**MOVIE RATING ANALYSIS**

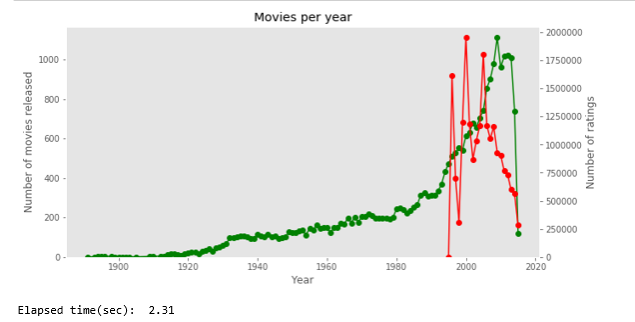
**This is just a basic (but decently comprehensive) data exploration exercise, trying to cover as much Pandas DataFrames and Matplotlib crunching as possible, all for my own learning process.**

**My first contact with this dataset is from an online course in EDX (UCSanDiegoX: DSE200x Python for Data Science), and comes to show how many questions and insights can be derived from very basic information (and I've only used 2 of the 4 data files available).**

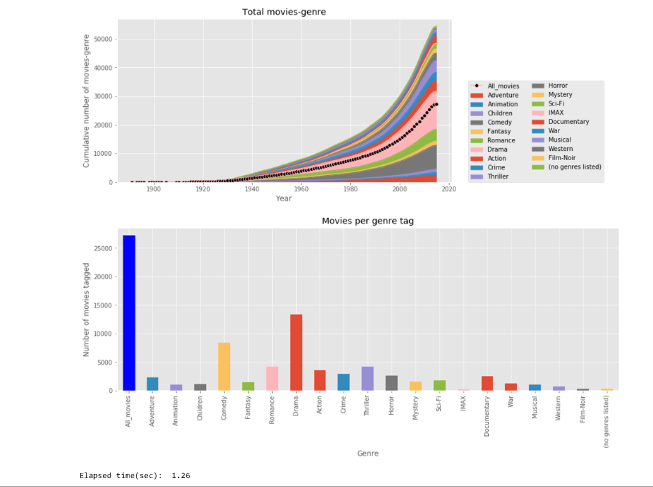
**Along the way I discovered another important learning objective, as my laptop (8Gb RAM) was not able to properly handle the 20MM rows of the ratings data when combined with the movies data (expanding the number of columns significanty), so I had to apply even more Pandas tricks to filter and combine the two data files on the fly, and also continually doing some housekeeping on the active namespace deleting unnecesary variables when they were no longer required.**

## **Explore the data with some basic plots**

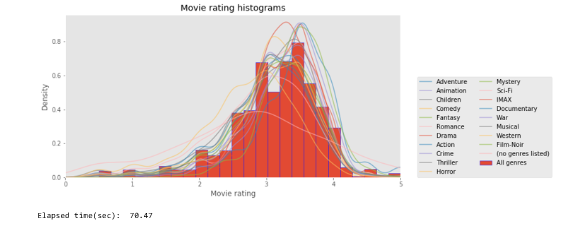
**\*\*PLOT#1\*\*: Number of movies and ratings per year.   
\*\*INSIGHT#1\*\*: Number of movies released per year increasing almost exponentially until 2009, then flattening and dropping signifincantly in 2014 (2015 data is incomplete). Does this confirm expontential growth (i.e. bubbles) is seldom sustainable in the long term? No ratings before 1995, likely to do with the availability of Internet to general public.**

****

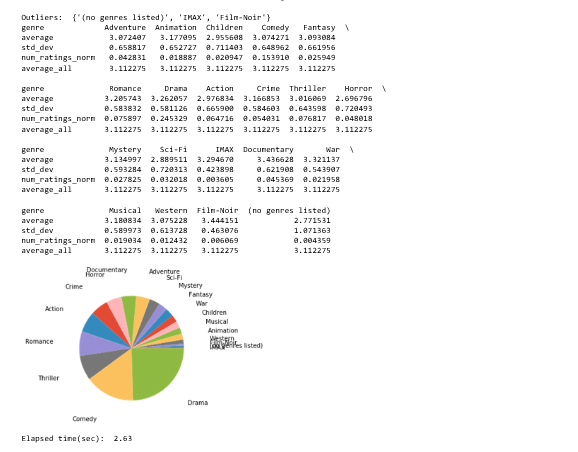
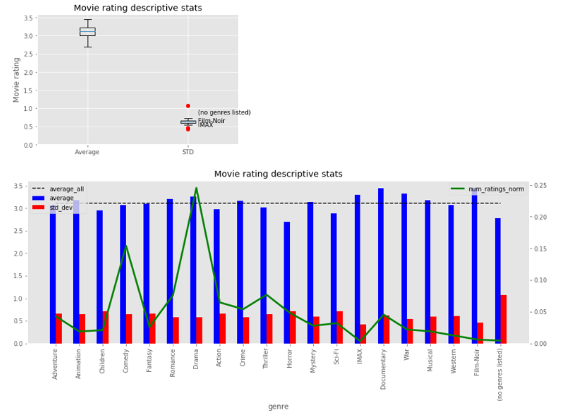
**\*\*PLOT#2\*\*: Cumulative number of movies, in total and per genre.   
\*\*INSIGHT#2\*\*: On average, movies are categorized into 2 genres (i.e. number of movies-genres 54k doubles the number of movies 27k). Comedy 8.3k and Drama 13.3k are the top genres used.**

****

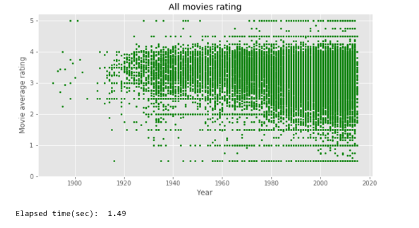
**\*\*PLOT#3\*\*: Distributions by genre, on top of total rating distribution. This will help identifying consitent ratings or outliers (e.g. Comedies being rated higher in general).   
\*\*INSIGHT#3\*\*: All genres show a similar pattern (right-skewed log-normal distribution??), except perhaps Horror movies which are a bit skewed to the left (poorer ratings)...people don't like being scared, no matter how good the movie is fro a technical point of view? Movies without a tagged genre (no-genres listed) are also outliers, but likely due to the low number of ocurrences.**

****

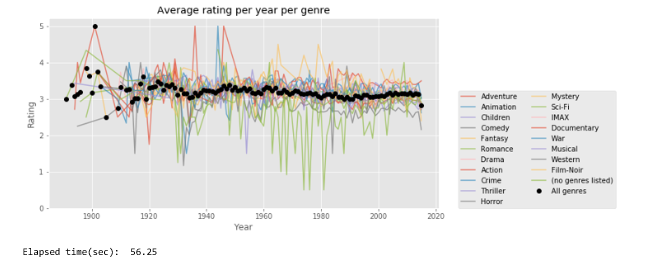
**\*\*PLOT#4\*\*: Compute basic statistics (avg, std) per genre. Plot dispersion (box plot).   
\*\*INSIGHT#4\*\*: As highlighted in the prevous plot, all genres really show similar behaviour. Using numbers in this case, mean rating is sligtly above 3, with a standard deviation around 0.65.**

****

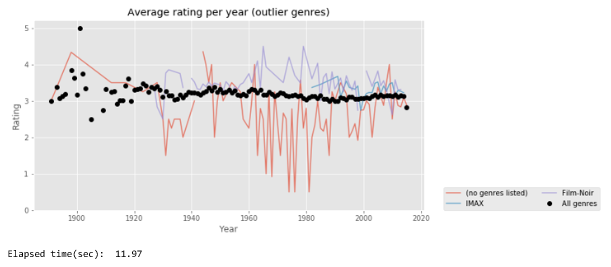
**\*\*PLOT#5\*\*: Average rating for all individual movies.   
\*\*INSIGHT#5\*\*: Especially after 1970, it seems there are more lower ratings, but also more higher (4.5-5.0)...it could just an effect of having more movies. Not many insights from this plot.**

****

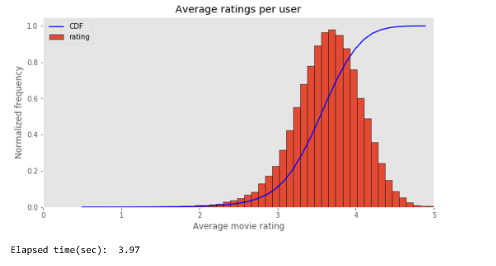
**\*\*PLOT#6\*\*: Average rating for all movies in each year, and also per genre.   
\*\*INSIGHT#6\*\*: Slight decline in average movie ratings after 1960, but still remains above 3. Range quite narrow, except for a few outliers.**

****

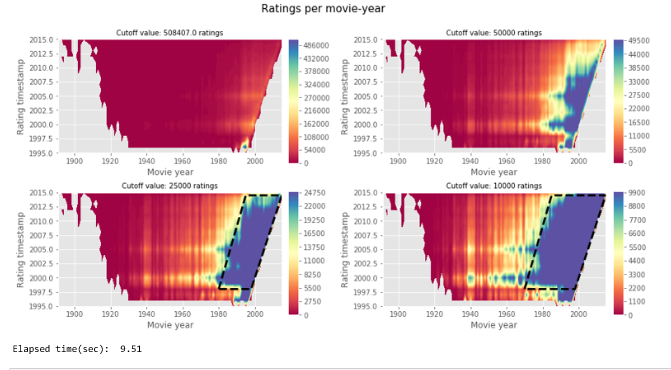
**\*\*PLOT#7\*\*: Same as #6, but only the outliers now.   
\*\*INSIGHT#7\*\*: All the outliers present vary few occurrences: IMAX 196, Film-Noir 330, no-genre 237. In any case, Film-Noir movies are generally rated well above average, and the ones without any classification tend to do very poorly...perhaps having no genre to anchor or relate to impairs critics?.**

****

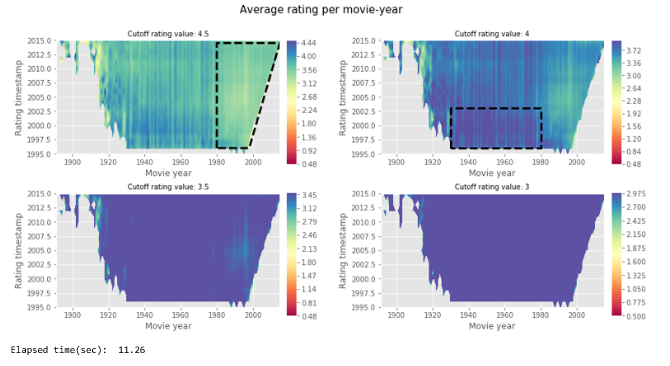
**\*\*PLOT#8\*\*: Average ratings per user.   
\*\*INSIGHT#8\*\*: Users have a positive bias in general, with roughly 95% of their average ratings above the mid-point of 2.5. This is to be expected, and could have many explanations: users actually watch the better movies due to available ratings (and this should get better over time, as the rating system expands); users don't bother that much to rate bad movies as they do with the good ones (i.e. we don't want other to know we watched such a piece of s\*\*\*), etc.**

****

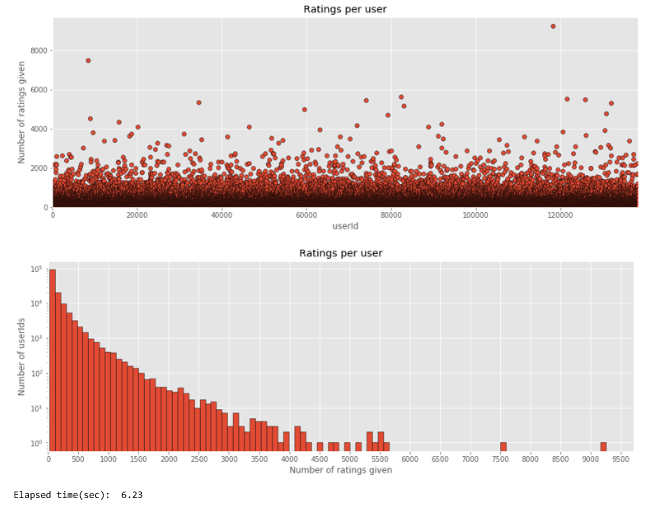
**\*\*PLOT#9\*\*: Rating timestamp vs. movie year vs. rating count  
\*\*INSIGHT#9\*\*: Besides the evident insight that newer movies get more ratings, and that older movies get a number of ratings inversely proportional to their age, we can also see than the oldest movies have just received ratings very recently (2010+), implying those were not readily for users to watch available before. There's also sort of a "block behaviour", where at any point in time, movies with X number of ratings stop some Y years before (e.g. in 2005, there are no movies older than 1980 with +10000 ratings).**

****

**\*\*PLOT#10\*\*: Rating timestamp vs. movie year vs. average rating  
\*\*INSIGHT#10\*\*: After 1980, average rating declines clearly; a lower ratings count for sure has an impact, but it seems -perceived- movies quality did step-decline at that point. Also, movies released 1930-1980 and rated 1996-2003 seem to comform the better rated block.**

****

**\*\*PLOT#11\*\*: Ratings per user.  
\*\*INSIGHT#11\*\*: As expected, exponentially declining function, with no users with less than 20 ratings give, and some outliers with 7.5k and 9.2k ratings...that's insane, 9,200 movies watched and rated, at 2 hours each, more than two years 100% dedicated to watching movies!!!.**

****

**\*\*PLOT#12\*\*: Ratings per movie.  
\*\*INSIGHT#12\*\*: My favorite movie (no kidding) is the "best-rated & most-popular" movie ever!!! But I'm so pissed of Gladiator didn't make it anywhere close to the top....** 