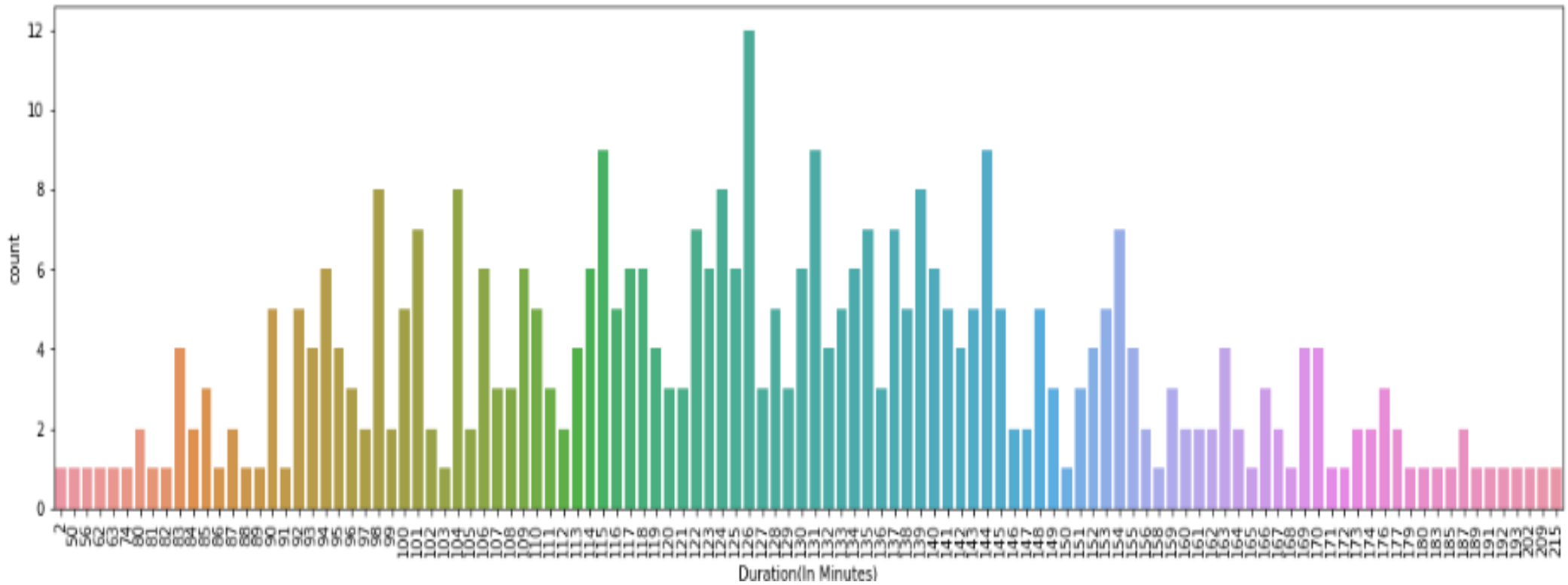
Movies and their Releasing Year

Out[148]: <AxesSubplot:>



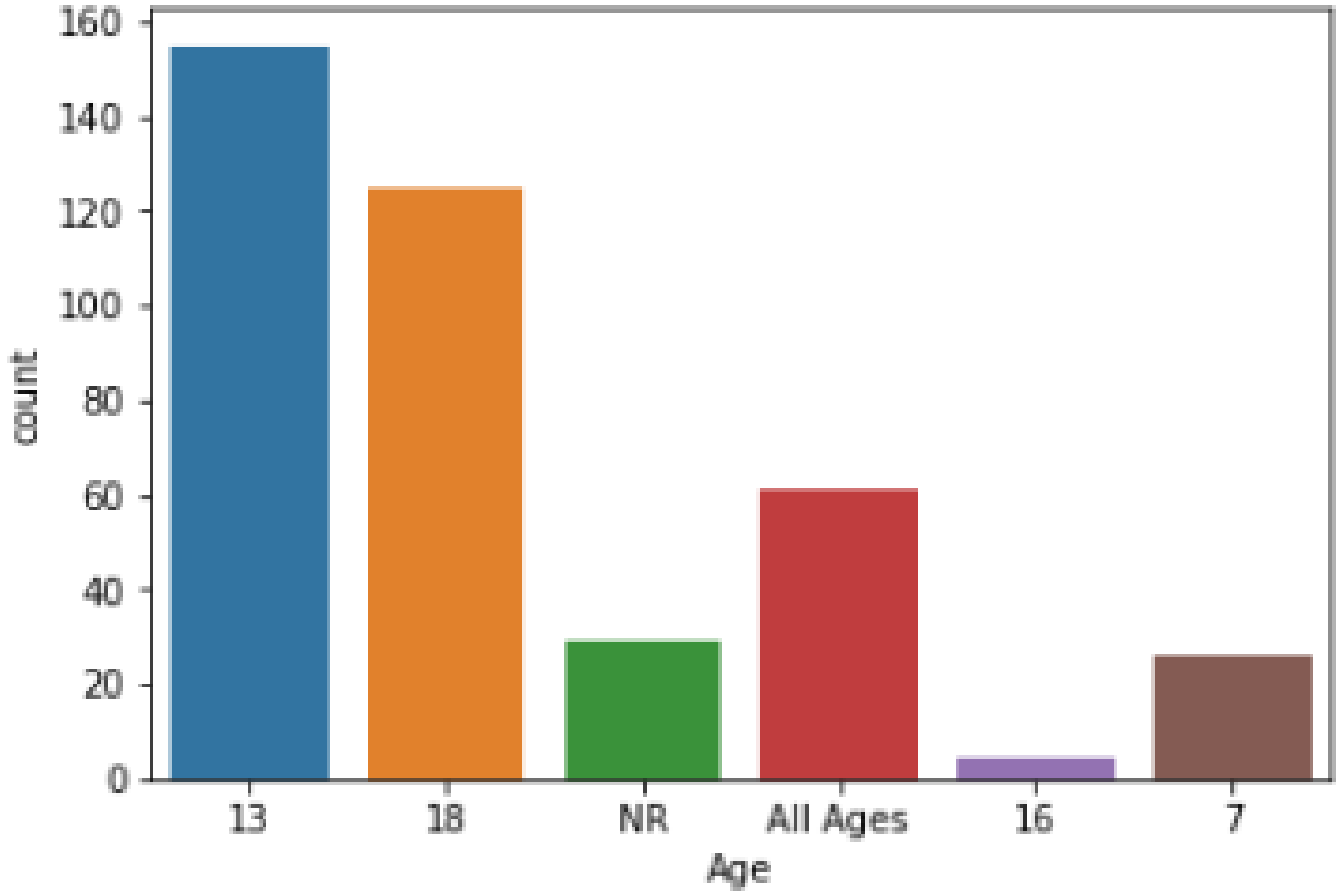
Duration of the movies

```
Out[72]: <AxesSubplot:xlabel='Duration(In Minutes)', ylabel='count'>
```



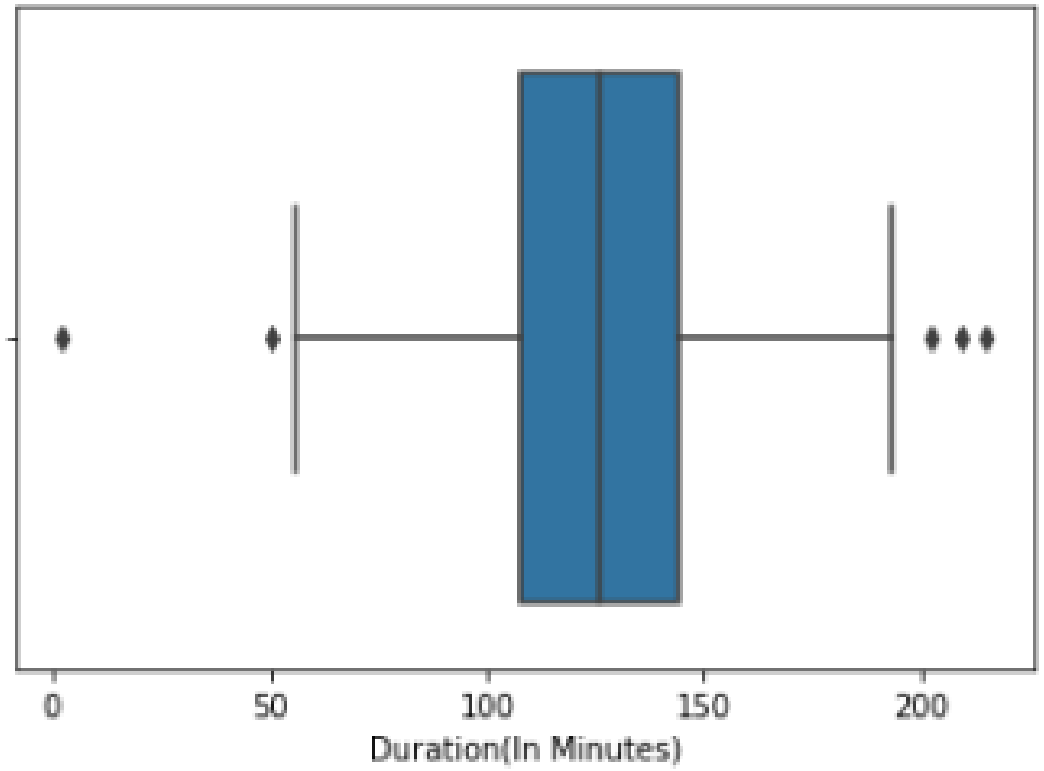
Movies according to Age Limit

```
Out[88]: <AxesSubplot:xlabel='Age', ylabel='count'>
```



Boxplot

Out[87]: <AxesSubplot:xlabel='Duration(In Minutes)'>



conclusion

- From the above observation we have found similar rated movies and the releasing year.
- We have concluded movies for various Age Group.

THANK YOU