Wrangle report

WeRateDogs is a twitter account that rate's people with humorous comment about the dogs.

The Objectives of project:

1- **Gathering Data**
2- **Assessing Data**
3- **Cleaning Data**

Gathering Data

The first step we will gather the data from several sources

- The Twitter Archive file have been downloading manually.
- The Image Prediction file have been downloading Programmatically by using requests library.
- Downloading the tweet json file.

Assessing and Cleaning Data

Since we are done with gathering data, now we will start with assessing and cleaning data. We observed some of issues of the data and solve it.

Quality and Tidiness Issues

Table name	Issue	Solution
Twitter archive Image prediction Tweet data	The datatype of tweet_id is integer in all 3 tables	Convert the datatype to string in all 3 tables

Twitter archive	The datatype of timestamp is object	Convert the datatype to datetime
Twitter archive	Only keep the original tweet	Remove all retweets
Twitter archive	We have a lot of columns that not necessary	Drop the columns that we don't need in twitter archive table
Twitter archive	Not all the names of the dogs are correct	Remove the incorrect names
Twitter archive	Doggo, floofer, pupper and puppo columns has 'None' for missing value	Replace 'None' with 'np.nan' for these columns: doggo, floofer, pupper and puppo
Twitter archive	Some of rating_denominator is different of 10	Remove each tweet that the rating_denominator of it is not equal to 10
Tweet data	id column in TweetData table is not the same name of the other tables	Rename the id column to tweet_id
Twitter archive	The doggo, floofer, pupper and puppo columns should be merge into one column	In twitter archive table: merge these columns into one column (dog_stage)
Twitter archive Image prediction Tweet data	All tables should be merge into one table	Merge all tables