Data Wranlgling Report

Data Collection/Gathering:

The dataset I'll be working with is the Twitter archive of @dog_rates (https://twitter.com/dog_rates), also known as WeRateDogs. A total of 2356 basic tweets are included in this archive/dataset from November 2015 to August 2017. An additional dataset was generated based on the images from the above dataset (I.e., WeRateDogs Twitter archive) that includes image predictions (top three only), each tweet ID, image URL, and the image number (numbered 1 to 4) for their most confident prediction.

Tsv Data

-- To gather tweet image predictions, I downloaded the tweet image predictions data from Udacity's servers using Python's Requests library. Following that, I imported the file into a Python Pandas DataFrame (dog_pred). URL:(https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)

Twitter API alternative(JSON.txt) Data

-- As an alternative to using Twitter API, I accessed the entire tweet summary by reading the JSON data provided by Udacity. From this JSON, I created the Tweet_data DataFrame, which contained only tweet_id, retweet_count, favorite_count, and create_date.

CSV Data

--In order to collect the Twitter archive, I used the link provided by Udacity to download the Twitter archive in CSV format. I then read the (twitter-archive-enhanced.csv) using Pandas directly into a DataFrame (Twitter_archive) URL: https://video.udacity-data.com/topher/2018/November/5be5fb7d_tweet-json/tweet-json.txt

Assessing Data

Visual Assessment

Progammatic Assessment

My visual and programmatic assessment of the data indicated that there were particular issues with the quality and tidiness of each piece of data

QUALITY

Twitter_archive

--Timestamp and retweeted_status_timestamp datatype should be datatime

--[tweet_id, retweeted_status_id, retweeted_status_user_id] should be object and not float[avoid operation being perform]

--floofe, pupper,puppo,doggo has alot of missing value, indicated as None instead of NaN

--Name has irrelevant names like "a","not","all","by","the","my" etc

--there are more than 1 dog stage e.g doggo and pupper(12), doggo and floofer(1), doggo and puppo(1)

--Duplicated value in expanded_url

Tweet_data

--tweet_id, create_date has incorrect datatype

dog_pred

--duplicated values in jpg_url

TIDINESS

Twitter_archive

--[Doggo, puppo, pupper,floofer] should be in one column as "stages"

--source having unnecessary html tags 'Twitter for iPhone'

--in_reply_to_status_id and in_reply_to_user_id are not original data

--Only the original tweet is needed, drop all retweet columnsi.e retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp

Tweet_data

--create_date exist in Twitter_archive as timestamp. therefore, create date should be drop

dog_pred

--Ensure that all rows contain at least one true prediction

Cleaning Data

A copy of each DataFrame was created (Twitter_archive_df, Tweet_data_df, and dog_pred_df).

In order to fix each quality/tidiness issue, I followed a three-stage model of programmatic data cleaning - Define, Code and Test.

Storing Data

Following the cleaning process, I merged the 3 cleaned datasets into 1 using pandas.merge(how=inner, on=tweet_id) library.

Then, I saved the file as a CSV

In [1]:
```
#pip install nbconvert[webpdf]
```

In [2]:
```
#jupyter nbconvert --to webpdf wrangling_report.ipynb
```

In [ ]:
```

```