Wrangle Report

Mariam Abouzeid

Absolutely the most challenging project in this nanodegree so far. I've had so much fun and learned a lot from it.

Dealing with actual data was such an experience. I got to see real data problems and how to solve them by passing through the wrangling process.

1. Gathering Data

First of all, I have started with gathering data. I used the data provided by Udacity and Twitter API, parsing the JSON files and started to work on them. These files are (twitter-archive-enhanced.csv, image_predictions.tsv and tweet_json.txt)

2. Assessing Data

After importing the data into dataframes, I started to take a closer look to the data and see what its problems is

2.1. Twitter Archive data:

- ids columns must be int/string not fload
- retweeted_status_timestamp, timestamp should be datetime not string
- the numerator and denominator columns have invalid values
- in several columns null objects are non-null (None to NaN)
- name column have invalid names i.e 'None', 'a', 'an' and less than 3 characters
- we dont need tetweeted tweets, only the original ones
- We may want to drop in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp columns as we won't do operations on them and they have a lot of null values

2.2. Image Prediction data:

- the dataset should be 2356 records instead of 2075, which means that there is missing data
- dataset contains retweets
- some records their tweet_ids is the same jpg_url

2.3. Tweets data:

- the tweet_id (666020888022790149) duplicated 8 times

2.4. General Structure problems:

- we need to join tweet info and images datasets with tweet archive dataset

- dog stages doggo, floofer, pupper and puppo are divided in 4 columns
- we only need tweet_id and jpg_url from image dataset

3. Data Cleaning

After taking a closer look at the data and notice its problems, I have started to perform data cleaning and followed these steps

- Copy every dataframe to a new one, where we will do our operations into
- Join tweet info and image predictions dataframes to tweets archive
- Merge dog stages doggo, floofer, pupper and puppo in one column
- Clean the dataframe from duplicated IDs, tweets with no images, retweets and drop unnesseccary columns
- Fixing the rating columns (rating_numerator and rating_denominator)
- Before dropping image prediction data except tweet_id and jpg_url, we need to store the "True" prediction and its confidence, looking for TRUE prediction values and append in the px value and its ps_conf values in the list, else, append NaN
- After calcultaing prediction and confidence, remove the unnesseccary columns from the image prediction table
- Extract from the dataset dogs gender and add column for gender
- Adjust datatypes

4. Export Data

After cleaning the data, I exported it to a csv file so I can move to the next step. Visualizations.