Data Wrangling Report

Introduction

In this data wrangling project, the goal is to clean up the data quality and tidiness issues using both visual and programmatic assessments

Data Gathering

There are 3 main datasets for this project.
1. Twitter Archive: This file which provided by Udacity in the beginning of this project.
2. Retweet favorite: This file is query extract into JSON file from the Twitter AP by using the Twitter_id columns inside the twitter_archive dataset that was provided in the project introduction. The file was then store under the name tweet json.txt.
3. Image prediction: This file was also provided by Udacity which was downloaded under the name is image prediction.tsv, it contains the tweet image prediction breed of dogs by using neural network.

Data Accessing

The main goal of data accessing is to clean up the data and turn it into reader friendly. Two most important key factors during data assessing is to check for data quality issues and tidiness issues.

- Quality are the contents inside the data, check for missing data, duplicate, or incorrect data
- Untidy data would be the structural of the data, such as datatype.

We also should keep in mind when performing data quality assessment for the four dimensions, completeness, validity, accuracy, and consistency. These are fours important elements that will optimize the data quality.

Tidiness issues:
- Float in reply status_id (archive)
- Float in reply user_id (archive)
- Float in retweeted_status_user_id (archive)
- Float in retweeted_status_id (archive)
- Timestamp column in string (archive)
- Retweet stat timestamp in string (archvie)
- floofer, doggo, puppo, pupper columns belong in one column (archive)
- Tweet_image and retweet and favorites share dataframe (favorites)

Quality issues:
- Upper and lower in p1, p2, and p3 columns (image)
- column name for p1,p2 not clear (image)
- Retweet status_id in decimal (archive)

- Retweet status_user_id in decimal (archive)
- Rating numerator max is with index 979 (archive)
- Missing data in in_reply_to_status_id and in_reply_to_user_id (archive)
- Missing data in retweet status id, user_id, timestamp (archive)
- Missing data in expanded urls (archive)
- Under name column, names lowercase is not actual name (archive)
- Rating denominator and numerator are inconsistent (archive)
- Source code difficult to comprehend

Data Cleaning

- Use dropna on the missing data's column, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp. Therefore, we will drop these columns since we dont need them to complete our analysis because our focus of this project is dog's breed rating.
- Timestamp data type is in string. We will need to convert the data type to timestamp.
- Use melt to organize dog's stage under one column
- Use merg function to combine all 3 datasets into one
- Capitalize dog's name under the p1,p2,p3 column
- Track down and drop the outlier under numerator and denominator column
- Replace out lowercase names under name column by iteration and replace method
- Correct display format for floats
- Use drop duplicates function to drop the duplicate rows
- correct values under sources column into 3 types of tweeter sources, iPhone, web client and TweetDeck.

Store

Finalize and save the clean data into the file under the name twitter_archive_master.csv