Anh Tran Thi Ngoc
Data Analysis Nano Degree

DATA WRAGLING REPORT

The purpose of this report is to briefly describe the wrangling process of WeRateDogs Twitter Data Analysis Project. This report consists of three main parts: gathering data, assessing data and cleaning data which will be explained as follow.

I. Gathering Data

Data for this project is gathered from three main sources:

1. The twitter archive data contains basic tweet data of Twitter user @dog_rates, also known as WeRateDogs, from 11/2015 to 8/2017. This data set can be manually downloaded by clicking this link.
2. The tweet image prediction data, which provides information about the breed of dog, is hosted by Udacity's server. We can programmatically download this data set by using the Python Requests library using the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. Data relating to each tweet's retweeted count and favorite count is obtained by querying Twitter API for each tweet's JSON data using Python's Tweepy library based on tweet ID in the twitter archive data set. Each tweet's JSON data should be written to its own line and then stored in a file called tweet_json.txt file.

II. Assessing Data

After successfully gathering, data will be visually and programmatically assessed for quality and tidiness issues.

1. Visual Assessment

In visual assessment, each table will be displayed in its entirety in a pandas DataFrame that it was gathered into. The column descriptions were also collected to understand the meaning of each variable in the table.

Quality issues

- *Twitter archive table*
  - Columns with missing values: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls
  - Some values in rating numerator and rating denominator are wrongly extracted from text
  - Rating denominator has different value, which is normally 10.
  - Some dog's names are improperly extracted.

- *Image prediction table*
  - Some data in columns p1, p2, p3 are written in lowercase.

Tidiness issue: four columns (doggo, floofer, pupper, puppo) should be integrated in to one column dog_stage.

2. Programmatic Assessment

In programmatic assessment, the following assessment methods in pandas are used to figure out the quality and tidiness issue of data: .info(), .head(), .sample(), .describe(), .value_counts(), .isnull(), .isin(), .duplicated() and so on.

Quality issue

- *Twitter archive table*
  - 181 records for retweets (we only focus on original tweets).
  - Timestamp has a data type of datetime instead of object.
  - Source should be written in text instead of url. There are only four values for sources: Twitter for Iphone, Vine - Make a Scene, Twitter Web Client, TweetDeck.
  - Many tweet_id in twitter archive table are missing in image prediction table.

- *Image prediction table*
  - There are some predictions of non-dog images

- *JSON table*
  - Created_at is duplicated with timestamp in twitter archive table
  - Many tweet_id in JSON table are missing in image prediction table

Tidiness issue

- Parse year, month, date from timestamp for analysis.
- Merge 3 tables on tweet_id for further analysis.

III. Cleaning Data

A copy of each table is created before cleaning data. For each quality and tidiness issue, the cleaning process which consists of three steps define, code and test are performed. After that, three cleaned tables are inner joined on tweet id to form a cleaned master data. This data set is saved in the **twitter_archive_master.csv** file.