Data Wrangling Report

Abha Ramchandani

May 10, 2020

Data Gathering

I gathered data from 3 sources and created 3 separate data frames for these datasets.

1. Twitter Archive Data provided in CSV format was stored in `twtr_archv_df`.
2. Downloaded Tweet Image Predictions from the url - url - and stored the TSV data in `img_prdctns_df` programmatically.
3. For the 3rd source, Twitter API Data, we had the option of using Twitter Developer Account to download the Tweet Data from Twitter or use Udacity provided tweet-json.txt file from here. I chose to use the Udacity provided file, since I have still not received approval for my Twitter Developer Account and I have used APIs in the past for Udacity's Full Stack Nanodegree course. I also read through Udacity provided code twitter-api.py to extract data using Twitter API. This data was stored in `tweets_df`.

Assessing Data

I used visual and programmatic inspection to identify following issues in the datasets:

(Source #1: Archive Data, Source #2: Predictions Data, Source #3: API Data)

Quality Issues

- (Archive Data) Retweets are included in the dataset. This means there are duplicates.
- (Predictions Data) The lower number of entries in Predictions DF compared to Archive DF means that some posts don't have images
- (Archive Data) Incompatible data types - *in_reply_to_status_id and in_reply_to_user_id must be integer; and timestamp must be datetime*
- (Archive Data) source column must tell us the source of the tweet and not have any unnecessary HTML tags
- (Archive Data) text column should contain only introduction and rating and show full text, which is not the case. It also includes a short version of some link
- (Archive Data) *rating_denominator* column has values other than 10
- (Archive Data) *rating_numerator* column has values less than 10
- (Archive Data) Incorrect Dog names, e.g., a, an, etc.

- (Archive Data) More than 1 stage (doggo, floofer, pupper or puppo) applies to some Dogs

Tidiness Issues

- (Archive Data) doggo, floofer, pupper and puppo columns in the table should, ideally, be merged into one column called 'stage'
- (Predictions Data) p1, p2 and p3 columns in the table should, ideally be merged into one column called 'breed' and then merged with Archive Data
- (API Data) retweets and favorites columns from the table should be joined with archive data table

Cleaning Data

Most of the issues identified corresponded to `twtr_archv_df`. I worked through one issue at a time and cleaned the data programmatically. For each issue, I stated the issue, defined the solution, built code to fix the issue and tested my solution. These steps were performed on copies of dataframes created during data gathering: `cln_twtr_archv_df, cln_img_prdctns_df, cln_tweets_df`.

Storing Data

DataFrame `cln_twtr_archv_df` contained the clean data and this data was exported to `twitter_archive_master.csv` file.