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.

8 data quality issues:

twitter-Archive-enhanced data set :

1. Checked null values in the data frame , noticed 'expanded_urls' column has 59 missing values. Removed this column.
2. After checked the statistic for the data frame, I noticed the rating_numerator is within a very wide range of numbers 10-1776, and it's not evenly distributed, so I categorize these values into different levels based on the statistics, and created a number categorized data column as rating_levl

tweet_json.txt file:

3. After checked the statistic for the data frame, I noticed the favorite_count and retweet_count are within a very wide range of numbers , and they're not evenly distributed, so I categorize these values into different levels based on the statistics, and created a number categorized data columns.
4. There are some duplicated columns like the retweet_count and retweeted is the same column, dropped retweeted columns.
5. Noticed the 'full_text' column is not going to be used for analysis, delete the column.
6. There are some null_values in favor_count, drop the null_value rows.
7. In order to merge the tweet_json and twitter-archive-enhaced data later, I changed the 'id' to 'tweet_id' to match the same column name.

Merged data

8. After merged the data, noticed there're null values , so I removed the null_value rows.

2 tidiness issues:

1. Removed the useless columns that are not part of the analysis
2. Noticed there are 4 columns for 'dog stage' information, I combined them into one column as stage, and delete the 4 columns.

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.