**IMDB**

The first thing we decided to do was work off a similar database. We would then ask different questions about the data.

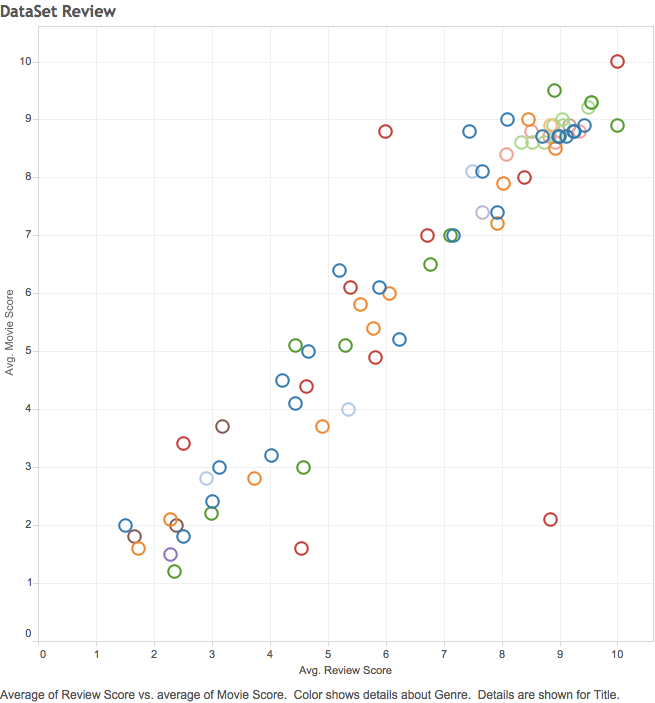
Mike : Can you predict the genre a user is discussing in a review?

Robbie: Can you predict the overall movie rating based on the summary of a review?

Robbie: Can you predict the user’s movie rating based on the summary of a review?

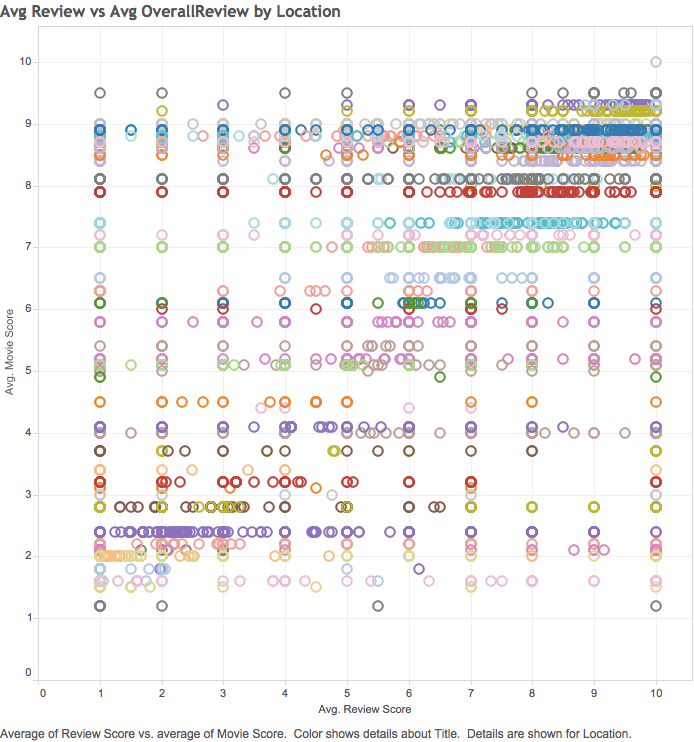
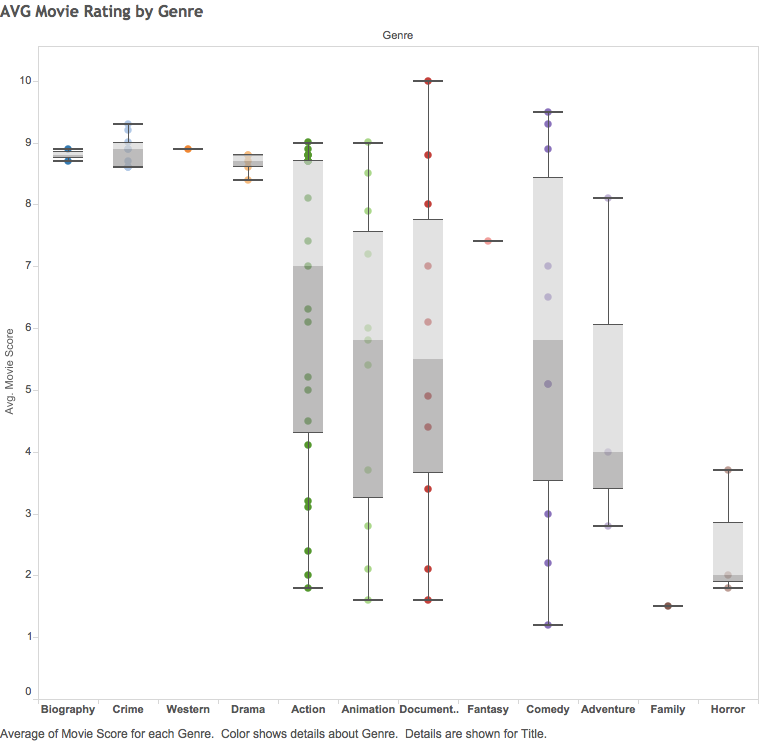
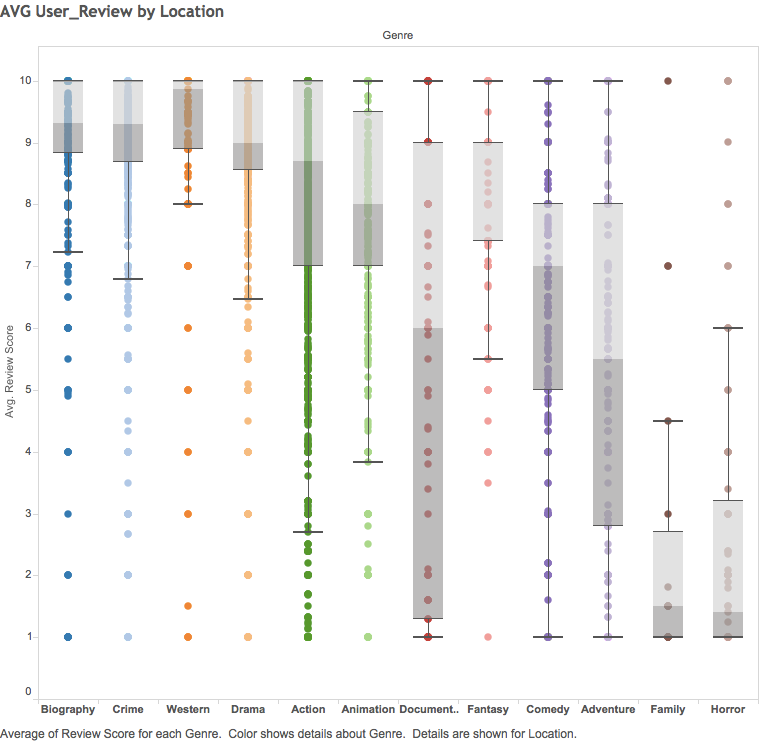
In order to get the best data, we wanted an evenly distributed data set based on genre and movie rating. We randomly chose 9 movies from 6 different genres.

We sampled Action, Comedy, Horror, Sci-Fi, Animation and Documentaries.

We decided to choose one movie from each rating score. IE. One for rating of 1, one for rating of 2, etc.

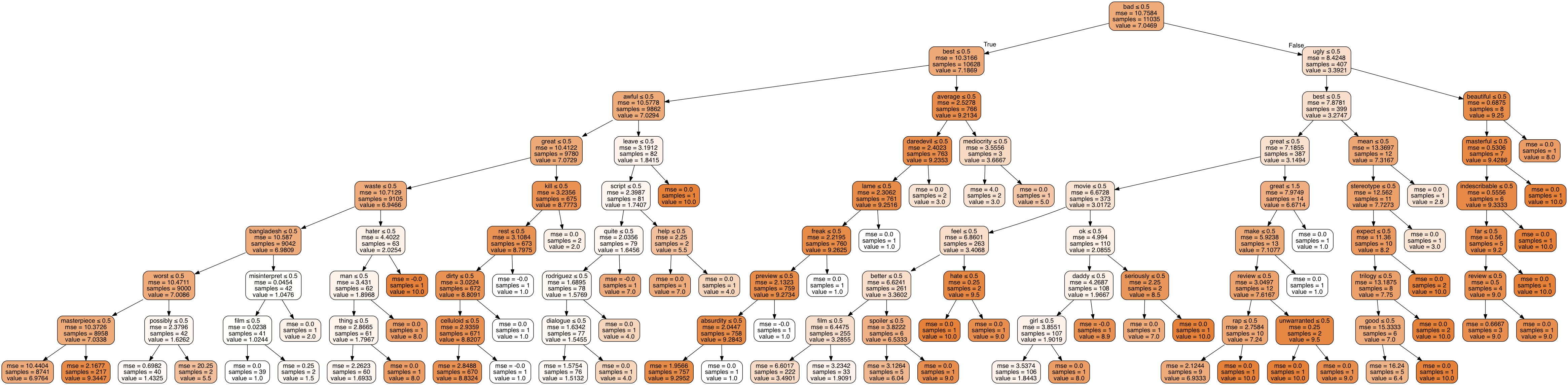
We added a sleep timer to our api function because we were encountered a 502 error and this was very helpful when we were downloading more than 200 reviews. We downloaded as high as 2000 but working with that much data was too much for our computers and we decided to move forward with the 200 reviews per movie.

This graph shows that we did a pretty good job choosing a strongly correlated data set to start with.

The work we did to acquire the data can be seen at Project6\_GET\_summary\_200.ipynb. Additional visuals can be seen below.

Paul:

We originally focused on the reviews and because I had the weakest computer, I decided to download summary’s instead and see if that would help. These summary’s potentially represented the same view they were expressing in the longer winded reviews. I encountered several issues like not knowing when to vectorize and could not get lemmatization or stemming to work. I was able to figure out the proper order and get to a decision tree. You can see this at Project6\_Fail\_FirstAttempt.ipynb. I started over and did the rest of my work in Project6\_Paul.ipynb.



Major blockers was lack of computing power.

Having an underpowered computer, I was constantly waiting for my computer which has probably been pushed to it’s limits. It would be nice if we could cover cloud computing as I have noticed a lot of companies are looking for people with AWS experience. It would be nice to work with a data set that is already clean so we can practice doing NLP.

Robbie:

Mike: