Wrangle Report

Introduction

The purpose of this project is to wrangle twitter data from WeRateDogs to create interesting and accurate analysis and visualization.

Project Details

The tasks in this project are as follows:

- Data wrangling, which consists of:
  - Gathering data (downloadable file in the Resources tab in the left most panel of your classroom and linked in step 1 below).
  - Assessing data
  - Cleaning data
- Storing, analyzing, and visualizing your wrangled data

1. Data Gathering

gathering the data from various sources

- WeRateDogs Twitter Archive:

download "twitter_archive_enhanced.csv " file manually

- image_predictions.tsv:

is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

- additional data via Twitter API:

each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

2. Data Assessing

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues

Quality issues :

- ### twitterarchive_clean

1. There are useless rows that have non-empty retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp
2. [in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp] are useless columns
3. Null value in expanded_urls
4. The type of timestamp is object
5. The type of tweet_id is int64
6. the type of rating_numerator is int64
7. the type of rating_denominator is int64
8. There are rows whose name == a, an, the, such

- ### *imgpredictions_clean file*

The type of tweet_id is int64

- ### *getdata_clean file*

The type of id is int64

Tidiness issues :

1. variable in 4 columns in df_twit_archive table (doggo, floofer, pupper, puppo)
2. image_pred dataset condence the columns p1,p1_dog_p1_conf,...etc to dog_breed, confidence
3. tweet_json and image_pred datasets should be part of our main dataset twitter_archive.

3.Data Cleaning

Clean each of the issues you documented while assessing

First thing first

The dataframes are copied to new dataframes before the cleanup begins

Clean procees phase:

Define: **act as an instruction list**

Code: **convert those definitions and run the code**

Test: **test if the dataset cleaning operations have performed**

The end: stored the wrangled data in twitter_archive_master.csv file

Analyze and visualize your wrangled data