Wrangle Report

By Ibrahim Badr

July-2020

This is like an internal documentation for the data wrangling process, that describes the wrangling effort.

Brief Introduction:

This is the second project in the Data Analyst professional track from Udacity in collaboration with ITIDA.

The dataset that was wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

The wrangling process:

The data wrangling process is divided into 3 steps:

- ### Gathering data

The data was gathered from 3 different sources:

1. **Twitter archive file:** the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.

2. **Image prediction file:** This file (`image_predictions.tsv`) is 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

3. **Twitter API & JSON.txt:** This was a query of Twitter API for each tweet's JSON data using Python's (tweepy) library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file.

- ### Assessing data

After gathering each of the above pieces of data, the data was visually and programmatically assessed for quality and tidiness issues. The target was to detect and document at least eight (8) quality issues and two (2) tidiness issues.

There are some important methods to use when assessing the data.

(.info(), .head(), .tail(), .sample(), .value_counts(), etc……)

- **Summary of issues after assessing the data:**

a. quality issues:

1. Twitter Archive

- 'tweet_id' should be a string not an integer.
- 'timestamp' should be date-time format not a string.
- 'retweets' need to be excluded as we need only the original tweets.
- 'expanded_urls' has 59 missing values.
- 'name' some names are not valid as a dog names like(such, my, a, the, ...).
- 'in_reply_to_status_id', 'in_reply_to_user_id ', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp' are not useful for the analysis and should be removed.
- 'name' dog names have upper case in the first letter.
- 'source' can be modified to be easy to read and used in analysis.

2. Image Prediction

- 'tweet_id' should be a string not an integer.
- 'jpg_url' has (66) duplicated values that should be dropped.

3. Twitter API

- 'tweet_id' should be a string not an integer.

b. tidiness issues:

- Dog stages combined into one column rather than four.

- Create one column for image prediction and another for confidence level and delete columns p1, p2, p3

- all three datasets needs to be merged into one using inner join.

• Cleaning data

The final step in the wrangling process is cleaning data, now it is the time to clean our data to make it ready for our analysis and visualization by following three important steps:

i. **Define**
ii. **Code**
iii. **Test**