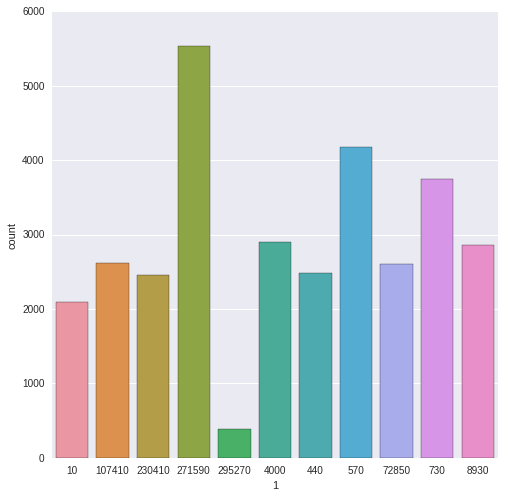
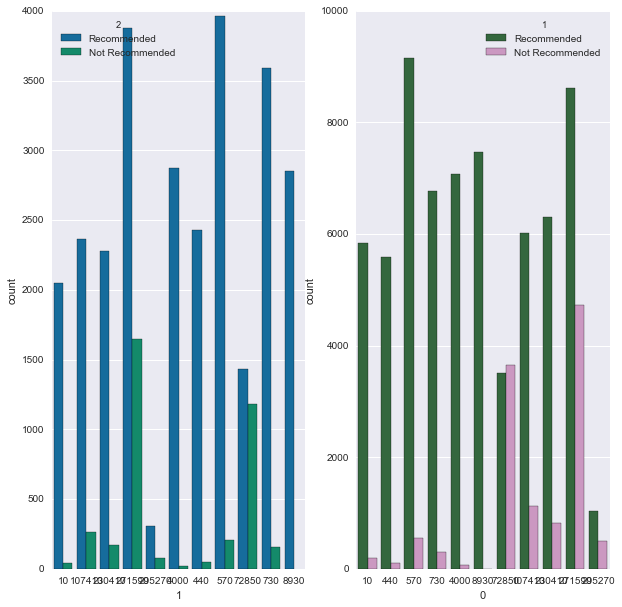
EDA Report

To try and answer whether or not the valence of reviews for a video-game is related to the overall success of a video-game, I had to find data that would encompass multiple different domains of video-game (first person shooter, role-playing game, etc.) Using Matt Mulholland’s scraped data, <https://github.com/mulhod/steam_reviews/>, in JSON format was easy to reclaim and get straight to work on. These JSON files contain 79,437 total reviews over eleven very different games. The posting dates of the reviews range between the earliest: 2010, and the latest, 2015. The rows of data consist of the following: Time stamp for the review, number of hours played total for the game being reviewed, steam ID, steam URL, and various other metrics that are attuned to either the user account or the review itself.

The cleaning process was fairly simple given the column of Steam URLs, I used an already written class in order to munge this data because the steam ID column did not always give a numerical representation of the steam ID as required for the steam web API. The web API specifically asks for a '64-bit' ID in the API calls. To combat this I used <https://github.com/annasapek/steamid>. This class allowed for converting what is referred to as a ‘vanity URL’ (steam URL) to a ’64-bit’ steam ID to then retrieve how many hours the reviewer had played the game thus far. Through the process, retrieving the hours played was unsuccessful about half of the time, either because the reviewer had changed their vanity URL (you can do this through your Steam settings), or in the JSON response received from the API, the game ID for the game they reviewed did not exist, indicating that they had reviewed a game without having played the game in the first place.  The games are denoted by their specific application ID on the steam market. Other than games 271590 (Grand Theft Auto), 570 (DoTa 2), 730 (Counter-Strike: Global Offensive), and game 295270 (Football Manager), I have between two thousand and three thousand reviews for each video-game. The count-bars are representative of what was left, post-cleaning. I don’t think performing analysis on the Football Manager dataset will be relevant as there are fewer than one thousand remaining reviews, of which, only 77 reviews remain that don't recommend the game, 305 recommending it.

 Further exploring the data, I noticed a very obvious class imbalance that could potentially create problems when trying to depict commonalities between the two classes recommending the video-game or not recommending the video-game. Under the case of using Naive Bayes to define likelihood of words between the classes it would be an issue given the minuet amount of reviews for each video game not recommending the video-game. This would also create problems in defining valence for different dialogue between the two types of reviews.

Another issue about this data that could be worked or expanded on somewhat easily is the small amount of games I'm taking observations from. Steam encompasses a large amount of games and a large variety of game types, given this I could potentially automate a process to clean, train, and score models on review data for a much larger number of games, and perhaps find more general similarities between the two classes of reviews. For this data generalizing for future data seems somewhat impossible because of the vastly different language in idioms, names, and jargon surrounding the games.

Another thing to take into account about the data itself is all of these games are out of alpha, beta, or anything of the sort, they are all released. To really predict the success of video-games for profit it would be better to have reviews of games when they are in beta, and after they come out of beta if they're successful or unsuccessful would be what I'd be predicting. To predict the blow up of a game from the beta reviews would be so ridiculously profitable to investors just because these games have their initial blow up, and the communities still continue to grow after the game initially being released.

There is definitely more aspects of NLP in this data beyond just sentiment and word counts. Perhaps the different parts of speech, name recognition the frequency of words, the frequency of certain parts of speech. There definitely would be more dimensions to the data if the reviews I was looking at had numerical representations of a rating. This is problematic as when a user recommends a game, there is no indication to how much they recommend or don't recommend the game.