Data Wrangling Report

Project objectives

The project main objectives were:

- Perform data wrangling (gathering, assessing and cleaning) on the provided sources of data.
- Store, analyze, and visualize the wrangled data.
- Reporting on

    1. data wrangling efforts.
    2. data analyses and visualizations.

Step 1: Gathering Data

In this phase, the three pieces of data were gathered and represented as pandas dataframes:

- The We Rate Dogs Twitter archive (file on hand, manual download of 'twitter-archive enhanced.csv')
- The tweet image predictions ('image-predictions.tsv'). This file was be downloaded programmatically using the Requests library from a provided URL.
- Each tweet's entire set of JSON data (with at minimum tweet ID, retweet count,and favorite count) in a file called 'tweet_json.txt' were stored using Twitter API code submitted by udacity.

Step 2 and 3: Assessing and Cleaning Data

While working with data, a number of observations were made. In the below table there are the observations along with actions taken in the Cleaning Step.

Tidiness

Dataset	Observation	Solution
ALL	Dog stage data is separated into 4 columns.	Merge the 4 columns into 1 called dog_stage.
	All data is related but divided into 3 separate dataframes.	Merge all dataframes into 1 based on tweet_id

Quality

Dataset	Observatin	Solution
twitter archive	There are 181 retweets as indicated by retweeted status_id.	Delete rows that represent retweets and all the relatted columns
	Name column have None instead of NaN and too many unvalid values.	Replace 'None' with np.name in t_archive name column. Remove any rows with invalid names which starts with lower letter.
	invalid tweet_id data type(integer instead of string).	correct invalid data type by converting tweet_id to string.
	invalid timestamp data type(string not datetime).	correct invalid data type by converting timestamp to datamine
Image predictions	underscores are used in multi-world names in columns p1,p2&p3 instead of spaces.	convert underscores to spaces
	some P names started with an uppercase letter while other started with lowercase	convert lowercase letters to uppercase
Tweet data(api)	missing entires (only 2354 entires instead of 2356)	delete rowes without retweet_count entiers

Result

A combined data set with all needed information was stored in a data base.