Report On Data Wrangling Project.

By Caleb Adelaitar

Introduction

The dataset I worked with is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Data gathering

I began my project by downloading 'twitter-archive-enhanced.csv' manually. This data was provided by Udacity. Using the Udacity's server's request library, I downloaded 'image-predictions.tsv' programmatically. I then wrote it into image _predictions.tsv.

'Twitter_data', my third dataset which I accessed by downloading Twitter's JSON data using the tweepy library. I began by extracting a list of tweet ID from the 'twitter-archive-enhanced.csv' file, then looped through each ID and query Twitter's API with the ID to get each tweet's JSON data.

I then recorded the data in a text file named 'tweet-json.txt', with each tweet's data written in a new line. After the query was completed and all data was written in the text file, I read the text file line by line, obtained each tweet's information using the json library, and appended the information into an empty list. I finally converted the list of dictionaries into pandas DataFrame and saved it into 'twitter_data'

Assessing Data

Once the data was gathered, I assessed the data on two factors; quality and tidiness issues.

Under the data Quality I looked out for the following;

1. Completeness
2. Validity
3. Accuracy
4. Consistency

Three requirements were met under the Tidiness issues

1. Each variable forms a column
2. Each observation forms a row
3. Each type of unit forms a table

Quality Issues

1. Some dog names were incorrect and needed to be dropped.

2. Some dogs had more than one dog stages.

3. Erroneous datatypes which include;

    - Timestamp column is not of the correct datatype
    - Tweet_id column should be an object not integer
    - convert source to datatype categoty

4. Source column is in HTML-formatted string, not a normal string

5.Columns not being used should be dropped

6.Some images are not pictures of dogs

7. The dog rates should be standardized

8.Keep original ratings that have images

Tidiness Issues

1. The twitter api table and image prediction table should be merged to twitter archive table
2. Create one column for the various dog types: doggo, floofer, pupper, puppo.