Siddhant Sareen

Udacity’s Data Analyst Nanodegree

Wrangle and Analyse Data

# Data Wrangling Report

## Gathering Data

Data Gathering is done through three steps :

- Downloading given file as provided by Udacity (twitter-archive-enhanced.csv) and loading this into a Dataframe (df)
- Downloading an image predictions file using the requests library and the link provided within Udacity resources, this file was written into a separate Dataframe (img_df)
- Downloading additional data from the Twitter API to obtain mainly retweets and favourites for corresponding tweet id’s , a separate twitter developer account was created for access to consumer and access tokens and keys.

## Assessing Data

Data assessment was done using two methods:

- Visual Assessment - The data frames were looked at as a whole to get a feel of all the columns and their meanings
- Programmable Assessment - Functions such as info, describe, length and value counts were used to get a closer look at retweets, missing values and incorrect datatypes.
- Quality and Tidiness issues are documented in this section.

Data Cleaning

These Quality and Tidiness issues were cleaned and Define, code and testing phase was documented.

***Quality***¶

- contains retweets and therefore, duplicates
- many *tweet_id*(s) of df table are missing in img_df (image predictions) table
- change in datatypes (*in_reply_to_status_id, in_reply_to_user_id and timestamp* columns)
- unnecessary html tags in *source* column in place of actual source name
- *rating_denominator* column has values other than 10
- replace dog names starting with lowercase characters (e.g. a, an, actually, by)
- some records have more than one dog stage

**Tidiness**¶

- df without any duplicates (i.e. retweets) will have empty *retweeted_status_id, retweeted_status_user_id* and *retweeted_status_timestamp* columns, which can be dropped
- *doggo, floofer, pupper* and *puppo* columns in df table should be merged into one column named "*stage*"
- *retweet_count* and *favorite_count* columns from status_df (tweet status) table should be joined with df table