Thank you Stella, so to find out the answers for this 2 task, we will be analysing the data from IMDb movie dataset. For the first task, we remove and replace some of the plot keywords of the 5,000 movie and the terms are summarised in this this wordcloud here. Rather than using all this terms, we have pulled out the top 30 terms which have atleast 70 times of occurance. Some terms are like nudity.. love…. Sex.. and murder.

After running some cluster analysis and comparing the relevance and relationship of the terms, we decided to assign these 30 key words into 4 different cluster. The First cluster is the mature content. For are terms like nudity and sex, which are commonly found in R-rated movie. The second cluster is the terms related family and relationship. Love and gay might be seen to be related to mature content, however, the association of the terms used in the plot keywords shows that it is more associated to this cluster. The next cluster is the crime and violence content and the last cluster is the Science Fiction content, which are the content that are fictional, such as alien and time travel.

To answer Task 2, which is what type of plot elements are associated with better IMDb rating, we also run a regression analysis, including a number of other factors such as no. of critic reviews, duration, no of face In poster to see how are this associated to the rating as well.

The outcome of the analysis shows that Cluster 2 and most of the other independent variables are significantly associated to the IMDb\_score at 0.1% significant level.

Now, Yan Chao will analyse this insights.