Sami Farooqui
Wrangle data report

This project had me take tweets from a the WeRateDogs twitter account and clean to data to a more tidy dataset and perform statistical analysis. My first step was to obtain the data given to us by Udacity by loaded in onto my Jupyter notebook. Next I loaded the python libraries and asked python to read the twitter archive file. I also signed up for a twitter developer account and after going through udacity‘ steps to obtain permission was able to get the access tokens easily. The next step was to use the access tokens I was granted to load the twitter api json file, this was for me the most difficult task in the project, partially because I need to write for loop to iterate over the json file and also because twitter would lock me out for extended periods of time, so loading all the data took a long time.

Once I received all the data I had three datasets, the archived file, tweet file, and the image breed prediction file. I viewed the datasets and found a few problems with each. The archived file has many columns some of which were not necessary, the names column had too many non-names, the retweets column was not necessary for this particular dataset, and the stage columns needed to be grouped into just one stage column rather than the individual ones. I was able to remove the non name outliers and replace with none if no other names were present using the replace function, and used the drop function to drop all retweets. My next step was one of the more difficult ones, as used the datetime ibrary to convert the time stamp column into an object, and then use the apply function to convert these into two separate columns labeled time and date. I learned a lot for this procedure from the SoloLearn App for IPad, and it was a big help for me to be able to learn some other capabilities of python in a more portable way. Next I replaced the 4 columns that would indicate dog stage and store them in a single neat column indication dog stage or none if none were listed. I replaced the column id with tweet_id in the tweet dataset therefore allowing me to merge all 3 datasets into a master dataset later on. I dropped the retweet status in tweet dataframe and removed unwanted columns then went on to the breed image prediction dataset, there I found that underscores were present in the p1, p2 and p3 column, so those were removed, and the same columns were also given titles, so the dataset appears cleaner and neater. I used inner join to merge the 3 dataframes into one master dataframe, and saved to a master file. There was an unnamed column which wasnt needed and therefore also dropped. Next I printed some statistics of the dataset, using the describe function. Then checked the tweet with the highest rating, as well as tweets with ratings over 10. I used the counter library to count the most often ratings, as well as most often names. Lastly I plotted two graphs one showing the relationship between the rating ratio and favorite tweets, and the other showing the relationship between favorite tweet and retweets. I imported seaborn for the retweet favorite plot, and used matplotlib for the ratio favorite plot.