Wrangle Report

Introduction

The purpose of this project is to put in practice what I learned in data wrangling data section from Udacity Data Analysis Nanodegree program. The dataset that is wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10.

This report briefly describes my wrangling efforts.

Project Details

The tasks of this project are as follows:

- Gathering data
- Assessing data
- Cleaning data

Gathering Data

The data for this project consist on three different dataset that were obtained as following:

- Twitter archive file: the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.

- The tweet image predictions, i.e., what breed of is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information
- Twitter API & JSON: by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url.

Assessing Data

Once the three tables were obtained I assessed the data as following:

- Visually, I used two tools. One was by printing the three entire dataframes separate in Jupyter Notebook and two by checking the csv files in Excel.
- Programmatically, by using different methods (e.g. info, value_counts, sample, duplicated, groupby, etc).
- Then I separated the issues encountered in quality issues and tidiness issues. Key points to keep in mind for this process was that original ratings with images were wanted.

Cleaning Data

This part of the data wrangling was divided in three parts: Define, code and test the code. These three steps were on each of the issues described in the assess section

There were a couple of cleaning steps that were very challenging. One of them was in the image prediction table. I had to create a 'nested if' inside a function in order to capture the first true prediction of the type of dog. The original table had three predictions and confidence levels. I filtered this into one column for dog type and one column for confidence level.

One very challenging cleaning step was when I had to correct some numerators that were actual decimals. This issue was brought to my attention after the first Udacity review. Using Excel and visual assessment was not sufficient to verify those decimals. Therefore, I had to run a code in order to check those actual tweets (decimals numerators).

Used basic python function like duplicates, drop, sort, value_count, describe, info and others to comply with above mentioned point. I struggle with few issues and had to spend a lot of time to get my understand. As little help was provided its first time I used so many websites for checking syntax and possible solutions.

Conclusion

Data wrangling is a core skill that whoever handles data should be familiar with.

I think only after learning thoroughly from Udacity platform I have started to grasp coding mindset but as I am changing career so still a lot more to learn.