1.Wrangling Report
by Mohamed Alsayed

1.1 Gathering and extracting data

In this section we collected 3 different tables using different ways:

1.downloaded twitter-archive-enhanced.csv manually from Udacity
2.downloaded image-prediction.tsv programmatically using requests library in python.
3.scraped data of all tweets using each tweet id represented in the twitter-archive-enhanced.csv with the Twitter API and connected to the API using Tweepy library in python.

we gathered all files in the same directory and starting the assessing process.

1.2 Assessing

In Assessing we explored the data to find quality issues for all the collected data. we looked at each file info and coloumns types and structure then tried to find redundancy related issues by assessing all the data together.

1.2.1 Quality Issues

api_df table (data pulled using Twitter API)

1. Zero Values columns (possibly_sensitive , possibly_sensitive_appealable) which adds no value to our data
2. values in column Source contains href a html tag which we won't be using in any of our analysis.
3. not related columns data to the scope of our analysis (user, favorited, retweeted)
4. Null values columns (contributors, coordinates, geo, place, quoted status id, quoted status id str) which adds no value to our data

image_predict table (data downloaded programmatically using requests library)

1. data is separated in many different ways in columns (p1, p2, p3) like the use of (-, _)

twitter_archive_df table (data downloaded from Udacity manually)

1. different prefix of dog names (a , an)
2. empty cells are not defined as null but instead defined as string
3. rating is not correctly defined/entered as mentioned
4. Column(doggo, floofer, puppo, and pupper)values either None or its column name
5. timestamp is not defined as datetime but instead defined as string

1.2.2 Tidiness Issues

1. the p1, p2, p3 contain redundancy since the p1, p2, p3 is not unique throughout the row in image_predict
2. created_at / timestamp, source, text, in_reply_to_status_id, in_reply_to_user_id are duplicated in api_df table and twitter_archive table
3. api_df and image_predict should be part of twitter_archive table
4. Entities data seems to contain image information which are already contained in the twitter archive data, like the image_url and extended url
5. Extended entities column contains duplicate information of the entities column

1.3 Cleaning

In this part we are going to clean the issues we found in our data above

1.making a copy of api obtained data and dropping each retweet value so that we only keep genuine tweets.
2.fixing data by separating them in unique way not in many different ways as it presents in columns (p1, p2, p3) like the use of (-, _)
3.timestamp is not defined as datetime but instead defined as string so we defined it as timestamp.
4.Merging the api_df and image_predict table to the twitter_archive table, joining on tweet_id and id.
5.Dropping Null values columns (contributors, coordinates, geo, place, quoted status id, quoted status id str) which adds no value to our data
6.Column in twitter_archive_df (doggo, floofer, puppo, and pupper)values either None or its column name
7.dropping Zero Values columns (possibly_sensitive , possibly_sensitive_appealable) which adds no value to our data
8.empty cells are not defined as null but instead defined as string so we defined them as null
9.rating is not correctly defined/entered as mentioned