Introduction:

The goal of this project is to wrangle, analyze and visualize data from 3 sources associated with WeRateDogs Twitter account. People send their dogs photos to the account, then the account tweets selected photos with humorous comment and a rating that almost often higher than 10/10. After collecting the data, assessing and cleaning data issues were done. Finally, insights and visualizations were produced. There are in the act_report.pdf document.

Gathering:

Data was gathered from the following sources

1. Enhanced Twitter Archive given by Udacity in csv format.
   The data was stored in a dataFrame called twitter_df.
2. Image Predictions File:
   It was downloaded programmatically using URL given by Udacity and then it was stored in a dataFrame called images.
3. Additional Data via the Twitter API:
   It was retrieved by querying Twitter's API to gather retweet count and favourite count using tweet_id. Then, a dataFrame tweet_json1 was created.

Assessing:

Multiple Quality and Tidiness issues were identified and assessed.

Tidiness issues that were cleaned:

1. Merge all tables since they are information describing one tweet.
2. (doggo, floofer, pupper, and puppo) are 4 different columns for one variable (dog stage).
3. image prediction table has different columns (p1_dog, p2_dog, p3_dog,'p1', 'p1_conf', 'p2', 'p2_conf','p3', 'p3_conf') for prediction and confidence level variables.

Quality issues that were cleaned:

1. "None" string in both "name" and "dog_type" columns.
2. 'retweeted_status_user_id' and 'retweeted_status_timestamp' columns will be removed since some entries are retweets.
3. "tweet_id" column has the wrong data type, it is an integer. It has to be changed to string.

4. 'in_reply_to_status_id', 'in_reply_to_user_id' and 'timestamp' columns have wrong data type.
5. "breed" column has capitalization consistency issues.
6. 'name' column has naming issues.
7. Unstandardized ratings.
8. 'source' column has to be readable.
9. delete columns no longer needed.

Cleaning:

Executing a clean job requires three steps: Define, code and test. During the cleaning process, the following methods and techniques were used:

.head()

.unique()

.info()

.value_counts()

.drop()

. astype()

.str.slice()

.to_datetime()

fillna()

.lower()

.replace()

.apply()

Round()

re.findall()

Storing:

The dataset is stored in a csv file called twitter_archive_master.csv .