Introduction

The dataset I have wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. This Report will briefly describe my Wrangling efforts done on the data associated with WeRateDogs Twitter account.

Image via Boston Magazine

Data Wrangling Process

1- Gathering Data

In this project I gathered three pieces of data as described below in a Jupyter Notebook titled wrangle_act.ipynb:

1) The WeRateDogs Twitter archive. This was a kind of a file on hand. I download this file manually by clicking on the link provided on the project details page. The file was named as : twitter_archive_enhanced.csv

2) The tweet image predictions. That is associated with predictions regarding what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) was hosted on Udacity's servers and I downloaded it programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3) The third piece of data was supposed to be gathered using Tweepy to query Twitter's API for additional data beyond the data included in the WeRateDogs Twitter archive. This additional datawas required to include retweet count and favorite count. Gathering data using the Twitter API was requiring me to get a Twitter Developer Account. However, I applied (three times from three different twitter accounts) for a Twitter Developer Account , but unfortunately my applications were NOT APPROVED

So as a Plan B, I accessed the Twitter data without actually creating a Twitter Developer account By Downloading the two files that were included on Twitter API Page in Udacity's Project Section:

1. **twitter_api.py:** I copied the code that was included in this python document and pasted it on a Code Cell on my Jupyter Notebook wrangle_act.ipynb.
2. **tweet_json.txt:** The data included in this txt file are the one that was supposed to be the resulting data from successfully running the code included in the twitter_api.py. Then , I read this tweet_json.txt file line by line into a pandas DataFrame to get include retweet count and favorite count data.

At this point, I was successfully gathered the three pieces of data and was ready to move on to the Second Step: Assessing Data.

2- Assessing Data

I assessed the gathered data both visually and programmatically for quality and tidiness issues. The Programmatic Assessment required using a range of Pandas functions and methods such as : .info(), .describe(), .head(), .sample(), .value_counts() and .unique(). I detected and documented **a total of 20 quality issues and 3 tidiness issues** in my wrangle_act.ipynb Jupyter Notebook.

Detected Quality Issues

In twitter_arch Dataset:

- Missing values in most of the cells of the `in_reply_to_status_id` and `in_reply_to_user_id` columns (78 instead of 2356).
- Missing values in most of the cells of `retweeted_status_id`, `retweeted_status_user_id` and `retweeted_status_timestamp` columns (181 instead of 2356).
- We are ONLY interested in ((ORIGINAL)) tweets of DOG ratings i.e. we don't want any retweet or reply tweet to an original tweet.
- We are ONLY interested in ORIGINAL tweets of DOG ratings ((that have images)) i.e. we don't want any tweet without an image/images.
- Erroneous datatypes (`timestamp` and `retweeted_status_timestamp`).
- Object is a better datatype for `tweet_id` column.

- All cells that belongs to `rating_denominator` column should have the value of 10.
- Inaccurate values in the `rating_numerator` column i.e. some ratings are not out of a rating_denominator 10.
- Misrepresenting the null values in columns (`doggo`,`floofer`,`pupper`,`puppo`) i.e. Nulls represented as "None".
- Content of the `source` column is too long and can not be analyzed easily.
- Irregular and illogical values in the`name` column.
- Some entries in the `name` column are in lowercase.
- Misrepresenting the null values in the `name` i.e. Nulls represented as "None".
- Undescriptive label for the `name` column.

In img_predict Dataset:

- "True" and "False" values in the `p1_dog`, `p2_dog` and `p3_dog` columns are not that much handy in the analysis.
- Some images are not for dogs while we are ONLY interested in "dogs" ratings.
- Object data type is a better datatype choice for `tweet_id` column.
- Undescriptive columns' labels for `p1`, `p2`, `p3`, `p1_conf`, `p2_conf`, `p3_conf`, `p1_dog`, `p2_dog` and `p3_dog`.

In tweet_json Dataset:

- `tweet_id` as a column label instead of `id`.
- Object data type is a better datatype choice for `id` column.

<u>**Detected Tidiness Issues**</u>

- The values of four columns (`doggo`, `floofer`, `pupper`, `puppo`) in `twitter_arch` table should be represented in one column `dog_stage` with a `category` datatype.
- `tweet_json` and `img-predict` tables need to be part of our main dataset `twitter_arch`.
- `rating_numerator` and `rating_denominator` columns in `twitter_arch` table should form one column `rating_out_of_10`.

3- Cleaning Data

I grouped the issues according the priopritiy into three groups **: Missing Data , Tidiness and Other Quality Issues.** It is important to state here that some of the missing data and tidiness issues required the cleaning of some other issues first . So I tried my best to think logically and clean the issues in preplanned and prioritized ordered. I cleaned each of the issues I've documented while assessing.

The following are some of the Resources I used to successfully Clean the Data

- https://stackoverflow.com/questions/10665889/how-to-take-column-slices-of-dataframe-in-pandas
- https://pandas.pydata.org/pandas-docs/version/0.23.4/generated/pandas.Timestamp.html
- https://www.akc.org/dog-breeds/?letter=O
- http://pbpython.com/pandas_dtypes.html
- https://stackoverflow.com/questions/11346283/renaming-columns-in-pandas
- https://stackoverflow.com/questions/44869327/find-index-of-all-rows-with-null-values-in-a-particular-column-in-pandas-datafra
- https://stackoverflow.com/questions/46893386/as-typecategory-not-yielding-the-desired-datatype-change-from-float64
- https://stackoverflow.com/questions/25125168/array-shape-giving-error-tuple-not-callable
- https://www.geeksforgeeks.org/python-pandas-dataframe-rename/
- https://pandas.pydata.org/pandas-docs/version/0.23/generated/pandas.to_datetime.html
- http://pbpython.com/pandas_dtypes.html
- https://stackoverflow.com/questions/15891038/change-data-type-of-columns-in-pandas
- https://stackoverflow.com/questions/38101009/changing-multiple-column-names-but-not-all-of-them-pandas-python
- https://stackoverflow.com/questions/29960733/how-to-convert-true-false-values-in-dataframe-as-1-for-true-and-0-for-false
- https://kaijento.github.io/2017/04/22/pandas-create-new-column-sum/
- https://stackoverflow.com/questions/50847374/convert-multiple-columns-to-string-in-pandas-dataframe

Important Note : I assessed and cleaned (if necessary) the data upon which my analyses and visualizations are based.

My Final product from this Data Wrangling Process was a high quality and tidy master pandas titled: twitter_archive_master.csv containing the combined and cleaned data.