

# • **NETNAIJA**

## **Data Wrangling and Analysis**

**Achimugu A . S**  
**20<sup>th</sup> May 2022**

# Outline

---

- **Introduction**
- **Executive Summary**
- **Methodology**
- **Results**
- **Conclusion**
- **Appendix**

# Introduction

---

Netnaija is media download center created in March of 2016. The project aims at scraping data from the movies section that has been uploaded from 2016 till date and answering some questions from the data. Questions intended to be answered are

1. Which year has the highest number of upload?
2. number of movies uploaded by month in the year 2019
3. What movies has highest movie length?
4. Stars with most number of movies
5. Find movies which Scott Adkin was an actor
6. How long from on average from official release of movie to site upload
7. Trend between release and Update
8. Compared top 4 genre with others genre in a pie chart
9. What movie got movie people talking most and what is the summary
10. Trend in upload

# Executive Summary

---

Data was scraped from [Netnaija](#) with beautifulsoup and request library. Scraped data contains 3073 rows and 13 columns.

Pandas was used to clean wrangle the data and after the process 82% of row and 11 columns were left for the analysis

Answer were gotten to initially proposed questions after the anlaysis. Question like star with the most number of movie and How long from on average from offical release of movie to site upload were found to Nicolas Cage and 7 days

Section 1

# Methodology

# Data Collection

---

For the purpose of the analysis, Each movie title, movie link, movie type, time of upload, movie length, number of comment, movie summary, Genre, Release Date, Stars, movie Languages, movie Subtitles, IMDB links were scraped. Not all not all feature listed were present in some movie in which case I replaced them with missing

# Data Wrangling

---

After getting the scraped data, pandas package was used to do some cleaning by dropping duplicate rows, removing parentheses and semi colon present in the data. Columns that were not significant to the analysis were also dropped. Regex was used to parse the movie from the title added it to the upload date and each column was casted to the appropriate data type. Nan values were removed, replaced or left alone depending on the column and final index was set to be the upload date.

After the process 82% of row and 11 columns were left for the analysis

# EDA with Data Visualization

---

Visualisation was done with Seaborn and matplotlib Pie graph plot, Bar plot and Line plot;

**Pie Chart** was used to compare the genre of movies. Action, drama, comedy, Thriller and the rest genre which was grouped as other

**Bar chart** was used to show the comparisons between Total Upload by Month(2019), Total Upload by Year. Year and months are both categorical data hence the choice of Bar chart.

**Line plot** was used to show the Yearly\_avg between Release and Upload and Trend in number of upload



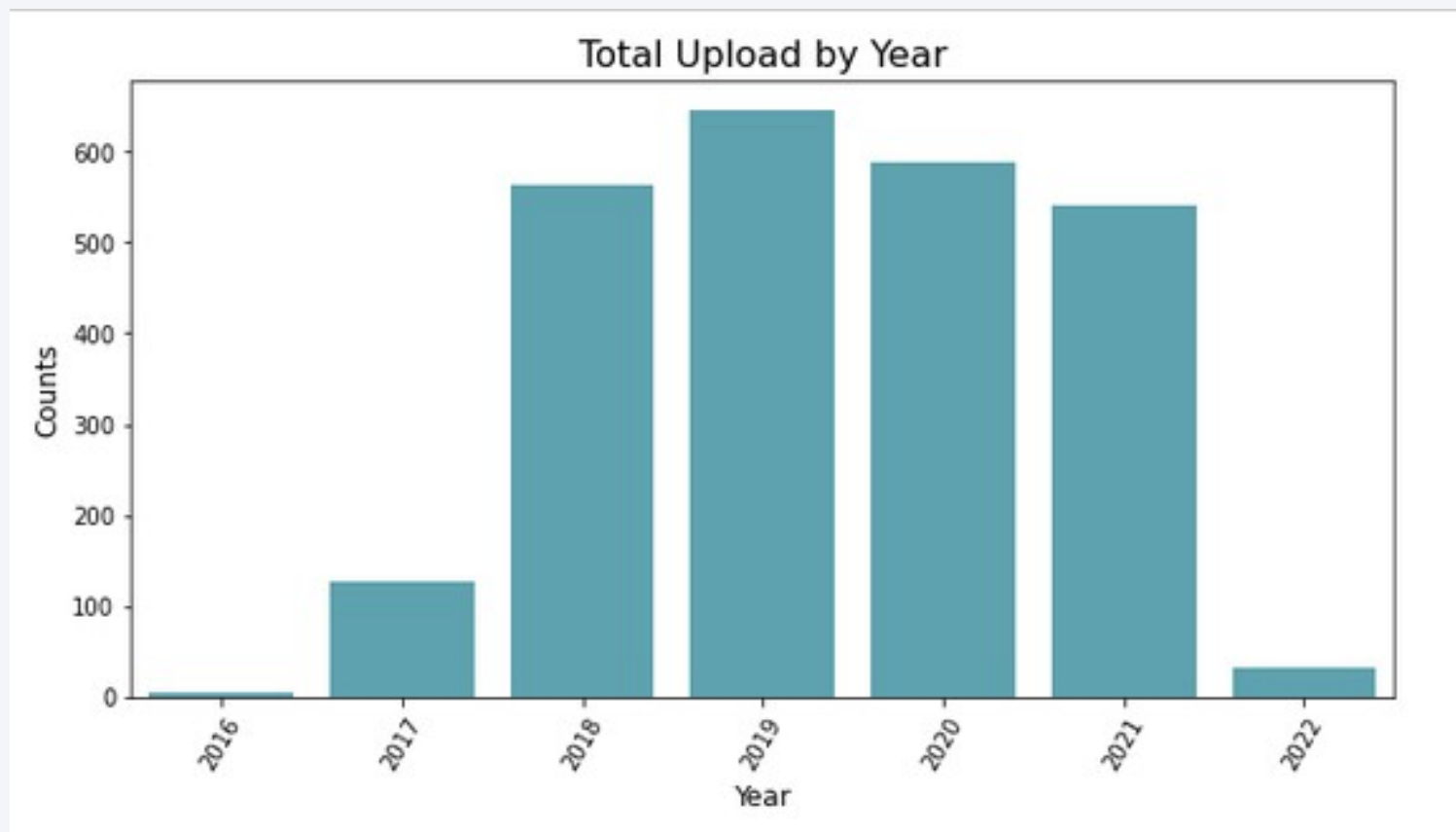
Section 2

# **Insights drawn from EDA**

# Uploads by Year

---

1. Which year has the highest number of upload?

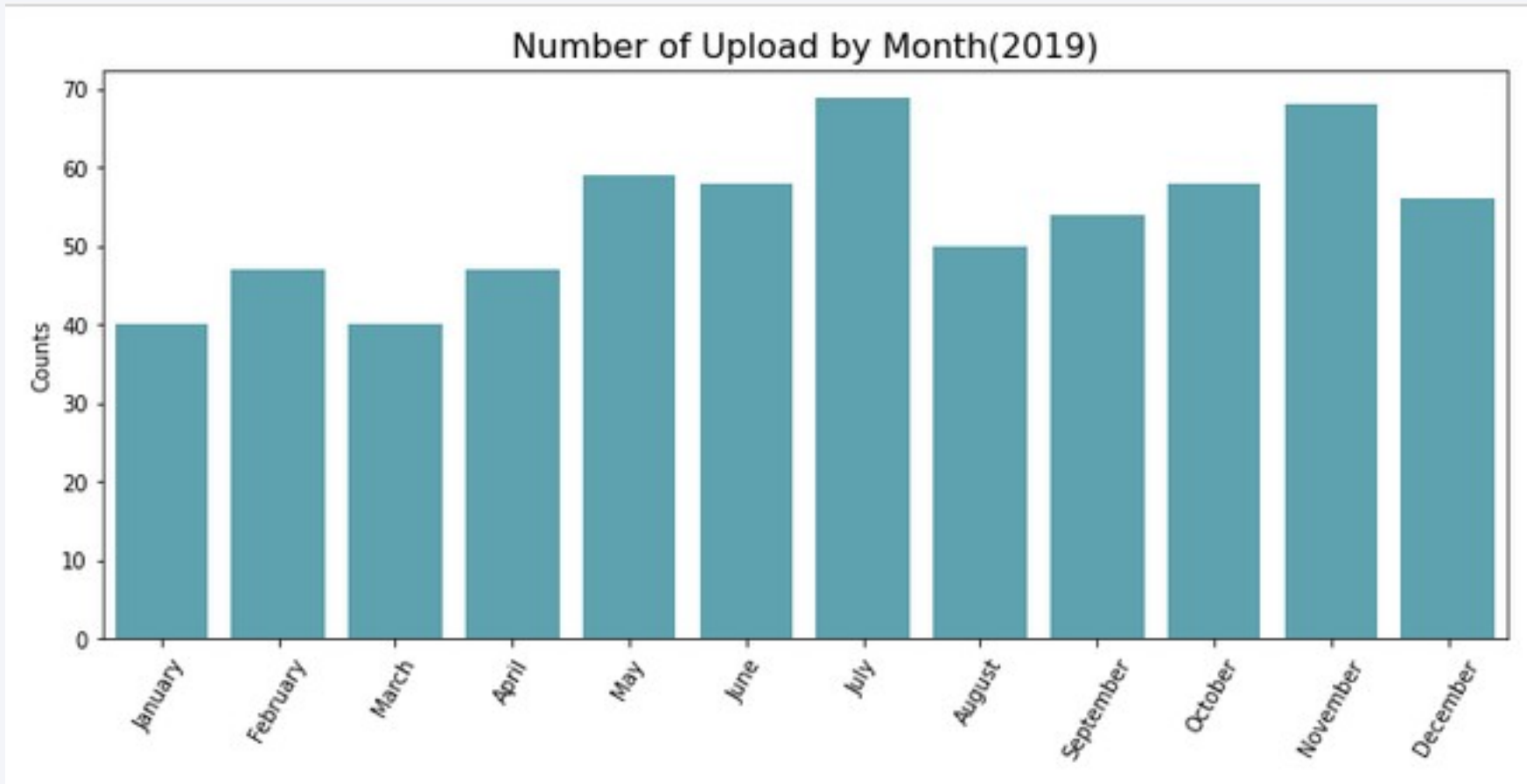


2019 has the highest number of upload with 646 movies followed by a decline in 2020 and 2021 with a count of 589 and 541 respectively

# Number of Upload by Month in 2019

---

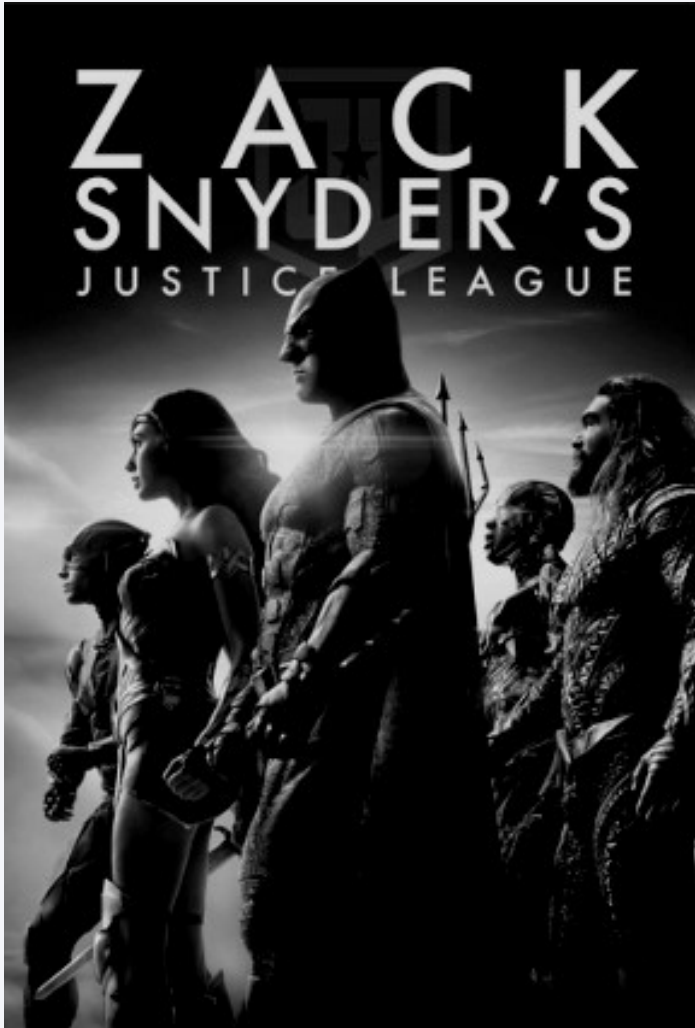
2. Find the number of movies by month in the year 2019



July is the highest with a count of 68 while January and March are the lowest with 40

# Movie with the highest length

---



3. What movies has highest movie length

**Title** Zack Snyder's Justice League

**Lenght** 04:02:03

## **Summary**

Determined to ensure Superman's ultimate sacrifice was not in vain, Bruce Wayne aligns forces with Diana Prince with plans to recruit a team of metahumans to protect the world from an approaching threat of catastrophic proportions

**Stars** Henry Cavill, Ben Affleck, Gal Gadot

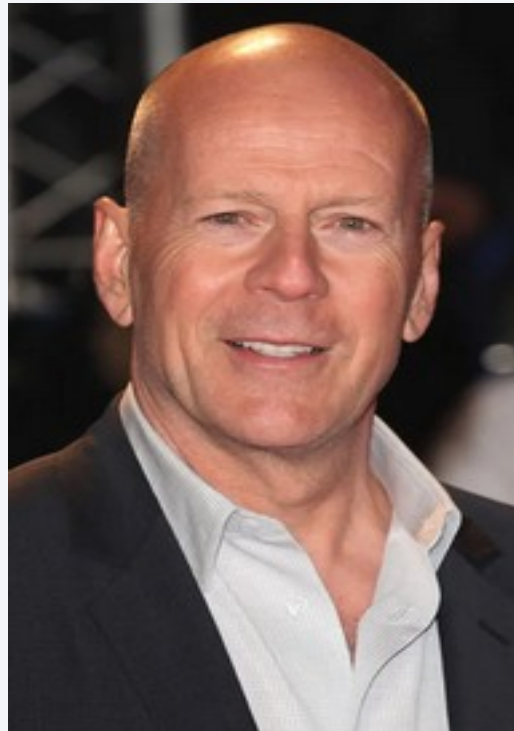
# Stars with most number of movies

---

## 4. 3 stars with most number of movies



Nicolas Cage



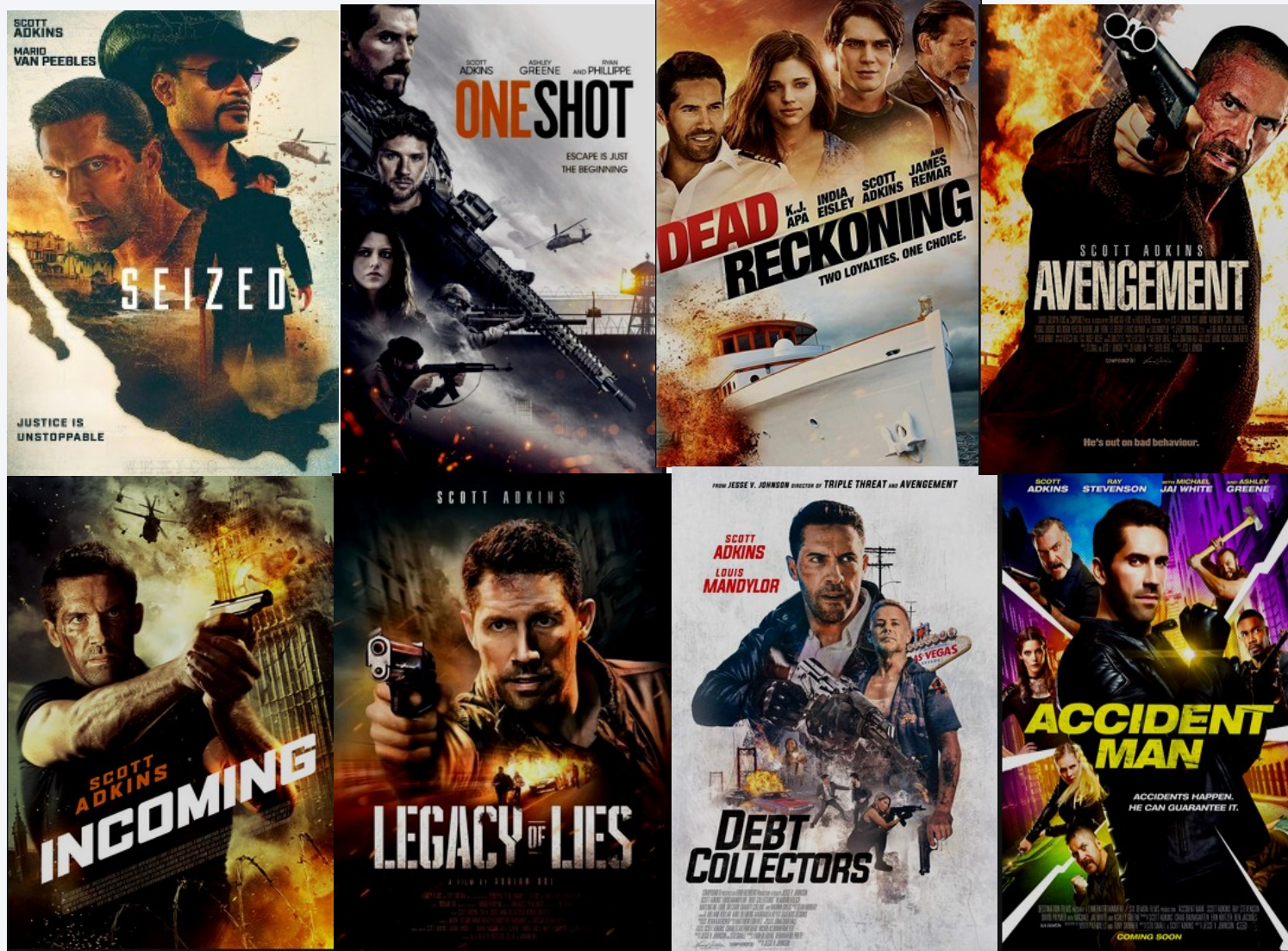
Bruce Willis



Samuel L. Jackson



# Movies Scott Adkins acted



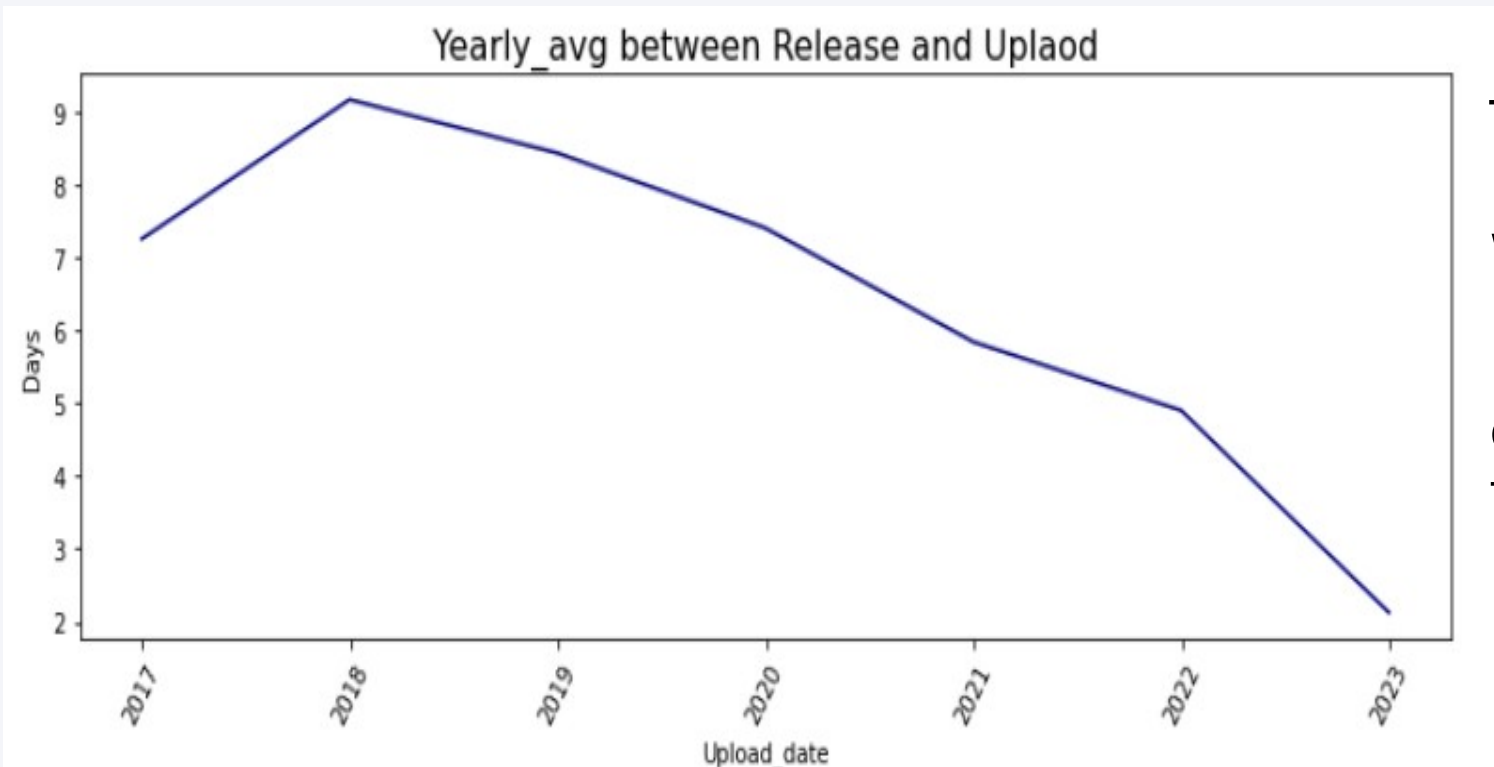
## Titles

1. One Shot
2. Dead Reckoning
3. Seized
4. Legacy of Lies
5. Debt Collectors
6. Avengement
7. Incoming
8. Accident Man

# Wait Time

---

6. How long from on average from official release of movie to site upload
7. Trend between release and Upload

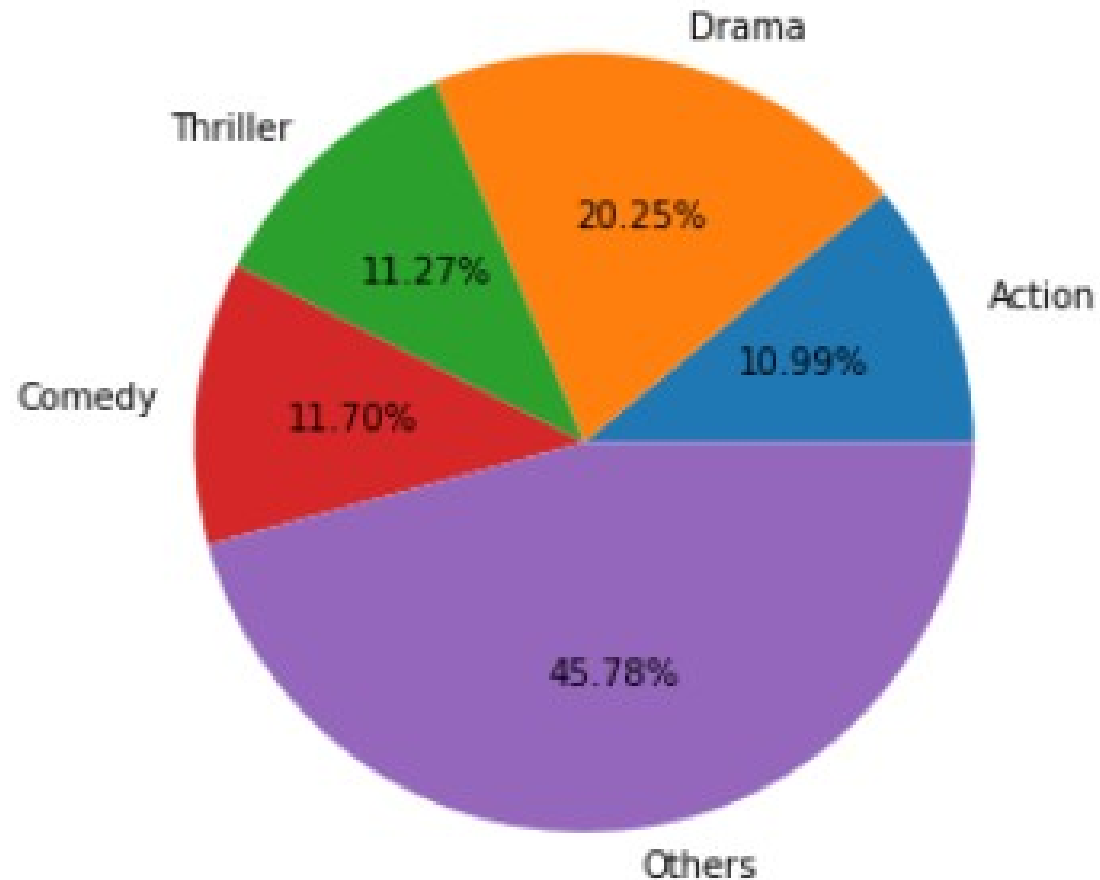


They has been a progressive decrease in the wait time from official release of a movie to when it's available on the site for download. The average wait time is 7 days

# Genre

---

8. Compared top 4 genre with others genre in a pie chart



**Comedy:** 11.70%

**Thriller:** 11.27%

**Drama:** 20.25%

**Action:** 10.99%

**Other:** 45.78%



# Movie with the Most Comment

---



9. What movie got movie people talking most

**Title** Fast and Furious 9

## **Summary**

Dominic Toretto and his crew battle the most skilled assassin and high-performance driver they've ever encountered: his forsaken brother.

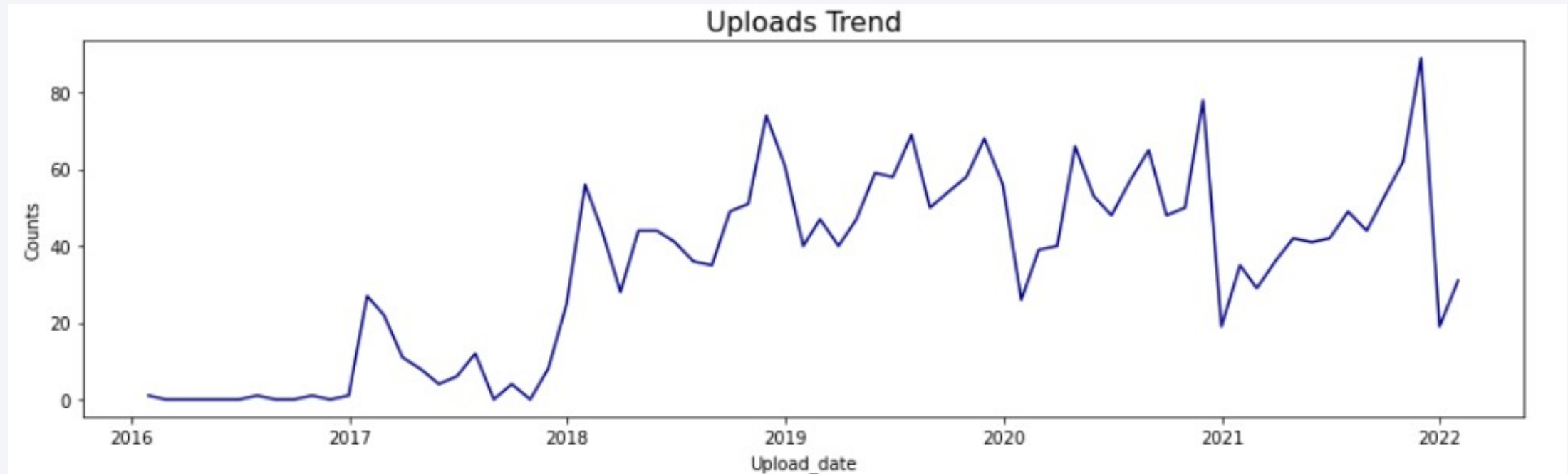
Cypher enlists the help of Jakob, Dom's younger brother to take revenge on Dom and his team.

**Stars** Vin Diesel, Michelle Rodriguez, Jordana Brewster

# Trend in Upload

---

## 10. Trend in upload



Around the first quarter of each year they seem to be a high surge in number of uploads.

# Conclusions and Limitations

---

Most of the initial questions were answered.

## **Limitations**

Number of upload per has progressively incread from 2016 till 2019 and decrease in 2020 and 2021. What has caused the decrease? Could it be covid 19 or something else wwe need post covid data to confirm.

Why is they a surge in the number of upload every first quater of the year? That a question this current data can't answer among other limitations.

# Appendix

---

- <https://www.thenetnaija.com/videos/movies/>
- [https://en.wikipedia.org/wiki/Nicolas\\_Cage](https://en.wikipedia.org/wiki/Nicolas_Cage)
- [https://en.wikipedia.org/wiki/Bruce\\_Willis](https://en.wikipedia.org/wiki/Bruce_Willis)
- [https://en.wikipedia.org/wiki/Samuel\\_L.\\_Jackson](https://en.wikipedia.org/wiki/Samuel_L._Jackson)
- [https://en.wikipedia.org/wiki/Fast\\_%26\\_Furious](https://en.wikipedia.org/wiki/Fast_%26_Furious)
- [https://en.wikipedia.org/wiki/Scott\\_Adkins](https://en.wikipedia.org/wiki/Scott_Adkins)

**Thank You!**