Data Wrangling Report

Introduction

This report is written for the Project 4 of the Udacity Nanodegree Fundamentals in Data Analytics.

Project Instructions

Gather each of the three pieces of data as described below in a Jupyter Notebook titled wrangle_act.ipynb:

- The WeRateDogs Twitter archive: twitter_archive_enhanced.csv

- The tweet image predictions (image_predictions.tsv) hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

- Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count. Note: do not include your Twitter API keys, secrets, and tokens in your project submission.

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues. Detect and document at least eight (8) quality issues and two (2) tidiness issues in your wrangle_act.ipynb Jupyter Notebook.

Clean each of the issues you documented while assessing. Perform this cleaning in wrangle_act.ipynb as well. The result should be a high quality and tidy master pandas DataFrame (or DataFrames, if appropriate).

Wrangle Report

Gathering

Collecting information from different sources and in different formats could be quite challenging if you are just starting programming. The first two sources (csv file and tsv file) was pretty simple to gather. Connecting to Tweeter API was more difficult because it was something I am not familiar with and getting the code to read each line as a single row and extract only the 3 columns that we needed (tweet_id, retweet_count and favorite_count) required a hugue amount of effort and time.

Assessing

Using visual and programmatic assessment is really useful to get to know he data you are working with and detect quality and tidiness issues. In this datasets, I identify the following issues:

Quality

df table

- Contains retweets, that are not required to this analysis
- Contains replies, that are not required to this analysis
- Columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp have a lot of missing values and will not be required because retweets are not being used.
- Column source should be Category: Twitter for iPhone, Vine, Twitter Web Client, TweetDeck
- 745 dogs have None as a name, 55 dogs have 'a' as a name
- Review value extracted from text to rating_numerator and rating_denominator
- Column timestamp has +0000 extra
- Column timestamp datatype should be datetime instead of object

image_pred table

- Column img_num is not useful

Tidiness

df table

- Dog stage variable is in 4 columns, instead of one

df_api table

- retweet_count and favorite_count from df_api could be part of df table

image_pred table

- Merge p1 column with True values in p1_dog to df_clean

Cleaning

In order to perform a high quality and tidy dataframe, a lot of different pandas functions and methods must be used. This is because there were all sort of quality and tidiness issues to be solved before we can analyze our dataset properly to get consistent conclusions.