Project 4: Wrangle and Analyze Data

# Wrangle Report

**Introduction**

Data wrangling is the process of transforming and mapping data.
Wrangling consists of basically three steps:
• Gathering.
• Assessing.
• Cleaning.

**Gathering**

Data was gathered from 3 files:

1. *twitter_archive_enhanced.csv*
(Provided by WeRateDogs and hosted on Udacity server in csv format)
This archive contained basic tweet data like tweet ID, rating, dog stage, timestamp, text, etc. I imported the file into df_twitter dataframe using pandas read_csv function.

2. *image_predictions.tsv*
(Image prediction algorithm file provided by Udacity)
I imported the file into df_prediction dataframe using pandas read_csv function (with separator as \t) after getting the file by using request library from the URL provided by Udacity.

3. *tweet_json.txt*
As I didn't want to create an account on twitter, I downloaded the file from Udacity server and extracted required information using json library (loads function) and then storing into df_count dataframe.

**Assessment**

After gathering, I assessed the data by both visually and programmatically approach for quality (based on completeness, validity, accuracy and consistency) and tidiness issues.
Following assessments were made:

Quality Issues:

df_twitter:

- The timestamp field is in object (pointer to string)
- Missing expanded_urls for some records
- Some name are not correct like a, None
- Rating_numerator max value is 1776 (this is wrongly captured)
- Rating_denominator minimum values is 0 and max value is 170 (this is wrongly captured)
- Contains retweets
- Contains replies

- Extra columns like in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id
- Multiple stages for few dogs.

df_prediction:

Prediction columns contain mixed string i.e., some are starting from capital and some are from lower.

Tidiness Issues:

- There are 4 columns for representing dog stages (doggo, floofer, pupper, puppo) - indicating one variable in df_twitter.
- This dataset should be merged with df_twitter dataset

**Cleaning**:
Here I fixed the issues illustrated in above step i.e., Assessment.
The common approach is followed (not for few simple cleaning activities):

1. Define: Defining how to address the assessed issue
2. Code: Writing code to solve the issue
3. Test: Checking whether the issue is solved.

Finally cleaned the data and store in new csv files after few iterations of cleaning and assessments.

Analysis & Visualization
Data wrangling is mandatory before Analysis & Visualization as dirty or messy data could result into wrong insights and observations. As I've wrangled the data, I performed some Analysis using matplotlib and seaborn (will be better in this section after last lesson ☺)

Analysis is shared in separate report: act_report.

Thanks & Regards,
Bharat.