NYC Data Science Academy
Web scraping project

Exploratory analysis of the correlation between variables of Netflix originals and its paid membership

July 2019

FRED (LEFAN) CHENG

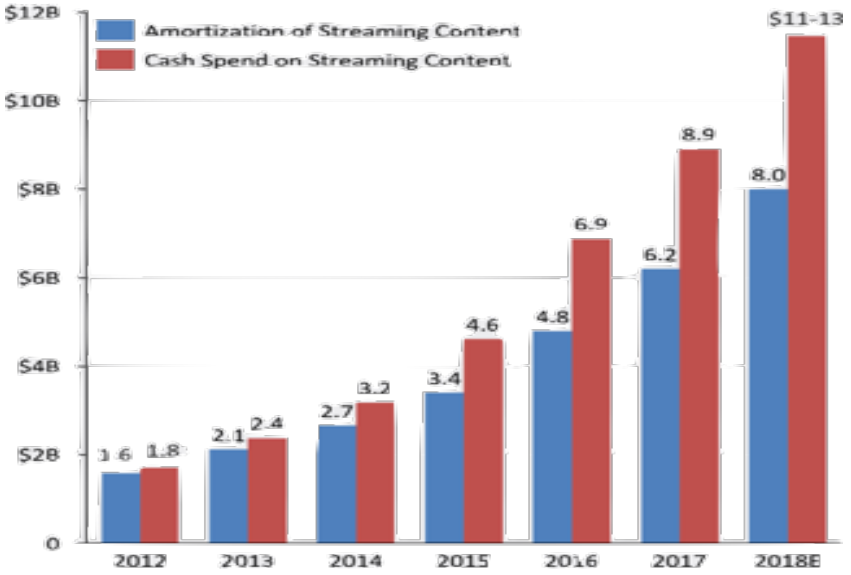
Netflix has maintained a long-run growth over the last number of years

The sustainable capability of producing high-quality original content is a core competitiveness

Shares have steadily rise over 10 years with over 40 times growth since 2012

A $1,000 investment made on Jan, 2007, would be worth more than $110,000 as of April, 2019,

A whopping $13 billion was spent on content in 2018, comprising around 85% of spending



-> Exploring Netflix original data

Sources: Google Finance, Ball Math: https://redef.com/original/netflix-misunderstandings-pt-1-netflixs-content-budget-is-bigger-than-it-seems

The goal is to briefly explore variables influencing paid memberships and their relationships

Revenue from paid memberships comprises 98.22% of total revenue in the 1st Q of 2019.

Domestic Streaming

Paid memberships at end of period	60,229
Paid net membership additions	1,743
Free trials	1,563
Revenues	$ 2,073,555

International Streaming

Paid memberships at end of period	88,634
Paid net membership additions	7,861
Free trials	5,003
Revenues	$ 2,366,749

Consolidated Statements of Operations

Revenues	$ 4,520,992

Being able to forecast paid memberships would be highly valuable

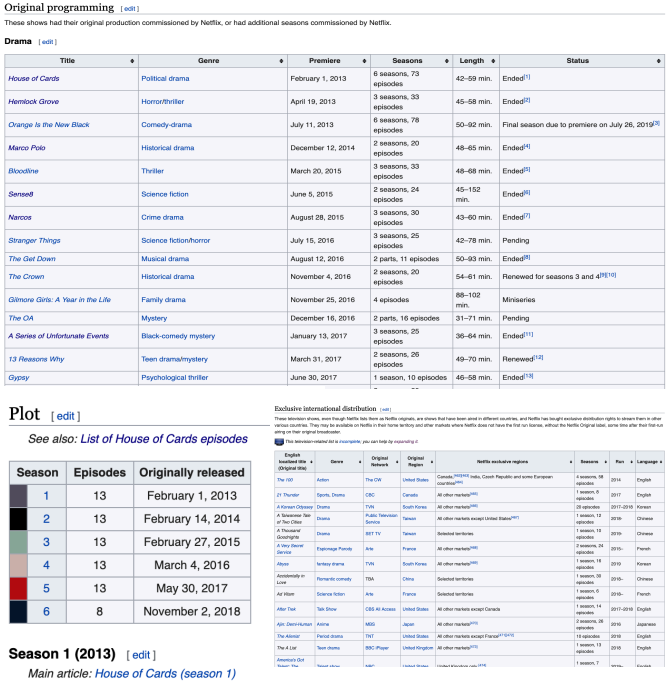
Revenue + Stock Price

Sources: Netflix Quarter report of 1st Q in 2019 from official website

Data sources from IMBD, Wikipedia, and Netflix Media Center

Tool: Scrapy

Wikipedia



Original programming [edit]

These shows had their original production commissioned by Netflix, or had additional seasons commissioned by Netflix.

Drama [edit]

Title	Genre	Premiere	Seasons	Length	Status
House of Cards	Political drama	February 1, 2013	6 seasons, 73 episodes	42–59 min.	Ended[1]
Hemlock Grove	Horror/thriller	April 19, 2013	3 seasons, 33 episodes	45–58 min.	Ended[2]
Orange Is the New Black	Comedy-drama	July 11, 2013	6 seasons, 78 episodes	50–92 min.	Final season due to premiere on July 26, 2019[3]
Marco Polo	Historical drama	December 12, 2014	2 seasons, 20 episodes	48–65 min.	Ended[4]
Bloodline	Thriller	March 20, 2015	3 seasons, 33 episodes	48–68 min.	Ended[5]
Sense8	Science fiction	June 5, 2015	2 seasons, 24 episodes	45–152 min.	Ended[6]
Narcos	Crime drama	August 28, 2015	3 seasons, 30 episodes	43–60 min.	Ended[7]
Stranger Things	Science fiction/horror	July 15, 2016	3 seasons, 25 episodes	42–78 min.	Pending
The Get Down	Musical drama	August 12, 2016	2 parts, 11 episodes	50–93 min.	Ended[8]
The Crown	Historical drama	November 4, 2016	2 seasons, 20 episodes	54–61 min.	Renewed for seasons 3 and 4[9][10]
Gilmore Girls: A Year in the Life	Family drama	November 25, 2016	4 episodes	88–102 min.	Miniseries
The OA	Mystery	December 16, 2016	2 parts, 16 episodes	31–71 min.	Pending
A Series of Unfortunate Events	Black-comedy mystery	January 13, 2017	3 seasons, 25 episodes	36–64 min.	Ended[11]
13 Reasons Why	Teen drama/mystery	March 31, 2017	2 seasons, 26 episodes	49–70 min.	Renewed[12]
Gypsy	Psychological thriller	June 30, 2017	1 season, 10 episodes	46–58 min.	Ended[13]

Plot [edit]

See also: List of House of Cards episodes

Season	Episodes	Originally released
1	13	February 1, 2013
2	13	February 14, 2014
3	13	February 27, 2015
4	13	March 4, 2016
5	13	May 30, 2017
6	8	November 2, 2018

Season 1 (2013) [edit]

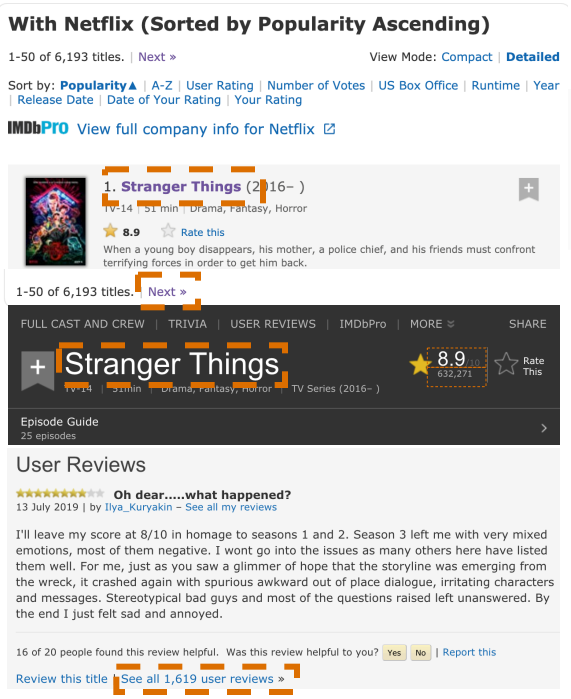
Main article: House of Cards (season 1)

Exclusive international distribution [edit]

List of Netflix Original programming and film

Title, Genre, Premiere of each season, Length, Language, Distribution

IMBD



With Netflix (Sorted by Popularity Ascending)

1-50 of 6,193 titles. | Next »

View Mode: Compact | Detailed

Sort by: Popularity▲ | A-Z | User Rating | Number of Votes | US Box Office | Runtime | Year | Release Date | Date of Your Rating | Your Rating

IMDbPro View full company info for Netflix

1. Stranger Things (2016–)

TV-14 | 51 min | Drama, Fantasy, Horror

8.9 Rate this

When a young boy disappears, his mother, a police chief, and his friends must confront terrifying forces in order to get him back.

1-50 of 6,193 titles. Next »

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE | SHARE

Stranger Things

TV-14 | 51min | Drama, Fantasy, Horror | TV Series (2016–)

8.9 /10 632,271 Rate This

Episode Guide

25 episodes

User Reviews

Oh dear.....what happened?

13 July 2019 | by Ilya_Kuryakin – See all my reviews

I'll leave my score at 8/10 in homage to seasons 1 and 2. Season 3 left me with very mixed emotions, most of them negative. I wont go into the issues as many others here have listed them well. For me, just as you saw a glimmer of hope that the storyline was emerging from the wreck, it crashed again with spurious awkward out of place dialogue, irritating characters and messages. Stereotypical bad guys and most of the questions raised left unanswered. By the end I just felt sad and annoyed.

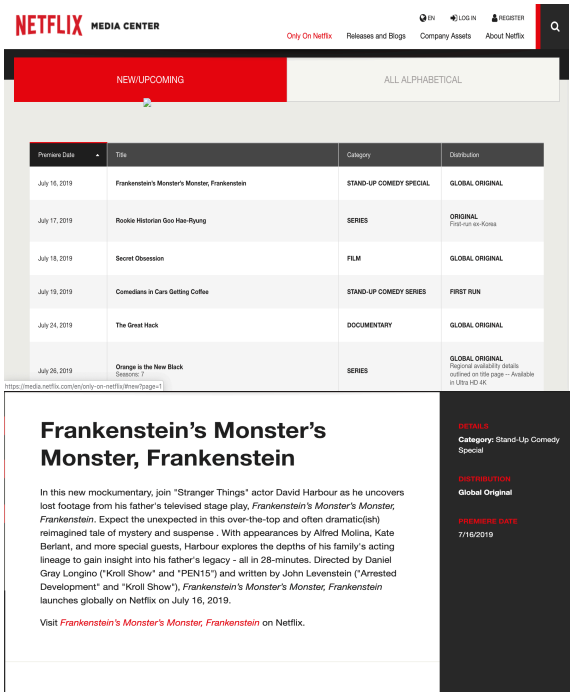
16 of 20 people found this review helpful. Was this review helpful to you? Yes No | Report this

Review this title | See all 1,619 user reviews »

Shows sorted by Netflix as distributor

Title, Number of reviews, Count of rating, Average rating

Netflix Media Center



NETFLIX MEDIA CENTER

Only On Netflix | Releases and Blogs | Company Assets | About Netflix

NEW/UPCOMING | ALL ALPHABETICAL

Premiere Date	Title	Category	Distribution
July 16, 2019	Frankenstein's Monster's Monster, Frankenstein	STAND-UP COMEDY SPECIAL	GLOBAL ORIGINAL
July 17, 2019	Rookie Historian Goo Hae-Ryung	SERIES	ORIGINAL First-run ex-Korea
July 18, 2019	Secret Obsession	FILM	GLOBAL ORIGINAL
July 19, 2019	Comedians in Cars Getting Coffee	STAND-UP COMEDY SERIES	FIRST RUN
July 24, 2019	The Great Hack	DOCUMENTARY	GLOBAL ORIGINAL
July 26, 2019	Orange is the New Black Seasons: 7	SERIES	GLOBAL ORIGINAL Regional availability details outlined on title page -- Available in Ultra HD 4K

https://media.netflix.com/en/only-on-netflix#new?page=1

Frankenstein's Monster's Monster, Frankenstein

In this new mockumentary, join "Stranger Things" actor David Harbour as he uncovers lost footage from his father's televised stage play, *Frankenstein's Monster's Monster, Frankenstein*. Expect the unexpected in this over-the-top and often dramatic(ish) reimagined tale of mystery and suspense . With appearances by Alfred Molina, Kate Berlant, and more special guests, Harbour explores the depths of his family's acting lineage to gain insight into his father's legacy - all in 28-minutes. Directed by Daniel Gray Longino ("Kroll Show" and "PEN15") and written by John Levenstein ("Arrested Development" and "Kroll Show"), *Frankenstein's Monster's Monster, Frankenstein* launches globally on Netflix on July 16, 2019.

Visit *Frankenstein's Monster's Monster, Frankenstein* on Netflix.

DETAILS

Category: Stand-Up Comedy Special

DISTRIBUTION

Global Original

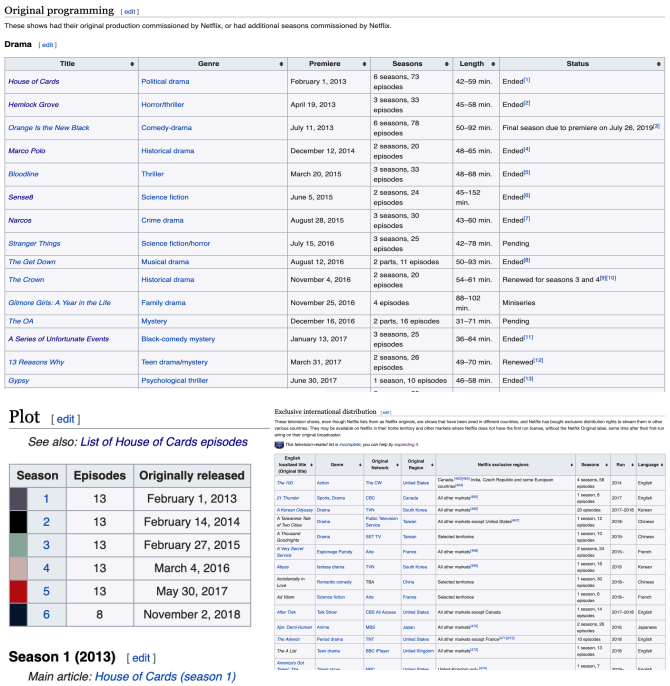
PREMIERE DATE

7/16/2019

Upcoming shows
For future analysis

Challenges

Wikipedia



Original programming [edit]

These shows had their original production commissioned by Netflix, or had additional seasons commissioned by Netflix.

Drama [edit]

Title	Genre	Premiere	Seasons	Length	Status
House of Cards	Political drama	February 1, 2013	6 seasons, 73 episodes	42–59 min.	Ended[1]
Hemlock Grove	Horror/thriller	April 19, 2013	3 seasons, 33 episodes	45–58 min.	Ended[2]
Orange Is the New Black	Comedy-drama	July 11, 2013	6 seasons, 78 episodes	50–92 min.	Final season due to premiere on July 26, 2019[3]
Marco Polo	Historical drama	December 12, 2014	2 seasons, 20 episodes	48–65 min.	Ended[4]
Bloodline	Thriller	March 20, 2015	3 seasons, 33 episodes	48–68 min.	Ended[5]
Sense8	Science fiction	June 5, 2015	2 seasons, 24 episodes	45–152 min.	Ended[6]
Narcos	Crime drama	August 28, 2015	3 seasons, 30 episodes	43–60 min.	Ended[7]
Stranger Things	Science fiction/horror	July 15, 2016	3 seasons, 25 episodes	42–78 min.	Pending
The Get Down	Musical drama	August 12, 2016	2 parts, 11 episodes	50–93 min.	Ended[8]
The Crown	Historical drama	November 4, 2016	2 seasons, 20 episodes	54–61 min.	Renewed for seasons 3 and 4[9][10]
Gilmore Girls: A Year in the Life	Family drama	November 25, 2016	4 episodes	88–102 min.	Miniseries
The OA	Mystery	December 16, 2016	2 parts, 16 episodes	31–71 min.	Pending
A Series of Unfortunate Events	Black-comedy mystery	January 13, 2017	3 seasons, 25 episodes	36–64 min.	Ended[11]
13 Reasons Why	Teen drama/mystery	March 31, 2017	2 seasons, 26 episodes	49–70 min.	Renewed[12]
Gypsy	Psychological thriller	June 30, 2017	1 season, 10 episodes	46–58 min.	Ended[13]

Plot [edit]

See also: List of House of Cards episodes

Season	Episodes	Originally released
1	13	February 1, 2013
2	13	February 14, 2014
3	13	February 27, 2015
4	13	March 4, 2016
5	13	May 30, 2017
6	8	November 2, 2018

Season 1 (2013) [edit]

Main article: House of Cards (season 1)

Exclusive international distribution [edit]

Title, Genre, Premiere of each season, Length, Language, Distribution

Challenges

Inconsistent column order, variables

- Last columns neither share the same variable nor unique id.
- Premiere Data is only for the first season of series

Spread series to different rows according to its seasons

- Dive into each detail page and crawl the plot table if it's series
- Plot table is not always unique and as the first one

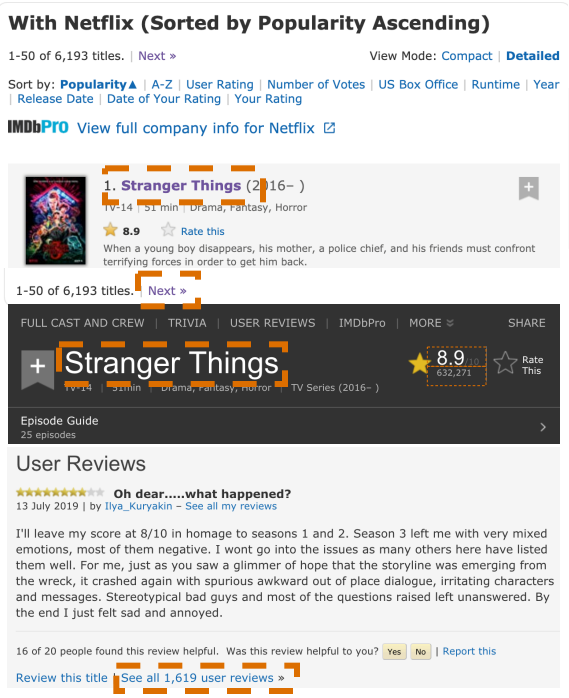
Hard to catch the parent category of tables and lots of wrong and missing values in the result csv

- Lost many variables eventually

This part of Web Scraping consumed me tons of time

Challenges

IMBD



Shows sorted by Netflix as distributor

Title, Number of reviews, Count of rating, Average rating

Challenges

The list includes all the shows distributed by US Netflix

- Not original ones and distribution can not be distinguished
- The title of show might be different in websites so around 1500 rows were lost when merged with dataset from wiki.

Missing values

- When the reviews are not enough the ‘see all __ reviews’ won’t include the numbers.

A glance of datasets

Raw datasets before cleaning and merging

	year	quarter	paid_membership	quarter_revenue
7	2013	4	41.43	1175.0
26	2018	3	130.42	3999.0
4	2013	1	34.24	1024.0

	avg_rating	num_reviews	rating_count	title
248	5.5	13	371	Catching Feelings
2392	7.3	432	180,160	Blue Jasmine
81	5.5	3	1,121	Yucatán

	episodes	genre	language	length	premiere	seasons	status	title
1345	3	Documentary	English\n	NaN	2016-10-25	2	English\n	Tales by Light
987	16	Crime drama	French\n	NaN	2017	2	French\n	Black Spot
190	NaN	NaN	English	23 min.	2015-04-03	NaN	Ended	All Hail King Julien

	distribution	genre	language	length	premiere	title
112	original film	Comedy	English\n	2 hours, 11 min.	April 14, 2017	Sandy Wexler
183	original film	Variety show	\n	1 hour, 3 min.	February 8, 2019	Kevin Hart's Guide to Black History
139	original film	Teen/comedy	English\n	1 hour, 39 min.	June 8, 2018	Alex Strangelove

After cleaning and merging (565 rows)

	avg_rating	genre	language	num_reviews	premiere	rating_count	title	quarter	year	paid_membership	quarter_revenue	year_quarter	month
36	8.8	Political drama	English	708	2017-05-30	433959	House of Cards	2	2017	99.04	2785.0	2017_02	5
329	7.9	Docu-series	English	9	2019-05-31	626	Killer Ratings	2	2019	151.56	4923.0	2019_02	5
187	8.1	Dark fantasy	English	153	2018-10-26	24596	Castlevania	4	2018	139.26	4187.0	2018_04	10
181	8.2	Comedy	English	177	2018-11-16	12133	The Kominsky Method	4	2018	139.26	4187.0	2018_04	11
183	5.7	Superhero	English	40	2018-11-09	1145	Super Drags	4	2018	139.26	4187.0	2018_04	11

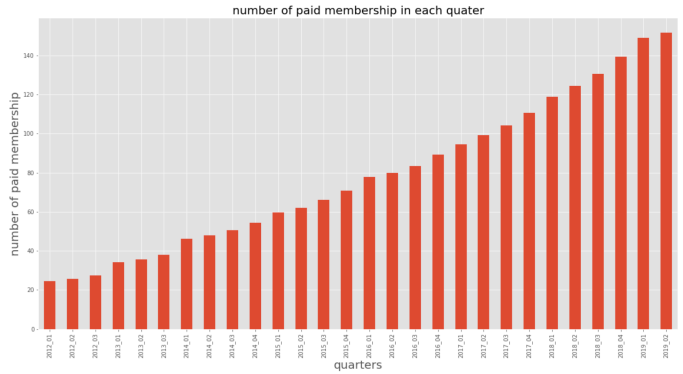
```
infj.describe()
```

	avg_rating	num_reviews	rating_count	quarter	year	paid_membership	quarter_revenue
count	565.000000	565.000000	565.000000	565.000000	565.000000	565.000000	565.000000
mean	7.121770	154.769912	33034.224779	2.442478	2017.269027	112.825735	3379.126903
std	1.164412	200.372057	75695.313481	1.134079	1.481031	31.100952	1082.601753
min	3.300000	1.000000	53.000000	1.000000	2012.000000	24.430000	869.800000
25%	6.300000	25.000000	2124.000000	1.000000	2016.000000	89.090000	2478.000000
50%	7.300000	73.000000	7877.000000	2.000000	2018.000000	118.900000	3701.000000
75%	8.100000	192.000000	28878.000000	4.000000	2018.000000	139.260000	4187.000000
max	9.400000	961.000000	631000.000000	4.000000	2019.000000	151.560000	4923.000000

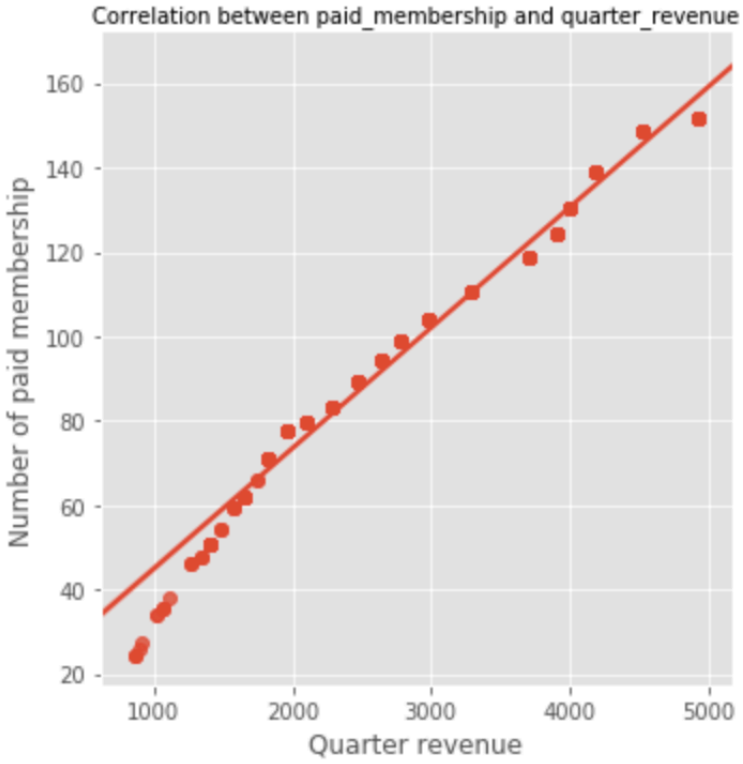
Lost more than 2000 rows and variables of distribution and length

Paid membership does act as a main driver of revenue

Perfect stable growing pattern of quarter paid membership and quarter revenue



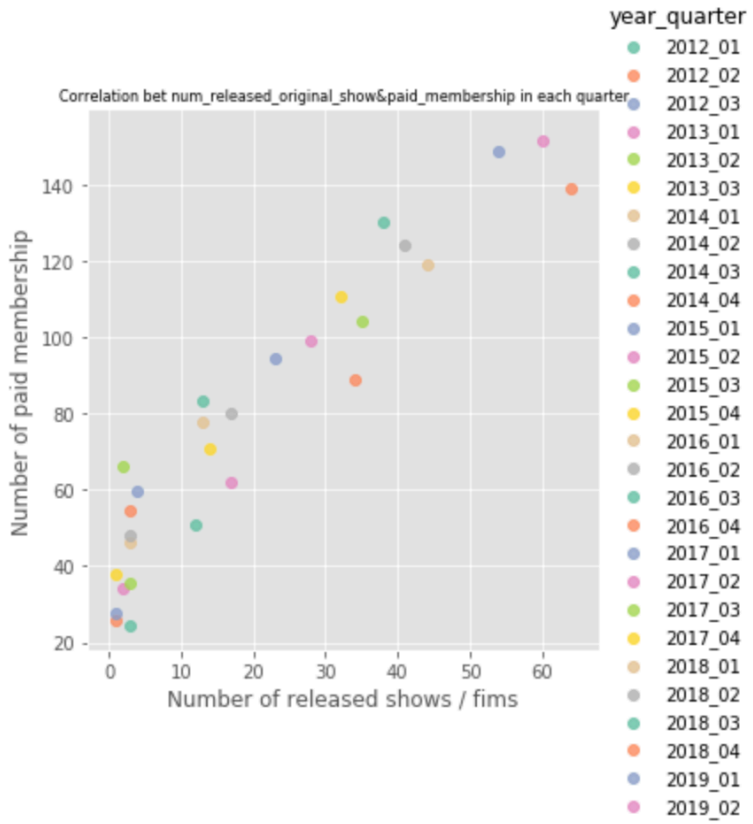
Linear positive linear correlation between them



Proving the previous argument

The positive correlation between the number of released show and paid membership in quarters can be a predictor of revenue

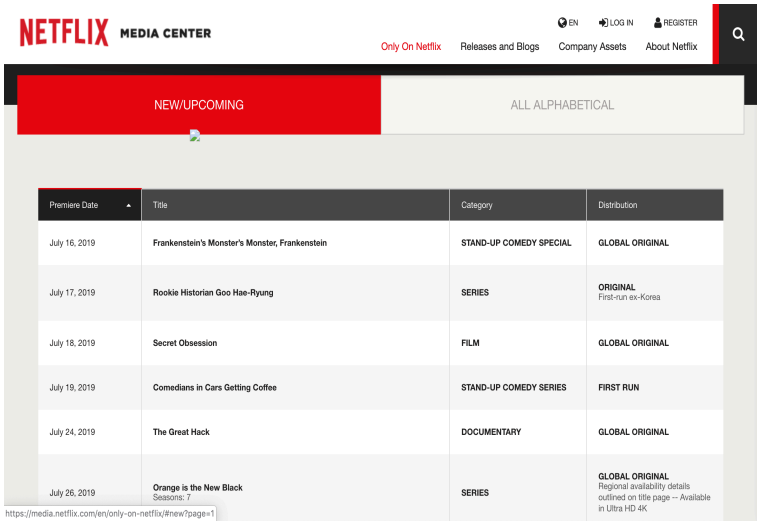
Positive correlation between the number of released show and paid membership in quarters



Significance

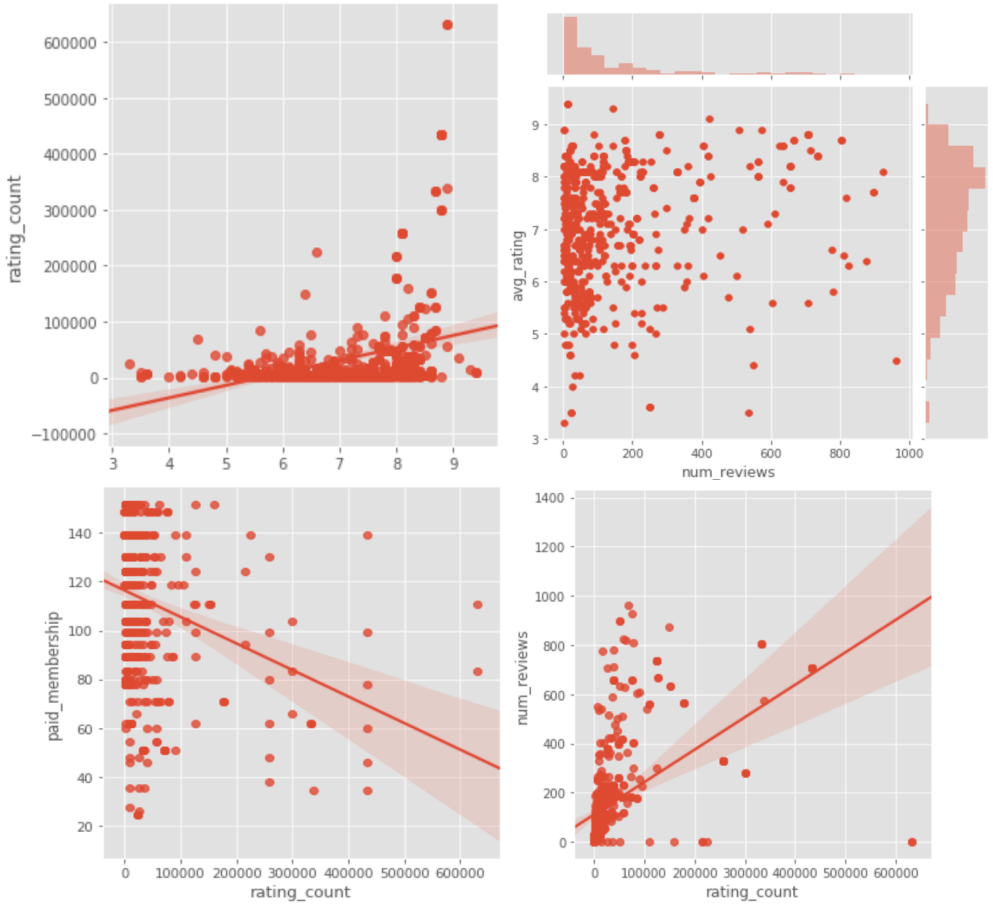
Netflix release upcoming shows at media center

- This number could be a predictor used in the further research building models
- A real practice in Hedge funds



Fail trials

Failures

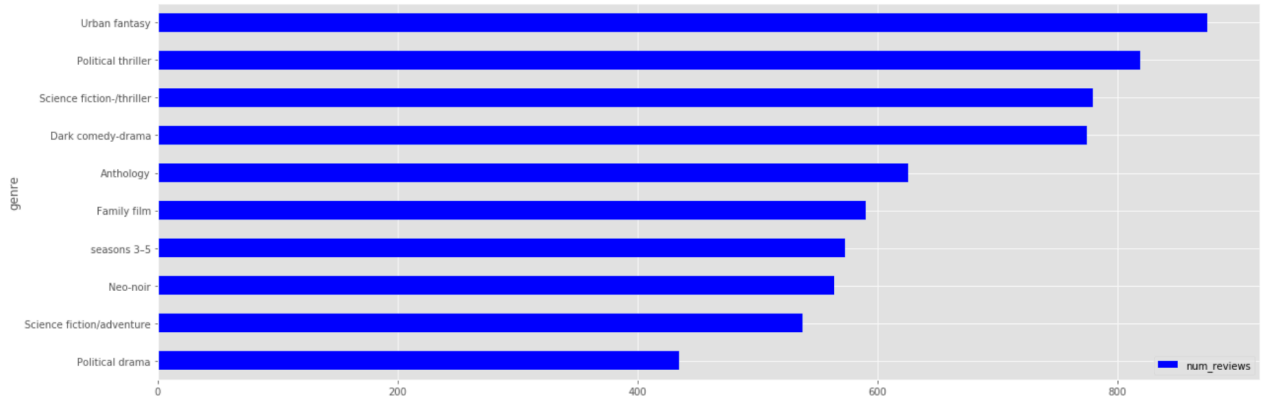


Problems

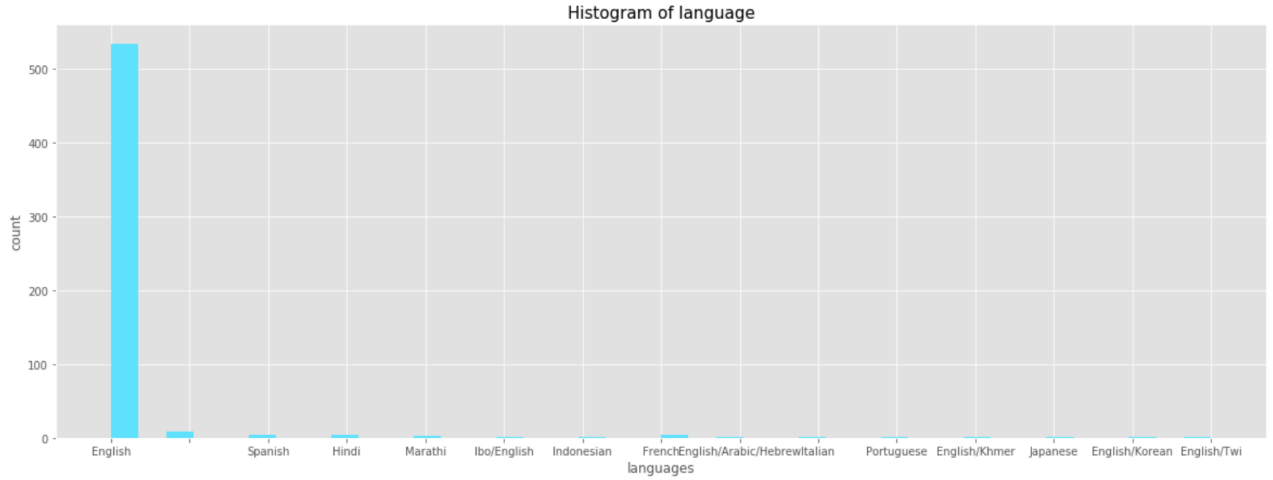
- **Compared variables with different time range**
- **No correlations themselves**
- **Data is incomplete**

Other explorations: popular genres and languages

Urban fantasy, political thriller, and Science fiction/thriller are the most popular genres of Netflix originals

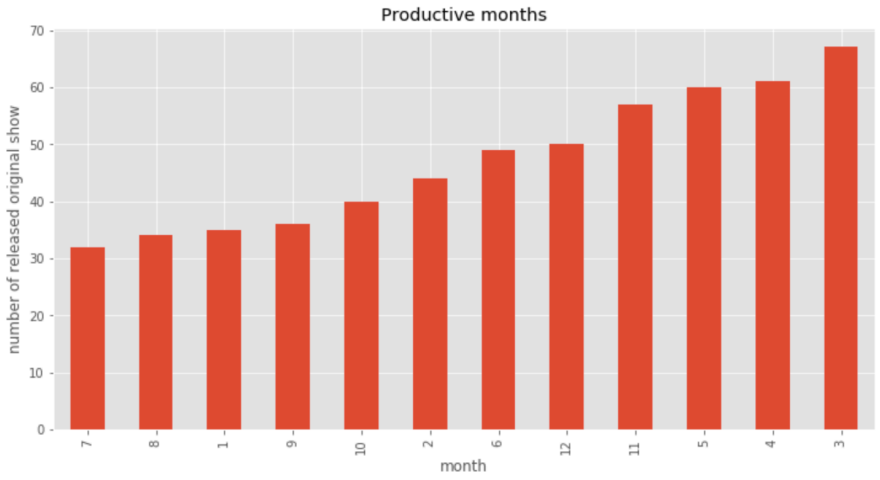


English dominates and is followed by Spanish and Hindi

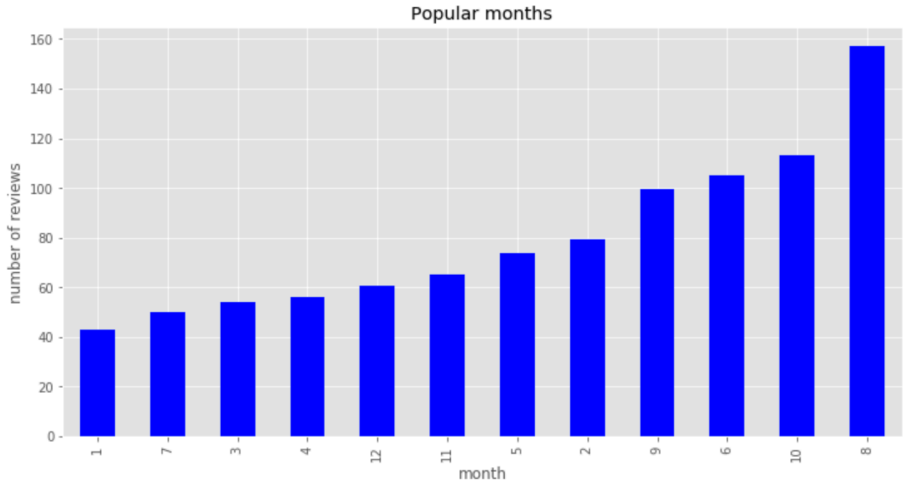


Other explorations: popular and productive months

March, April, and May are the most productive months



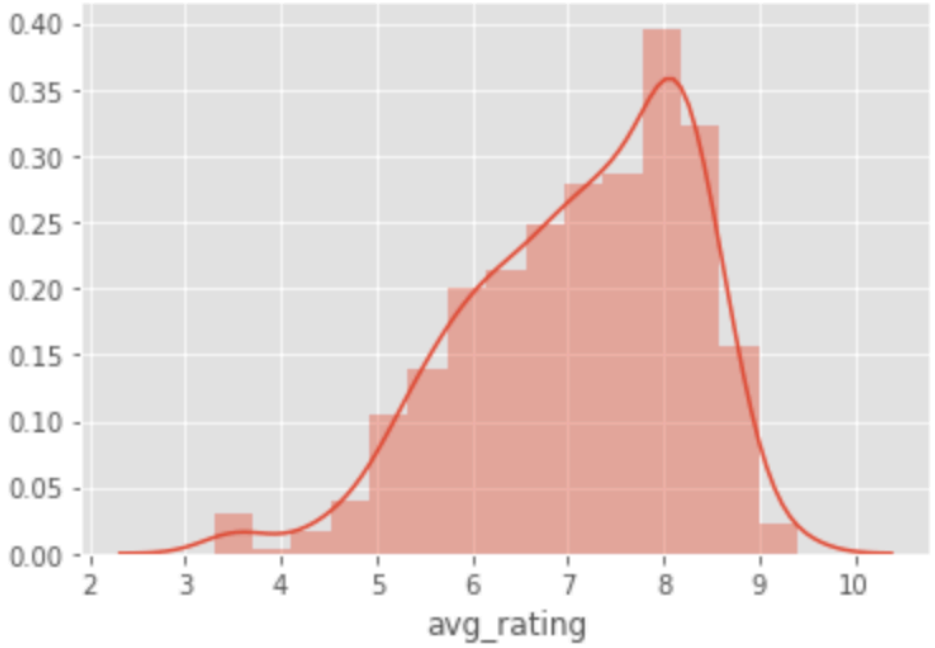
While August, October, and June are the months with most number of reviews



Future work: Find popular and productive months for different genres. Like animation and cartoon should be more popular in vacations of schools, and the reviews should from their parents.

Other explorations: Distribution of the average rating of Netflix originals in IMDB

(Conclusion of the comparison)



Distribution of the average rating in IMBD to compare

Future works

Find more meaningful variables to test the their correlation with paid membership and build up prediction model

Improve the web scraping code to catch more completed datasets

More data analysis