Wrangle Report

Introduction

The purpose of this report is describing the wrangling efforts for this project

Details

Provided data:

The data to be wrangled was provided in 2 files, twitter-archive-enhanced.csv (The WeRateDogs Twitter archive) and image-predictions.tsv (The result of running every picture attached to a tweet through 3 different neural networks to classify dog breeds)

Wrangling process:

First, Twitter API was used to collect more data, the twitter data was provided in json format.

The Wrangling process was followed starting from gathering where data was collected from sources and imported to Pandas data frame, then assessing, both visually and programmatically for both quality and tidiness issues using different pandas function.

The found quality and tidiness issues was cleaned using pandas functions too.

The following quality issue was noticed:

- Many data were assigned wrong data type (i.e. tweet id , in_reply_to_status_id and in_reply_to_user_id) in twitter archive data
- Some dogs name are missing
  dogs names, some are not correct , other are null
- rating_denominator values other than 10
- some dogs with no dog stage
- some dogs with 2 dog stage
- null values represeted as the string "None" in multiple colomns (doggo,floofer,pupper,puppo)
- tweet_id sould be string instead of integer in dog prediction data
- p1, p2 , p3 columns names are not descriptive enough
- possibly_sensitive_appealable and possibly_sensitive are string, it should be Boolean in twitter data collected using API

The following tidiness issues were noticed

- One variable in 4 colomns: doggo, floofer , pupper , puppo should be one column
- Multiple dog predictions available in different columns.
- id , in_reply_to_status_id and in_reply_to_user_id are all duplicated in the json_df
- All three tables should be combined

One assessing is done, the cleaning phase started to deal with the noticed issues for only those that is useful to the analysis.

The selected data was combined from the 3 tables into single one.

Finally, the cleaned data was stored to a csv file (twitter_archive_master.csv)