Wrangling Report

Introduction

The gathered data, resided in a total of three (3) DataFrames,

- An archive of Tweets from the account: “weRateDogs”
- The results of an image predictions algorithm
- Further Twitter data such as retweet count, and favorite count

In order to wrangle the data, it was first visually assessed by printing out the data, and observing the structure of each DataFrame, and the content of the different columns.

After which the regular DataFrame functions such as head, tail, sample, info, and shape were ran on the data to gain further insight into the structure of the gathered data, and how said data is represented.

Some of the Quality issues found are:

- Not all tweets have pictures of dogs,
- Not all tweets are original tweets (some are retweets),
- Not all tweets are of dogs,
- Not all tweets have a Dog Classification, that follows the “Dogtionary “, and some tweets have more than a single classification.
- Many Values are missing,
- Values such as Timestamps are presented as strings,
- Tweet Source (device) is represented as a link.
- Missing Classifications are entered as None, which is misleading.

Aside from quality issues, there are tidiness issues as well, a sample of those issues are

- Classification Values are columns in the archive DataFrame
- Predictions are multiple columns in the Predictions DataFrame.
- Their exists 3 DataFrames, for the Data, that is all referenced by the same value (tweet_id), and no extra value is gained from having three DataFrames.
- Column Names are not very descriptive.

Fixing the quality and Tidiness issues is documented within the attached Jupyter notebook “wrangle_act.ipynb”, in details, organized in the three phases of Defining the issues, coding a fix, and finally testing that fix. Yet the following shall skim over some of the steps taken to clean the data.

Fixing Quality Issues

Cleaning was split into two directives, the first reducing the data present to clean, that was achieved by ensuring data are of original tweets and had pictures.

Therefore, first we delete all tweets from the archive DataFrame that are not present in the image_predictions DataFrame.

Next We delete all retweets, by checking the fields

- in_reply_to_status_id

- in_reply_to_user_id
- retweeted_status_id
- retweeted_status_user_id

if they are NaN, the tweet is original, otherwise it's a retweet, and deleted from all dataframes.

The second cleaning directive deals with some of the quality issues above such as:

- Dog Classification uses None (String) instead of empty string, which can be corrected by replacing None, with an empty string,
- also Tweet source record include both the source, and a link for downloading the source application. this issue can be fixed by means of looking for and removing the link part.
- Tweet time is represented as a string, instead of a timestamp. This is fixed by means of converting the DataFrame type for the timestamp record.
- Not all Classifications, are Mutually Exclusive, this issue is resolved by dropping multiple classification records, unless the classifications are doggo and pupper (according to the dogtionary) hence for those records, we can discard one of the classifications.
- Not all Tweets are of dogs this issue can be fixed by means of dropping rows of archive_df, and subsequent tables that include the statement "Please only send in dogs"
- Some predictions are false for all 3 attemps, those are fixed by dropping the record.

Fixing Tidiness Issues

Tidiness issues revolve around merging all DataFrames, renaming columns to be more descriptive, melting columns so that observations are not columns, and dropping columns that does not help with analysis