WeRateDogs Data wrangling

1. *Introductions:*

This document explaining the process I have taken in order to analyze WeRateDogs tweeter account.

2. *Gathering Data:*

Gathering is the first step of analyzing the data, I have been requested to gather three pieces of data Twitter Archive, Image Predictions and Twitter API as below.

- Twitter Archive has been downloaded manually and added in data frame called ‘archive_df’.
- Image Predictions has been downloaded Programmatically from a given link using request function and stored in data frame called ‘image_predictions_df’.
- Twitter API has been collected from the Twitter using Twitter API and created a Json file, then from this file new data frame has been created called ‘api_df’, I have considered only three columns from this data *Id*, *Retweet count*s and *Favorite counts.*

3. *Data Assessing:*

The second step is to assess the data both visually and programmatically to detect any tidiness or Quality issues and below what I have found.

Quality Issues:

- **archive_df table**
  - tweet_id column should be string not int.
  - Timestamp is an object not a Datetime.
  - Tweets without expanded URL should be drop.
  - All retweeted tweets should be drop.
  - Some rating_denominator values are not equal to 10 as it's should be.
  - Some rating_numerator values are too high values like (1776, etc.) need investigation.
  - Some Dog names are not correct extracted from the text.
  - There is a `None` value in the last 5 columns.
  - Some Doges are not classified and some other classified two times.
  - We should change Dog_Stage type to category.
  - Tweets without image should be dropped.

- **image_predictions_df table**
  - In case all predications are false then these pictures are not for dogs, so we can drop them.
  - Columns name not descriptive so I will replace them
  - Some names in P1 ,P2 & P3 lower case or has _ in between
- **api_df table**
  - Id column name should be match `archive_df`
  - Id column should be string not int.

Tidiness Issues:

- Drop unnecessary columns
- `archive_df` we can combine doggo, floofer, pupper, puppo coulmns in one column as Dog_Stage
- `image_predictions_df` P1, P1_con, P1_dog are repeated three times we need to make it in one coulmn.
- The Three Data frames need to be merged.

4. *Cleaning Data:*

Using the programmatic techniques to clean the data (Define, Code, Test) I have went through each issue to start cleaning the issues for example I have dropped all the tweets that has retweets and created dog stage and breed columns merging the data frames to filter more tweets.