Data Wrangling Steps Report by Anas A. Mohaisin

Data wrangling for "WeRateDogs" twitter account, divided into three stages:

I. Gather
II. Assess
III. Clean

Data gathering stage concerned with reading date from three sources: Twitter-Archive-Enchanced.csv (provided by Udacity), Tweet-Json.txt (provided by Udacity), and Twitter API Image-predictions.tsv through Requests and Tweepy library. We have accessed Twitter data without actual creation for a Twitter account.

Then moved moved to the next stage "Assess" where we have drawn five samples from each Data-Frame; df.sample(5); and explored the variables types df.info(); and the descriptive statisticas df.describe(). At this stage, we have identified data quality and tidiness, as follows:
Quality Issues:

- df_enhanced_twitter_archive
    1. tweet_id is int
    2. timestamp` is obj/str
    3. drop demominator zero to avoid ∞∞ results
    4. name contains invalid name such as a, the, an
    5. timestamp has +0000
    6. missing data in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id,retweeted_status_user_id, and retweeted_status_timestamp
- df_image_predictions
    1. Extract source type from links in source column

2. tweet_id is int
3. rename P1, P2, P3 columns
4. remove URLs from text

- df_json
    1. tweet_id is int
    2. rename id,retweet_count,favorite_count

Tidiness Issues:

1. Each dog stage has its own column.
2. Join gathered dataframes into a single dataframe.

To make sure that the Cleaning process is correctly implemented, we had three strict procedures: Define, Code, then Test.

At this stage we have begun with a copy creation from the imported assessed data-frames then we dropped, cleaned, and extract data to prepare them for merging. Data merging concerned with a single data-frame based tweet_id as the primary key. Mathematical we have interactions of three sets:

$$n(A \cap B \cap C) = n(A) + n(B) + n(C), where\ n\ is\ the\ cordinality\ of\ the\ set$$

Thus, we have got 2073 records in one data-frame df_enhanced_cln.

We have started with replacing dog stages columns with one column dog_stage by using Melt command which, necessarily, created duplicates. Dog stages are divided into four types: doggo, pupper, puppo, and floofer. maked Initial data-frame had 2073 records and the new melted records df has 8292; i.e. 4x. Therefore, we have dropped additional and duplicated records.

The last thing we have done in this stage is data type conversion.