Project 1- Are Streaming Services Killing Movies?

By: Bethany, Rohith, Velindia, Vlad
March 30, 2019

1) Introduction

The overall quality of cinema movies has declined over the past 10-20 years based on viewer and critic ratings. This project evaluates the decline in movie ratings available across all platforms and if it is declining as expected examines what could be some possible reasons for the decline.

Has the quality of movies declined over time?
Quality is determined by ticket sales and movie ratings by viewers and critics. We need to evaluate how many tickets have been sold in theatres by year since 2000 to see if there is a decline as expected. Likewise we need to find how both critics and viewers have rating these movies to determine if the quality has decreased. If we can show that there is a decrease in ticket sales and in reviews, we can conclude that the quality of movies have decreased.

Is the increase in streaming services and other viewing alternatives associated with a decrease in overall cinema quality?
This question will tell us if consumers are choosing alternate entertainment sources, like Netflix and TV programs more frequently and if they choose these forms of entertainment over cinema viewing. This will determine if there is a relationship between the decline in movie quality and increase in quality of other entertainment sources.

What trends are shown between different genres as it relates to quality of movies?
We looked at the percentage of movies for each genre that were released per year as well as the ratings by genre to determine if there are negative trends in specific categories. This would break down the decline in quality of movies in a way that determine which the genres are compared to overall trend.

2) Description of data and technology used

The bulk of the data used for this project was retrieved from the OMDB API through a series of calls.http://www.omdbapi.com/

The imdb titles in the first call was taken from a list of movies posted on kaggle.com and formatted before being placed in a list and called with a for loop that built a dataframe from the data retrieved from the API. After closer examination of the dataset it was it was decided that the data still desired included details on tv series and movies from the year 2018 as the kaggle csv included neither. These titles were downloaded from imdb as a tab separated and gzipped file.(tsv.gz) The gzip and shutil libraries were used to unzip the files and save as csv files so they could be read into jupyter notebook as a dataframe. The imdb file with IDs and votes was combined to complete the list for the second and third API calls.

The OMDB data was cleaned by correcting the formats of both the rotten tomatoes and imdb ratings so they were consistent for easy comparison.

Because most movies had multiple genres stored as a comma separated string, some work needed to be done to narrow down what was included for analysis purposes. Each row was first split and expanded to have each genre on its own column and run through a loop to collect a list of unique genres. The resulting list of genres was narrowed from 23 to 7 to make a succinct analysis possible.

In addition to the above data, aggregated data was retrieved from several sources to aid in answering additional questions that arose during the course of analysis.

For revenue analysis:
Box Office Revenue Source: https://www.the-numbers.com/market/
US Population Source: https://www.multpl.com/united-states-population/table/by-year

For accurate movie releases (North America):
https://www.statista.com/statistics/187122/movie-releases-in-north-america-since-2001/

Alternative Genre Share (for illustrative purposes only):
https://www.the-numbers.com/market/genres

Number of original scripted TV series in the United States from 2009 to 2018
https://www.statista.com/statistics/444870/scripted-primetime-tv-series-number-usa/

See applicable data dictionary for more information on the datasets. (located in cleaned data folder)

3) Methodology

To answer the first question the data was grouped by year and plotted against the mean imdb ratings. The vertical lines represent the approximate years that Netflix began to offer streaming and released their first original series. Ratings were further broken out by genre to determine if any trends could be observed. This was achieved by grouping the cleaned genre data by genre and year and plotting against mean imdb ratings after some minor data munging using unstack and list comprehension to rename the columns. To confirm if there were any large changes in genre ratings, the percentage change between 2000 and 2018 was calculated and visualized with a bar graph.

Though we were most interested in the change in user ratings over the observed time period, a bubble plot was created to see if there is a large difference in user and critic ratings. This was achieved by grouping the movie dataset by year, filtering by genre and plotting mean imdb ratings against mean rotten tomato ratings.

We wanted to determine if movie and TV production has changed over the same time frame. TV series data was next grouped by year and year and tv count plotted to show change over time. Because of the limitations in the OMDB dataset and to ensure accuracy, a pre-aggregated statistic was used to show the Movie releases over the same period. The genre dataset was grouped by genre and year and divided by total per year to produce a change overtime. The percentage change was visualized in a bar graph to check for noticeable shifts.

We then needed to determine if the trend for box office revenue and compare it to the production of movies and tv series to identify possible correlations supporting our initial hypothesis. Box office revenue can easily be visualized by plotting figures adjusted for inflation with years to produce the below line graph. To check that the line was not unduly affected by change in population, a line graph was produced that shows the average tickets purchased per person in the US based on the tickets sold per year and the US population. The graphs show a very similar trend, showing that the decrease in revenue is not caused by changes in the population or change in ticket prices, but is actually declining as a share of the population.

At each phase of above process we had performed normality test and ttest to ensure the data we are working is accurate and is in par for comparison.

4) Results/Output

We were able to get sufficient information to answer our question, as this information is readily available on the internet, though we did turn to previously aggregated data where noted when the raw dataset we had access to was insufficient to answer the questions. Although we saw a drop in movie ratings and a rapid growth of online streaming services such as Netflix in the last decade, there was not enough evidence to conclude that the growth of streaming service is the reason for this decline.

We observed that movies ratings have been steadily declining since 2007 while TV series ratings have been consistently increasing over the same period. We further observed that users and critics tend to agree on the quality of movies.

We also noted that the average number of movie tickets sold per person and annual box office revenue have been decreasing since 2003 with a small rebound since 2011.

Finally, we observed the shift in genres popularity over the last 20 years. We noted large increase in Action, Adventure and horror movies produced and large decrease in Dramas and Comedies. Ratings for genres remained relatively steady.

Although, the data shows a decline in quality of movies over time, we recommend that this not deter individuals from going to see movies. Movies are about having fun and being entertained, so you should see whatever you like. We would encourage everyone to try an action movie as they have been increasing in supply and many people are likely to enjoy the content.

5) Limitations of data and analysis

Initially, there was difficulty gathering information. Most sites required a fee to access data or limited how much data that could be accessed for free, thus restricting our options for data sources.

We discussed how to best analyze movie genres when a single movie could be categorized as 10 different genres. We went with a method that would counts movies with multiple genres more than once which does have a possibility to skew the data but provides some insight on trends.

The OMDB dataset presented several challenges. The rotten tomatoes ratings were received in an inconsistent format and required a unique solution though in the end it was clean and usable. The Metacritic ratings and box office series were both missing enough values that they were was not usable for analysis.

If starting this project again, it is likely that collection of data would have been approached differently. By starting with a pre-curated list from kaggle instead of going straight to the imdb text files the list of movies was not as pure as it could have been.

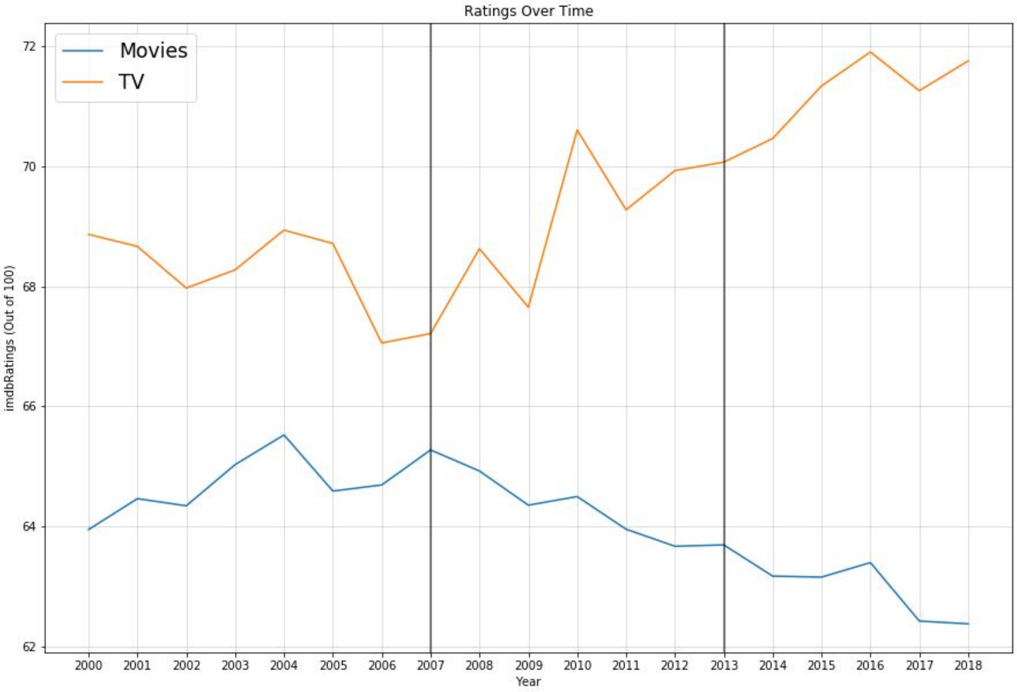
With more time we could have taken a similar approach to the production houses that we took with the genres and reviewed trends by producer. Additionally, with a dataset that included information such as budget and box office it would be possible to determine which genres and even plot points are most profitable.

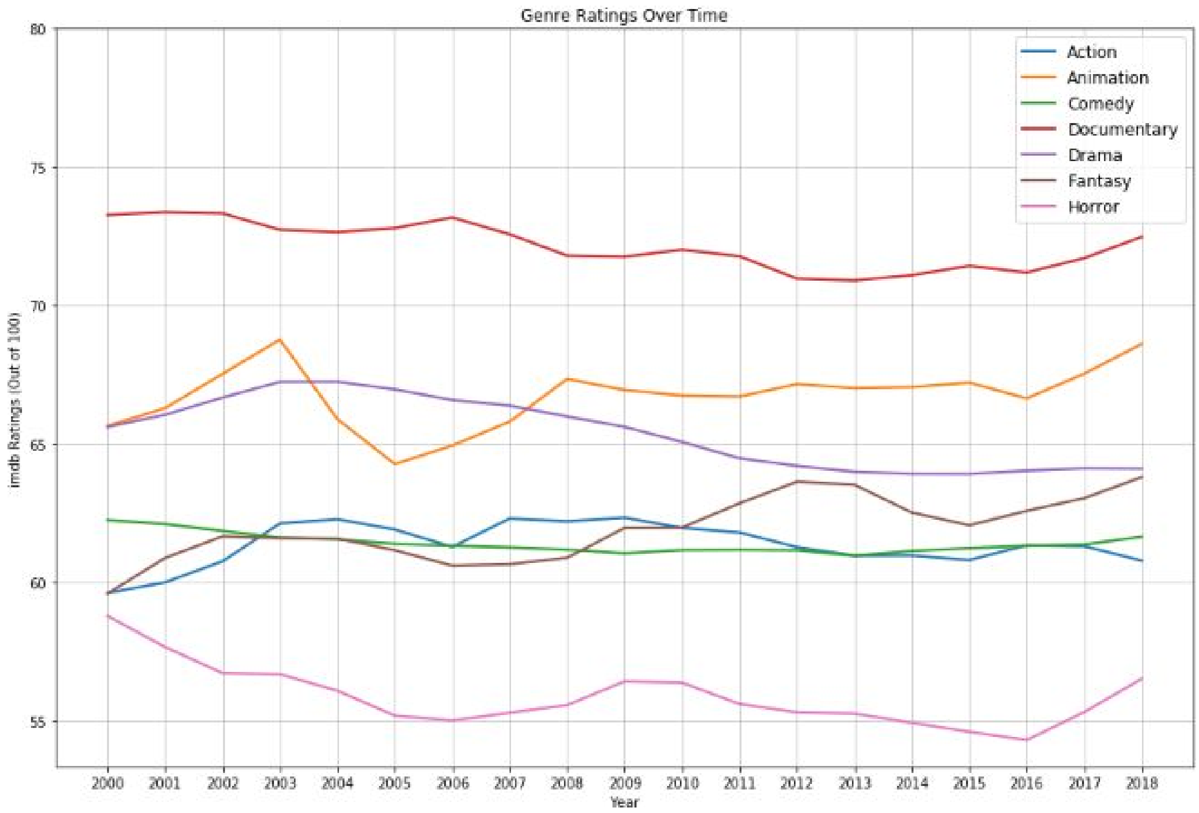
6) Conclusions/Recommendations

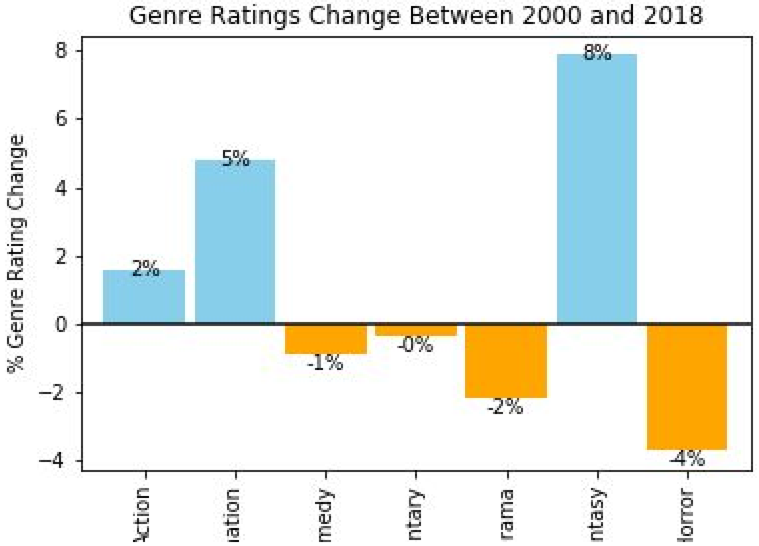
We expected to find a clear decline in movie quality and a correlation between the decline and the rise of other entertainment sources such as Netflix. However, this is not what we found. Although we saw a drop in movie ratings and a rapid growth of online streaming services such as Netflix in the last decade, there was not enough evidence to conclude that the growth of streaming service is the reason for this decline.

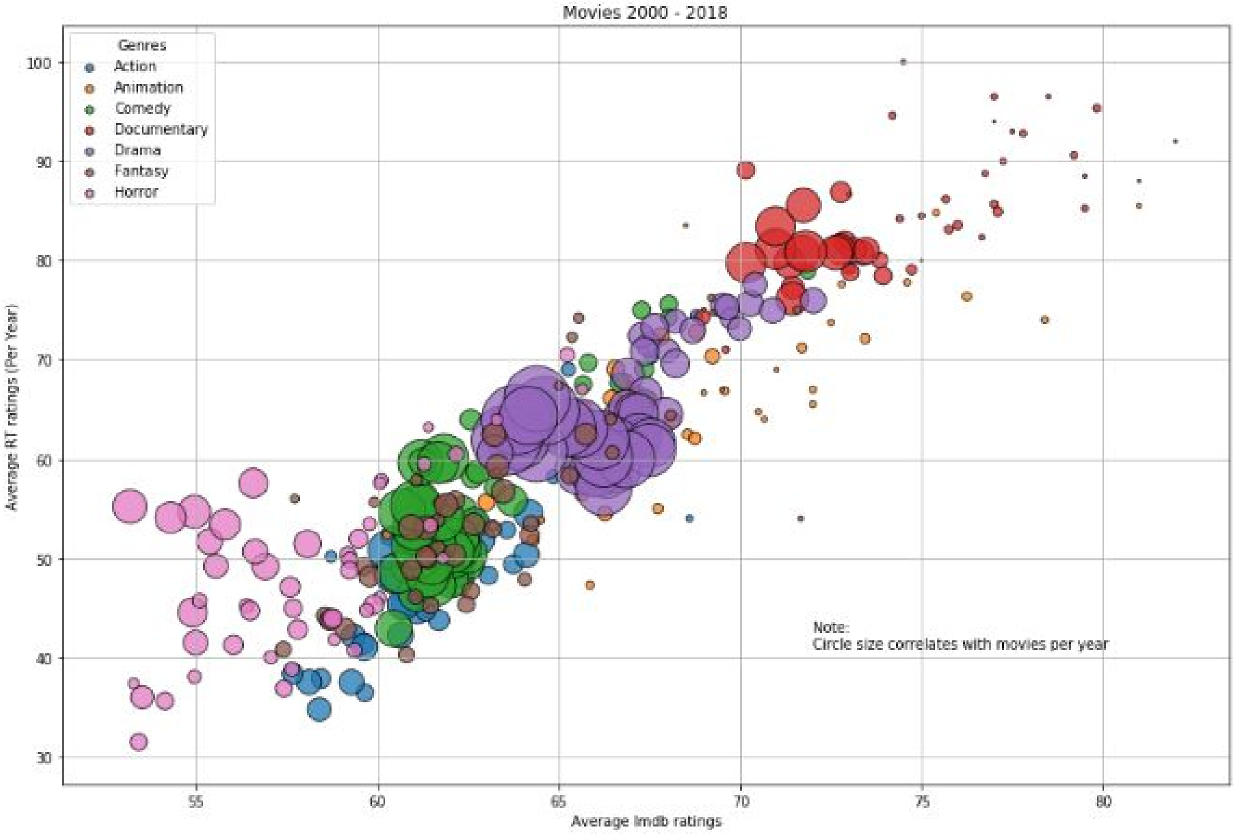
We observed the shift in genres popularity over the last 20 years. We noted large increase in Action and Adventure movies produced and large decrease in Dramas and Comedies. We concluded that the shift in genres popularity may be the reason for perceived drop in quality of movies.

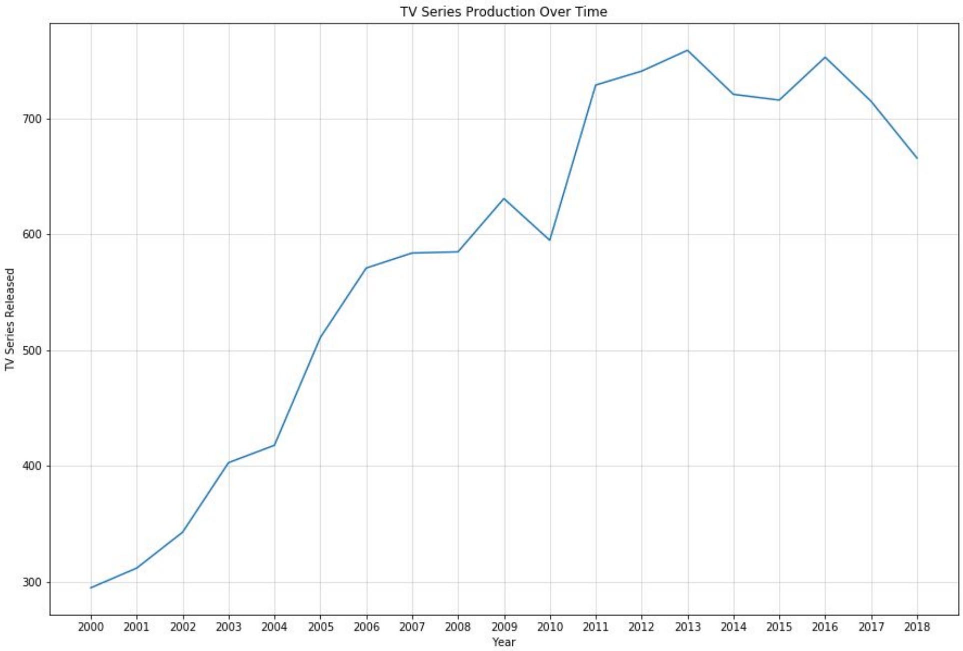
7) Appendix of All Vizzes

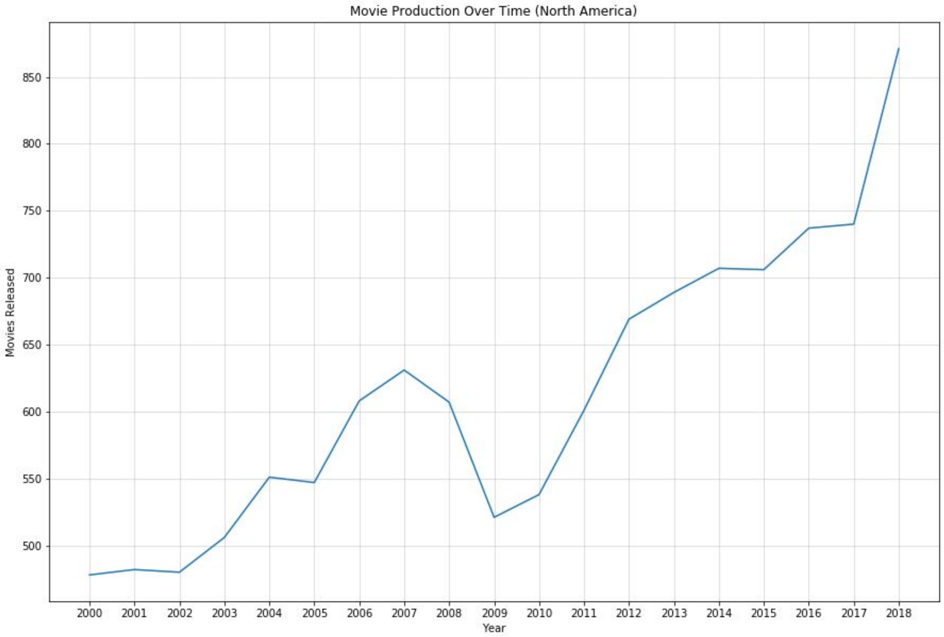


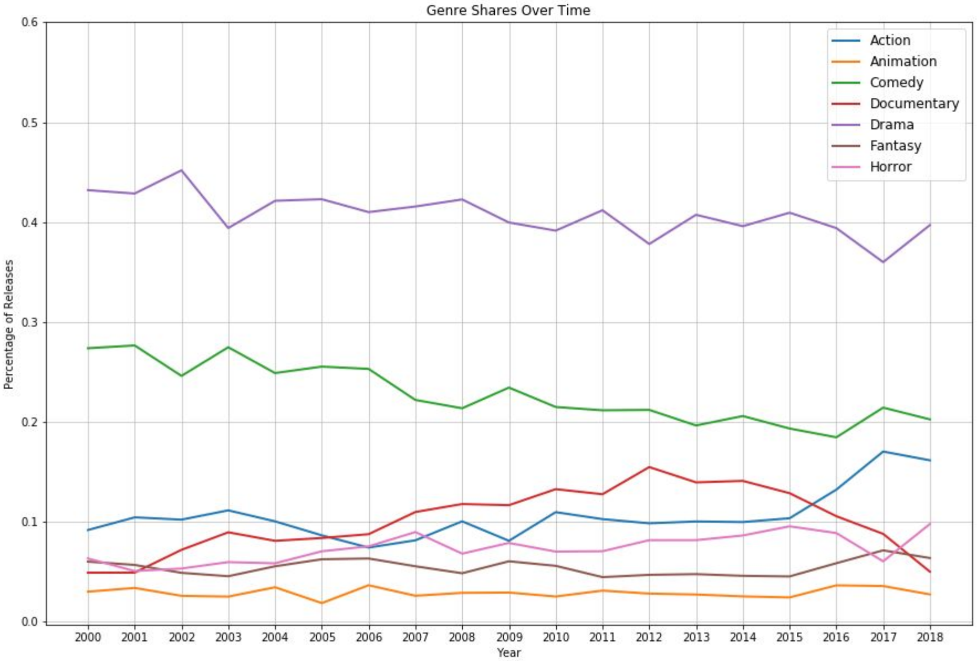

Genre Ratings Over Time
80
75
70
65
60
55
imdb Ratings (Out of 100)
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Year
Action
Animation
Comedy
Documentary
Drama
Fantasy
Horror

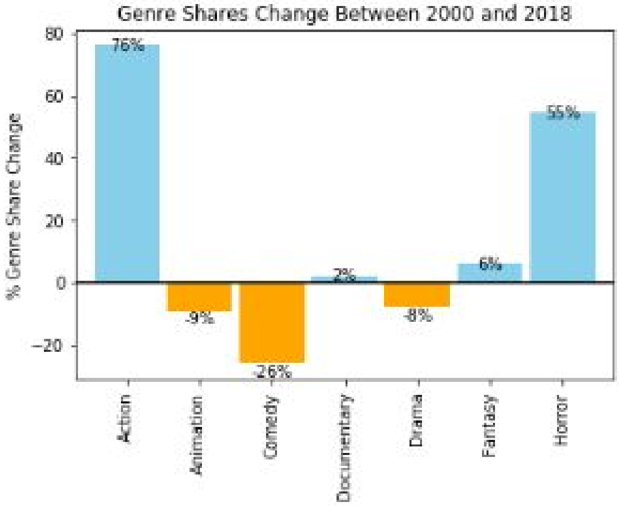

Genre Ratings Change Between 2000 and 2018
8
6
4
2
0
−2
−4
% Genre Rating Change
2%
5%
-1%
-0%
-2%
8%
-4%
ction
ation
medy
ntary
rama
ntasy
lorror

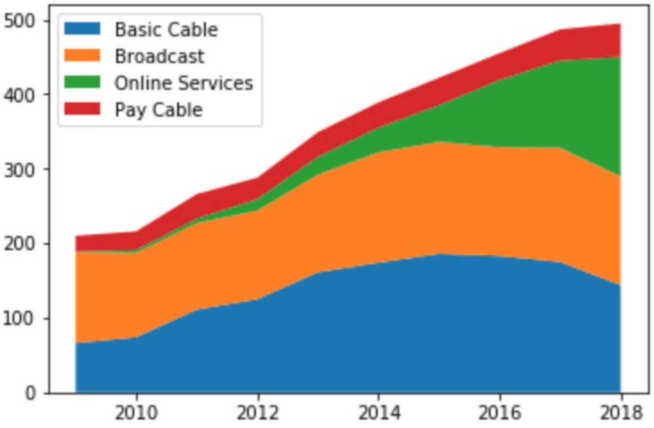

Movies 2000 - 2018
Genres
Action
Animation
Comedy
Documentary
Drama
Fantasy
Horror
Average RT ratings (Per Year)
100
90
80
70
60
50
40
30
55
60
65
70
75
80
Average Imdb ratings
Note:
Circle size correlates with movies per year

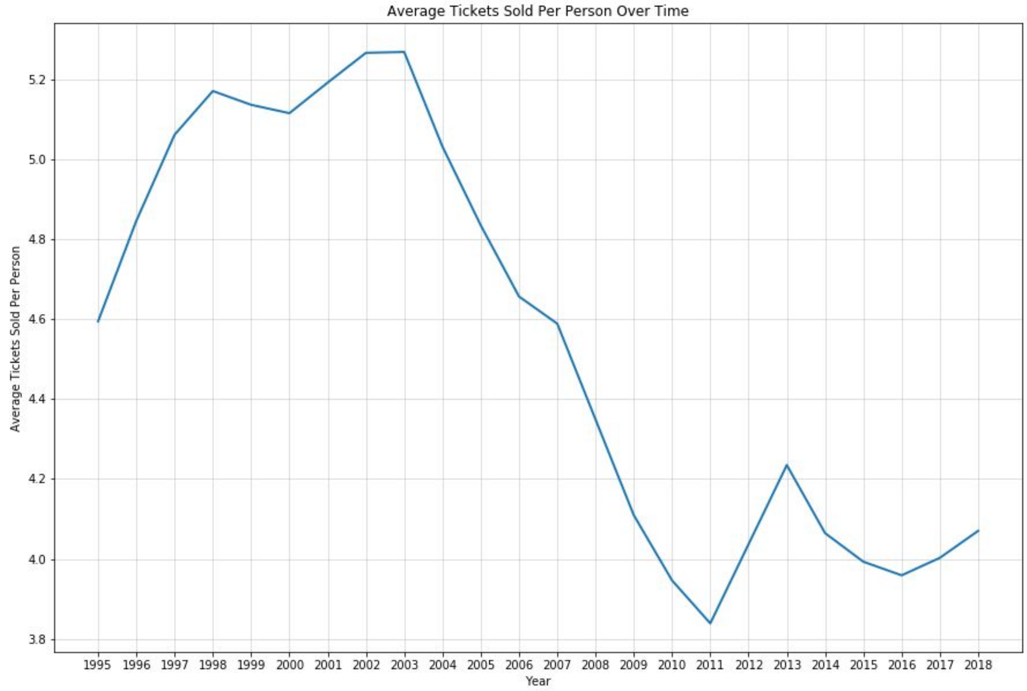

TV Series Production Over Time
TV Series Released
300
400
500
600
700
Year
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018


Movie Production Over Time (North America)
Movies Released
500
550
600
650
700
750
800
850
Year
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018

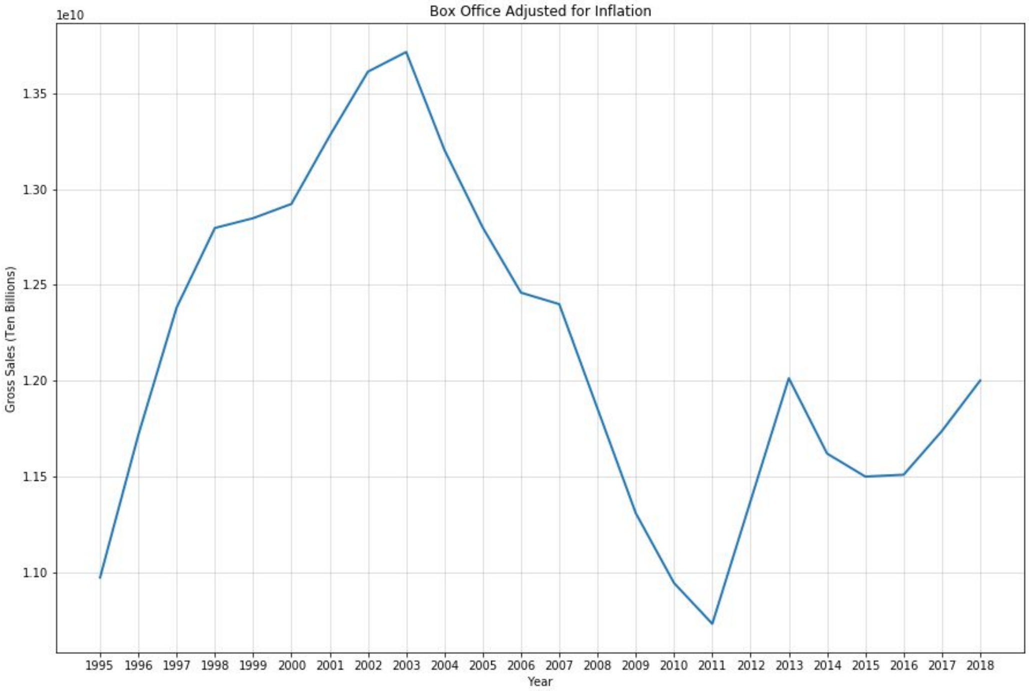

Genre Shares Over Time
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Percentage of Releases
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Year
Action
Animation
Comedy
Documentary
Drama
Fantasy
Horror


Genre Shares Change Between 2000 and 2018
80
60
40
20
0
−20
% Genre Share Change
76%
-9%
-26%
2%
-8%
6%
55%
Action
Animation
Comedy
Documentary
Drama
Fantasy
Horror

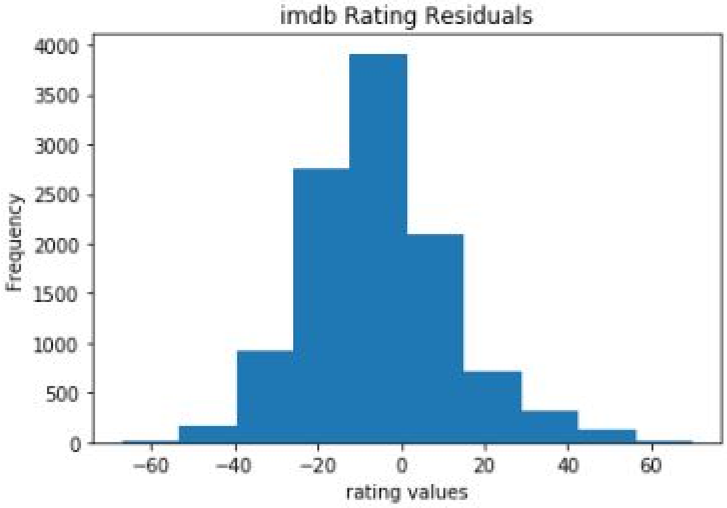

Basic Cable
Broadcast
Online Services
Pay Cable
500
400
300
200
100
0
2010
2012
2014
2016
2018



Average Tickets Sold Per Person Over Time
Average Tickets Sold Per Person
5.2
5.0
4.8
4.6
4.4
4.2
4.0
3.8
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Year



Box Office Adjusted for Inflation
1e10
Gross Sales (Ten Billions)
1.35
1.30
1.25
1.20
1.15
1.10
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Year


imdb Rating Residuals
Frequency
4000
3500
3000
2500
2000
1500
1000
500
0
−60
−40
−20
0
20
40
60
rating values