Data Wrangling Report

By Amir Reda Halim

August 2020

As a student of the Udacity Data Analysis Nanodegree, I'm providing the below Data Wrangling report to illustrate the steps taken in the 2nd project 'WeRateDogs'

The work was divided into

- **Data Gathering**
- **Data Assessment**
- **Data Cleaning**
- **Storing Master file**
- **Analysis and Visualization**

Data Gathering:

In this step, I imported all the needed files for the whole project. As requested in the Project Motivation there are 3 main files to import in different ways as specified

1. Enhanced Twitter Archive : The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets. The file was imported as dataframe from csv format
2. Additional Data via the Twitter API: since allowing the access to twitter API took 4 day, I downloaded the file 'tweet-json.txt' provided byUdacty and converted it into dataframe
3. Image Predictions File: file was imported using the provided URL then converted into a data frame while saving a copy of the file in a newly created directory called 'dogs'

Data Assessment :

In this step, the imported files were investigated both visually by opening them on MS Excel and going through them for better understanding of the files content and Programmatically on Jupiter notebook using code to examine the data provided. From this step I was able to capture some of the quality and tidiness issues that must be sorted before I could proceed with my analysis and visualization

Notes were taken to be a guideline for the Data Cleaning step and some were added during the work

Data Cleaning:

In this step code was used to clean any quality or tidiness issues that were addressed during the assessment. All action were done on 3 steps: Define, Code and Test

The first step done was to remove the repeated column after merging the archive with API's data. After that came excluding any retweets using in_reply_to_status_id & retweeted_status_id.

Then all tweets with no images were excluded. Regarding the dog classification, the 4 fields were grouped under 'Class' then dropped the old fields and replacing None with NaN.

Some useless columns were dropped afterwards ('in_reply_to_status_id' , 'in_reply_to_user_id', 'retweeted_status_id' , 'retweeted_status_user_id' , 'retweeted_status_timestamp').

Later I fixed the types of the archive columns.

One of the visual observations was that name was extracted from text so some mistakes were found like capturing 'such' as a name of a dog, which was corrected manually. Then the 'a' and 'an' were presented as none

Also the rating_numerator was extracted from the text

One of the columns name was change from expanded_urls to Image_url to be clear

After that I moved to the image prediction dataframe, started by changing the names to more descriptive ones and grouped the predictions into 3 fields instead of 9. Since some of the outcomes were not dog breeds, I had to remove them

Data Storing:

As requested the final Twitter archive was saved in the csv format under the name twitter_archive_master.csv

Data Analysis, Viz & Reporting:

To be able to visualize the data some tweaking had to be done like grouping and recalculation. Then comes the final step of adding visuals to be later presented