wrangle_report

February 9, 2021

1 Project 4: "WeRateDoge" Data Wrngling

This project is wrangle "WeRateDogs". The main followed steps to wrangle them are the following:
1. Gather Data 2. Assess 3. Clean

1.1 1. Gather Twitter Data Information:

There are three resources the data were gathered from: 1. Given `twitter-archive-enhanced.csv` file - This file provides lots of information about each tweet, as its text, the related tweeting time, the dog rating and the stage. This file will be read from the current working directory 2. The retweets and favorite counts of each tweet are not in `twitter-archive-enhanced.csv` file. Therefore we will use the given tweet_id in the file to retrieve the counts using Twitter API 3. Additionally a hosted `image-predictions.tsv` file give us the top three breeds predictions of each dog image for the related tweets. This file is downloaded programtically from the given URL

1.2 2. Assess Datasets

We will need to see datasets samples, preview theier information and descriptions to provide our assessment #### Quality Content Issues: Completeness, Validity, Accuracy, Consistency `archive_df` **Table**: The dataframe of `twitter-archive-enhanced.csv` file 1. Remove retweets rows and keep original tweets only 2. Drop `retweeted_status_id`, `retweeted_status_user_id` and `retweeted_status_timestamp` columns as they contian only None for the original tweets 3. Change `tweet_id` from int64 to object 4. Change `timestamp` type from object to datetime 5. Remove `in_reply_to_status_id` and `in_reply_to_user_id` columns as they present only non null values in 3.3% of the entire dataset and will not help in any future analysis 6. Correct the wrong values in `rating_denominator` and `rating_numerator` columns 7. Change the datatypes for `rating_denominator` and `rating_numerator` to floats. As the correct ratings contain float values 8. Replace the lowercase names with "UNKNOWN" as they are for stopwords not the correct names

`image_df` **Table**: The dataframe of `image-predictions.tsv` 9. Change `tweet_id` from int64 to object 10. Correct the names capitalization in `p1`, `p2` and `p3` by converting them all to lowercase

`tweets_json_df` **Table**: The resulted dataframe of gathering tweets infromation from Twitter and keeping only the important information 11. Remove extra columns by keeping only `id`, `retweet_count` and `favorite_count` 12. Rename the `id` column to `tweet_id` to be consistent with the other tables `tweet_id` column 13. Change `tweet_id` from int64 to object

Tidiness Each variable forms a column. Each observation forms a row. Each type of observational unit forms a table.

1. Join the three tables using the unique column value `tweet_id`
2. Melt the four dogs stages `doggo`, `floofer`, `pupper` and `puppo` columns into `stage` column and drop the four columns from `archive_df` table
3. Add `rating` column resulted from dividing the correct `rating_numerator` on the correct `rating_denominator`, and remove these two columns

- Changing Tweet_id in the three tables will be done after merging the three tables to reduce redunduncy

1.3 Clean

All the steps mentioned in the ASSESSMENT were passed through in CLEANING, by defining each step, executing it using one of the found methods found on stackoverflow or github, aand then tested by previewing the dataframe info, describtion or the related column value_counts. After finishing the cleaning process, some statisics were visualized to see thier common values and the relations between the different varibles in the resulted clean dataframe.