Wrangle Report

The Data Wrangling is the 5th project, part of Udacitys Data Analyst Nanodegree, we have 3 data frames to work on them and to assessing them quality and tidiness preparing to create a clean data frame to start analyses and visualizations.

We will take our data from the Twitter (@dog_rates) and the Twitter account WeRateDogs.

First Step: Gathering Data

We gathered the data from three sources:

1. Twitter archive file by name (twitter_archive_enhanced.csv).
2. Downloading programmatically the (image_predictions.tsv) file from Udacitys servers
3. The (tweer_json.txt) file from the Twitter API using Tweepy library.

Second Step: Assessing Data

In this step, we will identify quality and tidiness issues as a step before the cleaning and to know exactly what do we want to change and drop

Quality

csv_file:
* the colums (in_reply_to_status_id) , (in_reply_to_user_id), (retweeted_status_id) , (retweeted_status_user_id) is supposed to be integers instead of float.

* the (retweeted_status_timestamp)and the (timestamp) supposed to be datetime instead of object (string).
* ther is sum sources difficult to read.
* the numerator and denominator columns have invalid values.
* In several columns null objects are non-null (None to NaN).

tsv_file:
* Some of the tweet_ids have the same jpg_url.
* Missing values from some images dataset.

json_file:
* some of the tweet_id are duplicated.

Third step: Cleaning Data

In this step, we cleaned the data from the duplicate rows, ‘None’ and missing values yousing a different code in the jupyter notebook and different libraries like numpy, pandas and tweepy and others, until we got a cleaned Data Frame.