Wrangle report

WeRateDogs is a famous twitter account that rate dogs and create humorous comments.

In order to better understand and analysis the data, three steps were performed.

a. Gathering
b. Assessing
c. Cleaning

Gathering

In this project, we used datasets from three data resources.

tweet archive: a local csv file

tweet_image: query from website via file url link

twet_df: retrieved via the twitter API connection.

Assessing

All three datasets have their quality and tidiness issues.

Quality

tweet_archive:

- tweet_id should be string not int
- incorrect rating for numerator with decimal points
- incorrect ratings, some have rating_denominator < 10
- Missing Dog names
- Incorrect Dog names, name is 'an', 'the'.
- Need to exclude retweets
- Timestamp should be a datetime object

tweet_image:

- tweet_id should be string not int
- p1, p2, p3 dog type names is inconsistant, need to remove under score, make all lower cases
- missing records
- duplicate image urls, same images used multiple times

tweet_df:

- tweet_id should be string not int
- Need to exclude retweets
- created time is not datetime type
- missing records

Tideness

tweet_archive:

- Urls should be separated from text
- Rating_numerator, and rating_denominator should be float data types rather than int
- Multiple dog names in the table, doggo, floofer, pupper, puppo should all be combined into one stage column
- Tweet_id should be string type

tweet_image:

- 6 variables to represent predictions, should be summarized into two columns, prediction and confidence
- Not all tweet ids in the archive table have images

tweet_df:

- column id needs to be changed to 'tweet_id'
- Duplicate ids in the dataframe, id and ids, both of them need to be converted to one tweet_id column
- Data index need to be reset to tweet_id
- tweet_df should be part of tweet_archive table

Cleaning

For the consistence of datasets, three new datasets were crated, tweet_archive_clean, tweet_image_clean, and tweet_df_clean.

Both manual and programmatic methods are being used here to clean the datasets, and quality and tidiness issues were solved in the cleaning section.

For the consistency of the analysis, all three datasets are joined together using unique tweet_ids, and the master csv file is created.

Conclusion

Data wrangling provided the very solid foundation of the further data analysis, and using data wrangling methods in this project, we are able to fix quality and tidiness issues and crate a much cleaner dataset.