UDACITY DATA ANALYTICS NANODEGREE PROGRAM

DATA WRANGLING REPORT FOR @WeRateDogs Twitter

By Egwuda Ojonugwa Everest

INTRODUCTION

The objective of this project was to gather, clean and analyze over 5000+ tweets from the twitter account of @dog_rates (WeRateDogs), draw some insights and make visualizations to communicate one of the insights drawn. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The account was started in 2015 by college student Matt Nelson, and has received international media attention both for its popularity and notoriety.

The data wrangling process contains three major steps, namely:

- Gathering Data
- Assessing Data
- Cleaning Data

During this project, we will gather data belonging to Twitter user @dog_rates from a variety of sources and in different formats, assess its quality and tidiness, then clean it.

DATA GATHERING

In this phase, three datasets were indicated to be necessary for this analysis.

- One was provided beforehand (twitter_archive_enhanced.csv), we downloaded this file and read it in as a CSV file in a Pandas DataFrame using the pandas' read_csv function.

- The second file (image-predictions.tsv), i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network was to be downloaded programmatically using the 'requests' library and then stored into a file on the pc before being read into a Pandas DataFrame using the read_csv function.

- The third file proved difficult to gather, it required applying for Twitter Developer access which I applied for but I didn't get the approval even after waiting for over a week. Thoughtfully Udacity had an alternative option to this which involved downloading a provided tweet-json.text file and reading the file line by line.

ASSESSING DATA

Assessing Data Involves looking at the dataset visually and programmatically so as to notice possible quality or tidiness issues in the datasets before Data Cleaning them. A couple of issues were noted and documented in preparation for the cleaning phase, with two major types of issues with the dataset, **QUALITY** and **TIDINESS** issues.

QUALITY Issues

1. retweeted_status.timestamp and timestamp in ratingDF are objects

2. The rating_numerator and rating_denominator columns in ratingDF should be a float.

3. Some of the names in ratingDF are inaccurate and can't be names.

4. retweets and replies in ratingDF are not necessary

5. Some of the jpg_urls in image_predict are incomplete (endswith (...))

6. The remaining IDs should all be qualitative

7. We will get rid of "None" values in the doggo - puppo columns in ratingDF, replace with null and then melt the columns together.

8. Source column is overpopulated with irrelevant information, making it difficult for us to tell the source.

9. The rating numerator and denominator column was extracted well, with some ratings as de cimals

TIDINESS Issues

1. Merge data frames on shared tweet_id

2. There are a lot of null values in rd_df columns doggo, floofer, puppo, and pupper.

DATA CLEANING

Data cleaning was the most time consuming and tricky part of wrangling the data. I made good use of classroom resources and also google to better understand codes and adjust some of the issues with the dataset. After cleaning the datasets, I combined the three datasets into one master dataset (twitter_archive_master.csv) and stored it in preparation for analysis and visualization.