WRANGLE REPORT

This report briefly describes the data wrangling efforts exerted in this project.

The tweet archive of the Twitter account @dog rates, better known as WeRateDogs, was wrangled, analyzed, and visualized.

The entirety of this project was completed on the Udacity Project workspace. However, the reports were created as PDFs using Google Docs.

The wrangling process is divided into three steps:

1. Gathering Data
2. Assessing Data
3. Cleaning Data

The steps involved in the project are explained below:

1. GATHERING DATA

The data used was gathered from three different sources:

i. ENHANCED TWITTER ARCHIVE

This file contains data retrieved programmatically from Twitter data sent to Udacity by WeRateDog via email for use in this project. The information includes the rating, dog name, dog stage, and other pertinent details.

This file was downloaded manually from the provided link and uploaded manually to the project workspace twitter-archive-enhanced (1).csv.

ii. IMAGE PREDICTION FILE

This is created by using a neural network to classify dog breeds on every image in the WeRateDogs Twitter archive. This algorithm produced a table with picture predictions next to each tweet ID, Image URL, and image number.

This is hosted on Udacity saver and was downloaded programmatically using Request Library and the URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv.

One of the Images from WeRateDog

iii. ADDITIONAL DATA VIA TWITTER API

This is obtained by querying Twitter’s API and then stored in a text file called tweet_json.

The ready-made version of the file provided by Udacity was used in this project and was read line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count and was later saved as ‘tweet_data.csv file for future use.

2. ASSESSING DATA

After gathering the above data they were assessed visually and programmatically for quality and tidiness issues.

While working with data, a number of observations were made the below findings were discovered along with the actions taken in the cleaning steps

A. TIDINESS

OBSERVATION	SOLUTION
Dog stage data is separated into four columns	Merge the columns into one called dog_stage
All data is related but divided into 3 separate data frames.	Merge all data into one tweet_id

B. QUALITY

DATASET	OBSERVATION	SOLUTION
Enhanced Twitter Archive	There are 181 retweets as seen in the retweeted_status_id.	Delete rows that present retweets and all related columns
	Some dog names are not valid since some names are "None, a, an, the" instead of name.	Covert invalid names to None and extract the correct names from the text column (after the word 'named')
	Invalid tweet_id data (this is an integer instead of a string).	Correct invalid data type
	Timestamp data is a string instead of dog time.	Convert timestamp to DateTime
	The source column is HTML tag "< a >"	Extract tweet source from source column.
Tweet Image Prediction	Missing photos for some ids. There are 2075 rows instead of 2356.	Delete rows with missing photos
	Underscores are used instead of spaces in Multi-words	Correct the Multi-words names and use spaces instead of underscores
	Some P names start with a capital letter while others start with a small letter.	Convert small letters to Capital letters
Tweet Data from Twitter API	Have 2354 entries instead of 2356 entries (missing entries).	Delete row without retweet count entries

3. CLEANING DATA

The first step in the cleaning method was to copy all the three DataFrame using the .copy() method.

- clean_df_1 = df_1.copy()
- clean_i_predictions = i_predictions.copy()
- clean_tweet_data = tweet_data.copy()

Then the quality and tidiness issues were cleaned as appropriate resulting in high quality and tidy master DataFrame