Wrangle Report

This document has been created to briefly describe my wrangling efforts. This is to be framed as an **internal** document.

Gathering Data

Frist, I had to import the archived flat file from the local disk, called "twitter-archive-enhanced.csv".

Then from the Udacity website using requests library download another flat file. Last, gather tweet data from the API called “tweety”. I saved this dataset into a text file, because it takes a lot of time to get data from the API every time the Kernel restarts. So there is an option to get the data from the text file.

Asses & Clean

I have found 8 quality issues and 3 tidiness issues.

Quality

I have found some unrealistic dog names, like “a”, “an”

I changed these names into “none”.

There are some erroneous data types only in the archive table

- Float ID columns to integer:
    1. Replace NaN values with 0
    2. Change float type to integer
    3. Change integer types to string
    4. Fill missing values with None (now it is 0)
- Change columns that display date time, from string to timestamp

Retweets should not be there in the dataset, and there are a lot of them. There are duplicated and missing jpg URL values.

We can solve all of these 3 assesses with the same solution. Select all retweet ID and remove them from the archive table. Duplicated and missing jpg URLs are the causations of the retweet issue.

Non descriptive column headers

Change column names, **from** *p1, p1_conf, p1_dog* … p3_dog **to** *prediction_1, prediction_1_confidence, is_prediction_1_dog.* I did this by simply call the rename function.

Snake case dog predictions

Words are separated by “_” not “ “. I changed every “_” character with the blank space (“ “) character. This solved the problem.

Tidiness

Text and twitter link in the same column (text column), fix numerator issues

Separate these 2 values with regex (regular expression).

With also regex, separate the numerators, then copy and paste them into the dataset. This will fix the numerators that have decimal values.

Doggo, floofer, pupper, puppo columns are variable names

1. I simply made a new list, storing the dog “types” in it by every row. (Every row can have only one type in the original table, so this is why I could use a list).
2. Then I deleted the original columns and connect the new list to the data frame.

Unuseful tweet_count table

1. Add its columns to the archive table by Pandas.merge function.
2. Drop the duplicated id column (good to have first for test reasons).

Add image prediction info to the final dataset

Because they are related to the same observation, I merged them into the archive table.