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. This table should contain only original tweets to avoid duplication of rows.
- Erroneous datatypes (assigned timestamp, in_reply_to_status_id, in_reply_to_user_id, rating_numerator column). Their data types should be changed to proper data type representation of their content.
- There is a row with the *rating_denominator* equal to zero, which is an invalid value for a denominator. Calculation of ratings for such rows will not yield meaningfull result.
- The *rating_numerator* contains some incorrect values which do not correspond with those *text* column.
- The *source* column has trailing strings. Anchor tags used to display 'source' values are not part of the source itself. They make values of this column to be incorrect.
- The *name* column has erroneous values (like a, an, this, etc), some names start with Capital letter while other start with small leter. This is an inaccuracy problem and 'fake' names should be removed.
- Name variable has "None" values instead of NaN. This is a wrong representation of none existing data leading to inaccuracy problem

In `img_predictions` table

- Erroneous datatypes (assigned tweet_id column). Values in this column are not numbers. Therefore its data type should be changed to proper data type representation of its content.
- There are duplicated jpg_url. Two or more different dogs should not have same image in the dataset. These are wrong but valid data.

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. They are no longer needed.
- One variable **'stage'** scattered into four columns doggo, floofer, pupper and puppo. Hence they are about one variable, they should form one column.
- These three datasets have to be combined in one dataset because every row in each of them has information about one single tweet.

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.