Wrangle Data: WeRateDogs

Introduction

This project is about data wrangling and visual analysis of data from "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 Jupyter notebook demonstrates the detailed process to gather, assess and clean data. And this wrangle report provides a brief description of the steps taken during the data wrangling process.

Data Wrangling Steps

The data wrangling effort can be broken into three steps:

- Gather data
- Assess data
- Clean data

Gather Data

The data for this project comes from three different sources:

- **Twitter archive file**: The WeRateDogs Twitter archive file (twitter-archive-enhanced.csv) is directly provided by Udacity and manually downloaded.
- **Tweet image prediction file**: The dog breed prediction file (image_predictions.tsv) is hosted on Udacity's servers and is downloaded programmatically using the Requests library and the given URL information
- **Twitter API and JSON**: I use the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. I then read this .txt file line by line into a pandas DataFrame.

Assess Data

Once data from different sources are loaded into pandas DataFrame, I assessed the data quality and tidiness issue programmatically using Python in the Jupyter notebook. I identified following quality and tidiness issue, and fixed them item by item.

Quality Issue

twitter_archive

- Column tweet_id should be converted to str type.

- Columns timestamp need to be converted to timestamp type
- Un-useful columns need to be removed, such as in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp
- Columns rating_numerator and rating_denominator should be converted to float datatype, and they have invalid values
- Invalid None value in column name, should be NaN

image_prediction

- Column tweet_id should be converted to str type.
- Un-useful columns need to be removed
- Duplicated jpg_url values, which possibly indicate retweet vs original tweet

tweets_info

- Column tweet_id should be converted to str type.
- Columns favorite_count and retweet_count should be converted to int type

Tidiness Issue

- Data from separate tables need to be merged
- Need to melt columns doggo, floofer, pupper and puppo, and create a single column 'stage' to consolidate the information

Clean Data

In order to tackle each issue mentioned above, I split each task into three parts: Define, Code and Test. I start with defining the scope of each underlying issue, then write Python code to resolve the issue, and print result to test if the issue has been resolved.

Conclusion

With the minimum required data gathered from different sources, I was able to apply data wrangling techniques and provide preliminary analysis and insights about tweet data with the help of visualization. In order to accomplish a more comprehensive analysis, it would be helpful to collect additional data via Twitter API.