wrangle_report

October 15, 2020

Abdullah Othman

0.1 Gathering data

- Data was collected from three different sources:

1. twitter-archive-enhanced.csv
    - This file holds the main data about WeRateDogs page tweets
    - this file was downloaded manually from udacity servers then uploaded to the notebook
2. image-predictions.tsv
    - This was generated by a neural network that predicts that whether the images on tweets twitter-archive-enhanced.csv are for a dog or not then predicting dog breed if the image is for a dog
    - It was downloaded programmatically from Udacity's servers too
3. tweet_json.txt
    - This file holds extra information about each tweet in twitter-archive-enhanced.csv file: Number of retweets of and number of likes on this tweet
    - It was queried from Twitter API After I created a twitter developer account

- All the three files were loaded into the jupyter notebook and saved as pandas DataFrames

0.2 Assessing

0.2.1 Gatherd data was assessed to determine if it has any Quality or Tidiness issues, the following issues were found:

- Quality:

1. Compleetness issues: - The data had multiple missing values e.g. Name of dogs - Some columns was not significant for the sake of analysis

2. Validity issues: - Some tweets is not meeting the criteria of the rating tweet, the criteria is: 1. It must be an original tweet, Not a reply or a retweet 2. The tweet must have images of dogs 3. The tweet must have a rating for the dog in the image

3. Accuracy issus: - Rating values of some data was inacuarte

4. Consistency issues: - Data types of timestamps and tweets IDs were not right so we can not deal with them in that form - Some Nan values were not presented as Nans but 'None' strings - Tidiness: - Some column headers were values not variables e.g column headers for dog stage - The three datasets are representing only one observational unit

0.3 Cleaning

0.3.1 After assessing the data and detecting its issues, I started the process of cleaning trying to solve each issue

- 1. Started by deleting tweets that doesn't meet the criteria or have important missing values
- 2. Fixing inacurate rating values by sitting specifications of acurate ratings and modifying inacurate ratings to meet those specs and deleting values that can't be modified so not being outliers

- 3. Then fixing Tidiness issues by merging *value column headers* into one column and merging the three datasets into one master dataset that contains all the data needed
- 4. Then dealing with inconsistent values by altering there datatypes to the needed ones and turning "None" strings into Nans

0.4 Storing

The master dataset was saved and stored as csv file named: twitter_archive_master.csv