Data Wrangling Report

Introduction

This project is a data wrangling project, which mainly focus on fixing the data quality and tidiness issues using python.

Data Gathering

1. `twitter_archive`: The WeRateDogs Twitter archive, which is provides by the Udacity Course and I use pd.read_csv() to import them into dataframe.
2. `image_predictions`: The tweet image predictions, i.e., what breed of dog (or other objects, animal, etc.) 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 requests library and the provided url.
3. `tweet_data`: 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 is written to its own line.

Data Assessing

Inspecting data set for two things: data quality issues and lack of tidiness

- Quality Issues means content issues like missing, duplicate, or incorrect data
- Untidy Data has specific structural issues

In addition, four dimensions of data quality assessment help me guide the thought process while assessing the data. For example, Completeness: are there any missing data in specific rows or columns? Validity: are there any records not correct due to any reason? Accuracy: are there any extreme data or unusual data? Consistency: are they keep the consistence of scale standard or data type?

Tidiness Issues

1. Columns 'doggo', 'floofer', 'pupper', 'puppo' in twitter_archive should belong to one colomn – stage
2. The tweet_data table need to merge into the twitter_archive table.

Quality Issues

< twitter_archive >

1. Some columns have huge amount of missing values, for example, "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp". Since I don't need in_reply and retweet data in this project, I prefer to delete those columns directly.
2. The variable "expanded_urls" also has few missing values, which means some records had no images. Any ratings without images should not be taking into account.
3. The datatype of "timestamp" is not correct.
4. Optimize the source contents for human reading habit; change the long URL links to certain words.
5. The standard for "rating_denominator" is 10, but it includes some other numbers, which could be the mis-parse.
6. The "rating_numerator" also has some incorrect values.
7. Some dog names are incorrect.
8. The dog names are sometimes first letter capital but sometimes not. Keep the name format consistent.

< image_predictions >

1. The columns' names are not clear and straightforward such as p1, p2.
2. The prediction dog breeds involve both uppercase and lowercase for the first letter.

Data Cleaning

- Tidiness Issue 1: Create a new variable – 'stage' to show the four dog stages, drop the four columns, and fill the empty with NaN.

- Tidiness Issue 2: Merge the tweet_data into the twitter_archive using inner join.

- Quality Issue 1: Remove all the unnecessary columns directly ('retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'in_reply_to_status_id', 'in_reply_to_user_id', 'in_reply_to_user_id).

- Quality Issue 2: Remove the records with no images information ('expanded_urls' is NaN).

- Quality Issue 3: Change the datatype of 'timestamp' to datetime.

- Quality Issue 4: Optimize the source content by 'Twitter for iphone', 'Vine - Make a Scene', 'Twitter Web Client', and 'TweetDeck'.

- Quality Issue 5: 10 is the default value of ‘rating_denominator’, then correct the wrong values based on the corresponding text information.

- Quality Issue 6: Correct the ‘rating_numerator’ values from the text information.

- Quality Issue 7: Change the frequent incorrect dog name to None.

- Quality Issue 8: Capitalize the first letter of dog name for consistence.

- Quality Issue 9: Change the column names for better readability in image_predictions.

- Quality Issue 10: Capitalize the first letter of first prediction in image_predictions (I could do that for all the predictions, but I decide to only apply to the first prediction since this variable is the important one).

Finally, I conduct a final test for the datasets and store the twitter_archive_clean to the file ‘twitter_archive_master.csv’.

Reference

- Twitter API Guide: https://www.slickremix.com/docs/how-to-get-api-keys-and-tokens-for-twitter/

- Tweepy documentation: https://media.readthedocs.org/pdf/tweepy/latest/tweepy.pdf

- WeRateDogsTwitter: https://twitter.com/dog_rates?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor