Mini Project

UCSanDiegoX: DSE200x

Python for Data Science

Arvind Karir – Mar 27-2018

Datasets and background

I am new to Data Science and this is my first course/project. I wanted to explore the relationship between various world indicators but found that exercise a bit challenging. For example, I saw some numbers for India which showed an increasing trend in the land area under forests, which seemed counter-intuitive since my regular visits and travels in the country indicate increasing stress on the environment. I decided to explore further and found that there is no one way to measure the land covered by forests; for example every bit of green in satellite image is not forest – it could be farmland, plantations or forests. And their intended use is also unknown – plantations can be harvested for trees when they are ready to be harvested. Plantations also suffer from the lack of bio-diversity typically found in forests; while they may have trees, they are not the best in terms of nature's healing process. Soil Quality Index is a good indicator of the forest and is measured by a Spectral method which captures a spectrum of radiations and not just RGB. I felt that this is too advanced for me at this stage and decided not to pursue it.

I explored world indicators to some degree but more in nature of exploration.

Movielens database was explored in more detail.

Motivation

- I am motivated to learn Data Science. I started learning Python four months ago and then found this excellent course on EdX by UCSD.

- My motivation is to learn and explore based on the teaching imparted in this course. There was no world-problem as such that I wanted to solve, but I wanted to get my hands dirty and gather and communicate the ‘insights’. I will leave it to the community at large to provide feedback and evaluate my work.

Research Questions

I have chosen to explore the Movielens database to see if it can provide insights into our times and if and how movies represent not just our culture, but our economic situation, our political stance, and their relationship to life in general.

Do the narratives explore our problems or cultural anxieties?

What do the genres tell us about our times?

Is there a relationship between genres?

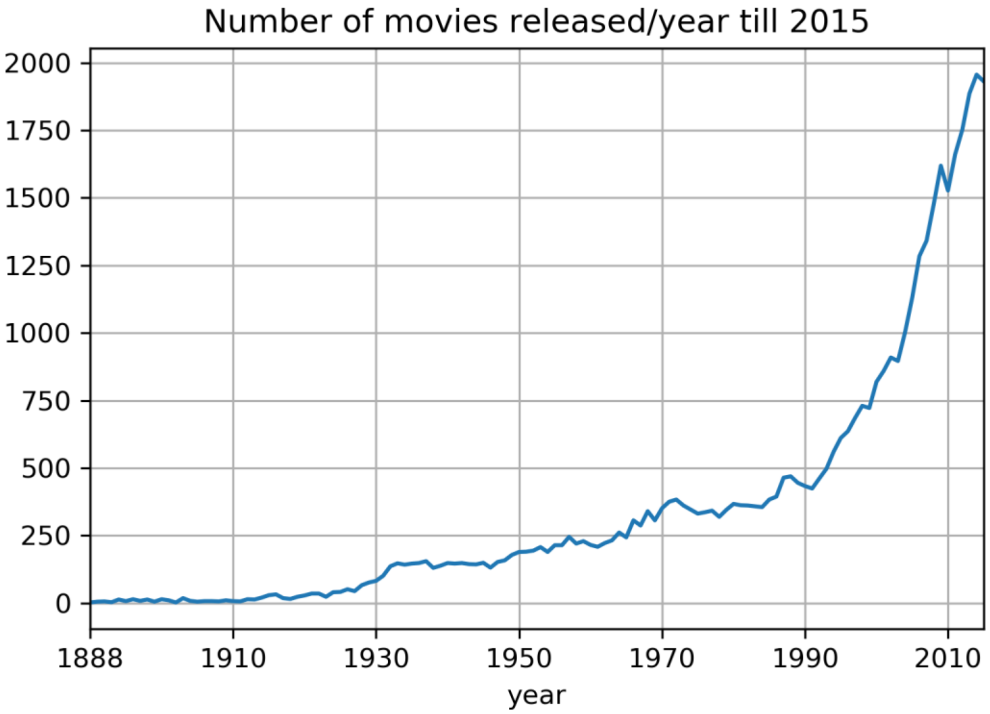
How do movies as an industry relate to our economic prosperity?

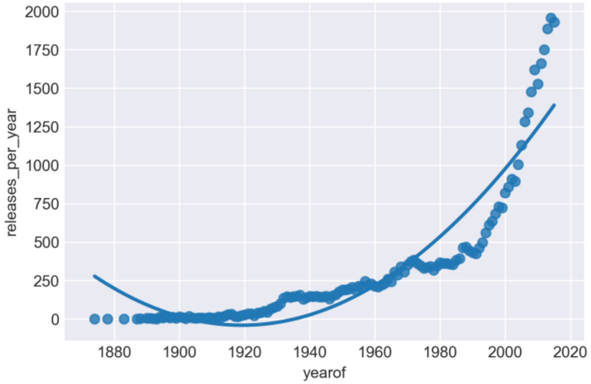
What can we learn from history?

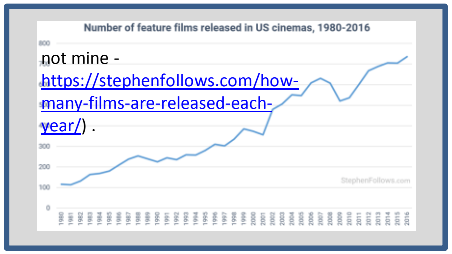
Findings

Following slides present some of the findings

Visualizations - I







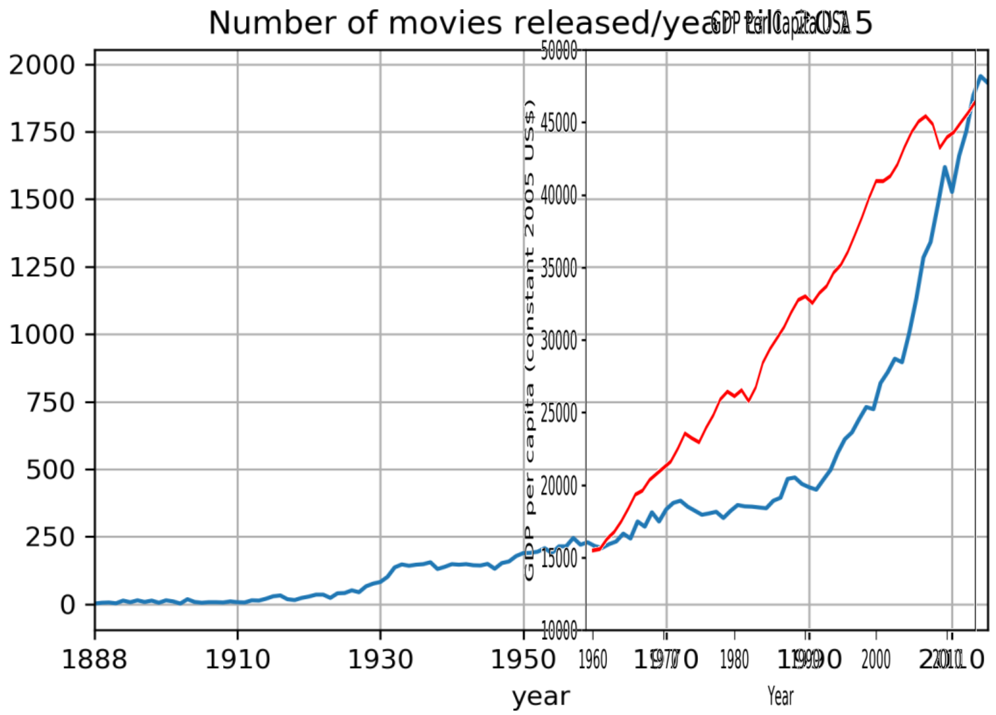
not mine - https://stephenfollows.com/how-many-films-are-released-each-year/) .

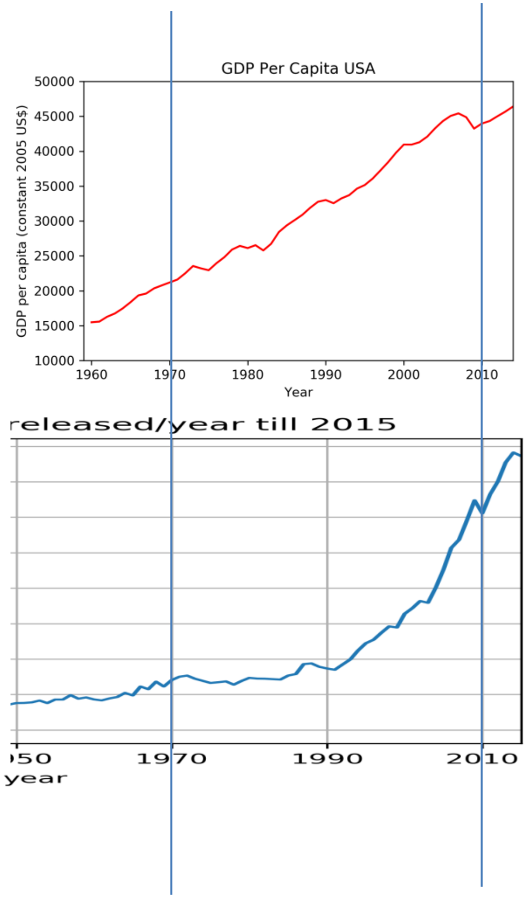
2016-2017 data was deleted (perhaps incomplete) as it showed a decline in number of movies released (corroborated by info in graph in gray box).

There is a dip in 2010 possibly as an impact of the economic downturn in 2008, which would indicate that the time gap between financing and release of movie is over a year.

It appears that the graph suddenly jumps around 1990, it changes its shape. I did a scatter diagram and plotted the line of best fit using a second order polynomial. Possible explanation in https://en.wikipedia.org/wiki/1990s_in_film. I used seaborn to plot the scatter diagram and line of best fit and had to manually clean the data in excel as df.drop command was not working.

Visualizations – I, contd.





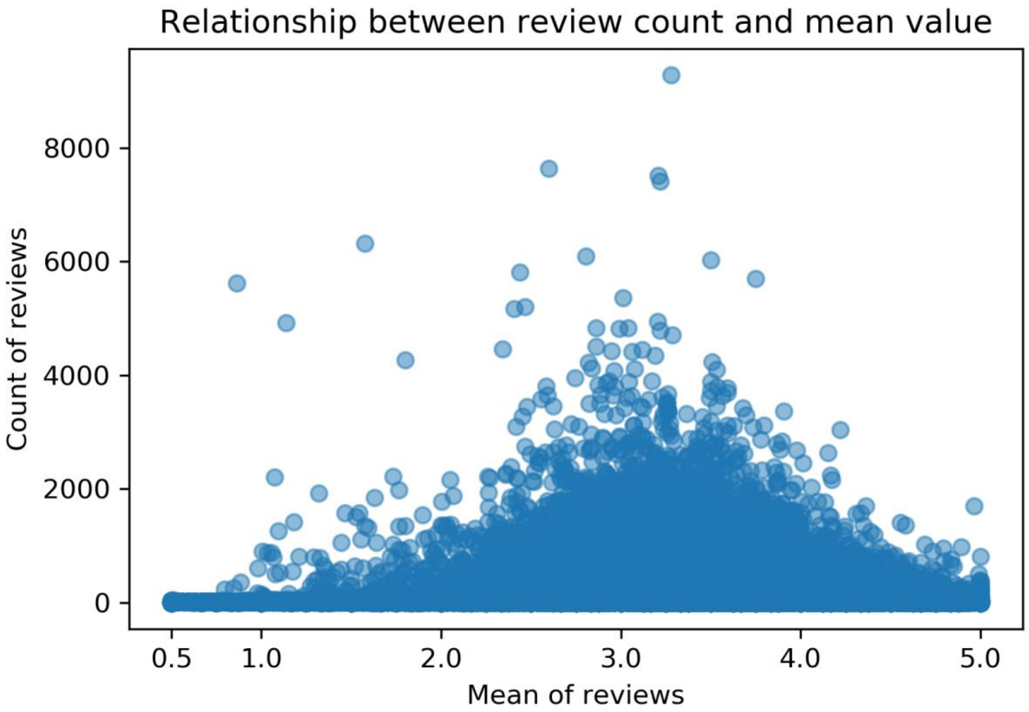
Plot between US GDP – red line and number of movies released/year.

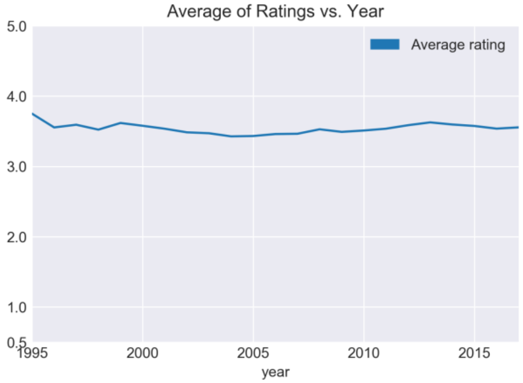
- GDP from world indicators dataset
- Movies from Movielens dataset

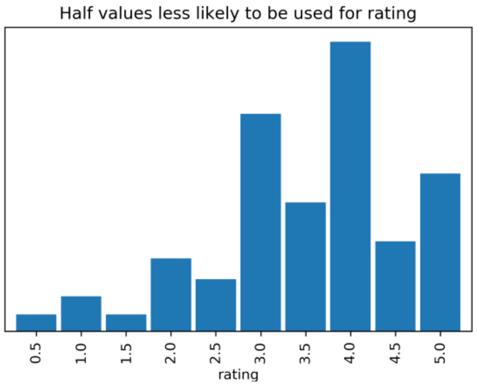
Impact of the economic downturn of 2008 is pretty evident.

Also, that there are other factors besides GDP which caused a different slope in graph of movies after 1990.

Visualizations - II


Relationship between review count and mean value


Average of Ratings vs. Year

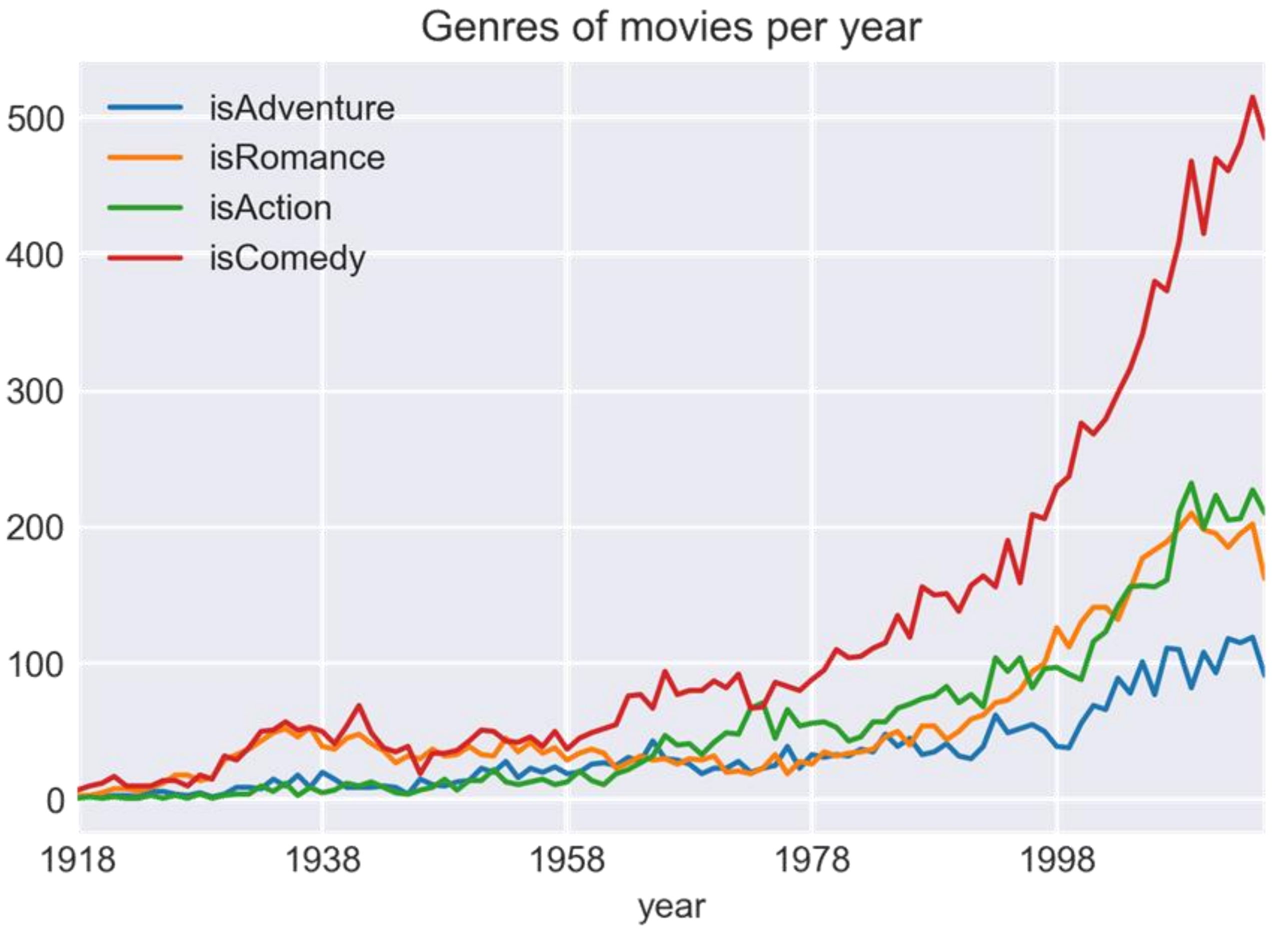

Half values less likely to be used for rating

The scatter plot supports the theory of ‘Mean Reversion’.

In general ‘round number’ ratings (like 1, 2, 3, 4, 5) are much greater than ‘half numbers ‘ (0.5, 1.5. 2.5, 3.5, 4.5), which

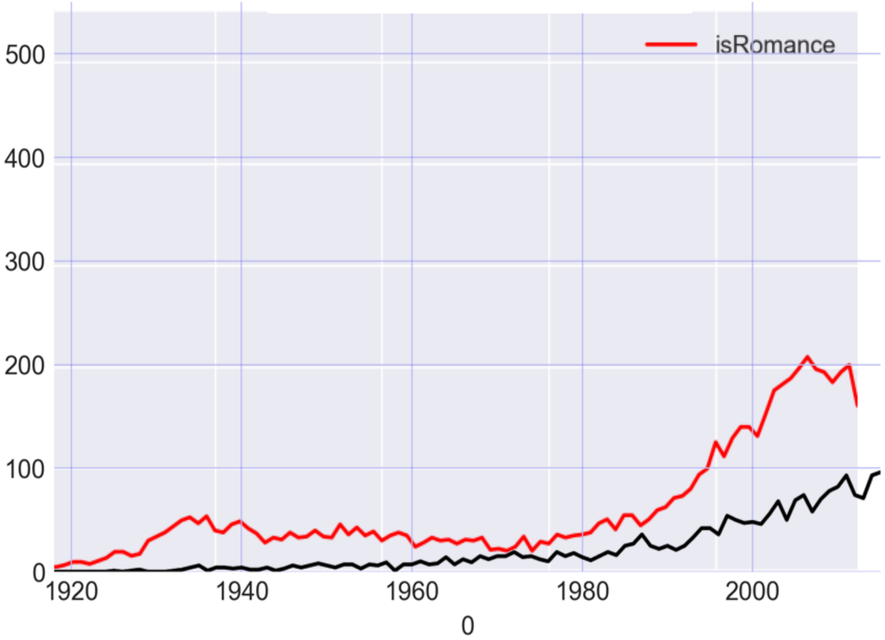
indicates that reviewers find it easier to distinguish between (say) , ★★ vs. ★★⯨

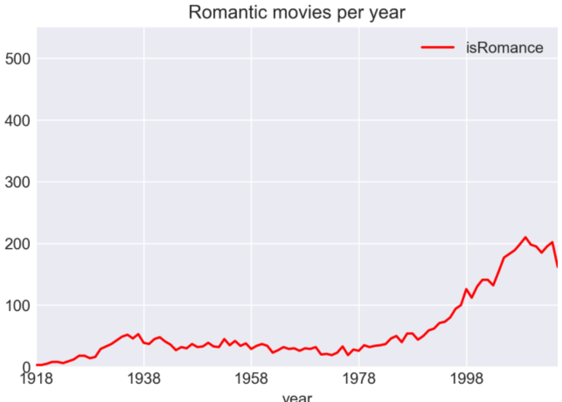
Visualizations - III

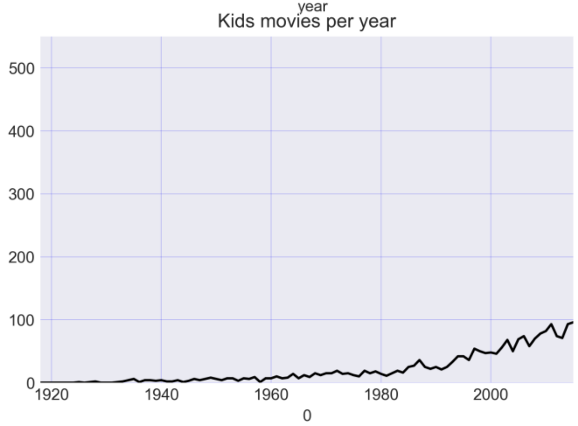


This graph shows the relationship of some of the genres with time. Noticeable jump in Romance and Action around 2001 timeframe.

Visualizations - IV

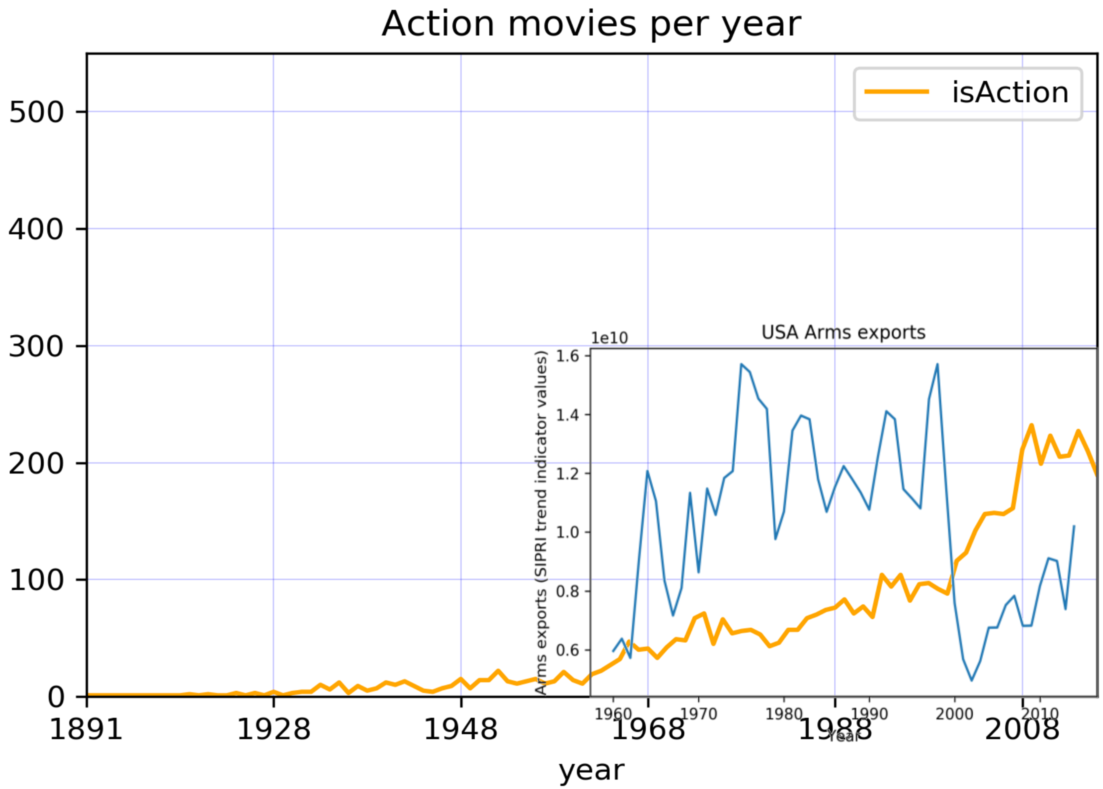






Do the number of romantic movies increase the likelihood of romance and hence more children (used proxy of kids movies) – maybe!
There was a spike in romantic movies around the 1930's depression and the reasons could be -
https://www.independent.co.uk/arts-entertainment/films/features/the-rom-com-a-new-look-of-love-1781324.html
https://www.independent.ie/entertainment/movies/how-the-great-depression-inspired-hollywoods-golden-age-26481978.html
If so, did history repeat itself post '9/11' or in 2008?
(There was something not right in plotting multiple genres so I super-imposed the plots in PowerPoint)

Visualizations - V



Is there any correlation between US Arms Exports and number of action films released? Post-2000 the trend appears to be following. Somehow the dips and plateaus seem correlated.

Challenges

I faced the following challenges, and I am sure most of them can be attributed to my inexperience, but thought I will share in any case.

- 1. Jupyter requires all the cells to be run from very beginning on startup, and in some cases the data has to be re-read for the commands to work.
- 2. Once I delete something and run the cell again, it causes error, of course! So I am now very careful when deleting the columns in a data frame.
- 3. To save on reading time (from csv file), I decided to use Pickle and generated the pickle file after reading the data once. Pickle files are generally smaller in size than csv file, and faster to read, except in case of text files. For instance ratings.csv file is 692 Mb but ratings.pickle file is 813 Mb. And they sometimes gave errors – I am sure it is my inexperience.
- Matplotlib is not very intuitive, and the documentation/examples was not easy to follow along. Again, I think it is because of my inexperience. I use Stack Overflow a lot to look up how people have solved their coding problems.
- In some cases the data is read as strings and then assigned as index (‘years’ for example). I found it hard to deal with in some cases and dropna did not do the trick either. And in some cases years had other non-numeric values which caused errors in plotting
- Kids movies data had a couple of row indices like ‘2009-’ and ‘2006-2007’ which I found impossible to replace or drop