Wrangle Report

1. Introduction

Data wrangling is the practice of converting and then plotting data from one “raw” form into another. The aim is to make it ready for downstream analytics. Using Python and its libraries, we used data wrangling skills to pull real-world data from Twitter, clean it, and did some analysis. The dataset that we wrangled (and also analyzed and visualized) 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.

WeRateDogs is a popular Twitter hash tag, as the name tells, people rate dogs with a denominator of 10 and the numerator is usually higher than 10 to show how lovely the dog is.

2. Gathering data

The data was gathered from three sources:

- **Enhanced Twitter archive**: This part is kind of on-hand data, stored in the twitter_archive_enhanced.csv
- **Image prediction**: We get the image prediction data from web scraping
- **Twitter API**: This dataset is archived from Twitter's API and parsed from JSON to csv

3. Assessing Data

After assessing the datasets, we summarized several quality and tidiness problems of them:

Quality problems:

- Missing values identified in in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp columns
- In p1, p2, and p3 columns dog names were both lowercase and uppercase
- timestamp, retweeted_status_timestamp columns were not in date-time format
- tweet_id, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id columns were not in a string
- source column needed to be cleaned so that we might have four categories
- Suspicious character identified like 'a' 'actually' 'all' 'an' 'by' ect. in the name column
- Missing values were presented both as None and NaN in the name column
- Duplicates identified in the expanded_urls column
- Original ratings (no retweets) needed to be kept that have images (i.e. rows where retweeted_status_id and retweeted_status_user_id are not NaN) and the tweets up to the August 1st, 2017

Tidiness problems:

- Multiple timestamp columns after merging - drop the one and rename the other one so that we have one dataset
- doggo, floofer, pupper, puppo columns had a tidiness problem
- Drop other useless columns as well

4. Cleaning data

The data quality and tidiness problems mentioned above were cleaned by multiple methods including pandas join, regular expression, combining multiple columns, pandas subsetting, removing missing values and so on.

In the end of this part, we stored the cleaned version of the data into a csv file for future usage.

5. Analysis & Visualization

These steps are not part of the data-wrangling process. However, it cannot reflect correct and accurate insights without performing data wrangling first. Visualizations and insights are provided in act_report.pdf