Wrangler and Analyze Data

Introduction

This project is part of Udacity's Data Analyst Nanodegree. This project is focused on wrangling data from the WeRateDogs Twitter account that rates people's dogs with humorous commentary. These ratings always have a denominator of 10 and the numerators are usually greater than 10. The goal is to gather data from a different sources and formats (csv, tsv, and json files), assessing data quality and tidiness, then perform cleaning data to create high quality and tidy pandas DataFrames for insight analyses and visualization.
This report briefly describes my wrangling efforts.

Project details

- Gathering data
- Assessing data
- Cleaning data

Gathering data

Data are from 3 sources:

- **Twitter Archive file:** the twitter_archive_enhanced.csv. contains tweet IDs and their tweets.
  This file was provided by Udacity and downloaded manually.
- **Additional Data via the Twitter API and JSON:** by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data file into tweet_json.txt file. Then, each tweet's JSON data was written to its own line. I created a dataframe to read this txt file line by line into a pandas DataFrame with the columns of my interest such as the tweet ID, retweet count, and favorite count.
- **The tweet image predictions:** contains what breed of dog is present in each tweet according to a neural network. This file image_predictions.tsv is hosted on Udacity's servers and downloaded programmatically using the Request library and URL information.

Assessing data

In this process, it allows us to identify quality and tidiness issues. Low quality data are considered as dirty data and have content issues. Untidy data are considered as messy data and are structural issues. We need to assess at least eight quality issues and at least two of structural issues.

Two types of assessment:

- **Visual assessment** is when you review data for completeness, validity, accuracy, and consistency. You can perform the visual review through Excel or Google Sheets.
- **Programmatic assessment** is when you use code methods in pandas such as info, head, tail, sample, describe, value_counts, and other various methods of indexing and selecting data.

List of issues:
Data Quality Issues:
twitter_1 dataframe:

1. retweeted_status_id and retweeted_status_user_id contains some numbers instead of NaN as majority.
2. timestamp and retweeted_status_timestamp are both type 'object'.
3. source is in HTML format with an a and \a tags before and after the text.
4. Some names with values of 'None' instead of NaN.
5. Some names are inaccurate like "a", "the", "an", and "quiet".
6. There is one name as "O" instead of "O'Malley".
7. Incorrect values in rating numerators. The current pipeline captures incorrect values when rating numerators contain decimals.
8. There are several columns in this dataframe are unnecessary and not needed for analysis.

df_tweet2 dataframe:

1. There are 23 missing tweets compared to the twitter_1 dataframe. By checking the collected status data on tweets, the 23 missing tweets are resulted to 'No status found with these IDs.'

images dataframe:

1. The first letter on the labels in p1, p2, p3 are inconsistent with upper case and lower case.
2. There are three dog predictions.
3. There is a total of 2075 images and there are 2356 tweets in the twitter_1 dataframe.
   This probably means not all 2356 tweets had pictures.

Tidiness Issues:

twitter_1 dataframe:

1. df_tweet2 dataframe and images dataframe should be combined with the twitter_1 dataframe since they are information about the same tweet.
2. Four different columns (doggo, floofer, pupper, and puppo) with 1 variable.

Cleaning data

Data wrangling process steps are define (instruction list), code (extract, drop, isnan, islower, replace, and etc.) and then test the code to assure the cleaning operations work correctly.

Very first step is to create a copy of the original dataframes and then I merged the three dataframes since they are the information about the same tweet. The combined and cleaned data is saved as twitter_archive_master.csv

Conclusion

Generated reports as part of this project.

- wrangler_act_ipynb: code for gathering, assessing, cleaning, analyzing and visualizing data.
- wrangler_report_pdf: documentation for data wrangling steps for gather, assess, and clean.
- act_report_pdf: documentation of analyses and insights of final data.

List of libraries, applications, Twitter access, and resources to successfully completed the operations.

- Jupiter Notebook to run wrangler_act_ipynb
- Developer account to get approval access and get the keys and tokens to query from Twitter API.
- tweepy import OAuthHandler
- tweepy
- request
- pandas
- json
- matplotlib.pyplot
- %matplotlib inline
- seaborn
- re
- datetime
- functools reduce

Resources:

https://www.quora.com/What-does-t-do-in-Python
https://www.pythoncentral.io/introduction-to-tweepy-twitter-for-python
https://stackoverflow.com/questions/47334968/pandas-keyerror-value-not-in-index
https://stackoverflow.com/questions/23668427/pandas-three-way-joining-multiple-dataframes-on-columns
https://developer.twitter.com/en/docs/ads/creatives/api-reference/tweets.html
https://stackoverflow.com/questions/47612822/how-to-create-pandas-dataframe-from-twitter-search-api
https://docs.oracle.com/en/cloud/paas/integration-cloud-service/icstw/getting-invalid-or-expired-token-error-response.html
https://stackoverflow.com/questions/13635215/convert-a-string-such-that-the-first-letter-is-uppercase-and-everythingelse-is-l
https://stackoverflow.com/questions/32364127/store-function-result-into-variable
https://www.geeksforgeeks.org/apply-uppercase-to-a-column-in-pandas-dataframe