wrangle_report

April 27, 2022

1 Project:Wrangling Report

1.1 Data Source: WeRateDogs From Twitter Archive

1.2 Presented by: Paul Anugwom

1.3 Wrangling Process

The data source is from WeRateDogs Twitter archive The following process were carried out to complete this project:

1.4 Gathering of datasets

The Three datasets were gathered for this project. One dataset(twitter_archive_enhanced.csv) was provided by Udacity. I directly downloaded this file and uploaded the file.I then went ahead and read the data in pandas dataframe.Before doing this, I did import all necessary libraries required to complete this project.

For the second dataset,I used the Requests library to download the tweet image prediction (image_predictions.tsv),saved the file, and read it in pandas dataframe.

For the third dataset,I used the Udacity provided jason file,I have no access to tweeter to extract the file on mine own.I downloaded the file provided by Udacity and uploaded into the notebook.I read the file and converted it to pandas dataframe.

1.5 Assessing the Datasets

Using visual and programmatic methods, the following quality and tidiness issues were discovered:

Quality issues

1.Clean dog_stage column with multiple entry

2.Remove retweets and replies since we are interested in original ratings,and drop columns not required for analysis

3.Missing URLs in expanded_urls

4.Drop 66 duplicated jpg_url rows

5.Erroneus Numerator and denominator values in archive dataset

6.Erroneous datatypes in these columns (tweet_id,in_reply_to_status_id, in_reply_to_user_id, timestamp, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)

7.Convert dog_stage datatype to category

8.Erroneous name 'a','an,'the' in Archive dataset

9.Drop expanded urls and change tweet_id dtype to string

Tidiness issues

1.Pupper,Daggo,floofer and Puppo columns in archive should be on a column named 'dog_stage'

2.All dataset should be merged as they represents same observation unit

1.6 Cleaning/Storage of Dataset

Here I attended to the identified issues during assessment.I made a copy of the original datasets before performing the cleaning.After cleaning and merging the three datasets into one dataset.I saved the gathered, assessed, and cleaned master dataset to a CSV file named "twitter_archive_master.csv" for futher analysis and visualization.