Wrangling Report - @WeRateDogs on Twitter

Background

This report is an analysis of the data wrangling process - gathering, assissing, cleaning, analyzing and visualizing data for the @WeRateDogs Twitter Project.

Datasets analyzed were:

- **twitter-archive-enhanced.csv** - An enhanced twitter archive basic tweet data pulled from tweet text
- **tweet_json.txt** - Additional tweet data which I downloaded from the Twitter API
- **image-predictions.tsv** - Image prediction data on what kinds of dogs were in the photos, as well as dog categories (dogger,fluffo,pupper,puppo) and names extracted from tweet text using an ML model

Data Wrangling Process

Gathering

I gathered the data from 3 data sources both manually & programmatically :

- manually downloaded the enhanced twitter archive from Udacity (twitter-archive-enhanced.csv)
- programatically downloaded image prediction data .tsv file from Udacity(image-predictions.tsv)
- programatically downloaded Twitter data using the Twitter API using my credentials (tweet_json.txt)
- programatically downloaded Thumbnail images from Twitter in order to visualize composite representations of machine learning model results in a single image.

Assessing

I explored the datasets visually as well as programmatically. During this process I started to get to know the data and start to form ideas about which pieces of data would be relevant to my project.

During this process, I became most interested in : dog names, what the ML model predicted for dog breeds (and how funny it was to see what it predicted when it was wrong!), interesting dog categories like (dogger, fluffo, pupper, puppo) and how prevalent they were, what languages were indicated in tweet data, and the ways dogs are rated (numerator / denominator).

Once I had an idea of which data might be the most relevant to me, I focused on noting what issues would need to be fixed in the cleaning phase. As explained in the course, I had to come back to the assessing part of the project many times as my questions and assumptions about the data changed. Items to change from the assessment are described in the Cleaning section below.

Cleaning

Some of the ways that I cleaned the data:

- Removed dog names that were very unlikely like : "a, the, and, etc"
- Converted columns like dogger,fluffo,pupper,puppo I converted to 1 & 0 values so that I could count them.
- Renamed columns for consistency across tables so that I could merge them, ie. created_at -> timestamp, id -> tweet_id
- Renamed columns so that they are easier to interpret - (ie p1 as pred_1, lang as language, etc)
- Capitalized breed names and removed underscores.
- Removed columns that I wasn't interested in reporting on, or that didn't have much data for ie. in_reply_to_status_id, in_reply_to_userid, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls
- Merged the df_archive & df_twitter_api datasets on tweet_id and timestamp
- Fixed numerating column to reflect correct value gathered from the text column

Cleaned data sets.

- ./Data/dfa_clean.csv
- ./Data/dfi_clean.csv

Analyzing and Visualizing

Once the data was cleaned I analyzed and visualized it using Pandas in the Jupyter notebooks. It was quite challenging for me to figure out how to use Seaborn and Matplotlib at times, however I managed to create some interesting graphs with it. I had the most fun visualizing the machine learning model predictions by filtering on correctly and incorrectly identified dog breeds, then creating composite images of the results vs. creating charts representing things numerically.