Wrangle and Analyze Data Project

Introduction:

In this project I wrangled data from twitter account **WeRateDogs.**
This account is specialized in rating dogs but it's only start the rating from 10 because they say all dogs are good.

Libraries I used:

1. Pandas
2. Numpy
3. Matplotlib
4. OS
5. Tweepy
6. Requests
7. Json
8. IPython

Project steps:

1. Gathering Data
2. Assessing Data
3. Cleaning Data
4. Visualization and Analyze Data

1. Gathering Data:

I gathered data from 3 different recourses

- **csv file(twitter_enhanced_archive.csv)**
  Was provided as a supported material from udacity
- **tsv file(Tweet image prediction.tsv)**
  This file from a neural networks a table full of image predictions has columns: ID, image URL, and the image number that corresponded to the most confident prediction
- API &json file
  In my case I did not download the data from the API I get the tweet_jason.text from udacity

2. Assessing Data:

After I assessed the 3 data frames I found:

DATA IUSSES

Tidiness

Data atr divied into three dataframes (csv,,tsv,test)

All dataframes are related

Quality

1- Archive data frame:

- There are many useless columns and NAN columns.
- wrong data types ex:timestarp.
- correct numerators and denominator float rate.
- Some denominator rating column not =10.

2- Image data frame:

- massing rows in image data frame has only 2075 and archive as 2356.
- many columns are useless (well be dropped).
- 66 jpg_url duplicated.
- some names are lowercase and some are upper case.
- tweet_id datatype is wrong.

3- text data frame:

- massing rows in text datagram has only 2354 and archive as 2356

3. Cleaning Data:

I cleaned what I found when I assessed the data frames and I merged them into one data frame

4. Visualization and Analyze Data:

From my Visualization and Analyzing I could determine:

a. The common dog stage
b. Common dog Names
c. Relation between Retweets count& Favorite count
d. Top retweeted & favorite dog