DATA WRANGLING PROJECT

BY MUTHOLIB YUSIRA

INTRODUCTION

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. The goal is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering from other sources, then assessing and cleaning will be required for "Wow!"-worthy analyses and visualizations.

DATA GATHERING

I started my project by gathering data from 3 different sources:

- The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided to me by Udacity. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets. I downloaded the 'twitter-archive-enhanced.csv' file manually into the workspace and loaded it into a dataframe titled, 'twitter_archive'
- The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file was provided to me by Udacity. I downloaded the 'image-predictions.tsv' programmatically from Udacity's servers using the requests library. Then I read it into 'image_predictions' Pandas dataframe.
- I accessed the third data using the 'tweet-json.txt' file provided by Udacity as an alternative. I read the text file line by line, obtained each tweet's information (tweet ID, retweet count, and favorite count) using the json library, and appended the information into an empty list. Finally, I convert the list of dictionaries to a pandas DataFrame and saved it into 'twitter_data'.

ASSESSSING AND CLEANING

Some quality and tidiness issues were identified for the three tables. Details of the issues identified and solutions are in the table below:

Quality Issues:

Twitter_archive Dataset:

Issues	Solutions
• Rows with 'retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_user_id, in_reply_to_status_id' are retweets and not original ratings	• Keep only original ratings and no retweets, by deleting rows where the values in these columns are not null
• Dataset includes columns not needed for the analysis	• Drop the columns that are not needed
• Erroneous datatypes(tweet_id, timestamp, source)	• Convert columns to correct dataype
• Some rows in the 'rating_denominator' column are not equal to 10	• Drop rows where the rating_denominator is not equal to 10
• The string in the 'source' column is written in HTML format	• Convert string in 'source' column to normal format by removing the parts of the text with HTML format
• Rows in the 'text' column end with short links	• Remove the unnecessary links from the ends of text in the 'text' column
• 'text' column includes ratings again, and some do not correspond to the 'rating_numerator' column	• Correct the incorrectly exrtracted 'rating_numerator' colum and remove ratings from the 'text' column
• Null values in the 'expanded_urls' column	• Remove rows with null 'expanded_urls'
• Some urls in the 'expanded_urls' column are written multiple times in the same row	• Remove duplicated urls in the individual rows in the 'expanded_urls' column
• Incorrect dog names e.g 'a', 'an', 'not', 'very', 'the', etc	• Change dog names that are not proper nouns to None

Image_predictions and Twitter_data Dataset:

Issues	Solutions
• Erroneous datatypes(tweet_id)	• Convert 'tweet_id' to string

Tidiness Issues:

Twitter_archive Dataset:

Issues	Solutions
• 4 different columns for the 4 dog stages (doggo, floofer, pupper, and puppo)	• Combine all the 4 columns into a single column and drop the other columns not needed

Image_predictions and Twitter_data Dataset:

Issues	Solutions
• Multiple columns for image predictions and confidence intervals	• Use a function to select the first true image predictions and confidence intervals, and deleting the rest of the columns
• Columns should be part of the twiiter_archive dataset	• Merge the image_predictions table to the twitter_archive_clean table