Wrangle report – Thanh Dinh

The purpose of this project is data wranging practical. We will collect data which rate the dogs from three sources: incomplete flat file, the tweet image predictions and Twitter API & JSON based on tweet's source; manage to bring the signal over noises and data some initial insight about the cleaned & combined data.

1. Data Collection

We-rate-dogs data was obtained from threes sources:

- twitter_archive_enhanced.csv: This file has been downloaded manually on Udacity's assignment
- image_predictions.tsv: Downloaded programmatically using the Requests library and URL information. The content of this file is about the breed which goes along with picture to be able to predict dog's types.
- Twitter API & JSON: Captured entire of tweet's comments about @weratedog. I used my tweet's account to extract useful information & save it to a file name tweet_json.txt. be noted that my tweet's authentication will be then remove from the submitted assignment.

After extracting data from three sources, I will combine them into one file only to take into the next stage (I call this file df_combined)

2. Data Cleaning

In this stage, the combined data, which has been obtained from previous stage, would be cleaned up & verified. The outcome of this stage is the reliable, cleaned, more pattern of information about @weratedog competition on Tweet.

The combined data from the first stage would be taken into consideration in order to see if we're able to enhance to quality of this data. (Checking duplication, data information, shape of data…)

Based on the data exploration process, there are some points which I have applied to tackle the issues of data quality:

- Drop down rows do not have picture include
- Drop down re-tweets rows
- Convert Timestamp to a Datetime to made it human-reaing
- Revock other nolonger fields (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)
- Define new 'rating' field as numerator and denominator made not much sense
- Set null values for multiples fields
- Tackle name issue as there are lot of dogs named "a"

And structural issues:

- The 'Unnamed: 0' field fell in to this file by the change. Need to get rid off
- There are quite a lot of fields relate to dog breed's prediction. Those need to be combined

3. Data Analytics

Data wragling can be easy. But, if we done have much attention on what we are doing, some important points will be washed away. A good data analyst will tell a good story about data based on what about data pattern which we want to extract.

To rate dogs, there are some points I'm concerned in are:

- Which facets made a high rate dog, mean favourite dog? Is it because outnumber of tweeter rate & re-rate for it?
- People want to see the ranking of breed, which made more sense
- Is there any one who is crazy to overrate in any kind of dog.

The story telling would be written in act_report