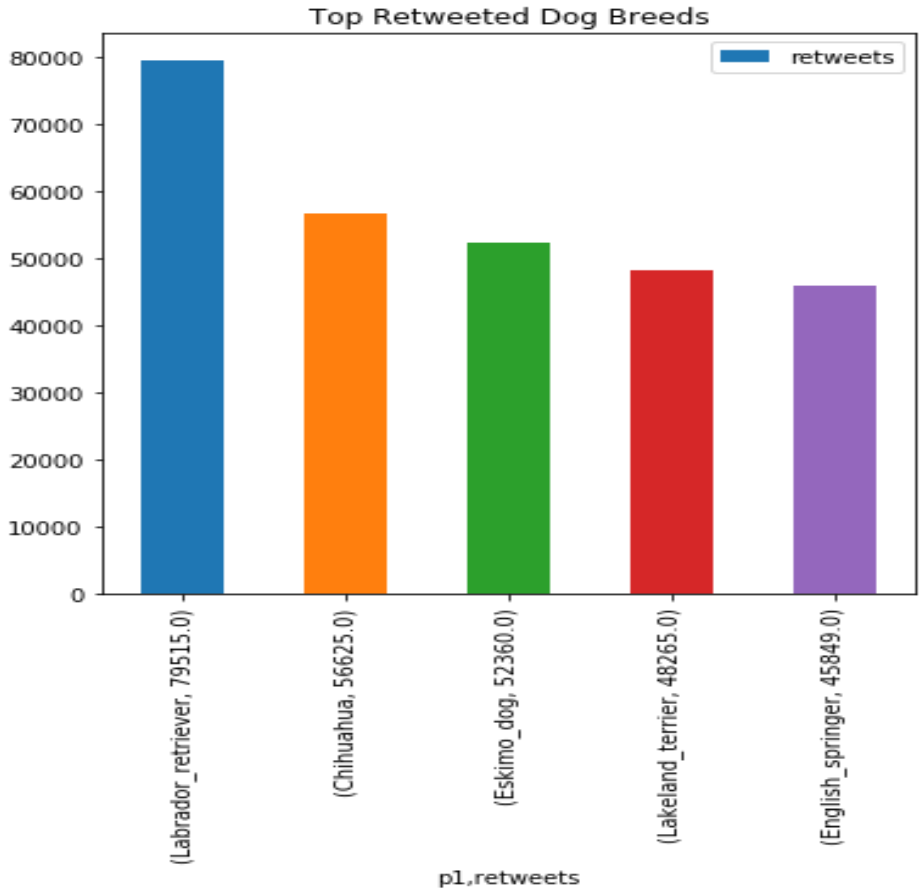
Act Report

In this project, we have performed the data wrangling on the collected data from 3 different sources. The collected data were processed in gather, assess and clean phases. Cleaned data for all datasets was merged and stored into a single master dataset and performed further analysis and visualization to conclude the project.

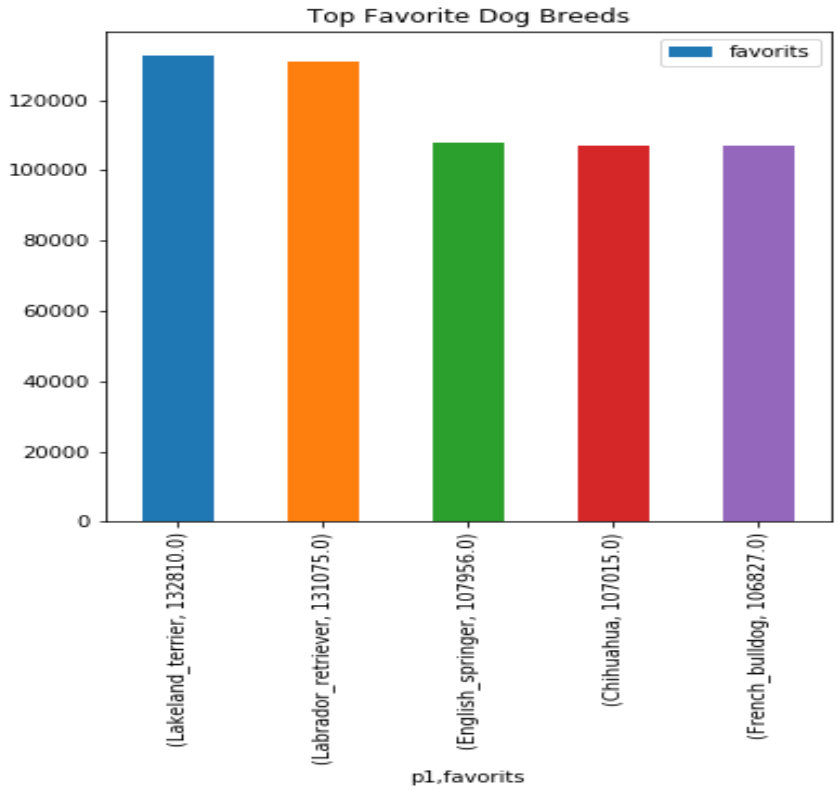
While analyzing the master dataset, the visual and programmatic assessment were performed with help of pandas and matplotlib libraries. In the programmatic assessment, the sorting and grouping operations were performed on variables such as p1, stages, retweets, favorits and p1_confs to collect various outputs and get some insights on the dataset. The variables retweet and favorits played major role.

Insights and Visualization:

1. The aim was to find out the most liked and retweeted dog breed and verify if the same was captured by neural network algorithm. The grouping of p1, retweets and favorits were performed on p1_conf confidence variable and sorted by retweets and favorits. The same kind of operation was performed by adding the stages variable to check which dog breed was on top in all or max stages in their life. The output of dog breeds confirmed that people really love the Labrador related posts in spite of their stages and the algo captures the same with good confidence level.



2. The next aim was to find out least liked and retweeted dog breed and again verify if the same was captured by neural network algorithm. This time the same grouping and sorting operations were performed but with ascending the result. We've found that English setter dogs were least liked, retweeted and the algorithm worked perfectly in this case as well.



3. As the favorits and retweets variables played major role, so checked the relationship between them. The correlation coefficient between favorits and retweets was 0.88. Since this correlation was positive, we can confirm on, if people like something on tweeter, they usually retweet.