wrangle_report

June 21, 2022

1 We Rate Dog: Data Wragling Report

The data went through the 3 wrangling stages

1. Data Gathering
2. Accessing the Data
3. Data Cleaning

Step 1: Data Gathering The following data gather and used for this project

1. Twitter-archive-enhanced.csv was given. It was manually download and the read into the jupyter's workspace programmatically
2. Image_predictions.tsv: This file which contains tweet images and image predictions was downloaded programmatically using the request library
3. tweet_json: This was manually downloaded from the Udacity's resource library as I was not able to open a twitter's developer account and get a twitter API. The file was programmatically read and necessary data such as tweet_id, favorite_ciunt and retweet_id was extracted.

Step 2: Assessing the Gathered Data

1. A visual inspection of the three datasets was first done in order to identify any quality and structural/ tidiness issues in the datasets
2. Next, I carried out a Programmatic inspection using .info(), .describe(), .head(), in order to identify issues with the datatypes, check for missing values, check for outliers amongst others

The following Quality and Tidiness issues were identified after programmatically assessing the data

Quality issues

1. Erroneous datatypes in these columns (tweet_id, rating_denominator,rating_numerator, in_reply_to_status_id, in_reply_to_user_id, timestamp, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, doggo, floofer, pupper, and puppo)
2. Errors in Dogs name
3. 181 records do not contain original tweets from WeRateDogs
4. Drop columns not needed for our analysis
5. Text column includes a text and a short link
6. Source column in the archieve dataset contains HTML-formatted string,this should be categorical
7. Some tweets have no image.
8. Some values in rating_numerator and rating_denominator seem to be in error or suspicious outliers

Tidiness issues

1. The twitter API table and the image prediction dataset should be merged to twitter_archive dataframe
2. The dog stage is being spread across 4 columns

Step 3: Data Cleaning A step by step process for cleaning the data was provided – Define, Code and Test. Each of the identified quality and tidiness issues were fixed in the following manner:

Correcting the Quality Issues

1. The erroneous data types were converted to the appropriate data type.
2. Typographical errors in dog names were fixed
3. Rows not containing original tweets were removed
4. Colunms such as ['in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'] not needed for the analysis were removed
5. Hyperlinks in tweets were removed
6. I archive datasets which contained HTML formatted string was formatted and extracted from source
7. I removed the tweets that had no images
8. Rating_numerators which seemed like an error was corrected

Correcting the Tidiness issues

1. The twitter API table and the image prediction dataset were merged with the twitter_archive dataframe
2. Merged the dog stage to one column and replaced ’None’ with ’NaN’

Step 4: Data Storage The cleaned master dataset was saved as CSV file named "twitter_archive_master.csv"