Second Capstone Project Milestone Report: Bob Ross’ Happy Trees

1. **What is the problem you want to solve?**

Understand patterns in viewership of Bob Ross painting shows.

2. **Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?**

YouTube or the manager of the Bob Ross YouTube channel could decide which episodes to promote to YouTube visitors based on content.

3. **What data are you going to use for this? How will you acquire this data? Describe your data set, and how you cleaned/wrangled it.**

I used data containing episode information, and paintings categorized (‘tagged’) by content, e.g. mountains, streams, etc. These data are found here: https://github.com/fivethirtyeight/data/tree/master/bob-ross. I combined these data with data pulled from the Bob Ross YouTube channel, such as number of views: https://www.youtube.com/channel/UCxcnsr1R5Ge_fbTu5ajt8DQ.

ELEMENTS: The painting categorizations data was already very clean, but I removed categories with less than 5 elements, and then made sure categories were relevant (e.g. removed tags like ‘guest host’ or describing the painting frame).

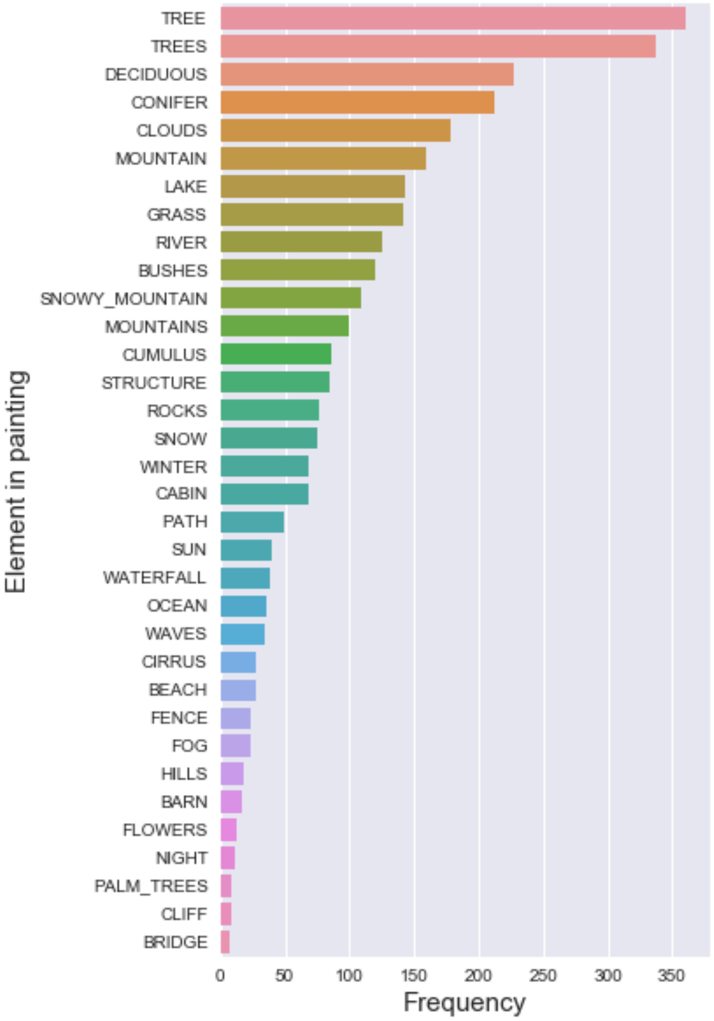
STATISTICS: I also scraped video information from the YouTube channel using the YouTube Data API, and then read through the resulting json text for information like episode title and view count.

Finally, I performed a join on these two tables, dropping any episodes that weren’t contained in the elements dataset, and removed rows with missing data in viewership.

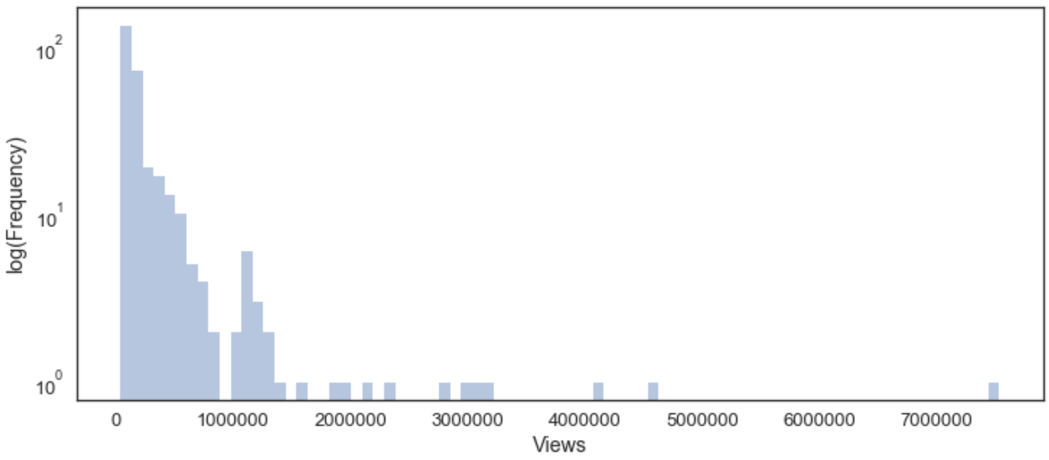
Other potential datasets? I plan to also look at the titles of the paintings and make a predictive tool to develop ‘new’ viral titles.

4. **Explain your initial findings**

The most frequent element in a Bob Ross painting is a tree, followed closely by the presence of multiple trees, or trees of specific types (deciduous or conifer). He is, after all, known for his ‘happy trees’! Mountains, lakes, and grasslands also figure prominently.



The final joined dataset contained information about 302 episodes. The greatest number of views for any given episode is 7,529,170. The top 10% of episode viewership is greater than 727,790 views. There seems to be a natural break in the views that you can see in this histogram:



I have categorized videos with 900,000 or more views as ‘viral’, and there are 26 of these in the prepared dataset.