wrangle_report

June 28, 2022

0.1 Reporting: wrangle_report

In my Data Wrangling process, I followed three main steps: 1. Gathering 2. Assessing 3. Cleaning

In the first step: Gathering; - Dataset 1: I loaded the provided Twitter Archive Data into a Dataframe. - Dataset 2: I used requests library to download the tweet image prediction file using the provided url.I then loaded it into a pandas dataframe. - Dataset 3: I queried the Twitter API to gather additional data on the tweets. This also involved reading the tweet_json.txt into a dataframe line by line.

In the second step: Assessing; - I carried out both visual and programmatic assessments and documented 8 quality issues and 2 tidiness issues.

Here are the documented issues: **Quality issues** ##### Dataset 1: Twitter Archive (twitter_archive_df) 1. Missing data. More than 60% of values missing in these columns (in_reply_to_status_id, in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp). More missing values in expanded_urls column. Also, the source column is unnecessary.

2. The column, timestamp, is object/string datatype instead of datetime. tweet ids are also integers instead of object/string.
3. Tweets beyond August 1st, 2017 are not necessary as they'll not be used.
4. The dog "stages" (doggo, floofer, pupper, and puppo) are not all correct. eg. Dog id 25 is a floof a.k.a floofer but dog name is None.
5. There are retweets when we only want original tweets that have images.

Dataset 2: The Tweet Image Predictions (image_predictions_df)

6. tweet_id (s) are integers instead of object/string data type.

Dataset 3: Additional data from Twitter API (tweets_df)

7. Missing data. More than 60% of values missing in these columns (quoted_status,quoted_status_id_str, quoted_status_permalink,quoted_status_id,retweeted_status,contribu place, 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).

- Also these columns are unnecessary (display_text_range, entities, full_text, truncated, id_str, is_quote_status, source, truncated, lang).

8. Both id and id_str columns contain the same data. One is unnecessary. Also, the retained should be changed to tweet_id to be consistent with the data in the other tables.

Tidiness issues ##### Dataset 1: Twitter Archive (twitter_archive_df) 1. The dog stages (doggo, floofer, pupper, puppo) are in different columns instead of 1 column, dog_stage. 2. All the three tables should be merged into one.

In the third step: Cleaning; - I first made copies of the data so as to reserve the original data. - I then cleaned the above documented issues using the Define,Code and Test Frame.

I then made copies of the master cleaned dataset and saved into a csv file.

Having completed these 3 steps of Data Wrangling successfully, I proceeded to analyzing and visualization. Here, I documented 3 insights and had 2 visualizations.

The highlighted Insights were: 1. The most retweeted tweets also have very high favourite count. They are clearly quite popular.

2. The most retweeted dog stage is doggo

3. The more favourited a tweet is, the more likely it is to be retweeted again and again.

There is still more work to be done on the dataset, more assessing and cleaning then visualizations since the Data Wrangling Process can always be iterated.