glancing at the movie\_titles

Release\_yr Title

Movie\_ID

1 2003.0 Dinosaur Planet

2 2004.0 Isle of Man TT 2004 Review

3 1997.0 Character

4 1994.0 Paula Abdul's Get Up & Dance

5 2004.0 The Rise and Fall of ECW

6 1997.0 Sick

7 1992.0 8 Man

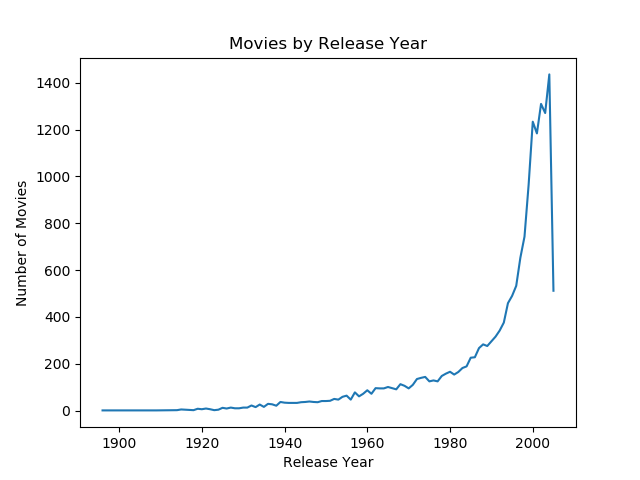
8 2004.0 What the #$\*! Do We Know!?

9 1991.0 Class of Nuke 'Em High 2

10 2001.0 Fighter

Shape of movie\_titles: (17770, 2)

There were three columns on the movie titles data. This data refers the movie id in the training file to the release year and the title of the movie. I changed the movie id to the index of the data frame so the shape of the data has two columns instead of three.



When I plot the number of movies per their release year, most of the movies were released around 2000. There is further to observe whether there is a correlation between this distribution and the dates users left their ratings.

glancing at the training data 1

User\_ID Rating Date

0 1: NaN NaN

1 1488844 3.0 2005-09-06

2 822109 5.0 2005-05-13

3 885013 4.0 2005-10-19

4 30878 4.0 2005-12-26

5 823519 3.0 2004-05-03

6 893988 3.0 2005-11-17

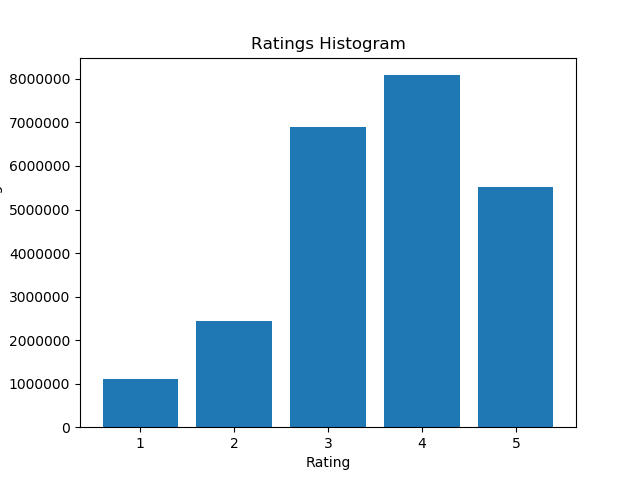
7 124105 4.0 2004-08-05

8 1248029 3.0 2004-04-22

9 1842128 4.0 2004-05-09

Shape of training data 1: (24058263, 3)

When I look at the shape of the training data, there are 24million customer ratings along with User ID and rated dates in the training data set 1. There are total 5 training data sets but I only included the first training dataset to save time and space for the data exploring purpose. Moving forward, we will need to incorporate all the training data sets.



Percentage of Ratings

5.0 22.89

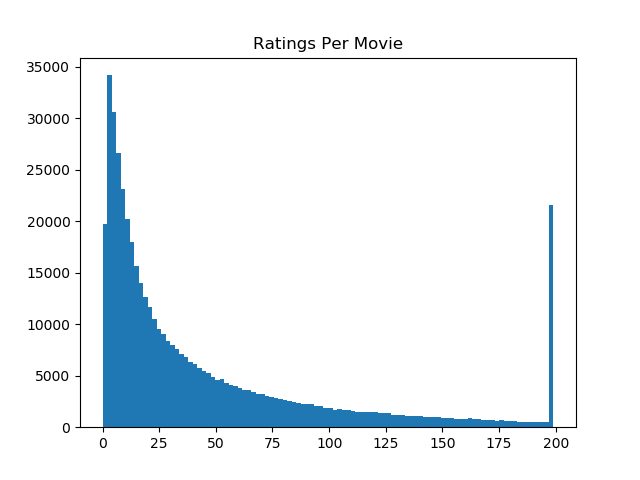
4.0 33.62

3.0 28.70

2.0 10.14

1.0 4.65

As you can see from the table and the histogram, there are majority of ratings in the higher ratings such as 3 to 5, where as there are only 10% and 5% of ratings in rating 2 and 1 respectively.



The third histogram shows number of ratings per user. As you can see it shows highly positively skewed distribution of how many ratings each user had. Mostly, each user has less than 10 ratings, except there was a group that had 200 ratings. I need to explore whether this is an outlier that I need to exclude from my data or from an error in my analysis method.