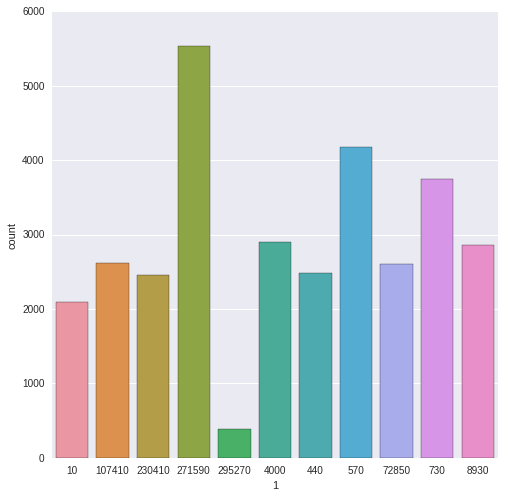
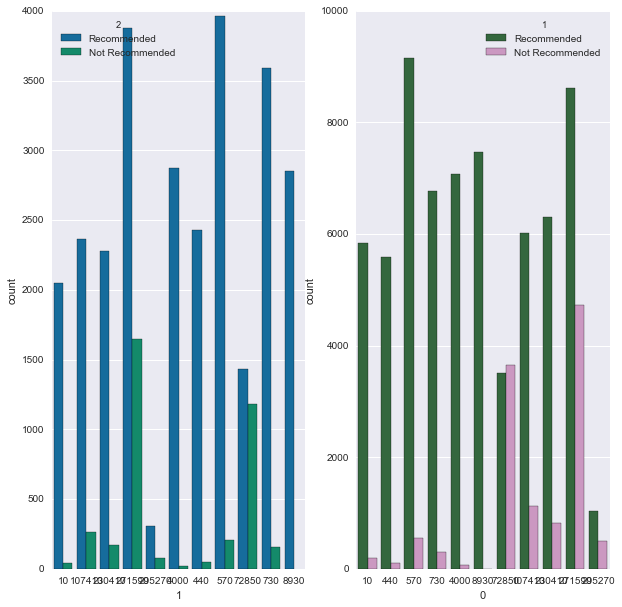
EDA

To try and answer whether or not the valence of reviews for a videogame was related to the overall success of a videogame, I had to find data that would encompass multiple different videogames (first person shooter, roleplaying 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 with. These files contain 79,437 total reviews over eleven very different games. The reviews’ posting dates range between the earliest being 2010, and the latest, 2015. The rows consist of the following: Number of overall achievements earned, not just in the game reviewed, but for all of your videogames owned on Steam, timestamp 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 steam URL column. 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. 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 that could be passed to the steam web API to 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 videogame. 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.

Further exploring the data, I noticed a very noticeable class imbalance that could potentially create problems when trying to depict commonalities between the two classes recommending the videogame or not recommending the videogame.