DATA WRANGLING

The data that needed to be gathered for the wrangling comes from three different sources. The csv file which was already given, the TSV file which I have to download programmatically, and data from Twitter in the form of JSON Format. The request library allowed me to extract out the content and csv reader allowed me to read the contents in tab separated format and hence therefore It allowed me wrote the contents in separate file as comma separated objects effectively making it csv file. Tweepy Library allowed me to extract out the metadata of the tweet given a tweet id and I choose to extract out the language, the retweet count and the like count. These metadata of the tweet were written in the form of JSON in new Lines which I then was able to import in the form of data frame using pandas. Having gathered the data in the form of frames in appropriate variables I turned towards assessing the dataset for possible consistency accuracy completion issues (aka Dirty Data) and eventually looked for the structural issues in the data frames. The dataset imported from JSON apart from converting language to categorical column dosen' t require much cleaning . The dataset which was giving me was grossly untidy and required much work to translate it into an appropriate format. Some of the work I did on twitter archive dataset included

- Removing the redundant source column which consist of only 4 links which didn't went nowhere and does not uniquely define the tweet in any way, also one link in particular which was on more than half of observations
- Removing out the padded zero from the timestamps . The string so formed eventually can be translated into date time object
- Removing out the ridiculous names out from the name column like a , an infurating etc
- Removing the 0 from the numerator and denominator rating. However I suspect further work can be done on it and many outlier can still be removed if analyzed by a person domain knowledge of we rate Dogs
- **The Numerator Column contains the Floating point Variable and it was necessary the correct no should be extracted from the Text through the application of Regular Expression. No difficulty of same kind was observed in Denominator**

The Image prediction Dataset have a categorical column indicating the choice of photo, but it would be treated by the pandas as ordered variable and can create problems in Visualizing and thus needed changing

Tidying the dataset required much effort particularly because the values of the variable makes up the columns which make the whole data messy. On top of that most of the stages were missing and since it is in large no **it required me to make a separate category as NA for them . After than through the melt function of pandas I was able to make required columns for the pandas but thrice the no of rows , one for each variable which then I filtered out by removing out the values which didn't have 0 in front of them which in this case indicates duplication**

Having Cleaned the data I combined the dataset of (archive and like- retweet) dataset through internal join. I let the image dataset to be as it is (saved in the file PREDICTION_breed_dog_CLEANFILE) and not

joined with the main dataset as it indicates information totally different from the metadata of tweet . However I made it sure that all the data frames contains the information about the common Tweet id’s