Gathering the 3 sets of data:

For the twitter archive (archive) given to us by Udacity I used pd.read_csv in order to read the data.

For the image predictions (images) again given to us by Udacity I used the requests library and then imported the data using pd.read_csv(StringIO(r.text), sep='\t')

For the twitter API (retweet_and_favourite) - I queried the API for each tweet’s JSON data using Tweepy and then stored these all in a text file (tweet_json.txt). I created a dataframe using the queries using the tweet id, retweet count and favourite counts.

I now had my 3 data sources.

Assessing -

This meant looking at data quality and tidiness of the data.

Tidiness issues

1. Having 3 sources of data is untidy so I decided to merge them.

2. The source was taking up too much room in the table so this may have to be looked at.

3. Look into the probability columns to see if they are needed.

Quality issues

1. Rows that were not consistent over the 3 sources will have to be deleted in order to not have any missing data.

2. We do not want to look at retweets or replies so these can be deleted. These have a lot of missing values too.

3. I wanted to see if there were any duplicate rows.

4. I wanted to see if there was any missing data.

5. The names column had several names that were not names so this would have to be looked at too.

6. The column headers in images (p1 etc) were very vague.

7. Timestamp had seconds and minutes which I did not think was relevant and was an object not a date time.

8. Dog rating was over 2 columns so this will have to be looked at and merged. Also assessed to see if any of the ratings were wrong / not accurate. I saw that the denominator should be 10 - when it was not there seemed to be more than 1 dog in the photo.

9. The dog stages in ‘archive’ were over 4 different columns - these will have to be made into one dog stage column.

10. Floofer is not a dog stage so this column will have to be removed.

Cleaning:

Tidiness issues:

1. In order to see all the data I decided to merge all of the sources.
2. I looked at the sources column and there were only 3 types of sources. Therefore I changed the names to shorter and more legible names.
3. I realised that the probability columns were not that relevant as only the true or false was. Therefore these columns were dropped.

Quality issues

1. I merged the sources using an inner merge so that all rows with missing data could be deleted.
2. The retweet columns and in reply to columns were dropped as we only want to look at original tweets. These also included a lot of missing values.
3. I looked to see if there were any duplicates - there were none once the sources were merged.
4. The expanded url column had missing data so these rows were dropped
5. The names that started with a lowercase letter were looked at, they all seemed to not be names and therefore these rows were dropped.
6. The column names in images were changed to Guess 1 etc to make it more readable.
7. Timestamp was changed to a datetime and the time was removed as I felt that the date was relevant and not the time.
8. I made it so that the denominator was 10 as any bigger it seemed that there were more than 1 dog in the photo. Therefore the rows that were not 10 were deleted. I then dropped the denominator row as if they were all 10 then it was not needed and renamed the numerator column to dog rating.

9. For dog stages, I merged the columns into one dog_stage column to make it more succinct. This also made the table more tidy. Some had 2 different stages of dog in the column so these were looked at to be more accurate.
10. I dropped the floofer column as this is not a dog stage.