Data Gathering- I used pandas to open the csv and tsv files. Based on the tweet_id from the csv file, I used tweepy gathered data (id, full_text, retweet_count, favorite_count) from tweeter API, and saved the data into a json file, then saved the json file into a txt file called tweet_json.txt.

Assessing & Cleaning Data- I used pandas to get the basic information(shape, data type, head, null_values, duplicated value) from each data set, so I can find out which columns are going to be used for my analysis and which columns are not. Then, I deleted the duplicated rows, columns and the null values. Based on the statistics of the numerical columns, since there are some very large values for some column, I created a new categorical columns for those columns, so that would be better for visualization for the next step. For the “dog stage” columns, I combined them into one dog_stage column with the same information to make it easier to access. Since df_tweet and df_archive share the same column tweet_id, I merged them into one data frame which I used to create the master data set.

Bootstrapping for the prediction data set- For the dog image_predictions dataset, I used bootstrapping method for a hypothesis test the prediction mean for prediction 1. The null is the whole sample prediction mean is the same as the 20 sample prediction. By simulating 10000 times of random sampling 10 from the 20 samples, I got the prediction mean, mean distribution and p value. Since the p_value is 1.0, we fail to reject the null which the mean of the sample data set is the mean of the whole population.