Wrangle report

Introduction

In this project, I have wrangled data about the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a comment about the dog.

Gathering Data

I have gathered from three different resources:

1. I have downloaded the file WeRateDogs Twitter archive manually by clicking the following link: twitter_archive_enhanced.csv. Then, I imported data into a DataFrame (twitter).
2. I have downloaded programmatically the file image_predictions.tsv, which is hosted on Udacity's servers, using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv. This file contains tweet image predictions, i.e., what breed of dog is present in each tweet according to a neural network. Then, I imported data into a DataFrame (twitter_predictions).
3. I have gathered third part of data using the tweet IDs in the WeRateDogs Twitter archive, queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called `tweet_json.txt` file. I gathered data about tweet ID, retweet count, and favorite count.Then, I imported data into a DataFrame (twitter).

Assessing Data

I performed assessing using the following methods:

- head()
- value_counts()
- info()

- sample()

I have detected eight quality issues and two tidiness issues:

- Table Twitter contains useless data about retweets.
- The timestamp column in twitter table is an object, not a datetime.
- The Twitter table contains 2356 rows, whereas twitter_predictions table contains 2075.
- Missing values in Twitter table: retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.
- The tweet_id is an integer, not an object.
- Table Twitter contains useless columns, i.e. in_reply_to_status_id, in_reply_to_user_id.
- Some values of rating_numerator in the Twitter table are inaccurate.
- Some values of rating_denominator in the Twitter table are not equal to 10.

I also detected two tidiness issues:

- Information about one type of observational unit (tweets) is spread across three different dataframes. So these three dataframes should be merged as they are part of the same observational unit.
- Doggo, floofer, pupper, puppo columns in twitter table should be one variable that identifies the stage of dog.

Cleaning Data

I have cleaned each of the issues that I documented while assessing. Each problem was solved in three steps: define, code and test. I perform the cleaning using the following methods: to_datetime(), isnull(), isin(), astype(), drop(), iteritems(), melt() and merge().

Storing Data

The result was saved into CSV file twitter_archive_master.csv.