WRANGLE REPORT

I. INTRODUCTION

This project is one of the requirements to complete and graduate from Udacity's 'Data Analyst Nanodegree' program. The project mainly focuses on the classical three steps of data wrangling process, namely: gathering, assessing and cleaning data.

A twitter archive file was provided by Udactiy and it contains 2356 users' tweets each of which having 17 attributes. Other files were required to complete the project and instructions were given on how to gather them from different sources. All of these files are related to tweets from WeRateDogs twitter user account. WeRateDogs rates people's dogs with a humorous comment about the dog[1].

II. DATA GATHERING

As stated in introductory part of this report, different files from different sources were needed to complete the project. These files include:

- **twitter_archive_enhanced.csv**: this file has been provided by Udacity and can be manually downloaded from 'resources' section for the project part of the course.
- **image_predictions.tsv**: this file is hosted on Udacity's servers and is downloaded programmatically using the Requests library and this URL.
- **tweet-json.txt**: this file can be accessed by querying the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data. As we faced some technical challenges to set up a developer account on Tweeter (*application under review* for too long, till the time of submission of this project), we decided to use the version provided by the instructor which can be manually downloaded from the course's project page.

III. ASSESSING DATA

With all needed files at hand, the next logical step was to assess them for quality and tidiness. The assessment of the files was done both manually and programmatically to come up with the following issues:

[1] https://classroom.udacity.com/nanodegrees/nd002/parts/af503f34-9646-4795-a916-190ebc82cb4a/modules/14d9f5f1-9e7b-4bfb-97f3-bcdbf4a3699c/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/5e3db54a-1a5f-41a6-8e20-fd99f201861d

Quality issues

In `tweet_archive` table

- Has rows for tweets and for retweets
- Erroneous datatypes (assigned ***timestamp***, ***in_reply_to_status_id***, ***in_reply_to_user_id*** column)
- There is a row with the ***rating_denominator*** equal to zero
- The ***source*** column has trailing strings
- The ***name*** column has erroneous values (like a, an, this, etc.), some names start with Capital letter while other start with small letter. This is inconsistency problem.
- Name variable has "None" values instead of NaN

In `img_predictions` table

- There are fewer tweets IDs in `***img_predictions***` table than in `***tweet_archive***` table suggesting that there could be missing data or inconsistent rows in one of the tables.

In `tweets_data` table

- There are fewer tweets IDs in `***tweets_data***` table than in `***tweet_archive***` table suggesting that there could be missing data or inconsistent rows in one of the tables.

Tidiness issues

- ***retweeted_status_id***, ***retweeted_status_user_id*** and ***retweeted_status_timestamp*** columns are empty if all retweets are removed from `tweet_archive` table.
- One variable ***stage*** is scattered into four columns ***doggo***, ***floofer***, ***pupper*** and ***puppo***.
- Retweet counts and favorite counts should be part ***tweet_archive*** table.
- Breed of dogs computed based on predictions should be part of ***tweet_archive*** table.
- ***retweet_counts*** and ***favorite_counts*** columns from ***tweets_data*** table should be part of ***tweet_archive*** table.
- Columns from ***img_predictions*** should also be part of ***tweets_data*** table

IV. DATA CLEANING

In this section of the project we cleaned all issues identified in previous section using the Define-Code-Test strategy.

Clean data frame was saved as a comma separated file (***twitter_archive_master.csv***) as well as in a Sqlite database called ***twitter_db*** as ***twitter_master*** table.