Wrangle Report

Data Gathering:
In this data wrangling project, I gathered data from three different sources for data analysis. The Twitter archive is given in the form of a CSV file. This archive contains the basic tweet information such as tweet id, tweet text, tweet time stamp etc. Tweet image prediction file contains the dog breed predicted by a neural network, this information is stored on Udacity server. The most challenging part of data gathering is using Python Tweepy to collect extra tweets information (retweets and favorite count) via Twitter API. By using the tweet id from tweet archive, I queried the Tweet API and obtained the JSON data. In my program, the query was initially sent out at a very high rate which triggered the sleeping mechanism and made the query process very slow. To solve this problem, I forced the program to wait 0.1s after sending every 100 queries..

Data Access/cleaning:
I evaluated the data set for quality and tidiness issues and then come up solutions to fix them. It is easy to see that several variable types are wrong, for example, tweet_ID should be string instead of number, timestamp is not in the form of timestamp, etc. There are many faulty dog names, my current solution only capable of fixing faulty names individually. Another issue is that some ratings are extracted from texts incorrectly, because 1) the first fraction is always treated as the rating, 2) date in the form of x/y/z is used as rating, 3) or any other number in the form of x/y. To tackle 1), I first filtered out the data with rating_denominator larger than 10 and then select the last fraction as the correct rating. This method doesn't work for 2) and 3). But they can always be fixed individually. Entries which are retweets are deleted. Original data set has separate column for each dog stage which are merged into one column. The next step is to left join the prediction data set and tweet API to the main tweet archive. The lase step is to add a column for the predicted dog breeds.