Gathering:

Gathering the data and reading them in pandas dataframes for each.

- Twitter_archive was provided as csv and directly read. Containing an archive of tweets from WeRateDogs
- Image_prediction dataset was downloaded programatically using requests library. A model that predicts the type of dog using images.
- Tweet_json was downloaded from Udacity due to difficulty of getting an API Developer access from Twitter. Which is an updated version of twitter_archive which also containts retweet and favorite counts

Assessing and Cleaning

Assessing the data's tidiness and quality was based on the importance of each issue to my analyses.

- 2 tidiness issues were identifed:
    - Transforming the four columns that indicate dog stage into a single column "stage". Relying on regular expressions to read tweet texts and find the dog stages and fill the stage column
    - Merging all dataframes together since they all represent tweets. Done by merging twitter_archive with image_prediction using the "tweet_id" column. Then renaming the "id" column in tweet_json to "tweet_id" to merge the result of the first operation with this dataframe resulting in a master dataframe.

- 9 quality issues were identified:
    - The datatype of tweet_id should be casted to string because of int limits. Done straightforwardly by using pandas' "astype"
    - Dog name with value "none" should be replaced with NaNs, done by using Numpy's nan.
    - Timestamp should be casted to datetime instead of string to assist us in our analyses. Using pandas' "to_datetime" and providing a date format that matches the one in our dataset.
    - Removed tweet replies and retweets since they are not related to the dataset of dog picture tweets. By first removing the records then dropping the columns.
    - Fixed some dog names and replaced the ones that were given unrealistic names with NaN values.

- Removed tweets that had no urls linked to by first checking using pandas’ “isnull” function.
- Standardized ratings by dividing the numerator by the denominator to assist us in our analyses