Wrangle and Analyze Data Project

By: Hanan Al-Tuwaijri

Introduction:

Here is brief of my efforts in wrangling "WeRateDogs "Twitter account as project in Data Analysis Nano degree program from Udacity.
In this project, I'll gather, assess, and clean data then act on it through analysis, visualization.

Gathering :

Gather data from a variety of sources and in a variety of formats

Source	Data Type
extract Local file	CSV
Download file from the internet "is hosted on Udacity's servers"	TSV
Twitter APIs	JSON

Assessing:

- Quality

1. There is 181 Retweets in the dataset
2. In the name column Wrong name was taken after (this is) and it's not names, e.g. "a", "an", "in"
3. Wrong data type for timestamp feature
4. There are some tweets beyond August 1st, 2017
5. Source data have an extra HTML format.
6. No need to the retweets columns after filtering
7. Tweets with this ID 835152434251116546, 746906459439529985 have no rating
8. 59 tweets without image (expended URL) image predictions Dataset
9. Some of the image not for dogs

- Tidiness
    1. Column headers(doggo,floofer,pupper,puppo) are values, not variable names.
    2. Data for same tweets is stored in 3 tables (twet_df , df_image , df_twit_arch) .
    3. There is so many Confident columns and breed prediction (**image predictions Dataset**)

Cleaning :

By defining how to clean this issue then, coding the solutions then testing the Cleaned Dataset .

All this steps provided in (wrangle_act.ipynb) jupyter notebook.