We Rate Dogs

A data-wrangling project by Chinomnso Chinedum

Wrangle Report

Background

This project is part of the Udacity Data Analysis Nanodegree program. The main aim of the project was to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. Three datasets were used

- The twitter archive for WeRateDogs which was manually downloaded from Udacity's servers.
- Image predictions dataset that was hosted in Udacity's servers and downloaded programatically,
- Each tweet's retweet count and favorite ("like") count; gathered by querying Twitter's APIs using the tweet IDs in the WeRateDogs twitter archive dataset.

Issues Identified

A total of 12 quality and tidiness issues where identified. They include:

Quality issues

1. Some tweets are replies and retweets, not original tweets.
2. The source_column in `twitter_feed` has some unnecessary HTML code in it.
3. Incorrect rating numerators (decimal issues).
4. Wrong data types for the following columns:
    - a. `timestamp` column in `twitter_feed` has object data type instead of datetime64.
    - b. `tweet_id` in all the datasets
    - c. `source` in `twitter_feed` - should be changed to category data type.
5. Data about dogs and tweets are in the same dataset.
6. Multiple dog stages for some individual records.
7. `create_date` column in fave_retweet_data is not needed (after merge) - see below.
8. Incorrect Names in `dogs_df` .
9. Delete enteries without images

Tidiness issues ¶

1. Dog stage is split into 4 columns in the `twitter_feed` dataset
2. The predictions, confidence and dog are split into multiple columns in the `image_predictions` dataset.
3. Merge datasets to create a master dataset.

Solutions

All the 12 quality and tidyness issues identified were solved and here is how i solved them

Quality issues

1. Some tweets are replies and retweets, not original tweets: I dropped all the rows that have values in the `in_reply_to_status_id` , and `retweeted_status_id` columns. Then, I dropped the `in_reply_to_status_id` , `in_reply_to_user_id` , `retweeted_status_id` , `retweeted_status_user_id` , and the `retweeted_status_timestamp` columns themselves.
2. The source_column in `twitter_feed` has some unnecessary HTML code in it: I used str.replace with some regex to remove the HTML code in `source` column
3. Incorrect rating numerators (decimal issues): I used the with, display method to confirm that out of range rating numerators are errors. 4 cases were identified as errors and 1 case was an outlier and not an error. For the error cases, i used `.loc()` method to update rating_numerator values.
4. Wrong data types for the following columns:
    - a. `timestamp` column in `twitter_feed` has object data type instead of datetime64.
    - b. `tweet_id` in all the datasets
    - c. `source` in `twitter_feed` - should be changed to category data type. Here, I used the astype() function to change the datatype for each identified column.
5. Data about dogs and tweets are in the same dataset: This issue was addressed primarily to aid the overall cleaning of the dataset. To address this issue, I created a new dataset for dogs by copying out the relevant dog columns i.e; `tweet_id` , `name` , `doggo` , `floofer` , `pupper` , `puppo` , `rating_numerator` , `rating_denominator` . Then, I dropped dog specific columns i.e; `name` , `doggo` , `floofer` , `pupper` , `puppo` , `rating_numerator` , `rating_denominator` .
6. Multiple dog stages for some individual records: I created a new column for that identifies unknown and multiple dog stages, as well as the correct ones using the `np.select()` function.
7. `create_date` column in fave_retweet_data is not needed (after merge): Droped the column
8. Incorrect Names in `dogs_df` : I confirmed the incorrect names and used `str.match()` fuction to replace incorrect names with `None`
9. Rows without images in `image_predictions` dataset: Dropped the image predictions without images.

Tidiness issues

1. Dog stage is split into 4 columns in the `twitter_feed` dataset: Dropped the 4 columns since they had been collapsed into one.
2. The predictions, confidence and dog are split into multiple columns in the `image_predictions` dataset: I collapsed the `p1` , `p2` , and `p3` columns into prediction round, organized the values in `p1` , `p2` , and `p3` into a single column `breed_predictions` , collapsed the `p1_conf` , `p2_conf` , and `p3_conf` into a single column `confidence` and melted the `p1_dog` , `p2_dog` , and `p3_dog` into column `dog` .
3. Merge datasets to create a master dataset: Used the pd.merge() function to create a master dataset

Reflection

I did not anticipate how challenging this project was going to be, but i appreciate how much learning that completing the project made me experience. I was able to query Twitter's API, and this was the first time i was doing that by myself. I also ensured i used functions and methods that I was not familiar with like `np.select` and `str.match`. It was also very useful to get more practice with using `seaborn` for data visualisation. I found the practice of define, code, test to be very useful in organising my wrangling efforts. It also made sense to document all the issues identified first, before going on to solve them. In all, I am very happy about the technical things i learned, and the knowledge on best practices that i gained while completing the project.