**Team E Milestone Report 2**

**Project Goal:**

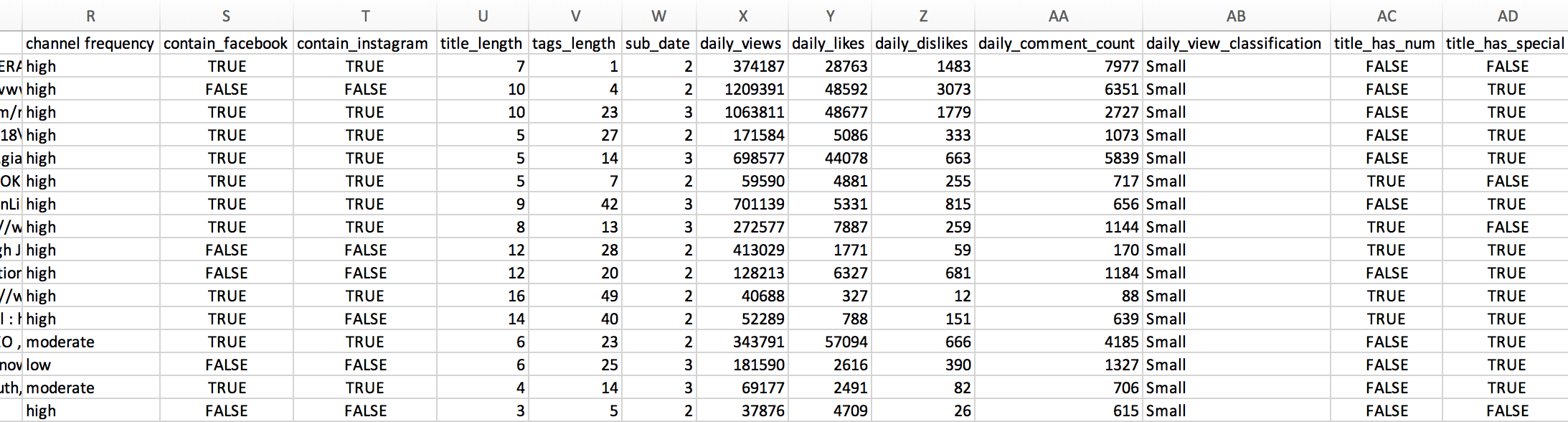
1. Analyze factors that highly affect on Youtube video popularity.
2. Find out the top ten rules and attributes that contribute to high views on YouTube.

**Current Objective**: The original feature contains video title, category, channel title, like counts, dislike counts, view counts, comment counts, publish date, data collected date, description, tag words, and video cover picture. Since our data has a relatively low number of features, in order to determine the association rules that contribute to high views, we did feature engineering and added more features.

**Features Engineering**: The following are the current new features we added into our dataset:

1. Channel frequency: According to the dataset, during 213 days, different channels have different number of video posting. We calculate the video posting frequency and categorize it into three categories: low, moderate and high.
2. Facebook link: This is a boolean categorical feature that demonstrates whether the description of the video contains facebook links and this features has true or false as values.
3. Instagram link: Same as the facebook link feature. It lists true or false for whether the video contains instagram link.
4. Length of title: This column contains integer values, indicates the length of words of the title.
5. Number of tags: This column contains integer values, indicates the number of tags used.
6. Number in title: This column contains if the title contains number, this feature has true or false as values.
7. Special characters in title: This column contains the title contains special characters(such as “!” “?”), this feature has true or false as values.
8. Days between the published date and data collection date: This column contains how long it is from the published date of the video to the date that we get the data.
9. Daily views/likes/dislikes/comments: This column contains the count of daily views/likes/dislikes/comments.
10. Classification of the daily views: In this column, we divide daily views into 3 tags “small”, ”middle”, ” large”, according to the values, each of tag has ⅓ of the total data.

**Screenshot of new columns:**



**Future plan**: The rest of our work is focusing on exploring more informative features and find out the association rules that contribute to high views. Also, we can use regression to predict the outcome.