WeRateDogs Wrangle Report

Udacity Data Analyst Nanodegree
Jean Carlos da Cruz

We had access to three different data sources to work on this project. Which are:

- WeRateDogs™ Twitter Archive (twitter-archive-enhanced-2.csv) File downloaded from Udacity repository.
- Tweet image predictions (image_predictions-3.tsv) File downloaded from request to Udacity URL repository.
- Additional Twitter data (tweet_json.txt) File downloaded from Udacity repository.

I **gathered** the data from udacity repository for all three needed data sources. I decided to not use Twitter API since I don't want to have a Twitter dev account. It was very smart from Udacity for providing the dataset separately. After gathering data I then **assessed** it trying to find issues related for both quality and tidiness. I was able to find the following issues to each dataset:

WeRateDogs twitter archive Assessing Summary

Quality

- tweet_id is integer and should be str, since is a categorical information;
- retweeted_status_timestamp is object and should be datetime;
- Columns related to reply and retweets have lots of missing data and they should be removed;
- Not all tweets could be defined as doggo,floofer, pupper or puppo and all columns;
- name has values defined wrongly as "a", "an" and "None";
- Only 17% of tweets have dog classification
- Some of the ratings from both rating_numerator and rating_denominator are different from the original tweet

Tidiness

- doggo, floofer,pupper and puppo should be in one column.

Tweet image predictions data Assessing Summary

Quality

- tweet_id is integer and should be str, since is a categorical information;
- Data contains retweets since there is the same picture in different tweets;
- There are pictures in tweets that are not dogs;
- p1,p2 and p3 show "_" instead of space in the names;
- Predictions have inconsistent casing;

Tidiness

- the prediction and confidence columns should be reduced to two columns considering the one with the highest confidence (dog)

Additional Twitter data Assessing Summary

Quality

- tweet_id is integer and should be str, since is a categorical information;

After I found the issues listed above I started the **cleaning** process.To make it more clear I defined the following steps of cleaning:

Cleaning Steps:

- 1. Create one single source with all data combined;
- 2. Remove tweets that are actually retweets;
- 3. Remove empty and unnecessary columns;
- 4. Fix the wrong datatypes of the columns;
- 5. Fix the wrong numerators and denominators;
- 6. Remove the "None" out of the doggo, floofer, pupper and puppo column and combine them into one column;
- 7. Remove the wrong names of name column;
- 8. Reduce the prediction columns into two based on the highest confidence;
- 9. Remove tweets where the prediction was not a dog;
- 10. Fix the predictions that have inconsistent casing (mix of uppercase and lowercase)

After cleaning I then saved the cleaned dataset to a new csv file called twitter_archive_master.csv that can be found in this repository. The final dataset has the following structure:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1443 entries, 0 to 1442
Data columns (total 16 columns):
tweet_id            1443 non-null object
timestamp           1443 non-null datetime64[ns]
source              1443 non-null object
text                1443 non-null object
expanded_urls       1443 non-null object
name                1035 non-null object
retweet_count       1443 non-null int64
favorite_count      1443 non-null int64
jpg_url             1443 non-null object
img_num             1443 non-null int64
new_numerator       1443 non-null float64
new_denominator     1443 non-null int64
dog_class           216 non-null object
breed               1443 non-null object
conf                1443 non-null float64
check_dog           1443 non-null bool
dtypes: bool(1), datetime64[ns](1), float64(2), int64(4), object(8)
memory usage: 170.6+ KB
```