After conducting EDA it was found that movies with budget less than 1000 were randomized data and were removed for the sake of data integrity.

Using matplotlib

1. I plotted ‘budget’ against ‘release\_year’ and found that the movie industry has be growing successfully and saw a spike in the years around 2010.
2. When I plotted ‘vote\_average’ against ‘release\_year’ it was apparent that , though the quality of the ‘good’ movies did not vary by a large amount, the number of movies which were not ‘good’in the public opinion were getting worse. This could indicate that the criteria for judging movies is getting tougher.

Q1]

After sorting the data and removing all data with less than 1000 as their budget entries, the third lowest budget was found to be 5000 $

Three movies came under this category:

1. The Hunt for Gollum
2. The Party at Kitty and Stud's
3. The American Matrix - Age Of Deception

No cleaning was done on the higher budget under assumption that they were accurate.

The third highest was found to be:

Pirates of the Caribbean: At World's End - 300000000 $

Q2]

Sorted data was filtered to only contain movies released between 2000 and 2005(both inclusive)

After calculation it was found that the average number of words in these movie titles is

2.879668049792531 words

Q3]

Q4]

The values were sorted and revenues were that were less than 1000 were discarded.

The lowest and highest revenue movies are as follows:

Lowest: The Butterfly Room - 1081 $

Highest: Avatar - 237000000 $

Q5]

All movies released in 2006 were accessed and the total runtime was calculated and divided by the number of movies released, and hence the average runtime of movies released in 2006 was calculated to be:

101.68382352941177 minutes