Wrangle Report

September 17, 2020

0.1 Introduction

As per wikipedia, Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. The data wrangling processes are performed in the juptyer notebook and overall processes are well documented with the use of proper comments and markdown text.

Data Set

The data set used are extracted from the twitter user @dogs_rates (WeRateDogs), whose three year period data from 2015 to 2017 were used to wrangle and analyze in the project. They have a comic way of rating dogs with denominator being primarily 10 while the dogs are rated more than denominator so rating can be 11, 12 or higher.

0.2 Wrangling Steps

- Gathering
- Assessing
- Cleaning

0.2.1 Gathering

There are three different raw data sets: - **Twitter Archive data:** This data is the twitter's archive of @dogs_rate user, named `twitter-archive-enhanced.csv` - **Twitter Api data:** This data was extracted using tweepy querying the twitter api. `twitter-json.txt` - **Image predictions:** This data was obtained manually from udacity.This data set has predicted the tweeted dog's breed with the help of deep learning's convolution neural network algorithm.`image-predictions.tsv`

0.2.2 Assessing

The gathered data are assessed using spreadsheet tool: Excel ,and programmatically using Pandas library. The data has both quality and tidiness issues that were detected, defined and cleaned. The quality and tidiness issues are reported as mentioned below:

Quality Issues

- `twitter_archive_enhanced`
  - `timestamp` is a string (i.e object) which should be `datetime` type.
  - Unnecessary retweet records.

  - Many records of the tweet's columns: doggo, floofer, pupper, puppo , do not contain any value excpet 'None'. and they are string data type. They could be of category data type.
  - The `source` column contain redundant HTMl informations.
  - The `name` column contain wrong names or literary words/characters like 'a', 'Bo','the',and 'an' .
  - The numerator and denominator contains the wrong values when compared with the `text` column.

- `image_predictions`

  - There are duplicated jpg_urls.
  - There is inconsisitency in the predicted string's letter case. Some are lower case while others are not.
  - The predicted dog breed is of `object` (i.e string) datatype, category type may be suitable.
  - The underscore symbol in place of space in the dog bredd's name

- `tweet-json`

  - Unwanted columns
  - There are 0 non-null data in `contributors` , `coordinates` and `geo` columns whereas only 1 non-null data in `place` columns.

Tidiness Issues

- `twitter_archive_enhanced`

  - The dog stages: `doggo`, `floofer`, `pupper`, `puppo` are in different columns.

- `twitter_json`

  - the `display_text_range` contains a list of numerical data

- `image_predictions`

  - There are three different predictions breeds and their confidence values in three different columns

- `All three`

  - Merge all three dataframe into a master dataframe

0.3 Cleaning

This process involved working massivley with pandas library. At first, all three differnt data sets were cleaned separately, then the necessary columns were taken from each of them to finally merge into a master data set.

Further,the cleaned data was saved to a CSV file into the local directory.

0.4 Visulaization

After the data wrangling process, the features of the data were analyzed and visualized.