I started by downloading the dataset necessary for this project both programmatically and directly from the web and then I also scraped data using the twitter API.

After downloading and loading my datasets, I moved on to importing the necessary librabries needed for wrangling and cleaning the data. I visually assessing and programmatically assessing the data and right off the bat I spotted some Quality issues in the twitter_archive and image_predictions dataset. The Quality issues I spotted are as follows:

1. Drop the columns with too many null values (**in_reply_to_status_id**, **in_reply_to_user_id**, **retweeted_status_id**, **retweeted_status_user_id**, **retweeted_status_timestamp**) as they won't be neccesary
2. Missing values represented as "None" instead of "NaN" in the **twitter_archive** dataset
3. **timestamp** & **retweeted_status_timestamp** as **object** instead of **datetime** in the **twitter_archive** dataset
4. **in_reply_to_status_id**, **retweeted_status_id**, **retweeted_user_id** & **in_reply_to_user_id** as **float** instead of **int** in the **twitter_archive** dataset
5. The **source** column should be of type **category** instead of **object** in the **twitter_archive** dataset
6. The **source** column needs to be sliced to get the actual source of the tweet rather than the whole HTML
7. The **ratin_numerator** & **rating_denominator** has values outside the **15/10** rating standard
8. We don't need the exact time of the tweet in the **twitter_archive** dataset, the date is okay
9. We have predicions of **False** and we need just **True** values since it's for dogs in the **image_predictions** dataset
10. Inconsistent dog type name predictions in the **image_predictions** dataset

Here are the Tidiness issues I spotted after assessing the data also:

1. The dog categories should be a categorical variable in a single column rather than 4 columns in the **twitter_archive** dataset
2. Retweeted posts not needed for this projected, should be deleted from the **twitter_archive** dataset
3. We need just 2 tables ie One with the whole tweet info (joining **twitter_archive** & **tweet_info** on tweet_id) and the second table should be the image predictions

After spotting all the above issues with the datasets I moved on to use the following steps in cleaning the data and getting it ready for analysis

- I made a copy of all the datasets first.

Then for the Quality issues I:

- This columns(**in_reply_to_status_id**, **in_reply_to_user_id**, **retweeted_status_id**, **retweeted_status_user_id**, **retweeted_status_timestamp**) won't be necessary for this project and coupled with having too many null values they need to be dropped

- I'll be replacing the "None" with *NaN* so it can be used properly during analysis or further cleaning in the **twitter_archive** dataset

- The dates in the “**timestamp**” column are represented as *object* and needs to be changed to *datetime* data type

- This would've been changed to int64 but since we dropped the columns no need for that anymore. Nothing was done.

- I changed the **source & dog_type(doggo, floofer, puppo, pupper)**column to *category* data type

- I sliced out the html tag”<>” part of the text so I’ll have just the inner text left.

- I dropped the rows where **rating_numerator** & **rating_denominator** are outside the rating standard ie greater than 15 & 10 respectively.

- I stripped the time from the timestamp column so the date alone will be left since the exact time of the post/tweet won’t be necessary.

- I dropped the rows where **p1_dog** prediction is false since it has the highest prediction accuracy, from visual assessment the prediction accuracy drops after each prediction so the first prediction is the most accurate.

- I removed the "_" and make everything lower case letters

- The Tidiness issues

- I reshaped the data frame and make the dog type a single column and using fillna I'll fill the places where the dog type was specified

- I deleted the rows with the retweeted posts "RT" at the beginning of the tweets

- I moved to my final step of the cleaning and wrangle process by joining the 3 tables on the "tweet_id" column to get a master dataset

After cleaning the dataset and joining the as one in a single master dataset, I saved the new master dataset and used it for my analysis.