UDACITY – WRANGLE REPORT

TO: UDACITY REVIEWER

FROM: JHONATAN NAGASAKO

SUBJECT: WRANGLE REPORT (300-600 WORDS)

DATE: 27-FEB-2021

PURPOSE

The purpose of this project was to analyze real-world data. This data analyzed took the form of:

1. Meta twitter data (@dog_rates)
2. WeRateDogs twitter
3. Udacity Neural network algorithm (to predict dog breed)

The meta twitter data or referred to as "twitter-archive-enhanced.csv" was a complication of over 5000+ tweets from @dog_rates. This data was then filtered for tweets that only contained ratings of dogs, which was about little under half (2356 tweets). This data was already provided to the data scientist, for example as if it were given as a file in a flash drive.

In addition to this data, WeRateDogs repository of dog images data was also provided for analysis, "image_prediction.tsv". This data was extracted via the internet, which was then fed through a proprietary neural network program that predicated the breed of a given dog by only its picture. The analysis and creation of this neural network code is out of scope for this project.

Lastly, the implementation of a Twitter API (Tweepy) was explored—but not used—for this project to extract additional data to supplement the "flash drive" data provided earlier. Tweepy was not used because of cyber security and privacy reasons. Instead, the given "tweet-json.txt" file was used, which would have been the same file extracted from the Twitter API. This data had a multitude of meta information (which took the form of a JSON file) and additional processing was required to extract relevant information (e.g., minimum the retweet_count and favorite_count) for the analysis completed in this project.

GATHERING PROCESS

The analysis of the first type of data, the data wrangling of the “twitter-archive-enhanced.csv”, was not an issue. Analysis of this .csv file was like lessons and projects completed in the past. However, the processing of the second type of data file from the internet proved challenging. Specially, the understanding of HTTP library request process was required to ensure the correct status code (e.g., status code 200) and proper encoding (e.g., “utf-8”) was checked to successfully extra the data required for analysis. Lastly, the coding to extract important information from “tweet-json.txt” was also a challenging task because it was difficult understanding both the Python and HTML code. Regardless, all the necessary files were extracted successfully for the data cleaning process.

CLEANING PROCESS

As with most data science projects, 80% of the time was spent cleaning and tiding data. There were multiple instances where during the data *explanatory* process that additional data *exploration* was required to visually or programmatically review the data for further processing. In general, the following tips was used to clean and tidy the datasets:

Tips for Tidying

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table

Tips for Common Data Quality Issues

1. Missing data
2. Invalid data (e.g., state a negative height, or other datatype validation errors--str vs int vs float, think there can only be 2 people in a room... not 2.54 people in a room... *unless there's ghosts lol*)
3. Inaccurate data (e.g., specifying a foot = 5 inches, which is WRONG. A foot = 12 inches)
4. Inconsistent data (e.g., mixing up units, some data captured as cm instead of inches)

CONCLUSION

The data scientist was able to download/process multiple twitter datasets and perform necessary cleaning operations for later analysis. Please see the supplement “act_report” for the data exploration and data explanatory phases of this project.