Wrangle Report

Introduction

The purpose of this project is to put in practice what I learned in data wrangling data section from Udacity Data Analysis Nanodegree program.

Project Details

The tasks of this project are as follows:

- Gathering data
- Assessing data
- Cleaning data

Gathering Data

The data for this project consist on three different dataset that were obtained as following:

- **Twitter archive file:** I managed to read the data stored in the file 'twitter-archive-enhanced.csv'. I stored it in a DataFrame called 'df_twitter'. The data has many issues that will be cleaned and resolved later.
- **The tweet image predictions:** I have downloaded the file and open it using the panda library and stored it in a Data Frame called ‘df_image’.

- **Twitter API & JSON:** Using the list of tweet_id's in dataframd 'twitter_archive', I made a loop through each tweet and query Twitter's APIs with the tweet ID to get each tweet's JSON data. Then, I retrieved the required data ('favorite_count', 'retweet_count', 'followers_count', 'favourites_count', 'created_at') and store it in a list called 'df_api'.

Assessing Data

After gathering the data and storing them in DataFrames, the following step was assessing the data for quality and tidiness. Data were assessed programmatically and visually.

Quality:

df_twitter dataframe:

- tweet_id is an integer
- timestamp and retweeted_status_timestamp are currently of type 'object'
- name has values that are the string "None" instead of NaN
- Data contains retweets (ie. rows where retweeted_status_id and retweeted_status_user_id have a number instead of NaN)
- Some ratings with decimals such as 13.5/10, 9.5/10 have been incorrectly exported as 5/10 (in addition to other numbers with decimals such as 11.26 and 11.27).

df_image dataframe:

- tweet_id is an integer
- p1,p2 and p3 have unnessary underscore instead of space.
- drop duplicate jpg_url.

df_api dataframe:

- rename id to tweet_id so can merge later.

Tidiness Issues

df_twitter dataframe:

- 1 variable (dog stage) in 4 different columns (doggo, floofer, pupper, and puppo)

df_api dataframe and df_image dataframe:

- merging of all 3 data file into master file on the tweet_id

Cleaning Data

This part of the data wrangling was divided in three parts: Define, code and test the code. These three steps were on each of the issues described in the assess section.