WRANGLE REPORT: WE RATE DOGS PROJECT

Overview

This project aims to conduct useful information about dogs. Some questions that we are interested in finding the answers are:

1. Which stage of the dogs is the most preferred?
2. Which predicted dog'breed have highest rating?
3. Which is the most common name?
4. Which breed has features that can be predicted with highest confidence?

To answer these questions, data is the most important and it will be gathered, assessed and cleaned before doing any analysis and visualization. This report will describe the gather, assess and clean processes.

Gather

Data is gathered from three sources:

1. The WeRateDogs Twitter archive: `twitter_archive_enhanced.csv`

This dataset is downloaded manually from the link

https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv)

and loaded to the `twee_arx` table.

2. The tweet image predictions: `image_predictions.tsv`

This dataset is downloaded programmatically using the `request` library from the link

https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)

and loaded to `image_pred` table

3. WeRateDogs Twitter: `tweet_json.txt`

This dataset is downloaded from WeRateDogs Twitter using Twitter API and `tweepy` library as follows:

1. Use *tweet_id* in `twee_arx` table. For each *tweet_id* using `api.get_status` to pull out all information related to the status.
2. Wrtie the JSON data of the status to `tweet_json.txt`, each tweet in one line.
3. Handle the exceptions, usually because the tweet had been deleted.
4. After finish downloading all tweets' JSON data, we load `tweet_json.txt` line by line to `twee_data` table

Assess

By using `pandas` built-in functions, we have the following finds:

Quality

`twee_data` table

- Missing records (2330 instead of 2356)

`twee_arx` table

- Missing data in columns `in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id`, `retweeted_status_timestamp`, `expanded_urls`
- `timestamp` should be in Datetime datatype not string
- `retweeted_status_timestamp` should be in Datetime datatype not string
- Errors in `rating_numerator` and `rating_denominator`
- Errors in `name` : such, a, not, one, my, his, this, unacceptable, ... They are all in lower case.
- stage `doggo` contains values None and doggo. They should be `0` for not doggo and `1` for being a doggo.
- Missing a lot of data in stage variable (doggo, floofer, pupper, puppo) (can't clean)
- Unnecessary columns `source`, `timestamp`, `expanded_urls`

`image_pred` table

- Missing records (2075 instead of 2356)
- Dog's breeds are lower case and sometime upper case.

Tidiness

- One varible in four columns in `twee_arx` table (`doggo`, `floofer`, `pupper`, `puppo`)
- Columns `retweet_count` and `favorite_count` should also be part of the `twee_arx` table
- column `text` in `twee_arx` duplicated in `twee_data` table
- `text` column in `twee_data` table should be split into `name`, `rating_numerator` and `rating_denominator`

- `twee_data` and `twee_arx` contain retweet record. These record should be remove because they are not part of our analysis.

Clean

Missing data

1. We merge the tables in order to get a new table that containing the common records in all tables.
2. Some missing data due to the poorly represented values, we reform these values to get complete data.

Unnecessary columns

We remove some unnecessary columns.

Tidiness

1. Create column contains information of the four doggo,floofer,pupper,puppo
2. Merge the retweet_count and favorite_count columns to the twee_arx table, joining on tweet_id
3. Merge the p1 to p3_dog columns to the twee_arx table, joining on tweet_id
4. Remove the tweets which are retweets from other status or reply to others

Quality

1. Correct the name of the dogs
2. Fix the datatype of some columns
3. Correct the rating
4. Correct the dog's breeds

Store

The final cleaned dataset `twitter_archive` is store to a file named `twitter_archive_master.csv` .