WRANGLE REPORT

This is the report of the processes involved in wrangling the data used for the analysis of WeRateDog Twitter archive.

INTRODUCTION

Data Wrangling is the second step of the Data Analysis Process. It is an iterative process which consists of three steps:

- Gathering data
- Assessing data
- Cleaning data

DATA GATHERING

I gathered the datasets used for the analysis from three different sources:

- WeRateDog Twitter Archive: This is a csv file provided on the Udacity learning platform. It was downloaded and then uploaded into the jupyter notebook by using the Pandas read_csv() method.
- Image Prediction Data: This was downloaded programmatically from the web by using Python requests library with a url provided by Udacity.
- Twitter API dataset: I used Python Tweepy library to query Twitter API to gather some additional data from WeRateDog Twitter archive.

ASSESSING DATA

In this stage, I assessed the gathered data visually and programmatically to check possible quality and tidiness issues that need to be cleaned.

The following quality issues were observed:

1. Only 78 entries in in_reply_to_status_id and in_reply_to_user_id
2. Timestamp column in archive is a string not a datetime data
3. Invalid entries in the names column e.g. 'getting', 'a', 'actually', etc.
4. retweeted_status_id and retweeted_status_user_id and retweeted_status_timestamp have 181 entries.
5. Invalid data type of tweet_id columns in tweet and image_prediction tables.
6. Some denominators are greater than 10.
7. Source column in tweet table can be cleaned properly to have the possible sources of tweets.
8. Rename some columns in image_prediction to be more descriptive.

The following tidiness issues were also observed:

1. tweet_count and favorite_count should be part of tweet table
2. doggo, floofer, pupper, puppo columns in tweet table should be collapsed into a single dog_stage column.
3. Drop columns that are not needed from the tweet table.

DATA CLEANING

The last step of the data wrangling process is cleaning data. This is where the issues observed in the assessment phase will be addressed. Before proceeding with cleaning, I created a copy of the original datasets.

Cleaning Quality Issues:

1. Addressing the first quality issue, the rows where the columns (("in_reply_to_status_id" "in_reply_to_user_id") were not empty. This will leave us with tweets which are not replies to original tweets.
2. The timestamp column was converted to a datetime data type.
3. I converted the invalid entries in the name column to null entries.
4. I dropped the rows where "retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp" were not null. This will leave us with only tweets which are original tweets.
5. I converted the data types of this columns to strings because I will not be doing any calculation with the ID column.
6. Rows where rating denominator are greater than 10 dropped.
7. The exact source of tweets were extracted from the urls in the source column by using regex to extract the value just before the closing tag </a>.
8. The columns p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog in the image_prediction table were renamed to be more descriptive.

Cleaning Tidiness Issues:

1. I merged tweet_count and favorite_count to the tweet table.
2. The four columns 'doggo', 'floofer', 'pupper', 'puppo' were collapsed into a single 'dog_stage' column
3. Some columns in the tweet table which are now redundant were dropped.

Finaly, I merged the image_prediction table to the tweet table using an inner merge so that we have only original tweets with images.

The cleaned data was then stored into a new file: "twitter_archive_master.csv".d