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 - 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 column – stage

2. The tweet_data table and image predictions table need to merge into the twitter_archive table.

Quality issues:

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". I perfer to delete those columns directly,since not needed.

2. The varaible "expanded_urls" also has few missing values. Any ratings without images should not be considered.

3. The datatype of "timestamp" is incorrect.

4. Change the long url links to certain words.

5. The standard for "rating_denominator" is 10, but it includes some other numbers.

6. The "rating_numerator" also has some incorrect values.

7. Remove all invalid dog names.

8. Change the column names for better readability in twitter_archive_clean and image_predictions_clean.

9. Capitalize the first letter of first prediction.

Data Cleaning

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

2. Tidiness Issue 2: Merge the tweet_data and image_predictions table into twitter_archive table using inner join.

3. 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) and their related rows.

4. Quality Issue 2: Remove the records with no images information (‘expanded_urls’ is NaN).

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

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

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

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

9. Quality Issue 7: Remove all the incorrect dog name.

10. Quality Issue 9: Change the column names for better readability.

11. Quality Issue 10: Capitalize the first letter of first prediction.

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

References:

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