Data Wrangling Report

Project Objectives

The main project objective was to help the student engage in the following:

- Perform the first step in the data wrangling tree which is data gathering from three different file formats and sources
- Access the data gathered both visually and programmatically
- Clean the data for all visible errors that could be gathered
- Make insights into the data and create visualizations when appropriate
- Create a report to discuss wrangling efforts and insights and visualizations created.

Step 1: Gather The Data

The weratedogs Twitter archive

The file was given to me by udacity via a link, which I had to download programmatically, upload, and read into a pandas data frame.

The tweet image predictions

The file was also downloaded programmatically using the Request library, using the URL that was provided by udacity

Additional data from Twitter API

Query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.Each tweet's JSON data should be written to its own line.

Steps 2 and 3: Assess and Clean The Data

While working with the data some assessments were made and errors were recorded and clean, the table below shows a record of this

QUALITY

Dataset	Observation	Solution
Tweet_df	Not all are dog ratings and some are retweets	I Removed rows that are retweets and not dog ratings as required by the project.
Tweet_df	Inconsistent URL format between the source and extended URL column	I removed the HTML tags that are with the source URL.
Tweet_image	Inconsistent input of lower and upper cases in the (P1, P2,P3) columns and German_short-haired_pointer instead of German_short_haired_pointer	Changed the case of the various columns to lower case and changed the value with the correct format.
Tweet_df	Rating are incorrect	Inspect the ratings denonmianor and check out the text for this values to extract the proper ratings
Tweet_df	The timestamp column is not in the right data type.	convert the datatype of the timestamp column datetime
Tweet_df	missing values in the in_reply_to_status_id and in_reply_to_user_id	Drop the columns
Tweet_df	Multiple dog stages in a row	Check the text and extract the right stage for each row.
Tweet_df	Names starting with capital letter are valid name	Replace the names that do not stat with capital letters
Tweet_df	Decimals are not captured in the rating numerator and denominator	Extract the decimals from the text
Tweet_df	The null values are None	Change to empty values

TIDINESS

Dataset	Observation	Solution
Tweet count	the tweet count table should be in the tweet_df table	I merged the tweet count table to the tweet_df.
Tweet df	values(doggo,pupper,puppo,floffer) are variables	convert the columns to rows and name the column dog_stage.
Tweet_image	the tweet_image table should be in the tweet_df	I merged the tweet image table to the tweet_df.

RESULT:

As a result, I was a to create one tidy data analytical table that was read for analysis.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1990 entries, 0 to 1989
Data columns (total 22 columns):
tweet_id              1990 non-null int64
timestamp             1990 non-null object
source                1990 non-null object
text                  1990 non-null object
expanded_urls         1990 non-null object
rating_numerator      1990 non-null float64
rating_denominator    1990 non-null int64
name                  1990 non-null object
stage                 304 non-null object
retweet_count         1990 non-null int64
favorite_count        1990 non-null int64
jpg_url               1990 non-null object
img_num               1990 non-null int64
p1                    1990 non-null object
p1_conf               1990 non-null float64
p1_dog                1990 non-null bool
p2                    1990 non-null object
p2_conf               1990 non-null float64
p2_dog                1990 non-null bool
p3                    1990 non-null object
p3_conf               1990 non-null float64
p3_dog                1990 non-null bool
dtypes: bool(3), float64(4), int64(5), object(10) memory usage: 301.3+ KB
```