Data Wrangling Report

By Amr Mohamed Helal

August2020

This report illustrates the main data analytics process steps to wrangle Twitter “We Rate Dogs” data.

Data Gathering:

In this step we collected data from different resources:

1. twitter-archive-enhanced.csv file which is downloaded manually and then imported to our project notebook using “pandas.read_csv()”.
2. image-predictions.tsv file which is exported from a given web page link using “requests.get()” and then imported to our notebook using “pandas.read_csv()”.
3. tweet-json file which is extracted using tweepy API and then imported at out notebook using “json.loads()”.

Data Assessment:

In this step we performed both quality and tidiness assessment on our 3 data frames visually and programmatically as follows:

1. Quality Assessment:

    a. **archive_df**

        1. Should include only original ratings tweets (remove retweets and replies)
        2. Missing values at "expanded_urls" are for tweets without photos (could be dropped safely)
        3. Most of records are null at these columns "in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp" (could be dropped)
        4. Datatype of "tweet_id" is int (should be str format)

    b. **image_predictions_df**

        1. Column name of "p1, p1_conf, p1_dog ..." are not descriptive
        2. Datatype of "tweet_id" is int (should be str format)
        3. Tweets with no images should be dropped (total tweets at archive_df more than images at image_predictions_df)

    c. **api_df**

        1. Most of records are null at these columns "contributors, coordinates, geo, in_reply_to_screen_name, in_reply_to_status_id, in_reply_to_status_id_str, in_reply_to_user_id, in_reply_to_user_id_str, place" (could be dropped)

2. Tidiness Assessment:

    a. **archive_df**

        1. Column names are values for "doggo, floofer, pupper, puppo"

    b. **api_df**

        2. Multiple variables are listed in one column "display_text_range" (should be splitted into 2 columns "display_text_min & display_text_max")

Data Cleaning:

In this step we proceeded with Define, Code, Test process to define each solution we worked on to solve issues discovered in assessment step with 3 outcome data frames:

1. archive_df_clean
2. image_predictions_df_clean
3. archive_df_clean

Data Storing:

In this step we stored the final cleaned data frames in CSV files format as below:

1. Twitter_archive_master.csv
2. Twitter_image_predictions_master.csv
3. Twitter_api_master.csv